close print preview

The New Palgrave Dictionary of Economics Online

Abbott, Edith (1876 – 1957)

P. Kerr

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Abbott, Edith; social policy; women, economic role of

Article

Social reformer, economic historian and a pioneer in America of the study of the economic position of women, Edith Abbott was born on 26 September 1876 in Nebraska, and graduated from the University of Nebraska in 1901. She enrolled in a summer session at the University of Chicago in 1902, attracting the attention of James Lawrence Laughlin and Thorstein Veblen, and on their recommendation returned to Chicago in 1903 on a fellowship in political economy, taking her PhD in 1905 with a dissertation on the wages of unskilled labour in the USA between 1850 and 1900 (Abbott, 1905). It was during this period at Chicago that she met Sophonisba Breckinridge who became her mentor and lifelong friend. In 1906, on a Carnegie Fellowship, she went to the LSE to carry out research on women in industry. In London she was influenced by the social reformers of the day, including Charles Booth and Sydney and Beatrice Webb. She returned to the USA in 1907 and taught political economy at Wellesley. In 1908 Breckinridge, now Director of Research at the newly established Chicago School of Civics and Philanthropy, invited her to become her assistant.

Abbott's work there involved her directly in action for the protection and education of juveniles and immigrants, for improvements in housing, and for the reform of correctional institutions. She also worked towards women's suffrage, the ten-hour law to protect women in employment, and the admission of women into trades unions. In the 1930s she was to become a staunch advocate of social insurance measures and the welfare state. Although sympathetic to the New Deal, she felt it to be entirely inadequate when it came to welfare policies.

Her publications ranged over a number of areas in social and public policy, and with Breckinridge, she was an influential proponent of the role of the state as the key element in any extensive programme of social welfare. The journal they jointly established in 1927, *Social Science Review*, was immediately recognized as a highly esteemed professional journal. Her main writings on economics were collected in her *Women in Industry* (1910), where a recurring theme was the distinction between the progress of ‘professional’ women (and the women's movements with which they were associated) and the relatively

unchanged position of working-class women.
After 1920, although social work came increasingly to dominate her time, Abbott continued her role as an applied economist. She was a member of the advisory committee of the ILO on immigration, and succeeded Breckinridge as Dean of the School of Social Studies Administration at Chicago. She remained in the post until 1942, and continued editing the *Social Science Review* until 1953. She died at the age of 80 at her family home in Grand Island.

Selected works

1905. Wages of unskilled labour in the United States, 1850–1900. *Journal of Political Economy* 13, 321–67.

1906. Industrial employment of women in the United States. *Journal of Political Economy* 14, 461–501.

1908. Study of early history of child labour in America. *American Journal of Sociology* 14, 15–37.

1910. *Women in Industry: A Study in American Economic History*. London: Appleton & Co; last reprinted in 1970.

1915. A forgotten minimum wage bill. *Life and Labour* 5, 13–16.

How to cite this article

Kerr, P. "Abbott, Edith (1876–1957)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000001> doi:10.1057/9780230226203.1851

close print preview

The New Palgrave Dictionary of Economics Online

Abramovitz, Moses (1912 – 2000)

Richard A. Easterlin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Abramovitz, M.; aggregate demand theory; business cycles; economic growth in the very long run; inventories; Kuznets cycles

Article

Born in Brooklyn, New York, Abramovitz was educated at Harvard (AB, 1932) and Columbia (Ph.D., 1939). He held faculty appointments at Columbia (1940–2, 1946–8) and Stanford University (1948–77) and was a member of the research staff of the National Bureau of Economic Research from 1938 to 1969. From 1942 to 1946 he worked as an economist for several organizations within the United States government. He was elected president of the American Economic Association in 1979–80.
Abramovitz's work, which was particularly influenced by Wesley C. Mitchell and Simon Kuznets, centres on the study of long-term economic growth and fluctuations in industrialized market economies. His first major contribution was an empirical study of business inventories that demonstrated the importance of inventory change in the shorter swings of the business cycle, and showed how the classification of inventories by stage of processing aided in the explanation of their behaviour (Abramovitz, 1950). From this, Abramovitz went on to the study of longer-term fluctuations, Kuznets cycles of 15 to 20 years duration, and formulated the most widely accepted interpretation of these cycles. Using Keynesian aggregate demand theory, Abramovitz developed a model linking Kuznets cycles to long swings in building cycles and demographic variables, and to shorter-term business cycles (Abramovitz, 1959a; 1961; 1964; 1968).
Contemporaneously with his work on fluctuations, Abramovitz made important contributions to long-term economic growth. He was one of the first to demonstrate that only a small share of long-term output growth in the United States was explained by factor inputs (Abramovitz, 1956). He documented and analysed the increasing role of government during long-term economic growth (Abramovitz, 1957; 1981) and directed and coordinated a comparative study of the post-war economic growth of a number of industrialized market nations (Abramovitz, 1979b; 1986). Finally, he challenged in characteristically perceptive fashion the facile linkage made by many economists between economic growth and improving human welfare (Abramovitz, 1959b; 1979a; 1982).

Selected works

1950. *Inventories and Business Cycles*. New York: NBER.

1956. Resource and output trends in the United States since 1870. *American Economic Review, Papers and Proceedings* 46(2), 5–23.

1957. (With V. Eliasberg.) *The Growth of Public Employment in Great Britain*. Princeton: Princeton University Press.

1959a. Long swings in U.S. economic growth. Statement presented to Joint Economic Committee of the Congress. Hearings before Joint Economic Committee of the Congress of the U.S. on *Employment, Growth and Price Levels*, Part 2, 11–66, 10 April.

1959b. The welfare interpretation of secular trends in national income and production. In M. Abramovitz et al., *The Allocation of Economic Resources: Essays in Honor of Bernard F. Haley*. Stanford: Stanford University Press.

1961. The nature and significance of Kuznets cycles. *Economic Development and Cultural Change* 9, 225–48.

1964. *Evidence of Long Swings in Aggregate Construction since the Civil War*. Occasional Paper No. 90. New York: NBER.

1968. The passing of the Kuznets cycle. *Economica*, 349–67.

1979a. Economic growth and its discontents. In *Economics and Human Welfare: Essays in Honor of Tibor Scitovsky*, ed. M. Boskin. New York: Academic Press.

1979b. Rapid growth potential and its realization: the experience of capitalist economies in the postwar period. In *Economic Growth and Resources. Proceedings of the Fifth World Congress of the International Economic Association*, vol. 1. London and New York: Macmillan.

1981. Welfare quandaries and productivity concerns. Presidential Address to the American Economic Association. *American Economic Review* 71, 1–17.

1982. The retreat from economic advance. In *Progress and its Discontents*, ed. G.A. Almond, M. Chodorow and R.H. Pearce. Berkeley: University of California Press.

1986. Catching up, forging ahead and falling behind. *Journal of Economic History* 46, 385–406.

How to cite this article

Easterlin, Richard A. "Abramovitz, Moses (1912–2000) ." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000002> doi:10.1057/9780230226203.0001

close print preview

The New Palgrave Dictionary of Economics Online

absolute and exchangeable value

John Eatwell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The notion of absolute (as distinct from exchangeable or relative) value arises in classical economics from the image of a given magnitude of output being distributed between the social classes. Ricardo posited that the value of the social surplus could be expressed in terms of labour regardless of how the surplus was distributed. But since changes in distribution affect exchangeable value, the value of the surplus will typically vary as distribution varies, even though its physical magnitude remains unchanged. In 1823 Ricardo concluded that ‘there is no such thing in nature as a perfect measure of value’.

Keywords

absolute and exchangeable value; Cairnes, F. E.; Cairnes, J. E.; class; classical economics; invariable standard of value; labour theory of value; Marx, K.H.; rate of profit; Ricardo, D.; Sraffa, P.; surplus

Article

No one can doubt that it would be a great desideratum in political economy to have such a measure of absolute value in order to enable us to know, when commodities altered in relative value, in which the alteration in value had taken place. (David Ricardo, 1823, p. 399n)

The idea that changes in the relative or exchangeable value of a pair of commodities might usefully be attributed to alterations in the ‘absolute value’ of one or the other of them will appear rather odd to anyone accustomed to thinking of the basic problem of price theory as being the determination of sets of relative prices, with any consideration of ‘absolute’ value being confined to problems in monetary theory and the determination of the overall price level. Since in neoclassical theory it is the *relative* scarcity of commodities, or of the factor services which are used to produce them, which is the key to relative price formation, no conception of ‘absolute’ value, that is, a price associated with the conditions of production of a single commodity, is either relevant or necessary.

Yet the notion of absolute value arose naturally within Ricardo's analysis of value and distribution. The central problem of classical theory is to relate the physical magnitude of surplus (defined as the social

output *minus* the replacement of materials used in its production and the wage goods paid to the labourers employed) to the general rate of profit and the rents in terms of which the surplus is distributed. The key image is the distribution of a given magnitude of output between the classes of the society. ‘After all’, as Ricardo put it, ‘the great questions of Rent, Wages and Profits must be explained by the proportions in which the whole produce is divided between landlords, capitalists, and labourers, and which are not essentially connected with the doctrine of value’ (1820, p. 194). Ricardo was able to sustain this ‘material’ view of distribution only in the *Essay on Profits*, and only there by the implicit device of a sector in which all inputs and all output consist of the same commodity, corn, which is also used to pay wages in the other sectors of the economy. In the corn sector the division of the product may be expressed in physical terms, and the rate of profit expressed as a ratio of physical magnitudes.

This clear and direct analysis is no longer possible once the strong assumption of a self-reproducing sector is dropped.

The need to express heterogeneous surplus (net of rent) and heterogeneous capital as homogeneous magnitudes in order to determine the rate of profit created the need for a theory of value. Ricardo's materialist approach led him to the labour theory of value. The quantity of labour embodied directly and indirectly in the production of a commodity is determined by the conditions of production of that commodity, or as Ricardo put it, by the difficulty or facility of production, and will change only when the technique changes. Hence the aggregates of social surplus and capital advanced may be expressed as quantities of labour, these quantities being invariant to changes in the distribution of social product. So the rate of profit is determined as the ratio of surplus (on the land last brought into use) to the means of production, including wages.

Once, however, the impact of changes in distribution on exchangeable value is taken into account the picture is far less clear. The value of social output, and of the surplus, measured in any given standard, will typically now vary as distribution varies, even though the physical magnitude of social output remains unchanged. The direct deductive relationship between wages, surplus, and hence, the rate of profit, is no longer self-evident, or indeed, evident at all. It was Ricardo's desire to restore clarity to his analysis which led to his search for an invariable standard of value (a standard in terms of which the size of the aggregate would not vary as distribution was changed) and for what Sraffa describes as ‘for Ricardo its necessary complement’, absolute value (Sraffa, 1951, p. xlvi).

The term ‘absolute value’ was used by Ricardo but once in the first edition of the *Principles* and occasionally in letters. It was clarified in the papers on ‘Absolute Value and Exchangeable Value’, written in 1823 in the last few years of his life. These were discovered in a locked box at the home of F. E. Cairnes, the son of the economist John Elliot Cairnes, in 1943, and published for the first time in Sraffa's edition of Ricardo's Works and Correspondence.

There are two versions of the essay. One, a rough draft, is written on odd pieces of paper, some of them the covers of letters addressed to Ricardo. The other is a scarcely corrected draft, written on uniform sheets of paper. This clean draft breaks off, unfinished.

The importance of the essay derives from the reinforcement it provides to that interpretation of Ricardo's theory of value and distribution which suggests that the problem of the determination of the relative values of commodities stemmed from Ricardo's desire to relate his image of the division of social product as a physical magnitude to the wages, rents, and rate of profit of a market economy. Ricardo was not interested for its own sake in the problem of why two commodities produced by the same quantities of labour are not of the same exchangeable value. He was, rather, concerned by the fact that as

distribution of social output *changes* exchangeable value *changes*, disrupting and obscuring an otherwise clear vision. It was this emphasis on the fact that *changes* in distribution lead to changes in exchangeable value, even though the quantity of social output and the method by which it is produced are unchanged, which led Ricardo into the intellectual cul-de-sac of the search for an invariable standard of value.
The absolute value of a commodity is the value of that commodity measured in terms of an invariable standard. An invariable standard of value may be found

> … if precisely the same length of time and neither more nor less were necessary to the production of all commodities. Commodities would then have an absolute value directly in proportion to the quantity of labour embodied in them. (Ricardo, 1823, p. 382)

Changes in the absolute values of commodities could then derive only from changes in the amount of labour embodied in them, and the value of social output would be invariate to its distribution.
Yet precisely because all commodities are not produced under the same circumstances, ‘difficulty or facility of production is not absolutely the only cause of variation in value, there is one other, the rise or fall of wages’ since commodities cannot ‘be produced and brought to market in precisely the same time’ (1823, p. 368). Hence Ricardo must conclude, rather sadly, that ‘there is no such thing in nature as a perfect measure of value’ (1823, p. 404) – there is no such thing as an invariable standard of value.
Marx (1883), who could not, of course, have seen the papers on Absolute and Exchangeable Value, was critical of Ricardo's absorption with the search for an invariable standard. The focus on changes in relative value obscured the fact that commodities do not exchange at rates proportional to their labour values (labour embodied). Yet Marx's attempt to restore clarity to the analysis of distribution by first determining the rate of profit as the ratio of quantities of labour, and then ‘transforming’ labour values into prices of production, encounters difficulties which derive from exactly the same source as those which bedevilled Ricardo – the difference in production conditions or ‘organic composition of capital’ of commodities.
The data of classical theory can be used to determine the rate of profit, as Sraffa (1960) has shown. But the determination cannot be ‘sequential’ – first specifying a theory of value and then evaluating the ratio of surplus to capital advanced by means of that predetermined theory of value. Rather the rate of profit and the rates at which commodities exchange must be determined simultaneously.

See Also

- Ricardo, David

Bibliography

Marx, K. 1883. *Capital*, vol. 3, London: Lawrence and Wishart, 1976.

Ricardo, D. 1820. Letter to J.R. McCulloch, 13 June 1820. In *Works and Correspondence of David Ricardo*, vol. 8, ed. P. Sraffa. Cambridge: Cambridge University Press, 1953.

Ricardo, D. 1823. Paper on ‘Absolute and exchangeable value’ (rough draft, and unfinished clean version). In *Works and Correspondence of David Ricardo*, vol. 4, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Sraffa, P. 1951. *Introduction to Works and Correspondence of David Ricardo*, vol. 1, Cambridge: Cambridge University Press.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

How to cite this article

Eatwell, John. "absolute and exchangeable value." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000004> doi:10.1057/9780230226203.0002

close print preview

The New Palgrave Dictionary of Economics Online

absorption approach to the balance of payments

David Vines

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The absorption approach to the balance of payments states that a country's balance of trade will only improve if the country's output of goods and services increases by more than its absorption, where the term ‘absorption’ means expenditure by domestic residents on goods and services. This approach was first put forward by Alexander (1952, 1959).

The novelty of this approach may be appreciated by considering the particular question ‘will a devaluation improve a country's balance of trade?’ The elasticities approach, popular when Alexander was writing, answers this question by focusing on the price elasticities of supply and demand for exports and imports. It holds that the devaluation will be successful if the price elasticities of demand for exports and imports are large enough so that the increase in exports sold to foreigners and the reduction in imports bought by domestic residents together more than offset the terms of trade loss caused by the devaluation. (A special case of this result is formalized in the Marshall–Lerner conditions.) The absorption approach argues, by contrast, that the devaluation will only be successful if it causes the gap between domestic output and domestic absorption to widen. In effect Alexander criticizes the elasticities approach for focusing on the movement along given supply and demand curves in the particular markets for exports and imports (a microeconomic approach), instead of looking at the production and spending of the nation as a whole which shift these curves (a macroeconomic approach).

Alexander's criticism of the elasticities approach is valid. But without further elaboration the absorption approach is unhelpful in rectifying the inadequacy. This is because, taken at face value, the absorption approach merely states an identity. Let the symbols, Y, C, I, G, X and M stand for output, consumption, investment, government expenditure, exports and imports respectively. Then the Keynesian income-expenditure *identity* states that

$$Y = C + I + G + X - M \tag{1}$$

which may be rewritten

$$X - M = Y - (C + I + G). \tag{2}$$

This *identity* states precisely that the trade balance will improve if output, Y, increases by more than absorption ($C+I+G$).

What is needed, and what Alexander helped to provide, is an analysis of exactly how output and absorption change, in response to a devaluation, and indeed in response to other developments in the economy. Such a gap was also being filled at the time by Keynesian writers (Robinson, 1937; Harrod, 1939; Machlup, 1943; Meade, 1951; Harberger, 1950; Laursen and Metzler, 1950; see also Swan, 1956). All of these authors grafted the Keynesian multiplier onto the elasticities approach. The resulting hybrid construct can be used to analyse the effects of a devaluation as follows. Suppose that the price elasticity effects do improve the balance of trade, $X–M$, by ‘switching’ expenditures towards domestic goods. Then these ‘expenditure-switching’ effects provide a positive stimulus to the Keynesian multiplier process, and drive up output Y and absorption $C+I+G$. Let x be the expenditure-switching effects on the trade balance of a devaluation of the currency by one unit, and let the overall effects of this devaluation on the trade balance be y. Let the propensity to consume be c, the tax rate be t and the propensity to import m, so that the Keynesian multiplier is $k=1/[1–c(1–t)+m]$. The increase in output resulting from the devaluation is kx and the increase in absorption is $c(1–t)kx$. And so

$$y = k[1 - c(1 - t)]x. \tag{3}$$

If the propensity to consume c is less than unity and the tax rate t is positive then absorption increases by less than output, and, as equation (3) shows the trade balance is improved by the devaluation. The above sketch shows how the combination of the elasticities approach and Keynesian theory is able to provide the needed analysis of how output and absorption change following a devaluation. And instead of describing the outcomes in terms of output and absorption, as Alexander did, it is possible to give a more conventional Keynesian description, which would proceed as follows. Since the multiplier $k=1/[1–c(1–t)+m]$ times the propensity to import m is less than unity, the increase in imports induced by the multiplier, mkx, is less than the positive ‘expenditure-switching effects’, x, and so the trade balance improves.

We can also show how output and absorption change after an ‘expenditure-changing’ adjustment of policy. For example, a one unit increase in government spending will cause output to increase by k whereas absorption increases by the sum of the increase in government expenditure and the induced increase in consumption $(1–t)ck$; the trade balance thus worsens by an amount z where

$$z = k - [1 + (1 - t)ck] = k - [1 - c(1 - t) + m + c(1 - t)]k = -mk. \quad (4)$$

Again this outcome can be described in the more conventional Keynesian way: high government expenditure drives up output by the multiplier, *k*, and sucks in imports of an amount *mk*.
The combination of the elasticities approach and Keynesian multiplier theory was used to produce a theory of economic policy for an open economy, which involved the pursuit of full employment as well as a satisfactory balance of trade as policy objectives (Meade, 1951; see especially Swan 1956). This theory can be stated just as well in terms of Alexander's absorption approach. For example an improvement in the balance of trade at full employment requires a reduction in absorption, without any change in output. It is obvious from the previous two paragraphs that this, in turn, requires both expenditure-switching policies and expenditure-changing policies, since both of these policies and influence output as well as absorption. Johnson (1956) put this point masterfully, and I now express it algebraically. Let the desired increase in the trade balance be *w*, let the required devaluation of the currency be α units and let the required change in government expenditure be β. Then from equations (3) and (4)

$$w = [1 - c(1 - t)]kx\alpha - mk\beta \quad (5)$$

whereas, since output is not to be affected,

$$0 = kx\alpha + k\beta \quad (6)$$

Solving for β from equation (6) and substituting into equation (5), nothing that $1–c(1–t)=1/k–m$, gives

$$w = [1 / k - m]kx\alpha + mkx\alpha = x\alpha.$$

Thus the required devaluation is simply $\alpha = w/x$ and substituting in equation (6) the required change in government expenditure is simply $\beta = –w$. This states what is obvious: government absorption must be reduced enough to release resource from domestic use – the expenditure-changing component of policy

– and the devaluation must ensure that these resources are actually used to improve the trade balance, rather than leading to a fall in domestic output – the expenditure-switching component of policy. Laursen and Metzler (1950) show that what is obvious must in fact be qualified. A more careful analysis would show that the positive expenditure switching effect of a devaluation on the trade balance is slightly smaller than the positive expenditure switching stimulus which devaluation imparts to the Keynesian multiplier process (whereas we have assumed both of these effects to be equal, and have denoted them by β). See also Harberger (1950) and Svensson and Razin (1983).

Modern balance of payments theory has carried criticisms much further than this. It has shown that the hybrid of the Keynesian multiplier and elasticities approaches is inadequate in providing a full analysis of how output and absorption change. First it does not deal with the inflationary effects of devaluation. But one way in which devaluation depresses absorption relative to output is through engendering rises in costs and prices which depress the real incomes (particularly real wages) of domestic consumers (Diaz Alexandro, 1966). Furthermore, devaluation may also engender a wage-price spiral so strong as to preserve the real incomes of domestic consumers, with the end result that prices rise by the full extent of the devaluation and there is no relative price change for the price elasticities effects to work on (Ball, Burns and Laury, 1977). In that case positive effects of devaluation on the trade balance can *only* emerge as a result of the effects of higher prices on absorption. (Higher prices lower the real wealth of consumers and perhaps also increase the tax burden if tax rates are progressive and not indexed with inflation.) Second, the multiplier-plus-elasticities analysis is not appropriate in analysing the effects of a devaluation not accompanied by any expenditure changing policy if the economy is at full employment, for in that case output cannot be expanded through the multiplier, and the effects of the devaluation must primarily work through the influence of inflation on absorption described above. Third, the multiplier-plus-elasticities analysis does not deal with monetary conditions. A devaluation, because it raises prices, may initially also cause higher interest rates which helps to curtail absorption. But if the improvement in the trade balance caused by the devaluation is allowed to lead to an expansion of the domestic money supply, then gradually interest rates will fall, absorption will rise, and the effects of the devaluation may turn out to be temporary. This issue has been analysed by the Monetary Approach to the Balance of Payments (Frenkel and Johnson, 1976; Kyle, 1976; McCallum and Vines, 1981). Alexander made many of these points in his articles whereas the authors cited at the end of the fourth paragraph tended to skate over them. For that reason his work prefigures much subsequent balance of payments theory.

In conclusion, the absorption approach provides a useful perspective from which to view the trade balance. But it must be supplemented by a theory both of what determines absorption and of what determines output. And of course, the absorption approach only deals with the trade balance; a full theory of the balance of payments requires a theory of capital account movements (and a discussion of how the exchange rate itself is determined).

See Also

- elasticities approach to the balance of payments
- monetary approach to the balance of payments

Bibliography

Alexander, S.S. 1952. Effects of devaluation on a trade balance. *International Monetary Fund Staff Papers*, Vol. 2, 263–78.

Alexander, S.S. 1959. A simplified synthesis of elasticities and absorption approaches. *American Economic Review* 49, 22–42.

Ball, J., Burns, T. and Laury, J.S.E. 1977. The role of exchange rate change in balance of payments adjustment – the United Kingdom case. *Economic Journal* 87, 1–29.

Caves, R.E. and Johnson, H.G., ed. 1968. *Reading in International Economics*. London: George Allen & Unwin.

Diaz Alexandro, C. 1966. *Exchange Rate Devaluation in a Semi-Industrialized Country: The Experience of Argentina 1955–1961*. Boston: MIT Press.

Frenkel, J.A. and Johnson, H.G., eds. 1976. *The Monetary Approach to the Balance of Payments*. London: George Allen & Unwin.

Harberger, A.C. 1950. Currency depreciation, income and the balance of trade. In Caves and Johnson (1968).

Harrod, R.F. 1939. *International Economics*. 2nd edn, Cambridge Economic Handbooks, VIII, London: Nisbet & Co.

Kyle, J.F. 1976. *The Balance of Payments in a Monetary Economy*. Princeton: Princeton University Press.

Laursen, S. and Metzler, L. 1950. Flexible exchange rates and the theory of employment. *Review of Economics and Statistics* 32, 281–99.

Machlup, F. 1943. *International Trade and the National Income Multiplier*. Philadelphia: Blakiston Co.

McCallum, J. and Vines, D. 1981. Cambridge and Chicago on the balance of payments. *Economic Journal* 91, 439–53.

Meade, J.E. 1951. *The Balance of Payments*. London: Oxford University Press.

Robinson, J.V. 1937. The Foreign Exchanges. In J. Robinson, *Essays in The Theory of Employment*, 2nd edn, 1947. Reprinted in *Reading in the Theory of International Trade*, ed. H.S. Ellis and L.A. Metzler,

Philadelphia: Blakiston, 1949, 83–103.

Svensson, L. and Razin, A. 1983. The terms of trade and the current account: the Harberger–Laursen–Metzler Effect. *Journal of Political Economy* 91, 97–125.

Swan, T.W. 1956. Longer run problems of the balance of payments. In *The Australian Economy, a Volume of Readings*, ed. H. Arndt and W.M. Corden, Melbourne: Cheshire Press, 1963. Reprinted in Caves and Johnson (1968).

How to cite this article

Vines, David. "absorption approach to the balance of payments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000008>
doi:10.1057/9780230226203.0003

close print preview

The New Palgrave Dictionary of Economics Online

acceleration principle

P.N. Junankar

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The acceleration principle holds that the demand for capital goods is a derived demand and that changes in the demand for output lead to changes in the demand for capital stock and, hence, lead to investment. The flexible accelerator, which includes both demand and supply elements, allows for lags in the adjustment of the actual capital stock towards the optimal level. The principle neglects technological change but has been used successfully in explaining investment behaviour and cyclical behaviour in a capitalist economy. Almost all macroeconomic models of the economy employ some variant of it to explain aggregate investment.

Keywords

acceleration principle; Aftalion, A.; aggregate demand; aggregate investment; business cycles; capital–output coefficient; Chenery, H. B.; Clark, J. M.; depreciation; derived demand; distributed lag accelerator; Eisner, R.; expectations; Haberler, G.; Harrod, R. F.; Harrod–Domar growth model; Marx, K. H.; Pigou, A.C.; technical change

Article

The acceleration principle has been proposed as a theory of investment *demand* as well as a theory determining the *supply* of capital goods. When combined with the multiplier, it has played a very important role in models of the business cycle as well as in growth models of the Harrod–Domar type. The acceleration principle has been used to explain investment in capital equipment, the production of durable consumer goods and investment in inventories (or stocks). In general, it has been used to explain aggregate investment, although it is sometimes used to explain investment by firms (micro-investment behaviour). The main idea underlying the acceleration principle is that the demand for capital goods is a derived demand and that changes in the demand for output lead to changes in the demand for capital stock and, hence, lead to investment. Its distinctive feature, then, is its emphasis on the role of (expected) demand and its de-emphasis on relative prices of inputs or interest rates.

The acceleration principle is a relatively new concept: it is possible to find its antecedents in Marx's *Theories of Surplus Value*, Part II (1863, p. 531). Amongst the earliest exponents of the acceleration principle is Albert Aftalion in *Les Crises périodiques de surproduction* (1913). Later contributions by J. M. Clark (1917), A.C. Pigou (1927) and R.F. Harrod (1936) discussed the acceleration principle both as a determinant of investment and in its role in explaining business cycles. Haberler (1937) provides a fairly comprehensive account of the acceleration principle up to that date. Since then the contributions by Chenery (1952) and Koyck (1954) provide important extensions and developments of the theory. In recent years work by Eisner (1960) has employed the acceleration principle in econometric work. Almost all macroeconomic models of the economy employ some variant of the acceleration principle to explain aggregate investment.

Underlying the acceleration principle is the notion that there is some optimal relationship between output and capital stock: if output is growing, an increase in capital stock is required. In the simplest version of the acceleration principle,

$$K_t^* = vY_t$$

where K_t^* is planned capital stock, Y_t is output and v is a positive capital–output coefficient. On the assumption that the capital stock is optimally adjusted in the initial period (that is $K_t = K_t^*$ where K_t is the actual capital stock) an increase in output (or planned output) leads to an increase in planned capital stock,

$$K_{t+1}^* = vY_{t+1}$$

and again on the assumption of an optimal adjustment in the unit period

$$K_{t+1}^* - K_t^* = K_{t+1} - K_t = I_t = v(Y_{t+1} - Y_t) = v\Delta Y_t.$$

In other words, for net investment to be positive, output must be growing: v is called the accelerator. The acceleration principle can be derived from a cost-minimizing model on the assumption of either fixed (technical) coefficients and exogenous output, or variable coefficients with constant relative prices of inputs and exogenous output.

Some of the shortcomings of this simple model were well known; for example, the problem of being optimally adjusted: this was discussed in the context of whether or not the economy (or the firm) was working at full capacity. If the economy was operating with surplus capacity, an increase in aggregate demand would not lead to an increase in investment. Similarly, it was well known that the accelerator may work in an asymmetric fashion because of the limitations imposed on decreasing aggregate capital stock by the rate of depreciation: the economy as a whole could only decrease its capital stock by not replacing capital goods that were depreciating. Another important qualification to the simple accelerator model was than an increase in (expected) output would lead to an increase in investment only if it was believed that, in some way, the increase was ‘permanent’ or at least of long duration.

A generalization of the simple accelerator is provided by the flexible accelerator or the capital stock adjustment principle (also known as the distributed lag accelerator). It overcomes one of the major shortcomings of the simple accelerator, namely, the assumption that the capital stock is always optimally adjusted. The flexible accelerator also assumes that there is an optimal relationship between capital stock and output but allows for lags in the adjustment of the actual capital stock towards the optimal level. This is written as

$$I_t = b(K_t^* - K_{t-1})$$

where b is a positive constant between zero and one and K_t^* equals vY_t. This equation implies that the adjustment path of actual capital stock towards the optimal level is asymptotic. In this version, the adjustment is not instantaneous either since, because of uncertainty, firms do not *plan* to make up the difference between K_t^* and K_{t-1} and/or because the *supply* of capital goods does not allow the adjustment to be instantaneous. A similar equation was derived by assuming increasing marginal costs of adjusting capital stock by Eisner and Strotz (1963).

In evaluating the acceleration principle it is worth stressing that, in some versions, it is used as an explanation of investment demand with the implicit assumption that the supply of capital goods will always satisfy that demand. In models where the acceleration principle is used to explain the *supply* of capital goods, it is assumed that they always satisfy the demand for them. The flexible accelerator is a hybrid version which includes both demand and supply elements. Although there is no formal treatment of replacement investment, it is usually postulated to be determined in the same way as net investment.

A major shortcoming of the acceleration principle is its simplistic treatment of *expectations* of future demand as well as its neglect of expectations of the time paths of output and input prices. Although most of the work in this field treats the acceleration principle as applying to the aggregate economy, it has also been used to explain investment by firms. It is especially important that the supply of capital goods is formally modelled along with the acceleration principle determining investment demand. Aggregation over firms is usually assumed to be a simple exercise of ‘blowing up’ an individual firm's investment demand. However, it should not be forgotten that in a modern capitalist economy an individual firm may invest by simply taking over an existing firm rather than by buying new capital goods. An important shortcoming of the acceleration principle is its neglect of technological change.

The acceleration principle is an important concept and has been used successfully in explaining investment behaviour as well as cyclical behaviour in a capitalist economy. It will continue to play an important role in macroeconometric models as well as in models of business cycles.

See Also

- Clark, John Maurice
- multiplier–accelerator interaction

Bibliography

Aftalion, A. 1913. *Les crises périodiques de surproduction*. Paris: Rivière.

Chenery, H.B. 1952. Overcapacity and the acceleration principle. *Econometrica* 20(1), 1–28.

Clark, J.M. 1917. Business acceleration and the law of demand: a technical factor in economic cycles. *Journal of Political Economy* 25, 217–35.

Eisner, R. 1960. A distributed lag investment function. *Econometrica* 28(1), 1–29.

Eisner, R. and Strotz, R. 1963. Determinants of business investment. In Commission on Money and Credit, *Impacts of Monetary Policy*. Englewood Cliffs, NJ: Prentice-Hall.

Haberler, G. 1937. *Prosperity and Depression*. Geneva: League of Nations.

Harrod, R.F. 1936. *The Trade Cycle*. Oxford: Oxford University Press.

Junankar, P.N. 1972. *Investment: Theories and Evidence*. London: Macmillan.

Knox, A.D. 1952. The acceleration principle and the theory of investment: a survey. *Economica* 19(75), 269–97.

Koyck, L. 1954. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.

Marx, K.H. 1863. *Theories of Surplus Value*, Part II. Moscow: Progress Publishers.

Pigou, A.C. 1927. *Industrial Fluctuations*. 2nd edn. London: Macmillan, 1929.

How to cite this article

Junankar, P. N. "acceleration principle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000014> doi:10.1057/9780230226203.0004

close print preview

The New Palgrave Dictionary of Economics Online

access to land and development

Alain de Janvry and Elisabeth Sadoulet
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Access to land can be an effective policy instrument for poverty reduction. This article shows how different types of property rights can affect access and use, analyses different modes of access, especially the role of land markets, and sets out some of the policy implications. It argues that making land an effective tool for development requires more than policing access: access must be secure, combined with the use of complementary inputs, and achieved in a context of institutions, public goods, and policies that allow the sustainable competitiveness of beneficiaries.

Keywords

access to land and development; agriculture and economic development; common property resources; democracy and economic development; family farm system; Geographical information systems; human capital; Inheritance rights; land markets; land reform; land registration; land use; natural capital; open-access resources; poverty alleviation programmes; property rights; rent control; rental markets; total factor productivity; tragedy of the commons

Article

Access to land, and the conditions under which it happens, play a fundamental role in economic development. This is because the way the modes of access to land and the rules and conditions of access are set, as policy instruments, has the potential of increasing agricultural output and aggregate income growth, helping reduce poverty and inequality, improving environmental sustainability, and providing the basis for effective governance and securing peace. This potential role is, however, difficult to capture, and there are many cases of failure. History is indeed replete with serious conflicts over access to land and with instances of wasteful use of the land, both privately and socially. Governments and development agencies have for this reason had to deal with the ‘land question’ as an important item on their agendas (de Janvry et al., 2002). We explain in this article: (*a*) why access to land, and the conditions under which it is accessed and used, are important for economic development, (*b*) how

different types of property rights can affect access and use, (*c*) the different modes of access, and in particular the role of land markets, and (*d*) some of the policy implications, in order to show how access to and use of the land can contribute to economic development. We stress in this article that access to land may be a difficult policy question, but that access will translate into development only if the harder question of influencing the way it is used is effectively resolved.

Importance of access to land for development

Land is not only a factor of production, and as such a source of agricultural output and income; it is also an asset, and hence a source of wealth, prestige, and power. Because it is a natural asset, its use affects environmental sustainability or degradation. For these reasons, the link between access to land and development is quite multidimensional and complex, with many trade-offs involved.

If land is to serve as an instrument for output and income growth, investments have to be made to improve its productivity. For this to happen, incentives have to be provided. Some of these investments are short-term, but many others are tied to the land for long periods of time. As a result, security of access is a central policy issue as it is necessary for these investments to be made. Security can be guaranteed through formal means such as titles and legal enforcement, but also through informal mechanisms such as community recognition and enforcement of rights. Whichever way it is achieved, security of access must be credible if it is to induce investment (Deininger, 2003).

To result in output and income growth, access to land must not only be secure, it must also be accompanied by access to complementary inputs and occur in a context favorable to productive use of the land. Empirically well-established complementary inputs include other types of natural capital such as water, working capital, and human capital. Access to land without these complementary inputs in the agricultural production function is not useful for development. In addition, the context where land is used affects its productivity. This includes institutions (such as credit, insurance, and product and factor markets with low transactions costs), public goods (such as infrastructure, market intelligence, research and extension, land registration, and contract enforcement mechanisms), and policies (macroeconomic and agricultural policies favorable to the activities in which the land is used). If complementary inputs and a favorable context for land use are not provided, it is quite evident that access to land will achieve little for output and income. Access to land is thus necessary but not sufficient. Providing what it takes beyond access to achieve income and growth – complementary inputs and a favorable context – can be highly demanding.

Secure access to land and to complementary inputs in a context that allows productive use can be a powerful instrument for poverty reduction. The family farm, with its labour cost advantage when there are transactions costs in labour markets and incomplete incentives to hired labour, can be particularly effective for this (Bardhan, 1984). The inverse relation between farm size and total factor productivity, derived from the labour cost advantage of the family farm, has been cited as the empirical regularity justifying redistributive land reforms towards a family farm system. Access to even a small plot of land can be a source of security in the face of food market and labour market risks. Women's control over land can be a source of empowerment, helping them consolidate their decision-making status over household expenditures that will often favour children (Agarwal, 1994).

Finally, as a good in limited supply, the distribution of access to land can have a powerful influence on

social inclusion and local governance. More egalitarian access can be the basis for greater political participation, more respect for the rule of law, and the ability to raise local fiscal revenues from a land tax, and provide the basis for the consolidation of democracy (Binswanger, Deininger and Feder, 1995). While these relations are far from direct, it is impossible to ignore the role that access to land plays in affecting these outcomes.

Property rights over land

The benefits that can be derived from access to land depend on the property rights that codify access and use. Property rights become increasingly complete as they allow the following functions to accumulate: entry, extraction, management, exclusion, and sale (Ostrom, 2002). Open-access resources grant to all the rights of entry and extraction. They typically induce over-extraction, leading to the 'tragedy of the commons'. Common property resources grant to members of a defined group, such as a community, the rights of entry, extraction, management, and exclusion of non-community members. This form of property right can result in socially optimal resource use if community members have the ability to cooperate in defining and enforcing rules for individual extraction and maintenance (Baland and Platteau, 1996). Public ownership with centralized management also gives leaders these same rights. Socially optimum resource use can be achieved if controls and incentives can be aligned between leaders and workers, which has historically proved to be difficult in agriculture, despite many attempts. Finally, individual or corporate property rights give owners the full bundle of rights, including those of rental and sale. The effectiveness of this form of property right in land use depends on the existence of efficient land rental and sales markets, as well as the ability to internalize externalities, achieve economies of scale, and access mechanisms for risk spreading. Common property resources with cooperation may be a superior form of property right when individual tenures are unable to fulfil these functions.

Whether property rights correspond to common property or to individual or corporate forms of tenure, these rights have desirable aspects that need to be realized for access to be efficient. One is duration of the rights: long-term investments require sustained access and clear specification of how rights are transferred to others. Inheritance rights are thus a fundamental aspect not only of access to land but also of land use. A second is precise demarcation of land boundaries and clear specification of rights. Geographical information systems based land demarcation, land registries and record keeping of transactions, and adjudication of rights mechanisms are thus fundamental aspects of land management. A third is availability of conflict-resolution mechanisms, where conflicts over access to land can be resolved through informal or formal procedures that are fair and expedient. Uncertain rights and unresolved conflicts over access rights are the norm rather than the exception in developing countries, requiring major investments in regularizing these situations. Finally, property rights must be evolutive, and it must be possible to individualize or consolidate rights as opportunities and needs arise.

Modes of access to land

With open-access resources, entry is granted to all. Access to common property resources is usually given by birthright in a particular community. Clear demarcation of boundaries and clear determination

of membership are important to permit the definition and enforcement of rules. Individual encroachment on public lands and establishing adverse possession rights through occupation is an important form of access where public lands remain plentiful. Finally, individual inheritance is also one of the most prevalent forms of access to land, with eventually discriminatory rights due to primogeniture and to gender and kinship privileges in inheritance.

Access to land through rental markets is often constrained by insecurity of property rights, confining transactions to narrow circles of confidence (family, friends, social peers), thus segmenting markets. While fixed-rent contracts are first-best efficient, sharecropping contracts may be the most efficient way of accessing land when there are market failures in insurance, credit, and non-traded inputs such as management and supervision (Hayami and Otsuka, 1993). In general, the role of land rental markets as a mode of access to land for the poor has been under-appreciated in land policy, and these markets have all too often been atrophied by misguided rent controls.

Finally, the land sales market should expectedly be the most effective way of providing access to land to the most efficient entrepreneurs. This may not be the case, however, because these markets suffer from serious distortions that limit the fulfilment of this role. Land tends to be overpriced relative to its value in productive use due to its function as a store of wealth, speculation on land appreciation, tax advantages, use as collateral in accessing credit, and the status and power it conveys. Overpricing implies that even full credit lines using the land as collateral will not be sufficient to allow poor people to access land without subsidies.

Access to land and development: policy implications

In managing their ‘land question’, most countries have experimented with some type of land reform programme (Dorner, 1992). This includes land reforms that have used the threat of expropriation to induce extensively used large farms to modernize or subdivide into smaller farms (Brazil). Other reforms have collectivized the land, either as state farms or as cooperatives. This has generally, as in Russia and eastern Europe, been based on the belief in economies of scale in farming and the superior efficiency of centralized management. In other cases, as in Latin America, collective farms have been used to facilitate transitions between large haciendas and subsequent distribution of the land as individual tenures (Mexico, Peru, Chile). Finally, the inverse relation between total factor productivity and farm size has been invoked in implementing redistributive land reforms that have established family farms out of former large farms (Taiwan, South Korea) or out of state farms or cooperatives (Albania, Bulgaria).

Because the land sales market should be the most effective way of codifying access to land, land reforms have recently taken the form of ‘market-assisted land reforms’, with examples in Brazil, Colombia, and South Africa (Deininger, 2003). In this case, transactions occur between willing sellers and willing buyers, and subsidies are granted to the poor in addition to credit so they can afford purchases at market prices that are in excess of the productive value of the land. These interesting experiments are still in progress and in much need of evaluation.

Conclusion

Access to and use of the land is a fundamental instrument for successful development, both economically and socially. History shows both success stories and resounding failures. In general, making land an effective tool for development requires more than policing access: access must be secure, combined with the use of complementary inputs, and achieved in a context of institutions, public goods, and policies that allow the sustainable competitiveness of beneficiaries. Many policies and programmes have been put in place to achieve this goal, but the complexity of the task explains why success requires extensive control and commitment (Warriner, 1969). A fundamental lesson derived from the history of the ‘land question’ is thus that, while reforming the pattern of access to land is difficult, it is far more difficult to make access complete in the sense of securing the competitiveness of beneficiaries so that they achieve income growth, poverty reduction, and sustainable use.

See Also

- common property resources
- land markets
- peasant economy
- poverty alleviation programmes
- property rights
- tragedy of the commons

Bibliography

Agarwal, B. 1994. *A Field of One's Own: Gender and Land Rights in South Asia*. Cambridge, Mass.: Cambridge University Press.

Baland, J.-M. and Platteau, J.-P. 1996. *Halting Degradation of Natural Resources: Is There a Role for Rural Communities?* Oxford: Oxford University Press.

Bardhan, P. 1984. *Land, Labor, and Rural Poverty: Essays in Development Economics*. New York: Columbia University Press.

Binswanger, H., Deininger, K. and Feder, G. 1995. Power, distortions, revolt, and reform in agricultural land relations. In *Handbook of Development Economics*, vol. 3B, ed. J. Behrman and T. Srinivasan. Amsterdam: North-Holland.

de Janvry, A., Gordillo, G., Platteau, J.-P. and Sadoulet, E. 2002. *Access to Land, Rural Poverty, and Public Action*. Oxford, UK: Oxford University Press.

Deininger, K. 2003. *Land Policies for Growth and Poverty Reduction*. Washington, DC: World Bank.

Dorner, P. 1992. *Latin American Land Reform in Theory and Practice: A Retrospective Analysis*.

Madison, WI: University of Wisconsin Press.

Hayami, Y. and Otsuka, K. 1993. *The Economics of Contract Choice: An Agrarian Perspective*. Oxford, UK: Oxford University Press.

Ostrom, E. 2002. The puzzle of counterproductive property rights reforms: a conceptual analysis. In *Access to Land, Rural Poverty, and Public Action*, ed. A. de Janvry et al. Oxford, UK: Oxford University Press.

Warriner, D. 1969. *Land Reform in Principle and Practice*. Oxford: Clarendon Press.

How to cite this article

de Janvry, Alain and Elisabeth Sadoulet. "access to land and development." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000212> doi:10.1057/9780230226203.0005

close print preview

The New Palgrave Dictionary of Economics Online

accounting and economics

Joel S. Demski
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Accounting provides an important source of economic measures, yet consistently falls short of the economist's conceptual ideal. This shortfall is fodder for economic research, is the result of economic forces, and is the key to making the best possible use of these measures.

Keywords

accounting and economics; auditing regulation; depreciation; Financial Accounting Standards Board (FASB); Generally Accepted Accounting Principles (GAAP); historical-cost accounting; information school of accounting; International Accounting Standards Board (IASB); measurement school of accounting

Article

Broadly viewed, economics is concerned with the production and allocation of resources, and accounting is concerned with measuring and reporting on the production and allocation of resources. Corporate financial reporting, income tax reporting, and product cost analysis at the firm level are familiar accounting activities. Of course, accounting itself is a production process, and the production and allocation of its output is even regulated; for example, how a firm measures and reports its financial progress and how a firm communicates with outsiders are regulated, and auditing of a firm's public financial statements is mandatory. This suggests two interrelated themes: accounting is useful in a wide variety of activities, including economics research, and accounting itself is a fascinating and important area of economics research.

Using or researching the accountant's products, however, rests on an understanding of what those products are and how they are produced. Accounting, in fact, uses the language of economics (for example, value, income and debt) and the algebra of economic valuation (as income is change in value adjusted for dividends and stock issues). But it falls far short of how an economist would approach these matters. For example, the accounting value of a firm is usually well below its market value, as measured

by the market price of its outstanding equity securities.
This disparity is related to the institutional setting in which accounting products are produced, and to the economic forces operating on and within those institutions.

Institutional highlights

Accounting cannot be divorced from its institutional setting. Were firms truly single-product entities, and were markets complete and perfect, economic measurement would be well defined, the nirvana of classical income measurement (for example, Hicks, 1946) would be operational. Unfortunately, in such a setting no one would pay for the services of an accountant simply because the underlying fundamentals would be assumed to be common knowledge. But firms are multi-product entities, markets are neither perfect nor complete, and the underlying fundamentals are far from common knowledge. Here we find a demand for accounting services, such as measuring a firm's periodic income, the performance of the divisions within that firm, and the cost of each of its products. We also find considerable ambiguity over how best to perform those services.
Firms’ published financial reports are the most visible accounting product. They entail a reporting entity (the organization about which the financial reports purport to speak), a listing of resources and obligations in its balance sheet, and a listing of the flow of resources during the reporting period in its income statement. Ambiguity is omnipresent. The reporting entity is not an economically defined firm, as its economic relationships are likely to be more extensive than those identified by its formal reporting; for example, implicit economic arrangements are generally ignored in these reports. Nor is the reporting entity simply a legally defined firm, as it often includes, say, a number of wholly or partially owned though legally free-standing legal entities aggregated into its public reports. Even with an unambiguous reporting entity, that entity's control of economic resources would be incompletely and inaccurately measured. Some assets, such as proprietary knowledge or capital assets acquired through lease arrangements, would not be included. And among those included we would find a mixture of current prices (for example, cash and some financial instruments) and historical cost (for example, most real assets).
The flow measure is equally ambiguous. It is broadly based on what customers have paid minus the resources that were consumed in the process of satisfying those customers. Such wide-ranging phenomena as product warranties and potential product liabilities, uncollectible accounts, pension plans, advertising, research and development and employee training render precise identification of what customers have paid or what resources were consumed largely the product of art as opposed to science.
Regulation, to no one's surprise, now enters the picture. Public financial reports are typically required to be produced according to Generally Accepted Accounting Principles (GAAP). These reports are also typically required to be audited, where the auditor attests to the claim the reports are in compliance with GAAP. One reason for regulations is that the noted ambiguity places a premium on coordinated measurement approaches, a classic example of a network externality (Wilson, 1983). A second reason, based on investor protection concerns and again related to the ambiguity, is the potential for opportunism. Absent auditing, the public financial report is simply management's self-report of its financial results and the unverified claim that those results were measured according to GAAP. Of course the auditor's verification is statistical and judgemental; to no one's surprise, the auditor himself is also regulated.

GAAP itself is fluid, varied, contentious and political at the margin. Two major, competing boards, the Financial Accounting Standards Board (FASB) in the United States and the International Accounting Standards Board (IASB) outside the United States, are largely but not entirely responsible for the definition of GAAP. Historically, the two boards have differed (though inter-board coordination has become a priority in recent years), and have tended to lag behind innovations in transaction design. Moreover, firms design transactions with an eye towards how they will be rendered under GAAP. Leases, as noted above, are largely absent from firms' balance sheets. This reflects careful transaction design so the acquisition and financing of capital assets can be excluded, according to GAAP, from the firm's balance sheet – in effect lowering the officially measured debt. Similarly, compensating employees with equity options was, until most recently, a form of compensation that, according to GAAP, is absent from firms' income statements. (While GAAP is defined outside explicit governmental agencies, compliance with GAAP is legally required. The Securities and Exchange Commission in the United States has statutory authority to define GAAP, and has delegated this task, by and large, to the FASB. The European Union, in turn, has delegated this task to the IASB. Auditing regulations, in turn, are more varied, as is enforcement.)

The least visible accounting activity is what transpires inside the firm. Here we again find measures of stocks and flows of resources, aimed now at divisions, plants, departments, product lines, and so forth. The noted ambiguities remain, and extend to such arenas as tracing services from a common provider, such as human resources or cash management, to the consuming units inside a firm or dividing the accounting profit on some particular product line among the various units within the firm whose combined activities produced it. Here we also find less, but far from nil, reliance on GAAP. These measurement activities are not, literally speaking, regulated; but they do rely on the same underlying financial history. We also find a variety of non-financial measures, such as customer and employee satisfaction or student course evaluations. We also find occasional wholesale redesign of a firm's internal accounting activity (Anderson, Hesford and Young, 2002). (Tax accounting is yet another activity, though the measurement rules are often more directly statutory in nature, and diverge from GAAP.)

Importantly, now, the question is: how are we to make sense of these patterns? Two approaches have emerged through the years, the measurement school and the information school.

The measurement school

The measurement school takes its cue from classical economics. In a fully developed general equilibrium model, with complete and perfect markets (for example, Debreu, 1959), value and income are well defined, as is the value of a firm's assets and obligations. The measurement school takes this as a desideratum and emphasizes the importance of approaching this economic ideal reasonably well. This is the source of accounting's intellectual history, its underlying definitions of asset, liability, income, revenue and expense, and the rhetoric used by its regulators. (Important contributors to this school of thought include Paton, 1922; Clark, 1923; Canning, 1929; Edwards and Bell, 1961; Solomons, 1965; and Chambers 1966.)

The advantage of the measurement approach is its (relative) clarity. Foreign currency translation at contemporaneous exchange rates, economic depreciation, and market value of complex financial instruments, for example, all take on a natural conceptual clarity at this point. Indeed, at least in the

United States, we find the national income accounts are not mere consolidations of GAAP measures, but are produced with an eye on the economic fundamentals. (See Petrick, 2002. More broadly, this leads us to the theory of measurement in general – for example, existence, uniqueness and meaningfulness of a measure – and the axiomatic characterization of additive structures; Krantz et al., 1971; and Mock, 1976. Unfortunately, adding up the value of a firm's assets views the firm as the sum of its assets, so to speak, and is inconsistent with synergies among the asset groups. In parallel fashion, marginal cost is the only meaningful product-cost statistic in a multi-product firm, absent separability. Yet accounting requires accounting product costs to sum to the total cost, which implies that the accounting product costs can be reasonably viewed as marginal-cost estimates only under conditions of separability and constant returns. This suggests theoretical limits to the measurement approach.)

Likewise, with the advent of financial engineering it is natural, from the measurement school perspective, that GAAP require fair value (that is, as if market value) estimates of these instruments. In short, with the measurement school we at least know what it is, conceptually, we are trying to measure. The disadvantage of the measurement approach is that it relies on economics to identify the conceptual ideal, but ignores economics when the time comes to worry about resources devoted to the measurement enterprise. (Audit fees alone exceed $6 billion annually in the United States.) It also raises such questions as why international differences persist, why accounting does such a poor job of tracking economic value and why, given this presumptively poor performance, it continues to survive. (Flawed as it is, from this perspective, we also know foreknowledge of firms' annual reports would allow highly profitable speculation; Ball and Brown, 1968.) It also fails to capture the accountant's stock in trade of eschewing economic measurement and embracing historical-cost allocation. Capital assets are not measured at economic value, and no attempt is made to measure economic depreciation. Rather, the historical cost of the capital asset is allocated, is divided among multiple uses in some formula-driven manner. For example, the initial cost of a real asset is divided among periods (accounting depreciation) and from there among products, resulting in an allocated portion hitting the income statement and the net balance being the asset value on the balance sheet. Moreover, when accounting reports the cost of a firm's product, it is reporting not marginal cost but an allocated accounting cost. Morgenstern (1965, p. 79) is particularly eloquent:

> But it is clear that in the absence of a convincing and complete theory there is no unique and objective way of accounting for costs when overhead, amortization and joint costs have to be taken into consideration … 'Cost' is merely one aspect of a valuation process of great complexity.

The measurement school, then, focuses on economic measurement as the ideal, but ignores economic forces that impinge on the measurement process.

The information school

The information school, in contrast, focuses on these economic forces and takes its cue from the economics of uncertainty. It views the accounting product not literally as measures of resources but as information that purports to inform about these resources. Abstractly, then, accounting is a mapping

from underlying acts and events into the real numbers. In this view, accounting is one among many sources of information. Analysts, the financial press and trade associations are familiar sources of financial information, as are government statistics themselves. Moreover, firms often engage in voluntary disclosures; for example, new product announcements, major investment announcements, and even so-called earnings warnings where they reveal that a forthcoming earnings measure will be lower than originally anticipated. In addition, the typical financial report reports cash flow, an utterly reliable, unambiguous measure. (Important contributors to this school of thought include Butterworth, 1972; Feltham, 1972; Ijiri, 1975; Beaver, 1998; and Christensen and Demski, 2002.)

The advantage of this view is it forces us to think in terms of complements and substitutes when dealing with this vast array of sources, and to look for economic forces that drive the disparity that bedevils the measurement school. And it is here that the comparative advantage of the accounting channel comes into focus: it is purposely designed and managed so that it is difficult to manipulate (Ijiri, 1975). This is why it often resorts to historical-cost measurement, as this removes major elements of subjectivity and manipulation potential. It is also why, in organized financial markets, most valuation information arrives before the firm's financial reports; and in this sense the financial reports provide a veracity check on the earlier reporting sources. In addition, cost allocation now enters as a natural phenomenon, either as a simple scaling device or – to use an analogy with informationally efficient markets – as a cousin to an information-based pricing kernel in a financial market (Christensen and Demski, 2002; Ross, 2004).

Libraries are organized in coordinated fashion, as are phone books; and the same can be said about accounting. A curiosity is the political side of the regulatory apparatus. It is difficult, for example, for the incumbent government to alter a government-provided statistical series, yet it is routine for the incumbent government to intervene in the accounting regulatory process. A second curiosity is the seemingly episodic nature of financial reporting frauds (Demski, 2003), although at the micro level it is well understood that opportunistic reporting is part of the game. For example, an ability to shift income from a later to an earlier period may be an inexpensive signal or, to speak more cynically, less costly to the firm than shifting real resources.

The disadvantage of the information school is its sheer breadth. The institutional context includes a vast array of information sources and actors, and sorting out first-order effects remains problematic.

Conclusion

Accounting, then, is simultaneously an important source of economic data and a collection of institutional regularities that provide research economists with yet another venue for documentation and exploration of economic forces. Why do we see episodic regulatory interventions? Why do we see forecasts of forthcoming accounting measures? Why do we not see supplementary estimation of economic depreciation? Why do we see the mix of historical-cost and market values that characterize modern financial reporting? Questions of this sort motivate much of the current research in accounting and finance.

See Also

- assets and liabilities

- capital measurement
- cost functions
- depreciation
- double-entry bookkeeping
- human capital
- measurement
- pensions
- present value

Helpful comments by Haijin Lin and David Sappington are gratefully acknowledged.

Bibliography

Anderson, S., Hesford, J. and Young, M. 2002. Factors influencing the performance of activity based costing teams: a field study of ABC model development time in the automobile industry. *Accounting Organizations and Society* 27, 195–211.

Ball, R. and Brown, P. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6, 159–78.

Beaver, W. 1998. *Financial Reporting: An Accounting Revolution*. Englewood Cliffs, NJ: Prentice-Hall.

Butterworth, J. 1972. The accounting system as an information function. *Journal of Accounting Research* 10, 1–27.

Canning, J. 1929. *The Economics of Accountancy*. New York: Ronald Press.

Chambers, R. 1966. *Accounting, Evaluation and Economic Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Christensen, J. and Demski, J. 2002. *Accounting Theory: An Information Content Perspective*. New York: McGraw-Hill/Irwin.

Clark, J. 1923. *Studies in the Economics of Overhead Costs*. Chicago: University of Chicago Press.

Demski, J. 2003. Corporate conflicts of interest. *Journal of Economic Perspectives* 17(2), 51–72.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New Haven, CT: Yale University Press.

Edwards, E. and Bell, P. 1961. *The Theory and Measurement of Business Income*. Berkeley: University of California Press.

Feltham, G. 1972. *Information Evaluation*. Sarasota, FL: American Accounting Association.

Hicks, J. 1946. *Value and Capital*. Oxford: Clarendon Press.

Ijiri, Y. 1975. *Theory of Accounting Measurement*. Sarasota, FL: American Accounting Association.

Krantz, D., Luce, R., Suppes, P. and Tversky, A. 1971. *Foundations of Measurement*. New York: Academic Press.

Mock, T. 1976. *Measurement and Accounting Information Criteria*. Sarasota, FL: American Accounting Association.

Morgenstern, O. 1965. *On the Accuracy of Economic Observations*. Princeton, NJ: Princeton University Press.

Paton, W. 1922. *Accounting Theory*. New York: Ronald Press.

Petrick, K. 2002. Corporate profits: profits before tax, profits tax liability, and dividends. Methodology Paper. Washington, DC: Bureau of Economic Analysis. Online. Available at http://www.bea.gov///bea/ARTICLES/NATIONAL/NIPA/Methpap/methpap2.pdf, accessed 18 November 2005.

Ross, S. 2004. *Neoclassical Finance*. Princeton, NJ: Princeton University Press.

Solomons, D. 1965. *Divisional Performance: Measurement and Control*. New York: Financial Executives Research Foundation, 1965.

Wilson, R. 1983. Auditing: perspectives from multiperson decision theory. *Accounting Review* 58, 305–18.

How to cite this article

Demski, Joel S. "accounting and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000019> doi:10.1057/9780230226203.0006

close print preview

The New Palgrave Dictionary of Economics Online

adaptive estimation

Douglas G. Steigerwald

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Adaptive estimation arises in the context of partially specified models. Partially specified models occur with some frequency in econometrics. For example, a linear regression model in which the error distribution is unknown is a partially specified model. So too are many of the diffusion models employed in empirical finance. One active research area is to understand the conditions under which the lack of full specification does not affect the asymptotic efficiency of the estimator, in which case the estimator is termed ‘adaptive’.

Keywords

adaptive estimation; kernel estimator; linearized likelihood estimation; maximum likelihood; nonparametric estimation; semiparametric estimation; spline functions

Article

An adaptive estimator is an efficient estimator for a model that is only partially specified.
For example, consider estimating a parameter that describes a sample of observations drawn from a distribution *F*. One natural question is: is it possible that an estimator of the parameter constructed without knowledge of *F* could be as efficient (asymptotically) as any well-behaved estimator that relies on knowledge of *F*? For some problems the answer is ‘yes’, and the estimator that is efficient is termed an adaptive estimator.
Consider the familiar scalar linear regression model (in which we let *t* rather than *i* index observations)

$$Y_t = \beta_0 + \beta_1 X_t + U_t,$$

where the regressor is exogenous and $\{U_t\}$ is a sequence of n independent and identically distributed random variables with distribution F. The parameter vector $\beta = (\beta_0, \beta_1)'$ is often of interest rather than the distribution of the error, F. If we assume that F is described by a parameter vector λ (that is, we parameterize the distribution), then the resultant (maximum likelihood or ML) estimator of β is parametric. If we assume only that F belongs to a family of distributions, then the resultant estimator of β is *semi*parametric. Because the OLS estimator does not require that we parameterize F, the OLS estimator is semiparametric. If the population error distribution is Gaussian, we know that the OLS estimator is equivalent to the ML estimator, and so is efficient. Although the OLS estimator is generally inefficient if F is not Gaussian, it may be possible to construct an alternative (semiparametric) estimator that retains asymptotic efficiency if F is not Gaussian. If we find that, for a family of distributions that includes the Gaussian, this estimator is asymptotically equivalent to the ML estimator, then this estimator is adaptive for that family.

The question then is: how can we verify that an estimator is adaptive? As there will generally be an arbitrarily large number of distributions in the family, it is not feasible to algebraically verify asymptotic equivalence for each distribution. In a creative paper, Stein (1956) first proposed a solution to this problem. Let $\{F_\lambda, \lambda \in \Lambda\}$ define a subset of the family of distributions, each member of which is parameterized by a value of λ (each member of this family must satisfy certain technical conditions, such as absolute continuity, which will not be explicitly defined). Although primary interest centers on β, the full set of parameters includes λ. The information matrix, evaluated at the population parameter values, is

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\lambda} \\ \mathcal{I}_{\lambda\beta} & \mathcal{I}_{\lambda\lambda} \end{pmatrix},$$

where $\mathcal{I}_{\beta\beta}$ corresponds to the elements of β. Estimators of β (again, the estimators must satisfy technical conditions, such as $\sqrt{n}$ consistency, which are also not explicitly defined) will have covariance matrix that is at least as large as $\mathcal{I}^{\beta\beta}$, which is the upper left component of $\mathcal{I}^{-1}$. If the partial derivative of the log-likelihood with respect to β (the score for β) is orthogonal to the score for λ, then $\mathcal{I}_{\beta\lambda} = 0$ and $\mathcal{I}^{\beta\beta} = \mathcal{I}_{\beta\beta}^{-1}$. Because $\mathcal{I}_{\beta\beta}$ corresponds only to the parameter β, the asymptotically efficient estimator of β can be constructed without knowledge of λ. Stein argued that, if the condition $I_{\beta\lambda} = 0$ holds for all the elements of $\{F_\lambda\}$, then β is adaptively estimable.

While Stein's condition has intuitive appeal, it is not straightforward how to use the condition to define estimators that are adaptive. In an invited lecture, Bickel (1982) laid out a simpler condition that does yield a straightforward link to the construction of adaptive estimators. To understand the condition, let E_F denote expectation with respect to the population error distribution and let $E_{\tilde{F}}$ denote expectation with respect to an arbitrary distribution $\tilde{F} \in \mathcal{F}$. Let l be the log-likelihood for the regression model with data $z = (y, x)$ and let $l(z, \beta, F)$ denote the score for β, constructed from the model in which F is the

error distribution. A familiar condition that arises in the context of likelihood estimation is that the expected population score $E_F\left[\dot{l}(z, \beta, F)\right]$ equal 0. Bickel's condition is simply that the population score must have expectation zero over the entire family $\mathfrak{F}$, that is, for any $\tilde{F} \in \mathfrak{F}$,

$$E_{\tilde{F}}\left[\dot{l}(z, \beta, F)\right] = 0.$$

The two conditions are linked: if $\mathfrak{F}$ is a convex family, then Stein's condition is implied by Bickel's condition. In detail, if $\mathfrak{F}$ is a convex family, then $F_\lambda = \lambda F + (1 - \lambda)\tilde{F}$ with λ an element of $\Lambda = (0, 1)$. Bickel's condition then arises from Stein's condition by taking the limit as $\lambda \to 0$. For the linear regression model, an adaptive estimator of β exists for the family $\mathfrak{F}$ that consists of all distributions that are symmetric about the origin (and several other technical conditions). If interest centres on the slope coefficient alone, then one need not restrict attention to distributions that are symmetric about the origin, as an adaptive estimator of β_1 can exist even if β_0 is not identified.

Bickel's score condition leads naturally to estimators that contain nonparametric estimators of the distribution, $\hat{F}$. In consequence, adaptive estimation requires a second condition: the nonparametric estimator of the score must converge in quadratic mean to the population score. The resulting estimators of β are two-step estimators. The estimators require, as the first step, a $\sqrt{n}$-consistent estimator such as the OLS estimator. To understand the estimator's form, note that, if the distribution were known, then the two-step (linearized likelihood) estimator is

$$\hat{\beta}_{OLS} + n^{-1}\sum_{t=1}^{n} s\left(Z_t, \hat{\beta}_{OLS}, F\right),$$

with $s\left(Z_t, \hat{\beta}_{OLS}, F\right) = \mathcal{I}^{11}\left(\hat{\beta}_{OLS}, F\right)\dot{l}\left(Z_t, \hat{\beta}_{OLS}, F\right)$. The linearized likelihood estimator is asymptotically efficient. To form an adaptive estimator of β, we must replace F with a nonparametric estimator $\hat{F}$. If $\hat{F}$ is constructed so that $s\left(Z_t, \hat{\beta}_{OLS}, \hat{F}\right)$ converges in quadratic mean to $s\left(Z_t, \hat{\beta}_{OLS}, F\right)$, then

$$\hat{\beta}_{AD} = \hat{\beta}_{OLS} + n^{-1}\sum_{t=1}^{n} s\left(Z_t, \hat{\beta}_{OLS}, \hat{F}\right)$$

is an adaptive estimator of β for the family $\mathfrak{F}$.

For the linear regression model, as for numerous other models, nonparametric estimation of *F* entails nonparametric estimation of the density *f*. One popular nonparametric density estimator is the kernel estimator, which is employed by Portnoy and Koenker (1989) in their proof that semiparametric quantile estimators are also adaptive for β. If $\{\hat{U}_t\}$ denotes the OLS residuals, then a kernel density estimator is defined for all *u* in a small neighbourhood of each value of $\hat{U}_t$ as

$$\hat{f}_t(u) = (n-1)^{-1} \sum_{\substack{s=1 \\ s \neq t}}^{n} \xi_\sigma\left(u - \hat{U}_s\right),$$

where ξ_σ is a weight function that depends on the smoothing parameter σ. In Steigerwald (1992), ξ_σ corresponds to a Gaussian density with mean 0 and variance σ^2. The variance controls the amount of smoothing; as σ^2 declines, the weight given to residuals that lie some distance from $\hat{U}_t$ tends to zero. Of course, there are many other ways to form the nonparametric score estimator. Newey (1988) approximates the score by a series of moment conditions, which arise from exogeneity of the regressor and symmetry of *F*. Faraway (1992) uses a series of spline functions to approximate the score. Chicken and Cai (2005) use wavelets to form the basis for nonparametric estimation of *f*.

Recent results in adaptive estimation have focused on problems in which the error distribution is known, but other features are modelled nonparametrically. Some of the most intriguing results concern the type of stochastic differential equation often encountered in financial models. The price of an asset that is measured continuously over time, P_t, is often modelled as

$$dP_t = m_t dt + v_t dB_t.$$

The presence of standard Brownian motion, B_t, makes the model of price a stochastic differential equation. The function m_t captures the deterministic movement or drift while v_t is the potentially time-varying scale of the random component. Lepski and Spokoiny (1997) study the model in which v_t is constant and m_t is unknown. They establish that a nonparametric estimator of *m* is pointwise adaptive. Yet an estimator that is pointwise adaptive – that is, for a given point t_0 the nonparametric estimator of *m* (t_0) is asymptotically efficient – may not perform well for all values within the range of the function *m*. Such an idea is intuitive; without knowledge of the smoothness of *m*, estimators designed to be optimal for one value of *t* may be very different from optimal estimators for another value of *t*. Cai and Low (2005) study efficient estimation of *m* over neighbourhoods of t_0 and show that an estimator constructed from wavelets is adaptive. The restriction that the scale is constant is often difficult to support with

financial data. A more realistic model, which Mercurio and Spokoiny (2004) study, models the asset return as a stochastic differential equation with drift 0 and ν_t varying over time. The time-varying scale is assumed to be constant over (short) intervals of time, but is otherwise unspecified. They construct a nonparametric estimator of the volatility from a kernel that performs local averaging and show that the resultant estimator is adaptive.

See Also

- efficiency bounds
- partial linear model
- semiparametric estimation

Bibliography

Bickel, P. 1982. On adaptive estimation. *Annals of Statistics* 10, 647–71.

Cai, T. and Low, M. 2005. Nonparametric estimation over shrinking neighborhoods: superefficiency and adaptation. *Annals of Statistics* 33, 184–213.

Chicken, E. and Cai, T. 2005. Block thresholding for density estimation: local and global adaptivity. *Journal of Multivariate Analysis* 95, 76–106.

Faraway, J. 1992. Smoothing in adaptive estimation. *Annals of Statistics* 20, 414–27.

Lepski, O. and Spokoiny, V. 1997. Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics* 25, 2512–46.

Mercurio, D. and Spokoiny, V. 2004. Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics* 32, 577–602.

Newey, W. 1988. Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics* 38, 301–39.

Portnoy, S. and Koenker, R. 1989. Adaptive L-estimation for linear models. *Annals of Statistics* 17, 362–81.

Steigerwald, D. 1992. On the finite sample behavior of adaptive estimators. *Journal of Econometrics* 54, 371–400.

Stein, C. 1956. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley*

Symposium on Mathematical Statistics and Probability, ed. J. Neyman. Berkeley: University of California Press.

How to cite this article

Steigerwald, Douglas G. "adaptive estimation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000235> doi:10.1057/9780230226203.0007

close print preview

The New Palgrave Dictionary of Economics Online

adaptive expectations

Michael Parkin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The adaptive expectations hypothesis may be stated most succinctly in the form of the equation:

$$E_t x_{t+1} = \sum_{i=0}^{\infty} \lambda(1-\lambda)^i x_{t-i};\ 0 < \lambda < 1 \qquad (1)$$

where E denotes an expectation, x is the variable whose expectation is being calculated and t indexes time. What this says is that the expectation formed at the present time, E_t of some variable, x, at the next future date, t+1, may be viewed as a weighted average of all previous values of the variable, x_{t-i}, where the weights, $\lambda\ (1-\lambda)^i$, decline geometrically. The weight attaching to the most recent, or current, observation is λ The above equation can be manipulated readily to deliver:

$$E_t x_{t+1} = E_{t-1} x_t + \lambda(x_t - E_{t-1} x_t). \qquad (2)$$

What this equation says is that, viewed from time t, the expected value of the variable, x at t+1, is equal to the value which, at time t–1 was expected for t, plus an adjustment for the extent to which the variable turned out to be different at t from the value which, viewed from date t – 1, had been expected. The change in the expectation is simply the fraction λ multiplied by the most recently observed forecast error. In this formulation, the adaptive expectations hypothesis is sometimes called the error learning hypothesis (see Mincer, 1969, pp. 83–90).

The adaptive expectations hypothesis was first used, though not by name, in the work of Irving Fisher (1911). The hypothesis received its major impetus, however, as a result of Phillip Cagan's (1956) work on hyperinflations. The hypothesis was used extensively in the late 1950s and 1960s in a variety of applications. L.M. Koyck (1954) used the hypothesis, though not in name, to study investment behaviour. Milton Friedman (1957), used it as a way of generating permanent income in his study of the consumption function. Marc Nerlove (1958) used it in his analysis of the dynamics of supply in the agricultural sector. Work on inflation and macro-economics in the 1960s was dominated by the use of this hypothesis. The most comprehensive survey of that work is provided by David Laidler and Michael Parkin (1975).

The adaptive expectations (or error learning) hypothesis became popular and was barely challenged from the middle-1950s through the late-1960s. It was not entirely unchallenged but it remained the only extensively-used proposition concerning the formation of expectations of inflation and a large number of other variables for something close to two decades. In the 1970s the hypothesis fell into disfavour and the rational expectations hypothesis became dominant.

The adaptive expectations hypothesis became and remained popular for so long for three reasons. First, in its error learning form it had the appearance of being an application of classical statistical inference. It looked like classical updating of an expectation based on new information.

Second, the adaptive expectations hypothesis was empirically easy to employ. Koyck (1954) showed how a simple transformation of an equation with an unobservable expectation variable in it could be rendered observable by performing what became a famous transformation bearing Koyck's name. If some variable, *y*, is determined by the expected future value of *x*, that is:

$$y_t = \alpha + \beta E_t x_{t+1} \tag{3}$$

where α and β are constants, then we can obtain an estimate of α and β by using a regression model in which equation (1) [or equivalently (2)] is used to eliminate the unobservable expected future value of *x*. To do this, substitute (1) into (3). Then write down an equation identical to (3) but for one period earlier. Multiply that second equation by $1-\lambda$ and subtract the result from (3) (Koyck, 1954, p. 22), to give:

$$y_t = \alpha\lambda + \beta\lambda x_t + (1-\lambda) y_{t-1} \tag{4}$$

An equation like this may be used to estimate not only the desired values of α and β but also the value of λ, the coefficient of expectations adjustment. Thus, economists seemed to have a very powerful way

of modelling situations in which unobservable expectational variables were important and of discovering speeds of response both of expectations to past events and of current events to expectations of future events.

Third, the adaptive expectations hypothesis seemed to work. That is, when equations like (4) were estimated in the wide variety of situations in which the hypothesis was applied (see above), ‘sensible’ parameter values for α, β, λ were obtained and, in general, a high degree of explanatory power resulted.

If the adaptive expectations hypothesis was so intuitively appealing, easy to employ, and successful, why was it eventually abandoned? There are three key reasons. First, the interpretation of the hypothesis as an application of classical inference came to be questioned, notably by John Muth (1960). Muth pointed out that the adaptive expectations hypothesis would only be optimal in the sense of delivering unbiased and minimum mean square error forecasts for a variable whose first difference was a first-order moving average process. Since this is likely to be a limited class of variables, the general validity of interpreting the adaptive expectations hypothesis as being consistent with classical inference came to be questioned. Second, in the area of macroeconomics, the adaptive expectations hypothesis was seen to be logically inconsistent with what came to be called the ‘natural rate hypothesis’ (Lucas, 1972). The latter hypothesis, that unemployment and other real variables are ultimately determined by real forces and not influenced by anticipations of inflation (at least not to a first-order) is so deeply entrenched in economics that the logical clash of the two hypotheses had to result in the modification of adaptive expectations (see Friedman, 1968, and Phelps, 1970). Third, and as almost always happens in scientific developments, a new, rational expectations alternative to adaptive expectations became available. The new theory had all the intuitive appeal of the old and, eventually, became equally tractable in empirical studies and began to show signs of success.

See Also

- cobweb theorem
- expectations
- hyperinflation
- Phillips curve
- rational expectations

Bibliography

Cagan, P. 1956. The monetary dynamics of hyper-inflation. In *Studies in the Quantity Theory of Money*, ed. Milton Friedman. Chicago: University of Chicago Press.

Fisher, I. 1911. *The Purchasing Power of Money*. New York: Macmillan (latest edition, A.M. Kelley, New York, 1963).

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(March), 1–17.

Koyck, L.M. 1954. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.

Laidler, D. and Parkin, M. 1975. Inflation: a survey. *Economic Journal* 85(December), 741–809.

Lucas, R.E. Jr. 1972. Econometric testing of the natural rate hypothesis. In *The Econometrics of Price Determination*, ed. Otto Eckstein, Washington, DC: Board of Governors of the Federal Reserve System.

Mincer, J. 1969. *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Muth, J.F. 1960. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55, 299–306.

Nerlove, M. 1958. *The Dynamics of Supply: Estimation of Farmers' Response to Price*. Baltimore: Johns Hopkins Press.

Phelps, E.S. (ed.) 1970. *Microeconomic Foundations of Employment and Inflation Theory*. New York: Norton.

How to cite this article

Parkin, Michael. "adaptive expectations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000028> doi:10.1057/9780230226203.0008

close print preview

The New Palgrave Dictionary of Economics Online

addiction

George Loewenstein and Scott Rick
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Research on addiction had already yielded a wide range of interesting and important findings when economists first arrived on the scene. The economic study of addiction was initiated by a seminal paper by Becker and Murphy (1988) which challenged the prevailing view of addiction as self-destructive, proposing instead a ‘rational account of addiction’. Although some empirical research has confirmed the model's critical prediction that anticipated increases in future prices will decrease current demand for a drug, more recent research by economists, stimulated by the prior work from other disciplines, has challenged some of the rational account's assumption and predictions.

Keywords

addiction, rational account of; Becker, G.; drugs; excise tax; forward price elasticity; rational behaviour; sin taxes; time consistency

Article

Economists were latecomers to the study of addiction, a concept which researchers in other disciplines usually define as including a loss of self-control, continuation of behaviour despite adverse consequences, and preoccupation or obsession with the substance or activity one is addicted to. Economists came late to the subject perhaps because the first two of these characteristics seem inconsistent with economists’ rational choice paradigm.
This may be exactly what spurred Gary Becker, along with coauthor Kevin Murphy, to propose, in 1988, a ‘rational account of addiction’, which stimulated much subsequent research and theorizing by economists. Although not the first economic account of addiction, Becker and Murphy's model (referred to henceforth as B&M) was certainly the most influential, and has spawned a very lively line of research, theorizing and debate about addiction by economists.

Contributions of disciplines other than economics

Prior to B&M, scientists in a range of disciplines had already developed a rich tradition of research on addiction. For example, early studies by psychopharmacologists identified the actions of addictive drugs in the brain, and subsequent research by neuroscientists has uncovered the neural pathways through which addictive activities derive their motivational power (see, for example, Gardner and James, 1999; Lyvers, 2000). Sociologists have also been major contributors, conducting ethnographic and life-course studies of drug users that have identified many of the social influences on drug use. Psychologists have studied the widest range of different facets of drug abuse, including biological underpinnings and social, cognitive and emotional dimensions, and have also been in the forefront when it comes to treatment. Psychologists, as well as other health professionals, have tested a great diversity of treatments for addiction, including residential treatment, counselling, psychotherapy, drug therapies such as methadone, nicotine patches and antidepressants, aversive conditioning, and hypnosis. Taken together, these diverse lines of research have yielded a number of important, and often counter-intuitive, findings.

- Historic use of different types of drugs exhibits ‘fads’, rising then falling in popularity, sometimes repeatedly for a specific drug.
- Most drug users do not just use a single drug, but many different drugs.
- Many if not most drug abusers also suffer from other psychiatric conditions, such as anxiety or mood disorders, schizophrenia or antisocial personality disorder.
- Much if not most quitting occurs outside of treatment.
- It is not short-term withdrawal from drugs (for example, for a few days) that most addicts find difficult, but long-term abstinence, which tends to be punctuated by episodes of ‘craving’ which create an almost overwhelming motivation for drug use.
- Episodes of craving are often triggered by ‘cues’ – people or other stimuli that the addict associates with drug use.
- While approximately 20 per cent of a sample of veterans reported being addicted to heroin in Vietnam, and 45 per cent reported narcotic use, only one per cent remained addicted, and two per cent reported using narcotics after returning home (Robins, 1973); this finding radically changed prevailing views of the incidence of recovery from heroin addiction.
- Humans and other mammals voluntarily self-administer most of the same chemical compounds. (Hallucinogens, which some humans seek out but most animals avoid, are a major exception.)
- Although a small number of intense users account for a large fraction of drug use, most drug users consume at moderate or low rates, and do not become addicted in the sense of losing control, suffering adverse consequences or becoming obsessed with drug-taking.
- Many of the adverse health effects of illicit drugs, such as opiates, do not stem from physical effects of the drugs themselves, but from the difficulty of financing an illegal, and hence typically expensive, habit.
- Most addictions begin when people are in their teens or early twenties, and addicts often ‘mature out’ – quitting when they reach middle age. People rarely become addicted for the first time in middle or old age.

In addition to generating a wide range of interesting and important findings, researchers in disciplines other than economics have proposed a variety of theoretical perspectives on addiction. Some

perspectives place great importance on the pleasure of drug-taking, the pain of withdrawal, or the motivational force of ‘cue-conditioned’ craving, while others view drug use as a form of self-medication for psychiatric conditions such as depression.

For better or for worse, economists’ focus on addiction has been much narrower, at both the theoretical and the empirical levels. Most empirical work has involved estimating price elasticities of demand for drugs (often using aggregate consumption data), and most theoretical work has involved some type of generalization of Becker and Murphy's perspective.

Becker and Murphy's model

In Becker and Murphy's rational model of addiction, utility from an addictive good, $c(t)$, is assumed to depend on consumption of that good and on the degree of addiction $S(t)$. $S(t)$ changes according to the function $\dot{S}(t) = c(t) - \delta S(t)$, where the first term represents the impact of engaging in the addictive good on one's level of addiction, and the second represents the natural decline in addictedness when one desists. The individual is assumed to trade off consumption of the addictive good against consumption of other (non-addictive) goods, discounting for time delay in the conventional (exponential) fashion. The central insight of B&M is that people treat addictive goods no differently from the way they treat any good whose utility depends on consumption over time, trading them off against other goods based on current and future (anticipated) prices.

This model can accommodate a number of features of classical addiction, such as that being addicted lowers instantaneous utility $u_S < 0$, that it increases the instantaneous marginal utility of taking the drug $u_{cS} > 0$. Solving the model yields a number of implications, most importantly that it can be rational for an individual to maintain a positive rate of consumption of an addictive good.

Empirical tests of B&M have focused on the strong prediction that anticipated changes in future prices affect the current behaviour of addicts, which is counter-intuitive given that addicts are commonly seen as behaving myopically. The model is therefore typically tested by estimating what could be called the ‘forward price elasticity’ of various addictive substances. Consistent with Becker and Murphy's model, negative forward price elasticities have been found for alcohol, cigarettes, marijuana, opium, heroin and cocaine (for a review, see Pacula and Chaloupka, 2001), although the effect appears to be more consistent for adults than for youth.

Moving beyond Becker and Murphy

In proposing their rational account of addiction, Becker and Murphy initiated the study of addiction among economists, and made the key point that it is useful to think of addicts as solving a forward-looking optimization problem. However, the B&M model fails to incorporate a number of important features of addiction, and is either inconsistent with or fails to predict many salient features of addiction, including some of the stylized facts listed above. Responding to these limitations, economists have built upon the B&M model by relaxing some of its most extreme assumptions or incorporating more realistic assumptions that are often inspired by research in other disciplines.

One important generalization has been to examine the implications of relaxing the assumption of exponential time discounting. Gruber and Koszegi (2001; 2004), for example, propose a model in which

time-inconsistent addicts have self-control problems: they would like to quit using but cannot force themselves to do so (see also O'Donoghue and Rabin, 1997). As in B&M, Gruber and Koszegi's model predicts that a rise in current or anticipated excise taxes will reduce use of addictive substances. However, although the models make similar behavioural predictions, they interpret the hedonic consequences of altered usage behaviour differently. B&M predicts that taxes on addictive substances – 'sin taxes' – make addicts worse off since the price of a good that they enjoy has risen. Gruber and Koszegi's model, on the other hand, predicts that the tax makes time-inconsistent addicts better off since it provides a valuable self-control device.

Since behavioural data cannot distinguish between the models, Gruber and Mullainathan (2005) bypassed the standard practice of measuring the impact of policy interventions by estimating price elasticities in favour of directly examining the impact of these interventions on subjective well-being. They did so by matching cigarette excise taxation data to surveys from the United States and Canada that contain data on self-reported happiness. Consistent with Gruber and Koszegi's model, Gruber and Mullainathan (2005) found that excise taxes on cigarettes make smokers happier.

Another implication of time inconsistency involves purchasing patterns. The B&M model predicts that addicts will behave in a time-consistent fashion and hence will buy in bulk to save time and money in satisfying their anticipated long-term habit. Wertenbroch (1998; 2003), however, found that consumers – even those who are not liquidity-constrained – often purchase 'vice' items, such as cigarettes, in small quantities in an attempt to control their intake of the harmful substance.

Other research has questioned the assumption that addicts begin drug taking with full knowledge of the consequences. For example, Slovic (2000a; 2000b) has argued that people take up cigarette smoking in part because they underestimate the health risks, although Viscusi (2000) counters that any error is actually in the opposite direction – that smokers *overestimate* the health risks of smoking. Pointing to a somewhat different type of underestimation, Loewenstein (1999) has argued, based on a wide range of evidence, that potential drug users underestimate their own proneness to addiction because they underestimate the motivational force of drug craving.

Finally, a recent line of theoretical models, while also building on the insights of Becker and Murphy, has incorporated evidence from the psychological literature on cue-conditioned craving and from neuroscience. For example, Laibson (2001) proposes a model of addiction that incorporates the role of cue-conditioned craving. In his model, environmental cues that become associated with drug use, when encountered by an ex-addict, produce surges of craving (like sudden changes in $S(t)$ in B&M). Bernheim and Rangel (2004) develop a model of addiction that is particularly closely grounded in neuroscience research and that is perhaps the most radical departure from B&M. Their model is based on the idea that repeated experience with drugs sensitizes individuals to environmental cues that trigger mistaken usage.

So far, economists are still playing catch-up with researchers in other disciplines when it comes to their understanding of addiction or their influence on policy. Thus, a large fraction of empirical research on drug use by economists has focused on price elasticities. While price is one determinant of drug use, it is arguably not the most important, or even the most amenable to manipulation through the instruments of policy. Nevertheless, economic models of addiction have made great strides, building on Becker and Murphy's seminal contribution with new models that incorporate many of the insights and findings generated by research in other disciplines.

See Also

- Becker, Gary S.
- intertemporal choice
- rational behaviour

We thank Caroline Acker, Ted O'Donoghue and Antonio Rangel for helpful suggestions.

Bibliography

Becker, G. and Murphy, K. 1988. A theory of rational addiction. *Journal of Political Economy* 96, 675–700.

Bernheim, B. and Rangel, A. 2004. Addiction and cue-triggered decision processes. *American Economic Review* 94, 1558–90.

Gardner, E. and James, D. 1999. The neurobiology of chemical addiction. In *Getting Hooked: Rationality and Addiction*, ed. J. Elster and O.-J. Skog. Cambridge: Cambridge University Press.

Gruber, J. and Koszegi, B. 2001. Is addiction 'rational?' Theory and evidence. *Quarterly Journal of Economics* 116, 1261–305.

Gruber, J. and Koszegi, B. 2004. A theory of government regulation of addictive bads: tax levels and tax incidence for cigarette excise taxation. *Journal of Public Economics* 88, 1959–87.

Gruber, J. and Mullainathan, S. 2005. Do cigarette taxes make smokers happier? *Advances in Economic Analysis & Policy* 5, 1–43.

Laibson, D. 2001. A cue-theory of consumption. *Quarterly Journal of Economics* 116, 81–119.

Loewenstein, G. 1999. A visceral account of addiction. In *Getting Hooked: Rationality and Addiction*, ed. J. Elster and O.-J. Skog. Cambridge: Cambridge University Press.

Lyvers, M. 2000. 'Loss of control' in alcoholism and drug addiction: a neuroscientific interpretation. *Experimental and Clinical Psychopharmacology* 8, 225–45.

O'Donoghue, T. and Rabin, M. 1997. Addiction and self control. In *Addiction: Entries and Exits*, ed. J. Elster. New York: Russell Sage Foundation.

Pacula, R. and Chaloupka, F. 2001. The effects of macro-level interventions on addictive behavior. *Substance Use and Misuse* 36, 1901–22.

Robins, L. 1973. *The Vietnam Drug User Returns*. Washington, DC: U.S. Government Printing Office.

Slovic, P. 2000a. What does it mean to know a cumulative risk? Adolescents' perceptions of short-term and long-term consequences of smoking. *Journal of Behavioral Decision Making* 13, 259–66.

Slovic, P. 2000b. Rejoinder: the perils of Viscusi's analyses of smoking risk perceptions. *Journal of Behavioral Decision Making* 13, 273–76.

Viscusi, W. 2000. Comment: the perils of qualitative smoking risk measures. *Journal of Behavioral Decision Making* 13, 267–71.

Wertenbroch, K. 1998. Consumption self-control via purchase quantity rationing of virtue and vice. *Marketing Science* 17, 317–37.

Wertenbroch, K. 2003. Self-rationing: self-control in consumer choice. In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, ed. G. Loewenstein, D. Read and R. Baumeister. New York: Russell Sage Foundation.

How to cite this article

Loewenstein, George and Scott Rick. "addiction." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000252> doi:10.1057/9780230226203.0009

close print preview

The New Palgrave Dictionary of Economics Online

adjustment costs

Aubhik Khan and Julia K. Thomas

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys the use of adjustment frictions in macroeconomic research, exploring the consequences of convex and non-convex adjustment costs for firm-level decisions and the dynamics of macroeconomic aggregates. The mechanics of these frictions are illustrated using several prominent examples including the partial adjustment model of employment, the q-theoretic investment model, and lumpy adjustment models of investment and employment. We also review the (S,s) inventory model, where stock accumulation is explained as the result of fixed delivery costs, and briefly discuss (S,s) decision rules arising from piecewise-linear costs in the context of capital irreversibility and firing taxes.

Keywords

adjustment costs; adjustment hazards; aggregate nonlinearities; business cycles; capital irreversibility; convex cost functions; distributed lags; dynamic stochastic equilibrium analysis; Euler equations; equilibrium; frictions; intermediate goods; inventory policies; investment theory; linear quadratic inventory models; lumpy investment; market-clearing relative prices; monetary non-neutralities; neoclassical investment theory; nonlinear microeconomic adjustment; partial adjustment; piecewise-linear adjustment costs; production functions; quadratic cost functions; rational expectations; (S,s) decision rule; (S,s) policies; Tobin's q; total factor productivity

Article

Across a wide body of macroeconomic research, the interest in adjustment costs has been largely utilitarian. In designing theoretical models to organize our understanding of patterns observed in the data, we make hard choices about which of the many elements affecting the decisions of actual firms and households and the outcomes of their market interactions to include. Given their necessary simplicity, we often find that the predictions of the theoretical economies we are able to analyse are too stark relative to the behaviour observed in actual economies. Thus, in a variety of settings we have adopted adjustment costs in our economic laboratories to summarize omitted frictional elements that

reduce, delay or protract changes in the demand and supply of final goods and their factor inputs in response to changes in economic conditions.

In these few pages, we describe the mechanics of commonly used adjustment costs and briefly discuss their role in several leading macroeconomic applications. Since a comprehensive survey is beyond the scope of this article, many important applications have been excluded. However, where possible we direct the reader to influential research on these topics.

1 Convex costs

Until relatively recently, most macroeconomic research involving adjustment costs emphasized the use of convex cost functions to penalize swift changes in aggregate variables and thereby induce gradual movements over time. Historically, models with convex adjustment costs were developed as a theoretical foundation to explain why the inclusion of lagged dependent variables in empirical models of factor demand led to sharp improvements in their econometric performance. While early researchers had found decision-theoretic models based on static demand theory unable to account for the serial correlation observed in aggregate employment and investment, these same models performed relatively well when they were augmented with ad hoc distributed lags of the dependent variable or its theoretical determinants (as in the flexible accelerator model of Koyck, 1954, or the flexible user-cost model of Hall and Jorgenson, 1967). These lags were broadly motivated by the idea that certain frictions prevent firms from immediately attaining their chosen employment or capital levels, instead engendering gradual, *partial adjustment* towards these target levels over time.

For example, by assuming that firms adjusted their workforces at constant rate $\lambda \in (0, 1)$ towards the target implied by static demand theory, N_t^*, current employment could be written as a distributed lag of previous target employments:

$$N_t = \lambda N_t^* + (1-\lambda) N_{t-1} = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j N_{t-j}^*. \tag{1}$$

To implement such partial adjustment models, researchers replaced the distributed lag of unobservable targets with distributed lags of each observable series the theory suggested should influence them – for instance, real wages. In this way, lags of the determinants of demand were introduced into the estimation equation, thus introducing the empirically desirable serial correlation.

Without some theoretical basis to explain their empirical success, partial adjustment models might have been abandoned quickly. A partial resolution arrived in the mid- to late 1960s with the application of capital adjustment costs in models of investment (see Eisner and Strotz, 1963; Lucas, 1967; Gould, 1968; Treadway, 1971). There, gradual aggregate adjustment broadly consistent with the analogue to (1) was obtained by assuming that, beyond other costs associated with the acquisition of capital (for

example, user costs), the very act of adjusting the capital stock incurred real output costs. These costs, $\Phi(k', \cdot k)$, were strictly increasing and convex in the distance between the chosen new level of capital and the current level, $|k' - k|$, thereby implying a smoothly rising marginal adjustment cost in the size of the current adjustment. As such, they introduced dynamic elements into the firm's previously static decision problem and led it to smooth its investment activities over time. Nonetheless, so long as the treatment of expectations was incomplete, the mapping to a partial adjustment equation could not be robustly established.

The work of Sargent (1978) extended the theory in the context of employment adjustment by showing how, under rational expectations, the partial adjustment model could be derived from the profit maximization problem of a firm facing quadratic adjustment costs. To simplify the problem somewhat, consider a firm that enters any period with employment n_{t-1} and incurs costs, $\Phi(n_t, n_{t-1}) \equiv \frac{\phi}{2}(n_t - n_{t-1})^2$, in altering its workforce for production. Next, assume that the firm's production function is quadratic, $f(n_t, z_t) \equiv (f_0 + z_t)n_t - \frac{f_1}{2}n_t^2$, where $f_0 > 0$, $f_1 > 0$, and z is a serially correlated exogenous productivity process, as is the real wage, w. Discounting its future earnings by $\beta \in (0, 1)$ and given initial employment n_{-1}, the firm selects $\{n_t\}_{t=0}^{\infty}$ to maximize its expected present discounted value, $E\left[\sum_{t=0}^{\infty}\beta^t(f(n_t - z_t) - w_t n_t - \Phi(n_t, n_{t-1}))|z_0, w_0\right]$, arriving at a sequence of Euler equations:

$$\beta E_t n_{t+1} - \left(1 + \beta + \frac{f_1}{\phi}\right)n_t + n_{t-1} = \frac{w_t - a_t - f_0}{\phi}.$$

If we isolate the two real roots of this second-order stochastic difference equation, the solution is precisely (1) above, with target employment in each date given by

$$N_t^* = \left[E_t\sum_{j=0}^{\infty}(\beta/\lambda)^j(x_z z_{t+j} - x_w w_{t+j})\right] \quad (2)$$

and the parameters λ, χ_z and χ_w determined by the adjustment cost parameter φ, the discount factor β, and the parameters of the production function.

For researchers implementing equations like (1), an important contribution of Sargent's model was in illustrating how the very features that linked current employment to its lagged determinants also

necessarily divorced each date's target, N_t^*, from the statically derived optima assumed in early partial adjustment estimations. Notice that the firm's target in (2) involves expectations of each variable affecting the future value marginal product of labour, because, given adjustment costs, this current choice influences its future level of employment. Moreover, as an increase in the adjustment cost parameter, φ, shifts the marginal adjustment cost schedule upward at all dates, it not only implies a slower adjustment rate (lower λ) but also increases the influence of these expectations of future variables in the determination of the current target.

Across the many models including convex adjustment costs, quadratic cost functions have been by far the most common specification, essentially for sake of tractability. Note that, given the quadratic form of $\Phi(n_t, n_{t-1})$ above, firms' decision rules described by (1) and (2) are linear. As such, they aggregate conveniently to represent economy-wide factor demand in partial adjustment models. (Hamermesh, 1989, and Hamermesh and Pfann, 1996, discuss the role of these costs in partial adjustment models of employment demand. Chirinko, 1993, Hassett and Hubbard, 1997, and Caballero, 1999, survey their use in empirical investment equations. Hall, 2004, estimates an industry-level model of production with quadratic adjustment costs applied to both labour and capital.)

A similar cost function appears in the history of q-theoretic investment models, unifying neoclassical investment theory with the theory of Brainard and Tobin (1968) and Tobin (1969), which holds that investment should be positively related to average Q, the ratio of the value of the firm relative to its capital stock. Appending the neoclassical model with a general convex adjustment cost function, Abel (1979) moved to reconcile the two theories by showing that the expected discounted marginal value of capital for a firm, marginal q, is sufficient to determine its investment rate. The reconciliation was complete when Hayashi (1982) showed that average Q is identical to marginal q if firms are perfectly competitive and both the production function and $\Phi(k', k)$ are linearly homogenous (for example, $\Phi(k', k) = \frac{\phi}{2}\frac{(k'-k)^2}{k}$).

Since the mid-1980s, macroeconomic analysis has become firmly grounded in dynamic stochastic equilibrium analysis. Nonetheless, the gradual movements implied by equilibrium relative price changes have often proven inadequate in reconciling models to data; thus, convex costs have continued to appear. A famous early application to capital adjustment is the industry equilibrium study of investment by Lucas and Prescott (1971). More recently, examples of general equilibrium models adopting these frictions may be found in almost every field of macroeconomics.

2 Non-convex costs

Despite their relative success in reproducing the persistence of aggregate series, empirical models based on convex adjustment costs have fared poorly along other dimensions. For example, estimations of the neoclassical investment model attribute very low explanatory power to average Q and assign large coefficients to adjustment cost parameters in explaining changes in investment (Chirinko, 1993; Caballero, 1999). Large estimates of adjustment costs, which in turn imply implausibly slow adjustment speeds, are also a recurring problem for linear quadratic inventory models (Ramey and West, 1999). Elsewhere, the sharp difference between rates of employment adjustment estimated from high-frequency

firm-level data and those estimated from low-frequency aggregate data suggests spatial and temporal bias inconsistent with the common assumption of symmetric quadratic adjustment costs (Hamermesh and Pfann, 1996). Moreover, there is mounting microeconomic evidence suggesting that the predominant adjustment frictions confronting firms in actual economies may be non-convex, rather than convex, in nature.

Contrary to the smooth, continual adjustments implied by convex cost models, recent microeconomic studies reveal that firm-level factor adjustment exhibits long periods of relative inactivity punctuated by infrequent and large, or lumpy, changes in stocks. Examining capital adjustment in a 17-year sample of large, continuing US manufacturing plants, Doms and Dunne (1998) find that roughly 25 per cent of the typical plant's cumulative investment occurs in a single year, and more than half of plants exhibit capital adjustment of at least 37 per cent within one year. Using a similar dataset, Cooper, Haltiwanger and Power (1999) provide additional evidence of lumpy investment, and they show that the conditional probability of a large investment episode rises in the time since the last such episode. Microeconomic evidence of non-smooth employment adjustment is abundant (see Hamermesh and Pfann, 1996). For example, examining monthly data on employment and output across seven US manufacturing plants between 1983 and 1987, Hamermesh (1989) finds that plant-level employment remains roughly constant over long periods while production fluctuates. These long episodes of constancy are broken by infrequent but large jumps, at times roughly coinciding with the largest output fluctuations. (Interestingly, while the convex cost model is inconsistent with the lumpy employment adjustments at each plant, Hamermesh finds that it represents the aggregate of employment – and production – across plants reasonably well.) Beginning with Scarf (1960), a number of theoretical studies have shown that precisely this variety of nonlinear microeconomic adjustment can arise when firms are confronted with non-convex adjustment technologies.

2.1 (S,•s) stock adjustment

Scarf (1960) provided the earliest formal analysis of microeconomic adjustment behaviour in the presence of non-convex adjustment costs. There, the adjustment cost was a simple fixed cost, $\Phi > 0$, incurred at any time a firm wished to adjust its stock of inventories. (Beginning with the work of Barro, 1972, and Sheshinski and Weiss, 1977, fixed costs have also been used to develop models of (*S*,•*s*) firm-level price adjustment. Early studies examining the potential for monetary non-neutralities in such settings include Sheshinski and Weiss, 1983; Caplin and Spulber, 1987; and Caplin and Leahy, 1991. More recent general equilibrium analyses include Caplin and Leahy, 1997; Dotsey, King and Wolman, 1999; Gertler and Leahy, 2006; and Golosov and Lucas, forthcoming.) We briefly review the model below.

Consider a retail firm entering any period with inventories, $y > 0$, of a homogenous good available for sale. The firm faces stochastic demand, ξ, drawn from a time-invariant distribution $F(\xi)$, and the value of its sales is p min $\{y, \xi\}$. At the end of the period, it may place an order $x > 0$ to increase its available stock for the next period; $y' = y - \min\{y, \xi\} + x$. The cost of any such order is $\Phi + cx$, where $c > 0$ represents the unit cost of the good held in inventory. By proving K-concavity of the value function,

Scarf was able to establish that the firm's optimal decision rule takes the following one-sided (*S*,•*s*) form. (Scarf, 2005, shows this decision rule generalizes to a setting where the firm selectively sells its inventories with the option of leaving some demand unsatisfied. See Dixit, 1993, for a characterization of two-sided (*S*,•*s*) policies arising in continuous time settings involving fixed and piecewise linear adjustment costs.)

$$x = \begin{cases} 0 & \text{for} \quad y \in (s, S] \\ S - y & \text{for} \quad y \leq s \end{cases}.$$

To avoid repeatedly incurring fixed costs, the firm places no orders so long as its sales do not move its stock outside the interval (*s*,•*S*]. Only when its inventories have fallen to the lower threshold, *s*, does it take action, resetting its stock to *S*. Thus, the increasing returns adjustment technology implied by fixed order costs induces infrequent and relatively large, or lumpy, orders.

Just as firm-level data indicates lumpiness in microeconomic capital and employment adjustment, there are a number of studies suggesting that firms in both manufacturing and trade manage their inventories according to (*S*,•*s*) policies resembling that obtained in Scarf's path-breaking analysis (for example, Mosser, 1991; Hall and Rust, 2000). Nonetheless, despite the empirical difficulties associated with convex cost inventory models (Blinder and Maccini, 1991; Ramey and West, 1999), the implications of firm-level inventory policies under non-convex adjustment costs have been left relatively unexplored by macroeconomists. To reproduce the relatively smooth changes observed in the aggregate, such models necessarily involve a distribution of firms over inventory levels. As this distribution becomes part of the economy's aggregate state vector, the resulting high dimensionality makes it difficult to determine equilibrium prices, including real wages and interest rates. It is this basic problem that has generally dissuaded researchers from undertaking dynamic stochastic general equilibrium analyses of environments involving non-convexities, among them the (*S*,•*s*) inventory model.

One exception to this is found in Fisher and Hornstein (2000). Building on the work of Caplin (1985) and Caballero and Engel (1991), who study the aggregate implications of exogenous (*S*,•*s*) policies across firms, Fisher and Hornstein construct an environment that endogenously yields time-invariant one-sided (*S*,•*s*) adjustment rules and a constant order size per adjusting firm. This allows them to tractably study (*S*,•*s*) inventory policies in general equilibrium without confronting substantial heterogeneity across firms. More generally, in models involving time-varying two-sided (*S*,•*s*) policies, the heterogeneity becomes more cumbersome, as in Khan and Thomas' (2006a) general equilibrium business cycle study. There, at the start of any period, each firm observes the current state and then chooses whether to order intermediate goods for use in production. Given this timing, alongside positive real interest rates, inventories would never be held in the absence of some friction. However, by confronting firms with idiosyncratic order costs independent of their chosen order sizes, continual orders are deterred, and (*S*,•*s*) inventory adjustment adopted. Based on the results of their calibrated model, Khan and Thomas conclude that such non-convex costs can be quite successful in explaining not only the existence of aggregate inventories but also their cyclical dynamics.

2.2 Implications for aggregate investment

Non-convex adjustment costs imply distributed lags in aggregate series similar to those generated by convex costs, because they stagger the lumpy adjustments undertaken by individual firms in response to shocks (King and Thomas, 2006). However, they are distinguished by their potential for aggregate nonlinearities, which has generated particular interest within investment theory. A number of influential partial equilibrium studies (Caballero and Engel, 1999; Cooper, Haltiwanger and Power, 1999; Caballero, Engel and Haltiwanger, 1995) have argued that investment models with non-convex costs empirically outperform convex cost models because they can deliver disproportionately sharp changes in aggregate investment demand following large aggregate shocks. (Caballero and Engel, 1993, and Caballero, Engel, and Haltiwanger, 1997, arrive at similar conclusions in the context of employment adjustment.)

Caballero and Engel (1999) examine generalized (S,•s) policies rationalized by stochastic fixed adjustment costs, φ, distributed i.i.d. across firms and over time. In this environment, a firm's capital, k, becomes part of its state vector alongside its total factor productivity, z. Moreover, microeconomic adjustment becomes probabilistic; firms with the same current gap between actual and target capital do not necessarily behave identically; rather, those with relatively low φ draws are more likely to alter their capital than those drawing high costs. If we transform Caballero and Engel's gap-based analysis to reflect the firm-level state, (k, z), the implication is an adjustment hazard, $\Lambda(k, z)$, indicating what fraction of each group of firms sharing a common current state will choose to adjust their capital to a common target, $k^*(z)$. The resulting generalized (S,•s) adjustment model allows convenient aggregation and has been studied in a variety of settings. (Dotsey, King and Wolman, 1999, apply this basic framework to price adjustment, Thomas, 2002, adopts it in an equilibrium business cycle model with lumpy investment, and King and Thomas, 2006, use it to examine employment adjustment.)

To understand how this mechanism can affect the dynamics of aggregate investment, consider the following simple partial equilibrium example described by Khan and Thomas (2003). Assume that total factor productivity, z, is a Markov process common to all firms. If there have been no aggregate shocks for many periods, the distribution of firms will have support at $k^*(z)$, $(1-\delta)k^*(z)$, $(1-\delta)^2k^*(z)$, and so on. As a firm's capital stock depreciates further below the target, $k^*(z)$, the maximum adjustment cost it will accept to reset its capital stock to that target, $\varphi(k, z)$, rises. Thus, the adjustment hazard, $\Lambda(k, z)$, is increasing in the distance $\left|k^*(z) - k\right|$. Finally, the total measure of adjusting firms is $\int\Lambda(k, z)\mu(dk)$, and aggregate investment is $I = \int\Lambda(k, z)(k^*(z) - (1-\delta)k)\mu(dk)$.

Suppose that a negative aggregate shock reduces z to z_L, thereby reducing expected future marginal productivity of capital. This causes a downward shift in the target stock, placing it strictly within the existing range of capital held by firms. Thus, $\Lambda(k, z)$ falls for many firms, rising only for those with the highest levels of capital. As a result, the total adjustment rate can actually fall, thereby dampening the fall in aggregate investment demand implied by the reduced target. By contrast, when a positive technology shock raises z to z_H, the target capital rises above that currently held by any firm. This increases the total adjustment rate, compounding the effect of the raised target to which firms adjust.

More generally, this example illustrates that, when there is an aggregate shock, and thus a change in the target, higher moments of the distribution of capital across firms matter in determining movements in aggregate investment, because the adjustment hazard is a non-trivial function of capital. (This is an important distinction relative to the convex cost/ partial adjustment model. Rotemberg, 1987, shows its aggregate dynamics are reproduced by a model where individual firms adjust infrequently, but all face a common probability of undertaking adjustment, independent of their individual states. Given this constant hazard, only the first moment of the distribution is relevant in determining aggregate changes.)
Alternatively, in the language of Caballero (1999, p. 841), microeconomic non-convexities can generate an important ‘time-varying/history-dependent aggregate elasticity’ of investment to shocks by allowing changes in the synchronization of firms' capital adjustments.
Although findings like those above echo throughout partial equilibrium studies involving lumpy adjustments, the omission of market-clearing relative prices (for example, equilibrium interest rates) may be critical to the inferred macroeconomic importance of non-convex factor adjustment costs. Significant aggregate nonlinearities can only occur if adjustment hazards exhibit large changes in response to shocks. Clearly, from the example above, such changes depend entirely on the extent to which $k^*(z)$ responds to changes in z. However, just as the capital adopted by a representative firm facing no adjustment costs varies far less when prices adjust to clear all markets, Thomas (2002) and Khan and Thomas (2003; 2006b) show that the target capital(s) selected by firms facing non-convex costs exhibit changes an order of magnitude smaller in general equilibrium. Because large movements in target capital, and hence in aggregate investment demand, would imply intolerable consumption volatility for households (at least in the closed-economy settings examined in these studies), they do not occur in equilibrium. Instead, small changes in relative prices serve to discourage sharp changes in $k^*(z)$, thereby preventing large synchronizations in firms' investment timing and leaving the aggregate series largely unaffected by the microeconomic lumpiness caused by non-convex adjustment costs.

3 Piecewise-linear costs

Among the adjustment frictions commonly applied in macroeconomic research, we have thus far omitted an important type of convex costs, namely, piecewise-linear adjustment costs, which are often associated with partial irreversibilities in investment and employment. As these costs have quite different implications from those described in section 1, we briefly discuss them here. Like non-convex costs, piecewise-linear costs lead to (S,•s) decision rules. However, as they yield no increasing returns in the adjustment technology, they do not in themselves cause lumpiness. Rather, when the firm's relevant state variable reaches the lower or upper bound of its tolerated region of inaction, the firm undertakes small adjustments to maintain it at that bound. (To explore the extreme case of complete irreversibility, see Pindyck, 1988, for an analysis that emphasizes the option value of waiting to invest, or Bertola, 1998, for a characterization of firm decision rules using standard dynamic programming. Dixit and Pindyck (1994) provide a comprehensive survey of this literature.)
Partial irreversibilities have been widely examined in investment theory as an explanation for the common empirical finding that investment is insensitive to Tobin's q. Abel and Eberly (1994) characterize firm-level investment when the purchase price of capital, p_K^+, exceeds its sale price, p_K^-

(and there are flow-fixed and convex adjustment costs). They show that this cost structure makes investment a nonlinear function of marginal *q*, implying a range of values over which the firm does not invest. (Veracierto, 2002, solves a general equilibrium business cycle model where the resale price of capital goods is a constant fraction of the purchase price. Examining a wide range of values for this irreversibility parameter, he concludes that such frictions have no quantitatively significant effects for business cycle dynamics.) Elsewhere, in the context of employment adjustment, a simple example of piecewise-linear costs is an environment where firms incur no adjustment costs in increasing their employment, but pay a tax of $\phi > 0$ per worker fired. The implications of such firing costs for aggregate employment are theoretically ambiguous. While their direct effect is to discourage firing, they also induce a reluctance to hire. Bentolila and Bertola (1990) provide an early analysis suggesting that the direct effect dominates, while Hopenhayn and Rogerson (1993) find the converse.

4 Conclusion

Throughout the history of their use, the primary purpose of adjustment costs has been to reduce the distance between model-generated and actual economic time series. Because they largely represent implicit costs of forgone output, we have little ability to directly measure adjustment frictions. Thus, when we adopt them to enhance the empirical performance of our models, the resulting improvements are, in some sense, a measure of our ignorance.

As suggested by the discussion above, the existence and size of particular adjustment frictions has typically been inferred from the extent to which they modify dynamic behaviour within a specific model to more closely resemble that in the data. This raises an obvious, but sometimes forgotten, point. Adjustment costs derived within a given class of model may be quite inappropriate in a second, distinct class of model. For example, the relative sizes of various types of adjustment frictions needed to reconcile theoretical and actual microeconomic data can differ sharply depending on the specification of equilibrium and firm-level shocks.

See Also

- inventory investment
- irreversible investment

Bibliography

Abel, A. 1979. *Investment and the Value of Capital*. New York: Garland.

Abel, A. and Eberly, J. 1994. A unified model of investment under uncertainty. *American Economic Review* 84, 1369–84.

Abel, A. and Eberly, J. 2002. Investment and q with fixed costs: an empirical analysis. Working paper, Kellogg School of Management, Northwestern University.

Barro, R. 1972. A theory of monopolistic price adjustment. *Review of Economic Studies* 39, 17–26.

Bentolila, S. and Bertola, G. 1990. Firing costs and labour demand: how bad is Eurosclerosis? *Review of Economic Studies* 57, 381–402.

Bertola, G. 1998. Irreversible investment. *Research in Economics* 52, 3–37.

Bertola, G. and Caballero, R. 1992. Irreversibility and aggregate investment. *Review of Economic Studies* 61, 223–46.

Blinder, A. and Maccini, L. 1991. Taking stock: a critical assessment of recent research on inventories. *Journal of Economic Perspectives* 5(1), 73–96.

Brainard, W. and Tobin, J. 1968. Pitfalls in financial model-building. *American Economic Review* 58, 99–122.

Caballero, R. 1999. Aggregate investment. In *Handbook of Macroeconomics*, vol. IB, ed. M. Woodford and J. Taylor. Amsterdam: North-Holland.

Caballero, R. and Engel, E. 1991. Dynamic (S,s) economies. *Econometrica* 59, 1659–86.

Caballero, R. and Engel, E. 1993. Microeconomic adjustment hazard and aggregate dynamics. *Quarterly Journal of Economics* 82, 359–83.

Caballero, R. and Engel, E. 1999. Explaining investment dynamics in U.S. manufacturing: a generalized (S,s) approach. *Econometrica* 67, 783–826.

Caballero, R., Engel, E. and Haltiwanger, J. 1995. Plant-level adjustment and aggregate investment dynamics. *Brookings Papers on Economic Activity* 1995(2), 1–39.

Caballero, R., Engel, E. and Haltiwanger, J. 1997. Aggregate employment dynamics, building from microeconomic evidence. *American Economic Review* 87, 115–37.

Caplin, A. 1985. The variability of aggregate demand with (S,s) inventory policies. *Econometrica* 53, 1395–410.

Caplin, A. and Leahy, J. 1991. State-dependent pricing and the dynamics of money and output. *Quarterly Journal of Economics* 106, 683–708.

Caplin, A. and Leahy, J. 1997. Aggregation and optimization with state-dependent pricing. *Econometrica* 65, 601–26.

Caplin, A. and Spulber, D. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102, 703–25.

Chirinko, R. 1993. Business fixed investment spending: a critical survey of modelling strategies, empirical results, and policy implications. *Journal of Economic Literature* 31, 1875–911.

Cooper, R. and Haltiwanger, J. 2006. On the nature of capital adjustment costs. *Review of Economic Studies* 73, 611–33.

Cooper, R., Haltiwanger, J. and Power, L. 1999. Machine replacement and the business cycle, lumps and bumps. *American Economic Review* 89, 921–46.

Dixit, A. 1993. Choosing among alternative discrete investment projects under uncertainty. *Economics Letters* 41, 265–68.

Dixit, A. and Pindyck, R. 1994. *Investment Under Uncertainty*. Princeton: Princeton University Press.

Doms, M. and Dunne, T. 1998. Capital adjustment patterns in manufacturing plants. *Review of Economic Dynamics* 1, 409–29.

Dotsey, M., King, R. and Wolman, A. 1999. State-dependent pricing and the general equilibrium dynamics of money and output. *Quarterly Journal of Economics* 114, 655–90.

Eisner, R. and Strotz, R. 1963. Determinants of business investment. In Commission on Money and Credit, *Impacts of Monetary Policy*. Englewood Cliffs, NJ, Prentice–Hall.

Fisher, J. and Hornstein, A. 2000. (S,•s) inventory policies in general equilibrium. *Review of Economic Studies* 67, 117–45.

Gertler, M. and Leahy, J. 2006. A Phillips curve with an Ss foundation. Working Paper No. 11971. Cambridge. MA: NBER.

Golosov, M. and Lucas, R. (forthcoming). Menu costs and Phillips curves. *Journal of Political Economy*.

Gould, J. 1968. Adjustment costs in the theory of investment of the firm. *Review of Economic Studies* 35, 47–56.

Hall, G. and Rust, J. 2000. An empirical model of inventory investment by durable commodity intermediaries. In *Carnegie-Rochester Conference Series on Public Policy* 52, 171–214.

Hall, R. 2004. Measuring factor adjustment costs. *Quarterly Journal of Economics* 119, 899–927.

Hall, R. and Jorgenson, D. 1967. Tax policy and investment behavior. *American Economic Review* 57, 391–414.

Hamermesh, D. 1989. Labor demand and the structure of adjustment costs. *American Economic Review* 79, 674–89.

Hamermesh, D. and Pfann, G. 1996. Adjustment costs in factor demand. *Journal of Economic Literature* 34, 1264–92.

Hassett, K. and Hubbard, R. 1997. Tax policy and investment. In *Fiscal Policy: Lessons from Economic Research*, ed. A.J. Auerbach. Cambridge, MA: MIT Press.

Hayashi, F. 1982. Tobin's marginal q and average q: a neoclassical interpretation. *Econometrica* 50, 213–24.

Hopenhayn, H. and Rogerson, R. 1993. Job turnover and policy evaluation: a general equilibrium analysis. *Journal of Political Economy* 101, 915–38.

Jorgensen, D. 1963. Capital theory and investment behavior. *American Economic Review* 53, 247–57.

Khan, A. and Thomas, J. 2003. Nonconvex factor adjustments in equilibrium business cycle models: do nonlinearities matter? *Journal of Monetary Economics* 50, 331–60.

Khan, A. and Thomas, J. 2006a. Inventories and the business cycle: an equilibrium analysis of (S,s) policies. Working paper.

Khan, A. and Thomas, J. 2006b. Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics. Working paper.

King, R. and Thomas, J. 2006. Partial adjustment without apology. *International Economic Review* 47, 779–809.

Koyck, L. 1954. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.

Lucas, R. 1967a. Adjustment costs and the theory of supply. *Journal of Political Economy* 75, 321–34.

Lucas, R. 1967b. Optimal investment policy and the flexible accelerator. *International Economic Review* 8, 78–85.

Lucas, R. and Prescott, E. 1971. Investment under uncertainty. *Econometrica* 39, 659–81.

Mendoza, E. 1991. Real business cycles in a small open economy. *American Economic Review* 81, 797–818.

Mosser, P.C. 1991. Trade inventories and (S,s). *Quarterly Journal of Economics* 106, 1267–86.

Pindyck, R. 1988. Irreversible investment, capacity choice, and the value of the firm. *American Economic Review* 78, 969–85.

Ramey, V.A. and West, K.D. 1999. Inventories. In *Handbook of Macroeconomics*, vol. 1B, ed. J. B. Taylor and M. Woodford. Amsterdam: North-Holland.

Rotemberg, J. 1987. The New Keynesian microfoundations. In *NBER Macroeconomics Annual 1987*, ed. S. Fischer. Cambridge, MA: MIT Press.

Sargent, T.J. 1978. Estimation of dynamic labor demand schedules under rational expectations. *Journal of Political Economy* 86, 1009–44.

Scarf, H. 1960. The optimality of (S,s) policies in the dynamic inventory problem. In *Mathematical Methods in the Social Sciences*, ed. K.J. Arrow, S. Karlin and H. Scarf. Stanford: Stanford University Press.

Scarf, H. 2005. Optimal inventory policies when sales are discretionary. *International Journal of Production Economics* 93–94 , Special Issue, 111–19.

Sheshinski, E. and Weiss, Y. 1977. Inflation and costs of price adjustment. *Review of Economic Studies* 54, 287–303.

Sheshinski, E. and Weiss, Y. 1983. Optimum pricing policy under stochastic inflation. *Review of Economic Studies* 50, 513–29.

Thomas, J. 2002. Is lumpy investment relevant for the business cycle? *Journal of Political Economy* 110, 508–34.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1, 15–29.

Treadway, A. 1971. The rational multivariate flexible accelerator. *Econometrica* 39, 845–55.

Veracierto, M. 2002. Plant-level irreversible investment and equilibrium business cycles. *American*

Economic Review 92, 181–97.

How to cite this article

Khan, Aubhik and Julia K. Thomas. "adjustment costs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000151> doi:10.1057/9780230226203.0010

The New Palgrave Dictionary of Economics Online

adverse selection

Charles Wilson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A market exhibits adverse selection when the inability of buyers to distinguish among products of different quality results in a bias towards the supply of low-quality products. Typically, the average quality of a product supplied by the market depends on the price, possibly resulting in multiple Walrasian equilibria and even equilibria with rationing. Agents have an incentive to trade multidimensional contracts so that informed agents can reveal their quality by the contracts they purchase. Various mechanisms such as price floors and mandatory partial insurance may be used to reduce the market inefficiencies resulting from adverse selection.

Keywords

adverse selection; Akerlof, G.; asymmetric information; Bertrand game; credit rationing; incentive compatibility; insurance markets; Pareto improvement; rationing; resale markets; reservation price; self-selection; separating equilibrium; signalling; Walrasian equilibria

Article

Adverse selection refers to a negative bias in the quality of goods or services offered for exchange when variations in the quality of individual goods can be observed by only one side of the market. For instance, suppose sellers of high-quality goods have a higher reservation price than sellers of low-quality goods, but that buyers cannot directly determine the quality of a specific good offered for sale. Then any mix of goods offered for sale at the market price must include the low-quality goods. That is, the market adversely selects for low-quality products.

Adverse selection may appear in any market where either the buyer or the seller has difficulties ascertaining the quality of the product to be exchanged. Examples include resale markets for durable goods where it is difficult for the buyer to identify defects known to the seller, labour markets where the seller has a better idea of his productivity than his potential employer, credit markets where the borrower knows more about her credit worthiness than the seller, and insurance markets where the insured have knowledge about their riskiness that is unavailable to the insurer.

The theoretical study of adverse selection began with the seminal paper by George Akerlof, ‘The Market for “Lemons”’ (1970). In this paper, Akerlof demonstrated how adverse selection could eliminate all trade in otherwise efficient markets. As the title suggests, he illustrated his argument in a stylized model of a market for used cars. Suppose there is a potential supply of n_s cars indexed by a quality parameter q that is uniformly distributed between 0 and 1. Assume that q measures the reservation price of the owner, but that the reservation value of each of the potential buyers is $\frac{3}{2}q$. If both buyers and sellers can observe the quality of each car and there are enough potential buyers, efficiency requires that all cars be exchanged. However, if buyers can observe only the average quality of cars offered for sale at each price, there is no positive price at which cars will be demanded.

The argument is as follows. If buyers cannot observe the quality of individual cars and prices adjust to clear the market, then all cars must sell at the same price p. Since an owner offers a car of quality q for sale only if $q < p$, it follows that the supply of cars is $S(p) = n_s p$ at any price p between 0 and 1 and the average quality of cars at that price is $q^a(p) = \frac{p}{2}$. But since a buyer's reservation value of a car with expected value q is $\frac{3}{2}q$, he purchases at price p only if $q^a(p) > \frac{2}{3}p$. Consequently, demand is $D(p) = 0$ at each price p and the only market clearing price is $p = 0$ with no trade occurring at all.

Akerlof's example of a zero-trade equilibrium illustrates the most extreme consequence of adverse selection. As demonstrated below, not all trade is necessarily eliminated. However, if goods of different quality are treated as a homogeneous good, several sources of inefficiency may persist. One problem is that the marginal value of a trade may not be equated between buyers and sellers. Since sellers offer any good for exchange that they value less than its price, the value to the sellers of the average product offered for sale is generally lower than the price. In contrast, the uninformed buyers purchase the product to the point where their value of the average car offered for sale equals the price so that their value of the marginal car offered by sellers exceeds the price.

A second source of inefficiency is that the wrong set of cars may be exchanged. In the example above, the net gain from trade of a car with quality q is $\frac{q}{2}$ so that the highest-quality cars should be exchanged first. However, if all cars are sold at the same price, lower-quality cars will always be supplied before higher-quality cars. In general, this inefficiency depends on our assumptions regarding preferences. In a dynamic model in which the market for used cars arises endogenously, Hendel and Lizzeri (1999) argue that buyers of used cars generally value increases in quality less than sellers. Consequently, in their model the sale of the lowest-quality cars is relatively efficient and measures to increase the volume of trade may be counterproductive.

A third source of inefficiency emerges when the preferences of buyers are heterogeneous so that high-quality cars should be allocated to quality-intensive buyers. In this case, even if the efficient set of goods were exchanged, the random allocation of cars among buyers implies that buyers and sellers would not be correctly matched.

All of these sources of inefficiency can be illustrated with a slight modification to Akerlof's example. Suppose we change the distribution of the n_s cars so that q is uniformly distributed between 1 and 5. Then, at any price p between 1 and 5, the supply of cars is $S(p) = \frac{p-1}{4} n_s$ and average quality is $q^a(p) = \frac{1+p}{2}$. At any price $p > 5$, $S(p) = n_s$ and $q^a(p) = 3$. On the demand side, we suppose there are two types of buyers. For a car of quality q, low-intensity buyers are willing to pay $\frac{3}{2}q$ and high-intensity buyers are willing to pay $2q$. Consequently, the demand function has two steps. Low-intensity buyers are just indifferent to buying a car at price $p = 3$ where $\frac{3}{2}q^a(p) = p$. For high-intensity buyers, the point of indifference is at $p = 6$. Consequently, if there are n_L low intensity buyers and n_H high intensity buyers, demand is

$$D(p) = \begin{cases} n_L + n_H & for \quad p < 3 \\ n_H & for \quad 3 < p < 6 \\ 0 & for \quad p > 6 \end{cases}$$

At prices 3 and 6, demand is a correspondence.

Figure 1 illustrates two possible relations between supply and demand depending on the relative number of buyers and sellers. The supply curve labelled $S'(p)$ corresponds to a case where $n_S < n_H$ so that the market clears at price $p = 6$. At this price, all cars are sold to high-intensity buyers, and the corresponding allocation is Pareto efficient. The supply curve labelled $S(p)$ corresponds to the case where $n_H < \frac{n_S}{2} < n_L + n_H$ so that the market clears at price $p = 3$. At this price, only cars of quality $q < 3$ are sold and every active buyer receives a car of expected quality $q^a(p) = 2$.

Figure 1

An inefficient Walrasian allocation

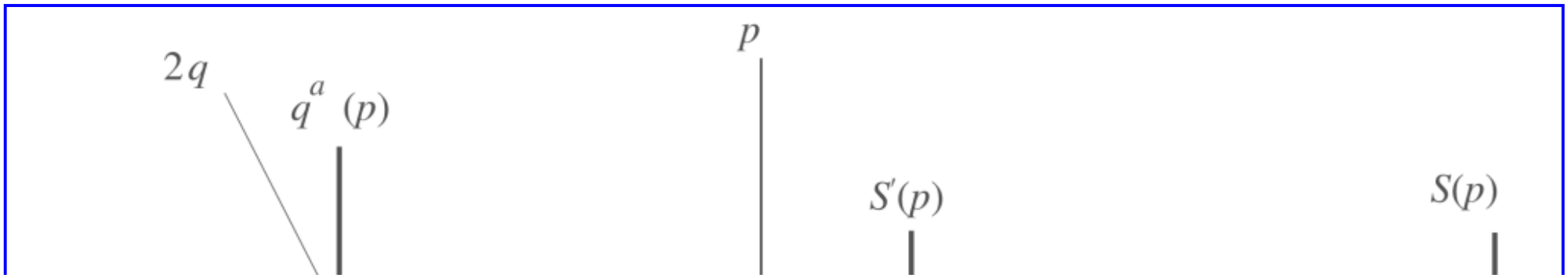

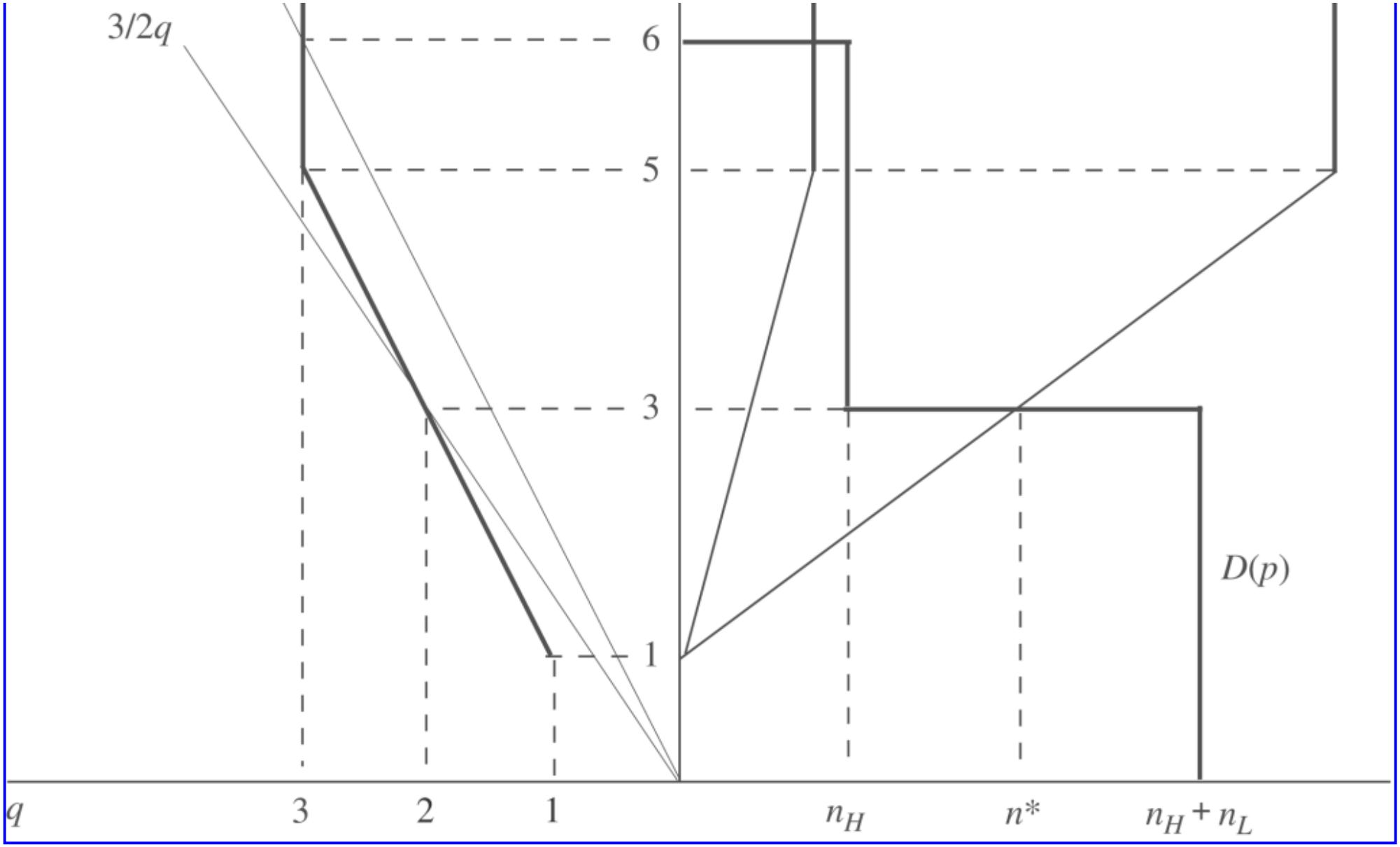

Observe that this allocation exhibits all of the sources of inefficiency that were identified above. First, not all potential buyers purchase a car even though half of the cars remain unsold, all of which are more valuable to buyers than to owners. Second, the cars that are sold provide the least possible net benefit to buyers. If only half of the cars are to be sold, efficiency requires they be the highest-quality cars. Third, since all buyers purchase from the same pool of cars, the cars that are sold are not efficiently allocated among buyers. Since high-intensity buyers value quality more than low-intensity buyers, the efficient allocation of these cars requires that the high-intensity buyers receive the cars with the highest quality. Given the asymmetry in information, there is typically no incentive-compatible mechanism that achieves first-best efficiency. However, there may be instruments or mechanisms that may increase net surplus and in some cases even generate a Pareto improvement. For instance, for supply curve $S(p)$ a subsidy on sales would increase the volume of trade. However, the resulting allocation would not be completely efficient since low-quality cars are still sold before high-quality cars and both types of buyers still purchase from the same pool of cars. We explore below some other mechanisms that may be used to further improve efficiency.

Multiple Walrasian equilibria

The examples above have a unique Walrasian equilibrium. However, since average quality increases with price, it is possible that over some range of prices demand may also increase with price. As a consequence, there may be multiple market clearing prices, which can be Pareto ranked. We can illustrate this possibility in an example with one type of buyer and just two types of sellers.

Suppose half of the n_s sellers own cars of quality $q = 1$ and half own a car of quality $q = 2$. Since low-quality sellers supply cars at any price p at or above $p = 1$, and high-quality sellers supply cars at any price p at or above $p = 2$, it follows that average quality jumps from 1 to $\frac{3}{2}$ at price $p = 2$. As above, suppose that each of the n_B buyers is willing to pay $\frac{3}{2}q$ for a car of quality q. Then $D(p) = n_B$ for $p < \frac{3}{2}$, but then falls to zero until the high-quality sellers enter the market at price $p = 2$. At this price, $q^a(p)$ rises to $\frac{3}{2}$ and all buyers again enter the market until p rises to $\frac{9}{4}$, after which price exceeds the buyers' reservation value and $D(p)$ falls back to zero. The result is a non-monotonic demand function and consequently it is possible that there is more than one market clearing price.

In this example, multiple Walrasian equilibria arise whenever the number of buyers exceeds the total number of cars. Such a case is illustrated in Figure 2, where demand $D(p)$, indicated by the heavy dotted line, intersects $S(p)$ at prices $\frac{3}{2}$, 2, and $\frac{9}{4}$. All cars are sold at price $p = \frac{9}{4}$, while only low-quality cars are sold at price $p = \frac{3}{2}$. In both cases, $p = \frac{3}{2}q^a(p)$ so that buyers are just indifferent to purchasing a car. There is also a Walrasian equilibrium at price $p = 2$, but to clear the market only half of the owners of high-quality cars supply their cars. As a result, average quality is reduced to $\frac{4}{3}$ so that buyers are again just indifferent to purchasing at that price.

Figure 2

Multiple Walrasian equilibria

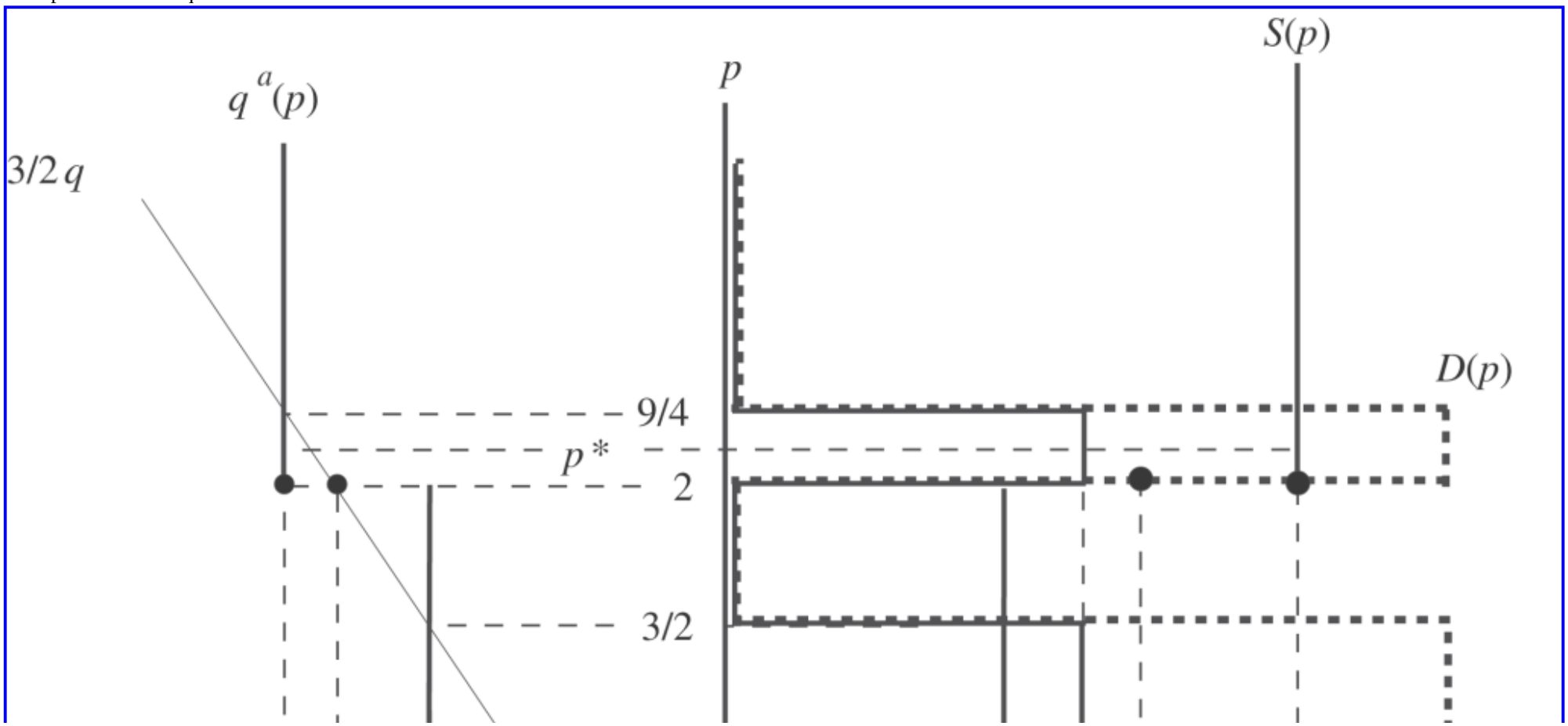

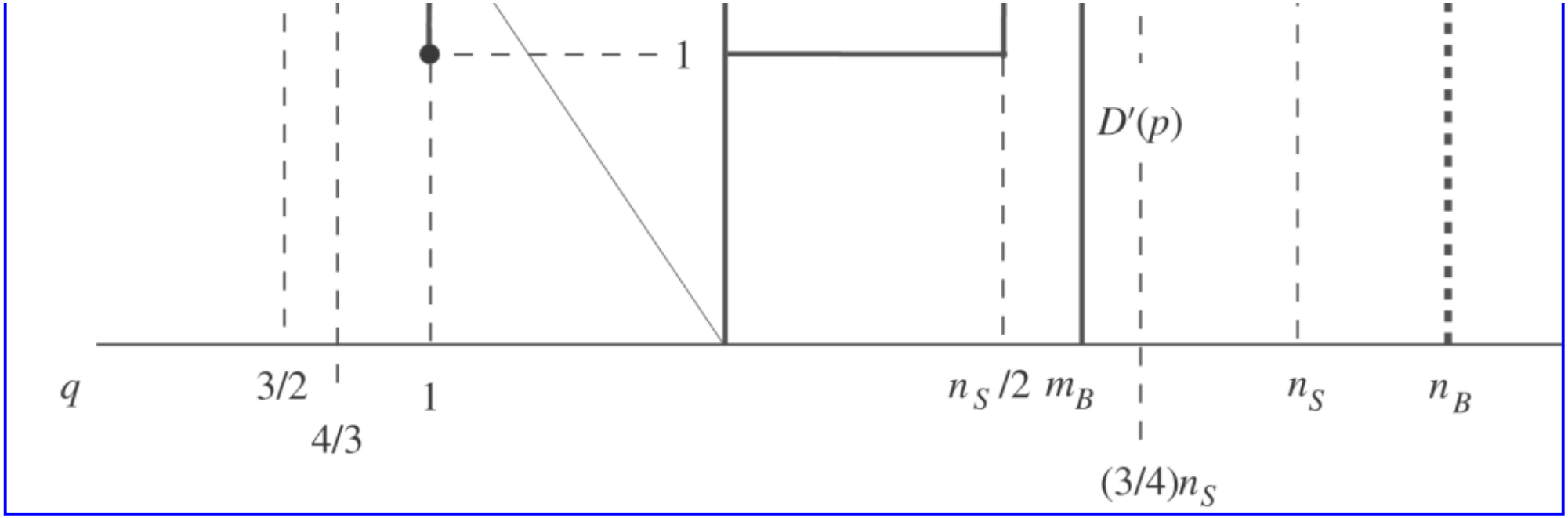

Observe that the allocations at these three prices may be Pareto ranked. Although buyers are indifferent to each of the prices, some or all sellers strictly benefit from selling at a higher price. In a more general model with heterogeneous buyers, Wilson (1980) shows that buyers also benefit from buying at a higher price.

Pareto improving price floors

Because of the dependence of average quality on price, it is sometimes possible to achieve an additional Pareto improvement by setting a price floor and rationing the excess supply of cars. Consider again the example illustrated in Figure 2. If we reduce the number of buyers to m_B where $\frac{n_S}{2} < m_B < n_S$, then we obtain a demand curve like $D'(p)$, illustrated by a heavy solid line. In this case, there is only one Walrasian equilibrium at price $p = \frac{3}{2}$. At this price, only low-quality cars are offered for sale and buyers gain no net benefit.
Now suppose that we impose a floor ceiling at some price p^* between 2 and $\frac{9}{4}$. Since high-quality cars are also supplied at this price, average quality rises to $q^a(p^*) = \frac{3}{2}$ which provides any buyer with a positive net benefit. Since there are more sellers than buyers at this price, sales must be rationed. Nevertheless, owners of both low-quality and high-quality cars benefit from selling at this price. Owners of high-quality cars benefit because the Walrasian price is below their reservation value. And since more than half of the cars are sold at this price, the expected return to low-quality sellers is also higher at price p^*. At the Walrasian price $p = \frac{3}{2}$, their net benefit from a sale is $\frac{1}{2}$, while at the price floor $p^* > 2$, their net benefit from a sale is at least 1.

Uninformed price setters and rationing

Our analysis so far has focused on primarily on Walrasian allocations. In a frictionless economy with perfect information and a large number of competing agents, this solution is generally robustly independent of the mechanism or conventions by which the price is set. However, once we introduce asymmetries in information, the opportunity for market participants to exploit the relation between quality and price or to indirectly identify products of different quality may lead to different market behaviour. To study these effects, we need to be more explicit in specifying the mechanism by which trade takes place.
Consider a market mechanism in which each buyer fixes a price at which he is willing to buy. To sell their cars, sellers first queue at the highest announced price. Any excess supply then spills over to successively lower-price offers until the supply of cars is exhausted or there are no more offers to buy. Buyers who announce a price below the point at which supply is exhausted do not obtain a car.

Suppose that all buyers value a car of quality q at $\frac{3}{2}q$. Then, without regard to market conditions, each buyer prefers the price p that maximizes his or her net benefit $\frac{3}{2}q^a(p) - p$. However, such a price p is an equilibrium only if there is no excess demand at that price. As in a standard Bertrand game, rather than face rationing, buyers prefer a small increase in the price so that they can buy a car with certainty. Consequently, the equilibrium strategy for buyers is to set the price that maximizes net benefit $\frac{3}{2}q - p$ subject to the constraint $D(p) \leq S(p)$.

Figure 2 illustrate two types of solution to this problem. For the case where the number of buyers is $n_B > n_S$, represented by the heavy dotted demand curve $D(p)$, the equilibrium price is $p = \frac{9}{4}$, which is the highest Walrasian price. At this price, all cars are sold to buyers who are just indifferent to purchasing a car. For the case where the number of buyers m_B satisfies $\frac{n_S}{2} < m_B < n_S$, the equilibrium price is $p = 2$ (or slightly above to ensure that all owners supply their cars). All buyers demand a car and all owners supply a car. But since there are more sellers than buyers, the sellers must be rationed. With heterogeneous buyers, Wilson (1980) shows that more than one price may be announced in equilibrium. In this case, sellers are rationed at all but possibly the lowest announced price.

A mechanism in which uninformed agents set the price may not be applicable for most resale markets for durable goods. However, it may explain some pricing strategies in financial markets where the uninformed agents are large institutions such as banks. Stiglitz and Weiss (1981) implicitly use this price-setting mechanism in their study of credit rationing. In their model, banks supplying loans correspond to the uninformed buyers of the used car market, and the creditors, who know better their idiosyncratic riskiness, correspond to the car owners. Because creditors have only limited liability in the case of default, risky borrowers demand loans at higher interest rates than do less risky borrowers. So, if the demand for loans is sufficiently large, only risky borrowers are served at the Walrasian rate of interest. In such a case, it may be more profitable for banks to lower their interest rate to attract low-risk borrowers, even though they must ration their limited supply of funds among the resulting increased demand.

Informed price setters

In markets for products such as used cars, a mechanism in which sellers are responsible for setting the price may be of more interest. For example, consider the price-setting convention in which all sellers simultaneously announce prices for their cars, after which each buyer submits a bid at one of these prices. If demand does not equal supply at any price, the long side of the market is rationed. Since the informed agents act first, this mechanism is essentially a signalling game, first introduced by Spence (1973) and later formalized by Cho and Kreps (1987) and others.

Consider again the example above with two types of sellers, half with cars of quality $q = 1$ and half with cars of quality $q = 2$, and one type of buyer who is willing to pay $\frac{3}{2}q$ for a car of quality q. Assume also that there are more potential buyers than sellers. As in many signalling models, there is a continuum of sequential equilibria for this game. We focus here on two possible outcomes. One possibility is a pooling equilibrium in which each seller announces price $p = \frac{9}{4}$, and exactly n_S buyers bid to purchase at that price, resulting in a Walrasian allocation. Buyer behaviour is optimal since each buyer is indifferent between buying and not buying, and seller behaviour is optimal if buyers believe that average quality will not increase at higher prices.

A second possibility is a separating equilibrium that involves rationing at some prices. In this equilibrium, low-quality sellers announce price $p_L = \frac{3}{2}$ and high-quality sellers announce price $p_H = 3$. Exactly $\frac{n_S}{2}$ buyers bid at price p_L, so that demand exactly matches supply and low-quality sellers sell with probability 1. However, at price p_H, only $\frac{n_S}{8}$ (or fewer) buyers bid so that high-quality sellers sell with probability at most $\frac{1}{4}$. Observe that at each price buyers are just indifferent between purchasing and not purchasing. Each seller is also acting optimally, since high-quality sellers would suffer a loss by selling at p_L, while low-quality sellers prefer to earn a net gain of $\frac{1}{2}$ with certainty at price p_L rather than a net gain of 2 with probability less than or equal to $\frac{1}{4}$ at price p_H. A general analysis with heterogeneous buyers is provided in Wilson (1980).

It is not obvious how expectations and prices would adjust to sustain the separating equilibrium in this example. However, the example does illustrate how market participants may use another dimension, in this case the probability of selling, to identify products of different quality, albeit at some cost. The key ingredient is that sellers of different-quality cars face a different tradeoff between price and the probability of selling. In general, there may be other dimensions in which the preferences of informed agents differ. In such a case the market may exploit multidimensional contracts to identify product quality. A market for insurance provides a good example.

Self-selection in insurance markets

In its most primitive form, an insurance policy consists of two elements, the price of coverage and the level of coverage. Although all consumers prefer a lower price to a higher price and prefer more coverage to less coverage, their tradeoff between price and quantity depends on the probability of a payout. Consequently, by offering contracts which differ in both price and the level of indemnity, sellers may be able to indirectly identify different risk classes of consumers who otherwise appear to be a homogeneous population. Some of the implications of competition in these kinds of contract can be illustrated in a simple model first studied by Rothschild and Stiglitz (1996) and Wilson (1977).

Suppose there are two types of insurance consumers. Each consumer has the same risk-averse von Neumann–Morgenstern utility u, the same initial wealth W and the same reduction in wealth to $W - 1$ in case of an accident. Low-risk types have an accident with probability π_L and high-risk types have an accident with probability π_H, where $\pi_L < \pi_H$. An insurance policy may be represented as pair $(p,\bullet t)$, where t is the indemnity in case of an accident and p is the premium. Therefore, a consumer who purchases policy $(p,\bullet t)$ is left with wealth $W - 1 - p + t$ if he has an accident and $W - p$ if he does not. Suppose that each individual can identify his own risk type but that firms know only the proportion α of low-risk types. Let $\pi^a = \alpha\pi_L + (1 - \alpha)\pi_H$ denote average probability of an accident among both types of consumers. To allocate the policies, we suppose that the uninformed firms are Bertrand price setters that earn zero profit for any policy that is actuarially fair.

The model is illustrated in Figure 3, where the vertical axis represents the premium and the horizontal axis represents the level of coverage. The vertical line at $t = 1$ represents the set of policies that provide full indemnity. The lines labelled π_L and π_H represent the set of actuarially fair policies for the low- and high-risk types respectively. The line labelled π^a represents the set of policies that break even if both types purchase it. The curves labelled v_L and v_H represent typical indifference curves for the two risk types. Although both risk types prefer more coverage and a smaller premium, high-risk types have a higher marginal rate of substitution (MRS) of premium for indemnity than do low-risk types at any policy. At any full insurance policy, the MRS of each type is equal to their probability of an accident.

Figure 3

Equilibrium in an insurance market

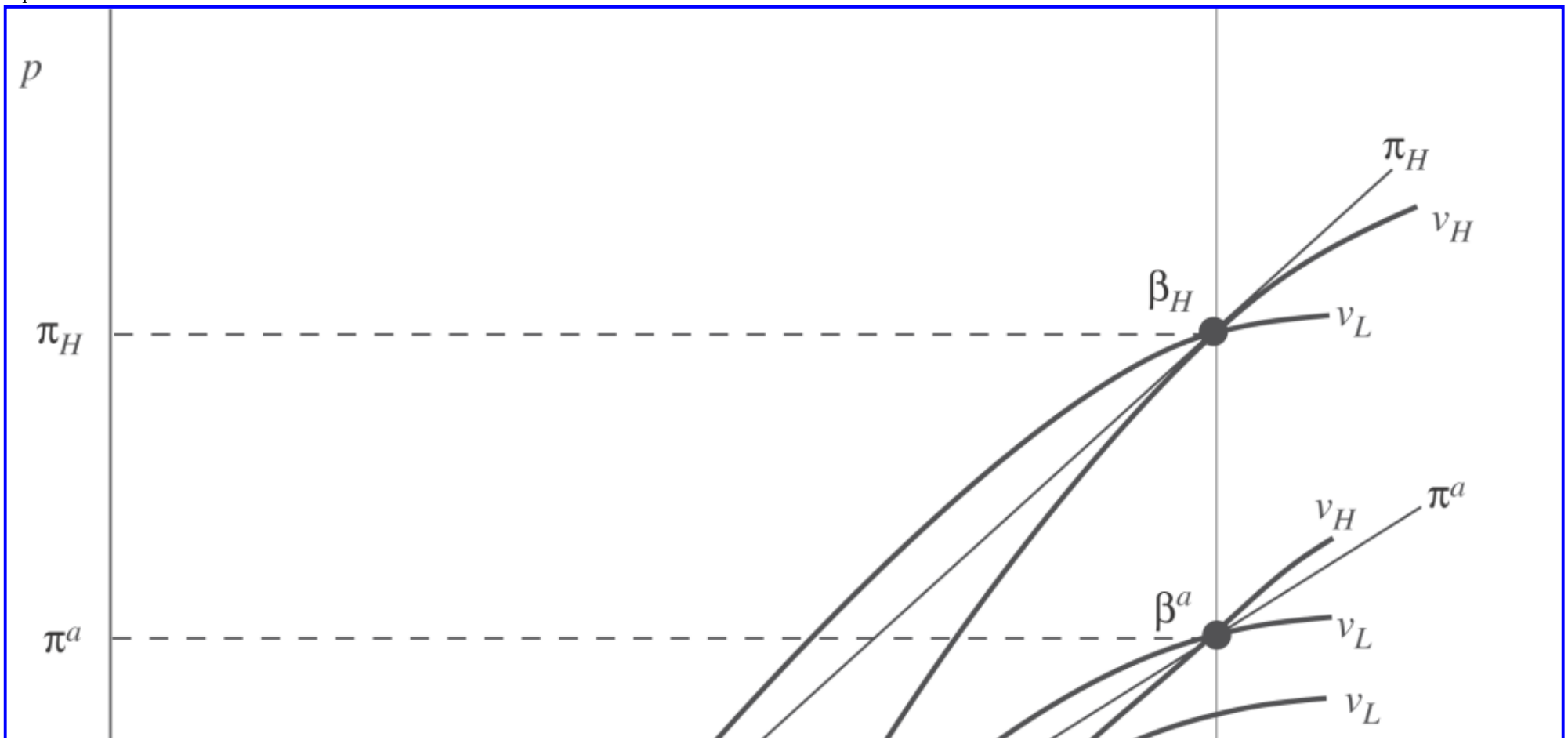

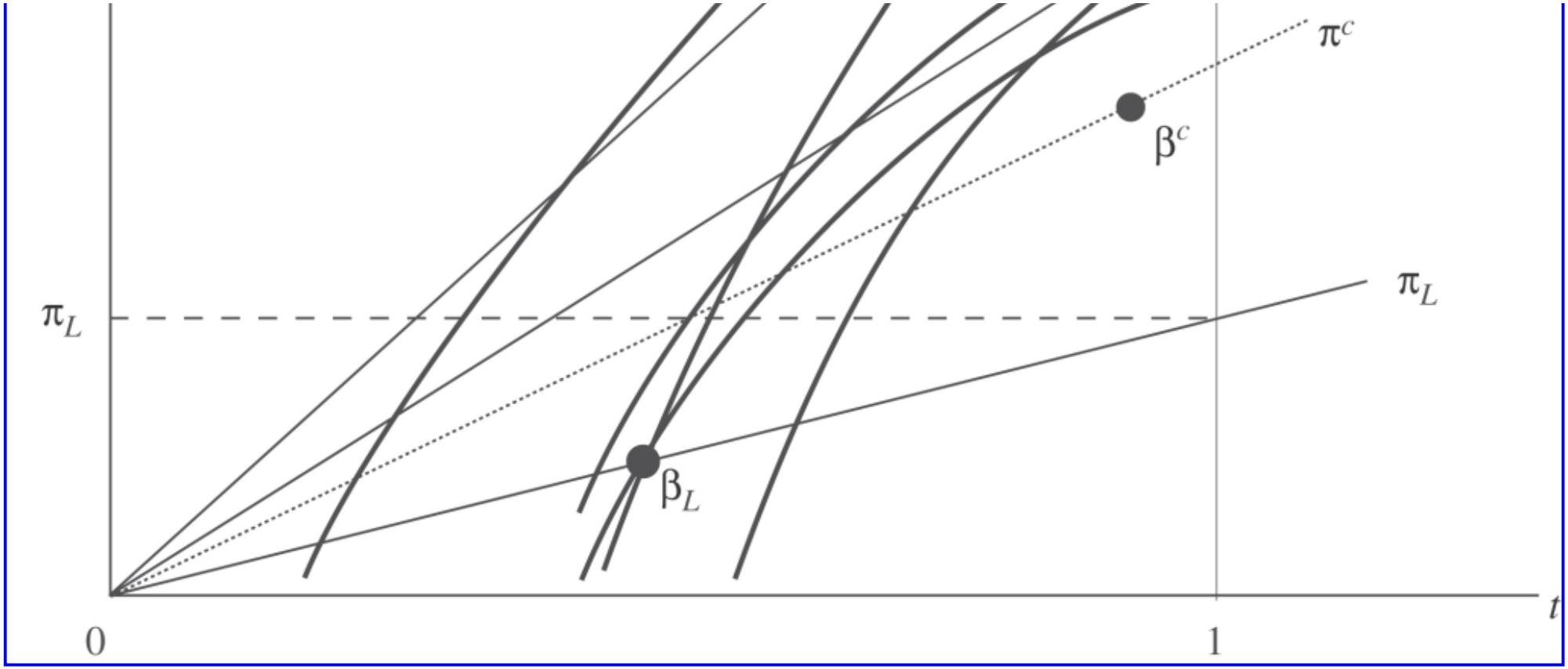

Suppose first that firms may offer only policies that provide full coverage so that $t = 1$. In this case, the model is exactly analogous to the used-car example above when the uninformed buyers are price setters and there are more buyers than sellers. Consumers demand insurance policy (p,1) only if their expected utility from purchasing exceeds their expected utility from remaining uninsured. The policy $\beta_H = (\pi_H, 1)$ represents the full insurance policy that just breaks even for the high-risk types. For the case illustrated here, the low-risk types would also demand insurance at this price. Consequently, the unique Bertrand equilibrium is the policy $\beta^a = (\pi^a, 1)$, which just breaks even when purchased by both risk types. In effect, low-risk types are subsidizing the high-risk types.

Now suppose that firms may also compete in the indemnity dimension. To begin, we also suppose that each firm may offer only one insurance policy to its customers. Observe that the equilibrium policy under mandatory full coverage is not an equilibrium for this game. The reason is that, if some firm deviates and offers a policy near β_L, above the π_L line and behind the v_H curve, it attracts only low-risk types and earns a positive profit. But if low-risk types are attracted away from policy β^a, it earns negative profits.

The only possible equilibrium is a separating allocation in which some firms offer policy β_H, which is purchased by high-risk types, and some firms offer policy β_L, which is purchased by the low-risk types. Equilibrium requires that the policy purchased by each risk type lie on its own zero profit line. Otherwise, firms may exploit the differences in the preferences of the two risk types to offer a policy that attracts only the risk class that earns positive profits. Competition among firms must then lead to the best zero-profit policy for the high-risk types and the best zero-profit policy for the low-risk types, subject to the self-selection constraint for high-risk types to choose policy β_H.

If the aggregate zero profit line π^a lies above the low-risk indifference curve that passes through the low-risk policy β_L, as illustrated in Figure 3, then equilibrium exists. Both policies lie on their respective zero-profit lines and each consumer selects his optimal policy from the available set. If any firm deviates with a new policy offer that attracts only the high-risk types, it must lie below the π_H line and consequently earn negative profits. However, any new policy that attracts the low-risk types cannot earn positive profits unless it also attracts the high-risk types. But any such policy earns positive profit only if it lies above the π^a line, which in turn attracts only the high-risk types.

If the aggregate zero-profit line intersects the low-risk indifference curve passing through π_L, as illustrated by the dotted line labelled π^c in Figure 3, then there is no equilibrium for this game. In this case, a firm may offer a policy just above β^c that attracts both types of consumers and still makes positive profits in the aggregate. If we permit individual firms to offer a menu of contracts as in Miyazaki (1977), then equilibrium fails to exist under an even wider range of parameters. A number of authors have suggested alternative solution

concepts, incorporating non-Nash behaviour, that generate an equilibrium for this case. Wilson (1977) defines a solution concept in which both types purchase a policy like β^c. Riley (1975) proposes an alternative solution concept for which the separating allocation β_L and β_H is an equilibrium.

Efficient public provision of insurance

Consider the case where (β_L, β_H) is an equilibrium. The low-risk types are made better off than under the equilibrium with mandatory full coverage by lowering their indemnity to segregate themselves from the high-risk types. But high-risk types are worse off since they must now pay the actuarially fair value of their coverage. Clearly, this allocation is not first-best efficient since an increase in the coverage of the low-risk types at an actuarially fair rate makes them better off. Consequently, it may be possible to increase the welfare of both types by introducing a menu of policies in which the low-risk types subsidize the high-risk types. Such an allocation is represented by policies γ_L and γ_H as illustrated in Figure 4.

Figure 4

The public provision of insurance

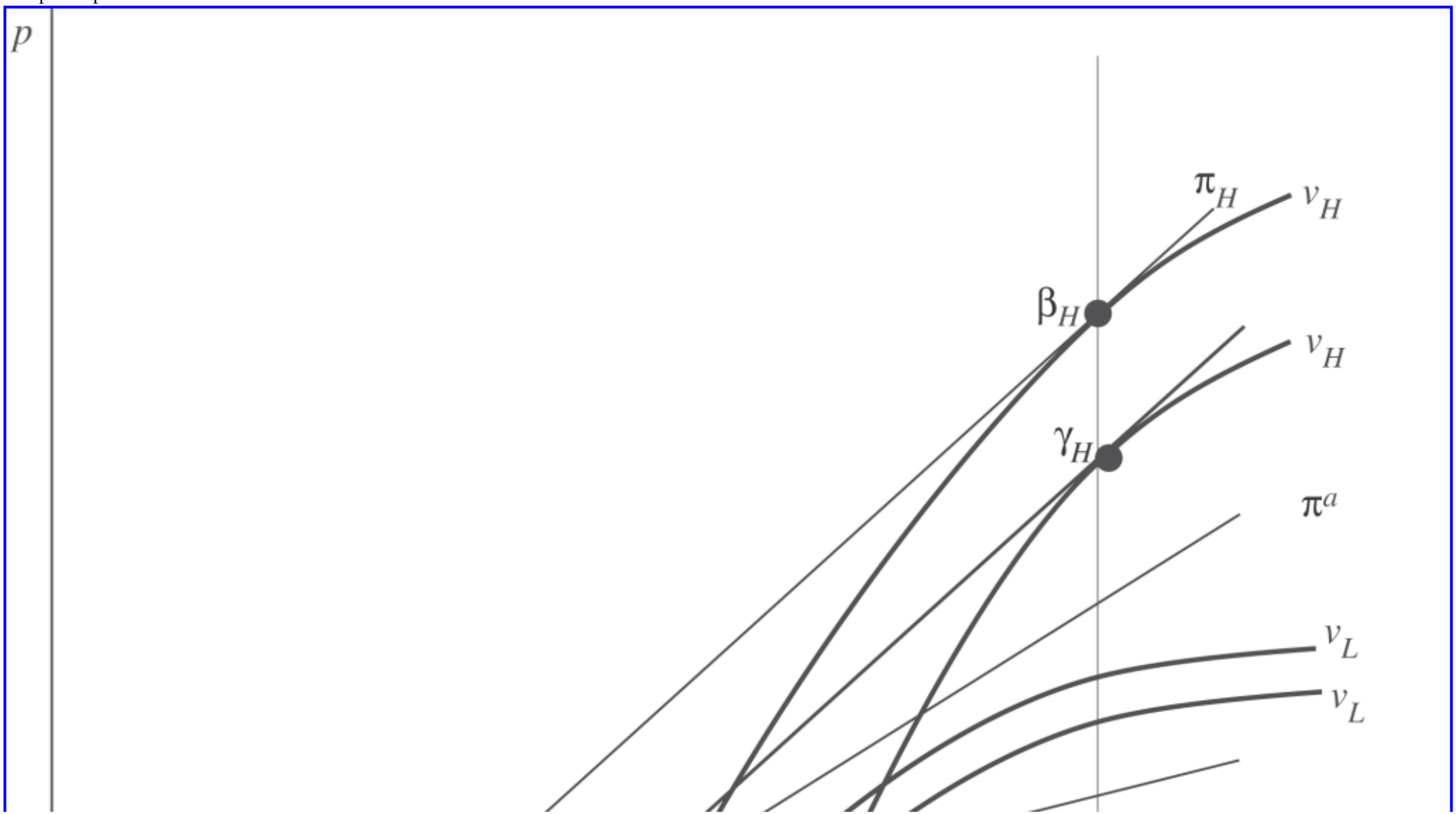

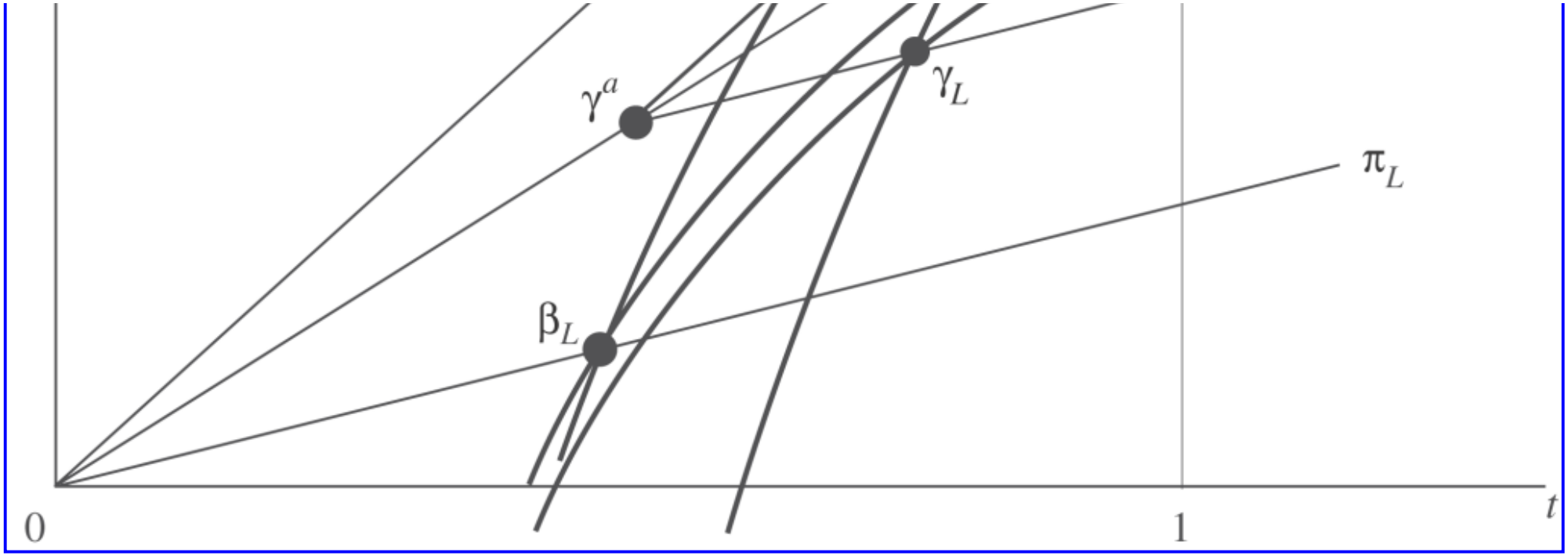

To see that the policies are actuarially fair in the aggregate, observe that they can be constructed by decomposing each policy into a common policy γ^a that lies on the aggregate zero-profit line and then supplementing the coverage of each risk type with an additional policy that lies on their respective isoprofit line that passes through policy γ^a. One way to implement such an allocation is for the government to provide policy γ^a to all consumers and then let the market supply the supplementary policies. Furthermore, by choosing the appropriate policy γ^a, this mechanism may be used to attain any constrained Pareto-optimal allocation (subject to the self-selection constraints and aggregate zero-profit condition). In this case, the supplementary pair of policies required to attain allocation (γ_L, γ_H) is necessarily an equilibrium so there is no need to appeal to alternative solution concepts to ensure the existence of an equilibrium.

See Also

- credit rationing
- implicit contracts
- incomplete contracts
- moral hazard
- selection bias and self-selection
- signalling and screening

Bibliography

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Cho, I. and Kreps, D. 1987. Signalling games and stable equilibria. *Quarterly Journal of Economics* 102, 179–221.

Hendel, I. and Lizzeri, A. 1999. Adverse selection in durable goods markets. *American Economic Review* 89, 1097–115.

Miyazaki, H. 1977. The rat race and internal labor markets. *Bell Journal of Economics* 8, 394–418.

Riley, J.G. 1975. Competitive Signalling. *Journal of Economic Theory* 10, 175–86.

Rothschild, M. and Stiglitz, J. 1996. Equilibrium in competitive insurance markets: an essay on the economics imperfect information. *Quarterly Journal of Economics* 90, 629–49.

Spence, M. 1973. Job market signalling. *Quarterly Journal of Economics* 87, 355–74.

Stiglitz, J. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

Wilson, C. 1977. A model of insurance with incomplete information. *Journal of Economic Theory* 16, 167–207.

Wilson, C. 1980. The nature of equilibrium in markets with adverse selection. *Bell Journal of Economics* 11, 108–30.

How to cite this article

Wilson, Charles. "adverse selection." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000040>
doi:10.1057/9780230226203.0011

The New Palgrave Dictionary of Economics Online

advertising

Original 1987 article by Richard Schmalensee, revised by David Genesove
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Empirical studies suggest that advertising is not an important determinant of consumer behaviour and that advertising follows rather than leads cultural trends. On the core issue of whether advertising is anti- or pro-competitive, the evidence suggests that advertising is associated with lower prices.

Keywords

advertising; concentration; market structure; entry

Article

Advertising has been controversial, probably more so that its economic importance would justify, at least since the emergence of the mass media in the 19th century. In the United States, advertising spending in the second half of the 20th century was just above two per cent of GDP. This ratio grew slowly over time; it is much lower in most other countries, especially in developing nations. In the United States and elsewhere, the ratio of advertising to sales varies dramatically among industries, even if attention is limited to industries selling consumer goods and services.
Chamberlin's Theory of Monopolistic Competition (1933) was the first major work in economics to treat advertising formally, but its analysis led to few definite positive or normative conclusions. Perhaps reflecting the traditional distaste for advertising in the intellectual community, most early discussions of advertising by economists were generally critical, describing it as wasteful, manipulative, and anti-competitive. Its main redeeming feature was that it provided a source of revenue for the press (Kaldor, 1950, is a leading example). Most writers are less enthusiastic about the relation between advertising and the media, perhaps because of the rise of television.

Consumer demand

We still know relatively little about how advertising affects consumer behaviour. Some writers distinguish between informative and persuasive advertising. Buyers are assumed to respond rationally to informative advertisements, while persuasive advertisements are somehow manipulative. But this distinction is of little value empirically: few if any advertisements present facts in a neutral fashion with no attempt to persuade, and even those with no obvious factual content signal to consumers that the seller has invested money to get their attention.

Following Nelson (1974), a number of authors have explored the possibility that advertising affects behaviour through such signals. The core of the argument is that advertising is more profitable for high-quality than low-quality producers, all else equal, since the former are more likely to enjoy repeat sales. In sharp contrast, information processing models of human behaviour, explored in the marketing literature, suggest that advertising may affect behaviour mainly by enhancing a brand's chances of being on the short list (‘evoked set’) from which final choices are made.

It seems likely that the role of advertising varies considerably, depending on the characteristics of products and distribution systems. In some markets advertised brands sell for substantially more than physically identical unadvertised brands; in others, restrictions on advertising serve to increase prices (Benham, 1972). Porter (1976) has argued that advertising is less powerful when retailers are an important source of consumer information. The extent to which a buyer can judge quality prior to purchase (Nelson, 1974) should also affect the role of advertising. Similarly, buyers need more information to make decisions about new products than about established products, and advertising by retailers generally provides more price information than advertising by manufacturers.

Econometric analysis of the effects of advertising on consumer spending patterns is difficult because advertising is endogenous; it reflects sellers’ decisions. This gives rise to simultaneous equations problems (Schmalensee, 1972). Survey evidence suggests that firms often follow percentage-of-sales decision rules in determining advertising budgets. If this were strictly true, the effect of advertising on sales would be impossible to identify. In fact, advertising–sales ratios are not constant over time, but it is difficult to find seller-related variables that explain the variations well. To the extent that advertising spending is based to some extent on actual or anticipated sales, but demand equations are estimated via least squares because the advertising spending decision cannot be modelled adequately, the importance of advertising as a determinant of consumer behaviour will be overstated.

Borden's (1942) massive study of the effects of advertising on demand concluded that advertising is not generally an important determinant of industry sales. Exceptions arise in new and growing sectors, where advertising can serve to accelerate growth that would occur in any case. Recent work seems generally to support these conclusions (see, for instance, Lambin, 1976). At the aggregate level, advertising tends to lag cyclical changes in total consumption slightly, not to lead those changes (Schmalensee, 1972, ch. 3). At the other extreme, while advertising is generally found to affect market shares, dollar advertising spending typically explains little of the variation in shares over time. This presumably reflects in part the fact that designing effective advertising themes and campaigns remains much more an art than a science.

Overall, advertising does not emerge from the empirical literature on consumer demand as an important determinant of consumer behaviour. Some have argued that advertising has fostered the long-run growth of materialism, but nobody has offered anything like a rigorous test of this proposition. Most practitioners contend that advertising follows rather than leads cultural trends, in part because most

advertisers are reluctant to appear out of step with society.

Seller behaviour

All else equal, one would expect sellers to spend more on advertising in markets in which demand is more responsive to advertising, and one might expect demand to be more responsive when consumers need more information to make rational decisions (see Schmalensee, 1972, ch. 2). But we observe very intensive advertising, without much obvious factual content, of some products with which consumers are generally familiar, such as beer and soft drinks.

To the extent that advertising's effects persist over time, advertising outlays are an investment, and advertising budgets must be set using dynamic optimization methods (Sethi, 1977). The greater the profit on additional sales (that is, the greater the gap between price and marginal cost), the more intensively it pays to advertise. Finally, advertising decisions by oligopolists must take into account the strategies of their rivals.

Consideration of these last two points indicates that the intensity of advertising may rise or fall with increases in market concentration (Schmalensee, 1972, ch. 2). On the one hand, reductions in the number of sellers would be expected to reduce the intensity of all forms of rivalry, and thus to reduce advertising spending. On the other hand, if sellers in concentrated markets manage to raise prices far above marginal costs, they thereby enhance incentives to advertise.

Advertising competition can serve to erode excess profits. With a fixed number of sellers, it is likely to be more effective at doing so the more sensitive market shares are to differences in advertising outlays. Greater sensitivity encourages all sellers to advertise more without necessarily increasing the size of the market for which they are competing.

The evidence on scale economies in advertising is mixed. On the one hand, there is little or no evidence that doubling the number of advertisements seen by buyers will more than double the impact on demand. On the other hand, some media offer bulk discounts. And some media, particularly network television in the United States, are such that the minimum required outlay is large in absolute terms. This may serve to disadvantage small sellers by effectively denying them the use of these media.

Economic welfare

One must distinguish between global and local welfare analysis in this context. Global analysis is concerned with questions like ‘could one ban advertising (everywhere or in some particular market) and make society better off?’ Local analysis deals with questions like ‘would society be made better off by a reduction in the level of advertising spending (everywhere or in some particular market)?’

Global questions are difficult to treat formally and thus have not been answered rigorously. Since advertising provides some information, one must specify how information would be provided if advertising were banned. In principle an omniscient bureaucrat can provide information to perfectly rational consumers optimally, so that a properly administered advertising ban can do no harm.

In practice, bureaucrats are far from omniscient, and the way in which information is presented to consumers affects the extent to which they retain and use it. Advertisers have every incentive to present information effectively, though they rarely have any incentive to present all information that might

affect decisions. Advertising, like democracy, is terrible in principle but better than any known alternative in practice. Note also that advertising is practised, though not intensively by US standards, in socialist economies.

Local questions about the optimality of advertising are more susceptible of formal treatment. There are as many answers to these questions as there are papers that address them, however. The answers depend critically on exactly how advertising is assumed to affect behaviour. Butters (1977), for instance, assumes that advertising simply provides price information. He concludes that market-determined advertising levels are optimal if buyers cannot engage in search but are excessive if search is possible. Dixit and Norman (1978) assume that advertising simply changes tastes. If pre-advertising tastes are assumed to be socially ‘correct’, a value-laden assumption, they show that advertising is generally socially excessive.

In general the literature offers no support for a presumption that market-determined advertising levels are socially optimal. But it also fails to provide any workable scheme for regulating those levels in the public interest.

Market structure

Discussions of the effects of advertising spending on the evolution of market structure have been dominated by two extreme views. Advertising's critics (for example, Kaldor, 1950) stress its persuasive nature, argue that it builds loyalties and thus reduces price elasticities of demand within markets, and contend that it is a source of barriers to entry. Beginning with Telser (1964), advertising's defenders stress its role as a source of information, argue that it provides knowledge of alternatives and thus increases elasticities, and contend that it is a means of effecting, not deterring, entry. Since the role of advertising seems to vary considerably among markets, neither of these extreme views is likely to be universally correct.

As a theoretical matter, the impact of advertising spending on price elasticities and barriers to entry depends, once again, on exactly how advertising is assumed to affect consumer behaviour. A good deal of empirical work has attempted to choose between the two extreme views outlined above, without producing any definitive results (see Camanor and Wilson, 1979, for a survey).

Many studies have examined the cross-section correlation between advertising and seller concentration; none has provided a satisfactory interpretation of this statistic. Telser (1964) found market shares to be less stable in markets with heavy advertising than in other markets, and Lambin (1976) found price elasticities to be lower in such markets. But neither study controlled for the product characteristics that affect share stability, price elasticity, and sellers’ advertising spending decisions.

The clearest empirical regularity to emerge from this work is the strong, positive cross-section correlation between industry-level measures of advertising intensity (typically the advertising–sales ratio) and accounting measures of profitability. This stylized fact would seem to favour advertising's critics.

But profits are high when price–cost margins are large, and large margins encourage advertising (Schmalensee, 1972, ch. 7). Since it is difficult to model advertising spending decisions empirically, it is difficult to deal adequately with this simultaneous equations problem. Moreover, accounting measures of profit treat advertising as an expense, but it should be treated as a durable investment if its effects on

demand persist over time. If those effects are assumed to be very long-lived, correcting the accounting profitability figures eliminates the correlation with advertising. Unfortunately, like so much in this area, the longevity of the impact of advertising on demand remains controversial.

New empirical developments

The core empirical question in the economics of advertising is whether its presence is anti- or pro-competitive. Beginning with Benham (1972), a number of studies have compared prices across US states that do and do not prohibit advertising (for example, Cady, 1976; Kwoka, 1984). Because of the concern that advertising prohibitions may be the result of concerted effort among firms, the effectiveness of which may be correlated with their ability to collude, other studies have considered changes in advertising regimes over time. Thus Glazer (1981) exploits a newspaper strike in New York City, which impeded advertising by supermarkets (but not small grocery stores, which do not generally advertise) in most but not all of the city, while Milyo and Waldfogel (1999) trace the pattern of prices in Rhode Island and neighbouring Massachusetts around the time the US Supreme Court struck down a law prohibiting liquor store advertising in Rhode Island. Devine and Marion (1979) published supermarket prices in Ottawa during a five-week period, and compared prices during that period to prices before and after and in Winnipeg. In none of these studies, whether cross-section or event study, are prices higher in the advertising regime. Typically they are lower, and, typically within the advertising regime, prices of advertised products are lower than those not advertised. A different approach is taken in Ackerberg (2001), where it is shown that only consumers who have not previously purchased a newly introduced yogurt are affected by advertising, and from which the author concludes that advertising in this instance is informative.

See Also

- Chamberlin, Edward Hastings
- market structure
- monopolistic competition

Bibliography

Benham, L. 1972. The effects of advertising on the price of eye-glasses. *Journal of Law and Economics* 15, 337–52.

Borden, N.H. 1942. *The Economic Effects of Advertising*. Chicago: Irwin.

Butters, G. 1977. Equilibrium distribution of sales and advertising prices. *Review of Economic Studies* 44, 465–91.

Chamberlin, E.H. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Camanor, W.S. and Wilson, T.A. 1979. The effect of advertising on competition: a survey. *Journal of Economic Literature* 17, 453–76.

Dixit, A. and Norman, V. 1978. Advertising and welfare. *Bell Journal of Economics* 9, 1–17.

Kaldor, N. 1950. The economic aspects of advertising. *Review of Economic Studies* 18, 1–27.

Lambin, J.J. 1976. *Advertising, Competition, and Market Conduct in Oligopoly over Time*. Amsterdam: North-Holland.

Nelson, P. 1974. Advertising as information. *Journal of Political Economy* 82, 729–54.

Porter, M. 1976. *Interbrand Choice, Strategy, and Bilateral Market Power*. Cambridge, MA: Harvard University Press.

Schmalensee, R. 1972. *The Economics of Advertising*. Amsterdam: North-Holland.

Sethi, S.P. 1977. Dynamic optimal control models of advertising. *SIAM Review* 19, 685–725.

Telser, L.G. 1964. Advertising and competition. *Journal of Political Economy* 72, 537–62.

Bibliographic addendum

Ackerberg, D.A. 2001. Empirically distinguishing informative and prestige effects of advertising. *RAND Journal of Economics* 32, 316–33.

Benham, L. 1972. The effect of advertising and the price of eyeglasses. *Journal of Law and Economics* 15, 337–52.

Cady, J.F. 1976. An estimate of the price effects of restrictions on drug price advertising. *Economic Inquiry* 14, 493–510.

Devine, D.G. and Marion, B.W. 1979. The influence of consumer price information on retail pricing and consumer behavior. *American Journal of Agricultural Economics* 61, 228–37.

Glazer, A. 1981. Advertising, information and prices: a case study. *Economic Inquiry* 19, 661–71.

Kwoka, J.E., Jr. 1984. Advertising and the price and quality of optometric services. *American Economic*

Review 74, 211–6.

Milyo, J. and Waldfogel, J. 1999. The effect of price advertising on prices: evidence in the wake of 44 Liquormart. *American Economic Review* 89, 1081–96.

How to cite this article

Schmalensee, Richard and David Genesove. "advertising." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000042> doi:10.1057/9780230226203.0012

The New Palgrave Dictionary of Economics Online

affine term structure models

Michael W. Brandt and David A. Chapman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An affine term structure model hypothesizes that interest rates, at any point in time, are a time-invariant linear function of a small set of common factors. This class of models has proven to be a remarkably flexible structure for examining the dynamics of default-risk free bonds, and as a result affine modelling has become the dominant framework for term structure research since the early 1980s.

Keywords

affine term structure models; arbitrage opportunities; bonds; Brownian motion; continuous-time models; expectations hypothesis; generalized method of moments; Itô integral; liquidity preference; maximum likelihood; multifactor models; Ornstein–Uhlenbeck processes; preferred habitat theory; principal components; single-factor models; state price deflators; term premiums; term structure of interest rates; zero-coupon bonds

Article

The term structure of interest rates refers to the relationship between the yields-to-maturity of a set of bonds and their times-to-maturity. It is a simple descriptive measure of the cross-section of bond prices observed at a point in time. An affine term structure model hypothesizes that the term structure of interest rates at any point in time is a time-invariant linear function of a small set of common state variables or factors. Once the dynamics of the state variables and their risk premiums are specified, the dynamics of the term structure are determined.
For the term structure of interest rates to be meaningful, the bonds being compared must have similar risk and payout characteristics. The literature we examine in this article focuses on the term structure of default-risk free nominal bonds that make a single payment at a pre-specified future date – so-called zero-coupon bonds. The models described below can be applied to other types of bonds, but zero-coupon bonds are particularly important because they represent the fundamental discount rates

embedded in all financial claims that make payments through time.

The literature on term structure modelling is large and reaches back to some of the giants of early 20th century economics: Fisher, Hicks, and Keynes. The pre-eminent model of the term structure, prior to the advent of affine models, was the expectations hypothesis. While the expectation hypothesis exists in a variety of forms (see Cox, Ingersoll and Ross, 1981), most researchers today use the definition of Campbell (1986) and Campbell and Shiller (1991) that the expected returns, or so-called term premiums, on default-risk-free zero-coupon bonds are constant through time. Other commonly espoused early term structure models, namely, the liquidity preference and preferred habitat theories, can be viewed as extensions of the expectation hypothesis that make additional predictions about the size of term premiums as a function of time-to-maturity. Most empirical tests of the expectations hypothesis, including Fama and Bliss (1987) and Campbell and Shiller (1991), find strong evidence against the prediction that term premiums are constant through time. This rejection of the expectations hypothesis implies that the prices of default-risk-free zero-coupon bonds embed time-varying term premiums. Explaining the dynamics of these term premiums is an important goal of affine term structure models.

Any affine term structure model starts from the assumption that there are no arbitrage opportunities in financial markets. This assumption implies the existence of a strictly positive stochastic process, Λ, that prices all assets. (See Duffie, 2001, for a textbook treatment of the implications of absence of arbitrage for asset pricing in general and term structure modelling in particular.) This process is typically referred to as a state price deflator in continuous-time models of asset pricing or as a stochastic discount factor in discrete-time models. We follow the more common approach in the literature and develop the affine term structure models in continuous time. The existence of a state price deflator also implies that there exists a risk-neutral measure, $\mathbb{Q}$, which is distinct from the physical measure, $\mathbb{P}$, that generates observed variation in asset prices.

Independent of any specific model of bond prices, it is always possible to express the price at time *t* of a zero coupon bond that matures at time $t + \tau$ as

$$P_t(\tau) = E_t^{\mathbb{Q}}\left[\exp\left(-\int_0^{\tau} r_s ds\right)\right], \tag{1}$$

where $E_t^{\mathbb{Q}}[\cdot]$ denotes the expected value at time *t* under the risk-neutral measure, and *r* is the instantaneous rate of interest (or short rate). The short rate can be defined as

$$r_t = \lim_{\tau \downarrow 0} \ln P_t(\tau), \tag{2}$$

but it is also related to the expected value of the instantaneous rate of change of the state price deflator because

$$\frac{d\Lambda_t}{\Lambda_t} = -r_t dt + \sigma_\Lambda(\Lambda_t, t)\, dW_t^{\mathbb{Q}}, \tag{3}$$

where $W_t^{\mathbb{Q}}$ is a Brownian motion under $\mathbb{Q}$, $\sigma_\Lambda(\cdot)$ is the possibly time-and state-dependent instantaneous volatility of the state price deflator, and the second term in (3) is a common shorthand notation for an Itô stochastic integral. (See Duffie, 2001, for a textbook treatment of continuous-time stochastic processes, including the definitions of Brownian motion and the Itô integral.)

As eq. (1) clearly shows, pricing zero-coupon default-risk-free bonds boils down to specifying a model for the dynamics of the short rate under the risk-neutral measure. In choosing models for r_t, there are two paramount considerations: (*a*) a flexible specification that does a reasonable job of capturing the dynamics of proxies for the short rate (since r_t itself is unobservable), and (*b*) a specification that yields a convenient form for the bond prices that are the ultimate objects of interest.

The dynamic of the short rate, when modelled in continuous time, are completely determined by the drift function, which defines the instantaneous expected value of the short rate, and the diffusion function, which determines the instantaneous volatility of the short rate. What is not clear from eq. (1) is that, in order to move from the theoretical risk-neutral measure, $\mathbb{Q}$, to the actual (or physical measure), $\mathbb{P}$, that generates the observed data, a term structure model must also specify a structure for the risk premium functions controlling the transformation between the measures $\mathbb{Q}$ and $\mathbb{P}$. While the risk-neutral measure is sufficient for pricing, researchers wanting to fit affine term structure models to observed time-series data or wanting to use these models to forecast future interest rates require also the actual measure.

We can now turn to the basic building blocks (that is, short rate dynamics and market price of risk assumptions) and the main pricing results (that is, exponentially linear bond prices) of affine term structure models. We first present the main points in the context of single-factor models and then generalize the discussion to the multifactor case. Chapman and Pearson (2001), Dai and Singleton (2003), and Piazzesi (2005) are all recent, more detailed, and more technical examinations of the material that follows.

Single-factor models

In a single-factor affine model, the determinant of bond prices is the short rate itself. The model is constructed by specifying a continuous-time process for the short rate and a form of the risk premium function. As Cox, Ingersoll, and Ross (1985) note, these choices must be mutually consistent in order to avoid accidentally introducing arbitrage opportunities into a (supposedly) arbitrage-free model. The

fundamental building blocks of all affine models are the single-factor models due to Vasicek (1977) and Cox, Ingersoll, and Ross (1985) (hereafter CIR).

The Vasicek model assumes that the short rate evolves as an Ornstein–Uhlenbeck process under the risk-neutral measure

$$dr_t = \kappa(\theta - r_t)dt + \sigma dW_t^{\mathbb{Q}}, \tag{4}$$

where $\kappa > 0$ determines the speed of reversion to the constant mean, $\theta > 0$, and σ is the unconditional instantaneous volatility of the process. The conditional and unconditional distributions of interest rate changes are Gaussian in this model. Accordingly, it is possible for the short rate to be negative. The risk premium function is a constant, λ_0, which implies that the short rate is also Gaussian under the physical measure, $\mathbb{P}$. Solving the conditional expectation in (1) under these assumptions generates an explicit expression for the price of a default-risk free zero coupon bond

$$P_t(\tau) = \exp[a(\tau) + b(\tau)r_t], \tag{5}$$

where

$$a(\tau) = \left(\theta - \frac{\lambda_0}{\kappa} - \frac{1}{2}\frac{\sigma^2}{\kappa^2}\right)\left[\frac{1}{\kappa}(1 - \exp(-\kappa\tau)) - \tau\right] - \frac{\sigma^2}{4\kappa^3}[1 - \exp(-\kappa\tau)]^2 \tag{6}$$

and

$$b(\tau) = -\frac{1}{\kappa}[1 - \exp(-\kappa\tau)]. \tag{7}$$

Equation (5) is the first statement of an exponential-affine pricing function. It implies a simple structure

where continuously compounded yields are Gaussian with constant volatility. The term structure of forward rates implied by this simple model can assume most (but not all) of the commonly observed shapes of the term structure. In particular, the term structure of forward rates can be upward sloping, downward sloping, or humped shaped, although the model cannot generate an inverted humped shape. Since prices at all maturities are driven by a single stochastic factor, this model implies that all yield levels are perfectly correlated. In the data, yield levels are very highly, but not perfectly, correlated.
In the single-factor CIR term structure model, the short rate evolves as

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t^Q \tag{8}$$

where $\kappa > 0$ and $\theta > 0$ have the same interpretation as in the Vasicek case, but the short rate is no longer Gaussian. The parameter restriction $2\kappa\theta \geq \sigma^2$ is imposed in order to ensure that the short rate process does not get trapped at zero. r_t has a conditional non-central chi-square distribution (and an unconditional Gamma distribution). The instantaneous conditional variance of the short rate is linear in the level of the rate. The risk premium specification that is consistent with no-arbitrage in the single-factor CIR specification is $\lambda(r_t) = \lambda_1 r_t$, and the no-arbitrage bond price is, again, of the form (5) with

$$a(\tau) = \frac{2\kappa\theta}{\sigma^2}\log\left[\frac{2\gamma\exp(\frac{1}{2}\tau(\kappa + \lambda_1 + \gamma))}{(\kappa + \lambda_1 + \gamma)[\exp(\gamma\tau) - 1] + 2\gamma}\right] \tag{9}$$

$$b(\tau) = \frac{-2[\exp(\gamma\tau) - 1]}{(\kappa + \lambda_1 + \gamma)[\exp(\gamma\tau) - 1] + 2\gamma}, \tag{10}$$

where $\gamma \equiv \sqrt{(\kappa + \lambda_1)^2 + 2\sigma^2}$. The CIR model can generate the most common shapes of the term structure, but it still implies that all yield levels are perfectly correlated.
The Vasicek and CIR models are the most common forms of single-factor affine models, but Duffie and Kan (1996) provide the conditions on the drift, diffusion, and risk premium functions of a short rate

specification, like (4) or (8), that ensure that the bond pricing function is exponential-affine under the risk neutral measure. In particular, a pricing function of the form of (5) will follow if

$$\mu(r_t) - \lambda(r_t) = \rho_0 + \rho_1 r_t \tag{11}$$

and

$$\sigma(r_t) = \sqrt{\beta_0 - \beta_1 r_t} \tag{12}$$

hold, where $\mu(r_t)$ is a general expression for the drift of the short rate and $\sigma(r_t)$ is a general expression for the instantaneous volatility of the short rate. For example, in the CIR case $\rho_0 = \kappa\theta$, $\rho_1 = -(\kappa + \lambda_1)$, $\beta_0 = 0$, and $\beta^1 = \sigma^2$. In this more general case, the $a(\tau)$ and $b(\tau)$ functions do not generally have explicit closed-form expressions. Rather, they are defined as the solutions to a pair of ordinary differential equations.

The empirical evidence clearly shows that a single-factor specification is not sufficient to describe the dynamics of the default-risk-free term structure. As such, empirical analysis of simple specifications, like (4) and (8), have shifted away from attempting to completely characterize yields on all maturities and, instead, have concentrated on explaining the dynamics of a proxy for the unobservable short rate. Chan et al. (1992) pioneered this approach, using a simple generalized method of moments estimation scheme. Durham (2003) is the natural evolution of this literature using state-of-the-art approximate maximum likelihood estimation. The conclusions of this literature are: (*a*) the evidence of mean reversion in the short rate is weak, at best, but (*b*) there is little consistent evidence of nonlinear mean reversion; and (*c*) there are complicated volatility dynamics that are not consistent with either constant volatility (Vasicek) or instantaneous conditional variances that are linear in the short rate (CIR).

Multifactor models

If single-factor models are insufficient to explain the observed term structure, then how many factors are needed and what are the dynamics of these factors? The common answer to the first question is provided by the analysis of Litterman and Scheinkman (1991). Using a simple principal components approach, they argue that three factors, extracted from bond yields or returns themselves, can explain well over 95 per cent of the variation in weekly changes of US Treasury bond prices, for maturities of up to 18 years. The answer to the second question – in the most general form consistent with an exponential-affine pricing function – is provided by Dai and Singleton (2000) and extended by Duffee (2002).

The multifactor affine term structure model consists of the following components. First, there is linear relation between the short rate and the factors:

$$r_t = \delta_0 + \delta' Y_t, \tag{13}$$

where Y_t denotes the N-vector of time t factor realizations. The factor dynamics conform to an affine diffusion

$$dY_t = K(\theta - Y_t)dt + \Sigma\sqrt{S_t}dW_t^{\mathbf{Q}}, \tag{14}$$

where K and Σ are $N \times N$ matrices (with no general restrictions) and S_t is a diagonal matrix with the i-th diagonal element equal to

$$[S_t^{ii}] = \alpha_i + \beta_i' Y_t. \tag{15}$$

The S_t matrix allows for the instantaneous conditional variance of the factors to be linear functions of factor levels. If every element of Y_t can affect the conditional volatility of every other factor, then (14) is a multifactor generalization of the CIR model from the last section. Of course, the fact that volatility is linear in the level of Y requires strong restrictions on the parameters of the model in order to ensure that variances are non-negative.

If no elements of Y affect the conditional volatility, then (14) is a multifactor generalization of the Vasicek model. If $m < N$ factors affect the conditional volatility, then the multifactor affine model is a mixture of the CIR and Vasicek forms. Dai and Singleton (2000) define different classes of affine models by the number of factors that affect the conditional factor volatilities, with $A_m(N)$ being the general notation for an N-factor model with m-factors driving conditional volatilities.

Under these assumptions, bond prices satisfy a multivariate generalization of (5) given by

$$P_t(\tau) = \exp[A(\tau) + B(\tau)' Y_t].$$

(16)

The functions $A(\tau)$ and $B(\tau)$ are the solutions to the ordinary differential equations

$$\frac{dA(\tau)}{d\tau} = -\theta\kappa' B(\tau) + \frac{1}{2}\sum_{i=1}^{N}[\Sigma' B(\tau)]_i^2 \alpha_i - \delta_0$$

(17)

and

$$\frac{dB(\tau)}{d\tau} = -\kappa' \beta(\tau) + \frac{1}{2}\sum_{i=1}^{N}[\Sigma' B(\tau)]_i^2 \beta_i - \delta.$$

(18)

The final component of the general multifactor affine model is the specification of the market prices of risk, which connects pricing under the risk-neutral measure to pricing under the physical measure:

$$\Lambda_t = \sqrt{S_t}\lambda_0 + \sqrt{S_t^-}\lambda Y_t,$$

(19)

where λ_0 is an *N*-vector of constants, λ is an $N \times N$ matrix of constants, and S_t^- is an *N*-dimensional diagonal matrix with diagonal elements equal to

$$S_t^-(ii) = \begin{cases} (\alpha_i + \beta_i' Y_t)^{-1/2}, & \text{if } \inf\ (\alpha_i + \beta_t' Y_t) > 0; \\ 0, & \text{otherwise} \end{cases}$$

(20)

The first term in (19) is a straightforward generalization of the single-factor risk premium specifications: risk premiums are proportional to factor volatilities. The second component is an important source of additional flexibility in multifactor affine models. It allows these models to provide a better fit to the distribution of bond excess returns, and it is also useful in rationalizing the observed violations of the expectations hypothesis discussed above.

The general multifactor affine model can be viewed as a blending of the Vasicek and CIR forms. These extreme specifications also reveal a critical trade-off in multifactor term structure modelling. The CIR form offers the greatest flexibility in specifying the volatility dynamics of bond prices. However, this flexibility comes at a cost. The parameter restrictions that are required to ensure that (15) provides a valid description of factor variances impose substantial restrictions on the permissible correlations between the factors. In the extreme case of the pure multifactor CIR model, the factors must be uncorrelated to ensure an admissible volatility specification.

Dai and Singleton (2002), Duffee (2002) and Brandt and Chapman (2005) fit multifactor affine term structure models to more than 25 years of monthly US bond data. Each paper considers the ability of different versions of $A_m(3)$ models to both explain the rejections of the expectations hypothesis and to provide accurate forecasts of future yields. Both Dai and Singleton (2002) and Brandt and Chapman (2005) find that a Gaussian version (an $A_0(3)$ model) can rationalize the risk premiums dynamics revealed by expectations hypothesis tests. Duffee (2002) demonstrates that an $A_0(3)$ model with the expanded risk premium specification of (19) can produce more accurate yield forecasts than a random walk benchmark model.

Although the ability to explain risk premiums and yield movements is an important success for multifactor affine models, their biggest failing to date is that the favoured Gaussian specifications require that conditional yield volatilities are constant. Essentially, the flexibility in factor correlations that are required to explain these features of the data require a stochastic structure that precludes the volatility dynamics that are an equally important feature of interest rate data.

Concluding remarks

Affine models have two important strengths compared with the earlier theories of the term structure. They explicitly rule out arbitrage opportunities in the cross-section of bond prices, and they simultaneously allow for flexible specifications of term premiums and their dynamics. Weaknesses of affine models include the fact that they are typically not easy to estimate, that model specifications which can explain the rejection of the expectations hypothesis are inconsistent with observed volatility dynamics, and that there is generally limited intuition as to the economic interpretation of the factors. Ang and Piazzesi (2003) and Ang, Dong, and Piazzesi (2005) are recent attempts to combine affine term structure modelling with elements of the macroeconomy. This line of research holds out the promise of greater intuition behind the factors as well as a greater understanding of how capital markets perceive the actions of monetary authorities.

See Also

- arbitrage
- continuous and discrete time models
- finance
- finance (new developments)
- linear models
- Markov processes
- term structure of interest rates
- Wiener process

Bibliography

Ang, A. and Piazzesi, M. 2003. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics* 50, 745–87.

Ang, A., Dong, S. and Piazzesi, M. 2005. No-arbitrage Taylor rules. *Federal Reserve Bank of San Francisco, Proceedings* 2005(12).

Brandt, M. and Chapman, D. 2006. Comparing multifactor models of the term structure. Manuscript. Duke University and Boston College.

Campbell, J. 1986. A defense of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 41, 183–93.

Campbell, J. and Shiller, R.J. 1991. Yield spreads and interest rate movements: a bird's eye view. *Review of Economic Studies* 58, 495–514.

Chan, K., Karolyi, G., Longstaff, F. and Sanders, A. 1992. An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* 47, 1209–27.

Chapman, D. and Pearson, N. 2001. Recent advances in estimating term structure models. *Financial Analysts Journal* 57, 77–95.

Cox, J.C., Ingersoll, J., Jr. and Ross, S. 1981. A re-examination of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 36, 769–99.

Cox, J., Ingersoll, J., Jr. and Ross, S. 1985. A theory of the term structure of interest rates. *Econometrica* 53, 385–408.

Dai, Q. and Singleton, K. 2000. Specification analysis of affine term structure models. *Journal of Finance* 55, 1943–978.

Dai, Q. and Singleton, K. 2002. Expectation puzzles, time-varying risk premia, and affine models of the term structure. *Journal of Financial Economics* 63, 415–41.

Dai, Q. and Singleton, K. 2003. Term structure dynamics in theory and reality. *Review of Financial Studies* 16, 631–78.

Duffee, G. 2002. Term premia and interest rate forecasts in affine models. *Journal of Finance* 57, 405–43.

Duffie, D. 2001. *Dynamic Asset Pricing Theory*, 3rd edn. Princeton: Princeton University Press.

Duffie, D. and Kan, R. 1996. A yield factor model of interest rates. *Mathematical Finance* 6, 379–406.

Durham, G. 2003. Likelihood-based specification analysis of continuous-time models of the short-term interest rate. *Journal of Financial Economics* 70, 463–87.

Fama, E. and Bliss, R. 1987. The information in long maturity forward rates. *American Economic Review* 77, 680–92.

Litterman, R. and Scheinkman, J. 1991. Common factors affecting bond returns. *Journal of Fixed Income* 3, 54–61.

Piazzesi, M. 2005. Affine term structure models. In *Handbook of Financial Econometrics*, ed. Y. Aït-Sahalia and L. Hansen. Amsterdam: North-Holland.

Vasicek, O. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5, 177–88.

How to cite this article

Brandt, Michael W. and David A. Chapman. "affine term structure models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000239> doi:10.1057/9780230226203.0013

The New Palgrave Dictionary of Economics Online

affirmative action

Harry J. Holzer and David Neumark

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Affirmative action practices go beyond non-discrimination to enhance employment, education, and business-ownership opportunities for minorities and women. Critics argue that affirmative action does this at the expense of white males who might be more qualified, and so could be both unfair and inefficient. Supporters claim that affirmative action is necessary to overcome the many inherent disadvantages faced by minorities and women, and could enhance efficiency by expanding the pool of available talent or because diversity itself has positive impacts. This article summarizes the evidence for these arguments and claims.

Keywords

affirmative action; black–white wage differences; efficiency; labour market discrimination; labour-market institutions; National Educational Longitudinal Study (NELS); redistribution; women's work and wages

Article

‘Affirmative action’ refers to a set of practices undertaken by employers, university admissions offices, and government agencies to go beyond non-discrimination, and actively improve the economic status of minorities and women with regard to employment, education, and business ownership and growth.

Legal underpinnings and controversies

The roots of affirmative action in employment lie in a set of Executive Orders issued by US Presidents since the 1960s. Executive Order 10925 (issued in 1961) introduced the term ‘affirmative action’, encouraging employers to take action to ensure non-discrimination. Executive Order 11246 (1965) required federal contractors and subcontractors (currently, with contracts of $50,000 or more) to identify

underutilized minorities, to assess availability of minorities, and if available, to set goals and timetables for reducing the underutilization. Executive Order 11375 (1967) extended this requirement to women. Federal contractors may be sued and barred from contracts if they are judged to be discriminating or not pursuing affirmative action, although this outcome is rare (Stephanopoulos and Edley, 1995). But affirmative action is not just limited to contractors; it can be imposed on non-contractor employers by courts as a remedy for past discrimination, and it can be undertaken voluntarily by employers.

While universities may be bound by affirmative action in employment in their role as federal contractors, there are no explicit federal policies regarding affirmative action in university admissions. Rather, universities have voluntarily implemented affirmative action admissions policies that are widely regarded as giving preferential treatment to women and minority candidates. Court decisions have shaped (and continue to shape) what universities can and cannot do. Preferential admissions policies initially came under attack in *Bakke v. University of California Regents* (1978), in which the Supreme Court declared that policies that set aside a specific number of places for minority students violated the 14th Amendment of the US Constitution, which bars states from depriving citizens of equal protection of the laws. However, while this decision is viewed as declaring strict quotas illegal, it is also interpreted as ruling that race can be used as a flexible factor in university admissions.

Most recently, the Supreme Court in 2003 struck down the undergraduate admissions practices at the University of Michigan in the case of *Gratz v. Bollinger* et al., finding that the point system used by the university in its consideration of race (and other criteria) was too rigid. At the same time, in *Grutter v. Bollinger* et al., it upheld the university's law school admissions procedures, finding that the more flexible treatment of race in this case satisfied the state's compelling interest in expanding the pool of minority candidates admitted to this prestigious school. Affirmative action can also be limited by popular referenda; voters passed Proposition 209 in California in the 1990s, barring the use of racial preferences in admissions to public universities (as well as in state employment and contracting).

The third major component of affirmative action is contracting and procurement programmes. At the federal level, these have principally taken the form of preferential treatment in bidding for Small/Disadvantaged Businesses (SDBs), and Small Business Administration programmes of technical assistance. These contracting and procurement programmes focus more on minorities than on women (Stephanopoulos and Edley, 1995, Section 9). In addition to the federal government, numerous states and localities have used programmes aimed at increasing the share of contracts awarded to minority-owned businesses.

As with affirmative action in education, court rulings since the late 1980s have challenged the legal standing of such programmes. *City of Richmond v. J. A. Croson Co*. (1989) established that the legal standard of ‘strict scrutiny’ for compelling state interests must be met for state programmes to be legal under the 14th Amendment to the Constitution. In *Adarand Constructors, Inc. v. Pena* (1995), the Supreme Court ruled that strict scrutiny could apply to federal programmes as well, invoking the Fifth Amendment (which guarantees that citizens shall not ‘be deprived of life, liberty, or property, without due process of law’), instead of the 14th (which explicitly applies to states).

Affirmative action remains vastly more controversial than anti-discrimination activity, even though the distinctions between them are clearer in theory than in practice (Holzer and Neumark, 2000a). The critics of affirmative action argue that it transfers jobs, university admissions, and business contracts to minorities and women at the expense of white males who might be more qualified and therefore more deserving. If so, it might constitute a form of ‘reverse discrimination’ against white males, which could

be both inefficient and unfair. In contrast, the supporters of affirmative action claim that extra efforts beyond just the removal of explicit discrimination are necessary to overcome the many inherent disadvantages that minorities and women face in universities, the labour market, and the business sector. On this view, affirmative action is necessary for equal opportunity (or ‘fairness’), and would not necessarily reduce efficiency. Indeed, it might even raise overall efficiency by making available a wider pool of talent on which businesses and universities could draw, or because diversity itself has positive impacts.

The economic impacts of affirmative action largely centre on two issues: (*a*) the actual magnitudes of the *redistribution* of jobs, university admissions, or business contracts from white males to minorities or women attributable to affirmative action; and (*b*) any effects of affirmative action on *efficiency*, as measured (for example) by the credentials or performance of those who receive preferential treatment relative to those who do not. Evidence on these issues does not settle the ‘fairness’ question, which ultimately depends on personal values. But the evidence can and should inform the debate. A comprehensive review of the evidence is provided in Holzer and Neumark (2000a).

Redistributive effects

At this point, there seems to be little doubt that racial or gender preferences redistribute certain jobs or university admissions away from white men towards minorities and women. The question, instead, involves the magnitudes of these shifts. In terms of the labour market, a wide range of studies have demonstrated that affirmative action has shifted employment within the contractor sector from white males to minorities and women. But the magnitudes of these shifts are not necessarily large. For instance, Leonard (1990) found that employment of black males grew about five per cent faster at contractor establishments in the critical period of 1974–80 (when affirmative action requirements on contractors were rigorously enforced for the first time) than did employment of white males, while for white females and black females there were somewhat more modest effects. Looking at cross-sectional differences across establishments that did and did not use affirmative action in hiring (rather than using actual contractor status), Holzer and Neumark (1999) found that the share of total employment accounted for by white males was about 15–20 per cent lower in establishments using affirmative action than in those that did not – which is broadly consistent with the findings of Leonard and others. This does not necessarily imply that employment of white males overall is reduced by affirmative action, but only that it is redistributed to the non-affirmative action sector (where wages and benefits are likely lower).

The magnitude of the redistribution of university admissions from white males to minorities or women generated by affirmative action has been debated. On the one hand, test scores of those admitted are considerably higher among whites than minorities across the full spectrum of colleges and institutions (Datcher Loury and Garman, 1995). But at least some of these differences could be generated even with a common test score cut-off, given the racial gaps in test scores that exist in the population. And, if test scores are worse predictors of subsequent performance among blacks than whites, it might be perfectly rational for schools to put less weight on test scores in the admissions process for blacks (Dickens and Kane, 1999).

Furthermore, analyses of micro-level data on applications and admissions by Kane (1998) and by Long

(2004) suggest somewhat modest effects of affirmative action on overall admissions of minorities, but both studies suggest that the magnitudes rise with the overall level of scores at universities. Using data from the High School and Beyond Survey, Kane found significant racial differences in admissions (conditional on test scores and many other personal characteristics) only in the top quintile of colleges and universities by test scores. Long, using data from the National Educational Longitudinal Study (NELS), found significant effects on admissions in all quintiles. But the magnitudes of these differences were not large in absolute terms – the probability that minorities are accepted at their top choice would decline by less than two percentage points (14.7 per cent against 16.4 per cent) in the aggregate and about 2.5 percentage points in the top quintile in the absence of affirmative action.

That affirmative action is more important as college quality rises is further established by Bowen and Bok (1998), who find quite large effects at a set of the most prestigious colleges and universities. Indeed, their work suggests that admissions rates among minorities at these schools would fall from 42 per cent to 13 per cent if affirmative action were abolished, a view consistent with the initial effects of Proposition 209 in California on admissions at Berkeley. The magnitudes of racial preferences in admissions in a variety of graduate programmes are also fairly large (Attiyeh and Attiyeh, 1997; Davidson and Lewis, 1997), while gender preferences are much more modest.

Overall, the elimination of affirmative action in admissions to elite schools or graduate programmes would likely generate large reductions in minority student enrolments, but only modest improvements in overall grades and test scores at these institutions, as the whites who would be admitted in place of them appear to perform only marginally better in terms of these measures (Bowen and Bok, 1998). Implementing the reforms that have been recently adopted in Texas, Florida, and elsewhere, where admissions are based only on class rank rather than minority status, would likely generate major reductions as well in the presence of minorities on campus (Long, 2004). And using preferences based on family income instead of race or gender in admissions would also result in large declines in minority representation at universities.

As for the redistribution of contracts from white-owned to minority- or female-owned businesses, we know of no study that has attempted to carefully measure the magnitude of this shift, though some summary studies suggest that the effects might be substantial.

Efficiency and performance effects

Regarding labour markets, it is fairly clear in theory that affirmative action could reduce efficiency in well-functioning labour markets in the short run if minorities or women were assigned to jobs for which they were not fully qualified, while it could increase efficiency if it opened up to minorities or women jobs from which they had been excluded in favour of less qualified white males. On the other hand, affirmative action might also lead minorities and women to invest in more education and training if the rewards to this investment would be increased; however, whether affirmative action would change incentives in this way is uncertain (Coate and Loury, 1993). The positive benefits on skill development across generations might be important as well. Finally, diversity per se may bring benefits, such as fostering mentoring relationships (Athey, Avery and Zemsky, 2000). To a large extent, the more important the imperfection in the labour market associated with the lower relative status of minorities – such as negative externalities generated for other members of the community, or imperfect information

driving the outcome – the greater is the chance that affirmative action will not reduce efficiency, and might even raise it.

A similar point can be made regarding university admissions. Significant market imperfections are likely to impede university admissions for some groups – such as imperfect information among university officials about individual candidates (or vice versa), and capital market problems that limit the access of lower-income groups to finance. Furthermore, important externalities might exist in the education process, at least along certain dimensions. For instance, students might learn more from one another in more diverse settings; indeed, the value of being able to interact with those of other ethnicities or nationalities might be growing over time, as product and labour markets become more diverse and more international. Alternatively, race-specific or gender-specific role models might be important for some individuals in the learning process.

What does the empirical evidence on the efficiency and performance of affirmative action beneficiaries show? One approach is to look at measures of individual employee credentials or performance, by race and/or sex, to see whether affirmative action generates major gaps in performance between white males and other groups. An earlier paper (Holzer and Neumark, 1999) compares a variety of measures of employee credentials and performance, where the former include educational attainment (absolute levels and those relative to job requirements), and the latter include wage or promotion outcomes as well as a subjective performance measure across these groups. The study inquired whether observed gaps in credentials and performance between white males and females or minorities are larger among establishments that practice affirmative action in hiring than among those that do not. The results indicated virtually no evidence of weaker credentials or performance among females in the affirmative action sector, relative to those of males within the same racial groups. In comparisons between minorities and whites, there was clear evidence of weaker educational credentials among the former group, but relatively little evidence of weaker performance.

But how could affirmative action result in minorities with weaker credentials but not weaker performance, if educational credentials generally are meaningful predictors of performance? In a separate paper, Holzer and Neumark (2000b) considered various mechanisms by which firms engaging in affirmative action might offset the productivity shortfalls among those hired from ‘protected groups’ that would otherwise be expected. The study found that firms engaging in affirmative action: (*a*) recruit more extensively; (*b*) screen more intensively and pay less attention to characteristics such as welfare recipiency or limited work experience that usually stigmatize candidates; (*c*) provide more training after hiring; and (*d*) evaluate worker performance more carefully.

Thus, these firms tend to cast a wider net with regard to job applicants, gather more information that might help uncover candidates whose productivity is not fully predicted by their educational credentials, and then invest more heavily in the productivity of those whom they have hired. This view is consistent with a variety of case studies (for example, Badgett, 1995), and other work in the literature on employee selection, suggesting that affirmative action works best if employers use a broad range of recruitment techniques and predictors of performance when hiring, and when they make a variety of efforts to enhance performance of those hired. In these studies, affirmative action need not just ‘lower the bar’ on expected performance of employees hired, and generally does not appear to do so (though some exceptions exist).

A variety of other studies have been undertaken within specific sectors of the workforce, where it is easier to define employee performance. Among the sectors that have been studied are police forces

(Carter and Sapp, 1991), physicians (Davidson and Lewis, 1997), and university faculties (Kolpin and Singell, 1996). The results of these studies again show no evidence of weaker performance among women, and generally limited evidence of weaker performance among minorities. In contrast, there is evidence of potential social benefits from affirmative action in the medical sector, as minority doctors appear more likely to locate in poor neighbourhoods and treat minority or low-income patients.

Thus, the existing research finds evidence of weaker credentials but only limited evidence of weaker labour market performance among the beneficiaries of affirmative action, and evidence (at least in one important sector) consistent with positive externalities.

Regarding university admissions, there are gaps in high school grades and test scores between white and minority students admitted at universities, and the college grades of minorities lag behind as well. Black students fail to complete their college degrees at significantly higher rates, especially at institutions with higher average test scores (Datcher Loury and Garman, 1995). Similar findings have been generated for law schools (Sander, 2004). On the other hand, there is some evidence that the lower college completion rates among blacks at more selective institutions disappear once one controls for the effects of attending the historically black colleges and universities (Kane, 1998). And earnings are generally higher among blacks (as well as whites) who attend more prestigious and highly ranked schools, despite their higher rates of failure there.

The more challenging question is whether affirmative action actually hurts minority students by admitting them to colleges and universities for which some of them are unqualified, generating a poor ‘fit’ between them and the colleges or universities that they attend that may actually lead to worse outcomes. Sander (2004) claims to show evidence that affirmative action in law schools worsens outcomes for blacks, although this conclusion is disputable. Conversely, dropout rates of minorities at the most prestigious institutions are generally lower than elsewhere (Bowen and Bok, 1998). More decisive evidence on this question requires adequate comparison with counterfactuals of what would be observed absent affirmative action.

Along some other dimensions, the benefits of affirmative action in generating greater understanding and positive interactions across racial groups have been documented at these schools (Bowen and Bok, 1998). There is limited evidence of direct educational benefits of the diversity that affirmative action promotes (Antonio et al., 2004), although not yet in terms of the economic returns to education on which economists tend to rely in assessing educational outcomes. And evidence on the effects of minority or female faculty ‘mentoring’ and ‘role models’ is mixed (for example, Neumark and Gardecki, 1998).

Finally, the evidence on the performance of female- or minority-owned businesses that obtain more contracts as a result of affirmative action rules is somewhat inconclusive as well. Amendments to Section 8(a) rules on federal contracting do not allow companies to receive contracts under these provisions for longer than nine years, and apparently those who ‘graduate’ from the programme seem to perform (at least in terms of staying in business) as well as firms more generally (Stephanopoulos and Edley, 1995). On the other hand, there is some evidence of higher failure rates among firms that currently receive a high percentage of their revenues from sales to local government (Bates and Williams, 1995). The higher failure rates may be attributable to the fact that a significant fraction of the latter are ‘front’ companies that have formed or reorganized in an attempt to gain Section 8(a) contracts. There is also evidence that failure rates can be limited with the right kinds of certification and technical

assistance, especially if the reliance of the companies on governmental revenues is limited as well. In any event, this evidence suggests that failing companies are not being ‘propped up’ by government contracts, as is commonly alleged. But stronger data and analysis are needed in this area before conclusions can be drawn with a greater degree of confidence on the issue of the efficiency of minority contracting programmes.

See Also

- black–white labour market inequality in the United States
- labour market institutions

Bibliography

Antonio, A. et al. 2004. Effects of racial diversity on complex thinking in college students. *Psychological Science* 15, 507–10.

Athey, S., Avery, C. and Zemsky, P. 2000. Mentoring and diversity. *American Economic Review* 90, 765–86.

Attiyeh, G. and Attiyeh, R. 1997. Testing for bias in graduate school admissions. *Journal of Human Resources* 32, 524–48.

Badgett, L. 1995. Affirmative action in a changing legal and economic environment. *Industrial Relations* 34, 489–506.

Bates, T. and Williams, D. 1995. Preferential procurement programs and minority-owned businesses. *Journal of Urban Affairs* 17, 1–17.

Bowen, W. and Bok, D. 1998. *The Shape of the River*. Princeton, NJ: Princeton University Press.

Carter, D. and Sapp, A. 1991. *Police Education and Minority Recruitment: The Impact of a College Requirement*. Washington, DC: Police Executives Research Forum.

Coate, S. and Loury, G. 1993. Will affirmative action policies eliminate negative stereotypes? *American Economic Review* 83, 1220–40.

Datcher Loury, L. and Garman, K. 1995. College selectivity and earnings. *Journal of Labor Economics* 13, 289–308.

Davidson, R. and Lewis, E. 1997. Affirmative action and other special considerations admissions at the University of California, Davis School of Medicine. *Journal of the American Medical Association* 278,

1153–58.

Dickens, W. and Kane, T. 1999. Racial test score differences as evidence of reverse discrimination: less than meets the eye. *Industrial Relations* 28, 331–63.

Holzer, H. and Neumark, D. 1999. Are affirmative action hires less qualified? *Journal of Labor Economics* 17, 534–69.

Holzer, H. and Neumark, D. 2000a. Assessing affirmative action. *Journal of Economic Literature* 38, 483–568.

Holzer, H. and Neumark, D. 2000b. What does affirmative action do? *Industrial and Labor Relations Review* 53, 240–71.

Kane, T. 1998. Racial preferences and higher education. In C. Jencks and M. Phillips, ed. *The Black-White Test Score Gap*. Washington, DC: Brookings Institution.

Kolpin, V. and Singell, L. 1996. The gender composition and scholarly performance of economics departments: a test for employment discrimination. *Industrial and Labor Relations Review* 49, 408–23.

Leonard, J. 1990. The impact of affirmative action regulation and equal opportunity law on employment. *Journal of Economic Perspectives* 4, 47–63.

Long, M. 2004. Race and college admissions: an alternative to affirmative action? *Review of Economics and Statistics* 86, 1020–33.

Neumark, D. and Gardecki, R. 1998. Women helping women? role model and mentoring effects on female Ph.D. students in economics. *Journal of Human Resources* 33, 220–46.

Sander, R. 2004. A systematic analysis of affirmative action in American law schools. *Stanford Law Review* 57, 368–483.

Stephanopoulos, G. and Edley, C. 1995. Review of federal affirmative action programs. Unpublished White House document. Online. Available at http://clinton4.nara.gov/textonly/WH/EOP/OP/html/aa/aa-lett.html, accessed 25 August 2005.

How to cite this article

Holzer, Harry J. and David Neumark. "affirmative action." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_A000228> doi:10.1057/9780230226203.0014

The New Palgrave Dictionary of Economics Online

agency problems

Luca Anderlini and Leonardo Felli
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We illustrate agency problems with the aid of heavily stripped-down models which can be explicitly solved. Variations on a principal–agent model with both actors risk-neutral allow us to illustrate a canonical benchmark case, multi-tasking problems and informed-principal ones. We illustrate intertemporal agency problems using a two-period model with a risk-averse agent, which yields linear incentives. We conclude by briefly looking at more recent developments of the field such as present-biased preferences and motivated agents.

Keywords

agency problems; commitment; common values; continuous-time models; contract theory; discrete-time models; first-order approach; incentive design; insurance–incentives trade-off; intertemporal incentives; limited liability; linear incentive schemes; menu contracts; noisy tasks; non-profit organizations; pooling equilibria; principal and agent; separating equilibria; signalling; soft incentives

Article

Within modern economic analysis, early recognition of the importance of agency problems goes back to at least Marschak (1955), Arrow (1963) and Pauly (1968). These early works are followed by the classical contributions of Mirrlees (1975), Holmström (1979), Shavell (1979) and Grossman and Hart (1983).

The canonical form of the principal–agent problem still in use crystallizes in Holmström (1979) and Grossman and Hart (1983). A risk-neutral Principal $\mathcal{P}$ hires a risk-averse Agent $\mathcal{A}$. Both actors are necessary to generate output, which depends stochastically on $\mathcal{A}$'s actions. These are generally referred to as ‘effort’ (e) and, crucially are *not observable* by $\mathcal{P}$ or any third party like a Court. In jargon, effort is neither *observable* nor *verifiable*, and hence no contractual arrangements can depend on e. (Anderlini and Felli, 1998, consider a principal–agent problem in which e is in principle contractible, but where the equilibrium contract does not include it because of complexity considerations arising from the difficulties of *describing* it.) The interests of $\mathcal{P}$ and $\mathcal{A}$ are not aligned because e causes disutility to $\mathcal{A}$.

$\mathcal{P}$ makes a take-it-or-leave-it offer of a contract to $\mathcal{A}$ that specifies a schedule of output-contingent wages. $\mathcal{P}$'s offer is rejected unless it meets $\mathcal{A}$'s *individual rationality* constraint (henceforth *IR*), stating that $\mathcal{A}$'s expected utility cannot be less than that yielded by his next-best alternative employment. In addition, the problem may or may not include an explicit *limited liability* constraint (henceforth *LC*) stating that, regardless of output, $\mathcal{A}$'s wage cannot go below a given level. After a contract is signed, $\mathcal{A}$ chooses e, then the uncertain output is realized, and finally payments are made according to the contract.

In the canonical model there is a trade-off between *insurance* and *incentives*. Optimal risk-sharing would require $\mathcal{P}$ to insure $\mathcal{A}$ against output uncertainty. However, doing so would leave $\mathcal{A}$ without any *incentives* to exert effort: $\mathcal{A}$ would be guaranteed a constant wage and hence would choose that e which gives minimal disutility. Typically, $\mathcal{P}$'s choice is instead to offer a contract that does *not* fully insure $\mathcal{A}$, so as to give him incentives to exert effort. The contract compensates $\mathcal{A}$ for the risk he bears in order to satisfy the *IR* (and possibly the *LC*). If e is sufficiently productive in the stochastic technology, $\mathcal{P}$'s expected profit increases as a result. The need to generate effort via incentives yields an *agency problem*. The equilibrium contract may be far from the ‘first-best’ world in which a social planner can choose e at will. A lower than ‘socially efficient’ e is selected and $\mathcal{A}$ is not fully insured.

When both $\mathcal{P}$ and $\mathcal{A}$ are risk-neutral, an agency problem also arises if the *LC binds* (and typically the *IR* does not). (If the reverse is true, then giving incentives to $\mathcal{A}$ has no cost since he does not mind risk and the *IR* binds on his *expected* payoff. In fact in this case, the ‘social optimum’ coincides with the ‘constrained social optimum’ in which a social planner can

choose *e*, but only subject to giving the appropriate incentives to $\mathcal{A}$.) In this case in order to give $\mathcal{A}$ incentives $\mathcal{P}$ can pay him more when output indicates that effort is higher. This drives a wedge between $\mathcal{P}$'s marginal cost for increased *e* and its social marginal cost. This in turn dictates that the equilibrium contract will differ from the first-best, and a 'second-best' 'constrained-inefficient' outcome obtains.

Because of its tractability, the case in which both $\mathcal{P}$ and $\mathcal{A}$ are risk-neutral and the *LC* binds while the *IR* does not is a good benchmark to illustrate the mechanics of the problem and some of the more recent developments of the theory.

A simple benchmark

$\mathcal{P}$ hires $\mathcal{A}$ to carry out a task that requires unobservable non-contractible effort $e \in [0, 1]$. $\mathcal{A}$'s effort determines the probability that the task is successful in generating output. Output equals 1 with probability *e* and 0 with probability $1 - e$. Output is observable and contractible. First, $\mathcal{P}$ offers a contract to $\mathcal{A}$, then $\mathcal{A}$ accepts or rejects it. After a contract is signed, $\mathcal{A}$ chooses *e*.

A *contract* is a pair of reals (w_1,w_0), with the first being the wage (in units of output) that $\mathcal{P}$ pays $\mathcal{A}$ if output is 1, and the second being the wage if output is 0. Importantly, $\mathcal{A}$ has *limited liability*. He cannot be paid a negative wage in any state of the world. This generates the two *LC*s $w_1 \geq 0$ and $w_0 \geq 0$.

Both $\mathcal{P}$ and $\mathcal{A}$ are risk-neutral, and $\mathcal{A}$ dislikes effort which generates disutility $e^2/2$. Given (w_1,w_0) and *e*, $\mathcal{P}$'s payoff is $e(1 - w_1) - (1 - e)w_0$, while $\mathcal{A}$'s is given by $ew_1 + (1 - e)w_0 - e^2/2$. The outside options of both $\mathcal{P}$ and $\mathcal{A}$ are normalized to zero, so that in equilibrium both expected payoffs must be non-negative. These are the *IR*s.

Given (w_1,w_0), $\mathcal{A}$'s choice of *e* is immediately computed as $e = w_1 - w_0$, this is the *incentive* constraint (henceforth *IC*) of the agent. If both w_0 and w_1 are lowered by the same amount *e* does not change. Hence in equilibrium $w_0 = 0$ and $e = w_1$. Taking into account *IC*, $\mathcal{P}$ maximizes $e(1 - e)$. Therefore, in equilibrium, $e = w_1 = 1/2$. Hence $\mathcal{P}$'s equilibrium payoff is $\Pi^{\mathcal{P}} = 1/4$, while $\mathcal{A}$'s is $\Pi^{\mathcal{A}} = 1/8$, so that the *IR* does not bind for either $\mathcal{A}$ or $\mathcal{P}$.

If a social planner were able to choose *e* at will, this would be chosen so as to maximize $e - e^2/2$, expected output minus cost of effort. So the first-best level of effort is $e = 1$. In this hypothetical world, $\Pi^{\mathcal{P}} + \Pi^{\mathcal{A}} = 1/2$, while in equilibrium $\Pi^{\mathcal{P}} + \Pi^{\mathcal{A}} = 3/8$. This gap is the result of the *agency problem*; $\mathcal{A}$ is motivated by the difference $w_1 - w_0$. Because of limited liability, the only way for $\mathcal{P}$ to motivate $\mathcal{A}$ is to raise w_1. This makes $\mathcal{A}$'s effort too costly at the margin for $\mathcal{P}$: the (expected) cost of effort *e* is $w_1 e = e^2$, so that the marginal cost is 2*e*. This exceeds the social marginal cost, which is $\partial / \partial e[e^2/2] = e$, thus inducing an inefficient second-best outcome.

Multi-tasking

Starting with Holmström and Milgrom (1991), the theory evolved to encompass the multi-tasking case in which $\mathcal{A}$ has to carry out multiple tasks that affect output. (See also Holmström and Milgrom, 1994.) Some of the insights can be conveyed adapting the simple benchmark model above.

$\mathcal{A}$ now has two tasks; one is 'standard' (*S*) and one is 'noisy' (*N*). He chooses two effort levels: e_S and e_N, both in [0, 1]. Choosing (e_S,e_N) costs $\mathcal{A}$ a disutility of $(e_S^2 + e_N^2)/4$. The two tasks are perfect complements in the stochastic technology. Given (e_S,e_N), output equals 1 with probability min {e_S,e_N}, and 0 with probability 1–min{e_S,e_N}. As in the benchmark, $\mathcal{P}$'s payoff is expected output minus expected wage, while $\mathcal{A}$'s payoff equals his expected wage minus the disutility of effort. The *LC* and *IR* are as before.

Task *N* is noisier than task *S* in the following sense. Output is *not* contractible. Instead, each task yields a *binary signal* that can be contracted on. The signal σ_S for the *S* task is equal to 1 with probability e_S, and 0 with probability $1 - e_S$. The signal σ_N for the *N* task is equal to 1 with probability $[e_N p + (1 - e_N)(1 - p)]$ and equal to 0 with the complementary probability, with $p \in [1/2, 1]$. So, if $p = 1/2$ then σ_N contains no information about e_S, while if $p = 1$, the signals σ_S and σ_N are equally informative about the respective tasks.

Because of the signal structure, a contract is now a quadruple of wages (w_{S1},w_{S0},w_{N1},w_{N0}), one for each task, and for each possible value of the corresponding signal. As in the benchmark, in equilibrium we must have $w_{S0} = w_{N0} = 0$. Given (w_{S1},w_{S0},w_{N1},w_{N0}), the *IC*s pin down e_S and e_N as satisfying $e_S = 2w_{S1}$ and $e_N = 2w_{N1}(2p - 1)$. Maximizing $\mathcal{P}$'s profit using these restrictions gives that in equilibrium $e_S = e_N = \max\{0, 1/2 - (1 - p)/(8p - 4)\}$. When $p = 1$ this model yields the same first best and the equilibrium payoffs as the benchmark above. When $p = 3/5$ or less then $e_S = e_N = 0$.

The literature highlights some features of the equilibrium for values of $p \in [1/2, 1]$. As *p* decreases, so that task *N* becomes more noisy, two changes occur. In equilibrium, e_N decreases. This is not very surprising, given the increased noise. What is less straightforward is that e_S decreases as well: increased noise yields *softer* incentives on the *standard task*, as well as the noisy one. The complementarity between the tasks (extreme in the version used here, but this is not necessary) dictates that, as e_N becomes more expensive for $\mathcal{P}$ because of the noise, he will choose to induce lower values of e_S as well. Another way to check this is that the equilibrium values of both w_S and w_N decrease as *p* goes down. When $p \leq 3/5$, σ_N

is not informative enough. In this case $e_S = e_N = w_{S1} = w_{N1} = 0$. This has been interpreted as *no contract* being signed. The no-contract outcome obtains even though an informative contractible signal for *both* tasks is available.

Informed principal

Myerson (1983) and Maskin and Tirole (1990; 1992) examine the case in which $\mathcal{P}$ has private information, creating a potential signalling role for the contract offer. Despite the intricacies involved, the simple benchmark model above can be adapted again to illustrate some of the key points. (The computations below all pertain to the case of 'common values' analysed in Maskin and Tirole, 1992.)

There are two *types* of principal, $\mathcal{P}_H$ and $\mathcal{P}_L$. $\mathcal{P}$ is of type *H* with probability $\varphi = 18/29$ and of type *L* with probability $1 - \varphi = 11/29$. The principal's type is his *private information*. If $\mathcal{P}$ is of type *H*, $\mathcal{A}$'s outside option is $k = 9/32$, while if $\mathcal{P}$ is of type *L* then $\mathcal{A}$'s outside option is 0, as in the benchmark above. Hence, if $\mathcal{P}_H$ and $\mathcal{P}_L$ *separate* in equilibrium, there are two *IR*s for $\mathcal{A}$, while if pooling obtains $\mathcal{A}$'s expected outside option is $\varphi k = 81/464$, and he faces a single *IR*. $\mathcal{A}$'s *LC*s are as in the benchmark above.

First $\mathcal{P}$ learns his type. Then he offers a contract to $\mathcal{A}$, which may take the form of a *menu* (wages contingent on output and $\mathcal{P}$'s type). At this point $\mathcal{A}$ updates his beliefs about $\mathcal{P}$'s type and then decides whether to accept or reject. (As in any signalling game, the issue of off-the-equilibrium-path beliefs arises. The simplest way to deal with this issue is to assume that $\mathcal{A}$'s beliefs after observing an 'unexpected' offer are that $\mathcal{P}$ is of type *H* with probability 1. This is implicitly assumed in all computations below.) After a contract is signed $\mathcal{P}$ tells $\mathcal{A}$ which part of the menu applies in his case (if the contract is in fact a menu). Finally, $\mathcal{A}$ chooses effort, output is realized and payoffs are obtained.

There is a single task requiring effort which stochastically produces output as in the benchmark model. Output is contractible. $\mathcal{P}$'s payoffs and *IR* are also as above. $\mathcal{A}$'s payoff is also as in the benchmark above, except that he takes expectations using his beliefs.

In a separating equilibrium $\mathcal{P}_H$ and $\mathcal{P}_L$ offer two distinct pairs of output-contingent wages: (w_{H1},w_{H0}) and (w_{L1},w_{L0}) respectively. $\mathcal{A}$'s *IC*s dictate that after being offered (w_{H1},w_{H0}) effort is $e_H = w_{H1} - w_{H0}$, while after being offered (w_{L1},w_{L0}) effort is $e_L = w_{L1} - w_{L0}$.

Separation requires that neither $\mathcal{P}_H$ nor $\mathcal{P}_L$ has an incentive to offer the other type's wage pair. Since $\mathcal{P}$'s private information does not enter directly his payoff, this can be true only if the expected profits for the two types of principals, Π_H and Π_L, are the same. This is the *truth telling* (henceforth *TC*) constraint, which, using *IC*, since w_{H0} can be shown to be 0, reads $\Pi_H = e_H(1 - e_H) = e_L(1 - e_L) - w_{L0} = \Pi_L$. Since $k = 9/32$, one of the two *IR*s for the agent does bind. Using *IC* this yields $e_H = w_{H1} = 3/4$. Using *TC*, this implies $e_L = 1/2$, $w_{L0} = 1/16$ and $w_{L1} = 9/16$. With these values $\Pi_H = \Pi_L = 3/16$.

With informed principals, the literature highlights the possibility of *pooling* equilibria, in which the contract is a *menu*. Both $\mathcal{P}_H$ and $\mathcal{P}_L$ offer a menu $(w_{H1}^M, w_{H0}^M, w_{L1}^M, w_{L0}^M)$, which $\mathcal{A}$ has to accept or reject based on his expected *IR*. After a contract is signed, $\mathcal{P}$ tells $\mathcal{A}$ which pair of output-contingent wages applies. The *TC* constraint still applies, since both $\mathcal{P}_H$ and $\mathcal{P}_L$ have to be willing to indicate to $\mathcal{A}$ the appropriate wage pair. In fact, using *IC* and $w_{H0}^M = 0$, *IC* still reads $\Pi_H^M = e_H^M(1 - e_H^M) = e_L^M(1 - e_L^M) - w_{L0}^M = \Pi_L^M$. Using the single binding expected *IR* and the *IC*s, which are unchanged, yields $(18/58)(e_H^M)^2 + (11/29)[(e_L^M)^2 + w_L^M] = 81/464$. Using the *TC* constraint this gives $e_H = w_{H1} = 5/8$, $e_L = 1/2$, $w_{L0} = 1/64$ and $w_{L1} = 33/64$. With these values $\Pi_H = \Pi_L = 15/64$. Thus *both* types of $\mathcal{P}$ enjoy strictly higher profits than under separation. Pooling relaxes $\mathcal{A}$'s *IR* which binds in expectation. $\mathcal{P}_H$ can lower w_{H1} which increases Π_H^M relative to the separation case. The increased profit for $\mathcal{P}_H$ affects $\mathcal{P}_L$ via the *TC* constraint. $\mathcal{P}_L$ lowers both output-contingent wages to satisfy the *TC* constraint, which in turn increases Π_L^M to keep it in line with Π_H^M.

Intertemporal incentives

Holmström and Milgrom (1987) analyse the case of a relationship between $\mathcal{P}$ and $\mathcal{A}$ that extends over time. Some of the main insights can be gained in the following simple set-up.

There are two time periods – the first denoted *F* and the second denoted *S*. $\mathcal{A}$ chooses an effort in [0, 1] in both periods. Output can be either 1 or 0, and output draws are independent across the two periods. The first period effort is denoted e_F. The second period effort if output is 1 in the first period is e_{1S}, while the second period effort if output in the first period is 0 is e_{0S}. The probability that output is 1 is $\sqrt{e_F}$ in the first period, and $\sqrt{e_{iS}}$ (with $i \in \{0, 1\}$) in the second period.

$\mathcal{A}$ is paid at the end of the two periods, as a function of observed output in the two periods. The wage paid if output is $i \in \{0, 1\}$ in period *F* and $j \in \{0, 1\}$ in period *S* is denoted w_{ij}. Neither $\mathcal{P}$ nor $\mathcal{A}$ discounts the future. While $\mathcal{P}$ is risk-neutral, $\mathcal{A}$ is risk-averse with an *exponential* utility with a constant absolute risk-aversion coefficient equal to 1/2. His effort in the two periods is perfectly substitutable. Given a wage scheme w_{ij} and effort levels e_F and e_{iS} his expected utility is

$$\Pi^{\mathcal{A}} = -\sqrt{e_F}\left[\sqrt{e_{1S}}\exp\left\{-\frac{1}{2}(w_{11} - e_F - e_{1S})\right\} + (1-\sqrt{e_{1S}})\exp\left\{-\frac{1}{2}(w_{10} - e_F - e_{1S})\right\}\right] - (1-\sqrt{e_F})\left[\sqrt{e_{0S}}\exp\left\{-\frac{1}{2}(w_{01} - e_F - e_{0S})\right\} + (1-\sqrt{e_{0S}})\exp\left\{-\frac{1}{2}(w_{00} - e_F - e_{0S})\right\}\right]$$

while $\mathcal{P}$'s expected payoff is

$$\Pi^{\mathcal{P}} = \sqrt{e_F}\left[\sqrt{e_{1S}}(2 - w_{11}) + (1-\sqrt{e_{1S}})(1 - w_{10})\right] + (1-\sqrt{e_F})\left[\sqrt{e_{0S}}(1 - w_{01}) + (1-\sqrt{e_{0S}})(-w_{00})\right]$$

The optimal incentive scheme is found by maximizing $\Pi^{\mathcal{P}}$ subject to *IR* constraints imposing that $\Pi^{\mathcal{A}} \geq -1$ and $\Pi^{\mathcal{P}} \geq 0$ (these levels of reservation payoff can be taken to be a normalization for $\mathcal{P}$ and an assumption that $\mathcal{A}$ can earn a certain payoff of 0 elsewhere, yielding a utility level of –1), and subject to the *IC* constraints which now impose that e_F, e_{0S} and e_{1S} should jointly maximize $\Pi^{\mathcal{A}}$ given the incentive scheme w_{ij}.

The *IR* constraint is binding for $\mathcal{A}$ while it is not binding for $\mathcal{P}$. The *IC* constraint can be subsumed in the *first order* conditions obtained by differentiating $\Pi^{\mathcal{A}}$ with respect to e_F and e_{iS} and setting these equal to 0 which are sufficient for a maximum. This way to proceed is known in the literature as taking the *first-order approach*. In the more general case considered for instance by Holmström and Milgrom (1987) this is not viable. In the simple case considered here, the first-order approach works because we are assuming that the exponent of effort variables – 1/2 in this case – plus $\mathcal{A}$'s constant absolute risk-aversion coefficient – also 1/2 in this case – sum to 1. Even in single-period agency models, whether the first-order approach is valid or not is an intricate question first uncovered by Mirrlees (1975). Subsequent contributions on this topic can be found in Grossman and Hart (1983), Rogerson (1985) and Jewitt (1988). To characterize the optimal incentive scheme for the two-period problem it is useful to first consider the second period (*S*) sub-problem after output $i \in \{0, 1\}$ has been realized in the first period (*F*). These problems are obtained considering (continuation) payoffs for $\mathcal{A}$ and $\mathcal{P}$ given by the relevant square bracket term of $\Pi^{\mathcal{A}}$ and $\Pi^{\mathcal{P}}$ above, and with an *IR* constraint for $\mathcal{A}$ given by his utility level (contingent on output in *F*) in the solution to the two-period problem, after factoring out the common term $\exp\{e_F/2\}$.

If we use these binding *IR* constraints and the first-order *IC* constraints it can be seen that the difference $(w_{i1} - w_{i0}) > 0$ is independent of *i* – the second-period incentive premium $\Delta_S = (w_{i1} - w_{i0})$ does not depend on first-period output. Hence, if we use the first-order *IC* constraints it is also the case that $e_{0S} = e_{1S} = e_S \in (0, 1)$. $\mathcal{A}$'s *IR* constraints in each period *S* sub-problem determines w_{i0}.

The period *S* sub-problems can then be plugged into the two-period problem. Viewed from period *F* we can think of $\mathcal{P}$ as offering $\mathcal{A}$ two *certainty equivalent* wages c_i for each period *F* output. Notice that we can write $c_i = \tilde{w}_i - \pi_i$ where $\tilde{w}_i$ is the expected period *S* wage when the realized period *F* output is *i* and π_i is the associated risk-premium. Since $(w_{i1} - w_{i0}) = \Delta_S$ is independent of *i*, and $\mathcal{A}$'s utility exhibits constant absolute risk-aversion we then get $\pi_0 = \pi_1 = \pi$. Hence factoring out the common term $\exp\{\pi/2\}$ from $\mathcal{A}$ utility, the period *F* problem can be seen as having the same form as the two period *S* sub-problems with a different *IR* constraint for $\mathcal{A}$. Hence, as before, the difference $\Delta_F = (\tilde{w}_1 - \tilde{w}_0)$ does not depend on $\mathcal{A}$'s reservation utility and in fact $\Delta_F = \Delta_S = \Delta$. For the same reason $e_F = e_S = e$.

Using $\Delta_F = \Delta_S = \Delta$ and $e_F = e_S = e$ we then get that the optimal incentive scheme is *linear in output* in the sense that $w_{01} = w_{10} = w_{00} + \Delta$ and $w_{11} = w_{00} + 2\Delta$. Given w_{00}, the wage increases by a fixed amount Δ for each unit of realized output over the two periods.

In the simple model we have used here output is either 1 or 0. The linearity result holds in the same model (with an arbitrary finite number of periods) when there are *N* possible output realizations each period. In this case the incentive scheme is *linear in accounts* – in essence linear in a vector of variables that count the number of realizations of each possible output level.

Hellwig and Schmidt (2002) clarify that linearity in accounts need not imply *linearity in aggregate output*, and in fact some additional assumptions are needed for the latter to hold. They show that if $\mathcal{A}$ can destroy output unnoticed, and $\mathcal{P}$ only observes aggregate output at the end of the last period, then the (approximately) optimal incentive scheme is indeed linear in aggregate output.

Both Holmström and Milgrom (1987) and Hellwig and Schmidt (2002) are principally concerned with a *continuous-time* model in which $\mathcal{A}$ controls the drift of a (multi-dimensional) Brownian motion process that represents output. The continuous-time version of the problem yields elegant closed-form solutions that confirm the linearity result. Hellwig and Schmidt (2002) analyse in detail the status of the continuous-time model as the limit of discrete-time models.

The linearity of incentive schemes is of great interest in applications because of the prominence in practice of linear (or approximately linear) incentive schemes. In all known theoretical settings, linear optimal incentive schemes rely on exponential utility functions for both $\mathcal{A}$ and $\mathcal{P}$, whenever the latter is not risk-neutral. Stochastically independent periods also play a crucial role.

Finally, the tight linear characterizations of intertemporal incentive schemes also rely on $\mathcal{P}$'s ability to *commit* in advance to an incentive scheme, and on $\mathcal{A}$'s ability to commit not to *quit* before the end. The question of whether a full-commitment long-term contract can be implemented via a sequence of short-term contracts has been analysed in a general context by Malcomson and Spinnewyn (1988), Fudenberg, Holmström and Milgrom (1990) and Rey and Salanié (1990). A common thread of this literature is that $\mathcal{P}$'s ability to monitor $\mathcal{A}$'s savings decisions plays a key role in the possibility of short-term implementation of long-term contracts.

Recent developments

Since its inception the literature on agency problems and applications has grown dramatically, influencing many areas of economics ranging from development to finance. Agency theory has found a prominent place in many graduate and undergraduate programs in economics. Recent texts that provide a comprehensive treatment of the field include Salanié (2000), Laffont and Martimort (2002) and Bolton and Dewatripont (2005). Recent developments in the actual analytical framework relax some of the basic assumptions of the canonical model. Eliaz and Spiegler (2006) and O'Donoghue and Rabin (2005) focus on the underlying *behavioural* assumptions. The first paper tackles an environment in which agents may differ in their *cognitive abilities*, which generates dynamically inconsistent behaviour. The second paper is concerned with the effect of *present bias* in the agent's preferences on the optimal incentive scheme. In both cases the optimal incentive scheme becomes more realistically 'sensitive to detail' than in the standard case.
Besley and Ghatak (2005) focus on the case of *motivated agents* in the provision of a public good. Motivated agents do not always regard effort as a cost. This has important effects on incentive design, which in turn sheds light on the nature of non-profit organizations.

See Also

- contract theory
- incentive compatibility
- incomplete contracts
- mechanism design
- moral hazard

Bibliography

Anderlini, L. and Felli, L. 1998. Describability and agency problems. *European Economic Review* 42, 35–59.

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53, 941–73.

Besley, T. and Ghatak, M. 2005. Competition and incentives with motivated agents. *American Economic Review* 95, 616–36.

Bolton, P. and Dewatripont, M. 2005. *Contract Theory*. Cambridge, MA: MIT Press.

Eliaz, K. and Spiegler, R. 2006. Contracting with diversely naive agents. *Review of Economic Studies* 73, 689–714.

Fudenberg, D., Holmström, B. and Milgrom, P. 1990. Short-term contracts and long-term agency relationships. *Journal of Economic Theory* 51, 1–31.

Grossman, S. and Hart, O. 1983. An analysis of the principal–agent problem. *Econometrica* 51, 7–45.

Hellwig, M. and Schmidt, K. 2002. Discrete-time approximations of the Holmström–Milgrom Brownian-motion model of intertemporal incentive provision. *Econometrica* 70, 2225–64.

Holmström, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Holmström, B. and Milgrom, P. 1987. Aggregation and linearity in provision of intertemporal incentives. *Econometrica* 55, 303–328.

Holmström, B. and Milgrom, P. 1991. Multitask principal–agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics & Organization* 7, 24–52.

Holmström, B. and Milgrom, P. 1994. The firm as an incentive system. *American Economic Review* 84, 972–91.

Jewitt, I. 1988. Justifying the first-order approach to principal–agent problems. *Econometrica* 56, 1177–1190.

Laffont, J.-J. and Martimort, D. 2002. *The Theory of Incentives*. Princeton: Princeton University Press.

Malcomson, J. and Spinnewyn, F. 1988. The multiperiod principal–agent problem. *Review of Economic Studies* 55, 391–407.

Marschak, J. 1955. Elements for a theory of teams. *Management Science* 1, 127–37.

Maskin, E. and Tirole, J. 1990. The principal–agent relationship with an informed principal: the case of private values. *Econometrica* 58, 379–409.

Maskin, E. and Tirole, J. 1992. The principal-agent relationship with an informed principal II: common values. *Econometrica* 60, 1–42.

Mirrlees, J. 1975. The theory of moral hazard and unobservable behavior: part I. Mimeo, Nuffield College, Oxford University. Published in *Review of Economic Studies* 66 (1999), 3–21.

Myerson, R. 1983. Mechanism design by an informed principal. *Econometrica* 51, 1767–98.

O’Donoghue, T. and Rabin, M. 2005. Incentives and self control. Mimeo, University of California, Berkeley.

Pauly, M. 1968. The economics of moral hazard. *American Economic Review* 58, 531–37.

Rey, P. and Salanié, B. 1990. Long-term, short-term and renegotiation: on the value of commitment in contracting. *Econometrica* 58, 597–619.

Rogerson, W. 1985. The first-order approach to principal–agent problems. *Econometrica* 53, 1357–67.

Salanié, B. 2000. *The Economics of Contracts*. Cambridge, MA: MIT Press.

Shavell, S. 1979. On moral hazard and insurance. *Quarterly Journal of Economics* 93, 541–62.

How to cite this article

Anderlini, Luca and Leonardo Felli. "agency problems." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000227> doi:10.1057/9780230226203.0015

The New Palgrave Dictionary of Economics Online

agent-based models

Scott E. Page
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Agent-based models consist of purposeful agents who interact in space and time and whose micro-level interactions create emergent patterns. Agent-based models consist not of real people but of computational objects that interact according to rules. The four primary features of agent-based models – learning, networks, externalities, and heterogeneity – though previously far from neoclassical economics, have become part of the mainstream. Agent-based models allow us to consider richer environments that include these features with greater fidelity than do existing techniques. They occupy a middle ground between stark, dry rigorous mathematics and loose, possibly inconsistent, descriptive accounts.

Keywords

agent-based models; behavioural game theory; central limit theorem; complexity; Conway's game of life; economic complexity; emergence; equilibrium; interaction structures; learning and information aggregation in networks; mathematics and economics; prisoner's dilemma; rule-based behaviour

Article

An economy consists of agents who interact in space and time and who act purposefully choosing their actions, their strategies, and their locations with some objective in mind. This purposefulness implies that they respond to incentives and information in predictable ways at the individual level, but it makes for complex aggregation. The aggregation of micro-level behaviours and interactions can create trading patterns, price bubbles and business cycles that were not built into the economy. They emerge from the bottom up. It is these patterns and regularities which economists seek to understand, explain, and predict, and which policymakers try to alter for the better.
Agent-based models of economies, like real economies, consist of computational objects that interact according to rules. Agent-based modelling allows us to consider richer environments with greater

fidelity than do existing techniques (Tesfatsion, 1997). This increased fidelity results from the inductive nature of the modelling enterprise. When constructing an agent-based model, we are constrained only by our imagination and interest. In contrast, when constructing a mathematical model, we must always be concerned with analytic tractability. This constrains our endeavours. The set of models that one believes to be tractable is small when compared with the set of models worth exploring. Thus, the flexibility and potential for realism enlarge the set of questions economists can explore (Anderson, Arrow and Pines, 1988; Arthur, Durlauf and Lane, 1997).

By freeing us from considerations of provability, agent-based models focus us on those aspects of the world that we believe most relevant. We can then encode the relevant assumptions in a computer program and allow the logical implications to iterate. Owing to the inductive nature of the enterprise, we do not know results a priori. Some agent models produce a chaotic mess and their assumptions need to be rethought. But often agent-based models produce interesting results, and these results can then be supplemented with analytic ones. We can much more easily prove a result when we know the answer. Thus, at a minimum, agent-based models can be thought of as a powerful engine for generating insights. Many mathematical theorists even admit that they use agent-based models for this purpose. But agent-based models can do far more.

The benefits of agent-based models

Proponents claim that agent-based models will advance the discipline because they can include more realistic assumptions about behaviour, structure and timing – that they have greater resonance. These claims ring true. Agent-based models look and feel more like real economies. All else equal, more realism improves models. The benefits of greater fidelity and realism in modelling behaviour can also be seen in the contributions of behavioural economics (Camerer, 2003). Agent-based models go further than behavioural models by also taking a realistic approach to modelling interaction structures and the timing of events (Kirman, 1997).

The four primary features of agent-based models – learning, networks, externalities and heterogeneity – which once lied outside of the mainstream have all received growing interest from economists over the past two decades. That said, despite what their advocates claim, agent-based models are not likely to lead to a complete rethinking of economics or of social science. No matter how they are implemented, be it mathematically or computationally, economic models will always have consumers and producers. Consumers will still choose bundles of goods with an eye towards getting high utility. Producers will still try to buy low and sell high. And markets, most of the time, will come close to efficiently allocating goods and services.

As Holland and Miller (1991) stated early on, agent-based models occupy a middle ground between stark, dry rigorous mathematics and loose, possibly inconsistent, descriptive accounts. We should not expect that middle ground to differ in kind from the two end points. We might, though, expect a better, more comprehensive economics. Thus, the real contribution of agent-based models will more likely be to push theory into places it has heretofore ignored or avoided. Thus, we should not expect a revolution based on this new methodology, but we should expect absorption. Like experimental economics, agent-based modelling should become one more row of street lights for economists to stand underneath (de Marchi, 2005).

When first introduced, agent-based models were somewhat controversial. This was caused by claims that they combined the precision of Samuelson with the scope and breadth of Keynes. Critics responded by dismissing agent-based models as simulations, as mere examples or sets of examples, to be contrasted with the general truths revealed by mathematics-based theory. Both sides were partly correct. Agent-based models are logically consistent. Agent behaviour is encoded in computer programs and the model proceeds according to the rules embedded in those programs. An agent-based model can be thought of as an enormous recursive equation being cranked over and over. What could be more logical and rigorous than that? Of course, codes can contain errors, as can computer software, but this is hardly a damning critique. The modern practice of programming and testing minimizes those errors and, fortuitously, most coding errors become apparent in the implementation stage.

I noted above that agent-based models can include diverse agents, geographic and social space, externalities, and learning. Many agent-based models include *all* of these features. These models can generate equilibria, emergent patterns and structure, and complexity. All of these can even occur in the same model but on different dimensions, just as in the real economy. Prices may attain something close to an equilibrium, information and trade networks may form patterns, and the inventory levels of suppliers may be complex and unpredictable.

The output flexibility of agent-based models leads some to jump to the inaccurate and unfortunate conclusion that agent-based models preclude equilibrium analysis. True, agent-based models naturally allow for dynamics, but this does not mean that they cannot attain equilibria. These equilibria are not assumed by *generated* (Epstein, 2003). The generative claim that ‘if you didn't grow it, you didn't show it’ should be ignored at our peril. Proving that an equilibrium exists and showing that it can be attained and maintained are separate findings. But not all agent-based models generate the equilibria predicted by mathematics. They fail because attaining equilibrium often requires slow learning rates and lots of agents. Sometimes, though, they fail because the mathematics contains errors (Page and Tassier, 2004). Attaining equilibria to complement mathematical analyses (Judd, 1997) is not the reason to use agent-based models. They are better suited to exploring those parts of the economy that are complex or on the boundary between complexity and equilibrium. Even critics of agent-based modelling admit the appeal of exploring complexity, but they question what we learn from individual models. Mathematical theorems *prove* results for entire classes of functions. Arrow, Debreu and McKenzie proved theorems for any convex preferences, not just for preferences derived from Cobb–Douglas utility functions. Agent-based models, at least for now, assume particular functional forms. Mathematics therefore gives us the kind of general results on which a science has traditionally been built. Agent-based models do not. This is only partly true. These critics are less than honest about the current state of our knowledge (Leombruni and Richiardi, 2005). Although mathematical theorems are general and agent-based models are particular, that is not the whole story. In economics, general results are few and far between. Many papers (*a*) assume specific functional forms rendering them examples not general truths, or (*b*) consider restrictive classes of functional forms such as quasi-linear preferences, or (*c*) rely on dubious assumptions such as the monotone likelihood ratio property or independent signals.

Imagine the space of all possible economic environments as a room. Far too many theorems create small boxes in the corner of that room. Those boxes may not contain many real economies. Agent-based models, though only points (of light perhaps), can be scattered throughout the room wherever we like. We may need boxes to build a science, but a room full of light is better than a stack of boxes in the corner. And ideally, we can use the lights to construct boxes that fill the room.

Several excellent surveys describe the contributions of agent-based modelling as well as the enormous potential of this new methodology (see Tesfatsion and Judd, 2006, for surveys of several fields). This affords me the opportunity to use these pages to explore ideas related to agent-based models. I take three ideas that are fundamental to agent-based models and at the same time not familiar to most economists: *people as objects*, *complexity*, and *emergence*. In discussing these ideas, I explain why each is important to the study of economics.

Economic actors as objects

As I mentioned, agent-based models contain agents who follow rules. In the language of computer science, these agents are *objects* that exhibit rule-based behaviour. These objects can represent people, families, or firms. In constructing an object, the modeller must consider (*a*) the nature of the rules, (*b*) how the rules interact, and (*c*) the determinants of agent activation (Kirman, 1997). The behavioural rules can vary in their sophistication. The economic agents can follow *simple fixed rules* that are naive and routine. In a spatial Prisoner's Dilemma game, agents can play a strategy that always cooperates, or they can be extremely sophisticated. Incidentally, if agents play an equilibrium strategy in a game, they follow a fixed rule as well, but that simple fixed rule may take some effort to find.

It is in the region between primitive rule following and full cognitive closure where we might expect to find real people and firms. An assumption of naive rules understates human abilities and an assumption of full rationality overstates them, at least in non-trivial contexts. Human behaviour is more dynamic. We adapt and change our behaviours according to what works well. Sometimes we follow higher-order rules that allow us to learn to change our behavioural rules. But this learning algorithm – be it fictitious play, Hebbian learning or experience-weighted learning (Camerer, 2003) – is nothing more than a fixed rule. Sometimes we even apply learning rules on top of learning rules: we learn how to learn. These are all types of individual learning. We also learn socially. We mimic more successful people. Social learning is also rule-based. We have a rule for how we learn from others. Individual and social learning create different dynamics (Vriend, 2000). Social learning supports less diversity than does individual learning.

Agent-based modellers must also make explicit assumptions about the intelligence and adaptability of agents. Regardless though of how sophisticated or adaptive these agents may be, they still follow rules embedded in the computer code. So the agent-based models can be thought of as the recursive accumulation of those rules. Lest this seem unrealistic, economies can also be thought of as accumulated rules. People and firms follow rules, those rules may change, but, nevertheless, the total output of an economy and its allocation are determined by the accumulation of those rules, as are prices.

The conception of agents as objects requires explicit rules for how objects interact with one another. The agents must be situated in an interaction structure (Epstein and Axtell, 1996). These interaction structures can be represented in space or in networks that encode geographic, sociological, or feature-based differences (Riolo, Axelrod and Cohen, 2001). Feature-based, social and geographic spaces are more similar than might be thought. Two agents with similar features or social standing are more likely to interact than two agents with diverse features or social standings, just as two agents at nearby locations are more likely to interact than two agents who are far apart.

Finally, the idea of agents as objects demands explicit consideration of agent activation. In what order

do the agents get called to take their action? Do they get called simultaneously or sequentially? If the former, how are conflicts settled – what if two agents choose the same trading partner? If the latter, is that order independent of the agents' incentives to update, or do the agents who benefit most by updating their behaviour move first (Page, 1997)? The nature of results can often hinge on how timing is implemented and timing interacts with other features (Nowak, Bonhoeffer and May, 1994).

The interactions between timing, interaction structures, and rules can alter the performance of a model. These interaction effects support the idea of richer model. This last observation leads into what I call the *irony of robustness*. Agent-based models are considered to be less robust because 'you can get any result' by changing a few assumptions (Miller, 1998). Seemingly minor changes in the timing of events or the network structure can have large effects on the outcomes of some models. Herein lies the irony. Results that depend crucially on these assumptions should not be seen as a weakness of agent-based models, as evidence that they have too many moving parts. Instead, the lack of robustness of these models can be seen as a critique of the starker mathematical models. The starker models ignore the very features of the economy that have been shown in the agent-based model to matter (Andreoni and Miller, 1995). As Mason and Wellman (2005) point out in their survey of the market design literature, many mathematical theorems lack detail about how, where, and when trade takes place. We should therefore think of theorems that exclude assumptions about time and place as incomplete. Decades of experiments with human subjects confirm this insight. Minor changes in how we run experiments can have enormous effects on outcomes.

Emergence

Modellers implement agent-based models in computational platforms that permit graphical representations of outcomes. This has had profound implications (both good and bad) for the growth and direction of the methodology. The graphical interfaces have revealed what are called 'emergent phenomena': meso- and macro-level phenomena that arise from the micro-level interactions of agents. Agent-based models produce emergent patterns and structures. Emergence was thought by some to be a clever bit of marketing but logically vacuous. And any initial tests for emergent phenomena were based on ocular statistics (Bankes, 2002). Look! Emergence! But since the mid-1990s emergence has become a scientific concept with several definitions.

To understand emergence, we must first recognize that a structure or entity can have multiple levels of explanation. A crowd's movements can be explained as if the crowd were a single entity or as the accumulation of individuals' movements. If a entity's actions can be explained equally accurately at a higher level – if the individuals really move as a crowd – then it is emergent. One of the simplest examples of emergence arises in Conway's Game of Life (Poundstone, 1985). Fixed automata rules on a lattice produce gliders. These gliders move diagonally across the space. The movement of the gliders can be explained by an appeal to the micro-level rules of the automata, but it can be more succinctly explained at the level of glider. Hence, the glider can be said to emerge.

In economies and societies, many things emerge: prices, cities, trade patterns, information networks, and cultural norms, to name just a few (Tesfatsion and Judd, 2006). These features of our world matter for economies. Cities matter. Trade networks matter. Culture matters. Social science needs ways of understanding how these things come to be as well as how they influence the performance of economic

and political systems. Agent-based models offer a route to those understandings that complements our mathematical approaches.

Complexity

Agent-based models can generate complexity and allow us to explore its causes, thereby interweaving the methodology of agent-based models with the theoretical idea of complexity. The four main features of agent-based models are diverse agents, situated in an interaction structure, whose actions create interactive effects (externalities), which adapt, evolve or learn each contribute to the level of complexity a model produces (Axelrod and Cohen, 2000). These features can be thought of as choice variables. We can imagine a knob for each feature – a diversity knob, a connectedness knob, an externality knob, and a learning rate knob. The agents can be nearly homogeneous or very diverse. The space can be sparsely connected or highly connected. The interactions can be few and small or numerous and large, and the agents can adapt not at all or instantaneously. By turning these knobs, we can create complexity.

If we set all of the knobs at low levels, the resulting model usually settles into an equilibrium or a simple pattern. Wolfram's amazing cellular automata models and the Game of Life notwithstanding, most models with identical agents loosely connected with mild externalities and little learning do not produce much complexity. They tend to settle into equilibria or cycles. Turning up individual knobs creates complexity: complicated patterns and elaborate interacting emergent structures, such as trading patterns. As we turn the knobs further one of two things happens: equilibrium or chaos.

Often, by turning up the connectedness knob, we lead the system back towards equilibrium. When every agent connects to every other agent the environment becomes simpler for reasons explained by the central limit theorem. Diversity, externalities, and learning all get averaged out and the system stabilizes. In contrast, in many of these same models turning up the externality knob creates to chaos. If agents' actions have large external effects on other agents, the system does not settle down, but spins out of control. Complexity then can lie either between order and order or between order and chaos.

The existence of complexity depends upon having the right level of *interplay* between the agents. Interplay is a measure of how often and how much the behaviour of other agents influences the behaviour of any individual agent. The four knobs all adjust the level of interplay. As agents become more diverse, they take more extreme actions, increasing interplay. As agents become more connected and more interactive, interplay also increases. More agents have larger effects on each individual agent. Finally, the more agents change their behaviour, the more they cause other agents to change. This too increases interplay.

Social systems differ from physical systems in that these knobs are not fixed. In human systems, the agents can tune these knobs. They can choose to be more or less diverse, connected, interdependent, or reactive. The idea of adjustable levels of interplay raises the question of whether we should expect social systems to generate equilibrium, complexity or chaos. Changes in the level of interplay can transport a system out of equilibrium and into complexity. Alternatively, if agents want order, they can have it by slowing down or becoming less interdependent. Whether equilibrium or whether complexity may be a choice. We might assume that agents seek out equilibria, that they want stability. But agents may also desire complexity, for with complexity comes opportunity. Probably no one wants chaos though, and the ability to dial the knobs back to prevent it is invaluable. Thus, the fact that some parts of the economy appear more complex than others may be predictable based upon the incentives for ramping up or

dampening levels of interplay between the agents.

The future of modelling

To summarize, agent-based modelling offers a new methodology, a new tool for economists and social scientists. One cannot resist the temptation to talk about how existing research presents just the tip of the iceberg, that we have just begun to scratch the surface, but these metaphors fail. Some icebergs should remain sunk and some surfaces should remain unmarred. The case for agent-based modelling cannot be simply one of opportunity – we have a new tool, let's build something with it. We need reasons to believe that the submerged part of the iceberg merits exploring.

Resonance provides one strong reason. Agent-based models contain people and firms embedded in interaction structures. These people and firms have conceptualizations of problems and situations. At times, they adhere to routines. At times, they experiment.

And at times, they learn from those who are most successful. Real people and real firms behave similarly. These models also produce emergent structures. And, they sometimes result in complexity and sometimes settle into equilibria. Herein lies a second reason for agent-based models. We should not think of the economy as either having attained equilibrium or to be exhibiting complex dynamics, for it has both properties simultaneously. Parts of the economy equilibrate. Shares of oil production across OPEC members resemble sequences of equilibria that respond to shocks. Other parts do not. The monthly, weekly, daily, hourly, and second-by-second fluctuations of the stock market create complex patterns (Palmer et al., 1994). Agent-based models allow us to explore this complexity, a large and important part of the iceberg.

See Also

- behavioural game theory
- learning and information aggregation in networks
- mathematics and economics

I would like to thank Ken Kollman and Rick Riolo for comments on earlier drafts.

Bibliography

Anderson, P., Arrow, K. and Pines, D., eds. 1988. *The Economy as an Evolving Complex System*. Redwood City, CA: Addison Wesley.

Andreoni, J. and Miller, J. 1995. Auctions with adaptive artificial agents. *Games and Economic Behavior* 10, 39–64.

Arthur, B., Durlauf, S. and Lane, D. 1997. *The Economy as an Evolving Complex System II*. Redding, MA: Addison Wesley.

Axelrod, R. and Cohen, M. 2000. *Harnessing Complexity Organizational Implications of a Scientific Frontier*. New York: Free Press.

Bankes, S. 2002. Agent-based modeling: a revolution? *Proceedings of the National Academy of Sciences* 99, 7199–200.

Camerer, C. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

de Marchi, S. 2005. *Computational and Mathematical Modeling in the Social Sciences*. Cambridge: Cambridge University Press.

Epstein, J. 2003. *Generative Social Science*. Princeton: Princeton University Press.

Epstein, J. and Axtell, R. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, MA: MIT Press.

Holland, J. and Miller, J. 1991. Artificial agents in economic theory. *American Economic Review papers and proceedings* 81, 365–70.

Judd, K. 1997. Computational economics and economic theory: complements or substitutes? *Journal of Economic Dynamics and Control* 21, 907–42.

Judd, K. and Tesfatsion, L., eds. 2005. *Handbook of Computational Economics Volume 2: Agent-Based Computational Economics*. Amsterdam: North-Holland.

Kirman, A. 1997. The economy as an interactive system. In *The Economy as a Complex Evolving System II*, ed. W. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison Wesley.

Leombruni, R. and Richiardi, M. 2005. Why are economists sceptical about agent-based simulations? *Physica A* 355, 103–9.

Mason, J. and Wellman, M. 2005. Automated markets and trading agents. In *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*, ed. L. Tesfatsion and K. Judd. Amsterdam: North-Holland.

Miller, J. 1998. Active nonlinear tests ANTs of complex simulations models. *Management Science* 44, 820–30.

Nowak, M., Bonhoeffer, S. and May, R. 1994. Spatial games and the maintenance of cooperation. *Proceedings of the National Academy of Sciences* 91, 4877–81.

Page, S. 1997. On incentives and updating in agent-based models. *Computational Economics* 10, 67–87.

Page, S. and Tassier, T. 2004. On the existence and stability of Groves Ledyard equilibria. *Journal of Public Economic Theory* 62, 311–35.

Palmer, R., Arthur, W., Holland, J., Lebaron, B. and Tayler, P. 1994. Artificial economic life: a simple model of a stock market. *Physica D* 75, 264–74.

Poundstone, W. 1985. *The Recursive Universe*. Chicago: Contemporary Books.

Riolo, R., Axelrod, R. and Cohen, M. 2001. Evolution of cooperation without reciprocity. *Nature* 414, 441–3.

Tesfatsion, L. 1997. How economists can get A-Life. In *The Economy as a Complex Evolving System II*, ed. W. Arthur, S. Durlauf and D. Lane. Redding, MA: Addison Wesley.

Tesfatsion, L. and Judd, K., eds. 2006. *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*. Amsterdam: North-Holland.

Vriend, N. 2000. An illustration of the essential difference between individual and social learning, and its consequences for computational analyses. *Journal of Economic Dynamics and Control* 24, 1–19.

How to cite this article

Page, Scott E. "agent-based models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000218> doi:10.1057/9780230226203.0016

The New Palgrave Dictionary of Economics Online

aggregate demand theory

H. Sonnenschein
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand theory; consumer demand functions; demand theory; individual demand functions; market demand functions

Article

Aggregate demand theory investigates the properties of market demand functions. These functions are obtained by summing the preference maximizing actions of individual agents. The study of aggregate demand theory is primarily motivated by the fact that market demand functions, rather than individual demand functions, are the data of economic analysis. In general, market demand functions do not inherit the structure which is imposed on individual demand functions by the utility hypothesis. Such structure, when present, enables us to obtain stronger predictions from available data.

Here we focus on three aspects of market demand functions. The first is that in certain special cases, market demand functions can be shown to satisfy the classical restrictions that characterize individual demand functions. The second is that aside from these very special cases, the economy cannot be expected to behave as an ‘idealized’ or ‘representative’ consumer. Finally, we verify that when the economy is modelled as a continuum of infinitesimally sized agents market demand functions may in some respects be better behaved than individual demand functions. For an elaboration of the material through Example 3 see Shafer and Sonnenschein (1982).

1. This section presents the notation and briefly reviews the properties of individual demand functions. There are *n* consumers and *l* commodities. The consumption set of each consumer is R^l_+. The preferences of a consumer are described by a weak ordering $\succsim$ of R^l_+. If $x \succsim y$ we say ‘*x* is at least as good as *y*’; if $x \succsim y$ and not $y \succsim x$, then we write $x \succ y$ and say ‘*x* is preferred to *y*’; if $x \succsim y$ and $y \succsim x$, we write *x*~*y* and say ‘*x* is indifferent to *y*’. The preference relation $\succsim$ is continuous if $\{(x, y): x \succsim y\}$ is closed; $\succsim$ is locally non-satiated if for each $x \in R^l_+$ and every $\eta > 0$ there exists a *y* such that $y \succ x$ and $\|x - y\| < \eta$; $\succsim$ is strictly convex if $x \succsim y$, $x \neq y$ and $0 < \alpha < 1$ implies that $\alpha x + (1-\alpha) y \succ y$; $\succsim$ is representable if there exists a ‘utility function’ $U: R^l_+ \to R$ such that $x \succsim y$ if and only if $u(x) \geq u(y)$; $\succsim$ is homothetic if it is representable by a utility function which is homogeneous of degree 1. It is assumed throughout that preference relations for all consumers are continuous, locally non-satiated and strictly convex. A continuous function $f: R^l_{++} \times R_+ \to R^l_+$ is a candidate consumer demand function if it satisfies (Budget balance) $p \cdot f(p, I) = I$ for all $(p, I) \in R^l_{++} \times R_+$ and (Homogeneity) $f(\lambda p, \lambda I) = f(p, I)$ for all $\lambda > 0$ and $(p, I) \in R^l_{++} \times R_+$. At prices *p* and income *I*, $f(p, I)$ denotes the commodity bundle purchased. If there exists a preference relation $\succsim$ such that for each $(p, I) \in R^l_{++} \times R_+$, $f(p, I)$ is the $\succsim$ maximal element in the set $\{x: p \cdot x \leq I\}$, then *f* is a consumer demand function.

Let *f* be a differentiable candidate consumer demand function. The Slutsky matrix associated with *f* is an $l \times l$ matrix denoted by $\Sigma(p, I)$ whose $(h, k)^{th}$ term is defined by

$$\sigma_{hk}(p, I) = \frac{\partial f_h}{\partial p_k}(p, I) + f_k(p, I) \cdot \frac{\partial f_h}{\partial I}(p, I).$$

The classical theorems of demand theory state that, if f is a consumer demand fucntion, then for all $(p,\bullet I)$ $\Sigma(p,\bullet I)$ is symmetric and negative semi-definite. The integrability theorem establishes the converse (see Hurwicz and Uzawa, 1971).

Let $\Delta^{n-1} = \{(x_1, x_2, \ldots, x_n) | x_i \geq 0$ for all i and $\Sigma x_i = 1\}$. Given prices p and income I, the distribution of income among consumers is defined by a mapping $\delta: R^l_{++} \times R_+ \to \Delta^{n-1}$. Thus $\delta^i(p, I)I$ is the ith individual's income when prices are p and income is I. A candidate demand function F is a market demand function relative to the distribution of income mapping δ if there exists n consumer demand functions $f^1,\ldots, f^n$ such that $F(p, I)= \Sigma_i f^i [p, \delta^i (p, I)I]$ holds for all $(p, I) \in R^l_{++} \times R_+$. If $(f^1,\ldots,f^n)$ are individual demand functions and if for all $\delta, \bar{\delta} \in \Delta^{n-1}$, $\Sigma_i f^i(p, \delta^i I) = \Sigma_i f^i(p, \bar{\delta}^i I)$, then market demand is independent of the distribution of income.

2. This section considers the conditions under which market demand functions belong to the class generated by a single consumer. The following classic result, due to Antonelli (1886) and later independently discovered by Gorman (1953) and Nataf (1953), gives necessary and sufficient conditions for a market demand function to be both independent of the distribution of income and generated by a preference relation.

Theorem 1: (Antonelli). Market demand is independent of the distribution of income and is preference generated if and only if there is a homothetic preference relation $\succsim$ such that each consumer demand function f^i is derived from $\succsim$. In this case, market demand is also generated by f^i.

Examples 1 and 2 demonstrate that if either the condition that preferences are homothetic or the condition that preferences of all consumers are identical is dropped, then market demand may depend on the distribution of income (for elaboration of, these examples, and of Example 3, see Shafer and Sonnenschein 1982).

Example 1: Let two consumers have identical preferences on R^2_+ that are represented by $U(x, y)=xy+y$ and let prices be (1, 1). If the distribution of income is $I_1=1$, $I_2=1$, then aggregate demand for x and y is 0 and 2 respectively. If the distribution of income is $I_1=2$, $I_2=0$, then aggregate demand for x and y is $\frac{1}{2}$ and $1\frac{1}{2}$ respectively.

Example 2: Let two consumers have homothetic preferences on R^2_+ represented by $U_1(x, y)=x$ and $U_2(x, y)=y$. Then market demand depends completely on the distribution of income. If the income share of each consumer is fixed [that is, $\delta(p, I)$ is a constant vector $(\delta^1,\ldots, \delta^n)$ for all (p, I)], then homotheticity of each individual preference relation is sufficient for market demand to be utility generated. This result is due to Eisenberg (1961).

Theorem 2: (Eisenberg). If the preferences of each agent can be represented by a homogeneous of degree one utility function U^i on R^l_+, and if income shares are fixed at $(\delta^1,\ldots, \delta^n) \in \Delta^{n-1}$, then market demand is generated by the homogeneous of degree one utility function U

$$U(x) = \max \prod_{i=1}^{n} \left[U^i\left(x^i\right)\right]^{\delta^i} \quad \text{s.t.} \quad \sum_i x^i = x.$$

Under the hypothesis of Theorem 2 market demand is determined by maximizing a social welfare function that gives each individual's preferences, a weight equal to his share of total income. The following example indicates that a fixed distribution of income, but no restrictions on agents' preferences, is not sufficient to ensure that market demand is utility generated.

Example 3: (Hicks, 1957). There are two consumers who share market income equally. Market budgets for two different price ratios are indicated with dotted lines. The choices of the first individual are indicated by a cross and those of the second by a circle. Market demand at the steeper budget is denoted by D while demand at the flatter budget is denoted by D'. The choice of each individual is consistent with utility maximization; however, since D is chosen in the aggregate when D' is available and since D' is chosen when D is available, market demand is not utility generated (Figure 1).

Figure 1

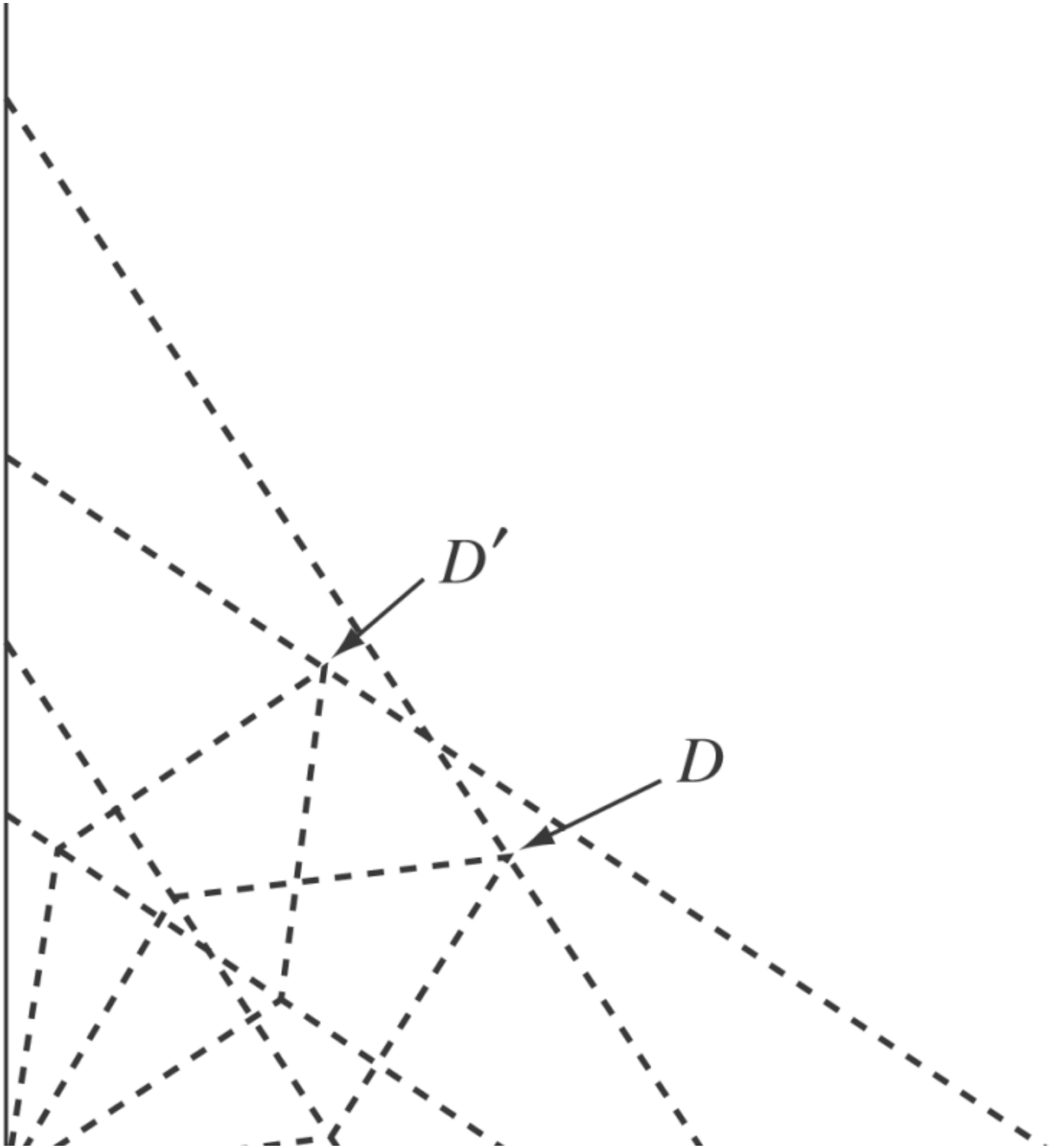
D′
D

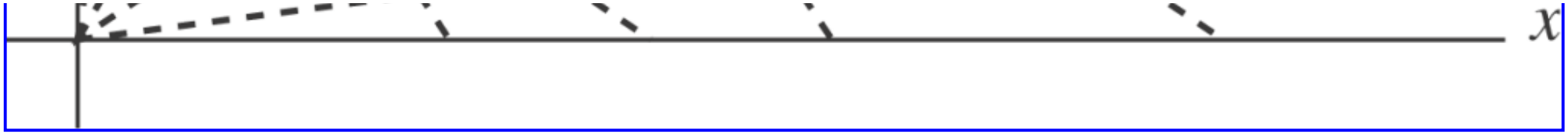

Theorems 1 and 2 referred to situations in which the distribution of income was determined exogeneously. In a much referenced paper, Samuelson (1956) presented a theorem in which the distribution of income is determined as a solution to a maximization problem. Specifically, it is assumed that for every price–income combination, the government distributes income so as to maximize a Bergsonian social welfare function: let δ denote the distribution of income function determined by this process. Samuelson's theorem asserts that under these conditions, market demand relative to δ is utility generated. Proofs of the result may also be found in Chipman and Moore (1979), and Dow and Sonnenschein (1983).

Theorem 3: Suppose that f^i is generated by U^i for i=1,…,n. If there exists a Bergsonian social welfare function $W(U^1,\ldots,U^n)$ that is increasing in all its arguments and such that for all $(p, I) \in R^l_{++} \times R_+$,

$$\delta(p, I) \in \underset{(d^1,\ldots,d^n)\in\Delta^{n-1}}{\arg\max}\; W\left\{U^1\left[f^1\left(p, d^1 I\right)\right], \ldots, U^n\left[f^n\left(p, d^n I\right)\right]\right\},$$

then aggregate demand $\Sigma_i f^i [p, \delta^i (p, I)I]$ is generated by the utility function

$$U(x) = \max\; W\left[U^1\left(x^1\right), \ldots, U^n\left(x^n\right)\right] \quad \text{s.t.} \quad \sum_i x^i = x.$$

3. Theorems 1–3 identify sets of assumptions under which market demand functions belong to the same class as consumer demand functions. Theorem 4 indicates that in the absence of these assumptions, none of the classical restrictions holds for market demand functions. In particular any values of demand and its derivatives that are consistent with Homogeneity and Budget balance are possible.

Theorem 4: (Sonnenschein). Let F be an arbitrary C^1 candidate demand function for l commodities and let $n \geq l$. Then, for any $(p, I) \in R^l_{++} \times R_+$, there exists a market demand function generated by n consumers with demand functions $f^1,\bullet\ldots,\bullet f^n$ such that

$$F(p, I) = \sum_{i=1}^{n} f^i\left(p, \frac{I}{n}\right)$$

and

$$\frac{\partial F_k}{\partial P_j}(p, I) = \sum \frac{\partial f^i_k}{\partial p_j}\left(p, \frac{I}{n}\right), \quad \text{for each } k, \; j.$$

More general results of this nature exist for market excess demand functions; see Sonnenschein (1973a), Debreu (1974) and Shafer and Sonnenschein (1982, section 4).

4. In this section an example of an economy with a continuum of infinitesimally sized agents is presented in which market demand is continuous despite the fact that individual demand functions are discontinuous: market demand is better behaved than individual demand. The point that is made here is quite general and is of importance in establishing the existence of competitive equilibrium without need for the assumption that preferences are convex; see Debreu (1982, section 4).

Example 4: There are two commodities *x* and *y* and the preferences of a consumer of type *a* are represented by the utility function $U(x, y, a) = x^2 + a^2 \cdot y^2$. The income of each consumer is fixed at unity and the consumption set of each consumer if R^2_+. The price of commodity *y* in terms of the numeraire commodity *x* is denoted by *p*. The distribution of agent types is specified by defining the following density function *g*, over the domain of *a*:

$$g(a) = \begin{cases} 2 & \text{if } a \in \left[\frac{1}{4}, \frac{3}{4}\right] \\ 0 & \text{otherwise} \end{cases}.$$

Strict convexity of preferences is violated for each *a*, and consequently, the demand function of each consumer type is not single valued. The demand function for *y* as a function of *p* is given by

$$f^a(p) = \begin{cases} \frac{1}{p} & \text{if } p < a \\ 0 & \text{if } p > a \\ \left[\frac{1}{a}, 0\right] & \text{if } p = a. \end{cases}$$

The graph of f^a is drawn in Figure 2.

Figure 2

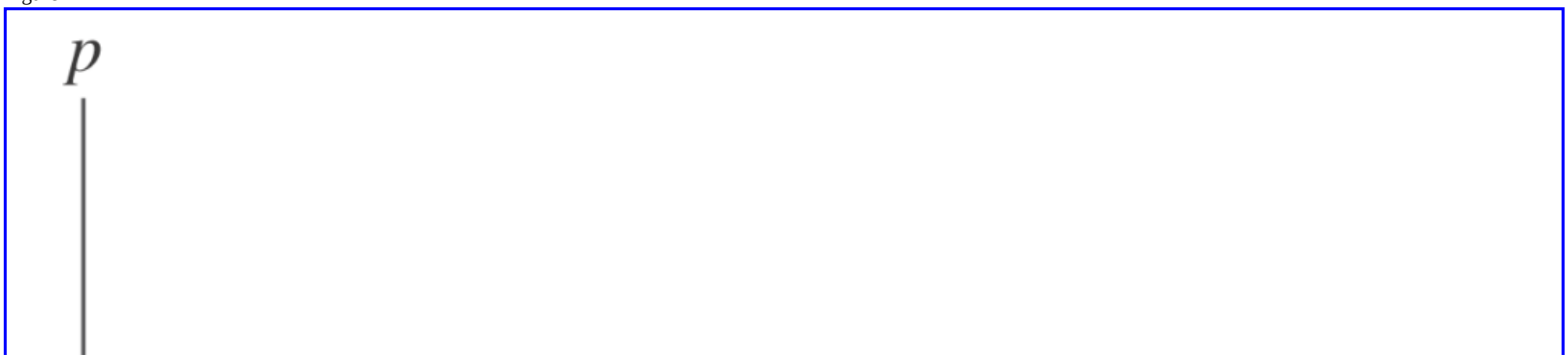

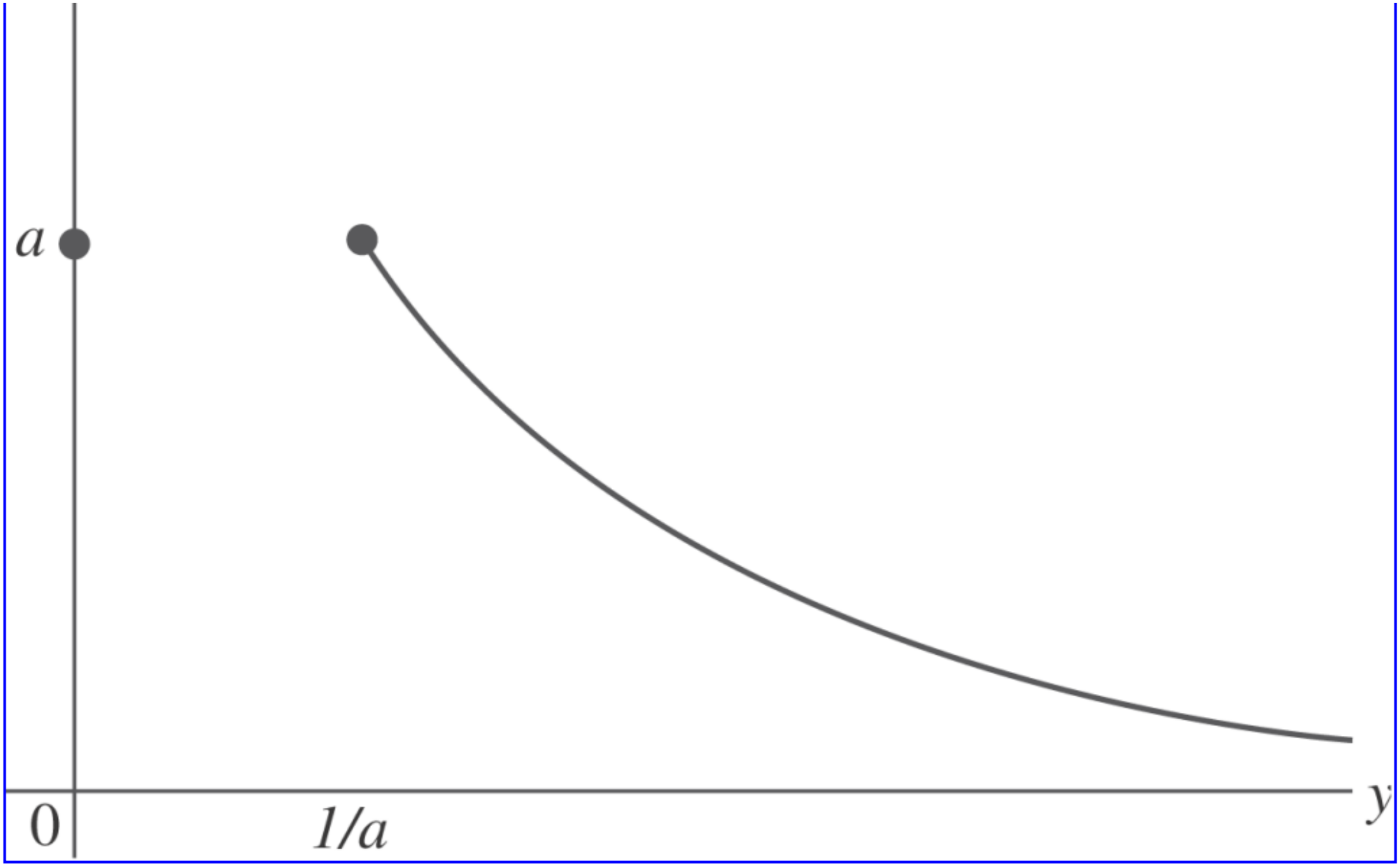

The multi-valued function f^a is not well-behaved in the sense that it jumps at a. Let $F(p)$ denote market demand at price p. By definition

$$F(p) = 2\int_{a=1/4}^{a=3/4} f^a(p)\,\mathrm{d}a = 2\int_{a=1/4}^{a=p} (0)\,\mathrm{d}a + 2\int_{a=p}^{a=3/4} \frac{1}{p}\,\mathrm{d}a = \frac{3}{2p} - 2.$$

Thus, market demand is single-valued and differentiable in the entire domain of p, despite the fact that these properties do not hold for any given a. One way to understand the result

is to observe that for each p, the relative mass of consumers whose demand is discontinuous at p is zero. This observation also illustrates the importance of the assumption that each agent is a 'small' part of the market and that preferences are dispersed. The result would not hold if the density function was assumed to be

$$h(a) = \begin{cases} 1 & \text{if } a \in \left[\frac{1}{4}, \frac{3}{4}\right] \\ \frac{1}{2} & \text{if } a = 1 \\ 0 & \text{otherwise.} \end{cases}$$

A final result, which illustrates a theorem due to Hildenbrand (1983), gives conditions under which market demand is necessarily downward sloping. Again, the point is that with the continuum of agents market demand may be better behaved than individual demand.

Theorem 5: Consider an economy in which all individuals have identical preferences but differ in their incomes. In particular, assume that income is uniformly distributed over the interval [0, 1] and let $f(p, I)$ denote the identical demands of the individuals with income I who face prices p. Under the above conditions, the mean demand for each commodity has a nonpositive slope.

A sketch of a proof of the theorem follows: It is well known from consumer demand theory that the sign of the term $\partial f_k(p, I)/\partial p_k$ can be either positive or negative. Since individual substitution effects are nonpositive, to prove the result it is sufficient to demonstrate that the mean income effect is nonpositive.

The income effect as a result of a change in the price of commodity k on the demand for k, for an individual with income I, is given by

$$-f_k(p, I)\frac{\partial}{\partial I}f_k(p, I).$$

Therefore, the mean income effect is given by

$$-\int_0^1 f_k(p, I)\frac{\partial}{\partial I}f_k(p, I)\mathrm{d}I = -\frac{1}{2}\int_0^1 \frac{\partial}{\partial I}\left[f_k^2(p, I)\right]\mathrm{d}I = -\frac{1}{2}\left[f_k^2(p, 1) - f_k^2(p, 0)\right] = -\frac{1}{2}f_k^2(p, 1) \leq 0,$$

which establishes the result.

See Also

- aggregation (theory)
- demand theory
- integrability of demand
- law of demand

Bibliography

Antonelli, G.B. 1886. *Sulla Teoria Matematica della Economia Politica*. Pisa: Nella Tipognafia del Folchetto. Trans. J.S. Chipman and A.P. Kirman in *Preferences, Utility and Demand*, ed. J.S. Chipman et al., New York: Harcourt Brace Jovanovich, 1971.

Arrow, K.J. and Intriligator, M.D., eds. 1981–6. *Handbook of Mathematical Economics*, Vols 1, 2 and 3. Amsterdam: North-Holland.

Chipman, J.S., Hurwicz, L. Richter, M.K. and Sonnenschein, H.F., eds. 1971. *Preferences, Utility and Demand*. New York: Harcourt Brace Jovanovich.

Chipman, J.S. and Moore, J. 1979. On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21, 111–39.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–23.

Debreu, G. 1982. Existence of competitive equilibrium. In *Handbook of Mathematical Economics*, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Dow, J. and Sonnenschein, H. 1983. Samuelson and Chipman–Moore on utility generated community demand. In *Prices, Competition and Equilibrium*, ed. M. Peston and R. Quandt. Oxford: Philip Allan Publishers.

Eisenberg, B. 1961. Aggregation of utility functions. *Management Science* 7, 337–50.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21, 63–80.

Hicks, J.R. 1957. *A Revision of Demand Theory*. Oxford: Clarendon Press.

Hildenbrand, W. 1983. On the ‘law of demand’. *Econometrica* 51, 997–1020.

Hurwicz, L. and Uzawa, H. 1971. On the integrability of demand functions. In *Preferences, Utility and Demand*, ed. J.S. Chipman et al. New York: Harcourt Brace Jovanovich.

Nataf, A. 1953. Sur des qsts d’agrégation en économétrie. *Publications de l’Institute de Statistique de l’Université de Paris* 2, 5–61.

Samuelson, P. 1956. Social indifference curves. *Quarterly Journal of Economics* 70, 1–22.

Shafer, W. and Sonnenschein, H. 1982. Market demand and excess demand functions. In *Handbook of Mathematical Economics*, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Sonnenschein, H. 1973a. Do Walras’ identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6, 345–54.

Sonnenschein, H. 1973b. The utility hypothesis and market demand theory. *Western Economic Journal* 11, 404–10.

How to cite this article

Sonnenschein, H. "aggregate demand theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000055> doi:10.1057/9780230226203.0017

The New Palgrave Dictionary of Economics Online

aggregation (econometrics)

Thomas M. Stoker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The econometrics of aggregation is about modelling the relationship between individual (micro) behaviour and aggregate (macro) statistics, so that data from both levels can be used for estimation and inference about economic parameters. Practical models must address three types of individual heterogeneity – in income and preferences, in wealth and income risk, and in market participation. This entry discusses recent solutions to these problems in the context of demand analysis, consumption modelling and labour supply. Also discussed is work that uses aggregation structure to solve microeconometric estimation problems, and work that addresses whether macroeconomic interactions provide approximate solutions to aggregation problems.

Keywords

aggregate demand models; aggregation (econometrics); aggregation factors; approximate aggregation; calibration; computable stochastic growth models; constant relative risk aversion; demand models; exact aggregation; Gorman, W. (Terence); household demand models; identification; income-risk insurance; individual heterogeneity; industrial organization; law of demand; Mills ratio; reservation wage; selection effects; Theil, H.; uncorrelated transfers

Article

Aggregation refers to the connection between economic interactions at the micro and the macro levels. The micro level refers to the behaviour of individual economic agents. The macro level refers to the relationships that exist between economy-wide totals, averages or other economic aggregates. For instance, in a study of savings behaviour refers to the process that an individual or household uses to decide how much to save out of current income, whereas the aggregates are total or per-capita savings and income for a national economy or other large group. The econometrics of aggregation refers to modelling with the individual–aggregate connection in mind, creating a framework where information

on individual behaviour together with co-movements of aggregates can be used to estimate a consistent econometric model.

In economic applications one encounters many types and levels of aggregation: across goods, across individuals within households, and so on. We focus on micro to macro as outlined above, and our ‘individual’ will be a single individual or a household, depending on the context. We hope that this ambiguity does not cause confusion.

At a fundamental level, aggregation is about handling detail. No matter what the topic, the microeconomic level involves purposeful individuals who are dramatically different from one another in terms of their needs and opportunities. Aggregation is about how all this detail distils in relationships among economic aggregates. Understanding economic aggregates is essential for understanding economic policy. There is just too much individual detail to conceive of tuning policies to the idiosyncrasies of many individuals.

This detail is referred to as individual heterogeneity, and it is pervasive. This is a fact of empirical evidence and has strong econometric implications. If you ignore or neglect individual heterogeneity, then you can't get an interpretable relationship between economic aggregates. Aggregates reflect a smear of individual responses and shifts in the composition of individuals in the population; without careful attention, the smear is unpredictable and uninterpretable.

Suppose that you observe an increase in aggregate savings, together with an increase in aggregate income and in interest rates. Is the savings increase primarily arising from wealthy people or from those with moderate income? Is the impact of interest rates different between the wealthy and others? Is the response different for the elderly than for the young? Has future income for most people become more risky?

How could we answer these questions? The change in aggregate savings is a mixture of the responses of all the individuals in the population. Can we disentangle it to understand the change at a lower level of detail, like rich versus poor, or young versus old? Can we count on the mixture of responses underlying aggregate savings to be stable? These are questions addressed by aggregation.

Recent progress on aggregation and econometrics has centred on explicit models of individual heterogeneity. It is useful to think of heterogeneity as arising from three broad categories of differences. First, individuals differ in tastes and incomes. Second, individuals differ in the extent to which they participate in markets. Third, individuals differ in the situations of wealth and income risk that they encounter depending on the market environment that exists. Our discussion of recent solutions is organized around these three categories of heterogeneity. For deeper study and detailed citations, see the surveys by Blundell and Stoker (2005), Stoker (1993) and Browning, Hansen and Heckman (1999).

The classical aggregation problem provides a useful backdrop for understanding current solutions. We now review its basic features, as originally established by Gorman (1953) and Theil (1954). Suppose we are studying the consumption of some product by households in a large population over a given time period t. Suppose that the quantity purchased q_{it} is determined by household resources m_{it}, or ‘income’ for short, as in the formula:

$$q_{it} = \alpha_i + \beta_i m_{it}$$

Here α_i represents a base level consumption, and β_i represents household *i*'s marginal propensity to spend on the product.
For aggregation, we are interested in what, if any, relationship there is between average quantity and average income:

$$\bar{q}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} q_{it} \text{and} \bar{m}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} m_{it}$$

where all households have been listed as $i=1,\ldots,n_t$. Let's focus on one version of this issue, namely, what happens if some new income becomes available to households, either through economic growth or a policy. How will the change in average quantity purchased $\Delta\bar{q}$ be related to the change in average income $\Delta\bar{m}$?
Suppose that household *i* gets Δm_i in new income. Their change in quantity purchased is the difference between purchases at income $m_{it}+\Delta m_i$ and at income m_{it}, or

$$\Delta q_i = \beta_i \cdot \Delta m_i$$

Now, the average quantity change is $\Delta\bar{q} = \Sigma\, \Delta q_i / n_t$, so that

$$\Delta\bar{q} = \frac{1}{n_t}\sum_{i=1}^{n_t} \beta_i \cdot \Delta m_i \quad (1)$$

In general, it seems we need to know a lot about who gets the added income – which *i*'s get large values of Δm_i and which *i*'s get small values of Δm_i. With a transfer policy, any group of households could be targeted for the new income, and their specific set of values of β_i would determine $\Delta\bar{q}$. A full schedule of how much new income goes to each household *i* as well as how they spend it (that is, Δm_i and β_i), seems like a lot of detail to keep track of, especially if the population is large. Can we ever get by knowing just the change in average income $\Delta\bar{m} = \Sigma\, \Delta m_i / n_t$?
There are two situations where we can, where a full schedule is not needed:

1. 1. Each household spends in exactly the same way, namely, $\beta_I = \beta$ for all *i*, so that who gets the new income doesn't affect $\Delta\bar{q}$.
2. 2. The distribution of income transfers is restricted in a convenient way.

Situation 1 is (common) micro linearity, which is termed *exact aggregation*. Another way to understand the structure is to write (1) in the covariance formulation:

$$\Delta\bar{q} = \bar{\beta} \cdot \Delta\bar{m} + \frac{1}{n_t}\sum_{i=1}^{n_t} (\beta_i - \bar{\beta}) \cdot (\Delta m_i - \Delta\bar{m})$$
(2)

where we denote the average spending propensity as $\bar{\beta} = \Sigma_i \beta_i / n_t$. With exact aggregation there is no variation in β_i, so that $\beta_i = \beta = \bar{\beta}$ and the latter term always vanishes. That is, it doesn't matter who gets the added income because everyone spends the same way. When there is variation in β_i, matters are more complicated unless it can be assured that the new income were always given to households in a way that is uncorrelated with the propensities β_i. ‘Uncorrelated transfers’ provide an example of a Situation 2, but that is a distribution restriction that is hard to verify with empirical data.
Under uncorrelated transfers, we can also interpret the relationship between $\Delta\bar{q}$ and $\Delta\bar{m}$, that is, the macro propensity is the average propensity $\bar{\beta}$. There are other distributional restrictions that give a constant macro propensity, but a different one from the parameter produced by uncorrelatedness. For instance, suppose that transfers of new income always involved fixed shares of the total amount. That is, household *i* gets

$$\Delta m_i = s_i \Delta\bar{m}$$
(3)

In this case, average purchases are

$$\Delta\bar{q} = \frac{1}{n_t}\sum_{i=1}^{n_t} \beta_i \cdot (s_i \Delta\bar{m}) = \tilde{\beta}_{wtd} \cdot \Delta\bar{m}$$
(4)

where $\ddot{\beta}_{wtd}$ is the weighted average $\ddot{\beta}_{wtd} \equiv \Sigma_i \beta_i s_i / n_t$. This is a simple aggregate relationship, but the coefficient $\ddot{\beta}_{wtd}$ applies only for the distributional scheme (3); it matters who gets what share of the added income. Aside from being a weighted average of $\{\beta_i\}$, there is no reason for $\ddot{\beta}_{wtd}$ to be easily interpretable – for instance, if households with low β_i's have high s_i's, then $\ddot{\beta}_{wtd}$ will be low. If your aim was to estimate the average propensity $\bar{\beta}$, there is no reason to believe that the bias $\ddot{\beta}_{wtd} - \bar{\beta}$ will be small.

Empirical models that take aggregation into account apply structure to individual responses and to allowable distributional shifts. Large populations are modelled, so that compositional changes are represented via probability distributions, and expectations are used instead of averages (for example, mean quantity $E_t(q)$ is modelled instead of the sample average $\bar{q}_t$). Individual heterogeneity is the catch-all term for individual differences, and they must be characterized. Distribution restrictions must be applied where heterogeneity is important. For instance, in our example structure on the distribution of new income is required for dealing with the heterogeneity in β_i, but not for the heterogeneity in α_i.

Progress in empirical modelling has come about because of the enhanced availability of micro data over time. The forms of behavioural models in different research areas have been tightly characterized, which is necessary for understanding how to account for aggregation. That is, when individual heterogeneity is characterized empirically, the way is clear to understanding what distributional influences are relevant and must be taken into account. We discuss recent examples of this below.

Some solutions to aggregation problems

Demand models and exact aggregation

It is well known that demand patterns of individual households vary substantially with whether households are rich or poor, and vary with many observable demographic characteristics, such as household (family) size, age of head and ages of children, and so on. As surveyed in Blundell (1988), traditional household demand models relate household commodity expenditures to price levels, total household budget (income) and observable household characteristics. Aggregate demand models relate (economy-wide) aggregate commodity expenditures to price levels and the distribution of income and characteristics in the population. Demand models illustrate exact aggregation, a practical approach for accommodating heterogeneity at the micro and macro levels. These models assume that demand parameter values are the same for all individuals, but explicitly account for observed differences in tastes and income.

For instance, suppose we are studying the demand for food and we are concerned with the difference in demands for households of small size versus large size. We model food purchases for household i as part of static allocation of the budget m_{it} to j=1,…,J expenditure categories, where food is given by j=1, and price levels at time t are given by $P_t=(p_{1t},\ldots,p_{Jt})$. Small families are indicated by z_{it}=0 and large families by z_{it}=1.

Expenditure patterns are typically best fit in budget share form. For instance, a translog model of the food share takes the form

$$w_{1it} = \frac{p_{1t}q_{1it}}{m_{it}} = \frac{1}{D(p_t)}\left[\alpha_1 + \sum_{i=1}^{J}\beta_{1j}\ln p_{jt} + \beta_m \ln m_{it} + \beta_z z_{it}\right]$$
(5)

where $D(p_t) = 1 + \Sigma_{i=1}^{J}\beta_j \ln p_{jt}$. The parameters (α_1 and all β's) are the same across households, and the price levels (p_{jt}'s) are the same for all households but vary with t. Individual heterogeneity is represented by the budget m_{it} and the family size indicator z_{it}. We have omitted an additive disturbance for simplicity, which would represent another source of heterogeneity. The important thing for aggregation is that model (5) is intrinsically linear in the individual heterogeneity. That is, we can write

$$w_{1it} = b_1(p_t) + b_m(p_t)\cdot \ln m_{it} + b_z(p_t)\cdot z_{it}$$
(6)

The aggregate share of food in the population is the mean of food expenditures divided by mean budget, or

$$W_{1t} = \frac{E_t(m_{it}w_{1it})}{E_t(m_{it})} = b_1(p_t) + b_m(p_t)\cdot\frac{E_t(m_{it}\ln m_{it})}{E_t(m_{it})} + b_z(p_t)\cdot\frac{E_t(m_{it}z_{it})}{E_t(m_{it})}$$
(7)

The aggregate share depends on prices, the parameters (α_1 and all β's) and two statistics of the joint distribution of m_{it} and z_{it}. The first,

$$S_{mt} = \frac{E_t(m_{it}\ln m_{it})}{E_t(m_{it})}$$
(8)

is an entropy term that captures the size distribution of budgets, and the second

$$S_{zt} = \frac{E_t(m_{it}z_{it})}{E_t(m_{it})} \tag{9}$$

is the percentage of total expenditure accounted for by households with z_{it}=1, that is, large families. The expressions (6) and (7) illustrate *exact aggregation* models. Heterogeneity in tastes and budgets (incomes) are represented in an intrinsically linear way. For aggregate demand, all one needs to know about the joint distribution of budgets m_{ti} and household types z_{it} is a few statistics; here S_{mt} and S_{zt}. The obvious similarity between the individual model (6) and the aggregate model (7) raises a further question. How much bias is introduced by just fitting the individual model with aggregate data, that is, putting $E_t(m_{it})$ and $E_t(z_{it})$ in place of m_{it} and z_{it}, respectively? This can be judged by the use of *aggregation factors*. Define the factors π_{mt} and π_{zt} as

$$\pi_{mt} = \frac{S_{mt}}{\ln E_t(m_{it})} \text{ and } \pi_{zt} = \frac{S_{zt}}{E_t(z_{it})}$$

so that the aggregate share is

$$W_{1t} = \frac{E_t(m_{it}w_{1it})}{E_t(m_{it})} = b_1(p_t) + b_m(p_t)\cdot\pi_{mt}\cdot\ln E_t(m_{it}) + b_z(p_t)\cdot\pi_{zt}\cdot E_t(z_{it})$$

One can learn about the nature of aggregation bias by studying the factors π_{mt} and π_{zt}. If they are both roughly equal to 1 over time, then no bias would be introduced by fitting the individual model with aggregate data. If they are roughly constant but not equal to 1, then constant biases are introduced. If the factors are time varying, more complicated bias would result. In this way, with exact aggregation models, aggregation factors can depict the extent of aggregation bias.

The current state of the art in demand analysis uses models in exact aggregation form. The income (budget) structure of shares is adequately represented as quadratic in ln m_{it}, as long as many demographic differences are included in the analysis. This means that aggregate demand depends explicitly on many statistics of the income-demographic distribution, and it is possible to gauge the nature and sources of aggregation bias using factors as we have outlined. See Banks, Blundell and Lewbel (1997) for an example of demand modelling of British expenditure data, including the computation of various aggregation factors.

Exact aggregation modelling arises naturally in situations where linear models have been found to

provide adequate explanations of empirical data patterns. This is not always the case, as many applications require models that are intrinsically nonlinear. We now discuss an example of this kind where economic decisions are discrete.

Market participation and wages

Market participation is often a discrete decision. Labourers decide whether to work or not, firms decide whether to enter a market or exit a market. There is no 'partial' participation in many circumstances, and changes are along the extensive margin. This raises a number of interesting issues for aggregation.

We discuss these issues using a simple model of labour participation and wages. We consider two basic questions. First, how is the fraction of working (participating) individuals affected by the distribution of factors that determine whether each individual chooses to work? Second, what is the structure of average wages, given that wages are observed only for individuals who choose to work? The latter question is of interest for interpreting wage movements: if average wages go up, is that because (*a*) most individual wages went up or (*b*) low-wage individuals become unemployed, or leave work? These two reasons give rise to quite different views of the change in economic welfare associated with an increase in average wages.

The standard empirical model for individual wages expresses log wage as a linear function of time effects, schooling and demographic (cohort) effects. Here we begin with

$$\ln w_{it} = r(t) + \beta \cdot S_{it} + \varepsilon_{it} \tag{10}$$

where $r(t)$ represents a linear trend or other time effects, S_{it} is the level of training or schooling attained by individual i at time t, and ε_{it} are all other idiosyncratic factors. This setting is consistent with a simple skill price model, where $w_{ti}=R_tH_{it}$ with skill price $R_t=e^{r(t)}$ and skill (human capital) level $H_{it} = e^{\beta S_{it}+\varepsilon_{it}}$. We take eq. (10) to apply to all individuals, with the wage representing the available or offered wage, and β the return to schooling. However, we observe that wage only for individuals who choose to work.

We assume that individuals decide whether to work by first forming a reservation wage

$$\ln w_{it}^{*} = s^{*}(t) + \alpha \ln B_{it} + \beta^{*} \cdot S_{it} + \zeta_{it}$$

where $s(t)$ represents time effects, B_{it} is the income or benefits available when individual i is out of work at time t, S_{it} is schooling as before, and ζ_{it} are all other individual factors. Individual i will work at time

t if their offered wage is as big as their reservation wage, or $w_{it} \geq w_{it}^*$. We denote this by the participation indicator I_{it}, where I_{it}=1 if i works and I_{it}=0 if i doesn't work. This model of participation can be summarized as

$$I_{it} = 1\,[w_{it} \geq w_{it}^*] = 1\,[\ln w_{it} - \ln w_{it}^* \geq 0] = 1\,[s(t) - \alpha \ln B_{it} + \gamma \cdot S_{it} + v_{it} \geq 0] \tag{11}$$

where $s(t) \equiv r(t) - s^*(t)$, $\gamma \equiv \beta - \beta^*$ and $\nu_{It} \equiv \varepsilon_{it} - \zeta_{it}$.

If the idiosyncratic terms ε_{it}, ν_{it} are stochastic errors with zero means (conditional on B_{it},S_{it}) and constant variances, then (10) and (11) is a standard selection model. That is, if we observe a sample of wages from working individuals, they will follow (10) subject to the proviso that I_{it}=1. This can be accommodated in estimation by assuming that ε_{it}, ν_{it} have a joint normal distribution. That implies that the log wage regression of the form (10) can be corrected by adding a standard selection term as

$$\ln w_{it} = r(t) + \beta \cdot S_{it} + \frac{\sigma_{\varepsilon\nu}}{\sigma_\nu}\lambda\left[\frac{s(t) - \alpha \ln B_{it} + \gamma S_{it}}{\sigma_\nu}\right] + \eta_t. \tag{12}$$

Here, σ_ν is the standard deviation of ν and $\sigma_{\varepsilon\nu}$ is the covariance between ε and ν. $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the 'Mills ratio', where ϕ and Φ are the standard normal p.d.f. and c.d.f respectively. This equation is properly specified for a sample of working individuals – that is, we have $E(\eta_t | S_{it}, B_{it}, I_{it}=1) = 0$. For a given levels of benefits and schooling, eq. (11) gives the probability of participating in work as

$$E_t[I_{it} | B_{it}, S_{it}] = \Phi\left[\frac{s(t) - \alpha \ln B_{it} + \gamma \cdot S_{it}}{\sigma_\nu}\right] \tag{13}$$

where $\Phi[\cdot]$ is the normal c.d.f.

For studying average wages, the working population is all individuals with I_{it}=1. The fraction of workers participating is therefore the (unconditional) probability that $\alpha \ln B_{it} - \gamma \cdot S_{it} - \nu_{it} \leq s(t)$. This probability is the expectation of I_{it} in (11), an intrinsically nonlinear function in observed heterogeneity B_{it} and S_{it}

and unobserved heterogeneity ν_{it}, so we need some explicit distribution assumptions. In particular, assume that the participation index $\alpha \ln B_{it} - \gamma \cdot S_{it} - \nu_{it}$ is normally distributed with mean $\mu_t = \alpha E_t(\ln B_{it}) - \gamma E_t(S_{it})$ and variance

$$\sigma_t^2 = \alpha^2 Var_t(\ln B_{it}) + \beta^2 Var_t(S_{it}) - 2\alpha\beta \cdot Cov_t(\ln B_{it}, S_{it}) + \sigma_v^2. \quad (14)$$

Now we can derive the labour participation rate (or one minus the unemployment rate) as

$$E_t[I_{it}] = \Phi\left[\frac{s(t) - \alpha E_t(\ln B_{it}) + \gamma E_t(S_{it})}{\sigma_t}\right] \quad (15)$$

where again $\Phi[\cdot]$ is the normal c.d.f. This formula relates the participation rate to average out-of-work benefits $E_t(\ln B_{it})$ and average training $E_t(S_{it})$, as well as their variances and covariances through σ_t. The specific relation depends on the distributional assumption adopted; (15) relies on normality of the participation index in the population.

For wages, a similar analysis applies. Log wages are a linear function (10) applicable to the full population. However, for participating individuals, the intrinsically nonlinear selection term is introduced, so that we need explicit distributional assumptions. Now suppose that log wage $\ln w_{it}$ and the participation index $\alpha \ln B_{it} - \gamma \cdot S_{it} - \nu_{it}$ are joint normally distribution. It is not hard to derive the expression for average log wages of working individuals

$$E_t[\ln w_{it} | I_{it} = 1] = r(t) + \beta \cdot E_t(S_{it} | I = 1) + \frac{\sigma_{\varepsilon v}}{\sigma_t}\lambda\left[\frac{s(t) - \alpha E_t(\ln B_{it}) + \gamma E_t(S_{it})}{\sigma_t}\right]. \quad (16)$$

This is an interesting expression, which relates average log wage to average training of the workers as well as to the factors that determine participation.

However, we are not interested in average log wages, but rather average wages $E_t(w_{it})$. The normality structure we have assumed is enough to derive a formulation of average wages, although it is a little complex to reproduce in full here. In brief, Blundell, Reed and Stoker (2003) show that the average

wages of working individuals $E\,[w_{it}|I_{it}=1]$ can be written as

$$\ln E[w_{it}|I_{it}=1] = r(t) + \beta \cdot E_t(S_{it}) + \Omega_t + \Psi_t \tag{17}$$

where Ω_t, Ψ_t are correction terms that arise as follows. Ω_t corrects for the difference between the log of an average and the average of a log, as

$$\Omega_t \equiv \ln E_t(w_{it}) - E_t(\ln w_{it}) + \Omega_t.$$

Ψ_t corrects for participation, as

$$\Psi_t \equiv \ln E[w_{it}|I_{it}=1] - \ln E_t(w_{it}).$$

Recall our original question, about whether an increase in average wages is due to an increase in individual wages or to increased unemployment of low-wage workers. That is captured in (17). That is, Ψ_t gives the participation effect, and the other terms capture changes in average wage $E_t(w_{it})$ when all are participating. As such, this analysis provides a vehicle for separating overall wage growth from compositional effects due to participation.

Blundell, Reed and Stoker (2003) analyse British employment using a framework similar to this, but also allowing for heterogeneity in hours worked. Using out-of-work benefits as an instrument for participation, they find that over 40 per cent of observed aggregate wage growth from 1978 to 1996 arises from selection and other compositional effects.

We have now discussed aggregation and heterogeneity with regard to tastes and incomes, and market participation. We now turn to heterogeneity with regard to risks and market environments.

Consumption and risk environments

Consumption and savings decisions are clearly affected by preference heterogeneity, as we discussed earlier. The present spending needs of a large family clearly differ from those of a small family or a single individual, the needs of teenage children differ from those of preschoolers, the needs of young adults differ from those of retirees, and on and on. These aspects are very important, and need to be addressed as they were in demand models above. Browning and Lusardi (1996) survey the extensive

evidence on heterogeneity in consumption, and Attanasio (1999) is an excellent comprehensive survey of work on consumption.

We use consumption and savings to illustrate another type of heterogeneity, namely, that of wealth and income risks. That is, with forward planning under uncertainty, the risk environment of individuals or households becomes relevant. There can be individual shocks to income, such as a work layoff or a health problem, or aggregate shocks, such as an extended recession or stock market boom. Each of these shocks can differ in its duration – a temporary layoff can be usefully viewed as transitory, whereas a debilitating injury may affect income for many years. In planning consumption, it is important to understand the role of income risks and wealth risks. When there is no precautionary planning, such as when consumers have quadratic preferences, income risks do not become intertwined with other heterogeneous elements. However, when there is risk aversion, then the precise situation of individual income risks and insurance markets is relevant.

A commonly used model for income is to assume multiplicative permanent and transitory components, with aggregate and individual shocks, as in

$$\Delta \ln y_{it} = (\eta_t + \Delta u_t) + (\varepsilon_{it} + \Delta v_{it}).$$

Here $\eta_t + \Delta u_t$ is the common aggregate shock, with η_t a permanent component and Δu_t transitory. The idiosyncratic shock is $\varepsilon_{it} + \Delta v_{it}$, where ε_{it} is permanent and Δv_{it} transitory.

For studying individual level consumption with precautionary planning, it is standard practice to assume constant relative risk aversion (CRRA) preferences and assume that the interest rate r_t is small. This, together with the income process above, gives a log-linear approximation to individual consumption growth

$$\Delta \ln c_{it} = \rho r_t + (\beta + \phi r_t)' z_{it} + k_1 \sigma_{At} + k_2 \sigma_{it} + \kappa_1 \eta_t + \kappa_2 \varepsilon_{it}. \quad (18)$$

Here, z_{it} reflects heterogeneity in preferences, such as differences in demographic characteristics. σ_{At} is the variance of aggregate risk and σ_{it} s the variance of idiosyncratic risk (with each conditional on what is known at time $t - 1$), so that these terms reflect precautionary planning. Finally, η_t and ε_{it} arise because of adjustments that are made as permanent shocks are revealed. At time $t - 1$ these shocks are not possible to forecast, but then they are incorporated in the consumption plan once they are revealed. In terms of the level of consumption c_{it}, eq. (18) is written as

$$c_{it} = exp(\ln c_{it-1} + \rho r_t + (\beta + \phi r_t)' z_{it} + k_1 \sigma_{At} + k_2 \sigma_{it} + \kappa_1 \eta_t + \kappa_2 \varepsilon_{it}).$$

This is an intrinsically nonlinear model in the following heterogeneous elements: $\ln c_{it-1}$, z_{it}, σ_{it} and ε_{it}. For aggregation, it seems we would need a great deal of distributional structure.
Here is where we can see the role of the risk environment, or markets for insurance for income risks. That is, if there were complete markets with insurance for all risks, then all risk terms vanish from consumption growth. When complete insurance exists for idiosyncratic risks only, then the idiosyncratic terms σ_{it} and ε_{it} vanish from consumption growth, since less precautionary saving is needed. Otherwise, the idiosyncratic risk terms σ_{it} and ε_{it} represent heterogeneity that must be accommodated just like preference differences (and in other settings, participation differences).
In the realistic situation where risks are not perfectly insurable, we require distributional assumptions in order to formulate aggregate consumption. For instance, suppose that we assume that $(\ln c_{it-1}, (\beta + \phi r_t)' z_{it}, \varepsilon_{it})$ is joint normally distributed with $E_t(\varepsilon_{it})=0$, and that idiosyncratic risks are drawn from the same distribution for each consumer (so $\sigma_{it}=\sigma_{It}$ for each i), and that a stability assumption applies to the distribution of lagged consumption. Blundell and Stoker (2005) show that aggregate consumption growth is

$$\Delta \ln E_t(c_{it}) = \rho r_t + (\beta + \phi r_t)' E_t(z_{it}) + k_1 \sigma_{At} + k_2 \sigma_{It} + \kappa_1 \eta_t + \Lambda_t.$$

This model explains aggregate consumption growth in terms of the mean of preference heterogeneity, risk terms, and an aggregation factor Λ_t. The factor Λ_t is comprised of variances and covariances of the heterogeneous elements $\ln c_{it-1}$, z_{it} and ε_{it}. Thus, this model reflects how aggregate consumption will vary as the individual incomes become more or less risky, and captures how the income risk interplays with previous consumption values.
In overview, as micro consumption models are nonlinear, distributional restrictions are essential. On this point, an empirical fact is that the distribution of household consumption is often observed to be well approximated by a lognormal distribution, and so such lognormal restrictions may have empirical validity. Also relevant here is the empirical study of income and wealth risks, which has focused on earnings processes; see Meghir and Pistaferri (2004) for a recent contribution.

Micro to macro and vice versa

We now turn to two related uses of aggregation structure that have emerged in the literature.

Aggregation as a solution to microeconometric estimation

Consider a situation where the estimation of a model at the micro level is the primary goal of empirical work. Some recent work uses aggregation structure to enhance or permit micro-level parameter identification and estimation. Since aggregation structure provides a bridge between models at the micro level and the aggregate level, it permits all data sources – individual-level data and aggregate-level data – to be used for identification and estimation of economic parameters. Sometimes it is necessary to combine all data sources to identify economic effects (for example, Jorgenson, Lau and Stoker, 1982), and sometimes one can study (micro) economic effects with aggregate data alone (for example, Stoker, 1986). Recent work has developed more systematic methods of using aggregate data to improve micro-level estimates. In particular, one can match aggregate data with simulated moments from the individual data as part of the estimation process.

To see how this can work, suppose we have data on labour participation over several time periods (or groups). We assume that the participation decision is given by the model (11) with normal unobserved heterogeneity, as discussed above. We normalize $\sigma_\nu=1$ and take $s(t)=\psi$, a constant, so that the unknown parameters of the participation model are α, γ and ψ. The data situation is as follows; for each group $t=1,\ldots,T$, we observe the proportion of labour participants P_t and a random sample of benefits and schooling values, $\{B_{it}, S_{it}, i=1,\ldots,n_t\}$. Given the (probit) expression (13), estimation can be based on matching the observed proportion P_t to the simulated moment

$$\bar{P}_t(\alpha, \gamma, \psi) = \frac{1}{n_t}\sum_{i=1}^{n_t} \Phi\left[\psi - \alpha \ln B_{it} + \gamma \cdot S_{it}\right].$$

For instance, we could estimate by least squares over groups, by choosing $\hat{\alpha}, \hat{\gamma}, \hat{\psi}$ to minimize

$$\sum_{t=1}^{T} \left(P_t - \bar{P}_t(\alpha, \gamma, \psi)\right)^2.$$

Note that this approach does not require a specific assumption on the joint distribution of B_{it} and S_{it} for each t, as the random sample provides the distributional information needed to link the parameters to the observed proportion P_t.

It turns out that this approach for estimation is extremely rich, and was essentially mapped out by Imbens and Lancaster (1994). It has become a principal method of estimating demands for differentiated products, for use in structural models of industrial organization. See Berry, Levinsohn and Pakes (2004) for good coverage of this development.

Can macroeconomic interaction solve aggregation problems?

The basic heuristic that underlies much macroeconomic modelling is that, because of markets, individuals are very coordinated in their actions, so that individual heterogeneity likely has a secondary impact. In simplest terms, the notion is that common reactions across individuals will swamp any behavioural differences. This idea is either just wrong or, at best, very misleading for economic analysis. But that is not to deny that in real world economies there are many elements of commonality in reactions across individuals. Households face similar prices, interest rates and opportunities for employment. Extensive insurance markets effectively remove some individual differences in risk profiles. Optimal portfolio investment can have individuals choosing the same (efficient) basket of securities.

The question whether market interactions can minimize the impact of individual heterogeneity is a classic one, and by and large the answers are negative. However, there has been some recent work with calibrated stochastic growth models that raises some possibilities. A principal example of this is Krusell and Smith (1998), which we now discuss briefly. The Krusell–Smith set-up has infinitely lived consumers, with the same preferences within each period, but with different discount rates and wealth holdings. Each consumer has a chance of being unemployed each period, so there are transitory individual income shocks. Production arises from labour and capital, and there are transitory aggregate productivity shocks. Consumers can insure for the future by investing in capital only. Thus, insurance markets are incomplete, and consumers cannot hold negative capital amounts.

To make savings and portfolio decisions, consumers must predict future prices. To do this, each consumer must keep track of the evolution of the entire distribution of wealth holdings, in principle. This is a lot of information to know, just like what is needed for standard aggregation solutions as discussed earlier. Krusell–Smith's simulations show, however, that this forecasting problem is much easier than one would suspect. That is, for consumer planning and for computing equilibrium, consumers get very close to optimal solutions by keeping track of only two things: mean wealth in the economy and the aggregate productivity shock. This is approximate aggregation, a substantial simplification of the information requirements that one would expect.

The source of this simplification, as well as its robustness, is a topic of active current study. One aspect is that most consumers, especially those with lowest discount rates, save enough to insure their risk so that their propensity to save out of wealth is essentially constant. Those consumers also hold a large fraction of the wealth, so that saving is essentially linear in wealth. This means that there is (approximate) exact aggregation structure, with the mean of wealth determining how much aggregate saving is undertaken. That is, the nature of savings and wealth accumulation approximately solves the aggregation problem for individual forecasting. Aggregate consumption, however, does not exhibit the same simplification. Many low-wealth consumers become unemployed and encounter liquidity constraints. Their consumption is much more sensitive to current output than that of wealthier consumers.

These results depend on the specific formulation of the growth model. Krusell and Smith (2006) survey work that suggests that their type of approximate aggregation can be obtained under a variety of variations of the basic model assumptions. As such, this work raises a number of fascinating issues on the interplay between economic interaction, aggregation and individual heterogeneity. However, it remains to be seen whether the structure of such calibrated models is empirically relevant to actual economies, or whether forecasting can be simplified even with observed variation in saving propensities

of wealthy households.

Future progress

Aggregation problems are among the most difficult in empirical economics. The progress that has been made recently is arguably due to two complementary developments. First is the enormous expansion in the availability of data on the behaviour of individual agents, including consumers, households, firms, and so on, in both repeated cross-section and panel data form. Second is the enormous expansion in computing power that facilitates the study of large data sources. These two trends can be reasonably expected to continue, which makes the prospects for further progress quite bright.

There is sufficient variety and complexity in the issues posed by aggregation that progress may arise from many approaches. For instance, we have noted how the possibility of approximate aggregation has arisen in computable stochastic growth models. For another instance, it is sometimes possible to derive properties of aggregate relationships with very weak assumptions on individual behaviour, as in Hildenbrand's (1994) work of the law of demand.

But is seems clear to me that the best prospects for progress lie with careful microeconomic modelling and empirical work. Such work is designed to ferret out economic effects in the presence of individual heterogeneity, and can also establish what are ‘typical’ patterns of heterogeneity in different applied contexts. Knowledge of typical patterns of heterogeneity is necessary for characterizing the distributional structure that will facilitates aggregation, and such distributional restrictions can then be refuted or validated with actual data. That is, enhanced understanding of the standard structure in the main application areas of empirical economics, such as with commodity demand, consumption and saving and labour supply, will lead naturally to an enhanced understanding of aggregation problems and accurate interpretation of aggregate relationships. There has been great progress of this kind in the past few decades, and there is no reason to think that such progress won't continue or accelerate.

Bibliography

Attanasio, O. 1999. Consumption. In *Handbook of Macroeconomics*, vol. 1B, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Banks, J., Blundell, R. and Lewbel, A. 1997. Quadratic Engel curves, indirect tax reform and welfare measurement. *Review of Economics and Statistics* 79, 527–39.

Berry, S., Levinsohn, J. and Pakes, A. 2004. Differentiated products demand systems from a combination of micro and macro data: the new car market. *Journal of Political Economy* 112, 68–105.

Blundell, R. 1988. Consumer behaviour: theory and empirical evidence. *Economic Journal* 98, 16–65.

Blundell, R., Reed, H. and Stoker, T. 2003. Interpreting aggregate wage growth. *American Economic Review* 93, 1114–31.

Blundell, R. and Stoker, T. 2005. Aggregation and heterogeneity. *Journal of Economic Literature* 43, 347–91.

Browning, M., Hansen, L. and Heckman, J. 1999. Micro data and general equilibrium. In *Handbook of Macroeconomics*, vol. 1A, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Browning, M. and Lusardi, A. 1996. Household saving: micro theories and micro facts. *Journal of Economic Literature* 34, 1797–855.

Gorman, W.M. (Terence). 1953. Community preference fields. *Econometrica* 21, 63–80.

Hildenbrand, W. 1994. *Market Demand: Theory and Empirical Evidence*. Princeton: Princeton University Press.

Imbens, G. and Lancaster, T. 1994. Combining micro and macro data in microeconometric models. *Review of Economic Studies* 61, 655–80.

Jorgenson, D., Lau, L. and Stoker, T. 1982. The transcendental logarithmic model of aggregate consumer behavior. In *Advances in Econometrics*, vol. 1, ed. R. Basmann, and G. Rhodes. Greenwich, CT: JAI Press.

Krusell, P. and Smith, A. 1998. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy* 106, 867–96.

Krusell, P. and Smith, A. 2006. Quantitative macroeconomic models with heterogeneous agents. In *Advances in Economics and Econometrics, Proceedings of the Ninth World Congress of the Econometric Society*, ed. R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press.

Meghir, C. and Pistaferri, L. 2004. Income variance dynamics and heterogeneity. *Econometrica* 72, 1–32.

Stoker, T. 1986. Simple tests of distributional effects on macroeconomic equations. *Journal of Political Economy* 94, 763–95.

Stoker, T. 1993. Empirical approaches to the problem of aggregation over individuals. *Journal of Economic Literature* 31, 1827–74.

Theil, H. 1954. *Linear Aggregation of Economic Relations*. Amsterdam: North-Holland.

How to cite this article

Stoker, Thomas M. "aggregation (econometrics)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000234> doi:10.1057/9780230226203.0018

The New Palgrave Dictionary of Economics Online

aggregation (production)

Jesus Felipe and Franklin M. Fisher

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Aggregation concerns the conditions under which several variables can be treated as one, or macro-relationships derived from micro-relationships. This problem is especially important in production, where, without proper aggregation, one cannot interpret the properties of the aggregate production function. The conditions under which aggregate production functions exist are so stringent that real economies surely do not satisfy them. The aggregation results pose insurmountable problems for theoretical and applied work in fields such as growth, labour or trade. They imply that intuitions based on micro variables and micro production functions will often be false when applied to aggregates.

Keywords

aggregation (production); Cambridge capital theory debates; capital aggregation; Cobb–Douglas functions; endogenous growth; growth accounting; Hicks, J.; Hicks–Leontief aggregation; labour aggregation; Leontief, W.; National Income and Products Account (NIPA); neoclassical growth theory; output aggregation; production functions; productivity (measurement problems); total factor productivity

Article

Aggregation in production concerns the conditions under which macro production functions can be derived from micro production functions. Microeconomic theory elegantly treats the behaviour of optimizing individual agents in a world with an arbitrarily long list of individual commodities and prices. However, the desire to analyse the great aggregates of macroeconomics – gross national product, inflation, unemployment, and so forth – leads to theories that treat such aggregates directly. The aggregation ‘problem’ matters because without proper aggregation one cannot interpret the properties of such macroeconomic models. This is particularly true as regards the production sector.

Leontief's theorem

Underlying many results on aggregation is a theorem of Leontief (1947a; 1947b). Let x and y be vectors of variables and $F(x, y)$ a twice-differentiable function. It is desired to aggregate over x, that is, to replace x with a scalar aggregator function, $g(x)$, such that $F(x) = H[g(x), y]$. This can be done if and only if, along any surface on which $F(x, y)$ is constant, the marginal rate of substitution between each pair of elements of x is independent of y. (For a proof, see Fisher, 1993, pp. xiv–xvi.)

Hicks – Leontief aggregation

Since optimizing, price-taking agents equate marginal rates of substitution to price ratios, one restriction permitting aggregation over commodities is the assumption that the prices of all goods to be included in an aggregate always vary proportionally. This is called ‘Hicks–Leontief aggregation’ (Leontief, 1936; Hicks, 1939) and is a powerful expository tool. It requires no special assumptions as to the form of utility or production functions, but is applicable only in relatively artificial situations. Under more general circumstances, restrictions on utility or production functions become essential.

Aggregation in consumption

Consider a single household. Suppose that we wish to describe behaviour in terms of aggregate commodities such as ‘food’ or ‘clothing’. By Leontief's Theorem, a food aggregate exists if and only if the marginal rate of substitution between any two kinds of food is independent of consumption of any non-food commodity. If a similar restrictive condition is satisfied for all the aggregates to be constructed, then the household's utility function can be written in aggregate terms.

Even such restrictive conditions will not always suffice. If we wish to represent the household as maximizing the aggregate utility function subject to an aggregate budget constraint, we must have aggregate prices as well as aggregate consumption goods. This requires that aggregates such as ‘food’ be homothetic in their component variables, again considerably restricting the household's utility function (Gorman, 1959; Blackorby et al., 1970).

Aggregation over agents presents a different set of questions. Suppose that we wish to treat the aggregate demands of a collection of households as the demands of a single, aggregate household. Then, only aggregate income and not its distribution can influence demand. At given prices, this makes the income derivative of every household's demand for a given commodity the same constant. Engel curves must be parallel straight lines. If zero income implies zero consumption, then all households must have the same homothetic utility function (Gorman, 1953).

In general, the only consumer-theoretic restrictions obeyed by aggregate demand functions are those of continuity, homogeneity of degree zero, and the various restrictions implied by the budget constraint (cf. Sonnenschein, 1972; 1973).

Aggregation in production

A more detailed survey of much of what follows in this section is given in Felipe and Fisher (2003). The analysis of aggregation conditions for production functions is far richer and the conditions even

more demanding than in the case of demand functions. Moreover, the subject has a complicated history and bears on the very foundations of neoclassical *macro*economics, negatively implicating the use of such important concepts as ‘total factor productivity’, ‘natural rate of growth’, ‘capital–labour ratio’, and even such terms as ‘investment’, ‘capital’, ‘labour’, and ‘output’.

To take a simple example, suppose we have two production functions $Q^A = f^A(K_1^A, K_2^A, L^A)$ and $Q^B = f^B(K_1^B, K_2^B, L^B)$ for firms *A* and *B*, where $K_1 = K_1^A + K_1^B$, $K_2 = K_2^A + K_2^B$ and $L = L^A + L^B$(*K* refers to capital – two types – and *L* to labour – assumed homogeneous). The problem is to determine whether and in what circumstances there exists a function $K = h(K_1, K_2)$ where the aggregator function *h*(·) has the property that $G(K, L) = G[h(K_1, K_2), L] = \Psi(Q^A, Q^B)$, and the function Ψ is the production possibility curve for the economy. Note that we have implicitly assumed that a production function exists for the firm. Further, even within the firm there is a problem of aggregation over factors. Here, we concentrate on aggregation over firms.

Klein (1946a; 1946b) initiated the first debate on aggregation in production functions. He argued that the aggregate production function should be strictly a technical relationship, akin to the micro production function, and objected to utilizing the entire micro model with the assumption of profit-maximizing behaviour by producers in deriving the production functions of the macro model.

However, Kenneth May (1947) pointed out that this program is not generally achievable and, indeed, rests on a misunderstanding of what production functions actually are – even at the micro level. A production function does not tell us what outputs are or can be produced from a given set of inputs. It tells us what the *maximum* output is of a particular commodity, given a vector of inputs and the other outputs that are also to be produced from them.

That Klein's aggregation program is generally unachievable was specifically proved by André Nataf (1948). He showed that such aggregation is possible if and only if all micro production functions are additively separable in capital and labour.

The problem here is as follows. Suppose there are n firms indexed by $\nu = 1, \ldots, n$. Each produces the same output *Y*(ν) using the same type of labour *L*(ν), and a single type of capital *K*(ν). The ν th firm has a two-factor production function $Y(\nu) = f^{\nu}\{K(\nu), L(\nu)\}$. The total output of the economy is $Y = \Sigma_{\nu} Y(\nu)$, total labour is $L = \Sigma_{\nu} L(\nu)$. Capital, on the other hand, may differ from firm to firm. Under what conditions can total output Y be written as $Y = \Sigma_{\nu} Y(\nu) = F(K, L)$ where $K = K\{K(1), \ldots, K(n)\}$ and $L = L\{L(1), \ldots, L(n)\}$ are indices of aggregate capital and labour, respectively? Nataf showed that, where the variables *K*(ν) and *L*(ν) are free to take on all values, the aggregate production function $Y = F(K, L)$ exists, if *and only if* every firm's production function is additively separable in labour and capital, that is, if every f^{ν} can be written in the form $f^{\nu}\{K(\nu), L(\nu)\} = \varphi^{\nu}\{K(\nu)\} + \psi^{\nu}\{L(\nu)\}$. Moreover, if one insists that labour aggregation be ‘natural’, with the L appearing in the aggregate production function, then all the $\psi^{\nu}\{L(\nu)\} = c\{L(\nu)\}$, where c is the same for all firms.

Nataf's theorem provides an extremely restrictive condition for inter-sectoral or even inter-firm aggregation. Evidently, aggregate production functions will not exist unless there are some further restrictions on the problem.

In fact, such restrictions are available; they stem from the requirement that a production function describe *efficient* production possibilities.

Capital aggregation

Consider the simplest case of two factors, with physically homogeneous capital (*K*) and homogeneous labour (*L*), where total capital can be written as $K = \Sigma_\nu K(\nu)$, efficient production requires that aggregate output *Y* be maximized given aggregate labour (*L*) and aggregate capital (*K*). Under these simplified circumstances, it follows that $Y^M = F(K, L)$ where Y^M is maximized output, since, as was pointed out by May (1946; 1947), individual allocations of labour and capital to firms would be determined in the course of the maximization problem. This holds even if all firms have different production functions and whether or not there are constant returns.

In the (somewhat) more realistic case where only labour is homogeneous and technology is embodied in capital, Fisher (1965) proposed to treat the problem as one of labour being allocated to firms so as to maximize output, with capital being firm-specific. Here, no ‘natural’ aggregate of capital exists.

Given that output is maximized with respect to the allocation of labour to firms, with such maximized output denoted by Y^*, the question becomes: under what circumstances is it possible to write total output as $Y^* = F(J, L)$ where $J = J\{K(1), \ldots, K(n)\}$, where $K(\nu)$, $\nu = 1, \ldots, n$, represents the stock of capital of each firm (that is, one kind of capital per firm)? Since the values of $L(\nu)$ are determined in the optimization process there is no labour aggregation problem. The entire problem in this case lies in the existence of a capital aggregate. Since Leontief's condition is both necessary and sufficient for the existence of a group capital index, the previous expression for Y^* is equivalent to $Y^* = G\{K(1), \ldots, K(n), L\}$ if and only if the marginal rate of substitution between any pair of the $K(\nu)$ is independent of *L*.

Fisher drew the implications of this condition. He showed that, under strictly diminishing returns to labour ($f^\nu_{LL} < 0$), if any one firm has an additively separable production function (that is, $f^\nu_{KL} \equiv 0$), then a necessary and sufficient condition for capital aggregation is that *every* firm have such a production function. (Throughout, such subscripts denote partial differentiation in the obvious manner.) This means that capital aggregation is impossible if there is both a firm which uses labour and capital in the same production process, and another one which has a fully automated plant. Fisher found that a necessary and sufficient condition for capital aggregation is that every firm's production function satisfy a partial differential equation in the form $f^\nu_{KL} / f^\nu_K f^\nu_{LL} = g(f^\nu_L)$, where *g* is the same function for all firms. More important, on the assumption of constant returns to scale, the case of capital-augmenting technical differences (that is, embodiment of new technology can be written as the product of the amount of capital times a coefficient) turns out to be the *only case* in which a capital aggregate exists. This means that each firm's production function must be writeable as $F(b_\nu K_\nu, L_\nu)$, where the function $F(\cdot, \cdot)$ is common to all firms, but the parameter b_ν can differ. Under these circumstances, a unit of one type of new capital equipment is the exact duplicate of a fixed number of units of old capital equipment (‘better’ is equivalent to ‘more’). As we would expect, given constant returns to scale, the aggregate stock of capital can be constructed with capital measured in efficiency units. Fisher (1965)

could not come up with a closed-form characterization of the class of cases in which an aggregate stock of capital exists when the assumption of constant returns is dropped. Nevertheless, as he showed, there do exist classes of non-constant returns production functions which do allow construction of an aggregate capital stock. On the other hand, if constant returns are not assumed there is no reason why perfectly well-behaved production functions cannot fail to satisfy Fisher's partial differential equation given above. Capital aggregation is then impossible if any firm has one of these ‘bad apple’ production functions. To sum up: aggregate production functions exist if and only if all micro production functions are identical except for the capital efficiency coefficient – an extremely restrictive condition.

Working with the profits function rather than with the production function, Gorman (1968) reached similar conclusions to those of Fisher.

Fisher extended his original work. First of all, he analysed (1965) the case where each firm produces a single output with a single type of labour, but two capital goods, that is, $Y(\nu) = f^{\nu}(K_1, K_2, L)$. Here Fisher distinguished between two different cases. The first is that of aggregation across firms over one type of capital (for example, plant or equipment). Fisher concluded that the construction of a sub-aggregate of capital goods requires even more stringent conditions than for the construction of a single aggregate. For example, if there are constant returns in K_1, K_2 and L, there will not be constant returns in K_1 and L, so that the difficulties of the two-factor non-constant returns case appear. Further, if the ν th firm has a production function with all three factors as complements, then no K_1 aggregate can exist. Thus, for example, if any firm has a generalized Cobb–Douglas production function (with the ν argument omitted) in plant, equipment, and labour $Y = AK_1^{\alpha}K_2^{\beta}L^{1-\alpha-\beta}$, one cannot construct a separate plant or separate equipment aggregate for the economy as a whole (although this does not prevent the construction of a full capital aggregate).

The other case Fisher (1965) considered was that of the construction of a complete capital aggregate. In this case, a necessary condition is that it be possible to construct such a capital aggregate for each firm taken separately; and a necessary and sufficient condition (with constant returns), given the existence of individual firm aggregates, is that all firms differ by at most a capital augmenting technical difference. They can differ *only* in the way in which their individual capital aggregate is constructed.

Second, Fisher (1982) asked whether the crux of the aggregation problem derives from the fact that capital is considered to be an immobile factor. He showed that the aggregation problem seems to be due only to the fact that capital is fixed and is not allocated efficiently. That is true in the context of a two-factor production function. However, if one works in terms of many factors, all mobile over firms, and asks when it is possible to aggregate them into macro groups, the mobility of capital has little bearing on the issue. In fact, where there are several factors, each of which is homogeneous, optimal allocation across *firms* does not guarantee aggregation across *factors*. The conditions for the existence of such aggregates are still very stringent, but this has to do with the necessity of aggregating over firms rather than with the immobility of capital. A possible way of interpreting the existence of aggregates at the firm level is that each firm could be regarded as having a two-stage production process. In the first one, the factors to be aggregated, $X_i(\nu)$, are combined to produce an intermediate output, $\varphi^{\nu}(X(\nu))$. This intermediate output is then combined with the other factor, $L(\nu)$, to produce the final output. Aggregation of X can be done if and only if firms are either all alike as regards the first stage of production, or all alike as regards the second stage. If they are all alike as regards the first stage, then the

fact that L is mobile plays no role. If they are all alike as regards the second stage, then the fact that the X_i are mobile plays no role.

Finally, Fisher (1983) is another extension of the original problem to study the conditions under which full and partial capital aggregates, such as ‘plant’ or ‘equipment’, would exist simultaneously. Not surprisingly, the results are as restrictive as those above. See also Blackorby and Schworm (1984).

Labour and output aggregation

Fisher (1968) went on to study the problems involved in labour and output aggregation, pointing out that the aggregation problem is not restricted to capital. Output aggregation and labour aggregation are also necessary if one wants to use a sector-wide or economy-wide aggregate production function.

Fisher again studied aggregation over firms, with labours and outputs shifted over firms to achieve efficient production, given the capital stocks. In the simplest case of constant returns, a labour aggregate will exist if and only if a given set of relative wages induces all firms to employ different labours in the same proportions. Similarly, where there are many outputs, an output aggregate will exist if and only if a given set of relative output prices induces all firms to produce all outputs in the same proportion. Thus, the existence of a labour aggregate requires the absence of specialization in employment; and the existence of an output aggregate requires the absence of specialization in production – indeed, all firms must produce the same market basket of outputs differing only in their scale. (Blackorby and Schworm, 1988, is an extension of Fisher, 1968.)

Houthakker – Sato aggregation conditions

Whereas Fisher sought to develop conditions where aggregate production functions would always work, Houthakker (1955–56) and Sato (1975) considered two-factor cases in which the problem was restricted by assuming that the distribution of capital over firms remains constant. In such cases it is obvious that one can aggregate over capital. Houthakker and Sato's contributions (see also Levhari, 1968) were to show the relationships between the fixed distribution of capital and the form of the aggregate production function.

Fisher's simulations

But, if aggregate production functions do not exist, how is it that they appear to ‘work’ in the sense that they fit the data well, that the estimated elasticities are close to the factor shares, and that wage rates are approximate the calculated marginal product of labour? We shall have more to say on this below, but here consider another result of Fisher (1971). This paper reports the results of simulations in a simple (heterogeneous capital, homogeneous labour and output) economy in which the aggregation conditions are known not to be satisfied. The principal result is that when, despite this, calculated factor shares just happen to be roughly constant, then the Cobb–Douglas aggregate production function ‘works’ in the above sense, even though the approximate constancy of factor shares *cannot* be caused by the non-existent aggregate production function. (See Fisher, Solow and Kearl, 1977 for the case of the CES

production function.)

Implications for empirical work

Empirically, the non-existence of the aggregate production function poses a conundrum. If aggregate production functions do not exist, there must be some other reason why they seem to work empirically. The answer has been in the literature for a long time (Simon and Levy, 1963; Simon, 1979; Shaikh, 1980), and more recently Felipe (2001) and Felipe and McCombie (2001; 2002; 2003; 2005; 2006a; 2006b) have elaborated upon it. (For an in-depth discussion of these issues see the papers in the *Eastern Economic Journal*, 2005.) However, like the theoretical arguments underlying the non-existence of the aggregate production function, these arguments have largely been ignored.

The argument is that, because the data used in aggregate empirical applications are not physical quantities but values, the accounting identity that relates definitionally the value of total output to the sum of the value of total inputs can be rewritten as a form that resembles a production function.

More specifically, the National Income and Products Account (NIPA) identity states that value added equals the wage bill plus total profits, that is,

$$V_t \equiv W_t + \Pi_t \equiv w_t L_t + r_t J_t \tag{1}$$

where V is real value added, W is the total wage bill in real terms, Π denotes total profits ('operating surplus', in the NIPA terminology), also in real terms, w is the average real wage rate, L is employment, r is the average *ex post* real profit rate, and J is the deflated or constant-price value of the stock of capital. (Expression (1) is an accounting identity, not the result of Euler's Theorem.) In applied aggregate work, the measures of output and capital used are the constant-price values, not physical quantities. We denote them by V and J, respectively. These are different from Y and K used above, which denoted physical quantities. The symbol $\equiv$ indicates that expression (1) is an accounting identity. Expressing the identity (1) in growth rates yields:

$$\hat{V}_t \equiv a_t \hat{w}_t + (1 - a_t)\hat{r}_t + a_t \hat{L}_t + (1 - a_t)\hat{J}_t \tag{2}$$

where ^ denotes a proportional growth rate, $a_t \equiv w_t L_t / V_t$ is the share of labour in output, and $1 - a_t \equiv r_t J_t / V_t$ is the share of capital. So far no assumption of any kind has been made.

Suppose now that factor shares in the economy are relatively stable. This could be due, for example, to the fact that firms set prices according to a mark-up on unit labour costs. Assume also that w_t and r_t

grow at constant rates. Then

$$\hat{V}_t \equiv \lambda + a\hat{L}_t + (1-a)\hat{J}_t \tag{3}$$

where $\lambda \equiv a\hat{w} + (1-a)\hat{r}$. Integrating (3) and taking antilogarithms,

$$V_t \equiv A_0 \exp(\lambda t) L_t^{a} J_t^{1-a} \tag{4}$$

Expression (4) is simply the NIPA accounting identity, expression (1), rewritten under the two assumptions mentioned above. It is certainly not a Cobb–Douglas production function, as such does not exist.

What are the implications of this argument? Suppose one estimates the standard Cobb–Douglas regression $V_t \equiv C_0 \exp(\gamma t) L_t^{\alpha_1} J_t^{\alpha_2}$ and in this economy factor shares are approximately constant and wage and profit rate growth is approximately constant. Then, this regression will yield very good results, since it approximates the identity (4). The statistical fit will be close to unity, $\alpha_1 \cong a$, $\alpha_2 \cong 1-a$, and $\gamma \cong \lambda$. However, the aggregate production function may not exist, or firms in this economy may be subject to increasing returns to scale, although the regression results might lead us to believe otherwise. On the other hand, if the assumptions about the path of the factor shares and the growth rates of *w* and *r* are incorrect, the regression $V_t \equiv C_0 \exp(\gamma t) L_t^{\alpha_1} J_t^{\alpha_2}$ will not yield good results. Felipe and Holz, 2001, showed using Monte Carlos simulations that the main reason why the Cobb–Douglas regression $V_t \equiv C_0 \exp(\gamma t) L_t^{\alpha_1} J_t^{\alpha_2}$ often fails is that the approximation of $[a_t\hat{w}_t + (1-a_t)\hat{r}_t]$ through the constant term λ is incorrect. Such widely discussed problems as unit roots or endogeneity of the regressors are not the key issues. This simply means that we have to search for better approximations to the identity. (See Felipe and McCombie, 2001; 2003, for the derivations of the CES and translog approximations to the accounting identity.)

These results have devastating implications for empirical neoclassical macro growth theory, including endogenous growth, and total factor productivity measurement and growth accounting exercises. Indeed, Felipe and McCombie (2006b) have shown using simulations that the true rate of technical progress, computed with the use of firm-level data, is very different from that obtained with the use of aggregate data. Indeed, the two measures of productivity are so far apart that it is concluded that total factor productivity growth calculated with aggregate data is in no way a proxy for the true rate of technological progress.

Why do economists continue using aggregate production functions?

Most economists are not aware of these results, but simply think of the aggregate production function as part of their basic toolkit. Others use such concepts as total productivity growth without realizing that they are assuming the existence of a non-existent construct.

Some economists, on the other hand, are aware of the aggregation results and yet continue using aggregate production functions. The reasons for doing so fall under three broad categories:

1. 1. Aggregate production functions are seen as useful parables (Samuelson, 1961–62).
2. 2. So long as aggregate production functions appear to give empirically reasonable results, why shouldn't they be used?
3. 3. For the applications where aggregate production functions are used, there is no other choice.

However, in the light of the aggregation results, none of these reasons seems valid.

Samuelson's parable argument was stated in the context of the so-called Cambridge capital theory debates. (It should not be thought that the aggregation problems have no bearing on the Cambridge–Cambridge debates. The discovery that aggregate production functions can violate properties that one expects of production functions, so-called reswitching and reverse capital-deepening, was at bottom a discovery that the aggregate concept used is not a production function at all. The aggregation problem literature shows that this was to be expected.) Samuelson showed that even in cases with heterogeneous capital goods some rationalization could be provided for the validity of the neoclassical parable, which assumes that there is a single homogenous factor referred to as capital, whose marginal product equals the interest rate. But Samuelson's results hold only in very restrictive cases, as we should expect from the aggregation literature. (See also Garegnani, 1970.)

A variation of the parable argument is that the aggregate production function should be understood as an *approximation*. It is evident that Fisher's (exact) aggregation conditions are so stringent that one can hardly believe that actual economies will satisfy them even approximately. Fisher (1969), therefore, asked: What about the possibility of a *satisfactory approximation*? Thus, suppose the values of capitals and labours in the economy lie in a bounded set and the requirement is that an aggregate production function lie within some specified distance of the true production surface for all points in the bounded set. Can this happen without the approximate satisfaction of the aggregation conditions? Fisher showed that this cannot reasonably happen by proving that the *only* way for approximate aggregation to hold without approximate satisfaction of the Leontief conditions is for the derivatives of the functions involved to wiggle violently up and down, an unnatural property not exhibited by the aggregate production functions used in practice.

The second argument is that, despite the aggregation results, neoclassical macroeconomic theory generally deals with macroeconomic aggregates derived by analogy with the micro concepts. Then, the argument goes, why not continue using them? Naturally, the aggregation problem appears in all areas of economics, including consumption theory, where a well-defined micro consumption theory exists. The neoclassical aggregate production function is also built by analogy (Ferguson, 1971).

This argument is untenable. Employing macroeconomic production functions on the unverified premise that inference by analogy is correct is inadmissible. Further, as opposed to the (already suspect) case of

the consumption function, the conditions for successful aggregation of production functions seem far more outlandish.

The third and final argument given for the use of aggregate production functions is that there is no other option if one is to answer the questions for which the aggregate production function is used, for example to discuss productivity differences across nations. But, ‘It's crooked, but it's the only wheel in town’ is not a scientific argument. The profession needs to find a different ‘wheel’.

See Also

- aggregation (theory)
- cost functions
- endogenous growth theory
- growth accounting
- neoclassical growth theory
- production functions
- total factor productivity

Bibliography

Blackorby, C., Lay, G., Nissen, D. and Russell, R. 1970. Homothetic separability and consumer budgeting. *Econometrica* 38, 468–72.

Blackorby, C. and Schworm, W. 1984. The structure of economies with aggregate measures of capital: a complete characterization. *Review of Economic Studies* 51, 633–50.

Blackorby, C. and Schworm, W. 1988. The existence of input and output aggregates in aggregate production functions. *Econometrica* 56, 613–43.

Eastern Economic Journal. 2005. Symposium on the aggregate production function. Vol. 31(3).

Felipe, J. 2001. Endogenous growth, increasing returns, and externalities: an alternative interpretation of the evidence. *Metroeconomica* 52, 391–427.

Felipe, J. and Fisher, F. 2003. Aggregation in production functions: what applied economists should know. *Metroeconomica* 54, 208–62.

Felipe, J. and Holz, C. 2001. Why do aggregate production functions work? Fisher's simulations, Shaikh's identity, and some new results. *International Review of Applied Economics* 15, 261–85.

Felipe, J. and McCombie, J. 2001. The CES production function, the accounting identity, and Occam's

razor. *Applied Economics* 33, 1221–32.

Felipe, J. and McCombie, J. 2002. A problem with some recent estimations and interpretations of the mark-up in manufacturing industry. *International Review of Applied Economics* 16, 187–215.

Felipe, J. and McCombie, J. 2003. Some methodological problems with the neoclassical analysis of the East Asian miracle. *Cambridge Journal of Economics* 54, 695–721.

Felipe, J. and McCombie, J. 2005. Why are some countries richer than others? A skeptical view of Mankiw–Romer–Weil's test of the neoclassical growth model. *Metroeconomica* 56, 360–92.

Felipe, J. and McCombie, J. 2006a. Is a theory of total factor productivity really needed? *Metroeconomica* (forthcoming).

Felipe, J. and McCombie, J. 2006b. The tyranny of the identity: growth accounting revisited. *International Review of Applied Economics* (forthcoming).

Ferguson, C. 1971. Capital theory up to date: a comment on Mrs. Robinson's article. *Canadian Journal of Economics* 4, 250–54.

Fisher, F. 1965. Embodied technical change and the existence of an aggregate capital stock. *Review of Economic Studies* 32, 263–88.

Fisher, F. 1968. Embodied technology and the existence of labor and output aggregates. *Review of Economic Studies* 35, 391–412.

Fisher, F. 1969. Approximate aggregation and the Leontief conditions. *Econometrica* 37, 457–69.

Fisher, F. 1971. Aggregate production functions and the explanation of wages: a simulation experiment. *Review of Economics and Statistics* 53, 305–25.

Fisher, F. 1982. Aggregate production functions revisited: the mobility of capital and the rigidity of thought. *Review of Economic Studies* 49, 615–26.

Fisher, F. 1983. On the simultaneous existence of full and partial capital aggregates. *Review of Economic Studies* 50, 197–208.

Fisher, F. 1993. *Aggregation: Aggregate Production Functions and Related Topics*. Cambridge, MA: MIT Press.

Fisher, F., Solow, R. and Kearl, J. 1977. Aggregate production functions: some CES experiments.

Review of Economic Studies 44, 305–20.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37, 407–36.

Gorman, W. 1953. Community preference fields. *Econometrica* 21, 63–80.

Gorman, W. 1959. Separable utility and aggregation. *Econometrica* 27, 469–81.

Gorman, W. 1968. Measuring the quantities of fixed factors. In *Value, Capital, and Growth: Papers in Honour of Sir John Hicks*, ed. J. Wolfe. Edinburgh: Edinburgh University Press.

Hicks, J. 1939. *Value and Capital*. Oxford: Clarendon Press.

Houthakker, H. 1955–56. The Pareto distribution and the Cobb–Douglas production function in activity analysis. *Review of Economic Studies* 23, 27–31.

Klein, L. 1946a. Macroeconomics and the theory of rational behavior. *Econometrica* 14, 93–108.

Klein, L. 1946b. Remarks on the theory of aggregation. *Econometrica* 14, 303–12.

Leontief, W. 1936. Composite commodities and the problem of index numbers. *Econometrica* 4, 39–59.

Leontief, W. 1947a. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15, 361–73.

Leontief, W. 1947b. A note on the interrelationship of subsets of independent variables of a continuous function with continuous first derivatives. *Bulletin of the American Mathematical Society* 53, 343–50.

Levhari, D. 1968. A note on Houthakker's aggregate production function in a multi-firm industry. *Econometrica* 36, 151–4.

May, K. 1946. The aggregation problem for a one-industry model. *Econometrica* 14, 285–98.

May, K. 1947. Technological change and aggregation. *Econometrica* 15, 51–63.

Nataf, A. 1948. Sur la possibilité de construction de certains macromod•les. *Econometrica* 16, 232–44.

Samuelson, P. 1961–62. Parable and realism in capital theory: the surrogate production function. *Review of Economic Studies* 29, 193–206.

Sato, K. 1975. *Production Functions and Aggregation*. Amsterdam: North-Holland.

Shaikh, A. 1980. Laws of production and laws of algebra: Humbug II. In *Growth, Profits and Property: Essays in the Revival of Political Economy*, ed. E. Nell. Cambridge: Cambridge University Press.

Simon, H. 1979. On parsimonious explanations of production relations. *Scandinavian Journal of Economics* 81, 459–74.

Simon, H. and Levy, F. 1963. A note on the Cobb–Douglas function. *Review of Economic Studies*, 93–4.

Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40, 549–63.

Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6, 345–54.

How to cite this article

Felipe, Jesus and Franklin M. Fisher. "aggregation (production)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000059>
doi:10.1057/9780230226203.0019

The New Palgrave Dictionary of Economics Online

aggregation (theory)

Werner Hildenbrand
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The aim of aggregation theory is to link the micro and macroeconomic notions of aggregate demand. One would like such a link to exist for any heterogeneous population, for a large set of all conceivable income assignments, and for a small number of statistics of the income distribution. This cannot be achieved. What can be achieved is critically discussed in Section 2. In Section 3, another important topic of aggregation theory is considered: how does mean demand react to price changes? As an example, the ‘law of demand’ is discussed.

Keywords

aggregate demand; aggregation; behavioural heterogeneity; exact income aggregation; law of demand; monotonicity; revealed preferences; Slutzky substitution effect

Article

1 Introduction

Aggregation theory of demand aims at identifying observable explanatory variables for aggregate demand starting from a microeconomic description of the underlying population of households. In the simple case, where the demand decision of a household is the choice of a commodity vector in a budget set, which is determined by the price vector p and income x (total expenditure), the demand behaviour of a household h is modelled by a demand function $f^h(p, x) \in \mathbb{R}^l_+$ (commodity space), which is defined for every strictly positive price vector $p \in P$ and every income level $x \geq 0$. The demand function f^h might, but need not be derived from preference maximization under the budget constraint.
Aggregate demand is defined as *mean demand across the population H*, that is to say, $\frac{1}{\#H} \Sigma_{h \in H} f^h(p, x^h)$. The population H is viewed as heterogeneous in income and demand behaviour. Thus, mean demand is determined by the price vector p and the joint distribution of income x^h and demand function f^h across the population H.
This general microeconomic definition of mean demand is sufficiently specific for certain problems in pure theory, for example for the existence problem in general equilibrium theory.
In macroeconomics or in applied demand analysis the notion of aggregate demand is quite different. There the explanatory variables for aggregate demand are the price vector and certain statistics $S(G_x)$ of the income distribution function G_x such as mean income, a measure of income inequality (for example, the variance of log income) or higher moments of the income distribution. In any case, no household specific variable is used in the aggregate demand function. The aim of the aggregation theory is to link the micro and macroeconomic notions of aggregate demand. More specifically, given an assignment $(f^h)_{h \in H}$ of demand functions and a set $\mathcal{X} \subset \mathbb{R}^H_+$ of income assignments $(x^h)_{h \in H}$, one seeks for a representation of mean demand of the following form: there exists a function F from $P \times \mathbb{R}^N$ into $\mathbb{R}^l_+$ and N statistics $S_1(G_x), \ldots, S_N(G_x)$ of the income distribution function G_x, such that

$$\frac{1}{\#H} \sum_{h \in H} f^h(p, x^h) = F(p, S_1(G_x), \ldots, S_N(G_x))$$

(1)

for all income assignments $(x^h)_{h\in H}$ in $\mathcal{X}$ *and all* price vectors p in P.

One would like such a representation to exist for any heterogeneous population H, for a large set $\mathcal{X}$, ideally for all conceivable income assignments, that is, $\mathcal{X} = \mathbb{R}_+^H$ and for a small number N of statistics. This, of course, cannot be achieved.
The theory of income aggregation is surveyed in Section 2, where also basic references are given. The main results are:

- a representation of the form (1), which must hold in the case $\mathcal{X} = \mathbb{R}_+^H$ is an unreasonable strong requirement. Indeed, if a representation exists, then the population H must be homogeneous in demand behaviour, that is, $f^h = f$ for all $h \in H$, and furthermore
- if N is less than the number of households in H and the common demand function f has the basic properties of demand theory (budget identity and homogeneity), then either f is linear in income or at least for one commodity i, the income share function $w_i(p, x) := p_i f_i(p, x) / x$ is oscillating (that is, the derivative $\partial_x w_i(p, \cdot)$ changes its sign infinitely often).

Thus, households' behaviour which is modelled by the common demand function is either unreasonably simple or incredibly sophisticated. These results clearly show that the requirement $\mathcal{X} = \mathbb{R}_+^H$ leads to an ill-posed problem.

For a heterogeneous population H there exists (see Example 3) a finite partition $\{\mathcal{X}_k\}_{k\in K}$ of the set $\mathbb{R}_+^H$ of all conceivable income assignments and for every $k \in K$ there is a function $F^k(p, G)$, where G denotes an income distribution function, such that

$$\frac{1}{\# H} \sum_{h\in H} f^h(p, x^h) = F^k(p, G_x)$$

(2)

for every income assignment $(x^h)_{h\in H}$ in the set $\mathcal{X}^k$ and for every $p \in P$.

Thus, for a heterogeneous population H, there is no closed-form definition of an aggregate demand function; there is only a piecewise one, since the aggregate demand functions F^k and F^j are different for $k \neq j$. The less heterogeneous the population the coarser the partition, that is, the smaller is $\#K$. The sets $\mathcal{X}^k$ of the partition are large (see Example 3), in particular, if $(x_0^h) \in \mathcal{X}^k$, then for every strictly increasing function φ the income assignment $x^h = \phi(x_0^h)$, $h \in H$, also belongs to $\mathcal{X}^k$ (see Figures 3 and 4).

The aggregate demand functions $F^k(p, G)$ in (2) require the knowledge of the entire income distribution. In many applications one might assume that the distribution of relevant income assignments in the set $\mathcal{X}^k$ can be modelled by some few parameters (structural stability of income distributions). For example, if the population is 'very large' one might restrict attention to those (x^h) in $\mathcal{X}^k$ whose distributions are (approximately) log normal. Then, on this subset of $\mathcal{X}^k$, mean demand has a representation of the form $F^k(p, \bar{x}, \sigma)$, where $\bar{x}$ denotes mean income across the population and σ^2 is the variance of log income, which can be interpreted as a measure of income inequality.
Another important topic of aggregation theory is to analyse how mean demand of a heterogeneous population reacts to price changes under the ceteris paribus clause that households' income and demand functions remain fixed. In this case mean demand is denoted by $F(p)$. Among the various desirable dependence structures is certainly the 'law' of demand, which asserts that the vector $\Delta p \in \mathbb{R}^l$ of price changes and the resulting vector $\Delta F \in \mathbb{R}^l$ of mean demand changes point in opposite directions, that is, the scalar product $\Delta p \cdot \Delta F := \sum_{i=1}^{l} \Delta p_i \Delta F_i$ is negative.
Certainly, the 'law' is not meant to be an empirical law, but a monotonicity property of the mean demand function $F(p)$ which is defined under a ceteris paribus clause in a mathematical model of a population of households. Thus, the 'law' asserts that the mean demand function F is *strictly monotone*, that is,

$$(p - q) \cdot (F(p) - F(q)) < 0 \text{ for all } p \neq q \text{ in } P.$$

In particular, every partial mean demand curve is strictly decreasing. This partial monotonicity property, however, is not sufficient for proving the uniqueness and stability of the equilibria for a multi-commodity demand-supply system; one needs strict monotonicity in the multi-commodity version.

Which behavioural assumption on the household level and/or which form of heterogeneity of the population lead to monotone mean demand? To answer this question one assumes that demand functions f^h satisfy the weak axiom of revealed preferences or, more specifically, that they are derived from preference maximization. Then, partial monotonicity is easily obtained, for example, by excluding inferior goods. However, multi-commodity monotonicity is more difficult to obtain. Trivially, mean demand is monotone if all demand functions $f^h(p, x^h)$ were monotone in p. This, however, requires that either $f^h(p, \cdot)$ is linear in income or that the Slutzky substitution effect is sufficiently strong. (For a precise formulation, see the Theorem of Mitjuschin and Polterovich, 1978; law of demand.) Since the Slutzky substitution effect might be arbitrarily small, one is interested in finding alternative assumptions, which do not rely on a strong Slutzky substitution effect. These assumptions should not require that households' demand functions are monotone. Obviously, to obtain the desirable aggregation effect, the population must be heterogeneous. Thus, in contrast to the problem of income aggregation, heterogeneity does not complicate the analysis, yet it is necessary to obtain monotonicity of mean demand by aggregation. More details are given in Section 3. For example, let H be a population which is homogeneous in demand behaviour, that is, $f^h = f$, $h \in H$ and the common demand function is not monotone. However, the population is heterogeneous in income. Then, for a given income assignment $(x^h)_{h \in H}$, mean demand $F^H(p)$ is not monotone in p. If one increases now the population size, that is, the number $\#H$ of households tends to infinity and if for increasing $\#H$ the income distribution functions G^H of households in H converge to a concave distribution function G, then, for $\#H$ sufficiently large, mean demand $F^H(p)$ is 'approximately' monotone, that is to say, $F^H(p)$ converges to a monotone function. Consequently, in the limit, that is, for an indefinitely large population which admits a concave income distribution function, mean demand is monotone. The mathematical model for such a limit population cannot be a finite or countably infinite set; it must be an atomless measure space of households, for example, the unit interval [0,1] with Lebesgue measure (continuum of households).

If these large populations are heterogeneous in income and demand behaviour, then one can meaningfully pose the problem of 'smoothing by aggregation': is mean demand continuous or differentiable without assuming these properties on the household level? The basic reference is Trockel (1984).

Finally, one should mention the literature on 'behavioural heterogeneity' initiated by Grandmont (1992). Here the goal is to obtain a stronger property than strict monotonicity of mean demand: diagonal dominance of the Jacobian $\partial_p F(p)$ of mean demand in the sense that

$$p_i \partial_{p_i} F_i(p) > \sum_{j \neq i} p_j |\partial_{p_j} F_i(p)|$$

and

$$p_i \partial_{p_i} F_i(p) > \sum_{j \neq i} p_j |\partial_{p_i} F_j(p)|.$$

This diagonal dominance models a strong restriction on the interdependence among the various commodity markets and is the basis for partial equilibrium analysis. For a general discussion of 'behavioural heterogeneity' see Hildenbrand and Kneip (2005).

2 Income aggregation

The demand behaviour of every household h in a population H is modelled by a demand function $f^h \in \mathcal{F}$. In this section it is not required that demand functions are derived by preference maximization under budget constraints. One only needs that demand functions $f \in \mathcal{F}$ are continuous functions from $P \times \mathbb{R}_+$ into $\mathbb{R}^l_+$ with $f(p, 0) = 0$, where P denotes the set of all strictly

positive price vectors in $\mathbb{R}^l$.

For every income assignment $(x^h)_{h\in H}$, $x^h \geq 0$, we consider *mean demand* $\frac{1}{\#H}\sum_{h\in H} f^h(p, x^h)$. The 'problem of income aggregation' has been defined in the literature by the qst: does there exist a function F from $P \times \mathbb{R}_+$ into $\mathbb{R}^l_+$ such that

$$\frac{1}{\#H}\sum_{h\in H} f^h(p, x^h) = F(p, \bar{x}), \text{ where } \bar{x} = \frac{1}{\#H}\sum_{H} x^h, \tag{3}$$

for all income assignments in a given set $\mathcal{X} \subset \mathbb{R}^H_+$ and all $p\in P$?

If one asks this question for all conceivable income assignments, that is, $\mathcal{X} = \mathbb{R}^H_+$, then this is an ill-posed problem since it allows only a trivial solution.

Theorem: (Antonelli, 1886): *There exists a function* $F(p, \bar{x})$ *such that (3) holds on* $\mathbb{R}^H_+ \times P$ *if and only if the population H is homogeneous in demand behaviour, that is,* $f^h = f$, *and furthermore* $f(p, x)$ *is linear in x, that is,* $f(p, x) = \alpha(p)x$, $\alpha(p) \in \mathbb{R}^l_+$. *Thus,* $F(p, \bar{x}) = \alpha(p)\bar{x}$.

One might ask whether a less restrictive condition than (3) allows for a nontrivial solution. That is to say, one might consider mean demand functions that depend on a wider set of aggregate income variables than just mean income, for example, the variance or higher moments of the distribution of income. The answer is definitely negative.

For every income assignment $(x^h)_{h\in H}$, let G_x denote its distribution function, that is,

$$G_x(\xi) := \frac{1}{\#H}\#\left\{h\in H \middle| x^h \leq \xi\right\}, \ \xi\in\mathbb{R}.$$

Proposition 1: There exists a function $F(p, G)$ *such that*

$$\frac{1}{\#H}\sum_{h\in H} f^h(p, x^h) = F(p, G_x) \tag{4}$$

for all conceivable income assignments, that is, $\mathcal{X} = \mathbb{R}^H_+$ *and all* $p\in P$, *if and only if the population H is homogeneous in demand behaviour, that is, all households in H have the same demand function. Then* $F(p, G_x) = \int f(p, \xi)\, dG_x(\xi)$.

Proof: Consider any two households k and j in H, and an income assignment $(x^h)_{h\in H}$ with $x^k > 0$ and $x^j = 0$. Now one interchanges the income of households k and j. This does not change the distribution function of income. Hence property (4) and the fact that $f^k(p, 0) = f^j(p, 0) = 0$ implies that $f^k(p, x^k) = f^j(p, x^k)$. Since this holds for all $x^k > 0$ and $p\in P$ one obtains $f^k = f^j$. On the other hand, if $f^h = f$ for all $h\in H$, then $\frac{1}{\#H}\sum_{h\in H} f^h(p, x^h) = \int f(p, x)\, dG_x =: F(p, G_x)$.

The justification for considering the generalized problem of income aggregation as defined by (4) is based on the view that for large populations, which this survey emphasizes, income distribution functions can often be modelled by some few parameters, for example, log-normal distributions.

By Proposition 1 it is clear that one is forced to restrict the set $\mathcal{X}$ of admissible income assignments if one wants to escape the case of trivial solutions, $f^h = f$, to the aggregation problem as

defined by (4). Motivated by the special role which zero income and the assumption $f(p, 0) = 0$ play in the proofs of Antonelli's Theorem or Proposition 1 one has considered in the literature (for example, Nataf, 1948, or Gorman, 1953) a restriction on the domain of individual income:

$$x(a, b) := \left\{(x^h) \in \mathbb{R}_+^H \mid 0 < a \leq x^h \leq b \leq \infty\right\}, \; a < b.$$

Proposition 2 shows that this restriction allows merely for some very limited and quite special heterogeneity in demand behaviour of the population *H*.
Proposition 2:

1. 1. *There exists a function* $F(p, G)$ *such that (2) holds on* $x(a, b) \times P$ *if and only if for every commodity i and p∈P the income expansion paths* $f_i^h(p, \cdot)$, $h \in H$, *are parallel (vertically) on the interval (a,b); (with differentiability)* $\partial_x f_i^h(p, x)$ *does not depend on h∈H (*Figure 1*).*
   Figure 1

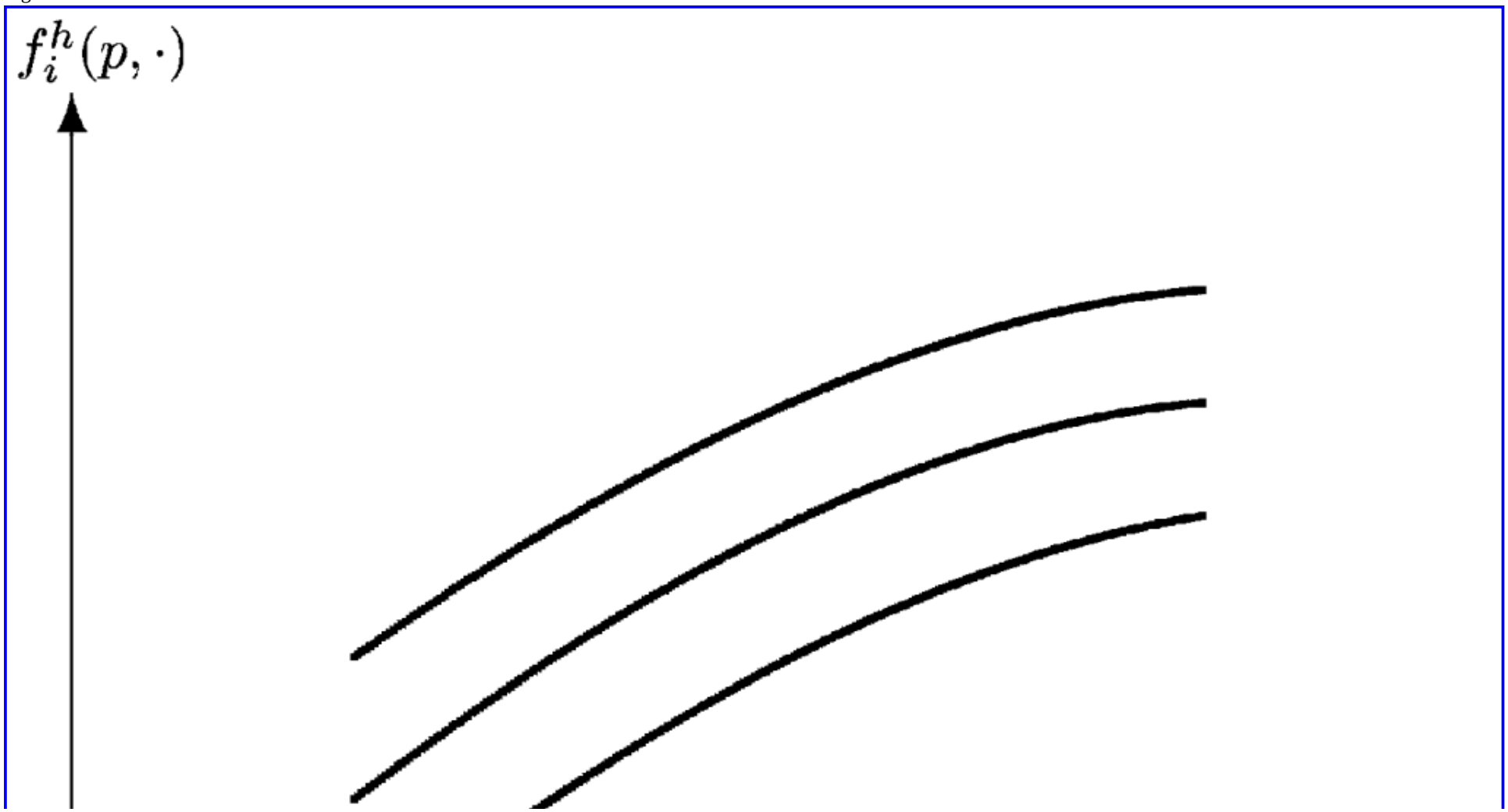

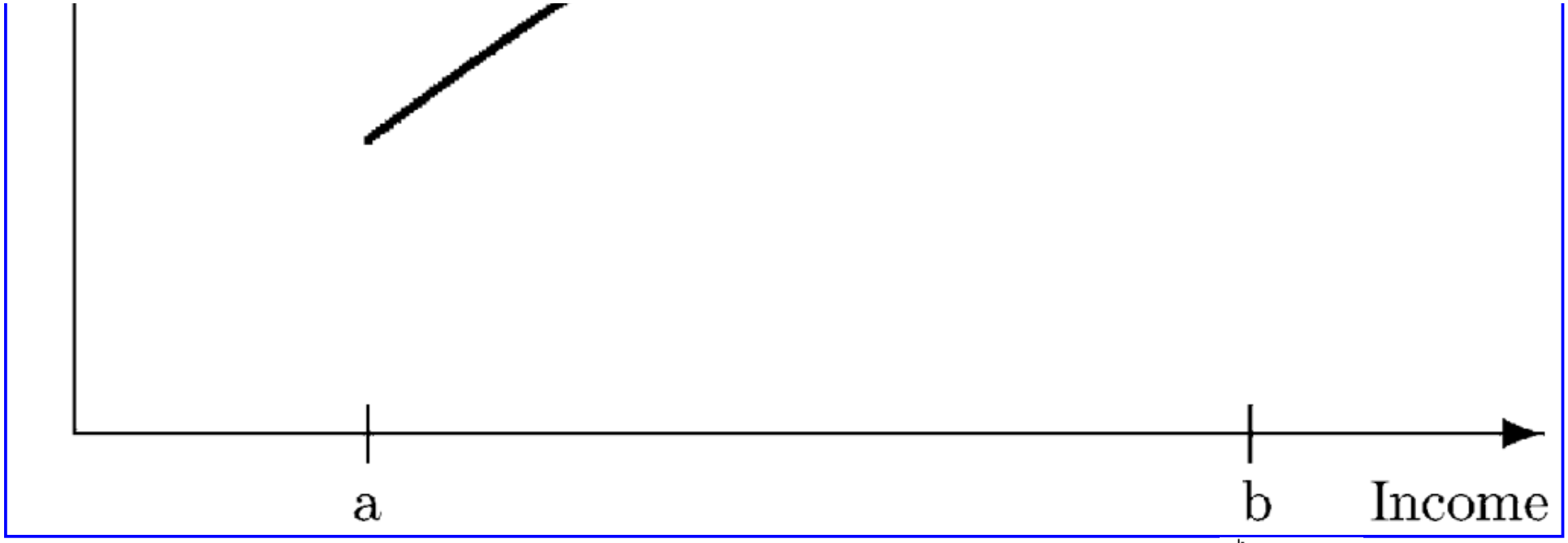

2. 2. *There exists a function* $F(p, \bar{x})$ *such that (1) holds on* $x(a, b) \times P$ *if and only if for every commodity i and p∈P the income expansion paths* $f_i^h(p, \cdot), h \in H$, *are affine and parallel on the interval (a,b); (with differentiability)* $\partial_x f_i^h(p, x)$ *does not depend on h∈H and x∈(a,b) (Figure 2).*

Figure 2

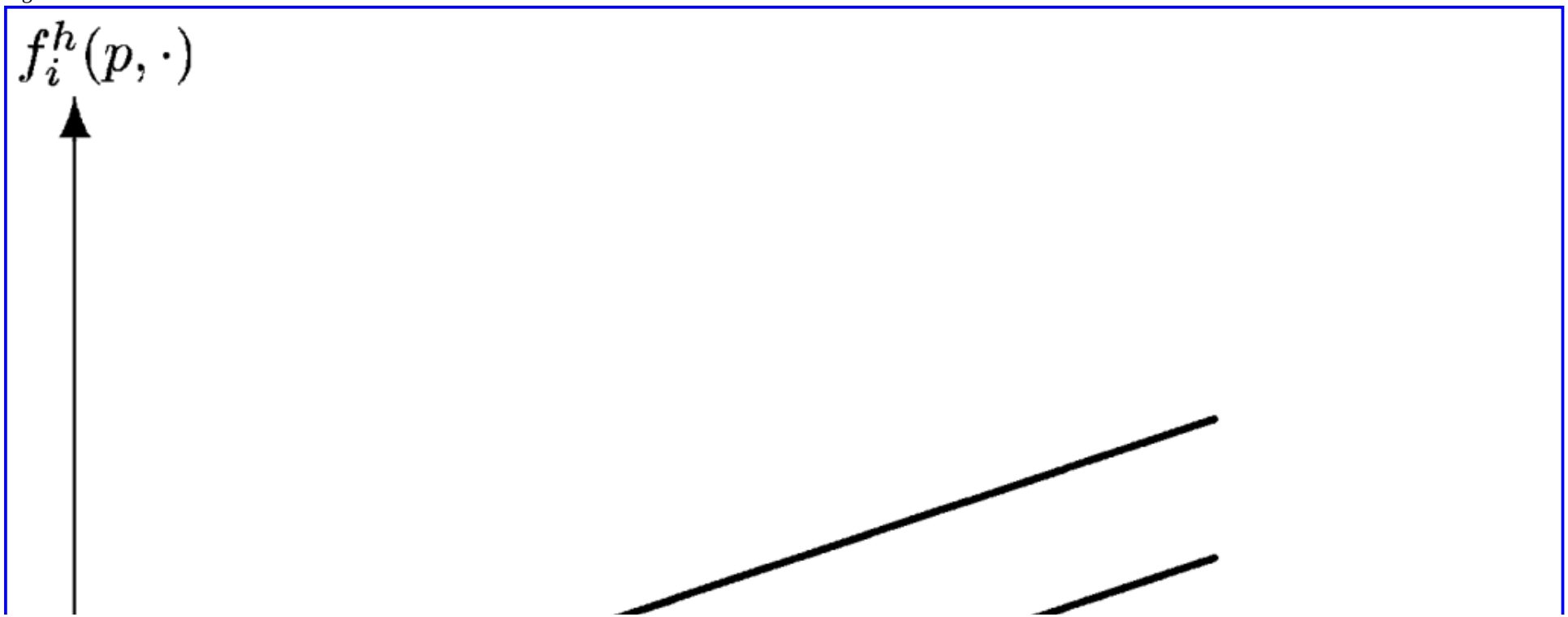

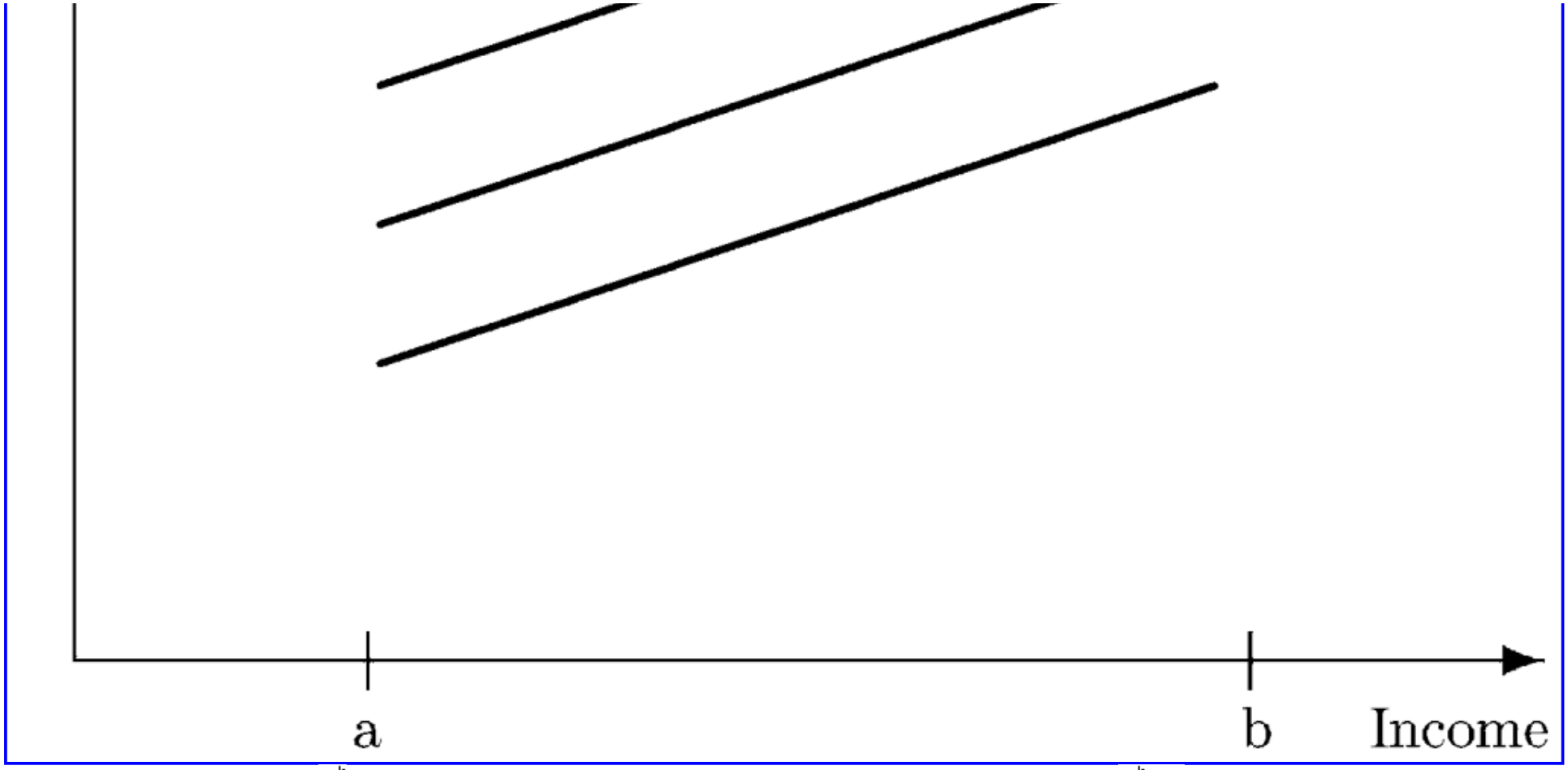

3. 3. *If all individual demand functions* f^h *belong to* $\mathcal{F}$ *and are homogeneous in (p,x), then the necessary condition in (i) implies that* $f^h \equiv f$.

Proof:

1. (i) Consider any two households *k* and *j* in *H* and an income assignment in $\mathcal{X}(a, b)$ with $x^k \neq x^j$. Now one interchanges the income of households *k* and *j*. This does not change the income distribution function. Hence, property (2) implies $f^k(p, x^k) + f^j(p, x^j) = f^k(p, x^j) + f^j(p, x^k)$. Thus $f^k(p, x^k) - f^k(p, x^j) = f^j(p, x^k) - f^j(p, x^j)$. Since it holds for all $x^k, x^j \in (a, b)$ and all $p \in P$ one obtains the claimed property in (i). The converse is trivial.
2. (ii) Instead of interchanging the income of households *k* and *j* one chooses $x^k + \Delta$ and $x^j - \Delta \in (a, b)$ for sufficiently small Δ. Property (1) then implies

$$f^k(p, x^k + \Delta) - f^k(p, x^k) = f^j(p, x^j) - f^j(p, x^j - \Delta) = f^k(p, x^j) - f^k(p, x^j - \Delta)$$

   by (i), which implies the claimed property in (ii). The converse is trivial.
3. (iii) If the expansion paths $f_i^k(p, \cdot)$, $h \in H$, are parallel on (*a*,*b*) for every $p \in P$, then homogeneity implies that they are also parallel on ($\lambda a, \lambda b$) for all $\lambda > 0$ and $p \in P$. Hence they

are parallel on $(0,\infty)$ for all $p \in P$. Continuity and $f^h(p, 0) = 0$ then implies the claim.

An alternative approach to allow for heterogeneous populations consists of considering, in addition to income, further explanatory variables for household demand. For example, in applications it is standard practice to stratify the whole population H by a certain profile $a = (a_1, a_2, \ldots)$ of observable household attributes, such as household size, age of household head, etc. Let $H(a)$ denote the sub-population of all households in H with attribute profile a. Without loss of generality one can assume that $a \in \mathbb{R}^m$. Let $G_{x,a}$ denote the joint distribution of function of x^h, a^h across H. Analogously to Proposition 1 one shows

Proposition 1 ' : There exists a function $F(p, G_{x,a})$ *such that*

$$\frac{1}{\# H} \sum_{h \in H} f^h(p, x^h) = F(p, G_{x,a})$$

for all conceivable income-attribute assignments and all $p \in P$ *if and only if all sub-populations H(a) are homogeneous in demand behaviour, that is,* $f^h = f^a$ *for all* $h \in H(a)$.

Thus, the whole population need not be homogeneous, yet the joint distribution of x^h and a^h across H has typically a complex dependence structure, and hence, it cannot be modelled by some few parameters, as in the case of income.

Exact income aggregation

In the literature on 'exact income aggregation', as initiated by Gorman (1953), Lau (1982), and Jorgensen, Lau and Stoker (1982), one seeks for a representation of mean demand which is less restrictive than (3), yet more demanding than (4), that is to say, $\frac{1}{\# H} \Sigma_{h \in H} f^h(p, x^h) = F(p, S_1(G_x), \ldots, S_N(G_x))$ on $\mathbb{R}_+^H \times P$ for some continuous function F from $P \times \mathbb{R}^N$ into $\mathbb{R}^l$ (the commodity space) and some vector of distributional statistics $S_1(G_x), \ldots, S_N(G_x)$ with $N < \# H$. This representation is more demanding than (4); it does not require the knowledge of the entire income distribution since $N < \# H$.

If such a representation exists, then by Proposition 1, $f^h = f$, $h \in H$, and f is called 'exactly aggregable'. Thus, the question is whether there are exactly aggregable demand functions which are not linear in income and satisfy the basic restrictions of demand theory?

To simplify the presentation one assumes that all distributional statistics are 'generalized moments', that is, $S_n(G_x) = \int s_n(\xi)\, dG_x(\xi)$, with continuous functions $s_n(\cdot)$. Without loss of generality one can require that $s_n(0) = 0$.

Proposition 3: There exists a representation of mean demand of the form

$$\int f(p, \xi)\, dG_x(\xi) = F(p, \int s_1(\xi)\, dG_x(\xi), \ldots, \int s_N(\xi)\, dG_x(\xi)), \tag{5}$$

which holds for every income distribution function G_x *of every finite population H and every price vector in P if and only if the function f is of the form*

$$f(p, \xi) = \alpha_1(p) s_1(\xi) + \cdots \alpha_N(p) s_N(\xi), \quad p \in P \text{ and } \xi \in \mathbb{R}_+, \tag{6}$$

where $\alpha_n(p) \in \mathbb{R}^l$.

Proof: Trivially, (6) implies (5). Assume that (5) holds. Let $\mathcal{G}$ denote the set of all income distribution functions for every finite population. Note that for every $G^1, G^2 \in \mathcal{G}$ and any rational λ with $0 \leq \lambda \leq 1$ it follows that $G^\lambda = \lambda G^1 + (1-\lambda) G^2 \in \mathcal{G}$.

The representation (5) implies for every commodity *i*

$$f_i(p, \xi) = F_i(p, s_1(\xi), \ldots, s_N(\xi)), \quad p \in P \text{ and } \xi \in \mathbb{R}_+. \tag{7}$$

Now one shows that the function $F_i(p, \cdot)$ has a 'linear structure' on its relevant domain $\mathcal{D} := \left\{ y \in \mathbb{R}^N \mid y_n = \int s_n(\xi)\, dG(\xi),\ G \in \mathcal{G},\ n = 1, \ldots, N \right\}$, that is,

$$F_i(p, \lambda y^1 + (1-\lambda) y^2) = \lambda F_i(p, y^1) + (1-\lambda) F_i(p, y^2) \tag{8}$$

for every $y^1, y^2 \in \mathcal{D}$ and any rational λ with $0 \leq \lambda \leq 1$. Indeed, $y_n^k = \int s_n(\xi)\, dG^k(\xi)$, k=1,2 for some $G^1, G^2 \in \mathcal{G}$. Let $G^\lambda = \lambda G^1 + (1-\lambda) G^2$. Then $\int s_n(\xi)\, dG^\lambda(\xi) = \lambda \int s_n(\xi)\, dG^1(\xi) + (1-\lambda) \int s_n(\xi)\, dG^2(\xi)$. Hence $\lambda y^1 + (1-\lambda) y^2 \in \mathcal{D}$ since $G^\lambda \in \mathcal{G}$ for rational λ. Consequently, the closure $\overline{\mathcal{D}}$ of $\mathcal{D}$ is convex. Since $G^\lambda \in \mathcal{G}$, one obtains from (5)

$$\int f_i(p, \xi)\, dG^\lambda(\xi) = F_i\left(p, \int s_1(\xi)\, dG^\lambda(\xi), \ldots, \int s_N(\xi)\, dG^\lambda(\xi)\right) = F_i(p, \lambda y^1 + (1-\lambda) y^2).$$

The left hand is equal to $\lambda \int f_i(p, \xi)\, dG^1(\xi) + (1-\lambda) \int f_i(p, \xi)\, dG^2(\xi) = \lambda F_i(p, y^1) + (1-\lambda) F_i(p, y^2)$ by (5), which proves (8). Since F_i is continuous, the 'linear structure' (8) also holds for any y^1, y^2 in the closure $\overline{\mathcal{D}}$ of $\mathcal{D}$ and any λ with $0 \leq \lambda \leq 1$. Since $s_n(0) = 0$ and $f(p, 0) = 0$ it follows from (7) that $F_i(p, 0) = 0$. Consequently, by (8), the restriction of the function $F_i(p, \cdot)$ on the convex domain $\overline{\mathcal{D}}$ can be extended to a function $\tilde{F}_i(p, \cdot)$, which is linear in *y*, that is, $\tilde{F}_i(p, y) = \alpha_1^i(p) y_1 + \cdots \alpha_N^i(p) y_N$. Thus (7) implies (6). The extension is unique if the dimension of the convex domain $\overline{\mathcal{D}}$ is equal to *N*.

Remark: The proof of Proposition 3 is quite simple since it was assumed that the representation (5) must hold for all income distribution functions for *all* finite populations. This case is also treated in Heineke and Shefrin (1988), their proof, however, requires differentiability. If one only requires (5) to hold for all income distribution functions of a *given* population *H* with $N < \# H$, then it is much more difficult to obtain (6). See Lau (1982) and Heineke and Shefrin(1988).

Note that the global structural specification (6) is very restrictive if the demand function $f \in \mathcal{F}$ has the basic properties of static demand theory. In fact, Heineke and Shefrin (1987) show the following result: *if* $f \in \mathcal{F}$ *satisfies the budget-identity, is homogeneous in p and x and if no budget share function* $w_i(p, x) := p_i f_i(p, x) / x$ *is oscillating (that is, the derivative* $\partial_x w_i(p, x)$ *changes infinitely often its sign), then (6) implies* $f(p, x) = \alpha(p) x$.

Indeed, if $f \in \mathcal{F}$ satisfies the budget identity, then $0 \leq w_i(p, x) \leq 1$. Let the budget share function $w_k(\overline{p}, \cdot)$ be non-constant and non-oscillating. Consider the function $\varphi_\lambda(\cdot), \lambda > 0$, defined by $\varphi_\lambda(x) = w_k(\overline{p}, \lambda x)$, and the linear function space which is generated by all functions $\varphi_\lambda(\cdot), \lambda > 0$. Heineke and Shefrin (1987) argue that the dimension of this linear space is infinite. By homogeneity, $\varphi_\lambda(x) = w_k(\overline{p} / \lambda, x)$; thus, the linear space $\mathcal{L}$ which is generated by all budget share functions $w_k(p, \cdot)$, $p \in P$ has infinite dimension. Consequently, the demand function *f* cannot satisfy (6), since (6) implies that $\dim \mathcal{L} \leq N$. Thus, if *f* satisfies (6) and $w_k(\overline{p}, \cdot)$ is non-oscillating, then it must be constant, that is, $f_k(\overline{p}, \cdot)$ is linear.

As a consequence, for demand functions which have the basic properties of atemporal demand theory including non-oscillating budget share functions, one either has to be satisfied with a representation as in Proposition 1 or one is in the trivial case of Antonelli's Theorem.

Heterogeneous populations

The representations (3), (4), and (5) of mean demand which have been considered up to now imply that the population of households must be homogeneous in demand behaviour, that is, $f^h \equiv f, \ h \in H$. The reason for this unsatisfactory fact is due to the very strong requirement that the representations must hold for every conceivable income assignment. This is more demanding than is needed in many applications, since there, changes in individual income are not entirely arbitrary; they might be the result of an underlying process. This point was emphasized by Malinvaud (1956) and (1993). To capture this idea, one starts from an initial income assignment (x_0^h) (status quo), and then one considers a sequence $(x_n^h), \ n = 1, 2, \ldots$ or a set $\mathcal{X}(x_0)$ of income assignments which are viewed as the result of the underlying (unspecified) process. Which properties must the sequence (x_n^h) or the set $\mathcal{X}(x_0)$ have such that for any assignment of demand functions f^h the representations of mean demand hold along this sequence or on the set $\mathcal{X}(x_0)$?
We give three examples. The first one is well-known. The second and third example generalize substantially the first one.
Example 1: Fixed income shares
Starting from an initial income assignment (x_0^h), one defines the set $\mathcal{X}(\delta) \subset \mathbb{R}_+^H$ of income assignments

$$\mathcal{X}(\delta) := \left\{ (x^h) \in \mathbb{R}_+^H \middle| x^h / \bar{x} = x_0^h / \bar{x}_0 =: \delta^h \right\}.$$

where $\bar{x}$ denotes mean income across H.
Given any assignment of demand functions f^h, $h \in H$, there exists a function F from $P \times \mathbb{R}_+$ into $\mathbb{R}_+^l$ such that mean demand has the representation

$$\frac{1}{\# H} \sum_{h \in H} f^h(p, x^h) \equiv F(p, \bar{x}) \text{on } \mathcal{X}(\delta) \times P. \tag{9}$$

The function *F* is defined by $F(p, \bar{x}) = \frac{1}{\# H} \Sigma_{h \in H} f^h(p, \delta^h \bar{x})$. If all f^h are linear in income then $F(p, \bar{x})$ is linear in mean income $\bar{x}$. Moreover, Eisenberg (1961) and Chipman and Moore (1979) have shown: if all f^h are generated by a utility function homogeneous of degree one then $F(p, \bar{x})$ is also generated by a utility function homogeneous of degree one given by

$$u(z) = \max_{z^h \in \mathbb{R}_+^l, \sum_H z^h = z} \prod_{h \in H} (u^h(z^h))^{x_0^h / \bar{x}_0}$$

Example 2: Rank preserving income changes
Starting from an initial income assignment $(x_0^h)_{h \in H}$ one defines the set $\mathcal{X}(x_0) \subset \mathbb{R}_+^H$ of income assignments (x^h) which have the property that every household keeps his rank position of income, that is, if for two households *j* and *k*, $x_0^j = x_0^k$ then $x^j = x^k$ and if $x_0^j < x_0^k$ then $x^j < x^k$. For any (x_1^h) and (x_2^h) in $\mathcal{X}(x_0)$ there is a strictly increasing function φ such that

$\phi(x_1^h) = x_2^h$, $h \in H$. Examples for $\varphi(\cdot)$ are given in Figure 3 (low income is increased, high income is decreased) and Figure 4 (low and high incomes are decreased, middle ones increased) below.

Figure 3

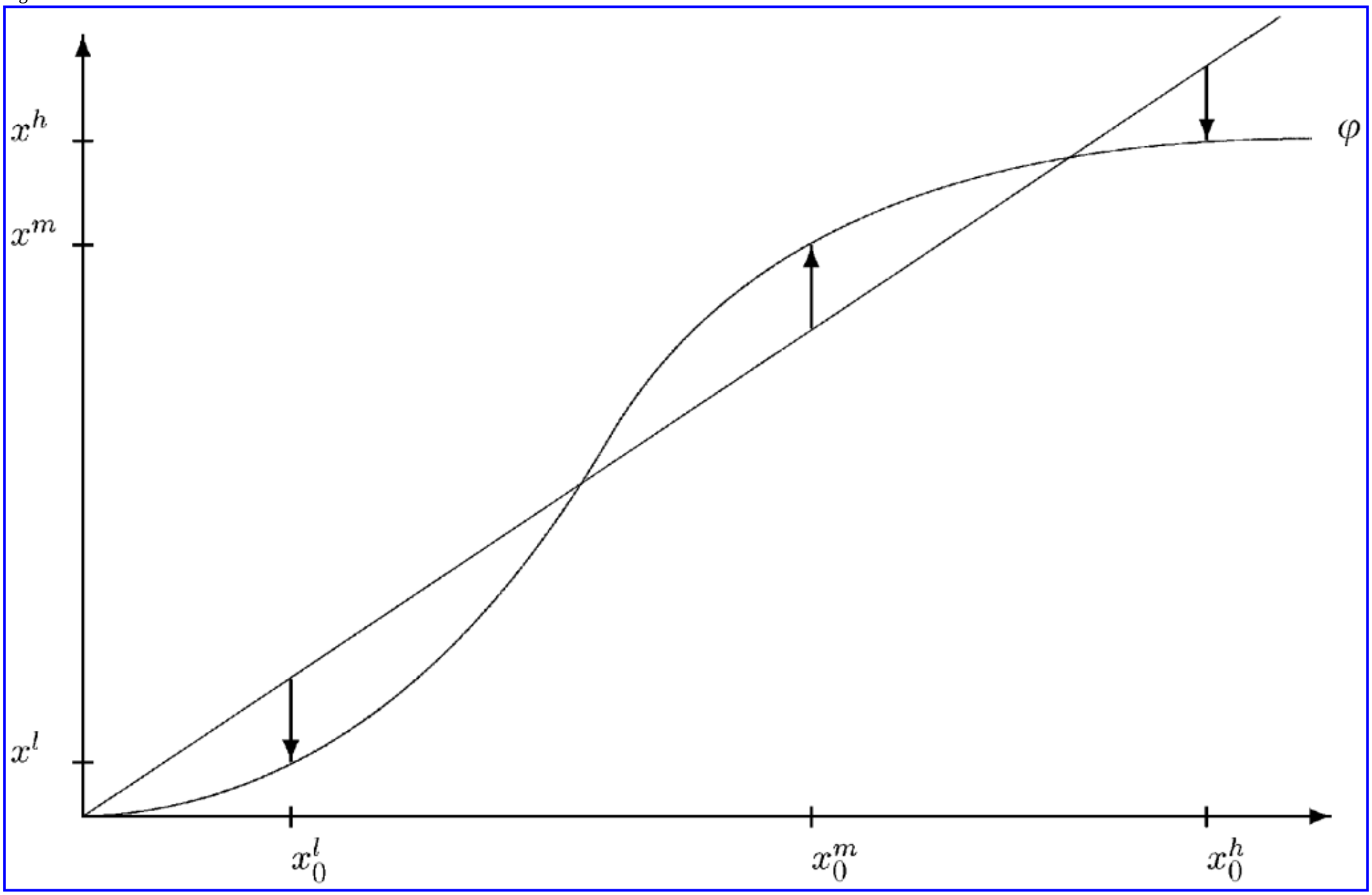

Figure 4

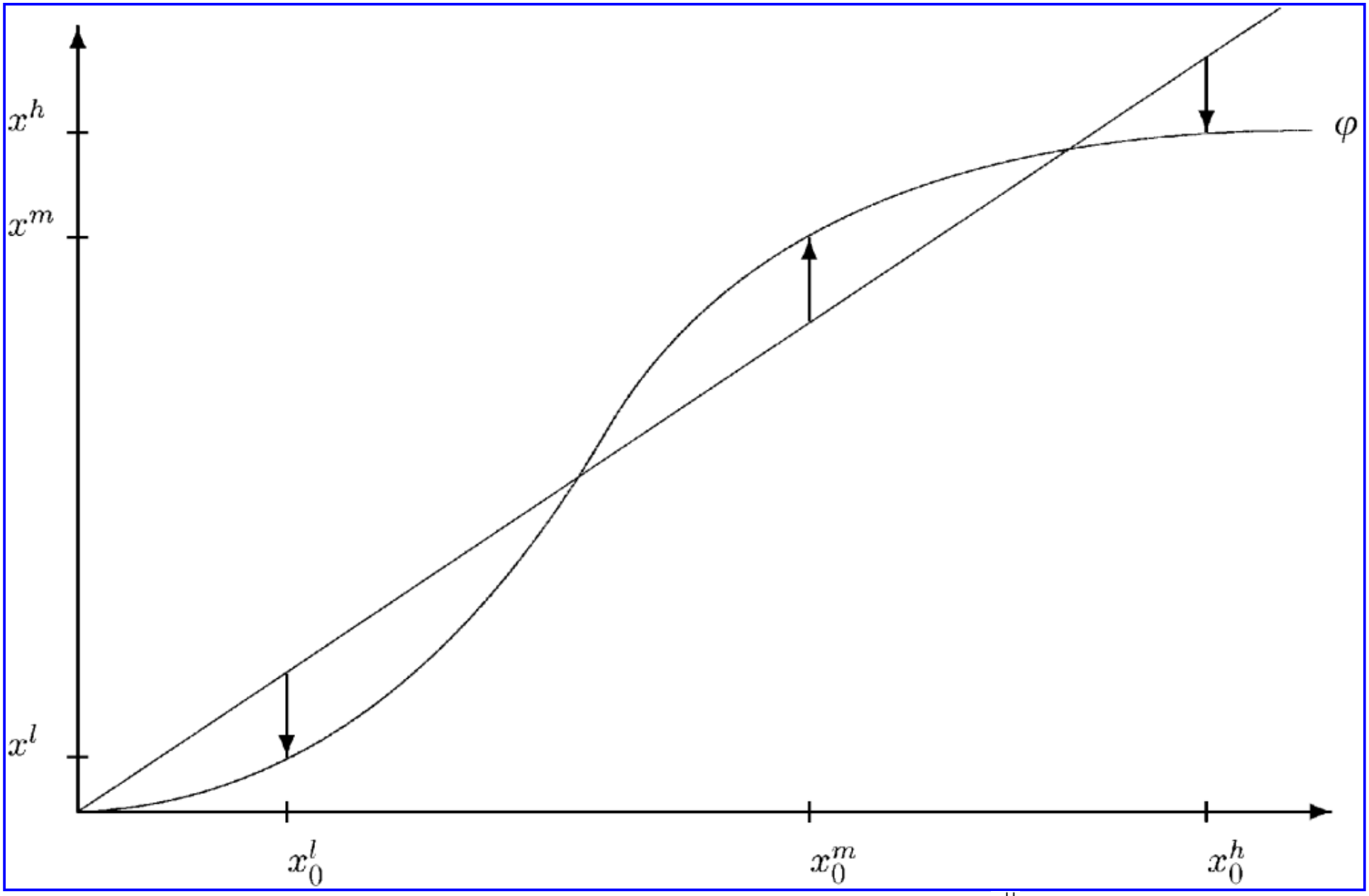

Note that $(x^h) \in \mathcal{X}(x_0)$ implies $\mathcal{X}(x) = \mathcal{X}(x_0)$ and $(x^h) \notin \mathcal{X}(x_0)$ implies $\mathcal{X}(x) \cap \mathcal{X}(x_0) = \varnothing$. Thus, there is a finite partition $\{\mathcal{X}_i\}$ of $\mathbb{R}_+^H$ into sets $\mathcal{X}_i$ of rank preserving income assignments. Note that for any rank preserving income assignments (x^h) in $\mathcal{X}(x_0)$ one can recover the income assignment from knowing only its distribution function G_x, since $x^h = G_x^{-1} G_0(x_0^h)$ for every $h \in H$, where G^{-1} denotes the quantile function (quasi-inverse) of the distribution function G, which is defined by $G^{-1}(q) := \inf\{x \in \mathbb{R}_+ | G(x) \geq q\}$. Consequently, one obtains:

Given any assignment of demand functions f^h, $h \in H$, there exists a function $F(p, G)$ such that mean demand has the representation

$$\frac{1}{\#H}\sum_{h\in H} f^h(p, x^h) \equiv F(p, G_x) \text{on } \mathcal{X}(x_0) \times P. \tag{10}$$

The function F is defined by $F(p, G_x) = \frac{1}{\#H}\Sigma_{h\in H} f^h(p, G_x^{-1} G_0(x_0^h))$.

There might be larger sets than $\mathcal{X}(x_0)$ for which the representation (10) holds. For example, if households *k* and *j* have the same demand function then one can interchange their rank position. Thus, in defining a set $\mathcal{X}$ for which (10) holds, one should take into account the heterogeneity structure of $(f^h)_{h\in H}$. This is done in the next example

Example 3: Common copula

Let $\{f_1, \ldots, f_N\}$ be the set of distinct demand functions of the given assignment $(f^h)_{h\in H}$. Thus, for $h\in H$ there is an integer $n(h) \leq N$ such that $f^h = f_{n(h)}$. For every income assignment $(x^h)_{h\in H}$ consider the bivariate distribution function D_x, which is defined by

$$D_x(\xi, \eta) := \frac{1}{\#H}\#\left\{h\in H \middle| x^h \leq \xi \text{and} n(h) \leq \eta\right\}, \xi, \eta \in \mathbb{R}.$$

The distribution function D_x and the price vector *p* determines mean demand $\frac{1}{\#H}\Sigma_{h\in H} f^h(p, x^h)$. The marginal distribution functions of D_x are denoted by G_x and *V*.

By Sklar's Theorem (see, for example, Nelson, 1999), for every bivariate distribution function *D* with marginals *G* and *V*, there exists a copula *C* (a function from $[0,1]^2$ into [0,1] with certain properties) such that $D(\xi, \eta) = C(G(\xi), V(\eta))$ for all $\xi, \eta \in \mathbb{R}$. Conversely, if *C* is a copula and *G* and *V* are distribution functions, then $C(G(\xi), V(\eta))$ is a bivariate distribution function. Thus, a copula ‘couples’ the marginals to the bivariate distribution. The copula models the dependence structure of the bivariate distribution function.

Starting from an initial income assignment (x_0^h), one considers the set $\mathcal{X}(x_0, f) \subset \mathbb{R}_+^H$ of income assignments (x^h) such that the corresponding bivariate distribution functions D_x have a common copula. Thus, the dependence structure of (x^h, f^h) across *H* is the same for all (x^h) in $\mathcal{X}(x_0, f)$. It follows that income assignments in the set $\mathcal{X}(x_0)$ of rank preserving income assignments is contained in the set $\mathcal{X}(x_0, f)$. Furthermore, *given any assignment of demand functions* $(f^h)_{h\in H}$, *there exists a function* $F(p, G)$ *such that mean demand has the representation*

$$\frac{1}{\#H}\sum_{h\in H} f^h(p, x^h) \equiv F(p, G_x) \text{on} \mathcal{X}(x_0, f) \times P.$$

There is a very simple, however, special case which is worthwhile to be mentioned (and could have been discussed at the beginning). If the initial income x_0^h and the demand function f^h of household *h* are independently distributed across *H*, that is, $D_{x_0}(\xi, \eta) \equiv G_{x_0}(\xi)V(\eta)$ (the copula of D_{x_0} is equal to $C(u, v) = u \cdot v$), then the set $\mathcal{X}(x_0, f) =: \mathcal{Z}(x_0)$ is very large; it consists of all income assignments $(x^h) \in \mathbb{R}_+^H$ with the property: $x_0^k = x_0^j$ implies $x^k = x^j$. Then, one obtains

$$\frac{1}{\#H}\sum_{h\in H} f^h(p, x^h) \equiv F(p, G_x) \text{on} \mathcal{Z}(x_0) \times P$$

with $F(p, G) = \int \bar{f}(p, \xi) dG(\xi)$ where $\bar{f}(p, \xi) = \frac{1}{\#H} \Sigma_{h \in H} f^h(p, \xi)$.

3 Monotone mean demand

The *'law' of demand* for a population of households asserts that the vector of price changes $\Delta p \in \mathbb{R}^l$ and the resulting vector of mean demand changes $\Delta F \in \mathbb{R}^l$ point in opposite directions, provided the price changes do not affect households' incomes (total expenditure) and demand functions (preferences). Thus, the 'law' asserts that the mean demand function *F*(*p*) is *strictly monotone*, that is,

$$(p - q) \cdot (F(p) - F(q)) < 0 \text{for all } p, q \in \mathbb{R}^l_{++}, \ p \neq q.$$

Strict monotonicity of mean demand implies, in particular, that for every commodity *i* the partial mean demand function F_i is strictly decreasing in its own price p_i and that the mean demand function *F*(·) is invertible (existence of an inverse demand function).

The goal of aggregation theory is to derive strict monotonicity of mean demand without assuming that households' demand functions $f^h(p, x)$ are strictly monotone in *p*.

Demand functions $f^h \in \mathcal{F}$ are assumed to be continuous in *p* and *x* and satisfy the budget-identity $p \cdot f(p, x) = x$. The function $f \in \mathcal{F}$ satisfies the *Weak Axiom* of revealed preferences if for every price-income pair (*p*,*x*) and (*p* ′ ,*x* ′), $p \cdot f(p', x') \leq x$ implies $p \cdot f(p, x) \geq x'$, and satisfies the *Axiom* of revealed preferences, if $f(p, x) \neq f(p', x')$ and $p \cdot f(p', x') \leq x$ implies $p' \cdot f(p, x) > x'$.

Every demand function which is derived from a continuous, strictly convex and non-saturated preference relation satisfies the Axiom, yet it is not necessarily monotone.

Theorem: (Hildenbrand, 1983)

1. 1. The function $F(p) := \int_0^\infty f(p, x) \rho(x) dx$ is monotone, that is, $(p - q) \cdot (F(p) - F(q)) \leq 0$ for all *p*, *q* in $\mathbb{R}^l_{++}$, if $f \in \mathcal{F}$ satisfies the Weak Axiom of revealed preferences and ρ is a density which is non-increasing on $\mathbb{R}_+$ with $\int_0^\infty \rho(x) dx < \infty$.
2. 2. The function *F* is strictly monotone, if, in addition, *f* satisfies the Axiom of revealed preferences and the expansion paths $f(p, \cdot)$ and $f(q, \cdot)$ have only 0 in common for any *p*, *q* that are not collinear.

Interpretation: The underlying micro-model is a population *H* of households which is 'indefinitely large'; mathematically, an atomless measure space, for example, the unit interval [0,1] with Lebesgue measure. Every household *h*∈[0,1] is modelled by its income $x(h) \geq 0$ and the common demand function *f*. The income assignment *x*(·) is an integrable function whose distribution admits a density ρ. Thus, mean demand $F(p) = \int_0^1 f(p, x(h)) dh = \int_0^\infty f(p, x) \rho(x) dx$.

Three qsts are relevant:

1. 1. Why a continuum of households? Does the result still hold approximately for a large but finite population?
2. 2. Why a non-increasing income density? Does monotonicity of *F* fail if the density is first increasing and then decreasing?
3. 3. Why a common demand function? Does the result extend to heterogeneous populations in income and demand behaviour?

The discussion of these qsts is simplified by assuming that *f* is continuously differentiable in *p* and *x*. Then monotonicity of *F* is equivalent with negative semi-definiteness (n.s.d.) of the Jacobian matrix $\partial_p F(p)$ for all *p*, that is, $\Sigma_{i,j=1}^l v_i v_j \partial_{p_i} F_j(p) \leq 0$ for all $v \in \mathbb{R}^l$, and the Weak Axiom for *f* is equivalent with n.s.d. of the Slutzky substitution matrix. Consequently, monotonicity of *F* follows from the positive semi-definiteness (p.s.d.) of the mean income effect matrix $I(f, \rho) = \int I(f, x) \rho(x) dx$, where $I(f, x) = (f_i(p, x) \partial_x f_j(p, x))_{i,j=1,\ldots,l}$.

Question 1: The mean income effect matrix for a finite population *H*, that is, $\frac{1}{\#H} \Sigma_H (f_i(p, x^h) \cdot \partial_x f_j(p, x^h))_{i,j} = I_H$ is p.s.d. if and only if for every $v \in \mathbb{R}^l$, $v' \cdot I_H v = \frac{1}{\#H} \Sigma_H g'(x^h) \geq 0$

where $g(x) := \frac{1}{2}(v \cdot f(p, x))^2$. Assume that income x^h is measured in multiples of Δ (euro). Let $\pi_n := \frac{1}{\#H} \# \left\{ h \in H \middle| x^h = n \cdot \Delta =: x_n \right\}$, $n = 0, 1, \ldots$ Then

1. (1) $\frac{1}{\#H} \Sigma_H g'(x^h) = \Sigma_{n=0}^{\infty} \pi_n g'(x_n) = \Sigma_{n=1}^{\infty} \frac{1}{\Delta}(\pi_{n-1} - \pi_n) g(x_n) + o(\Delta)$ using the approximation
2. (2) $g'(x_n) = \frac{1}{\Delta}(g(x_{n+1}) - g(x_n)) + o(\Delta)$.

Consequently, one needs $\pi_{n-1} \geq \pi_n$, $n = 1, \ldots$, to obtain a non-negative first term on the right hand side of (1); this is the finite analogue of a non-increasing density. *Thus, for a finite population with a small* Δ (which requires by $\pi_{n-1} \geq \pi_n$ a large population) *one obtains the desired result up to the small term o*(Δ). For a population *H*=[0,1] one does not need the approximation (2) and hence *o*(Δ), since (1) becomes $\int g'(x) \rho(x) dx = - \int g(x) \rho'(x) dx$ (by partial integration), which is non-negative for a non-increasing differentiable density ρ.

Question 2: The mean income effect matrix $I(f, \rho)$ is p.s.d. in each of the two extreme cases: either, ρ is non-increasing and no assumption on the shape of the income expansion path $f_i(p, \cdot)$ or, no assumption on ρ yet linearity of $f_i(p, \cdot)$. There must be results in between. Indeed, if the curvature of all income expansion paths $f_i(p, \cdot)$ is limited and the unimodal density ρ is sufficiently skewed, then $I(f, \rho)$ is p.s.d.

Example: All income expansion paths restricted to the interval $[0, \bar{x}]$ are polynomials of degree *n* (note that, no non-linear $f_i(p, \cdot)$ can be a polynomial on $\mathbb{R}_+$) and ρ is concentrated on $[0, \bar{x}]$. Then, $I(f, g)$ is p.s.d. if and only if the matrix $M(n, \rho) := ((i + j) m_{i+j-1})_{i,j=1,\ldots,n}$ is p.s.d. where $m_k := \int x^k \rho(x) dx$ (Hildenbrand, 1994, Appendix 6).

Let the densities ρ_m be as in Figure 5.

Figure 5

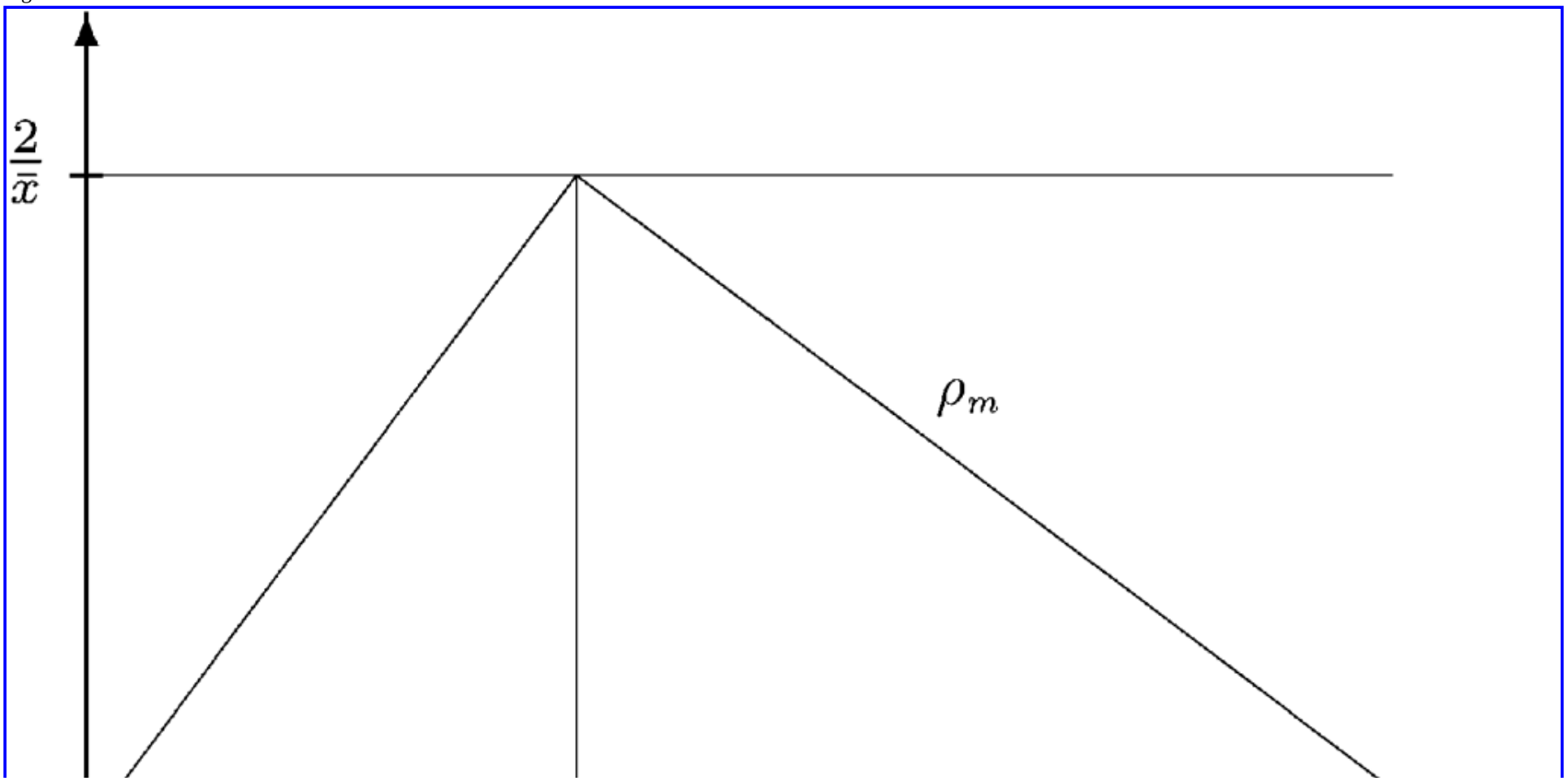

For every n *there exists* $m(n) > 0$ *such that* $l(f, p_m)$ *is p.s.d. if* $m \leq m(n)$; for example, $n = 2$, $m(2) = 0.38\bar{x}$ or $n = 3$, $m(3) = 0.14\bar{x}$.
For a more general analysis see Chiappori (1985) and Hildenbrand (1994).

Question 3: A population of households that is *heterogeneous in income and demand* functions is described by a joint distribution μ of income and demand functions, that is, μ is a distribution on $\mathbb{R}_+ \times \mathcal{F}$. (A reader not familiar with distributions on function spaces might replace $\mathcal{F}$ by a finite set $\mathcal{F}_0$.) As before, the marginal distribution of income admits a density ρ. The conditional distribution of demand functions given the income level *x* is denoted by $\nu(x)$. Then *mean demand*

$$F(p) := \int_{\mathbb{R}_+ \times \mathcal{F}} f(p, x)\, d\mu = \int_0^\infty \bar{f}(p, x)\rho(x)\, dx$$

where $\bar{f}(p, x) := \int_{\mathcal{F}} f(p, x)\, d\nu(x)$. Consequently, the Theorem or the extensions discussed under Question 2 imply that *F*(*p*) is monotone provided the function $\bar{f}$ satisfies the Weak Axiom. This approach to derive monotonicity for a heterogeneous population is the most direct, yet not the most general way (see Hildenbrand, 1994).

It is well-known (Hicks, 1956, p. 53) that $\bar{f}$ does not necessarily satisfy the Weak Axiom, even if individual demand functions are derived from utility maximization. The following two assumptions (which, again, are not the most general ones) imply that $\bar{f}$ satisfies the Weak Axiom

1. (a) *independence*: $\nu(x)$ does not depend on *x*
2. (b) *increasing dispersion*: the distribution $D(x + \Delta)$, $\Delta > 0$, is more dispersed than the distribution *D*(*x*), where *D*(ξ) denotes the distribution (in the commodity space $\mathbb{R}^l$) of individual demand of all households with income ξ at the price *p* (that is, *D*(ξ) is the image distribution of ν under the mapping $f \mapsto f(p, \xi)$).

Generalizing the one-dimensional case where the variance is a measure of dispersion one chooses the positive definiteness of the covariance matrix as a measure of dispersion for distributions on $\mathbb{R}^l$. Thus, increasing dispersion means that for $\Delta > 0$, $covD(x + \Delta) - covD(x)$ is positive semi-definite.
Assumptions (a) and (b) are quite restrictive, in particular, the independence assumption. Therefore one partitions the whole population *H* into sub-populations *H*(*a*) by stratifying with respect to a certain vector *a* of household attributes (household size, age, …) and then one requires assumptions (a) and (b) for each sub-population *H*(*a*). The role of stratifying is to reduce the heterogeneity in demand behaviour. In the extreme case, where stratifying leads to a homogeneous sub-population in demand behaviour, assumptions (a) and (b) are trivially satisfied. If the income density of each sub-population *H*(*a*) is non-increasing on $\mathbb{R}_+$ or if the extension discussed in Question 2 apply, the mean demand of each sub-population is monotone and hence also the mean demand of the whole population, since monotonicity is additive.
A more general definition of 'increasing dispersion' and a detailed discussion is given in Hildenbrand (1994). For an empirical study of the law of demand, see Härdle, Hildenbrand. and Jerison (1991).
A broader discussion of the law of demand and related properties including cases where income is price dependent is contained in the entry law of demand.

See Also

- aggregation (econometrics)
- copulas
- law of demand

Bibliography

Antonelli, G.B. 1886. *Sulla teoria matematica della economia politica*. Pisa: Nella Tipografia del Folchetto. Trans. J.S. Chipman and A.P. Kirman in *Preferences, Utility* and Demand, ed. J. S. Chipman, L. Hurwicz and M.K. Richter. New York: Harcourt Brace Jovanovich, 1971.

Chiappori, P. 1985. Distribution of income and the law of demand. *Econometrica* 53, 109–27.

Chipman, J.S. and Moore, J. 1979. On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21, 111–39.

Eisenberg, B. 1961. Aggregation of utility functions. *Management Science* 7, 337–50.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21, 63–80.

Grandmont, J.M. 1992. Transformations of the commodity space, behavioural heterogeneity, and the aggregation problem. *Journal of Economic Theory* 57, 1–35.

Härdle, W., Hildenbrand, W. and Jerison, M. 1991. Empirical evidence on the law of demand. *Econometrica* 59, 1525–49.

Heineke, J. and Shefrin, H. 1987. On some global properties of Gorman class demand systems. *Economics Letters* 25, 155–160.

Heineke, J. and Shefrin, H. 1988. Exact aggregation and the finite basis property. *International Economic Review* 29, 525–38.

Hicks, J.R. 1956. *A Revision of Demand Theory*. London: Oxford University Press.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51, 997–1019.

Hildenbrand, W. 1994. *Market Demand*. Princeton: Princeton University Press.

Hildenbrand, W. and Kneip, A. 2005. On behavioral heterogeneity. *Economic Theory* 25, 155–69.

Jorgensen, D.W., Lau, L. and Stoker, T.M. 1982. The transcendental logarithmic model of aggregate consumer behavior. In *Advances in Econometrics*, ed. R.L. Basmann and G.F. Rhodes. Greenwich, CT: JAI Press.

Lau, L. 1982. A note on the fundamental theorem of exact aggregation. *Economics Letters* 9, 119–26.

Malinvaud, E. 1956. L'agregation dans les modèles économiques. *Cahiers du Seminaire d'Econométrie* 4, 69–146.

Malinvaud, E. 1993. A framework for aggregation theories. *Ricerche Economiche* 47(2), 107–35.

Mitjuschin, L.G. and Polterovich, W.M. 1978. Criteria for monotonicity of demand functions [in Russian]. *Ekonomika i Matematicheskie Metody* 14, 122–8.

Nataf, A. 1948. Sur la possibilité de construction de certains macromodèles. *Econometrica* 16, 232–44.

Nelson, R. 1999. *An Introduction to Copulas*. Lecture Notes in Statistics 139. New York: Springer Verlag.

Trockel, W. 1984. *Market Demand: An Analysis of Large Economies with Non-convex Preferences*. Lecture notes in economics and mathematical systems 223. Heidelberg: Springer Verlag.

How to cite this article

Hildenbrand, Werner. "aggregation (theory)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000247>
doi:10.1057/9780230226203.0020

The New Palgrave Dictionary of Economics Online

agricultural economics

C. Ford Runge

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Agricultural economics arose in the late 19th century, combined the theory of the firm with marketing and organization theory, and developed throughout the 20th century largely as an empirical branch of general economics. The discipline was closely linked to empirical applications of mathematical statistics and made early and significant contributions to econometric methods. From the 1960s, as agricultural sectors in the OECD countries contracted, agricultural economists were drawn to the development problems of poor countries, to the trade and macroeconomic policy implications of agriculture in richer countries, and to a variety of issues in production, consumption, environmental and resource economics.

Keywords

agricultural economics; agricultural subsidies; agriculture and economic development; American Agricultural Economics Association; American Economic Association; American Farm Economics Association; Australian Agricultural Economics Society; Bellman's dynamic programming principle; Black, J. D.; Boulding, K.; Center for Agricultural Research and Development; Chayanov, A.; Chenery, H.; Clark, J. B.; cobweb theorem; common property resources; Commons, J. R.; computable general equilibrium models; corn–hog cycle; Cournot, A. A.; diminishing returns; Ely, R. T.; environmental economics; European Association of Agricultural Economists; Ezekiel, M.; factor mobility; fisheries; general equilibrium; Georgescu-Roegen, N.; Griliches, Z.; Heady, E.; health; hedonic prices; Hicks, J. R.; horizontal integration; human capital; index numbers; indicative planning; industrial organization; innovation; input–output analysis; Johnson, D. Gale; land; land grant model; Laur School (Switzerland); Leontief, W.; neoclassical economics; neoclassical synthesis; Nourse, E.; nutrition; optimization theory; Physiocracy; pollution; poverty alleviation; price stabilization; production economics; property rights; Quesnay, F.; recursive models; recursive programming; rent; Ricardo, D.; risk preferences; Samuelson, P. A.; Schultz, T. W.; Sering School (Germany); Serpieri School (Italy); Shackle, G. L. S.; Smith, A.; spatial economics; species loss; Stigler, G.; stochastic programming methods; subjective probability;

supply controls; surplus production; Taylor, H. C.; technical change; Thünen, J. H. von; treadmill hypothesis; Veblen, T.; vertical integration; Warren, G.; Waugh, F.; Working, E.; Working, H.; Young, A.

Article

Agricultural economics arose in the late 19th century, combined the theory of the firm with marketing and organization theory, and developed throughout the 20th century largely as an empirical branch of general economics. This emphasis was due to the historical importance of agriculture, and in the United States was made possible by the rich data compiled by the US Department of Agriculture beginning in the mid-19th century. The discipline was closely linked to empirical applications of mathematical statistics and made early and significant contributions to econometric methods. From the 1960s on, as agricultural sectors in the OECD countries contracted, agricultural economists were drawn to the development problems of poor countries, to the trade and macroeconomic policy implications of agriculture in richer countries, and to a variety of issues in production, consumption, environmental and resource economics. This ramified the subject and enlarged its international focus, at the same time as its microeconomic, empirical and policy orientation distanced it from developments in general equilibrium theory, macroeconomic modelling, game theory and axiomatic social choice, which preoccupied many departments of economics throughout the late 20th century.

Retracing the evolution of agricultural economics, especially in the United States, requires an explanation of institutional innovation in 19th-century America (see Taylor and Taylor, 1952). In the midst of the Civil War, President Lincoln created the Federal Department of Agriculture (later the US Department of Agriculture, USDA), empowered to collect a wide range of farm statistics. At the same time, legislation introduced by Vermont's Justin Morrill (previously blocked by the seceded South) was signed in 1862 by Lincoln. The Morrill Act established the Land Grant Colleges (financed through sales of government land) especially in the states of the Old Northwest Territory: Illinois, Indiana, Michigan, Ohio and Wisconsin. Their creation reflected both vast surpluses of land and the drive to improve plant and animal husbandry through applications of chemistry and biology. Eventually, the land grant model was replicated in every state as well as in some other countries. In 1887 the Hatch Act created the Agricultural Experiment Stations of USDA, which functioned together with the Land Grant Colleges to form a system of research, instruction and outreach to farmers (Cochrane, 1993; Kerr, 1987; Moore, 1988). In 1914, extension education and outreach was formalized under the Smith–Lever Act. By the beginning of the 20th century, the application of scientific management to agricultural production created the foundations of the discipline.

Intellectual origins

Agricultural economics in the United States derived from two intellectual streams. The first was neoclassical political economy and the theory of the firm applied to farm production. The second, borne of an economic crisis in American agriculture in the late 19th century, focused on strategies for organized marketing of agricultural commodities through collective bargaining and cooperatives. The first stream may be traced to the 18th-century Enlightenment and a preoccupation with land as a factor

by the French Physiocrats. Francois Quesnay's *Tableau économique* (1758) organized a logical explanation of the conversion of land inputs to agricultural outputs and profit, anticipating modern production economics, input–output analysis and general equilibrium theory. His emphasis on surplus production was a touchstone of classical economics and exercised a direct influence over Adam Smith (Eltis, 1975; Smith, 1776, book II, ch. 9).

Like all 18th-century political economists, Smith could not ignore agricultural questions, even if he gave them less primacy than the Physiocrats. Together with Ricardo, Von Thünen and Malthus, he provided commentary on the difficulties of agricultural specialization, returns to land as a factor, issues of space and distance to market, and the long-run relation between arithmetic increases in food supply and geometric increases in demand due to population growth. Many pages of the *Wealth of Nations* dealt with agricultural questions, including the differential capacity for specialization and routinization of agriculture versus industry and the arts of husbandry at the microeconomic level (1776, pp. 16, 143). Echoing the Physiocrats, Smith emphasized the central role of agriculture as a store of national wealth, and noted that compared with manufacturing, agriculture ‘is much more durable, and cannot be destroyed by [the] violent convulsions’ of war and political instability (1776, p. 427). In the same period, Arthur Young assembled comprehensive data on production, rents and land tenure in Great Britain. Serving as editor of the *Annals of Agriculture* from 1768 to 1770, he collected his data and observations into nine volumes of 4,500 pages, which have proved to be of continuing value especially to economic historians (for example, Allen, 1992). Ricardo (1821, p. 44) was famously concerned with returns to land as a fixed factor ‘for the use of the original and indestructible powers of soil’. He also distinguished between productivity enhancements due to augmentation of the soil and improvements in machinery and the capitalization of various investments or policies (such as taxes) into the value of land (1821, pp. 57–61; 246). Von Thünen's (1828) analysis of the extensive margin and the relationship between distance to market and rent made him, in Marshall's view, the first agricultural economist among economists, who with Cournot provided the inspiration for marginalist economics (Day and Sparling, 1977, p. 93).

It was the neoclassical developments of the late 19th century, however, that provided the main foundations for agricultural economics. Marshall's *Principles* (1890) first clearly established the link from diminishing marginal utility in exchange to decreasing marginal productivity on the supply side. Veblen (1900) dubbed Marshall's work ‘neoclassical’ to distinguish it from classical labour theories of value. The elaboration of Marshall's theory of the firm, and attempts to measure and statistically validate the relationship between input costs, output prices, and farm profits distinguished agricultural economics well into the 20th century, and linked it firmly to the neoclassical syntheses of Hicks (1939) and Samuelson (1947).

To this was added a second stream of marketing and organizational issues growing out of the extended farm depression from the 1870s to the 1890s. Joined with labour interests, farmers sought marketing outlets and modes of organization that would give them greater bargaining power, notably cooperatives popular in northern Europe and Scandinavia, where many recently arrived American farmers originated (Jesness, 1923). Even after the business cycle turned upward after 1897, the Land Grant colleges emphasized farm management. The result was the organization in 1910 of the American Farm Management Association. Farm managers were focused on the physical, technical and scientific aspects of production, especially the new field of agronomy.

Many early agricultural economists regarded farm management as a sub-field, and agricultural economics as an applied version of general economics. Beginning in 1907, at the tenth American Economic Association (AEA) meetings, a session was devoted to ‘What is agricultural economics?’ Thereafter, the AEA regularly included sessions on the economics of agriculture. In 1915 the National Association of Agricultural Economists was formed. In 1917 the AEA meeting was held jointly with the National Agricultural Economics Association and the American Farm Management Association, and talks began on a merger of the latter two. This was realized in 1919 in the form of the American Farm Economics Association, with Henry C. Taylor of the University of Wisconsin as President (Taylor, 1922; Cochrane, 1983). It retained this title until 1968, when it became the American Agricultural Economics Association (AAEA).

The discipline expands

As Cochrane (1983, p. 66) observed, ‘the first flowering of agricultural economics as an applied field of economics occurred at the University of Wisconsin in the period of 1900–1920. The second flowering occurred at the University of Minnesota in the period of 1918–1928.’ A department of agricultural economics was established at Wisconsin in 1909 by Henry C. Taylor and colleagues such as Benjamin Hibbard. Taylor's text, *An Introduction to the Study of Agricultural Economics* (1905), applied Marshallian principles to farm production, and developed production functions showing increasing, steady and diminishing returns. Among the most influential leaders in the young subject was Taylor's student at Wisconsin, John D. Black, who also studied under John R. Commons and Richard T. Ely (who himself authored an influential, though unpublished, 1904 study on the economics and property rights of irrigation). Their emphasis on land and institutions permeated the discipline and was reflected in the journal *Land Economics*, which began publication at Madison in 1925.

Black, a follower of Marshall and John Bates Clark, received his Ph.D. in 1918 and moved to the University of Minnesota, where he remained a dominant force until hired by Harvard in 1927. By the mid-1920s Black's leadership had marked him, together with George F. Warren of Cornell and Edwin G. Nourse of Iowa State, as ‘the most influential economist in the United States dealing with the problems of agriculture’ (Galbraith, 1959, p. 10). Together with a cadre of other young economists working with the Bureau of Agricultural Economics (BAE), created in USDA in 1921, Black set the tone for research in the field from the 1920s until the advent of the Second World War.

Black's text, *Introduction to Production Economics* (1926), became the standard. His emphasis on the theory of the firm was complemented by his colleague Holbrook Working's econometric explorations. Working's 1922 bulletin, ‘Factors Determining the Price of Potatoes in St. Paul and Minneapolis’, was among the first to derive an empirical demand curve (H. Working, 1922; 1925). It was followed by his brother E. J. Working's widely cited 1927 article, ‘What Do Statistical “Demand Curves” Show?’ The Workings and colleague Warren Waite continued to expand research into price analysis in the interwar years. Minnesota's Frederick V. Waugh contributed the first quantitative study of quality characteristics as determinants of prices, recognized as a forerunner of hedonic price analysis. Appearing as ‘Quality Factors Influencing Vegetable Prices’ (1928), it noted that if ‘a premium for certain qualities and types of products is more than large enough to pay the increased cost of growing a superior product, the individual can and will adapt his production and marketing policies to market demand’ (quoted in

Berndt, 1991, p. 106).

Taylor, Black, Warren and Nourse were followed by a group of young empiricists and econometricians who continued to develop the USDA Bureau of Agricultural Economics (BAE). Tolley, Black and Ezekiel (1924) showed how production surfaces in three dimensions could express diminishing returns to inputs, a concept readily grasped by agricultural field scientists. They then derived cost surfaces showing the relationship between costs, relative prices, and profit maximization. Ezekiel followed this empirical work with his 1930 volume *Methods of Correlation Analysis*, which became a standard text on regression analysis, and in 1938 with a state-of-the-art description of cobweb and recursive models illustrated by the corn–hog cycle. Leontief (1971, p. 5) would call this and other early agricultural economists' work 'An exceptional example of a healthy balance between theoretical and empirical analysis …' and 'the first among economists to make use of the advanced methods of mathematical statistics'.

By the 1930s departments of agricultural economics had been established in many US universities, where technical and institutional issues affecting agricultural production formed the core subjects. In addition to the leading roles played by Cornell, Illinois, Iowa State, Minnesota, Purdue and Wisconsin, a major research programme was established at the University of California-Berkeley (and a later campus at Davis) with the endowment of the Giannini Foundation. At Iowa State, future Nobel Laureate T.W. Schultz arrived in 1930 with a Ph.D. from Wisconsin, and then served as department head from 1934 to 1943 until leaving for Chicago. Schultz attracted numerous talents including Kenneth Boulding, George Stigler, D. Gale Johnson and Earl O. Heady, several of whom would also leave for Chicago following controversy surrounding oleomargarine and the Iowa butter industry (Beneke, 1998). The butter–margarine dispute was typical of agricultural economists' conflicts with interest groups in a profession seldom sheltered from political winds, especially at state universities. Partly for this reason, several private universities also made substantial contributions to agricultural economics research. In addition to Black (and later Galbraith) at Harvard, the University of Chicago remained a center of research excellence. At Vanderbilt, Nicholas Georgescu-Roegen, a demand theorist and econometrician, expressed path-breaking insights into the physical process underlying economic activity, and contributed a deep critique of agrarianism and Marxian misunderstandings of agricultural production (Georgescu-Roegen, 1960).

Earl O. Heady remained at Iowa State, creating a post-war engine of applied research, the Center for Agricultural Research and Development (CARD), in 1957. He pioneered the application of programming methods first developed for war planning, analysing how inputs could most efficiently be employed in producing agricultural outputs. This made the discipline a centre for research in applications of optimization theory. Heady authored or oversaw hundreds of mainly empirical production studies, exemplified by Heady and Dillon (1961) and Heady and Candler (1958). He also pioneered the application of computing power to problem-solving in applied economics. This included work on human and animal diet rations and consumption (for example, Waugh 1951; Heady 1951). Farm management also saw optimization applications in work by Hildreth (1957a) among others. By the late 1950s Bellman's dynamic programming principle was applied to optimal wheat rotations by Burt and Allison (1963). Agricultural economics also began to grapple empirically with uncertainty through stochastic programming methods, including Hildreth's (1957b) work and Hazell's applications (1971).

French economists Boussard and Petit applied Shackle's ‘focus loss’ concept of uncertainty to agriculture (1967). The application of subjective probability concepts to agriculture was surveyed by Dillon (1971) and Anderson, Dillon and Hardaker (1977).

Yet another outgrowth of optimization theory was analysis of the growth and decline of farms in modern economies, including contributions by German agricultural economists Heidhues (1966) and De Haen (De Haen and Heidhues, 1973). Behavioural adjustment (‘supply response’) in agriculture was studied using recursive programming models (Henderson, 1959), and generalized by Day (1963), following the path set by Nerlove (1958). Optimal storage rules were analysed by Gustafson (1958). Spatial issues in agriculture analysed best-location decisions (Egbert and Heady, 1961), and interregional supply–demand equilibrium issues (for example, Fox, 1953). An extensive bibliography of spatial and temporal equilibrium models was published by Judge and Takayama (1973).

New frontiers

Two additional applications of optimization theory pushed agricultural economics in the 1960s and 1970s toward new frontiers: natural resources and agricultural development in developing countries. These helped attract a new generation of economists concerned less with domestic farm production than with environmental issues and poverty alleviation in the Third World. Natural resources were analysed as problems of materials shortages and treated as a form of capital, following the early analytical leads of Hotelling (1931) and Ciriacy-Wantrup (1952). Especially after the Paley Commission Report of 1952 led to the creation of Resources for the Future in Washington, DC, a new group of economists applied themselves to these issues. Fisheries were studied by Scott (1955) and Crutchfield and Zellner (1962); groundwater allocation over time was considered as a dynamic programme with stochastic state variables in a series of articles by Burt (for example, Burt, 1966; Burt and Cummings, 1970). These dynamic models were extended to interregional investments in water in studies such as Cummings and Winkelmann (1970). By the 1970s, environmental pollution became a major subject of applied economics, pulling many in the profession away from a restricted view of agricultural issues as matters of yields and production in acknowledgement of the sector's negative external effects and market failures.

Agricultural development in developing countries, meanwhile, was an important area of applied economics in project evaluation, supported by multilateral and bilateral aid agencies such as the World Bank, the Food and Agriculture Organization of the UN (FAO) and US Agency for International Development. At Stanford, the Food Research Institute (1921–95) established an internationally focused research programme. The development problem in the Third World was seen largely as an imbalance between agricultural and manufacturing sectors, with a need to right this balance by drawing low-productivity resources out of agriculture (Lewis, 1954; Mellor, 1966; Timmer 2002). Hollis Chenery at the World Bank exemplified the analysis of agriculture's sectoral role (Chenery and Syrquin, 1975). However, unlike the United States and some other OECD countries, data limitations in poor countries restricted the early application of optimization models at the microeconomic level. Indeed, T.W. Schultz's famous *Transforming Traditional Agriculture* (1964) relied mainly on stylized representations of ‘rational but poor’ farmers and descriptive analysis from anthropologists.

Throughout the 1950s and 1960s the agricultural sector continued to contract in the OECD countries, setting the tone for policy debates. Many agricultural economists saw the ‘farm problem’ as one of surplus labour supplying farm commodities in excess of domestic demand. Analysing low agricultural prices as a matter of chronic oversupply, aggravated by rapid technological improvements and productivity gains in the face of inelastic demand, Cochrane (1958) proposed his treadmill hypothesis: rapid and early adopters of productivity-improving technology will reap the lion's share of rents to innovation, as laggards are forced off the farm, while Brewster (1959) considered the social and policy implications of these trends. In the early 1960s, serving as presidential adviser, Cochrane advocated a solution to excess production in the form of federally mandated supply control. When it became clear that the major commodity groups would vote down the enabling referenda, and that its success would raise prices to consumers, President Kennedy abandoned the scheme. Thereafter, although mandated supply control retained adherents (not including Cochrane), US agricultural policy shifted towards exports as a vent-for-surplus.

This opened the way to consideration of agriculture in an open economy, and a new policy emphasis on the macroeconomics of the food sector (Schuh, 1974; 1976; Cochrane and Runge, 1992; Ardeni and Freebairn, 2002; Abbott and McCalla, 2002). In the 1980s, this open economy analysis was supported by the development of large-scale computable general equilibrium models linking agriculture to trade (for example, Hertel, 1997) as well as more traditional macroeconomic sectoral forecasting models (for example, Myers et al., 1987). Together, the large-scale models allowed alternative trade and agricultural policy approaches to be simulated and compared to the *status quo* (for example, Cochrane and Runge, 1992).

International reach

The intellectual antecedents of agricultural economics make clear that the field has never been restricted to the United States. In 1905, the International Agricultural Institute was founded in Rome as the forerunner of the FAO. In Great Britain, an Agricultural Economics Research Institute was established at Oxford in 1913, and in 1945 became part of the School of Rural Economy, merging with Queen Elizabeth House and the Institute for Commonwealth Studies in 1986. Oxford led the creation of the International Association of Agricultural Economists and helped coordinate its first conference in 1929 at Dartington Hall, Devon and a second in 1930 at Cornell. These were largely Anglo-American meetings, although by the third meeting in Germany in 1934, 19 different countries were represented. At Cambridge, a Department of Estate Management was transformed into a Department of Land Economy in the 1960s. At Wye, an agricultural college was founded in 1894. The college was awarded a royal charter in 1948 and in 2000 its agricultural economics department became part of Imperial College London.

On the Continent, followers of Von Thünen had developed marginalist principles and farm accounting methods in the late 19th and early 20th century represented by the Laur School in Switzerland and the Sering and Serpieri Schools in Germany and Italy. However, their capacity was limited by poor data, few marketing studies, and a weak connection to production economics (N ō u, 1967; Raeburn and Jones, 1990, p. 13). In 1948 a French professional association began, and a Department of Agricultural Economics was created at the Institut National de la Recherche Agronomique (INRA) in 1955 (Petit,

1982). A European Association of Agricultural Economists was founded in 1975 in Uppsala, Sweden. By the late 1980s, it was estimated that 3,000–5,000 European professionals were engaged in full-time agricultural economics research dispersed in hundreds of research institutes, universities and government offices (Hanf, 1988). Among the leaders were the French government's INRA, the Universities of Goettingen and Kiel in Germany, the University of Padova in Italy, Wageningen University in the Netherlands, and the aforementioned activities in Great Britain.

In Canada, agricultural economics began at the Ontario Agricultural College (now the University of Guelph) in 1907. Noteworthy research departments of agricultural economics were established at the University of Guelph, Ontario, McGill University in Montreal, Laval University in Quebec, and the Universities of Manitoba, Alberta, Saskatchewan and British Columbia.

The Australian Agricultural Economics Society was founded in Sydney in 1957, following the models of the US, British and Canadian associations. In 1975, a New Zealand branch of the association was established at a meeting in Christchurch. The leading Australian institution in creating a separate department was the University of New England at Armidale, which in 1958 began a four-year course. Supported by grants from the Commonwealth Bank, a chair of agricultural economics was appointed at the University of Sydney in 1951 (Campbell, 1985). While maintaining the specialty within economics rather than a separate department, major research was also undertaken beginning in the 1950s and 1960s at the University of Adelaide and at the University of Melbourne, and later at the Australian National University in Canberra and the University of Western Australia in Perth. All of these universities were closely linked to the national Bureau of Agricultural Economics (BAE), which became the Australian Bureau of Agriculture and Resource Economics (ABARE) in 1987 (Miller, 1985).

In Russia, interest in agricultural economics may be traced to the establishment in 1865 of the Moscow Agricultural Academy. In 1929 Lenin created the Russian Academy of Agricultural Sciences, following conflicts between Chayanov and Marxist agriculturalists. After Stalin's rise to power in 1930, agricultural research was fully politicized with well-known results, including the purge of many academic researchers (Nazarenko, 2004). In the 1950s, concepts such as profit and cost were revived, and central planners embraced modelling and forecasting. Since the 1990s, agricultural reforms have led to dissension in the Russian discipline (Klyukach, 2004).

In Brazil, the Rockefeller and Ford Foundations and the US Agency for International Development provided core support for agricultural economics research, beginning in the late 1950s. Four US universities were directly involved: Purdue, Wisconsin, Ohio State and Arizona.

In India, a Society of Agricultural Economics was established in 1939. The advent of indicative economic planning in the 1950s stimulated analytical studies to assist in the Plan. Due to the overwhelming importance of agriculture as a supplier of wage goods, the sector attracted considerable analysis, in which Indian agricultural universities, established on the land-grant model, consciously borrowed methods from their US counterparts, notably Earl O. Heady and the CARD group at Iowa State (Bhide, 1994, p. 119).

In China, missionary efforts to promote agricultural research and development by the Presbyterian Church of New York during the first quarter of the 20th century resulted in a Cornell University–University of Nanking collaboration, led beginning in 1914 by John Lossing Buck (Buck, 1973). J. L. Buck's contributions included early agricultural surveys and analysis of Communist production into the 1960s (Buck, 1943; Buck, Dawson and Wu, 1966).

Late 20th century

Since the 1970s, seven broad subjects have defined the most distinctive contributions of agricultural economics: technical change and the returns to human capital investments; environmental and resource issues; trade and economic development; agricultural risk and uncertainty; price determination and income stabilization; market structure and the organization of agricultural businesses; and consumption and food supply chains.

The study of technical change, innovation and returns to investments in human capital in agriculture attracted some of the most talented economists of the post-war generation, such as Zvi Griliches (1957; 1958; 1963; 1964). Anticipating debates among economic growth theorists over ‘embodied’ technical change due to improvement in the quality of capital inputs (versus ‘disembodied’ changes without new net capital investments), Cochrane (1953) criticized Schultz (1953) for failing to account for capital requirements in agriculture and a resulting overemphasis on weather variations in describing growth in yields. Focusing on the direction of agricultural innovation, Ruttan (1956) and Hayami and Ruttan (1971) emphasized the Hicks-non-neutrality of technical change in both labour-saving US and land-saving Japanese agriculture. This approach was extended in a formal framework by Binswanger (1974). Based on Hicks's (1932) analysis of relative factor prices as the inducement to alternative paths of innovation, the induced innovation argument was extended into an explanation of priority setting by public sector agencies, leading research towards abundant factor use that lowered social costs of production (Peterson and Hayami, 1977, p. 504). How to measure productivity and technical change in agriculture using alternative index numbers attracted both theorists and applied econometricians (for example, Jorgenson and Griliches, 1967; Lau and Yotopoulos, 1971). Finally, analysts considered the welfare gains and losses resulting from farm mechanization (Schmitz and Seckler, 1970).

Agricultural economists also delved into the role of productivity embodied in labour as ‘human capital’, a natural reflection of the huge public investments in research and education by the US land grant system. Surveyed by T. W. Schultz (1971), this line of research attracted work by Peterson (1969), Huffman (1974) and general economists such as Nelson and Phelps (1966), and led to widening emphasis on private and social returns to research including Peterson (1967), Evenson (1967), Evenson and Kislev (1976) and Alston et al. (2000). It also led to analysis of how research ought to be organized in order to maximize its aggregate benefits. Alston, Norton and Pardey (1998) developed a comprehensive summary of this priority-setting problem (see Huffman, 2002; Sunding and Zilberman, 2002).

Environmental and resource issues, as noted, became a significant focus of the profession in the 1970s and beyond, partly in recognition of the pollution and species losses resulting from modern agricultural systems. Surveyed by Lichtenberg (2002), the economics of agriculture and the environment analysed the perverse incentives created by agricultural subsidies and the agency problems of monitoring agricultural practices (for example, Chambers and Quiggin, 1996; Just and Antle, 1990; Segerson, 1988). Induced innovation theory was broadened to explain how technical innovations such as irrigation might give rise to new water quality issues and thus new institutional responses (for example, Runge,

1987; Caswell, Lichtenberg and Zilberman, 1990). Apart from specific agriculture–environment interactions, resource economists emphasized the critical role of property rights in the use and management of resources, especially those held publicly or in common, notably in developing countries (Runge, 1981; Bromley, 1991; Walker, Gardner and Ostrom, 2000).

Trade and development also dominated agricultural economics research, especially after the mid-1980s, as global trade negotiations increasingly hinged on struggles between heavily subsidized farm sectors in OECD countries and the highly taxed sectors of the developing world (Anderson and Hayami, 1986; Kreuger, Schiff and Valdes, 1991–2; Sumner and Tangermann, 2002). An overview of post-war agricultural trade policy was given by D. G. Johnson (1977); a synthetic treatment of agriculture–trade interactions was provided by Karp and Perloff (2002). Meanwhile, a major share of agricultural economics literature was devoted to microeconomic studies of agricultural change and food insecurity in developing countries, and to macroeconomic linkages with other sectors and global trade (for example, Barrett, 2002; Runge et al., 2003).

Risk and uncertainty are inherent in agriculture and their relevance has drawn interest from many agricultural economists, especially in developing-country decision environments (see Moschini and Hennessey, 2002). Roumassett (1976) conducted an early assessment of risk aversion and the adoption of hybrid rice in the Philippines. Dillon and Scandizzo (1978) analysed risk preferences among small farmers in Brazil, while Moscardi and de Janvry (1977) analysed Mexican maize production and the response to risk. Antle (1987) and Myers (1989) provided econometric tests for risk aversion by farmers while Goodwin and Smith (1995) and Miranda and Glauber (1997) considered why crop insurance contracts fail effectively to pool risk without reinsurance.

Price determination and stabilization of agricultural prices as a focus of research arose as a direct consequence of widespread instability in agricultural commodities markets. Tomek and Robinson (1977) surveyed the post-war literature through the 1970s, including the analysis of Cochrane (1958) and Gray and Rutledge (1971). In response to widespread calls for buffer stocks and other mechanisms to affect prices counter-cyclically, Newberry and Stiglitz (1981) offered a comprehensive (and sceptical) assessment of the advantages of stabilization policy. A more recent survey was developed by Wright (2002).

The organizational structure of farms and the role of economies of scale, scope, technological change, capital and labour mobility were reviewed by Chavas (2002). Farm size was analysed as a function of the opportunity cost of labour and the price of machinery (Kislev and Peterson, 1982). Farm structure and the economics of contracting was also an additional area of risk and agency studies (Allen and Lueck, 1998; Hueth and Ligon, 2001; Knoeber and Thurman, 1995). Despite their declining importance in many rural markets, cooperatives continued to attract analysis (for example, Sexton, 1990).

A final area of broad interest was food consumption and supply chains in the food industry. Taking an industrial organization approach, Sexton and Lavoie (2002) provided an overview, emphasizing vertical and horizontal integration and imperfect competition as forces driving the sector, with implications for consumer choice, nutrition and health.

In the 21st century, the profession has continued to reach beyond the agricultural sector, expanding its scope through numerous applications of relevant economic theory. Meanwhile, the high level of

abstraction in economics characteristic of the last half of the 20th century appears to have given way to new interest in empirical and experimental studies, suggesting that the distance between agricultural economics and its mother discipline may narrow in the years ahead.

See Also

- agriculture and economic development
- econometrics

Bibliography

Note: This article can only gesture to the agricultural economics literature. The reader is referred to L. R. Martin (general editor) and the four-volume *Survey of Agricultural Economics Literature*, Minneapolis, University of Minnesota Press, 1977–1987, and Wallace C. Olsen et al., *Agricultural Economics and Rural Sociology: The Contemporary Core Literature*, Ithaca: Cornell University Press, 1991. A compilation of analytical and interpretive essays is B. L. Gardner and G. C. Rausser (eds.), *Handbook of Agricultural Economics*, vols. 1A, 1B, 2A and 2B, Amsterdam: North-Holland, 2002. An internet-based open source of information is AgEcon Search, maintained at the University of Minnesota: http://www.apec.umn.edu/AgEcon.html (accessed 16 August 2006).

Abbott, P. and McCalla, A. 2002. Agriculture in the macroeconomy. In *Handbook of Agricultural Economics*, vol. 2A, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Allen, R. 1992. *Enclosure and the Yeoman*. Oxford: Clarendon Press.

Allen, D. and Lueck, D. 1998. The nature of the farm. *Journal of Law and Economics* 41, 343–86.

Alston, J., Norton, G. and Pardey, P. 1998. *Science Under Scarcity: Principles and Practices for Agricultural Evaluation and Priority Setting*. New York: CAB International.

Alston, J., Chan-Kang, C., Marra, M., Pardey, P. and Wyatt, T. 2000. A meta-analysis of rates of return to agricultural R&D. Research Report No. 113. Washington, DC: International Food Policy Research Institute.

Anderson, J., Dillon, J. and Hardaker, B. 1977. *Agricultural Decision Analysis*. Ames: Iowa University Press.

Anderson, K. and Hayami, Y. 1986. *The Political Economy of Agricultural Protection: East Asia in International Perspective*. London: Allen and Unwin.

Antle, J. 1987. Econometric estimation of risk producers' risk attitudes. *American Journal of*

Agricultural Economics 69, 509–22.

Ardeni, P. and Freebairn, J. 2002. The macroeconomics of agriculture. In *Handbook of Agricultural Economics*, vol. 2A, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Barrett, C.B. 2002. Food security and food assistance. In *Handbook of Agricultural Economics*, vol. 2B, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Beneke, R. 1998. T.W. Schultz and Pamphlet No. 5: The Oleo Margarine War and Academic Freedom. *Choices* 2nd Quarter, 4–8.

Berndt, E. 1991. *The Practice of Econometrics: Classical and Contemporary*. Reading, MA: Addison-Wesley.

Bhide, S. 1994. The science of agricultural economics in India: implications of Heady's contributions. In *Earl O. Heady: His Impact on Agricultural Economics*, ed. J. Langley, G. Vocke and L. Whiting. Ames: Iowa State University Press.

Binswanger, H. 1974. A microeconomic approach to induced innovation. *Economic Journal* 84, 940–58.

Black, J. 1926. *Introduction to Production Economics*. New York: Henry Holt and Co.

Boussard, J. and Petit, M. 1967. Representation of farmers’ behavior under uncertainty with a focus loss constraint. *Journal of Farm Economics* 49, 869–80.

Brewster, J. 1959. The impact of technical advance and migration on agricultural society and policy. *Journal of Farm Economics* 41, 1169–84.

Bromley, D. 1991. *Environment and Economy: Property Rights and Public Policy*. Blackwell: Oxford.

Buck, J. 1943. *An Agricultural Survey of Szechwan Province*. Chungking: Farmers Bank of China.

Buck, J. 1973. Development of agricultural economics at the University of Nanking, Nanking, China 1920–1946. Agricultural Development Bulletin No. 25. Ithaca, NY: Cornell University.

Buck, J., Dawson, O. and Wu, Y. 1966. *Food and Agriculture in Communist China*. New York: Praeger.

Burt, O. 1966. Economic control of groundwater reserves. *Journal of Farm Economics* 48, 632–47.

Burt, O. and Allison, J. 1963. Farm management decisions with dynamic programming. *Journal of Farm Economics* 45, 121–36.

Burt, O. and Cummings, R. 1970. Production and investment in natural resource industries. *American Economic Review* 60, 576–90.

Campbell, K. 1985. Some reflections on the development of agricultural economics in Australia. In *Agricultural Economics in Australia*, ed. S. Bearman. Armidale: Department of Agricultural Business and Management, University of New England.

Caswell, M., Lichtenberg, E. and Zilberman, D. 1990. The effects of pricing policies on water conservation and drainage. *American Journal of Agricultural Economics* 72, 883–90.

Chambers, R. and Quiggin, J. 1996. Non-point-source Pollution Regulation as a Multi-task principal–agent problem. *Journal of Public Economics* 59, 95–116.

Chavas, J. 2002. Structural Change in agricultural production, economics, technology and policy. In *Handbook of Agricultural Economics*, vol. 1A, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Chenery, H. and Syrquin, M. 1975. *Development, 1950–1970*. Oxford: Oxford University Press.

Ciriacy-Wantrup, S. 1952. *Resource Conservation – Economics and Policies*. Berkeley: University of California Press.

Cochrane, W. 1953. Professor Schultz discovers the weather. *Journal of Farm Economics* 35, 280–3.

Cochrane, W. 1958. *Farm Prices: Myth and Reality*. Minneapolis: University of Minnesota Press.

Cochrane, W. 1983. Agricultural economics at the University of Minnesota 1886–1979. Miscellaneous Publication No. 21. Institute of Agriculture, Forestry and Home Economics, University of Minnesota.

Cochrane, W. 1993. *The Development of American Agriculture: A Historical Analysis*. Minneapolis: University of Minnesota Press.

Cochrane, W. and Runge, C. 1992. *Reforming Farm Policy: Toward a National Agenda*. Ames: Iowa State University Press.

Crutchfield, J. and Zellner, A. 1962. *Economic Aspects of the Pacific Halibut Industry*. Fishery Industrial Research No. 1. Washington, DC: United States Department of the Interior.

Cummings, R. and Winkelmann, D. 1970. Water resource management in arid environs. *Water Resources Research* 6, 1559–68.

Day, R. 1963. *Recursive Programming and Production Response*. Amsterdam: North-Holland.

Day, R. and Sparling, E. 1977. Optimization models in agricultural and resource economics. In *A Survey of Agricultural Economics Literature*, vol. 2, ed. L. Martin. Minneapolis: University of Minnesota Press.

De Haen, H. and Heidhues, T. 1973. Recursive programming models to simulate agricultural development – applications in West Germany. Working Paper No. 18. Göttingen: Institute for Agricultural Economics.

Dillon, J. 1971. An expository review of Bernoullian decision theory in Agriculture: is utility futility? *Review of Marketing and Agricultural Economics* 39, 3–80.

Dillon, J. and Scandizzo, P. 1978. Risk attitudes of subsistence farmers in Northeast Brazil: a sampling approach. *American Journal of Agricultural Economics* 60, 425–35.

Egbert, A. and Heady, E. 1961. Regional adjustments in grain production: a linear programming analysis. Technical Bulletin No. 1241. Washington, DC: USDA.

Eltis, W. 1975. François Quesnay: a reinterpretation. 1. The Tableau Economique. *Oxford Economic Papers* 27, 167–200.

Evenson, R. 1967. The contribution of agricultural research to production. *Journal of Farm Economics* 49, 1415–25.

Evenson, R. and Kislev, Y. 1976. A stochastic model of applied research. *Journal of Political Economy* 84, 265–81.

Ezekiel, M. 1930. *Methods of Correlation Analysis*. New York: Wiley.

Ezekiel, M. 1938. The cobweb theorem. *Quarterly Journal of Economics* 52, 255–80.

Fox, K. 1953. A spatial equilibrium model of livestock-feed economy in the United States. *Econometrica* 21, 547–66.

Galbraith, J. 1959. John D. Black: a portrait. In *Economics for Agriculture*, ed. J. Cavin. Cambridge, MA: Harvard University Press.

Georgescu-Roegen, N. 1960. Economic theory and agrarian economics. *Oxford Economic Papers* 12, 1–40.

Giles, A. 2000. The AES, 1926–2001: a view from the archives. *Journal of Agricultural Economics* 52,

1–19.

Goodwin, B. and Smith, V. 1995. *The Economics of Crop Insurance and Disaster Aid*. Washington, DC: AEI Press.

Gray, R. and Rutledge, D. 1971. The economics of commodity futures markets: a survey. *Review of Marketing and Agricultural Economics* 39, 57–108.

Griliches, Z. 1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25, 501–22.

Griliches, Z. 1958. Research costs and social returns: hybrid corn and related innovations. *Journal of Political Economy* 66, 419–31.

Griliches, Z. 1963. The sources of measured productivity growth: United States agriculture, 1940–1960. *Journal of Political Economy* 71, 331–46.

Griliches, Z. 1964. Research expenditures, education, and the aggregate agricultural production function. *American Economic Review* 54, 961–74.

Gustafson, R. 1958. Implications of recent research on optimal storage rules. *Journal of Farm Economics* 40, 290–300.

Hanf, C. 1988. Adjustment necessities in European agricultural economics research. *European Review of Agricultural Economics* 15, ix–xvi.

Hayami, Y. and Ruttan, V. 1971. *Agricultural Development: An International Perspective*. Baltimore, MD: Johns Hopkins University Press.

Hazell, P. 1971. A linear alternative to quadratic and semivariance programming for farm planning under uncertainty. *American Journal of Agricultural Economics* 53, 53–62.

Heady, E. 1951. A production function and marginal rates of substitution in the utilization of feed resources by dairy cows. *Journal of Farm Economics* 33, 485–98.

Heady, E. and Candler, W. 1958. *Linear Programming Methods*. Ames: Iowa State University Press.

Heady, E. and Dillon, J. 1961. *Agricultural Production Functions*. Ames: Iowa State University Press.

Heidhues, T. 1966. A recursive programming model of farm growth in northern Germany. *Journal of Farm Economics* 48, 668–84.

Henderson, J. 1959. The utilization of agricultural land: a theoretical and empirical inquiry. *Review of Economics and Statistics* 41, 242–59.

Hertel, T., ed. 1997. *Global Trade Analysis Modeling and Applications*. New York: Cambridge University Press.

Hicks, J. 1932. *The Theory of Wages*. London: Macmillan.

Hicks, J. 1939. *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory*. Oxford: Clarendon Press.

Hildreth, C. 1957a. Problems of uncertainty in farm planning. *Journal of Farm Economics* 39, 1430–41.

Hildreth, C. 1957b. Some problems and possibilities of farm programming. In *Fertilizer Innovations and Resource Use*, ed. E. Baum et al. Ames: Iowa State College Press.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Hueth, B. and Ligon, E. 2001. Agricultural markets as relative performance evaluation. *American Journal of Agricultural Economics* 83, 318–28.

Huffman, W. 1974. Decision making: the role of education. *American Journal of Agricultural Economics* 56, 85–97.

Huffman, W. 2002. Human capital, education and agriculture. In *Handbook of Agricultural Economics*, vol. 1A, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Jesness, O. 1923. *The Cooperative Marketing of Farm Products*. Philadelphia: Lippencott.

Johnson, D. 1977. Postwar policies relating to trade in agricultural products. In *A Survey of Agricultural Economics Literature*, ed. L. Martin. Minneapolis: University of Minnesota Press.

Jorgenson, D. and Griliches, Z. 1967. The explanation of productivity change. *Review of Economic Studies* 34, 249–83.

Judge, G. and Takayama, T. 1973. *Studies in Economic Planning over Space and Time*. Amsterdam: North–Holland.

Just, R. and Antle, J. 1990. Interactions between agricultural and environmental Policies: a conceptual framework. *American Economic Review* 80, 197–202.

Karp, L. and Perloff, J. 2002. A synthesis of agricultural trade economics. In *Handbook of Agricultural Economics*, vol. 2B, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Kerr, N. 1987. *The Legacy: A Centennial History of the State Agricultural Experiment Stations 1887–1987*. Columbia: Missouri Agricultural Experiment Station.

Kislev, Y. and Peterson, W. 1982. Prices, technology and farm size. *Journal of Political Economy* 90, 578–95.

Klyukach, V. 2004. Stages in the formation and modern development of Russian agricultural economics. *Ekonomika Sel'skokhozyaistvennykh I Pererabatyvayushchikh Predpriyatii* 5, 10–13.

Knoeber, C. and Thurman, W. 1995. ‘Don't Count Your Chickens… ’: risk and risk shifting in the broiler industry. *American Journal of Agricultural Economics* 77, 486–96.

Kreuger, A., Schiff, M. and Valdes, A. 1991–94. *The Political Economy of Agricultural Pricing Policy*, 4 vols. Baltimore, MD: Johns Hopkins University Press.

Lau, L. and Yotopoulos, P. 1971. A test for relative efficiency and application to Indian agriculture. *American Economic Review* 61, 94–109.

Leontief, W. 1971. Theoretical assumptions and non-observed facts. *American Economic Review* 61, 1–7.

Lewis, W. 1954. Economic development with unlimited supplies of labour. *The Manchester School* 22, 3–42.

Lichtenberg, E. 2002. Agriculture and the environment. In *Handbook of Agricultural Economics*, vol. 2A, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Marshall, A. 1890. *Principles of Economics*, 1st edn. London: Macmillan and Co.

Mellor, J. 1966. *The Economics of Agricultural Production*. Ithaca, NY: Cornell University Press.

Miller, G. 1985. Agricultural economics in Canberra. In *Agricultural Economics in Australia*, ed. S. Bearman. Armidale: Department of Agricultural Business and Management, University of New England.

Miranda, M. and Glauber, J. 1997. Systematic risk, reinsurance, and the failure of crop insurance markets. *American Journal of Agricultural Economics* 79, 206–15.

Moscardi, E. and de Janvry, A. 1977. Attitudes toward risk among peasants: an econometric approach.

American Journal of Agricultural Economics 59, 710–16.

Moschini, G. and Hennessey, D. 2002. Uncertainty, risk aversion and risk management for agricultural producers. In *Handbook of Agricultural Economics*, vol. 2B, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Moore, G. 1988. The involvement of experiment stations in secondary agricultural education, 1887–1917. *Agricultural History* 62, 164–76.

Myers, R. 1989. Econometric testing for risk averse behavior in agriculture. *Applied Economics* 21, 542–52.

Myers, W.H., Womack, A., Johnson, S., Brandt, J. and Young, R. 1987. Impacts of alternative programs indicated by FAPRI analysis. *American Journal of Agricultural Economics* 69, 972–9.

Nazarenko, V. 2004. Conflict in Soviet agricultural economics in the 1920s and 1930s. *Ekonomika Sel'skokhozyaistvennykh I Pererabatyvayushchikh Predpriyatii* 6, 8–11.

Nelson, R. and Phelps, E. 1966. Investment in humans, technological diffusion, and economic growth. *American Economic Review* 56(2), 69–75.

Nerlove, M. 1958. *The Dynamics of Supply, Estimation of Farmers Response to Price*. Baltimore, MD: Johns Hopkins University Press.

Newbury, D. and Stiglitz, J. 1981. *The Theory of Commodity Price Stabilization – A Study in the Economics of Risk*. Oxford: Clarendon.

N ō u, J. 1967. *Studies in the Development of Agricultural Economics in Europe*. Uppsala: Almquist and Wiksells.

Olsen, W. et al. 1991. *Agricultural Economics and Rural Sociology: The Contemporary Core Literature*. Ithaca, NY: Cornell University Press.

Peterson, W. 1967. Return to poultry research in the United States. *Journal of Farm Economics* 49, 656–69.

Peterson, W. 1969. The allocation of research, teaching and extension personnel in U.S. colleges of agriculture. *American Journal of Agricultural Economics* 51, 41–56.

Peterson, W. and Hayami, Y. 1977. Technical change in agriculture. In *A Survey of Agricultural Economics Literature*, vol. 1, ed. L. Martin. Minneapolis: University of Minnesota Press.

Petit, M. 1982. Is there a French School of Agricultural Economics? *Journal of Agricultural Economics* 23, 325–37.

Quesnay, F. 1758. *The Economical Table*. Milton Keynes: Lightning Source Books, 2004.

Raeburn, J. and Jones, J. 1990. *The History of the International Association of Agricultural Economists*. Brookfield, VT: Gower Publishing Company.

Ricardo, D. 1821. *The Principles of Political Economy and Taxation*. London: M. Dent and Sons, 1911.

Roumasset, J. 1976. *Rice and Risk: Decisionmaking Among Low Income Farmers*. Amsterdam: Elsevier.

Runge, C. 1981. Common property externalities: isolation, assurance and resource depletion in a traditional grazing context. *American Journal of Agricultural Economics* 63(4), 595–606.

Runge, C. 1987. Induced agricultural innovation and environmental quality: the case of groundwater regulation. *Land Economics* 63, 249–58.

Runge, C., Senauer, B., Pardey, P. and Rosegrant, M. 2003. *Ending Hunger in Our Lifetime: Food Security and Globalization*. Baltimore, MD: Johns Hopkins University Press.

Ruttan, V. 1956. The contribution of technological progress to farm output, 1950–75. *Review of Economics and Statistics* 38, 61–9.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Schmitz, A. and Seckler, D. 1970. Mechanized agriculture and social welfare: the case of the tomato harvester. *American Journal of Agricultural Economics* 52, 569–77.

Schuh, G. 1974. The exchange rate and U.S. agriculture. *American Journal of Agricultural Economics* 56, 1–13.

Schuh, G. 1976. The new macroeconomics of agriculture. *American Journal of Agricultural Economics* 58, 802–11.

Schultz, T. 1953. *The Economic Organization of Agriculture*. New York: McGraw Hill.

Schultz, T. 1964. *Transforming Traditional Agriculture*. New Haven, CT: Yale University Press.

Schultz, T. 1971. *Investment in Human Capital*. New York: Free Press.

Scott, A. 1955. The fishery: the objectives of sole ownership. *Journal of Political Economy* 63, 116–24.

Segerson, K. 1988. Uncertainty and incentives for nonpoint pollution control. *Journal of Environmental Economics and Management* 15(1), 87–98.

Sexton, R. 1990. Imperfect competition in agricultural markets and the role of Cooperatives: a spatial analysis. *American Journal of Agricultural Economics* 72, 709–20.

Sexton, R. and Lavoie, N. 2002. Food processing and distribution: an industrial organization approach. In *Handbook of Agricultural Economics*, vol. 2B, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Indianapolis, IN: Liberty Classics, 1981.

Sumner, D. and Tangermann, S. 2002. International trade policy and negotiations. In *Handbook of Agricultural Economics*, vol. 2B, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Sunding, D. and Zilberman, D. 2002. The agricultural innovation process: research and technology adoption in a changing agricultural sector. In *Handbook of Agricultural Economics*, vol. 1A, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Taylor, H. 1905. *An Introduction to the Study of Agricultural Economics*. New York: Macmillan.

Taylor, H. 1922. The history of the development of the Farm Economic Association. *Journal of Farm Economics* 4(2), 92–100.

Taylor, H. and Taylor, A. 1952. *The Story of Agricultural Economics in the United States, 1940–1932*. Ames: Iowa State University Press.

Timmer, C. 2002. Agriculture and economic development. In *Handbook of Agricultural Economics*, vol. 2A, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

Tomek, W. and Robinson, K. 1977. Agricultural price analysis and outlook. In *A Survey of Agricultural Economics Literature*, vol. 1, ed. L. Martin. Minneapolis: University of Minnesota Press.

Tolley, H., Black, J. and Ezekiel, M. 1924. Input as related to output in farm organization and cost of production studies. Bulletin No. 1277. Washington, DC: USDA.

Veblen, T. 1900. The preconception of economic science, III. *Quarterly Journal of Economics* 14, 240–69.

Von Thünen, J. 1828. *Isolated State* (English translation). Oxford: Oxford University Press, 1966.

Walker, J., Gardner, R. and Ostrom, E. 2000. Collective choice in the commons: experimental results on proposed allocation rules and votes. *Economic Journal* 110, 212–34.

Waugh, F. 1928. Quality factors influencing vegetable prices. *Journal of Farm Economics* 10(2), 185–96.

Waugh, F. 1951. The minimum-cost dairy feed. *Journal of Farm Economics* 33, 299–305.

Working, E. 1927. What do statistical ‘demand curves’ show? *Quarterly Journal of Economics* 41, 212–35.

Working, H. 1922. Factors determining the price of potatoes in St. Paul and Minneapolis. Agricultural Experiment Station Technical Bulletin No. 10, University of Minnesota.

Working, H. 1925. The Statistical determination of demand curves. *Quarterly Journal of Economics* 39 (August), 503–43.

Wright, B. 2002. Storage and price stabilization. In *Handbook of Agricultural Economics*, vol. 1B, ed. B. Gardner and G. Rausser. Amsterdam: North-Holland.

How to cite this article

Runge, C. Ford. "agricultural economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000243> doi:10.1057/9780230226203.0021

The New Palgrave Dictionary of Economics Online

agricultural finance

Michael R. Carter

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic analysis of agricultural finance has traditionally focused on access to capital in the agricultural sector. Key concerns have included patterns of non-price rationing in agricultural credit markets, the institutions and contracts that provide credit to agricultural producers, the implications of the conditions of capital access on agricultural growth and rural income distribution, and the role of the public sector in agricultural credit markets. More recently, the analysis of agricultural finance has expanded beyond these credit-centred concerns to consider systemic approaches to rural finance that address risk and insurance, savings services and the provision of credit.

Keywords

adverse selection; agricultural banks; agricultural finance; agricultural markets in developing countries; asymmetric information; credit markets in developing countries; credit rationing; Grameen Bank; Green Revolution; microcredit; moneylenders in developing countries

Article

Several structural features of the agricultural sector make agricultural finance and financial markets distinctive. First, the demand for agricultural finance is potentially high. Agricultural production processes are roundabout, with outputs and returns coming months or even years (in the case of vineyards and tree crops) after expenditures on productive inputs. The extreme riskiness of agriculture further increases the demand for credit or other contracts that share the risk of the production process. A second distinguishing feature of agricultural finance is that the organization of agricultural production makes it difficult to supply with financial services. In a classic paper, John Brewster (1950) noted that agricultural production differed from industrial production because of its spatial dispersion and its heavy dependence on inherently random inputs provided by nature. These features create what more contemporary economic analysis would call agency problems, meaning that it is difficult for an outsider

to either monitor directly the quality of labour and management on a farm, or to infer *ex post* the qualities of those inputs from final agricultural output. As Brewster and others have remarked, the result is that agriculture tends to be organized in small-scale units, with much of the labour and management provided by the residual claimant to the production process (that is, it is rare to find large-scale ‘factories in the field’ except in special historical circumstances, as discussed by Binswanger, Deininger and Feder, 1995).

Excess demand for financial services

Agriculture thus stands as a sector with potentially high demand for financial services coming from relatively small-scale, spatially disperse, hard-to-monitor firms. In the contemporary low-income countries of Asia, Africa and Latin America, where the vast majority of farming households operate tiny holdings of an acre or two, between 5 and 15 per cent of producers have formal financial contracts (Braverman and Huppi, 1991). Others are observed to borrow from a variety of informal sources, typically at nominal interest rates well in excess of those charged by formal financial institutions (Braverman and Guasch, 1986).

While these observations are not by themselves sufficient to identify an excess demand for financial services in agriculture, they are consistent with it. Bolstering this interpretation is the fact that the characteristics of agriculture conform closely to the assumptions that underlie the formal economic theory of credit rationing. The seminal analysis of Stiglitz and Weiss (1981) assumes precisely the sorts of information costs and asymmetries that typify an agricultural sector comprising numerous, spatially disperse firms producing a highly random output. As extended by Carter (1988), this theoretical perspective suggests that adverse incentive and selection effects will prevent competitive formal lenders raising interest rates to market clearing levels (because higher rates result in lower expected profits for lenders as the borrowers still left in the market become increasingly less desirable as clients as interest rates increase). The result, according to this theory, is an agricultural credit market characterized by excess demand for formal credit and by a skewed allocation of (relatively cheap) formal credit toward larger farm units.

Some of this residual excess demand would be expected to spill over to locally based informal agents (moneylenders, input suppliers and processors). These lenders typically enjoy the twin advantages of cheaper information (because they are local) and the capacity to accept collaterals that could not be easily claimed by distant lenders (such as standing crops). Whether these agents are competitive suppliers of credit, or whether they enjoy spatial monopolies that grant them real market power, remains an open question (see, for example, Kochar, 1997; Bell, Srinivasan and Udry, 1997).

Implications of excess demand

While there is thus still debate about the degree of excess demand for financial services in agriculture, its implications are potentially large at two levels. First, excess demand for finance may result in slower agricultural technological change and growth. Again, examples from low-income countries make this point most easily. A study of new, input-intensive agricultural export products in Central America found that annual working capital requirements per hectare exceeded the total annual incomes that farm

families had been earning (Barham, Carter and Sigelko, 1995). The questionable ability of these families to self-finance investments of this magnitude, and to self-insure against the estimated 25 per cent failure rate of these activities, makes clear the economic costs of excess demand for agricultural finance. The deep and well-developed literature on the constrained adoption of input-intensive Green Revolution technologies ratifies this point.

In addition to its effects on the level and growth of agricultural incomes, excess demand for agricultural finance may also have impacts on income distribution within the rural economy. The theoretical analysis of Eswaran and Kotwal (1986) is especially instructive in this regard. Using a single-period general equilibrium model, they show that skewed access to capital, which leaves lower-wealth producers with excess demand for credit, will shift land access and income away from small-scale producer households, despite the intrinsic labour monitoring advantages enjoyed by these producers. The result is an agricultural economy that produces less, and distributes it less equally, than it would in a world of perfect financial markets. Eswaran and Kotwal go on to show that, under these conditions, an agricultural economy can become a prisoner of its own history. Economies that begin with relatively unequal wealth distributions tend to maintain them, while initially more egalitarian economies create more equal income distributions.

More recent theoretical analysis has used dynamic methods to extend the Eswaran and Kotwal analysis, asking whether the effects of excess demand for credit will be so long-lived and dramatic when credit-constrained and other agents have the option of building up their own sources of self-finance via savings over time. While not explicitly focused on agriculture, the analysis of Banerjee and Newman (1993) was an important demonstration that inadequate access to capital can fundamentally distort the occupational and production structure of an economy over the long term. Subsequent work has continued to build on this analytical tradition and has, among other things, shown that inadequate access to capital (in the presence of risk) can lead to a type of structural bifurcation in the agricultural economy. Initially wealthier producers move to a higher level of equilibrium well-being, while the initially poor become mired in a low-level poverty trap (see, for example, Dercon, 1998; Mookherjee and Ray, 2000; Zimmerman and Carter, 2003).

Policy debates

While much of this literature on the costs of inadequate access to capital in agriculture is relatively recent, the sense that agricultural financial markets are fundamentally imperfect has driven generations of policy interventions in both high- and low-income nations. Historically, these interventions have included the direct provision of agricultural credit by public lenders, often at subsidized rates. For example, in the United States in 2002 more than 40 per cent of all farm debt to institutional lenders was held by two public entities, the Farm Credit System and the Farm Service Agency (USDA, 2004). While still large, the public provision of agricultural credit in the United States has been trending downward for sometime, signalling the even larger role played by state credit in an earlier era when farms in the United States were smaller and more numerous.

In the low-income countries of Asia, Africa and Latin America, state agricultural banks and other mechanisms of public credit provision became a common feature of the agricultural landscape in the 1960s and 1970s. Interest rates were typically subsidized, and these interventionist policies were

justified on the grounds that private provision of capital was either inadequate, priced at extortionate terms, or simply unavailable, especially for smaller farmers.

However, by the early 1980s, a coherent critique of these policies had emerged, arguing that state banks were financially unsustainable, crowded out private financial institutions, and did not even succeed in channelling credit to small-scale agricultural producers (see Adams, Graham and von Pischke, 1984).

Under the pressure of structural adjustment and the broader move toward economic liberalization, state agricultural banks began to disappear from the developing country landscape, and in Latin America, at least, were almost completely gone by the mid-1990s.

While commercial lending to agriculture continues to expand in the United States, the prediction by some that private institutional lenders would fill the gap left by public banks in Latin America and elsewhere in the developing world has been largely unfulfilled (Wenner, Alvarado and Galarza, 2003).

While in a few instances there has been renewed interest in public provision of agricultural finance, contemporary policy discussion largely focuses on three alternatives. The first is the provision of agricultural credit by non-financial businesses, such as input suppliers and commodity warehouses. The informational advantages of these informal lenders that permit them to monitor borrowers and lend where formal banks cannot has been more fully developed in recent theoretical literature (Conning, 1999). As mentioned above, this sector remains enigmatic in terms of its efficiency and competitiveness.

Nonetheless, there is increasing interest in the reform of collateral laws that might open the door to an expansion of lending by these businesses (Fleisig and de la Peña, 2003). Others have argued that a general strengthening of legally weak landownership rights through systematic land titling programmes will induce greater entry into agricultural markets by private financial institutions (Feder and Akihiko, 1999). However, evidence to date that land title bolsters formal credit supply to agricultural producers (especially small-scale producers) remains thin (Carter and Olinto, 2003).

Micro-finance providers are a second alternative for the future provision of agricultural finance. Like informal lenders, micro-finance institutions can tap into cheap, locally available information about borrowers and their behaviour. They also utilize non-standard collateral assets, including group repayment guarantees in the case of micro-finance programmess that build on the Grameen Bank model of sequenced group loans. However, as Zeller and Meyer (2002) and others have discussed, the very localness of micro-finance institutions (which is the informational key to their ability to lend to small-scale, dispersed borrowers) can become a liability in weather-dependent agriculture where risks across borrowers are strongly correlated. Unlocking the potential for micro-finance lending to provide agricultural credit may thus require mechanisms to insure microfinance lenders, or their clients, against correlated weather risks. Pilot programmes to do just that are currently under development by the World Bank and others (Skees and Barnett, 1999).

The third and final approach to the conundrum of agricultural finance is a more general systemic approach to developing rural (not necessarily agricultural) financial institutions. Motivated in part by the observation that farm families in both wealthy and developing nations derive much of their income from non-agricultural sources, this systemic approach advocates legal and institutional reforms designed to promote the expansion of full-service financial intermediaries in rural areas (Gonzalez-Vega, 2003).

Among these reforms are efforts to establish credit bureaus and other institutions that share borrowers' credit history across multiple lenders. Work such as that by Jappelli and Pagano (2002) suggests that the credit expansion effects of such institutions can be substantial. However, as with the other novel

approaches described here, there is much yet to learn about whether these systemic approaches will suffice to improve the operation of financial markets in agriculture.

See Also

- agricultural markets in developing countries
- credit rationing
- micro-credit
- moneylenders in developing countries

Bibliography

Adams, D., Graham, D. and von Pischke, J. 1984. *Undermining Rural Development with Cheap Credit*. Boulder, CO: Westview Press.

Banerjee, A. and Newman, N. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101, 274–98.

Barham, B., Carter, M. and Sigelko, W. 1995. Agro-export production and peasant land access, examining the dynamic between adoption and accumulation. *Journal of Development Economics* 46, 85–107.

Bell, C., Srinivasan, T. and Udry, C. 1997. Rationing, spillover and interlinking in credit markets: the case of rural Punjab. *Oxford Economic Papers* 49, 557–85.

Binswanger, H., Deininger, K. and Feder, G. 1995. Power, distortions, revolt and reform in agricultural and land relations. In *Handbook of Development Economics*, vol. IIIB, ed. J. Behrman and T. Srinivasan. Amsterdam: North-Holland.

Braverman, A. and Guasch, J. 1986. Rural credit markets and institutions in developing countries: lessons for policy analysis from practice and modern theory. *World Development* 14, 1253–67.

Braverman, A. and Huppi, M. 1991. Improving rural finance in developing countries. *Finance and Development* 28, 42–4.

Brewster, J. 1950. The machine process in agriculture and industry. *Journal of Farm Economics* 32, 69–81.

Carter, M. 1988. Equilibrium credit rationing of small farm agriculture. *Journal of Development Economics* 28, 83–103.

Carter, M. and Olinto, P. 2003. Getting institutions right for whom? Credit constraints and the impact of property rights on the quantity and composition of investment. *American Journal of Agricultural Economics* 85, 173–86.

Conning, J. 1999. Outreach, sustainability and leverage in monitored and peer-monitored lending. *Journal of Development Economics* 60, 51–77.

Dercon, S. 1998. Wealth, Risk and Activity Choice, Cattle in Western Tanzania. *Journal of Development Economics* 55, 1–42.

Eswaran, M. and Kotwal, A. 1986. Access to capital and agrarian production organization. *Economic Journal* 96, 482–98.

Feder, G. and Akihiko, N. 1999. The benefits of land titling and registration: economic and social perspectives. *Land Policy Studies* 15, 25–43.

Fleisig, H. and de la Peña, N. 2003. Legal and regulatory requirements for effective rural finance markets. Paper presented at the Paving the Way Forward for Rural Finance Conference, Washington, DC, 2–3 June. Online. Available at http//www.basis.wisc.edu/rfc/documents/theme_legal.pdf, accessed 21 November 2005.

Gonzalez-Vega, C. 2003. Deepening rural financial markets: macroeconomic, policy and political dimensions. Paper presented at the Paving the Way Forward for Rural Finance Conference, Washington, DC, 2–3 June. Online. Available at http//www.basis.wisc.edu/rfc/documents/theme_macro.pdf, accessed 21 November 2005.

Jappelli, T. and Pagano, M. 2002. Information sharing, lending and defaults: cross-country evidence. *Journal of Banking and Finance* 26, 2017–45.

Kochar, A. 1997. An empirical investigation of rationing constraints in rural markets in India. *Journal of Development Economics* 53, 339–72.

Mookherjee, D. and Ray, D. 2000. Contractual structure and wealth accumulation. *American Economic Review* 92, 818–49.

Skees, J. and Barnett, B. 1999. Conceptual and practical considerations for sharing catastrophic / systemic risks. *Review of Agricultural Economics* 21, 424–41.

Stiglitz, J. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

USDA (US Department of Agriculture). 2004. *Agriculture Income and Finance Outlook/AIS–82*. Washington, DC: Economic Research Service, USDA.

Wenner, M., Alvarado, J. and Galarza, F. 2003. *Promising Practices in Rural Finance, Experiences from Latin America and the Caribbean*. Lima: Inter-American Development Bank.

Zeller, M. and Meyer, R. 2002. *The Triangle of Microfinance, Outreach, Financial Sustainability and Impact*. Baltimore, MD: Johns Hopkins University Press.

Zimmerman, F. and Carter, M. 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. *Journal of Development Economics* 71, 233–60.

How to cite this article

Carter, Michael R. "agricultural finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000064> doi:10.1057/9780230226203.0022

The New Palgrave Dictionary of Economics Online

agricultural markets in developing countries

Christopher B. Barrett and Emelly Mutambatsere

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The history of agricultural markets in developing countries reflects attempts to establish the appropriate government responses to the inefficiencies created by incomplete institutional and physical infrastructure and imperfect competition. Government intervention in the 1960s and 1970s to resolve market failures gave way in the 1980s to market-oriented liberalization to ‘get prices right’ and, more recently, to ‘get institutions right’. But market openness may accentuate the latent dualism of a modern, efficient marketing sector, accessible only to those with adequate scale and capital, alongside a traditional, inefficient marketing channel to which the poor are effectively restricted.

Keywords

agricultural markets in developing countries; agriculture and economic development; border parity pricing; commodity prices; contract enforcement; contract farming; cooperatives; corruption; credit unions; dual economies; enforceability of contracts; first fundamental theorem of welfare economics; foreign direct investment (FDI); General Agreement on Tariffs and Trade (GATT); Green Revolution; imperfect competition; imperfect information; incomplete contracts; infrastructure; liberalization: of agriculture; liquidity constraints; market failure; marketing boards; monopsony; non-tariff barriers; oligopsony; price controls; price stabilization; property rights; public goods; rule of law; structural adjustment; subsidies; supermarkets; tariffs; transaction costs; World Trade Organization (WTO)

Article

Markets aggregate demand and supply across actors at different spatial and temporal scales. Well-functioning markets ensure that macro and sectoral policies change the incentives and constraints faced by micro-level decision makers. Macro policy commonly becomes ineffective without market transmission of the signals sent by central governments. Similarly, well-functioning markets underpin important opportunities at the micro level for welfare improvements that aggregate into sustainable

macro-level growth. For example, without good access to distant markets that can absorb excess local supply, the adoption of more productive agricultural technologies typically leads to a drop in farm-gate product prices, erasing all or many of the gains to producers from technological change and thereby dampening incentives for farmers to adopt new technologies that can stimulate economic growth. Markets also play a fundamental role in managing risk associated with demand and supply shocks by facilitating adjustment in net export flows across space and in storage over time, thereby reducing the price variability faced by consumers and producers. Markets thus perform multiple valuable functions: distribution of inputs (such as fertilizer, seed) and outputs (such as crops, animal products) across space and time, transformation of raw commodities into value-added products, and transmission of information and risk. Per the first welfare theorem, competitive market equilibria help ensure an efficient allocation of resources so as to maximize aggregate welfare.

The micro-level realities of agricultural markets in much of the developing world, however, include poor communications and transport infrastructure, limited rule of law, and restricted access to commercial finance, all of which make markets function much less effectively than textbook models typically assume. A long-standing empirical literature documents considerable commodity price variability across space and seasons in developing countries, with various empirical tests of market integration suggesting significant and puzzling forgone arbitrage opportunities, significant entry and mobility barriers, and highly personalized exchange (Barrett, 1997; Platteau, 2000; Fackler and Goodwin, 2001; Fafchamps, 2004). Widespread inefficiencies result from incomplete or unclear property rights, imperfect contract monitoring and enforcement, high transactions costs, and binding liquidity constraints. Such failures often motivate government intervention in markets, although interventions have often done more harm than good, either by distorting incentives or by creating public sector market power. The history of agricultural markets in developing countries reflects evolving thinking on the appropriate role for government in trying to address the inefficiencies created by incomplete institutional and physical infrastructure and imperfect competition. The emphasis in the 1960s and 1970s on government intervention to resolve market failures gave way in the 1980s to market-oriented liberalization to ‘get prices right’ and, more recently, to a focus on ‘getting institutions right’.

Past approaches

Agricultural marketing of most major export and food commodities and of modern inputs – such as fertilizer, machinery and hybrid seed – was historically highly regulated by developing country governments into the 1980s, via input price controls and subsidies, oligopolistic input markets, monopsonistic produce marketing boards, pan-seasonal and pan-territorial administrative commodity pricing, oligopolistic processing industries, and fixed wholesale and retail prices. Commodity prices were generally set below market levels, implicitly taxing producers while subsidizing consumers. Marketing channels were typically very inefficient, with centralized storage and processing facilities and government-imposed grades and standards for product quality, although these were not always and everywhere enforced. Sometimes these inefficient systems provided satisfactory coordination of marketing channels, but that was by no means universal. Heavy government presence, especially pan-seasonal and pan-territorial producer pricing, and fixed retail pricing systems and bans on private commerce effectively eliminated most incentives for private arbitrage or investment in fixed capital by marketing intermediaries. Meanwhile, management by government fiat too often facilitated corruption,

which often had a devastating long-run impact on economic governance.

In addition to state-run marketing boards, producer marketing cooperatives were prevalent in developing countries at all levels of the marketing chain, ranging from credit unions through farmer cooperatives to wholesale-level cooperatives. Credit unions commonly accumulated funds for input purchase or served as intermediaries for government-subsidized credit programmes. Farmer marketing cooperatives typically facilitated bulk input procurement, price negotiation, and sharing of transportation costs. Wholesale cooperatives mainly assembled bulk commodity lots for sale into government processing and distribution channels. Cooperatives have often worked well in specialized production areas distant from major markets, and with homogenous production of not-so-perishable commodities such as coffee. However, due to high administrative and coordination costs, free-rider problems and political interference, cooperative systems have not lived up to expectations in most developing countries, and many have collapsed.

In contrast to the major export and domestic staple food crops, smaller-scale food commodities for domestic consumption, such as indigenous fruits and vegetables, have almost always operated on a free market basis, with little history of state intervention or price regulation. These markets are characterized by many cash, spot market transfers of product between intermediaries en route from producer to consumer, many small, non-specialized and unorganized buyers and sellers, few if any grades or standards, one-on-one (dyadic) price negotiations, poor market information systems, and mostly informal contracts, largely enforced through social networks (Fafchamps, 2004). Such marketing channels depend disproportionately on rural periodic markets prevalent in most of the developing world, arguably the closest one ever gets to a true ‘free market’: free of government regulation, subsidies and taxes, and lacking public goods such as physical infrastructure, contract law, public market price information systems, or codified product grades and standards. Indeed, they have been termed the ‘flea market economy’ by Fafchamps and Minten (2001).

The emerging problems of state agricultural market control

Given the inherent variability of agricultural production and the significance of agriculture in economic activity and general well-being in developing countries, price stabilization policies were long considered necessary for economic stability. However, a number of problems emerged. First, the fixing of commodity prices below market levels inevitably created a disincentive for agricultural producers. By the late 1970s, low producer prices had led to the stagnation of production and exports and to increased parallel market activity, including cross-border smuggling, in many developing countries, especially in those areas of Africa and Central America that were largely bypassed by the Green Revolution.

The second major problem was the fiscal and political sustainability of government agricultural market interventions. The inefficiencies of parastatal marketing boards, along with the repression of private market intermediation, led to unreliable supplies of consumer goods for politically important urban populations. Moreover, those inefficiencies, combined with the numerous subsidies and frequent corruption within government-controlled marketing channels, became too costly for central governments, which faced massive pressure from international donors in the 1980s and 1990s to trim expenditures and to eliminate price controls (Timmer, 1986).

Economic liberalization: market relaxation and state compression

Market-oriented agricultural policy reforms were a centrepiece of economic liberalization in developing countries in the 1980s and 1990s, commonly within the context of broader structural adjustment programmes designed to restore fiscal and current account balance, to reduce or eliminate price distortions, and to facilitate efficient price transmission so as to stimulate investment and production. The new focus was on re-establishing a close correspondence between local and world market prices, so-called border parity pricing. The withdrawal of the state from agricultural market intermediation, specifically price discovery, was seen as a necessary condition in getting prices right, itself a necessary condition for improving market efficiency and stimulating investment and productivity growth (Timmer, 1986).

The market-oriented reforms typically implemented by developing country governments included, on the input side, the liberalization of land and labour markets, decontrol and de-licensing of input production, supply and distribution, removal of input subsidies and price controls, closure of loss-making credit schemes, liberalization of credit markets, and reform of agricultural extension. On the output markets side, reforms included commodity price liberalization, the removal of parastatal monopoly power and commodity movement restrictions, and reduction in tariffs and quotas on imports. The net result of these reforms typically turned on the balance between the pro-competitive effects of reduced government interference in marketing operations – what Lipton (1993) termed ‘market relaxation’ – and the anti-competitive effects of reduction of public goods and services that underpin private market transactions – what Lipton (1993) termed ‘state compression’. Since the two phenomena were typically inextricable in agricultural liberalization initiatives, experiences varied markedly.

The empirical evidence suggests that commodity prices generally increased after market reforms, often stimulating an increase in production, especially of export crops. These price increases also facilitated the emergence of supermarket chains, export-oriented outgrower schemes and export processing zones, and a generalized stimulus to agro-industrialization in developing countries (Reardon and Barrett, 2000; Sahn, Dorosh and Younger, 1997). Increased investment in the downstream marketing channel has transformed the orientation of many agricultural markets from raw commodity towards processed product markets, and with this increased investment came increased competition. In countries such as Chile, India and South Africa, private firms now play a leading role in development of improved seed varieties, producing and distributing inputs, post-harvest processing and modern retailing through supermarkets and restaurant chains (Reardon et al., 2003; Reardon and Timmer, 2005). Both formal and informal traders entered agricultural commodity marketing channels as government controls fell away, from rural periodic markets all the way through urban retail markets.

However, market entry has tended to be limited to certain marketing niches not protected by capital, information or relationship barriers, with substantial bottlenecks in other areas such as inter-seasonal storage and motorized transportation. Neither widespread entry into market intermediation activities nor workably competitive markets emerged everywhere, let alone quickly. For example, because long-haul motorized transportation in rural markets tends to involve considerable sunk costs and some economies of scale due to poor road conditions and high vehicle maintenance costs, entry into this sector of the markets has often been limited after the removal of legal and policy barriers to entry (Barrett, 1997). Meanwhile, the end of pan-seasonal and pan-territorial administrative pricing has brought increased

price risk, with consequences for investment incentives facing both producers and market intermediaries (Barrett and Carter, 1999).

The elimination of input subsidies and removal of government monopsony power in crop marketing has also often led to reduced access to input financing and increased input prices. The withdrawal of parastatals from core input marketing activities created a void that the private sector often failed to fill due to underdeveloped physical communications, power and transport infrastructure, credit constraints and continued bureaucratic impediments that increased transactions costs for input suppliers. In addition, periodic state and donor-funded input programmes have often reduced profitability and frustrated private investments. Input credit schemes by processors have been used in the post-reform period in an attempt to overcome the low input use resulting from these access problems, for example in the cotton sectors of Mali and Uganda and horticultural export sectors of Kenya and Zimbabwe.

Although the level of reform implementation differed from country to country, in many cases reform was only partially implemented and policy reversals were common (Jayne and Jones, 1997, Kherallah et al., 2002). In important food and export markets, liberalization efforts have been prolonged and incomplete, reflecting the difficulty in relinquishing government control in the face of uncertainty and political pressures to intervene in order to resolve perceived inequities or inefficiencies in market performance. For example, parastatals remain active in the West African cotton sector, the southern African maize sector has not been fully liberalized, and in Indonesia BULOG continues to operate amid private marketing companies. The ebb and flow of market-oriented reforms and the frequency with which governments have engaged in policy reversals has made it terribly difficult to tease out clear patterns in the impact of liberalization measures on the performance of agricultural markets in developing countries.

Post-structural adjustment market reforms

As the weaknesses of reformed agricultural markets in developing countries became evident, development agencies' and governments' focus began to shift from merely 'getting prices right' to 'getting institutions right' so as to address market failures arising from imperfect information, contract enforcement and property rights, and insufficient provision of public goods. Such reforms have used non-price measures in an attempt to develop the public and private institutions necessary for efficient market operations and to reduce transactions costs and business risk.

The post-structural adjustment era has also coincided with international market deregulation through the GATT and its successor, the WTO. Bilateral, regional and global trade agreements have reduced tariff and non-tariff barriers to cross-border flows of raw and processed agricultural commodities, and increased the openness of financial markets, leading to increased capital flow into developing countries, especially in the form of foreign direct investment (FDI). Where structural adjustment reforms had substantially reduced state control over input and output markets, trade and FDI liberalization has paved the way for major investment in post-harvest processing and retailing in developing countries since the 1990s. This 'new' capital investment differs from the structural adjustment era reforms in that whereas the focus previously was upstream, in the input, production, and wholesale sectors, more recent emphasis, especially in private investment, has tended to be downstream, in food processing, retail and restaurant markets. The exceptionally rapid diffusion of supermarkets in developing countries, in particular, has also been driven by improved coordination and communication technologies in addition

to increased urbanization, lower prices of processed goods, increased per capita incomes in developing countries, as well as saturation and intense competition in foreign firms' home markets (Reardon and Barrett, 2000; Reardon et al., 2003). In Latin America, for example, supermarkets currently account for 50–60 per cent of national food retail sale, compared with only 10–20 per cent in the 1980s (Reardon et al., 2003; Reardon and Timmer, 2005).

The rise of supermarket and restaurant chains has changed the fundamental structure and operations of agricultural markets significantly, directing far more market power downstream, often to chains wholly or partly owned by multinational corporations. Commodity procurement by retailers has become more centralized, with consolidated buying points at a regional, even global, level. It is not uncommon for a major supermarket chain located in three different countries to consolidate its procurement in a few large growers in just one of those countries. Global food chains have also established regional procurement nodes – for example, Walmart throughout Asia and Latin America – and in-country commodity procurement for regional firms such as the China Resource Enterprise has been centralized from individual store level to provincial systems (Reardon et al., 2003). These structural shifts have increased contract farming and outgrower schemes between agro-industrial firms and farmers in developing countries, and production of non-staple foods has increased.

Increased foreign investment in agricultural markets in developing countries, however, has produced conflicting results. Increased industrialization of agricultural markets has fostered improved market efficiency and competitiveness, integration of formerly fragmented markets, product diversification through differentiation, and value addition and technology transfer. However, the rapid pace of structural change, with some developing countries accomplishing in a few years what developed countries accomplished over decades, has left limited room for adjustment by smaller, less well-informed and poorly capitalized market actors to new ways of doing business. There is thus growing concern that market openness may lead to the replacement of traditional processors by oligopsonistic multinationals, accentuating the latent dualism of a modern, efficient marketing sector accessible only to those with adequate scale and capital, alongside a traditional, inefficient marketing channel to which the poor are effectively restricted. The tendency towards selection of a few medium- to large-scale firms or producers capable of delivering consistent quality product at large volumes has toughened competition for structurally inefficient producers, and seems to have led to some crowding out of smaller producers (Reardon and Timmer, 2005). Local informal wholesalers and retailers have found themselves having to compete with bigger firms, both for the more efficient producers offering consistent product quality and throughput volumes, and for consumers seeking more services. The emergence of big, concentrated downstream private marketing intermediaries could also potentially lead, once again, to non-competitive agricultural marketing channels, effectively replacing government with private market power.

Increased contract farming, while offering significant potential for smaller growers in the form of guaranteed markets and prices for their produce often coupled with input credit and extension service, has evidently also reduced farmer bargaining power in negotiating contract conditions. These negotiations now take place bilaterally, between individual farmers and the large contracting firm, rather than via collective bargaining by farmer associations with government parastatals.

Conclusion

Agricultural markets play a crucial role in the process of economic development. Yet, by virtue of the spatial dispersion of producers and consumers, the temporal lags between input application and harvest, the variable perishability and storability of commodities, and the political sensitivity of basic food staples, agricultural markets are prone to high transactions costs, significant risks and frequent government interference. The relative power of developing country governments and private domestic or multinational firms in agricultural markets has varied over time. But the fundamental functions of input and output distribution, post-harvest processing and storage, as well as the persistent challenges of liquidity constraints, contract enforcement and imperfect information, have characterized agricultural markets in developing countries under all forms of organization.

See Also

- agriculture and economic development
- development economics
- dual economies
- foreign direct investment
- marketing boards
- spatial market integration

Bibliography

Barrett, C. 1997. Food marketing liberalization and trader entry: evidence from Madagascar. *World Development* 25, 763–77.

Barrett, C. and Carter, M. 1999. Microeconomically coherent agricultural policy reform in Africa. In *African Economies in Transition, Volume 2: The Reform Experiences*, ed. J. Paulson. London: Macmillan.

Fafchamps, M. and Minten, B. 2001. Property rights in a flea market economy. *Economic Development and Cultural Change* 49, 229–68.

Fafchamps, M. 2004. *Market Institutions in Sub-Saharan Africa*. Cambridge, MA: MIT Press.

Fackler, P. and Goodwin, B. 2001. Spatial price analysis. In *Handbook of Agricultural Economics*, Vol. 1B, ed. B. Gardner and G. Rausser. Amsterdam: Elsevier.

Jayne, T. and Jones, S. 1997. Food marketing and pricing policy in eastern and southern Africa: a survey. *World Development* 25, 1505–27.

Kherallah, M., Delgado, C., Gabre-Madhin, E., Minot, N. and Johnson, M. 2002. *Reforming Agricultural Markets in Africa*. Washington, DC: International Food Policy Research Institute.

Lipton, M. 1993. Market relaxation and agricultural development. In *States and Markets: Neo-liberalism and the Development Policy Debate*, ed. C. Colclough and J. Manor. Oxford: Oxford University Press.

Platteau, J.-P. 2000. *Institutions, Social Norms, and Economic Development*. London: Harwood Academic Publishers.

Reardon, T. and Barrett, C. 2000. Agroindustrialization, globalization, and international development: an overview of issues, patterns, and determinants. *Agricultural Economics* 23(3), 195–205.

Reardon, T. and Timmer, P. 2005. Transformation of markets for agricultural output in developing countries since 1950: how has thinking changed? In *Handbook of Agricultural Economics*, vol. 3A, ed. R. Evenson, P. Pingali and T. Schultz. Amsterdam: Elsevier.

Reardon, T., Timmer, P., Barrett, C. and Berdegue, J. 2003. The rise of supermarket chains in Africa, Asia and Latin America. *American Journal of Agricultural Economics* 85, 1140–6.

Sahn, D., Dorosh, P. and Younger, S. 1997. *Structural Adjustment Reconsidered*. Cambridge: Cambridge University Press.

Timmer, P.C. 1986. *Getting Prices Right – The Scope and Limits of Agricultural Price Policy*. Ithaca, NY: Cornell University Press.

How to cite this article

Barrett, Christopher B. and Emelly Mutambatsere. "agricultural markets in developing countries." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000209> doi:10.1057/9780230226203.0023

The New Palgrave Dictionary of Economics Online

agricultural research

Robert E. Evenson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews contributions made by agricultural research programmes in historical context. Before 1850, when the agricultural experiment station (AES) model was developed, most crop and livestock improvement was due to farmer selection of seeds and livestock breeding. By 1875, a number of plant breeding programmes were in place. Developed countries achieved a green revolution in the first half of the 20th century, developing countries in the second half. A number of countries are now benefiting from the gene revolution. An assessment of social returns to public spending on agricultural research shows these returns to be high.

Keywords

agricultural research; Ehrlich, P.; genetically modified (GM) crops; green revolution; green revolution modern varieties (GRMVs); Griliches, Z.; internal rate of return (IRR); international agricultural research centers (IARCs); Mendel, G.; national agricultural research system (NARS); patents; precautionary principle; recombinant DNA (rDNA) gene revolution; returns to research

Article

Before 1850

The earliest form of agricultural research was agricultural invention. The patent systems in Europe date back to the Statute of Monopolies in 1623 in England. During the 18th century, England and France further developed their patent systems. Article 1, Section B of the US Constitution, drawn up in 1787, states that ‘Congress shall have the power to promote the progress of science and useful arts, by securing for limited times for authors and inventors the exclusive right to their respective writings and discoveries.’ The first Patent Act in the United States was enacted in 1790. Many of the earliest inventions, including Eli Whitney's cotton gin, were agricultural inventions.

Prior to the development of the modern agricultural experiment station in 1843, the ‘botanic garden’ served as the chief research vehicle for plants. Botanic gardens were established in many countries, preserving and further classifying plants and trees in the tradition of Linnaeus. (Today there are 1,500 botanical gardens worldwide. Of these, 698 have germplasm collections for the conservation of ornamental species, indigenous crop relatives and medicinal and forest species, and 119 conserve germplasm of cultivated species, including landraces – that is, distinct types – and wild food plants.)
Both plant and animal improvement prior to the modern experiment station was achieved by farmers themselves. Prior to the 18th century, farmers selected seed from each crop to improve the productivity of crop species. (There are approximately 300,000 species of higher plants, that is, flowering and cone-bearing plants. Of these, 270,000 have been identified and described. About 30,000 species are edible and about 7,000 have been cultivated or collected by humans for food; 120 species are important cultivated crops, but 90 per cent of the world's caloric intake is provided by only 30 species.)
As populations moved to new locations and production conditions, they created new landraces in each cultivated species. As new landraces were created, three distinct classes were identified. Landraces created in the centre of origin of cultivation were the first class. For rice, as many as three or four centres of origin (that is, locations of first cultivation) for the two cultivated species *Oryza sativa* and *Oryza glaberrima* have been identified. The second class includes landraces created in centers of diffusion (that is, locations where populations diffused the crop). The third class comprised landraces created in the New World countries in the Americas and Oceania.
These landraces were later collected and, along with mutants and uncultivated species in the genus, they constitute the genetic resources used in modern plant breeding programmes based on conventional methods of crossing parental plants. Table 1 summarizes contemporary *ex situ* genebank collections.

Genebank collections (*ex situ*)

Crops	Estimated numbers of landraces (000's)	Major collections (number)	Genebank accessions (000's)	Percent in genebanks
Cereals				
•Wheat	150	36	844	95
•Rice	130	20	420	90
•Maize	65	22	277	90
•Sorghum	45	19	169	80
•Millets	30	18	90	80
Legumes				
•Beans	n/a	15	268	50
•Soybeans	30	23	174	60
•Lentils	n/a	5	26	n/a
•Groundnuts	15	16	81	n/a
Root crops				
•Cassava	n/a	5	28	35
•Potato	30	16	31	95

•Sweet potato	5	7	52	50
Other				
•Sugar cane	20	20	20	70

Source: FAO (1998)

Animal improvement actually pre-dates crop improvement. It, too, was achieved by farmers and herdsmen. Most of the breeds of cattle, pigs, poultry, horses, sheep, and so forth were developed in the 16th through 18th centuries. Most were developed in Europe. Work animals, including oxen, horses and water buffalo, were particularly important in agriculture prior to the 20th century, when tractors became the dominant source of power in many countries. Work animals, including the powerful workhorses, important to cultivation, are sensitive to climatic conditions. Animal breeds used in Asia range from the powerful bullocks in North India, weighing more than a ton, to much smaller cattle in the Himalayan mountains.

1850 – 1900

Agricultural research programmes were changed dramatically with the development of the agricultural experiment station. It is generally accepted that the first truly scientific experiment stations were located in the UK, in the Rothamsted Experiment Station, established in 1843, and in Saxony, where several experiment stations were established in the 1850s.

With the experiment station and its formal structure of experiments with ‘treatments’ and ‘controls’, agricultural research became scientific, and by 1900 agricultural science was established as a mature applied science. The application of statistical methods to experiments furthered this development. R. A. Fisher, the statistician at the Rothamsted Experiment Station in the UK from 1919 to 1933, is credited with numerous methodological developments, many of them relevant to modern-day econometrics.

Early experiments focused on agricultural chemistry, including the application of chemical fertilizers and related soil amendments. By 1875 or so, formal plant breeding programmes were beginning to be established. It is often thought that formal plant breeding did not take place until after the ‘rediscovery’ of Gregor Mendel's work, first published in 1856, in 1900. But that is not the case: breeding programmes in sugar cane, wheat and many other crops were established before 1900. Sugar cane breeders in Java and Barbados simultaneously discovered techniques to induce flowering in sugar cane plants in 1878, and by 1900 the ‘noble’ canes from their breeding programmes were beginning to transform sugar cane production in several countries.

In the United States, the Hatch Act of 1887 provided funds for experiment stations in every state. Most state experiment stations recognized the synergistic relationship between research and graduate teaching, and formally linked experiment stations with land grant college programmes. It is widely thought that legislation such as the Hatch Act reflected exceptional wisdom on the part of legislators. This was not the case. Prior to the Hatch Act, many states had considerable experience with experiment stations. This was also true for the Land Grant College Act – the Morill Act – in 1862. Some 20 states had established colleges of agriculture prior to 1862. As these programmes matured, veterinary medicine colleges were established in land grant colleges. By 1900, sufficient experimental data were available from state agricultural experiment stations to answer many questions of importance to farmers in the US.

1900 – 1940

The period 1900 to 1940 was a one of extraordinary achievements by agricultural experiment stations. Plant breeding gains were achieved in most crops planted in temperate zone countries (in effect, temperate-zone developed countries realized a green revolution in this period). Plant breeding gains in sugar cane, coffee, tea and spices (the Mother Country crops) were also achieved in tropical regions. Brazil and Argentina in Latin America realized major gains (Brazil became the world's major producer of coffee and sugar; Argentina the major exporter of beef).

Two major scientific developments in plant breeding were achieved during this period. The first was the development of techniques to produce hybrid crop varieties to take advantage of the ‘heterosis’ effect in crops. The early development of hybrid techniques took place at Harvard and Yale Universities, but the major achievement was made by Donald Jones at the Connecticut Agricultural Experiment Station in New Haven. Jones developed the ‘double cross’ method for seed production. Hybrid seed production requires ‘selfing’ or ‘inbreeding’ for several generations. Prior to Jones, a single cross was made between two inbred lines to produce hybrid seed; the seed cannot be saved by farmers because the heterosis effect is present only in the hybrid generation. Jones used four inbred lines in a double-cross to produce seed more efficiently. Since Connecticut is not a major corn production state, it was several years before hybrid corn was available to farmers in Iowa. Henry A. Wallace, later a vice-president of the US, was an early leader in developing private industry production of hybrid corn. He established the Pioneer Hybrid Seed Company in 1926.

Zvi Griliches (1957) analysed the adoption of hybrid corn by farmers in different US states. Farmers in Alabama had access to hybrid corn varieties 20 years after farmers in Iowa. This was not because hybrids suited to Iowa farmers were not exhaustively evaluated in Alabama. Alabama farmers did not have hybrid varieties until seed companies established breeding programmes in Alabama to develop varieties suited to Alabama production conditions. Corn has a high degree of photo-period sensitivity. Varieties suited to Alabama were also varieties with longer growing seasons. This same principle applies to the green revolution (see below). No country without a functioning plant breeding programme has realized a green revolution.

The second scientific development was another form of hybridization, inter-specific hybridization or ‘wide crossing’. Until the gene revolution, based on ‘recombinant DNA’ techniques, all plant breeding entailed a ‘sexual’ cross between two ‘parent’ cultivars (this continues to be the case for achieving continuous plant improvement). Inter-specific hybridization entails a sexual cross between different species, usually members of the same genus. This was first achieved in sugar cane in 1919 when breeders achieved crosses between *Saccharum officianaram*, the cultivated species, and *Saccharum spontaneum*, an ornamental species of sugar cane. Later a third species, *Saccharum barberie*, was added. By the 1980s, inter-specific hybridization techniques (chiefly embryo rescue techniques) had been developed for most crop species. With these techniques, sexual crosses have been achieved between cultivated species and most or all uncultivated species in the same genus for all important crop species.

During 1900–40, developed country agriculture (and some developing country agriculture) was also being affected by the development of farm machinery and tractor power. Stationary tractors and steam engines were developed before 1900. After 1900 the row crop tractor was developed along with improved harvesting and planting machinery. By the 1930s these developments were changing the structure (farm size, off-farm work) of US agriculture. These developments were produced largely by

private sector firms in the farm machinery and farm chemical industries. Patent incentives existed for mechanical, electrical and chemical inventions in this period. They were not developed for genetic inventions until after 1980.

1940 – 1965

At the end of the Second World War, agricultural research experienced a renaissance in developed countries. This was at least in part because of synergism between public sector agricultural research and private sector R&D in the farm machinery and farm chemical industries. By 1965 supermarkets had crowded out the ‘mom and pop’ grocery stores in most US cities. Poultry production was effectively industrialized by 1965 as confined housing units became the norm. Dairy production was subject to scale economies, and herd size was increasing. Feed management had improved greatly. The widespread use of United States Department of Agriculture grades and standards for livestock was transforming the meat packing industry. By 1965, in all OECD countries total factor productivity growth was faster in the agricultural sector than in the rest of the economy, and this continues to be the case today.

In developing economies, a sense of alarm had been created by the growing recognition that developing countries were in for a population explosion. With improvements in public health measures, death rates, particularly among children, began to decline and life expectancy began to increase. With even modest delays until the birth rate declined, this meant rapid increases in population. The alarm in question centred on food security. Many alarmists of the 1950s, notably Paul Ehrlich (1968), concluded that food production growth could not keep pace with population growth.

The international community (including the World Bank, regional banks, foundations and bilateral aid organizations) responded by developing a system of international agricultural research centers (IARCs). The first two IARCs were the International Rice Research Institute (IRRI) in the Philippines and the International Wheat and Maize Improvement Center (CIMMYT) in Mexico. These two centres were credited with creating a ‘green revolution’ based on high-yielding varieties of rice and wheat introduced to farmers in 1965. Other IARCs, however, contributed to green revolutions in all major food crops.

The green revolution: 1965 – 2004

The period 1965–2004 was truly extraordinary for agriculture. In 1991 the Soviet Union collapsed, leaving the former Soviet republics in severe recession. This included the agricultural sector. Most, but not all, developing countries experienced a green revolution during this period.

Table 2 summarizes the production of green revolution modern varieties (GRMVs) by five-year period. These data show that the production of GRMVs is increasing over time. Thirty-six per cent of all GRMVs were crossed in an IARC programme. Twenty-two per cent of GRMVs crossed in national agricultural research system (NARS) programmes utilized an IARC-crossed parent or other ancestors. Non-government organizations (NGOs) did not produce GRMVs. None were crossed in developed country programmes and transferred to developing countries. Private sector firms did produce hybrid maize, sorghum and millet varieties (five per cent of GRMVs) but only after improved open-pollinated varieties (OPVs) had been produced by IARC programmes. GRMVs were produced in public sector IARC programs and in NARS programmes in developing countries.

Average annual varietal releases by crop and region, 1965–2000

Crop	1965–70	1971–75	1976–80	1981–85	1986–90	1991–95	1996–00
Wheat	40.8	54.2	58.0	75.6	81.2	79.3	80
Rice	19.2	35.2	43.8	50.8	57.8	54.8	58.5
Maize	13.4	16.6	21.6	43.4	52.7	108.3	71.3
Sorghum	6.9	7.2	9.6	10.6	12.2	17.6	14.3
Millets	0.8	0.4	1.8	5.0	4.8	6.0	9.7
Barley	0.0	0.0	0.0	2.8	8.2	5.6	7.3
Lentils	0.0	0.0	0.0	1.8	1.8	3.9	5.0
Beans	4.0	7.0	12.0	18.5	18.0	43.0	40.0
Cassava	0.0	1.0	2.0	15.8	9.8	13.6	14.0
Potatoes	2.0	10.4	13.0	15.9	18.9	19.6	20.0
All crops							
Latin America	37.8	55.9	65.9	92.5	116.2	177.3	139.2
Asia	27.2	59.6	66.8	86.3	76.7	81.2	79.9
Middle East–North Africa	4.4	8.0	10.2	12.2	28.4	30.5	82.2
Sub-Saharan Africa	17.7	18.0	23.0	43.2	46.2	50.1	55.2
All regions	87.1	132.0	161.8	240.2	265.8	351.7	320.5

Source: Evenson (2003a)

Table 3 summarizes the economic consequences of the green revolution. Production increases are separated into increases from higher crop area planted and increases from higher yields. Yield increases are further separated into GRMV contributions and other input (fertilizer, labour) contributions. In the early green revolution period, production increased by 3.2 per cent a year. Yield increases account for 2.5 per cent a year. In the late green revolution period, production increased by 2.2 per cent per year. Yield increases accounted for 1.8 per cent per year. The sub-Saharan Africa region was an outlier in both periods, with low modern varieties (MV) contributions. The green revolution for sub-Saharan Africa was not accompanied by increased inputs, as it was in Asia and Latin America. (At least 12 countries – Afghanistan, Angola, Burundi, Central African Republic, Congo (Brazzaville), Gambia, Guinea Bissau, Mauritania, Mongolia, Niger, Somalia and Yemen – did not have a Green Revolution. Most are in sub-Saharan Africa.)

Economic consequences of the green revolution (growth rates of food production, area, yield, and yield components, by region and period)

	Early green revolution 1961–80	Late green revolution 1981–2000
Latin America		
Production	3.083	1.631

Area	1.473	–0.512
Yield	1.587	2.154
•MV contributions to yield	0.463	0.772
•Other input/ha	1.124	1.382
Asia		
Production	3.649	2.107
Area	0.513	0.020
Yield	3.120	2.087
•MV contributions to yield	0.682	0.968
•Other input/ha	2.439	1.119
Middle East–North Africa		
Production	2.529	2.121
Area	0.953	0.607
Yield	1.561	1.505
•MV contributions to yield	0.173	0.783
•Other input/ha	1.389	0.722
Sub-Saharan Africa		
Production	1.697	3.189
Area	0.524	2.818
Yield	1.166	0.361
•MV contributions to yield	0.097	0.471
•Other input/ha	1.069	–0.110
All developing countries		
Production	3.200	2.192
Area	0.683	0.386
Yield	2.502	1.805
•MV contributions to yield	0.523	0.857
•Other input/ha	1.979	0.948

Notes: Data on food crop production and area harvested are taken from FAOSTAT (2003) on total cereals, total roots and tubers, and total pulses. Asia: Developing Asia minus the countries of the Near East in Asia.

Africa: Developing Africa minus the countries of the Near East in Africa and the countries of North-West Africa.

Middle East–North Africa: Near East in Africa, Near East in Asia, and North-West Africa.

Latin America: Latin America and the Caribbean.

Crop production is aggregated for each region using area weights from 1981.

Estimates of production increases due to MVs are from Evenson (2003b). Growth rates of other inputs are taken as a residual. Growth rates are compound and are computed by regressing time series data on a constant and trend variable. The totals for All developing countries are derived by weighting the regional figures by 1981 area shares.

Source: Evenson and Gollin (2003)

The recombinant DNA (rDNA) gene revolution

In 1953 Watson and Crick published work (Watson, 1968) that identified the ‘double helix’ structure of DNA and established DNA as the carrier of genetic information. In 1974 Cohen at Stanford and Boyer at the University of California at San Francisco achieved recombinant DNA ‘transformation’ or insertion of ‘alien’ DNA into organisms, and the field of genetic engineering was born (Cohen, 1997).

Within a few years many ‘crop biotech’ companies were established. Large agricultural chemical companies were early entries into the field. Today seven life science firms (Monsanto, DuPont, and Dow in the US, Syngenta, BASF, and Bayer in Europe, and Savia in Mexico) dominate the genetically modified (GM) crop products industry. The first GM products introduced in the late 1980s were commercial failures. But bovine somatotrophin hormone (BsT), a product to stimulate milk production, was successfully introduced in 1993.

In 1995 several companies introduced GM crop products for canola (rapeseed), soybeans, maize and cotton. These products fall into two classes: herbicide tolerance and insect resistance (*Bacillus thuriengensis*, B_T). Herbicide tolerance (soybeans, canola and maize) enables weed control with traditional herbicides. This trait has been highly valued by farmers and rapidly adopted. Most of the world's canola and soybeans now have this trait, as does considerable acreage of maize. Insect resistance is achieved by engineering maize and cotton plants to produce B_T toxins that limit insect damage to the plant. This has a particularly important effect on cotton, where insects cannot readily be controlled by insecticides.

GM crop products enable farmers to reduce production costs. Cost reductions depend on mechanization status and insect pest status. Estimates of cost reduction vary by country, with Western European countries having negligible cost reduction potential (less than one per cent, because they produce little cotton, canola or soybeans). The US has significant cost reduction potential, as do many developing countries. It should be noted, however, that cost reduction gains are ‘static’ in nature (that is, they do not cumulate over time). Dynamic gains can be produced only by the development of generations of modern varieties, as reflected in Table 2 for GRMVs. The gene revolution is not a substitute for the green revolution.

The gene revolution has become strongly politicized in recent years. A clear division has emerged between the original European Union countries and North American countries. The European Union position is that the ‘precautionary principle’ should apply, while the North American position is that, in the absence of scientific evidence to the contrary, farmers should be allowed to adopt GM crops (see FAO, 2004).

Returns to agricultural research

Griliches (1958) was the first economist to measure ‘returns to research’ by computing returns to hybrid corn research. To do this, he created a cost stream and a benefit stream, and applied present value methods to them. (At a five per cent discount rate the present value of benefits was roughly seven times the present value of costs. Some interpreted this as a 700 per cent rate of return. Of course, it was in fact a benefit–cost ratio.) Griliches computed an internal rate of return to hybrid corn research of 43 per cent. Evenson (2001) reviewed more than 300 studies of returns to research in the decades after the Griliches studies. Table 4 reports a summary of internal rates of return reported in these studies. The project evaluation studies utilized methods similar to those used by Griliches. The statistical studies generally regressed measures of total factor productivity on research stock variables. Some studies were focused on specific commodities, others on aggregate research programmes. Several studies made a distinction between pre-invention science and applied science, and several studies were undertaken of the private sector contribution to agriculture.

Returns to agricultural research studies

	No. of IRRs	Distribution of internal rates of return (% distribution)						Median IRR
		0–20	21–40	41–60	61–80	81–100	100+	
Project evaluation methods	121	.25	.31	.14	.18	.06	.07	40
Statistical methods	254	.14	.20	.21	.12	.10	.20	50
Aggregate programmes	126	.16	.27	.29	.10	.09	.09	45
Pre-invention science	12	0.00	.17	.38	.17	.17	.17	60
Private sector R&D	11	.18	.09	.45	.09	.18	0.00	50
By region								
••OECD	146	.15	.35	.21	.10	.07	.11	40
••Asia	120	.08	.18	.21	.15	.11	.26	67
••Latin America	80	.15	.29	.29	.15	.07	.06	47
••Africa	44	.27	.27	.18	.11	.11	.05	37

Source: Evenson (2001)

The studies are characterized by great diversity in internal rates of return (IRRs), ranging from IRRs of zero to very high levels. Median IRRs are high for all categories. This diversity is consistent with the fact that research is a highly uncertain activity.

Finally, Table 5 utilizes data from the green revolution where GRMV adoption rates were available. The method applied was similar to that which Griliches originally used. These data confirm the estimates in Table 4. Very high returns to IARC research are shown. Returns to NARS programmes are lower, especially in sub-Saharan Africa where many countries did not achieve a green revolution.

Green revolution returns to research

Countries	IARCs	NARS
Latin America	39	31
Asia	115	33
West Asia–North Africa	165	22
Sub-Saharan Africa	68	9

Source: Evenson (2003b)

See Also

- agriculture and economic development
- population and agricultural growth
- technology

Bibliography

Cohen, J. 1997. The genomics gamble. *Science* 275, 767–72.

Ehrlich, P. 1968. *The Population Bomb*. New York: Sierra Club/Ballantine Books.

Evenson, R. 2001. Economic impacts of agricultural research and extension. In *Handbook of Agricultural Economics*, vol. 1, ed. B. Gardner and G. Rausser. Amsterdam: Elsevier Science B.V.

Evenson, R. 2003a. Modern variety production: a synthesis. In *Crop Variety Improvement and Its Effect on Productivity: TheImpact of International Research*, ed. R. Evenson and D. Gollin. Wallingford, UK: CABI Publishing.

Evenson, R. 2003b. Production impacts of crop genetic improvement. In *Crop Variety Improvement and Its Effect on Productivity: The Impact of International Research*, ed. R. Evenson and D. Gollin. Wallingford, UK: CABI Publishing.

Evenson, R. and Gollin, D. 2003. Assessing the impact of the Green Revolution, 1960 to 2000. *Science* 300, 758–62.

FAO (Food and Agriculture Organization of the United Nations). 1998. *The State of the World's Plant Genetic Resources for Food and Agriculture*. Rome: FAO.

FAO. 2004. *The State of Food and Agriculture, 2003–2004*. Rome: FAO.

FAOSTAT. 2003. Agricultural Data. FAO Statistical Databases. Online. Available at http://apps.fao.org/

page/collections?subset=agriculture, accessed 15 September 2005.

Griliches, Z. 1957. Hybrid corn: an exploration in the economics of technical change. *Econometrica* 25, 501–22.

Griliches, Z. 1958. Research costs and social returns: hybrid corn and related innovations. *Journal of Political Economy* 66, 419–31.

Watson, J. 1968. *The Double Helix: A Personal Account of the Discovery and Structure of DNA*. New York: Athenium.

How to cite this article

Evenson, Robert E. "agricultural research." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000210> doi:10.1057/9780230226203.0024

The New Palgrave Dictionary of Economics Online

agriculture and economic development

John W. Mellor

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Agriculture plays a vital role in economic development by facilitating the transition from a low-income subsistence to a high-income commercial economy. Agriculture promotes economic transformations by supplying food, foreign exchange, labour, and effective demand to the non-farm sectors, and is the dominant force in poverty reduction. A land constraint makes agricultural growth unusually dependent on technological change, while geographically dispersed production units favour a family-size labour force. These in turn lead to a special role for government in achieving rapid agricultural growth.

Keywords

agriculture and economic development; economic development; economic growth; employment growth; family; Food and Agriculture Organizational (FAO); foreign aid; government role; Green Revolution; International Food Policy Research Institute (IFPRI); non-tradable commodities; poverty reduction; productivity; productivity; protection; research expenditure; rural non-farm sector; rural-urban income disparities; technological change

Article

Economic development is characterized by three transformations: from domination by agriculture to domination by manufacturing and then services; from domination by non-tradable goods and services to a much larger weight of tradable goods and services; and, from a high proportion of poor people living at the edge of basic subsistence to one with few or no such people. If those transformations are to proceed rapidly and efficiently, agriculture must play a vital role. In the course of playing that role, the relative size of agriculture declines drastically while its absolute size increases.

Agriculture has several characteristics that define not only its ability to influence the various transformations but also the means by which it grows and facilitates those transformations. The most important of these are threefold: first, dependence on land and a land constraint that yields rapidly

diminishing returns to increased inputs, making agriculture unusually dependent on technological change for its growth; second, geographically dispersed production units that favour a family-size labour force, with the amount of land and capital per family increasing immensely with rising incomes; and third, derived from the first two, a special role for government in meeting the conditions of rapid agricultural growth. Reinforcing the need for good governance is the increasing need for government-provided institutions for ensuring a healthy, educated labour force as agriculture modernizes.

The size of agriculture

Initially humankind produced the basic means of substance at such low levels of productivity that there was time for little else. Agriculture dominated those subsistence activities. From that initial base, progress could be made only by increased productivity in agriculture, thereby releasing resources for other needs and eventually for luxuries. Even in lower middle-income countries agriculture remains sufficiently large that it continues to play a critical role in transforming the economy. Agriculture's role in employment growth, raising real wage rates and hence reducing poverty is even greater than its role in GDP growth. It continues to be dominant in employment growth at least through upper middle-income status.

Share of GDP

In low-income countries, such as those in most of contemporary Africa, significant parts of Latin America, and, until recently, most of Asia, agriculture accounts for in the order of half of GDP. By the time middle-income status is reached, as it has been in most of contemporary Asia, Latin America and the Middle East, agriculture's relative importance has declined to between 15 and 25 per cent of GDP. With high-income status it declines to under five per cent.

However, as the economy is transformed agriculture can still grow rapidly in absolute size. Indeed, the faster agriculture grows in absolute terms the faster its relative importance declines. This is because high-income elasticity of demand by farmers for non-farm goods and services causes those sectors to grow faster than agriculture and all the more so at high rates of agricultural growth (see Mellor, 1995).

The decline in the relative size of agriculture is further hastened by the appearance of scale economies in many of the production and marketing services for modern agriculture. As development proceeds, many tasks performed on farms in the early stages of development are more economically produced by large-scale firms. Initially farmers produced their own plant nutrients from composting and manure, but it became much cheaper to buy inorganic fertilizers from immense petrochemical plants. Power initially is derived from humans and animals raised on the farm but eventually from tractors and other machines produced off the farm. The examples are endless.

Share of employment and employment growth

Statistics for agriculture's share of employment in low- and middle-income countries are always far larger than those for its share of GDP. That is substantially because of misclassification. Persons with very small holdings that are insufficient in size to provide even half of family employment or income are

normally classified as farmers, but of course they are more properly classified as rural non-farm population given the way they make their living. Thus, typically even in low-income countries the rural population is divided about equally between those who make their living primarily in farming and those in other rural occupations. Seen this way, farmers represent a similar proportion of employment and GDP. This is not surprising since farm income derives substantially from land ownership, not just from labour, just as in the urban sector income derives substantially from return on capital as well as from labour.

Thus, in a low-income country 80–90 per cent of the population may be rural, half with farming as their principal occupation. By the time middle-income status arrives the share of population that is rural has declined to around 40 per cent and the share principally occupied in farming to less than 20 per cent. In high-income countries the farm population is less than five per cent, two-thirds of those producing the bulk of the farm output.

Agriculture and economic growth

Because of its initially dominant size, agriculture makes several large initial contributions to overall growth (Mellor, 1976.) Growth in agricultural productivity releases labour for the fast-growth non-farm sectors. Agriculture earns foreign exchange that is utilized to import capital goods for the non-farm sector. It provides low-cost food to keep labour costs down as employment in the non-farm sector grows rapidly. Rapid growth in non-farm employment faces rapidly rising, competitiveness-destroying increases in real wage rates if agricultural production does not grow rapidly. Even in an open economy with rapid growth in urban incomes, increased food imports would be so great with a failing agriculture that the real exchange rate would change sharply and push up the cost of food and therefore of labour. Finally, fast-growth agriculture plays the dominant role in employment growth and poverty reduction. In the context of modern open economies and free capital flows, the latter contribution remains the most important for agriculture.

Agriculture and poverty reduction

Statistical data from diverse cross-sectional analyses show that in low- and middle-income countries it is agricultural growth that drives poverty reduction (Ravallion and Datt, 1996.) Further, there is a significant lag in that poverty-reducing impact. The lack of immediate impact led to an incorrect view that agricultural growth does not reduce poverty. It is now known that the lag is due to the large indirect impact of agricultural growth on poverty reduction. There is, however, a major exception to this relation. When land ownership is highly skewed, as for example in much of Latin America, agricultural growth does not significantly reduce poverty. That is because very rich people with large landholdings spend additional income not on employment-intensive rural non-farm goods and services but on capital and import-intensive urban goods and services.

When agricultural incomes are broadly distributed, agricultural growth reduces urban poverty more than does urban growth. This is because urban poverty is a product of rural-urban wage disparities. If rural incomes are stagnant, the rural-urban disparity increases and poor rural people migrate to the cities. If the disparity is large, rural people will be willing to wait a long time in the urban area for a job, living in

poverty in slums. The return to waiting is made up once they get the good urban job. Thus, the greater the income disparity, the longer the queue and hence the greater the number in urban poverty. Measures that make waiting cheaper, such as subsidized housing or even normal urban amenities such as potable water, simply increase the rural-urban disparity and hence the queue. Thus, the way to reduce urban poverty is to raise rural incomes and amenities as rapidly as those in urban areas.

There are three means by which agricultural growth contributes to reduced poverty: lower food prices; increased agricultural employment; and farm income-driven rural non-farm employment.

Food prices

Poor people in low-income countries spend in the order of 80 per cent of their income on food. It follows that the real price of food is a primary determinant of the real income of the poor. In a neoclassical economy, increased domestic food production does not reduce the price of food because the international price rules. However, high transfer costs in low-income countries somewhat insulate domestic food prices from international prices. This may be reinforced by trade restrictions. In that case, increasing food production faster than domestic demand will reduce domestic food prices and greatly benefit poor people. The high-yielding rice varieties that brought the Green Revolution to Asia were of low quality, depressing the price of rice consumed by the poor. Market forces may depress the nominal wage as food prices decline, but those same market forces will then increase employment. Hence, the poor tend to benefit from rapid growth in agriculture either through lower food prices or through increased employment (Mellor, 1976.)

Of course, these same processes work in reverse. If agricultural production grows more slowly than domestic demand, food prices tend to rise, reducing the real incomes of the poor. Unfavorable weather reduces agricultural production; prices rise and the poor suffer. Wage rates rarely adjust in the short run, although they do in the long run, in which case higher wage rates reduce employment. In either case the poor lose.

Of course, increasing the supply of food faster than demand is difficult in low-income countries in which population growth is rapid and in which incomes may also be rising. The income elasticity of demand for food is much less inelastic in low-income countries than in high-income countries, and hence income growth has a major effect on the demand for food. For example, the Food and Agriculture Organizational of the United Nations (FAO) and the International Food Policy Research Institute (IFPRI) both show for Africa continued shortfall in supply into the indefinite future. The African poor will continue to suffer from such trends (Eicher and Staatz, 1998.)

Increased farm employment

Because agriculture is initially so large, rapid growth does add directly and substantially to employment. However, direct employment growth is small compared with the growth in output. This is because productivity-increasing technological change is the primary source of high growth rates in agriculture. Even though the technology is generally designed to be land-saving, it also increases labour productivity. Thus, for each ten per cent increase in agricultural output employment increases by between less than three per cent and at most six per cent. Thus, the big impact of agricultural growth on employment comes indirectly through the rural non-farm sector.

Increased rural non-farm employment – driven by rising farm incomes

In an open economy, agricultural output that grows faster than demand does not depress prices significantly because of access to international markets. A small decrease in prices brings increased exports. A high growth rate in output without depression of prices raises farm income and reduces poverty in a quite different manner from that of reduced prices.

Farmers spend a large and increasing proportion of increments to their income on the goods and services produced by local, rural non-farm workers. Numerous studies show that the bulk of the poor are rural non-farm workers. They largely produce non-tradable goods and services. Because of low quality and high transaction costs they cannot export as an alternative to meeting local demand.

When farmers prosper they enlarge their homes and buy local furniture, local tailoring, and a vast panoply of services. That increases employment and eventually real wages in the rural non-farm sector. This is the source of poverty reduction in a low- or medium-income open economy.

Because of the strong multiplier on those expenditures, there is a significant lag in the full effect of agricultural growth on poverty reduction as successive rounds of expenditure occur. Similarly, rich, and especially absentee, landowners spend incremental income largely on capital and import-intensive commodities and services and so have little effect on poverty reduction. These two relations are consistent with the data cited earlier.

In very poor agricultures that are growing little or not at all, those in the rural non-farm sector are exceedingly poor because of lack of local demand. In that situation outmigration of the principal male worker and sending back of remittances are important factors holding poverty in check. This is, of course, a socially disruptive means of holding off poverty. Thus it is not surprising that when farm incomes rise rapidly migration beyond commuting range is sharply reduced.

Rural-urban income disparities

It is not uncommon for the urban sector to grow rapidly in low-income countries, even while agriculture stagnates. Foreign aid may be spent largely in the cities, as in Africa, or macro policy stimulates manufacturing growth while the complex processes of agricultural growth are neglected. In that case, urban and rural poverty both surge. At that stage of development it is critical that agricultural production grows rapidly in order to prevent rapid widening of rural urban disparities.

As countries move to middle-income status, the problem of rural-urban disparities changes. The rate of growth of urban incomes accelerates – to around six per cent per year. The capacity to absorb migration also increases as the urban proportion of the population increases. Concurrently, the potential for accelerating the agricultural growth rate improves. The demand for high-value agricultural commodities, such as livestock products and fruits and vegetables, grows at a rate of between six and eight per cent, much of which can be efficiently met from domestic production. Thus, in middle-income countries the agricultural growth rate may pick up to between four and six per cent. That would allow rural incomes to roughly keep pace with urban incomes. While not uncommon amongst middle-income countries, such growth rates are by no means universal and require carefully selected government actions.

Characteristics of agriculture that determine the means of growth

Agriculture has very different characteristics from urban industry and therefore different requirements for growth (Eicher and Staatz, 1998.) If those divergent characteristics are not recognized then not only does agriculture grow slowly but poverty reduction halts and-income disparities between rural and urban areas widen. A family-size labour force, the importance of technological change and rural infrastructure, and the consequent importance of government are the dominant characteristics that distinguish the process of agricultural growth from that of other sectors.

The most obvious characteristic of agriculture is that each farm is spread over a wide area. This disperses the workforce and, combined with the complex biological nature of the production process, puts a premium on family-size operating units (commonly including one hired worker) with minimal supervision costs. Size of farm measured by land area or capital investment varies immensely among countries; but the labour force per farm is a virtual constant.

Particularly in low- and middle-income countries in which both land and capital holdings are small as well, the small-scale unit requires support from activities with scale economies. Most of these activities are most efficiently pursued by the private sector. But some are public goods and require public sector activity.

The balance between public and private sectors gradually shifts towards the private sector as development occurs and the private sector cultivates a broadened set of skills. Particularly in low-income countries, such a substantial burden falls on the public sector, in research, extension, enforcement of grades and standards (especially for export), and some aspects of finance and of market information systems, that the government must set difficult priorities. In that context it must continually press to turn activities over to, and encourage, the private sector as that sector's capacity increases.

The key role of government in agricultural growth, in turn, makes the role of the agriculture ministry important as it diagnoses needs and facilitates and complements the private sector. Particularly in early stages of accelerated agricultural growth, the agriculture ministry must have an explicit strategy with clear priorities and sequences in which to take up key activities. When the fashion in development swings towards minimizing the role of government, agriculture is more likely to suffer than other sectors.

Key forces in agricultural growth

Much of what is required for rapid agricultural growth is most appropriately and efficiently undertaken in the private sector, but even the minimum set of required public sector activities is long and complex. Government can do only a few major things at a time. Thus, one of the most important elements of a high growth rate is an at least implicit strategy within which a small number of limiting priorities will be set with an efficient sequence of activities guiding the moving on to new priorities as earlier ones are fulfilled and institutionalized.

The immediate priorities differ from country to country depending on the physical circumstances and the history of interventions. Hence, setting priorities and sequences and even the broad strategy are highly country-specific. A few generalizations are possible. Physical infrastructure and technology institutions are critical in all growth plans, and government is essential to the provision of both. They are also both never-ending tasks, requiring constant improvement, and thus are always a priority. The other constant is the growing importance of the private sector to agricultural growth and the increasing importance of

public sector facilitation of that growth. For agriculture to growth rapidly, good governance is critical – technically competent and committed to agricultural growth and rural development.

Technological change

Basic science-based, institutionalized research is essential to thwart the diminishing returns incident to a limited land area, and in any case provides a high rate of return. The varied biological and physical environment of agriculture limits the transfer of technology and thus requires area-specific research systems. Because research results are often public goods, public sector research is critical to agricultural advance. As the private sector expands it will increasingly take on research activities. But even in high-income countries public sector research is a major component of private-public sector partnerships.
As farming becomes more complex and dynamic, the educational requirement of farmers increases. Concurrently, many farm children will leave agriculture for education, demanding urban jobs. Thus technology-based agricultural growth creates a strong demand-pull for increased rural education.
Because research is so important, and because it is becoming increasingly expensive, depending on expensive equipment and large coordinated teams, low-income countries must set difficult, narrow priorities for their research activities. That is one of the most important and difficult priority-setting exercises in economic development. Typically it is not done well and so research expenditure is not efficient and agricultural growth does not reach its full potential. In parallel with research are systems for the dissemination of research results. These too start heavily in the public sector and then move to a complementary mix with the private sector.

Physical infrastructure

Agriculture's contribution to overall economic development is dependent on a steady flow of technology that requires increased inputs and produces increased output. For those processes to proceed rapidly, transaction costs must be reduced. This requires constantly upgraded roads, electrification, and telecommunications. While such physical infrastructure is naturally provided to urban areas, the dispersal of agriculture increases infrastructure costs in rural areas and makes it necessary to sequence provision geographically.
Rapid agricultural growth requires educated people in villages to provide agricultural extension, financial institutions, and modern marketing systems. Schools and clinics are of no use without trained staff. These educated people will not live in places without the full set of physical infrastructure. Thus, there is synergy between the requirements of agriculture and the social services for physical infrastructure.

Private sector input supply and output marketing

Rising agricultural productivity depends on massive increases in purchased input supplies as the cost of those inputs decreases and the cost of on-farm sources increases. This in turn requires rural financial markets that can mobilize national and international savings for innovating farmers and provide an outlet for farmers’ savings when they reap the income benefits of improved technology.
Rising incomes and technological advance in marketing require increased quality of farm output,

especially for high-value perishable commodities, and large volumes. Thus, the size and complexity of agricultural marketing increase rapidly. While family labour force-size farms preserve their competitive position in production they are at an increasing disadvantage in meeting quality and volume requirements of modern marketing systems. This challenge is best met by organizing farmers into large units for marketing purposes. This may occur through contract farming provided by large agricultural business firms, or cooperatives, or farmers' organizations. For the latter, government may play an important role in facilitating farmer organization, but must be careful not to stifle efficiency by making them in effect government institutions.

In setting their own priorities, governments must seek the means to assist the private sector in providing the input and output supply activities, and be careful not to stifle private development with onerous regulation, even while protecting consumer interests and helping to build a favorable reputation for exports.

Change over time in pace and composition of agricultural growth

The sources of agricultural growth change greatly over time. Yield rapidly increases in importance compared with land area. This is because of the combined effect of loss of the land frontier with population growth and exploitation and rapid increase in the efficiency of producing improved technology. The input composition switches to purchased inputs such as fertilizer and chemical pest control and off-farm marketing and processing. This rapidly increases productivity of labour and raises income.

The output composition commences with domination by cereals and root crops as the low-cost sources of calories. As incomes rise the demand for income-elastic livestock and horticultural products grows very rapidly. These are labour-intensive commodities for which physical conditions in low- and middle-income countries are usually suitable. These commodities are little restricted by land area since a modest shift of area from extensive crops allows a large increase in their production, and so the overall growth rate accelerates. An agriculture dominated by cereals is unlikely to exceed a three per cent growth rate for more than a few years. But when livestock and horticulture come to occupy over half the agricultural GDP, as happens in middle- and high-income countries, the growth rate can accelerate to between four and six per cent.

The importance of trade to agricultural growth

In low-income countries demand for agricultural products grows slowly. Consumption is largely of cereals, incomes grow slowly and demand is inelastic. At that early stage of development, agriculture has considerably greater capacity to grow than domestic markets can absorb, and achieving that growth is vital to poverty reduction and also to overall GDP growth rates. Thus, what Hla Myint (1958) referred to, as 'vent for surplus' is important to agriculture playing its role. That is to say, agricultural production must grow faster then domestic demand and the surplus exported. This drives the domestic employment multipliers as well as paying for imported capital equipment critical to overall growth.

For agricultural exports to grow a country must produce efficiently, providing constantly improving physical infrastructure to bring down transaction costs, and constantly increasing productivity through

technological change and an effective private sector capable of adapting to rapidly changing markets and constantly rising quality standards. However, these favourable policies can be nullified by unfavourable macro policy, particularly including overvalued exchange rates. Those are the most important requisites of export success. Globalization, based on declining costs of transport, facilitates access to markets, but also brings competition. Countries lagging in provision of physical infrastructure, technological change, and efficient macro policy will be losers from globalization.

Trade protection by high-income countries has been an important barrier to export success even when low- and middle-income countries become efficient and productive. Protection is particularly onerous for cotton, widely grown in quite poor countries and heavily protected and subject to export subsidies from high-income countries. Protection may also be subtle, using health rules to make it difficult for poor countries to enter high-income markets. Thus, the rate of growth of agriculture is dependent in part on negotiations to reduce both trade barriers erected by high-income countries against high-value agricultural commodities and agricultural subsidies more generally.

Foreign aid, agriculture and development

Successful late starters in economic development exceed the growth rate of the front-runners because they can catch up by drawing capital and, more important, technology and the pure science base for creating technology from their now wealthier predecessors. Foreign aid can play an important role in those transfers. This has been dramatically the case in agriculture. In Asia, the scientific base for the startling technological breakthroughs of the Green Revolution was laid by foreign aid that sponsored the key research institutions, in Mexico, then the Philippines, and finally in many other countries. These efforts were complemented by assistance to development of a host of national institutions vital to the spread of the Green Revolution and to increasing the effectiveness of agriculture ministries.

A variety of factors, including the rise of specialized lobbies that distort the distribution of foreign aid between directly productive and social activities and away from national institutions to local institutions and, most important, from national institution building to local activities, caused foreign aid to lose its effectiveness.

The late starters, particularly in Africa, were the big losers from this shift. For the late starters to achieve faster growth than their immediate predecessors will require a return to basics. A great deal has been learned about the details of agricultural growth and its contribution to overall economic development. That new information can accelerate growth beyond previous levels. But the basic principles have not changed and there must be a reversion to these if the new knowledge is to be useful. Africa and a few low-income countries in Asia and Latin await that renaissance.

See Also

- agricultural finance
- agricultural markets in developing countries
- agricultural research
- foreign aid
- family economics

- growth and international trade

Bibliography

Eicher, C. and Staatz, J. 1998. *International Agricultural Development*, 3rd edn. Baltimore, MD: Johns Hopkins University Press.

Mellor, J. 1976. *The New Economics of Growth*. Ithaca, NY: Cornell University Press.

Mellor, J. 1995. *Agriculture on the Road to Industrialization*. Baltimore, MD: Johns Hopkins University Press.

Myint, H. 1958. The ‘classical theory’ of international trade and the underdeveloped countries. *Economic Journal* 68, 317–37.

Ravallion, M. and Datt, G. 1996. How important to India's poor is the sectoral composition of economic growth? *World Bank Economic Review* 10(1), 1–25.

How to cite this article

Mellor, John W. "agriculture and economic development." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000068> doi:10.1057/9780230226203.0025

The New Palgrave Dictionary of Economics Online

airline industry

Severin Borenstein and Nancy Rose

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The 1978 US airline deregulation benefited passengers through lower fares and expanded service. Airline privatization and liberalization elsewhere in the developed world has since had similar effects. Still, there have been some unanticipated effects: hub-and-spoke networks have efficiency appeal, but they also increase congestion and confer market power on dominant airlines; price discrimination is widespread; loyalty programmes exacerbate market power concerns; airline finances are subject to extreme cyclic volatility; and labour is a significant residual claimant on profits. Airline competition and industry structure remain in flux: entry and exit are commonplace, as is experimentation with new pricing and products.

Keywords

airline deregulation; airline industry; market power; price discrimination; principal–agent conflict

Article

Since the mid-1970s, privatization and deregulation have transformed domestic passenger airline markets in many developed economies.

From its infancy through the early 1970s, scheduled passenger air service was considered a public utility nearly everywhere in the world. In most countries, this took the form of state-owned national airlines, often operating with significant government subsidies. US airlines were privately owned, but prices and entry decisions were controlled by federal regulators. California and Texas provided limited but notable exceptions, where small airlines providing only intra-state service operated free of most economic regulation. Their substantially lower fares and higher load factors relative to regulated operations foreshadowed the possible impact of deregulation.

The United States legislated federal airline deregulation in 1978, replacing government decision-making with carrier determination of pricing, entry, and network configuration. Within 20 years, similar reforms faced newly privatized and entrant carriers operating within Europe, Asia and Australia. Most

international air travel, however, remains heavily regulated through bilateral government agreements, apart from intra-European Union flights and a few examples of ‘open skies’ pacts that allow broad freedom in entry and pricing,

Deregulation yielded numerous benefits, best documented for the US domestic market due to public availability of detailed, high-quality data. The most striking and robust finding is that fares are substantially lower and passengers are better off under deregulation than they would have been under continued regulation (in the United States) or state ownership (in many other countries); see, for example, Borenstein (1992), Morrison and Winston (1995) and Borenstein and Rose (2006). Facilitating lower prices were decreased costs per available seat–mile and increased load factors, resulting from a mix of operational reorganization, service changes, and efficiency gains. In the United States deregulation-induced transfers from labour to consumers were initially modest, though labour costs and contract negotiations have since become focal in competition between formerly regulated ‘legacy’ carriers and discount airline entrants in many markets. Labour transfers generally account for a more substantial share of cost reductions for newly privatized carriers.

While price declines conformed to expectations, not all responses to deregulation were anticipated. First, legacy airlines in the United States rapidly reconfigured their operations from point-to-point to hub-and-spoke networks, in which coordinated ‘banks’ of flights arrive at a centrally located airport, allow passengers to change planes, and depart a short time later. This allows airlines to offer relatively frequent, albeit connecting, service on a large number of city pairs without dedicating aircraft to serving each route non-stop. Legacy carriers outside the United States generally operated some form of hub-based network even prior to reform, due largely to relatively thin domestic markets and bilateral agreements that restricted international service to operate through a few gateway airports. Hub-and-spoke operations initially were thought to confer significant efficiency improvements, facilitating greater flight frequency and higher load factors for all but the most dense markets, though it was recognized that passengers preferred non-stop service, all else equal.

Over time, the benefits of hubs have been called into question. Coordinated banks of flights increase congestion costs and delays at hub airports and reduce system-wide aircraft utilization rates; airline dominance of local traffic in and out of their hubs raises concerns about market power; many hubs have been created, then abandoned, as airlines attempted to discern the optimal number and characteristics of hub airports.

Second, average real price declines masked an explosion in pricing complexity. From a pair of distance-based coach and first-class fares on each route, airlines sprouted a dozen or more fare offerings. Prices on a single carrier-route may differ by the time or day of travel, how far in advance a ticket is purchased, the length of stay, and whether the stay includes a Saturday night. Economists have debated the extent to which fare variation reflects efficient competitive peak-load pricing or potentially less efficient price discrimination, but both effects are undoubtedly significant in most markets.

Third, market power concerns, focal at hub airports generally dominated by a single carrier, have been exacerbated by the diffusion of various loyalty programmes. Best-known are frequent flyer programmes, which reward passengers for concentrating their business with a single carrier, but similar programmes were also created for travel agents, who booked about 85 per cent of all tickets in the early deregulation days. Nonlinear reward schemes benefit the largest carrier in a market and increase switching costs among their participants. These programmes also generate principal–agent conflicts: travel agents benefit from directing passengers to flights that may be slightly more expensive or less desirable in

exchange for side payments from the carrier. Similarly, in exchange for free personal travel, business passengers choose flights for which their employer may have to pay more.

Fourth, extreme cyclic volatility of airline finances has raised concerns about the ‘core’ of the competitive equilibrium. The industry reaped large profits when demand was strong relative to capacity and fuel prices were low (the late 1980s and late 1990s) and reported huge losses when fuel prices rose and demand weakened, generating excess aircraft capacity and a wave of bankruptcies (the early 1980s, 1990s and 2000s). Debate continues over whether this profit volatility should spark concern or is part of the normal functioning of an industry with high fixed costs, slow capacity adjustment, fluctuating operating costs (particularly fuel), and highly cyclical and unpredictable demand. Is this any different from the steel, computer memory chip, or software industries which also have exhibited extreme swings? Economic research has provided few answers as yet.

Finally, airline labour has been at the heart of continuing concern and stress. At most legacy carriers, pilots and mechanics have negotiated very lucrative contracts during good times, effectively sharing in the high profits. When profits declined, however, downward adjustment of wages has been slow. Entry or expansion by new airlines with substantially lower labour pay scales is fairly easy, particularly during downturns when excess capacity makes aircraft leases cheap and easily available. During downturns, wages at established carriers may differ most from competitive wages, leaving incumbents vulnerable to new competition and financially constrained in their ability to respond aggressively. The rise of low-cost carriers and intensity of legacy carrier wage and benefit cuts in the most recent industry downturn raise significant questions for the future position of airline employees.

Many of the research results from early post-deregulation studies have been reopened in the face of dramatic industry evolution over recent years. The challenge to both economists and industry participants is to infer the long-run equilibrium structure of the industry. What is the stable number of airlines in a given geographic market? What sort of competition is feasible? Are hub networks viable in the face of point-to-point competition? What is the long-run role of labour as a quasi-equity holder? These questions remain for future researchers to address.

See Also

- agency problems
- bankruptcy, economics of
- network goods (empirical studies)
- network goods (theory)
- price discrimination (empirical studies)
- price discrimination (theory)

Bibliography

Borenstein, S. 1992. The evolution of US airline competition. *Journal of Economic Perspectives* 6(2), 45–73.

Borenstein, S. and Rose, N. 2006. Airline deregulation and liberalization: lessons learned. Working Paper. Cambridge, MA: NBER.

Morrison, S. and Winston, C. 1995. *The Evolution of the Airline Industry*. Washington, DC: Brookings Institution.

How to cite this article

Borenstein, Severin and Nancy Rose. "airline industry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000226> doi:10.1057/9780230226203.0026

The New Palgrave Dictionary of Economics Online

Aiyagari, S. Rao (1952 – 1997)

Zvi Eckstein and Dan Peled

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

generational altruism; borrowing constraints; consumption inequality; dynamic macroeconomics; dynastic models; income inequality; idiosyncratic income shocks; market frictions; mobility; numerical solution techniques; optimal taxation; overlapping generations models; precautionary saving; risk aversion; taxation of capital; taxation of capital income; time preference; uninsurable idiosyncratic risks

Article

S. Rao Aiyagari was 45 years old when he died in 1997, just as his approach to dynamic macroeconomic research was gaining recognition. Rao's vision was motivated by empirical observations and academic debates stemming from the different implications of aggregate and individual economic data. In particular, individual earnings, saving, wealth and labour exhibit much larger fluctuations over time than per-capita averages, and accordingly significant individual mobility is hidden within these cross-sectional distributions. Rao became convinced that this kind of heterogeneity and individual dynamics has important implications for the understanding of aggregate economic data and can provide new insights on the role of various economic policies.

The Aiyagari–Bewley economic model, proposed by Bewley (1986) and developed further in Aiyagari (1994; 1995), has become a leading model for modern dynamic macroeconomics. The economy is populated with heterogeneous infinitely lived agents subject to uninsurable idiosyncratic income risks. Possible long sequences of adverse income shocks naturally lead to borrowing constraints on individuals, and consequently fluctuations in consumption can be mitigated only by precautionary individual savings. Since agents' histories of income shocks are different, the model generates equilibrium cross-section distributions of wealth, saving and consumption, which reflect the fact that borrowing constraints are tighter for wealth-poor agents. These cross-sectional distributions are contrasted with or calibrated to fit their empirical counterparts in the data, and their responses to various policy changes can be analysed. Solving for the equilibrium in dynamic models with heterogeneous agents is complicated, and Rao was among the pioneers in developing and applying numerical solution techniques for that purpose.

In his most influential paper (Aiyagari, 1994), Rao investigated the implications of precautionary saving due to individual earning risks and borrowing constraints for aggregate savings. He found that the contribution of uninsured idiosyncratic risks to aggregate saving is modest for plausible values of risk aversion, variability and persistence of earnings (at most three per cent), but can be significantly larger with higher variability and persistence parameters of the earning stochastic process. Access to asset markets in that model enables agents to cut consumption volatility by half, and enjoy a welfare gain of 14 per cent of per-capita consumption compared with the equilibrium with no access to assets markets. The model generates a wealth distribution that is positively skewed, more dispersed than income distribution, and inequality is significantly higher for wealth than for income.

Precautionary savings generated by uninsured idiosyncratic shocks and borrowing constraints motivated Rao to examine the recommendation to eliminate tax on capital income (Lucas, 1990). Aiyagari (1995) showed that for the Aiyagari–Bewley economies this dictum may be wrong because the frictions in these models result in agents' behaviour that is closer to that in overlapping generations (OLG) models. Precautionary saving can lead to over-accumulation of capital in equilibrium, so that positive taxes on capital are needed to bring the pre-tax return on capital to equality with the rate of time preferences, at any point in time as well as in the long run. In contrast to OLG models, where government debt can also be used to reduce excessive saving, in Aiyagari–Bewley economies the demand for such assets becomes infinite when the interest rates approaches the rate of time preferences. The suitability of the model for addressing such fundamental issues is evidenced by the fact that a decade later it was still being used to study the same issue, albeit with different conclusions (Werning, 2005).

Rao has examined many other implications of cross-sectional distributions generated by frictions in capital markets and uninsurable idiosyncratic risks, such as asset pricing and trading patterns (Aiyagari and Gertler, 1991), setting taxes in a median-voter context (Aiyagari and Peled, 1995), marriage patterns and investment in children (Aiyagari, Greenwood and Guner, 2000; Aiyagari, Greenwood and Ananth, 2002). He also studied the equilibrium implications of market frictions and borrowing constraints that emerge endogenously from private information on individual earnings (Aiyagari and Williamson, 2000). Many other influential papers have adopted his framework of uninsurable idiosyncratic risks for the study of various phenomena, including, for instance, Kocherlakota (2005) on optimal taxation, Krueger and Perri (2006) on the joint evolution of income and consumption, and Storesletten, Telmer and Yaron (2004) on age-dependent income and consumption inequality.

Rao's earlier theoretical work focused on the links between dynastic and OLG models, and provided the deep theoretical understanding of dynamic models that he applied in his subsequent work. He examined whether the two models become similar in terms of equilibrium existence, optimality and cyclicality, with and without money, when the life of each generation and the period of overlap across generations are sufficiently long, or when generations are linked through altruism (for example, 1985; 1988; 1989). Additional work with Wallace and others examined the role for policy in search equilibrium models of money (for example, Aiyagari, Wallace and Wright, 1996; Aiyagari and Wallace, 1997).

Aiyagari published more than 30 influential papers during his 18-year career as an economist. The force of his work and ideas and their impact on his colleagues are evidenced by the continued appearance of his co-authored papers for many years after his unexpected death, exhibiting some of the most innovative dynamic macroeconomic research.

See Also

- income taxation and optimal policies
- incomplete markets

Selected works

1985. Observational equivalence of the overlapping generations and the discounted dynamic programming frameworks for one-sector growth. *Journal of Economic Theory* 35, 202–21.

1988. Nonmonetary steady states in stationary overlapping generations models with long lived agents and discounting: multiplicity, optimality, and consumption smoothing. *Journal of Economic Theory* 45, 102–27.

1989. Can there be short-period deterministic cycles when people are long lived? *Quarterly Journal of Economics* 104, 163–85.

1991. (With M. Gertler.) Asset returns with transaction costs and uninsured individual risks. *Journal of Monetary Economics* 27, 311–31.

1994. Uninsured idiosyncratic risks and aggregate saving. *Quarterly Journal of Economics* 109, 659–84.

1995. Optimal capital taxation with incomplete markets, borrowing constraints, and constant discounting. *Journal of Political Economy* 103, 1158–75.

1995. (With D. Peled.) Social insurance and taxation under sequential majority voting and utilitarian regimes. *Journal of Economic Dynamics and Control* 19, 1511–28.

1996. (With N. Wallace and R. Wright.) Coexistence of money and interest-bearing securities. *Journal of Monetary Economics* 37, 397–419.

1997. (With N. Wallace.) Government transaction policy, the medium of exchange, and welfare. *Journal of Economic Theory* 74, 1–18.

2000. (With J. Greenwood and N. Guner.) On the state of the union. *Journal of Political Economy* 108, 213–44.

2000. (With S. Williamson.) Money and dynamic credit arrangements with private information. *Journal of Economic Theory* 91, 248–79.

2002. (With J. Greenwood and S. Ananth.) Efficient investment in children. *Journal of Economic*

Theory 102, 290–321.

Bibliography

Bewley, T. 1986. Stationary monetary equilibrium with a continuum of independently fluctuating consumers. In *Contributions to Mathematical Economics in Honor of Gérard Debreu*, ed. W. Hildenbrand and A. Mas-Colell. Amsterdam: North-Holland.

Kocherlakota, N. 2005. Zero expected wealth taxes: a Mirrlees approach to dynamic optimal taxation. *Econometrica* 73, 1587–621.

Krueger, D. and Perri, F. 2006. Does Income Inequality Lead to Consumption Inequality? Evidence and Theory. *Review of Economic Studies* 73, 163–93.

Lucas, R., Jr. 1990. Supply-side economics: an analytical review. *Oxford Economic Papers* 42, 293–316.

Storesletten, K., Telmer, C. and Yaron, A. 2004. Consumption and risk sharing over the life cycle. *Journal of Monetary Economics* 51, 609–33.

Werning, I. 2005. Tax smoothing with redistribution. Staff Report No. 365, Federal Reserve Bank of Minneapolis.

How to cite this article

Eckstein, Zvi and Dan Peled. "Aiyagari, S. Rao (1952–1997)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000245>
doi:10.1057/9780230226203.0027

The New Palgrave Dictionary of Economics Online

Akerlof, George Arthur (born 1940)

Brian G.M. Main

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

George Akerlof is forever associated with his landmark 1970 paper, ‘The market for “lemons”’, which transformed the way economists approach markets where there is a difference between the transacting agents in the information they possess. This concept of asymmetric information, with its major impact on many fields of economics, was singled out when, in 2001, he was awarded the Nobel Memorial Prize in Economics (along with Michael Spence and Joseph Stiglitz). A more comprehensive assessment of his contribution to economics would be as providing a better behavioural underpinning for macroeconomics as a major figure in the New Keynesian movement.

Keywords

Akerlof, G.; asymmetric information; caste system; efficiency wage theory; Friedman, M.; neuroeconomics; New Keynesian Economics; inflation-unemployment trade-off; market for lemons; social norms

Article

George Akerlof's father came to the United States from Sweden to obtain a Ph.D. at the University of Pennsylvania, and remained in the country to pursue a career as a research chemist. He met George's mother while she was a graduate student in chemistry. Hers was an academic family. George's great grandfather was among the earliest graduates from the University of California at Berkeley (in 1873), and his grandfather also graduated from Berkeley. Other members on that side of the family also established successful academic careers. George grew up on the East Coast, where his father held a series of posts, variously at Yale University, at the Mellon Institute in Pittsburgh and at Princeton University, before running his own independent research firm in the Princeton area. Indeed, it was witnessing the uncertainty surrounding his father's continuing employment, dependent as it was on securing government research grants, which first turned George Akerlof's mind to macroeconomic

themes such as unemployment. As an undergraduate at Yale he majored in mathematics and economics, and in the fall of 1962 he entered graduate school at MIT, where he had the good fortune to find himself one of an exceptionally talented cohort of students. His doctoral supervisor was Robert Solow (Nobel Laureate 1987). Akerlof joined the Berkeley faculty in the fall of 1966 and, although he has spent extended periods away from Berkeley – at the Indian Statistical Institute in New Delhi, the Council of Economic Advisors, the Federal Reserve Board (where he met his wife, Janet Yellen), the LSE, and the Brookings Institution – he has remained closely identified with Berkeley ever since.

The ‘Market for “lemons”’ paper

For the generations of economics students trained since 1970, when asked to single out a favorite economics article, it is a pretty safe bet that the most popular article would be George Akerlof's (1970) paper on asymmetric information, ‘The market for “lemons”’. Part of this paper's appeal lies in its modelling approach. While mathematically rigorous, it is derived from close observation of the world. Care is taken to incorporate realistic economic detail, yet the results obtained provide tremendously powerful insights. The reader is left with an understanding of an important market situation that was previously obscure and, in addition, is offered policy options whereby economic well-being can be improved. This general approach characterizes all of Akerlof's work.

The ‘lemons’ paper starts by offering an analysis of the second-hand car market in which the existence of lower-quality vehicles (the eponymous ‘lemons’) can disrupt the workings of the market – to the extent that the usual economic law of lowering the price in the face of an excess of supply (or difficulty experienced in selling into the market) simply makes matters worse. Rather than bringing about a market equilibrium through matching supply and demand, the lower price drives out the better-quality cars remaining in the market and this further depresses demand.

The problem arises from an asymmetry of information that exists between those supplying used cars into the market (they know, in considerable detail, just how good or otherwise their present car is) and those who are buying in the market (they can obviously inspect the car, but are left with substantially less knowledge than the seller). If those on the demand side use the price as an indication of the average quality of car traded, this can cause demand to decline in the face of falling prices – if, as seems reasonable, the suppliers with better-quality cars withhold them as the price falls, leaving only the poorer-quality cars to be offered at lower prices. Note that this problem does not arise in the new car market. While this market is, unfortunately, not free from ‘lemons’, the probability of being stuck with a lemon can be ascertained from sources such as consumer reports. The fraction of new cars entering the market as lemons does not vary with the price or discount offered on new cars.

Varian (1992, p. 469) offers the following simple characterization of the model. Assume there is a quality-of-car index q, which is uniformly distributed between 0 and 1. Additionally, assume the demand for cars is a function of this quality to the extent that the price offered for cars of quality q is exactly $(3/2)q$ and that, on the other side of the market, suppliers with a car of quality q would be willing to sell for price q or better. There is clearly scope for mutually beneficial trade in this market, as any price between q and $(3/2)q$ leaves both the buyer and seller of a car with quality q better off.

On the other hand, if the buyer is unable to perceive the quality of the car but has to rely on the average quality of cars traded in the second-hand market as a measure of the expected quality of any car purchased, then the price offered is $(3/2)q^*$, where q^* is the average quality in the market.

But on the supply side, of course, sellers know the exact quality of their cars and, for any price p, only those with quality p or lower will offer cars for sale. Thus, the observed quality of cars traded at price p will be $p/2$. However, at quality $p/2$ there will be no cars demanded, as cars of this average quality fetch an offer of only $(3/2)q^*=(3/2)(p/2)=(3/4)p$. So no cars will traded at this price. But nor will a fall in the price offer any improvement because, if price falls, then so too will the quality of car offered to the market and the average quality of cars observed. As things stand, there is no price that will allow cars to be traded. Potentially mutually advantageous trades are not made. Economic welfare is lower than it might be. The culprit is, of course, asymmetric information.

It is the inability of the supply side of the market (which possesses the hidden information about car quality) to meaningfully communicate this information to the buyers that undermines the potential for mutually advantageous trades. The existence of lemons inhibits the proper functioning of the market. Akerlof points out that the inability of older people to secure health-care insurance, the inability of minorities to secure decent employment prospects, the external costs of dishonest business practices, and the difficulty developing countries experience in establishing capital markets can all be viewed as manifestations of the same ‘lemons’ problem, i.e., asymmetric information.

In awarding the 2001 Nobel Memorial Prize in Economics to George Akerlof, Michael Spence and Joseph Stiglitz, the Royal Swedish Academy of Sciences cited ‘their analyses of markets with asymmetric information’. In reviewing the contributions of these prize winners, Rosser (2003) identifies a nascent discussion of this idea in the earlier economics literature, but there is little doubt that it was with the publication of Akerlof's 1970 ‘Market for “lemons”’ paper that the metaphorical light bulb was switched on in the economics community and the idea of asymmetric information started to become integrated into economics. As a recent survey by Riley (2001) makes clear, this concept is now an important feature of modern approaches to development economics, financial economics, industrial organization, international economics, labour economics, and many other areas. It is now difficult to imagine the world of economics without this insight.

Other work

While for many people the ‘lemons’ paper stands as a seminal example of the power of microeconomic analysis, the underlying motivation that led Akerlof to investigate this area was actually macroeconomic. Cyclical fluctuations in the car market were seen as a major destabilizing factor in the macroeconomy: hence the original research effort. Throughout his career Akerlof has been driven by a desire to develop macroeconomics in a way that allows problems such as unemployment to be better understood. Never happy with the neoclassical synthesis and distinctly critical of the New Classical economics, Akerlof has been a major contributor to the development of New Keynesian Economics (2002). Indeed, his work can be seen as a lifetime effort to create a better behavioural micro-foundation to macroeconomics – continuing in the tradition started by Keynes’ (1936) *General Theory*.

Caste and identities

In subsequent work the ‘lemons’ paper was soon developed into an analysis of caste systems (1976; 1985), in which irrational and economically inefficient belief systems can be sustained out of a concern

for individual well-being, albeit at the cost of society's overall welfare. This work is typical of Akerlof's approach to economic theory in that it seeks to broaden our view of economic exchange from the simplistic dyad of buyer and seller (the focus of so much economic analysis) to admit the real possibility that such exchanges are heavily conditioned by the existence of wider social forces. In this specific case, people adhere to what are obviously dysfunctional behaviours because, in their individual calculus, the costs of being seen to break such conventions (and hence being outcaste) outweigh any individual short-term gains. Thus, individually rational action leads to a macroeconomically inefficient outcome.

More generally, people can be seen as exhibiting patterns of behaviour that are consistent with chosen identities but would be otherwise difficult to explain (Akerlof and Kranton, 2000). Such identities are chosen in an attempt to fit most comfortably into society, given people's individual circumstances. The choice of identity brings with it a set of behaviours and an exposure to the behaviour of others with whom one identifies. This stream of work represents a major step in bridging the gap between economics and sociology that is so aptly summarized by James Duesenberry (quoted in Granovetter, 1985, p. 485): ‘economics is all about how people make choices; sociology is all about how they don't have any choices to make.’

This approach led Akerlof to empirical analyses of the dramatic rise in out-of-wedlock births (Akerlof, Yellen and Katz, 1996) and the marked increase in the number of men living without children (1998). These papers demonstrate that the rise of children born to unmarried mothers and the increase in men living outside of households with children can each be ascribed to changing norms (the notion of the shotgun marriage and the destigmatization of out-of-wedlock births) that have more to do with changing technology (birth control) and the social reaction to these changes than to any wealth or incentive effects arising from welfare programmes.

This enthusiasm to engage with real-world data and empirical work is another salient characteristic of Akerlof's work. Somewhat unusually, for a theorist of major repute, he has throughout his career undertaken empirical studies of the major social and economic policy issues of the day. Thus, in addition to the analysis of family structure and poverty mentioned above, he has studied the distribution of employment and unemployment experience (Akerlof and Main, 1980, 1981), job mobility (Akerlof, Rose and Yellen, 1988), German reunification (Akerlof, Rose, Yellen, and Hessenius, 1991), financial malfeasance (Akerlof and Romer, 1993), and the inflation-unemployment trade-off (Akerlof, Dickens and Perry, 1996, 2000). Akerlof's intellectually open and outgoing approach to his work also shows in the wide range of co-authors involved in his theoretical work, including, for example, Akerlof and Miyazaki (1980), Akerlof and Milbourne (1980), Akerlof and Katz (1989), Akerlof and Yellen (1990), and Akerlof and Kranton (2005). As will be seen below, his collaboration with Janet Yellen has been the most sustained and intellectually productive.

Near-rational economic behaviour

While the ‘lemons’ paper is undoubtedly his most famous, the stream of papers that best demonstrates Akerlof's New Keynesian pedigree starts with Akerlof (1969). This paper investigates structural unemployment in a framework that sees firms as being in monopolistic competition and having staggered price setting, with wages emerging as bargains struck between firms and workers. With Taylor's (1979) incorporation of rational expectations, this links directly to the overlapping contracts approach that now lies at the heart of the New Keynesian model. Akerlof also deployed this approach in

the study of monetary policy (1973; 1978; 1979). Here, simple monitoring rules by agents of their bank balances are shown to make both monetary and fiscal policy effective.

Extending this approach more generally, Akerlof and Yellen (1985) demonstrate that what appear as rule-of-thumb behavioural rules deployed in economic decision-making actually bring with them substantial savings in computational costs (and deal with the bounded rationality problem) while, at the same time, imposing only second-order costs on the agent by way of lost economic efficiency. In this sense, such rules of thumb are quite sustainable and sensible modes of behaviour. The insights of this paper have far-reaching implications. Accepting the existence of such behaviour not only points to why monetary policy might be effective but also explains why there can, indeed, be significant trade-offs between inflation and unemployment, particularly at low rates of inflation (Akerlof, Dickens and Perry, 1996, 2000).

Friedman's (1968) original attack on the notion of a long-run trade-off between inflation and unemployment was further strengthened by the incorporation of rational expectations by the New Classical economists, Lucas (1972) and Sargent (1971). Deploying the Akerlof and Yellen (1985) insight of near-rational behaviour towards inflation, Akerlof, Dickens and Perry (2000) demonstrate that at low rates of inflation, such as were typical in the 1950s and are now prevalent once again, there can be an empirically significant trade-off between inflation and unemployment. The fact is that in setting wages and prices economic agents (business people, wage negotiators and so on) do not behave exactly as economic models of rational expectations would suggest – at least not when inflation is moderate and the costs of deviating from such rationality are modest when compared with the informational and computational costs involved.

Sociologically based efficiency wage theory

In attempts to explain the unemployment that fiscal and monetary policy is often deployed to remedy, a standard question is why in the face of unemployment wages do not simply decline, so restoring equilibrium in the market. The answer is, of course, that cheaper is not always better. In a paper evocatively titled ‘Jobs as dam sites’, Akerlof (1981) explains that, just as it makes poor economic sense to construct a lower-quality dam on a prime site (no matter that it may be cheaper), so it may not make economic sense to hire cheaper labour even when available. These ideas, further developed in Akerlof (1982) and most elegantly expressed in Akerlof and Yellen (1990), provide a sociologically rooted explanation for efficiency wages.

The key idea here is that the exchange between employer and employee is rich and complex, extending well beyond the narrow instrumental delivery of labour in return for wages. Workers who display ‘consummate’ cooperation in playing their part to achieve the objectives of the organization are much preferred to those exhibiting ‘perfunctory’ cooperation (see Williamson, Wachter and Harris, 1975, p. 266). Part of the key to ensuring the higher-productivity outcome is being seen to pay a fair wage. The concept of fair wage-effort is socially determined, and both equity theory from social psychology and social exchange theory from sociology offer explanations of how workers react when this balance is disturbed. From this perspective, the financial savings from lowering wages can be a poor bargain when set against the impact on the productivity of the workforce. In the face of such rigidity coming about through the individually rational decisions of employers, there is clear scope for macroeconomic policy

to effect a coordinated move to a higher level of employment. This is a key insight of the efficiency wage model of the labour market (Akerlof and Yellen, 1986).

Psychologically based models

The incorporation of psychological insights into economics has proved highly successful in recent years, as indicated by the award of the Nobel Prize in 2002 to Daniel Kahneman. Akerlof and Dickens (1982) is an early contribution to this movement, drawing on the notion of cognitive dissonance whereby individuals choose their beliefs or view of a situation in such a way that renders them the greatest comfort or happiness. In this way, it is possible to explain many common phenomena that otherwise seem to make little economic sense, such as the widespread flouting of workplace safety standards. In some ways the more recent work in Akerlof and Kranton (2005) on choice of identity can be seen as a sociological version of this same phenomenon. The common theme is that social actors are capable of choosing the frame through which they view their circumstances and, unsurprisingly, can be expected to choose an approach that, given the situation in which they find themselves, offers them the greatest comfort. To an external observer this can often result in behaviours that are perplexing.

Thus, in Akerlof (1991) a psychologically based explanation is offered for the widely documented phenomenon of people acting in ways that seem too short-sighted to be in their interest. This is seen in the widespread failure to make adequate provision for retirement or to save enough in general. Drawing on a personal experience during a year living in India during the late 1960s, Akerlof recounts how day after day he procrastinated over mailing off a promised package to Joseph Stiglitz. This is developed into a model that demonstrates why in repeatedly opting for what appears as the best short-term course of action (to procrastinate) one is often left in a situation that in retrospect one may regret. The insights offered by this model of economic behaviour are both powerful and far-reaching, and later proponents, such as David Laibson (1997), have extended the area into neurological studies of the brain under the heading ‘neuroeconomics’.

Conclusion

If economists were ever to adapt the psychologist's stimulus-response technique into a game of declaring a famous economist's name as a stimulus and then noting the response, it seems clear that the overwhelming response to ‘George Akerlof’ would be ‘lemons’. This would, at the same time, be both a sufficient response and an insufficient response. As the above discussion has shown, it is insufficient to try to capture such a major body of important studies by reference to one paper. Akerlof has not only dealt with asymmetric information but, as a major contributor to modern Keynesian economics, has also confronted the major macroeconomic issues of the day, most notably by providing the behavioural underpinnings to explain the efficacy of interventionist economic policy.

Yet the ‘lemons’ response could arguably be judged sufficient in the sense that the ‘lemons’ paper contains all of the elements that make Akerlof's approach to economic theory so different and so potent. Mark Granovetter (1985) criticizes economic models as either totally ignoring the influence of social structures and relations or else going to the other extreme, by being oversocialized in the sense that there are really no choices left for agents to make. Akerlof is one of a small but growing set of economists

who manage to position their models on the middle ground. Far from Friedman's (1953) positive economics approach, which regards assumptions as something to be minimized and whose realism is of no consequence as long as the predictive power of the model holds up, Akerlof adheres to an approach that utilizes models based on closely observed empirical examples. The fact that the most observers believe that monopolistic competition is the norm means to Akerlof that such a feature must appear in the model. A model utilizing perfect competition might be able to do just as well, but would be rejected in the face of Akerlof's pragmatic goal of making his models as near to the observed reality as possible while still being tractable.

‘The market for “lemons”’ will almost certainly stand as Akerlof's best-known contribution, having provided the impetus for radical new ways of looking at events in so many areas of economics. But it is also an excellent exemplar of a different approach to economic modelling. It is this pragmatic approach to economic modelling that makes all of Akerlof's contributions so worthwhile.

See Also

- caste system
- efficiency wages
- information aggregation and prices
- social norms
- economic sociology

Selected works

1969. Relative wages and the rate of inflation. *Quarterly Journal of Economics* 83, 353–74.

1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

1973. The demand for money: a general-equilibrium inventory-theoretic approach. *Review of Economic Studies* 40, 115–30.

1976. The economics of caste and of the rat race and other woeful tales. *Quarterly Journal of Economics* 90, 599–617.

1978. The microfoundations of a flow of funds theory of the demand for money. *Journal of Economic Theory* 18, 190–215.

1979. Irving Fisher on his head: the consequences of constant target-threshold monitoring for the demand for money. *Quarterly Journal of Economics* 93, 169–87.

1980 (With B. Main.) Unemployment durations and unemployment experience. *American Economic*

Review 70, 885–93.

1980 (With R. Milbourne.) Irving Fisher on his head II: The consequences of the timing of payments for the demand for money. *Quarterly Journal of Economics* 94, 145–57.

1980 (With H. Miyazaki.) The implicit contract theory of unemployment meets the wage bill argument. *Review of Economics Studies* 47, 321–38.

1981. Jobs as dam sites. *Review of Economic Studies* 48, 37–49.

1981 (With B. Main.) An experience-weighted measure of employment and unemployment durations. *American Economic Review* 71, 1003–11.

1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97, 543–69.

1982. (With W. Dickens.) The economic consequences of cognitive dissonance. *American Economic Review* 72, 307–19.

1985. Discriminatory status-based wages among tradition-oriented stochastically based coconut producers. *Journal of Political Economy* 93, 265–76.

1985. (With J. Yellen.) A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100, 823–38.

1986. (With J. Yellen, eds.) *Efficiency Wage Models of the Labor Market*. New York: Cambridge University Press.

1988. (With A. Rose and J. Yellen.) Job switching and job satisfaction in the US labor market. *Brookings Papers on Economic Activity*, 2, 495–592.

1989. (With L. Katz.) Workers' trust funds and the logic of wage profiles. *Quarterly Journal of Economics* 103, 525–36.

1990. (With J. Yellen.) The fair-wage effort hypothesis and unemployment. *Quarterly Journal of Economics* 105, 255–283.

1991. Procrastination and obedience. *American Economic Review* 81(2), 1–19.

1991. (With A. Rose, J. Yellen and H. Hessenius.) East Germany in from the cold: The economic aftermath of currency union. *Brookings Papers on Economic Activity*, 1, 1–105.

1993. (With P. Romer.) Looting: The economic underworld of bankruptcy for profit. *Brookings Papers on Economic Activity*, 2, 1–60.

1996. (With W. Dickens and G. Perry.) The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1, 1–59.

1996. (With J. Yellen and M. Katz.) An analysis of out-of-wedlock childbearing in the United States. *Quarterly Journal of Economics* 111, 277–317.

1998. Men without children. *Economic Journal* 108, 287–309.

2000. (With W. Dickens and G. Perry.) Near-rational wage and price setting and the optimal rates of inflation and unemployment. *Brookings Papers on Economic Activity* 2000: 1, 1–60.

2002. Behavioral macroeconomics and macroeconomic behavior. *American Economic Review* 92, 365–94.

2000. (With R. Kranton.) Economics and identity. *Quarterly Journal of Economics* 115, 715–53.

2005. (With R. Kranton.) Identity and the economics of organizations. *Journal of Economic Perspectives* 19, 9–32.

Bibliography

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91, 481–510.

Keynes, J. 1936. *The General Theory of Employment, Interest and Money*. London: Macmilan.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112, 443–77.

Lucas, R. Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Riley, J. 2001. Silver signals, twenty-five years of screening and signaling. *Journal of Economic Literature* 39, 432–78.

Rosser, J. Jr. 2003. A Nobel Prize for Asymmetric Information: the economic contributions of George Akerlof, Michael Spence and Joseph Stiglitz. *Review of Political Economy* 15, 3–21.

Sargent, T. 1971. A note on the‘accelerationist’ controversy. *Journal of Money, Credit and Banking* 3, 721–5.

Taylor, J. 1979. Staggered wage setting in a macro model. *American Economic Review* 69, 108–13.

Varian, H. 1992. *Microeconomic Analysis*, 3rd edn. New York: W. W. Norton & Company.

Williamson, O., Wachter, M. and Harris, J. 1975. Understanding the employment relationship: the analysis of idiosyncratic exchange. *Bell Journal of Economics* 6, 250–78.

How to cite this article

Main, Brian G. M. "Akerlof, George Arthur (born 1940)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000208> doi:10.1057/9780230226203.0028

The New Palgrave Dictionary of Economics Online

Åkerman, Johan Gustav (1888 – 1959)

K. Velupillai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Åkerman, Johan Gustav; capital theory; fixed capital; Wicksell, Knut

Article

Gustav Åkerman received perhaps the supreme accolade for any economist working in the theory of capital, Knut Wicksell's endorsement. Wicksell concluded his masterly review of the first part of Åkerman's doctoral dissertation with the following acknowledgement: ‘I am convinced that on the whole the author has made a really significant contribution to the theory of capital’ (Wicksell, 1934, Appendix 2(a), p. 273).

Born in 1888, Åkerman obtained his doctoral degree from the law faculty of the University of Lund in the days before economics as a subject had independent status, in 1923. He was appointed Docent (Associate Professor) in Lund the same year, on the strength of his brilliant doctoral dissertation, ‘Realkapital und Kapitalzins’. He was subsequently appointed Professor of Political Economy and Sociology at what was later to become the University of Gothenburg in 1931, and remained there until his retirement. He died in 1959.

Wicksell's famous two-part review article (the second part being on ‘Åkerman's Problem’) of the first volume of his dissertation assured him international fame. The first volume of his dissertation dealt with the static problems of fixed-capital systems and the second volume with dynamic problems for analogous systems. His method of analysis, in the Austrian tradition, was very similar to Böhm-Bawerk's approach: copious numerical and special examples to illustrate subtle and deep general propositions. It is to his great credit that he seldom went wrong in deriving propositions by this primitive method; as a testimony to his insights we can cite concepts and issues at the frontiers of capital theoretic debates that owe much to the results of his dissertation of 1923–4: Wicksell effects, truncation of production flows, transverse flows, to name but a few.

He was perhaps also the first (after the early classical economists) to try to approach the problem of fixed capital as joint products – a method made famous by von Neumann and Sraffa in more recent times.

Even before his capital theoretic writings, he had engaged the grand old man of Swedish economics, Knut Wicksell, in a debate in the pages of the *Ekonomik Tidskrift* (1922) on the latter's proposals on norms for price stabilization.
His later work was mostly on practical problems of economic policy.

Selected works

1923. *Realkapital und Kapitalzins: Heft* I. Stockholm: Centraltryckeriet.

1924. *Realkapital und Kapitalzins: Heft* II. Stockholm: Centraltryckeriet.

1931. Om den Industriella Rationaliseringen och dess Verkningar (bilaga 2 till Arbetslöshetsutredningens Betänkande). *SOU* 42, Stockholm.

1947. *Engelsk Arbetslöshet och Arbetslöshetspolitik*. Göteborg.

1956. Marginal productivity with different agricultural products. In *Twenty Five Economic Essays in Honour of Erik Lindahl*. Stockholm: Ekonomisk Tidskrift.

Bibliography

Wicksell, K. 1934. Real capital and interest: Dr. Gustav Åkerman's *Realkapital und Kapitalzins*. Appendix 2(a) of K. Wicksell, *Lectures on Political Economy*, London: Routledge & Kegan Paul.

How to cite this article

Velupillai, K. "Åkerman, Johan Gustav (1888–1959)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000070> doi:10.1057/9780230226203.1852

The New Palgrave Dictionary of Economics Online

Åkerman, Johan Henrik (1896 – 1982)

K. Velupillai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Åkerman, Johan Henrik; business cycle; Kalkyl models; Lund School; stochastic; Stockholm School

Article

Somewhat lesser known internationally than his elder brother Gustaf, Johan Henrik Åkerman was, however, much better known inside Sweden. He was a prolific contributor to the theoretical, methodological, epistemological and policy debates in Sweden for almost 50 years. He challenged almost single-handedly (at least inside Sweden) the methodological position of the so-called Stockholm School and made valiant (but unsuccessful) attempts to provide an alternative vision which he described as the ‘Lund School’ method.

Johan Henrik Åkerman was born in Stockholm in 1896, graduated from the Stockholm Business School in 1918, and then spent two terms at Harvard University (1919–20) working with Warren M. Persons. On his return to Sweden, he continued his postgraduate studies in the Universities of Uppsala and Sweden. He obtained his PhD (Fil.Dr.) in 1929 from the University of Lund, where he was appointed Associate Professor in Political Economy and Economic Statistics in 1932. In 1943 he was appointed Professor in Political Economy in Lund, and retained that position until his retirement in 1961. His scientific publications of more than 150 items included several books, some of which were translated into English and German. He was almost totally deaf from a very early stage in his life and totally deaf during his tenure as Professor in Lund. He died in 1982.

Johan Åkerman's outstanding doctoral dissertation had the title *On the Rhythm of Economic Life* (*Om Det Ekonomiska Livets Rytmik*). It was an ambitious attempt to codify, theoretically and empirically, all aspects of the problem of fluctuations in economic life. It was based on the theoretical framework of Wicksell's *Geldzins und G•terpreise* (1898) and Cassel's ‘Om kriser och dåliga tider’ (On crises and bad times), *Ekonomisk Tidskrift*, 1904; and on the empirical methodology of the budding NBER work. Åkerman's dissertation was perhaps the earliest attempt to apply spectral analysis for studying time series phenomena in economics. His main examiner for the doctoral degree was Ragnar Frisch, whose more influential later work on ‘Propagation problems and impulse problems in dynamic

economics' (1933) owes much to Åkerman's specific considerations of Wicksell's celebrated 'rocking-horse' example. This latter example, delineating one influential strand in business cycle methodology – the stochastic approach – stressed the important distinction between sources of propagation and impulse mechanisms. It is to Åkerman's great credit that he was able to revive and place in the centre of discussion on business-cycle methodology this important distinction, which was initially stressed by Wicksell in an obscure footnote to a review article in the *Ekonomisk Tidskrift*, 1918. It is both important and topical in view of recent developments in equilibrium business cycle theories, where these issues are central. Indeed, Åkerman's dissertation could claim to be an early manifesto of aspects of the New Classical Economics.

In the 1930s and 1940s Åkerman's research interests shifted towards methodological and epistemological problems – mainly under the influence and impact of the works of members of the Stockholm School (and later the Keynesians). He was severely critical of the rationality and individualistic assumptions underlying the then popular macroeconomic theories (and their microeconomic underpinnings). He developed a highly original alternative modelling strategy for macroeconomics based on a so-called dual principle of 'causal' and 'computing' ('Kalkyl') models where institutional details and socio-economic classes were explicit factors. His research and reflections on these matters, spread over a period of 30 years, were summarized and elegantly delivered as a lecture on the occasion of his retirement ('Avskedsföreläsning') from the Professorship in Lund on 9 May 1961 ('Fyra methodologiska moment', *Ekonomisk Tidskrift*, 1961). The depth of his understanding of recent developments in economic analysis, and the scope of his comprehensive references to epistemological developments in theoretical physics and relevance to economic theory, were displayed in that last masterly lecture.

His lifelong interest in the political economy of business cycles was also reflected in a highly original work on political business cycles, *Ekonomiskt Skeende och Politiska Förändringar*. He was continuing a Swedish tradition on this subject – and quite independently of Kalecki's important work on political business cycles – initiated by Herbert Tingsten's inter-war work on *Political Behaviour: Studies in Election Statistics* (1937).

Retrospectively, it is significant that Johan Åkerman's two pioneering studies on problems of fluctuations in mixed economies have their counterpart in research in the frontiers of the theory and empirical analysis of business, political and economic cycles even today.

Selected works

1928. *Om Det Ekonomiska Livets Rytmik*. Stockholm: Nordiska Bokhandeln.

1932. *Industriförbundets Produktionsindex: Motiv och Principer*. Stockholm: Sveriges Industriförbund.

1934. *Konjunkturteoretiska Problem*. Lund: C.W.K. Gleerup.

1936. *Ekonomisk Kausalitet*. Lund.

1939. *Ekonomisk Teori, I: De Ekonomiska Kalkylerna*. Lund and Leipzig.

1944. *Ekonomisk Teori, II: Kausalanalys av det Ekonomiska Skeendet*. Lund and Leipzig.

1945. *Banbrytare och Fulföljare inom Nationalekonomien*. Lund: C.W.K. Gleerup.

1946. *Ekonomiskt Skeende och Politiska Förändringar*. Lund: C.W.K. Gleerup.

1952. Innovationer och Kumulativa Förlopp. *Ekonomiskt Tidskrift*.

1960. Samhällsstruktur och Ekonomisk Teori. *Samhällsvetenskapliga studier* 18 (Lund).

1961. Fyra Metodologiska Moment. *Ekonomiskt Tidskrift*.

Bibliography

Frisch, R. 1933. Propagation problems and impulse problems in economic dynamics. In *Economic Essays in Honour of Gustav Cassel*. London: Allen & Unwin.

Tingsten, H. 1937. *Political Behaviour: Studies in Election Statistics*. Stockholm Economic Studies, No. 7, London: P.S. King & Son.

How to cite this article

Velupillai, K. "Åkerman, Johan Henrik (1896–1982)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000071> doi:10.1057/9780230226203.1853

The New Palgrave Dictionary of Economics Online

Albert the Great, Saint Albertus Magnus (c.1200 – 1280)

Odd Langholm

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Albert the Great; Aquinas, St. T.; Aristotle; just price; labour; private property; usury; value

Article

Albert the Great, *doctor universalis*, was the foremost German philosopher and theologian of the Middle Ages. He was born in the village of Lauingen on the Danube and became a member of the Dominican Order while studying at Padua. He subsequently studied at Paris, and eventually taught there as well as in Dominican houses in Germany, primarily Cologne, where he became Regent Master of Studies and where he died. Albert served as Bishop of Regensburg, was German Provincial of his Order and Master of the Sacred Palace of the Pope, but repeatedly returned to Cologne to devote himself to study and teaching. He composed a comprehensive set of commentaries on the works of Aristotle and is considered the founder of Christian Aristotelianism. He was canonized and named a Doctor of the Church in 1931. Ten years later he was declared patron ‘of all who cultivate the natural sciences’, which indicates his main area of interest. In what is now called economics he is overshadowed by his famous student Thomas Aquinas, but in fact he made important contributions of his own. They are found in his comments on Scripture and on the theological *Sentences* of Peter Lombard as well as in some of his Aristotelian works. On the *Nicomachean Ethics* he composed a close textual commentary, and later a freer *Ethica*. His *Politica* is the first complete Latin commentary on Aristotle's *Politics*.
Two striking features of Albert the Great's discussions of matters relating to material wealth and economic activity are his empirical orientation and the store he sets by human labour. He argues that private property is the best arrangement in civil society because common ownership engenders strife, pointing to the observable fact that those who reap less than their labour share under communism are likely to protest and cause trouble (*Politica*, II.2). In Book V of the *Nicomachean Ethics*, Aristotle discusses justice in relation to barter between persons of different occupations and states obscurely that as one person is to another person, thus are their respective products to each other. Albert the Great interprets this formula in terms of respective input: as a farmer is to a shoemaker in labour and expenses, thus the product of the shoemaker is to the farmer's product (*Ethica*, V.2.9). This solution is explained

by a factual observation: unless a carpenter receives for a bed what it cost him to make it, he will not make any more beds (*Ethica*, V.2.7). In his commentary on the *Sentences*, Albert's approach and conclusion are different. In the absence of economic coercion and fraud, the just price is that at which a good sold can be valued according to the estimation of the market at the time of the sale (*Comm*. *Sent*., IV.16.46). If these arguments are combined, what Albert asserts is that the competitive market determines value but that unprofitable goods will be withdrawn from the market.

Albert discussed the purposes and properties of money and warns against debasement of the currency. Examining usury in the same context, he rejects the ‘barren metal’ theory falsely attributed to Aristotle. Lending for profit is a perverse use of money, which makes it *seem as though* money reproduces itself (*Politica*, I.7). Usury is a form of economic coercion because it is paid with a conditional, not an absolute, will. The payment is voluntary only in the sense in which, according to Aristotle, the captain of a ship in peril jettisons cargo voluntarily (*Comm*. *Sent*., III.37.13). But the full force of Albert the Great's denunciation of usury comes through in one of his Gospel commentaries: ‘By hard labour [the borrower] has acquired something on which he could live, and this the usurer, suffering no distress, spending no labour, fearing no loss of capital by misfortune, takes away, and through the distress and labour and changing luck of his neighbour collects and acquires riches for himself’ (*Super Lucam*, 6.35).

See Also

- Aquinas, St Thomas
- just price
- scholastic economics

Selected works

Opera Omnia, ed. A. Borgnet. Paris: Louis Vives.

Ethica, vol. 7.

Politica, vol. 8.

Super Lucam, vols. 22–3.

Super Tertium Sententiarum, vol. 28.

Super Quartum Sententiarum, vols 29–30.

Bibliography

Langholm, O. 1992. *Economics in the Medieval Schools*. Leiden: Brill.

Weisheipl, J.A. 1980. The life and works of St Albert the Great. In *Albertus Magnus and the Sciences: Commemorative Studies 1980*, ed. J.A. Weisheipl.Toronto: Pontifical Institute of Mediaeval Studies.

How to cite this article

Langholm, Odd. "Albert the Great, Saint Albertus Magnus (c.1200–1280)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000255> doi:10.1057/9780230226203.0029

The New Palgrave Dictionary of Economics Online

alienation

Jonathan Wolff

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘alienation’ is associated especially with the early writings of Karl Marx, for whom the core idea was that of human beings becoming detached from part of their ‘essence’. At one stage Marx hoped to demonstrate that all of the concepts of classical economics could be derived from the concept of alienation. As he matured, exploitation and surplus value replaced alienation at the heart of his analysis. Nevertheless, Marx's observations regarding alienation remain insightful to this day.

Keywords

alienation; capitalism; classical economics; communism; Engels, F.; exploitation; Feuerbach, L.; Hegel, G. W. F.; historical materialism; Marx, K. H.; Smith, A.; surplus value

Article

Although the word ‘alienation’ is commonly used to express an idea of, perhaps, resentful dislocation, within social theory its central use is to be found in the early writings of Karl Marx (1818–83), and especially his *Economic and Philosophical Manuscripts*, also known as the *Paris Manuscripts*, of 1844. Marx did not invent the concept; it was widely used by the group of Young Hegelian philosophers with whom he associated in the early 1840s, and especially by Ludwig Feuerbach, in his account of religious alienation. In turn, these thinkers had been influenced by Hegel's concept of externalization.
The term itself cannot be given a single, uncontroversial definition; rather. it seems a marker for a constellation of ideas, not always present in every use. A common understanding sees alienation as a subjective feeling. For Marx, however, alienation is an objective fact about the world, and in its core use we can often distinguish three constitutive elements. The most easily observable aspect is that human beings become detached from something that properly belongs to them. This implies, of course, a second element; a normative claim about how things ought to be, that is, their non-alienated state. Finally, and most metaphysically ambitious, that from which man has become separated nevertheless returns in some ‘alien’ form; by this means human beings are not only estranged from but also

dominated by their own essence or products.

Marx's use of the idea of alienation went through a number of phases. The first takes over and extends Feuerbach's concept of religious alienation. The second is the most ambitious: alienation is used as an explanatory concept in the sense that it is claimed that all the categories of economics can be generated from an analysis of the concept of alienation. This neo-Hegelian phase, however, was short-lived, not surviving beyond the *Economic and Philosophical Manuscripts*; Marx was shortly to become aware that a priori philosophy was not the best tool for economic analysis. In a third phase the idea of alienation was retained as a central concept in the understanding of the effect of capitalism on human beings, and held out the promise of emancipation. This, however, faded in to a fourth and final phase where, although, the same ideas were present, the term itself was used less and less, and Marx's key concept for the analysis of capitalism became surplus value or exploitation.

Religious alienation: the influence of Feuerbach

The young Marx wrote for a philosophical audience which had accepted Feuerbach's reversal of traditional theology in which he asserted that human beings had created God in their own image; indeed this is a view with a long history. Feuerbach's distinctive contribution was to argue that worshipping God diverted human beings from enjoying their own human powers. While accepting much of Feuerbach's account, Marx criticized Feuerbach on the grounds that he had failed to understand why people fell into religious alienation and so was unable to explain how it could be transcended. Marx's explanation, of course, was that religion was a response to alienation in material life, and could not be removed until human material life was emancipated, at which point religion would wither away. This was discussed in Marx's 1843 essay *Contribution to the Critique of Hegel's Philosophy of Right: Introduction*, and, very briefly, in the *Theses on Feuerbach* of 1845.

Precisely what it is about material life that creates religion was not set out by Marx with complete clarity. However, it seems that at least two aspects of alienation are responsible. One is alienated labour, which will be explored shortly. A second is the need for human beings to assert their communal essence. Marx argued that, whether or not we explicitly recognize it, human beings exist as a community, and what makes human life possible is our mutual dependence on the vast network of social and economic relations which engulf us all, even though this is rarely acknowledged in our day-to-day life. Marx's view appeared to be that we must, somehow or other, acknowledge our communal existence in our institutions. At first it is ‘deviously acknowledged’ by religion, which creates a false idea of a community in which we are all equal in the eyes of God. After the post-Reformation fragmentation of religion, when religion is no longer able to play the role even of a fake community of equals, the state fills this need by offering us the illusion of a community of citizens, all equal in the eyes of the law. But the state and religion will both be transcended when a genuine community of social and economic equals is created.

Here we see all three aspects of alienation. Human communal existence has come apart from its essence through the invention of God. The normatively correct situation for humans, however, is one in which they enjoy their essence on earth. Finally, our own communal essence returns to dominate us in the alien form first of religion and then of the political state.

Alienated labour as the foundation of economics: the Neo-Hegelian phase

It is commonplace to observe that Marx transformed a critique of religion into a critique of society. The *Economic and Philosophical Manuscripts* is an important element in this early critique. Here Marx famously depicts workers under capitalism as suffering from four types of alienated labour. First, they are alienated from their products, in at least two ways: they may not understand what they are making, and, as soon as it is created, is taken away from them. Second, they are alienated in productive activity (work) which is experienced as a torment, often requiring the performance of mindless or back-breaking toil. Third, they are alienated from their species-being. The distinctive feature of human beings is their productive and creative power. Yet under capitalism humans produce blindly and not in accordance with their truly human powers. Consequently, argues Marx, workers feel free only when away from work, engaged in activities that they share with animals; eating, drinking and having sex. Hence they are alienated from their distinctively human powers. Finally, they are alienated from other human beings, where the relation of exchange replaces mutual need.

These categories overlap in some respects, but this is no surprise given Marx's remarkable methodological ambition in these writings. Essentially he attempted to apply a Hegelian deduction of categories to economics, trying to demonstrate that all the categories of bourgeois economics – wages, rent, exchange, profit, and so on – were ultimately derived from an analysis of the concept of alienation. Consequently, each category of alienated labour was supposed to be deducible from the previous one. However, Marx got no further than a rather unconvincingly attempt to deduce categories of alienated labour from each other. Quite possibly in the course of writing he came to understand that a different methodology was required for approaching economic issues. Nevertheless, we are left with a very rich text on the nature of alienated labour.

Alienation and emancipation

Marx based his account of capitalism not, at this stage, on independent empirical study, but on his readings of the works of the classical economists, most notably Adam Smith; much of the descriptive content of the idea of alienated labour from was derived his reading of *The Wealth of Nations*. However, by setting it within the theory of alienation he was able to depict capitalism as a world which was by its nature contrary to the human essence, and therefore with an inbuilt tendency to its own destruction.

The bridge between Marx's early analysis of alienation and his later social theory is the idea that the alienated individual is ‘a plaything of alien forces’, albeit alien forces which are themselves a product of human action. In our daily lives we take decisions that have unintended consequences, which then combine to create large-scale social forces which may have an utterly unpredicted effect. In Marx's view the institutions of capitalism – themselves the consequences of human behaviour – come back to structure our future behaviour, determining the possibilities of our action. For example, for as long as a capitalist intends to stay in business he must exploit his workers to the legal limit. Whether racked by guilt or not, the capitalist must act as a ruthless exploiter. Similarly, the worker must take the best job on offer; there is simply no other sane option. But by doing this we reinforce the very structures that oppress us. The urge to transcend this condition, and to take collective control of our destiny – whatever that would mean in practice – was one of the motivating and sustaining elements of Marx's attraction to communism.

However, Marx's idea of emancipation – of a non-alienated society – has largely to be inferred from its

negative. There are, however, two short passages in the early writings which are often cited in this context. The more famous is from the *German Ideology*, co-authored with Engels in 1845, and like many of their works unpublished in their lifetime. Here Marx and Engels described future society as a rural idyll, lived in complete freedom to order one's own life. Recent scholarship, however, casts doubt on whether this passage, which is quite unlike anything else written by Marx and Engels, was intended as a serious contribution to the development of their view (Carver, 1998).

A second short passage appears at the end of the text ‘On James Mill’ (1844) in which non-alienated labour is briefly described in terms which emphasize both the producer's immediate enjoyment of production as a confirmation of his or her powers, and the idea that production is to meet the needs of others, thus confirming for all parties our human essence as mutual dependence. Both sides of our species essence are revealed here: our individual human powers and our membership in the human community.

Alienation and the rise of ‘surplus value’

As Marx turned to economics he found philosophy of decreasing use and interest, and as he matured as a social thinker the concept of alienation becomes less and less prominent. This has led some commentators, notably Althusser, to argue that there was an ‘epistemological break’ between Marx's early, humanist, phase, and a later scientific phase, incorporating the first volume of *Capital* (1867). Although the publication, since Althusser's famous essay (Althusser, 1970), of many of Marx's writings of the 1850s shows that there is something closer to a natural development of ideas rather than a decisive break, it is true that the concept of alienation does not play the central role in Marx's later economic writings that it did in his early writings. Nevertheless, even in *Capital* there are descriptions of the labour process under capitalism which bear close comparison with the arguments of the 1844 manuscripts, and a discussion of ‘commodity fetishism’ in *Capital* is very close indeed to the idea of alienation.

Conclusion

Although Marx's economic theories play little role in contemporary economic analysis, and his theory of historical materialism is valued more for its small-scale insights rather than its long-term predictions, Marx's theory of alienation remains of great interest. On a descriptive level, Marx's account of the conditions of work under capitalism remain highly relevant if not to the developed world, then clearly to the major developing economies. Furthermore, the idea that human beings can become trapped within structures they have created for themselves is a deep insight that is constantly being rediscovered especially within the feminist and environmental movements. Marx's ideas concerning alienation are an inspiration even to those who are unaware of their source.

See Also

- Braverman, Harry
- capitalism

- commodity fetishism
- Engels, Friedrich
- exploitation
- Marx, Karl Heinrich
- socialism

Bibliography

All the writings of Marx and Engels discussed are available in McLellan (2000) and online at http://www.marxists.org (accessed 4 November 2006).

Althusser, L. 1970. *For Marx*. London: Verso, 2005.

Avineri, S. 1970. *The Social and Political Thought of Karl Marx*. Cambridge: Cambridge University Press.

Braverman, H. 1975. *Labor and Monopoly Capital*. New York: Monthly Review Press.

Carver, T., (ed.) 1991. *The Cambridge Companion to Marx*. Cambridge: Cambridge University Press.

Carver, T. 1998. *The Post-Modern Marx*. Manchester: Manchester University Press.

Colletti, L., (ed.) 1992. *Karl Marx: Early Writings*. London: Penguin.

Elster, J. 1985. *Making Sense of Marx*. Cambridge: Cambridge University Press.

Hook, S. 1950. *From Hegel to Marx*. New York: Humanities Press.

Feuerbach, L. 1957. *The Essence of Christianity*. New York: Harper.

Kolakowski, L. 1978. *Main Currents of Marxism*, vol. 1. Oxford: Oxford. University Press.

McLellan, D. 1970. *Marx Before Marxism*. London: Macmillan.

McLellan, D., (ed.) 2000. *Karl Marx: Selected Writings*, 2nd edn. Oxford: Oxford University Press.

Maguire, J. 1972. *Marx's Paris Writings*. Dublin: Gill and Macmillan.

Meszaros, I. 1975. *Marx's Theory of Alienation*. London: Merlin Press.

Ollman, B. 1971. *Alienation: Marx's Conception of Man in Capitalist Society*. Cambridge: Cambridge University Press.

Wolff, J. 2002. *Why Read Marx Today*? Oxford: Oxford University Press.

Wood, A. 1981. *Karl Marx*. London: Routledge.

How to cite this article

Wolff, Jonathan. "alienation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000073> doi:10.1057/9780230226203.0030

The New Palgrave Dictionary of Economics Online

Allais paradox

Maurice Allais
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The ‘Allais paradox’ is that risk-averse persons’ choices between alternatives tend to vary according to the absolute amounts of potential gain involved in different pairs of alternatives, even though rational choice between alternatives should depend only on how the alternatives differ. But there is no paradox once we accept the non-identity of monetary and psychological values and the importance of the distribution of cardinal utility about its average value.

Keywords

Allais paradox; Allais, M.; Bernoulli, N.; cardinal utility; de Finetti, B.; expectations; Harsanyi, J. C.; independence; preferences; propensity for risk; psychological versus monetary value; random choice; psychology of risk; Samuelson, P. A.; Savage, L. J.; St Petersburg paradox; von Neumann and Morgenstern

Article

The St Petersburg paradox and the Bernoullian formulation

Let there be a random prospect $g_1, \ldots, g_i, \ldots, g_n, \ldots, p_1, \ldots, p_i, \ldots, p_n (\Sigma_i p_i = 1)$ giving the probability p_i of positive or negative gains g_i. The early theorists of games of chance considered that a game was advantageous when the mathematical expectation

$$M = \sum_i p_i g_i (1 \leq i \leq n)$$

(1)

was positive (Allais, 1952b, pp. 68–9).

The principle of the mathematical expectation of monetary gains has proven to be open to question in the case of the *St Petersburg Paradox* outlined by Nicolas Bernoulli. For this game, we have: $g_i = 2^i$, $p_i = 1/2^i$, $n = \infty$ so that $M = \infty$. However, if the unit of value is the dollar, it can be seen that for most subjects, the psychological monetary value of the game (that is the price they are ready to pay for this random prospect) is generally lower than 20 dollars. This, at first sight, involves a paradox. To explain this paradox, Daniel Bernoulli (1738) considered the mathematical expectation of cardinal utilities $u(C + g_i)$ instead of the mathematical expectation of monetary gains, *C* being the player's capital. Thus the formulation (1) is replaced by the Bernoullian formulation

$$u(C + V) = \sum_i p_i u(C + g_i) \tag{2}$$

in which *V* is the psychological monetary value of the random prospect. He proposed to take the logarithmic expression $u = \log(C + g)$ as cardinal utility (Bernoulli, 1738; Allais, 1952b, p. 68; 1977, pp. 498–506; 1983, p. 33). It can then be shown that we have approximately $V \sim a + [\log C / \log 2]$ with $a = 0.942$, which yields $V \sim 14$ or 18 US $ for *C* equal to 10,000 or 100,000 dollars respectively (Allais, 1977, p. 572).

The neo-Bernoullian formulation

In order to measure cardinal utility from random choice, von Neumann and Morgenstern demonstrated in the *Theory of Gamevs* (1947), on the basis of a set of more or less appealing postulates, the existence of an index $B(C + g)$, such that

$$B(C + V) = \sum_i p_i B(C + g_i) \tag{3}$$

in which the index $B(C + g)$ is independent of the random prospect considered, but depends on the subject (von Neumann and Morgenstern, 1947, pp. 8–31 and 617–32; Allais, 1952b, p. 74; 1977, pp. 521–3, 591–603; 1983, p. 34).

Using other sets of postulates, Marschak, Friedman and Savage, Samuelson, Savage, etc. (Marschak,

1950 and 1951; Friedman and Savage, 1948; Samuelson, 1952; Savage, 1952 and 1954; Allais, 1952b, pp. 74–5, 88–92, and 99–103; 1977, pp. 464–5, 508–14; 1983, pp. 33–5) came to the same formulation (3), which may be referred to as the neo-Bernoullian formulation, but its interpretation differs depending on the postulates adopted. While von Neumann and Morgenstern believed, at least initially, that $B \equiv u$, the p_i being objective probabilities (Allais, 1952b, p. 74; 1977, pp. 591–2), Savage held that cardinal utility is a myth (Savage, 1954, p. 94), and that the neo-Bernoullian index B alone is real, the p_i being subjective probabilities, the existence of the function B and the p_i being proven on the basis of the axioms considered. Some authors (e.g. de Finetti, Krelle, Harsanyi) admit the existence of cardinal utility u, but they consider that $B \neq u$, and the index B is deemed to take account of the relative propensity for risk corresponding to the distribution of cardinal utility (de Finetti, 1977; Allais, 1952b, pp. 123–4; 1983, pp. 30–31).

Whereas von Neumann's and Morgenstern's opinion, accepted by most authors, is that the crucial axiom of their theory is axiom 3 Cb, I consider that their axioms 3 Ba and 3 Bb are the crucial ones (Allais, 1977, pp. 596–8). However, one way or another, irrespective of the nature of the axioms from which it is derived, the neo-Bernoullian formulation boils down to assuming the independence of the B_i for given values of the p_i. This is the principle of independence (Allais, 1952b, pp. 88–90 and 98–9; 1977, pp. 466–7).

The Allais paradox

When I read the *Theory of Games* in 1948, formulation (3) appeared to me to be totally incompatible with the conclusions I had reached in 1936 attempting to define a reasonable strategy for a repetitive game with a positive mathematical expectation (Allais, 1977, pp. 445–6). Consequently, I viewed the principle of independence as incompatible with the preference for security in the neighbourhood of certainty shown by every subject and which is reflected by the elimination of all strategies implying a non-negligible probability of ruin, and by a preference for security in the neighbourhood of certainty when dealing with sums that are large in relation to the subject's capital (Allais, 1952b, pp. 84–6, 88–90, 92–5; 1977, pp. 451, 466–7, 491–8).

This led me to devise some counter-examples. One of them, formulated in 1952, has become famous as the ‘Allais Paradox’. Today, it is as widespread as its real meaning is generally misunderstood.

This counter-example consists of two questions, the gains considered being expressed in (1952) francs [one million (1952) francs is roughly equivalent to 10,000 (1985) dollars].

- Do you prefer Situation A to Situation B?
- Situation A
  - certainty of receiving 100 million.
- Situation B
  - a 10 per cent chance of winning 500 million,
  - an 89 per cent chance of winning 100 million,
  - a 1 per cent chance of winning nothing.
- Do you prefer Situation C to Situation D?

- Situation C
  - an 11 per cent chance of winning 100 million,
  - an 89 per cent change of winning nothing.
- Situation D
  - a 10 per cent chance of winning 500 million,
  - a 90 per cent chance of winning nothing.

It can be shown that, according to the neo-Bernoullian formulation, the preference $A > B$ should entail the preference $C > D$, and conversely (Allais, 1952b, pp. 88–90; 1977, pp. 533–41).

However, it is observed that for very careful persons, well aware of the probability calculus and considered as rational, and whose capital C is relatively low by comparison with the gains considered, the preference $A > B$ can be observed in parallel to the preference $C < D$. Since the neo-Bernoullians consider the axioms from which they deduce the neo-Bernoullian formulation as evident, they consider this result a paradox.

In 1952, Savage's answers to these two questions contradicted his own axioms. The explanation he gave is somewhat surprising. It boiled down to stating: ‘Since my axioms are totally evident, my answers, which are indeed incompatible with my axioms, are explained by the fact that I did not give the matter enough thought’ (Savage, 1954, pp. 101–103).

Empirical research

After analysing the answers to the 1952 Questionnaire (Allais, 1952d). I found that the rate of violation of the neo-Bernoullian formulation coresponding to the Allais Paradox was approximately 53 per cent (Allais, 1977, p. 474).

This violation example is not an isolated one (Allais, 1977, pp. 636–6, n. 15). There is even one test for which the rate of violation is 100 per cent. It is based on the comparative analysis of, on the one hand, the monetary value x' attributed to a probability of 1/2 of winning a sum between 0.0001 and 1000 million, with a probability of 1/2 of winning nothing at all; and, on the other hand, of the monetary value x'' attributed to a probability p_i between 0.25 and 0.999 of winning 200 million, with a probability $1 - p_i$ of winning nothing at all. The two indexes $B_{1/2}$ and B_{200} deduced from these two series of questions, which according to the neo-Bernoullian formulation should be totally identical up to a linear transformation, in fact are completely different for all the subjects who answered the questions. Such was in particular the case of de Finetti (Allais, 1977, pp. 612–13, 620–31; 1983; pp. 61–2 and 110–11, n. 146).

Much empirical research has been carried out since 1952. It has shown that many subjects who can be viewed as rational may behave in contradiction with the neo-Bernoullian formulation (e.g. MacCrimmon and Larsson, 1975; Allais, 1977, pp. 507–8, pp. 611–54). Confronted with these results, the neo-Bernoullians always explain these violations as ‘anomalies’, ‘errors’, ‘insufficient thought by the subjects’, or ‘ill constructed and inconclusive’ experiments made by incompetent persons, ‘inexperienced in experimental psychology’ (e.g. Amihud, 1974 and 1977; Morgenstern, 1976). But these statements do not hold in the face of the very numerous violations observed by the many

researchers, following different methods and operating in different countries at different times (Allais, 1977, pp. 541–2; 1983, p. 66).

The Allais paradox, a simple illustration of Allais's general theory of random choice

These violations can be explained very simply. Limiting consideration to the mathematical expectation of the B_i involves neglecting the basic element characterizing psychology vis-à-vis risk, namely the distribution of cardinal utility about its mathematical expectation (Allais, 1952b, pp. 51–5, 96–7; 1977, pp. 481–2, 520–23, 550–52; 1983, pp. 30–31), and in particular, when very large sums are involved in comparison with the psychological capital of the subject, the strong dependence between the different eventualities (g_i,•p_i), and the very strong preference for security in the neighbourhood of certainty.

My 1952 inquiry (Allais, 1952d; 1977, pp. 447–9, 451–4, 604–54; 1983, pp. 28 and 41) showed that all the subjects questioned were able to answer questions on the intensity of their preferences for different possible gains, setting aside any consideration of random choices (only a few neo-Bernoullian authors refused to answer these questions) (Allais, 1943, pp. 156–77; 1952b, pp. 43–6; 1977, pp. 460–61, 475–80, 614–17, 632–3). The analysis of the answers made it possible to design a well defined cardinal utility curve, the structure of which is the same for all the subjects up to a linear transformation. It portrays their answers on average remarkably well (Allais, 1984a and 1984c).

This result is all the more significant in that this expression of cardinal utility shows a very striking similarity to the expression for psychophysiological sensation as a function of luminous stimulus, determined by Weber's and Fechner's successors (Allais, 1984c, § 4.3 and Charts III and XXV).

The existence of a cardinal utility $u(C + g)$ being proven and the neo-Bernoullian index $B(C + g)$, if it exists, being defined also up to a linear transformation, it can be shown that the two indexes are necessarily identical up to a linear transformation (Allais, 1952b, pp. 97–8, 103, 128–30; 1977; pp. 465, 483, 604–607; 1983, pp. 29–30; 1985).

As a consequence the neo-Bernoullian formulation reduces to considering the mathematical expectation of cardinal utility alone, neglecting its dispersion about the average. In so doing, it neglects what may be considered as the specific element of risk (Allais, 1952b, pp. 49–56; 1983; pp. 35–41).

In fact the cardinal utility corresponding to a monetary value V of a random prospect should be considered as a function

$$u(C + V) = F[u(C + g_1), \ldots, u(C + g_i), \ldots\ldots, u(C + g_n), p_1, \ldots, p_i, \ldots, p_n] \tag{4}$$

of cardinal utilities u_i corresponding to the different gains g_i. Since utilities u_i are defined up to a linear transformation, it can be shown that (Allais, 1977, pp. 481–3, 550–52, 607–609; 1985, § 12 and 22)

$$u + \Delta = F(u_1 + \Delta, \ldots, u_i + \Delta, \ldots\ldots, u_n + \Delta, p_1, \ldots, p_i, \ldots, p_n)$$
(5)

in which Δ is any constant (property of cardinal isovariation). Consequently it can be shown that relation (4) can be written

$$u(C + V) = \bar{u} + R(\mu_2, \ldots, \mu_l, \ldots, \mu_{2n-1})$$
(6)

in which $\bar{u}$ represents the mathematical expectation of the u_i and the μ_l represent the moments of order l:

$$\mu_l = \sum_i p_i(u_i - \bar{u})^l.$$
(7)

The ratio $\rho = R / \bar{u}$ can be considered as an index of the propensity for risk. For $\rho = 0$, the behaviour is Bernoullian; for $\rho > 0$, there is a propensity for risk; for $\rho < 0$, there is a propensity for security. For a given subject, ρ can be nil, positive or negative, depending on the domain of the field of random choices considered (Allais, 1983, pp. 35–41; 1985).

The mistake made by the proponents of the neo-Bernoullian formulation is to want to impose restrictions on the preference index

$$I = f\,[g_1, \ldots, g_i, \ldots, g_n, p_1, \ldots, p_i, \ldots, p_n]$$
(8)

of any subject other than those corresponding to conditions of rationality, such as the existence of a field of ordered random choice or the axiom of absolute preference. According to this axiom, taking two random prospects $g_i, \bullet p_i$ and g'_i, p_i such that $g_i > g'_i$ for any p_i, the first is obviously preferable to the second (Allais, 1952b, pp. 38–41; 1977, pp. 457–8, 530–35; 1985, § 31.3).

Imposing other restrictions would, in the case of certain goods (A),(B),…,(C), reduce to imposing special restrictions on the preference index I(A,B,…,C) which no author has ever envisaged. In fact, to have a marked preference for security in the neighbourhood of certainty together with a preference for risk far

from certainty is not more irrational than preferring roast beef to chicken (Allais, 1952b, pp. 65–7; 1977, pp. 527–33; 1983, pp. 39–40; 1985, § 31.3).

From The St Petersburg paradox to the Allais paradox

In sum, just as the St Petersburg Paradox led Daniel Bernoulli to replace the principle of maximization of the mathematical expectation of monetary values by the Bernoullian principle of maximization of cardinal utilities, the Allais Paradox leads to adding to the Bernoullian formulation a specific term characterizing the propensity to risk which takes account of the distribution as a whole of cardinal utility (Allais, 1978, pp. 4–7; 1977, pp. 548–52; 1983, pp. 35–42).

Neither the St Petersburg nor the Allais Paradox involves a paradox. Both correspond to basic psychological realities: the non-identity of monetary and psychological values and the importance of the distribution of cardinal utility about its average value.

For nearly forty years the supporters of the neo-Bernoullian formulation have exerted a dogmatic and intolerant, powerful and tyrannical domination over the academic world; only in very recent years has a growing reaction begun to appear. This is not the first example of the opposition of the ‘establishments’ of any kind to scientific progress, nor will it be the last (Allais, 1977, pp. 518–46; 1983, pp. 69–71, 112–14).

The Allais Paradox does not reduce to a mere counter-example of purely anectodal value based on errors of judgement as too many authors seem to think without referring to the general theory of random choice which underlies it. It is fundamentally an illustration of the need to take account not only of the mathematical expectation of cardinal utility, but also of its distribution as a whole about its average, basic elements characterizing the psychology of risk.

See Also

- expected utility hypothesis

Bibliography

Allais, M. 1943. *A la recherche d’une discipline économique*, Première partie: l’économie pure. Ateliers Industria, 920 pp. Second edition under the title *Traité d’économie pur* Paris: Imprimerie Nationale, 1952, 5 vols. (The second edition is identical to the first, apart from the addition of a new introduction, 63 pp.)

Allais, M. 1952a. Fondements d’une théorie positive des choix comportant un risque et critique des postulats et axiomes de l’école Américaine. International Conference on Risk, Centre National de la Recherche Scientifique, May 1952. *Colloques Internationaux XL, Econométrie*, Paris, 1953, 257–332.

Allais, M. 1952b. The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American school. English translation of 1952a. In Allais and Hagen (1979),

27–145.

Allais, M. 1952c. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* 21(4) (1953), 503–46. This paper corresponds to some parts of Allais, 1952a.

Allais, M. 1952d. La psychologie de l'homme rationnel devant le risque – la théorie et l'expérience. *Journal de la Société de Statistique de Paris*, January–March 1953, 47–73.

Allais, M. 1977. The so-called Allais' Paradox and rational decisions under uncertainty. In Allais and Hagen (1979), 437–699.

Allais, M. 1978. Editorial Introduction, Foreword. In Allais and Hagen (1979), 3–11.

Allais, M. 1983. The foundations of the theory of utility and risk. In *Progress in Decision Theory*, ed. O. Hagen and F. Wenstop. Dordrecht: Reidel, 1984, 3–131.

Allais, M. 1984a. L'utilité cardinale et sa détermination – hypothèses, méthodes et résultats empiriques. Memoir presented to the Second International Conference on Foundations of Utility and Risk Theory, Venice, 5–9 June 1984.

Allais, M. 1984b. The cardinal utility and its determination – hypotheses, methods and empirical results. English version of 1984a, in *Theory and Decision*, 1987.

Allais, M. 1984c. Determination of cardinal utility according to an intrinsic invariant model. Abridged version of 1984a in *Recent Developments in the Foundations of Utility and Risk Theory*, ed. L. Daboni et al., Dordrecht: Reidel, 1985, 83–120.

Allais, M. 1985. Three theorems on the theory of cardinal utility and random choice. In *Essays in Honour of Werner Leinfellner*, ed. H. Berghel. Dordrecht: Reidel, 1986.

Allais, M. and Hagen, O., eds. 1979. *Expected Utility Hypotheses and the Allais' Paradox; Contemporary Discussions and Rational Decisions under Uncertainty with Allais' Rejoinder*. Dordrecht: Reidel.

Amihud, Y. 1974. Critical examination of the new foundation of utility. In Allais and Hagen (1979), 149–60.

Amihud, Y. 1977. A reply to Allais. In Allais and Hagen (1979), 185–90.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. Trans. as 'Exposition of a new theory

on the measurement of risk' *Econometrica* 22(1954), 23–36.

de Finetti, B. 1977. A short confirmation of my standpoint. In Allais and Hagen (1979), 161.

Friedman, M. and Savage, J.L. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56, August, 279–304.

MacCrimmon, K. and Larsson, S. 1975. Utility theory: axioms versus paradoxes. In Allais and Hagen (1979), 333–409.

Marschak, J. 1950. Rational behavior, uncertain prospects and measurable utility. *Econometrica* 18(2), 111–41.

Marschak, J. 1951. Why 'should' statisticians and businessmen maximize moral expectation? In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

Marschak, J. 1977. Psychological values, and decision makers. In Allais and Hagen (1979), 163–75.

Morgenstern, O. 1976. Some reflections on utility. In Allais and Hagen (1979), 175–83.

Neumann, J.von. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*. 2nd edn., Princeton: Princeton University Press.

Samuelson, P. 1952. Utility, preference and probability. International Conference on Risk, Centre National de la Recherche Scientifique, Paris, May 1952. *Colloques Internationaux XL, Econométrie*, Paris (1953), 141–50.

Savage, L. 1952. An axiomatization of reasonable behavior in the face of uncertainty. International Conference on Risk, Paris, May 1952. Centre National de la Recherche Scientifique, *Colloques Internationaux XL, Econométrie*, Paris (1953), 29–33.

Savage, L. 1954. *The Foundations of Statistics*. New York: Wiley.

How to cite this article

Allais, Maurice. "Allais paradox." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000074> doi:10.1057/9780230226203.0032

The New Palgrave Dictionary of Economics Online

Allais, Maurice (born 1911)

Bernard Belloc and Michel Moreaux
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Allais, M.; arbitrage; Bernoullian principle; Boiteux, M.; business cycles; capitalistic optimum; choice under risk; Debreu, G.; Desrousseaux, J.; Divisia, F. J. M.; Dupuit, A.-J.-E.; expected utility hypothesis; Friedman, M.; functional analysis; golden rule of accumulation; Hicks, J. R.; hyperinflation; industrial policy; interdependence; intertemporal general equilibrium; intertemporal optimality; Jevons, W.S.; Malinvaud, E.; Marschak, J.; Morgenstern, O.; neoclassical growth theory; optimum population; probability; Samuelson, P.A.; Savage, L. J.; steady state equilibrium; surplus; tâtonnement; uncertainty; von Neumann, J.; Walras, L.

Article

Maurice Allais was born on 31 May 1911, in Paris. Originally a student at the Ecole Polytechnique he moved later to the Ecole Nationale des Mines (ENMP hereafter). He gained the doctorate of engineering of the University of Paris in 1949. He is currently director of Research at the Centre National de la Recherche Scientifique (CNRS) and Professor of Economic Analysis at the ENMP. The CNRS awarded him a gold medal in 1978, the first time this award was given to an economist. He was awarded the Nobel Prize for Economics in 1988.

His initial professional activity led him toward problems of applied economics and regulation. In France, the corps of mining engineers, one of the greatest branches of the civil service, is entrusted with the regulation of mining and energy and is very influential in the definition and control of public industrial policy. In some sense Allais' theoretical works are an attempt to find rational public economic public decisions. The title of his first book, *A la recherche d'une discipline économique. Première Partie: l'économie pure* (1943) is very significant in this respect. One feels in Allais' thought a deep reluctance to accept any theory which cannot be made operative (1978a). Thus a very important part of his activity, which will not be surveyed here, is devoted to applied economic studies, always, directly supported by a theoretical analysis (see 1954; 1956a; 1977). In the brilliant tradition of Dupuit, Colson and Divisia this aspect of Allais' work has been essential for the development of the school of French economist

engineers. Allais educated several generations of researchers and public managers: M. Boiteux, G. Debreu and E. Malinvaud were among his students.

In the line of descent from Walras, Fisher and Pareto, Allais' theoretical contributions are basic in four fields: general equilibrium and optimal allocation of resources ('*rendement social*' or 'efficacité maximale' in Allais' terminology), capital and growth, money and business cycle, risky choices.

Allais is primarily a theorist of interdependence and optimum. It is impressive to observe that the research programme defined at the start in Allais (1943) has been almost wholly fulfilled, even though some of the initial basic assumptions have been drastically revised. When published in 1943, Allais' book was one of the most complete reports on general equilibrium and optimum theories, comparable to Hicks's *Value and Capital* and Samuelson's *Foundations of Economic Analysis*. Let us emphasize its differences. Allais gives the earliest formalization of an intertemporal general equilibrium and, in particular, all the arbitrage conditions between capital goods and land are made explicit. Then, the first results on global stability of Walrasian *tâtonnement* are proved by means of Lyapunov's second method under assumptions equivalent to gross substitutability (see Negishi, *Econometrica* (1962), for a report in English). The book also contains a complete account of optimum theory in terms of distributable surpluses and a precise and correct statement of the two welfare theorems. Finally, Allais outlined a theory of optimum population. Later, Allais' opinion on the relevance of the Walrasian model changed markedly (1967b; 1968; 1971; 1981). He would now define a state of general equilibrium as a position in which no distributable surplus can be obtained, and describes the whole motion of the system as governed by the search for such surpluses. In some way this new view is a true merging of general equilibrium and optimum theories (1981).

His main contributions to capital and growth theory are expressed in Allais (1947; 1960; 1962). First, and sometimes with a lead of 15 years, he found most of the results of so-called neoclassical theory of growth, including the famous golden rule of accumulation. Allais worked out a complete theory of capitalistic processes with a rigorous formalization of the concept of characteristic function first proposed by Jevons in 1871, by which is meant the sequence of past expenditures on primary inputs which have generated the present national income. The systematic use of this concept allowed Allais to build up a theory of economic growth. But its use has been even more fruitful in the analysis of capitalistic efficiency. Allais proved in 1947 that, in a stationary state, a zero rate of interest maximizes real income. This is the first version of the golden rule of accumulation obtained by Phelps some 14 years later. In 1962 Allais widened this result and demonstrated that in steady states a capitalistic optimum is attained when the rate of interest is equal to the rate of growth (it is to be noted that Allais himself acknowledges that J. Desrousseaux had been the first to get this result in 1959, in a non-published paper). Thus Allais was completing his theory of optimal allocation of resources with a theory of capitalistic optimum.

To analyse intertemporal optimality, he assumes that each agent has preferences, on present and future consumption, possibly different in different periods. Hence it becomes possible to consider the psychological evolution of an individual over his lifetime, unlike the usual approach. In other respects Allais has been very careful to test the explicative power of his capitalistic optimum theory, by comparing the growth processes in different countries and trying to evaluate in every case the gap between the capitalistic optimum and the real state of accumulation.

Allais must be also considered as a major actor in the revival of the quantity theory of money (1956b; 1956c; 1965a; 1966; 1969; 1970; 1972; 1974). The reduced form of the model explaining the dynamics

of national monetary expenditure is very similar to Cagan's contemporary formulation. But Allais claims that his model has very different foundations because it is supported by an alleged psychological law of the perception of time. The solutions of the integro-differential equation describing the evolution of income are shown to have three limit cycles, depending upon initial conditions. It is then possible to explain local stability of a steady state equilibrium, business cycles and hyperinflation state with the same basic model.

The last aspect of Allais' work concerns choice under risk (1953b; 1953c; 1979). As usual, Allais' approach is both theoretical and empirical. He builds up his analysis on the basis of experimental psychological tests conducted in 1952 (see Allais, 1953c, for a partial statement). For Allais the theory of choice under risk went, historically, through four steps. At first it was assumed that the mathematical expectation of the monetary gain was the natural evaluation of a lottery. Then the mathematical expectation of the gain in utility was used. The third step then considered subjective probabilities. The American school (Friedman, Marschak, von Neumann, Morgenstern, Samuelson and Savage) takes into account only these three steps. So Allais claims that a fourth step must be reached: the value of a lottery is a functional depending upon the probability density parameterized by the gains. In effect the expected utility hypothesis implies a special such functional, so this last step seems very natural. Allais systematically criticizes the axioms on which the Bernoullian principle is based. According to him such axioms cannot help to define rationality in an uncertain environment. Through convincing examples he specially refutes Savage's independence and Samuelson's substitutability axioms. The major argument is in short that in the neighbourhood of certainty, a rational agent will prefer absolute safety. Then Allais proposes an alternative definition of rationality in risky situations: the set of choices must be ordered, an absolute preference axiom must be satisfied (that is, if a lottery gives in every case larger gains than another, then any agent will prefer the first one) and only objective probabilities must be considered. The first two axioms seem quite reasonable and it is difficult, according to Allais, to disprove the last one. But it is clear that a decision rule following the Bernoullian principle cannot be deduced from these three axioms. They imply the use of a functional of more general form than the mathematical expectation of the psychological evaluation of gains. In fact Allais argues that the Bernoullian principle only takes into account the dispersion of the gains whereas the dispersion of their psychological values is pertinent.

Finally, Allais applies his theory of behaviour under uncertainty to a general equilibrium model (1953a). He demonstrates this through an example where a competitive allocation of risks leads to an optimal allocation of resources, and where such an allocation can be obtained as a competitive equilibrium with an appropriate redistribution of initial endowments.

See Also

- expected utility hypothesis

Selected works

For a complete record of Allais's work on the period 1943–78 and an analysis by Allais himself, see Allais (1978a; 1978b; 1978c).

1943. *A la recherche d'une discipline économique. Première partie: I'économie pure*. Paris: Ateliers Industria.

1947. *Economie et intérêt*. Paris: Imprimerie Nationale.

1953a. L'extension des théories de l'équilibre économique général et du rendement social au cas du risque. *Econometrica* 21, 269–90.

1953b. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'Ecole Américaine. *Econometrica* 21, 503–46. Trans. in *Expected Utility Hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen. Dordrecht: Reidel, 1979.

1953c. La psychologie de l'homme rationnel devant le risque. La théorie et l'expérience. *Journal de la Société de Statistique de Paris*, January, 47–72.

1954. Méthode d'évaluation des perspectives économiques de la recherche minière sur des grands espaces. Application au Sahara Algérien. Paris, mimeo. Trans. in *Management Science* 3(4) (1957), 285–347.

1956a. *La gestion des houillères nationalisées et la théorie économique*. Paris: Imprimerie Nationale.

1956b. Explication des cycles économiques par un modèle non linéaire à régulation retardée. *Metroeconomica* 8, 4–83.

1956c. Explication des cycles économiques par un modèle non linéaire à retardée. Mémoire complémentaire. In *Les modèles dynamiques en économétrie*, Collection des Colloques Internationaux du CNRS, 52. Paris: CNRS.

1960. L'influence du coefficient capitalistique sur le revenu réel par tête. *International Statistical Institute Bulletin* 38(2), 3–27.

1962. The influence of the capital-output ratio on real national income. *Econometrica* 30, 700–28.

1965a. Reformulation de la theórie quantitative de la monnaie. *Bulletin SEDEIS*, No. 928, Supplement.

1965b. The role of capital in economic development. In *The Economic Approach to Economic Development*. Amsterdam: North-Holland.

1966. A restatement of the quantity theory of money. *American Economic Review* 56, 1123–57.

1967a. Some analytical and practical aspects of the theory of capital. In *Activity Analysis in the Theory*

of Growth and Planning, ed. E. Malinvaud and M. Bacharach. London: Macmillan.

1967b. *Les fondements du calcul économique*, vol. 1. Paris: Ecole Nationale Supérieure des Mines (mimeo).

1968. *Les fondements du calcul économique*, vols. 2 and 3. Paris: Ecole Nationale Supérieure des Mines (mimeo).

1969. Growth and inflation. *Journal of Money, Credit and Banking* 1, 355–426.

1970. A reply to Michael R. Darby's comment on Allais' restatement of the quantity theory. *American Economic Review* 60, 447–56.

1971. Les théories de l'équilibre économique général et de l'efficacité économique. Imapasses récentes et nouvelles perspectives. *Revue d'Economie Politique*, May. Trans. in *Equilibrium and Didsequilibrium in Economic Theory*, ed. G. Schwödiauer. Dordrecht: Reidel.

1972. Forgetfulness and interest. *Journal of Money, Credit and Banking* 4(1), Part I, 40–73.

1974. The psychological rate of interest. *Journal of Money, Credit and Banking* 6, 285–331.

1975. The hereditary and relativisitc formulation of the demand for money. Circular reasoning or a real structural relation? (A reply to Scadding's Note.) *American Economic Review* 65, 454–64.

1977. *L'impôt sur le capital et la réforme monétaire*. Paris: Hermann.

1978a. *Contributions à la science économique. Vue d'ensemble 1943–1978*. Paris: Centre d'Analyse Economique.

1978b. *Principaux ouvrages et mémoires 1943–1978*. Paris: Centre d'Analyse Economique.

1978c. *Titres et travaux scientifiques*. Paris: Centre d'Analyse Economique.

1979. The so-called Allais Paradox and rational decisions under uncertainty. In *Expected Utility Hypotheses and the Allais Paradox*, ed. M. Allais and T. Hagen. Dordrecht: Reidel.

1981. La théorie générale des surplus. Vol. II, Modèle économie et sociétés. *Cahiers de l'ISMEA*, Series EM No. 9.

How to cite this article

Belloc, Bernard and Michel Moreaux. "Allais, Maurice (born 1911)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000075>
doi:10.1057/9780230226203.0031

The New Palgrave Dictionary of Economics Online

Allen, Roy George Douglas (1906 – 1983)

J.R.N. Stone
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Allen, R. D. G.; Bowley, A. L.; Econometric Society; econometrics; Hicks, J. R.; index numbers; mathematics and economics; national accounting; positive economics; Royal Statistical Society; statistics and economics; utility; value

Article

Allen was born on 3 June 1906 at Stoke-on-Trent, and died on 29 September 1983 at Southwold. He was knighted in 1966 and made a Fellow of the British Academy in 1952. He was educated at the Royal Grammar School, Worcester, and Sidney Sussex College, Cambridge. From 1928 he was assistant, then lecturer, then reader in economic statistics at the London School of Economics, becoming professor of statistics in 1944 and emeritus professor in 1973.

During the war, he was a statistician in H.M. Treasury from 1939 to 1941; from 1941 to 1942 he was Director of Records and Statistics for the British Supply Council in Washington, and from 1942 to 1945, he became British Director of Research and Statistics for the Combined Production and Resources Board in Washington. His other principal activities were as statistical adviser for H.M. Treasury (1947–8); member of the Air Transport Licensing Board (1960–72); and member of the Civil Aviation Authority (1972–3). He was President of the Econometric Society in 1951 and President of the Royal Statistical Society in 1969–70. He was also consultant to many international and professional organizations.

Allen was an economic statistician, mathematical economist and econometrician of exceptional competence and breadth of knowledge. His early and most original research, carried out in part with J.R. Hicks and A.L. Bowley, was on the theory of value, utility and consumers' behaviour: for example, Hicks and Allen (1934), Allen (1935), and Allen and Bowley (1935), the last an outstanding work on the econometrics of family budgets.

In the late 1930s he embarked on a series of successful textbooks based on his lectures. His *Mathematical Analysis for Economists* (1938) was intended to help students of economics whose training in mathematics was typically much less thorough than it is now. After the war, in addition to

numerous papers on economic and statistical topics, including one reflecting his wartime work in Washington (Allen, 1946), and a compilation of papers on international trade statistics (Allen and Ely, 1953), he continued the good work begun in 1938 with a succession of books on macroeconomics and the mathematical and statistical tools required in its study. Thus *Statistics for Economists* (1949) is an introduction to statistical methods in their application to economic material; *Mathematical Economics* (1956) is a text on economic theory, written in mathematical terms, which takes account of the growth of econometrics and the use of increasingly sophisticated mathematics by economists; *Basic Mathematics* (1962) provides a general introduction to mathematical ideas, applicable in both the natural and the social sciences; *Macro-Economic Theory* (1967) treats deterministic models from a positive rather than an optimizing or policy-oriented point of view; his 1975 work deals comprehensively with the design, construction and use of index numbers, paying full attention to both the economic and the statistical aspects of the subject; his last book (1980) is an introduction to national accounting, concentrating on the main aggregates at current and constant prices and illustrated by means of recent British official estimates.

Allen was an assiduous disseminator of ideas. His textbooks were translated into many languages and he continued to lecture until shortly before his death. As head of the Statistics Department of the LSE he was instrumental, with the help of M.G. Kendall, in expanding it from a staff of five in 1944 to one of 28, of whom seven were professors.

Selected works

1934. (With J.R. Hicks.) A reconsideration of the theory of value: parts I and II. *Economica*, February, 52–76; May, 196–219.

1935. A note on the determinateness of the utility function. *Review of Economic Studies* 2, February, 155–8.

1935. (With A.L. Bowley.) *Family Expenditure*. London: P.S. King.

1938. *Mathematical Analysis for Economists*. London: Macmillan.

1946. Mutual aid between the US and the British Empire, 1941–45. *Journal of the Royal Statistical Society* 109, 243–71.

1949. *Statistics for Economists*. London: Hutchinson.

1953. (With J.E. Ely, eds.) *International Trade Statistics*. London: Chapman and Hall.

1956. *Mathematical Economics*. London: Macmillan.

1962. *Basic Mathematics*. London: Macmillan.

1967. *Macro-Economic Theory*. London: Macmillan.

1975. *Index Numbers in Theory and Practice*. London: Macmillan.

1980. *An Introduction to National Accounts Statistics*. London: Macmillan.

Bibliography

Cairncross, A. 1985. Roy Allen, 1906–1983. In *Proceedings of the British Academy* 70, 379–85.

Grebenik, E. 1984. Roy George Douglas Allen, 1906–1983. *Journal of the Royal Statistical Society* 147, 706–7.

How to cite this article

Stone, J. R. N. "Allen, Roy George Douglas (1906–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000077> doi:10.1057/9780230226203.0033

The New Palgrave Dictionary of Economics Online

altruism in experiments

James Andreoni William T. Harbaugh and Lise Vesterlund

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We call an act altruistic when it is a sacrifice that benefits others. We discuss how experiments have demonstrated that altruistic choices appear to follow the same regularity conditions as those assumed for private goods. In particular they vary rationally in response to changes in prices and circumstances. We show how experiments have distinguished between different economic models of how concern for others enters utility functions, and have explored the implications of those models for charitable giving, labour markets, and trust. We also discuss the experimental evidence for differences in altruism by gender, and work on altruism's cultural, developmental, and neural foundations.

Keywords

altruism; altruism in experiments; centipede game; cooperation; crowding out; dictator game; efficiency wages; experiments; gift exchange; moral hazard game; Prisoner's Dilemma; public goods games; Rawls, J.; reciprocity; repeated games; reputation; revealed preference; trust game; ulterior motives; ultimatum game; utilitarianism; warm-glow

Article

Unlike experiments on markets or mechanisms, experiments on altruism are about an individual motive or intention. This raises serious obstacles for research. How do we define an altruistic act, and how do we know altruism when we see it?

The philosopher Thomas Nagel provides this definition of altruism: ‘By altruism I mean not abject self-sacrifice, but merely a willingness to act in the consideration of the interests of other persons, without the need of ulterior motives’ (1970, p. 79). Notice that there are two parts to this definition. First, the act must be in the consideration of others. It may or may not imply sacrifice on one's own part, but it does require that the consequences for someone else affect one's own choice. The second aspect is that one does not need ‘ulterior motives’ rooted in selfishness to explain altruistic behaviours. Of course, ulterior

motives may exist alongside altruism, but they cannot be the only motives.

If this is our definition of altruism, then how do we know altruism when we see it? The answer, unfortunately, is necessarily a negative one – we only know when we do not see it. Altruism is part of the behaviour that you cannot capture with a specifically defined ulterior motive. Experimental investigation of altruism is thus focused around eliminating any possible ulterior motives rooted in selfishness. One of the central motives that potentially confounds altruism is the warm-glow of giving, that is, the utility one gets simply from the act of giving *without* any concern for the interests of others (Andreoni, 1989; 1990). While it is possible that warm-glow exists apart from altruism, it seems most likely that the two are complements – the stronger your desire to act unselfishly, the greater the personal satisfaction from doing so. Indeed, the two may be inextricably linked. Having a personal identity as an altruist may necessarily precede altruistic acts, and maintaining that identity can only come from actually being generous.

In what follows we will highlight the main experimental evidence regarding choices made in the interests of others, and the systematic attempts in the literature to rule out ulterior motives for these choices. Since these serious and repeated attempts to rule out ulterior motives have not been totally successful, the experimental evidence, like Thomas Nagel, favours the possibility of altruism.

Laboratory experiments with evidence of altruism

In describing the games below, we adopt the convention of using Nash equilibrium to refer to the prediction that holds if all subjects are rational money-maximizers.

Prisoner's Dilemma

There have been thousands of studies using Prisoner's Dilemma (PD) games in the psychology and political science literatures, all exploring the stubborn nature of cooperation (Kelley and Stanelski, 1970). Roth and Murnigham (1978) explored PD games under paid incentives and with a number of different payoff conditions. Their study confirmed to economists that cooperation is robust.

Sceptics noted, however, that cooperation need not be caused by altruism. First, inexperience and initial confusion may cause subjects to cooperate. Second, subjects in a finitely repeated version of the game may cooperate if they each believe there is a chance someone actually is altruistic. Behaviourally this ‘sequential equilibrium reputation hypothesis’ (Kreps et al., 1982) does not actually require subjects to be altruistic, but only that they believe that they are sufficiently likely to encounter such a person. Andreoni and Miller (1993) explore these two factors by asking subjects to play 20 separate ten-period repeated PD games. A control treatment had subjects constantly changing partners, thus unable to build reputations. They find significant evidence for reputations, but that these alone cannot explain the level of cooperation, especially at the end of the experiment. Rather, they estimate that about 20 per cent of subjects actually need to be altruistic to support the equilibrium findings. This finding is corroborated in other repeated games, such as Camerer and Weigelt's (1988) moral hazard game, McKelvey and Palfrey's (1992) centipede game, and in a two-period PD of Andreoni and Samuelson (2006).

Public goods

Linear public goods games have incentives that make them resemble a many-person PD game. Individuals have an endowment *m* which they each must allocate between themselves and a public account. Each of the *n* members of the group earns α for each dollar allocated to the public account. By design, $0<\alpha<1$, so giving nothing is a dominant strategy, but $\alpha n>1$, so giving *m* is Pareto efficient. The results of these games are that average giving is significantly above zero, even as we change *n*, *m* and α (Isaac and Walker, 1988; Isaac, Walker and Williams; 1994) and whether the play is with the same group of 'partners' or with randomly changing groups of 'strangers' (Andreoni, 1988). Hence, reputations play little role in public goods games (Andreoni and Croson, 2008; Palfrey and Prisbrey, 1996).

In his review of this literature, Ledyard (1995) notes that, with a dominant strategy of giving zero, any error or variance in the data could mistakenly be viewed as altruism. Thus, to determine what drives giving one needs to confirm that subjects understand the dominant strategy but choose to give anyway. Andreoni (1995) develops a design to separate 'kindness' from 'confusion' in linear public goods games. Rather than paying subjects for their absolute performance, in one treatment he paid subjects by their relative performance. Converting subjects' ranks into their payoffs converts a positive-sum game to a zero-sum game. It follows that even altruists have no incentive to cooperate when paid by rank (that is, under the usual definition of altruism where people love themselves at least as much as they love others). Cooperation by subjects in the treatment group, therefore, provides a measure of confusion. Andreoni finds that both kindness and confusion are significant, and about half all cooperation in public goods games is from people who understand free riding but choose to give anyway.

To establish that giving is deliberate, however, does not necessarily mean it is based in altruism; it could, instead, be from warm-glow. Two papers, using similar experimental designs but different data analysis methods, explore this question by separating the marginal net return that a gift to the public good has for the giver and for the recipient. The 'internal return' experienced by the giver should affect warm-glow and altruism, but the 'external return' received by the others affects only altruism. Palfrey and Prisbrey (1997) find that warm-glow dominates altruism, while Goeree, Holt and Laury (2002) find mostly altruism. Combining this evidence, it appears that both motives are likely to be significant.

Another way to test for the presence of altruism and warm-glow is to choose a manipulation that would have different predictions in the two regimes. Andreoni (1993) looks at the complete crowding out hypothesis, which states that a lump-sum tax, used to increase government spending on a public good, will reduce an altruist's voluntary contributions by the amount of the tax. He employs a public goods game with an interior Nash equilibrium. Suppose subjects care only about the payoffs of other subjects (altruism). Then if we force subjects to make a minimum contribution below the Nash equilibrium, this should simply crowd out their chosen gift, leaving the total gift unchanged. If they get utility from the act of giving (warm-glow), by contrast, crowding out should be incomplete. Andreoni finds crowding at 85 per cent, which is significantly different from both zero and 100 percent. This confirms the findings from the last paragraph; both warm-glow and altruism are evident in experiments on public goods. Similar findings are presented in Bolton and Katok (1998) and Eckel, Grossman and Johnston (2005).

Dictator games

This line of research began with the ultimatum game, where a proposer makes an offer on the split of a sum of money. If the responder accepts, the offer is implemented, while if she rejects both sides get nothing. Guth, Schmittberger and Schwarze (1982) find that proposers strike fair deals and leave money on the table. Is this altruism, or just fear of rejection? To answer this question Forsythe et al. (1994) also examine behaviour in a dictator game that cuts out the second stage, leaving selfish proposers free to keep the whole pie for themselves, and leaving altruists unconstrained to give a little or a lot. While keeping the entire endowment is the modal choice in the dictator game, a significant fraction of people give money away. On average, people share about 25 per cent of their endowment. This seems to indicate significant altruism.

Again, researchers have explored numerous non-altruistic explanations. One is that, while the dictator's identity is unknown to the recipient, it is not unknown to the researcher. This lack of ‘social distance’ could cause the selfish but self-conscious subjects to give when they would prefer not to. Hoffman et al. (1994) take elaborate steps to increase the anonymity and confidentiality of the subjects so that even the researcher cannot know their choices for sure. They find that this decreases giving to about 10 per cent of endowments. However, this ‘double anonymous’ methodology creates problems of its own. Bolton, Katok and Zwick (1998) argue that greater anonymity makes the participants sceptical about whether the transfers will be carried out. Bohnet and Frey (1999) find that reducing the social distance increases equal splits greatly, but in their anonymous treatments giving again averages 25 per cent (see also Rege and Telle, 2004).

Andreoni and Miller (2002) take a different approach. They note that, if altruism is a deliberate choice, then it should follow the neoclassical principles of revealed preference. They gave subjects a menu of several dictator ‘budgets’, each with different ‘incomes’ and different ‘prices’ of transferring this income to another anonymous subject. By checking choices against the generalized axiom of revealed preference, they show that indeed most subjects are rational altruists, that is, they have consistent and well-behaved preferences for altruistic giving in a dictator game. They also show substantial heterogeneity across subjects, with preferences ranging from utilitarian (maximizing total payments to both subjects) to Rawlsian (equalizing payments to both subjects). Interestingly, men and women are on average equally altruistic in this study, but vary significantly in response to price. Andreoni and Vesterlund (2001) show that men are more likely to be utilitarian, and women are more likely to be Rawlsian. This implies that men are significantly more generous when giving is cheap (that is, it costs the giver less than one to give one), but women are significantly more altruistic when giving is expensive (costs greater than or equal one to give one). Which is the fairer sex, therefore, depends on the price of giving (see also Eckel and Grossman, 1998, on dictator games when the price is one).

Trust games and gift exchange

When someone buys a loaf of bread from a baker, there is a moment when one party has both the bread and the money and the incentive to take both. Why don't they? Similarly, why are some car mechanics truthful, and why do some workers put in an honest effort even when they are not monitored? These questions have been studied under names of trust games and gift exchange.

In the trust game, two players are endowed with M each. A sender chooses to pass x to a receiver. A

receiver receives kx, where $k>1$. The receiver then chooses a y to pass back to the sender. Senders earn $M-x+y$, while receivers earn $M+kx-y$. Since $y=0$ is a dominant strategy for receivers, $x=0$ is the subgame perfect equilibrium strategy for senders. That is, since the baker keeps both the bread and the money, no exchange is attempted. Despite this dire prediction, x and y are often positive, and y is typically increasing in x. While there is tremendous variance, the average y is often slightly below the average x (Berg, Dickhaut and McCabe, 1995).

The gift exchange game is a nonlinear version of the trust game above. Fehr, Kirchsteiger and Riedl (1993) adapted the Akerlof (1982) labour market model of efficiency wages. Some subjects play the roles of firms and offer labour contracts to workers. The contracts stipulate a wage and an expected effort level of workers. Since effort is costly and unobservable, it should be minimal. The subjects playing the role of firms should expect low effort, and offer low wages. However, in the experiment wages are high and effort rises with the wage offer, just as Akerlof predicted.

Trust and gift exchange games are often used to argue for the importance of reciprocity. Reciprocity is, however, an ulterior motive – giving in order to either generate or relieve an obligation is not altruism by the definition in our introduction. How much of the exchange can be attributed to altruism alone? Cox (2004) separates these motives by comparing senders in a trust game with those in a dictator game. As dictators have no ulterior motive of generating an obligation, their behaviour can be used to estimate the altruism of senders. For receivers he uses a control group whose x is determined at random by the experimenter. These receivers have no obligation to the sender, thus their transfers serve as a measure of the receivers' altruism. Cox finds that 60 per cent of an average sender's x and 42 per cent of the average receiver's y is motivated by altruism. Thus, while reciprocity is clearly present, altruism is not replaced in this exchange (see also Charness and Haruvy, 2002; Gneezy, Guth and Verboven, 2000).

While some have criticized whether gift exchange in the laboratory is robust to small changes in parameters and presentation (Charness, Frechette and Kagel, 2006), others have challenged gift exchange in the field. List (2006) looks for gift exchange on the trading floor of a sports card market. He conducts a series of experiments that move incrementally from a standard laboratory game with a neutral presentation to actual exchanges on the floor. While he finds that gift exchange (higher-quality product in return for higher price) is not totally extinguished in the actual market, he also finds that reputation is far more important in determining the quality provided by sellers. Gneezy and List (2006) follow up with a labour market experiment. They recruited students to do a one-day job working in a library. The treatment group was told, unexpectedly, that their wage would be 167 per cent of the agreed wage. These subjects were significantly more productive in the first 90•minutes of work than the control subjects. However, after a one-hour lunch break, there was no difference between the productivity of treatment and control. They conclude that gift exchange in actual labour markets may have no long-term effects.

Conclusion

There is ample consistent evidence of altruism in experiments. This follows both from studies that have taken great effort to remove any ulterior motives, as well as studies that provide manipulations that should influence altruism. While the existence and importance of altruism seem well established in the laboratory, many questions that could help us understand and amplify altruism remain unanswered.

First, where do altruistic preferences come from? One notion is that they come from culture. Evidence of this is suggested by differences in behaviour in experiments in different countries (Roth et al., 1991; Henrich et al., 2001). Another notion is that they are acquired as part of psychological development and socialization, as seen in economic experiments using children as subjects (Harbaugh and Krause, 2000). A third possibility for altruism is that we are innately wired to care. Harbaugh, Mayr and Burghart (2007) use fMRI to show that neural activation in the ventral striatum is very similar when money goes to the subject and when it goes to a charity, and that the relative activations actually predict who will give. Tankersley, Stowe and Huettel (2007) show that posterior superior temporal sulcus activation is higher for people who report more helping behaviour outside the lab.

Second, is altruism significant outside the laboratory? The laboratory is, after all, a unique environment. Field experiments on fundraising, such as List and Lucking-Reiley (2002), show the potential of this method for finding good evidence of altruism outside the laboratory, but without giving up all experimental control.

Finally, how does altruism combine with other ulterior motives? Are warm-glow and altruism inextricably linked, and can we use mechanisms that act on warm-glow to amplify altruism and overcome free riding? Does voting to force everyone to provide a public good provide a warm-glow benefit to the voters? Economic experiments may be a productive method for answering these questions, and for using the knowledge of altruism that results to improve the institutions within which altruist economic agents interact.

See Also

- altruism, history of the concept
- charitable giving
- experimental economics
- experimental economics, history of
- public goods experiments
- public goods

Bibliography

Akerlof, G.A. 1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97, 543–69.

Andreoni, J. 1988. Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics* 37, 291–304.

Andreoni, J. 1989. Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy* 97, 1447–58.

Andreoni, J. 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. *Economic Journal* 100, 464–77.

Andreoni, J. 1993. An experimental test of the public-goods crowding-out hypothesis. *American Economic Review* 83, 1317–27.

Andreoni, J. 1995. Cooperation in public-goods experiments: kindness or confusion? *American Economic Review* 85, 891–904.

Andreoni, J. and Croson, R. 2008. Partners versus strangers: the effect of random rematching in public goods experiments. In *Handbook of Experimental Economics Results*, ed. V. Smith and C. Plott. New York: North-Holland, forthcoming.

Andreoni, J. and Miller, J.H. 1993. Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. *Economic Journal* 103, 570–85.

Andreoni, J. and Miller, J.H. 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. *Econometrica* 70, 737–53.

Andreoni, J. and Samuelson, L. 2006. Building rational cooperation. *Journal of Economic Theory* 127, 117–54.

Andreoni, J. and Vesterlund, L. 2001. Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics* 116, 293–312.

Berg, J., Dickhaut, J. and McCabe, K. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10, 122–42.

Bohnet, I. and Frey, B.S. 1999. Social distance and other-regarding behavior in dictator games: comment. *American Economic Review* 89, 335–9.

Bolton, G.E. and Katok, E. 1998. An experimental test of the crowding out hypothesis: the nature of beneficent behavior. *Journal of Economic Behavior and Organization* 37, 315–31.

Bolton, G.E., Katok, E. and Zwick, R. 1998. Dictator game giving: rules of fairness versus acts of kindness. *International Journal of Game Theory* 27, 269–99.

Camerer, C. and Weigelt, K. 1988. Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56, 1–36.

Charness, G., Frechette, G.R. and Kagel, J.H. 2004. How robust is laboratory gift exchange?

Experimental Economics 7, 189–205.

Charness, G. and Haruvy, E. 2002. Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach. *Games and Economic Behavior* 40, 203–31.

Cox, J.C. 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46, 260–81.

Eckel, C.C. and Grossman, P.J. 1998. Are women less selfish than men? Evidence from dictator experiments. *Economic Journal* 108, 726–35.

Eckel, C.C., Grossman, P.J. and Johnston, R.M. 2005. An experimental test of the crowding out hypothesis. *Journal of Public Economics* 89, 1543–60.

Fehr, E., Kirchsteiger, G. and Riedl, A. 1993. Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108, 437–59.

Forsythe, R., Horowitz, J.L., Savin, N.E. and Sefton, M. 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior* 6, 347–69.

Gneezy, U. Guth, W. and Verboven, F. 2000. Presents or investments? An experimental analysis. *Journal of Economic Psychology* 21, 481–93.

Gneezy, U. and List, J.A. 2006. Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments. *Econometrica* 74, 1365–84.

Goeree, J.K., Holt, C.A. and Laury, S.K. 2002. Private costs and public benefits: unraveling the effects of altruism and noisy behavior. *Journal of Public Economics* 83, 255–76.

Guth, W. Schmittberger, R. and Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3, 367–88.

Harbaugh, W.T., Mayr, U. and Burghart, D. 2007. Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, forthcoming.

Harbaugh, W.T. and Krause, K. 2000. Children's altruism in public good and dictator experiments. *Economic Inquiry* 38, 95–109.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. and McElreath, R. 2001. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review* 91, 73–8.

Hoffman, E., McCabe, K., Shachat, K. and Smith, V.L. 1994. Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior* 7, 346–80.

Isaac, M.R. and Walker, J.M. 1988. Group size effects in public goods provision: the voluntary contributions mechanism. *Quarterly Journal of Economics* 103, 179–99.

Isaac, M.R., Walker, J.M. and Williams, A.W. 1994. Group size and the voluntary provision of public goods: experimental evidence utilizing large groups. *Journal of Public Economics* 54, 1–36.

Kelley, H.H. and Stanelski, A.J. 1970. Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology* 16, 66–91.

Kreps, D., Milgorm, P., Roberts, J. and Wilson, R. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory* 27, 245–52.

Ledyard, J.O. 1995. Public goods: a survey of experimental research. In *The Handbook of Experimental Economics*, ed. J.H. Kagel and A.E. Roth. Princeton: Princeton University Press.

List, J.A. 2006. The behavioralist meets the market: measuring social preferences and reputation effects in actual transactions. *Journal of Political-Economy* 114, 1–37.

List, J.A. and Lucking-Reiley, D. 2002. The effects of seed money and refunds on charitable giving: experimental evidence from a university capital campaign. *Journal of Political Economy* 110, 215–33.

McKelvey, R.D. and Palfrey, T.R. 1992. An experimental study of the centipede game. *Econometrica* 60, 803–36.

Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Clarendon Press.

Palfrey, T.R. and Prisbrey, J.E. 1996. Altruism, reputation and noise in linear public goods experiments. *Journal of Public Economics* 61, 409–27.

Palfrey, T.R. and Prisbrey, J.E. 1997. Anomalous behavior in public goods experiments: how much and why? *American Economic Review* 87, 829–46.

Rege, M. and Telle, K. 2004. The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics* 88, 1625–44.

Roth, A.E. and Murnigham, J.K. 1978. Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology* 17, 189–98.

Roth, A.E. Prasnikar, V. Okuno-Fujiwara, M. and Zamir, S. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review* 81, 1068–95.

Tankersley, D. Stowe, C.J. and Huettel, S.A. 2007. Altruism is associated with an increased neural response to agency. *Nature Neuroscience* 10, 150–1.

How to cite this article

Andreoni, James, William T. Harbaugh and Lise Vesterlund. "altruism in experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000240> doi:10.1057/9780230226203.0035

The New Palgrave Dictionary of Economics Online

altruism, history of the concept

Philippe Fontaine

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes the incorporation from the early 1960s of seemingly unselfish behaviour into economics. Faced with the problem of accounting for such behaviour in a discipline that often relies on the selfishness assumption, some economists used the notion of sympathetic preferences within the self-interest model, whereas others tried instead to supplement that model with an ethically inspired model. It is unclear that in investigating seemingly unselfish behaviour, economists have gained a better understanding of its actual motivations, but in the process they have been led to take more seriously other conceptions of human being than economic man.

Keywords

altruism; Becker, G.; biology and economics; Boulding, K.E.; Buchanan, J.M.; charitable giving; choice; commitment; economic man; ethics and economics; family decision making; family economics; free-rider problem; *homo economicus* ; Institute of Economic Affairs; non-economic behaviour; Phelps, E.; philanthropy; preference; reciprocity; Ricardian equivalence theorem; Robertson, D.; rotten kid theorem; self-interest; Sen, A.; sociobiology; Titmuss, R.M.; utility interdependence; Vickrey, W.S.; warm-glow hypothesis

Article

Dennis Robertson once asked: ‘What does the economist economize?’ (1955, p. 154). His answer was: ‘[T]hat scarce resource Love – which we know, just as well as anybody else, to be the most precious thing in the world’. He meant that a better understanding of the economy had the happy consequence of allowing people to conduct their business without having to rely excessively on social virtues. For upholders of economic man, that was certainly a good justification for doing without ‘love’. And had it not been for a study of philanthropy, conducted in the late 1950s, they might well have continued to ignore ‘love of human kind in general’, as dictionaries usually define it.

The reintroduction of what is regarded today as ‘altruism’ into contemporary economics, following Edgeworth's (1881, p. 53n) first modern formulation in the late 19th century, came out of this effort to understand philanthropy, and not, as conventional wisdom suggests, from the publication of Gary Becker's (1974) ‘A Theory of Social Interactions’, which was but one tardy sequel to it. In writing a history of recent work on unselfishness, therefore, it is crucial that Becker's two chief results in that article, namely, the invariance proposition and the ‘ “rotten kid” theorem’, do not mask the sheer diversity of research before the mid-1970s nor the tensions that persisted afterwards. In what follows, we describe the key moments that preceded the inclusion of an ‘altruism’ heading in the *Journal of Economic Literature* (JEL) classification system for journal articles at the end of 1993.

Understanding philanthropy

After private foundations fell under increasing regulatory scrutiny in the early 1950s (Hall, 1999) and their tax status was attacked in the early 1960s (Frumkin, 1999), their leaders found it opportune to approach economists for advice. In effect, Donald Young, President of the Russell Sage Foundation (RSF), asked Solomon Fabricant, Director of Research at the National Bureau of Economic Research (NBER), to think about the possibility of investigation into the economic aspects of philanthropy. The RSF eventually funded a study of this phenomenon in the American economy, which was conducted by the NBER between 1959 and 1962 under the supervision of the economist Frank G. Dickinson, who was assisted by an advisory committee. The first meeting of the committee took place in late 1959.

A few members of the NBER staff, notably Becker, attended the meeting. Some work stemming from it, notably by Fabricant and Dickinson, and dealing mostly with definitional and empirical aspects, benefited from a limited circulation. That explains why Becker wrote the obscure and unpublished ‘Notes on an Economic Analysis of Philanthropy’ in April 1961. He later identified this article as the first expression of his interest in social interactions, but it was originally just another effort to extend the utility maximization assumption to the study of ‘non-economic’ topics. Becker's ‘Notes’ was not the only outgrowth of the NBER project. In addition, a conference, envisioned by Dickinson, took place in June 1961, bringing together a number of economists, among whom William Vickrey and Kenneth Boulding gave papers and James Buchanan simply attended.

By the early 1960s, then, the study of philanthropy had provided an opportunity for a handful of economists to explore aspects of seemingly unselfish behaviour. Following Dickinson's remark, in late 1959, that philanthropy was not in the mainstream of economic analysis, Becker (1961), Vickrey (1962) and Boulding (1962) suggested that there were no theoretical impediments to its understanding. ‘It can be dealt with quite easily in utility theory’, wrote Boulding, ‘by considering the utility of one person a function not only of his own wealth or his own income, but a function of the wealth and income of others’ (1962, p. 61). Essentially Becker and Vickrey agreed. Utility interdependence, in its modern form, had long been around and it appeared to be the proper tool to tackle philanthropic behaviour, even if there could be variations in the arguments to be included in the giver's utility function. There was, however, a more significant difference. Becker was not especially concerned with the motivations of philanthropic behaviour, whereas Boulding and Vickrey were: they believed utility theory could not elucidate the variety of motivations for philanthropy. Accordingly, Boulding emphasized the sense of community (and the associated capacity for empathy) as the essence of ‘genuine philanthropy’, while

Vickrey saw social distance as the significant factor.

Following the work of Becker, Vickrey and Boulding, various research efforts gave momentum to the study of philanthropy. Interested as he was in the effects of fiscal systems on income redistribution, Buchanan could easily relate to the theme of the philanthropy conference. As was the case for Becker and Boulding, his work at the intersection of economics and other social sciences made the whole undertaking of studying a form of seemingly unselfish behaviour especially appealing to him. His own views on the free-rider problem led him to distinguish between the expediency criterion and moral law as the two main determinants of an individual's choice, and to connect their relative strength to group size (Buchanan, 1965). The individual was said to follow moral law in small-group interactions and turn into a utility maximizer as soon as group size grew – the ‘large-group ethical dilemma’. In his presidential address to the American Economic Association a few years later, in December 1968 when social crisis was at its height, Boulding (1969), too, felt it timely to contrast two sets of common values guiding human behaviour: the ‘economic ethic’ and the ‘heroic ethic’, with the former centring on cost–benefit analysis and the latter emphasizing the sense of identity.

After his resignation from the University of Virginia in 1968, Buchanan visited UCLA where a number of economists, including Armen Alchian and William Allen (1964) and Jack Hirshleifer (1967), had studied seemingly unselfish behaviour. These authors had reached the conclusion that its treatment did not require a fundamental reconsideration of the behavioural assumptions of economic theory. Buchanan had doubts. Although he recognized the merits of enriched utility functions for the study of seemingly unselfish behaviour, Buchanan warned that they did not unveil the variety of human motivations. Consequently, he argued, the inclusion of ‘non-economic’ arguments, such as love or concern for the welfare of others, into the utility function did not necessarily improve the predictive power of theory.

Understanding altruism

In the context of adverse circumstances for foundations, the 1962 conference was meant to correct the inadequacy of knowledge about the economic aspects of philanthropy. In 1971, Edmund Phelps sent a grant proposal to Orville Brim, Jr., then President of the RSF, to ask support for the organization of a conference to be held in New York City. Here, too, the idea of the conference emerged in a difficult political environment. With the Tax Reform Act of 1969 imposing new regulations, foundations leaders were under pressure to defend philanthropy from any further threat. Unlike its predecessor, however, the conference contemplated by Phelps would not deal with an instance of seemingly unselfish behaviour, but with altruistic behaviour in general.

Pointing to the extension of the domain of economics to neglected topics such as crime and war, to the disenchantment with classical liberalism that accompanied the intensification of economic problems and the deepening of social crisis in the United States, and to new developments in the analysis of markets such as the relaxation of the assumption of perfect information, Phelps concluded: ‘the time has arrived for a theory of altruism’ (Phelps to Brim, 19 October 1971). That the conference was meant to deal with a topic, the definition of which was still unclear to many, including Phelps himself, speaks volumes about the appeal of seemingly unselfish behaviour in social science at a time when ‘the amount of divisiveness and conflict in a society’ – to use Mancur Olson's (1971, p. 173) words – occasioned serious concern.

Phelps's consideration of possible participants reveals that what the profession has come to call ‘altruism’ was in the early 1970s a heterogeneous body of knowledge comprising disparate analyses of human behaviour. Phelps first contacted Kenneth Arrow, Paul Samuelson and Vickrey, who all agreed to present papers. In his proposal, Phelps mentioned Boulding, Thomas Schelling, Becker, James Mirrlees, Peter Hammond, Sydney Winter, Alchian, Duncan Foley and Scott Boorman. Among non-economists, philosophers had the lion's share in an otherwise odd group including John Rawls, Tom Nagel, Marshall Cohen, Erving Goffman, Edward Banfield, Bernhard Lieverman and Sydney Morgenbesser. Several of these researchers were part of a movement in the late 1960 and early 1970s to connect moral philosophy with economics and other social sciences. And many of them were concerned with the respective role of self-interest and ethics in the explanation of human behaviour.

Amartya Sen did not appear in the list above but he attended the conference. His call for reconsidering the economic theory of human behaviour fitted in well with the overall preoccupation of the conference with ethics. In his LSE inaugural lecture, ‘Behaviour and the Concept of Preference’, Sen (1973) offered valuable insights into the relationships between choices and individual preferences, showing that the same choice (use and reuse of glass bottles) could correspond to four distinct cases in terms of the agent's underlying preferences. The first three cases represented the preferences of a selfish, sympathetic and socially conscious individual, respectively; they were consistent with utility theory. The fourth case, which Sen (1977) later associated with the notion of ‘commitment’, was of a different sort, however. It shows that moral considerations could influence individual choice in such a way as to undermine the correspondence between choice and preference on the one hand and preference and welfare on the other. The maximization framework with utility interdependence told some truth about seemingly unselfish behaviour, but not the only truth.

However, not all students of seemingly unselfish behaviour found ethics illuminating. At about the same time as Sen's LSE lecture was published in August 1973, Arthur Seldon, from the Institute of Economic Affairs (IEA), the London-based think-tank, was completing the preface to *The Economics of Charity: Essays on the Comparative Economics and Ethics of Giving and Selling, with Applications to Blood*. Unlike Sen and others, the main contributors to the collection, including Alchian, Allen and Gordon Tullock, were doubtful about the possibility of learning something significant economically from an ethical approach to unselfish behaviour. They preferred instead to explore the potentialities of utility theory.

In the early 1970s, economists were undoubtedly showing greater interest in what was now occasionally called ‘altruism’, but a unified theory was still lacking. The plurality of viewpoints reflected varied motivations, with some striving to renew the understanding of small-group interactions and others discussing either the moral dimension of economic behaviour or the economic dimension of moral behaviour. In the literature, there emerged a dividing line between the advocates of *homo economicus* and the supporters of *homo ethicus*, which became more pronounced with the publication of Becker's ‘A Theory of Social Interactions’ (1974) and Phelps's *Altruism, Morality, and Economic Theory* (1975), a collection of essays resulting from the New York conference.

The polarization of the mid-1970s

Becker's 1974 article was originally titled ‘Interdependent preferences: charity, externalities and income taxation’: it was renamed in September 1969 – a change that revealed Becker's intention to broaden his

frame of analysis from the issue of charity to the treatment of seemingly unselfish behaviour in general. The article was published in the same issue of the *Journal of Political Economy* as Robert Barro's ‘Are Government Bonds Net Wealth?’ (1974). It would be unreasonable to think of Barro's analysis of government budget deficits as a simple application of Becker's ‘rotten kid’ statement, but some cross-fertilization occurred, especially since Becker's manuscript had spent some six years in his files and Barro had commented on it. Becker also knew Barro's article, a draft of which had been presented, in 1973, in the Money and Banking workshop run by Milton Friedman in Chicago. That Becker and Barro discussed seemingly unselfish behaviour is evidenced by the fact that the latter's former wife suggested the phrase ‘rotten son’ to the former who later turned it into ‘rotten kid’ in his eponymous ‘theorem’ (Barro to Fontaine, 3 April 2001)

Becker proposed, in contrast to what he called the ‘usual theory of consumer choice’, which places in the utility function of the giver his own consumption together with the amount of his charitable giving, a ‘social interactions’ approach, which replaces the amount of charitable giving with the consumption of beneficiaries, as financed by their income and the amount of charitable giving they receive. In the context of the family, Becker reached the conclusion ‘that if a [benevolent] head exists, *others members also are motivated to maximize family income and consumption, even if their welfare depends on their own consumption alone*. This is the ‘“rotten kid” theorem’ (1974, p. 1080).

Against the background of a family break-up, Becker showed that the conditions for family cohesion were not so demanding as to require that all family members have sympathetic preferences or so unrealistic as to imply that all family members are selfish. Regarding the recipients of the head's generosity, he endorsed Friedman's (1953) influential argument and made it clear that only ‘*as-if* altruism’ was involved. Yet the head had sympathetic (‘altruistic’) preferences. In other words, his transfers were said to result from sympathy, which was explained by the fact that ‘the marriage market is more likely to pair a person with someone he cares about than with an otherwise similar person that he does not care about’ (Becker, 1974, p. 1074n).

Assuming continuity between family and other groups, Becker extended his results to the ‘synthetic “family” ’, consisting of a charitable person and all recipients of his or her charity, and to a number of other multi-person interactions. Here again, due to offsetting changes in transfers from the sympathetic benefactor, a redistribution of income among ‘members’ left their own welfare unchanged. To the problem represented by the possibility that opportunistic tendencies can surface in groups characterized by the interactions of selfish individuals and therefore prevent socially desirable outcomes, in the mid-1970s Becker offered a solution centred on the sympathetic preferences of certain individuals in society. To many today, this answer will seem ad hoc, but, at a time when much was said about the unresponsiveness of people to each other's lot, it went against the stream. With the increase in macroeconomic volatility, its policy implications were, however, straightforward: due to offsetting private transfers, one could hardly count on social and economic policies to change the distribution of resources (see Barro, 1974).

Though it can be argued that ‘A Theory of Social Interactions’ played a significant role in the history of unselfishness research, it should be remembered that its main objective was to analyse the economic implications of interactions within various groups. Phelps (1975), by contrast, meant to offer a contribution to the ‘theory of altruism’. As such he aimed at understanding a variety of behaviours, the motivations of which were seemingly unselfish. While Becker had provided a coherent framework

centred on maximization with utility interdependence to analyse social interactions, the essays in Phelps's collection illustrated the complexity, indeed vagueness, of ‘altruism’ as soon as one ventures beyond the self-interest model.

In dealing with unselfishness, Phelps's book actually considered a great variety of behaviours and motivations. Accordingly, contributors strove to classify them so as to identify their similarities and differences. When Arrow (1975) discussed Richard Titmuss's analysis of blood giving and its motives, for instance, he introduced a distinction between benefiting from the satisfactions obtained by others, benefiting from one's contributions to these satisfactions and the idea that ‘each performs duties for the other in a way calculated to enhance the satisfaction of all’ (1975, p. 17), but he refrained from providing an economic translation of Titmuss's reference to a sense of obligation to strangers. Arrow acknowledged the possibility that individuals act according to a categorical imperative, but noted: ‘I should add that, like many economists, I do not want to rely too heavily on substituting ethics for self-interest’ (1975, p. 22).

Others in the volume were probably more willing to take note of ethical motivations if only because they could serve to justify opposition to governmental regulation in various areas. In ‘The Samaritan's Dilemma’, Buchanan (1975) showed that the expectation of other-oriented behaviour could lead the potential beneficiary to behave opportunistically. Of particular interest in Buchanan's approach was the association of the undesirable consequences of other-oriented behaviour with the prevalence of the expediency criterion (the selfishness of agents) in society and the conclusion that commitment à la Schelling offered a solution to that problem. This solution, Buchanan realized, was threatened by the weakening adherence to ethical rules resulting from increase in group size.

The last three essays in Phelps's volume came back to the issue of philanthropy. Of particular interest was Bruce Bolnick's (1975) acknowledgement that a number of writers had ‘rendered such behavior susceptible to the traditional tools of economic analysis’ and his concomitant remark that ‘a more fundamental issue is uncovered: What types of motivation underlie philanthropic activity?’ (1975, p. 197). In the same vein, Bolnick pointed to the difference between trying to understand seemingly unselfish behaviour and studying the consequences of the inclusion of utility interdependence in the maximization framework in terms of optimality conditions (see, for example, Hochman and Rodgers, 1969; Kolm, 1969; Thurow, 1971). The latter approach Bolnick saw as ‘unsatisfying as a behavioral theory’ (1975, p. 198) and accordingly argued that social rewards and psychological consistency had to be taken into account not only for small groups but also for larger ones. In the process, Bolnick mentioned the justification in terms of empathetic identification, as suggested by Boulding (1962) and Vickrey (1962), but expressed uneasiness with its limitation to close-knit groups.

Despite notable efforts to go beyond the self-interest model, *Altruism, Morality, and Economic Theory* failed to identify the main features of the ‘commitment model’. The fact that ethical considerations had to be taken into account in the analysis of seemingly unselfish behaviour did not mean that the self-interest model failed on most accounts or that another model could claim greater explanatory power. It is understandable therefore that in his Introduction to the volume Phelps (1975) wavered: ‘Can altruistic behavior be fit into some version of the economist's beloved model of utility maximization subject to constraints? Or must that model be importantly modified and hooked up to some complementary body of analysis to yield a satisfactory product?’ (1975, p. 2). Jean-Jacques Laffont (1975) conveyed some of these tensions when he uncharacteristically defined the behaviour of *homo economicus* as selfish, not

self-interested, and contrasted it with ‘Kantian’ behaviour.

The self-interest view of unselfishness

With the studies of seemingly unselfish behaviour within the framework of utility maximization with interdependence, the question of the arguments to be included in the utility function became more relevant than that of the actual motivations of behaviour, though these arguments have occasionally been equated with motives for action.

The malleability of utility functions made it possible for economists to consider a variety of influences on the satisfaction of the individual besides own consumption. It even allowed for the inclusion of biological arguments into the utility function. Becker's (1976a) review article on Edward Wilson's (1975) controversial *Sociobiology* provides an interesting illustration. To Wilson, who suggested that biology might enlighten the analysts of social behaviour, Becker, who by that time saw himself as one of them, replied that economics too had its merits in terms of explaining the ‘social’ (for illustrations of the ‘economic approach’, see Becker, 1976b). Thus, though Becker accepted Wilson's definition of *altruism* as behaviour that reduces one's genetic fitness to the benefit of another's, he also pointed out that ‘altruism’, because of its effects on the behaviour of beneficiaries, could increase the genetic fitness of the ‘altruist’. In emphasizing the positive outcome of unselfish behaviour for the ‘altruist’, Becker complicated the emerging discourse on the essentially selfish nature of human behaviour, as derived from the view that ‘altruism’ is detrimental to its author (Dawkins, 1976).

There was indeed something accidental about Becker's considering the biological basis of social behaviour and writing about sociobiology, but for economists taken by the ‘economic approach’ there was good reason to address unselfishness: economics could not hope to embrace anthropological, sociological and political subjects without at the same time breaking away from the advocacy of behavioural assumptions that pictured the economic agent as a non-social being.

The second half of the 1970s offered several examples of authors, among whom were Hirshleifer and Tullock, who advocated the expansion of the ‘economic’ and wrote on unselfishness as well. It is hardly surprising therefore that these two commented on Becker's article in the *Journal of Economic Literature*. Hirshleifer (1977a), whose extremely well-documented ‘Economics from a Biological Viewpoint’ had just appeared in the *Journal of Law and Economics*, another symbol of the expansionist ambitions of economics, noted that the ‘ “rotten kid” theorem’ obtained only if the ‘altruistic’ head had the last word in the decision sequence (Hirshleifer to Becker, 13 December 1976; see also Hirshleifer, 1977b). Hirshleifer's proviso suggested paradoxical implications. If the ‘head’ did not have the last word, the theorem lost its strength as a demonstration that selfish individuals were dissuaded from behaving opportunistically in groups; if he or she did, on the other hand, it might be presumed that some of the problems dealt with in the ‘theorem’ lost significance.

Unlike Becker, Tullock (1977) preferred a model of unselfishness in which the giver derives utility from the mere act of giving. In his comment, he made the interesting point that in Becker's model the giver does not necessarily know the preference ordering of recipients. Becker thought this problem irrelevant since his model was concerned with family, not government, transfers (Becker to Tullock, 14 December 1976). Such a justification, it should be noted, could undermine the claim that his argument reached beyond the kin selection explanation of unselfishness by biologists.

As Hirshleifer's and Tullock's reactions to Becker's inroad into sociobiology illustrate, some economists were interested in biology. Though the impetus came from the heated debates surrounding the publication of *Sociobiology*, the ongoing redefinition of territories in social science was the determining factor. In his review of the literature on the relationships between economics and biology, Hirshleifer (1977a) noted that ‘the social sciences generally can be regarded as in the process of coalescing’ (1977a, p. 3) and he concluded that ‘economics can be regarded as the general field, whose two great subdivisions consist of the natural economy studied by the biologists and the political economy studied by economists proper’ (1977a, p. 52). Clearly, economists were unwilling to see their attempts at investigating the ‘social’ threatened by similar ambitions on the side of natural scientists (see Hirshleifer, 1985, who later spoke of ‘competing imperialisms’ but acknowledged their complementarities), especially since these attempts continued to be regarded suspiciously by some in the profession. Accordingly, economists took every occasion to emphasize economics’ lessons for the natural sciences. Becker did this and so did others, including Boulding (1978), Hirshleifer (1977a), Schelling (1978), Tullock (1978; 1979), who all took an interest in studying ‘non-economic’ behaviour. Though these various initiatives enjoyed greater visibility with the organization of a session on ‘Economics and Biology: Evolution, Selection, and the Economic Principle’ at the meeting of the American Economic Association in December 1977, from the early 1980s unselfishness research was conducted independently of sociobiology. With ‘economics imperialism’ gradually entering the mainstream (see Stigler, 1984; Hirshleifer, 1985), the interest of economists turned to the more general study of the relationships between economics and biology (see for example, Hirshleifer, 1982; Nelson and Winter, 1982; Samuelson, 1985), and it is only in the early 1990s that the question of unselfishness surfaced again in this kind of literature (Tullock, 1990; Simon, 1990; 1992; 1993; Bergstrom and Stark, 1993; Samuelson, 1993).

By the early 1980s, the self-interest view of unselfishness was well established in the profession: it associated ‘altruism’ with the fact that an individual's utility function depended on another's well-being. Becker's (1981, p. 2) ‘Altruism in the Family and Selfishness in the Market Place’ illustrated the main orientations of that view when he noted that his was a definition of *altruism* that concerned behaviour, not ‘a philosophical discussion of what “really” motivates people’, and that ‘altruism’ was more common in the family than in the market place because of its greater relative efficiency in the former (1981, p. 10).

The departures from ‘altruism’ à la Becker were encouraged by the political debates of the 1980s. With the macroeconomic volatility of the 1970s, the bearing of economics on policy matters began to be challenged. The conclusion, that due to offsetting transfers from ‘altruists’ one could hardly count on social and economic policies to change the distribution of resources, found continuation in various remarks about the ‘ungovernability’ of modern societies (see Olson, 1982, p. 8). And with the beginning of Ronald Reagan's first presidency and its economic programme turning away from demand management, the link between the ineffectiveness of governmental redistribution and the existence of sympathetic transfers took up a broader significance; it could be taken as another argument for lesser state intervention.

With significant changes in economic and social policies looming on the horizon in the first half of the 1980s, notably ‘the control of federal spending, the reduction or elimination of a wide variety of social entitlement and redistributive schemes … and the aggressive reduction of tax rates on

incomes' (Bernstein, 2001, p. 164), a number of economists were led to re-examine the strength of the ' "Ricardian equivalence" theorem' and the ' "rotten kid" theorem', two results that were closely associated with the unselfishness literature.

Becker's (1981) article appeared in February at a time when President Reagan's programme was being presented. That programme carried with it a vision of the workings of society that some of Reagan's predecessors considered mistaken, precisely because it gave inadequate weight to the failures of the invisible hand of the market. In the 1960s and 1970s, some may have thought seemingly unselfish behaviour a solution to the opportunistic tendencies capable of emerging in groups – small and perhaps large as well – but in the 1980s there was growing scepticism towards that possibility as well as gradual realization that 'altruism' à la Becker was not necessarily a positive force (see, for instance, Wintrobe 1983).

Building on Becker's model, B. Douglas Bernheim, Andrei Shleifer and Lawrence H. Summers (1985) included a strategic component into family transfers. The authors did not reject the possibility of sympathetic transfers from parents (or testators), but stressed above all their intention to control the beneficiaries' behaviour. In departing from Becker's model, the authors noted that the ' "Ricardian equivalence" theorem' did not hold in theirs (1985, p. 1046) and that the ' "rotten kid" theorem' was valid only under special circumstances (p. 1048). At least from that perspective, there was ground for reconsidering the presumed ineffectiveness of public policies.

Yet the authors preferred instead to review some macroeconomic implications of their model. When contrasted with Becker's, theirs was especially interesting because it reached the conclusion that the influence of parents over their children went further than simply dissuading opportunism within the family. While Becker's model was turned towards the absorption of the negative effects of economic and social change by the 'head' of the family, Bernheim, Schleifer and Summers, in emphasizing parents' influence on 'decisions by their children concerning education, migration and marriage' (1985, p. 1073), identified family as a factor of economic and social change. In the context of the breakdown of the traditional family unit, that conclusion could surprise, but it could also appear as the recognition that, with the loosening of family bonds, not only sympathy but also strategy was needed to prevent opportunism.

Further clarification in terms of policy implications came from Bernheim (1986) and Bernheim and Bagwell (1988), who instead of directly challenging the neutrality implications of Barro's (and Becker's) analytical framework pointed to its unsuitability to analyse the effects of public policies. In rejecting the 'Ricardian equivalence hypothesis', these authors suggested a different analytical framework in which the linkages between families, more than the 'dynastic family' à la Barro, were especially important. On the basis of these linkages, Bernheim and Bagwell (1988) established strong neutrality results, the practical implications of which they eventually dismissed on the grounds of being unrealistic. Perhaps because changes affecting family since the 1970s gained more visibility by the end of the 1980s, a number of presuppositions, characterizing Becker's and Barro's notion of family as that of a 'big happy family' behaving as if it maximized a single utility function (Bernheim and Bagwell, 1988, p. 333), became gradually untenable. At the very least, the complexity of intra-family relationships seemed to call for alternative representations.

The changes in perspective can easily be realized when one considers Assar Lindbeck and Jörgen W. Weibull's (1988, p. 1165) argument about the inefficient outcomes generated by 'altruism'. In an

intertemporal setting, the authors argued, gift-giving leads to social inefficiencies because the recipient can act strategically and thus induces the donor to give more than he or she was prepared to (see also Bruce and Waldman, 1990). Though reminiscent of Buchanan's ‘Samaritan's Dilemma’ of the mid-1970s, the argument differed in that it allowed for unselfish preferences on both sides (see also Kimball, 1987). Like Buchanan's, it suggested a solution in terms of commitment à la Schelling, with the donor making a binding commitment to the level of support provided to the recipient; and, like Buchanan's, the argument included the proviso of the difficult practical enforceability of that solution. Unlike Becker's suggestion, unselfishness did not suffice to remove opportunistic tendencies in social interactions; it could even encourage them.

In the same vein, Bernheim and Stark (1988, p. 1034) saw the ‘ “rotten kid” theorem’ as rather ‘special’ and even identified ‘a variety of circumstances in which members of a group would actually prefer to interact with less altruistic individuals, and in which the efficiency of resource allocation is inversely related to the prevailing degree of altruism’ (for a perhaps more positive, though nuanced, view, see Bergstrom, 1989). In addition to the criticisms levelled at Becker, that article called into question the customary distinction between family and the market in terms of behavioural assumptions. To the extent that ‘altruism’ tended to induce exploitability, it was suggested that ‘family decisions were more properly modelled as negotiations among primarily self-interested (read: ‘selfish’) agents (Bernheim and Stark, 1988, p. 1044). As far as society was concerned, similar conclusions apply: ‘altruism’ did not necessarily limit negative externalities. Worse still, unless it reached high levels, there were indications of its being a ‘counterproductive social force’.

In view of the above, it may be concluded that a decade and a half after Becker and Barro had produced their results, there were serious misgivings about the generality of their application. Given that unselfishness research owed some of its impetus to the realization of the undesirable consequences of selfish behaviour in terms of the provision of public goods and considering that government intervention could be regarded as a solution to that problem, there was some irony in James Andreoni's (1990) conclusion that economic and social programmes could increase the total provision of public goods because not merely sympathetic but also selfish considerations motivated giving.

In studying privately provided public goods, Andreoni (1988) interpreted various neutrality results as many limitations of the ‘pure altruism model’, which he identified with the definition of the utility function of the giver as including his own consumption and the total supply of public good. Citing in passing Margolis (1982), Sugden (1984) and Bernheim, Shleifer and Summers (1985), he called for a new approach characterized by ‘non-altruistic motives for giving’ (Andreoni, 1988, p. 72). In subsequent works, however, Andreoni (1989; 1990) clarified his own alternative model by resorting to the warm-glow hypothesis, whereby he meant that the utility function of the giver also included his personal contribution to the public good. Combining altruism à la Tullock with altruism à la Becker, this ‘impure altruism model’ was said to be more consistent with empirical evidence contradicting neutrality.

Throughout the Reagan years, there were a variety of results in economics contradicting neutrality. Given the increase in the government debt over that period, it was clear that ‘lesser state intervention’ meant not so much strict control of federal spending as its reorientation in the context of tax reduction. From that perspective, the results obtained by Andreoni and others suggested that the existence of sympathetic transfers could not be taken as a serious justification for the ineffectiveness of national policies. Accordingly, the emphasis was shifted towards examining the power of government

intervention to remedy the undesirable social consequences not only of selfish but also self-interested behaviour.

When it is remembered that Becker presented the existence of a sympathetic head as a solution to the difficulty of achieving socially desirable outcomes in various groups of otherwise selfish individuals, it is hardly surprising that the literature emphasizing the limits to ‘altruism’ was led to confront Becker's work on the family. In their variety, these critics did not call into question the utility maximization framework. For others, however, that framework showed significant inadequacies when it came to explaining seemingly unselfish behaviour.

Alternative views of unselfishness

Just as Becker's (1974) ‘A Theory of Social Interactions’ epitomizes the self-interest view of unselfishness, so Sen's (1977) ‘Rational Fools’ represents the alternative views though the latter go beyond the well-known distinction between ‘sympathy’ and ‘commitment’. While Sen delivered his ‘Rational Fools’ lecture at Oxford University in October 1976, Margaret Thatcher was already the leader of the Conservative Party and when the lecture was published in the summer of 1977 she was only a couple of years from being Prime Minister. That was a time of transition to economic liberalism. Thatcher's intention to dismantle collectivist public policies raised doubts within her own party and in society at large. The fact that Sen, a professor at the London School of Economics since 1971, proposed ‘a critique of the behavioral foundations of economic theory’ (the subtitle of his 1977 article) was a reminder that from the 1960s the debates on public policy in Britain had been marked by the strengthening of a vision endorsing the invisible hand of the market and economic man.

For Sen, sympathy or concern for others’ welfare (‘altruism’ for most economists) was part of the self-interest model, whereas ‘commitment’ was not. He wrote:

> The former corresponds to the case in which the concern for others directly affects one's own welfare. If the knowledge of torture of others makes you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment … It can be argued that behavior based on sympathy is in an important sense egoistic, for one is oneself pleased at others’ pleasure and pained at others’ pain, and the pursuit of one's own utility may thus be helped by sympathetic action. It is action based on commitment rather than sympathy which would be non-egoistic in this sense. (Sen, 1977, p. 326)

Perhaps because it was difficult for economists to think of an unselfish person as someone who is motivated by the welfare of others and yet benefits personally from his or her action, Sen stressed exaggeratedly both the interestedness of sympathetic agents and the indifference of committed ones. Another aspect of Sen's approach was to link commitment to groups and then distinguish it from ‘impartial concern for all’, as illustrated by ethical preferences à la Harsanyi (Sen, 1977, p. 336). In following that lead, Sen was echoing the earlier distinction between two sets of values, the ‘economic ethic’ and the ‘heroic ethic’, which Buchanan (1978) was now presenting under the guise of two motivational forces, ‘self-interest’ and ‘community’, the latter of which he continued to connect with

group size. In the context of a changing society, which some saw as regressing economically because of the inadequate attention being given to the invisible hand mechanism, Sen felt the need to remind his readers that in addition to contributing to social harmony the economy required a degree of social cohesion and that the latter was facilitated by the individuals' sense of commitment to groups. Accordingly, economics' behavioural assumptions needed to be reconsidered so as to allow for commitment. David Collard (1978), in one of the first monographs on the subject of unselfishness, illustrated this orientation when he argued that once all self-interested motivations were allowed for, there was still room for 'a truly altruistic residual' (1978, p. 5).

By the early 1980s, it was clear that the sympathy-based view of seemingly unselfish behaviour was not the whole story. In two voluminous articles, the French economist Serge-Christophe Kolm (1981a; 1981b) showed the complexities of 'altruism' and linked them to the prevailing schizophrenia associated with *Das Adam Smith Problem*. To some extent, Margolis's *Selfishness, Altruism, and Rationality* (1982) shifted the problem to the coexistence of two selves (or two utility functions representing an individual's self-interested preferences and his group-interested preferences, respectively) in economic man. For economists accustomed to distinguishing between economic man and moral man, Margolis's approach was disturbing. Olson, who reviewed the manuscript for Cambridge University Press, urged Margolis to reframe the argument so as to bring it within standard economic theory (Margolis to author, 17 May 2001), but Margolis felt that his model of individual choice was more 'consistent with the way human beings are observed to behave' (1982, p. 3).

It is unclear whether Margolis's overall approach influenced economists. Yet his distinction between 'participation altruism' – in which the economic agent gains satisfaction from giving resources away to the benefit of others – and 'goods altruism' – in which the economic agent gains satisfaction from an increase in the goods available to others – gave structure to later attempts, such as Andreoni's, to combine these two kinds of altruism.

Among the alternative views of unselfishness, the British economist Robert Sugden's (1982) deserves special mention since it proposed to reconstruct the public good theory of philanthropic behaviour, which assumed that 'the total amount of a charitable activity is an argument in the utility functions of its donors' (1982, p. 350). Having in mind the British context in which large charities exist, Sugden saw one promising option as the dropping of the utility maximization assumption and the concomitant admission that 'some individuals act on moral principles rather than on pure self-interest' (1982, p. 349). He reached the conclusion that 'the conventional argument that private philanthropy leads to the under-supply of charitable activities cannot be sustained' (1982, p. 350). In the highly charged political environment of Thatcher's first administration, such a conclusion could easily be read as another argument for lesser government intervention.

As we have seen, in the mid-1970s the ineffectiveness of economic and social policies was often justified by the existence of sympathetic transfers, but by the mid-1980s some doubted the suitability of Becker's (and Barro's) 'altruism' theories to analyse the effects of public policies. Interestingly, in a later article, Sugden (1984) explicitly dissociated his effort from 'theories of *altruism*' – by which he meant representations of behaviour in terms of concern for others. He proposed a theory of reciprocity in which, because of a Kantian rule, an individual feels obliged to make an effort (in the production of some public good) that matches others' in the group (on a more general perspective on reciprocity, see Kolm, 1984). Here again, the British context was of some significance, as Sugden made clear when he

mentioned the role of unpaid donors in blood procurement as an example of the supply of public goods through voluntary contributions. Sugden made the ‘assumption that most people believe free riding to be morally wrong’ (1984, p. 772).

The above approaches rely on groups as a relevant level of analysis between the individual and society. Recourse to ethical variables in that context makes sense as the rejection of ethics from economics has long been encouraged by its focus on impersonal relationships in the market as opposed to interactions in close-knit groups, with frequency of interactions as the main factor constitutive of sense of belongingness. More recently, however, another factor has been considered. Sen (1985), for instance, studied the influence of identification with others in the determination of a person own welfare (for an earlier attempt in that direction, see Boulding's, 1962, notion of empathy in relation to groups). Sen recognized that ‘[o]ne of the ways in which the sense of identity can operate is through making members of a community accept certain rules of conduct as part of obligatory behavior towards others in the community’ (1985, p. 349). Likewise, Herbert Simon (1992) allowed for loyalty in and identification with groups, and even accepted the working of these notions at the level of the city or nation.

In these approaches, one feels a growing uneasiness as economists move from close-knit groups, such as the family, to more informal groups, such as the country, society or humanity, in which the more obvious associations in terms of behavioural assumptions are with self-interest and not those ‘perceptions of a shared humanity’ which Kristen Monroe (1996) in *The Heart of Altruism* saw as central to unselfishness. There remains that in theory nothing prevents individuals from empathizing with strangers, feeling sympathy for them and behaving altruistically towards them. To date, however, this line of research has not attracted much attention.

The question may therefore be asked whether economists entertaining alternative views of unselfishness have really been able to get over the dichotomy, to be found in the mainstream view, between the family/altruism and the market/selfishness (see, for example, Becker, 1981). Considering the slight impact of Philip Wicksteed on modern economics, it can be argued that economists have yet to digest his crucial distinction between the nature of an economic relation – the fact that the agent enters it without expressing concern for the purposes of his or her partner (‘non-tuism’) – and the agent's motives, which are either selfish or altruistic depending on whether the economic relation is meant to further the agent's own welfare or that of a third party (Steedman, 1989; Fontaine, 2000). The lack of appreciation for that distinction in modern economic theories of unselfishness and the resulting derivation of motivation (selfishness or unselfishness) from the nature of economic relation itself (impersonal or personal), explain why economists find it so unnatural to explore seemingly unselfish behaviour outside families or groups even if a number of other social scientists have shown less reluctance in that respect (see, for example, some contributions in Mansbridge, 1990).

1993: annus mirabilis

Following attempts to investigate philanthropy in the early 1960s, unselfishness theories experienced a dramatic growth. When it is remembered that in the late 1950s economists complained about the lack of attention to love of humankind (*philanthropy*), Collard's (1992) late addition to the debate on unselfishness, ‘Love is Not Enough’, signalled a sea change. By early 1990, the weaknesses of research in that area could no longer be attributed to inadequate scrutiny of seemingly unselfish behaviour.

In striking contrast with the early 1960s, 1993 was a prolific year: it saw the publication of a session on the 'Economics of Altruism' in the *Papers and Proceedings* of the *American Economic Review* (Samuelson, 1993; Bergstrom and Stark, 1993; Simon, 1993); a collection of essays, *Beyond Economic Man*, edited by Marianne Ferber and Julie Nelson (1993), which challenged the masculine foundations of economics' behavioural assumptions; and, outside economics, another collection including two essays by economists Sugden (1993) and Tyler Cowen (1993); and finally a special issue of the *Social Service Review* including interdisciplinary studies, among which was Dasgupta (1993), on the concept of 'altruism'. And to crown this achievement, Becker (1993) published a revised version of his Nobel Lecture in which he tellingly observed: 'Along with others, I have tried to pry economists away from narrow assumptions about self-interest [read: 'selfishness']. Behavior is driven by a much richer set of values and preferences' (1993, p. 385).

This list is not meant to be comprehensive, though it reflects the increasing volume of publication in this area and explains in turn the addition of an 'altruism' heading to the JEL classification system for journal articles in December 1993. Since then, research on seemingly unselfish behaviour has not slowed down, giving more room to economic experiments. There have been a reader (Zamagni, 1995), several monographs and collections of essays (Stark, 1995; Gérard-Varet, Kolm and Mercier-Ythier, 2000) and a handbook investigating the foundations and applications of altruism research (Kolm and Mercier-Ythier, 2006). If this remarkable development speaks to something it is certainly for economics' remarkable capacity to absorb and digest the most foreign subjects and notably those that present a serious challenge to its most central behavioural assumption. Whether this should be taken as a sign of strong intellectual identity is an open question.

See Also

- altruism in experiments
- charitable giving
- economic man
- ethics and economics
- rationality, history of the concept

Bibliography

Alchian, A.A. and Allen, W.R. 1964. *University Economics*. Belmont, CA: Wadsworth.

Andreoni, J. 1988. Privately provided public goods in a large economy: the limits of altruism. *Journal of Public Economics* 35, 57–73.

Andreoni, J. 1989. Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy* 97, 1447–58.

Andreoni, J. 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. *Economic Journal* 100, 464–77.

Arrow, K.J. 1975. Gifts and exchanges. In Phelps (1975).

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1095–117.

Barro to Fontaine. 3 April 2001. (personal communication).

Becker, G.S. 1961. Notes on an economic analysis of philanthropy. Mimeo. NBER, April.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82, 1063–93.

Becker, G.S. 1976a. Altruism, egoism, and genetic fitness: economics and sociobiology. *Journal of Economic Literature* 14, 817–26.

Becker, G.S. 1976b. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Becker to Tullock. 14 December 1976. Correspondence, *Journal of Economic Literature* Series, box 30, Mark Perlman Papers, Duke University, Rare Book, Manuscript, and Special Collections Library.

Becker, G.S. 1981. Altruism in the family and selfishness in the market place. *Economica* 48, 1–15.

Becker, G.S. 1993. Nobel lecture: the economic way of looking at behavior. *Journal of Political Economy* 101, 385–409.

Bergstrom, T.C. 1989. A fresh look at the rotten kid theorem – and other household mysteries. *Journal of Political Economy* 97, 1138–59.

Bergstrom, T.C. and Stark, O. 1993. How altruism can prevail in an evolutionary environment. *American Economic Review: Papers and Proceedings* 83, 149–55.

Bernheim, B.D. 1986. On the voluntary and involuntary provision of public goods. *American Economic Review* 76, 789–93.

Bernheim, B.D. and Bagwell, K. 1988. Is everything neutral? *Journal of Political Economy* 96, 308–338.

Bernheim, B.D., Shleifer, A. and Summers, L.H. 1985. The strategic bequest motive. *Journal of Political Economy* 93, 1045–76.

Bernheim, B.D. and Stark, O. 1988. Altruism within the family reconsidered: do nice guys finish last? *American Economic Review* 78, 1034–45.

Bernstein, M.A. 2001. *A Perilous Progress: Economists and Public Purpose in Twentieth-Century America*. Princeton and Oxford: Princeton University Press.

Bolnick, B.R. 1975. Toward a behavioral theory of philanthropic activity. In Phelps (1975).

Boulding, K.E. 1962. Notes on a theory of philanthropy. In *Philanthropy and Public Policy*, ed. F.G. Dickinson. New York: NBER.

Boulding, K.E. 1969. Economics as a moral science. *American Economic Review* 59, 1–12.

Boulding, K.E. 1978. Sociobiology or biosociology? In *Sociobiology and Human Nature: An Interdisciplinary Critique and Defense*, ed. M.S. Gregory, A. Silvers and D. Sutch. San Francisco: Jossey-Bass.

Bruce, N. and Waldman, M. 1990. The rotten-kid theorem meets the Samaritan's dilemma. *Quarterly Journal of Economics* 105, 155–65.

Buchanan, J.M. 1965. Ethical rules, expected values, and large numbers. *Ethics* 76, 1–13.

Buchanan, J.M. 1975. The Samaritan's dilemma. In Phelps (1975).

Buchanan, J.M. 1978. Markets, states, and the extent of morals. *American Economic Review* 78, 364–8.

Collard, D. 1978. *Altruism and Economy: A Study in Non-Selfish Economics*. Oxford: Martin Robertson.

Collard, D. 1992. Love is not enough. In *Thoughtful Economic Man: Essays on Rationality, Moral Rules and Benevolence*, ed. G. Meeks. Cambridge: Cambridge University Press.

Cowen, T. 1993. Altruism and the argument from offsetting transfers. In *Altruism*, ed. E. Frankel Paul, F. D. Miller, Jr. and J. Paul. Cambridge: Cambridge University Press.

Dasgupta, P. 1993. Altruism and the allocation of resources. *Social Service Review* 67, 374–87.

Dawkins, R. 1976. *The Selfish Gene*, 2nd edn. Oxford and New York: Oxford University Press, 1989.

Edgeworth, F.Y. 1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. New York: Augustus M. Kelley, 1967.

Ferber, M.A. and Nelson, J.A. 1993. *Beyond Economic Man: Feminist Theory and Economics*. Chicago and London: University of Chicago Press.

Fontaine, P. 2000. Making use of the past: theorists and historians on the economics of altruism. *European Journal of the History of Economic Thought* 7, 407–22.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago and London: University of Chicago Press.

Frumkin, P. 1999. Private foundations as public institutions: regulation, professionalization, and the redefinition of organized philanthropy. In *Philanthropic Foundations: New Scholarship, New Possibilities*, ed. E. Condliffe Lagemann. Bloomington and Indianapolis: University of Indiana Press.

Gérard-Varet, L.-A., Kolm, S.-C. and Mercier-Ythier, J. 2000. *The Economics of Reciprocity, Giving and Altruism*. Basingstoke: Palgrave Macmillan.

Hall, P.D. 1999. Resolving the dilemmas of democratic governance: the historical development of trusteeship in America, 1636–1996. In *Philanthropic Foundations: New Scholarship, New Possibilities*, ed. E. Condliffe Lagemann. Bloomington and Indianapolis: University of Indiana Press.

Hirshleifer, J. 1967. Disaster behavior: altruism or alliance? In *Economic Behaviour in Adversity*. Chicago: University of Chicago Press, 1987.

Hirshleifer to Becker. 13 December 1976. Correspondence, *Journal of Economic Literature* Series, box 31, Mark Perlman Papers, Duke University, Rare Book, Manuscript, and Special Collections Library.

Hirshleifer, J. 1977a. Economics from a biological viewpoint. *Journal of Law and Economics* 20, 1–52.

Hirshleifer, J. 1977b. Shakespeare vs. Becker on altruism: the importance of having the last word. *Journal of Economic Literature* 15, 500–2.

Hirshleifer, J. 1982. Evolutionary models in economics and law: cooperation versus conflict strategies. *Research in Law and Economics* 4, 1–60.

Hirshleifer, J. 1985. The expanding domain of economics. *American Economic Review* 75, 53–68.

Hochman, H.M. and Rodgers, J.D. 1969. Pareto optimal redistribution. *American Economic Review* 59, 542–57.

Kimball, M.S. 1987. Making sense of two-sided altruism. *Journal of Monetary Economics* 20, 301–26.

Kolm, S.-C. 1969. The optimal production of social justice. In *Public Economics: An Analysis of Public Production and Consumption and their Relations to the Private Sectors*, ed. J. Margolis and H. Guitton. London and New York: Macmillan.

Kolm, S.-C. 1981a. Efficacité et altruismes: les sophismes de Mandeville, Smith et Pareto. *Revue économique* 32, 5–31. Repr. in Kolm (1983).

Kolm, S.-C. 1981b. Altruismes et efficacités: le sophisme de Rousseau. *Information sur les sciences sociales* 20, 293–344. Repr. in Kolm (1983).

Kolm, S.-C. 1983. Altruism and efficiency. *Ethics* 94, 18–65.

Kolm, S.-C. 1984. *La Bonne Économie: La Réciprocité générale*. Paris: PUF.

Kolm, S.-C. and Mercier-Ythier, J. 2006. *Handbook of the Economics of Giving, Altruism and Reciprocity*, 2 vols. Amsterdam: North-Holland.

Laffont, J.-J. 1975. Macroeconomic constraints, economic efficiency and ethics: an introduction to Kantian economics. *Economica* 42, 430–37.

Lindbeck, A. and Weibull, J.W. 1988. Altruism and time consistency: the economics of fait accompli. *Journal of Political Economy* 96, 1165–82.

Mansbridge, J.J. 1990. *Beyond Self-Interest*. Chicago and London: University of Chicago Press.

Margolis, H. 1982. *Selfishness, Altruism, and Rationality: A Theory of Social Choice*. Cambridge: Cambridge University Press.

Margolis to Fontaine. 17 May 2001. (personal communication).

Monroe, K.R. 1996. *The Heart of Altruism: Perceptions of a Common Humanity*. Princeton, NJ: Princeton University Press.

Nelson, R.R. and Winter, S.G. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap.

Olson, M. 1971. Appendix. *The Logic of Collective Action: Public Goods and the Theory of Groups*, 2nd edn. Cambridge, MA: Harvard University Press.

Olson, M. 1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. New Haven and London: Yale University Press.

Phelps to Brim. 19 October 1971. Russell Sage Foundation Archives: Rockefeller Archive Center, Studies in Philanthropy, Folder 493, Box 57. New York, North Tarrytown.

Phelps, E.S. 1975. *Altruism, Morality, and Economic Theory*. New York: Russell Sage.

Robertson, D.H. 1955. What does the economist economize? In *Economic Commentaries*. London: Staple Press, 1956.

Samuelson, P.A. 1985. Modes of thought in economics and biology. *American Economic Review: Papers and Proceedings* 75, 166–72.

Samuelson, P.A. 1993. Altruism as a problem involving group versus individual selection in economics and biology. *American Economic Review: Papers and Proceedings* 83, 143–8.

Schelling, T. 1978. Altruism, meanness, and other potentially strategic behaviors. *American Economic Review: Papers and Proceedings* 68, 229–30.

Sen, A. 1973. Behaviour and the concept of preference. *Economica* 40, 241–59.

Sen, A. 1977. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs* 6, 317–44.

Sen, A. 1985. Goals, commitment, and identity. *Journal of Law Economics & Organization* 1, 341–55.

Simon, H.A. 1990. A mechanism for social selection and successful altruism. *Science* 250, 1665–8.

Simon, H.A. 1992. Altruism and economics. *Eastern Economic Journal* 18, 73–83.

Simon, H.A. 1993. Altruism and economics. *American Economic Review* 83, 156–61.

Stark, O. 1995. *Altruism and Beyond: An Economic Analysis of Transfers and Exchanges within Families and Groups*. Cambridge: Cambridge University Press.

Steedman, I. 1989. Rationality, economic man and altruism. In Zamagni (1995).

Stigler, G.J. 1984. Economics – The imperial science? *Scandinavian Journal of Economics* 86, 301–13.

Sugden, R. 1982. On the economics of philanthropy. *Economic Journal* 92, 341–50.

Sugden, R. 1984. Reciprocity: the supply of public goods through voluntary contributions. *Economic*

Journal 94, 772–87.

Sugden, R. 1993. Thinking as a team: towards an explanation of nonselfish behavior. In *Altruism*, ed. E. Frankel Paul, F.D. Miller, Jr. and J. Paul. Cambridge: Cambridge University Press.

Thurow, L.C. 1971. The income distribution as a pure public good. *Quarterly Journal of Economics* 85, 327–36.

Tullock, G. 1977. Economics and sociobiology: a comment. *Journal of Economic Literature* 15, 502–6.

Tullock, G. 1978. Altruism, malice and public goods. *Journal of Social and Biological Structures* 1, 3–9.

Tullock, G. 1979. Sociobiology and economics. *Atlantic Economic Journal* 7, 1–10.

Tullock, G. 1990. The economics of (very) primitive societies. *Journal of Social and Biological Structures* 13, 151–62.

Vickrey, W.S. 1962. One economist's view of philanthropy. In *Philanthropy and Public Policy*, ed. F.G. Dickinson. New York: NBER.

Wilson, E.O. 1975. *Sociobiology*, Abridged edn. Cambridge, MA: Belknap, 1980.

Wintrobe, R. 1983. Taxing altruism. *Economic Inquiry* 21, 255–70.

Zamagni, S. 1995. *The Economics of Altruism*. Aldershot and Brookfield: Edward Elgar.

How to cite this article

Fontaine, Philippe. "altruism, history of the concept." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000236> doi:10.1057/9780230226203.0034

The New Palgrave Dictionary of Economics Online

ambiguity and ambiguity aversion

Marciano Siniscalchi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Experimental evidence strongly suggests that subjects facing a decision under uncertainty often find it difficult to assess the relative likelihood of certain events; decision theorists deem such events ‘ambiguous’. Furthermore, subjects generally dislike options (acts) whose final outcome depends upon the realization of such ambiguous events; that is, they are ‘ambiguity-averse’. This article surveys the main decision-theoretic models developed since the mid-1980s to accommodate ambiguity and ambiguity aversion, including Choquet-expected utility (Schmeidler, 1989) and maxmin expected utility (Gilboa and Schmeidler, 1989). More recent developments in the theory of ambiguity are also briefly summarized.

Keywords

ambiguity; ambiguity aversion; Bernoulli utility function; choice under uncertainty; Choquet-expected utility; decision theory; decision weights; Ellsberg paradox; Incomplete preferences; maximum likelihood; maxmin expected utility; multiple priors; preference reversals; probability; Savage, L.; second-order probabilities; subjective expected utility; sure-thing principle; von Neumann–Morgenstern utility function

Article

Consider the following choice problem, known as ‘Ellsberg's three-colour urn example’, or simply the ‘Ellsberg paradox’ (Ellsberg, 1961). An urn contains 30 red balls, and 60 green and blue balls, in unspecified proportions; subjects are asked to compare (*a*) a bet on a red draw with a bet on a green draw, and (*b*) a bet on a red or blue draw with a bet on a green or blue draw. If the subject wins a bet, she receives ten dollars; otherwise, she receives zero dollars. To model this situation as a problem of choice under uncertainty, let the state space be $\{s_r, s_g, s_b\}$, in obvious notation, and consider the bets in Figure 1.

Figure 1
Ellsberg's three-colour urn

	S_r	S_g	S_b
f_r	10	0	0
f_g	0	10	0
f_{rb}	10	0	10
f_{gb}	0	10	10

The modal preferences in this example are $f_r \succ f_g$ and $f_{rb} \prec f_{gb}$, where '$\succ$' denotes strict preference. (Ellsberg did not conduct actual experiments, but similar patterns of behaviour have been reported in subsequent experimental studies; see Camerer and Weber, 1992, for an exhaustive survey.) A common rationalization runs as follows: betting on red is 'safer' than betting on green, because the urn may actually zero green balls; on the other hand, betting on green or blue is 'safer' than betting on red or blue, because the urn may contain zero blue balls. Equivalently, when one evaluates f_r and f_{gb}, the fact that the relative likelihood of green as against blue balls is unspecified is irrelevant; on the other hand, this consideration looms large when one evaluates the acts f_g and f_{rb}.

While these preferences seem plausible, they are inconsistent with subjective expected utility maximization (SEU). Indeed, they are inconsistent with the weaker assumption that the decision-maker's (DM) qualitative beliefs, as revealed by her betting behaviour, can be numerically represented by a probability measure. Note that $f_r \succ f_g$ indicates that r is deemed strictly more likely than g, so any probability P that represents the individual's likelihood ordering of events must satisfy $P(\{r\}) > P(\{g\})$; on the other hand, $f_{rb} \prec f_{gb}$ indicates that $\{r, b\}$ is strictly less likely than $\{g, b\}$, which would require $P(\{r\}) + P(\{b\}) = P(\{r, b\}) < P(\{g, b\}) = P(\{g\}) + P(\{b\})$, hence $P(\{r\}) < P(\{g\})$.

The key to Ellsberg's example is the fact that the composition of the urn is incompletely specified; in particular, the relative likelihood of a green as against a blue draw is

‘ambiguous’. More generally, in the words of Daniel Ellsberg, *ambiguity* is

> a quality depending on the amount, type, reliability and ‘unanimity’ of information, and giving rise to one's ‘degree of confidence’ in an estimate of relative likelihoods. (1961, p. 657).

To borrow Ellsberg's terminology, the modal preferences $f_r \succ f_g$ and $f_{rb} \prec f_{gb}$ indicate that the DM would rather have the ultimate outcome of her choices (that is, whether she receives 10 or 0) depend upon events about whose relative likelihood she is more confident. In other words, these preferences denote *ambiguity aversion*.
Since the mid-1980s, several decision models that can accommodate ambiguity and ambiguity aversion (or appeal) have been axiomatized; other contributions have addressed the behavioural manifestations and implications of ambiguity, as well as updating and dynamic choice. Furthermore, there is an ever-growing collection of applications to contract theory, auctions, finance, macroeconomics, political economy, insurance and other areas of economic inquiry.
The following section reviews two of the most influential models of ambiguity-sensitive preferences in a static setting, while the succeeding section briefly discusses additional models, updating, and dynamic choice.

‘Classical’ models of ambiguity-sensitive preferences

Preliminaries

Fix a finite or infinite state space *S* and an algebra Σ of its subsets. A *probability charge* is set function $P: \Sigma \to [0, 1]$ that satisfies $P(S) = 1$ and $P(E \cup F) = P(E) + P(F)$ for all $E, F \in \Sigma$ with $E \cap F = \varnothing$; that is, *P* is normalized and *finitely* additive. The set of probability charges on (*S*, Σ) is denoted $\Delta(S, \Sigma)$.
The decision models discussed in this section were first axiomatized in the framework introduced by Anscombe and Aumann (1963); it is convenient to adopt the same set-up here. (Alternative axiomatizations that do not rely on lotteries have also been obtained: see, for example, Gilboa, 1987; Chew and Karni, 1994; Casadesus-Masanell, Klibanoff and Ozdenoren, 2000; Ghirardato et al., 2003). Fix a set of prizes *X*, and let $\Delta(X)$ be the collection of all lotteries (probability distributions) on *X* with finite support. An act is a Σ-measurable map $f: S \to \Delta(X)$. The set $\Delta(X)$ is closed under mixtures, that is, convex combinations; mixtures of acts are then defined pointwise, so that the set $\mathcal{F}$ of all acts is also closed under mixtures (that is, for every $\alpha \in [0, 1]$ and every pair of acts f, g, $\alpha f + (1 - \alpha) g$ is the act that yields the lottery $\alpha f(s) + (1 - \alpha) g(s)$ in state $s \in S$).
A preference is a binary relation $\succsim$ on $\mathcal{F}$: its symmetric and asymmetric parts are denoted by ~ and $\succ$ respectively. It is customary to identify every lottery $p \in \Delta(X)$ with the constant act that yields *p* in every state.
A (von Neumann–Morgenstern, or Bernoulli) utility function is a map $u: \Delta(X) \to \mathbb{R}$ that satisfies $u(\alpha p + (1 - \alpha) q) = \alpha u(p) + (1 - \alpha) u(q)$ for all $\alpha \in [0, 1]$ and $p, q \in \Delta(X)$. All axiomatizations discussed below ensure that preferences over lotteries can be represented by a utility function.
A function $a: S \to \mathbb{R}$ is simple if its range is finite; write $a = (a_1, E_1; \ldots; a_n, E_n)$, where $a_1, \ldots, a_N \in \mathbb{R}$ and $E_1, \ldots, E_N$ is a partition of *S*, to indicate that, for all $n = 1, \ldots, N$, $a(s) = a_n$ for all $s \in E_n$. An act is simple if its range can be partitioned into finitely many indifference classes. The set of simple Σ-measurable acts is denoted by $\mathcal{F}_0$.
Virtually all substantive decision-theoretic issues can be analysed by restricting attention to preferences over $\mathcal{F}_0$; the reader is urged to consult the references cited for a discussion of preferences over non-simple acts.

Capacities and Choquet-expected utility

The modal preferences in the three-colour urn example are inconsistent with a probabilistic representation of beliefs essentially because probabilities are finitely additive. Specifically, if the probability charge *P* represents the individual's qualitative beliefs, $f_{rb} \prec f_{gb}$ requires that $P(\{r, b\}) < P(\{g, b\})$; since *P* is additive, this implies $P(\{r\}) < P(\{g\})$. However, $f_r \succ f_g$ implies the reverse inequality. Thus, formally, the Ellsberg paradox can be ‘resolved’ if a weaker, non-additive representation of the individual's qualitative beliefs is allowed. This approach is pursued in Schmeidler (1986; 1989).
A *capacity* is a set function $v: \Sigma \to [0, 1]$ such that $v(S) = 1$ and $v(A) \leq v(B)$ for all events $A, B \in \Sigma$ such that $A \subseteq B$. Thus, a capacity is not required to be additive, although it must satisfy a monotonicity property that has a natural interpretation in terms of qualitative beliefs: ‘larger’ events are ‘more likely’.
To define expectation with respect to capacities, a suitable notion of integration is required. Consider a simple function $a = (a_1, E_1; \ldots; a_N, E_N)$, with $a_1 > a_2 > \ldots > a_N$. The *Choquet integral* of *a* with respect to a capacity *v* (Choquet, 1953) is the quantity

$$\int a\,dP = \sum_{n=1}^{N-1} (a_n - a_{n+1}) v\left(\bigcup_{m=1}^{n} E_m\right) + a_N. \tag{1}$$

With the convention that $\cup_{m=1}^{0} E_m = \varnothing$, Eq. (1) can be rewritten as follows:

$$\int a\,dP = \sum_{n=1}^{N} a_n \left[v\left(\bigcup_{m=1}^{n} E_m\right) - v\left(\bigcup_{m=1}^{n-1} E_m\right)\right]. \tag{2}$$

Thus, Choquet integration performs a ‘weighted average’ of the values $a_1, \ldots, a_N$, with non-negative weights $v(E_1), v(E_1 \cup E_2) - v(E_1), \ldots, 1 - v(E_1 \cup \ldots \cup E_{N-1})$ that add up to one. If *v* is additive, Eq. (1) reduces to $\int a\,dP = \sum_{n=1}^{N} a_n v(E_n)$. However, in general, the ordering of the values $a_1, \ldots, a_N$ affects the decision weights: for instance, suppose $a = (\alpha, E; \beta, S\backslash E)$, with $\alpha \neq \beta$: then $\int a\,dv$ equals $\alpha v(E) + \beta[1 - v(E)]$ if $\alpha > \beta$, and $\beta v(S\backslash E) + \alpha[1 - v(S\backslash E)]$ if $\beta > \alpha$. These expressions are different unless $v(E) + v(S\backslash E) = 1$.
A preference admits a *Choquet-expected utility* (CEU) representation if there exists a utility function *u* and a capacity *v* such that, for all simple acts $f, g \in \mathcal{F}_0$, $f \succcurlyeq g$ if and only if $\int u(f(s))\,dv \geq \int u(g(s))\,dv$, where the integrals are as in Eq. (1).
Preferences in the Ellsberg paradox are consistent with CEU. Let *u* satisfy $u(10) > u(0)$, and observe that $f_r \succ f_g$ requires $v(\{r\}) > v(\{g\})$, whereas $f_{rb} \prec f_{gb}$ implies that $v(\{r, b\}) < v(\{g, b\})$; since *v* is not required to be additive, these inequalities can be mutually consistent: for instance, let

$$v(\{r\}) = v(\{r, b\}) = v(\{r, g\}) = \frac{1}{3}, v(\{b\}) = v(\{g\}) = 0, \text{and} v(\{b, g\}) = \frac{2}{3}. \tag{3}$$

Recall that the key axiom in the Anscombe–Aumann axiomatization of SEU is *Independence*: for all triples of (simple) acts *f*, *g*, *h*, and all $\alpha \in (0, 1)$, $f \succ g$ implies $\alpha f + (1 - \alpha)h \succ \alpha g + (1 - \alpha)h$. Schmeidler (1989) shows that CEU preferences are instead characterized by a weaker independence property. Say that two acts *f* and *g* are *comonotonic* if there is no pair of states *s*, *s* ′ such that $f(s) \succ f(s')$ and $g(s) \prec g(s')$; the key axiom in Schmeidler's characterization of CEU preferences, *Comonotonic Independence*, requires that $f \succ g \Rightarrow \alpha f + (1 - \alpha)h \succ \alpha g + (1 - \alpha)h$ only if *f*, *g*, *h* are pairwise comonotonic.
To illustrate the rationale behind this weakening of Independence, consider the acts f_r and f_g in the Ellsberg paradox, and define a third act f_b by $f_b(r) = f_b(g) = 0$ and $f_b(b) = 10$. For the CEU preferences defined above, $f_r \succ f_g$, but $\frac{1}{2} f_r + \frac{1}{2} f_b \prec \frac{1}{2} f_g + \frac{1}{2} f_b$. This is consistent with the notion that the DM dislikes ambiguity, and hence would rather have the ultimate outcome of her choices depend upon events about whose relative likelihood she is more confident; in particular, notice that the mixture $\frac{1}{2} f_g + \frac{1}{2} f_b$ yields the same outcome in states *g* and *b*, so the DM need not worry about her lack of confidence in her assessment of their relative likelihood.
This example also suggests that mixtures of non-comonotonic acts can be appealing for an individual who might informally be described as ‘ambiguity-averse’. As was just noted, mixtures of f_g and f_b can reduce or eliminate the dependence of the final outcome upon the realization of *g* rather than *b*, and hence provide a *hedge against ambiguity*. The DM under consideration finds this appealing: $\frac{1}{2} f_g + \frac{1}{2} f_b \succ f_g \sim f_b$.
Schmeidler (1989) suggests that this ‘preference for mixtures’ may be taken as a behavioural definition of *ambiguity aversion*. Formally, say that an individual is ambiguity-averse if,

for all $f, g \in \mathcal{F}_0$, $f \succcurlyeq g$ implies $\alpha f + (1-\alpha) g \succcurlyeq g$. Schmeidler then shows that a CEU individual is ambiguity-averse if and only if the capacity representing her preferences is *convex*: that is, for all events $E, F \in \Sigma$, $v(E \cup F) + v(E \cap F) \geq v(E) + v(F)$. For instance, the capacity in Eq. (3) is convex.

Multiple priors and maxmin expected utility

Gilboa and Schmeidler (1989, p. 142) propose an alternative rationalization of the preferences $f_r \succ f_g$ and $f_{rb} \prec f_{gb}$ in the Ellsberg paradox:

> One conceivable explanation of this phenomenon which we adopt here is as follows: …the subject has too little information to form a prior. Hence (s)he considers a *set* of priors as possible. Being [ambiguity] averse, s(he) takes into account the *minimal* expected utility (over all priors in the set) while evaluating a bet. For an analysis of this interpretation of multiple priors, see Siniscalchi (2006).

Formally, preferences admit a *maxmin expected utility* (MEU) decision rule if, given a utility function *u* and a weak* closed, convex set *C* of probability charges on *S*, for all $f, g \in \mathcal{F}_0$, $f \succcurlyeq g$ if and only if

$$\min_{P \in C} \int u(f)\, dP \geq \min_{P \in C} \int u(g)\, dP,$$

where integration has the usual meaning. For instance, the modal rankings in the Ellsberg paradox are consistent with MEU, with $u(10) > u(0)$ and

$$C = \left\{ P \in \Delta(S, \Sigma) : P(\{r\}) = \frac{1}{3} \right\} \tag{4}$$

(other choices of *C* are possible).

Gilboa and Schmeidler's axiomatization of the MEU decision rule features two key axioms: *C-Independence* and Ambiguity Aversion. The latter was stated in the previous subsection; C-Independence requires that, for all acts $f, g \in \mathcal{F}_0$ and all *constant* acts, or lotteries, $p \in \Delta(X)$, $f \succcurlyeq g$ if and only if $\alpha f + (1-\alpha) p \succcurlyeq \alpha g + (1-\alpha) p$. Thus, relative to the full Independence axiom, preference reversals are ruled out only for mixtures with constant acts.

Intuitively, mixing an act with a constant does not provide any hedging opportunities; rather, such mixtures change only the ‘scale and location’ of an act's utility profile. Thus, the requirement formalized by C-Independence is consistent with the discussion in the preceding subsection; indeed, CEU preferences satisfy C-Independence. On the other hand, C-Independence allows for violations of Comonotonic Independence (see Klibanoff, 2001, for an example and further discussion).

Ambiguity-averse CEU preferences satisfy both C-Independence and Ambiguity Aversion (in addition to other structural axioms); thus, they are MEU preferences. Schmeidler (1989) shows that, in particular, the convex capacity *v* representing an ambiguity-averse CEU preference is the *core* of the set *C* of priors in the MEU representation of the same preferences: that is, $C = \{P \in \Delta(S, \Sigma) : \forall E \in \Sigma, P(E) \geq v(E)\}$. For instance, the capacity *v* in Eq. (3) is the core of the set *C* in Eq. (4).

Other models, updating, and dynamic choice

A generalization of the MEU model, related to Hurwicz's α-maxmin criterion (cf. Luce and Raiffa, 1957, p. 304), sometimes appears in applications; given a utility function *u*, a weak*-closed, convex set *C* of priors, and a number $\alpha \in [0, 1]$, $f \succcurlyeq g$ if and only if

$$\alpha \min_{P \in C} \int u(f)\, dP + (1-\alpha) \max_{P \in C} \int u(f)\, dP \geq \alpha \min_{P \in C} \int u(g)\, dP + (1-\alpha) \max_{P \in C} \int u(g)\, dP,$$

thus, MEU corresponds to the case $\alpha = 1$. An axiomatization and further discussion can be found in Ghirardato, Maccheroni and Marinacci (2004).

Truman Bewley (2002) proposes an alternative approach to ambiguity. In both the CEU and MEU models, the DM responds to ambiguity by essentially evaluating different acts using different ‘decision weights’. Bewley suggests that, alternatively, the DM may simply be unable to rank certain acts in the presence of ambiguity; in other words, preferences may be *incomplete*. He axiomatizes the following partial decision rule: for a given utility function *u* and weak* closed, convex set *C* of priors, $f \succcurlyeq g$ if and only if

$$\forall P \in C, \int u(f)\, dP \geq \int u(g)\, dP.$$

For instance, in Ellsberg's three-colour-urn example, if the set *C* is chosen as above, the DM is unable to rank the acts f_r and f_g, as well as the acts f_{rb} and f_{gb}. Notice that preferences satisfy the full Independence axiom in Bewley's model: ambiguity manifests itself solely through incompleteness.

Ambiguity can also be modelled by introducing *second-order probabilities*. For instance, Klibanoff, Marinacci and Mukerji (2005) axiomatize the following decision rule:

$$\forall f, g \in F_0,\ f \succcurlyeq g \Leftrightarrow \int_{\Delta(S)} \phi\left(\int_S u(f)\, dP\right) d\mu \geq \int_{\Delta(S)} \phi\left(\int_S u(g)\, dP\right) d\mu,$$

where μ is a probability measure over the set $\Delta(S)$ of probability charges on the finite state space *S*, and φ is a ‘second-order utility function’. A notion of ambiguity aversion is characterized by concavity of φ. See also Ergin and Gul (2004).

Recent contributions aim at characterizing ambiguity without restricting attention to specific decision models, and without relying on functional-form considerations. Epstein and Zhang (2001) propose a definition of ‘unambiguous event’ that is based solely on preferences. Under suitable structural axioms, preferences over acts that are measurable with respect to such ‘subjectively unambiguous’ events are *probabilistically sophisticated* in the sense of Machina and Schmeidler (1992); this indicates that the proposed behavioural definition characterizes absence of ambiguity. See also Epstein (1999) for a related assessment of Schmeidler's definition of ambiguity aversion.

Ghirardato, Maccheroni and Marinacci (2004) note that, in models such as CEU and MEU, ambiguity manifests itself via violations of the Anscombe–Aumann Independence axiom. Thus, they propose to deem an act *f* ‘unambiguously preferred’ to an act *g* if $\alpha f + (1-\alpha) h \succcurlyeq \alpha g + (1-\alpha) h$ for all $\alpha \in (0, 1)$ and all $h \in \mathcal{F}_0$. They show that unambiguous preference admits a Bewley-style representation, characterized by a set *C* of priors which is a singleton if and only if the original preference is SEU. In light of this result, they suggest that the DM perceives ambiguity whenever *C* is not a singleton. See also Ghirardato and Marinacci (2002).

To highlight the differences between these definitions, consider a probabilistically sophisticated, non-SEU preference. According to the Epstein–Zhang definition, all events are subjectively unambiguous, whereas the Ghirardato–Maccheroni–Marinacci approach concludes that some ambiguity is perceived.

The modal preferences in the Ellsberg paradox constitute a violation of the *sure-thing principle*, which is arguably the centrepiece of Leonard Savage's (1954) axiomatization of SEU; indeed, this was a main focus of Ellsberg's seminal article. However, the sure-thing principle also plays a key role in ensuring that conditional preferences are well-defined and ‘dynamically consistent’; finally, it provides a foundation for Bayesian updating. Thus, since ambiguity leads to violations of the sure-thing principle, defining updating and ensuring a suitable form of dynamic consistency for MEU, CEU and similar decision models presents some challenges.

Gilboa and Schmeidler (1993) axiomatize Dempster–Shafer updating of capacities (cf. Dempster, 1968; Shafer, 1976) and ‘maximum-likelihood updating’ of multiple priors for ambiguity-averse CEU preferences. Prior-by-prior updating for MEU preferences is axiomatized in Jaffray (1994).

All these updating rules may lead to ‘dynamic inconsistencies’, that is, preference reversals: the ranking of two acts may be different before and after learning than a (typically ambiguous) event has occurred. Epstein and Schneider (2001) instead axiomatize a model of recursive MEU preferences by explicitly imposing dynamic consistency with respect to a

pre-specified filtration. The recursive formulation is especially convenient in applications; on the other hand, dynamic consistency imposes some restrictions on the set of MEU priors: see Epstein and Schneider (2001) for further discussion. Wang (2003) provides related results. Dynamic choice under ambiguity is currently an area of active research.

See Also

- decision theory in econometrics
- expected utility hypothesis
- measure theory
- non-expected utility theory
- risk aversion
- Savage's subjective expected utility model
- uncertainty

Bibliography

Aliprantis, C. and Border, K. 1994. *Infinite Dimensional Analysis*. Berlin: Springer Verlag.

Anscombe, F. and Aumann, R. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34, 199–205.

Bewley, T. 2002. Knightian decision theory: Part I. *Decisions in Economics and Finance* 25(2), 79–110.

Camerer, C. and Weber, Martin. 1992. Recent developments in modeling preferences: uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5, 325–70.

Casadeus-Masanell, R., Klibanoff, P. and Ozdenoren, E. 2000. Maxmin expected utility over savage acts with a set of priors. *Journal of Economic Theory* 92, 33–65.

Chew, H. and Karni, E. 1994. Choquet expected utility with a finite state space: commutativity and act-dependence. *Journal of Economic Theory* 62, 469–79.

Choquet, G. 1953. Theory of capacities. *Annales de l'Institut Fourier (Grenoble)* 5, 131–295.

Dempster, A. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B* 30, 205–47.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Epstein, L. 1999. A definition of uncertainty aversion. *Review of Economic Studies* 66, 579–608.

Epstein, L. and Schneider, M. 2001. Recursive multiple-priors. *Journal of Economic Theory* 113, 1–31.

Epstein, L. and Zhang, J. 2001. Subjective probabilities on subjectively unambiguous events. *Econometrica* 69, 265–306.

Ergin, H. and Gul, F. 2004. A subjective theory of compound lotteries. Mimeo. *Econometric Society 2004 North American Summer Meetings*, No. 152.

Ghirardato, P., Maccheroni, F. and Marinacci, M. 2004. Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* 118, 133–73.

Ghirardato, P., Maccheroni, F., Marinacci, M. and Siniscalchi, M. 2003. A subjective spin on roulette wheels. *Econometrica* 71, 1897–908.

Ghirardato, P. and Marinacci, M. 2002. Ambiguity made precise: a comparative foundation. *Journal of Economic Theory* 102, 251–89.

Gilboa, I. 1987. Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics* 16, 65–88.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Gilboa, I. and Schmeidler, D. 1993. Updating ambiguous beliefs. *Journal of Economic Theory* 59, 33–49.

Jaffray, J.-Y. 1994. Dynamic decision making with belief functions. In *Advances in the Dempster–Shafer Theory of Evidence*, ed. R. Yager, J. Kacprzyk and M. Fedrizzi. New York: Wiley.

Klibanoff, P. 2001. Characterizing uncertainty aversion through preference for mixtures. *Social Choice and Welfare* 18, 289–301.

Klibanoff, P., Marinacci, M. and Mukerji, S. 2005. A smooth model of decision making under ambiguity. *Econometrica* 73, 1849–92.

Luce, R. and Raiffa, H. 1957. *Games and Decisions*. New York: Wiley.

Machina, M. and Schmeidler, D. 1992. A more robust definition of subjective probability. *Econometrica* 60, 745–80.

Savage, L. 1954. *The Foundations of Statistics*. New York: Wiley.

Schmeidler, D. 1986. Integral representation without additivity. *Proceedings of the American Mathematical Society* 97, 255–61.

Schmeidler, D. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57, 571–87.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.

Siniscalchi, M. 2006. A behavioral characterization of plausible priors. *Journal of Economic Theory* 128(1), 91–135.

Wang, G. 2003. Conditional preferences and updating. *Journal of Economic Theory* 108, 286–321.

How to cite this article

Siniscalchi, Marciano. "ambiguity and ambiguity aversion." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000211> doi:10.1057/9780230226203.0036

The New Palgrave Dictionary of Economics Online

American Economic Association

Michael A. Bernstein

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Following its foundation, the American Economic Association (AEA) cultivated a unique professional visibility, struggling to establish its professional credentials and to demonstrate the usefulness of economists' ostensible skills. While celebrating the virtues of 'free markets', the AEA was itself shaped by government and collective action. In the latter half of the 20th century, the AEA promoted a 'New Economics' focused on macroeconomic intervention and regulation. However, these developments fostered a new generation of specialists with different views of public purpose, the appropriate role of government, and how professional economists could participate in the formulation and implementation of public policy.

Keywords

American Economic Association; American Finance Association; American Home Economics Association; business cycles; economics, definition of; home economics; Royal Economic Society; Union for Radical Political Economics

Article

The American Economic Association (AEA) was inaugurated by a miscellaneous group of scholars, university administrators and public figures, in September 1885, in the early stages of a sustained expansion in American academic life. Its original objectives of encouraging research, publications on economic subjects, and perfect freedom in economic discussions have been consistently maintained, sometimes not without difficulty given the disagreements among its members, and the persistent tension between the desire for scientific objectivity and non-partisanship on the one hand and the urge to make an impact on public policy on the other. This problem was especially acute during the AEA's early years, when economic questions were at the forefront of public discussion. A number of prominent American economists were then under attack, and some were dismissed from or forced out of their university posts

because of their opinions. However, under its first President, F.A. Walker, an internationally known figure who served for the first seven years, the AEA gradually lost some of its initial reformist tone and concentrated increasingly on more strictly scholarly issues. Unlike the British Royal Economic Society, which has frequently had a non-professional president, the AEA has invariably been dominated by academic economists, although in recent decades prominent government professional economists have occasionally held the office – for example, Alice Rivlin, the first woman President, in 1985.

Early challenges and strategies

While the AEA's contributions to economic knowledge through its periodicals – the *American Economic Review* (from 1911), the *Journal of Economic Literature* (from 1963), and the *Journal of Economic Perspectives* (from 1987) – and in various other ways are undeniable, its services to the profession have perhaps been unnecessarily restricted because of the heterogeneity of its constituency, which has always included a substantial proportion of non-academic members, and its commitment to non-partisanship. Thus, for example, the AEA's reactions to the conflicts and tensions in American society have been distinctly more cautious than those of some other learned societies, both within and outside the social sciences, with respect to academic freedom issues. However, in both world wars the AEA played a notable and constructive part by organizing professional expertise for government service, and by conducting open debates and issuing publications on the economic problems of war and peace. The Association has also since 1945 occupied a leading role in the internationalization of the economic profession. It has always been an ‘open’ society, with no significant membership restrictions, partly because of the objections to control by a limited elite or coterie. Consequently it has only occasionally had any direct influence on doctrinal developments in the field. Nevertheless, there have been periodic protests about the organization's unrepresentativeness and oligarchic management, a state of affairs reflecting the size, diversity, and geographical dispersion of its membership, which now stands at a little over 22,000 (including subscribers).

Under its charter of incorporation, the AEA committed itself to ‘the encouragement of economic research, especially the historical study of the actual conditions of industrial life’ as well as to ‘the encouragement of perfect freedom of economic discussion’. In particular, ‘the Association as such [took] no partisan attitude, nor commit[ed] its members to any position on practical economic questions’. While the formal organization was thus made distinct from the individual activities and convictions of its members, nevertheless the stresses and strains attendant upon the struggles over its initial establishment were, in its earliest years, never far from the surface. These anxieties in turn framed the process by which major decisions were ultimately made concerning AEA membership criteria, annual meetings, publications, and operational procedures; what is more, they made the Association's leadership particularly eager to seize upon whatever opportunities and circumstances within the public arena might enhance the prestige and sway of their field.

From its earliest days, the AEA faced certain difficulties associated with maintaining the separation between professional image and individual values. One of these involved continuing struggles over academic freedom issues, involving economists at certain educational institutions across the nation. The most celebrated of these, although by no means the only ones, were the cases of Richard Ely at the University of Wisconsin, Edward Bemis at the University of Chicago, and Edward Ross at Stanford. All three scholars had been accused in the 1890s, in different contexts and in various ways, of poisoning the

minds of their students with ideas and beliefs inimical to corporate interests and private wealth. Two of them, Ely and Ross, managed to bring their careers back from the brink of the abyss; Bemis was not as fortunate and, in the end, was condemned to oblivion. Whether in success or failure, however, the defence of colleagues placed in jeopardy for their political convictions and beliefs relied more on the *individual* support of powerful champions within the profession rather than on the *collective* imprimatur of the AEA.

Fretting over the size of their professional society was, for the early AEA leadership, one thing; firmly articulating the Association's *raison d'être* was something else. Declarations of purpose, no matter how frequently or even stridently made, served only to a point. It was in actual practice, and in the decisions that animated it, that the professional community of the AEA truly explained and revealed itself. No amount of enforcement of particular boundaries of expertise could substitute for the rigorous refinement of colleagues that would result from the inculcation of specific ways of doing the community's business. Whether self-consciously or not, Association members and officials were, from the earliest years of the 20th century, concerned to frame the interests, activities, and procedures of their group in ways that would, more powerfully and vividly than any set of membership standards might, decisively create and preserve the profession that it was their goal to foster.

Creating a professional journal was also quite challenging. With no debate among AEA secretariat colleagues, Davis Dewey, the founding editor of the *American Economic Review*, rejected a suggestion from the Theodora B. Cunningham in 1916 that the journal include 'a Women's Department of household economics'. Dewey's decision in this regard was thoroughly consistent with not one but two strategies of professionalization in early-20th-century America. On the one hand, it furthered the conscious effort of AEA founders to secure a distinctive place for economics as a scientifically grounded enterprise that avoided the lesser prestige of feminized occupations like 'home economics'. On the other, it actually dovetailed with efforts dating from 1900 to constitute home economics as a separate discipline in its own right. Women professionals eager to find in the home economics field the same authority and influence that their male counterparts struggled for in an array of other disciplines had worked assiduously to establish collegiate degree programmes, journals, and a national association – the American Home Economics Association (AHEA). Their very success made the 'defeminization' of economics, at the hands of professional communities like the AEA, rather easy.

In fact, the question of publication standards threatened to destabilize the general consensus about the desirability of creating the *American Economic Review* in the first place. Argument over the implementation of standards not only raised questions of intellectual freedom and openness but also drew attention back to the general and often delicate matter of the journal's purpose. Not simply value as to method and technique, but significance and appropriateness as to subject figured prominently in the deliberations of the AEA Executive Council regarding the new journal and the Association's annual meetings. These discussions continued for years and ultimately decades to come. They were, in fact, often intertwined, touching upon related concerns about professional status and prestige, scientific conduct and codes, and the boundaries (topical and methodological) of economics itself. Stoutly defining what economics was involved being clear-minded about what it was not. Prominent AEA members, at the very moment they were wrestling with the nature of a new publication for the Association, vigorously protested to President Seligman that sociologists be kept at bay from the annual meeting and even the quarterly itself. 'We have heard [the sociologists] so many times', Henry Carter Adams wrote Seligman in the spring of 1902, 'that we know absolutely what each one of the[m] will say

upon any subject'. When gathered in an annual convention, Thomas Carver argued, 'Economists would prefer to stick to the subject of Economists. [One] should especially doubt whether the members of [the] association would easily find a common ground of discussion with Miss [Jane] Addams or Mr. Felix Adler, admirable as these persons are and valuable as their work is. [One] should be afraid that there would be difficulty in trying to think in the same language.' The same, Carver believed, was true for the *Review*. He doubted very much if 'it would be wise to include much sociology, except such as has a distinctly economic coloring'. (All quotations of AEA minutes and correspondence are from the AEA Archives, Northwestern University Library, Box 8.)

Enforcing disciplinary boundaries, in both publication strategies and convention planning, also involved making precise decisions about the relationship between scholarly research and contemporary policy debate. With apparently little discussion or debate, the AEA Executive Committee formally chose in 1915 to exclude from the pages of the *American Economic Review* a 'department of current economic events'. Even if contemporary policy concerns found their way into the submissions to the Association's quarterly, the editors were determined 'that current economic questions…be treated by scholarly men and not left to the sensational magazine writer'. In some respects this was a curious position for the leadership to assume given the additional concern that the work of economists be made visible and influential in the world of public affairs. The notion that the *Review* should be 'a craftsman's tool' had, after all, animated a great deal of the effort of the editorial office from the earliest days. Maintaining a dispassionate, scholarly tone while encouraging a wide and even diverse readership was neither a simple nor an obvious task. Editor Davis Dewey put it well to the distinguished English theorist Francis Edgeworth in January 1911 when he wrote, 'We are trying to appeal to a somewhat varied membership who are interested in current questions. We do not, however, wish to be popular in a commonplace way, but shall endeavor to have our articles prepared by men of scholarly standards.' The problem of attracting 'a somewhat varied membership' while adhering to 'scholarly standards' that would guard against being 'popular in a commonplace way' was truly vexing.

The impact of national mobilizations and emergencies

The coming of the Great War stimulated the professionalization of AEA ranks. In the spring of 1914, the AEA secretariat fashioned a special opportunity to bring the potential benefits of professional economics expertise to the attention of federal officials. Not surprisingly, it involved concerns with the ways in which the United States Department of Agriculture (DOA) calculated and reported statistical data on the performance of the nation's farms. Cornell University Professor Allyn Young contacted the secretary of agriculture, David F. Houston, to express the fear of the AEA leadership that 'much of the statistical work … issu[ed] from government offices [wa]s of disgracefully poor quality'. He noted that the failures of the DOA in this regard were by no means unique. Clearly, 'many of the activities of [federal] government bureaus furnish[ed] statistical by-products that [c]ould be of the greatest usefulness'. There was a clear need, in Young's opinion, that these data be 'properly tabulated and published'.

By the interwar period, additional federal legislation also gave the AEA a unique opportunity to define itself. For example, passed by the Sixty-seventh Congress in 1923, the Classification Act provided for the categorization and grading of technical and professional employees in the civilian branches of the federal government. Like their counterparts in many other fields, the leaders of the American Economic Association succeeded in linking this particular federal effort to their own continuing pursuit of

professional cultivation. An early 1924 resolution of the AEA Executive Committee began steps to ‘secure the classification of the technical economists in the professional and scientific services’ of the federal government. The findings of a committee tasked to collate the results of this survey were reported to the Personnel Classification Board (of the US Civil Service Commission), the Committees on the Civil Service of the two houses of the Congress, and to the Executive Office of the President. In many respects the classification survey powerfully resonated with what had begun a decade earlier as part of the effort to support national mobilization for war. Yet here, in peacetime, it extended beyond the confines of an emergency canvass and became instead the basis of a continuing and ever more specific detailing of economics subspecialties. Indeed, for some older members of the profession the steps taken to stipulate as precisely as possible the expertise of individual practitioners could at times appear to narrow, and thereby adulterate, what the discipline as a whole had to offer. For most colleagues, however, that governmental needs melded so well with professionalizing strategies was cause for satisfaction rather than regret.

By the late 1930s, a segment of the AEA membership dissatisfied with the Association's perceived lack of attention to financial issues worked to create the American Finance Association (AFA). At the 1939 AEA Annual Meeting, the formal steps were taken to create the AFA. Although the Second World War slowed the evolution of the new organization, by 1942 the new journal *American Finance* appeared. It ultimately evolved into the well-known *Journal of Finance* just after war's end. Over 1,000 members populated the AFA ranks by the early 1950s.

In so far as a desire to distil professional opinion dated back to the early years of the Association's founding, it is not surprising to find that renewed interest along these lines emerged as economists turned their attention to planning for another war and its aftermath, and anticipating the role of economists in government during peacetime. During the Second World War the AEA leadership began deliberations ‘to [consider ways of] making the informed opinion of our membership more effective in matters of public policy’. Because the Association, by the terms of its charter, could take no partisan positions, the trio nevertheless believed that the ‘technical competence’ of members could be expressed on ‘matters of public importance’. This would require of course that ‘all academically respectable views on any posed controversial question be represented’ on committees formed to pronounce on policy matters.

While striving to adhere to its strictures against partisan endorsements, a task made all the more difficult in the highly charged politics of the immediate post-war era, the leadership of the American Economic Association turned its attention to engagement with seemingly more ‘objective’ needs of the national security state. In these efforts, their work was paralleled by that of colleagues already assigned to some of Washington's highest echelons. Over the course of the 1950s, for example, government economists made frequent visits to the military service academies, and to such institutions as the War College of the Air Force and the Industrial College of the Armed Forces (of the National Defense University, Fort McNair, Washington, DC) to discuss (and participate in conferences on) such matters as ‘mobilization of the national economy in the face of atomic attack’, ‘economic stabilization after attack’ and ‘domestic economies and their relation to national power’.

AEA officials also worked closely with colleagues on government duty to assist the national service academies in fully integrating an increasingly rigorous and operational discipline within their curricula. On behalf of the Armed Forces Institute, Secretary-Treasurer James Washington Bell coordinated the efforts of several scholars to oversee textbook selections in the field for cadets and midshipmen, thus

‘prov[iding] the Armed Forces of the United States with educational materials which [we]re in accord with the best civilian practices’ in economics as a whole. By the mid-1950s it had also become common for AEA functionaries to help designate particular professionals for work in special seminars on international organization and security convened by the transnational diplomatic and military alliance known as the North Atlantic Treaty Organization (NATO). It was a short step from these activities to involvement with the recruitment of undergraduate and graduate economics students for work within the now greatly expanded domain of the national security apparatus – including the Central Intelligence Agency (CIA).

The post-war and Cold War eras

Post-war reconstruction also brought the Association into the business of aiding professionals in devastated areas overseas. In addition to contributing free books and copies of the *American Economic Review* along with cash donations to scholarly libraries in Europe and East Asia, the AEA became involved in the revision of curricula and the rehabilitation and vetting of foreign faculties. American economists going overseas, on either official or personal tours, were asked by government authorities to check up on colleagues who had perhaps been imprisoned, wounded or otherwise victimized by German national socialism or Japanese imperialism. Letters to Association members from economists abroad often contained information regarding colleagues who either had or had not collaborated with the enemy. Efforts were made to raise money for the relief of those who had opposed fascism and militarism. A note from a German colleague to former AEA President Paul Douglas was forwarded to the Association offices because in it there was ‘a very valuable list of economists who either opposed Hitler or kept their honor clean’. American economists were now in a position not only to secure greater influence and prestige at home but also to reconstitute virtually from scratch the European and Asian branches of the guild.

The reconstruction of foreign scholarly libraries prompted the American Library Association (ALA) to ask professional societies to provide book lists in their fields to guide rebuilding efforts. AEA officials canvassed the membership for suggestions and ultimately provided such lists, with regard to economics, to the ALA. With such recommended titles as *Stalin, A Critical Survey of Bolshevism* and *Marxism: An Autopsy*, the ideological content of the library aid effort seems clear. This is of course hardly surprising. The point here is not that American economists would generally be loath to suggest books that extolled Marxism or Stalin – indeed, AEA members and the AEA leadership utterly failed to defend beleaguered colleagues victimized by the anti-communist hysteria stoked by McCarthyism – but that Allied victory had the added impact of giving them a great deal of influence on the future course of foreign scholarship in the field. If post-war reconstruction served to recast Europe and Asia in America's image, as some scholars have suggested, the representations of that process in the academic and intellectual world should not be overlooked.

Participation of the American economics profession in the emergent Pax Americana of the 1950s also expressed itself in a continuation and evolution of links between economists and the military–industrial establishment that had necessarily arisen in the 1940s. Economists of course participated both in the private sector and at the government level in the mobilization and allocation of resources for war. In addition, the profession became increasingly involved in establishing curricula at the nation's armed service academies on the economics of national security and defence. Defence-related research and

support of basic economics investigations by armed forces agencies became more and more common. Moreover, the emergence of wholly new aspects of the discipline – such as ‘linear programming’ and ‘input–output analysis’ – was inherent in the association of professional economics with the national security state. The AEA even helped the U.S. Information Agency in securing prominent and competent personnel to do radio broadcasts on economic subjects for the Voice of America.

Curriculum revision and reform was a project that lasted well into the 1950s. Two months before the opening of a second front in western Europe the Association Executive Committee asked that the new Committee on Undergraduate Teaching and the Training of Economists concern itself with ‘the long-run postwar period’. Ultimately, of particular interest to this committee with regard to the matter of undergraduate instruction were ‘problems of indoctrination [of students] as to social consciousness and professional responsibility’. Four months after the surrender of Japan, 160 college and university economics departments around the country received questionnaires from the AEA soliciting information on undergraduate instruction. By the autumn of 1950 the AEA secretariat initiated plans for a conference on social science teaching at the pre-collegiate and collegiate levels. At the same time, the Committee on Graduate Training in Economics began its work, seeking to formalize in detail the professional requirements for the Ph.D. degree. To this effort, the Rockefeller Foundation donated $16,000. When the committee transmitted its findings to university deans and presidents, return correspondence was grateful and enthusiastic. War-related agendas thus carried over into long-standing peacetime activities.

Interestingly enough, and not surprisingly, concerns with the content and delivery of economics curricula emerged directly from Second World War experience. Wartime efforts on behalf of the National Roster of Scientific and Specialized Personnel (NRSSP) had made the leadership of the American Economic Association both particularly sensitive and responsive to requests for information about the discipline and its specialists. Moving from a focus on calculating the profession's numbers and activities, as the NRSSP had requested, to a self-conscious assessment of teaching methods, course content, and educational performance standards was altogether understandable and clear-cut. AEA initiatives in this regard were only further stimulated by the desire of the Veterans Administration and related agencies to facilitate the re-entry of armed forces personnel to civilian life after the Second World War and the Korean conflict.

Defining what an economist was, and what he or she did for a living, was one thing; stipulating how an economist was to be trained, not to mention evaluating his or her professional skills, was something else. In a series of studies, the first of which was launched in 1949, with follow-ups taking place throughout the 1950s, AEA task forces conducted wide-ranging surveys of undergraduate and graduate curricula throughout the country. Of particular importance to these committees were the ‘opinions of leaders in graduate training’ in the field at the nation's foremost research institutions. Recognizing that ‘[t]he Association ha[d] a definite professional responsibility in this [regard]’, the Ad Hoc Committee on Graduate Training in Economics made its first report to the AEA Executive Committee late in 1950. Determined to guide universities in the establishment and maintenance of ‘good graduate program[s] in economics at various levels’, the committee particularly encouraged institutions to improve standards for the selection of incoming students, articulate precise objectives for advanced study in the field, and vet subject matter and course content with a view towards the rigorous training of new colleagues. Specifically, the committee believed that the ‘important tools’ in all graduate economics instruction were ‘mathematics, accounting, statistics, history, logic, scientific method, and foreign language’.

Not least of the historical forces that shaped the continuing evolution of the American economics

profession in the latter half of the 20th century was the unique prosperity the nation enjoyed throughout the 1950s and 1960s. If the application of a new learning to the management of a ‘mixed economy’ provided an exceptional opportunity for social scientific expertise to demonstrate its rigour and effectiveness, the context within which that display took place set the terms of both its practice and its success. Having proved its mettle in the extraordinary years of world wars, and having continued to do so in the early stages of what would be an even longer cold war, modern economic theory was now deployed in an altogether novel exercise: the pursuit and maintenance of full employment growth in peacetime. That, owing to history itself, the national economy was singularly well positioned for sustained expansion in the post-war period made that task all the more tractable.

Unlike any other industrialized nation in the world at the time, the United States met the 1950s with an economy not only physically intact but also organizationally and technologically robust. The demographic echoes of war set the stage for an acceleration in the rate of population growth, while the labour market effects of demobilization surprisingly sparked a rise in wages and incomes. Rapid and profitable conversion to domestic production was further stimulated by foreign demand – most vividly and poignantly emanating from those regions most devastated by the war itself – for the products of American industry and agriculture. As for international finance, the nation stood as creditor virtually to the entire world, and the dollar, both by default and by a multilateral agreement first reached by the Allied nations at Bretton Woods, had become a kind of numeraire to a newly emergent system of global commerce. With no small justification, the 1950s and 1960s came to be regarded as a golden age of American capitalism.

The era of the ‘New Economics’ and beyond

Macroeconomic management, demanding under any circumstances, was made substantially easier for a post-war generation that found itself the beneficiaries of historical circumstance. Farm from solving the cruel puzzle of idle capacity and widespread unemployment that had characterized the Great Depression, and unlike the challenge to rationalize allocation and maximize production in the emergency of war, the task that lay before American economists by the mid-1950s was both more straightforward and less difficult. More straightforward because, thanks to both the ‘Keynesian revolution’ in economic thought and the policy experience derived from mobilization and war, the relationship between individual market behaviour and aggregate outcomes was finally subject to systematic understanding. Less difficult because, given the sturdy rebound of the economy in the wake of the Second World War, there existed both the confidence (most especially exemplified by the moderate rates of return in the markets for Treasury bills and other government obligations) and the means (most vividly represented by rising income tax receipts) to realize fiscal spending targets with a minimum of redistributive implications.

So optimistic were politicians and the vast majority of economists concerning the effectiveness of stabilization policy techniques that it became fashionable by the early 1960s to speak of the ‘end of the business cycle’ and of the ability of policymakers to ‘fine-tune’ macroeconomic performance. In the *Economic Report of the President, 1965*, President Lyndon Johnson made it clear that he ‘d[id] not believe recessions [we]re inevitable’ (Council of Economic Advisers, 1965, p. 10). Similarly, in what was arguably the most influential economics textbook ever published, Paul Samuelson (1972, p. 250) wrote that his colleagues ‘kn[ew] how to use monetary and fiscal policy to keep any recessions that br[oke] out from snowballing into lasting chronic slumps’. He went on to claim that the business cycle was

thus a thing of the past. Expert knowledge buttressed by a healthy and resilient economy could now make the periodic deprivation and hardship once believed to be the inevitable consequence of the cycle truly a thing of the past.

Cultivating a politics of aggregate productivity and a discourse about sustained prosperity was not solely the result of professional self-assurance and self-promotion, nor was it simply the manifestation of a particular politician's (or a particular party's) strategy to procure votes. The focus on growth and accumulation so characteristic of the new economics of the post-war era represented as well a transformation in the nation's political culture that had been in the making for decades. For 19th-century convictions regarding the probity of thrift and self-improvement, mid-20th-century Americans had swapped a fascination with, and a virtual anxiety about, the individuation and comfort associated with consumption. Production was no longer an end in itself, nor could it alone provide meaning and dignity to one's life. Rather, it was the goods and services of the material world that afforded freedom and amenities, setting one's self off from others and liberating all from both the overt and the hidden injuries of class, ethnicity, and gender. What came to be known as the ‘economic growthmanship’ practised by a new social scientific elite was, on the one side, a particular aspect of a stage in the evolution of a professional community; on the other, it distilled, within a set of seemingly unassailable aspirations and beliefs, a society's unself-conscious embrace of an altogether new set of cultural ideals.

Within an economics of abundance and stability rested the ingredients of a prosperous commonwealth devoid of the class antagonisms and struggles over normative values that were a threat to both the legitimacy of social scientific policymaking and social tranquility and political cohesion. If an ‘emphasis on an ever-growing pie, rather than on slicing up a given pie in a new way, [wa]s well designed … to attract widespread support’ for particular policies (Tobin, 1966, p. 42), it was also true that the depiction of the economy as a kind of positive-sum game from which all could benefit independent of their *relative* shares in particular outcomes was an essential part of the political-economic ideology of post-war America from the time of Truman's Fair Deal through that of Lyndon Johnson's Great Society, up to and including the early stages of Richard Nixon's New Federalism. Their specific analytical differences aside, virtually all mainstream American economists both embraced and relied upon this ‘depoliticization’ of the marketplace in their determination to separate positive economic ‘science’ from normative assertions. So long as the profession could retain this image of its work as a calculation of optimal means to a given end rather than the comparison of different and possibly incompatible goals, its claims to the authority and influence devoutly sought since the late 1890s were secure. As soon as that archetype was jettisoned or challenged, modern economics would find itself in a world, not of rigour and logic, but rather of ideological belief and political power.

Indeed, in December 1968 the Union for Radical Political Economics (URPE) held its first national conference in Philadelphia. This was done in opposition to the AEA's Annual Meeting in Chicago, which URPE interpreted as an endorsement of that city's violent response to anti-war demonstrations that summer. The AEA Executive Committee, chaired by then AEA President Kenneth Boulding, concluded that moving the Meeting would have violated the Association's policy of political neutrality. A year later, an activist disrupted the AEA Annual Meeting by reading a statement, at a plenary session, denouncing the Association for ‘perpetuating professionalism, elitism, and petty irrelevance’. This led to a mass walk of ‘radical economists’. In partial response to these insurgencies from within the ranks, the AEA established a Committee on the Status of Minority Groups in the Economics Profession (CSMGEP) in 1968 – and, by 1971, a Committee on the Status of Women in the Economics Profession

(CSWEP) and a working group on the status of minorities. The social change and turmoil of American society in the Vietnam War era had come home to the AEA itself.

In the mid-1980s, concerns regarding the training of new generations of economists came to the fore in AEA deliberations. At a National Science Foundation symposium held late in 1986, many participants argued that graduate curricula in economics had become exceedingly esoteric and abstract, of little use in the resolution of contemporary economic problems. A Commission on Graduate Education in Economics (COGEE) was subsequently charged to study the problem. It issued a report in 1991 that identified a number of problems in the profession such as a lack of focus on the inculcation of applied research skills, untoward emphasis on mathematics and axiomatic reasoning instead of analysing institutions and historical change, inadequate attention to the training with respect to communication and writing skills, an absence of creativity, and excessive emphasis on conformity and homogeneity in professional discourse. The COGEE report was so controversial that it was never accepted as an official AEA document.

Over a century ago American scholars eager to understand the economic world in which they lived embraced a project of both theoretical and social import. In doing so, they yoked the insights of an intellectual revolution in the ways social scientists understood human behaviour in commercial settings to a specific agenda of professional advancement. A late-19th-century transformation in economic thought afforded these investigators a powerful and versatile set of tools with which to situate human rationality at the centre of a remarkable and immensely influential human institution – the marketplace. A ‘science’ of individual behaviour and social organization was thus established, the implications of which played no small part in the creation of a respected and ultimately quite accomplished community of professional experts – as exemplified by the AEA.

But an authoritative community does not, precisely because it cannot, subsist on its own. American economists were most eager to place their skills at the service of the state. Here history proved both a blessing and a curse, for the profession's great achievements of the 20th century, especially but not solely during years of global conflict and war, were also paralleled by failures and betrayals emanating from the same source. Indeed, it would be these negative moments in the century-long progress of their self-realization that would drive economists and their discipline farther and farther from engagement with the affairs of state in favour of an increasingly introverted and surprisingly opaque discourse. At the same time, eager like most professionals to retain an influence and visibility in public affairs that would cultivate a continued appreciation of their virtues and skills, later generations of economists would make themselves – whether consciously or not – useful servants of those, in both the political and the commercial worlds, who had an altogether different view of public purpose and of the appropriate role of government.

Bibliography

AEA (American Economic Association). Archives, Northwestern University Library, Evanston, IL., and Secretary's Office, Vanderbilt University, Nashville, TN. *Records of the American Economic Association*.

Bernstein, M.A. 2001. *A Perilous Progress: Economists and Public Purpose in Twentieth-Century America*. Princeton: Princeton University Press.

Coats, A.W. 1960. The first two decades of the American Economic Association. *American Economic Review* 50, 555–74.

Coats, A.W. 1964. The American Economic Association, 1904–1929. *American Economic Review* 54, 261–85.

Coats, A.W. 1985. The American Economic Association and the economics profession. *Journal of Economic Literature* 23, 1697–728.

Coats, A.W. 1993. *The Sociology and Professionalization of Economics*, 2 vols. New York: Routledge.

Council of Economic Advisers. 1965. *Economic Report of the President, 1965*. Washington, DC: U.S. Government Printing Office.

Dorfman, J. 1936. The founding and early history of the American Economic Association. *American Economic Review* 26, 141–50.

Dorfman, J. 1949, 1959. *The Economic Mind in American Civilization*, vols 3–5. New York: Viking Press.

Furner, M.O. 1975. *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905*. Lexington: University Press of Kentucky.

Ross, D. 1991. *The Origins of American Social Science*. New York: Cambridge University Press.

Samuelson, P.A. 1972. *Economics: An Introductory Analysis*. New York: McGraw-Hill.

Tobin, J. 1966. *National Economic Policy: Essays*. New Haven: Yale University Press.

U.S. Council of Economic Advisers. 1965. *Economic Report of the President: 1965*. Washington, DC: US Government Printing Office.

How to cite this article

Bernstein, Michael A. "American Economic Association." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000089> doi:10.1057/9780230226203.0037

The New Palgrave Dictionary of Economics Online

American exceptionalism

Louise C. Keely

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘American exceptionalism’ refers to significant differences between the United States and Western Europe, first identified by European commentators in the 19th century, including the circumstances surrounding the founding and settlement of the United States, as well as a concept of nationhood based on immigration rather than a shared history. Economists have contributed in important ways to the documentation and evaluation of exceptionalism's economic effects. One important example of this research is on contemporary differences in social policy between the United States and Western Europe.

Keywords

American exceptionalism; crime; de Tocqueville, A.; equality of opportunity; fertility; higher education; individualism; inequality; international migration; productivity; racism; redistribution; religion; slavery; socialism; trade unions; voluntarism; Western Europe; workfare

Article

The term ‘American exceptionalism’, which has been current among scholars since Alexis de Tocqueville coined it, captures the idea that America is different in important ways from Western European countries. This exceptionalism is, at first glance, surprising given that the United States was initially settled and governed by persons from Europe and that in many ways the two regions appear relatively similar. The term suggests a set of reasons for the differences in institutions and individual choices in the realms of politics, economics, and social interactions. (For some insightful and broad discussions of American exceptionalism, see Lipset, 1996; Shafer, 1999.)

Comparing America with Western Europe is somewhat arbitrary. The attention paid to American exceptionalism does not suggest that other countries are not also exceptional. Indeed, other examples of exceptionalism have been studied by social scientists.

Interest in the United States is due in part to its economic and military power. Simply put, the US

government and economy exert a significant influence on all countries, including those of Western Europe. But there is an historical reason for the comparison with Western Europe. Europeans settled and governed the region that became the United States of America. It was Europeans who in the 19th century visited and wrote about the United States, comparing it with their native lands. De Toqueville is the best known, but he is only one of several Europeans who were interested in what they saw as a profound contrast between the United States and Western Europe.

It is useful to define American exceptionalism in terms of origins rather than consequences. Political, economic and cultural outcomes, whether observed today or in the 19th century, are endogenous. However, there may be circumstances distinguishing the United States from Western Europe that can be treated as fundamental, or exogenous, to the United States as a sovereign state. Those circumstances may have led to the differences in outcomes observed in the 19th century and today.

Before the American Revolution that began in 1776, the British governed the colonies that came to constitute the original United States. The constitution of the United States can be understood as a product of both the trauma of the revolution and the fact that 13 geographical areas, with distinct identities, were creating a single federal government. Moreover, the framers of the constitution were themselves diverse not only in place of origin but also in social and economic background (Mee, 1987).

The constitution contains features reflecting a certain distrust of centralized public authority. The increase in popular political participation beyond that which existed under colonial administration, the checks and balances across the three branches of government, and the restrictions on the powers of the federal government are prominent indications of this concern.

Europeans, many with a specific religious agenda, initially settled the area that became the United States. They aimed to create a society directed by divine providence. These settlers faced unusual circumstances in modern history, having the opportunity not only to establish a government largely from scratch but also to settle a large geographic area that was either uninhabited or inhabited by people they could displace, albeit sometimes with difficulty.

The historical circumstance of the United States as a state whose citizens' families came from other countries within recent memory led, in part, to a notion of nationality that was flexible from the beginning. What it is to be American has never, with the important exception of slaves who were not treated as full citizens, been dependent on ethnic background or common historical circumstances. This is not to deny that racism or ethnic prejudice have existed in the United States, but rather to say that what it is to be *American* has never been predicated on a particular origin or history.

This notion of nationality lent itself to the United States' openness to immigrants from many countries, until recently mostly Europeans. Some immigrants came to the United States to escape political or religious persecution, such as the Jews during the pogroms of the late 19th century and during and after the fascist regimes that held sway in Europe in the first half of the 20th century. However, many more came in pursuit of economic opportunity. Some immigrants, such as the Irish in the mid-19th century, faced terrible economic circumstances in their home countries. Others chose to emigrate under less dire constraints.

Across these diverse circumstances of immigration, it is generally the case that immigrants to the United States were self-selected into this group. The important exception to this self-selection is the immigrants from Africa and the Caribbean who were brought as slaves to the United States.

While self-selecting immigrants left their countries of origin for a variety of reasons, they would all have believed that in the United States their lives would be better in economic, political or religious terms. By

seizing the opportunity to become American, they could lead better lives (loosely defined). This possibility is attributable to exogenous circumstances: the physical expansiveness of the United States and the related expandable notion of American nationality. The populations who chose to move to the United States also did so in large part because they believed that self-determination was possible in the United States.

Such self-determination is part of the ideology on which (rather than on a common history) the United States was founded, and subscription to which makes one American. That ideology includes a set of values and institutions that are immediately familiar as distinctly American. Americans are viewed (and are thought to view themselves) as relatively distrustful of public authority and as embracing self-reliance. Broadly speaking, they subscribe to the ideals of equal socio-economic opportunity (as distinct from equality in outcomes), a classless society, and an inclusive democratic process. American institutions are relatively fragmented and public services are generally viewed to be less comprehensive than in countries with similar per capita incomes. Americans are more religious than Europeans. The concept of American nationality is relatively inclusive.

Why should American exceptionalism matter to economists? What role does economics have to play in understanding the consequences of American exceptionalism?

At least three distinct avenues of enquiry are of interest to economists. The first is positive: to document outcomes that may be attributable to American exceptionalism. The second is evaluative: to examine whether the exceptional circumstances under which the United States and its citizenry were constituted have led to differences between both the institutions and the values and beliefs (or culture) of Americans and those of Western Europeans. (A substantial political science literature debates the relative importance of institutional differences and cultural differences in defining American exceptionalism. The present author finds that discussion unclear, and thinks that it is more useful to view both types of differences are outcomes of exceptionalism rather than manifestations of it. That is, both types of differences may exist and are not mutually exclusive.) The third avenue of enquiry is normative: given evidence of exceptionalism, the task is to examine the context in which economic policies in the United States are to be designed and evaluated relative to Western Europe.

Existing research focuses on American-European differences in political, cultural, and economic outcomes, and asks questions including those in the following non-exhaustive list:

- Why was there not a socialist movement in the United States? (Jacoby, 1991; Lipset and Marks, 2000; Voss, 1993)
- Why have labour unions been weaker in the United States than in Western Europe? (Currie and Ferrie, 1995; Freeman, 1994; Jacoby, 1991; Voss, 1993)
- Why do Americans publicly redistribute income less than Europeans do? (Alesina and Glaeser, 2004; Benabou and Tirole, 2004; Shafer, 1991)
- Why do Americans perceive a higher probability of socio-economic mobility within and across generations than those in Western Europe? (Keely, 2005a)
- Why is the US higher education system larger than those in European countries? (Shafer, 1991)
- Why is there more violent crime in the United States? (Shafer, 1991)
- Why is productivity in the United States higher than in Western Europe? (Abramoviz and David, 1994; Gordon, 2002; Romano, 1993)

- Why do Americans participate in volunteer activity more than Western Europeans? (Lipset in Shafer, 1991; Lipset, 1996)
- Why did the institution of slavery persist in the United States long after it disappeared from Western Europe? (Shafer, 1991)
- Why are Americans more religiously observant than Western Europeans? (Shafer, 1991)
- Why is fertility higher in the United States than in Western Europe? (Keely, 2004; 2005b)
- Why has the United States been able to assimilate immigrants at levels well beyond those of Western Europe? (Glazer, 1999)

Proposed answers to these questions have one common element: American exceptionalism. These issues are all directly or indirectly related to economic policy, and pose questions that economists' tools can help to answer. Consider for example the question: Why do Americans publicly redistribute income less than Europeans do? Economists have recently tried to answer this question.

A first step is to document the differences in redistribution. OECD data indicate that, while public spending on social services amounted on average to 24 per cent of GDP in Western European countries, in the United States it amounted to 15 per cent. In the United States private social spending as a share of the total in 1995 is reported by the OECD to be 41 per cent, while for European Union countries it varied from 1.5 per cent (Spain) to 16.9 per cent (the United Kingdom) (OECD, 2005).

Second, how can this difference be attributed to American exceptionalism rather than to some other source? Identifying the effects of exceptionalism as such is extremely difficult. Competing hypotheses about the same outcome can be observationally equivalent. However, the models that lead to the same predicted outcome may also contain secondary predictions that do vary across models. That variation may be exploited to compare hypotheses. One suggested approach has been predicated on the higher level of ethnic heterogeneity in the United States than in Western Europe. The institution of slavery, which led to the existence of a minority of citizens of African origin, and the flow of ethnically varied immigrants into the United States have been attributed to American exceptionalism.

Heterogeneity itself doesn't explain why there is less income redistribution. Some authors have proposed that heterogeneity may matter in terms of its interaction with preferences. (This hypothesis has been proposed by Alesina, Baqir and Easterly, 1999, and Luttmer, 2001. See Keely and Tan, 2005, for related discussion.) The assumption regarding preferences is that agents experience disutility when they observe people who differ from them in some salient dimension such as race to be more likely recipients of public income redistribution. Such preferences capture a notion of racism.

Racism is not a feature or direct consequence of American exceptionalism as I have defined it. Nor are norms regarding interracial interactions exogenous or unchanging variables. Interactions between, and socio-economic outcomes across, racial and ethnic groups in the United States have changed enormously (though perhaps still not enough) over the past century. Therefore, this preference-based hypothesis regarding different levels of income redistribution in the United States and Europe is only partially based on an observation directly attributable to American exceptionalism.

An alternative hypothesis that relies more squarely on exceptionalism, rather than on other cultural or political assumptions, is as follows. People face uncertainty about future income and whether they will be net beneficiaries of income redistribution policies. In order to form the expectations that are necessary to determine preferred income redistribution policy, people may use information about others

who are similar to them in ways that are relevant to income determination. In a society where racial or ethnic characteristics are correlated with income, race and ethnicity can be a factor determining similarity. If the size of the minority group (in this case, blacks) is sufficiently large and/or the difference in the groups' income distributions is sufficiently large (in some well-defined way), then it can be the case that whites, who have higher average income, are less likely to be in favour of income redistribution than are blacks.

This hypothesis relies on three factors that have been traced directly to American exceptionalism: (*a*) the ethnic heterogeneity of agents; (*b*) income inequality linked to the legacy of the institution of slavery (given the presence of relatively large amounts of arable land); and (*c*) the focus on individualism rather than communal obligation.

Both hypotheses lead to a prediction that the United States has lower levels of redistribution than Western European countries. How can competing hypotheses be evaluated? As suggested above, one strategy is to look for secondary and testable predictions that differ across hypotheses. While there is a history in the United States of racism connected to whites and blacks, there is also a history of racism against other ethnic groups such as Asians and Hispanics. Certainly there is a widely recognized ethnic distinction between those groups on the one hand and people of European descent on the other. If differences in income redistribution preferences are due to racism, then it should be the case that exposure to ethnic heterogeneity of these types should also lead to stronger opposition to redistribution overall.

In contrast, if differences in redistribution preferences stem from differences in income distributions conditional on ethnic group, then an effect of heterogeneity might not be uniform. For instance, if the conditional distribution of whites and Asians is not statistically significantly different, then income redistribution preferences are predicted to be lower in areas with more heterogeneity in the white-Asian dimension only under the first 'racism' hypothesis.

The third way in which American exceptionalism matters to economists is its impact on political economy parameters. Every public authority is policy constrained, for instance by cultural values and economic circumstances. America was founded on an ideology that, it has been argued, persists. While its details and interpretation may change, its essence is constant. Any normative statement regarding the political economy of the United States should, in the face of strong evidence of American exceptionalism, take account of those constraints. More specifically, one of the ways in which American exceptionalism manifests itself and has been summarized is the claim that individualism and anti-statism lead to a notion of egalitarianism based on opportunity rather than outcomes.

In this light, it is completely unsurprising that the United States has a smaller welfare state than those of Western Europe. Moreover, the types of welfare reform that have been instituted since 1995 and the rhetoric used to promote them are also consistent with American exceptionalism. Welfare is now sometimes called workfare; there is a push to move welfare towards a policy that provides opportunity through job training and work rather than providing a guaranteed outcome through direct transfers. Private involvement in a publicly administered welfare programme also seems more politically feasible than a purely public model as in Western Europe.

American exceptionalism is an old idea. In his now famous 1630 'City on a Hill' speech, John Winthrop spoke thus of the newly settled land:

> [W]ee shall finde that the God of Israell is among us, when tenn of us shall be able to

> resist a thousand of our enemies, when hee shall make us a prayse and glory, that men shall say of succeeding plantacions: the lord make it like that of New England: for wee must Consider that wee shall be as a Citty upon a Hill, the eies of all people are uppon us….

Economists have a perspective and set of skills to contribute towards understanding the extent to which American exceptionalism exists and its implications for Americans and people in other countries.

See Also

- equality of opportunity

Bibliography

Abramovitz, M. and David, P. 1995. Convergence and deferred catch-up: productivity leadership and the waning of American exceptionalism. In *Growth and Development: The Economics of the 21st* Century, ed. R. Landau, T. Taylor and G. Wright. Stanford: Stanford University Press.

Alesina, A. and Glaeser, E. 2004. *Fighting Poverty in the US and Europe: A World of Difference*. Oxford: Oxford University Press.

Alesina, Al., Baqir, R. and Easterly, W. 1999. Public goods and ethnic divisions. *Quarterly Journal of Economics* 114, 1243–84.

Benabou, R. and Tirole, J. 2004. Belief in a just world and redistributive politics. Working paper. Princeton, NJ: Princeton University.

Currie, J. and Ferrie, J. 1995. Strikes and the law in the US, 1881–1884: new evidence on the origins of American exceptionalism. Working paper. UCLA.

Freeman, R. 1994. American exceptionalism in the labor market: union-nonunion differentials in the United States and other countries. In *Labor Economics and Industrial Relations: Markets and Institutions*, ed. C. Kerr and P. Staudohar. Cambridge, MA: Harvard University Press.

Glazer, N. 1999. Multiculturalism and American exceptionalism. In *Multicultural Questions*, ed. C. Joppke and S. Lukes. Oxford: Oxford University Press.

Gordon, R. 2002. Technology and economic performance in the American economy. Working Paper No. 8771. Cambridge, MA: NBER.

Jacoby, S. 1991. American exceptionalism revisited: the importance of management. In *Masters to*

Managers: Historical and Comparative Perspectives on American Employers, ed. S. Jacoby. New York: Columbia University Press.

Keely, L. 2004. Perceived return to human capital in fertility. Working paper. University of Wisconsin-Madison.

Keely, L. 2005a. Mobility beliefs and socioeconomic mobility. Working paper. University of Wisconsin-Madison.

Keely, L. 2005b. Intergenerational socioeconomic mobility and fertility: a new look at an old question. Working paper. University of Wisconsin-Madison.

Keely, L. and Tan, C.M. 2005. Understanding preferences for income redistribution. Working paper. Medford, MA: Tufts University.

Lipset, S. 1996. *American Exceptionalism: A Double-Edged Sword*. New York: W.W. Norton.

Lipset, S. and Marks, G. 2000. *It Didn't Happen Here: Why Socialism Failed in the United States*. New York: W.W. Norton.

Luttmer, E. 2001. Group loyalty and the taste for redistribution. *Journal of Political Economy* 109, 500–28.

Mee, C. 1987. *The Genius of the People*. New York: Harper and Row.

OECD (Organisation for Economic Co-operation and Development). 2005. *Social Expenditure Database*. Paris: OECD.

Romano, R. 1993. *The Genius of American Corporate Law*. Washington, DC: AEI Press.

Shafer, B., ed. 1991. *Is America Different?* Oxford: Oxford University Press.

Shafer, B. 1999. American exceptionalism. *Annual Review of Political Science* 2, 445–63.

Voss, K. 1993. *The Making of American Exceptionalism: The Knights of Labor and Class Formation in the Nineteenth Century*. Ithaca, NY: Cornell University Press.

Winthrop, J. 1639. City upon a Hill. Online. Available at http://www.mtholyoke.edu/acad/intrel/winthrop.htm, accessed 30 August 2005.

How to cite this article

Keely, Louise C. "American exceptionalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000215> doi:10.1057/9780230226203.0038

The New Palgrave Dictionary of Economics Online

Amoroso, Luigi (1886 – 1965)

Giancarlo Gandolfo

From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Amoroso, Luigi; consumer equilibrium; Pareto efficiency

Article

A mathematician by training (at the Normale, Pisa), Amoroso was assistant professor of mathematics in Rome, then professor of financial mathematics in Bari, but soon turned to economics, which he taught from 1921 in Naples and then Rome. He was a fellow of the Econometric Society.
Leaving aside his contributions to pure mathematics (e.g. 1910), financial mathematics (e.g. 1921a), statistics (e.g. 1916), demography (e.g. 1929), four books (1921b, 1938, 1942, 1949) well summarize his contributions to economics, also contained in over 100 articles.
Inspired by Pareto, his mathematical background led him to develop the analogy between pure economics and classical mechanics: the principle of minimum (use of scarce) means is the equivalent of the principle of least action. He also saw analogies between Heisenberg's uncertainty principle and economic phenomena, but did not develop this idea. His existence and uniqueness proof (1928) of a meaningful solution to the system of equations defining consumers' equilibrium is the first modern treatment of existence and uniqueness problems in economics.
Amoroso stressed the need to analyse all optimum conditions in a dynamic context: for example, the consumer maximizes a function under the balance constraint expressed as a differential equation; the problem is solved by applying the calculus of variations. He thus derived the extension of Pareto's static optimum conditions to a dynamic context. By considering the market determination of prices and introducing relationships between inventories and prices, he obtained systems of integro-differential equations capable of causing cycles around a trend, thus giving an explanation for crises and secular movements.

Selected works

A full bibliography of Amoroso's works up to 1959 and an evaluation of his scientific contributions by

various authors is contained in: Onoranze al Prof. Luigi Amoroso, *Annali dell'Istituto di Statistica*, Vol. 30, Università di Bari, 1959.

1910. Sulla risolubilità della equazione lineare integrale di prima specie. *Rendiconti della Accademia dei Lincei*. Classe di Scienze fisiche, matematiche e naturali. Nota presentata dal socio Castelnuovo, nella seduta del 16.1.1910, Roma.

1916. Contributo al metodo delle minime differenze. *Giornale degli Economisti e Rivista di Statistica*, July, 50–86.

1921a. *Lezioni di matematica finanziaria*, Vol. 1 (1921), Vol. 2 (1923). Naples: Majo.

1921b. *Lezioni di economia matematica*. Bologna: Zanichelli.

1928. Discussione del sistema di equazioni che definiscono l'equilibrio del consumatore. *Annali di Economia*, March, 31–41.

1929. L’equazione differenzialle del movimento della popolazione. *Rivista Italiana di Statistica*, April, 151–7.

1935. La dynamique de la circulation. *Econometrica* 3(4), October, 400–10.

1938. *Principi di economica corporativa*. Bologna: Zanichelli.

1942. *Meccanica economica*. Città di Castello: Macri.

1949. *Economia di mercato*. Bologna: Zuffi.

How to cite this article

Gandolfo, Giancarlo. "Amoroso, Luigi (1886–1965) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000049> doi:10.1057/9780230226203.1876

The New Palgrave Dictionary of Economics Online

amortization

Charles R. Hulten

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

amortization; balloon loan; depreciation; investment; asset value

Article

'Amortization' is an accounting term meaning the allocation of a cost to several time periods. The term is derived from the Latin word for 'death' and literally means to 'kill off' the liability. Debts which are paid off gradually are said to be amortized.

The term is also applied to the depreciation costs of the cost of certain assets which are used up in producing income. Amortization in this second sense is illustrated by the following example (Table 1).

A firm spends $10,000 to invent and patent a new product which is expected to yield revenue (net of operating expenses) of $5,000 in the first year of production, $2,000 in each of the next three years, and $1,500 in the fifth year (see column (3) of Table 1). The product is assumed to become obsolete at the end of five years and to generate no additional revenue. The patent thus becomes valueless at that time.

Amortization of hypothetical asset

(1)	(2)	(3)	(4)	(5)	(6)
End of:	Outlay	Net revenue	Present value*	Loss in value	Profit
yr 0	$10,000	0	$10,000	0	0
yr 1	0	$5,000	$6,000	$4,000	$1,000
yr 2	0	$2,000	$4,599	$1,401	$599
yr 3	0	$2,000	$3,058	$1,541	$459
yr 4	0	$2,000	$1,364	$1,694	$306
yr 5	0	$1,500	0	$1,364	$136

*Present value of remaining net revenue calculated using discount rate of 9.992%.

The present value of the net revenue stream associated with the invention is initially $10,000 at an

approximate ten per cent rate of discount. However, the present value of the remaining net revenue falls to $6,000 at the end of the first year, to $4,599 at the end of the second year, to $3,058 and $,1364 at the end of the third and fourth years, and to zero at the end of the product's useful life (see column (4)). This implies that the original $10,000 investment has been eroded by $4,000 at the end of the first year, $1,401 in the second year, and so on (see column (5)). In considering how much profit is earned in the first year, the loss in the value of the investment must be subtracted from revenue in order to keep the original value of the investment intact. Thus, profit in the first year is $1,000, or ten per cent of the original investment. Inspection of columns (4) and (6) reveals that the ratio of profit to remaining present value in the previous year is always ten per cent.

If, on the other hand, the reduction in value is not recognized as a cost, one would erroneously conclude that the investment yielded $12,500 over the life of the asset (the sum of column (3)) rather than $2,500 (the sum of column (6)). However, the value of the investment would have fallen from $10,000 to zero. To avoid a misstatement of profit for tax and financial accounting purposes, investors are allowed to amortize the cost of the asset over its useful life. A pattern of amortization that matches the actual yearly loss in asset value is usually termed ‘economic depreciation’, although this typically (but not always) applies to tangible capital like plant and equipment, while ‘amortization’ is often used in the context of intangible assets. The actual loss in value is often hard to measure and, in practice, reasonable assumptions about useful asset life and about the pattern of value loss are used (for example, the straight-line and declining-balance patterns).

The graduation write-off of a debt is another context in which the term ‘amortization’ is frequently used. The level-payment home mortgage is, for example, a common type of amortized loan. In the level-payment mortgage, the sum of the interest and principal payments is constant. During the early life of the loan, the bulk of this constant (or ‘level’) payment is for interest on the outstanding balance of the loan. The proportion of the level payment allocated to the repayment of principal gradually increases as time goes by, since interest is paid on the outstanding balance of the loan. In the fully amortized loan, the sum of the period-by-period repayments of principal over the life of the loan is equal to the original value of the debt.

This type of arrangement may be contrasted with the case of the ‘balloon’ loan, in which the entire principal is repaid at the termination date of the loan. Loans may be a mixture of the two types: amortization of part of the principal with a balloon payment equal to the unamortized balance.

See Also

- capital measurement
- depreciation

How to cite this article

Hulten, Charles R. "amortization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000095> doi:10.1057/9780230226203.0039

The New Palgrave Dictionary of Economics Online

analogy and metaphor

Rom Harré

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

analogy; metaphor; models; simile

Article

We say that something A is analogous to something B if, in some relevant respect, A is similar to but not identical with B. This is the basic relation upon which the use of analogy in various kinds of reasoning depends. We speak of reasoning by analogy when on the basis of some similarity which we discern between two things or processes or properties, or what you will, we infer some other similarity. Reasoning by analogy is a special case of inductive reasoning since we must be wary of the possibility that the further similarities which are presupposed in our inference may not actually obtain. Like all inductive inference reasoning by analogy is stepping from the known to the unknown. Clearly, then, analogical reasoning is not demonstrative or deductive.

A more refined analysis of the structure of the analogy can be made by distinguishing between those respects in which the analogues are similar, called the positive analogy, those respects in which they are different, called the negative analogy, and those respects in which we are unsure whether the property in question marks a similarity or a difference – the neutral analogy (Hesse, 1963). Once we have introduced the idea of neutral analogy the relation between the analogues is no longer symmetrical. If we think of analogy simply in terms of similarities and differences then if A is similar to B, B is similar to A, and if A is dissimilar to B, B is dissimilar to A. It does not matter which of A and B we say is analogous to which. But once we introduce the idea of neutral analogy we are obliged to decide which of the items under comparison is the one from which our reasoning will take a start and usually this decision is dependent on which of A or B we are confident we know. For example, if we argue that an illness is analogous to the invasion of a country by a hostile army, as van Helmont proposed in the 17th century, it seems reasonable to take the invasion by the hostile army as the term about which we can in principle know a great deal and the cause of illness as the term about whose properties we are less certain. In reasoning by analogy, then, about the cause of disease, the idea of an invasion is the given term and the illness is the unknown. We can then take the known properties of invasions and armies and

set out on an experimental programme to decide how many properties similar to them are to be found in the causes of disease. Thus: ‘Soldiers are organisms’, ‘Are the causes of disease micro-organisms?’ The logic of analogy then consists in picking out sets of properties and making comparisons between the members of the one set and of the other.

In judging the force of an analogy we must have some way of deciding which properties are important and which are not. If two things are similar only in unimportant or inessential ways and differ in other respects, then we generally take the analogy between them to be weak. Unlike deductive reasoning, analogy is, therefore, highly sensitive to context and to the interests of whoever is making use of it. It can hardly be said that there is anything intrinsic about a property which makes it important. Rather its importance depends upon the context and interests of the user. Furthermore, we need also to assume that we can make some sort of quantitative assessment of the degrees of similarities and differences between the analogues and this may be quite difficult to do in any principled way.

I have described the relation of analogy in terms of concrete relations of similarity and difference between the properties of analogous things. However, there are important linguistic phenomena which are in some ways like an analogy. The most obvious is simile. When we use that figure of speech we explicitly invite a comparison between the referents of the terms between which the simile is drawn by reference to likenesses. We tacitly assume that we draw a simile only where there are also differences. There are plenty of literary examples to illustrate this relationship.

The analogy relation seems to have another realization in language in metaphor. In a metaphorical use of a term an expression is employed in a novel context. Words which are customarily used for discussing one kind of subject matter, are used to describe some other. Some have said that in metaphor the sense of a word is displaced. In order for a metaphor to have any bite it must reflect some similarity. The metaphor ‘life's journey’ would hardly have had the currency that it enjoys in improving discourses, such as the speeches which accompany school prize-givings, had there been no way in which life could be seen as a journey. But unlike simile, metaphorical uses do not leave words unaffected. It has been pointed out by many students of metaphor that when a concept is displaced into a new domain it not only serves to highlight some hitherto unnoticed similarity between its old and new referents, but it changes its significance through coming to be used in a new domain. So the term ‘current’ was first used in the description of electricity, to highlight similarities between electricity and more easily observable fluids. The two centuries of use of this term in the electrical domain have certainly led to a change in its meaning (Martin and Harré, 1982).

Analogies and models

The recent trend in philosophy of science to look more closely at actual examples of scientific reasoning has disclosed the quite central role that analogical reasoning plays in both the physical sciences and the social sciences. A special terminology has grown up in the sciences by which the term ‘model’ is appropriated for concrete analogues (Bunge, 1973).

Scientific models are of two main kinds. There are heuristic or homoeomorphic models and explanatory or paramorphic models. Each kind has a specific use.

Many phenomena are too complicated for ready examination. Salient features can be brought out by abstracting a simpler form from the original complexity and idealizing its properties. A homoeomorphic or heuristic model is a convenient representation of its subject. It may be a concrete thing, such as the

scale models used in engineering. But it may be an abstract conceptual representation embodied in something like the ‘rational actor’ assumption in economics. Heuristic models are conservative. In a sense they merely represent what we already know but in some useful or convenient form.

Explanatory models (paramorphic analogues) are used creatively. They enable scientists to conceive of new kinds of beings and so far unobserved processes. Their main use is to complete theories by standing in for unobserved and so currently unknown causal processes. The kinetic theory depends on the idea of a swarm of molecules which are a model or analogue of the unknown constitution of real gases. The hypothetical behaviour of the molecular analogue must be like (analogous to) the behaviour of the real gas. Such models are of great interest to methodologists since they not only form the core of most scientific theories, but are also the vehicles for much creative scientific thinking. They are not devised at random. Their construction is always controlled by some implicit metaphysical assumptions (in the gas model case Newtonian atomism) which ensure their plausibility to the scientific community. This means that they are balanced between two analogy relations. They must behave analogously to the real thing they are a model for; and they are constructed by analogy with the real thing they are modelled on. For instance, the popular rule-following models in social psychology should replicate the behaviour of the unknown cognitive systems they are models for while they must lie within the constraints imposed by the real cases of rule-following, say in ceremonial action, which they are modelled on. Both analogy relations are usually open, that is, though they exhibit positive and negative aspects, similarities and differences, there is usually a degree of unexplored neutral analogy. Theories develop by the conceptual exploration and, in favourable cases, the empirical testing of the neutral analogy.

Explanatory and heuristic models can be neatly distinguished by reference to their constitutive analogies. For a heuristic model source and subject are identical. A model plane is a model of a plane. But for an explanatory model source and subject are distinct. The idea of an implicit rule is modelled on that of an explicit rule, but the former is an analogue of some unknown regulative cognitive process.

Bibliography

Bunge, M. 1973. *Method, Model and Matter*. Dordrecht: Reidel.

Hesse, M. 1963. *Models and Analogies in Science*. London: Sheed and Ward.

Martin, J. and Harré, R. 1982. Metaphor in science. In *Metaphor: Problems and Perspectives*, ed. D. Miall. Brighton: Harvester Press.

How to cite this article

Harré, Rom. "analogy and metaphor." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000097> doi:10.1057/9780230226203.0040

The New Palgrave Dictionary of Economics Online

Anderson, James (1739 – 1808)

J.M.A. Gee

From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Anderson, James; Corn Laws; development economics; Edinburgh school; Poor Laws

Article

Anderson farmed from the age of 15, first at Hermiston near Edinburgh, then at Monkshill, Aberdeenshire. Aberdeen honoured him with an LL.D. in 1780. He settled in Leith (near Edinburgh) in 1783, and founded *The Bee* (1790–94), a miscellany weekly magazine including literary, political and economic topics. He moved to London in 1797 and set up the magazine *Recreations* … (1799–1802) along the same lines as *The Bee*. The most important primary and secondary sources are listed below.

A contemporary of Adam Smith and James Steuart, James Anderson was second to none as a development economist. His writings lay great stress on the deadening effects of outmoded (feudal) institutions, adverse political and historic legacies, poor communications allied with sparse population, and repressive English-inspired taxation – especially the duties on salt and coal – on Scottish development. His proposals for improvement emphasized the gradualist approach – abstract economic models and grandiose schemes attracted his scorn – where the latent desire of man to improve his lot was freed from constraint and encouraged by state action and private self-interested philanthropy. Thus, though Anderson in general supported laissez-faire as being an essential requisite of optimal development, he believed the paternalistic encouragement of such development was frequently necessary, especially in the early stages. That he was no doctrinaire free-trader is seen in his espousal of the Corn Laws, on developmental grounds (see *An Inquiry into the Nature of the Corn Laws* …). He took issue with Smith on this, and also on Smith's notion that corn regulates the price of all commodities (see especially his ‘Postscript to Letter Thirteen’ in his *Observations* …). Smith never properly answered Anderson's criticisms (see Dow, 1984).

Anderson is regarded as an anticipator of Ricardo's rent theory (see, e.g., Schumpeter, 1954), but cannot in fact be cast in the narrowly abstract Ricardian mould. True, for Anderson an increase in corn price would have the differential effects on land rent as described by Ricardo; but this would be the first stage only of a development process. At the end of the process all land would have increased in fertility, and

what was previously the least fertile cultivated land could well be now as fertile as the previously most fertile land (see *The Bee*, vol. 6, 28 December 1791).
Anderson was convinced of the harm caused by the Poor Laws, and was responsible for a successful appeal against the introduction of the poor rate in Leith.
In addition to his writings on agriculture and economic development and his literary magazine pieces, Anderson also wrote on slavery, archaeology and greenhouse and chimney design!

Selected works

1777a. *Observations on the Means of exciting a spirit of National Industry*…. Edinburgh.

1777b. *An Inquiry into the Nature of the Corn Laws*…. Edinburgh.

1785. *An Account of the present State of the Hebrides, and Western Coasts of Scotland*…. Edinburgh.

1791–4. *The Bee*. Edinburgh.

1794. *A General View of the Agriculture and Rural Economy of the County of Aberdeen*…. Edinburgh.

1799–1802. *Recreations* …. London.

Bibliography

Anderson, William. 1865. *The Scottish Nation*. Vol. 1, Edinburgh: A. Fullarton & Co., 126–9.

Dow, A. 1984. The hauteur of Adam Smith: an unpublished letter from James Anderson of Monkshill. *Scottish Journal of Political Economy* 31(3), November, 284–5.

Mullet, C.F. 1968. A village Aristotle and the harmony of interests: James Anderson of Monks Hill. *Journal of British Studies* 8(1), November, 94–118.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: George Allen & Unwin.

How to cite this article

Gee, J.M.A. "Anderson, James (1739–1808) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000053> doi:10.1057/9780230226203.1877

The New Palgrave Dictionary of Economics Online

Anderson, Oskar Nikolayevich (1887 – 1960)

Heinrich L. Strecker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Anderson, O. N.; Econometric Society; econometrics; index numbers; quantity theory of money; sample surveys; statistics and economics; Tschuprow, A. A.; variate difference method

Article

Anderson was born on 2 August 1887 in Minsk, Russia, and died on 12 February 1960 in Munich, Federal Republic of Germany. As a disciple of Aleksandr A. Tschuprow the younger in St Petersburg, Anderson was a pioneer in statistics and econometrics. After leaving Russia in 1920 he became professor of statistics at the universities of Varna and Sofia in Bulgaria (until 1942), Kiel (until 1947) and Munich.

His oeuvre includes two textbooks and more than 150 articles in Russian, Bulgarian, English and German. Anderson participated during 1913–17 in the theoretical preparation and actual conduct of a sample on agricultural production in the Syr-Darja river area of Russia, one of the very earliest sample surveys. Later, he designed the sample plan for the processing of the Bulgarian Agricultural Census of 1926, with very good results which were decisive for further propagation and acceptance of sampling (1929; 1949).

Before and after the First World War Anderson developed, independently of W.S. Gossett, the variate difference method, a procedure to separate the smooth component (trend, business cycles) from the residual component, without making further assumptions about the underlying type of function (1929).

Anderson wrote one of the first, much-noticed econometric papers, an effort to verify statistically the quantity theory of money, which was a very early analysis of causes by means of economic data (1931).

Regarding index numbers, Anderson pointed particularly to the problem of chain index numbers, caused by error accumulation (1949; 1952).

Anderson was a charter member of the Econometric Society, a fellow or honorary member of numerous scientific associations, and held honorary doctorates from Vienna and Mannheim.

Selected works

1929a. *Über die repräsentative Methode und deren Anwendung auf die Aufarbeitung der Ergebnisse der bulgarischen landwirtschaftlichen Betriebszählung vom 31. Dezember 1926*. Munich: Fachausschuss für Stichprobenverfahren der Deutschen Statistischen Gesselschaft, 1949.

1929b. *Die Korrelationsrechung in der Konjunkturfoschung. Ein Beitrag zur Analyse von Zeitreihen*. Bonn: Schroeder-Verlag.

1931. Ist die Quantitätstheorie statistisch nachweisbar? *Zeitschrift für Nationalökonomie* 2, Vienna.

1935. *Einführung in die Mathematische Statistik*. Vienna: Springer-Verlag.

1949. Mehr Vorsicht mit Indexzahlen! *Allgemeines Statistisches Archiv* 33, 71–83, Munich.

1952. Wieder eine Indexverkettung? *Mitteilungsblatt für Mathematische Statistik* 4, Munich.

1954a. *Probleme der statistischen Methodenlehre*, 4th edn. Würzburg: Physica-Verlag, 1964.

1954b. Über den Umgang mit systematischen statistichen Fehlern. *Statistische Vierteljahresschrift* 7, Vienna.

Bibliography

Fels, E.M. 1968. Anderson, Oskar N. In *International Encyclopedia of Statistics*, vol. 1. London: Macmillan.

Strecker, H.L. 1965. Anderson, Oskar. In *Handwörterbuch der Sozialwissenschaften* (HDSW), vol. 12. Stuttgart/Tübingen: Fischer-Verlag/Mohr-Verlag.

Strecker, H., Kellerer, H., et al. 1963. *Oskar Anderson Ausgewählte Schriften*. 2 vols. Tübingen: Mohr-Verlag, with a complete bibliography of Anderson's publications.

How to cite this article

Strecker, Heinrich L. "Anderson, Oskar Nikolayevich (1887–1960)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000101>
doi:10.1057/9780230226203.0041

The New Palgrave Dictionary of Economics Online

Ando, Albert K. (1929 – 2002)

Charles Yuji Horioka

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Ando, A.; life-cycle hypothesis of saving; Modigliani, F.; Simon, H.

Article

Albert K. Ando was an eminent Japanese-born American economist who made many seminal contributions in a broad range of areas of economics. Born in Tokyo, Japan, on 15 November 1929, Ando went to the United States after the Second World War instead of joining the family business (ANDO Corporation, a major construction company). He received his BS in economics from the University of Seattle in 1951, his MA in economics from St Louis University in 1953, an MS in economics in 1956 and a Ph.D. in mathematical economics in 1959 from Carnegie Institute of Technology (now Carnegie Mellon University). After teaching at Carnegie and the Massachusetts Institute of Technology, Ando moved to the University of Pennsylvania in 1963 and remained there until his death from leukaemia on 19 September 2002, first as an associate professor of economics and finance, and from 1967 as a professor of economics and finance.

Ando held visiting appointments at universities in Louvain, Bonn and Stockholm, and consulted with the International Monetary Fund, the Federal Reserve Board, the Bank of Italy, and the Economic Planning Agency of Japan.

During his long and productive career, Ando received many honours and awards. For example, he was named Fellow of the Econometric Society, Ford Foundation Faculty Research Fellow, Guggenheim Fellow, and Japan Foundation Fellow, and was given the Alexander von Humboldt Award for Senior American Scientists.

Ando made important contributions in such diverse fields as econometrics (theory and applications), stochastic optimal control, the theory of aggregation and partitions in dynamic systems, monetary economics, macroeconomic modelling, and policy design, with an emphasis on interactions between economic growth and cyclical fluctuations, investment behaviour, theoretical and empirical investigations of household saving and consumption behaviour, and demography. His geographic breadth was equally great, with particular focus on Italy, Japan, and the United States. Ando

collaborated, among others, with Nobel laureate Herbert Simon on questions regarding aggregation and causation in economic systems (see, for example, Simon and Ando, 1961, and Ando, Fisher, and Simon, 1963) and with another Nobel laureate, Franco Modigliani, on extending the life-cycle hypothesis of saving (see, for example, Ando and Modigliani, 1963), and constructing large-scale macroeconomic models (see, for example, Ando and Modigliani, 1969).

A common thread in much of Ando's work is the care with which he analysed data. He subjected all of the data he used (whether national accounts data, data from household surveys, or company data) to careful scrutiny, was constantly on the lookout for inconsistencies, conceptual deficiencies, and so on, in the data, and made the necessary adjustments to the data to correct for any inconsistencies and conceptual deficiencies. He then analysed the resulting data meticulously and creatively to shed light on important questions such as the causes of the decade-long recession in Japan in the 1990s (he found that it was due primarily to the massive capital losses on household holdings of corporate equities; see, for example, Ando, 2002a), whether aged households dissave (he found that they dissave relatively rapidly in Italy and the United States but moderately or not at all in Japan; see, for example, Ando and Kennickell, 1987; Hayashi, Ando and Ferris, 1988; and Ando and Nicoletti-Altimari, 2004), how the cost of capital compares in the United States and Japan (he found that it is considerably higher in the United States if individual company data are used but not if national accounts data are used; see, for example, Ando and Auerbach (1988; 1990) and Ando, Hancock and Sawchuk (1997).

Ando played a central role in the construction of the Massachusetts Institute of Technology, the University of Pennsylvania, and the Social Science Research Council (MPS) model, an early large-scale macroeconomic model of the US economy, as well of the Bank of Italy's macroeconomic model of the Italian economy (see, for example, Ando and Modigliani, 1969, and Ando, 1974), and in his later years he devoted considerable energy to constructing a dynamic micro-simulation model of demographic structure for Italy, Japan and the United States, which he used to project future trends in the saving rate (he projected that Japan's saving rate would increase slightly in the immediate future as the number of children per family declined sharply, then fall moderately as the proportion of older persons in the population increased; he projected similar trends in Italy as well: see, for example, Ando et al., 1995, and Ando and Nicoletti-Altimari, 2004).

See Also

- Modigliani, Franco
- Simon, Herbert A.

Selected works

1961. (With H. Simon.) Aggregation of variables in dynamic systems. *Econometrica*. 29, 111–38.

1963. (With F. Modigliani.) The ‘life cycle’ hypothesis of saving: aggregate implications and tests. *American Economic Review* 53, 55–84.

1963. (With F. Fisher, M. Franklin and H. Simon.) *Essays on Social Science Models*. Cambridge, MA:

MIT Press.

1965. (With F. Modigliani.) The relative stability of monetary velocity and the investment multiplier. *American Economic Review* 55, 693–728.

1965. (With G. Kaufman.) Bayesian analysis of the independent multinomial process – neither mean nor precision known. *Journal of the American Statistical Association* 60, 347–58.

1969. (With F. Modigliani.) Econometric analysis of stabilization policies. *American Economic Review* 59, 296–314.

1974. Some aspects of stabilization policies, the monetarist controversy, and the MPS model. *International Economic Review* 15, 541–71.

1974. (With F. Modigliani, R. Rasche and S. Turnovsky.) On the role of expectations of price and technological change in an investment function. *International Economic Review* 15, 384–414.

1985. (With M. Blume, E. Marshall and I. Friend.) *The Structure and Reform of the US Tax System*. Cambridge, MA: MIT Press.

1987. (With A. Kennickell.) How much or little life cycle is there in micro data? The cases of the United States and Japan. In *Macroeconomics and Finance: Essays in Honor of Franco* Modigliani, ed. R. Dornbusch. Cambridge, MA and London: MIT Press.

1988. (With A. Auerbach.) The cost of capital in the United States and Japan: a comparison. *Journal of the Japanese and International Economies* 2, 134–58.

1988. (With F. Hayashi and R. Ferris.) Life cycle and bequest savings: a study of Japanese and US households based on data from the 1984 NSFIE and the 1983 Survey of Consumer Finances. *Journal of the Japanese and International Economies* 2, 450–91.

1990. (With A. Auerbach.) The cost of capital in Japan: recent evidence and further results. *Journal of the Japanese and International Economies* 4, 323–50.

1992. (With L. Guiso, D. Terlizzese and D. Dorsainvil.) Saving among young households: evidence from Japan and Italy. *Scandinavian Journal of Economics* 94, 233–50.

1995. (With A. Moro, J. Cordoba and G. Garland.) Dynamics of demographic development and its impact on personal saving: case of Japan. *Ricerche Economiche* 49, 179–205.

1997. (With J. Hancock and G. Sawchuk.) Cost of capital for the United States, Japan and Canada: an

attempt at measurement based on individual company records and aggregate national accounts data. In *Financing Growth in Canada*, ed. P. Halpern. Calgary: University of Calgary Press.

2002a. Missing household saving and valuation of corporations: inquiry into Japanese national accounts I. *Journal of the Japanese and International Economies* 16, 147–76.

2002b. The elusive total budget outlay of the Japanese government: an inquiry into the Japanese national accounts II. *Journal of the Japanese and International Economies* 16, 177–93.

2004. (With S. Nicoletti-Altimari.) A micro simulation model of demographic development and households' economic behavior in Italy. Economic Working Paper No. 533. Economic Research Department, Bank of Italy.

How to cite this article

Horioka, Charles Yuji. "Ando, Albert K. (1929–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000207> doi:10.1057/9780230226203.0042

The New Palgrave Dictionary of Economics Online

animal spirits

Roger E. A. Farmer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘animal spirits’ was used by Keynes to refer to the idea that business cycles might be caused by crowd psychology. Recent work, in the aftermath of rational expectations, has focused on incorporating this idea into general equilibrium theory by exploiting the fact that dynamic general equilibrium models often contain a continuum of indeterminate equilibria. In stochastic models, production may differ across states of nature solely because of differences in the rational self-fulfilling beliefs of investors. This dependence of outcomes on beliefs provides a modern interpretation of the idea that the business cycle may be driven by animal spirits.

Keywords

animal spirits; autarky; capital–labour substitution; commodity space; comparative statics; complete and incomplete participation; determinacy and indeterminacy of equilibria; dynamic inefficiency; dynamic stochastic general equilibrium (DSGE) models; externalities in preferences; extrinsic uncertainty; general equilibrium; golden rule; Great Depression; Hume, D.; infinite horizon models; insurance contracts; intrinsic uncertainty; Keynes, J. M.; Kydland, F.; leisure; Lucas, R.; Markov processes; overlapping generations models; Prescott, E.; rational expectations; real business cycles; returns to scale; self-fulfilling expectations; stabilization policy; sunspot equilibrium; technology shocks; Thornton, H.; total factor productivity; variable capacity utilization

Article

The term ‘animal spirits’ is closely associated with John Maynard Keynes, who used it in his 1936 book, *The General Theory of Employment Interest and Money*, to capture the idea that aggregate economic activity might be driven in part by waves of optimism or pessimism (although Robin Mathews, 1984, p. 212, points out that Keynes would have been aware of its use by David Hume, 1739, pp. 60–1).

> Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of animal spirits – a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities. (Keynes, 1936, pp. 161–2).

The idea that waves of spontaneous optimism might drive business cycles was not new to Keynes and can be traced at least as far back as Henry Thornton, who attributed a central role in his theory of credit to '… that confidence which subsists among commercial men in respect to their mercantile affairs …' (Thornton, 1802, p. 75).

The advent of rational expectations

The early writers, including Keynes, did not develop fully worked-out dynamic models in which expectations of agents are related to outcomes that are later realized. The development of complete artificial economies of this kind occurred first with the rational expectations revolution in the 1970s in which the static macroeconomic disequilibrium model of Keynes's *General Theory* was replaced by modern dynamic general equilibrium models rooted in Chapter 7 of Gerard Debreu's *Theory of Value* (1959). This development began with the work of Robert E. Lucas, Jr., and early examples of rational expectations models include Lucas and Leonard Rapping (1969) and Lucas (1972; 1973). Lucas's 1972 and 1973 papers were attempts to understand the business cycle as a monetary phenomenon. Monetary models gave way to exclusively real models of the business cycle following the publication of influential papers by Fynn Kydland and Edward C. Prescott (1982) and John B. Long and Charles Plosser (1983), and modern macroeconomics theories, based on these early contributions, are referred to as 'dynamic stochastic general equilibrium (DSGE) models'.

Early DSGE models were restricted to examples in which there exists a finite number of agents (often only one) choosing consumption, investment and employment sequences in an economy with complete markets. Infinite horizon (IH) models of this kind have the same structure as the finite general equilibrium model studied by Kenneth Arrow and Gerard Debreu (1954) and Lionel McKenzie (1959), with the exception that the commodity space is infinite dimensional. Timothy Kehoe and David Levine (1985) showed that the competitive equilibria of IH exchange economies satisfy the first and second theorems of welfare economics; and from applying their methods to production economies it follows that that consumption, investment and employment sequences can be treated 'as if' they were chosen by a social planner maximizing a concave objective function subject to a set of linear constraints. Social planning problems have a unique solution in which all fluctuations in investment must occur as a direct consequence of fluctuations in the fundamentals of the economy; typically taken to consist of preferences, endowments and technologies. It follows that, if expectations are rational, there is no room in these economies for animal spirits to exert an independent influence on economic activity.

The infinite horizon model under constant returns to scale

The modern use of DSGE models has followed two routes. One class of models, following the IH approach, assumes that all decisions are taken by a finite set of infinitely lived households each of which makes decisions for current and future family members. This class includes the real business cycle (RBC) model, currently dominant in the profession, which has a history dating back to Frank Ramsey (1928), David Cass (1965) and Tjalling Koopmans (1965).

In simple representations of the IH model, one assumes that a single representative agent allocates output, Y_t between consumption, C_t and next period's capital stock, K_{t+1}. Output is produced from capital, K_t and labour L_t using a constant returns to scale technology that is subject to a productivity shock which is modelled as a random variable A_t. The representative agent ranks alternative probability distributions over consumption and labour supply using an additively separable utility function. This problem can be represented as follows:

$$\max_{\{C_t, L_t, K_{t+1}\}} \sum_{t=1}^{\infty} \left(\frac{1}{1+\rho}\right)^{t-1} E_1[U(C_t, L_t)], \tag{1}$$

$$Y_t = A_t K_t^a L_t^b, \tag{2}$$

$$K_{t+1} = K_t(1-\delta) + Y_t - C_t, \; K_1 = \bar{K}_1. \tag{3}$$

Here, $\rho > 0$ is the agent's discount rate and $0 \leq \delta < 1$ represents depreciation. The parameters *a* and *b* represent the elasticities of capital and labour in production and the assumption of constant returns to scale implies that

$$a + b = 1. \tag{4}$$

$E_1[\cdot]$ is the expectations operator, and the interpretation of this problem is that the agent chooses sequences $\left\{C_t(A^t), L_t(A^t)K_{t+1}(A^t)\right\}_{t=1}^{\infty}$ where $A^t = \left\{A_1, A_2 \ldots A_t\right\}$ is the history of shocks from date 1 to date t. A_t is a random variable, generated by an autocorrelated stochastic process.

In standard IH models one assumes that $U(x, y)$ is increasing in x, decreasing in y, strictly concave and twice continuously differentiable, and under these assumptions the programming problem defined in eq. (1) is concave and has a unique solution. Under the commonly assumed functional form,

$$U(C, L) = \log(C) - \frac{L^{1+\gamma}}{1+\gamma},$$

this solution is characterized by the first order conditions:

$$C_t L_t^{\gamma} = b\frac{Y_t}{L_t}, \tag{5}$$

$$\frac{1}{C_t} = E_t\left\{\frac{1}{(1+\rho)C_{t+1}}\left(1 - \delta + a\frac{Y_{t+1}}{K_{t+1}}\right)\right\}, \tag{6}$$

$$\lim_{T\to\infty}\left(\frac{1}{1+\rho}\right)^T E_1\left[\frac{K_{T+1}}{C_T}\right] = 0. \tag{7}$$

For the real business cycle programme it is critical to assume that the production function is linearly homogenous and preferences are strictly concave, since these assumptions imply that the problem of the representative agent has a unique solution. More generally, if there are multiple agents one can write down the problem of a social planner who maximizes a social welfare function, defined as a weighted sum of individual utilities.

The OLG model and how it differs

In contrast to the IH model, in overlapping generations (OLG) economies one assumes that the set of agents is infinite and that each agent lives for a finite number of periods; this model was developed first in English by Paul Samuelson (1958), although Maurice Allais' book (1948), written in French, predates Samuelson's contribution.

In OLG models, unlike in the IH model with concave preferences and technologies, there may exist equilibria that are dynamically inefficient. In equilibria of this kind the economy has 'too much capital', and a benevolent social planner could improve social welfare for all generations by consuming part of the capital stock (thereby raising consumption for the current generation) and diverting future output from investment to consumption (thereby raising consumption for all future generations).

After the publication of Samuelson's article in 1958, a considerable literature developed discussing the source of dynamic inefficiency. The question was finally settled with the publication of Shell's (1971) paper, 'Notes on the Economics of Infinity'. Shell argued that both IH and OLG models are special cases of Debreu's (1959) formulation of general equilibrium. In both cases the commodity space is infinite dimensional. In the IH model the number of agents is finite; in the OLG model it is infinite. This apparently innocuous difference is the key to understanding why there may be inefficient equilibria in the OLG model since, in an inefficient equilibrium, no single agent can make a welfare-improving trade. In contrast, dynamic inefficiency in an IH economy would imply the existence of an agent with infinite wealth at equilibrium prices.

Both IH and OLG models have been used as vehicles to develop the idea that animal spirits may independently influence economic activity. Since the IH model with concave preferences and technologies leads to equilibria that are efficient, it was the OLG model that was first exploited to develop the modern version of the 'animal spirits hypothesis'. However, since the period length of the two-period OLG model is typically interpreted as 25 or 30 years, and since the average period of a business cycle is six to eight years, it was easy to dismiss the early work, based on the OLG structure, on the grounds that the equilibria that it led to were theoretical curiosities that are not relevant in the real world. This criticism was addressed by a second generation of animal spirits economies, in which the OLG model was replaced by an IH framework that relaxed the assumption that the technology is subject to constant returns to scale.

Animal spirits, sunspots and incomplete participation

In DSGE models the term 'animal spirits' (Azariadis, 1981; Howitt and McAfee, 1992; Farmer and Guo, 1994) is used interchangeably with 'sunspots' (Cass and Shell, 1983), 'self-fulfilling

prophecies' (Azariadis, 1981; Farmer, 1993) and most recently 'irrational exuberance' by Alan Greenspan (1996) at an after-dinner speech.

Jevons (1884) used the term 'sunspots' to refer to the literal possibility that astronomical events could influence the trade cycle through the intermediating effect of the weather on agriculture. In their 1983 article, Cass and Shell meant something different. They constructed a two-period general equilibrium model with complete markets in which some agents are unable to enter into insurance contracts. They referred to this restriction as 'incomplete participation' to distinguish it from a potentially more serious market breakdown in which some kinds of insurance contracts cannot be entered into *by anyone*. Cass and Shell distinguished between *intrinsic* uncertainty, which can influence fundamentals of the economy, and *extrinsic* uncertainty, under which the fundamentals are unchanged across alternative extrinsic events. They showed that the inability of a subset of agents to enter into insurance contracts is a sufficient departure from standard general equilibrium assumptions to permit the existence of equilibria in which allocations differ across states of the world in which all uncertainty is extrinsic. When this occurs, they said that *sunspots matter*.

In an economy with a complete set of insurance markets and risk-averse agents, all of whom can participate in these markets, sunspots cannot matter. Since agents are risk averse, they would prefer the mean of a random allocation to the allocation itself. But if all uncertainty is extrinsic then the mean allocation is feasible; hence a sunspot allocation cannot be an equilibrium of a complete markets economy with complete participation. Sunspot equilibria are Pareto-inefficient, but for a different reason from the dynamic inefficiency associated with over-accumulation of capital in deterministic OLG models. Sunspot inefficiency arises from the addition of unnecessary randomness to an economy in which agents prefer to avoid fluctuations in their consumption allocations.

Animal spirits in an OLG model

The first application of sunspots to a DSGE model is due to Azariadis (1981). He constructed a two-period overlapping generations model with no intrinsic uncertainty. This model possesses a unique steady state in which money has value. Under typical assumptions about preferences, the linearized dynamics of equilibrium price sequences in the neighbourhood of the steady state obey a functional equation of the form

$$p_t = \alpha E_t[p_{t+1}] + c, \ |\alpha| < 1. \tag{8}$$

Azariadis looked for equilibria that follow a two-state Markov process: that is, equilibria of the form

$$\begin{bmatrix} p_t(s_t = 1) \\ p_t(s_t = 2) \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} \alpha p_{t+1}^1(s_t = 1) \\ \alpha p_{t+1}^1(s_t = 2) \end{bmatrix} + \begin{bmatrix} c \\ c \end{bmatrix}$$
(9)

where $s_t \in \{1, 2\}$ is the state at date *t* and π_{ij} is the probability that $s_t = i$ conditional on $s_{t-1} = j$. For the linearized model, the fact that $|\alpha| < 1$ implies that the only equilibrium in this class is one for which

$$p(s_t = i) = \frac{c}{1 - \alpha}, \; i = 1, 2,$$
(10)

that is, the price is constant and independent of the non-fundamental uncertainty. But in the nonlinear model the equation that defines equilibrium price sequences takes the form

$$p_t(s_t) = E_t[g(p_{t+1}(s_{t+1})|s_t)],$$
(11)

where the function $g(\cdot)$ depends on assumptions about the form of the utility function. The equation defining a two-state Markov equilibrium takes the more general form

$$\begin{bmatrix} p_t(s_t = 1) \\ p_t(s_t = 2) \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} g\left(p_{t+1}^1(s_t = 1)\right) \\ g\left(p_{t+1}^1(s_t = 2)\right) \end{bmatrix}.$$
(12)

In this case, Azariadis showed that, as long as consumption and leisure are not gross substitutes, it is possible to find positive numbers p_1, p_2 such that $p_1 \neq p_2$ and positive probabilities π_{11}, π_{12}, π_{21} and π_{22} such that

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} g(p_1) \\ g(p_2) \end{bmatrix}. \tag{13}$$

In other words, prices (and implicitly employment, consumption and GDP) in this economy fluctuate between two different levels based purely on the occurrence of self-fulfilling expectations or, in Keynes's terminology, ‘animal spirits’. As with the Cass–Shell example of sunspots, however, the Azariadis example could easily be dismissed as a model of a real economy since it required the assumption that consumption and leisure are gross complements – an assumption that was widely believed to be implausible and inconsistent with other evidence. The challenge was to develop a quantitative model of the business cycle in which aggregate fluctuations are driven by animal spirits, expectations are rational, and the model can capture the observed volatilities of output, consumption, GDP and hours.

Animal spirits and indeterminacy

The example of sunspots provided in the Cass–Shell (1982) paper relied on constructing an economy in which there are multiple equilibria. They showed that, when some agents are unable to participate in the insurance markets that occur before they are born, randomizations across deterministic allocations can also be sustained as equilibria. In the presence of complete participation in insurance markets these randomized equilibria would be ruled out since they are associated with unnecessary uncertainty that risk-averse agents would prefer to avoid.

In addition to the fact that an OLG equilibrium can be dynamically inefficient, there is a second key way in which OLG and IH models differ. In the IH model the set of equilibria is generically finite whereas OLG economies can contain a continuum of equilibria. (Roughly speaking, ‘generically finite’ means that for almost all IH economies there is a finite number of equilibria, and ‘almost all’ means that this statement is true for an open dense set of parameters in a parameterized family of economies.) The fact that there is a finite number of equilibria implies that each equilibrium of the IH model is locally unique, that is, there is no other equilibrium that is arbitrarily close to it.

A locally unique equilibrium is also called ‘determinate’. Determinacy of equilibrium is an important property since, if one is interested in comparative statics, it is important that small changes in exogenous variables lead to predictable small changes in endogenous variables. If the equilibrium is one of a continuous set of equilibria (as would happen if the equilibrium were indeterminate) then the model does not make a clear prediction as to how prices and quantities would be expected to change in response to a change in policy or in some other fundamental of the economy.

Under some assumptions about preferences (a sufficient condition is that the endowment of the agents is sufficiently tilted towards youth), the one-good two-period OLG model possesses two steady states. Each of these steady states is a stationary equilibrium with a constant real rate of interest; in one stationary equilibrium money has positive value and in the other it does not. David Gale (1973) refers to economies that possess a monetary steady state as ‘Samuelson’ to distinguish them from those that do

not (he calls these ‘Classical’). In a Samuelson economy the two steady states are respectively ‘generationally autarkic’ (money has no value) and ‘golden rule’ (the real rate of interest equals the population growth rate). In Samuelson economies there exists a continuum of non-stationary equilibria and, when consumptions in adjacent periods are gross substitutes, each of these non-stationary equilibria converges to the autarkic steady state.

The non-stationary equilibria in the OLG model provide a rich source of equilibria over which to randomize; however, they all converge to an autarkic equilibrium in which money has no value. This property makes it difficult to construct stationary stochastic equilibria around the autarkic steady state since there are no non-stationary paths that approach the steady state from below. To get around this difficulty, Farmer and Woodford (1984) showed that, by adding government spending to the OLG model, one can construct randomizations over a set of non-stationary equilibria that converge to a stationary state in which money has value. The addition of positive inflation-financed government expenditure shifts the set of stationary equilibria, and the indeterminate non-monetary equilibrium of the OLG model becomes a second monetary equilibrium. By adding a zero mean random variable to the model, Farmer and Woodford were able to construct a new set of *stationary* sunspot equilibria. Locally, these equilibria obey a difference equation of the form of eq. (8), but the parameter α is greater than 1 in absolute value. It follows that one can construct equilibria in this model of the form:

$$p_{t+1} = \frac{1}{\alpha} p_t - \frac{c}{\alpha} + u_{t+1}, \tag{14}$$

where u_{t+1} is *any* random variable with zero conditional mean. Further, the unconditional probability distribution of the price level can be shown to converge to an invariant probability measure that depends on the distribution of the sequence of sunspot shocks, $\{u_t\}$. This is an important property of a rational expectations equilibrium since, arguably, stationarity is necessary for agents to learn about the world in which they live and to find ways of making unbiased forecasts of the moments of future prices.

Real business cycles and the animal spirits hypothesis

The examples of stationary sunspot rational expectations equilibria, originally constructed in the OLG model, did not have much impact on mainstream macroeconomics. Although the first rational expectations models were constructed as monetary examples within the two-period OLG structure (for example, Lucas's seminal 1972 paper), the profession soon moved on to real models based on IH economies. The IH structure is more amenable to confrontation with data since the period of the model can easily be mapped into the period of data collection. Further, the examples of Azariadis and Farmer–Woodford were constructed in models that relied on assumptions widely believed to be unrealistic; these included the assumption of gross complements and two-period lives (in the case of the Azariadis model) and the assumption that sunspots exist close to a dynamically inefficient steady state in the Farmer–Woodford model (this assumption can be shown to generate counter-intuitive responses of inflation to

expansionary fiscal policy).

To confront these criticisms, Howitt and McAfee (1992), Benhabib and Farmer (1994) and Farmer and Guo (1994) constructed examples of animal spirits equilibria within the IH paradigm by dropping the assumption that the technology is subject to constant returns to scale. At the time that this work was published, a number of authors (Caballero and Lyons, 1993, are prominent examples) had estimated the degree of increasing returns to scale in US manufacturing industries and found it to be large.

In their 1994 paper, Benhabib and Farmer took a relatively standard IH model and added externalities and increasing returns to scale. Farmer and Guo (1994) constructed a discrete time version of the Benhabib–Farmer model and showed that it can be used to generate business cycle fluctuations driven by animal spirits. They argued that the animal spirits-driven model is *more* successful than the real business cycle model at capturing the observed dynamics of output, employment, investment and consumption because it can replicate the hump-shaped response of output and investment to shocks that is observed in US data.

The Benhabib–Farmer–Guo (BFG) model has the same form as the IH model described in eqs. (11) to (17) but it distinguishes between the *private* technology and the *social* technology. BFG assume that the economy contains a large number of identical firms, each of which produces output using the production function

$$Y_t = A_t K_t^a L_t^b. \tag{15}$$

In BFG, the term A_t is not exogenous. Instead, it represents an input externality of the form

$$A_t = \bar{K}_t^{\alpha - a} \bar{L}_t^{\beta - b}, \tag{16}$$

where $\bar{K}_t$ and $\bar{L}_t$ represent the economy-wide average use of capital and labour. Replacing (1.16) in (1.15) and imposing the assumption that the economy is in a symmetric equilibrium in which $\bar{K}_t = K_t$ and $\bar{L}_t = L_t$ leads to the *social technology*

$$Y_t = K_t^{\alpha} L_t^{\beta}. \tag{17}$$

BFG assumed that

$$\alpha + \beta > 1, a + b = 1, \tag{18}$$

which implies that there are increasing returns to scale in the social technology but constant returns to scale at the level of the individual firm. Since increasing returns enter the economy as an external effect, each firm maximizes a concave profit function, and the equilibrium of the competitive economy is well defined. BFG showed that equilibria in their IH economy with increasing returns are characterized by the following system of equations.

$$Y_t = A_t K_t^{\alpha} L_t^{\beta}, \tag{19}$$

$$K_{t+1} = K_t(1 - \delta) + Y - C_t, \tag{20}$$

$$C_t L_t^{Y} = b\frac{Y_t}{L_t}, \tag{21}$$

$$\frac{1}{C_t} = E_t\left\{\frac{1}{(1+\rho)C_{t+1}}\left(1 - \delta + a\frac{Y_{t+1}}{K_{t+1}}\right)\right\}, \tag{22}$$

$$\lim_{T \to \infty} \left(\frac{1}{1+\rho} \right)^T E_1 \left[\frac{K_{T+1}}{C_T} \right] = 0. \tag{23}$$

When $a = \alpha$ and $b = \beta$, this model collapses to the real business cycle version of the IH economy. But if $\alpha > a$, $\beta > b$ and $\alpha + \beta$ is greater than 1 and 'large enough', Benhabib and Farmer showed that the dynamics of the IH model change character, and the model contains a continuum of indeterminate equilibria, just as the OLG model does. Farmer and Guo calibrated the model to US data and, by choosing parameters that appeared consistent with contemporary estimates of returns to scale, they showed that the model exhibits business cycles driven by self-fulfilling waves of optimism and pessimism.

To provide a degree of discipline to the calibration exercise, real business cycle economists estimate the volatility of real productivity shocks by constructing an estimate of total factor productivity (TFP). This is an accurate measure of TFP under the maintained assumptions of competitive markets and constant returns to scale. Farmer–Guo provided discipline to their calibration exercise by constructing the measure of TFP that would be estimated from data generated by an animal spirits economy by an econometrician who assumed incorrectly that the technology was driven by technology shocks, and imposed the incorrect identifying assumption of constant returns to scale. They showed that this measure has very similar properties to that of the TFP estimates from US data.

Animal spirits, business cycles and welfare

Much recent business cycle research assumes that business cycles are driven by technology shocks; but we do not have a very good explanation of what these shocks represent. The BFG model represents a plausible alternative to the real business cycle model. It recaptures an old idea and recasts it in modern language.

Why should we care if shocks arise in the productivity of the technology or in the minds of entrepreneurs? The answer is connected to the efficiency question. If business cycles arise as the consequence of the optimal allocation of resources in the face of unavoidable fluctuations in the technology, then there is not much that government can or should do about them. But, if they arise as the consequence of avoidable fluctuations in the animal spirits of investors, then the fluctuations that result are avoidable and the allocations are Pareto-suboptimal. Animal spirit-driven business cycles provide a reason for countercyclical stabilization policy, and the cause of cycles is therefore an important question. In 1996 Takashi Kamihigashi showed that the RBC economy (driven by TFP shocks) and the Benhabib–Farmer model (driven by animal spirits) are observationally equivalent when estimated on aggregate

data and that, if one uses aggregate evidence alone, constant returns to scale is an identifying assumption. The empirical literature since the publication of volume 63 of the *Journal of Economic Theory* in 1994 suggests that early estimates of the degree of returns to scale were overstated, and more recent estimates (for example, Basu and Fernald, 1997) are more modest. This has led to renewed developments by theorists who have constructed modifications of the basic animal spirits model that are able to bring down the required degree of returns to scale to well within the tolerance of the best econometric estimates. Innovations to this literature include the construction of multi-sector models (Benhabib and Farmer, 1996; Weder, 1998; Benhabib, Nishimura and Meng, 2000; Harrison, 2001), externalities in preferences (Farmer and Bennett, 2000; Hintermaier, 2003), capital–labour substitution (Grandmont, Pintus and de Vilder, 1998), stabilization policy (Schmitt-Grohé and Uribe, 1997; Guo and Lansing, 1998; Lloyd Braga, 2003), alternative explanations of the Great Depression (Harrison and Weder, 2006) and variable capacity utilization (Wen, 1998; Benhabib and Wen, 2004). Benhabib and Farmer (1999) provide a survey of this literature and references to additional related papers.

Recent examples of animal spirits-driven models are able to explain a wide range of phenomena and, when supplemented by the assumption of variable capacity utilization, the animal-spirits explanation of business cycles outperforms the RBC model in most dimensions. Since the two models have very different policy conclusions, research that addresses the question of whether business cycles are driven by animal spirits is likely to remain a lively and important focus of research for some time to come.

See Also

- Keynes, John Maynard
- Keynesianism
- Keynesian revolution
- overlapping generations model of general equilibrium
- rational expectations
- sunspot equilibrium

Bibliography

Allais, M. 1948. *Economie et intérêt*. Paris: Imprimerie Nationale.

Arrow, K. and Debreu, G. 1954. Existence of a competitive equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25, 380–96.

Basu, S. and Fernald, J. 1997. Returns to scale in US production: estimates and implications. *Journal of Political Economy* 105, 249–83.

Benhabib, J. and Farmer, R. 1994. Indeterminacy and increasing returns. *Journal of Economic Theory* 63, 19–46.

Benhabib, J. and Farmer, R. 1996. Indeterminacies and sector specific externalities. *Journal of Monetary Economics* 37, 421–44.

Benhabib, J. and Farmer, R. 1999. Indeterminacy and sunspots in macroeconomics. In *The Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Benhabib, J., Nishimura, K. and Meng, Q. 2000. Indeterminacy under constant returns to scale in multisector economies. *Econometrica* 68, 1541–8.

Benhabib, J. and Wen, Y. 2004. Indeterminacy, aggregate demand and the real business cycle. *Journal of Monetary Economics* 51, 503–30.

Caballero, R. and Lyons, R. 1993. The case for external economies: an overview. In *Political Economy, Growth and Business Cycles*, ed. A. Cukierman, Z. Hercowitz and L. Leiderman. Cambridge, MA: MIT Press.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32, 233–40.

Cass, D. and Shell, K. 1982. Do sunspots matter? *Journal of Political Economy* 91, 193–227.

Debreu, G. 1959. *The Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New York: Wiley.

Farmer, R. 1993. *The Macroeconomics of Self-Fulfilling Prophecies*. Cambridge, MA: MIT Press.

Farmer, R. and Bennett, R. 2000. Indeterminacy with non-separable utility. *Journal of Economic Theory* 93, 118–43.

Farmer, R. and Guo, J.-T. 1994. Real business cycles and the animal spirits hypothesis. *Journal of Economic Theory* 63, 42–73.

Farmer, R. and Woodford, M. 1984. Self-fulfilling prophecies and the business cycle. Working Paper No. 84-12, Caress, University of Pennsylvania. Reproduced in *Macroeconomic Dynamics* 1 (1997), 740–69.

Gale, D. 1973. Pure exchange equilibrium of dynamic economic models. *Journal of Economic Theory* 5, 12–36.

Grandmont, J.-M., Pintus, P. and de Vilder, R. 1998. Capital–labor substitution and nonlinear endogenous business cycles. *Journal of Economic Theory* 80, 14–59.

Greenspan, A. 1996. The challenge of central banking in a democratic society. Remarks at the Annual Dinner and Francis Boyer Lecture of The American Enterprise Institute for Public Policy Research, Washington, D.C, December 5, 1996. Online. Available at http://www.federalreserve.gov/BoardDocs/speeches/1996/19961205.htm, accessed 7 June 2006.

Guo, J. and Lansing, K. 1998. Indeterminacy and stabilization policy. *Journal of Economic Theory* 82, 481–90.

Harrison, S. 2001. Indeterminacy in a model with sector-specific externalities. *Journal of Economic Dynamics and Control* 25, 747–64.

Harrison, S. and Weder, M. 2006. Did sunspot forces cause the Great Depression? *Journal of Monetary Economics* (forthcoming).

Hintermaier, T. 2003. On the minimum degree of returns to scale in sunspot models of the business cycle. *Journal of Economic Theory* 110, 400–9.

Howitt, P. and McAfee, P. 1992. Animal spirits. *American Economic Review* 82, 493–507.

Hume, D. 1739. *A Treatise of Human Nature*, ed. L. Selby-Bigge. Oxford: Clarendon Press, 1888.

Jevons, W. 1884. On the study of periodic commercial fluctuations. In *Investigations in Currency and Finance*, ed. H. Foxwell. London: Macmillan.

Kamahigashi, T. 1996. Real business cycles and sunspot fluctuations are observationally equivalent. *Journal of Monetary Economics* 37, 105–17.

Kehoe, T. and Levine, D. 1985. Comparative statics and perfect foresight in infinite horizon economies. *Econometrica* 53, 433–53.

Keynes, J.M. 1936. *The General Theory of Employment Interest and Money*. London: Macmillan.

Koopmans, T. 1965. On the concept of optimal economic growth. In *The Econometric Approach to Development Planning*. Amsterdam: North-Holland.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lloyd Braga, T. 2003. Endogenous business cycles and systematic stabilization policy. *International*

Economic Review 44, 895–915.

Long, J., Jr. and Plosser, C. 1983. Real business cycles. *Journal of Political Economy* 91, 39–69.

Lucas, R., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R., Jr. 1973. Some international evidence on output–inflation tradeoffs. *American Economic Review* 63, 326–34.

Lucas, R., Jr. and Rapping, L. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77, 721–54.

Mathews, R. 1984. Animal spirits. In The Proceedings of the British Academy 70, 209–29.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27, 51–71.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Samuelson, P. 1958. An exact consumption–loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66, 467–82.

Schmitt-Grohé, S. and Uribe, M. 1997. Balanced-budget rules, distortionary taxes, and aggregate instability. *Journal of Political Economy* 105, 976–1000.

Shell, K. 1971. Notes on the economics of infinity. *Journal of Political Economy* 79, 1002–11.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. Fairfield, NJ: August Kelley, 1978.

Weder, M. 1998. Fickle consumers, durable goods and business cycles. *Journal of Economic Theory* 81, 37–57.

Wen, Yi. 1998. Capacity utilization under increasing returns to scale. *Journal of Economic Theory* 81, 7–36.

How to cite this article

Farmer, Roger E. A. "animal spirits." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/

article?id=pde2008_A000105> doi:10.1057/9780230226203.0043

The New Palgrave Dictionary of Economics Online

anthropometric history

John Komlos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Anthropometric history is the study of human size as an indicator of how well the human organism fared during childhood and adolescents in its socio-economic and epidemiological environment. The development of this field has opened up new windows on the ways in which economic processes affected the populations experiencing it, such as the hidden costs of industrialization and urbanization.

Keywords

antebellum puzzle; anthropometric history; biology; cliometrics; gross national product; height; biological welfare; Human Development Index; labour productivity; longevity; an indicator of biological welfare; nutrition; population density; Sen, A.; slaves; weight

Article

Anthropometric history is the study of human size, primarily physical stature, weight, and the body mass index [$\frac{weight(kg)}{height(m)^2}$] in order to ascertain how well the human organism thrived in its socio-economic and epidemiological environment.
As early as 1829 scholars recognized that the economy had a profound influence on human physical growth. In the 1960s French historians resurrected this tradition and explored the socio-economic correlates of height (Le Roy Ladurie, Bernageau and Pasquet, 1969), but the true expansion of the field began simultaneously in the mid-1970s among development economists and cliometricians. The former were interested in measuring malnutrition and its synergistic effect on economic performance in the Third World (Scrimshaw, 2003). In cooperation with the United Nations, they expanded the work of nutritionists in combating poverty (Strauss and Thomas, 1998) and measuring the impact of nutrition on labour productivity. Their effort culminated in the United Nation's formulation of the Human Development Index (HDI), which incorporates income, mortality, and schooling, in a superior measure of welfare (Sen, 1987). In contrast, cliometricians analyse secular changes and cross-sectional patterns in biological welfare as well as the effect of economic development on the growth of the human organism. Initial research in this vein was influenced by the controversial finding that American slaves were relatively well nourished (Fogel and Engerman, 1974), and was followed up by investigations of the height of slaves as an indicator of their nutritional status (Engerman, 1976). The results implied that slaves were indeed well-nourished once they reached working age, as they were markedly taller than the European lower classes (Figure 1) as well as their brethren in Africa (Steckel, 1979). This astounding discovery prompted further research along these lines at a time when there was increased dissatisfaction with relying exclusively on gross national product (GNP) per capita as a welfare indicator, as it is not adjusted for income distribution or for externalities such as pollution; moreover, it pertains only indirectly to children and others not in the labour market, such as self-sufficient peasants and women for much of human history. Hence, GNP is only a rough indicator of well-being in a society.
Figure 1
International comparison of height profiles (cm), 18th and 19th centuries. *Sources:* Steckel (1979), Komlos and Cuff (1998).

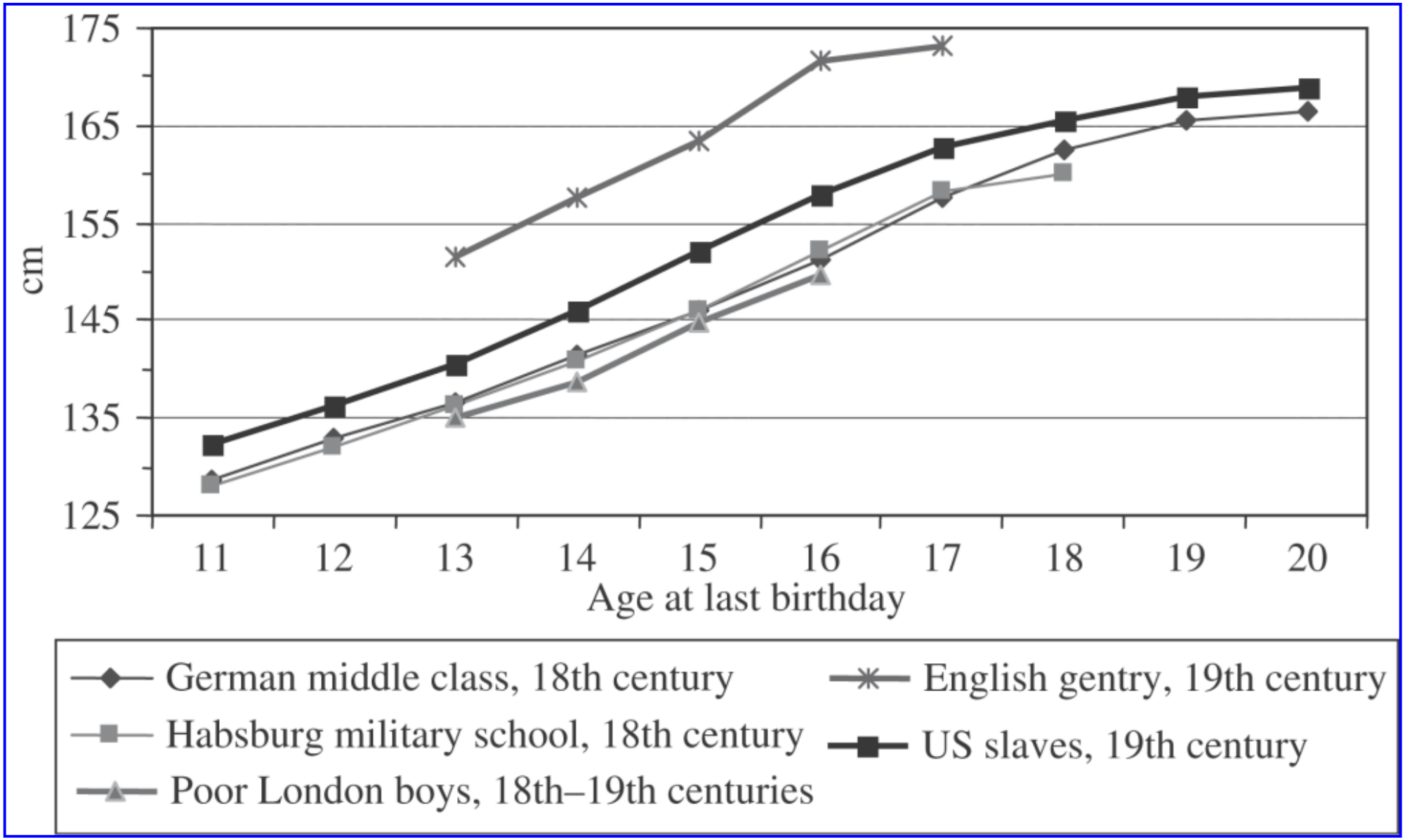

The average height of a birth cohort – until adulthood is given approximately by:

$$H(x)_t = H_{min}(x) + \int_{age=0}^{x} g\left[s_t \cdot Y_t, \left(\frac{P_f}{P_{aog}}\right)_t, W_t, D_t, \sigma_t, \theta_t, M_t, T_t, E_t\right] dt < H_{max}(x),$$

where $H(x)_t$=physical stature at age=x for a particular birth cohort, for x<25, Y_t=real disposable family income; s_t=share of income dedicated to children; P_f=price of nutrients;

P_{aog}=price of all other goods (aog), W_t=work effort; D_t=epidemiological environment, σ_t=detrended variance of income longitudinally from t=0 to t=x (unpredictable income fluctuations might hinder the maintenance of an adequate diet). In turn, children sufficiently deprived will be forced off of their growth profile and may never catch up to their previous growth path; θ_t=cross-sectional inequality of income, M_t=cost of medical services, T_t=transfer payments from governments to families, E_t=environmental conditions (climate), and $H_{min}(x)$ and $H_{max}(x)$ are genetically determined minimum and maximum heights attainable by a given age; with

$$\frac{\partial g}{\partial Y} > 0,\ \frac{\partial^2 g}{\partial Y^2} < 0,\ \frac{\partial g}{\partial \frac{P_f}{P_{aog}}} < 0,\ \frac{\partial g}{\partial W} < 0,\ \frac{\partial g}{\partial D} < 0,\ \frac{\partial g}{\partial \sigma} < 0,\ \frac{\partial g}{\partial \theta} < 0,\ \frac{\partial g}{\partial M} < 0,\ \frac{\partial g}{\partial T} > 0$$

Diminishing returns to income imply that higher income volatility results in shorter stature for a given amount of average income over time. In practice, the analysis frequently pertains to the changes in height over time of adjacent cohorts of adults or of sub-adults of the same age in order to eliminate possible genetic components relevant to H_{min} (x) and H_{max} (x). Thereby one analyses how height is affected by the variables inside the integral over time (Komlos 1985; WHO, 1995). Thus, adult height of a cohort reflects the history of its net-nutritional status during the growing years.

This innovative perspective opened up new windows to understanding of the impact of economic processes on the human organism and vice versa. According to archaeological evidence it is now evident that health of the natives of the New World '…was on a downward trajectory long before Columbus arrived' (Steckel and Rose, 2002, p. 578). There were cycles in physical stature of about a generation long, brought about by demographic growth, urbanization, or changes in relative prices, market structure, income, inequality, and climate (Baten, 2002; Baten and Murray, 2000; Komlos, 1998). There were also shorter cycles in height associated with business cycles (Woitek, 2003); only in the 20th century were these cycles attenuated due to improvements in medicine, increases in labour productivity, and the substantial decline in the relative price of nutrients. The socio-economic crisis of the 17th century is evident in the height of the French population, as men measured about 162 cm on average (Komlos, 2003). Europeans were never as short thereafter. The rapid population growth during the demographic revolution of the late 18th century brought about a decline in height everywhere in Europe as technological change in the agricultural sector did not suffice to maintain the nutritional status of the populations. The French Revolution was preceded by a decline in nutritional status, but no worse than in other parts of Europe, and not to the previous trough of the 17th century. Malthusian crisis generally began with a decline in heights even before mortality rates increased, as human organisms attempted to adjust their size to the available nutrition before the onset of subsistence crisis.

Social status has been related positively to height everywhere and at all times without exception. This generalization holds for 18th-century Germany as well as for the German Democratic Republic in the 20th century. The greatest social gradient in height ever recorded was found in early industrial England, where the difference between upper and lower class 15-year-olds reached 20•cm (Figure 1). Height was related negatively to population density, as denser populations tended to have a higher disease load, as well as higher prices of nutrients. Urban populations tended to be shorter because of higher food prices and because of the higher incidence of diseases until the turn of the 20th century, when perishables became transportable longer distances due to refrigeration, and improvements in urban sanitation improved the epidemiological environment of towns. The degree of commercialization of the economy had an effect on human growth, as propinquity to nutrients invariably conferred considerable nutritional advantages in the early industrial period in so far as self-sufficient consumers did not have to pay for transportation costs of nutrients. Hence, self-sufficient (protein-producing) farmers tended to be tall. This was true in such widely separated places Tennessee, Japan or Bavaria (Cuff, 1998; Craig and Weiss, 1998; Haines, 1998). Americans were the tallest in the world until the middle of the 20th century as resource abundance translated into higher wages, lower food prices, and a more equal distribution of income than prevailed elsewhere.

A transformation in the economic system put a hitherto unknown stress on the human organism. This was the case not only during the neolithic agricultural revolution but also during the Industrial Revolution, during the onset of modern economic growth as well as during the transition from socialism to capitalism. Thus, height declined (in the 1830s) at the onset of modern economic growth even in the resource-abundant United States, a phenomenon that has come to be known as the 'antebellum puzzle'. Average heights declined although real incomes increased (at a rate of 1.4 per cent per annum) because the relative price of nutrients and the degree of inequality were increasing and because self-sufficiency in agriculture was declining (Figure 2).

Figure 2

The puzzling trend in the height of Americans during the antebellum period. *Sources*: Margo and Steckel (1983), Komlos (1987; 1998), Komlos and Coclanis (1997), Weiss (1994).

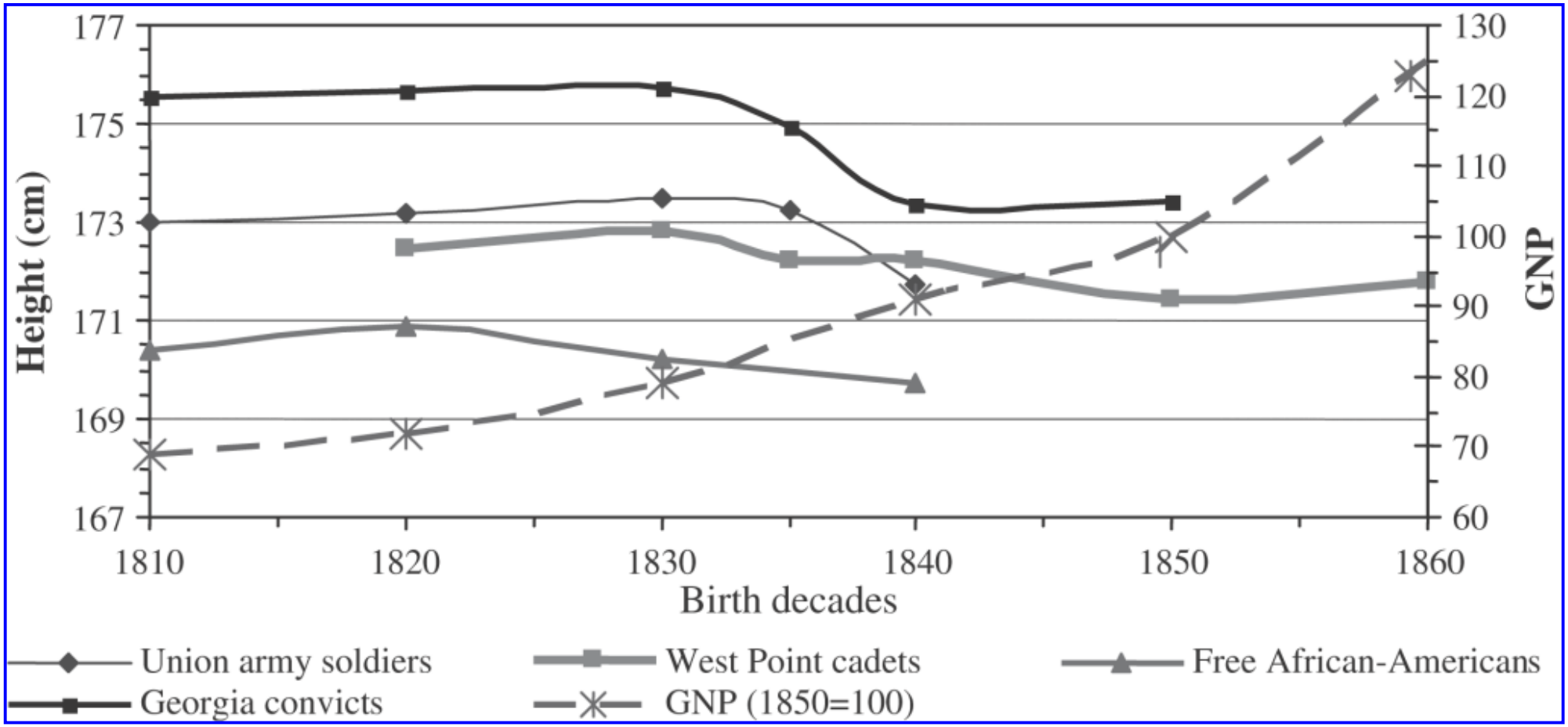

Slaves were well nourished relative to the European lower classes (Figure 1), even if they were not particularly tall in the US context (Figure 3). Income was protective of nutritional status, as one would expect. High-status Americans did not experience a decline in height at the onset of modern economic growth, and the height of aristocrats did not decline during the Industrial Revolution. As Kuznets (1966) demonstrated, the anthropometric record also shows an increase in inequality with industrialization. Heights did not begin to improve substantially and reach their 18th-century levels until the end of the 19th century. Heights tended to correlate positively with wages except in the presence of countervailing forces. Height was associated positively with life expectancy up to about 185 cm; underweight and overweight individuals tended to have lower life expectancy; populations were underweight prior to the mid-20th century as food was relatively expensive and there was a lot of physical activity associated with daily life. Much of the increase in life expectancy in the 20th century is associated with an increase in body size; however, for the first time in its existence, because of technological and cultural changes the human specie is facing an obesity epidemic that threatens to slow down the rate of increase of life expectancy.

Figure 3

Height of US youth, early 19th century. *Sources*: Komlos (1987; 1998), Komlos and Coclanis (1997), Steckel (1979).

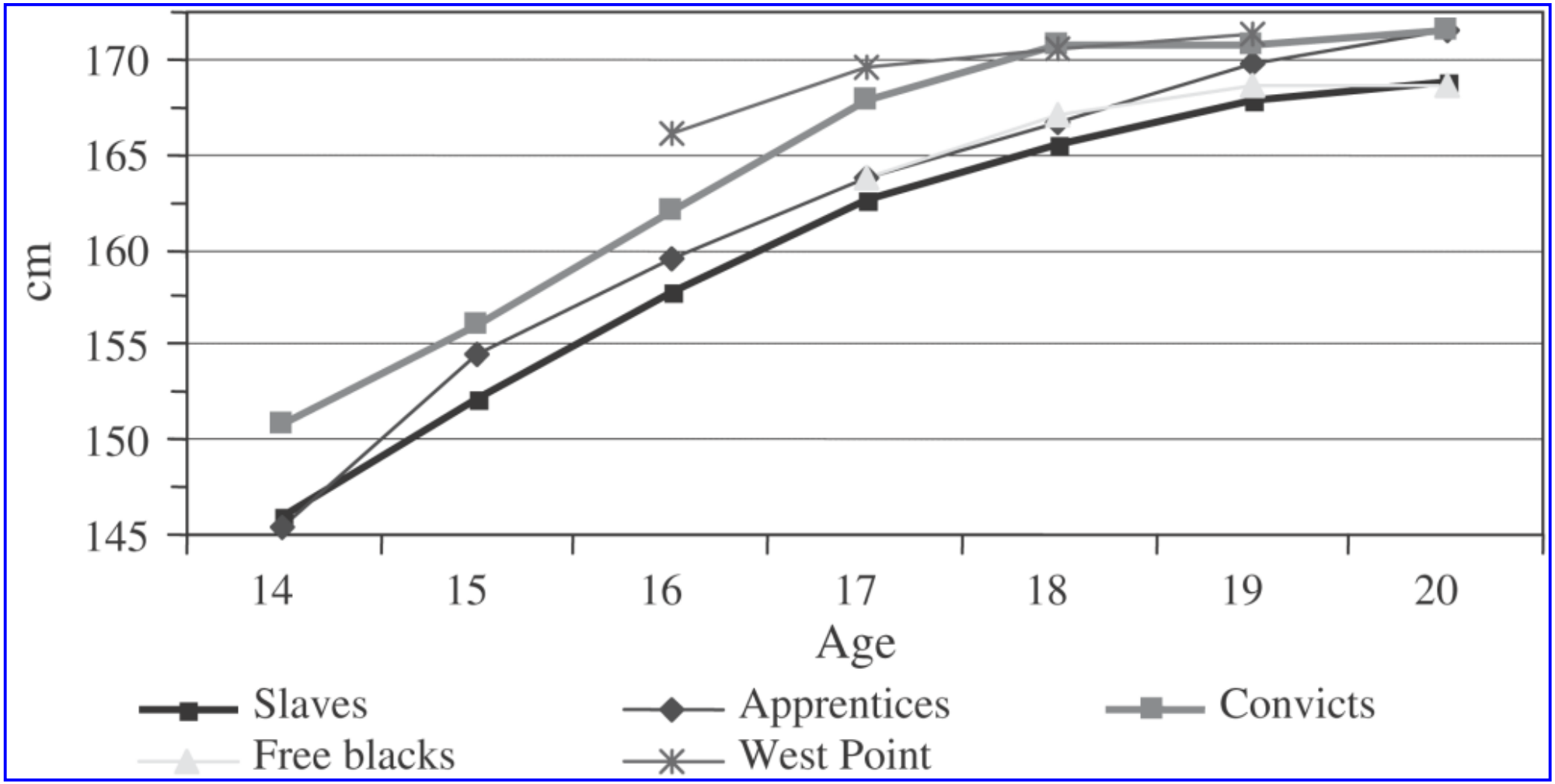

The citizens of the western and northern European welfare states are the tallest in the world now, having overtaken the Americans about a generation ago. That implies that these welfare states provide a higher biological standard of living than the more free-market-oriented American society (Komlos and Baur, 2004).

With the development of the concept of the 'biological standard of living' as distinct from conventional indicators of well-being, and with the founding of the new journal *Economics and Human Biology* in 2003, biology became integrated into mainstream economics. Height and weight are components and relatively easily measured indicators of biological welfare. In addition, we gain new insights of the effect of economic processes on the human organism. Hence, anthropometric history emphasizes that well-being encompasses more than the command over goods and services. Rather, it is multidimensional, and height, weight, health in general, and longevity all contribute to it – independently of purchasing power. In many ways, such indexes provide a more nuanced view of the impact of dynamic economic processes on the quality of life than income or GNP per capita alone. Anthropometric indicators are not meant to be substitutes for, but complements to, such conventional measures of living standards as income per capita.

See Also

- cliometrics
- development economics
- economic history
- environmental Kuznets curve

- family economics
- Fogel, Robert William
- Industrial Revolution
- nutrition and development
- Sen, Amartya

Bibliography

Baten, J. 2002. Climate, grain production and nutritional status in 18th century southern Germany. *Journal of European Economic History* 30, 9–47.

Baten, J. and Murray, J. 2000. Heights of men and women in nineteenth century Bavaria: economic, nutritional, and disease influences. *Explorations in Economic History* 37, 351–69.

Craig, L. and Weiss, T. 1998. Nutritional status and agricultural surpluses in the antebellum United States. In *The Biological Standard of Living in Comparative Perspective*, ed. J. Komlos and J. Baten. Stuttgart: Franz Steiner.

Cuff, T. 1998. Variation and trends in stature of Pennsylvanians, 1820–1860. In *TheBiological Standard of Living in Comparative Perspective*, vol. 1, ed. J. Komlos and J. Baten. Stuttgart: Franz Steiner.

Engerman, S. 1976. The height of US slaves. *Local Population Studies* 16, 45–9.

Fogel, R. and Engerman, S. 1974. *Time on the Cross: The Economics of American Negro Slavery*, 2 vols. Boston: Little, Brown and Co.

Haines, M. 1998. Health, height, nutrition, and mortality: evidence on the ‘antebellum puzzle’ from Union army recruits for New York State and the United States. In *The Biological Standard of Living in Comparative Perspective*, ed. J. Komlos and J. Baten. Stuttgart: Franz Steiner.

Komlos, J. 1985. Stature and nutrition in the Habsburg Monarchy: the standard of living and economic development in the eighteenth century. *American Historical Review* 90, 1149–61.

Komlos, J. 1987. The height and weight of West Point Cadets: dietary change in antebellum America. *Journal of Economic History* 47, 897–927.

Komlos, J. 1998. Shrinking in a growing economy: the mystery of physical stature during the Industrial Revolution. *Journal of Economic History* 58, 779–802.

Komlos, J., in collabaration with Hau, M. and Bourguinat, N. 2003. An anthropometric history of early modern France, 1666–1766. *European Review of Economic History*, 7, 159–89.

Komlos, J. and Baur, M. 2004. From the tallest to (one of) the fattest: the enigmatic fate of the size of the American population in the twentieth century. *Economics and Human Biology* 2, 57–74.

Komlos, J. and Coclanis, P. 1997. On the puzzling cycle in the biological standard of living: the case of the antebellum Georgia. *Explorations in Economic History* 34, 433–59.

Komlos, J. and Cuff, T. 1998. *Classics of Anthropometric History*. St. Katharinen: Scripta Mercaturae.

Kuznets, S. 1966. *Modern Economic Growth: Rate, Structure, and Spread*. New Haven: Yale University Press.

Le Roy Ladurie, E., Bernageau, N. and Pasquet, Y. 1969. Le conscrit et l’ordinateur: perspectives de recherches sur les archives militaires du XIXe siècle français. *Studi Storici* 10,

260–308.

Margo, R. and Steckel, R. 1983. Heights of native born Northern whites during the antebellum period. *Journal of Economic History* 43, 167–74.

Scrimshaw, N. 2003. Historical concepts of interactions, synergism and antagonism between nutrition and infection. *Journal of Nutrition*, 133(1), 316S–21S.

Strauss, J. and Thomas, D. 1998. Health, nutrition, and economic development. *Journal of Economic Literature* 36, 766–817.

Sen, A. 1987. *The Standard of Living*. Cambridge: Cambridge University Press.

Steckel, R. 1979. Slave height profiles from coastwise manifests. *Explorations in Economic History* 16, 363–80.

Steckel, R. 1995. Stature and the standard of living. *Journal of Economic Literature* 33, 1903–40.

Steckel, R. and Rose, J., eds. 2002. *The Backbone of History: Health and Nutrition in the Western Hemisphere*. New York: Cambridge University Press.

Weiss, T. 1994. Economic growth before 1860: revised conjectures. In *Economic Development in Historical Perspective*, ed. D. Schaefer and T. Weiss. Stanford: Stanford University Press.

WHO (World Health Organization). 1995. *Physical Status: The Use and Interpretation of Anthropometry*. Technical Report Series No. 854. Geneva: WHO.

Woitek, U. 2003. Height cycles in the 18th and 19th centuries. *Economics and Human Biology* 1, 243–58.

How to cite this article

Komlos, John. "anthropometric history." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000222> doi:10.1057/9780230226203.0044

The New Palgrave Dictionary of Economics Online

anti-discrimination law

John J. Donohue III

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the major legislative initiatives outlawing discrimination, discusses the theoretical arguments for and against such initiatives, and assesses the impact of these laws on the groups they try to protect. The significant effects of federal law in the first decade after passage of the 1964 Civil Rights Act are contrasted with the less optimistic findings from subsequent anti-discrimination interventions. Insights about the social benefits and the costs of the unintended consequences of employment discrimination law apply equally to other types of anti-discrimination legislation, such as mortgage lending and policing.

Keywords

anti-discrimination law; Becker, G.; black–white labour market inequality in the United States; efficient capital markets hypothesis; Friedman. M.; human capital; insider–outsider problems; Jim Crow South; labour market discrimination; sex discrimination; statistical discrimination

Article

In the aftermath of the Second World War, New York and New Jersey became the first in a series of non-Southern states to pass laws prohibiting racial discrimination in employment. Almost two decades later Congress passed, over strong Southern opposition, the momentous Civil Rights Act of 1964, which banned discrimination on the basis of race, sex, religion and national origin in employment and public accommodations. Over the ensuing 40 years, the reach of federal and state anti-discrimination law has extended beyond intentional discrimination (disparate-treatment discrimination) to ostensibly neutral practices that have an adverse impact on selected groups (disparate-impact law), and to protect those over age 40 (the Age Discrimination in Employment Act) and those with disabilities (the Americans with Disabilities Act). Anti-discrimination law has come to play an increasingly important role in employment, government contracting, policing and criminal justice, mortgage lending, retail and

marketing practices, and education.

The Becker model, federal anti-discrimination law, and the end of the Jim Crow era

In 1957, Gary Becker launched the serious economic evaluation of discrimination when he developed a model based on individual animus towards a certain class of workers (see Becker, 1957). The analysis had a number of shortcomings when applied to the real world, not least of which was that it assumed that the psychological burden of discrimination fell only on the *discriminators* (they were the ones who suffered the distaste), and the only cost borne by the victims of discrimination was any resulting decrease in wages or employment. In the early 1960s, Milton Friedman, in part influenced by Becker's work, argued against employment discrimination law on the grounds that it was unnecessary since competitive markets would protect workers from discrimination, and undesirable since government should not interfere with the personal preferences of discriminating employers. Although it is now clear that Friedman's position was incomplete, both arguments carry some weight.

First, frictionless competitive markets should offer protection from discriminatory employers. This means that, even in the presence of substantial employer animus, highly competitive markets reduce the need for law if a sufficient number of non-discriminators are available to bid up the wages of, say, black labour. The efficient capital markets hypothesis assumes that prices of financial assets will always tend to be close to their underlying value. Workers are also valuable assets, so Friedman believed that competitive markets would similarly push wages towards underlying productivity (‘true value’) in the labour market as well. But, even under the best of circumstances, one would not expect labour markets to be as efficient as capital markets with their homogeneous products, low transaction costs, ability to sell short and hordes of analysts whose job it is to identify the true value of certain securities. The resulting trades will tend to push these stock prices towards their true value (Donohue, 1994). In the labour market, workers are not homogeneous, transaction costs associated with hiring and dealing with labour are high, there is no ability to sell short, and the value in ascertaining the true productivity of a modal worker is relatively small. If one adds in labour market imperfections posed by unions, minimum wage laws, high information costs and the racist and segregationist Jim Crow laws – laws requiring strict racial segregation in many aspects of public life including schooling and accommodations that led to inferior treatment of blacks despite the supposed legal requirement of equality under the ‘separate but equal’ doctrine – it is not hard to imagine that, in the absence of anti-discrimination legislation, blacks would be unfairly excluded from a range of good jobs or paid less than their marginal product.

Moreover, while competitive markets would be hostile to employer discrimination, they would actually encourage an employer to discriminate if that is the preference of fellow workers or customers.

Moreover, the empirical evidence demonstrated clearly that, whether from the pressures of racist norms or governmental encumbrances, the market afforded little protection to black workers in major industries of the South, such as Southern textiles (Heckman and Payner, 1989). The major federal intervention directed against the Jim Crow policies of the South beginning with the 1964 Civil Rights Act did what competitive markets had failed to achieve – open up entire industries to qualified black workers and substantially dampen the black shortfall in earnings vis-à-vis white workers (Donohue and Heckman, 1991).

Second, under the Becker model, net utility will be decreased by an employment discrimination law if

one gives weight to the preferences of discriminators, as Friedman and Becker were wont to do. But Donohue (1986; 1989) argued that driving the discriminators out of business could actually enhance welfare by eliminating the Beckerian social cost. Moreover, while Becker conceived of discrimination as a stable taste, the evidence again suggests that the federal prohibition ultimately changed the attitudes (tastes) of millions of Americans. Rather than relentlessly and constantly imposing the burdens of inefficient interactions on unwilling discriminators, the Civil Rights Act aided a social process of integration that ultimately reduced the prior Beckerian taste for discrimination. While short-run costs were undoubtedly high, in the long run an entire region of the country was energized by the disruption of previously regimented views of racial inferiority – to the benefit of both blacks *and whites*. Since the Beckerian discriminatory tastes represented social costs, the reduction in the magnitude of these social costs constituted a major social benefit.

Did federal law improve the economic status of blacks and others?

Perhaps the most important question concerning federal anti-discrimination law is whether it has aided its primary intended beneficiaries – black Americans (particularly in the South). James Smith and Finis Welch (1989) argued that the Civil Rights Act of 1964 was not responsible for substantial gains in black economic welfare. They conceded that black economic welfare improved at about the time of the federal initiatives in the 1960s, but they contended that the gains were the result of human capital enhancement, not of demand-side policies addressed to ameliorate the impact of discrimination. To buttress their view that Title VII – the section in the Civil Rights Act prohibiting employment discrimination based on, inter alia, race or colour – generated no benefits for black workers, Smith and Welch argued that the economic gains of blacks during the period 1940–60 were the same as those in the 1960–80 period (thereby suggesting that the Civil Rights Act of 1964 had been unimportant). The major response to Smith and Welch came from Donohue and Heckman (1991), who argued that Title VII did indeed generate a decade of economic gains for blacks:

> …the evidence of sustained economic advance for blacks over the period 1965–1975 is not inconsistent with the fact that the racial wage gap declined by similar amounts in the two decades following 1940 as in the two decades following 1960. The long-term picture from at least 1920–1990 has been one of black relative stagnation with the exception of two periods – that around World War II and that following the passage of the 1964 Civil Rights Act. (Donohue and Heckman, 1991, p. 1614)

It is now widely accepted that, in helping to break down the extreme discriminatory patterns of the Jim Crow South, Title VII considerably increased the demand for black labour, leading to both greater levels of employment and higher wages in the decade after its adoption (see also, Freeman et al., 1973; Conroy, 1994; and Orfield and Ashkinaze, 1991). Chay (1998) shows that, when the reach of the 1964 Civil Rights Act was expanded in 1972, the demand for black labour was further stimulated. But the good news in terms of law-induced efforts to improve the economic status of blacks through anti-discrimination policy has probably run its course. A series of papers by Oyer and Schaefer (2000; 2002a; 2002b) offers little support for the view that the strengthening of federal anti-discrimination law in 1991

stimulated black or female employment, as occurred with the federal laws passed in 1964 and 1972. (The CRA actually changed race discrimination law in a relatively minor way – restoring the standards that had existed in June 1989 with respect to discriminatory discharge and the standards for employer justification of practices with disparate racial impacts. For non-race cases, however, the 1991 Act expanded the damages available and authorized punitive damage awards for intentional discrimination.) Moreover, papers by Acemoglu and Angrist (2001), and DeLiere (2000) hold that another piece of anti-discrimination legislation, the Americans with Disabilities Act (ADA), actually harms employment. This very pessimistic conclusion may be too strong. Attributing the poorer employment experience of the disabled in a short period after the federal law passed in 1990 turns out to be a tricky proposition, given the downturn in the economy and the substantial growth in those collecting disability benefits at roughly the same time. Burkhauser, Houtenville and Rovba (2006) extend the time period of Acemoglu and Angrist's analysis, and conclude that the decline in relative employment of the disabled actually began in the mid-1980s, roughly the time at which rules for disability benefits eligibility were loosened. But even if the ADA did not hurt, there is no strong evidence that it helped on the macro level, even if it did assist in securing small micro-level accommodations for the disabled. Jolls and Prescott (2004) argue that disability laws having a reasonable accommodation requirement may generate an insider–outsider problem. Those who gain the accommodation are better off, but at the expense of some disabled workers who end up out of the labour force.

Is employment discrimination a first-order problem for US blacks today?

Is the black – white earnings differential fully explained?

Heckman (1998) contends that labour market discrimination no longer substantially contributes to the black–white wage gap (as it once clearly did), and therefore he doubts that four decades after the Civil Rights Act racial discrimination in the labour market is a first-order problem in the United States. Rather, Heckman looks to other factors (namely, those that promote skill formation) to explain the black–white earnings gap – a theme that he builds on in Carneiro, Heckman, and Masterov (2005).

An important paper that informs Heckman's analysis of the current reasons for the black–white wage gap is Neal and Johnson (1996). If factors that exist prior to workers' entry into the labour market largely explain the black–white wage gap, then the contribution of racial discrimination to this wage gap is presumably small. Neal and Johnson note that many studies have examined the black–white wage gap and found that it could not be explained with standard measures such as age, years of education, marital status and so forth, implying that the contribution of discrimination was sizable. Neal and Johnson note that years of education may exaggerate the true skill level attained by blacks, given the poorer-quality schools that many blacks attend. They argue that scores on the Armed Forces Qualification Test (AFQT) are a better measure than innate ability of acquired skill brought to the labour market.

The authors begin by showing that the *unadjusted* wage gap between blacks and whites is minus 24.4 per cent for black men and minus 8.5 per cent for black women. Using National Longitudinal Surveys of Youth (NLSY) data, Neal and Johnson found that the unexplained wage gap fell to minus 7.2 per cent for black men and plus 3.5 per cent (although insignificant) for black women, once they controlled for race, age and AFQT score. In other words, the AFQT test score can explain a very large portion of the

black–white wage gap for men, and the entire gap for women. One source of continuing debate in the literature is whether these wage regressions should include controls for years of education as well as AFQT score. Neal and Johnson say it should not since the test better captures ability, and so they exclude the education measure from their regressions. Others have included years of education and find that the unexplained wage gap re-emerges when this control is added.

A potential problem with their approach is the possibility of black underinvestment in human capital due to the presence of statistical discrimination. Neal and Johnson reject this concern, finding that the return to higher AFQT scores is significantly higher for black men (although not for black women), so that blacks seem to have adequate incentive to invest in developing human capital.

Evidence of racial discrimination in entry level hiring from audit-pair studies

The view that racial discrimination seems to have largely been wrung from the labour market is in apparent conflict with a number of audit studies that document differential treatment of blacks and whites. For example, a recent study by Devah Pager concludes that the degree of discrimination in employment is so great that blacks without criminal records are treated as badly as whites with criminal records (Pager, 2003). Pager's audit experiment involved four male participants, two blacks and two whites, applying for entry-level job openings. The auditors formed two teams so that the members of each team were of the same race (the only difference in the application was that one of the testers in each team was assigned a criminal record, a felony drug conviction, and 18 months of prison). The teams applied for 15 jobs per week and the final data included 150 applications by the white pair and 200 by the black pair. The auditors applied for the jobs and advanced as far as they could during the first visit. The application was considered a success only if the auditors were called back for a second interview or hired.

The results showed that 34 per cent of whites with no criminal record were called back, compared with only 17 per cent of those with a criminal record; 14 per cent of blacks without a criminal record were called back, compared to only 5 per cent with a criminal record. Notably, the black auditor without a criminal record received a smaller percentage of callbacks than the white auditor with a criminal record, suggesting the presence of substantial discrimination against blacks in general. Note that Pager found a greater disparity than that found in other audit pair studies in the employment realm. Pager's approach has one notable advantage: the black pair and the white pair were able to use *identical* sets of résumés, which would not have been possible had they been visiting the same employers (the résumés of test partners were similar but not identical). Some have also raised concerns about whether experimenters might have been influenced by the goals of the study to ‘find discrimination’. (This is the ‘experimenter’ effect that Heckman and Siegelman, 1993, discuss in the context of the Urban Institute audit studies and that social psychologists have long recognized.)

Bertrand and Mullainathan (2004) also try to measure the extent of race-based labour market discrimination using a slightly different audit strategy that avoids some of the potential pitfalls of direct applicant auditing. Employing a so-called correspondence test methodology, they submitted about 5,000 fictitious résumés in response to employment advertisements appearing in Boston and Chicago newspapers. Their experiment was designed to estimate the racial gap in response rates, measured by phone calls or e-mails requesting an interview. Random application of traditional black or white names to résumés ensures (*a*) that race remains the only component that varies for a given résumé and (*b*) that

heterogeneous responses to behaviour or appearance do not affect outcomes (as often occurs with human auditors).

The Bertrand and Mullainathan paper differs from Pager's audit study in that no personal contact with the potential employer takes place in their experiment, so perceived problems with auditor behaviour are eliminated. Bertrand and Mullainathan find significant differences in callback rates for whites and blacks: ‘applicants with White names need to send about 10 résumés to get one callback whereas applicants with African–American names need to send about 15 résumés’ (Bertrand and Mullainathan, 2004, p. 3). Put differently, the advantage of having a distinctly white name translates into roughly eight additional years of experience in the eyes of a potential employer. Whites also appear to benefit much more than blacks from possessing the skills and attributes of a high-quality applicant and from living in a wealthier or whiter neighbourhood. (The difference in callback rates between high and low quality whites is 2.3 percentage points, while for blacks the difference is a meagre one half a percentage point.)

Although these results represent compelling evidence of unlawful discriminatory conduct by employers, the question remains whether the markets are robust enough to reduce or eliminate the apparent disadvantage in the initial hiring process. Fryer and Levitt (2004) indicate that distinctive names do not disadvantage blacks for a variety of adult outcomes. They offer some potential arguments for reconciling their findings with those of Bertrand and Mullainathan (2004). First, if names are considered a noisy initial indicator of race, then they should have no effect once a candidate arrives for the interview. Second, if distinctively black names damage labour market prospects, one might observe more name changes than appear to occur. Finally, with only about ten per cent of jobs being secured through formal résumé-submission processes, the disadvantage of being screened out by certain employers may not be high when other employers and other job search paths remain open.

The combination of the audit studies and the better regression studies seems to tell us that (*a*) there are enough discriminators around for blacks to have to search harder to find employment, (*b*) there are enough non-discriminators around for the resulting unexplained earnings shortfall to be not very high, and (*c*) the unexplained earnings shortfall will overstate discrimination if other legitimate factors are omitted, but will understate the cost of discrimination to blacks because they bear the added search costs of the higher level of employer rejection and any attendant psychological burden that it imposes. To eliminate discrimination would narrow the unexplained earnings gap and remove the added search costs, but this would still leave a substantial unadjusted disparity in black and white earnings.

Statistical discrimination

A number of theoretical articles have explored whether statistical discrimination contributes to the black–white earnings gap (Arrow, 1973; Phelps, 1972). This seems unlikely. If, say, blacks are on average treated as their productivity would warrant, then as a class there should be no earnings shortfall, apart from the issue of underinvestment that was discussed above with reference to the Neal and Johnson paper. David Autor and David Scarborough (2004) explore the impact on the hiring and productivity of minority workers, using data from a large nationwide retail firm that changed from an informal worker selection process to one based on standardized testing in 1999. Given that minorities and underprivileged groups on average score lower on such standardized tests, one would expect that this change in the firm's hiring scheme would disadvantage minority workers.

The company originally used informal, paper applications to select candidates for entry level positions. Starting in June 1999, the firm began instituting a computer-based application system that included a personality test for selecting compatible and potentially productive candidates. Autor and Scarborough's sample contains information on test scores, worker demographics, termination date and termination reason (if applicable) for hires made between January 1999 and May 2000 in all the firm's outlets; their sample consists of 34,257 observations. The question they address is how the introduction of testing and the ensuing improvement in the firms' applicant selection procedure affected minority hiring and productivity.

Autor and Scarborough show that if employers statistically discriminate before the test is introduced – that is, if they already use demographic characteristics as a signal for expected productivity of the candidate – then adding testing to the model does not hurt minority hiring but still increases the average productivity of both minority and non-minority workers. The empirical evidence supports this last scenario, revealing uniform increased productivity across demographic groups along with no negative effects on minority hiring.

While we must be careful not to extrapolate the Autor and Scarborough results too far from their context of entry level, near-minimum wage jobs, the paper suggests that before testing was implemented the retail firm either selected workers based on (*a*) some non-race proxy that was correlated (imperfectly) with productivity, or (*b*) statistically discriminated on the basis of race (in violation of federal law), which was itself (imperfectly) correlated with productivity. The evidence from this one firm confirms the intuition of many economists that statistical discrimination should not be unlawful since on average it should not disadvantage minority workers. One should query, though, whether the legal regime is nuanced enough to legitimize statistical discrimination while prohibiting intentional, animus-based discrimination. Judicial and jury determinations of such issues would presumably be subject to high levels of Type I (incorrectly finding discrimination) and Type II (incorrectly failing to find discrimination) errors.

Sex discrimination in employment

Many of the issues discussed above with respect to race discrimination are also relevant to other types of discrimination, including sex discrimination. First, there are questions about whether anti-discrimination law has helped the protected worker. Second, there are issues about whether discrimination can be accurately established. Almost all the groups that seek the aid of anti-discrimination law – minorities, women, the disabled, the elderly – have attributes that non-discriminatory employers might be legitimately concerned about. Under such circumstances, it is difficult to prove that under-representation of any of these groups is caused by discrimination rather than some legitimate factor. The original goal of employment discrimination law in the United States was to eliminate any gap between a worker's productivity and pay caused by discrimination. Today, some argue that the goal of mimicking the outcome of perfectly competitive labour markets is insufficient and that employment discrimination law should more aggressively pursue broader goals of social fairness that will enhance the economic status of disadvantaged groups beyond what a perfect market would provide. According to this view, women should be treated differently to ensure that their role in child-bearing does not disadvantage them in the labour market even if it imposes costs on employers.

Claudia Goldin and Cecelia Rouse (2000) offer an interesting illustration of establishing labour market

discrimination in the context of auditions and hiring of musicians for the major US orchestras. To test for sex discrimination in the hiring process, they exploit the changes in the audition process introduced by all major US orchestras in the 1970s and 1980s. Of particular interest for their study was the change to ‘blind’ auditions, which effectively hid the identity and gender of the applicant from the hiring committee for certain rounds of the audition process. Using audition and roster data spanning several decades and employing an individual fixed effect strategy, they found that the likelihood of female hiring and advancement was increased by the introduction of blind auditions.

More specifically, using audition data from the late 1950s to 1995, Goldin and Rouse found that in blind audition rounds women were as much as 50 per cent more likely to advance from preliminary to final rounds. Furthermore, the likelihood of women winning the finals increased by 33 percentage points if the final round was blind. Using official roster data from 1970 to 1996, they found that completely blind auditions – defined as auditions in which all rounds are conducted with a screen hiding the gender of the applicant – increased the likelihood of a women being hired by 25 per cent. Based on the roster data, blind auditions explain 30 per cent of the increase in female hiring and 25 per cent of the increase in overall female representation in the orchestras. There are, however, some caveats with respect to these findings: first, some estimates have relatively large standard errors that render them statistically insignificant; second, in one scenario – auditions with blind semi-finals – the effect on females is persistently strongly negative.

The issue of gender differences in aptitude, specifically aptitude in competitive environments, is explored in an article by Uri Gneezy, Muriel Niederle, and Aldo Rustichini (2003). Unlike previous studies that tried to explain the gender gap either through occupational self-selection due to differences in abilities and preference or through employer discrimination, Gneezy, Niederle and Rustichini explore the possibility of gender-differentiated performance in competition, which could ‘reduce the chance of success for women when they compete for new jobs, promotions, etc’. In a series of controlled experiments the authors examine the performance of men and women in a computerized maze game as they vary the incentive schemes and group composition for different treatments. They find that, while men receive a significant performance boost in competitive environments such as tournaments, the response of women in competitive environments is more nuanced: they do not significantly change their performance in mixed-sex tournaments, but they do increase their performance in single-sex competitions.

The authors find that under a piece-rate payment scheme men perform only slightly (and not significantly) better than women on average in terms of number of mazes solved. However, when the authors introduce their main competitive treatment of mixed-sex tournaments, they find that men increase their performance significantly, while women's performance remains relatively unchanged. While women do not seem to receive a performance boost in mixed competitive environments, Gneezy, Niederle and Rustichini also use single-sex tournaments to show that there *are* competitive situations where women increase their performance in response to competition. Both women and men significantly increase their performance in single-sex tournaments, suggesting that women do not dislike competition in general; rather, they dislike competing against men. To explain this, the authors also test for varying feelings of competence across gender. Indeed, once they allow men and women to choose the level of difficulty of the mazes that they are to solve, men choose a higher level of difficulty on average than women do. Whether such factors could explain different pay levels between male and female workers operating under merit pay systems – such as, the lower pay of female stockbrokers, which has been a

subject of sex discrimination litigation – is a question that will probably be further explored in the courtroom as well as in the academy.

Conclusion

Anti-discrimination law has generated a number of important social benefits. The elimination of the oppressive race code of the South has been a major benefit of law and policy, opening up all jobs to the most highly qualified candidates. The development of a strong anti-discrimination norm has been an important social asset, and one that merits preservation. To the extent that employers find it natural to be fair to all applicants and workers, the burdens on workers, courts, and employers will be lessened, to the benefit of all.

At the same time, anti-discrimination law has generated some unfortunate unintended consequences, some of which may even threaten the important anti-discrimination norm by undermining its widespread acceptance. I have already alluded to the perverse effects of the situation where an employer might avoid hiring a particular protected worker because of the presence of a governing anti-discrimination law, as some have argued with respect to the protections mandated by the Americans with Disabilities Act. In a regime where the difficulties in ascertaining the existence of discrimination lead to Type I error, firms might find that they are being compelled to hire and compensate certain workers at wages above their levels of productivity. Similarly, as with any negligence-type standard where being adjudicated to have been below a certain level of care can lead to substantial damage awards (including punitive damages), firms have an incentive to take costly measures to be above the threshold that might lead to a finding of discrimination. Tests that may be useful in selecting a high-quality workforce may be avoided if they have, or are thought to have, a disparate impact on certain protected workers that could provide the basis for costly litigation. Note that all these employer adjustments involve costs, but they would appear to involve the benefit of enhancing the employment of groups that are relatively disadvantaged. One might argue that this is a positive development in terms of distributive justice even if it is not actually furthering a corrective justice rationale of eliminating discrimination.

But of course if costs are being imposed on businesses, they will have an incentive to avoid them in the cheapest way possible, which might be through compliance with the legal mandates but could also involve efforts to circumvent the legal mandates. Indeed, because movements in either direction from the ‘non-discriminatory equilibrium’ can lead to litigation by whites or blacks or males or females, firms may at times take measures to avoid the litigation risks by using temporary help or by sending their jobs offshore. If these issues were to arise in a racial discrimination context, firms might decide to move offices out to suburban areas or locate where the requirements for hiring black workers would be lessened by the smaller minority benchmark percentages in the relevant labour markets.

As Donohue and Siegelman (1991) noted, the nature of employment discrimination litigation has changed very dramatically in a way that was not anticipated and which may not be entirely desirable. Specifically, most early cases of discrimination complained of failure to hire. These suits tended to open up whole industries or occupations to formerly excluded workers, thereby furthering the objectives of the law. Over time, however, there has been a massive shift in the direction of discharge lawsuits where protected workers claim that they were discriminated against when they were fired. This change sometimes means that low productivity workers can threaten Title VII litigation to hold up an employer for a higher severance package when they are fired for cause. Even worse, firms may find that, at the

margin, it is safer not to hire additional protected workers because, at the margin, firms face greater risks from possible, future wrongful discharge discrimination lawsuits than from failure to hire cases. An overall assessment of the impact of anti-discrimination law needs to examine not only the obvious benefits in the form of better treatment of workers through greater professionalization in hiring and human resource management and the productivity enhancements from selecting workers in non-discriminatory ways, but also the array of costs in terms of non-optimal employee selection and retention and firm location decisions, more costly selection processes, and greater litigation costs and legal consulting fees. When every discharge carries the potential for an award of punitive damages, the costs of getting rid of even quite poor workers becomes high. Thus, it may not be surprising that, once the extreme forms of discriminatory conduct were eliminated in the wake of the initial passage of the 1964 Act, further efforts at ratcheting up enforcement of anti-discrimination law seem not to have generated added benefits. Similar arguments about the costs of unintended consequences apply to anti-discrimination enforcement in the realms of mortgage lending, consumer purchases, policing and fighting terrorists.

See Also

- black–white labour market inequality in the United States
- gender roles and division of labour
- Jim Crow South
- real wage rates (historical trends)
- social networks in labour markets
- search models of unemployment

Bibliography

Acemoglu, D. and Angrist, J.D. 2001. Consequences of employment? The case of the Americans with Disabilities Act. *Journal of Political Economy* 109, 915–57.

Arrow, K.J. 1973. The theory of discrimination. In *Labor Economics*, vol. 4, ed. O.C. Ashenfelter and K. F. Hallock. Aldershot: Edward Elgar.

Autor, D. and Scarborough, D. 2004. Will job testing harm minority workers? Working Paper No. 10763. Cambridge, MA: NBER.

Becker, G. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bertrand, M. and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94, 991–1011.

Burkhauser, R.V., Houtenville, A.J. and Rovba, L. 2006. Accounting for the declining fortunes of

working-age people with disabilities. Paper presented at the New York University School of Law Center for Labor and Employment Law Symposium on Assessing the Employment Provisions of the Americans with Disabilities Act, 7 April.

Carneiro, P., Heckman, J. and Masterov, D. 2005. Labor market discrimination and racial differences in premarket factors. *Journal of Law and Economics* 48, 1–40.

Chay, K. 1998. The impact of Federal civil rights policy on black economic progress: evidence from the Equal Employment Opportunity Act of 1972. *Industrial and Labor Relations Review* 51, 608–32.

Conroy, M. 1994. *Faded Dreams: The Politics and Economics of Race in America*. Cambridge: Cambridge University Press.

DeLiere, T. 2000. The wage and employment effects of the Americans with Disabilities Act. *Journal of Human Resources* 35, 693–715.

Donohue, J.J. 1994. Employment discrimination law in perspective: three concepts of equality. *Michigan Law Review* 92, 2583–612.

Donohue, J.J. 1986. Is Title VII efficient? *University of Pennsylvania Law Review* 134, 1411–31.

Donohue, J.J. 1989. Prohibiting sex discrimination in the workplace: an economic perspective. *University of Chicago Law Review* 56, 1337–68.

Donohue, J.J. and Heckman, J. 1991. Continuous versus episodic change: the impact of civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29, 1603–43.

Donohue, J.J. and Siegelman, P. 1991. The changing nature of employment. *Stanford Law Review* 43, 983–1033.

Freeman, R., Gordon, R.A., Bell, D. and Hall, R.E. 1973. Changes in the labor market for Black Americans, 1948–72. *Brookings Papers on Economic Activity* 1973(1), 67–131.

Fryer, R.G. and Levitt, S.D. 2004. The causes and consequences of distinctively black names. *Quarterly Journal of Economics* 119, 767–805.

Gneezy, U., Niederle, M. and Rustichini, A. 2003. Performance in competitive environments: gender differences. *Quarterly Journal of Economics* 118, 1049–74.

Goldin, C. and Rouse, C. 2000. Orchestrating impartiality: the impact of ‘blind’ auditions on female musicians. *American Economic Review* 90, 715–41.

Heckman, J.J. 1998. Detecting discrimination. *Journal of Economic Perspectives* 12(2), 101–16.

Heckman, J. and Payner, B. 1989. Determining the impact of federal antidiscrimination policy on the economic status of blacks: a study of South Carolina. *American Economic Review* 79, 138–77.

Heckman, J. and Siegelman, P. 1993. The Urban Institute Audit Studies: Their Methods: Response to Comments by John Yinger. In *Clear and Convincing Evidence: Measurement of Discrimination in America*, ed. M. Fix and R. Struyck. Washington, DC: Urban Institute Press.

Jolls, C. and Prescott, J.J. 2004. Disaggregating employment protection: the case of disability discrimination. Working Paper No. 10740. Cambridge, MA: NBER.

Neal, D.A. and Johnson, W.R. 1996. The role of premarket factors in black–white wage differences. *Journal of Political Economy* 104, 869–95.

Orfield, G. and Ashkinaze, C. 1991. *The Closing Door: Conservative Policy and Black Opportunity*. Chicago: University of Chicago Press.

Oyer, P. and Schaefer, S. 2000. Layoffs and litigation. *RAND Journal of Economics* 31, 345–58.

Oyer, P. and Schaefer, S. 2002a. Litigation costs and returns to experience. *American Economic Review* 92, 683–705.

Oyer, P. and Schaefer, S. 2002b. Sorting, quotas, and the Civil Rights Act of 1991: who hires when it's hard to fire? *Journal of Law and Economics* 45, 41–68.

Pager, D. 2003. The mark of a criminal record. *American Journal of Sociology* 108, 937–75.

Phelps, E.S. 1972. The statistical theory of racism and sexism. *American Economic Review* 62, 659–61.

Smith, J.P. and Welch, F.R. 1989. Black Economic Progress after Myrdal. *Journal of Economic Literature* 27, 519–64.

How to cite this article

Donohue, John J., III. "anti-discrimination law." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000254> doi:10.1057/9780230226203.0045

The New Palgrave Dictionary of Economics Online

antidumping

Bruce A. Blonigen Thomas J. Prusa

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Antidumping is a legal statute that allows for a remedy to offset the effects of dumped imports. Antidumping has emerged as the preferred method of trade protection, accounting for more disputes than all the other trade statutes combined. The economic rationale for current antidumping statutes is weak and generally inconsistent with competition policies. Empirical evidence suggests that antidumping activity is motivated by the same political economy considerations that lead to other forms of trade protection. The economic impact of antidumping remedies can be significant, often dramatically reducing import flows and imposing welfare costs as great as any current trade distortion.

Keywords

antidumping; cartels; collusion; competition policy; fair value; predatory pricing; price discrimination; protection; strategic behaviour; trade diversion; voluntary export restraints

Article

Antidumping refers to a legal statute that allows for a remedy (typically an import duty) to offset the effects of dumped imports. Under the General Agreement on Trade and Tariffs/World Trade Organization (GATT/WTO) rules, two tests must be satisfied before a country may impose an antidumping duty on subject imports. First, the imports must be shown to be sold at price that is ‘less than fair value’. Second, the dumped imports must be shown to have caused or threaten to cause ‘material’ injury to a domestic industry.

History and institutions

The first antidumping (AD) statutes were established in Canada and the United States in the early 1900s. Ultimately, these statutes have been codified into the GATT/WTO statutes. Until the mid-1980s almost

all AD activity was confined to four major countries/regions – the United States, the European Union, Australia and Canada (Finger, 1993). By the early 1990s countries with newly adopted antidumping statutes accounted for almost one-quarter of AD cases and, since the mid-1990s, new antidumping countries have accounted for well over half of AD complaints (Miranda, Torres and Ruiz, 1998; Prusa, 2001). These new antidumping countries are also far more likely to make an affirmative determination and, consequently, now account for far more than half of all measures in place. Since 1980 GATT/WTO members have filed more complaints under the AD statute than under all other trade laws combined. Worldwide, more AD duties are now levied in any one year than were levied in the entire period from 1947 to 1970.

An antidumping investigation generally proceeds as follow, though there are differences across countries. First, an investigation is initiated when an interested party (often a domestic industry that competes with the imported product) files a petition with the appropriate government agency contending dumping of a particular product(s) from certain import-source countries. The administering government agency (or agencies) then collects data from petitioners and foreign firms that are alleged to be the source of dumped imports and calculates the extent to which imports have been dumped and have injured the domestic industry. Findings of dumping and material injury lead to the imposition of an antidumping duty, which is often equal to the per cent difference between the price of the dumped imports and fair value (that is, the dumping margin). Under WTO statutes, antidumping cases must be reviewed at least every five years to determine whether an antidumping remedy is still appropriate given recent import activity in the subject product.

It is important to understand that antidumping arises from legal concepts. Thus, the meaning of ‘less-than-fair-value’, causation, and material injury are examined from a legal perspective where previous rulings establish precedence in interpreting the legal definitions. Legal bodies have been active in adjusting these statutes over time. The GATT/WTO antidumping code has undergone significant revisions in nearly every negotiating round, and most countries with these statutes also make periodic legislative changes to their antidumping codes. Many economists have noted that the increase in antidumping activity after these legislative changes is not coincidental. For example, the Tokyo GATT Round contained numerous amendments to the antidumping statute. Of particular importance was the broadening of the definition of the ‘less-than-fair-value’ concept to capture not only price discrimination, but also sales below cost. Cost-based allegations now account for between one-half and two-thirds of US AD cases (Clarida, 1996); an even greater share of EU cases is prosecuted using cost-based methodology (Messerlin, 1989).

Given its legal foundation, perhaps it is not surprising that the economic rationale for antidumping statutes is far from clear. A possible rationale is to address predatory pricing practices, where foreign firms are pricing low to induce exit by the domestic firms, allowing monopoly prices in future periods. Economists generally agree that predatory pricing will lead to a welfare loss for a country, but they are sceptical about how often such a strategy is feasible or successful. More importantly, antidumping statutes and practices do not apply the stringent standard used by antitrust (or competition) agencies to determine if pricing is predatory, that is, pricing below marginal cost. Instead, depending on the typical definitions of fair value used by agencies, simple price discrimination across markets or pricing below a level that would return a significant profit to the foreign firm will lead to findings of dumped imports. Such practices are not generally seen as anticompetitive and, in fact, there is often clear tension between antidumping and competition policy. For example, Staiger and Wolak (1992) have shown that domestic

firms can use AD actions to punish foreign firms for refusing to join in collusive actions to raise prices, including the enforcement of price-fixing cartels; examples of price-fixing behaviour in conjunction with AD activity include Ferrovandium and DRAMs. Thus, economists generally believe there is little connection between national welfare considerations and antidumping protection (Stiglitz, 1997).

Instead, most economists find evidence that antidumping activity is motivated by the same political economy considerations that lead to other forms of trade protection. While the studies documenting this vary in what proxies they construct to measure political pressure, all find that such non-statutory factors are significant in ultimate antidumping decisions. These studies include Moore (1992), DeVault (1993) and Hansen and Prusa (1996; 1997). Industries with production facilities in politically important districts fare better. There is also some evidence that financial contributions to politicians by industries seeking antidumping protection improve the chance of an affirmative determination. In a related vein, these studies find that antidumping duties are more likely to be levied against particular trading partners. Blonigen and Bown (2003) argue that this finding does not so much reflect a bias against certain countries, but rather reflects that the inability of certain countries to effectively use the threat of retaliation to deter others from using antidumping against it.

In addition, studies of US antidumping activity have found that changes in legal statutes and agency discretion have led to ever greater dumping margins and the likelihood of determining material injury. For example, Hansen and Prusa (1996) show that the US legal change to allow government agencies to consider the all import sources named in an investigation cumulatively (not individually) makes a material injury decision much more likely. This US legal change was later adopted by WTO antidumping statutes in the Uruguay Round and led to both a dramatic increase in the incidence of multi-country cases and also a sharp increase in affirmative determinations (Hansen and Prusa, 1996; Tharakan, Greenway and Tharakan, 1998; Irwin, 2005). Another example is the documentation by various studies of how the antidumping statutes allow substantial latitude to agencies in how they practically determine dumping margins. Blonigen's (2006) statistical analysis finds that changes in agency discretionary practices is the primary factor behind the rise in average US dumping margins from around 15 per cent in the early 1980s to 60 per cent by 2000.

Direct economic effects of antidumping statutes and remedies

The direct economic result of antidumping remedies is to reduce import flows. Such import declines can happen once an investigation is begun and when antidumping remedies are uncertain. In addition, Staiger and Wolak (1994) emphasize that about half the trade impact occurs before the final determination. They argue that trade impact is sufficiently large for the benefits accruing during the investigation to often exceed the costs of filing the petition. Ethier and Fischer (1987), Fischer (1992), Reitzes (1993), and Prusa (1994) also emphasize the dampening impact on trade created by the threat of AD investigation.

From a welfare perspective, a number of studies have documented that domestic firms can gain from such trade-dampening effects, including Hartigan, Kamma and Perry (1989), Blonigen, Tomlin and Wilson (2004), and Konings and Vandenbussche (2005). However, the latter paper shows that such positive gains are eliminated when foreign firms locate production of the investigated product in country

and, thus, avoid the antidumping duties. Prusa (1997) also documents the substantial trade diversion effects that can take place from investigated import sources to non-investigated sources, which provides another reason why such antidumping remedies may not benefit the domestic industry.

Other studies have used computable equilibrium analysis to examine the total welfare consequences of antidumping remedies for a country. As is typical of trade policy welfare analysis, such losses to consumers are typically estimated to outweigh the gains to the protected producers for antidumping protection. For example, using a computable general equilibrium model, Gallaway, Blonigen and Flynn (1999) estimate that the cumulative effect of all antidumping duties in place leads to an annual four billion dollar welfare loss for the United States. This figure places this form of trade protection as second only to the restrictive and comprehensive quotas on textiles and apparel (Multifiber Arrangement) in terms of welfare costs.

Indirect economic effects of antidumping statutes and remedies

Beyond these typical trade and welfare considerations, economists have pointed to a number of features of antidumping programmes that may cause a greater range of ancillary (or indirect) effects that are often unique to this form of trade protection. In fact, this is where the bulk of recent economic literature has centred its attention, and insights often come from thinking about strategic considerations applying game theoretic techniques.

Such issues are pervasive in analysing the decision to file an antidumping case and its likely chance of success. A foreign industry can almost guarantee it will not be subject to antidumping duties if it charges sufficiently high prices in its export markets. On the other hand, a domestic industry has incentives to look ‘weak’ to make an injury determination more likely, which could lead it to charge higher prices (produce less) than optimal, or lay off more workers than it otherwise would. Ethier and Fischer (1987), Fischer (1992), Reitzes (1993), and Prusa (1994) are examples of applied game theory pieces that document these possible strategic decisions by domestic and foreign firms to influence future antidumping outcomes. Anderson (1992; 1993) examines the potential interdependence of antidumping with another form of trade protection: voluntary export restraints (VERs). The artificial scarcity created by the VERs generates rents for foreign firms that are typically divided up by their market shares. This perversely gives the foreign firms incentives to ‘dump’ their products to garner larger market shares, which makes antidumping investigations and remedies more likely.

The strategic interactions described above are non-cooperative in nature, but a number of papers have examined how antidumping can elicit various forms of cooperative strategic behaviour. These studies primarily provide theoretical analysis, showing how antidumping law can facilitate or sustain collusive cartel pricing by foreign and domestic firms; such studies include Staiger and Wolak (1989), Prusa (1992), and Veugelers and Vandenbussche (1999). Taylor (2004) and Zanardi (2004) provide empirical examinations of collusive behaviour in antidumping activity using US data.

Strategic interactions surrounding antidumping petitions may also occur amongst domestic firms. Cassing and To (2004) show that the decision by a domestic firm to join an antidumping petition can signal its efficiency to other firms in the market. Thus, for example, some domestic firms may not join a petition to signal to others that they have low costs.

Once antidumping remedies are in place, other strategic reactions are possible too. As mentioned above,

a foreign firm can ‘jump’ the antidumping duties and relocate its production to either the domestic market or to a third country that is not subject to the duties. Belderbos (1997) and Blonigen (2002) document significant tariff-jumping of antidumping duties in Europe and the United States. Interestingly, if foreign firms differ in their ability to make such investments, then antidumping might particularly burden firms who cannot make such adjustments. Ironically, this means the foreign firms who are most able to ‘jump’ the AD duty potentially have an incentive to encourage antidumping actions (Blonigen and Ohno, 1998).

The ability of firms to reduce their antidumping duties in subsequent administrative reviews also provides interesting incentives to firms. Such reviews examine recent data to recalculate antidumping duties, which creates a dynamic environment for price setting by the foreign firm. Blonigen and Park (2004) develop a model of dynamic pricing decisions by foreign firms facing the possibility of antidumping duties and subsequent recalculations in future periods. They first show that, if antidumping duties are a certainty when a foreign firm dumps, then the only firms that will dump care very little about the future (high discount rates). Over time the punitive antidumping duties will cause them to dump even more. However, if antidumping remedies are uncertain, foreign firms that have *ex ante* low expectations of antidumping remedies will quickly reduce their dumping once, to their surprise, they become subject to antidumping duties. Blonigen and Park confirm these hypotheses using data on US antidumping investigations. In a related paper, Blonigen and Haynes (2002) find that foreign firms subject to antidumping duties alter their behaviour to fully pass through exchange rate changes and also pass through greater than 100 per cent of the antidumping duty onto the prices in their export market.

Blonigen and Prusa (2003) provide a detailed review of the economics literature on antidumping and also point towards what they consider fruitful areas for future research. These include the treatment of antidumping in competition policy, effects on downstream industries and import/export companies, and comparisons of antidumping statutes across various WTO member countries. The U.S. Antidumping and Countervailing Duty Database and the Global Antidumping Database should play an important role in facilitating future research in antidumping.

See Also

- international trade theory
- tariffs
- trade costs

Bibliography

Anderson, J.E. 1992. Domino dumping I: competitive exporters. *American Economic Review* 82, 65–83.

Anderson, J.E. 1993. Domino dumping II: anti-dumping. *Journal of International Economics* 35, 133–50.

Belderbos, R.A. 1997. Antidumping and tariff jumping: Japanese firms. DFI in the European Union and

the United States. *Weltwirtschaftliches Archiv* 133, 419–57.

Blonigen, B.A. 2002. Tariff-jumping antidumping duties. *Journal of International Economics* 57, 31–50.

Blonigen, B.A. 2006. Evolving discretionary practices of U.S. antidumping activity. *Canadian Journal of Economics*.

Blonigen, B.A. and Ohno, Y. 1998. Endogenous protection, foreign direct investment, and protection-building trade. *Journal of International Economics* 46, 205–27.

Blonigen, B.A. and Haynes, S.E. 2002. Antidumping investigations and the pass-through of exchange rates and antidumping duties. *American Economic Review* 92, 1044–61.

Blonigen, B.A. and Bown, C.P. 2003. Antidumping and retaliation threats. *Journal of International Economics* 60, 249–73.

Blonigen, B.A. and Prusa, T.J. 2003. Antidumping. In *Handbook of International Economics*, ed. E. Kwan Choi and J. Harrigan. Malden, MA: Blackwell.

Blonigen, B.A. and Park, J.-H. 2004. Dynamic pricing in the presence of antidumping policy: theory and evidence. *American Economic Review* 94, 134–54.

Blonigen, B.A., Tomlin, K. and Wilson, W.W. 2004. Tariff-jumping FDI and domestic firms' profits. *Canadian Journal of Economics* 37, 656–77.

Cassing, J. and To, T. 2004. Antidumping, signaling and cheap talk. Working paper, Department of Economics, University of Pittsburgh. Clarida, R.H. 1996. Dumping in theory, in policy, and in practice. In *Fair Trade and Harmonization*, ed. J. Bhagwati and R. Hudec. Cambridge, MA: MIT Press.

DeVault, J.M. 1993. Economics and the international trade commission. *Southern Economic Journal* 60, 463–78.

Ethier, W.J. and Fischer, R.D. 1987. The new protectionism. *Journal of International Economic Integration* 2, 1–11.

Finger, J.M. 1993. *Antidumping: How It Works and Who Gets Hurt*. Ann Arbor, MI: University of Michigan Press.

Fischer, R.D. 1992. Endogenous probability of protection and firm behavior. *Journal of International Economics* 32, 149–63.

Gallaway, M.P., Blonigen, B.A. and Flynn, J.E. 1999. Welfare costs of US antidumping and countervailing duty laws. *Journal of International Economics* 49, 211–44.

Global Antidumping Database. Online. Available at http://people.brandeis.edu/~cbown/global_ad, accessed 12 May 2007.

Hansen, W.L. and Prusa, T.J. 1996. Cumulation and ITC decision making: the sum of the parts is greater than the whole. *Economic Inquiry* 34, 746–69.

Hansen, W.L. and Prusa, T.J. 1997. The economics and politics of trade policy: an empirical analysis of ITC decision making. *Review of International Economics* 5, 230–45.

Hartigan, J.C., Kamma, S. and Perry, P.R. 1989. The injury determination category and the value of relief from dumping. *Review of Economics and Statistics* 71, 183–86.

Irwin, D.A. 2005. The rise of U.S. antidumping activity in historical perspective. *The World Economy* 28, 651–68.

Konings, J. and Vandenbussche, H. 2005. Antidumping protection and markups of domestic firms. *Journal of International Economics* 65, 151–65.

Messerlin, P.A. 1989. The EC antidumping regulations: a first economic appraisal, 1980–85. *Weltwirtschaftliches Archiv* 125, 563–87.

Miranda, J., Torres, R.A. and Ruiz, M. 1998. The international use of antidumping: 1987–1997. *Journal of World Trade* 32, 5–71.

Moore, M.O. 1992. Rules or politics? An empirical analysis of ITC anti-dumping decisions. *Economic Inquiry* 30, 449–66.

Prusa, T.J. 1992. Why are so many antidumping petitions withdrawn? *Journal of International Economics* 33, 1–20.

Prusa, T.J. 1994. Pricing behavior in the presence of antidumping law. *Journal of Economic Integration* 9, 260–89.

Prusa, T.J. 1997. The trade effects of US antidumping actions. In *The Effects of US Trade Protection and Promotion Policies*, ed. R.C. Feenstra. Chicago: University of Chicago Press.

Prusa, T.J. 2001. On the spread and impact of antidumping. *Canadian Journal of Economics* 34, 591–611.

Reitzes, J.D. 1993. Antidumping policy. *International Economic Review* 34, 745–63.

Staiger, R.W. and Wolak, F.A. 1989. Strategic use of antidumping law to enforce tacit international collusion. Working Paper, No. 3016. Cambridge, MA: NBER.

Staiger, R.W. and Wolak, F.A. 1992. The effect of domestic antidumping law in the presence of foreign monopoly. *Journal of International Economics* 32, 265–87.

Staiger, R.W. and Wolak, F.A. 1994. Measuring industry specific protection: antidumping in the United States. *Brookings Papers on Economic Activity: Microeconomics* 1994, 51–118.

Stiglitz, J.E. 1997. Dumping on free trade: the US import trade laws. *Southern Economic Journal* 64, 402–24.

Taylor, C.T. 2004. The economic effects of withdrawn antidumping investigations: is there evidence of collusive settlements? *Journal of International Economics* 62, 295–312.

Tharakan, P.K.M., Greenaway, D. and Tharakan, J. 1998. Cumulation and injury determination of the European community in antidumping cases. *Weltwirtschaftliches Archiv* 134, 320–39.

U.S. Antidumping and Countervailing Duty Database. Online. Available at http://darkwing.uoregon.edu/~bruceb/adpage.html, accessed 12 May 2007.

Veugelers, R. and Vandenbussche, H. 1999. European anti-dumping policy and the profitability of national and international collusion. *European Economic Review* 47, 1–28.

Zanardi, M. 2004. Antidumping law as a collusive device. *Canadian Journal of Economics* 37, 95–122.

How to cite this article

Blonigen, Bruce A. and Thomas J. Prusa. "antidumping." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000253> doi:10.1057/9780230226203.0046

The New Palgrave Dictionary of Economics Online

anti-poverty programmes in the United States

Robert A. Moffitt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic theory suggests that the extent of redistribution should be constrained by its direct and indirect costs, including disincentive effects. The emphasis in the United States has been on programmes that emphasize employment as well as in-kind rather than cash redistribution, and that provide benefits to populations with special needs. Research on their effects has shown them to decrease poverty rates and the poverty gap but to have labour-supply disincentives as well. Reforms to the main cash programme in the 1990s have increased earnings and employment.

Keywords

Aid to Families with Dependent Children (AFDC) (USA); altruism; anti-poverty programmes in the United States; child care subsidies; crowding out; Earned Income Tax Credit (USA); food stamps; free-rider problem; Head Start (USA); in-kind transfers; Job Corps (USA); labour supply; low-income housing policy; marginal utility of consumption; means-tested transfers; Medicaid (USA); negative income tax; poverty gap; Supplemental Security Income (SSI) (USA); Temporary Assistance for Needy Families (TANF) (USA); Working Families Tax Credit (UK)

Article

Anti-poverty programmes in the United States have received much attention from the economics profession since the 1970s. Economists have studied their effectiveness in reducing poverty and increasing well-being among the poor, their rationale and goals, and trends in their caseloads and expenditures. Scholars have also extensively studied the effects of anti-poverty programmes on a wide range of individual and family behaviours.

Rationale and design issues

Anti-poverty programmes are generally considered to arise from altruism on the part of non-poor voters, who wish to transfer resources, for charitable reasons, to those who have low incomes or assets. Such charitable support is generally considered to be suboptimally provided if left to the private sector because of the free-riding problem that arises when one individual's contribution to the poor makes other givers better off, so individuals have an incentive to let others contribute rather than contribute themselves.

However, the exact nature of the preferences of the non-poor – let us call them voters, since the United States is a democracy – are not well understood. In the classic utilitarian model, the social welfare function equals the sum of individual utilities and the marginal utility of income is assumed to decline with income, so that a dollar redistributed from a high-income person to a low-income person raises social utility. One issue with this framework is whether the ‘weights’ that the voters assign to the poor are the same as marginal utility of income weights, and today most analysts assume those weights to deviate in an arbitrary way and to simply reflect voter preferences that will vary from group to group and from country to country. Another important distinction is whether the voters desire to increase the utility of the poor per se, as the utilitarian model implies, or to increase their consumption of specific goods like food, housing, and medical care. Redistributing in the latter fashion, resulting in what are termed ‘in-kind’ transfers, is quite common in practice, and economists have often assumed that it implies that voters are paternalistic in the sense that they wish to override the spending preferences of the poor themselves. Redistributing purely in the form of income, for example, would allow recipients to allocate the transfer in a way that maximizes their utility as they see it. Another rationale for in-kind transfers is that they induce only those with the highest marginal utility of consumption of those goods to accept such transfers, which induces a desirable (from the voter's point of view) selection from the low-income population to those who need it most (Nichols and Zeckhauser, 1982; Blackorby and Donaldson, 1988), and yet another is that they reduce the incentive of the recipient to alter behaviour to increase later transfers (Bruce and Waldman, 1991).

Whatever the preferences of the voters, the main issue in models of optimal provision of anti-poverty benefits to the poor is the trade-off between the benefits of redistribution and the direct and indirect costs of the transfer. The direct costs arise because taxation has its own resource cost and the indirect costs arise because the transfer distorts the behaviour of the recipients. As in the classic models of taxation, lump-sum transfers are not possible and so transfers alter the prices of various goods in the utility function. In the well-known Mirrlees (1971) model, the main margin examined is work effort, which is reduced by transfers, and optimal redistribution proceeds up to the point where the marginal benefits of additional redistribution are counterbalanced by the marginal losses arising from reductions in work effort. However, one of the main areas of research on anti-poverty programmes, particularly those that are empirical in nature, has been on other possible margins of adjustment by programme recipients. Transfers may reduce incentives to invest in human capital, reduce incentives to save if assets are taxed by the programme, increase incentives to have additional children if benefits are tied to family size, change incentives to marry if marital status affects benefits, or increase incentives to migrate from one jurisdiction to another to obtain higher benefits if benefits vary within a country. For in-kind programmes, there is also potential ‘leakage’ in the consumption effects. For example, giving a family either a lump-sum amount of food or a subsidy to the price of food may lead them to reduce their own expenditures on food in order to spend more on other consumption items.

The prototype of a transfer programme that aims to balance redistribution and disincentives is the

negative income tax (NIT) (Watts, 1987). In an NIT, recipients who have no income receive a maximal benefit but the size of the transfer declines as income rises. Thus those with lower incomes receive greater benefits than those with higher incomes, as most models imply should occur, but the rate at which benefits are reduced as income rises is generally taken to be less than 100 per cent. This provides some incentive to work, and work disincentives are therefore controlled by the rate of benefit reduction. A transfer system with a 100 per cent reduction rate, particularly one that extends relatively high into the income distribution, is said to create a ‘poverty trap’ because individuals cannot escape poverty through modest increases in income. The first formal demonstration of the optimality of an NIT was provided again by Mirrlees (1971), who showed that such a programme results from an optimal utilitarian model. This general paradigm applies to the other margins mentioned above as well, for in each case a programme can be designed to provide the highest benefits to those with the lowest resources while paying attention to the effect of the programme on the price of changing behaviour (undertaking human capital investment, saving, and so on). An important modification of the Mirrlees models appears in Diamond (1980) and Saez (2002), who showed that consideration of the ‘extensive’ margin of work – namely, the decision to work at all rather than the decision of how many hours to work, which was the focus in the Mirrlees model – can lead to earnings subsidies, where the marginal ‘tax rate’ on earnings at the bottom of the income distribution is negative rather than positive for some range. The Earned Income Tax Credit in the United States and the Working Families Tax Credit in the United Kingdom are important examples of such earnings subsidies.

Finally, a benefit-provision issue that economists have studied is the relative merits of redistribution by a central government versus local governments within a country. For many years it was assumed that the utility of the poor in all jurisdictions should affect the utility of voters in all jurisdictions equally, which leads to a central government programme. But Pauly (1973) and others have argued that local voters care more about the poor in their own jurisdictions, making redistribution partly a local public good, although they may care to some extent about the poor in other jurisdictions as well. This leads to a mixed central–local system in which the central government subsidizes local governments because of the limited interest of all voters but allows localities to spend on redistribution out of their own resources as well. This leads to subsidy mechanisms such as block grants, matching grant programmes and related funding mechanisms. This structure is found in the United States but also in some European countries.

Anti-poverty programmes

There are a large number of anti-poverty programmes in the United States whose structure and expenditure have changed over time (Moffitt, 2003). We shall ignore Social Security, which has a major impact on poverty rates of the elderly but which is generally considered to be a social insurance programme rather than a means-tested transfer programme. The most well-known and heavily studied programme, and that which historically most resembled an NIT, is the Temporary Assistance for Needy Families (TANF) programme, which was called the Aid to Families with Dependent Children (AFDC) programme prior to 1996. The TANF programme provides monthly cash benefits to families with low income and assets, but primarily to those headed by a single parent (mostly single mothers). The benefit-reduction rate in the programme varies across states but is most often around 50 per cent. However, the TANF programme also has some non-NIT features – specifically, it has work requirements that mandate

that most able-bodied parents work at least some minimum number of hours per week as a condition of receiving benefits, and time limits, which stipulate that parents can receive benefits for only a limited number of years over their lifetimes. These latter provisions were enacted in 1996.

While the AFDC programme was one of the leading US anti-poverty programmes in the 1960s and 1970s, when its caseloads and expenditures were among the largest of US programmes, in 2007 it ranked only sixth in terms of expenditure and fifth in terms of caseload (Moffitt, 2007). It is smaller than the Medicaid programme, which provides medical subsidies to the poor; the Supplemental Security Income (SSI) programme, which provides benefits to poor families with aged adults and disabled adults and children; the Earned Income Tax Credit (EITC), which provides tax credits to working families; Food Stamps, which provides food subsidies to the poor; and housing programmes for the poor. Per capita expenditures on AFDC–TANF have steadily declined since the late 1970s, whereas those on the other programmes have grown by amounts much greater in magnitude. In 2007, total real per capita expenditures in the largest means-tested transfer programmes in the United States had more than quadrupled since 1968 and had grown by 60 per cent just since 1990 as a result of the growth in many of these programmes.

The Medicaid programme, the largest programme in the United States, is a diverse programme covering several different populations. The four primary groups served are low-income single mothers and their children; the low-income elderly; the low-income disabled; and individuals in nursing homes or long-term care with low income and assets. Expenditures and caseloads in the programme grew rapidly in the late 1980s and early 1990s as a result of expansions of eligibility for low-income mothers and children and growth of disabled recipients, and have continued to grow secularly because of growth in the demands for long-term care of the elderly. The United States does not have national health insurance and the size and growth of the Medicaid programme partially reflects that fact. With a few exceptions in certain parts of the programme, there is no benefit-reduction rate in the programme; either the full package of benefits is provided or none at all.

The SSI programme pays cash benefits to low-income individuals who are blind or disabled, and to the low-income elderly. The programme also saw very rapid growth in the early 1990s as a result of increases in disabled, child, and non-citizen recipients. The definition of disability for adults is quite stringent; 60 per cent of applications are denied. The disability definition for children is more elastic and has fluctuated in stringency over time. The programme has a nominal 50 per cent benefit-reduction rate.

The EITC also grew rapidly in the late 1980s and early 1990s, while the Food Stamp programme grew most rapidly after its introduction in the late 1960s and early 1970s, but also most recently (since 2000). The EITC has a subsidy rate of up to 40 per cent and a maximum clawback rate of 21 per cent, while the Food Stamp programme has a nominal 30 per cent benefit-reduction rate.

Other important programmes include those covering housing, child care and training programmes. Housing programmes, which have a typical benefit-reduction rate of 30 per cent, grew most rapidly in the late 1970s and early 1980s, and have seen only modest growth since that time. Child care subsidies in the United States are spread over several different programmes serving overlapping populations, including the welfare poor but also the ‘working’ poor. Expenditures have grown modestly since 2000 as the need for employment support has become increasingly recognized. Included in the child care framework is the Head Start programme, whose goal is to assist child development in pre-school children of low-income families but which also serves a child care function. The United States spends relatively little on training programmes, and has changed the name and nature of its programme for

adults several times since the 1970s in an attempt to make the programmes more effective. Perhaps the most popular programme is the Job Corps, a high-cost residential-based programme for disadvantaged young men and women.

Several patterns can be discerned in the US transfer programme system. First, in-kind transfers are preferred to cash transfers. The only programme that is a pure cash transfer programme is the AFDC–TANF programme, which has declined in importance because of its unpopularity and is now coupled with work requirements in any case. The most popular programmes are those that subsidize medical and food expenditures; those which subsidize housing and child care expenditures are large as well. Second, subsidies that serve specialized populations with specific identifiable needs are preferred to subsidies based on low income per se. The SSI programme, which is cash in nature, is the best example of this preference. However, even the EITC could be argued to fit this category, for it provides cash but only to a specific population viewed as meritorious, namely, low-wage workers. Third, an increasing emphasis on employment is apparent. The EITC reflects this emphasis as do the recent reforms in the AFDC–TANF programme and increases in child care subsidies. Fourth, US voters dislike providing subsidies to low-income single-mother families, who are viewed unfavourably because of US views towards marriage. All four of these features are in explicit conflict with the original idea of an NIT as espoused by Milton Friedman, Robert Lampman, James Tobin, and others, who saw the ideal transfer programme as one that provided only cash benefits, on the basis of income only, and without preference for family structure or type.

Research findings on the effects of US anti-poverty programmes

One overriding issue of interest in research on US anti-poverty programmes is whether such programmes have, in fact, reduced poverty. The evidence indicates that they have (Scholz and Levine, 2001). In 1997, the system of means-tested transfer programmes in the United States reduced the poverty rate of families from 29 per cent to 26 per cent, a modest amount. However, the programmes also raised the incomes of many poor families even if not by enough to cross the poverty line, for the programmes filled in 27 per cent of the poverty gap (defined as the total dollar gap between the poverty line and the incomes of poor families). The most important programme in reducing poverty was Medicaid; SSI and the EITC were also important. It is often noted that these estimates should be considered to be an upper bound for the true effect of transfer programmes on poverty because the work disincentives of the programmes themselves cause a reduction in income, which widens the poverty gap and increases the poverty rate to some offsetting extent.

In addition to this issue, there has been a very large amount of research on the behavioural effects of US anti-poverty programmes. By far the most research has been conducted on the AFDC–TANF programme, where the primary focus prior to 1996 was on its effects on labour supply, marriage and fertility, and a few other behaviours (Moffitt, 1992). Most research on labour supply indicated, as economic theory would predict, negative effects of the programme as a whole. However, the effects of reducing the benefit-reduction rate have been shown to be mostly zero or negligible, with the general interpretation being that such changes bring in new recipients who experience labour-supply reductions that offset the labour supply increases of those initially on the programme. Research on marriage and fertility effects of AFDC has shown mostly small but non-zero effects in reducing marriage and

increasing childbearing. Research conducted on the effects of the 1996 reform of the programme (Blank, 2002; Moffitt, 2003; Grogger and Karoly, 2005) has shown the reform, whose major elements were work requirements and time limits, to have had positive effects on average employment, earnings, and family income and negative effects on welfare usage. However, some research also suggests that there is a group of very disadvantaged families who were made worse off by the reform. The research also has shown the reform to have had little if any effect on marriage and fertility behaviour and to have had modest effects, if any, on children in low-income families.

There has been a fair amount of research on other programmes as well. The Medicaid programme appears to have modest negative effects on labour supply and expansions in the programme have led to ‘crowdout’ of private health insurance, but the programme has also been shown to have had many favourable effects on health, particularly that of children (Gruber, 2003). Research on the SSI programme has focused particularly on reasons for fluctuations in the size of the caseload, but has also concerned work incentives, where both benefit-reduction rates and other employment-incentive programmes have been shown to have had little effect (Daly and Burkhauser, 2003). Research on child care programmes have shown them to have had positive effects on female employment, and Head Start has been shown to have some positive effects on child outcomes, but which fade out over time (Blau, 2003). Work on training programmes has shown them to have different effectiveness for different groups, with several low-cost programmes found to be effective in increasing earnings for single mothers and with the high-cost Job Corps programme found to be effective for disadvantaged youth, but with no type of programme having been found to have a significantly positive rate of return for adult men (Lalonde, 2003).

See Also

- nutrition and public policy in advanced economies
- poverty alleviation programmes
- taxation and poverty
- welfare state

Bibliography

Blackorby, C. and Donaldson, D. 1988. Cash versus kind, self-selection, and efficient transfers. *American Economic Review* 78, 691–700.

Blank, R. 2002. Evaluating welfare reform in the United States. *Journal of Economic Literature* 40, 1105–66.

Blau, D. 2003. Child care subsidy programs. In Moffitt (2003).

Bruce, N. and Waldman, M. 1991. Transfers in kind: why they can be efficient and nonpaternalistic. *American Economic Review* 81, 1345–51.

Daly, M. and Burkhauser, R. 2003. The supplemental security income program. In Moffitt. (2003).

Diamond, P. 1980. Income taxation with fixed hours of work. *Journal of Public Economics* 13, 101–10.

Grogger, J. and Karoly, L. 2005. *Welfare Reform: Effects of a Decade of Change*. Cambridge, MA: Harvard University Press.

Gruber, J.D. 2003. Medicaid. In Moffitt (2003).

Lalonde, R. 2003. Employment and training programs. In Moffitt (2003).

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Moffitt, R. 1992. Incentive effects of the U.S. welfare system: a review. *Journal of Economic Literature* 30, 1–61.

Moffitt, R. (ed.) 2003. *Means-Tested Transfer Programs in the United States*. Chicago: University of Chicago Press.

Moffitt, R. 2007. Four decades of anti-poverty policy: past developments and future directions. *Focus*. Institute for Research on Poverty, University of Wisconsin (forthcoming).

Nichols, A. and Zeckhauser, R. 1982. Targeting transfers through restrictions on recipients. *American Economic Review* 78, 240–4.

Pauly, M. 1973. Income redistribution as a local public good. *Journal of Public Economics* 2, 35–58.

Saez, E. 2002. Optimal income transfer programs: intensive versus extensive labor supply responses. *Quarterly Journal of Economics* 117, 1039–73.

Scholz, J.K. and Levine, K. 2001. The evolution of income support policy in recent decades. In *Understanding Poverty*, ed. S. Danziger and R. Haveman. New York: Russell Sage.

Watts, H. 1987. Negative income tax. In *The New Palgave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

How to cite this article

Moffitt, Robert A. "anti-poverty programmes in the United States." The New Palgrave Dictionary of

Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000250>
doi:10.1057/9780230226203.0047

The New Palgrave Dictionary of Economics Online

antitrust enforcement

Joseph E. Harrington, Jr.

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article explores the enforcement of those laws intended to promote competitive markets through the prohibition of certain practices such as price-fixing, welfare-reducing mergers, and monopolization. The discovery and prosecution of violations are examined, including the role of leniency programmes. The determination of penalties is investigated with an assessment of their relationship to optimal penalties. Enforcement policy is found to vary over time and its determinants are reviewed. Finally, the efficacy of enforcement is assessed.

Keywords

antitrust enforcement; antitrust penalties; cartels; collusion; Corporate Leniency Program; price fixing

Article

Antitrust enforcement is the process whereby a more competitive environment is created through the prohibition of certain practices deemed illegal by antitrust laws.
Restraints of trade such as price-fixing and bid-rigging are prohibited in the United States under section 1 of the Sherman Act of 1890 and in the European Union under article 81 of the Treaty of the European Communities of 1999. Practices designed to create monopolies (such as predatory pricing and tying) are prohibited in the United States under section 2 and in the European Union under article 82. Mergers that are harmful to competition are prohibited in the United States under section 7 of the Clayton Act of 1914 and in the European Union under article 2(3) of the Merger Regulation. Although this article adopts a US focus, much of what is described is applicable to many OECD countries. (For a more general treatment on antitrust policy, see Motta, 2004, for the European Union and Viscusi, Harrington and Vernon, 2005, for the United States.)

Detection of antitrust offences

Enforcement can involve three stages: (*a*) discovery and evaluation of a possible antitrust violation; (*b*) prosecution when it is deemed there is a violation; and (*c*) levying of penalties and enacting of remedies when prosecution is successful. Antitrust cases can arise in a variety of ways. With a recent exception noted below, cartels are generally discovered not by the antitrust authorities but rather by customers, employees, and even competitors. Though not yet widely used, economic and econometric methods for detecting collusion include determining whether: (*a*) firm behaviour is inconsistent with competition; (*b*) there is a structural break in behaviour; (*c*) the behaviour of suspected colluding firms differs from that of some benchmark competitive firms; and (*d*) a collusive model fits the data better than a competitive model (Harrington, 2006). In contrast, prospective merger cases are brought by the participants themselves to the antitrust authorities, as mandated by the Hart–Scott–Rodino Act of 1976. In evaluating a proposed merger, the primary considerations are the extent to which it would raise price and whether there are offsetting cost savings.

Antitrust penalties

In the case of price-fixing, the government levies fines at the corporate level which, as a result of the Sentencing Reform Act of 1984, can be as high as twice the gross pecuniary gain of the defendant or twice the pecuniary loss of the victims (though a Supreme Court decision in 2005 has since put these guidelines into jeopardy). The most significant financial penalty comes from private damages which, due to the Clayton Act, allow direct buyers to receive compensation equal to three times the damages. At the individual level, the government imposes fines and prison sentences; since 1970, 53 per cent of convicted individuals have been imprisoned (Gallo et al., 2000). The use of government fines is common in many other countries, although prison sentences and civil damages are unique to the United States and Canada.

Are these penalties optimal? An optimal penalty is one that deters only those activities that are welfare-reducing. If the gain to the offenders is *g*, the loss to other agents is *l*, the probability of being penalized is *p*, and the penalty is *f* then optimality requires: $g - pf \geq 0$ if and only if $g \geq l$ (Polinsky and Shavell, 2000). Therefore, the optimal penalty is $f = l/p$. In practice, private damages are calculated as $(P^c - P^{bf})Q^c$ where P^c is the observed (collusive) price, Q^c is the number of units sold, and P^{bf} is the 'but for' price, that is, the price that would have been charged but for collusion. $P^c - P^{bf}$ is referred to as the 'overcharge'. A major source of contention in many price-fixing cases is the determination of P^{bf}, for which reduced-form estimation methods are largely deployed with the use of data encompassing both the cartel and non-cartel regimes. The 'before and after' approach is quite common and entails estimating: $P(t) = \delta + \beta X(t) + \gamma v(t) + \varepsilon(t)$ where $P(t)$ is price, $X(t)$ is a vector of demand and cost shifters, and $\nu(t)$ is a dummy variable that equals one in those periods that firms were colluding (Page, 1996). If $\hat{\delta}$ and $\hat{\beta}$ are the parameter estimates, then $P^{bf}(t) = \hat{\delta} + \hat{\beta}X(t)$. Since damages, as calculated in practice, ignore deadweight loss, penalties are neither optimally punitive nor compensatory: $g < (P^c - P^{bf})Q^c < l$. Government fines also suffer from this deficiency as they tend to be proportional to sales, P^cQ^c.

Of course, if collusion serves only to reduce supply, then $l > g$ and thus we should prevent all collusion,

in which case $f \geq g/p$ is desired. As cartels continue to form, penalties clearly fall short. But how far away are they from being an effective deterrent? In practice, cases are largely settled out of court and single (not treble) damages are typical (Lande, 1993). For international cartels over 1990–2003, Connor (2004) calculates private and public recovery in the United States was only 115 per cent of damages. Bryant and Eckard (1991) infer from observed cartel lengths that the chances of a price-fixing cartel being indicted in a 12-month period is 11–15 per cent. Though that estimate relies on a properly specified functional form for the distribution on cartel lifetimes, it is safe to say that the probability of a cartel being discovered and paying penalties is well below one, so that financial penalties are woefully inadequate. What may be more effective is the use of prison sentences (Werden and Simon, 1987).
Although remedies have been used in price-fixing cases (for example, a ten-year consent decree in 1994 placed restrictions on announcements of future price changes by airlines), they are typically more important in merger and monopolization cases. Some proposed mergers receive government approval only after restructuring, such as the selling of assets that, if retained by the newly merged firm, would significantly harm competition. In rare cases, the authorities seek to prevent the merger entirely. In the case of monopolization, remedies may be either behavioural or structural. Behavioural remedies could, for example, require a firm to license intellectual property to competitors (as with Xerox) or prohibit certain contractual arrangements (as with Microsoft). Structural remedies are typically quite draconian and accordingly rare. Notable examples include the break-up of Standard Oil in 1911 and AT&T in 1984. A lower court initially ordered Microsoft to be broken into two companies – one with the operating system and the other with applications – though it was later remanded by the US Court of Appeals, and the Department of Justice (DOJ) stopped pursuing it as a remedy.

Corporate leniency program

One of the most significant innovations in antitrust enforcement in recent years is the 1993 revision of the DOJ's Corporate Leniency Program and the institution of a similar programme by the European Commission in 1996. The first member of a cartel to come forward and cooperate receives full amnesty with respect to government penalties and liability for only single damages. As a condition of entering the programme, company representatives must answer an ‘omnibus question’ which asks them whether they know of any collusion in other markets. Failure to truthfully answer that question results in the loss of all amnesty. This policy has proven useful for both the discovery and the prosecution of cartels.
Under the standard repeated game framework, a leniency programme affects the stability of collusion through the usual equilibrium condition: the expected payoff from continuing to collude must be at least as great as the payoff to a firm from (optimally) cheating on the cartel. (The discussion here is based on Harrington, 2005; see also Motta and Polo, 2003, and Spagnolo, 2003.) More leniency enhances the payoff to cheating because a firm that does so can simultaneously apply for amnesty and thereby reduce expected penalties. However, leniency also affects the expected collusive payoff because firms anticipate the possibility of using the programme in the future. More leniency lowers penalties in the event that leniency is received and thus can raise the payoff from continuing to collude. But it is also possible that waiving a higher fraction of penalties *increases* future expected penalties. The reason is that there can be two equilibria: one in which all firms apply for amnesty and one in which none does. The latter can Pareto-dominate because only one firm can receive amnesty and use of the programme

results in certain conviction. More leniency can destabilize the Pareto-preferred equilibrium in which all firms refrain from using the programme because it becomes too attractive for a firm to apply (given that other firms do not). Although there are then several countervailing forces, it is generally optimal to provide some leniency, and conditions are not too restrictive for it to be optimal to waive all penalties.

Intensity of antitrust enforcement

An enforcement policy is described not just by the types of cases pursued but also by its intensity. One might expect the socially optimal level of enforcement to vary with economic activity as, for example, there are more merger notifications during booms and possibly more cartels during periods of weak demand. Furthermore, government preferences regarding the level and focus of enforcement may vary with the incumbent presidential administration.

The budgets of the DOJ and the Federal Trade Commission are indeed increasing in GDP (Kwoka, 1999) but antitrust case activity is counter-cyclical (Ghosal and Gallo, 2001). Although most studies do not find case activity to be related to the administration's political party, Ghosal (2004) shows that this is due to aggregation and mis-specification. He disaggregated data for 1958–2002 into criminal and civil cases and allowed there to be a structural break in the relationship between the usual independent variables – such as GDP, the DOJ's budget, and the president's political party – and the number of DOJ cases. Reasons for a break comprise the growing influence among economists and judges of the Chicago School – which argued that a number of previously considered antitrust offences may be profitable for firms to pursue for competitive reasons – and the fact that the Supreme Court had a two-thirds majority of Republican-nominated justices starting in 1972. Both of these forces would give less credence to certain practices – such as vertical restraints and monopolization practices – being treated as antitrust violations. A break in the number of civil cases (such as mergers and vertical restraints) occurred around the mid-1970s, which resulted in a significant decline, while a significant rise in the number of criminal cases (collusion) occurred around the late 1970s. There is also a post-regime rise in polarization between Republican and Democratic presidential administrations with Republicans pursuing more (less) criminal (civil) cases.

Impact of antitrust enforcement

Is enforcement having an effect? This is a difficult question for which hard facts are lacking, and sharply divergent views have been expressed. (See Baker, 2003, and Crandall and Winston, 2003; the latter should be read with caution as their review of some literatures is seriously deficient – Kwoka, 2003, and Werden, 2004, provide a critique.) With respect to the most egregious offence – namely, collusion – we pose three questions. Do cartels actually charge higher prices? Does prosecution lower prices? And, does successful prosecution have a deterrent effect?

The evidence is overwhelming that cartels raise prices. Connor and Lande (2005) have provided an exhaustive survey and found the median overcharge is 25 per cent. The evidence on how prices respond after indictment and conviction is mixed. A price decline was found in the break-up of cartels in white pan bread (Block, Nold and Sidak, 1981); and Feinberg (1984) found that, for four of five cartels, the Producer Price Index for the cartelized market fell by 6.6–11.4 per cent relative to a broader industry

price index. Evidence to the contrary is provided in Sproul (1993) where, for 25 price-fixing cases over 1973–84, price (measured relative to that of a related good) *rose* by seven per cent in the four-year period after the indictment, although in some cases the immediate response was a nine to ten per cent fall in price. In light of the well-established evidence of an overcharge, the natural interpretation is that, although prosecution may reduce prices in the short run, in the longer run collusion may re-establish itself either explicitly or tacitly.

Even if prices do rebound from a conviction, prosecution and penalties are still useful because they reduce the profitability of collusion and thus may deter some cartels from forming. Indeed, there is some evidence of deterrence. The general method of testing for it is to have a reduced form equation explaining markups over time and to include a dummy variable when an action has been filed for collusion in a related market. In the case of white pan bread, markups fell for cities in a region for which the DOJ had filed an action that year in some other city in that region (Block, Nold and Sidak, 1981). Similar evidence of deterrence holds for highway construction procurement auctions, which are notorious for bid-rigging (Block and Feinstein, 1986).

In sum, the evidence is that cartels exist, they substantially raise price, and the indictment and conviction of firms may result in lower prices and may have a deterrent effect. Finally, financial penalties fall significantly short of making collusion unprofitable.

See Also

- cartels
- merger analysis (United States)
- merger simulations

I appreciate the comments of Vivek Ghosal.

Bibliography

Baker, J. 2003. The case for antitrust enforcement. *Journal of Economic Perspectives* 17(4), 27–50.

Block, M. and Feinstein, J. 1986. The spillover effect of antitrust enforcement. *Review of Economics and Statistics* 68, 122–31.

Block, M., Nold, F. and Sidak, J. 1981. The deterrent effect of antitrust enforcement. *Journal of Political Economy* 89, 429–45.

Bryant, P. and Eckard, E. 1991. Price fixing: the probability of getting caught. *Review of Economics and Statistics* 73, 531–6.

Connor, J. 2004. Effectiveness of antitrust sanctions on modern international cartels. Working paper, Purdue University.

Connor, J. and Lande, R. 2005. How high do cartels raise prices? Implications for optimal cartel fines. *Tulane Law Review* 80, 513–70.

Crandall, R. and Winston, C. 2003. Does antitrust policy improve consumer welfare? Assessing the evidence. *Journal of Economic Perspectives* 17(4), 3–26.

Feinberg, R.M. 1984. Strategic and deterrent pricing responses to antitrust investigations. *International Journal of Industrial Organization* 2, 75–84.

Gallo, J., Dau–Schmidt, K., Craycraft, J. and Parker, C. 2000. Department of Justice antitrust enforcement, 1955–1997: An empirical study. *Review of Industrial Organization* 17, 75–133.

Ghosal, V. 2004. Regime shifts in antitrust. Working paper, Georgia Institute of Technology.

Ghosal, V. and Gallo, J. 2001. The cyclical behavior of the Department of Justice's antitrust enforcement activity. *International Journal of Industrial Organization* 19, 27–54.

Harrington, J. Jr. 2005. Optimal corporate leniency programs. Working paper, Johns Hopkins University.

Harrington, J. Jr. 2006. Detecting cartels. In *Handbook of Antitrust Economics*, ed. P. Buccirossi. Cambridge, MA: MIT Press.

Kwoka, J. 1999. Commitment to competition: an assessment of antitrust agency budgets since 1970. *Review of Industrial Organization* 14, 295–302.

Kwoka, J. 2003. The attack on antitrust policy and consumer welfare: a response to Crandall and Winston. Working Paper No. 03-008. Northeastern University.

Lande, R. 1993. Are antitrust ‘treble’ damages really single damages? *Ohio State Law Journal* 54, 115–74.

Motta, M. 2004. *Competition Policy: Theory and Practice*. Cambridge: Cambridge University Press.

Motta, M. and Polo, M. 2003. Leniency programs and cartel prosecution. *International Journal of Industrial Organization* 21, 347–79.

Page, W., ed. 1996. *Proving Antitrust Damages*. Chicago: American Bar Association.

Polinsky, A. and Shavell, S. 2000. The economic theory of public enforcement of law. *Journal of Economic Literature* 38, 45–76.

Spagnolo, G. 2003. *Divide et Impera*: optimal deterrence mechanisms against cartels and organized crime. Working paper, University of Mannheim.

Sproul, M. 1993. Antitrust and prices. *Journal of Political Economy* 101, 741–54.

Viscusi, W., Harrington, J. Jr. and Vernon, J. 2005. *Economics of Regulation and Antitrust*, 4th edn. Cambridge, MA: MIT Press.

Werden, G. 2004. Comment. *Journal of Economic Perspectives* 18(3), 224–5.

Werden, G. and Simon, M. 1987. Why price fixers should go to prison. *Antitrust Bulletin* 32, 917–37.

How to cite this article

Harrington, Joseph E., Jr. "antitrust enforcement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000223> doi:10.1057/9780230226203.0048

The New Palgrave Dictionary of Economics Online

Antonelli, Giovanni Battista (1858 – 1944)

A.P. Kirman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Antonelli, G. B.; Engel curves; indirect demand function; integrability problem

Article

Antonelli was born near Pisa in 1858. He studied mathematics and then went on to qualify as an engineer. Although his life was devoted to civil engineering, he made an important contribution to early mathematical economics. His *Sulla teoria matematica dell'economia politica* (1886), intended to be the first part of a book, is remarkable, in particular for the conditions he gives for the 'integrability problem'. This asks under what conditions single valued demand functions are generated by the maximization of a utility function. Antonelli studied the 'local' aspects of this problem. He started from what is now called the indirect demand function:

$$p = M[q]$$

where q is the vector of goods and p the vector of prices. He gave the symmetry of the matrix of the price substitution terms $\partial p_i / \partial q_j$ as a condition for the recoverability of the utility function but should have also required the negative semi-definiteness of this matrix. The importance of this work has been recognized by Samuelson (1950) and later authors, but passed unappreciated if not unnoticed at the time. In the same work Antonelli derives a condition for a market demand function to be derivable from a market utility function, that is, that individuals have linear parallel Engel curves. This condition was found much later by Gorman (1953) and Eisenberg (1961). Antonelli had an active and productive career in engineering and what would now be called 'operations research' but never came back to theoretical economics. He died in 1944.

Selected works

1886. *Sulla teoria matematica dell'economia politica*. Pisa. Reprinted, with an introduction by G. Demaria, Milan: Malfasi, 1952.

Bibliography

A detailed and careful account of Antonelli's contributions and a translation of his economic paper is given in:

Chipman, J.S., Hurwicz, L., Richter, M.K. and Sonnenschein, H.F., eds. 1971. *Preferences, Utility and Demand*. New York: Harcourt Brace and Jovanovich. (See in particular Introduction to Part II by J.S. Chipman, chapter 16: a translation of Antonelli and chapter 9 on the integrability problem by L. Hurwicz.)

Eisenberg, E. 1961. Aggregation of utility functions. *Management Science* 7, 337–50.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21, 63–80.

Samuelson, P.A. 1950. The problem of integrability in utility theory. *Economica* 17, 355–85.

How to cite this article

Kirman, A. P. "Antonelli, Giovanni Battista (1858–1944)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000114> doi:10.1057/9780230226203.0049

The New Palgrave Dictionary of Economics Online

approximate solutions to dynamic models (linear methods)

Harald Uhlig
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article explains how to obtain an approximate solution to dynamic stochastic discrete-time (DSGE) models by first log-linearizing the relevant equations and then obtaining a recursive law of motion, by using the method of undetermined coefficients. Calculations are provided based on both an eigenvector decomposition and the QZ or Schur decomposition. The role of sunspots and the relationship to the method of Blanchard and Kahn are discussed. The base example is a generic real business cycle model, for which log-linearization is described generally and in detail. The method described should be easily implemented. Further literature references and software sources are provided.

Keywords

approximate solutions to dynamic models (linear methods); dynamic stochastic general equilibrium (DSGE) models; linearization; log-linearization; QZ decomposition; real business cycles; representative agent; recursive law of motion; sunspots

Article

Linear methods are often used to compute approximate solutions to dynamic models, as these models often cannot be solved analytically. While a plethora of advanced numerical methods exist, the most popular ‘bread-and-butter’ method for solving them is linearization. It is described here first with the example of a simple real business cycle model, but is applicable generally to dynamic stochastic general equilibrium (DSGE) models. It is shown how to easily generate the log-linearized equations needed. The linear system is then solved for the recursive law of motion, by using the method of undetermined coefficients. The classic reference for solving linear difference models under rational expectations is Blanchard and Kahn (1980), while Kydland and Prescott (1982) is the origin of the modern approach of calculating numerically approximate solutions to dynamic stochastic models in order to obtain quantitative results. Much of the material here is taken from Uhlig (1999), which builds on the method of undetermined coefficients in King, Plosser and Rebelo (2002).

A basic example

As a basic example, consider a version of the real business cycle model of Hansen (1985). A social planner or representative agent chooses c_t, k_t, y_t, l_t and n_t to maximize the utility function $U = E[\sum_{t=0}^{\infty} \beta^t u(c_t, l_t)]$ for some twice differentiable utility function u(.), satisfying the usual conditions, subject to the constraints

$$c_t + k_t = y_t + (1-\delta)k_{t-1} y_t = \gamma_t f(k_{t-1}, n_t) 1 = n_t + l_t$$

as well as a given initial capital stock k_{-1},where c_t denotes consumption, k_t denotes capital, y_t denotes output, l_t denotes leisure, n_t denotes labour, $f(k,n)$ denotes a twice differentiable production function, typically assumed to obey constant returns to scale, β is the discount factor and γ_t is total factor productivity, with $z_t = \log(\gamma_t) - \log(\gamma^*)$ evolving according to $z_t = \rho z_{t-1} + \varepsilon_t$ where $E_t[\varepsilon_{t+1}] = 0$ for some values γ^* and ρ, with $-1 < \rho < 1$. A solution is a stochastic sequence $(c_t, k_t, y_t, l_t, n_t)$, $t \geq 0$ where all variables dated t are independent of all ε_s for $s > t$ and satisfies all constraints, and which maximizes the utility function given above within the set of all such sequences.
The necessary first-order conditions for this problem are given by

$$u_c(c_t, l_t) = \lambda_t u_l(c_t, l_t) = f_n(k_{t-1}, n_t)\lambda_t = \beta E_t[\lambda_{t+1} R_{t+1}] R_t = f_k(k_{t-1}, n_t) + 1 - \delta$$

Linearization

The first step towards solving the model by linear approximation is to linearize all the constraints and necessary equations (possibly after substituting out some variables, if so desired). Linearization amounts to finding a first-order approximation to all equations. Formally, linearization amounts to replacing a set of equations $0 = g(x_t)$ in a vector x_t of variables with its linearized counterpart around some point of approximation x^*, $0 = g(x^*) + g'(x^*)\tilde{x}_t$ where $\tilde{x}_t = x_t - x^*$ is the deviation of x_t from the approximation point x^* and where $G'(x^*)$ is the matrix of first derivatives of $G(.)$. As point of approximation x^*, the nonstochastic steady state is often chosen, that is, one solves the equations $0 = g(x^*)$ under the assumption that all exogenous stochastic variables are constant (here: $Y_t = Y^*$ and all $\varepsilon_s = 0$). Then, the remaining linearized system consists of $0 = g'(x^*)\tilde{x}_t$.

Since many economic variables are constrained to be positive, it is often more attractive to log-linearize the equations rather than to linearize them. The difference between linearization and log-linearization is that entries in x_t denote the original variable (for example, consumption c_t) in the case of linearization and the log of these variables (for example, $\log(c_t)$) in the case of log-linearization. There is no need to choose either linearization or log-linearization for all entries in x_t. One may choose to linearize some and log-linearize others or take other transformations. Indeed, for variables such as trade balances it is better to use linearization rather than log-linearization, if they can take negative values. Also, tax rates, for example, are often more appropriately linearized than log-linearized to provide a more useful interpretation.

This makes no difference as far as the linearized solution is concerned. More generally, differentiable and differentiable invertible transformations (that is, homeomorphisms) of the variables (for example, taking ratios of variables) make no difference to the properties of the linearized solution. The differences always lie only in the recalculation of the original variables, where one may want to take into account the nonlinearities originally inherent in the model. To see, more generally, that any homeomorphism (that is, differentiable and differentiably invertible transformation) $y_t = h(x_t)$ of the variables makes no difference to remaining calculations, note that the equations can be restated as $0 = g(h^{-1}(y_t))$. The linearized version is now $0 = g(h^{-1}(y^*)) + g'(x^*)(f^{-1})'(y^*)\tilde{y}_t$, which coincides with the previous linearization if $y^* = F(x^*)$, noting that $\tilde{y} = f'(x^*)\tilde{x}_t$ as well as $I = f'(x^*)(f^{-1})'(y^*)$.

While linearization can be performed numerically or with the usual rules of calculus, one can often ‘read’ the log-linearized version of an equation from its original form, exploiting $x_t = \exp(y_t) \approx x^* + x^*\tilde{y}_t$, where now $y_t = \log(x_t)$. Write $\hat{x}_t$ instead of $\tilde{y}_t$ for the loglinear deviation.

For log-linearization, the following useful ‘rules’ can easily be derived. Let a_t, b_t, c_t be three variables, with $c_t = h(a_t)$ for some monotone and differentiable function h(.), and let B be some constant. Then,

$$a_t + Bb_t \approx (a^* + Bb^*) + (a^*\hat{a}_t + Bb^*\hat{b}_t)Ba_tb_t \approx (Ba^*b^*) + (Ba^*b^*)(\hat{a}_t + \hat{b}_t)\hat{c}_t \approx \frac{h'(a^*)a^*}{h(a^*)}\hat{a}_t.$$

Either with these rules or directly, the equations in the example log-linearize to

$$c^*\hat{c}_t + k^*\hat{k}_t = y^*\hat{y}_t + (1-\delta)k^*\hat{k}_{t-1}\hat{y}_t = z_t + \frac{f_k k^*}{f}\hat{k}_{t-1} + \frac{f_n n^*}{f}\hat{n}_t 0 = n^*\hat{n}_t + \left(1 - n^*\right)\hat{l}_t\hat{\lambda}_t = \frac{u_{cc}c^*}{u_c}\hat{c}_t + \frac{u_{cl}l^*}{u_c}\hat{l}_t\frac{u_{cl}l^*}{u_l}\hat{c}_t + \frac{u_{ll}l^*}{u_l}\hat{l}_t = \frac{f_{nk}k^*}{f_n}\hat{k}_{t-1} + \frac{f_{nn}n^*}{f_n}\hat{n}_t\hat{\lambda}_t = E_t\left[\hat{\lambda}_{t+1} + \hat{R}_{t+1}\right]R^*\hat{R}_t = \frac{f_{kk}k^*}{f_k}\hat{k}_{t-1} + \frac{f_{kn}n^*}{f_k}\hat{n}_t.$$

Solving for the recursive law of motion

With some further algebra, one can turn this system into a second-order one-dimensional difference equation, $0 = E_t[Fx_{t+1} + Lz_{t+1}] + Gx_t + Mz_t + Hx_{t-1}$ plus the evolution of the exogenous state, $z_t = Nz_{t-1} + O\varepsilon_t$, where $x_t = k_t$ is the capital stock, and F, L, G, M, H, N and O are real numbers (here, with $N = \rho$ and $O = 1$). Alternatively, use the system of equations above directly (or with some variables substituted out) and stack all variables into a vector x_t to reformulate it in this form, where now F, L, G, M and H are matrices of coefficients. Indeed, if there is more than one predetermined variable like k_{t-1} in the system of equations, one will need to use such a matrix restatement of the equations anyways. More generally, z_t may also be a vector, and N and O matrices.

Anderson et al (1996) as well as Binder and Pesaran (1997) contain detailed and general results for solving linearized systems. In most cases, the system has a solution in the form of a recursive law of motion, $x_t = Px_{t-1} + Qz_t$, for some coefficient matrices P and Q. Most models require the solution to be stable, that is, all eigenvalues of P to be less than unity in absolute value. Often, one also allows for roots equal to unity in absolute value, as this arises easily in, for example, models of international trade or with multiple agents: one may then want to think of the linear approximation as a local solution. In many models, this uniquely determines the matrix P and usually also Q. The solutions can be found by substituting the recursive law of motion in for x_{t+1} and again for all x_t into the second-order difference equation above, exploiting $Nz_t = E_t[z_{t+1}]$ so that only x_{t-1} and z_t and some coefficient matrices remain.

Examine first the equation by matching coefficients on x_{t-1}. One obtains the equation $0 = FP^2 + GP + H$ for P. In case of a one-dimensional difference equation (as can be obtained for the example above and $x_t = k_t$), this is a quadratic equation in the feedback coefficient P, which has two solutions. The system is said to be saddle-path stable if only one of the two roots is smaller than unity in absolute value. Thus, if a stable solution is desired, this is the unique solution for P.

Generally, the equation above is a matrix quadratic equation, which can be solved per computing generalized eigenvalues or by QZ decomposition as follows. Let m be the dimensionality of x_t. Define the matrices

$$A = \begin{bmatrix} -G & -H \\ I_m & 0_m \end{bmatrix}, B = \begin{bmatrix} F & 0_m \\ 0_m & I_m \end{bmatrix}$$

where I_m is the m-by-m identity matrix and 0_m the *m*-by-*m* matrices of only zeros. Recall that a generalized eigenvector *s* with eigenvalue λ for the matrices *A* and *B* is defined as satisfying $\lambda Bs = As$. The generalized eigenvector problem reduces to the standard eigenvector problem of $B^{-1}A$, if *B* is invertible. If *s* is a generalized eigenvector with eigenvalue λ for the matrices *A* and *B* above, it can be written as $s' = [\lambda x', x']$ for some *m*-dimensional vector *x*. If there are *m* generalized eigenvalues $\lambda_1, \ldots, \lambda_m$ together with generalized eigenvectors $s_i = [\lambda_i x_i', x_i']$ such that $C = [x_1, \ldots, x_m]$ is of full rank, then $P = C\Lambda C^{-1}$ is a solution to the matrix quadratic equation, where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_m \end{bmatrix}$$

is the diagonal matrix of the eigenvalues for the generalized eigenvectors used as well as of *P*. The system is said to be saddle-path stable if there are exactly *m* generalized eigenvalues smaller than unity in absolute value. In that case, the matrix *P* is unique, if one requires all eigenvalues of *P* to be stable. If there are fewer than *m* eigenvalues smaller than (or equal to) unity in absolute value, then there is no solution, such that the difference equation $x_t = Px_{t-1}$ remains bounded for all x_0. In that case, the set of bounded solution is characterized by $e'x_0 = 0$ as well as $e'Qz_t = 0$ for all *t* for all eigenvectors *e* of *P* corresponding to explosive eigenvalues. The second of these two constraints may impose restrictions on the exogenous shock process. If there are more than *m* eigenvalues smaller than (or equal to) unity in absolute value, then sunspot solutions may arise, that is, there are additional solutions. In the one-dimensional case and if *F* is nonzero, the general solution is now given by the original equation, that is, as $x_t = -F^{-1}Gx_{t-1} - F^{-1}Hx_{t-2} - F^{-1}(LN + M)z_{t-1} + v_t$ where ν_t is any stochastic process with $E_t[v_{t+1}] = 0$ and which is independent of all ε_s for $s > t$, but not necessarily independent of ε_t. Note that the recursive law of motion now includes an additional lag of the state variable, as well as the possibility for additional random influences ('sunpots') via ν_t, which are not part of the original system of equations. Farmer (1999) provides a detailed treatment of sunspots in linearized solutions.

Equivalently, consider the stacked variable $s_t' = [x_t', x_{t-1}']$, and note that the second half of this vector is 'predetermined', that is, must be independent of all ε_s for $s > t - 1$. The linearized system can be rewritten as

$$BE_t[s_{t+1}] = As_t + \begin{bmatrix} -M - LN \\ 0 \end{bmatrix} z_t.$$

If *B* is invertible, the solutions can now be characterized in terms of the eigenvalues and eigenvectors of $B^{-1}A$. This is the approach taken in the classic reference of Blanchard and Kahn (1980).

Alternatively, find the QZ decomposition (or generalized Schur decomposition) of *A* and *B* (see Sims, 2002), that is, find unitary matrices *U* and *V* as well as upper triangular matrices *K* and *L* such that

$$A = U'LVB = U'KV$$

(and recall that a matrix is unitary, if the product with its complex conjugate transpose is the identity matrix). Such a Schur decomposition always exists, although it may not be unique. Partition *U* and *V* into *m*-by-*m* submatrices,

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}, V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

If U_{21} and V_{21} are invertible, then $P = -V_{21}^{-1}V_{22}$ solves the matrix quadratic equation. Suppose furthermore, that the QZ decomposition has been chosen so that the ratios $|L_{ii}/K_{ii}|$ are in ascending order. Furthermore, suppose $|L_{mm}/K_{mm}| < 1$. Then *P* is stable.

To solve for *Q*, given a solution to *P*, compare the coefficients on z_t to find $V\mathrm{vec}(Q) = -\mathrm{vec}(LN + M)$ where vec(.) denotes columnwise vectorization and where $V = N' \otimes F + I_k \otimes (FP + G)$ with *k* the dimensionality of z_t. If *V* is invertible, the solution is unique.

Note: Many links for codes for solving dynamic stochastic models are available from QM&RBC Codes Online, Department of Economics, University of Connecticut, http://dge.repec.org/codes.html (accessed 4 September 2006). The procedure outlined above has been used in particular in the author's 'A toolkit for analyzing nonlinear economic dynamic models easily: MATLAB programs', http://www.wiwi.hu-berlin.de/wpol/html/toolkit.htm (accessed 4 September 2006). For a discussion of the accuracy of linearized solutions, see, for example, Taylor and Uhlig (1990) and Aruoba, Fernández-Villaverde and Rubio-Ramírez (2006).

See Also

- multiple equilibria in macroeconomics
- numerical optimization methods in economics
- Prescott, Edward Christian
- real business cycles
- simulation estimators in macroeconometrics
- sunspot equilibrium
- vector autoregressions

Bibliography

Anderson, E., McGrattan, E., Hansen, L. and Sargent, T. 1996. Mechanics of forming and estimating dynamic linear economies. In *Handbook of Computational Economics*, vol. 1, ed. H. Amman et al. Amsterdam: North-Holland.

Aruoba, B., Fernández-Villaverde, J. and Rubio-Ramírez, J. 2006. Comparing solution methods for dynamic equilibrium economies. *Journal of Economic Dynamics and Control*.

Binder, M. and Pesaran, M. 1997. multivariate linear rational expectations models: characterization of the nature of the solutions and their fully recursive computation. *Econometric Theory* 13, 877–88.

Blanchard, O. and Kahn, C. 1980. The solution of linear difference models under rational expectations. *Econometrica* 48, 1305–12.

Farmer, R. 1999. *Macroeconomics of Self-fulfilling Prophecies*. Cambridge, MA: MIT Press.

Hansen, G. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16, 309–27.

King, R., Plosser, C. and Rebelo, S. 2002. Production, growth and business cycles: technical appendix. *Computational Economics* 20, 87–116.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Sims, C. 2002. Solving linear rational expectations models. *Computational Economics* 20, 1–20.

Taylor, J. and Uhlig, H. 1990. Solving nonlinear stochastic growth models: a comparison of alternative solution methods. *Journal of Business Economics and Statistics* 8, 1–17.

Uhlig, H. 1999. A toolkit for analysing nonlinear dynamic stochastic models easily. In *Computational Methods for the Study of Dynamic Economies*, ed. R. Marimon and A. Scott. Oxford: Oxford University Press.

How to cite this article

Uhlig, Harald. "approximate solutions to dynamic models (linear methods)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000237> doi:10.1057/9780230226203.0050

The New Palgrave Dictionary of Economics Online

Aquinas, St Thomas (1225 – 1274)

Barry Gordon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Aquinas, St Thomas; Aristotle; commutative justice; compensation; just price; justice; market price; restitution; usury; value

Article

St Thomas Aquinas is generally acknowledged as the outstanding theologian of the high Middle Ages. A member of the Dominican order and a pupil of Albertus Magnus (1206–80), St Thomas taught at a number of centres including Paris, Anagni, Orvieto, Rome, Viterbo and Naples. In his research he drew on an extensive range of sources, from the Christian tradition (based on the Scriptures, the Fathers and the Roman writers) to Greek philosophy including the thought of the newly ‘rediscovered’ Aristotle. The writings of Aquinas are also wide-ranging, including commentaries on Aristotle's *Politics* and *Ethics*. Most celebrated among his major works is the *Summa Theologica*, which was set down between 1265 and 1273.

For St Thomas, economic reasoning is integrated with moral philosophy and the establishment of legal precepts. Analysis of economic activity is undertaken for the sake of determining appropriate standards in dealings between citizen and citizen, and so is an aspect of the inquiry into justice. The category of justice which Aquinas finds most relevant to economic life is commutative justice (from *commutatio*, that is, transaction). Hence the focal points for his economic reasoning are value and price, money and interest.

On money, St Thomas stresses its roles as a medium for the exchange of commodities and as a unit of account, that is, a standard of value or measuring rod for comparing the relative worths of exchangeable things. In his treatments of compensation for delay in repayment of a money loan and of restitution of stolen money Aquinas also recognizes that money may have economic significance when held in balance (especially when held by businessmen). The stress on money as a medium of exchange and unit of account leads to a condemnation of most forms of interest-taking as usury, hence unjust. However, the analysis of restitution and compensation help pave the way for the later acceptance by theologians of *lucrum cessans* and *damnum emergens* as phenomena offering bases for a legitimate positive rate of

interest.

The just price of any commodity for St Thomas is its current market price, established in the absence of fraud or monopolistic trading practices. It is a price established by *communiter venditur*, the price generally charged in the community concerned, rather than the price dictated by the preferences or needs of any one individual in that community. The value of a commodity will depend on subjective estimates of the utility of the good in question. It will also depend, in part, on cost of production, in that the latter influences supply conditions in any particular market. Aquinas does not achieve an effective synthesis of the utility and cost elements in his analysis of value, nor does he extend the analysis into a theory of distribution. These latter problems, however, were addressed by some of his Scholastic successors, often with reference to the analytical framework devised by St Thomas.

See Also

- scholastic economics

Selected works

An English translation of Aquinas' most celebrated work is: St Thomas Aquinas, *Summa Theologiae*, translated and edited by M. Lefebure, New York: Oxford University Press, 1975. There is also a translation of one of his commentaries on Aristotle, *Commentary on the Nicomachean Ethics*, Chicago: Library of Living Catholic Thought, 1964. Selected passages from the writings of St Thomas which are of interest for economists are included in A.E. Monroe, *Early Economic Thought*, Cambridge, MA: Harvard University Press, 1924, and in A.C. Pegis, ed., *Basic Writings of St Thomas Aquinas*, 2 vols, New York: Random House, 1945. A Latin edition of Aquinas' works is: St Thomas Aquinas, *Opera Omnia*, 34 vols, ed. P. Mare and S.E. Frette, Paris: Vives, 1871–80.

Bibliography

Baldwin, J.W. 1959. *The Medieval Theories of the Just Price*. Philadelphia: American Philosophical Society.

Blaug, M., ed. 1981. *St. Thomas Aquinas*. London: Edward Elgar.

Gilchrist, J.T. 1969. *The Church and Economic Activity in the Middle Ages*. London: Macmillan.

Gordon, B. 1975. *Economic Analysis Before Adam Smith*. London: Macmillan.

Langholm, O. 1979. *Price and Value Theory in the Aristotelian Tradition*. Bergen: Universitetsforlaget.

Langholm, O. 1984. *The Aristotelian Analysis of Usury*. Bergen: Universitetsforlaget.

Langholm, O. 1998. *The Legacy of Scholasticism in Economic Thought: Antecedents of Choice and Power*. London: Cambridge University Press.

Noonan, J.T. 1957. *The Scholastic Analysis of Usury*. Cambridge, MA: Harvard University Press.

Stark, W. 1956. *The Contained Economy: An Interpretation of Medieval Economic Thought*. London: Blackfriars.

Torrell, J.P. 1996. *St. Thomas Aquinas. Volume I: The Person and His Work*. Washington DC: Catholic University of America Press.

Torrell, J.P. 2003. *St. Thomas Aquinas. Volume II: Spiritual Master*. Washington, DC: Catholic University of America Press.

Viner, J. 1978. *Religious Thought and Economic Society*. Durham, NC: Duke University Press.

Weisheipl, J.A. 1974. *Friar Thomas D'Aquino: His Life, Thought and Work*. Garden City, NY: Doubleday.

Worland, S.T. 1967. *Scholasticism and Welfare Economics*. Notre Dame: University of Notre Dame Press.

How to cite this article

Gordon, Barry. "Aquinas, St Thomas (1225–1274)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000117> doi:10.1057/9780230226203.0051

The New Palgrave Dictionary of Economics Online

arbitrage pricing theory

Gur Huberman and Zhenyu Wang

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Focusing on asset returns governed by a factor structure, the APT is a one-period model, in which preclusion of arbitrage over static portfolios of these assets leads to a linear relation between the expected return and its covariance with the factors. The APT, however, does not preclude arbitrage over dynamic portfolios. Consequently, applying the model to evaluate managed portfolios contradicts the no-arbitrage spirit of the model. An empirical test of the APT entails a procedure to identify features of the underlying factor structure rather than merely a collection of mean-variance efficient factor portfolios that satisfies the linear relation.

Keywords

arbitrage; arbitrage pricing theory; Arrow–Debreu security pricing; asset allocation; asset pricing; Black–Scholes model; capital asset pricing model; cost of capital; factor models; generalized method of moments; Hilbert space techniques; mean-variance efficiency; portfolio analysis; stochastic discount factor

Article

The arbitrage pricing theory (APT) was developed primarily by Ross (1976a; 1976b). It is a one-period model in which every investor believes that the stochastic properties of returns of capital assets are consistent with a factor structure. Ross argues that, if equilibrium prices offer no arbitrage opportunities over static portfolios of the assets, then the expected returns on the assets are approximately linearly related to the factor loadings. (The factor loadings, or betas, are proportional to the returns' covariances with the factors.) The result is stated in section 1.

Ross's (1976a) heuristic argument for the theory is based on the preclusion of arbitrage. This intuition is sketched in Section 2. Ross's formal proof shows that the linear pricing relation is a necessary condition for equilibrium in a market where agents maximize certain types of utility. The subsequent work, which

is surveyed below, derives either from the assumption of the preclusion of arbitrage or the equilibrium of utility maximization. A linear relation between the expected returns and the betas is tantamount to an identification of the stochastic discount factor (SDF). Sections 3 and 4, respectively, review this literature.

The APT is a substitute for the capital asset pricing model (CAPM) in that both assert a linear relation between assets' expected returns and their covariance with other random variables. (In the CAPM, the covariance is with the market portfolio's return.) The covariance is interpreted as a measure of risk that investors cannot avoid by diversification. The slope coefficient in the linear relation between the expected returns and the covariance is interpreted as a risk premium. Such a relation is closely tied to mean-variance efficiency, which is reviewed in section 5.

Section 5 also points out that an empirical test of the APT entails a procedure to identify at least some features of the underlying factor structure. Merely stating that some collection of portfolios (or even a single portfolio) is mean-variance efficient relative to the mean-variance frontier spanned by the existing assets does not constitute a test of the APT, because one can always find a mean-variance efficient portfolio. Consequently, as a test of the APT it is not sufficient to merely show that a set of factor portfolios satisfies the linear relation between the expected return and its covariance with the factors portfolios.

A sketch of the empirical approaches to the APT is offered in section 6, while section 7 describes various procedures to identify the underlying factors. The large number of factors proposed in the literature and the variety of statistical or ad hoc procedures to find them indicate that a definitive insight on the topic is still missing.

Finally, section 8 surveys the applications of the APT, the most prominent being the evaluation of the performance of money managers who actively change their portfolios. Unfortunately, the APT does not necessarily preclude arbitrage opportunities over dynamic portfolios of the existing assets. Therefore, the applications of the APT in the evaluation of managed portfolios contradict at least the spirit of the APT, which obtains price restrictions by assuming the absence of arbitrage.

1 A formal statement

The APT assumes that investors believe that the $n \times 1$ vector, *r*, of the single-period random returns on capital assets satisfies the factor model

$$r = \mu + \beta f + e, \tag{1}$$

where *e* is an $n \times 1$ vector of random variables, *f* is a $k \times 1$ vector of random variables (factors), μ is an $n \times 1$ vector and β is an $n \times k$ matrix. With no loss of generality, normalize (1) to make $E[f] = 0$ and $E[e] = 0$, where $E[\cdot]$ denotes expectation and 0 denotes the matrix of zeros with the required dimension. The factor model (1) implies $E[r] = \mu$.

The mathematical proof of the APT requires restrictions on β and the covariance matrix $\Omega = E[ee']$. An additional customary assumption is that $E[e|f] = 0$, but this assumption is not necessary in some of the APT's developments.

The number of assets, *n*, is assumed to be much larger than the number of factors, *k*. In some models, *n* is infinity or approaches infinity. In this case, representation (1) applies to a sequence of capital markets; the first *n* assets in the (*n*+1)st market are the same as the assets in the *n*th market and the first *n* rows of the matrix β in the (*n*+1)st market constitute the matrix β in the *n*th market.

The APT asserts the existence of a constant *a* such that, for each *n*, the inequality

$$(\mu - X\lambda)Z^{-1}(\mu - X\lambda) \leq a \tag{2}$$

holds for a $(k+1) \times 1$ vector λ, and an $n \times n$ positive definite matrix *Z*. Here, $X = (\iota, \beta)$, in which ι is an $n \times 1$ vector of ones. Let λ_0 be the first component of λ and λ_1 consists of the rest of the components. If some portfolio of the assets is risk-free, then λ_0 is the return on the risk-free portfolio.

The positive definite matrix *Z* is often the covariance matrix $E[ee']$. *Exact* arbitrage pricing obtains if (2) is replaced by

$$\mu = X\lambda = \iota\lambda_0 + \beta\lambda_1. \tag{3}$$

The vector λ_1 is referred to as the risk premium, and the matrix β is referred to as the beta or loading on factor risk.

The interpretation of (2) is that each component of μ depends approximately linearly on the corresponding row of β. This linear relation is the same across assets. The approximation is better, the smaller the constant *a*; if $a = 0$, the linear relation is exact and (3) obtains.

2 Intuition

The intuition behind the model draws from the intuition behind Arrow–Debreu security pricing. A set of *k* fundamental securities spans all possible future states of nature in an Arrow–Debreu model. Each asset's payoff can be described as the payoff on a portfolio of the fundamental *k* assets. In other words, an asset's payoff is a weighted average of the fundamental assets' payoffs. If market clearing prices allow no arbitrage opportunities, then the current price of each asset must equal the weighted average of the current prices of the fundamental assets.

The Arrow–Debreu intuition can be couched in terms of returns and expected returns rather than payoffs and prices. If the unexpected part of each asset's return is a linear combination of the unexpected parts of the returns on the *k* fundamental securities, then the expected return of each asset is the same linear combination of the expected returns on the *k* fundamental assets.

To see how the Arrow–Debreu intuition leads from the factor structure (1) to exact arbitrage pricing (3), set the idiosyncratic term *e* on the right-hand side of (1) equal to zero. Translate the *k* factors on the right-hand side of (1) into the *k* fundamental securities in the Arrow–Debreu model. Then (3) follows immediately.

The presence of the idiosyncratic term *e* in the factor structure (1) makes the model more general and realistic. It also makes the relation between (1) and (3) more tenuous. Indeed, ‘no arbitrage’ arguments typically prove the weaker (2). Moreover, they require a weaker definition of arbitrage (and therefore a stronger definition of no arbitrage) in order to get from (1) to (2).

The proofs of (2) augment the Arrow–Debreu intuition with a version of the law of large numbers. That law is used to argue that the average effect of the idiosyncratic terms is negligible. In this argument, the independence among the components of *e* is used. Indeed, the more one assumes about the (absence of) contemporaneous correlations among the component of *e*, the tighter the bound on the deviation from exact APT.

3 No-arbitrage models

Huberman (1982) formalizes Ross's (1976a) heuristic argument. A portfolio *v* is an $n \times 1$ vector. The cost of the portfolio *v* is $v'\iota$, the income from it is $v'r$, and its return is $v'r/v'\iota$ (if its cost is not zero). Huberman defines arbitrage as the existence of zero-cost portfolios such that a subsequence $\{w\}$ satisfies

$$\lim_{n\to\infty} E[w'r] = \infty \quad \text{and} \quad \lim_{n\to\infty} \operatorname{var}[w'r] = 0, \tag{4}$$

where $\operatorname{var}[\cdot]$ denotes variance. The first requirement in (4) is that the expected income associated with *w* becomes large as the number of assets increases. The second requirement in (4) is that the risk (as measured by the income's variance) vanishes as the number of assets increases. Accordingly, a sequence of capital markets offers no arbitrage if there is no subsequence $\{w\}$ of zero-cost portfolios that satisfy (4).

Huberman shows that, if the factor model (1) holds and if the covariance matrix $E[ee']$ is diagonal for all *n* and uniformly bounded, then the absence of arbitrage implies (2) with $Z = I$ and a finite bound *a*. The idea of his proof is as follows. Consider the orthogonal projection of the vector μ on the linear space spanned by the columns of *X*:

$$\mu = X\hat{\lambda} + \alpha, \tag{5}$$

where $\alpha' X = 0$ and $\hat{\lambda}$ is a $k \times 1$ vector. The projection implies

$$\alpha'\alpha = \min_{\lambda} (\mu - X\lambda)'(\mu - X\lambda). \tag{6}$$

A violation of (2) is the existence of a subsequence of $\{\alpha'\alpha\}$ that approaches infinity. The vector α is often referred to as a pricing error and it can be used to construct arbitrage. For any scalar *h*, the portfolio $w = h\alpha$ has zero cost because the first column of *X* is ι. The factor model (1) and the projection (5) imply $E[w'r] = h(\alpha'\alpha)$ and $\mathrm{var}[w'r] = h^2(\alpha' E[ee']\alpha)$. If σ^2 is the upper bound of the diagonal elements of $E[ee']$, then $\mathrm{var}[w'r] \leq h^2(\alpha'\alpha)\sigma^2$. If *h* is chosen to be $(\alpha'\alpha)^{-2/3}$, then $E[w'r] = (\alpha'\alpha)^{1/3}$ and $\mathrm{var}[w'r] \leq (\alpha'\alpha)^{-1/3}\sigma^2$, which imply that (4) is satisfied by a subsequence of the zero-cost portfolios $\{(\alpha'\alpha)^{-2/3}\alpha\}$.

Using the no-arbitrage argument, the exact APT can be proven to hold in the limit for well-diversified portfolios. A portfolio *w* is well diversified if $w'\iota = 1$ and $\mathrm{var}[w'e] = 0$, that is, if the portfolio's return contains only factor variance. A sequence of portfolios, $\{w\}$, is well diversified if $w'\iota = 1$ and $\lim_{n\to\infty} \mathrm{var}[w'e] = 0$. Suppose there are *m* sequences of well-diversified portfolios and *m* is a fixed number larger than $k + 1$. For each *n*, let *W* be an $n \times m$ matrix, in which each column is one of the well-diversified portfolios. The exact APT holds in the limit for the well-diversified portfolios if and only if there exists a sequence of $k \times 1$ vectors, $\{\lambda\}$, such that

$$\lim_{n\to\infty} (W'\mu - \tilde{X}\lambda)'(W'\mu - \tilde{X}\lambda)) = 0, \tag{7}$$

where $\tilde{X} = (\iota, W'\beta)$ and ι is an $m \times 1$ vector of ones. The projection of $W'\mu$ on the columns of $\tilde{X}$ gives $W'\mu = \tilde{X}\hat{\lambda} + \alpha$, in which $\alpha'\tilde{X} = 0$. If eq. (7) does not hold, a subsequence of α satisfies $\alpha'\alpha > \delta$ for some

positive constant δ. This sequence of α can be used to construct arbitrage as follows. For any scalar *h*, define a portfolio as $v = hW\alpha$, which is then costless because $v'\iota = h\alpha'W'\iota = h\alpha'\jmath = 0$. It follows from $\alpha'\tilde{X} = 0$ that $E[v'r] = h\alpha'\alpha$ and $\mathrm{var}[v'r] = h^2\alpha'W'E[\varepsilon\varepsilon']W\alpha$. If *h* is chosen to be $(\alpha'W'E[\varepsilon\varepsilon']W\alpha)^{-1/3}$, then $\mathrm{var}[v'r] = h^{-1}$. Since $\{W\}$ is well-diversified and $E[\varepsilon\varepsilon']$ is diagonal and uniformly bounded, it follows that $\lim_{n\to\infty} h = \infty$. This implies that portfolio sequence $\{v\}$ is arbitrage because it satisfies (4).

Ingersoll (1984) generalizes Huberman's result, showing that the factor model, uniform boundedness of the elements of β and no arbitrage imply (2) with $Z = E[\varepsilon\varepsilon']$, which is not necessarily diagonal. A variant of Ingersoll's argument is as follows. Write the positive definite matrix *Z* as the product $Z = UU'$, where *U* is an $n \times n$ non-singular matrix. Then, consider the orthogonal projection of the vector $U^{-1}\mu$ on the column space of $U^{-1}X$:

$$U^{-1}\mu = U^{-1}X\lambda + \alpha, \tag{8}$$

where $\alpha'U^{-1}X = 0$. The rest of the argument is similar to those presented earlier.

Chamberlain and Rothschild (1983) employ Hilbert space techniques to study capital markets with (possibly infinitely) many assets. The preclusion of arbitrage implies the continuity of the cost functional in the Hilbert space. Let *L* equal the maximum eigenvalue of the limit covariance matrix $E[\varepsilon\varepsilon']$ and *d* equal the supremum of all the ratios of expectation to standard deviation of the incomes on all costless portfolios with a non-zero weight on at least one asset. Chamberlain and Rothschild demonstrate that (2) holds with $a = Ld^2$ and $Z = I$ if asset prices allow no arbitrage.

With two additional assumptions, Chamberlain (1983) provides explicit lower and upper bounds on the left-hand side of (2). He further shows that exact arbitrage pricing obtains if and only if there is a well-diversified portfolio on the mean-variance frontier. The first of his additional assumptions is that all the factors can be represented as limits of traded assets. The second additional assumption is that the variances of incomes on any sequence of portfolios that are well diversified in the limit and that are uncorrelated with the factors converge to zero.

4 Utility-based arguments

In utility-based arguments, investors are assumed to solve the following problem:

$$\max_{c_0, c_T, w} E[u(c_0, c_T)] \text{ subject to } c_0 \le b - w'\iota \text{ and } c_T \le w'r,$$

(9)

where b is the initial wealth, and $u(c_0, c_T)$ is a utility function of initial and terminal consumption c_0 and c_T. The utility function is assumed to increase with initial and with terminal consumption. The first order condition is

$$E[rM] = 1,$$
(10)

where $M = (\partial u / \partial c_T) / (\partial u / \partial c_0)$. The random variable M satisfying (10) is referred to as the stochastic discount factor (SDF) by Hansen and Jagannathan (1991; 1997). Substitution of the factor model (1) into the first order condition gives

$$\mu = 1\lambda_0 + \beta\lambda_1 + \alpha,$$
(11)

where $\lambda_0 = 1 / E[M]$, $\lambda_1 = -E[fM] / E[M]$ and $\alpha = -E[eM] / E[M]$. It follows from (11) that

$$(\mu - X\lambda)'(\mu - X\lambda) = \alpha'\alpha,$$
(12)

where $X = (1, \beta)$ and $\lambda = (\lambda_0, \lambda_1')'$.
Clearly, the APT (2) holds for $Z = 1$ and a if $\alpha'\alpha$ is uniformly bounded by a. Ross (1976a) is the first to set up an economy in which $\alpha'\alpha$ is uniformly bounded. The exact APT (3) holds if and only if

$$E[eM] = 0.$$
(13)

If the SDF is a linear function of the factors, then eq. (13) holds. Conversely, if eq. (13) holds, there exists an SDF, which is a linear function of factors, such that eq. (10) is satisfied. However, the SDF does not have to be a linear function of factors for the purpose of obtaining the exact APT. A nonlinear function, $M = g(f)$, of factors for the SDF would also imply (13) under the assumption $E[e|f] = 0$.
Connor (1984) shows that, if the market portfolio is well diversified, then every investor holds a well-diversified portfolio (that is, a $k + 1$ fund separation obtains; the funds are associated with the factors and with the risk-free asset, which Connor assumes to exist). With this, the first order condition of any investor implies exact arbitrage pricing in a competitive equilibrium.
Connor and Korajczyk (1986) extend Connor's previous work to a model with investors who have better information about returns than most other investors. The former class of investors is sufficiently small, so the pricing result remains intact and it is used to derive a test of the superiority of information of the allegedly better informed investors.
Connor and Korajczyk (1988) extend Connor's single-period model to a multi-period model. They assume that the capital assets are the same in all periods, that each period's cash payoffs from these assets obey a factor structure, and that competitive equilibrium prices are set as if the economy had a representative investor who maximizes exponential utility. They show that exact arbitrage pricing obtains with time-varying risk premium (but, similar to Stambaugh, 1983, with constant factor loadings.)
Chen and Ingersoll (1983) argue that, if a well-diversified portfolio exists and it is the optimal portfolio of some utility-maximizing investor, then the first order condition of that investor implies exact arbitrage pricing.
Dybvig (1983) and Grinblatt and Titman (1983) consider the case of finite assets and provide explicit bounds on the deviations from exact arbitrage pricing. These bounds are functions of the per capita asset supplies, individual bounds on absolute risk aversion, variance of the idiosyncratic risk, and the interest rate. To derive his bound, Dybvig assumes that the support of the distribution of the idiosyncratic term *e* is bounded below, that each investor's coefficient of absolute risk aversion is non-increasing and that the competitive equilibrium allocation is unconstrained Pareto optimal. To derive their bound, Grinblatt and Titman require a bound on a quantity related to investors' coefficients of absolute risk aversion and the existence of *k* independent, costless and well diversified portfolios.

5 Mean-variance efficiency

The APT was developed as a generalization of the CAPM, which asserts that the expectations of assets' returns are linearly related to their covariances (or betas, which in turn are proportional to the covariances) with the market portfolio's return. Equivalently, the CAPM says that the market portfolio is mean-variance efficient in the investment universe containing all possible assets. If the factors in (1) can be identified with traded assets, then exact arbitrage pricing (3) says that a portfolio of these factors is mean-variance efficient in the investment universe consisting of the assets *r*.
Huberman and Kandel (1985b), Jobson and Korkie (1982; 1985) and Jobson (1982) note the relation between the APT and mean-variance efficiency. They propose likelihood-ratio tests of the joint hypothesis that a given set of random variables are factors in model (1) and that exact arbitrage pricing (3) obtains. Kan and Zhou (2001) point out a crucial typographical error in Huberman and Kandel

(1985b). Peñaranda and Sentana (2004) study the close relation between the Huberman and Kandel's spanning approach and the celebrated volatility bounds in Hansen and Jagannathan (1991).

Even when the factors are not traded assets, (3) is a statement about mean-variance efficiency: Grinblatt and Titman (1987) assume that the factor structure (1) holds and that a risk-free asset is available. They identify *k* traded assets such that a portfolio of them is mean-variance efficient if and only if (3) holds. Huberman, Kandel and Stambaugh (1987) extend the work of Grinblatt and Titman by characterizing the sets of *k* traded assets with that property and show that these assets can be described as portfolios if and only if the global minimum variance portfolio has non-zero systematic risk. To find these sets of assets, one must know the matrices $\beta\beta'$ and $E[ee']$. If the latter matrix is diagonal, factor analysis produces an estimate of it, as well as an estimate of $\beta\beta'$.

The interpretation of (3) as a statement about mean-variance efficiency contributes to the debate about the testability of the APT. (Shanken, 1982; 1985, and Dybvig and Ross, 1985, however, discuss the APT's testability without mentioning that (3) is a statement about mean-variance efficiency.) The theory's silence about the factors' identities renders any test of the APT a joint test of the pricing relation and the correctness of the factors. As a mean-variance efficient portfolio always exists, one can always find 'factors' with respect to which (3) holds. In fact, any single portfolio on the frontier can serve as a 'factor'.

Thus, finding portfolios which are mean-variance efficient – or failure to find them – neither supports nor contradicts the APT. It is the factor structure (1) which, combined with (3), provides refutable hypotheses about assets' returns. The factor structure (1) imposes restrictions which, combined with (3), provide refutable hypotheses about assets' returns. The factor structure suggests looking for factors with two properties: (*a*) their time-series movements explain a substantial fraction of the time-series movements of the returns on the priced assets, and (*b*) the unexplained parts of the time series movements of the returns on the priced assets are approximately uncorrelated across the priced assets.

6 Empirical tests

Empirical work inspired by the APT typically ignores (2) and instead studies exact arbitrage pricing (3). This type of work usually consists of two steps: an estimation of factors (or at least of the matrix β) and then a check to see whether exact arbitrage pricing holds. In the first step, researchers typically use the following regression model to estimate the parameters in the factor model:

$$r_t = \alpha + \beta f_t + e_t, \tag{14}$$

where r_t, f_t and e_t are the realization of the variables in period *t*. The factors observed in empirical studies often have a non-zero mean, denoted by δ. Let *T* be the total number of periods and Σ the summation over $t = 1, \ldots, T$. The ordinary least-square (OLS) estimates are

$$\hat{\mu} = \frac{1}{T}\sum r_t \text{ and } \hat{\delta} = \frac{1}{T}\sum f_t$$
(15)

$$\hat{\beta} = \left(\sum (r_t - \hat{\mu})(f_t - \hat{\delta})'\right)\left(\sum (f_t - \hat{\delta})(f_t - \hat{\delta})'\right)^{-1}$$
(16)

$$\hat{\alpha} = \hat{\mu} - \hat{\beta}\hat{\delta}$$
(17)

$$\hat{\Omega} = \frac{1}{T}\sum \hat{e}_t \hat{e}'_t \text{ where } \hat{e}_t = r_t - \hat{\alpha} - \hat{\beta} f_t.$$
(18)

These are also maximum-likelihood estimators if the returns and factors are independent across time and have a multivariate normal distribution.
In the second step, researchers may use the exact pricing (3) and (14) to obtain the following restricted version of the regression model,

$$r_t = \iota\lambda_0 + \beta(f_t + \lambda_1) + e_t.$$
(19)

Under the assumption that returns and factors follow identical and independent normal distributions, the maximum-likelihood estimators are

$$\bar{\beta} = \left(\sum (r_t - \iota\bar{\lambda}_0)(f_t + \bar{\lambda}_1)'\right)\left(\sum (f_t + \bar{\lambda}_1)(f_t + \bar{\lambda}_1)'\right)^{-1} \tag{20}$$

$$\bar{\Omega} = \frac{1}{T}\sum \bar{e}_t\bar{e}'_t \text{where} \bar{e}_t = r_t - \iota\bar{\lambda}_0 - \bar{\beta}(f_t + \bar{\lambda}_1) \tag{21}$$

$$\bar{\lambda} = (\bar{X}'\bar{\Omega}^{-1}\bar{X})^{-1}\bar{X}'\bar{\Omega}^{-1}(\hat{\mu} - \bar{\beta}\hat{\delta}) \text{where} \bar{X} = (\iota, \bar{\beta}). \tag{22}$$

These estimators need to be solved simultaneously from the above three equations. Notice that $\bar{\beta}$ and $\bar{\Omega}$ are the OLS estimators in (19) for a given $\bar{\lambda}$. The last equation shows that $\bar{\lambda}$ is the generalized least-square estimator in the cross-sectional regression of $\hat{\mu} - \bar{\beta}\hat{\delta}$ on $\bar{X}$ with $\bar{\Omega}$ being the weighting matrix. To test the restriction imposed by the exact APT, researchers use the likelihood-ratio statistic,

$$\mathrm{LR} = T(\log|\bar{\Omega}| - \log|\hat{\Omega}|), \tag{23}$$

which follows a χ^2 distribution with $n - k - 1$ degrees of freedom when the number of observations, *T*, is very large. When factors are payoffs of traded assets or a risk-free asset exists, the exact APT imposes more restrictions. For these cases, Campbell, Lo and MacKinley (1997, ch. 6) provide an overview. If the observations of returns and factors do not follow independent normal distribution, similar tests can be carried out using the generalized method of moments (GMM). Jagannathan and Wang (2002) and

Jagannathan, Skoulakis and Wang (2002) provide an overview of the application of the GMM for testing asset pricing models including the APT.

Interest is sometimes focused only on whether a set of specified factors are priced or on whether their loadings help explain the cross section of expected asset returns. For this purpose, most researchers study the cross-sectional regression model

$$\hat{\mu} = \hat{X}\lambda + v \text{ or } \hat{\mu} = \iota\lambda_0 + \hat{\beta}\lambda_1 + v, \tag{24}$$

where $\hat{X} = (\iota, \hat{\beta})$ and v is an $n \times 1$ vector of errors for this equation. The OLS estimator of λ in this regression is tested to see whether it is different from zero. To test this specification, asset characteristics z, such as firm size, that are correlated with mean asset returns are added to the regression:

$$\hat{\mu} = \iota\lambda_0 + \hat{\beta}\lambda_1 + z\lambda_2 + v. \tag{25}$$

A significant λ_1 and insignificant λ_2 are viewed as evidence in support of the specified factors being part of the exact APT. Black, Jensen and Scholes (1972) and Fama and MacBeth (1973) pioneered this cross-sectional approach to test the CAPM. Chen, Roll and Ross (1986) used it to test the exact APT. Shanken (1992) and Jagannathan and Wang (1998) developed the statistical foundations of the cross-sectional tests. The cross-sectional approach is now a popular tool for analysing risk premiums on the loadings of proposed factors.

7 Specification of factors

The tests outlined above are joint tests that the matrix β is correctly estimated and that exact arbitrage pricing holds. Estimation of the factor loading matrix β entails at least an implicit identification of the factors. The three approaches listed below have been used to identify factors.

The first consists of an algorithmic analysis of the estimated covariance matrix of asset returns. For instance, Roll and Ross (1980), Chen (1983) and Lehman and Modest (1988) use factor analysis, and Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986; 1988) recommend using principal component analysis.

The second approach is one in which a researcher starts at the estimated covariance matrix of asset returns and uses his judgement to choose factors and subsequently estimate the matrix β. Huberman and Kandel (1985a) note that the correlations of stock returns of firms of different sizes increase with a

similarity in size. Therefore, they choose an index of small firms, one of medium-size firms and one of large firms to serve as factors. In a similar vein, Fama and French (1993) use the spread between the stock returns of small and large firms as one of their factors. Echoing the findings of Rosenberg, Reid and Lanstein (1984), Chan, Hamao and Lakonishok (1991) and Fama and French (1992) observe that expected stock returns and their correlations are also related to the ratio of book-to-market equity. Based on these observations, Fama and French (1993) add the spread between stock returns of value and growth firms as another factor.

The third approach is purely judgemental in that it is one in which the researcher primarily uses his intuition to pick factors and then estimates the factor loadings and checks whether they explain the cross-sectional variations in estimated expected returns (that is, he checks (3)). Chan, Chen and Hsieh (1985) and Chen, Roll and Ross (1986) select financial and macroeconomic variables to serve as factors. They include the following variables: the return on an equity index, the spread of short- and long-term interest rates, a measure of the private sector's default premium, the inflation rate, the growth rates of industrial production and the aggregate consumption. Based on economic intuition, researchers continue to add new factors, which are too many to enumerate here.

The first two approaches are implemented to conform to the factor structure underlying the APT: the first approach by the algorithmic design and the second because researchers check that the factors they use indeed leave the unexplained parts of asset returns almost uncorrelated. The third approach is implemented without regard to the factor structure. Its attempt to relate the assets’ expected returns to the covariance of the assets’ returns with other variables is more in the spirit of Merton's (1973) inter-temporal CAPM than in the spirit of the APT.

The empirical work cited above examines the extent to which the exact APT (with whatever factors are chosen) explains the cross-sectional variation in assets’ mean returns better than the CAPM. It also examines the extent to which other variables – usually those that include various firm characteristics – have marginal explanatory power beyond the factor loadings to explain the cross section of assets’ mean returns. The results usually suggest that the APT is a useful model in comparison with the CAPM. (Otherwise, they would probably have gone unpublished.) However, the results are mixed when the alternative is firm characteristics. Researchers who introduce factors tend to report results supporting the APT with their factors and test portfolios. Nevertheless, different tests and construction of portfolios often reject the proposed APT. For example, Fama and French (1993) demonstrate that exact APT using their factors holds for portfolios constructed by sorting stocks on firm size and book-to-market ratio, whereas Daniel and Titman (1997) demonstrate that the same APT does not hold for portfolios that are constructed by sorting stocks further on the estimated loadings with respect to Fama and French's factors.

The APT often seems to describe the data better than competing models. It is wise to recall, however, that the purported empirical success of the APT may well be due to the weakness of the tests employed. Some questions come to our mind: which factors capture the data best; what is the economic interpretation of the factors; what are the relations among the factors that different researchers have reported? As any test of the APT is a joint test that the factors are correctly identified and that the linear pricing relation holds, a host of competing theories exist side by side under the APT's umbrella. Each fails to reject the APT but has its own factor identification procedure. The number of factors, as well as the methods of factor construction, is exploding. The multiplicity of competing factor models indicates ignorance of the true factor structure of asset returns and suggests a rich and challenging research

agenda.

8 Applications

The APT lends itself to various practical applications due to its simplicity and flexibility. The three areas of applications critically reviewed here are: asset allocation, the computation of the cost of capital, and the performance evaluation of managed funds.

The application of the APT in asset allocation is motivated by the link between the factor structure (1) and mean-variance efficiency. Since the structure with *k* factors implies the existence of *k* assets that span the efficient frontier, an investor can construct a mean-variance efficient portfolio with only *k* assets. The task is especially straightforward when the *k* factors are the payoffs of traded securities. When *k* is a small number, the model reduces the dimension of the optimization problem. The use of the APT in the construction of an optimal portfolio is equivalent to imposing the restriction of the APT in the estimation of the mean and covariance matrix involved in the mean-variance analysis. Such a restriction increases the reliability of the estimates because it reduces the number of unknown parameters.

If the factor structure specified in the APT is incorrect, however, the optimal portfolio constructed from the APT will not be mean-variance efficient. This uncertainty calls for adjusting, rather than restricting, the estimates of mean and covariance matrix by the APT. The degree of this adjustment should depend on investors' prior belief in the model. Pastor and Stambaugh (2000) introduce the Bayesian approach to achieve this adjustment. Wang (2005) further shows that the Bayesian estimation of the return distribution results in a weighted average of the distribution restricted by the APT and the unrestricted distribution matched to the historical data.

The proliferation of APT-based models challenges an investor engaging in asset allocation. In fact, Wang (2005) argues that investors averse to model uncertainty may choose an asset allocation that is not mean-variance efficient for any probability distributions estimated from the prior beliefs in the model.

Being an asset pricing model, the APT should lend itself to the calculation of the cost of capital. Elton, Gruber and Mei (1994) and Bower and Schink (1994) used the APT to derive the cost of capital for electric utilities for the New York State Utility Commission. Elton, Gruber and Mei specify the factors as unanticipated changes in the term structure of interest rates, the level of interest rates, the inflation rate, the GDP growth rate, changes in foreign exchange rates, and a composite measure they devise to measure changes in other macro factors. In the meantime, Bower and Schink use the factors suggested by Fama and French (1993) to calculate the cost of capital for the Utility Commission. However, the Commission did not adopt any of the above-mentioned multi-factor models but used the CAPM instead (see DiValentino, 1994).

Other attempts to apply the APT to compute the cost of capital include Bower, Bower and Logue (1984), Goldenberg and Robin (1991) who use the APT to study the cost of capital for utility stocks, and Antoniou, Garrett and Priestley (1998) who use the APT to calculate the cost of equity capital when examining the impact of the European exchange rate mechanism. Different studies use different factors and consequently obtain different results, a reflection of the main drawback of the APT – the theory does not specify what factors to use. According to Green, Lopez and Wang (2003), this drawback is one

of the main reasons that the US Federal Reserve Board has decided not to use the APT to formulate the imputed cost of equity capital for priced services at Federal Reserve Banks.

The application of asset pricing models to the evaluation of money managers was pioneered by Jensen (1968). When using the APT to evaluate money managers, the managed funds' returns are regressed on the factors, and the intercepts are compared with the returns on benchmark securities such as Treasury bills. Examples of this application of the APT include Busse (1999), Carhart (1997), Chan, Chen and Lakonishok (2002), Cai, Chan and Yamada (1997), Elton, Gruber and Blake (1996), Mitchell and Pulvino (2001), and Pastor and Stambaugh (2002).

The APT is a one-period model that delivers arbitrage-free pricing of existing assets (and portfolios of these assets), given the factor structure of their returns. Applying it to price derivatives on existing assets or to price trading strategies is problematic, because its stochastic discount factor is a random variable which may be negative. Negativity of the SDF in an environment which permits derivatives leads to a pricing contradiction, or arbitrage. Consider, for instance, the price of an option that pays its holder whenever the SDF is negative. Being a limited liability security, such an option should have a positive price, but applying the SDF to its payoff pattern delivers a negative price. (The observation that the stochastic discount factor of the CAPM may be negative is in to Dybvig and Ingersoll, 1982, who also studied some of the implications of this observation.)

Trading and derivatives on existing assets are closely related. Famously, Black and Scholes (1973) show that dynamic trading of existing securities can replicate the payoffs of options on these existing securities. Therefore, one should be careful in interpreting APT-based excess returns of actively managed funds because such funds trade rather than hold on to the same portfolios. Examples of interpretations of asset management techniques as derivative securities include Merton (1981) who argues that market-timing strategy is an option, Fung and Hsieh (2001) who show that hedge funds using trend-following strategies behave like a look-back straddle, and Mitchell and Pulvino (2001) who demonstrate that merger arbitrage funds behave like an uncovered put.

Motivated by the challenge of evaluating dynamic trading strategies, Glosten and Jagannathan (1994) suggest replacing the linear factor models with the Black–Scholes model. Wang and Zhang (2005) study the problem extensively and develop an econometric methodology to identify the problem in factor-based asset pricing models. They show that the APT with many factors is likely to have large pricing errors over actively managed funds, because empirically these models deliver SDFs which allow for arbitrage over derivative-like payoffs.

It is ironic that some of the applications of the APT require extensions of the basic model which violate its basic tenet – that assets are priced as if markets offer no arbitrage opportunities.

See Also

- arbitrage
- capital asset pricing model
- factor models

The views stated here are those of the authors and do not necessarily reflect the views of the Federal

Reserve Bank of New York or the Federal Reserve System.

Bibliography

Antoniou, A., Garrett, I. and Priestley, R. 1998. Calculating the equity cost of capital using the APT: the impact of the ERM. *Journal of International Money and Finance* 14, 949–65.

Black, F., Jensen, M. and Scholes, M. 1972. The capital-asset pricing model: some empirical tests. In *Studies in the Theory of Capital Markets*, ed. M. Jensen. Praeger Publishers: New York.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–654.

Bower, D., Bower, R. and Logue, D. 1984. Arbitrage pricing and utility stock returns. *Journal of Finance* 39, 1041–54.

Bower, R. and Schink, G. 1994. Application of the Fama–French model to utility stocks. *Financial Markets, Institutions and Instruments* 3, 74–96.

Busse, J. 1999. Volatility timing in mutual funds: evidence from daily returns. *Review of Financial Studies* 12, 1009–41.

Cai, J., Chan, K. and Yamada, T. 1997. The performance of Japanese mutual funds. *Review of Financial Studies* 10, 237–73.

Campbell, J., Lo, A. and MacKinley, C. 1997. *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

Carhart, M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52, 57–82.

Chamberlain, G. 1983. Funds, factors and diversification in arbitrage pricing models. *Econometrica* 51, 1305–23.

Chamberlain, G. and Rothschild, M. 1983. Arbitrage, factor structure, and mean variance analysis on large asset markets. *Econometrica* 51, 1281–304.

Chan, K., Chen, N. and Hsieh, D. 1985. An exploratory investigation of the firm size effect. *Journal of Financial Economics* 14, 451–71.

Chan, L., Chen, H. and Lakonishok, J. 2002. On mutual fund investment styles. *Review of Financial Studies* 15, 1407–37.

Chan, L., Hamao, Y. and Lakonishok, J. 1991. Fundamentals and stock returns in Japan. *Journal of Finance* 46, 1739–64.

Chen, N. 1983. Some empirical tests of the theory of arbitrage pricing. *Journal of Finance* 38, 1393–414.

Chen, N. and Ingersoll, J. 1983. Exact pricing in linear factor models with infinitely many assets: a note. *Journal of Finance* 38, 985–8.

Chen, N., Roll, R. and Ross, S. 1986. Economic forces and the stock markets. *Journal of Business* 59, 383–403.

Connor, G. 1984. A unified beta pricing theory. *Journal of Economic Theory* 34, 13–31.

Connor, G. and Korajczyk, R. 1986. Performance measurement with the arbitrage pricing theory: a framework for analysis. *Journal of Financial Economics* 15, 373–94.

Connor, G. and Korajczyk, R. 1988. Risk and return in an equilibrium APT: application of a new test methodology. *Journal of Financial Economics* 21, 213–54.

Daniel, K. and Titman, S. 1997. Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance* 52, 1–33.

DiValentino, L. 1994. Preface. *Financial Markets, Institutions and Instruments* 3, 6–8.

Dybvig, P. 1983. An explicit bound on deviations from APT pricing in a finite economy. *Journal of Financial Economics* 12, 483–96.

Dybvig, P. and Ingersoll, J. 1982. Mean-variance theory in complete markets. *Journal of Business* 55, 233–51.

Dybvig, P. and Ross, S. 1985. Yes, the APT is testable. *Journal of Finance* 40, 1173–88.

Elton, E., Gruber, M. and Blake, C. 1996. Survivorship bias and mutual fund performance. *Review of Financial Studies* 9, 1097–120.

Elton, E., Gruber, M. and Mei, J. 1994. Cost of capital using arbitrage pricing theory: a case study of nine New York utilities. *Financial Markets, Institutions and Instruments* 3, 46–73.

Fama, E. and French, K. 1992. The cross-section of expected stock returns. *Journal of Finance* 47, 427–86.

Fama, E. and French, K. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 3–56.

Fama, E. and MacBeth, J. 1973. Risk, return and equilibrium: empirical tests. *Journal of Political Economy* 71, 607–36.

Fung, W. and Hsieh, D. 2001. The risks in hedge fund strategies: theory and evidence from trend followers. *Review of Financial Studies* 14, 313–41.

Glosten, L. and Jagannathan, R. 1994. A contingent claim approach to performance evaluation, *Journal of Empirical Finance* 1, 133–60.

Goldenberg, G. and Robin, A. 1991. The arbitrage pricing theory and cost-of-capital estimation: the case of electric utilities. *Journal of Financial Research* 14, 181–96.

Green, E., Lopez, J. and Wang, Z. 2003. Formulating the imputed cost of equity for priced services at Federal Reserve Banks. *Economic Policy Review* 9, 55–8.

Grinblatt, M. and Titman, S. 1983. Factor pricing in a finite economy. *Journal of Financial Economics* 12, 495–507.

Grinblatt, M. and Titman, S. 1987. The relation between mean-variance efficiency and arbitrage pricing. *Journal of Business* 60, 97–112.

Hansen, L. and Jagannathan, R. 1991. Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99, 225–262.

Hansen, L. and Jagannathan, R. 1997. Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52, 557–590.

Huberman, G. 1982. A simple approach to arbitrage pricing. *Journal of Economic Theory* 28, 183–91.

Huberman, G. and Kandel, S. 1985a. A size based stock returns model. Center for Research in Security Prices Working Paper 148, University of Chicago.

Huberman, G. and Kandel, S. 1985b. Likelihood ratio tests of asset pricing and mutual fund separation. Center for Research in Security Prices Working Paper 149, University of Chicago.

Huberman, G., Kandel, S. and Stambaugh, R. 1987. Mimicking portfolios and exact arbitrage pricing. *Journal of Finance* 42, 1–9.

Ingersoll, J. 1984. Some results in the theory of arbitrage pricing. *Journal of Finance* 39, 1021–1039.

Jagannathan, R., Skoulakis, G. and Wang, Z. 2002. Generalized method of moments: applications in finance. *Journal of Business and Economic Statistics* 20, 470–481.

Jagannathan, R. and Wang, Z. 1998. An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *Journal of Finance* 53, 1285–309.

Jagannathan, R. and Wang, Z. 2002. Empirical evaluation of asset pricing models: a comparison of the SDF and beta methods. *Journal of Finance* 57, 2337–67.

Jensen, M. 1968. The performance of mutual funds in the period 1945–1964. *Journal of Finance* 23, 389–416.

Jobson, J. 1982. A multivariate linear regression test of the arbitrage pricing theory. *Journal of Finance* 37, 1037–42.

Jobson, J. and Korkie, B. 1982. Potential performance and tests of portfolio efficiency. *Journal of Financial Economics* 10, 433–66.

Jobson, J. and Korkie, B. 1985. Some tests of linear asset pricing with multivariate normality. *Canadian Journal of Administrative Sciences* 2, 114–38.

Kan, R. and Zhou, G. 2001. Tests of mean-variance spanning. Working paper, Washington University in St Louis.

Lehman, B. and Modest, D. 1988. The empirical foundations of the arbitrage pricing theory. *Journal of Financial Economics* 21, 213–54.

Merton, R. 1973. An intertemporal capital asset pricing model. *Econometrica* 41, 867–87.

Merton, R. 1981. On market timing and investment performance I: any equilibrium theory of values for markets forecasts. *Journal of Business* 54, 363–406.

Mitchell, M. and Pulvino, T. 2001. Characteristics of risk and return in risk arbitrage, *Journal of Finance* 56, 2135–75.

Pastor, L. and Stambaugh, R. 2000. Comparing asset pricing models: an investment perspective. *Journal of Financial Economics* 56, 335–81.

Pastor, L. and Stambaugh, R. 2002. Mutual fund performance and seemingly unrelated assets. *Journal of*

Financial Economics 63, 315–49.

Peñaranda, F. and Sentana, E. 2004. Spanning tests in return and stochastic discount factor mean-variance frontiers: a unifying approach. Working paper No. 0410. Madrid: CEMFI.

Roll, R. and Ross, S. 1980. An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35, 1073–103.

Rosenberg, B., Reid, K. and Lanstein, R. 1984. Persuasive evidence of market inefficiency. *Journal of Portfolio Management* 11, 9–17.

Ross, S. 1976a. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13, 341–60.

Ross, S. 1976b. Risk, return and arbitrage. In *Risk Return in Finance*, ed. I. Friend and J. Bicksler. Cambridge, MA: Ballinger.

Shanken, J. 1982. The arbitrage pricing theory: is it testable? *Journal of Finance* 37, 1129–240.

Shanken, J. 1985. A multi-beta CAPM or equilibrium APT?: A reply. *Journal of Finance* 40, 1189–96.

Shanken, J. 1992. On the estimation of beta-pricing models. *Review of Financial Studies* 5, 1–33.

Stambaugh, R. 1983. Arbitrage pricing with information. *Journal of Financial Economics* 12, 357–69.

Wang, Z. 2005. A shrinkage approach to model uncertainty and asset allocation. *Review of Financial Studies* 18, 673–705.

Wang, Z. and Zhang, X. 2005. Empirical evaluation of asset pricing models: arbitrage and pricing errors over contingent claims. Working paper. Federal Reserve Bank of New York.

How to cite this article

Huberman, Gur and Zhenyu Wang. "arbitrage pricing theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000119>
doi:10.1057/9780230226203.0053

The New Palgrave Dictionary of Economics Online

arbitrage

Philip H. Dybvig and Stephen A. Ross
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The absence of arbitrage is the unifying concept for much of finance. Absence of arbitrage is more general than equilibrium because it does not require all agents to be rational. The Fundamental Theorem of Asset Pricing asserts the equivalence of absence of arbitrage, existence of a positive linear pricing rule, and existence of some hypothetical agent who prefers more to less and has an optimum. Equivalent representations of the pricing rule are the martingale measure (risk-neutral pricing), and a positive state price density. Applications of no arbitrage and these representations include Modigliani–Miller theory, option pricing, investments, and forward exchange parity.

Keywords

arbitrage; Arrow–Debreu model; arbitrage pricing theory; capital asset pricing model; capital budgeting; capital structure; dividend discount model; dividend policy; dominance; duality; efficient allocation; efficient market hypothesis; equivalent martingale measure; Farkas' Lemma; forward exchange, parity theory of; fundamental theorem of asset pricing; Hahn–Banach th; hyperplanes; interest rates; law of one price; linear pricing rules; linear programming; Markov processes; martingales; Modigliani–Miller th; no arbitrage; noise; option; option pricing; pricing rule representation th); purchasing power parity; risk premium; risk-neutral probabilities; separation theorems; state price density; state spaces; trade costs; von Neumann and Morgenstern

Article

An arbitrage opportunity is an investment strategy that guarantees a positive payoff in some contingency with no possibility of a negative payoff and with no net investment. By assumption, it is possible to run the arbitrage possibility at arbitrary scale; in other words, an arbitrage opportunity represents a money pump. A simple example of arbitrage is the opportunity to borrow and lend costlessly at two different fixed rates of interest. Such a disparity between the two rates cannot persist: arbitrageurs will drive the rate together.

The modern study of arbitrage is the study of the implications of assuming that no arbitrage opportunities are available. Assuming no arbitrage is compelling because the presence of arbitrage is inconsistent with equilibrium when preferences increase with quantity. More fundamentally, the presence of arbitrage is inconsistent with the existence of an optimal portfolio strategy for any competitive agent who prefers more to less, because there is no limit to the scale at which an individual would want to hold the arbitrage position. Therefore, in principle, absence of arbitrage follows from individual rationality of a single agent. One appeal of results based on the absence of arbitrage is the intuition that absence of arbitrage is more primitive than equilibrium, since only relatively few rational agents are needed to bid away arbitrage opportunities, even in the presence of a sea of agents driven by 'animal spirits'.

The absence of arbitrage is very similar to the zero economic profit condition for a firm with constant returns to scale (and no fixed factors). If such a firm had an activity which yielded positive profits, there would be no limit to the scale at which the firm would want to run the activity, and no optimum would exist. The theoretical distinction between a zero profit condition and the absence of arbitrage is the distinction between commerce, which requires production, and trading under the price system, which does not. In practice, the distinction blurs. For example, if gold is sold at different prices in two markets, there is an arbitrage opportunity but it requires production (transportation of the gold) to take advantage of the opportunity. Furthermore, there are almost always costs to trading in markets (for example, brokerage fees), and therefore a form of costly production is required to convert cash into a security. For the purposes of this article, we will tend to ignore production. In practical applications the necessity of production will weaken the implications of absence of arbitrage and may drive a wedge between what the pure absence of arbitrage would predict and what actually occurs.

The assertion that two perfect substitutes (for example, two shares of stock in the same company) must trade at the same price is an implication of no arbitrage that goes under the name of the law of one price. While the law of one price is an immediate consequence of the absence of arbitrage, it is not equivalent to the absence of arbitrage. An early use of a no-arbitrage condition employed the law of one price to help explain the pattern of prices in the foreign exchange and commodities markets.

Many economic arguments use the absence of arbitrage implicitly. In discussions of purchasing power parity in international trade, for

example, presumably it is an arbitrage possibility that forces the spot exchange rate between currencies to equal the relative prices of common baskets of (traded) goods. Similarly, the statement that the possibility of repackaging implies linear prices in competitive product markets is essentially a no-arbitrage argument.

Early uses of the law of one price

The parity theory of forward exchange based on the law of one price was first formulated by Keynes (1923) and developed further by Einzig (1937). Let *s* denote the current spot price of, say, euros, in terms of dollars, and let *f* denote the forward price of euros one year in the future. The forward price is the price at which agreements can be struck currently for the future delivery of euros with no money changing hands today. Also, let r_s and r_m denote the one year dollar and euro interest rates, respectively. To prevent an arbitrage possibility from developing, these four prices must stand in a particular relation.

To see this, consider the choices facing a holder of dollars. The holder can lend the dollars in the domestic market and realize a return of r_s one year from now. Alternatively, the investor can purchase euros on the spot market, lend for one year in the German market, and convert the euros back into dollars one year from now at the fixed forward rate. By undertaking the conversion back into dollars in the forward market, the investor locks in the prevailing forward rate, *f*. The results of this latter path are a return of

$$f(1 + r_m) / s$$

dollars one year from now. If this exceeds $1+r_s$, then the foreign route offers a sure higher return than domestic lending. By borrowing dollars at the domestic rate r_s and lending them in the foreign market, a sure profit at the rate

$$f(1 + r_m) / s - (1 + r_s)$$

can be made with no net investment of funds. Alternatively, if

$$f(1 + r_m) / s - (1 + r_s) < 0,$$

the arbitrage works in reverse. By borrowing in euros, investing in dollars, and buying euros forward, a sure profit at the rate

$$(1 + r_s) - f(1 + r_m) / s$$

can be made with no investment in funds.

Thus, the prevention of arbitrage will enforce the forward parity result,

$$(1 + r_s) / (1 + r_m) = f / s.$$

This result takes on many different forms as we look across different markets. In a commodity market with costless storage, for example, an arbitrage opportunity will arise if the following relation does not hold:

$$f \leq s(1 + r).$$

In this equation, f is the currently quoted forward rate for the purchase of the commodity – for example, silver, one year from now – s is the current spot price, and r is the interest rate. More generally, if c is the up-front proportional carrying cost, including such items as storage costs, spoilage and insurance, absence of arbitrage ensures that

$$f \leq s(1 + c)(1 + r).$$

(We normally would expect these relations to hold with equality in a market in which positive stocks are held at all points in time, and perhaps with inequality in a market which may not have positive stocks just before a harvest. However, proving equality is based on equilibrium arguments, not on the absence of arbitrage, since to short the physical commodity one must first own a positive amount.)
The above applications of the absence of arbitrage (via the law of one price) share the common characteristic of the absence of risk. The law of one price is less restrictive than the absence of arbitrage because it deals only with the case in which two assets are identical but have different prices. It does not cover cases in which one asset dominates another but may do so by different amounts in different states. The most interesting applications of the absence of arbitrage are to be found in uncertain situations, where this distinction may be important.

The fundamental theorem of asset pricing

The absence of arbitrage is implied by the existence of an optimum for any agent who prefers more to less. The most important implication of the absence of arbitrage is the existence of a positive linear pricing rule, which in many spaces including finite state spaces is the same as the existence of positive state prices that correctly price all assets. Taken together with their converses, we refer collectively to these results as the *Fundamental Theorem of Asset Pricing*. Traditionally, the emphasis has been on the linear pricing rule as an implication of the absence of arbitrage. Including the existence of an optimum (introduced in the version of this article in the first edition of *The New Palgrave*) is useful both because it reminds us why we are interested in arbitrage, and because the converse tells us that, absent other restrictions, consistency with equilibrium is equivalent to the absence of arbitrage. We state the theorem verbally here; the formal meanings of the words and the proof are given later in this section.
Theorem: (*Fundamental Theorem of Asset Pricing*) The following are equivalent:

1. (i) absence of arbitrage;
2. (ii) existence of a positive linear pricing rule;
3. (iii) existence of an optimal demand for some agent who prefers more to less.

Beja (1971) was one of the first to emphasize explicitly the linearity of the asset pricing function, but he did not link it to the absence of arbitrage. Beja simply assumed that equilibrium prices existed and observed ‘that equilibrium properties require that the functional q be linear’, where q is a functional that assigns a price or value to a risky cash flow. The first statement and proof that the absence of arbitrage implied the existence of non-negative state space prices and, more generally, of a positive linear operator that could be used to value risky assets appeared in Ross (1976a; 1978). Besides providing a formal analysis, Ross showed that there was a pricing rule that prices *all* assets and not just those actually marketed. (In other words, the linear pricing rule could be extended from the marketed assets to all hypothetical assets defined over the same set of states.) The advantage of this extension is that the domain of the pricing function does not depend on the set of marketed assets. We will largely follow Ross's analysis with some modern improvements.
Linearity for pricing means that the price functional or operator q satisfies the ordinary linear condition of algebra. If we let x and y be two random payoffs and we let q be the operator that assigns values to prospects, then we require that

$$q(ax + by) = aq(x) + bq(y),$$

where a and b are arbitrary constants. Of course, for many spaces (including a finite state space), any linear functional can be represented as a sum or integral across states of state prices times quantities.
To simplify proofs in this article, we will make the assumption that there are finitely many states, each of which occurs with positive probability, and that all claims purchased today pay off at a single future date. Let Θ denote the state space,

$$\Theta = \{1, \ldots, m\},$$

where there are *m* states and the state of nature θ occurs with probability π_θ. Applying *q* to the 'indicator' asset e_θ whose payoff is 1 in state θ and 0 otherwise, we can define a price q_θ for each state θ as the value of e_θ;

$$q_\theta = q(e_\theta).$$

Now, if there were linearity, the value of any payoff, *x*, could be written as

$$q(x) = \sum_\theta q_\theta x_\theta.$$

Of course, this argument presupposes that $q(e_\theta)$ is well defined, which is a strong assumption if e_θ is not marketed.
We want to make a statement about the conditions under which all marketed assets can be priced by such a linear pricing rule *q*. We assume that there is a set of *n* marketed assets with a corresponding price vector, *p*. Asset *i* has a terminal payoff $X_{\theta i}$ (inclusive of dividends, and so on) in state of nature θ. The matrix $X \equiv [X_{\theta i}]$ denotes the state space tableau whose columns correspond to assets and whose rows correspond to states. Lower-case *x* represents the random vector of terminal payoffs to the various securities. An arbitrage opportunity is a portfolio (vector) η with two properties. It does not cost anything today or in a state in the future. And, it has a positive payoff either today or in some state in the future (or both). We can express the first property as a pair of vector inequalities. The initial cost is not greater than zero, which is to say that it uses no wealth and may actually generate some,

$$p\eta \leq 0, \tag{1}$$

and its random payoff later is never negative,

$$X\eta \geq 0. \tag{2}$$

(We use the notation that $\geq$ denotes greater or equal in each component, > denotes $\geq$ and greater in some component, and • denotes greater in all components. Note that writing the price of $X\eta$ as $p\eta$ for arbitrary η embodies an assumption that investment in marketed assets is divisible.) The second property says that the arbitrage portfolio η has a strict inequality, either in (1) or in some component of (2). We can express both properties together as

$$X_*\eta \equiv \begin{bmatrix} -p \\ X \end{bmatrix} \eta > 0. \tag{3}$$

Here, we have stacked the net payoff today on top of the vector of payoffs at the future date. This is in the spirit of the Arrow–Debreu model in which consumption in different states, commodities, points of time and so forth, are all considered components of one large consumption vector.
The absence of arbitrage is simply the condition that no η satisfies (3). A consistent positive linear pricing rule is a vector of state prices *q*•0 that correctly prices all marketed assets, that is, such that

$$p = qX.$$

(4)

We have now collected enough definitions to prove the first half (that (i) ⇔ (ii)) of the Fundamental Theorem of Asset Pricing.
Theorem: (*first half of the Fundamental Theorem of Asset Pricing*) There is no arbitrage if and only if there exists a consistent positive linear pricing rule.
Proof: The proof that having a consistent positive linear pricing rule precludes arbitrage is simple, since any arbitrage opportunity gives a direct violation of (4). Let η be an arbitrage opportunity. By (4),

$$p\eta = qX\eta,$$

or equivalently

$$0 = -p\eta + q(X\eta) = [1q]X*\eta.$$

By definition of an arbitrage opportunity (3) and positivity of *q*, we have a contradiction.
The proof that the absence of arbitrage implies the existence of a consistent positive linear pricing rule is more subtle and requires a separation theorem. The mathematical problem is equivalent to Farkas' Lemma of the alternative and to the basic duality theorem of linear programming. We will adopt an approach that is analogous to the proof of the second theorem of welfare economics that asserts the existence of a price vector which supports any efficient allocation, by separating the aggregate Pareto optimal allocation from all aggregate allocations corresponding to Pareto preferable allocations. Here we will find a price vector that 'supports' an arbitrage-free allocation by separating the net trades from the set of free lunches (the positive orthant).
The absence of arbitrage is equivalent to the requirement that the linear space of net trades defined by

$$s \equiv \{y | \text{for some } \eta,\ y = X*\eta\},$$
(5)

does not intersect the positive orthant $\Re_+^{m+1} \equiv \{y | y \geq 0\}$ except at the origin, that is $S \cap \Re_+^{m+1} = \{0\}$.
Since *S* is a subspace (and is therefore a convex closed cone), a simple separation theorem (Karlin, 1959, Theorem B3.5) implies that there exists a nonzero vector q_* such that for all $y \in S$ and all $z \in \Re_+^{m+1}$, $z \neq 0$, we must have

$$q*z > 0 \geq q*y.$$
(6)

Letting *z* be each of the unit vectors in turn, the first inequality in (6) implies that $q*$ is a strictly positive vector.
Since *S* is a subspace, the second inequality in (6) must hold with equality for all $y \in S$ Define

$$q \equiv (q*_2, q*_3, \ldots, q*_n) / q*_1.$$

Since q_*•0, likewise *q*•0.
Dividing the second equality in (6) (which we now know to be an equality) by $q*_1$ and expanding using the definition of X_* [from (3)], we

have that

$$0 = -p + qX,$$

or

$$p = qX,$$

which shows that q is a consistent positive linear pricing rule.

Before we can prove the second half of the pricing theorem, we need to define the maximization problem faced by a typical investor. In this problem, all we really need to assume is that more is preferred (strictly) to less, that is, that increasing initial consumption or random consumption later in one or more states always leads to a preferred outcome. In fact, this is literally all we need: we do not need completeness or even transitivity of preferences, let alone a utility function representation or any restriction to a functional form. However, for concreteness, we will write down preferences using a state-dependent utility function of consumption now and in the future. The assumption that the investor prefers more to less is satisfied if the utility function in each state is increasing in consumption at both dates.

The state-dependent restriction implies that the maximization problem faced by a particular agent is the maximization of the expectation of the state-dependent utility function $u_0(\cdot,\cdot)$ of initial wealth and terminal wealth, given initial wealth w_0 and the possibility of trading in the security market. Then the maximization problem faced by a typical agent is the unconstrained choice of a vector α of portfolio weights to maximize

$$\sum_0 \pi_0 u_0[w_0 - p\alpha, (X\alpha)_0].$$

The quantity $p\alpha$ is the price of the portfolio, and therefore $w_0 - p\alpha$ is the residual amount of the initial wealth available for initial consumption. The preferences of the agent are said to be increasing if each $u_0(\cdot, \cdot)$ is (strictly) increasing in both arguments. Saying the agent prefers more to less is just another way of saying that preferences are increasing.

Here is the rest of the proof of the Fundamental Theorem of Asset Pricing.

Theorem: (*second half of the Fundamental Theorem of Asset Pricing*) There is no arbitrage if and only if there exists some (at least hypothetical) agent with increasing preferences whose choice problem has a maximum.

Proof: If there is an arbitrage opportunity, η, then clearly the choice problem for an agent with increasing preferences cannot have a maximum, since for every α,

$$\sum_\theta \pi_\theta u_\theta\{w_0 - p(\alpha + k\eta), [X(\alpha + k\eta)]_\theta\}$$

increases as k increases.

Conversely, if there is no arbitrage, by the first half of the Fundamental Theorem of Asset Pricing (proven earlier), there exist a consistent positive linear pricing rule q. Let $w_0=0$ and $\alpha=0$. Consider the particular utility function

$$u_{*\theta}(c_0, c_1) \equiv -\exp[-(c_0 - w_0)] - (q_\theta / \pi_\theta)\exp(-c_1). \tag{7}$$

Each function $u_{*\theta}$ is strictly increasing and also happens to be strictly concave, infinitely differentiable, and additively separable over time. Using $p=qX$, it is easy to show that this utility function satisfies the first-order conditions for a maximum, which are necessary and sufficient

by concavity. (Note: by a more complicated argument, it can be shown that the von Neumann–Morgenstern ‘state independent’ utility function $-\exp(-c_0) - \exp(-c_1)$ has a maximum, but the maximum will not necessarily be achieved at $\alpha = 0$).

As should be clear from the proof, it is not really important what class of preference we use, so long as all agents having preferences in the class prefer more to less and the class includes the particular preferences used in the proof (which are additive over states and time, increasing, concave, and infinitely differentiable).

Recent research on arbitrage, starting with Ross (1978) and Harrison and Kreps (1979), has focused on extending these results to more general state spaces in which there are many time periods and, more importantly, infinitely many states. In these spaces, deriving a positive linear pricing rule for marketed claims is still straightforward (one can prove the algebraic linearity condition and positivity directly from the no-arbitrage condition), but extending the pricing rule from the priced claims to all non-marketed claims requires some sort of extension theorem, such as a Hahn–Banach theorem. Obtaining a truly general result is complicated by the fact that the positive orthant is not typically an open set in these general spaces, and openness is a condition of the Hahn–Banach theorems. One part of the result that goes through in general is the implication that existence of an optimum implies existence of a linear pricing rule: so long as preferences are continuous in our topology, the preferred set will be open, and the linear pricing rule will be a hyperplane that separates the optimum from the preferred set.

Alternative representations of linear pricing rules

There are many equivalent ways of representing a linear pricing rule. Which representation is simplest depends on the context. In one representation, the price is the expected value under artificial ‘risk-neutral’ probabilities discounted at the riskless rate. (The risk-neutral probability measure is also referred to as an equivalent martingale measure.) In another representation, the price is the expectation of the quantity times the state price density, which is the state price per unit probability. In yet another representation, the price is the expected value discounted at a risk-adjusted rate. The purpose of this section is to show the fundamental equivalence of these representations.

The motive for using a particular representation is usually found in the study of intertemporal models or models with a continuum of states. Nonetheless, we will continue our formal analysis of the single-period model with finitely many states, leaving the more general discussion of the merits of the various approaches until afterwards. Now, we have already seen the basic linear pricing rule representation. For any portfolio α,

$$p\alpha = qX\alpha = \sum_{\theta} q_{\theta}(X\alpha)_{\theta}, \tag{8}$$

that is, the sum across states of state price times the payoff.

The risk-neutral or martingale representation asserts the existence of a vector Π of artificial probabilities and a shadow riskless rate r such that

$$p\alpha = (1+r)^{-1}\Pi X\alpha = (1+r)^{-1}E_{\Pi}(x\alpha), \tag{9}$$

that is, the expectation E_{Π} of the payoff under the risk-neutral (martingale) probabilities Π, discounted at the riskless rate. It is easy to see the shadow riskless rate is equal to the riskless rate if one exists. The risk neutral approach is trivially equivalent to the positive linear pricing rule approach. Simply let

$$\Pi = q / \sum_{\theta} q_{\theta} \tag{10}$$

and

$$(1+r)^{-1} = \sum_{\theta} q_{\theta} \tag{11}$$

For the converse, let

$$q = (1 + r)^{-1} = \sum_{\theta} q_{\theta}$$
(12)

Therefore, the existence of a positive linear pricing rule is the same as the existence of positive risk-neutral probabilities. (The risk-neutral measure is equivalent to the original probability measure, that is, Π has the same null sets as α. Here, that is simply the requirement that the list of states with positive probability is the same for both measures.)
A third approach emphasizes the role of the state price density, ρ_0. In this case, the price is given by

$$p\alpha = \sum_{\theta} \pi_{\theta}\rho_{\theta}(X\alpha)_{\theta} = E(\rho x\alpha).$$
(13)

To see that this is equivalent to the linear pricing rule, simply let

$$\rho_{\theta} = q_{\theta} / \pi_{\theta},$$
(14)

or, conversely, let

$$q_{\theta} = \rho_{\theta}\pi_{\theta}.$$
(15)

Clearly, p is positive in all states if and only if q is.
We have shown the equivalence of these three approaches. This equivalence is stated in the following theorem.
Theorem: (*Pricing Rule Representation Theorem*) The following are equivalent:

- existence of a positive linear pricing rule;
- existence of positive risk-neutral probabilities and an associated riskless rate (the martingale property);
- existence of a positive state price density.

The remaining representation is that the value is equal to the terminal value discounted at a risk-adjusted interest rate r_a.

$$p\alpha = (1 + r_a)^{-1}E(x\alpha)$$
(16)

While this might at first appear to be inconsistent with the other representations, the risk-adjusted rate r_a is typically proportional to the covariance of return ($=x\alpha/p\alpha$) with some random variable, and consequently solving this equation for px yields a linear rule. (See Beja, 1971, and Rubinstein, 1976, for general results concerning pricing rules using covariances.) For example, in the capital asset pricing model,

$$r_a = r + \lambda \mathrm{cov}(x\alpha / p\alpha, r_m), \tag{17}$$

where r_m is the random return on the market and λ is the market price of risk. Solving these two equations for *px*, we obtain

$$p\alpha = (1 + r)^{-1} E[x\alpha\{1 - \lambda[r_m - E(r_m)]\}], \tag{18}$$

which is certainly linear in $x\alpha$. The subtle question is whether or not this is positive, and this hinges on whether the market return can get larger than $E(r_m) + 1/\lambda$ (Dybvig and Ingersoll, 1982). In any case, the important observation is that the basic form of the representation is linear even if verification of positivity depends on the exact form of the risk premium.

Now we return to the question of the comparative advantages of the various representations. The risk-neutral or martingale representation was first employed by Cox and Ross (1976a) for use in option pricing problems and was later developed more formally by Harrison and Kreps (1979) and a number of others. The risk-neutral representation is particularly useful for problems of valuation or optimization without reference to individual preferences, since under the martingale probabilities we can ignore risk altogether and maximize discounted expected value. In fact, for some problems this approach tells us that risk-neutral results generalize immediately to worlds where risk is priced. However, this approach tends to be complicated when preferences are introduced, since von Neumann–Morgenstern (state independent) preferences under ordinary probabilities become state dependent under the martingale probabilities. As an aside, we note that, in intertemporal contexts in which the interest rate is stochastic, the price is the risk-neutral expectation of the future value discounted by the rolled-over spot rate (which is stochastic).

The state price density representation (Cox and Leland, 2000; Dybvig, 1980; 1988) is most useful when we want to look at choice problems. Samuelson (1947) emphasized the value of deriving equilibrium conditions from first- and second-order conditions for optimization. In asset pricing the first-order condition for an agent with von Neumann–Morgenstern preferences is that the agent's marginal utility of consumption is proportional to a consistent state price density (not necessarily unique) for the security market (Dybvig and Ross, 1982). (Note that if there is a non-atomic continuum of states, the state price density will typically be well-defined even though all primitive states have probability zero and state price zero.) For the CAPM, this fact was used implicitly by Sharpe (1964) and Lintner (1965), and was made explicit by Dybvig and Ingersoll (1983).

The representation of discounting expected returns using a risk-adjusted discount rate is most useful when we can get some independent assessment of the risk premium involved. Otherwise, it is needlessly complicated, since the price appears not only on the left-hand side of the equation but also in the denominator on the right-hand side. Discounting using a risk-adjusted rate is usually the method of choice for capital budgeting, since the risk adjustment is usually determined from comparables (for example, from past returns on assets in similar firms). For capital budgeting, there may also be a pedagogical advantage that (so far) it has been easier to communicate to practitioners than the other methods. Furthermore, focusing on the risk-adjusted discount rate sharpens the comparison of competing approaches (such as the capital asset pricing model and the dividend discount model).

It is useful to note how the various representations evolve over time. State prices are simply the product of state prices over sub-periods. For example, for $t<s<T$, the state price of a state at T given the state at t is equal to the state price of the state at T given the state at s times the state price of the state at s given the state at t. (The state at s is determined by the state at T given the pervasive assumption of perfect recall, that is, the assumption that the family of sigma-algebras is increasing. If we use some reduced specification of the state – as when looking at Markov processes – the state price is the product of the two, summed over all possible intermediate states.)

The martingale representation yields a price equal to the expected value under the martingale measure of the product of the terminal value times a discount factor that corresponds to rolling over shortest maturity default-free bonds. This representation makes particularly clear the interaction between term structure effects and other effects. If there is a significant term structure, the discount factor is random, and we cannot ignore the interplay between term structure risk and random terminal value unless the terminal value of the asset under consideration is independent of interest rates (under the martingale measure). If the terminal value is independent of interest rate movements, then the value of the asset today is the risk-neutral expected terminal value of the asset discounted at the riskless discount factor (which equals the risk-neutral expected discount factor from rolling over shorts).

The state price density has an evolution over time similar to that of the state price, namely, the state price density over a long interval is the product of the state price density over short intervals. Since the state price density equals the state price divided by the probability, the ratio of the two evolutions gives us a relation involving only probabilities, which is Bayes' law.

Finally, the discounted expected value approach is more complicated than the others. The exact evolution over time depends on whether uncertainty is multiplicative, linear, a distributed lag, or whatever. This difficulty is usually overlooked in capital budgeting applications,

which is probably not so bad in practice, given the imprecision of our estimates of risk premia and future cash flows.

Modern results based on the absence of arbitrage

Most of modern finance is based on either the intuitive or the actual theory of the absence of arbitrage. In fact, it is possible to view absence of arbitrage as the one concept that unifies all of finance (Ross, 1978). In this section, we will try to provide a sample of how arbitrage arguments are used in diverse areas in finance. We will touch on applications in option pricing, corporate finance, asset pricing and efficient markets.

The efficient market hypothesis says that the price of an asset should fully reflect all available information. The intuition behind this hypothesis is that, if the price does not fully reflect available information, then there is a profit opportunity available from buying the asset if the asset is underpriced or from selling it if it is overpriced. Clearly this is consistent with the intuition of the absence of arbitrage, even if what we have here is only an approximate arbitrage possibility, that is, a large profit at little risk. Approximate arbitrage is always profitable to a risk-neutral investor. More generally, the issue is clouded somewhat by qsts of risk tolerance and what is the appropriate risk premium. Happily, empirical violation of efficiency of the market (for example, in event studies) is not significantly affected by the procedure for measuring the risk premium (Brown and Warner, 1980; 1985). Therefore, an empirical violation of efficiency is an approximate arbitrage opportunity that presumably would be attractive at large scale to many investors.

The Modigliani–Miller propositions tell us that, in perfect capital markets, changing capital structure or dividend policy without changing investment is a matter of irrelevance to the shareholders. The original proofs of the Modigliani–Miller propositions used the law of one price and assumed the presence of a perfect substitute for the firm that was altering its capital structure. As an illustration of the Fundamental Theorem of Asset Pricing, Ross (1978) demonstrated that these propositions could be derived directly from the existence of a positive linear pricing rule.

To illustrate this argument, consider the proposition that the total value of the firm does not depend on the capital structure. The original argument assumed that there is another identical firm. If we change the financing of our firm, then the value of holding a portfolio of all the parts will give a final payoff equal to that of the identical firm, and must therefore have the same value under the law of one price.

Alternatively, suppose that there exists a positive linear pricing rule q. Let x represent the total terminal value of a firm in a one-period model and x_i the payoff to financial claim i on the assets of the firm. Then the sum of all the payoffs must add up to the total terminal value.

$$x = \sum_i x_i \tag{19}$$

Using the positive linear operator, q, which values assets, we have that the value of the firm,

$$v \equiv \Sigma_i q(x_i) = q(\Sigma_i x_i) = q(x_i), \tag{20}$$

which is independent of the number of structure of the financial claims.

Note that both proofs make an implicit assumption that goes beyond what absence of arbitrage promises, namely, that changing the capital structure of the firm does not change the way in which prices are formed in the economy. In the original proof this is the assumption that the other firm's price will not change when the firm changes its capital structure. In the linear pricing rule proof this is the assumption that the state price vector q does not change.

Another application of the absence of arbitrage is to asset pricing. The most obvious application is the derivation of the arbitrage pricing theory (Ross, 1976a; 1976b). We will consider the special case without asset-specific noise. Assume that the mechanism generating the per dollar investment rates of return for a set of assets is given by

$$R_i = E_i + \beta_{i1} f_1 + \ldots + \beta_{ik} f_k, \; i = 1, \ldots, n \tag{21}$$

where E_i is the expected rate of return on asset i per dollar invested and f_i is an exogenous factor. This form is an exact factor generating mechanism (as opposed to an approximate one with an additional asset specific mean zero term).

Applying the pricing operator, q, to equation (21) we have that

$$1 = q(1 + R_i) = q(1 + E_i + \beta_{i1} f_1 + \ldots + \beta_{ik} f_k) = q(1 + E_i) + \beta_{i1} q(f_1) + \ldots + \beta_{ik} q(f_k) = (1 + E_i) / (1 + r) + \beta_{i1} q(f_1) + \ldots + \beta_{ik} q(f_k),$$

which implies that

$$E_i - r = \lambda_1 \beta_{i1} + \ldots + \lambda_k \beta_{ik}, \quad (22)$$

where $\lambda_j \equiv -(1 + r) q(f_j)$ is the risk premium associated with factor j. Equation (22) is the basic equation of the arbitrage pricing theory. We have derived it using absence of exact arbitrage in the absence of asset-specific noise. More general derivations account for asset-specific noise and use absence of approximate arbitrage.

The most important paper in option pricing, Black and Scholes (1973), is based on the absence of arbitrage, as is the whole literature it has generated. At any point in time, the option is priced by duplicating the value one period later using a portfolio of other assets, and assigning a value using the law of one price. We will illustrate this procedure using the binomial process studies by Cox, Ross and Rubinstein (1979). During each period, the stock price either goes up by 20 per cent or it goes down by 10 per cent, and for simplicity we take the riskless rate to be zero. Assume that we are one period from the maturity of a call option with an exercise price of $100, and that the stock price is now $100 (the call is at the money).

How much is the option worth? To figure this out, we must find a portfolio of the stock and the bond that gives the same terminal value. This is the solution of two linear equations (one for each state) in two unknowns (the two portfolio weights). Explicitly, the terminal call value is the larger of 0 and the stock price less 100. In the good state, the stock value will be $120 and the option will be worth $20. In the bad state, the stock price will be $90 and the option will be worthless. If α_S is the amount of stock and α_B the amount of $100 face bond to hold in the duplicating portfolio, then we have that

$$20 = 120\alpha_S + 100\alpha_B$$

to duplicate the option value in the good state, and

$$0 = 90\alpha_S + 100\alpha_B$$

to duplicate the option value in the bad state. The solution to the two equations is given by

$$\alpha_S = 2/3 \quad \alpha_B = -3/5.$$

Therefore, each option is equivalent to holding 2/3 shares of stock and shorting (borrowing) 3/5 bonds. By the law of one price, the option value is the value of this portfolio, or $100\alpha_S + 100\alpha_B = 6\ 2/3$. In this context, we used arbitrage to value the option exactly. More generally, if less is known about the form of the stock price process, absence of arbitrage still places useful restrictions on the option price (Merton, 1973; Cox and Ross, 1976b). For example, the price of a call option is less than the current stock price, and the price of a European put option is no smaller than the present value of the stock price less the current stock price.

Absence of arbitrage also implies a surprising feature of the behaviour of long interest rates in the limit as maturity increases. Let $V(t,T)$ denote the zero-coupon bond price, namely, the price at t of a riskless claim for $1 at T. Equivalently, we can describe bond prices in terms of the zero-coupon rate $z(t,T)$ where $V(t,T)=1/(1+z(t,T))^{T-t}$. Defining the long zero-coupon rate, $zL(t) \equiv \lim_{T \uparrow \infty} z(t, T)$, absence of arbitrage implies that the probability is zero that this rate will ever fall. This is because the bond price today is an average of bond prices tomorrow weighted by (positive) state prices, and the bond price in any state declines asymptotically at the rate $zL(t)$ in that state tomorrow. Thus, the weighted

average of prices today declines at a rate equal to the smallest rate under our maintained assumption of finitely many states (and perhaps more slowly given infinitely many states). As a consequence, $zL(t)$ at time t is always less than or equal to its value $zL(t)$ at any future date, $s>t$, in every realization (with probability one). For details see Dybvig, Ingersoll and Ross (1996).

Dominance is a useful concept to combine with the absence of arbitrage. A dominance argument gives features of a strategy that are optimal independent of preferences and, often, independent of distributions as well. For example, when we write the payoff on a call as $\max(S-X, 0)$, we are implicitly assuming it is a chosen strategy to exercise the option when it is in the money and not to exercise it when it is out of the money. Absent frictions, this is a dominant strategy and the assumption is without loss of generality. A more subtle dominance argument, relying on the absence of frictions and on a non-negative riskless rate, gives the classical result that an American call option (which can be exercised at or before maturity) has the same value as the corresponding European call option (which can only be exercised at maturity), because waiting to exercise is a dominant strategy (Merton, 1973; Cox and Ross, 1976b). Another dominance argument can be used to show that it is optimal to exercise certain reload options used in executive compensation again and again, whenever they are in the money (Dybvig and Loewenstein, 2003).

An alternative to option pricing by arbitrage is to use a ‘preference-based’ model and price options using the first-order conditions of an agent (Rubinstein, 1976). While using this alternative approach is very convenient in some contexts, the Fundamental Theorem of Asset Pricing tells us that we are not really doing anything different, and that the two approaches are simply two different ways of making the same assumption. The same point is true of the distinction some authors have made between the ‘equilibrium’ derivations of the arbitrage pricing theory and the ‘arbitrage’ derivations: there is no substance in this distinction. One derivation may give a tighter approximation than another, but all derivations require similar assumptions in one form or another.

See Also

- finance
- Modigliani–Miller theorem
- options
- present value

Bibliography

Beja, A. 1971. The structure of the cost of capital under uncertainty. *Review of Economic Studies* 38, 359–68.

Black, F. and Scholes, M.S. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Brown, S. and Warner, J. 1980. Measuring security price performance. *Journal of Financial Economies* 8, 205–58.

Brown, S. and Warner, J. 1985. Using daily stock returns: the case of event studies. *Journal of Financial Economics* 14, 3–31.

Cox, J. and Leland, H. 2000. On dynamic investment strategies. *Journal of Economic Dynamics and Control* 24, 1859–80.

Cox, J. and Ross, S.A. 1976a. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3, 145–66.

Cox, J. and Ross, S.A. 1976b. A survey of some new results in financial option pricing theory. *Journal of Finance* 31, 383–402.

Cox, J., Ross, S. and Rubinstein, M. 1979. Option pricing: a simplified approach. *Journal of Financial Economics* 7, 229–63.

Dybvig, P. 1980. Some new tools for testing market efficiency and measuring mutual fund performance. Unpublished manuscript.

Dybvig, P. 1988. Distributional analysis of portfolio choice. *Journal of Business* 61, 369–93.

Dybvig, P. and Ingersoll, J. Jr. 1982. Mean-variance theory in complete markets. *Journal of Business* 55, 233–51.

Dybvig, P., Ingersoll, J. and Ross, S.A. 1996. Long forward and zero-coupon rates can never fall. *Journal of Business* 69, 1–25.

Dybvig, P. and Ross, S. 1982. Portfolio efficient sets. *Econometrica* 50, 1525–46.

Dybvig, P. and Loewenstein, M. 2003. Employee reload options: pricing, hedging, and optimal exercise. *Review of Financial Studies* 16, 145–

71.

Einzig, P. 1937. *The Theory of Forward Exchange*. London: Macmillan.

Harrison, J.M. and Kreps, D. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20, 381–408.

Karlin, S. 1959. *Mathematical Methods and Theory in Games, Programming, and Economics*. Reading, MA: Addison-Wesley.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London: Macmillan.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 13–37.

Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and management Science* 4, 141–83.

Ross, S.A. 1976a. Return, risk and arbitrage. In *Risk and Return in Finance*, ed. I. Friend and J. Bicksler. Cambridge, MA: Ballinger.

Ross, S.A. 1976b. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13, 341–60.

Ross, S.A. 1978. A simple approach to the valuation of risky streams. *Journal of Business* 51, 453–75.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics and Management Science* 7, 407–25.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

How to cite this article

Dybvig, Philip H. and Stephen A. Ross. "arbitrage." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000123> doi:10.1057/9780230226203.0052

The New Palgrave Dictionary of Economics Online

ARCH models

Oliver B. Linton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The ARCH model and its many generalizations are very important in analysing discrete time financial data. We review the properties of the original model and discuss many of the subsequent developments.

Keywords

ARCH models; ARMA models; estimation; exponentially weighted moving average model; factor models; GARCH models; generalized error distribution; heteroskedasticity; IGARCH models; linear models; long memory models; multivariate models; news impact curve; nonparametric models; semiparametric models; stationarity; time series analysis; unit roots

Article

Introduction of model and basic properties

The key properties of financial time series appear to be that: (*a*) marginal distributions have heavy tails and thin centres (leptokurtosis); (*b*) the scale appears to change over time; (*c*) return series appear to be almost uncorrelated over time but to be dependent through higher moments (see Mandelbrot, 1963; Fama, 1965). Linear models like the autoregressive moving average (ARMA) class cannot capture well all these phenomena, since they only really address the conditional mean $\mu_t = E(y_t | y_{t-1}, \ldots)$ and in a rather limited way. This motivates the consideration of nonlinear models. For a discrete time stochastic process y_t, the conditional variance $\sigma_t^2 = \mathrm{var}(y_t | y_{t-1}, \ldots)$ of the process is a natural measure of risk for an investor at time t–1. Empirically it appears to change over time and so it is important to have a model for it. Engle (1982) introduced the autoregressive conditional heteroskedasticity (ARCH) model

$$\sigma_t^2 = \omega + \gamma y_{t-1}^2, \; t = 0, \pm 1, \ldots,$$

where for simplicity we rewrite $y_t \mapsto y_t - \mu_t$ and suppose that the process started in the infinite past. This model makes σ_t^2 vary over time depending on the realization of past squared returns. For σ_t^2 to be a valid conditional variance it is necessary that $\omega > 0$ and $\gamma \geq 0$, in which case $\sigma_t^2 > 0$ for all *t*. Suppose also that $y_t = \varepsilon_t \sigma_t$ with ε_t i.i.d. mean zero and variance one. Provided $\gamma < 1$, the process y_t is weakly (covariance) stationary and has finite unconditional variance $\sigma^2 = E(\sigma_t^2) = E(y_t^2) = \omega / (1 - \gamma)$. This can be proven rigorously under a variety of assumptions on the initialization of the process (see Nelson, 1990). The meaning of this is that the process fluctuates about the long-run value σ^2 and forecasts converge to this value as the forecast horizon lengthens.

The ARCH process is dynamic like ARMA models and indeed we can write the process as an AR(1) in y_t^2, that is,

$$y_t^2 = \omega + \gamma y_{t-1}^2 + \eta_t,$$

where $\eta_t = y_t^2 - \sigma_t^2 = \sigma_t^2 (\varepsilon_t^2 - 1)$ is a mean zero, uncorrelated sequence, that is heteroskedastic. Therefore, we generally have dependence in σ_t^2, y_t^2, and because of the parameter restrictions, positive dependence that is, $\mathrm{cov}(\sigma_t^2, \sigma_{t-j}^2) > 0$ and $\mathrm{cov}(y_t^2, y_{t-j}^2) > 0$. As far as the second order properties (that is, the covariance function) of the process y_t^2, this is identical to that of an AR(1) process. However, it should be remembered that y_t^2 is heteroskedastic itself and that the form of the heteroskedasticity has to be particularly extreme since y_t^2 is kept non-negative.

One feature of linear models like the ARMA class is that the marginal distribution of the variable is normally distributed whenever the shocks are i.i.d. normally distributed. This is not the case for the ARCH class of processes. Specifically, the marginal distribution of y_t will be heavy tailed even if $\varepsilon_t = (y_t - \mu_t) / \sigma_t$ is standard normal. Suppose ε_t is standard normal (and the process is weakly stationary), then the excess kurtosis of y_t is $\kappa_4 = 6\gamma^2 / (1 - 3\gamma^2) \geq 0$ provided $\gamma^2 < 1/3$. If $\gamma \geq 1/3^{1/2}$, then $E(y_t^4) = \infty$. For leptokurtic ε_t, the restriction on γ for finite fourth moment is even more severe. Although the ARCH(1) model implies heavy tails and volatility clustering, it does not in practice generate enough of either. The constraint on γ for finite fourth moment severely restricts the amount of persistence; it is an undesirable feature that the same parameter controls both persistence and heavy tailedness, although if one allows non-normal distributions for ε_t, this link is broken on one side

at least.

The extension to the ARCH(p) process with p lags, while more flexible, becomes very complicated to estimate without restrictions on the coefficients. Bollerslev (1986) introduced the GARCH(p,q) process

$$\sigma_t^2 = \omega + \sum_{k=1}^{p} \beta_k \sigma_{t-k}^2 + \sum_{j=1}^{q} \gamma_j (y_{t-j} - \mu_{t-j})^2,$$

whose p=1, q=1 GARCH(1,1) special case contains only three parameters and usually does a better job than an unrestricted ARCH(12), say, according to a variety of statistical criteria. The GARCH(1,1) process is probably still the most widely used model. As with the ARCH process one needs restrictions on the parameters to make sure that σ_t^2 is positive with probability one. For the GARCH(1,1) it is necessary that $\gamma, \beta \geq 0$ and $\omega > 0$. Interestingly, for higher order processes it is not necessary that $\omega, \gamma_j, \beta_j \geq 0$ for all j: see Nelson and Cao (1992). For example, in GARCH(1,2) the conditions are that $\beta, \gamma_1 \geq 0$ and $\beta\gamma_1 + \gamma_2 \geq 0$. Provided $\Sigma_{k=1}^{p} \beta_k + \Sigma_{j=1}^{q} \gamma_j < 1$, the process y_t is weakly stationary and has finite unconditional variance

$$\sigma^2 = E(\sigma_t^2) = \frac{\omega}{1 - \Sigma_{k=1}^{p} \beta_k - \Sigma_{j=1}^{q} \gamma_j}.$$

As for the ARCH process, the series y_t has higher kurtosis than ε_t.

Drost and Nijman (1993) provide an important classification of ARCH models according to the precise properties required of the error terms. The strong GARCH process is where

$$\varepsilon_t = \frac{y_t - \mu_t}{\sigma_t} \text{ i.i.d. } E(\varepsilon_t) = 0 \text{ and } E(\varepsilon_t^2) = 1.$$

It is generally this case that has been investigated in the literature. It is a very strong assumption by the standards of most modern econometrics, where usually only conditional moment restrictions are imposed, but is a complete specification that is useful for deriving properties like stationarity. The strong Gaussian case is where ε_t is additionally normally distributed. The semi-strong GARCH process is where

$$E[\varepsilon_t | y_{t-1}, y_{t-2}, \ldots] = 0 \text{ and } E\left[\varepsilon_t^2 | y_{t-1}, y_{t-2}, \ldots\right] = 1.$$

These assumptions are weaker and turn out to be sufficient in many cases for consistent estimation. They are quite weak assumptions and restrict only the conditional mean and conditional variance of the process, allowing a variety of behaviour in the potentially time varying distribution of ε_t. Drost and Nijman (1993) show that conventional strong and semi-strong GARCH processes are not closed under temporal aggregation, meaning that if a process is GARCH at the daily frequency that the weakly or monthly data may not be GARCH, either weak or strong.

Strong stationarity and mixing

Consider the GARCH(1,1) process

$$y_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma y_{t-1}^2$$

with ε_t i.i.d. and $\omega > 0$ and $\beta, \gamma \geq 0$. A sufficient condition for strong stationarity is that $E[\ln(\beta + \gamma \varepsilon_t^2)] < 0$ (see Nelson, 1990). If additionally, $E(\varepsilon_t)=0$ and $\mathrm{var}(\varepsilon_t)=1$, then the necessary and sufficient condition for weak stationarity is that $\beta + \gamma < 1$. By Jensen's inequality $E[\ln(\beta + \gamma \varepsilon_t^2)] < \ln E[(\beta + \gamma \varepsilon_t^2)] = \ln(\beta + \gamma)$, so it can be that $E[\ln(\beta + \gamma \varepsilon_t^2)] < 0$ even when $\beta + \gamma \geq 1$, that is, there are strongly stationary processes that are not weakly stationary.

There are many measures of dependence in time series. Mixingness is the property that dependence dies out with horizon. It can be measured in different ways: covariance mixing, strong mixing, and beta mixing are the main concepts. A stationary sequence $\{X_t, t=0, \pm 1, \ldots\}$ is said to be covariance mixing if $\mathrm{cov}(X_t, X_{t+k}) \to 0$ as $k \to \infty$. A stationary sequence $\{X_t, t=0, \pm 1, \ldots\}$ is said to be strong mixing (α-mixing) if

$$\alpha(k) = \sup_{A \in \mathcal{F}_{-\infty}^{n}, B \in \mathcal{F}_{n+k}^{\infty}} |P(AB) - P(A)P(B)| \to 0$$

as $k \to \infty$, where $\mathcal{F}_{-\infty}^{n}$ and $\mathcal{F}_{n+k}^{\infty}$ are two σ-fields generated by $\{X_t, t \leq n\}$ and $\{X_t, t \geq n+k\}$,

respectively. We call $\alpha(\cdot)$ the mixing coefficient. A stationary sequence $\{X_t, t=0, \pm 1, \ldots\}$ is said to be β-mixing if

$$\beta(k) = \sup_{A \in \mathcal{F}_{-\infty}^{n},\, B \in \mathcal{F}_{n+k}^{\infty}} |P(A|B) - P(A)| \to 0$$

as $k \to \infty$. We call $\beta(\cdot)$ the mixing coefficient. We have $2\alpha(k) \leq \beta(k)$. The covariance mixing property is only well defined for weakly stationary processes, so it is natural here to work with the more general notions of α and β mixing. A sufficient condition that a GARCH(1,1) process is β-mixing with exponential decay is that it is weakly stationary, Carrasco and Chen (2002), but this is not necessary. More recently it has been shown that IGARCH is strong mixing under some conditions (see Meitz and Saikkonen, 2004). One problem is that when you combine a GARCH process with other processes for the mean, the mixingness is not preserved and has still to be established. The weaker concept of near epoch dependence can be established, though in quite a general class of models (Hansen, 1991). Why does mixing matter? It is a key property that allows one to learn from the data through the law of large numbers and central limit theorems.

IGARCH models

In practice, estimated GARCH parameters lie close to the boundary of the weakly stationary region. This prompts consideration of the process where $\sum_{k=1}^{p} \beta_k + \sum_{j=1}^{q} \gamma_j = 1$, which is called the integrated GARCH or IGARCH. In this case, the process y_t with i.i.d. Gaussian innovations is strongly stationary but not covariance stationary, since the unconditional variance is infinite (although the conditional variance is finite with probability 1). This is in contrast to linear unit root processes in which the process is neither weakly nor strongly stationary and these two notions coincide. Also, in contrast to the linear case, differencing does not induce weak stationarity, that is, $y_t^2 - y_{t-1}^2$ is not weakly stationary (although its mean is constant over time).
The exponentially weighted moving average model (sometimes called the J.P. Morgan model) is a variant on the IGARCH model in which there is no intercept ω and a unit root:

$$y_t = \sigma_t \varepsilon_t,\ \sigma_t^2 = \beta \sigma_{t-1}^2 + (1-\beta) y_{t-1}^2.$$

It is a very simple process with only one parameter and is widely used by practitioners, with particular values of the parameter β. Write $\sigma_t^2 = \sigma_{t-1}^2 [\beta + (1-\beta)\varepsilon_{t-1}^2]$, so that $\ln \sigma_t^2$ is a random walk, that is,

$$\ln \sigma_t^2 = \ln \sigma_{t-1}^2 + \eta_{t-1}, \quad \eta_{t-1} = \ln(\beta + (1-\beta)\varepsilon_{t-1}^2),$$

and hence is not strongly stationary. On the other hand, the process y_t is informally weakly stationary since $E[y_t^2 | \mathcal{F}_{-\infty}^0] = E[\sigma_t^2 | \mathcal{F}_{-\infty}^0] = \Pi_{s=1}^t E[\beta + (1-\beta)\varepsilon_s^2]\sigma_0^2 = \sigma_0^2$ for all t. The properties of this process depend on the moments of η_{t-1}. If $E[\eta_{t-1}] > 0$, then $\sigma_t^2 \to \infty$ with probability 1. If $E[\eta_{t-1}] < 0$, then $\ln \sigma_t^2 \to -\infty$ with probability 1 as $t \to \infty$ and so $\sigma_t^2 \to 0$ with probability 1. If $E[\eta_{t-1}]=0$, then $\ln \sigma_t^2$ is a driftless random walk and the process just wanders everywhere. If we assume $E[\varepsilon_t^2] = 1$, then by Jensen's inequality $E[\eta_{t-1}] < 0$, and the process $\sigma_t^2 \to 0$ with probability 1 as $t \to \infty$ whatever the initialization. Thus the process is essentially degenerate and is not plausible, despite being widely used.

Functional form

The news impact curve is the relationship between σ_t^2 and $y_{t-1}=y$ holding past values σ_{t-1}^2 constant at some level σ^2. This is an important relationship that describes how new information affects volatility. For the GARCH process, the news impact curve is

$$m(y, \sigma^2) = \omega + \gamma y^2 + \beta \sigma^2.$$

It is separable in σ^2, it is an even function of news y, $m(y, \sigma^2)=m(-y, \sigma^2)$, and it is a quadratic function of y. The symmetry property implies that $\mathrm{cov}(y_t^2, y_{t-j}) = 0$ for symmetric about zero ε_t.
The GARCH process does not allow ‘leverage effects’ or asymmetric news impact curves. Because of limited liability, we might expect that negative and positive shocks have different effects on volatility. Nelson (1991) introduced the exponential GARCH model. Let $h_t = \log \sigma_t^2$ and let

$$h_t = \omega + \sum_{j=1}^{p} \gamma_j [\theta \varepsilon_{t-j} + \delta |\varepsilon_{t-j}|] + \sum_{k=1}^{q} \beta_k h_{t-k},$$

where $\varepsilon_t = (y_t - \mu_t) / \sigma_t$ is i.i.d. with mean zero and variance one. Nelson's paper contains four innovations. First, it models the log, not the level. Therefore there are no parameter restrictions to ensure that $\sigma_t^2 \geq 0$. Second, it allows asymmetric effect of past shocks ε_{t-j} on current volatility, that is, the news impact curve is allowed to be asymmetric. For example, $\mathrm{cov}(y_t^2, y_{t-j}) \neq 0$ even when ε_t is symmetric about zero. Third, it makes the innovations ε_t i.i.d. It follows that h_t is a linear process so that strong and weak stationarity coincide where they ought to (for h_t anyway). On the other hand estimation and forecasting is quite tricky because of the repeated exponential/logarithmic transformations involved. The final innovation was to allow heavy tailed innovations based on the so-called generalized error distribution (GED) that nests the Gaussian as a special case.

An alternative approach to allowing asymmetric news impact curve is the Glosten, Jagannathan and Runkle (1993) model

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 + \delta y_{t-1}^2 1(y_{t-1} < 0).$$

In this case, the news impact curve is asymmetric but still has quadratic tails. It is a simple enough modification, that it has similar probabilistic properties to the GARCH(1,1) process. There are many other variations on the basic GARCH model, too many to list here, but the interested reader can find a fuller description in the survey paper of Bollerslev, Engle and Nelson (1994).

One might expect that risk and return should be related: see Merton (1973) for an example. The GARCH-in-Mean process captures this idea. This process is

$$y_t = g(\sigma_t^2; b) + \varepsilon_t \sigma_t,$$

for various functional forms of *g*, for example, linear and log-linear and for some given GARCH specification of σ_t^2. Engle, Lilien and Robins (1987) used this model on interest rate data (see also Pagan and Hong, 1991). Here, *b* are parameters to be estimated along with the parameters of the error variance. Some authors find small but significant effects.

Estimation

The standard approach to estimation of these models has been through estimation of the (conditional) Gaussian quasi-likelihood criterion

$$L_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta) = -\sum_{t=1}^{T} \log \sigma_t^2(\theta) - \frac{1}{2}\sum_{t=1}^{T}\left(\frac{y_t - \mu_t(\theta)}{\sigma_t(\theta)}\right)^2,$$

where $\sigma_t^2(\theta)$ and perhaps $\mu_t(\theta)$ are built up by recursions from some starting values. There are several possibilities regarding starting values: (a) $\sigma_0^2(\theta) = \omega / 1 - \beta - \gamma$, (b) $\sigma_0^2(\theta) = T^{-1}\Sigma_{t=1}^{T} y_t^2$, and (c) $\sigma_0^2(\theta) = y_1^2$. Approach (a) imposes weak stationarity and would not be appropriate were IGARCH to be thought plausible, while value (b) sort of requires weak stationarity for the asymptotic properties to follow through. The likelihood function is maximized with respect to the parameter values usually using some derivative-based algorithm like BHHH and sometimes imposing inequality restrictions (like those required for $\sigma_t^2 \geq 0$ with probability 1 or for σ_t^2 to be weakly stationary) and sometimes not.

The (quasi) MLE (QMLE) can be expected to be consistent provided only the conditional mean and the conditional variance are correctly specified (Bollerslev and Wooldridge, 1992), that is, semi-strong not strong GARCH is required and conditional normality is certainly not required. This is true because the score function $\partial \ell_t(\theta_0) / \partial \theta$ is a martingale difference sequence. Robust standard errors can be constructed in the usual way

$$\left[\frac{\partial \ell_T(\hat{\theta})}{\partial \theta \partial \theta^{\mathrm{T}}}\right]^{-1}\left[\sum_{t=1}^{T} \frac{\partial \ell_t}{\partial \theta}\frac{\partial \ell_t}{\partial \theta^{\mathrm{T}}}(\hat{\theta})\right]\left[\frac{\partial \ell_T(\hat{\theta})}{\partial \theta \partial \theta^{\mathrm{T}}}\right]^{-1}, \quad (1)$$

although the default option in many software packages is to compute standard errors as if Gaussianity held.

The distribution theory is difficult to establish from primitive conditions even for simple models. There is one important point about these asymptotics – that one does not need moments on y_t (for example, one does not need weak stationarity). Lumsdaine (1996) established consistency and asymptotic normality allowing the IGARCH case but under strong stationarity and symmetric unimodal i.i.d. ε_t with $E[\varepsilon_t^{32}] < \infty$. Lee and Hansen (1994) proved the same result under weaker conditional moment conditions and allows for semi-strong processes with some higher-level assumptions. Jensen and Rahbek (2004) established consistency and asymptotic normality of the QMLE in strong GARCH model without strict stationarity. Hall and Yao (2003) assume weak stationarity and show that if $E(\varepsilon_t^4) < \infty$ the asymptotic normality holds, but also establish limiting behaviour (non-normal) under weaker moment conditions. No results have yet been published for consistent and asymptotically normality of EGARCH

from primitive conditions, although simulation evidence does suggest normality is a good approximation in large samples.

Typically, one finds small intercepts and a large parameter on the lagged dependent volatility; see Lumsdaine (1995) and Brooks, Burke and Persand (2001) for simulation evidence. These two parameter estimates are often highly correlated. Engle and Sheppard (2001) suggested a method they called target variance to obviate the computational difficulties sometimes encountered in estimating GARCH models. For a weakly stationary GARCH(1,1) process we have $E(v_t^2) = \omega / (1 - \beta - \gamma)$ so that $\omega = E(v_t^2)(1 - \beta - \gamma)$. They suggest replacing $E(v_t^2)$ by $\Sigma_{t=1}^{T} v_t^2 / T$ in the likelihood so that one only has two parameters to chose. This results in a much more stable performance of most algorithms. The downside with this approach is that distribution theory is much more complicated due to the lack of martingale property, and in particular one needs to use Newey–West standard errors.

It is quite common now to estimate GARCH models using different objective functions suggested by alternative specifications of the error distribution like the *t* or the GED distribution that Nelson (1991) favoured. These objective functions often have additional parameters such as the degrees of freedom that have to be computed. They lead to greater efficiency when the chosen specification is correct, but otherwise can lead to inconsistency, as was shown by Newey and Steigerwald (1997).

Long memory

The GARCH(1,1) process $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma v_{t-1}^2$ is of the form

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j v_{t-j}^2 \tag{2}$$

for constants c_j satisfying $c_j = \gamma\beta^{j-1}$, provided the process is weakly stationary, which requires $\gamma + \beta < 1$. These coefficients decay very rapidly so the actual amount of memory is quite limited. There is some empirical evidence on the autocorrelation function of v_t^2 for high frequency data that suggests a slower decay rate than would be implied by these coefficients. Long memory models essentially are of the form (2) but with slower decay rates. For example, suppose that $c_j = j^{-\theta}$ for some $\theta > 0$. The coefficients satisfy $\Sigma_{j=1}^{\infty} c_j^2 < \infty$ provided $\theta > 1/2$. Fractional integration (FIGARCH) leads to such an expansion. There is a single parameter called *d* that determines the memory properties of the series, and

$$(1-L)^d\sigma_t^2 = \omega + \gamma\sigma_{t-1}^2(\varepsilon_{t-1}^2 - 1),$$

where $(1–L)^d$ denotes the fractional differencing operator. When d=1 we have the standard IGARCH model. For $d \neq 1$ we can define the binomial expansion of $(1–L)^{–d}$ in the form given above. See Robinson (1991) and Bollerslev and Mikkelson (1996) for models and evidence of long memory. The evidence for long memory is often based on sample autocovariances of y_t^2, and this may be questionable due to a paper of Mikosch and St•ric• (2000).

Multivariate models

In practice we observe many closely related series, and so it may be important to model their behaviour jointly. Define the conditional covariance matrix

$$\Sigma_t = E(y_t y_t^{\mathrm{T}} | \mathcal{F}_{-\infty}^{t-1})$$

for some n×1 vector of mean zero series y_t. Bollerslev, Engle and Wooldridge (1988) introduced the most general generalization of the univariate GARCH(1,1) process

$$h_t = \mathrm{vech}(\Sigma_t) = A + Bh_{t-1} + C\mathrm{vech}(y_{t-1}y_{t-1}^{\mathrm{T}}),$$

where A is an $n(n+1)/2$×1 vector, while B, C are $n(n+1)/2$×$n(n+1)/2$ matrices. In practice, there are too many parameters. Also, the restrictions on the parameters to ensure that Σ_t is positive definite are very complicated in this formulation. For weak stationarity one requires that the matrix $I–B–C$ is nonsingular and positive definite in which case the unconditional variance matrix is unvech($(I–B–C)^{–1}A$). The conditions for strong stationarity are rather complicated to state.
The so-called BEKK model is a special case that addresses these issues. It is of the form

$$\Sigma_t = AA^{\mathrm{T}} + B\Sigma_{t-1}B^{\mathrm{T}} + Cy_{t-1}y_{t-1}^{\mathrm{T}}C^{\mathrm{T}}$$

for n×n matrices A, B, C. This gives a big reduction in number of parameters and imposes symmetry and

positive definiteness automatically. There are still many parameters that have to be estimated simultaneously, of the order n^2, and this limits the applicability and interpretability of this model. Bollerslev (1990) introduced the constant conditional covariance (CCC) model, which greatly reduces the parameter explosion issue. This involves standard univariate dynamic models for each of the conditional variances and a constant correlation assumption, that is,

$$\Sigma_t = D_t R D_t,\ D_t = \mathrm{diag}\{\sigma_{it}\} \tag{3}$$

$$\sigma_{it}^2 = \omega_i + \beta_i \sigma_{i,t-1}^2 + \gamma_i y_{i,t-1}^2 \tag{4}$$

and $R=(R_{ij})$ is a time invariant matrix

$$R_{ij} = \frac{E[\varepsilon_{it}\varepsilon_{jt}]}{\left(E[\varepsilon_{it}^2]E[\varepsilon_{jt}^2]\right)^{1/2}} = E[\varepsilon_{it}\varepsilon_{jt}],$$

where $\varepsilon_{it}=y_{it}/\sigma_{it}$. The values R_{ij} are restricted to lie in [−1,1] and the matrix R is symmetric and positive definite but otherwise unrestricted. This model generates time varying conditional covariances, but the dynamics are all driven by the conditional variances as the correlations are constant. The estimation of R is quite straightforward: use the sample correlation matrix of the standardized residuals $\varepsilon_{it} = y_{it} / \hat{\sigma}_{it}$. The estimated matrix R is guaranteed to be symmetric and positive definite because it is a correlation matrix and consequently the estimated Σ_t shares these properties.

Engle and Sheppard (2001) introduced the dynamic conditional covariance (DCC) model where we replace in (3) and (4)

$$R_{ij,t} = \frac{q_{ij,t}}{\left(q_{ii,t}q_{jj,t}\right)^{1/2}}$$

$$q_{ij,t} = c_{ij} + b_{ij} q_{ij,t-1} + a_{ij} \varepsilon_{i,t-1} \varepsilon_{j,t-1}.$$

If we assume also that $a_{ij}=a$, $b_{ij}=b$, and $c_{ij}=c$ for all $i \neq j$ one can show that the resulting covariance matrix Σ_t is guaranteed to be symmetric and positive definite. This model allows slightly more flexibility in allowing the correlations to vary over time, but because of the need to impose positive definiteness it still imposes common dynamics on the correlations, which may be too restrictive.

The approach that brings the most flexible dimensionality reduction is based on the ideas of factor analysis. Suppose that for $y_t \in \mathbb{R}^n$, $f_t \in \mathbb{R}^k$:

$$y_t = C f_t + u_t \tag{5}$$

$$\begin{pmatrix} f_t \\ u_t \end{pmatrix} \Big| I_{t-1} \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Lambda_t & 0 \\ 0 & \Gamma \end{pmatrix}, \tag{6}$$

where $Y_{t-1}=\{y_{t-1}, \ldots\}$ is the observed information and $I_t = \{y_t, f_t, y_{t-1}, f_{t-1}, \ldots\}$ contains both observed series and the latent factors $F_{t-1} = \{f_t, f_{t-1}, \ldots\}$. Suppose that rank($C$)=$k$ and that Λ_t is a $k \times k$ positive definite time varying matrix. It follows that $y_t | I_{t-1} \sim 0, C\Lambda_t C^T + \Gamma$ (Sentana, 1998). The implied Σ_t is of reduced rank and depends on only order nK (time-varying associated) parameters so there is a big reduction in dimensionality. This model includes as a special case the Diebold and Nerlove (1989) model where Γ, Λ_t are diagonal and $\lambda_{jjt} = var[f_{jt} | I_{t-1}] = \omega_j + \beta_j \lambda_{jj,t-1} + \gamma_j f^2_{j,t-1}$, in which case $\lambda_{jjt} \notin Y_{t-1}$. This process is closed under block marginalization – that is, subsets of y_t do not have the same structure. Estimation is complicated by the latent variables. This framework also includes the Engle, Ng and Rothschild (1990) factor GARCH model $\Sigma_t = \Sigma_0 + \Sigma_{k=1}^{K} \delta_k \delta_k^T \sigma^2_{kt}$, where $K<n$, and

σ_{kt}^2 is the conditional variance of a certain portfolio *k*, with time invariant weights α_k, that is, $y_{kt}^p = \alpha_k^T y_t$ with $\alpha_k^T i = 1$. They assume also that σ_{kt}^2 are standard univariate GARCH(1,1) processes, that is, for some parameters $(\omega_k, \beta_k, \gamma_k)$, $\sigma_{kt}^2 = \omega_k + \beta_k \sigma_{k,t-1}^2 + \gamma_k (\gamma_k^T y_{t-1})^2$. This model is written in terms of observables and consequently its estimation is somewhat easier, but it suffers from the fact that it is not closed under block marginalization – that is, subsets of y_t do not have the same structure. Sentana (1998) shows how it is nested in the general model (5) and (6).

Nonparametric and semiparametric models

There have been a number of contributions to ARCH modelling from the nonparametric or semiparametric point of view; see Hafner (1998) for an overview. Engle and González-Rivera (1991) suggested treating the error distribution in a GARCH process nonparametrically, that is,

$$y_t = \mu_t + \varepsilon_t \sigma_t \sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma (y_{t-1} - \mu_{t-1})^2,$$

where μ_t depends on observed covariates and parameters, while ε_t is i.i.d. with density *f* that is not restricted in shape. This is motivated by the great deal of evidence that the density of the standardized residuals $\varepsilon_t = (y_t - \mu_t) / \sigma_t$ is non-Gaussian. They proposed an estimation algorithm that involved estimating *f* from the data. Linton (1993) and Drost and Klaassen (1997) have shown that one can achieve significant efficiency improvements depending on the shape of the error density.

An alternative line of research has been to treat the functional form of $\sigma_t^2 (y_{t-1}, y_{t-2}, \ldots)$ nonparametrically. In particular, suppose that

$$\sigma_t^2 = g(y_{t-1}, \ldots, y_{t-p})$$

for some unknown function *g* and fixed lag length *p*. This allows for a general shape to the news impact curve and nests all the usual parametric ARCH processes. See Pagan and Hong (1991) and Härdle and Tsybakov (1997) for some applications. This model is somewhat limited in the dependence it allows in comparison with the GARCH(1,1) process, which is a function of all past $y's$. Also, the curse of dimensionality means that the usual estimation methods do not work well in practice for large *p*, that is, $p>4$.

One compromise approach to avoiding the curse of dimensionality is to use additive models, whence

$$\sigma_t^2 = \sum_{j=1}^{p} g_j(y_{t-j}) \tag{7}$$

for some unknown functions g_j. The functions g_j are allowed to be of general functional form but only depend on y_{t-j}. This class of processes nests many parametric ARCH models. The functions g_j can be estimated by kernel regression techniques (see Masry and Tjøstheim, 1995). Yang, Härdle and Nielsen (1999) proposed an alternative nonlinear ARCH model in which the conditional mean is again additive, but the volatility is multiplicative $\sigma_t^2 = c_y \Pi_{j=1}^{d} \sigma_j^2(y_{t-j})$. Kim and Linton (2004) generalize this model to allow for arbitrary, but known, transformations, that is, $G(\sigma_t^2) = c_y + \Sigma_{j=1}^{d} \sigma_j^2(y_{t-j})$, where G(.) is known function like log or level. Linton and Mammen (2005) considered the case where $\sigma_t^2 = \Sigma_{j=0}^{\infty} \beta^{j-1} g(y_{t-j})$, which nests the GARCH(1,1) process when $g(y) = \omega + \gamma y^2$.
One final semiparametric approach has been to model the coefficients of a GARCH process as changing over time, thus

$$\sigma_t^2 = \omega(x_{tT}) + \beta(x_{tT})\sigma_{t-1}^2 + \gamma(x_{tT})(y_{t-1} - \mu_{t-1})^2,$$

where ω, β, and γ are smooth functions of a variable x_{tT}, for example, $x_{tT}=t/T$. This class of processes is non-stationary but can be viewed as locally stationary along the lines of Dahlhaus (1997).

See Also

- continuous and discrete time models
- factor models
- finance
- local regression models
- martingales
- time series analysis

The author would like to thank the Economic and Social Science Research Council of the United Kingdom for financial support through a research fellowship.

Bibliography

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–27.

Bollerslev, T. 1990. Modelling the coherence in short-run nominal exchange rates: a multivariate generalized autoregressive conditional heteroskedasticity. *Review of Economics and Statistics* 72, 498–505.

Bollerslev, T., Engle, R.F. and Wooldridge, J.M. 1988. A capital asset pricing model with time varying covariances. *Journal of Political Economy* 96, 116–31.

Bollerslev, T. and Mikkelson, H.O. 1996. Modelling and pricing long memory in stock market volatility. *Journal of Econometrics* 73, 151–184, 498–505.

Bollerslev, T. and Wooldridge, J.M. 1992. Quasi maximum likelihood estimation and inference in dynamic models with time varying covariances. *Econometric Reviews* 11, 143–72.

Bollerslev, T., Chou, R.Y. and Kroner, K. 1992. ARCH modelling in finance. *Journal of Econometrics* 52, 5–59.

Bollerslev, T., Engle, R.F. and Nelson, D. 1994. ARCH models. In *The Handbook of Econometrics*, vol. 4, ed. D.F. McFadden and R.F. Engle III. Amsterdam: North-Holland.

Brooks, C., Burke, S.P. and Persand, G. 2001. Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting* 17, 45–56.

Carrasco, M. and Chen, X. 2002. Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* 18, 17–39.

Dahlhaus, R. 1997. Fitting time series models to nonstationary processes. *Annals of Statistics* 25, 1–37.

Diebold, F.S. and Nerlove, M. 1989. The dynamics of exchange-rate volatility: a multivairate latent-factor ARCH model. *Journal of Applied Econometrics* 4, 1–22.

Drost, F.C. and Klaassen, C.A.J. 1997. Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* 81, 193–221.

Drost, F.C. and Nijman, T.E. 1993. Temporal aggregation of GARCH processes. *Econometrica* 61, 909–27.

Engle, R.F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987–1008.

Engle, R.F. and Bollerslev, T. 1986. Modeling the persistence of conditional variances. *Econometric Reviews* 5, 1–50.

Engle, R.F. and González-Rivera, G. 1991. Semiparametric ARCH models. *Journal of Business and Economic Statistics* 9, 345–59.

Engle, R.F., Lilien, D.M. and Robins, R.P. 1987. Estimating time varying risk premia in the term structure: the ARCH-M model. *Econometrica* 19, 3–29.

Engle, R.F. and Ng, V.K. 1993. Measuring and testing the impact of news on volatility. *Journal of Finance* 48, 1749–78.

Engle, R.F. and Ng, V.K. and Rothschild, M. 1990. Asset pricing with a FACTOR-ARCH covariance structure: empirical estimates for treasury bills. *Journal of Econometrics* 45, 213–37.

Engle, R.F. and Sheppard, K. 2001. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Working Paper No. 8554. Cambridge, MA: NBER.

Fama, E.F. 1965. The behavior of stock market prices. *Journal of Business* 38, 34–105.

Glosten, L.R., Jagannathan, R. and Runkle, D.E. 1993. On the relation between the expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance* 48, 1779–801.

Hafner, C. 1998. *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility*. Heidelberg: Physica.

Hall, P. and Yao, Q. 2003. Inference in ARCH and GARCH models with heavy tailed errors. *Econometrica* 71, 285–317.

Hansen, B.A. 1991. GARCH(1,1) processes are near epoch dependent. *Economics Letters* 36, 181–6.

Hansen, P.R. and Lunde, A. 2005. A forecast comparison of volatility models: does anything beat a GARCH(1,1). *Journal of Applied Econometrics* 20, 873–89.

Härdle, W. and Tsybakov, A.B. 1997. Locally polynomial estimators of the volatility function. *Journal of Econometrics* 81, 223–42.

Härdle, W., Tsybakov, A.B. and Yang, L. 1996. Nonparametric vector autoregression. Discussion Paper,

SFB 373. Berlin: Humbodt-Universität.

Jensen, S.T. and Rahbek, A. 2004. Asymptotic normality of the QMLE of ARCH in the nonstationary case. *Econometrica* 72, 641–6.

Kim, W. and Linton, O. 2004. A local instrumental variable estimation method for generalized additive volatility models. *Econometric Theory* 20, 1094–139.

Lee, S.-W. and Hansen, B.E. 1994. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory* 10, 29–52.

Linton, O.B. 1993. Adaptive estimation in ARCH models. *Econometric Theory* 9, 539–69.

Linton, O.B. and Mammen, E. 2005. Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica* 73, 771–836.

Lumsdaine, R. 1995. Finite-sample properties of the maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models: a Monte Carlo investigation. *Journal of Business and Economic Statistics* 13, 1–10.

Lumsdaine, R.L. 1996. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica* 64, 575–96.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Masry, E. and Tjøstheim, D. 1995. Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality. *Econometric Theory* 11, 258–89.

Meitz, M. and Saikkonen, P. 2004. Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models. Working Paper Series in Economics and Finance No. 573. Stockholm School of Economics.

Merton, R.C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41, 867–87.

Mikosch, T. and St•ric•, C. 2000. Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Annals of Statistics* 28, 1427–51.

Nelson, D.B. 1990. Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory* 6, 318–34.

Nelson, D.B. 1991. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* 59,

347–70.

Nelson, D.B. and Cao, C.Q. 1992. Inequality constraints in the univariate GARCH model. *Journal of Business and Economic Statistics* 10, 229–35.

Newey, W.K. and Steigerwald, D.G. 1997. Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. *Econometrica* 65, 587–99.

Pagan, A.R. and Schwert, G.W. 1990. Alternative models for conditional stock volatility. *Journal of Econometrics* 45, 267–90.

Pagan, A.R. and Hong, Y.S. 1991. Nonparametric estimation and the risk premium. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. W. Barnett, J. Powell and G.E. Tauchen. Cambridge: Cambridge University Press.

Robinson, P.M. 1991. Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics* 47, 67–84.

Robinson, P.M. and Zaffaroni, P. 2006. Pseudo-maximum likelihood estimation of ARCH(∞) models. *Annals of Statistics* 34, 1049–1074.

Sentana, E. 1998. The relation between conditionally heteroskedastic factor models and factor GARCH models. *Econometrics Journal* 1, 1–9.

Yang, L., Härdle, W. and Nielsen, J.P. 1999. Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis* 20, 579–604.

How to cite this article

Linton, Oliver B. "ARCH models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000124> doi:10.1057/9780230226203.0054

The New Palgrave Dictionary of Economics Online

arms races

Dagobert L. Brito and Michael D. Intriligator
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We analyse arms races for an environment in which social, human and intellectual capital are more important than physical capital. The Richardson model can be used to analyse the Anglo–German naval race before the First World War and the US–Soviet missile race during the Cold War; in both cases the economic constraint associated with acquiring weapons was the binding constraint. Previously, human and social capital were more important components of military power. Modern technology has reduced the importance of the economic constraints associated with acquiring physical capital. Our model of such a process suggests that a stable equilibrium is unlikely.

Keywords

arms races; arms trade; Cold War; human capital; increasing returns to scale; lotteries; public goods; returns to scale; Richardson model of arms races; risk; slavery; social capital; technical change; terrorism; uncertainty

Article

The traditional literature on arms races starts with the Richardson model (named after Lewis Fry Richardson, 1881–1953, British polymath who made fundamental contributions to the mathematical analysis of war, to weather forecasting, and to measuring the length of coastlines and borders). The Richardson model is a descriptive model of the dynamic processes of interaction in an arms race. The model is summarized by two differential equations describing the rate of change over time of weapon stocks in each of two countries, 1 and 2. Let $w_1(t)$ represent the stock of weapons for country 1 and $w_2(t)$ represent the stock of weapons for country 2 at time t. In the Richardson model the rate of change of weapon stocks at time t is given by

$$\dot{w}_1(t) = a_1 w_2(t) + b_1 w_1(t) + c_1$$
$$\&$$
$$\dot{w}_2(t) = a_2 w_1(t) + b_2 w_2(t) + c_2 \qquad (1)$$

According to these coupled differential equations, the accumulation of weapons in country 1 can be described as the sum of three separate influences. First is the ‘defence term’, a_1, where the accumulation of weapons is influenced positively by the stock of weapons of the opponent, $w_2(t)$, representing the need to defend oneself against the opponent. Second is the ‘fatigue term’, b_1, where the accumulation of weapons is influenced negatively by one's own stock of weapons, representing the economic and administrative burden of

conducting the arms race. Third is the 'grievance term', c_1, representing all other factors influencing the arms race, whether historical, institutional, cultural, or derived from some other source. The dynamics of the arms accumulation equation for country 2 are symmetrical.

During the Cold War, Richardson's equations attracted much interest among political scientists, economists and others interested in the arms race. One of the questions of interest was the stability of the arms race. There are three schools of thought about the stability of armament races. One is that armaments races have a stable equilibrium. A second belief is that armaments races are unstable, a belief often seen in the popular press, which holds that unless some agreement is reached weapon stocks will increase in an ever-accelerating spiral that must ultimately lead to bankruptcy or nuclear holocaust. A third view is that a stable equilibrium may exist, but that the stability may only be a local property, so that a large disturbance of the system, such as the introduction of a new weapons system, which may set off an armaments race, either positive (leading to larger and larger weapons stocks) or negative (leading to major decreases in weapons).

The first two questions could be addressed by using the parameters of the model to calculate the roots and check for stability. The third question requires that the underlying process that led to these differential equations be modelled. Much of the theoretical work on the arms race is in the Richardson tradition of explaining the arms race was attempt to estimate these parameters empirically or to find theoretical reasons for constraining the magnitudes (as discussed in the Intriligator, 1982, survey paper). The third question was addressed by research that derived the dynamics of arms accumulation in a model based on the axioms of rational choice, on the assumption that each country can be modelled as a single rational actor. Brito (1972) and Intriligator (1975) each obtained a general set of equations describing an arms race, of which the Richardson model is one special case.

In Figure 1, $\hat{x}_1$ and $\hat{x}_2$ are the stable equilibrium.

Figure 1

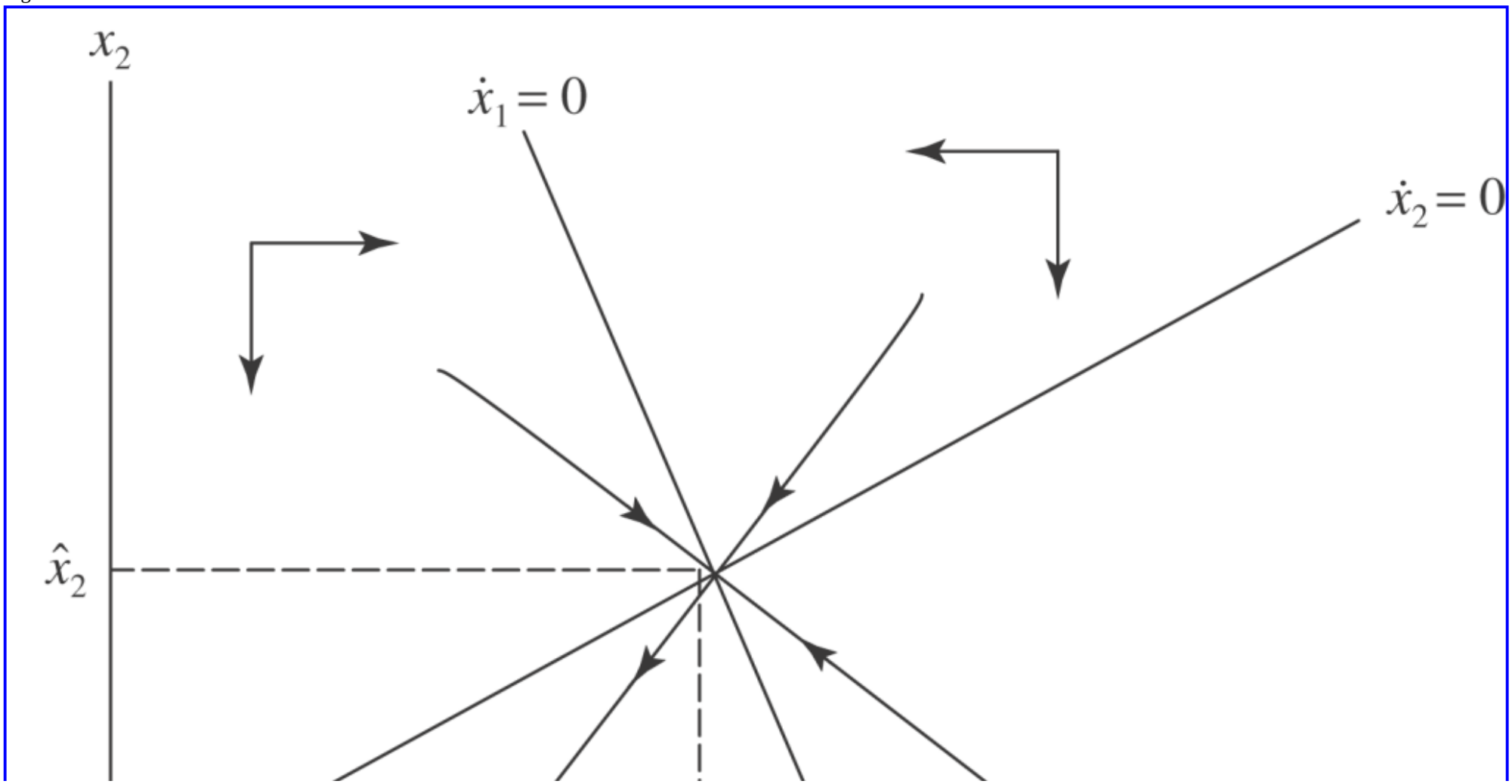

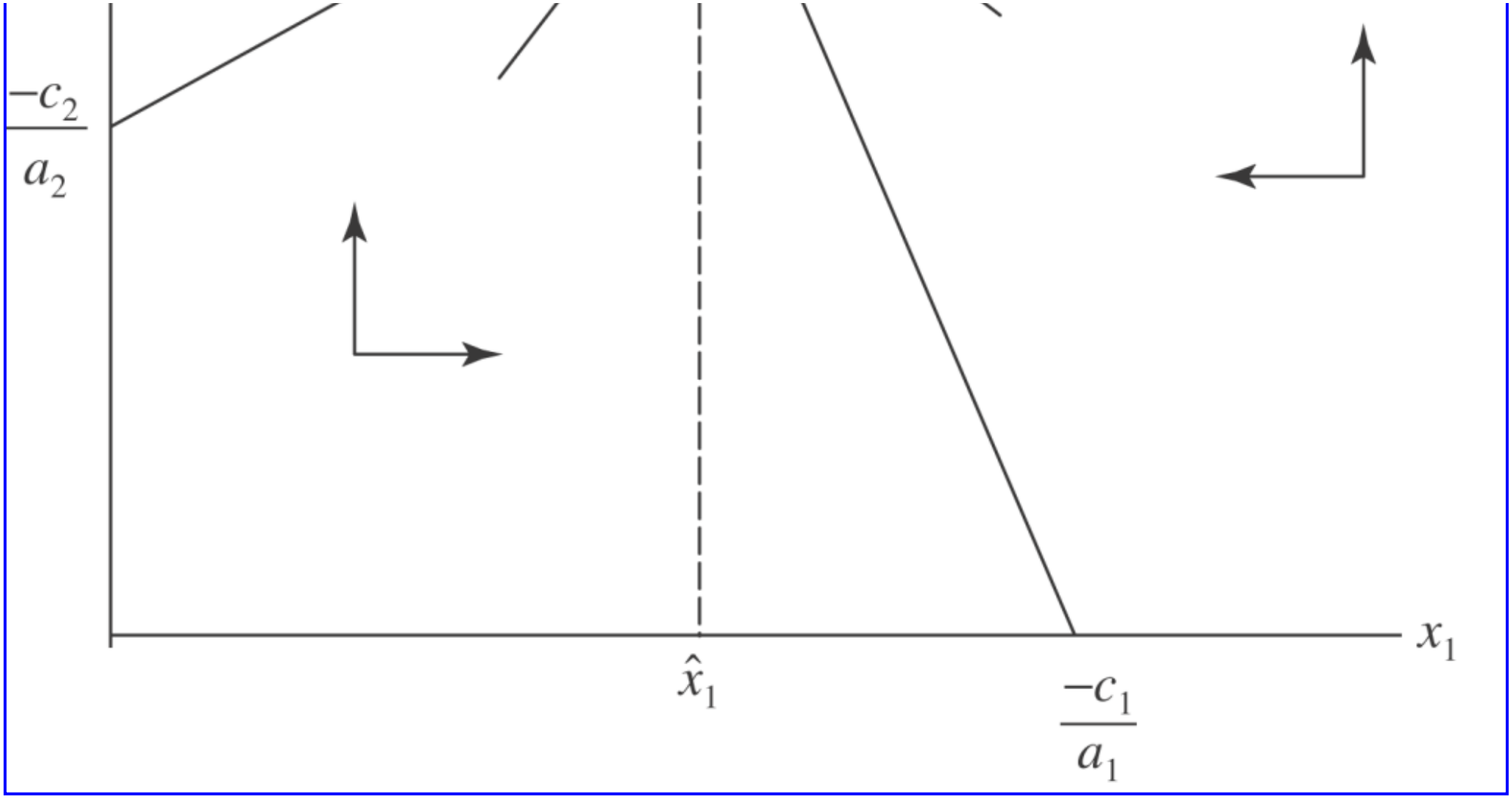

The Richardson paradigm was the central focus of research on arms races. During the Cold War, the build-up of bombers and then missiles by the United States and the Soviet Union was, or should have been, the most important concern as it had the potential of destroying civilization as we know it, if not mankind. This danger was reduced with the end of the Cold War and the later dissolution of the Soviet Union, which ended the US–Soviet arms race. The new environment may not be well characterized by the Richardson paradigm; this article describes the changes and suggests a new approach to the formulation of a model of arms races.

The changing nature of arms races

There have been several major changes in the nature of the arms race since the early 1990s. The most important has clearly been the end of the Cold War. This epochal change began with the demise of the Warsaw Pact in 1989 and ended with the dissolution of the Soviet Union in December 1991. The result has been the end of the global East–West arms race of the Cold War period, when it dominated global politics. Among the implications of this profound change have been drastic reductions in arms expenditures by the member states of the former Soviet Union and its former allies, accompanied by relatively smaller reductions in arms expenditures by the United States prior to the Afghanistan and Iraq wars and by its allies in NATO. As a result, the United States is currently by far the world leader in expenditures on arms, spending almost as much as the rest of the world combined.

Another major change since the mid-1990s has been the substantial increases in arms expenditures by China and its neighbouring states in east and south-east Asia. In China, the reforms that started as a result of Deng Xiaoping's four modernizations of 1978 profoundly changed the course of the country and its economy and society. The last of these four

modernizations was that of the military, which led to the rapid modernization of the Chinese People's Liberation Army (PLA), involving the deployment of newer weapons and major expenditures on arms. The neighbouring nations of east and south-east Asia have reacted to these developments in China by increasing their own arms expenditures. As a result, this region is witnessing major increases in arms, including substantial arms imports.

The India–Pakistan arms race also continues with both qualitative and quantitative arms developments, both nations having demonstrated their nuclear weapons capabilities in tests conducted in May 1998. In both cases, third parties have played an important role. China has shared nuclear and missile technology with Pakistan, and Pakistan, in turn, has been a major actor in the proliferation to nuclear technology to North Korea and Iran.

In the Middle East, the United States has provided Saudi Arabia with weapons, given financial and military assistance to both Israel and Egypt, and has shared anti-missile defence technology with Israel. While Russia can no longer afford to support the former client states of Soviet Union, it appears to be willing to sell weapons technology to any country that can afford it for purely commercial, as opposed to diplomatic or military, purposes.

An important change of recent years has been the appearance of certain newer or evolving regional arms races or arms build-ups. One is the important arms race is that involving the nations of the Gulf, including Iraq, Iran, Syria, Saudi Arabia, Kuwait and the Gulf States, that both was stimulated by and resulted in wars in the region, including the Iran–Iraq war and the Iraqi invasion and annexation Kuwait, resulting in a war to liberate it and the subsequent US-led invasion and occupation of Iraq. The major suppliers of weapons to all parties in the region except Iran are the United States and its European allies. Second, there have also been arms build-ups among the states of the former Soviet Union that are seeking to preserve their independence through their military capabilities. A third type of arms build-up is that in the former Warsaw Pact states of central and eastern Europe that have joined NATO, or hope to do so, and that have to upgrade their weapons capabilities to become members of the alliance.

The major weapons states have played an important role in fuelling these and other regional arms races through arms exports, including the disposal of surplus weapons in the post-cold war period. The United States, Russia, Germany, Britain and France are the leading suppliers of surplus weapons, while Turkey, Greece, Pakistan, Morocco and a number of Middle East countries are the main recipients of such weapons.

Impacts of recent changes on stability

These changes in arms races since the mid-1990s have had important impacts on the stability of both the regional and global systems. As a result of these changes, we believe that there are probably greater instabilities today than those of the earlier Cold War period.

Consider first the principal antagonists of the Cold War. Where there had earlier been two ‘superpowers’, now there is only one as measured by arms expenditures and military capabilities, namely, the United States. Russia has assumed most of the Soviet weapons of mass destruction and the associated responsibilities involved with such weapons. The continued presence of nuclear weapons in Russia and the United States, albeit at lower levels, is probably adequate for mutual deterrence, but there are great dangers inherent in the current unstable political, economic, and social situation in Russia. The result could be a loss of effective control of weapons of mass destruction, with the possibility of an accidental or inadvertent launch of such weapons. The disquieting similarities between Russia today and Germany in the Weimar Republic period between the wars, including loss of empire, inflation, depression and the destruction of the middle class, suggest the possibility of the emergence of a new authoritarian leader in Russia, which would create additional instabilities.

Another major threat to stability at both global and regional levels is the proliferation of weapons of mass destruction. There is now much greater worldwide access to technology and the required material for nuclear, chemical, and biological weapons stemming, in part, from the collapse of the Soviet Union and the desperate situation of its military and scientific establishment. There are also the chains of proliferation that started with the United States and continued with the Soviet Union, the United Kingdom, France, China, India, and Pakistan, and that could continue to other nations, including Iran and other nations of the Gulf region.

Yet another threat to stability in the post-Cold War world is that of terrorists using various weapons of mass destruction. Sub-national groups, motivated by extreme ideologies, religious fanaticism, or other causes, have much greater access to such weapons on world markets. Large urban centres and freedoms of speech, travel, assembly, and the press have made modern societies highly vulnerable to possible terrorist attack. This was clearly demonstrated on September 11, 2001.

Beyond the Richardson paradigm

Until the East–West arms race of the Cold War period, most arms races were naval. Until the 20th century, armies were highly labour-intensive institutions with relatively little capital. Roman soldiers furnished their own equipment until the late Republic. Feudal armies also furnished their own equipment, where the obligation of a fief holder under military tenure was to furnish a certain number of knights and men at arms for a given number of days a year and to provide arms and horses for these men. The key element in deploying military power at that time was the organization of the state and its ability to raise revenue. The possibility of organizing and disciplining free men to serve as heavy infantry was the key to the Greek and Republican Roman armies. Heavy infantry required a body of free men willing to serve. It is very difficult to find examples of heavy infantry manned by professional soldiers except in circumstance where the state had the ability to tax effectively, such as the early Roman Empire and European states after the 16th century.

In hindsight, however, the Richardson paradigm of competitive accumulation of weapons, though important, was limited. The Anglo–German naval race that first attracted Richardson's attention played a very minor role in the First World War. After the indecisive battle of Jutland in 1916, both battle fleets were inactive and the important naval element was the German use of U-boats.

The other important arms race of the 20th century that fits the Richardson paradigm was the nuclear arms race between the United States and the Soviet Union. Fortunately, because of mutual assured destruction, these weapons were never used and the downfall of the Soviet Union was largely the result of the failure of its institutions.

Arms races did not play a major role in the Second World War. British aircraft manufacturers increased the stock of fighter planes during the Battle of Britain. The United States did not fully gear up for a war economy until after Pearl Harbor, and Soviet war production came from factories they moved east of the Urals. Even German production was increasing until the very end of the war.

In recent years technological change has also called into question the Richardson paradigm. Constant or increasing returns to scale have always created difficulties for economic theory. An economy with constant returns to scale is indeterminate with respect to the scale size of firms, and it is necessary to appeal to some fixed factor to determine the size of the economy. Increasing returns to scale leads to monopolies constrained only by demand. Firm behaviour then becomes strategic and none of the standard welfare theorems that hold in competitive markets apply. Thus, it is not surprising that increasing returns to scale in an arms race can lead to very different results than constant or decreasing returns to scale.

Increasing returns to scale in the technology of arms production is more likely to occur with newer types of ‘smart’ weapons that rely heavily on electronics, computers, software, and so forth. In producing weapons with such a large informational component, it is likely that increasing the scale of the production process will make production more efficient. Nations producing arms may sell weapons even when these sales may be contrary to their foreign policy. The drive to lower weapons unit costs through greater sales gives momentum to foreign arms sales that can even conflict with diplomatic or political goals. An example may be the decision of the United States to lift its embargo on arms sales to Latin America at the urging of weapons producers.

Another consequence of technological change is that new technologies have made nuclear weapons and missiles feasible for most nation states, and some of these technologies have valid non-military applications. North Korea with an annual GDP of US$40 billion has acquired nuclear weapons and is ready to test the Taepondong-2, a missile that can reach the United States or, as the North Koreans claim, put a satellite in orbit. Iran is developing the capability of enriching uranium, a capability that can be used to produce fuel or bombs. As of 2006, the developed world is trying to prevent the test of the missile by North Korea and the acquisition of the capability of enriching uranium by Iran. Technological change has forced the developed world into the position of trying to deny countries in the developing world technologies that the developed world possesses and that have plausible non-military use.

As discussed above, social capital has been a very important element in the ability of a state to mobilize its resources and project power. Social capital includes not only the tangible institutions that the state has to tax, to conscript and to mobilize resources, but also less tangible institutions such as the relations of the members of the state to each other and to the state. States with sharp class, ethnic or caste distinctions may find it difficult to mobilize effectively to project force. During the American Civil War, the institution of slavery kept the South from mobilizing the members of its population that were black, and gave President Lincoln the political advantage of defining the war to be against the institution of slavery. In present day Iraq, ethnic differences have made it very difficult to organize an Iraqi national army.

Among the components of social capital are the common values of the society and its institutions. One important element of social capital familiar to most economists, but largely neglected in the arms race literature, is the attitude of the society towards risk and uncertainty. One very important question is how a society views a lottery that will cost a specific member of society his or her life with certainty to be equivalent to a lottery in which 1,000 individuals face one chance in a thousand of dying. It has long been noted by scholars in such fields as public finance, law, and economics that people in the United States are willing to spend more resources to save a specific individual than an individual who is a statistical abstraction. This element of social capital is reflected in how the United States conducts war, but it is not shared by other cultures.

There is widespread use of suicide bombers in current conflicts in Palestine and Iraq. Although this is a new phenomenon in recent history, most of the elements are not new. In the Second World War Japan sent young pilots on kamikaze suicide missions while the United States was willing to send bomber crews over Germany knowing that few would survive and there would be civilian casualties. The probability that a bomber crew would survive a full tour of duty was small. There may be some substantive difference between the Palestinians being willing to send a young man to kill himself to induce terror among the Israelis and the Doolittle raid where 16 bombers attacked the Japanese home islands in 1942 for psychological purposes; but it seems that the difference is that, whereas Western cultures are willing to sacrifice individuals for the common good as long as the sacrifice is a lottery, some other cultures are willing to sacrifice specific individuals. This difference changes the war-making potential of the different cultures.

To illustrate with another example, the Japanese supply of trained pilots was seriously depleted during the battle of Midway in 1942 and subsequent naval engagements. The Japanese were not able to compete with the Americans in training new pilots. By the Marianas campaign the Japanese were no match for the Americans, and the Japanese resorted to using untrained pilots as kamikazes to attack the American fleet. This example illustrates the role of various forms of social capital in war. The more open and egalitarian American society allowed the United States to train pilots as it had a larger pool to draw from than the more structured and hierarchical Japanese society. However, the advantage of this type of American social capital was offset in part by the fact that Japanese society was willing to sacrifice specific individuals. American pilots were better trained and had more human capital; the willingness of Japanese society to sacrifice specific Japanese pilots was a different form of social capital.

Richardson's world was one in which dreadnoughts and battlecruisers would steam into battle planned by admirals who had studied Admiral Thayer Mahan (1840–1914, US naval officer and geostrategist who was influential on the US building a modern naval fleet, acquiring overseas naval bases, and building the Panama Canal) and other theorists. The US–

Soviet arms race was also a very intellectual process that was based on very sophisticated doctrines and involved weapons systems that were highly quantifiable. The conflicts we now face, by contrast, are very different. They involve state and non-state entities, and the means of deploying force are highly asymmetrical. Fighter planes carrying GPS guided bombs are used against terrorists who employ suicide bombers and can use the internet to transmit pictures of the decapitation of prisoners. Modelling such phenomena is the task for the next generation. What we propose to do is offer a conjecture as to the nature of such processes.

A conjecture on arms race theory

Assume that the war-making potential of the *i*-th country can be described by a vector of physical, human, intellectual and social capital, $\mathbf{k_i}$ and a vector of strategies $\mathbf{v_i}$. Its war-making potential, x_i, is given by

$$x_i = \max_{v_i} \theta(\mathbf{k_i}, \mathbf{v_i}) \tag{2}$$

where the cost of the strategies and other tradeoffs is reflected in the social capital. We conjecture that the intertemporal optimization results in a differential equation of the form

$$\dot{x}_1 = a_1(x_1) + b_1(x_2) + c_1(x_1, x_2)$$
$$\&$$
$$\dot{x}_2 = a_2(x_2) + b_2(x_1) + c_2(x_1, x_2) \tag{3}$$

The first term $a_i(x_i)$ reflects the role of the *i*-th country's war-making potential on the rate of growth of *i*-th country's war-making potential. In the Richardson model the derivative of this term is negative as it represents the fatigue term. In this model it could well be positive as many of the components of the war-making potential – social, intellectual and human capital – are productive. The second term $b_i(x_j)$ reflects the role of the *j*-th country's war-making potential on the rate of growth of *i*-th country's war-making potential. This is analogous to the defence term in the Richardson model. As in the Richardson model, this term is positive. In this model such an assumption is made for two reasons. First, as in the Richardson model, an increase in the war-making potential of the *j*-th country will be viewed as a threat. Second, and perhaps more important, some of the inputs in the production of x_j, particularly intellectual capital and social capital, are public goods and can be transferred to the competing country. Meiji Japan acquired from the West the technology to build warships and organize a modern navy, and at the present time the technology the North Koreans are using to build nuclear bombs can be traced from the United States through various intermediaries to China, to Pakistan and then to North Korea. The problem of technological transfer is more difficult to control when it is dual use, that is, could be used for civilian as well as military purposes. After all, the Taepondong-2 *could* be used to launch weather satellites. The term, $c_i(x_i, x_j)$ is different from the grievance term in the Richardson model in that it represents the competition of the parties for resources or perhaps even ecological space, and is assumed to be quadratic in order. The derivative is assumed to be positive. If we consider the equation

$$\dot{x}_1 = a_1(x_1) + b_1(0) + c_1(x_1, 0) \tag{4}$$

and if $a_1(0) = 0$ and $b_1(0) = 0$, we would assume that eq. (4) would behave in a way similar to a biological population growth equation (Figure 2).

Figure 2

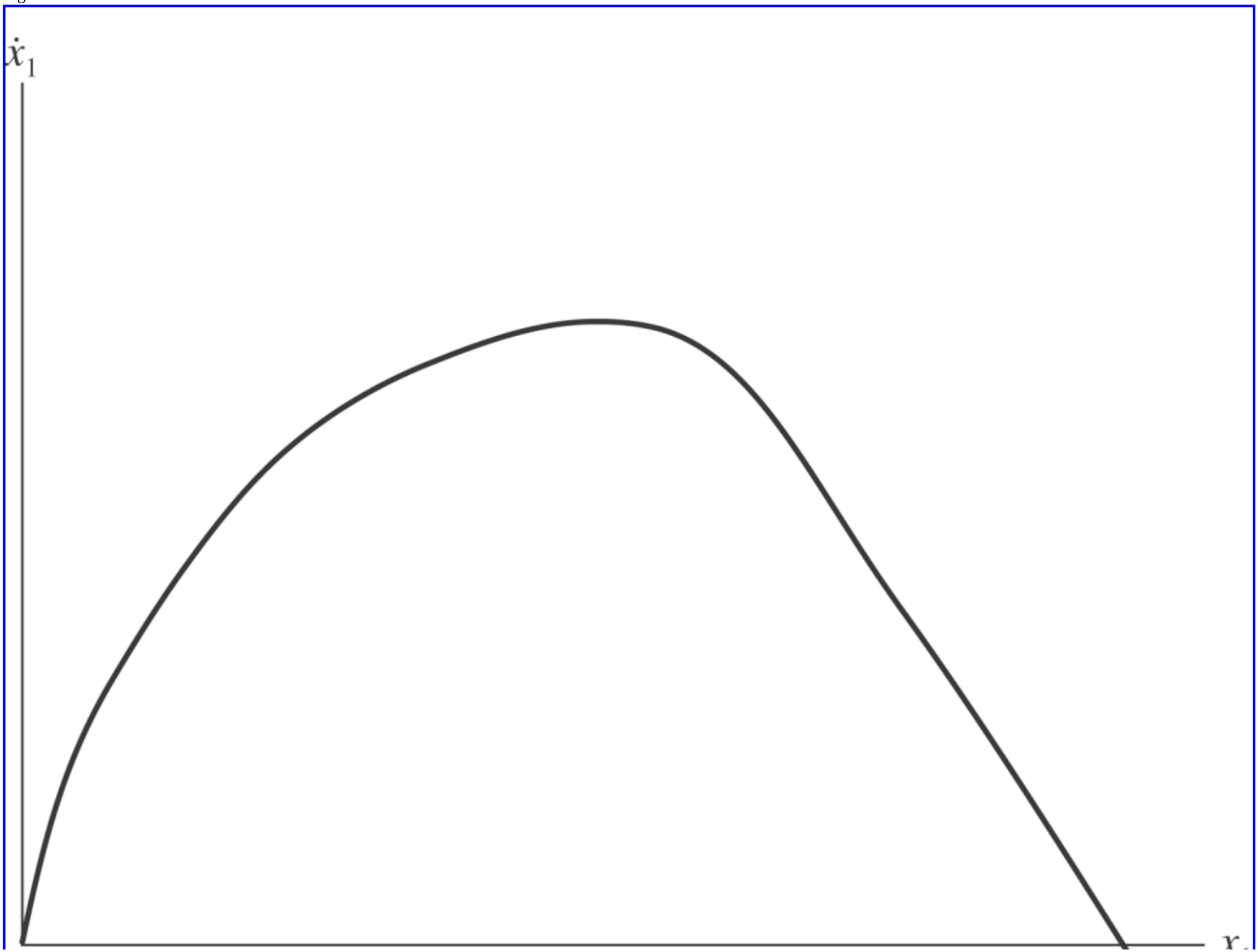

$\bar{x}_1$ is the maximum potential size of Country 1 in the absence of competition. A linear approximation of eq. (3) is given by

$$\dot{x}_1 = a_1 x_1 + b_1 x_2 + c_1 (x_1 + x_2)^2$$
$$\&$$
$$\dot{x}_2 = a_2 x_2 + b_2 x_1 + c_2 (x_1 + x_2)^2 \quad (5)$$

This is similar to the Richardson equation except for the quadratic term of the common resource constraint. If we assume that a_i, the 'fatigue term', is negative, b_i, the 'defence term', is positive and c_i, the 'resource term', is negative, then we can represent the dynamics of this nonlinear system in the phase diagram in Figure 3.

Figure 3

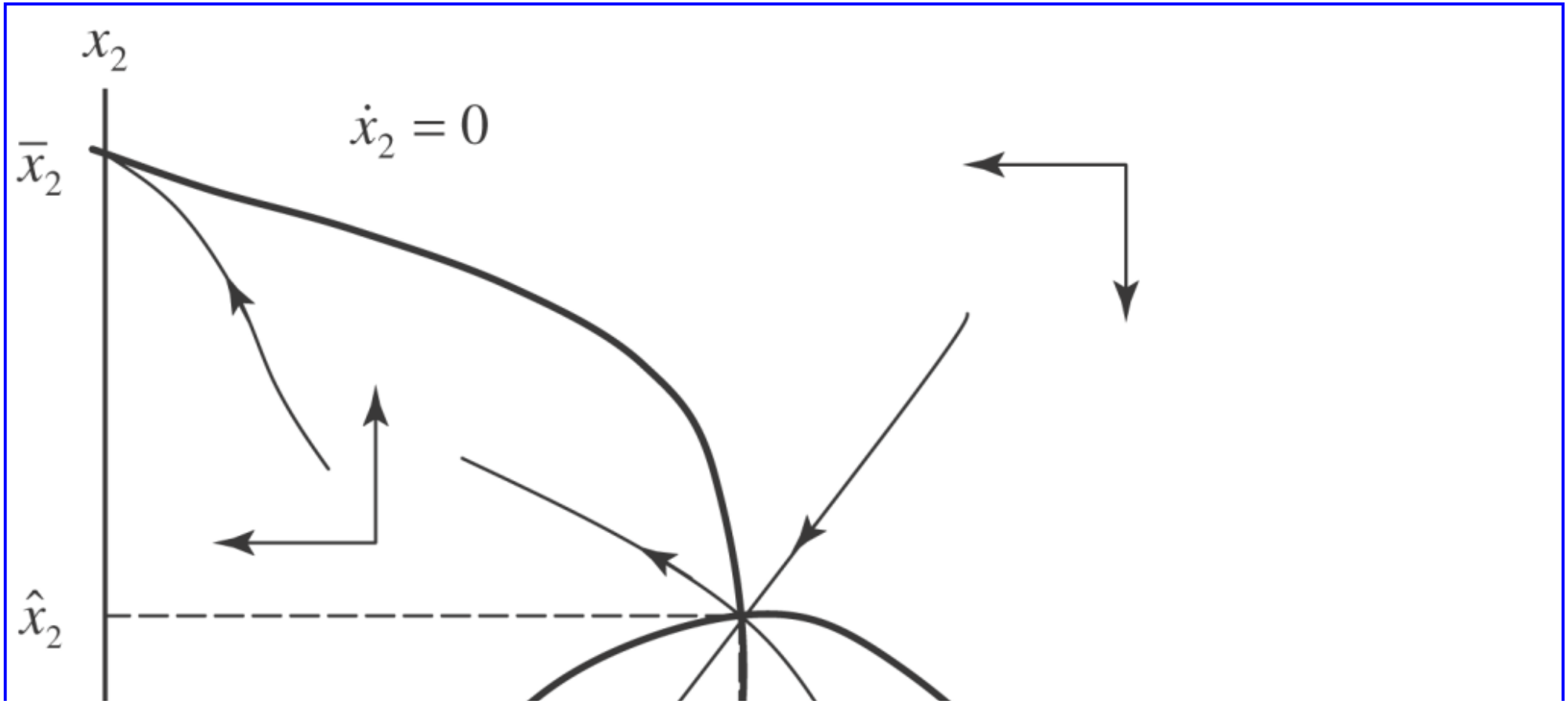

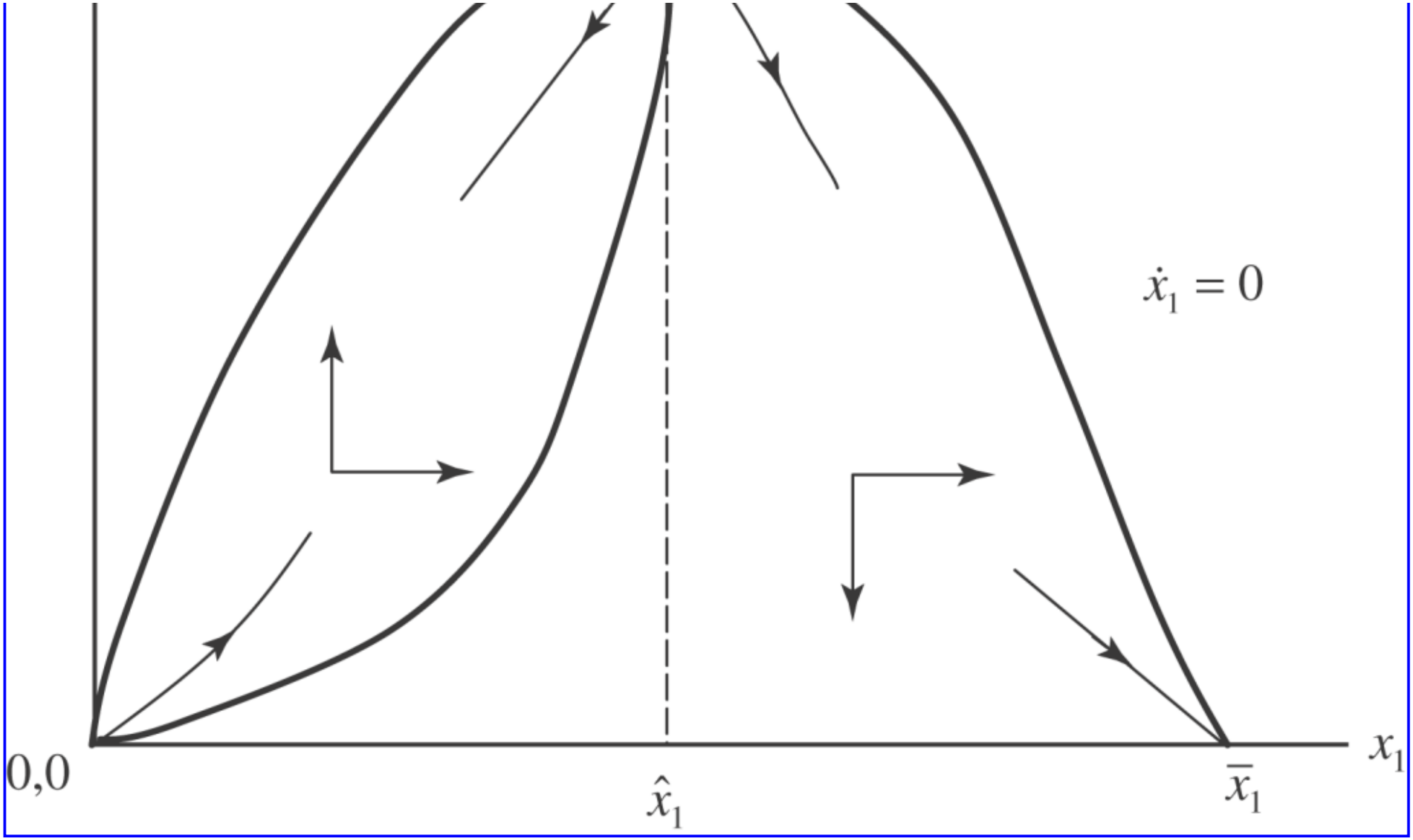

Although on the surface this appears to be very similar to the Richardson equation, the variable x_i is war-making potential that is the result of a prior optimization. One of the elements of the prior optimization is social capital, which includes among its elements moral values.

The differential equation system has four equilbria, of which two are stable and two are unstable. The two that are stable, $(\bar{x}_1, 0)$ and $(0, \bar{x}_2)$, involve the elimination of one of the parties. Whether this is good or bad depends on the process of the optimizations underlying the dynamical system. Recall that one of the important components of the process is social capital. One realization could be that the social capital of the competing parties would evolve in such a fashion as to eliminate conflict. An example is the transformation of the nation states of Europe, with a thousand-year history of wars, into the European Union. A second, less optimistic, scenario is the complete destruction of the weaker party. Again, the crucial element is social capital. Initially, the weaker power many threaten the stronger power by using tactics that are not acceptable to the values of the stronger power – for example, the use of suicide bombers. However, civilization has a thin veneer. Historically, if a country feels that its survival or vital interests are at stake, it will quickly shed its inhibitions. The tactics the British used to suppress the sepoy mutiny were brutal. At Peshawar, 40 sepoys were stood before cannons and blown apart in a public execution. The countries that condemned the German bombing of Guernica in the Spanish Civil War (fewer than 2,000 casualties) firebombed Hamburg (50,000 casualties), Dresden (25,000–35,000 casualties)

and Tokyo (100,000 casualties) in the Second World War, and ultimately used atomic weapons on Japanese cities. Before the start of the Gulf War of 1991, US Secretary of State James A. Baker III warned Iraq that the use of weapons of mass destruction by Iraq would result the destruction of Iraq as a modern state.

The third alternative is decoupling. This results in a stable equilibrium (see Figure 4). The French in Algeria, the United States in Vietnam and the Soviet Union in Afghanistan withdrew because the game was not worth the candle. The partition that appears to be imminent in Palestine, where Israel is building a wall to minimize its interaction with the Palestinians, may be an omen of things to come. If interaction with the developing world becomes too costly, the developed world has the alternative of disengaging. Without oil, the Middle East would be no more important than Africa, and conflicts between the Sunnis and Sh'ias would receive the same attention as conflict between the various African tribes. At prices greater than US$45.00 a barrel, technologies exist for the developed world to be self-sufficient in oil. History could repeat itself. An argument can be made that the Muslim world started to decline in the 16th century partly because the opening of alternative trade routes to Asia destroyed the Muslim monopoly on such trade.

Figure 4

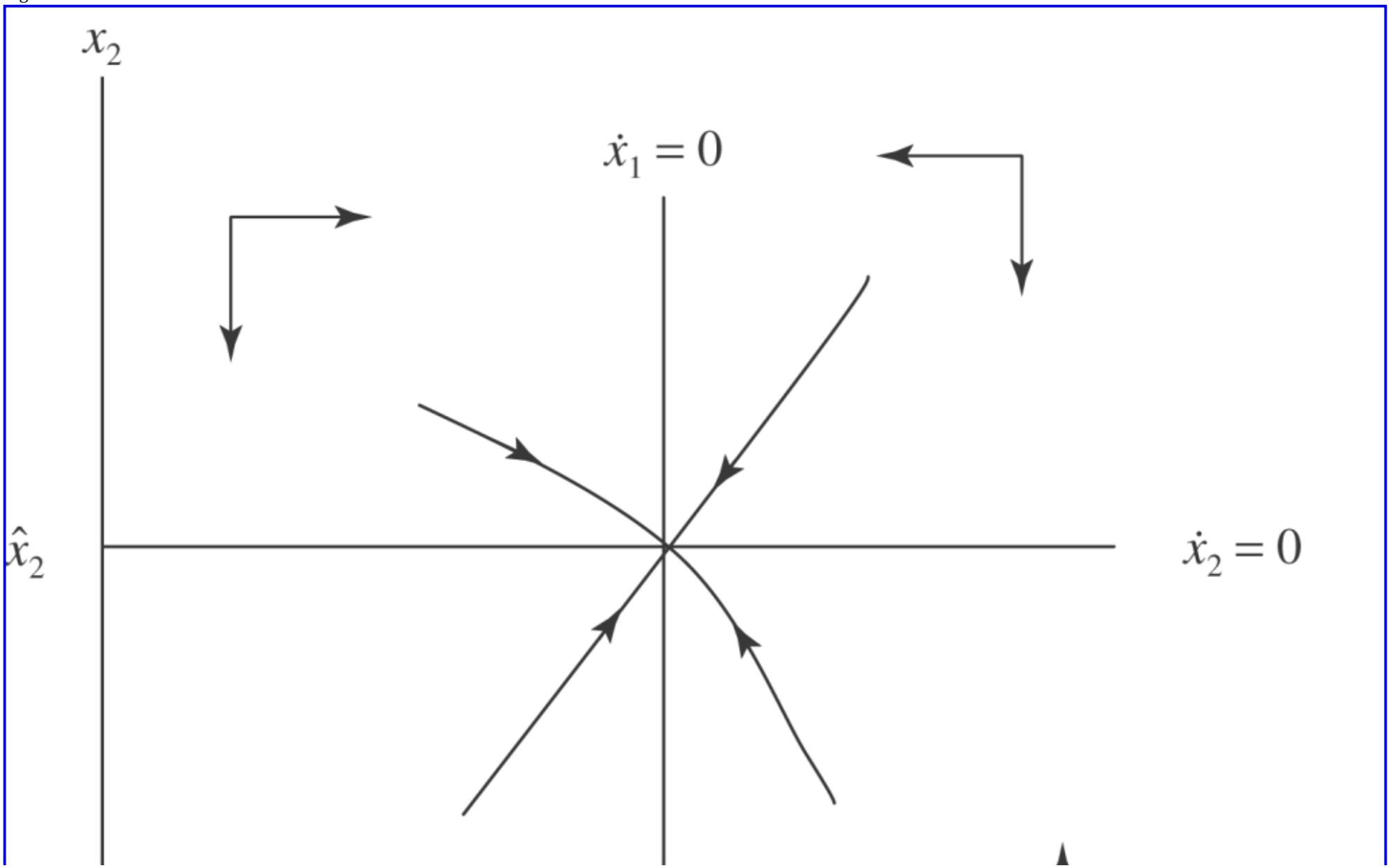

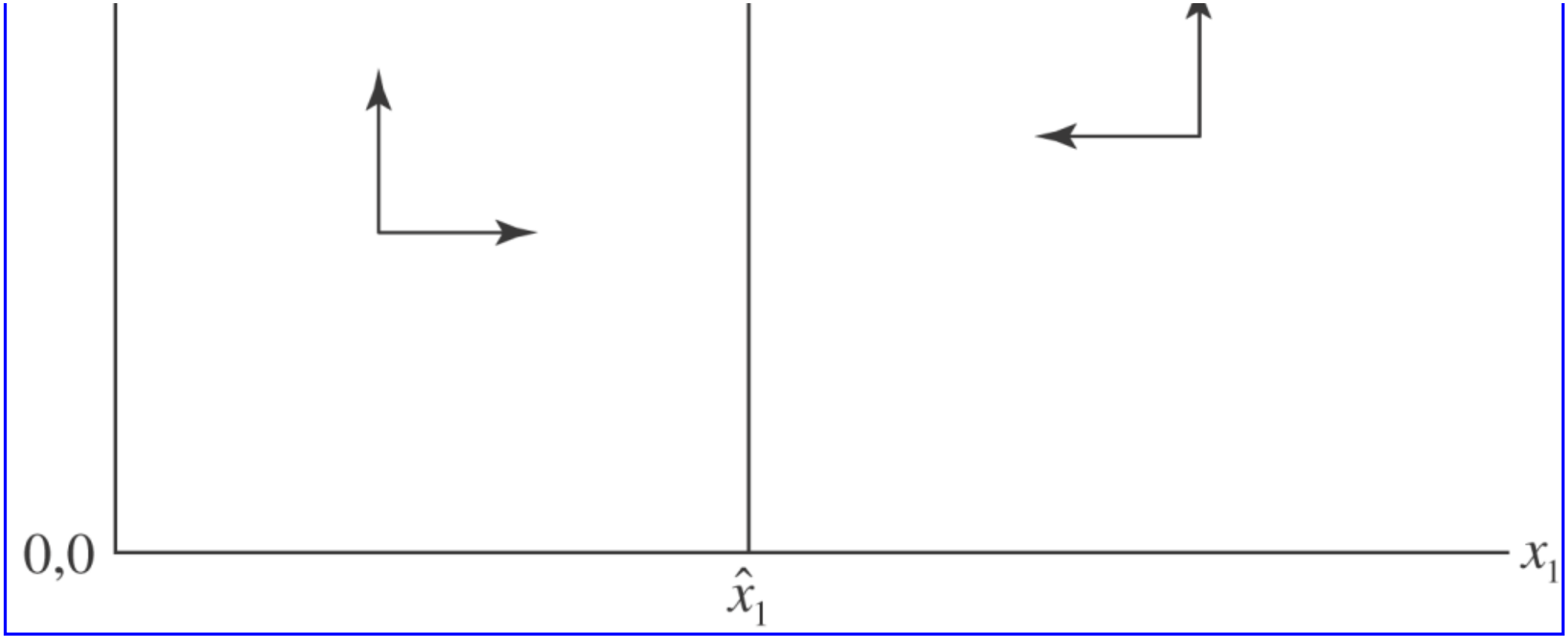

Conclusions

The arms race as described by the Richardson paradigm, where nation states arm in a competitive fashion, is a phenomenon that starts with the naval arms race at the end of the 19th century and may have ended along with the US–Soviet arms race after the dissolution of the Soviet Union. Before that time, warfare was not very capital-intensive, and the most important elements in the projection of military power were the human capital of the population and the social capital that enabled countries to mobilize their resources in war. Richardsonian arms races reflect competition that is constrained by economic resources. Recent developments in technology have broken that link. Technological change has made it possible for a country like North Korea, with an annual GNP of US$40 billion, to acquire nuclear weapons and a missile that may be capable of attacking the United States. The link between economic power and the ability to project military power has been broken. The Richardson paradigm no longer applies. We conjecture the structure of an alternative model. This model suggests three alternatives: cultural convergence, destruction of the weaker party, and decoupling of the conflict. It should be clear that the model is a conjecture based on our intuition, and much work is needed to develop the theoretical foundations of the next arms race paradigm.

See Also

- arms trade
- defence economics
- war and economics

Bibliography

Anderton, C. 1985. A Selected bibliography of arms race models and related subjects. *Conflict Management and Peace Science* 8, 99–122.

Anderton, C. 1989. Arms race modeling: problems and prospects. *Journal of Conflict Resolution* 33, 346–67.

Boulding, K. 1978. Future Directions in Conflict and Peace Studies. *Journal of Conflict Resolution* 22, 342–54.

Brito, D. 1972. A dynamic model of an armaments race. *International Economic Review* 13, 359–75.

Brito and Intriligator, M. 1995. Arms races and proliferation. In *Handbook of Defence Economics*, ed. K. Hartley and T. Sandler. Amsterdam: North-Holland.

Brito, D. and Intriligator, M. 1996. Proliferation and the probability of war: a cardinality theorem. *Journal of Conflict Resolution* 40, 204–12.

Brito, D. and Intriligator, M. 1999. Increasing returns to scale and the arms race: the end of the Richardson paradigm? *Defence and Peace Economics* 10, 39–54.

Downs, G. and Rocke, D. 1990. *Tacit Bargaining, Arms Races, and Arms Control*. Ann Arbor: University of Michigan Press.

Gleditsch, N. and Njølstad, O., eds. 1990. *Arms Races, Technological and Political Dynamics*. London: Sage Publications.

Intriligator, M. 1975. Strategic considerations in the Richardson model of arms races. *Journal of Political Economy* 83, 339–53.

Intriligator, M. 1982. Research on conflict theory: analytic approaches and areas of application. *Journal of Conflict Resolution* 26, 307–27.

Intriligator, M. and Brito, D. 1984. Can arms races lead to the outbreak of war? *Journal of Conflict Resolution* 28, 63–84.

Isard, W. 1988. *Arms Races, Arms Control and Conflict Analysis*. New York: Cambridge University Press.

Isard, W. and Anderton, C. 1985. Arms race models: a survey and synthesis. *Conflict Management and Peace Science* 8, 27–98.

Nicholson, M. 1999. Review article: Lewis Fry Richardson and the study of the causes of war. *British Journal of Political Science* 29, 541–63.

Richardson, L. 1960. *Arms and Insecurity; A Mathematical Study of the Causes and Origins of War*. Pittsburgh: Boxwood Press.

How to cite this article

Brito, Dagobert L. and Michael D. Intriligator. "arms races." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000129> doi:10.1057/9780230226203.0055

The New Palgrave Dictionary of Economics Online

arms trade

Charles H. Anderton and John R. Carter

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Arms trade is the transfer of weapons systems, components, technologies, and services across national and territorial borders. Contemporary arms trade occurs in three product categories: major conventional weapons; small arms and light weapons; and weapons of mass destruction. This article briefly surveys the theoretical and empirical arms trade literature. Topics include arms trade data sources, commercial and security motives for weapons exports, competitive and imperfectly competitive models of arms trade, empirical studies of the economic and political effects of arms trade, and arms export controls.

Keywords

arms trade; arms exports; arms export restraints

Article

Arms trade is the transfer of weapons systems, components, technologies and services across national and territorial borders. Contemporary arms trade occurs in three product categories: major conventional weapons (MCW), such as fighter aircraft and destroyers; small arms and light weapons (SALW), such as assault rifles, machine guns and improvised explosive devices; and weapons of mass destruction (WMD), such as nuclear, biological and chemical weapons technologies and long-range missile systems. MCW are the dominant form of weapons in interstate wars, while SALW are used intensively by non-state actors in intra-state wars (for example, civil wars) and extra-state conflicts (for example, transnational terrorism). WMD components and technologies proliferate by spreading to states or possibly non-state actors via trade or indigenous production.

Major sources of arms trade data include the US Congressional Research Service for all categories of weapons and arms-related services to developing nations; the Stockholm International Peace Research Institute for MCW; the Norwegian Initiative on Small Arms Transfers and the Graduate Institute of International Studies (Geneva) Small Arms Survey for SALW; and the Monterey Institute's Center for Nonproliferation Studies for WMD proliferation. These sources indicate that, of the world's total arms

exports, more than one-half originates in the United States and Russia, and close to two-thirds goes to developing nations (Brauer, 2007).

Theories of arms trade have shifted in emphasis over time. Pre-Cold War literature emphasized economic motives, often from a condemnatory ‘merchants of death’ perspective (see, for example, Engelbrecht and Hanighen, 1934). During the Cold War, classic texts focused on domestic and international politics, with some coverage of economic incentives (see, for example, Pierre, 1982). Post-Cold War models of arms trade highlight both commercial and security concerns. For example, in Levine and Smith's (1995) model, a few suppliers export weapons to a large number of price-taking buyers who are involved in dyadic arms rivalries. Suppliers’ utility depends on security and producers’ profits, while recipients’ utility depends on security and consumption. Under certain conditions, commercial gains to arms exporters are offset by security losses because the arms exports create a greater risk of war among recipients. Under other conditions, arms exports reduce war risk, implying both commercial and security gains to suppliers from weapons exports.

Theoretical models of international trade and industrial organization often apply to arms trade (see, for example, Anderton, 1996). Competitive models are useful for the study of SALW trade because such weapons are relatively homogeneous and the number of buyers and sellers is large. For MCW and WMD, the number of suppliers is relatively small and products within weapons classes are differentiated. For these weapons, models incorporating economies of scale, technological differences, intermediate products and strategic behaviour are more appropriate.

Some empirical studies investigate the determinants of arms trade (for example, Smith and Tasiran, 2005), but most focus on economic and political effects, including the impact on employment, growth and development, arms rivalries, and human rights (see, for example, Grober, Stern and Deardorff, 1990; Yakovlev, 2005; Sanjian, 1999; and Blanton, 1999). Perhaps the most important empirical relationship considered is the effect of arms trade on the risk of war. Craft and Smaldone (2003) report that arms imports significantly increase the risk of interstate or intrastate conflict for sub-Saharan African nations. Krause (2004) finds that arms transfers that occur outside of defence pacts increase the risk that recipients will become involved in militarized interstate disputes. Most other studies likewise find that arms exports increase the risk of conflict, but there are exceptions (see Anderton, 1995).

Arms exports are typically subject to extensive government influence. Arms trade offsets require an exporting firm to use some of the revenue from arms sales to invest in activities in the importing nation. Brauer and Dunne (2004) report that there is little empirical or case study evidence that arms trade offsets enhance economic development. Some interventions, like subsidies and diplomatic lobbying on behalf of weapons firms, enhance arms exports. Virtually all governments limit arms exports to particular recipients, and various multilateral arms export limitation regimes exist including the Wassenaar Arrangement, the EU Code of Conduct on Arms Exports, the Nuclear Suppliers Group, the Missile Technology Control Regime, and the Australia Group. Brzoska (2004) argues in favour of a multilateral arms export tax in order to reduce arms exports.

Because production and trade are jointly determined economic activities, arms export restraints cannot be understood in isolation from arms production (Brauer, 2000). In a competitive market model, reduction of weapons supply through production or export controls can raise the equilibrium world price, creating an incentive for new arms suppliers to enter the market or existing suppliers to

circumvent the controls. This suggests that a reduction in weapons demand or an increase in the cost structure of weapons firms is necessary to reduce the number of weapons in the international system in the long run (see, for example, Anderton, 1996; Brauer, 2000). In Levine and Smith's (1995) imperfect competition model, arms export restraints can benefit suppliers by raising prices and also reduce inefficiencies associated with recipients' arms rivalries. On the assumption that arms sales are taxed, proceeds could be distributed to recipients so that the control regime would Pareto-dominate the outcome with no controls. Such a regime would, however, be vulnerable to cartel-like defections of individual suppliers.

Arms trade involves many direct and indirect economic and political costs and benefits, which suggest a number of broad research themes going forward. First, for the sake of tractability, partial equilibrium analyses of arms trade determinants and effects will continue to dominate the literature. Second, general equilibrium perspectives are beginning to emerge which promise a richer assessment of the nature and effects of arms trade and arms export restraints (see, for example, Levine, Sen and Smith, 2000). Third, efforts by governments, NGOs, and multilateral organizations to implement Pareto-improving arms trade policies require collective action solutions (Sandler, 2000).

See Also

- arms races
- war and economics

Bibliography

Anderton, C.H. 1995. Economics of arms trade. In *Handbook of Defense Economics*, vol. 1, ed. K. Hartley and T. Sandler. Amsterdam: North-Holland.

Anderton, C.H. 1996. What can international trade theory say about the arms trade? *Peace Economics, Peace Science, and Public Policy* 4, 7–30.

Blanton, S.L. 1999. Instruments of security or tools of repression? Arms imports and human rights conditions in developing countries. *Journal of Peace Research* 36, 233–44.

Brauer, J. 2000. Potential and actual arms production: implications for the arms trade debate. *Defence and Peace Economics* 11, 461–80.

Brauer, J. 2007. Arms industries, arms trade, and developing countries. In *Handbook of Defense Economics*, vol. 2, ed. T. Sandler and K. Hartley. Amsterdam: North-Holland.

Brauer, J. and Dunne, J.P., eds. 2004. *Arms Trade and Economic Development: Theory, Policy, and Cases in Arms Trade Offsets*. New York: Routledge.

Brzoska, M. 2004. Taxation of the global arms trade? An overview of the issues. *Kyklos* 57, 149–72.

Craft, C. and Smaldone, J.P. 2003. Arms imports in Sub-Saharan Africa: Predicting conflict involvement. *Defence and Peace Economics* 14, 37–49.

Engelbrecht, H.C. and Hanighen, F.C. 1934. *Merchants of Death: A Study of the International Armament Industry*. New York: Dodd, Mead.

Grobar, L.M., Stern, E.M. and Deardorff, A.V. 1990. The economic effects of international trade in armaments in the major Western industrialized and developing countries. *Defence Economics* 1, 97–120.

Krause, V. 2004. Hazardous weapons? Effects of arms transfers and defense pacts on militarized disputes, 1950–1995. *International Interactions* 30, 349–71.

Levine, P., Sen, S. and Smith R., eds. 2000. Special issue: Arms exports, controls and production. *Defence and Peace Economics* 11, 443–548.

Levine, P. and Smith, R. 1995. The arms trade and arms control. *Economic Journal* 105, 471–84.

Pierre, A.J. 1982. *The Global Politics of Arms Sales*. Princeton, NJ: Princeton University Press.

Sandler, T. 2000. Arms trade, arms control, and security: collective action issues. *Defence and Peace Economics* 11, 533–48.

Sanjian, G.S. 1999. Promoting stability or instability? Arms transfers and regional rivalries, 1950–91. *International Studies Quarterly* 43, 641–70.

Smith, R.P. and Tasiran, A. 2005. The demand for arms imports. *Journal of Peace Research* 42, 167–81.

Yakovlev, P. 2005. Do arms exports stimulate economic growth? Ph.D. thesis. Department of Economics, West Virginia University.

How to cite this article

Anderton, Charles H. and John R. Carter. "arms trade." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000232> doi:10.1057/9780230226203.0056

The New Palgrave Dictionary of Economics Online

Arndt, Heinz Wolfgang (1915 – 2002)

R.D. Freeman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Arndt, H. W.; Indonesia; Keynesianism; planning

Article

Born February 1915 in Breslau, Germany (now Wroclaw, Poland), Arndt was educated at Oxford University (1933–8) and London School of Economics (1938–41). After two years as a research assistant at the Royal Institute of International Affairs, Arndt was Assistant Lecturer in Economics, University of Manchester (1943–6), Senior Lecturer, University of Sydney (1946–50) and then Professor of Economics in the School of General Studies and Research School of Pacific Studies, Australian National University (1951–80). He became Emeritus Professor of Economics, Australian National University, in 1981. His many prestigious appointments include Member, Governing Council, United Nations Asian Institute for Economic Development and Planning (1969–75); Deputy Director, OECD (1972) and Chairman, Expert Group on Structural Change and Economic Growth Commonwealth Secretariat (1980).

Arndt first came to prominence in 1944 with his analytical economic analysis of the interwar period in which he argued the structuralist thesis that market forces could not correct the existing major disequilibria in the world economy. He recommended cooperative planning in the post-war period, involving controls on the volume and directions of international trade and investment and international cooperation if not supranational economic authorities.

His major contributions were in policy-oriented economic research with particular reference to developing countries in the Pacific Basin. A leading authority on the Indonesian economy as well as other Asian economies, led Arndt to start the *Bulletin of Indonesian Economic Studies* in 1965; he was also instrumental in the establishment in the Australian National University of a major research school on Asian economic development.

A prolific writer, Arndt was an important influence in Australian academic and policy circles in developing post-war understanding and acceptance of Keynesian macroeconomic analysis.

Selected works

1944. *The Economic Lessons of the Nineteen-Thirties*. Oxford: Oxford University Press. Reprinted London: Frank Cass & Co., 1963; Italian trans., 1949; Japanese trans. 1978.

1954. A suggestion for simplifying the theory of international capital movements. *Economia Internazionale* 7, 469–81.

1955. External economies in economic growth. *Economic Record* 31(61), 192–214.

1957. *The Australian Trading Banks*. Melbourne: Cheshire 2nd edn, 1960; 3rd edn (with C.P. Harris), 1965; 4th edn (with D.W. Stammer), 1973; 5th edn (with W.J. Blackert), Melbourne: Melbourne University Press, 1977.

1978. *The Rise and Fall of Economic Growth: A Study of Contemporary Thought*. Melbourne: Longman Cheshire.

1979. The modus operandi of protection. *Economic Record* 55(149), 149–55.

1981. Economic development: a semantic history. *Economic Development and Cultural Change* 29, 457–66.

1985. *A Course Through Life*. Canberra: Australian National University.

How to cite this article

Freeman, R. D. "Arndt, Heinz Wolfgang (1915–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000131> doi:10.1057/9780230226203.0057

The New Palgrave Dictionary of Economics Online

arrears

Mark E. Schaffer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Arrears, in both common and general economic parlance, are overdue payments of any sort. The comparative economics literature has focused on the large-scale arrears of all sorts that emerged when central and eastern Europe and the former USSR began the transition to market economies. Soft budget constraints have been invoked in explanations of the growth of overdue trade credit or ‘inter-enterprise arrears’ in early transition, and in analyses of arrears to banks and tax arrears; studies of wage arrears in transition economies have focused on differential impacts across workers and firms and on weak institutions.

Keywords

arrears; inter-enterprise arrears; soft budget constraint; tax arrears; trade credit arrears; transition economies; wage arrears

Article

Arrears, in both common and typical economic parlance, are overdue payments of any sort. In its last previous appearance in this dictionary, *Palgrave's Dictionary of Political Economy*, the term is defined simply as ‘sums remaining unpaid after they are due’ (Higgs, 1925, p. 58). The context is usually one in which a payment is required by a contract or by law; hence the cross-references in the 1925 *Palgrave* entry to ‘law of contract’ and ‘wages’. The same is true of contemporary usage: Internet search engines at the time of writing indicate the most commonly used term by far is ‘mortgage arrears’, followed by ‘wage arrears’ and ‘tax arrears’. In most of economic science, arrears are generated by some behaviour or event, and it is the latter which is typically the focus of analysis. The analytical framework varies hugely with the object of analysis, and there is no theme that unites, for example, the analysis of consumer debt arrears and that of sovereign debt arrears.
The main exception to this, and the reason ‘arrears’ has reappeared in the *New Palgrave*, is the arrears

phenomenon that emerged on a large scale when the countries of central and eastern Europe and the former USSR abandoned the socialist economic system and began the transition to market economies. The arrears phenomenon in transition economies arose when firms accumulated non-payments of obligations to various creditors, often on a very large scale. The natural way to analyse this phenomenon is to distinguish between the main categories of creditors to the firms that have accumulated arrears – other firms, banks, the state and employees – and between stocks and flows, late payments and non-payments.

In the comparative and transition economics literature, overdue debts of firms to other firms has often been termed ‘inter-enterprise arrears’, though a more standard term from mainstream economics would be ‘trade credit arrears’. The rapid emergence of large volumes of overdue trade credit in many formerly socialist countries in the early phase of the transition (1989–93) took many economists in both the policy and academic communities by surprise. In retrospect, this surprise partly reflects the fact that trade credit is an understudied phenomenon in general. After early, rapid growth, the volumes of both total trade credit and overdue trade credit in the transition economies stabilized at levels similar to those found in developed market economies – the equivalent of roughly 20–40 per cent and 10–20 per cent of GDP, respectively (Schaffer, 1998). The eventual stabilization at levels found in normal market economies implies an approximate matching of inflows and outflows, and follows partly from the fact that late payment of trade credit is an endemic problem in market economies generally, as a reading of the business press and reports by factoring agencies will confirm. It also implies that firms in transition economies, including state-owned firms that had previously been unexposed to market forces, learned fairly rapidly to impose hard budget constraints on each other.

The early phase of rapid growth of trade credit arrears is a somewhat different matter. First, the payment systems that were used in socialist economies were typically very inflexible. Ickes and Ryterman (1992) argue that in Russia, the most studied country case of trade credit arrears, the combination of a lack of liquidity following price liberalization in January 1992 and a first-in-first-out (FIFO) queuing system for clearing payments generated ‘payments gridlock’ and thus rapid growth in arrears on payments to suppliers. The government's response in mid-1992 was to abandon the payment queuing system and, separately, to try to clear the accumulated backlog of payments with an accompanying injection of credit, amounting to a bail-out of the enterprise sector. Second, the model of Perotti (1998) suggests that collusive non-payment by the enterprise sector can force a government bail-out via a ‘too-big-to-fail’ mechanism. Both explanations are examples of soft-budget constraints in action. This early phase of rapid growth also took place in the moderate- to high-inflation environments that followed price liberalization in these countries. The effective interest rate subsidy that accompanied trade credit thus involved a substantial discount to buyers, though it has also been suggested that sellers anticipated both inflation and payment delays, and incorporated a corresponding markup in their prices.

Arrears of firms to banks in transition economies is the phenomenon that is least specific to the transition experience. The large bad-debt problems that emerged following the start of transition have been analysed in the literature using the standard frameworks and tools for analysing systemic banking-sector problems. The limited evidence from these economies suggests that connected lending and directed state credits became a primary mechanism in the slower reformers for bailing out firms and softening budget constraints into the 1990s and beyond. Large-scale tax arrears of firms, by contrast, are peculiar to the transition experience. In developed market economies, tax arrears of firms are a phenomenon largely associated with exit of insolvent firms, and the scale is relatively small; New

Zealand has been cited as an example, with a stock of tax arrears amounting to one or two percentage points of GDP, and annual write-offs of uncollectible taxes coming to less than one-half of one percentage point of GDP. In the first five or ten years of transition, however, available evidence suggests that government toleration of non-payment of taxes was common even in the more rapidly reforming countries. Rough estimates of the scale of tax arrears range from two to 12 percentage points of GDP for the stock, and one to seven percentage points for the annual flow (Schaffer, 1998), and the empirical evidence suggests they were one of the main mechanisms governments used to soften the budget constraints of firms.

Lastly, large-scale and persistent wage arrears are also peculiar to transition economies, though in this case mostly limited to the countries of the former USSR. The scale of the wage arrears of firms at their peak – in aggregate, several percentage points of GDP – was typically smaller than trade credit and even tax arrears, but substantial in comparison with monthly wages. Payment of wages to employees several months in arrears was commonplace, and the absence of indexation imposed an extra cost in the high-inflation period of the early 1990s and following the burst of inflation that accompanied the collapse of the rouble in mid-1998. Wage arrears have sometimes been an important adjustment mechanism for labour markets in transition economies, partially absorbing negative shocks that would otherwise be fully reflected in actual wages or employment levels. The empirical evidence suggests that most wage arrears were late payments rather than non-payments, and with important distributional impacts with respect to household income. The social consequences of uncertainty and irregularity of wage payments were substantial, since workers in these countries had limited savings to fall back on and even less access to consumer credit markets, and thus faced great difficulties in smoothing income. Patterns across firms and workers in wage arrears have been related to firm, worker, and economy-wide characteristics (state-owned, poorly performing firms; workers in rural areas, outside options; tight credit policies; workers in sectors such as health and education, funded by the government budget), and to weak institutional environments that made it possible for firms to violate wage contracts at relatively low cost (see, for example, Lehmann, Wadsworth and Acquisti, 1999; Earle and Sabirianova, 2002).

See Also

- assets and liabilities
- soft budget constraint
- transition and institutions

Bibliography

Earle, J.S. and Sabirianova, K.Z. 2002. How late to pay? Understanding wage arrears in Russia. *Journal of Labor Economics* 20, 661–707.

Higgs, H., ed. 1925. *Palgrave's Dictionary of Political Economy*. London: Macmillan and Co.

Ickes, B.W. and Ryterman, R. 1992. The interenterprise arrears crisis in Russia. *Post-Soviet Affairs*

(formerly *Soviet Economy*) 8, 331–61.

Lehmann, H., Wadsworth, J. and Acquisti, A. 1999. Grime and punishment: job insecurity and wage arrears in the Russian Federation. *Journal of Comparative Economics* 27, 595–617.

Perotti, E.C. 1998. Inertial credit and opportunistic arrears in transition. *European Economic Review* 42, 1703–25.

Schaffer, M.E. 1998. Do firms in transition economies have soft budget constraints? A reconsideration of concepts and evidence. *Journal of Comparative Economics* 26, 80–103.

How to cite this article

Schaffer, Mark E. "arrears." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000230> doi:10.1057/9780230226203.0058

The New Palgrave Dictionary of Economics Online

Arrow, Kenneth Joseph (born 1921)

Ross M. Starr

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Kenneth Arrow is the author of key post-Second World War innovations in economics that have made economic theory a mathematical science. The Arrow Possibility Theorem created the field of social choice theory. Arrow extended and proved the relationship of Pareto efficiency with economic general equilibrium to include corner solutions and non-differentiable production and utility functions. With Gerard Debreu, he created the Arrow–Debreu mathematical model of economic general competitive equilibrium including sufficient conditions for the existence of market-clearing prices. Arrow securities and contingent commodities extend the model to cover uncertainty and provide a cornerstone of the modern theory of finance.

Keywords

American Economic Association; Arrow, K.; Arrow–Debreu model of general equilibrium; Arrow-Pratt index of risk aversion; Arrows' Theorem; Bergson social welfare function; Black, D.; Condorcet, J.-A.-N.; constant-elasticity-of-substitution production function; contingent commodities; control theory; convexity: in theorems of welfare economics; corner solutions; Cowles Commission for Research in Economics; Debreu, G.; discrimination: racial; Econometric Society; endogenous growth; Expected Utility Theorem; first fundamental theorem on welfare economics; fixed-point theorem; game theory; growth models; Hahn, F.; Hicks, J.; Hotelling, H.; impossibility theorem; independence of irrelevant alternatives; inventory policy: optimal; knowledge: as externality; Kuhn-Tucker Theorem; marginal rate of substitution; mathematical economics; medical insurance; moral hazard; Nash equilibrium; Pareto efficiency; paradox of voting; partial equilibrium theory; principal-agent problem; in medical care; racial discrimination; Arrow–Pratt index of risk aversion; Russell, B.; Samuelson, P.; second fundamental theorem on welfare economics; securities markets; social choice; Solow, R.; Stigler, G.; Tarski, A.; technical change; Tinbergen model; Tobin, J.; Uzawa, H.; Walras's Law

Article

Kenneth Arrow is a legendary figure, with an enormous range of contributions to 20th-century economics, responsible for the key post-Second World War innovations in economic theory that allowed economics to become a mathematical science. His impact is suggested by the number of major ideas that bear his name: Arrow's Theorem, the Arrow–Debreu model, the Arrow–Pratt index of risk aversion, and Arrow securities.

Four of his most distinctive achievements, all published in the brief period 1951–54, are as follows:

Arrow Possibility Theorem. *Social Choice and Individual Values* (1951a) created the field of social choice theory, a fundamental construct in theoretical welfare economics and theoretical political science.

Fundamental Theorems of Welfare Economics. ‘An extension of the basic theorems of classical welfare economics’ (1951b) presents the First and Second Fundamental Theorems of Welfare Economics and their proofs without requiring differentiability of utility, consumption, or technology, and including corner solutions (zeroes in quantities of inputs or outputs).

The Arrow–Debreu model of general economic equilibrium. ‘Existence of equilibrium for a competitive economy’ (with Gerard Debreu, 1954) creates the mathematical model of a competitive economy. The article formalizes the cross-effects between markets (effect of one market's price on another's demand and supply) and provides sufficient conditions for the existence of prices allowing decentralized market-clearing general equilibrium of a market economy. This model is central to the study of markets and welfare economics; it is now a standard of the field.

Securities markets and risk-bearing. ‘Le rôle des valeurs boursières pour la répartition la meilleure des risques’ (1953) introduces the concept of a ‘contingent commodity’. The article formalizes the role of markets, including financial markets, insurance and the stock market, in resource allocation; it is a cornerstone of the modern theory of finance.

Personal and intellectual history

Kenneth Arrow was born in New York City on 23 August 1921. He describes his family circumstances as financially comfortable during the 1920s, but ‘my father lost everything in the great depression and we were very poor for about ten years … When it came to college, my family's poverty constrained me to attend the City College’ (Breit and Spencer, 1986, p. 45). Free tuition at City College of New York (CCNY) gave a generation of New Yorkers their start on success.

The searing experience of the Depression affected career ambitions. Arrow thought he should pursue the safe career of a high-school mathematics teacher. He took education courses and he had a very successful period of practice teaching in mathematics, preparing students for the New York State Regents examination. However, the roster of applicants for New York City teachers’ positions was already filled.

Arrow graduated from CCNY in 1940 with the unusual combination of a mathematics major and a Bachelor of Science in Social Science. While at CCNY he studied with Alfred Tarski in a course on the calculus of relations. Arrow was a proofreader for Tarski's *Introduction to Logic* (1941). He entered Columbia University for graduate study and received an MA in mathematics in June 1941. Harold Hotelling, a statistician with an appointment in the economics department, was the decisive influence.

Arrow notes, ‘When I took [Hotelling's] course in mathematical economics, I realized I had found my niche’ (Breit and Spencer, 1986, p. 45). With the inducement of a fellowship in economics, Arrow transferred to the economics department for the rest of his graduate study.

Arrow's graduate work at Columbia was interrupted by the Second World War. During the war Arrow was a weather officer in the US Army Air Corps achieving the rank of Captain, working in the Long Range Forecasting Group. Arrow's first published paper comes from that period, ‘On the Use of Winds in Flight Planning’ (1949a). The group's principal task was to forecast the number of rainy days in air combat areas – a month in advance. The young statisticians in the Weather Division subjected the prediction techniques in use to statistical test against a simple null hypothesis based on historical data. Finding that prevailing techniques were not significantly more reliable than the null, the junior officers sent a memo to the General of the Air Corps suggesting that the group be disbanded. Six months later, the General's secretary replied on his behalf: ‘The general is well aware that your forecasts are no good. However, they are required for planning purposes.’ The group remained intact.

In 1946 Arrow returned to graduate study at Columbia. Harold Hotelling had by then left for the University of North Carolina's newly formed statistics department. The concern about making a living persisted. Arrow considered a non-academic career as a life insurance actuary. Tjalling Koopmans (at a Cowles Commission meeting in Ithaca, New York) advised him that actuarial statistics would prove unrewarding, saying, with characteristic reticence, ‘There is no music in it.’ Fortunately for economic science, Arrow followed this advice and decided to continue a research career.

In 1947 Arrow joined the (now legendary – then fledgling) research group at the Cowles Commission for Research in Economics at the University of Chicago. It seemed a golden age – all the ideas of mathematical economic theory and econometrics were being newly discovered. The close friendships and collaborations among colleagues of the Cowles Commission lasted a lifetime. Arrow describes the setting as a ‘brilliant intellectual atmosphere … with eager young econometricians and mathematically inclined economists under the guidance of Tjalling Koopmans and Jacob Marschak’ (Lindbeck, 1992, p. 107).

Jacob Marschak, the Cowles Commission Research Director, arranged for the Commission to administer the Sarah Frances Hutchinson Cowles Fellowship for women pursuing quantitative work in the social sciences (the Fellowship had originally specified a preference that fellows be women of the Episcopal Church of Seneca Falls, New York [reported in conversation with Jacob Marschak]). The fellows were Sonia Adelson (subsequently married to Lawrence Klein) and Selma Schweitzer. Kenneth Arrow and Selma Schweitzer were married in 1947.

Graduate study 1946–50, through Columbia, Chicago, Cowles, RAND and Stanford, included a daunting search for a worthy dissertation topic. Prospects considered and rejected included revising and restating the Tinbergen model (Tinbergen, 1939), and revising and restating Hicks's *Value and Capital* (1939). No topic seemed worthy. Then lightning struck: Arrow invented an entire field of economics with his dissertation ‘Social Choice and Individual Values’. The Columbia Ph.D., with Professor Albert Hart as dissertation advisor, was granted in 1951. As an econometrician, T. W. Anderson of Columbia (subsequently Arrow's colleague at Stanford) was called upon to pass judgement on a draft thesis unrecognizable as economics to Ken's advisors; Anderson pronounced the work sound.

The summer of 1948 and several summers thereafter were spent at the recently formed RAND Corporation in Santa Monica, California, a major centre of the newly emerging specialities of game theory and mathematical programming. In 1949 Arrow was appointed Acting Assistant Professor of

Economics and Statistics at Stanford University, and rapidly became Professor of Economics, and of Statistics, with the eventual additional title of Professor of Operations Research. He moved to Harvard in 1968 (returning regularly to Stanford for summer workshops), and rejoined the Stanford faculty in 1979. He retired in 1991.

In the 1950s and 1960s at Stanford, economic theory and econometrics faculty and graduate students were located in Serra House (converted from the retirement residence of the first president of the university) under the auspices of the Institute for Mathematical Studies in the Social Sciences (IMSSS) organized under the leadership of Patrick Suppes. In his memorial remarks for his student, Walter P. Heller (1942–2001), Arrow describes the *esprit de corps*: ‘Economic theory backed by serious mathematical reasoning was just beginning to be recognized. Our group of faculty and students in economic theory at Serra House. felt ourselves a community. Not an oppressed minority, but rather a vanguard. We were taking over!’

Stanford and UC Berkeley were centres of research in statistics and economic theory. The joint Berkeley–Stanford Mathematical Economics Seminar met biweekly at alternate campuses. The Berkeley group included Gerard Debreu, Roy Radner, Peter Diamond and Dan McFadden. Stanford's included Herbert Scarf and Hirofumi Uzawa. Uzawa came to Stanford on fellowship arranged by Arrow. Working on his own in Japan, he had written the manuscript eventually published as ‘Gradient method for concave programming, II: Global stability in the strictly convex case’ (Arrow, Hurwicz and Uzawa, 1958a, ch. 7). It was a successful global stability analysis of gradient adjustment, following Arrow and Hurwicz's local analysis (available to Uzawa in manuscript, published in the same volume). Arrow read the manuscript and enthusiastically invited Uzawa to accept a fellowship at Stanford.

Although the profession is now used to mathematical expression, in the 1950s and 1960s the mathematical complexity of Arrow's work was regarded as forbidding. Although Arrow was the pre-eminent economic theorist at Stanford, he was not designated to teach in the required first-year graduate microeconomic theory course; it was presumed that the treatment would be excessively abstract for this general audience. His reputation for mathematical abstraction provided the excuse for a jest when Arrow received the 1957 John Bates Clark Award of the American Economic Association (presented to a leading economist under the age of 40). At the presentation ceremony, introductory remarks were made by George Stigler, who reportedly advised Arrow, in a stage whisper, ‘You should probably say, “Symbols fail me”.’

Under the administration of President J.F. Kennedy, Arrow and Robert Solow served on the research staff of the Council of Economic Advisers. That was a remarkable group: Walter W. Heller, chair, Kermit Gordon and James Tobin. The Council and its staff then included three future Nobel laureates: Arrow, Solow and Tobin.

Academic travels abroad included visits to the Institute for Advanced Studies in Vienna in the summers of 1964 and 1971, and productive years at Churchill College, Cambridge, in 1963–64 and 1970, for collaboration with Frank Hahn on *General Competitive Analysis* (1971a).

To no one's surprise, Arrow received the 1972 Nobel Prize in Economic Sciences (jointly with the distinguished British economic theorist, John Hicks of Oxford). Aged 51 at the time of the award, he is (at this writing) by far the youngest recipient of the Nobel Prize in Economics.

Testimony to Arrow's qualities as a dissertation advisor, a teacher of the next generation of economists, is abundant. The flurry of former students volunteering to contribute to the Festschrift by Heller, Starr and Starrett (1986) was overwhelming. The most personal tribute is the number of leading colleagues

whose children have studied with Arrow. Jacob Marschak's son Thomas Marschak and Walter W. Heller's son Walter P. Heller wrote their doctoral dissertations with Arrow as principal advisor. Any list of Arrow's students (dissertation advisees, postdocs, and so forth) is a partial listing. They are numerous and are enthusiastically devoted to him, playing leading roles in academic and research economics. A selection includes: Theodore Bergstrom (UC Santa Barbara), David Bradford (Princeton University), Michael Bruno (Hebrew University, Bank of Israel), Graciela Chichilnisky (Columbia University), Peter Coughlin (University of Maryland), John Geanakoplos (Yale University), Louis Gevers (Université de Namur, Belgium), John Harsanyi (UC Berkeley), Walter P. Heller (UC San Diego), Peter Huang (University of Minnesota Law School), Takatoshi Ito (University of Tokyo), Jean-Jacques Laffont (Université des Sciences Sociales, Toulouse, France), Robert Lind (Cornell University), Thomas Marschak (UC Berkeley), Eric Maskin (Institute for Advanced Study, Princeton), Roger Myerson (University of Chicago), Hajime Oniki (Osaka-Gakuin University, Osaka, Japan), Heraklis Polemarchakis (Brown University), Karl Shell (Cornell University), Ross Starr (UC San Diego), David Starrett (Stanford University), Nancy Stokey (University of Chicago), Laurence Weiss (Goldman Sachs Corp.), Ho-Mou Wu (National Taiwan University), and Menahem Yaari (Hebrew University, Jerusalem).
A range of stories depict Arrow as a legendary larger-than-life figure:
‘Arrow is personally accessible and unpretentious, addressed as “Ken” by students, colleagues, and staff… Arrow thinks faster than he – or anyone else – can talk. Conversation takes place at such a rapid pace that no sentence is ever actually completed’ (Heller, Starr and Starrett, 1986, v. 1, p. xvii). The breadth of Arrow's knowledge is repeatedly a surprise, encompassing Chinese art, English history and the works of Shakespeare. At the 80th birthday celebration, Eric Maskin related the following example:

> On almost any subject arising in conversation, Arrow turns out to know a lot more than you do. Tired of being repeatedly shown up by their senior colleague, a group of junior faculty once concocted a plan. They first read up thoroughly on the most arcane topic they could think of – the breeding habits of gray whales. On the appointed day they gathered in the coffee room and waited for Ken to come in. Then they started talking about whales, concentrating on the elaborate theory of a marine biologist named Turner on how gray whales found their way back to the same breeding spot year after year. Ken was silent … they had him at last! With a sense of delicious triumph, they continued to discuss whales, and Ken looked more and more perplexed. Finally, he couldn't hold back: ‘But I thought that Turner's theory was entirely discredited by Spencer, who showed that the hypothesized homing mechanism couldn't possibly work.’

Arrow's presence in seminars is distinctive. He may open his (copious) mail, juggle a pencil, seem inattentive. He will then make a comment demonstrating that he is several steps ahead of the speaker. He will make clear that the history of economic thought includes abundant antecedents (which he can readily cite from memory) for the issues under discussion.

Social Choice and Individual Values: The General Possibility Theorem

Social Choice and Individual Values was Arrow's doctoral dissertation, published as a Cowles

Commission monograph. There are very few new ideas in economics. Arrow's General Possibility Theorem is as novel and fundamental as they come. The paradox of voting (cyclic majorities) appears to have been well-known, though not well formalized; Arrow (1951a) and Duncan Black (1948) both take it as understood. A review of the literature shows that it is attributable to Condorcet (1785). The paradox – intransitivity of choice from majority vote based on voters with transitive preferences – can be stated simply.

Think of three voters trying to decide by majority vote among three possibilities, A, B and C. Each of the individual voters has transitive (rational) preferences. Voter 1 prefers A to B and prefers B to C. Voter 2 prefers B to C and C to A. Voter 3 prefers C to A and A to B. Then there is a majority of voters preferring A to B (voters 1 and 3), and a majority preferring B to C (voters 1 and 2). If group decision-making is also transitive (rational), then the group should prefer A to C. But just the opposite occurs; there is a majority preferring C to A (voters 2 and 3). Despite the transitivity of individual preferences, the group preference on pairs of alternatives, as expressed by majority vote, is intransitive (irrational).

Arrow's General Possibility Theorem (also known as ‘Arrow's Theorem’, the ‘Arrow Possibility Theorem’ or the ‘Arrow Impossibility Theorem’) shows that the paradox is not merely an anomaly but intrinsic to group decision-making. The theorem has been a focus of vigorous study for generations. An elegant proof in Sen (1986) is particularly striking since it is framed as a generalization of the Condorcet paradox.

The Possibility Theorem suggests four reasonable criteria for a group decision-making mechanism, all of which are fulfilled by majority voting (assume at least three possible choices and at least three voters):

1. 1. *Unrestricted Domain*. The decision-making mechanism can accommodate all logically possible preferences on the available choices.
2. 2. *Pareto Principle*. If everyone prefers one alternative over another, the group decision should have that preference as well.
3. 3. *Independence of Irrelevant Alternatives*. In choosing between any two alternatives, group decision-making takes account only of individual preferences on those alternatives; preferences on a third possibility do not enter the choice between those two.
4. 4. *Non-dictatorship*. There is no single person whose preferences will always be followed by the group decision-making mechanism.

The Possibility Theorem says that no decision-making mechanism that fulfils all four of the above conditions results in transitive (rational) group choices based on transitive (rational) individual preferences. The Condorcet paradox is not merely an anomaly. It is unavoidable. It represents a fundamental defect in group decision-making.

Each of the four above conditions is essential to the theorem; there are examples of transitive group decision-making mechanisms that fulfil any three but not four. Of the four, the most controversial is Independence of Irrelevant Alternatives; it prevents voluntary misstatement by a voter of his preferences from being an attractive strategy (overstating dislike of a third option to make a preferred one of two succeed in a weighted voting scheme).

At the time *Social Choice and Individual Values* was published, the logic of group decision-making was

not even recognized as an economic issue. Since then there has been an overwhelming blossoming of the ‘social choice’ field. It is a topic for the *Handbook of Mathematical Economics* (Sen, 1986); thousands of journal articles deal with it; every graduate student in economics is introduced to it. Kenneth Arrow created the field by formalizing a result that says the object of the field is unachievable.

The book also had a significant impact in a second direction: treating economic theory as an axiomatic logical field rather than as a sphere of calculation. *Social Choice* was one of the first essays, certainly the first monograph, to treat economics with the same generality and logical rigour as classical geometry. This approach was to be repeated in the next of Arrow's several major works in general equilibrium theory and classical welfare economics.

How did Arrow come to develop this structure? It was during the first summer, in 1948, at RAND that several strands of thought came together. The Condorcet paradox of cyclic majorities was common knowledge (though not the attribution to Condorcet). Independently of Duncan Black (1948), Arrow developed the restriction of individual preferences to the single-peaked format as a solution, but then realized that he'd been scooped when he read Black's result in the *Journal of Political Economy*. He was aware of the ambiguity in describing the optimizing policy of a business firm under uncertainty: profit maximization is no longer well-defined and majority voting of shares is subject to the Condorcet paradox. Arrow's techniques of logical formalization were ready. As a high-school student he had read Russell's *Introduction to Mathematical Philosophy* (1920); at CCNY he became familiar with Tarski's *Introduction to Logic* (1941) and the calculus of relations. With that preparation, it was obvious that the indifference curve approach used by economists was a form of a logical ordering. Axiomatic treatment came naturally.

RAND was the centre of the developing field of game theory, which was being used to formalize discussions of strategic behaviour in international relations. During a coffee break the logician Olaf Helmer posed the following problem. Game theory supposes rational strategic behaviour among optimizing agents. The maximand of an individual may be well-defined, perhaps as a utility function; but what is the maximand of a country? Arrow replied that a Bergson social welfare function should represent a country's maximand. That set him to work. Demonstrating that his answer to Helmer was fundamentally and necessarily inadequate is the meaning of the Possibility Theorem. Arrow started the inquiry by looking at a variety of group decision-making mechanisms. They all looked wrong; either they led to intransitivity or they violated the Independence of Irrelevant Alternatives, so that preferences for an alternative that was out of the running nevertheless entered the group's decision. He was led to formalize the conditions of group decision-making, reflecting a long-standing interest in axiomatic reasoning. ‘The development of the theorems and their proofs then required only about three weeks, although writing them as a monograph took many months’ (1983a, p. 4).

Extension of the fundamental theorems of welfare economics

In the 1940s welfare economics in mathematical form (the relationship of market equilibrium to economically efficient allocation) was very much a matter of the calculus (Samuelson, 1947). Marginal rates of substitution (ratios of marginal utilities) were equated to marginal rates of transformation (ratios of marginal products of factors) which were equated to price ratios. This is a sound viewpoint so long as the underlying functions are differentiable and the quantities of goods and factors are in a range where

they can be varied. Arrow's view was that there is a fundamental weakness to this approach in the presence of non-negativity constraints on quantities. It works only when quantities are strictly positive. That is, the calculus doesn't treat corner solutions. But almost every practical economic solution is a corner solution: it is rare to find that all quantities of all possible goods and all possible inputs are used in strictly positive quantities. This is particularly true when differing qualities or varieties of similar goods are treated distinctly (white, sourdough and rye breads are distinct commodities, as are luxury and efficiency apartments). There must be a welfare economics that includes corner solutions; it must be possible to present welfare economics without the calculus.

Arrow attributes his insight to a seminar presentation on the fundamental theorems of welfare economics given by Paul Samuelson at the University of Chicago, in Samuelson's style using the calculus (1983b, p. 14). The diagrams that illustrated the equations depicted a separating hyperplane. Arrow had learned of the fundamental role of convexity and the separating hyperplane theorem at RAND in the summer of 1948. The result of these reflections is ‘An extension of the basic theorems of classical welfare economics’ appearing in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The conference was held in the summer of 1950 in Berkeley, and the proceedings appeared a year later. There, the First and Second Fundamental Theorems of Welfare Economics are stated in terms of real analysis and convex sets, without the use of the calculus and including corner solutions.

At the level of the firm and the household, characterizing optimizing behaviour at corner solutions is the job of the Kuhn–Tucker Theorem. In a case of simultaneous discovery of related ideas, that theorem was first publicly presented at the same Berkeley Symposium (Kuhn and Tucker, 1951).

First Fundamental Theorem of Welfare Economics: Every competitive equilibrium allocation is Pareto efficient. This result does not require convexity of tastes or technology, though convexity may be useful in establishing the existence of equilibrium prices.

Second Fundamental Theorem of Welfare Economics: In an economy with convex technology and preferences, every Pareto-efficient allocation can be sustained as a competitive equilibrium with appropriate prices subject to a redistribution of ownership shares in firms and redistribution of endowment (except that some low-income households may be expenditure minimizers subject to utility constraint, rather than utility maximizers subject to budget).

Neither of these results depends on positivity of quantities or on differentiability of the functions or relations. The generality of the results, the use of a formal mathematical structure of assumptions, theorems and proofs was again novel. It meant that economics was becoming closer to formal mathematics.

General equilibrium theory

In the early 1950s, Arrow (at Stanford) pursued, largely by correspondence, joint work on general equilibrium theory with Gerard Debreu, who was then at the Cowles Commission in Chicago. The theory of general economic equilibrium recognizes that the economy is an interactive system. Decisions and prices in one market have a direct impact on supply and demand in other markets. The question Arrow and Debreu treated is: under what (sufficiently general and formalized) conditions can there be prices so that all markets simultaneously clear? This issue is known as ‘the existence of economic general equilibrium’. The term ‘general’ equilibrium refers to the many markets simultaneously clearing, as opposed to ‘partial’ equilibrium where a single market is considered in isolation. Moreover,

the theory allows – or forces – the theorist to formulate relatively complete models of the economy. The result of these inquiries has been an intellectual revolution and an intellectual foundation for market economics. A half-century after it was introduced to economics, the Arrow–Debreu model is the cornerstone and workhorse of our theory of markets and resource allocation.

Abraham Wald, with whom Arrow had studied at Columbia, had written several papers in the field (while in Vienna in the 1930s before emigrating to avoid the Nazi takeover) but had run up against fundamental mathematical difficulties (Wald, 1934–35, 1936). He explained to Arrow that the problem was ‘very difficult’, advice that was enough to discourage the young economic theorist for some years. It was the recognition by Arrow and Debreu of the importance of using a fixed point theorem that led to major progress in this area. (Credit for independent discovery of the importance of fixed point theorems in this context is due to Lionel McKenzie, 1954. The use of a fixed point theorem for demonstrating the existence of an equilibrium [of a game] was pioneered by John Nash, 1950. See Debreu, 1983).

Arrow describes his early thoughts on the subject and the interaction with ideas current at the time (particularly the Nash equilibrium of N-person games) thus:

> My original approach, for what it is worth, was to formulate competitive equilibrium as the equilibrium of a suitably chosen game. The players of this fictitious game were the consumers, a set of ‘anticonsumers’ (one for each consumer), producers, and a price chooser. Each consumer chose a consumption vector, each anticonsumer a nonnegative number (interpretable as the marginal utility of income), each firm a production vector, and the price chooser a price vector on the unit simplex. The payoff to a consumer was the utility of his consumption vector plus the budgetary surplus (possibly negative, of course) multiplied by the anticonsumer's chosen number. The payoff to an anticonsumer was the negative of the payoff to the corresponding consumer. The payoff to the firm was profit and to the price chooser the value of excess demand at the chosen prices. This is a well-defined game. The existence of equilibrium does not follow mechanically from Nash's theorem, since some of the strategy domains are unbounded.
>
> Debreu and I sent our manuscripts to each other and so discovered our common purpose. We also detected the same flaw in each other's work; we had ignored the possibility of discontinuity when prices vary in such a way that some consumers’ incomes approach zero. [The possibility of discontinuity in demand at incomes where household consumption is on the boundary of the possible consumption set is known as the ‘Arrow corner’.]. We then collaborated, mostly by correspondence, until we had come to some resolution of this problem. In the main body of the work we followed more closely Debreu's more elegant formulabased on the concept of generalized games, which eliminated the need for ‘anticonsumers.’ (1983b, pp. 58–9)

The papers of Arrow and Debreu (1954) and McKenzie (1954) were presented to the 1952 meeting of the Econometric Society. Publication of ‘Existence of equilibrium for a competitive economy’ represents a fundamental step in the revision of economic analysis and modelling, demonstrating the power of a formal axiomatic approach with relatively advanced mathematical techniques. The approach of the field is revolutionary: it fundamentally changes our way of thinking. Once we see things this way,

it is hard to conceive of them otherwise.

Sufficient conditions for the existence of market-clearing prices – consistent with one-another – for N distinct commodities are: (a) demand and supply are continuous as a function of prices, and (b) Walras's Law. These properties are derived from fundamental assumptions on the structure of preferences and endowments of households and the technology of firms. The theory is general enough to include point-valued and (convex) set-valued demand and supply.

Debreu's *Theory of Value* (1959) made the Arrow–Debreu general equilibrium model accessible to the wider profession. The implications for economic theory as a discipline were multifaceted: *general* equilibrium, treating all markets as interacting together, became systematic; the axiomatic method was set firmly in place as part of economic theory. Economic theory could be as precise and logically demanding as geometry. The potential of formal theory to generalize could be brought to bear. The Arrow–Debreu treatment proved, with full mathematical rigour, that any economy fulfilling the model's clearly and generally specified assumptions would produce its specified results.

A number of articles (principally co-authored with Leonid Hurwicz, 1958b, 1959) treat the stability of general equilibrium. Though Arrow and Debreu (1954) establishes the existence of market clearing prices, it does not derive ‘equilibrium’ as the rest point of a dynamic system. The stability question focuses on how a price adjustment system will lead to market clearing prices. Since prices in each market (at least potentially) enter into the excess demands of all markets, there is plenty of room for price adjustments to go awry. This body of literature sorts out and proves sufficient conditions for adjustment to be successful. Bottom line: a sufficient condition is that other markets do not excessively interfere with excess demands on any single market; if the principal determinant of excess demands for each good is the price of that good, then price adjustment to market clearing will be successful.

The effect of the introduction of the Arrow–Debreu model on economic theory has been overwhelming. Every graduate-level textbook in microeconomic theory discusses it. Whole classes of economic theorists describe their speciality as ‘general equilibrium theory’. In the 15 years following publication of *Theory of Value*, a major focus of pure theory was understanding and extending the model. This included its relationship to bargaining (Debreu and Scarf, 1963), to large economies (Aumann, 1966) and to computing general equilibrium prices (Scarf and Hansen, 1973). It was further elaborated by Arrow and Hahn (1971a).

Contingent commodities

Part of the power of mathematics is generalization. If you've solved a problem once, you don't have to solve it again – even in different circumstances if you can show that the previous treatment applies. This was the brilliantly simple insight in the creation of the concept of ‘contingent commodity’.

Arrow's thought had been influenced by Hicks's *Value and Capital*, including understanding the power of defining a commodity to include specification of time and location, and by L.J. Savage's lectures on mathematical statistics at Chicago, including a notion of the ‘state of the world’ as defining a random variable. (The ‘state of the world’ concept for defining a random variable is attributable to Kolmogorov [1933]). It was a fundamental step to combine these notions so that a commodity might be defined by what it is, where and when deliverable, *and by the ‘state of the world’ in which it is deliverable*.

By redefining a ‘commodity’ in this way as a ‘contingent commodity’, the complete structure of the

Arrow–Debreu model of general equilibrium and economic efficiency could be applied. This is now typically described in the literature as ‘a full set of Arrow–Debreu futures contracts’. The concept of an efficient (or ‘optimal’) allocation of risk-bearing is immediately evident as a consequence of the modelling structure. The next step is to suggest a security contract contingent on the state of the world payable in money – to economize on the number of actively traded commodities – now known as an ‘Arrow security’ or ‘Arrow insurance contract’. This has been an extremely powerful concept, allowing researchers to formulate their ideas clearly; the Arrow security is a staple of 21st century theoretical finance.

The paper ‘Le rôle des valeurs boursières pour la répartition la meilleure des risques’, originally written in English, was translated into French for a conference at Centre National de Recherche Scientifique, Paris, in June 1952. Other conference participants included Jacob Marschak, Maurice Allais, L.J. Savage, Milton Friedman and Pierre Massé. It was published in French in *Econométrie* and the original English version appeared (as a ‘translation’) a decade later in *Review of Economic Studies*, after the notions had been introduced to English-speaking readers in *Theory of Value*.

Individual behaviour towards risk, economics of medical care, learning by doing

Treatment of uninsurable risk (where contingent commodities and Arrow securities are not available or correctly priced) has been a focus of Arrow's work for decades. It appears in the *Collected Papers*, the *Aspects of the Theory of Risk Bearing (Yrjo Jahnsson lectures)* (1965a), and in *Essays in the Theory of Risk Bearing* (1971b). These essays provide for many readers the most systematic treatment available of the statement and proof of the Expected Utility Theorem, derivation of the Arrow–Pratt risk aversion index, and a systematic framework for considering decision-making in an uncertain world.

Several papers (1963, 1965b) treat the economics of medical care, a setting where uncertainty, information as a scarce resource, and insurance all play a part. An element of the contribution is to state the issues in an abstract analytic economic framework. This reminds economists of why these problems are not textbook economics, and reminds non-economists that the economics textbook is useful. The historical setting in which these articles were written is pre-1990, that is, before health maintenance organizations (HMOs) became popular, when the principal form of medical insurance available was fee for service. They contain several insights (probably not unique to or first from Arrow, but effectively presented). For example, medical needs are uncertain so medical insurance is not merely a form of payment but is a response to risk. Again, medical insurance reduces the marginal cost of care as seen by the patient below actual cost, encouraging increased use (moral hazard consequence of insurance). Finally, medical care is distinct (but not unique) among commodities in that the decisions to incur care and the form that it should take are made to a large extent by the provider (the medical doctor) who is paid for providing care rather than by the buyer (patient). There is a resulting conflict of interest and reliance on professional norms. Arrow's treatment of the doctor-patient relationship as a seller-buyer interaction is an early appearance in the literature of the conflict we now recognize as the ‘principal-agent problem’ with an attendant family of issues.

In the 18th century Adam Smith noted that one of the benefits of specialization in production was that workers at specialized tasks learned how most effectively to perform them. Arrow's ‘The economic implications of learning by doing’ (1962) reflects in part the temper of the time – economic growth and growth models were a principal focus of theory and policy. In addition, it is a leap several decades ahead

in growth theory. In contrast to growth models in the 1960s, it presents endogenous growth, a research topic that became an active focus decades later (Romer, 1994). The study brings together two apparently disparate strands of economic modelling: technical change and the theory of external effects. The benefits of production in a particular line of work include not only output but the greater experience of the firm and the workforce in production. Through production, workers and firms learn how to produce more with fewer inputs. To the extent that this knowledge is inappropriable or non-marketable, it provides an external benefit to the economy. This on-the-job experience will typically be under-provided relative to an economically efficient allocation.

Optimal programming, control theory, mathematical statistics, racial discrimination, and the CES production function

In 16 books (not including the *Collected Papers*) and 250 technical articles, there are significant contributions to a breadth of issues in economics, mathematical programming and public policy. There's even some mathematical statistics (with Blackwell and Girshick, 1949b).

One of the most useful – to other economists – is ‘Capital-labor substitution and economic efficiency’ by Arrow, Chenery, Minhas and Solow (1961). It introduced the constant-elasticity-of-substitution (CES) production function, spawning an immense empirical literature.

Public Investment, the Rate of Return, and Optimal Fiscal Policy and several papers with Mordecai Kurz (1970) introduced control theory to the theory of the firm, to the theory of the household, and to public finance. A variety of books and articles treat mathematical programming and optimal inventory policy.

Several papers formally model racial discrimination in employment (1973). This is a tricky problem, and not merely because it is politically controversial. Pure microeconomic theory would suggest that there should be no racial discrimination by rational profit-maximizing employers; significant discrimination should result in below-market wage rates for the discriminated-against workers with resultant extra incentive for employers to hire them. How then can an economic model of optimizing behaviour explain the prevalence of racial discrimination? The answers (based on the racial views of employers, employees, customers) provide clues to locating the points of leverage that may lead to amelioration or policy.

What have we learned?

Arrow, along with Debreu, was a decisive figure in introducing the axiomatic method to economic theory. *Social Choice and Individual Values* and ‘Existence of equilibrium for a competitive economy’ fundamentally changed the agenda of economic theory. Formal logical reasoning and formal statement of assumptions and conclusions became the standard of pure theory (Suppes, 2005). The axiomatic method need not be a straitjacket. Arrow's less formal work demonstrates the role of insight: observing actual economic activity and asking ‘why?’, where the acceptable class of answers reflects underlying principles of economic analysis. The result is a rich understanding of the nuance and power of economics.

Celebrations

Dedicated colleagues and students have done their best to show adulation and gratitude to Arrow. There has been a succession of public celebrations.

On Arrow's 65th birthday in August 1986, an immense birthday conference and party, known as the ‘Arrowfest’, took place at Stanford. It reunited colleagues and students from all over the world. There were two days of conference papers and testimonial remarks. A three-volume Festschrift was presented (on time) (Heller, Starr and Starrett, 1986), including papers by 35 of Arrow's students and colleagues. Among the contributing authors were three (eventual) Nobel laureates: John Harsanyi, Amartya Sen and Robert Solow. The observance included a gala dinner with testimonial remarks and an expression of thanks from Arrow.

To observe his 70th birthday, the celebration was at the doctoral alma mater, a conference and social gathering in October 1991 titled ‘Columbia Celebrates Arrow's Contributions’. The Festschrift volume (Chichilnisky, 1999) included papers by 22 colleagues and students. The 70th birthday was also the occasion of formal retirement from active faculty status at Stanford. That rite of passage was observed with a reception, including testimonials from colleagues, among them the senior colleagues who had been clever enough to recruit Arrow to Stanford two generations earlier. Stanford's Arrow Lecture Series was initiated, annually inviting distinguished speakers in economic theory in Arrow's honour.

A 40th anniversary party for general equilibrium theory was held in June 1993 at Center for Operations Research and Econometrics (CORE) of the Université Catholique de Louvain in Louvain-la-Neuve, Belgium. For several days and nights hundreds of professors, researchers and students from around the world presented papers, discussions and reminiscences of the speciality they had pursued for years. At the centre of the celebration were the 20th-century founders of the field, Kenneth Arrow, Gerard Debreu and Lionel McKenzie.

There was a happy coincidence in 2001, when the 50th anniversary of *Social Choice and Individual Values* approximately coincided with Arrow's 80th birthday. A panel discussed the book's impact over the previous half century: Pat Suppes (Stanford University) on philosophy, John Ferejohn (Stanford University) on political science, and Eric Maskin (Institute for Advanced Study) on economics. The gathering included Professor Ted Anderson, who was at Columbia when *Social Choice* was submitted as Arrow's dissertation.

A dinner that evening featured moving toasts of appreciation by colleagues from around the world and presentations by Arrow's sons, Andy and David. The conclusion – sending the audience out singing into the evening – was the ad hoc musical group, the Economy Singers, singing advice to rising young economists: ‘Brush Up Your Arrow, Start Quoting Him Now.’

To many students and colleagues, Kenneth Arrow is a source of inspiration and a focus of friendship and respect:

> … an inspirational teacher and colleague … The intellectual standards he set and the enthusiasm with which he approaches our subject are surely part of all of us. Those of us who have had a chance to know him well are particularly fortunate. We are far richer for the experience. (Heller, Starr and Starrett, 1986, vol. 1, pp. xi, xvii)

See Also

- Arrow's theorem
- Debreu, Gerard
- equilibrium (development of the concept)
- general equilibrium (new developments)
- Sen, Amartya
- uncertainty

Selected works

1949a. On the use of winds in flight planning. *Journal of Meteorology* 6, 150–9.

1949b. (With D. Blackwell and M. A. Girshick.) Bayes and minimax solutions of sequential decision problems. *Econometrica* 17, 213–44.

1951a. *Social Choice and Individual Values*. New York: Wiley.

1951b. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley and Los Angeles: University of California Press.

1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Économetrie. Colloques Internationaux du Centre National de la Recherche Scientifique* 11, 41–7. Translated into English as ‘The role of securities in the optimal allocation of risk-bearing’, *Review of Economic Studies* 31(1964), 91–6.

1954. (With G. Debreu) Existence of equilibrium for a competitive economy. *Econometrica* 22, 265–90.

1958a. (With L. Hurwicz and H. Uzawa.) *Studies in Linear and Non-Linear Programming*. Stanford: Stanford University Press.

1958b. (With L. Hurwicz.) On the stability of the competitive equilibrium. *Econometrica* 26, 522–52.

1959. (With H. Block and L. Hurwicz.) On the stability of the competitive equilibrium, II. *Econometrica* 27, 82–109.

1961. (With H. Chenery, B. Minhas and R. Solow.) Capital-labor substitution and economic efficiency. *Review of Economics and Statistics* 43, 225–50.

1962. The economic implications of learning by doing. *Review of Economic Studies* 29, 155–73.

1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53, 941–73.

1965a. *Aspects of the Theory of Risk-Bearing*. Helsinki: Yrjö Jahnssonin säätiö.

1965b. Uncertainty and the welfare economics of medical care: reply (The implications of transaction costs and adjustment lags). *American Economic Review* 55, 154–8.

1970. (With M. Kurz.) *Public Investment, the Rate of Return, and Optimal Fiscal Policy*. Baltimore and London: The Johns Hopkins University Press.

1971a. (With F. Hahn.) *General Competitive Analysis*. San Francisco: Holden-Day; Edinburgh: Oliver & Boyd.

1971b. *Essays in the Theory of Risk-Bearing*. Chicago: Markham; Amsterdam and London: North-Holland.

1973. The Theory of Discrimination. In *Discrimination in Labor Markets*, ed. O. Ashenfelter and A. Rees. Princeton: Princeton University Press.

1983a. *Collected Papers of Kenneth J. Arrow, Volume 1: Social Choice and Justice*. Cambridge, MA: The Belknap Press of Harvard University Press.

1983b. *Collected Papers of Kenneth J. Arrow, Volume 2: General Equilibrium*. Cambridge, MA: The Belknap Press of Harvard University Press.

1984a. *Collected Papers of Kenneth J. Arrow, Volume 3: Individual Choice under Certainty and Uncertainty*. Cambridge, MA: The Belknap Press of Harvard University Press.

1984b. *Collected Papers of Kenneth J. Arrow, Volume 4: The Economics of Information*. Cambridge, MA: The Belknap Press of Harvard University Press.

1985a. *Collected Papers of Kenneth J. Arrow, Volume 5: Production and Capital*. Cambridge, MA: The Belknap Press of Harvard University Press.

1985b. *Collected Papers of Kenneth J. Arrow, Volume 6: Applied Economics*. Cambridge, MA: The Belknap Press of Harvard University Press.

1991. The origins of the Impossibility Theorem. In *History of Mathematical Programming*, ed. J. Leenstra, A. Rinnooy Kan and A. Schrijver. Amsterdam, New York, Oxford, Tokyo: CWI Amsterdam,

North-Holland.

Bibliography

Aumann, R. 1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34, 1–17.

Black, D. 1948. On the rationale of group decision-making. *Journal of Political Economy* 56, 23–34.

Breit, W. and Spencer, R., eds. 1986. Kenneth J. Arrow. *Lives of the Laureates: Seven Nobel Economists*. Cambridge, MA and London: MIT Press.

Chichilnisky, G., ed. 1999. *Markets, Information, and Uncertainty: Essays in Economic Theory in Honor of Kenneth J. Arrow*. New York: Cambridge University Press.

Condorcet, J.-A.-N. de Caritat, marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie royale.

Debreu, G. 1959. *Theory of Value*. Cowles Foundation Monograph No. 17. New York: John Wiley.

Debreu, G. 1983. Mathematical economics at Cowles. Online. Available at http://cowles.econ.yale.edu/about-cf/50th/debreu.htm, accessed 9 July 2005.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

Heller, W., Starr, R. and Starrett, D., eds. 1986. *Essays in Honor of Kenneth J. Arrow*, 3 vols. Cambridge: Cambridge University Press.

Hicks, J. 1939. *Value and Capital*. Oxford: Clarendon Press.

Kolmogorov, A. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: J. Springer.

Kuhn, H. and Tucker, A. 1951. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Lindbeck, A., ed. 1992. *Nobel Lectures: Economic Sciences, 1969–1980*. Singapore: World Scientific Publishing Co.

McKenzie, L. 1954. On equilibrium in Graham's model of world trade and other competitive systems.

Econometrica 22, 147–61.

Nash, J. 1950. Equilibrium points of N-person games. *Proceedings of the National Academy of Sciences of the USA* 36, 48–9.

Romer, P. 1994. The origins of endogenous growth. *Journal of Economic Perspectives* 8(1), 3–22.

Russell, B. 1920. *Introduction to Mathematical Philosophy*. London: Allen and Unwin; New York: Macmillan.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Scarf, H. and Hansen, T. 1973. *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Sen, A. 1986. Social choice theory. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M. Intriligator. New York: Elsevier.

Suppes, P. 2005. The pre-history of Kenneth Arrow's *Social Choice and Individual Values*. Unpublished manuscript. Stanford: Center for the Study of Language and Information, Stanford University.

Tarski, A. 1941. *Introduction to Logic*. New York: Oxford University Press.

Tinbergen, J. 1939. *Business Cycles in the United States of America, 1919–1932*. Geneva: Economic Intelligence Service, League of Nations.

Wald, A. 1934–35. Uberdie Produktionsgleichungen der Oekonomischen Wertlehre. *Ergebnisse eines Mathematischen Kolloquiums* 7, 1–6.

Wald, A. 1936. Uber einige Gleichungssystem der mathematischen Okonomie. *Zeitschrift fur Nationaloekonomie* 7, 637–70. Translated by Otto Eckstein as ‘On some systems of equations of mathematical economics’, *Econometrica* 19(1951) 368–403.

How to cite this article

Starr, Ross M. "Arrow, Kenneth Joseph (born 1921)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000067> doi:10.1057/9780230226203.0060

The New Palgrave Dictionary of Economics Online

Arrowâ€ “ Debreu model of general equilibrium

John Geanakoplos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the 1950s Kenneth Arrow and Gerard Debreu showed that the market system could be comprehensively analysed in terms of the neoclassical methodological premises of individual rationality, market clearing, and rational expectations, using the two mathematical techniques of convexity and fixed point theory. In so doing they greatly advanced the use of mathematics in economics.

Keywords

agent optimization; Arrowâ€“Debreu model of general equilibrium; asymmetric information; bankruptcy; bounded rationality; Brouwer's fixed point theorem; coalitions; commodities; comparative statics; competitive equilibrium; consumption loan model; convexity; core; Cournot's duopoly model; differential pay principle; existence of equilibrium; externalities; Fixed point theorems; general equilibrium; Hicks, J. R.; incomplete markets; individualism; insurance markets; interpersonal utility comparisons; Kakutani's fixed point theorem; Lindahl equilibrium; local uniqueness; market clearing; mathematical economics; mathematics and economics; microfoundations; Minkowski's theorem; monotonicity; negative prices; neoclassical economics; neoclassical production function; non-satiation hypothesis; ordinal utility; overlapping generations model; Pareto optimality; preference; price; rational expectations; rational expectations equilibrium; rationality; risk allocation; Samuelson, P. A.; Sard's theorem; separating hyperplane theorem; tÃ¢tonnement; transversality theory; uncertainty; utility; Walras's Law; welfare economics; welfare theorems

Article

IÂ Introduction

It is not easy to separate the significance and influence of the Arrowâ€“Debreu model of general

equilibrium from that of mathematical economics itself. In an extraordinary series of papers (Arrow, 1951; Debreu, 1951; Arrow and Debreu, 1954), two of the oldest and most important questions of neoclassical economics, the viability and efficiency of the market system, were shown to be susceptible to analysis in a model completely faithful to the neoclassical methodological premises of individual rationality, market clearing, and rational expectations, through arguments at least as elegant as any in economic theory, using the two techniques (convexity and fixed point theory) that are still, after 30 years, the most important mathematical devices in mathematical economics. Fifteen years after its birth (for example, Arrow, 1969), the model was still being reinterpreted to yield fresh economic insights, and 20 years later the same model was still capable of yielding new and fundamental mathematical properties (for example, Debreu, 1970; 1974). When we consider that the same two men who derived the most fundamental properties of the model (along with McKenzie, 1954) also provided the most significant economic interpretations, it is no wonder that its invention has helped earn for each of its creators, in different years, the Nobel Prize for economics.

In the next few pages I shall try to summarize the primitive mathematical concepts, and their economic interpretations, that define the model. I give a hint of the arguments used to establish the model's conclusions. Finally, on the theory that a model is equally well described by what it cannot explain, I list several phenomena that the model is not equipped to handle.

IIÂ The model

Commodities and Arrowâ€ “ Debreu commodities

(A.1) Let there be L commodities, l=1,â€‰â€¦,â€‰L. The amount of a commodity is described by a real number. A list of quantities of all commodities is given by a vector in $\mathbb{R}^L$.

The notion of commodity is the fundamental primitive concept in economic theory. Each commodity is assumed to have an objective, quantifiable, and universally agreed upon (that is, measurable) description. Of course, in reality this description is somewhat ambiguous (should two apples of different sizes be considered two units of the same commodity, or two different commodities?) but the essential quantitative aspect of commodity cannot be doubted. Production and consumption are defined in terms of transformations of commodities that they cause. Conversely, the set of commodities is the minimum collection of objects necessary to describe production and consumption. Other objects, such as financial assets, may be traded, but they are not commodities. General equilibrium theory is concerned with the allocation of commodities (between nations, or individuals, across time, or under uncertainty, and so on). The Arrowâ€“Debreu model studies those allocations which can be achieved through the exchange of commodities at one moment in time.

It is easy to see that it is often important to the agents in an economy to have precise physical descriptions of commodities, as for example when placing an order for a particular grade of steel or oil. The less crude the categorization of commodities becomes, the more scope there is for agents to trade, and the greater is the set of imaginable allocations. Two agents may each have apples and oranges. There is no point in exchanging one man's fruit for the other man's fruit, but both might be made better off if one could exchange his apples for the other's oranges. Of course there need not be any end to the distinctions which in principle could be drawn between commodities, but presumably finer details

become less and less important. When the descriptions are so precise that further refinements cannot yield imaginable allocations which increase the satisfaction of the agents in the economy, then the commodities are called Arrowâ€“Debreu commodities.

A field is better allocated to one productive use than another depending upon how much rain has fallen on it; but it is also better allocated depending on how much rain has fallen on other fields. This illustrates the apparently paradoxical usefulness of including in the description of an Arrowâ€“Debreu commodity characteristics of the world, for example the commodity's geographic location, its temporal location (Hicks, 1939), its state of nature (Arrow, 1953; Debreu, 1959; Radner, 1968), and perhaps even the name of its final consumer (Arrow, 1969), which at first glance do not seem intrinsically connected with the object itself (but which are in principle observable).

Hicks, perhaps anticipated by Fisher and Hayek, was the first to suggest an elaborate notion of commodity; this idea has been developed by others, especially Arrow in connection with uncertainty. Hicks was also the first to understand apparently complicated transactions, perhaps involving the exchange of paper assets or other non-commodities, over many time periods, in terms of commodity trade at one moment in time. Thus saving, or the lending of money, might be thought of as the purchase today of a particular future dated commodity. The second welfare theorem, which we shall shortly discuss, shows that an â€˜optimalâ€™ series of transactions can always be so regarded. By making the distinction between the same physical object depending, for example, on the state of nature, the general equilibrium theory of the supply and demand of commodities at one moment in time can incorporate the analysis of the optimal allocation of risk (a concept which appears far removed from the mundane qualities of fresh fruit) with exactly the same apparatus used to analyse the exchange of apples and oranges. Classifying physical objects according to their location likewise allows transportation costs to be handled in the same framework. Distinguishing commodities by who ultimately consumes them could allow general equilibrium analysis to systematically include externalities and public goods as special cases, though this has not been much pursued.

In reality, it is very rare to find a market for a pure Arrowâ€“Debreu commodity. The more finely the commodities are described, the less likely are the commodity markets to have many buyers and sellers (that is, to be competitive). More commonly, many groups of Arrowâ€“Debreu commodities are traded together, in unbreakable bundles, at many moments in time, in â€˜second bestâ€™ transactions. Nevertheless, this understanding of the limitations of real world markets, based on the concept of the Arrowâ€“Debreu commodity, is one of the most powerful analytical tools of systematic accounting available to the general equilibrium theorist. Similarly, the model of Arrowâ€“Debreu, with its idealization of a separate market for each Arrowâ€“Debreu commodity, all simultaneously meeting, is the benchmark against which the real economy can be measured.

Consumers

(A.1) Let there be H consumers, h=1,â€¦, H.â–¡

Each consumer h can imagine consumption plans $x \in \mathbb{R}^L$ lying in some consumption set X^h.(A.2)X^h is a closed subset in $\mathbb{R}^L$ which is bounded from below.â–¡

Each consumer h also has well-defined preferences $\succsim h$ over every pair $(x, y) \in X^h \times X^h$, where $x \succsim y$ means x is at least as desirable as y. Typically it is assumed that (A.3) $\succsim$ is a complete, transitive,

continuous ordering.â–¡

Notice that in general equilibrium consumers make choices between entire consumption plans, not between individual commodities. A single commodity has significance to the consumer only in relation to the other commodities he has consumed, or plans to consume. Together with transitivity and completeness, this hypothesis about consumer preferences embodies the neoclassical ideal of rational choice.

Rationality has not always been a primitive hypothesis in neoclassical economics. It was customary (for example, for Bentham, Jevons, Menger, Walras) to regard satisfaction, or utility, as a measurable primitive; rational choice, when it was thought to occur at all, was the consequence of the maximization of utility. And since utility was often thought to be instantaneously produced, sequential consumer choice on the basis of sequential instantaneous utility maximization was sometimes explicitly discussed as irrational (see, for example, BÃ¶hm-Bawerk on saving and the reasons why the rate of interest is always positive).

Once utility is taken to be a function not of instantaneous consumption, but of the entire consumption plan, then rational choice is equivalent to utility maximization. Debreu (1951) proved that any preference ordering $\succcurlyeq_h$ defined on X^hÃ—X^h satisfies (A.1)â€“(A.3) if and only if there is a utility function $u^h : X^h \to \mathbb{R}$ such that $x \succcurlyeq_h y$ exactly when $u^h(x) \geqq u^h(y)$.

Under the influence of Pareto (1909), Hicks (1939) and Samuelson (1947), neoclassical economics has come to take rationality as primitive, and utility maximization as a logical consequence. This has had a profound effect on welfare economics, and perhaps on the scope of economic theory as well. In the first place, if utility is not directly measurable, then it can only be deduced from observable choices, as in the proof of Debreu. But at best this will give an â€˜ordinalâ€™ utility, since if $f : \mathbb{R} \to \mathbb{R}$ is any strictly increasing function, then u^h represents $\succcurlyeq_h$ if and only if $v^h \equiv f \circ u^h$ represents $\succcurlyeq_h$. Hence there can be no meaning to interpersonal utility comparisons; the Benthamite sum $\Sigma_{h=1}^{H} u^h$ is very different from the Benthamite sum $\Sigma_{h=1}^{H} f^{h \circ} u^h$. In the second place, the ideal of rational choice or preference, freed from the need for measurement, is much more easily extended to domains not directly connected to the market and commodities such as political candidates or platforms, or â€˜social statesâ€™. The elaboration of the nature of the primitive concepts of commodity and rational choice, developed as the basis of the theory of market equilibrium, prepared the way for the methodological principles of neoclassical economics (rational choice and equilibrium) to be applied to questions far beyond those of the market.

Although the rationality principle is in some respects a weakening of the hypothesis of measurable utility and instantaneous utility maximization, when coupled with the notion of consumption plan it is also a strengthening of this hypothesis, and a very strong assumption indeed. For example there is not room in this theory for the Freudian split psyche (or self-deception), or for Odysseus-like changes of heart. Perhaps more importantly, a consumer's preferences (for example how thrifty he is) do not change according to the role he plays in the process of production (for example, on whether he is a capitalist or landowner), nor do they change depending on other consumersâ€™ preferences, or the supply of commodities. As an instance of this last case, note that it follows from the rationality hypothesis that the surge in the microcomputer industry influenced consumer choice between typewriters and word processors only through availability (via the price), and not through any learning effect. (Consumers can

â€˜learnâ€™ in the Arrowâ€“Debreu model, for example their marginal rates of substitution can depend on the state of nature, but the rate at which they learn is independent of production or consumption â€“ it depends on the exogenous realization of the state. We shall come back to this when we consider information.) If for no other reason, the burden of calculation and attention which rational choice over consumption plans imposes on the individual is so large that one expects rationality to give way to some kind of bounded rationality in some future general equilibrium models.
Two more assumptions on preferences made in the model of Arrowâ€“Debreu are nonsatiation and convexity:

- (A.4) For each $x \in X^h$, there is a $y \in X^h$ with $y \succ_h x$, that is, such that $y \succsim_h x$ and not $x \succsim_h y$ â–¡
- (A.5) X^h is a convex set, and $\succsim_h$ is convex, that is, if $y \succ_h x$ and 0<*t*â‰¤1, then $[ty + (1-t)x] \succ_h x$.â–¡

The nonsatiation hypothesis seems entirely in accordance with human nature. The convexity hypothesis implies that commodities are infinitely divisible, and that mixtures are at least as good as extremes. When commodities are distinguished very finely according to dates, so that they must be thought of as flows, then the convexity hypothesis is untenable. In a standard example, a man may be indifferent between drinking a glass of gin or of scotch at a particular moment, but he would be much worse off if he had to drink a glass of half ginâ€“half scotch. On the other hand, if the commodities were not so finely dated, then they would be more analogous to stocks, and a consumer might well be better off with a litre of gin and a litre of scotch, than two litres of either one. In any case, as we shall remark later, if every agent is small relative to the market (that is, if there are many agents) then the non-convexities in preferences are relatively unimportant.
Each agent *h* is also characterized by a vector of initial endowments

- (A.6) $e^h \in X^h \subset \mathbb{R}^L$ for all $h = 1, \ldots, H$.â–¡

The endowment vector e^h represents the claims that the consumer has on all commodities, not necessarily commodities in his physical possession. The fact that $e^h \in X^h$ means that the consumer can ensure his own survival even if he is deprived of all opportunity to trade. This is a somewhat strange hypothesis for the modern world, in which individuals often have labour but few other endowments, e.g. land. Doubtless the hypothesis could be relaxed; in any case, survival is not an issue that is addressed in the Arrowâ€“Debreu model.
Each individual *h* is also endowed with an ownership share of each of the firms *j*=1, â€¦, *J*

- (A.7) For all $h = 1, \ldots, H$, $j = 1, \ldots, J$, $\theta_{hj} \geq 0$, and for all $j = 1, \ldots, J$, $\sum_{h=1}^{H} \theta_{hj} = 1$.â–¡

- *Firms*. (A.8) Let there be *J* firms, *j*=1,â€¦, *J*.â–¡

The firm in Arrowâ€“Debreu is characterized by its initial distribution of owners, and by its

technological capacity $Y_j \subset \mathbb{R}^L$ to transform commodities. Any production plan $y \in \mathbb{R}^L$, where negative components of *y* refer to inputs and positive components denote outputs, is feasible for firm *j* if $y \in Y_j$. A customary assumption made in the Arrowâ€“Debreu model is free disposal: if *l*=1,â€¦, *L* is any commodity, and v_1 is the unit vector in $\mathbb{R}^L$, with one in the *l*th coordinate and zero elsewhere, then

- (A.9) For all $l = 1, \ldots, L$ and $k > 0$, $-kv_1 \in Y_j$, for some $j = 1, \ldots, J$.â–¡

Although it is strange, when thinking of nuclear waste etc., to think that any commodity can be disposed without cost (i.e. without the use of any other inputs), as we shall remark later, this assumption can be relaxed, if negative prices are introduced (or if weak monotonicity is assumed).
The empirically most vulnerable assumption to the Arrowâ€“Debreu model, and one crucial to its logic, is:

- (A.10) For each *j*, Y_j is a closed, convex set containing 0.â–¡

This convexity assumption rules out indivisibilities in production (e.g. half a tunnel), increasing returns to scale, gains from specialization, etc. As with consumption, if the indivisibilities of production are small relative to the size of the whole economy, then the conclusions we shall shortly present are not much affected. But when they are large, or when there are significant increasing returns to scale, the model of competitive equilibrium that we are about to examine is simply not applicable. Nevertheless, convexity is consistent with the traditionally important cases of decreasing and constant returns to scale in production.
We conclude by presenting three final assumptions used in the Arrow-Debreu model.

- (A.11) Let $e = \Sigma_{h=1}^{H} e^h$,
- â€ƒâ€ƒlet $F = \left\{ y \in \mathbb{R}^L \mid y = \Sigma_{j=1}^{J} y_j,\ y_j \in Y^j,\ j = 1, \ldots, J \right\}$,
- â€ƒâ€ƒlet $\bar{F} = \{ y \in F \mid y + e \geqq 0 \}$, and
- â€ƒâ€ƒlet $K = \left\{ (y_1, \ldots, y_J) \in Y_1 \times \ldots \times Y_J \mid \equiv y \Sigma_{j=1}^{J} y^j \in \bar{F} \right\}$.
- Then $\bar{F} \cap \mathbb{R}^L_{++} \neq \varphi$, and *K* is compact.â–¡

Assumption (A.11) requires that the level of productive activity that is possible even if the productive sector appropriates all the resources of the consuming sector is bounded (as well as closed).
Notice that these assumptions are consistent with firms owning initial resources, as well as individuals. In the original Arrowâ€“Debreu model (1954), the firms were prohibited from owning initial resources (they were assigned to the firm owners: with complete markets there is little difference, but with incomplete markets the earlier assumption is restrictive).

- (A.12) The economy is irreducible.â–¡

We shall not elaborate this assumption here. It means that for any two agents *h* and *h*â€², the endowment e^h of agent *h* is positive in some commodity *l*, which (taking into account the possibilities of production) agent *h*â€² could use to make himself strictly better off. It certainly seems reasonable that each agent's labour power could be used to make another agent better off.
Lastly, we assume that

- (A.13) The commodities are not distinguished according to which firm produces them, or who consumes them.â–¡

Assumption (A.13) is made simply for the purposes of interpretation. When put together with the definition of competitive equilibrium, it implies that there are no externalities to production or consumption, no public goods, etc. Mathematically, however, (A.13) has no content. In other words, if we dropped assumption (A.13), the Arrowâ€“Debreu notion of competitive equilibrium would still make sense (even in the presence of externalities and public goods) and it would still have the optimality properties we shall elaborate in Section III, but it would require an entirely different interpretation. Consumers, for example, would be charged different prices for the same physical commodities (same, that is, according to date, location and state of nature). In more technical language, a Lindahl equilibrium is a special case of an (A.1)â€“(A.12) Arrowâ€“Debreu equilibrium, with the commodity space suitably expanded and interpreted. Thus each physical unit of a public good is replaced by *H* goods, one unit for the public good indexed by which agent consumes it. Also the physical technology set describing the production of the public good is replaced by a different set in the Arrowâ€“Debreu model, lying in a higher dimensional space, where the output of the one physical public good is replaced by the joint output of the same amount of *H* goods. In an Arrowâ€“Debreu equilibrium, consumers will likely pay different prices for these *H* goods, i.e. for what in reality represents the same physical public good. Hence the differential pay principle for the optimal provision of public goods elucidated by Samuelson, which appeared to point to a qualitative difference between the analytical apparatus needed to describe optimality in public goods and private goods economies, is thus shown to be explicable by exactly the same apparatus used for private goods economies, simply by multiplying the number of commodities. The same device can also be used for analysing the optimal provision of goods when there are externalities, provided that negative prices are allowed. Assumption (A.13) thus seriously limits the normative conclusions that can be drawn from the model. From a descriptive point of view, however, rationality and the price taking behaviour which equilibrium implies make (A.13) necessary.

III**Â** Equilibrium

Price is the final primitive concept in the Arrowâ€“Debreu model. Like commodity it is quantifiable and directly measurable. As Debreu has remarked, the fundamental role which mathematics plays in economics is partly owing to the quantifiable nature of these two primitive concepts, and to the rich mathematical relationship of dual vector spaces, into which it is natural to classify the collections of price values and commodity quantities. Properly speaking, price is only sensible (and measurable) as a relationship between two commodities, i.e. as relative price. Hence there should be $L^2 - L$ relative prices

in the Arrowâ€“Debreu model. But the definition of Arrowâ€“Debreu equilibrium immediately implies that it suffices to give *L*â^'1 of these ratios, and all the rest are determined.

For mathematical convenience (namely to treat prices and quantities as dual vectors), one price is specified for each unit quantity of each commodity. The relative price of two commodities can be obtained by taking the ratio of the Arrowâ€“Debreu prices of these commodities. I shall proceed by specifying the definition of Arrowâ€“Debreu equilibrium, and then I make a number of remarks emphasizing some of the salient characteristics of the definition. The longest remark concerns the differences between the historical development of general equilibrium, up until the time of Hicks and Samuelson and the particular Arrowâ€“Debreu model of general equilibrium.

An Arrowâ€“Debreu economy *E* is an array $E = \{L, H, J(X^h, e^h, \succ_h), (Y^j), (d^{hj}), h = 1, \ldots, H, j = 1, \ldots, J\}$ satisfying assumptions (A.1)â€“(A-13). An Arrowâ€“Debreu equilibrium is an array $[(\bar{p}_l), (\bar{x}_l^h), (\bar{y}_l^j), l = 1, \ldots, L, h = 1, \ldots, H, j = 1, \ldots, J]$ satisfying:

For all $j = 1, \ldots, J, \bar{y}_j \in$

$$\arg\max \left\{ \sum_{l=1}^{L} \bar{p}_l y_1 | (y = y_1, \ldots, y_L) \in Y^j \right\}$$

(1)

For all *h*=1,â€¦, *H* $\bar{x}^h \in B^h(\bar{p})$, where $B^h(\bar{p})$

$$\equiv \left\{ x \in X^h | \sum_{l=1}^{L} \bar{p}_l x_l \leqq \sum_{l=1}^{L} \bar{p}_l e_l^h + \sum_{j=1}^{J} d^{hj} \sum_{l=1}^{L} \bar{p}_1 \bar{y}_l^j \right\}$$

(2)

and if $x \in B^h(\bar{p})$, then not $x \succ_h \bar{x}^h$,

$$\text{For all } l = 1, \ldots, L, \sum_{h=1}^{H} \bar{x}_l^h = \sum_{h=1}^{H} e_l^h + \sum_{j=1}^{J} y_l^{-j}.$$

(3)

The most striking feature of general equilibrium is the juxtaposition of the great diversity in goals and resources it allows, together with the supreme coordination it requires. Every desire of each consumer, no matter how whimsical, is met precisely by the voluntary supply of some producer. And this is true for all markets and consumers simultaneously.

There is a symmetry to the general equilibrium model, in the way that all agents enter the model individually motivated by self-interest (not as members of distinct classes motivated by class interests), and simultaneously, so that no agent acts prior to any other on a given market (e.g. by setting prices). If workersâ€™ subsistence were not assumed, for example, that would break the symmetry; workers income could have to be guaranteed first, otherwise demand would (discontinuously) collapse. As it is, at the aggregate level, supply and demand equally and simultaneously determine price; in equilibrium, both the consumersâ€™ marginal rates of substitution and the producersâ€™ marginal rates of transformation are equal to relative prices (assuming differentiability and interiority). There are gains to trade both through exchange and through production. This point of view represents a significant break with the classical tradition of Ricardo and Marx. We shall come to the main difference between the classical and neoclassical approaches shortly. Another difference is that there need not be fixed coefficients of production in the Arrowâ€“Debreu model â€“ the sets Y are much more general. Also in an Arrowâ€“Debreu equilibrium, there is no reason for there to be a uniform rate of profit. There is none the less one aspect of the model which these authors would have greatly approved, namely the shares d^{hj} which allow the owners of firms to collect profits even though they have contributed nothing to production.

Notice that in general equilibrium each agent need only concern himself with his own goals (preferences or profits) and the prices. The implicit assumption that every agent â€˜knowsâ€™ all the prices is highly non-trivial. It means that at each date each agent is capable of forecasting perfectly all future prices until the end of time. It is in this sense that the Arrowâ€“Debreu model depends on â€˜rational expectationsâ€™. Each agent must also be informed of the â€˜price q_j of each firm j, where $q_j = \Sigma_{l=1}^{L} \bar{p}_l \bar{y}_l^j$. (Firms that produce under constant returns to scale must also discover the level of production, which cannot be deduced from the prices alone.) Assuming that the â€˜man on the spotâ€™ (Hayek's expression) knows much better than anyone else what he wants, or best how his changing environment is suited to producing his product, decentralized decision making would seem to be highly desirable, if it is not incompatible with coordination. Indeed, harmony through diversity is one of the sacred doctrines of the liberal tradition.

The greatest triumph of the Arrowâ€“Debreu model was to lay out explicitly the conditions (roughly (A.1)â€“(A.13)) under which if is possible to claim that a properly chosen price system must always exist that, like the invisible hand, can guide diverse and independent agents to make mutually compatible choices. The idea of general equilibrium had gradually developed since the time of Adam Smith, mostly through the pioneering work of Walras (1874), von Neumann (1937), Hicks (1939) and Samuelson (1947). By the late 1940s the definition of equilibrium, including ownership shares in the firms, was well-established. But it was Arrowâ€“Debreu (1954) that spelled out precise microeconomic assumptions at the level of the individual agents that could be used to show the model was consistent. The axiomatic and rigorous approach that characterized the formulation of general equilibrium by Arrowâ€“Debreu has been enormously influential. It is now taken for granted that a model is not

properly defined unless it has been proved to be logically consistent. Much of the clamour for â€˜microeconomic foundations to macroeconomicsâ€™, for example, is a desire to see an axiomatic clarity similar to that of the Arrowâ€“Debreu model applied to other areas of economics. Of course, there were other earlier economic models that were similarly axiomatic and rigorous; one thinks especially of von Neumannâ€“Morgenstern's *Theory of Games* (1944). But game theory was, at the time, on the periphery of economics. Competitive equilibrium is at its heart.

The central mathematical techniques, convexity theory (separating hyperplane theorem) and Brouwer's (Kakutani's) fixed point theorem, used in Arrowâ€“Debreu are, 30 years later, still the most important tools used in mathematical economics. Both elements had played a (hidden) role in von Neumann's work. Convexity had been prominent in the work of Koopmans (1951) on activity analysis, in the work of Kuhn and Tucker (1951) on optimization, and in the papers of Arrow (1951) and Debreu (1951) on optimality. Fixed point theorems had been used by von Neumann (1937), by Nash (1950) and especially by McKenzie (1954), who one month earlier than Arrowâ€“Debreu had published a proof of general equilibrium using Kakutani's theorem, albeit in a model where the primitive assumptions were made on demand functions, rather than preferences. McKenzie (1959) also made an early contribution to the notion of an irreducible economy (assumption (A.9)).

The first fruit of the more precise formulation of equilibrium that began to emerge in the early 1950s was the transparent demonstration of the first and second welfare theorems that Arrow and Debreu simultaneously gave in 1951. Particularly noteworthy is the proof that every equilibrium is Pareto optimal. So simple and illuminating is this demonstration that it is no exaggeration to call it the most frequently imitated argument in all of neoclassical economic theory.

Among the confusions that were cleared away by the careful axiomatic treatment of equilibrium was the reliance of the discussions by Hicks and Samuelson on interior solutions and differentiability. When discussing the optimal allocation of housing, for example, it is evident that most agents will consume nothing of most houses, but this does not affect the Pareto optimality of a free (and complete) market allocation of housing. Similarly, it is not necessary to either the existence of Arrowâ€“Debreu equilibrium, nor to the first and second welfare theorems, that preferences or production sets be either differentiable or strictly convex. In particular, it is possible to incorporate the â€˜neoclassical production functionâ€™ with constant returns to scale with variable inputs, the classical fixed coefficients methods of production, and the strictly concave production functions of the Hicksâ€“Samuelson vintage, all in the same framework.

This is not to say that differentiability has no role to play in the Arrowâ€“Debreu model. In his seminal paper (1970), Debreu resurrected the role of differentiability by showing, via the methods of transversality theory (a branch of differential topology) that almost every differentiable economy is regular, in the sense that small perturbations to the economic data (e.g. the endowments) make small changes in all the equilibrium prices. Before Debreu, comparative statics could be handled only under specialized hypotheses, for example, the invertibility of excess demand at all prices, etc. We shall give a fuller discussion of the three crucial mathematical results of the Arrowâ€“Debreu model â€“ existence, optimality and local uniqueness â€“ in the next section.

Observe finally, that although the commodities may include physical goods dated over many time periods, there is only one budget constraint in an Arrowâ€“Debreu equilibrium. The income that could be obtained from the sale of an endowed commodity, dated from the last period, is available already in

the first period.

IVÂ Pareto optimality

The first theorem of welfare economics states that any Arrowâ€“Debreu equilibrium allocation $\bar{x} = (\bar{x}^h),\ h = 1, \ldots, H$ is Pareto optimal in the sense that if [(x^h), (y^j)] satisfies $y^j \in Y^j, \Sigma_{h=1}^{H} x^h = \Sigma y_h^j + e$, then it cannot be the case that $x^h \succ_h \bar{x}^h$ for all *h*. The second theorem of welfare analysis states the converse, namely that any Pareto optimal allocation for an Arrowâ€“Debreu economy *E* is a competitive equilibrium allocation for an Arrowâ€“Debreu economy *ÃŠ* obtained from *E* by rearranging the initial endowments of commodities and ownership shares.

The first welfare theorem expresses the efficiency of the ideal market system, although it makes no claim as to the justice of the initial distribution of resources. The second welfare theorem implies that any income redistribution is best effected through a lump sum transfer, rather than through manipulating the market, e.g. through rent control, etc.

The connection between competitive equilibrium and Pareto optimality has been perceived for a long time, but until 1951 there was a general confusion between the necessity and sufficiency part of the arguments. The old proof of Pareto optimality (see Lange, 1942) assumed differentiable utilities of production sets, and a strictly positive allocation $\bar{x}$. It noted the first order conditions to the problem of maximizing the *i*th consumer's utility, subject to maintaining all the others at least as high as they got under $\bar{x}$, and feasibility, are satisfied at $\bar{x}$, if and only if $\bar{x}$; is a competitive equilibrium allocation for a â€˜rearrangedâ€™ economy *ÃŠ*. This first order, or infinitesimal, proof of equivalence between competitive equilibrium and Pareto optimality could have been made global by postulating in addition that preferences and production sets are convex.

The Arrow and Debreu (1954) proofs of the equivalence between competitive equilibrium and Pareto optimality, under global changes, do not require differentiability, nor do they require that all agents consume a strictly positive amount of every good. In fact the proof of the first welfare theorem, that each competitive equilibrium is Pareto optimal, does not even use convexity.

The only requirement is local nonsatiation, so that every agent spends all his income in equilibrium. If (*x*, *y*) Pareto dominates the equilibrium allocation $(\bar{p}, \bar{x}, \bar{y})$, then for all $h,\ \bar{p} \cdot x^h < \bar{p} \cdot \bar{x}^h$. Since profit maximization implies that for all $j,\ \bar{p} \cdot \bar{y}^j \geqq \bar{p} \cdot y^j$, it follows that $\bar{p} \cdot (\Sigma_h x^h - \Sigma_j y^j) > \bar{p} \cdot (\Sigma_h \bar{x}^h - \Sigma_h \bar{y}^j)$, contradicting feasibility.

The proof of the second welfare theorem, on the other hand, does require convexity of the preferences and production sets (though not their differentiability, nor the interiority of the candidate allocation $\bar{x}$). Essentially it depends on Minkowski's theorem, which asserts that between any two disjoint convex sets in $\mathbb{R}^L$ there must be a separating hyperplane.

In this connection let us mention one more remarkable mathematical property of the Arrowâ€“Debreu model. Let us suppose that all production takes place under constant returns to scale: if $y \in Y^j$, then so is *Î»y*, for $\lambda \geq 0$. We say that a feasible allocation $\bar{x}$ for the economy *E* is in the core if there is no coalition of consumers $S \subset \{1, \ldots, H\}$ such that using only their initial endowments of resources, as well as access to all the production technologies, they cannot achieve an allocation for themselves which they all prefer

to $\bar{x}$. The core is meant to reflect those allocations which could be maintained when bargaining (the formation of coalitions) is costless. In a status quo core allocation, any labour union or cartel of owners that threatens to withhold its goods from the market knows that another coalition could form and by withholding its goods, prevent some members of the original coalition from being better off than they were under the status quo. It is easy to see that any competitive equilibrium is in the core. Debreuâ€“Scarf (1963), building on earlier work of Scarf, showed by using the separating hyperplane theorem, that if agents are small relative to the market, in the sense they made precise through the notion of replication, then the core consists only of competitive allocations. Such a theorem can also be proved even if there are small nonconvexities in preferences (see Aumann, 1964, for a different formulation of the small agent).

Existence of equilibrium

Suppose that agentsâ€™ preferences and firmsâ€™ production sets are strictly convex, and that agents strictly prefer more of any commodity to less (strict monotonicity) and that they all have strictly positive endowments. Let Î” be the set of *L*-price vectors, all non-negative, summing to one. Let $f^h(p)$ be the commodity bundle most preferred by agent *h*, given the strictly positive prices $p \in \Delta ++$. Similarly let gâ€²(p) be the profit maximizing choice of firm *j*, given prices $p \in \Delta ++$. Finally, let $f(p) = \Sigma_{h=1}^{H} f^h(p) - \Sigma_{j=1}^{j} g^j(p) - e$. It is easy to show that *f* is a continuous function at all $p \in \Delta ++$. A price $\bar{p} \in \Delta ++$ is an Arrowâ€“Debreu equilibrium price if and only if $f(\bar{p}) = 0$.
In general there is no reason to expect a continuous function to have a zero. Thus Wald could prove only with great difficulty in a special case that an equilibrium necessarily exists. Now observe that the function must satisfy Walras's Law, $p \cdot f(p) = 0$, for all *p*. So *f* is not arbitrary.
Consider the convex, compact set Δ_e of prices $p \in \Delta$ with $p_l \geqq \varepsilon > 0$, for all *l*. Consider also the continuous function $\varphi : \Delta_e \to \Delta_e$ mapping *p* to the closest point $\hat{p}$ in Δ_e to $f(p) + p$. By Brouwer's fixed point theorem, there must be some $\bar{p}$ with $\varphi(\bar{p}) = \bar{p}$. From strict monotonicity, it follows that $\bar{p}$ cannot be on the boundary of Δ_e, if μ is chosen sufficiently small. From Walras's Law it follows that if $\bar{p}$ is in the interior of Δ_e, then $f(\bar{p}) = 0$. The demonstration of the existence of equilibrium by Arrow and Debreu, as modified later by Debreu (1959), followed a similar logic.
Note the essential role of convexity in two parts of the above proof. It was used with respect to agentsâ€™ characteristics to guarantee that their optimizing behaviour is continuous. And it was also used to ensure that the space Î” has the fixed point property. Smale (1976) has given a path-following proof (related to Scarf's, 1973 algorithm) that on closer inspection does not require convexity of the price space. (Dierker, 1974, and Balasko, 1986, have given homotopy proofs.) This is not only of computational importance. It appears that there may be economic problems, dealing with general equilibrium with incomplete markets, in which the price space is intrinsically nonconvex, and in which the existence of equilibrium can only be proved using path-following methods (see Duffieâ€“Shafer, 1985).
To weaken the assumption of strict convexity, in the above proof, one can replace Brouwer's fixed point

theorem with Kakutani's. An important conceptual point arises in connection with strict monotonicity. If that is dropped, and the production sets do not have free disposal, then, in order to guarantee the existence of equilibrium, the definition must be revised to require either $f_l(\bar{p}) = 0$, or $f_l(\bar{p}) < 0$ and $\bar{p}_l = 0$. There may be free goods, like air, in excess supply. One cannot drop monotonicity and free disposal without allowing for negative prices.

Finally, it can be shown that if there are small nonconvexities in either preference or production, and if all the agents are small relative to the market (either in the replication sense of Debreuâ€“Scarf, or the measure zero sense of Aumann), then there will be prices at which the markets nearly clear. On the other hand, increasing returns to scale over a broad range is definitely incompatible with equilibrium.

Local uniqueness and comparative statics

Another property of the excess demand function $f(p)$ is that it is homogeneous of degree zero. So instead of taking pâ^^Î”, let us fix p_1=1. Similarly, let $F(p)$ be the Lâ^'1 vector of excess demands for goods l=2, â€¦, L. If $F(p)$=0, then by Walras's Law, $f(p)$=0.

Suppose furthermore that agent characteristics are smooth. Then $F(P)$ is a differentiable function. If $D_p F(\bar{p})$ has full rank at an equilibrium $\bar{p}$, then $\bar{p}$ is locally unique. Moreover, the equilibrium $\bar{p}$ will move continuously, given continuous, small changes in the agentsâ€™ characteristics, such as their endowments e. If $D_p F(\bar{p})$ has full rank at all equilibria $\bar{p}$, then there are only a finite number of equilibria. Debreu (1970) called an economy E regular if $D_p F(\bar{p})$ has full rank at all equilibrium $\bar{p}$ of E. The problem of trying to give sufficient conditions on preferences etc. to guarantee that $D_p F$ has full rank in equilibrium has proved intractable (except for restrictive, special cases). But Debreu (1970) solved the problem in classic style, appealing to the transversality theorem of differential topology (or Sard's theorem), to show that if one were content with regularity for â€˜almost allâ€™ economies, then the problem is simple. He proved that for almost all economies, $D_p F$ has full rank at every equilibrium. Hence, in almost all economies comparative statics (the change in equilibrium, given exogenous changes to the economy) is well defined.

Observe that excess demand F depends on the agentsâ€™ characteristics, including their endowments, so we could write $F(e, p)$. Now the transversality theorem says that (given some technical conditions) if $D_e F(e, \bar{p})$ has full rank at all equilibria $\bar{p}$ for the economy $E(e)$ with endowments e, for all e, then for â€˜almost allâ€™ e, $D_e F(e, \bar{p})$ has full rank at all equilibrium $\bar{p}$ for $E(e)$. But it is easy to show that $D_e F(e, \bar{p})$ always has full rank. Along similar lines, Debreu proved that the â€˜generic regularityâ€™ of equilibrium.

There is one unfortunate side to this comparative statics story. One would like to show not only that comparative statics are well defined, but also that they have a definite form. In a concave programming problem, for example, a small increase in an input results in a decrease in that input's shadow price, and an increase in output approximately equal to the size of the input increase multiplied by its original shadow price. Given the strong rationality hypothesis of the Arrowâ€“Debreu model, one would hope for some sort of analogous result. Following a conjecture of Sonnenschein, Debreu proved in 1974 that given any function $f(p)$ on $\hat{I}''_e$, satisfying Walras's Law, he could find an Arrowâ€“Debreu economy such that $f(p)$ is its aggregate excess demand on $\hat{I}''_e$. The assumptions (A.1)â€“(A.13) do not permit any a

priori predictions about the changes that must occur in equilibrium given exogenous changes to the economy. An increase in the aggregate endowment of a particular good, for example, might cause its equilibrium price to rise. The possibility of such pathologies is disappointing. It means that to make even qualitative predictions, the economist needs detailed data on the excess demands *F*.

VÂ What the model doesnâ€™t explain

We have already discussed the implications of the notion of Arrowâ€“Debreu commodities and the second welfare theorem for insurance, namely that since every Pareto optimal allocation is supportable as an Arrowâ€“Debreu equilibrium, every optimal allocation of risk bearing can be accomplished by the production and trade of Arrowâ€“Debreu commodities, i.e. without recourse to additional kinds of insurance markets specializing in risks. Every Arrowâ€“Debreu commodity is as much a diversifier in location, or time, or physical quality as it is for risk. This leads to a great simplification and economy of analysis. But it also means that, from the positive point of view, the Arrowâ€“Debreu economy cannot directly provide an analysis of insurance markets (except as a benchmark case). In this section I shall try to point out a few of the other phenomena which needle into the background in the Arrowâ€“Debreu model, but which would emerge if the assumption of a finite, but complete set of Arrowâ€“Debreu commodities, and consumers was dropped.

There are four currently active lines of research which attempt to come to grips in a general equilibrium framework with some of these phenomena, while preserving the fundamental neoclassical Arrowâ€“Debreu principles of agent optimization, market clearing, and rational expectations, that I think are particularly worthy of attention. They are the theory of general equilibrium with incomplete asset markets which can be traced back to Arrow's (1953) seminal paper on securities; overlapping generations economies, whose study was initiated by Samuelson (1958) in his classic consumption loan model; the Cournot theory of market exchange with few traders, first adapted to general equilibrium by Shapleyâ€“Shubik (1977), and the model of rational expectations equilibrium, pioneered by Lucas (1972).

Let us note first of all that in Arrowâ€“Debreu equilibrium there is no trade in shares of firms. A stock certificate is not an Arrowâ€“Debreu commodity, for its possession entitles the owner to additional commodities which he need not obtain through exchange. Note also that in Arrowâ€“Debreu equilibrium, the hypothesis that all prices will remain the same, no matter how an individual firm changes its production plan, guarantees that firm owners unanimously agree on the firm objective, to maximize profit. If there were a market for firm shares, there would not be any trade anyway, since ownership of the firm and the income necessary to purchase it would be perfect substitutes. In an incomplete markets equilibrium, different sources of revenue are not necessarily perfect substitutes. There could be active trade on the stock market. Of course, such a model would have to specify the firm objectives, since one would not expect unanimity. The theory of stock market equilibrium is still in its infancy, although some important work has already been done. (See DrÃ¨ze, 1974, and Grossman and Hart, 1979.)

Bankruptcy is not allowed in an Arrowâ€“Debreu equilibrium. That follows from the fact that all agents must meet their budget constraints. In a game theoretic formulation of equilibrium (such as I shall discuss shortly), it is achieved by imposing an infinite bankruptcy penalty. Since every Arrowâ€“Debreu

equilibrium is Pareto optimal, there would be no benefit in reducing the bankruptcy penalty to the point where someone might choose to go bankrupt. But with incomplete markets, such a policy might be Pareto improving, even allowing for the deadweight loss of imposing, the penalties.

Money does not appear in the Arrowâ€“Debreu model. Of course, all of the reasons for its life existence: transactions demand, precautionary demand, store of value, unit of account, etc. are already taken care of in the Arrowâ€“Debreu model. One could imagine money in the model: at data zero every agent could borrow money from the central bank. At every date afterwards he would be required to finance his purchases out of his stock of money, adding to that stock from his sales. At the last data he would be required to return to the bank exactly what he borrowed (or else face an infinite bankruptcy penalty). In such a model the Arrowâ€“Debreu prices would appear as money prices. The absolute level of money prices and the aggregate amount of borrowing would not be determined, but the allocations of commodities would be the same as in Arrowâ€“Debreu. There is no point in making the role of money explicit in the Arrowâ€“Debreu model, since it has no effect on the real allocations. However, if one considers the same model with incomplete asset markets, the presence of explicitly financial securities can be of great significance to the real allocations.

In the Arrowâ€“Debreu model, all trade takes place at the beginning of time. If markets were reopened at later dates for the same Arrowâ€“Debreu commodities, then no additional trade would take place anyway. At the other extreme, one might consider a model in which at every date and state of nature only those Arrowâ€“Debreu commodities could be traded which were indexed by the corresponding (date, state) pair. An intermediate case would also permit the trade of some (but not all) differently indexed Arrowâ€“Debreu commodities. Now the Arrowâ€“Debreu proofs of the existence and Pareto optimality of equilibrium do not apply to such an incomplete markets economy, as Hart (1975) first pointed out. We have already noted the existence problem. As for efficiency, the Pareto optimality of Arrowâ€“Debreu equilibria might suggest the presumption that, though there might be a loss to eliminating markets, trade on the remaining markets would be as efficient as possible. In fact, it can be shown (generically) that equilibrium trade does not make efficient use of the existing markets.

The Arrowâ€“Debreu model of general equilibrium is relentlessly neoclassical; in fact it has become the paradigm of the neoclassical approach. This stems in part from its individualistic hypothesis, and its celebrated conclusions about the potential efficacy of unencumbered markets. (Although Arrow, for example, has always maintained that a proper understanding of Arrowâ€“Debreu commodities is also useful for showing how inefficient is the limited real world market system.) But still more telling is the fact that the assumption of a finite number of commodities (and hence of dates) forces upon the model the interpretation of the economic process as a one-way activity of converting given primary resources into final consumption goods. If there is universal agreement about when the world will end, there can be no question about the reproduction of the capital stock. In equilibrium it will be run down to zero. Similarly when the world has a definite beginning, so that the first market transaction takes place after the ownership of all resources and techniques of production, and the preferences of all individuals have been determined, one cannot study the evolution of the social norms of consumption in terms of the historical development of the relations of production. One certainly cannot speak about the production of all commodities by commodities (Sraffa, 1960) (since at date zero there must be commodities which have not been produced by commodities, i.e. by physical objects which are traded).

It seems natural to suppose that as L becomes very large, so that the end of the world is put off until the distant future, that this event cannot be of much significance to behaviour now. But let us not forget the

rationality imposed on the agents. Far off as the end of the world might be, it is perfectly taken into account. Thus, for example, social security (funded as it is in the US by taxes on the young) could not exist if rational agents agreed on a final stopping time to transactions.

Consider a model satisfying all the assumptions (A.1)â€“(A.13), except that *L* and *H* are allowed to be infinite, such as the overlapping generations model. It can be shown that there is a robust collection of economies which have a continuum of equilibria, most of which are Pareto sub-optimal, which differ enormously in time 0 behaviour. Thus in a model where time does not have a definite end, the optimality and comparative statics properties of equilibria are radically different. (For example, there may be a continuum of equilibria, indexed by the level of period 0 real wages â€“ inversely related to the rate of profit â€“ or the level of output or employment. The interested reader can consult the entry on the overlapping generations model of general equilibrium. A systematic study of economies where only *L* is allowed to be infinite was begun by Bewley, (1972). Such economies tend to have properties similar to those of Arrowâ€“Debreu.)

There is no place in the Arrowâ€“Debreu model for asymmetric information. The second welfare theorem, for example, relies on lump sum redistributions, i.e. redistributions that occur in advance of the market interactions. But if agents cannot be distinguished except through their market behaviour, then the redistribution must be a function of market behaviour. Rational agents, anticipating this, will distort their behaviour and the optimality of the redistribution will be lost.

Similarly, in the definition of equilibrium no agent takes into account what other agents know, for example about the state of nature. Thus it is quite possible in an Arrowâ€“Debreu equilibrium for some ignorant agents to exchange valuable commodities for commodities indexed by states that other agents know will not occur. This problem received enormous attention in the finance literature, and some claim (see Grossman, 1981) that it has been solved by extending the Arrowâ€“Debreu definition of equilibrium to a â€˜rational expectations equilibriumâ€™ (Lucas, 1972; see also Radner, 1979). But this definition is itself suspect; in particular, it may not be implementable.

Even if rational expectations equilibrium (REE) were accepted as a visible notion of equilibrium, it could not come to grips with the most fundamental problems of asymmetric information. For like Arrowâ€“Debreu equilibrium, in REE all trade is conducted anonymously through the market at given prices. Implicit in this definition is the assumption of large numbers of traders on both sides of every market. But what has come to be called the incentive problem in economics revolves around individual or firm specific uncertainty, i.e. trade in commodities indexed by the names of the traders, which by definition involves few traders.

This brings us to another major riddle: how are agents supposed to get to equilibrium in the Arrowâ€“Debreu model? The pioneers of general equilibrium never imagined that the economy was necessarily in equilibrium; Walras, for example, proposed an explicit tÃ¢tonnement procedure which he conjectured converged to equilibrium. But that idea is flawed in two respects: in general, it can be shown not to converge, and more importantly, it is an imaginary process in which no exchange is permitted until equilibrium is reached. This illustrates a grave shortcoming of any equilibrium theory, namely that it cannot begin to specify outcomes out of equilibrium. The major crisis of labour market clearing in the 1930s, and again recently, argues strongly that there are limits to the applicability of equilibrium analysis.

One is led naturally to consider market games, in which the outcomes are well-specified even when agents do not make their equilibrium moves. The most famous market game is Cournot's duopoly

model, which has been extended to general equilibrium by Shapleyâ€“Shubik (1977). When there are a large number of agents of each type, the Nash equilibria of the Shapleyâ€“Shubik game give nearly identical allocations to the competitive allocations of Arrowâ€“Debreu. This justifies (to first approximation) the price taking behaviour of the Arrowâ€“Debreu agents. But note that the informational requirements of Nash equilibrium are at least twice that of Arrowâ€“Debreu competitive equilibrium (each agent must know the aggregates of birds and offers on each market). It is also extremely interesting that trade takes place in the Shapleyâ€“Shubik game even if there is only one trader on each side of the market. Hence many problems in asymmetric information which have no place in the Arrowâ€“Debreu model, because they involve too fine a specification of the commodities to be consistent with price taking, might be sensible in a market game context. Finally, it can be shown that REE is not consistent with the Shapleyâ€“Shubik game, or indeed with any continuous game.

We have indicated some of the ways in which it is possible to extend general equilibrium analysis to phenomena outside the scope of the Arrowâ€“Debreu model, while at the same time preserving the neoclassical methodological premises of agent optimization, rational expectations, and equilibrium. It is important to note that these variations have extended the definition of equilibrium as well; this is most obvious in the case of market games, where Nash equilibrium replaces competitive equilibrium. All of the models have retained, on the other hand, more or less the same notion of rationality, sometimes at the cost of increasing the demands on the rationality of expectations. A great challenge for future general equilibrium models is how to formulate a sensible notion of bounded rationality, without destroying the possibility of drawing normative conclusions.

See Also

- existence of general equilibrium
- general equilibrium
- intertemporal equilibrium and efficiency
- overlapping generations model of general equilibrium

Bibliography

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press, 507â€“32.

Arrow, K.J. 1953. Le rÃ´le des valeus boursiÃ¨res pour la rÃ©partition la meilleure des risques. *EconomÃ©trie*. Paris: Centre National de la Recherche Scientifique, 41â€“8.

Arrow, K.J. 1969. The organization of economic activity: issues pertinent to the choice of market vs nonmarket allocation. Reprinted in *Collected Papers of Kenneth Arrow*, Cambridge, MA: Belknap Press, Vol. II, 133â€“55.

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265â€“90.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39â€“50.

Balasko, Y. 1986. *Foundations of the Theory of General Equilibrium*. New York: Academic Press.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4, 514â€“40.

Cournot, A. 1838. *Recherches sur les principes mathÃ©matiques de la thÃ©orie des richesses*. Paris: L. Hachette.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19, 273â€“92.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New York: Wiley.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387â€“92.

Debreu, G. 1974. Excess functions. *Journal of Mathematical Economics* 1, 15â€“21.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235â€“46.

Dierker, E. 1974. *Topological Methods in Walrasian Economics*. Berlin: Springer.

DrÃ¨ze, J. 1974. Investment under private ownership: optimality, equilibrium, and stability. In *Allocation under Uncertainty*, ed. J. DrÃ¨ze. New York: Macmillan.

Duffie, D. and Shafer, W. 1985. Equilibrium in incomplete markets: I-A basic model of generic existence. *Journal of Mathematical Economics* 14(3), 285â€“300.

Geanakoplos, J.D. and Polemarchakis, H.M. 1987. Existence, regularity, and constrained suboptimality of equilibrium with incomplete asset markets. In *Essays in Honor of Kenneth J. Arrow*, vol. 3, ed. W. Heller, R. Starr and D. Starrett. New York: Cambridge University Press.

Grossman, S. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48, 541â€“60.

Grossman, S. and Hart, O. 1979. A theory of competitive equilibrium in stock market economies. *Econometrica* 47, 293â€“329.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11(3), 418â€“33.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Koopmans, T.C., ed. 1951. *Activity Analysis of Production and Allocation*. New York: Wiley.

Kuhn, H.W. and Tucker, A.W. 1951. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press, 481â€“92.

Lange, O. 1942. The foundations of welfare economics. *Econometrica* 10, 215â€“28.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103â€“24.

McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22, 147â€“61.

McKenzie, L.W. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27, 54â€“71.

Nash, J.F. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the USA* 36, 48â€“9.

Neumann, J. von. 1928. Zur theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295â€“320.

Neumann, J. von. 1937. Uber ein Ã¶konomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, 73â€“83.

Neumann, J. von. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Pareto, V. 1909. *Manuel dâ€™Ã©conomie politique*. Paris: Giard.

Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36(1), 31â€“58.

Radner, R. 1979. Rational expectations equilibrium: generic existence and the information revealed by prices. *Econometrica* 17, 655â€“78.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66, 467â€“82.

Scarf, H. (with the collaboration of T. Hansen.) 1973. *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Shapley, L. and Shubik, M. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 85, 937â€“68.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3, 107â€“20.

Sonnenschein, H. 1973. Do Walrasâ€™ identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6, 345â€“54.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Walras, L. 1874â€“7. *ElÃ©ments dâ€™Ã©conomie politique pure*. Lausanne: L. Corbaz.

How to cite this article

Geanakoplos, John. "Arrowâ€“Debreu model of general equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000133>
doi:10.1057/9780230226203.0059

The New Palgrave Dictionary of Economics Online

Arrow's theorem

Kenneth J. Arrow
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Any satisfactory method of making a social choice should be in some measure representative of the individual criteria which enter into it, should use the range of possible actions, and should observe consistency conditions among the choices made for different data sets. Arrow's Theorem, or the Impossibility Theorem, states that there is no social choice mechanism which satisfies such reasonable conditions and which will be applicable to any arbitrary set of individual criteria. This article sets out the proof of the theorem.

Keywords

Arrow's th; Bentham, J.; compensation principle; Condorcet, Marquis de; constitutions; Edgeworth, F.; Hicks, J.; impossibility th; independence of irrelevant alternatives; intransitivity; Kaldor, N.; Pareto principle; preference orderings; Rawls, J.; Scitovsky, T.; Sidgwick, H.; single-peaked preferences; social choice; social welfare function; voting

Article

Economic or any other social policy has consequences for the many and diverse individuals who make up the society or economy. It has been taken for granted in virtually all economic policy discussions since the time of Adam Smith, if not before, that alternative policies should be judged on the basis of their consequences for individuals; political discussions are less uniform in this respect, the welfare of an abstract entity, the state or nation, playing a role occasionally even in economic policy.
It follows that there are as many criteria for choosing social actions as there are individuals in the society. Furthermore, these individual criteria are almost bound to be different in some measure so that there will be pairs of policies such that some individuals prefer one and some the other. In the economic context, policies invariably imply distributions of goods, and in most policy choices, some individuals will receive more goods under one policy and others under the other. Individuals may also have different

evaluations because of different concepts of justice or other social goals.

The individual criteria may be based on individual preferences over bundles of goods or individual preferences of a more social nature, with preferences over goods supplied to others. From the viewpoint of the formal theory of social choice, the criteria may even be judgements by others as to the welfare of individuals. The only assumption is that there is associated with each individual a criterion by which social actions are evaluated for that individual. Whatever their origin, these criteria differ from individual to individual.

Every society has a range of actions, more or less wide, which are necessarily made collectively. Much of the debate on the foundations of social decision theory began with criteria for evaluating alternative tariff structures, including as the most famous illustration moving from a tariff to free trade. The redistribution of income through governmental taxes and subsidies provides another important case of an inherently collective decision which would be judged differently by different individuals.

If every individual prefers one policy to another, it is reasonable to postulate, as is always done by economists, that the first policy should be preferred. The problem arises of making social choices (between alternative collective policies) when some individual criteria prefer one policy and some another.

The fundamental question of social choice theory, then, is the following: given a range of possible social decisions, one of which has to be chosen, and given the criteria associated with the individuals in the society, find a method of making the choice. Not all methods of decision would be regarded as satisfactory. The method should be in some measure representative of the individual criteria which enter into it. For example, we would want the Pareto condition to be satisfied, that an alternative not be chosen if there is another preferred by all individuals. The method should use all the data, that is, both the range of possible actions and the individual criteria, and there are consistency conditions among the choices made for different data sets.

A pure case of social choice in action is voting, whether for the election to an office or a legislative decision. Here, the candidates or alternative legislative proposals are evaluated by each voter, and the evaluations lead to messages in the form of votes. The social decision, which candidate to elect or which bill to pass, is made by aggregating the votes according to the particular voting scheme used. The social decision then depends on both the range of alternatives (candidates or legislative proposals) available and the ranking each voter makes of the alternatives.

Voting procedures have one very important property which will play a key role in the conditions required of social choice mechanisms: only individual voters’ preferences about the alternatives under consideration affect the choice, not preferences about unavailable alternatives.

Arrow's Theorem, or the Impossibility Theorem, states that there is no social choice mechanism which satisfies a number of reasonable conditions, stated or implied above, and which will be applicable to any arbitrary set of individual criteria.

Some terminology will be introduced in section 1 of this entry. In section 2, there will be a brief review of the relevant literature as it was known to me prior to the discovery of the theorem. In section 3, I state the theorem with some variants and discuss the meaning of the conditions on the social choice mechanisms.

1 The language of choice

The formulation of choice and the criteria for it are those standard in economic theory since the ‘marginalist revolution’ of the 1870s as subsequently refined. There is a large set of conceivable *alternatives*; in any given decision situation, some given subset of these alternatives is actually available or feasible. This subset will be referred to as the *opportunity set*. Each individual can evaluate all alternatives. This is expressed by assuming that each individual has a *preference ordering* over the set of all alternatives. That is, for each pair of alternatives, the individual either prefers one to the other or else is indifferent between them (completeness), and these choices are consistent in the sense that if alternative x is preferred or indifferent to alternative y and y is preferred or indifferent to z, then x is preferred or indifferent to z (transitivity). This preference ordering is analogous to the preference ordering over commodity bundles in consumer demand theory. I have adopted the ordinalist viewpoint that only the ordering itself and not any particular numerical representation by a utility function is significant.

The *profile* of preference orderings is a description of the preference orderings of all individuals. For a given profile, the social choice mechanism will determine the choice of an alternative from any given opportunity set. In the case of an individual, it is assumed that the choice made from any given set of alternatives is that alternative which is highest on the individual's preference ordering. Analogously, it is assumed that social choices can be similarly rationalized. The social choice mechanism will have to be such that there exists a *social ordering* of alternatives such that the choice made from any opportunity set is the highest element according to the social ordering.

Therefore, a social choice mechanism or *constitution* is a function which assigns to each profile a social ordering.

2 The relevant literature

I will here review the literature on the justification of economic policy as I knew it in 1948–50. There was some work in economics and more in the theory of elections of which I was unaware, which I will briefly note.

The best-known criterion for what is now known as social choice was Jeremy Bentham's proposal for using the sum of individuals’ utilities. Curiously, despite its natural affinity with marginal economics, it received very little serious use, possibly because its distributional implications were unacceptably extreme. Edgeworth applied the criterion to taxation (1925: originally published in 1897): see also Sidgwick (1901, ch. 7).

The use of the sum-of-utilities criterion required interpersonally comparable cardinal utility. A reluctance to make interpersonal comparisons led to the proposal of the compensation principle by Kaldor (1939) and Hicks (1939). Consider a choice between a current alternative x and a proposed chance to another alternative y. In general, some individuals will gain by the change and some will lose. The compensation principle asserts that the change should be made if the gainers *could* give up some of their goods in y to the losers so as to make the losers better off than under x without completely wiping out the gains to the winners. Notice that the compensation is potential, not actual. Since the only information used is the preference relation of each individual among three different alternatives, x, y, and a potential alternative derived from y by transfers of goods, no interpersonal comparisons are needed. However, it turns out that the compensation principle does not define a social ordering. Indeed,

Scitovsky (1941) showed that it was possible that the compensation principle would call for changing from x to y and then from y to x.

A different approach which sought to avoid not only interpersonal comparisons but also cardinal utility was the social welfare function concept of Bergson (1938). For each individual, first choose a utility function which represents his or her preference ordering. Then define social welfare as a prescribed function $W(U_1,\ldots, U_n)$ of the utilities of the n agents. For a given profile of preference orderings, if one of the utility functions is replaced by a monotone transformation (which represents therefore the same preference ordering), the function W has to be transformed correspondingly, so that social preferences defined by W are unchanged. In this formulation, a given social welfare function is associated with a given profile. There are no necessary relations among social welfare functions associated with different profiles.

It was also known to me, though I do not know how, that majority voting, which could be considered as a social decision procedure, might lead to an intransitivity. Consider three voters A, B and C and three alternatives, a, b and c. Suppose that A has preference ordering abc, B has ordering bca, and C the ordering cab. Then a majority prefer a to b, a majority prefer b to c, and a majority prefer c to a. Therefore, if we interpret a majority for one alternative to another as defining social preference, the relation is not an ordering. This paradox had in fact been discovered by Condorcet (1785), and there had been a small and sporadic literature in the intervening period (for an excellent survey, see Black, 1958, Part II; also, Arrow, 1973), but all of this literature was unknown to me when developing the Impossibility Theorem.

There was one further very important paper, which I did know, the remarkable paper of Black (1948) on voting under single-peaked preferences. Suppose the set of alternatives can be represented in one dimension, for example, a choice among levels of expenditure (this was the case studied by Bowen (1943) who anticipated part of Black's results). Suppose individuals have different preference orders over the alternatives, but these preferences have a common pattern; namely, there is a most preferred alternative from which preference drops steadily in both directions. Put another way, of any three alternatives, the one in the intermediate position is never inferior to both of the others. Under this *single-peakedness* condition, majority voting defined a transitive relation and therefore an ordering. Hence, if the preferences of individuals are restricted to satisfy the single-peakedness condition, there does exist a constitution as defined earlier.

3 Statement of the impossibility th

I now state formally the conditions to be imposed on constitutions and then state the Impossibility Theorem, which simply asserts the non-existence of constitutions satisfying all of the conditions. The theorem as stated in the original paper (Arrow, 1950) and in a subsequent book (Arrow, 1951) is not correct as written, as shown by Blau (1957). To avoid confusion, I give a corrected statement and then explain the error.

Condition U: The constitution is defined for all logically possible profiles of preference orderings over the set of alternatives.

Condition M (Monotonicity): Suppose that x is socially preferred to y for a given profile. Now suppose a

new profile in which x is raised in preference in some individual orderings and lowered in none. Then x is preferred to y in the social ordering associated with the new profile.

Condition I (Independence of Irrelevant Alternatives): Let S be a set of alternatives. Two profiles which have the same ordering of the alternatives in S for every individual determine the same social choice from S.

To state the next condition, it is necessary to define an imposed constitution as one in which there is some pair of alternatives for which the social choice is the same for all profiles.

Condition N (Non-imposition): The constitution is not imposed.

A constitution is said to be *dictatorial* if there is some individual, any one of whose strict preferences is the social preference according to that constitution.

Condition D (Non-dictatorship): The constitution is not dictatorial.

Theorem 1: : There is no constitution satisfying Condition *U, M, I, N and C*.

A sketch of the argument can be given. From Condition *I*, the preference between any two alternatives depends only on the preferences of individuals between them and not on preferences about any other alternatives. Define a set of individuals to be *decisive* for alternative x against alternative y if the social preference is for x against y whenever all the individuals in the set prefer x to y. First, it can be shown that a set which is decisive for one alternative against one other is decisive for any alternative against any other. Hence, we can speak of a set of individuals as being decisive or not without reference to the alternatives being considered. If a set is not decisive, its complement (the voters not in the given set) can guarantee a weak preference, that is, preference or indifference. The set of all voters can easily be shown to be decisive, so there are decisive sets. The second stage in the proof is to take a decisive set with as few members of possible. If there were only one member, then by definition there would be a dictator, contrary to Condition *D*. Therefore, split the smallest decisive set so chosen into two subsets, say V_1 and V_2, and let V_3 contain all other voters. We now use an argument similar to that which showed the intransitivity of majority voting. Take any three alternatives, x, y and z. Suppose the members of V_1 all have the preference ordering, xyz, the members of V_2 the ordering yzx, and the members of V_3 the ordering zxy. Since V_1 and V_2 each have fewer members than the smallest decisive set, neither is decisive. Since all voters other than those in V_2 prefer x to y, x must be preferred or indifferent socially to y. Since V_1 and V_2 together constitute a decisive set and y is preferred to z in both sets, y must be preferred socially to z. By transitivity, then, x is socially preferred to z. But x is preferred to z only by the members of V_1, which would therefore be decisive for x against z and hence a decisive set. This, however, contradicts the construction that V_1 is a proper subset of the smallest decisive set and therefore is not a decisive set. The theorem is therefore proved.

Notice that Condition *U*, that the constitution be defined for all profiles, is essential to the argument. We consider the consequences of particular profiles.

In Arrow (1951, p. 59), the theorem is stated with a weaker version of Condition *U* (and a corresponding restatement of Condition *M*).

Condition U′ : The constitution is defined for a set of profiles such that, for some set of three alternatives, each individual can order the set in any way.

Since the contradiction requires only three alternatives, I supposed that the more general assumption would be sufficient. This is not so, as first pointed out by Blau (1957). The reason is that the non-

dictatorship Condition *D* may hold for the set of all alternatives and not hold for a subset, such as the triple of alternatives just described. To illustrate, suppose there are four alternatives altogether. Let *S* be a set of three of them, and let *w* be the fourth. Suppose each individual may have any ordering such that *w* is either best or worst. There are two individuals in the society. The constitution provides that the social preference between any pair in *S* follows the preferences of individual 1, but *w* is best or worst according to individual 2's preference ordering. This constitution would satisfy all the conditions of the Arrow 1951 version and therefore provides a counter-ex. What is true, of course, is that individual 1 is a dictator over the alternatives in *S*. If we still wish to retain the weaker Condition *U′*, the theorem remains valid if a stronger non-dictatorship condition is imposed (see Murakami, 1961).

Condition D′: No individual shall be a dictator over any three alternatives.

The conditions are fairly straightforward and need little comment. If it is reasonable to limit the range of possible individual orderings because of prior knowledge about the range of possible beliefs, then Condition *U* or *U′* could be replaced by a corresponding range condition. As has already been rmked, if preference orderings are restricted to the single-peaked type, then majority voting defines a constitution. There has been a considerable literature on range restrictions which imply that majority voting defines a constitution and some on more general voting methods. In a world of multi-dimensional issues, these restrictions are not particularly persuasive.

Conditions *M* and *N* embody different aspects of the value judgement that social decisions are made on behalf of the members of the society and should shift as values shift in a corresponding way. Condition *D* expresses a very minimal degree of democracy.

Condition *I* (independence of irrelevant alternatives) is central to the social choice approach whether in the Impossibility Theorem or in other, more positive, results. It is implicit in Rawls's difference principle of justice (Rawls, 1971), as well as in utilitarianism or methods based on voting.

The above conditions have not included the Pareto principle explicitly.

Condition P: If every individual prefers *x* to *y*, then *x* is socially preferred to *y*.

It is not hard to prove, however, that this condition is implied by some of the previous conditions, specifically Conditions *M*, *I* and *N*. Further, if the Pareto condition is imposed, then the Impossibility Theorem holds without assuming Monotonicity or Non-imposition. Of course, it is obvious that the Pareto principle implies Non-imposition, since any choice can be enforced by unanimous agreement.

Theorem 2: : there is no constitution satisfying Conditions *U, P, I, and D*.

This entry has dealt with Arrow's theorem itself and not with subsequent developments, which have been very abundant. The reader is referred to the entry on social choice in this work, and the surveys by Sen (1986) and Kelly (1978).

See Also

- social choice
- social welfare function
- welfare economics

Bibliography

Arrow, K.J. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy* 58, 328–46.

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley. (2nd edn, New Haven: Yale University Press, 1963, identical except for an additional chapter.)

Arrow, K.J. 1973. Formal theories of social welfare. In *Dictionary of the History of Ideas*, ed. P.P. Wiener, New York: Scribner's, Vol. 4, 276–84.

Bergson (Burk), A. 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52, 310–34.

Black, D. 1948. On the rationale of group decision-making. *Journal of Political Economy* 56, 23–34.

Black, D. 1958. *Theory of Committees and Elections*. Cambridge: Cambridge University Press.

Blau, J. 1957. The existence of social welfare functions. *Econometrica* 25, 302–13.

Bowen, H.R. 1943. The interpretation of voting in the allocation of economic resources. *Quarterly Journal of Economics* 58, 27–48.

Condorcet, M. de Caritat, Marquis de. 1785. Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix. Paris: Imprimerie Royale.

Edgeworth, F.Y. 1897. The pure theory of taxation. In F.Y. Edgeworth, *Papers Relating to Political Economy*, London: Macmillan, 1925, Vol. II, 63–125.

Hicks, J.R. 1939. The foundations of welfare economics. *Economic Journal* 49, 696–700, 711–12.

Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49, 549–52.

Kelly, J.S. 1978. *Arrow Impossibility Theorems*. New York: Academic Press.

Murakami, Y. 1961. A note on the general possibility theorem of social welfare function. *Econometrica* 29, 244–6.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Scitovsky, T. 1941. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9, 92–5.

Sen, A.K. 1986. Social choice theory. In *Handbook of Mathematical Economics*, ed. K.J. Arrow and M.

Intriligator, Amsterdam, North-Holland, Vol. III, 1073–81.

Sidgwick, H. 1901. *Principles of Political Economy*. 3rd edn, London: Macmillan.

How to cite this article

Arrow, Kenneth J. "Arrow's theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000136> doi:10.1057/9780230226203.0061

The New Palgrave Dictionary of Economics Online

art, economics of

David Throsby
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The application of economic theory and analysis to problems in the performing arts (music, theatre, dance), the visual and literary arts and other art forms has expanded greatly over the last 30 years. A basic issue has been to identify the ways in which artistic goods and services differ from other goods and services in the economy, thereby warranting particular attention. This article considers the economic analysis of demand and supply conditions in the arts, market structures (including factor markets) and a range of public policy issues in the area of cultural policy.

Keywords

advertising; and cultural goods; art, economics of; Baumol's disease; capital goods; cultural goods as; charitable donations; and tax; contingent valuation; copyright; and ‘fair use’; cultural capital; cultural goods and services; defined; *droit de suite* ; education; and arts demand; herding; heritage assets; intellectual property; and cultural goods; joint consumption; market structure; monopolistic competition; in live performing arts; moral rights; non-participant benefits; non-pecuniary motives; in the arts; not-for-profit firms; output quality; vs price as determinant of demand; path dependence; product differentiation; superstars; tax deductibility

Article

The definition of art has been a philosophical conundrum for centuries, but there is probably a reasonable consensus on what comprises ‘the arts’. These include the performing arts (music, dance, opera and theatre), the visual and plastic arts (painting, drawing, print-making, photography, sculpture, craft, and so on), the literary arts (poetry, fiction, drama, screenplays, and some forms of non-fiction such as biography), certain types of film, and some emerging practices such as video art that derive from new information and communications technologies. The application of economic theory and analysis across these various art forms comprises the discipline that has come to be known as cultural economics,

although the ambit of this field has expanded in recent years to embrace wider economic questions relating to culture in an anthropological sense, such as the role of culture in economic development. Apart from some issues relating to the definition of cultural goods, this contribution does not deal with culture in the broader sense but rather is confined to the arts as defined above, and considers the conditions of demand, supply and exchange of artistic products, and some consequent issues for policy.

Characteristics of cultural goods

The goods and services produced by the arts, as well as some neighbouring commodities such as television programmes, video games and heritage services, can be called cultural goods and services. A fundamental question is whether such goods have unique characteristics that distinguish them as a commodity class from other goods and services in the economy. A reasonable definition of cultural goods attributes to them three necessary features: they require some input of human creativity in their manufacture; they possess or convey some symbolic meaning or messages; and they contain, at least potentially, some form of intellectual property. This definition extends to include a wide range of goods with only minor cultural content, such as fashion design, some forms of advertising, and some architectural services. Nevertheless, while there may be some blurring of boundaries at the cultural edges, there is little doubt that goods and services produced by the arts, as a subset of cultural goods, fit this definition nicely.

An alternative (or perhaps additional) definitional approach has been to portray cultural goods as embodying or giving rise to a form of value that lies beyond the reach of conventional economic assessment, and is not expressible (or is only imperfectly expressible) in market prices or in individual willingness-to-pay judgements. In the case of art works, such ‘cultural’ value might derive from ineffable aesthetic or spiritual qualities that such works of art are known to possess. These sources of value are only partially comprehensible within standard neoclassical price theory; indeed, they can be fully understood only by extending the analytical range to wider areas of economics, and beyond economics into other disciplines such as philosophy, psychology and aesthetics.

A further distinctive characteristic of the arts as consumption goods is that they are subject to the phenomenon of path dependence or, more specifically, rational addiction; that is, they are commodities for which an individual's present consumption depends on his or her past consumption, and patterns of demand tend to be cumulative. Although it is generally agreed that increased exposure to the arts in the past and the present will generate increased demand in the future (with consequent lessons for arts in education), this is hardly a sufficient condition for defining artistic goods, since a number of other commodities, not least addictive drugs, share a similar characteristic.

As economic commodities it is appropriate to categorize cultural goods as being capital goods, intermediate goods, or goods for final consumption. When classified as capital items (reusable goods whose services are combined with other inputs to produce further outputs), cultural goods have come to be known within economics as cultural capital, distinguished from other forms of capital by reference to either or both of the above definitions. This concept is especially relevant in the analysis of artworks and cultural heritage, where the interpretation of tangible or intangible cultural property as long-lasting assets created by the investment of resources, subject to depreciation unless properly maintained and yielding a rate of return over time, is readily understood.

It is important to note that cultural goods are generally very heterogeneous, suggesting that working in

characteristics space may be a preferred way to analyse their demand and supply. For instance, demand for paintings can be thought of in Lancastrian terms as determined by the works' colour, size, style, school, and so on, and similar collections of characteristics can readily be imagined for other types of artistic commodities. Nevertheless, such heterogeneity does not vitiate the application of the tools of demand and supply analysis to the arts, as demonstrated further below.

Demand

A demand function for any type of artistic good or service could be expected to contain the usual sorts of explanatory variables: own price, price of substitutes, product quality characteristics and socio-demographic indicators relating to consumers' age, gender, income, education, and so forth. Within standard demand models, interest has focused on empirical questions: price and income elasticities, the relative importance of education and income, the cost of time, and the influence of quality aspects (to the extent that they can be measured). Results from a variety of art forms, time periods, geographical locations and data sources have varied widely, and even apparently plausible hypotheses, such as that the arts are a luxury good, have been by no means universally upheld. Nevertheless, the weight of evidence suggests, inter alia, that education is generally a more powerful predictor of arts demand than is income, and that output quality characteristics exert a strong influence on consumption patterns, perhaps overshadowing price as a determinant of demand behaviour in particular circumstances.

One topic of considerable interest in the demand for the performing arts is the emergence of so-called superstars, performers such as rock musicians and film actors whose incomes are greater than those of their competitors by a much larger differential than marginal productivity theory would suggest. Rosen (1981) attributed this phenomenon to two features of the demand for superstars' services. First, since consumers rationally prefer one good performance to two mediocre ones, particular types of services (such as rock music) are imperfect substitutes on the demand side, leading to convexity in sellers' returns and to a skewness in the distribution of earnings. Second, scale economies in joint consumption allow relatively few sellers to supply the entire market. Add to this the possible 'herding' behaviour of consumers, who follow the lead of others in making their demand decisions, and a plausible explanation as to why some performers command excessively high rents is obtained. Paradoxically, however, having broken away from the pack, superstars may finish up receiving *less* than their full earnings potential because some of their incremental contribution may have to be shared with employers, agents, managers and other beneficiaries of their superstardom.

Compared with the performing arts, the demand for art objects such as paintings – occurring in what is generally known as 'the art market' – raises some quite different questions. Durable works of art are sought by buyers not just for their aesthetic qualities but also because they are financial assets whose value may appreciate over time. Demand for paintings, prints, drawings, movable sculptures and other collectables such as silverware and rare books is readily separable into demand for art as a source of aesthetic gratification and demand for art as financial instrument. Both demands are affected by some of the same sorts of considerations – the reputation of the artist, the opinion of critics and market analysts, fashions in taste, past prices, and so on. At the same time other influences affect one or other aspect of demand specifically; for instance, demand for art as asset is constrained by some unattractive features of works of art as investments compared with alternative instruments, in particular their indivisibility, their illiquidity and their riskiness. In freely functioning markets, prices are expected to reflect all these

influences, providing in equilibrium a means of balancing their respective importance. Since quite extensive and detailed data on prices in various art markets are available, a substantial econometric effort has been devoted to analysing price patterns across time and space for a wide range of types and styles of works of art. While much of this research yields results of interest only to art market specialists and connoisseurs – for example, do prices for paintings and prints by the same artist follow similar trends? – some of it addresses the more general issue of rates of return to art investment over time. Although contrary examples can be found, the general conclusion is that a collection of works of art will yield a lower return over the long term than a corresponding portfolio of stocks and bonds, the differential being attributable in part to the consumption services provided by the art for the period for which it is held.

Finally on the demand side, we can point to the demand for museum and heritage services. This demand includes attendances at art museums and heritage sites which provide private consumption experiences to the visitor, the specialist demand for conservation and restoration services provided by curators, art historians, and so on who staff the institutions concerned, and the demand for the public-good output of these cultural facilities, seen in the form of non-participant benefits accruing to the local and wider communities. With regard to direct visits to museums and sites, empirical experience suggests some price sensitivity, leading to arguments for free admission to publicly funded or operated facilities on the grounds that their educational and access benefits outweigh their potential for revenue raising. Nevertheless, in some instances, especially in the heritage field, revenue from visitors such as tourists is the only reliable source of ongoing funds for restoring or maintaining the facility concerned. However, regardless of the income-earning prospects of museum and heritage assets, the demand for their public-good output may well prove more decisive than the private-use demand for their services in rationalizing their existence in economic terms. In this respect demand estimation methods using stated preference techniques such as contingent valuation methods have proved useful in evaluating option, existence and bequest demands for these items of cultural capital and in quantifying willingness to pay for their services.

Supply

Artistic goods and services for final consumption are produced by a variety of types of enterprises ranging from single-person firms through small for-profit and not-for-profit companies to large corporate organizations in both private and public sectors. At the simplest end of this spectrum is the individual artist who produces goods or services for direct sale to the public – the visual artist selling paintings from her home, or the busker playing his saxophone in the shopping mall. From an economic viewpoint these artists can be seen as single-proprietor firms, probably unincorporated and subject to more than the usual vagaries of production, cost and market uncertainties that attend such producers elsewhere in the economy. Their labour time and their talent are likely to be their principal inputs, and their production functions are likely to relate as much to the quality as to the quantity of their output. We return to the economic circumstances of individual artists below.

Across many fields in the arts – including opera, theatre, dance, classical music, jazz, independent film-making, small-scale literary publishing, contemporary visual art and craft, and so on – the predominant firm types, in terms of numbers of firms, are small and medium-sized enterprises, constituted on either a for-profit or a not-for-profit basis. Microeconomic theory offers straightforward means for

characterizing the production and cost conditions under which all these firms operate, with differences according to specific features of the various industries. For example, in the performing arts the unit of output in both production and cost function estimations is generally taken as paid attendances, in a manner similar to the way output is measured in other service-providing firms such as hospitals and universities. Standard functional forms can be used to investigate elasticities of output with respect to various inputs, economies of scale and scope, technical and allocative efficiencies, and productivity growth.

While production and cost conditions may be expected to be similar for these firms whether they are profit-oriented or otherwise, the structure and behaviour of for-profit and not-for-profit firms will differ markedly. Much attention in the economics of the arts has been focused on the latter because of the prevalence of not-for-profit firms at the ‘serious’ end of the artistic spectrum, producing innovative output or work which, though judged artistically worthy, does not appeal to a mass audience. Not only is there insufficient demand to sustain commercial production of this sort of work, but also the motives of the firms producing it are artistic rather than pecuniary. They can therefore be modelled as constrained maximizers of output quality (and possibly of the quantity of output as well if they wish to spread their art to as wide an audience as possible); the constraint is a break-even restriction whereby earned plus unearned revenue must at least cover costs over some specified period. Other model specifications have also been investigated, for example incorporating an objective of maximizing revenues from sponsorship and donations.

An issue of continuing interest in the economics of the performing arts is that of productivity lag, first identified by Baumol and Bowen (1966) and subsequently labelled ‘Baumol's disease’ or ‘the cost disease’. Essentially the hypothesis states that labour productivity in the live arts remains static over time – it still takes the same number of workers the same amount of time to perform *Hamlet* today as it did in Shakespeare's day. In a two-sector model in which one sector suffers from this technological disadvantage, wage rises in the productive sector are transmitted to the stagnant sector, causing a widening gap in the latter between revenues and costs, since firms in the stagnant sector cannot cover wage rises with improved labour productivity. Applying this to the live arts, Baumol and Bowen predicted that performing firms would have to access increasing levels of non-box-office revenue over time in order to stay in business. Empirical studies of this phenomenon have confirmed that costs of live performances have indeed risen as the model implies, but that the impact of these cost increases on firms has been somewhat muted; most performing companies have been able to mitigate the effects of slow productivity growth through a variety of strategies, including tapping new sources of unearned revenue, exploiting the potential of new recording and distribution technologies, expanded ancillary activities such as merchandising, and so on.

Finally in this section we turn to large-scale production in the arts. There are certainly some not-for-profit firms in the arts with multi-million dollar budgets, including major art museums, the world's principal opera companies and symphony orchestras, national theatre companies in several countries, and so on. In almost all cases some level of public funding is involved, together with significant levels of private-sector support from foundations, corporations and individual donors to supplement box-office revenue. In some countries these large-scale enterprises are government business undertakings, subject to varying degrees of independence or control in their governance and their operational decision-making. However, the majority of large-scale producers of artistic goods are profit-seeking firms operating in commercial markets where complex production processes are required and/or where

substantial scale economies exist. These firms include theatre companies staging popular shows, commercial and independent film producers, music publishers, record companies, major book publishers, art auction houses and so on. Taken together, these firms form a significant component (measured in terms of value of output) of the so-called creative or copyright industries, terms reflecting two of the necessary characteristics of cultural goods discussed earlier. From an economic point of view, these industries are notable for their peculiar contractual arrangements that reflect, among other things, the inherent uncertainties that attend every stage of artistic production processes whereby ‘nobody knows’ what the quality or market potential of the final product will be (Caves, 2000).

Market structures

It is perhaps surprising that there is little in the industrial organization literature dealing with structure, conduct and performance in the arts. There are many interesting questions concerning competition, market efficiency and pricing behaviour in the arts that await the attention of economists. As may be evidenced from the preceding section, the range of market structures in the arts is quite wide, providing considerable scope for empirical investigation.

At one extreme can be found instances of almost atomistic competition, as in the so-called primary market for visual art. Here there are many small producers, mostly individual artists selling on their own or through small local galleries, art fairs, and so on. Although the product is not exactly homogeneous, buyers tend to be not very discriminating, and prices may well be competed down to little more than cost of production plus some modest return to labour. Moving further across the market structure spectrum, we can suggest that the live performing arts in medium-to-large towns and cities show some evidence of monopolistic competition: a relatively large number of small firms competing through product differentiation and other non-price strategies for customers drawn from a single pool. Higher levels of concentration appear in other areas of the arts, especially in local markets for live performance characterized by one or two dominant firms when close substitutes are not available; the markets for opera or orchestral music in a given city may be examples. In all of the above cases, market conditions affect the pricing and output decisions of participating firms. Given that non-pecuniary motives play an important role in influencing the behaviour of economic agents in the arts, the competitive outcomes in the markets discussed might be expected to diverge somewhat from those predicted under more conventional conditions.

Factor markets

The input into artistic production processes that provides the unique qualities of artistic goods and services is, of course, the creative labour of artists themselves. Labour markets in the arts have been widely studied in both theoretical and empirical terms in an effort to understand whether and in what ways they differ from conventional labour markets. A principal finding relates again to the non-pecuniary motives for artistic production. Artists in general do not regard work as a chore whose only purpose is to earn an income. Rather, their commitment to making art means that they have a positive preference for working at their chosen profession, and empirical evidence indicates that they often forgo lucrative alternative employment in order to spend more time pursuing their creative work. This can be

modelled as a time allocation problem where the worker has to choose between preferred but less remunerative work in the arts on the one hand and better-paid but less desired non-arts work on the other. The choice is subject to a minimum-income constraint, necessary to prevent starvation, a condition often romantically associated with artists but rarely observed in practice. Such a ‘work preference’ model of labour supply yields predictions of behaviour at variance with the usual textbook construct – for example, a wage rise in the non-arts occupation may induce *less* work in that occupation because it enables more time to be devoted to the arts, a phenomenon akin to the backward-bending supply curve of labour in the conventional model.

The generally low levels of average earnings available from artistic practice mean that arts labour markets are characterized by ubiquitous multiple job-holding and much fluidity in career paths. The distribution of earnings across any population of arts workers is almost always skewed towards the lower end. Some attention has been paid to the role of risk in affecting entry and exit decisions in arts labour markets. Given the superstar phenomenon noted above, where extremely high incomes are earned by very few, some writers have portrayed these labour markets as winner-take-all lotteries to which artists submit themselves willingly. An alternative explanation of persistent labour market participation when expected monetary returns are low lies in the supposition that artists earn a sufficient level of psychic income to offset the meagre levels of their pecuniary rewards.

Turning to capital markets, we note simply that a similar psychic component may be present in rewarding suppliers of capital to the arts. For example, investors willing to back a theatre company putting on a new show may perhaps do so in expectation that the show will be a hit and they will earn a handsome return on their investment; however, a more plausible explanation for such a risky decision may be that these donors are motivated by a love of the theatre and hence that their satisfaction will derive largely if not entirely from the psychic rewards from helping to make it happen. Indeed, much private capital flows to the arts not as investments or loans but as untied donations with no strings attached, as discussed further below.

Policy issues

Government provision of financial assistance to the arts is widespread across the developed world, though the extent of intervention varies substantially between countries and between jurisdictions within countries. It is not clear whether such assistance is in accord with the wishes of voters or whether it is a case of imposed preferences whereby the arts are seen by governments as a merit good. It is also entirely possible that public subsidies to the arts are consistent with the restoration of Pareto optimality in an economy subject to market failure, if it is indeed the case that the arts give rise to public goods or positive externalities. Some economists remain sceptical of the latter proposition on empirical rather than theoretical grounds, and there is as yet not a great deal of evidence to resolve the issue one way or the other. In these circumstances more attention has been focused on the appropriate means for intervention once a normative rationale is accepted. The instruments governments have at their disposal include public-sector provision of artistic services (for example, through public art galleries); direct subsidies to cultural production or consumption; indirect support through the tax system; regulation; provision of information; assistance through the education system; and so on. An issue of considerable interest is the specification of optimal decision rules for allocation of public financing among competing avenues of artistic activity, a process apparently driven as much by rent-seeking or political expediency

as by the pursuit of economic efficiency.

The use of the tax system as a means of providing assistance has been of particular significance to the arts, especially via the tax deductibility allowed to philanthropic donors who give money to not-for-profit performing companies, museums, galleries, and so on. Such giving is likely to be motivated by a desire to secure the sorts of public-good benefits of the arts mentioned earlier, in circumstances where direct government support is regarded as inadequate. In some countries, most notably the United States, the cost of indirect support for the arts, measured in terms of tax revenue forgone, greatly exceeds the amount of direct financing by the public sector. Given that governments can manipulate the incentives facing donors by changing marginal tax rates, by raising or lowering thresholds and ceilings on allowable donations, and so on, much interest has focused on elasticities of giving with respect to variables such as the tax price. The critical issue from a policy viewpoint is whether the price elasticity is greater or less than unity in absolute terms, since a price elastic response would imply that lowering the tax price would increase recipients' revenue by more than the tax receipts forgone. However, despite many empirical studies, no clear consensus as to the size of these elasticities has emerged. Other policy issues of concern in this field include whether increased government support for the arts crowds out or crowds in private donations, and whether it is good or bad policy to use an instrument that allows private individuals to direct the allocation of public resources via their charitable-giving decisions.

One way in which public policy can assist the functioning of markets in the arts is via the creation and enforcement of property rights in artistic goods and services. Efficient copyright regimes aim to facilitate public access to information, at the same time as allowing creators to regulate the use of their work and to capture remuneration that would otherwise be lost to piracy, free-riding, unauthorized commercial exploitation, and the like. While often seen as a purely legal matter, copyright has a number of economic implications for the arts. In particular, artistic output in the form of literary works, paintings, photographs, musical compositions, and so forth can generally be reproduced at low or negligible cost, and in the absence of copyright protection their price would be driven down to marginal cost, so reducing or eliminating the incentive to the artist to create further output. Nevertheless, some exceptions to universal copyright coverage exist, for example in the 'fair use' provisions of copyright law, which allow free access for certain scholarly or public-interest purposes, or where high transactions costs of enforcement outweigh the potential gains to the rights holder. Other intellectual property issues of interest to economists include the market effects of moral rights (the rights that artists have over attribution and integrity of their works) and, in the visual arts, the phenomenon of *droit de suite* (the payment of a royalty to the artist or his or her heirs each time a given work is resold).

An area of growing importance in policy terms in recent years has been the role of the arts in urban and regional development. This role may be evident in a specific sense, for example in the impact of an arts festival on the local economic base, or in the use of community arts projects to engage and motivate disaffected youth in areas of high unemployment. In a wider context, the creative industries may be seen as a source of new enterprise, income growth and employment creation in depressed industrial regions. Empirical studies have looked at the impact of arts events, facilities, and so forth on a local or regional economy, and at the more general contribution that the arts industries make to economic activity, as a basis for policy formulation in a field increasingly engaging the attention of governments at both national and local levels.

Public policy towards the arts, heritage, the creative industries, cultural trade, and so forth can be gathered together under the somewhat fuzzy heading of 'cultural policy'. Given the significant economic

content of all of these areas, it can be expected that economic theory and analysis will continue to make an important contribution to policy-making in this field in the future.

Further reading

Recent surveys of the economics of the arts include Throsby (1994), Blaug (2001) and Ginsburgh (2001). Major contributions to the literature on the economics of the arts from the mid-1960s to the mid-1990s are collected together in Towse (1997). A broader view of cultural economics is contained in Throsby (2001). An accessible account of the principal topics in contemporary cultural economics is provided in Towse (2003), while a comprehensive research-oriented coverage of the economics of art and culture is contained in Ginsburgh and Throsby (2006).

See Also

- books, economics of
- intellectual property
- music markets, economics of
- superstars, economics of

Bibliography

Baumol, W. and Bowen, W.G. 1966. *Performing Arts: The Economic Dilemma*. New York: Twentieth Century Fund.

Blaug, M. 2001. Where are we now in cultural economics? *Journal of Economic Surveys* 15, 123–43.

Caves, R. 2000. *Creative Industries: Contracts Between Art and Commerce*. Cambridge, MA: Harvard University Press.

Ginsburgh, V. 2001. Economics of arts and culture. In *International Encyclopaedia of the Social and Behavioural Sciences*, ed. N. Smelser and P. Baltes. Amsterdam: Elsevier.

Ginsburgh, V. and Throsby, D., eds. 2006. *Handbook of the Economics of Art and Culture*. Amsterdam: Elsevier/North Holland.

Rosen, S. 1981. The economics of superstars. *American Economic Review* 71, 845–58.

Throsby, D. 1994. The production and consumption of the arts: a view of cultural economics. *Journal of Economic Literature* 32, 1–29.

Throsby, D. 2001. *Economics and Culture*. Cambridge: Cambridge University Press.

Towse, R., ed. 1997. *Cultural Economics: The Arts, the Heritage and the Media Industries*, 2 vols. Cheltenham: Edward Elgar.

Towse, R., ed. 2003. *A Handbook of Cultural Economics*. Cheltenham: Edward Elgar.

How to cite this article

Throsby, David. "art, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000224> doi:10.1057/9780230226203.0062

The New Palgrave Dictionary of Economics Online

artificial neural networks

Chung-Ming Kuan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Artificial neural networks (ANNs) constitute a class of flexible nonlinear models designed to mimic biological neural systems. In this article we introduce ANN using familiar econometric terminology and provide an overview of the ANN modelling approach and its implementation methods.

Keywords

ARMA models; artificial neural networks; back propagation algorithm; Bayesian information criterion; distributed lags; feedforward neural networks; learning; model misspecification; Newton algorithm; nonlinear least squares (NLS) method; nonlinear models; nonparametric functional analysis; nonparametric methods; predictive stochastic complexity; recurrent neural networks; stochastic approximation method

Article

1 Introduction

Artificial neural networks (ANNs) constitute a class of flexible nonlinear models designed to mimic biological neural systems. Typically, a biological neural system consists of several layers, each with a large number of neural units (neurons) that can process the information in a parallel manner. The models with these features are known as ANN models. Such models can be traced back to the simple input–output model of McCulloch and Pitts (1943) and the ‘perceptron’ of Rosenblatt (1958). The early yet simple ANN models, however, did not receive much attention because of their limited applicability and also because of the limitation of computing capacity at that time. In seminal works, Rumelhart, McClelland and PDP Research Group (1986) and McClelland, Rumelhart and PDP Research Group (1986) presented the new developments of ANN, including more complex and flexible ANN structures and a new network learning method. Since then, ANN has become a rapidly growing research area.

As far as model specification is concerned, ANN has a multi-layer structure such that the middle layer is built upon many simple nonlinear functions that play the role of neurons in a biological system. By allowing the number of these simple functions to increase indefinitely, a multi-layered ANN is capable of approximating a large class of functions to any desired degree of accuracy, as shown in, for example, Cybenko (1989), Funahashi (1989), Hornik, Stinchcombe and White (1989; 1990), and Hornik (1991; 1993). From an econometric perspective, ANN can be applied to approximate the unknown conditional mean (median, quantile) function of the variable of interest without suffering from the problem of model misspecification, unlike parametric models commonly used in empirical studies. Although nonparametric methods, such as series and polynomial approximators, also possess this property, they usually require a larger number of components to achieve similar approximation accuracy (Barron, 1993). ANNs are thus a parsimonious approach to nonparametric functional analysis.

ANNs have been widely applied to solve many difficult problems in different areas, including pattern recognition, signal processing, and language learning. Since White (1988), there have also been numerous applications of ANN in economics and finance. Unfortunately, the ANN literature is not easy to penetrate, so it is hard for applied economists to understand why ANN works and how it can be implemented properly. Fortunately, while the ANN jargon originated from cognitive science and computer science, they often have econometric

interpretations. For example, a 'target' is, in fact, a dependent variable of interest, an 'input' is an explanatory variable, and network 'learning' amounts to the estimation of unknown parameters in a network. The purpose of this article is thus twofold. First, it introduces ANN using familiar econometric terminology and hence serves to bridge the gap between the fields of ANN and economics. Second, it provides an overview of ANN modelling approach and its implementation methods. For an early review of ANN from an econometric perspective, we refer to Kuan and White (1994).

This article proceeds as follows. We introduce various ANN model specifications and the choices of network functions in Section 2. We present the 'universal approximation' property of ANN in Section 3. Model estimation and model complexity regularization are discussed in Section 4. Section 5 concludes.

2 ANN model specifications

Let *Y* denote the collection of *n* variables of interest with the *t*-th observation $\mathbf{y}_t(n \times 1)$ and *X* the collection of *m* explanatory variables with the *t*-th observation $\mathbf{x}_t(m \times 1)$. In the ANN literature, the variables in *Y* are known as *targets* or *target variables*, and the variables in *X* are *inputs* or *input variables*. There are various ways to build an ANN model that can be used to characterize the behavior of $\mathbf{y}_t$ using the information contained in the input variables $\mathbf{x}_t$. In this section, we introduce some network architectures and the functions that are commonly used to build an ANN.

2.1 Feedforward neural networks

We first consider a network with an *input* layer, an *output* layer, and a *hidden* layer in between. The input (output) layer contains *m* input units (*n* output units) such that each unit corresponds to a particular input (output) variable. In the hidden layer, there are *q* hidden units connected to all input and output units; the strengths of such connections are labelled by (unknown) parameters known as the network *connection weights*. In particular, $\boldsymbol{\gamma}_h = (\gamma_{h,1}, \ldots, \gamma_{h,m})'$ denotes the vector of the connection weights between the *h*-th hidden unit and all *m* input units, and $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,q})'$ denotes the vector of the connection weights between the *j*-th output unit and all *q* hidden units. An ANN in which the sample information (signals) are passed forward from the input layer to the output layer without feedback is known as a *feedforward neural network*. Figure 1 illustrates the architecture of a three-layer feedforward network with three input units, four hidden units and two output units.

Figure 1

A feedforward network with three input units, four hidden units and two output units

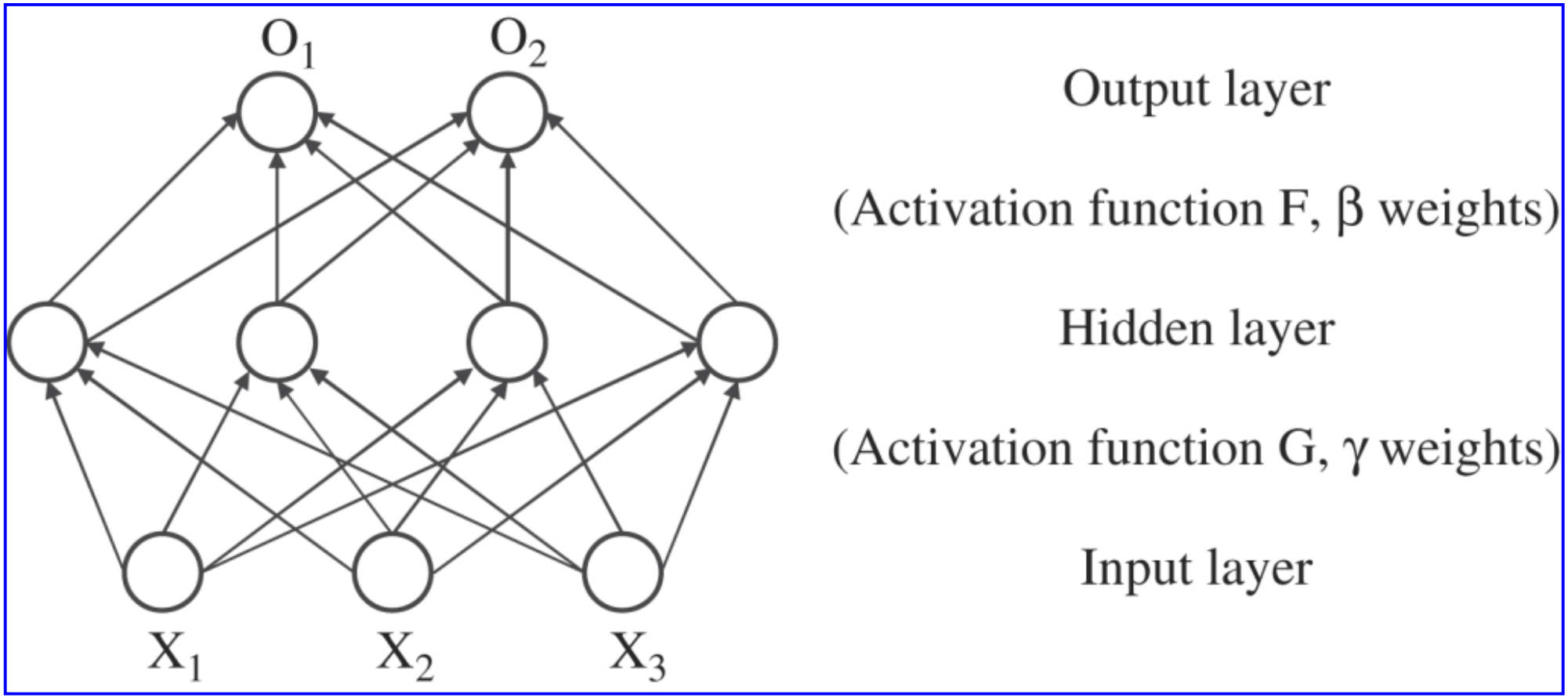

This multi-layered structure of a feedforward network is designed to function as a biological neural system. The input units are the neurons that receive the information (stimuli) from the outside environment and pass them to the neurons in a middle layer (that is, hidden units). These neurons then transform the input signals to generate neural signals and forward them to the neurons in the output layer. The output neurons in turn generate signals that determine the action to be taken. Note that all information from the units in one layer is processed simultaneously, rather than sequentially, by the units in an 'upper' layer. (This concept, also known as *parallel processing* or *massive parallelism*, differs from the traditional concept of sequential processing and has led to a major advance in designing computer architecture.)

Formally, the input units receive the information $\mathbf{x}_t$ and send to all hidden units, weighted by the connection weights between the input and hidden units. This information is then transformed by the *activation function G* in each hidden unit. That is, the *h*-th hidden unit receives $\mathbf{x}_t'\gamma_h$ and transforms it to $G(\mathbf{x}_t'\gamma_h)$. The information generated by all hidden units is further passed to the output units, again weighted by the connection weights, and transformed by the activation function *F* in each output unit. Hence, the *j*-th output unit receives $\sum_{h=1}^{q}\beta_{j,h}G(\mathbf{x}_t'\gamma_h)$ and transforms it into the network output:

$$o_{t,j} = F\left(\sum_{h=1}^{q}\beta_{j,h}G(\mathbf{x}_t'\gamma_h)\right),\ j = 1, \ldots, n. \quad (1)$$

The output O_j is used to describe or predict the behaviour of the *j*-th target Y_j.

In practice, it is typical to include a constant term, also known as the *bias* term, in each activation function in (1). That is,

$$o_{t,j} = F\left(\beta_{j,0} + \sum_{h=1}^{q} \beta_{j,h} G(\gamma_{h,0} + \mathbf{x}_t' \boldsymbol{\gamma}_h)\right), \; j = 1, \ldots, n, \tag{2}$$

where $\gamma_{h,0}$ is the bias term in the *h*-th hidden unit and $\beta_{j,0}$ is the bias term in the *j*-th output unit. A constant term in each activation function adds flexibility to hidden-unit and output-unit responses (activations), in a way similar to the constant term in (non)linear regression models. Note that when there is no transformation in the output units, *F* is an identity function (that is, *F*(*a*)=*a*) so that

$$o_{t,j} = \beta_{j,0} + \sum_{h=1}^{q} \beta_{j,h} G(\gamma_{h,0} + \mathbf{x}_t' \boldsymbol{\gamma}_h), \; j = 1, \ldots, n. \tag{3}$$

It is also straightforward to construct networks with two or more hidden layers. For simplicity, we will focus on the three-layer networks with only one hidden layer.
While parametric econometric models are typically formulated using a given function of the input $\mathbf{x}_t$, the network (2) is a class of flexible nonlinear functions of $\mathbf{x}_t$. The exact form of a network model depends on the activation functions (*F* and *G*) and the number of hidden units (*q*). In particular, the network function in (3) is an affine transformation of *G* and hence may be interpreted as an expansion with the 'basis' function *G*.
The networks (2) and (3) can be further extended. For example, one may construct a network in which the input units are connected not only to the hidden units but also directly to the output units. This leads to networks with *short-cut* connections. Corresponding to (2), the outputs of a feedforward network with short cuts are

$$o_{t,j} = F\left(\beta_{j,0} + \mathbf{x}_t' \boldsymbol{\alpha}_j + \sum_{h=1}^{q} \beta_{j,h} G(\gamma_{h,0} + \mathbf{x}_t' \boldsymbol{\gamma}_h)\right), \; j = 1, \ldots, n,$$

where α_j is the vector of connection weights between the output and input units, and, corresponding to (3), the outputs are

$$o_{t,j} = \beta_{j,0} + \mathbf{x}_t' \boldsymbol{\alpha}_j + \sum_{h=1}^{q} \beta_{j,h} G(\gamma_{h,0} + \mathbf{x}_t' \boldsymbol{\gamma}_h), \; j = 1, \ldots, n.$$

Figure 2 illustrates the architecture of a feedforward network with two input units, three hidden units, one output unit and short-cut connections. Thus, parametric econometric models may be interpreted as feedforward networks with short-cut connections but no hidden-layer connections. The linear combination of hidden-unit activations, $\sum_{h=1}^{q} \beta_{j,h} G(\gamma_{h,0} + \mathbf{x}_t' \boldsymbol{\gamma}_h)$, in effect characterizes the nonlinearity not captured by the linear function of $\mathbf{x}_t$.

Figure 2
A feedforward neural network with short cuts

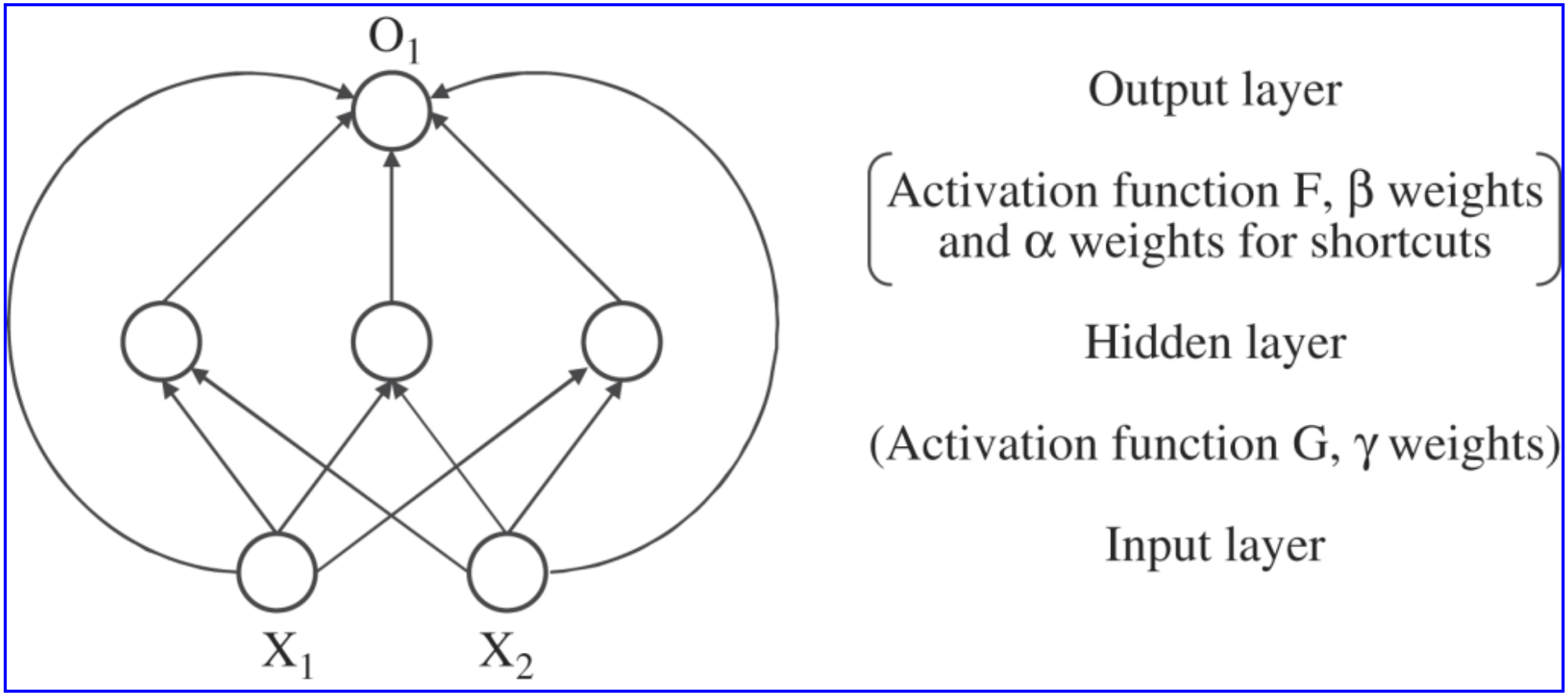

2.2 Recurrent neural networks

From the preceding section we can see that there is no ‘memory’ device in feedforward networks that can store the signals generated earlier. Hence, feedforward networks treat all sample information as ‘new’; the signals in the past do not help to identify data features, even when sample information exhibits temporal dependence. As such, a feedforward network must be expanded to a large extent so as to represent complex dynamic patterns. This causes practical difficulty because a large network may not be easily implemented. To utilize the information from the past, it is natural to include lagged target information $\mathbf{y}_{t-k}$, $k = 1, \ldots, s$, as input variables, similar to linear AR and ARX models in econometric studies. Yet such networks do not have any built-in structure that can ‘memorize’ previous neural responses (transformed sample information). The so-called *recurrent neural networks* overcome this difficulty by allowing internal feedbacks and hence are especially appropriate for dynamic problems.
Jordan (1986) first introduced a recurrent network with feedbacks from output units. That is, the output units are connected to input units but with *time delay*, so that the network outputs at time $t - 1$ are also the input information at time t. Specifically, the outputs of a Jordan network are

$$o_{t,j} = F\left(\beta_{j,0} + \sum_{h=1}^{q} \beta_{j,h} G(\gamma_{h,0} + \mathbf{x}_t' \boldsymbol{\gamma}_h + \mathbf{o}_{t-1}' \boldsymbol{\delta}_h)\right), \; j = 1, \ldots, n, \tag{4}$$

where δ_h is the vector of the connection weights between the *h*-th hidden unit and the input units that receive lagged outputs $\mathbf{o}_{t-1} = (o_{t-1,1}, \ldots, o_{t-1,n})'$. The network (4) can be further extended to allow for more lagged outputs $\mathbf{o}_{t-2}$, $\mathbf{o}_{t-3}$, ….

Similarly, Elman (1990) considered a recurrent network in which the hidden units are connected to input units with time delay. The outputs of an Elman network are:

$$o_{t,j} = F\left(\beta_{j,0} + \sum_{h=1}^{q} \beta_{j,h} a_{t,h}\right),\ j = 1, \ldots, n, a_{t,h} = G(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h + \mathbf{a}_{t-1}'\boldsymbol{\delta}_h),\ h = 1, \ldots, q, \tag{5}$$

where $\mathbf{a}_{t-1} = (a_{t-1,1}, \ldots, a_{t-1,q})'$ is the vector of lagged hidden-unit activations, and δ_h here is the vector of the connection weights between the *h*-th hidden unit and the input units that receive lagged hidden-unit activations $\mathbf{a}_{t-1}$. The network (5) can also be extended to allow for more lagged hidden-unit activations $\mathbf{a}_{t-2}$, $\mathbf{a}_{t-3}$, ⋯. Figure 3 illustrates the architectures of a Jordan network and an Elman network.

Figure 3

Recurrent neural networks: Jordan (left) and Elman (right)

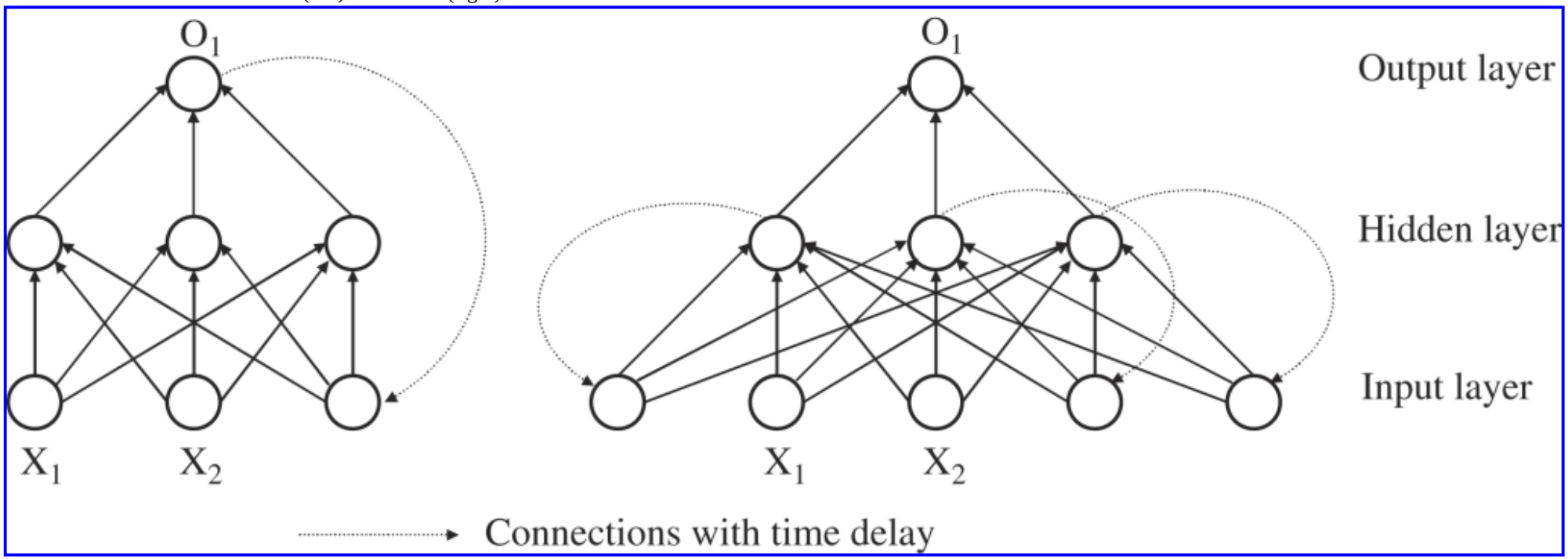

From (4) and (5) we can see that, by recursive substitution, the outputs of these recurrent networks can be expressed in terms of current and all past inputs. Such expressions are analogous to the distributed lag model or the AR representation of an ARMA model (when the inputs are lagged targets). Thus, recurrent networks incorporate the information in the past input variables without including all of them in the model. By contrast, a feedforward network requires a large number of inputs to carry such information. Note that the Jordan network and the Elman network summarize past input information in different ways and hence have their own merits. When the previous ‘location’ of a network is crucial in determining the next move, as in the design of a robot, a Jordan network seems more appropriate. When the past internal neural responses are more important, as in language learning

problems, an Elman network may be preferred.

2.3 Choices of activation function

As far as model specifications are concerned, the building blocks of an ANN model are the activation functions *F* and *G*. Different choices of the activation functions result in different network models. We now introduce some activation functions commonly employed in empirical studies.
Recall that the hidden units play the role of neurons in a biological system. Thus, the activation function in each hidden unit determines whether a neuron should be turned on or off. Such an on/off response can be easily represented using an indicator (threshold) function, also known as a *heaviside* function in the ANN literature, that is,

$$G(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h) = \begin{cases} 1, & \text{if } \gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h \geq c, \\ 0, & \text{if } \gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h < c, \end{cases}$$

where *c* is a pre-determined threshold value. That is, depending on the strength of connection weights and input signals, the activation function *G* will determine whether a particular neuron is on $(G(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h) = 1)$ or off $(G(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h) = 0)$.
In a complex neural system, neurons need not have only an on/off response but may be in an intermediate position. This amounts to allowing the activation function to assume any value between zero and 1. In the ANN literature, it is common to choose a *sigmoid* (*S*-shaped) and *squashing* (bounded) function. In particular, if the input signals are ‘squashed’ between zero and 1, the activation function is understood as a smooth counterpart of the indicator function. A leading example is the logistic function:

$$G(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h) = \frac{1}{1 + \exp(-[\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h])},$$

which approaches 1 (zero) when its argument goes to infinity (negative infinity). Hence, the logistic activation function generates a partially on/off signal based on the received input signals.
Alternatively, the hyperbolic tangent (tanh) function, which is also a sigmoid and squashing function, can serve as an activation function:

$$G(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h) = \frac{\exp(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h) - \exp(-[\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h])}{\exp(\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h) + \exp(-[\gamma_{h,0} + \mathbf{x}_t'\boldsymbol{\gamma}_h])}.$$

Compared with the logistic function, this function may assume negative values and is bounded between –1 and 1. It approaches 1 (–1) when its argument goes to infinity (minus infinity). This function is more flexible because the negative values, in effect, represent ‘suppressing’ signals from the hidden unit. See Figure 4 for an illustration of the logistic and tanh functions. Note that for the logistic function *G*, a re-scaled function $\breve{G}$ such that $\breve{G}(a) = 2G(a) - 1$ also generates values between –1 and 1 and may be used in place of the tanh function. (A choice of the activation function in classification problems is the so-called *radial basis function*. We do not discuss this choice because its argument is not an affine transformation of inputs and hence does not fit in our framework here. Moreover, the networks with this activation function provide only *local* approximation to unknown functions, in contrast with the approximation property discussed in Section 3.)

Figure 4
Activation functions: logistic (left) and tanh (right)

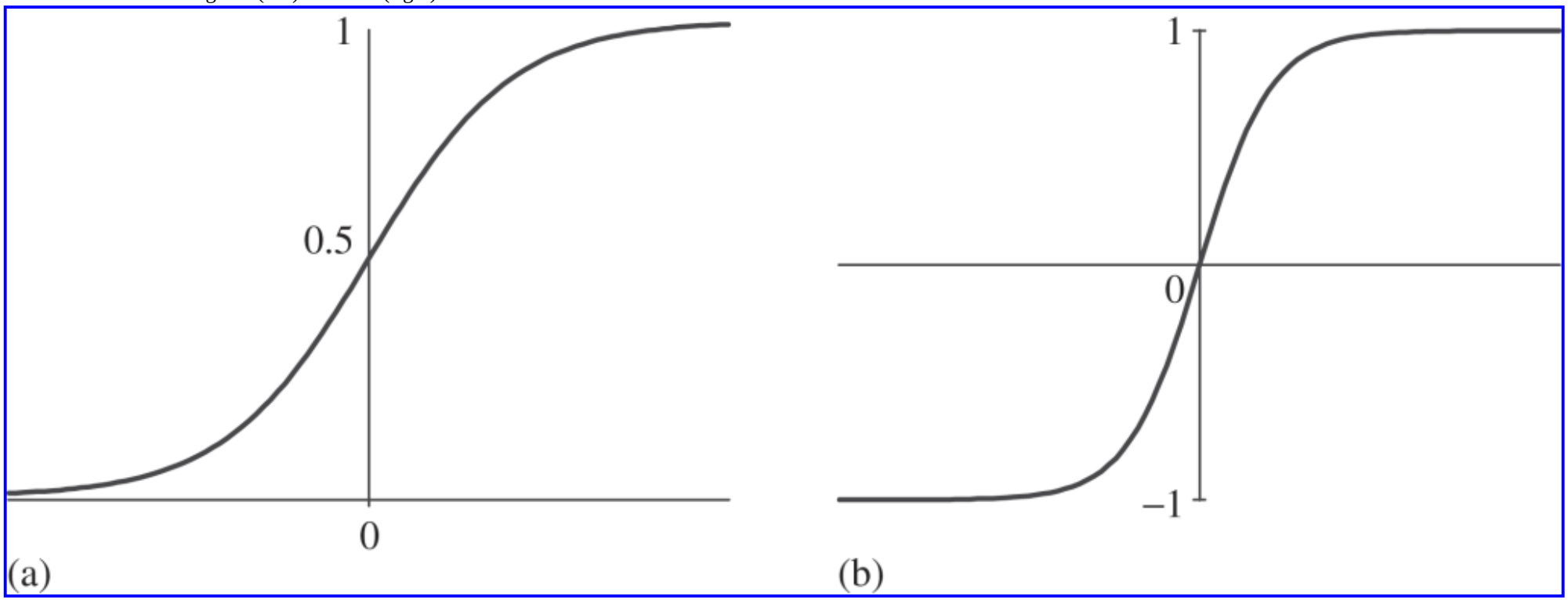

The aforementioned activation functions are chosen for convenience because they are differentiable everywhere and their derivatives are easy to compute. In particular, when G is the logistic function,

$$\frac{dG(a)}{da} = G(a)[1 - G(a)];$$

when G is the tanh function,

$$\frac{dG(a)}{da} = \left[\frac{2}{\exp(a) + \exp(-a)}\right]^2 = \operatorname{sech}^2(a).$$

These properties facilitate parameter estimation, as will be seen in Section 4.1. Nevertheless, these functions are not necessary for building proper ANNs. For example, smooth cumulative distribution functions, which are sigmoidal and squashing, are also legitimate candidates for activation function. In Section 3, it is shown that, as far as network approximation property is concerned, the activation function in hidden units does not even have to be sigmoidal, yet boundedness is usually required. Thus, sine and cosine functions can also serve as an activation function.

As for the activation function F in the output units, it is common to set it as the identity function so that the outputs of (3) enjoy the freedom of assuming any real value. This choice

suffices for the network approximation property discussed in Section 3. When the target is a binary variable taking the values zero and one, as in a classification problem, *F* may be chosen as the logistic function so that the outputs of (2) must fall between zero and 1, analogous to a logit model in econometrics.

3 ANN as an universal approximator

What makes ANN a useful econometric tool is its *universal approximation* property, which basically means that a multi-layered ANN with a large number of hidden units can well approximate a large class of functions. This approximation property is analogous to that of nonparametric approximators, such as polynomials and Fourier series, yet it is not shared by parametric econometric models.

To present the approximation property, we consider the network function element by element. Let $f_{G,q}: \mathbb{R}^m \times \Theta_{m,q} \to \mathbb{R}$ denote the network function with *q* hidden units, the output activation function *F* being the identity function, and the hidden-unit activation function *G*, that is,

$$f_{G,q}(\mathbf{x}; \boldsymbol{\theta}) = \beta_0 + \sum_{h=1}^{q} \beta_h G(\gamma_{h,0} + \mathbf{x}'\boldsymbol{\gamma}_h),$$

as in (3), where $\Theta_{m,q}$ is the parameter space whose dimension depends on *m* and *q*, and $\theta \in \Theta_{m,q}$ (note that the subscripts *m* and *q* for θ are suppressed). Given the activation function *G*, the collection of all $f_{G,q}$ functions with different *q* is:

$$\mathcal{F}_G = \bigcup_{q=1}^{\infty} \left\{ f_{G,q}: f_{G,q}(\mathbf{x}; \boldsymbol{\theta}) = \beta_0 + \sum_{h=1}^{q} \beta_h G(\gamma_{h,0} + \mathbf{x}'\boldsymbol{\gamma}_h) \right\};$$

when the union is taken up to a finite number *N*, the resulting collection is denoted as $\mathcal{F}_G^N$. Intuitively, $\mathcal{F}_G$ is capable of functional approximation because $f_{G,q}$ can be viewed as an expansion with the 'basis' function *G* and hence is similar to a nonparametric approximator.

More formally, we follow Hornik (1991) and consider two measures of the closeness between functions. First define the *uniform distance* between functions *f* and *g* on the set *K* as

$$d_K(f, g) = \sup_{\mathbf{x} \in K} |f(\mathbf{x}) - g(\mathbf{x})|.$$

Let *K* denote a compact subset in $\mathbb{R}^m$ and *C*(*K*) denote the space of all continuous functions on *K*. Then, when the activation function *G* is continuous, bounded and nonconstant, the collection $\mathcal{F}_G$ is dense in *C*(*K*) for all *K* in $\mathbb{R}^m$ in terms of d_K (Theorem 2 of Hornik, 1991). (Hornik, 1991, considered the network without the bias term in the output unit, that is, β_0=0. Yet as long as *G* is not a constant function, all the results in Hornik, 1991, carry over; see Stinchcombe and White, 1998, for details.) That is, for any function g in *C*(*K*) and any $\varepsilon > 0$, there is a network function $f_{G,q}$ in $\mathcal{F}_G$ such that $d_K(f_{G,q} - g) < \varepsilon$. As $\mathcal{F}_G^N$ is not dense in *C*(*K*) for any finite number *N*, this result shows that any continuous function can be approximated arbitrarily well on compacta by a three-layered feedforward network $f_{G,q}$, provided that *q*, the number of hidden units, is sufficiently large.

Taking **x** as random variables, defined in the probability space with the probability measure $\mathbb{P}$, we consider the L_r-*norm* of *f*(**x**)–*g*(**x**):

$$\| f - g \|_r = \left(\int_{\mathbb{R}^m} | f(\mathbf{x}) - g(\mathbf{x}) |^r d\mathbb{P}(\mathbf{x}) \right)^{1/r},$$

$1 \leq r < \infty$. For $r = 2$ ($r = 1$), this is the well-known measure of mean squared error (mean absolute error). Then, when the activation function G is bounded and nonconstant, the collection $\mathcal{F}_G$ is dense in the L_r space (Theorem 1 of Hornik, 1991). That is, any function g (with finite L_r-norm) can also be well approximated by a three-layered feedforward network $f_{G,q}$ in terms of L_r-norm when q is sufficiently large.

It should be emphasized that the universal approximation property of a feedforward network hinges on the three-layered architecture and the number of hidden units, but not on the activation function per se. As stated above, the activation function in the hidden unit can be a general bounded function and does not have to be sigmoidal. Hornik (1993) provides results that permit even more general activation functions. Moreover, a feedforward network with only one hidden layer suffices for such approximation property. More hidden layers may be helpful in certain applications but are not necessary for functional approximation.

Barron (1993) further derived the rate of approximation in terms of mean squared error $\| f - g \|_2^2$. It was shown that three-layered feedforward networks $f_{G,q}$ with G a sigmoidal function can achieve the approximation rate of order $O(1/q)$, for which the number of parameters grows linearly with q (with the order $O(mq)$). This is in sharp contrast with other expansions, such as polynomial (with p the degree of the polynomial) and spline (with p the number of knots per coordinate), which yield suitable approximation when the number of parameters grows exponentially (with the order $O(p^m)$). Thus, it is practically difficult for such expansions to approximate well when the dimension of the input space, m, is large.

4 Implementation of ANNs

In practice, when the activation functions in an ANN are chosen, it remains to estimate its connection weights (unknown parameters) and to determine a proper number of hidden units. Given that the connection weights of an ANN model are unknown, this network must be properly ‘trained’ so as to ‘learn’ the unknown weights. This is why parameter estimation is referred to as network *learning* and the sample used for parameter estimation is referred to a *training sample* in the ANN literature. As the number of hidden units q determines network complexity, finding a suitable q is known as network *complexity regularization*.

4.1 Model estimation

The network parameters can be estimated by either *online* or *offline* methods. An online learning algorithm is just a *recursive estimation* method which updates parameter estimates when new sample information becomes available. By contrast, offline learning methods are based on fixed training samples; standard econometric estimation methods are typically offline.

To ease the discussion of model estimation, we focus on the simple case that there is only one target variable y and the network function $f_{G,q}$. Generalization to the case with multiple target variables and vector-valued network functions is straightforward. Once the activation function G is chosen and the number of hidden units is given, $f_{G,q}$ is a nonlinear parametric model for the target y; the network with multiple outputs is a system of nonlinear models. If we take mean squared error as the criterion, the parameter vector of interest θ^* thus minimizes

$$\mathbb{E}[y - f_{G,q}(\mathbf{x}; \boldsymbol{\theta})]^2. \tag{6}$$

It is well known that

$$\mathbb{E}[y - f_{G,q}(\mathbf{x}; \boldsymbol{\theta})]^2 = \mathbb{E}[y - E(y|\mathbf{x})]^2 + \mathbb{E}[\mathbb{E}(y|\mathbf{x}) - f_{G,q}(\mathbf{x}; \boldsymbol{\theta})]^2.$$

As $\mathbb{E}(y|\mathbf{x})$ is the best L_2 predictor of y, θ^* must also minimize the mean squared approximation error: $\mathbb{E}[\mathbb{E}(y|\mathbf{x}) - f_{G,q}(\mathbf{x};\boldsymbol{\theta})]^2$. This shows that, among all three-layered feedforward networks with the activation function G and q hidden units, $f_{G,q}(\mathbf{x};\theta^*)$ provides the best approximation to the conditional mean function.

Given a training sample of T observations, an estimator of θ^* can be obtained by minimizing the sample counterpart of (6):

$$\frac{1}{T}\sum_{t=1}^{T}[y_t - f_{G,q}(\mathbf{x}_t;\boldsymbol{\theta})]^2,$$

which is just the objective function of the nonlinear least squares (NLS) method. The NLS method is an offline estimation method because the size of the training sample is fixed. Under very general conditions on the data and nonlinear function, it is well known that the NLS estimator is strongly consistent for θ^* and asymptotically normally distributed (see, for example, Gallant and White, 1988).

In many ANN applications (for example, signal processing and language learning), the training sample is not fixed but constantly expands with new data. In such cases, offline estimation may not be feasible, but online estimation methods, which update the parameter estimates based solely on the newly available data, are computationally more tractable. Moreover, online estimation methods can be interpreted as 'adaptive learning' by biological neural systems. It should be emphasized that when there is only a given sample, as in most empirical studies in economics, recursive estimation is *not* to be preferred because it is, in general, statistically less efficient than the NLS method in finite samples.

Note that the parameter of interest θ^* is the zero of the first order condition of (6):

$$\mathbb{E}[\nabla f_{G,q}(\mathbf{x};\boldsymbol{\theta})(y - f_{G,q}(\mathbf{x};\boldsymbol{\theta}))] = \mathbf{0},$$

where $\nabla f_{G,q}(\mathbf{x};\theta)$ is the (column) gradient vector of $f_{G,q}$ with respect to θ. To estimate θ^*, a recursive algorithm proposed by Rumelhart, Hinton and Williams (1986) is

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t + \eta_t \nabla f_{G,q}(\mathbf{x}_t;\hat{\boldsymbol{\theta}}_t)[y_t - f_{G,q}(\mathbf{x}_t;\hat{\boldsymbol{\theta}}_t)], \quad (7)$$

where $\eta_t > 0$ is a parameter that re-scales the adjustment term in the square bracket. It can be seen from (7) that the adjustment term is determined by the gradient descent direction and the error between the target and network output: $y_t - f_{G,q}(\mathbf{x}_t;\hat{\boldsymbol{\theta}}_t)$, and it requires only the information at time t, that is, y_t, $\mathbf{x}_t$, and the estimate $\hat{\boldsymbol{\theta}}_t$. (The algorithm (7) is analogous to the numerical steepest-descent algorithm. However, (7) utilizes only the information at time t, whereas numerical optimization algorithms are computed using all the information in a given sample and hence are offline methods.)

The algorithm (7) is known as the *error back-propagation* (or simply back-propagation) algorithm in the ANN literature, because the error signal $[y_t - f_{G,q}(\mathbf{x}_t;\hat{\boldsymbol{\theta}}_t)]$ is propagated back through the network to determine the change of each weight. The underlying idea of this algorithm can be traced back to the classical *stochastic approximation* method introduced in Robbins and Monro (1951). White (1989) established consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_t$ in (7). Note that the parameter η_t in the algorithm is known as a

learning rate. For consistency of $\widehat{\boldsymbol{\theta}}_t$, it is required that η_t satisfies $\sum_{t=1}^{\infty}\eta_t = \infty$ and $\sum_{t=1}^{\infty}\eta_t^2 < \infty$, for example, $\eta_t = 1/t$. The former condition ensures that the updating process may last indefinitely, whereas the latter implies $\eta_t \to 0$ so that the adjustment in the parameter estimates can be made arbitrarily small. (In many applications of ANN, the learning rate is often set to a constant η_o; the resulting estimate $\widehat{\boldsymbol{\theta}}_t$ loses consistency in this case. Kuan and Hornik (1991) established a convergence result based on small-η_o asymptotics.)
Instead of the gradient descent direction, it is natural to construct a recursive algorithm with a Newton search direction. Kuan and White (1994) proposed the following algorithm:

$$\widehat{\mathbf{H}}_{t+1} = \widehat{\mathbf{H}}_t + \eta_t[\nabla f_{G,q}(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t)\nabla f_{G,q}(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t)' - \widehat{\mathbf{H}}_t], \widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \eta_t\widehat{\mathbf{H}}_{t+1}^{-1}\nabla f_{G,q}(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t)[y_t - f_{G,q}(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t)], \tag{8}$$

where $\widehat{\mathbf{H}}_{t+1}$ characterizes a Newton direction and is recursively updated via the first equation. Kuan and White (1994) showed that $\widehat{\boldsymbol{\theta}}_t$ in (8) is $\sqrt{t}$-consistent, statistically more efficient than $\widehat{\boldsymbol{\theta}}_t$ in (7), and asymptotically equivalent to the NLS estimator. The algorithm (8) may be implemented in different ways; for example, there is an algorithm that is algebraically equivalent to (8) but does not involve matrix inversion. See Kuan and White (1994) for more discussions on the implementation of the Newton algorithms.
On the other hand, estimating recurrent networks is more cumbersome. From (4) and (5) we can see that recurrent network functions depend on θ directly and also indirectly through the presence of internal feedbacks (that is, lagged output and lagged hidden-unit activations). The indirect dependence on parameters must be taken into account in calculating the derivatives with respect to θ. Thus, NLS optimization algorithms that require analytic derivatives are difficult to implement. Kuan, Hornik and White (1994) proposed the *dynamic back-propagation* algorithm for recurrent networks, which is analogous to (7) but involves more updating equations. Kuan (1995) further proposed a Newton algorithm for recurrent networks, analogous to (8), and showed that it is $\sqrt{t}$-consistent and statistically more efficient than the dynamic back-propagation algorithm. We omit the details of these algorithms; see Kuan and Liu (1995) for an application of these estimation methods for both feedforward and recurrent networks.
Note that the NLS method and recursive algorithms all require computing the derivatives of the network function. Thus, a smooth and differentiable activation function, as the examples given in Section 2.3, are quite convenient for network parameter estimation. Finally, given that ANN models are highly nonlinear, it is likely that there exist multiple optima in the objective function. There is, however, no guarantee that the NLS method and the recursive estimation methods discussed above will deliver the global optimum. This is a serious problem because the dimension of the parameter space is typically large. Unfortunately, a convenient and effective method for finding the global optimum in ANN estimation is not yet available.

4.2 Model complexity regularization

Section 3 shows that a network model $f_{G,q}$ can approximate unknown function when the number of hidden units, q, is sufficiently large. When there is a fixed training sample, a complex network with a very large q may over fit the data. Thus, there is a trade-off between approximation capability and over-fitting in implementing ANN models.
An easy approach to regularizing the network complexity is to apply a model selection criterion, such as Schwarz (Bayesian) information criterion (BIC), to the network models with various q. (Alternatively, one may consider testing whether some hidden units may be dropped from the model. This amounts to testing, say, $\beta_h = 0$ for some h. Unfortunately, the parameters in that hidden-unit activation function ($\gamma_{h,0}$ and γ_h) are not identified under this null hypothesis. It is well known that, when there are unidentified nuisance parameters, standard econometric tests are not applicable.) As is well known, BIC consists of two terms: one is based on model fitness, and the other penalizes model complexity. Hence, it is suitable for regularizing network complexity; see also Barron (1991). A different criterion introduced in Rissanen (1986; 1987) is *predictive stochastic complexity* (PSC) which is just an average of squared prediction errors:

$$\mathrm{PSC} = \frac{1}{T-k}\sum_{t=k+1}^{T}[y_t - f_{G,q}(\mathbf{x}_t,\widehat{\boldsymbol{\theta}}_t)]^2,$$

where $\hat{\theta}_t$ is the predicted parameter estimate based on the sample information up to time $t-1$, and k is the total number of parameters in the network. Given the number of inputs, the network with the smallest BIC or PSC gives the desired number of hidden units q^*. Rissanen showed that both BIC and PSC can be interpreted as the criteria for 'minimum description length' in the sense that they determine the shortest code length (asymptotically) that is needed to encode a sequence of numbers. In other words, these criteria lead to the least complex model that still captures the key information in data. Swanson and White (1997) showed that a network selected by BIC need not perform well in out-of-sample forecasting, however.

Clearly, PSC requires estimating the parameters at each t. It would be computationally demanding if the NLS method is to be used, even for a moderate sample. For simplicity, Kuan and Liu (1995) suggested a two-step procedure for implementing ANN models. In the first step, one estimates the network models and computes the resulting PSCs using the recursive Newton algorithm, which is asymptotically equivalent to the NLS method. When a suitable network structure is determined, the Newton parameter estimates can be used as initial values for NLS estimation in the second step. This approach thus maintains a balance between computational cost and estimator efficiency.

5 Concluding remarks

In this article, we introduce ANN model specifications, their approximation properties, and the methods for model implementation from an econometric perspective. It should be emphasized that ANN is neither a magical econometric tool nor a 'black box' that can solve *any* difficult problems in econometrics. As discussed above, a major advantage of ANN is its universal approximation property, a property shared by other nonparametric approximators. Yet, compared with parametric econometric models, a simple ANN need not perform better, and a more complex ANN (with a large number of hidden units) is more difficult to implement properly and cannot be applied when there is only a small data-set. Therefore, empirical applications of ANN models must be exercised with care.

See Also

- nonparametric structural models
- stochastic adaptive dynamics

I would like to express my sincere gratitude to Steven Durlauf for his patience and constructive comments on early drafts of this article. I also thank Shih-Hsun Hsu and Yu-Lieh Huang for very helpful suggestions. The remaining errors are all mine.

Bibliography

Barron, A. 1991. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas. Dordrecht: Kluwer Academic Publishers.

Barron, A. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* 39, 930–45.

Cybenko, G. 1989. Approximation by superpositions of a sigmoid function. *Mathematics of Control, Signals, and Systems* 2, 303–14.

Elman, J. 1990. Finding structure in time. *Cognitive Science* 14, 179–211.

Funahashi, K. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks* 2, 183–92.

Gallant, A. and White, H. 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell.

Hornik, K. 1991. Approximation capabilities of multilayer feedforward nets. *Neural Networks* 4, 231–42.

Hornik, K. 1993. Some new results on neural network approximation. *Neural Networks* 6, 1069–72.

Hornik, K., Stinchcombe, M. and White, H. 1989. Multi-layer feedforward networks are universal approximators. *Neural Networks* 2, 359–66.

Hornik, K., Stinchcombe, M. and White, H. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks* 3, 551–60.

Jordan, M. 1986. Serial order: a parallel distributed processing approach. Report No. 8604. Institute for Cognitive Science, University of California San Diego.

Kuan, C.-M. 1995. A recurrent Newton algorithm and its convergence properties. *IEEE Transactions on Neural Networks* 6, 779–83.

Kuan, C.-M. and Hornik, K. 1991. Convergence of learning algorithms with constant learning rates. *IEEE Transactions on Neural Networks* 2, 484–9.

Kuan, C.-M., Hornik, K. and White, H. 1994. A convergence result for learning in recurrent neural networks. *Neural Computation* 6, 420–40.

Kuan, C.-M. and Liu, T. 1995. Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometric* 10, 347–64.

Kuan, C.-M. and White, H. 1994. Artificial neural networks: an econometric perspective (with discussions). *Econometric Reviews* 13, 1–91, 139–43.

McClelland, J., Rumelhart, D. and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2, Cambridge, MA: MIT Press.

McCulloch, W. and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5, 115–33.

Rissanen, J. 1986. Stochastic complexity and modeling. *Annals of Statistics* 14, 1080–100.

Rissanen, J. 1987. Stochastic complexity (with discussions). *Journal of the Royal Statistical Society, B* 49, 223–39, 252–65.

Robbins, H. and Monro, S. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22, 400–7.

Rosenblatt, F. 1958. The perception: A probabilistic model for information storage and organization in the brain. *Psychological Reviews* 62, 386–408.

Rumelhart, D., Hinton, G. and Williams, R. 1986. Learning internal representation by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, ed. D. Rumelhart, J. McClelland and the PDP Research Group. Cambridge, MA: MIT Press.

Rumelhart, D., McClelland, J. and the PDP Research Group, eds. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, Cambridge, MA: MIT Press.

Stinchcombe, M. and White, H. 1998. Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* 14, 295–325.

Swanson, N. and White, H. 1997. A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics* 79, 540–50.

White, H. 1988. Economic prediction using neural networks: the case of IBM stock prices. *Proceedings of the Second Annual IEEE Conference on Neural Networks II*. New York:

IEEE Press.

White, H. 1989. Some asymptotic results for learning in single hidden layer feedforward network models. *Journal of the American Statistical Association* 84, 1003–13.

How to cite this article

Kuan, Chung-Ming. "artificial neural networks." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000134> doi:10.1057/9780230226203.0063

The New Palgrave Dictionary of Economics Online

artificial regressions

James G. MacKinnon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An artificial regression is a linear regression that is associated with some other econometric model, which is usually nonlinear. It can be used for a variety of purposes, in particular computing covariance matrices and calculating test statistics. The best-known artificial regression is the *Gauss–Newton regression*, whose key properties are shared by all artificial regressions. The chief advantage of artificial regressions is conceptual: because econometricians are very familiar with linear regression models, using them for computation reduces the chance of errors and makes the results easier to comprehend intuitively.

Keywords

artificial regressions; binary response model regression; bootstrap; double-length artificial regression; efficient score tests; Gauss–Newton regression; generalized method of moments; heteroskedasticity; heteroskedasticity-consistent covariance matrices; instrumental variables; Lagrange multiplier tests; multivariate nonlinear regression models; non-nested hypotheses; outer product of the gradient regression; RESET test; score tests; specification

Article

An *artificial regression* is a linear regression that is associated with some other econometric model, which is usually, but not always, nonlinear. It can be used for a variety of purposes, in particular, computing covariance matrices and calculating test statistics. The best-known artificial regression is the *Gauss–Newton regression* (GNR), which is discussed in the next section. All artificial regressions share the key properties of the GNR.

The Gauss – Newton regression

A univariate nonlinear regression model may be written as

$$y_t = x_t(\beta) + u_t,\ u_t \sim \mathrm{IID}(0, \sigma^2),\ t = 1, \ldots, n, \tag{1}$$

where y_t is the tth observation on the dependent variable, and β is a k-vector of parameters to be estimated. Here the scalar function $x_t(\boldsymbol{\beta})$ is a nonlinear regression function which may depend on exogenous and/or predetermined variables. The model (1) may also be written using vector notation as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u},\ \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{2}$$

where $\boldsymbol{y}$ is an n-vector with typical element y_t, $\boldsymbol{x}(\boldsymbol{\beta})$ is an n-vector with typical element $x_t(\boldsymbol{\beta})$, and $\mathbf{I}$ is an $n\times n$ identity matrix.
The Gauss–Newton regression that corresponds to (2) is

$$\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}(\boldsymbol{\beta})\boldsymbol{b} + \text{residuals}, \tag{3}$$

where $\boldsymbol{b}$ is an n-vector of regression coefficients, and the matrix $\boldsymbol{X}(\boldsymbol{\beta})$ is $n\times k$ with tith element the derivative of $x_t(\boldsymbol{\beta})$ with respect to β_i, the ith component of β. The regressand here is a vector of residuals, and the regressors are matrices of derivatives. When regression (3) is evaluated at the least-squares estimates $\hat{\boldsymbol{\beta}}$, it becomes

$$\hat{\boldsymbol{u}} \equiv \boldsymbol{y} - \hat{\boldsymbol{x}} = \hat{\boldsymbol{X}}\boldsymbol{b} + \text{residuals}, \tag{4}$$

where $\hat{\boldsymbol{x}} \equiv \boldsymbol{x}(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}})$. Since the regressand of this artificial regression must be orthogonal to all the regressors, running the GNR (4) is an easy way to check that the NLS estimates actually satisfy the first-order conditions.

The usual OLS covariance matrix for $\hat{\boldsymbol{b}}$ from regression (4) is

$$s^2(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}, \text{ where } s^2 = \frac{1}{n-k}(\boldsymbol{y}-\hat{\boldsymbol{x}})'(\boldsymbol{y}-\hat{\boldsymbol{x}}). \tag{5}$$

This is also the usual estimator of the covariance matrix of the NLS estimator $\hat{\boldsymbol{\beta}}$ under the assumption that the errors are IID. If that assumption were relaxed to allow for heteroskedasticity of unknown form, then (5) would be replaced by a heteroskedasticity-consistent covariance matrix (HCCME) of the form

$$(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}, \tag{6}$$

where $\widehat{\boldsymbol{\Omega}}$ is an $n\times n$ diagonal matrix with squared residuals, probably rescaled, on the principal diagonal. The matrix (6) is precisely what a regression package would give if we ran the GNR (4) and requested an HCCME. Similar results hold if we relax the independence assumption and use a HAC estimator. In every case, a standard estimator of the covariance matrix of $\hat{\boldsymbol{b}}$ from the artificial regression (4) is also perfectly valid for the NLS estimates $\hat{\boldsymbol{\beta}}$.

If we evaluate the GNR (3) at a vector of restricted estimates $\tilde{\boldsymbol{\beta}}$, we can use the resulting artificial regression to test the restrictions. For simplicity, assume that $\tilde{\boldsymbol{\beta}} = [\tilde{\boldsymbol{\beta}}_1' \ \boldsymbol{0}']'$, where β_1 is a k_1-vector and β_2, which is equal to $\mathbf{0}$ under the null hypothesis, is a k_2–vector. In this case, the GNR becomes

$$\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{X}}_1\boldsymbol{b}_1 + \tilde{\boldsymbol{X}}_2\boldsymbol{b}_2 + \text{residuals}. \tag{7}$$

The ordinary F statistic for $\boldsymbol{b}_2=\mathbf{0}$ is asymptotically valid as a test for $\beta_2=\mathbf{0}$, and it is asymptotically equal, under the null hypothesis, to the F statistic for $\beta_2=\mathbf{0}$ in the nonlinear regression (1). Of course, when $\boldsymbol{X}_2$ has just one column, the t statistic for the scalar b_2 to equal zero is also asymptotically valid. Yet another test statistic that is frequently used is n times the uncentred R^2 from regression (7), which is asymptotically distributed as $\chi^2(k_2)$ under the null hypothesis.

The GNR (3) can also be used as part of a quasi-Newton minimization procedure if it is evaluated at any

vector, say $\beta_{(j)}$, where *j* denotes the *j*th step of an iterative procedure. In fact, this is where the name of the GNR came from. It is not hard to show that the vector

$$\boldsymbol{b}_{(j)} \equiv (\boldsymbol{X}_{(j)}^{'}\boldsymbol{X}_{(j)})^{-1}\boldsymbol{X}_{(j)}(\boldsymbol{y} - \boldsymbol{x}_{(j)}),$$

where the notation should be obvious, is asymptotically equivalent to the vector that defines a Newton step starting at $\beta_{(j)}$. The vector $\boldsymbol{b}_{(j)}$ is asymptotically equivalent to what we would get by postmultiplying minus the inverse of the Hessian of the sum of squared residuals function by the gradient. Because of this, the GNR has the same *one-step property* as Newton's method itself. If we evaluate (3) at any consistent estimator, say $\dot{\boldsymbol{\beta}}$, then the one-step estimator $\boldsymbol{\beta}^{'} = \dot{\boldsymbol{\beta}} + \dot{\boldsymbol{b}}$ is asymptotically equivalent to the NLS estimator $\hat{\boldsymbol{\beta}}$.
For more detailed treatments of the Gauss–Newton regression, see MacKinnon (1992) and Davidson and MacKinnon (2001; 2004).

Properties of artificial regressions

A very general class of artificial regressions can be written as

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}, \tag{8}$$

where θ is a parameter vector of length *k*, $\boldsymbol{r}(\boldsymbol{\theta})$ is a vector of length an integer multiple of the sample size *n*, and $\boldsymbol{R}(\boldsymbol{\theta})$ is a matrix with *k* columns and as many rows as $\boldsymbol{r}(\boldsymbol{\theta})$. In order to qualify as an artificial regression, the linear regression (8) must satisfy three key properties.

1. 1. The regressand $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$ is orthogonal to every column of the matrix of regressors $\boldsymbol{R}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ denotes a vector of unrestricted estimates. That is,

$$\boldsymbol{R}^{'}(\hat{\boldsymbol{\theta}})\boldsymbol{r}(\hat{\boldsymbol{\theta}}) = 0. \tag{9}$$

2. 2. The asymptotic covariance matrix of $n^{1/2}(\hat{\theta} - \theta_0)$ is given either by

$$\operatorname*{plim}_{n\to\infty} (n^{-1}\boldsymbol{R}'(\hat{\boldsymbol{\theta}})\boldsymbol{R}(\hat{\boldsymbol{\theta}}))^{-1}, \text{ or by} \tag{10}$$

$$\operatorname*{plim}_{n\to\infty} s^2(n^{-1}\boldsymbol{R}'(\hat{\boldsymbol{\theta}})\boldsymbol{R}(\hat{\boldsymbol{\theta}}))^{-1}, \tag{11}$$

where s^2 is the OLS estimate of the error variance obtained by running regression (8) with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Of course, this is also true if $\hat{\boldsymbol{\theta}}$ is replaced by any other consistent estimator of θ.

3. 3. If $\dot{\boldsymbol{\theta}}$ denotes a consistent estimator, and $\dot{\boldsymbol{b}}$ denotes the vector of estimates obtained by running regression (8) evaluated at $\dot{\boldsymbol{\theta}}$, then

$$\operatorname*{plim}_{n\to\infty} n^{1/2}(\dot{\boldsymbol{\theta}} + \dot{\boldsymbol{b}} - \boldsymbol{\theta}_0) = \operatorname*{plim}_{n\to\infty} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{12}$$

This is the one-step property, which holds because the vector $\dot{\boldsymbol{b}}$ is asymptotically equivalent to a single Newton step.

There exist many artificial regressions that take the form of (8) and satisfy conditions 1, 2, and 3. Some of these will be discussed in the next section. We have seen that the GNR satisfies these conditions and that its asymptotic covariance matrix is given by (11).

The most widespread use of artificial regressions is for specification testing. Of course, any artificial regression can be used to test restrictions on the model to which it corresponds. We simply evaluate the artificial regression for the unrestricted model at the restricted estimates, as in (7). However, in many cases, we can also use artificial regressions to test model specification without explicitly specifying an alternative. Consider the artificial regression

$$\boldsymbol{r}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}(\hat{\boldsymbol{\theta}})\boldsymbol{b} + Z(\hat{\boldsymbol{\theta}})c + \text{residuals}, \tag{13}$$

which is evaluated at unrestricted estimates $\hat{\boldsymbol{\theta}}$. Here $Z(\boldsymbol{\theta})$ is a matrix with *r* columns, each of which is supposed to be asymptotically uncorrelated with $\boldsymbol{r}(\boldsymbol{\theta})$, that has certain other properties which ensure that standard test statistics for $\boldsymbol{c}=\mathbf{0}$ are asymptotically valid. In effect, regression (13) must have the same properties as if it corresponded to an unrestricted model. See Davidson and MacKinnon (2001; 2004) for details.

When the artificial regression (13) is a GNR, $\boldsymbol{r}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{u}}$ and $\boldsymbol{R}(\hat{\boldsymbol{\theta}}) = \hat{X}$. Such a GNR can be used to implement a number of well-known specification tests, including the following ones.

- If we let $Z(\hat{\boldsymbol{\theta}})$ be a vector of squared fitted values, then the *t* statistic for the coefficient on the test regressor to be zero can be used to perform one version of the well-known RESET test (Ramsey, 1969).
- If we let $Z(\hat{\boldsymbol{\theta}})$ be an $n\times p$ matrix containing the residuals lagged once through *p* times, either the *F* statistic for $\boldsymbol{c}=\mathbf{0}$ or *n* times the uncentred R^2 can be used to perform a standard test for *p*th order serial correlation (Godfrey, 1978).
- If we let $Z(\hat{\boldsymbol{\theta}})$ be the vector $\hat{\boldsymbol{w}} - \hat{\boldsymbol{x}}$, where $\hat{\boldsymbol{w}}$ denotes the fitted values from a non-nested alternative model, then the *t* statistic on the test regressor can be used to perform a non-nested hypothesis test, namely, the *P* test proposed by Davidson and MacKinnon (1981).

Like all asymptotic tests, the three tests just described may not have good finite-sample properties. This is particularly true for the *P* test and other non-nested hypothesis tests. Finite-sample properties can often be greatly improved by bootstrapping, which is quite easy to do in these cases. For a recent survey of bootstrap methods in econometrics, see Davidson and MacKinnon (2006).

More artificial regressions

A great many artificial regressions have been proposed over the years, far more than there is space to discuss here. Some of them apply to very broad classes of econometric models, and others to quite narrow ones.

One of the most widely applicable and commonly used artificial regressions is the *outer product of the gradient* (OPG) regression. It applies to every model for which the log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}),$$

(14)

where ℓ_t is the contribution to the log-likelihood made by the *t*th observation, and θ is a *k*-vector of parameters. The *n*×*k* matrix of contributions to the gradient, $G(\theta)$, has typical element

$$G_{ti}(\theta) \equiv \frac{\partial \ell_t(\theta)}{\partial \theta_i}.$$
(15)

Summing the elements of the *i*th column of this matrix yields the *i*th element of the gradient. The OPG regression is

$$\iota = G(\theta)b + \text{residuals},$$
(16)

where ι is an *n*-vector of ones.
It is easy to see that the OPG regression satisfies condition 1, since the inner product of ι and $G(\theta)$ is just the gradient, which must be zero when evaluated at the maximum likelihood estimates $\hat{\theta}$. That it satisfies condition 2 follows from the fact that the plim of the matrix $n^{-1}G'(\theta)G(\theta)$ is the information matrix, which implies that the asymptotic covariance matrix is give by (10). The OPG regression also satisfies condition 3, and it is therefore a valid artificial regression.
Because it applies to such a broad class of models, the OPG regression is easy to use in a wide variety of contexts. This includes information matrix tests (Chesher, 1983; Lancaster, 1984) and conditional moment tests (Newey, 1985), both of which may be thought of as special cases of regression (13). However, because $n^{-1}G'(\hat{\theta})G(\hat{\theta})$ tends to be an inefficient estimator of the information matrix, tests based on the OPG regression often have poor finite-properties, iterative procedures based on it may converge slowly, and covariance matrix estimates may be poor. Davidson and MacKinnon (1992) contains some simulation results which show just how poor the finite-properties of tests based on the OPG regression can be. However, these properties can often be improved dramatically by bootstrapping.
Another artificial regression that applies to a fairly general class of models estimated by maximum likelihood is the *double-length artificial regression* (DLR), proposed by Davidson and MacKinnon (1984). The class of models to which it applies may be written as

$$f_t(y_t, \boldsymbol{\theta}) = \varepsilon_t,\ t = 1, \ldots, n,\ \varepsilon_t \sim \mathrm{NID}(0, 1), \tag{17}$$

where $f_t(\cdot)$ is a smooth function that depends on the random variable y_t, on a k-vector of parameters θ, and, implicitly, on exogenous and/or predetermined variables. This class of models is much more general than may be apparent at first. It includes both univariate and multivariate linear and nonlinear regression models, as well as models that involve transformations of the dependent variable. The main restrictions are that the dependent variable(s) must be continuous and that the distribution(s) of the error terms must be known.

As its name suggests, the DLR has $2n$ observations. It can be written as

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{\theta}) \\ \iota \end{bmatrix} = \begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y}, \boldsymbol{\theta}) \end{bmatrix} \boldsymbol{b} + \text{residuals}. \tag{18}$$

Here $\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{\theta})$ is an n-vector with typical element $f_t(y_t, \boldsymbol{\theta})$, ι is an n-vector of ones, $\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{\theta})$ is an $n \times k$ matrix with typical element $\partial f_t(y_t, \theta) / \partial \theta_i$, and $\boldsymbol{K}(\boldsymbol{y}, \boldsymbol{\theta})$ is an $n \times k$ matrix with typical element $\partial k_t(y_t, \boldsymbol{\theta}) / \partial \theta_i$, where

$$k_t(y_t, \boldsymbol{\theta}) \equiv \log \left| \frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial y_t} \right|$$

is a Jacobian term that appears in the log-likelihood function for the model (17). The information matrix associated with the DLR (18) has the form

$$\frac{1}{n}(\boldsymbol{F}'(\boldsymbol{\theta})\boldsymbol{F}(\boldsymbol{\theta}) + \boldsymbol{K}'(\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{\theta})). \tag{19}$$

In most cases, this is a much more efficient estimator than the one associated with the OPG regression. As a result, inferences based on the DLR are generally more reliable than inferences based on the OPG regression. See, for example, Davidson and MacKinnon (1992). The DLR is not the only artificial

regression for which the number of ‘observations’ is a multiple of the actual number. For other examples, see Orme (1995).

Ideally, an information matrix estimator should depend on the data only through estimates of the parameters. A Lagrange multiplier, or score, test based on such an estimator is often called an *efficient score test*. Because (19) often does not satisfy this condition, using the DLR generally does not yield efficient score tests. In contrast, at least for models with no lagged dependent variables, the GNR does yield efficient score tests, as do several other artificial regressions.

A number of somewhat specialized artificial regressions can be obtained as modified versions of the Gauss–Newton regression. These include two different forms of GNR that are robust to heteroskedasticity of unknown form, a variant of the GNR for models estimated by instrumental variables, a variant of the GNR for models estimated by the generalized method of moments, a variant of the GNR for multivariate nonlinear regression models, and the binary response model regression (BRMR), which applies to models like the logit and probit model. See Davidson and MacKinnon (2001; 2004) for detailed discussions and references.

Of course, any quantity that can be computed using an artificial regression can also be computed directly by using a matrix language. Why then use artificial regressions for computation? This is, to some extent, simply a matter of taste. One potential advantage is that most statistics packages perform least squares regressions efficiently and accurately. In my view, however, the chief advantage of artificial regressions is conceptual. Because econometricians are very familiar with linear regression models, using them for computation reduces the chance of errors and makes the results easier to comprehend intuitively.

See Also

- non-nested hypotheses
- serial correlation and serial dependence
- testing

Bibliography

Chesher, A. 1983. The information matrix test: simplified calculation via a score test interpretation. *Economics Letters* 13(1), 45–8.

Davidson, R. and MacKinnon, J.G. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49, 781–93.

Davidson, R. and MacKinnon, J.G. 1984. Model specification tests based on artificial linear regressions. *International Economic Review* 25, 485–502.

Davidson, R. and MacKinnon, J.G. 1992. A new form of the information matrix test. *Econometrica* 60, 145–57.

Davidson, R. and MacKinnon, J.G. 2001. Artificial regressions. In *Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Davidson, R. and MacKinnon, J.G. 2004. *Econometric Theory and Methods*. New York: Oxford University Press.

Davidson, R. and MacKinnon, J.G. 2006. Bootstrap methods in econometrics. In *Palgrave Handbooks of Econometrics. Volume 1: Econometric Theory*, ed. T.C. Mills and K.D. Patterson. Basingstoke: Palgrave Macmillan.

Godfrey, L.G. 1978. Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica* 46, 1303–10.

Lancaster, T. 1984. The covariance matrix of the information matrix test. *Econometrica* 52, 1051–3.

MacKinnon, J.G. 1992. Model specification tests and artificial regressions. *Journal of Economic Literature* 30, 102–46.

Newey, W.K. 1985. Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047–70.

Orme, C.D. 1995. On the use of artificial regressions in certain microeconometric models. *Econometrica* 11, 290–305.

Ramsey, J.B. 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31, 350–71.

How to cite this article

MacKinnon, James G. "artificial regressions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000233> doi:10.1057/9780230226203.0064

The New Palgrave Dictionary of Economics Online

Ashton, Thomas Sutcliffe (1889 – 1968)

Phyllis Deane

From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Ashton, T.S.; economic history

Article

T.S. Ashton was born in Lancashire in 1889, graduated from Manchester University in 1909 and returned there in 1921 (after some years at the Universities of Sheffield and Birmingham) to teach political economy and economic history in the Faculty of Commerce. By the time he took up Eileen Power's chair of economic history at the London School of Economics in 1944, he had made a substantive and distinctive contribution to the history of the industrial revolution in three research monographs: *Iron and Steel and in the Industrial Revolution* (1924), *The Coal Industry of the Eighteenth Century* (1929, written with Joseph Sykes), and *An Eighteenth Century Industrialist: Peter Stubs of Warrington* (1939). Over the next decade this unassuming, humane, passionately non-dogmatic scholar had become the leader of a new generation of economic historians, a generation whose members had been schooled in the theories and analytical techniques of economics rather than in the thinking habits of a history faculty.

The two industrial studies and the business history published while Ashton was in Manchester were exercises in applied economics, based on detailed investigation of primary sources (including a mass of business ledgers, letters and accounts) and of a wide range of 18th-century material reflecting economic and social events, transactions and opinions. These researches gave him a formidable armoury of qualitative and quantitative data from which he set out in the 1940s explicitly to ‘find answers (partial and provisional though these may be) to the questions economists ask, or should ask, of the past’.

Ashton's last three books constituted a coherent and cumulative contribution to the economic history of the first country to make the transition to modern economic growth. His highly original essay *The Industrial Revolution* (1948) appeared just when the industrialization problems of developing countries were assuming major importance on the applied economists' research agenda and became a long-running bestseller. His *Economic History of England: The Eighteenth Century* (1955), the prime example of a new genre of economic history, contained the first systematic attempt to use standard economic theory

to explain long-term changes in the general level of prices and economic activity over that century, and also injected a characteristic objectivity into the perennial controversy over the standard of living of workers during the industrial revolution. In his last book, *Economic Fluctuations in England 1700–1800* (1959), he shifted his analysis of 18th-century economic change to a short-run focus. But by then only the pure theorists and the econometricians were actively interested in cyclical analysis, and Ashton was effectively distanced from both groups by his persistent concern with taking account of social as well as economic factors in economic change and by his realistically discriminating approach to the use of either abstract concepts or statistical evidence.

Selected works

1924. *Iron and Steel in the Industrial Revolution*. Manchester: Manchester University Press.

1929. (With Joseph Sykes). *The Coal Industry of the Eighteenth Century*. Manchester: Manchester University Press.

1939. *An Eighteenth Century Industrialist: Peter Stubs of Warrington*. Manchester: Manchester University Press.

1948. *The Industrial Revolution*. London: Oxford University Press.

1949. The standard of life of the workers of England, 1790–1830. *Journal of Economic History*, Supplement 9, 19–38.

1955. *An Economic History of England: The Eighteenth Century*. London: Methuen.

1959. *Economic Fluctuations in England, 1700–1800*. Oxford: Clarendon Press.

How to cite this article

Deane, Phyllis. "Ashton, Thomas Sutcliffe (1889–1968) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000076> doi:10.1057/9780230226203.1878

The New Palgrave Dictionary of Economics Online

assets and liabilities

Kenneth E. Boulding

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Assets and liabilities come into prominence when double-entry bookkeeping and the balance sheet were invented, a prerequisite for the development of complex market economies. Production is a function of the size and structure of real assets, including human capital. We derive more satisfaction from the use of assets rather than from their consumption. Like any business, the household has a balance sheet of assets and liabilities. The lack of capital accounting in government means that many of its activities have no real ‘bottom line’, and their value is usually assessed in non-economic terms, sometimes resulting in catastrophic mistakes of judgement.

Keywords

assets and liabilities; balance sheet; bank deposits; bankruptcy; capital accounting; depreciation; double-entry bookkeeping; equity; fixed capital; Great Depression; household capital; human capital; net national product; working capital

Article

The concepts of assets and liabilities are very closely related. Liabilities can be regarded as negative assets. The term ‘assets’ is related to the French ‘*assez*’, meaning ‘enough’. It emerges as a legal concept, particularly in laws relating to bankruptcy, the question being whether in bankruptcy assets are enough to meet all the liabilities. Historically, there has been a tendency to distinguish between real, personal and equitable assets, but these distinctions are now of little importance.

In accounting, assets and liabilities come into prominence with the invention of double-entry bookkeeping and the balance sheet, a concept which seems to have originated in northern Italy at least by the 12th or 13th century. This concept was important as a prerequisite for the development of complex markets and profit-oriented economies as an improvement in the information system. Before the invention of the balance sheet it was hard for a merchant to know whether he had made any profit or

not.

It is the convention of the balance sheet that assets are listed on one side and liabilities and equity on the other side, equity being defined fundamentally as net assets; that is, assets minus liabilities, which are negative assets. Accounting practice divides both assets and liabilities into a number of categories. Assets are commonly divided into current, deferred and fixed assets. Current assets consist of cash, bank deposits, short-term notes, accrued interest, inventories of goods in process or finished goods which are expected to be sold within the accounting period, usually six months or a year. Sometimes items like repair parts are included in this category, even though their life on the shelf may be longer. Another item may be deferred assets, such as insurance, advertising payments which are paid in advance where the services have not yet been performed. Finally, there are fixed assets of a lasting nature, such as buildings and machines. There is also a category of intangible assets, like goodwill, value of patents, and so on. These tend to have a rather dubious status in accounting practice.

Liabilities have a somewhat similar categorization. Current liabilities are those which are expected to be paid off in the accounting period – wage claims, short-term loans, accounts payable, and so on. Current assets minus current liabilities is sometimes called ‘working capital’. Somewhat corresponding to fixed assets are long-term loan obligations. The sum of all assets minus the sum of all liabilities is the equity or net worth. This is usually divided into paid-up capital and undistributed profits.

Every time an event happens to an organization that has a balance sheet, the items in the balance sheet change. Thus, in production, when wheat is ground into flour the stock of wheat diminishes and of flour increases. Likewise, the stock of money may diminish as wages are paid, and the product of the work is added to assets. Assets diminish as machinery and buildings depreciate. Exchanges, purchases and sales are reflected in an increase in what is acquired and a decrease in what is given up for it. When money is borrowed, cash is increased on the asset side and the debt is increased on the liability side. It is a convention of cost accounting that both exchange and production represent transfers of equal values. When something is purchased, it is valued at the amount paid for it, so that the net worth does not change. Similarly, in production, the value of what is produced is equal to what has been consumed (i.e. destroyed) in the process, whether this is the money used to pay wages, raw materials used up or depreciation.

Profit is the growth of net worth, which happens when some asset is revalued, usually at the moment of sale. If it is sold for more than the accounting cost, the difference is an increase in net worth. Before sale, the asset is valued at cost. After the sale, if it is profitable, the asset disappears from the accounts but a larger sum of money than the value of the asset is entered, and this is why the net worth increases. When profits are distributed the liquid assets are diminished and the net worth diminishes by the same amount. Interest-bearing liabilities grow at the rate of interest, which accrues. This diminishes the net worth, this being the growth of a negative asset. Interest paid, cash or some liquid asset, diminishes by the same amount as accrued interest diminishes. There is no change in the net worth. Profit is made by constant manipulation of the assets through production and exchange to increase the total value of assets at a greater rate than interest on liabilities is accruing. Debt is presumably incurred because of a belief that it will increase the total volume of assets sufficiently so that some kind of economies of scale will permit a rate of growth of the increased assets more rapid than the rate of interest on the liabilities that are incurred in order to expand the assets.

An important problem in accounting, by no means satisfactorily solved, is how to deal with inflation and deflation. In order to get a net worth or ‘bottom line’, both assets and liabilities have to be expressed in

terms of the monetary unit. In the case of physical assets, this means multiplying the quantity of the assets by some valuation coefficient which will turn it into a number of monetary units. Where the asset is constantly being bought and sold, the price, or ratio of exchange, is generally used as a valuation coefficient. In the case of fixed capital, the value is usually reckoned by taking an original purchase price and depreciating it over time by various methods, either at a constant percentage rate or at a constant amount per year. This figure is very arbitrary in any case and in periods of inflation and deflation becomes extremely misleading. Inflation tends to increase accounting profits because fixed capital tends to be undervalued.

Another element in the situation is that all profit-making involves buying something at a certain price or cost at one time and selling it at a later time. If in the time interval all prices have risen, there is a spurious profit, which is not really represented by purchasing power. Thus there is much to be said for having a profit figure indexed, although the technical difficulties in this have so far prevented very much application of this principle. Inflation, therefore, produces illusory high profits; deflation, likewise, produces illusory low profits. This happened in the Great Depression, when accounting profits in 1932 and 1933 were negative. Unfortunately, it is accounting profits rather than real profits which tend to govern business expectations and decisions.

Beyond accounting, assets and liabilities make a very important contribution to the understanding of both the description and the dynamics of the economic system. Every liability is or should be an asset in some other balance sheet, for every debt is an asset to the creditor and a liability to the debtor. When we sum all the balance sheets in society, therefore, we should come out with an overall balance sheet that consists merely of real assets on one side and the total net worth of the society on the other. There is some question as to whether we should include money of various kinds in real assets. Bank deposits, of course, are assets to the holder and liabilities to the bank, so if we sum all assets, including banks, deposits would disappear. Even paper money is in a certain sense a liability of the government, although it is not usually reckoned as such, for it has to be accepted by government in payment of taxes. An important proposition follows from the concept of the aggregate balance sheet, that an increase in net assets, that is, investment, will produce an increase in the total of net worth, which is profit. This may be offset by other events. This is an important clue, however, to the dynamics of a great depression, which exhibits positive feedback: a decline in investment produces a decline in profits, a decline in profits produces a further decline in investment, a further decline in profits, and so on. This is clearly what happened between 1929 and 1933 in the capitalist world.

The relation of assets and liabilities to income, production and consumption is very important. Real assets can be regarded as a kind of ecosystem of goods, with the stock of each good representing a population. Production is then equivalent to births, consumption to deaths. Production minus consumption is the increase in the total stock of a particular good. The net national product is equal to the total production of goods, which is equal to the total consumption, plus an increase in the total stock of goods, just as an increase in any population is equal to the number of births minus the number of deaths in a given period.

Production is a function of the size and structure of real assets themselves, which is particularly clear if we include the value of the human bodies and minds (i.e. human capital) in the total, as ideally we should. Economists have an unfortunate way of regarding households as a kind of black box outside the economy proper. Actually they are very much a part of it, and household capital – houses, furniture, automobiles, clothing, and so on – is very close to half of the total in a modern society. When we fly

over a city we see far more houses than factories. If we compare the capital around us at our workplace with the capital around us in our home, for a considerable part of the population the home capital is much larger than the capital at work.

Another very important problem is the contribution of assets, particularly household assets to economic welfare. There is a long tradition in economics that regards consumption as the main method of measurement of riches. It is clear, however, that we get most of our satisfaction from the use and enjoyment of assets rather than from their consumption. I get no satisfaction out of the fact that my car, house and clothing are wearing out. What I get satisfaction out of is using them. An increase in durability, especially of household capital, therefore, is an addition to economic welfare. This is a point much neglected by economists. Consumption, then, can usually be seen as a bad thing, and production as what is necessary to offset it. There are exceptions to this rule. We like eating. We like the activity of producing in itself, even though it involves the using up of raw materials and so on. Thus the economic welfare function would include both assets of all kinds and certain forms of production and consumption, that is, income. Economists have often confused consumption with household expenditure or purchases, again because they regard the household as outside the economy. In modern society this can be very misleading, for household purchases are governed in no small degree by the depreciation of household capital to the point where it has to be replaced, so this depreciation is a very important aspect of consumption and income. Household purchases are exchange, not consumption. The production of assets include households also tends to be neglected, and it is an important part of the total economy in terms of cooking, mending, painting and repairing. The household has a balance sheet of assets and liabilities just as much as a business does and cannot be understood without it.

Human capital, both in terms of assets and liabilities, is a concept which has achieved some recognition. Economic development is primarily a process in human learning and the increase in human capital. A natural catastrophe or a war which destroys physical capital is restored remarkably quickly if the human capital remains intact and the knowledge and the know-how are unimpaired. We often do not realize that an enormous destruction of capital takes place every year just by depreciation and consumption. Even spectacular disasters are often just a relatively small addition to this annual destruction. The fact that some human beings have a negative human capital, both for themselves and for society, cannot be overlooked, though our social accounting system is ill-equipped to deal with this problem. In political decisions, however, we do recognize it. The criminal justice system is at least intended to diminish negative human capital; the educational system, to increase positive human capital. The fact that there is very little capital accounting in government means that considerable parts of its activity, like unilateral national defence organizations, do not really have a ‘bottom line’, and their value is usually assessed in non-economic terms, which can easily lead into catastrophic mistakes of judgement.

See Also

- accounting and economics
- double-entry bookkeeping

How to cite this article

Boulding, Kenneth E. "assets and liabilities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000150> doi:10.1057/9780230226203.0065

The New Palgrave Dictionary of Economics Online

assortative matching

Li, Hao
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the simple economics of matching by characteristics. The goal is to understand sorting patterns in the marriage market and other matching markets by focusing on the nature of the gain from match and the mechanism of the market force of competition.

Keywords

assortative matching; Becker, G.; Beckmann, M.; complements and substitutes; incomplete information; Koopmans, T.; marriage markets; matching frictions; matching markets; mixed matching; price discrimination; self-matching; substitutes and complements; transferable utility; wage heterogeneity, sources of

Article

In a marriage market the competition for spouses leads to sorting of mates by characteristics such as wealth and education. ‘Positive assortative matching’ refers to a positive correlation in sorting between the values of the traits of husbands and wives (matching of likes); ‘negative assortative matching’ refers to a negative correlation (matching of unlikes). While it has been long recognized that sorting of husbands and wives by characteristics occurs in all cultures and societies, economists have tried to understand sorting patterns in the marriage market and other matching markets by focusing on the nature of the gain from match and the mechanism of the market force of competition.

The basic framework

A simple framework to illustrate the economic approach to sorting in matching markets is a two-sided marriage market with an equal number of men and women, who differ in one-dimensional characteristics called ‘type’ and have common preferences for higher types over lower types. In positive

assortative matching, the highest-type man mates the highest-type woman, and the second-highest-type man mates the second-highest-type woman, and so on. Negative assortative matching is between the highest-type man and the lowest-type woman, between the second-highest-type man and the-second-lowest type woman, and so on. We assume transferable utility and zero reservation utility from remaining single for each market participant. Then, the gain from a match can be represented by an increasing, positive-valued function *f*, which gives the match output $f(x, y)$ of any pair of type x man and type *y* woman. Consider two men, with types $x_H > x_L$, and two women, with types $y_H > y_L$. If type x_H and type x_L command the same price in terms of the utility transfer they demand from the wife for the match, then both type y_H and type y_L would prefer the higher-type man because *f* is increasing in male type. Competition for type x_H naturally leads to a higher price for type x_H than for type x_L. Whether the higher female type y_H can outbid type y_L for type x_H or vice versa depends on whether the male type and the female type are complements or substitutes in the match output function *f*. If

$$f(x_H, y_H) - f(x_H, y_L) > f(x_L, y_H) - f(x_L, y_L), \tag{1}$$

then the male type and the female type are complementary, because the marginal product of the female type is greater when matched with a higher male type (the left-hand side of inequality (1) than with a lower male type (the right-hand side of (1)). In this case, type y_L is willing to offer type x_H at most $f(x_H, y_L) - f(x_L, y_L)$ more than she offers type x_L, but by inequality (1) this difference is smaller than $f(x_H, y_H) - f(x_L, y_H)$, which is the most type y_H is willing to offer. Thus, type y_L will be outbid by type y_H for type x_H when the male type and the female type are complements. Since the argument is valid for any two pairs of men and women, the competition for spouses must lead to positive assortative matching. Conversely, if inequality (1) is reversed, male type and female type are substitutes. A lower female type can outbid a higher type for any male type, and the competition for spouses leads to negative assortative matching.

The differentiable version of inequality (1) is

$$\frac{\partial^2 f(x, y)}{\partial x \partial y} > 0. \tag{2}$$

Conditions (1) and (2) are commonly referred to as the (strict) ‘supermodularity’ condition of the match output function *f*. See Topkis (1998) for a comprehensive mathematical treatment of supermodularity, and Milgrom and Roberts (1990) and Vives (1990) for applications in game theory and economics. Inequality (1) can be rewritten as

$$f(x_H, y_H) + f(x_L, y_L) > f(x_L, y_H) + f(x_H, y_L). \quad (3)$$

Condition (3) suggests that positive assortative matching maximizes the sum of match outputs in the marriage market when male type and female type are complements in the match output function. This result is a direct application of Koopmans and Beckmann's (1957) theorem of equivalence between efficient matching, which maximizes the sum of match outputs among all feasible pairwise matchings, and competitive equilibrium matching, which obtains when each woman y takes as given a schedule of utility transfers $u(x)$ to men and chooses the male type that maximizes her utility. Competitive equilibrium matching can also be obtained as each man *x* takes as given a schedule of utility transfers $v(y)$ to women and chooses the female type that maximizes his utility. Shapley and Shubik (1972) model the marriage market with transferable utilities as a cooperative game. They show that a pair of transfer schedules that support an equilibrium matching correspond to the core of the game, so that no pair of a man and a woman not matched in equilibrium can form a blocking coalition that produces a match output greater than the sum of their respective transfers.

Applications of assortative matching

The results of Koopmans and Beckmann (1957) and Shapley and Shubik (1972) are obtained in a matching market without any hierarchical ordering of types. By introducing one-dimensional, heterogeneous types, Becker (1973) seeks to explain why sorting of mates by wealth, education and other characteristics is similar in the marriage market. He constructs a household production function and derives condition (1) for each of the characteristics separately by considering how the characteristic affects household output while holding other characteristics fixed. Becker's model can accommodate dissimilar sorting of mates by some characteristics as well; for example, negative assortative matching by wage rates may arise because the benefits from the division of labour within a household can make the earning abilities of the man and the woman substitutes for each other.

Sattinger (1980) uses condition (2) to explain why the distribution of earnings of workers is skewed to the right relative to the distribution of their measured skills. In a market that matches a continuum of workers with different skills to a continuum of positions of different capital investment, the distribution of earnings would have the same shape as the distribution of skills if matching is random. In Sattinger's theory of differential rents, positive assortative matching of worker skill and job capital investment occurs because skill and capital investment are complements. In this case, the distribution of earnings will not resemble the distributions of outputs of workers at a job with the average capital investment. Instead, workers with higher skills are paid more than those with lower skills both because they are more productive at any job and because they occupy positions with greater capital investments. Formally, in equilibrium the wage schedule *u* satisfies the first-order condition of type *y*'s maximization problem of choosing *x* to maximize $f(x, y) - u(x)$

$$\frac{\partial f(x, m(x))}{\partial x} = u'(x),$$

where $m(x)$ is the capital investment of the job occupied by the worker with skill *x* in equilibrium. It can shown that condition (2) and positive assortative matching imply that $f(x, y) - u(x)$ is concave in *x* at $x = m^{-1}(y)$, so the second-order condition is satisfied for each y. The first-order condition implies that the worker's wage increases at the rate of the marginal product of the worker's skill *x* at his equilibrium job, so that the rate of increase of *u* is augmented by the complementarity (condition 2) and positive assortative matching ($m'(x) > 0$). Therefore, with positive assortative matching, the distribution of earnings will be positively skewed relative to the distribution of skills.

Kremer (1993) highlights the role of positive assortative matching in economic development. In his model of a one-sided, many-to-many matching market, each firm consists of a fixed number of workers, each employed for a production task. Workers have different skills, with a higher-skilled worker less likely to make mistakes in performing his task. Condition (1) is assumed to capture the complementarity among worker skills in the sense that the production process of a firm requires completion of each task without mistakes. Self-matching obtains in equilibrium where each firm employs workers of identical skills. Kremer uses this form of positive assortative matching to explain the large wage and productivity differences between developing and developed countries that cannot be accounted for by their differences in levels of physical or human capital.

Self-matching will generally be inefficient and will not occur in equilibrium if production tasks in a firm differ in skill requirements. In Kremer and Maskin (1996), a firm consists of two workers with a match output function $f(x, y)$ that satisfies the supermodularity conditions (1) and (2) but is asymmetric in that $f(x, y) > f(y, x)$ for any $x > y$. The interpretation of the asymmetry is that the first argument in *f* represents the skill of the worker who does the manager's job, while the second argument represents the skill of the worker who performs the assistant's job. In any given firm, it is optimal to make the higher-skilled worker the manager and the lower-skilled worker the assistant, but it is no longer generally true that self-matching maximizes the total match outputs. Indeed, we can have

$$2f(z_H, z_L) > f(z_H, z_H) + f(z_L, z_L) \quad (4)$$

for some $z_H > z_L$, so that two firms each with the higher type z_H as the manager and the lower type z_L as the assistant produce more in total than two firms with the manager and the assistant having the same skill level. Note that inequality (4) does not contradict inequality (3) due to the asymmetry in *f*. Mixed matching may do better than self-matching because it can be more important to exploit the asymmetry in the match output function and have each high-skill worker as the manager of a firm than to exploit the

complementarity in *f* and have one high-skill worker as the assistant to the other high-skill worker. Kremer and Maskin find that efficient matching in their model depends on the skill distribution in the matching market, because the trade-off between the asymmetry and the complementarity in the match output function depends on the relative scarcity of high-skilled workers.

Frictions in matching markets

Assortative matching may be hindered by the presence of frictions in the matching market. For example, if there is a moral hazard problem in producing the match output by each matched pair, transferability of utilities will be restricted by incentive compatibility constraints. Legros and Newman (2002) discuss this and other examples of transaction costs, and find that equilibrium matching in these examples can be inefficient. Frictions can also arise due to incomplete information about type. Roth and Xing (1994) provide detailed descriptions of labour markets for entry-level professionals (such as lawyers and medical interns) in which early matches are sometimes made before complete information about matching characteristics, such as qualifications of job candidates and desirability of job positions becomes available. The complementarity in the match output function between the type of the applicant and the type of the job implies that there will be matching efficiency loss if matches are formed before the uncertainty about types is resolved. If all market participants are risk neutral, this efficiency loss is sufficient to rule out early matches as applicants compete for job positions. However, when some participants are risk averse, early matches provide them with some insurance against the payoff risks associated with late matches formed after complete information about types becomes available. Li and Suen (2000) apply competitive equilibrium analysis to the early matching market to determine the pattern of early matching, the terms of early matches, and the distribution of benefits in the early market. Early matching need not be positive-assortative in terms of expected type. Higher expected types of workers may face greater payoff risks from late matches due to the complementarity in the match output function. In this case, they may be willing to match with lower expected types of jobs to insure against the risks, while owners of higher expected types of jobs are content with waiting for late matches if they are risk neutral.

Private information about type may also result in frictions in the matching market. For example, many users of Internet dating agencies complain about the problems of misrepresentation and exaggeration by some users in the information they provide to the agencies. This problem arises because current matching services adopt a uniform pricing policy, and this in practice results in almost random matching. Damiano and Li (2007) point out that the complementarity in the match output function implies a version of the standard single-crossing condition in mechanism design problems, and an intermediary can use price discrimination to improve matching efficiency and generate greater revenue. They consider the problem of a monopoly matchmaker that uses a pair of fee schedules to sort different types of agents on the two sides into exclusive meeting places. The revenue-maximizing sorting need not be positive assortative (that is, efficient in the first-best sense). Conditions necessary and sufficient to recover positive assortative matching require that the complementarity in the match output function to be sufficiently strong to overcome the incentive cost to the matchmaker of eliciting private type information.

Matching frictions can arise also because finding type information about potential partners takes time or

involves costly effort. In the search and matching framework, each market participant randomly meets a currently unmatched agent from the other side of the market, and decides whether to form a match or to search again in the next period. Search is costless, but agents must trade off the benefit from starting to produce with the encountered partner right away against the opportunity cost of waiting for a better partner. With an exogenous probability of separation of matched agents who then re-enter the market, Shimer and Smith (2000) characterize the stationary search and matching equilibrium where the matching decisions of each type and the type distributions of unmatched agents are time-invariant. Types *x* and *y* in an agreeable match are assumed to use the Nash bargaining solution to split the net surplus, defined as the match output $f(x, y)$ minus the sum of the (endogenous) continuation payoffs $g(x)$ to *x* and $h(y)$ to *y* as unmatched agents. Shimer and Smith modify the definition of positive assortative matching in the frictionless world to allow for set-valued mutually agreeable matches. The match set of a type *x* is the intersection of the set of types that type *x* agrees to match with and the set of types that agree to match with *x*. In Shimer and Smith's definition, matching is positive-assortative where, if for any male types $x_H > x_L$ and female types $y_H > y_L$ such that y_H is in the match set of x_L and y_L is in the match set of x_H, then y_H is in the match set of x_H and y_L is in the match set of x_L. When match sets are convex, positive assortative matching requires the lowest and the highest type of the match set to be increasing in *x*. However, match sets need not be convex even though the match output function is supermodular. This is because the net surplus $f(x, y) - g(x) - h(y)$ is not necessarily quasi-concave in *y* for fixed *x*, so one cannot to say anything about how match sets vary across different *x*. Shimer and Smith provide conditions on *f* in addition to supermodularity to ensure convexity of match sets and re-establish positive assortative matching in a stationary equilibrium.

The stationary search and matching equilibrium does not capture the dynamics of matching in markets where there is no entry of a new cohort in each period and each matched pair receives their match output after the market closes for all participants. For example, many entry-level markets for professionals (such as academic economists) are organized around annual recruitment cycles. In these markets, matches are formed sequentially without centralized matching procedures. Damiano, Li and Suen (2005) consider such markets by constructing a two-sided, finite-horizon search and matching model with heterogeneous types and complementarity between types. The quality of the pool of potential partners deteriorates as agents who have found mutually agreeable matches exit the market. When search is costless and all agents participate in each matching round, the market performs a sorting function in that high types of agents have multiple chances to match with their peers. The matching efficiency measured by the total expected match outputs improves as the number of matching rounds increases; positive assortative matching is achieved if there are as many matching rounds as there are types. However, this sorting function is lost if agents incur an arbitrarily small cost in order to participate in each round. With a sufficiently rich type space relative to the number of matching rounds, the market unravels as almost all agents rush to participate in the first round, and match and exit with anyone they meet.

See Also

- marriage markets
- matching and market design

Bibliography

Becker, G. 1973. A theory of marriage: part I. *Journal of Political Economy* 81, 813–46.

Damiano, E. and Li, H. 2007. Price discrimination and efficient matching. *Economic Theory* 30, 243–63.

Damiano, E., Li, H. and Suen, W. 2005. Unravelling of dynamic sorting. *Review of Economic Studies* 72, 1057–76.

Koopmans, T.C. and Beckmann, M. 1957. Assignment problems and the location of economic activities. *Econometrica* 25, 53–76.

Kremer, M. 1993. The o-ring theory of economic development. *Quarterly Journal of Economics* 108, 551–75.

Kremer, M. and Maskin, E. 1996. Wage inequality and segregation by skill. Working Paper No. 5718. Cambridge, MA: NBER.

Legros, P. and Newman, A. 2002. Monotone matching in perfect and imperfect worlds. *Review of Economic Studies* 69, 925–42.

Li, H. and Suen, W. 2000. Risk-sharing, sorting, and early contracting. *Journal of Political Economy* 108, 1058–91.

Milgrom, P. and Roberts, J. 1990. Rationalizability, learning and equilibrium in games with strategic complementarities. *Econometrica* 45, 101–14.

Roth, A. and Xing, X. 1994. Jumping the gun: imperfections and institutions related to the timing of market transactions. *American Economic Review* 84, 992–1044.

Sattinger, M. 1980. *Capital and the Distribution of Labor Earnings*. Amsterdam: North-Holland.

Shapley, L. and Shubik, M. 1972. The assignment game I: the core. *International Journal of Game Theory* 1, 111–30.

Shimer, R. and Smith, L. 2000. Assortative matching and search. *Econometrica* 68, 343–69.

Topkis, D. 1998. *Supermodularity and Complementarity*. Princeton, NJ: Princeton University Press.

Vives, X. 1990. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics* 19, 305–21.

How to cite this article

Hao, Li. "assortative matching." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000214> doi:10.1057/9780230226203.0066

The New Palgrave Dictionary of Economics Online

assumptions controversy

Lawrence A. Boland

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

While the controversy concerning the realism of the assumptions of economic theory goes back to the mid-19th century, the issue today concerns only a reaction to Milton Friedman's famous 1953 methodology essay. That essay advanced the position that the assumptions of economic theory do not have to be realistic so long as they and the theories they form are useful. Many ideologically motivated critiques of this essay were published in the 1950s, 1960s and 1970s but all were problematic since they did not recognize that Friedman's essay was nothing more than a restatement of the common methodology called Instrumentalism.

Keywords

assumptions controversy; Cairnes, J. E.; induction; instrumentalism; methodology of economics; Mill, J. S.; New Deal; operationalism; perfect competition; Robbins, L. C.; Senior, N. W.; Testing

Article

Today, any reference to an ‘assumptions controversy’ immediately calls to mind the many critical reactions to Milton Friedman's famous 1953 essay. But historians of economic thought will also point out that there was an assumptions controversy going back to the mid-19th century involving John Stuart Mill, John Elliot Cairnes and Nassau Senior (for an excellent review of this ‘old’ assumptions controversy, see Hirsch, 1980). This old controversy was mainly between Mill and Senior and was about whether economics was an empirical science or a hypothetical one. The controversy was mediated by Cairnes and ultimately decided in his favour. For Cairnes, economic theory was true ‘because it rested on premises which were undeniably true’ (Hirsch, 1980, p. 105). But any application of theory can be compromised by ‘disturbing causes’ and so the application needed ‘to be compared with the facts’ to see just what disturbing causes needed ‘to be added in specific instances to make theory and facts correspond’ (1980, p. 105). According to Abraham Hirsch, Cairnes's position reigned for over three-

quarters of a century.
Friedman's essay was defending the use of perfect competition assumptions in applied economics against criticism of the assumption of universal maximization. The critics could easily find support in the philosophy of science of the day that claimed science is concerned with propositions that are meaningful because they are verifiable. But Friedman argued that, even in science, assumptions did not have to be true – only the logically derived results matter and theory should be judged according to whether these work or are useful. Friedman even argued it was acceptable to use simple assumptions that were obviously false on the grounds that one's theory might otherwise be so complex as to be useless.

Ideology as method

Given the strong objections of most economists of this period to Friedman's views on markets, the suspicion must arise that ideology accounted for much of the interest in his methodology (Boland, 2003). In particular, in the 1960s when Keynesian policies were thought by most mainstream economists to be obviously correct, Friedman's advocacy of a very limited role for the government was seen as a throwback to before the programmes of US President Franklin Roosevelt's New Deal that many other people thought helped overcome the Great Depression. But ideological arguments are not what academia is about. Instead, if one objects to Friedman's methodology, one must provide philosophical or scientific arguments against it to win the day. So, between 1957 and 1971 the controversy raged, not in the field of ideology but in the fields of semantics and methodology.
Ideology aside, it is difficult to understand why anyone would see Friedman's position to be very strong. After all, as I argued in Boland (1979), one can easily see Friedman's methodological position as nothing more than an up-to-date version of Instrumentalism (see instrumentalism and operationalism). And as such, if one were to ask Friedman or any Instrumentalists to defend their methodology – the methodology that claims the truth status of assumptions do not matter, only whether possibly false assumptions are useful – their only defence is to say that the Instrumentalist methodology itself works and hence is useful. There does not seem to be any other possible defence. But leading critics prior to 1979 seemed to think telling criticism could be provided. Unfortunately, none of their critiques was logically successful even though many opponents of Friedman's ideology wished to think so. To be effective, criticism of a doctrine must be in terms that a proponent of that doctrine would accept. Changing terms or imposing different objectives for the doctrine will not yield an effective or fair critique. All of the famous critiques published in the 1950s, 1960s and 1970s failed in this way.

Friedman's Instrumentalist methodology

As explained in Boland (1979), any theory, in terms of Friedman's viewpoint, is an argument for some given propositions or towards specific predictions. As such a theory consists only of a conjunction of assumption statements, that is, statements, each of which is *assumed* (or asserted) to be true; in order for the argument to be sufficient it must be a deductive argument. To be logically sufficient, an argument must satisfy the requirements of what logicians call *modus ponens*. To do so means that *whenever* all of the statements that make up the argument are true, all logically derived statements *must* be true. But

quantificational logic also requires that, for a sufficient deductive argument in favour of some proposition, at least some of the assumptions must be in the form of universal general statements (in the form: ‘all X have property Y’). With these two requirements in mind it should be evident that no purely inductive argument (one consisting *only* of particular statements such as observation reports) can be sufficient. The reason is simply that there is no purely inductive logic that satisfies *modus ponens*; that is, no inductive argument can *guarantee* that whenever all of the statements or assumptions that make up the argument are true that the conclusions will necessarily be true. Philosophers call this the problem of induction. It is a problem because without an inductive logic one cannot prove the truth status of any needed assumption in the form of a universal general statement (for example, ‘*all* firms are profit maximizers’). Friedman's 1953 essay attempts to overcome this key methodological problem. Friedman's method simply dismisses the need to know that one's assumptions are true before deriving one's conclusions. The argument of his essay is that we are explaining given observation statements (for example, statements about the state of the economy) that are known already to be true. This means that the only requirement for any explanatory theory is that it does logically entail the truth of the observation statements – hence it forms a sufficient argument in favour of those observation statements. Moreover, there is no claim that the assumptions of the theory are necessarily true – only that, if they are true, the observed statements would be true. In other words, it is the *sufficiency* of the argument formed by any theory's assumption that matters, not the *necessity* of the theory's assumptions. In this sense, theories are tools or instruments for deriving known true statements. The test of an instrument can be only whether it works or is useful. This view of the *role* of theories is the essence of the doctrine of Instrumentalism. Proponents of Instrumentalism seem to think they have solved the problem of induction by ignoring the truth status of assumptions and thus they also imply that *modus ponens* will be of limited use. This is because Instrumentalist methodology does not begin with a search for the true assumptions but rather for true or useful (that is, successful) conclusions. Instrumentalist analysis of the sufficiency of a set of assumptions always begins by assuming the conclusion is true and then asks what set of assumptions will do the logical job of yielding that conclusion.

The failed critiques

Any valid or fair criticism of an Instrumentalist argument can only be about the argument's sufficiency. As a result, to refute an Instrumentalist argument one must show that the theory in question is insufficient, and thus inapplicable. The failure to recognize the logical requirements of any refutation of Friedman's 1953 methodology led to several failed critiques that nevertheless perpetuated the assumptions controversy. The first prominent shots fired in the assumptions controversy were by Tjalling Koopmans (1957) and Eugene Rotwein (1959), and the last – before the pot was stirred up again by Boland (1979) – was by Louis De Alessi (1971). In between were the critiques by Paul Samuelson (1963), Jack Melitz (1965) and Donald Bear and Daniel Orr (1967). As explained in Boland (1979), none of them dealt fairly or effectively with the Instrumentalism underlying Friedman's methodology as presented in his 1953 article. It should be acknowledged that the title of his article (‘The methodology of positive economics’) can be misleading. However, most misunderstandings are likely the result of his introduction, where he seems to be giving another contribution to the traditional discussions about methodology. Traditional discussions were about issues such as the verifiability or

refutability of truly scientific theories. But Friedman's essay does not do this. Instead, he actually gives an alternative to that type of discussion.

Following traditional discussion, Koopmans sees all theorists seeking to develop or analyse the ‘postulational structure of economic theory’ so as to obtain ‘those implications that are verifiable or otherwise interesting’ (1957, p. 133). Unlike Friedman's essay, which presumes that what one assumes depends on one's purposes, Koopmans presumes all theories are directly analysable independently of their uses. Koopmans's critique of Friedman's essay is based on a restatement of Lionel Robbins's methodological position (1935) which itself seems to be a restatement of what Cairnes argued.

Koopmans's basic concern (but not Friedman's) is the sources of the basic premises or assumptions of economic theory. For the followers of Robbins, the assumptions of economic analysis are promulgated and used *because* they are (obviously) true. The truth of the assumptions is never in doubt. The only question is whether they are necessary for the mathematical derivation of the interesting implications. Koopmans objects to Friedman's dismissal of the problem of clarifying the truth of the premises, the problem that Koopmans wishes to solve using mathematics. Koopmans is an inductivist and as such defines successful explanation as being logically based on inductively and observably true premises. Friedman does not consider assumptions or theories to be the embodiment of truth but only as instruments for the generation of useful (because successful) predictions.

In order to criticize Friedman's argument, Koopmans offers an *interpretation* of his own theory of the logical structure of Friedman's view. His interpretation contradicts Friedman's purpose (that *some*, but not necessarily all, conclusions need to be successful). It is most important to keep in mind that Friedman's methodology is concerned only with the *sufficiency* of a theory's set of assumptions. Koopmans falsely assumes that Friedman's methodology has a concern for *necessity*. In other words, Koopmans's theory of Friedman's methodology is itself void because (by Koopmans's own rules) at least one of its assumptions is false (for more, see Boland, 1979, pp. 515–17).

Many self-proclaimed ‘empiricists’ accept the obviousness of the premises of economic theory. For them, the truth of one's conclusions (or predictions) rests *solely* (and firmly) on the demonstrable truth of the premises – and the presumption that one *must* also justify every claim for the truth of one's conclusions or predictions arrived at by *modus ponens*. Needless to say, such empiricists do not see a problem of induction. Friedman clearly does, and in this sense he is not an orthodox empiricist (despite the term ‘positive’ in his title, which usually means ‘empirical’). According to the empiricist critic Rotwein, Friedman is criticizing views such as his by claiming that they represent ‘a form of naive and misguided empiricism’ (Rotwein, 1959, p. 555). Actually, Rotwein sees the thrust of Friedman's essay as a family dispute among empiricists.

Obviously, there is ‘good’ and ‘bad’ naivety. Good naivety exposes the dishonesty or ignorance of others. But Friedman's essay does not join with the empiricist's pretence that there is an inductive logic, one that would serve as a foundation for Rotwein's verificationist empiricism. Rotwein twists the meaning of ‘validity’ into a matter of probabilities so that he can use something like *modus ponens* (1959, p. 558). But *modus ponens* will not work with statements whose truth status is a matter of probabilities (see Haavelmo, 1944), and thus Friedman is correct in rejecting this approach to empiricism (for more, see Boland, 1979, pp. 517–18).

A more sophisticated critique of Friedman's methodology is the one by Bear and Orr (1967). They criticize only certain aspects while accepting others. In particular, they dismiss Friedman's

Instrumentalism while simultaneously recommending what they call his ‘as if’ principle. Their reason is that they too accept the view that the problem of induction is still unsolved but they see his principle as an adequate means of dealing with that problem. Their main complaint is that Friedman erred by ‘confounding… abstractness and unrealism’ (1967, p. 188, n. 3). Each part of Friedman's argument is, of course, designed only to be *sufficient*, but they ignore this and just claim Friedman's arguments against the *necessity* of testing and against the *necessity* of ‘realism’ of assumptions are both wrong. They go further to claim, ‘all commentators except Friedman seem to agree that the testing of the whole theory (and not just the predictions of theory) is a constructive activity’ (1967, p. 194, n. 15). However, this criticism is unfair because Friedman's concept of testing (as verifying) does not correspond to theirs. Of course, it is not always clear what various writers mean by ‘testing’, mostly because its meaning is too often taken for granted. Where Friedman sees testing only in terms of verification or ‘confirmation’, Bear and Orr appear to adopt Karl Popper's view that a *successful* test is a refutation (Bear and Orr, 1967, pp. 189 ff.). In a similar vein, another critic, Melitz (1965, pp. 48 ff.), seems to be saying that a successful test is confirmation or disconfirmation. In both critiques, the logic of the criticism is an allegation of an inconsistency between *the critic's* concepts of testing and Friedman's rejection of the necessity of testing assumptions. The logic of such criticism may be valid, but in each case the criticism is based on a rejection of Instrumentalism even though it is an absolutely essential part of Friedman's essay. Consequently, the critics are wrong as the alleged inconsistency does not exist *within* Friedman's Instrumentalist methodology. Moreover, it is unfair for critics to assert criticisms only on the basis of an inconsistency between *their* concept of testing and Friedman's methodological judgements which are based on *his* concept (for more, see Boland, 1979, pp. 520–1).

De Alessi (1965; 1971) offered more friendly criticisms. First, he meekly criticizes Friedman for seeing only *two* attributes of theories; a theory can be viewed as a language and as a set of substantive hypotheses. De Alessi says, ‘Unfortunately, Friedman's analysis has proved to be amenable to quite contradictory interpretations’ (1965, p. 477). And, like Koopmans's criticism, it is presumed that Friedman is relying on *modus ponens*. But Instrumentalism, by not requiring true assumptions, cannot use *modus ponens*. So, such a presumption is false.

In his later article, De Alessi says Friedman argues that some assumptions and conclusions are ‘interchangeable’. De Alessi notes that such ‘reversibility’ of an argument allows it to be tautological. Moreover, whenever an argument is tautological, it cannot also be empirical, that is, positive. The logic of De Alessi's argument may be correct – but it is not clear that Friedman was indicating ‘reversibility’ of (entire) arguments with the term ‘interchangeable’. The only methodological point Friedman was making was that the status of a statement's being an ‘assumption’ is not necessarily automatic.

The most celebrated criticism of Friedman's methodology was presented by Samuelson (1963) in his discussion of Ernest Nagel (1963). Samuelson claims that Friedman is in effect saying that a ‘theory is vindicable if (some of) its consequences are empirically valid to a useful degree of approximation; the (empirical) unrealism of the theory “itself”, or of its “assumptions”, is quite irrelevant to its validity and worth’ (1963, p. 232). Samuelson labels this the ‘F-Twist’. And about this he says it is ‘fundamentally wrong in thinking that unrealism in the sense of factual inaccuracy even to a tolerable degree of approximation is anything but a demerit for a theory or hypothesis (or set of hypotheses)’ (1963, p. 233). But Samuelson admits that his characterization of Friedman's view may be ‘inaccurate’ – supposedly why he labelled it the ‘F-Twist’ rather than the ‘Friedman-Twist’. Nevertheless, Samuelson willingly

applies his potentially false assumption in his explanation of Friedman's view. His justification for using a false assumption is Friedman's own ‘as if’ principle. In this way, Samuelson argues that followers of Friedman's methodology must concede defeat if one can discredit or refute Friedman's view by using Friedman's view. Samuelson admits there is ‘cheap humor’ in this line of argument. Nevertheless, he is attempting to criticize Friedman by using Friedman's own methodology. But by Samuelson's own mode of argument, his assumption that attributes the F-Twist to Friedman is false and the attempt to apply this by means of *modus ponens* is thus logically invalid.

Surely it is illogical (and at best pointless) to criticize someone's view with an argument that gives different meanings to the essential terms. But this is just what the prominent critics do. Similarly, using assumptions that are allowed to be false while relying on *modus ponens*, as Samuelson does, is also illogical. Beyond preaching to the choir, an effective criticism must deal properly with Friedman's Instrumentalism. Any criticism that ignores his Instrumentalism will be an irrelevant critique. For this reason, the critiques of Koopmans, Rotwein and De Alessi are clear failures. None of the famous critics was willing to straightforwardly criticize Instrumentalism.

Towards resolving the assumptions controversy

The obvious critique that might succeed is to dispute the success of the observations that Friedman and his followers choose to explain by using his Instrumentalist methodology. For example, it is all too easy to find special cases where maximum dependence on the market can solve social problems. Of course, many people would still not accept Friedman's advocacy of policies involving minimum government if based only on selected examples. But any dispute about Friedman's policy views would open the door to straightforward ideological arguments on the floor of academia. Without this (or at least a critique of the positive claims that are claimed to underlie Friedman's policy views), the controversy will never be decided in favour of Friedman's critics other than to simply recognize – as argued in Boland (1979) – that the only justification for Instrumentalist methodology is a self-serving appeal to Instrumentalism itself. Surely this would be a weak if not dishonest defence.

See Also

- instrumentalism and operationalism

Bibliography

Bear, D.V.T. and Orr, D. 1967. Logic and expediency in economic theorizing. *Journal of Political Economy* 75, 188–96.

Boland, L. 1979. A critique of Friedman's critics. *Journal of Economic Literature* 17, 503–22.

Boland, L. 2003. Methodological criticism vs. ideology and hypocrisy. *Journal of Economic Methodology* 10, 521–6.

De Alessi, L. 1965. Economic theory as a language. *Quarterly Journal of Economics* 79, 472–7.

De Alessi, L. 1971. Reversals of assumptions and implications. *Journal of Political Economy* 79, 867–77.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Haavelmo, T. 1944. Probability approach to econometrics. *Econometrica* 12 (Supplement), 1–118.

Hirsch, A. 1980. The ‘assumptions’ controversy in historical perspective. *Journal of Economic Issues* 14, 99–118.

Koopmans, T. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Melitz, J. 1965. Friedman and Machlup on the significance of testing economic assumptions. *Journal of Political Economy* 73, 37–60.

Nagel, E. 1963. Assumptions in economic theory. *American Economics Review* 53, 211–19.

Robbins, L. 1935. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Rotwein, E. 1959. On the methodology of positive economics. *Quarterly Journal of Economics* 73, 554–75.

Samuelson, P.A. 1963. Problems of methodology: discussion. *American Economic Review, Papers and Proceedings* 53, 231–6.

How to cite this article

Boland, Lawrence A. "assumptions controversy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000231> doi:10.1057/9780230226203.0067

The New Palgrave Dictionary of Economics Online

Attwood, Thomas (1783 – 1856)

B.A. Corry

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Attwood, M.; Attwood, T.; Birmingham School; Chartism; Cobbett, W.; full employment; Mill, J. S.; paper currency; Ricardo, D.

Article

In British social and political history the name of Thomas Attwood is usually connected with the Birmingham Political Union, of which he was a founder, and hence the part that movement played in the peaceful enactment of the great Reform Act of 1832. Later he was also associated with the Chartist movement. However, Attwood also has a place in the history of economic thought as an early exponent of anti-classical monetary and macroeconomic ideas and as the leading member of the so-called Birmingham School.

Thomas Attwood was born in 1783, the son of a banker and into whose profession he followed. From an early age he was also active in public affairs in the City of Birmingham. In 1811 he was elected high Bailiff of that town and the following year, with Richard Spooner (later to be another notable member of the Birmingham School) he represented Birmingham manufacturers' interests against the Orders in Council that had restricted UK trade with the USA and the Continent.

He was first drawn into monetary controversy by the depression that followed the ending of the Napoleonic wars in 1815. Birmingham was then an important manufacturing town and had become the centre of small arms manufacture during the wars. Hence the abrupt reduction in government demand had a quick and sharp effect on the local economy. Attwood was particularly incensed by the cavalier attitude adopted by some orthodox classical economists towards the distress brought about by the post-war depression. Ricardo, for example, expressed little knowledge of it and doubted the claims of Birmingham industrialists. Attwood's first pamphlet – *The Remedy* – appeared anonymously in 1816 and this was followed in 1817, under his own name, by *A Letter to Nicholas Vansittart on the Creation of Money, and its Action upon National Prosperity*.

Those early pamphlets give us the theme that was to dominate all of Thomas Attwood's writings in the field of monetary economics. His prime object was the abolition of the metallic standard and its

replacement with a flexible, managed, currency which, he believed was essential for a full employment policy. Throughout his many subsequent writings he never wavered from this position.

In 1830 Attwood was a founder of the Birmingham Political Union for the Protection of Public Rights: its aim was to secure middle and lower class representation in the House of Commons and the Union played a crucial role in supporting the Grey administration during the passage of the Reform Bill of 1832. In the same year together with Joshua Scholefield he was returned unopposed as a Member of Parliament for the new Parliamentary Borough of Birmingham. He continued to agitate for further Parliamentary reform and in 1839 was a presenter of the mammoth Chartist Petition to Parliament.

His place in the Chartist movement was uneasy and ambiguous. He never endorsed the use of physical force that was advocated by some of the more extreme leaders of the movement. More fundamentally the central tenet of Attwood's monetary proposals – the introduction of an inconvertible paper currency – was utterly rejected by the Chartists who attacked what they termed ‘rag botheration’ (paper currency) as enthusiastically as Cobbett.

Attwood felt, and rightly so, that his monetary ideas were never taken seriously by the establishment and he undoubtedly suffered from what may be termed a persecution complex. He was for example, caricatured by Disraeli in the *Runnymede Letters* and by J.S. Mill in the *Currency Juggle*.

Attwood died in 1856 a disappointed man. Birmingham honoured him with a statue in Stephenson's Place (1859).

His brother Matthias also wrote some important pamphlets in monetary matters but never took up the extreme position of his brother Thomas.

Selected works

1964. *Selected Economic Writings*. Edited with an Introduction by F.W. Fetter. London: LSE Reprints of Scarce Works on Political Economy.

Bibliography

Briggs, A. 1948. Thomas Attwood and the economic background of the Birmingham Political Union. *Cambridge Historical Journal* 9(2), 190–216.

Checkland, S.G. 1948. The Birmingham economists 1815–1850. *Economic History Review*, Second Series 1(1), 1–19.

Corry, B.A. 1962. *Money, Saving and Investment in English Economics*. London: Macmillan.

Wakefield, C.M. 1885. *Life of Thomas Attwood*. Printed privately.

How to cite this article

Corry, B. A. "Attwood, Thomas (1783–1856)." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000159> doi:10.1057/9780230226203.0068

The New Palgrave Dictionary of Economics Online

auctioneer

F. Hahn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The auctioneer is a fictitious agent, introduced by Leon Walras, who matches supply and demand in a market with perfect competition. The process is called ‘tâtonnement’, finding the market clearing price for all commodities, resulting in general equilibrium. No actual trading occurs during this process. The concept of the auctioneer sidesteps the important question of the coordinating power of the price mechanism. There are in fact only a few special cases for which the auctioneer process leads the economy to an equilibrium.

Keywords

auctioneer; auctions; competitive equilibrium; Keynesianism; monopolistic competition; perfect competition; tâtonnement; Walras, L.; Walras's Law

Article

Walras (1874) introduced the idea of a tâtonnement to provide a theoretical account of the formation of equilibrium prices. This account was not meant to be taken descriptively but rather as a ‘Gedanken Experiment’. It was hoped that its study would provide insights into the actual *modus operandi* of the price mechanism.

Consider an economy of *H* households, *F* firms and *n* goods. Let $p \in \Delta \subset R^n_+$, where *p* is a price vector and Δ the simplex. Given the endowments of households $(e^h \in R^n_+)$, $x^h - e^h$ is the net trade vector of household *h* where $x^h \in R^n_+$ is the vector of demand of household *h*. Assume that

$$x^h - e^h = \xi_h(p)$$

where $\xi_h(p)$ is a continuous function from Δ to R^n. Let $y^f \in R^n$ be an activity of firm *f*, where $y_i^f > 0$ is interpreted as 'the firm supplies good *i*' and $y_i^f < 0$ is interpreted as 'the firm demands good *i* as an input'. Let $y = \Sigma_f y^f$ and assume that

$$y = \eta(p)$$

is a continuous function from Δ to R^n. Then define

$$z = \sum (x^h - e^h) - y$$

which by our assumptions can be written as, say

$$z = \sum_h \xi_h(p) - \eta(p) = \theta(p) .$$

It is known that addition of budget constraints implies

$$p \cdot z = 0 \text{all } p \in \Delta.$$

(Walras's Law). An equilibrium of the economy is $p^* \in \Delta$ such that

$$\theta(p^*) \leq 0.$$

It should be added that the net trades $\xi_h(p)$ are assumed to be utility maximizing for each household under the budget constraint:

$$p \cdot \xi_h(p) \leq \sum_f \lambda_{hf}(p \cdot y^f)$$

where $1 \geq \lambda_{hf} \geq 0, \Sigma_h \lambda_{hf} = 1,$ is the share of *h* in the profits of firm *f*. Similarly $\eta^f(p) = y^f$ satisfies for all *f*: $p \cdot \eta^f(p) \geq p \cdot y^f$ all *y* which the firm can choose amongst.

A tâtonnement is now described as follows. A fictitious agent called the *auctioneer* announces $p \in \Delta$. Households now report to this auctioneer their desired net trades $[\xi_h(p)]$ and firms report to him their desired activities $[\eta^f(p)]$. From these reports the auctioneer can deduce θ (*p*). In its light he calculates a new price vector p' as follows:

$$\frac{p_i'}{\Sigma p_i'} = \frac{p_i}{\Sigma p_i} \text{if } \theta_i(p) = 0 \text{ or if } \theta_i(p) < 0 \text{ and } p = 0 \frac{p_i'}{\Sigma p_i'} > \frac{p_i}{\Sigma p_i} \text{if } \theta_i(p) > 0 \frac{p_i'}{\Sigma p_i'} < \frac{p_i}{\Sigma p_i} \text{if } \theta_i(p) < 0.$$

He announces p' agents send back messages which allow him to calculate $\theta(p')$. The process continues until and if the rule for calculating a new price vector yields the preceding price vector. *No actual trading occurs* during this process.

The rule which we have supposed the auctioneer follows in changing his price announcement is only one of a number of possible ones. Indeed, it is not the one proposed by Walras. He supposed the auctioneer to concentrate on one market at a time; specifically he changes only one price. Suppose he changes the *i*th price. Then he changes it until, given all other prices which are held constant, the *i*th market is in equilibrium. (He assumed that there always is such a price and that it is unique.) Thereafter he moves on to the next market. Of course, this process may never terminate in an equilibrium.

In all of this one ought to specify what it is that the auctioneer knows. So far we have assumed that he does not know the function θ (*p*). If, however, he does know this function we may think of the auctioneer as being concerned to find a solution to $\theta(p) \leq 0$ for $p \in \Delta$. He is then no more than a programmer. In this case, for instance, he may adopt Newton's method (Arrow and Hahn, 1971; Smale, 1976). That is he proceeds as follows: Let *J*(*p*) be the $(n-1) \times (n-1)$ Jacobian of the first (*n* – 1) excess demand functions $[\hat{\theta} = \theta_1(p) \ldots, \theta_{n-1}(p)]$. The price of the *n*th good is set identically equal to unity (it is the numeraire). Then define $\hat{p} = (p_1, \ldots, p_{n-1})$ and let $\hat{q} = (q_1, \ldots, q_{n-1})$ solve:

$$\hat{\theta}(\hat{p}) - J(\hat{p})(\hat{q} - \hat{p}) = 0$$

where it is assumed that a solution exists:

$$(\hat{q} - \hat{p}) = J(\hat{p})^{-1}\hat{\theta}(\hat{p}).$$

The auctioneer now follows the rule: raise p_i, if $q_i - p_i > 0$, lower p_i if $q_i - p_i < 0$ if $q_i - p_i < 0$ and $p_i > 0$ and leave p_i unchanged if either $q_i = p_i$ or $q_i < p_i$ and $p_i = 0$. Under certain technical assumptions this way of calculating will lead the auctioneer to an equilibrium (see Arrow and Hahn, 1971).

This example demonstrates that it is possible to think of a tâtonnement as a kind of computer program. If one adopts this view, however, one will certainly not be mimicking the invisible hand. For instance, in the Newton method the price change in any one market depends on the excess demand functions in all markets and that is not what any version of 'the law of supply and demand' stipulates. Moreover the proposal violates the supposed economy in information of decentralized economies – that is, much more is known to the auctioneer than can be known to any one agent. From the point of view of positive theory, therefore, this second interpretation of the auctioneer is not helpful, although it has found application in the theory of planning (e.g. Heal, 1973).

Assuming that the auctioneer only knows aggregate excess demands at the announced p, it has been customary ever since a famous paper by Samuelson (1941; 1942) on Hicksian stability to formulate the rule followed by the auctioneer dynamically. For instance:

$$\frac{\mathrm{d}\,p_i}{\mathrm{d}t} = 0 \qquad \text{if } \theta_i(p) < 0 \text{ and } p_i = 0$$

$$\frac{\mathrm{d}\,p_i}{\mathrm{d}t} = k_t\theta_i(p) \quad \text{otherwise with } k_i > 0\,.$$

Even if this process leads to p^* it will do so only as $t \to \infty$. This is awkward since no one is allowed to trade while the process is still in motion. Some economists have by-passed this by saying that the time here involved is not calendar, but 'model-time'. On reflection it is not clear what that means unless it is 'computer time' which is meant and, if it is, one must again ask whether the construction will then have anything to do with any actual price mechanism.

Arrow (1959) has suggested an alternative interpretation which, however, much restricts the applicability of the tâtonnement. Suppose we think of time as divided into trading periods and let the auctioneer follow the rule:

$$p_i(t) = p_i(t-1) + k_i\theta_i[\,p(t-1)\,]\,k_i > 0$$

(with the usual boundary condition to avoid negative prices). Now suppose (*a*) that one is concerned with a pure exchange economy and (*b*) that all goods last for only one period so that agents in each period receive new endowments (identical for each period). Then we can allow the agents to trade during the process

without the trade in any one period affecting the excess demand at any *p* in a subsequent period. So now (*a*) we think of the process in real time and (*b*) even if it converges to p^* only as $t \to \infty$ or does not converge at all, agents can trade.

This very restrictive case clarifies the reason why in general the tâtonnement prohibits trade out of equilibrium. Let $\hat{e} = (e^1, \ldots, e^H)$, the endowment matrix of a pure exchange economy in which goods are durable. Let us now take explicit note of $\hat{e}$ in the excess demand function (since it was constant it was omitted hitherto) and write

$$\sum_h (x^h - e^h) = \hat{\theta}(p, \hat{e}).$$

Assuming that $\theta(p, \hat{e}) = 0$ has a unique solution, the latter will depend on $\hat{e}$ and may be written as $p^*(\hat{e})$. If now trading takes place out of equilibrium, $\hat{e}$ will be changing and so therefore will $p^*(\hat{e})$. Thus when there is such out of equilibrium trading, the equilibrium which the tâtonnement is groping for will depend on the manner of the groping. To exclude this dependence was the purpose of excluding out of equilibrium trade. But there was another reason, namely, the lack of any clear theory of how trade would proceed when either some prospective buyers or sellers could not carry out their trading intentions.

The fictitious auctioneer is also a consequence of theoretical lacunae and indeed of a certain logical difficulty. If prices are to be changed by the economic agents of the theory, that is either by households or firms or both then it is not easy to see how those same agents are also to treat prices as given exogenously as is required by the postulate of perfect competition. This difficulty was first noted by Arrow (1959) who argued that out of equilibrium price changes not brought about by an auctioneer require a departure from the perfect competition assumption if they are to be understood. Take for instance a situation for which $\hat{\theta}_i(p, \hat{e}) > 0$. Then at *p* there will be unsatisfied buyers. But that means that any firm raising its price for good *i* by a little will not, as in the usual perfect competition setting, lose all its customers. The reason is that buyers cannot be sure of obtaining the good from any of the other firms which have not yet raised their price. Hence the demand curve for good *i* facing a producer of that good is not perfectly elastic. (On the other hand, in equilibrium it well might be.) The postulate of the auctioneer sidesteps these problems at the cost of an understanding of how prices are actually changed. It has enabled theorists to ignore the role of monopolistic competition in the process of price formation – a circumstance which until recently has left the whole matter without proper theoretical foundations.

But it must also be admitted that there are formidable theoretical difficulties to be faced in banishing the auctioneer. Whether we think of prices as formed by a bargaining process or by monopolistic competition or in some form of auction process, strategic considerations, that is to say, game theoretic tools, will be required. In addition, careful attention will have to be given to the information available to each of the agents involved in the process. Some progress has been made (e.g. Roth, 1979; Schmeidler, 1980; Rubinstein and Wolinsky, 1985) but there is a very long way to go. (Some economists have banished the auctioneer without considering these matters by the simple device of treating it as axiomatic that at all times the economy is in competitive equilibrium. There is nothing favourable to be said for this move.)

There is now also a somewhat subtler point to consider: the behaviour postulated for the auctioneer will implicitly define what we are to mean by an equilibrium: that state of affairs when the rules tell the auctioneer

to leave prices where they are. But the auctioneer's pricing rules are not derived from any consideration of the rational actions of agents on which the theory is supposed to rest. Thus the equilibrium notion becomes arbitrary and unfounded. If, on the other hand, we had a theory of price formation based on the rational calculations of rational agents then the equilibrium notion would be a natural corollary of such a theory. For instance, one might then be led to describe a situation in which there is unemployment as one of equilibrium because neither firms nor workers, given their information and beliefs, find it advantageous to change the wage.

This line of reasoning leads one to a central objection to the auctioneer and indeed the tâtonnement: it sidesteps the important question of the coordinating power of the price mechanism. Here is an example. In an oligopolistic industry with excess supply it may not be advantageous for any one firm to reduce its price given its beliefs as to the strategies of its competitors. Yet it may be to all of the firms' advantage to have the price reduced: there is a cooperative solution which dominates the competitive one. Put another way, there are significant externalities in price signalling. To leave these unstudied is to leave very important matters in darkness. The auctioneer is a coordinator *deus ex machina* and hides what is central.

These considerations are most striking in the context of Keynesian theory. As long as the auctioneer is in the picture no state of the economy in which there is involuntary unemployment can qualify as an equilibrium – the auctioneer would be reducing wages. But without the auctioneer the observation that a worker would prefer to work at the going real wage to being idle does not logically entail the proposition that the wage will be reduced. That proposition would require a great deal of further theoretical underpinning turning on the beliefs of workers, the strategies of other workers and the strategies of employers. It would also turn on the information available to agents. For instance, if lowering one's wage is regarded as a signal of lower quality of work then one may be reluctant to offer to work at a lower wage. The fictitious auctioneer makes sure that none of these matters is studied or understood. The use of this fiction encourages the view that all Pareto-improving moves will, in a competitive economy, be undertaken. This view, however, lacks any foundations other than the auctioneer himself.

One might just about convince oneself that, notwithstanding all these objections, the tâtonnement and its auctioneer are worthwhile, if it were the case that it provided one story which showed how equilibrium was brought about. Unfortunately, however, it does not do this for there are only a few special cases for which the auctioneer process leads the economy to an equilibrium. In many others it will not do so. Indeed, in so far as one holds the view that an equilibrium is the normal state of an economy one should not be tempted to understand this circumstance by means of a tâtonnement.

See Also

- tâtonnement and recontracting
- Walras, Léon

Bibliography

Arrow, K.J. 1959. Towards a theory of price adjustment. In *The Allocation of Economic Resources*, ed. M. Abramovitz, et al. Stanford: Stanford University Press, 41–51.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day; Edinburgh: Oliver and Boyd.

Heal, G.M. 1973. *The Theory of Economic Planning*. Amsterdam, London: North-Holland.

Roth, A.E. 1979. *Axiomatic Models of Bargaining*. Lecture Notes in Economics 170. Berlin: Springer-Verlag.

Rubinstein, A. and Wolinsky, A. 1985. Equilibrium in a market with sequential bargaining. *Econometrica* 53, 1133–50.

Samuelson, P.A. 1941; 1942. The stability of equilibrium. *Econometrica* 9, 97–120; 10, 1–25. Reprinted in Samuelson (1966), vol. 1, 539–62, 565–89.

Samuelson, P.A. 1966. In *The Collected Scientific Papers of Paul A. Samuelson*, ed. J.E. Stiglitz. Cambridge, MA: MIT PressIn.

Schmeidler, D. 1980. Walrasian analysis via strategic outcome functions. *Econometrica* 48, 1585–93.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3(2), 107–20.

Walras, L. 1874–7. *Eléments d'économie politique pure*. Definitive edn, Lausanne, 1926. Translated by W. Jaffé as *Elements of Pure Economics*, London: George Allen and Unwin, 1954.

How to cite this article

Hahn, F. "auctioneer." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000161> doi:10.1057/9780230226203.0069

The New Palgrave Dictionary of Economics Online

auctions (applications)

Patrick Bajari
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We survey some recent empirical work concerning the analysis of auctions. We begin by describing a two-step nonparametric approach for estimating bidding models that is commonly used in the applied literature. Two applications of this approach are considered: empirical work on bidding in Treasury markets, and empirical tests for collusion in auctions.

Keywords

auctions; bid rigging; cartels; collusion; conditional independence; discriminatory auctions; exchangeability; first-price auctions; sealed-bid auctions; statistical decision theory; structural estimation; uniform price auctions

Article

In this article, we survey some recently developed methods for the econometric analysis of action data and related applications. Since the mid-1990s, auctions have been an active area of research in empirical industrial organization. Auctions are an attractive setting for empirically testing game theory, for three reasons. First, real-world auctions have well-defined rules, which often correspond closely to game forms in economic theory. The mapping between the data and economic theory is typically less ambiguous in auctions than in other applications in empirical industrial organization. Second, the theoretical literature on auctions is well developed and offers many testable implications. Third, there are many high quality, easily accessible data sets. For example, detailed data sets from public sector procurements or online auctions can easily be collected from the Internet.
In this survey, we shall describe the estimation strategy proposed in Guerre, Perrigne and Vuong (2000) (henceforth GPV) and two substantive applications. The empirical literature in auctions is diverse. Numerous useful alternative approaches have been proposed, so it is impossible to cover all of them in a short survey. However, the work of GPV and related extensions is widely viewed as one of the most

important recent additions to the literature. This survey will omit many of the technical details which are required to correctly implement these estimators. Instead, we discuss the estimators somewhat informally, focusing on what we believe is the key intuition behind these methods. Fortunately, there are several excellent surveys that discuss these estimators and related applications in considerable detail. See, in particular, Athey and Haile (2007), Hendricks and Porter (2007) and Hong and Paarsch (2006).

1 The first-price auction

Following GPV, consider a first-price sealed-bid auction with independent private values. There are i=1, …,N bidders. Bidder i's valuation for winning the auction is denoted by v_i and is private information. The bidders are symmetric in the sense that each bidder's valuation is an i.i.d. draw from a distribution $F(v)$, which is common knowledge. After learning their valuations, each bidder independently and simultaneously submits a bid b_i. Bidders are risk neutral, and bidder i receives utility $v_i - b_i$ if i is the high bidder and zero otherwise. The equilibrium bid function is symmetric and strictly increasing under fairly mild regularity conditions. Let $b=\mathbf{b}(v)$ denote the equilibrium bid function and $\phi(b)=\mathbf{b}^{-1}(v)$ denote the inverse bid function.
Bidder i's expected utility from bidding b_i is equal to

$$(v_i - b_i)F(\varphi(b_i))^{N-1}. \tag{1}$$

Bidder i wins the auction when the other N–1 bidders bid less than b_i. Bidder $j \neq i$ bids less than b_i when j's valuation is less than $\phi(b_i)$. The probability of this event is $F(\phi(b_i))$. Therefore the probability that bidders $j \neq i$ bid less than b_i is $F(\phi(b_i))^{N-1}$. Expected utility is the product of the surplus bidder i receives conditional on winning, $(v_i - b_i)$, times the probability that i wins the auction. Given v_i, the first-order condition for utility maximization is

$$(v_i - b_i)(N-1)f(\varphi(b_i))\varphi'(b_i) - F(\varphi(b_i) = 0. \tag{2}$$

Suppose that the econometrician observes t=1,…,T independent repetitions of the auction described above. For each auction t, the econometrician observes all of the bids $b_{i,t}$. The object that GPV wish to estimate is the distribution of bidder valuations, $F(v)$. GPV's approach is structural in the sense that they attempt to recover the economic primitives of the model. As we shall discuss in our applications, structural estimation of the model may allow the economist to answer a number of substantive questions.

For example, we can assess the efficiency of the observed auction mechanism or test between competing models, such as competition versus collusion.

GPV note that an econometric approach based directly on evaluating eq. (2) may be difficult. This equation involves the inverse bid function, ϕ, and its derivative, ϕ', which in turn are complicated, nonlinear functions of the unknown $F(v)$. In principle, it is possible to estimate parametric auction models based on eq. (2), as in Paarsch (1992), Donald and Paarsch (1993), Hong and Shum (2002) and Bajari and Hortaçsu (2003). However, these methods rely on restricting attention to carefully chosen parametric distributions or require the use of reasonably sophisticated numerical methods. (Despite these limitations, it is worth noting that many parametric approaches generate superconsistent estimators, which converge much more quickly than the nonparametric rate of convergence as in GPV. This may be useful when the sample size available to the econometrician is limited. See Donald and Paarsch, 1993, and Hirano and Porter, 2003, for a discussion.)

A key insight of GPV is that the econometric analysis of the first-price auction is greatly simplified by a change of variables. Let $G(b)=F(\phi(b_i))$ denote the equilibrium distribution of the bids. If we substitute $G(b)$ into (1), we can write expected utility as

$$(v_i - b_i)\,G(b_i)^{N-1}.$$

The first-order conditions now become

$$(v_i - b_i)(N-1)g(b_i) - G(b_i) = 0 \tag{3}$$

$$v_i = b_i + \frac{G(b_i)}{(N-1)g(b_i)}. \tag{4}$$

The right-hand side of eq. (4) involves the bid, b_i, the distribution of the bids, G, and the density of the bids, g. GPV observe that if we have access to a large number of independent repetitions of the same auction, then both G and g can be consistently estimated using standard techniques. Given estimate $\hat{G}$ and $\hat{g}$ of G and g, we can form an estimate $\hat{v}_{i,t}$ of bidder i's private information $v_{i,t}$ in auction t by

evaluating the empirical analogue of eq. (4):

$$\hat{v}_{i,t} = b_{i,t} + \frac{\hat{G}(b_{i,t})}{(N-1)\hat{g}(b_{i,t})}. \quad (5)$$

To summarize, the estimator proposed by GPV is as follows:

1. 1. Given bids $b_{i,t}$ for i=1,…,N and t=1,…,T, estimate the distribution and density of bids $\hat{G}(b)$ and $\hat{g}(b)$.
2. 2. Compute $\hat{v}_{i,t}$ for i=1,…,N and t=1,…,T using eq. (5). Use the empirical cdf of the $\hat{v}_{i,t}$ to estimate F.

This procedure is attractive for three reasons. First, it does not impose parametric assumptions on F during estimation. Since the economist is likely to have poor a priori information about the distribution of values, this is desirable for empirical work. Second, the procedure described above is computationally simple to implement since it does not require evaluation of ϕ and ϕ'. Finally, it is possible to demonstrate that $F(v)$ is nonparametrically identified. The intuition is quite simple. As T grows arbitrarily large, the economist will be able to estimate G and g very precisely under standard regularity conditions. Equation (4) implies that for any given bid b_i we can recover the latent valuation v_i that generates this bid (that is, $v_i=\phi(b_i)$). Since the distribution of b_i is known, it can easily be demonstrated that $F(v)$ is therefore identified.

GPV also demonstrate that the first-price auction model can be tested. Given estimates $\hat{G}$ and $\hat{g}$, define $\xi(b)$ as

$$\xi(b) = b + \frac{\hat{G}(b)}{(N-1)\hat{g}(b)}.$$

Theoretical models of bidding imply that the bid function should be increasing, that is, bidders with higher valuations should submit higher bids. Therefore, if $\hat{G}(b)/(N-1)\hat{g}(b)$ is sufficiently close to $G(b)/(N-1)g(b)$, $\xi(b)$ should be monotonically increasing if the model is correctly specified. This prediction of the theory could be rejected by the data since $\hat{G}$ and $\hat{g}$ are estimated nonparametrically and do not impose a priori that $\xi(b)$ is increasing.

2 Generalizations and applications

Following GPV, a large number of authors have proposed similar estimators for other auction models. In these papers, a key step is typically to rewrite the first-order conditions in terms of the equilibrium distribution of the bids (for example *G* and *g*). Next, as in eq. (4), the economist attempts to isolate private information on the left-hand side as a function of the bids on the right-hand side. Following GPV, the economist then nonparametrically estimates the distribution of the bids from the data and recovers the latent private information by evaluating the empirical analogue of the first-order condition. This basic algorithm often needs to be modified for different auctions. However, attempting to follow these steps as a first pass will typically take the economist a long way towards deriving an estimator. Listed in Table 1, in alphabetical order, are some recent papers which build on the insights of GPV in other auction models.

Related papers

Paper	Topic
Athey and Haile (2002)	Identification in auctions
Bajari and Ye (2003)	First-price auctions with collusion
Brendstrup and Paarsch (2003)	Dutch and first-price auctions with asymmetric bidders
Campo et al. (2002)	Auctions with risk aversion
Campo, Perrigne and Vuong (2003)	Asymmetric first-price auctions with affiliated values
Flambard and Perrigne (2006)	Asymmetric first-price auctions
Hendricks, Pinkse and Porter (2003)	Common value auction models
Hortaçsu (2002)	Treasury auctions
Li and Perrigne (2003)	Random reserve prices
Li, Perrigne and Vuong (2002)	Affiliated private values
Pesendorfer and Jofre-Bonet (2003)	Dynamic first-price auctions

Next, in order to illustrate how these techniques are used in practice, we briefly summarize Hortaçsu (2002) who analyses bidding in Treasury bill auctions, and Bajari and Ye (2003) who test for collusion in procurement auctions.

2.1 Auctions for Treasury bills

Hortaçsu (2002) asks how governments should conduct auctions for Treasury bills. Treasury bill auctions are an example of a multiple unit auction since large numbers of T-bills are typically sold during a single auction. Since there are multiple units, a ‘bid’ in a Treasury auction is a demand curve, instead of a scalar as in the example of Section 1. Two commonly used mechanisms for conducting a Treasury bill auction are the uniform price auction and the discriminatory auction. In a uniform price auction, the auctioneer begins by aggregating all of the individual demand curves into a market demand

curve. The supply curve is vertical, with an intercept equal to the number of T-bills that the government wishes to sell. The market-clearing price is determined by the intersection of the supply and demand curve. Each bidder pays his demanded quantity at the market-clearing price, analogous to a competitive market. By contrast, in a discriminatory auction, the intersection of the supply and market demand curves determines the price for the last unit purchased. Analogous to first-degree price discrimination, bidder *i* pays the area under his demand curve, so that the price for the first unit purchased will be higher than for the last unit purchased.

There is no general consensus about which auction mechanism should be preferred. Since the equilibria to these auctions are quite complicated, it is difficult to characterize revenue in each auction. Each year, nearly $4 trillion dollars of securities are sold in T-bill auctions. Given the size of these markets, econometrically modelling the determination of the bids and comparing revenue from alternative auction mechanisms is an interesting public policy question.

The particular market that Hortaçsu examines is the short-term (13-week) market for T-bills in Turkey. This market is run using a discriminatory auction. Hortaçsu uses the Wilson (1979) auction of shares model as a starting point for his econometric analysis. He assumes that bidders have private values. According to surveys of bidders, 42 per cent of purchases in the auctions are to meet reserve requirements imposed by the Turkish Central bank. Thirty-seven per cent of purchases are for resale in the secondary market. Ten per cent are to fulfil customer orders and ten per cent are to fulfil collateral requirements, for investment funds administered by the bank, and for buy-and-hold purposes. Other than those shares purchased for resale, the other sources of demand are probably best modelled as private values.

Let s_i denote bidder *i*'s private information about her willingness to pay for government debt and $v_i(q,\bullet s_i)$ denote bidder *i*'s valuation for the *q*th unit. Assume that private information is distributed i.i.d. $s_i \sim F(s)$. Let $y_i(p)$ denote the demand curve submitted by bidder *i*. Hortaçsu assumes that $y_i(p)$ is strictly decreasing and differentiable. If there are *N* bidders and *Q* units of debt for sale, the market-clearing price p^c will satisfy

$$Q = \sum_i y_i(p^c).$$

The cdf of the market-clearing price, conditional on *i*'s bid function $y_i(p)$ is

$$H(p, y_i(p)) = \Pr\left\{ y_i(p) \le Q - \sum_{j \ne i} y_j(p) \right\} = \Pr\left\{ p^c \le p \mid y_i(p) \right\}. \tag{6}$$

Equation (6) is analogous to a residual supply curve. The term $H(p,y_i(p))$ is the probability that the market-clearing price will be less than p given i's own bid, $y_i(p)$. However, unlike a residual supply curve in a model with certainty, the bidder has to take into account her uncertainty about the bids of others.

Given a bid $y_i(p)$, the surplus that a bidder gets, conditional on p^c is equal to

$$\int_0^{y_i(p^c)} v_i(q, s_i)\mathrm{d}q - \int_0^{y_i(p^c)} y_i^{-1}(q)\mathrm{d}q.$$

There are two terms in the above sum. The first term is the integral of $v_i(q, s_i)$ from 0 to $y_i(p^c)$. This is bidder i's valuation for the units that she wins. The second term is the integral of i's inverse demand curve. This determines the total payment that i just made for the units that she won. Therefore, i's expected profit from submitting a bid of $y_i(p)$ is equal to

$$\int_0^{\infty}\left\{\int_0^{y_i(p^c)}\left\{v_i(q, s_i) - y_i^{-1}(q)\right\}\mathrm{d}q\right\}\mathrm{d}H(p^c, y_i(p)).$$

Following Wilson (1979), the first-order condition for maximization implies that

$$v_i(y_i(p), s_i) = p + \frac{H(p, y_i(p))}{\frac{\partial}{\partial p}H(p, y_i(p))}. \tag{7}$$

That is, a bidder's valuation will be equal to the price on the submitted demand curve plus a bid-shading factor, $H(p, y_i(p)) / (\partial / \partial p)H(p, y_i(p))$. Just as in the first-price auction example in Section 1, Hortaçsu notes that $H(p,y_i(p))$ is the cdf of the equilibrium distribution of bids given $y_i(p)$. Given a large number of repetitions of the same, or similar auctions, this object can be estimated from the observed bidding data. And, similar to the first-price auction example above, an estimate of bidder i's valuation, $v_i(y_i(p),s_i)$ can be recovered by evaluating the empirical analogue of eq. (7). While the econometric details are somewhat involved, a key economic insight was expressing the first-order conditions in terms of a function of the bids which, in principle, can be recovered from the data.

Using his estimates of bidder valuations, Hortaçsu examines two applied questions. The first is to

explore the impact of reserve requirements on bidding behaviour. He constructs a variable, $\%SHORTFALL_{i,t-1}$, which is the fraction of orders in the previous Treasury auction that were unfulfilled. He finds that when bidders have a large shortfall in previous auctions, they are more likely to bid aggressively in upcoming auctions. Using his survey on bidder demands, he interprets this as derived demand from satisfying reserve requirements to hold a required portfolio of Turkish Treasury notes. For instance, he finds that the R^2 of a regression of the intercept of the submitted bid function on $\%SHORTFALL_{i,t-1}$, $\%SHORTFALL_{i,t-2}$ and an auction fixed effect is 0.61. Bidder-fixed effects only increase R^2 to 0.64.

A second applied question Hortaçsu examines is whether a uniform price auction would generate increased revenue. This is complicated to answer since changing to a uniform price auction would generate an entirely new equilibrium in this market. However, Hortaçsu demonstrates that it is possible to construct a simple upper bound on revenue given estimates of $v_i(q,\bullet s_i)$ for $i=1,\ldots,N$. Since bidders typically engage in demand reduction in a uniform price auction, they will bid at most $v_i(q,\bullet s_i)$ so that $v_i(q,\bullet s_i)$ is an upper bound on i's bid. Assuming that this upper bound is binding for all bidders, he generates an upper bound on the market-clearing price in the auction. Using his structural estimates, Hortaçsu finds that switching to a uniform price auction would generate a revenue loss of at least 3.8 per cent on average in the auctions in his sample.

Hortaçsu therefore argues that the discriminatory price auction generates higher revenue since bidders are being forced to pay the area under their demand curves. Even after accounting for changes in the strategic incentives to shade bids, discriminatory auctions generate more revenue. However, this conclusion is subtle. Recall that bids are the steepest when shortfalls are the highest. It is hard to argue that forcing banks to hold Turkish Treasury debt is optimal for securing deposits. More likely, this policy was implemented in order to guarantee that there is a constant demand for government debt even if the government engages in irresponsible fiscal or monetary policies. These results suggest that the reserve requirements plus the discriminatory mechanism may be imposing a burden on the banking sector by forcing banks to hold more than the optimal number of domestic T-bills.

2.2 Collusion application

Next, we briefly discuss an application by Bajari and Ye (2003) that tests for collusive bidding behaviour in procurement auctions. Bid rigging is an important antitrust problem. For instance, Pesendorfer (2000) notes that 55 per cent of the criminal antitrust cases filed by the US Department of Justice involved bid rigging. One well-known example of bid rigging was the ‘concrete club’ in New York where organized crime figures placed an implicit ‘tax’ of two per cent on every ton of concrete used in certain construction jobs in the 1980s. However, the costs of collusion were likely much larger than two per cent. Mafia informer Sammy ‘The Bull’ Gravano, who was involved in bid-rigging in the concrete industry, stated ‘If one of them (contractor) gets a contract for, say, thirteen million, the next thing you know, after he knows he's got it, he jacks up the whole thing before it's over to a sixteen- or seventeen-million-dollar job. Now he's increased the cost thirty-three percent. So our greed (the Mafia) is compounded by the greed of them so-called legitimate guys (contractors)’ (Maas, 1997, p. 271).

While bid-rigging is an important antitrust problem, it can be difficult to detect. Bajari and Ye (2003),

expanding on the methods in Section 1, and on the work of Porter and Zona (1993; 1999), propose three statistical tests that can be used to potentially detect bid rigging in procurement auctions. Certainly, no test for bid-rigging can hope to be foolproof. However, it may be a basis for determining which sets of bids are most worrisome and whether further investigation of certain firms is warranted.

Bajari and Ye apply their methods to a set of contracts in the highway construction industry for ‘seal coating’ jobs in Minnesota, North Dakota and South Dakota. Seal coating is a type of highway repair that attempts to extend the life of the road by sealing surface cracks. The surface of the highway is initially sprayed with a coating of oil. Next, a ‘chip spreader’ distributes a uniform layer of sand and aggregate on the road. Finally, rollers are used to bind the oil, sand and aggregate. Bidding is conducted using sealed bids. While there are a large number of fringe firms in the industry, the market is dominated by a few large bidders that regularly compete against each other. Since all of the bids are publicly available shortly after they are submitted, collusion has occurred in seal coating in many markets. Bajari and Ye note that three of the largest bidders in their data have been fined for previous attempts to rig bids. The owner of the largest firm in the data set served prison time for a bid rigging conviction.

Bajari and Ye consider a first-price auction model similar to the example discussed in Section 1. However, they drop the assumption that all bidders are *ex ante* identical. In the construction industry, they argue it is important to allow for asymmetric bidders for three reasons. First, transportation costs are substantial in this market so that firms located closest to the project will tend to have lower cost. Second, there is a skewed size distribution of firms in the industry. Therefore, it is important to allow for firm specific difference in productivity. Third, project backlog increases the opportunity cost of taking on additional work and is likely therefore to be an additional source of *ex ante* asymmetries.

In the model, N firms compete for a contract to build a single and indivisible public works project. Firm i's cost to complete the project, c_i, is a random variable with cumulative distribution function $F_i(\cdot : z_i; \theta_i)$ and probability density function $f_i(\cdot : z_i; \theta_i)$. Here z_i reflects publicly observed cost shifters from firm i. For instance, in the application, these include distance to the project, a firm fixed effect to capture differences in productivity, backlog at the time bids are submitted and an engineering cost estimate. The term θ_i is a set of firm specific parameters. In the model, firm i is risk neutral and has profits of $b_i - c_i$ if it is the low bidder and zero otherwise.

Let $G_i(b;z)$ be the equilibrium distribution of bids submitted by firm i. Note that the distribution of the bids depends on $z=(z_1,\ldots,z_N)$, the publicly observed information for all firms in the industry. Then i's expected profits from submitting a bid of b_i when i's costs are c_i is equal to

$$(b_i - c_i)\prod_{j \neq i}(1 - G_j(b_i; z)) \tag{8}$$

It can easily be shown that the first-order condition to the model must satisfy

$$c_i = b_i - \left[\sum_{j \neq i} \frac{g_j(b_i; z)}{1 - G_j(b_i; z)} \right]. \tag{9}$$

As in Section 1, if the economist has estimates of $\hat{G}_i$ and $\hat{g}_i$, it is possible to generate an estimate of c_i by evaluating the empirical analogue of the above equation for all bidders in the sample.

Bajari and Ye (2003) propose three tests for collusive bidding. We next describe the basic spirit of these tests, referring the interested reader to the text for complete details. The first test for competitive bidding is that conditional on z, the bids of all firms I=1,…,N must be distributed independently. This is a fairly robust prediction of the theory of competitive bidding and is in fact more general than the particular model described above. Because bidders have private information which is independently distributed, their bids, which are a deterministic function of this private information, must also be independently distributed. Obviously, one limitation of such a test is if some component of z is observed by the firms, but not by the econometrician. Following Porter and Zona (1993; 1999), their estimation strategy allows for the inclusion of an auction-specific fixed effect. Thus, they control for project specific cost shifters which are common to all of the firms.

Second, they demonstrate that the equilibrium distribution of competitive bids must be exchangeable. Let π be a permutation of the bidder identities {1,…,N}, that is, a one-to-one map from {1,…,N} to {1, …,N}. If the equilibrium bid function is unique, the bid distribution must be exchangeable: that is, $G_i(b; z_1, z_2, z_3, \ldots, z_N) = G_{\pi(i)}(b; z_{\pi(1)}, z_{\pi(2)}, z_{\pi(3)}, \ldots, z_{\pi(N)})$. In words, exchangeability means that if you permute the cost shifters of all the bidders, then the equilibrium bids must also permute in a symmetric fashion. Conditional independence and exchangeability are necessary for equilibrium bidding. If other regularity conditions hold, conditional independence and exchangeability are also sufficient for competitive bidding: that is, the economist can reverse engineer a competitive bidding model that rationalizes the observed bids.

Porter and Zona (1993, 1999) study the bidding behaviour of known cartels in construction and in the supply of school milk. Many of the irregular patterns of bidding that they describe can be characterized as failures of conditional independence and exchangeability. For instance, the bids of cartel members are more correlated with each other than with non-cartel members. Also, cartel members do not shift their bids aggressively in response to shifts in the z_i of other cartel members which is a failure of exchangeability.

Bajari and Ye (2003) test for conditional independence and exchangeability in their data set. Given the limited number of observations available to them, they test these conditions in a regression framework. Essentially, they run a regression of b_i on z_i and z_{-i}, including auction fixed effects and bidder fixed effects. Conditional independence is tested by asking whether the fitted residuals from bidder i's bid function is correlated with the fitted residuals from $j \neq i$'s bid function. Exchangeability is formulated as a test of the equality of certain regression coefficients. In total, 46 separate hypothesis tests are conducted. Forty-one of these tests are consistent with the implications of competitive bidding (that is, conditional independence and exchangeability). Therefore, they argue that most of the bids in the market

appear to be competitive. However, reduced form tests suggest that bidding by two coalitions of firms appear to be suspicious. They label these coalitions ‘candidate cartels’. Interestingly, all of the members of the candidate cartels had previously been convicted of bid rigging.

The third and final test for bid rigging uses structural estimates based on eq. (9). Bajari and Ye consider a non-nested hypothesis test between three models. Model M1 is that the data-generating process is the no collusion model. Model M2 is that the first candidate cartel is engaged in efficient collusion, but that other firms in the industry are competitive. Model M3 is that the second candidate cartel engages in bid rigging. The costs c_i can be estimated under each of these three alternatives using the empirical analogue of eq. (9). The different models generate different first-order conditions and hence, different estimated costs, c_i.

Bajari and Ye then ask which set of markups is ‘most reasonable’. To answer this question, they consulted with two managers at one of the biggest firms in this market (which was not in a candidate cartel). From each manager, they elicited their beliefs about the distribution of markups in this industry. Bajari and Ye argue that it is reasonable to suppose that these managers have informative priors about markups for two reasons. First, all bidders in this industry must be bonded. The bonding companies are contractually liable to complete the project if the contractors go bankrupt. Contractors are typically required to give weekly profit and loss statements to the bonding companies. The bonding companies are therefore well informed about profit margins for firms in the industry. Profit margins in the industry are a common topic of conversation between contractors and bonding companies and are one source of information.

Second, the contractors in this industry compete against each other quite frequently and over many years. The contractors have access to similar cost information and study the bids of competing contractors in detail after the bids are publicly opened. Given that contractors closely follow cost conditions and bids in the industry, they will have a lot of information about their competitors’ markups. There is an issue, of course, about whether the contractors would lie about their beliefs. However, Bajari and Ye shared their estimates with the contractor, which included empirical analysis of the behaviour of competing firms. Lying about the industry would reduce the value of these estimates. Also, the information from the contractor that was verifiable from external sources about the industry did seem to be accurately reported.

The stated beliefs of the experts were quite close. Below, we average the elicited beliefs from the contractors:

$$\text{25th percentile} = 3\%\ \text{50th percentile} = 5\%\ \text{75th percentile} = 7\%\ \text{99th percentile} = 15\%. \tag{10}$$

For example, the 25th percentile of the bids has a markup of three per cent and the median bid has a markup of five per cent.

Table 2 shows the estimated distribution of markups from the three alternative structural models, M1, M2 and M3.

Distribution of markups under

	alternative models		
Percentile	M1	M2	M3
10	0.01229	0.01273	0.0114
20	0.01597	0.01818	0.0182
30	0.02077	0.02422	0.0256
40	0.02536	0.03201	0.0343
50	0.03329	0.04126	0.0447
60	0.04227	0.05434	0.0584
70	0.05692	0.0754	0.0930
80	0.1000	0.1621	0.1756
90	0.2381	0.3354	0.5826

Bajari and Ye note that the markups under M1 (competitive bidding) correspond most closely to the elicited prior beliefs. The markups under models M2 and M3 seem to be too large, particularly on the tails. They argue that this is evidence against the collusive models since they generate markups that seem implausibly large compared to the beliefs of an informed party. Bajari and Ye formalize this intuition by posing the selection of M1, M2 or M3 as a problem in statistical decision theory. As the table above suggests, the competitive model M1 is most favoured. Therefore, they cautiously interpret the data as being consistent overall with non-collusive behaviour.

3 Conclusion

In this short survey, I have attempted to provide an overview of recent empirical papers concerning auctions. Many recent papers build on the pioneering work of Guerre, Perrigne and Vuong (2000). A key insight of this paper was that a first-price sealed-bid auction model can be simply estimated using a two-step procedure. In the first step, the economist flexibly estimates the empirical distribution of the bids. In the second step, the economist evaluates the empirical analogue of the first-order condition for utility maximization. The method of Guerre, Perrigne and Vuong estimates the structural primitives of the model without imposing ad hoc parametric restrictions. We also discussed two applications of these recently developed estimators. Hortaçsu (2002) studied bidding in Treasury auctions in Turkey. His model predicted that discriminatory auctions generate higher revenue than uniform price auctions. Bajari and Ye (2003) applied these methods to test for collusion in sealed-bid auctions. They applied these methods to searching for suspicious bidding patterns in a market where the largest firms had recently been sanctioned for collusion.

See Also

- auctions (empirics)
- auctions (experiments)
- auctions (theory)

- cartels
- epistemic game theory: incomplete information
- game theory
- nonparametric structural models

Bibliography

Athey, S. and Haile, P. 2002. Identification of standard auction models. *Econometrica* 70, 2107–40.

Athey, S. and Haile, P. 2007. Nonparametric approaches to auctions. In *Handbook of Econometrics*, vol. 6, ed. J.J. Heckman and E.E. Leamer. Amsterdam: North-Holland.

Bajari, P. and Hortaçsu, A. 2003. The winner's curse, reserve prices, and endogenous entry: empirical insights from eBay auctions. *RAND Journal of Economics* 34, 329–55.

Bajari, P. and Ye, L. 2003. Deciding between competition and collusion. *Review of Economics and Statistics* 85, 971–89.

Brendstrup, B. and Paarsch, H. 2003. Nonparametric estimation of Dutch and first-price, sealed-bid auction models with asymmetric bidders. Working paper, University of Iowa.

Campo, S.E., Guerre, I., Perrigne and Vuong, Q. 2002. Semiparametric estimation of first-price auctions with risk averse bidders. Working paper, Pennsylvania State University.

Campo, S., Perrigne, I. and Vuong, Q. 2003. Asymmetry in first-price auctions with affiliated private values. *Journal of Applied Econometrics* 18, 197–207.

Donald, S. and Paarsch, H. 1993. Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *International Economic Review* 34, 121–48.

Flambard, V. and Perrigne, I. 2006. Asymmetry in procurement auctions: evidence from snow removal contracts. *Economic Journal* 116, 1014–36.

Guerre, E., Perrigne, I. and Vuong, Q. 2000. Optimal nonparametric estimation of first-price auctions. *Econometrica* 68, 525–74.

Hendricks, K., Pinkse, J. and Porter, R. 2003. Empirical implications of equilibrium bidding in first-price, symmetric, common value auctions. *Review of Economic Studies* 70, 115–45.

Hendricks, K. and Porter, R. 2007. Lectures on auctions: an empirical perspective. In *Handbook of Industrial Organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: Elsevier.

Hirano, K. and Porter, J. 2003. Asymptotic efficiency in parametric structural models with parameter dependent support. *Econometrica* 71, 1307–38.

Hong, H. and Paarsch, H.J. 2006. *An Introduction to the Structural Econometrics of Auction Data*. Cambridge, MA: MIT Press.

Hong, H. and Shum, M. 2002. Increasing competition and the winner's curse: evidence from procurement. *Review of Economic Studies* 69, 871–98.

Hortaçsu, A. 2002. Mechanism choice and strategic bidding in divisible good auctions: an empirical analysis of the Turkish treasury auction market. Working paper, University of Chicago.

Li, T. and Perrigne, I. 2003. Timber sale auctions with a random reserve price. *Review of Economics and Statistics* 85, 189–200.

Li, T., Perrigne, I. and Vuong, Q. 2002. Structural estimation of the affiliated private value auction model. *RAND Journal of Economics* 33, 171–93.

Maas, P. 1997. *Underboss: Sammy the Bull Gravano's Story of Life in the Mafia*. New York: HarperCollins.

Paarsch, H. 1992. Deciding between common and private values paradigms in empirical models of auctions. *Journal of Econometrics* 51, 191–215.

Pesendorfer, M. 2000. A study of collusion in first-price auctions. *Review of Economic Studies* 67, 381–411.

Pesendorfer, M. and Jofre-Bonet, M. 2003. Estimation of a dynamic auction game. *Econometrica* 71, 1443–89.

Porter, R. and Zona, J.D. 1993. Detection of bid rigging in procurement auctions. *Journal of Political Economy* 101, 518–38.

Porter, R. and Zona, J.D. 1999. Ohio school milk markets: an analysis of bidding. *RAND Journal of Economics* 30, 263–88.

Wilson, R. 1979. Auctions of shares. *Quarterly Journal of Economics* 93, 675–89.

How to cite this article

Bajari, Patrick. "auctions (applications)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000216> doi:10.1057/9780230226203.0070

The New Palgrave Dictionary of Economics Online

auctions (empirics)

Isabelle Perrigne and Quang Vuong

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The structural analysis of auction data relying on game theoretic models has undergone a tremendous development since the mid-1990s. This article reviews some important contributions for first-price and ascending auctions. It stresses identification of the structure and the development of tractable econometric methods, while addressing bidders' asymmetry, common value, bidders' risk aversion, endogenous entry, dynamic and multi-unit auctions as well as the choice of the reserve price and the auction mechanism. Various domains are studied, such as auctions of timber, gas lease, treasury bills, agricultural products, electricity and construction procurements.

Keywords

affiliated private value model; asymmetric information; auctions; Bayesian Nash equilibrium; bidding; collusion; game theory; log-normal distribution; maximum likelihood; maximum likelihood; nonlinear least squares; nonparametric estimation; nonparametric methods; reserve price; risk aversion

Article

Auctions and procurements are widely used market mechanisms for allocating public contracts, financial securities, agricultural products, natural resources, artwork, and electricity, to name a few commodities. Recent years have also witnessed the developments of auction websites and business-to-business auctions. In general, auctions have well-defined rules that can be captured by an economic model. Relying on the concept of the Bayesian Nash equilibrium, game theory has greatly contributed to the modelling of auctions, where a seller or buyer faces a limited number of bidders who behave strategically. The auction is typically an incomplete information game where the asymmetry of information between the seller/buyer and the bidders and among the bidders themselves plays a crucial role.
While auctions are largely used in economic life and data are rich and accessible, until recently the

empirical analysis of auction data has been confined to testing some predictions generated by game theoretic models. One influential example of the reduced form approach is the work by Porter and his coauthors on the role of private information in oil and gas auctions, as surveyed in Porter (1995). This approach has also been used to test for collusive behaviour in timber and milk auctions. Although important, this approach does not allow for policy evaluations that require knowledge of the informational structure of the game such as the choice of the reserve price and the auction mechanism that would generate greater revenue for the seller/buyer.

The structural approach addresses such questions by assuming that observed bids are the equilibrium bids of some auction model. Specifically, $b_i = s_i(v_i)$ where b_i and v_i are bidder's *i* (observed) bid and (unobserved) private information, respectively, and $s_i(\cdot)$ is bidder's *i* equilibrium strategy in the corresponding auction game. Bidders' private information is assumed to be derived from some distribution that is common knowledge to all bidders. This distribution and the bidders' preferences are the key elements that explain bidding behaviour. They are the structural elements of the induced econometric model for the observed bids. The structural approach then exploits the equilibrium relations $b_i = s_i(v_i)$ to recover bidders' private information, which can be exploited for policy purposes. A major difficulty in implementing this approach arises from the numerical complexity or the implicit form of the equilibrium strategies. Of its nature, the structural approach raises challenging questions. One question is related to identification, namely, whether the auction structure can be uniquely recovered from observables while minimizing parametric restrictions. This question relates to whether auction models can be distinguished from observables. A second question concerns the model validity, namely, whether an auction model imposes testable restrictions on observables. A third difficulty is to develop tractable estimation methods. Since ascending (English) auctions and first-price sealed-bid auctions involve different equilibrium strategies and different identification and estimation problems, they are treated separately.

Econometrics of first-price auctions and applications

Two kinds of methods can be distinguished. Direct methods start from a parameterization of the private information distribution $F(\cdot)$ and sometimes require the computation of equilibrium strategies. Indirect methods exploit the first-order condition(s) to estimate $F(\cdot)$ from the observed bid distribution without computing the equilibrium strategies. Direct methods require explicit forms for the equilibrium strategies, while indirect methods can be considered when no explicit form exists. The structural approach was initiated by Paarsch (1992) using a direct method to analyse tree planting contract auctions with symmetric bidders. If the latent distribution is parameterized as $F(\cdot;\theta)$, then $b_i = s(v_i;\theta)$, which is distributed as $G(\cdot;\theta) = F[s^{-1}(\cdot;\theta);\theta]$. This raises two difficulties. First, a limited number of distributions lead to tractable equilibrium strategies. Second, the standard regularity conditions of maximum likelihood (ML) estimation are violated because the bid distribution support depends on θ. Paarsch and coauthors have extended ML estimation to this problem. Laffont, Ossard and Vuong (1995) propose an alternative direct method based on simulations while analysing Dutch auctions of vegetables. This method allows a large family of distributions to be entertained. It exploits the revenue equivalence theorem for independent private value models to write the expectation $m_\ell(\theta)$ of the winning bid in a Dutch auction as the expectation of the second highest value. The authors develop a simulated nonlinear

least squares estimator based on minimizing $Q_L(\theta) = (1/L)\Sigma_{\ell=1}^{L}[b_\ell^w - m_\ell(\theta)]$, where $m_\ell(\theta)$ is replaced by a simulator, *L* is the number of auctions and b^w is the winning bid, while correcting for its inconsistency. This idea has been extended by others when the expected winning bid can be simulated. This limits the number of models to be considered. Bayesian estimation methods, though computationally demanding, have also been developed.

In contrast, the indirect method initiated by Guerre, Perrigne and Vuong (2000) requires neither the computation nor the simulation of equilibrium strategies. It uses the differential equation (s) or first-order condition(s) to express each private value as a function of its corresponding bid. Within the symmetric independent private value paradigm, the differential equation is $s'(v_i) = [v_i - s(v_i)](I-1)[f(v_i)/F(v_i)]$, where *I* is the (known) number of bidders, $s'(\cdot)$ is the derivative of $s(\cdot)$ and $f(\cdot)$ is the private value density. Because $b_i = s(v_i)$, bids are also i.i.d. with $G(b) = F[s^{-1}(b)] = F(v)$ leading to $g(b) = f(v)/s'(v)$. Hence, the differential equation can be written as

$$v_i = b_i + \frac{1}{I-1}\frac{G(b_i)}{g(b_i)} \equiv \xi(b_i, G, I). \tag{1}$$

Relying on (1), the authors show that the model is nonparametrically identified: that is, one can recover uniquely the distribution $F(\cdot)$ from the observed bid distribution without parametric restrictions. Moreover, they derive the restrictions imposed by the model on observables: that is, bids must be i.i.d. (since private values are i.i.d.) and $\xi(\cdot)$ should be strictly increasing (since $s(\cdot)$ is strictly increasing). These two restrictions can be used to test the validity of the model. Equation (1) calls for a two-step estimation procedure. The first step consists in estimating nonparametrically $G(\cdot)$ and $g(\cdot)$, while the second step estimates nonparametrically $f(\cdot)$ from the estimated private values $\hat{v}_i$ using (1). In practice, auctioned goods are heterogeneous. Observed characteristics can be introduced in the econometric model by writing (1) with conditional bid distribution and density. Nonparametric estimation can be a drawback when a limited number of auctions is available and/or when the number of exogenous variables is relatively large. It can, however, provide a preliminary estimate of the underlying density, which can be used later to specify $F(\cdot)$ when using a parametric two-step estimation procedure.

In addition to not parameterizing $F(\cdot)$, the indirect method does not require an explicit form for the equilibrium strategy, as it relies on the first-order condition(s). The method provides key insights on questions at the core of the structural approach, as discussed above. It can be easily extended to the case of a binding reserve price, where the number of actual (observed) bidders is smaller than the number *I* of potential bidders as only bidders with private values above the reserve price effectively participate. Alternatively, the seller may not announce his reserve price, keeping it secret as in timber and wine auctions. Although the equilibrium strategy in such a model does not have an explicit form, the above method allows a simple expression to be obtained for the inverse equilibrium strategy, which can be

used to develop a two-step estimation procedure as above. Likewise, the method can be easily extended to situations in which only the winning bids are observed, as in Dutch auctions, which are widely used for agricultural products such as vegetables and flowers.

Independence among private values can be restrictive. One can expect some affiliation or positive correlation among private values and some common value ν affecting all bidders' utilities, that is, bidder's *i* utility becomes $v_i = U(\sigma_i, \nu)$. In the private value paradigm $v_i = \sigma_i$, while in the pure common value paradigm $v_i = \nu$. The vector $(\sigma_1, \ldots, \sigma_I, \nu)$ is distributed as $F(\cdot, \ldots, \cdot)$, which is affiliated and exchangeable in its first *I* arguments under bidders' symmetry. Affiliation means that, if one bidder values the auctioned object highly, other bidders are also likely to value it highly. In the common-value model bidders receive signals about the value of the object, which is unknown at the time of the auction. This model has been widely used to explain bidding behaviour in gas lease auctions where firms have imperfect information about the amount of oil. The general framework is considered by Laffont and Vuong (1996), who study the problem of identification and theoretical restrictions. They show that any symmetric affiliated value model is observationally equivalent to some symmetric affiliated private value (APV) model because $U(\cdot)$ is unidentified, as any dependence across utilities arising from ν can be replaced by a dependence among private values. Similarly, the pure common value is unidentified from observed bids. If some additional information is available, such as the *ex post* common value, identification can be achieved. On the other hand, the symmetric APV model is identified.

Regarding estimation, a two-step estimation procedure can be developed. Let $B_1 = s(V_1)$ with $V_1 = \max_{j \neq 1} \sigma_j$. When $v_i = \sigma_i$, (1) becomes

$$v_i = b_i + \frac{G_{B_1|b_1}(b_i|b_i)}{g_{B_1|b_1}(b_i|b_i)} \equiv \xi(b_i, G). \tag{2}$$

Regarding theoretical restrictions, $\xi(\cdot)$ needs to be strictly increasing and the bid distribution $G(\cdot, \ldots, \cdot)$ must be affiliated and exchangeable. An interpretation of the APV model is that affiliation arises from some latent variable ν. Building on this interpretation, Li, Perrigne and Vuong (2000) propose a model with private information conditionally independent upon some common component. Specifically, each piece of private information is the product of two unobserved independent components, one specific to the auctioned object and common to all bidders, the other specific to each bidder, that is, $\sigma_i = x\eta_i$. Hence, $\log \sigma_i = \log x + \log \varepsilon_i$ with $\log x = [\log \nu + E(\log \eta)]$ and $\log \varepsilon_i = [\log \eta_i - E(\log \eta)]$ showing that $\log \varepsilon_i$ can be interpreted as an error term in a measurement error model with log *x* unobserved. Because the v_i can be recovered from (2) when $v_i = \sigma_i$, the densities for log *x* and log ε are nonparametrically identified and estimated with the use of characteristic functions. When $v_i = \nu$, (2) gives $E[V|\sigma_1 = \sigma, V_1 = \sigma]$. Under loglinearity of the latter, that is, $\log E[V|\sigma_1 = \sigma, V_1 = \sigma] = C + D\log \sigma$, the pure common value model is identified up to location and scale. It is important to test whether a common value or private value paradigm is the more appropriate.

Recent developments exploit how $E[V|\sigma_1 = \sigma, V_1 = \sigma]$ varies with the number of bidders to formulate such tests.

Several auction data provide evidence of bidders' asymmetry, which can arise from, for example, different firms' sizes, different access to information such as the drainage auctions, and different capacity constraints and locations as in construction procurements. Collusion may also lead to asymmetry as a cartel of bidders behaves differently from other bidders. Asymmetry is *ex ante* known to all bidders. A common feature of asymmetric auction models is that they lead to intractable systems of differential equations. Hence, the direct approach is difficult to implement as it requires the numerical determination of the equilibrium strategies for any trial parameter value. Let $F_1(\cdot), \ldots, F_I(\cdot)$ be the private value distributions of the *I* bidders whose identities are observed. For simplification, independent private values are considered, though the method can be easily extended to affiliated private values. Let $G_1(\cdot), \ldots, G_I(\cdot)$ be the corresponding bid distributions. The intractable system of differential equations can be rewritten as

$$v_i = b_i + \frac{1}{\sum_{j \neq i} \frac{g_j(b_j)}{G_j(b_j)}}, \; i = 1, \ldots, I. \tag{3}$$

This method has been used to analyse joint bidding in gas lease auctions and snow removal procurements, where asymmetry arises from a firm's location relative to contract location.

Bidders' risk neutrality is often assumed because the value of the object is small relative to bidders' assets. Recent studies have suggested that bidders may be risk averse in timber auctions. The experimental literature has noted a tendency to bid above the Bayesian Nash equilibrium, which can be rationalized by risk aversion. In a private value framework, the bidder's utility becomes $U(v_i - b_i)$ with $U(\cdot)$ strictly increasing and concave. Campo et al. (2006) study the identification and estimation of risk aversion. Using an indirect approach and omitting wealth to simplify, the differential equation defining the equilibrium strategy becomes

$$v_i = b_i + \lambda^{-1}\left(\frac{1}{I-1}\frac{G(b_i)}{g(b_i)}\right) \equiv \xi(b_i, U, G, I), \tag{4}$$

where $\lambda^{-1}(\cdot)$ denotes the inverse of $\lambda(\cdot) = U(\cdot) / U'(\cdot)$. The model is not identified only from observed bids. In fact, any bid distribution can be rationalized by a constant relative or absolute risk aversion model. Additional restrictions, such as parameterizing either the utility function or the private value distribution, are not sufficient to identify the model as an increase in the risk aversion parameter

can be compensated by a shrinkage of all the quantiles of $F(\cdot)$. Consequently, the authors parameterize a single quantile of $F(\cdot)$ to achieve identification of the model while exploiting auction heterogeneity. Under parameterization of $U(\cdot)$ and a conditional quantile, (4) at any quantile provides an estimating equation for the parameters of the utility function and the quantile of $F(\cdot)$. The method can be easily extended to affiliated private values and bidders' asymmetry in private values. Alternatively, if the number of bidders is exogenous, that is, $F(\cdot)$ is independent of *I*, nonparametric identification can be achieved. More generally, exclusion restrictions help in identifying the model. Regarding asymmetry, bidders may have heterogeneous preferences, that is, they may have different attitudes towards risk given their assets, experience, and so on. Thus, (4) evaluated at any quantile for two different bidders provides additional identifying restrictions since the corresponding quantile of $F(\cdot)$ is equal. Construction procurement data show that firms with more experience tend to be less risk averse. Risk aversion has important implications for several policy issues including the announcement of the reserve price and the auction format. These results allow more advanced auction models to be considered, in which risk aversion plays a key role. Examples includes stochastic values when uncertainties affect bidders' *ex post* value and financially constrained bidders.

Identical commodities such as treasury bills and electricity are sold sometimes through multi-unit auctions. A bidder acquires a share of the quantity supplied. Each bidder submits several (quantity, price) pairs. Hortaçsu (2002) studies discriminatory share auctions of treasury bills while considering private values in light of empirical evidence. Each bidder strategy is a demand function $y(p, \sigma_i)$ where σ_i is bidder's *i* private information. The clearing price P_c equates the bidder's demand function with the residual supply curve $Q - \Sigma_{j \neq i}^{I} y(p, \sigma_j)$, where *Q* is the total supply. Let $G(p, x)$ be the distribution of the residual supply faced by bidder *i* at price *p* given $y(p, \sigma_i) = x$, that is, $G(p, x) = \Pr[x \leq Q - \Sigma_{j \neq i}^{I} y(p, \sigma_j) | y(p, \sigma_i = x)] = \Pr[P_c \leq p | y(p, \sigma_i) = x]$. The optimal bid *p* for the quantity $y(p, \sigma_i)$ is

$$v[y(p, \sigma_i), \sigma_i] = p + \frac{G[p, y(p, \sigma_i)]}{\partial G[p, y(p, \sigma_i)] / \partial p},$$

where $v[y(p, \sigma_i), \sigma_i]$ is bidder's *i* marginal utility from winning the $y(p, \sigma_i)$th unit. With the use of a re-sampling strategy to estimate $G(\cdot, \cdot)$, the results are used to compare the discriminatory price mechanism with the uniform price mechanism. The problems of identification of the private information distribution and the restrictions imposed by the model on observables remain to be solved. This method has also been applied to electricity auctions.

The preceding developments ignore dynamic considerations, while bidders frequently participate in several auctions over time. Jofre-Bonet and Pesendorfer (2003) consider a dynamic auction to analyse highway construction procurements where previously won uncompleted contracts introduce capacity constraints affecting firms' actual costs. This involves inter-temporal optimization, while introducing asymmetry among bidders arising from different capacity constraints, location and size. If we use an

indirect approach, the inverse equilibrium strategies solve

$$c_i = b_i - \frac{1}{\sum_{j \neq i} \frac{g_j(b_j)}{1 - G_j(b_j)}} + \beta \sum_{j \neq i} \frac{\frac{g_j(b_j)}{1 - G_j(b_j)}}{\sum_{k \neq j} \frac{g_k(b_j)}{1 - G_k(b_j)}} [V_i(\omega(i)) - V_i(\omega(j))], \; i = 1, \ldots, I, \tag{5}$$

where β is a discount factor, $\omega(i)$ is a transition function indicating the sizes and remaining times of all current projects for bidder i, $V_i(\cdot)$ is the value function determining the discounted sum of expected future profits. The system (5) is similar to (3) with cost c_i and $1 - G_j(\cdot)$ as the firm with the lowest bid wins the procurement. Because the value function can be written as a function of the bid distributions, identification comes down to whether the cost distributions and the discount factor can be uniquely recovered from observed bids. Identification is obtained when the discount factor is known. Relying on standard numerical methods to approximate the value function, a two-step parametric procedure allows us to estimate the cost distributions $F_1(\cdot), \ldots, F_I(\cdot)$.

Econometrics of ascending auctions and applications

In the private value paradigm, a dominant strategy for every bidder is to exit the auction at his valuation. The bidding process ends when a single bidder remains. In the button auction model, the winning bid can be interpreted as the second highest among I values. Athey and Haile (2002) study identification of ascending auctions while emphasizing data requirements. When private values are independent and the number of bidders is observed, the transaction price is the (I–1)th order statistic $v^{(I-1:I)}$ whose distribution is

$$F^{(I-1:I)}(v) = \frac{I!}{(I-2)!} \int_0^{F(v)} t^{I-2}(1-t)\,dt,$$

from which the distribution $F(\nu)$ is recovered. When bidders are asymmetric and bidders' identities are known, a similar argument can be used to show that $F_1(\cdot), \ldots, F_I(\cdot)$ are identified. Nonparametric estimation can be performed. The problem becomes complicated when one considers more general frameworks. When private values are affiliated, the winning bid is not sufficient to recover affiliation among bids and hence $F(\cdot, \ldots, \cdot)$. Additional observations are needed.

However, many ascending auctions do not match the button auction model. In practice, bidders do not continuously indicate whether they are still participating. Moreover, because bid increments are often

used, bidders may fail to reveal their willingness to pay or even to bid. In the empirical literature it is agreed that, at most, the winning bid can be rationalized by the ascending auction model. An alternative approach is proposed by Haile and Tamer (2003), who formulate an incomplete model based on two simple assumptions: (*a*) bidders do not bid more than they are willing to pay; and (*b*) bidders do not allow an opponent to win at a price they can beat. These assumptions do not allow us to identify the private value distribution but provide some bounds on this distribution. Assumption (*a*) implies $b^{(l:l)} \leq v^{(l:l)}$ or equivalently $F^{(l:l)}(v) \leq G^{(l:l)}(v)$ for $i = 1, \ldots, l$. This inequality is used to construct the upper bound for $F(\cdot)$ as

$$F^{U}(v) = \min_{i,l} \varphi[G^{(i,l)}(v); i, l],$$

where $\varphi(\cdot)$ is a strictly increasing function defined as $F(v) = \varphi[F^{(i:l)}(v); i, l]$. Assumption (*b*) implies that all losing bidders have valuations no higher than the winning bid plus a bid increment Δ, i.e. $v_i \leq b^{(l:l)} + \Delta$ if $b_i < b^{(l:l)}$. Let $G_{\Delta}^{(l:l)}(\cdot)$ be the distribution of $b^{(l:l)} + \Delta$. Thus $G_{\Delta}^{(l:l)}(v) \leq F^{(l-1:l)}(v)$, which is used to construct the lower bound for $F(\cdot)$ as

$$F^{L}(v) = \max_{l} \varphi[G_{\Delta}^{(l:l)}(v); l-1, l].$$

Nonparametric estimation of $F^{U}(\cdot)$ and $F^{L}(\cdot)$ is proposed. Tight estimated bounds suggest that the data do not deviate much from the button auction model. Bounds for the optimal reserve price can also be derived. The method is illustrated on timber auction data, and can be extended to affiliated private values and asymmetric bidders.

In a common value paradigm, bidding takes a more complex form as bidders obtain information during the auction when their rivals drop out. The auction can be modelled as a game with several rounds with *I* –1 rounds indexed by $k = 0, 1, \ldots, I-2$. Bidders are indexed in the inverse order of their dropping out. Each bidder observes a signal σ_i of his value ν_i. An interesting feature of the ascending common value auction is that bidder's *j* dropping out is useful to bidder *i* for evaluating his own ν_i. In this game, every bidder has $I-1$ bidding functions $s_{ik}(\cdot)$, $k = 0, \ldots, I-2$. With asymmetric bidders, the equilibrium bid functions at round *k* are given by

$$s_{ik}(\sigma_i) = \mathrm{E}[v_i | \sigma_i; \sigma_j = s_{jk}^{-1}(s_{ik}(\sigma_i)), j = 1, \ldots, I-k, j \neq i, \Omega_k], i = 1, \ldots, I-k,$$

where $\Omega_k = \left\{\sigma_j = s_{j,l-j}^{-1}(P_{l-j}),\ j = l-k+1, \ldots, l\right\}$ is the public information set containing the observed signals of the bidders who have dropped out prior to round *k* and P_k is the (observed dropping out) price. Thus, at round *k* the $l-k$ inverse bidding strategies are solutions of the system of nonlinear equations

$$P_k = E[v_i | \sigma_i = s_{ik}^{-1}(P_k);\ \sigma_j = s_{jk}^{-1}(P_k), j = 1, \ldots, l-k,\ j \neq i, \Omega_k]. \quad (6)$$

Using log-normal distributions and a multiplicative form for ν_i and σ_i, Hong and Shum (2003) develop a tractable econometric model based on (6) that is estimated by either maximum likelihood or simulated nonlinear least squares. An illustration of the method is proposed on spectrum auctions which are organized in multiple rounds.

The recent development of auction websites provides new data opportunities. Bajari and Hortaçsu (2003) analyse coin auctions within a common value framework in light of resale opportunities, while bidders face an entry cost leading to endogenous entry. Another interesting characteristic is that the reserve price can be either posted or secret. As is well known, bidding activity is concentrated at the very end of the auction. The authors show that this practice, known as ‘sniping’, can be explained by a two-stage game in which no bidding is an equilibrium in the first stage, while second stage bids are the equilibrium bids in a sealed-bid second-price auction. Empirical results show that bidders’ entry increases with a secret reserve price.

Concluding remarks

The structural approach to analysing bidding data has been a field of extremely active research in the recent years. It has also contributed to the development of new econometric techniques. Many interesting problems remain to be addressed. Since auction models can be viewed as simple forms of asymmetric information, one can expect that more progress will be made in the analysis of complex asymmetric information models such as contracts.

See Also

- auctions (applications)
- auctions (theory)
- nonparametric structural models

Bibliography

Athey, S. and Haile, P. 2002. Identification of standard auction models. *Econometrica* 70, 2107–40.

Bajari, P. and Hortaçsu, A. 2003. The winner's curse, reserve prices and endogenous entry: empirical insights from eBay auctions. *Rand Journal of Economics* 34, 329–55.

Campo, S., Guerre, E., Perrigne, I. and Vuong, Q. 2006. Semiparametric estimation of first-price auctions with risk averse bidders. Working paper. University Park: Pennsylvania State University.

Guerre, E., Perrigne, I. and Vuong, Q. 2000. Optimal nonparametric estimation of first-price auctions. *Econometrica* 68, 525–74.

Haile, P. and Tamer, E. 2003. Inference with an incomplete model of English auctions. *Journal of Political Economy* 111, 1–51.

Hong, H. and Shum, M. 2003. Econometric models of asymmetric ascending auctions. *Journal of Econometrics* 112, 327–58.

Hortaçsu, A. 2002. Mechanism choice and strategic bidding in divisible good auctions: an empirical analysis of the Turkish treasury auction market. Working paper. Chicago: University of Chicago.

Jofre-Bonet, M. and Pesendorfer, M. 2003. Estimation of a dynamic auction game. *Econometrica* 71, 1443–89.

Laffont, J.-J., Ossard, H. and Vuong, Q. 1995. Econometrics of first-price auctions. *Econometrica* 63, 953–80.

Laffont, J.-J. and Vuong, Q. 1996. Structural analysis of auction data. *American Economic Review, Papers and Proceedings* 86, 414–20.

Li, T., Perrigne, I. and Vuong, Q. 2000. Conditionally independent private information in OCS wildcat auctions. *Journal of Econometrics* 98, 129–61.

Paarsch, H. 1992. Deciding between the common and private value paradigms in empirical models of auctions. *Journal of Econometrics* 51, 191–215.

Porter, R. 1995. The role of information in US offshore oil and gas lease auctions. *Econometrica* 63, 1–27.

How to cite this article

Perrigne, Isabelle and Quang Vuong. "auctions (empirics)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000219>
doi:10.1057/9780230226203.0071

The New Palgrave Dictionary of Economics Online

auctions (experiments)

John H. Kagel and Dan Levin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Experiments permit rigorous investigations of auction theory generating a dialogue with theorists and policymakers. In single-unit private value auctions the revenue equivalence theorem fails, but the comparative static predictions of Nash bidding theory hold, indicating that bidders are responsive to the primary economic forces present in the theory. In single-unit common value auctions inexperienced bidders invariably suffer from a ‘winner's curse’, and the comparative static predictions of the theory fail, but more experienced bidders do substantially better. Recent research dealing with Internet auctions, mixed private and common value auctions and multi-unit demand auctions are surveyed as well.

Keywords

adverse selection; auctions; auctions (experiments); Becker–DeGroot–Marshak procedure; common value auctions; Dutch auctions; English auctions; English clock auctions; first-price auctions; independent private values model; Internet auctions; learning direction theory; mechanism design; multi-unit demand auctions; revenue equivalence theorem; risk aversion; risk-neutral Nash equilibrium; second-price auctions; Vickrey auction; winner's curse

Article

Experimental work in auctions interacts with theory, providing a basis for testing and modifying theoretical developments. It has advantages and disadvantages relative to empirical work with field data, so that we view the two as complimentary. Experimental work is used increasingly as a test bed for new auction formats such as the Federal Communication Commission's (FCC) sale of spectrum (air-wave) rights.

Until recently most of theoretical and experimental work was devoted to single-unit demand auctions. With the success of the FCC's spectrum auctions, much of the interest has shifted to auctions in which

individual bidders demand multiple units. Experimental work in this area is still in its infancy. In keeping with the historical development of the field, we first report on single-unit demand auctions and then move to multi-unit demand auctions and Internet auctions.

Single-unit, private-value auctions

Initial experimental research on auctions focused on the independent private values (IPV) model investigating the revenue equivalence theorem. In the IPV model each bidder knows his valuation of the item with certainty, bidders' valuations are drawn identically and independently from each other, and bidders know the distribution from which their rivals' values are drawn (but not their values) and the number of bidders. Under the revenue equivalence theorem the four main auction formats – first- and second-price sealed-bid auctions, English and Dutch auctions – yield the same average revenue for risk neutral bidders. Further, first-price sealed-bid and Dutch auctions are theoretically isomorphic – they yield the same revenue for each auction trial regardless of risk preferences – as are second-price sealed-bid and English clock auctions. These isomorphisms are particularly attractive as it is hard to control bidders' risk preferences. These theoretical results are also quite surprising and counter-intuitive as the Dutch auction starts with a high price which is lowered until a bidder accepts at that price. And in the English auctions the price starts low and increases until only one bidder is left standing and pays the price where the next-to-last bidder dropped out; while in a first- (second-) price sealed-bid auction the high bidder wins the item and pays the highest (second-highest) bid.

An experimental session typically consists of 20–40 auction periods under a given auction institution. Subjects' valuations are determined randomly prior to each auction period (by the experimenter) and are private information. Valuations are typically independent and identical draws (i.i.d) from a *uniform* distribution. In each period the high bidder earns a profit equal to his value less the auction price; other bidders earn zero profit. Bids are commonly restricted to be non-negative and rounded to the nearest penny. Theory does not specify what information feedback bidders ought to get after each auction. Although such information is unimportant in a one-shot auction, it may be important, even critical, to learning given that experimental sessions typically consist of a number of auction periods. Information feedback usually differs between different experimenters, with almost all experimenters reporting back the auction price to all bidders and own earnings to the winning bidder.

Strategic equivalence usually fails between the relevant auction formats: Coppinger, Smith and Titus (1980) and Cox, Roberson and Smith (1982) found higher prices in first-price than in Dutch auctions (about five per cent higher) with these differences holding across auctions with different numbers of bidders. Further, bidding was significantly above the risk-neutral Nash equilibrium (RNNE) in the first-price auctions for all numbers of bidders $n>3$, which is consistent with risk-averse bidders.

Kagel, Harstad and Levin (1987) reported failures of strategic equivalence in second-price and English clock auctions, with winning bids in the second-price auctions averaging 11 per cent *above* the predicted equilibrium price. In contrast, market prices converge rapidly to the predicted equilibrium in the clock auctions. Bidding above value in second-price auctions is widespread, with 62 per cent of all bids above values, 30 per cent of all bids essentially equal to value (within five cents of it), and eight per cent of all bids below it (Kagel and Levin, 1993). (In clock auctions price rises by fixed increments with bidders counted as active until they drop out – and are not permitted to re-enter the auction. This format insures

clear information flows as a consequence of announcing irrevocable drop-out prices.)

Bidding above value in second-price auctions is attributable to a number of factors: (*a*) it is sustainable since average profits are positive, (*b*) figuring out the dominant strategy is not that obvious, and (*c*) the feedback from losses that would promote the dominant bidding strategy is weak (Kagel, Harstad and Levin, 1987). Subsequent research generalizes the superiority of the (dynamic) clock auction format compared to the (static) sealed-bid format to Vickrey-style auctions in which bidders demand multiple units. The closer conformity to equilibrium outcomes in the clock auctions results from the clock format in conjunction with bidders knowing that the auction ends when the next-to-last bidder drops out. This induces bidders to remain active as long as the clock price is less than their value (as they have nothing to lose by remaining active and might win the item) and to drop out once the price is greater than their value (as they will lose money for sure should they win the item) (Kagel and Levin, 2006).

Efficiency in private value auctions can be measured by the percentage of auctions won by the high-value holder. In Cox, Roberson and Smith (1982) 88 per cent of the first-price auctions were Pareto efficient compared with 80 per cent of the Dutch auctions. In contrast, efficiency in first- and second-price auctions may be quite comparable; for example, 82 per cent of the first-price auctions and 79 per cent of the second-price auctions reported in Kagel and Levin (1993) were Pareto efficient. More work needs to be devoted to comparing efficiency across auction institutions.

A number of papers have explored bidding above the RNNE in first-price sealed-bid auctions, questioning the risk-aversion interpretation. This has generated some heated debate (see the December 1992 issue of the *American Economic Review*). Isaac and James (2000) compare estimates of risk preferences from first-price auctions with estimates using the Becker–DeGroot–Marshak (BDM) procedure for comparably risky choices. The Spearman rank–correlation coefficient between individual subject risk parameters is significantly *negatively* correlated under the two procedures. Subjects whose bids in the first-price auction are relatively risk neutral remain risk neutral under BDM, but those who are relatively risk averse in the first-price auction become relatively risk loving under BDM. The net result is that *aggregate* measures of risk preferences show that bidders are risk averse in the first-price auction but risk neutral, or moderately risk loving, under the BDM procedure. Although it is well known from the psychology literature that different elicitation procedures will yield somewhat different quantitative predictions, a negative correlation between measures seems rather astonishing. (See Dorsey and Razzolini, 2003, for a similar investigation.)

Neugebauer and Selten (2006) compare treatments with different information feedback**:** (i) a bidder only learns if s/he won the auction or not, (ii) the winning bid (market price) is revealed to bidders whether they win or not; and (iii) the winning bid is revealed to bidders and the winner learns the second highest bid as well. They find that average bids are highest under treatment (ii) and exceed the RNNE for every given market size. In contrast, bidding above the RNNE does not occur consistently, or is not as strong, in the other two treatments. They use ‘learning direction theory’ to argue that the information feedback in (ii) promotes bidding above the RNNE. However, the result for treatment (iii) contrasts with results from Kagel, Harstad and Levin (1987) and Dyer, Kagel and Levin (1989a), who find consistent bidding above the RNNE when providing bidders with all bids and valuations following each auction. Perhaps the best conclusion at this point is that subjects typically act ‘as if’ they are risk averse in first-price auctions, while the underlying basis of their behaviour remains open to interpretation.

In spite of the reported deviations from equilibrium outcomes reported above, the comparative static

implications of the IPV model tend to hold (albeit with varying levels of noise). Bidding in first-price auctions increases regularly in response to increased numbers of bidders. For example, in a series of first-price sealed-bid auctions, 86 per cent of subjects increased their bids when the number of bidders increased from five to ten, with the majority of these increases (60 per cent) being statistically significant, with no subjects decreasing their bids by a statistically significant amount (Battalio, Kogut and Meyer, 1990). More aggressive bidding in response to increased numbers of rivals would seem to be a natural reaction, and can be rationalized by plausible ad hoc rules of thumb.

Kagel and Levin (1993) provide a more stringent test of the comparative static implications of the IPV model using a third-price auction in which the high bidder wins the item and pays the third-highest bid. In this case the model predicts that bids will be *above* values and will be *reduced* in response to increases in n. They find that 85–90 per cent of all bids are above value compared with 58–67 per cent in second-price auctions and less than 0.5 per cent in first-price auctions. Further, comparing auctions with n=5 and n=10 (i) in first-price auctions *all* bidders increased their bids on average (average increase of \$0.65 per auction; $p<.01$), (ii) in second-price auctions the majority of bidders did not change their bids on average (average decrease of \$0.04; $p>.10$), and (iii) in third-price auctions 46 per cent of all subjects *decreased* their bids on average (average decrease of \$0.40 per auction; $p<.05$). Even stronger qualitative support for the theory is reported when the calculations are restricted to valuations lying in the top half of the domain of valuations (where bidders have a realistic chance of winning and might be expected to take bidding more seriously). Thus, although a number of bidders in third-price auctions clearly err in response to increased numbers of rivals by increasing, or not changing, their bids, the change in pricing rules has relatively large and statistically significant effects on bidders' responses in the *direction* that Nash equilibrium bidding theory predicts. This experiment also illustrates one of the great strengths of the experimental method as there are no third-price auctions outside the lab, where it was developed for the explicit purpose of providing unusual, counter-intuitive predictions to use in testing the theory. The results are increased confidence in the fundamental 'gravitational' forces underlying the theory, in spite of violations of its point predictions. The latter could be the result of some uncontrolled factor impacting on behaviour and/or simple miscalibration on subjects' part.

Single-unit common value auctions

In common value auctions (CVA) the value of the item is the same to all bidders. What makes common value auctions interesting is that bidders receive signals (estimates**)** that are correlated (affiliated) with the value of the item but they do not know its true value. Mineral rights auctions (for example, outer continental shelf – OCS – oil lease auctions) are usually modelled as a common value auction. There is a common value element to most auctions. Bidders for a painting may purchase it for their own pleasure, a private value element, but also for investment and eventual resale, the common value element.

Experimental research on CVAs has focused on the 'winner's curse'. Although all bidders obtain unbiased estimates of the item's value, they typically win in cases where they have (one of) the highest signal value. Unless this adverse selection problem is accounted for, it will result in winning bids that are systematically too high, earning below normal or negative profits – a disequilibrium phenomenon. Oil companies claim they fell prey to the winner's curse in early OCS lease sales, with similar claims made in a variety of other settings (for example, free agency markets for professional athletes and

corporate takeovers). Economists are naturally sceptical of such claims as they involve out-of-equilibrium play. Experiments clearly show the presence of a winner's curse for inexperienced bidders under a variety of circumstances and with different experimental subjects: average undergraduate or MBA students (Bazeramn and Samuelson, 1983; Kagel and Levin, 1986), extremely bright (Cal Tech) undergraduates (Lind and Plott, 1991), experienced professionals in a laboratory setting (Dyer, Kagel and Levin, 1989b), and auctions in which it is common knowledge that one bidder knows, with certainty, the value of the item (Kagel and Levin, 1999). Further, these deviations from equilibrium predictions cannot be explained by simple miscalibration on bidders' part as the theory's comparative static implications are systematically violated when bidders suffer from a winner's curse; for example, bidder responses to additional information or increased numbers of rivals.

Kagel et al. (1989) find that inexperienced bidders suffer a pervasive winner's curse in first-price, sealed-bid auctions. For the first nine auctions, profits averaged minus $2.57 compared with the RNNE prediction of $1.90, with only 17 per cent of all auctions having positive profits. This is not a simple matter of bad luck as 59 per cent of all bids, and 82 per cent of the high bids, were above the expected value of the item conditional on winning the auction. Although public information in first-price auctions is predicted to raise sellers' revenue, it reduces it for inexperienced bidders as subjects use the public information to help overcome the winner's curse (Kagel and Levin, 1986). Similarly, 'public information' reduces revenue in English clock auctions when bidders suffer from a winner's curse (Levin, Kagel and Richard, 1996). Further, experienced bidders appear to adjust to the winner's curse through a 'hot stove' learning process**:** with the losses, bids are lowered and losses are mitigated, or eliminated, but there is no real understanding of the adverse selection problem. For example, an increase in *n* generates higher individual bids, although theory predicts a slight reduction (Kagel and Levin, 1986). Efforts to explain the winner's curse in terms of limited liability for losses and/or the 'joy of winning' fail as well (Kagel and Levin, 1991; Holt and Sherman, 1994). In short, inexperienced subjects do not perform well in pure common value auctions.

Experienced subjects learn to overcome the worst effects of the winner's curse, earning positive average profits. But these rarely exceed 65 per cent of the RNNE profit, and virtually all subjects are *not* best responding to their rivals' overly aggressive bids (Kagel and Richard, 2001). However, once bidders overcome the worst effects of the winner's curse, public information raises sellers' revenue, English auctions raise more revenue than sealed-bid auctions, and a number of other comparative static implications of the theory are satisfied as well (Kagel and Levin, 2002). Experienced bidders learn to overcome the winner's curse through a combination of individual learning and market selection process whereby bankrupt bidders self-select out of further experimental sessions. Ability as measured by composite SAT/ACT scores (standardized college entrance exam scores) matters in terms of avoiding the winner's curse, with the biggest and most consistent impact resulting from those with *below median* scores being more susceptible to the winner's curse. Economics and business majors consistently bid more aggressively than others (thus, lose more), and women, at least initially, are much more susceptible to a winner's curse than men. However, there is still a winner's curse even for the best-calibrated demographic and ability groups (Casari, Ham and Kagel, 2007).

Experiments combining common-value and private-value elements

Goeree and Offerman (2002) provide the only experimental study to date in which the object's expected value depends on both private and common value elements. (The difficulty here is in combining private and common value information into a single statistic that maps into a bid.) Actual bids lie in between the RNNE benchmark of fully rational bidding and the naive benchmark in which subjects completely fail to account for the winner's curse. The winner's curse effect is more pronounced the less important a bidder's private value is relative to the common value. Realized efficiency is roughly at the level predicted under the RNNE, with the winner's curse only raising seller revenue and cutting into bidder profits. This occurs because (*a*) almost all bidders suffer from a winner's curse and (*b*) the degree of suffering is roughly the same across bidders, so that the size of the private value element serves to dictate who wins the item.

In an almost common value auction one bidder, the advantaged bidder, has an added private value for the item, unlike all the other (regular) bidders who care only about the common value. With only two bidders, even a tiny private value advantage is predicted to have an explosive effect in second-price sealed-bid auctions: the advantaged bidder always wins and revenue decreases dramatically as the regular bidder lowers her bid to protect against a winner's curse. This effect extends to a variety of English auctions that start with more than two bidders, raising serious concerns about the English auction format (Klemperer, 1998). Three experiments have looked at almost common value auctions using both second-price sealed-bid and clock auctions (Avery and Kagel, 1997; Rose and Levin, 2005; and Rose and Kagel, 2005). In all cases the response to the private value advantage has been proportional rather than explosive. This is true even with experienced bidders who earn a respectable share of RNNE profits in pure common value first-price and clock auctions (Rose and Kagel, 2005). The apparent reason for these failures is that bidders do not fully appreciate the adverse selection effect conditional on winning, which is exacerbated for regular bidders with an advantaged rival. As such, the behavioural mechanism underlying the explosive effect is not present, and there are no forces at work to replace it.

Internet auctions

Internet auctions provide new opportunities to conduct experiments to study old and new puzzles. Lucking-Reiley (1999) has used the Internet to sell collectable trading cards under the four standard auction formats, testing the revenue equivalence theorem. He finds that Dutch auctions produce 30 per cent higher revenue than first-price auctions, a reversal of previous laboratory results, and that English and second-price auctions produce roughly equivalent revenue. These results are interesting but lack the controls present in more standard laboratory experiments; that is, there may well be a common value element to the trading cards, and Dutch auctions provide an opportunity to use the game cards immediately, which cannot be done until the fixed closing date in the first-price auctions. Garratt, Walker and Wooders (2004) conduct a second-price auction, recruiting subjects with substantial experience bidding on eBay. Using induced valuations, they find that average bids are close to valuations, but those with prior experience as *sellers* tend to underbid and those with prior experience as *buyers* tend to overbid.

In eBay auctions which have a fixed closing time many bidders *snipe* (submit bids seconds before the closing time), while other bidders increase their bids over time in response to higher bids. This seems

puzzling since eBay has a number of characteristics similar to a second-price auction. In addition, there is substantially more last-minute bidding for comparable (private-value) items in eBay than in Amazon auctions, which automatically extend the deadline in response to last-minute bids. Roth and Ockenfels (2002) argue that sniping results from the fixed deadline in eBay, suggesting at least two rational reasons for sniping. Because there are differences between eBay and Amazon other than their ending rules, they conduct a laboratory experiment in which the only difference between auction institutions is the ending rule – a dynamic eBay auction with a .8 (1.0) probability that a late bid will be accepted (eBay.8 and eBay1, respectively) and an Amazon-style auction with a .8 probability that a late bid will be accepted, in which case the auction is automatically extended (Ariely, Ockenfels and Roth, 2005). The results show quite clearly that there is more late bidding in both eBay auctions than in the Amazon auction. Further, there is significantly more late bidding in eBay1 than in eBay.8, which at least rules out one possible rational explanation for sniping – implicit collusion on the part of snipers in an effort to get the item at rock-bottom prices since not all last-minute bids will be recorded (due to congestion) at the website.

Salmon and Wilson (2008) investigate the Internet practice of second-chance offers to non-winning bidders when selling multiple (identical) items. They compare a two-stage game with a second-price auction followed by an ultimatum game between the seller and the second-highest bidder with a sequential English auction. As predicted, the auction-ultimatum game mechanism generates more revenue than the sequential English auction.

Multi-unit demand auctions

Most of the work on multi-unit demand auctions has been devoted to mechanism design issues, in particular dealing with problems created by complementarities, or synergies, between items. Absent package bidding, the latter can create an ‘exposure’ problem whereby efficient outcomes require submitting bids above the stand-alone values for individual units since the value of the package is more than the sum of the individual values. Correcting for this problem by permitting package bids increases the complexity of the auction significantly, and creates a ‘threshold’ problem whereby ‘small’ bidders (for example, those with only local markets) could, in combination, potentially outbid a large competitor who can internalize the complementarities. But the small bidders have no means to coordinate their bids. Leading examples of this line of research are Porter et al. (2003), Kwasnica et al. (2005), and Goeree, Holt and Ledyard (2006). Much more work remains to be done in this area.

See Also

- auctions (applications)
- auctions (empirics)
- auctions (theory)

Bibliography

Ariely, D., Ockenfels, A. and Roth, A.E. 2005. An experimental analysis of ending rules in internet auctions. *RAND Journal of Economics* 36, 890–907.

Avery, C. and Kagel, J.H. 1997. Second-price auctions with asymmetric payoffs: an experimental investigation. *Journal of Economics and Management Strategy* 6, 573–604.

Battalio, R.C., Kogut, C.A. and Meyer, D.J. 1990. Individual and market bidding in a Vickrey first-price auction: varying market size and information. In *Advances in Behavioral Economics*, vol. 2, ed. L. Green and J.H. Kagel. Norwood, NJ: Ablex Publishing.

Bazeramn, M.H. and Samuelson, W.F. 1983. I won the auction but don't want the prize. *Journal of Conflict Resolution* 27, 618–34.

Casari, M., Ham, J.C. and Kagel, J.H. 2007. Selection bias, demographic effects and ability effects in common value auction experiments. *American Economic Review* 97, 1278–304.

Coppinger, V.M., Smith, V.L. and Titus, J.A. 1980. Incentives and behavior in English, Dutch and sealed-bid auctions. *Economic Inquiry* 43, 1–22.

Cox, J., Roberson, B. and Smith, V.L. 1982. Theory and behavior of single object auctions. In *Research in Experimental Economics*, ed. V.L. Smith. Greenwich: JAI Press.

Dorsey, R. and Razzolini, L. 2003. Explaining overbidding in first price auctions using controlled lotteries. *Experimental Economics* 6, 123–40.

Dyer, D., Kagel, J.H. and Levin, D. 1989a. Resolving uncertainty about the number of bidders in independent private-value auctions: an experimental analysis. *RAND Journal of Economics* 20, 268–79.

Dyer, D., Kagel, J.H. and Levin, D. 1989b. A comparison of naive and experienced bidders in common value offer auctions: a laboratory analysis. *Economic Journal* 99, 108–15.

Garratt, R., Walker, M. and Wooders, J. 2004. Behavior in second-price auctions by highly experienced eBay buyers and sellers. Working Paper No. 1181, Department of Economics, UC Santa Barbara.

Goeree, J.K., Holt, C.A. and Ledyard, J.O. 2006. *An Experimental Comparison of the FCC's Combinatorial and Non-Combinatorial Simultaneous Multiple Round Auctions*. Prepared for the Wireless Telecommunications Bureau of the Federal Communications Commission. Online. Available at http://wireless.fcc.gov/auctions/data/papersAndStudies/fcc_final_report_071206.pdf, accessed 1 February 2007.

Goeree, J.K. and Offerman, T. 2002. Efficiency in auctions with private and common values: an

experimental study. *American Economic Review* 92, 625–43.

Holt, C.A., Jr. and Sherman, R. 1994. The loser's curse and bidder's bias. *American Economic Review* 84, 642–52.

Isaac, M. and James, D. 2000. Just who are you calling risk averse? *Journal of Risk and Uncertainty* 20, 177–87.

Kagel, J.H., Harstad, R.M. and Levin, D. 1987. Information impact and allocation rules in auctions with affiliated private values: a laboratory study. *Econometrica* 55, 1275–304.

Kagel, J.H. and Levin, D. 1986. The winner's curse and public information in common value auctions. *American Economic Review* 76, 894–920.

Kagel, J.H. and Levin, D. 1991. The winner's curse and public information in common value auctions: reply. *American Economic Review* 81, 362–9.

Kagel, J.H. and Levin, D. 1993. Independent private value auctions: bidder behavior in first-, second- and third-price auctions with varying numbers of bidders. *Economic Journal* 103, 868–79.

Kagel, J.H. and Levin, D. 1999. Common value auctions with insider information. *Econometrica* 67, 1219–38.

Kagel, J.H. and Levin, D. 2001. Behavior in multi-unit demand auctions: experiments with uniform price and dynamic Vickrey auctions. *Econometrica* 69, 413–54.

Kagel, J.H. and Levin, D. 2002. Bidding in common value auctions: a survey of experimental research. In *Common Value Auctions and the Winner's Curse*. Princeton: Princeton University Press.

Kagel, J.H. and Levin, D. 2006. Implementing efficient multi-object auction institutions: an experimental study of the performance of boundedly rational agents. Mimeo, Ohio State University.

Kagel, J.H., Levin, D., Battalio, R. and Meyer, D.J. 1989. First-price common value auctions: bidder behavior and the winner's curse. *Economic Inquiry* 27, 241–58.

Kagel, J.H. and Richard, J.F. 2001. Super-experienced bidders in first-price common value auctions: rules of thumb, Nash equilibrium bidding and the winner's curse. *Review of Economics and Statistics* 83, 408–19.

Klemperer, P. 1998. Auctions with almost common values: the ‘wallet game’ and its applications. *European Economic Review* 42, 757–69.

Kwasnica, A.M., Ledyard, J.O., Porter, D. and DeMartini, C. 2005. A new and improved design for multiobject iterative auctions. *Management Science* 51, 419–34.

Levin, D., Kagel, J.H. and Richard, J.F. 1996. Revenue effects and information processing in English common value auctions. *American Economic Review* 86, 442–60.

Lind, B. and Plott, C.R. 1991. The winner's curse: experiments with buyers and with sellers. *American Economic Review* 81, 335–46.

Lucking-Reiley, D. 1999. Using field experiments to test equivalence between auction formats: magic on the internet. *American Economic Review* 89, 1062–80.

Neugebauer, T. and Selten, R. 2006. Individual behavior of first-price auctions: the importance of information feedback in computerized experimental markets. *Games and Economic Behavior* 54, 183–204.

Porter, D., Rassenti, S., Roopnarine, A. and Smith, V. 2003. Combinatorial auction design. *Proceedings of the National Academy of Sciences* 100, 11153–7.

Rose, S.L. and Kagel, J.H. 2005. Bidding in almost common value auctions: an experiment, mimeographed. Mimeo, Ohio State University.

Rose, S.L. and Levin, D. 2005. An experimental investigation of the explosive effect in common value auctions. Mimeo, Ohio State University.

Roth, A.E. and Ockenfels, A. 2002. Last-minute bidding and the rules for ending second-price auctions: evidence from eBay and Amazon auctions on the internet. *American Economic Review* 92, 1093–103.

Salmon, T.C. and Wilson, B.J. 2008. Second chance offers versus sequential auctions: theory and behavior. *Economic Theory* 34, 47–67.

How to cite this article

Kagel, John H. and Dan Levin. "auctions (experiments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000241> doi:10.1057/9780230226203.0072

The New Palgrave Dictionary of Economics Online

auctions (theory)

Lawrence M. Ausubel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Auction theory has undergone two waves of innovation. The first, which originated with Vickrey (1961) and was completed in the early 1980s, focused on single-item auctions. Results included: guiding principles such as revenue equivalence; the derivation of the optimal auction; and comparisons of first-price, second-price and English auctions. The second, influenced by Treasury and spectrum auctions, emerged in the 1990s and dealt particularly with multi-item auctions. Research has studied: static auctions, including pay-as-bid and uniform-price auctions; dynamic auctions such as simultaneous ascending and clock auctions; combinatorial auctions; and efficient auction design. Much progress has been made, but outstanding problems remain.

Keywords

auctions; auctions (theory); bidding; clock auction; coalitional form game; combinatorial auctions; complements; core; demand reduction; Dutch auction; efficient auctions; English auction; envelope theorem; equilibrium; experimental economics; first-price auction; game theory; games of complete information; games of incomplete information; incentive compatibility; linkage principle; optimal auctions; pay-as-bid auction; private information; revenue equivalence theorem; sealed-bid auctions; second-price auction; simultaneous ascending auctions; sincere bidding; spectrum auctions; subgame perfection; substitutes; tâtonnement; Treasury auctions; uniform-price auction; Vickrey auction; Vickrey, W. S.; Walrasian auctioneer; winner's curse

Article

Auctions occupy a deservedly prominent place within microeconomics and game theory, for at least three reasons.
First, the auction is, in its own right, an important device for trade. Auctions have long been a common way of selling diverse items such as works of art and government securities. In recent years, their

importance in consumer markets has increased through the ascendancy of eBay and other Internet auctions. At the same time, the use of auctions for transactions between businesses has expanded greatly, most notably in the telecommunications, energy and environmental sectors, and for procurement purposes generally.

Second, auctions have become the clearest success story in the application of game theory to economics. In most applications of game theory, the modeller has considerable (perhaps excessive) freedom to formulate the rules of the game, and the results obtained will often be highly sensitive to the chosen formulation. By way of contrast, an auction will typically have a well-defined set of rules, yielding clearer theoretical predictions.

Third, there has been an increasing wealth of auction data available for empirical analysis in recent years. In conjunction with the available theory, this has led to a growing body of empirical work on auctions. Moreover, auctions are very well suited for laboratory experiments and they have been a very fruitful area for experimental economics.

This article is limited in its scope to *auction theory*. Other related articles, reviewing empirical and experimental work on auctions and the theoretical analysis of mechanism design, are cross-referenced at the end.

1 Introduction

Auction theory is often said to have originated in the seminal 1961 article by William Vickrey. While Vickrey's insights were initially unrecognized and it would be many years before his work was followed up by other researchers, it eventually led to a formidable body of research by pioneers including Wilson, Clarke, Groves, Milgrom, Weber, Myerson, Maskin and Riley. The first wave of theoretical research into auctions was concluded in the mid-1980s, by which time there was a widespread sense that it had become a relatively complete body of work with very little remaining to be discovered. See McAfee and McMillan (1987) for an excellent review of the first wave of auction theory.

However, the perception that auction theory was complete began to change following two pivotal events in the 1990s: the Salomon Brothers scandal in the US government securities market in 1991, and the advent of the Federal Communications Commission (FCC) spectrum auctions in 1994. In the aftermath of the former, the Department of the Treasury sought input from academia concerning the US Treasury auctions. In the preparation for the latter, the FCC encouraged the active involvement of auction theorists in the design of the new auctions.

Each of these two episodes undoubtedly benefitted from the participation of academics. In particular, the FCC introduced an innovative dynamic auction format – the simultaneous ascending auction – whose empirical performance appears far superior to previous static sealed-bid auctions. The Treasury's experimentation with, and eventual adoption of, uniform-price auctions in place of pay-as-bid auctions also appears to have resulted from economists’ input.

At the same time, these two pivotal events underscored some extremely serious limitations in auction theory as it existed in the early to mid-1990s. It became apparent then that the theory that had been developed was almost exclusively one of single-item auctions, and that relatively little was established concerning multi-item auctions. As the flip side of the same coin, these episodes made it obvious that many of the empirically important examples of auctions involve a multiplicity of items. As a result, a second wave of theoretical research into auctions, focusing especially on multi-item auctions, emerged

in the middle of the 1990s and continued into the 21st century.

This article begins by reviewing the theory of single-item auctions, largely completed during the first period of research. It continues by reviewing the theory of multi-unit auctions, still a work in progress as of 2007.

The scope and detail of the present article is necessarily quite limited. For deeper and more comprehensive treatments of auctions, three notable books, by Krishna (2002), Milgrom (2004) and Cramton, Shoham and Steinberg (2006), are especially recommended to readers. Earlier survey articles by McAfee and McMillan (1987) and Wilson (1992) also provide excellent treatments of the literature on single-item auctions. A compendium by Klemperer (2000) brings together many of the best articles in auction theory.

2 Sealed-bid auctions for single items

Much of the analysis within traditional auction theory has concerned sealed-bid auctions (that is, static games) for single items. Bidders submit their sealed bids in advance of a deadline, without knowledge of any of their opponents' bids. After the deadline, the auctioneer unseals the bids and determines a winner. The following are the two most commonly studied sealed-bid formats:

- *First-price auction*: the highest bidder wins the item, and pays the amount of his bid.
- *Second-price auction*: the highest bidder wins the item, and pays the amount bid by the second-highest bidder.

Note that the above auction formats (and, indeed, all of the auctions described in this article) have been described for a regular auction in which the auctioneer offers items for sale and the bidders are buyers. Each can easily be restated for a 'reverse auction' (that is, procurement auction) in which the auctioneer solicits the purchase of items and the bidders are sellers. For example, in a second-price reverse auction, the lowest bidder is chosen to provide the item and is paid the amount bid by the second-lowest bidder.

2.1 The private values model

A seller wishes to allocate a single unit of a good or service among n bidders (i=1,•…•, n). The bidders bid simultaneously and independently as in a non-cooperative static game. Bidder i's payoff from receiving the item in return for the payment y is given by $v_i - y$ (whereas bidder i's payoff from not winning the item is normalized to zero). Each bidder i's valuation, v_i, for the item is private information. Bidder i knows v_i at the time he submits his bid. Meanwhile, the opposing bidders $j \neq i$ view v_i as a random variable whose realization is unknown, but which is drawn according to the known joint distribution function $\hat{F}(v_1, \ldots, v_i, \ldots, v_n)$.

This model is referred to as the *private values* model, on account that each bidder's valuation depends only on his own – and not the other bidders' – information. (By contrast, in a *pure common values* model, $v_i = v_j$, for all i, j=1,•…•, n; and in an *interdependent values* model, bidder i's valuation is allowed

to be a function of $v_{-i} = \{v_j\}_{j \neq i}$, as well as of v_i.) With private values, some especially simple and elegant results hold, particularly for the second-price auction.

Two additional assumptions are frequently made. First, we generally assume that bidders are *risk neutral* in evaluating their payoffs under uncertainty. That is, each bidder seeks merely to maximize the mathematical expectation of his payoff. Second, we often assume *independence* of the private information. That is, the joint distribution function, $\hat{F}(v_1, \ldots, v_n)$, is given by the product of separate distribution functions, $F_i(\cdot)$, for each of the v_i. However, both the risk neutrality and independence assumptions are unnecessary for solving the second-price auction, which we analyse first.

2.2 Solution of the second-price auction

Sincere bidding (that is, the truthful bidding of one's own valuation) is a Nash equilibrium of the sealed-bid second-price auction, under private values. That is, if each bidder *i* submits the bid $b_i = v_i$, then there is no incentive for any bidder to unilaterally deviate. Moreover, sincere bidding is a weakly dominant strategy for each bidder; and sincere bidding by all bidders is the unique outcome of elimination of weakly dominated strategies. These facts make the sincere bidding equilibrium an especially compelling outcome of the second-price auction.

Let $\hat{b}_{-i} = \max_{j \neq i} \{b_j\}$, the highest among the opponents' bids. The dominant strategy property is easily established by comparing bidder *i*'s payoff from the sincere bid of $b_i = v_i$ with his payoff from instead bidding $b_i' < v_i$ ('shading' his bid). If $\hat{b}_{-i}$ is less than b_i' or greater than v_i, then bid-shading has no effect on bidder *i*'s payoff; in the former case, bidder *i* wins either way, and in the latter case, bidder *i* loses either way. However, in the event that $\hat{b}_{-i}$ is between b_i' and v_i, the bid-shading makes a difference: if bidder *i* bids v_i, he wins the auction and thereby achieves a positive payoff of $v_i - \hat{b}_{-i} > 0$; whereas, if bidder *i* bids b_i', he loses the auction and receives zero payoff. Thus, $b_i = v_i$ weakly dominates any bid $b_i' < v_i$. A similar comparison finds that $b_i = v_i$ weakly dominates any bid $b_i' > v_i$. Sincere bidding is optimal, regardless of the bidding strategies of opposing bidders.

Note that the above argument in no way uses the risk neutrality or independence assumptions, nor does it require any form of symmetry. Sincere bidding may also be viewed as an *ex post equilibrium* of the second-price auction, in the sense that the strategy would remain optimal even if the bidder were to learn his opponents' bids before he was required to submit his own bid. Indeed, one of the strengths of the result that sincere bidding is a Nash equilibrium in weakly dominant strategies is that it basically relies only upon the private values assumption, and is otherwise extremely robust to the specification of the model.

2.3 Incentive compatibility in any sealed-bid auction format

Consider any equilibrium of *any* sealed-bid auction format, in the private values model. Given that bidder *i*'s valuation is private information, observe that there is nothing to force bidder *i* to bid according

to his true valuation v_i instead of some other valuation w_i. As a result, the equilibrium must have a structure that gives bidder *i* the incentive to bid according to his true valuation. This requirement is known as *incentive compatibility*.

In the following derivation, we assume that the support of each bidder *i*'s valuation is the interval $[\underline{v}_i, \overline{v}_i]$. We will make both the risk neutrality and independence assumptions. Let $\Pi_i(v_i)$ denote bidder *i*'s expected payoff, let $P_i(v_i)$ denote bidder *i*'s probability of winning the item, and let $Q_i(v_i)$ denote bidder *i*'s expected payment in this equilibrium, when his valuation is v_i. The reader should note that $Q_i(v_i)$ refers here to bidder *i*'s unconditional expected payment, *not* to his expected payment conditional on winning. Given the risk-neutrality assumption, $\Pi_i(v_i)$ is given by:

$$\Pi_i(v_i) = P_i(v_i)v_i - Q_i(v_i). \tag{1}$$

Next, we pursue the observation that there is nothing forcing bidder *i* to bid according to his true valuation v_i rather than according to another valuation w_i. Define $\pi_i(w_i, v_i)$ to be bidder *i*'s expected payoff from employing the bidding strategy of a bidder with valuation w_i when his true valuation is v_i. Observe that:

$$\pi_i(w_i, v_i) = P_i(w_i)v_i - Q_i(w_i), \tag{2}$$

since bidder *i*'s probability of winning and expected payment depend exclusively on his bid, not on his true valuation. Bidder *i* will voluntarily choose to bid according to his true valuation only if his expected payoff is greater than from bidding according to another valuation w_i, that is, if:

$$\Pi_i(v_i) \geq \pi_i(w_i, v_i), \text{ for all } v_i, w_i \in [\underline{v}_i, \overline{v}_i] \text{ and all } i = 1, \ldots, n. \tag{3}$$

Inequality (3), referred to as the *incentive-compatibility constraint*, has very strong implications. Next, note that $\Pi_i(v_i) = \pi_i(v_i, v_i) = \max_{w_i \in [\underline{v}_i, \overline{v}_i]} \pi_i(w_i, v_i)$. It is straightforward to see that $\Pi_i(\cdot)$ is monotonically non-decreasing and continuous. Consequently, it is differentiable almost everywhere and equals the integral of its derivative. Applying the envelope theorem at any v_i where $\Pi_i(\cdot)$ is

differentiable yields:

$$\frac{d\Pi_i(v_i)}{dv_i} = \left.\frac{\partial \pi_i(w_i, v_i)}{\partial v_i}\right|_{w_i=v_i} = P_i(w_i)|_{w_i=v_i} = P_i(v_i). \tag{4}$$

Integrating eq. (4), we have:

$$\Pi_i(v_i) = \Pi_i(\underline{v}_i) + \int_{v_i}^{v_i} P_i(x)\,dx, \text{ for all } v_i \in [\underline{v}_i, \overline{v}_i] \text{ and all } i = 1, \ldots, n. \tag{5}$$

2.4 Solution of the first-price auction

The sealed-bid first-price auction requires two symmetry assumptions in order to yield a fairly simple solution. First, we assume *symmetric bidders*, in the sense that the joint distribution function $\hat{F}(v_1, \ldots, v_i, \ldots, v_n)$ governing the bidders' valuations is a symmetric function of its arguments. This assumption and the associated notation are simplest to state if independence is assumed. In this case, we write $F_i(\cdot)$ for the distribution function of each v_i; symmetry is the assumption that $F_i = F$, for all $i = 1, \ldots, n$, or, in other words, the assumption that the various v_i are identically distributed, as well as independent, random variables. However, a similar derivation with only slightly more cumbersome notation is possible if the bidders are symmetric but the v_i are affiliated random variables. We write $[\underline{v}, \overline{v}]$ for the support of $F(\cdot)$. In addition, we assume that $F(\cdot)$ is a continuous function, so that there are no mass points in the common probability distribution of the bidders' valuations.

Second, we restrict attention to *symmetric, monotonically increasing equilibria* in pure strategies. The assumed symmetry of bidders opens the possibility for existence of a symmetric equilibrium. (Meanwhile, asymmetric equilibria are also possible in symmetric games, but Maskin and Riley, 2003, establish that, under slightly stronger assumptions, the construction here gives the unique equilibrium of the auction.) Any pure-strategy equilibrium can be characterized by the bid functions $\{B_i(\cdot)\}_{i=1}^{n}$, which give bidder *i*'s bid $B_i(v_i)$ when his valuation is v_i. Our assumption is that $B_i = B$, for all i=1,•…•, n, where $B(\cdot)$ is a strictly increasing function.

Observe that, in any symmetric equilibrium, bidder *i* wins against bidder *j* if and only if $B(v_j) < B(v_i)$ and, given strict monotonicity, if and only if $v_j < v_i$. (We can ignore the event $v_j = v_i$; this is a zero-probability event, since we have assumed the distribution of valuations has no mass points.) Consequently, bidder *i* wins the item if and only $v_j < v_i$ for all $j \neq i$. Since the $\{v_j\}_{j \neq i}$ are i.i.d. random

variables, bidder *i* has probability $F(v_i)^{n-1}$ of winning the auction when his valuation is v_i. We write: $P_i(v_i) = F(v_i)^{n-1}$, for all $v_i \in [\underline{v}, \overline{v}]$ and all i=1 , … , n.
Moreover, in a first-price auction, the bidder's payoff equals $v_i - B(v_i)$ if he wins the auction and zero if he loses. Consequently his expected payoff equals:

$$\Pi_i(v_i) = P_i(v_i)[v_i - B(v_i)] = F(v_i)^{n-1}[v_i - B(v_i)]. \tag{6}$$

Observe from eq. (6) that, if $v_i = \underline{v}$, bidder *i*'s probability of winning equals zero and, hence, $\Pi_i(\underline{v}) = 0$. Substituting this fact and $P_i(v_i) = F(v_i)^{n-1}$ into Eq. (5) yields:

$$\Pi_i(v_i) = \int_{\underline{v}}^{v_i} F(x)^{n-1} dx, \text{ for all } v_i \in [\underline{v}, \overline{v}] \text{ and all } i = 1, \ldots, n. \tag{7}$$

Combining eq. (6) with eq. (7), and solving for $B(\cdot)$, yields the equilibrium bid function:

$$B(v_i) = v_i - \frac{\Pi_i(v_i)}{F(v_i)^{n-1}} = v_i - \frac{\int_{\underline{v}}^{v_i} F(x)^{n-1} dx}{F(v_i)^{n-1}}. \tag{8}$$

The posited strict monotonicity is verified by differentiating eq. (8) with respect to v_i, which shows that $B'(v_i) > 0$. Thus, eq. (8) provides us with the unique symmetric equilibrium in pure strategies of the sealed-bid first-price auction. This result holds for arbitrary continuous distribution functions $F(\cdot)$ with support on an interval $[\underline{v}, \overline{v}]$.

3 Revenue equivalence, efficient auctions and optimal auctions

Standard practice in auction theory is to evaluate auction formats according to either of two criteria: efficiency and revenue optimization. With the quasi-linear utilities generally assumed in auction theory, efficiency means putting the items in the hands of those who value them the most. Revenue

maximization means maximizing the seller's expected revenues or, in a procurement auction, minimizing the buyer's expected procurement costs. In auctions of government assets such as spectrum licenses, the explicit objective is often efficiency. In auctions by private parties, the explicit objective is often revenue optimization.

3.1 Efficient auctions

The above solutions to the second-price and first-price auctions both yield full efficiency. In the symmetric increasing equilibrium of the first-price auction, the highest bid corresponds to the highest valuation, and so the item is assigned efficiently for every realization of the random variables. In the dominant strategy equilibrium of the second-price auction, the identical conclusion holds. Thus, in a symmetric private values model, an objective of efficiency looks kindly upon both auction formats – but does not prefer one over the other.

3.2 Revenue equivalence

One of the classic and most far-reaching results in auction theory is revenue equivalence, which provides a set of assumptions under which the sellers' and buyers' expected payoffs are guaranteed to be the same under different auction formats.

Revenue equivalence (Vickrey, 1961; Myerson, 1981; Riley and Samuelson, 1981) may be stated as follows. Assume that the random variables representing the bidders' valuations are independent, and assume that bidders are risk neutral. Consider any two auction formats satisfying both of the following properties: (*a*) the two auction formats assign the item(s) to the same bidder(s), for every realization of random variables; and (*b*) the two auction formats give the same expected payoff to the lowest valuation type, $\underline{v}_i$, of each bidder *i*. Then each bidder earns the same expected payoff under each of the two auction formats and, consequently, the seller earns the same expected revenues under each of the two auction formats.

For an auction of a single item, the result follows directly from eq. (5) above. Recall that this equation holds for any equilibrium of *any* sealed-bid auction format. If for every realization of the random variables the two auction formats assign the item to the same bidder, then each bidder's probability, $P_i(\cdot)$, of winning is the same under the two auction formats. If in addition, $\Pi_i(\underline{v}_i)$ is the same under the two auction formats, then eq. (5) implies that the entire function $\Pi_i(\cdot)$ is the same under the two auction formats. Since this holds for every bidder *i*, and since the expected gains from trade are the same under the two auction formats, it follows from an accounting identity that the seller's expected revenues are also the same under the two auction formats.

One of the most important applications of revenue equivalence is that the above solutions to the second-price and first-price auctions give the seller the same expected revenues (and also give each buyer the same expected payoffs). Revenue equivalence is applicable because, as argued above, the item is assigned efficiently for every realization of the random variables in each of these auction formats. Moreover, when $v_i = \underline{v}$, the expected payoff of bidder *i* equals zero in each of these auction formats. To understand this result, observe that (all other things equal) a bidder in a first-price auction will bid lower than in a second-price auction, since the payment rule is less generous. Expected revenues will be

greater in the first-price or the second-price auction depending on whether the highest of a collection of smaller bids or the second-highest of a collection of larger bids is greater in expectation. The revenue equivalence theorem establishes that, in the symmetric private values model, the two effects exactly offset one another.

3.3 Optimal auctions

Another classic result of auction theory is the determination of the auction format that optimizes revenues. This result, known in the literature as the *optimal auction*, is due to Harris and Raviv (1981), Myerson (1981), and Riley and Samuelson (1981). Any possible auction format is considered – the item may be assigned to the bidder who submitted the highest bid (as in the second-price or first-price auction), but it may alternatively be allocated to another bidder, randomized in its allocation, or withheld from sale entirely, depending on the collection of bids submitted. At the outset, this might be viewed as a very complicated problem, since it requires selecting simultaneously the probability of winning and a payment that optimizes revenues. However, by using analysis similar to the treatment of incentive compatibility, above, it can be shown that the expected payment is determined up to a constant by the probability of winning. Consequently, the problem simplifies to determining the probability of each bidder winning (for every realization of the random variables) that optimizes revenues.

For symmetric bidders, each of whose distributions satisfies a regularity condition, a particularly simple characterization of the optimal auction can be obtained. Let $F(\cdot)$ be the distribution function of the valuation v_i of each bidder i, let $f(\cdot)$ be the associated density function and suppose that $v_i - \frac{1-F(v_i)}{f(v_i)}$ is strictly increasing in v_i for all $v_i \in [\underline{v}, \overline{v}]$. Then the optimal auction assigns the item to the bidder i with the highest v_i, if and only if the highest v_i exceeds the reserve valuation r, where r is defined by $r - \frac{1-F(r)}{f(r)} = v_0$ and where v_0 is the seller's valuation for the item.

In other words, with symmetric bidders, both the second-price and the first-price auctions become optimal auctions, once a reserve price of r is inserted.

3.4 Full rent extraction

The optimal auctions problem can be reconsidered without the independence assumption. However, Crémer and McLean (1985) demonstrate that, if the bidders' private information is correlated, then there exists a mechanism that enables the seller to extract *all* of the gains from trade. The mechanism includes a procedure for allocating the item efficiently. Superimposed on this, the mechanism provides rewards to bidders if their reports of private information 'agree' with each other, and penalties to bidders if their reports 'disagree' with each other. The amounts of the rewards and penalties – both potentially quite large – are set so as to make the bidders indifferent between participating and not participating in the mechanism. As such, the mechanism enables the seller to extract the entire surplus, including the informational rents that the bidders are able to obtain under the independence assumption. This is referred to as *full rent extraction*.

Crémer and McLean's result may be viewed as fundamentally negative, in that it suggests that the

optimal auctions analysis may be of limited relevance. Real-world auction mechanisms appear to be broadly consistent with the predictions of the optimal auctions theory under the independence assumption, but they look nothing like the full rent-extracting mechanisms possible with correlated private information. Given that there are good reasons to believe that bidders' private signals are correlated with one another, it would appear that the optimal auctions analysis does not provide us with great insight into real-world auctions. Some subsequent research has attempted to weaken the extreme conclusion of full rent extraction by positing that bidders have limited liability or by introducing opportunities for auctioneer collusion or cheating, but in many respects these devices appear to be ineffectual patches for an elegant theory (optimal auctions) that suffers from only limited empirical relevance.

4 Dynamic auctions for single items

The next two formats considered for auctioning single items are dynamic auctions: participants bid sequentially over time and, potentially, learn something about their opponents' bids during the course of the auction. In the first dynamic auction, the price *ascends*; and in the second dynamic auction, the price *descends*:

- *English auction*: bidders dynamically submit successively higher bids for the item. The final bidder wins the item, and pays the amount of his final bid.
- *Dutch auction*: the auctioneer starts at a high price and announces successively lower prices, until some bidder expresses his willingness to purchase the item by bidding. The first bidder to bid wins the item, and pays the current price at the time he bids.

Note that, as in Section 2, each of these auction formats has been described for a regular auction in which the auctioneer offers items for sale, but can easily be restated for a 'reverse auction'. For example, in an English reverse auction the bids would descend rather than ascend, while in a Dutch reverse auction the auctioneer would offer to buy at successively higher prices.

4.1 Solution of the Dutch auction

An insight due to Vickrey (1961) is that the Dutch auction is strategically equivalent to the sealed-bid first-price auction. To see the equivalence, consider the real meaning of a strategy b_i by bidder i in the Dutch auction: 'If no other bidder bids for the item at any price higher than b_i, then I am willing to step in and purchase it at b_i.' Just as in the sealed-bid first-price auction, the bidder i who selects the highest strategy b_i in the Dutch auction wins the item and pays the amount b_i. Furthermore, although the Dutch auction is explicitly dynamic, there is nothing that can happen that would lead any bidder to want to change his strategy while the auction is still running. If strategy b_i was a best response for bidder i evaluated at the starting price p_0, then b_i remains a best response evaluated at any price $p<p_0$, on the assumption that no other bidder has already bid at a price between p_0 and p. Meanwhile, if another bidder has already bid, then there is nothing that bidder i can do; the Dutch auction is over. Hence, any

equilibrium of the sealed-bid first-price auction is also an equilibrium of the Dutch auction, and vice versa.

4.2 Solution of the English auction

By way of contrast, some meaningful learning and/or strategic interaction is possible during an English auction, so the outcome is potentially different from the outcome of the sealed-bid second-price auction. We model the English auction as a ‘clock auction’: the auctioneer starts at a low price and announces successively higher prices. At every price, each bidder is asked to indicate his willingness to purchase the item. The price continues to rise so long as two or more bidders indicate interest. The auction concludes at the first price such that fewer than two bidders indicate interest, and the item is awarded at the final price. This clock-auction description is used instead of a game where bidders successively announce higher prices, since it yields simpler arguments and clean results.

With pure private values, the reasonable equilibrium of the English auction corresponds to the dominant-strategy equilibrium of the sealed-bid second-price auction. A bidder's strategy designates the price at which he will drop out of the auction (on the assumption that at least one opponent still remains); in equilibrium, the bidder sets his drop-out price equal to his true valuation. However, matters become more complicated in the case of interdependent valuations, where each bidder's valuation depends not only on his own information, v_i, but also on the opposing bidders’ information, v_{-i}. We turn to this case next.

4.3 The winner's curse and revenues under interdependent values

One of the most celebrated phenomena in auctions is the ‘winner's curse’. Whenever a bidder's valuation depends positively on other bidders’ information, winning an item in an auction may confer ‘bad news’ in the sense that it indicates that other bidders possessed adverse information about the item's value. The potential for falling victim to the winner's curse may induce restrained bidding, curtailing the seller's revenues. In turn, some auction formats may produce higher revenues than others, to the extent that they mitigate the winner's curse and thereby make it safe for bidders to bid more aggressively.

The basic intuition, which is often referred to as the ‘linkage principle’ and is due to Milgrom and Weber (1982), is that the winner's curse is mitigated to the extent that the winner's payment depends on the opposing bidders’ information. Thus, under appropriate assumptions, the second-price auction will yield higher expected revenues than the first-price auction: the price paid by the winner of a second-price auction depends on the information possessed by the highest losing bidder, while the price paid by the winner of a first-price auction depends exclusively on his own information. Moreover, the English auction will yield higher expected revenues than the second-price auction: the price paid by the winner of an English auction may depend on the information possessed by *all* of the losing bidders (who are observed as they drop out), while the price paid by the winner of a (sealed-bid) second-price auction depends only on the information of the highest losing bidder.

These conclusions require an assumption known as ‘affiliation’, which intuitively means something very close to ‘non-negative correlation’. More precisely, let $v = (v_1, \ldots, v_n)$ and $v' = (v'_1, \ldots, v'_n)$ be possible realizations of the *n* bidders’ random variables, and let $f(\cdot, \ldots, \cdot)$ denote the joint density function. Let

$v \vee v'$ denote the component-wise maximum of v and v', and let $v \wedge v'$ denote the component-wise minimum. The random variables v and v' are said to be *affiliated* if:

$$f(v \vee v')\, f(v \wedge v') \geq f(v)\, f(v'), \quad \text{for all } v, v' \in [\underline{v}, \ldots \overline{v}]^n. \tag{9}$$

Affiliation provides that two high realizations or two low realizations of the random variables are at least as likely as one high and one low realization, and so on, meaning something close to non-negative correlation. Independence is included (as a boundary case) in the definition: for independent random variables, the affiliation inequality (9) is satisfied with equality. To obtain strict revenue rankings, the affiliation inequality must hold strictly.

These conclusions also rely on several symmetry assumptions. Bidders are symmetric, the equilibria considered are symmetric, and each bidder's valuation depends on all of its opponents' information in a symmetric way. Each bidder's valuation increases (weakly) in its own and its opponents' information, and attention is restricted to equilibria in monotonically increasing strategies. As before, each bidder is risk neutral in evaluating its payoff under uncertainty.

These conclusions also rely on a monotonicity assumption: each bidder's valuation increases (weakly) in its own and in the opposing bidders' information. In addition, as before, each bidder is risk-neutral in evaluating its payoff under uncertainty. Furthermore, the two symmetry assumptions of Section 2.4 are made: bidders are symmetric in the sense that the joint distribution governing the bidders' information is a symmetric function of its arguments; and attention is restricted to symmetric, monotonically increasing equilibria in pure strategies.

Under these assumptions, the sealed-bid first-price and second-price auctions and the English auction possess symmetric, monotonic equilibria. However, while these equilibria are all efficient, Milgrom and Weber (1982) establish that they may be ranked by revenues: the English auction yields expected revenues greater than or equal to those of the sealed-bid second-price auction, which in turn yields expected revenues greater than or equal to those of the sealed-bid firstvprice auction. Their theorem provides one of the most powerful results of auction theory, justifying the conventional wisdom that dynamic auctions yield higher revenues than sealed-bid auctions.

5 Auctions of homogeneous goods

5.1 Sealed-bid, multi-unit auction formats

The defining characteristic of a homogeneous good is that each of the *M* individual items is identical (or a close substitute), so that bids can be expressed in terms of quantities without indicating the identity of the particular good that is desired. Treating goods as homogeneous has the effect of dramatically simplifying the description of the bids that are submitted and the overall auction procedure. This simplification is especially appropriate in treating subject matter such as financial securities or energy

products. Any two $10,000 US government bonds with the same interest rate and the same maturity are identical, just as any two megawatts of electricity provided at the same location on the electrical grid at the same time are identical.

There are three principal sealed-bid, multi-unit auction formats for *M* homogeneous goods. In each of these, a bid comprises an inverse demand function, that is, a (weakly) decreasing function $p_i(q)$, for $q \in [0, M]$, representing the price offered by bidder *i* for a first, second, and so on, unit of the good. (Note that this notation may be used to treat situations where the good is perfectly divisible, as well as situations where the good is offered in discrete quantities.) The bidders submit bids; the auctioneer then aggregates the bids and determines a clearing price. Each bidder wins the quantity demanded at the clearing price, but his payment varies according to the particular auction format:

- *Pay-as-bid auction*. Each bidder wins the quantity demanded at the clearing price, and pays the amount that he bid for each unit won.
- *Uniform-price auction*. Each bidder wins the quantity demanded at the clearing price, and pays the clearing price for each unit won.
- *Multi-unit Vickrey auction*. Each bidder wins the quantity demanded at the clearing price, and pays the opportunity cost (relative to the bids submitted) for each unit won.

(Pay-as-bid auctions are also known as ‘discriminatory auctions’ or ‘multiple-price auctions’. Uniform-price auctions are often referred to in the financial press as ‘Dutch auctions’, generating some confusion with respect to the standard usage of the auction theory literature. They are also known as ‘nondiscriminatory auctions’, ‘competitive auctions’ or ‘single-price auctions’.)

Sealed-bid, multi-unit auction formats are best known in the financial sector for their long-time and widespread use in the sale of government securities. For example, a survey of OECD countries in 1992 found that Australia, Canada, Denmark, France, Germany, Italy, Japan, New Zealand, the United Kingdom and, of course, the United States then used sealed-bid auctions for selling at least some of their debt. The pay-as-bid auction was the traditional format used for US Treasury bills, as well as for government securities of most other countries. The uniform-price auction was first proposed seriously as a replacement for the pay-as-bid auction by Milton Friedman in testimony at a 1959 Congressional hearing. Wilson (1979) gave the first theoretical analysis of a uniform-price auction. In 1993 the United States began an ‘experiment’ of using the uniform-price auction for two- and five-year government notes and, beginning in 1998, the United States switched entirely to the uniform-price auction for all issues. Meanwhile, the multi-unit Vickrey auction was introduced and first analysed in Vickrey's 1961 paper.

The pay-as-bid auction can be correctly viewed as a multi-unit generalization of the first-price auction. However, it is quite difficult to calculate Nash equilibria of the pay-as-bid auction, unless efficient equilibria exist. *Three* symmetry assumptions together guarantee the existence of efficient equilibria. First, bidders are assumed to be symmetric, in the sense that the joint distribution governing the bidders’ information is symmetric with respect to the bidders. Second, bidders regard every unit of the good as symmetric: that is, each bidder *i* has a constant marginal valuation for every quantity $q_i \in [0, \lambda_i]$, up to a capacity of λ_i, and a marginal valuation of zero thereafter. Third, the bidders are symmetric in their capacities: that is, $\lambda_i = \lambda$, for all bidders *i*. With these assumptions, the pay-as-bid auction has a solution

very similar to that of the first-price auction for a single item. However, without these assumptions, it inherits an undesirable property from the single-item auction: absent symmetry, all Nash equilibria of the pay-as-bid auction will generally be inefficient (Ausubel and Cramton, 2002, Theorems 3 and 4).

The uniform-price auction bears a superficial resemblance to the second-price auction of a single item, in that a high winning bid gains the benefit of a lower marginal bid. However, any similarity is indeed only superficial as, except under very restrictive assumptions, all equilibria of the uniform-price auction are inefficient. The argument is simplest in the same model of constant marginal valuations as in the previous paragraph. If the capacities of all bidders are equal (that is, if $\lambda_i = \lambda$ for all *i*) and if the supply is an integer multiple of λ, then there exists an efficient Bayesian-Nash equilibrium of the uniform-price auction. (For example, if there are *M* identical units available and if every bidder has a unit demand, then sincere bidding is a Nash equilibrium in dominant strategies.) However, if the bidders' capacities are unequal or if the supply is not an integer multiple of λ, then all equilibria of the uniform-price auction are inefficient (Ausubel and Cramton, 2002, Theorems 2 and 5).

The intuition for inefficiency in the uniform-price auction can be found by taking a close look at optimal bidding strategies. Sincere bidding is weakly dominant for a *first* unit: if a bidder's first bid determines the clearing price, then the bidder wins zero units. However, the bidder's *second* bid may determine the price he pays for his first unit, providing an incentive to shade his bid. The extent of *demand reduction*, as this bid shading is known, increases in the number of units, since the number of infra-marginal units whose price may be affected increases. Further, note that the allocation rule in the auction has the effect of equating the amounts of the bidders' marginal bids. Since a large bidder will likely have shaded his marginal bid more than a small bidder, the large bidder's marginal value is probably greater than a small bidder's. Consequently, the bidders' marginal values will be unequal, contrary to efficiency.

Meanwhile, the Vickrey auction is the correct multi-unit generalization of the second-price auction. As in the pay-as-bid and uniform-price auctions, bidders simultaneously submit inverse demand functions and each bidder wins the quantity demanded at the clearing price. However, rather than paying the bid price or the clearing price for each unit won, a winning bidder pays the *opportunity cost*. If a bidder wins *K* units, he pays the *K*th highest rejected bid of his opponents for his first unit, the (*K*–1)st highest rejected bid of his opponents for his second unit, … , and the highest rejected bid of his opponents for his *K*th unit. The dominant strategy property of the sealed-bid second-price auction generalizes because a bidder's payment is determined solely by his opponents' bids. Consequently, given pure private values and non-increasing marginal values, sincere bidding is an efficient equilibrium in weakly dominant strategies.

5.2 Efficiency and revenue comparisons

Under pure private values, the dominant strategy equilibrium of the Vickrey auction attains full efficiency. It can be shown that neither the pay-as-bid nor the uniform-price auction generally attains efficiency; moreover, the efficiency ranking of these two formats is inherently ambiguous. To continue the argument of the previous subsection, it is sufficient to examine environments in which bidders have constant marginal valuations. If $F_i = F$ and $\lambda_i = \lambda$ for all bidders *i*, but the supply is *not* an integer multiple of λ, then the pay-as-bid auction has an efficient equilibrium while all equilibria of the uniform-price auction are inefficient. Conversely, if $\lambda_i = \lambda$ for all bidders *i* and if the supply is an integer

multiple of λ, but $F_i \neq F_j$ for two bidders *i* and *j*, then the uniform-price auction has an efficient equilibrium while all equilibria of the pay-as-bid auction are generally inefficient (Ausubel and Cramton, 2002).

On revenues, the policy literature has generally assumed that the uniform-price auction outperforms the pay-as-bid auction; however, the argument of the previous paragraph can be extended to reverse the assumed ranking. Maskin and Riley (1989) extend Myerson's (1981) characterization of the optimal auction to multiple homogeneous goods: with symmetric bidders and constant marginal valuations, their characterization requires allocating items efficiently. Thus, as in the previous paragraph, if $F_i = F$ and $\lambda_i = \lambda$ for all bidders *i*, but the supply is *not* an integer multiple of λ, then the efficient equilibrium of the pay-as-bid auction outranks all equilibria of the uniform-price auction on *revenues* (as well as efficiency).

5.3 Uniform-price clock auctions

The ‘clock auction’ – a practical design for dynamic auctions of one or more types of goods, with its origins in the ‘Walrasian auctioneer’ from the classical economics literature – has seen increasing use as a trading institution since 2001. A fictitious auctioneer is often presented as a device or thought experiment for understanding convergence to a general equilibrium. The Walrasian auctioneer announces a price vector, *p*; bidders report the quantity vectors that they wish to transact at these prices; and the auctioneer increases or decreases each component of price according as excess demand is positive or negative (*Walrasian tâtonnement*). This iterative process continues until a price vector is reached at which excess demand is zero, and trades occur only at the final price vector. In real-world applications, instead of a fictitious auctioneer serving as a metaphor for a market-clearing process, the process is taken literally; a real auctioneer announces prices and accepts bids of quantities. Applications, to date, have largely been in the electricity, natural gas, and environmental sectors.

The basic clock auction differs from the standard Sotheby's or eBay auction in that bidders do not propose prices. Rather, the auctioneer announces prices, and bidders’ responses are limited to the reporting of quantities desired at the announced prices, until clearing is attained. As such, it is closest to the auction-theorist's depiction of the English auction for a single item (or the traditional Dutch auction), but generalized, so that, instead of bidders merely giving binary responses of whether they are ‘in’ or ‘out’ as prices ascend, they indicate their quantities desired.

Observe that the uniform-price clock auction is correctly viewed as a dynamic version of the sealed-bid uniform-price auction reviewed in the previous two subsections. The important difference is that, in the dynamic auction, bidders will typically receive repeated feedback as to the aggregate demand at the various prices.

As such, the clock auction may inherit the advantages that dynamic auctions have over sealed-bid auctions. First, under conditions that can be made precise, the insight from single-item auctions that feedback about other bidders’ valuations would ameliorate the winner's curse and lead to more aggressive bidding carries over to the multi-unit environment. Second, clock auctions, better than sealed-bid auctions, allow bidders to maintain the privacy of their valuations for the items being sold. Bidders never need to submit any indications of interest at any prices beyond the auction's clearing price. Third, when there are two or more types of items, auctioning them simultaneously enables bidders to submit

bids based on the substitution possibilities or complementarities among the items at various price vectors. At the same time, the iterative nature of the auction economizes on the amount of information submitted: demands do not need to be submitted for all price vectors, but only for price vectors reached along the convergence path to equilibrium.

Unfortunately, the uniform-price clock auction also inherits the demand reduction and inefficiency of the sealed-bid uniform-price auction. Indeed, as a theoretical proposition, the problem of bidders optimally reducing their quantities bid well below their true demands can become substantially worse in the dynamic version of the auction. The *reductio ad absurdum* is provided by Ausubel and Schwartz (1999), who analyse a two-bidder clock auction game of complete information in which the bidders alternate in their moves. For a wide set of environments, the unique subgame perfect equilibrium has the qualitative description that, at the first move, the first player reduces his quantity to approximately half of the supply and, at the second move, the second player reduces his quantity to clear the market. Thus, the outcome is inefficient and the revenues barely exceed the starting price.

As a practical matter, demand reduction may not undermine the outcome of a uniform-price clock auction where there is substantial competition for every item being sold. However, if one or more of the bidders has considerable market power, it may become important to use an auction format which avoids creating incentives for demand reduction.

5.4 Efficient clock auctions

Ausubel (2004; 2006) proposes an alternative clock auction design, which utilizes the same general structure as the uniform-price clock auction, but adopts a different payment rule that eliminates the incentives for demand reduction. In essence, the design provides a dynamic version of the (multi-unit) Vickrey auction, and thereby inherits its incentives for truth-telling.

The Ausubel auction is easiest described for a homogeneous good. After each set of bidder reports, the auctioneer determines whether any bidder has ‘clinched’ any of the units offered (that is, whether any bidder is mathematically guaranteed to win one or more units). For example, in an auction with a supply of 5 units, and three bidders demanding 3, 2 and 2 units, respectively, the first bidder has clinched 1 unit, as his opponents’ total demand of 4 is less than the supply of 5. Rather than awarding units only at a final uniform price, the auction awards units at the current price whenever they are newly clinched.

If this alternative clock auction is represented as a static auction, it collapses to the Vickrey auction in the same sense that an English auction collapses to the sealed-bid second-price auction. Consequently, it can be proven that sincere bidding is an equilibrium and, in a suitable discrete specification of the game under incomplete information, sincere bidding is the unique outcome of iterated elimination of weakly dominated strategies. Thus, unlike the uniform-price clock auction, there is no incentive for demand reduction.

6 Auctions of heterogeneous goods

In many significant applications, the multiple items offered within an auction are each unique, so it is not adequate for bidders merely to indicate the quantities that they desire. For example, an FCC spectrum auction might include a New York licence, a Washington licence and a Los Angeles licence.

Moreover, there might be synergies in owning various combinations: for example, a New York and a Washington licence together might be worth more together than the sum of their values separately. Such environments pose particular challenges for auction theory.

6.1 Simultaneous ascending auctions

The simultaneous ascending auction, proposed in comments to the FCC by Paul Milgrom, Robert Wilson and Preston McAfee, has been used in auctions on six continents allocating more than $100 billion worth of spectrum licenses. Some of the best known applications of the simultaneous ascending auction include: the Nationwide Narrowband Auction (July 1994), the first use of the simultaneous ascending auction; the PCS A/B Auction (December 1994–March 1995), the first large-scale auction of mobile telephone licences, which raised $7 billion; the United Kingdom UMTS Auction (March–April 2000), which raised 22.5 billion British pounds; and the German UMTS Auction (July–August 2000), which raised 50 billion euro.

In the simultaneous ascending auction, multiple items are put up for sale at the same time and the auction concludes simultaneously for all of the items. As such, it is a modern version of the ‘silent auction’ that is frequently used in fundraisers by charitable institutions. Bidders submit bids in a sequence of rounds. Each bid comprises a single item and an associated price, which must exceed the standing high bid by at least a minimum bid increment. After each round, the new standing high bids for each item are determined. The auction concludes after a round passes in which no new bids are submitted, and the standing high bids are then deemed to be winning bids. Payments equal the amounts of the winning bids.

The critical innovation in the simultaneous ascending auction is the inclusion of *activity rules* into the auction design. Activity rules are bidding constraints that limit a bidder's bidding activity in the current round based on his past bidding activity (that is, his standing high bids and new bids). Without activity rules, bidders would tend to wait as ‘snakes in the grass’ until nearly the end of the auction before placing their serious bids, thwarting any price discovery (the main reason for conducting a dynamic auction in the first place). Conversely, activity rules have the effect of forcing bidders to place meaningful bids in early rounds of the auction and thereby to reveal information to their opponents.

6.2 Walrasian equilibria as outcomes of simultaneous ascending auctions

A Walrasian equilibrium – consisting of prices for the various items and an allocation of the items to the bidders such that each item with a non-zero price is assigned to exactly one bidder and such that each bidder prefers his assigned allocation to any alternative bundle at the given prices – is a plausible outcome for the simultaneous ascending auction. On the assumption that a Walrasian equilibrium was reached, no bidder would have any incentive to attempt to upset the allocation, even if he believed he could obtain additional items without further increasing their prices. Thus, it becomes interesting to identify the conditions needed for existence of Walrasian equilibria with discrete items.

Kelso and Crawford (1982) show that the substitutes condition is sufficient for the existence of Walrasian equilibrium. ‘Substitutes’ literally refers to the price-theoretic condition that if the price of one item is increased while the price of every other item is held fixed, then the demand for every other item weakly increases. Moreover, the substitutes condition is ‘almost necessary’ for existence. Suppose

that the set of possible bidder preferences includes all valuation functions satisfying the substitutes condition, but also includes at least one valuation function violating the substitutes condition. Then if there are at least two bidders, there exists a profile of valuation functions such that no Walrasian equilibrium exists (Gul and Stacchetti, 1999; Milgrom, 2000).

The reader should avoid losing sight of the fact that, just because a Walrasian equilibrium exists for a discrete environment, it does not necessarily follow that the simultaneous ascending auction will terminate at a Walrasian equilibrium. The strongest statement that can be made is that, if bidders bid ‘straightforwardly’ (that is, if they demand naively the bundle of items that maximizes their utility, while ignoring strategic considerations), then a Walrasian equilibrium will be reached. However, observe that, even with homogeneous goods, consumers with weakly diminishing marginal valuations satisfy the substitutes condition. Nonetheless, the uniform-price auction is susceptible to demand reduction – meaning that bidders are likely to reduce their demands and thereby end the auction before reaching a Walrasian equilibrium. Indeed, we know from the Fundamental Theorem of Welfare Economics that the Walrasian equilibrium is efficient, so that any conclusion of inefficiency in a uniform-price auction implies that the outcome must be non-Walrasian.

6.3 Static pay-as-bid combinatorial auctions

Let us consider an example with two bidders, 1 and 2, and two items, A and B, where the substitutes condition is not satisfied and the existence of Walrasian equilibrium fails. Bidder 1 has a valuation of 3 for the package of A and B, but has a valuation of 0 for each item separately. (Thus, for Bidder 1, the goods are complements – not substitutes.) Bidder 2 has a valuation of 2 for item A, 2 for item B, and only 2 for the package of A and B. The efficient allocation assigns both items to Bidder 1. Consequently, any Walrasian equilibrium (if it exists) must assign both items to Bidder 1. However, to dissuade Bidder 2 from purchasing either item, the prices p_A and p_B of items A and B, respectively, must satisfy $p_A>2$ and $p_B>2$. Consequently, $p_A+p_B>4$, exceeding Bidder 1's valuation for the package of two items and yielding a contradiction.

Given the argument of the previous paragraph, we should not expect the simultaneous ascending auction – or any auction format with bids for individual items – to generate the efficient allocation in this example. Bidder 1's dilemma is often referred to as the exposure problem: a bidder may refrain from bidding more than his stand-alone valuations for each of the individual items, knowing that, if he is outbid on some of the individual items, he will remain ‘exposed’ as the high bidder on the remaining items. This may prevent the available synergies from being realized. Indeed, if Bidder 1 understands this example, he may be unwilling to bid any positive price for either item, since Bidder 2 is sure to win one of the items, and therefore Bidder 1 would obtain zero value from the item that he wins.

The exposure problem can be avoided by using a *combinatorial auction*. The rules are modified to permit bidders to place *package bids*, each comprising a *set* of items and a price. For example, the bid ({A, B}, p) is interpreted as an all-or-nothing offer in the amount of p for the package of A and B – with no requirement that the bidder is willing to accept a part of the package for a part of the price. The allocation is determined by a combination of compatible bids that maximizes the seller's revenues. In this example, Bidder 2 is unwilling to bid any more than 2 for any combination of items, while Bidder 1 is able to exceed 2 for {A, B}. Consequently, the solution has Bidder 1 receiving both items, the

efficient allocation.

To the extent that bidders value some of the items in the auction as substitutes, then it may be important for any two bids by the same bidder to be treated as *mutually exclusive*. For example, Bidder 2 in the above example may have been willing to bid 1.5 for item A and 1.5 for item B – but *not* if there was a significant risk that both bids would be accepted. This difficulty is avoided if the auction rules permit at most one of his bids to be accepted. (Such mutually exclusive bids are sometimes referred to as 'XOR' bids.) Observe that a rule of mutual exclusivity is *fully expressive* in the sense that it enables the bidder to express any arbitrary preferences. For example, if Bidder 2 in the above example wished to allow both of his bids to be accepted, he could effectively opt out of the mutual exclusivity by submitting a third bid comprising the package {A, B} at a price of 3.

In a static pay-as-bid combinatorial auction, each bidder simultaneously and independently submits a collection of package bids. The auctioneer then solves the *winner determination problem*: find a combination of bids (at most one from each bidder) that maximizes the seller's revenues subject to the constraint that each item can be allocated to at most one bidder. The submitter of each bid selected in the winner determination problem wins the items specified in the bid and pays the amount of the bid. Rassenti, Smith and Bulfin (1982) are credited with the first experimental study of combinatorial auctions. They studied a static combinatorial auction treating the problem of allocating airport time slots, a natural application given that landing and takeoff slots are strong complements. Bernheim and Whinston (1986) provided an important characterization of equilibria of static pay-as-bid combinatorial auctions under complete information.

6.4 The Vickrey – Clarke – Groves (VCG) mechanism

Just as the payment rule of a pay-as-bid auction for a single item or for homogeneous goods can be modified to be 'second-price', an analogous modification can be done in the case of a combinatorial auction for heterogeneous goods. This generalization is due to Clarke (1971) and Groves (1973). Let N be an arbitrary finite set of items and let L be the set of bidders. In the *Vickrey–Clarke–Groves (VCG) mechanism*, each bidder $\ell \in L$ submits $2^{|N|}$ package bids, for all subsets of set N. After the bids are submitted, the auctioneer finds a solution, $(x_\ell)_{\ell \in L}$, to the winner determination problem. While bidder • is allocated the subset $x_\ell \subset N$, he does not pay his bid $b_\ell(x_\ell)$. Rather, his payment $v_\ell \in \mathbb{R}$ is calculated so that $b_\ell(x_\ell) - v_\ell = R^*(L) - R^*(L \setminus \ell)$, where $R^*(L)$ denotes the maximized revenue of the winner determination problem with bidder • present and $R^*(L/\bullet)$ denotes the maximized revenue of the winner determination problem with bidder • absent. With sincere bidding, each bid $b_\ell(x_\ell)$ corresponds to the bidder's valuation $v_\ell(x_\ell)$, and $R^*(L)$ corresponds to the (maximized) social surplus. Thus, bidder • is allowed a payoff equaling the *incremental surplus* that he brings to the auction. As in the Vickrey auction for homogeneous goods, a bidder's payment thus equals the opportunity cost of assigning the items to the bidder.

Applied to a setting with a single item, observe that the VCG mechanism reduces to the sealed-bid second-price auction. Applied to a setting of homogeneous goods and non-increasing marginal valuations, the VCG mechanism reduces to the (multi-unit) Vickrey auction. By the same reasoning as before, the dominance properties of these special cases extend to the setting with heterogeneous items: if bidders have pure private values, sincere bidding is a weakly dominant strategy for every bidder,

yielding an efficient allocation.

6.5 Dynamic combinatorial auctions

In auctions for a single item, we have seen that a close relationship exists between a dynamic procedure with a pay-as-bid payment rule (that is, the English auction) and a static procedure with a second price rule (that is, the sealed-bid second-price auction). Furthermore, for homogeneous goods with non-increasing marginal values, an analogous relationship holds between the dynamic Ausubel auction and the static Vickrey auction. An important question for heterogeneous goods is the extent to which outcomes of a dynamic combinatorial auction with a pay-as-bid rule map to the static VCG mechanism. Banks, Ledyard and Porter (1989) conducted an early and influential study of dynamic combinatorial auctions. They defined several alternative sets of rules for the auction, developing some theoretical results and conducting an experimental study. Other important contributions have included Parkes and Ungar (2000), who independently provided a formulation of the ascending proxy auction described below, and Kwasnica et al. (2005).

Ausubel and Milgrom (2002) give two formulations of a combinatorial auction and use them to provide a partial answer to the relationship between dynamic combinatorial auctions and the VCG mechanism:

- *Ascending package auction*. Bidders submit package bids in a sequence of bidding rounds. Each new bid must exceed the bidder's prior bids for the same package by at least a minimum bid increment. After each round, the winner determination problem is solved, on all past and present bids, to determine a provisional allocation and provisional payments. The auction concludes after a round in which no new bids are submitted.
- *Ascending proxy auction*. Each bidder enters his valuations for the various packages into a *proxy bidder*. The proxy bidders then bid on behalf of the bidders in an ascending package auction in which the minimum bid increment is taken arbitrarily close to zero.

The second formulation may be viewed both as a new auction format which greatly speeds the progress of the auction, as well as a modelling device for obtaining results about the first formulation. While the first formulation is an extremely complicated dynamic game, efficiency results and a partial equilibrium characterization are available for the second formulation.

A bidder • in the ascending proxy auction is said to bid *sincerely* if he submits his true valuation, $v_{\bullet}(S)$, for every package $S \subset N$; and he is said to bid *semi-sincerely* if he submits his true valuation less a positive constant, $v_{\bullet}(S) - c$, where the same constant c is used for all packages S with valuations of at least c. The following results refer to the coalitional form game (with transferable utility) corresponding to the package economy: the value of any coalition that includes the seller is the total value associated with an efficient allocation among the buyers in the coalition; and the value of any coalition without the seller equals zero. The *core* is defined as the set of all payoff allocations that are feasible and upon which no coalition of players can improve.

Ausubel and Milgrom (2002) establish that the payoff allocation from the ascending proxy auction, given any reported preferences, is an element of the core (relative to the reported preferences). Furthermore, for any payoff vector π that is a bidder-Pareto-optimal point in the core, there exists a

Nash equilibrium of the ascending proxy auction with associated payoff vector π. Conversely, for any Nash equilibrium in semi-sincere strategies at which losing bidders bid sincerely, the associated payoff vector is a bidder-Pareto-optimal point in the core.
Furthermore, the set of all economic environments essentially dichotomizes into two cases. First, if all bidders' preferences satisfy the substitutes condition, then a single point in the core dominates all other points in the core for every bidder, and it equals the payoff vector from the Vickrey–Clarke–Groves mechanism. Thus, in this first case, the outcome of the ascending proxy auction coincides with the outcome of the VCG mechanism. Second, if at least one bidder's preferences violate the substitutes condition, then there exists an additive preference profile for the remaining bidders such that there is more than one bidder-Pareto-optimal point in the core. In this second case, the VCG payoff vector is *not* an element of the core; and the low revenues of the VCG mechanism may become problematic.

7 Conclusion

The proportion of goods and services transacted by auction processes has dramatically increased in recent years and is likely to increase further, making the understanding of auctions and the improvement of their designs increasingly important. At the same time, auctions will remain one of the most useful test beds for game theory, since the rules of the game are better defined than in most other markets. Consequently, auction theory will almost certainly continue to be a central area of study in economics.

See Also

- auctions (applications)
- auctions (empirics)
- auctions (experiments)
- incentive compatibility
- mechanism design
- Vickrey, William Spencer

Bibliography

Ausubel, L.M. 2004. An efficient ascending-bid auction for multiple objects. *American Economic Review* 94, 1452–75.

Ausubel, L.M. 2006. An efficient dynamic auction for heterogeneous commodities. *American Economic Review* 96, 602–29.

Ausubel, L.M. and Cramton, P. 2002. Demand reduction and inefficiency in multi-unit auctions. Working Paper 96–07, Department of Economics, University of Maryland.

Ausubel, L.M. and Milgrom, P.R. 2002. Ascending auctions with package bidding. *Frontiers of*

Theoretical Economics 1(1), Article 1.

Ausubel, L.M. and Schwartz, J. 1999. The ascending auction paradox. Mimeo, University of Maryland.

Banks, J.S., Ledyard, J.O. and Porter, D.P. 1989. Allocating uncertain and unresponsive resources: an experimental approach. *RAND Journal of Economics* 20, 1–25.

Bernheim, B.D. and Whinston, M. 1986. Menu auctions, resource allocation and economic influence. *Quarterly Journal of Economics* 101, 1–31.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11, 17–33.

Cramton, P., Shoham, Y. and Steinberg, R. 2006. *Combinatorial Auctions*. Cambridge, MA: MIT Press.

Crémer, J. and McLean, R.P. 1985. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica* 53, 345–61.

Groves, T. 1973. Incentives in teams. *Econometrica* 41, 617–31.

Gul, F. and Stacchetti, E. 1999. Walrasian equilibrium with gross substitutes. *Journal of Economic Theory* 87, 95–124.

Harris, M. and Raviv, A. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49, 1477–99.

Kelso, A.S., Jr. and Crawford, V.P. 1982. Job matching, coalition formation, and gross substitutes. *Econometrica* 50, 1483–504.

Klemperer, P. 2000. *The Economic Theory of Auctions*. Cheltenham: Edward Elgar.

Krishna, V. 2002. *Auction Theory*. San Diego: Academic Press.

Kwasnica, A.M., Ledyard, J.O., Porter, D. and DeMartini, C. 2005. A new and improved design for multiobject iterative auctions. *Management Science* 51, 419–34.

Maskin, E. and Riley, J. 1989. Optimal multi-unit auctions. In *The Economics of Missing Markets, Information, and Games*, ed. F. Hahn. Oxford: Oxford Univ. Press.

Maskin, E. and Riley, J. 2003. Uniqueness of equilibrium in sealed high-bid auctions. *Games and Economic Behavior* 45, 395–409.

McAfee, R.P. and McMillan, J. 1987. Auctions and bidding. *Journal of Economic Literature* 25, 699–738.

Milgrom, P. 2000. Putting auction theory to work: the simultaneous ascending auction. *Journal of Political Economy* 108, 245–72.

Milgrom, P. 2004. *Putting Auction Theory to Work*. Cambridge: Cambridge University Press.

Milgrom, P.R. and Weber, R.J. 1982. A theory of auctions and competitive bidding. *Econometrica* 50, 1089–122.

Myerson, R.B. 1981. Optimal auction design. *Mathematics of Operations Research* 6, 58–73.

Parkes, D.C. and Ungar, L.H. 2000. Iterative combinatorial auctions: theory and practice. *Proceedings of the 17th National Conference on Artificial Intelligence*, 74–81.

Rassenti, S.J., Smith, V.L. and Bulfin, R.L. 1982. A combinatorial auction mechanism for airport time slot allocation. *Bell Journal of Economics* 13, 402–17.

Riley, J.G. and Samuelson, W.F. 1981. Optimal auctions. *American Economic Review* 71, 381–92.

Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16, 8–37.

Wilson, R. 1979. Auctions of shares. *Quarterly Journal of Economics* 93, 675–89.

Wilson, R. 1992. Strategic analysis of auctions. In *Handbook of Game Theory*, vol. 1, ed. R. Aumann and S. Hart. Amsterdam: North-Holland.

How to cite this article

Ausubel, Lawrence M. "auctions (theory)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000217> doi:10.1057/9780230226203.0073

The New Palgrave Dictionary of Economics Online

Aumann, Robert J. (born 1930)

Abraham Neyman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robert Aumann has played an essential and indispensable role in shaping game theory and much of economic theory. He promotes a unified view of the very wide domain of rational behaviour, a domain that encompasses areas of many apparently disparate disciplines, like economics, political science, biology, psychology, mathematics, philosophy, computer science, law and statistics. His contributions have had a most profound impact on the social sciences.

Keywords

Aumann, R. J.; bounded rationality,; coalitions; common knowledge; continuum of traders; contract curve; convexity; core; core equivalence; correlated equilibrium; folk theorem; game theory; non-transferable utility; perfect competition; rational behaviour; repeated games with incomplete information; Shapley value; strong equilibrium; supergames; transferable utility.

Article

Robert J. Aumann, Professor Emeritus of Mathematics at the Hebrew University of Jerusalem, and member of the interdisciplinary Center for Rationality there, shares (with Thomas C. Schelling) the 2005 Nobel Prize in Economics (Aumann and Schelling, 2005).

Aumann was born in Frankfurt, Germany, in 1930, and moved to New York with his family in 1938. In 1955 he completed his Ph.D. in mathematics at MIT under the supervision of George Whitehead. His thesis, in knot theory, was published in the *Annals of Mathematics* (Aumann, 1956).

In 1955, Aumann joined the Princeton University group that worked on industrial and military applications, where he realized the importance and relevance of game theory, then in its infancy. In 1956 Aumann joined the Institute of Mathematics at the Hebrew University.

Since the mid-1950s, Aumann has played an essential and indispensable role in shaping game theory, and much of economic theory, to become the great success it is today. He promotes a unified view of the

very wide domain of rational behaviour, a domain that encompasses areas of many apparently disparate disciplines, like economics, political science, biology, psychology, mathematics, philosophy, computer science, law, and statistics. Aumann's research is characterized by an unusual combination of breadth and depth. His scientific contributions are path-breaking, innovative, comprehensive and rigorous, ranging from the discovery and formalization of the basic concepts and principles, through the development of the appropriate tools and methods for their study, to their application in the analysis of various specific issues. Some of his contributions require very deep and complex technical analysis; others are (as he says at times) ‘embarrassingly trivial’ mathematically, but very profound conceptually. He has influenced and shaped the field through his pioneering work. There is hardly an area of game theory today where his footprint is not readily apparent. Most of Aumann's research is intimately connected to central issues in economic theory; on the one hand, these issues provided the motivation and impetus for his work; on the other, his results produced novel insights and understandings in economics. No less important than his own pioneering work is Aumann's indirect impact through his many students, collaborators and colleagues. He inspired them, excited them with his vision, and led them to further important results.

Here we must confine ourselves to brief commentary touching on only a small part of his output. It is important to note that the scope of each description is not indicative of the importance of the contribution. Further and more detailed accounts of Aumann's contributions may be found in Hart and Neyman (1995).

We start with Aumann's study of long-term interactions, which had a most profound impact on the social sciences. The mathematical model enabling a formal analysis is a *supergame* G^*, consisting of an infinite repetition of a given one-stage game G. (A game G in strategic form consists of a set of players N, pure strategy sets Ai for each player i, and payoff functions gi, which describe the payoff to player i as a function of the strategy profiles $a \in A{:=}Xi{\in}NAi$.) A pure strategy in G^* assigns a pure strategy in G to each period/stage, as a function of the history of play up to that stage. A profile of supergame strategies, one for each player, defines the play, or sequence of stage actions. The payoff associated with a play of the supergame is essentially an average of the stage payoffs.

In 1959 Aumann defined the notion of a *strong equilibrium* – a strategy profile where no group of players can gain by unilaterally changing their strategies – and characterized the strong equilibrium outcomes of the supergame by showing that it coincides with the so-called β*-core* of G. When Aumann's 1959 methodology is applied to *Nash equilibrium* – a strategy profile where no single player can gain by unilaterally changing his strategy – the result is essentially the so-called *folk theorem* for supergames: the set of Nash equilibria of the supergame G^* coincides with the set of feasible and individual rational payoffs in the one-stage game. In 1976, Aumann and Shapley (and Rubinstein, 1976, in independent work) proved that the equilibrium payoffs and the perfect equilibrium payoffs of the supergame G^* coincide.

Supergames are repeated games of complete information; it is assumed that all players know precisely the one-shot game that is being repeatedly played.

> The theory of repeated games of complete information is concerned with the evolution of fundamental patterns of interaction between people (or for that matter, animals; the problems it attacks are similar to those of social biology). Its aim is to account for

> phenomena such as cooperation, altruism, revenge, threats (self-destructive or otherwise), etc. – phenomena which may at first seem irrational – in terms of the usual 'selfish' utility-maximizing paradigm of game theory and neoclassical economics. (Aumann, 1981, p. 11)

The model of repeated games with incomplete information, introduced in 1966 by Aumann and Maschler (Aumann and Maschler, 1995), analyses long-term interactions in which some or all of the players do not know which stage game G is being played. The game $G=Gk$ depends on a parameter k; at the start of the game a commonly known lottery $q(k)$ with outcomes in a product set $S=\times iSi$ is performed and player i is informed of the i-th coordinate of the outcome. The repetition enables players to infer and learn information about the other players from their behaviour, and therefore there is

> a subtle interplay of concealing and revealing information: concealing, to prevent the other players from using the information to your disadvantage; revealing, to use the information yourself, and to permit the other players to use it to your advantage. (Aumann, 1985, pp. 46–47)
> The stress here is on the strategic use of information – when and how to reveal and when and how to conceal, when to believe revealed information and when not, etc. (Aumann, 1981, p. 23)

This problem of the optimal use of information is solved in an explicit and elegant way in Aumann and Maschler (1995).

Another substantial line of contributions of Aumann is the introduction and study of the continuum idea in game theory and economic theory.

A perfectly competitive economic model is meant to describe a situation in which there are many participants, and the influence of each one individually is negligible. The state of the economy is thus insensitive to the actions of any single agent; only the aggregate behaviour matters. For instance, in a pure exchange economy in which the initial endowment of each trader is very small relative to the whole, the quantities of goods traded by any one agent cannot essentially affect the total supply and demand.

The first question is: What is the correct way of modelling perfect competition? Aumann introduced the model of economies with a continuum of participants, as the appropriate model where each individual is indeed insignificant:

> Indeed, the influence of an individual participant on the economy cannot be mathematically negligible, as long as there are only finitely many participants. Thus a mathematical model appropriate to the intuitive notion of perfect competition must contain infinitely many participants. We submit that the most natural model for this purpose contains a continuum of participants, similar to the continuum of points on a line or the continuum of particles in a fluid. (Aumann, 1964, p. 39)

The introduction of the 'continuum' idea in economic theory has been indispensable to the advancement of this discipline. In the same way as in most of the natural sciences, it enables a precise and rigorous

analysis, which otherwise would have been very hard or even impossible. Specifically,

> the continuum can be considered an approximation to the ‘true’ situation in which there is a large but finite number of particles (or traders, or strategies, or possible prices). The purpose of adopting the continuous approximation is to make available the powerful and elegant methods of the branch of mathematics called ‘analysis,’ in a situation where treatment by finite methods would be much more difficult or even hopeless (think of trying to do fluid mechanics by solving *n*-body problems for large *n*. (Aumann, 1964, p. 41)

Once the basic model is specified, the next question is: What does perfect competition lead to? The classical economic approach is that there are prices for all goods, which every agent takes as given (he is, after all, insignificant, so his decision cannot affect the prices). In order for the economy to be in a stable situation the prices must be such that the total demand equals the total supply. This is the Walrasian competitive equilibrium. That it exists and is well defined in markets with a continuum of traders was shown by Aumann in 1966; moreover, unlike in finite markets, no convexity assumptions were required.

Another approach considers the possible trades that groups of agents – called coalitions – can make among themselves, in such a way that they all benefit. This leads to the core, a game-theoretic concept that generalizes Edgeworth's famous ‘contract curve’: the core consists of all those allocations that no coalition can improve upon. These are clearly different concepts:

> The definition of competitive equilibrium assumes that the traders allow market pressures to determine prices and that they then trade in accordance with these prices, whereas that of core ignores the price mechanism and involves only direct trading between the participants. (Aumann, 1964, p. 40)

Aumann (1964) showed that the core and the set of competitive allocations coincide in markets with a continuum of traders. By introducing the model of the continuum that expresses precisely the idea of perfect competition, he succeeded in making precise also this equivalence (originally suggested by Edgeworth, 1881, and proved in various other models – Shubik, 1959; Debreu and Scarf, 1963), which has since become one of the basic tenets of economic theory.

Aumann then turned to the study of other concepts in the context of perfectly competitive markets. A traditional idea in economics is that of ‘marginal worth’ or ‘marginal contribution’. This idea is embodied in the concept of value due to Lloyd Shapley (1953). It may be interpreted as follows:

> The Shapley value is an a priori measure of a game's utility to its players; it measures what each player can expect to obtain, ‘on the average,’ by playing the game. Other concepts of cooperative game theory … predict outcomes (or sets of outcomes) that are in themselves stable, that cannot be successfully challenged or upset … The Shapley value … can be considered a mean, which takes into account the various power relationships and possible outcomes. (Aumann, 1978, p. 995)

While the definition of competitive equilibrium or core generalizes in a straightforward manner to the continuum of players case, this is not so in the case of value. This led to a most prolific collaboration between Aumann and Shapley, starting in the late 1960s and culminating in 1974 with the publication of their book *Values of Non-Atomic Games*. They addressed deep problems, both conceptual – how to define the correct notions – and technical, and solved them masterfully. In consequence, most important and beautiful insights were obtained. One example is the 'diagonal principle', stating that in games with many players one need consider only coalitions whose composition constitutes a good sample of the grand coalition of all participants. It is important to note that, unlike the core (or the competitive equilibrium), the value solution is applicable in almost every interactive set-up. For instance, political contexts usually lead to situations where the core is empty, whereas the value is well defined and yields most significant insights.

Returning to perfectly competitive economies, in 1975 Aumann obtained another equivalence result, this time between the competitive allocations and the value allocations – on the assumption that the market is 'sufficiently smooth'. (Again, the continuum of traders model allows Aumann to obtain a precise and general result; the first such result, in transferable utility markets only, is due to Shapley, 1964.) This is perhaps even more surprising than the core equivalence, since the concept of value does not capture, by its definition, considerations of stability and equilibrium.

This equivalence is indeed striking. In Aumann's view:

> Perhaps the most remarkable single phenomenon in game and economic theory is the relationship between the price equilibria of a competitive market economy, and all but one of the major solution concepts for the corresponding game. … Intuitively, the equivalence principle says that the institution of market prices arises naturally from the basic forces at work in a [perfectly competitive] market, (almost) no matter what we assume about the way in which these forces work. (From game theory)

This nicely exemplifies Aumann's view on the universality of the game theoretic approach:

> The more conventional approaches take institutions as given, and ask where they lead. The game-theoretic approach asks how the institutions came about, what led to them? Thus general equilibrium theory takes the idea of market prices for granted; it concerns itself with their existence and properties, calculating them, and so on. Game Theory asks, why are there market prices? How did they come about? (From game theory)

The fundamental insights and understandings obtained in the analysis of perfect competition enabled and facilitated the study of basic economic issues that go beyond perfect competition. We mention a few where Aumann's contributions and influence are most noticeable: monopolistic and oligopolistic competition, modelled by a continuum of traders together with one or more large participants (Shubik, 1959); public economics – models of taxation based on the interweaving of the economic activities with a political process, such as voting (Aumann and Kurz, 1977a; 1977b; Aumann, Gardner and Rosenthal, 1977; Aumann, Kurz and Neyman, 1983; 1987); fixed-price models (Aumann and Drèze, 1986).

Another fundamental contribution of Aumann is ‘Agreeing to Disagree’ (1976): it formalizes the notion of *common knowledge* and shows (the somewhat unintuitive result) that, if two agents start with the same prior beliefs and their posterior beliefs (about a specific event), which are based on different private information, are common knowledge, then these posterior beliefs coincide. This paper had a major impact; it led to the development of the area known as *interactive epistemology* and has found many applications in different disciplines like economics and computer science.
Other fundamental contributions include the introduction and study of *correlated equilibrium*, the study of *bounded rationality*, and many important contributions to cooperative game theory: extending the theory of *transferable utility* (TU) games to general *nontransferable utility* (NTU) games, formulating a simple set of axioms that characterize the NTU-value (introduced in Shapely, 1969) and the ‘Game-Theoretic Analysis of a Bankruptcy Problem from the Talmud’ (Aumann and Maschler, 1985).
Aumann has been a Member of the US National Academy of Sciences since 1985, a Member of the Israel Academy of Sciences and Humanities since 1989, a Foreign Honorary Member of the American Academy of Arts and Sciences since 1974, and a corresponding fellow of the British Academy since 1995. He received the Harvey Prize in Science and Technology in 1983, the Israel Prize in Economics in 1994, the Lanchester Prize in Operations Research in 1995, the Nemmers Prize in Economics in 1998, the EMET prize in Economics in 2002, the von Neumann prize in Operations Research in 2005, and the Nobel Memorial Prize in Economic Sciences in 2005. He was awarded honorary doctorates by the University of Bonn in 1988, by the Université Catholique de Louvain in 1989, and by the University of Chicago in 1992.

See Also

- convexity
- game theory

Selected works

1956. Asphericity of alternating knots. *Annals of Mathematics* 64, 374–92.

1959. Acceptable points in general cooperative *n*-person games. In *Contributions to the Theory of Games*, vol. 4 (AM-40), ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34, 1–17.

1974. (With L.S. Shapley.) *Values of Non-Atomic Games*. Princeton: Princeton University Press.

1975. Values of markets with a continuum of traders. *Econometrica* 43, 611–46.

1976. Agreeing to disagree. *Annals of Statistics* 4, 1236–9.

1976. (With L.S. Shapley.) Long-term competition: a game-theoretic analysis. Mimeo, Hebrew University;repr. In *Essays in Game Theory in Honor of Michael Maschler*, ed. N. Megiddo. Berlin: Springer, 1994.

1977. (With R.J. Gardner and R.W. Rosenthal.) Core and value for a public goods economy: an example. *Journal of Economic Theory* 15, 363–5.

1977a. (With M. Kurz.) Power and taxes. *Econometrica* 45, 1137–61.

1977b. (With M. Kurz.) Power and taxes in a multi-commodity economy. *Israel Journal of Mathematics* 27, 185–234.

1978. Recent developments in the theory of the Shapley value. In *Proceedings of the International Congress of Mathematicians*. Helsinki: Academia Scientiarum Fennica, 1980.

1981. Survey of repeated games. In *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern*, vol. 4. Mannheim: Bibliographisches Institut.

1983. (With M. Kurz and A. Neyman.) Voting for public goods. *Review of Economic Studies* 50, 677–694.

1985. What is game theory trying to accomplish? In *Frontiers of Economics*, ed. K.J. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

1985. (With M. Maschler.) Game-theoretic analysis of a bankruptcy problem from the Talmud. *Journal of Economic Theory* 36, 195–213.

1986. (With J.H. Drèze.) Values of markets with satiation or fixed prices. *Econometrica* 54, 1271–318.

1987. (With M. Kurz and A. Neyman.) Power and public goods. *Journal of Economic Theory* 42, 108–27.

1995. (With M. Maschler and the collaboration of R.E. Stearns.) *Repeated Games of Incomplete Information*. Cambridge, MA: MIT Press.

2000. *Collected Papers*, vols 1 and 2. Cambridge, MA: MIT Press.

2005. (With T. Schelling.) *Robert Aumann's and Thomas Schelling's Contributions to Game Theory: Analyses of Conflict and Cooperation*. Stockholm: Royal Swedish Academy of Sciences. Online.

Available at http://nobelprize.org/economics/laureates/2005/ecoadv05.pdf, accessed 5 June 2007.

Bibliography

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 236–46.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Hart, S. and Neyman, A. 1995. Introduction. In *Games and Economic Theory, Selected Contributions in Honor of Robert J. Aumann*, ed. S. Hart and A. Neyman. Ann Arbor: University of Michigan Press. Online. Available at http://ratio.huji.ac.il/dp/neyman/bookintroduction95.pdf, accessed 5 June 2007.

Neyman, A. 2006. Aumann awarded Nobel Prize. *Notices of the AMS* 53, 44–6.

Rubinstein, A. 1976. Equilibrium in supergames. RM-26. Hebrew University of Jerusalem.

Shapley, L.S. 1953. A value for *n*-person games. In *Contributions to the Theory of Games*, vol. 2, ed. H. W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Shapley, L.S. 1964. *Values of Large Games VII: A General Exchange Economy with Money*. RM-4248. Santa Monica, CA: RAND Corporation.

Shapley, L.S. 1969. Utility comparison and the theory of games. In *La Décision*. Paris: CNRS.

Shubik, M. 1959. Edgeworth market games. In *Contributions to the Theory of Games*, vol. 4, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

How to cite this article

Neyman, Abraham. "Aumann, Robert J. (born 1930)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000251> doi:10.1057/9780230226203.0074

The New Palgrave Dictionary of Economics Online

Aupetit, Albert (1876 – 1943)

Roger Dehem
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Aupetit, A.; bimetallism; composite money standards; exchange rate determination; index numbers; quantity theory of money; Walras, L.

Article

Aupetit was born in Sancerre (Cher). His two doctoral theses at the Faculté de Droit were respectively entitled *Théorie générale de la monnaie* (1901) and *Les accidents du travail dans l'agriculture*. Having twice failed the *concours d'agrégation*, the narrow gateway to a professorship at the Faculté de Droit, he entered the research department at the Banque de France, where he served as secretary-general from 1920 to 1926. He then entered private business. In 1936 he was elected a member of the Institut de France. His teaching was restricted to the Ecole Pratique des Hautes Etudes (1910–14) and to the Ecole des Sciences Politiques, from 1921 on.
Considered by Walras as his first disciple in France, Aupetit can best be judged by the master himself: 'He is in agreement with my social economics as well as with my pure and applied economics. He is the best and most brilliant disciple and successor I may wish to have' (Jaffé, 1965, p. 353). Aupetit's *Essai sur la théorie générale de la monnaie* is a faithful though simpler and more precise reformulation of Walras' general equilibrium and monetary theories. The postulates sustaining the quantity theory are made remarkably explicit. Questions of composite monetary standards, bimetallism, exchange rate determination and index numbers are also thoroughly discussed.

Selected works

1901. *Essai sur la théorie générale de la monnaie*. Paris: Guillaumin. A truncated version of this book was published by Marcel Rivière, Paris, in 1957.

1905. L'oeuvre économique de Cournot. *Revue de métaphysique et de morale* 13, 377–93.

Bibliography

Jaffé, W. 1965. *Correspondence of Léon Walras and Related Papers*, vol. 3, Amsterdam: North-Holland.

How to cite this article

Dehem, Roger. "Aupetit, Albert (1876–1943)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000165> doi:10.1057/9780230226203.0075

The New Palgrave Dictionary of Economics Online

Auspitz, Rudolf (1837 – 1906)

Jürg Niehans
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Auspitz, R.; Austrian School; Böhm-Bawerk, E. von; Cartels; competitive equilibrium; consumer surplus; disutility; Edgeworth, F. Y.; forward markets; gains from trade; indifference curves; indivisibilities; inventories; Launhardt, C. F. W.; Lieben, R.; marginal utility of money; Marshall, A.; mathematical economics; Menger, C.; multiple equilibria; options; Pareto, V.; partial equilibrium; producer surplus; reciprocal demand curves; security markets; speculation; substitutes and complements; technical progress; Walras, L.; Wicksell, J. G. K.

Article

Auspitz was born on 7 July 1837 in Vienna, where he died on 8 March 1906. He grew up in a well-educated Jewish family and studied mathematics and physics but without acquiring a degree. At the age of 26, apparently with some reluctance, he became a businessman and founded one of the first sugar refineries of the Austrian empire. As a lifelong opponent of cartels, he used to donate the extra profits he obtained from the sugar cartel to the employees' pension fund. Auspitz was also Richard Lieben's partner in the family bank, Auspitz, Lieben & Co.

A successful Liberal politician, Auspitz was a member of the Moravian Diet (1871–1900) and of the Austrian lower chamber (1873–90 and 1892–1905), where he acquired a reputation and influence as a financial expert. His first wife was Lieben's sister and a first cousin. They had two children, but the marriage was dissolved after 20 years because of the wife's insanity, whereupon Auspitz married his children's governess. He seems to have been a man of quiet energy and balanced judgement, untiring but of frail health. In some respects his life reminds one of Ricardo's.

All of Auspitz's significant scientific work was done jointly with Lieben; nothing seems to be known about their relative contributions. In 1889 appeared the *Researches on the Theory of Price*, the book that assured its authors of a place among the eminent mathematical economists. It is essentially an exhaustive partial-equilibrium analysis of price in terms of an ingenious geometrical apparatus.

The fundamental first chapter, preprinted in 1887 to fix priorities relative to Böhm-Bawerk, provides the

basic tools. For every quantity of a given commodity, the ‘curve of total satisfaction’ indicates the maximum amount of money the buyer is willing to pay. The ‘total cost curve’, on the other hand, plots the minimum amount of money for which the seller (producer) is willing to supply each quantity. In modern terminology, these are indifference curves. The corresponding marginal curves, called respectively demand and supply curves, give the maximum (minimum) amount of money for which the buyer (seller) is willing to buy (sell) an additional unit.

On the assumption of a constant marginal utility of money, both parties choose the quantity in such a way that this marginal value is equal to the market place. The two marginal curves are thus equivalent to Marshall's reciprocal demand curves as applied to the exchange of one commodity against money. Auspitz and Lieben did not know Marshall's privately printed paper of 1879, however.

Competitive equilibrium is established where the demand curve intersects the supply curve. The vertical distances between the equilibrium point and the two indifference curves then measure the gains from trade, which leads to an analysis of consumer's and producer's surplus (but without these terms).

In subsequent chapters this apparatus is applied to a wide range of microeconomic problems and cases, including substitutes and complements, indivisibilities, disutility, technical progress, inventories, security markets, forward markets and options. Among many notable pieces of analysis one finds the argument that speculation is socially beneficial if it is profitable, and a derivation of long-run curves as envelopes of short-run curves which was not surpassed until Harrod and Viner. An important final chapter extends the analysis to monopoly, monopolistic competition, excise taxes and international trade, and includes a brilliant discussion of optimal tariffs (which disturbed free-trader Pareto; see *Giornale degli Economisti*, 1892).

Four appendices present the main argument in terms of univariate differential calculus, concluding with an extension to general equilibrium. In contrast to Launhardt, who, as an engineer, loved to computed numerical results for special functional forms, Auspitz and Lieben emphasize the logic of the problem.

Auspitz and Lieben, though highly regarded by men like Edgeworth, Pareto and Fisher, never received the credit they deserved. In their local environment, in view of the Austrian School's intolerance for mathematics, they were academic outcasts. This is illustrated by Menger's critical review (*Wiener Zeitung*, 8 March 1889, quoted in Weinberger, 1931) and by Auspitz's exchange with Böhm-Bawerk of 1894, which also shows Auspitz's analytical superiority. More importantly, Auspitz and Lieben, cut off from direct scholarly intercourse, were prisoners of their idiosyncrasy, never developing the knack for felicitous terminology and expository devices that in economics is so important for academic success. It also turned out that for partial analysis Cournot's price/quantity diagram is often more illuminating than the reciprocal demand curves.

Despite their gentle, scholarly personalities, Auspitz and Lieben also managed to stir up a controversy with Walras (see *Correspondence of Léon Walras and Related Papers*, ed. William Jaffé, 3 vols, Amsterdam, 1965). As early as 1887, Launhardt had warned Walras of the ‘plagiarism’ of those ‘insolent Jewish pirates’. The preface to the *Researches*, while revealing Launhardt's diatribes as entirely unfounded, added a more substantive irritant by arguing that (1) Walras’ simultaneous demand curves were not correctly constructed, in as much as the curve for one good presupposes a given price for the other, and (2) there cannot be multiple equilibria. This criticism stung Walras all the more since Edgeworth, in his presidential address of 1889, described Auspitz and Lieben as more accurate than Walras (an unwarranted observation, deleted in *Papers Relating to Political Economy*). Walras tried to mobilize Pareto and Bortkiewicz in his defence (without success) and began to polemicize against those

who ‘make bad theory in mathematical language’. His own reply, however (reprinted in the 4th edition of the ‘*Eléments*’), missed the essential point and only added to the confusion. Wicksell, as usual, got things right (*Wert, Kapital und Rente*, 1893). Auspitz and Lieben had overlooked the fact that Walras’ curves, in effect, related to the demand and supply of one good in terms of the other, and the impossibility of multiple equilibria depended on the constancy of the marginal utility of money. After Auspitz's death Lieben graciously acknowledged their error (to which Walras, ungraciously, replied that the point was not important after all).

See Also

- Lieben, Richard

Selected works

1885. *Meine parlamentarische Thätigkeit während der Reichsraths-Session 1879–85*. Vienna.

1887. (With R. Lieben.) *Zur Theorie des Preises*. Leipzig: Duncker & Humblot.

1889. (With R. Lieben.) *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot. French translation by Louis Suret, Paris: M. Giard & E. Brière, 1914.

1890a. (With R. Lieben.) Reply. (To article by L. Walras in same publication.) *Revue d’économie politique* 4.

1890b. Die klassische Werttheorie und die Theorie von Grenznutzen. *Jahrbücher für Nationalökonomie und Statistik* 55.

1894. Der letzte Maasstab des Güterwertes und die mathematische Methode. *Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung* 3.

Bibliography

Weinberger, O. 1931. Rudolf Auspitz und Richard Lieben. *Zeitschrift für die gesamte Staatswissenschaft* 91.

Weinberger, O. 1935. Rudolf Auspitz. *Neue Oesterreichische Biographie 1815–1918*, vol. 8. Vienna.

Winter, J. 1927. *Fünfzig Jahre eines Wiener Hauses*. Vienna: F. Jasper.

How to cite this article

Niehans, Jürg. "Auspitz, Rudolf (1837–1906)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000166> doi:10.1057/9780230226203.0076

The New Palgrave Dictionary of Economics Online

Australasia, economics in

Selwyn Cornish

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Writing on economic subjects began in Australasia (Australia and New Zealand) within a decade or two of the commencement of European settlement. There are many examples of innovative and influential contributions to economics from these countries, but there has never been a ‘school’ of Australasian economics. Between the two world wars, economics in Australia experienced a golden age, when a small group of economists influenced economic policy and advanced economic thought. Since the 1940s, however, Australasian economics has been dominated by ideas and methods associated with work in the United States.

Keywords

Arndt, H. W.; Australasia, economics in; Brigden, J.B.; Butlin, N.; capital theory; Clark, C. G.; Coghlan, T.A.; Condliffe, J.B.; Copland, D.; Corden, W. M.; Crawford, J. G.; De Lissa, A.; Dutch disease; Economic Society of Australia; effective protection; equation of exchange; Fisher, A.G.B.; full employment; Garnaut, R.; general equilibrium trade models; Giblin, L. F.; Gregory thesis; growth, models of; Harcourt, G. C.; Hearn, W. E.; Hight, J.; inflation; international monetary economics; international trade theory; Irvine, R. F.; Jevons, W. S.; Kemp, M.; Keynes, J. M.; Marshall, A.; measurement; Mill, J. S.; money; multiplier; Musgrave, A.; national income accounting; New Zealand Economic Association; Phillips, A.W.; Premiers’ Plan (Australia); protection; Salter, W.E.G.; Shann, E.; Snooks, G. D.; Swan, T. W.; Syme, D.; tariffs; tradable and non-tradable commodities; unemployment; Viner, J.; Wakefield, E.G.; Walker, F. A.; Wilson, R.

Article

There has never been a ‘school’ of Australasian economics in the sense that English, German, Austrian, Italian, American and Swedish schools are said to have existed.
This is not to say that Australians and New Zealanders have contributed little or nothing to the history of

economics. On the contrary, an economics literature commenced from the early decades of the 19th century. For the most part, economic analysis was derived from ideas originating outside the region, though imported ideas were adapted, extended and refashioned to meet peculiar Australasian conditions and circumstances. Between the two world wars, economics in Australia experienced a golden age when a remarkable group of economists exerted a profound impact on economic policy, and in the process advanced economic thought. Since the Second World War, Australasian economics has been dominated by approaches and methods that are characteristically associated with the discipline in the United States, a phenomenon by no means unique to Australia and New Zealand.

The 19th century

Survival was difficult and far from guaranteed for some years immediately after the establishment of European settlement in Australia in 1788. In these circumstances there was little time to write about economics. But as private activity evolved from the original penal settlements, economic issues were debated more frequently. By the 1840s, a flourishing private economy had developed around the wool export trade with Britain. The pastoral industry was land intensive, giving rise to discussion about the occupation and alienation of crown land. The growth of domestic production led to an interest in its measurement and the contributions made by different industries. The creation of private institutions, especially those catering to foreign trade, including banks and other financial institutions, wholesaling and retailing, shipping and inland transport, became subjects of interest among those who wrote and talked about economic matters. Population growth and immigration were other subjects that drew attention. With the rise of domestic and foreign trade, instability occasioned by excessive optimism and pessimism was manifested in booms and slumps; this, too, engaged the interest of writers.
E.G. Wakefield, though he never visited the antipodes, wrote in 1829 that the Australian colonies were

> in a barbarous condition, like that of every people scattered over a territory immense in proportion to their numbers; every man is obliged to occupy himself with questions of daily bread; there is neither leisure nor reward for investigation of abstract truth; money-getting is the universal object; taste, science, morals, manners, abstract politics are subjects of little interest unless they bear on the wool question. (Quoted in Nadel, 1957, p. 36)

There is some truth in this, but, by the time Wakefield wrote, pamphlets and books by colonists on economic topics had started to appear. In 1819, for example, W.C. Wentworth published *A Statistical, Historical, and Political Description of the Colony of New South Wales and its Dependent Settlements in Van Diemen's Land*. Wentworth estimated the national income of New South Wales and Van Diemen's Land (since renamed Tasmania), and discussed processes of economic development that borrowed heavily from Adam Smith. Another early writer of some significance was the Reverend John Dunmore Lang. In 1834 he published *An Historical and Statistical Account of New South Wales* which provided a description of economic progress in the colony and an analysis of the nature and causes of the depressions of the late 1820s and the early 1840s.
William Stanley Jevons spent some years in Australia in the 1850s as assayer to the Royal Mint in

Sydney. He wrote on railways and land development, and commenced a social survey of Sydney, revealing some of the promise that later was to emerge in his work in economics. Perhaps the most important writer on economics in Australia during the second half of the century was William Edward Hearn. Born in Ireland and educated at Trinity College, Dublin, an exact contemporary of Cairnes and Cliffe Leslie, Hearn in 1854 was appointed foundation Professor of Modern History, Modern Literature, Logic and Political Economy in the University of Melbourne. As an academic (he later became a Member of Parliament), Hearn published a number of books, of which the most important was *Plutology* (1863). Written as a university textbook, it was widely known in Britain and elsewhere as an outstanding summary of the state of economic knowledge. Hearn believed that the satisfaction of wants, and the efforts to meet them, constituted the chief problems of economics.

Another prominent writer of the second half of the 19th century was Sir Anthony Musgrave, Governor of South Australia and later of Queensland. His major work, *Studies in Political Economy* (1875), contained six essays critical of J.S. Mill. He claimed that Mill had failed to explore adequately the role of money as a store of value and there were deficiencies in Mill's discussion of capital. Though Musgrave's work was often quoted, his jaundiced view of Mill's writing won him few friends among authorities overseas. David Syme, proprietor of *The Age*, a Melbourne newspaper, was yet another writer with a reputation beyond Australia. Better known for his powerful advocacy of protection, and for his writing on the disposal of crown land, Syme published as well on economic methodology and other abstract topics. His *Outlines of an Industrial Science* (1876) seems to have been known in Europe, notably in Germany. Syme supported the application of inductive approaches to economics and criticized Mill for arguing that economics should be based on deduction. He wrote as well on economic motivation and on supply and demand analysis, criticizing as he did Mill's theory of value.

Towards the end of the 19th century a number of factors combined to encourage greater scrutiny of economic issues. One was the banking and financial crisis and collapse of economic activity in eastern Australia in the 1890s. As a consequence of the depression, debate sharpened on subjects such as the causes of fluctuations in economic activity, the role of government in moderating booms and slumps, the need for a central or government bank, unemployment and tariff policy. Another issue was the projected federation of the Australian colonies. Hitherto the six colonies of Australia had acted independently, having their own administrations, including armies and navies. Ever since the middle of the 19th century there had been calls for an Australian federation; during the 1890s several inter-colonial conventions were held to draft a federal constitution, at which economic and financial considerations, including tariffs, taxation, federal–state finance, money and banking, were debated at length.

Reflecting the heightened interest in economics for these and other reasons, an Australian Economic Association was formed in Sydney in 1887. Between March 1888 and December 1898 the Association published a monthly periodical (for a short time it was published fortnightly). Contributors to the *Australian Economist* were interested principally in the issues of the day, including unemployment, wage rates, tariff policy, recovery measures, control of banks and money, land tenure, federation, socialism, state banks, education, immigration, the role of women, democracy, bimetallism, old age pensions and industrial arbitration. Short extracts from the works of prominent economists, including Jevons, Marshall and F.A. Walker, were often included, as were articles about the work of these and other economists.

The most original of the local contributors to the *Australian Economist* was Alfred De Lissa, whose work sometimes is heralded as a forerunner of the multiplier. In March 1890 he read to the Australian

Economic Association a paper on The Law of the Incomes (1890), in which he noted that incomes arising from primary production led to an increase in income in other sectors. Using production data, and taking into account leakages abroad, he concluded that, as a general rule, incomes of primary producers equalled incomes of secondary producers; the original primary income, in other words, had a general tendency to multiply by a factor of two. De Lissa later argued that the relationship between primary and secondary income would diminish progressively until the additional income reached zero.
An area where Australia was clearly at the forefront of work internationally by the end of the 19th century was the official collection and interpretation of economic and social statistics. The most acclaimed of the colonial statisticians was Timothy Coghlan, the New South Wales Statistician, who pioneered the measurement of the national income using income, output and expenditure methods, an approach similar in many ways to modern national income accounting. Coghlan later worked in London as Agent-General for New South Wales. There he wrote a four-volume economic history of Australia – *Labour and Industry in Australia* (1918) – that drew upon quantitative information he had assembled when he was in Sydney. Later work in Australia by Colin Clark (1940), H.W. Arndt (1949), N.G. Butlin (1962) and G.D. Snooks (1994) acknowledged the ground-breaking statistical work, including national income estimation, of Coghlan and other 19th-century colonial statisticians.

Economics in the universities

When the first universities were established in Sydney in 1851 and in Melbourne in 1854, economics was not a subject that attracted much attention. At the University of Sydney, the Professor of Classics (John Woolley) and the Professor of Philosophy (Francis Anderson) took occasional classes in economics. The Professor of Mathematics (Morris Birbeck Pell) and a later Professor of Classics (Walter Scott) gave some lectures in economics outside the university. But, as a result of growing interest in the subject by business organizations, chambers of commerce, and professional associations of bankers and accountants, courses in economics over three years began at the University of Sydney in the early 1900s. A department of economics was established in 1912, to which R.F. Irvine was appointed Professor of Economics, the first separate chair of economics in Australasia. A graduate of Canterbury University College, New Zealand, Irvine had been a pupil of James Hight. Earlier, at the University of Melbourne, Hearn had taught courses in economics for both the BA and the MA. His successor, J.S. Elkington, however, seems not to have taken the same interest in economics, and as a consequence the subject languished for a time in Melbourne.
A final year course in political economy for the BA had been offered at the University of Tasmania since the university's creation in 1889. Later a lectureship in philosophy and economics was established, but the lecturer taught courses mainly in philosophy rather than in economics. The major breakthrough in Tasmania – and, as it turned out, for economics in Australia – occurred in 1917 when Douglas Copland was appointed lecturer in history and economics. In 1920 he was appointed to a chair in economics, and later was elevated to the deanship of a new Faculty of Economics and Commerce. Like Irvine, Copland was a graduate of Canterbury University College, where he, too, had been a pupil of Hight's. In 1924 Copland was the leading force behind the establishment of the Economic Society of Australia and New Zealand, which, in the following year, published the first issue of its journal, *The Economic Record*. In the same year, 1925, Copland was appointed Professor of Commerce in the

University of Melbourne.

In the University of Adelaide, founded in 1874, courses in political economy were taught by William Mitchell in the 1890s, and by Herbert Heaton in the early 1920s; in 1929 L.G. Melville was appointed to the foundation chair of economics. Meanwhile, the universities of Queensland and Western Australia, founded just before the First World War, had established combined chairs of history and economics; Henry Alcock was appointed to the chair at Queensland, and Edward Shann to the chair at the University of Western Australia. In New Zealand by the early 1920s, chairs in economics had been established at four universities: Auckland (Horace Belshaw), Canterbury (J.B. Condliffe), Otago (A.G. B. Fisher) and Wellington (Barney Murphy).

In 1914 Irvine wrote: ‘When one considers the political and economic evolution of Australia, one cannot but be astonished at the neglect of these studies [that is, economics] in Australian universities’ (Goodwin, 1966: 636). That was certainly true of Australia prior to the First World War, but it was not true of New Zealand. By the 1890s, economics had become an important subject of study at Canterbury. There, James Hight was the foundation Professor of History and Economics. More a political historian than an economist, Hight nevertheless promoted economics as a significant field of study. A number of able students were attracted to the subject, including the first two professors of economics in Australia. By the 1920s, John Maynard Keynes could justly write that training in economics at Canterbury ‘was as good as any place in the world’ (Harper, 1986, p. 41).

The golden age of Australian economics

Yet it was in Hobart where the so-called golden age of Australian economics had its origins. Soon after his arrival at the University of Tasmania, Copland became a protégé of L.F. Giblin, a graduate in mathematics of King's College, Cambridge. Born in Tasmania, Giblin had fought on the western front in the First World War, and on leave in England had met Keynes through mutual friends. When he returned to Hobart, Giblin was appointed Tasmanian Statistician. As a member of the Council of the University of Tasmania, he was instrumental in Copland's appointment to the newly established chair in economics and for the creation of the Faculty of Economics and Commerce. Copland then attracted J.B. Brigden to fill the lectureship that he had vacated. Copland's star pupil at Hobart was Roland Wilson, who later completed doctorates in economics at Oxford and Chicago. Wilson was to become Commonwealth Statistician and later head of the Australian Treasury. The four – Giblin, Copland, Brigden and Wilson – were at the centre of the most important work undertaken in economics in Australia from the 1920s to the 1940s.

The early promise of this group, and the coming of age of Australian economics, can be seen in Copland's paper, ‘Currency Inflation and Price Movements in Australia’, published in the *Economic Journal* in 1920. Using Australian data for 1901–17, and invoking Fisher's equation of exchange, Copland derived P as a residual after applying data for M, V and T. He then compared an actual price series with the hypothetical series for P, showing that the two series exhibited close agreement. Copland concluded that the ‘equation of exchange may be regarded as true for Australia’. Keynes praised Copland for this work, referring as he did to Copland's ‘masterly article’ (Coleman, Cornish and Hagger, 2006, p. 51).

Later in the 1920s, Giblin, Copland and Brigden were appointed to the committee of enquiry into the

Australian tariff (*The Australian Tariff: An Economic Enquiry*, often known as the Brigden Report) established by the federal government in 1927 (Brigden et al., 1929). The *Enquiry* concluded that, in Australian circumstances, protection had raised the ‘standard of living’. This controversial conclusion, and the analysis upon which it was based, is said to have been significant for the emergence of modern international trade theory (Coleman, Cornish and Hagger, 2006, 65–73); Keynes adjudged that the *Enquiry* was ‘a brilliant effort of the highest interest’ (Millmow, 2005, p. 1013). Similarly, Giblin's inaugural lecture in April 1930, upon his appointment to the first research chair in economics in Australia (the Ritchie Chair in the University of Melbourne), in which he produced a multiplier based on the repercussions of a decline in exports on total domestic output, is thought to have been an important stepping-stone to the eventual formulation of the Cambridge multiplier. When Giblin sent an early version of his multiplier to Keynes in August 1929, Keynes admitted that Giblin's ‘method of argument’ was ‘novel’ (Coleman, Cornish and Hagger, 2006, p. 83).

The youngest member of ‘Giblin's Platoon’, Roland Wilson, published a book in 1931 that attracted the attention of Viner, Harrod, Hicks, Robertson and Pigou. In *Capital Imports and the Terms of Trade*, Wilson disputed Mill's contention that the import of capital would improve a borrowing country's terms of trade. More importantly, Wilson focused on the consequences of capital imports for the price ratio of tradables to non-tradables. He showed that the ratio would decline. This conclusion was taken up in the 1970s, when it was incorporated in notions such as the Dutch disease and the Gregory thesis (named after R.G. Gregory, an Australian economist who argued in the 1970s that Australia's massive export of minerals would serve to push up the Australian dollar exchange rate with adverse consequences for other industries, particularly manufacturing industry in Australia).

Giblin's group, supported by other economists, played a decisive role in furnishing advice to Australian governments and banks during the early 1930s. The economists were critical of the central bank's policy to retain a fixed rate of exchange with sterling, advising the Bank of New South Wales early in 1931 that it should use its power and prestige as Australia's largest and oldest commercial bank to devalue the Australian pound. The economists’ advice was accepted and the Australian pound was devalued. The federal and state governments then appointed Copland and Giblin to a committee (the ‘Copland Committee’) charged with the responsibility of formulating policies to deal with the depression. The committee's recommendations formed the core of measures included in the famous Premiers’ Plan of 1931. A common theme running through the anti-depression measures proposed by Australian economists was that the loss of income occasioned by the decline in exports should be spread among all income groups and not be confined to export and related trades. Their work was highly praised by foreign observers. Keynes, for example, wrote in 1932 that: ‘I am sure that the Premiers’ Plan last year saved the economic structure of Australia’ (1932, p. 94). As a measure of the influence of Australian economists, Copland was invited to present the inaugural Alfred Marshall Memorial Lectures in Cambridge in 1933; the lectures were published under the title *Australia in the World Crisis, 1929–1933* (1934).

Australian economists were prominent again during and immediately after the Second World War. Shortly before the outbreak of war, the federal government established an Economic and Financial Committee (the F&E) to advise it on economic questions that might arise in the event of war. Giblin was appointed chairman of the committee, which included Copland, Brigden and Wilson. When the war came, the F&E formulated the government's approach to war finance, following principles that Keynes

had put to the British government.

When it came to formulating plans for post-war reconstruction, Australian economists prepared at the government's request a domestic employment policy based on demand management. Their proposals were published in the famous government white paper of 1945, *Full Employment in Australia* (Cornish, 1981). The economists supported Keynes's Clearing Union, opposing as they did the rival Stabilization Fund of the United States Treasury. In fact, they went further than Keynes by formulating what they called the ‘international full employment approach’ or ‘positive approach’, sometimes known as ‘Australia's Keynesian crusade’ (Cornish, 1993). This policy arose from Article VII of the Mutual Aid Agreement signed in 1942. In return for United States assistance during the war, recipient countries pledged to enter discussions aimed at liberalizing foreign trade and international payments. Given uncertainty about the restoration of world trade, and concerned about the impact on employment of abolishing preferential trade arrangements, the ‘positive approach’ maintained that Australia would support Article VII provided the United States and other major economic powers committed themselves to policies aimed at maintaining full employment in their domestic economies. Such policies, it was believed, would provide a buoyant demand for Australian exports. Australian representatives promoted the ‘positive approach’ at major international conferences during the 1940s, including those at Bretton Woods, San Francisco and Havana.

Australasian economics since the Second World War

The numbers working in economics increased enormously after the Second World War. It is estimated that, in Australia, whereas 5,000 persons graduated in economics between 1916 and 1947, 50,000 graduated between 1947 and 1986 (Butlin, 1987). While there had been no increase in Australian universities between the two world wars, between 1945 and the early 1990s the number rose from six to more than 30. Some of the newer universities offered economics simply as a subsidiary course in business studies programmes; most, however, offered specialist degrees in economics (Groenewegen, 1996). In the 1970s, reflecting the growth of economists, the Economics Society of Australia and New Zealand was divided into two professional organizations – the Economic Society of Australia, and the New Zealand Economic Association. Yet another indicator of the expanding scale of the discipline was the increase in the number of journals dedicated to economics, from one in 1945 (*Economic Record*) to four by the mid-1960s (the additions were *Australian Economic Papers*, *Australian Economic Review* and *New Zealand Economic Papers*).

However distinctive the character of Australasian economics may have been in the interwar period, it disappeared after the Second World War as the American approach, with its emphasis on model building, mathematics and econometrics, began to dominate the discipline (Groenewegen and McFarlane, 1990). It is understandable perhaps that economists seeking to publish their work in leading international journals, many of them American-based, would want to incorporate the latest ideas and methods arising in the United States. The Americanization of the discipline also stemmed in part from the increasing number of students from Australasia going to the United States for postgraduate studies; previously the United Kingdom (Cambridge in particular) had been the destination for graduate studies in economics. Yet the American dominance of economics did not inhibit Australian and New Zealand economists from making important contributions to the subject. For example, there was the work of T.

W. Swan (1956; 1963) and W.E.G. Salter (1959) in growth theory and on issues of internal–external balance in small dependent economies; W.M. Corden's work in the theory and measurement of effective protection, tariff policy and international monetary economics (1971); Murray Kemp's formulation of general equilibrium trade models (1964); G.C. Harcourt's writing on capital theory (1986); A.W. Phillips's contributions to the theory and measurement of inflation, and the relation between wages and unemployment (1958); and the writing on Australia–Asia economic relations by J.G. Crawford (Evans and Miller, 1987), H.W. Arndt (1972) and Ross Garnaut (2001).

See Also

- Arndt, Heinz Wolfgang
- Butlin, Noel George
- Clark, Colin Grant
- Swan, Trevor W.

Bibliography

Arndt, H.W. 1949. A pioneer in national income estimates. *Economic Journal* 59, 616–25.

Arndt, H.W. 1972. *Australia and Asia: Economic Essays*. Canberra: Australian National University Press.

Brigden, J.B., Copland, D.B., Dyason, E.C., Giblin, L.F. and Wickens, C.H. 1929. *The Australian Tariff: An Economic Enquiry*. Melbourne: Melbourne University Press.

Butlin, N.G. 1962. *Australian Domestic Product, Investment and Foreign Borrowing 1861–1938/39*. Cambridge: Cambridge University Press.

Butlin, N.G. 1987. Human or inhuman capital? The economics profession 1916–87. Working Papers in Economic History No. 91, Australian National University.

Clark, C. 1940. *Conditions of Economic Progress*. London: Macmillan.

Coghlan, T.A. 1918. *Labour and Industry in Australia from the First Settlement in 1788 to the Establishment of the Commonwealth in 1901*. London and New York: Oxford University Press.

Coleman, W., Cornish, S. and Hagger, A. 2006. *Giblin's Platoon: The Trials and Triumph of the Economist in Australian Public Life*. Canberra: Australian National University Press.

Copland, D.B. 1920. Currency inflation and price movements in Australia. *Economic Journal* 30, 484–

505.

Copland, D.B. 1934. *Australia in the World Crisis, 1929–1933*. Cambridge: Cambridge University Press.

Corden, W.M. 1971. *The Theory of Protection*. Oxford: Clarendon Press.

Cornish, S. 1981. *Full Employment in Australia: The Genesis of a White Paper*. Research Papers in Economic History No. 1. Canberra: Australian National University.

Cornish, S. 1993. The Keynesian Revolution in Australia: fact or fiction? *Australian Economic History Review* 33(2), 42–68.

De Lissa, A. 1890. The law of the incomes. *Australian Economist* 2(1), 6–17.

Endres, A.M. 1991. J.B. Condliffe and the early Canterbury tradition in economics. *New Zealand Economic Papers* 25(2), 171–97.

Evans, L.T. and Miller, J.D.B., eds. 1987. *Policy and Practice: Essays in Honour of Sir John Crawford*. Canberra: Australian National University Press.

Garnaut, R. 2001. *Social Democracy in Australia's Asian Future*. Canberra: Asia Pacific Press.

Giblin, L.F. 1930. *Australia 1930*. Melbourne: Melbourne University Press.

Goodwin, C.D.W. 1966. *Economic Enquiry in Australia*. Durham, NC: Duke University Press.

Groenewegen, P. 1996. The Australian experience. In *The Post-1945 Internationalization of Economics*, ed. A.W. Coats. Durham, NC and London: Duke University Press. Annual Supplement to *History of Political Economy* 28.

Groenewegen, P. and McFarlane, B. 1990. *A History of Australian Economic Thought*. London and New York: Routledge.

Harcourt, G.C. 1972. *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.

Harper, M. 1986. Melbourne economists in the public arena: From Copland to the Institute. In *Victoria's Heritage. Lectures to Celebrate the 150th Anniversary of European Settlement in Victoria*, ed. A.G.L. Shaw. Sydney: Allen & Unwin.

Hearn, W.E. 1863. *Plutology*. Melbourne: George Robertson.

Kemp, M. 1964. *The Pure Theory of International Trade*. Englewood Cliffs, NJ: Prentice Hall.

Keynes, J.M. 1932. Report of the Australian experts. In *The Collected Writings of John Maynard Keynes*, vol. 21, ed. D. Moggridge, 94–100. London: Macmillan and Cambridge University Press for the Royal Economic Society, 1982.

Lang, J.D. 1834. *An Historical and Statistical Account of New South Wales*. London: Longman, Brown, Green and Longmans, 1852.

Millmow, A. 2005. Australian economics in the twentieth century. *Cambridge Journal of Economics* 29, 1011–26.

Musgrave, A. 1875. *Studies in Political Economy*. London: Henry S. King and Co.

Nadel, G. 1957. *Australia's Colonial Culture. Ideas, Men and Institutions in Mid-Nineteenth Century Eastern Australia*. Melbourne: F.W. Cheshire.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25, 283–99.

Salter, W.E.G. 1959. Internal and external balance: the role of price and external expenditure effects. *Economic Record* 35, 226–38.

Snooks, G.D. 1994. *Portrait of the Family Within the Total Economy: A Study in Longrun Dynamics, Australia 1788–1990*. Cambridge: Cambridge University Press.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Swan, T.W. 1963. Longer run problems of the balance of payments. In *The Australian Economy: A Volume of Readings*, ed. H.W. Arndt and W.M. Corden. Melbourne: F.W. Cheshire.

Syme, D. 1876. *Outlines of an Industrial Science*. London: Henry S. King and Co.

Wentworth, W.C. 1819. *A Statistical, Historical and Political Description of the Colony of New South Wales and its Dependent Settlements in Van Diemen's Land*. London: G. & W.B. Whittaker.

Wilson, R. 1931. *Capital Imports and the Terms of Trade, Examined in the Light of Sixty Years of Australian Borrowings*. Melbourne: Melbourne University Press.

How to cite this article

Cornish, Selwyn. "Australasia, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000249> doi:10.1057/9780230226203.0077

The New Palgrave Dictionary of Economics Online

Austrian economics

Israel M. Kirzner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian capital theory; Austrian economics; Austrian School; Austrian theory of the business cycle; Böhm-Bawerk, E. von; exploitation; German Historical School; Haberler, G.; Hayek, F. A. von; higher-order goods; Jaffé, W.; Jevons, W. S.; labour theory of value; Lange, O. R.; libertarianism; Machlup, F.; marginal productivity theory; marginal revolution; marginal utility; marginalism; market process; Marxist economics; Mayer, H.; Menger, C.; Methodenstreit; methodological individualism; methodological subjectivism; Mises, L. E. von; Morgenstern, O.; non-market prices; opportunity cost; Robbins, L. C.; Rosenstein-Rodan, P. N.; Rothbard, M. N.; roundabout production processes; Schams, E.; Schonfield, L.; Schumpeter, J. A.; Shackle, G. L. S.; socialist calculation debate; Strigl, R.; subjective theory of value; surplus value; time preference; uncertainty; utility; value and price; Walras, L.; Wicksell, J. G. K.; Wieser, F. F. von.

Article

The birth of the Austrian School of economics is usually recognized as having occurred with the 1871 publication of Carl Menger's *Grundsätze der Volkwirthschaftslehre*. On the basis of this work Menger (hitherto a civil servant) became a junior faculty member at the University of Vienna. Several years later, after a stint as tutor and travelling companion to Crown Prince Rudolph, he was appointed to a professorial chair at the University. Two younger economists, Eugen von Böhm-Bawerk and Friedrich von Wieser (neither of whom had been a student of Menger), became enthusiastic supporters of the new ideas put forward in Menger's book. During the 1880s a vigorous outpouring of literature from these two followers, from several of Menger's students, and in particular a methodological work by Menger himself, brought the ideas of Menger and his followers to the attention of the international community of economists. The Austrian School was now a recognized entity. Several works of Böhm-Bawerk and Wieser were translated into English; and by 1890 the editors of the US journal *Annals of the American Academy of Political and Social Science* were asking Böhm-Bawerk for an expository paper explaining the doctrines of the new school. What follows seeks to provide a concise survey of the history of the

Austrian School with special emphasis on (*a*) the major representatives of the school; (*b*) the central ideas identified with the school; (*c*) the relationship between the school and its ideas, and other major schools of thought within economics; (*d*) the various meanings and perceptions associated today with the term Austrian economics.

The founding Austrians

Menger's 1871 book is recognized in the history of economic thought (alongside Jevons's 1871 *Theory of Political Economy*, and Walras's 1874 *Eléments d'économie politique pure*) as a central component of the 'marginalist revolution'. For the most part, historians of thought have emphasized the features in Menger's work that parallel those of Jevons and Walras. More recently, following especially the work of W. Jaffé (1976) attention has come to be paid to those aspects of Menger's ideas which set them apart from those of his contemporaries. A series of recent studies (Grassl and Smith, 1986) have related these unique aspects of Menger and the early Austrian economists to broader currents in the late 19th-century intellectual and philosophical scene in Austria.

The central thrust of Menger's book was unmistakable; it was an attempt to rebuild the foundations of economic science in a way which, while retaining the abstract, theoretical character of economics, offered an understanding of value and price which ran sharply counter to classical teachings. For the classical economists value was seen as governed by past resource costs; Menger saw value as expressing judgements concerning future usefulness in meeting consumer wants. Menger's book, offered to the German-speaking scholarly community of Germany and Austria, was thus altogether different, in approach, style and substance, from the work coming from the German universities. That latter work, while also sharply critical of classical economics, was attacking its theoretical character, and appealing for a predominantly historical approach. At the time Menger's book appeared, the 'older' German Historical School (led by Roscher, Knies and Hildebrand) was beginning to be succeeded by the 'younger' Historical School, whose leader was to be Gustav Schmoller. Menger, the 31-year-old Austrian civil servant, was careful not to present his work as antagonistic to that of German economic scholarship. In fact he dedicated his book – with 'respectful esteem' – to Roscher, and offered it to the community of German scholars 'as a friendly greeting from a collaborator in Austria and as a faint echo of the scientific suggestions so abundantly lavished on us Austrians by Germany …' (Menger, 1871, Preface). Clearly Menger hoped that his theoretical innovations might be seen as reinforcing the conclusions derived from historical studies of the German scholars, contributing to a new economics to replace a discredited British classical orthodoxy.

Menger was to be bitterly disappointed. The German economists virtually ignored his book; where it was noticed in the German language journals it was grossly misunderstood or otherwise summarily dismissed. For the first decade after the publication of his book, Menger was virtually alone; there was certainly no Austrian 'school'. And when the enthusiastic work of Böhm-Bewark and Wieser began to appear in the 1880s, the new literature acquired the appellation 'Austrian' more as a pejorative epithet bestowed by disdainful German economists than as an honorific label (Mises, 1969, p. 40). This rift between the Austrian and German scholarly camps deepened most considerably after the appearance of Menger's methodological challenge to the historical approach (Menger, 1883). Menger apparently wrote that work having been convinced by the unfriendly disinterest with which his 1871 book had been

received in Germany, that German economics could be rescued only by a frontal attack on the Historical School. The bitter *Methodenstreit* that followed is usually (but not invariably, see Bostaph, 1978) seen by historians of economics as constituting a tragic waste of scholarly energy. Certainly this venomous academic conflict helped bring the existence of an Austrian School to the attention of the international economics fraternity – as a group of dedicated economists offering a flood of exciting theoretical ideas reinforcing the new marginalist literature, sharply modifying the hitherto dominant classical theory of value. Works by Böhm-Bawerk (1886), Wieser (1884; 1889), Komorzynski (1889) and Zuckerkandl (1889) offered elaborations or discussions of Menger's central, subjectivist ideas on value, cost, and price. Works on the theory of pure profit, and on such applications as public finance theory, were contributed by writers such as Mataja (1884), Gross (1884), Sax (1887), and R. Meyer (1887). The widely used textbook by Philippovich (1893), who was a professor at the University of Vienna (but more sympathetic towards the contributions of the German School), is credited with an important role in spreading Austrian marginal utility theory among German-language students.

In these early Austrian contributions to the theory of value and price, emphasis was (as in the Jevonsian and Walrasian approaches) placed both on marginalism and on utility. But important differences set the Austrian theory apart from other early marginalist theories. The Austrians made no attempt to present their ideas in mathematical form, and as a consequence the Austrian concept of the margin differs somewhat from that of Jevons and Walras. For the latter, and for subsequent microeconomic theorists, the marginal value of a variable refers to the instantaneous rate of change of the ‘total’ variable. But the Austrians worked, deliberately, with discrete variables (see K. Menger, 1973). More importantly the concept of marginal utility, and the sense in which it decreases, referred for the Austrians not to psychological enjoyments themselves, but to (ordinal) marginal *valuations* of such enjoyments (McCulloch, 1977). In any event, as has been urged by Streissler (1972), what was important for the Austrians in marginal utility was not so much the adjective as the noun. Menger saw his theory as demonstrating the unique and exclusive role played, in the determination of economic value, by subjective, ‘utility’, considerations. Values are not seen (as they are in Marshallian economics) as *jointly* determined by subjective (utility) and objective (physical cost) considerations. Rather values are seen as determined *solely* by the actions of consumers (operating within a given framework of existing commodity and/or production possibilities). Cost is seen (by Menger, and especially by Wieser, whose name came to be associated closely with this insight) merely as prospective utility deliberately sacrificed (in order to command more highly preferred utility). Whereas in the development of the other marginalist theories, it took perhaps two decades for it to be seen that marginal utility value theory points directly to marginal productivity distribution theory. Menger at least glimpsed this insight immediately. His theory of ‘higher-order’ goods emphasizes how both the economic character and the value of factor services are derived exclusively from the valuations placed by consumers upon the consumers products to whose emergence these higher-order goods ultimately contribute. Böhm-Bawerk contributed not only to the exposition and dissemination of Menger's basic subjective value theory, but most prominently also to the theory of capital and interest. Early in his career he published a massive volume (Böhm-Bawerk, 1884) in the history of doctrine, offering an encyclopedic critique of all earlier theories of interest (or ‘surplus value’ or ‘normal profit’). This he followed up several years later with a volume (Böhm-Bawerk, 1889) presenting his own theory. At least part of the renown of the Austrian School at the turn of the century derived from the fame of these contributions. As we shall note later on,

a number of subsequent and modern writers (such as Hicks, 1973; Faber, 1979; and Hausman, 1981) have indeed seen these Böhm-Bawerkian ideas as constituting the enduring element of the Austrian contribution. Others, taking their cue from an oft-repeated critical remark attributed to Menger (Schumpeter, 1954, p. 847 n. 8), have seen Böhm-Bawerk's theory of capital and interest as separate from, or even as somehow inconsistent with, the core of the Austrian tradition stemming from Menger (Lachmann, 1977, p. 27)). Certainly Böhm-Bawerk himself saw his theory of capital and interest as a seamless extension of basic subjectivist value theory. Once the dimension of time has been introduced into the analysis of both consumer and producer decisions, Böhm-Bawerk found it possible to explain the phenomenon of interest. Because production takes time, and because economizing men systematically choose earlier receipts over (physically similar) later receipts, capital-using production processes cannot fail to yield (even after the erosive forces of competition are taken into account) a portion of current output to those who in earlier periods invested inputs into time-consuming, ‘roundabout’ production processes.

Böhm-Bawerk became, indeed, so prominent a representative of the Austrian School prior to World War I that, largely due to his work, the Marxists came to view the Austrians as the quintessential bourgeois, intellectual enemy of Marxist economics (Bukharin, 1914). Not only did Böhm-Bawerk offer his own theory explaining the phenomenon of the interest ‘surplus’ in a manner depriving this capitalist income of any exploitative character, he had emphatically and mercilessly refuted Marxist theories of this surplus. In his 1884 work Böhm-Bawerk had systematically deployed the Austrian subjective theory of value to criticize witheringly the Marxist labour theory underlying the exploitation theory. A decade later (Böhm-Bawerk, 1896) he offered a patient, but relentless and uncompromising elaboration of that critique (in dissecting the claim that Marx's posthumously published Volume 3 of *Capital* could be reconciled with the simple labour theory forming the basis of Volume 1). This tension between the Marxists and the Austrians was to find later echoes in the debate which Mises and Hayek (third- and fourth-generation Austrians) were to conduct, during the 1920–40 interwar period, with socialist economists concerning the possibility of economic calculation in a centrally planned economy.

Menger retired from his University of Vienna professorship in 1903. His chair was assumed by Wieser. Wieser has been justly described as

> the central figure of the Austrian School: central in time, central in the ideas he propounded, central in his intellectual abilities, that is to say neither the most outstanding genius nor one of those also to be mentioned … He had the longest teaching record … (Streissler, 1986)

Wieser had been an early and prolific expositor of Menger's theory of value. His general treatise on economics, summing up his life's contributions (Wieser, 1914), has been hailed by some (but certainly not all) commentators as a major achievement. (Hayek, 1968, sees the work as a personal achievement rather than as representative of the Austrian School.) In the decade prior to the First World War, it was Böhm-Bawerk's seminar (begun when Böhm-Bawerk rejoined academic life after a number of years as Finance Minister of Austria) that became famous as the intellectual centre of the Austrian School. Among the subsequently famous economists who participated in the seminar were Josef A. Schumpeter and Ludwig von Mises, both of whom published books prior to the war (Schumpeter, 1908; 1912;

Mises, 1912).

After the First World War

The scene in Austrian economics after the war was rather different than it had been before. Böhm-Bawerk had died in 1914. Menger, who even in his long seclusion after retirement, used to receive visits from the young economists at the university, died in 1921. Although Wieser continued to teach until his death in 1926, the focus shifted to younger scholars. These included particularly Mises, the student of Böhm-Bawerk, and Hans Mayer, who succeeded his teacher Wieser, to his chair. Mises, although an ‘extraordinary’ (unsalaried) faculty member at the university, never did obtain a professorial chair. Much of his intellectual influence was exercised outside the university framework (Mises, 1978, ch. ix). Other notable (pre-war-trained) scholars during the 1920s included Richard Strigl, Ewald Schams, and Leo Schonfeld (later Illy). In the face of these changes the Austrian tradition thrived. New books were published, and a new crop of younger students came to the fore, many of whom were to become internationally famous economists in later decades. These included particularly Friedrich A. Hayek, Gottfried Haberler, Fritz Machlup, Oskar Morgenstern, and Paul N. Rosenstein-Rodan. Economic discussion among the Austrians was vigorously carried on, during the 1920s and early 1930s, within two partly overlapping groups. One, at the university, was led by Hans Mayer. The other centred on Mises, whose famed *privatseminar* met in his Chamber of Commerce office and drew not only the gifted younger economists, but also such philosophers, sociologists and political scientists as Felix Kaufmann, Alfred Schutz and Erik Voegelin. It was during this period that British economist Lionel Robbins came decisively under the influence of the intellectual ferment going on in Vienna. A distinctly important outcome of this contact was Robbins's highly influential book (Robbins, 1932). It was largely through this work that a number of key Austrian ideas came to be absorbed into the mainstream literature of 20th-century Anglo-American economics. In 1931 Robbins invited Hayek to lecture at the London School of Economics, and this led to Hayek's appointment to the Tooke chair at that institution.

Hayek's arrival on the British scene contributed especially to the development and widespread awareness of the ‘Austrian’ theory of the business cycle. Mises had sketched such a theory as early as 1912 (Mises, 1912, pp. 396–404). This theory attributed the boom phase of the cycle to intertemporal misallocation stimulated by ‘too low’ interest rates. This intertemporal misallocation consisted of producers initiating processes of production that implicitly anticipated a willingness on the part of the public to postpone consumption to a degree in fact inconsistent with the true pattern of time preferences. The subsequent abandonment of unsustainable projects constitutes the down phase of the cycle. Mises emphasized the roots of this theory in Wicksell, and in earlier insights of the British Currency School. Indeed Mises was tempted to challenge the appropriateness of the ‘Austrian’ label widely attached to the theory (Mises, 1943). But, as he recognized, the Austrian label had become firmly attached to the doctrine. Hayek's vigorous exposition and extensive development of the theory (Hayek, 1931; 1933; 1939) and his introduction (through the theory) of Böhm-Bawerkian capital-theoretic insights to the British public, unmistakably left Hayek's imprint on the fully developed theory, and taught the profession to see it as a central contribution of the Austrian School. Given all these developments it is apparent that we must consider the early 1930s as constituting in many ways the period of greatest Austrian School influence upon the economics profession generally. Yet this triumph was to be short-lived indeed.

With the benefit of hindsight it is perhaps possible to understand why and how this same period of the early 1930s constituted, in fact, a decisive, almost fatal, turning point in the fortunes of the School. Within a few short years the idea of a distinct Austrian School – except as an important, but bygone, episode in the history of economics – virtually disappeared from the economics profession. While Hans Mayer continued to occupy his chair in Vienna until after the Second World War, the group of prominent younger economists who had surrounded Mises soon dispersed (for political or other reasons), many of them to various universities in the United States. With Mises migrating in 1934 to Geneva and later to New York, with Hayek in London, Vienna ceased to be a centre for the vigorous continuation of the Austrian tradition. Moreover, many of the group were convinced that the important ideas of the Austrian School had now been successfully absorbed into mainstream economics. The emerging ascendancy of theoretical economics, and thus the eclipse of historicist and anti-theoretical approaches to economics, no doubt permitted the Austrians to believe that they had finally prevailed, that there was no longer any particular need to cultivate a separate Austrian version of economic theory. A 1932 statement by Mises captures this spirit. Referring to the usual separation of economic theorists into three schools of thought, 'the Austrian and the Anglo-American Schools and the School of Lausanne', Mises (citing Morgenstern) emphasized that these groups 'differ only in their mode of expressing the same fundamental idea and that they are divided more by their terminology and by peculiarities of presentation than by the substance of their teachings' (Mises, 1933, p. 214). Yet the survival and development of an Austrian tradition during and subsequent to the Second World War, largely through the work of Mises himself and of Hayek, deserves and requires attention.

Fritz Machlup has, on several occasions (Machlup, 1981; 1982) listed six ideas as central to the Austrian School prior to the Second World War. There is every reason to agree that it was these six ideas that expressed the Austrian approach as understood, say, in 1932. These ideas were: (*a*) methodological individualism (not to be confused with political or ideological individualism, but referring to the claim that economic phenomena are to be explained by going back to the actions of individuals); (*b*) methodological subjectivism (recognizing that the actions of individuals are to be understood only by reference to the knowledge, beliefs, perception and expectations of these individuals); (*c*) marginalism (emphasizing the significance of prospective *changes* in relevant magnitudes confronting the decision maker); (*d*) the influence of utility (and diminishing marginal utility) on demand and thus on market prices; (*e*) opportunity costs (recognizing that the costs that affect decisions are those that express the most important of the alternative opportunities being sacrificed in employing productive services for one purpose rather than for the sacrificed alternatives); (*f*) time structure of consumption and production (expressing time preferences and the productivity of 'roundaboutness').

It seems appropriate, however, to comment further on this list. (1) With varying degrees of emphasis most modern microeconomics incorporates all of these ideas, so that (2) this list supports the cited Morgenstern–Mises statement emphasizing the common ground shared by *all* schools of economic theory. However (3) subsequent developments in the work of Mises and Hayek suggest that the list of six Austrian ideas was not *really* complete. While few Austrians at the time (of the early 1930s) were perhaps able to identify additional Austrian ideas, such additional insights were in fact implicit in the Austrian tradition and were to be articulated explicitly in later work. From this perspective, then, (4) important *differences* separate Austrian economic theory from the mainstream developments in microeconomics, particularly as these latter developments proceeded from the 1930s onwards. It was left for Mises and Hayek to articulate these differences and thus preserve a unique Austrian 'presence' in the

profession.

Later developments in Austrian economics

One early expression of such differences between the Austrian understanding of economic theory and that of other schools, was Hans Mayer's paper criticizing ‘functional price theories’ and calling for the ‘genetic-causal’ method (Mayer, 1932). Here Mayer was criticizing equilibrium theories of price that neglected to explicate the *sequence* of actions leading to market prices. To understand this sequence one must understand the causal genesis of the component actions in the sequence. In the light of the later writings of Mises and Hayek, it seems reasonable to recognize Mayer as having placed his finger on an important and distinctive element embedded in the Austrian understanding. Yet the Austrians themselves during the 1920s (and such students of their works as Lionel Robbins) seemed to have missed this insight. What appears to have helped Hayek and Mises articulate this hitherto overlooked element was the well-known interwar debate concerning the possibility of economic calculation under central planning. A careful reading of the contributions to that debate suggests that it was in reaction to the ‘mainstream’ equilibrium arguments of their opponents that Mises and Hayek made explicit the emphasis on process, learning and discovery to be found in the Austrian understanding of markets (Lavoie, 1985).

Mises had argued that economic calculation calls for the guidance supplied by prices; since the centrally planned economy has no market for productive factors, it cannot use factor prices as guides. Oskar Lange and others countered that prices need not be market prices; that guidance could be provided by non-market prices, announced by the central authorities, and treated by socialist managers ‘parametrically’ (just as prices are treated by producers in the theory of the firm, in perfectly competitive factor and product markets). It was in response to this argument that Hayek developed his interpretation of competitive market processes as processes of discovery during which dispersed information comes to be mobilized (Hayek, 1949, chs 2, 4, 5, 7, 8, 9). An essentially similar characterization of the market process (without the Hayekian emphasis on the role of knowledge, but with an accent on entrepreneurial activity in a world of open-ended, radical uncertainty) was presented by Mises during the same period (Mises, 1940; 1949). In the light of these Mises–Hayek developments in the theory of market process (and recognizing that these developments constituted the articulation of insights taken for granted in the early Austrian tradition: Kirzner, 1985; Jaffé, 1976), it seems reasonable to add the following to Machlup's list of ideas central to the Austrian tradition: (*g*) markets (and competition) as processes of learning and discovery; (*h*) the individual decision as an act of choice in an essentially uncertain context (where the identification of the relevant alternatives is part of the decision itself). It is these latter ideas that have come to be developed in and made central to the revived attention to the Austrian tradition that, stemming from the work of Mises and Hayek, has emerged in the United States in recent decades.

Austrian economics today

As a result of these somewhat varied developments in the history of the Austrian School since 1930, the term ‘Austrian economics’ has come to evoke a number of different connotations in contemporary professional discussion. Some of these connotations are, at least partly, overlapping; others are, at least

partly, mutually inconsistent. If seems useful, in disentangling these various perceptions, to identify a number of different meanings that have come to be attached to the term ‘Austrian economics’ in the 1980s. The present status of the Austrian School of economics is, for better or for worse, encapsulated in these current perceptions.

1. For many economists the term ‘Austrian economics’ is strictly a historical term. In this perception the existence of the Austrian School did not extend beyond the early 1930s: Austrian economics was partly absorbed into mainstream microeconomics, and partly displaced by emerging Keynesian macroeconomics. To a considerable extent this view seems to be that held by economists in Austria today. Economists (and other intellectuals) in Austria today are thoroughly cognizant of – and proud of – the earlier Austrian School, as evidenced by several commemorative conferences held in Austria in recent years, and by several related volumes (Hicks and Weber, 1973; Leser, 1986), but see themselves today simply as a part of the general community of professional economists. Erich Streissler, holder of the chair occupied by Menger, Wieser and Mayer, has written extensively, and with the insights and scholarship of one profoundly influenced by the Austrian tradition, concerning numerous aspects of the Austrian School and its principal representatives (Streissler, 1969; 1972; 1973; 1986).

2. For a number of economists the adjective ‘Austrian’ has come to mark a revival of interest in Böhm-Bawerkian capital-and-interest theory. This revival has emphasized particularly the time dimension in production and the productivity of roundaboutness. Among the contributors to this literature should be mentioned Hicks (1973), Bernholz (1971; 1973), Faber (1979) and Orosel (1981). In this literature, then, the term ‘Austrian’ has very little to do with the general subjectivist Mengerian tradition (which had, as noted earlier, certain reservations in regard to the Böhm-Bawerkian theory).

3. For other economists (and non-economists) the term ‘Austrian economics’ has come to be associated less with a unique methodology, or with specific economic doctrines, than with libertarian ideology in political and social discussion. For these observers, to be an Austrian economist in the 1980s is simply to be in favour of free markets. Machlup (1982) has noted (and partly endorsed) this perception of the term ‘Austrian’. He has ascribed it, particularly, to the impact of the work of Mises. Mises’ championship of the market cause was so prominent, and his identification as an Austrian was at the same time so unmistakable, that it is perhaps natural that his strong policy pronouncements in support of unhampered markets came to be perceived as the core of Austrianism in modern times. This has been reinforced by the work of a leading US follower of Mises, Murray N. Rothbard, who was also prominent in libertarian scholarship and advocacy. Other observers, however, would question this identification. While, as earlier noted, many of the early contributions of the Austrian School were seen as sharply antagonistic to Marxian thought, the school on the whole maintained an apolitical stance. Among the founders of the school, Wieser was in fact explicit in endorsing the interventionist conclusions of the German Historical School (Wieser, 1914, pp. 490 ff). While both Mises and Hayek provocatively challenged the possibility of efficiency under socialism, they too, emphasized the *wertfrei* character of their economics. Both writers would see their free market stance at the policy level as related to, but not as central to, their Austrianism.

4. For many in the profession the term ‘Austrian economics’ has come, since about 1970, to refer to a revival of interest in the ideas of Carl Menger and the earlier Austrian School, particularly as these ideas have been developed through the work of Mises and Hayek. This revival has occurred particularly in the United States, where a sizeable literature has emerged from a number of economists. This literature

includes, in particular, works by Murray N. Rothbard (1962), Israel Kirzner (1973), Gerald P. O'Driscoll (1977; 1985), Mario J. Rizzo (O'Driscoll and Rizzo, 1985), and Roger W. Garrison (1978; 1982; 1985). The thrust of this literature has been to emphasize the differences between the Austrian understanding of markets as processes, and that of the equilibrium theorists whose work has dominated much of modern economic theory. As a result of this emphasis, this sense of the term 'Austrian economics' has often (and only partly accurately; see White, 1977, p. 9) come to be understood as a refusal to adopt modern mathematical and econometric techniques – which standard economics adopted largely as a result of its equilibrium orientation. The economists in this group of modern Austrians (sometimes called neo-Austrian) do see themselves as continuators of an earlier tradition, sharing with mainstream neoclassical economics an appreciation for the systematic outcomes of markets, but differing from it in its understanding of how these outcomes are in fact achieved. Largely as a result of the activity of this group, many classic works of the early Austrians have recently been republished in original or translated form, and have attracted a considerable readership both inside and outside the profession.

5. Yet another current meaning loosely related to the preceding sense of the term has come to be associated with the term 'Austrian economics'. This meaning refers to an emphasis on the radical uncertainty that surrounds economic decision making, to an extent that implies virtual rejection of much of received microeconomics. Ludwig Lachmann (1976) has identified the work of G.L.S. Shackle as constituting in this regard the most consistent extension of Austrian (and especially of Misesian) subjectivism. Lachmann's own work (1973; 1977; 1986) has, in the same vein, stressed the indeterminacy of both individual choices and market outcomes.

This line of thought has come to imply serious reservations concerning the possibility of systematic theoretical conclusions commanding significant degrees of generality. This connotation of the term 'Austrian economics' thus associates it with a stance sympathetic, to a degree, towards historical and institutional approaches. Given the prominent opposition of earlier Austrians to these approaches, this association has, as might be expected, been seen as ironic or even paradoxical by many observers (including, especially, modern exponents of the broader tradition of the Austrian School of economics).

[An earlier article on the Austrian School of economics was begun and substantially drafted by Professor Friedrich A. Hayek – himself a Nobel laureate in economics whose celebrated contributions are deeply rooted in the Austrian tradition. The present author gratefully acknowledges his indebtedness (in the writing of this essay) to the characteristic scholarship and treasure trove of facts contained in Professor Hayek's unfinished article, as well as to Professor Hayek's other numerous studies that relate to the history of the Austrian School.]

See Also

- Böhm-Bawerk, Eugen von
- competition, Austrian
- Hayek, Friedrich August von
- imputation
- Menger, Carl
- Mises, Ludwig Edler von

- Wiesner, Friedrich Freiherr, (Baron) von

Bibliography

Bernholz, P. 1971. Superiority of roundabout processes and positive rate of interest. A simple model of capital and growth. *Kyklos* 24, 687–721.

Bernholz, P. and Faber, M. 1973. Technical superiority of roundabout processes and positive rate of interest. A capital model with depreciation and n-period horizon. *Zeitschrift für die gesamte Staatswissenschaften* 129, 46–61.

Böhm-Bawerk, E. von. 1884. *Geschichte und Kritik der Kapitalzins-Theorien*. English trans. as vol. 1 of *Capital and Interest*, South Holland, IL: Libertarian Press, 1959.

Böhm-Bawerk, E. von. 1886. Grundzuge der Theorie des Wirtschaftlichen Guterwerths. *Conrad's Jahrbuch* 13, 1–88, 477–541.

Böhm-Bawerk, E. von. 1889. *Positive Theorie des Kapitales*. Innsbruck: Wagner.

Böhm-Bawerk, E. von. 1891. The Austrian economists. *Annals of the American Academy of Political and Social Science* 1, 361–84.

Böhm-Bawerk, E. von. 1896. *Zum Abschluss des Marxschen Systems*. In Trans. (1898) as *Karl Marx and the Close of his System*, ed. P. Sweezy, New York: Kelley, 1949.

Bostaph, S. 1978. The methodological debate between Carl Menger and the German Historicists. *Atlantic Economic Journal* 6(3), 3–16.

Bukharin, N. 1914. *The Economic Theory of the Leisure Class*. Trans. from Russian, London: M. Lawrence, 1927; reprinted, New York: Monthly Review Press, 1972.

Faber, M. 1979. *Introduction to Modern Austrian Capital Theory*. Berlin: Springer.

Garrison, R.W. 1978. Austrian macroeconomics: a diagrammatical exposition. In *New Directions in Austrian Economics*, ed. L.M. Spadaro. Kansas City: Sheed, Andrews & McMeel.

Garrison, R.W. 1982. Austrian economics as the middle ground: comment on Loasby. In *Method, Process, and Austrian Economics: Essays in Honor of Ludwig von Mises*, ed. I.M. Kirzner. Lexington, MA: Lexington Books.

Garrison, R.W. 1985. Time and money: the universals of macroeconomic theorizing. *Journal of*

Macroeconomics 6, 197–213.

Grassl, W. and Smith, B., eds. 1986. *Austrian Economics, Historical and Philosophical Background*. New York: New York University Press.

Gross, G. 1884. *Die Lehre von Unternehmergewinn*. Leipzig.

Hausman, D.M. 1981. *Capital, Profits, and Prices*. New York: Columbia University Press.

Hayek, F.A. 1931. *Prices and Production*. London: Routledge & Sons.

Hayek, F.A. 1933. *Monetary Theory and the Trade Cycle*. London: Jonathan Cape.

Hayek, F.A. 1939. *Profits, Interest and Investment: and Other Essays on the Theory of Industrial Fluctuations*. London: Routledge & Kegan Paul.

Hayek, F.A. 1949. *Individualism and Economic Order*. London: Routledge & Kegan Paul.

Hayek, F.A. 1968. Economic thought VI: the Austrian School. In *International Encyclopedia of the Social Sciences*, ed. D.L. Sills. New York: Macmillan.

Hicks, J. 1973. *Capital and Time: A Neo-Austrian Theory*. Oxford: Clarendon Press.

Hicks, J.R. and Weber, W. 1973. *Carl Menger and the Austrian School of Economics*. Oxford: Clarendon Press.

Jaffé, W. 1976. Menger, Jevons and Walras de-homogenized. *Economic Inquiry* 14, 511–24.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan.

Kauder, E. 1965. *A History of Marginal Utility Theory*. Princeton: Princeton University Press.

Kirzner, I.M. 1973. *Competition and Entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, I.M. 1981. Mises and the renaissance of Austrian Economics. In *Homage to Mises, the First Hundred Years*, ed. J.K. Andrews, Jr. Hillsdale, MI: Hillsdale College Press.

Kirzner, I.M. 1985. Comment on R.N. Langlois, ‘From the Knowledge of Economics to the Economics of Knowledge: Fritz Machlup on Methodology and on the “Knowledge Society”’. In *Research in the History of Economic Thought and Methodology*, ed. Warren J. Samuels. Greenwich, CT: JAI.

Komorzynski, J. von 1889. *Der Werth in der isolirten Wirthschaft*. Vienna: Manz.

Lachmann, L. 1973. *Macro-economic Thinking and the Market Economy*. London: Institute of Economic Affairs.

Lachmann, L. 1976. From Mises to Shackle: an essay on Austrian economics and the Kaleidic Society. *Journal of Economic Literature* 14, 54–62.

Lachmann, L. 1977. Austrian economics in the present crisis of economic thought. In *Capital, Expectations, and the Market Process*. Kansas City: Sheed, Andrews & McMeel.

Lachmann, L. 1986a. Austrian Economics under fire: the Hayek-Sraffa duel in retrospect. In Grassl and Smith (1986).

Lachmann, L. 1986b. *The Market as a Process*. Oxford: Basic Blackwell.

Lavoie, D. 1985. *Rivalry and Central Planning: The Socialist Calculation Debate Reconsidered*. Cambridge: Cambridge University Press.

Leser, N., ed. 1986. *Die Wiener Schule der Nationalökonomie*. Vienna: Hermann Böhlau.

Machlup, F. 1981. Ludwig von Mises: the academic scholar who would not compromise. *Wirtschaftspolitischen Blätter* 28(4), 6–14.

Machlup, F. 1982. Austrian economics. In *Encyclopedia of Economics*, ed. D. Greenwald. New York: McGraw-hill.

Mataja, V. 1884. *Der Unternehmergewinn*. Vienna.

Mayer, H. 1932. Der Erkenntniswert der Funktionellen Preistheorien. In *Die Wirtschaftstheorie der Gegenwart*, ed. H. Mayer. Vienna: Springer.

McCulloch, J.H. 1977. The Austrian theory of the marginal use and of ordinal marginal utility. *Zeitschrift für Nationalökonomie* 37(3–4), 249–80.

Menger, C. 1871. *Grundsätze der Volkwirthschaftslehre*. In Trans. as *Principles of Economics*, ed. J. Dingwall and B.F. Hoselitz, 1950; reprinted, New York: New York University Press, 1981.

Menger, C. 1883. *Untersuchungen über die Methode der Sozialwissenschaften und der politischen Ökonomie insbesondere*. In Trans. F.J. Nock as *Problems of Economics and Sociology*, ed. L. Schneider, Urbana: University of Illinois Press, 1963. Repr. as *Investigations into the Method of the Social Sciences*

with Special Reference to Economics with a new Introduction by L.H. White, New York and London: New York University Press, 1985.

Menger, K., Jr. 1973. Austrian marginalism and mathematical economics. In *Carl Menger and the Austrian School of Economics*, ed. J.R. Hicks and W. Weber. Oxford: Clarendon Press.

Meyer, R. 1887. *Das Wesen des Einkommens: Eine volkswirthschaftliche Untersuchung*. Berlin: Hertz.

Mises, L. von. 1912. *Theorie des Geldes und der Umlaufsmittel*. Trans. as *Theory of Money and Credit* (1934), Indianapolis: Liberty Classics, 1980.

Mises, L. von. 1933. *Grundprobleme der Nationalökonomie*. Trans. as *Epistemological Problems of Economics*, Princeton: Van Nostrand, 1960.

Mises, L. von. 1940. *Nationalökonomie, Theorie des Handelns und Wirtschaftens*, Geneva: Editions Union.

Mises, L. von. 1943. ‘Elastic expectations’ and the Austrian theory of the trade cycle. *Economica* 10, 251–2.

Mises, L. von. 1949. *Human Action: A Treatise on Economics*. New Haven: Yale University Press.

Mises, L. von. 1969. *The Historical Setting of the Austrian School of Economics*. New Rochelle: Arlington House.

Mises, L. von. 1978. *Notes and Recollections*. South Holland, IL: Libertarian Press.

O’Driscoll, G.P., Jr. 1977. *Economics as a Coordination Problem: The Contributions of Friedrich A. Hayek*. Kansas City: Sheed, Andrews & McMeel.

O’Driscoll, G.P., Jr. and Rizzo, M.J. 1985. *The Economics of Time and Ignorance*. Oxford: Basil Blackwell.

Orosel, G.O. 1981. Faber's modern Austrian capital theory: a critical survey. *Zeitschrift für Nationalökonomie* 41, 141–55.

Philippovich, E. von Philippsberg 1893. *Grundriss der Politischen Ökonomie*. Freiburg: Mohr.

Robbins, L. 1932. *The Nature and Significance of Economic Science*. London: Macmillan.

Rothbard, M.N. 1962. *Man, Economy, and State: A Treatise on Economic Principles*. Princeton: Van

Nostrand.

Sax, E. 1887. *Grundlegung der Theoretischen Staatswirtschaft*. Vienna: Holder.

Schumpeter, J.A. 1908. *Das Wesen und der Hauptinhalt der Theoretischen Nationalökonomie*. Leipzig: Duncker & Humblot.

Schumpeter, J.A. 1912. *Theorie der wirtschaftlichen Entwicklung*. Leipzig: Duncker & Humblot. English trans. as *The Theory of Economic Development*, Cambridge, MA: Harvard University Press, 1934.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Streissler, E. 1969. Structural economic thought: on the significance of the Austrian School today. *Zeitschrift für Nationalökonomie* 29(3–4), 237–66.

Streissler, E. 1972. To what extent was the Austrian School marginalist? *History of Political Economy* 4, 426–61.

Streissler, E. 1973. The Mengerian tradition. In *Carl Menger and the Austrian School of Economics*, ed. J.R. Hicks and W. Weber. Oxford: Clarendon Press.

Streissler, E. 1986. Arma virumque cano. Friedrich von Wieser, the bard as economist. In Leser (1986).

Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz.

White, L.H. 1977. *The Methodology of the Austrian School Economists*. Rev. edn, Auburn, AL: The Ludwig von Mises Institute of Auburn University, 1984.

Wieser, F. von. 1884. *Ursprung des Wirtschaftlichen Wertes*. Vienna: Hölder.

Wieser, F. 1889. *Der Naturliche Werth*. Vienna: Hölder. In Trans. as *Natural Value*, ed. W. Smart, London: Macmillan, 1893; reprinted, New York: Kelley, 1956.

Wieser, F. von. 1914. *Theorie der Gesellschaftlichen Wirtschaft*. Tübingen: Mohr. Trans. as *Social Economics*, London: G. Allen & Unwin, 1927; reprinted, New York: Kelley, 1967.

Zuckerkandl, R. 1889. *Zur Theorie des Preises*. Leipzig: Stein.

How to cite this article

Kirzner, Israel M. "Austrian economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000170> doi:10.1057/9780230226203.0078

The New Palgrave Dictionary of Economics Online

Averch – Johnson effect

H.A. Averch
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Averch–Johnson effect is produced when fair rate of return regulation encourages a firm to invest more than is consistent with the minimization of its costs. This can happen when the allowed rate of return exceeds the cost of capital, since the difference between the two represents pure profit. Detailed descriptions of actual regulatory processes may be useful in suggesting guides for action, since actual outcomes depend as much on political and bureaucratic necessity as they do on economic analysis and ‘rational’ benefit–cost estimates.

Keywords

adjustment costs; allocational efficiency; Averch, H. A.; Averch–Johnson effect; bounded information; bounded rationality; capital–labour ratio; cost of capital; cost–benefit analysis; electricity utilities; innovation; input–output analysis; Lagrange multiplier; operational gaming; production functions; public utility pricing and finance; rate of return; rate of return regulation; regulation, political economy of; research and development; simulation; stochastic demand; technical change; technical efficiency

Article

The Averch–Johnson effect explores some unintended consequences of fair rate of return regulation (Averch and Johnson, 1962). Such regulation may cause the firm to select excessively capital-intensive technologies, and, thereby, not produce its output at minimum social cost. Specifically, the main Averch–Johnson result is that the capital–labour ratio selected by a profit-maximizing, regulated firm will be greater than that consistent with a cost-minimizing one for any output it chooses to produce. If the fair rate of return is greater than the cost of capital, a firm will have an incentive to invest as much as it can consistent with its production possibilities, because the difference between the allowed rate and its actual cost of capital is pure profit.

This brief overview discusses (1) the effects of rate of return regulation on a monopolist's inputs and outputs; (2) the effects on incentives to innovate; (3) the empirical evidence on the existence and strength of the Averch–Johnson effect; and (4) some of the main theoretical extensions. Since 1962, the Averch–Johnson literature has been extended to include objectives other than profit maximization, more subtle interactions between regulators and firms and more complex market conditions. By making the models more complex, the number of possible regulatory outcomes has been enlarged. But the basic Averch–Johnson result, as stated above, has proven remarkably robust. So the discussion here focuses on this result and some of the main corollary results.

Choice of inputs in the basic Averch – Johnson model

Suppose there exists a single-product, profit-maximizing monopolist subject to rate of return regulation. The firm's production function is

$$Q = F(K, L),\ K, L > 0,\ F(0, L) = F(K, 0) = 0,$$
$$F_1, F_2 > 0,\ \ F_{11}, F_{22} < 0\ .$$
(1)

Suppose the firm's inverse demand function is

$$P = P(Q),\ \ P'(Q) < 0.$$
(2)

Profit is

$$\prod = PQ - rK - wL.$$
(3)

Assuming, as is standard, that there is no depreciation and that the acquisition cost of capital is adjusted to one, the rate of return constraint can be written

$$(PQ - wL)\ /\ K \leq s \text{ or } PQ - wL - sK \leq 0,$$
(4)

or

$$\prod \leq (s - r)K,$$
(5)

where s is the allowed rate of return. The fair rate of return is taken to be at least as great as the cost of capital ($s > r$) and less than the rate the firm could earn if it were unconstrained. Consequently, the constraint is effective, and the firm maximizes

$$\prod = PQ - rK - wL$$
(6)

subject to (4) or (5). Letting R equal total revenue PQ, the necessary first order conditions are

$$(1-\lambda)R'F_1 - \lambda(s-r) = 0 \tag{7}$$

$$(1-\lambda)R'F_2 - (1-\lambda)w = 0 \tag{8}$$

$$R - wL - sK = 0. \tag{9}$$

λ is the standard Averch–Johnson Lagrange multiplier. Given that the constraint is effective, that $s > r$, and that the revenue function $R = PQ$ is concave, the multiplier λ is greater than zero and less than one. Consequently, the marginal rate of substitution of capital for labour for the regulated firm is

$$-dL/dK = [r - (\lambda/1-\lambda)(s-r)]/w < r/w. \tag{10}$$

For any given output, the firm will not minimize cost, since this requires that the firm's marginal rate of technical substitution be equal to r/w.
This result can be shown graphically in several different ways (Baumol and Klevorick, 1970; Zajac, 1970). Zajac's formulation is shown here. Fig. 1 shows the regulatory constraint (9) in relation to the firm's isoquants.

Figure 1
The Averch–Johnson (A-J) effect

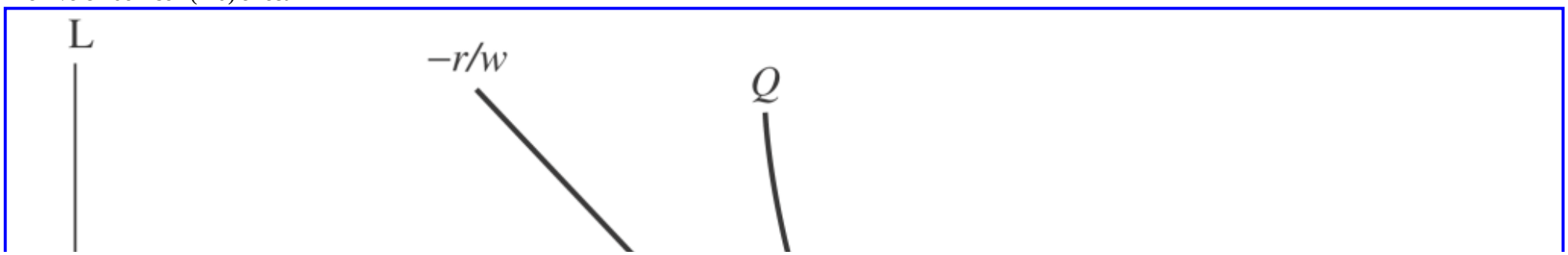

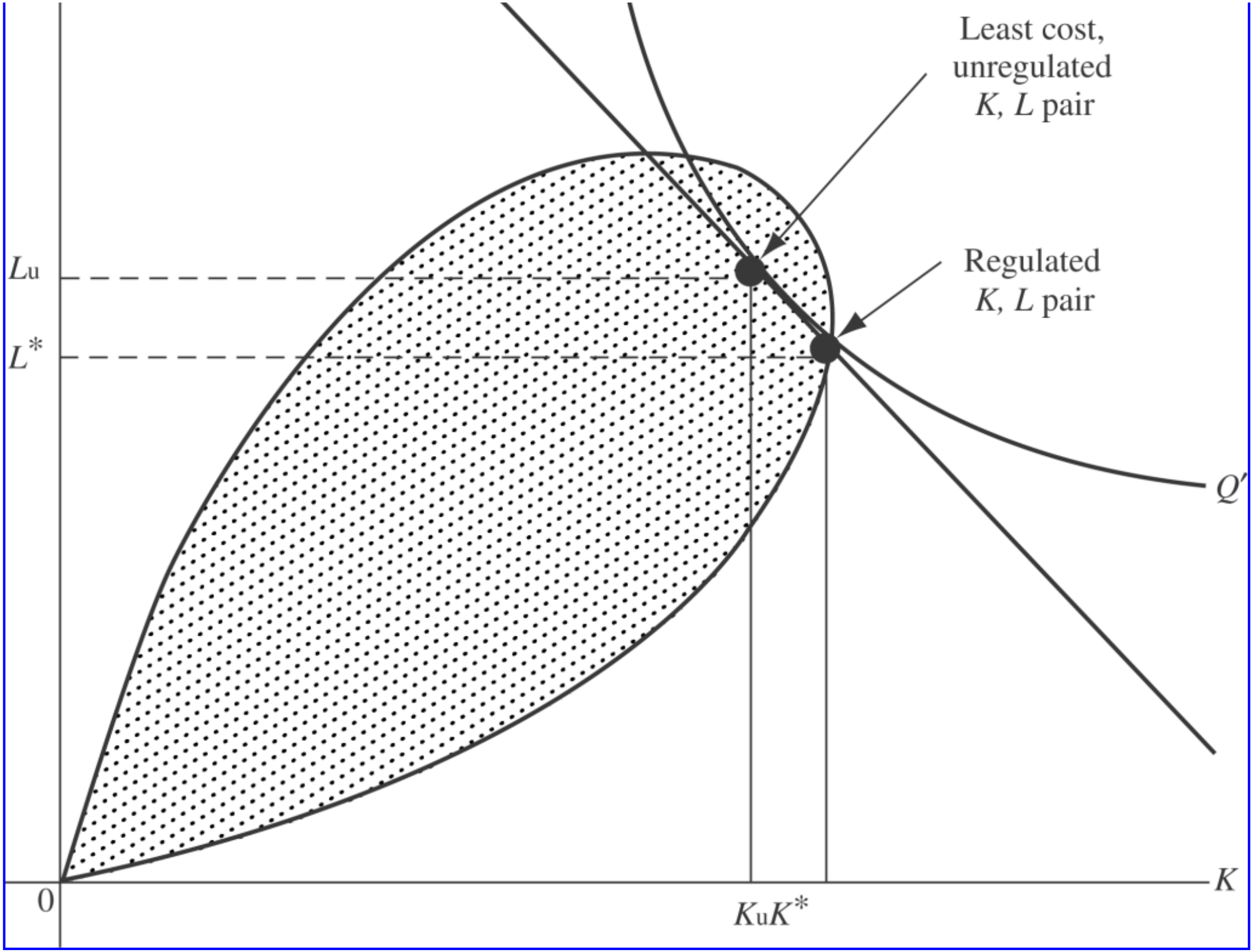
Least cost,
unregulated
K, L pair
Regulated
K, L pair
Lu
L*
Q′
K
0
KuK*

The shaded region inside the constraint curve shows input combinations resulting in rates of return greater than *s*. The firm wants to be as far up to the right on the constraint curve as possible, because, from (5), every increment of capital increases profit. Consequently, the firm will operate at the rightmost point of the constraint curve.

The output for this rightmost point can be obtained from the isoquant that intersects the constraint curve at its rightmost point. However, the least cost combination of capital and labour for producing this output, where $-dL/dK = r/w$, lies inside the proscribed shaded area, on the firm's efficient expansion path. For any given output, the firm cannot simultaneously be on the cost-minimizing price line with slope $-r/w$ and on the constraint curve.

The output of the regulated firm

One of the original rationales for regulation was that it would increase allocative efficiency by forcing monopolists to offer more output than ordinarily they would. If a larger output were always the result of rate of return regulation, then decreases in technical efficiency would be compensated by increases in allocational efficiency. In principle, regulatory agencies could seek an *s* that just balanced the marginal benefits of increased output against the marginal costs of decreased efficiency (Klevorick, 1971; Sheshinski, 1971; Bailey, 1973; Callen, Mathewson, and Mohring, 1976).

Increasing output, however, is not inevitable. The firm will use greater quantities of capital as *s* falls towards *r*, but the amount of labour the firm chooses to use will not necessarily be larger, and so output need not be larger. However, if labour is not an inferior input – the most likely case – then the optimal amount of labour for a regulated monopoly will also increase over the unregulated one, and, consequently, so will output (Baumol and Klevorick, 1970; Bailey, 1973). Firms with linear, homogeneous production functions will produce greater output. Given two firms with identical positive, homogeneous production functions – one regulated, one unregulated – the output of the unregulated one becomes a lower bound on regulated output, and the output such that ‘regulated’ average cost equals price becomes an upper bound (Murphy and Soyster, 1982).

Technological change and the regulated firm

Even if regulated firms are inefficient in static situations, technological change conceivably could induce more output through cost reductions. And rate of return regulation might conceivably induce regulated firms to be more innovative than unregulated firms. Regulation usually guarantees some profits, if not maximal ones, and these could be used for innovation.

If technological change is exogenous to the firm, but is factor-augmenting, then the optimal constrained K^* rises (Westfield, 1971; Magat, 1976). However, factor-augmenting technological advance will not necessarily result in increased output, since the firm may again use less labour to produce its output. Technological change, of course, is not usually entirely exogenous. Through their own research and development (R&D), firms gain knowledge of feasible innovation possibilities. Profit-maximizing firms subject to both a rate of return constraint and their own innovation possibilities constraint can, depending on production conditions, choose more labour-augmenting technologies than they would without regulation, reinforcing the bias the regulated firm has towards relatively capital-intensive technologies (Smith, 1974, 1975; Okuguchi, 1975).

In any case, regulation does not unambiguously increase innovation possibilities. The R&D expenditures of the regulated firm are not always larger than those that an unregulated firm would select under the same production and demand conditions (Magat, 1976). Furthermore, there is no systematic evidence that regulated firms select more high payoff R&D projects than unregulated ones and much anecdotal evidence to indicate that they are highly conservative.

Empirical tests

In the mid-1970s and early 1980s there were a number of attempts to determine whether Averch–Johnson effects actually existed and whether, if they existed, they imposed significant social costs. The empirical investigations used different tests for the effect and different data sets, most, however, relating to electric utilities. Unsurprisingly, the empirical evidence from these efforts was mixed. But overall the number of empirical investigations that find some evidence for the Averch–Johnson effect or its behavioural consequences outnumber those that find no evidence.

Using different methods but similar data, Courville (1974) and Spann (1974) concluded that Averch–Johnson effects existed. Petersen (1975), using a cost-minimizing version of the Averch–Johnson model, found that as the allowed rate of return approached the market cost of capital, capital costs increased as did the share of those costs in total costs. Hayashi and Trapani (1976) confirmed that regulated firms have a capital–labour ratio greater than the cost-minimizing one and that tightening *s* decreases efficiency. However, Boyes (1976) concluded that there was no effect.

Smithson (1978) reported that there was static inefficiency among electric utilities, but he could not confirm that lowering the rate of return caused the optimal capital stock to increase. Tapon and Van der Weide (1979) found that only strictly regulated electric utility firms exhibit Averch–Johnson effects, but that less than half of the industry appears to be so regulated. Regulatory lag permits firms to avoid Averch–Johnson effects, but raises the question of the worth of public investments in regulatory institutions.

Gollop and Karlson (1980), using data on electric utilities and an intertemporal model, found no evidence of input distortions. But Filer and Hallas (1983), testing for the effects of regulation in the interruptable gas industry, found rate of return regulation induced investment in additional storage capacity. Giordano (1983), examining utilities during 1964–77, concluded that there was capital bias during the 1960s, but not in the 1970s, because increasing regulatory lag and rapidly rising factor prices wiped it out. Such a finding was consistent with Averch–Johnson predictions, but it made Averch–Johnson effects perhaps less relevant in the 1980s. However, Averch–Johnson effects continue to be reported. Mirucki (1984), for example, concludes that the Canadian Bell system overinvests in capital and does not minimize costs.

Some investigators have argued that even if Averch–Johnson effects exist, their impact may be small, for there may be deterrents to technical inefficiency such as open entry (Sharkey, 1982). Others have argued that even if Averch–Johnson effects existed in the 1960s and 1970s, the relevant problem for utilities in the 1980s has been one of avoiding actual rates of return that fall below the allowed rate *s*. The 1980s problem is under-investment, because consumers are now able to prevent regulatory agencies from granting the price increases necessary to cover rising input costs (Navarro, 1983; Nelson, 1984; Rozek, 1984).

Theoretical extensions

The Averch–Johnson results have been extended and generalized in many ways. Three of the more significant extensions are discussed below.

Regulatory lag and stochastic review: The original Averch–Johnson result implicitly assumed regulatory agencies were always effective in enforcing the *s* they chose. In fact, regulators have great difficulty in keeping actual rates close to target rates. The regulatory process does its work episodically, through adjustments in price. Occasional adjustments, the regulator hopes, will bring the actual *s* to a tolerable level, if not back to the one originally set. Since regulation is a political, bureaucratic and legal process, there are almost always lags in enforcement. Consequently, firms may be able to escape the constraint for long periods of time (Bailey and Coleman, 1971; Klevorick, 1973).

Sufficient regulatory lag may allow the firm to be technically efficient at an unregulated monopolist's output, and it may induce more technological innovation than the case without enforcement lag. Continuous, effective regulation would prevent the firm from gaining the windfall profits that innovation may require, although Nelson argues that most technological change in the utilities industry is disembodied and has little relation to regulation (Nelson, 1984).

Demand uncertainty: Some authors argue that Averch–Johnson results hold only under some specifications of a stochastic demand function, but not others (Perrakis, 1976; Peles and Stein, 1976). Most of this discussion goes to whether the optimal capital stock would be larger, if regulated firms faced stochastic demands. If, as in the original Averch–Johnson discussion, we assume that the firm selects K and L as part of a simultaneous, ex ante optimization process, then the basic Averch–Johnson result, the inefficient capital–labour ratio, still holds under stochastic demand (Das, 1980).

Dynamic analysis: Some authors have introduced time explicitly into the original static Averch–Johnson model. For example, El-Hodiri and Takayama (1981) interpret the ‘Averch–Johnson effect’ to be a larger optimal K^* for a regulated firm than an unregulated one, and they show that this is true even with the adjustment costs attributable to time. However, much of this dynamic literature has been devoted to showing that, given a firm that maximizes the present value of profits over any number of time periods, one or more Averch–Johnson results do not hold or hold only under special conditions (Niho and Musaccio, 1983; Dechert, 1984).

The significance of the Averch – Johnson effect

From the stand-point of microeconomic theory, the original Averch–Johnson results provided impetus for increasingly complex, analytical models of the regulatory process. The Averch–Johnson approach suggested that much of the conventional, qualitative wisdom about regulation could be modelled and tested and that it was necessary to do so. Without thinking through all the potential consequences, actions and rules could be quite flawed without anyone intending them to be so. But flaws generally become apparent only after actions and rules have become entrenched, difficult to change or reverse. So explicit modelling of regulatory rules became part of the economist's stock in trade.

From a public policy perspective, the Averch–Johnson results and the very large volume of follow-on research have made economists, legislators and administrators far more sensitive to the potential unintended consequences of regulatory alternatives in general and not just rate of return alternatives. The Averch–Johnson effect has also figured directly in rate cases with utilities sometimes forced to defend themselves against charges of inefficiency.

Future lines of development

By injecting changes into the Averch–Johnson formulation one at a time, theoretical work has sought to make the model more representative of the actual regulatory process. One set of writers has pursued the effects of stochastic demand. Another set has worked on regulatory lag and stochastic review processes, but without stochastic demand. Yet another set has had the firm making global optimizations over time without either stochastic demands or random review. Economists interested in welfare issues have tried to determine an optimal fair rate of return from a strict economics perspective, but neglected politics and bureaucratic behaviour in setting rates. No model builders to date have addressed firms and regulators

as interacting organizations both suffering from bounded rationality and bounded information, although there is some recent work on what regulators might do when a firm's costs are unknown and it has incentives to lie (Baron and Meyerson, 1982).

Regulatory systems are so complex and interactive that the standard strategy of a priori modelling with a minimum number of plausible assumptions may no longer have sufficient pay off. In complex, interactive, relatively poorly understood situations, other analytical styles such as simulation or operational gaming can be useful. They have not been tried and probably should be. In fact, brute force, detailed descriptions of actual regulatory processes may be highly useful in suggesting guides for action. Regulation remains a problem in political economy. Actual outcomes depend as much on political and bureaucratic necessity as they do on economic analysis and ‘rational’ benefit–cost estimates.

See Also

- marginal and average cost pricing

Bibliography

Averch, H.A. and Johnson, L.L. 1962. Behavior of the firm under regulatory constraint. *American Economic Review* 52, 1052–69.

Bailey, E.E. 1973. *Economic Theory of Regulatory Constraint*. New York: D.C. Health.

Bailey, E.E. and Coleman, R.D. 1971. The effect of lagged regulation in an Averch–Johnson model. *Bell Journal of Economics and Management Science* 2(1), 278–92.

Baron, D.P. and Meyerson, R.B. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50(4), 911–30.

Baumol, W.J. and Klevorick, A.K. 1970. Input choices and rate-of-return regulation. *Bell Journal of Economics and Management Science* 1(2), 162–90.

Boyes, W.J. 1976. An empirical examination of the Averch–Johnson effect. *Economic Inquiry* 14(1), 25–35.

Callen, J., Mathewson, G.F. and Mohring, H. 1976. The benefits and costs of rate of return regulation. *American Economic Review* 66(3), 290–97.

Courville, L. 1974. Regulation and efficiency in the electric utility industry. *Bell Journal of Economics and Management Science* 5(1), 53–74.

Das, S.P. 1980. On the effect of rate of return regulation under uncertainty. *American Economic Review* 70, 456–60.

Dechert, W. 1984. Has the Averch–Johnson effect been theoretically justified? *Journal of Economic Dynamics and Control* 8(1), 1–17.

El-Hodiri, M. and Takayama, A. 1981. Dynamic behavior of the firm with adjustment costs under regulatory constraint. *Journal of Economic Dynamics and Control* 3(1), 29–41.

Filer, J.E. and Hallas, D.R. 1983. Empirical tests for the effect of regulation on firms and interruptible gas service. *Southern Economic Journal* 50(1), 195–205.

Giordano, J.N. 1983. The changing impact of regulation on the U.S. electric utility industry, 1964–1977. *Eastern Economic Journal* 9(2), 91–101.

Gollop, F.M. and Karlson, S.H. 1980. The electric power industry: an econometric model of intertemporal behavior. *Land Economics* 56(3), 299–314.

Hayashi, P.M. and Trapani, J.M. 1976. Empirical evidence on the Averch–Johnson model. *Southern Economic Journal* 42(3), 384–98.

Klevorick, A.K. 1971. The ‘optimal’ fair rate of return. *Bell Journal of Economics and Management Science* 2(1), 122–53.

Klevorick, A.K. 1973. The behavior of a firm subject to stochastic regulatory review. *Bell Journal of Economics and Management Science* 4(1), 57–88.

Magat, W. 1976. Regulation and the rate of direction of induced technical change. *Bell Journal of Economics and Management Science* 7(2), 478–96.

Mirucki, J. 1984. A study of the Averch–Johnson effect in the telecommunications industry. *Atlantic Economic Journal* 12(1), 121.

Murphy, F.H. and Soyster, A.L. 1982. Optimal output of the Averch–Johnson model. *Atlantic Economic Journal* 10(4), 77–81.

Navarro, P. 1983. Save now, freeze later: the real price of cheap electricity. *Regulation* 7, 31–6.

Nelson, R.A. 1984. Regulation, capital vintage, and technical change in the electric utility industry. *Review of Economics and Statistics* 66(1), 59–69.

Niho, Y. and Musaccio, R.A. 1983. Effects of regulation and capital market imperfection on the dynamic behavior of a firm. *Southern Economic Journal* 49, 625–36.

Okuguchi, K. 1975. The implications of regulation for induced technical change: comment. *Bell Journal of Economics and Management Science* 6, 703–5.

Peles, Y.C. and Stein, J.L. 1976. The effect of rate of return regulation is highly sensitive to the nature of uncertainty. *American Economic Review* 66, 278–89.

Perrakis, S. 1976. On the regulated price-setting monopoly firm with a random demand curve. *American Economic Review* 66, 410–16.

Petersen, H.C. 1975. An empirical test of regulatory effects. *Bell Journal of Economics and Management Science* 6(1), 111–26.

Rozek, R.P. 1984. The over-capitalization effect with diversification and cross subsidization. *Economics Letters* 6(1–2), 159–63.

Sharkey, W.W. 1982. *The Theory of Natural Monopoly*. New York: Cambridge University Press.

Sheshinski, E. 1971. Welfare aspects of a regulatory constraint: note. *American Economic Review* 61, 175–8.

Smith, V.K. 1974. The implications of regulation for induced technical change. *Bell Journal of Economics and Management Science* 5, 623–32.

Smith, V.K. 1975. The implications of regulation for induced technical change: reply. *Bell Journal of Economics and Management Science* 6, 706–7.

Smithson, C.W. 1978. The degree of regulation and the monopoly firm: further empirical evidence. *Southern Economic Journal* 44, 568–80.

Spann, R.M. 1974. Rate of return regulation and efficiency in production: an empirical test of Averch–Johnson thesis. *Bell Journal of Economics and Management Science* 5(1), 38–52.

Tapon, F. and Van der Weide, J. 1979. Effectiveness of regulation in the electric utility industry. *Journal of Economics and Business* 31(3), 180–9.

Westfield, F. 1971. Innovation and monopoly regulation. In *Technological Change in Regulated Industries*, ed. W.M. Capron. Washington, DC: Brookings Institution.

Zajac, E.E. 1970. A geometric treatment of Averch–Johnson's behavior of the firm model. *American Economic Review* 60, 117–25.

How to cite this article

Averch, H. A. "Averch–Johnson effect." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000186> doi:10.1057/9780230226203.0079

The New Palgrave Dictionary of Economics Online

Ayres, Clarence Edwin (1891 – 1972)

Warren J. Samuels
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Ayres, C. E.; culture; industrialization; institutional economics; methodological collectivism; technology; values; Veblen, T.

Article

Ayres was born on 6 May 1891 in Lowell, Massachusetts, and died on 25 July 1972 in Alamogordo, New Mexico. Trained as a philosopher, with degrees from Brown and Chicago (PhD, 1917), Ayres taught at Chicago, Amherst and Reed before moving to the University of Texas at Austin in 1930, from which he retired in 1968. For one year, 1924–5, he was an associate editor of *The New Republic*, associated with Herbert Croly, John Dewey, Alvin Johnson and R.H. Tawney. He had a lifelong correspondence with another philosophically oriented, but more traditional economist, Frank H. Knight. He was profoundly influenced by Thorstein Veblen and Dewey and became a, if not the, leader of institutional economics after World War II. A truly charismatic lecturer, at Texas he had long-lasting influence on a coterie of students who continued his teachings in their own careers. As his ideas evolved, particularly with regard to the nature of and relations between institutions and technology, his students came away with coherent but varying substantive understandings.

Ayres' formulation of institutionalism stressed that science was a system of belief, that human values were only means to the continuation and enhancement of the life process, that technology, as he defined it, was a (largely) beneficent driving force in social change, and that considerations of rightness tended in practice to be matters of tradition and custom.

Technology, to Ayres, meant the use of tools, but he defined tools increasingly broadly to include intangible symbols and organizations. Technology was the surging force governing economic welfare, and constituted what he considered to be an objective industrial or developmental process. His conception of technologically instrumental value and truth emphasized the transcultural values of workability and efficiency which form a continuum. Opposed to technology was the binding force of established institutions which, through sanctioning ceremonial behaviour in favour of established or vested interests, were hostile to the conceptual and economic progress generated by technology.

Economic progress was thus fundamentally a matter of industrialization; the logic of industrialization, or technological advancement in all respects, was continually at war with outworn, inhibitive institutions. Mankind's task was to develop new institutional forms and revise old ones in order to keep pace with evolving technology.

Ayres insisted that human behaviour was socially formed, and that for such behaviour to be explained and understood the economist had to study existing behaviour patterns (institutions) and general culture. In common with other institutionalists, Ayres insisted upon methodological collectivism and challenged what he considered to be the narrow focus on market equilibrium conditions maintained by mainstream economics.

Ayres influenced many development economists, who similarly perceived that modernization was inhibited by the continuance of traditional institutions or by the maintenance of positions of power antagonistic to modernization. More generally, Ayres, again like other institutionalists, argued that to understand the allocation of resources one had to go beyond the market to the institutions and cultural forces which, in part through adaptation to and incorporation of technology, constitute the real allocational mechanism. In a sense, the neoclassical juxtaposition between cost of production and utility became for Ayres something different, a juxtaposition between technology and the institutions which formed and weighted individual and collective choice.

See Also

- institutional economics

Selected works

1938. *The Problem of Economic Order*. New York: Farrar & Rinehart.

1944. *The Theory of Economic Progress*. Chapel Hill: University of North Carolina Press.

1946. *The Divine Right of Capital*. Boston: Houghton Mifflin.

1952. *The Industrial Economy*. Boston: Houghton Mifflin.

1961. *Toward a Reasonable Society*. Austin: University of Texas Press.

Bibliography

Breit, W. 1973. The development of Clarence Ayres's theoretical institutionalism. *Social Science Quarterly* 54, 244–57.

How to cite this article

Samuels, Warren J. "Ayres, Clarence Edwin (1891–1972)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000188>
doi:10.1057/9780230226203.0080

The New Palgrave Dictionary of Economics Online

Babbage, Charles (1791 – 1871)

Maxine Berg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Babbage, Charles; computing; demographics; division of labour; political economy; profit sharing; science policy; Statistical Society of London; technological underemployment; work organization

Article

Charles Babbage is rarely regarded as a major contributor to economic thought. His name is synonymous with the early origins of the computer, and he was an important figure in early 19th-century scientific circles. He was educated at Trinity College and Peterhouse, Cambridge, and while still a student started the Analytical Society with Herschel and Peacock, for reforming mathematics in Britain. His interest in mathematics was the foundation for his later contributions to science, economics and statistics. After Cambridge, Babbage moved to London, where he began his lifelong work on his analytical engine and became a leading participant in scientific circles. He joined the Royal Society and was a founding member of the Cambridge Philosophical Society and the Royal Astronomical Society. Later he was to be one of Newton's illustrious successors in the Lucasian Chair of Mathematics at Cambridge. But he was also a radical if maverick intellectual and political critic. He wanted to see science reformed, to see British science play a leading part in theoretical advance, and to see this science related closely to applied technology. He also demanded a role for the state in providing support for science and university education, and for establishing a policy on technology. He wrote a controversial attack on the Royal Society, *Reflections on the Decline of Science and Some of its Causes* (1830), and was one of the founding trustees of the British Association for the Advancement of Science, with the purpose of bringing science and technology, from the provinces as well as the metropolis, into the forefront of culture and society.
Babbage was an early promoter of industrial exhibitions as a part of meetings of the British Association; he participated in the Mechanics Section of the Association and later wrote a book on the Great Exhibition of 1851. He took part in the great controversies over religion and science in the period, and wrote the *Ninth Bridgewater Treatise* in 1837 (2nd edn 1838), conveying his belief in a Newtonian universe, with a scientific Deity.

Politically, Babbage was a liberal Whig; he chaired an election committee and stood twice for Finsbury. He denounced election corruption and bribery, attacked church preferments and tithes, and was a firm supporter of the Reform Bill. His political pamphlet on income tax showed a concept of moderate reform. He identified an electoral system based on one man, one vote with the ‘advance of socialism’, for where the poor were in the majority they would vote for low taxes for themselves and high taxes for the rich, ‘thus destroying private enterprise’.

Babbage's social and academic context was clearly that of early 19th-century liberal-scientific circles, and he participated in the salon culture of the day. But there was another very important component to his intellectual make-up: an abiding interest in practical mechanics and a fascination with contemporary industrial technology. He learned from manufacturers, large and small, mechanical engineers and above all from the skilled artisans he never ceased to praise. Developing the analytical engine was itself a task of scientific and mathematical reasoning combined with practical invention. The continental tour he made in 1827–8, which was to be so formative to his later work, was not in the company of a scientific friend or even a servant, but with one of the artisans who had worked on the building of the analytical engine. Travelling through the Low Countries, Germany, Austria and Italy with a prolonged stay in Naples, Babbage lost no opportunity to visit local workshops and factories.

His transcendence of contemporary social and intellectual boundaries was the real basis for his brilliant and utterly original foray into political economy. *On the Economy of Machinery and Manufactures* (1832) was immensely popular: there were four editions in two and a half years, it was reprinted in the United States and translated into four continental languages. Babbage wanted to present his readers with the mechanical principles of arts and manufactures, and he hoped also to be read by the intelligent working man. To this extent the book fell within the contemporary genre of industrial-technological literature; indeed, part of it had been published in 1829 as a part of the *Encyclopedia Metropolitana*. Tracts on the steam engine, histories of the cotton industry and industrial manuals, dictionaries and encyclopedias were very popular at the time. Andrew Ure's later *Philosophy of Manufactures* (1835), an extraordinary panegyric on the factory system and steam-powered machinery, was very much a product of this genre, but it completely lacked the analysis of Babbage's contribution. The latter was much more than popular industrial observation. It was an analysis based on economic principles, especially the Smithian account of the division of labour, of manufacturing technology and the organization of industrial work. Babbage's obvious first-hand knowledge of a wide variety of industrial and business processes, combined with general analysis of production systems, made the work a tour de force. At a time of anxiety and ambiguity over the reception of new technology, he also offered authoritative policy statements on a wide range of machinery issues including patent reform, export of machinery, crises of overproduction, and technological unemployment.

The book's intellectual situation in relation to political economy was not, however, easily apparent, and apart from Mill and Marx few appreciated its significance to their discipline. Before he wrote the book Babbage had intended to deliver a series of lectures in Cambridge on the Political Economy of Manufactures, but this never materialized. He himself conceded that his first edition did not profess to examine questions of political economy, and he attempted to correct this in the next edition by introducing three new chapters: ‘The new system of manufactures’, ‘The effects of machinery in reducing the demand for labour’, and ‘On money as a medium of exchange’. But most of the topics raised by Babbage were also foreign to contemporary classical political economy. Moving back to Smith, he analysed industrial organization and the microeconomics of the manufacturing firm, never

losing sight of technological constraints and opportunities.
The book was initially criticized for failing to give due attention to the factory system and steam-powered textile technology. But this was precisely its strength, for it analysed the factory and the workshop as parts of the more general organization of work, and examined machinery in the context of a more general discussion of technology, including skill. Babbage's close observation of skills and hand processes as well as machinery, of the workshop as well as the factory, was anyway a more accurate perception of contemporary industrial practice than a work concentrating on the outstanding and atypical phenomenon of the factory would have been.

> Babbage analysed what he called ‘the domestic economy of the factory’. He sought to specify what arrangement of production would succeed in selling articles at a minimum price, and he made a careful analysis of economies of scale in relation to the division of labour, distinguishing the dynamics of the factory from those of the workshop. He developed Smith's principle of the division of labour to a further refinement, introducing the significance of the division of skill, or the division of mental and manual labour. Vital, he believed, to the success of any organization of work was his ‘Babbage Principle’: that the master manufacturer by dividing the work to be executed into different processes, each requiring different degrees of skill or of force, can purchase exactly that precise quantity of both which is necessary for each process.

From this emphasis on the economy of skill, Babbage introduced novel discussions of the role of accounting, time and motion studies, communications innovations, and an analysis of machine functions. He was particularly concerned with the significance of precision and measurement in all processes, with the regularity of production, and with the planning of layout. He thus regarded as some of the greatest innovations not the celebrated power techniques themselves, but the processes which helped to make the new machines work properly, for example the steam engine governor and lubrication or grease. His interest in measurement led him to support all manner of instruments for counting machines and human actions; and he devised a detailed questionnaire as a basis for job studies and early time and motion studies. He also analysed as had no one before him the role of the speed of production and the intensity of labour in increasing output. Introducing machinery was only one incomplete route to increasing productivity; the productivity of labour could be rapidly improved through greater order, precision and labour discipline. Babbage noticed the convergence of technological and economic principles on topics such as velocity and copying; in a long discussion of the significance of copying techniques he pointed out the parallels between printing, casting and moulding, stamping and turning. This core analysis of workshop organization was complemented by topical commentary on profit sharing, technological unemployment and trade unions. An important radical departure on wages and labour was provided in his ‘New System of Manufactures’ which argued for a piece-rate wage system and profit sharing, if not cooperation, as the key to overcoming the long-standing worker opposition to machinery. This was the problem which Babbage along with many of his contemporaries believed to be the major brake on Britain's industrial progress. The system was a far-reaching proposal for a worker's stake in increasing productivity, for collective decision-making on hiring, dismissal and the organization of the works. Where the system prevailed, modern methods would be chosen and an extensive division of labour introduced, not to control and subordinate labour, but as a cooperative decision by workers for

the most efficient methods. The lengths to which his suggestions went were probably surpassed only by radical and Owenite cooperatives. When it came to practical implementation Babbage held out little hope of any appeal of the system to large established firms, but thought that groups of artisans and small firms would lead the way. Babbage's chapters on trade unions and machinery and employment were, however, the comments of a reformer not a radical. He attacked the truck system, but warned that trade unions could well lead to more rapid displacement of labour through machinery or industrial relocation. Dealing with technological underemployment, he used the case of hand-weaving and the power loom, arguing that the only solution lay in better workers' planning through such institutions as savings banks and friendly societies.

Babbage's use of practical observation and statistical data, and his critique of political economy's ‘closest philosophers’, induced him, with Richard Jones, J.E. Drinkwater, Malthus and Quetelet, to form a Statistics Section of the British Association in 1833, followed later by the Statistical Society of London. The Statistical Section, Section F, was confined to the presentation of statistical data, avoiding areas of political controversy. But the less restrictive London Society made its brief the connection of political economy to the statistical investigation of economic improvement. Babbage was the first President of Section F, and wrote several statistical papers. His earlier ‘Letter to the Right Hon T.P. Courtenay on the proportional number of births of the two sexes under different circumstances’ (1829) compared the demographic structures of the Kingdom of Naples, France, Prussia and Westphalia. Much later he wrote ‘On the statistics of lighthouses’ for the Brussels Congress of Statistics in 1853, and ‘The clearing house’, read to the London Statistical Society and printed in its memoirs in 1856. Babbage also wrote a book on insurance, *A Comparative View of the Various Institutions for the Assurance of Lives* (1826), and is remembered for his revised actuarial tables and his popular presentation of a difficult subject.

Babbage certainly produced an original and far-seeing economic analysis of industry in *On the Economy of Machinery and Manufactures*. He applied the principles of the division of labour he elaborated to his perception of the sciences. The ultimate result of the division of skills and especially the mental division of labour was the ‘science of calculation’. He argued that the science of calculation, like any technology, would be developed to a degree where machinery would take over all numerical calculation. Arithmetical exercise would thus be separated from mathematical reasoning, and the ‘science of calculation’ harnessed to the analytical engine would become the science of all sciences. Babbage's ultimate vision for Britain's industrial progress was one of a computer-run technology.

Selected works

1830. *Reflections on the Decline of Science and Some of its Causes*. London: B. Fellowes.

1832. *On the Economy of Machinery and Manufactures*. London: Charles Knight. 4th edn, 1835.

1838. *The Ninth Bridgewater Treatise*, 2nd edn. London: John Murray.

1851. *The Exposition of 1851*. London: John Murray.

1864. *Passages from the Life of a Philosopher*. London: Longman & Co.

Bibliography

Hyman, A. 1982. *Charles Babbage, Pioneer of the Computer*. Oxford: Oxford University Press.

Dubbey, J.M. 1978. *The Mathematical Work of Charles Babbage*. Cambridge: Cambridge University Press.

Rosenberg, Nathan. 1994. Charles Babbage: pioneer economist, chapter 2 of *Exploring the Black Box: Technology, Economics, and History*. Cambridge: Cambridge University Press.

Rosenberg, Nathan. 2000. Charles Babbage in a complex world, in *Complexity and the History of Economic Thought*, ed. D. Colander, London; Routledge.

Swade, Doron. 2001. *The Difference Engine: Charles Babbage and the Quest to Build the First Computer*. New York: Viking Press.

How to cite this article

Berg, Maxine. "Babbage, Charles (1791–1871)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000001> doi:10.1057/9780230226203.1870

The New Palgrave Dictionary of Economics Online

Bachelier, Louis (1870 – 1946)

Benott B. Mandelbrot

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bachelier, L.; Brownian motion; efficient markets; Levy, P.; martingales; Poincaré, H.; probability; random walk model; speculation; Wiener process

Article

Bachelier was born in Le Havre, France, on 11 March 1870 and died in Saint-Servan-sur-Mer, Ille-et-Vilaine, on 28 April 1946. He taught at Besançon, Dijon and Rennes and was professor at Besançon from 1927 to 1937.

The unrecognized genius is one of the stock figures of popular history, and it is also a platitude of which many examples dissolve upon careful examination. But the story of Louis Bachelier is in perfect conformity to all the clichés. He invented efficient markets in 1900, 60 years before the idea came into vogue. He described the random walk model of prices, ordinary diffusion of probability – also called Brownian motion – and martingales, which are the mathematical expression of efficient markets. He even attempted an empirical verification. But he remained a shadowy presence until 1960 or so, when his major work was revived in English translation.

This major work was his doctoral dissertation in the mathematical sciences, defended in Paris on 19 March 1900. Things went badly from the start: the committee failed to give it the ‘mention très honorable’, key to a university career. It was very late, after repeated failures, that Bachelier was appointed to the tiny University of Besançon. After he had retired, the university archives were accidentally set on fire and no record survives, not even one photograph. Here are a few scraps I have managed to put together.

We begin with the proverbial episode of the grain of sand, or the lack of a nail. Bachelier made a mathematical error that is recounted in a letter the great probabilist Paul Levy wrote me on 25 January 1964:

> I first heard of him around 1928. He was a candidate for a professorship at the University of Dijon. Gevrey, who was teaching there, came to ask my opinion. In a work published in

> 1913, Bachelier had defined Wiener's function (prior to Wiener) as follows: In each interval $[n\tau, (n+1)\tau]$, he considered a function $X(t|\tau)$ that has a constant derivative equal to either $+v$ or $-v$, the two values being equiprobable. He then proceeded to the limit $\tau \to 0$, keeping v constant, and claimed he was obtaining a proper function $X(t)$! Gevrey was scandalized by this error. I agreed with him and Bachelier was blackballed.

I had forgotten it when in 1931, reading Kolmogorov's fundamental paper, I came to ‘der Bacheliers Fall’. I looked up Bachelier's works, and saw that this error, which is repeated everywhere, does not prevent him from obtaining results that would have been correct if only he had written $v = C\tau^{-1/2}$, and that, prior to Einstein [1905] and prior to Wiener [circa 1925], he has seen some important properties of the Wiener function, namely, the diffusion equation and the distribution of $\max_{0<\tau<t} X(\tau)$.

We became reconciled. I had written to him that I regretted that an impression, produced by a single initial error, should have kept me from going on with my reading of a work in which there were so many interesting ideas. He replied with a long letter in which he expressed great enthusiasm for research.

That Levy should have played this role is tragic, for his own career also nearly foundered because his papers were not sufficiently rigorous for the mathematical extremists.

The second and deeper reason for Bachelier's career problems was the topic of his dissertation: ‘Mathematical theory of speculation’ – not of (philosophical) speculation on the nature of chance, rather of (money-grubbing) speculation on the ups and downs of the market for consolidated state bonds: ‘*la rente*’. The function $X(t)$ mentioned by Levy stood for the price of *la rente* at time t. Hence, the delicately understated comment by Henri Poincaré, who wrote the official report on this dissertation, that ‘the topic is somewhat remote from those our candidates are in the habit of treating’. One may wonder why Bachelier asked for the judgement of unwilling mathematicians (assigning a thesis subject was totally foreign to French professors of that period), but he had no choice: his lower degree was in mathematics and probability was taught by Poincaré.

Bachelier's tragedy was to be a man of the past and of the future but not of his present. He was a man of the past because gambling is the historical root of probability theory; he introduced the continuous-time gambling on *La Bourse*. He was a man of the future, both in mathematics (witness the above letter by Levy) and in economics. Unfortunately, no organized scientific community of his time was in a position to understand and welcome him. To gain acceptance for himself would have required political skills that he did not possess, and one wonders where he could have gained acceptance for his thoughts.

Poincaré's report on the 1908 dissertation deserves further excerpting:

> The manner in which the candidate obtains the law of Gauss is most original, and all the more interesting as the same reasoning might, with a few changes, be extended to the theory of errors. He develops this in a chapter which might at first seem strange, for he titles it ‘Radiation of Probability’. In effect, the author resorts to a comparison with the analytical theory of the propagation of heat. A little reflection shows that the analogy is real and the comparison legitimate. Fourier's reasoning is applicable almost without change to this problem, which is so different from that for which it had been created. It is regrettable that [the author] did not develop this part of his thesis further.

While Poincaré had seen that Bachelier had advanced to the threshold of a general theory of diffusion, he was notorious for lapses of memory. A few years later, he took an active part in discussions concerning Brownian diffusion, but had forgotten Bachelier.
Comments in a *Notice* Bachelier wrote in 1921 are worth summarizing:

> 1906: *Théorie des probabilités continues*. This theory has no relation whatsoever with the theory of geometric probability, whose scope is very limited. This is a science of another level of difficulty and generality than the calculus of probability. Conception, analysis, method, everything in it is new. 1913: *Probabilités cinématiques et dynamiques*. These applications of probability to mechanics are the author's own, absolutely. He took the original idea from no one; no work of the same kind has ever been performed. Conception, method, results, everything is new.

The hapless authors of academic *Notices* are not called upon to be modest, but Louis Bachelier had no reason for being modest. Does anyone know more about him?

See Also

- Wiener process

Selected works

1900. Théorie de la spéculation. *Annales de l'Ecole normale supérieure*, 3rd series 17, 21–86, trans. A. Boness. In *The Random Character of Stock Market Prices*, ed. P.H. Cootner. Cambridge, MA: MIT Press, 1967.

1901. Théorie mathématique des jeux. *Annales de l'Ecole normale supérieure*, 3rd series 18, 143–210.

1906. Théorie des probabilités continues. *Journal des mathématiques pures et appliquées*, 6th series 2, 259–327.

1910a. Les probabilités à plusieurs variables. *Annales de l'Ecole normale supérieure*, 3rd series 27, 340–60.

1910b. Mouvement d'un point ou d'un système soumis à l'action des forces dépendant du hasard. *Comptes rendus de l'Académie des sciences* 151, 852–5.

1912. *Calcul des probabilités*. Paris: Gauthier-Villars.

1913. Les probabilités cinématiques et dynamiques. *Annales de l'Ecole normale supérieure*, 3rd series 30, 77–119.

1924. *Le jeu, la chance et le hasard*. Paris: E. Flammarion.

1937. *Les lois des grands nombres du calcul des probabilités*. Paris: Gauthier-Villars.

1938. *La spéculation et le calcul des probabilités*. Paris: Gauthier-Villars.

1939. *Les nouvelles méthodes du calcul des probabilités*. Paris: Gauthier-Villars.

How to cite this article

Mandelbrot, Benott B. "Bachelier, Louis (1870–1946)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000003> doi:10.1057/9780230226203.0081

The New Palgrave Dictionary of Economics Online

Bagehot, Walter (1826 – 1877)

Asa Briggs
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bagehot, W.; Bank of England; bimetallism; British classical economics; expectations; free trade; Giffen, R.; Jevons, W. S.; mathematics and economics; Mill, J. S.; Ricardo, D.; socialism; statistics and economics; trade unions; Walras, L.

Article

Editor and literary critic as well as banker and economist, Bagehot was described in retrospect by Lord Bryce as ‘the most original mind of his generation’ (Buchan, 1959, p. 260). It is a difficult claim to sustain, certainly as far as his scattered economic writings are concerned. There was no doubt, however, about his intellectual versatility: there was an immediacy, a clarity and an irony – what he said of his friend Arthur Hugh Clough's poems, ‘a sort of truthful scepticism’ – about Bagehot's essays in different fields which make them still pre-eminently readable. Bagehot saw connections, too, between economics, politics, psychology, anthropology and the natural sciences – ‘mind and character’ – refusing to draw rigid boundaries between most of these subjects and ‘literary studies’, while recognizing in his later years that the frontiers of political economy needed to be more carefully marked. ‘Most original’ or not, he was, as the historian G.M. Young (1948) has observed, *Victoranum maxime*, if not *Victoranum maximus*: ‘he was in and of his age, and could have been of no other.’ He pre-dated academic specialization and professionalization, and he was never didactic in his approach.
His first writing on economics, a revealing if not a searching review of John Stuart Mill's *Principles of Political Economy*, appeared in 1848 before the sense of a Victorian age had taken shape. His last and most voluminous writing on the subject appeared posthumously in a volume of essays, the first on ‘the postulates of English political economy’, which his editor-friend Richard Holt Hutton entitled *Economic Studies* (1879). By then the economic confidence of the mid-Victorian years was over, and there were many signs both of economic and social strain, some of which Bagehot had predicted. It was in 1859, the *annus mirabilis* of mid-Victorian England, however, the year of Darwin's *Origin of Species*, Mill's *On Liberty* and Smile's *Self Help*, that Bagehot became editor of *The Economist*, a periodical founded by his father-in-law James Wilson, and it was through his lively editorship, which continued until his death,

that he was in regular touch with an interesting and influential, if limited, section of his contemporaries. ‘The politics of the paper’, he wrote simply, ‘must be viewed mainly with reference to the tastes of men of business.’

The mid-Victorian years constituted, in his own phrase, ‘a period singularly remarkable for its material progress, and almost marvellous in its banking development’. It was the latter aspect of the period which provided him with the theme of his best-known and brilliantly written book *Lombard Street*, which was begun in 1870 and appeared in 1873. It dealt, however, as it was bound to do, not only with the ‘marvellous development’, but with the ‘panics’ of 1857 and 1866 to which the Bank of England, the central institution in the system, had to respond. Indeed, the germ of *Lombard Street* was an article written in *The Economist* in 1857, 13 years after Peel's Bank Charter Act, and it was in 1866 that he took up the theme again.

Bagehot's conviction that the Bank of England neither fully understood nor fully lived up to its responsibilities was the product of years of experience which went back to his own early life between 1852 and 1859 as a country banker with Stuckey's at Langport, his birthplace, in the West of England, where his father also was a banker. The chapter on deposit banking reflects this. So, too, does his complaint that the directors of the Bank of England were ‘amateurs’, and his insistence that the ‘trained banking element’ needed to be augmented.

Lombard Street is a book with a distinctive purpose rather than an essay in applied economics; and, as Schumpeter has observed, ‘it does not contain anything that should have been new to any student of economics’. The main stress in it is on confidence as a necessary foundation of London's banking system. ‘Credit – the disposition of one man to trust another – is singularly varying. In England after a great calamity, everybody is suspicious of everybody; as soon as that calamity is forgotten everybody again confides in everybody.’ Bagehot underestimated the extent to which through joint stock banks’ cheques trade was expanding without increases in note issue and the extent to which the Bank of England itself was beginning to develop techniques of influencing interest rates. He also overestimated the extent to which in ‘rapidly growing districts’ of the country ‘almost any amount of money can be well employed’. In the last resort, too, his policy recommendations were deliberately restricted. He was disposed in principle to a ‘natural system’ in which each bank kept its own reserves of gold and legal tender, but in English circumstances he saw no more future in seeking to change the system fundamentally than in changing the political system. ‘I propose to retain this system because I am quite sure that it is of no manner of use proposing to alter it.’ With a characteristic glance across the Channel to France for a necessary comparison – things were done very differently there – he noted how the English system had ‘slowly grown up’ because it had ‘suited itself to the course of business’ and ‘forced itself on the habits of men’. It would not be altered, therefore, ‘because theorists disapprove of it, or because books are written against it’.

Bagehot had little use for ‘theorists’ and disdained the French for what he called their ‘morbid appetite for exhaustive and original theories’. He described political economy ‘as we have it in England’ as ‘the science of business’ and did not object to the fact that it was ‘insular’. Yet he talked of the ‘laws of wealth’ and believed that they had been arrived at in the same way as the ‘laws of motion’. Free trade was such a law. It was impossible, he argued, to write the history of ‘similar phenomena like those of Lombard Street’ without ‘a considerable accumulation of applicable doctrine’: to do so would be like ‘trying to explain the bursting of a boiler without knowing the theory of steam’, a not very helpful analogy since the invention of the steam engine preceded the discovery of the laws of thermodynamics.

Bagehot relied considerably on analogies. ‘Panics’, for example, were ‘a species of neuralgia’. The ‘unconscious “organization of capital”’ in the City of London, described by Bagehot as a ‘continental phrase’, depended on the entry into City business of a ‘dirty crowd of little men’; and this ‘rough and vulgar structure of English commerce’ was ‘the secret of its life’ because it contained ‘the propensity to variation’ which was ‘the principle of progress’ in the ‘social as in the animal kingdom’.

Such an approach to political economy was radically different from that of W.S. Jevons who, like Bagehot, had been educated at University College, London, or ‘M. Walras, of Lausanne’ who, according to Bagehot himself, had worked out ‘a mathematical theory’ of political economy ‘without communication and almost simultaneously’. There were however three defects, Bagehot maintained, in the British tradition of political economy, which started with Adam Smith but was sharpened and ‘mapped’ by David Ricardo. First, it was too culture-bound; for example, it took for granted the free circulation of labour, unknown in India. Second, its expositors did not always make it clear that they were dealing not with real men but with ‘imaginary’ ones. Abstract political economy did not focus on ‘the entire man as we know him in fact, but … a man answering to pure definition from which all impairing and conflicting elements have been fined away’. It was not concerned with ‘middle principles’. Third, considered as a body of knowledge, English political economy was ‘not a questionable thing of unlimited extent but a most certain and useful thing of limited extent’. It was certainly not ‘the highest study of the mind’. There were others ‘which are much higher’.

Bagehot did not push such criticism far. He had much to say about primitive and pre-commercial economies, but he put forward no theory of economic development. Nor, despite an interest in methodology, did he draw out the full implications of his own behaviourist (and in places institutionalist) approach to economics. Finally, he offered no agenda for political economists in the future. He noted, as others noted, that during the 1870s political economy lay ‘rather dead in the public mind. Not only does it not excite the same interest as it did formerly, but there is not exactly the same confidence in it.’ His own precoccupations in that decade were more practical than theoretical despite the writing of such essays as ‘The Postulates of English Political Economy’, which first appeared in article form in the *Fortnightly* in 1876. He never completed a new essay on Mill, and an essay on Malthus, whom he took along with Smith, Ricardo and Mill to be the founders of British political economy, revealed more interest in the man than in his thought. In the year when the ‘Postulates’ appeared, he successfully suggested to the Chancellor of the Exchequer the value to the Treasury of short-term securities resembling as much as possible commercial bills of exchange. The result was the Treasury Bill. The fact that the Chancellor was then a Conservative mattered little to the liberal-conservative Bagehot, who was described by his Liberal admirer W.E. Gladstone as a ‘sort of supplementary Chancellor of the Exchequer’.

Bagehot was as out of sympathy with the liberal radicals of the 1870s as he was with the bimetallists, and he had never shown any sympathy for socialist political economy. He saw the capitalist as ‘the motive power in modern production’ in the ‘great commerce’, the man who settled ‘what goods shall be made, and what not’. Nonetheless, he stated explicitly in several places that he had ‘no objection whatever to the aspiration of the workmen for more wages’, and he came to appreciate more willingly than Jevons the role of trade unions and collective bargaining. In his first review of Mill in 1848 he had stated that ‘the great problem for European and especially for English statesmen in the nineteenth century is how shall the [wage] rate be raised and how shall the lower orders be improved’. Some of the views he expressed on this subject – and on expectations – were not dissimilar to those of the

neoclassical Alfred Marshall. He did not use the term ‘classical’ himself in charting the evolution of British political economy.

Bagehot left no school of disciples. He was content to persuade his contemporaries. His sinuous prose style was supremely persuasive. So, too, was his skill in sifting and assessing inside economic intelligence. Yet while he devoted little attention to precise quantitative evidence in *Lombard Street* and, unlike Jevons, saw little point in developing economics in mathematical form, he was always interested in numbers as well as in words. One of his closest collaborators on the staff of *The Economist*, the statistician Robert Giffen, his first full-time assistant, paid tribute to ‘his knowledge and feeling of the “how much” in dealing with the complex workings of economic tendencies’. ‘He knew what tables could be made to say, and the value of simplicity in their construction.’ Bagehot always maintained, however, that while ‘theorists take a table of prices as facts settled by unalterable laws, a stockbroker will tell you such prices can be made’. Statistics were ‘useful’: they needed to be interpreted by ‘men of business’ who possessed the grasp of ‘probabilities’ and the ‘solid judgement’ which Bagehot most admired and which he sought to express. Indeed, business for him was ‘really a profession often requiring for its practice quite as much knowledge, and quite as much skill, as law and medicine’. Businessmen did not go to political economy: political economy, as in the case of Ricardo, came to them.

Selected works

All Bagehot's economic writings are collected in N. St. John Stevas, ed., *The Collected Works of Walter Bagehot*, vols 1–15 (1978–86), London: The Economist.

Bibliography

Buchan, A. 1959. *The Spare Chancellor. The Life of Walter Bagehot*. London: Chatto & Windus.

Giffen, R. 1880. Bagehot as an economist. *The Fortnightly*, April, 549–67.

Young, G.M. 1948. The greatest Victorian. In G.M. Young, *Today and Yesterday*, London: Rupert Hart-Davis.

How to cite this article

Briggs, Asa. "Bagehot, Walter (1826–1877)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000009> doi:10.1057/9780230226203.0082

The New Palgrave Dictionary of Economics Online

Bailey, Samuel (1791 – 1870)

R.M. Rauner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

absolute and exchangeable value; Cotterill, C.F.; De Quincey, T.; index numbers; labour supply; labour theory of value; Lowe, J.; McCulloch, J. R.; Macleod, H.D.; Malthus, T. R.; Mill, J.; Mill, J. S.; relative value; Ricardo, D.; Scrope, P.; Seligman, E. R. A.; Torrens, R.

Article

Samuel Bailey was born in Sheffield, England, one of 11 children. His father was a cutler and merchant of substance. Samuel also became a merchant and banker. Throughout his life, he served on the Sheffield Town Trust (a quasi-governmental agency) and was twice a candidate for Parliament in the Reform elections of 1832 and 1835. Writing widely on banking, politics and philosophy, he lived his entire life in Sheffield, unmarried, and died there in 1870.

Bailey published his principal economic work, *A Critical Dissertation*…, in 1825, a time when Ricardian theory was nearing its peak of popularity and acceptance. The *Westminster Review* (1826) thought the *Critical Dissertation* inconsequential, and J.R. McCulloch (1845) later claimed that it had not shaken the foundations of Ricardo's labour theory of value. Robert Torrens, however, praised Bailey's book in 1831 at the London Political Economy Club, and John Stuart Mill brought it before his bi-weekly reading group. This attention, nevertheless, did not keep Bailey on front stage, and he had to be rediscovered later by E.R.A. Seligman (1903); the London School of Economics republished the *Critical Dissertation* in 1931. Schumpeter (1954) judged Bailey's tract to be a ‘masterpiece of criticism’ and to lie near the ‘front rank in the history of scientific economics’. R.M. Rauner (1961) re-examined Bailey's work from a larger perspective.

The centrepiece of Bailey's argument was his definition of value as ultimately ‘esteem’ or a ‘mental affection’. The ‘specific feeling of value’, however, arose only when items were subject to preference or exchange. This defined value as relative, not something intrinsic like labour in Ricardo's theory. Value is the amount of one commodity exchanged for another; it is measured in terms of a third commodity with which the two exchange if they are not directly bartered. From this position Bailey attacked Ricardo's

postulate that labour effort defined value. He showed that, despite Ricardo's claim to the contrary, constancy of labour used in production could not assure constancy in exchange value – unless value were defined differently. This, of course, is what Ricardo had done in shifting from exchange to ‘real’ or ‘absolute’ value.

Ricardo's conception of value as an absolute and his endless search for a standard of invariable value opened him to Bailey's stricture that constancy of value meant constancy in exchange ratios. Evidence and observation showed that exchange values rarely stayed constant. To the Ricardians, however, constancy of value meant constancy of labour cost of production; this, they believed, was necessary in the determination of whether individual economic welfare had changed over time. Bailey objected that exchange of commodities cannot take place between two different time periods. Exchanges occur at different times and these exchanges can be compared. But such comparisons are the only way economic welfare in different times or places can be assessed. In a later tract (1844), Bailey used this same argument, making the point that interperiod contracts could be fixed only in terms of quantities, not constancy of values. This enabled him to oppose the index number proposals (then called ‘tabular standards’) of Joseph Lowe and Poulett Scrope. Such standards could not assure constancy of quantities exchanged in different times, a criticism of index numbers that is still valid today.

Using relative value as his anchor, Bailey then demonstrated that Ricardo's theory of wages was faulty. He insisted that labour value – wages – was definitionally the same as all other value, namely, what a unit of labour exchanged for. Ricardo's theorem, that wages and profits varied inversely, was wrong since it implied that wages could be high (i.e. taking a large proportionate share of production) while labour value was low, wages exchanged for little and workers were near starvation.

The relative value concept applied to wages allowed Bailey an easy application of the principles of rent to labour. Just as with land, different values for labour were caused by the monopoly characteristics of labour supply, as well as by differential productivity due to varying labour skill or dexterity. This contrasted sharply with Ricardian–Malthusian subsistence wages. Unfortunately, Bailey did not use the same reasoning against capital and merely denoted profits as the gain over capital employed.

The *Critical Dissertation* prompted some serious attempts to clear up the loose ends in Ricardo, most notably by McCulloch (1845); by the anonymous *Westminster Review* article (1826), probably written by James Mill (1826); and by Thomas De Quincey (1844). But Ricardo's *system* held fast. Malthus (1827) devoted the largest part of his work on definitions to Bailey, mainly quarrelling over the purely relative value notion. He reaffirmed the importance of a constant, unvarying measure of value, defined as the quantity of labour commanded by commodities in exchange. Samuel Read (1829) drew on Bailey's destruction of the Mill–McCulloch theory that time used in production is congealed labour, but he did not follow Bailey on the relativity of value or the measure of value. C.F. Cotterill (1831) and H. D. Macleod (1863; 1866) both praised Bailey's work and used his treatment of the nature and measure of value in their own studies.

From a larger perspective, by stressing relative value exclusively, Bailey pulled economic analysis back from the Smith–Ricardo stream that sought a principal cause of value to explain the production and distribution of material wealth among the labouring, rentier and capitalist classes. In Bailey's argument relative values – prices – vary for all kinds of reasons affecting demand (‘esteem’) and supply (production under constant or increasing cost, supply-limiting) conditions. Hence, his view involves no notion of long-run growth, tendencies toward equilibrium, stationary states or other systemic visions.

Everything is relative; individual economic welfare is expressed period-by-period solely in terms of relative values.

Bailey's is an incomplete treatment if one demands that value theory be integral with the determination of social, institutional and economic forces in an inter-dependent production system. On the other hand, Bailey's work freed analysis from the need to link production and distribution to socioeconomic class relationships. It pointed instead towards relationships between individual needs and perceptions, and the material goods that can satisfy them.

Selected works

1821. *Essays on the Formation and Publication of Opinions and Other Subjects*. London.

1823. *Questions on Political Economy, Politics, Metaphysics, Polite Literature and Other Branches of Knowledge*. London.

1825. *A Critical Dissertation on the Nature, Measures and Causes of Value; chiefly in reference to the writings of Mr Ricardo and his followers*. London.

1826. *A Letter to a Political Economist; occasioned by an article in the Westminster Review on the subject of Value*. London.

1830. *A Discussion of Parliamentary Reform*. London.

1835. *The Rationale of Political Representation*. London.

1837. *Money and Its Vicissitudes in Value; as they affect National Industry and Pecuniary Contracts; with a Postscript on Joint-Stock Banks*. London.

Bibliography

[Anon.] 1826. Letter to a Political Economist. *Westminister Review*, January.

Cotterill, C.F. 1831. *An Examination of the Doctrines of Value as set forth by Adam Smith, Ricardo, McCulloch, the Author of 'A Critical Dissertation', etc., Torrens, Malthus, Say, etc., being a reply to those distinguished authors*. London.

De Quincey, T. 1844. *The Logic of Political Economy*. Edinburgh.

McCulloch, J.R. 1825. *Principles of Political Economy*. London.

McCulloch, J.R. 1845. *The Literature of Political Economy*. London.

Macleod, H.D. 1863. *A Dictionary of Political Economy*. London.

Macleod, H.D. 1866. *The Theory and Practice of Banking*. 2nd edn. London.

Malthus, T.R. 1827. *Definitions in Political Economy*. London.

Mill, J. 1826. *Elements of Political Economy*. 3rd edn. London.

Rauner, R.M. 1961. *Samuel Bailey and the Classical Theory of Value*. London: London School of Economics and Political Science, G. Bell & Sons.

Read, S. 1829. *An Inquiry into the Natural Grounds of Right to Vendible Property or Wealth*. Edinburgh.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Seligman, E.R.A. 1903. On some neglected British economists. *Economic Journal* 13, 335–63, 511–35.

How to cite this article

Rauner, R. M. "Bailey, Samuel (1791–1870)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000010> doi:10.1057/9780230226203.0083

The New Palgrave Dictionary of Economics Online

Bain, Joe Staten (1912 – 1991)

William G. Shepherd
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bain, J. S.; barriers to entry; concentration; economies of scale; industrial organization; limit pricing; market power; market share; oligopoly group; profitability

Article

Joe S. Bain was born in Spokane, Washington, on 4 July 1912. After graduating from the University of California at Los Angeles in 1935 and gaining the doctorate from Harvard in 1940 (under Joseph Schumpeter), he spent his entire career at the University of California at Berkeley, retiring in 1975. He was appointed Distinguished Fellow of the American Economic Association in 1982.

A prolific and seminal writer, Bain helped to shape the field of industrial organization in its modern form, with special attention to market structure. Bain's analysis focused on the oligopoly group within an industry, and on barriers to new competition. He also worked on natural resource development by public enterprise, concentrating on the oil industry.

Bain's empirical work on economies of scale, entry barriers, and limit pricing broke new ground. He developed the field's intellectual format, in which technical factors may determine structure, and structure then influences behaviour and performance. Some of these concepts were already current as early as 1900. During 1925–40, as the field took shape, attention shifted to the industry and the oligopoly group within it.

In the 1930s, Bain entered a formative field which was rich in possibilities for giving new rigour to older concepts, for developing new ones, and for shaping the framework. That has been his main role and contribution. Though he did not create concepts, nor indeed the framework, he selected from among them and carried their scientific analysis further than anyone else.

The analysis grew after 1940 in a series of articles and chapters, culminating in *Barriers to New Competition* in 1956 and *Industrial Organization* in 1959. His analysis was verbal and graphical rather than mathematical. In Bain's analysis of the conditions of entry, the barriers have three possible economic sources; absolute cost advantages, product differentiation and size. Barriers then permit ‘limit pricing’ by a firm or firms which consciously apply their strategy towards entry.

Bain drew the main conclusions, and he noted the difficulties of empirical tests. The definition of barriers as a single, general phenomenon posed special problems, which are still unsolved. Since 1960, over seven new barrier ‘sources’ have been proposed, and the concept of barriers has tended to acquire just that ad hoc character which Bain frequently reproved in others’ theories.

Measurement has also proven to be difficult. It requires a merging of disparate objective and subjective data about the barriers’ causes. Whether these sources of barriers are additive, multiplicative or merely parallel was also left unclear by Bain (and all others).

Bain's measurement of scale economics was pioneering. Earlier studies had suffered from data problems and from a mingling of technical and pecuniary elements. Bain centred unerringly on technical economies. Thereby he gave the first solid normative basis for evaluating excess concentration.

By estimating ‘best practice’ conditions for scale for new capacity, Bain neatly avoided the normative–positive confusion which infects cross-section studies of past costs and survivor tests of emerging sizes. His ‘engineering’ estimates supply a normative basis for appraising how much concentration is socially ‘necessary’.

Profitability was also analysed closely by Bain. He tried nearly every available method to factor out the concentration–profitability relationship. In a 1949 article (later extended in *Barriers*), Bain put the study of profitability on a firm scientific and normative basis. His findings of a step function, with a break at 70 per cent for eight-firm concentration, has tended to be replaced in recent research by a continuously sloping concentration–profitability relationship. Still, Bain set the basis for all good later research on the subject.

Bain's architectural choices in using and emphasizing individual elements were distinctive. Three features stand out – the triad, the industry basis, and the stress on the oligopoly group behind an entry barrier. (1) Bain developed the three-tier format of structure, behaviour and performance with what may be called a ‘soft structuralist’ emphasis. Bain used it as a broad set of concepts, by which the whole subject (theory, tests, policy lessons) is organized, not as just a format for individual cases. (2) Bain used the *industry* as the basic unit behaviour. It was a choice that shaped the images and methodology in distinctive ways. (3) The *oligopoly group*, setting limit price strategy behind an entry barrier, came to be the most distinctive part of Bain's analysis. As of 1949–50, Bain regarded concentration as the key determinant of market power and profitability.

By 1951 he appeared to regard barriers as the decisive element, which could be both necessary and sufficient to govern profitability. Yet Bain later suggested frequently that barriers would be highly correlated with the degree of concentration. In fact, all of the sources of barriers are also sources of high market shares and concentration. Do barriers shape the dominant firm's share, or do they operate jointly? Any eventual resolution of barriers’ role will probably assign barriers at least a significant role, thanks to Bain's stress on them. He put the concepts and relationships in testable form, and he began the testing of them. To a large extent he rescued the subject from a preoccupation with oligopoly interactions and games, and he gave it a strong framework.

Yet Bain's most durable contribution lies deeper, in the methods and research standards of the field. By 1960, he had helped to give it structure, precision, and high standards of research quality. He selected the main concepts and relationships, gave them extended analysis, tested them, and drew policy lessons. The individual parts were related within a framework of causation and performance.

His more specific methods and results have also continued to be valid because they met these standards. Beyond the individual concepts and tests is the fact that they fit together in a system, and that this

system was carefully developed and tested. That is the way to scientific permanence.

Selected works

1944, 1945, 1947. *The Economics of the Pacific Coast Petroleum Industry*. 3 vols. Berkeley: University of California Press.

1948. *Pricing, Distribution and Employment: Economics of an Enterprise System*. New York: Henry Holt; Italian and Spanish translations.

1956. *Barriers to New Competition: Their Character and Consequences in Manufacturing Industries*. Cambridge, MA: Harvard University Press.

1959. *Industrial Organization*. New York: Wiley; revised edn, 1968.

1966a. *International Differences in Industrial Structure: Eight Nations in the 1950s*. New Haven: Yale University Press.

1966b. (With R.E. Caves and J. Margolis.) *Northern California's Water Industry: The Comparative Efficiency of Public Enterprise in Developing a Scarce Natural Resource*. Baltimore: Johns Hopkins Press.

1972. *Essays on Price Theory and Industrial Organization*. Boston: Little, Brown & Co.

How to cite this article

Shepherd, William G. "Bain, Joe Staten (1912–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000011> doi:10.1057/9780230226203.0084

The New Palgrave Dictionary of Economics Online

Bairoch, Paul (1930 – 1999)

Elise S. Brezis

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agricultural revolution; Bairoich, P.; city and economic development; cliometrics; colonialism; diffusion of technology; economic development; foreign aid; free trade; inequality (global); protection; technological progress; trade and economic growth; urbanization

Article

Paul Bairoch was born in Antwerp in 1930. He was the son of a Jewish family that emigrated from Poland to Belgium in the 1920s, and that later went into exile in a small village in the Gers, France, during the Second World War. After the war, Bairoch moved to Brussels, later spent a short period in Israel, and upon his return to Belgium began to study economic history. While a research fellow at the University of Brussels, Bairoch developed statistical time series on the national statistics of Belgium, worked on his doctorate, and in 1963, presented his thesis, ‘The Starting Process of Economic Growth’. He then went on to teach in a number of universities and even worked at General Agreement on Tariffs and Trade (GATT) for a time. From 1972 onwards, Bairoch was a member of the faculty at the University of Geneva, where he was director of the Center of International Economic History until his death in 1999.

A trait common to all Bairoch's research in economic history from his thesis onwards was that he based his opinions on data, and, when the data did not exist, he found a way to collect or construct new data. Bairoch can be seen as a pioneer of cliometrics, and believed that economic history cannot survive without data and statistical information. David Landes (1998, p. xiii) even gave Bairoch the nickname ‘collector and calculator of the numbers of growth and productivity’. Another characteristic typical of Bairoch's research is that he was not afraid to be nonconformist and present views that ran against the mainstream.

Bairoch worked in three main subjects: economic development and growth, urban studies and international trade.

Population, cities, and urban research

Bairoch was interested in the relationship between urbanization and economic development, and examined urban evolution from the Neolithic period to 1900. He developed series on sizes of cities from ad 800 to 1850.
Bairoch's main achievement in this field was showing that there was a typical pattern of urbanization: traditional societies reached their maximum urban population rapidly, levelling off at somewhere between 8 and 15 per cent (Europe reached this level around 1300), and maintained this proportion until the onset of industrialization, when the urban population then surged. He also observed that for non-developed countries urbanization has negative consequences for agricultural development.

Development, industrialization, and inequality

One of the main topics of Bairoch's research was the dynamics of development and the inequality between developed and developing countries. In his last book, *Victoires et déboires* (1997), a formidable synthesis of the economic and social history of the world, Bairoch tried to explain the pre-eminence of the West, and the setbacks (*déboires*) suffered by the Third World.
Regarding the mechanism of development of the West, Bairoch insisted on the necessity of an agricultural revolution, and also on the importance of institutions. He had also a strong interest in the development of technological progress in the 19th century, and stressed the differences between it and the diffusion of the science-based technology of the 20th century.
Bairoch also analysed at length the reasons for the backwardness of the Third World, and through the use of comparative statistics his analysis includes a comparison between its present economic progress and that of developed countries at the times of their take-offs. Bairoch's conclusions were that the absence of an agricultural revolution and failure to reduce fertility rates were among the most binding facts impeding development. He was therefore pessimistic about the prospects for development of the lagging countries, especially those in Africa.
Regarding inequality, Bairoch stressed that before the Industrial Revolution no appreciable difference in per capita income separated western Europe from the rest of the world, while the gap between the developed and the developing world increased thereafter. Moreover, regarding the effect of colonialism, Bairoch stressed that colonialism was not only largely unprofitable for the West but also harmed the Third World. Bairoch was a proponent of foreign aid to reduce inequalities.

International trade

Probably Bairoch's best-known work is *Economics and World History: Myths and Paradoxes* (1993), in which he sets the record straight on 20 commonly held myths about economic history, among them that free trade has historically led to periods of economic growth; a myth associated with those who ‘could be described as a conservative group that romanticizes the 19th century and makes free trade almost into a sacred doctrine’ (1993, p. xiv).
Bairoch claimed that the idea that free trade was the rule during the 19th century is a myth based on insufficient knowledge and misguided interpretations of the economic history of the United States, Europe, and the Third World, since protection is the rule and free trade the exception. Moreover,

Bairoch expressed doubts that free trade leads to economic growth. His thesis was that during development countries use protectionist policies, which they dismantle once they industrialize. He showed that Britain protected its home market until British firms in the main sectors dominated the market, and only later on did Britain advocate free trade.
I cannot conclude without mentioning Bairoch's personality: he combined the best of open-minded curiosity and a powerful intellect with warmth, humanity and overwhelming kindness to all who knew him.

See Also

- development economics
- economic history
- international trade theory

Selected works

1975. *The Economic Development of the Third World since 1900*. London: Methuen.

1988. *Cities and Economic Development from the Dawn of History to the Present*. Chicago: University of Chicago Press.

1993. *Economics and World History: Myths and Paradoxes*. Chicago: University of Chicago Press.

1997. *Victoires et déboires: histoire economique et sociale du monde du 16ème siècle à nos jours*, 3 vols. Paris: Gallimard.

How to cite this article

Brezis, Elise S. "Bairoch, Paul (1930–1999)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000335> doi:10.1057/9780230226203.0085

The New Palgrave Dictionary of Economics Online

balanced growth

Jonathan Temple

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Balanced growth’ has at least two different meanings in economics. In macroeconomics, balanced growth occurs when output and the capital stock grow at the same rate. This growth path can rationalize the long-run stability of real interest rates, but its existence requires strong assumptions. In development economics, balanced growth refers to the simultaneous, coordinated expansion of several sectors. The usual arguments for this development strategy rely on scale economies, so that the productivity and profitability of individual firms may depend on market size. The article reviews the balanced growth debate and the extent to which it has influenced development policies.

Keywords

balanced growth; Big Push; coordination failures; development strategies; economic geography; economies of scale; Hirschman, A.; increasing returns; industrialization; Krugman, P.; multiple equilibria; Nurkse, R.; pecuniary external economies; Rosenstein-Rodan, P.; Scitovsky, T.; Solow, R.,; Swan, T.; unbalanced growth

Article

In macroeconomics, ‘balanced growth’ refers to classes of equilibrium growth paths, while in development economics the term refers to a particular development strategy.
These two uses of the term are clearly distinct, and each is discussed in turn.
The concept of a balanced growth path is a central element of macroeconomics. It refers to an equilibrium in which major aggregates, usually but not exclusively output and the capital stock, grow at the same rate over time, and the real interest rate is constant. Most textbook growth models are constructed in a way that delivers this outcome. This is motivated partly by theoretical convenience but also by historical observation. The conventional wisdom is that real interest rates and the capital-output ratio are surprisingly stable over long spans of time, at least in developed countries.

Balanced growth is not an inevitable property of growth models. It was not until the publication of classic papers by Solow (1956) and Swan (1956) that economists saw how a balanced growth path might arise from relatively appealing assumptions. The key insight is that a stable equilibrium path requires the possibility of substitution between capital and labour. The Solow–Swan model has subsequently underpinned much empirical work on economic growth, and has also influenced short-run macroeconomics.

The existence of a balanced growth path requires strong assumptions. The usual derivation assumes that aggregate output can be written as a function of the total inputs of capital and labour, with diminishing returns to each input and constant returns to scale overall. In addition to the conditions needed for aggregation, either the production function should be Cobb–Douglas, or technical progress should be restricted to the labour-augmenting type. In other words, when technology advances, it should be ‘as if’ the economy had more labour than before, and not ‘as if’ it had more capital.

Because these assumptions are strong, any use of balanced growth to rationalize the data tends to create new puzzles. For example, why should technical progress be exclusively labour-augmenting, as stability of real interest rates would require? Acemoglu (2003) has examined this question using an incentives-based model of technical change, but in general balanced growth seems a less than inevitable outcome of a real-world growth process. The picture is even more complicated when there are multiple sectors, whether differentiated as capital and consumer goods, or as different types of final goods. As might be expected, where multiple sectors are present, the conditions needed for balanced growth become even stricter. Greenwood, Hercowitz and Krusell (1997) and Kongsamut, Rebelo and Xie (2001) are two useful references on multi-sector growth models.

None of this is to deny that balanced growth is a useful concept. The idea plays an important role in teaching and research in macroeconomics because of its simplicity and explanatory power. As with all organizing frameworks, however, it is sensible to be aware of its limitations and the possibilities that lie outside it.

In macroeconomics, balanced growth is usually associated with constant returns to scale. For most development economists, the term is more strongly associated with increasing returns and a debate that began with Rosenstein-Rodan (1943). He argued that the post-war industrialization of eastern and south-eastern Europe would require coordinated investments across several industries. The idea is that expansion of different sectors is complementary, because an increase in the output of one sector increases the size of the market for others. A sector that expands on its own may make a loss but, if many sectors expand at once, they can each make a profit. This tends to imply the need for coordinated expansion, or a ‘Big Push’, and potentially justifies a role for state intervention or development planning. Another influential contribution by Nurkse (1953) made similar points, giving more emphasis to the links between market size and the incentives to accumulate capital.

In Rosenstein-Rodan's paper the argument is set out informally, and with many digressions. But the central point will have a familiar ring to students of modern game theory and the literature on coordination failures. Essentially, Rosenstein-Rodan was setting out assumptions that might give rise to multiple equilibria in levels of development. Papers by Fleming (1955) and Scitovsky (1954) further clarified some of the necessary assumptions. Fleming emphasized the importance of Rosenstein-Rodan's assumption that the industrializing sectors can draw on labour from other sectors without forcing up wages. Scitovsky noted that the proponents of balanced growth appeared to see externalities everywhere,

but under perfect competition, external effects that are mediated through markets (‘pecuniary external economies’) do not preclude Pareto efficiency. This result hints at the importance of scale economies to the balanced growth hypothesis, since then market size can influence unit costs, and Scitovsky's logic no longer applies.

The key ideas of the balanced growth hypothesis were formalized in a much-admired paper by Murphy, Shleifer and Vishny (1989). In their multi-sector model, firms in each sector use constant returns-to-scale technologies, but one firm in each sector also has access to an increasing returns-to-scale technology. This technology will be profitable to operate only given a sufficiently large market. The structure of the model, with a competitive fringe of small-scale producers, ensures that wages are independent of labour demand in the industrializing sectors. The model yields multiple equilibria that can be Pareto-ranked.

The assumptions needed for multiplicity are more complicated than earlier authors believed, however. For example, increasing returns and an elastic supply of labour are not sufficient in themselves to generate multiple equilibria. Consider an equilibrium in which no sectors have industrialized (meaning that none is using the increasing returns-to-scale technique). If a single firm then adopts the modern technique and makes a loss, this will reduce rather than increase the size of the market for other sectors, so the necessary complementarity is absent. For multiple equilibria to arise, the industrializing firm must somehow raise the size of the market for other sectors, even though it makes a loss when acting alone. In one of the models considered by Murphy, Shleifer and Vishny (1989), this is achieved by an extra assumption, namely, that industrializing firms must pay higher wages than other firms.

Although the balanced growth hypothesis has been widely discussed, it has a number of limitations. The ideas are difficult to test empirically. From a purely theoretical point of view, the argument does not generalize straightforwardly to open economies. If firms can sell their output abroad, the role of domestic market size appears much less important. The balanced growth hypothesis then requires a more complex story, perhaps one in which firms are especially reliant on domestic markets in the early stages of their development.

The ideas have also been criticized on other grounds. The most prominent sceptic was Hirschman (1958), who argued that simultaneous, coordinated investment asked too much of developing countries. He regarded growth as a necessarily unbalanced dynamic process, in which successive disequilibria create the conditions for development in other sectors. Unbalanced growth could occur either through forward and backward linkages to downstream and upstream industries or by drawing out latent capacities needed for growth, such as the application of entrepreneurial skills.

Importantly, this process is seen as too complex and unpredictable to lend itself readily to a government-inspired ‘Big Push’, partly because governments may lack the relevant information, and partly because simultaneous investment would place too many demands on limited organizational resources. Hirschman (1958, pp. 53–4) summarized his objections by saying: ‘if a country were ready to apply the doctrine of balanced growth, then it would not be underdeveloped in the first place’.

But his preferred vision has echoes of the balanced growth doctrine in its appeal to complementarities and increasing returns; Krugman (1995) discusses this point in more detail. Arguably it is not so much the assumptions that differ, but the view of equilibrium selection. One interpretation of Hirschman's critique is that the multiplicity of equilibria is illusory, because the earlier authors had missed out relevant state variables.

In practice, balanced growth ideas have had less influence on development strategies than a more general commitment to state-led industrialization and import substitution. A perceived need for balanced growth may have motivated some attempts at indicative planning, but state interventions have usually tried to focus on particular sectors rather than attempting the more ambitious task of simultaneous expansion across many industries. The reasons for this are likely to be complex, including uncertainty over which sectors should be encouraged to expand, and the lack of obvious ways to coordinate this without direct state control. In the academic literature, the difficulty of testing the main ideas has been another factor limiting their influence.

For reasons like these, the balanced growth hypothesis is currently at the margins of development thinking and policy advice. The ideas are still interesting, however, and their neglect is partly due to the accidents of intellectual history. Formalizing Rosenstein-Rodan's original insights proved a difficult task. The reasons for this are discussed in Krugman (1995) as part of an illuminating account of the balanced growth debate and the role of formal models. He shows the continuing relevance of the main ideas to economic geography and regional science, and his book can be highly recommended to anyone interested in balanced growth, or the methods of modern economics more generally. Another useful reference is the special issue of the *Journal of Development Economics* on increasing returns and economic development (April 1996).

See Also

- development economics
- growth models, multisector
- linkages
- new economic geography
- poverty traps
- regional development, geography of
- structural change

Bibliography

Acemoglu, D. 2003. Labor- and capital-augmenting technical change. *Journal of the European Economic Association* 11, 1–37.

Fleming, M. 1955. External economies and the doctrine of balanced growth. *Economic Journal* 65, 241–56.

Greenwood, J., Hercowitz, Z. and Krusell, P. 1997. Long-run implications of investment-specific technological change. *American Economic Review* 87, 342–62.

Hirschman, A. 1958. *The Strategy of Economic Development*. New Haven: Yale University Press.

Kongsamut, P., Rebelo, S. and Xie, D. 2001. Beyond balanced growth. *Review of Economic Studies* 68, 869–82.

Krugman, P. 1995. *Development, Geography, and Economic Theory*. Cambridge, MA and London: MIT Press.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Industrialization and the Big Push. *Journal of Political Economy* 97, 1003–26.

Nurkse, R. 1953. *Problems of Capital Formation in Underdeveloped Countries*. Oxford: Basil Blackwell.

Rosenstein-Rodan, P. 1943. Problems of industrialisation of eastern and south-eastern Europe. *Economic Journal* 53, 202–11.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62, 143–51.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Swan, T. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

How to cite this article

Temple, Jonathan. "balanced growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000023> doi:10.1057/9780230226203.0086

The New Palgrave Dictionary of Economics Online

Balassa, B é la (1928 – 1991)

Jaime de Melo and Carl F. Christ

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Balassa, B.; Bhagwati, J.; Chenery, H.; Christ, C.; effective rate of protection; exchange rate regimes; import substitution; Institute for International Economics; international trade; Fischer, S.; Johns Hopkins University; Krueger, A.; Melo, J. de; Noland, M.; protection; purchasing power parity; revealed comparative advantage; Summers, L.; structural adjustment lending; tariffs; World Bank

Article

Béla Balassa, holding degrees in law and economics, left his native Hungary when the Soviet tanks put down the 1956 revolution. In 1959 he received his Ph.D. in economics from Yale. From 1966 until his death in 1991 he was professor of economics at Johns Hopkins and a consultant to the World Bank. Influenced by events in his youth, Béla held a deep lifelong belief in political and economic freedom. At the World Bank, Béla was very active as Research Advisor to the Vice-President for Research, first to Hollis Chenery, then to his successors, Anne Krueger, Stanley Fischer and Larry Summers. He held this position until his death, and those of us who were then at the Bank will remember him as the Bank's most influential economic advisor during his 25 years involvement at the institution. His commitment to economic policy was extended by his involvement in his later years at the Institute for International Economics, where he wrote on trade policy issues of developed countries, notably on Japan (Balassa and Noland, 1988).

Béla was among the most prolific international trade economists of his generation, contributing several books that are still widely cited. Early in his career he made several lasting contributions, among which was his famous paper on purchasing power parity (1964) in which he used a Ricardian model to show that a country's real exchange rate would appreciate as its productivity gap narrowed. Béla also made lasting contributions to the theory of economic integration (1962) and to empirical methods, proposing a measure of ‘revealed comparative advantage’ and ways to measure rates of effective protection (1965). As research advisor of the Development Research Center at the World Bank, Béla fulfilled many roles during the three days a week he spent there. Whereas most members at the centre would devote most of their time to research, in addition to his highly productive research activities Béla participated very

actively in the Bank's policy dialogue, commenting on the vast majority of country reports, and invariably on all those that contained advice on trade policies. In those days trade policy was a major issue in virtually all countries. Then, import-substitution policies supported by highly restrictive trade regimes were the rule. With a handful of trade economists, including Jagdish Bhagwati and Anne Krueger, Béla would tirelessly recommend a simplification of the trade regime, moderate protection of industrial activities supported by uniform tariffs, a removal of quantitative restrictions, and a unification of the then prevailing multiple exchange rate regimes.

Béla's advice on trade policy was supported by his research carried out under the Bank's auspices. He directed and edited an influential book that examined the trade regimes of several countries in Latin America and East Asia, documenting systematically the patterns of effective rates of protection in these countries (Balassa and Associates, 1971).

Béla's research output was not only prolific but also timely. His ability to be the first to deliver relevant research on the policy issue of the day was uncanny. In the late 1970s, when developing countries were hit by oil, commodity and interest rate shocks, Béla was the first to implement a useful decomposition formula to assess the extent of purchasing power loss. Later, when the Bank launched structural adjustment lending activities and wanted to assess performance of countries having received adjustment loans, Béla again delivered the first assessment of adjustment lending.

Béla's work capacity was legendary. Despite his influential research and his sage and realistic policy advising at the World Bank, which left him only two days a week for Johns Hopkins, his contribution to teaching, thesis supervision and academic governance at Hopkins was enormous. He taught most of the courses in international and development economics. He supervised more students than almost anyone else, and he responded to their papers and thesis drafts almost instantly with demanding but constructive comments. For ten years he was an elected and re-elected member of the faculty governing council. As chair of the faculty budget committee, he persuaded the university to reverse the decline that had been permitted to occur in the real value of its tuition charge, its faculty compensation levels and its academic expenditures.

Besides all this, Béla was an informed lover of art, opera, French literature and food (his guide to Paris restaurants was prized), and he always made time for his friends and for his family.

See Also

- purchasing power parity

Selected works

1962. *The Theory of Economic Integration*. Homewood, IL: Richard D. Irwin Inc.

1964. The purchasing power parity doctrine: a reappraisal. *Journal of Political Economy* 72, 584–96.

1965. Tariff protection in industrial countries: an evaluation. *Journal of Political Economy* 73, 573–94.

1971. (With Associates.) *The Structure of Protection in Developing Countries*. Baltimore: Johns

Hopkins University Press.

1988. (With M. Noland.) *Japan and the World Economy*. Washington, DC: Institute for International Economics.

How to cite this article

de Melo, Jaime and Carl F. Christ. "Balassa, Béla (1928–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000327>
doi:10.1057/9780230226203.0087

The New Palgrave Dictionary of Economics Online

bandit problems

Dirk Bergemann and Juuso Välimäki
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The multi-armed bandit problem is a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. This classic problem has received much attention in economics as it concisely models the trade-off between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff).

Keywords

asset pricing; bandit problems; branching bandit problem; continuous-time models; corporate finance; descending price auction; experimentation; free-rider problem; Gittins index th; informational efficiency; learning; liquidity; Markov equilibria; matching markets; moral hazard; noise trader; probability distribution; product differentiation; regime switch

Article

The multi-armed bandit problem, originally described by Robbins (1952), is a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. In the multi-arm bandit problem, the gambler has to decide which arm of *K* different slot machines to play in a sequence of trials so as to maximize his reward. This classical problem has received much attention because of the simple model it provides of the trade-off between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff). Each choice of an arm results in an immediate random payoff, but the process determining these payoffs evolves during the play of the bandit. The distinguishing feature of bandit problems is that the distribution of returns from one arm only changes when that arm is chosen. Hence the rewards from an arm do not depend on the rewards obtained from other arms. This feature also implies that the distributions of returns do not depend explicitly on calendar time.
The bandit framework found early applications in the area of clinical trials where different treatments

need to be experimented with while minimizing patient losses and in adaptive routing efforts for minimizing delays in a network. In economics, experimental consumption is a leading example of an intertemporal allocation problem where the trade-off between current payoff and value of information plays a key role.

Basic model

It is easiest to formulate the bandit problem as an infinite horizon Markov decision problem in discrete time with time index $t = 0, 1, \ldots$ At each *t*, the decision maker chooses amongst *K* arms and we denote this choice by $a_t \in \{1, \ldots, K\}$. If $a_t = k$, a random payoff x_t^k is realized and we denote the associated random variable by X_t^k. The state variable of the Markovian decision problem is given by s_t. We can then write the distribution of x_t^k as $F^k(\cdot\,; s_t)$. The state transition function ϕ depends on the choice of the arm and the realized payoff:

$$s_{t+1} = \varphi(x_t^k, s_t)$$

Let S_t denote the set of all possible states in period *t*. A feasible Markov policy $a = \{a_t\}_{t=0}^{\infty}$ selects an available alternative for each conceivable state s_t, that is,

$$a_t : S_t \to \{1, \ldots, K\}$$

The following two assumptions must be met for the problem to qualify as a bandit problem.

1. 1. Payoffs are evaluated according to the discounted expected payoff criterion where the discount factor δ satisfies $0 \leq \delta < 1$.
2. 2. The payoff from each *k* depends only on outcomes of periods with $a_t = k$. In other words, we can decompose the state variable s_t into *K* components $(s_t^1, \ldots, s_t^K)$ such that for all *k*:

$$s_{t+1}^k = s_t^k \quad \text{if} \quad a_t \neq k,$$
$$s_{t+1}^k = \varphi(s_t^k, x_t) \quad \text{if} \quad a_t = k,$$

and

$$F^k(\cdot, s_t) = F^k(\cdot; s_t^k).$$

Notice that when the second assumption holds, the alternatives must be statistically independent.
It is easy to see that many situations of economic interest are special cases of the above formulation. First, it could be that $F^k(\cdot; \theta^k)$ is a fixed distribution with an unknown parameter θ^k. The state variable is then the posterior probability distribution on θ^k. Alternatively, $F^k(\cdot; s^k)$ could denote the random yield per period from a resource k after extracting s^k units.

The value function $V(s_0)$ of the bandit problem can be written as follows. Let $X^k(s_t^k)$ denote the random variable with distribution $F^k(\cdot; s_t^k)$. Then the problem of finding an optimal allocation policy is the solution to the following intertemporal optimization problem:

$$V(s_0) = \sup_a \left\{ \mathbb{E} \sum_{t=0}^{\infty} \delta^t X^{a_t}(s_t^{a_t}) \right\}.$$

The celebrated index theorem due to Gittins and Jones (1974) transforms the problem of finding the optimal policy into a collection of k stopping problems. For each alternative k, we calculate the following index $Y^k(s_t^k)$, which depends only on the state variable of alternative k:

$$m^k\left(s_t^k\right) = \sup_{\tau} \left\{ \frac{\mathbb{E} \sum_{u=t}^{\tau} \delta^t X^k\left(s_u^k\right)}{\mathbb{E} \sum_{u=t}^{\tau} \delta^t} \right\}, \quad (1)$$

where τ is a stopping time with respect to $\left\{s_t^k\right\}$. The idea is to find for each k the stopping time τ that results in the highest discounted expected return per discounted expected number of periods in operation. The Gittins index theorem then states that the optimal way of choosing arms in a bandit

problem is to select in each period the arm with the highest Gittins index, $m^k(s_t^k)$, as defined by (1).

Theorem 1: **Gittins-Jones (1974)**

The optimal policy satisfies $a_t = k$ *for some k such that*

$$m^k(s_t^k) \geq m^j(s_t^j) \quad \text{for all } j \in \{1, \ldots, K\}.$$

To understand the economic intuition behind this theorem, consider the following variation on the original problem. This reasoning follows the lines suggested in Weber (1992). The arms are owned and operated by separate risk-neutral agents. The owner can rent a single arm at a time to the operators and there is a competitive market of potential operators. As time is discounted, it is clearly optimal to obtain high rental incomes in early periods of the model. The rental market is operated as a descending price auction where the fee for operating an arbitrary arm is lowered until an operator accepts the price. At the accepted price, the operator is allowed to operate the arm as long as it is profitable. Since the market for operators is competitive, the price is such that, under an optimal stopping rule, the operator breaks even. Hence the highest acceptable price for arm *k* is the Gittins index $m^k(s_t^k)$, and the operator operates the arm until its Gittins index falls below the price, that is, its original Gittins index. Once an arm is abandoned, the process of lowering the price offer is restarted. Since the operators get zero surplus and they are operating under optimal rules, this method of allocating arms results in the maximal surplus to the owner and thus the largest sum of expected discounted payoffs.

The optimality of the index policy reduces the dimensionality of the optimization problem. It says that the original *K*-dimensional problem can be split into *K* independent components, and then be knitted together after the solutions of the indices for the individual problems have been computed, as in eq. (1). In particular, in each period of time, at most one index has to be re-evaluated; the other indices remain frozen.

The multi-armed bandit problem and many variations are presented in detail in Gittins (1989) and Berry and Fristedt (1985). An alternative proof of the main theorem, based on dynamic programming can be found in Whittle (1982). The basic idea is to find for every arm a retirement value M_t^k, and then to choose in every period the arm with the highest retirement value. Formally, for every arm *k* and retirement value *M*, we can compute the optimal retirement policy given by:

$$V^k(s_t^k, M) \triangleq \max\left\{\mathbb{E}[X^k(s_u^k) + \delta V^k(s_t^{k+1}, M), M]\right\} \tag{2}$$

The auxiliary decision problem given by (2) compares in every period the trade-off between continuation with the reward process generated by arm *k* or stopping with a fixed retirement value *M*. The index of arm *k* in the state s_t^k is the highest retirement value at which the decision is just indifferent between continuing with arm *k* or retiring with $M = M(s_t^k)$:

$$M^k(s_t^k) = V^k(s_t^k, M^k(s_t^k)).$$

The resulting index $M^k(s_t^k)$ is equal to the discounted sum of flow index $m^k(s_t^k)$, or $M^k(s_t^k) = m^k(s_t^k) / (1 - \delta)$.

Extensions

Even though it is easy to write down the formula for the Gittins index and to give it an economic inpt, it is normally impossible to obtain analytical solutions for the problem. One of the few settings where such solutions are possible is the continuous-time bandit model where the drift of a Brownian motion process is initially unknown and learned through observations of the process. Karatzas (1984) provides an analysis of this case when the volatility parameter of the process is known.

From an analytical standpoint, the key property of bandit problems is that they allow for an optimal policy that is defined in terms of indices that are calculated for the individual arms. It turns out that this property does not generalize easily beyond the bandit problem setting. One instance where such a generalization is possible is the branching bandit problem where new arms are born to replace the arm that was chosen in the previous period (see Whittle 1981).

An index characterization of the optimal allocation policy can still be obtained without the Markovian structure. Varaiya, Walrand and Buyukkoc (1985) give a general characterization in discrete time, and Karoui and Karatzas (1997) provide a similar result in a continuous time setting. In either case, the essential idea is that the evolution of each arm depends only on the (possibly entire) history and running time of the arm under consideration, but not on the realization nor the running time of the other arms. Banks and Sundaram (1992) show that the index characterization remains valid under some weak additional condition even if the number of indices is countable, but not necessarily finite.

On the other hand, it is well known that an index characterization is not possible when the decision maker must or can select more than a single arm at each *t*. Banks and Sundaram (1994) also show further that an index characterization is not possible when an extra cost must be paid to switch between arms in consecutive periods. Bergemann and Välimäki (2001) consider a stationary setting in which there is an infinite supply of *ex ante* identical arms available. Within the stationary setting, they show that an optimal policy follows the index characterization even when many arms can be selected at the same time or when a switching cost has to be paid to move from one arm to another.

Market learning

In economics, bandit problems were first used to model search processes. The first paper that used a one-armed bandit problem in economics is Rothschild (1974), in which a single firm is facing a market with unknown demand. The true market demand is given by a specific probability distribution over consumer valuations. However, the firm initially has a prior probability over several possible market demands. The problem for the firm is to find an optimal sequence of prices to learn more about the true demand while maximizing its expected discounted profits. In particular, Rothschild shows that *ex ante* optimal pricing rules may well end up using prices that are *ex post* suboptimal (that is, suboptimal if the true distribution were to be known). If several firms were to experiment independently in the same market, they might offer different prices in the long run. Optimal experimentation may therefore lead to price dispersion in the long run as shown formally in McLennan (1984).

In an extension of Rothschild, Keller and Rady (1999) consider the problem of the monopolist facing an unknown demand that is subject to random changes over time. In a continuous time model, they identify conditions on the probability of regime switch and discount rate under which either very low or very high intensity of experimentation is optimal. With a low-intensity policy, the tracking of the actual demand is poor and the decision maker eventually becomes trapped, in contrast with a high-intensity policy demand, which is tracked almost perfectly. Rustichini and Wolinsky (1995) examine the possibility of mis-pricing in a two-armed bandit problem when the frequency of change is small. Nonetheless, they show that it is possible that learning will cease even though the state of demand continues to change.

The choice between various research projects often takes the form of a bandit problem. In Weitzman (1979) each arm represents a distinct research project with a random prize associated with it. The issue is to characterize the optimal sequencing over time in which the projects should be undertaken. It shows that as novel projects provide an option value to the research, the optimal sequence is not necessarily the sequence of decreasing expected rewards (even when there is discounting). Roberts and Weitzman (1981) consider a richer model of choice between R&D processes.

Many-agent experimentation

The multi-armed bandit models have recently been used as a canonical model of experimentation in teams. In Bolton and Harris (1999) and Keller, Rady and Cripps (2005) a set of players choose independently between the different arms. The reward distributions are fixed, but characterized by parameters that are initially unknown to the players. The model is one of common values in the sense that all players receive independent draws from the same distribution when choosing the same arm. It is assumed that outcomes in all periods are publicly observable, and as a result a free riding problem is created. Information is a public good and each individual player would prefer to choose the current payoff maximizing arm and let other players perform costly experimentation with currently inferior arms. These papers characterize equilibrium experimentation under different assumptions on the reward distributions. In Bolton and Harris (1999) the model of uncertainty is a continuous time model with unknown drift and know variance, whereas in Keller, Rady and Cripps (2005) the underlying uncertainty

is modelled by an unknown Poisson parameter.

Experimentation and matching

The bandit framework has been successfully applied to learning in matching markets such as labour and consumer good markets. An early example of this is given in the job-market matching model of Jovanovic (1979), who applies a bandit problem to a competitive labour market. Suppose that a worker must choose employment in one of *K* firms and her (random) productivity in firm *k* is parametrized by a real variable θ^k. The bandit problem is then a natural framework for the study of learning about the match-specific productivities. For each *k*, s_0^k is then simply the prior on θ^k and s_t^k is the posterior distribution given s_0^k and x_s^k for $s < t$. Over time, a worker's productivity in a specific job becomes known more precisely. In the event of a poor match, separation occurs in equilibrium and job turnover arises as a natural by-product of the learning process. On the other hand, over time the likelihood of separation eventually decreases as, conditional on being still on the job, the likelihood of a good match increases. The model hence generates a number of interesting empirical implications which have since been investigated extensively. Miller (1984) enriches the above setting by allowing for a priori different occupations, and hence the sequence in which a worker is matched over time to different occupations is determined as part of the equilibrium.

Experimentation and pricing

In a related literature, bandit problems have been taken as a starting point for the analysis of division of surplus in an uncertain environment. In the context of a differentiated product market and a labour market respectively, Bergemann and Välimäki (1996) and Felli and Harris (1996) consider a model with a single operator and a separate owner for each arm. The owners compete for the operator's services by offering rental prices. These models are interested in the efficiency and the division of the surplus resulting from the equilibrium of the model. In both models, arms are operated according to the Gittins index rule, and the resulting division of surplus leaves the owners of the arms as well as the operator with positive surpluses. In Bergemann and Välimäki (1996), the model is set in discrete time and a general model of uncertainty is considered. The authors interpret the experiment as the problem of choosing between two competing experience goods, in which both seller and buyer are uncertain about the quality of the match between the product and the preferences of the buyer. In contrast, Felli and Harris (1996) consider a continuous model with uncertainty represented by a Brownian motion and interpret the model in the context of a labour market. Both models show that, even though the models allow for a genuine sharing of the surplus, allocation decisions are surplus maximizing in all Markovian equilibria, and each competing seller receives his marginal contribution to the social surplus in the unique cautious Markovian equilibrium. Bergemann and Välimäki (2006) generalize the above efficiency and equilibrium characterization from two sellers to an arbitrary finite number of sellers in a deterministic setting. Their proof uses some of the techniques first introduced in Karoui and Karatzas (1997). On the other hand, if the market consists of many buyers and each one of them is facing the

same experimentation problem, then the issue of free-riding arises again. Bergemann and Välimäki (2000) analyse a continuous time model as in Bolton and Harris (1999), but with strategic sellers. Surprisingly, the inefficiency observed in the earlier paper is now reversed and the market equilibrium displays too much information. As information is a public good, the seller has to compensate an individual buyer only for the impact his purchasing decision has on his own continuation value, and not for its impact on the change in continuation value of the remaining buyers. As experimentation leads in expectation to more differentiation, and hence less price competition, the sellers prefer more differentiation, and hence more experimentation to less. As each seller has to compensate only the individual buyers, not all buyers, the social price of the experiment is above the equilibrium price, leading to excess experimentation in equilibrium.

Experimentation in finance

Recently, the paradigm of the bandit model has also been applied in corporate finance and asset pricing. Bergemann and Hege (1998; 2005) model a new venture or innovation as a Poisson bandit model with variable learning intensity. The investor controls the flow of funding allocated to the new project and hence the rate at which information about the new project arrives. The optimal funding decision is subject to a moral hazard problem in which the entrepreneur controls the unobservable decision to allocate the funds to the project. Hong and Rady (2002) introduce experimentation in an asset pricing model with uncertain liquidity supply. In contrast to the standard noise trader model, the strategic seller can learn about liquidity from past prices and trading volume. This learning implies that strategic trades and market statistics such as informational efficiency are path-dependent on past market outcomes.

See Also

- competition
- diffusion of technology

Bibliography

Banks, J. and Sundaram, R. 1992. Denumerable-armed bandits. *Econometrica* 60, 1071–96.

Banks, J. and Sundaram, R. 1994. Switching costs and the Gittins index. *Econometrica* 62, 687–94.

Bergemann, D. and Hege, U. 1998. Dynamic venture capital financing, learning and moral hazard. *Journal of Banking and Finance* 22, 703–35.

Bergemann, D. and Hege, U. 2005. The financing of innovation: learning and stopping. *RAND Journal of Economics* 36, 719–52.

Bergemann, D. and Välimäki, J. 1996. Learning and strategic pricing. *Econometrica* 64, 1125–49.

Bergemann, D. and Välimäki, J. 2000. Experimentation in markets. *Review of Economic Studies* 67, 213–34.

Bergemann, D. and Välimäki, J. 2001. Stationary multi choice bandit problems. *Journal of Economic Dynamics and Control* 25, 1585–94.

Bergemann, D. and Välimäki, J. 2006. Dynamic price competition. *Journal of Economic Theory* 127, 232–63.

Berry, D. and Fristedt, B. 1985. *Bandit Problems*. London: Chapman and Hall.

Bolton, P. and Harris, C. 1999. Strategic experimentation. *Econometrica* 67, 349–74.

Felli, L. and Harris, C. 1996. Job matching, learning and firm-specific human capital. *Journal of Political Economy* 104, 838–68.

Gittins, J. 1989. *Allocation Indices for Multi-Armed Bandits*. London: Wiley.

Gittins, J. and Jones, D. 1974. A dynamic allocation index for the sequential allocation of experiments. In *Progress in Statistics*, ed. J. Gani. Amsterdam: North-Holland.

Hong, H. and Rady, S. 2002. Strategic trading and learning about liquidity. *Journal of Financial Markets* 5, 419–50.

Jovanovic, B. 1979. Job search and the theory of turnover. *Journal of Political Economy* 87, 972–90.

Karatzas, I. 1984. Gittins indices in the dynamic allocation problem for diffusion processes. *Annals of Probability* 12, 173–92.

Karoui, N. and Karatzas, I. 1997. Synchronization and optimality for multi-armed bandit problems in continuous time. *Computational and Applied Mathematics* 16, 117–52.

Keller, G. and Rady, S. 1999. Optimal experimentation in a changing environment. *Review of Economic Studies* 66, 475–507.

Keller, G., Rady, S. and Cripps, M. 2005. Strategic experimentation with exponential bandits. *Econometrica* 73, 39–68.

McLennan, A. 1984. Price dispersion and incomplete learning in the long run. *Journal of Economic Dynamics and Control* 7, 331–47.

Miller, R. 1984. Job matching and occupational choice. *Journal of Political Economy* 92, 1086–120.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 55, 527–35.

Roberts, K. and Weitzman, M. 1981. Funding criteria for research, development and exploration of projects. *Econometrica* 49, 1261–88.

Rothschild, M. 1974. A two-armed bandit theory of market pricing. *Journal of Economic Theory* 9, 185–202.

Rustichini, A. and Wolinsky, A. 1995. Learning about variable demand in the long run. *Journal of Economic Dynamics and Control* 19, 1283–92.

Varaiya, P., Walrand, J. and Buyukkoc, C. 1985. Extensions of the multiarmed bandit problem: the discounted case. *IEEE Transactions on Automatic Control* AC-30, 426–39.

Weber, R. 1992. On the Gittins index for multi-armed bandits. *Annals of Applied Probability* 2, 1024–33.

Weitzman, M. 1979. Optimal search for the best alternative. *Econometrica* 47, 641–54.

Whittle, P. 1981. Arm-acquiring bandits. *Annals of Probability* 9, 284–92.

Whittle, P. 1982. *Optimization Over Time*, vol. 1. Chichester: Wiley.

How to cite this article

Bergemann, Dirk and Juuso Välimäki. "bandit problems." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000312> doi:10.1057/9780230226203.0088

The New Palgrave Dictionary of Economics Online

Bank of England

Charles A.E. Goodhart
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Bank of England, founded in 1694 to finance war against France, soon became Britain's largest bank. It became responsible for maintaining the gold standard and acting as lender of last resort. To do so, it had to withdraw from commercial banking. After failing to stay on gold (1931) the Bank became subservient to the Chancellor in macro-monetary policy and was nationalized in 1946. Operational independence to set interest rates in pursuit of an inflation target was restored in the 1990s, while its previous functions, notably bank supervision, debt management, and foreign exchange intervention, fell away.

Keywords

Bagehot, W.; Bank for International Settlements; Bank of England; Bank of Scotland; Bank rate; banking crises; banking supervision; Basel Committee on Banking Regulation and Supervisory Practices; Big Bang; Bretton Woods system; Bullion; central bank independence; central banking; cost-push theory of inflation; Dutch East India Company; European Monetary System; exchange controls; exchange rate mechanism; exchange rate targets; financial intermediaries; financial liberalization; financial repression; Financial Services Authority; Friedman, M.; gold standard; incomes policies; inflation; inflation targeting; London Clearing House; Medium Term Financial Strategy (UK); monetarism; monetary policy; monetary targets; natural rate of unemployment; Phillips curve; Radcliffe Report; Royal Bank of Scotland; South Sea Company; stagflation; Thornton, H.; trade unions; treasury bills; velocity of circulation

Article

The primary motivation for the establishment of the Bank of England was the need to raise funds to help the government finance the then current war against France, although the view had also developed that a bank could help to ‘stabilize’ financial activity in London given periodic fluctuations in the availability

of currency and credit. An original proposal by William Paterson in 1693 for a government ‘fund of perpetual interest’ was turned down in favour of another proposal by Paterson in 1694 to establish a company known as the Governor and Company of the Bank of England, whose capital, once raised, would be lent in its entirety to the government.

An ordinary finance act, now known as the Bank of England Act (1694), stipulated that the Bank was to be established via stock subscriptions which were to be lent to the government. A governor, deputy governor and 24 directors were to be elected by stockholders (holding £500 or more of stock).

The evolution of the Bank's objectives and functions, 1694 – 1914

Under its original charter the Bank was allowed to issue bank notes, redeemable in silver coin, as well as to trade in bills and bullion. The notes of the Bank competed with other paper media of exchange, which comprised notes issued by the Exchequer and by private financial companies. In addition, customers could maintain deposit accounts with the Bank, which were transferable to other parties via notes drawn against deposit receipts (known as accomptable notes), thus providing an early form of cheque.

An early customer of the Bank was the Royal Bank of Scotland, which made arrangements to keep cash at the Bank from its outset in 1727. Loans were extended, predominantly in the form of discounting of bills, to individuals and companies, and the Bank undertook a large amount of lending (often via overdrafts) to the Dutch East India Company and, from 1711, to the South Sea Company. The Bank also acted as a mortgage lender, although this business never took off, and ceased some years later. Finally, an important function of the Bank was the remittance of cash to Flanders and elsewhere for the wars against Louis XIV, which was facilitated through correspondent arrangements with banks in Holland.

In 1697 the renewal of the Bank's charter for another ten years involved the passage of a second Bank Act, which increased the capital of the Bank and prohibited any other banks from being chartered in England and Wales. This monopoly was strengthened at the next renewal of the Bank's charter in 1708, when any association of six or more persons was forbidden to engage in banking activity, thereby precluding the establishment of any other joint stock banks. The Bank's position as banker to the government was consolidated in 1715 when it was decided that subscriptions for government debt issues would be paid to the Bank, and further that the Bank was to manage the government debt (the Ways and Means Act). The Bank then acted as manager of the government's debts from that date until 1997.

The Bank also encouraged the use of its own notes in preference to other media of exchange by persuading the Treasury to increase the denomination of Exchequer bills. By 1725 the Bank's notes had become sufficiently widely used as to be pre-printed for the first time. Although a number of private banks had developed by 1750, both within and outside London, none competed seriously with the Bank in the issue of notes. By 1770 most London bankers had ceased to issue notes, using Bank of England notes (and cheques) to settle balances among themselves in what had become a well-developed clearing system. Furthermore, in 1775 Parliament raised the minimum denomination for any non-Bank of England notes to one pound and, two years later, to five pounds, effectively guaranteeing the use of Bank of England notes as the dominant form of currency. Problems relating to counterfeiting, and to the harsh treatment of those caught in the act, were, however, perennial.

In Scotland, by contrast, no note issuing monopoly existed, and banks were free to issue notes, although two banks dominated, namely, the Bank of Scotland and the Royal Bank of Scotland. Furthermore, several private note-issuing banks were in business in Ireland, and the Bank of Ireland was established in

1783. These banks relied on the Bank of England to obtain silver and gold, particularly during times of financial stress, such as 1783 and 1793.

Following a dramatic rise in government expenditures after 1793 due to the war against France, which caused a large rise in the Bank's note issue, the Bank's gold holdings fell sharply. After a scare about a French invasion convertibility was suspended in 1797, and resumed only in 1821. In view of the financial exigencies of the war, and the fact that there was in such circumstances no limit to the expansion of its note issue, now effectively legal tender, by the Bank, a privately owned company, what is in retrospect surprising about the period of suspension is how comparatively low the resulting inflation was. Even so, it was high enough to set off a major debate on its causation, for example in the Parliamentary Committee on the High Price of Bullion (1810). This period saw a further consolidation of the Bank as a note issuer, since it began to issue small denomination notes (given the shortage of silver and gold coin), which became legal tender in 1812. Furthermore, in 1816 silver coin ceased to be legal tender for small payments. The government also moved most of its accounts to the Bank in 1805 (in 1834 all government accounts were finally moved to the Bank).

During the 18th century and early part of the 19th century, smaller country banks had proliferated throughout England and Wales, many issuing their own notes. Given the prohibition on joint stock banking, the capital of these banks was usually small, and they regularly became insolvent, especially when the demand for cash (coin) became strong. This contrasted sharply with Scotland, where joint stock banking and branch banking were permitted, and relatively few failures occurred. Following a severe banking crisis in 1825, during which many English country banks failed, an Act renewing the Bank's charter (in 1826) abolished the restrictions on banking activity more than 65 miles outside of London. This led to the establishment of several joint stock banks, while the Bank countered by opening several branches throughout England.

Thus, a semblance of a banking ‘system’ began to emerge by 1830, with the Bank of England as the ‘central’ bank. By far the best book on such nascent central banking at this time was that written by Henry Thornton, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802).

The practice of banks placing surplus funds with bill brokers also developed, with the Bank beginning to extend secured loans to these brokers on a more or less regular basis. In 1833 joint stock banks were finally allowed to operate in London, although they were not permitted to issue notes and thus were essentially deposit-taking banks only. The same Act specified that Bank of England notes were legal tender, and the Bank was also given the freedom to raise its discount rate freely (until then usury laws had placed a ceiling on interest rates) in response to cash outflows. The Bank's reaction (an early reaction function), in varying its interest rate, to cash inflows and outflows became codified around this time in what became known as the Palmer rule, after Horsley Palmer, Governor 1830–33, though the rule itself is usually dated from 1827.

The position of Bank of England notes was consolidated in an important Act, passed in 1844, generally known as the Bank Charter Act, preventing all note issuers from expanding their note issue above existing levels, and prohibiting the establishment of any new note-issuing banks. The 1844 Act also separated the issue and banking functions of the Bank into different departments, and required the Bank to publish a weekly summary of accounts.

Given that it did not pay interest on its deposits, the deposit activity of the Bank could never really compete with that of other banks, which expanded rapidly from 1850 onwards. In 1854, joint stock banks in London joined the London Clearing House, and it was agreed that clearing by transfer of Bank

of England notes would be abandoned in favour of cheques drawn on bank accounts held at the Bank. Ten years later the Bank of England itself entered this clearing arrangement, and cheques drawn on bankers' accounts at the Bank became considered as paid.

Although the Bank had, from the beginning of the 19th century, periodically bought or sold exchequer bills to influence the note circulation, explicit open-market borrowing operations to support its discount rate began in 1847. From 1873 until 1890 the Bank almost always acted as a borrower rather than a lender of funds, as there were typically cash surpluses. As a result, the Bank introduced the systematic issue of Treasury bills via a regular tender offer in 1877. Treasury bills had a much shorter maturity (three to twelve months) than Exchequer bills (five or more years), and were to play an important role in raising funds from the outset of the First World War onwards.

By 1890, the Bank's role as lender of last resort became undisputed when it orchestrated the rescue of Baring Brothers and Co., a bank whose solvency had become suspect, threatening to cause systemic problems. Earlier, in 1866, the failure of a discount house, Overend, Gurney and Co., had precipitated a financial panic, during which the Bank discounted large amounts of bills and extended considerable loans. The Bank, however, was criticized for not doing more to prevent the onset of such a panic, not least by Walter Bagehot in his famous book *Lombard Street* (1873).

Throughout the 19th century, the Bank streamlined its discount facilities. In 1851 it overhauled its discount rules, stipulating that only those parties having a discount account could present bills, and that these bills had to have a maturity of fewer than 95 days and be endorsed by two creditworthy firms. In the latter part of the century, however, the Bank gradually came to favour discount houses, often by presenting them with better rates of discount, and the range of firms doing discount business with the Bank declined. Discount houses were favoured because there was tension then between the Bank and the rapidly growing commercial banks – there was much banking consolidation via mergers between the 1870s and 1914 – and dealing via the intermediation of the discount houses enabled the Bank to influence market rates without having to interact directly with the joint-stock banks as counterparties.

Until the First World War the Bank pursued a discount policy which was primarily aimed at maintaining its gold reserves (as noted earlier) and which was conducted largely independently of the government. During the First World War, however, a clash occurred between the Bank Governor (Cunliffe) and the Chancellor (Law), during which the government made clear that it bore the ultimate responsibility for monetary policy, and that the Bank was expected to act on its direction.

A subservient Bank, 1914 – 1992

The First World War was a major watershed not only in the history of the Bank but in the world more widely. It ushered in a half-century of increasing government intervention in every country, of a move towards socialist economies in most, and of communism in a wide swathe of countries. Under these circumstances the Bank became increasingly subservient to the government, in practice to the Chancellor of the Exchequer and to the Treasury, in the conduct of macro-monetary policy, its previous primary function.

Initially, however, there was little perception that the war and the rise of socialist ideas had irretrievably altered the context for policy. There was a desire to return to the previous regime, the gold standard, with its tried and true verities, as expressed in the Cunliffe Committee Report (the first report of the Committee on Currency and Foreign Exchange, 1919). That was probably inevitable under the

circumstances, but a much more questionable decision was to return at the pre-war parity (against gold) despite the war-induced loss of markets (especially for the UK's main staples, textiles, coal, and iron and steel) and of competitiveness. Several of the other belligerent states, notably France, had inflated, and allowed their exchange to float downwards by so much that they did not seek to re-peg at the previous parity, but could choose a more suitable and competitive rate. While the decision to return to gold at the pre-war parity, steadfastly supported by the Bank, has been much criticized, the modern theory of time inconsistency provides some defence, namely, if the Bank had started to change the chosen rate to suit the immediate conjuncture it would have been expected to do so again in future, making commitment to the regime less credible.

Be that as it may, conditions after the First World War, with a weak balance of payments and a massively inflated money stock and floating debt, were hardly conducive to the re-establishment of gold standard conditions. Indeed, the authorities initially felt forced to move in the other direction, to unpeg the sterling–dollar rate that had been established since 1916 and formally to leave the gold standard in March 1919. The ending of the war led then to an extremely sharp and short boom and bust, in which tight monetary policy played a major role in the subsequent deflation (see Howson, 1975). From then until the return to gold at the pre-war parity of $4.86 to the pound in 1925, the Bank advocated keeping the Bank rate high enough to facilitate that regime change, but decisions on Bank rate and on the conduct of monetary policy were joint, in that no proposal by the Bank could be activated without the agreement of the Chancellor and HM Treasury; the Treasury view, however, then was in line with classical thought, namely, that monetary policy could and should impinge primarily on nominal prices, with real output affected by real factors.

Despite the boom in the USA, growth in the UK was perceived as remaining low and unemployment high, at least as compared with its main comparator countries, in the 1920s. This was in part due to the continuing problems of restoring a successful economic regime in Europe, wherein German reparations had a malign effect. Although the Bank had lost much of its power to direct domestic monetary policy (to Whitehall), the Bank and its Governor, Montagu Norman, played a leading role in the various international exercises to try to restore Europe to normality and to the gold standard, (Sayers, 1976, ch. 8); and Sir Otto Niemeyer, a top Bank official, spread the gospel of establishing central banks to maintain price stability to the Dominions.

This whole structure came apart in the crisis that started in the USA in 1929 and then engulfed the rest of the world progressively through the subsequent four years. How far that collapse was itself exacerbated by the attempt to restore the gold standard has been explored by Eichengreen (1992). The UK was not in a strong economic position to avoid the world recession, but suffered a much smaller decline in output than in the USA or much of Continental Europe. The struggle to maintain the gold standard had required the maintenance of high interest rates, despite the imposition of controls on new issues in sterling by foreign governments. Despite high unemployment, wages and prices remained too sticky to allow the restoration of international competitiveness, though quite why this was so remains a debated issue.

With the gold standard collapsing in Europe and social pressures rising in the UK, there was diminishing political will to take the measures that appeared necessary to maintain the gold standard. The government decided to abandon it (in Norman's absence) in September 1931. From that moment onwards, until May 1997, the decision to alter the Bank rate moved decisively to Whitehall, effectively into the hands of the Chancellor, advised by HM Treasury. Of course, the Bank could, and did, make

suggestions and played a major role in all the discussions, but the Chancellor took the decisions. Indeed, from June 1932 until November 1951 a policy of cheap money was followed whereby Bank rate was held constant at two per cent. Norman stated in 1937, ‘I am an instrument of the Treasury’.

Meanwhile, the Bank was becoming more professional. The old system of circulating the Governor's chair in turn among the directors of the Bank, who were appointed from city (but not commercial bank) institutions, was superseded by the continuing governorship of Montagu Norman from 1920 until 1944. While this arose by happenstance rather than intention (see Sayers, 1976, ch. 22), it gave the Bank highly skilled, even if also highly idiosyncratic, leadership. Moreover, Norman introduced economists and other able officials into both the staff and the Court (the largely ceremonial board) of the Bank, although it is (apocryphally) recorded that Norman told one such economist, ‘You are not here to tell me what to do, but to explain why I have done what I have already decided to do.’

In effect, the Bank had already become nationalized by the end of the Second World War. So the formal act of nationalization in 1946 brought about no real substantive changes, except that the Governor and his deputy (there has as yet been no woman Governor, although Rachel Lomax became the first female Deputy Governor in 2003), were appointed by the government for five years, renewable once more in most cases. Indeed, the more profound changes were brought about by Governor Gordon Richardson (1973–83) in the early 1980s. Until then, the Governor had been rather akin to a chairman, with the deputy and other internal directors as members of the board, setting strategy. Much of the executive power still lay with the Chief Cashier, who acted as leader of the heads of department, who ran the Bank. There was a clear break, a division, between the staff in the departments on the one hand and the Governors and Directors on the other. Richardson changed all that, concentrating power in the Governors’ hands, sharply demoting the role of Chief Cashier, and underlining the precedence of (internal) directors over heads of department in all policy matters.

So, as power to decide the course of monetary policy – and to set the Bank rate – passed to Whitehall, what did these professional central bank officials do? The Bank came to have three main areas of responsibility. The first was the management of markets, notably the money market, the bond (gilts) market and the foreign exchange market. The UK had come out of the Second World War with a massively inflated ratio of debt to GDP, and its management had remained difficult and delicate, at least until after the War Loan Conversion of 1932. No sooner, however, had debt management been thereby put on a sounder foundation than the Second World War led to a further upsurge in the debt ratio, which led once again to debt management becoming a major preoccupation of policy. Thereafter, a combination of generally prudent fiscal policies, so that the debt ratio fell steadily, and then unexpected inflation in the 1970s, which accelerated the decline in the debt ratio, and market reforms in the 1980s, enabled the procedures of debt management to become simpler and standardized. Similarly, the floating exchange rate in the 1930s, followed by attempts to maintain pegged exchange rates both during the Second World War and thereafter under the Bretton Woods system, against a background of perennially weak balance of payments conditions, made the management of the UK's foreign exchange reserves and intervention on the foreign exchange market a crucial function of the Bank until 1992, when the UK was forced out of the European exchange rate mechanism. During crises the officials in charge of such foreign exchange operations were in telephone communication with the Chancellor and, occasionally, the Prime Minister at frequent intervals.

The Bank held that such market operations required a special professional expertise (though HM Treasury remained sceptical). The Bank threw itself into such activities with enthusiasm, and defended

its pre-eminent role in this respect stoutly against all outside encroachment or criticism. Indeed, its market ‘savvy’ was its most powerful lever to persuade the Chancellor to its views in any debate; ‘I am sorry, Chancellor, but the market will not accept that policy’ was the strongest card it had to play, and that card was played often and with alacrity.

Although ultra-cheap money, with Bank rate held at two per cent, was abandoned in 1951, when the Conservative Party was returned to office, monetary policy in general, and interest rates in particular, were still seen as both more ineffective and uncertain in their impact on domestic demand than the supposedly more reliable fiscal policy, a conclusion upheld by the controversial Radcliffe Report (1959).

Consequently, fiscal policy was used to try to steer domestic demand while interest rates were raised to protect the balance of payments during the regular bouts of external weakness, and otherwise held low both to ease government finance and to support fixed investment. The outcome was a system in which inflationary pressures regularly threatened both the internal and external value of the currency. The chosen solution was to supplement market measures by direct interventions, in the case of external pressure via exchange controls, in the case of monetary expansion via direct controls on bank lending to the private sector. In both instances the Bank acted as the administrative agent of HM Treasury.

Such direct controls were introduced (on bank lending), or greatly extended and tightened (exchange controls), with the onset of the Second World War in 1939, but were continued, for the reasons outlined above, until 1971 for bank lending and 1979 for exchange controls. The administration of exchange controls required a large staff, but, unlike with its market operations, the Bank had little enthusiasm for acting in this guise. The Bank hoped to restore London to its former role as an international financial centre. While it succeeded in this through its encouragement of the Eurodollar market, aided by inept US policies, the continued administration of exchange controls remained an unwelcome burden. The same was true for direct controls on bank lending. Such controls were regarded by politicians as a comparatively painless way of dampening demand and inflation, while they were resented by commercial bankers. The Bank found itself in the middle of these disputes, and grew painfully aware of such controls’ stultifying effect on efficiency, dynamism and growth. The Bank, inspired by John Fforde (the then executive director in charge of domestic finance, and subsequent Bank historian), pressed hard for these controls to be dismantled, and succeeded with the liberalizing reform of Competition and Credit Control (Bank of England, 1971).

As with many other cases of banking liberalization, such as in Scandinavia at the end of the 1980s, this was followed by an expansionary boom and then a bust, the fringe (secondary) bank crisis of 1973/74 (Reid, 1982). While there remain questions about how monetary policy could have been better applied to prevent the prior monetary boom (1972/73), there was no question but that the financial crisis found both the Bank and the banks unskilled in risk management and unprepared for adverse shocks to financial stability. The long period of financial repression – that is, controls on bank lending to the private sector and force-feeding with government debt – had had the by-product of making the (core) commercial banking system safe between the mid-1930s and the early 1970s. The central banking function of maintaining financial stability, via regulation and supervision, had atrophied.

This had not been so earlier, and the Bank had been closely involved in the rescue of Williams Deacon's Bank by the Royal Bank of Scotland in 1930 (Sayers, 1976, ch. 10), and in helping to shape the structure of both the commercial banking system and the London Discount Market Association. Williams Deacon's had got into trouble largely because of bad debts from Lancashire cotton companies. Norman, and the Bank, extended their structural interventions beyond banking to try to encourage strategic

amalgamations to shore up the positions of weakened companies in a variety of industries, such as cotton, steel, shipping, armaments (Sayers, 1976, ch. 14). The Bank's involvement in structural matters outside of banking itself was episodic depending on both circumstances and personalities. Another example of such Bank involvement was the considerable role it played in the reform of the UK capital market in the 1980s, more familiarly known as ‘Big-Bang’. But views on whether the Bank has any locus in such wider structural issues vary over time; the early 2000s saw a major withdrawal by the Bank from any such involvement.

The fringe bank crisis in the early 1970s was, however, a clarion call to put more emphasis on its third main function, bank supervision and regulation. The immediate result was a reorganization in the Bank. Initially a nucleus of a new specialized department was established in the Discount Office where the limited staff assigned to this role had sat, which rapidly absorbed staff and resources. Thereafter this became a separate department devoted to banking supervision and regulation (its first head was George Blunden, later to become Deputy Governor, who handed it on to Peter Cooke in 1976). Its position was regularized in the Banking Act (1979) which gave formal powers to the Bank to authorize, monitor, supervise, control and, under certain circumstances, withdraw prior authorization (tantamount to closure) for banks. No such powers had been available before that date. Meanwhile, other financial intermediaries, such as building societies or insurance companies, remained (lightly) regulated by various government departments.

The fringe bank crisis was almost entirely domestic, confined to British headquartered companies. Meanwhile, however, the onwards march of liberalization (involving the removal of direct controls, notably exchange controls in 1979) and of information technology were leading to a growing internationalization of financial business. For a variety of reasons, mostly relating to the innovation of the Eurodollar and Euro-markets, London regained its role as an international financial centre in the 1960s, and thus international monetary problems became of particular importance to the Bank, which took a leading role in such matters from the 1970s onwards.

Central bankers had met regularly at the headquarters of the Bank for International Settlements (BIS) in Basel for many years. It was, therefore, a logical step for supervisory officials also to come together at Basel on regular occasions to discuss matters of common interest. Thus was born (in 1974), as a result of an initiative from Gordon Richardson, the Basel Committee on Banking Regulation and Supervisory Practices. For the first 15 years of its existence it was chaired by the participant from the Bank of England, and was usually known by his name; thus, the Blunden Committee (1974–77) gave way in due course to the Cooke Committee (1977–88). The failures of Franklin National and Herstatt prompted the First Basel Concordat, which allocated responsibility for supervising internationally active banks to home and host authorities.

So by the mid-1970s, a need was perceived for banking supervision at both the domestic and, via consolidation, at the international levels. The purpose of these initiatives was to clarify where responsibility lay for the supervision of international banks, to prevent fragile, and possibly fraudulent, banking leading to avoidable failures and potential systemic crises.

Despite the growing number of bank supervisors, and notable success in reversing prior declines in capital ratios, the history of banking in the subsequent decades in the UK was spotted by occasional bank failures. Unlike the fringe bank crisis, none was, or was allowed to become, systemic, nor did individual depositors lose any money, except in the case of Bank of Credit and Commerce International (BCCI), and even in that case the deposit protection scheme provided some relief. The failures of

Johnson–Matthey (in 1984), BCCI (in 1991) and Barings (in 1995) were all isolated cases of bad, in some respects fraudulent, banking.

The main problem of the 1970s and 1980s was, however, that of combating inflation, which soared to heights previously unknown, not only in peacetime but even in wartime, during the 1970s, up to 25 per cent per annum. There were three main theories, though divisions between them were never completely distinct. The first was the cost-push theory, that inflation was driven by over-mighty trade unions, seeking to increase the relative real pay of their members; the appropriate remedy was then prices and incomes policies plus reform (and constraint) of trades unions. The second was the (vertical) Phillips curve analysis; the remedy here was to raise unemployment above the ‘natural’ rate to reduce inflation. The third was that inflation was a monetary phenomenon; the remedy was to control the rate of growth of the (appropriate) monetary aggregate.

Until the mid-1970s, both major political parties, the Bank and HM Treasury all professed some combination of theories 1 (cost-push) and 2 (Phillips curve). Left-leaning politicians, academics and officials tended to put more weight on cost-push. In the 1960 and 1970s the third, monetarist, view seemed to explain events better and gained strength, not only in the USA (Milton Friedman) but also in the UK. In particular, the surge in inflation in the UK in 1973–75 followed closely behind the rapid expansion of broad (but not narrow) money in 1972–73. So, when in opposition, the leading Conservative politicians Keith Joseph and Margaret Thatcher embraced a version of monetarism. When they came to power in 1979, they tried to commit monetary policy to follow a target for broad money, via the Medium Term Financial Strategy. In order to achieve this, nominal, and real, interest rates were kept high, and the exchange rate appreciated sharply, partly under the influence of North Sea oil and confidence in Thatcherite policies. Inflation duly declined, as planned, but broad money growth did not. This latter was partly due to the abolition of the ‘corset’ in 1980. The ‘corset’ was a reformulated, and somewhat disguised, direct control over commercial bank expansion that had been pressed into service on several occasions during the 1970s. The Bank was glad to see the end of exchange controls and direct controls over bank lending, but had never shared the government's monetarist faith in trying to set, and stick to, targets for the growth of (the various) monetary aggregates. The empirical demonstration of the unpredictability of the relationship between (broad) money and nominal incomes in the early 1980s soon weakened the government's own faith. After moving from one monetary target to several joint targets, and an attempt to hit the broad money target by ‘overfunding’, an exercise criticized by many as artificial, the government abandoned its monetary targetry in 1986. That left the question of how monetary policy, and with it control of inflation, was to be managed or, in the standard phrase, ‘anchored’. The then Chancellor, Nigel Lawson, wanted to ‘anchor’ by joining the exchange rate mechanism (ERM) of the European Monetary System and leaving the steering of monetary policy to the Bundesbank. The Prime Minister, Mrs Thatcher, and her adviser, Alan Walters, were opposed, both on economic grounds (that such a pegged system was ‘half-baked’) and for wider political reasons. There was a battle royal in which the Bank was left on the sidelines. Lawson was sacked, but eventually Mrs Thatcher was, grudgingly, persuaded to allow the UK to join the ERM in October 1990.

This was in the aftermath of German reunification, and the expenditures connected with that led the Bundesbank to keep interest rates higher than was tolerable for the UK (or Italy). The UK was in the throes of a sharp downturn in housing prices, following an unstable housing boom in the late 1980s. With the Conservatives having become politically weaker, there was just no stomach to raise interest

rates to the levels necessary to sustain the ERM. The UK was forced out in September 1992.

Independent and focused, 1992 –

The ejection of the UK from the ERM left the government and HM Treasury with the recurrent problem of how to manage, to ‘anchor’, monetary policy. Both monetary and exchange rate targets had been tried, and both had been found wanting. While the economic experience of the 1980s was better than that of the stagflationary 1970s, it was hardly stellar, with a boom–bust cycle at the end of the decade. Meanwhile, a new approach had been adopted in New Zealand, whereby the central bank was given administrative freedom to vary interest rates for the purpose of hitting a target for the inflation rate, jointly set by the government and the central bank: that is, inflation targetry. This obviated one of the shortcomings of monetary targetry, namely, the unpredictability of the velocity of money; it left setting the goals of policy, the overall strategy, in the hands of government, but shifted the (constrained) discretion to vary interest rates to the professional and technical judgement of the central bank. This procedure soon generated a strong body of academic support (for example, Fischer, 1994).

Although Conservative Chancellors (both Lawson and Lamont) had toyed with the idea of giving the Bank operational independence, consecutive Prime Ministers (Thatcher and Major) refused, primarily on political grounds. Nevertheless Lamont wanted to move to an inflation target. But there was a problem of governmental credibility. To foster credibility, Lamont now encouraged (in 1992/93) the Bank to prepare and to publish an independent forecast of the likely projection for inflation, the *Inflation Report* (on the assumption of unchanged policies); this was a reversal of prior habits whereby HM Treasury and Ministers customarily censored Bank publications and discouraged any publication of internal Bank forecasts. The process of gradually giving the Bank a more independent role in setting monetary policy took a step further when the next Chancellor, Clarke, not only held a meeting with the Governor, and the Bank, to discuss future changes in interest rates, but published the minutes of the meeting, including the Governor's initial statement, verbatim; this was termed the Ken (Clarke) and Eddie (George) show. That said, Clarke had strong views on the appropriate policy and on a couple of occasions overruled the Governor's suggestions.

At that time – the mid-1990s – there were still question marks over the Labour Party's ability to manage the economy; financial markets are inherently suspicious of left-leaning governments. So Labour had more to gain (than the Conservatives), in terms of confidence and lower interest rates, by granting operational independence (back) to the Bank. In advance of the 1997 election the then shadow Chancellor, Gordon Brown, was cautious; while indicating general support for both inflation targetry and operational independence, he stated that he wanted time to see how well the Bank performed before granting such independence. But, within days of winning the election, he made that strategic change to the monetary regime.

This was, of course, a great prize for the Bank, but it did not come without cost. In the same month as operational independence was awarded to the Bank, both debt management and banking supervision were hived off, to a separate Debt Management Office (DMO) and Financial Services Authority (FSA) respectively. With the government debt to GDP ratio having declined and capital markets strengthened, debt management had become more of a routine and standardized exercise. Nevertheless, its departure to the DMO, and the fact that the float of the exchange rate after 1992 was kept ‘clean’, that is, without intervention, meant that much of the market operations which had been so central to the Bank in the post-

Second World War period disappeared, though its money market operations, of course, continued. The administration of direct controls had gone at the beginning of the 1980s. And now banking supervision was also taken away. This meant that almost *all* the prime functions that the Bank had undertaken in its post-Second World War period of subservience had now gone. Instead, the Bank was now focused on varying interest rates to achieve the inflation target set for it by the Chancellor.

There are numerous arguments, quite evenly balanced, for whether bank supervision should be kept within a central bank or put with a separate Financial Services Authority (FSA), covering both banks and other financial intermediaries (see Goodhart, 2000). Be that as it may, there are various aspects of the financial system, such as oversight of the payments' system, and of crisis management, such as lender of last resort functions, which cannot be delegated to an FSA. Moreover, the achievement of price stability is likely to be seriously compromised by any serious bout of financial instability – and vice versa, with financial stability adversely affected by price instability. So the removal of individual bank supervision does not absolve the Bank from concern with financial stability issues more widely; indeed, the Bank is specifically charged with maintaining overall systemic stability in the financial system. But exactly what that means when responsibility for the conduct of individual bank supervision is located elsewhere is not yet entirely clear.

What it certainly does mean is that the FSA, the Bank, and the political authorities as the ultimate source of any needed fiscal support have to work extremely closely together, in advising on any new regulations (whether domestic or international), in monitoring developments (as in the Financial Stability Review), and in crisis management. This latter task would be done via the Tripartite Standing Committee (FSA, Bank, and HM Treasury), set up in 1997, although so far no such financial (as contrasted with simulated 'war games') crisis has occurred, though the Committee did meet after the terrorist attacks on 7 July, 2005. How successful crisis management by such a committee may be has yet to be seen.

The monetary policy function of the Bank, now its central preoccupation, has, however, been very successful by all the usual criteria. In several papers Luca Benati (for example, Benati, 2005) has demonstrated that the variance of both GDP and of inflation around its target has been lower under the inflation targetry regime (whether taken as starting in 1992 or in 1997) than under any previous historical regime. The procedures of having a Monetary Policy Committee consisting of five senior Bank officials and four outside experts (appointed by the Chancellor), with the Committee serviced by Bank staff, has worked generally smoothly and well. So the Bank's reputation and credibility have rarely been higher, although now tightly focused on one main function.

See Also

- banking crises
- bullionist controversies (empirical evidence)
- gold standard
- inflation targeting
- monetary policy, history of

Bibliography

Acres, W. 1931. *The Bank of England from Within*. London: Oxford University Press.

Andréadès, A. 1909. *A History of the Bank of England*. London: P. S. King and Sons.

Bagehot, W. 1873. *Lombard Street*. London: Kegan, Paul and Co.

Bank of England. 1971. *Competition and Credit Control*. London: Bank of England.

Bank for International Settlements. 1963. ‘Bank of England’, in *Eight European Central Banks*. Basle: Bank for International Settlements.

Benati, L. 2005. The Inflation-targeting framework from an historical perspective. *Bank of England Quarterly Bulletin* 45(2), 160–8.

Bowman, W. 1937. *The Story of the Bank of England: From its Foundation in 1694 until the Present Day*. London: Herbert Jenkins.

Chapham, R. 1968. *Decision Making: A Case Study of the Decision to Raise the Bank Rate in September 1957*. London: Routledge and Kegan Paul.

Clapham, J. 1944. *The Bank of England: A History*. Cambridge: Cambridge University Press.

Clay, H. 1957. *Lord Norman*. London: Macmillan.

Committee on Currency and Foreign Exchange After the War (Cunliffe Committee). 1918. *First Interim Report*, Cmnd. 9182; and 1919. *Final Report*, Cmnd 464, London: HMSO.

Eichengreen, B. 1992. *Golden Fetters: The Gold Standard and the Great Depression*. New York: Oxford University Press.

Feavearyear, A. 1963. *The Pound Sterling: A History of English Money*, 2nd edn, rev. E. Morgan. Oxford: Clarendon.

Fforde, J. 1992. *The Bank of England and Public Policy 1941–1958*. Cambridge: Cambridge University Press.

Fischer, S. 1994. ‘Modern central banking’, in F. Capie, C. Goodhart, S. Fischer and N. Schnadt, *The Future of Central Banking*. Cambridge: Cambridge University Press.

Geddes, P. 1987. *Inside the Bank of England*. London: Boxtree.

Giuseppi, J. 1966. *The Bank of England: A History from its Foundation in 1694*. London: Evans Brothers Limited.

Hennessey, E. 1992. *A Domestic History of the Bank of England 1930–1960*. Cambridge: Cambridge University Press.

Goodhart, C. 2000. The organisational structure of banking supervision. Special Paper No. 127. London: Financial Markets Group Research Centre, London School of Economics. Subsequently published in *Economic Notes* 31, 1–32.

Howson, S. 1975. *Domestic Monetary Management in Britain, 1919–38*. Cambridge: Cambridge University Press.

Radcliffe Report. 1959. *Report: Committee on the Working of the Monetary System*. Cmnd 827. London: HMSO.

Reid, M. 1982. *The Secondary Banking Crisis, 1973–75: Its Causes and Course*. London: Macmillan.

Richards, R. 1929. *The Early History of Banking in England*. London: Frank Cass and Co.

Rogers, J. 1887. *The First Nine Years of the Bank of England*. Oxford: Clarendon.

Sayers, R. 1936. *Bank of England Operations, 1890–1914*. London: P.S. King and Son.

Sayers, R. 1957. *Central Banking After Bagehot*. Oxford: Clarendon.

Sayers, R. 1976. *The Bank of England, 1891–1944*. Cambridge: Cambridge University Press.

Smith, V. 1936. *The Rationale of Central Banking*. London: P.S. King and Son.

Steele, H. and Yerbury, F. 1930. *The Old Bank of England*. London: Ernest Benn.

Stockdale, E. 1967. *The Bank of England in 1934*. London: Eastern Press.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. New York: Kelley, 1962.

Ziegler, D. 1990. *Central Bank, Peripheral Industry: The Bank of England in the Provinces 1826–1913*. London: Leicester University Press.

How to cite this article

Goodhart, Charles A. E. "Bank of England." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000188> doi:10.1057/9780230226203.0089

The New Palgrave Dictionary of Economics Online

banking crises

Charles W. Calomiris

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Banking crises take a variety of forms ranging from temporary liquidity crises to massive insolvencies. They sometimes coincide with other financial crises in currency and sovereign debt markets, and sometimes occur in isolation. These differences reflect the variety of causal influences that give rise to problems for banks. The unusually crisis-prone experience of the United States historically reflected its unique industrial organization of banking. Policies intended to reduce the incidence of banking crises (especially deposit insurance) have instead often increased the risk of crises, as safety-net protection reduces market discipline, allowing banks to undertake imprudent risks.

Keywords

banking crises; central banks; currency crises; deposit insurance; devaluation; Federal Reserve System; Great Depression; liquidity crises; panic of 1907; prudential bank regulation; sovereign debt

Article

There are two distinct phenomena associated with banking system distress: exogenous shocks that produce insolvency, and depositor withdrawals during ‘panics’. These two contributors to distress often do not coincide. For example, in the rural United States during the 1920s many banks failed, often with high losses to depositors, but those failures were not associated with systemic panics. In 1907, the United States experienced a systemic panic, originating in New York. Although some banks failed in 1907, failures and depositor losses were not much higher than in normal times. As the crisis worsened, banks suspended convertibility until uncertainty about the incidence of the shock had been resolved. The central differences between these two episodes relate to the commonality of information regarding the shocks producing loan losses. In the 1920s, the shocks were loan losses in agricultural banks, geographically isolated and fairly transparent. Banks failed without resulting in system-wide concerns. During 1907, the ultimate losses for New York banks were small, but the incidence was unclear *ex ante* (loan losses reflected complex connections to securities market transactions, with uncertain

consequences for some New York banks). This confusion hit the financial system at a time of low liquidity, reflecting prior unrelated disturbances in the balance of payments (Bruner and Carr, 2007).

Sometimes, large loan losses, and confusion regarding their incidence, occurred together. In Chicago in mid-1932 losses resulted in many failures and also in widespread withdrawals from banks that did not ultimately fail. Research has shown that the banks that failed were exogenously insolvent; solvent Chicago banks experiencing withdrawals did not fail. In other episodes, however, bank failures may reflect illiquidity resulting from runs, rather than exogenous insolvency.

Banking crises can differ according to whether they coincide with other financial events. Banking crises coinciding with currency collapse are called ‘twin’ crises (as in Argentina in 1890 and 2001, Mexico in 1995, and Thailand, Indonesia and Korea in 1997). A twin crisis can reflect two different chains of causation: an expected devaluation may encourage deposit withdrawal to convert to hard currency before devaluation (as in the United States in early 1933); or, a banking crisis can cause devaluation, either through its adverse effects on aggregate demand or by affecting the supply of money (when a costly bank bail-out prompts monetization of government bail-out costs). Sovereign debt crises can also contribute to bank distress when banks hold large amounts of government debt (for example, in the banking crises in the United States in 1861, and in Argentina in 2001).

The consensus views regarding banking crises’ origins (fundamental shocks versus confusion), the extent to which crises result from unwarranted runs on solvent banks, the social costs attending runs, and the appropriate policies to limit the costs of banking crises (government safety nets and prudential regulation) have changed dramatically, and more than once, over the course of the 19th and 20th centuries. Historical experience played a large role in changing perspectives toward crises, and the US experience had a disproportionate influence on thinking. Although panics were observed throughout world history (in Hellenistic Greece, and in Rome in ad 33), prior to the 1930s, in most of the world, banks were perceived as stable, large losses from failed banks were uncommon, banking panics were not seen as a great risk, and there was little perceived need for formal safety nets (for example, deposit insurance, or programmes to recapitalize banks). In many countries, ad hoc policies among banks, and sometimes including central banks, to coordinate bank responses to liquidity crises (as, for example, during the failure of Barings investment bank in London in 1890), seemed adequate for preventing systemic costs from bank instability.

Unusual historical instability of US banks

The unusual experience of the United States was a contributor to changes in thinking that led to growing concerns about banks runs, and the need for aggressive safety net policies to prevent or mitigate runs. In retrospect, the extent to which US banking instability informed thinking and policy outside the United States seems best explained by the size and pervasive influence of the United States; in fact, the US crises were unique and reflected peculiar features of US law and banking structure.

The US panic of 1907 (the last of a series of similar US events, including 1857, 1873, 1884, 1890, 1893, and 1896) precipitated the creation of the Federal Reserve System in 1913 as a means of enhancing systemic liquidity, reducing the probability of systemic depositor runs, and mitigating the costs of such events. This innovation was specific to the United States (other countries either had established central banks long before, often with other purposes in mind, or had not established central banks), and reflected the unique US experience with panics – a phenomenon that the rest of the world had not experienced

since 1866, the date of the last British banking panic (Bordo, 1985).

For example, Canada did not suffer panics like those of the United States and did not establish a central bank until 1935. Canada's early decision to permit branch banking throughout the country ensured that banks were geographically diversified and thus resilient to large sectoral shocks (like those to agriculture in the 1920s and 1930s), able to compete through the establishment of branches in rural areas (because of the low overhead costs of establishing additional branches), and able to coordinate the banking system's response in moments of confusion to avoid depositor runs (the number of banks was small, and assets were highly concentrated in several nationwide institutions). Outside the United States, coordination among banks facilitated systemic stability by allowing banks to manage incipient panic episodes to prevent widespread bank runs. In Canada, the Bank of Montreal would occasionally coordinate actions by the large Canadian banks to stop crises before the public was even aware of a possible threat.

The United States, however, was unable to mimic this behaviour on a national or regional scale (Calomiris, 2000; Calomiris and Schweikart, 1991). US law prohibited nationwide branching, and most states prohibited or limited within-state branching. US banks, in contrast to banks elsewhere, were numerous (for example, numbering more than 29,000 in 1920), undiversified, insulated from competition, and unable to coordinate their response to panics (US banks established clearing houses, which facilitated local responses to panics beginning in the 1850s, as emphasized by Gorton, 1985).

The structure of US banking explains why the United States uniquely had banking panics in which runs occurred despite the health of the vast majority of banks. The major US banking panics of the post-bellum era (listed above) all occurred at business cycle peaks, and were preceded by spikes in the liabilities of failed businesses and declines in stock prices; indeed, whenever a sufficient combination of stock price decline and rising liabilities of failed businesses occurred, a panic *always* resulted (Calomiris and Gorton, 1991). Owing to the US banking structure, panics were a predictable result of business cycle contractions that, in other countries, resulted in an orderly process of financial readjustment.

The United States, however, was not the only economy to experience occasional waves of bank failures before the First World War. Nor did it experience the highest bank failure rates, or bank failure losses. None of the US banking panics of the pre-First World War era saw nationwide banking distress (measured by the negative net worth of failed banks relative to annual GDP) greater than the 0.1 per cent loss of 1893. Losses were generally modest elsewhere, but Argentina in 1890 and Australia in 1893, where the most severe cases of banking distress occurred during this era, suffered losses of roughly ten per cent of GDP. Losses in Norway in 1900 were three per cent and in Italy in 1893 one per cent of GDP, but with the possible exception of Brazil (for which data do not exist to measure losses), there were no other cases in 1875–1913 in which banking loss exceeded one per cent of GDP.

Loss rates tended to be low because banks structured themselves to limit their risk of loss, by maintaining adequate equity-to-assets ratios, sufficiently low asset risk, and adequate asset liquidity. Market discipline (the fear that depositors would withdraw their funds) provided incentives for banks to behave prudently. The picture of small depositors lining up around the block to withdraw funds has received much attention, but perhaps the more important source of market discipline was the threat of an informed (often ‘silent’) run by large depositors (often other banks). Banks maintained relationships with each other through interbank deposits and the clearing of public deposits, notes and bankers’ bills. Banks often belonged to clearing houses that set regulations and monitored members’ behaviour. A bank

that lost the trust of its fellow bankers could not long survive.

Changing perceptions of banking instability

This perception of banks as stable, as disciplined by depositors and interbank arrangements to act prudently, and as unlikely to fail was common prior to the 1930s. The banking crises of the Great Depression changed this perception. US Bank failures resulted in losses to depositors in the 1930s in excess of three per cent of GDP. Bank runs, bank holidays (local and national government-decreed periods of bank closure to attempt to calm markets and depositors), and widespread bank closure suggested a chaotic and vulnerable system in need of reform. The Great Depression saw an unusual raft of banking regulations, especially in the United States, including restrictions on bank activities (the separation of commercial and investment banking, subsequently reversed in the 1980s and 1990s), targeted bank recapitalizations (the Reconstruction Finance Corporation), and limited government insurance of deposits.

Academic perspectives on the Depression fuelled the portrayal of banks as crisis-prone. The most important of these was the treatment of the 1930s banking crises by Milton Friedman and Anna Schwartz in their book, *A Monetary History of the United States* (1963). Friedman and Schwartz argued that many solvent banks were forced to close as the result of panics, and that fear spread from some bank failures to produce failures elsewhere. They saw the early failure of the Bank of United States in 1930 as a major cause of subsequent bank failures and monetary contraction. They lauded deposit insurance: ‘federal deposit insurance, to 1960 at least, has succeeded in achieving what had been a major objective of banking reform for at least a century, namely, the prevention of banking panics’. Their views that banks were inherently unstable, that irrational depositor runs could ruin a banking system, and that deposit insurance was a success, were particularly influential coming from economists known for their scepticism of government interventions.

Since the publication of *A Monetary History of the United States*, however, other scholarship (notably, the work of Elmus Wicker, 1996, and Charles Calomiris and Joseph Mason, 1997; 2003a) has led to important qualifications of the Friedman–Schwartz view of bank distress during the 1930s, and particularly of the role of panic in producing distress. Detailed studies of particular regions and banks’ experiences do not confirm the view that panics were a nationwide phenomenon during 1930 or early 1931, or an important contributor to nationwide distress until very late in the Depression (that is, early 1933). Regional bank distress was often localized and traceable to fundamental shocks to the values of bank loans. Not only does it appear that the failure of the Bank of United States had little effect on banks nationwide in 1930, one scholar has argued that there is evidence that the bank was, in fact, insolvent when it failed (Lucia, 1985).

Other recent research on banking distress during the pre-Depression era has also de-emphasized inherent instability, and focused on the historical peculiarity of the US banking structure and panic experience, noted above. Furthermore, recent research on the destabilizing effects of bank safety nets has been informed by the experience of the US Savings and Loan industry debacle of the 1980s, the banking collapses in Japan and Scandinavia during the 1990s, and similar banking system debacles occurring in 140 developing countries in the last quarter of the 20th century, all of which experienced banking system losses in excess of one per cent of GDP, and more than 20 of which experienced losses in excess of ten

per cent of GDP (data are from Caprio and Klingebiel, 1996, updated in private correspondence with these authors). Empirical studies of these unprecedented losses concluded that deposit insurance and other policies that protect banks from market discipline, intended as a cure for instability, have become instead the single greatest source of banking instability.

The theory behind the problem of destabilizing protection has been well known for over a century, and was the basis for US President Franklin Roosevelt's opposition to deposit insurance in 1933 (an opposition shared by many). Deposit insurance was seen as undesirable special interest legislation designed to benefit small banks. Numerous attempts to introduce it failed to attract support in Congress (Calomiris and White, 1994). Deposit insurance removes depositors' incentives to monitor and discipline banks, and frees bankers to take imprudent risks (especially when they have little or no remaining equity at stake, and see an advantage in 'resurrection risk taking'). The absence of discipline also promotes banker incompetence, which leads to unwitting risk taking.

Empirical research on late 20th-century banking collapses has produced a consensus that the greater the protection offered by a country's bank safety net, the greater the risk of a banking collapse (see, for example, Caprio and Klingebiel, 1996, and the papers from a 2000 World Bank conference on bank instability listed in the bibliography). Empirical research on prudential bank regulation emphasizes the importance of subjecting some bank liabilities to the risk of loss to promote discipline and limit risk taking (Shadow Financial Regulatory Committee, 2000; Mishkin, 2001; Barth, Caprio and Levine, 2006).

Studies of historical deposit insurance reinforce these conclusions (Calomiris, 1990). The basis for the opposition to deposit insurance in the 1930s was the disastrous experimentation with insurance in several US states during the early 20th century, which resulted in banking collapses in all the states that adopted insurance. Government protection had played a similarly destabilizing role in Argentina in the 1880s (leading to the 1890 collapse) and in Italy (leading to its 1893 crisis). In retrospect, the successful period of US deposit insurance, from 1933 to the 1960s, to which Friedman and Schwartz referred, was an aberration, reflecting limited insurance during those years (insurance limits were subsequently increased), and the unusual macroeconomic stability of the era.

Models of banking crises followed trends in the empirical literature. The understanding of bank contracting structures, in light of potential crises, has been a consistent theme. Banks predominantly hold illiquid assets ('opaque,' non-marketable loans), and finance those assets mainly with deposits withdrawable on demand. Banks are not subject to bankruptcy preference law, but rather, apply a first-come, first-served rule to failed bank depositors (depositors who are first in line keep the cash paid out to them). These attributes magnify incentives to run banks. An early theoretical contribution, by Douglas Diamond and Philip Dybvig (1983), posited a banking system susceptible to the constant threat of runs, with multiple equilibria, where runs can occur irrespective of problems in bank portfolios or any fundamental demand for liquidity by depositors. They modelled deposit insurance as a means of avoiding the bad (bank run) equilibrium. Over time, other models of banks and depositor behaviour developed different implications, emphasizing banks' abilities to manage risk effectively, and the beneficial incentives of demand deposits in motivating the monitoring of banks in the presence of illiquid bank loans (Calomiris and Kahn, 1991).

The literatures on banking crises also rediscovered an older line of thought emphasized by John Maynard Keynes (1931) and Irving Fisher (1933): market discipline implies links between increases in

bank risk, depositor withdrawals and macroeconomic decline. As banks respond to losses and increased risk by curtailing the supply of credit, they can aggravate the cyclical downturn, magnifying declines in investment, production, and asset prices, whether or not bank failures occur (Bernanke, 1983; Bernanke and Gertler, 1990; Calomiris and Mason, 2003b; Allen and Gale, 2004; Von Peter, 2004; Calomiris and Wilson, 2004). New research explores general equilibrium linkages among bank credit supply, asset prices and economic activity, and adverse macroeconomic consequences of ‘credit crunches’ that result from banks’ attempts to limit their risk of failure. This new generation of models provides a rational-expectations, ‘shock-and-propagation’ approach to understanding the contribution of financial crises to business cycles, offering an alternative to the endogenous-cycles, myopic-expectations view pioneered by Hyman Minsky (1975) and Charles Kindleberger (1978).

See Also

- credit rationing
- currency crises
- deposit insurance
- Great Depression
- moral hazard

Bibliography

Allen, F. and Gale, D. 2004. Financial fragility, liquidity, and asset prices. *Journal of the European Economic Association* 2, 1015–48.

Barth, J.R., Caprio, G. and Levine, R. 2006. *Rethinking Bank Regulation: Till Angels Govern*. Cambridge: Cambridge University Press.

Bernanke, B.S. 1983. Nonmonetary effects of the financial crisis in the propagation of the great depression. *American Economic Review* 73, 257–76.

Bernanke, B.S. and Gertler, M. 1990. Financial fragility and economic performance. *Quarterly Journal of Economics* 105, 87–114.

Bordo, M. 1985. The impact and international transmission of financial crises: some historical evidence, 1870–1933. *Revista di Storia Economica* 2(2d), 41–78.

Boyd, J., Gomis, P., Kwak, S. and Smith, B. 2000. A user's guide to banking crises. Conference paper. Washington, DC: World Bank.

Bruner, R.F. and Carr, S.D. 2007. Money Panic: Lessons from the Financial Crisis of 1907. New York:

Wiley.

Calomiris, C.W. 1990. Is deposit insurance necessary? A historical perspective. *Journal of Economic History* 50, 283–95.

Calomiris, C.W. 2000. *U.S. Bank Deregulation in Historical Perspective*. Cambridge: Cambridge University Press.

Calomiris, C.W. and Gorton, G. 1991. The origins of banking panics: models, facts, and bank regulation. In *Financial Markets and Financial Crises*, ed. R.G. Hubbard. Chicago: University of Chicago Press.

Calomiris, C.W. and Kahn, C.M. 1991. The role of demandable debt in structuring optimal banking arrangements. *American Economic Review* 81, 497–513.

Calomiris, C.W. and Mason, J.R. 1997. Contagion and bank failures during the great depression: the June 1932 Chicago banking panic. *American Economic Review* 87, 863–83.

Calomiris, C.W. and Mason, J.R. 2003a. Fundamentals, panics and bank distress during the depression. *American Economic Review* 93, 1615–47.

Calomiris, C.W. and Mason, J.R. 2003b. Consequences of bank distress during the great depression. *American Economic Review* 93, 937–47.

Calomiris, C.W. and Schweikart, L. 1991. The panic of 1857: origins, transmission, and containment. *Journal of Economic History* 51, 807–34.

Calomiris, C.W. and White, E.N. 1994. The origins of federal deposit insurance. In *The Regulated Economy: A Historical Approach to Political Economy*, ed. C. Goldin and G. Libecap. Chicago: University of Chicago Press.

Calomiris, C.W. and Wilson, B. 2004. Bank capital and portfolio management: the 1930s ‘capital crunch’ and scramble to shed risk. *Journal of Business* 77, 421–55.

Caprio, G. and Klingebiel, D. 1996. Bank insolvencies: cross country experience. Working Paper No. 1620. Washington, DC: World Bank.

Cull, R., Senbet, L. and Sorge, M. 2000. Deposit insurance and financial development. Conference paper. Washington, DC: World Bank.

Demirguc-Kunt, A. and Detragiache, E. 2000. Does deposit insurance increase banking system stability? Conference paper. Washington, DC: World Bank.

Demirguc-Kunt, A. and Huizinga, H. 2000. Market discipline and financial safety net design. Conference paper. Washington, DC: World Bank.

Diamond, D. and Dybvig, P. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91, 401–19.

Fisher, I. 1933. The debt deflation theory of great depressions. *Econometrica* 1, 337–57.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press.

Gorton, G. 1985. Clearing houses and the origin of central banking in the United States. *Journal of Economic History* 45, 277–83.

Honohan, P. and Klingebiel, D. 2000. Controlling fiscal costs of banking crises. Conference paper. Washington, DC: World Bank.

Keynes, J.M. 1931. The consequences to the banks of the collapse of money values. In *Essays in Persuasion*. New York: W.W. Norton, 1963.

Kindleberger, C.P. 1978. *Manias, Panics, and Crashes: A History of Financial Crises*. New York: Basic Books.

Lucia, J.L. 1985. The failure of the bank of United States: a reappraisal. *Explorations in Economic History* 22, 402–16.

Minsky, H.P. 1975. *John Maynard Keynes*. New York: Columbia University Press.

Mishkin, F.S. 2001. *Prudential Supervision: What Works and What Doesn't*. Chicago: University of Chicago Press.

Shadow Financial Regulatory Committee. 2000. *Reforming Bank Capital Regulation*. Washington, DC: American Enterprise Institute.

Von Peter, G. 2004. Asset prices and banking distress: a macroeconomic approach. BIS Working Paper No. 167. Basel: Bank of International Settlements.

Wicker, E. 1996. *The Banking Panics of the Great Depression*. Cambridge: Cambridge University Press.

How to cite this article

Calomiris, Charles W. "banking crises." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000051> doi:10.1057/9780230226203.0090

The New Palgrave Dictionary of Economics Online

banking industry

Dario Focarelli and Alberto Franco Pozzolo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adverse selection; assets and liabilities; asymmetric information; bank deregulation; banking crises; banking industry; bankruptcy; barriers to entry; capital controls; credit; excessive risk taking; Glass–Stegall Act 1933 (USA); information costs; information monopolies; liquidity; mergers and acquisitions; moral hazard; relationship lending; technological innovation

Article

The distinctive function of banks is the transformation of short-term deposits into longer-term, less liquid and riskier loans (Fama, 1980; 1985; Diamond and Rajan, 2001; Gorton and Winton, 2003). By raising funds from depositors and providing credit, banks avoid the duplication of monitoring, which reduces the overall cost of transferring funds from capital suppliers to its users (Leland and Pyle, 1977; Diamond, 1984). At the same time, however, the greater liquidity of liabilities than of assets, which are typically longer-term and riskier, makes bank balance sheets vulnerable. Not only may banks fail if they are unable to obtain repayment of their loans, but depositors might even decide to withdraw their assets simply anticipating that others will do so. Such a ‘bank run’ can drive an otherwise sound bank to insolvency (Diamond and Dybvig, 1983). The need to protect depositors and so guarantee a stable monetary transaction system explains why the banking industry is so heavily regulated. It is harder for a depositor to protect his interests than for an average investor, because judging the financial condition of a bank is difficult and costly, even for specialists. For this reason, the typical instruments adopted by bank regulators include restrictions on the amount of risk that a bank can take, and compulsory deposit insurance schemes that prevent runs.

Regulatory intervention affects the shape of the banking industry and its degree of competition. Until the mid-1960s, governments deliberately limited competition in the interest of ‘safety and soundness’ by regulating deposit rates, entry, branching and mergers. The traditional view is of a trade-off between soundness and competition, with more intense competition reducing franchise values and increasing incentives to take on risky projects, since forgone future profits in the case of bankruptcy are lower

(Keeley, 1990). By increasing the equity at risk, capital controls reduce (although perhaps not entirely) excessive risk-taking (Hellman, Murdock and Stiglitz, 2000).

Recently, a more comprehensive view has been put forward, suggesting that regulation interacts dynamically with pervasive information asymmetries, and that the relationship between competition and stability is accordingly complex and multifaceted (Allen and Gale, 2003). The cost of acquiring information in order to mitigate moral hazard and adverse selection is a strong endogenous barrier to the entry of new banks, allowing incumbents to gain monopoly rents (Broecker, 1990), making competitive equilibria unsustainable (Dell'Ariccia, 2001; Dell'Ariccia, Friedman and Marquez, 1999), and forcing new entrants to take a higher-risk clientele (Shaffer, 1998).

The problems of information asymmetries can be attenuated if a bank deals repeatedly with the same customer, a practice known as ‘relationship lending’. However, as Sharpe (1990) and Rajan (1992) show, this gives relationship banks a monopoly on information about their borrowers, further reducing competition, especially in the short run (Petersen and Rajan, 1995). In this case, deregulation aimed at fostering inter-bank competition in transaction lending could have the effect of augmenting the scope for relationship banking, which permits banks to retain some monopoly power. As Boot and Thakor (2000) show, this is not the case if stronger competition comes from capital market financing, which drives some banks out of the market, reducing competition and consequently relationship lending.

Since the mid-1980s, the banking industry has been transformed by a series of events: deregulation of deposit accounts, which forced US banks to compete on interest rates; branching liberalization, which led to a sharp decline in the number of banks; the changes in capital requirements introduced with the Basel accords of 1988, which pushed banks towards newer and less regulated off-balance-sheet activities; the introduction of the euro, which created a unique wholesale banking market within Europe (Berger, Kashyap and Scalise, 1995); and the substantial repeal of the Glass–Steagall Act of 1933, allowing banks to supply financial services previously offered only by other intermediaries, such as investment firms and insurance companies.

One of most important consequences of deregulation has been the unprecedented numbers of mergers and acquisitions during the 1990s, which sharply reduced the number of banks in many industrial countries and often heightened concern over possible anti-competitive effects. However, there is no clear evidence that the consolidations have harmed consumers or diminished competition, as would have been predicted from the observed negative correlation between the degree of concentration in local banking markets and the level of deposit rates (Berger and Hannan, 1989). Rather, the available evidence indicates a positive effect stemming from the larger and more efficient banks taking over the smaller and less efficient (Berger, Kashyap and Scalise, 1995; Focarelli, Panetta and Salleo, 2002). And while there may be some contraction of credit to smaller clients due to consolidation, this effect appears to be largely offset by increased lending by other banks (Berger et al., 1998). Indeed, there is evidence that in the medium term mergers increase the efficiency of the target bank, benefiting depositors (Focarelli and Panetta, 2003).

The future of the banking industry is likely be determined by the interaction of three major forces: international competition, innovation in information technology and regulation. At present, all three factors are heightening competition in banking. International competition, while still limited, tends to display the same pattern as domestic consolidation, with larger and more efficient banks in more

developed countries taking over less efficient banks in financially less developed areas (Focarelli and Pozzolo, 2005). Technological innovation is lessening the importance of close lending relationships, enlarging the size of local credit markets and further reducing the role of small banks (Petersen and Rajan, 2002). Worldwide regulatory systems are moving to allow more competition and to assign a more important role to market evaluation (Basel Committee on Banking Supervision, 2005).

See Also

- agency problems
- banking crises
- financial intermediation
- market structure
- merger analysis (United States)
- micro-credit
- payment systems

Bibliography

Allen, F. and Gale, D. 2003. Competition and financial stability. *Journal of Money, Credit, and Banking* 36, 433–80.

Basel Committee on Banking Supervision. 2005. *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Basel: BIS.

Berger, A. and Hannan, T. 1989. The price–concentration relationship in banking. *Review of Economics and Statistics* 71, 291–9.

Berger, A., Kashyap, A. and Scalise, J. 1995. The transformation of the US banking industry: what a long trip it's been. *Brookings Papers on Economic Activity* 1995(2), 55–201.

Berger, A., Saunders, A., Scalise, J. and Udell, G. 1998. The effects of bank mergers and acquisitions on small business lending. *Journal of Financial Economics* 50, 187–229.

Boot, A. and Thakor, A. 2000. Can relationship banking survive competition? *Journal of Finance* 55, 679–713.

Broecker, T. 1990. Credit-worthiness tests and interbank competition. *Econometrica* 58, 429–52.

Dell'Ariccia, G. 2001. Asymmetric information and the structure of the banking industry. *European Economic Review* 45, 1957–80.

Dell'Ariccia, G., Friedman, E. and Marquez, R. 1999. Adverse selection as a barrier to entry in the banking industry. *RAND Journal of Economics* 30, 515–34.

Diamond, D. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51, 393–414.

Diamond, D. and Dybvig, P. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91, 401–19.

Diamond, D. and Rajan, R. 2001. Liquidity risk, liquidity creation and financial fragility: a theory of banking. *Journal of Political Economy* 109, 287–327.

Fama, E. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6, 39–57.

Fama, E. 1985. What's different about banks? *Journal of Monetary Economics* 15, 29–34.

Focarelli, D. and Panetta, F. 2003. Are mergers beneficial to consumers? Evidence from the market for bank deposits. *American Economic Review* 93, 1152–72.

Focarelli, D., Panetta, F. and Salleo, C. 2002. Why do banks merge? *Journal of Money, Credit, and Banking* 34, 784–803.

Focarelli, D. and Pozzolo, A. 2005. Where do banks expand abroad? An empirical analysis. *Journal of Business* 78, 2435–64.

Gorton, G. and Winton, A. 2003. Financial intermediation. In *Handbook of the Economics of Finance*, vol. 1, ed. G. Constantinides, M. Harris and R. Stulz. Amsterdam: North-Holland.

Hellman, T., Murdock, K. and Stiglitz, J. 2000. Liberalization, moral hazard in banking and prudential regulation: are capital requirements enough? *American Economic Review* 90, 147–65.

Leland, H. and Pyle, D. 1977. Informational asymmetries, financial structure and financial intermediation. *Journal of Finance* 32, 371–87.

Keeley, M. 1990. Deposit insurance, risk, and market power in banking. *American Economic Review* 80, 1183–200.

Petersen, M. and Rajan, R. 1995. The effect of credit market competition on lending relationships. *Quarterly Journal of Economics* 110, 407–43.

Petersen, M. and Rajan, R. 2002. Does distance still matter? The information revolution in small business lending. *Journal of Finance* 57, 2533–70.

Rajan, R. 1992. Insiders and outsiders: the choice between relationship and arms length debt. *Journal of Finance* 47, 1367–400.

Shaffer, S. 1998. The winner's curse in banking. *Journal of Financial Intermediation* 7, 359–92.

Sharpe, S. 1990. Asymmetric information, bank lending and implicit contracts: a stylized model of customer relationships. *Journal of Finance* 45, 1069–87.

How to cite this article

Focarelli, Dario and Alberto Franco Pozzolo. "banking industry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000309>
doi:10.1057/9780230226203.0091

The New Palgrave Dictionary of Economics Online

Banking School, Currency School, Free Banking School

Anna J. Schwartz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The doctrines of the three 19th century schools differed. The Currency School believed that note issues should vary one-to-one with the Bank of England's gold reserves. The Banking School believed that real bills, needs of trade and the law of reflux should govern bank operations. The Free Banking School believed that competitive private banks would not overissue, whereas a monopoly issuer did so. Other issues were debated. Was a central bank needed? Should a central bank be subject to rules or allowed discretion? How should money be defined? No one point of view carried the day and several of the issues that divided the schools are still debated today.

Keywords

balance of payments; Bank Charter Act 1833 (UK); Bank Charter Act 1844 (UK); Bank of England; Bank of Ireland; Bank of Scotland; Banking School; bullion reserve; central banking; convertibility; country banks; credit; currency principle; Currency School; Free Banking School; Fullarton, J.; Gilbart, J.; gold standard; inflation; joint stock banking; law of reflux; Longfield, M.; McCulloch, J. R.; Mill, J. S.; money supply; money, definition of; monopoly of note issue; needs of trade doctrine; Norman, G.; note issue; overissue; Overstone, Lord; Parnell, H.; private banks; real bill doctrine; reserve–deposit ratio; rules versus discretion; Schwartz, A. J.; Scrope, G. P.; stocks and flows; Tooke, T.; Torrens, R.; trade cycles; Wilson, J.

Article

Historians of economic thought conventionally represent British monetary debates from the 1820s on as centred on the question of whether policy should be governed by rules (espoused by adherents of the Currency School), or whether authorities should be allowed discretion (espoused by adherents of the Banking School). In fact many other questions were in dispute, including those raised by neglected or misidentified participants in the debates – adherents of the Free Banking School.

Among the questions in dispute were the following: (1) Should the banking system follow the Currency School's principle that note issues should vary one-to-one with the Bank of England's gold holdings? (2) Were the doctrines of the Banking School – real bills, needs of trade and the law of reflux – valid? (3) Was a monopoly of note issue desirable or, as the Free Banking School contended, destabilizing? (4) Was overissue a problem and, if so, who was responsible? (5) How should money be defined? (6) Why do trade cycles occur? (7) Should there be a central bank? No, was the Free Banking School answer to the final question; yes, was the answer of the other two schools, with disparate views, as indicated, on the question of rules vs. authorities. What was not in dispute was the viability of the gold standard system with gold convertibility of Bank of England notes.

On what grounds did the schools oppose each other? Each of the first three questions identifies the central doctrines that the adherents of one of the schools shared; on the remaining questions, individual views within each school varied. Before establishing the positions of each school in the monetary debates, we introduce the institutional background and the principal participants.

Institutional background

The Bank of England, incorporated in 1694 as a private institution with special privileges, stood at the head of the British banking system at the time of the debates. Until 1826 the Bank's charter was interpreted to mean the prohibition of other joint stock banks in England. As a result banking establishments were either one-man firms or partnerships with not more than six members. Two types of banks predominated in England: the wealthy London private banks which had voluntarily surrendered their note-issuing privilege, and the country banks which depended almost exclusively on the business of note issues. Numerous failures among the country banks demonstrated that the effect of the Bank's charter was to foster the formation of banking units of uneconomical size.

Banking in Ireland was patterned on English lines. The Bank of Ireland, chartered in 1783 with the exclusive privilege of joint stock banking in Ireland, surrendered its monopoly in 1821 in places farther than fifty miles from Dublin. Joint-stock banking in the whole of Ireland was legalized in 1845.

The Bank of Scotland was founded in 1695 with privileges similar to those of the Bank of England, except that it was formed to promote trade, not to support the credit of the government. It lost its monopoly in 1716, and no further monopolistic banking legislation was enacted in Scotland. With free entry possible, many local private and joint stock banks, most of the latter well capitalized, were established, and a nationwide system of branch banking developed. Unlike the English system, overissue was not a problem in the Scottish system. The banks accepted each other's notes and evolved a system of note exchange. Shareholders of Scottish joint stock banks (except for three chartered banks) assumed unlimited liability. At the time of the debates banking in Scotland was at a far more advanced stage than in England.

Principals in the debates

The leading spokesmen for the Currency School side in the debates were McCulloch, Loyd (later Lord Overstone), Longfield, George Warde Norman, and Torrens. Norman, a director of the Bank of England for most of the years 1821–72, and of the Sun Insurance Company, 1830–64, was active in the timber trade with Norway. The principal Banking School representatives were Tooke, Fullarton, and John

Stuart Mill, while James Wilson held views that straddled Banking and Free Banking School doctrines. The most prominent members of the Free Banking School were Parnell (later Baron Congleton), James William Gilbart, and Poulett Scrope. Gilbart, a banker, was general manager of the London and Westminster Bank, the first of the joint stock banks authorized by the Bank Charter Act of 1833.

Currency School principle

The objective of the Currency School was to achieve a price level that would be the same whether the money supply were fully metallic or a mixed currency including both paper notes and metallic currency. According to Loyd, gold inflows or outflows under a fully metallic currency had the immediate effect of increasing or decreasing the currency in circulation, whereas a mixed currency could operate properly only if inflows or outflows of gold were exactly matched by an increase or decrease of the paper component. He and others of the Currency School regarded a rise in the price level and a fall in the bullion reserve under a mixed currency as symptoms of excessive note issues. They advocated statutory regulation to ensure that paper money was neither excessive nor deficient because otherwise fluctuations in the currency would exacerbate cyclical tendencies in the economy. They saw no need, however, to regulate banking activities other than note issue.

The Banking School challenged these propositions. Fullarton denied that overissue was possible in the absence of demand, that variations in the note issue could cause changes in the domestic price level, or that such changes could cause a fall in the bullion reserve ([1844] 1969, pp. 57, 128–9). Under a fully metallic as well as under a mixed currency bank, deposits, bills of exchange, and all forms of credit might influence prices. Moreover, inflows and outflows of gold under a fully metallic currency might change bullion reserves but not prices. If convertibility were maintained, overissue was not feasible and no statutory control of note issues was required. An adverse balance of payments was a temporary phenomenon that was self-correcting when, for example, a good harvest followed a bad one. According to the Free Banking School, the possibility of overissue and inflation applied only to Bank of England notes but could not occur in a competitive banking system.

Banking School principle

The Banking School adopted three principles that for them reflected the way banks actually operated as opposed to the Currency School principle which they dismissed as an artificial construct of certain writers (White, 1984, pp. 119–28).

The first Banking School principle was the doctrine that liabilities of deposits and notes would never be excessive if banks restricted their earning assets to real bills. One charge levelled by modern economists against the doctrine is that it leaves the quantity of money and the price level indeterminate, since it links the money supply to the nominal magnitude of bills offered for discount. Some members of the school may be exculpated from this charge if they regarded England as a small open economy, its domestic money stock a dependent variable determined by external influences. However, because it ignored the role of the discount rate in determining the volume of bills generated in trade, the doctrine was vulnerable. In addition, the Banking School confused the flow demand for loanable funds, represented by the volume of bills, with the stock demand for circulating notes, although the two magnitudes are non-

commensurable.

Free Banking School members who also adopted the real bills doctrine erroneously attributed overissue by the Bank of England to its purchase of assets other than real bills, when overissue was possible with a portfolio limited to real bills, acquired at an interest rate that led to a stock of circulating medium inconsistent with the prevailing price level (Gilbart, 1841, pp. 103–5; 119–20). The Currency School regarded the real bills doctrine as misguided since it could promote a cumulative rise in the note issue and hence in prices.

A second Banking School principle was the ‘needs of trade’ doctrine, to the effect that the note circulation should be demand-determined – curtailed when business declined and expanded when business prospered, whether for seasonal or cyclical reasons. An implicit assumption of the doctrine was that banks could either vary their reserve ratios to accommodate lower or higher note liabilities, or else offset changes in note liabilities by opposite changes in deposit liabilities. For non-seasonal increases in demand for notes, the doctrine implied that expanding banks could obtain increased reserves from an interregional surplus of the trade balance. The Currency School regarded an increase in the needs of trade demand to hold notes accompanying increases in output and prices as unsound because it would ultimately produce an external drain. The Free Banking School countered that such an objection by the Currency School was paradoxical since the virtue of a metallic currency according to the latter was that it accommodated the commercial wants of the country, and therefore for a mixed currency to respond to the needs of trade could not be a vice. The modern objection to the needs of trade doctrine as procyclical is an echo of the Currency School view.

The third Banking School principle was the law of the reflux according to which overissue was possible only for limited periods because notes would immediately return to the issuer for repayment of loans. This was a modification of the real bills doctrine that Tooke and Fullarton advanced, since adherence to the doctrine supposedly made overissue impossible. They made no distinction between the speed of the reflux for the Bank of England and for competitive banks of issue – a distinction at the heart of the Free Banking position. For the latter, reflux of excess notes was speedy only if the notes were deposited in rival banks. These would then return the notes to the issuing banks and accordingly bring an end to relative overissue by individual banks. The Bank of England, on the contrary, could overissue for long periods because it had no rivals. Fullarton, however, made the unwarranted assumption that notes would be returned to the Bank to repay previous loans at a faster rate than the Bank was discounting new loans, hence correcting the overissue. Moreover, he believed that if the Bank overissued by open market purchases, the decline in interest rates would quickly activate capital outflows, reducing the Bank's bullion and forcing it to retreat. Tooke was sounder in arguing for the law of reflux on the ground that excess issues would not be held if they did not match the preferences of holders for notes rather than deposits.

The Banking School had no legislative programme for reform of the monetary system. Good bank management, in the view of the school, could not be legislated.

Free Banking School principle

As the name suggests, the principle the Free Banking School advocated was free trade in the issue of currency convertible into specie. Members of the school favoured a system like the Scottish banking system, where banks competed in all banking services, including the issue of notes, and no central bank

held a monopoly of note issue. They argued that in such a system banks did not issue without limit but indeed provided a stable quantity of money, Although the costs of printing and issuing were minimal, to keep notes in circulation required restraint in their issue. The profit-maximizing course for competitive banks was to maintain public confidence in their issues by maintaining convertibility into specie on demand, which required limiting their quantity.

Loyd's response to the argument for free trade in currency was that unlike ordinary trades, what was sought was not the greatest quantity at the cheapest price but a regulated quantity of currency. The Free Banking School denied that free banking would debase the currency, and contended that the separation of banking from note issue, the Banking School proposal, was impractical. Scrope (1833a, pp. 32–3) asked why the Currency School objected to unregulated issue of notes but not to that of deposits, questioning Loyd's assumption that an issuing bank's function was to produce money, when in fact its function was to substitute its bank notes for less well-known private bills of exchange that were the bank's assets. Scrope and other Free Banking adherents (Parnell, 1827, p. 143) neglected the distinction between a banknote immediately convertible into gold and a commercial bill whose present value varied with time to maturity and the discount rate. Contrary to Loyd, they reasoned that free trade and competition were applicable to currency creation because the business of banks was to produce the scarce good of reputation.

Loyd's second disagreement with the argument for free trade in banking was that miscalculations by the issuers were borne not by them but by the public. Moreover, individuals had no choice but to accept notes they received in ordinary transactions, and trade in general suffered as a result of overissue. The Free Banking School answer to this externalities argument turned on the ability of holders to refuse notes of issuers without reputation. Protection against loss could also be provided if joint stock banks were allowed to operate in place of country banks limited to six or fewer partners. In addition, if banks were required to deposit security of government bonds or other assets, noteholders would be further protected (Scrope, 1832, p. 455; 1833b, p. 424; Parnell, 1827, pp. 140–4). Free Banking School members who argued in this vein failed to recognize that they were thereby acknowledging a role for government intervention in currency matters.

In the 1820s the Free Banking School championed joint stock banking both in the country bank industry and in direct competition in note issue with the Bank of England in London. Although the six-partner rule for banks of issue at least 65 miles from London was repealed in 1826 after a spate of bank failures, the Bank retained its monopoly of note circulation in the London area. In addition, the Bank was permitted to establish branches anywhere in England. The Parliamentary inquiry in 1832 on renewal of the Bank's character was directed to the question of prolonging the monopoly. The Act of 1833 eased entry for joint stock banks within the 65-mile limit but denied them the right of issue and made the Bank's notes legal tender for redemption of country bank notes, in effect securing the Bank's monopoly. The doom of the Free Banking cause was finally pronounced by the Bank Charter Act of 1844. It restricted note issues of existing private and joint stock banks in England and Wales to their average circulation during a period in 1843. Note issue by banks established after the Act was prohibited.

Was overissue a problem?

Participants in the debates understood overissue to mean a stock of notes, whether introduced by a single issuer or banks in aggregate, in excess of the quantity holders voluntarily chose to keep as assets, given

the level of prices determined by the world gold standard. Was overissue of a convertible currency possible? According to the Free Banking School, interbank note clearing by competitive banks operated to eliminate excess issued by a single bank. The check to excess issues by the banking system as a whole was an external drain through the price-specie flow mechanism. In this respect the school acknowledged that the result of overissue by a competitive banking system as a whole was the same as for a monopoly issuer. However, they held that overissue was a phenomenon that the monopoly of the Bank of England encouraged but a competitive system would discourage.

The Currency School, on the other hand, regarded both the Bank of England and the Scottish and country banks as equally prone to overissue and did not grant that a check to overissue by a single bank or banks in the aggregate was possible through the interbank note clearing mechanism. For them, regulation of a monopoly issuer promised a stable money supply that was not attainable with a plural banking system.

The Free Banking School's explanation of the Bank of England's ability to overissue rested on the absence of rivals for the Bank's London circulation, so no interbank note clearing took place; the absence of competition in London from interest-bearing demand deposits; and the fact that London private banks held the Bank's notes as reserves. Hence the demand for its notes was elastic. The Free Banking and Currency Schools agreed that there was a substantial delay before an external drain checked overissue, so the Bank's actions inescapably inflicted damage on the economy. Scrope (1830, pp. 57–60), who attributed the Bank's willingness to overexpand its note issues to its monopoly position, advocated abrogating that legal status.

The Banking School dismissed the question of overissue as irrelevant, for noteholders could easily exchange unwanted notes by depositing them. What they failed to examine was the possibility that a broader monetary aggregate could be in excess supply resulting in an external drain.

How should money be defined?

Currency School members favoured defining money as the sum of metallic money, government paper money, and bank notes (Norman, 1833, pp. 23, 50; McCulloch, 1850, pp. 146–7). The Free Banking School, like the Currency School, focused on bank notes as the common medium of exchange, ignoring demand deposits that were not usually subject to transfer by check outside London. The Banking School definition of money is sometimes represented as broader than that of the other schools, but in fact was narrower – money was restricted to metallic and government paper money. Bank notes and deposits were excluded, since they were regarded as means of raising the velocity of bank vault cash but not as adding to the quantity of money (Tooke [1848] 1928, pp. 171–83; Fullarton [1844] 1969, pp. 29–36; Mill [1848] 1909, p. 523). In the short run, the school held that all forms of credit might influence prices, but only money as defined could do so in the long run, because the domestic price level could deviate only temporarily from the world level of prices determined by the gold standard.

Why do trade cycles occur?

The positions of the three schools on the impulses initiating trade cycles were not dogma for their members. In general the Currency and Banking Schools held that nonmonetary causes produced trade

cycles, whereas the Free Banking School pointed to monetary causes, but individual members did not invariably hew to these analytical lines. McCulloch (1837, p. 63), Loyd (1857, p. 317), and Longfield (1840, pp. 222–3) essentially attributed cycles to waves of optimism and pessimism to which the banks then responded by expanding and contracting their issues. Banks accordingly never initiated the sequence of expansion and contraction. Hence the Currency School principle of regulating the currency to stabilize prices and business did not imply that cycles would thereby be eliminated. Cycles would, however, no longer be amplified by monetary expansion and contraction, if country banks were denied the right to issue and the Bank of England's circulation were governed by the ‘currency principle’. Torrens (1840, pp. 31, 42–3), unlike other Currency School members, attributed trade cycles to actions of the Bank of England. That was also the position of the Free Banking School, although in an early work Parnell (1827, pp. 48–51) of that school held that cycles were caused by nonmonetary factors. For the Banking School, however, nonmonetary factors accounted for both the origin and spread of trade cycles. Tooke (1840, pp. 245, 277), for example, believed that overoptimism would prompt an expansion of trade credit for which the banks were in no way responsible. Collapse of optimism would then lead to shrinkage of trade credit. For Fullarton ([1844] 1969, p. 101) nonmonetary causes produced price fluctuations to which changes in note circulation were a passive response. Proponents of the nonmonetary theory of the onset of trade cycles provided no explanation of the waves of optimism and pessimism themselves. For the Free Banking School the waves were precipitated by the Bank of England's expansion and ultimate contraction of its liabilities. Initially, the Bank's actions depressed interest rates and ultimately forced them up, as loanable funds increased in supply and then decreased. The Bank's monopoly position enabled it to create such monetary disturbances, whereas competitive country banks had no such power.

Should there be a central bank?

The Currency and Banking Schools were in agreement that a central bank with the sole right of issue was essential for the health of the economy. McCulloch (1831, p. 49) regarded a system of competitive note issuing institutions as one of inherent instability. Tooke (1840, pp. 202–7) favoured a monopoly issuer as promoting less risk of overissue and greater safety because it would hold sufficient reserves. The two schools differed on the need for a rule to regulate note issues, the Currency School pledged to a rulebound authority, the Banking School to an unbound authority. The Free Banking School disapproved of both a rule and a central bank authority, instead favouring a competitive note-issuing system that it held to be self-regulating. For that school proof that centralized power was inferior to a competitive system was revealed by cyclical fluctuations that had been caused by errors of the Bank of England.

A continuing debate

The Bank Charter Act of 1844 ended the right of note issue for new banks in England and Wales. Scottish banks, however, were treated differently from Irish banks by the Act of 1845 and from English provincial banks by the Act of 1844. Like the latter, authorized circulation for the Scottish banks was determined by the average of a base period, but they could exceed the authorized circulation provided they held 100 per cent specie reserves against the excess – a provision also imposed on the Bank of

England.

The Free Banking School thus lost its case for an end of the note issue monopoly of the Bank of England. The death of Parnell in 1842, a leading Parliamentary spokesman, had hurt the cause. Others of the school were mainly country and joint stock bankers. The Acts conferred benefits on them by restricting entry into the note-issuing industry and by freezing market shares (White, 1984, pp. 78–9). Their voices were not raised in opposition. Only Wilson was critical of the privileges the Bank of England was accorded ([1847] 1859, pp. 34–66).

The Banking School objected not only to the Act but claimed vindication for its point of view by the necessity to suspend it in 1847, 1857 and 1866. The Currency School responded that the suspensions were of no great significance (Loyd, 1848, pp. 393–4). The recommendations of the Currency School prevailed to set a maximum for country bank note issues and the eventual transfer of their circulation to the Bank of England.

The monetary debates that were initiated in the 1820s were not conclusive. No point of view carried the day. Long after the original participants had passed from the scene, the doctrines of the schools found supporters. Even the Free Banking School position in opposition to monopoly issue of hand-to-hand currency that seemed to be buried has recently been revived by new adherents (White, 1984, pp. 137–50). The debate on all the questions in dispute in the 19th century continues to be live.

See Also

- Boyd, Walter
- bullionist controversies (empirical evidence)
- Fullarton, John
- money, classical theory of
- Overstone, Lord [Samuel Jones Loyd]
- real bills doctrine
- Tooke, Thomas

Bibliography

Fullarton, J. 1844. *On the Regulation of Currencies*. London: John Murray. Reprinted, New York: Augustus M. Kelley, 1969.

Gilbart, J.W. 1841. Testimony before the Select Committee of the House of Commons on Banks of Issue. British Sessional Papers, vol. 5 (410).

Gregory, T.E. 1928. *Introduction to Tooke and Newmarch's A History of Prices*. London: P.S. King.

Longfield, S.M. 1840. Banking and currency. *Dublin University Magazine*.

Loyd, S.J. 1848. Testimony before the Secret Committee of the House of Commons on Commercial Distress. British Sessional Papers, 1847–8, vol. 8, part 1 (584).

Loyd, S.J. 1857. *Tracts and Other Publications on Metallic and Paper Money*. London.

McCulloch, J.R. 1831. *Historical Sketch of the Bank of England*. London: Longman.

McCulloch, J.R. 1837. The Bank of England and the country banks. *Edinburgh Review*, April.

McCulloch, J.R. 1850. *Essays on Interest, Exchange, Coins, Paper Money, and Banks*. London.

Mill, J.S. 1848. *Principles of Political Economy*, ed. W.J. Ashley. Reprinted London: Longmans & Co., 1909.

Norman, G.W. 1833. *Remarks upon Some Prevalent Errors, with Respect to Currency and Banking*. London: Hunter.

Parnell, H.B. 1827. *Observations on Paper Money, Banking and Overtrading*. London: James Ridgway.

Scrope, G.P. 1830. *On Credit-Currency, and its Superiority to Coin, in Support of a Petition for the Establishment of a Cheap, Safe, and Sufficient Circulating Medium*. London: John Murray.

Scrope, G.P. 1832. The rights of industry and the banking system. *Quarterly Review*, July, 407–55.

Scrope, G.P. 1833a. *An Examination of the Bank Charter Question*. London: John Murray.

Scrope, G.P. 1833b. *Principles of Political Economy*. London: Longman.

Tooke, T. 1840. *A History of Prices and of the State of the Circulation in 1838 and 1839*. London: Longman. Reprinted, London: P.S. King, 1928.

Tooke, T. 1848. *History of Prices and of the State of the Circulation, from 1839 to 1847 inclusive*. London: Longmans. Reprinted, London: P.S. King, 1928.

Torrens, R. 1840. *A Letter to Thomas Tooke, Esq. in Reply to His Objections against the Separation of the Business of the Bank into a Department of Issue and a Department of Discount: With a Plan of Bank Reform*. London: Longman.

White, L.H. 1984. *Free Banking in Britain: Theory, Experience, and Debate, 1800–1845*. Cambridge: Cambridge University Press.

Wilson, J. 1847. *Capital, Currency, and Banking; being a collection of a series of articles published in the Economist in 1845 … and in 1847*. London: The office of the Economist. 2nd edn., London: D.M. Aird, 1859.

How to cite this article

Schwartz, Anna J. "Banking School, Currency School, Free Banking School." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000057> doi:10.1057/9780230226203.0092

The New Palgrave Dictionary of Economics Online

bankruptcy law, economics of corporate and personal

Michelle J. White

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bankruptcy is the legal process whereby financially distressed firms, individuals, and occasionally governments resolve their debts. The bankruptcy process for firms plays a central role in economics, because competition tends to drive inefficient firms out of business, thereby raising the average efficiency level of those remaining. Bankruptcy also has an important economic function for individual debtors, since it provides them with partial consumption insurance and supplements the government-provided safety net. This article discusses the economic objectives of bankruptcy and surveys theoretical and empirical research on corporate and personal bankruptcy.

Keywords

absolute priority rule (APR); auctions; bankruptcy; bankruptcy contracting; bankruptcy law, economics of corporate and personal; bankruptcy, economics of; consumption insurance; equity finance; fresh start; limited liability; liquidation; options; reorganization; risk and return; strategic default

Article

Bankruptcy is the legal process whereby financially distressed firms, individuals, and occasionally governments resolve their debts. The bankruptcy process for firms plays a central role in economics, because competition tends to drive inefficient firms out of business, thereby raising the average efficiency level of those remaining. Consumers benefit because the remaining firms produce goods and services at lower costs and sell them at lower prices. The legal mechanism through which most firms exit the market is bankruptcy. Bankruptcy also has an important economic function for individual debtors, since it provides them with partial consumption insurance and supplements the government-provided safety net. Local governments occasionally also use bankruptcy to resolve their debts, and there has been discussion of establishing a bankruptcy procedure for financially distressed countries (see White, 2002).

Bankruptcy law

For both corporate and individual debtors, bankruptcy law provides a collective framework for simultaneously resolving all debts when debtors' assets are less valuable than their liabilities. This includes both rules for determining which of the debtor's assets must be used to repay debt and rules for dividing the assets among creditors. Thus bankruptcy is concerned with both the size of the pie – the total amount paid to creditors – and how the pie is divided.

For financially distressed corporations, both the size and the division of the pie depend on whether the corporation liquidates or reorganizes in bankruptcy, and bankruptcy law also includes rules for deciding whether reorganization or liquidation will occur. When corporations liquidate under Chapter 7 of US bankruptcy law, the pie includes all of the firm's assets but none of its owners' other assets. This reflects the doctrine of limited liability, which exempts owners of equity in corporations from personal liability for the corporation's debts beyond loss of the value of their shares. The corporation's assets are liquidated and the proceeds are used to repay creditors according to the absolute priority rule (APR). The APR carries into bankruptcy the non-bankruptcy rule that debt must be repaid in full before equity receives anything. The APR also determines how the pie is divided among creditors. Classes of creditors are ranked and each class receives full payment of its claims until funds are exhausted.

When corporations reorganize under Chapter 11 of US bankruptcy law, the reorganized corporation retains most or all of its assets and continues to operate – generally under the control of its pre-bankruptcy managers. Bankruptcy law again provides a procedure for determining both the size and the division of the pie in reorganization, but the procedure involves a negotiation process rather than a formula.

Funds to repay creditors come from the firm's future earnings rather than from liquidating its assets. The rule for the division of the pie in reorganization is also different. Instead of creditors receiving either full payment or nothing, most classes of creditors receive partial payment regardless of their rank, and pre-bankruptcy equity receives some of the reorganized firm's new shares. This priority rule is referred to as 'deviations from the APR' since equity receives a positive payoff even though creditors are repaid less than 100 per cent. Creditors and equity negotiate a reorganization plan that specifies what each group will receive, and the plan must be adopted by a super-majority vote of each class of creditors and equity.

For individuals in financial distress, bankruptcy law also includes both rules for determining which of the individual's assets must be used to repay debt (the size of the pie) and rules for dividing the assets among creditors (the division of the pie). In determining the size of the pie, personal bankruptcy law plays a role similar to that of limited liability for corporate equity-holders, since it limits the amount of assets that individual debtors must use to repay. It does this by specifying exemptions, which are maximum amounts of both financial wealth and post-bankruptcy earnings that individual debtors are allowed to keep. Only amounts in excess of the exemption levels must be used to repay. An important feature of US bankruptcy law is the 100 per cent exemption for post-bankruptcy earnings, known as the 'fresh start', which greatly limits individual debtors' obligation to repay. (Note that in 2005 Congress adopted limits on the availability of the fresh start.) In personal bankruptcy, the rule for dividing repayment among creditors is also the APR.

An important difference between personal and corporate bankruptcy law is that, while corporations may either liquidate or reorganize in bankruptcy, individuals can only reorganize (even though the most

commonly used personal bankruptcy procedure in the United States is called liquidation). This is because part of individual debtors' wealth is their human capital, and the only way to liquidate human capital is to sell debtors into slavery – as the Romans did. Since slavery is no longer used as a penalty for bankruptcy, all personal bankruptcy procedures are forms of reorganization in which individual debtors keep their human capital and the right to decide whether to use it.

Economic objectives

The economic objectives are similar in corporate and personal bankruptcy. One important objective of bankruptcy is to require sufficient repayment that lenders will be willing to lend – not necessarily to the bankrupt debtor but to other borrowers. Reduced access to credit makes debtors worse off because businesses need to borrow in order to grow and individuals benefit from borrowing to smooth consumption. On the other hand, repaying more to creditors harms debtors by making it more difficult for financially distressed firms to survive and by reducing financially distressed individuals' incentive to work. Both the optimal size and the division of the pie in bankruptcy are affected by this trade-off. A second important objective of both types of bankruptcy is to prevent creditors from harming debtors by racing to be first to collect. When creditors think that a debtor is in financial distress, they have an incentive to collect their debts quickly, since the debtor will be unable to repay all creditors in full. But aggressive collection efforts by creditors may force debtor firms to shut down even when the best use of their assets is to continue operating, and may cause individual debtors to lose their jobs (if creditors repossess their cars or garnish their wages). A third objective of personal bankruptcy law that has no counterpart in corporate bankruptcy is to provide individual debtors with partial consumption insurance. If consumption falls substantially, long-term harm may occur, including debtors' children leaving school prematurely in order to work or debtors' medical conditions going untreated and becoming disabilities. Discharging debt in bankruptcy when debtors' consumption would otherwise fall reduces these costs.
An additional objective that applies only to corporate bankruptcy is to reduce filtering failure. Financially distressed firms may be economically either efficient or inefficient, depending on whether the best use of their assets is the current use or some alternative use. Filtering failure in bankruptcy occurs when efficient but financially distressed firms shut down and when inefficient financially distressed firms reorganize and continue operating. The cost of filtering failure is either that the firm's assets remain tied up in an inefficient use or that they move to an alternative use when the current one is the most efficient. Many researchers have argued that reorganization in Chapter 11 tends to save economically inefficient firms that should shut down.
Research on corporate and personal bankruptcy is discussed separately below. Small-business bankruptcy is included with personal bankruptcy, because small businesses are often unincorporated and therefore their debts are legal liabilities of the business owner. When these businesses fail, their owners can file for bankruptcy and both their business and personal debts will be discharged. (Note that most of the research on bankruptcy is focused on US law and US data. For a longer survey of research on corporate and personal bankruptcy that includes many references, see White, 2006.)

Corporate bankruptcy

Theory

A central theoretical question in corporate bankruptcy is how priority rules affect the efficiency of decisions made by managers (who are assumed to represent the interests of equity), particularly whether the firm invests in safe or risky projects and whether and when it files for bankruptcy. Inefficient investment decisions lower the firm's return, and inefficient bankruptcy decisions result in filtering failure. Both reduce creditors' returns and cause them to raise interest rates or to reduce the amount they are willing to lending.

Bebchuk (2002) compares the efficiency of corporate investment decisions when the priority rule in bankruptcy is the APR with those when deviations from the APR occur, where use of the APR represents liquidation in bankruptcy and deviations from the APR represent reorganization in bankruptcy. A well-known result in finance is that equity prefers risky to safe investment projects, because equity gains disproportionately when risky projects succeed and bears only limited losses when risky projects fail. If the priority rule in bankruptcy is changed from the APR to deviations from the APR, then equity's preference for risky projects becomes even stronger. This is because equity now receives a positive return rather than nothing when risky projects fail, and the same high return when risky projects succeed. This change makes risky projects even more attractive relative to safe ones, since the latter rarely fail and so their return is unaffected by the change in the priority rule. Thus, when the bankruptcy regime is reorganization rather than liquidation, investment decisions become less efficient because equity over-invests in risky projects.

But Bebchuk argues that the results are reversed when firms are already in financial distress. Here, deviations from the APR reduce rather than increase equity's bias towards choosing risky investment projects. This is because, when the project is likely to fail and the firm to file for bankruptcy, equity's main return comes from the share that it receives of the firm's value in bankruptcy – the deviations from the APR. And since safe projects have higher downside returns, they generate more for equity. Thus the overall result is that neither priority rule in bankruptcy always leads to efficient investment incentives. Similar models have shown that none of the standard priority rules always leads to efficient bankruptcy decisions.

Bankruptcy law also affects other economically important decisions, including whether managers default strategically, whether they reveal important information about the firm's condition to creditors, and how much effort they expend. Strategic default occurs when firms default on their debt even though they are financially solvent. In the financial contracting literature, there is a trade-off between strategic default and filtering failure (see Bolton and Scharfstein, 1996). Suppose a firm borrows D in period 0 to finance an investment project. The firm will either succeed or fail. If it succeeds, it earns $R_1 > D$ in period 1 and an additional $R_2 > L$ in period 2. If it fails, then its period 1 earnings are zero, but it still earns R_2 in period 2. Regardless of whether the firm succeeds or fails, the liquidation value of its assets is L in period 1 and 0 in period 2. The firm's earnings are assumed to be observable but unverifiable. The loan contract calls for the firm to repay D in period 1 and it gives lenders the right to liquidate the firm in period 1 and collect L if default occurs. The contract does not call for any repayment in period 2, since promises to repay are not credible when the firm's liquidation value is zero. Liquidating the firm in period 1 is inefficient, since the firm would earn more than L if it continued to operate. Under these assumptions, the firm's owners always repay in period 1 when the firm is successful, since they benefit from retaining control and collecting R_2 in the following period. But if the firm fails, then its owners

default and creditors liquidate it. Thus there is no strategic default, but filtering failure occurs since there is inefficient liquidation. If lenders instead allowed owners to remain in control following default, then there would be no filtering failure but a high level of strategic default. Because of incomplete information, strategic default and filtering failure cannot both be eliminated.

Bankruptcy law also affects managers' choice of how much effort to expend and whether to delay filing for bankruptcy. Povel (1999) analyses a model in which managers make an effort-level decision and also receive an early signal on whether the firm will succeed. When the signal is bad, managers decide whether to file for bankruptcy or continue operating outside of bankruptcy. Filing for bankruptcy is assumed to be economically efficient in this situation, since it allows creditors to rescue the firm. Neither the effort-level decision nor the signal is observed by creditors. Povel considers two different bankruptcy laws: reorganization and liquidation. In the model, if the bankruptcy procedure is reorganization, the result is that managers choose low effort and file for bankruptcy when the signal is bad. Filing for bankruptcy is economically efficient, but low effort by managers is inefficient. Conversely, if the bankruptcy procedure is liquidation, the result is that managers choose high effort and avoid bankruptcy when the signal is bad. This trade-off suggests that the better bankruptcy procedure could be either reorganization or liquidation, depending on parameter values. See Berkovitch, Israel and Zender (1998) for a similar model that explores the efficiency of auctions as an alternative bankruptcy procedure.

There is a large literature on reforms of bankruptcy law. Most studies start from the premise that too many firms reorganize in bankruptcy under current law, since reorganization under Chapter 11 has both high transactions costs and high costs of filtering failure. One proposal is to auction all bankrupt firms and use the proceeds to repay creditors according to the APR. This procedure has the dual advantages that it would be quick and that the new owners would make efficient decisions on whether to save or liquidate each firm (see Baird, 1986). Another proposal is to use options to divide the value of firms in reorganization (Bebchuk, 1988). Both auctions and options would establish a market value of the firm's assets, so that creditors could be repaid according to the APR and deviations from the APR could be eliminated. Another proposal, called bankruptcy contracting, would allow debtors and creditors to adopt their own bankruptcy procedure when they write their loan contracts, rather than requiring them to use the state-supplied mandatory bankruptcy procedure. Schwartz (1997) showed that bankruptcy contracting could improve efficiency in particular circumstances. But whether bankruptcy contracting or any of the other reform proposals would work well in a general model that takes account of other complications – such as the existence of multiple creditor groups and strategic default – has not been established.

Empirical research

Now we turn to empirical research on corporate bankruptcy. It has focused on measuring the costs of bankruptcy and the size and frequency of deviations from the APR. Studies of the costs of bankruptcy include only the legal and administrative costs of the bankruptcy process; that is, the costs of bankruptcy-induced disruptions are excluded. Most studies have found that bankruptcy costs as a fraction of the value of firms' assets are higher in liquidation than in reorganization, but this may reflect the fact that bankruptcy costs are subject to economies of scale and larger firms tend to reorganize rather than

liquidate in bankruptcy. Unsecured creditors generally receive nothing in liquidation, but are repaid one-third to one-half of their claims in reorganization. This higher return in reorganization could be due to selection bias, if firms that reorganize are in relatively better financial condition. Other studies provide evidence that Chapter 11 filings are associated with an increase in managers' and directors' turnover, suggesting that the process is very disruptive. In addition, many firms that reorganize in Chapter 11 end up requiring additional financial restructuring within a short period. This is consistent with the theoretical prediction that too many financially distressed firms reorganize. Deviations from the APR have been found to occur in around three-quarters of all reorganization plans of large corporations in bankruptcy (see Bris, Welch and Zhu, 2006, for a recent study and references).

Personal bankruptcy

When an individual or a married couple files for bankruptcy under Chapter 7 (the most commonly used procedure), most unsecured debts are discharged. Debtors are obliged to use their non-exempt assets to repay debt, but their future earnings are entirely exempt under the 'fresh start'. Exemption levels, unlike other features of US bankruptcy law, differ across states. The most important exemption is the 'homestead' exemption for equity in owner-occupied homes, which varies widely from zero to unlimited. Because debtors can convert non-exempt assets such as bank accounts into home equity before filing for bankruptcy, high homestead exemptions protect all types of wealth for debtors who are homeowners.

There is also a second personal bankruptcy procedure, Chapter 13, under which debtors' assets are completely exempt, but they must use some of their future earnings to repay their debt. Until recently, debtors had the right to choose between the two procedures and, since most debtors have few non-exempt assets, Chapter 7 was almost always the more favourable. It was also the more heavily used – about 70 per cent of all personal bankruptcy filings were under Chapter 7. Those debtors who filed under Chapter 13 often repaid only token amounts, since the value of their non-exempt assets was zero. However, in late 2005 bankruptcy reforms went into effect that will force some debtors having higher incomes to file for bankruptcy under Chapter 13 and to repay more.

Theory

From an economic standpoint, the main reason for having a personal bankruptcy procedure is to provide individual debtors with consumption insurance by discharging debt when the obligation to repay would cause a substantial reduction in their consumption levels. This is because sharp falls in consumption can have permanent negative effects – debtors may become homeless, their illnesses may become disabilities for lack of medical care, and their children may leave school prematurely and have lower future earnings. Consumption insurance is mainly provided by the public sector in the form of the social safety net – welfare payments, food stamps and health insurance for the poor. But bankruptcy reduces the cost to the public sector of providing the safety net, since discharge of debt in bankruptcy frees up funds for consumption that debtors might otherwise use to repay debt.

The higher the exemption levels for wealth and earnings in bankruptcy, the more the consumption insurance that bankruptcy provides. Theoretical research on personal bankruptcy has focused on deriving optimal exemption levels. Higher levels of both exemptions benefit debtors by providing them

with extra consumption insurance, but harm those who repay their debts by reducing the availability of credit and increasing interest rates. However, the two exemptions have differing effects on debtors' incentives to work after bankruptcy. A higher wealth exemption is likely to have little effect on work incentives, while a higher earnings exemption increases debtors' incentive to work as long as the positive substitution effect outweighs the negative income effect. The model suggests that the optimal earnings exemption is 100 per cent – that is, the 'fresh start' – while the optimal wealth exemption is an intermediate level. This is because a higher earnings exemption both encourages debtors to work more after bankruptcy and provides better consumption insurance than a higher wealth exemption. See White (2005).

An important feature of personal bankruptcy law is that it encourages opportunistic behaviour by debtors. Although bankruptcy debt relief is intended for debtors whose consumption has fallen sharply due to factors such as job loss or illness, in fact debtors' incentive to file is hardly affected by these adverse events. Debtors' financial benefit from bankruptcy equals the amount of debt discharged minus the sum of non-exempt assets that must be used to repay plus the costs of bankruptcy. White (1998b) calculated that at least one-sixth of US households would benefit financially from filing for bankruptcy, and this figure rose to more than one-half if households were assumed to pursue various strategies, such borrowing more on an unsecured basis, converting non-exempt assets into exempt home equity, and moving to states with high homestead exemptions. White (1998b) also found that these calculations understate the proportion of households that would benefit from bankruptcy, since some households that would not benefit from filing immediately could benefit from filing in the future. She calculated the value of the option to file for bankruptcy and found that it is particularly valuable for high-wealth households and those in high-exemption states. These features of bankruptcy law are probably responsible for high filing levels (more than 1.6 million US households filed for bankruptcy in 2003) and for the fact that the US Congress recently changed Chapter 7 to make bankruptcy less attractive to many debtors.

Empirical research

Most of the empirical research on personal bankruptcy makes use of the variation in exemption levels that causes bankruptcy law to differ across US states. Gropp, Scholz and White (1997) found that, if households live in states with high rather than low exemptions, they are more likely to be turned down for credit, they borrow less, and they pay higher interest rates. They also found that in high-exemption states credit is redistributed from low-asset to high-asset households. Households in high-exemption states demand more credit because borrowing is less risky, but lenders respond by offering larger loans to high-asset households while rationing credit more tightly to low-asset households. Fay, Hurst and White (2002) found that households are more likely to file for bankruptcy when their financial benefit from filing is higher. Since households' financial benefit from filing is positively related to the size of the exemption, this means that households are more likely to file if they live in states with higher bankruptcy exemptions. Fay, Hurst and White did not find that recent job loss or health problems were significantly related to whether households filed for bankruptcy. But they found that households were more likely to file when they live in regions that have higher average bankruptcy filing rates – which suggests the existence of network effects.

Personal bankruptcy exemption levels also affect small businesses, since business debts often are personal obligations of the business owner and these debts are discharged in bankruptcy. Fan and White (2003) found that individuals are more likely to own or start businesses in states with higher exemption levels, presumably because the additional consumption insurance in these states makes going into business more attractive by lowering the cost of failure. But Berkowitz and White (2004) found that small businesses are more likely to be turned down for credit and to pay higher interest rates if they are located in states with higher exemption levels. Overall, higher exemption levels have mixed effects on small business.

Finally, since higher exemption levels provide households with additional consumption insurance, the variance of household consumption is predicted to be smaller in states that have higher exemption levels. Grant (2006) found macro-level support for this hypothesis using data on the variance of consumption across state-years.

See Also

- bankruptcy, economics of

Bibliography

Baird, D. 1986. The uneasy case for corporate reorganizations. *Journal of Legal Studies* 15, 127–47.

Bebchuk, L. 1988. A new method for corporate reorganization. *Harvard Law Review* 101, 775–804.

Bebchuk, L. 2002. The ex ante costs of violating absolute priority in bankruptcy. *Journal of Finance* 57, 445–60.

Berkovitch, E., Israel, R. and Zender, J. 1998. The design of bankruptcy law: a case for management bias in bankruptcy reorganizations. *Journal of Financial and Quantitative Analysis* 33, 441–67.

Berkowitz, J. and White, M. 2004. Bankruptcy and small firms' access to credit. *RAND Journal of Economics* 35, 69–84.

Bolton, P. and Scharfstein, D. 1996. Optimal debt structure and the number of creditors. *Journal of Political Economy* 104, 1–25.

Bris, A., Welch, I. and Zhu, N. 2006. The costs of bankruptcy: Chapter 7 cash auctions vs. Chapter 11 bargaining. *Journal of Finance* 61.

Fan, W. and White, M. 2003. Personal bankruptcy and the level of entrepreneurial activity. *Journal of Law and Economics* 46, 543–68.

Fay, S., Hurst, E. and White, M. 2002. The household bankruptcy decision. *American Economic Review* 92, 706–18.

Grant, C. 2006. Evidence on the effect of U.S. bankruptcy exemptions. In *Economics of Consumer Credit: European Experience and Lessons From the U.S.*, ed. G. Bertola, R. Disney and C. Grant. Cambridge, MA: MIT Press.

Gropp, R., Scholz, K. and White, M. 1997. Personal bankruptcy and credit supply and demand. *Quarterly Journal of Economics* 112, 217–52.

Povel, P. 1999. Optimal soft or tough bankruptcy procedures. *Journal of Law, Economics, and Organization* 15, 659–84.

Schwartz, A. 1997. Contracting about bankruptcy. *Journal of Law, Economics, and Organization* 13, 127–46.

White, M. 1998a. Why don't more households file for bankruptcy? *Journal of Law, Economics, and Organization* 14, 205–31.

White, M. 1998b. Why it pays to file for bankruptcy: a critical look at incentives under U.S. bankruptcy laws and a proposal for change. *University of Chicago Law Review* 65, 685–732.

White, M. 2002. Sovereigns in distress: do they need bankruptcy? *Brookings Papers on Economic Activity* 2002(1), 287–319.

White, M. 2005. Personal bankruptcy: insurance, work effort, opportunism and the efficiency of the ‘Fresh Start’. Paper presented to the American Law and Economics Association conference, New York, May.

White, M. 2006. Bankruptcy law. In *Handbook of Law and Economics*, ed. A. Mitchell Polinsky and S. Shavell. Amsterdam: North-Holland.

How to cite this article

White, Michelle J. "bankruptcy law, economics of corporate and personal." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000220> doi:10.1057/9780230226203.0094

The New Palgrave Dictionary of Economics Online

bankruptcy, economics of

Arturo Bris

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bankruptcy is the formal procedure to resolve the disputes among creditors, shareholders, and managers of a company in financial distress. Countries have designed bankruptcy procedures that differ in the control that is given to the existing management relative to creditors. These differences determine the incentives that are given to the parties before, during, and after the bankruptcy proceedings. They also determine how expensive the bankruptcy process is. Ultimately, bankruptcy costs are borne by firms' shareholders.

Keywords

absolute priority rule (APR); administrative receiverships; bankruptcy; bankruptcy law; bankruptcy, economics of; Chapter 11 reorganizations; Chapter 7 liquidations (USA); default; liquidations; workouts

Article

Bankruptcy is the legal procedure whereby the assets of a debtor are distributed among its creditors. The debtor can be either an individual or a firm. In corporations, bankruptcy happens when either the firm or its creditors delegate a third party – be it a judge or other public official – to determine the amount of the creditors' claims, as well as the way to distribute the firm's assets among them. In essence, bankruptcy results from financial distress, which happens when the market value of the assets is insufficient to satisfy the debt claims, or when the firm does not generate enough cash flow to meet the coupon and interest payments. An alternative to bankruptcy is an informal reorganization, or *workout*, whereby creditors relax debt covenants, possibly exchanging their claims for a package of new claims. Bankruptcy is an old European institution that derives its name from the Italian '*banca rotta*' (broken bench). It refers to the boards from which traders in medieval towns traded coins, and which they broke whenever they defaulted on their payments. Nowadays, countries have implemented different procedures to deal with the distribution of the assets of a firm that cannot meet its debt obligations. In the United States, firms and creditors can opt into two forms of restructuring. Under a Chapter 7

liquidation, assets are sold piecemeal and the proceeds distributed according to the *absolute priority rule* (APR), whereby debt and equity are paid according to a predetermined order: secured debt first, then unsecured claims, and finally common stock. The distinction between *senior* and *junior* claims refers to the priority of secured debt (senior) over unsecured debt (junior). The firm ceases to exist after a Chapter 7. Under a Chapter 11 reorganization, shareholders and creditors agree on a reorganization plan, which allows the company to continue. When the company enters a Chapter 11, the firm becomes a ‘debtor-in-possession’, a term that recognizes that the management retains control of the company's operations, under court supervision. In a Chapter 11, APR may be violated if secured creditors give up part of their claims in favour of unsecured debtors, or if shareholders receive some interest in the restructured firm at the expense of debtholders (Herbert, 1998).

Under the absolute priority rule, unsecured claims are classified into priority claims and general unsecured claims. Priority claims are further classified into three groups: administrative claims, wages and employee benefits, and taxes. This means that, under APR – which is always upheld in Chapter 7 cases – wages cannot be paid unless administrative expenses (compensation of lawyers and other professionals) have been satisfied in full. Moreover, tax claims include only those taxes that the firm owes at the time it files for bankruptcy.

The practice in the United States is to reimburse administrative expenses incurred by the committee of unsecured creditors. A Chapter 11 creditors’ committee is composed of creditors ‘that hold the seven largest claims against the debtor of the kinds represented on such committee’ (Bankruptcy Code §1102 (b)(1)). The bankruptcy court is authorized to reimburse a substantial portion of the expert expenses that juniors incur. However, the United States code does not authorize the bankruptcy court to compensate the expenses of creditors whom it defines as ‘senior.’ This cost allocation fails to encourage the seniors to spend on activities that increase the value of the firm, but encourages the juniors to spend on activities that maximize only the value of their own claims.

In the United States the debtor has an exclusivity period of 120 days to file a plan of reorganization. This period can be, and usually is, extended upon the debtor's requests. In the plan, each class of creditors is classified as *impaired* or *unimpaired*. An unimpaired class of creditors is paid in full, and does not vote on the reorganization plan. The plan requires the approval of each impaired class of creditors and equity security holders. Approval requires dual majority: more than one-half of the votes, and more than two-thirds of the amount of the claims.

In the United Kingdom and other countries with British legal traditions, such as Canada, Australia and New Zealand, bankrupt companies are restructured via an *administrative receivership*. White (1996) and Franks and Davydenko (2006) provide a comparison between the bankruptcy codes in the United States and some European countries. Under an administrative receivership, the secured creditors appoint an expert (the administrative receiver) whose objective is to obtain sufficient funds to repay the secured creditors. To do that, the receiver can either liquidate some assets or sell the company as a going concern. The receiver does not have any obligation with respect to other creditors or shareholders, as long as absolute priority is respected. Unlike with a United States Chapter 11, in a receivership control is transferred from the management to the secured creditors.

Under the old French system neither the firm nor the creditors retained control. The court appointed an administrator who managed the day-to-day operations of the firm, and whose objectives were, first, to preserve the estate and employment, and then to satisfy creditors. Most systems in Continental Europe have followed this tradition. In the new Loi de Sauvegarde des Enterprises enacted in 2005, France has

moved towards the Chapter 11 system in the United States.
In Germany, the system introduced in 1999 establishes an automatic stay of three months, which means that creditors cannot dispose of the firm's assets during that period. Moreover, and similar to a Chapter 7 in the United States, the court appoints an administrator who monitors the process and determines a plan of reorganization.
Auctions are a very efficient alternative to court-administered procedures. In Sweden, the court appoints an independent trustee who is in charge of selling the firm's assets to the highest bidder. The winning bidder can pay only in cash, as described in Thorburn (2000), and the trustee distributes the proceeds respecting the APR. Stromberg (2000) shows that in one out of three cases in Sweden the assets are sold back to the incumbent managers (because they have the highest valuation of the assets), and the remaining cases are liquidated.

Controversy over Chapter 11

In recent years, there has been a convergence in bankruptcy laws towards a Chapter 11-type reorganization. Countries in western and eastern Europe, Asia and Latin America have enacted regulations that allow managers to retain control of defaulted firms. Regulators have moved from a system that favours liquidations to a legal procedure that tends to maximize the probability of firm survival. However, the efficiency of Chapter 11 has been questioned by scholars like Bebchuk (1988), Adler (1993), Schwartz (1998), Baird and Rasmussen (2002), and Baird and Morrison (2005). They promote a *contractual approach* to bankruptcy, or a formal scheme of bargained bankruptcy. Under this view, the parties should be free to bargain in advance over a set of rules that will govern their rights in the event of bankruptcy, with Chapter 11 being only a default system. Bebchuk (1988), for instance, proposes that firms can issue derivative securities, contingent on the firm being in default. The contractual view attacks the Chapter 11 system on several fronts, first of all on the grounds that it leads to inefficient outcomes (Baird and Morrison, 2005; Franks and Loranth, 2006). In particular, Franks and Loranth show that Chapter 11 in Hungary is biased in favour of inefficient going concerns. The argument is that most bankrupt firms should be liquidated rather than reorganized. Chapter 11 is also attacked because it is considered a more lengthy process than other systems (Stromberg, 2000; Thorburn, 2000). Additionally, it is extremely expensive (Bris, Welch and Zhu, 2006).
The opponents of such a private bankruptcy system (Warren and Westbrook, 2005) make two important arguments to defend Chapter 11. In principle, a private system would have only redistributive effects, with some creditors (secured and large creditors) shifting risks to others. Also, Chapter 11 is a mechanism by which benevolent large creditors give up part of their claims in favour of small, empowered creditors. Therefore it has a positive redistributive effect. Finally, a private system is inefficient because of the duplication of transaction costs.
Most of the theoretical and empirical research on bankruptcy addresses the conflicts that arise among creditors, shareholders, firm managers and bankruptcy specialists. These conflicts arise during the bankruptcy proceedings, but also when the company is in financial distress and before it files for bankruptcy. The design of the bankruptcy system can affect the interaction among all these agents, the efficiency of the bankruptcy process and, therefore, the costs of bankruptcy.

Incentives before filing for bankruptcy

Financial distress may lead to bankruptcy if either the firm management or the creditors opt into a legal procedure to resolve their disputes. But, if the distressed firm is economically viable, managers have an incentive to delay filing for bankruptcy and to maintain operations, especially if the legal procedure gives control to a third party. Self-interested managers will then preserve their jobs at the expense of shareholders and creditors. Jensen and Meckling (1976) show that in distressed firms there is a *debt overhang* problem. Managers have an incentive to bypass positive net present value (NPV) projects (a problem known as *underinvestment*) because they benefit only current creditors (Myers, 1977). Instead, when choosing between less and more risky projects managers prefer to invest in more risky projects because managers act on behalf of shareholders, and shareholders, because of limited liability, are interested only in the upside of the investments (*excess risk taking* or *overinvestment*). These incentives in turn reduce the value of the debtor's claims and ultimately the value of the firm because creditors take them into account when pricing their securities.

Recently, Adler, Capkun and Weiss (2005) have shown that a change in regulation in the United States around 2000, which gave more control to creditors during the filing period, induced managers to delay the bankruptcy filing. Indeed, they show that after 2000 firms that file for Chapter 11 in the United States display a worse financial and operating condition. This can explain why, in countries with secured creditor control of the bankruptcy process, the number of bankruptcy filings is much lower, and firm managers prefer liquidation (Claessens and Kappler, 2005).

Conversely, and depending on the debt structure, managers may have an incentive to default strategically even if the firm is still economically viable. Bolton and Scharfstein (1996) argue that managers will always prefer to default strategically so as to divert cash to themselves. In order to avoid that distortion, creditors should have the right to liquidate the firm in case of default. However, this induces inefficient liquidations because the value of the firm as a going concern may exceed its liquidation value. Bolton and Scharfstein (1996) show that borrowing from multiple creditors solves the problem by increasing the liquidation value of the firm.

Incentives during bankruptcy proceedings

The efficiency of the bankruptcy process and a firm's capital structure are closely related because, for a firm with multiple creditors, bankruptcy results in coordination problems among creditors, as well as conflicts between secured and unsecured, or between senior and junior, claimants. Regarding coordination problems, and in contrast to Bolton and Scharfstein (1996), Bris and Welch (2005) argue that, when competing for the firm's assets, multiple creditors (similar to public bonds) waste the firm's resources in fighting with each other; hence, it is more efficient to issue highly concentrated debt (bank debt). Indeed, Welch (1997) shows that bank debt should be senior because a single creditor fights better with shareholders, thereby increasing the *ex ante* value of the debt.

Conflicts between secured and unsecured creditors depend on the bankruptcy system and the priority rules. If unsecured creditors can extract rents at the expense of more senior debtors (that is, if absolute priority can be violated), then a firm may prefer to liquidate its assets because unsecured creditors will expend the firm's resources in order to satisfy part of their claim. Eberhart, Moore and Roenfeldt (1990)

and Franks and Torous (1994) show that APR is often violated under Chapter 11.

Firms in bankruptcy are allowed sometimes to issue new financing that can be senior to the already outstanding debt (*debtor-in-possession, DIP, financing*). The ability to raise DIP financing is priced *ex ante* by the firm's creditors. Therefore, it increases the value of the firm *ex post* but it reduces shareholder value *ex ante*. This trade-off has been extensively considered in the literature.

Life after bankruptcy

The design of the bankruptcy process can also affect the performance of firms when they emerge from Chapter 11. Hotchkiss (1995) reports that over 40 per cent of the firms in her sample still experience operating losses in the three years following the bankruptcy case, while another 32 per cent re-file for bankruptcy or restructure their debt.

Bankruptcy costs

Bankruptcy costs encompass not only the explicit payments made to bankruptcy specialists (lawyers, trustees, accountants, investment bankers) but also the indirect costs of being in default. Among the latter, we can include loss of customers when the company is in financial distress, adverse payment terms enforced by suppliers when the viability of the firm is not guaranteed, loss of key personnel and waste of management time.

Measuring the indirect costs of bankruptcy is very difficult. Altman (1984) uses forgone profits as a proxy, while Opler and Titman (1994) focus on losses of trade credit. However, because of the nature of the indirect costs, any proxy tends to underestimate their extent. Other researchers have used the length of the proceedings as a proxy for indirect bankruptcy costs, under the assumption that, the longer the firm stays in bankruptcy, the larger the collateral effects (Franks and Torous, 1994). Bris, Welch and Zhu (2006) show that both liquidations under Chapter 7 and reorganizations under Chapter 11 take about two years to resolve. In exploring the Swedish system, Thorburn (2000) shows that the Swedish auction system is much faster than the United States Chapter 11 process, since auctions take only two months on average.

The evidence on direct costs is more extensive. Warner (1977) finds that the direct costs of bankruptcy are about four per cent of the market value of the firm one year prior to the default. This result is based on a sample of 11 bankrupt railroads. Altman (1984) calculates these costs to be about 7.5 per cent of firm value, using a broader sample of 19 bankrupt companies from 1974 to 1978. Using 105 Chapter 11 cases, Ang, Chua and McConnell (1982) report that administrative fees are about 7.5 per cent of the total liquidating value of the bankrupt corporation's assets. Lubben (2000) calculates in his sample of 22 firms from 1994 that the cost of legal counsel in Chapter 11 bankruptcy represents 1.8 per cent of the distressed firm's total assets, and in some cases more than five per cent. In his average case, the debtor spends $500,000 on lawyers, and creditors spend $230,000. LoPucki and Doherty (2004) study a sample of 48 cases from 1998 to 2002, mostly from Delaware and New York. They report that professional fees were 1.4 per cent of the debtors’ total assets at the beginning of the bankruptcy case. Bris, Welch and Zhu (2006) compare the costs of bankruptcy for Chapter 7 and Chapter 11 cases. They report that the

mean ratio of total expenses to assets is 9.5 per cent for Chapter 11, and 8.1 per cent for Chapter 7. However, they warn against simple averages because cost measures depend on the value of the assets (pre-bankruptcy or post-bankruptcy) one uses.

Conclusion

The design of a bankruptcy system is very important because it determines shareholder value for all firms, whether or not they are in financial distress. The reason is that any conflict that can arise among creditors of different classes, and any coordination problem in the bankruptcy proceedings among creditors in a similar class, are both priced in the debt securities that a company issues. Moreover, the bankruptcy system can impose distortions on a firm's policies when it is in financial distress; in particular it can induce managers to make suboptimal decisions at the expense of shareholders. Countries' legal systems differ in terms of who controls the firm's assets during bankruptcy. Because control shapes the conflicts set out above, this feature of the bankruptcy system is one of the most important considered by the academic literature. Additionally, scholars have studied the issue of bankruptcy costs in detail. While we have extensive evidence on the direct costs of bankruptcy, the indirect costs of being in distress are very difficult to measure.

See Also

- bankruptcy law, economics of corporate and personal
- default and enforcement constraints
- extremal quantiles and value-at-risk

Bibliography

Adler, B. 1993. Financial and political theories of American corporate bankruptcy. *Stanford Law Review* 45, 311–46.

Adler, B., Capkun, V. and Weiss, L. 2005. Theory and evidence on the bankruptcy initiation problem. Working paper, University of Lausanne.

Altman, E. 1984. A further empirical investigation of the bankruptcy cost question. *Journal of Finance*, 1067–89.

Ang, J., Chua, J. and McConnell, J. 1982. The Administrative costs of corporate bankruptcy: a note. *Journal of Finance* 37, 219–26.

Baird, D. and Morrison, E. 2005. Serial entrepreneurs and small business bankruptcies. *Columbia Law Review* 105, 2310–68.

Baird, D. and Rasmussen, R. 2002. The end of bankruptcy. *Stanford Law Review* 55, 751–90.

Bebchuk, L. 1988. A new approach to corporate reorganizations. *Harvard Law Review* 101, 775–804.

Bolton, P. and Scharfstein, D. 1996. Optimal debt structure and the number of creditors. *Journal of Political Economy* 104, 1–25.

Bris, A. and Welch, I. 2005. The optimal concentration of creditors. *Journal of Finance* 60, 2193–212.

Bris, A., Welch, I. and Zhu, N. 2006. The costs of bankruptcy: Chapter 7 liquidations vs. Chapter 11 reorganizations. *Journal of Finance* 61, 1253–303.

Claessens, S. and Kappler, L. 2005. Bankruptcy around the world: explanations of its relative use. *American Law and Economics Review* 7, 253–83.

Eberhart, A., Moore, W. and Roenfeldt, R. 1990. Security pricing and deviations from absolute priority rule in bankruptcy proceedings. *Journal of Finance* 45, 1457–69.

Franks, J. and Davydenko, S. 2006. Do bankruptcy codes matter? A study of defaults in France, Germany and the UK. Finance Working Paper No. 89/2005. Brussels: European Corporate Governance Institute.

Franks, J. and Loranth, G. 2006. A study of inefficient going concerns in bankruptcy. Working paper, London Business School.

Franks, J. and Torous, W. 1994. A comparison of financial recontracting in distressed exchanges and Chapter 11 reorganizations. *Journal of Financial Economics* 35, 349–70.

Herbert, M. 1998. *Understanding Bankruptcy*. New York: Matthew Bender & Company.

Hotchkiss, E. 1995. Post-bankruptcy performance and management turnover. *Journal of Finance* 50, 3–22.

Jensen, M. and Meckling, W. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

LoPucki, L. and Doherty, J. 2004. The determinants of professional fees in large bankruptcy reorganization cases. *Journal of Empirical Legal Studies* 1, 111–41.

Lubben, S. 2000. The direct costs of corporate reorganization: an empirical examination of professional fees in large Chapter 11 cases. *American Bankruptcy Law Journal* 74, 508–52.

Myers, S. 1977. The determinants of corporate borrowing. *Journal of Financial Economics* 5, 147–75.

Opler, T. and Titman, S. 1994. Financial distress and corporate performance. *Journal of Finance* 49, 1015–40.

Schwartz, A. 1998. A contract theory approach to business bankruptcy. *Yale Law Journal* 107, 1807–52.

Stromberg, P. 2000. Conflicts of interest and market illiquidity in bankruptcy auctions: theory and tests. *Journal of Finance* 55, 2641–92.

Thorburn, K. 2000. Bankruptcy auctions: costs, debt recovery and firm survival. *Journal of Financial Economics* 58, 337–68.

Warner, J. 1977. Bankruptcy cost: some evidence. *Journal of Finance* 32, 337–47.

Warren, E. and Westbrook, J. 2005. Contracting out of bankruptcy: an empirical intervention. *Harvard Law Review* 118, 1197–224.

Welch, I. 1997. Why is bank debt senior? A theory of asymmetry and claim priority based on influence costs. *Review of Financial Studies* 10, 1203–36.

White, M. 1996. The costs of corporate bankruptcy: a U.S.–European comparison. In *Corporate Bankruptcy: Economic and Legal Perspectives*, ed. J. Bhandari and L. Weiss. Cambridge: Cambridge University Press.

How to cite this article

Bris, Arturo. "bankruptcy, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000324> doi:10.1057/9780230226203.0093

The New Palgrave Dictionary of Economics Online

Banks, Jeffrey Scot (1958 – 2000)

David Austen-Smith
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adverse selection; agency models; Banks Set; Banks, J.; divinity; incomplete information; industrial organization; moral hazard; Nash equilibrium; pretrial bargaining; social choice; spatial theory of elections; subgame perfection

Article

Jeff Banks received his BA from University of California, Los Angeles, in 1982 and his Ph.D. from California Institute of Technology in 1986. He arrived as a new assistant professor of political science and economics at the University of Rochester with two significant and influential publications in hand, reflecting his principal interests in social choice theory (1985) and game theory (1987) respectively. By the time he died of complications from treating leukemia, Banks had published (or had forthcoming) more than 50 papers in economics, game theory and formal political theory, edited one conference volume, published a review monograph and coauthored two books.

In the 1985 paper, Banks completely characterized the set of subgame perfect Nash equilibrium outcomes achievable through an amendment agenda on a voting tournament. In effect, this set (which came to be called the Banks Set through no fault of its author) defines the consequential limits of an agenda-setter's power under the amendment procedure. Banks went on to write a series of influential papers on a variety of topics in social choice theory (for example, 1995; 1996; 2000; 2006) and in more applied positive political theory (for example, 1988; 1989; 1990a; 1990b). Indeed, it is difficult to identify any area within the field to which Banks did not make some significant contribution.

In (1987), Banks addressed the equilibrium refinement problem. Their proposed refinement, ‘divinity’, is on out-of-equilibrium beliefs and is closely related to the Cho and Kreps (1987) D1 refinement. Like D1, a virtue of divinity (in particular of its stronger variant, universal divinity) is that it is widely applicable and easy to compute, especially in games with a continuum of types and actions. Banks was a pioneer in developing strategic theories of collective decision-making under incomplete information, and his (1990a) paper is both the seminal contribution to the spatial theory of elections under incomplete information and the first application of divinity to an applied problem. Subsequently, the refinement has

been used profitably by others on a variety of problems in industrial organization, pretrial bargaining and so forth. Along with incomplete information, Banks contributed some of the earliest formal papers dealing with problems of time and dynamics in politics. For example, he explored dynamic agency models that exhibit both moral hazard and adverse selection simultaneously (1993; 1998). Such environments are notoriously complicated and, as a step towards developing an appropriate toolbox for handling them, Banks (1992) made an important contribution to theory of denumerably armed bandits. Banks's professional career barely spanned 15 years, yet the footprint he has left on (especially) positive political theory is considerable. He was a fine teacher and a remarkable colleague; he is, and will continue to be, much missed.

Selected works

1985. Sophisticated voting outcomes and agenda control. *Social Choice and Welfare* 1, 295–306.

1987. (With J. Sobel.) Equilibrium selection in signaling games. *Econometrica* 55, 647–61.

1988. (With D. Austen-Smith.) Elections, coalitions and legislative outcomes. *American Political Science Review* 82, 405–22.

1989. Agency budgets, cost information and auditing. *American Journal of Political Science* 33, 670–99.

1990a. A model of electoral competition with incomplete information. *Journal of Economic Theory* 50, 309–25.

1990b. Monopoly agenda control and asymmetric information. *Quarterly Journal of Economics* 105, 445–64.

1992. (With R. Sundaram.) Denumerable-armed bandits. *Econometrica* 60, 1071–96.

1993. (With R. Sundaram.) Adverse selection and moral hazard in a repeated elections model. In *Political Economy: Institutions, Information and Competition*, ed. W. Barnett, M. Hinich and N. Schofield. New York: Cambridge University Press.

1995. Acyclic social choice from finite sets. *Social Choice and Welfare* 12, 293–310.

1996. (With D. Austen-Smith.) Information aggregation, rationality and the Condorcet jury theorem. *American Political Science Review* 90, 405–22.

1998. (With R. Sundaram.) Optimal retention in agency problems. *Journal of Economic Theory* 82, 293–323.

2000. (With J. Duggan.) A bargaining model of collective choice. *American Political Science Review* 94, 73–88.

2006. (With J. Duggan.) A multidimensional model of repeated elections. *Quarterly Journal of Political Science* 1, 49–85.

Bibliography

Cho, I. and Kreps, D. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102, 179–221.

How to cite this article

Austen-Smith, David. "Banks, Jeffrey Scot (1958–2000)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000330> doi:10.1057/9780230226203.0095

The New Palgrave Dictionary of Economics Online

Baran, Paul Alexander (1910 – 1964)

Paul M. Sweezy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Baran, P. A.; capitalism; feudalism; India, economics in; Japan, economics in; monopoly capitalism; potential surplus; surplus; surplus value; Sweezy, P. M.; underdevelopment

Article

Paul Baran, the eminent Marxist economist, was born on 8 December 1910 in Nikolaev, Russia, the son of a medical doctor who was a member of the Menshevik branch of the Russian revolutionary movement. After the October Revolution the family moved to Germany, where Baran's formal education began. In 1925 the father was offered a position in Moscow and returned to the USSR. Baran began his studies in economics at the University of Moscow the following year. Both his ideas and his politics were deeply and permanently influenced by the intense debates and struggles within the Communist Party in the late 1920s. Offered a research assignment at the Agricultural Academy in Berlin in late 1928, he enrolled in the University of Berlin, and when his assignment at the Agricultural Academy ended he accepted an assistantship at the famous Institute for Social Research in Frankfurt. This experience too had a lasting influence on his intellectual development.

Leaving Germany shortly after Hitler's rise to power, Baran sought without success to find academic employment in France. He therefore moved to Warsaw, where his paternal uncles had a flourishing international lumber business. During the next few years he travelled widely as a representive of his uncles’ business, ending up in London in 1938. With the approach of World War II, however, he decided to take what savings he had been able to accumulate, move to the United States, and resume his interrupted academic career.

Arriving in the United States in the fall of 1939, he was accepted as a graduate student in economics at Harvard. From there he went to wartime Washington, where he served in the Office of Price Administration, the Research and Development branch of the Office of Strategic Services, and the United States Strategic Bombing Survey, ending in 1945–6 as Deputy Chief of the Survey's mission to Japan. Back in the United States, he took a job at the Department of Commerce and gave lectures at George Washington University before being offered a position in the Research Department of the

Federal Reserve Bank of New York. After three years in New York, he accepted an offer to join the economics faculty at Stanford University and was promoted to a full professorship in 1951, a position he retained until his death of a heart attack on 26 March 1964.

Baran was not a prolific writer, but his two main books, *The Political Economy of Growth* (1957) and (in collaboration with Paul M. Sweezy) *Monopoly Capital: An Essay on the American Economic and Social Order* (1966), are generally considered to be among the most important works in the Marxian tradition of the post-World War II period.

The Political Economy of Growth is concerned with the processes and condition of economic growth (or development, the terms are used interchangeably) in both industrialized and underdeveloped societies, with a special emphasis throughout on the ways the two relate to and interact with each other. It is at once an outstanding work of scholarship weaving an intricate pattern of theory and history, and a passionate polemic against mainstream economics. Its chief (innovative) analytical concept is that of ‘potential surplus’, defined as ‘the difference between the output that *could* be produced in a given natural and technological environment with the help of employable productive resources, and what might be regarded as essential consumption’. (This concept presupposes Marx's ‘surplus value’, extending and modifying it for the particular purposes of the study in hand.) Two long chapters, totalling 90 pages, apply the concepts of surplus and potential surplus to the analysis of monopoly capitalism in ways that would later be refined and elaborated in *Monopoly Capital*. Three chapters (115 pages) follow on ‘backwardness’ (also called underdevelopment), and it is for these that the book has become famous, especially in the Third World.

Baran begins this analysis with a question which may be said to define the focus of the whole work: ‘Why is it that in the backward capitalist countries there has been no advance along the lines of capitalist development that are familiar from the history of other capitalist countries, and why is it that forward movement there has been slow or altogether absent?’ His answer, in briefest summary, is as follows: all present-day capitalist societies evolved from precapitalist conditions which Baran for convenience labels ‘feudal’ (explicitly recognizing that a variety of social formations are subsumed under this heading). Viable capitalist societies could have emerged in various parts of the world; actually the decisive breakthrough occurred in Western Europe (Baran speculates on the reasons, but in any case they are not crucial to the subsequent history). Having achieved its headstart, Europe proceeded to conquer weaker precapitalist countries, plunder their accumulated stores of wealth, subject them to unequal trading relations, and reorganize their economic structures to serve the needs of the Europeans. This was the origin of the great divide in the world capitalist system between the developed and the underdeveloped parts. As the system spread into the four corners of the globe, new areas were added, mostly to the underdeveloped part but in a few cases to the developed (North America, Australia, Japan). One of the highlights of Baran's study is the brilliant historical sketch of the contrasting ways India and Japan were incorporated into the world capitalist system, the one as a hapless dependency, the other as a strong contender for a place at the top of the pyramid of power. Baran's message to the Third World was loud and clear: once trapped in the world capitalist system, there is no hope for genuine progress; only a revolutionary break can open the road to a better future. The message has been widely heard. Most of the revolutionary movements of the Third World have been deeply influenced, directly or indirectly, by Paul Baran's *Political Economy of Growth*.

The economic analysis of *Monopoly Capital* is a development and systematization of ideas already contained in *the Political Economy of Growth* and Paul Sweezy's *The Theory of Capitalist Development*

(1942). The central theme is that in a mature capitalist economy dominated by a handful of giant corporations the potential for capital accumulation far exceeds the profitable investment opportunities provided by the normal *modus operandi* of the private enterprise system. This results in a deepening tendency to stagnation which, if the system is to survive, must be continuously and increasingly counteracted by internal and external factors. In the authors' estimation – not always shared, or even understood by critics – the new and original contributions of *Monopoly Capital* had to do mainly with these counteracting factors and their far-reaching consequences for the history, politics, and culture of American society during the period from roughly the 1890s to the 1950s when the book was written. They intended it, in other words, as much more than a work of economics in the usual meaning of the terms.

See Also

- monopoly capitalism

Selected works

There is a comprehensive bibliography of Baran's writings in English in a special issue of *Monthly Review*, 'In Memory of Paul Alexander Baran. Born at Nikolaev, the Ukraine, 8 December 1910. Died at San Francisco, California, 26 March 1964', 16(11), March 1965. This also includes statements on his life and work by more than three dozen contributors, most of whom had been his friends or colleagues.

1957. *The Political Economy of Growth*. New York: Monthly Review Press. 2nd edn, with a new preface, 1962.

1966. (With P.M. Sweezy.) *Monopoly Capital: An Essay on the American Economic and Social Order*. New York: Monthly Review Press.

1970. *The Longer View: Essays Toward a Critique of Political Economy*. Edited by J. O'Neill, preface by P.M. Sweezy, New York: Monthly Review Press. This volume, which follows an outline prepared before his death by the author, brings together his most important hitherto scattered essays and reviews.

How to cite this article

Sweezy, Paul M. "Baran, Paul Alexander (1910–1964)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000069> doi:10.1057/9780230226203.0096

The New Palgrave Dictionary of Economics Online

Barbon, Nicholas (1637/40 – ?1698)

Douglas Vickers
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Barbon, N.; Child, J.; money supply; North, D.; use value

Article

Nicholas Barbon, son of Praisegod Barbon, a London leather merchant, was born in 1637 (or 1640), and after studying medicine at Leyden and Utrecht and taking the MD at Utrecht in 1661, was admitted an Honorary Fellow of the College of Physicians at London in 1664. He was elected a Member of Parliament in 1690 and 1695. His successful career in various mercantile activities is reported in the autobiography of Roger North, the brother, biographer, and co-author of Sir Dudley North. He was engaged in the building trade in London following the great fire of 1666, and in 1685 he published a pamphlet *Apology for the Builder: or a Discourse showing the Cause and Effects of the Increase of Building*. In 1681 he established the first fire insurance company, and in 1684 published an *Account of two insurance offices*. Barbon also established a large financial venture in banking. With John Asgill he operated a land bank in 1695 and in the same year published *An Account of the Land Bank, showing the design and manner of the settlement*, and prepared a scheme for a national land bank which did not, however, come into existence.

Barbon's place in the history of economics is due to his *Discourse of Trade* (1690) and his more important *Discourse concerning coining the new money lighter: An answer to Mr Locke's Considerations about raising the value of money*. Taking the same position as Josiah Child and arguing, against Locke, for a legal reduction of the maximum rate of interest, he published in 1694 An *Answer to … reasons against reducing interest to four per cent*. His argument against trade restrictions and for international free trade principles places him in the front rank of anticipators of the doctrines that developed in the following century. He exhibited clearly the connection between the supply of money and the effective level of trade. Against the proposals to recoin the currency at the old standard he pointed out the potential deflationary effects of the reduction in the money supply that would result, ‘the consequence whereof will be that trade will be at a stand’.

Barbon's concern with the ‘disorder … that attends a nation that want money to drive their trade and

commerce' and the 'prejudice to the state by making money scarce' led him to argue, in contexts that elevated to priority the functional significance of money, that 'it is not absolutely necessary that money should be made of gold or silver'. 'Banks of credit … are of great advantage to trade.' 'Money is the instrument and measure of commerce and not silver.' Barbon held a supply and demand theory of market price, based on a logically prior notion of use values, and what he called 'time and place' value. He argued that 'interest is the rent of stock and is the same as the rent of land', claiming that a lower interest rate would raise capital values, indirectly by remedying 'the decay of trade' and directly by increasing the capitalized value of income streams.

Consumption expenditures, Barbon argued, provided employment. In his argument that 'prodigality is a vice that is prejudicial to the man but not to trade … covetousness is a vice prejudicial to both man and trade', he anticipated the prodigality and employment-creating expenditure argument of the following century.

Selected works

1690. *A Discourse of Trade*. London, Ed. J.H. Hollander, Baltimore: Johns Hopkins University Reprint, 1905.

1696. *A Discourse concerning coining the new money lighter*. London.

Bibliography

Letwin, W. 1963. *The Origins of Scientific Economics: English Economic Thought 1660–1776*. London: Methuen.

Vickers, D. 1959. *Studies in the Theory of Money 1690–1776*. Philadelphia: Chilton.

How to cite this article

Vickers, Douglas. "Barbon, Nicholas (1637/40–?1698)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000070> doi:10.1057/9780230226203.0097

The New Palgrave Dictionary of Economics Online

bargaining

Roberto Serrano
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article is a survey on bargaining theory. The focus is the game theoretic approach to bargaining, both on its axiomatic and strategic counterparts. The application of bargaining theory to large markets and its connections with competitive allocations are also discussed.

Keywords

asymmetric information; bargaining; Bayesian equilibrium; Edgeworth, F.; game theory; Harsanyi, J.; Hicks, J.; incomplete information; independence of irrelevant alternatives; Kalai–Smorodinsky bargaining solution; Morgenstern, O.; Nash equilibrium; Nash solution; Nash's demand game; Pareto efficiency; rational behaviour; risk aversion; subgame perfect equilibrium; ultimatum game; Von Neumann, J.; Walrasian outcomes; Zeuthen, F.

Article

In its simplest definition, ‘bargaining’ is a socio-economic phenomenon involving two parties, who can cooperate towards the creation of a commonly desirable surplus, over whose distribution the parties are in conflict.

The nature of the cooperation in the agreement and the relative positions of the two parties in the status quo before agreement takes place will influence the way in which the created surplus is divided. Many social, political and economic interactions of relevance fit this definition: a buyer and a seller trying to transact a good for money, a firm and a union sitting at the negotiation table to sign a labour contract, a couple deciding how to split the intra-household chores, two unfriendly countries trying to reach a lasting peace agreement, or out-of-court negotiations between two litigating parties.

In all these cases three basic ingredients are present: (*a*) the status quo, or the disagreement point, that is, the arrangement that is expected to prevail if an agreement is not reached; (*b*) the presence of mutual gains from cooperation; and (*c*) the multiplicity of possible cooperative arrangements, which split the

resulting surplus in different ways.

If the situation involves more than two parties, matters are different, as set out in von Neumann and Morgenstern (1944). Indeed, in addition to the possibilities already identified of either disagreement or agreement among all parties, it is conceivable that an agreement be reached among only some of the parties. In multilateral settings, we are therefore led to distinguish pure bargaining problems, in which partial agreements of this kind are not possible because subcoalitions have no more power than individuals alone, from coalitional bargaining problems (or simply coalitional problems), in which partial agreements become a real issue in formulating threats and predicting outcomes. An example of a pure bargaining problem would be a round of talks among countries in order to reach an international trade treaty in which each country has veto power, whereas an example of a coalitional bargaining problem would be voting in legislatures. In this article we concentrate on pure bargaining problems, leaving the description of coalitional problems to other articles in the dictionary. We are likewise not concerned with the vast informal literature on bargaining, which conducts case studies and tries to teach bargaining skills for the ‘real world’ (for this purpose, the reader is referred to Raiffa, 1982).

Approaches to bargaining before game theory

Before the adoption of game theoretic techniques, economists deemed bargaining problems (also called bilateral monopolies at the time) indeterminate. This was certainly the position adopted by important economic theorists, including Edgeworth (1881) and Hicks (1932). More specifically, it was believed that the solution to a bargaining problem must satisfy both individual rationality and collective rationality properties: the former means that neither party should end up worse than in the status quo and the latter refers to Pareto efficiency. Typically, the set of individually rational and Pareto-efficient agreements is very large in a bargaining problem, and these theorists were inclined to believe that theoretical arguments could go no further than this in obtaining a prediction. To be able to obtain such a prediction, one would have to rely on extra-economic variables, such as the bargaining power and abilities of either party, their state of mind in negotiations, their religious beliefs, the weather and so on.

A precursor to the game theoretic study of bargaining, at least in its attempt to provide a more determinate prediction, is the analysis of Zeuthen (1930). This Danish economist formulated a principle whereby the solution to a bargaining problem was dictated by the two parties’ risk attitudes (given the probability of breakdown of negotiations following the adoption of a tough position at the bargaining table). The reader is referred to Harsanyi (1987) for a version of Zeuthen's principle and its connection with Nash's bargaining theory. The remainder of this article deals with game theoretic approaches to bargaining.

The axiomatic theory of bargaining

Nash (1950) and Nash (1953) are seminal papers that constitute the birth of the formal theory of bargaining. Two assumptions are central in Nash's theory. First, bargainers are assumed to be fully rational individuals, and the theory is intended to yield predictions based exclusively on data relevant to them (in particular, the agents are equally skilful in negotiations, and the other extraneous factors mentioned above do not play a role).

Second, a bargaining problem is represented as a pair (S, d) in the utility space, where *S* is a compact and convex subset of $\mathbb{R}^2$ – the feasible set of utility pairs – and $d \in \mathbb{R}^2$ is the disagreement utility point. Compactness follows from standard assumptions such as closed productions sets and bounded factor endowments, and convexity is obtained if one uses expected utility and lotteries over outcomes are allowed. Also, the set *S* must include points that dominate the disagreement point, that is, there is a positive surplus to be enjoyed if agreement is reached and the question is how this surplus should be divided. As in most of game theory, by 'utility' we mean von Neumann–Morgenstern expected utility; there may be underlying uncertainty, perhaps related to the probability of breakdown of negotiations. We shall normalize the disagreement utilities to 0 (this is without loss of generality if one uses expected utility because any positive affine transformation of utility functions represents the same preferences over lotteries). The resulting bargaining problem is called a normalized problem.

With this second assumption, Nash is implying that all information relevant to the solution of the problem must be subsumed in the pair (S, d). In other words, two bargaining situations that may include distinct details ought to be solved in the same way if both reduce to the same pair (S, d) in utility terms. In spite of this, it is sometimes convenient to distinguish between feasible utility pairs (points in *S*) and feasible outcomes in physical terms (such as the portions of a pie to be created after agreement).

Following the two papers by Nash (1950; 1953), bargaining theory is divided into two branches, the so-called axiomatic and strategic theories. The axiomatic theory, born with Nash (1950), which most authors identify with a normative approach to bargaining, proposes a number of properties that a solution to any bargaining problem should have, and proceeds to identify the solution that agrees with those principles. Meanwhile, the strategic theory, initiated in Nash (1953), is its positive counterpart: the usual approach here is the exact specification of the details of negotiation (timing of moves, information available, commitment devices, outside options and threats) and the identification of the behaviour that would occur in those negotiation protocols. Thus, while the axiomatic theory stresses how bargaining *should* be resolved between rational parties according to some desirable principles, the strategic theory describes how bargaining *could* evolve in a non-cooperative extensive form in the presence of common knowledge of rationality. Interestingly, the two theories connect and complement one another.

The Nash bargaining solution

The first contribution to axiomatic bargaining theory was made by John Nash in his path-breaking paper published in 1950. Nash wrote it as a term paper in an international trade course that he was taking as an undergraduate at Carnegie, at the age of 17. At the request of his Carnegie economics professor, Nash mailed his term paper to John von Neumann, who had just published his monumental book with Oskar Morgenstern. John von Neumann may not have paid enough attention to a paper sent by an undergraduate at a different university, and nothing happened with the paper until Nash arrived in Princeton to begin studying for his Ph.D. in mathematics.

According to Nash (1950), a solution to bargaining problems is simply a function that assigns to each normalized utility possibility set *S* one of its feasible points (recall that the normalization of the disagreement utilities has already been performed). The interpretation is that the solution dictates a specific agreement to each possible bargaining situation. Examples of solutions are: (*a*) the disagreement solution, which assigns to each normalized bargaining problem the point (0,0), a rather pessimistic

solution; and (*b*) the dictatorial solution with bargainer 1 as the dictator, which assigns the point in the Pareto frontier of the utility possibility set in which agent 2 receives 0 utility. Surely, neither of these solutions looks very appealing: while the former is not Pareto efficient because it does not exploit the gains from cooperation associated with an agreement, the latter violates the most basic fairness principle by being so asymmetric.

Nash (1950) proceeds by proposing four desirable properties that a solution to bargaining problems should have.

1. 1. *Scale invariance or independence of equivalent utility representations*. Since the bargaining problem is formulated in von Neumann–Morgenstern utilities, if utility functions are re-scaled but they represent the same preferences, the solution should be re-scaled in the same fashion. That is, no fundamental change in the recommended agreement will happen following a re-normalization of utility functions; the solution will simply re-scale utilities accordingly.
2. 2. *Symmetry*. If a bargaining problem is symmetric with respect to the 45 degree line, the solution must pick a point on it: in a bargaining situation in which each of the threats made by one bargainer can be countered by the other with exactly the same threat, the two should be equally treated by the solution. This axiom is sometimes called ‘equal treatment of equals’ and it ensures that the solution yields ‘fair’ outcomes.
3. 3. *Pareto efficiency*. The solution should pick a point of the Pareto frontier. As elsewhere in welfare economics, efficiency is the basic ingredient of a normative approach to bargaining; negotiations should yield an efficient outcome in which all gains from cooperation are exploited.
4. 4. *Independence of irrelevant alternatives (IIA)*. Suppose a solution picks a point from a given normalized bargaining problem. Consider now a new normalized problem, a subset of the original, but containing the point selected earlier by the solution. Then, the solution must still assign the same point. That is, the solution should be independent of ‘irrelevant’ alternatives: as in a constrained optimization programme, the deleted alternatives are deemed irrelevant because they were not chosen when they were present, so their absence should not alter the recommended agreement.

With the aid of these four axioms, Nash (1950) proves the following result:

Theorem 1: There is a unique solution to bargaining problems that satisfies properties (1–4): it is the one that assigns to each normalized bargaining problem the point that maximizes the product of utilities of the two bargainers.

Today we refer to this solution as the ‘Nash solution’. Although some of the axioms have been the centre of some controversy – especially his fourth, IIA, axiom – the Nash solution has remained as the fundamental piece of this theory, and its use in applications is pervasive.

Some features of the Nash solution ought to be emphasized. First, the theory can be extended to the multilateral case, in which there are $n \geq 3$ parties present in bargaining: in a multilateral problem, it continues to be true that the unique solution that satisfies (1–4) is the one prescribing that agreement in which the product of utilities is maximized. See Lensberg (1988) for an important alternative axiomatization.

Second, the theory is independent of the details of the negotiation-specific protocols, since it is

formulated directly in the space of utilities. In particular, it can be applied to problems where the utilities are derived from only one good or issue, as well as those where utility comes from multiple goods or issues.

Third, perhaps surprisingly because risk is not explicitly part of Nash's story, it is worth noting that the Nash solution punishes risk aversion. All other things equal, it will award a lower portion of the surplus to a risk-averse agent. This captures an old intuition in previous literature that risk aversion is detrimental to a bargainer: afraid of the bargaining breakdown, the more risk-averse a person is, the more he will concede in the final agreement. For example, suppose agents are bargaining over how to split a surplus of size 1. Let the utility functions be as follows: $u_1(x_1) = x_1^{\alpha}$ for $0 < \alpha \leq 1$, and $u_2(x_2) = x_2$, where x_1 and x_2 are the non-negative shares of the surplus, which add up to 1. The reader can calculate that the Pareto frontier of the utility possibility set corresponds to the agreements satisfying the equation $u_1^{1/\alpha} + u_2 = 1$. Therefore, the Nash solution awards the utility vector $(u_1^*, u_2^*) = ((\frac{\alpha}{\alpha+1})^{\alpha}, \frac{1}{\alpha+1})$, corresponding to shares of the surplus $(x_1, x_2) = (\frac{\alpha}{\alpha+1}, \frac{1}{\alpha+1})$. Note how the smaller α is, the more risk-averse bargainer 1 is.

Fourth, Zeuthen's principle turns out to be related to the Nash solution (see Harsanyi, 1987): in identifying the bargainer who must concede next, the Nash product of utilities of the two proposals plays a role. See Rubinstein, Safra and Thomson (1992) for a related novel interpretation of the Nash solution.

Fifth, the family of asymmetric Nash solutions has also been used in the literature as a way to capture unequal bargaining powers. If the bargaining power of player *i* is $\beta_i \in [0, 1]$, $\Sigma_i \beta_i = 1$, the asymmetric Nash solution with weights (β_1, β_2) is defined as the function that assigns to each normalized bargaining problem the point where $u_1^{\beta_1} u_2^{\beta_2}$ is maximized.

The Kalai – Smorodinsky bargaining solution

Several researchers have criticized some of Nash's axioms, IIA especially. To see why, think of the following example, which begins with the consideration of a symmetric right-angled triangle *S* with legs of length 1. Clearly, efficiency and symmetry alone determine that the solution must be the point (1/2,1/2). Next, chop off the top part of the triangle to get a problem $T \subset S$, in which all points where $u_2 > 1/2$ have been deleted. By IIA, the Nash solution applied to the problem *T* is still the point (1/2,1/2).

Kalai and Smorodinsky (1975) propose to retain the first three axioms of Nash's, but drop IIA. Instead, they propose an individual monotonicity axiom. To understand it, let $a_i(S)$ be the highest utility that agent *i* can achieve in the normalized problem *S*, and let us call it agent *i*'s aspiration level. Let $a(S) = (a_1(S), a_2(S))$ be the utopia point, typically not feasible.

1. 5. *Individual monotonicity*. If $T \subset S$ are two normalized problems, and $a_j(T) = a_j(S)$, the solution must award *i* a utility in *S* at least as high as in *T*.

We can now state the Kalai–Smorodinsky theorem:

Theorem 2: There is a unique solution to bargaining problems that satisfies properties (1, 2, 3, 5): it is the one that assigns to each normalized bargaining problem the intersection point of the Pareto frontier and the straight line segment connecting 0 and the utopia point.

Note how the Kalai–Smorodinsky solution awards the point (2/3,1/3) to the problem *T* of the beginning of this subsection. In general, while the Nash solution pays attention to local arguments (it picks out the point of the smooth Pareto frontier where the utility elasticity $(du_2/u_2)/(du_1/u_1)$ is 1), the Kalai–Smorodinsky solution is mostly driven by 'global' considerations, such as the highest utility each bargainer can obtain in the problem.

Other solutions

Although the two major axiomatic solutions are Nash's and Kalai–Smorodinsky's, authors have derived a plethora of other solutions also axiomatically (see, for example, Thomson, 1994, for an excellent survey). Among them, one should perhaps mention the egalitarian solution, which picks out the point of the Pareto frontier where utilities are equal. This is based on very different principles, much more tied to ethics of a certain kind and less to the principles governing bargaining between two rational individuals. In particular, note how it is not invariant to equivalent utility representations, because of the strong interpersonal comparisons of utilities that it performs.

The strategic theory of bargaining

Now we are interested in specifying the details of negotiations. Thus, while we may lose the generality of the axiomatic approach, our goal is to study reasonable procedures and identify rational behaviour in them. For this and the next section, some major references include Osborne and Rubinstein (1990) and Binmore, Osborne and Rubinstein (1992).

Nash's demand game

Nash (1953) introduces the first bargaining model expressed as a non-cooperative game. Nash's demand game, as it is often called, captures in crude form the force of commitment in bargaining. Both bargainers must demand simultaneously a utility level. If the pair of utilities is feasible, it is implemented; otherwise, there is disagreement and both receive 0. This game admits a continuum of Nash equilibrium outcomes, including every point of the Pareto frontier, as well as disagreement. The first message that emerges from Nash's demand game is the indeterminacy of equilibrium outcomes, commonplace in non-cooperative game theory. In the same paper, advancing ideas that would be developed a couple of decades later, Nash proposed a refinement of the Nash equilibrium concept based on the possibility of uncertainty around the true feasible set. The result was a selection of one Nash equilibrium outcome, which converges to the Nash solution agreement as uncertainty vanishes.

The model just described is referred to as Nash's demand game with fixed threats: following an incompatible pair of demands, the outcome is the fixed disagreement point. Nash (1953) also analysed a variable threats model. In it, the stage of simultaneous demands is preceded by another stage, in which bargainers choose threats. Given a pair of threats chosen in the first stage, the refinement argument is

used to obtain the Nash solution of the induced problem in the ensuing subgame (where the threats determine an endogenous disagreement point). Solving the entire game is possible by backward induction, appealing to logic similar to that in von Neumann's minimax theorem; see Abreu and Pearce (2002) for a connection between the variable threats model and repeated games.

The alternating offers bargaining procedure

The following game elegantly describes a stylized protocol of negotiations over time. It was studied by Stahl (1972) under the assumption of an exogenous deadline (finite horizon game) and by Rubinstein (1982) in the absence of a deadline (infinite horizon game). Players 1 and 2 are bargaining over a surplus of size 1. The bargaining protocol is one of alternating offers. In period 0, player 1 begins by making a proposal, a division of the surplus, say $(x, 1-x)$, where $0 \leq x \leq 1$ represents the part of the surplus that she demands for herself. Player 2 can then either accept or reject this proposal. If he accepts, the proposal is implemented; if he rejects, a period must elapse for them to come back to the negotiation table, and at that time (period 1) the roles are reversed so that player 2 will make a new proposal $(y, 1-y)$, where $0 \leq y \leq 1$ is the fraction of surplus that he offers to player 1. Player 1 must then either accept the new proposal, in which case bargaining ends with $(y, 1-y)$ as the agreement, or reject it, in which case a period must elapse before player 1 makes a new proposal. In period 2, player 1 proposes $(z, 1-z)$, to which player 2 must respond, and so on. The *T*-period finite horizon game imposes the disagreement outcome, with zero payoffs, after *T* proposals have been rejected. On the other hand, in the infinite horizon version, there is always a new proposal in the next period after a proposal is rejected. Both players discount the future at a constant rate. Let $\delta \in [0, 1)$ be the per period discount factor. To simplify, let us assume that utility is linear in shares of the surplus. Therefore, from a share *x* agreed in period *t*, a player derives a utility of $\delta^{t-1}x$. Note how utility is increasing in the share of the surplus (monotonicity) and decreasing in the delay with which the agreement takes place (impatience).

A strategy for a player is a complete contingent plan of action to play the game. That is, a strategy specifies a feasible action every time a player is called upon to act in the game. In a dynamic game, Nash equilibrium does little to restrict the set of predictions: for example, it can be shown that in the alternating offers games, any agreement $(x, 1-x)$ in any period *t*, $0 \leq t \leq T < \infty$, can be supported by a Nash equilibrium; disagreement is also a Nash equilibrium outcome.

The prediction that game theory gives in a dynamic game of complete information is typically based on finding its subgame perfect equilibria. A subgame perfect equilibrium (SPE) in a two-player game is a pair of strategies, one for each player, such that the behaviour specified by them is a best response to each other at every point in time (not only at the beginning of the game). By stipulating that players must choose a best response to each other at every instance that they are supposed to act, SPE rules out incredible threats: that is, at an SPE players have an incentive to carry out the threat implicit in their equilibrium strategy because it is one of the best responses to the behaviour they expect the other player to follow at that point.

In the alternating offers games described above, there is a unique SPE, in both the finite and the infinite horizon versions. The SPE in the finite horizon game is found by backward induction. For example, in the one-period game, the so-called ultimatum game, the unique SPE outcome is the agreement on the split (1,0): since the outcome of a rejection is disagreement, the responder will surely accept any share of

$\in > 0$, which implies that in equilibrium the proposer ends up taking the entire surplus. Using this intuition, one can show that the outcome of the two-period game is the immediate agreement on the split $(1-\delta, \delta)$: anticipating that if negotiations get to the final period, player 2 (the proposer in that final period) will take the entire surplus, player 1 persuades him not to get there simply by offering him the present discounted value of the entire surplus, that is, δ , while she takes the rest. This logic continues and can be extended to any finite horizon. The sequence of SPE outcomes so obtained as the deadline $T \to \infty$ is shown to converge to the unique SPE of the infinite horizon game. This game, more challenging to solve since one cannot go to its last period to begin inducting backwards, was studied in Rubinstein (1982). We proceed to state its main theorem and discuss the properties of the equilibrium (see Shaked and Sutton, 1984, for a simple proof).

Theorem 3: Consider the infinite horizon game of alternating offers, in which both players discount the future at a per period rate of $\delta \in [0, 1)$. There exists a unique SPE of this game: it prescribes immediate agreement on the division $\left(\frac{1}{1+\delta}, \frac{\delta}{1+\delta}\right)$.

The first salient prediction of the equilibrium is that there will not be any delay in reaching an agreement. Complete information – each player knows the other player's preferences – and the simple structure of the game are key factors to explain this.

The equilibrium awards an advantage to the proposer, as expressed by the discount factor: note how the proposer's share exceeds the responder's by a factor of $1/\delta$. Given impatience, having to respond to a proposal puts an agent in a delicate position, since rejecting the offer entails time wasted until the next round of negotiations. This is the source of the proposer's advantage. Of course, this advantage is larger, the larger the impatience of the responder: note how if $\delta=0$ (extreme impatience), the equilibrium awards all the surplus to the proposer because her offer is virtually an ultimatum; on the other hand, as $\delta \to 1$, the first-mover advantage disappears and the equilibrium tends to an equal split of the surplus.

To understand how the equilibrium works and in particular how the threats employed in it are credible, consider the SPE strategies. Both players use the same strategy, and it is the following: as a proposer, each player always asks for $1/(1+\delta)$ and offers $\delta/(1+\delta)$ to the other party; as a responder, a player accepts an offer as long as the share offered to the responder is at least $\delta/(1+\delta)$. Note how rejecting a share lower than $\delta/(1+\delta)$ is credible, in that its consequence, according to the equilibrium strategies, is to agree in the next period on a split that awards the rejecting player a share of $1/(1+\delta)$, whose present discounted value at the time the rejection occurs is exactly $\delta/(1+\delta)$.

To appreciate the difference from Nash equilibrium, let us argue, for example, that the split (0,1) cannot happen in an SPE. This agreement happens in a Nash equilibrium, supported by strategies that ask player 1 to offer the whole pie to player 2, and player 2 to reject any other offer. However, the threat embodied in player 2's strategy is not credible: when confronted with an offer $(\in, 1-\in)$ for $\delta < 1-\in < 1$, player 2 will have to accept it, contradicting his strategy. Can the reader argue why the Nash equilibrium split (1,0) is not an SPE outcome either (because to do so one would need to employ non-credible threats)? Rubinstein (1982) shows that the same non-credible threats are associated with any division of the pie other than the one identified in the theorem.

The Rubinstein–Stahl alternating offers game provides an elegant model of how negotiations may take place over time, and its applications are numerous, including bargaining problems pertaining to international trade, industrial organization, or political economy. However, unlike Nash's axiomatic theory, its predictions are sensitive to details. This is no doubt one of its strengths because one can

calibrate how those details may influence the theory's prediction, but it is also its weakness in terms of lack of robustness in predictive power.

Incomplete information

In a static framework, Chatterjee and Samuelson (1983) study a double auction. A buyer and a seller are trying to transact a good. Each proposes a price, and trade takes place at the average of the two prices if and only if the buyer's price exceeds the seller's. Each trader knows his own valuation for the good. However, there is incomplete information on each side concerning the other side's valuation. It can be shown that in any equilibrium of this game there are inefficiencies: given certain *ex post* valuations of buyer and seller, there should be trade, yet it is precluded because of incomplete information, which leads traders to play ‘too tough’.

Let us now turn to bargaining over time. As pointed out above, one prediction of the Rubinstein–Stahl model is immediate agreement. This may clash with casual observation; one may simply note the existence of strikes, lockouts and long periods of disagreement in many actual negotiations. As a consequence, researchers have suggested the construction of models in which inefficiencies, in the form of delay in agreement, occur in equilibrium. The main feature of bargaining models with this property is incomplete information. (For delay in agreement that does not rely on incomplete information, see Fernandez and Glazer, 1991, Avery and Zemsky, 1994, and Busch and Wen, 1995.)

If parties do not know each other's preferences (impatience rate, per period fixed cost of hiring a lawyer, profitability of the agreement, and so on), the actions taken by the parties in the bargaining game may be intended to elicit some of the information that they do not have, or perhaps to reveal or misrepresent some of the information privately held.

One technical remark is in order. The typical approach is to reduce the uncertainty to a game of imperfect information through the specification of types in the sense of Harsanyi (1967–8). In such games, SPE no longer constitutes an appropriate refinement of Nash equilibrium. The relevant equilibrium notions are perfect Bayesian equilibrium and sequential equilibrium, and in them the off-equilibrium path beliefs play an important role in sustaining outcomes. Moreover, these concepts are often incapable of yielding a determinate prediction in many games, and authors have in these cases resorted to further refinements. One problem of the refinements literature, however, is that it lacks strong foundations. Often the successful use of a given refinement in a game is accompanied by a bizarre prediction when the same concept is used in other games. Therefore, one should interpret these findings as showing the possibilities that equilibrium can offer in these contexts, but the theory here is far from giving a determinate answer.

Rubinstein (1985) studies an alternating offers procedure in which there is one-sided incomplete information (that is, while player 1 has uncertainty regarding player 2's preferences, player 2 is fully informed). Suppose there are two types of player 2: one of them is ‘weaker’ than player 1, while the other is ‘stronger’ (in terms of impatience or per period costs). This game admits many equilibria, and they differ as a function of parameter configurations. There are pooling equilibria, in which an offer from player 1 is accepted immediately by both types of player 2. More relevant to the current discussion, there are also separating equilibria, in which player 1's offer is accepted by the weak type of player 2, while the strong type signals his true preferences by rejecting the offer and imposing delay in

equilibrium. These equilibria are also used to construct other equilibria with more periods of delay in agreement. Some authors (Gul and Sonnenschein, 1988) argue that long delays in equilibrium are the product of strong non-stationary behaviour (that is, a player behaves very differently in and out of equilibrium, as a function of changes in his beliefs). They show that imposing stationary behaviour limits the delay in agreement quite significantly. One advantage of stationary equilibria is their simplicity, but one problem with them is that they impose stationary beliefs (players hold beliefs that are independent of the history of play).

The analysis is simpler and multiplicity of equilibrium is less of a problem in games in which the uninformed party makes all the offers. Consider, for example, a version of the model in Sobel and Takahashi (1983). The two players are a firm and a union. The firm is fully informed, while the union does not know the true profitability of the firm. The union makes all offers in these wage negotiations, and there is discounting across periods. In equilibrium, different types of the firm accept offers at different points in time: firms whose profitability is not very high can afford to reject the first high wage offers made by the union to signal their private information, while very profitable firms cannot because delay in agreement hurts them too much.

Most papers have studied the case of private values asymmetric information (if a player knows her type, she knows her preferences), although the correlated values case has also been analysed (where knowing one's type is not sufficient to know one's utility function); see Evans (1989) and Vincent (1989). The case of two-sided asymmetric information, in which neither party is fully informed, has been treated, for example, in Watson (1998). In all these results, one is able to find equilibria with significant delay in agreement, implying consequent inefficiencies. Uncertainty may also be about the rationality of the opponent: for example, one may be bargaining with a ‘behavioural type’ who has an unknown threshold below which he will reject all proposals (see Abreu and Gul, 2000).

A more general approach is adopted by studies of mechanism design. The focus is not simply on explaining delay as an equilibrium phenomenon in a given extensive form. Rather, the question is whether inefficiencies are a consequence of equilibrium behaviour in any bilateral bargaining game with incomplete information. The classic contribution to this problem is the paper by Myerson and Satterthwaite (1983). In a bilateral trading problem in which there is two-sided private values asymmetric information and the types of each trader are drawn independently from overlapping intervals, there does not exist any budget-balanced mechanism satisfying incentive compatibility, interim individual rationality and *ex post* efficiency. All these are desirable properties for a trading mechanism. Budget balance implies that payoffs cannot be increased with outside funds. Incentive compatibility requires that each type has no incentive to misrepresent his information. Interim individual rationality means that no type can be worse off trading than not trading. Finally, *ex post* efficiency imposes that trade takes place if and only if positive gains from trade exist. This impossibility result is a landmark of the limitations of bargaining under incomplete information, and has generated an important literature that explores ways to overcome it (see for example Gresik and Satterthwaite, 1989, and Satterthwaite and Williams, 1989).

Indivisibilities in the units

One important way in which Rubinstein's result is not robust happens when there is only a finite set of

possible offers to be made (see van Damme, Selten and Winter, 1990, and Muthoo, 1991). Indivisibilities make it impossible for an exact adjustment of offers to leave the responder indifferent; as a result, multiple and inefficient equilibria appear. The issue concerns how fine the grid of possible instantaneous offers is with respect to the time grid in which bargaining takes place. If the former is finer than the latter, Rubinstein's uniqueness goes through; otherwise it does not. There will be circumstances for which one or the other specification of negotiation rules will be more appropriate.

Multi-issue bargaining

The following preliminary observation is worth making: if offers are made in utility space or all issues must be bundled in every offer, Rubinstein's result obtains. Thus, the literature on multi-issue bargaining has looked at procedures that depart from these assumptions.
The first generation of papers with multiple issues assumed that the agenda – that is, the order in which the different issues are brought to the table – was exogenously given. Since each issue is bargained over one at a time, Rubinstein's uniqueness and efficiency result obtains, simply proceeding by backward induction on the issues. Fershtman (1990; 2000) and Busch and Horstmann (1997) study such games, from which one learns the comparative statics of equilibrium when agendas are exogenously fixed. The next group of papers studies more realistic games where the agenda is chosen endogenously by the players. The main lesson from this line of work is that restricting the issues that a proposer can bring to the table is a source of inefficiencies. Inderst (2000) and In and Serrano (2003) study a procedure where agenda is totally unrestricted, that is, the proposer can make offers on any subset of remaining issues and, by exploiting trade-offs in the marginal rates of substitution between issues, Rubinstein's efficiency result is also found. In contrast, Lang and Rosenthal (2001) and In and Serrano (2004) construct multiple and inefficient equilibria (including those with arbitrarily long delay in agreement) when agenda restrictions are imposed. Finally, Weinberger (2000) considers multi-issue bargaining when the responder can accept selectively subsets of proposals and also finds inefficiencies if issues are indivisible.

Multilateral bargaining

Even within the case of pure bargaining problems, one needs to make a distinction between different ways to model negotiations. The first extension of the Rubinstein game to this case is due to Shaked, as reported in Osborne and Rubinstein (1990, p. 63); see also Herrero (1985). Today we refer to the Shaked/Herrero game as the ‘unanimity game’. In it, one of the players, say player 1, begins by making a public proposal to the others. A proposal is a division of the unit of surplus available when agreement is reached. Players 2,•…,•*n* then must accept or reject this proposal. If all agree, it is implemented immediately. If at least one of them rejects it, time elapses and in the next period another player, say player 2, will make a new proposal, and so on. Note how these rules reduce to Rubinstein's when there are only two players. However, the prediction emerging from this game is dramatically different. For values of the discount factor that are sufficiently high (if $\delta \geq 1/(n-1)$), every feasible agreement can be supported by an SPE and, in addition, equilibria with an arbitrary number of periods of delay in agreement show up. The intuition for this extreme result is that the unanimity required by the rules in

order to implement an agreement facilitates a plethora of equilibrium behaviours. For example, let us see how in the case of $n = 3$ it is possible to sustain an agreement where all the surplus goes to player 3. If player 2 rejects it, the same split will be repeated in the continuation, so it is pointless to reject it. If player 1 changes her proposal to try to obtain a gain, it will be rejected by that responder who in the proposal receives less than 1/2 (there must be at least one). This rejector can be bribed with receiving the entire surplus in the continuation, whose present discounted value is at least 1/2 (recall $\delta \geq 1/2$), thereby rendering his rejection credible. Of course, the choice of player 3 as the one receiving the entire surplus is entirely arbitrary and, therefore, one can see how extreme multiplicity of equilibrium is a phenomenon inherent to the unanimity game. This multiplicity relies on non-stationary strategies, as it can be shown that there is a unique stationary SPE.

An alternative extension of the Rubinstein rules to multilateral settings is given by exit games; see Jun (1987), Chae and Yang (1994), Krishna and Serrano (1996). As an illustration, let us describe the negotiation rules of the Krishna–Serrano game. Player 1 makes a public proposal, a division of the surplus, and the others must respond to it. Those who accept it leave the game with the shares awarded by the proposer, while the rejectors continue to bargain with the proposer over the part of the surplus that has not been committed to any player. A new proposal comes from one of the rejectors, and so on. These rules also reduce to Rubinstein's if $n = 2$, but now the possibility of exiting the game by accepting a proposal has important implications for the predictive power of the theory. Indeed, Rubinstein's uniqueness is restored and the equilibrium found inherits the properties of Rubinstein's, including its immediate agreement and the proposer's advantage (the equilibrium shares are $1/[1+(n-1)\delta]$ for the proposer and $\delta/[1+(n-1)\delta]$ for each responder). Note how, given that the others accept, each responder is *de facto* immersed in a two-player Rubinstein game, so in equilibrium he receives a share that makes him exactly indifferent between accepting and rejecting: this explains the ratio $1/\delta$ between the proposer's and each responder's equilibrium shares. The sensitivity of the result to the exact specification of details is emphasized in other papers. Vannetelbosch (1999) shows that uniqueness obtains in the exit game even with a notion of rationalizability, weaker than SPE; and Huang (2002) establishes that uniqueness is still the result in a model that combines unanimity and exit, since offers can be made both conditional and unconditional to each responder. Baliga and Serrano (1995; 2001) introduce imperfect information in the unanimity and exit games (offers are not public, but made in personalized envelopes), and multiplicity is found in both, based on multiple off-equilibrium path beliefs. Merlo and Wilson (1995) propose a stochastic specification and also find uniqueness of the equilibrium outcome. In a model often used in political applications, Baron and Ferejohn (1989) study a procedure with random proposers in which the proposals are adopted if approved by simple majority (between the unanimity and exit procedures described).

Bargaining and markets

Bargaining theory provides a natural approach to understand how prices may emerge in markets as a consequence of the direct interaction of agents. One can characterize the outcomes of models in which the interactions of small groups of agents are formulated as bargaining games, and compare them with market outcomes such as competitive equilibrium allocations. If a connection between the two is found, one is giving an answer to the long-standing question of the origin of competitive equilibrium prices

without having to resort to the story of the Walrasian auctioneer. If not, one can learn the importance of the frictions in the model that may be preventing such a connection. Both kinds of results are valuable for economic theory.

Small markets

Models have been explored in which two agents are bargaining, but at least one of them may have an outside option (see Binmore, Shaked and Sutton, 1988). Thus, the bargaining pair is part of a larger economic context, which is not explicitly modelled. In the simplest specification, uniqueness and efficiency of the equilibrium is found. In the equilibrium, the outside option is used if it pays better than the Rubinstein equilibrium; otherwise it is ignored. Jehiel and Moldovanu (1995) show that delays may be part of the equilibrium when the agreement between a seller and several buyers is subject to externalities among the buyers: a buyer may have an incentive to reject an offer in the hope of making a different buyer accept the next offer and free-ride from that agreement. In general, these markets involving a small number of agents do not yield competitive allocations because market power is retained by some traders (see Rubinstein and Wolinsky, 1990).

Large markets under complete information

The standard model assumes a continuum of agents who are matched at random, typically in pairs, to perform trade of commodities. If a pair of agents agrees on a trade, they break the match. In simpler models, all traders leave the market after they trade once. In the more general models agents may choose either to leave and consume, or to stay in the market to be matched anew. Some authors have studied steady-state versions, in which the measure of traders leaving the market every period is offset exactly by the same measure of agents entering the market. In contrast, non-steady state models do not keep the measure of active traders constant (one prominent class of non-steady state models is that of one-time entry, in which after the initial period there is no new entry; certain transacting agents exit every period, so the market size dwindles over time). The analysis has been performed with discounting (where δ is the common discount factor that is thought of as being near 1) or without it: in both cases the idea is to describe frictionless or almost frictionless conditions (for example, Muthoo, 1993, considers several frictions and the outcomes that result when some, but not all, of them are removed).

The first models were introduced by Diamond and Maskin (1979), Diamond (1981), and Mortensen (1982), and they used the Nash solution to solve each bilateral bargaining encounter. Later each pairwise meeting has been modelled by adopting a procedure from the strategic theory.

The most general results in this area are provided by Gale (1986a; 1986b; 1986c; 1987). First, in a partial equilibrium set-up, a market for an indivisible good is analysed in Gale (1987), under both steady state and non-steady-state assumptions. The result is that all equilibrium outcomes yield trade at the competitive price when discounting is small: in all equilibria trade tends to take place at only one price, and that price must be the competitive price because it is the one that maximizes each trader's expected surplus. This generalizes a result of Binmore and Herrero (1988) and clarifies an earlier claim made by Rubinstein and Wolinsky (1985). Rubinstein and Wolinsky analysed the market in steady state and claimed that the market outcome was different from the competitive one. Their claim is justified if one

measures the sets of traders in terms of the stocks present in the market, but Gale (1987) argues convincingly that, given the steady state imposed on the solution concept, it is the flow of agents into the market every period, not the total stock, that should comprise the relevant demand and supply curves. When this is taken into account, all prices are competitive because the measure of transacting sellers is the same as that of the transacting buyers.

In a more general model, Gale (1986a; 1986b; 1986c) studies an exchange economy with an arbitrary number of divisible goods. Now there is no discounting and agents can trade in as many periods as they wish before they leave the market place. Only after an agent rejects a proposal can he leave the market. Under a number of technical assumptions, Gale shows once again that all the equilibrium outcomes of his game are Walrasian:

Theorem 4: At every market equilibrium, each agent leaves the market with the bundle x_k with probability 1, where the list of such bundles is a Walrasian allocation of the economy.

Different versions of this result are proved in Gale (1986a; 1986c) and in Osborne and Rubinstein (1990). Also, Kunimoto and Serrano (2004) obtain the same result under substantially weaker assumptions on the economy, thereby emphasizing the robustness of the connection between the market equilibria of this decentralized exchange game and the Walrasian allocations of the economy. There are two key steps in this argument: first, one establishes that, since pairs are trading, pairwise efficiency obtains, which under some conditions leads to Pareto efficiency; and second, the equilibrium strategies imply budget balance so that each agent cannot end up with a bundle that is worth more than his initial endowment (given prices supporting the equilibrium allocation, already known to be efficient).

Dagan, Serrano and Volij (2000) also show a Walrasian result, but in their game the trading groups are coalitions of any finite size: in their proof, the force of the core equivalence theorem is exploited. One final comment is pertinent at this point. Some authors (for example, Gale, 2000) question the use of coalitions of any finite size in the trading procedure because the ‘large’ size of some of those groups seems to clash with the ‘decentralized’ spirit of these mechanisms. On the other hand, one can also argue that for the procedure to allow trade only in pairs, some market authority must be keeping track of this, making sure that coalitions of at least three agents are ‘illegal’. Both trading technologies capture appealing aspects of decentralization, depending on the circumstances, and the finding is that either one yields a robust connection with the teachings of general equilibrium theory in frictionless environments. This is one more instance of the celebrated equivalence principle: in models involving a large number of agents, game theoretic predictions tend to converge, under some conditions, to the set of competitive allocations.

Large markets under incomplete information

If the asymmetric information is of the private values type, the same equivalence result is obtained between equilibria of matching and bargaining models and Walrasian allocations. This message is found, for example, in Rustichini, Satterthwaite and Williams (1994), Gale (1987) and Serrano (2002). In the latter model, for instance, some non-Walrasian outcomes are still found in equilibrium, but they can be explained by features of the trading procedure that one could consider as frictions, such as a finite set of prices and finite sets of traders’ types.

The result is quite different when asymmetric information goes beyond private values. For example,

Wolinsky (1990) studies a market with pairwise meetings in which there is uncertainty regarding the true state of the world (which determines the true quality of the good being traded). Some traders know the state, while others do not, and there are uninformed traders among buyers and sellers (two-sided asymmetric information). The analysis is performed in steady state. To learn the true state, uninformed traders sample agents of the opposite side of the market. However, each additional meeting is costly due to discounting. The relevant question is whether information will be transmitted from the informed to the uninformed when discounting is removed. Wolinsky's answer is in the negative: as the discount factor $\delta \to 1$, a non-negligible fraction of uninformed traders transacts at a price that is not *ex post* individually rational. It follows that the equilibrium outcomes do not approximate those given by a fully revealing rational expectations equilibrium (REE). The reason for this result is that, while as $\delta \to 1$ sampling becomes cheaper and therefore each uninformed trader samples more agents, this is true on both sides, so that uninformed traders end up trying to learn from agents that are just as uninformed as they are. Serrano and Yosha (1993) overturn this result when asymmetric information is one-sided: in this case, although the noise force behind Wolinsky's result is not operative because of the absence of uninformed traders on one side, there is a negative force that works against learning, which is that misrepresenting information becomes cheaper for informed traders as $\delta \to 1$. The analysis in Serrano and Yosha's paper shows that, under steady state restrictions, the learning force is more powerful than the misrepresentation one, and convergence to REE is attained. Finally, Blouin and Serrano (2001) perform the analysis without the strong steady-state assumption, and show that with both information structures (one-sided and two-sided asymmetries) the result is negative: Wolinsky's noise force in the two-sided case continues to be crucial, while misrepresentation becomes very powerful in the one-sided model because of the lack of fresh uninformed traders. In these models, agents have no access to aggregate market signals; information is heavily restricted because agents observe only their own private history. It would be interesting to analyse other procedures where information may flow more easily.

See Also

- Nash program
- Shapley value

Bibliography

Abreu, D. and Gul, F. 2000. Bargaining and reputation. *Econometrica* 68, 85–118.

Abreu, D. and Pearce, D. 2002. Bargaining, reputation and equilibrium selection in repeated games. Mimeo. Princeton University.

Avery, C. and Zemsky, P. 1994. Money burning and multiple equilibria in bargaining. *Games and Economic Behavior* 7, 154–68.

Baliga, S. and Serrano, R. 1995. Multilateral bargaining with imperfect information. *Journal of*

Economic Theory 67, 578–89.

Baliga, S. and Serrano, R. 2001. Multilateral negotiations with private side-deals: a multiplicity example. *Economics Bulletin* 3(1), 1–7.

Baron, D. and Ferejohn, J. 1989. Bargaining in legislatures. *American Political Science Review* 83, 1181–206.

Binmore, K. and Herrero, M. 1988. Matching and bargaining in dynamic models. *Review of Economic Studies* 55, 17–31.

Binmore, K., Osborne, M. and Rubinstein, A. 1992. Non-cooperative models of bargaining. In *Handbook of Game Theory with Economic Applications*, vol. 1, ed. J. Aumann and S. Hart. New York: Elsevier.

Binmore, K., Shaked, A. and Sutton, J. 1988. An outside option experiment. *Quarterly Journal of Economics* 104, 753–70.

Blouin, M. and Serrano, R. 2001. A decentralized market with common values uncertainty: non-steady states. *Review of Economic Studies* 68, 323–46.

Busch, L.-A. and Horstmann, I. 1997. Bargaining frictions, bargaining procedures and implied costs in multiple-issue bargaining. *Economica* 64, 669–80.

Busch, L.-A. and Wen, Q. 1995. Perfect equilibria in a negotiation model. *Econometrica* 63, 545–65.

Chae, S. and Yang, J.-A. 1994. An *n*-person pure bargaining game. *Journal of Economic Theory* 62, 86–102.

Chatterjee, K. and Samuelson, W. 1983. Bargaining under incomplete information. *Operations Research* 31, 835–51.

Dagan, N., Serrano, R. and Volij, O. 2000. Bargaining, coalitions and competition. *Economic Theory* 15, 279–96.

Diamond, P. 1981. Mobility costs, frictional unemployment, and efficiency. *Journal of Political Economy* 89, 798–812.

Diamond, P. and Maskin, E. 1979. An equilibrium analysis of search and breach of contract I: steady states. *Bell Journal of Economics* 10, 282–316.

Edgeworth, F. 1881. Mathematical psychics. In *F. Y. Edgeworth's Mathematical Psychics and Further Papers on Political Economy*, ed. P. Newman. Oxford: Oxford University Press, 2003.

Evans, R. 1989. Sequential bargaining with correlated values. *Review of Economic Studies* 56, 499–510.

Fernandez, R. and Glazer, J. 1991. Striking for a bargain between two completely informed agents. *American Economic Review* 81, 240–52.

Fershtman, C. 1990. The importance of the agenda in bargaining. *Games and Economic Behavior* 2, 224–38.

Fershtman, C. 2000. A note on multi-issue two-sided bargaining: bilateral procedures. *Games and Economic Behavior* 30, 216–27.

Gale, D. 1986a. Bargaining and competition. Part I: characterization. *Econometrica* 54, 785–806.

Gale, D. 1986b. Bargaining and competition. Part II: existence. *Econometrica* 54, 807–18.

Gale, D. 1986c. A simple characterization of bargaining equilibrium in a large market without the assumption of dispersed characteristics. Working Paper 86–05. Philadelphia: CARESS, University of Pennsylvania.

Gale, D. 1987. Limit theorems for markets with sequential bargaining. *Journal of Economic Theory* 43, 20–54.

Gale, D. 2000. *Strategic Foundations of General Equilibrium – Dynamic Matching and Bargaining Games*. Cambridge: Cambridge University Press.

Gresik, T. and Satterthwaite, M. 1989. The rate at which a simple market converges to efficiency as the number of traders increases: an asymptotic result for optimal trading mechanisms. *Journal of Economic Theory* 48, 304–32.

Gul, F. and Sonnenschein, H. 1988. On delay in bargaining with one-sided uncertainty. *Econometrica* 56, 601–11.

Harsanyi, J. 1967–8. Games with incomplete information played by Bayesian players. Parts I, II and III. *Management Science* 14, 159–82, 320–34, 486–502.

Harsanyi, J. 1987. Bargaining. In *The New Palgrave Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Herrero, M. 1985. N-player bargaining and involuntary underemployment. In A strategic bargaining approach to market institutions. Ph.D. thesis. London: London School of Economics.

Hicks, J. 1932. *The Theory of Wages*. London: Macmillan; New York: St. Martin Press, 1963.

Huang, C.-Y. 2002. Multilateral bargaining: conditional and unconditional offers. *Economic Theory* 20, 401–12.

In, Y. and Serrano, R. 2003. Agenda restrictions in multi-issue bargaining II: unrestricted agendas. *Economics Letters* 79, 325–31.

In, Y. and Serrano, R. 2004. Agenda restrictions in multi-issue bargaining. *Journal of Economic Behavior and Organization* 53, 385–99.

Inderst, R. 2000. Multi-issue bargaining with endogenous agenda. *Games and Economic Behavior* 30, 64–82.

Jehiel, P. and Moldovanu, B. 1995. Cyclical delay in bargaining with externalities. *Review of Economic Studies* 62, 619–37.

Jun, B. 1987. A structural consideration on 3-person bargaining. In Essays on topics in economic theory. Ph. D. thesis. Philadelphia: Department of Economics, University of Pennsylvania.

Kalai, E. and Smorodinsky, M. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43, 513–18.

Krishna, V. and Serrano, R. 1996. Multilateral bargaining. *Review of Economic Studies* 63, 61–80.

Kunimoto, T. and Serrano, R. 2004. Bargaining and competition revisited. *Journal of Economic Theory* 115, 78–88.

Lang, K. and Rosenthal, R. 2001. Bargaining piecemeal or all at once? *Economic Journal* 111, 526–40.

Lensberg, T. 1988. Stability and the Nash solution. *Journal of Economic Theory* 45, 330–41.

Merlo, A. and Wilson, C. 1995. A stochastic model of sequential bargaining with complete information. *Econometrica* 63, 371–99.

Mortensen, D. 1982. Property rights and efficiency in mating, racing, and related games. *American Economic Review* 72, 968–79.

Muthoo, A. 1991. A note on bargaining over a finite set of feasible agreements. *Economic Theory* 1, 290–92.

Muthoo, A. 1993. Sequential bargaining and competition. *Economic Theory* 3, 353–63.

Myerson, R. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29, 265–81.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Nash, J. 1953. Two person cooperative games. *Econometrica* 21, 128–40.

Osborne, M. and Rubinstein, A. 1990. *Bargaining and Markets*. San Diego, CA: Academic Press. Online. Available at http://www.economics.utoronto.ca/osborne/bm, accessed 4 August 2005.

Raiffa, H. 1982. *The Art and Science of Negotiation*. Cambridge, MA: Harvard University Press.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50, 97–109.

Rubinstein, A. 1985. A bargaining model with incomplete information on time preferences. *Econometrica* 53, 1151–72.

Rubinstein, A., Safra, Z. and Thomson, W. 1992. On the interpretation of the Nash bargaining solution and its extension to non-expected utility preferences. *Econometrica* 60, 1171–86.

Rubinstein, A. and Wolinsky, A. 1985. Equilibrium in a market with sequential bargaining. *Econometrica* 53, 1133–50.

Rubinstein, A. and Wolinsky, A. 1990. Decentralized trading, strategic behavior and the Walrasian outcome. *Review of Economic Studies* 57, 63–78.

Rustichini, A., Satterthwaite, M. and Williams, S. 1994. Convergence to efficiency in a simple market with incomplete information. *Econometrica* 62, 1041–63.

Satterthwaite, M. and Williams, S. 1989. The rate of convergence to efficiency in the buyer's bid double auction as the market becomes large. *Review of Economic Studies* 56, 477–98.

Serrano, R. 2002. Decentralized information and the Walrasian outcome: a pairwise meetings market with private values. *Journal of Mathematical Economics* 38, 65–89.

Serrano, R. and Yosha, O. 1993. Information revelation in a market with pairwise meetings: the one-

sided information case. *Economic Theory* 3, 481–99.

Shaked, A. and Sutton, J. 1984. Involuntary unemployment as a perfect equilibrium in a bargaining model. *Econometrica* 52, 1351–64.

Sobel, J. and Takahashi, I. 1983. A multi-stage model of bargaining. *Review of Economic Studies* 50, 411–26.

Stahl, I. 1972. *Bargaining Theory*. Stockholm: Stockholm School of Economics.

Thomson, W. 1994. Cooperative models of bargaining. In *Handbook of Game Theory with Economic Applications*, vol. 2, ed. R. Aumann and S. Hart. New York: Elsevier.

van Damme, E., Selten, R. and Winter, E. 1990. Alternating bid bargaining with a smallest money unit. *Games and Economic Behavior* 2, 188–201.

Vannetelbosch, V. 1999. Rationalizability and equilibrium in n-person sequential bargaining. *Economic Theory* 14, 353–71.

Vincent, D. 1989. Bargaining with common values. *Journal of Economic Theory* 48, 47–62.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Watson, J. 1998. Alternating-offer bargaining with two-sided incomplete information. *Review of Economic Studies* 65, 573–94.

Weinberger, C. 2000. Selective acceptance and inefficiency in a two-issue complete information bargaining game. *Games and Economic Behavior* 31, 262–93.

Wolinsky, A. 1990. Information revelation in a market with pairwise meetings. *Econometrica* 58, 1–23.

Zeuthen, F. 1930. *Problems of Monopoly and Economic Warfare*. London: Kegan Paul.

How to cite this article

Serrano, Roberto. "bargaining." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000073> doi:10.1057/9780230226203.0098

The New Palgrave Dictionary of Economics Online

Barone, Enrico (1859 – 1924)

F. Caffè

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Barone, E.; Edgeworth, F. Y.; Einaudi, L.; general equilibrium; indirect taxation; mathematics and economics; Pantaleoni, M.; Paretian income curve; Pareto, V.; public finance; socialism; tâtonnement; Walras, L.; Wicksell, J. G. K.; Wieser, F. F. von

Article

Barone was born in Naples on 22 December 1859 and died in Rome on 14 May 1924. His education provided him with a solid grounding in the classics and in mathematics, with a view to embarking on a military career. He was appointed in 1894 to the Officers' Training School, where he was 'teacher in charge of military history'. He remained in this position until 1902, when he became the head of the historical office of the General Staff, and was given the rank of colonel.

He resigned in 1906, having already published an excellent series of biographical and historical military studies which altered the traditional concept of historical study in that field, by applying to it a method of successive approximation to which his growing interest in economics had introduced him.

His acquaintance with Maffeo Pantaleoni and Vilfredo Pareto provided him with the opportunity of collaborating with the *Giornale degli Economisti*. This association proved to be extremely valuable and productive and was to last from 1894 right up to the year of his death. It was in this periodical that in September/October 1908 he published the article 'II Ministro della Produzione nello Stato Collettivista'. This article was for a long time considered to be a mere 'curiosum'. However, after its publication in English in a volume edited by Hayek in 1935, it was destined to place its author, together with von Wieser and Pareto, alongside the founders of the pure theory of a socialist economy.

The whole discussion on collective economic planning, as it had developed since the 1920s, had ideological motivations and implications. These were totally excluded from Barone's article. The paper was, above all, a very ingenious illustration of one of Barone's deep beliefs: the usefulness of mathematical tools in clarifying questions which otherwise remain intricate and obscure. In fact it was Barone's use of equations which established the formal equivalence of the basic economic categories between a society based on private ownership in perfectly competitive conditions and a socialist society,

in which the distinct need to establish the relative distribution of income was recognized. As Samuelson writes, the innovative meaning of Barone's contribution was that ‘by avoiding all mention of utility and indeed without introducing even the notion of indifference curves, Barone was able to break new ground along lines which have in recent years become associated with the economic theory of index numbers’. The importance of Barone's arguments in the 1930s debate on the economics of socialism in which he used the idea of a Pareto optimum and improved its application, was also not fully appreciated. It remained for Samuelson's *Foundations of Economic Analysis* (1948) to give a complete acknowledgement of Barone's development (adding different products after they have been weighted by their respective prices through a process of tâtonnement) of the Paretian optimum conditions as they relate to the planning of production under collectivism.

In addition to his connections with the economists already mentioned, Barone was acquainted with the famous academics of the time, both Italian and foreign (in particular, Walras) and they all in various ways underlined the enormous potential of Barone's intellect, his clever use of analytical tools, and the extreme clarity of his graphics. Walras, for example, wrote to him saying that

> Providence has singled you out to write the historical review of the various attempts made at mathematical economics over the last centuries, which promise to offer a doctrine which will become generally accepted in the next century. I strongly urge you to recognize this as your vocation and I hope that circumstances will allow you to undertake the task.

Alongside this appreciation, however, is the impression that Barone was overstretching his interests, a feeling which was stated in no uncertain terms by Luigi Einaudi: ‘Because of the various vicissitudes of a life torn between activity, journalism, learning and the cinema … Barone, who was not inclined to laborious and painstaking research, produced far fewer fruits than his supporters had anticipated.’ The comment on the cinema refers to the fact that Barone, pressed by financial necessity and using his historical and military background, prepared treatments for the booming early Italian film industry. This division of interests delayed until 1910 Barone's appointment to a chair in political economy at the Advanced Institute of Economics and Commerce in Rome, which later became the Faculty of Economics and Commerce. But with hindsight it cannot be said that Barone's admirers were justified in ‘asking for more’. It is nearer the truth to say that he had not taken the trouble to put together his often very original and therefore extremely important papers on various subjects. As often happens, however, the very fact that his work on the pure theory of socialism received so much international acclaim was the cause for inadequate recognition of his other notable contributions. Of these, the much revised *Principi di Economia Politica* (1908) was an excellent textbook, which, together with the booklet *Moneta e Risparmio* (1920), indicated that dynamic market forces constituted the main area of his intellectual interest. See also his works entitled *Economia coloniale* (1912), ‘I Costi Connessi e l’Economia dei Trasporti’ (1921), and ‘Sindacati (Cartelli e Trust)’ (1921). Of comparable importance are Barone's investigations in the field of financial studies, demonstrating an approach different from that of De Viti de Marco and Einaudi. Barone assumed an autonomous position in as much as he availed himself of Pareto's contributions on the stability of the distribution of incomes, using it as the basis of the distribution of taxes amongst the members of the community. There have been numerous criticisms of the statistical foundation of the Paretian income curve, and even Barone admitted that its shape could

undergo change according to variations in social composition. Nevertheless, using its formulation, he provided an inductive basis for the study of a central issue in public finance. Barone's other research of recognized theoretical relevance was on the adverse welfare effects of indirect taxes on taxpayers as compared with direct taxes, for the same given tax returns. Barone was also a severe critic of the alternative versions of the financial theories of savings, in particular that of Edgeworth on minimum saving.

Although Barone was at the centre of the major theoretical debates of his time, he suffered from a conflicting loyalty to the two main formulators of general equilibrium theory, Walras and Pareto. Having been one of the first to grasp the logical aspects of general equilibrium theory, Barone was able to suggest ideas which Walras used to improve his formulation of the production function and the theory of distribution. When Pareto criticized the Walrasian formulation, Barone refrained from taking sides between the two exponents of general equilibrium theory, and as a result Walras refused to recognize the suggestions Barone had given him. Barone himself confided to Wicksell that much of his work had aimed at ‘bringing peace’ between the two great antagonists. He considered their ‘heated disputes’ to be ‘utterly and completely’ deplorable. In spite of this show of fidelity, Barone should not be thought of merely as a follower of Walras and Pareto. As Gustavo del Vecchio, an excellent judge of both Italian and international economic thought, observed,

> Barone understood the deep systematic and critical significance of general equilibrium theory, but because he had been brought up on philosophy and history, he was able to fully appreciate how great were the writers who followed the partial approach, of whom Marshall was pre-eminent. For them, economic science existed only where it could be related to concrete and immediate reality by means of our instruments of observation.

See Also

- economic calculation in socialist countries
- Pareto, Vilfredo
- social welfare function

Selected works

A complete bibliography of Barone's military studies is provided by a symposium on the 50th anniversary of his death, published in the periodical *L’Amministrazione della Difesa*, July/October 1974. The economic works of Enrico Barone have been reprinted in three volumes by Zanichelli (Bologna), 1937. The latest partial reprinting was carried out by Cedam (Padua), 1970. Of his works on economics, see:

1894a. Di alcuni problemi fondamentali per la teoria matematica dell’imposta. *Giornale degli Economisti*, March.

1894b. Sul trattamento di questioni dinamiche. *Giornale degli Economisti*, November.

1895. Studi sull distribuzione. *Giornale degli Economisti*, February/March.

1908a. *Principi di economia politica*. 7th edn, Rome: Sampaolesi, 1929.

1908b. II Ministro della produzione nello stato collectivista. *Giornale degli Economisti*, September/October. Reprinted as 'The Ministry of Production in the Collectivist State' in *Collectivist Economic Planning*, ed. F.A. Jaffé. London: Routledge, 1935; translated into many other languages.

1912a. *Economia coloniale*. Rome: Sampaolese.

1912b. Studi de economia e finanza. *Giornale degli Economisti*, April/July.

1920. *Moneta e Risparmio*. Rome: Armani.

1921a. I costi connessi e l'economia dei trasporti. *Giornale degli Economisti*, February.

1921b. Sindacati (cartelli e trust). In *Nuova Collana di Economisti*, vol. 7. Turin: Utet, 1956.

Bibliography

Del Vecchio, G. 1925. L'opera scientifica di Enrico Barone. *Giornale degli Economisti*, November.

Einaudi, L. 1939. *Prefazione di Principi di Economia finanziaria di A. De Viti de Marco*. Turin: Einaudi.

Jaffé, W., ed. 1965. *Correspondence of Léon Walras and related papers*. 3 vols. Amsterdam: North-Holland.

Samuelson, P.A. 1948. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Spinedi, F. 1924. Di un metodo nello studio della scienza sociale. *Rivista di Scienze Sociali e Discipline Ausiliarie*.

How to cite this article

Caffè, F. "Barone, Enrico (1859–1924)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000076> doi:10.1057/9780230226203.0099

The New Palgrave Dictionary of Economics Online

barriers to entry

Luís M.B. Cabral
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The precise definition of barriers to entry is controversial; different versions have been proposed over the years. The issue is not one of pure semantics, since evidence of barriers to entry plays an important role in merger review and other areas of antitrust policy. One definition that seems to reflect current thought and practice is as follows: barriers to entry are structural, institutional and behavioural conditions that allow established firms to earn economic profits for a significant length of time.

Keywords

absolute cost advantage; antitrust; asymmetric information; barriers to entry; collusion; contracts as barriers to entry; excess capacity; game theory; limit pricing; merger analysis; oligopolistic competition; patents; predatory pricing; product differentiation; scale economies; sunk costs

Article

Scholars usually debate theories, proofs, frameworks and the like. Rarely does controversy arise over a definition, as it does in the case of ‘barriers entry’.
Economists tend to agree on the relevant issues, for example, what the market outcome is given a set of assumptions regarding costs, demand, and the nature of competition. So why so much argument over a definition? One answer is that words and definitions play an important role in antitrust analysis. For example, the Federal Trade Commission and U.S. Department of Justice's *Antitrust Guidelines for Collaborations Among Competitors* (2000) suggests that evidence of substantial barriers to entry leads to closer scrutiny of the practice being challenged. Entry conditions play a similar role in other areas of antitrust policy (for example, merger analysis) in the United States, the European Union and other parts of the world. So, like it or not, we must address the issue of what barriers to entry are.
Bain (1956) defined an entry barrier as the set of technology or product conditions that allow incumbent firms to earn economic profits in the long run. Bain identified three sets of conditions: economies of

scale, product differentiation, and absolute cost advantages of established firms. Stigler (1968) criticized this approach, especially the idea of scale economies as a barrier to entry. He offered an alternative definition: a production cost that must be borne by an entrant but not by an incumbent.

Both of these approaches are incomplete, as a simple example will show. I will consider a series of different markets with the same structural conditions: a demand $D(p)$ and a technology that consists of a fixed cost F and zero variable costs. In market A, potential entrants sequentially decide whether to pay F, which is sunk; and then active firms compete à la Bertrand. Market B is like market A, but entrants collude at the monopoly price. Market C differs from market A in that potential entrants simultaneously decide whether to pay the fixed cost F, and moreover F is committed only for a short period of time. Finally, in market D potential entrants first simultaneously commit to their price level for a given short period, and then decide whether to pay the fixed cost F, to which they are committed during the same period as they are committed to price.

All of these scenarios feature the same structural conditions, and so the Bain and Stigler tests would yield the same answer. Under the Bain approach, there would be barriers to entry, namely, the scale economies implied by the fixed-cost technology. Under the Stigler definition, there would be no barriers to entry, since all firms face the same cost conditions. But both approaches would miss the substantial differences between the various markets. In market A, the equilibrium is for the first potential entrant to become a monopolist. In market B, firms will enter to the point where each firm makes zero profits (I am ignoring here the integer constraint). In market C there are multiple Nash equilibria. A reasonable equilibrium is for firms to enter with a probability such that their expected profit is zero. However, with positive probability the outcome of this equilibrium is for one firm to be a monopolist, just as in market A. Finally, in market D the equilibrium is for one firm to enter with a price equal to average cost.

The above example, while simplistic, shows the importance of looking beyond costs and demand to include behavioural conditions. What is the timing of moves – that is, what are firms committed to and for how long? The toughness of oligopolistic competition, one of the key differences across the cases in the above example, is largely the result of the assumed timing of moves. The length of time over which costs are committed (how sunk costs are) is also a crucial factor. In fact, the issue of time reveals an additional limitation of the Bain approach, with its emphasis on the long-run equilibrium. What use is it to know that the long-run equilibrium is a symmetric duopoly if it takes years for an entrant to catch up with an established firm?

If we take these considerations into account, and bear in mind the practical antitrust use of the concept of barriers to entry, a reasonable definition seems to be: *the set of structural, institutional and behavioural conditions that allow incumbent firms to earn economic profits for a significant length of time*. Admittedly, this is a fairly general definition, but necessarily so: the problem with other definitions is that, in attempting to be more specific, they become incomplete and potentially misleading.

Strategic entry deterrence

In the analysis of entry conditions and barriers to entry, a greater emphasis was initially placed on structural (or exogenous) entry conditions, such as economies of scale or incumbent cost advantages. The game theory ‘revolution’ of the 1970s and 1980s, however, shifted the focus to firm behaviour. This led to a coherent story of why structural conditions may turn into barriers to entry. Consider, for example, market A in the above example. If two firms imply zero prices, as the Bertrand assumption and

zero variable costs imply, then the equilibrium outcome is for one firm to enter and set a monopoly price, no matter how low *F* is. However, if price competition is not vigorous (market B), then no matter how high *F* is incumbent firms never earn economic profits. More generally, it's the combination of entry cost levels, the irreversibility assumption and the oligopolistic competition assumption that, together, lead to a barrier to entry.

Once the game theory apparatus was developed, the number of applications blossomed, frequently with particular models formalizing particular instances of entry barriers endogenously created by incumbents. So in the 1970s DuPont increased its capacity in the titanium dioxide industry as a way to preempt entry or expansion by rival firms. From the 1950s to the 1970s, established firms in the ready-to-eat breakfast cereal industry rapidly increased the number of brands they offered, possibly as an entry pre-emption strategy. In the late 1960s and early 1970s, Xerox developed hundreds of patents that it never used (‘sleeping patents’), their purpose being allegedly to make it more difficult for an entrant to challenge its plain-paper photocopy monopoly. Before the expiry of its patent on aspartame, Monsanto signed exclusive contracts with its major customers of Nutrasweet (Coke and Pepsi), effectively reducing the residual demand to a potential entrant. And so on.

Gilbert (1989) provides an excellent, if slightly dated, survey of the game-theoretic work in this area.

What is common to all of these examples of strategic entry deterrence is a prior action by incumbents that decreases the probability of subsequent entry. This may result from an increase in entry costs (Xerox's sleeping patents, Nutrasweet's contracts) or a decrease in the entrant's post-entry profits (Dupont's excess capacity, excess number of cereal brands). In fact, it suffices that the entrant's *beliefs* regarding costs and profits shift in the appropriate direction, even if there is no direct effect. In a world of asymmetric information, a low price by the incumbent may be interpreted as an absolute cost advantage and thus discourage entry; and repeated aggressive reaction to past entry episodes may increase the expectation of aggressive reaction to future entry. So the strategies of limit pricing or predatory pricing may also create barriers to entry.

Conclusion

The game theory revolution had the benefit of revealing the rich interaction between structural conditions and behavioural conditions. But it also complicated the task of deriving a simple, general definition of barriers to entry. In other parts of the field of industrial organization, the reaction to the ‘embarrassment of riches’ created by game theory has been to focus on particular industries. I believe a similar approach must be taken with respect to the concept of barriers to entry and its application.

See Also

- antitrust enforcement
- contestable markets
- market structure

Bibliography

Bain, J. 1956. *Barriers to New Competition*. Cambridge, MA: Harvard University Press.

Baumol, W., Panzar, J. and Willig, R. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovitch.

Carlton, D. 2004. Why barriers to entry are barriers to understanding. *American Economic Review* 94, 466–70.

Demsetz, H. 1982. Barriers to entry. *American Economic Review* 72, 47–57.

Federal Trade Commission and U.S. Department of Justice. 2000. *Antitrust Guidelines for Collaborations Among Competitors*. Online. Available at http://www.ftc.gov/os/2000/04/ftcdojguidelines.pdf, accessed 18 January 2007.

Gilbert, R. 1989. Mobility barriers and the value of incumbency. In *Handbook of Industrial Organization*, ed. R. Schmalensee and R. Willig. Amsterdam: North-Holland.

McAfee, R.P., Mialon, H. and Williams, W. 2004. What is a barrier to entry? *American Economic Review* 94, 461–5.

Schmalensee, R. 2004. Sunk costs and antitrust barriers to entry. *American Economic Review* 94, 471–5.

Stigler, G.J. 1968. *The Organization of Industry*. Homewood, IL: Richard D. Irwin.

Sutton, J. 1991. *Sunk Costs and Market Structure*. Cambridge, MA: MIT Press.

Weizsäcker, C.C.von. 1980. *Barriers to Entry*. New York: Springer-Verlag.

How to cite this article

Cabral, Luís M. B. "barriers to entry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000077> doi:10.1057/9780230226203.0100

The New Palgrave Dictionary of Economics Online

barter in transition

Barry W. Ickes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

One of the striking features of the transition in Russia was the enormous growth in the use of barter and other non-monetary means of payment. The transition from command initially led to a monetization of the economy, but a subsequent re-demonetization was a surprise. Barter was a passing phase in most transition economies but became endemic in Russia. Barter proliferated as inflation was tamed and reached its zenith prior to the August 1998 financial crisis. Various theories of why barter exploded in Russia are discussed and empirical findings are assessed.

Keywords

arrears; barter; barter in transition; contract enforcement; credit rationing; double coincidence of wants; hold-up problem; inflation; institutional trap; liquidity constraints; market power; multilateral barter; natural monopoly; price discrimination; relational capital; tax evasion; tax offsets; trust; virtual economy.

Article

One of the striking features of the transition in Russia was the enormous growth in the use of barter and other non-monetary means of payment. In addition to conventional barter – goods exchanged for goods – non-monetary transactions were also prevalent in this period. These involved non-monetary IOUs, *veksels*, which were claims on goods from other enterprises or offsets on future taxes. (The literature often treats these as equivalent, and indeed, they do arise from similar causes, but the nature of the transactions is clearly distinct.) What was a passing phase of transition in Central Europe became, by 1997, an endemic feature of the Russian situation. The explosion in barter culminated in the August 1998 Russia crisis, and since then the importance of barter has declined.

The growth in the use of barter has been characterized as ‘re-demonetization’ (Ickes, Murrell and Ryterman, 1997). The Soviet economy (with the partial exception of the household sector) was essentially a non-monetary economy. Central planners’ decisions, not purchasing power, determined the

production and allocation of goods and services. Money was mainly a record-keeping instrument. A main objective of economic reform was to transform the economy from a partially demonetized planned economy to a monetized market economy. Hence the growth in barter represented a return to a non-monetary economy, or a re-demonetization. By 1997 barter accounted for nearly half of all enterprise transactions: see Aukutsionek (1998), Commander and Mummsen (2000) and Noguera and Linz (2006). Not only was barter used in payments between enterprises (estimates of the share of barter in inter-enterprise transactions ranged from 30 per cent to 80 per cent) but it was also widely used in paying taxes to local, regional, and even federal governments. Even wages were occasionally paid in kind.
The emergence of barter in Russia in the mid-1990s presents a challenge to economic theory. Textbook analysis suggests that barter is inferior to monetary exchange. Barter requires a double coincidence of wants and hence is more costly than monetary transactions. Moreover, in Russia barter exploded as inflation was declining. Hence, the growth of barter was not the result of a flight from money as its store-of-value services declined. Indeed, one indication of this is the fact that this explosion in barter was almost exclusively within the enterprise and budget sectors of the economy. Households were typically involved only to the extent they received wages in kind. This suggests that the growth of barter had something to do with what was happening to enterprises.
Explanations of the prevalence of barter in Russia and other transition economies tend to divide into two types. One group of explanations focuses on circumstances external to the firm and views barter as an involuntary decision. The other group of explanations views the use of barter as a strategic decision by the enterprise to reduce its costs or increase its profitability (survivability).

Barter as a passive response

A leading argument of the passive theory views barter as the result of a lack of liquidity. Enterprises engage in barter because they simply lack the cash to use money. This could be due to underdeveloped financial systems (Hendley, Ickes and Ryterman, 1998) or to the effects of macroeconomic tightening (Commander and Mummsen, 2000; Noguera and Linz, 2006). In either case, the premise is that barter will only be used by enterprises that cannot afford to pay with cash; that is, barter is the result of a liquidity constraint. Hence, as argued in Woodruff (1999), barter is an instrument for cutting prices to enterprises that cannot pay the nominal price for inputs using money. Barter thus allows production to continue for those enterprises that are liquidity constrained. Barter is thus an instrument used to price discriminate. Models with this feature are developed by Ericson and Ickes (2001) and Guriev and Kvasov (2004). The liquidity explanation of barter has the advantage of getting the timing right: barter began to increase as real interest rates rose in response to the switch in policy from monetization to borrowing to finance fiscal deficits. Most of the empirical support for the theory, however, comes from survey responses of directors who state that they accept barter because their customers lack liquidity. That is, the information on the buyers’ lack of liquidity stems from surveys of sellers. A problem with this evidence, however, is that, if it is advantageous to the buyer to pay with goods rather than with money, then buyers will act strategically. That is, they will pretend to be liquidity constrained when they may not be, in order to qualify for barter. What sellers observe is the financial condition that the buyers want the sellers to believe. Hence, the liquidity of an enterprise may be endogenous. If enterprises act strategically, then the seller's information may not be the most accurate indicator of the liquidity position

of the buyer.

Some empirical evidence that casts doubt on the liquidity hypothesis comes from a study by Guriev and Ickes (2000) that avoids the problem of uninformed sellers and strategic buyers. To get around the problem of strategic signalling, Guriev and Ickes matched data on the proportions of revenues in cash and non-cash form taken from a survey of directors with the *Goskomstat* database of Russian enterprises, which contains the financial accounts of all large and medium-size industrial enterprises in Russia. This allowed them to compare the share of non-cash payments with the enterprise's financial position. They could find no discernible relationship between the use of barter and the financial condition of the enterprise. The only explanatory variable they found that predicted barter was share of export sales. (This also, perhaps, explains why barter fell dramatically when the ruble depreciated and exports increased.) Most interestingly, they found that the best predictor of whether an enterprise would use barter was lagged barter. This suggests that barter was an institutional trap. Once non-cash payments became a widespread phenomenon, it became part of the strategies of all agents. As barter proliferated it became a ‘normal’ way of doing business.

Barter as a choice

The notion that barter is a choice that an enterprise makes presumes that it results in a lowering of its net costs of production or an increase in its net revenues. Employing barter clearly increases the costs of transactions, so it must have some other offsetting benefits. For example, it may afford the buyer the opportunity to pay an effectively lower price, or it may enable enterprises to avoid taxation or reduce the cost of paying taxes. This still begs the question of why the seller is willing to accept lower-priced goods. Presumably, the key reason is the ability to pass these off for payment in taxes. This begs the further question of why governments are willing to allow tax offsets. The prevalence of tax offsets, especially at the regional level, is an accepted fact. But the motivation is more complex. (See Gaddy and Ickes, 2002 for a discussion.) Barter may also be used as a means of hiding revenues and avoiding restructuring (see virtual economy).

If we suppose that the effective price of purchasing inputs is cheaper using barter, it follows that enterprises will prefer to pay with barter than with money. There must be some way for sellers to limit the use of barter. One method would be to limit barter to enterprises with which there are good relations (see virtual economy for a discussion of relational capital and its importance in the Russian economy). Indeed, as it may be more difficult to enforce contracts using barter, a high level of relational capital or trust may be needed to enable barter to occur. An alternative method is price discrimination by those with market power.

Barter and tax evasion

One reason why enterprises may prefer to use barter is that it reduces the effective burden of taxation. In Russia, the traditional banking system served as a key part of the tax collection system. An enterprise in tax arrears would have its bank account blocked, and all receipts would go directly to the tax service. Such an enterprise thus faced 100 per cent marginal tax rates on revenues paid with money. Monetary transactions between enterprises in Russia were required by law to operate through the banking system.

Cash withdrawals could only be made for payment of wages and other incidental uses. Using barter allowed a seller in tax arrears to receive payment and circumvent the tax authorities. Hence, for such enterprises sufficient surplus would be generated by barter to offset the costs.

Evidence on the role of tax evasion as a motivation for barter is mixed. Some studies (for example, Hendley, Ickes and Ryterman, 1998) find survey evidence in favour of the tax-evasion hypothesis, while others do not (for example, Commander and Mummsen, 2000). But in most cases these studies focus the question too narrowly. They typically ask whether enterprises use barter to evade taxes. A more appropriate question would ask whether enterprises use barter to reduce the effective tax burden. Enterprises often use barter not to evade taxes but in order to pay taxes, only in a way advantageous to the enterprise. This is the practice of tax offsets.

The practice of using tax offsets as a means of reducing tax incidence became widespread prior to the 1998 crisis and was a key feature of the virtual economy (see virtual economy). Consider, for example, an enterprise that is able to supply the local government with services in lieu of taxes. The enterprise could pay its tax liability in money, but this would require selling its output for cash. Alternatively, the enterprise can negotiate with the government to supply some service as an offset for taxes. If the enterprise has resources that are not fully utilized, the latter alternative is likely to reduce the effective tax burden on the enterprise. Gaddy and Ickes (2002) provide an abundance of examples of the use of tax offsets.

Any comprehensive theory of barter in Russia in the 1990s must also explain one particularly vexing question: why governments are willing to accept tax payments in kind. It is easy to understand why enterprises would want to pay taxes in kind: this lowers the burden of their payments. It is harder to understand why governments would be willing to accept in-kind payments of taxes.

One explanation for the government's willingness to participate in barter is the virtual economy thesis. The proliferation of tax offsets is a mechanism for the distribution of subsidies in a non-transparent manner. Although more costly than a cash distribution of subsidies, non-transparency provides a more durable means of providing subsidies. They are less likely to be attacked as wasteful. This is especially true when subsidies are distributed through production, by keeping open enterprises that ought to be shut down. Thus it may be in the interest of government officials to keep subsidies non-transparent (see virtual economy).

Multilateral barter

A key problem with barter is the difficulty in finding a double coincidence of wants. It is thus interesting that in Russia multilateral barter chains appeared. Barter was often not bilateral, but part of a chain (see Humphrey, 2000). As one report described it:

> The barter chain itself turned out to be a special kind of consumer of the output. But its needs differed from the needs of liquid demand. The barter chains frequently reminded one of the ‘production for production's sake’ of the [Soviet] planned economy, when a quasi-cooperation gave rise to closed autonomous systems that served only themselves. In a number of enterprises which we surveyed, the share of output necessary simply to support the viability of the chain itself was as high as 30 per cent. (Institute of the

Economy in Transition, cited in Gaddy and Ickes, 2002)

Thus enterprises engaged in production of goods that were useful for maintaining the barter chain. The network character of barter also means that a web of relationships is crucial to maintaining it. This implies that barter was a conservative force, preserving relationships among enterprises.

Barter and market power

A robust finding among students of barter in Russia is that the large natural monopolies (Gazprom, UES, and the State Railways system, *tri tolstayaka*, 'The Three Fat Boys') were heavily involved with barter. This suggests that price discrimination may be a motive for barter. Guriev and Kvasov (2004) develop a model where firms can choose to pay in cash or in barter, and natural monopolies use barter to engage in price discrimination across customers. Unlike the model of Ericson and Ickes (2001) the Guriev–Kvasov model does not require the natural monopolies to receive any benefit from the government in exchange for the lower prices it charges to low-profitability purchasers. Rather, barter simply facilitates price discrimination and is thus profitable for monopolists. Barter allows enterprises with market power to extract higher prices from those that can afford to pay more. Of course, such discrimination can only occur if markets are not competitive.

Guriev and Ickes (2000) tested the predictions of this model and found that the use of barter increases with concentration. Industries where market concentration is very low display lower prevalence of barter than in other industries. Similarly, larger enterprises that operate in concentrated industries (and do not sell to foreign markets) are much more likely to engage in barter. Similar findings with respect to Russia (but not to Central Europe) were found in an EBRD study (Carlin et al., pp. 247–8).

Barter and efficiency

As barter is costly it is often assumed that the welfare effects of widespread barter are negative. Barter is typically viewed as a means of avoiding restructuring. An enterprise that successfully restructures may be unable to credibly signal that it is in distress, and thus it may be forced to use cash instead of barter. Ericson and Ickes (2001) developed a general equilibrium model where a restructuring trap exists: enterprises refuse to restructure because they are afraid of losing the benefits of cheap energy supplied via barter. Indeed, a form of this mechanism is at work in most price discrimination models of barter (for example, Guriev and Kvasov). Guriev and Ickes (2000) found empirical support for this hypothesis: in their sample an increase in the share of barter resulted in a decrease in labour productivity.

If barter is the result of liquidity problems external to the enterprise then access to this technology can be welfare enhancing (Noguera and Linz, 2006). The basic idea is that in a credit-rationing equilibrium higher interest rates do not provide access to capital; so cash-poor firms that have no access to barter may have to reduce production when real interest rates rise due to crowding out. With access to barter, however, they can maintain production. Of course, to evaluate the welfare consequences one must examine why the enterprises are cash poor in the first place. If this is purely external to the firm then higher production is welfare improving. If the reason they are cash poor is that they produce goods that

destroy value then barter actually is welfare decreasing (see virtual economy).

It has also been argued that barter enhances efficiency in an environment of weak contract enforcement. Marin, Kaufmann and Gorochowskiy (2000) argue that barter creates ‘deal-specific collateral’. They argue that this alleviates the hold-up problem that appears when credit enforcement is prohibitively costly. In such environments transactions that are mutually beneficial take place via barter but would not take place if cash were required. They argue that barter ‘is a *self-enforcing arrangement* which makes intermediate producers along the chain of production lose from reneging on the contract’ (2000, p. 222). The main difficulty with this theory, however, is to understand how barter creates deal-specific collateral. Presumably, an enterprise can always pledge collateral, and a promise to trade the good to a supplier to is no more credible than a promise to deliver the good if a loan cannot be repaid. The key point is that relational capital among enterprises supports barter, but barter itself does not create relational capital (see Gaddy and Ickes, 2002). The agreement between a buyer and a seller to engage in barter does not preclude the buyer from defecting anymore than a pledge of collateral to a supplier would. Thus, it is not easy to see how barter enhances transactions possibilities (though one can see how this might work with *veksels*: see below).

Veksels

As barter is costly, Russian enterprises developed an alternative institution, the use of non-monetary IOUs, or *veksels*. These were claims on output or offsets of future taxes, and their use proliferated prior to the August 1998 crisis. These promissory notes, issued by commercial banks, governments and enterprises, serve as an alternative medium of exchange. The use of *veksels* has become widespread: by one estimate the outstanding stock of these instruments had grown by the spring of 1997 to be roughly two-thirds of the value of all rubles in circulation (ruble M2) (OECD, 1997, p. 178). Enterprise *veksels* are issued by large established firms (for example, Gazprom, UES). These notes circulate among chains of enterprises that owe goods to the issuer. Eventually the note is redeemed by some customer of the issuer.

Veksels had two important characteristics that were similar to conventional barter. First, by operating out of the normal channels of the banking system they enabled enterprises to avoid taxation. Second, the use of *veksels* had the effect of keeping enterprises as part of a chain of production. The value of a *veksel* would be much lower outside the chain; hence, they had the effect of keeping enterprises from defecting. A *veksel*, for example, would be issued by a bank to support transactions among suppliers in a chain of production. If one of the suppliers chose not to produce the inputs but defect with the credit the discount on the paper may be quite large. If the credit had been issued in cash, on the other hand, it would be much easier to defect from the production chain. Hence, *veksels* may have served as a means of preserving production relations and extending credit with weak contract enforcement possibilities (Hendley, Ickes and Ryterman, 1998).

Consequences of barter

Barter raises the private costs of transactions for those engaged in it. Barter becomes prevalent when the institutional and macroeconomic environment is such that it is profitable for enterprises to bear these

costs. Hence, it is not barter per se, but the institutional and environmental constraints that generate it that are the problem. The fact that barter locks enterprises into a chain of production and inhibits restructuring is costly to the economy. But it is not the barter that is the cause of the problem, but rather a result of the peculiar economic conditions that make such an equilibrium sustainable.

After the Russian crisis, as the ruble depreciated in real terms and oil prices recovered, the barter equilibrium seems to have broken down. Cash transactions became less costly than they were prior to the crisis. Enhanced government revenues, due to tax reforms and export revenues, led to a decline in tax offsets. Hence, the relative cost of barter increased. The economy re-monetized. Whether barter will return if economic conditions return to their mid-1990s setting is an open question.

See Also

- arrears
- institutional trap
- soft budget constraint
- virtual economy

Bibliography

Aukutsionek, S. 1998. Barter in Russian industry. *Voprosy Economiki* 70, 51–60.

Carlin, W., Fries, S., Schaffer, M. and Seabright, P. 2000. Barter and non-monetary transactions in transition economies: evidence from a cross-country survey. In Seabright (2000).

Commander, S. and Mummsen, C. 2000. The growth of non-monetary transactions in Russia: causes and effects. In Seabright (2000).

Ericson, R.E. and Ickes, B.W. 2001. A model of Russia's virtual economy. *Review of Economic Design* 6, 185–214.

Gaddy, C. and Ickes, B.W. 2002. *Russia's Virtual Economy*. Washington, DC: Brookings Institution Press.

Guriev, S. and Ickes, B.W. 2000. Barter in Russian firms. In Seabright (2000).

Guriev, S. and Kvasov, D. 2004. Barter for price discrimination. *International Journal of Industrial Organization* 22, 329–50.

Hendley, K., Ickes, B.W. and Ryterman, R. 1998. Remonetizing the Russian economy. In *Russian Enterprise Reform: Policies to Further the Transition*, ed. H.G. Broadman. Washington, DC: World Bank.

Humphrey, C. 2000. An anthropological view of Barter in Russia. In Seabright (2000).

Ickes, B.W., Murrell, P. and Ryterman, R. 1997. End of the tunnel? The effects of financial stabilization in Russia. *Post-Soviet Affairs* (formerly *Soviet Economy*) 13, 105–33.

Marin, D., Kaufmann, D. and Gorochowskiy, B. 2000. Barter in transition economies: competing explanations confront Ukrainian data. In Seabright (2000).

Noguera, J. and Linz, S.J. 2006. Barter, credit and welfare: a theoretical inquiry into the Barter phenomenon in Russia. *Economics of Transition* 14, 719–45.

OECD (Organisation for Economic Co-operation and Development) 1997. *Economic Surveys, Russian Federation*. Paris: OECD.

Seabright, P. 2000. *The Vanishing Rouble*. New York: Cambridge University Press.

Woodruff, D. 1999. *Money Unmade: Barter and the Fate of Russian Capitalism*. Ithaca, NY, and London: Cornell University Press.

How to cite this article

Ickes, Barry W. "barter in transition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000079> doi:10.1057/9780230226203.0102

The New Palgrave Dictionary of Economics Online

barter

Keith Hart

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Aristotle; bargaining; barter; division of labour; exchange; gift-exchange; money; Plato

Article

Barter is a simultaneous exchange of commodities, whether goods or labour services, with bargaining and without using money. It is thus a form of trade in which credit is absent or weak, where buyers and sellers compete and rates are not fixed, and which lacks an abstract measure of value in exchange or payment.

There is no economy known to ethnographers in which barter is the only means of exchange; but there are some in which it is dominant (for example, Humphrey, 1985); and many marginal areas where barter plays a significant role alongside varieties of primitive trade and money transactions. Moreover barter is a major component of international trade, especially between east and west; it is an indispensable business tool of many modern corporations; and, with the rise of computerized exchange in the USA, it has begun to worry the Internal Revenue Service.

None of these contemporary examples, however, captures the interest of economists in barter. For it is as a central plank in the origin myth of classical and neoclassical economics that barter owes its prominence in modern thought. Adam Smith traced the ‘wealth of nations’ to division of labour:

> This division of labour, from which so many advantages are derived, is not originally the effect of any human wisdom! It is the necessary, though very slow and gradual, consequence of a certain propensity in human nature which has in view no such extensive utility; the propensity to truck, barter and exchange one thing for another. (Smith, 1776, I. ii, p. 13)

Linking this propensity to the faculties of reason and speech, Smith draws a line between ourselves and the animals: ‘Nobody ever saw a dog make a fair and deliberate exchange of one bone for another with another dog’ (1776, I. ii, p. 13). Given such a predisposition, mankind took advantage of differences in

geography and skill to establish interdependence through primitive barter. Eventually the difficulties inherent in barter led to the emergence of certain commodities as normal means of exchange and eventually to money proper. Barter, as an expression of a natural human tendency, is thus the forerunner of modern markets based on money. It follows that these markets should be allowed to be self-regulating and spared the interventions of political agents claiming to possess superior ‘wisdom’.

The founders of marginalist economics (Menger, Jevons) likewise traced the origins of money to the inefficiency of an earlier stage of barter. Most modern writers on money follow their example. In this they all echo a tradition first established by Plato and Aristotle. The Greek philosophers, however, imagined that, for money to come to express proportionate needs in a complementary division of labour, law rather than nature was required. To sum up the standard economists’ myth, a natural propensity to exchange led human beings to establish a division of labour articulated by individualized barter in local markets; eventually long-distance trade evolved and with it more efficient markets based on money. The absence of a guiding political agency is an important feature of this story.

The most elegant refutation of such a construct is made by Polanyi in *The Great Transformation* (1944). He suggests that a more plausible historical sequence is the reverse of the above. Starting from a geographically based division of labour, highly placed political agents trade goods over long distances and routinize means of payment in a process leading to the establishment of money. Local markets are sometimes a spinoff of these channels of *grand commerce*, ‘thus eventually, but no means necessarily, offering to some individuals an occasion to indulge in their alleged propensity for bargaining and haggling’ (Polanyi, 1944, p. 58). Clearly, evolutionary parables should be treated with caution, especially when they fall under one pole or the other of an ideological struggle between liberalism and socialism. Barter is invariably found in an economic context marked by several institutions of exchange. What matters is to identify its structural features in juxtaposition with alternative mechanisms. In the following discussion the evidence for barter in primitive or backward economies will be reviewed, before turning briefly to its revival in capitalist economies. The principal conclusion is that an understanding of barter requires a synthetic approach combining politics and markets.

Grierson's classic article on the silent trade (1903) is a compilation of evidence for barter without face-to-face contact which captures the early fascination of armchair anthropology with the subject. The first modern fieldwork monograph in anthropology was also devoted to institutions of exchange. In *Argonauts of the Western Pacific* (1922), Malinowski set out to challenge what he took to be prevailing models of ‘economic man’. His focus was the *kula*, a system of gift-exchange in the islands near New Guinea, involving armshells and necklaces. Under the cover of such an exchange between local leaders, the common people bartered for goods whose uneven distribution owed much to a geographically based division of labour. In addition maritime and inland villages exchanged fish for vegetables, sometimes through a formal rationing system organized by community leaders, sometimes through individual barter. Malinowski emphasized the contrast of styles and status honour between ceremonial exchange and ordinary barter, although in the first case cited they were spatially united and in the second were institutional alternatives. The Melanesians were as anxious as the ethnographer to stress their absolute antipathy to confusion of the two extreme forms of exchange. Gift-giving was formal, characterized by generosity and delay of a return (implying credit and trust); barter was informal, characterized by conflict in bargaining and immediacy of return (implying no projection of the relationship into the future). One conferred high social standing, the other low status. In practice, ceremonial exchange is a means of establishing a fragile political order for trade through a transfer of tokens of alliance between

leaders whose communities are on a footing akin to war, whereas individual barter and the appearance of hostility intrinsic to price negotiations can only be tolerated in a situation marked by peace and stable social order. Whatever the imputed social psychology, ceremonial exchange is a direct political intervention in the market, barter a manifestation of relatively free commodity exchange. Societies lacking states and money cannot rely exclusively on one form or the other. They must combine gift-exchange and barter pragmatically in response to variable degrees of ‘peace for the trade’.

More recently, Humphrey (1985) has linked barter to economic disintegration in the periphery. Her case study of a people living near the Nepal–Tibet border accounts for the dominance of barter by the low supply of money. Being very poor, they cannot afford to keep much wealth in the form of money, preferring to satisfy demand immediately in the one-to-one transactions of barter. Under these circumstances money itself becomes an item of barter. Humphrey relates this temporary phenomenon to a collapse of the local political order which has left the population in a fragmented and individuated state. They have a high level of mutual tolerance but no hierarchy through which to organize inter-local trade as they once did. There is sometimes ‘delayed barter’ involving more valuable items and the extension of credit between trading partners. But his looks like a weak version of that more formalized trade based on trust which perhaps ought not to be confused with barter. Delay in making a return and associated relations of credit/debt are antithetical to barter; for bargaining is impossible if either party does not have the option of withdrawing from the negotiation.

Recent anthropological research has focused on the tendency of bartered goods to fall into distinct ‘spheres of exchange’. In a classic article Bohannan (1955) argues that the Tiv of Nigeria prefer to exchange goods of the same broad category and look down on transactions across the boundaries between such spheres. Subsistence items are distinguished from prestige goods like cattle, slaves, metal bars and cloth. The highest level of exchange involves marriageable women only. In the colonial period money destroyed this compartmentalization of exchange by making conversion between spheres easier. Cultural disruption was the result.

This argument confuses several levels of analysis. First, as Marshall pointed out, utilities are never wholly commensurate: subsistence, luxury and prestige goods cannot be equalized simply by sharing a monetary medium of evaluation. It does not make any sense to ask how many sacks of potatoes an Eton education is worth, even though they both have a money price. Second, there are clearly problems of conversion in barter between low-bulk, high-value items and high-bulk, low-value items, typically between long-distance trade goods and small agricultural surpluses. Livestock and poultry offer one ready means of conversion, however. Again, nobody likes to sell a hi-fi set in order to pay the groceries bill, but such conversions are known to occur. Third – and most damaging – the main force restricting exchange to separate spheres is political and ideological, not economic in the technical sense. Tiv elders control commerce with the outside world and hold their junior kinsmen on the farm through a monopoly of marriageable women. Colonialism – not money as a fetishized abstraction – undermined that control by introducing markets for the young men's goods and labour.

The absence of money does not in itself present an insurmountable obstacle to efficient exchange. Much the most important precondition for barter lies in the forms of political order (or the lack of it); and it is this which is undermined by modern markets and by the states whose power is essential to their functioning. With this in mind we should consider briefly the survival of barter in the trading institutions of the advanced economies.

Much of the trade between the West and the Communist bloc took the form of barter for the obvious

reason that the East could not accumulate hard currency reserves. The end of Communism was also associated with substantial intra-country barter; see barter in transition. Third World countries, such as some West African states, barter the products of an ecological division of labour (meat for grain) owing to a general lack of cash. Such activities are similar to the early trade between political agents emphasized by Polanyi. The multinational corporations have treasuries larger than those of many nations, yet they often choose to barter commodities they would normally be unable to sell in open markets – so many thousand gallons of paint for several months' lease of a Bahamas hotel chain.

The laissez-faire economist's myth of barter as an expression of mankind's innate propensity to exchange ought to be replaced by a more complex historical appraisal of the institution's significance. Barter is an extremely widespread phenomenon, occurring in many times and places as a partial and often temporary solution to the problem of exchange. It is not abolished by money and indeed sometimes transforms money itself into an item of barter; and, if recent trends are a reliable indicator, it may now be undergoing a revival in the West. It was always a mistake to suppose that markets expanded without definite political conditions for their maintenance. Barter too rests on variable political conditions which are as much contemporary as they are primitive.

See Also

- barter in transition
- economic anthropology
- exchange

Bibliography

Bohannan, P. 1955. Some principles of exchange and investment among the Tiv. *American Anthropologist* 57, 60–70.

Grierson, P. 1903. The silent trade. In *Research in Economic Anthropology*, vol. 3, ed. G. Dalton. Greenwich, CT: JAI Press, 1980.

Humphrey, C. 1985. Barter and economic disintegration. *Man* 20(1), 48–72.

Malinowski, B. 1922. *Argonauts of the Western Pacific*. London: Routledge.

Polanyi, K. 1944. *The Great Transformation*. Boston: Farrar.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. E. Cannan. New York: The Modern Library, 1937.

How to cite this article

Hart, Keith. "barter." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000342> doi:10.1057/9780230226203.0101

The New Palgrave Dictionary of Economics Online

Barton, John (1789 – 1852)

Maxine Berg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Barton, J.; capital accumulation; circulating capital; colonization; Corn Laws; Fixed capital; Huskisson, W.; labour supply; London Statistical Society; machinery question; Malthus, T. R.; Malthus's theory of population; McCullough, J. R.; monopoly; Old Poor Law; population; Ricardo, D.; Sismondi, J. C. L. S. de; Smith, A.; technical change

Article

Barton is remembered in the history of economic thought for an early critical discussion of the impact of machinery on employment. A Sussex landowner, he combined an interest in statistical observation with a special concern for the impact of industrial and agrarian change on the condition of the labourer. He was the author of two important books, *Observations on the Circumstances which Influence the Condition of the Labouring Classes of Society* (1817) and *An Inquiry into the Causes of the Progressive Depreciation of Agricultural Labour in Modern Times* (1820). Later, in the 1830s, he wrote several tracts on the Corn Laws and on population and colonization. He was elected a fellow of the London Statistical Society in 1847 and read a paper in 1849, ‘The Influence of the Subdivision of the Soil on the Moral and Physical Well-being of the People of England and Wales’. His early manuscript essays show a wide and careful grounding in political economy based on Hume, Smith and Ricardo. His first books were, however, written as interventions in the contemporary debates on the Poor Laws.

Barton's primary purpose in writing both the *Observations* and the *Inquiry* was to challenge Malthusian population theory, and the prevailing opinion that the cause of excess population and falling wages was the support offered by the Old Poor Law. Barton combined abstract reasoning with statistical data in a critique of Malthus and Ricardo that so impressed Schumpeter that he judged it ‘a remarkable performance … far above the rest of the literature that currently criticized the class leaders for their lack of realism, actual or supposed’.

Barton drew on population figures from the 16th to the 18th century to challenge Malthusian propositions of the dependence of population growth on levels of capital accumulation. Using data gathered from the agricultural districts, he also challenged assumptions of flexible supplies of labour in

response to wage changes. His data provided no support for those who feared that population growth would follow on high wages. Custom and employment prospects, not changing wage rates, were the most important determinant of the age of marriage. Barton dissected the gap between population and labour supply, analysing age structure, apprenticeship, skills and labour immobility. His demographic work impressed Sismondi and induced McCulloch to give up Malthusianism.

The most influential analysis of the *Observations*, however, was Barton's critique of Ricardo's and Malthus' early optimistic assumptions of the impact of capital accumulation and machinery on the working classes. Another reason why high wages could not be blamed for inducing population growth, he argued, was that capital accumulation did not necessarily entail increases in employment. Capital had to be disaggregated into fixed (technological) and circulating (wage goods) capital before its impact on the labour market could be assessed. The demand for labour was dependent on circulating, not fixed, capital. And if wage rates rose relative to commodity prices, employers would substitute machinery for labour. The process of capital accumulation could, therefore, entail the release of rather than the demand for labour, and the amount of labour employed in the construction and repair of new machinery would provide only small compensation.

Barton's *Observations* was read by political economists and policy-makers – Huskisson and Malthus noted it, Sismondi praised it and McCulloch reviewed it. It was said to have induced Ricardo to make an about-turn in the third edition of his *Principles* and so to write his controversial chapter on machinery accepting the idea that the introduction of machinery could hurt the interests of manual labour. But Ricardo did not introduce this change until the third edition in 1821, and his analysis was rather different. Accepting Barton's point that the introduction of machinery might be induced by wage increases, he added his own novel analysis of autonomous technical change. It is likely that Ricardo changed his views on machinery not because he read Barton but because of contemporary political concern over the machinery issue combined with a timely reminder of Barton's work in a recent correspondence he had with McCulloch.

Barton's later pamphlets and newspaper articles of the 1830s and 1840s extended his early analysis into a general critique of industrialism. He defended the Corn Laws, arguing that labour thrown out of agriculture could not be transferred easily to manufacturing, and that the extension of manufacturing and machinery only concentrated wealth in fewer hands. He drew attention to an Adam Smith forgotten by his contemporaries – the Smith who conducted a radical critique of the monopoly spirit of merchants and manufacturers. John Barton's critique of industrialism and the introduction of machinery was a striking example of a special early 19th-century combination of traditional landed opinion with a radical concern for the condition of labour.

Selected works

1817. *Observations on the Circumstances which Influence the Condition of the Labouring Classes of Society*. London.

1820. *An Inquiry into the Causes of the Progressive Depreciation of Agricultural Labour in Modern Times*. London.

1962. *John Barton (1789–1852), Economic Writings*, 2 vols, ed. G. Sotiroff. Regina, Sask.: Lynn Publishing Company.

How to cite this article

Berg, Maxine. "Barton, John (1789–1852)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000082> doi:10.1057/9780230226203.0103

The New Palgrave Dictionary of Economics Online

Bastiat, Claude Fr é d é ric (1801 – 1850)

R.F. Hébert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bastiat, C. F.; Cobden, R.; Dupuit, A.-J.-E. J.; free trade; Harmonism; service theory of value

Article

French economist and publicist, born at Bayonne on 30 June 1801, the son of a merchant in the Spanish trade; died in Italy, at Rome, on 24 December 1850. Orphaned at the age of nine, Bastiat nevertheless received an encyclopedic education before entering his uncle's business firm in 1818. By 1824 he was expressing dissatisfaction with his employment. Upon inheriting his grandfather's estate in 1825, he left business and became a gentleman farmer at Mugron, but showed no more aptitude for agriculture than he had for commerce. So he became a provincial scholar, establishing a discussion group in his village and reading voraciously. His later writings show familiarity with the works of French, British, American and Italian authors, among them Say, Smith, Quesnay, Turgot, Ricardo, Mill, Bentham, Senior, Franklin, H.C. Carey, Custodi, Donato and Scialoja.

Bastiat left France in 1840 to study in Spain and in Portugal, where he tried unsuccessfully to establish an insurance company. Returning to Mugron, he learned (in the course of seeking information for his study club) of Cobden's Anti-Corn Law League and became an ardent free-trader (the ‘French Cobden’). As a complete unknown in economics, he submitted a stirring article to the *Journal des économistes* in 1844, dealing with the influence of protectionism on France and England. It created an immediate sensation and raised a clamour for more from the editors. This response encouraged Bastiat's *Economic Sophisms*, which quickly sold out upon its publication in 1845, and was soon thereafter translated into English and Italian. In 1846 Bastiat moved to Paris, where he established the Association for Free Trade and quickened his literary activity, endangering his frail health in the process. A torrent of articles, pamphlets and books now flowed from his talented pen, undoubtedly made possible in such short order by the preceding 20 years of practically uninterrupted reflection. Some scholars say the frenzy produced more heat than light, yet on the whole, economics is better off for Bastiat's Herculean efforts.

Bastiat was one of several writers (Quesnay, Smith, Say and Carey were the others) who formed the doctrine of Harmonism, or the optimistic idea that class interests naturally and inevitably coincide so as

to promote economic development. The major challenge to this view came from Ricardo and Malthus, whose theories cast a sinister shadow over the prospect of economic progress. As against Ricardo's system, Bastiat erected a theory of value based on the idea of service. He distinguished between utility and service, identifying the former as insufficient, of itself, to establish value, because certain free goods (sun, air, water) have utility. Bastiat considered all commercial transactions as exchanges of service, with value measured in terms of the trouble a buyer saves by making the purchase.

J.E. Cairnes complained that this merely confounded what Ricardo had sought to delineate, namely those cases in which value is proportioned to effort and sacrifice from those in which it is not. A more fundamental criticism is that Bastiat's theory, notwithstanding denials to the contrary, is simply a labour theory in different guise. It is noteworthy, however, that Bastiat's idea bears a close resemblance to the notion of ‘public utility’ which Dupuit applied so successfully to the measure of gain from transport improvements, and in which reduction of costs effected by the improved service became the central issue. Yet any connection between the two, tenuous as it may be, must be considered to run from Dupuit to Bastiat rather than the reverse, since Dupuit published his famous article on public works and marginal utility before Bastiat abandoned his earlier polemics in favour of more ‘constructive’ attempts at theory. Bastiat's theory of rent, also clearly aimed against Ricardo, denied the notion of unearned income, again advancing the view that the value of land (always in the absence of government interference) derives entirely from the services it renders.

Generally, judgement on Bastiat has been that he made no original contributions to economic analysis. Cairnes, Sidgwick and Böhm-Bawerk discounted his pure economics completely. Marshall said that he understood economics hardly better than the socialists against whom he declaimed. And Schumpeter declared that Bastiat was not a *bad* theorist, he was simply no theorist at all.

Schumpeter also described Bastiat as ‘the most brilliant economic journalist who ever lived’, and so weighty a thinker as Edgeworth praised Bastiat's genius for popularizing, in the best sense of the term, the economic discoveries of his predecessors. Almost all commentators agree that Bastiat was unrivalled at exposing economic fallacies wherever he found them, and he found them everywhere. He was quite simply a genius of wit and satire, frequently described as a combination of Voltaire and Franklin. He had the habit of exposing even the most complex economic principles in amusing parables that both charmed and educated his readers. His writings retain their currency, even today. And as Hayek has reminded us in his introduction to Bastiat's *Selected Essays*, his central idea continues to command attention: the notion that if we judge economic policy solely by its immediate and superficial effects, we shall not only not achieve the good results intended, but certainly and progressively undermine liberty, thereby preventing more good than we can ever hope to achieve through conscious design. This principle is exceedingly difficult to elaborate in all of its profundity, but it is one which has galvanized the thought of contemporary economists, Hayek and Friedman.

Over the long haul, Bastiat's influence has waxed and waned. In his own day he received the ready support of Dunoyer, Blanqui, Chevalier and Garnier. Francis A. Walker introduced his doctrines into America at about the time of the Civil War. Pre-First World War French liberals such as Leroy-Beaulieu, Molinari and Guyot relied on his authority. Bastiat's ideas subsequently went into a long decline, only to become resurgent in the late 20th century among libertarian economists dissatisfied with Keynesian orthodoxy and Marxist alternatives. Ironically, Bastiat's originality is exhibited most in his contribution to political theory, which has drawn surprisingly little attention to this day.

Selected works

1844. De l'influence des tarifs fran ē ais et anglais sur l'avenir des deux peuples. *Journal des économistes* 9, 244–71.

1964. *Economic Sophisms*, trans. A. Goddard Princeton: D. Van Nostrand.

1964. *Selected Essays on Political Economy*, trans. S. Caine. Princeton: D. Van Nostrand.

1964. *Economic Harmonies*, trans. W.H. Boyers. Princeton: D. Van Nostrand.

Bibliography

Baudin, L. 1962. *Frédéric Bastiat*. Paris: Dalloz.

Hayek, F.A. 1964. *Introduction to F. Bastiat, Selected Essays on Political Economy*. Princeton: D. Van Nostrand.

Russell, D. 1963. *Frédéric Bastiat: Ideas and Influence*. Irvington-on-Hudson, NY: Foundation for Economic Education.

How to cite this article

Hébert, R. F. "Bastiat, Claude Frédéric (1801–1850)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000092> doi:10.1057/9780230226203.0104

The New Palgrave Dictionary of Economics Online

Baudeau, Nicolas (1730 – c1792)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Baudeau, N.; free trade; Gournay, Marquis de; hoarding; luxury; Mirabeau, V. R., Marquis de; monopoly; net product; Physiocracy; productive vs. unproductive expenditure; sumptuary laws; surplus

Article

Born at Amoise, Baudeau entered the church, becoming a canon and professor of theology at the Chancelade Abbey. He was subsequently called to Paris in the service of Archbishop de Beaumont. In 1765, Baudeau founded the periodical *Ephémérides*, becoming its first editor till late 1768 and again during its two subsequent revivals. Converted to Physiocracy by Mirabeau in 1768, he became one of its most active propagandists through the many articles, pamphlets and books he produced. He died insane in Paris *circa* 1792 (Coquelin and Guillaumin, 1854, I, p. 148). Daire (1846, pp. 652–4) provides a bibliography of the economic writings and reprints his long introduction to economic philosophy (Baudeau, 1771) and his explanations of the *Tableau économique* (Baudeau, 1767–8), which Marx (1962, p. 324) found helpful for clarifying some of its more difficult points and which remains a most useful introduction to Physiocracy and the *Tableau's* intricacies. Baudeau (1771) is noteworthy for its concise definition of monopoly as ‘everything which by force limits the numbers and competition of buyers and sellers’ (p. 327) and its direct attribution to Gournay of the phrase, *laissez les faire* (p. 323). Perhaps the most interesting of Baudeau's many writings is his systematic exposition and development of the Physiocratic theory of luxury (Baudeau, 1767), the most complete version of that doctrine and as such wrongly ignored (Dubois, 1912, pp. v–vi). Inspired by the Swedish sumptuary laws of 1767, and bearing in mind the Physiocratic division of output between necessary expenses and disposable net product, the essay clearly defines luxury as ‘that subversion of the natural and essential order of national expenditure which increases the total of unproductive expenditure to the detriment of that which is used in production and at the same time to the detriment of production itself’ (Baudeau, 1767, p. 14). In other words, disposal of the net product when in direct agricultural investment or in spending which directly or indirectly enhances the demand for agricultural produce is productive: other uses of the surplus are

wasteful, luxury spending. For example, hoarding which detracts from demand for agricultural produce, is luxury; importing commodities from abroad, if this increases overseas demand for domestic produce and thereby augments productive expenses, is not. Sumptuary laws are therefore not appropriate for curtailing luxury; free trade and a more simple pattern of consumption channelling more demand to the agricultural sector, are much more effective. In short, ostentation in consumption is to be preferred to ostentation in display and ornament, since the former creates a greater market for agricultural produce and hence for all production. As Meek (1962, p. 318) points out, this 'theory of luxury, with its distinction between productive and unproductive expenditure out of revenue, was much more useful to Smith and Ricardo than it was to the underconsumptionists', despite its emphasis on consumption spending as a factor in stimulating production.

See Also

- Ephémérides du citoyen ou chronique de l'esprit national

Selected works

1767. Principes de la science morale et politique sur le luxe et les lois somptuaires. *Ephémérides* 1(1), January, and 1(3). Reprinted, ed. A. Dubois, Paris: Marcel Rivière, 1912.

1767–8. *Explication du Tableau économique à Madame de* ***. Paris: Delalain, 1776.

1771. *Première introduction à la Philosophie économique ou analyse des états policés*. In *Physiocrates*, ed. E. Daire. Paris: Guillaumin, 1846.

Bibliography

Coquelin, Ch. and Guillaumin, M. 1854. Baudeau. In *Dictionnaire de l'économie politique*. Paris: Guillaumin.

Daire, E. 1846. Notice sur la vie et les travaux de l'Abbé Baudeau. In E. Daire, *Physiocrates*. Paris: Guillaumin.

Dubois, A., ed. 1912. *Nicolas Baudeau, Principes de la science morale et politique*. Paris: Marcel Rivière.

Marx, K.H. 1962. *Theories of Surplus Value*, Part I. Moscow: Foreign Languages Publishing House.

Meek, R.L. 1962. *The Economics of Physiocracy*. London: Allen & Unwin.

How to cite this article

Groenewegen, Peter. "Baudeau, Nicolas (1730–c1792)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000093> doi:10.1057/9780230226203.0105

The New Palgrave Dictionary of Economics Online

Bauer, Peter Thomas (1915 – 2002)

James A. Dorn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bauer, P.T.; capital accumulation; central planning; democracy; development economics; economic freedom; foreign aid; free trade; freedom of contract; gains from trade; institutional economics; limited government; poverty; poverty traps; property rights; total factor productivity

Article

Peter (Lord) Bauer, one of the pioneers of early post-Second World War development economics, stood almost alone in the 1940s and 1950s in questioning the prevailing orthodoxy.
Born in Budapest on 6 November 1915, he was the son of a bookmaker. Bauer left Hungary in 1934 to study at Cambridge University, where he earned a first-class degree in economics from Gonville and Caius College in 1937. He returned home to complete his law degree at Budapest University, and then took a job in London with the trading firm of Guthrie & Company. In 1947 he was appointed a lecturer in agricultural economics at London University. From 1948 to 1956 he was a lecturer in economics at Cambridge University, and then became Smuts Reader in Commonwealth Studies. In 1960 Bauer accepted a professorship at the London School of Economics, and took emeritus status in 1983. Prime Minister Margaret Thatcher elevated Bauer to the House of Lords, as a life peer, in 1982. Lord Bauer was a fellow of the British Academy and of Gonville and Caius College. He was the first recipient of the Milton Friedman Prize for Advancing Liberty, a $500,000 prize awarded every two years by the Cato Institute. The award cited Bauer's ‘tireless and pioneering scholarly contributions to understanding the role of property and free markets in wealth creation’. Peter Bauer died on 2 May 2002 at the age of 86.
In the early post-war era, orthodox development economists held that there was a ‘vicious circle of poverty’. They assumed that low incomes in less developed countries would prevent sufficient domestic saving and capital accumulation, which were seen as essential for growth. Moreover, poor people were assumed to be incapable of readily responding to market incentives or to have the foresight to save and invest, investment opportunities were seen as narrowly limited, and external trade was viewed as ineffective or even harmful. Poverty was therefore regarded as self-perpetuating. The only escape was to generate a ‘big push’ by comprehensive central planning and by relying on external assistance.

Bauer's first-hand observations during his extensive work in south-east Asia and in British West Africa in the 1940s and 1950s led him to question the conventional wisdom. In his classic studies of the rubber industry in Malaya (Bauer, 1948) and small traders in West Africa (Bauer, 1954), he found strong evidence that poor people can lift themselves out of poverty by hard work, entrepreneurial activities, and internal and external trade – provided they have the freedom to do so. He was fond of saying, ‘If the notion of the vicious circle of poverty were valid, mankind would still be living in the Old Stone Age’. Rather than advocate a state-led development model, which was in high fashion at the time, Bauer argued that investment planning, compulsory saving, protectionist trade policies, marketing boards, and government-to-government transfers (foreign aid) would politicize economic life, empower the ruling class, and perpetuate poverty. His views have been vindicated by the failure of comprehensive economic planning and by the ineffectiveness of official aid to spur development.

For Bauer, the essence of economic development is to increase ‘the range of effective alternatives open to people’—that is, to increase economic freedom. Until recently, this classical-liberal view was largely invisible. Bauer was among the first to downplay the importance of physical capital accumulation as a precondition for growth. His focus was on institutions and incentives, and especially on the dynamic gains from trade. Total factor productivity is a black box that must be opened to understand the underlying forces of the development process. Bauer was sceptical that those forces could be precisely modelled or that there could be a general theory of development. The process was much too complex. The primary role of government, in Bauer's view, is to protect private property rights and freedom of contract so that individuals are free to choose and to trade. Conditions will then be conducive to develop and to prosper. Limited government is more important than democracy, in this respect. Hong Kong has few natural resources but has limited government and free trade, and was able to escape the ‘poverty trap’—without comprehensive planning or foreign aid.

Bauer, like Ronald Coase, relied on direct observation, an understanding of institutions and history, and sound economic logic to overturn conventional wisdom. When nearly everyone was focusing on capital accumulation as the primary determinant of growth, Bauer (1957a, p. 119) argued, ‘It is more meaningful to say that capital is created in the process of development, rather than that development is a function of capital’.

In his final book, *From Subsistence to Exchange and Other Essays* (2000), Bauer summarized his market-liberal vision of the development process:

- ‘Economic performance depends on personal, cultural, and political factors, on people's attitudes, motivations, and social and political institutions.’
- ‘Contacts through traders and trade are prime agents in the spread of new ideas, modes of behavior, and methods of production.’
- ‘Development aid is thus clearly not necessary to rescue poor societies from a vicious circle of poverty. Indeed, it is far more likely to keep them in that state.’

Those ideas were controversial for many years, but are now more readily accepted in the field of development economics. Bauer deserves much credit for that reversal.

See Also

- growth and institutions

Selected works

1948. *The Rubber Industry*. London: Longmans, Green and Co.

1954. *West African Trade*. Cambridge: Cambridge University Press.

1957a. *Economic Analysis and Policy in Underdeveloped Countries*. Durham, NC: Duke University Press.

1957b. (With B.S. Yamey.) *The Economics of Underdeveloped Countries*. Chicago: University of Chicago Press.

1976. *Dissent on Development*, revised ed. Cambridge, MA: Harvard University Press.

1991. *The Development Frontier*. Cambridge, MA: Harvard University Press.

2000. *From Subsistence to Exchange and Other Essays*. Princeton, NJ: Princeton University Press.

Bibliography

Dorn, J.A. 2002. Economic development and freedom: the legacy of Peter Bauer. *Cato Journal* 22, 355–71.

Yamey, B.S. 2005. Peter Bauer: an unusual applied economist. *Cato Journal* 25, 449–53.

How to cite this article

Dorn, James A. "Bauer, Peter Thomas (1915–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000261> doi:10.1057/9780230226203.0106

The New Palgrave Dictionary of Economics Online

Bayes, Thomas (1702 – 1761)

D.V. Lindley

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bayes, T.; De Finetti, B.; expectations; Inference; Laplace, P. S.; Price, R.; probability; subjective probability

Article

The Rev. Thomas Bayes was the eldest son of Joshua Bayes, a minister in the nonconformist church. He was probably educated at Coward's Academy. After assisting his father as pastor in Hatton Garden, London, he became, in 1731, Presbyterian minister at Mount Sion, Tunbridge Wells where he remained until his death on 17 April 1761. His fame today rests entirely on one paper, found by his friend Richard Price amongst Bayes' effects after his death and presented to the Royal Society (Bayes, 1763; a convenient recent reference is Bayes, 1958). The paper appears to have aroused little interest at the time and a proper appreciation was left to Laplace. Even today there is much discussion over just what Bayes meant, but the fact that so much interest is taken in a paper over 200 years old testifies to the importance of the problem and the brilliance of Bayes' argument.

The problem was this (as stated at the beginning of the paper): '*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.'

Bayes' solution depended on two original ideas. The first, in the modern notation where $p(A|B)$ means the probability of A given B, says

$$p(B|A) = p(A|B)\ p(B) \mid p(A)$$

and is always known as Bayes' theorem. The second idea is more controversial and open to many interpretations. The question is what 'rule is the proper one to be used in the case of an event concerning

the probability of which we absolutely know nothing antecedently to any trials made concerning it'? To solve the problem Bayes took *A* to be the event of *r* happenings and *s* failures; *B* to be the unknown value θ of 'its happening in a single trial' so that $p(r, s|\theta) = \theta^r(1-\theta)^s$; and supposed $p(r, s) = (r+s)^{-1}$ as a solution to the second question. This is equivalent to taking $p(\theta)$ as constant.

The importance of Bayes' ideas goes beyond the initial problem. Let *A* be any *particular* event and *B* some *general* proposition. Then his theorem enables one to pass from the probability of the particular given the general, $p(A|B)$, which, as above, is often straightforward, to the difficult probability of the general given the particular, $p(B|B)$. As such it provides a solution to the central problem of induction or inference, enabling us to pass from a particular experience to a general statement. This Bayesian inference applies generally in science, economics and law. A special case with statistical problems is called Bayesian Statistics. It has been shown by Ramsey (1931), De Finetti (1974/5) and others that this is the only coherent form of inference. Despite this, eminent philosophers like Popper (1959) still misunderstand Bayes and deny probabilistic induction.

Bayes' solution to the second question has not been generally accepted and the probability to be assigned to the general proposition before the particular is observed, $p(B)$, has been the subject of much discussion. Solutions by Jeffreys (1985), and by Jaynes (1983) using entropy ideas, have all met with difficulties. The best solution currently available is to accept that *all* probabilities are subjective so that, in particular, $p(B)$ is the subject's probability for the general proposition. This view is primarily due to De Finetti. Enough data (in the form of particular events) enable subjects, despite differences in $p(B)$, to have close agreement on $p(B|A)$.

An interesting feature of Bayes' approach is that he defines probability in terms of expectation. The amount you would pay for the expectation of one unit of currency were *B* to occur is $p(B)$. Because of its confusion with utility concepts, this approach has not been much used.

It is hard to think of a single paper that contains such important, original ideas as does Bayes'. His theorem must stand with Einstein's $E = mc^2$ as one of the great, simple truths.

Bibliography

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society London* 53, 370–418.

Bayes, T. 1958. Reprint of the above with biographical note by G.A. Barnard. *Biometrika* 45, 293–315.

De Finetti, B. 1974/5. *Theory of Probability*. 2 vols .New York: Wiley.

Jaynes, E.T. 1983. *Papers on Probability, Statistics and Statistical Physics*, ed. R.D. Rosenkrantz. Dordrecht: Reidel.

Jeffreys, H. 1985. *Theory of Probability*. Oxford: Clarendon Press.

Popper, K.R. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Ramsey, F.P. 1931. *The Foundations of Mathematics and Other Logical Essays*. London: Kegan, Paul, Trench, Trubner.

How to cite this article

Lindley, D. V. "Bayes, Thomas (1702–1761)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000096> doi:10.1057/9780230226203.0107

The New Palgrave Dictionary of Economics Online

Bayesian econometrics

Dale J. Poirier
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Bayesian econometrics’ consists of the tools of Bayesian statistics applicable to economic phenomena. The Bayesian paradigm interprets ‘probability’ as a measure of ‘uncertainty’ or ‘degree of belief’ associated with the occurrence of a particular uncertain event, given the available information and any accepted assumptions. It prescribes how an individual *should* act in the face of such uncertainty in order to avoid undesirable inconsistencies. The coherence of the Bayesian approach contrasts sharply with conventional statistical methods which sometimes advocate negative estimators of positive quantities to ensure unbiasedness, and confidence intervals which may be null or consist of the whole parameter space.

Keywords

Bayes, T.; Bayes’ theorem; Bayesian econometrics; Bernoulli, J.; collinearity; de Finetti, B.; empirical Bayes analysis; exchangeability; expected subjective utility; extreme bounds analysis; frequentist statistics; Good, I.J.; hypothesis testing; interval estimation; Jeffreys’ rule; Laplace, P.S.; likelihood principle; Lindley, D.; Markov chain Monte Carlo methods; maximum likelihood; model building; objective probability; point estimation; prediction; probability; regression; representation theorem; Savage, L. J.; statistical inference; subjective probability; uncertainty

Article

‘Bayesian econometrics’ consists of the tools of Bayesian statistics applicable to economic phenomena. Bayesian statistics traces its roots back to Reverend Thomas Bayes (born circa 1702 and died in 1761) who was an ordained nonconformist minister in England. His ideas appear to have been independently developed by James Bernoulli, and later popularized independently by Pierre Laplace later in the 18th century. After more than a century of neglect, a rebirth of Bayesian statistics occurred in the 1930s at the hands of Sir Harold Jeffreys and Bruno de Finetti, and momentum built in the 1950s as a result of the efforts of I.J. Good, Dennis Lindley and Leonard J. Savage. Bayesian econometrics started in the 1960s with the work of Jacque Dreze and Arnold Zellner. With the computational revolution sparked by Markov chain Monte Carlo (MCMC) techniques in the 1980s and 1990s, many computational constraints were removed, and Bayesian analysis was flourishing in a wide variety disciplines as the new millennium began.

The Bayesian paradigm interprets ‘probability’ as a measure of ‘uncertainty’ or ‘degree of belief’ associated with the occurrence of a particular uncertain event, given the available information and any accepted assumptions. It prescribes how an individual *should* act in the face of such uncertainty in order to avoid undesirable inconsistencies.

Consider an individual asked to quote probabilities on a set of uncertain events, and required to accept any wagers about these events. According to Bruno de Finetti's *coherency principle*, such an individual should never assign probabilities so that someone else can select stakes that guarantee a sure loss (*Dutch book*) for the individual whatever the eventual outcome. This simple principle implies the usual axioms of probability except that the additivity of probability for unions of disjoint events is required to hold only for finite unions.

Expected utility maximization (or loss minimization) provides a basis for rational decision making, and Bayes’ theorem describes how beliefs evolve as data are obtained. There are numerous axiomatic formulations leading to the central unifying Bayesian prescription of maximizing expected subjective utility as the guiding principle of Bayesian statistical analysis. Bernardo and Smith (1994, ch. 2) is a valuable introduction to this vast literature. While the *descriptive* accuracy of the Bayesian approach in capturing

the actual behaviours of individuals is questioned by many opponents, Bayesians claim that the Bayesian view provides only *normative* guidelines for behaviour.

The *subjective* interpretation of probability is based on an individual's personal assessment of a situation. For evidence of the use of subjectivity by history's most illustrious scientists, see Press and Tanur (2001). Accordingly, probability is a property of an individual's perception of reality. In contrast, according to objective interpretations, probability is a property of reality itself. For subjectivists there are no ‘true unknown probabilities’ in the world to be discovered. Instead, ‘probability’ is in the eye of the beholder. In de Finetti's words, ‘Probability does not exist’.

De Finetti assigned a fundamental role in Bayesian analysis to exchangeability. A finite sequence of random quantities is *exchangeable* if the joint probability of the sequence, or any subsequence, is invariant under permutations of the subscripts. An infinite sequence is exchangeable if any finite subsequence is exchangeable. Exchangeability involves recognizing symmetry in beliefs concerning *observables*, and presumably this is something about which a researcher may have intuition. It provides an operational meaning to the weakest possible notion of a sequence of ‘similar’ random quantities. It is operational because it requires only probability assignments of observable quantities, although admittedly this becomes problematic in the case of infinite exchangeability.

The links between exchangeable beliefs over uncertain *observables* and the parameters in statistical models are provided by various generalizations of Bruno de Finetti's celebrated representation theorem for infinite sequences of exchangeable Bernoulli random variables (see Bernardo and Smith, 1994, ch. 4). These theorems provide conditions under which exchangeability, and other symmetries, give rise to an isomorphic world consisting of i.i.d. observations with a given sampling distribution, conditional on a mathematical construct (a parameter), and guarantee the existence of a prior distribution for it. De Finetti put parameters in their proper perspective: they are mathematical constructs that provide a convenient index for a family of probability distributions, and they induce conditional independence in sequences of observables.

Bayesian inference involves updating prior beliefs into posterior beliefs conditional on observed data. Appealingly, Bayesian analysis requires only a few general principles that are applied over and over again in different settings. Bayesians begin by specifying a joint distribution for all quantities (denoted in bold italics) under consideration except known constants. The Bayesian paradigm reduces statistical inference to applied probability. Quantities that become known under sampling (data) are denoted by the T-dimensional vector $\boldsymbol{y} \in \mathbf{Y}$, and the remaining unknown (and unobserved) quantities (parameters) by the m-dimensional vector $\boldsymbol{\theta} \in \Theta \subseteq \Re^{m}$. Unless noted otherwise, $\boldsymbol{y}$ and θ are treated as continuous random variables. Working in terms of densities, consider

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{y})\, f(\mathbf{y}),\ \mathbf{y}, \boldsymbol{\theta} \in \mathbf{Y} \times \Theta, \tag{1}$$

where $f(\theta)$ is the *prior density*, $f(\mathbf{y}|\theta)$ viewed as a function of θ for known y is the *likelihood* function [denoted $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$], $f(\theta|\mathbf{y})$ is the *posterior density*, and

$$f(\mathbf{y}) = \int_{\Theta} f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})\, d\boldsymbol{\theta} = E_{\boldsymbol{\theta}}[\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})],\ \mathbf{y} \in \mathbf{Y}, \tag{2}$$

is the *marginal density* of the data y. From (1), Bayes’ theorem for densities follows:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})}{f(\boldsymbol{\theta})} \propto f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}),\ \boldsymbol{\theta} \in \Theta. \tag{3}$$

Hereafter, (3) is adopted as the way to update prior beliefs when $\boldsymbol{y}$=y is observed.

Fortunately, sometimes the integration in (2) can be performed analytically and so the updating of prior beliefs in light of the data to obtain the posterior beliefs is straightforward. These situations correspond mostly to cases where $\mathcal{L}(\theta; y)$ belongs to the exponential family of densities. In this case the prior density can be chosen so that the posterior density falls within the same elementary family of distributions as the prior. These prior families are called *conjugate families*. Conjugate priors are more flexible than they may appear at first since mixtures of conjugate priors are themselves conjugate, although they may be daunting to elicit.

The denominator in (3) serves as an integrating constant. Hence, when one considers experiments employing the same prior, and which yield proportional likelihoods for the observed data, identical posteriors will emerge, consistent with the *likelihood principle* (Berger and Wolpert, 1988). Unlike the inherent *ex ante* perspective of frequentist statistics, which seeks properties of procedures in repeated sampling, posterior density (3) is *ex post* – it conditions on the observed data $\boldsymbol{y}$=y, and dispenses with the part of the sample space Y that could have been observed but was not.

In most practical situations not all elements of θ are of direct interest. Let $\boldsymbol{\theta} = [\boldsymbol{\beta}', \boldsymbol{\delta}']' \in \mathbf{B} \times \boldsymbol{\Delta}$ be partitioned into *parameters of interest* β and *nuisance parameters* δ. Nuisance parameters are well-named for frequentists, because dealing with them in a general setting is one of the major problems non-Bayesian researchers face. In contrast, Bayesians adopt a universal approach to eliminating nuisance parameters from the problem: integrate them out of the joint posterior to obtain the marginal posterior density for β:

$$f(\boldsymbol{\beta}|\mathbf{y}) = \int_{\boldsymbol{\Delta}} f(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{y})\mathrm{d}\boldsymbol{\delta}, \ \boldsymbol{\beta} \in \mathbf{B}. \tag{4}$$

Point estimation

Consider a *loss* (*cost*) function $C(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ for the parameters of interest β, that is, a nonnegative function satisfying $C(b, b) = 0$ and which measures the consequences of using the estimate $\widehat{\boldsymbol{\beta}}$ when the parameter of interest is β. Both frequentists and Bayesians seek to 'minimize' (in some sense) $C(\widehat{\boldsymbol{\beta}}, b)$, but first its randomness must be eliminated.

From the frequentist point of view, β is a degenerate random variable equal to β, but $C(\widehat{\boldsymbol{\beta}}, b)$ is stochastic because $\widehat{\boldsymbol{\beta}}$ is viewed *ex ante* as the estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{y})$ depending on the data $\boldsymbol{y}$ which are random viewed *ex ante*. One way to circumscribe the randomness of $C(\widehat{\boldsymbol{\beta}}, b)$ is to focus on its expected value, assuming it exists. Frequentists consider the risk function

$$R(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \boldsymbol{\delta}) = E_{\boldsymbol{y}|\boldsymbol{\beta}=\boldsymbol{\beta}, \boldsymbol{\delta}=\boldsymbol{\delta}}[C(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}), \boldsymbol{\beta})], \tag{5}$$

where the expectation is taken with respect to the sampling density f(y|β, δ), y $\in$ Y.

In contrast, the Bayesian perspective is entirely *ex post*, and it seeks a function $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{y})$ of the observed data $\boldsymbol{y}$=y to serve as a point estimate of the parameter of interest β. Unlike the frequentist approach, no role is provided for data that could have been observed, but were not. Since β is unknown, the Bayesian perspective suggests formulation of subjective beliefs about it, given all the information at hand. Such information is fully contained in marginal posterior density (4). In contrast to (5), Bayesians focus on *expected posterior loss*:

$$c(\widehat{\boldsymbol{\beta}}|\mathbf{y}) = E_{b|\boldsymbol{y}=\mathbf{y}}[C(\widehat{\boldsymbol{\beta}}, b)] = \int_{\mathbf{B}} C(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})f(\boldsymbol{\beta}|\mathbf{y})\mathrm{d}\boldsymbol{\beta}. \tag{6}$$

The second Bayesian commandment (after Bayes' theorem) is: act so as to minimize expected posterior loss, that is, find $\hat{\boldsymbol{\beta}}_* \equiv \underset{\hat{\boldsymbol{\beta}}}{\mathbf{argmin}} E_{\beta|y}[C(\hat{\boldsymbol{\beta}}, b)]$. Frequentists emphasize the sampling distribution $y|\beta=\beta, \delta=\delta$ and Bayesians emphasize the posterior distribution $\beta|y=y$. The debate is about the desired conditioning – as are most debates in statistics. Posterior expectation (8) removes β from $C(\hat{\boldsymbol{\beta}}, b)$ yielding a criterion $c(\hat{\boldsymbol{\beta}}|y)$, unlike risk function (5), involving only known quantities.
For simplicity, consider univariate β and the following three loss functions in which c, c_1, c_2, and d are known constants: the *quadratic loss function* $C(\hat{\boldsymbol{\beta}}, b) = (\hat{\boldsymbol{\beta}} - b)^2$, the *asymmetric linear loss function* $C(\hat{\boldsymbol{\beta}}, b) = c_1|\hat{\boldsymbol{\beta}} - b|$, if $\hat{\boldsymbol{\beta}} \le b$, and $C(\hat{\boldsymbol{\beta}}, b) = c_2|\hat{\boldsymbol{\beta}} - b|$ if $\hat{\boldsymbol{\beta}} > b$, and the *all-or-nothing loss function* $C(\hat{\boldsymbol{\beta}}, b) = c$, if $|\hat{\boldsymbol{\beta}} - b| > d$, and $C(\hat{\boldsymbol{\beta}}, b) = 0$, if $|\hat{\boldsymbol{\beta}} - b| \le d$. The resulting Bayesian point estimates are the posterior mean, the qth posterior quantile where $q = \frac{c_1}{c_1 + c_2}$, and the centre of an interval of width 2d having maximum posterior probability (yielding the posterior mode as $d \to 0$), respectively. When β is a vector, the most popular loss functions are the *weighted squared error* generalization of quadratic loss, $C(\hat{\boldsymbol{\beta}}, b) = (\hat{\boldsymbol{\beta}} - b)'\mathbf{Q}(\hat{\boldsymbol{\beta}} - b)$, where Q is a positive definite matrix, or the all-or-nothing loss function. In these cases the Bayesian point estimates are again the posterior mean and mode (as $d \to 0$), respectively.
Minimum risk estimators do not exist in general because (5) depends on β and δ, and so an estimator that minimizes (5) will also depend on β and δ. Often extraneous side conditions are imposed (for example, unbiasedness) to sidestep the problem. In contrast, Bayesian point estimates are optimal by construction from the *ex post* standpoint. In general they also have good *ex ante* risk properties. Consider the minimizer of (6) viewed from the *ex ante* standpoint before the data are realized, that is, the *Bayesian point estimator* $\hat{\boldsymbol{\beta}}_* = \hat{\boldsymbol{\beta}}_*(y)$. Provided the prior distribution is *proper* (it integrates to unity), then $\hat{\boldsymbol{\beta}}_*(y)$ satisfies the minimal frequentist requirement of *admissibility* (its risk cannot be dominated by another estimator everywhere in the parameter space). Furthermore, in most interesting settings, all admissible estimators are either Bayes or limits thereof known as *generalized Bayes estimators* based on an *improper prior* whose integral diverges.

Interval estimation

Bayesian interval estimation follows directly from the posterior density $f(\beta|y)$. Because opinions about the unknown parameter are treated in a probabilistic manner, there is no need to introduce the additional concept of 'confidence'. For example, given a region $\mathbf{B}^\dagger \subset B$, it is meaningful to ask: given the data, what is the *probability* that β lies in $\mathbf{B}^\dagger$? The answer is direct:

$$\mathbf{Prob}(b \in \mathbf{B}^\dagger|y) = \int_{\mathbf{B}^\dagger} f(\boldsymbol{\beta}|y)\, d\boldsymbol{\beta}. \tag{7}$$

Alternatively, given a desired probability content of $1-\alpha$, it is possible to reverse this procedure and find a corresponding region $\mathbf{B}^\dagger$. The 'smallest' region $\mathbf{B}^\dagger$ satisfying (9), known as the *highest posterior density* (*HPD*) *region of content* ($1-\alpha$) *for* β corresponds to imposing the added condition that for all $\beta_1 \in \mathbf{B}^\dagger$ and $\beta_2 \notin \mathbf{B}^\dagger$, $f(\beta_1|y) \ge f(\beta_2|y)$.

Hypothesis testing

Consider a partition of the parameter space B for the parameter of interest β according to $B=B_1 \cup B_2$, where $B_1 \cap B_2$ is null. Suppose interest lies in testing H_1: $\beta \in B_1$ versus H_2: $\beta \in B_2$ based on a sample y yielding the likelihood $\mathcal{L}(\beta, \delta; y)$. The relevant decision space is $D=\{d_1,d_2\}$, where $d_j \equiv$ choose hypothesis H_j (j=1, 2). Extensions to cases involving more than two hypotheses are straightforward. Let $C(d; b) \ge 0$ denote the relevant loss function. Without loss of generality, assume that correct decisions yield zero loss.
From the Bayesian perspective a hypothesis is of interest only if the prior distribution assigns it positive probability. Therefore, assume $\pi_j \equiv \mathbf{Prob}(H_j) = \mathbf{Prob}(b \in \mathbf{B}_j) > 0$ **(j=1,2)** with $\pi_1 + \pi_2 = 1$. Let $f_j(\beta, \delta|H_j)$ be the prior density under H_j (j=1, 2).

Under H_j, the marginal data density (*expected likelihood*) is

$$f(\mathbf{y}|\mathbf{H}_j) = \int_{\boldsymbol{\Delta}} \int_{\mathbf{B}_j} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\delta}; \mathbf{y})\, dF_j(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{H}_j) = E_{b,d|\mathbf{H}_j}[\mathcal{L}(b, d; \mathbf{y})]\ (\mathbf{j} = 1, 2), \tag{8}$$

where $F_j(\cdot)$ denotes the c.d.f. corresponding to the distribution β, δ | H_j. From Bayes' theorem it follows that the posterior probability of H_j is

$$\bar{\pi}_j = \mathrm{Prob}(H_j|y) = \frac{\underline{\pi}_j f(y|H_j)}{f(y)}\ (j = 1, 2), \tag{9}$$

where the marginal density of the data is $f(\mathbf{y}) = \underline{\pi}_1 f(\mathbf{y}|H_1) + \underline{\pi}_2 f(\mathbf{y}|H_2)$. Under H_j, the posterior density of β and δ is (according to Bayes' theorem):

$$f(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{y}, \mathbf{H}_j) = \frac{f_j(\boldsymbol{\beta}, \boldsymbol{\delta}|H_j)\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\delta}; \mathbf{y})}{f(\mathbf{y}|H_j)},\ \boldsymbol{\beta} \in \mathbf{B}_j,\ \ \boldsymbol{\delta} \in \boldsymbol{\Delta}\ (j = 1, 2). \tag{10}$$

As in the case of estimation, the optimal Bayesian decision $\mathbf{d}_*$ in the hypothesis testing context minimizes expected posterior loss, that is, $\mathbf{d}_* \equiv \underset{d}{\mathrm{argmin}}\, c(d|\mathbf{y})$, where

$$c(d|\mathbf{y}) = \bar{\pi}_1 c(d|\mathbf{y}, H_1) + \bar{\pi}_2 c(d|\mathbf{y}, H_2), \tag{11}$$

and $c(d|y,H_j)=E_{\theta|y,H_j}[C(d;\theta)]$ (j=1,2). Specifically, $c(d_1|\mathbf{y}) = \bar{\pi}_2 c(d_1|\mathbf{y}, H_2)$, and $c(d_2|\mathbf{y}) = \bar{\pi}_1 c(d_2|\mathbf{y}, H_1)$. Therefore, it is optimal to choose $\mathbf{H}_2$[that is, $c(d_2|y)<c(d_1|y)$] iff

$$\mathbf{d}_* = d_2 \text{ iff } \frac{\bar{\pi}_2}{\bar{\pi}_1} > \frac{c(d_2|\mathbf{y}, H_1)}{c(d_1|\mathbf{y}, H_2)} \tag{12a}$$

The quantities $\frac{\underline{\pi}_2}{\underline{\pi}_1}$ and $\frac{\bar{\pi}_2}{\bar{\pi}_1}$ are the *prior odds and posterior odds*, respectively, of H_2, versus H_1. From (9) it follows immediately that these two odds are related by $\frac{\bar{\pi}_2}{\bar{\pi}_1} = B_{21}(\frac{\underline{\pi}_2}{\underline{\pi}_1})$, where $B_{21} = \frac{f(y|H_2)}{f(y|H_1)}$ is the *Bayes factor for H_2 versus H_1*. See Kass and

Raftery (1995) for an excellent review. In terms of the Bayes factor B_{21}, (12a) can also be written

$$\mathbf{d}_* = d_2 \text{ iff } B_{21} \geq \left[\frac{c(d_2|\mathbf{y}, H_1)}{c(d_1|\mathbf{y}, H_2)}\right]\left[\frac{\pi_1}{\pi_2}\right]. \quad (12b)$$

In general, expected posterior loss $c(d|y, H_j)$ depends on the data y, and hence, Bayes factor B_{21} does *not* serve as complete data summary because the right-hand side of the inequality in (12b) also depends on the data. One exception is when both hypotheses are simple. Another is when an all-or-nothing loss is used such that the loss $\mathbf{C}(d_i, b) = \overline{C}_i$ resulting from decision d_i when $\beta \in B_j$, $i \neq j$, is constant for all $\beta \in B_j$. In this case, for $i \neq j$, $c(d_1|y, H_j) = c(d_i|y, H_j) = \overline{C}_i$, and decision rule (12b) reduces to

$$\mathbf{d}_* = d_2 \text{ iff } B_{21} > \frac{\pi_1 \overline{C}_2}{\pi_2 \overline{C}_1}. \quad (12c)$$

The right-hand side of the inequality in (12c) is a known constant *Bayesian critical value*.

Prediction

The sampling distribution of an out-of-sample $\tilde{y} \in \tilde{Y}$ given y=y and θ, would be an acceptable predictive distribution if θ was known, but without knowledge of θ it cannot be used. In its place is the *Bayesian predictive density*

$$f(\tilde{y}|\mathbf{y}) = \frac{f(\tilde{y}, \mathbf{y})}{f(\mathbf{y})} = \int_{\Theta} \frac{f(\tilde{y}, \mathbf{y}, \boldsymbol{\theta})}{f(\mathbf{y})} d\boldsymbol{\theta} = \int_{\Theta} f(\tilde{y}|\mathbf{y}, \boldsymbol{\theta})\left[\frac{f(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y})}\right] d\boldsymbol{\theta} = \int_{\Theta} f(\tilde{y}|\mathbf{y}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = E_{q|\boldsymbol{y}}[f(\tilde{y}|\mathbf{y}, \boldsymbol{\theta})], \tilde{y} \in \tilde{Y} \quad (13)$$

If the past and future are independent conditional on θ (as in random sampling), then $f(\tilde{y}|\mathbf{y}, \boldsymbol{\theta}) = f(\tilde{y}|\boldsymbol{\theta})$ Letting $\tilde{C}(\tilde{y}_{\mathbf{p}}, \tilde{y})$ denote a *predictive loss function* measuring the performance of a predictor $\tilde{y}_{\mathbf{p}}$ of $\tilde{y}$, the optimal point predictor $\tilde{y}_*$ is defined to be $\tilde{y}_* \equiv \underset{\tilde{y}_{\mathbf{p}}}{\operatorname{argmin}} E_{\tilde{y}|\mathbf{y}}[\tilde{C}(\tilde{y}_{\mathbf{p}}, \tilde{y})]$. For example, if $\tilde{y}$ is a scalar and predictive loss is quadratic, then the optimal point estimate is the predictive mean $\tilde{y}_* = E(\tilde{y}|\mathbf{y})$. Predictive density (13) can also be used to generate forecast intervals analogous to HPD intervals. Predictive density (13) treats all parameters as nuisance parameters and integrates them out of the predictive problem. A similar strategy is used when adding parametric hypotheses to the analysis. Consider the hypothesis H_j and associated prior $f_j(\beta, \delta|H_j)$ (j=1, 2). Given data y leading to the posterior $f_j(\beta, \delta|y, H_j)$, the predictive density of $\tilde{y}$ conditional on H_j is

$$f(\tilde{y}|\mathbf{y}, H_j) = \int_{\Theta} f(\tilde{y}|\boldsymbol{\theta}, \mathbf{y}, H_j)\, f(\boldsymbol{\theta}|\mathbf{y}, H_j)\, d\boldsymbol{\theta} \tilde{y} \in \tilde{Y}. \quad (14)$$

Using the posterior probabilities (9), the marginal predictive density of $\tilde{y}$ is the mixture density

$$f(\tilde{y}|y) = \bar{\pi}_1 f(\tilde{y}|y, H_1) + \bar{\pi}_2 f(\tilde{y}|y, H_2), \tilde{y} \in \tilde{Y} \quad (15)$$

and it is the basis for interval and point prediction. For example, under quadratic loss the optimal Bayesian point prediction is the predictive mean

$$E(\tilde{y}|y) = \bar{\pi}_1 E(\tilde{y}|y, H_1) + \bar{\pi}_2 E(\tilde{y}|y, H_2), \quad (16)$$

which is a weighted average of the optimal point forecasts $E(\tilde{y}|y, H_1)$ under each hypothesis. The weights $\bar{\pi}_j (j = 1, 2)$ in (16) have an intuitive appeal: the forecast of the more probable hypothesis a posteriori receives more weight.

Choice of prior

Critics of Bayesianism find the choice of prior is the major stumbling block in adopting the Bayesian approach. In contrast, proponents see the required effort to be manageable and well worth it. Usually the likelihood is parameterized to facilitate thinking in terms of θ, and so subject matter considerations should suggest 'plausible' values of θ. Even when such direct thinking about θ is possible, it is also useful to think predictively (for example, see Kadane and Wolfson, 1998) about the observable ***y*** and use (2) to back out a parametric prior $f(\theta|\lambda)$ for a specific value of some hyperparameter $\lambda \in \Lambda$ in some space Λ. Usually such analyses restrict attention to conjugate priors. This ideal, however, is difficult to achieve.

Public research involving only a single prior is likely to draw few readers. Entertaining various professional positions in terms of θ can lead to different choices of λ. Rather than thinking of eliciting *the* prior, it is more useful to think in terms of a family $\mathcal{F} = \{f(\theta|\lambda), \lambda \in \Lambda\}$ of parametric priors. If a prior $f(\lambda)$ is available for λ, then we are back in the single prior case with *the* prior $f(\theta) = \int_\Lambda f(\theta) f(\lambda) d\lambda$. In most practical problems, however, there will be no agreed upon $f(\lambda)$, and the researcher is left with investigating the sensitivity of the analyses to different elements in $\mathcal{F}$. This is easier said than done, but in principle it can be done. For large dimensional θ, this can be difficult because the effects of the prior can be subtle: it may have little posterior influence on some functions of the data and have an overwhelming influence on other functions. Often a quantity of interest like the posterior mean $E(\theta|y)$ can be analytically restricted to a fairly small set of possible values for any given $\lambda \in \Lambda$. The *extreme bounds analysis* developed by Leamer (1982) is a leading example. In contrast, *empirical Bayes analysis* proceeds by using the data to estimate λ.

Kass and Wasserman (1996) survey formal rules that have been suggested for choosing a prior. Many of these rules reflect the desire to let the 'data speak for themselves'. This has led to variety of non-subjective priors intended to capture the elusive notion of *non-informativeness*. These priors are intended to lead to proper posteriors dominated by the data. They also serve as benchmarks for posteriors derived from ideal subjective considerations. At first many of these priors were also motivated on simplicity grounds. But as problems were discovered, and other features were seen to be relevant, derivation of such priors became more complicated, possibly even more so than a legitimate attempt to elicit an actual subjective prior.

One interpretation of letting the data speak for themselves is to use classical techniques. Maximum likelihood estimates are rationalizable in a Bayesian framework by appropriate choice of prior distribution and loss function, specifically a uniform prior and an all-or-nothing loss function. But in what parameterization should one be uniform?

In order to overcome the re-parameterization problem, Jeffreys sought a general rule for choosing a prior so that the same posterior inferences were obtained regardless of the parameterization chosen. Jeffreys (1961) made a general (but not dogmatic) argument in favor of choosing a prior proportional to the square root of the information matrix, that is, $f(\theta) \propto |J(\theta)|^{1/2}$, where $J(\theta) \equiv E_{y|\theta}[-\partial^2 L(\theta; y)/\partial\theta\partial\theta']$ is the information matrix of the sample. This prior has the desirable feature that if the model is reparameterized by a one-to-one transformation, say $\psi = h(\theta)$, then choosing the prior $f(\psi) \propto |E_{y|\alpha}[-\partial^2 L(\psi; y)/\partial\psi\partial\psi']|^{1/2}$

will lead to identical posterior inferences as using f(θ). Such priors are said to follow *Jeffreys' rule*.

Not all of Jeffreys' recommendations always followed Jeffreys rule: When Θ is finite, Jeffreys assigned equal probabilities to each of the values. When Θ is a bounded interval, Jeffreys assumed a constant proper prior. When $\Theta = \Re$, Jeffreys assumed a constant improper prior. When Θ=[0,∞), Jeffreys chose f(θ)= θ^{-1} because it is invariant under power transformations. When θ =[θ_1, θ_2] $'$ where θ_1 is a location parameter and θ_2 is a non-location parameter, Jeffreys chose $\mathbf{f}(\boldsymbol{\theta}) \propto |\mathbf{J}(\boldsymbol{\theta})|^{1/2}$, where J(θ) is calculated holding θ_1 fixed. In the case of mixture models, Jeffreys argued that the mixing parameters should be treated independently from the other parameters. There is a fair amount of agreement that such priors may be reasonable in one-parameter problems, but substantially less agreement (including Jeffreys) in multiple parameter problems.

Usually, Jeffreys' rule and other formal rules surveyed by Kass and Wasserman (1996), lead to *improper priors*, that is, priors which integrate to infinity rather than unity (a *proper prior*). When blindly plugged into Bayes' theorem as a prior they lead to proper posterior densities, but not always. They also produce proper predictive densities (13), but *not* proper marginal data densities (8). Furthermore, improper priors, in contrast to proper priors, are not guaranteed to lead to admissible Bayesian point estimators, and marginalization paradoxes can occur.

Bernardo (1979) suggested a method for constructing *reference priors* offering two innovations. First, he defined a notion of missing information in terms of the *Kullbach-Leibler distance* between the posterior and the prior density. Second, he developed a stepwise procedure for handling *nuisance parameters*. If there are no nuisance parameters, then his method usually leads to Jeffreys' rule. Subsequently, numerous refinements have been made in joint work with James O. Berger.

There are many candidates for non-subjective priors, and they often have properties that seem rather non-Bayesian. Most non-subjective priors depend on some or all of the following: (*a*) the form of the likelihood, (*b*) the sample size, (*c*) an expectation with respect to the sampling distribution, (*d*) the parameters of interest, and (*e*) whether the researcher is engaging in estimation, testing or predicting. The dependency in (*c*) of Jeffreys' prior on a sampling theory expectation makes it sensitive to a host of problems related to the likelihood principle. In light of (*d*), a *non-subjective* prior can depend on *subjective* choices such as which are the parameters of interest and which are nuisance parameters. Different quantities of interest require different non-subjective priors which cannot be combined in a coherent manner. My advice is use a non-subjective prior only with great care, and never alone. I include non-subjective priors in the class of priors over which I perform a sensitivity analysis.

One reaction to choice of prior is to not make one and proceed with an asymptotic analysis. The same way sampling distributions of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ in regular situations is asymptotically normal, posterior density (5) can be approximated as T $\rightarrow \infty$ by the multivariate normal density $\varphi_{\mathbf{m}}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{\mathbf{ML}}, [\mathbf{J}_{\mathbf{T}}(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})]^{-1}) = (2\pi)^{-m/2}|J_T(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})|^{1/2}\exp[-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})'J_T(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})]$, where $\mathbf{J}_{\mathbf{T}}(\cdot)$is the information matrix. This approximation does not depend on the prior. As an approximation to the posterior density of θ , the approximation usually improves by replacing the information matrix by the observed Hessian of the log-likelihood evaluated at $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$.The quality of this approximation can usually be improved by incorporating some information on the prior. For example, by using $\varphi_{\mathbf{m}}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}, [\overline{\mathbf{H}}_{\mathbf{T}}(\widehat{\boldsymbol{\theta}})]^{-1})$, where $\widehat{\boldsymbol{\theta}}$ is the posterior mode and $\overline{H}_T(\widehat{\boldsymbol{\theta}})$ is the Hessian of the log posterior evaluated at $\widehat{\boldsymbol{\theta}}$. Further asymptotic analysis using *Laplace approximations* (see Tierney and Kadane, 1986) often given remarkably accurate results.

Model building

A 'true model' is an oxymoron. An economic model is an abstract representation of reality that highlights what a researcher deems relevant to a particular economic issue. By definition an economic model is literally false, and so questions regarding its literal truth are trivial. Whether the model is useful is another matter.

A subjectivist's *econometric model* expresses probabilistically the researcher's beliefs concerning future observables of interest to economists. It has two components: a likelihood for viewing observables in the world, and a prior reflecting a professional position of interest. Poirier (1988) introduced the metaphor *window* for a likelihood function because it captures its essential role in de Finetti representation theorems: a parametric medium for viewing the observable world. Both model components are subjective, and both involve mathematical constructs called *parameters*. Parameters simply index distributions; any correspondence to physical reality is a rare side bonus.

In choosing the window $\mathcal{L}(\theta; Y)$ the researcher is torn in two directions: choosing the dimensionality of θ to be large increases the chances of getting a bevy of researchers *to agree to disagree* in terms of the appropriate priors for θ , but a large dimensional θ necessitate increasingly more informative priors if anything useful is to be learned from a finite sample. In one sense this dichotomy between prior and likelihood is tautological: if there is no agreement, then presumably the likelihood can always be expanded until agreement is obtained. The resulting window, however, may be hopelessly complex. The 'bite' in the statement comes from the

assertion that a researcher believes agreement is compelling in the case of a particular window. Despite the many arguments in the literature over the wisdom of ‘general to specific’ as opposed to ‘specific to general modelling’, observed behaviour suggests researchers start with a finite parameterization of the problem that can be both simplified and expanded. The arguments are really over a matter of emphasis rather than kind.

Diagnostic checking of the maintained initial window can help achieve agreement on it. If the diagnostic checks indicate window expansion, then rethinking is required, a new window must be introduced, and the diagnostic checking process repeated. The extent of diagnostic testing depends in part on the size of the initial window. Everything else being equal, small windows require more checking to convince others of their value than large windows. Reporting that the initial window passes diagnostic checks is intended to soothe the concerns of members of the research community. For good discussions of diagnostic checking, see Gelman et al. (2003) and Lancaster (2004). Such checking can be as much an art as a science.

Conscientious empirical researchers provide their readers with a variety of ways of looking at the data. This amounts to checking how the observed data fit marginal density (2), how out-of-sample observables fit predictive densities (13) or (15), and how posterior densities (3) or (10) are summarized and interpreted. This task is complicated when m is large or when many hypotheses are entertained. Furthermore, the question arises: ‘How should we bring together the results?’ Is one hypothesis is to be chosen after an ‘enlightened’ search of the data? If so, then the question is how to properly express uncertainty that reflects both sampling uncertainty from estimating the unknown parameters under a hypothesis and uncertainty over the hypothesis itself. The common practice of choosing a single hypothesis and then proceeding conditionally on it, is difficult to rationalize because the researcher's uncertainty is understated unless that hypothesis has a posterior probability near unity. Readers are interested in a clear articulation of the researcher's uncertainty because it can serve as a useful gauge or reference point for their own uncertainty.

When considering two hypotheses H_1 and H_2 it is possible to assign only $\underline{\pi}_1 + \underline{\pi}_2 = 1 - \varepsilon$ prior probability to them, and to reserve ε $(0< \varepsilon <1)$ probability for an unspecified H_3 representing ‘something else’. Then interpreting $\underline{\pi}_j$ relatively as $\text{Prob}(H_j|H_1 \text{ or } H_2)$ (j=1, 2), posterior probabilities (11) can be computed and also interpreted relatively as $\text{Prob}(H_j|y, H_1 \text{ or } H_2)$ *without specifying* ε. If in the process the researcher's creative mind has a new insight leading to specification of ‘something else’, then some fraction $\underline{\pi}_3$ of $1 - \varepsilon$ can be allocated to H_3 and the process repeated with the remaining portion allocated to a another unspecified H_4. The catch here is that H_3 is data-instigated (that is, created after looking at the data), and the researcher faces choice of a ‘post-data prior’ involving both $\underline{\pi}_3$ and any parameters unrestricted under H_3. However, the need for sensitivity analysis in public research implies the researcher is simply left with the usual task of presenting a variety of mappings from ‘interesting’ priors to posteriors. It is left to the reader to decide whether the priors are sufficiently plausible to warrant serious consideration of the data instigated hypothesis. Priors that have been contaminated by data can be presented as such – as always it remains for the reader to assess their plausibility.

Regression

To illustrate the preceding discussion, consider the standard normal linear regression model with fixed regressors X yielding likelihood function $\mathcal{L}(\theta; y) = \varphi_T(y|X\beta, \sigma^2 I_T)$, where $\theta = [\beta', \sigma^{-2}]'$ and β is K•×•1 parameter of interest. Working in terms of the precision σ^{-2}, the conjugate normal-gamma prior is

$$f(\beta, \sigma^{-2}) = \Phi_K(\beta|\underline{b}, \sigma^2\underline{Q})\gamma(\sigma^{-2}|\underline{s}^2, \underline{\nu}), \tag{17}$$

where $\gamma(\sigma^{-2}|\underline{s}^2, \underline{\nu}) = [2/(\underline{\nu}\underline{s}^2)^{\underline{\nu}/2}\Gamma(\underline{\nu}/2)]^{-1}(\sigma^{-2})^{(\underline{\nu}-2)/2}\exp[-(\frac{\underline{\nu}\underline{s}^2}{2})\sigma^{-2}]$ is a gamma density with mean $\underline{s}^{-2}$ and variance $\frac{2}{\underline{\nu}\underline{s}^4}$, $\Gamma(\cdot)$ denotes the gamma function, $\underline{b}$ is a K × 1 vector, $\underline{Q}$ is a K × K positive definite matrix, $\underline{s} > 0$, and $\underline{\nu} > 0$.

It is the straightforward to show that (5) implies the normal-gamma posterior distribution

$$f(\beta, \sigma^{-2}|y) = \varphi_K(\beta|\bar{b}, \sigma^2\bar{Q})\gamma(\sigma^{-2}|\bar{s}^2, \bar{\nu}), \tag{18}$$

where $\bar{b} = \bar{Q}(\underline{Q}^{-1}\underline{b} + X'Xb)$, $\bar{Q} = (\underline{Q}^{-1} + X'X)^{-1}$, $\bar{\nu} = \underline{\nu} + T$ and $\bar{\nu}\bar{s}^2 = \underline{\nu}\underline{s}^2 + (y - Xb)'(y - Xb) + (b - \underline{b})'[\underline{Q} + (X'X)^{-1}]^{-1}(b - \underline{b})$. The marginal posterior distribution of β is the multivariate-t density

$$f(\beta|y) = \left[\frac{\pi^{K/2}\Gamma\left(\frac{\nu}{2}\right)}{\bar{\nu}^{\nu/2}\Gamma\left(\frac{\nu+K}{2}\right)}\right]^{-1} \left|\bar{s}^2\bar{Q}\right|^{-1/2}\left[\bar{\nu} + (\beta - \bar{b})'(\bar{s}^2\bar{Q})^{-1}(\beta - \bar{b})\right]^{-(\nu+K)/2},$$

(19)

with mean $\bar{b}$ (if $\bar{\nu} > 1$), variance $\left(\frac{\nu}{\nu-1}\right)|\bar{s}^2\bar{Q}$ (if $\bar{\nu} > 2$), and $\bar{\nu}$ degrees of freedom. The marginal density of the data can be written

$$f(y) = \pi^{-T/2}\left[\frac{\Gamma[(T + \underline{\nu})/2]}{\Gamma[\underline{\nu}/2]}\right]\left[\frac{|\underline{Q}^{-1}|}{|\bar{Q}^{-1}|}\right]^{1/2}\left[\underline{s}^2\right]^{\underline{\nu}/2} \cdot \left[\underline{s}^2 + (y - Xb)'(y - Xb) + (b - \bar{b})'X'X(b - \bar{b}) + (\underline{b} - \bar{b})'\underline{Q}^{-1}(\underline{b} - \bar{b})\right].$$

Furthermore, the predictive density of an out-of-sample observation $\tilde{y} \in \tilde{Y}$ corresponding to the regressors $\tilde{x}$ is the university t density

$$f(\tilde{y}|y) = \left[\frac{\pi^{1/2}\Gamma\left(\frac{\nu}{2}\right)}{\bar{\nu}^{\nu/2}\Gamma\left(\frac{\nu+1}{2}\right)}\right]^{-1} \left|\tilde{s}^2\right|^{-1/2}\left[\bar{\nu} + \tilde{s}^{-2}(\tilde{y} - \tilde{x}'\bar{b})'(\tilde{y} - \tilde{x}'\bar{b})\right]^{-(\nu+1)/2},$$

(20)

where $\tilde{s}^2 = \bar{s}^2(1 + \tilde{x}'\bar{Q}^{-1}\tilde{x})$.

Note that no full column rank assumption for X is required for the preceding analysis. This reflects a general result that unidentifiability of a parameter, such as β when rank(X)<K, is not much of a problem for a Bayesian with a proper prior, because the posterior is guaranteed to be proper. There is no ‘free lunch’, however, because there will exist some quantity η about which no learning occurs, that is, $f(\eta|y)=f(\eta)$. For example, if Xc=0 for some nonzero K•×•1 vector c, then the prior and posterior distributions for $\eta = c'\underline{Q}^{-1}\beta$ given σ^2 is univariate normal with mean $c'\underline{Q}^{-1}\underline{b}$ and variance $\sigma^2 c'\underline{Q}^{-1}c$. Whether lack of updating is a problem depends on whether η is a quantity of interest. Note that η depends on both the nature of the collinearity (through c) and the prior (through) $\underline{Q}$.

Under weighted squared error loss, the Bayesian point estimate of β is the posterior mean $\bar{b}$. The matrix weighted average of $\underline{b}$ and is $\bar{b}$ is precisely the way a classicist combines two samples from the same distribution: a fictitious sample yielding an OLS estimate $\underline{b}$ with $\mathrm{Var}(\underline{b}|\sigma^2) = \sigma^2\underline{Q}$, and an actual sample yielding the OLS estimate b with $\mathrm{Var}(b|\sigma^2) = \sigma^2(X'X)^{-1}$. Elliptical HPD regions for β can be formed using (21). Bayes factors for hypothesis tests involving restrictions on β can be formed from versions of marginal likelihood (22). Finally, under quadratic loss the Bayesian point prediction of $\tilde{y}$ is $\tilde{y}_* = \tilde{x}'\bar{b}$ and forecast intervals can be

obtained directly from the predictive distribution (23).
The standard 'noninformative' prior is $f(\beta, \sigma^{-2}) \propto \sigma^{2}$, which, unlike the conjugate case, is predicated on the independence of prior beliefs concerning β and σ^{-2}. For this prior, under weighted squared error loss, the Bayesian point estimate of β is the OLS estimate b. HPD regions are numerically identical to frequentist confidence regions of the same level. Under quadratic loss the Bayesian point prediction of $\tilde{y}$ is $\tilde{y}_* = \tilde{x}'\bar{b}$ and forecast intervals are numerically identical to frequentist forecast intervals of the same level. Bayes factors, however, are not well defined in this case since the prior is improper, and as a result the Bayes factor involves a ratio of arbitrary constants. One class of alternatives in this case are the *intrinsic Bayes factors*, proposed by Berger and Pericchi (1996), which sometimes correspond to actual Bayes factors for particular proper priors known as *intrinsic priors*.

Conclusion

The coherence of the Bayesian approach contrasts sharply with the conventional statistical methods which sometimes advocate negative estimators of positive quantities to ensure unbiasedness, and confidence intervals which may be null or consist of the whole parameter space. Furthermore, Bayesian methods are completely general and do not require usual regularity conditions, asymptotics, sufficient statistics of finite dimension, or pivotal quantities.
There are now a number of textbook sources for Bayesian econometrics. Bayesian econometrics textbooks started with the major contribution of Zellner (1971). While not a textbook as such, Leamer (1978) remains a transparent introduction to Bayesian thinking. Poirier (1995) provides an intermediate level comparison of Bayesian and frequentist reasoning. More recently, Bauwens, Lubrano and Richard (1999), Koop (2003), Koop et al. (2007), Lancaster (2004), and Geweke (2005) have covered extensively the statistical models of direct interest to economists. These four texts also serve as excellent introductions to modern computational techniques. Finally, Koop, Poirier and Tobias (2006) provides extensive solved Bayesian exercises.

See Also

- Bayesian statistics
- Bayesian time series analysis
- Markov chain Monte Carlo methods

Bibliography

Bauwens, L., Lubrano, M. and Richard, J.-F. 1999. *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Berger, J.O. and Pericchi, L.R. 1996. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91, 109–22.

Berger, J.O. and Wolpert, R.L. 1988. *The Likelihood Principle*, 2nd edn. Hayward, CA: Institute of Mathematical Statistics.

Bernardo, J.M. 1979. Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society*, Series B 41, 113–47.

Bernardo, J.M. and Smith, A.F.M. 1994. *Bayesian Theory*. New York: Wiley.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 2003. *Bayesian Data Analysis*, 2nd edn. New York: Chapman & Hall.

Geweke, J. 2005. *Contemporary Bayesian Econometrics and Statistics*. Hoboken: Wiley.

Jeffreys, H. 1961. *Theory of Probability*, 3rd edn. London: Oxford University Press.

Kadane, J.B. and Wolfson, L.J. 1998. Experiences in elicitation. *Statistician* 47, 3–19.

Kass, R.E. and Raftery, A.E. 1995. Bayes factors. *Journal of the American Statistical Association* 90, 773–95.

Kass, R.E. and Wasserman, L. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91, 1343–70.

Koop, G. 2003. *Bayesian Econometrics*. Chichester: Wiley.

Koop, G., Poirier, D.J. and Tobias, J. 2007. Bayesian econometric methods. In *Econometrics Exercises Series*, vol. 7, ed. K. Abadir, J. Magnus, and P.C.B. Phillips. Cambridge: Cambridge University Press.

Lancaster, T. 2004. *An Introduction to Modern Bayesian Econometrics*. Oxford: Blackwell.

Leamer, E.E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Leamer, E.E. 1982. Sets of posterior means and bounded variance priors. *Econometrica* 50, 725–36.

Poirier, D.J. 1988. Frequentist and subjectivist perspectives on the problems of model building in economics (with discussion). *Journal of Economic Perspectives* 2(1), 121–70.

Poirier, D.J. 1995. *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press.

Press, S.J. and Tanur, J.M. 2001. *The Subjectivity of Scientists and the Bayesian Approach*. New York: Wiley.

Tierney, L. and Kadane, J.B. 1986. Accurate approximations for posterior moments and marginal posterior densities. *Journal of the American Statistical Association* 81, 82–6.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

How to cite this article

Poirier, Dale J. "Bayesian econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000322> doi:10.1057/9780230226203.0108

The New Palgrave Dictionary of Economics Online

Bayesian methods in macroeconometrics

Frank Schorfheide
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses how Bayesian methods can be used to cope with challenges that arise in the econometric analysis of dynamic stochastic general equilibrium models and vector autoregressions.

Keywords

Bayes' theorem; Bayesian methods in macroeconometrics; business cycles; calibration; Cowles Commission; dynamic macroeconomics; dynamic stochastic general equilibrium (DSGE) models; economic growth; estimation; expectations; identification; statistical inference; intertemporal optimization problems; joint probability distributions; likelihood functions; macroeconometrics; misspecification; monetary policy shocks; neoclassical growth model; probability; rational expectations; structural change; system-of-equations models; technology shocks; vector autoregressions; linear models; Markov chain Monte Carlo methods; stochastic growth models; habit formation; maximum likelihood; posterior probability; regime-switching models; latent state variables; state-space models

Article

Macroeconometrics encompasses a large variety of probability models for macroeconomic time series as well as estimation and inference procedures to study the determinants of economic growth, to examine the sources of business cycle fluctuations, to understand the propagation of shocks, to generate forecasts, and to predict the effects of economic policy changes. Bayesian methods are a collection of inference procedures that permit researchers to combine initial information about models and their parameters with sample information in a logically coherent manner by use of Bayes' theorem. Both prior and post-data information is represented by probability distributions.

Unfortunately, the term 'macroeconometrics' is often narrowly associated with large-scale system-of-equations models in the Cowles Commission tradition that were developed from the 1950s to the 1970s. These models came under attack on academic grounds in the mid 1970s. Lucas (1976) argued that the

models are unreliable tools for policy analysis because they are unable to predict the effects of policy regime changes on the expectation formation of economic agents in a coherent manner. Sims (1980) criticized the fact that many of the restrictions that are used to identify behavioural equations in these models are inconsistent with dynamic macroeconomic theories and proposed the use of vector autoregressions (VARs) as an alternative. Academic research on econometric models in the Cowles tradition reached a trough in the early 1980s and never recovered. The state-of-the-art is summarized in a monograph by Fair (1994).

I am adopting a modern view of macroeconometrics in this article and will portray an active research area that is tied to modern dynamic macroeconomic theory. Reviewing Bayesian methods in macroeconometrics in a short essay is a difficult task. My review is selective and not representative of Bayesian time-series analysis in general. I have chosen some topics that I believe are important, but the list is by no means exhaustive. I focus on the question how Bayesian methods are used to address some of the challenges that arise in the econometric analysis of dynamic stochastic general equilibrium (DSGE) models and VARs. A more extensive treatment can be found in the survey article by An and Schorfheide (2007).

DSGE models

The term ‘DSGE model’ is often used to refer to a broad class of dynamic macroeconomic models that spans the standard neoclassical growth model discussed in King, Plosser and Rebelo (1988) as well as the monetary model with numerous real and nominal frictions developed by Christiano, Eichenbaum and Evans (2005).

A common feature of these models is that decision rules of economic agents are derived from assumptions about preferences and technologies by solving intertemporal optimization problems. Moreover, agents potentially face uncertainty with respect to, for instance, total factor productivity or the nominal interest rate set by a central bank. This uncertainty is generated by exogenous stochastic processes or shocks that shift technology or generate unanticipated deviations from a central bank's interest-rate feedback rule. Conditional on distributional assumptions for the exogenous shocks, the DSGE model generates a joint probability distribution for the endogenous model variables such as output, consumption, investment, and inflation.

What are the goals?

While macroeconometric methods are used to address many different questions, several issues stand out. Business cycle analysts are interested in identifying the sources of fluctuations; for instance, how important are monetary policy shocks for movements in aggregate output? We would like to understand the propagation of shocks; for example, what happens to aggregate hours worked in response to a technology shock? Moreover, researchers ask questions about structural changes in the economy: has monetary policy changed in the early 1980s? Why did the volatility of many macroeconomic time series drop in the mid 1980s? Macroeconometricians are also interested in forecasting the future: how will inflation and output growth rates evolve over the next eight quarters? Finally, an important aspect of macroeconometrics is to predict the effect of policy changes: how will output and inflation respond to an

unanticipated change in the nominal interest rate? Is it desirable to adopt an inflation targeting regime?

What are the challenges?

In principle one could proceed as follows: specify a DSGE model that is sufficiently rich to address the substantive economic question of interest; derive its likelihood function and fit the model to historical data; answer the questions based on the estimated DSGE model. Unfortunately, this is easier said than done. A trade-off between theoretical coherence and empirical fit poses the first challenge to macroeconometric analysis.

Under certain regularity conditions DSGE models can be well approximated by VARs that satisfy particular cross-coefficient restrictions. The DSGE model is misspecified if these restrictions are at odds with the data and the model has difficulties in tracking and forecasting historical time series. Misspecification was quite apparent for the first generation of DSGE models and has led Kydland, Prescott, and their followers since the early 1980s to abandon formal econometric procedures and advocate a calibration approach, outlined for instance in Kydland and Prescott (1996). Recent Bayesian and non-Bayesian research, however, has resulted in formal econometric tools that are general enough to explicitly account for misspecification problems that arise in the context of DSGE models. Examples of Bayesian approaches are Canova (1994), Dejong, Ingram, and Whiteman (1996), Geweke (1999), Schorfheide (2000), Del Negro and Schorfheide (2004), and Del Negro et al. (2006).

The presence of misspecification might suggest that we should simply ignore the cross-coefficient restrictions implied by dynamic economic theories in the empirical work and try to answer the questions posed above directly by VARs. Unfortunately, there is no free lunch. VARs have many free parameters, and without restrictions on their coefficients tend to generate poor forecasts. VARs do not provide a tight economic interpretation of economic dynamics in terms of the behaviour of rational, optimizing agents. Moreover, it is difficult to predict the effects of rare policy regime changes on the expectation formation and the behaviour of economic agents since these are not explicitly modelled. While the most recent generation of DSGE models comes much closer to matching the empirical fit of VARs, as documented in Smets and Wouters (2003), a trade-off between theoretical coherence and empirical fit remains.

A second challenge is identification. The parameters of a model are identifiable if no two parameterizations of that model generate the same probability distribution for the observables. In VARs the mapping between the one-step-ahead forecast errors of the endogenous variables and the underlying structural shocks is not unique, and additional restrictions are necessary to identify, say, a monetary policy or a technology shock. Many of the popular identification schemes and the controversies surrounding them are surveyed in Cochrane (1994), Christiano and Eichenbaum (1999) and Stock and Watson (2001).

DSGE models can be locally approximated by linear rational expectations (LRE) models. While tightly parameterized compared to VARs, LRE models can generate delicate identification problems. Suppose a model implies that $y_t = \theta E_t[y_{t-1}] + u_t$, where u_t is an independently distributed random variable with mean zero. If $0 \leq \theta < 1$, then the only stable law of motion for y_t that satisfies the rational expectations restrictions is $y_t = u_t$, which means that θ is not identifiable. More elaborate examples are discussed in Beyer and Farmer (2004), Lubik and Schorfheide (2004; 2006), and Canova and Sala (2006).

Unfortunately, it is in many cases difficult to detect identification problems in DSGE models, since the mapping from the structural parameters into the autoregressive law of motion for y_t is highly nonlinear and typically can be evaluated only numerically.

Many regularities of macroeconomic time series are indicative of nonlinearities, for instance, the rise and fall of inflation in the 1970s and early 1980s and time-varying volatility of many macroeconomic time series; see, for example, Cogley and Sargent (2005), Sargent, Williams, and Zha (2006), and Sims and Zha (2006). In VARs nonlinear dynamics are typically generated with time-varying coefficients, whereas most DSGE models are nonlinear and only for convenience approximated by linear rational expectations models. Conceptually the analysis of nonlinear models is very similar to the analysis of linear models, but the implementation of the computations is often more cumbersome and poses a third challenge.

How can Bayesian analysis help?

Bayesian analysis is conceptually straightforward. Pre-sample information about parameters is summarized by a prior distribution $p(\theta)$. We can also assign discrete probabilities to distinct models although the distinction between models and parameters is somewhat artificial. The prior is combined with the conditional distribution of the data given the parameters (likelihood function) $p(Y|\theta)$. The application of Bayes' theorem yields the posterior model probabilities and parameter distributions $p(\theta|Y)$. Markov chain Monte Carlo methods can be used to generate θ draws from the posterior. Based on these draws one can numerically approximate the relevant moments of the posterior and make inference about taste and technology parameters as well as the relative importance and the propagation of the various shocks.

The literature on Bayesian estimation of DSGE models began with work by Landon-Lane (1998), DeJong, Ingram and Whiteman (2000), Schorfheide (2000), and Otrok (2001). DeJong, Ingram and Whiteman (2000) estimate a stochastic growth model and examine its forecasting performance, Otrok (2001) fits a real business cycle with habit formation and time-to-build to the data to assess the welfare costs of business cycles, and Schorfheide (2000) considers cash-in-advance monetary DSGE models. The Bayesian analysis of VAR dates at least back to Doan, Litterman and Sims (1984).

Since DSGE models are to some extent micro-founded, macroeconomists require their parameterization to be consistent with microeconometric evidence on, for instance, labour supply elasticities and the frequency with which firms adjust their prices. If information in the estimation sample were abundant and model misspecification were not a concern, then there would be little need for a prior distribution that summarizes information contained in other data-sets. However, in the estimation of DSGE model this additional information plays an important role.

The prior is used to down-weigh the likelihood function in regions of the parameter space that are inconsistent with out-of-sample information and in which the structural model becomes uninterpretable. The shift from prior to posterior can be an indicator of tensions between different sources of information. If the likelihood function peaks at a value that is at odds with, say, the micro-level information that has been used to construct the prior distribution then marginal data density $\int p(Y|\theta)\,p(\theta)\,d\theta$ will be low. If two models have equal prior probabilities, then the ratio of their marginal data densities determine the posterior model odds. Hence, in a posterior odds comparison a DSGE model

will automatically be penalized for not being able to reconcile two sources of information with a single set of parameters.

Identification problems manifest themselves through ridges and multiple peaks of equal height in the likelihood function. While Bayesian inference is based on the same likelihood function as classical maximum likelihood estimation, it can bring to bear additional information that may help to discriminate between different parameterizations of a model. If, for instance, the likelihood function is invariant to a subvector θ_1 of θ then the posterior distribution of θ_1 conditional on the remaining parameters will simply equal to the prior distribution. Hence, a comparison of priors and posteriors can provide important insights about the extent to which the data provide information about the parameters of interest. Regardless, the posterior provides a coherent summary of pre-sample and sample information and can be used for inference and decision making. This insight has been used, for instance, by Lubik and Schorfheide (2004) to assess whether monetary policy in the 1970s was conducted in a way that would allow expectations to be self-fulfilling and cause business cycle fluctuations unrelated to fundamental shocks.

Bayesian inference is well suited for model comparisons. Under a loss function that is zero if the correct model is chosen and 1 otherwise, it is optimal to select the model that has the highest posterior probability. However, in many applications, in particular related to the comparison of two possibly misspecified DSGE models, this zero–1 loss function is not very attractive because it does provide little insight into the dimensions along which the structural models should be improved. Schorfheide (2000) provides a framework for the comparison of two or more potentially misspecified DSGE models. A VAR plays the role of a reference model. If the DSGE models are indeed misspecified the VAR will attain the highest posterior probability and the model comparison is based on the question: given a particular loss function, which DSGE model best mimics the dynamics captured by the VAR? VARs typically have many more parameters than DSGE models and the role of prior distributions is mainly to reduce the effective dimensionality of this parameter space to avoid over-fitting. More interestingly, if one interprets the DSGE model as a set of restrictions on the VAR, then the DSGE model induces a degenerate prior for the VAR coefficients. If the researcher is concerned about potential misspecification of the DSGE model, a natural approach is to relax the DSGE model restrictions and construct a non-degenerate prior distribution that concentrates most of its mass near the restrictions. This approach was originally proposed by Ingram and Whiteman (1994) and has been further developed by Del Negro and Schorfheide (2004), who provide a framework for the joint estimation of VAR and DSGE model parameters. The framework generates a continuum of intermediate specifications that differ according to the degree by which the restrictions are relaxed. This degree is measured by a hyperparameter and the posterior distribution of the hyperparameter can be interpreted as a measure of fit.

Incorporating model and parameter uncertainty into a decision is straightforward in a Bayesian set-up. Levin et al. (2006), for instance, study the effect of optimal monetary policy under parameter uncertainty in the context of an estimated DSGE model. Let δ denote a decision, such as the choice of a monetary policy rule or a tax rate, and $L(\delta, \theta)$ be a loss function that is used to evaluate the decision. The optimal choice minimizes the posterior risk $\int L(\delta, \theta)\, p(\theta|Y)\, d\theta$. The calculation of the risk is facilitated by Markov chain Monte Carlo methods that enable a numerical evaluation of expected losses. If the parameter θ in the loss function is replaced by a future observation y' and $p(\theta|Y)$ is replaced by the

predictive distribution $p(y' \mid Y)$, the decision-theoretic framework can also be used to generate forecasts from the Bayes model.

Finally, with respect to the analysis of nonlinear models, Bayesian methods are in some instances very helpful. Data-augmentation techniques let researchers efficiently deal with numerical complications that arise in models with latent state variables, such as regime-switching models or VARs with time-varying coefficients as in Cogley and Sargent (2005) and Sims and Zha (2006). On the other hand, the need to compute a likelihood function can create serious obstacles. For instance, the computation of the likelihood function for a DSGE model solved with a nonlinear solution method requires a computational-intensive particle filter as in Fernández-Villaverde and Rubio-Ramírez (2006).

Conclusion

The Bayesian paradigm provides a rich framework for inference and decision making with modern macroeconometric models such as DSGE models and VARs. The econometric methods can be tailored to cope with the challenges in this literature: potential model misspecification and a trade-off between theoretical coherence and empirical fit, identification problems, and estimation of models with many parameters based on relatively few observations. Advances in Bayesian computations let the researcher efficiently deal with numerical complications that arise in models with latent state variables, such as regime-switching models, or nonlinear state-space models.

See Also

- Bayesian econometrics
- Markov chain Monte Carlo methods
- vector autoregressions

Bibliography

An, S. and Schorfheide, F. 2007. Bayesian analysis of DSGE models. *Econometric Reviews*, 26, pp. 113–72.

Beyer, A. and Farmer, R. 2004. On the indeterminacy of New-Keynesian economics. Working Paper No. 323, European Central Bank.

Canova, F. 1994. Statistical inference in calibrated models. *Journal of Applied Econometrics* 9, S123–S44.

Canova, F. and Sala, L. 2006. Back to square one: identification issues in DSGE models. Working Paper No. 583, European Central Bank.

Christiano, L. and Eichenbaum, M. 1999. Monetary policy shocks: what have we learned and to what

end? *Handbook of Macroeconomics*, vol. 1A, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–45.

Cochrane, J. 1994. Shocks. *Carnegie Rochester Conference Series* 41, 295–364.

Cogley, T. and Sargent, T. 2005. Drifts and volatilities: monetary policies and outcomes in the post WWII U.S. *Review of Economic Dynamics* 8, 262–302.

DeJong, D., Ingram, B. and Whiteman, C. 1996. A Bayesian approach to calibration. *Journal of Business Economics and Statistics* 14, 1–9.

DeJong, D., Ingram, B. and Whiteman, C. 2000. A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* 98, 201–23.

Del Negro, M. and Schorfheide, F. 2004. Priors from equilibrium models for VARs. *International Economic Review* 45, 643–73.

Del Negro, M., Schorfheide, F., Smets, F. and Wouters, R. 2006. On the fit of New Keynesian models. *Journal of Business and Economic Statistics* (forthcoming).

Doan, T., Litterman, R. and Sims, C. 1984. Forecasting and conditional projections using realistic prior distributions. *Econometric Reviews* 3, 1–100.

Fair, R. 1994. *Testing Macroeconomic Models*. Cambridge, MA: Harvard University Press.

Fernández-Villaverde, J. and Rubio-Ramírez, J. 2006. Estimating macroeconomic models: a likelihood approach. *Review of Economic Studies* (forthcoming).

Geweke, J. 1999. Computational experiments and reality. *Computing in Economics and Finance* 1999, No. 401. Society for Computational Economics, Department of Economics, Boston College.

Ingram, B. and Whiteman, C. 1994. Supplanting the Minnesota prior – forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics* 34, 497–510.

Levin, A., Onatski, A., Williams, J. and Williams, N. 2006. Monetary policy under uncertainty in micro-founded macroeconometric models. In *NBER Macroeconomics Annual 2005*, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

King, R., Plosser, C. and Rebelo, S. 1988. Production, growth, and business cycles: I. neoclassical model. *Journal of Monetary Economics* 81, 819–40.

Kydland, F. and Prescott, E. 1996. The computational experiment: An econometric tool. *Journal of Economic Perspectives* 10(1), 69–85.

Landon-Lane, J. 1998. Bayesian comparison of dynamic macroeconomic models. Ph.D. thesis, University of Minnesota.

Lubik, T. and Schorfheide, F. 2004. Testing for indeterminacy: an application to U.S. monetary policy. *American Economic Review* 94, 190–217.

Lubik, T. and Schorfheide, F. 2006. A Bayesian look at New Open Economy Macroeconomics. In *NBER Macroeconomics Annual 2005*, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Lucas, R., Jr. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

Otrok, C. 2001. On measuring the welfare cost of business cycles. *Journal of Monetary Economics* 47, 61–92.

Sargent, T., Williams, N. and Zha, T. 2006. Shocks and government beliefs: the rise and fall of American inflation. *American Economic Review* 96, 1193–224.

Schorfheide, F. 2000. Loss function-based evaluation of DSGE models. *Journal of Applied Econometrics* 15, 645–70.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Sims, C. and Zha, T. 2006. Were there regime switches in U.S. monetary policy? *American Economic Review* 96, 54–81.

Smets, F. and Wouters, R. 2003. An estimated stochastic dynamic general equilibrium model of the Euro area. *Journal of the European Economic Association* 1, 1123–75.

Stock, J. and Watson, M. 2001. Vector autoregressions. *Journal of Economic Perspectives* 15(4), 101–16.

How to cite this article

Schorfheide, Frank. "Bayesian methods in macroeconometrics." The New Palgrave Dictionary of

Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000316>
doi:10.1057/9780230226203.0109

The New Palgrave Dictionary of Economics Online

Bayesian nonparametrics

Stephen Graham Walker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses Bayesian nonparametric models, arguing that all Bayesians are constructing probability distributions (the prior) on spaces of density functions. The parametric Bayesian can be seen to be making restrictive assumptions about the choice of density for modelling data. In contrast, the nonparametric Bayesian constructs a probability distribution on as many densities as possible. The model is infinite dimensional, yet inference is possible, including density estimation and the implementation of decision rules, such as the maximization of expected utility. An example of a nonparametric model is given and a means by which to make inference provided by simulation techniques.

Keywords

Bayesian nonparametrics; density functions; expected utility; latent variables; likelihood; Markov chain Monte Carlo methods; parametric models; probability distribution; statistical inference; uncertainty

Article

Bayesian nonparametrics, and more generally the Bayesian approach to statistical inference, finds a theoretical justification via a set of axioms of rational behaviour in the presence of uncertainty. Bayesian decision theory establishes how decisions must be made if one desires to avoid irrational behaviour. Thus, coherence is a fundamental concept and is often used as the main argument against competing statistical approaches, such as those based on sampling or fiducial methods. See Lindley (1978) and Bernardo and Smith (1994, ch. 2), who provide a comprehensive discussion on the axiomatic approach to Bayesian inference.

Bayesian statistics is now commonplace among statistical procedures, and is routinely employed in many areas of science, including economics, medicine, biology and others. The use of a prior distribution is the distinguishing feature; the prior distribution updates to the posterior distribution when

the data are observed. The prior distribution is assumed to represent subjective beliefs about an unknown parameter; the data then provide further information about the parameter, and the revised beliefs are then to be found in the posterior distribution. The updating mechanism from prior to posterior is formalized through the procedure of multiplying the likelihood function by the prior density function. This idea was apparently first written down by Thomas Bayes in the 18th century.

The uncertainty which frustrates the choice of decision is to be assessed via the use of a probability distribution, and the coherent way to make progress with the inclusion of data is via Bayes's theorem. To elaborate, suppose θ is a parameter to be investigated, which if known would provide a decision, and that θ belongs to the parameter space Θ, which is a finite dimensional space. For example, Θ could represent the real line. Data arise from the density $f(x; \theta)$ in the form of independent and identically distributed observations, say $x_1, \ldots, x_n$, that is, a sample of size n. The likelihood function is any function of θ which is proportional to

$$l(\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Let $\pi(\theta)$ denote the prior density function. Then the posterior density function is given by

$$\pi(\theta | x_1, \ldots, x_n) = \frac{l(\theta)\, \pi(\theta)}{\int_{\Theta} l(\theta)\, \pi(d\theta)}.$$

Inference about θ is then performed using the posterior distribution. For example, an estimate of θ could be the posterior mean, which is

$$\hat{\theta} = \int_{\Theta} \theta \pi(d\theta | x_1, \ldots, x_n).$$

Alternatively, interest might be in the estimate of the density function $f(x; \theta)$ itself. In the Bayesian approach this would be provided by the predictive density function, which is given by

$$\hat{f}(x | x_1, \ldots, x_n) = \int_{\Theta} f(x; \theta)\, \pi(d\theta | x_1, \ldots, x_n).$$

Making decisions under uncertainty can be undertaken via the maximization of expected utility approach, see Hirshleifer and Riley (1992), which for the Bayesian would amount to maximizing, over the decision space,

$$\bar{u}(d) = \int_{\Theta} u(d, \theta)\, \pi(\mathrm{d}\theta | x_1, \ldots, x_n),$$

where $u(d, \theta)$ is the utility (reward) of selecting decision *d* from a set of possible decisions when the true parameter state is θ.

The key to the understanding of Bayesian nonparametrics is to think about the family of densities from which the data arose, which in the parametric case is represented as $f(x; \theta)$. Such a family may be known, or assumed to be known, for the data $\{x_1, \ldots, x_n\}$ and the family can be represented by a finite dimensional parameter θ. On the other hand, it may not be known, making assumptions about the family of densities problematic. In this case what is actually unknown is the density function which generated the data: not a parameter, but the entire density function itself. As a Bayesian it is incumbent on the experimenter to construct a prior distribution on the unknown, which is the entire density, and so a probability distribution, the prior, must be placed on the space of density functions. Let such a space be denoted by **F**, so a prior distribution Π must be constructed on **F**.

In fact, any parametric Bayesian model defines a probability measure on **F**. With a parametric model indexed by $\theta \in \Theta$, with family of densities $f(x; \theta)$ and prior $\pi(\theta)$, yields

$$P(f \in A) = \int_{\{\theta: f(\cdot;\theta) \in A\}} \pi(\mathrm{d}\theta),$$

for suitable sets of densities $A \subset \mathbf{F}$. If we let $\Pi(A) = P(f \in A)$, then Π is the prior distribution on **F**, and the pair $\{f(x; \theta), \pi(\theta)\}$ are a useful way to construct a probability on **F**. However, this approach of using the parametric model restricts the choice, the *A*'s for which $\Pi(A)>0$ form a very small set, and so, while it can be seen that all Bayesians are constructing probability measures on **F**, it is the parametric Bayesian who is making restrictive assumptions.

A consequence of the restrictive choice can be seen by considering $\Omega \subset \mathbf{F}$, which we define to be the smallest set of densities which are allocated probability 1, that is, $\Pi(\Omega)=1$. A parametric family is typically checked off with the data once it has been observed, to see if the model and the data are compatible. Yet this practice is clearly in contradiction (that is, incoherent) with the allocation of probability 1. See Lindsey (1999) for more on this aspect of Bayesian inference. It is the responsibility of the Bayesian to select Ω large enough to make any such checks redundant. This may mean having Ω to be the set of all densities, or at least having the set of *A*s' for which $\Pi(A)>0$ to be as large as can be

achieved.

In the Bayesian nonparametric approach, the prior distribution is placed on **F** directly and there is no finite-dimensional parameter characterizing the random density functions chosen from the prior. The model is infinite-dimensional. The prior is now written as $\Pi(\mathrm{d}f)$ to reflect the fact that there is no parametric θ generating the density *f*. The likelihood function simply becomes

$$l(f) = \prod_{i=1}^{n} f(x_i)$$

and so the posterior is given by

$$\Pi(\mathrm{d}f \mid x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} f(x_i)\Pi(\mathrm{d}f)}{\int_{\mathbf{F}} \prod_{i=1}^{n} f(x_i)\Pi(\mathrm{d}f)}.$$

Now, for example, the estimate of the density generating the data can be the predictive density, which is

$$\hat{f}(x \mid x_1, \ldots, x_n) = \int_{\mathbf{F}} f(x)\Pi(\mathrm{d}f \mid x_1, \ldots, x_n).$$

For decision theory, if $u(d, f)$ is the utility of decision *d* when *f* is the true density, that is, the true density function generating observations, then the maximization of the expected utility rule yields the decision *d* maximizing

$$\bar{u}(d) = \int_{\mathbf{F}} u(d, f)\Pi(\mathrm{d}f \mid x_1, \ldots, x_n).$$

So what has happened is that we have replaced the finite-dimensional θ with the infinite-dimensional *f*. Obviously, the important feature in Bayesian nonparametrics is to be able construct a probability distribution Π on **F** such that Ω is large. Suppose **F** is the space of density functions defined on the real line. Then, for example, we could choose Π by restricting attention to the normal family of density functions. That is, a random normal density function chosen from Π has the mean μ chosen from the

probability density $\pi(\mu)$ and the variance σ^2 chosen from the probability density $\pi(\sigma^2)$. However, the shape of densities constructed this way is restricted to the normal shape and Ω will not be large. To generate more shapes of density function, one needs to increase the number of parameters from two to a large number, even an infinite, but countable, number. This can be achieved by a mixture model, taking the normal distribution and mixing it over the parameters by using a random distribution function. If we let $\theta = (\mu, \sigma^2)$ and let N denote the normal density function, then a random density function can be obtained via

$$f_P(x) = \int_{\Theta} N(x|\theta)\, dP(\theta),$$

where P is a random distribution function defined on $(-\infty,+\infty)\times(0,+\infty)$. The variety of shapes for f_P as P varies over distribution functions is enlarged significantly.

The choice for the random distribution function P needs to be discussed. A common choice is the Dirichlet process model, introduced by Ferguson (1973). The model generates random distribution functions which are discrete. Essentially, a random path (stochastic process) is generated which behaves as a distribution function. That is, it starts at zero and moves to 1 in a non-decreasing way. It is possible to sample a Dirichlet process via the strategy of taking $\{\theta_i\}_{i=1}^{\infty}$ to be independent and identically distributed from some fixed distribution P_0 and $\{v_i\}_{i=1}^{\infty}$ to be independent and identically distributed from beta $(1,c)$ for some $c>0$. Then

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j},$$

where $w_1=v_1$ and for $j>1$,

$$w_j = v_j \prod_{l=1}^{j-1} (1 - v_l)\,.$$

It is straightforward to show that the sum of the w_j's is one. It is that $E(P)=P_0$ and for suitable sets B,

$$\mathrm{Var}\left\{P(B)\right\} = \frac{P_0(B)\{1 - P_0(B)\}}{c + 1}.$$

Using the Dirichlet process itself for modelling independent and identically distributed observations, say $\{y_1,\ldots,y_n\}$, can be done and the posterior is also a Dirichlet process with updated parameters $c \rightarrow c+n$ and

$$P_0 \rightarrow \frac{cP_0 + nP_n}{c + n},$$

where P_n is the empirical distribution function of $\{y_1, \ldots, y_n\}$. Hence, the Bayes estimate is a nice mixture of the prior choice and the empirical distribution.
However, the Dirichlet process is better placed to construct random density functions via mixtures, and we can write the random density function based on the mixture as

$$f_{w,\theta}(x) = \sum_{j=1}^{\infty} w_j N(x|\theta_j).$$

This is an infinite-dimensional model and is known as the mixture of Dirichlet process model. It was first studied by Lo (1984) and can really be estimated only by using recent advances in posterior simulation techniques based on Gibbs samplers and more generally Markov chain Monte Carlo methods (Smith and Roberts, 1993; Tierney, 1994). The original simulation technique was introduced by Escobar (1988), and since then a number of algorithms have been described. A nice approach, as is becoming usual with Bayesian nonparametric models, is to use latent variables. A slice variable can work well with the mixture of Dirichlet process model by introducing the latent variable u, which has joint density with x given by

$$f_{w,\theta}(x, u) = \sum_{j=1}^{\infty} 1(u < w_j) N(x|\theta_j).$$

Integrating over u returns us to the original model, and the usefulness of the latent variable is apparent in

that it makes the infinite sum finite. That is, there is only a finite number of the $\{w_j\}$ which are greater then u, for each $u>0$. A Gibbs sampler can now be employed on the model exactly. Typically one is interested in prediction, and at each iteration of the Markov chain it is possible to sample from the predictive density.

There is nowadays a wide range of Bayesian nonparametric models from which to select for any kind of statistical context. See, for example, Walker et al. (1999) for details. Analysis, in the way of inference or decision making, is then typically undertaken using simulation techniques such as Markov chain Monte Carlo methods.

Most Bayesian nonparametric priors are based on stochastic processes. The probability measure for the process acts as the prior distribution. One such example employed in survival models is based on independent increment processes; one has

$$S(t) = e^{-Z(t)},$$

where, with probability 1, Z is a non-decreasing process with $Z(0)=0$ and $\lim_{t\to+\infty} Z(t) = +\infty$. Here S is a random survival distribution, the law governing the path is the prior. The posterior is also based on an independent increment process (conjugate), and a limiting version of the Bayes estimate turns out to be the Kaplan–Meier nonparametric estimator for a survival function.

Bayesian nonparametric models support more outcomes than parametric models. Prior distributions are constructed on function spaces, such as density functions, survival distribution functions or even hazard functions. The prior distributions are the laws governing stochastic processes whose sample paths behave like these types of functions. Inference is typically reliant on Markov chain Monte Carlo methods, often following the introduction of latent variables.

See Also

- Bayesian statistics
- decision theory in econometrics
- utility

Bibliography

Bernardo, J.M. and Smith, A.F.M. 1994. *Bayesian Theory*. London: Wiley.

Escobar, M.D. 1988. Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Ph.D. thesis, Department of Statistics, Yale University.

Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1, 209–

30.

Hirshleifer, J. and Riley, J.G. 1992. *The Analysis of Uncertainty and Information*. Cambridge: Cambridge University Press.

Lindley, D.V. 1978. The Bayesian approach (with discussion). *Scandinavian Journal of Statistics* 5, 1–26.

Lindsey, J.K. 1999. Some statistical heresies. *The Statistician* 48, 1–40.

Lo, A.Y. 1984. On a class of Bayesian nonparametric estimates I. Density estimates. *Annals of Statistics* 12, 351–7.

Smith, A.F.M. and Roberts, G.O. 1993. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 55, 3–23.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *Annals of Statistics* 22, 1701–22.

Walker, S.G., Damien, P., Laud, P.W. and Smith, A.F.M. 1999. Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society*, Series B 61, 485–527.

How to cite this article

Walker, Stephen Graham. "Bayesian nonparametrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000321> doi:10.1057/9780230226203.0110

The New Palgrave Dictionary of Economics Online

Bayesian statistics

José M. Bernardo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Statistics is primarily concerned with analysing data, either to assist in appreciating some underlying mechanism or to reach effective decisions. All uncertainties should be described by probabilities, since probability is the only appropriate language for a logic that deals with all degrees of uncertainty, not just absolute truth and falsity. This is the essence of Bayesian statistics. Decision-making is embraced by introducing a utility function and then maximizing expected utility. Bayesian statistics is designed to handle all situations where uncertainty is found. Since some uncertainty is present in most aspects of life, Bayesian statistics arguably should be universally appreciated and used.

Keywords

asymptotic behaviour; Bayes, T.; Bayesian reference criterion; Bayesian statistics; exchangeability; expected utility; hypothesis testing; improper prior function; inference; likelihood; nonparametric models; nuisance parameters; point estimation; prediction; probability; reference analysis; region estimation; representation theorems; robustness; statistical decision theory; statistical inference; subjective probability; sufficiency; sure thing principle; uncertainty

Article

Bayesian statistics is a comprehensive approach to both statistical inference and decision analysis which derives from the fact that, for rational behaviour, all uncertainties in a problem must necessarily be described by probability distributions.

Unlike most other branches of mathematics, conventional methods of statistical inference do not have an axiomatic basis; as a consequence, their proposed desiderata are often mutually incompatible, and the analysis of the same data may well lead to incompatible results when different, apparently intuitive, procedures are tried. In marked contrast, the Bayesian approach to statistical inference is firmly based on axiomatic foundations which provide a unifying logical structure and guarantee the mutual consistency of the methods proposed. Bayesian methods constitute a complete paradigm for statistical inference, a scientific revolution in Kuhn's sense. Bayesian statistics require only the mathematics of probability theory and the interpretation of probability which most closely corresponds to the standard use of this word in everyday language: a conditional measure of uncertainty.

The main consequence of these axiomatic foundations is precisely the requirement to describe with probability distributions all uncertainties present in the problem. Hence, parameters are treated as random variables; this is not a description of their variability (parameters are typically fixed unknown quantities) but a description of the uncertainty about their true values.

The Bayesian paradigm is easily summarized. Thus, if available data *D* are assumed to have been generated from a probability distribution $p(D|\boldsymbol{\omega})$ characterized by an unknown parameter vector $\boldsymbol{\omega}$, the uncertainty about the value of $\boldsymbol{\omega}$ before the data have been observed must be described by a prior probability distribution $p(\boldsymbol{\omega})$. After data *D* have been observed, the uncertainty about the value of $\boldsymbol{\omega}$ is described by its posterior distribution $p(\boldsymbol{\omega}|D)$, which is obtained via Bayes's theorem; hence the adjective Bayesian for this form of inference. Point and region estimates for $\boldsymbol{\omega}$ may be derived from $p(\boldsymbol{\omega}|D)$ as useful summaries of its contents. Measures of the compatibility of the posterior with a particular set Θ_0 of parameter values may be used to test the hypothesis $H_0 \equiv \{q \in \Theta_0\}$. If data consist of a random sample $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from a probability distribution $p(\boldsymbol{x}|\boldsymbol{\omega})$, inferences about the value of a future observation $\boldsymbol{x}$ from the same process are derived from the (posterior) predictive distribution $p(\boldsymbol{x}|D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x}|\boldsymbol{\omega})\, p(\boldsymbol{\omega}|D)\, d\boldsymbol{\omega}$.

An important particular case arises when either no relevant prior information is readily available, or that information is subjective and an ‘objective’ analysis is desired, one that is

exclusively based on accepted model assumptions and well-documented data. This is addressed by reference analysis which uses information-theoretic concepts to derive the appropriate reference posterior distribution $\pi(\boldsymbol{\omega}|D)$, defined to encapsulate inferential conclusions about the value of $\boldsymbol{\omega}$ solely based on the assumed probability model $p(D|\boldsymbol{\omega})$ and the observed data *D*.

Pioneering textbooks on Bayesian statistics were Jeffreys (1961), Lindley (1965), Zellner (1971) and Box and Tiao (1973). For modern elementary introductions, see Berry (1996) and Lee (2004). Intermediate to advanced monographs on Bayesian statistics include Berger (1985), Bernardo and Smith (1994), Gelman et al. (2003), O'Hagan (2004) and Robert (2001). This article may be regarded as a very short summary of the material contained in the forthcoming second edition of Bernardo and Smith (1994). For a recent review of objective Bayesian statistics, see Bernardo (2005) and references therein.

Foundations

The central element of the Bayesian paradigm is the use of probabilities to describe all relevant uncertainties, interpreting $\Pr(A|H)$, the probability of *A* given *H*, as a conditional measure of uncertainty, on a [0,1] scale, about the occurrence of the event *A* in conditions *H*. There are two different independent arguments which prove the mathematical inevitability of the use of probabilities to describe uncertainties.

Exchangeability and representation theorems

Available data often consist of a finite set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of 'homogeneous' observations, in the sense that only their values matter, not the order in which they appear. Formally, this is captured by the notion of exchangeability. The set of random vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_j \in \mathcal{X}$, is exchangeable if their joint distribution is invariant under permutations. An infinite sequence of random vectors is exchangeable if all its finite subsequences are exchangeable. Notice that, in particular, any random sample from any model is exchangeable. The general representation theorem implies that, if a set of observations is assumed to be a subset of an exchangeable sequence, then it constitutes a random sample from a probability model $\{p(\mathbf{x}|\boldsymbol{\omega}),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, described in terms of some parameter vector $\boldsymbol{\omega}$; furthermore, this parameter $\boldsymbol{\omega}$ is defined as the limit (as $n \to \infty$) of some function of the observations, and available information about the value of $\boldsymbol{\omega}$ must necessarily be described by some probability distribution $p(\boldsymbol{\omega})$. This formulation includes 'nonparametric' (distribution free) modelling, where $\boldsymbol{\omega}$ may index, for instance, all continuous probability distributions on $\mathcal{X}$. Notice that $p(\boldsymbol{\omega})$ does not model a possible variability of $\boldsymbol{\omega}$ (since $\boldsymbol{\omega}$ will typically be a fixed unknown vector), but models the uncertainty associated with its actual value. Under exchangeability (and therefore under any assumption of random sampling), the general representation theorem provides an existence theorem for a probability distribution $p(\boldsymbol{\omega})$ on the parameter space $\boldsymbol{\Omega}$, and this is an argument which depends only on mathematical probability theory.

Statistical inference and decision theory

Statistical decision theory provides a precise methodology to deal with decision problems under uncertainty, but it also provides a powerful axiomatic basis for the Bayesian approach to statistical inference. A decision problem exists whenever there are two or more possible courses of action. Let $\mathcal{A}$ be the class of possible actions, let $\boldsymbol{\Theta}$ be the set of relevant events which may affect the result of choosing an action, and let $c(a, q) \in \mathcal{C}$, be the consequence of having chosen action *a* when event θ takes place. The triplet $\{\mathcal{A}, \boldsymbol{\Theta}, \mathcal{C}\}$ describes the structure of the decision problem. Different sets of principles have been proposed to capture a minimum collection of logical rules that could sensibly be required for rational decision-making. These all consist of axioms with a strong intuitive appeal; examples include the *transitivity* of preferences (if $a_1 \succ a_2$ and $a_2 \succ a_3$, then $a_1 \succ a_3$), and the *sure thing principle* (if $a_1 \succ a_2$ given *E*, and $a_1 \succ a_2$ given $\bar{E}$, then $a_1 \succ a_2$). Notice that these rules are not intended as a description of actual human decision-making, but as a normative set of principles to be followed by someone who aspires to achieve coherent decision-making. There are naturally different options for the set of acceptable principles, but they all lead to the same basic conclusions:

- Preferences among possible consequences should be measured with a *utility* function $u(c) = u(a, q)$ which specifies, on some numerical scale, their desirability.
- The uncertainty about the relevant events should be measured with a probability distribution $p(q|D)$ describing their plausibility given the conditions under which the decision must be taken (assumptions made and available data *D*).
- The best strategy is to take that action a^* with maximizes the corresponding expected utility, $\int_{\boldsymbol{\Theta}} u(a, q)\ p(q|D)\, dq$.

Notice that the argument described above establishes (from another perspective) the need to quantify the uncertainty about all relevant unknown quantities (the actual value of the

vector θ), and specifies that this must have the mathematical structure of a probability distribution. It has been argued that the development described above (which is not qsted when decisions have to be made) does not apply to problems of statistical inference, where no specific decision making is envisaged. Notice, however, that (*a*) a problem of statistical inference is typically considered worth analysing because it may eventually help make sensible decisions (as Ramsey put it in the 1930s, a lump of arsenic is poisonous because it may kill someone, not because it has actually killed someone), and (*b*) statistical inference on θ has the mathematical structure of a decision problem, where the class of alternatives is the functional space of all possible conditional probability distributions of θ given the data, and the utility function is a measure of the amount of information about θ which the data may be expected to provide.

In statistical inference it is often convenient to work in terms of the non-negative loss function $\ell(a, q) = \sup_{a \in \mathcal{A}}\{u(a, q)\} - u(a, q)$, which directly measures, as a function of θ, the *penalty* for choosing a wrong action. The undesirability of each possible action $a \in \mathcal{A}$ is then measured by its *expected loss*, $l(a|D) = \int_{\Theta} \ell(a, q)\, p(q|D)\, dq$, and the best action *a** is that with the minimum expected loss.

The Bayesian paradigm

The statistical analysis of some observed data set $D \in \mathcal{D}$ typically begins with some informal descriptive evaluation, which is used to suggest a tentative, formal probability model $\{p(D|\boldsymbol{\omega}, H),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ which, given some assumptions *H*, is supposed to represent, for some (unknown) value of $\boldsymbol{\omega}$, the probabilistic mechanism which has generated the observed data *D*. The arguments outlined above establish the logical need to assess a prior probability distribution $p(\boldsymbol{\omega}|H)$ over the parameter space $\boldsymbol{\Omega}$, describing the available knowledge about the value of $\boldsymbol{\omega}$ under the accepted assumptions *H*, prior to the data being observed. It then follows from Bayes's theorem that, if the probability model is correct, all available information about the value of $\boldsymbol{\omega}$ after the data *D* have been observed is contained in the corresponding *posterior distribution*,

$$p(\boldsymbol{\omega}|D, H) = \frac{p(D|\boldsymbol{\omega}, H)\, p(\boldsymbol{\omega}|H)}{\int_{\boldsymbol{\Omega}} p(D|\boldsymbol{\omega}, H)\, p(\boldsymbol{\omega}|H)\, d\boldsymbol{\omega}}, \quad \boldsymbol{\omega} \in \boldsymbol{\Omega} \tag{1}$$

It is this systematic use of Bayes's theorem to incorporate the information provided by the data that justifies the adjective ‘Bayesian’ by which the paradigm is usually known. It is obvious from Bayes's theorem that any value of $\boldsymbol{\omega}$ with zero prior density will have zero posterior density. Thus, it is typically assumed (by appropriate restriction, if necessary, of the parameter space $\boldsymbol{\Omega}$) that prior distributions are strictly positive. To simplify the presentation, the assumptions *H* are often omitted from the notation, but the fact that all statements about $\boldsymbol{\omega}$ given *D* are also conditional to *H* should always be kept in mind.

Computation of posterior densities is often facilitated by noting that Bayes's theorem may be simply expressed as $p(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\, p(\boldsymbol{\omega})$ (where $\propto$ stands for ‘proportional to’ and where, for simplicity, the assumptions *H* have been omitted from the notation), since the missing proportionality constant $[\int_{\boldsymbol{\Omega}} p(D|\boldsymbol{\omega})\, p(\boldsymbol{\omega})\, d\boldsymbol{\omega}]^{-1}$ may always be deduced from the fact that $p(\boldsymbol{\omega}|D)$, a probability density, must integrate to 1.

Improper priors

An improper prior function is defined as non-negative function $\pi(\boldsymbol{\omega})$ such that $\int_{\boldsymbol{\Omega}} \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ is not finite. The formal expression of Bayes's theorem remains, however technically valid if $p(\boldsymbol{\omega})$ is replaced by an improper prior function $\pi(\boldsymbol{\omega})$, provided the proportionality constant exists, thus leading to a well-defined *proper* posterior density $\pi(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\, \pi(\boldsymbol{\omega})$, which does integrate to 1.

Likelihood principle

Considered as a function of $\boldsymbol{\omega}$ for fixed data *D*, $p(D|\boldsymbol{\omega})$ is often referred to as the likelihood function. Thus, Bayes's theorem is simply expressed in words by the statement that the posterior is proportional to the likelihood times the prior. It follows from (1) that, provided the same prior $p(\boldsymbol{\omega})$ is used, two different data sets D_1 and D_2, with possibly different probability models $p_1(D_1|\boldsymbol{\omega})$ and $p_2(D_2|\boldsymbol{\omega})$ which yield proportional likelihood functions, will produce identical posterior distributions for $\boldsymbol{\omega}$. This immediate consequence of Bayes's theorem has been proposed as a principle on its own, the *likelihood principle*, and it is seen by many as an obvious requirement for reasonable statistical inference. In

particular, for any given prior $p(\boldsymbol{\omega})$, the posterior distribution does not depend on the set $\mathcal{D}$ of possible data values (the *outcome space*). Notice, however, that the likelihood principle applies only to inferences about the parameter vector $\boldsymbol{\omega}$ once the data have been obtained. Consideration of the outcome space is essential, for instance, in model criticism, in the design of experiments, in the derivation of predictive distributions, and in the construction of objective Bayesian procedures.

Sequential learning

Naturally, the terms ‘prior’ and ‘posterior’ are only relative to a particular set of data. As one would expect, if exchangeable data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are sequentially presented, the final result will be the same whether data are globally or sequentially processed. Indeed, $p(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1}|\boldsymbol{\omega})\, p(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$, for $i = 1, \ldots, n-1$, so that the ‘posterior’ at a given stage becomes the ‘prior’ at the next.

Sufficiency

For a given probability model, one may find that some particular function of the data $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathcal{T}$ is a sufficient statistic in the sense that, given the model, $\boldsymbol{t}(D)$ contains all information about $\boldsymbol{\omega}$ which is available in D. Formally, $\boldsymbol{t}$ is sufficient if (and only if) there exist non-negative functions f and g such that the likelihood function may be factorized in the form $p(D|\boldsymbol{\omega}) = f(\boldsymbol{\omega}, \boldsymbol{t})\, g(D)$. A sufficient statistic always exists, for $\boldsymbol{t}(D) = D$ is obviously sufficient; however, a much simpler sufficient statistic, with a fixed dimensionality which is independent of the sample size, often exists. In fact this is known to be the case whenever the probability model belongs to the *generalized exponential family*, which includes many of the more frequently used probability models. It is easily established that if $\boldsymbol{t}$ is sufficient, then the posterior distribution of $\boldsymbol{\omega}$ depends only on the data D through $\boldsymbol{t}(D)$, and $p(\boldsymbol{\omega}|D) = p(\boldsymbol{\omega}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\omega})\, p(\boldsymbol{\omega})$.

Robustness

As one would expect, for fixed data and model assumptions, different priors generally lead to different posteriors. Indeed, Bayes' theorem may be described as a data-driven probability transformation machine which maps prior distributions (describing prior knowledge) into posterior distributions (representing combined prior and data knowledge). It is important to analyse the robustness of the posterior to changes in the prior. Objective posterior distributions based on reference priors (see below) play a central role in this context. Investigation of the sensitivity of the posterior to changes in the prior is an important ingredient of the comprehensive analysis of the sensitivity of the final results to all accepted assumptions, which any responsible statistical study should contain.

Nuisance parameters

Typically, the quantity of interest is not the whole parameter vector $\boldsymbol{\omega}$, but some function $q = q(\boldsymbol{\omega})$ of possibly lower dimension than $\boldsymbol{\omega}$. Any valid conclusion on the value of θ will be contained in its posterior probability distribution $p(q|D)$, which may be derived from $p(\boldsymbol{\omega}|D)$ by standard use of probability calculus. Indeed, if $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\omega}) \in \Lambda$ is some other function of $\boldsymbol{\omega}$ such that $\boldsymbol{\psi} = \{\theta, \boldsymbol{\lambda}\}$ is a one-to-one transformation of $\boldsymbol{\omega}$, and $J(\boldsymbol{\omega}) = (\partial \boldsymbol{\psi} / \partial \boldsymbol{\omega})$ is the corresponding Jacobian matrix, one may change variables to obtain $p(\boldsymbol{\psi}|D) = p(q, \boldsymbol{\lambda}|D) = p(\boldsymbol{\omega}|D) / |J(\boldsymbol{\omega})|$, and the required posterior of θ is $p(q|D) = \int_{\Lambda} p(q, \boldsymbol{\lambda}|D)\, d\boldsymbol{\lambda}$, the marginal density obtained by integrating out the nuisance parameter $\boldsymbol{\lambda}$. Naturally, introduction of $\boldsymbol{\lambda}$ is not necessary if $\theta(\omega)$ is a one-to-one transformation of $\boldsymbol{\omega}$. Notice that elimination of unwanted nuisance parameters, a simple integration within the Bayesian paradigm, is a difficult (often polemic) problem for conventional statistics.

Restricted parameter space

Sometimes, the range of possible values of $\boldsymbol{\omega}$ is effectively restricted by contextual considerations. If $\boldsymbol{\omega}$ is known to belong to $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$, the prior distribution is positive only in $\boldsymbol{\Omega}_c$ and, if one uses Bayes's theorem, it is immediately found that the restricted posterior is

$$p(\boldsymbol{\omega}|D, \boldsymbol{\omega} \in \boldsymbol{\Omega}_c) = p(\boldsymbol{\omega}|D) \Big/ \int_{\boldsymbol{\Omega}_c} p(\boldsymbol{\omega}|D)\, d\boldsymbol{\omega},$$

for $\boldsymbol{\omega} \in \boldsymbol{\Omega}_c$ (and obviously vanishes if $\boldsymbol{\omega} \notin \boldsymbol{\Omega}_c$). Thus, to incorporate a restriction on the possible values of the parameters, it suffices to renormalize the unrestricted posterior distribution to the set $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$ of parameter values which satisfy the required condition. Incorporation of known constraints on the parameter values, a simple renormalization within the Bayesian paradigm, is another very difficult problem for conventional statistics.

Asymptotic behaviour

The behaviour of posterior distributions when the sample size is large is important, for at least two different reasons: (*a*) asymptotic results provide useful first-order approximations when actual samples are relatively large, and (*b*) objective Bayesian methods typically depend on the asymptotic properties of the assumed model. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \ \boldsymbol{x}_j \in \mathcal{X}$, be a random sample of size *n* from $\{p(\boldsymbol{x}|\boldsymbol{\omega}), \ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. It may be shown that, as $n \to \infty$, the posterior distribution $p(\boldsymbol{\omega}|D)$ of a discrete parameter $\boldsymbol{\omega}$ typically converges to a degenerate distribution which gives probability one to the true value of $\boldsymbol{\omega}$, and that the posterior distribution of a continuous parameter $\boldsymbol{\omega}$ typically converges to a normal distribution centred at its maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\omega}}$, with a covariance matrix $F^{-1}(\widehat{\boldsymbol{\omega}}) / n$, where $F(\boldsymbol{\omega})$ is Fisher information matrix, of general element

$$F_{ij}(\boldsymbol{\omega}) = -E_{\boldsymbol{x}|\boldsymbol{\omega}}[-\partial^2 \log[p(\boldsymbol{x}|\boldsymbol{\omega})] / (\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_j)].$$

Prediction

When data consist of a set $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of homogeneous observations, one is often interested in predicting the value of a future observation $\boldsymbol{x}$ generated by the same random mechanism that has generated the observations in *D*. It follows from the foundations arguments discussed above that the solution to this prediction problem must be a probability distribution $p(\boldsymbol{x}|D)$ which describes the uncertainty about the value that $\boldsymbol{x}$ will take, given the information provided by *D*, and any other available knowledge. In particular, if contextual information suggests that data *D* may be considered to be a random sample from a distribution in the family $\{p(\boldsymbol{x}|\boldsymbol{\omega}), \ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and $p(\boldsymbol{\omega})$ is a probability distribution which encapsulates all available prior information on the value of $\boldsymbol{\omega}$, the corresponding posterior will be (by Bayes's theorem) $p(\boldsymbol{\omega}|D) \propto \prod_{j=1}^{n} p(\boldsymbol{x}_j|\boldsymbol{\omega}) \ p(\boldsymbol{\omega})$. Since $p(\boldsymbol{x}|\boldsymbol{\omega}, D) = p(\boldsymbol{x}|\boldsymbol{\omega})$, the total probability theorem may then be used to obtain the desired posterior *predictive* distribution

$$p(\boldsymbol{x}|D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x}|\boldsymbol{\omega}) \, p(\boldsymbol{\omega}|D) \, d\boldsymbol{\omega} \tag{2}$$

which has the form of a *weighted average*: the average of all possible probability distributions of $\boldsymbol{x}$, weighted with their corresponding posterior densities. Notice that the conventional practice of plugging in some point estimate $\tilde{\boldsymbol{\omega}} = \tilde{\boldsymbol{\omega}}(D)$ and using $p(\boldsymbol{x}|\tilde{\boldsymbol{\omega}})$ to predict $\boldsymbol{x}$ may be seriously misleading, for this totally ignores the uncertainty about the true value of $\boldsymbol{\omega}$. If the assumptions on the probability model are correct, the posterior predictive distribution $p(\boldsymbol{x}|D)$ will converge, as the sample size increases, to the distribution $p(\boldsymbol{x}|\boldsymbol{\omega})$ which has generated the data. Indeed, a good technique to assess the quality of the inferences about $\boldsymbol{\omega}$ encapsulated in $p(\boldsymbol{\omega}|D)$ is to check against the observed data the predictive distribution $p(\boldsymbol{x}|D)$ generated from $p(\boldsymbol{\omega}|D)$. The argument used to derive $p(\boldsymbol{x}|D)$ may be extended to obtain the predictive distribution of any function $\boldsymbol{y}$ of future observations generated by the same process, namely, $p(\boldsymbol{y}|D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{y}|\boldsymbol{\omega}) \ p(\boldsymbol{\omega}|D)$.

Reference analysis

The posterior distribution combines the information provided by the data with relevant available prior information. In many situations, however, either the available prior information

is too vague to warrant the effort required to have it formalized in the form of a probability distribution, or it is too subjective to be useful in scientific communication or public decision making. It is therefore important to identify the mathematical form of a reference prior, a prior that would have a minimal effect, relative to the data, on the posterior inference. Much work has been done to formulate priors which would make this idea mathematically precise. This section summarizes an approach, based on information theory, which may be argued to provide the most advanced general procedure available. In this formulation, the reference prior is that which maximizes the missing information about the quantity of interest.

Reference distributions

Consider data *D*, generated by a random mechanism $p(D|\theta)$ which depends only on a real-valued parameter $\theta \in \Theta \subset \Re$, and let $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathcal{T}$ be any sufficient statistic (which may well be the complete data set *D*). In Shannon's general information theory, the amount of information $I\{\mathcal{T}, p(\theta)\}$ which may be expected to be provided by *D*, about the value of θ is

$$I\left\{\mathcal{T},\ p(\theta)\right\} = \int_{\mathcal{T}} \int_{\Theta} p(\boldsymbol{t},\ \theta) \log \frac{p(\boldsymbol{t},\ \theta)}{p(\boldsymbol{t})\, p(\theta)}\, d\theta\, d\boldsymbol{t} = E_t\left[\int_{\Theta} p(\theta|\boldsymbol{t}) \log \frac{p(\theta|\boldsymbol{t})}{p(\theta)}\, d\theta\right] \tag{3}$$

the expected logarithmic divergence of the prior from the posterior. This is a *functional* of the prior distribution *p*(θ): the larger the prior information, the smaller the information which the data may be expected to provide. The functional $I\{\mathcal{T},\ p(\theta)\}$ is concave, non-negative, and invariant under one-to-one transformations of θ. Consider now the amount of information $I\left\{\mathcal{T}^k,\ p(\theta)\right\}$ about θ which may be expected from the experiment which consists of *k* conditionally independent replications $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k\}$ of the original experiment. As $k \to \infty$, such an experiment would provide any *missing information* about θ which could possibly be obtained within this framework; thus, as $k \to \infty$, the functional $I\left\{\mathcal{T}^k,\ p(\theta)\right\}$ will approach the *missing information* about θ associated with the prior $p(\theta)$. Intuitively, the reference prior for θ is that which maximizes the missing information about θ. If $\pi_k(\theta|\mathcal{P})$ denotes the prior density which maximizes $I\left\{\mathcal{T}^k,\ p(\theta)\right\}$ in the class $\mathcal{P}$ of strictly positive prior distributions which are compatible with accepted assumptions on the value of θ (which may well be the class of all strictly positive proper priors), then the θ-reference prior $\pi(\theta|\mathcal{P})$ is the limit of the sequence of priors $\{\pi_k(\theta|\mathcal{P})\}_{k=1}^{\infty}$. The limit is taken in the precise sense that, for any value of the sufficient statistic $\boldsymbol{t}$, the reference posterior, the pointwise limit $\pi(\theta|\boldsymbol{t}, \mathcal{P})$ of the corresponding sequence of posteriors $\{\pi_k(\theta|\boldsymbol{t}, \mathcal{P})\}_{k=1}^{\infty}$, where $\pi_k(\theta|\boldsymbol{t}, \mathcal{P}) \propto p(\boldsymbol{t}|\theta)\, \pi_k(\theta|\mathcal{P})$, may be obtained from $\pi(\theta|\mathcal{P})$ by formal use of Bayes' theorem, so that $\pi(\theta|\boldsymbol{t}, \mathcal{P}) \propto p(\boldsymbol{t}|\theta)\, \pi(\theta|\mathcal{P})$.

The limiting procedure in the definition of a reference prior is not some kind of asymptotic approximation, but an essential element of the definition, required to capture the basic concept of missing information. Notice that, by definition, reference distributions depend only on the asymptotic behaviour of the assumed probability model, a feature which greatly simplifies their actual derivation.

Reference prior *functions* are often simply called reference priors, even though they are usually improper. They should not be considered as expressions of belief, but technical devices to obtain (proper) posterior distributions, which are a limiting form of the posteriors that would have been obtained from prior beliefs which, when compared with the information which data could provide, are relatively uninformative with respect to the quantity of interest.

If θ may take only a *finite* number *m* of different values, the missing information about θ associated to the prior *p*(θ) is its entropy, $H\left\{p(\theta)\right\} = -\sum_{j=1}^{m} p(\theta_j) \log p(\theta_j)$. Hence the reference prior $\pi(\theta|\mathcal{P})$ is in this case is the prior with *maximum entropy* within $\mathcal{P}$. In particular, if $\mathcal{P}$ contains all priors over $\{\theta_1, \ldots, \theta_m\}$, then the reference prior when θ is the quantity of interest is the uniform prior $\pi(\theta) = \{1/m, \ldots, 1/m\}$.

If the sufficient statistic $\boldsymbol{t}$ is a consistent, asymptotically sufficient estimator $\ddot{\theta}$ of a continuous parameter θ, and the class of priors is the set $\mathcal{P}_0$ of all strictly positive priors, then the reference prior is simply

$$\pi(\theta|\mathcal{P}_0) \propto p(\theta|\ddot{\theta})|_{\tilde{\theta}=\theta} \propto p(\theta|\ddot{\theta})|_{\tilde{\theta}=\theta}, \tag{4}$$

where $p(\ddot{\theta}|\theta)$ is any asymptotic approximation to the posterior distribution of θ, and $p(\ddot{\theta}|\theta)$ is the sampling distribution of $\ddot{\theta}$. Under conditions which guarantee asymptotic posterior normality, this reduces to *Jeffreys prior*, $\pi(\theta_0|p) \propto F(\theta)^{1/2}$, where $F(\theta)$ is Fisher information function. One-parameter reference priors are consistent under re-parametrization; thus, if $\psi = \psi(\theta)$ is a piecewise one-to-one function of θ, then the ψ-reference prior is simply the appropriate probability transformation of the θ-reference prior.

Example 1.: **Exponential data**. If $\mathbf{x} = \{x_1, \ldots, x_n\}$ is a random sample from $\theta e^{-\theta x}$, the reference prior is Jeffreys prior $\pi(\theta) = \theta^{-1}$, and the reference posterior is a gamma distribution $\pi(\theta|x) = Ga(\theta|n, t)$, where $t = \sum_{j=1}^{n} x_j$. With a random sample of size $n = 5$ (simulated from an exponential distribution with $\theta = 2$), which yielded a sufficient statistic $t = \sum_j x_j = 2.949$, the result is represented in the upper panel of Figure 1. Inferences about the value of a future observation from the same process may are described by the reference predictive posterior

$$\pi(x|\mathbf{t}) = \int_0^\infty \theta e^{-\theta x} \mathrm{Ga}(\theta|n, t)\, d\theta = n t^n (x + t)^{-(n+1)}.$$

Figure 1

Bayesian reference analysis of the parameter θ of an exponential distribution $p(x|\theta) = \theta e^{-x\theta}$, given a sample of size $n = 5$ with $t = \sum_j x_j = 2.949$

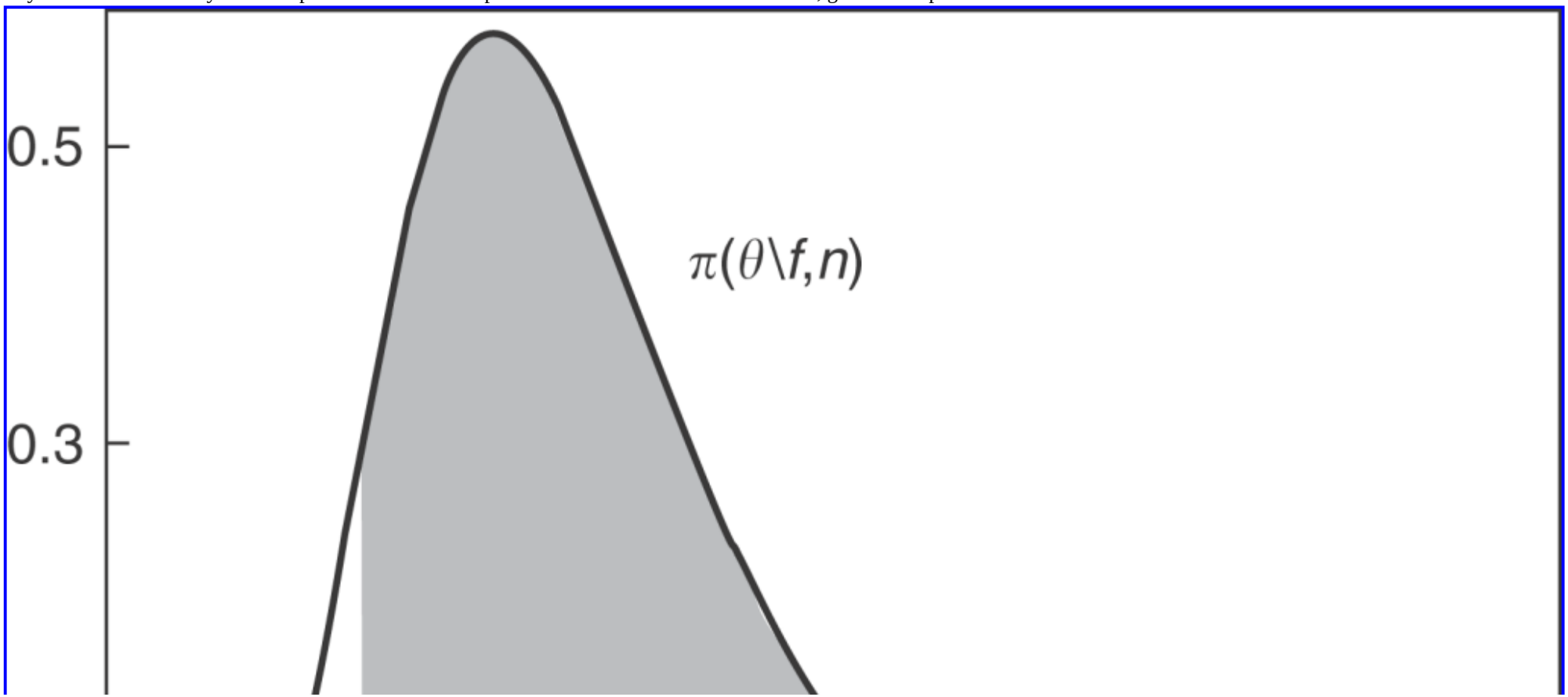

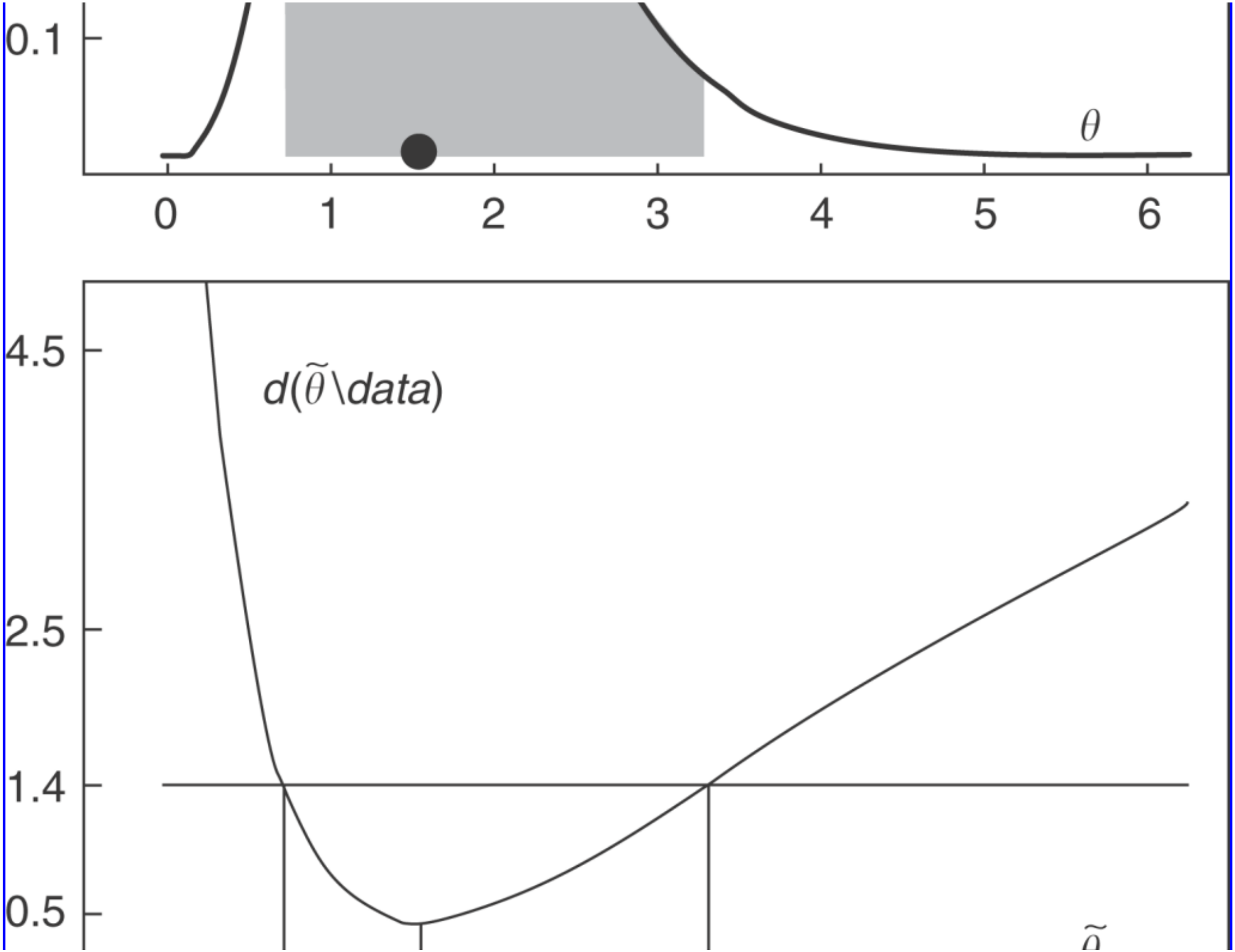
0.1
θ
0
1
2
3
4
5
6
4.5
d(θ̃\data)
2.5
1.4
0.5
θ̃

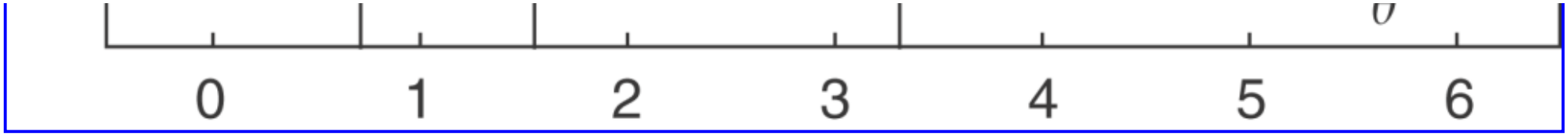

Nuisance parameters

The extension of the reference prior algorithm to the case of two parameters follows the usual mathematical procedure of reducing the problem to a sequential application of the established procedure for the single parameter case. Thus, if one drops explicit mention to the class $\mathcal{P}$ of priors compatible with accepted assumptions to simplify notation, if the probability model is $\{p(\mathbf{t}|\theta, \lambda),\ \theta\in\Theta,\ \lambda\in\Lambda\}$ and a θ-reference prior $\pi_\theta(\theta, \lambda)$ is required, the reference algorithm proceeds in two steps:

1. 1. Conditional on θ, $p(\mathbf{t}|\theta, \lambda)$ depends only on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the conditional reference prior $\pi(\lambda|\theta)$.
2. 2. If $\pi(\lambda|\theta)$ is proper, this may be used to integrate out the nuisance parameter, thus obtaining the one-parameter integrated model

$$p(\mathbf{t}|\theta) = \int_\Lambda p(\mathbf{t}|\theta, \lambda)\, \pi(\lambda|\theta)\, d\lambda$$

to which the one-parameter algorithm may be applied again to obtain $\pi(\theta)$. The θ-reference prior is then $\pi_\theta(\theta, \lambda) = \pi(\lambda|\theta)\, \pi(\theta)$, and the required reference posterior is $\pi(\theta|\mathbf{t}) \propto p(\mathbf{t}|\theta)\, \pi(\theta)$.

If the conditional reference prior $\pi(\lambda|\theta)$ is not proper, then the procedure is performed within an increasing sequence $\{\Lambda_i\}$ of subsets converging to Λ over which $\pi(\lambda|\theta)$ is integrable. This makes it possible to obtain a corresponding sequence of θ-reference posteriors $\{\pi_i(\theta|\mathbf{t})\}$ for the quantity of interest θ, and the required reference posterior is the corresponding pointwise limit $\pi(\theta|\mathbf{t}) = \lim_i \pi_i(\theta|\mathbf{t})$.

The θ-reference prior does not depend on the choice of the nuisance parameter λ. Notice, however, that the reference prior may depend on the parameter of interest; thus, the θ-reference prior may differ from the ϕ-reference prior unless either ϕ is a piecewise one-to-one transformation of θ or ϕ is asymptotically independent of θ. This is an expected consequence of the fact that the conditions under which the missing information about θ is maximized may be different from the conditions under which the missing information about some function $\varphi = \varphi(\theta, \lambda)$ is maximized.

The preceding algorithm may be generalized to any number of parameters. Thus, if the model is $p(\mathbf{t}|\omega_1, \ldots, \omega_m)$, a reference prior $\pi(\theta_m|\theta_{m-1}, \ldots, \theta_1) \times \cdot \times \pi(\theta_2|\theta_1) \times \pi(\theta_1)$ may sequentially be obtained for each ordered parametrization $\{\theta_1(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$ of interest, and these are invariant under re-parametrization of any of the $\theta_i(\omega)$'s. The choice of the ordered parametrization $\{\theta_1, \ldots, \theta_m\}$ precisely describes the particular prior required.

Flat priors

Mathematical convenience often leads to the use of ‘flat’ priors, typically some limiting form of a convenient family of priors; this may, however, have devastating consequences. Consider, for instance, that in a normal setting $p(\mathbf{x}|m) = N_k(\bar{x}_i|m, n^{-1}I)$, inferences are desired on $\theta = \Sigma_{i=1}^k \mu_i^2$, the squared distance of the unknown mean μ to the origin. It is easily verified that the posterior distribution of θ based on a uniform prior on μ (or in any ‘flat’ proper approximation) is strongly inconsistent (Stein's paradox). This is due to the fact that a uniform (or nearly uniform) prior on μ is highly informative about θ, introducing a severe bias on its marginal posterior. The reference prior which corresponds to a parametrization of the form $\{\theta, \boldsymbol{\lambda}\}$ produces, however, for any choice of the nuisance parameter vector $\boldsymbol{\lambda}$, a reference posterior $\pi(\theta|\mathbf{x}, \mathcal{P}_0) \propto \theta^{-1/2}\chi^2(n\theta|k, n\theta)$, where $t = \Sigma_{i=1}^k \bar{x}_i^2$, with appropriate consistency properties. Far from being specific to Stein's example, the inappropriate behaviour in problems with many parameters of specific marginal posterior

distributions derived from multivariate ‘flat’ priors (proper or improper) is indeed very frequent. Hence, sloppy, uncontrolled use of ‘flat’ priors (rather than the relevant reference priors) should be very strongly discouraged.

Inference summaries

From a Bayesian perspective, the final outcome of a problem of inference about any unknown quantity is the corresponding posterior distribution. Thus, given some data *D* and conditions *H*, all that can be said about any function $q = q(\boldsymbol{\omega})$ of the parameters which govern the model is contained in the posterior distribution $p(q|D, H)$, and all that can be said about some function $\boldsymbol{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\boldsymbol{y}|D, H)$. However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to summarize the information contained in the posterior distribution by (*a*) providing values of the quantity of interest which, in the light of the data, are likely to be a good proxy for its true (unknown) value, and by (*b*) measuring the compatibility of the results with hypothetical values of the quantity of interest which might have been suggested in the context of the investigation. The Bayesian counterparts of those of traditional problems of estimation and hypothesis testing are now briefly considered.

Point estimation

Let *D* be the available data, which are assumed to have been generated by a probability model $\{p(D|\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and let $q = q(\boldsymbol{\omega}) \in \Theta$ be the quantity of interest. A *point estimator* of θ is some function of the data $\tilde{q} = \tilde{q}(D)$ which could be regarded as an appropriate proxy for the actual, unknown value of θ. Formally, to choose a point estimate for θ is a decision problem, where the action space is the class Θ of possible θ values. As dictated by the foundations of decision theory, to solve this decision problem it is necessary to specify a loss function $\ell(\tilde{q}, q)$ measuring the consequences of acting as if the true value of the quantity of interest were $\tilde{q}$, when it is actually θ. The expected posterior loss if $\tilde{q}$ were used is

$$l(\tilde{q}|D) = \int_{\Theta} \ell(\tilde{q}, q)\, p(q|D)\, dq, \tag{5}$$

and the corresponding *Bayes estimator* is that function of the data, $q^* = q^*(D)$, which minimizes $l(\tilde{q}|D)$.

For any given model, data and prior, the Bayes estimator obviously depends on the loss function which has been chosen. The loss function is context specific, and should be selected in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for scientific communication. These loss functions produce estimates which may often be regarded as simple descriptions of the location of the posterior distribution. If the loss function is quadratic, so that $\ell(\tilde{q}, q) = (\tilde{q} - q)^t(\tilde{q} - q)$, the corresponding Bayes estimator is the posterior mean $E[q|D]$ (on the assumption that the mean exists). Similarly, if the loss function is a zero-one function, so that $\ell(\tilde{q}, q) = 0$ if $\tilde{q}$ belongs to a ball or radius ε centred in θ and $\ell(\tilde{q}, q) = 1$ otherwise, the corresponding Bayes estimator converges to the posterior mode as the ball radius ε tends to zero (on the assumption that a unique mode exists). If θ is univariate and the loss function is linear, so that $\ell(\tilde{\theta}, \theta) = c_1(\tilde{\theta} - \theta)$ if $\tilde{\theta} \geq \theta$, and $\ell(\tilde{\theta}, \theta) = c_2(\theta - \tilde{\theta})$ otherwise, the Bayes estimator is the posterior quantile of order $c_2 / (c_1 + c_2)$, so that $Pr[\theta < \theta^*] = c_2 / (c_1 + c_2)$. In particular, if $c_1 = c_2$, the corresponding Bayes estimator is the posterior median. The results quoted for linear loss functions clearly illustrate the fact that any possible parameter value may turn out be a Bayes estimator: it all depends on the loss function characterizing the consequences of the anticipated uses of the estimate.

Conventional loss functions are typically non-invariant under re-parametrization, so that the Bayes estimator ϕ^* of a one-to-one transformation $\varphi = \varphi(q)$ of the original parameter θ is not necessarily $\phi(\theta^*)$ (the univariate posterior median, which is invariant, is an interesting exception). Moreover, conventional loss functions focus on the discrepancy between the estimate $\tilde{q}$ and the true value θ, rather then on the more relevant discrepancy between the probability models which they label. Intrinsic losses directly focus on the discrepancy between the probability distributions $p(D|q)$ and $p(D|\tilde{q})$, and typically produce invariant solutions. An attractive example is the intrinsic discrepancy $\delta(\tilde{q}, q)$, defined as the minimum logarithmic divergence between a probability model labelled by θ and a probability model labelled by $\tilde{q}$. When there are no nuisance parameters, this is

$$\delta(\tilde{q}, q) = \min\left\{\kappa(\tilde{q}|q), \kappa(q|\tilde{q})\right\}, \kappa(q_i|\tilde{q}_j) = \int_T p(t|q_j)\log\frac{p(t|q_j)}{p(t|q_i)}dt, \quad (6)$$

where $t = t(D) \in T$ is *any* sufficient statistic (which may well be the whole data set *D*). The definition is easily extended to problems with nuisance parameters. The Bayes estimator is obtained by minimizing the corresponding posterior expected loss. An objective estimator, the *intrinsic estimator* $\tilde{q}_{int} = \tilde{q}_{int}(D)$, is obtained by minimizing the expected intrinsic discrepancy with respect to the *reference* posterior distribution,

$$d(\tilde{q}|d) = \int_\Theta \delta(\tilde{q}, q)\pi(q|D)dq \quad (7)$$

Since the intrinsic discrepancy is invariant under re-parametrization, minimizing its posterior expectation produces *invariant* estimators. Thus, the intrinsic estimator of say, the log of the speed of a galaxy is simply log of the intrinsic estimator of the speed of the galaxy.

Region estimation

To describe the inferential content of the posterior distribution of the quantity of interest $p(q|D)$ it is often convenient to quote *credible* regions, defined as subsets of the parameter space Θ of given posterior probability. For example, the identification of regions containing 50, 90, 95, or 99 per cent of the probability under the posterior may be sufficient to convey the general quantitative messages implicit in $p(q|D)$. Indeed, this is the intuitive basis of graphical representations of univariate distributions like those provided by boxplots. A posterior *q-credible region* for θ is any region $C \subset \Theta$ such that $\int_C p(q|D)dq = q$. Notice that this provides immediately a direct intuitive statement about the unknown quantity of interest θ in probability terms, in marked contrast to the circumlocutory statements provided by conventional confidence intervals. A credible region is invariant under re-parametrization; thus, for any *q*-credible region *C* for θ, $\phi(C)$ is a *q*-credible region for $\varphi = \varphi(q)$.
Clearly, for any given *q* there are generally infinitely many credible regions. Credible regions are often selected to have minimum size (length, area, volume), resulting in highest probability density (HPD) regions, where all points in the region have larger probability density than all points outside. However, HPD regions are not invariant under re-parametrization: the image $\phi(C)$ of an HPD region *C* will be a credible region for ϕ, but will not generally be HPD; indeed, there is no compelling reason to restrict attention to HPD credible regions. In one-dimensional problems, posterior quantiles are often used to derive credible regions. Thus, if $\theta_q = \theta_q(D)$ is the 100*q* per cent posterior quantile of $\theta \in \Theta \subset \Re$, then $C = \{\theta;\ \theta \le \theta_q\}$ is a one-sided, typically unique *q*-credible region, and it is invariant under re-parametrization; the similarly invariant probability centred *q*-credible regions of the form $C = \{\theta;\ \theta_{(1-q)/2} \le \theta \le \theta_{(1+q)/2}\}$ are easier to compute than HPD regions; this notion, however, does not extend to multivariate problems.
Choosing a *p*-credible region may be seen as a decision problem where the action space is the class of all *p*-credible regions. Foundations then dictate that a loss function $\ell(\tilde{q}, q)$ must be specified, and that the region chosen should consist of those $\tilde{\theta}$ values with the lowest expected posterior loss $l(\tilde{q}|D) = \int_\Theta \ell(\tilde{q}|q)\, p(q|D)dq$. By definition, lowest posterior loss (LPL) regions are credible regions where all points in the region have smaller expected posterior loss than all points outside. If the loss function is quadratic, so that $\ell(\tilde{q}, q) = (\tilde{q} - q)^t(\tilde{q} - q)$, the LPL *p*-credible region is a Euclidean sphere centred at the posterior mean $E[\theta|D]$. Like HPD regions, LDL quadratic credible regions are not invariant under re-parametrization; however, LDL intrinsic regions, which minimize the posterior expectation of the invariant intrinsic discrepancy loss (6) are obviously invariant. *Intrinsic p-credible* regions are LDL intrinsic regions which minimize the expected intrinsic discrepancy with respect to the reference posterior distribution. These provide a general, invariant, objective solution to multivariate region estimation. The notions of point and region parameter estimation described above may easily extended to prediction problems by using the posterior predictive rather than the posterior of the parameter.

Hypothesis testing

The posterior distribution $p(q|D)$ of the quantity of interest θ conveys immediate intuitive information on those values of θ which, given the assumed model, may be taken to be *compatible* with the observed data *D*, namely, those with a relatively high probability density. Sometimes, a *restriction* $q \in \Theta_0 \subset \Theta$ of the possible values of the quantity of interest (where Θ_0 may possibly consist of a single value θ_0) is suggested in the course of the investigation as deserving special consideration, either because restricting θ to Θ_0 would greatly simplify the model or because there are additional, context-specific arguments suggesting that $q \in \Theta_0$. Intuitively, the *hypothesis* $H_0 \equiv \{q \in \Theta_0\}$ should be judged to be *compatible* with the observed data *D* if there are elements in Θ_0 with a relatively high posterior density; however, a more precise conclusion is often required and, once again, this is possible with a decision-oriented approach. Formally, testing the hypothesis $H_0 \equiv \{q \in \Theta_0\}$ is a *decision problem* where the action space has only two elements, namely, to accept (a_0) or to reject (a_1) the proposed restriction. To solve this decision problem, it is necessary to specify an appropriate loss function, $\ell(a_i, q)$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value θ of the vector of interest. The optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_\Theta \ell(a_0, q)\, p(q|D)\, dq$, is larger than the expected posterior loss of rejecting, $\int_\Theta \ell(a_1, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$, that is, if (and only if)

$$\int_\Theta [\ell(a_0, q) - \ell(a_1, q)]\, p(q|D)\, dq = \int_\Theta \Delta\ell(q)\, p(q|D)\, dq > 0 \tag{8}$$

Therefore, only the loss difference $\Delta\ell(q) = \ell(a_0, q) - \ell(a_1, q)$, which measures the *advantage* of rejecting H_0 as a function of θ, has to be specified: the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive.

The simplest loss structure has the zero-one form given by $\{\ell(a_0, q) = 0, \ell(a_1, q) = 1\}$ if $q \in \Theta_0$ and, similarly, $\{\ell(a_0, q) = 1, \ell(a_1, q) = 0\}$ if $q \notin \Theta_0$, so that the *advantage* $\Delta\ell(q)$ of rejecting H_0 is 1 if $q \notin \Theta_0$ and it is –1 otherwise. With this, rather naive, loss function the optimal action is to reject H_0 if (and only if) $\Pr(q \notin \Theta_0|D) > \Pr(q \in \Theta_0|D)$. Notice that this formulation requires that $\Pr(q \in \Theta_0) > 0$, that is, that the hypothesis H_0 has a strictly positive prior probability. If θ is a continuous parameter and Θ_0 consists of a single point θ_0 (sharp null problems), this requires the use of a non-regular highly informative prior which places a positive probability mass at θ_0. This posterior probability approach is therefore only appropriate if it is sensible to condition on the assumption that θ is indeed concentrated around θ_0.

Frequently, however, the compatibility of the observed data with H_0 is to be judged without assuming such a sharp prior knowledge. In those situations, the advantage $\Delta\ell(q)$ of rejecting H_0 as a function of θ may be typically assumed to be of the general form $\Delta\ell(q) = \delta(\Theta_0, q) - d^*$, for some $d^* > 0$, where $\delta(\Theta_0, q)$ is some measure of the discrepancy between the assumed model $p(D|q)$ and its closest approximation within the class $\{p(D|q_0), q_0 \in \Theta_0\}$ and such that $\delta(\Theta_0, q) = 0$ whenever $q \in \Theta_0$, and d^* is a context dependent *utility constant* which measures the (necessarily positive) advantage of being able to work with the restricted model when it is true. For reasons similar to those supporting its use in estimation, an attractive choice for the loss function $\delta(\Theta_0, q)$ is an appropriate extension of the intrinsic discrepancy loss; when there are no nuisance parameters, this is given by $\delta(\Theta_0, q) = \inf_{q_0 \in \Theta_0} \delta(q_0, q)$ where $\delta(q_0, q)$ is the intrinsic discrepancy loss defined by (6). The corresponding optimal strategy, called the 'Bayesian reference criterion' (BRC), is then to reject H_0 if, and only if,

$$d(\Theta_0|D) = \int_\Theta \delta(\Theta_0, q)\, \pi(q|D)\, dq > d^*. \tag{9}$$

The choice of d^* plays a similar role to the choice of the significance level in conventional hypothesis testing. Standard choices for scientific communication may be of the form $d^* = \log k$ for, in view of (6) and of (7), this means that the data *D* are expected to be at least *k* times more likely under the true model than under H_0. This is actually equivalent to rejecting H_0 if Θ_0 is not contained in an intrinsic q_k-credible region for θ whose size q_k depends on *k*. Under conditions for asymptotic posterior normality,

$$q_k \approx 2\,\Phi[(2\log k - 1)^{1/2}] - 1,$$

where Φ is the standard normal distribution function. For instance, if $k = 100$, $q_k \approx 0.996$, while if $k = 11.25$, $q_k \approx 0.95$. The Bayesian reference criterion provides a general objective procedure for multivariate hypothesis testing which is invariant under re-parametrization.

Example 2.: **Exponential data, continued**. The intrinsic discrepancy loss for an exponential model is $\delta(\ddot{\theta}, \theta) = g(\varphi)$, if $\varphi \leq 1$, and $\delta(\ddot{\theta}, \theta) = g(1/\varphi)$, if $\varphi > 1$, where $g(\varphi) = \varphi - 1 - \log \varphi$, and $\varphi = \ddot{\theta} / \theta$. Using (7) with the data from Example 1, the expected intrinsic loss $d(\ddot{\theta}|x)$ is the function represented in the lower panel of Figure 1. The intrinsic estimate is the value which minimizes $d(\ddot{\theta}|x)$, $\ddot{\theta}_{int} = 1.546$ (marked with a solid dot in the figure), and the intrinsic 0.90-credible set is (0.720,3.290), the set of parameter values with expected loss below 1.407 (corresponding to the shaded area in the upper panel of the figure).

See Also

- Bayes, Thomas
- Bayesian econometrics
- Bayesian methods in macroeconometrics
- Bayesian nonparametrics
- Bayesian time series analysis
- de Finetti, Bruno
- Savage, Leonard J. (Jimmie)
- statistical decision theory

Bibliography

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Bernardo, J. 2005. Reference analysis. In *Handbook of Statistics* 25, ed. D. Dey and C. Rao. Amsterdam: North-Holland, 17–90.

Bernardo, J.M. and Smith, A.F.M. 1994. *Bayesian Theory*. Chichester: Wiley (2nd edn available in 2007).

Berry, D. 1996. *Statistics: A Bayesian Perspective*. Belmont, CA: Wadsworth.

Box, G. and Tiao, G. 1973. *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. 2003. *Bayesian Data Analysis*, 2nd edn. London: Chapman and Hall.

Jeffreys, H. 1961. *Theory of Probability*, 3rd edn. Oxford: University Press.

Lee, P. 2004. *Bayesian Statistics: An Introduction*, 3rd edn. London: Arnold.

Lindley, D. 1965. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: Cambridge University Press.

O'Hagan, A. 2004. *Bayesian Inference*, 2nd edn. London: Arnold.

Robert, C. 2001. *The Bayesian Choice*, 2nd edn. New York: Springer.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

How to cite this article

Bernardo, José M. "Bayesian statistics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000314> doi:10.1057/9780230226203.0111

The New Palgrave Dictionary of Economics Online

Bayesian time series analysis

Mark F.J. Steel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes the use of Bayesian methods in the statistical analysis of time series. The use of Markov chain Monte Carlo methods has made even the more complex time series models amenable to Bayesian analysis. Models discussed in some detail are ARIMA models and their fractionally integrated counterparts, state space models, Markov switching and mixture models, and models allowing for time-varying volatility. A final section reviews some recent approaches to nonparametric Bayesian modelling of time series.

Keywords

ARCH models; ARFIMA models; ARIMA models; ARMA models; Bayes factor; Bayesian inference; Bayesian methods in econometrics; Bayesian model averaging; Bayesian nonparametrics; Bayesian time series analysis; business cycles; cointegration; computational algorithms; conditional likelihood; continuous-time models; convergence clubs; data augmentation; Dirichlet processes; forecasting; GARCH models; Gibbs sampler; growth regressions; hidden Markov models; impulse response function; Kalman filter; latent states; Lévy processes; long-memory models; macroeconomic forecasting; Markov chain Monte Carlo methods; Markov switching models; Metropolis Hastings sampler; nonparametric models; Ornstein–Uhlenbeck processes; posterior odds; prediction; prior odds; regime switching models; regression; sequential learning; spatial statistics; state space models; stochastic volatility models; survival analysis; threshold autoregressive models; time series analysis; uncertainty; unit roots; vector autoregressions

Article

Bayesian methods

The importance of Bayesian methods in econometrics has increased rapidly since the early 1990s. This has, no doubt, been fuelled by an increasing appreciation of the advantages that Bayesian inference entails. In particular, it provides us with a formal way to incorporate the prior information we often possess before seeing the data, it fits perfectly with sequential learning and decision making, and it directly leads to exact small sample results. In addition, the Bayesian paradigm is particularly natural for prediction, since we take into account all parameter or even model uncertainty. The predictive distribution is the sampling distribution where the parameters are integrated out with the posterior distribution and provides exactly what we need for forecasting, often a key goal of time-series analysis. Usually, the choice of a particular econometric model is not pre-specified by theory, and many competing models can be entertained. Comparing models can be done formally in a Bayesian framework through so-called posterior odds, which is the product of the prior odds and the Bayes factor. The Bayes factor between any two models is the ratio of the likelihoods integrated out with the corresponding prior and summarizes how the data favour one model over another. Given a set of possible models, this immediately leads to posterior model probabilities. Rather than choosing a single model, a natural way to deal with model uncertainty is to use the posterior model probabilities to average out the inference (on observables or parameters) corresponding to each of the separate models. This is called Bayesian model averaging. The latter was already mentioned in Leamer (1978) and recently applied to economic problems in, for example, Fernández, Ley and Steel (2001) (for growth regressions) and in Garratt et al. (2003) and Jacobson and Karlsson (2004) (for macroeconomic forecasting).

An inevitable prerequisite for using the Bayesian paradigm is the specification of prior distributions for all quantities in the model that are treated as unknown. This has been the source of some debate, a prime example of which is given by the controversy over the choice of prior on the coefficients of simple autoregressive models. The issue of testing for a unit root (deciding whether to difference the series before modelling it through a stationary model) is subject to many difficulties from a sampling-theoretical perspective. Comparing models in terms of posterior odds provides a very natural Bayesian approach to testing, which does not rely on asymptotics or approximations. It is, of course, sensitive to how the competing models are defined (for example, do we contrast the stationary model with a pure unit root model or a model with a root larger than or equal to 1?) and to the choice of prior. The latter issues have lead to some controversy in the literature, and prompted a special issue of the *Journal of Applied Econometrics* with animated discussion around the paper by Phillips (1991). The latter paper advocated the use of Jeffreys' principles to represent prior ignorance about the parameters (see also the discussion in Bauwens, Lubrano and Richard, 1999, ch. 6).

Like the choice between competing models, forecasting can also be critically influenced by the prior. In fact, prediction is often much more sensitive than parameter inference to the choice of priors (especially on autoregressive coefficients) and Koop, Osiewalski and Steel (1995) show that imposing stationarity through the prior on the autoregressive coefficient in a simple AR(1) model need not lead to stabilization of the predictive variance as the forecast horizon increases.

Computational algorithms

Partly, the increased use of Bayesian methods in econometrics is a consequence of the availability of very efficient and flexible algorithms for conducting inference through simulation in combination with

ever more powerful computing facilities, which have made the Bayesian analysis of non-standard problems an almost routine activity. Particularly, Markov chain Monte Carlo (MCMC) methods have opened up a very useful class of computational algorithms and have created a veritable revolution in the implementation of Bayesian methods. Whereas Bayesian inference before 1990 was at best a difficult undertaking in practice, reserved for a small number of specialized researchers and limited to a rather restricted set of models, it has now become a very accessible procedure which can fairly easily be applied to almost any model. The main idea of MCMC methods is that inference about an analytically intractable posterior (often in high dimensions) is conducted through generating a Markov chain which converges to a chain of drawings from the posterior distribution. Of course, predictive inference is also immediately available once one has such a chain of drawings. Various ways of constructing such a Markov chain exist, depending on the structure of the problem. The most commonly used are the Gibbs sampler and the Metropolis Hastings sampler. The use of data augmentation (that is, adding auxiliary variables to the sampler) can facilitate implementation of the MCMC sampler, so that often the analysis is conducted on an augmented space including not only the model parameters but also things like latent variables and missing observations. An accessible reference to MCMC methods is, for example, Gamerman (1997).

As a consequence, we are now able to conduct Bayesian analysis of time series models that have been around for a long time (such as ARMA models) but also of more recent additions to our catalogue of models, such as Markov switching and nonparametric models, and the literature is vast. Therefore, I will have to be selective and will try to highlight a few areas which I think are of particular interest. I hope this can give an idea of the role that Bayesian methods can play in modern time series analysis.

ARIMA and ARFIMA models

Many models used in practice are of the simple ARIMA type, which have a long history and were formalized in Box and Jenkins (1970). ARIMA stands for ‘autoregressive integrated moving average’ and an ARIMA(p,d,q) model for an observed series $\{y_t\}$, $t = 1, \ldots, T$ is a model where the dth difference $z_t = y_t - y_{t-d}$ is taken to induce stationarity of the series. The process $\{z_t\}$ is then modelled as $z_t = \mu + \varepsilon_t$ with

$$\varepsilon_t = \varphi_1 \varepsilon_{t-1} + \cdots + \varphi_p \varepsilon_{t-p} + u_t - \theta_1 u_{t-1} - \cdots - \theta_q u_{t-q}$$

or in terms of polynomials in the lag operator L (defined through $L^s x_t = x_{t-s}$):

$$\varphi(L)\varepsilon_t = \theta(L) u_t$$

where $\{u_t\}$ is white noise and usually distributed as $u_t \sim N(0, \sigma^2)$. The stationarity and invertibility conditions are simply that the roots of $\phi(L)$ and $\theta(L)$, respectively, are outside the unit circle. An accessible and extensive treatment of the use of Bayesian methods for ARIMA models can be found in Bauwens, Lubrano and Richard (1999). The latter book also has a useful discussion of multivariate modelling using vector autoregressive (VAR) models and cointegration.

The MCMC samplers used for inference in these models typically use data augmentation. Marriott et al. (1996) use a direct conditional likelihood evaluation and augment with unobserved data and errors to conduct inference on the parameters (and the augmented vectors $\varepsilon_a = (\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{1-p})'$ and $u_a = (u_0, u_{-1}, \ldots, u_{1-q})$). A slightly different approach is followed by Chib and Greenberg (1994), who consider a state space representation and use MCMC on the parameters augmented with the initial state vector.

ARIMA models will either display perfect memory (if there are any unit roots) or quite short memory with geometrically decaying autocorrelations (in the case of a stationary ARMA model). ARFIMA ('autoregressive fractionally integrated moving average') models (see Granger and Joyeux, 1980) have more flexible memory properties, due to fractional integration which allows for hyperbolic decay. Consider $z_t = \Delta y_t - \mu$, which is modelled by an ARFIMA(p, δ ,q) model as:

$$\varphi(L)(1-L)^{\delta} z_t = \theta(L) u_t,$$

where $\{u_t\}$ is white noise with $u_t \sim N(0, \sigma^2)$, and $\delta \in (-1, 0.5)$. The fractional differencing operator $(1-L)^{\delta}$ is defined as

$$(1-L)^{\delta} = \sum_{j=0}^{\infty} c_j(\delta) L^j,$$

where $c_0(\cdot)=1$ and for $j>0$:

$$c_j(a) = \prod_{k=1}^{j} \left(1 - \frac{1+a}{k}\right).$$

This model takes the entire past of z_t into account, and has as a special case the ARIMA(p,1,q) for y_t (for $\delta=0$). If $\delta>-1$, z_t is invertible (Odaki, 1993) and for $\delta<0.5$ we have stationarity of z_t. Thus, we have three regimes:

$\delta \in (-1, -0.5)$: y_t trend-stationary with long memory

$\delta \in (-0.5,0)$: z_t stationary with intermediate memory

$\delta \in (0,0.5)$: z_t stationary with long memory.

Of particular interest is the impulse response function $I(n)$, which captures the effect of a shock of size one at time t on y_{t+n}, and is given by

$$I(n) = \sum_{i=0}^{n} c_i(-\delta-1)J(n-i),$$

with $J(i)$ the standard ARMA(p,q) impulse responses (that is, the coefficients of $\phi^{-1}(L)\,\theta(L)$). Thus, $I(\infty)$ is 0 for $\delta<0$, $\theta(1)/\phi(1)$ for $\delta=0$ and ∞ for $\delta>0$. Koop et al. (1997) analyse the behaviour of the impulse response function for real US GNP data using a set of 32 possible models containing both ARMA and ARFIMA models for z_t. They use Bayesian model averaging to conduct predictive inference and inference on the impulse responses, finding about one-third of the posterior model probability concentrated on the ARFIMA models. Koop et al. (1997) use importance sampling to conduct inference on the parameters, while MCMC methods are used in Pai and Ravishanker (1996) and Hsu and Breidt (2003).

State space models

The basic idea of such models is that an observable y_t is generated by an observation or measurement equation

$$y_t = F_t'\theta_t + v_t,$$

where $v_t \sim N(0, V_t)$, and is expressed in terms of an unobservable state vector θ_t (capturing, for example, levels, trends or seasonal effects) which is itself dynamically modelled through a system or transition equation

$$\theta_t = G_t\theta_{t-1} + w_t,$$

with $w_t \sim N(0, W_t)$ and all error terms $\{v_t\}$ and $\{w_t\}$ are mutually independent. Normality is typically assumed, but is not necessary and a prior distribution is required to describe the initial state vector θ_0. Models are defined by the (potentially time-varying) quadruplets $\{F_t, G_t, V_t, W_t\}$ and the time-varying states θ_t make them naturally adaptive to changing circumstances. This feature also fits very naturally with Bayesian methods, which easily allow for sequential updating. These models are quite general and include as special cases, for example, ARMA models, as well as stochastic volatility models, used in finance (see below).

There is a relatively long tradition of state space models in econometrics and a textbook treatment can already be found in Harvey (1981). Bayesian methods for such models were discussed in, for example, Harrison and Stevens (1976), and a very extensive treatment is provided in West and Harrison (1997), using the terminology ‘dynamic linear models’. An accessible introduction to Bayesian analysis with these models can be found in Koop (2003, Ch. 8).

Online sequential estimation and forecasting with the simple Normal state space model above can be achieved with Kalman filter recursions, but more sophisticated models (or estimation of some aspects of the model besides the states) usually require numerical methods for inference. In that case, the main challenge is typically the simulation of the sequence of unknown state vectors. Single-state samplers (updating one state vector at a time) are generally less efficient than multi-state samplers, where all the states are updated jointly in one step. Efficient algorithms for multi-state MCMC sampling schemes have been proposed by Carter and Kohn (1994) and de Jong and Shephard (1995). For fundamentally non-Gaussian models, the methods in Shephard and Pitt (1997) can be used. A recent contribution of Harvey, Trimbur and van Dijk (2006) uses Bayesian methods for state space models with trend and cyclical components, exploiting informative prior notions regarding the length of economic cycles.

Markov switching and mixture models

Markov switching models were introduced by Hamilton (1989) and essentially rely on an unobserved regime indicator s_t, which is assumed to behave as a discrete Markov chain with, say, K different levels. Given $s_t=i$ the observable y_t will be generated by a time series model which corresponds to regime i, where $i = 1, \ldots, K$. These models are often stationary ARMA models, and the switching between regimes will allow for some non-stationarity, given the regime allocations. Such models are generally known as hidden Markov models in the statistical literature.

Bayesian analysis of these models is very natural, as that methodology provides an immediate framework for dealing with the latent states, $\{s_t\}$, and a simple MCMC framework for inference on both the model parameters and the states was proposed in Albert and Chib (1993). A bivariate version of the Hamilton model is analysed in Paap and van Dijk (2003), who also examine the cointegration relations between the series modelled and find evidence for cointegration between US per capita income and consumption. Using a similar model, Smith and Summers (2005) examine the synchronization of business cycles across countries and find strong evidence in favour of the multivariate Markov switching

model over a linear VAR model.

When panel data are available, another relevant question is whether one can find clusters of entities (such as countries or regions) which behave similarly, while allowing for differences between the clusters. This issue is addressed from a fully Bayesian perspective in Frühwirth-Schnatter and Kaufmann (2006), where model-based clustering (across countries) is integrated with a Markov switching framework (over time). This is achieved by a finite mixture of Markov switching autoregressive models, where the number of elements in the mixture corresponds to the number of clusters and is treated as an unknown parameter. Frühwirth-Schnatter and Kaufmann (2006) analyse a panel of growth rates of industrial production in 21 countries and distinguish two clusters with different business cycles. This also feeds into the important debate on the existence of so-called convergence clubs in terms of income per capita as discussed in Durlauf and Johnson (1995) and Canova (2004).

Another popular way of inducing nonlinearities in time series models is through so-called threshold autoregressive models, where the choice of regimes is not governed by an underlying Markov chain but depends on previous values of the observables. Bayesian analyses of such models can be found in, for example, Geweke and Terui (1993) and are extensively reviewed in Bauwens, Lubrano and Richard (1999, ch. 8). The use of Bayes factors to choose between various nonlinear models, such as threshold autoregressive and Markov switching models is discussed in Koop and Potter (1999).

Geweke and Keane (2006) present a general framework for Bayesian mixture models where the state probabilities can depend on observed covariates. They investigate increasing the number of components in the mixture, as well as the flexibility of the components and the specification of the mechanism for the state probabilities, and find their mixture model approach compares well with ARCH-type models (as described in the next section) in the context of stock return data.

Models for time-varying volatility

The use of conditional heteroskedasticity initially introduced in the ARCH (autoregressive conditional heteroskedasticity) model of Engle (1982) has been extremely successful in modelling financial time series, such as stock prices, interest rates and exchange rates. The ARCH model was generalized to GARCH (generalized ARCH) by Bollerslev (1986). A simple version of the GARCH model for an observable series $\{y_t\}$, given its past which is denoted by I_{t-1}, is the following:

$$y_t = u_t\sqrt{h_t} \tag{1}$$

where $\{u_t\}$ is white noise with mean zero and variance one. The conditional variance of y_t given I_{t-1} is then h_t, which is modelled as

$$h_t = \omega + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_i h_{t-i} \quad (2)$$

where all parameters are positive and usually $p = q = 1$ is sufficient in practical applications. Bayesian inference for such models was conducted through importance sampling in Kleibergen and van Dijk (1993) and, with MCMC methods, in Bauwens and Lubrano (1998).

An increasingly popular alternative model allows for the variance h_t to be determined by its own stochastic process. This is the so-called stochastic volatility model, which in its basic form replaces (2) by the assumption that the logarithm of the conditional volatility is driven by its own AR(1) process

$$\ln(h_t) = \alpha + \delta \ln(h_{t-1}) + v_t,$$

where $\{v_t\}$ is a white noise process independent of $\{u_t\}$ in (1). Inference in such models requires dealing with the latent volatilities, which are incidental parameters and have to be integrated out in order to evaluate the likelihood. MCMC sampling of the model parameters and the volatilities jointly is a natural way of handling this. An MCMC sampler where each volatility was treated in a separate step was introduced in Jacquier, Polson and Rossi (1994), and efficient algorithms for multi-state MCMC sampling schemes were suggested by Carter and Kohn (1994) and de Jong and Shephard (1995). Many extensions of the simple stochastic volatility model above have been proposed in the literature, such as correlations between the $\{u_t\}$ and $\{v_t\}$ processes, capturing leverage effects, or fat-tailed distributions for u_t. Inference with these more general models and ways of choosing between them are discussed in Jacquier, Polson and Rossi (2004).

Recently, the focus in finance has shifted more towards continuous-time models, and continuous-time versions of stochastic volatility models have been proposed. In particular, Barndorff-Nielsen and Shephard (2001) introduce a class of models where the volatility behaves according to an Ornstein–Uhlenbeck process, driven by a positive Lévy process without Gaussian component (a pure jump process). These models introduce discontinuities (jumps) into the volatility process. Barndorff-Nielsen and Shephard (2001) also consider superpositions of such processes. Bayesian inference in such models through MCMC methods is complicated by the fact that the model parameters and the latent volatility process are often highly correlated in the posterior, leading to the problem of over-conditioning. Griffin and Steel (2006b) propose MCMC methods based on a series representation of Lévy processes, and avoid over-conditioning by dependent thinning methods. In addition, they extend the model by including a jump component in the returns, leverage effects and separate risk pricing for the various volatility components in the superposition. An application to stock price data shows substantial empirical support for a superposition of processes with different risk premiums and a leverage effect. A different approach

to inference in such models is proposed in Roberts, Papaspiliopoulos and Dellaportas (2004), who suggest a re-parameterization to reduce the correlation between the data and the process. The re-parameterized process is then proposed only in accordance with the parameters.

Semi- and nonparametric models

The development and use of Bayesian nonparametric methods has been a rapidly growing topic in the statistics literature, some of which is reviewed in Müller and Quintana (2004). However, the latter review does not include applications to time series, which have been perhaps less prevalent than applications in other areas, such as regression, survival analysis and spatial statistics.

Bayesian nonparametrics is sometimes considered an oxymoron, since Bayesian methods are inherently likelihood-based, and thus require a complete probabilistic specification of the model. However, what is usually called Bayesian nonparametrics corresponds to models with priors defined over infinitely dimensional parameter spaces (functional spaces) and this allows for very flexible procedures, where the data are allowed to influence virtually all features of the model.

Defining priors over collections of distribution functions requires the use of random probability measures. The most popular of these is the so-called Dirichlet process prior introduced by Ferguson (1973). This is defined for a space Θ and a σ-field B of subsets of Θ. The process is parameterized in terms of a probability measure H on (Θ, B) and a positive scalar M. A random probability measure, F, on (Θ, B) follows a Dirichlet process $DP(MH)$ if, for any finite measurable partition, $B_1, \ldots, B_k$, the vector $(F(B_1), \ldots, F(B_k))$ follows a Dirichlet distribution with parameters $(MH(B_1), \ldots, MH(B_k))$. The distribution H centres the process and M can be interpreted as a precision parameter.

The Dirichlet process is (almost surely) discrete and, thus, not always suitable for modelling observables directly. It is, however, often incorporated into semiparametric models using the hierarchical framework

$$y_i \sim g(y_i|u_i) \text{ with } u_i \sim F \text{ and } F \sim DP(MH), \tag{3}$$

where $g(\cdot)$ is a probability density function. This model is usually referred to as a ‘mixture of Dirichlet processes’. The marginal distribution for y_i is a mixture of the distribution characterized by $g(\cdot)$. This basic model can be extended: the density $g(\cdot)$ or the centring distribution H can be (further) parameterized, and inference can be made about these parameters. In addition, inference can be made about the mass parameter M. Inference in these models with the use of MCMC algorithms has become quite feasible, with methods based on MacEachern (1994) and Escobar and West (1995).

However, the model in (3) assumes independent and identically distributed observations and is, thus, not directly of interest for time series modelling. A simple approach followed by Hirano (2002) is to use (3) for modelling the errors of an autoregressive model specification. However, this does not allow for the distribution to change over time. Making the random probability measure F itself depend on lagged values of the variable under consideration y_t (or, generally, any covariates) is not a straightforward

extension. Müller, West and MacEachern (1997) propose a solution by modelling y_t and y_{t-1} jointly, using a mixture of Dirichlet processes. The main problem with this approach is that the resulting model is not really a conditional model for y_t given y_{t-1}, but incorporates a contribution from the marginal model for y_{t-1}. Starting from the stick-breaking representation of a Dirichlet process, Griffin and Steel (2006a) introduce the class of order-based dependent Dirichlet processes, where the weights in the stick-breaking representation induce dependence between distributions that correspond to similar values of the covariates (such as time). This class induces a Dirichlet process at each covariate value, but allows for dependence. Similar weights are associated with similar orderings of the elements in the representation and these orderings are derived from a point process in such a way that distributions that are close in covariate space will tend to be highly correlated. One proposed construction (the arrivals ordering) is particularly suitable for time series and is applied to stock index returns, where the volatility is modelled through an order-based dependent Dirichlet process. Results illustrate the flexibility and the feasibility of this approach. Jensen (2004) uses a Dirichlet process prior on the wavelet representation of the observables to conduct Bayesian inference in a stochastic volatility model with long memory.

Conclusion: where are we heading?

In conclusion, Bayesian analysis of time series models is alive and well. In fact, it is an ever growing field, and we are now starting to explore the advantages that can be gained from using Bayesian methods on time series data. Bayesian counterparts to the classical analysis of existing models, such as AR(F) IMA models, are by now well-developed and a lot of work has already been done there to make Bayesian inference in these models a fairly routine activity. The main challenge ahead for methodological research in this field is perhaps to further develop really novel models that not merely constitute a change of inferential paradigm but are inspired by the new and exciting modelling possibilities that are available through the combination of Bayesian methods and MCMC computational algorithms. In particular, nonparametric Bayesian time-series modelling falls in that category and I expect that more research in this area will be especially helpful in increasing our understanding of time series data.

See Also

- ARCH models
- Bayesian econometrics
- Bayesian methods in macroeconometrics
- Bayesian nonparametrics
- Bayesian statistics
- econometrics
- long memory models
- Markov chain Monte Carlo methods
- state space models

- statistics and economics
- stochastic volatility models
- time series analysis

Bibliography

Albert, J. and Chib, S. 1993. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics* 11, 1–15.

Barndorff-Nielsen, O. and Shephard, N. 2001. Non-Gaussian OU based models and some of their uses in financial economics. *Journal of the Royal Statistical Society Series B* 63, 167–241 (with discussion).

Bauwens, L. and Lubrano, M. 1998. Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal* 1, C23–C46.

Bauwens, L., Lubrano, M. and Richard, J.F. 1999. *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–27.

Box, G. and Jenkins, G. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

Canova, F. 2004. Testing for convergence clubs in income per capita: a predictive density approach. *International Economic Review* 45, 49–77.

Carter, C. and Kohn, R. 1994. On Gibbs sampling for state space models. *Biometrika* 81, 541–53.

Chib, S. and Greenberg, E. 1994. Bayes inference in regression models with ARMA(p,q) errors. *Journal of Econometrics* 64, 183–206.

de Jong, P. and Shephard, N. 1995. The simulation smoother for time series models. *Biometrika* 82, 339–50.

Durlauf, S. and Johnson, P. 1995. Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics* 10, 365–84.

Engle, R. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1008.

Escobar, M. and West, M. 1995. Bayesian density-estimation and inference using mixtures. *Journal of the American Statistical Association* 90, 577–88.

Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1, 209–230.

Fernández, C., Ley, E. and Steel, M. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16, 563–76.

Frühwirth-Schnatter, S. and Kaufmann, S. 2006. Model-based clustering of multiple time series. *Journal of Business and Economic Statistics* (forthcoming).

Gamerman, D. 1997. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton, FL: Chapman and Hall/CRC.

Garratt, A., Lee, K., Pesaran, H. and Shin, Y. 2003. Forecast uncertainties in macroeconometric modelling: an application to the UK economy. *Journal of the American Statistical Association* 98, 829–38.

Geweke, J. and Terui, N. 1993. Bayesian threshold autoregressive models for nonlinear time series. *Journal of Time Series Analysis* 14, 441–54.

Geweke, J. and Keane, M. 2006. Smoothly mixing regressions. *Journal of Econometrics* (forthcoming).

Granger, C. and Joyeux, R. 1980. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 15–39.

Griffin, J. and Steel, M. 2006a. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* 101, 179–94.

Griffin, J. and Steel, M. 2006b. Inference with non-Gaussian Ornstein-Uhlenbeck processes for stochastic volatility. *Journal of Econometrics* 134, 605–44.

Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–84.

Harrison, P. and Stevens, C. 1976. Bayesian forecasting. *Journal of the Royal Statistical Society Series B* 38, 205–47 (with discussion).

Harvey, A. 1981. *Time Series Models*. Oxford: Philip Allen.

Harvey, A., Trimbur, T. and van Dijk, H. 2006. Trends and cycles in economic time series: a Bayesian approach. *Journal of Econometrics* (forthcoming).

Hirano, K. 2002. Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* 70, 781–99.

Hsu, N. and Breidt, F. 2003. Bayesian analysis of fractionally integrated ARMA with additive noise. *Journal of Forecasting* 22, 491–514.

Jacobson, T. and Karlsson, S. 2004. Finding good predictors for inflation: a Bayesian model averaging approach. *Journal of Forecasting* 23, 479–96.

Jacquier, E., Polson, N. and Rossi, P. 1994. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics* 12, 371–417 (with discussion).

Jacquier, E., Polson, N. and Rossi, P. 2004. Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics* 122, 185–212.

Jensen, M.J. 2004. Semiparametric Bayesian inference of long-memory stochastic volatility models. *Journal of Time Series Analysis* 25, 895–922.

Kleibergen, F. and van Dijk, H. 1993. Non-stationarity in GARCH models: a Bayesian analysis. *Journal of Applied Econometrics* 8, S41–S61.

Koop, G. 2003. *Bayesian Econometrics*. Chichester: Wiley.

Koop, G., Ley, E., Osiewalski, J. and Steel, M. 1997. Bayesian analysis of long memory and persistence using ARFIMA models. *Journal of Econometrics* 76, 149–69.

Koop, G., Osiewalski, J. and Steel, M. 1995. Bayesian long-run prediction in time series models. *Journal of Econometrics* 69, 61–80.

Koop, G. and Potter, S. 1999. Bayes factors and nonlinearity: evidence from economic time series. *Journal of Econometrics* 88, 251–81.

Leamer, E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

MacEachern, S. 1994. Estimating Normal means with a conjugate style Dirichlet process prior. *Communications in Statistics, B* 23, 727–41.

Marriott, J., Ravishanker, N., Gelfand, A. and Pai, J. 1996. Bayesian analysis of ARMA processes: complete sampling-based inference under exact likelihoods. In *Bayesian Analysis in Statistics and Econometrics*, ed. D. Berry, K. Chaloner and J. Geweke. New York: Wiley.

Müller, P. and Quintana, F. 2004. Nonparametric Bayesian data analysis. *Statistical Science* 19, 95–110.

Müller, P., West, M. and MacEachern, S. 1997. Bayesian models for nonlinear autoregressions. *Journal of Time Series Analysis* 18, 593–614.

Odaki, M. 1993. On the invertibility of fractionally differenced ARIMA processes. *Biometrika* 80, 703–09.

Paap, R. and van Dijk, H. 2003. Bayes estimation of Markov trends in possibly cointegrated series: an application to U.S. consumption and income. *Journal of Business and Economic Statistics* 21, 547–63.

Pai, J. and Ravishanker, N. 1996. Bayesian modeling of ARFIMA processes by Markov chain Monte Carlo methods. *Journal of Forecasting* 16, 63–82.

Phillips, P. 1991. To criticize the critics: an objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics* 6, 333–473 (with discussion).

Roberts, G., Papaspiliopoulos, O. and Dellaportas, P. 2004. Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society Series B* 66, 369–93.

Shephard, N. and Pitt, M. 1997. Likelihood analysis of non-Gaussian measurement time series. *Biometrika* 84, 653–67.

Smith, P.A. and Summers, P.M. 2005. How well do Markov switching models describe actual business cycles? The case of synchronization. *Journal of Applied Econometrics* 20, 253–74.

West, M. and Harrison, P. 1997. *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer Verlag.

How to cite this article

Steel, Mark F. J. "Bayesian time series analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000100> doi:10.1057/9780230226203.0112

The New Palgrave Dictionary of Economics Online

Beccaria, Cesare Bonesana, Marchese di (1738 – 1794)

Peter Groenewegen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Beccaria, C. B.; Bentham, J.; Cantillon, R.; division of labour; Genovesi, A.; Hume, D.; Locke, J.; Montesquieu, C. de; Physiocracy; population; productive vs. unproductive labour; property; public finance; Quesnay, F.; smuggling; taxation; utilitarianism; wealth

Article

Italian economist, philosopher and statesman, Beccaria was born in Milan in 1738, educated at Parma and in law at Pavia, appointed Professor of Political (Public) Economy or Cameral Science in Milan (1768), resigned his chair to enter public service (1772), where he encouraged and implemented monetary, general economic and penal reforms and advocated a decimal system of weights, measures and coin. He died in Milan in 1794. Beccaria's greatest fame derives from his *Essay on Crimes and Punishment* (1764a), which made his European reputation almost overnight and ensured his magnificent reception when he visited Paris in 1766. Among others, it exerted considerable influence on Bentham's utilitarian philosophy (Halévy, 1928) and popularized the phrase, ‘the greatest happiness of the greatest number’ (Beccaria, 1764a, Introduction). He also enjoyed considerable reputation as an economist. This was based on his work on Milanese monetary problems of 1762 and the outline of his teaching programme and inaugural lecture of 1769 (translated into French and English). His most important economic work is an unfinished treatise, *Elementi di economia pubblica* (written in 1771 but not published till 1804), but his mathematical contribution to the economics of taxation and smuggling (1764b) is also of considerable interest (see Theocharis, 1961).

Beccaria (1764b) starts with a methodological point on the use of algebra in political and economic reasoning. He considered such use only legitimate when the analysis concerned quantities, hence not all subject matter of these sciences was amenable to mathematical reasoning. He then illustrates the use of algebra for solving an economic problem, namely, how much of a given quantity of merchandise must merchants smuggle in order to break even, even if the remainder of the goods is confiscated. The essay may have been inspired by Hume's ‘Of the Balance of Trade’ (1752, p. 76) with its comment on ‘Swift's

maxim' [that] 'in the arithmetic of the customs, two and two make not four, but often only one', because alterations in rates may alter revenue quite disproportionately.

Beccaria's plan for university instruction in economics and his inaugural lecture develop a classification of the subject matter into five, interconnected parts: general principles and overview, agriculture, trade, manufactures and public finance. Further subdivisions into chapters are reminiscent of the table of contents of Cantillon (1755), a work he appears to have studied closely, though the historical part of his inaugural lecture only acknowledges Vauban, Melon, Montesquieu, Uztariz, Ulloa, Hume and Genovesi. The last is described as the father of Italian economics (Beccaria, 1769). Groenewegen (1983) demonstrates that Beccaria's economic sources also included Locke and Quesnay's articles published in the French *Encyclopédie*. The last gave parts of the *Elementi* a Physiocratic flavour; for example, in the analysis of large- and small-scale farming, productive and unproductive labour and, more generally, its emphasis on the importance of agriculture.

Beccaria sees political economy as a highly practical subject, because it is part of the science of legislation and politics. Its purpose is to 'increase the wealth of the state and its subjects, by giving instruction on the most appropriate and useful management of the national revenue and that of the sovereign' (1769, p. 341). Although abstract treatment of the science is therefore largely rejected as inappropriate for such a practical subject, Beccaria maintains that serious discussion of its elements needs an introduction of general principles. A definition of wealth as 'things not only necessary but also convenient and elegant', starts these principles in Part I of the *Elementi*. Because wealth consists of goods designed to meet the needs of food, shelter and clothing, the science can be justifiably subdivided into parts derived from the sectors of production and exchange which supply the various wants of mankind. Raw materials are drawn from farming, pastoral activity, mineral exploitation and fishing, hence agriculture is the first part of political economy. Raw materials require work and preparation before they can be used, hence manufacturing is the second part. Efficient production of wealth creates a surplus available for exchange, hence commerce including value, money and credit constitutes the third part to be treated. Since protection of property is a prerequisite for efficient production and trade, public finance explaining how these expenses of government are met is the fourth element. Finally, Beccaria suggests a fifth topic to cover police and other government activity, but nothing of this nor the public finance part of his *Elementi* were ever completed. Having defined the scope of the subject in terms of wealth and the component parts helping its production, Beccaria elaborates on the principles in his theory of reproduction, or the combination of labour, time and capital which ensures the continuation of production activity. Here Beccaria demonstrates awareness of the links between division of labour and trade and recognizes that the prices which circulate commodities are regulated by necessary costs of production. A general analysis of the cost of labour or wages, of the advances and other means of production and of those incurred by the state in its essential protection of production activity, is therefore required. Beccaria further develops these general principles by examining the nature and interdependence of work and consumption, introducing considerations of thrift, value, profit, useful work, variability of wants and difficulties in measuring the subsistence wage of workers. A discussion of the principle of population concludes the analysis of the 'simple truths' and 'self-evident axioms' from which the whole science of political economy can be deduced, as Beccaria intended to demonstrate in the other parts of his work. Of these, the completed chapters in Part IV on value, money and exchange are of the greatest interest.

Selected works

1764a. *An Essay on Crime and Punishment*. London: J. Almon, 1767.

1764b. An Attempt at an Analysis of Smuggling. *II Caffè*, vol 1. Brescia, 118–19. In *Precursors in Mathematical Economics: An Anthology*, ed. W. Baumol and S. Goldfield. London: London School of Economics, 1968, 149–50.

1769. *A Discourse on Public Oeconomy and Commerce*. Translated from the Italian. London: J. Dodsley.

1771. Elementi di economia pubblica. In *Cesare Beccaria Opere*, ed. S. Romagnoli. Florence: Sansoni, 1958.

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Reprint with English translation edited by H. Higgs for the Royal Economic Society. London: Macmillan, 1931.

Groenewegen, P.D. 1983. Turgot, Beccaria and Smith. In *Italian Economics Past and Present*, ed. P.D. Groenewegen and J. Halevi. Sydney: Frederick May Foundation for Italian Studies.

Halévy, E. 1928. *The Growth of Philosophic Radicalism*. Trans. M. Morris. Boston: Beacon Press, 1955.

Hume, D. 1752. Of the balance of trade. Reprinted in *David Hume, Writings on Economics*, ed. E. Rotwein. London: Nelson, 1955.

Theocharis, R.D. 1961. *Early Developments in Mathematical Economics*. London: Macmillan.

How to cite this article

Groenewegen, Peter. "Beccaria, Cesare Bonesana, Marchese di (1738–1794)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000101> doi:10.1057/9780230226203.0113

The New Palgrave Dictionary of Economics Online

Becker, Gary S. (born 1930)

Casey B. Mulligan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Gary S. Becker has produced major economics books and articles for more than 50 years. His studies dominate labour economics and have significantly impacted studies of crime, habit formation, and other important behaviours once considered beyond the scope of economics. Some of Gary's lasting impact can be attributed to abstraction from institutional detail and his ‘thinking problems through fully’.

Keywords

Becker, G.; Buchanan, J.; Chicago School; Downs, A.; family economics; flat tax; Friedman, M.; human capital; Knight, F.; labour market discrimination; politics and economics; pressure groups; Ricardian equivalence theorem; rotten kid theorem; Schumpeter, J.; Stigler, G.; time allocation; Tullock, G.

Article

I walk over to my collection of *The American Economic Review*, and pick up the very first (and now disintegrating) issue, dated 1952, and notice an article entitled ‘A Note on Multi-Country Trade’. Its author is Gary S. Becker. By the time you read this, you probably can pick up the very latest issue from your collection and find an article by Gary S. Becker! If you did so in 2005, I can guarantee it: the article was entitled ‘The Quality and Quantity of Life and the Evolution of World Inequality’. Gary published an important article in the very first issue of the *Journal of Law and Economics*, ‘Competition and Democracy’ (1958). He published an article, ‘Deadweight Costs and the Size of Government’ in the 46th volume of the same journal (Becker and Mulligan, 2003a); it may have the same potential, although I must admit that its importance cannot yet be judged impartially.
Figure 1 quantitatively examines Gary's work over a half century. The vertical axis measures, from the Social Science Citation Index (SSCI), the number of articles citing each of Gary's books and major research projects. Each citation has a citer and a citee. The citees are Gary's *Economics of Discrimination* (1957, various editions), *Human Capital* (1964, various editions), *A Treatise on the Family* (1981, various editions), *Accounting for Tastes* (1996), ‘A Theory of the Allocation of Time’ (1965), ‘Crime and Punishment: An Economic Approach’ (1968), four journal articles on addiction (Becker and Murphy, 1988a; Becker, Grossman and Murphy, 1991; 1994; and Becker, 1992), and four journal articles on fertility (Becker and Lewis, 1973; Becker and Tomes, 1976; Becker and Barro, 1988; Barro and Becker, 1989). The citers are social science journal articles published in the year indicated on the horizontal axis. Since the articles are typically peer-reviewed and the journals are academic, the vertical axis is a measure (admittedly imperfect) of how important Gary's various works were in making intellectual progress, or in shaping the thinking behind intellectual progress, in social science. Notice the scale on the vertical axis – it reaches past 100 citations per year per work of Gary's – and remember that there are tenured professors at leading economics departments whose citations *combined for all of their works and all of their lives* do not reach these levels. Also notice the scale on the horizontal axis: it begins in 1960. (A fuller analysis of citations would separate year effects from other determinants of citations – for example, the number of journals covered by SSCI may increase over time; I owe this point to Bill Landes. However, the reader might make some guess at the year effects from the fact that *Human Capital*'s citation time series is quite similar to those of Schumpeter, 1942, and Downs, 1957. *Human Capital*'s citations significantly exceed and grow faster than those of Friedman, 1957, and Friedman and Schwartz, 1963.) *Discrimination* and *Treatise* are both heavily cited, but their first editions appeared 24 years apart. The addiction work first appeared 31 years after *Discrimination*. (The two pressure group papers, discussed later, appeared 26 and 28 years respectively after *Discrimination*, and surpassed 50 combined citations per year by 1990.) If Gary manages another big hit during the next few years, that would be a 50-year span.

Figure 1

Citations of Becker's major books and articles, excluding those on political economy. *Note*: For Addiction and Fertility, I sum citations for the four articles in each class; some double counting may occur due to articles that cite more than one of the four. Becker's political economy articles may be more important than the Fertility and Addiction articles, but for clarity the former are omitted from Figure 1 and deferred until later. Social Economics (Becker and Murphy, 2003b) is also omitted because, as of 2005, its annual citations were fewer than ten.

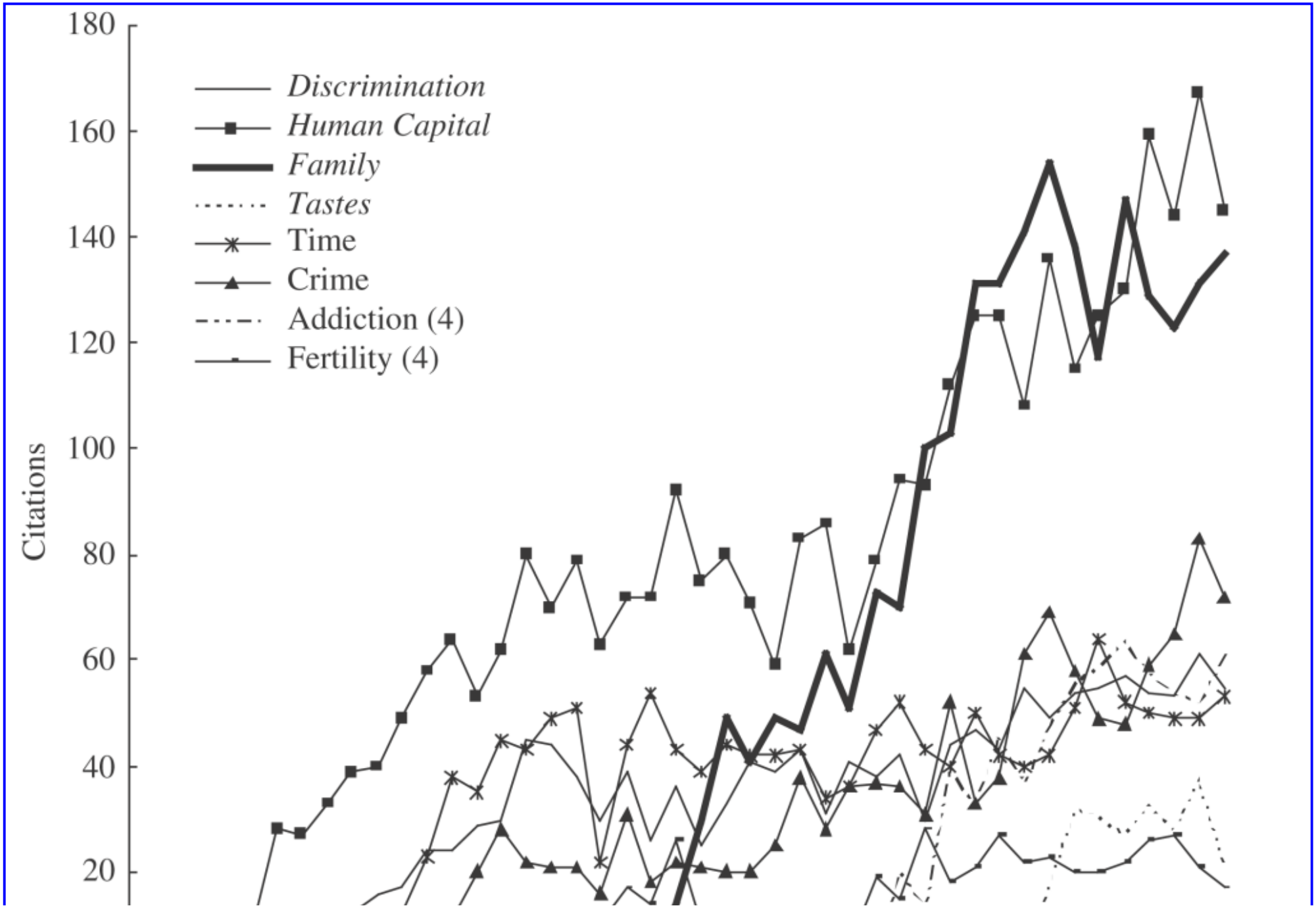

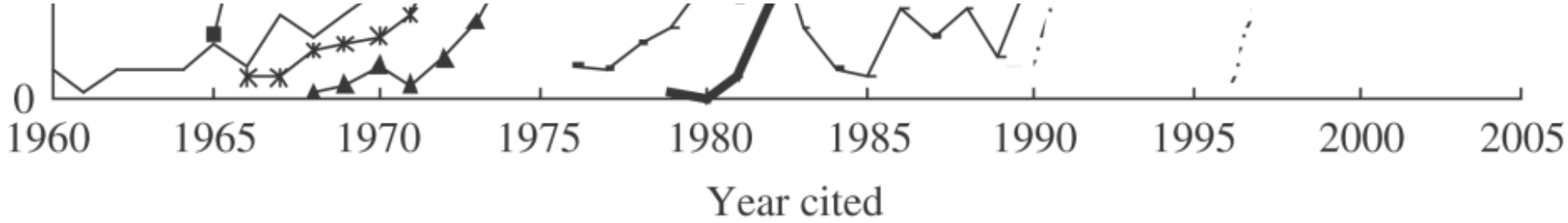

Note: For Addiction and Fertility, I sum citations for the four articles in each class; some doublecounting may occur due to articles that cite more than one of the four.Becker's political economy articles may be more important than the Fertility and Addiction articles, but for clarity the former are omitted from Figure 1 and deferred until later. *Social Economics* (Becker and Murphy, 2003b)is also omitted because, as of 2005, its annual citations were fewer than ten.

In 1999 – to me that seems a long time ago – I visited Wayne State University and met for the first time John Owen, a labour economist whom I knew by reputation. I was both flattered and wiser for this emeritus professor's making the trip to campus to meet me and hear my seminar. As we talked, his style of economic reasoning seemed familiar to me, so I asked him where he obtained his Ph.D. He replied 'I am one of Gary's students, of course'. Apparently Gary Becker alumni have been filling the emeritus professor ranks for a while now. Jack Nicklaus had better win the Masters a couple more times if he wants to be as good at golf as Gary is at economics.

It could be a hundred years or more before economics sees another iron man like Gary. Biographies about Becker should be written if for no other reason than people will ask 'How did he do it?' But why should I be writing a biography, and what could I possibly contribute to answering this difficult question? After all, Gary is closer in age to my grandfather than to my father, so I am certainly no authority on where he was born, what kind of student he was, and so on. By the time I first met Becker in 1991, his Nobel Prize was only one year away. On the other hand, I do know (some more closely than others) many of the important intellectual companions in his life, including Guity Nashat Becker, Aaron Director, Milton Friedman, Jacob Mincer, Sherwin Rosen, Gale Johnson, Jim Coleman, Bill Landes, Bob Lucas, Sam Peltzman, Dick Posner, Isaac Erlich, Kevin Murphy, Robert Barro, Eddie Lazear, Victor Fuchs, Ed Glaeser, Andy Rosenfield, and Tomas Philipson. The opportunity cost of time is certainly lower for me than for those on this list. (Becker's work is so widely applicable that it can even be used to predict who'd write his biography(ies).) Gary loves economics dearly, so perhaps my best tribute would exploit my perspective as a 14-year student, colleague, and friend of Gary's – who was always glad to hear stories about Gary's achievements and the University of Chicago from older students and colleagues such as John Owen and the other names mentioned above – in order to convey some information about Gary's life that is not readily found in a literal reading of his published work, and might help future economists progress a little faster.

The first section raises the question of whether and how the University of Chicago might have affected Gary's intellectual contributions. The second section discusses Gary's timing in the marketplace for economics ideas. Did Gary leave some potential unrealized? The third section addresses this question, with emphasis on economic approaches to political behaviour. Gary's results sometimes seem pretty obvious, but the fourth section explains how this judgement is usually the perspective of hindsight. It offers a number of remarkable examples of how economists, including Gary himself, took a while to fully understand the implications of his economic approach to the family, the labour market, and other areas.

Did Chicago matter?

I'm told that Becker first came to the University of Chicago in 1951 as a graduate student. How much did it matter that he came to Chicago rather than accepting a nice fellowship at Harvard? Some of Gary's undergraduate work at Princeton foreshadowed two of his important contributions to economics. First was the trade paper I mentioned above. Trade theory features prominently in *The Economics of Discrimination*, and even today is still an intense interest of Gary's, as his colleagues today can see any time a trade paper is presented in front of the economics faculty. I doubt that Chicago has done much to cultivate this interest. Second is Gary's 'A Theory of Competition among Pressure Groups for Political Influence' (1983). In one sense, Chicago was necessary for the production of this paper, because it grew out of a comment on Peltzman's 1976 paper in the *Journal of Law and Economics* and a dialogue with Stigler as to whether the political process favoured efficiency or special interests. However, Gary may have been thinking seriously about competition in the public sector during his Princeton days, since already in his first year at Chicago he was writing the first drafts of 'Competition and Democracy', which was published in the inaugural issue of the *Journal of Law and Economics* only after being squashed at the *Journal of Political Economy* by another important Chicago economist, Frank Knight. (Today Gary credits some of his early thinking on democracy to his reading of Schumpeter's *Capitalism, Socialism, and Democracy*, 1942, but he does not remember whether he read it

before coming to Chicago, or shortly after.)

Before coming to Chicago, Gary was already dissatisfied with the lack of applications of economics to important social problems, although his Princeton work does not yet show any success at resolving his discontent. Perhaps Chicago, and especially Milton Friedman, inspired or at least encouraged the application of economics beyond the usual areas. As Gary says, ‘[Friedman] emphasized that economic theory was not a game played by clever academicians, but was a powerful tool to analyse the real world. His course was filled with insights both into the structure of economic theory and its application to practical and significant questions’ (Becker, 1993). Gary is now known for his application of economic theory to practical and significant questions, from time allocation and fertility to inequality and addictions.

Gary sometimes explains, ‘I was such an outsider from the eastern and western establishments for so long’. Universities like Stanford, Harvard, and Yale have never showed any interest in hiring him, although Harvard granted him an honorary degree in 2003. Gary's abilities as an economist are so extraordinary that, despite being an outsider, and having such a large fraction of his productivity ahead of him, he was recognized in 1967 by the American Economic Association as the best young economist at the time (he won the their John Bates Clark Medal in that year). Gary's outsider position would have been different had he turned down Chicago's fellowship, but fortunately for him citations and academic job offers have very different production functions, at least as regards their use of personal acquaintances as inputs.

At Chicago Gary met, loved, and improved the workshop system. Columbia University was the first beneficiary of those improvements when he and Mincer created the Labor Economics workshop (Landes, 1998). Gary started a workshop when he returned to Chicago in 1970, which for many years was co-organized with Sherwin Rosen, and is now affectionately known as the ‘Applications Workshop’. By the time I began attending economics workshops in 1991, practically all had become (and maybe had always been?) something like lecture series, and were a form of *output* of the idea production process, namely, a process for disseminating finished research results. But the Applications Workshop was and is deliberately different; research papers are invited in their infancy, and 85 of the 90•minutes consist of the audience's (especially Gary's; Gary carefully reads the paper beforehand) trying to push the author in new and better directions. Students regularly come to the workshop to hear what Gary has to say, and, in the midst of a graduate programme that can easily overwhelm them with technical detail, learn that good choices of research question and basic strategy for seeking an answer are important and scarce academic skills. Gary later organized with the late James Coleman (Richard Posner continues the tradition) an interdisciplinary workshop on applications of rational models to economics, sociology, law, politics, anthropology, and so on. The success of these two workshops make Chicago a unique and highly stimulating experience for faculty and students, and probably would not be possible if it weren't for Gary's extraordinary breadth of knowledge, quickness of mind, and insatiable appetite for workshops.

Abstracting from institutional detail

The workshop system and Economics 301 (Chicago's first Ph.D. course in price theory) were important means by which Gary received his inheritance from Chicago, and made his bequest to students at Chicago and Columbia, where Becker was an economics professor from 1957 to 1970. I mentioned Friedman's lesson that economic theory was not a game played by clever academicians, but was a powerful tool to analyse practical and significant questions. Chicago was methodologically unique in two other ways. Despite their working on practical questions, Chicago economists were willing, and even eager, to abstract from institutional details, and view price theory as a general method to understanding many different behaviours. This approach was particularly novel in labour economics, where labour unions, marriage bars, and other personnel practices were often interpreted as having an independent influence on labour market outcomes, rather than as outcomes themselves of more basic and ubiquitous forces. Columbia's Jakob Mincer also practised this methodology in his enduring work on labour supply (for example, Mincer, 1962). Labour market institutions like trade unions and monster.com (an internet site where employers can read resumes posted by potential employees) come and go, but the fundamental economic forces include the income and substitution effects on labour supply featured at Columbia by Mincer and at Chicago by Lewis (Lewis, 1956), and are an important part of explanations of why labour market outcomes vary over time and across regions. It's no coincidence that Becker and Mincer together created the Labor Economics workshop at Columbia, and work appearing during these years by Becker, Mincer, and students continued the practice. (William Landes – Gary's student, colleague and friend during both the Chicago and Columbia years – wrote in 1998 an excellent biography of Becker which explains more about the Columbia days and Gary's influence on the law and economics field. To Landes’ account I would add that Gary still credits the City of New York with inspiring ‘Crime and punishment’. One day he illegally parked his car near Columbia's campus because he calculated it to be more important to attend a dissertation defense than to avoid the city's illegal parking fine.)

Human Capital also has some roots in Gary's time at Chicago before 1957. Chicago's agricultural economics group (Gary was one of the participants in those days), especially Ted Schultz, had attributed much of the underdevelopment problem to a lack of human capital investment. Gary's *Human Capital* explains why some people have more income from employment than others by viewing labour income as a dividend on historical investments, which in turn are understood as particular instances of capital accumulation. The basic concepts do not include labour market institutions, but rather the time value of money, ageing, the allocation of time, and other determinants of the costs and benefits of enhancing a person's productivity in the marketplace. Becker's abstractions facilitated applications of human capital theory beyond (perhaps) even what he had anticipated, including the determinants of sickness and health (Grossman, 1972, a Columbia Ph.D. student 1964–70), and the evolution of species (Robson and Kaplan, 2003).

‘A Theory of the Allocation of Time’ introduced the concepts of ‘full income’ and the ‘full price’ of a commodity. A commodity's full price combines the expenditures of money and time required to acquire one unit of its services. Because households differ in terms of the opportunity cost of their time, and perhaps also their time-efficiency in obtaining

commodities, they will face different full prices even though they face the same money price. For example, the substitution effect suggests that richer households (to the extent that the market rewards them highly for their time) would have fewer children and, per unit consumed, would replenish less often their inventories of household commodities (and currency: Karni, 1973). (For the same reason, Gary is perennially puzzled why rich people play golf; he plays tennis.) Full income is the money income that would be obtained if time were allocated in order to maximize money income. In many ways, full income permits time allocation to be studied as a particular application of consumer demand, because full income is spent on some combination of market expenditures on commodities and implicit expenditures on non-work time. Full income and full price are not institution-specific concepts, permitting 'Time' to be applied in so many different sub-fields, including monetary economics, fertility, lobbying (Mulligan and Sala-i-Martin, 1999), altruism (Mulligan, 1997), and even Communism (Boycko, 1992).

Public policy schisms

Milton Friedman's *Capitalism and Freedom* (1962) and *Free to Choose* (1981) clearly advertise the view that inefficient public policies are bad ideas unfortunately and inexplicably hatched by policy-makers, which can be rectified merely by giving some combination of voters, politicians, and bureaucrats a better economic education. If Gary continued that tradition, as with his *Business Week* column and internet blog, he did so with much less vigour. One of Chicago's important influences on Gary came from George Stigler, who often viewed public policies as the rational choices of politicians and the people who can influence them. Perhaps Stigler's influence was stronger because Friedman was there to contrast it, but in any case it's hard to see any Friedman in 'Pressure Groups' (1983) or 'The Family and the State' (1988b).

Interestingly, this schism persists today in Chicago's Economics Department and the economics profession more widely. A public finance group, embodied at Chicago in its macro group (for example, Lucas and Stokey, 1983; Shimer and Werning, 2003), aims at technical and normative public policy improvements, whereas political economists (for example, Becker and Murphy, 1988; Mulligan, Gil, and Sala-i-Martin, 2004) view public policies and their imperfections as the outcomes of other economic forces, such as demography, political competitiveness, and the technology of tax collection.

Becker (1983) also tries to bridge a gap among political economists – a gap defined according to whether they see special interests or efficiency as the primary determinant of actual public policies. He points out that a huge number of groups would like special favours from the government, but only a few can ultimately be successful. These groups compete with each other to obtain the favours. All else the same, groups advocating efficient public policies have an advantage because (by definition of efficiency) their policy proposals would hurt relatively little. Of course, group cohesion, political entry barriers, group size, and other variables may give particular groups an intrinsic advantage, but the competitive activity of special interest groups helps deliver efficient policies to the public sector rather than crowding out such policies with inefficient special favours.

Unfortunately, Becker has not (yet) bridged another gap among political economists – a gap defined by the degree of attention to institutional detail. It's interesting that labour economics work done by Gary and others at Chicago is praised for its lack of institutional detail (detail now considered unnecessary for understanding the major economic forces at work), whereas the political economics work is criticized, at least so far, for the same lack of detail.

Timing in the marketplace for ideas: human capital or luck?

Human Capital and 'Time' had some good fortune in their timing, both in terms of the ultimate demand for these ideas and in terms of the supply of intellectual building blocks. For example, *Human Capital*'s citations accelerated in the late 1980s as the profession came to realize the important wage structure changes that were occurring and began to write about them; human capital theory is probably the most common way of organizing and interpreting such observations. It may also be fortunate that, since 1940, the Census Bureau has been asking more people more questions about wages and schooling than about household expenditure, hence stimulating more empirical research on wages and schooling than empirical research on consumption.

Perhaps there was also good fortune on the input side. Mincer was making significant progress in the empirical analysis of labour supply and the empirical analysis of wage determinants. The economics of consumption was a very lively subject at Chicago in the 1950s, as evidenced by Friedman's *A Theory of the Consumption Function* (1957), work by Margaret Reid (1957), and the beginnings of Chicago's workshop system by Chicago's agricultural economics group. Gary's work on the value of time and life-cycle profiles must have been stimulated in this environment, in part because labour supply and human capital accumulation are such natural applications of the life-cycle way of thinking already apparent in *A Theory of the Consumption Function*. Remember also that Friedman (1957) was preceded by *Income From Independent Professional Practice* (1945), which straddled the fields we would now call consumption and labour economics. (Gregg Lewis was probably yet another Chicago influence in these days.) The economic concept of 'full income' first appeared in 'Time', where Becker credits the phrase to a conversation with Milton Friedman.

Gary adopted and improved the analytical style of *A Theory of the Consumption* Function and the methodology of positive economics more generally. Some consider Friedman's *A Theory of The Consumption Function* the best economics book since the 19th century, and perhaps earlier, because of its convincing and systematic applications of economic theory to important questions. But *Human Capital* may be even better. Both books clearly aim to develop refutable empirical implications from their theories, but *Human Capital* probably

does more to help its reader distinguish the important implications from the secondary ones. Gary always advises his students and colleagues to ‘think a problem through *fully*’ and apparently he followed his own advice in *Human Capital*. Not only is the importance of the basic ‘human capital’ concept appreciated several decades later, but modern analysis of the labour market still displays more detailed similarities, including attention to specific versus general human capital, comparisons between financial and human capital rates of return, the distinction between the forgone earnings and tuition components of human capital acquisition costs, and so on. Friedman's basic concept of permanent income and the details of his analysis of it (such as ‘distributed lags’) are less prevalent today, having been displaced by consumption Euler equations. (Almost immediately after *Human Capital*'s publication, its citation flow exceeded and grew faster than that of *A Theory of the Consumption Function*.) By thinking through the problem fully, Gary had produced in the early 1960s an analysis that would depreciate slowly, and thereby still be available in the 1980s to take advantage of the real-world events that drew attention to human capital questions.

Unrealized potential?

Only people who know Gary personally would know, or dare to believe, that he may have some regrets that he did not realize his full potential. His political economics work is an important instance. He regrets the obscurity of ‘Competition and democracy’, which has been cited only 33 times – less than once per year. He partly blames editor Aaron Director for forgetting to request revisions or proofs of the manuscript, and himself for not following up on work that he knew to be incomplete.

Political economics research has proliferated since the mid 1980s. Gary feels that progress might have been more significant if ‘A Theory of Competition Among Pressure Groups for Political Influence’ had received more attention. I am inclined to agree (Mulligan, Gil and Sala-i-Martin, 2004), but it would be much too extreme to say the article was ‘ignored’. Yes, it was rejected by the *American Economic Review* and perhaps another journal (Gary does not remember). Nevertheless, it may ultimately be the most cited article appearing in the *Quarterly Journal of Economics* since 1983. It has been cited almost 50 times every year since the 1980s. Only three articles – which happen to be from the economic growth literature: Heston and Summers (1991), Barro (1991), and Mankiw, Romer and Weil (1992) – have been cited more than 50 times per year for more than a couple of years, and their citation flows have regressed back to Gary's since 2000. (I thank Andrei Shleifer for suggesting comparisons between Becker, 1983, and other top *QJE* articles.) Two other *QJE* articles – Katz and Murphy (1992) since 1997 and Fehr and Schmidt (1999) since 2003 – enjoy about the same citation flow as Gary's, but over a much more recent period of time. ‘Pressure Groups’ citations are in the stratosphere in the universe of journal articles, but nevertheless it has been losing political economics market share as its annual cites have been pretty steady at 50 while the political economics literature has exploded. Figure 2 compares ‘Pressure Groups’ citations (summed here for the *QJE* and *Journal of Public Economics* articles) with some other political economics work. This time citations are displayed on a log scale. ‘Pressure Groups’ citations are shown as a thick solid line. Buchanan and Tullock's *Calculus of Consent* (1962) – maybe Nobel Laureate Buchanan's best known work – has had the same flow of citations since 1990, although of course the *Calculus of Consent* was published much earlier and deserves enormous credit for introducing to economics the principle of modelling policy-makers as self-interested. Perhaps more striking is the fact that ‘Pressure Groups’ citations have not grown with the political economics literature since 1985. For example, Alberto Alesina now accumulates about 200 citations per year (of all of his papers combined, see the dashed line in Figure 2). Meltzer and Richard's (1981) paper is actually older than Gary's, but it received very few citations until the late 1990s, when its citation flow increased by almost an order of magnitude. Downs (1957, dotted line) and Schumpeter (1942, circles) have also benefited from the growth of political economics. (Mancur Olson's *Logic of Collective Action*, 1965, might have been included in Figure 2; since 1980 its citation flow is about 50 per cent more than Downs's.)

Figure 2
Citations of major political economy works

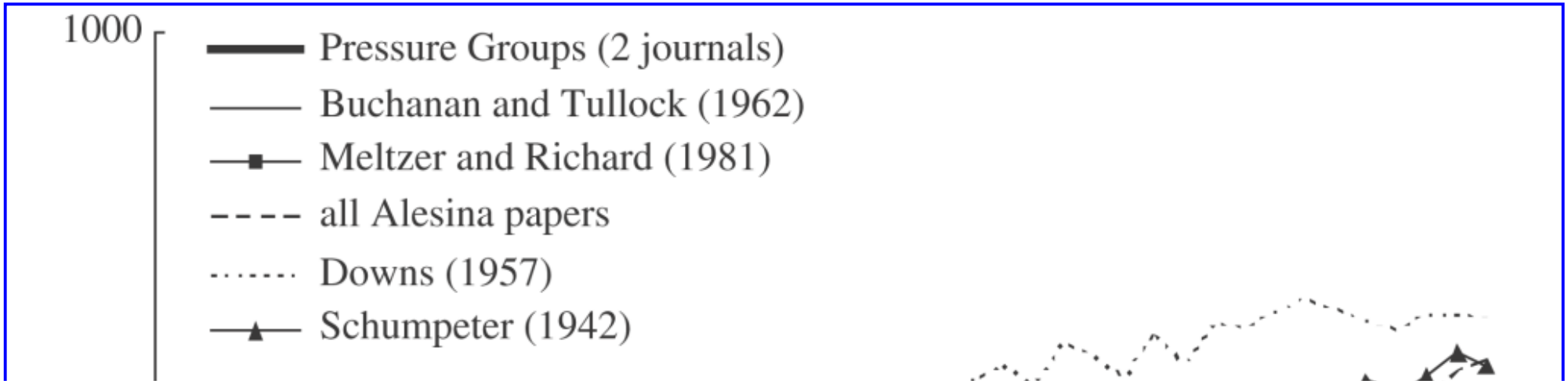

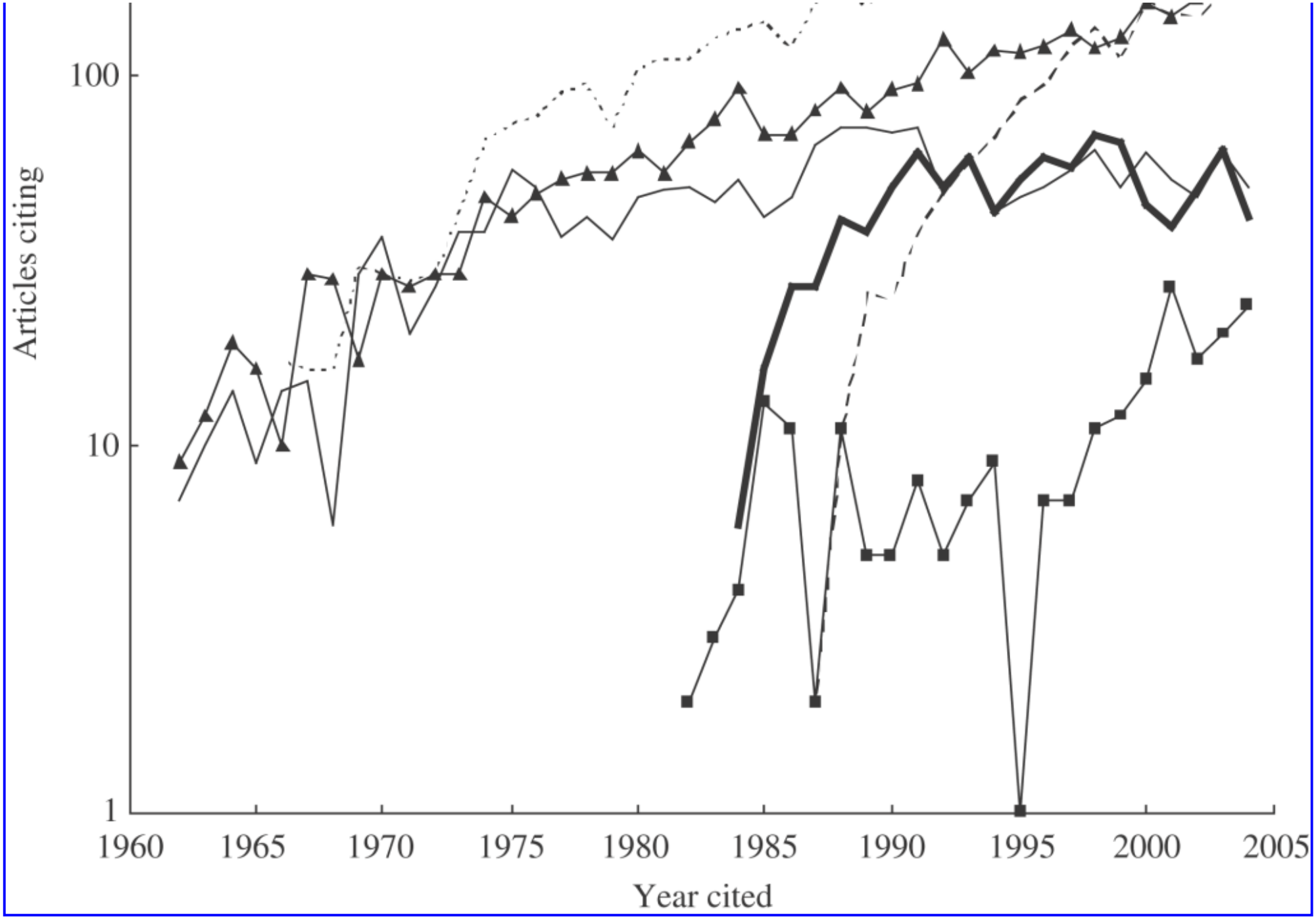

Perhaps ‘Pressure Groups’ should have been part of, or led to, a Becker political economics book that worked more fully through the implications of competition for the supply of public policies. Does it matter whether competition is time-intensive or goods-intensive? How competitive are authoritarian regimes? To judge from Gary's treatment of labour economics questions, it seems very likely that a Becker political economics book would have treated fundamental economics forces like deadweight costs, competition, and the allocation of time with little attention to institutional detail. Would such a book have succeeded in the current marketplace for political economics ideas? On the one hand, the answer

seems to be ‘no’ because the current literature prides itself on its analysis of those details; Persson and Tabellini (2004, p. 76) explain, ‘...the devil is in the details, especially the details of electoral systems’; see also Besley and Case (2003, p. 11). On the other hand, Gary's book may have pushed, or at least nudged, the literature in a different direction.

Wasn't it all obvious?

Perhaps this is a slight exaggeration, but some of Gary's results have been criticized as being too obvious, or adding too little value to simpler non-economic models or common-sense interpretations. I have to admit that I sometimes found it easier to remember the basic results of Gary's journal articles, and to produce simple derivations of my own (for example, Mulligan, 1997, ch. 3), than to follow Gary's published derivations. (I don't remember the derivations presented in Gary's University of Chicago courses to be so clear, either. But maybe I deserve much of the blame here; I am much better at following a geometric proof than an algebraic one, whereas Gary seems to prefer the latter.) To some extent, these critiques have the advantage of hindsight; it is quite normal for original ideas to be expressed later by followers in simpler terms, after a period of what Gary calls ‘cleaning up’. However, I believe that Gary's books are easier to follow than several of his journal articles, because the process of writing a whole book was complementary with some cleaning up on his own. This is also part of the reason why Becker and Murphy make such a good team; one of Murphy's extraordinary talents is to quickly conceive of a concise mathematical expression of a new economic idea.

Becker and Tomes (1979; 1986) reinterpret inter-temporal consumption theory and combine it with human capital theory to form a theory of the evolution of inequality from one generation to the next. In the model, altruistic parents allocate dynastic resources between themselves and their children. The opportunities for doing so depend on the process of monetary inheritance (for example, inheritance taxes) and on the technology for investing in the human capital of children. The model predicts that earnings regress to the mean across generations because ability, talent, and so forth (which determine the rate of return to human investment) regress to the mean. Perhaps the most explicit form of the ‘too obvious’ criticism appeared as Goldberger's (1989) contention that this approach to inheritance is an excessively complicated way of saying ‘economic characteristics regress to the mean’. Becker's (1989) reply lists some implications that are more than regression to the mean, although in some cases I think the results still derive from statistical rather then economic modelling assumptions (see Mulligan, 1997, and the references cited therein). Nevertheless, Becker's ‘micro-economic-optimizing approach’ is the only one, to my knowledge, predicting that consumption would regress to the mean more slowly than earnings. It's a nice bonus that, so far, the empirical evidence seems to support Gary in this regard.

For many years, and perhaps even now, it was far from obvious that wages are largely determined by human capital, as evidenced, for example, by the various debates on wage gaps by industry, race, and gender. The opponents of the human capital interpretation of industry gaps have, after several years, softened their view. Gender and race gaps are sometimes attributed to discrimination (Gary gets some credit under this interpretation, too), although there seem to be steady streams of new evidence showing that the effects of human capital have been too quickly misinterpreted as effects of discrimination (see, for example, Smith and Welch, 1989, and Neal and Johnson, 1996, on race gaps, and Mulligan and Rubinstein, 2005, on gender gaps).

As Gary began working on the family, he found ‘redistribution of income among members does not affect the consumption or welfare of any member because it simply induces offsetting changes in transfers from the head. As a result, each member is at least partially insured against disasters that may strike him’ (Becker, 1974, p. 1091). Put this way, the result seems obvious. However, the result could not have been fully understood at the time – otherwise the rotten kid theorem, the Ricardian equivalence result, and a number of other results would not have shaken the profession so much. Indeed, Gary himself did not fully appreciate its implications, because he admits not foreseeing how the macroeconomics of fiscal policy would change after 1974 thanks to Barro's (1974) article in the same issue of the *Journal of Political Economy*. (Barro's focus at the time was probably contemporaneous work on fiscal policy, such as Feldstein's famous 1974 article in the previous *JPE*. Barro, 1998, explains how the links between Ricardian Equivalence and the Rotten Kid Theorem began to be appreciated only when the *JPE* began preparing the November 1974 issue in which the two articles were to appear.) Peter Diamond's reaction (as reported second-hand by Barro, 1998) demonstrates the fallacy of dismissing these results as obvious, ‘[Ricardian equivalence is] obvious, of no practical significance, and surely not worth … research time.’ Professor Diamond was giving this advice in 1967 to student Bob Hall, who, if it weren't for his listening, was on the verge of scooping both Becker and Barro.

During the 1996 US presidential campaign, Republican primary candidate Steve Forbes revitalized the idea of replacing the current income tax with a ‘flat tax’: a tax with no deductions and low marginal rates. I was concerned that a painless tax would be a tax that Congress would exploit to obtain ever larger amounts of revenue, but to me this point was just something clever to publish in the op.-ed. pages or to make people pause at cocktail parties. I vividly remember mentioning this to Gary in March 1996. He was a flat tax fan at the time (see Becker et al., 1996), and told me ‘I'm not sure how you would analyse that formally and, besides, Hong Kong refutes your hypothesis: they have a flat tax and a small government’. A few days later he apparently saw the empirical evidence differently, and was excited enough to interrupt his trip in France to type a short first draft of our ‘Deadweight Costs and the Size of Government’ and attach it to an e-mail to me back at the University of Chicago. By then he was sure how to analyse it: using a simple version of his 1983 pressure group model. The lesson for the young assistant professor: think a problem through *fully*, regardless of how obvious the answer might seem at first glance. The rewards in this case were, among other things, a consistent analysis of tax reforms, spending reforms, and ‘flypaper effects’ (the tendency of governments to spend non-tax revenue rather than refund it to taxpayers), and a better understanding of the relations between democratic and authoritarian public sectors.

See Also

- family economics
- human capital

Selected works

1952. A note on multi-country trade. *American Economic Review* 42, 558–68.

1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

1958. Competition and democracy. *Journal of Law and Economics* 1, 105–9.

1964. *Human Capital*. New York: Columbia University Press for the NBER.

1965. A theory of the allocation of time. *Economic Journal* 75, 493–508.

1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

1973. (With H. Lewis.) 1973. On the interaction between the quantity and quality of children. *Journal of Political Economy* 81, Part II, S279–S288.

1974. A theory of social interactions. *Journal of Political Economy* 82, 1063–93.

1976. (With N. Tomes.) Child endowments and the quantity and quality of children. *Journal of Political Economy* 84, Part II, S143–S162.

1979. (With N. Tomes.) An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87, 1153–89.

1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 98, 371–400.

1985. Public policies, pressure groups, and deadweight costs. *Journal of Public Economics* 28, 329–47.

1986. (With N. Tomes.) Human capital and the rise and fall of families. *Journal of Labor Economics* 4, S1–S39.

1988. (With R. Barro.) A reformulation of the economic theory of fertility. *Quarterly Journal of Economics* 103, 1–25.

1988a. (With K. Murphy.) A theory of rational addiction. *Journal of Political Economy* 96, 675–700.

1988b. (With K. Murphy.) The family and the state. *Journal of Law and Economics* 31, 1–18.

1989. On the economics of the family: reply to a skeptic. *American Economic Review* 79, 514–18.

1989. (With R. Barro.) Fertility choice in a model of economic growth. *Econometrica* 57, 481–501.

1991. (With M. Grossman and K. Murphy.) Rational addiction and the effect of price on consumption. *American Economic Review* 81, 237–41.

1992. Habits, addictions, and traditions. *Kyklos* 45, 327–45.

1993. Gary S. Becker – Autobiography. In *Les Prix Nobel. The Nobel Prizes 1992*, ed. T. Frängsmyr. Stockholm: Almqvist and Wiksell International. Online. Available at http://nobelprize.org/economics/laureates/1992/becker-autobio.html, accessed 2 December 2005.

1994. (With M. Grossman and K. Murphy.) An empirical analysis of cigarette addiction. *American Economic Review* 84, 396–418.

1996. *Accounting for Tastes*. Cambridge, MA: Harvard University Press.

1996. (With others.) The flat-tax: ‘nutty’ it's not. *Wall Street Journal*. 22 February.

2003a. (With C. Mulligan.) Deadweight costs and the size of government. *Journal of Law and Economics* 46, 293–340.

2003b. (With K. Murphy.) *Social Economics: Market Behavior in a Social Environment*. Cambridge, MA: Belknap Press.

2005. (With T. Philipson and R. Soares.) The quality and quantity of life and the evolution of world inequality. *American Economic Review* 95, 277–91.

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1095–117.

Barro, R. 1991. Economic growth in a cross section of countries. *Quarterly Journal of Economics* 106, 407–43.

Barro, R. 1998. Reflections on Ricardian equivalence. In *Debt and Deficits: An Historical Perspective*, ed. J. Cheltenham: Edward Elgar.

Besley, T. and Case, A. 2003. Political institutions and policy choices: evidence from the United States. *Journal of Economic Literature* 41, 7–73.

Boycko, M. 1992. When higher incomes reduce welfare: queues, labor supply, and macro equilibrium in socialist economies. *Quarterly Journal of Economics* 107, 907–20.

Buchanan, J. and Tullock, G. 1962. *The Calculus of Consent.* Ann Arbor, MI: University of Michigan Press.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper.

Fehr, E. and Schmidt, K. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 817–68.

Feldstein, M. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82, 905–26.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton, NJ: Princeton University Press.

Friedman, M. 1962. *Capitalism and Freedom*. Chicago: University of Chicago Press.

Friedman, M. and Friedman, R.. 1981. *Free to Choose*. New York: Harcourt Brace Jovanovich.

Friedman, M. and Kuznets, S. 1945. *Income from Independent Professional Practice*. New York: National Bureau of Economic Research.

Friedman, M. and Schwartz, A. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press.

Goldberger, A. 1989. Economic and mechanical models of intergenerational transmission. *American Economic Review* 79, 504–13.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80, 223–55.

Karni, E. 1973. The transactions demand for cash: incorporation of the value of time into the inventory approach. *Journal of Political Economy* 81, 1216–25.

Katz, L. and Murphy, K. 1992. Changes in relative wages, 1963–1987: supply and demand factors. *Quarterly Journal of Economics* 107, 35–78.

Landes, W. 1998. Gary S. Becker Biography. In *The New Palgrave Dictionary of Economics and the Law*, ed. P. Newman. London: Macmillan Reference.

Lazear, E. 2000. Economic imperialism. *Quarterly Journal of Economics* 115, 99–146.

Lewis, H. 1956. Hours of work and hours of leisure. *Annual Proceedings of the Industrial Relations Research Association* 12, 196–206.

Lucas, R., Jr. and Stokey, N. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12, 55–93.

Mankiw, N., Romer, D. and Weil, D. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107, 407–37.

Meltzer,, Allan, H. Scott, F. and Richard. A rational theory of the size of government. *Journal of Political Economy*. 89(5), October 1981: 914–27.

Mincer,, Jacob. Labor force participation of married women: a study of labor supply. in H.G. Lewis, ed. *Aspects of Labor Economics*. Princeton: Princeton University Press, 1962.

Mulligan,, Casey B. *Parental Priorities and Economic Inequality*. Chicago: University of Chicago Press, 1997.

Mulligan, C., Gil, R. and Sala-i-Martin, X. 2004. Do democracies have different public policies than nondemocracies? *Journal of Economic Perspectives* 18(1), 51–74.

Mulligan, C. and Rubinstein, Y. 2005. Selection, investment, and women's relative wages since 1975. Working Paper No. 11159. Cambridge, MA: NBER.

Mulligan, C. and Sala-i-Martin, X. 1999. Gerontocracy, retirement, and social security. Working Paper No. 7117. Cambridge, MA: NBER.

Neal, D. and Johnson, W. 1996. The role of premarket factors in black-white wage differences. *Journal of Political Economy* 104, 869–95.

Olson, M. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Peltzman, S. 1976. Toward a more general theory of regulation. *Journal of Law and Economics* 19, 211–40.

Persson, T. and Tabellini, G. 2004. Constitutions and economic policy. *Journal of Economic Perspectives* 18(1), 75–98.

Reid, M. 1952. Effect of income concept upon expenditure curves of farm families. In *Studies of Income and Wealth*, vol. 15. New York: NBER.

Robson, A. and Kaplan, H. 2003. The evolution of human life expectancy and intelligence in huntereconomies. *American Economic Review* 93, 150–69.

Schumpeter, J. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper and Brothers.

Shimer, R. and Werning, Ivan. 2003. Optimal unemployment insurance with sequential search. Mimeo. University of Chicago.

Smith, J. and Welch, F. 1989. Black economic progress after Myrdal. *Journal of Economic Literature* 27, 519–64.

Summers, R. and Heston, A. 1991. The Penn World Table (Mark 5): an expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* 106, 327–68.

How to cite this article

Mulligan, Casey B. "Becker, Gary S. (born 1930)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000301> doi:10.1057/9780230226203.0114

The New Palgrave Dictionary of Economics Online

behavioural economics and game theory

Faruk Gul

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Behavioural economics, broadly defined, refers to the research programme that investigates the relationship between psychology and economic behaviour. The purpose of this article is to provide an outline of behavioural economics research and to describe where research in behavioural game theory stands within this outline. The aim is not to assess the impact of particular contributions or describe and interpret specific applications. Rather, the goal is to provide an organization of the literature based on the type of departures from standard theory.

Keywords

ambiguity; ambiguity aversion; asymmetric information; auctions; behavioural economics; behavioural economics and game theory; commitment; evolutionary games; expected utility hypothesis; fixed point theorems; framing effects; game theory; independence axiom; interdependent preferences; mechanism design; neuroeconomics; preference reversals; probability distributions; prospect theory; psychological games; social preferences; ultimatum game; uncertainty

Article

In traditional economic analysis, as well as in much of behavioural economics, the individual's motivations are summarized by a utility function (or a preference relation) over possible payoff-relevant outcomes while his cognitive limitations are described as incomplete information. Thus, the standard economic theory of the individual is couched in the language of constrained maximization and statistical inference.

The approach gains its power from the concise specification of payoff-relevant outcomes and payoffs as well as a host of auxiliary assumptions. For example, it is typically assumed that the individual's preferences are well behaved: that is, they can be represented by a function that satisfies conditions appropriate for the particular context such as continuity, monotonicity, quasi-concavity, and so on.

When studying behaviour under uncertainty, it is often assumed that the individual's preference obeys the expected utility hypothesis. More importantly, it is assumed that the individual's subjective assessments of the underlying uncertainty are reasonably close to the observed distributions of the corresponding variables. Even after all these bold assumptions, the standard model would say little if the only relevant observation regarding the utility function is one particular choice outcome. Thus, economists will often assume that the same utility function is relevant for the individual's choices over some stretch of time during which a number of related choices are made. One hopes that these observations will generate enough variation to identify the decision-maker's (DM's) utility function. If not, the analyst may choose to utilize choice observations from different contexts to identify the individual's preferences or make parametric assumptions. The analyst may even pool information derived from observed choices of different individuals to arrive at a representative utility function.

1 Experimental challenges to the main axioms of choice theory

The simplest type of criticism of the standard theory accepts the usual economic abstractions and the standard framework but questions specific assumptions within this framework.

1.1 The independence axiom

Allais (1953) offers one of the earliest critiques of standard decision-theoretic assumptions. In his experiment, he provides two pairs of binary choices and shows that many subjects violate the expected utility hypothesis, in particular, the independence axiom. Allais's approach differs from the earlier criticisms: Allais questions an explicit axiom of choice theory rather than a perceived implicit assumption such as 'rationality'. Furthermore, he does so by providing a simple and clear experimental test of the particular assumption.

Subsequent research documents related violations of the independence axiom and classifies them. Researchers have responded to Allais's critique by developing a class of models that either abandons the independence axiom or replaces it with weaker alternatives. The agents in these models still maximize their preference and still reduce uncertainty to probabilistic assessments (that is, they are probabilistically sophisticated), but have preferences over lotteries that fail the independence axiom.

Non-expected utility preferences pose a difficulty for game theory: because many non-expected utility theories do not lead to quasi-concave utility functions, standard fixed point theorems cannot be used to establish the existence of Nash equilibrium. Crawford (1990) shows that if one interprets mixed strategies not as random behaviour but as the opponents' uncertainty regarding this behaviour, then the required convex-valuedness of the best response correspondence can be restored and existence of Nash equilibrium can be ensured.

In dynamic games, abandoning the independence axiom poses even more difficult problems. Without the independence axiom, conditional preferences at a given node of an extensive form game (or a decision-tree) depend on the unrealized payoffs earlier in the game. The literature has dealt with this problem in two ways: first, by assuming that the DM maximizes his conditional preference at each node (for a statement and defence of this approach, see Machina, 1989). This approach leads to dynamically consistent behaviour, since the DM ends up choosing the optimal strategy for the reduced (normal form)

game. However, it is difficult to compute optimal strategies once conditional preference depends on the entire history of unrealized outcomes. The second approach rejects dynamic consistency and assumes that at each node the DM maximizes his unconditional preference given his prediction of future behaviour. Thus, in the second approach, each node is treated as a distinct player and a subgame perfect equilibrium of the extensive form game is computed. Game-theoretic models that abandon the independence axiom have favoured the second approach. Such models have been used to study auctions.

1.2 Redefining payoffs: altruism and fairness

The next set of behavioural criticisms question common assumptions regarding deterministic outcomes. Consider the *ultimatum game*: Player 1 chooses some amount $x \leq 100$ to offer to Player 2. If Player 2 accepts the offer, 2 receives x and 1 receives $100-x$; If 2 rejects, both players receive 0. Suppose the rewards are measured in dollars and Player 1 has to make his offer in multiples of a dollar. It is easy to verify that if the players care only about their own financial outcome, there is no subgame perfect Nash equilibrium of this game in which Player 1 chooses $x>1$. Moreover, in every equilibrium, any offer $x>0$ must be accepted with probability 1. Contrary to these predictions, experimental evidence indicates that small offers are often rejected. Hence, subjects in the Player 2 role resent either the unfairness of the (99,1) outcome, or Player 1's lack of generosity. Moreover, many experimental subjects anticipate this response and make more generous offers to ensure acceptance. Even in the version of this game in which Player 2 does not have the opportunity to reject (that is, Player 1 is a *dictator*), Player 1 often acts altruistically and gives a significant share to Player 2.

More generally, there is empirical evidence that suggests that economic agents care not only about their physical outcomes but also about the outcomes of their opponents and how the two compare. Within game theory, this particular behavioural critique has been influential and has led to a significant theoretical literature on social preferences (see, for example, Fehr and Schmidt, 1999).

1.3 Redefining the objects of choice: ambiguity, timing of resolution of uncertainty, and preference for commitment

The next set of behavioural criticisms points out how the standard definition of outcome or consequence is inadequate. The literature on ambiguity questions probabilistic sophistication; that is, the idea that all uncertainty can be reduced to probability distributions. Ellsberg (1961) provides the original statement of this criticism. Consider the following choice problem: there are two urns; the first contains 50 red balls and 50 blue balls; the second contains 100 balls, each of which is either red or blue. The DM must select an urn and announce a colour. Then a ball will be drawn from the urn he selects. If the colour of the ball is the same as the colour the DM announces, he wins 100 dollars. Otherwise the DM gets zero. Experimental results indicate that many DMs are indifferent between (urn 1, red) and (urn 1, blue) but they strictly prefer either of these choices to (urn 2, red) and (urn 2, blue). If the DM were probabilistically sophisticated and assigned probability p to choosing a red ball from urn 1 and q to choosing a red ball from urn 2, the preferences above would indicate that $p=1-p$, $p>q$, and $p>1-q$, a contradiction. Hence, many DMs are not probabilistically sophisticated.

Ellsberg's experiment has lead to choice-theoretic models where agents are not probabilistically sophisticated and have an aversion to ambiguity; that is, the type of uncertainty associated with urn 2.

Recent contributions have investigated auctions with ambiguity-averse bidders and mechanism design with ambiguity aversion.

Other developments in behavioural choice theory that fall into this category have had limited impact on game-theoretic research. For example, Kreps and Porteus (1978) introduce the notion of a temporal lottery to analyse economic agents' preference over the timing of resolution of uncertainty. The Kreps–Porteus model has been extremely influential in dynamic choice theory and asset pricing but has had less impact in strategic analysis.

Kreps (1979) takes as his primitive individuals' preferences over sets of objects. Hence, an object similar to the indirect utility function of demand theory defines the individual. Kreps uses this framework to analyse preference for flexibility. So far, there has been limited analysis of preference for flexibility in strategic problems.

Gul and Pesendorfer (2001) use preferences over sets to analyse agents who have a preference for commitment (an alternative approach to preference for commitment is discussed in Section (3.2)). The GP model has been used to analyse some mechanism design problems.

2 Limitations of the decision-maker

The work discussed in Section 1 explores alternative formulations of economic consequences to identify preference-relevant considerations that are ignored in standard economic analysis. The work discussed in this section provides a more fundamental challenge to standard economics. This research seeks alternatives to common assumptions regarding economic agents' understanding of their environments and their cognative/computational abilities.

2.1 Biases and heuristics

Many economic models are stated in subjectivist language. Hence probabilities, whether they represent the likelihood of future events or the individual's own ignorance of past events, are the DMs' personal beliefs rather than objective frequencies. Similarly, the DM's utility function is a description of his behaviour in a variety of contingencies rather than an assessment of the intrinsic value of the possible outcomes. Nevertheless, when economists use these models to analyse particular problems, the subjective probabilities (and sometimes other parameters) are often calibrated or estimated by measuring objective frequencies (or other objective variables).

Psychology and economics research has questioned the validity of this approach. Tversky and Kahneman (1974) identify systematic biases in how individuals make choices under uncertainty. This research has led to an extensive literature on heuristics and biases. Consider the following:

1. (a) Which number is larger $P(A|B)$ or $P(A \cap C|B)$? Clearly, $P(A|B)$ is the larger quantity; conditional on *B* or unconditionally, $A \cap C$ can never be more likely than *A*. Yet, when belonging to set *C* is considered 'typical' for a member of *B*, many subjects state that $A \cap C$ conditional on *B* is more likely than *A* conditional on *B*.
2. (b) Randomly selected subjects are tested for a particular condition. In the population, 95 per cent are healthy. The test is 90 per cent accurate; that is, a healthy subject tests negative and a subject

having the condition tests positive with probability 0.9. If a randomly chosen person tests positive, what is the probability that he is ill? In such problems, subjects tend to ignore the low prior probability of having the condition and come up with larger estimates than the correct answer (less than one-third in this example).

Eyster and Rabin's (2005) analysis of auctions offers an example of a strategic model of biased decision-making. This work focuses on DMs' tendency to overemphasize their own (private) information at the expense of the information that is revealed through the strategic interaction.

2.2 Evolution and learning

As in decision theory, it is possible to state nearly all the assumptions of game theory in subjectivist language (see, for example, Aumann and Brandenburger, 1995). Hence, one can define Nash equilibrium as a property of players' beliefs. Of course, Nash equilibrium beliefs (together with utility maximization) will impose restrictions on observable behaviour, but these restrictions will fall short of demanding that the observed frequency of actions profiles constitute a Nash equilibrium. The theory of evolutionary games searches for dynamic mechanisms that lead to equilibrium behaviour, where equilibrium is identified with observable decisions (as opposed to beliefs) of individuals. The objective is to describe how equilibrium may emerge and which equilibria are more likely to emerge through repeated interaction in a setting where the typical epistemic assumptions of equilibrium analysis fail initially. Thus, such models are used both to justify Nash (or weaker) equilibrium notions and to justify refinements of these notions.

2.3 Cognitive limitations and game theory

Some game theoretic solution concepts require iterative procedures. For example, computing rationalizable outcomes in normal form games or finding backward induction solutions in extensive form games involves an iterative procedure that yields a smaller game after each step. The process ends when the final game, which consists exclusively of actions that constitute the desired solution, is reached. In principle, the number of steps needed to reach the solution can be arbitrarily large. Ho, Camerer and Weigelt (1998) observe that experimental subjects appear to carry out at most the first two steps of these procedures.
This line of work focuses both on organizing observed violations of standard game theoretic solutions concepts and interpreting the empirical regularities as the foundation of a behavioural notion of equilibrium.

3 Alternative models of the individual

The work discussed in this section poses the most fundamental challenge to the standard economic model of the individual. This work questions the usefulness of constrained maximization as a framework of economic analysis, or at least argues for a fundamentally different set of constraints.

3.1 Prospect theory and framing effects

Consider the following pair of choices (Tversky and Kahneman, 1981): an unusual disease is expected to kill 600 people. Two alternative programmes to combat the disease have been proposed.
Programme A will save 200; with Programme B, there is a one-third probability that 600 people will be saved, and a two-thirds probability that no one will be saved.
Next, consider the following restatement of what would appear to be the same options:
If Programme C is adopted 400 people will die; with Programme D, there is a one-third probability that nobody will die, and a two-thirds probability that 600 people will die.
Among subjects given a choice between A and B, most choose the safe option A, while the majority of the subjects facing the second pair of choices choose the risky option D.
Kahneman and Tversky's (1979) prospect theory combines issues discussed in Sections (1.1) and (2.2), with a more general critique of standard economic models, or at least of how such models are used in practice. Thus, while a standard model might favour a level of abstraction that ignores the framing issue above, Kahneman and Tversky (1979) argue that identifying the particular frame that the individual is likely to confront should be central to decision theory. In particular, these authors focus on the differential treatment of gains and losses. Prospect theory defines preferences not over lotteries of terminal wealth but over gains and losses, measured as differences from a status quo. In applications, the status quo is identified in a variety of ways.
For example, Köszegi and Rabin (2005–6) provide a theory of the status quo and utilize the resulting model to study a monopoly problem. In their theory, the DM's optimal choice becomes the status quo. Thus, the simplest form of the Köszegi–Rabin model defines optimal choices from a set A as $C(A) = \{x \in A | U(x, x) \geq U(y, x)\ \forall y \in A\}$. Hence, $x \in A$ is deemed to be a possible choice from A if the DM who views x as his reference point does not strictly prefer some other alternative y.
The three lines of work discussed below all represent a fundamental departure from the standard modelling of economic decisions: they describe behaviour as the outcome of a game even in a single person problem.

3.2 Preference reversals

Strotz (1955–6) introduces the idea of dynamic inconsistency: the possibility that a DM may prefer to consume x in period 2 to consuming y in period 1, if he makes the choice in period 0, but may have the opposite preference if he makes the choice in period 1. Strotz suggests that the appropriate way to model dynamically inconsistent behaviour is to assume that the period 0 individual treats his period 1 preference (and the implied behaviour) as a constraint on what he can achieve. Thus, suppose the period 0 DM has a choice between committing to z for period 2 consumption, or rejecting z and giving his period 1 self the choice between x in period 2 and y in period 1. Suppose also that the period 0 self prefers x to z and z to y while the period 1 self prefers y to x. Then, the Strotz model would imply that the DM ends up consuming z in period 2: the period 0 self realizes that if he does not commit to z, his period 1 self will choose y over x, which, for the period 0 self, is the least desirable outcome. Therefore, the period 0 self will commit to z. Hence, dynamic inconsistency leads to a preference for commitment.
Peleg and Yaari (1973) propose to reconcile the conflict among the different selves of a dynamically

inconsistent DM with a strategic equilibrium concept. Their reformulation of Strotz's notion of consistent planning has facilitated the application of Strotz's ideas to more general settings, including dynamic games.

3.3 Imperfect recall

An explicit statement of the perfect recall assumption and analysis of its consequences (Kuhn, 1953) is one of the earliest contributions of extensive form game theory. In contrast, the analysis of forgetfulness, that is, extensive form games where the individual forgets his own past actions or information, is relatively recent (Piccione and Rubinstein, 1997).

Piccione and Rubinstein observe that defining optimal behaviour for players with imperfect recall is problematic and propose a few alternative definitions (1997). Subsequent work has focused on what they call the multi-selves approach. In the multi-selves approach to imperfect recall, as in dynamic inconsistency, each information set is treated as a separate player. Optimal behaviour is a profile of behavioural strategies and beliefs at information sets such that the beliefs are consistent with the strategy profile and each behavioural strategy maximizes the corresponding agent's payoff given his beliefs and the behaviour of the remaining agents. Hence, the multi-selves approach leads to a prediction of behaviour that is analogous to perfect Bayesian equilibrium.

3.4 Psychological games

Harsanyi (1967–8) introduces the notion of a type to facilitate analysis of the interaction of players' information in strategic problems. He argues that the notion of a type is flexible enough to accommodate all uncertainty and asymmetric information that is relevant in games. Geanakoplos, Pearce and Stacchetti (1989) observe that if payoffs are 'intrinsically' dependent on beliefs and beliefs are determined in equilibrium, then types cannot be defined independently of the particular equilibrium outcome. Their notion of a psychological game and type (for psychological games) allows for this interdependence between equilibrium expectations and payoffs.

Gul and Pesendorfer (2006) offer an alternative framework for dealing with interdependent preferences. In their analysis, players care not only about the physical consequences of their actions on their opponents, but also about their opponents' attitudes towards such consequences, and their opponents' attitudes towards others' attitudes towards such consequences, and so on. Gul and Pesendorfer provide a model of interdependent preference types similar to Harsanyi's interdependent belief types to analyse situations in which preference interdependence may arise not from the interaction of (subjective) information but from the interaction of the individuals' attitudes towards the well-being of others.

3.5 Neuroeconomics

The most comprehensive challenge to the standard economic modelling of the individual comes from research in neuroeconomics. Neuroeconomists argue that no matter how much the standard conventions are expanded to accommodate behavioural phenomena, it will not be enough: understanding economic behaviour requires studying the physiological, and in particular, neurological mechanisms behind

choice. Recent experiments relate choice-theoretic variables to levels of brain activity, the type of choices to the parts of the brain that are engaged when making these choices, and hormone levels to behaviour (Camerer, 2006) provide a concise summary of recent research in neuroeconomics). Neuroeconomists contend that ‘neuroscience findings raise questions about the usefulness of some of the most common constructs that economists commonly use, such as risk aversion, time preference, and altruism’ (Camerer, Loewenstein and Prelec, 2005). They argue that neuroscience evidence can be used directly to falsify or validate specific hypotheses about behaviour. Moreover, they claim that organizing choice theory and game theory around the abstractions of neuroscience will lead to better theories. Thus, neureconomics proposes to change both the language of game theory and what constitutes its evidence.

4 Conclusion

The interaction of behavioural economics and game theory has had two significant effects: first, it has broadened the subject matter and set of acceptable approaches to strategic analysis. New modelling techniques such as equilibrium notions that explicitly address biases have become acceptable and new questions such as the effect of ambiguity aversion in auctions have gained interest. More importantly, behavioural approaches have altered the set of empirical benchmarks – the stylized facts – that game theorists must address as they interpret their own conclusions.

See Also

- Allais paradox
- altruism in experiments
- ambiguity and ambiguity aversion
- learning and evolution in games: an overview
- prospect theory
- preference reversals
- neuroeconomics.

Bibliography

Allais, M. 1953. Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’ecole Americaine. *Econometrica* 21, 503–46.

Aumann, R.J. and Brandenburger, A. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63, 1161–80.

Camerer, C.F., Loewenstein, G. and Prelec, D. 2005. Neuroeconomics: how neuroscience can inform economics. *Journal of Economic Literature* 43, 9–64.

Camerer, C.F. 2007. Neuroeconomics: using neuroscience to make economic predictions. *Economic*

Journal.

Crawford, V.P. 1990. Equilibrium without independence. *Journal of Economic Theory* 50, 127–54.

Ellsberg, D. 1961. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Eyster, E. and Rabin, M. 2005. Cursed equilibrium. *Econometrica* 73, 1623–72.

Fehr, E. and Schimdt, K. 1999. A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114, 817–68.

Geanakoplos, J., Pearce, D. and Stacchetti, E. 1989. Psychological games and sequential rationality. *Games and Economic Behavior* 1, 60–80.

Gul, F. and Pesendorfer, W. 2001. Temptation and self-control. *Econometrica* 2001, 1403–35.

Gul, F. and Pesendorfer, W. 2006. The canonical type space for interdependent preferences. Working paper. Princeton University.

Harsanyi, J. 1967–8. Games with incomplete information played by Bayesian players. *Management Science* 14, 159–82, 320–34, 486–502.

Ho, T., Camerer, C. and Weingelt, K. 1998. Iterated dominance and iterated best responses in p-beauty contests. *American Economic Review* 88, 947–69.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–92.

Köszegi, B. and Rabin, M. 2006. A model of reference-dependent preferences, *Quarterly Journal of Economics* 121, 1133–66.

Kreps, D.M. 1979. A preference for flexibility. *Econometrica* 47, 565–76.

Kreps, D.M. and Porteus, E.L. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46, 185–200.

Kuhn, H.W. 1953. Extensive games and the problem of information. In *Contributions to the Theory of Games*, vol. 2. Princeton, NJ: Princeton University Press.

Machina, M. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 1622–68.

O'Donoghue, T. and Rabin, M. 1999. Doing it now or later. *American Economic Review* 103–24.

Peleg, B. and Yaari, M.E. 1973. On the existence of a consistent course of action when tastes are changing. *Review of Economic Studies* 40, 391–401.

Piccione, M. and Rubinstein, A. 1997. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior* 20, 3–24.

Strotz, R.H. 1955–6. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23, 165–80.

Tversky, A. and Kahnemann, D. 1974. Judgment under uncertainty: Heuristic and biases. *Science* 85, 1124–31.

Tversky, A. and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science* 211, 453–8.

How to cite this article

Gul, Faruk. "behavioural economics and game theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000210> doi:10.1057/9780230226203.0115

The New Palgrave Dictionary of Economics Online

behavioural finance

Robert Bloomfield

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Behavioural finance began as an attempt to understand why financial markets react inefficiently to public information. One stream of behavioural finance examines how psychological forces induce traders and managers to make suboptimal decisions, and how these decisions affect market behaviour. Another stream examines how economic forces might keep rational traders from exploiting apparent opportunities for profit. Behavioural finance remains controversial, but will become more widely accepted if it can predict deviations from traditional financial models without relying on too many ad hoc assumptions, and expand to settings (particularly corporate finance) in which arbitrage forces are weaker.

Keywords

accruals anomaly; anomalies; arbitrage; behavioural finance; book-to-market effect; capital asset pricing model; efficient markets hypothesis; equity premium puzzle; gambler's fallacy; home bias puzzle; hot-hand fallacy; incomplete revelation hypothesis; limited attention; market microstructure; momentum; miscalibration; mispricing; overconfidence; pattern recognition; post-earnings-announcement drift; prospect theory; reversal; risk; size effect

Article

Mounting evidence suggests that a variety of trading strategies generate returns that are larger than permitted by the reigning theory of efficient financial markets. Defenders of efficient markets theory argue that the anomalies represent methodological errors, and in many cases they appear to have been correct. In cases where the anomalies appear robust, the debates turn to two other questions. First, why would investors make systematic trading errors that could result in mispricing? Second, why wouldn't smarter traders exploit those errors, thereby driving prices to appropriate levels? Many answers to the first question have relied heavily on the branch of psychology called ‘behavioural decision theory’,

which has led to the entire body of research being dubbed ‘behavioural finance’ even though there is rarely much behavioural content in the literatures identifying pricing anomalies and explaining why price errors are not eliminated by smarter traders.

The next section of this article discusses the empirical evidence that market prices deviate from levels that would reflect perfectly rational traders acting in competitive markets (the ‘anomalies’ literature). I then discuss literatures that document how behavioural forces can explain these anomalies, and that examine why irrational traders might influence prices in competitive markets. I conclude by suggesting some promising future directions in behavioural finance.

Anomalies

In 1968, two accounting professors reported that markets react sharply to earnings announcements over the course of a few days, and then continue drifting in the same direction for the better part of a year (Ball and Brown, 1968). This post-earnings-announcement drift (PEAD) appeared to provide an easy opportunity for making money: one could create a hedged portfolio that is long in firms that have just announced good news and short in firms that have just announced bad news, so that it earns positive returns from no net investment.

The fact that prices react at all to earnings was surprising enough, given that earnings was then viewed as an accounting fiction describing past events, with no bearing on the future cash flows of the firm that should entirely determine firm value. (Accounting ‘fictions’ like earnings and book value are now known to provide important information about future cash flows, spawning a large field of financial accounting research.) But the subsequent drift was even more surprising, as it flew in the face of the recently developed efficient markets hypothesis (EMH), subsequently codified by Gene Fama (1970). The EMH relies on competition among investors to assert that strategies based on public information cannot earn returns after adjusting for risk. If all investors know that holding the PEAD portfolio would allow for excess returns, they would compete to hold the portfolio, and drive prices to the level needed to eliminate those returns.

PEAD has turned out to be one of the first – and most robust – of a large number of market anomalies. Initial explanations for PEAD were that the predictable returns simply reflect the expected returns that investors demand to compensate for the risk the PEAD portfolio would impose on them. Such arguments were made much more difficult by Bernard and Thomas (1990), who showed that about half the returns to the PEAD portfolio were experienced in the three-day windows surrounding the two subsequent earnings announcements. Thus, any risk-based explanation would require firms with extremely good or bad earnings news to experience dramatic changes in systematic risk for only a few days a year, several months in the future. The alternative explanation, proffered by Bernard and Thomas, was that investors simply did not understand the implications of current earnings for future earnings – an assertion that has been repeatedly supported by studies of analysts’ earnings estimates and laboratory experiments. Researchers were successful enough in ruling out the risk explanation, and in tying future returns to the information content of current earnings, so that Fama (1998, p. 304) concluded that PEAD ‘has survived robustness checks’, and was possibly ‘above suspicion’.

Three other robust anomalies seem more likely to reflect compensation for risk than mispricing: the book-to-market effect, the size effect and the momentum effect. The book-to-market ratio is the ratio of

a firm's net assets (as reported on the firm's balance sheet) to the total market value of the firm's outstanding stock. Firms with low book-to-market ratios earn substantially higher returns than those with high book-to-market ratios (the book-to-market effect), as if the market value reverts over time to the value indicated by the accounting statements. Firms with small market capitalization earn higher returns than firms with large market capitalization (the size effect), as if small firms are consistently underpriced. Stocks that move strongly upwards or downwards over a three- to six-month period are very likely to continue moving in that direction over a subsequent three to six months (the momentum effect), as if the market responds slowly to changes in value.

Distinguishing risk and mispricing is difficult for book-to-market and size and momentum effects because researchers have no hypothesis that the mispricing will be corrected at some particular moment. (In contrast, the theory explaining PEAD suggests that mispricing will be revealed and corrected upon subsequent earnings announcements). Proponents of efficient markets have provided evidence that book-to-market and size capture systematic risk, and have expanded the traditional asset pricing model to include book-to-market, size and (less frequently) momentum as risk factors. However, analysts appear to view book-to-market as an indicator of mispricing rather than risk, as indicated by examinations of analyst reports and controlled experiments.

Researchers in finance and accounting have identified a host of other pricing anomalies. Here is a selective sampling of some of the most well known, all of which remain controversial:

- *Long-term price reversal*. Stocks that move strongly over a three- to five-year period are very likely to reverse a portion of those movements over a following three- to five-year period (DeBondt and Thaler, 1985). Evidence for long-term reversal tends to be more controversial than evidence for short-term momentum, because longer horizons make it harder to guarantee appropriate computation of risk-adjusted returns.
- *The equity premium puzzle*. A diversified portfolio of equity securities should earn higher returns than a portfolio of bonds, because of the additional risk equities impose on investors. However, the equity premium appears far too large relative to the associated risk (Mehra and Prescott, 1985).
- *The home bias puzzle*. Both institutional and individual investors tend to hold a disproportionate amount of their portfolios in firms based in their own countries and regions. This may reflect a bias to purchase familiar stocks (Huberman, 2001), or the inside information held by local investors (Coval and Moskowitz, 2001).
- *Excessive volatility and excessive volume*. Shiller (1981) has argued that market prices are excessively volatile, relative to the volatility of fundamentals. Many others, including Kandel and Pearson (1995), have argued that trade volume is far too high to be explained by traditional theory, in light of the Milgrom and Stokey (1982) ‘no-trade theorem’, which proves that, in the absence of non-informational motivations for trade, such as a need for liquidity or sharing of risk, markets should not include any trade.
- *The accruals anomaly*. Firms’ earnings can be decomposed into cash flows and accruals (defined as earnings minus cash flows). Sloan (1996) showed that firms with large positive accruals earn lower future returns than firms with large negative accruals, as if investors are unaware that accruals – which do not represent cash flows and are easily manipulated by managers – reverse

rapidly.

Individual behaviour

The variety of market anomalies has led some to doubt the validity of the EMH, but few researchers are likely to let go of the efficient markets perspective without a coherent and parsimonious theory of when to predict which types of anomalies. One branch of psychology, called ‘behavioural decision theory’ (BDT), appears particularly well-suited to imposing regular structure on otherwise ad hoc results. BDT researchers have shown that a variety of apparently irrational behaviours can be explained by a relatively parsimonious set of theories. For their part, behavioural finance researchers have sought to use empirical and experimental studies to show that behavioural theories can describe the actions of individual investors (as well as managers), and to use theoretical methods to show that a small set of behavioural theories can account for the wide variety of market anomalies. Four streams of results feature most prominently in behavioural finance: prospect theory, miscalibration, pattern recognition and limited attention.

Prospect theory

Throughout the 1970s, Amos Tversky and Daniel Kahneman published a series of papers characterizing how people value outcomes. This research ultimately resulted in a mathematical representation of subjective (hedonic) value called ‘prospect theory’ (Kahneman and Tversky, 1979), for which Kahneman won the 2002 Nobel Prize in economics (Amos Tversky died in 1996). Prospect theory emphasizes three features of the value function: that the hedonic value of an outcome is determined by whether the outcome is a gain or loss relative to the agent's reference point; that the negative hedonic value of a loss more than offsets the positive hedonic value of a gain of the same size; and that the marginal effect of increasing a gain (or loss) is decreasing in the size of the gain (or loss).

Prospect theory yields a variety of predictions that describe individual behaviour well, and that can also account for several market anomalies. Prospect theory helps to explain a common behaviour termed the ‘disposition effect’ (Shefrin and Statman, 1985) – traders will close out profitable investments quickly, to lock in gains, while holding on to their losing investments or perhaps even invest more in them, in hopes that the investment will turn around. Let us assume that a trader has bought a stock at 50 dollars, and that it is now priced at 80 dollars. Using the 50-dollar purchase price as a reference point, the trader has a 30-dollar gain, and (because the marginal effect of increasing a gain is decreasing in the size of the gain) the agent is risk-averse, and will want to close the position quickly to avoid risk. If the price fell to 20 dollars, however, the trader has a 30-dollar loss, and (because the marginal effect of increasing a loss is decreasing in the size of the loss) the agent is risk seeking, and will want to keep the position open to take on more risk.

Terry Odean (1998a) has shown clear evidence of the disposition effect among thousands of individual investors at a brokerage firm. Unfortunately for the investors, selling winners and holding on to losers is nearly the opposite of the profitable momentum strategy, which involves buying recent winners and selling recent losers. As a result, the stocks the investors held subsequently underperformed the stocks they sold. The disposition effect does not seem restricted to amateurs. Coval and Shumway (2005) show

that professional commodity traders who have net losses near the end of the day tend to trade quite aggressively until trading closes, and take on significant risk. Finally, Frazzini (2006) ties the disposition effect back to price anomalies by providing evidence that disposition effects drive short-term momentum, because the relatively rapid selling of winners slows reactions to good news, while the tendency to hold losers slows reactions to bad news.

The disposition effect is driven by the different curvatures of the value function in the loss and gain realms. Curvature is important when investors evaluate the risk of relatively small changes in wealth. Investors who evaluate the risk of large wealth changes are influenced instead by the different average slopes of the value function in the loss and gain realms. Because the average slope is flatter in the realm of gains, investors with large gains in hand are likely to appear less risk-averse than those with losses or small gains. Evidence from experiments (Thaler and Johnson, 1990) and game show contestants (Gertner, 1993) are consistent with this ‘house money’ effect, named after the exaggerated risk tolerance of the behaviour of gamblers who have won money from the house, and therefore are risking only the house's money. Barberis, Huang and Santos (2001) show that the house money effect can account for both short-term momentum and long-term reversal. Short-term momentum arises because traders demand more compensation for risk after price declines, further depressing prices, while demanding less compensation for risk after price increases, further inflating prices. Similar reasoning shows that the house money effect can account for the book-to-market effect and an exaggerated equity premium.

While prospect theory is a relatively parsimonious and powerful theory, its predictions are highly sensitive to assumptions about how people identify benchmarks against which to measure gains and losses, and under what circumstances they might evaluate gains and losses of portfolios, rather than of individual securities. The field of ‘mental accounting’ (Barberis, Huang and Thaler, 2006) addresses such questions.

Miscalibrated confidence

Financial models of trade traditionally assume that agents have confidence calibrated to reflect the precision of their information. Experiments show that people rarely satisfy this requirement. People tend to be overconfident in their ability to predict events when they have very poor information, while people who are asked easy questions tend to be underconfident. Psychologists call this tendency the ‘hard–easy’ effect (Griffin and Tversky, 1992); Bloomfield, Libby and Nelson (2000) call it ‘moderated confidence’ because confidence is moderated from the optimal level towards a prior belief of moderate data reliability, as if people are rational Bayesians with imperfect information about the reliability of their data.

Because financial outcomes are so hard to predict, people are likely to be overconfident, rather than under-confident. Indeed, evidence of overconfidence is widespread. Odean (1999) finds that individual investors trade far too frequently, apparently overconfident in their ability to identify mispriced securities. Malmendier and Tate (2005) find that many executives are overconfident in their firms’ futures (as evidenced by their failure to exercise stock options before expiration), and further show that more overconfident executives are more likely to engage in value-reducing mergers.

Theoretical and experimental research has shown that calibration errors can account for a variety of known anomalies. Gervais and Odean (2001) and Odean (1998b) examine how overconfidence can lead

to excessive trading. Daniel, Hirshleifer and Subrahmanyam (1998) show that overconfidence can account for both overreactions and underreactions to information. In a similar vein, Bloomfield, Libby and Nelson (2003) show that overconfident inferences from old earnings numbers, which have little information content once newer numbers are available, lead to both post-earnings-announcement drift and overreactions to earnings trends.

Pattern recognition

The human mind has a gift for finding order in chaos, even when objective analysis shows no order to be found. In such cases, people show remarkable consistency in the order they perceive. People fall prey to the gambler's fallacy when they expect that a coin that has come up ‘heads’ many times in a row is then more likely to come up ‘tails’ because such streaks are typically short-lived. People fall prey to the ‘hot-hand’ fallacy when they mistakenly believe that basketball players who have made ten free throws in a row are especially likely to make the next, even though this is not the case (a professional basketball player's free throw performance is not distinguishable from a random series with a constant mean). The tendency to see patterns in random sequences is likely to be particularly important in financial markets, where competitive pressures force market prices to follow a random walk (after risk premia are accounted for). Despite the randomness in stock movements, many investors subscribe to ‘technical analysis’ trading strategies (and expensive newsletters) based on elaborate patterns like ‘head and shoulders’ and ‘cup with handle’, even though systematic research has found little evidence that such patterns can predict future stock movements.

Barberis, Shleifer and Vishny (1998) claim that people who observe a random walk are likely to fluctuate between beliefs in the gambler's fallacy (in which any trends are quickly reversed) and beliefs in the hot hand (in which trends continue), depending on how many reversals in price they have seen in recent periods. They then prove that such beliefs can account for both short-term price momentum and long-term price reversal. Bloomfield and Hales (2002) find experimental support for that assumption.

Limited attention

A fundamental tenet of cognitive science is that people have limited cognitive resources, implying that their attention to financial information and investment opportunities may be determined by economically irrelevant factors such as how information is presented or how often it is talked about by others. Experiments have found that even experienced analysts draw conclusions that are coloured by seemingly irrelevant aspects of how financial information is presented (Hirst and Hopkins, 1998). Employees’ decisions on how to invest their defined contribution pension funds are dramatically influenced by how the options are presented (Benartzi and Thaler, 2001), while their decision to enrol in such plans at all are dramatically increased by a policy that makes investment the default option, so that enrolment requires no attention at all (Benartzi and Thaler, 2004).

Limited attention may determine how stocks come in and out of favour, and provides a natural explanation for the home bias puzzle – people naturally notice local firms more readily than distant firms. Limited attention may also explain the tendency of firms to attract attention (and trading volume) when their earnings are growing rapidly, but be ignored when they perform poorly for long periods. Lee

and Swaminathan (2000) argue that such tendencies might explain short-term momentum, and support their argument by showing that firms with low volume and strong returns show strong momentum in returns (as if they are underpriced while still neglected), while those with high volume and strong returns show long-term reversal (as if they are overpriced at the peak of attention).

Accounting researchers have been particularly interested in the effects of limited attention, because they may explain why people care so much about accounting regulations that alter only how information is presented, and not the information content of the complete accounting disclosure. A highly publicized example is the controversy over whether employee stock option costs should be deducted from reported earnings per share; in both cases, investors could gather all relevant information from the footnotes to the financial statements. Bloomfield (2002) argues that fewer investors attend to footnotes than to earnings, and that standard models of information aggregation predict that market prices less completely reveal information that is held by fewer investors – a result repeatedly confirmed in laboratory markets. This ‘incomplete revelation hypothesis’ runs counter to the EMH, which is typically applied to all public information regardless of how it is presented. However, accounting researchers have made considerable progress in understanding how different presentation options, such as the formating, isolation and ordering of text can alter investors’ attention to and weighting of the information in that text (see, for example, Maines and McDaniel, 2000).

Limits to arbitrage

Studies of individual behaviour show that investors and managers make systematic errors of judgement, but do not explain how other investors fail to exploit, and thereby eliminate, any aggregate mispricing. A number of studies have noted that arbitrage may be limited by risks that cannot be captured as risk factors in traditional asset pricing models. Even if a pricing error must eventually converge (as when two securities representing claims on the same underlying assets have different prices), such convergence may not be rapid, and may even be preceded by additional divergence. While asset pricing models like the capital asset pricing model (CAPM) conclude that such idiosyncratic risk does not affect price levels, Pontiff (2006) has argued forcefully that idiosyncratic risk still hinders the correction of price errors by effectively imposing a ‘holding cost’ on arbitrageurs. Idiosyncratic risk restricts arbitrage most severely when a trader uses borrowed capital to engage in arbitrage, because a short-term loss may result in a margin call, or may lead the investors to infer that the arbitrageur has a poor strategy, and therefore withdraw their funds (Shleifer and Vishny, 1997). DeLong et al. (1990) take these arguments one step further: they assume that the noise in returns is driven by irrational traders, and then show that these traders still earn sufficient returns for them to survive indefinitely.

Another line of literature notes that rational arbitrageurs might earn greater profits by exacerbating price errors rather than disciplining them. Abreu and Brunnermeier (2002) construct a model in which irrational traders drive prices too high, a fact that eventually becomes known to every arbitrageur. Because arbitrageurs do not know whether other arbitrageurs have yet learned of the overpricing, each one continues to ‘ride the bubble’ after they learn of the overpricing, rather than pop it, because they expect others to do so as well. As a result, the arbitrageurs continue magnifying the bubble even after each individual arbitrageur knows that prices are too high.

The preceding explanations of limited arbitrage are largely devoid of behavioural content – the price

errors that fail to be corrected could arise from any cause, including completely random trading. However, researchers do occasionally examine how specific biases can limit arbitrage opportunities. Overconfidence, in particular, has been shown to be difficult to arbitrage. For example, Kyle and Wang (1997) show that overconfident traders can effectively gain ‘elbow room’ in a market, just as a trader in a Cournot oligopoly game can benefit by committing to aggressive production, and forcing others to produce less. As a result, overconfident traders earn enough trading gains to persist.

Conclusion and future directions

This history of behavioural finance fits well within Kuhn's (1962) narrative of scientific revolution. Early researchers uncovered results that were anomalous within the paradigm of efficient markets; as they became convinced that the anomalies were not simply the result of methodological error, researchers sought a new paradigm that could encompass the anomalies, as well as the predictions of the traditional theory. This new paradigm assumes that markets include some participants who optimize their expected utility, along with others whose susceptibility to psychological forces leads them to behave suboptimally.

No behavioural alternative will ever rival the coherence, parsimony and power of traditional efficient markets theory, because psychological forces are too complex. Thus, behavioural researchers in finance must devote themselves to the ‘normal science’ suggested by their new paradigm: documenting and refining our understanding of how psychological forces influence individual behaviour in financial settings, and how those behaviours affect market phenomena. This will require much more attention to behavioural psychology than is evident in the existing body of research. (As of 2007, few papers in behavioural finance rely on psychological research published after the 1970s.) Perhaps more importantly, advances in behavioural finance will require more attention to the details of market microstructure, which influence individual behaviour, and how those behaviours affect market-level phenomena. Finally, researchers in behavioural finance can expand their scope beyond describing the behaviour of investors and prices in highly competitive asset markets. Behavioural theories are likely to have greater ability to explain phenomena in settings that provide fewer opportunities for others to exploit (and thereby eliminate) suboptimal outcomes. For example, decisions on how to hire and compensate executives, and on when and how to raise and invest capital, seem particularly susceptible to behavioural analysis (as in Shefrin, 2005).

See Also

- arbitrage
- behavioural economics and game theory
- bubbles
- efficient markets hypothesis
- prospect theory

Bibliography

Abreu, D. and Brunnermeier, M.K. 2002. Synchronization risk and delayed arbitrage. *Journal of Financial Economics* 66, 341–60.

Ball, R. and Brown, P. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6, 159–78.

Barberis, N., Huang, M. and Thaler, R.H. 2006. Individual preferences, monetary gambles and stock market participation: a case for narrow framing. *American Economic Review* 96, 1069–90.

Barberis, N., Shleifer, A. and Vishny, R. 1998. A model of investor sentiment. *Journal of Financial Economics* 49, 307–43.

Barberis, N., Huang, M. and Santos, T. 2001. Prospect theory and asset prices. *Quarterly Journal of Economics* 116, 1–53.

Benartzi, S. and Thaler, R.H. 2001. Naïve diversification strategies in defined contribution savings plans. *American Economic Review* 91, 79–98.

Benartzi, S. and Thaler, R.H. 2004. Save more tomorrow™: using behavioral economics to increase employee saving. *Journal of Political Economy* 112, S164–S187.

Bernard, V.L. and Thomas, J. 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics* 13, 305–41.

Bloomfield, R. and Hales, J. 2002. Predicting the next step of a random walk: experimental evidence of regime-shifting beliefs. *Journal of Financial Economics* 65, 397–415.

Bloomfield, R. 2002. The incomplete revelation hypothesis: implications for financial reporting. *Accounting Horizons* 16, 233–44.

Bloomfield, R., Libby, R. and Nelson, M.W. 2003. Over-reliance on previous years' earnings. *Contemporary Accounting Research* 20, 1–31.

Bloomfield, R., Libby, R. and Nelson, M.W. 2000. Underreactions, overreactions and moderated confidence. *Journal of Financial Markets* 3, 113–37.

Coval, J. and Moskowitz, T. 2001. The geography of investment: informed trading and asset prices. *Journal of Political Economy* 109, 811–41.

Coval, J.D. and Shumway, T. 2005. Do behavioral biases affect prices? *Journal of Finance* 60, 1–34.

Daniel, K., Hirshleifer, D. and Subrahmanyam, A. 1998. Investor psychology and security market under- and overreaction. *Journal of Finance* 53, 1839–86.

DeBondt, W.F.M. and Thaler, R.H. 1985. Does the stock market overreact? *Journal of Finance* 40, 793–807.

DeLong, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98, 703–38.

Fama, E.F. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25, 383–417.

Fama, E.F. 1998. Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* 49, 283–306.

Frazzini, A. 2006. The disposition effect and underreaction to news. *Journal of Finance* 61, 2017–46.

Gertner, R. 1993. Game shows and economic behavior: risk taking on ‘card sharks’. *Quarterly Journal of Economics* 151, 507–21.

Gervais, S. and Odean, T. 2001. Learning to be overconfident. *Review of Financial Studies* 14, 1–27.

Griffin, D. and Tversky, A. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psychology* 24, 411–35.

Hirst, D.E. and Hopkins, P.E. 1998. Comprehensive income reporting and analysts’ valuation judgments. *Journal of Accounting Research* 36(Supplement), 47–75.

Huberman, G. 2001. Familiarity breeds investment. *Review of Financial Studies* 14, 659–80.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–92.

Kandel, E. and Pearson, N.D. 1995. Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy* 103, 831–72.

Kuhn, T.S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kyle, A. and Wang, F.A. 1997. Speculation duopoly with agreement to disagree: can overconfidence survive the market test? *Journal of Finance* 52, 2073–90.

Lee, C.M.C. and Swaminathan, B. 2000. Price momentum and trading volume. *Journal of Finance* 55, 2017–33.

Maines, L.A. and McDaniel, L.S. 2000. Effects of comprehensive-income characteristics on nonprofessional investors' judgments: the role of financial-statement presentation format. *Accounting Review* 75, 179–207.

Malmendier, U. and Tate, G. 2005. CEO overconfidence and corporate investment. *Journal of Finance* 60, 2661.

Mehra, R. and Prescott, E.C. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Milgrom, P. and Stokey, N. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26, 17–27.

Odean, T. 1998a. Are investors reluctant to realize their losses? *Journal of Finance* 53, 1775–98.

Odean, T. 1998b. Volume, volatility, price, and profit when all traders are above average. *Journal of Finance* 53, 1887–934.

Odean, T. 1999. Do investors trade too much? *American Economic Review* 89, 1279–98.

Pontiff, J. 2006. Costly arbitrage and the myth of idiosyncratic risk. *Journal of Accounting and Economics* 42, 35–52.

Shefrin, H. 2005. *Behavioral Corporate Finance*. New York: McGraw-Hill/Irwin.

Shefrin, H. and Statman, M. 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance* 40, 777–90.

Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71, 421–36.

Shleifer, A. and Vishny, R.W. 1997. The limits of arbitrage. *Journal of Finance* 52, 35–55.

Sloan, R. 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review* 71, 289–315.

Thaler, R.H. and Johnson, E.J. 1990. Gambling with the house money and trying to break even: the effects of prior outcomes on risky choice. *Management Science* 36, 643–60.

How to cite this article

Bloomfield, Robert. "behavioural finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000339> doi:10.1057/9780230226203.0116

The New Palgrave Dictionary of Economics Online

behavioural game theory

Colin F. Camerer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Behavioural game theory uses experimental regularities and psychology to model formally how limits on strategic thinking, learning, and social preferences interact when people actually play games. Emerging theories of behaviour in ultimatum and trust games (and others) focus on an aversion to inequality, reciprocity, or concern for social image. Learning models often focus on numerical updating of an unobserved propensity to choose a strategy (including fictitious play updating of beliefs as a special case). Models of limits on strategic thinking assume players are in equilibrium, but respond with error, or there is a cognitive hierarchy of increasingly sophisticated reasoning.

Keywords

adaptive expectations; altruism; analytical game theory; auctions; behavioural game theory; bounded rationality; cognitive hierarchy theories; communication; competition; contracting; coordination; dictator games; direction learning; disequilibrium behaviour; entry deterrence games; experience weighted attraction; individual learning in games; inequality aversion; learning; mixed strategy equilibrium; moral hazard; negative reciprocity; quantal response equilibrium; rational expectations; Rawls, J.; reciprocity; replicator dynamics; self-interest; signalling; social preferences; tit for tat; trust games; ultimatum games; utilitarianism; weighted fictitious play

Article

Analytical game theory assumes that players choose strategies which maximize the utility of game outcomes, based on their beliefs about what others players will do, given the economic structure of the game and history; in equilibrium, these beliefs are correct. Analytical game theory is enormously powerful, but it has two shortcomings as a complete model of behaviour by people (and other possible players, including non-human animals and organizations).

First, in complex naturally occurring games, equilibration of beliefs is unlikely to occur instantaneously. Models of choice under bounded rationality, predicting initial choices and equilibration with experience, are therefore useful.

Second, in empirical work, only received (or anticipated) payoffs are easily measured (for example, prices and valuations in auctions, or currency paid in an experiment). Since games are played over *utilities* for received payoffs, it is therefore necessary to have a theory of social preferences – that is, how measured payoffs determine players' utility evaluations – in order to make predictions.

The importance of understanding bounded rationality, equilibration and social preferences is provided by hundreds of experiments showing conditions under which predictions of analytical game theory are sometimes approximately satisfied, and sometimes badly rejected (Camerer, 2003). This article describes an emerging approach called 'behavioural game theory', which generalizes analytical game theory to explain experimentally observed violations. Behavioural game theory incorporates bounds on rationality, equilibrating forces, and theories of social preference, while retaining the mathematical formalism and generality across different games that has made analytical game theory so useful. While behavioural game theory is influenced by laboratory regularities, it is ultimately aimed at a broad range of applied questions such as worker reactions to employment terms, evolution of market institutions, design of auctions and contracts, animal behaviour, and differences in game-playing skill.

Social preferences

Let us start with a discussion of how preferences over outcomes of game can depart from pure material self-interest. In an ultimatum game a Proposer is endowed with a known sum, say ten dollars, and offers a share to another player, the Responder. If the Responder rejects the offer they both get nothing. The ultimatum game is a building block of more complex natural bargaining and a simple tool to measure numerically the price that Responders will pay to punish self-servingly unfair treatment.

Empirically, a large fraction of subjects rejects low offers of 20 per cent or so. Proposers fear these rejections reasonably accurately, and make offers around 40 per cent rather than very small offers predicted by perceived self-interest. (The earliest approximations of whether Proposers offer expected profit-maximizing offers, by Roth et al. 1991, suggested they did. However, those estimates were limited by the method of presenting Responders only with specific offers; since low offers are rare, it is hard to estimate the rejection rate of low offers accurately and hence hard to know conclusively whether offers are profit-maximizing. Different methods, and cross-population data used in Henrich et al., 2005, established that offers are too generous, even controlling for risk-aversion of the Proposers.) This basic pattern scales up to much higher stakes (the equivalent of months of wages) and does not change much when the experiment is repeated, so it is implausible to argue that subjects who reject offers (often highly intelligent college students) are confused.

It is crucial to note that rejecting two dollars out of ten dollars is a rejection of the *joint* hypothesis of utility-maximization and the auxiliary hypothesis that player *i*'s utility depends on only her own payoff x_i. An obvious place to repair the theory is to create a parsimonious theory of social preferences over (x_i, x_j) (and possibly of other features of the game) which predicts violations of self-interest across games with different structures. I will next mention some other empirical regularities, then turn to a discussion of such models of these regularities.

In ultimatum games, it appears that norms and judgements of fairness can depend on context and culture. For example, when Proposers earn the right to make the offer (rather than respond to an offer) by winning at a pre-play trivia game, they feel entitled to offer less – and Responders seem to accept less (Hoffman et al., 1994). Two comparative studies of small-scale societies show interesting variation across cultures. Subjects in a small Peruvian agricultural group, the Machiguenga, offer much less than those in other cultures (typically 15–25 per cent) and accept low offers. Across 15 societies, equality of average offers is positively related to the degree of cooperation in economic activity (for example, do men hunt collectively?) and to the degree of impersonal market trading (Henrich et al., 2005).

Ultimatum games tap negative reciprocity or vengeance. Other games suggest different psychological motives which correspond to different aspects of social preferences. In dictator games, a Proposer simply dictates an allocation of money and the Responder must accept it. In these games, Proposers offer less than in ultimatum games (about 15 per cent of the stakes on average), but offers vary widely with contextual labels and other variables (Camerer, 2003, ch. 2). In trust games, an Investor risks some of her endowment of money, which is increased by the experimenter (representing a return on social investment) and given to an anonymous Trustee. The Trustee pays back as much of the increased sum as she likes to the Investor (perhaps nothing) and keeps the rest. Trust games are models of opportunities to gain from investment with no legal protection against moral hazard by a business partner. Self-interested Trustees will never pay back money; self-interested Investors with equilibrium beliefs will anticipate this and invest nothing. In fact, Investors typically risk about half their money, and Trustees pay back slightly less than was risked (Camerer, 2003, ch. 2). Investments reflect expectations of repayment, along with altruism toward Investors (Ashraf, Bohnet and Piankov, 2006) and an aversion to ‘betrayal’ (Bohnet and Zeckhauser, 2004). Trustee payback is consistent with positive reciprocity, or a moral obligation to repay a player who risked money to benefit the group.

Importantly, competition has a strong effect in these games. If two or more Proposers make offers in an ultimatum game, and a single Responder accepts the highest offer, then the only equilibrium is for the Proposers to offer almost all the money to the Responder (the *opposite* of the prediction with one Proposer). In the laboratory this Proposer competition occurs rapidly, resulting in a very unfair allocation – almost no earnings for Proposers (for example, Camerer and Fehr, 2006). Similarly, when there is competition among Responders, at least one Responder accepts low offers and Proposers seem to anticipate this effect and offer much less. These regularities help explain an apparent paradox, why the competitive model based on self-interest works so well in explaining market prices in experiments with three or more traders on each side of the market. In these markets, traders with social preferences cannot make choices which reveal a trade-off of self-interest and concern for fairness. The parsimonious theory in which agents have social preferences can therefore explain both fairness-type effects in bilateral exchange and the absence of those effects in multilateral market exchange.

A good social preference theory should explain all these facts: rejections of substantial offers in ultimatum games, lower Proposer offers in dictator games than in ultimatum games, trust and repayment in trust games, and the effects of competition (which bring offers closer to the equilibrium self-interest prediction).

In ‘inequality-aversion’ theories of social preference, players prefer more money and also prefer that allocations be more equal (judged by differences in payoffs – Fehr and Schmidt, 1999 – or by deviations from payoff shares and equal shares – Bolton and Ockenfels, 2000). In a related ‘Rawlsitarian’ approach, players care about a combination of their own payoffs, the minimum payoff (à la Rawls) and the total payoff (utilitarian) (Charness and Rabin, 2002). These simple theories account relatively well for the regularities mentioned above across games, with suitable parameter values.

Missing from the inequality aversion and Rawlsitarian theories is a reaction to the intentions of players. Intentions seem to be important because players are much less likely to reject unequal offers that are created by a random device or third party than equivalently unequal offers proposed by a player who benefits from inequality (for example, Blount, 1995; Falk, Fehr and Fischbacher, 2007). In reciprocity theories which incorporate intentions, player A forms a judgement about whether another player B has sacrificed to benefit (or harm) her (for example, Rabin, 1993). A likes to reciprocate, repaying kindness with kindness, and meanness with vengeance. This idea can also explain the results mentioned above, and the

effects of intentions shown in other studies.

A newer class of theories focused on ‘social image’ – that is, player A cares about whether another player B believes A adheres to a norm of fairness. For example, Dufwenberg and Gneezy (2000) show that Trustee repayments in a trust game are correlated with the Trustee's perception of what he or she thought the Investor expected to be repaid. These models hinge on delicate details of iterated beliefs (A's belief about B's belief about A's fairness), so they are more technically complicated but can also explain a wider range of results (see Bénabou and Tirole, 2006; Dillenberger and Sadowski, 2006). Models of this sort are also better equipped to explain deliberate avoidance of information. For example, in dictator games where the dictator can either keep nine dollars or can play a ten-dollar dictator game (knowing the Recipient will *not know* which path was chosen), players often choose the easy nine dollar payment (Dana, Cain and Dawes, 2006). Since they could just play the ten-dollar game and keep all ten dollars, the ten-dollars sacrifice is presumably the price paid to avoid knowing that another person knows you have been selfish (see also Dana, Weber and Kuang, 2007).

Social preference utility theories and social image concerns like these could be applied to explain charitable contribution, legal conflict and settlement, wage-setting and wage dispersion within firms, strikes, divorces, wars, tax policy, and bequests by parents to siblings. Explaining these phenomena with a single parsimonious theory would be very useful and important for policy and welfare economics.

Limited strategic thinking and quantal response equilibrium

In complex games, equilibrium analysis may predict poorly what players do in unique games, or in the first period of a repeated game. Disequilibrium behaviour is important to understand if equilibration takes a long time, and if initial behaviour is important in determining which of several multiple equilibria will emerge. Two types of theories are prominent: cognitive hierarchy theories of different limits on strategic thinking; and theories which retain the assumption of equilibrium beliefs but assume players make mistakes, choosing strategies with higher expected payoff deviations less often.

Cognitive hierarchy theories describe a ‘hierarchy’ of strategic thinking and constrain how the hierarchy works to make precise predictions. Iterated reasoning surely is limited in the human mind because of evolutionary inertia in promoting high-level thinking, because of constraints on working memory, and because of adaptive motives for overconfidence in judging relative skill (stopping after some steps of reasoning, believing others have reasoned less). Empirical evidence from many experiments with highly skilled subjects suggests that 0–2 steps of iterated reasoning are most likely in the first period of play. A simple illustration is the ‘*p*-beauty contest’ game (Nagel, 1995; Ho, Camerer and Weigelt, 1998). In this game, several players choose a number in the interval [0,100]. The average of the numbers is computed, and multiplied by a value *p* (say 2/3). The player whose number is closest to *p* times the average wins a fixed prize.

In equilibrium players are never surprised what other players do. In the *p*-beauty contest game, this equilibrium condition implies that all players must be picking *p* times what others are choosing. This equilibrium condition only holds if everyone chooses 0 (the Nash equilibrium, consistent with iterated dominance). Figure 1 shows data from a game with p=7 and compares the Nash prediction (choosing 0) and the fit of a cognitive hierarchy model (Camerer, Ho and Chong, 2004). In this game, some players choose numbers scattered from 0 to 100, many others choose *p* times 50 (the average if others are expected to choose randomly) and others choose p^2 times 50. When the game is played repeatedly with the same players (who learn the average after each trial), numbers converge toward zero, a reminder that equilibrium concepts do reliably predict where an adaptive process leads, even if they do not predict the starting point of that process.

Figure 1

Number choices and theoretical predictions in beauty contest games. *Note*: Players choose numbers from 0 to 100 and the closest number to 0.7 times the average wins a fixed prize. *Source*: Camerer and Fehr (2006).

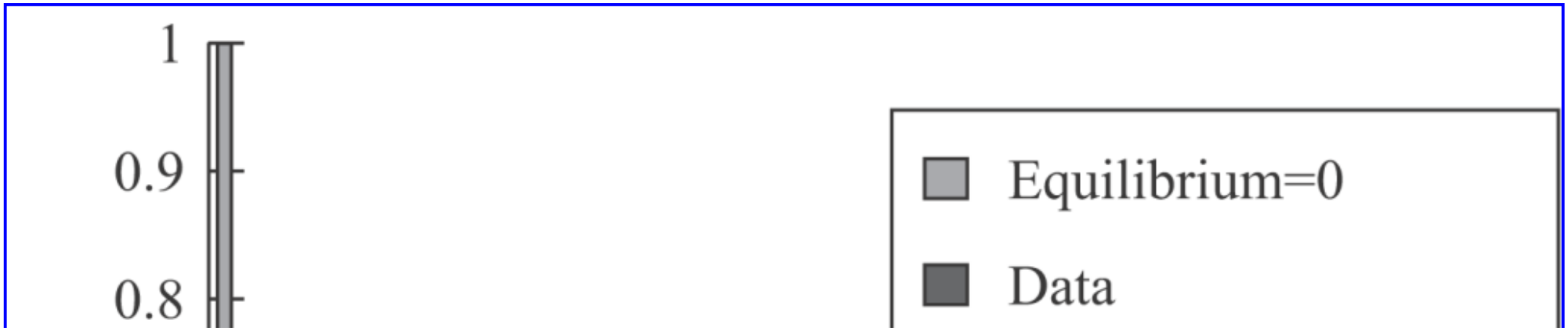

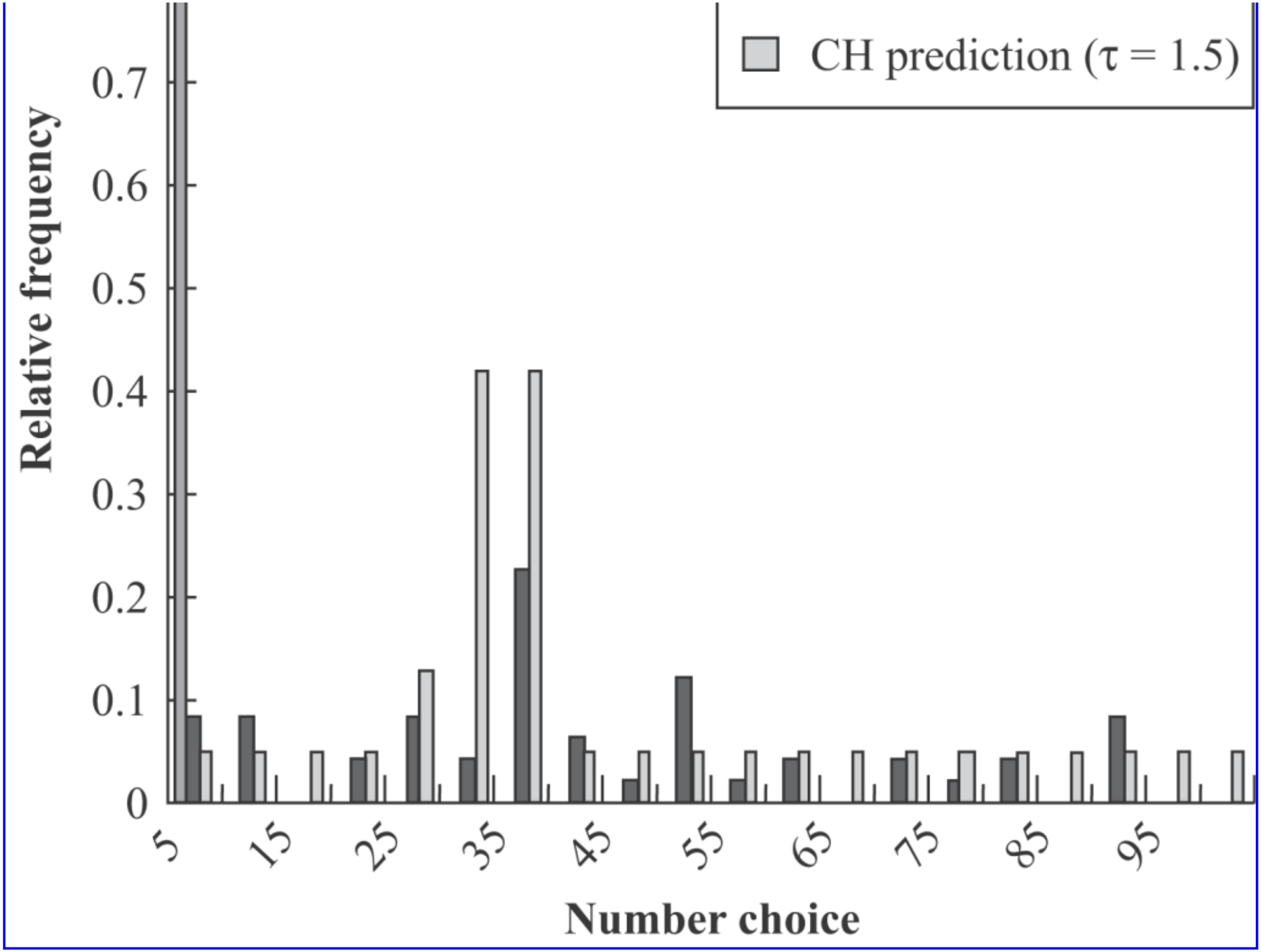

In cognitive hierarchy theories, players who do *k* steps of thinking anticipate that others do fewer steps. Fully specifying these theories requires specifying what 0-step players do,

what higher-step players think, and the statistical distribution of players' thinking levels. One type of theory assumes players who do *k* steps of thinking believe others do *k*-steps (Nagel, 1995; Stahl and Wilson, 1995; Costa-Gomes, Crawford and Broseta, 2001). This specification is analytically tractable (especially in games with *n*>two players) but implies that as players do more thinking their beliefs are further from reality. Another specification assumes increasingly rational expectations – *k*-level players truncate the actual distribution *f*(*k*) of *k*-step thinkers and guess accurately the relative proportions of thinkers doing 0 to *k*–1 steps of thinking. Camerer, Ho and Chong (2004) and earlier studies show how these cognitive hierarchy theories can fit experimental data from a wide variety of games, with similar thinking-step parameters across games.

These cognitive hierarchy theories ignore the benefits and costs of thinking hard. Costs and benefits can be included by relaxing Nash equilibrium, so that players respond stochastically to expected payoffs and choose better responses more often then worse ones, but do not maximize. Denote player *i*'s beliefs about the chance that other players *j* will choose strategy *k* by $P_i(s_j^k)$. The expected payoff of player *i*'s strategy s_i^h is $E(s_i^h) = \Sigma_k P_i(s_j^k)\pi_i(s_i^h, s_j^k)$ (where $\pi_i(x,y)$ is *i*'s payoff if *i* plays *x* and *j* plays *y*). If player *i* responds with a logit choice function, then $P_i(s_j^h) = \exp(\lambda E(s_i^h)) / \Sigma_k \exp(\lambda E(s_i^k))$. In this kind of 'quantal response' equilibrium (QRE), each player's beliefs about choice probabilities of others are consistent with actual choice probabilities, but *players do not always choose the highest expected payoff strategy (and* λ *parameterizes the degree of responsiveness; larger* λ *implies better* response). QRE fits a wide variety of data better than Nash predictions (McKelvey and Palfrey, 1995; 1998; Goeree and Holt, 2001). It also circumvents some technical limits of Nash equilibrium because players always tremble but the degree of trembling in strategies is linked to expected payoff differences.

Learning

In complex games, it is unlikely that equilibrium beliefs arise from introspection or communication. Therefore, theorists have explored the mathematical properties of various rules under which equilibration might occur when rationality is bounded.

Much research is focused on population evolutionary rules, such as replicator dynamics, in which strategies which have a payoff advantage spread through the population (for example, Weibull, 1995). Schlag and Pollock (1999) show a link between imitation of successful players and replicator dynamics.

Several individual learning rules have been fit to many experimental data-sets (see individual learning in games). Most of these rules can be expressed as difference equations of underlying numerical propensities or attractions of stage-game strategies which are updated in response to experience. The simplest rule is choice reinforcement, which updates chosen strategies according to received payoffs (perhaps scaled by an aspiration level or reference point). These rules fit surprisingly well in some classes of games (for example, with mixed strategy equilibrium, so that all strategies are played and reinforced relatively often) and in environments with little information, where agents must learn payoffs from experience, but can fit quite poorly in other games. A more complex rule is weighted fictitious play (WFP), in which players form beliefs about what others will do in the future by taking a weighted average of past play, and then choose strategies with high expected payoffs given those beliefs (Cheung and Friedman, 1997). Camerer and Ho (1999) showed that WFP with geometrically declining weights is mathematically equivalent to generalized reinforcement in which unchosen strategies are reinforced as strongly as chosen ones. Building on this insight, they create a hybrid called experience weighted attraction (EWA). The original version of EWA has many parameters because it includes all the parameters used in the various special cases it hybridizes. The EWA form fits modestly better in some games (it adjusts carefully for overfitting by estimating parameters on part of the data and then forecasting out-of-sample), especially those with rapid learning across many strategies (such as pricing). In response to criticism about the number of free parameters, Ho, Camerer, and Chong (2007) created a version with zero *learning parameters* (*just a response sensitivity* λ *as in QRE*) by replacing parameters by 'self-tuning' functions of experience.

Some interesting learning rules do not fit neatly into the class of strategy-updating difference equations. Often it is plausible to think that players are reinforcing learning *rules* rather than strategies (for example, updating the reinforcement rule or the WFP rule; see Stahl, 2000). In many game it is also plausible that people update history-dependent strategies (like tit for tat; see Erev and Roth, 2001; McKelvey and Palfrey, 2001). Selten and Buchta (1999) discuss a concept of 'direction learning' in which players adjust based on experience in a 'direction' when strategies are numerically ordered.

All the rules described above are naive (called 'adaptive') in the sense that they do not incorporate the fact that other players are learning. Models which allow players to be 'sophisticated' and anticipate learning by other players (Stahl, 1999; Chong, Camerer and Ho, 2006) often fit better, especially with experienced subjects. Sophistication is particularly important if players are matched together repeatedly – as workers in firms, firms in strategic alliances, neighbours, spouses, and so forth. Then players have an incentive to take actions that 'strategically teach' an adaptive player what to do. Models of this sort have more moving parts but can explain some basic stylized facts (for example, differences in repeated-game play with fixed 'partner' and random 'stranger' matching of players) and fit a little better than equilibrium reputational models in trust and entry deterrence games (Chong, Camerer and Ho, 2006).

Conclusion

Behavioural game theory uses intuitions and experimental evidence to propose psychologically realistic models of strategic behaviour under rationality bounds and learning, and incorporates social motivations in valuation of outcomes. There are now many mathematical tools available in both of these domains that have been suggested by or fit closely to many different experimental games: cognitive hierarchy, quantal-response equilibrium, many types of learning models (for example, reinforcement, belief learning, EWA and self-tuning EWA), and many different theories of social preference based on inequality aversion, reciprocity, and social image. The primary challenge in the years ahead is to continue to compare and refine these models – in most areas, there is still lively debate about which simplifications are worth making, and why – and then apply them to the sorts of problems in contracting, auctions, and signalling that equilibrium analysis has been so powerfully applied to.

A relatively new challenge is to understand communication. Hardly any games in the world are played without some kind of pre-play messages (even in animal behaviour). However, communication is so rich that understanding how communication works by pure deduction is unlikely to succeed without help from careful empirical observation. A good illustration is Brandts and Cooper (2007), who show the nuanced ways in which communication and incentives, together, can influence coordination in a simple organizational team game.

See Also

- adaptive expectations
- experimental economics
- individual learning in games

Bibliography

Ashraf, N., Bohnet, I. and Piankov, N. 2006. Decomposing trust and trustworthiness. *Experimental Economics* 9, 193–208.

Bénabou, R. and Tirole, J. 2006. Incentives and prosocial behavior. *American Economic Review* 96, 1652–78.

Blount, S. 1995. When social outcomes aren't fair – the effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes* 63, 131–44.

Bohnet, I. and Zeckhauser, R. 2004. Trust, risk and betrayal. *Journal of Economic Behavior & Organization* 55, 467–84.

Bolton, G.E. and Ockenfels, A. 2000. ERC: a theory of equity, reciprocity, and competition. *American Economic Review* 90, 166–93.

Brandts, J. and Cooper, A. 2007. It's what you say, not what you pay: an experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association* (forthcoming).

Camerer, C.F. 2003. *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton: Princeton University Press.

Camerer, C.F. and Fehr, E. 2006. When does ‘economic man’ dominate social behavior? *Science* 311, 47–52.

Camerer, C.F. and Ho, T.H. 1999. Experience-weighted attraction learning in normal form games. *Econometrica* 67, 827–74.

Camerer, C.F., Ho, T.-H. and Chong, J.-K. 2004. A cognitive hierarchy model of games. *Quarterly Journal of Economics* 119, 861–98.

Charness, G. and Rabin, M. 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117, 817–69.

Cheung, Y.-W. and Friedman, D. 1997. Individual learning in normal form games: some laboratory results. *Games and Economic Behavior* 19, 46–76.

Chong, J.-K., Camerer, C.F. and Ho, T.H. 2006. A learning-based model of repeated games with incomplete information. *Games and Economic Behavior* 55, 340–71.

Costa-Gomes, M., Crawford, V.P. and Broseta, B. 2001. Cognition and behavior in normal-form games: an experimental study. *Econometrica* 69, 1193–235.

Dana, J., Cain, D.M. and Dawes, R.M. 2006. What you don't know won't hurt me: costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes* 100, 193–201.

Dana, J.D., Weber, R.A. and Kuang, J.X. 2007. Exploiting moral wiggle room: behavior inconsistent with a preference for fair outcomes. *Economic Theory*. (forthcoming).

Dillenberger, D. and Sadowski, P. 2006. *Ashamed to Be Selfish*. Princeton: Princeton University Press.

Dufwenberg, M. and Gneezy, U. 2000. Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior* 30, 163–82.

Erev, I. and Roth, A.E. 2001. On simple reinforcement learning models and reciprocation in the prisoner dilemma game. In *The Adaptive Toolbox*, ed. G. Gigerenzer and R. Selten. Cambridge: MIT Press.

Falk, A., Fehr, E. and Fischbacher, U. 2007. Testing theories of intentions – fairness matters. *Games and Economic Behavior* (forthcoming).

Fehr, E. and Schmidt, K.M. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 817–68.

Goeree, J.K. and Holt, C.A. 2001. Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review* 91, 1402–22.

Henrich, J. et al. 2005. ‘Economic man’ in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28, 795–815.

Ho, T.H., Camerer, C.F and Chong, J.-K. 2007. Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory* 133, 177–98.

Ho, T.H., Camerer, C.F. and Weigelt, K. 1998. Iterated dominance and iterated best response in experimental ‘p-beauty contests’. *American Economic Review* 88, 947–69.

Hoffman, E., McCabe, K., Shachat, K. and Smith, V. 1994. Preferences, property-rights, and anonymity in bargaining games. *Games and Economic Behavior* 7, 346–80.

McKelvey, R.D. and Palfrey, T.R. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10, 6–38.

McKelvey, R.D. and Palfrey, T.R. 1998. Quantal response equilibria for extensive form games. *Experimental Economics* 1, 9–41.

McKelvey, R.D. and Palfrey, T.R. 2001. *Playing in the Dark: Information, Learning, and Coordination in Repeated Games*. Princeton: Princeton University Press.

Nagel, R. 1995. Unraveling in guessing games: an experimental study. *American Economic Review* 85, 1313–26.

Rabin, M. 1993. Incorporating fairness into game-theory and economics. *American Economic Review* 83, 1281–302.

Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review* 81, 1068–95.

Schlag, K.H. and Pollock, G.B. 1999. Social roles as an effective learning mechanism. *Rationality and Society* 11, 371–97.

Selten, R. and Buchta, J. 1999. Experimental sealed bid first price auctions with directly observed bid functions. In *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, ed. D.V. Budescu, I. Erev and R. Zwick. Mahwah, NJ: Lawrence Erlbaum.

Stahl, D. 1999. Sophisticated learning and learning sophistication. Working paper, University of Texas.

Stahl, D.O. 2000. Rule learning in symmetric normal-form games: theory and evidence. *Games and Economic Behavior* 32, 105–38.

Stahl, D.O. and Wilson, P. 1995. On players' models of other players: theory and experimental evidence. *Games and Economic Behavior* 10, 218–54.

Weibull, J. 1995. *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

How to cite this article

Camerer, Colin F. "behavioural game theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000302> doi:10.1057/9780230226203.0117

The New Palgrave Dictionary of Economics Online

behavioural genetics

William T. Dickens

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The study of how variation in genetic endowment affects behaviour has shown that a surprisingly wide range of human activities are subject to substantial genetic influence. Studies of the covariance of traits in more and less distant relatives that take into account the impact of family environment have been the main method used to demonstrate this. This article provides a brief introduction to the mechanisms of heredity, and then a discussion of the methods used by behavioural geneticists and their limitations. Both the traditional variance decomposition methods and the newer molecular genetic methods are described and discussed.

Keywords

adoption studies; behavioural genetics; chopstick problem; cognitive ability; environment vs heredity; evolutionary psychology; genome; heredity; heritability; linkage studies; twin studies

Article

While defining itself as the study of genetic influences on behaviour, behavioural genetics has been mainly concerned with demonstrating and quantifying the contribution of genetic variation to variation in human behavioural traits. As such, it contrasts with the related field of evolutionary psychology that attempts to understand how some behavioural traits common to all humans have been shaped by evolution.

The large and growing literature on the impact of genetic variation on behaviour leaves no room for doubt that genetic endowment is an important influence on a surprisingly wide range of behaviours. Behavioural genetics has relied mainly on the study of relatives with different degrees of relatedness or adoption to estimate the contributions of genetic variation and shared family environment to explaining cross-sectional variation in behavioural characteristics. More recently, behavioural geneticists have been extending their methodology to use relational studies to examine the covariation of different behavioural traits, and molecular genetic methodologies to trace the sources and causes of genetically induced

differences in behaviour.
Below I give a brief introduction to the mechanics of heredity. This is a necessary introduction to the methods of behavioural genetics, which I explain next.

Mechanics of heredity

The human genetic code is contained in 23 pairs of chromosomes made up of deoxyribonucleic acid or DNA. A DNA molecule consists of two backbone strands that are held apart by molecular pairs of four bases. A sequence of these four chemicals along one of the backbone strands encodes the plans for the different proteins from which our bodies are made. Other parts of the code are thought to control when proteins are created and in what quantities. There are about three billion base pairs on just one set of 23 chromosomes. A sequence of base pairs that codes the information for a protein or some other function is called a ‘gene’.
Of the three billion base pairs all but about three million are the same in all humans. Where base pairs differ it is said that a polymorphism exists. When a gene contains one or more polymorphic base pairs there will be different versions of the gene. Different versions of the same gene are referred to as alleles. A person's genotype is determined by what alleles that person has, while the physiological characteristics or behaviours that geneticists study are referred to as the phenotype. Any given phenotypic behaviour can be the result of having a particular genotype, a particular environmental influence, or some combination of the two. Phenotypic traits are said to be qualitative if they take a limited number of discrete forms and quantitative if they vary continuously. So the presence of the symptoms of Huntington's disease, a degenerative neurological disorder that affects older people, is a qualitative trait while one's score on an IQ test is a quantitative trait.
Genetic influence on a phenotype can involve one or more genes. For example, people who have the allele for Huntington's disease in the single gene encoding the huntingtin protein will contract it. Those who don't won't. Contrast that with the genetic influence on measured cognitive ability, which is thought to involve many genes, each of which has a very small effect on scores on tests of mental ability. When many genes influence a phenotypic trait, it is said to be polygenic.
Both qualitative and quantitative traits can be polygenic. A trivial example of a qualitative trait that is polygenic would be having an IQ score over 130. Other than some psychopathologies, most of the behaviours studied are thought to be polygenic with differences in each gene, making only a small contribution to differences in behaviour. In theory a quantitative trait could be influenced by a single gene that influenced the mean of the trait while environment determined the variance around the mean, but no examples of this have been identified.
Normally people inherit 46 chromosomes – 23 from their mothers and 23 from their fathers. Since there are many genes on any one chromosome, the inheritance of different traits can be linked if genes on the same chromosome influence the traits. However, the linkage is not perfect. In the process of creating the chromosomes that will be passed on to one's children in gamete cells (ova and sperm), contiguous parts of each pair of chromosomes can be swapped so that the chromosome that is passed onto one's child is a combination of parts from both of one's parents. This happens on average about once per chromosome in humans. Thus, traits that are influenced by genes located close together on the same chromosome are more likely to be inherited together than genes on the same chromosome that are at distant loci. As will be described later, this fact can be used to identify the location of the genes that affect a particular trait.

If one has different alleles for the same gene on each of a pair of chromosomes there are different possible impacts. In some cases, certain alleles will always be expressed (influence phenotype) if they are present. Such alleles are termed ‘dominant’. Other alleles for the same gene are called ‘recessive’ and will be expressed only if they are not paired with a dominant allele. In other cases, having two different alleles will have an effect on phenotype halfway between the effect of having two of the one allele and the effect of having two of the other. In this case genetic effects are termed ‘linear and additive’.

There can be interactions between multiple genes in creating effects on phenotype. The phenomenon is called ‘epistasis’. For example, there is epistasis if two different alleles of two different genes must be present for a phenotypic trait to be present. In this case, genetic effects on this trait will not be linear and additive.

Relational studies

Arguably the first behavioural genetics study was Galton's *Hereditary Genius* (1869) in which he looked at patterns of career success in English families. He showed that close relatives of prominent men were also likely to achieve distinction, but that the probability fell with more and more distant relatives. While a genetic basis for ability would explain this pattern, so would family connections and a host of other environmental factors. Modern behavioural genetics research uses relational data, but in a way that attempts to control for family environment.

The simplest version of this type of study looks at the behavioural similarity of identical (or monozygote) twins who are raised apart. Such twins are genetic copies of each other as they grew from the same fertilized egg, but, if they are reared apart, then environmental similarities can't explain any behavioural similarities. If one assumes that genetic and environmental influences on a trait are linear and additive, then one can write

$$P = hG + cS + eN \tag{1}$$

where P is a measure of the phenotypic behaviour, G is genetic endowment, S is an index of the influence of shared family environment, and N is an index of the influence of environmental factors not shared by family members. The variables G, S and N are not observed, but the parameters h, c and e can still be estimates. If all variables are measured as standard deviations from their means, and G, S and N are uncorrelated, then h, c and e will be the correlations of the respective variable with P and their squares will be the fraction of variance in P that is explained by each. The fractions of variance in P explained by genetic endowment, shared family environment, and non-shared environment are commonly denoted h^2, c^2 and e^2. The sum of the squared coefficients will be one. Under the assumptions that the S's and N's of identical twins raised apart are uncorrelated, the expected correlation of P for pairs of twins is h^2 or the fraction of variation in the population explained by differences in genetic endowments. This statistic is referred to as the heritability of the trait P.

If one also has data on the correlation of the behaviour for identical twins raised together, one can construct an estimate of the fraction explained by the two environmental components as well. Under the assumption that identical twins raised together have both the same *G* and the same value for *S*, the correlation of *P* across pairs of identical twins raised together will be h^2+c^2. So the difference between the correlation of *P* for identical twins raised apart and those raised together will be the fraction of variance explained by shared family environment, and 1•minus that correlation will equal the share explained by non-shared environment.

With one additional assumption it is not necessary for the adopted siblings to be identical twins. Since natural siblings receive half of their genes from each parent and the genes received from each parent are in some sense a random subset of the parents' genes, it is not unreasonable to assume that the correlation of *G* for siblings who are not identical twins will be .5. In that case the expected correlation of a phenotype behaviour for siblings raised apart will be $.5 \cdot h^2$, and multiplying that value by 2 yields an estimate of the fraction of variance in the population explained by variation in genetic endowments. Once again, the difference between the correlation for siblings raised apart and those raised together will provide an estimate of the fraction of variance explained by shared family environment. The share attributable to non-shared environment can be computed as 1•minus the sum of the shares of genetic endowment and family environment.

If the effects of genetic endowment are not linear, then heritability estimates derived from studying twins adopted apart will be larger than those for siblings raised separately. Since monozygote twins are genetically identical, they will be affected by dominant genes and interaction effects between genes (epistasis) in exactly the same way. Thus, studies of identical twins measure what is called 'broad-sense heritability' (denoted H^2) unless dominance and epistasis effects are absent. In the presence of dominance and epistasis effects the correlation of phenotypes between normal sibling pairs raised apart will be less than half of that of identical twins raised apart. Twice the correlation for normal siblings raised apart is said to measure narrow-sense heritability since it doesn't reflect the contribution of nonlinear genetic effects.

Estimated variance shares from adoption studies can be criticized on a number of grounds. Siblings raised apart, and particularly twins, will share aspects of their prenatal environment at least. They may also share their post-natal environment if they are not adopted away immediately. Also, siblings who are put up for adoption may end up in similar environments for a number of reasons. They may be adopted by relatives, or they may be adopted through the same agency that places children with parents of a particular social class in a particular geographic area. Adopting families may be matched to the socio-economic status of the biological mother. Similar environments will cause adoptees to resemble each other even if there is no effect of genetic endowment and will bias estimates of heritability upward. Adoption itself may affect the trait, leading to an overestimate of heritability and an underestimate of the role of shared environment.

Even if adoption doesn't place siblings in similar environments, it almost certainly restricts the range of environments compared with those occupied by children living with their natural parents, as adoption agencies rigorously screen parents wishing to adopt. Stoolmiller (1999) argues that this restriction of range leads adoption studies to underestimate the role of shared family environment and overestimate the importance of genetic differences in explaining variance in the general population, since there is much more variation in family environment in the general population than in adopting families. This illustrates an important characteristic of heritability estimates – they apply only to the population in

which they are estimated. Populations with different amounts of variation in environment or genetic endowment would exhibit different heritabilities. Finally, the assumption that the correlation of normal siblings with no environment in common will be exactly .5•h^2 is probably wrong for another reason. It assumes that each parent's genes for a trait are a random draw from the population – that is, that men and women don't choose each other as mates on the basis of the characteristic being studied or anything related to it. If parents are likely to have genes for the trait in common, then the expected correlation will be higher and multiplying it by 2 will overestimate heritability. If opposites attract, then multiplying the sibling correlation by 2 will understate heritability. Estimates of the variance explained by shared family environment will be affected and biased in the opposite direction to heritability.

An alternative to adoption studies are those that contrast the similarity of identical twins with that of fraternal twins. Identical twins are genetic copies of each other while fraternal twins are no more alike genetically than brothers and sisters. Thus we would expect identical twins to be more similar for traits that are subject to genetic influence. Again, under the standard assumptions, the correlation of identical twins in a population will be h^2+c^2. If one assumes that fraternal twins' genetic endowments have a correlation of .5, then their correlation will be .5 h^2+c^2. Thus, twice the difference between the correlation for identical and fraternal twins is an estimate of heritability. The fraction of variance explained by shared environment will be equal to the identical twin correlation minus the estimate of heritability, and that of non-shared environment will equal 1•minus the identical twin correlation.

Twin studies, too, can be criticized on a number of grounds. The assumption that the correlation of genetic endowment for fraternal twins will be .5 rests on random mating. If husbands and wives tend to have similar genetic endowments for the characteristic being studied, then the fraternal twin correlation will be greater than .5, and doubling the difference between fraternal and identical twins will understate heritability and overstate the role of shared environment. On the other hand, if there are dominance and epistasis effects, doubling the difference will overstate both broad and narrow sense heritability.

A common criticism of twin studies is that identical twins have more similar environments than fraternal twins and that accounts for some of their greater similarity. Whether or not this is a valid criticism, it certainly illustrates a common misunderstanding about the meaning of heritability. If identical twins have more similar environments because they behave in more similar ways and create for themselves more similar environments, some would say that it is legitimate to attribute the influence of environment of this sort to genetic endowment. In the same sense, natural siblings may have more similar environments than adopted siblings – even if they are raised apart – because their more similar genes induce more similar behaviour which induces more similar responses from their environment. If two siblings are both genetically predisposed to be taller, they may both end up playing on the high-school basketball team, where they receive professional coaching which greatly improves their skills. The similarity of their basketball skill is a direct effect of similar environments, but it is also an indirect effect of genetic endowment. Both twin and adoption studies will attribute such induced environmental effects to genetic endowment.

A common error in the interpretation of heritability estimates is the assumption that, if heritability is high, the effects of environment must be small and the trait not easy to change through environmental intervention. However, if heritability estimates attribute to genetic endowment indirect effects that come through environment, it's easy to see that this is not the case (see the discussion of malleability in the entry on cognitive ability). If a tall person is good at basketball mainly because he has received good coaching, then the skill of shorter people can probably be improved a great deal by coaching as well

(even if they can never be quite as good as the tall person). When genetic endowment has both direct physiological effects on a trait and indirect effects through induced environment, there is gene *x* environment correlation. Relaxing the assumption that genetic endowment and environmental influences are correlated doesn't invalidate heritability estimates, but it does change their interpretation as just explained. The fractions of variance explained by shared and non-shared environment in twin and adoption studies are not the full effect of environment, but the fractions explained by the residual environment – that part that can't itself be explained by differences in genetic endowment.

There is another reason why high heritability estimates do not mean that the effects of environment are necessarily weak. Recall that heritability estimates are valid only in the population in which they are estimated. If we were to study nearsightedness in a population of people who were not wearing corrective lenses, we would find it highly heritable. If we studied scores on an eye test allowing people to wear their corrective lenses, we would probably find very low heritability of test scores. The high heritability of nearsightedness in the first case certainly wouldn't mean that we couldn't treat it with corrective lenses.

Interaction of environment and genotype can create problems of interpretation similar to the just-described problems caused by the correlation between genotype and environment. Interaction is said to exist when environment has different effects depending on a person's genotype. In this case genetic effects are not linear and additive and the variance shares computed using standard behavioural genetic methods do not provide a meaningful measure of effects of genetic endowment and environment on the trait. None the less, high estimates of heritability for a population still indicate a substantial role for genetic variation in causing variation in the trait.

Some of the shortcomings of twin studies and adoption studies can be overcome by combining data from the two. Since they are subject to different biases, if results for the two types of studies are very similar, one can have some confidence that the biases are not important. Data from the two types of studies can be formally combined and used to estimate more elaborate models of inheritance that relax one or more assumptions such as linearity, random mating, or similar treatment of identical and fraternal twins. Information on other types of relations and more distant relations can be added to model building studies as well.

Of all the behaviours to which relational methods have been applied, the one that has received the most attention is scores on tests of cognitive ability. These studies have been extremely controversial -- at least in part because of the widespread misunderstanding that high heritability precluded an important role for environment. Today it is widely accepted that the heritability of cognitive test scores in adults is very high (0.6 or more; Neisser et al., 1996; Plomin et al., 2000, pp. 164–77), but it is understood that this does not imply a limited role for environment (as genetic endowment may be acting indirectly through the environment).

Besides cognitive ability, a wide range of other behaviours have been studied. The degree to which people display the symptoms of a number of psychopathologies has been shown to be subject to genetic influence (Plomin et al., 2000, chs 8 and 12). Major measurable aspects of personality (Loehlin, 1992), religiosity (Waller et al., 1990), attitudes towards one's job (Lykken et al., 1993), social attitudes (Martin et al., 1986) (including political conservatism; Eaves et al., 1997), education (Behrman and Taubman, 1989), earnings (Taubman, 1976), and even the amount of time spent watching television (Plomin et al., 1990), have all been shown to be subject to genetic influence. In most cases, studies find that the fraction

of variance explained by variation in genetic endowment is large and greater than the fraction explained by family environment (Turkheimer, 2000). Also interesting are the exceptions that have been found to this general pattern. For example, how often one attends church is influenced by one's genetic endowment, but not the type of church one attends.

A relatively recent development in relational studies is their use to analyse the sources of covariance between different measures of behaviour. By using similar assumptions to those used to identify variance shares, it is possible to tell whether correlations between variables are due mainly to common genetic factors, common environmental factors or both. For example, tests of cognitive ability are strongly correlated with scores on achievement tests and both are highly heritable. Are the same genetic factors responsible for both (as would be the case if genetic influence on achievement came entirely through its effects on cognitive ability)? For the most part they are, though some genetic influence is specific to achievement (Plomin et al., 2000, p. 201).

Animal models and molecular genetics studies

Work with animals allows behavioural geneticists to do many things that are impossible with human subjects. For example, animals can be bred for certain behavioural traits and then the specially bred animals can be used in experiments. One of the most interesting demonstrations of gene *x* environment interaction comes from a study of two strains of rats that had been bred for their performance in solving mazes (Cooper and Zubek, 1958). One strain was bred for superior performance and one for inferior performance. Rats raised in very sparse environments performed poorly in solving mazes no matter what their genetic endowment. Rats raised in enriched environments performed much better and there was little effect from their genetic endowment. However, rats raised in normal laboratory environments showed large differences consistent with their genetic endowments.

Animal studies can be particularly useful when combined with some of the new molecular genetic techniques. Certain genes can be turned off and the impact on behaviour studied. Genetic mutations can be created in experimental animals and the impact of the mutation on behaviour examined. Selectively bred animals can be compared for the frequency of different alleles to determine where genes that influence a trait are located.

Searches of this sort are facilitated by the previously described tendency for genes that are located close together on a chromosome to be inherited together. Suppose, for example, that animals that had been bred for an extreme form of some behaviour showed a much higher frequency of one allele on one chromosome than did the population from which they were bred. This would not mean that that allele played a role in the development of that trait, but it would make it more likely than not that one or more genes on the chromosome on which the gene was located played some role. The allele that is found to be associated with the trait being studied is said to be a marker for the trait, while the genes with the polymorphisms that matter for the trait are said to be trait loci. If the trait is a quantitative trait, each locus is referred to as a quantitative trait locus (QTL).

If several markers are studied on the same chromosome, some may be found to be more highly associated with the trait than others. The more highly associated markers are likely to be closer to one or more trait loci since, the closer two genes are together on the same chromosome the more likely it is that they will be inherited together.

This technique has been used to identify the location of genes with a large role in determining differences in fearfulness in mice. The same sequence of genes exists in the human genome and it is possible that variations in them may explain why some people develop anxiety disorders and some don't. Understanding the role of these genes may lead to more effective treatment.

Association techniques can also be used in humans, but are subject to a number of problems. In the example just discussed, the mice studied were all bred from the same homogenous population. The breeding for the trait is likely to have induced any association found between a marker and a phenotype trait. However, in human populations markers and traits could be associated even if there was no genetic influence on the behaviour. This is referred to as the ‘chopstick’ problem, which is named after a commonly cited example of a spurious association. In a population that included native Chinese and Europeans, using chopsticks would be associated with any marker more common in Chinese. This problem can be partially overcome by studying more homogenous populations or contrasting sibling pairs, as differences in marker frequency are more likely to signal genetic causation in these cases. In the extreme, studies can be done on large extended families. The families can be studied for co-transmission of the trait and particular alleles. These are termed ‘linkage studies’. Linkage studies were used to identify the gene responsible for Huntington's disease.

Linkage studies solve another problem of association studies in humans. Within a family, even markers fairly distant from a trait locus will have some degree of association with the trait. In the general population, markers are likely to be associated with traits only if they are trait loci themselves or are located very close to them, as recombination of chromosomes will eventually break down the association of any marker that is not a trait locus with the trait after a sufficient number of generations. A much smaller number of markers can be used to scan for the location of trait loci in a linkage study than in a study looking for association in the general population. However, linkage studies are not very good at finding QTLs when there are many genes contributing to a phenotype. Association studies in large populations are more promising, but only if the area of the genome to be examined can be narrowed on the basis of hypothesis about what systems might be involved. So far this approach has shown some promise. For example, associations have been found between a particular allele for a dopamine receptor gene and hyperactivity disorder in children (Thapar et al., 1999).

The future

Relational studies have demonstrated that variation in a surprisingly wide range of behaviours is substantially influenced by genetic differences. Molecular genetics has begun to discover some of the mechanisms by which genetic differences cause differences in behaviour, but work of this sort has barely scratched the surface, and further development faces some difficult obstacles. Most of the behaviours that have been studied are thought to be affected by many different genes, each of which has a small effect. This will make identifying QTLs difficult without some theory of what physiological processes might be involved and where the genes affecting those processes are in the genome. But what theory might one have about the location of physiological processes affecting, for example, time spent watching television?

When one begins to think about the many ways in which physiological differences could affect a wide range of behaviours, the task seems daunting. Suppose there was an allele that when present made people feel more discomfort when they were cold than others without the allele. Such people might be

inclined to spend more time inside watching TV. They might also be less athletic and/or more likely to spend a lot of time reading. If they read more, they might have larger vocabularies and score better on IQ tests. If their reading made them more sceptical, they might be less likely to attend church. Depending on how myriad and diffuse such cascading effects are, it might be impossible to understand how more than a small fraction of genetically induced differences in behaviour comes about. Still, that doesn't mean that valuable knowledge can't be gained from studying the pathways that can be identified. Such knowledge might accumulate faster if those studying the genetic influences on behaviour concentrated less on refining estimates of heritability and more on analysing the role of genetic differences in explaining the covariance of different behaviours.

See Also

- cognitive ability

Bibliography

Behrman, J.R. and Taubman, P. 1989. Is schooling mostly in the genes? *Journal of Political Economy* 97, 1425–46.

Cooper, R.M. and Zubek, J.P. 1958. Effects of enriched and restricted early environments on the learning ability of bright and dull rats. *Canadian Journal of Psychology* 12, 159–64.

Eaves, L.J., Silberg, J.L., Meyer, J.M., Maes, H.H., Simonoff, E., Pickles, A., Rutter, M., Neale, M.C., Reynolds, C.A., Erikson, M.T., et al. 1997. Genetics and developmental psychology: 2 The main effects of genes and environment on behavioral problems in the Virginia twin study of adolescent behavioral development. *Journal of Child Psychology and Psychiatry* 38, 965–80.

Galton, F. 1869. *Hereditary Genius: An Enquiry into its Laws and Consequences*. London: Macmillan.

Loehlin, J.C. 1992. *Genes and Environment in Personality Development*. Newbury Park, CA: Sage Publications, Inc.

Lykken, D.T., Bouchard, T.J., McGue, M. and Tellegen, A. 1993. Heritability of interests: a twin study. *Journal of Applied Psychology* 78, 649–61.

Martin, N.G., Eaves, L.J., Heath, A.C., Jardine, R., Feingold, L.M. and Eysenck, H.J. 1986. Transmission of social attitudes. *Proceedings of the National Academy of Science* 83, 4364–8.

Neisser, U., Boodoo, G., Bouchard, T.J., Jr., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J. and Urbina, S. 1996. Intelligence: knowns and unknowns. *American Psychologist* 51, 77–101.

Plomin, R., Corley, R., DeFries, J.C. and Fulker, D.W. 1990. Individual differences in television viewing in early childhood: nature as well as nurture. *Psychological Science* 1, 371–7.

Plomin, R., DeFries, J.C., McClearn, G.E. and McGuffin, P. 2000. *Behavioral Genetics*, 4th edn. Worth: New York.

Stoolmiller, M. 1999. Implications of the restricted range of family environments for estimates of heritability and non-shared environment in behavior-genetic adoption studies. *Psychological Bulletin* 125, 392–409.

Taubman, P. 1976. The determinants of earnings: genetics, family and other Environments: a study of male twins. *American Economic Review* 66, 858–70.

Thapar, A., Holmes, J., Poulton, K. and Harrington, R. 1999. Genetic basis of attention deficit and hyperactivity. *British Journal of Psychiatry* 174, 105–11.

Turkheimer, E. 2000. Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science* 9(5), 160–4.

Waller, N.G., Kojetin, B.A., Bouchard, T.J., Lykken, D.T. and Tellegen, A. 1990. Genetic and environmental influences on religious interests, attitudes, and values: a study of twins reared apart and together. *Psychological Science* 1, 138–42.

How to cite this article

Dickens, William T. "behavioural genetics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000165> doi:10.1057/9780230226203.0118

The New Palgrave Dictionary of Economics Online

behavioural public economics

B. Douglas Bernheim and Antonio Rangel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Behavioural public economics incorporates ideas from behavioural economics, psychology, and neuroscience in the analysis and design of public policies. This article provides an introduction to its methods and discusses its application to savings and addiction policy.

Keywords

addiction; behavioural public economics; budget constraints; compulsory saving; default options; imperfect decision processes; intertemporal choice; lump-sum taxes; myopia; neoclassical public economics; neuroscience; Pigouvian taxes; psychology and economics; tax incentives for saving; well-being

Article

Interest in the field of psychology and economics has grown in recent years, stimulated largely by accumulating evidence that the neoclassical model of consumer decision-making provides an inadequate description of human behaviour in many economic situations. Scholars have begun to propose alternative models that incorporate insights from psychology and neuroscience. Some of the pertinent literature focuses on behaviours commonly considered ‘dysfunctional’, such as addiction, obesity, risky sexual behaviour, and crime. However, there is also considerable interest in alternative approaches to more standard economic problems such as saving, investing, labour supply, risk-taking, and charitable contributions.
Behavioural public economics (BPE) is the label used to describe a rapidly growing literature that uses this new class of models to study the impact of public policies on behaviour and well-being (see Bernheim and Rangel, 2006a, for a more comprehensive review).

Background: the neoclassical approach to public economics

Public economic analysis requires us to formulate models of human decision-making with two components – one describing choices, and the other describing well-being. Using the first component, we can forecast the effects of policy reforms on individuals' actions, as well as on prices and allocations. Using the second component, we can determine whether these changes benefit consumers or harm them.
The neoclassical approach assumes that individuals' choices can be described *as if* generated by the maximization of a well-defined and stable utility function subject to feasibility and informational constraints. Neoclassical welfare analysis proceeds from the premise that, when evaluating policies, the government should act as each individual's proxy, extrapolating his preferred choices from observed decisions in related situations. This premise justifies the use of the *as-if* utility function as a gauge of well-being. In effect, this approach uses the same model for positive and normative analysis.
Within the neoclassical paradigm, government policy can affect behaviour and welfare only if it changes the decision maker's information or budget constraint. For example, vaccination campaigns may influence behaviour by providing information concerning the risks of a disease and the advantages of taking preventive action, while cigarette taxes may alter choices by raising the cost of smoking.
From the neoclassical perspective, government intervention in private markets is justified to enforce property rights, correct market failures, and address inequity by redistributing resources. Standard examples of interventions motivated by market failures include the use of taxes and subsidies to correct externalities, the provision of public goods, and the introduction of social insurance when private risk sharing is inefficient.
The accomplishments of neoclassical public economics, such as the theories of optimal income taxation and corrective environmental policy, are considerable. However, there is growing concern that this paradigm does not adequately address a number of important public policy challenges – for example, what to do about 'self-destructive' behaviours such as substance abuse, or about the apparently myopic choices of those who save 'too little' for retirement. Since the neoclassical welfare criterion respects all voluntary consumer choices (conditional on the information in the consumer's possession), it rules out the possibility of enhancing well-being by correcting 'poor' choices (except through the provision of information).

The behavioural approach to public economics

A key feature of BPE is the potential divergence of positive and normative models. Even when it is assumed that individuals are endowed with well-behaved lifetime preferences, decision processes may translate these preferences to choices imperfectly. To conduct positive analysis, one employs a model of the potentially imperfect decision process. To conduct normative analysis, one uses a well-defined welfare relation. In stark contrast to the neoclassical approach, the welfare relation may prescribe an alternative other than the one that the individual would choose for himself, at least under some conditions.
The analysis of addiction presented in Bernheim and Rangel (2004) illustrates this approach. Our model assumes that people attempt to optimize given their preferences, but randomly encounter conditions that trigger systematic mistakes, the likelihood of which evolves with previous substance use. The model is based on the following three premises. First, use among addicts is sometimes a mistake and sometimes rational. Second, experience with an addictive substance sensitizes an individual to environmental cues

that trigger mistaken usage. Third, addicts understand their susceptibility to cue-triggered mistakes and attempt to manage the process with some degree of sophistication. The first two premises are justified by a body of research in psychology and neuroscience, which shows that, after repeated exposure to an addictive substance, the brain tends to overestimate the hedonic consequences of drug consumption upon encountering environmental cues that are associated with past use. The third premise is justified by behavioural evidence indicating that users are often surprisingly sophisticated and forward looking.

The (β , δ)-model of intertemporal choice (Strotz, 1956; Phelps and Pollack, 1968; Laibson, 1997; O'Donoghue and Rabin, 1999; 2001) also illustrates the BPE approach. Psychologists have found that people often act as if they attach disproportionate importance to immediate rewards relative to future rewards, especially in situations where cognitive systems are overloaded. (For a recent review of this literature, see Frederick, Loewenstein and O'Donoghue, 2002; Loewenstein, Read and Baumister, 2003.) To capture this tendency, the (β , δ)-model assumes that, in each period *t*, individuals behave as if they maximize a utility function of the form

$$u(c_t) + \beta\left[\sum_{k=t+1}^{T} \delta^{k-t} u(c_k)\right],$$

where $0< \beta <1$. In this framework, the parameter β represents the degree of *present bias* or *myopia*. The neoclassical model corresponds to the special case where $\beta=1$. With $\beta<1$, behaviour is dynamically inconsistent. This complicates positive analysis, since behaviour no longer corresponds to the solution of single utility maximization problem.

Many analysts interpret present bias as a mistake. They argue that the individual's underlying well-being actually corresponds to the preferences revealed through choices that do not involve immediate rewards:

$$U(c_1, \ldots, c_T) = \sum_{t=0}^{T} \delta^t u(c_t).$$

Under this interpretation, $\beta<1$ creates a tendency to consume excessively in the present.

These examples illustrate some important conceptual and methodological aspects of BPE. First, with behaviour and welfare modelled separately, BPE allows for the possibility of mistakes. In contrast to a neoclassical analyst, a BPE analyst can pose questions that presuppose possible divergences between behaviour and preferences, such as whether Americans save too little for retirement, or whether addicts engage in self-destructive behaviour. Within the BPE framework, one can test the hypothesis that individuals maximize their well-being, and measure the magnitude of their errors. Second, to justify either a positive representation of choice or a particular welfare criterion, a BPE analyst relies on evidence from psychology and neuroscience. This evidence can help economists pin down underlying

preferences by identifying the mechanisms responsible for the decision-making errors. Good structural models of decision-making processes may also improve the quality of out-of-sample behavioural predictions, which are often required for policy evaluation.

Behavioural policy analysis

BPE models are extensions of neoclassical models. Thus, they imply that public policy can modify behaviour by changing budget constraints and/or information. For example, cigarette prices affect cigarette consumption in the Bernheim–Rangel addiction model, and savings are responsive to interest rates in most specifications of the (β , δ)-model.

In addition, the BPE framework introduces new channels through which public policy can affect behaviour and welfare. In particular, it allows for the possibility that some public policies can influence behaviour *directly* by activating particular cognitive processes, even when they leave budget constraints and information unchanged.

For example, Brazil and Canada require every pack of cigarettes to display a prominent, viscerally charged image depicting some deleterious consequences of smoking, such as lung disease and neonatal morbidity. Since the consequences of smoking are well known, this policy has no effect in information or budget constraints. And yet the Bernheim–Rangel theory of addiction allows for the possibility that a sufficiently strong counter-cue could reduce the probability of a mistake by triggering thought processes that induce users to resist cravings. When successful, this policy affects behaviour by activating particular cognitive processes.

Another striking example involves the effects of default options in employee-directed pension plans. A ‘default option’ is the outcome resulting from inaction. For a neoclassical consumer, choices depend only on preferences, information, and constraints. Consequently, in the absence of significant transaction costs, default options should be inconsequential. However, in the context of decisions concerning saving and investment, defaults seem to matter a great deal. For example, with respect to 401(k) plans (employer-sponsored retirement savings accounts in the United States that receive preferential tax treatment), there is considerable evidence that default options affect participation rates, contribution rates, and portfolios (Madrian and Shea, 2001; Choi, Laibson and Madrian, 2004). Yet, arguably, a default neither affects opportunities (since transaction costs are low) nor provides new information.

While BPE models admit traditional justifications for government intervention in private markets (the enforcement of property rights, the correction market failures, and the redistribution of resources), they also introduce novel justifications. For example, public policy may improve welfare by reducing the size, likelihood, or consequences of mistakes. As shown in the next two sections, this can lead to conclusions that are strikingly at odds with those generated by the neoclassical model.

Example: addiction policy

In the neoclassical theory of rational addiction (Becker and Murphy, 1988), government intervention may be justified *only* when it corrects market failures involving addictive substances, such as second-hand smoking, or when it combats ignorance or misinformation. In contrast, in our model of addiction (Bernheim and Rangel, 2004), government intervention may also be justified when it reduces the

frequency, magnitude, and consequences of mistakes. These considerations give rise to a number of non-standard policy implications.

Limitations of informational policy. In practice, public education campaigns (such as anti-smoking and anti-drug initiatives) have achieved mixed results. Our view of addiction highlights a fundamental limitation of informational policy: contrary to standard theory, one cannot assume that even a highly knowledgeable addict always makes informed choices. Information about the consequences of substance abuse may affect initial experimentation with drugs, but cannot alter the neurological mechanisms through which addictive substances subvert deliberative decision-making.

Beneficial harm reduction. If addiction results from randomly occurring mistakes, various interventions can serve social insurance objectives by ameliorating some of its worst consequences. For instance, subsidization of rehabilitation centres and treatment programmes (particularly for the indigent) can moderate the financial impact of addiction and promote recovery. Likewise, the free distribution of clean needles can moderate the incidence of diseases among heroin addicts. In some cases, it may even be beneficial to make substances available to severe addicts at low cost, a policy used in some European countries.

Counterproductive disincentives. Policies such as ‘sin taxes’ strive to discourage use by making substances costly. This is potentially justifiable on the grounds that use generates negative externalities. Even higher taxes (whether implicit or explicit) might be justified if they also reduce ‘unwanted’ use. Unfortunately, the compulsive use of addictive substances is probably much less sensitive to costs and consequences than is deliberative use. Consequently, imposing costs on users in excess of the standard Pigouvian levy will likely distort deliberate choices detrimentally, without significantly reducing problematic compulsive usage. In addition, policies that impose high costs on use may thwart social insurance objectives by exacerbating the consequences of uninsurable risks associated with the use of addictive substances, such as poverty and prostitution. Accordingly, for some substances the optimal rate of taxation for addictive substances may be significant *lower* than that the standard Pigouvian levy (see Bernheim and Rangel, 2005, for simulation results).

Policies affecting cues. Since environmental cues appear to trigger addictive behaviours, public policy can also influence use by changing the cues that people normally encounter. One approach involves the elimination of problematic cues. For example, advertising and marketing restrictions of the type imposed on sellers of tobacco and alcohol suppress one possible artificial trigger for compulsive use. Since one person's decision to smoke may trigger another, confining use to designated areas may reduce unintended use. Another approach involves the creation of counter-cues, which we discussed above. Policies that eliminate problematic cues or promote counter-cues are potentially beneficial because they combat compulsive use while imposing minimal inconvenience and restrictions on rational users.

Facilitation of self-control. Most behavioural theories of addiction potentially justify policies that provide better opportunities for self-regulation without making particular choices compulsory. In principle, this helps those who are vulnerable to compulsive use without encroaching on the freedoms of those who would deliberately choose to use. Laws that limit the sale of a substance to particular times, places, and circumstances may facilitate self-regulation. Well-designed policies could in principle accomplish this objective more effectively. For example, a number of states have enacted laws allowing problem gamblers to voluntarily ban themselves from casinos. Alternatively, if a substance is available only by prescription, and if prescription orders are filled on a ‘next day’ basis, then deliberate forward-looking planning becomes a prerequisite for availability. In the absence of a pervasive black market,

recovering heroin addicts could self-regulate problematic compulsive use by carefully choosing when, and when not, to file requests for refills.

Example: savings policy

The (β , δ)-model of savings also exemplifies the novel policy insights generated by the BPE approach. For example, this model implies that many individuals will save too little for retirement, and that there may be Pareto improving policy interventions even in the absence of capital market distortions – a conclusion that is at odds with the neoclassical framework. Other notable implications include the following:

Mandatory savings policies. Within the (β , δ) framework, compulsory saving may be welfare-enhancing if it fully crowds out private saving (in the form of liquid assets) at some point during the life cycle (Imrohoroglu, Imrohoroglu and Joines, 2003; Diamond and Koszegi, 2003). This provides a rationale for mandatory savings programmes, which are pervasive across the world, and which are more difficult to justify within the neoclassical framework.

Saving subsidies. On the assumption that (*a*) the population includes some individuals with self-control problems and (*b*) the social welfare function is continuous and concave, a small subsidy for saving financed with lump-sum taxes is welfare improving (O'Donoghue and Rabin, 2006; Krusell, Kuruscu and Smith, 2000; 2002). Intuitively, the subsidy produces a first-order improvement in the well-being of individuals with self-control problems (since they save too little), and only a second-order reduction in the well-being of those without self-control problems. This provides a possible rationale for tax-favoured savings programmes, such as, in the United States, 401(k) plans and Individual Retirement Accounts (IRAs).

Credit restrictions. Introducing restrictions on the availability of credit, for example, by regulating the distribution of revolving credit lines and mandating credit ceilings, can potentially enhance the well-being of those with self-control problems. For example, Laibson, Repetto and Tobacman (2004) estimate that the representative (β , δ) consumer would be willing to pay $2000 at the age of 20 to exclude himself from the credit card market.

Behavioural public economics circa 2006

As of 2006, the rapidly growing field of BPE has demonstrated its value by enhancing our understanding of public policy in several areas, including savings and addiction. Nevertheless, the literature is still in its infancy. As time passes, we anticipate that the methods and tools of BPE will contribute new insights in these areas, as well as to other difficult public policy issues involving poverty, crime, corruption, violence, obesity, and charitable giving, among others.

In addition to providing new insights concerning the effects of familiar policies, research in BPE can also guide the design of new policies. One obvious goal is to reduce the frequency of mistakes among those who behave suboptimally without interfering with the choices of those who behave optimally. Some recent fieldwork by Thaler and Bernartzi (2004), who advocate a savings programme called Save More Tomorrow, illustrates the potential value of this approach. In this programme, a worker can allocate a portion of her future salary increases towards retirement savings. Subsequently, she is allowed

to change this allocation at a negligible transaction cost. In practice, 78 per cent of those who were eligible for the plan chose to participate, 80 per cent of participants remained in the plan through the fourth pay raise, and the average contribution rate for programme participants increased from 3.5 per cent to 13.6 per cent over the course of 40 months.

To date, progress in BPE has been somewhat hampered by the absence of a general framework for behavioural welfare analysis. Analysts tend to devise and justify welfare criteria on a case-by-case basis, rather than through the application of general principles. Ongoing research aims to fill this gap (see Bernheim and Rangel, 2006b).

See Also

- addiction
- behavioural game theory
- charitable giving
- neuroeconomics
- public goods experiments

Bibliography

Becker, G. and Murphy, K. 1988. A theory of rational addiction. *Journal of Political Economy* 96, 675–700.

Bernheim, B.D. and Rangel, A. 2004. Addiction and cue-triggered decision processes. *American Economic Review* 94, 1558–90.

Bernheim, B.D. and Rangel, A. 2005. From neuroscience to public policy: a new economic view of addiction. *Swedish Economic Policy Review* 12, 11–46.

Bernheim, B.D. and Rangel, A. 2006a. Behavioral public economics: welfare and policy analysis with fallible decision-makers. In *Economic Institutions and Behavioral Economics*, ed. P. Diamond and H. Vartiainen. Princeton: Princeton University Press, (forthcoming).

Bernheim, B.D. and Rangel, A. 2006b. Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review Papers and Proceedings*, (forthcoming).

Choi, J., Laibson, D. and Madrian, B. 2004. Plan design and 401(k) savings outcomes. *National Tax Journal* 57, 275–98.

Diamond, P. and Koszegi, B. 2003. Quasi-hyperbolic discounting and retirement. *Journal of Public Economics* 87, 1839–72.

Frederick, S., Loewenstein, G. and O'Donoghue, T. 2002. Time discounting and time preference: a critical review. *Journal of Economic Literature* 40, 351–401.

Imrohoroglu, S., Imrohoroglu, A. and Joines, D. 2003. Time inconsistent preferences and social security. *Quarterly Journal of Economics* 118, 745–84.

Krusell, P., Kuruscu, B. and Smith, A. 2000. Tax policy with quasi-geometric discounting. *International Economic Journal* 14(3), 1–40.

Krusell, P., Kuruscu, B. and Smith, A. 2002. Equilibrium welfare and government policy with quasi-geometric discounting. *Journal of Economic Theory* 105, 42–72.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112, 443–77.

Laibson, D., Repetto, A. and Tobacman, J. 2004. Estimating discount functions from lifecycle consumption choices. Working paper, Harvard University.

Loewenstein, G., Read, D. and Baumister, R., eds. 2003. *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*. New York: Russell Sage Foundation.

Madrian, B. and Shea, D. 2001. The power of suggestion: inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics* 116, 1149–87.

O'Donoghue, T. and Rabin, M. 1999b. Doing it now or later. *American Economic Review* 89, 103–24.

O'Donoghue, T. and Rabin, M. 2001. Choice and procrastination. *Quarterly Journal of Economics*, 121–60.

O'Donoghue, T. and Rabin, M. 2006. Optimal sin taxes. *Journal of Public Economics* 90, 1825–49.

Phelps, E. and Pollack, R. 1968. On second-best national savings and game equilibrium growth. *Review of Economic Studies* 35, 185–99.

Strotz, R.H. 1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23, 165–80.

Thaler, R.H. and Bernartzi, S. 2004. Save more for tomorrow: using behavioral economics to increase employee savings. *Journal of Political Economy* 112, S164–S187.

How to cite this article

Bernheim, B. Douglas and Antonio Rangel. "behavioural public economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000331>
doi:10.1057/9780230226203.0119

The New Palgrave Dictionary of Economics Online

Bellman equation

Yongseok Shin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bellman equation; consumption smoothing; convergence; dynamic programming; Markov processes; neoclassical growth theory; value function

Article

Dynamic programming is a method that solves a complicated multi-stage decision problem by first transforming it into a sequence of simpler problems. Bellman equations, named after the creator of dynamic programming Richard E. Bellman (1920–84), are functional equations that embody this transformation.
Take, for example, a typical maximization problem in economics:

$$\max_{\{u_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t F(x_t, u_t), \tag{1}$$

s.t. $x_{t+1} = g(x_t, u_t)$ and $u_t \in \Gamma(x_t)$, with x_0 given.
The set $\Gamma(x_t)$ consists of admissible values of the control variable u_t given the state variable x_t. We assume that $\Gamma(x_t)$ is non-empty for all x_t. We also assume that $F(x_t,u_t)$ is concave and that the set $\{(x_t, x_{t+1}) : x_{t+1} = g(x_t, u_t),\ u_t \in \Gamma(x_t)\}$ is compact and convex. It is further assumed that $\beta \in (0,1)$.
This so-called sequence problem has an infinite number of controls $\{u_t\}_{t=0}^{\infty}$, and is generally intractable as it is. Dynamic programming reduces this infinite-dimensional problem into an infinite sequence of one-dimensional problems:

$$\max_{u \in \Gamma(x)} F(x, u) + \beta V(x'), \tag{2}$$

s.t. $x' = g(x, u)$.

The unknown function $V(x)$ represents the maximized value of the original problem starting from an arbitrary initial condition x, and is called the value function. In particular, $V(x_0)$ must be equal to the maximized value of the objective function in the original problem (1). Once $V(x)$ is known, the maximizer of (2) would take the form of an optimal decision rule, or a policy function: $u^* = h(x)$. Let the maximizer of the original problem (1) be $\{u_t^*\}_{t=0}^{\infty}$. Then $\{u_t^*\}_{t=0}^{\infty}$ can be generated from (2) recursively by $u_t^* = h(x_t)$ and $x_{t+1} = g(x_t, u_t^*)$, starting from the given x_0. Bellman called this connection between the sequence problem (1) and the recursive problem (2), the principle of optimality. Now we have to solve for $V(x)$ and, subsequently, $h(x)$. To this end, we re-write (2) as follows:

$$V(x) = \max_{u \in \Gamma(x)} F(x, u) + \beta V(g(x, u)). \tag{3}$$

This functional equation in $V(x)$ is the Bellman equation. From the definition of $h(x)$, it follows that $V(x) = F(x, h(x)) + \beta V(g(x, h(x)))$.

Typically, the Bellman equation can be solved for the unknown $V(x)$ by value function iteration. This method can be described as follows.

1. 1. Guess an arbitrary function $V_j(x)$, j=0.
2. 2. Given $V_j(x)$, compute $V_{j+1}(x) = \max_{u \in \Gamma(x)} F(x, u) + \beta V_j(g(x, u))$.
3. 3. Repeat Step 2 until the sequence of functions $\{V_j\}_{j=0}^{\infty}$ thus constructed converge. The limit of this sequence is the solution to the functional equation (3), $V(x)$.

Under some conditions (for example, Blackwell's sufficient conditions), it is proven that value function iteration recovers the unique solution to (3) starting from an arbitrary initial guess $V_0(x)$. See Bertsekas (1976) or Stokey and Lucas (1989) for detailed expositions on convergence. The procedure may sound straightforward, but, in practice, it is impossible (with few exceptions) to compute even one iteration of

Step 2 by hand. One has to use numerical approximation and maximization routines on computers.

It is known that the value function inherits monotonicity and concavity properties of the one-period return function *F*. In addition, Benveniste and Scheinkman (1979) showed that the value function is once differentiable under fairly general conditions. See Stokey and Lucas (1989) for more on the properties of the value function.

Dynamic programming enables researchers to analyse interesting economic problems that cannot be solved otherwise. Thus, it is no surprise that Bellman equations are widely used in economics. Below, I provide two examples of such usage.

Example 1: Neoclassical growth model

Brock and Mirman (1972) set up a neoclassical growth model with log preference and full depreciation. This example is one of the few cases where one can actually solve the Bellman equation by hand, using the value function iteration method. The planner's problem is to maximize $\sum_{t=0}^{\infty}\beta^t \ln(c_t)$, subject to the resource constraint of $c_t + k_{t+1} \leq Ak_t^{\alpha}$, with $A>0$, $\alpha \in (0, 1)$ and $\beta \in (0, 1)$. In this problem, k_t is the state variable and c_t is the control, with $\Gamma(k) = \{c: 0 < c \leq Ak^{\alpha}\}$, $g(k_t, c_t) = Ak_t^{\alpha} - c_t$ and $F(k_t, c_t) = \ln(c_t)$. The Bellman equation for this problem is:

$$V(k) = \max_{0<c\leq Ak^{\alpha}} \ln(c) + \beta V(Ak^{\alpha} - c).$$

Let's solve the Bellman equation by iterating on the value function. Begin by guessing $V_0(k) \equiv 0$. Following the procedure outlined above, we obtain:

$$V_1(k) = \ln(Ak^{\alpha}) = \ln(A) + \alpha\ln(k),$$

$$V_2(k) = \ln\frac{A}{1+\alpha\beta} + \beta\ln(A) + \alpha\beta\ln\frac{\alpha\beta A}{1+\alpha\beta} + \alpha(1+\alpha\beta)\ln(k).$$

Iterating onwards and using the summation formula for geometric series, we arrive at:

$$V(k) = \frac{1}{1-\beta}\left\{\ln(A(1-\alpha\beta)) + \frac{\alpha\beta}{1-\alpha\beta}\ln(A\alpha\beta)\right\} + \frac{\alpha}{1-\alpha\beta}\ln(k).$$

The optimal decision rule can now be easily computed: $c^* = h(k) = (1-\alpha\beta)Ak^{\alpha}$.

Example 2: Consumption smoothing

Our discussion of Bellman equations up to this point has been limited to deterministic models. However, as long as the objective function is additively separable over time and is linear in probability, we can easily accommodate uncertainty. For example, Miller (1974) analyses a consumer's utility maximization in the face of a stochastic income stream using dynamic programming. What follows is an adapted version of Miller's model.

Think of an infinitely lived consumer or dynasty that maximizes the discounted sum of the expected utility stream. The consumer derives utility from consumption c_t, and we denote the utility function with $U(c_t)$. Her income follows a Markov process $\{y_t\}_{t=0}^{\infty}$, and the distribution of y_{t+1} given y_t is represented by the cumulative density function $G(y_{t+1}|y_t)$. We assume that $y_t \in [0, y_{\max}]$, $\forall t$. The consumer's discount factor is $\beta \in (0,1)$ and the market interest rate is r. It is assumed that $\beta(1+r)<1$. She can borrow and lend at the market interest rate, but her debt cannot exceed $B_{\max} < \infty$. We denote her asset holdings at the beginning of period t with a_t. To be precise, c_t and a_t are measurable functions with respect to the σ-algebra generated by the income process. For notational convenience, we suppress this history dependence. Now we write down the consumer's problem:

$$\max_{\{c_t\}_{t=0}^{\infty}} E_0 \sum_{t=0}^{\infty} \beta^t U(c_t),$$

s.t. $c_t + \frac{a_{t+1}}{1+r} \le a_t + y_t$, $a_t \ge -B_{\max}$ and $y_{t+1} \sim G(y_{t+1}|y_t)$, with a_0 and y_0 given.

To obtain a recursive formulation, it must be noted that (a_t, y_t) are the relevant state variables. Without loss of generality, assume that there is no borrowing. The Bellman equation for this consumer's problem is then:

$$V(a, y) = \max_{0 \le c \le a+y} U(c) + \beta \int V((1+r)(a+y-c), y')\,dG(y'|y).$$

Unlike in the first example, this Bellman equation cannot be solved by hand in general, and necessitates numerical methods.

See Also

- dynamic programming

Bibliography

Bellman, R. 1957. *Dynamic Programming*. Princeton: Princeton University Press.

Benveniste, L. and Scheinkman, J. 1979. On the differentiability of the value function in dynamic models of economics. *Econometrica* 47, 727–32.

Bertsekas, D.P. 1976. *Dynamic Programming and Stochastic Control*. New York: Academic Press.

Blackwell, D. 1965. Discounted dynamic programming. *Annals of Mathematical Statistics* 36, 226–35.

Brock, W.A. and Mirman, L. 1972. Optimal economic growth and uncertainty: the discounted case. *Journal of Economic Theory* 4, 479–513.

Ljungqvist, L. and Sargent, T.J. 2004. *Recursive Macroeconomic Theory*, 2nd edn. Cambridge, MA: MIT Press.

Miller, B.L. 1974. Optimal consumption with a stochastic income stream. *Econometrica* 42, 253–66.

Stokey, N.L. and Lucas, R.E., Jr. 1989. *Recursive Methods in Economic Dynamics*. Cambridge: Harvard University Press.

How to cite this article

Shin, Yongseok. "Bellman equation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000340> doi:10.1057/9780230226203.0120

The New Palgrave Dictionary of Economics Online

Ben Porath, Yoram (1937 – 1992)

Reuben Gronau
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Ben Porath, Y.; child mortality; family economics; fertility; human capital investment; population growth; women and work

Article

Yoram Ben Porath's paper ‘The Production of Human Capital and the Life Cycle of Earnings’ (1967) is still regarded as one of the path-breaking papers in the economics of human resources. Following Mincer and Becker, the paper uses the framework of optimum control to analyse the joint decision of investment in human capital and market work over the life cycle. Diminishing marginal productivity in the investment process results in the process being spread over a lengthy period of time. A shrinking horizon results in the time devoted to the investment diminishing over the life cycle, an increasing fraction of time being diverted to market work. The model, part of Ben Porath's doctoral dissertation, provides an elegant economic explanation for the concentration of formal studies (that is, ‘full-time’ investment) early in life, and the concave shape of the age-earning profile.

Ben Porath's MA thesis (1966) was the most comprehensive economic study of the Arab labour force and the Arab sector in the Israeli economy at the time of its composition. Like his doctorate, it reflects Ben Porath's lifetime interest in the interaction between human resources and growth. In a series of studies on fertility patterns in Israel he explored the substitution between quality and quantity, sex preferences and family size (1976; 1981), the effect of child mortality on family size (1976), and the interaction between fertility and women's labour supply (1985), combining theory and empirical research.

Ben Porath's interest in the economics of fertility led him to widen the scope of investigation, focusing on the economic functions of the family. In his 1980 essay ‘The F-connection: Families, Friends and Firms and the Organization of Exchange’ he explored the social and economic role of families, contrasting the exchange taking place within the family (or other small socially knit groups) which are characterized by ‘specialization by identity’ and the conventional view of market exchange between anonymous buyers and sellers. In a world of imperfect information the transactional advantages of trade

within a small group plays an important role in explaining the shifting border between the family and the market.

In 1979, when Ben Porath became the director of the Maurice Falk Institute for Economic Research in Israel, he initiated a comprehensive study of the economy of Israel, an economy plagued by an uncontrollable inflation and halting growth. In the opening paper of the volume that he edited, *The Israeli Economy: Maturing through Crisis* (1986), he returned to tackle the question that puzzled him throughout his career – the interaction between output and population growth: is population growth the engine of output growth, or does output growth encourage immigration?

Yoram Ben Porath was born in Tel Aviv in 1937. He started his studies in economics at the Hebrew University in Jerusalem in 1957, and received his Ph.D. from Harvard in 1967, studying with Simon Kuznets. In 1986 he was elected Deputy Provost of the Hebrew University, and later became Provost. In 1990 he was elected president of the university. In 1992, during his term as president, he was killed in a car accident.

Selected works

1966. *The Arab Labor Force in Israel*. Jerusalem: Maurice Falk Institute for Economic Research in Israel.

1967. The production of human capital and the life cycle of earnings. *Journal of Political Economy* 75, (pt. 1), 352–65.

1976. Fertility response to child mortality: microdata from Israel. *Journal of Political Economy* 84, S163–S178.

1976. (With F. Welch.) Do sex differences really matter? *Quarterly Journal of Economics* 90, 285–307.

1980. The F-connection: families, friends, and firms and the organization of exchange. *Population and Development Review* 6, 1–30.

1981. (With F. Welch.) On sex preferences and family size. In *Research in Population Economics*, vol. 2, ed. J. Simon. Greenwich, CT: JAI Press.

1985. (With R. Gronau.) Jewish mother goes to work: trends in the labor force participation of women in Israel, 1955–80. *Journal of Labor Economics* 3, (pt. 2), S310–S327.

1986. *The Israeli Economy: Maturing through Crises*. Cambridge, MA: Harvard University Press.

How to cite this article

Gronau, Reuben. "Ben Porath, Yoram (1937–1992)." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000338> doi:10.1057/9780230226203.0121

The New Palgrave Dictionary of Economics Online

Bentham, Jeremy (1748 – 1832)

Philip Schofield

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Jeremy Bentham, English philosopher and reformer, was the founder of classical utilitarianism, the doctrine that an action was morally right to the extent that it promoted the greatest happiness of the greatest number. In Bentham's hands, the principle of utility provided a critical standard by which to test the value of existing practices, laws, and institutions, and to suggest reform and improvement. His basic premise in political economy was that wealth would be most effectively produced where the individual was left free from government intervention, though government had a crucial role in providing the background conditions of security without which civilized life was impossible.

Keywords

abundance; Bentham, J.; civil law; colonies; consumption taxation; democracy; diminishing marginal utility; economic freedom; equality; escheat; ethics; expectations; free press; happiness; inflation; moneylenders; national debt; Panopticon; political economy; population growth; possession; principle of utility; property; property taxation; psychological hedonism; public debt; punishment; redistribution of income and wealth; security; sinister interests; sinking fund; subsistence; usury; utilitarianism; wealth; well-being

Article

Jeremy Bentham, English philosopher and reformer, was the founder of classical utilitarianism, and, thereby, arguably the founder of the modern discipline of economics.
Bentham was born in Church Lane, Houndsditch, London on 15 February 1748. His father Jeremiah Bentham (1712–1792) was a solicitor, with a practice in the Court of Chancery, and wealthy and important clients in the City of London. Of his six siblings, only one younger brother Samuel (1757–1831) survived into adulthood, becoming a prominent naval architect and engineer. His mother Alicia died on 6 January 1759. A precocious child, he was educated at Westminster School until 1760 when his

father entered him, at the age of 12, into the University of Oxford, where he graduated in 1764, reputedly the youngest person ever to have done so. In the meantime, in accordance with his father's wish to see him pursue a career in the law, he had entered Lincoln's Inn in 1763, and was admitted to the bar in 1769. In that same year, however, he convinced himself that he should not practise law but rather devote himself to legal reform. Bentham thought of himself as ‘the Newton of legislation’—just as Isaac Newton (1642–1727) had brought order to the physical sciences, so would Bentham to the moral sciences. He adopted the principle of utility (an action was judged to be morally right to the extent that that it promoted the greatest happiness of the greatest number) as a critical standard by which to test the value of existing practices, laws, and institutions, and to suggest reform and improvement. He set about composing a comprehensive code of laws, to which his best-known work, *An Introduction to the Principles of Morals and Legislation* (printed 1780, published 1789), was intended to form a preface. He announced that his enterprise was ‘to rear the fabric of felicity by the hands of reason and of law’ (Bentham, 1970, p. 11).

Principle of utility

Bentham's critical standard, the principle of utility, was based on the psychological insight that sentient creatures were motivated by a desire for pleasure and an aversion to pain. An individual had a motive to perform an action – or, put another way, had an interest in performing it – if he expected to gain some pleasure or avert some pain from doing so, and the greater or more valuable the pleasure experienced or pain averted, the stronger the motive or greater the interest. The value of a pleasure or pain was determined by its quantity, which, in the case of a single individual was a product of its intensity, duration, certainty, and propinquity. Where the value of a pleasure or pain was considered in relation to more than one person, then, in addition to these circumstances, the circumstance of extent, that is, the number of persons affected by it, had to be taken into account. At this point, a statement of psychological fact became a statement of moral science. An act was morally good if, after calculating all the pains or pleasures produced in the instance of every individual affected, the balance was on the side of pleasure, and morally evil if on the side of pain. Psychology and ethics were both founded on, and therefore linked by their relation to, pleasure and pain. Hence, Bentham's statement that, ‘Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do.’ The ‘sovereign masters’ of pain and pleasure not only accounted for human motivation, ‘govern[ing] us in all we do, in all we say, in all we think’, but also provided ‘the standard of right and wrong’. (Bentham, 1970, p. 11).

Panopticon

The middle part of Bentham's life, from about 1790 to 1803, was dominated by his attempt to build a panopticon prison in London. The panopticon design was the brainchild of Bentham's brother Samuel, when employed in the 1780s on the estates of Prince Grigoriy Aleksandrovich Potemkin (1724–1791) at Krichev, in Russia. He found that, by organizing his workforce in a circular building, with himself at the centre, he could supervise its activities more effectively. On a visit to his brother in the late 1780s and seeing the design, Bentham immediately appreciated its potential. Enshrining the principle of inspection, the panopticon might be adapted as a mental asylum, hospital, school, poor house, factory, and, of

course, prison. The prison building would be circular, with the cells, occupying several storeys one above the other, placed around the circumference. At the centre of the building would be the inspector's lodge, which would be so constructed that the inspector would always be capable of seeing into the cells, while the prisoners would be unable to see whether they were being watched. The activities of the prisoners would be transparent to the inspector; his actions, in so far as the prisoners were concerned, were hidden behind a veil of secrecy. On the other hand, it was a cardinal feature of the design that the activities of the inspector and his officials should be laid open to the general scrutiny of the public, who would be encouraged to visit the prison. When the panopticon scheme effectively collapsed in 1803, Bentham was left embittered by what he regarded as the bad faith of successive ministries, and he became increasingly committed to political radicalism.

Defence of Usury

While in Russia, Bentham composed *Defence of Usury* (1787), which proved to be one of his most successful attempts to influence economic policy. Bentham greatly admired Adam Smith's *Wealth of Nations*, which he studied in detail. He was not, however, an uncritical admirer, and argued that Smith had contradicted his own free market principles by defending the legal prohibition against exorbitant rates of interest. Countering the popular sentiment which condemned the moneylender for his avarice and pitied the borrower, Bentham argued that the former embodied the virtues of frugality, thrift, and prudence, and the latter, whether described as an entrepreneur or a prodigal, should be allowed to decide for himself whether to enter into a particular money bargain. In other words, Bentham saw no reason why the freedom of commerce should not be extended to the lending and borrowing of money. At the same time, Bentham defended the projector from the criticisms of Smith, who had linked the projector with the prodigal, and contrasted both with the sober person. The projector (and Bentham, with his panopticon prison scheme, placed himself in this category) promoted utility by improving existing products and processes or by inventing new and better ones: in short, projectors were the agents of progress.

Political economy and the four sub-ends of utility

Bentham's most intense period of work on questions of political economy took place between 1793 and 1801. Political economy, like all other fields of knowledge, had a place in Bentham's classification of knowledge, and consequently a place in his conception of a comprehensive code of laws. It was the task of the utilitarian legislator to introduce measures which would increase the overall happiness (understood in terms of a balance of pleasure over pain), or, more centrally, which would prevent a decrease in happiness. This task would be undertaken by promoting what Bentham termed the four sub-ends of utility – subsistence, abundance, security and equality – using, where appropriate, sanctions (punishments and rewards), themselves composed of pain and pleasure, to discourage actions detrimental to the happiness of the community, and (to a lesser extent) to encourage those which were beneficial. More specifically, it was the task of the civil law to distribute rights and duties in such a way as to promote the four sub-ends of utility. Security consisted in the protection of the basic interests of the individual – his person, property, reputation, and condition in life – which constituted a major component of his well-being. Security was closely related to the notion of expectations, for it involved

both the present possession and the future expectation of possessing the property or other subject-matter in question. Without security, and thus the confidence to project oneself and one's plans into the future, there could be no civilized life. In short, security was a product of law, resulting from the imposition of rules on conduct.

The subject of political economy was more particularly concerned with subsistence and abundance, though the significance of security and equality should not be overlooked. For instance, without the security provided by law, no one would have an incentive to labour, and, therefore, to create wealth (abundance). Moreover, abundance itself was a security for subsistence, that is, the minimum quantity of resources which an individual needed to survive. Indeed, it was subsistence which had a prior claim on all resources in that an individual could be happy only if he were alive. Once wealth had been created, the principle of equality – in essence, the principle of diminishing marginal utility – demanded that it be distributed equally. Bentham argued that, if subsistence required £10 per annum, the most important £10 which an individual could possess was the first £10. Thereafter, each increment of £10 was worth something less than the previous increment. To put this another way, £10 given to an individual who had nothing constituted the difference between life and death, whereas £10 given to a rich man made hardly any difference at all. Bentham did not, however, advocate the levelling of property, for two reasons. First, if everyone began one morning with the same amount of property, by the end of the afternoon the intervening transactions would see inequality re-established. Second, the levelling of property would constitute an attack on security. Indeed, security, with its attendant expectations, was so important, that it was only in exceptional circumstances, such as providing subsistence to those who might otherwise starve to death, that it was legitimate to redistribute resources, and even here Bentham partly justified the redistribution on the grounds of security, in that such redistribution would render the property of the rich less liable to violent invasion by the poor.

In relation to abundance, or the creation of wealth, Bentham's basic principle was that of economic freedom. Each individual was most likely to be the best judge of his own interest, since he was most likely to be best informed about his own peculiar circumstances, and most likely to be motivated to act on that information in order to maximize his wealth, and thence his happiness. In a large number of areas in which government had traditionally intervened in economic matters, its intervention was counter-productive. Trade bounties, prohibitions, monopolies, and encouragements to population growth belonged to what Bentham termed the ‘non-agenda’ (although there might always be exceptions). Taking his lead from Smith, Bentham argued that since trade was limited by capital, government could not favour one branch of trade unless it discouraged another branch, since the capital applied to the former must be taken from the latter. In general, government was best advised not to interfere with the economy, and this included interference in the form of taxation. The imposition of taxation was a form of coercion, and all coercion was an evil in itself. As Bentham remarked: ‘The best use that government can make of money in the hands of the lawful possessors is: to leave it where it is’ (Bentham, 1989, p. 251). He argued that, in order to judge the utility of any element of public expenditure, one needed to compare the benefits produced by the expenditure with the burden produced by imposing an equivalent degree of taxation in the most aggravated form in which taxation was imposed. Hence, he recommended the immediate repeal of several particularly burdensome taxes – for instance those on legal proceedings, medicines, insurance, and newspapers (the latter constituting a tax on information). The taxation which remained should be imposed where there existed an ability to pay. Hence, the best form of taxation was that on consumption, followed by that on property and the transfer of property. As an alternative source

of public revenue, he advocated a revival of the medieval practice of escheat, whereby the state appropriated property where there was no other than a collateral heir. The money raised would be earmarked for a sinking fund, which would eventually redeem the national debt. The appropriation of collateral successions was a measure which Bentham believed could reconcile the otherwise conflicting demands of security and equality. Providing that individuals knew in advance that their potential to inherit would be limited according to law, they would not suffer any disappointed expectations, and their security would not be infringed. Apart from providing the background conditions of security which ensured that economic actors had the incentives to accumulate wealth (for instance security of person and property), there was, nonetheless, a limited ‘agenda’ for government, for instance to establish corn magazines to provide a security against dearth, to provide information, and to commission and disseminate research.

Monetary regulation

Following the suspension of payments in *specie* at the Bank of England in 1797, Bentham turned his attention to monetary regulation, devising his annuity note scheme, with the aim of redeeming the national debt. The annuity notes would in effect serve as paper currency, but at the same time earn compound interest, and, therefore, act as an investment. Depending on the prevailing rates of interest, holders of the notes would either use them as currency or horde them as savings. The government would issue the notes in order to buy up existing public debt, and thereafter successively reduce the rate of interest payable. The annuity notes as a circulating medium would replace an equivalent amount of bank notes, and lead to an earlier redemption of the national debt than would otherwise have been possible. It seems that Bentham abandoned the scheme because he did not, to his own satisfaction, solve the problem of inflation, which, he feared, would stifle the growth of national wealth and unfairly reduce the real value of fixed incomes.

In 1801 Bentham calculated that prices had increased by 50 per cent since 1760. He argued that this inflation had been caused by an increase in the amount of paper money in circulation. This increase was to be welcomed in that it represented a growth in national prosperity. However, it also represented an unfair tax on fixed incomes, and threatened a general bankruptcy. His remedy was to limit and to tax the issue of paper money by provincial banks, who were prone to over-issue bank notes since this was the main source of their profit. In return, a licensing system would be introduced which would, in effect, grant a monopoly to existing banks. In December 1801, in the extraordinary circumstances brought about by scarcity and dearth of provisions, he came to advocate legislative intervention in the economy in the form of the statutory imposition of a maximum price for wheat. This would have the immediate effect of bringing relief to the poor and security to the propertied, in that it would avoid the creation of a potentially revolutionary situation fuelled by the discontent of the destitute. Scarcity, he argued, could only permanently be remedied by the establishment of corn magazines and the promotion of emigration, both of population and of capital. In short, while favouring economic liberty as a leading principle, he was always prepared to consider state intervention should the principle of utility demand it.

Colonies

Bentham's opposition to the holding of colonies was grounded initially on economic arguments, though

he later developed political and constitutional objections to the practice. Given that the trade of a nation was limited by the quantity of capital it possessed, he argued that colony-holding could not bring any economic advantages. The extension of markets which the acquisition of colonies appeared to provide did not in itself affect the amount of trade. New markets were advantageous only to the extent that the profit made upon the capital employed in the new trade was greater than the profit made on the established trade. It was unlikely that the distant markets represented by colonies would offer a higher rate of return than those closer to home. Any benefit from a trade monopoly imposed on the produce of the colony was illusory, since a monopoly could not force the price of a commodity lower than the level to which it would be driven by competition, and it could not force anyone to produce a commodity at a loss. Finally, to the argument that trade with colonies was a source of revenue, Bentham responded that revenue could be raised on goods exchanged with all other countries, not just colonies, providing of course that the duties were not so high as to make smuggling attractive. The emancipation of colonies would also save the mother country the massive expense of defending them, particularly in time of war. Nonetheless, there were certain circumstances in which Bentham was prepared to defend the establishment of colonies. He approved the colonization of vacant lands in response to the pressure of population growth and the existence of an excess of capital in the mother country, and of colonial rule in countries where the native rulers were unfit to govern. The benefits, however, accrued to the colonists, and not to the mother country, and he recommended that dominion should be relinquished as soon as was practicable.

Political reform

By the 1820s Bentham was convinced that the only regime with an interest in enacting good legislation was a representative democracy. A crucial development took place around 1804 with the emergence in Bentham's thought of the notion of sinister interests, that is, the systematic development of the insight that rulers wished to promote not the happiness of the community, but their own happiness. There was no point in showing rulers what the best course of legislation might be unless they had an interest in adopting it. Only a legislature elected by a democratic suffrage had such an interest. Following the quashing of the panopticon scheme in 1803, Bentham became convinced that nothing worthwhile could be achieved through the existing political structure in Britain, or through similar regimes elsewhere. Having concentrated on questions of law reform from 1803, he was in the summer of 1809 prompted to compose material on political reform, eventually bearing fruit in *Plan of Parliamentary Reform* (1817). In this work he called for universal manhood suffrage (subject to a literacy test), annual parliaments, equal electoral districts, payment of MPs, and the secret ballot. Bentham then went a stage further and drew up a blueprint for representative democracy which would have abolished the monarchy, the House of Lords and any other second chamber, and all artificial titles of honour, and would have rendered government entirely open and, he hoped, fully accountable. These proposals were developed in astonishing detail in the magisterial *Constitutional Code* (partly printed 1827 and 1830, partly published 1830).

For Bentham the key principle of constitutional design was to ensure the dependence of rulers on subjects. Instead of the traditional theory of the separation of powers, he proposed lines of subordination, based on the ability of the superior to appoint and dismiss (in Bentham's terminology to locate and dislocate) the inferior, and to subject the inferior to punishment and other forms of ‘vexation’.

The supreme power or sovereignty in the state would be vested in the people, who held the constitutive power. Immediately subordinate to the people would be the legislature, elected by universal manhood suffrage, and subordinate to the legislature would be the administrative (that is, the executive) and judicial powers. The system of representative democracy was not an end in itself – the end was the greatest happiness – but was an indispensable means to that end, in that it was only under such a constitution that effective measures could be implemented to secure the good behaviour (appropriate aptitude) of officials and minimize the expense of government. The securities for official aptitude – otherwise termed securities against misrule – included the exclusion of factitious dignities (titles of honour), the economical auction (whereby officials made bids for the salary attached to the office), subjection to punishment at the hands of the legal tribunals of the state, the requirement to pass an examination, and, most importantly, publicity. Bentham went to great lengths to ensure that government would be open to public scrutiny, and thence subject to the force of the moral or popular sanction operating through the public opinion tribunal, which consisted in all those who commented on political matters, and of whom newspaper editors were the most important. Bentham saw the freedom of the press as a vital bulwark against misrule: hence his proposal to encourage the diffusion of literacy by making the suffrage dependent on a literacy test. These measures were intended to ensure that rulers would be so situated that the only way they could promote their own interest was by promoting the interest of the community.

Death and afterwards

Having lived in Lincoln's Inn from 1769 to 1792, he had then inherited his father's home in Queen's Square Place, Westminster, where he died on 6 June 1832. It was Bentham's wish that his body be dissected for the advancement of medical science, and that his remains then be used to create an ‘auto-icon’ or self-image. Bentham's auto-icon, assembled by his surgeon Thomas Southwood Smith (1788–1861), and consisting in a waxwork head mounted on Bentham's articulated skeleton and wearing his clothes, is now kept at University College London.

See Also

- utilitarianism and economic theory

Selected works

The Bentham Project, University College London, is preparing a new authoritative edition of *The Collected Works of Jeremy Bentham*, which, it is estimated, will run to 68 volumes. The 26th appeared in February 2006. The following volumes have been most extensively drawn upon in the compilation of this article:

1970. *An Introduction to the Principles of Morals and Legislation*, ed. J.H. Burns and H.L.A. Hart. London: Athlone Press.

1977. *A Comment on the Commentaries and A Fragment on Government*, ed. J.H. Burns and H.L.A. Hart. London: Athlone Press.

1989. *First Principles Preparatory to Constitutional Code*, ed. P. Schofield. Oxford: Clarendon Press.

1998. *'Legislator of the World': Writings on Codification, Law, and Education*, ed. P. Schofield and J. Harris. Oxford: Clarendon Press.

Where cited works have not appeared in *The Collected Works*, the standard source is the so-called Bowring edition: *The Works of Jeremy Bentham, published under the superintendence of his executor, John Bowring*, 11 vols. Edinburgh: William Tait, 1843. The standard source for Bentham's economic thought is *Jeremy Bentham's Economic Writings*, 3 vols., ed. W. Stark. London: George Allen & Unwin, 1952–54. A new authoritative edition is greatly needed.

Bibliography

Dinwiddy, J. 2004. *Bentham: Selected Writings of John Dinwiddy*, ed. W. Twining. Stanford: Stanford University Press.

Harrison, R. 1983. *Bentham*. London: Routledge and Kegan Paul.

Hart, H.L.A. 1982. *Essays on Bentham: Jurisprudence and Political Theory*. Oxford: Clarendon Press.

Kelly, P.J. 1990. *Utilitarianism and Distributive Justice: Jeremy Bentham and the Civil Law*. Oxford: Clarendon Press.

Lieberman, D. 2000. Economy and polity in Bentham's science of legislation. In *Economy, Polity, and Society: British Intellectual History, 1750–1950*, ed. S. Collini, R. Whatmore and B. Young. Cambridge: Cambridge University Press.

Postema, G.J. 1986. *Bentham and the Common Law Tradition*. Oxford: Clarendon Press.

Rosen, F. 2003. *Classical Utilitarianism from Hume to Mill*. London: Routledge.

Schofield, P. 2006. *Utility and Democracy: The Political Thought of Jeremy Bentham*. Oxford: Oxford University Press.

Semple, J. 1993. *Bentham's Prison: A Study of the Panopticon Penitentiary*. Oxford: Clarendon Press.

Warke, T. 2000. Multi-dimensional utility and the index number problem: Jeremy Bentham, J.S. Mill, and qualitative hedonism. *Utilitas* 12, 176–203.

How to cite this article

Schofield, Philip. "Bentham, Jeremy (1748–1832)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000109> doi:10.1057/9780230226203.0122

The New Palgrave Dictionary of Economics Online

bequests and the life cycle model

John Laitner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The standard life cycle model emphasizes a household's concerns over events within its lifetime, including providing for its own retirement and for its young children. However, in a more elaborate formulation, the household may care about its descendants when they are grown just as when they are young, causing the household to want to leave bequests. Its time horizon may expand to a dynastic scale, and new public policy implications, including so-called Ricardian neutrality, may emerge. Alternatively, bequests may signal non-market exchanges between parents and their adult children, perhaps arising to mitigate transactions costs or informational asymmetries.

Keywords

altruistic bequests; annuities; assortative mating; bequests; bequests and the life cycle model; implicit contracts; infinite horizons; *inter vivos* transfers; joy of giving; life cycle hypothesis; life-cycle model; non-market exchange; representative agent; retirement; Ricardian neutrality; strategic behaviour

Article

In the life-cycle model of household behaviour, each household expects a lifetime pattern of rising earnings in youth and middle age followed by retirement. Hence, households plan to save in their first segments of life in order to build resources to dissave, and from which to accrue interest income, during the last (Modigliani, 1986). The framework easily incorporates children, with consumption early in a household's life driven higher and saving for retirement perhaps delayed until middle age (Tobin, 1967). In a standard life-cycle model, parents plan for their own life and assume financial responsibility for their children until the latter reach adulthood (say, age 18 or 22) – but not beyond. Elaborations of the framework, on the other hand, extend parental concern, or interest in non-market transactions, to encompass a household's grown children. Such elaborations expand the scope of the life-cycle model to include bequests.

Conceptually, there are at least three broad categories of models in which bequests play a role. The first, which is often called the ‘altruistic model’, assumes that parents care about the well-being of their grown children. The second, which one might call the ‘joy of giving model’, assumes that parents derive pleasure from making transfers to their adult children's households but that the pleasure is not specifically dependent upon the children's utility gain. In the third formulation, parent-to-child emotional and social ties favour and facilitate non-market exchanges that may generate bequests – for example, bequests may emerge as payments to heirs for personal services rendered.

Altruistic model

A model with ‘altruistic bequests’ (Becker, 1974; Barro, 1974) extends to grown children parental concerns for minor children typical of standard life-cycle analyses.
Consider a specific example in which each household has one adult, raises one child, and lives two periods. Suppose that a household begun at time t has earnings y_t in youth but is retired in old age. It rears its child during its first stage of life; the child initiates its own household thereafter, with the descendant household passing its first stage of life as the parent household lives through its second stage. The time-t parent chooses consumption c_t^1 and c_t^2, respectively, for its two stages of life; derives utility $u(c_t^1, c_t^2)$ from this consumption; inherits i_t in youth; and transfers i_{t+1} in old age to its adult child. Let the interest rate be r. Given i_t and i_{t+1}, the parent household's *lifetime utility* is $U(\cdot)$ such that

$$U(i_t, i_{t+1}, y_t) \equiv \max_{c_t^1, c_t^2} u(c_t^1, c_t^2)$$

$$\text{subject to: } c_t^1 + \frac{c_t^2}{1+r} + \frac{i_{t+1}}{1+r} \leq i_t + y_t.$$

Let the parent household care δ times as much about its adult child's lifetime utility as about its own, δ^2 times as much about its grandchild's lifetime utility, and so on. Then the parent household's *dynastic utility* is

$$\sum_{s=0}^{\infty} \delta^{s} \cdot U(i_{t+s}, i_{t+s+1}, y_{t+s}).$$

If y_t=y all t, if institutions force bequests to be nonnegative, and if descendant households share the same preference ordering, we can characterize the time-t parent household's dynastic utility as $V(i_t, y)$ with

$$V(i_t, y) = \max_{i_{t+1} \geq 0} \left\{ U(i_t, i_{t+1}, y) + \delta \cdot V(i_{t+1}, y) \right\}.$$

(1)

If δ =0, we have a ‘pure life–cycle model’; if δ >0, we have an altruistic model in which positive bequests may emerge.

Laitner (1992) studies a second altruistic formulation, one allowing heterogeneous earning abilities. In terms of the framework above, a parent household with earnings y_t may know the random variable, say, $\tilde{y}$, from which the earnings of its descendants will be (independently, in the simplest case) sampled, but the parent cannot observe the sampling outcomes as it makes its bequest plans. Then dynastic utility is

$$V(i_t, y) = \max_{i_{t+1} \geq 0} \left\{ U(i_t, i_{t+1}, y) + \delta \cdot E[V(i_{t+1}, \tilde{y})] \right\},$$

(2)

where E[.] is the expectations operator.

Conceptually, a model with altruistic bequests provides an extension of the life-cycle model's parental concern for minor children's well-being to a more or less symmetric concern for grown children. Empirically, bequests and *inter vivos* transfers to adult children certainly occur in practice (Modigliani, 1986; Kotlikoff, 1988). The formulation with heterogeneous earnings predicts that bequests need not be universal but are most likely in the case of very prosperous parents. Social commentators frequently criticize bequests as a source of inequality, and the second point in the preceding sentence shows how bequests can contribute to cross-sectional dispersion of private wealth holdings. Bequests may have played a larger role in national wealth accumulation in the past, when long retirement spells were perhaps less common (Darby, 1979), and a model with both life-cycle saving and altruistic bequests can provide a framework for analysing the change (Laitner, 2001).

Loans for education fail to generate collateral for creditors; hence, parental and/or public support may be

important for ensuring efficient educational investment. Since benefits of education last long into adulthood, the model with altruistic bequests provides a logical framework for studying parental contributions (for example, Tomes, 1981). For instance, suppose that a child's earnings are an increasing, concave function f (.) of ability, a, and parental support for education, e, in the child's youth: $y_{t+1} = f(a_t, e_t)$. With homogeneous agents, $a_t = a$ all t, and (1) becomes

$$V(i_t, y) = \max_{i_{t+1} \geq 0, e_t \geq 0} \left\{ U(i_t, i_{t+1} + e_t \cdot (1 + r), y) + \delta \cdot V(i_{t+1}, f(a, e_t)) \right\}. \tag{3}$$

Then $i_{t+1}>0$ ensures efficient provision of education e_t regardless of the degree of parental concern for the child, δ. If, on the other hand, the tangible bequest is zero, investment in education can be inefficiently low.

A second prominent application of the altruistic model relates to fiscal policy. In a standard life-cycle model, when government turns from tax to deficit finance, national consumption may rise for a time, and the economy's long-run capital intensity may decline. Reformulating the life-cycle model to include altruistic bequests can overturn this result (for example, Barro, 1974). Debt service and repayment for current government borrowing may extend far beyond the life span of existing households, but not beyond the time horizon of dynasties. Maximization in (1) may yield an outcome in which the non-negativity constraint never binds, and Barro (1974) shows that in that case tax and deficit finance may have identical implications for aggregate consumption, capital accumulation, and interest rates. The latter equivalence is often referred to as ‘Ricardian neutrality’. (With heterogeneity of agents, as in formulation (2), non-negativity constraints will, on the other hand, tend to bind for some households – Laitner, 1992 – and then outcomes resembling Ricardian neutrality, while still possible, may be more in doubt – for example, Bernheim, 1987.)

Recent dynamic general equilibrium analyses of long-run growth and business cycles frequently employ the so-called ‘representative agent’ paradigm. Utility maximization over an infinite time horizon for a set of identical agents determines desired private consumption, saving, and labour supply. It seems fair to say that the life-cycle model with altruistic bequests, as in Barro (1974) and related papers, provides the most basic motivation for this approach.

Turning to empirical findings, the widespread existence of bequests (and *inter vivos* gifts) within family lines is well established (Modigliani, 1986; Kotlikoff, 1988). The pure life-cycle model does not seem able to explain as much national wealth as we see, and estate building seems a plausible explanation for the remainder (Kotlikoff, 1988). However, despite some consistency with the altruistic model, empirical evidence often seems to fail to support the implications of pervasive Ricardian neutrality (for example, Altonji, Hayashi and Kotlikoff, 1992; 1997). Long-standing evidence that households with multiple children tend in practice to divide their bequests equally (for example, Menchik, 1988) also seems contrary to implications of the simplest versions of the altruistic model. Perhaps altruistic bequest

behaviour is, in practice, concentrated among the highest-income households (as might be implied by formulation (2)).

Conceptually, as one considers couples instead of single parents, dynasties will interact through marriage. Assortative mating can preserve the logic of the analysis of the parthenogenetic theoretical construct (Laitner, 1991). Mating patterns that are random theoretically could, in contrast, expand to an overwhelming degree the scope of interpersonal connections that ‘neutralize’ incentives for self-interested behaviour (Bernheim and Bagwell, 1988).

The preceding formulations assume that a parent cares about his child but that the reverse is not true. A number of papers analyse two-sided altruism. Implicitly, in fact, all formulations with altruistic transfers are two sided – in model (1), for example, the parent cares about his child's utility relative to his own with a ratio of weights δ:1, while the child cares about his parent's utility relative to his own with weights in a ratio of 0:1. Unless parents and children agree on each other's relative importance, strategic behaviour may arise if agents have sufficient latitude in their set of feasible actions. In Laitner (1988), for instance, though parents and children care about each other, each may care less about the other than about itself – in which case a parent with low earnings may intentionally limit his life-cycle saving in youth in order to induce a larger transfer from his child during his retirement.

In the simplest life cycle model, a household saves before retirement in order to preserve an even level of consumption for the remainder of its life. An altruistic model extends the time frame of such behaviour: a household may use bequests (and *inter vivos* gifts) to promote evenness of consumption for its entire family line.

Joy of giving model

A joy-of-giving model provides a donor with pleasure that is independent of recipient utility and outside resources. For example, our two-period household above might solve

$$\max_{i_{t+1} \geq 0} \left\{ U(i_t, i_{t+1}, y_t) + W(i_{t+1}) \right\}, \tag{4}$$

with the new function $W(.)$ being unrelated to lifetime utility $U(.)$ or to recipient earnings y_{t+1}. In this approach, the parent household has preferences over its own lifetime consumption and the size of the bequest that it provides to its offspring, rather than over the descendant's consumption or utility. An example is Blinder (1974).

A possible advantage of this framework is that it does not require as great an ability on the part of donors to manifest empathy and rationality as the altruistic model. Another advantage is its analytic simplicity. In applications, authors may seek to specify the utility function $W(.)$ in a manner that can mimic, at least to some degree, the model with altruistic bequests (for example, Modigliani, 1986).

Exchange

The emotional ties of parents and their children may lead parents to prefer attentions from their grown children over services purchased in markets. Similarly, emotional bonds, tradition, or social norms may give trades between relatives lower transaction costs than those based on market contracts. Relatives may also have more complete information about one another than anonymous market participants do. Such factors may lead parents to make transaction and insurance arrangements with their grown children, and parental payments may take the form of bequests or *inter vivos* gifts.

In traditional societies, a household's eldest son might labour on his parents' farm, supporting his parents in their old age. In return, the son might expect to inherit the farm at his parents' death. One can view such a bequest as a payment for services, and neither altruistic nor joy-of-giving impulses on the part of parents (or their son) need be determinants of the transfer's size.

Bernheim, Shleifer and Summers (1985) provide a model in which elderly parents desire attention from their adult children, and the parents can be thought of as paying for the services through their bequest.

Many economists note the relative infrequency with which households purchase annuities. Transactions costs and adverse selection, due to private information about one's likely longevity, may be the underlying reason. In practice, parents may circumvent annuity markets by making implicit contracts with their grown children: in return for care and support in old age, the parents agree to bequeath their assets to their children. The children take the place of an insurance company: if their parents die young, the children's efforts receive generous remuneration; if the parents live a long time, their bequest may be small or non-existent, and the children's reward per hour of effort will be low. Kotlikoff and Spivak (1981) show that such arrangements can be surprisingly efficient. Friedman and Warshawsky (1990) illustrate a related point: they show that parents who have some inclination (either joy of giving or altruistic) to bequeath to their children may eschew market annuities with even modest transactions costs, preferring self-insurance, under which their children can inherit unspent parental resources.

See Also

- inheritance and bequests

Bibliography

Altonji, J.G., Hayashi, F. and Kotlikoff, L.J. 1992. Is the extended family altruistically linked? Direct tests using micro data. *American Economic Review* 82, 1177–98.

Altonji, J.G., Hayashi, F. and Kotlikoff, L.J. 1997. Parental altruism and *inter vivos* transfers: theory and evidence. *Journal of Political Economy* 105, 1121–66.

Barro, R.J. 1974. Are government bonds net worth? *Journal of Political Economy* 82, 1095–117.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82, 1063–93.

Bernheim, B.D. 1987. Ricardian equivalence: an evaluation of theory and evidence. In *NBER Macroeconomics Annual* 2, ed. S. Fischer. Cambridge, MA: MIT Press.

Bernheim, B.D. and Bagwell, K. 1988. Is everything neutral? *Journal of Political Economy* 96, 308–38.

Bernheim, B.D., Shleifer, A. and Summers, L.H. 1985. The strategic bequest motive. *Journal of Political Economy* 93, 1045–76.

Blinder, A.B. 1974. *Toward an Economic Theory of Income Distribution*. Cambridge, MA: MIT Press.

Darby, M.R. 1979. *The Effects of Social Security on Income and the Capital Stock*. Washington, DC: American Enterprise Institute.

Friedman, B.M. and Warshawsky, M.J. 1990. The cost of annuities: implications for saving behavior and bequests. *Quarterly Journal of Economics* 105, 135–54.

Kotlikoff, L.J. 1988. Intergenerational transfers and savings. *Journal of Economic Perspectives* 2, 41–58.

Kotlikoff, L.J. and Spivak, A. 1981. The family as an incomplete annuities market. *Journal of Political Economy* 89, 372–91.

Laitner, J. 1988. Bequests, gifts and social security. *Review of Economic Studies* 55, 275–99.

Laitner, J. 1991. Modeling marital connections among family lines. *Journal of Political Economy* 99, 1123–41.

Laitner, J. 1992. Random earnings differences, lifetime liquidity constraints and altruistic intergenerational transfers. *Journal of Economic Theory* 58, 135–70.

Laitner, J. 2001. Secular changes in wealth inequality and inheritance. *Economic Journal* 111, 691–721.

Menchik, P.L. 1988. Unequal estate division: is it altruism, reverse bequests, or simply noise? In *Modelling the Accumulation and Distribution of Wealth*, ed. D. Kessler and A. Masson. Oxford: Clarendon Press.

Modigliani, F. 1986. Life cycle, individual thrift and the wealth of nations. *American Economic Review* 76, 297–313.

Tobin, J. 1967. Life cycle saving and balanced growth. In *Ten Economic Studies in the Tradition of Irving Fisher*, ed. W. Fellner. New York: Wiley.

Tomes, N. 1981. The family, inheritance and the intergenerational transmission of inequality. *Journal of Political Economy* 89, 928–58.

How to cite this article

Laitner, John. "bequests and the life cycle model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000317> doi:10.1057/9780230226203.0123

The New Palgrave Dictionary of Economics Online

Bergson, Abram (1914 – 2003)

Michael Ellman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adjusted factor cost method; Arrow's theorem; Association for Comparative Economic Studies; Bergson gap; Bergson, A.; material product system (MPS); Russian Research Centre (Harvard); social welfare function; Soviet growth record

Article

Bergson was the intellectual father of US studies of the Soviet economy during the Second World War as chief of the Russian Economic subdivision of the Office of Strategic Services (OSS). After the war he played the major role in founding the US tradition of description and analysis of Soviet economic institutions, measurement of Soviet economic growth and evaluation of that growth. He had earlier made a major contribution to the development of welfare economics. His work on the Soviet economy was marked by a combination of encyclopaedic knowledge of Soviet statistics, theoretical analysis and immense industry. It had an enormous influence on the development of US studies of the Soviet economy and established itself as the dominant paradigm in that field.

Bergson's main contribution to the study of the Soviet economy concerned the measurement of Soviet economic growth. The result of the combination of the ‘propaganda of success’ with Soviet economic institutions and the material product system (MPS) method of calculating national income was that the data on economic growth published by the Soviet authorities were both incredible and clearly non-comparable with the data on economic growth of other countries. Bergson both developed a method which enabled internationally comparable national income statistics and growth rates to be calculated for the USSR and applied it to the USSR for 1928–55. The method was the ‘adjusted factor cost’ method. In essence it consisted of adjusting actual Soviet transactions prices so as to bring them into line with the prices that would have been observed if the USSR's prices had been determined in accordance with neoclassical theory. These adjusted prices were then used as weights to aggregate the physical output series of branches and sectors of the economy as known from Soviet official data into a system of national accounts (SNA)-type aggregate. This had the great advantage of producing data comparable to SNA data and hence suitable for international comparisons. At the same time, Bergson argued, this

procedure enabled a ‘production potential’ and possibly even a welfare interpretation to be given to the resulting national income data.

The development of this method and its application to the USSR for the period 1928–55 were enormous achievements. They clearly indicated that assessment of socialist economies did not have to remain at the level of ideological confrontation but was amenable to rational discourse and scientific inquiry. Both the method and its results were controversial. The rationality of the adjusted factor cost prices, the representativeness of the physical products selected, the huge data requirements and skilled labour inputs necessary to apply the method, the relevance of neoclassical theory for interpreting Soviet economic data, and the accuracy of the picture of the Soviet economy resulting from application of the method, all came under fire. Others used different methods of generating internationally comparable data (for example, the physical indicators method, or scaling up from net material product, NMP, to GNP using data for the missing sectors).

In welfare economics Bergson is famous for his 1938 paper which defined and discussed the concept of an individualistic social welfare function. The latter enables necessary conditions for an economic optimum to be calculated without the assumption of cardinal utility. This concept was subsequently utilized and developed by Samuelson and became an integral part of the welfare economics literature. Its usefulness remains a matter of controversy. According to Samuelson's contribution to the Bergson Festschrift it was a major contribution, a ‘flash of lightning’ after which ‘all was light’ in the hitherto extraordinarily confused subject of welfare economics. A number of opinions of a less positive kind can be found in M. Dobb (1969). Bergson also wrote on socialist economics and Arrow's Impossibility Theorem.

Besides his purely academic work on the Soviet economy, Bergson, with his OSS experience, played a major role in establishing and maintaining the close links between US academic studies of the Soviet economy and the intelligence community and other branches of the federal government. Besides being a professor of economics for many years, first at Columbia and then at Harvard, he was director of the Harvard Russian Research Center (1964–8, 1969–70), consultant to the RAND Corporation, member and subsequently chairman of the Social Science Advisory Board of the US Arms Control and Disarmament Agency, and consultant to various federal agencies. In addition, he served as president of the Association for Comparative Economic Studies and several times testified before the US Congress.

Many years after Bergson's publications, access to Soviet economic archives demonstrated the significance and accuracy of Bergson's analysis of discrepancies in Soviet labour statistics (‘the Bergson gap’). It also demonstrated the usefulness of his approach for studying the Soviet national accounts during the Second World War.

Bergson made a major contribution to 20th-century economics by establishing a school of economists who transformed the study of the Soviet economy, hitherto a reserve of partisan émigré and committed writers, into a field of sober academic inquiry.

See Also

- social welfare function
- Soviet growth record
- welfare economics

Selected works

Welfare economics

1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52, 310–34.

1966. *Essays in Normative Economics*. Cambridge, MA: Harvard University Press.

1982. *Welfare, Planning and Employment*. Cambridge, MA: MIT Press.

Soviet economic institutions

1944. *The Structure of Soviet Wages*. Cambridge, MA: Harvard University Press.

1964. *The Economics of Soviet Planning*. New Haven: Yale University Press.

1984. Income inequality under Soviet socialism. *Journal of Economic Literature* 22, 1052–99.

Measurement of Soviet economic growth

1950. Ruble prices and the valuation problem. *Quarterly Journal of Economics* 64, 408–41.

1953. *Soviet National Income and Product in 1937*. New York: Columbia University Press.

1954. (With H. Heymann, Jr.) *Soviet National Income and Product 1940–48*. New York: Columbia University Press.

1961. *The Real National Income of Soviet Russia since 1928*. Cambridge, MA: Harvard University Press.

Evaluation of Soviet economic growth

1968. *Planning and Productivity under Soviet Socialism*. New York: Columbia University Press.

1978. *Productivity and the Social System – the USSR and the West*. Cambridge, MA: Harvard University Press.

Bibliography

Birman, I. 1985. The Soviet economy: alternative views. *Survey* 29(2), 102–15.

Birman, I. 1986. The Soviet economy: alternative views. *Russia* 12, 60–74.

Birman, I. 1989. *Personal Consumption in the USSR and the USA*. Basingstoke, UK: Macmillan.

Comparative Economic Studies. 2005. Performance and efficiency under socialism: studies in honor of Abram Bergson. Vol. 47, 239–502.

Dobb, M. 1969. *Welfare Economics and the Economics of Socialism*. Cambridge: Cambridge University Press.

Hanson, P. 1971. East–West comparisons and comparative economic systems. *Soviet Studies* 22, 327–43.

Hardt, J. 2004. Abram Bergson and the development of Soviet economic studies. *Problems of Post-Communism* 51(4), 34–9.

Harrison, M. 1996. Measuring Soviet GNP. Ch. 3 in M. Harrison, *Accounting for War*. Cambridge: Cambridge University Press.

Holzman, F. 1957. The adjusted factor cost standard of measuring national income: comment. *Soviet Studies* 9(1), 32–6.

Khanin, G. 1993. *Sovetskii ekonomicheskii rost: analiz zapadnykh otsenok*. Novosibirsk: EKOR.

Marer, P. 1985. *Dollar GNPs of the USSR and Eastern Europe*. Baltimore, MD: Johns Hopkins University Press.

Rosefielde, S., ed. 1981. *Economic Welfare and the Economics of Soviet Socialism*. Cambridge: Cambridge University Press. (The Bergson Festschrift – a full bibliography of Bergson's work in 1936–80 can be found on pp. 334–7 of this book.).

Samuelson, P. 2005. Abram Bergson, economist. *Economic Journal* 115(501), F130–F133.

Wiles, P. 1955. Are adjusted roubles rational? *Soviet Studies* 7(2), 143–60.

How to cite this article

Ellman, Michael. "Bergson, Abram (1914–2003)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_B000110> doi:10.1057/9780230226203.0124

The New Palgrave Dictionary of Economics Online

Berle, Adolf Augustus, Jr. (1895 – 1971)

John Kenneth Galbraith
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Berle, A. A.; corporations; entrepreneurship; Keynes, J. M.; McConnell, C. R.; New Deal; ownership vs. control; Samuelson, P. A.; Schumpeter, J. A.; Twentieth Century Fund

Article

A graduate at an early age of Harvard College and the Harvard Law School, Berle served in Army Intelligence in World War I and on the American delegation to the Paris Peace Conference, from which he emerged to denounce the terms of the Treaty, as did Keynes, though to a lesser audience. After practising law in New York, he joined the law faculty of Columbia University, where he became a member of the famous Brains Trust of Franklin D. Roosevelt. He was a close adviser of Roosevelt's, both before and after the latter's election to the Presidency.

In the later New Deal years, Berle served as an Assistant Secretary of State, then a senior position in the Department, and thereafter as ambassador to Brazil. In the years following World War II, he was chairman of the Liberal Party in New York and the long-time head of the Twentieth Century Fund, a foundation engaged in the active sponsorship of research in economic and social issues.

Berle's major contribution to economics, made in 1932 in conjunction with Gardiner C. Means in *The Modern Corporation and Private Property*, was in showing that authority in the modern large business enterprise moves ineluctably away from the owners of property to the managers and that by the time of research for the book the process was already far advanced. As a conclusion for conventional economics this, it is not too much to say, ranked in inconvenience with that of Keynes. Ownership no longer conveyed power in the great enterprise. Profit maximization was now by managers, not on behalf of themselves but for others largely unknown or, in pay and perquisites, for the managers themselves. Berle's conclusions also denied the independent, self-motivated, heroic role of the entrepreneur as offered in conventional economics, notably by Schumpeter.

Berle's contribution came from outside the conventional boundaries of the profession – from, of all things, a lawyer. Perhaps for this reason its importance was discounted, even denied, by many economists. In recent times, however, the truth of Berle's contentions has been recognized as personal

profit maximization of managers – salaries, diverse perquisites, stock options, golden parachutes – has become one of the accepted scandals of the time. Nonetheless, Berle's role as one of the major innovating figures in economics has never been adequately recognized. In his textbook Paul Samuelson acknowledges *The Modern Corporation* as a classic; in Campbell R. McConnell's *Economics*, the most widely used text in the United States, Berle's name does not even appear.

In his later years Berle returned in a perceptive and informative way to the subject of power, though not with the innovative force of his earlier work.

Selected works

1932. (With G.C. Means.) *The Modern Corporation and Private Property*. New York: The Commerce Clearing House.

1959. *Power Without Property: A New Development in American Political Economy*. New York: Harcourt, Brace & World.

1963. *The American Economic Republic*. New York: Harcourt, Brace & World.

1969. *Power*. New York: Harcourt, Brace & World.

How to cite this article

Galbraith, John Kenneth. "Berle, Adolf Augustus, Jr. (1895–1971)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000112>
doi:10.1057/9780230226203.0125

The New Palgrave Dictionary of Economics Online

Bern á cer, Germ á n (1883 – 1965)

Mauro Boianovsky

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bernácer contributed to macroeconomics the concept of ‘disposable funds’ and a new theory of interest. A lag between received and disbursed income underlies his view that aggregate equilibrium in the goods market emerges only if the amount of disposable funds is the same at the beginning and at the end of the period. Bernácer also argued that the rate of interest was determined outside the production system by land purchases and sales in the assets market. Economic fluctuations are decided by oscillations in the amount of disposable funds determined by the interaction between the markets for goods and for old assets.

Keywords

aggregate equilibrium; Bernácer, G.; Böhm–Bawerk, E. von; business cycles; cash-in-advance constraint models; crowding out; disposable funds; economic geography; economic integration; effective demand; George, H.; interest rate determination; IS–LM model; Keynes, J. M.; liquidity constraints; money; natural rate and market rate of interest; Physiocracy; Robertson, D.; saving–investment equality; School of Salamanca; speculative markets; stabilization policies; stocks and flows; Turgot, A. R. J.

Article

Bernácer was born in Alicante, Spain, on 29 June 1883, and died on 22 May 1965 in the same city. He may be regarded as the first major monetary economist in the Spanish language since the School of Salamanca in the 16th century. Bernácer completed his studies at the Alicante School of Commerce (Escuela Superior de Comercio de Alicante) in 1901, where he was awarded the chair of industrial physics (Tecnología Industrial) in 1905. In that same year he started working on his big book *Sociedad y Felicidad – Ensayo de Mecánica Social*, which shows the influence of his physics background in the study of the economic aspects of social life, especially his distinction between the ‘static and dynamics of wealth’ in the study of ‘social problems’ such as business cycles and unemployment. That book was

eventually published in 1916, some time after a study tour of eight months that had taken him to several European countries in 1911. In the next ten years, some of the main ideas presented in incipient form in *Sociedad y Felicidad* were further developed in two publications by Bernácer. His 1922 essay introduced into the economic literature the concept of ‘disposable funds’ (‘disponibilidades’) and its implications for the treatment of the demand for money and monetary dynamics. Bernácer sent 150 copies of that essay (with a French summary) to prominent economists and journals around the world. His 1925 book advanced a new approach to the origins and determination of interest as a variable decided outside the production system.

In the early 1930s Bernácer moved to Madrid to become the first director of the Research Service of the Bank of Spain. His appointment was probably influenced by his long 1929 article about the determination of the exchange rate as an equilibrium variable, in which he discussed in detail how to stabilize the external and internal values of the Spanish peseta and the conditions for returning to the gold standard system. He continued to teach, this time as professor of physics and chemistry at the School of High Commercial Studies of Madrid (Escuela de Altos Estudios Mercantiles). In 1940 long extracts from Bernácer's 1922 article were translated into English and published in *Economica* with a commentary by Dennis Robertson, who had been one of the recipients of that article in the 1920s. Robertson's article made Bernácer known to the Anglo-Saxon world and led him to restate the main theoretical and methodological features of his approach to monetary economics in a volume published in 1945. In the 1950s he wrote his last two books, dealing with economic integration and economic geography (1953) and summing up his views about economic dynamics and economic reform (1955). At about this time Bernácer retired from both his appointments as professor in Madrid and as director of research at the Bank of Spain.

Period analysis and disposable funds

Bernácer's main contribution to economics is his analysis of the role played by money in the determination of economic variables such as income, employment, the rate of interest and the rate of exchange. He introduced the concept of a lag between received and disbursed income, which provided the starting-point of his discussion of aggregate disequilibrium in the market for goods. Bernácer's lag probably influenced the well-known Robertsonian related lag between received and disposable income. It follows from his concept of disposable funds (A) held at the beginning of the economic period, which, when added to the income (R) received during the period, give the upper limit of effective demand ($A+R$). Money balances are functionally classified into three grades, from minimum to maximum degree of disposability: (*a*) money demand by families to meet consumption; (*b*) money demand by businessmen for the conduct of their enterprises; and (*c*) new savings which have not yet been put by their owners to remunerative employment. Bernácer used the phrase ‘disponibilidades’ to refer to the last two classes. In order to determine the flow of ‘effective demand’ (D) it is necessary to subtract from A the amount of disposable funds left at the end of the period (A'), which gives the equation $R+(A-A')=D$, or, since R is identical with output P, the equation $P+(A-A')=D$. The last equation indicates that there is aggregate equilibrium (in the sense that production is equal to effective demand and the output produced is sold at the expected price) if the amount of disposable funds is the same at the beginning and at the end of the period $(\Delta A=0)$. The key to Bernácer's monetary economics is his notion that the spending

decisions of economic agents (firms and families alike) in any given period of time are constrained by the amount of money they possess at the outset of that period. Bernácer was probably the first to introduce the main elements of what would become known in the literature as the ‘cash-in-advance constraint’ models developed in the 1960s.

Bernácer's approach to the business cycle was based on his distinction between the market for goods (‘circulación productiva’), which decides the price level, and the market for ‘valores de renta’ or income-yielding assets (‘circulación especulativa’ or ‘circulación financeira’), where the rate of interest is determined. Similar distinctions between aggregate markets for flows and stocks respectively would be deployed later in macroeconomic models put forward by John Hicks (IS–LM model), James Tobin and others. The interplay between those two markets explains fluctuations in income and employment in Bernácer's framework. The use of disposable funds to buy ‘valores de renta’ in the financial or speculative market does not change the condition of disposable funds, as they remain disposable in the hands of the sellers of assets. On the other hand, the use of disposable funds to purchase consumption goods and new capital goods brings about a change in their degree of disposability, as they are turned into money income of the individuals involved in the production of goods. This constitutes ‘effective demand’, as opposed to ‘potential demand’ that does not involve a change in liquidity. Aggregate equilibrium can now be also described by the equality between saving and investment, which is the case if the saving flow is not directed to the purchase of ‘valores de renta’. Economic fluctuations result from the opposite effects on the price level and the rate of interest of changes in disposable funds. When ΔA is negative in the upswing, prices of consumption goods are higher than anticipated and, since wages and salaries are temporarily fixed, employers will see their ‘residual profits’ increase. The ensuing stimulus to production and employment will cease when, under the impact of an increasing shortage of disposable funds in the ‘speculative market’, the rate of interest rises and saving is gradually directed to that market. This way, ΔA becomes positive, which explains the upper turning point of the business cycle. During the downswing, unanticipated falling prices bring about losses, which contributes (together with the constraint represented by a reduction of firms’ liquidity) to a contraction in production and employment. The depression is characterized by widespread ‘forced [or involuntary] unemployment’ (‘paro forsozo’), which is not solved by money-wage reductions, since lower wages will bring about a further fall in consumption demand and ensuing price reductions.

The speculative market and the rate of interest

The main factor in Bernácer's account of the business cycle is not the variability of investment demand by entrepreneurs, but the savers’ decisions on how to allocate their disposable funds – purchase of new capital goods in the goods market or of old assets in the speculative market. The banking and credit system is incidental to Bernácer's framework, which is different from the well-known Wicksellian distinction between the ‘natural’ and the ‘market’ rates of interest. Bernácer's explanation of macroeconomic disequilibrium is based on another sort of divergence, that is, on differences between the rate of interest decided by the expected rate of return on new capital goods on one side, and the rate of interest determined by the relative yields of ‘valores de renta’ in the speculative market. The notion that the rate of interest is determined outside the system of current production is a crucial feature of the Bernácerian theoretical system. He argued that the rate of interest is determined not by the scarcity of capital goods as such, but by the scarcity of disposable funds. Moreover, given the identity between

aggregate income and output, the rate of interest cannot be determined simply by saving and investment: if the disposable funds were used only to purchase the current output (of consumption and capital goods), the saving flow would necessarily be identical with the output of new capital goods, with no scarcity of funds in that market. The rate of interest can be positive only if a scarcity of disposable funds comes about because of the possibility of employing them outside the production system, that is, in the speculative market. The problem of the origin and determination of interest, according to Bernácer, consists in the search for an asset able to yield a ‘free’ rent without any production costs. He found it in land (in the broad sense of agricultural and urban land, as well as mines), not because of its productivity, but because it has a price and is exchangeable for other assets through money. In particular, the rate of interest is the determined variable in the equation relating its value to the price and the rent of land. Land, however, is not capital, and its purchase is not a real investment, since money remains disposable; hence, Bernácer explained how land's ability to produce rent is transmitted to other applications of money – especially to new capital goods – through the equilibrium between the marginal rates of return of old and new assets in the market. Such a mechanism, however, cannot work if the rate of return of investment in new capital goods falls to zero or below (which, of course, cannot happen to land and other income-yielding assets) in the depression, as pointed out by Bernácer. After he had put forward the main elements of his interest theory in 1916, Bernácer noticed several similarities with what Böhm-Bawerk used to call Turgot's ‘fructification theory’ of interest, but observed that, in contrast with Turgot's, his approach was not based on the Physiocratic framework.

Bernácer would claim, after the publication of Robertson's article in 1940, that the dynamic approach to monetary economics introduced in his 1922 essay was the source of Robertson's own formulation of period analysis in 1926 and, via Robertson, of the ‘fundamental equations’ of Keynes's 1930 *Treatise on Money*. Whereas there are some grounds to substantiate Bernácer's claim, it should be noted that the economic policy conclusions he drew from his theoretical framework are far apart from those advocated by Robertson or Keynes. Bernácer was critical of attempted stabilization policies of both fiscal and monetary sorts, because of the crowding out effect and of the (destabilizing) impact of monetary and credit changes on prices. Instead, he believed that the market economy was an essentially efficient institution, except for the existence of the speculative market for income-yielding assets that kept the economy in a chronic state of unemployment. Bernácer's suggested solution was to make the amount of disposable funds constant by suppressing that market through the legal prohibition of the sale of land, which would bring the rate of interest to zero. Although this is somewhat reminiscent of Henry George's reform proposals in the 19th century, it should be noted that Bernácer supported neither George's tax reform nor George's approaches to economic fluctuations and the determination of interest. It is likely that Bernácer's idiosyncratic ideas about economic reform, as well as his rejection of macroeconomic stabilization policies, contributed to distracting interest from the depth of his economic theory and to explaining its relative lack of influence in Spain throughout his lifetime.

See Also

- Spain, economics in
- George, Henry
- Robertson, Dennis

- Turgot, Anne Robert Jacques, Baron de L'Aulne

Selected works

1916. *Sociedad y Felicidad – Ensayo de Mecánica Social*. Madrid: F. Beltrán.

1922. La teoría de las disponibilidades como interpretación de las crisis económicas y del problema social. *Revista Nacional de Economía* No. 40, 535–62.

1925. *Interés del Capital*. Alicante: Lucentum.

1926. El ciclo económico. *Revista Nacional de Economía* Nos. 66 and 67, 3–30 and 155–79.

1929. La técnica del retorno al patrón oro (I, II, III). *Revista Nacional de Economía* Nos. 83, 84 and 85, 3–15, 223–39 and 405–18.

1945. *La Doctrina Funcional del Dinero*. Madrid: Consejo Superior de Investigaciones Cientificas.

1950. Money and freedom. *Kyklos* 4(2/3), 123–39.

1953. *La Doctrina del Gran Espacio Económico*. Madrid: Aguilar.

1955. *Una Economía Libre, sin Crisis, y sin Paro*. Madrid: Aguilar.

2005. *En Torno a la Obra y Figura del Economista Germán Bernácer- Selección de Escritos y Conferencias*. Alicante: Caja de Ahorros del Mediterráneo.

Bibliography

Almenar, S. 1999. Keynes's economic ideas in Spain before the *General Theory*: spread, ‘anticipations’ and parallels. In *The Impact of Keynes on the Economics in the 20th Century*, ed. L. Pasinetti and B. Schefold. Cheltenham: Elgar.

Boianovsky, M., Dar, H., Presley, J. and Brañas-Garza, P. 2006. Cambridge and the Spanish connection: the contribution of Germán Bernácer. *History of Political Economy* 38, 407–36.

Freyre, J. 1957. Review of Bernácer (1955), *American Economic Review* 47, 436–8.

Pozuelo y Barnuevo, J. 1965. Germán Bernácer (1883–1965). *Revue d ' Économie Politique* 75, 1231–5.

Robertson, D. 1926. *Banking Policy and the Price Level*. London: P.S. King & Son Ltd.

Robertson, D. 1940. A Spanish contribution to the theory of fluctuations. *Economica*, n.s. 7(February), 50–65.

Ruiz, G. 1984. *Germán Bernácer – Un Economista Anticipativo*. Madrid: Pirámide.

Savall, H. 1975. *G. Bernácer – L ' Héterodoxie en Science Économique*. Paris: Dalloz.

Wallich, H. 1947. Review of Bernácer (1945). *Review of Economic Statistics* 29, 65–6.

How to cite this article

Boianovsky, Mauro. "Bernácer, Germán (1883–1965)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000313> doi:10.1057/9780230226203.0126

The New Palgrave Dictionary of Economics Online

Bernoulli, Daniel (1700 – 1782)

S.L. Zabell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bernoulli, D.; Bernoulli, N.; Cramer, G.; D'Alembert, C.; Euler, L.; Goldbach, C.; Laplace, P. S.; logarithmic utility; maximum likelihood; moral expectation; probability; St Petersburg paradox; utility

Article

Swiss mathematician and theoretical physicist; born at Groningen, 8 February 1700; died at Basel, 17 March 1782.

Daniel Bernoulli was a member of a truly remarkable family which produced no fewer than eight mathematicians of ability within three generations, three of whom – James 1 (1654–1705), John 1 (1667–1748) and Daniel – were luminaries of the first magnitude.

Although initially trained in medicine, in 1725 Daniel Bernoulli accepted a position in mathematics at the newly founded Imperial Academy in St Petersburg, but returned to Basel in 1733, holding successively the chairs in anatomy and botany, physiology (1743), and physics (1750–77). He was elected to membership in all of the major European learned societies of his day, including those of London, Paris, Berlin and St Petersburg, and maintained an extensive scientific correspondence which included both Euler and Goldbach.

Original in thought and prolific in output, Bernoulli worked in many areas but his most important contributions were to the fields of mechanics, hydrodynamics and mathematics. He enjoys with Euler, his close friend from childhood, the distinction of having won or shared no fewer than ten times the annual prize of the Paris Academy. His masterpiece, the *Hydrodynamica* (1738), contains a derivation of the *Bernoulli equation* for the steady flow of a non–viscous, incompressible fluid, and the earliest mathematical treatment of the kinetic theory of gases, including a derivation of Boyle's Law.

Bernoulli also made important contributions to probability and statistics, including an early application of the method of maximum likelihood to the theory of errors and an investigation of the efficacy of smallpox inoculation (Todhunter, 1865, ch. 11). Nevertheless, his best-known contribution to this subject is unquestionably his 1738 paper ‘Specimen theoriae novae de mensura sortis’, which discusses utility, ‘moral expectation’ and the St Petersburg paradox.

The St Petersburg paradox (so called because Bernoulli's paper appeared in the *Commentarii* of the St Petersburg Academy) concerns a game, first suggested by Nicholas Bernoulli (Daniel's cousin) in correspondence with Montmort: a coin is tossed n times until the first head appears; 2^n ducats are then paid out. Paradoxically, the mathematical expectation of gain is infinite although common sense suggests that the fair price to play the game should be finite.

Bernoulli proposed that the paradox could be resolved by replacing the mathematical expectation by a moral expectation, in which probabilities are multiplied by personal utilities rather than monetary prices. Arguing that incremental utility is inversely proportional to current fortune (and directly proportional to the increment in fortune), Bernoulli concluded that utility is a linear function of the logarithm of monetary price, and showed that in this case the moral expectation of the game is finite.

Strictly speaking, Bernoulli's advocacy of logarithmic utility did not ‘solve’ the paradox: if utility is unbounded, then it is always possible to find an appropriate divergent series. Nor was he the first to adopt such a line of attack; the Swiss mathematician Gabriel Cramer had earlier written to Nicholas Bernoulli in 1728, noting that if utility were either bounded or proportional to the square root of monetary price, then the moral expectation would be finite. But it was via Bernoulli's paper that the utility solution entered the literature, and despite initial (and eccentric) criticism by D'Alembert, by the 19th century most treatises on probability would contain a section on moral expectation and the paradox.

An English translation of Bernoulli's 1738 paper on the St Petersburg paradox was published in *Econometrica* 22 (1954), 23–36, and is reprinted in *Precursors in Mathematical Economics: An Anthology*, ed. W.J. Baumol and S.M. Goldfeld, Series of Reprints of Scarce Works on Political Economy, No. 19, London: London School of Economics and Political Science, 1968, pp. 15–26. An English translation of Bernoulli's paper on maximum likelihood estimation appears in *Biometrika* 48 (1961), 1–18.

For further biographical information about Daniel Bernoulli and a detailed scientific assessment of his work, see the article by Hans Straub in *Dictionary of Scientific Biography*, vol. 2 (1970). The DSB also contains excellent entries on several other members of the Bernoulli family. Eric Temple Bell's *Men of Mathematics* (1937) contains a spirited, if not necessarily reliable, account of the Bernoullis.

Todhunter (1865, ch. 11) is still valuable as a summary of Bernoulli's work in probability; Todhunter's book is, as Keynes justly remarked, ‘a work of true learning, beyond criticism’. For further information on Bernoulli's contributions to probability and statistics, see also Sheynin (1970; 1972) and Maistrov (1974, pp. 106–7, 110–18). The dispute with D'Alembert is discussed by Baker (1975, pp. 172–5); see also Pearson (1978, pp. 543–55, 560–65) and Daston (1979, pp. 259–79).

Useful discussions of Bernoulli's paper on the St Petersburg paradox include Leonard J. Savage (1954, pp. 91–5) and J.M. Keynes (1921, pp. 316–20). The mathematician Abel once wrote that one should read the masters and not the pupils; those who wish to follow Abel's advice will find challenging but rewarding Laplace's discussion of moral expectation in his *Théorie analytique des probabilités* (1812, ch. 10: ‘De l'espérance morale’).

The literature on the St Petersburg paradox up to 1934 is surveyed in Karl Menger (1934); an English translation of Menger's paper appears in M. Shubik (ed., 1967). For a discussion of the St Petersburg paradox in the context of an axiomatization of utility and probability other than that of Ramsey and Savage, see Jeffrey (1983, pp. 150–5). The paradox still continues to inspire interest and analysis; a

recent example is Martin-Lof (1985).

Bibliography

Baker, K.M. 1975. *Condorcet: From Natural Philosophy to Social Mathematics*. Chicago: University of Chicago Press.

Bell, E.T. 1937. *Men of Mathematics*. New York: Simon & Schuster; Harmondsworth: Pelican Books, 1953.

Bernoulli, D. 1738. *Hydrodynamica, sive de viribus et motibus fluidorum commentarii* …. Strasbourg: Johann Reinhold Dulsseker.

Daston, L.J. 1979. D'Alembert's critique of probability theory. *Historia Mathematica* 6, 259–79.

Jeffrey, R.C. 1983. *The Logic of Decision*, 2nd edn. Chicago: University of Chicago Press.

Keynes, J.M. 1921. *A Treatise on Probability*. London: Macmillan.

Laplace, P.S. 1812. *Théorie analytique des probabilités*. Paris. 2nd edn, 1814; 3rd edn, 1820.

Maistrov, L.E. 1974. *Probability Theory: A Historical Sketch*. New York: Academic Press.

Martin-Lof, A. 1985. A limit theorem which clarifies the ‘Petersburg paradox’. *Journal of Applied Probability* 22, 634–43.

Menger, K. 1934. Die Unsicherheitsmoment in der Wehrtlehre. *Zeitschrift für Nationalökonomie* 5, 459–85. Trans. in *Essays in Mathematical Economics in Honor of Oskar Morgenstern*, ed. M. Shubik, Princeton: Princeton University Press, 1967.

Pearson, K. 1978. *The History of Statistics in the 17th and 18th Centuries*. New York: Macmillan.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Sheynin, O.B. 1970. Daniel Bernoulli on the Normal Law. *Biometrika* 57, 99–102.

Sheynin, O.B. 1972. D. Bernoulli's work on probability. *RETE Strukturgeschichte der Naturwissenschaften* 1, 273–300.

Straub, H. 1970. Bernoulli, Daniel. In *Dictionary of Scientific Biography*, ed. C.C. Gillispie. New York: Scribner's, Vol. 2, 136–46.

Todhunter, I. 1865. *A History of Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Cambridge: Cambridge University Press, repr. New York: Chelsea, 1961.

How to cite this article

Zabell, S. L. "Bernoulli, Daniel (1700–1782)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000114> doi:10.1057/9780230226203.0127

The New Palgrave Dictionary of Economics Online

Bernstein, Eduard (1850 – 1932)

Tom Bottomore

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bernstein, E.; class; class conflict; Engels, F.; Kautsky, K.; Marx, K. H.; social democracy; socialism; Weber, M.; Winstanley, G.

Article

Born in Berlin, 6 January 1850; died in Berlin, 18 December 1932. The son of a Jewish railway engineer and the seventh child in a large family of 15 children, Bernstein grew up in a lower middle-class district of Berlin in ‘genteel poverty’. He did not complete his studies at the Gymnasium, and in 1866 he began an apprenticeship in a Berlin bank. Three years later he became a bank clerk and remained in this post until 1878, but he continued to study independently and for a time aspired to work in the theatre. He became a socialist in 1871, largely through sympathy with the opposition of Bebel, Liebknecht and others to the Franco–Prussian war, and strongly influenced by reading Marx's study of the Paris Commune, *The Civil War in France* (1871). In 1872 Bernstein joined the Social Democratic Workers’ Party, and in 1875 he was a delegate to the conference in Gotha which brought about the union of that party with Lassalle's General Union of German Workers to form a new Socialist Workers’ party, later the Social Democratic Party (SDP). From that time Bernstein became a leading figure in the socialist movement, and in 1878, just before Bismarck's anti-Socialist law was passed, he moved to Switzerland as secretary to a wealthy young socialist, Karl Höchberg, who expounded a form of utopian socialism in the journal *Die Zukunft* which he had founded. It was in 1878 also that Bernstein read Engels's *Anti-Dühring*, which, he said, ‘converted me to Marxism’, and he corresponded with Engels for the first time in June 1879. After some misunderstandings with Marx and Engels, who were suspicious of his relationship with Höchberg, Bernstein won their confidence during a visit to London and in January 1881, with their support, he became editor of *Der Sozialdemokrat* (the newspaper of the SDP, established in 1879). It was, as Gay (1952) notes, ‘the beginning of a great career’.

In 1888 the Swiss government, under pressure from Germany, expelled Bernstein and three of his colleagues on the *Sozialdemokrat*, and they moved to London to continue publication there. The period of exile in England, which lasted until 1901, was crucial in the formation of Bernstein's ideas. He

became a close friend of Engels, who made him his literary executor (jointly with Bebel), and developed a stronger interest in historical and theoretical subjects, contributing regularly to Kautsky's *Die Neue Zeit* and publishing in 1895 his first major work, a study of socialism and democracy in the English revolution (entitled *Cromwell and Communism* in the English translation). Bernstein's major contributions in this study, which he later described as ‘the only large scale attempt on my part to discuss historical events on the basis of Marx's and Engels's materialist conception of history’, were to analyse the civil war as a class conflict between the rising bourgeoisie and both the feudal aristocracy and the workers, and to give prominence to the ideas of the radical movements in the revolution (the Levellers and Diggers), and in particular those of Gerrard Winstanley, who had been ignored by previous historians.

At the same time Bernstein established close relations with the socialists of the Fabian Society and came to be strongly influenced by their ‘gradualist’ doctrines and their rejection of Marxism. In a letter to Bebel (20 October 1898) he described how, after giving a lecture to the Fabian Society on ‘What Marx really taught’, he became extremely dissatisfied with his ‘well-meaning rescue attempt’ and decided that it was necessary ‘to become clear just where Marx is right and where he is wrong’. Soon after Engels's death Bernstein began to publish in *Die Neue Zeit* (from 1896 to 1898) a series of articles on ‘problems of socialism’ which represented a systematic attempt to revise Marxist theory in the light of the recent development of capitalism and of the socialist movement. The articles set off a major controversy in the SDP, in which Kautsky defended Marxist orthodoxy and urged Bernstein to expound his views in a more comprehensive way, as he then proceeded to do in his book on ‘the premisses of socialism and the tasks of social democracy’ (1899; entitled Evolutionary Socialism in the English translation), which made him internationally famous as the leader of the ‘revisionist movement’.

Bernstein's arguments in *Evolutionary Socialism* were directed primarily against an ‘economic collapse’ theory of the demise of capitalism and the advent of socialism, and against the idea of an increasing polarization of society between bourgeoisie and proletariat, accompanied by intensifying class conflict. On the first point he was attacking the Marxist orthodoxy of the SDP, expounded in particular by Kautsky, rather than Marx's own theory, in which the analysis of economic crises and their political consequences was not fully worked out, and indeed allowed for diverse interpretations (Bottomore, 1985). The central part of Bernstein's study, however, concerned the changes in class structure since Marx's time, and their implications. In this view, the polarization of classes anticipated by Marx was not occurring, because the concentration of capital in large enterprises was accompanied by a development of new small and medium-sized businesses, property ownership was becoming more widespread, the general level of living was rising, the middle class was increasing rather than diminishing in numbers, and the structure of capitalist society was not being simplified, but was becoming more complex and differentiated. Bernstein summarized his ideas in a note found among his papers after his death:

‘Peasants do not sink; middle class does not disappear; crises do not grow ever larger; misery and serfdom do not increase. There *is* increase in insecurity, dependence, social distance, social character of production, functional superfluity of property owners’ (cited by Gay, 1952, p. 244).

On some points Bernstein was clearly mistaken. With the further development of capitalism, peasant production has declined rapidly and has been superseded to a great extent by ‘agri-business’; economic crises did become larger, at least up to the depression of 1929–33. It was his analysis of the changing class structure which had the greatest influence, becoming a major issue in the social sciences, and above all in sociology, in part through the work of Max Weber, whose critical discussion of Marxism in

his lecture on socialism (1918) largely restates Bernstein's arguments. There is a more general sense in which Bernstein's ideas have retained their significance; namely, in their assertion of the increasingly ‘social character’ of production and the likelihood of a gradual transition to socialism by the permeation of capitalist society with socialist institutions. In a different form the same notion is expressed by Schumpeter (1942) in his conception of a gradual ‘socialization of the economy’; a conception which can also be traced back to Marx (Bottomore, 1985).

One other aspect of Bernstein's thought should be noted. Influenced by the neo-Kantian movement in German philosophy and by positivism (in an essay of 1924 he noted that ‘my way of thinking would make me a member of the school of Positivist philosophy and sociology’) Bernstein made a sharp distinction between science and ethics and went on to argue, in his lecture ‘How is scientific socialism possible?’ (1901), that the socialist movement necessarily embodies an ethical or ‘ideal’ element: ‘It is something that *ought* to be, or a movement towards something that *ought* to be.’ From this standpoint he criticized in a more general way a purely economic interpretation of history, and especially the kind of ‘economic determinism’ that was prevalent in the orthodox Marxism of the SDP; but in so doing he cannot be said to have diverged radically from the conceptions of Marx and Engels (and indeed he cited Engels's various qualifications of ‘historical materialism’ in support of his own views).

Bernstein's book met with a vigorous and effective response in Rosa Luxemburg's *Sozialreform oder Revolution* (1899), and the SDP became divided between ‘radicals’, ‘revisionists’, and the ‘centre’ (represented by Bebel and Kautsky); and although the latter retained control Bernstein remained a leading figure in the party until 1914. But his growing opposition to the war led him to form a separate organization in 1916 and then to join the left-wing Independent Social Democratic Party of Germany (USPD) in 1917. After the war Bernstein became increasingly disillusioned with the ineffectualness of the SDP in countering the reactionary nationalist attacks on the Weimar Republic, his influence waned, and his last years were spent in isolation.

See Also

- social democracy

Selected works

1895. *Cromwell and Communism*. London: Allen & Unwin, 1930.

1899. *Evolutionary Socialism*. New York: Huebsch, 1909. Reprinted, New York: Schocken, 1961.

1901. Wie ist wissenschaftlicher Sozialismus möglich? *Sozialistische Monatshefte*.

Bibliography

Bottomore, T. 1985. *Theories of Modern Capitalism*. London: Allen & Unwin.

Gay. P. 1952. *The Dilemma of Democratic Socialism*. New York: Columbia University Press.

Luxemburg, R. 1899. *Sozialreform oder Revolution*. Trans. as *Reform or Revolution*, New York: Three Arrows, 1937.

Schumpeter, J.A. 1942. *Capitalism, Socialism and Democracy*. London: Allen & Unwin 5th edn, 1976.

Weber, M. 1918. Socialism. English trans. In *Max Weber The Interpretation of Social Reality*, ed. J.E.T. Eldridge. London: Michael Joseph, 1970.

Bibliographic addendum

A recent biography is Steger, M., *The Quest for Evolutionary Socialism: Eduard Bernstein and Social Democracy*, Cambridge: Cambridge University Press, 1997.

How to cite this article

Bottomore, Tom. "Bernstein, Eduard (1850–1932)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000116> doi:10.1057/9780230226203.0128

The New Palgrave Dictionary of Economics Online

Bertrand competition

Michael R. Baye and Dan Kovenock

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article presents the classic Bertrand model of oligopolistic price competition and shows how alternative assumptions on economic primitives – such as the structure of demand and cost functions, tie-breaking rules, and product differentiation – shape Nash equilibrium prices and profits. We also discuss the related Bertrand–Edgeworth model of price competition in which consumers may be rationed – either strategically or due to capacity constraints – and illustrate how alternative rationing rules influence equilibrium.

Keywords

Bertrand competition; Bertrand equilibrium; Bertrand paradox; Bertrand, J. L. F.; Bertrand–Edgeworth competition; best response (reply); capacity; Cournot, A. A.; duopoly; Edgeworth cycles; homogeneous products; mixed-strategy equilibria; monopoly; Nash equilibrium; oligopoly; price competition; product differentiation; pure-strategy equilibria; rationing rules; residual monopoly price; strategic complements; supermodularity; winner-take-all

Article

‘Bertrand competition’ refers to a model of oligopoly in which two or more firms compete by simultaneously setting prices and in which each firm is committed to provide consumers with the quantity of the firm's product they demand given these ‘posted prices’. The concept is named after the French mathematician Joseph Louis François Bertrand (1822–1900) who, in an 1883 review of Cournot (1838), was critical of Cournot's use of quantity as the strategic variable in his famous duopoly model of market rivalry. In his critique, Bertrand described how, in Cournot's duopoly environment where identical firms produce a homogeneous product under a constant unit cost technology, price competition would lead to price undercutting and a downward spiral of prices. Bertrand erroneously reasoned that this process would continue indefinitely, thereby precluding the existence of an equilibrium. It is now

widely recognized that an equilibrium exists not only in Bertrand's original formulation but in a plethora of other environments in which firms sell either homogeneous or differentiated products.

Formally, *Bertrand competition* is a normal form game in which each of $n \geq 2$ players (firms), $i = 1, 2, \ldots, n$, simultaneously sets a price $p_i \in P_i = [0, \infty)$. Under the assumption of profit maximization, the payoff to each firm *i* is $\pi_i(p_i, p_{-i}) = p_i D_i(p_i, p_{-i}) - C_i(D_i(p_i, p_{-i}))$, where p_{-i} denotes the vector of prices charged by all firms other than *i*, $D_i(p_i, p_{-i})$ represents the total demand for firm *i*'s product at prices (p_i, p_{-i}), and $C_i(D_i(p_i, p_{-i}))$ is firm *i*'s total cost of producing the output $D_i(p_i, p_{-i})$. A *Bertrand equilibrium* is a Nash equilibrium of this game; that is, a vector of prices (p_i^*, p_{-i}^*) such that, for each player *i*, $\pi_i(p_i^*, p_{-i}^*) \geq \pi_i(p_i, p_{-i}^*)$ for all $p_i \in P_i$.

The Bertrand paradox

In the ‘classic’ model of Bertrand competition, each of the *n* firms produces an identical product at a constant unit cost of *c*; that is, $C_i(q_i) = cq_i$. Since their products are perfect substitutes, firms effectively compete for the total demand, *D*(*p*), that a monopolist serving the entire market would obtain by pricing at *p*. The firm setting the lowest price gets all of this demand; in the event of a tie, the firms charging the lowest price share total demand equally. Total demand is sufficiently well-behaved to ensure that the corresponding monopoly profit function, $\pi(p) \equiv pD(p) - C(D(p))$, is not only continuous, but (*a*) has a unique maximizer, the monopoly price p^M; (*b*) satisfies $\pi(p) < \pi(c) = 0$ for $p < c$; and satisfies (*c*) $0 < \pi(p) < \pi(p^M) < \infty$ for all $p^M > p > c$. Despite the continuity of $\pi(p)$, each firm faces a discontinuous profit function

$$\pi_i(p_i, p_{-i}) = \begin{cases} (p_i - c)D(p_i) & \text{if } p_i < p_j \text{ for all } j \neq i \\ (p_i - c)D(p_i)/m & \text{if } i \text{ ties } m-1 \text{ other firms for low price} \\ 0 & \text{otherwise} \end{cases}$$

because a firm that prices even slightly above the lowest price gets no demand. In this classic setting with ‘well-behaved’ demand and constant marginal cost, (p_i^*, p_{-i}^*) is a *Bertrand equilibrium* if and only if $p_j^* \geq c$ for every firm *j* and at least two firms set price equal to *c*. Consequently, all firms earn zero profits in equilibrium, a result that has come to be known as the *Bertrand paradox*. The paradox stems from the fact that, while a monopolist would earn strictly positive profits by charging a price in excess of marginal cost, it takes only two firms to completely dissipate the monopoly profits and achieve the competitive outcome. In a Bertrand equilibrium, all transactions take place at marginal cost (*c*), and all firms earn zero profits.

The proof of this proposition follows in part from the original intuition of Bertrand. Since the products are perfect substitutes, consumers will purchase only from a firm that charges the lowest price in the

market, $p_L \equiv \min_j p_j$. First, $p_L \leq p^M$ in any equilibrium; otherwise, any firm could profitably deviate by lowering its price to p^M. Second, $p_L \geq c$ in any equilibrium; otherwise, a firm charging p_L (and thus earning strictly negative profits) could profitably deviate by increasing its price to c. Third, if $p^M \geq p_L > c$, then at least one firm could increase its profit by unilaterally undercutting p_L by a small amount. Hence, $p_L = c$ in any equilibrium. Fourth, if only a single firm charged a price of $p_L = c$, it would earn a payoff of zero, and could increase its price to $p' > c$ (but below the second-lowest price) to earn a positive profit. Thus, in any equilibrium at least two firms charge a price of $p_L = c$. Finally, since the only firms attracting any consumers are those pricing at $p_L = c$, all firms earn zero profits. Furthermore, no firm can unilaterally change its price to earn positive profits.

One consequence of this argument is that when $n = 2$ there is a unique Bertrand equilibrium in the classic model: both firms set the common price $p_1^* = p_2^* = c$. When $n > 2$, there is a unique symmetric equilibrium (in which $p_i^* = c$ for all i) and a continuum of asymmetric equilibria (where two or more firms price at c and one or more firms charge prices arbitrarily higher than c).

Although the Bertrand paradox result summarized above for the case of identical constant unit costs is stated in terms of pure strategies and a symmetric tie-breaking rule, the paradox also obtains for the extension of strategy spaces to allow for mixed-strategies as well as other tie-breaking rules. Alternative tie-breaking rules include ‘winner-take-all sharing’ (where a fair randomizing device is used to determine the identity of the firm that services the entire market in the event of a tie for the lowest price) and ‘unequal sharing’ (where firms tying for the lowest price receive an unequal fraction of total market demand in the event of a tie for the lowest price).

Baye and Morgan (1999) have shown that if the monopoly profit function, $\pi(p)$, is unbounded, there exists (in addition to the Bertrand paradox equilibria) a continuum of non-degenerate mixed strategy equilibria in which each firm earns positive profits. For instance, suppose market demand is given by $D(p) = p^{\alpha}$, where $\alpha \in (-\infty, -1/n)$ is the elasticity of market demand. In this case, one can show that there is a unique symmetric Cournot (quantity-setting) equilibrium in which each firm earns positive profits and the equilibrium market price is $p^* = [n\alpha/(1+n\alpha)]c$. In contrast, under Bertrand competition any symmetric profit level $\pi^* \in (0, \infty)$ (including profit levels above the Cournot profit) can be achieved in an (atomless) symmetric mixed strategy equilibrium. Equilibrium mixed strategies that support these positive profit levels are described by the cumulative distribution function $F(p) = 1 - \pi^*/\pi(p)$ on $[\pi^{-1}(\pi^*), \infty)$, where $\pi(p) = (p-c)p^{\alpha}$.

Even with a bounded monopoly profit function $\pi(p)$, the coexistence of positive profit equilibria and (zero profit) Bertrand paradox equilibria can arise for alternative cost functions and sharing rules. For instance, with a symmetric tie-breaking rule (see Dastidar, 1995), if firms have identical cost functions that are increasing and strictly convex in output, a symmetric zero profit equilibrium may exist in which each firm prices at p^0, where p^0 satisfies $p^0 D(p^0)/n - C(D(p^0)/n) = 0$. In addition, however, a continuum of positive profit symmetric pure-strategy equilibria can arise in which each firm charges a price contained in an interval above p^0. Intuitively, with strictly convex costs, a firm that deviates by undercutting such a price would increase its demand (and revenues) by a factor of n, but the firm's cost

would increase by a factor greater than *n*.

This result for bounded demand and identical convex costs is based on a symmetric tie-breaking rule; with convex costs, different results generally obtain for other tie-breaking rules. For instance, under the winner-take-all tie-breaking rule (see Baye and Morgan, 2002), any firm charging the price p_L earns a payoff of $\pi(p_L)/\#L$, where $\#L$ is the number of firms charging the price p_L. In this case, if $\pi(p_L) > 0$, some firm could gain by undercutting p_L by a small amount (a firm pricing above p_L could increase its payoff from zero to $\pi(p_L - \varepsilon) > 0$; a firm that tied another firm at p_L could increase its profits from $\pi(p_L)/\#L$ to $\pi(p_L - \varepsilon)$ by slightly undercutting p_L). Consequently, an argument similar to that for the case of constant unit costs implies that, with bounded demand and convex costs, any equilibrium under the winner-take-all sharing rule involves at least two firms charging a price p_L such that $\pi(p_L) = 0$, so that the (zero profit) Bertrand paradox is the only configuration of firm profits.

With bounded demand and identical concave costs, a similar argument reveals that any equilibrium under the winner-take-all sharing rule involves at least two firms charging a price p_L such that $\pi(p_L) = 0$ (Baye and Morgan, 2002). However, under a symmetric sharing rule, concave costs (increasing returns) are problematic for the existence of a Bertrand equilibrium in either pure or mixed strategies. To illustrate, consider a duopoly in which market demand is given by $D(p) = 1 - p$ for $p \in [0, 1]$, and in which each firm has an identical concave cost function

$$C_i(q_i) = \begin{cases} 0 & \text{if } q_i = 0 \\ f + cq_i & \text{if } q_i > 0 \end{cases}$$

where $1 > c > 0$ and $f < [(1-c)/2]^2$. Note that *c* represents marginal cost and *f* is a fixed cost that may be avoided by producing zero output. One may readily verify that a monopolist would earn strictly positive profits by pricing at the monopoly price $p^M = (1+c)/2$, and that the minimum 'breakeven price' is $p^0 = [(1+c) - [(1-c)^2 - 4f]^{1/2}]/2$; that is, $0 = \pi(p^0) > \pi(p)$ for all $p < p^0$. Under a winner-take-all sharing rule, $p_1 = p_2 = p^0$ is a pure-strategy Nash equilibrium and firms earn zero profits in this 'Bertrand paradox' equilibrium. In contrast, under a symmetric tie-breaking rule there does not exist an equilibrium (in pure or mixed strategies).

The intuition for the failure of existence of equilibrium with a symmetric tie-breaking rule in this example is as follows. Clearly, neither firm has an incentive to price below p^0 (since monopoly profits are negative for such prices and a firm can guarantee a payoff of zero by pricing at $p_i = 1$). If both firms priced at p^0 with probability one, symmetric sharing implies that they would earn negative profits, since $C_i(D(p^0)/2) > C_i(D(p^0))/2$. Thus, p^0 is strictly less than the upper bound of the support of at least one firm's (possibly degenerate) mixed strategy. Let $p^H > p^0$ denote highest of the upper bounds of the supports of the two firms' mixed strategies. In any equilibrium, at most one firm has a mass point at p^H; otherwise, there would be a positive probability of a tie at this price and a firm could gain by

reallocating mass to lower prices. If there is a mass point at p^H, the firm charging p^H with positive probability must earn its equilibrium profits at this price, which are necessarily zero since it is undercut with certainty. If there is no mass point at p^H, then a firm whose support includes p^H must achieve its equilibrium payoff when pricing at p^H, and since p^H is undercut with certainty, this equilibrium payoff is zero. Therefore, at least one firm i whose support includes p^H earns an equilibrium payoff of zero. Moreover, since firm i earns an equilibrium payoff of zero, p^0 must be the upper bound of the support of the other firm j's mixed strategy; if the upper bound of j's support was $p' \in (p^0, p^H]$, firm i could increase its profits by reallocating probability mass to some price below p'. Thus, if there is an equilibrium, at least one firm must charge a price of p^0 with probability one. However, since firm i charges prices in the interval $[p^0, p^H]$, and not all mass is at p^0, it follows that there exists some price $p'' \in (p^0, p^H]$ such that firm j could gain by reallocating mass from p^0 to p'', a contradiction. Hence, there does not exist an equilibrium in pure or mixed strategies.

Bertrand – Edgeworth competition

In an early critique of Bertrand and Cournot, Edgeworth (1925) observed that the Bertrand paradox may not obtain if firms are capacity constrained. Indeed, in the analysis above, if firm i's demand $D_i(p_i, p_{-i})$ is greater than firm i's largest competitive supply at p_i, $s_i(p_i) \equiv \max\{\arg\max_q p_i q - C_i(q)\}$, then firm i would earn higher profits by supplying a quantity strictly less than that demand and rationing customers. A variant of Bertrand competition, known as ‘Bertrand–Edgeworth competition’, allows any firm to ration the demand that it faces at given prices by only providing its optimal or competitive supply at its price. Rationing may stem from a physical capacity constraint, k_i, that prevents firm i from producing more than k_i units (as in Edgeworth's original formulation), or more generally, from a firm's strategic incentive to refuse to fulfil the quantity demanded of all consumers at a given price. Under Bertrand–Edgeworth competition one must therefore specify how demand is rationed when a firm's quantity demanded at given prices exceeds the amount of product it produces.

Two prominent rationing rules used in this context are efficient rationing (in which case the good is first allocated to consumers who most highly value the product) and proportional rationing (in which case the good is allocated to a fraction of consumers without regard to their valuations of the product). In the duopoly case, for instance, efficient rationing means that if $p_i > p_j$, firm i's ‘residual’ demand is $D_i(p_1, p_2) \equiv \max\{0, D(p_i) - s_j(p_j)\}$. Under proportional rationing, firm i's demand is $D_i(p_1, p_2) \equiv \max\{0, D(p_i)[1 - s_j(p_j) / D(p_j)]\}$. Under both rationing rules, the firm charging the lowest price enjoys a demand of $D(p_j)$. It is typically assumed that, in the event of a tie, total demand is allocated in proportion to firms’ competitive supplies; that is, if both firms charge a price of p, firm i gets a share $\alpha_i \equiv s_i(p) / (s_1(p) + s_2(p))$.

For the special case of a duopoly in which each firm has a constant marginal cost (c) up to a capacity of k_i, the cost functions are:

$$C_i(q_i) = \begin{cases} cq_i & \text{if } 0 \le q_i \le k_i \\ \infty & \text{if } q_i > k_i \end{cases}$$

In this case, under the assumption of well-behaved demand, $s_i(p_i) = k_i$ for all $p_i \ge c$; that is, each firm opts for a 'corner solution' at full capacity when price exceeds marginal cost. Under both efficient and proportional rationing, if $D(c) \le k_i$, $i = 1, 2$, then neither firm's capacity constraint ever binds and the Bertrand paradox arises under the same conditions as set forth above; the unique equilibrium is $p_1^* = p_2^* = c$. Characterization of equilibrium when one or more firms is capacity constrained at a price equal to *c* depends on whether each firm is capacity constrained at its 'residual monopoly price' when its rival sets $p_j = D^{-1}(k_1 + k_2)$. The term 'residual monopoly price' refers to a firm's optimal price, given its capacity constraint and residual demand (the demand that remains after the other firm has sold its capacity). Note that, in equilibrium, neither firm would ever set a price below $D^{-1}(k_1 + k_2)$, for at such a price total demand exceeds total capacity, and a firm could increase its price without losing sales. Characterization of equilibrium when $D(c) > k_i$ for one or more firms then depends on whether $p_1 = p_2 = D^{-1}(k_1 + k_2)$ is an equilibrium. If, for each firm *i*, $D^{-1}(k_1 + k_2)$ is the residual monopoly price when firm *j* sets $p_j = D^{-1}(k_1 + k_2)$, then $p_1^* = p_2^* = D^{-1}(k_1 + k_2)$ is the unique Bertrand–Edgeworth equilibrium. If some firm *i*'s residual monopoly price exceeds $D^{-1}(k_1 + k_2)$ when $p_j = D^{-1}(k_1 + k_2)$, then the unique equilibrium is in non-degenerate mixed-strategies.

The residual monopoly price depends on the rationing rule. For proportional rationing, $D_i(p_1, p_2) \equiv \max\{0, D(p_i)[1 - k_j / D(p_j)]\}$ for any given p_j, and hence firm *i*'s demand is proportional to $D(p_i)$. This implies that, ignoring firm *i*'s capacity constraint, the residual monopoly price based on $D_i(p_1, p_2)$ corresponds to the standard monopoly price, $p^M \equiv \arg\max_p \{(p - c)D(p)\}$. When $p_j = D^{-1}(k_1 + k_2) < p^M$, firm *i* has sufficient capacity to satisfy residual demand at p^M, and hence p^M is firm *i*'s residual monopoly price; if $p_j = D^{-1}(k_1 + k_2) \ge p^M$, concavity of the monopoly profit function implies that $p_i = D^{-1}(k_1 + k_2)$ is firm *i*'s residual monopoly price. It follows that, for proportional rationing, $p_1^* = p_2^* = D^{-1}(k_1 + k_2)$ is the unique Bertrand–Edgeworth equilibrium as long as $D^{-1}(k_1 + k_2) \ge p^M$.

Under efficient rationing, $D_i(p_1, p_2) \equiv \max\{0, D(p_i) - k_j\}$, so that ignoring firm *i*'s capacity constraint, the residual monopoly price is $p_i^R = \arg\max_{p_i}\{(p_i - c)\max(0, D(p_i) - k_j)\}$. It follows that $p_i^R < p^M$. When $p_j = D^{-1}(k_1 + k_2) < p_i^R$, firm *i* has sufficient capacity to satisfy residual demand at p_i^R, and hence p_i^R is firm *i*'s residual monopoly price; if $p_j = D^{-1}(k_1 + k_2) \ge p_i^R$, concavity of the monopoly profit function implies that $p_i = D^{-1}(k_1 + k_2)$ is firm *i*'s residual monopoly

price. Hence, $D^{-1}(k_1 + k_2)$ is firm *i*'s residual monopoly price when firm *j* sets $p_j = D^{-1}(k_1 + k_2)$ if and only if $D^{-1}(k_1 + k_2) \geq p_i^R$. This implies that the region in which a pure strategy equilibrium arises is larger for the case of efficient rationing than under proportional rationing. In fact, since the unconstrained residual profit-maximization problem faced by firm *i* under efficient rationing may be written in terms of either price or quantity, p_i^R is the price arising in a Cournot setting where firm *i*'s output is a best response to an output of k_j by the rival. Hence, if k_j is less than or equal to firm *i*'s Cournot best response to k_j, firm *i* is capacity constrained and its residual monopoly price equals $D^{-1}(k_1 + k_2)$. Consequently, $p_1^* = p_2^* = D^{-1}(k_1 + k_2)$ is the unique Bertrand–Edgeworth equilibrium when each firm's capacity is less than or equal to its Cournot best response (given unit cost *c*) to the other firm's capacity.

Outside of the above regions of capacity, the only Bertrand–Edgeworth equilibria are in non-degenerate mixed strategies in which firms randomize prices over a common interval of prices that exceed *c* and earn positive expected profits. This corresponds to the regions of capacities in which ‘Edgeworth cycles’ arise (Edgeworth, 1925). As before, these mixed strategies depend on the rationing rule. For proportional rationing, these mixed strategies are generally difficult to derive; see Davidson and Deneckere (1986) for a characterization. For efficient rationing, these mixed strategies have been characterized by Kreps and Scheinkman (1983), and entail the firm with the larger capacity earning an expected payoff that equals the monopoly profit associated with the residual demand (with symmetric capacities, each firm earns this expected payoff). The firm with the larger capacity earns the higher payoff.

To summarize, only two types of pure-strategy equilibria exist under Bertrand–Edgeworth duopoly with constant unit cost. When capacity constraints do not bind, the classic Bertrand equilibrium arises and the unique equilibrium is for each firm to price at marginal cost to earn zero profits. When capacities are sufficiently small, firms price above marginal cost (at a price that clears all capacity) and earn positive profits in the unique Bertrand–Edgeworth equilibrium. When capacities are in an intermediate range, the equilibrium is generally unique, but in non-degenerate mixed strategies. Firms’ prices exceed marginal cost with probability one, and firms earn positive profits.

Positive profit equilibria can also arise in homogeneous product Bertrand settings in which firms endogenously choose capacities. Specifically, consider a two-stage game where, in the first stage, firms simultaneously commit to a capacity, and in the second stage firms simultaneously engage in Bertrand–Edgeworth competition. Under both efficient and proportional rationing, capacity commitment in the first stage permits both firms to avoid the Bertrand paradox in the second stage to earn positive profits. Under efficient rationing, capacity choice followed by Bertrand–Edgeworth competition leads, under fairly general conditions, to equilibrium prices that are identical to those that would arise in a Cournot (quantity setting) duopoly where firms’ unit costs are the sum of capacity and production costs; see Kreps and Scheinkman (1983) and Deneckere and Kovenock (1996). Under proportional rationing, the Cournot outcome arises only if per unit capacity costs are sufficiently large. Otherwise, equilibria may arise in which capacities are asymmetric and non-degenerate mixed strategies are played at the pricing stage; see Davidson and Deneckere (1986).

Product differentiation

Bertrand competition with differentiated products is fundamentally different from Bertrand competition with homogenous products. With differentiated products, the demand for a firm's product is not generally discontinuous at p_L; a firm does not generally lose all of its demand by pricing slightly above p_L, nor does it steal all of rival firms' demands by pricing below p_L. In the classical model of differentiated-product Bertrand competition with downward sloping demands and costs that are non-decreasing in output, each firm's profit function, $\pi_i(p_i, p_{-i})$, is assumed to be twice continuously differentiable, with $\partial \pi_i / \partial p_i \partial p_j > 0$ (strategic complements) and $\partial^2 \pi_i / \partial p_i^2 < 0$.

With suitable assumptions on firms' demands and costs, a Bertrand equilibrium, (p_i^*, p_{-i}^*), is simply the solution to the system of first-order conditions implied by each firm's profit-maximizing pricing decision:

$$\frac{\partial \pi_i(p_i^*, p_{-i}^*)}{\partial p_i} = 0 \text{ for all } i = 1, 2, \ldots, n.$$

Alternatively, one may use the implicit function theorem and use firm *i*'s first-order condition to obtain firm *i*'s optimal price as a function of the prices charged by the other firms: $p_i = \rho_i(p_{-i})$. The function ρ_i is called firm *i*'s best-response (best-reply, reaction) function, and a Bertrand equilibrium in the case of differentiated products corresponds to the intersection of the firms' best-response functions. Total differentiation of firm *i*'s first-order condition reveals that $d p_i / d p_j = -\left(\partial \pi_i / \partial p_i \partial p_j\right) / \left(\partial^2 \pi_i / \partial p_i^2\right) > 0$; that is, strategic complementarities and the concavity of firm *i*'s profits in p_i imply that firm *i*'s best response function is upward sloping.

Notice that, at (p_i^*, p_{-i}^*),

$$\frac{\partial \pi_i(p_i^*, p_{-i}^*)}{\partial p_i} = [p_i^* - C_i'(D_i(p_i^*, p_{-i}^*))] \frac{\partial D_i(p_i^*, p_{-i}^*)}{\partial p_i} + D_i(p_i^*, p_{-i}^*) = 0.$$

Consequently, under mild regularity conditions firm *i*'s equilibrium price exceeds its marginal cost. Furthermore, firms may charge different prices and earn positive profits in a differentiated product Bertrand equilibrium. These results may be extended to the case where $\pi_i(p_i, p_{-i})$ is not differentiable by appealing to the more general notion of supermodularity (Vives, 1990; Milgrom and Roberts, 1990) rather than strategic complementarity (Bulow, Geanakoplos and Klemperer, 1985).

For the duopoly case with linear demands and constant unit costs, strategic complementarity $(\partial \pi_i / \partial p_i \partial p_j > 0)$ arises naturally when the duopolists' products are substitutes in consumption $(\partial D_i / \partial p_j > 0)$. In this case the firms' best-response functions are not only upward sloping (as is implied by strategic complementarity) but linear; consequently, there is a unique Bertrand equilibrium (see Cheng, 1985). Singh and Vives (1984) have shown that, in this linear duopoly case, even though each firm prices above its marginal cost in a differentiated-product Bertrand equilibrium, prices are lower under Bertrand competition than would arise in a differentiated-product Cournot (quantity setting) model. This result for linear demand and costs extends to markets with more than two firms when all firms' products are substitutes in consumption (Häckner, 2000).

See Also

- Cournot competition
- supermodularity and supermodular games

Bibliography

Baye, M.R. and Morgan, J. 1999. A folk theorem for one–shot Bertrand games. *Economics Letters* 65, 59–65.

Baye, M.R. and Morgan, J. 2002. Winner-take-all price competition. *Economic Theory* 19, 271–82.

Bertrand, J. 1883. (Review of) *Théorie Mathématique de la Richesse Sociale* par Léon Walras: *Recherches sur les Principes Mathématiques de la Théorie des Richesses* par Augustin Cournot. *Journal des Savants* 67, 499–508.

Bulow, J., Geanakoplos, J. and Klemperer, P. 1985. Multimarket oligopoly: strategic substitutes and complements. *Journal of Political Economy* 93, 488–511.

Cheng, L. 1985. Comparing Bertrand and Cournot equilibria: a geometric approach. *RAND Journal of Economics* 16, 146–52.

Cournot, A. 1838. *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. Paris: Hachette.

Dastidar, K.G. 1995. On the existence of pure strategy Bertrand equilibrium. *Economic Theory* 5, 19–32.

Davidson, C. and Deneckere, R. 1986. Long-run competition in capacity, short-run competition in price, and the Cournot model. *RAND Journal of Economics* 17, 404–15.

Deneckere, R. and Kovenock, D. 1996. Bertrand–Edgeworth duopoly with unit cost asymmetry. *Economic Theory* 8, 1–25.

Edgeworth, F.Y. 1925. The pure theory of monopoly. In *Papers Relating to Political Economy* 1, 111–42.

Häckner, J. 2000. A note on price and quantity competition in differentiated oligopolies. *Journal of Economic Theory* 93, 223–39.

Kreps, D.M. and Scheinkman, J.A. 1983. Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell Journal of Economics* 14, 326–37.

Milgrom, P. and Roberts, J. 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica* 58, 1255–77.

Singh, N. and Vives, X. 1984. Price and quantity competition in a differentiated duopoly. *RAND Journal of Economics* 15, 546–54.

Vives, X. 1990. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics* 19, 305–21.

How to cite this article

Baye, Michael R. and Dan Kovenock. "Bertrand competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000336>
doi:10.1057/9780230226203.0129

close print preview

The New Palgrave Dictionary of Economics Online

Bertrand, Joseph Louis François (1822 – 1900)

Martin Shubik

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bertrand, J. L. F.; Cournot, A. A.; duopoly; Edgeworth, F. Y.; mathematical economics; tâtonnement; Walras, L.

Article

Bertrand was born and died in Paris. He was an eminent but not great mathematician, graduate and professor of mathematics at the Ecole Polytechnique and from 1862 to 1900 a member of the Collège de France. His relevance to economic thought comes in his criticism of ‘pseudo-mathematicians’ in the *Journal des Savants* (1883) where he reviewed *Théorie mathématique de la richesse sociale* of Walras and *Recherches sur les principes mathématiques de la théorie des richesses* of Cournot. It is doubtful if Bertrand considered the problems of formal economic modelling more than casually, viewing the two works through the eyes of a mathematician with little substantive interest or understanding. His comments on Cournot were not only somewhat harsh, but as the subsequent developments in oligopoly theory and the theory of games have shown, both Cournot's model of duopoly and Bertrand's remodelling of duopoly with price rather than quantity as a strategic variable are worth investigation. Cournot's model has been (until recently) more generally treated than Bertrand's model. It remained for Edgeworth to point out the limitations of Bertrand's model (see Shubik, 1959). Bertrand also raised objections to the reference and realism of the process description of Walras of ‘tâtonnement’.
It has been suggested (Blaug and Sturges, 1983) that Bertrand's critical review was used by opponents of mathematical economics as the basis for their position. Although explicit proof of this is hard to establish the tone and force of Bertrand's critique makes this highly probable.

Selected works

1883. (Review of) *Théorie mathématique de la richesse sociale* par Léon Walras: *Recherches sur les principes mathématiques de la théorie des richesses* par Augustin Cournot. *Journal des Savants*, September, 499–508.

Bibliography

Blaug, M. and Sturges, P., eds. 1983. *Who's Who in Economics*. Brighton, England: Wheatsheaf Books.

Byron, G.H. 1899–1900. Joseph Bertrand. *Nature* 1591(61), 614–16.

Shubik, M. 1959. *Strategy and Market Structure*. New York: Wiley.

Storick, D.L. 1970. *Joseph Louis François Bertrand. Dictionary of Scientific Biography*, vol. 2. New York: Scribners.

How to cite this article

Shubik, Martin. "Bertrand, Joseph Louis François (1822–1900)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000119>
doi:10.1057/9780230226203.0130

The New Palgrave Dictionary of Economics Online

Beveridge curve

Eran Yashiv

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Beveridge curve depicts a negative relationship between unemployed workers and job vacancies, a robust finding across countries. The position of the economy on the curve gives an idea as to the state of the labour market. The modern underlying theory is the search and matching model, with workers and firms engaging in costly search leading to random matching. The Beveridge curve depicts the steady state of the model, whereby inflows into unemployment are equal to the outflows from it, generated by matching.

Keywords

Beveridge curve; Beveridge, W. H.; business cycle; excess demand and supply; frictions; information costs; job search; matching function; microfoundations; Phillips curve; unemployment; vacancies; wage inflation

Article

The Beveridge curve depicts a negative relationship between unemployed workers (u) and job vacancies (v). The interest in the curve is related to the role it plays in aggregate models, which study labour market outcomes and dynamics. The position of the economy on the curve gives an idea as to the state of the labour market; for example, a high level of vacancies and a low level of unemployment would indicate a ‘tight’ labour market. The literature has attempted to explain the coexistence of unemployment and vacancies, their negative relationship, and the implied dynamics.

The curve is named after William Beveridge, a British lord, lawyer, head of academic institutions, Member of Parliament, and founder of the modern British welfare state. In a 1944 report (Beveridge, 1944), Beveridge discussed the relationship between the demand for workers, captured by vacancies, and the rate of unemployment. While he did not plot a curve or present a table with a comparison of u and v, he offered detailed data on these variables and discussed them at some length. His analysis

implied that there is a negative relationship between them. In this early work he tackled many of the issues that remain under study in this field: the potential mismatch between unemployed workers and job vacancies, aggregate demand factors versus reallocation factors (for example, deficient overall demand for labour as opposed to low demand in particular industries), trend versus cyclical changes (for example, changes in *u* and *v* along the business cycle versus long-run changes), and measurement issues (such as the various possible ways of mismeasuring vacancies).

The negative $u - v$ relationship is a robust finding across countries, though shifts of the curve over time are often observed. This can be seen, for example, in a 16-country graphical description of the curve presented in Layard, Nickell and Jackman (2005, pp. 36–7). Detailed descriptions and analyses of the empirical findings concerning the Beveridge curve for the United States are to be found in Blanchard and Diamond (1989), and for the UK in Pissarides (1986).

What underlies this negative relationship? The early literature of the late 1950s and in the 1960s dealt with the curve in the context of exploring excess demand in the labour market and its influence on wage inflation. This was motivated by the extensive study of the Phillips curve that took place in those years. The literature typically defined excess demand as unfilled vacancies less unemployed workers, considered the data on these variables, and then looked at the relationship between measures of excess demand and wage behaviour. This literature recognized that, even when there is no excess supply, there is positive unemployment due to frictions. It derived a negatively sloped $u - v$ curve from a model of distinct labour markets, interacting at different levels of disequilibrium, with the markets at points off both labour supply and labour demand curves. The $u - v$ curve was shown to be stationary and observed *u* and *v* points were expected to cycle around it. Movements up and down the curve reflect increases and decreases in the excess demand for labour. The curve itself can shift as a result of changes in the speed of market clearing or changes in the sectoral composition of labour demand. The observed $u - v$ data may be a compound of structural shifts of the curve together with cyclical movements about it. Key contributions to this strand of work were progressively made by Dow and Dicks-Mireaux (1958), Lipsey (1960), Holt and David (1966), Hansen (1970), and Bowden (1980).

In the 1970s and 1980s an alternative approach was developed – the search and matching model. A key difference between this model and the early literature is its derivation of vacancies and unemployment as equilibria, rather than disequilibria, phenomena. The model was developed in the work of Peter Diamond, Dale Mortensen, and Christopher Pissarides (see Pissarides, 2000, for a detailed exposition, and Yashiv, 2006, for a recent survey). The model may be briefly described as follows. Workers and firms engage in costly search to find each other. Firms spend resources on advertising, on posting job vacancies, on screening and, subsequently, on training. Workers spend resources on job search, with costs pertaining to activities such as collecting information and applying for jobs. Workers and firms are assumed to be randomly matched. After matching, the worker and the firm engage in bilateral bargaining over the wage. The matching process assumes frictions such as informational or locational imperfections. It is formalized by a ‘matching function’ that takes searching workers and vacant jobs as arguments and produces a flow of matches (*m*), and is given by $m = m(u, v)$. It is continuous, non-negative, increasing in both its arguments, and concave. Typically, it is assumed to be constant returns to scale. The flow into unemployment results from job-specific shocks to matches that arrive at the Poisson rate λ. These shocks may be explained as shifts in demand or productivity shocks. Once a shock arrives, the firm closes the job down. The evolution of the unemployment rate ($\dot{u}$) is therefore given by

the difference between the separation flow (λ times the employment rate $1 - u$) and the matching flow:

$$\dot{u} = \lambda(1 - u) - m(u, v). \quad (1)$$

Denote the rate at which workers are matched to jobs (the job finding rate) by $p = \frac{m}{u}$ so that $m = pu$. In the steady state the rate of unemployment is constant, so setting $\dot{u} = 0$ the following obtains:

$$u = \frac{\lambda}{\lambda + p}. \quad (2)$$

This is the Beveridge curve: as *p* depends on *m*, it depends on both *u* and *v*, and this equation can be represented in vacancy (*v*) – unemployment (*u*) space by a downward-sloping curve. The mechanism is the following. When vacancies *v* rise, matching *m* rises, and so the job finding rate *p* rises. Workers find jobs at a faster rate and unemployment *u* declines. Vacancies themselves are determined by a firm optimality equation, equating vacancy costs and benefits at the margin.

As can be seen in the equations above, the matching function plays a crucial role in generating the Beveridge curve. Petrongolo and Pissarides (2001) provide a comprehensive survey of estimation of this function, finding the following main features: (*a*) the prevalent specification is Cobb–Douglas, that is, $m = \mu u^{\alpha} v^{\beta}$; (*b*) usually constant returns to scale ($\alpha + \beta = 1$) is found, though some studies have produced evidence in favour of increasing returns to scale; (*c*) many studies have added other variables – such as demographical or geographical variables, incidence of long-term unemployment, and UI – finding some of them significant, but not changing the preceding findings; (*d*) these general patterns are robust across countries and time periods.

Research along the lines of this model – in progress – is likely to provide a richer account of the Beveridge curve: the matching function is studied for microfoundations, heterogeneity is explicitly explored, endogenous separations are allowed for, interactions with capital investment are considered, and learning and on-the-job search leading to job-to-job movements are incorporated. Going beyond this strand of the literature, research is also beginning to explore equilibrium search models, which feature a Beveridge curve, with alternative $u - v$ meeting processes, not modelled as matching functions. Thus, the Beveridge curve remains a topic of active research in macroeconomics and labour economics, more than 60 years after it was first studied.

See Also

- Beveridge, William Henry

Bibliography

Beveridge, W. 1944. *Full Employment in a Free Society*. London: George Allen and Unwin.

Blanchard, O. and Diamond, P. 1989. The Beveridge curve. *Brookings Papers on Economic Activity* 1, 1–60.

Bowden, R. 1980. On the existence and secular stability of the u-v loci. *Economica* 47, 35–50.

Dow, J. and Dicks-Mireaux, L. 1958. The excess demand for labour: a study of conditions in Great Britain, 1946–56. *Oxford Economic Papers* 10, 1–33.

Hansen, B. 1970. Excess demand, unemployment, vacancies and wages. *Quarterly Journal of Economics* 84, 1–23.

Holt, C. and David, M. 1966. The concept of vacancies in a dynamic theory of the labor market. In *Measurement and Interpretation of Job Vacancies*, ed. NBER. New York: Columbia University Press.

Layard, R., Nickell, S. and Jackman, R. 2005. *Unemployment: Macroeconomic Performance and The Labour Market*, 2nd edn. Oxford: Oxford University Press.

Lipsey, R. 1960. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1862–1957: a further analysis. *Economica* 27, 1–31.

Petrongolo, B. and Pissarides, C. 2001. Looking into the black box: a survey of the matching function. *Journal of Economic Literature* 39, 390–431.

Pissarides, C. 1986. Unemployment and vacancies in Britain. *Economic Policy* 1, 499–559.

Pissarides, C. 2000. *Equilibrium Unemployment Theory*, 2nd edn. Cambridge, MA: MIT Press.

Yashiv, E. 2006. Labor search and matching in macroeconomics. Institute for the Study of Labour (IZA) Discussion Paper no. 2743.

How to cite this article

Yashiv, Eran. "Beveridge curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000318> doi:10.1057/9780230226203.0131

The New Palgrave Dictionary of Economics Online

Beveridge, William Henry (1879 – 1963)

Jose Harris

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Beveridge, W. H.; Keynesian analysis of unemployment; mixed economy; Pigou, A. C.; Poor Law; redistribution of income; Robbins, L. C.; social insurance; transfer payments; unemployment; welfare economics; welfare state

Article

Beveridge is chiefly remembered as a social and administrative reformer, whose *Social Insurance and Allied Services* (1942) set out the basic principles and structure of the post-war welfare state. Paradoxically, however, he thought of himself chiefly as an academic economist whose significance for posterity would lie in the fields of manpower policy and the theory of prices. Throughout his life his approach to economic problems was resolutely inductive and empirical, in contrast with the deductive and analytical method characteristic of most English economists. His early work, *Unemployment: A Problem of Industry* (1909), was based on detailed statistical analysis of the case-papers of applicants for unemployment relief. It drew attention to the structural, geographical and informational barriers that stood in the way of a perfect market for labour; and although its challenge to orthodox theory was practical rather than theoretical, it helped to erode belief in a natural economic equilibrium. Later editions of *Unemployment* (revised with the help of Lionel Robbins) were more strongly influenced by classical economic thought, but Beveridge never abandoned his belief that unemployment could only be cured by state intervention to organize and rationalize the market for labour. Beveridge in the 1930s was initially highly critical of the Keynesian analysis of unemployment; and although during the early 1940s he gradually absorbed many aspects of Keynesian thought, his *Full Employment in a Free Society* (1944) differed markedly from Keynes in its emphasis on the need for physical as well as fiscal controls over the economy and, in particular, on manpower planning.

Beveridge's early work on unemployment convinced him that there was a close and measurable connection between levels of economic activity and movements of prices. In the early 1920s he embarked upon what he came to see as his life's work; namely, the compilation of historical and statistical data relating to movements of prices since the 12th century. Beveridge's data convinced him

that unemployment was caused, both nationally and internationally, by falls in the prices of primary products (though he failed to consider the possibility that the sequence of causation might lie in the other direction). Beveridge's resistance to the use of analytical models meant that his data was of limited value to (and indeed often mocked by) economic theorists. Since his death, however, his material has been a seam of gold to many economic historians. Only one volume of the proposed project was ever published, *Prices and Wages in England from the Twelfth to the Nineteenth Century*, vol. I (1939), but much unpublished material survives among Beveridge's papers in the British Library of Political Science and the Institute of Historical Research.

Although Beveridge is often seen as a leading protagonist of the ‘mixed’ economy, his writings on economic policy displayed a recurrent scepticism about how far it was possible to reconcile state intervention with consumer sovereignty. His study of *British Food Control* (1928) suggested that there were advantages and disadvantages in both a ‘laissez faire’ and a ‘command’ economy, but that it was both logically and practically impossible to have the two in combination. Such doubts were partially allayed by the transformation of popular attitudes which appears to have occurred during the Second World War, but were never fully resolved. In his writings on social welfare, Beveridge appears to have been little influenced by, and indeed largely unconscious of, the growing body of contemporary writings on welfare economics produced by theorists like Pigou. His approach to social insurance, and to transfer payments generally, was that of an early 19th-century utilitarian, modified by a sociological and humanitarian perspective. All his proposals on social security display a concern to maintain some of the central economic tenets of the Poor Law (maintenance of incentives, encouragement to private saving, strict avoidance of relief-in-aid-of-wages) together with more ‘organic’ goals such as national efficiency and the maintenance of civilized minimum standards. His arguments for or against various methods and degrees of ‘redistribution’ were nearly always rooted in pragmatism or rule-of-thumb propositions about human behaviour, rather than in rigorous marginal analysis. Even in the most collectivist and ‘socialistic’ period of his career, he was insistent that claims to welfare should be rooted as far as possible in ‘contract’ rather than ‘status’. His general perception of social welfare should be seen as that of a popular political theorist rather than that of an academic economist; though clearly his ideas in this field were both influenced by, and had wider implications for, economic thought.

Selected works

1909. *Unemployment: a Problem of Industry*. London: Longmans & Co. Another edn, *Unemployment: a Problem of Industry, 1909 and 1930*. London: Longmans & Co., 1930.

1928. *British Food Control*. London: Oxford University Press.

1931. *Tariffs: the Case Examined*. London: Longmans & Co. Popular edn, 1932.

1936. *Planning under Socialism, and Other Addresses*. London: Longmans & Co.

1939. *Prices and Wages in England, from the Twelfth to the Nineteenth Century*, vol. 1, London: Longmans & Co.

1942. *Social Insurance and Allied Services: The Beveridge Report in Brief*. London: HMSO.

1944. *Full Employment in a Free Society: A Report*. Long: G. Allen & Unwin. 2nd edn, 1960.

Also numerous articles in *Economic Journal*, *Economica*, *Sociological Review*, *Politica*, and elsewhere.

How to cite this article

Harris, Jose. "Beveridge, William Henry (1879–1963)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000124> doi:10.1057/9780230226203.0132

The New Palgrave Dictionary of Economics Online

bias correction

Jinyong Hahn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

asymptotic theory; bias correction; empirical likelihood; generalized method of moments; limited information maximum likelihood; nuisance parameters; panel models; two-stage least squares

Article

Bias correction is a statistical technique used to remove the bias of an estimator. An unbiased estimator is such that its expectation is equal to the parameter of interest. Many introductory statistics textbooks discuss the desirability of having an unbiased estimator, although it is quickly pointed out that unbiasedness alone cannot be a good criterion for an estimator. This is usually illustrated by comparing two estimators with the use of a concrete loss function, where it is noted that an unbiased estimator with a large variance may be inferior to a biased estimator with a small variance.

Analysis of exact finite sample theory is difficult, or impossible, for many estimators. Therefore, sampling properties of econometric estimators are usually discussed in the context of asymptotic approximation. Many estimators used in econometrics are consistent and asymptotically efficient, so the bias is usually a non-issue in such first-order asymptotic theory. On the other hand, the first-order asymptotic theory may fail to provide a good approximation to the exact finite sample distribution of an estimator, and even an asymptotically unbiased estimator may have a significant bias under small sample sizes. Higher-order asymptotic approximation may then be used to understand the finite sample properties, including the approximate bias. To be more specific, suppose that we use an estimator $\hat{\theta}$ to estimate the parameter of interest θ_0. For many cases, $\hat{\theta}$ allows a three term stochastic expansion

$$\sqrt{n}(\hat{\theta} - \theta_0) = \hat{T}_1 + \hat{T}_2 / \sqrt{n} + \hat{T}_3 / n + O_p(n^{-3/2}),$$

where n is the sample size. The higher-order asymptotic bias of $\hat{\theta}$ is given by b_0/n, where

$$b_0 = \lim_{n \to \infty} E[\hat{T}_2].$$

In the recent literature, bias correction is usually understood to be a method of removing such approximate bias b_0/n. These methods include analytical corrections such as the standard textbook expansion for functions of sample means, and the more complicated formulas required for other estimators. They also include jackknife and bootstrap bias corrections. Correction of approximate bias is usually accompanied by increase of variance, and early literature such as Pfanzagl and Wefelmeyer (1978) focused on the efficiency aspects of bias correction. In general, bias correction cannot be always advocated on efficiency grounds.

Bias correction has received renewed attention in the more recent literature. When there are many nuisance parameters, the parameters of interest are typically estimated with significant biases. The biases are often so severe that removal of such biases almost always results in efficiency gain. Two strands of literature deal with models with many nuisance parameters. First, when a parameter of interest is estimated with many instruments, the resultant estimator may be quite biased. For example, the two-stage least squares estimator (2SLS) tends to be severely biased when there are many first-stage coefficients to be estimated; see for example Bekker (1994). It has been noted that some estimators are not sensitive to the presence of such nuisance parameters, and the instrumental variables literature is focused on developing such robust estimators. For linear simultaneous equations models, the limited information maximum likelihood estimator (LIML) was shown to have very little bias for linear models. For nonlinear models, it was shown that the empirical likelihood (EL) estimator tends to be less biased than the generalized method of moments estimator (GMM) when there are many moment restrictions; see Newey and Smith (2004).

The second strand of literature in which bias correction has played an important role is concerned with panel models. Parameters of interest in panel models are usually estimated with substantial bias when fixed effects are estimated; see Neyman and Scott (1948). The literature examined methods of removing such bias. Hahn and Newey (2004) proposed that the bias be estimated and subtracted from the estimator itself. Arellano (2003) and Woutersen (2002) proposed that the moment equation be modified.

See Also

- two-stage least squares and the *k*-class estimator

Bibliography

Arellano, M. 2003. Discrete choices with panel data. *Investigaciones Económicas* 27, 423–58.

Bekker, P. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62, 657–81.

Hahn, J. and Newey, W. 2004. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72, 1295–319.

Newey, W. and Smith, R. 2004. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72, 219–55.

Neyman, J. and Scott, E. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–31.

Pfanzagl, J. and Wefelmeyer, W. 1978. A third-order optimum property of the maximum likelihood estimator. *Journal of Multivariate Analysis* 8, 1–29.

Woutersen, T. 2002. Robustness against incidental parameters. Working Paper No. 20028. Department of Economics, University of Western Ontario.

How to cite this article

Hahn, Jinyong. "bias correction." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000308> doi:10.1057/9780230226203.0133

The New Palgrave Dictionary of Economics Online

biased and unbiased technological change

Peter L. Rousseau

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides working definitions of biased and unbiased technological change based on the relative responses of the marginal products of capital and labour that occur in the face of economic shocks. These Hicksian definitions are distinguished from others that focus on how technology augments the production function. The bias and augmentation of technical progress are then linked through the substitutability of labour and capital. Examples of ‘labour-biased’ and ‘capital-biased’ technological change from the 19th century to the present illustrate these ideas.

Keywords

biased and unbiased technological change; CES production function; Cobb–Douglas functions; elasticity of substitution; information technology; neutral production functions; skill-bias; technical change

Article

Among the central problems in growth economics is how to organize thinking about technological progress and its role in macroeconomic outcomes. In *The Theory of Wages* (1932), John Hicks offered a set of classifications for technical change that remains in common use. These classifications are based on the observation that inventions are unlikely to increase the marginal products of all factors of production in the same proportion, but rather will affect the marginal products of some factors more than others. Take, for example, the baseline two-factor neoclassical production function:

$$Y = F(K, L), \tag{1}$$

where Y is aggregate output, K is the capital stock, and L is labour. One way to introduce a technology parameter A is to place it at the front of the production function as

$$Y = AF(K, L). \tag{2}$$

Notice that A enters linearly, so that a doubling of the technology parameter also doubles output. Technological progress of this type is said to be ‘unbiased’ or ‘Hicks neutral’ in that the ratio of the marginal products of capital and labour used in the production process does not change. In this case, progress simply requires a renumbering of production isoquants.

Innovations are rarely neutral, however, and for this reason economists have naturally been more interested in cases where technological change alters the ratio of marginal products. When this occurs, technological change is said to be ‘biased’. Hicks defines the bias as ‘labour-saving’ when the marginal product of capital increases more than that of labour for a given capital-labour ratio, thereby increasing the demand for capital. ‘Capital-saving’ technical progress occurs when the marginal product of labour rises more than that of capital for a given capital–labour ratio, thereby increasing the demand for labour. Nowadays economists simply refer to technological change that is labour-saving in the Hicksian sense as having a ‘capital bias,’ and change that is capital-saving in the Hicksian sense as having a ‘labour bias.’ This avoids confounding the bias of a given technological change with the way that it enters the production function.

An alternative concept proposed by R.F. Harrod (1937; 1948) defines technological change as neutral if the marginal product of capital is unchanged at a given capital–output ratio. Another way of stating this is that, under a constant rate of interest and an infinite supply of capital at that rate, a technological change is ‘Harrod-neutral’ if it leaves the length of the production process unaltered. H. Uzawa (1961) shows that this implies a production function of the form

$$Y = F(K, AL), \tag{3}$$

where AL is a unit of ‘effective’ labour. Note that this formulation is *not* neutral in the Hicksian sense unless the production function is Cobb–Douglas. Economists commonly refer to (3) as a ‘labour-augmenting’ production function, but it does not follow that technological change is necessarily labour-biased in the Hicksian sense of relative marginal products.

The opposite symmetric case to Harrod-neutrality defines an invention as neutral if the wage rate remains unchanged at a constant labour-output ratio. This implies a production function of the form

$$Y = F(AK, L), \tag{4}$$

where *AK* is a unit of ‘effective’ capital. Economists often refer to this ‘capital-augmenting’ form of the production function as ‘Solow-neutral,’ but only because Robert Solow (1959) was first to use this form to model technological progress. Once again, this formulation is *not* neutral in the Hicksian sense unless the production function is Cobb–Douglas, and changes in *A* are not necessarily capital-biased in the Hicksian sense. R. Sato and M.J. Beckmann (1968) offer a useful taxonomy of these and other ‘neutral’ production functions.

Of the three output equations shown above, it turns out that only the second (that is, labour-augmenting) form is consistent with a settling down to constant growth under steady technological progress and assumptions of constant returns to scale and diminishing marginal rates of substitution in production. Thus, if we are interested in neoclassical models that move beyond Cobb–Douglas production and possess a steady state, it is useful for technology to multiply labour and make it more effective. Since US wages have risen over the past century while the rental rate has remained relatively steady, the labour-augmenting formulation is at least a priori consistent with the evidence from the United States.

To distinguish technological progress that is factor-augmenting from their underlying Hicksian factor-biases, it is necessary to consider the elasticity of substitution between the factors as technical change occurs. Daron Acemoglu (2002) illustrates this with a CES (that is, constant elasticity of substitution) production function of the form

$$Y = [w(A_L L)^{\frac{\sigma-1}{\sigma}} + (1-w)(A_K K)^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma}{\sigma-1}}, \tag{5}$$

where σ is the elasticity of substitution between capital and labour, A_L and A_K are factor-specific technology parameters, and *w* is a weight ($0 \leq w \leq 1$) that measures the relative importance of each factor. The factors are gross substitutes when $\sigma > 1$, whereas they are gross complements when $\sigma < 1$. With $\sigma > 1$, substitutability between factors allows both the augmentation and bias of technological change to lean towards the same factor. In the case where $\sigma < 1$, however, a capital-augmenting technological change (or a rise in A_K) actually increases demand for the complementary input (that is, labour) more than it increases the demand for capital. The excess demand for labour raises its marginal product more than that of capital, leading to a labour bias in production. Similarly, a labour-augmenting technological change (or a rise in A_L) leads to a capital-bias when $\sigma < 1$. When $\sigma = 1$ the production function is Cobb–Douglas and an increase in *A* does not produce a bias towards either factor.

Hicks and A.C. Pigou (1920) have contended that most technological change is capital biased, and the American experience in the latter half of the 19th century would seem to support this view. Innovations such as the Bessemer process of steelmaking, new distillation methods in petroleum refining, and the

adoption of European reduction methods in flour milling, as noted by John James (1983), led to capital deepening and economies of scale in these industries that increased concentration. Such technological changes seem so important that the rise of big business around the turn of the 20th century is sometimes attributed to them. Though this view probably overstresses the role of technology in the evolution of industrial structure over this period, it is interesting that the capital bias observed in industries for which the story fits were a result of labour augmentation (that is, a rise in A_L) and inelastic factor substitution (that is, $\sigma < 1$).

Electrification offers another example. Prior to its arrival, manufacturing had been designed around the rigidities of steel shafts that ran through the length of a factory and were turned in unison by a single water or steam-powered generator. Afterwards, as Warren Devine (1983) describes, the organization of work gradually evolved to exploit the open factory structure that electric unit drive made possible. Unit drive meant less time spent maintaining complex systems of leather straps and pulleys that transferred power from the rotating steel shafts to the machines, and less down time caused by the need to stop all production to repair a single machine. Electrification and unit drive also made it economical for factories to stay open longer. These innovations made labour more productive (that is, raising A_L), but more focused machinery also reduced the amount of labour that was needed to operate a factory ($\sigma < 1$), raising the marginal product of capital more quickly than that of labour and producing a capital bias. The bias leaned even more towards capital as the diffusion of electricity began to mature, and labour-saving innovations such as vacuum cleaners, toasters, and electric blast furnaces became commonplace.

But is the apparent capital-bias in technological change largely ‘induced’ by changes in factor prices? Charles Kennedy (1964) points out that falling capital prices will motivate individuals to build more inventions that economize on labour than they would build at constant factor prices. Since the prices of capital goods have declined fairly consistently for more than a century and a half, it seems natural that the vast majority of induced inventions would have been capital biased. At the same time, it is important to distinguish biased technological progress (that is, an outward movement and shift along an isoquant) from movements along a fixed isoquant that arise from changes in factor prices, since such changes do not represent technological progress at all. Noting these potential biases, Hicks concludes that ‘autonomous’ inventions, meaning those not prompted by decline of a relative factor price, need not be predominantly capital biased. Indeed, information technology (IT) presents an example where the bias may have moved in the opposite direction.

Computers reduced expenditures on specialized and/or mechanical office machines, thereby making capital more productive (that is, raising A_K). At the same time, labour also became more productive as skilled individuals learned how to use computers to perform complex tasks and less-skilled individuals accomplished routine tasks much more quickly (that is, raising A_L). Thus, there seem to be complementarities between IT and skilled workers, raising the return to skill and producing a ‘skill bias’, while there has been some substitution of computers for less skilled individuals, pressing towards a capital bias. On the whole, however, the complementarity effects so far have outweighed substitution effects, leading to a labour bias. As an invention in the method of inventing, IT has also led to a wide range of induced innovations, both capital- and labour-saving. Design tools used by engineers, for example, have improved the quality of capital goods and allowed more new products to be created. The availability of a broad base of knowledge on the World Wide Web from all over the globe has also transmitted the information needed to make labour more productive.

Is IT typical of the type of technological change that is likely to continue, starting with a labour bias but spawning new innovations that are for the most part labour-saving? If so, parsing out the components of labour bias, and particularly understanding the role of skill bias in the post-war US economy, seems at the core of understanding the role that technology will play in 21st century economic growth.

See Also

- Hicks, John Richard
- skill-biased technical change
- technical change

Bibliography

Acemoglu, D. 2002. Directed technical change. *Review of Economic Studies* 69, 781–809.

Devine, W.D., Jr. 1983. From shafts to wires: historical perspective on electrification. *Journal of Economic History* 43, 347–72.

Harrod, R.F. 1937. Review of Joan Robinson's *Essays in the Theory of Employment Economic Journal* 47, 326–30.

Harrod, R.F. 1948. *Towards a Dynamic Economics*. London: Macmillan.

Hicks, J. 1932. *The Theory of Wages*. London: Macmillan.

James, J.A. 1983. Structural change in American manufacturing, 1850–1890. *Journal of Economic History* 43, 433–59.

Kennedy, C. 1964. Induced bias in innovation and the theory of distribution. *Economic Journal* 74, 541–7.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Sato, R. and Beckmann, M.J. 1968. Neutral inventions and production functions. *Review of Economic Studies* 35, 57–66.

Solow, R.M. 1959. Investment and technical change. In *Mathematical Methods in the Social Sciences*, ed. K.J. Arrow, S. Karlin and P. Suppes. Palo Alto: Stanford University Press.

Uzawa, H. 1961. Neutral inventions and the stability of growth equilibrium. *Review of Economic Studies* 28, 117–24.

How to cite this article

Rousseau, Peter L. "biased and unbiased technological change." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000125>
doi:10.1057/9780230226203.0134

The New Palgrave Dictionary of Economics Online

Bickerdike, Charles Frederick (1876 – 1961)

John S. Chipman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

acceleration principle; Aftalion, A.; Bickerdike, C. F.; depreciation; durability of capital; Edgeworth, F. Y.; elasticities approach to the balance of payments; Frisch, R. A.; gestation period of capital; Hansen, A.; incipient tariff; Johnson, H. G.; Kahn, R. F.; Lerner, A.P.; local public finance; Metzler, L.A.; optimal tariffs; price discrimination; Robinson, J. V.

Article

Bickerdike was born in England (whereabouts unknown) on 15 May 1876 and died in Wallington, Surrey, on 3 February 1961. He studied at Oxford from 1895 to 1899 where he received his BA degree in 1899 and MA in 1910. Upon winning the Cobden Prize for an essay summarized in Bickerdike (1902) he became a protégé of Edgeworth. After serving briefly as Lecturer on Economics and Commerce at the University of Manchester (1910–12) he entered the civil service with a position in the Board of Trade, where he remained until his retirement in 1941.

Bickerdike's published work consists of 15 articles and 38 book reviews, all (save two of the articles) in the *Economic Journal*. He is chiefly known as the originator of the theory of incipient and optimal tariffs (1906; 1907), according to which a country can always gain by imposing a sufficiently small tariff on its imports and can maximize its welfare by imposing a suitable tariff. To derive these results he developed a model (1907) in which nominal import and export prices were expressed as functions of the quantities of imports and exports respectively (with no cross-effects), each country being assumed to stabilize the value of its currency. The elasticities of demand for imports and supply of exports were defined as the reciprocals of the elasticities of these functions (with opposite sign). This has come to be known as the ‘elasticity approach’. (For an interpretation of these demand and supply prices as prices relative to the price – assumed stabilized – of a non-tradable in a general-equilibrium model, see Chipman, 1978.)

Bickerdike derived formulas for the effect on national ‘advantage’ of a small tariff (p. 100n) and for the optimal tariff (p. 101n), and remarked – anticipating Lerner (1936) – that identical expressions would be obtained for an export tax. He noted that the optimal tariff depended only on the foreign elasticities (see

also Kahn, 1947); this apparent paradox was explained by Graaff (1949, p. 56). The now-familiar, simpler and more general optimal-tariff formula expressed in terms of Marshallian elasticity was first introduced by Johnson (1950), who showed its relation to Bickerdike's formula.

Edgeworth (1908, p. 544) showed that the positive sign of the denominator of Bickerdike's expression for the advantage from an incipient tariff followed from dynamic stability. A related stability condition was later derived by Bickerdike (1920) for the analysis of a regime of fluctuating exchange rates, and was obtained as a condition for a transfer to lower the paying country's exchange rate. Equivalent formulas were subsequently adopted by Robinson (1937, p. 194n) and Metzler (1948), and – for the special case indicated by Bickerdike of infinite elasticities of supply of exports – by Lerner (1944, p. 378).

Bickerdike's other contributions include two essays on local public finance (1902, 1912), a paper (1911) correcting a statement of Edgeworth's that price discrimination could improve upon competitive pricing, and papers on a number of other topics, the most noteworthy relating to business cycles and economic growth.

Although preceded by Carver (1903), Aftalion (1909, pp. 219–20) and Pigou (1912, pp. 144–5), Bickerdike (1914) may be considered one of the original developers of the acceleration principle (cf. Hansen, 1927, p. 112; Haberler, 1937, p. 87), providing a detailed numerical example and emphasizing (in contrast to Aftalion) the importance of durability of capital rather than the gestation period. Bickerdike regarded the phenomenon as an example of market failure. The paper was cited by Frisch (1931) – who erroneously attributed it to J.M. Clark – in the course of his criticism of Clark (1923) and reformulation according to which a deceleration of consumption will call forth a fall in gross investment only if it exceeds the rate of depreciation of capital. Bickerdike (1924; 1925) went on to develop an interesting mathematical model of economic growth according to which labour – the only factor – grows at a constant rate and produces only capital goods – of various durabilities and with various gestation periods – the services of which are consumed. On a path of balanced growth, the rate of interest is equal to the rate of growth, and interest is reinvested. The money supply grows at the same rate in order to maintain constant prices – or else it is constant and prices fall at a constant rate. Bickerdike's main object was to determine whether the process of saving benefited non-savers; in this he was not entirely successful, since his techniques limited him to balanced-growth paths. Nevertheless this work foreshadowed that of Lerner (1944, ch. 20) as well as many features of contemporary growth models, and attracted the attention of Hansen (1927, pp. 173ff).

Information on Bickerdike's life and work may be found in Jha (1963) and in Larson (1983; 1987), where other relevant literature is also cited. According to Larson, after Bickerdike's death his papers, including some 50 letters from Edgeworth and 20 from Edwin Cannan, passed into the hands of one Godfrey Alan Dick who died in Oxford in 1981. They are presumed lost.

See Also

- elasticities approach to the balance of payments
- Marshall–Lerner condition

Selected works

1902. Taxation of site values. *Economic Journal* 12, 472–84. Reprinted in Musgrave and Shoup (1959), 377–88.

1906. The theory of incipient taxes. *Economic Journal* 16, 529–35. Reprinted in Musgrave and Shoup (1959), 132–8.

1907. Review of *Protective and Preferential Import Duties* by A.C. Pigou. *Economic Journal* 17, 98–102.

1911. Monopoly and differential prices. *Economic Journal* 21, 139–48.

1912. The principle of land value taxation. *Economic Journal* 22, 1–15.

1914. A non-monetary cause of fluctuations in employment. *Economic Journal* 24, 357–70.

1920. The instability of foreign exchange. *Economic Journal* 30, 118–22.

1924. Individual and social interests in relation to saving. *Economic Journal* 34, 408–22.

1925. Saving and the monetary system, *Economic Journal* 35, 366–78.

Bibliography

Aftalion, A. 1909. La réalité des surproductions générales, 3rd instalment. *Revue d'économie politique* 23 , March, 201–29.

Carver, T.N. 1903. A suggestion for a theory of industrial depressions. *Quarterly Journal of Economics* 17, 497–500.

Chipman, J.S. 1978. A reconsideration of the 'elasticity approach' to balance-of-payments adjustment problems. In *Breadth and Depth in Economics*, ed. J.S. Dreyer. Lexington, MA: Heath.

Clark, J.M. 1923. *Studies in the Economics of Overhead Costs*. Chicago: University of Chicago Press.

Clark, J.M. 1931. Capital production and consumer-taking – a reply. *Journal of Political Economy* 39, December, 814–16. A further word, 40, October, 691–3.

Edgeworth, F.Y. 1908. Appreciations of mathematical theories, III. *Economic Journal* 18, September, 392–403; December, 541–56. Reprinted as: Mr. Bickerdike's theory of incipient taxes and customs

duties, in F.Y. Edgeworth, *Papers Relating to Political Economy*, vol. 2. London: Macmillan, 1925.

Frisch, R. 1931. The interrelation between capital production and consumer-taking. *Journal of Political Economy* 39, 646–54. A rejoinder, 40, 253–4. A final word, 40, 694.

Graaff, J. de V. 1949. On optimum tariff structures. *Review of Economic Studies* 17, 47–59.

Haberler, G. 1937. *Prosperity and Depression*. 3rd edn, Lake Success, NY: United Nations, 1946.

Hansen, A.H. 1927. *Business-Cycle Theory*. Boston: Ginn & Co.

Jha, N. 1963. *The Age of Marshall*. Patna: Novelty & Co; 2nd edn, London: Frank Cass, 1973.

Johnson, H.G. 1950. Optimum welfare and maximum revenue tariffs. *Review of Economic Studies* 19, 28–35.

Kahn, R.F. 1947. Tariffs and the terms of trade. *Review of Economic Studies* 15, 14–19.

Larson, B.D. 1983. The analysis of interests in the economics of Charles Frederick Bickerdike. Ph.D. dissertation, University of North Carolina, Chapel Hill, NC.

Larson, B. 1987. Bickerdike's life and work. *History of Political Economy* 19, 1–22.

Lerner, A.P. 1936. The symmetry between import and export taxes. *Economica*, n.s. 3, 306–13.

Lerner, A.P. 1944. *The Economics of Control*. New York: Macmillan.

Metzler, L.A. 1948. The theory of international trade. In *A Survey of Contemporary Economics*, ed. H.S. Ellis. Philadelphia: Blakiston.

Musgrave, R.A. and Shoup, C.S., eds. 1959. *Readings in the Economics of Taxation*. Homewood, IL: Irwin.

Pigou, A.C. 1912. *Wealth and Welfare*. London: Macmillan.

Robinson, J. 1937. The foreign exchanges. In Joan Dreyer, *Essays in the Theory of Employment*, London: Macmillan; 2nd edn, Oxford: Basil Blackwell, 1947.

How to cite this article

Chipman, John S. "Bickerdike, Charles Frederick (1876–1961)." The New Palgrave Dictionary of

Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000126>
doi:10.1057/9780230226203.0135

The New Palgrave Dictionary of Economics Online

bimetallism

Lawrence H. Officer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A bimetallic monetary standard is a combination of two metallic standards, each of which could in principle stand alone. Bimetallism has advantages over monometallism; but can be an unstable system, with legal bimetallism becoming de facto monometallism. The Persian and Roman Empires practised bimetallism. England's de facto bimetallism was short-lived, and US bimetallism difficult to maintain. French bimetallism in 1815–73 stabilized the gold–silver market price ratio and also exchange rates among gold, silver, and bimetallic countries. Bimetallism ended in the 1870s.

Keywords

bimetallic arbitrage; bimetallism; deflation; gold standard; Gresham's law; inflation; Latin Monetary Union; market ratio; mint ratio; monetary base; money supply; monometallism; seigniorage; silver standard; specie-flow mechanism

Article

A bimetallic monetary standard is a combination of two metallic standards, each of which could in principle stand alone, and often evolved into de facto monometallism.

The nature of bimetallism

Bimetallic metals are usually gold and silver, but there are exceptions. Ancient Rome was temporarily on a silver-bronze standard; in the 18th century Sweden and Russia experienced a silver–copper standard.
Under bimetallism, both gold and silver coins are full legal tender. The unit of account (dollar, franc, and so on) is defined in terms of a fixed weight both of pure gold and of pure silver. So there is a fixed legal (mint, coinage) gold–silver price ratio: number of grains or ounces of silver per grain or ounce of

gold. Both gold and silver enjoy free coinage (the government prepared to coin bars of either metal deposited by any party) and are full-bodied (have legal or face-value equal to metallic value). Token subsidiary (always silver) coins can exist. Subsidiary coins are fractions of (have face value less than) the unit of account; token coins have face value less than metallic (inherent) value, and invariably have restricted legal-tender power. Token coins were not adopted by bimetallic countries until late in their experience with bimetallism, and in conjunction with the process of terminating that standard.
Private parties may melt, import, and export coins (domestic or foreign) of either metal. There is no restriction on non-monetary uses of the monetary metals. Paper currency and deposits may exist; they are convertible into legal-tender coins, either directly or via government-issued paper currency (itself directly convertible into coin). Both private parties and the government may choose the metallic coin, or mixture of coins, in which to discharge debt (including paper currency). However, a private party does not have the right to a direct governmental exchange of gold for silver, or silver for gold. Logically, though, domestic gold and silver coin would exchange privately at the mint ratio.

Advantages and disadvantages of bimetallism

Bimetallism has four advantages. First, it embodies two sets of coins – one from a metal with a high value–weight ratio (gold), the other from a metal with a low ratio (silver). These provide a medium of exchange for a wide range of economic transactions. The range can be extended in both directions: upper, via paper currency and deposits; lower, via token subsidiary coins. Neither is incompatible with a bimetallic standard. Second, as does a monometallic standard, the bimetallic standard provides a constraint on the money supply and therefore inflation; for the legal-tender coins constitute the monetary base (given government-issued legal-tender paper, perhaps the ‘super monetary base’), and the government must acquire one or the other metal to increase the base. Because there is coinage on demand, there is also a check on reduction to the monetary base, and on deflation. Third, a bimetallic country or bloc of countries accommodates shocks so that resulting effects on monometallic countries’ money supplies are dampened. This is done by stabilizing the gold–silver price ratio (‘market ratio’) on the world market, the bullion market, where non-monetary gold and silver (generally bars) are traded either among themselves or individually for some important currency. Fourth, in stabilizing the market gold–silver price ratio, the bimetallic country or bloc also stabilizes the exchange rates between ‘gold currencies’ and ‘silver currencies’. Otherwise, these exchange rates would fluctuate, defeating one of the usual purposes of metallic standards.
The alleged disadvantage of bimetallism (relative to monometallism) is that it is unstable. Suppose the bimetallic-country's mint ratio initially is in the neighbourhood of the market ratio. A shock in the world supply of one metal can change the market ratio so that the mint ratio is now outside its neighbourhood. If the resulting market ratio is above (below) the mint ratio, then silver (gold) is ‘bad’ money, overvalued at the mint; domestic payments will tend to be made in that, relatively cheaper, coin rather than gold (silver), the ‘good’ money, undervalued at the mint and relatively expensive in the market. Good money will tend to be exported to settle balance-of-payments surpluses, bad money imported to finance balance-of-payments deficits. If the divergence between the market and mint ratio is large, ‘bimetallic arbitrage’ occurs, whereby good money is melted and traded on the bullion market for the bad metal, and the bad metal imported to be coined. In both situations, Gresham's law is operative: bad money drives out good.

Given sustained payments imbalances and/or a large and persistent divergence between the market and mint ratio, bad-money monometallism results. (The good money may be eliminated from the money supply, or circulate at a market-determined value –available only at a premium.) To avoid this, the mint ratio could be altered to remain in conformity with the market ratio. If the mint ratio is under-corrected, monometallism is not stemmed; if the mint ratio is over-corrected, monometallism in the opposite metal can occur. Successive changes in the market ratio can lead to alternating effective gold monometallism and silver monometallism, under the rubric of legal bimetallism. There are costs to such an alternating monetary standard; there are also costs in periodically altering the mint ratio.

Theories of bimetallic stabilization

Stabilizing bimetallic arbitrage occurs as follows. Suppose a shock occurs, new gold discoveries, that decrease the market ratio: the market price of non-monetary gold falls relative to silver. The market ratio now is below the mint ratio, so gold is ‘bad’ (overvalued) and silver ‘good’ (undervalued) money. Silver leaves the monetary system to be sold in the world (bullion) market, with gold purchased with the proceeds and coined. First, the arbitrageurs make a profit: the value of the gold coins they obtain is greater than the value of the silver coins they initially sold. Second, there is increased supply of silver (the appreciated metal) and increased demand for gold (the depreciated metal) in the bullion market – the two transactions constituting one arbitrage transaction. The result is an increase in the market ratio, which rises toward the mint ratio. Thus, the incentive for the arbitrage is eliminated. Third, the composition of the money supply of the bimetallic country changed, with a higher proportion of gold to silver. The bimetallic country stabilized the market ratio (and incidentally the exchange rates between gold and silver currencies), via the endogenous gold–silver composition of its money supply.

This mechanism is effective only to the extent that the bimetallic country has sufficient stock of the undervalued metal to return the market ratio close to the mint ratio, so that the incentive to arbitrage vanishes before monometallism in the overvalued metal results. However, the situation is not so dire, because costs of arbitrage imply ‘gold–silver price–ratio’ points that define a band for the market ratio within which the ratio can fluctuate without triggering bimetallic arbitrage. If the bimetallic-country's commitment to its mint ratio is absolutely credible, then stabilizing speculation exists within the bimetallic-arbitrage band, such that the market ratio turns away from its nearest bound and towards the mint ratio. The situation is analogous to stabilizing speculation within gold-point spreads, under the international gold standard.

Two other forces making for bimetallic stability have been suggested by Marc Flandreau. The first is ‘metal-specific arbitrage’ between the bullion and monetary markets. If a metal depreciates on the bullion market by more than coinage and associated costs, then owners of bars in that metal will coin them in lieu of supplying them to the bullion market. If a metal appreciates by more than melting and associated costs of bringing that coined metal to the market, then holders of coin of that metal will melt them and supply them to the market. The reduced supply of the depreciated metal and increased supply of the appreciated metal act to return the market ratio towards the mint ratio. Unlike bimetallic arbitrage, these are independent transactions. Therefore the costs of metal-specific arbitrage are below the costs of bimetallic arbitrage, and the former provide a ‘metal-specific band’ located within the ‘bimetallic arbitrage band.’ So metal-specific arbitrage is a stabilizing mechanism that becomes operative before bimetallic arbitrage.

The second force involves the bimetallic country (France) transacting with a gold-currency country (England) and a silver-currency country (Germany). There are franc–sterling gold points, and franc–mark silver points. Expressing exchange rates as percentage deviations from parity and specie points in percentage terms, the franc/sterling– franc/mark exchange-rate differential (via triangular arbitrage) proxies the mark/sterling exchange rate. Also, implicit mark–sterling parity (via franc bilateral parities) corresponds to the mint ratio. On the assumption of no bilateral specie-point violations, the mark–sterling exchange rate has as upper (lower) bound the *sum* (negative *sum*) of the franc–sterling export (import) point and the franc–mark import (export) point. Now, the mark–sterling exchange rate is itself a good representation of the gold–silver market price ratio, because the Bank of England (Bank of Hamburg) supports, within a narrow band, a fixed sterling (mark) price of gold (silver). For the market ratio above the mint ratio (parity), so that silver is overvalued, the upper bound correctly involves exporting gold (sterling) and importing silver (marks). The gold–silver market price ratio has a bimetallic-arbitrage band that is approximately double the width of the franc–sterling and franc–mark bilateral specie-point spreads. Hence specie flows to settle and adjust payments imbalances occur prior to bimetallic arbitrage.

Suppose that a bimetallic country has lost all its undervalued (‘good’) metal, so it has become monometallic in its overvalued coinage. Nevertheless, Oppers (2000) shows that a bimetallic-arbitrage band could exist, given that there is a second bimetallic country with a different mint ratio. The two countries’ mint ratios each constitute a bound to the market ratio, with, as usual, a market ratio beyond a bound giving rise to arbitrage that returns the market ratio to the band. For this mechanism to operate, both countries must actually or potentially have large amounts of both coined metals in their money stock, where ‘large’ means relative to shocks in the bullion market.

Bimetallism prior to the 19th century

The Persian Empire had the first bimetallic standard, with a mint ratio of $13\frac{1}{2}$ to 1 (all known mint ratios are in favour of gold) for a long time. This ratio undervalued silver relative to the ratio elsewhere, and presumably merchants took advantage of the price-ratio discrepancies in their regular dealings. The Roman Empire was often gold–silver bimetallic, but periodically debased the coinage. The likely reason was to increase seigniorage rather than to realign the mint ratio in conformity with the market ratio or the mint ratio in other lands. Until the mid-19th century, bimetallism was the legal standard in Europe (including England), though the mint ratio was often altered. Traditionally, the gold–silver price ratio was lower in China and India than in Europe.

England was legally on a bimetallic standard from the mid-13th century, when gold was first coined. The mint ratio was often changed. England was effectively on a silver standard until late in the 17th century, because the British mint ratio was generally below European gold–silver price ratios. Gold coins passed at a market price (in terms of the silver shilling) rather than face value, again indicative of a silver standard. In 1663 the (gold) guinea was coined, with a legal value of 20 (silver) shillings. The silver coins in circulation were in horrible condition, due in part to past debasement, in part to private clipping and sweating of the coins. So the market price of the guinea increased above 20 shillings – to as much as 30 shillings – implying a gold–silver price ratio that effectively overvalued gold relative to Continental ratios. England was in process of switching from an effective silver to an effective gold

standard.
In 1696 silver was recoined, so the coins became full-bodied again, and a ceiling (periodically reduced) was placed on the market price of the guinea. The result was that, for a brief period at the turn of the 18th century, England had effective bimetallism, with full-bodied coins of both metals in circulation. However, gold continued to be overvalued and silver undervalued; silver was exported, gold imported; and a de facto gold standard resulted. It became a de jure standard, via legislations restricting the legal-tender power of silver (1774) and effectively ending free coinage of silver (1816).
The Coinage Act of 1792 placed the United States on a legal bimetallic standard. The mint ratio (15 to 1) – selected because it was approximately the market ratio at the time – turned out to overvalue silver, because the market ratio increased. By 1823 gold had virtually gone from circulation, and an effective silver standard resulted. In 1834 Congress increased the ratio to 16.0022 (in 1837, revised slightly, to 15.9884). From 1834 to 1873, the world gold–silver price ratio was consistently below 16, so the new ratio overvalued gold, and an effective gold standard resulted. However, the export of full-bodied Mexican (silver) dollars and US subsidiary silver protected the circulation of underweight foreign silver pieces, which circulated at face value; so in a sense effective bimetallism continued. Only in the early 1850s, when the market gold–silver price ratio fell (due to gold discoveries and new production), did the United States begin to lose its remaining silver coins. In 1853, to retain the silver, Congress reduced subsidiary coins (below a dollar) to token status, with limited legal-tender power. The United States now was on a de facto gold standard. Legal bimetallism remained until 1873, when coinage of the silver dollar was terminated. One year later, silver was virtually demonetized; all silver coins (including the dollar) were restricted to maximum legal tender of five dollars in any payment.

Bimetallic France in the 19th century

In 1803 France made the franc the monetary unit, and solidified and made effective the mint ratio of $15\frac{1}{2}$ that had been established in 1785. From the end of the Napoleonic Wars until 1873, while France retained that bimetallism, the market gold–silver price ratio remained in the neighbourhood of $15\frac{1}{2}$. (Also, exchange rates among gold, silver, and bimetallic countries were stable.) The stability of the market ratio was remarkable in the face of severe shocks to the bullion market. In the 1850s gold production increased tremendously due to gold discoveries in California and Australia, putting strong downward pressure on the market price ratio. In the 1860s gold production stopped increasing, and exploitation of Nevada silver discoveries put strong upward pressure on the ratio.
The steady market gold–silver price ratio was due primarily to the continued bimetallism of France, which acted as a buffer to shocks and thus stabilized the gold–silver market price ratio. What gave France this power were its large economic size, the substantial amounts of both gold and silver coins in its circulation, and its credible commitment to bimetallism at an unchanged mint ratio. Therefore, French bimetallic arbitrage operated – in the 1850s and early 1860s via gold imported and coined and silver melted and exported, in the later 1860s via the opposite activities. Stabilizing speculation within the bimetallic-arbitrage band, stabilizing bilateral specie flows, and metal-specific arbitrage were also elements in the French stabilization service. In 1865 the French stabilizing force was enhanced by formation of the Latin Monetary Union (LMU), in which France, Belgium, Switzerland, and Italy adopted a common bimetallism.

Some scholars, especially Oppers (1995; 2000), believe, rather, that France underwent serial monometallism, with bimetallism transformed to a de facto silver standard in the 1830s and 1840s, and the latter yielding to a de facto gold standard in the 1860s. Yet a parity band (with stabilizing speculation within the band) existed, with the French mint ratio the lower bound and the US mint ratio the upper bound in 1834–61, followed subsequently by the French ratio the upper bound and the Russian ratio the lower bound. This interpretation of history is doubtful, for the strong propensity to use both metallic currencies was characteristic only of France. Also, Russia's mint ratio was inoperative at the time, as the country had an inconvertible paper currency.

In the early 1860s the future LMU countries, if not on a de facto gold standard, were certainly moving towards it. With the market ratio below the mint ratio, silver was being lost. To protect silver circulation, the individual countries made subsidiary coins token currency; while in 1866 the LMU came into effect, mandating reduction of the silver content and restriction of the legal-tender power of all silver coins except the largest, that is, the five-franc piece, which remained full-bodied.

French, LMU, and world bimetallism ended in the 1870s. The proximate cause was Germany's move to a gold standard, financed by the French indemnity that resulted from the Franco–Prussian War. Germany's release of silver put upward pressure on the gold–silver market price ratio. France was not prepared to accept the gold loss and silver inflow that would result from continued adherence to bimetallism. France (and Belgium) limited silver coinage in 1873, followed by the LMU mandating limits on coinage of the five-franc silver piece in 1874–6. In 1878 coinage of that piece was terminated. The existing five-franc coins retained full legal-tender power. France, along with Belgium and Switzerland, went on a ‘limping’ gold standard, redeeming government-issued paper money in either gold or silver at the discretion of the authority.

See Also

- gold standard
- Gresham's Law
- silver standard

Bibliography

Eichengreen, B. and Flandreau, M.R. 1996. Blocs, zones and bands: international monetary history in light of recent theoretical developments. *Scottish Journal of Political Economy* 43, 398–418.

Einzig, P. 1970. *The History of Foreign Exchange*, 2nd edn. London: Macmillan.

Flandreau, M.R. 1997. As good as gold? Bimetallism in equilibrium. In *Monetary Standards and Exchange Rates*, ed. M.C. Marcuzzo, L.H. Officer and A. Rosselli. London: Routledge.

Flandreau, M.R. 2002. ‘Water seeks a level’: modeling bimetallic exchange rates and the bimetallic band. *Journal of Money, Credit and Banking* 34, 491–519.

Flandreau, M.R. 2004. *The Glitter of Gold*. Oxford: Oxford University Press.

Friedman, M. 1990. Bimetallism revisited. *Journal of Economic Perspectives* 4(4), 85–104.

Gallarotti, G.M. 1994. The scramble for gold: monetary regime transformation in the 1870s. In *Monetary Regimes in Transformation*, ed. M.D. Bordo and F. Capie. Cambridge: Cambridge University Press.

Leavens, D.H. 1939. *Silver Money*. Bloomington, Indiana: Principia Press.

Martin, D.A. 1968. Bimetallism in the United States before 1850. *Journal of Political Economy* 76, 428–442.

Martin, D.A. 1973. 1853: The End of Bimetallism in the United States. *Journal of Economic History* 33, 825–844.

Officer, L.H. 1996. *Between the Dollar–Sterling Gold Points: Exchange Rates, Parity and Market Behavior*. Cambridge: Cambridge University Press.

Oppers, S.E. 1995. Recent developments in bimetallic theory. In *International Monetary Systems in Historical Perspective*, ed. J. Reis. Houndmills and London: Macmillan.

Oppers, S.E. 2000. A model of the bimetallic system. *Journal of Monetary Economics* 46, 517–33.

Redish, A. 1994. The Latin Monetary Union and the emergence of the international gold standard. In *Monetary Regimes in Transformation*, ed. M.D. Bordo and F. Capie. Cambridge: Cambridge University Press.

Redish, A. 2000. *Bimetallism: An Economic and Historical Analysis*. Cambridge: Cambridge University Press.

Velde, F.R. and Weber, W.E. 2000. A model of bimetallism. *Journal of Political Economy* 108, 1210–34.

Wilson, T. 2000. *Battles for the Standard*. Aldershot: Ashgate.

How to cite this article

Officer, Lawrence H. "bimetallism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/

article?id=pde2008_B000137> doi:10.1057/9780230226203.0136

The New Palgrave Dictionary of Economics Online

Black, Duncan (1908 – 1991)

Bernard Grofman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Black, D.; Borda, J.-C. de; Carroll, Lewis; Condorcet, Marquis de; game theory; majority rule; median voter; public choice; single-peaked preferences; spatial voting game; voting paradoxes

Article

Born on 23 May 1908 in Motherwell, Scotland, Black studied at the University of Glasgow, where he received an MA (Mathematics and Physics) in 1929, an MA (Economics and Politics) in 1932, and a Ph. D. (Economics) in 1937. He also served there as Senior Lecturer in Social Economics, 1946–52. The bulk of his teaching career was at the University College of North Wales, Bangor: Lecturer in Economics, 1934–45; Professor of Economics, 1952–68; and Professor Emeritus 1968 onwards.
Black's very early research was in public finance, of which the major work is Black (1939). It is, however, his work in the 1940s and early 1950s (notably Black, 1948a; 1948b; 1948c; 1949; 1950, and Black and Newing, 1951), work which was integrated and expanded in Black (1958), which is the basis for his status as a father of the modern theory of public choice.
More than two centuries ago Condorcet (1785) demonstrated that majority rule need not yield a stable outcome when there are more than two alternatives to be considered. Although periodically rediscovered or reinvented by succeeding generations of scholars, the ‘paradox of cyclical majorities’ was, for all practical purposes, unknown to modern students of democratic theory until called to their attention by Duncan Black (see especially Black, 1948a; 1958). Black demonstrated that the ‘paradox’ was not just a mathematical curiosity but rather was connected to important political issues such as manipulability of voting schemes (1958, p. 44; see also 1948a, p. 29) and the absence of strong similarity of citizen preference structures (Black, 1958, pp. 10–14).

> Although Black was not the first to discover this phenomenon, his work is the foundation of all subsequent research on the problem. The investigations in this field of his principal predecessors, Condorcet and Lewis Carroll, had made no impact on the intellectual community of their day and had been completely forgotten. Their work is known today

> only because Black, after discovering the phenomenon himself, discovered his predecessors. (Campbell and Tullock, 1965, p. 853)

Duncan Black's vision in the 1940s was a grand yet simple one: to develop a pure science of politics as a ramified theory of committees, so as to place political science on the same kind of theoretical footing as economics, with voters substituting for consumers. Because many of the basic ideas in his 1958 classic, *The Theory of Committees and Elections*, appear so ‘obvious’ in retrospect that it is hard to believe that they have not always been part of the stock of general human knowledge, and because this work understates by its silence the magnitude of Black's originality, the magnitude of Black's own contributions is often underappreciated. Black's great strength is that he has served as both synthesizer and pioneer. He rediscovered and reinterpreted for contemporary social science the strikingly modern probabilistic and game theoretic insights of long-dead theorists such as Dodgson (Lewis Carroll), Borda and Condorcet (for example, the paradox of cyclical majorities, the Condorcet criterion, the Borda criterion, optimizing strategies under the limited vote, results on manipulability of voting schemes, the Condorcet jury theorem); while himself developing such seminal ideas as single-peakedness, the importance of the median voter given ordinal preferences, and the notion of equilibrium in a spatial voting game (Black and Newing, 1951; Black, 1958; 1967; 1969; 1976). Black's work on Lewis Carroll (McLean, McMillan and Munroe, 1996) emphasizes Carroll's contributions to logic and the importance of his work on representation (under his real identity, that of the mathematician C.L. Dodgson) as a precursor to the modern theory of games and economic behaviour.

Underpinning virtually all of Black's work was the deceptively simple insight of modelling political phenomena in terms of the preferences of a given set of individuals in relation to a given set of motions, the same motions appearing on the preference schedule of each individual, where motions can be represented as points on a real line or in an N-dimensional space. Black's work on what (after him) has come to be called ‘the theory of committees and elections’ has been ‘one of the pillars on which rests the contemporary theory of public choice’ (Grofman, 1981).

See Also

- Arrow's theorem
- Borda, Jean-Charles de
- Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de
- democratic paradoxes
- social choice
- social choice (new developments)
- voting paradoxes

Selected works

1939. *The Incidence of Income Taxes*. London, Macmillan; reprinted, New York: Kelley, 1965.

1948a. On the rationale of group decision making. *Journal of Political Economy* 56, 23–34.

1948b. The decisions of a committee using a special majority. *Econometrica* 16, 245–61.

1948c. The elasticity of committee decision with an altering size of majority. *Econometrica* 16, 262–70.

1949. The theory of elections in single-member constituencies. *Canadian Journal of Economics and Political Science* 14(2), 158–75.

1950, 1964. The unity of political and economic science. *Economic Journal* 60, 506–14. Repr. in *Game Theory and Related Approaches to Social Behavior*, ed. M. Shubik. New York: Wiley, 1964.

1951. (With R.A. Newing.) *Committee Decisions with Complementary Valuation*. Glasgow: William Hodge.

1958. *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.

1966. A simple theory of non-cooperative games with ordinal utilities. In *Papers on Non-Market Decision Making*, ed. G. Tullock. Charlottesville: Thomas Jefferson Center for Political Economy, University of Virginia.

1967. The central argument in Lewis Carroll's ‘The Principles of Parliamentary Representation’. In *Papers on Non-Market Decision Making III*, ed. G. Tullock. Charlottesville: Thomas Jefferson Center for Political Economy, University of Virginia.

1969. Lewis Carroll and the theory of games. *American Economic Review* 59, 206–16.

1970. Lewis Carroll and the Cambridge Mathematical School of P.R.: Arthur Cohen and Edith Denman. *Public Choice* 8, 1–28.

1976. Partial justification of the Borda Count. *Public Choice* 28, 1–15.

Bibliography

Campbell, C.D. and Tullock, G. 1965. A measure of the importance of cyclical majorities. *Economic Journal* 75, 853–7.

Condorcet, N.C. de. 1785. *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix*. Paris.

Grofman, B. 1981. The theory of committees and elections: the legacy of Duncan Black. In *Towards a*

Science of Politics: Essays in Honor of Duncan Black, ed. G. Tullock. Blacksburg: Public Choice Center, Virginia Polytechnic Institute and State University.

McLean, I., McMillan, A. and Munroe, B.L., eds. 1996. *A Mathematical Approach to Proportional Representation: Duncan Black on Lewis Carroll*. Dordrecht: Kluwer.

How to cite this article

Grofman, Bernard. "Black, Duncan (1908–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000145> doi:10.1057/9780230226203.0137

The New Palgrave Dictionary of Economics Online

Black, Fischer (1938 – 1995)

Perry G. Mehrling
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Fischer Black is best known for the Black–Scholes option pricing formula, which he regarded as an application of the capital asset pricing model (CAPM). He understood the CAPM as a model of general economic equilibrium and extended it from finance to macroeconomics, including the theory of money and the theory of business cycles. His work reveals that finance was the origin of the dramatic changes in macroeconomic thinking in the last quarter of the 20th century.

Keywords

American Finance Association; BDT (Black–Derman–Troy) term structure model; Black, F.; Black–Litterman model; Black–Scholes formula; capital asset pricing model; futures markets; Lucas, R.; Merton, R.; noise trading; option pricing theory; Prescott, E.; real business cycles; Scholes, M.; zero-beta model

Article

Fischer Black is best known for the eponymous Black–Scholes option pricing formula that laid the foundations for so much of modern finance (Black and Scholes, 1973), a contribution that was recognized posthumously in the citation for the 1997 Nobel Prize in Economics that was awarded to Robert C. Merton and Myron Scholes. Today, the best known derivation of the famous formula follows the no-arbitrage argument laid out in Merton (1973), but Black approached the problem as simply an application of the capital asset pricing model (CAPM) developed by Sharpe (1964), Lintner (1965), and especially Jack Treynor (1962), whose version of CAPM was Black's first introduction to finance. Indeed, it is no exaggeration to say that not just the options formula but also everything Black ever wrote has its roots in CAPM, which Black always understood quite broadly as a model of general economic equilibrium, not just a model of how to price risky capital assets (Black, 1972b).
Born 11 January 1938, Fischer Black grew up in Bronxville, New York, before attending both college

and graduate school at Harvard University. After earning his Ph.D. in applied mathematics in 1964 for a thesis in the new area of artificial intelligence, Black took his first job as an analyst in the operations research section of the consulting firm Arthur D. Little, Inc. That's where he met Treynor and learned CAPM. Although he never took even a single course in either economics or finance, Black subsequently built a career as a financial consultant, a research professor (University of Chicago 1971–5, Massachusetts Institute of Technology 1975–83), and then a partner in the Wall Street investment firm Goldman Sachs (1984–95). He died prematurely on 30 August 1995, shortly after the publication of *Exploring General Equilibrium*, the book he considered to be his magnum opus.

Straddling the worlds of academia and business, Black developed his ideas by using practical problems in business as the stimulus for his abstract theorizing. The accessible early paper with Treynor, ‘How to use security analysis to improve portfolio selection’ (Treynor and Black, 1973) set the agenda that would occupy Black and the generation of financial engineers that grew up after him, namely, to find practical applications of the new academic theories of finance. Just so, Black's early work with Myron Scholes for the Wells Fargo Bank sought to develop a new ‘passive’ portfolio strategy from the implications of CAPM, a kind of leveraged index fund that anticipated the later development of portfolio insurance (Black and Scholes, 1974; Black, 1988a; Black and Perold, 1992). Similarly, his paper on ‘Bank funds management in an efficient market’ (1975) anticipated the eventual consequences of bank deregulation, and his paper ‘Toward a fully automated stock exchange’ (1971) anticipated the eventual consequences of computerized trading.

All of this was about remaking the world in the image of CAPM, an image that kept expanding in Black's mind as he worked to extend CAPM to a world without any riskless asset in his famous zero-beta model (1972a), to a world with long-term debt in the famous BDT term structure model (Black, Derman and Toy, 1990; Black, 1995b), and to an international environment in his controversial universal hedging model (1974; 1990) that formed the analytical core of the Black–Litterman model of global asset allocation (Black and Litterman 1991; 1992).

The irony is that the world of the original CAPM is a world of debt and equity only, no options at all. That explains why Black was not sure that the opening in April 1973 of the Chicago Board Options Exchange was a good thing, even though it provided an immediate application for the Black–Scholes formula. Similarly, Black's extension of the options analysis to the problem of pricing commodity futures (1976), although immediately useful in the currency futures markets that sprang up after the collapse of the Bretton Woods fixed exchange rate system, left him unsure whether he was helping to move the world toward CAPM or away from it. From this point of view, his work on pension fund investment policy, the theory of business accounting, and a practical method of capital budgeting more clearly contributed to the creation of a CAPM world (1980b; 1980a, 1993; 1988b).

Only after leaving academia for Goldman Sachs did Black come to fully appreciate the positive contribution of options and other derivatives to the brave new world of finance. The turning point was the theory of noise trading that he revealed for the first time in his presidential address to the American Finance Association (1986). Noise traders are people who trade, knowingly or not, without any information advantage. Earlier in his career, Black had assumed that such traders would eventually be driven out as markets become more and more efficient, but he changed his mind once he realized that ‘Noise trading actually puts noise into prices’. As a consequence, ‘we might define an efficient market as one in which price is within a factor of 2 of value; i.e. the price is more than half of value and less than twice value’ (1986, 532–3). Because of noise trading, psychology matters for asset pricing, and it is

in options prices that this effect can most clearly be seen; it shows up in the Black–Scholes formula as volatility.

Black's intellectual strategy to understand the world through the equilibrium lens of CAPM, as properly extended, was not confined to finance. He also used CAPM to lay the foundations of an alternative equilibrium understanding of macroeconomics, including the theory of money and the theory of business cycles, and he always considered this work at least as important as his work in finance. In this respect, his very first published paper, ‘Banking and interest rates in a world without money: the effects of uncontrolled banking’ (1970), set the agenda that would occupy him for the rest of his life. His two subsequent books *Business Cycles and Equilibrium* (1987) and *Exploring General Equilibrium* (1995a) had little impact on economics at the time they were published. In retrospect, however, they can be seen to have anticipated themes that eventually did enter economics, through the new classical revolution of Robert Lucas and his associates and the real business cycle revolution of Edward Prescott and his associates. More than anyone else, Fischer Black demonstrated that we must look to finance to discover the origin of the dramatic changes in macroeconomic thinking in the last quarter of the 20th century (Mehrling, 2005).

See Also

- accounting and economics
- capital asset pricing model
- efficient markets hypothesis
- futures markets, hedging and speculation
- growth and cycles
- Merton, Robert C.
- Miller, Merton
- money and general equilibrium
- noise traders
- options
- real business cycles
- Scholes, Myron
- term structure of interest rates

Selected works

1970. Banking and interest rates in a world without money: the effects of uncontrolled banking. *Journal of Bank Research* 1, 8–20. Reprinted as ch. 1 in Black (1987).

1971. Toward a fully automated stock exchange. *Financial Analysts Journal* 27, Part I, 28–35, 44; Part II, 24–28, 86–7.

1972a. Capital market equilibrium with restricted borrowing. *Journal of Business* 45, 444–5.

1972b. Equilibrium in the creation of investment goods under uncertainty. In *Studies in the Theory of Capital Markets*, ed. M. Jensen. New York: Praeger.

1973. (With M. Scholes.) The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

1973. (With J. Treynor.) How to use security analysis to improve portfolio selection. *Journal of Business* 46, 66–86.

1974. International capital market equilibrium with investment barriers. *Journal of Financial Economics* 1, 337–52.

1974. (With M. Scholes.) From theory to a new financial product. *Journal of Finance* 19, 399–412.

1975. Bank funds management in an efficient market. *Journal of Financial Economics* 2, 323–39.

1976. The pricing of commodity contracts. *Journal of Financial Economics* 3, 167–79.

1980a. The magic in earnings: economic earnings versus accounting earnings. *Financial Analysts Journal* 36, 19–24.

1980b. The tax consequences of long-run pension policy. *Financial Analysts Journal* 36, 21–8.

1986. Noise. *Journal of Finance* 41, 529–43.

1987. *Business Cycles and Equilibrium*. Cambridge, MA: Basil Blackwell.

1988a. Individual investment and consumption under uncertainty. In *Portfolio Insurance: A Guide to Dynamic Hedging*, ed. D. Luskin. New York: John Wiley and Sons.

1988b. A simple discounting rule. *Financial Management* 17, 7–11.

1990. Equilibrium exchange rate hedging. *Journal of Finance* 45, 899–907.

1990. (With E. Derman and W. Toy.) A one-factor model of interest rates and its application to Treasury bond options. *Financial Analysts Journal* 46, 33–9.

1991. (With R. Litterman.) Asset allocation: combining investor views with market equilibrium. *Journal of Fixed Income* 1, 7–18.

1992. (With R. Litterman.) Global portfolio optimization. *Financial Analysts Journal* 48, 28–43.

1992. (With A. Perold.) Theory of constant proportion portfolio insurance. *Journal of Economic Dynamics and Control* 16, 403–26.

1993. Choosing accounting rules. *Accounting Horizons* 7, 1–17.

1995a. *Exploring General Equilibrium*. Cambridge, MA: MIT Press.

1995b. Interest rates as options. *Journal of Finance* 50, 1371–6.

Bibliography

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 13–37.

Mehrling, P. 2005. *Fischer Black and the Revolutionary Idea of Finance*. Hoboken, NJ: John Wiley and Sons.

Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–83.

Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Treynor, J. 1962. Toward a theory of market value of risky assets. In *Asset Pricing and Portfolio Performance*, ed. R. Korajczyk. London: Risk Books.

How to cite this article

Mehrling, Perry G. "Black, Fischer (1938–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000300> doi:10.1057/9780230226203.0138

The New Palgrave Dictionary of Economics Online

black – white labour market inequality in the United States

Derek Neal

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

During much of the 20th century, each successive generation of black Americans came closer to its white counterparts in terms of educational achievement and labour market success. This pattern of black–white progress has stalled since the mid-1980s. This chapter documents the current levels of black-white inequality in terms of human capital and labour market outcomes and then discusses factors that may sustain and perpetuate current levels of black–white inequality. It is much easier to understand the record of black–white progress during earlier decades than to understand the lack of progress in recent years.

Keywords

affirmative action; black–white educational achievement; black–white labour market inequality in the United States; black–white skill gap; human capital investment; incarceration; inequality (explanations); intergenerational transmission of human capital; Jim Crow South; labour force participation rate; labour market discrimination; labour migration; Myrdal, G.; National Assessment of Educational Progress; National Center for Education Statistics; National Longitudinal Survey of Youth; skill investment; unemployment; women's work and wages

Article

Gunnar Myrdal won a Nobel Prize in economics in large measure for path-breaking work that documented the magnitude and scope of black–white inequality in the United States prior to the Second World War. Blending social science with social commentary, Myrdal argued that the contrast between American ideals and the existing legal and social institutions that oppressed blacks created *An American Dilemma* (1944) that was moral and social as well as economic.

A record of progress

In subsequent decades, blacks have made much relative economic progress in the United States, but the pace of this progress has not been steady. For example, during the 1940s, 1960s, and 1970s the earnings of black men rose rapidly relative to those of white men, but this did not occur during the 1950s, 1980s, or 1990s. In fact, in recent decades the pace of relative economic progress for blacks has slowed and may be on the verge of stalling completely.

Table 1 presents data on the black–white earnings gap. The data come from the 1940–2000 decennial census files. Inconsistencies in the survey instrument as well as data-quality problems in some years make it difficult to create a consistent measure of hourly wages across census years. Here, I present data on annual labour earnings for workers who report working at least 48 weeks in the previous calendar year. For each year, numbers are given, separately for men and women, of the black–white ratio of average earnings and of the average percentile rank that black workers would have occupied in the white earnings distribution. I restrict the samples to ages 26–46 to minimize the number of lost of observations due to schooling or early retirement.

Black–white ratio of average annual earnings and average black percentile in the white earnings distribution

	Men		Women	
Year	Ratio	Percentile	Ratio	Percentile
1940	0.45	0.167	0.39	0.126
1950	0.61	0.226	0.58	0.227
1960	0.60	0.214	0.63	0.268
1970	0.65	0.268	0.82	0.399
1980	0.73	0.343	0.97	0.494
1990	0.72	0.361	0.93	0.484
2000	0.70	0.367	0.88	0.464

Note: Data are from the Integrated Public Use Microdata Series (IPUMS) decennial census 1940–2000. The sample includes individuals between the ages of 26 and 45 who report positive wage and salary income and working at least 48 weeks in the previous calendar year. Sample weights ‘slwt’ are used for 1940 and 1950 and ‘perwt’ for 2000.

The results in Table 1 echo a common theme in the literature on black–white inequality. The 1960s and 1970s were decades when blacks made exceptional labour-market gains relative to whites both in terms of their position in the distribution of earnings and in terms of earnings levels. A significant literature debates whether government action during and after the civil rights era was a catalyst for black progress during the 1960s and into the 1970s. Smith and Welch (1989) and others emphasize the role of long-term improvements in the quantity and quality of black education as sources of black economic progress during the 20th century (see Card and Kruger, 1992; 1996). While not disputing the importance of relative improvements in black education, Donohue and Heckman (1991) build a compelling case that federal government intervention did play a significant role in black progress during the civil rights era. They stress that black relative earnings rose significantly during the 1960s and 1970s within cohorts who were already adults at the beginning of these decades. They also note that black relative earnings rose

during the 1960s primarily because of gains in the South, where civil rights laws were imposed on local communities by the federal government. Finally, they note that the decades-long wave of massive net black migration from the South to northern cities came almost to a complete stop around 1965. This one fact is strong prima facie evidence that the Civil Rights Act of 1964 did improve economic opportunity for blacks in the South.

Progress stalled

The results in Table 1 also indicate that black economic progress since 1980 has been mixed at best. The male black–white earnings ratio fell slightly between 1980 and 2000, but black men did enjoy modest improvements in their relative position in the male earnings distribution over the 1980s and 1990s. (A dramatic increase in earnings dispersion over the period accounts for the different trends in these two measures of black–white earnings inequality among men.) Black women actually lost ground relative to white women according to both relative earnings measures over the 1980–2000 period.

However, it is not clear that black men fared better than black women relative to their white peers over this period. Neal (2004) points out that, even though black and white women have had similar labour force participation rates for several decades, racial differences in patterns of selection suggest that measured black–white earnings and wage gaps among women understate actual gaps in earnings opportunities. This bias arises because white women who do not work are more likely to be well-educated and married to a working spouse while black women who do not work are more likely to be single, less educated mothers receiving means-tested public assistance. The importance of this bias may have diminished since 1980 as government assistance to single mothers has decreased and the number of married career women has increased.

Further, the results in Table 1 are likely to overstate how well black men have fared relative to white men since 1980. Table 2 presents employment rates and institutionalization rates for black and white men by age group and year of birth. Each diagonal row presents results from a particular census year, that is, 1980, 1990 or 2000. The employment rates refer to the past calendar year, and the institutionalization rates refer to the census date. Table 2 shows that the fraction of men who worked during the past calendar year has declined among both blacks and whites in recent decades (see Chandra, 2000, for more details on patterns of male labour force participation by race). However, the rate of decline is much more dramatic among black men. By 2000, roughly 30 per cent of prime-age black men did not report any market work in the previous year. Further, in all age groups the relative decline in black employment rates is more than five percentage points. Thus, while Table 1 shows that black male workers continued to improve their position in the earnings distribution relative to working white men during the 1980–2000 period, it is not certain that black men continued to make relative gains in the distribution of potential earnings.

(1) Fraction worked last calendar year (2) Fraction institutionalized

	White male age group				Black male age group			
Year of birth	26–30	31–35	36–40	41–45	26–30	31–35	36–40	41–45
1935–1939				0.938				0.820

				0.007					0.019
1940–1944			0.945				0.829		
			0.007				0.028		
1945–1949		0.947		0.927		0.822		0.779	
		0.007		0.008		0.039		0.041	
1950–1954	0.941		0.932		0.800		0.774		
	0.009		0.010		0.050		0.065		
1955–1959		0.933		0.888		0.756		0.709	
		0.013		0.012		0.081		0.068	
1960–1964	0.926		0.891		0.747		0.717		
	0.016		0.016		0.101		0.093		
1965–1969		0.898				0.715			
		0.018				0.116			
1970–1974	0.897				0.699				
	0.017				0.119				

Notes: Data for this table are from the decennial census IPUMS 1980–2000. The table displays the fraction of males who worked last year and fraction of males institutionalized. In order to be counted as working in the previous calendar year, a respondent must have (*a*) an affirmative, non-allocated response to the question ‘Did this person work …[during the previous calendar year]?’ or (*b*) positive, non-allocated weeks worked or (*c*) positive non-allocated earned income or (*d)* positive, allocated weeks worked and a non-allocated indication of working since 1 January of the census year in question. Sample weights ‘perwt’ are used for 2000.

The most certain inference that one can draw from Table 2 is that the population of institutionalized black men has grown dramatically since 1980. In addition, since most institutionalized young adult men are incarcerated, Table 2 suggests that roughly one in ten black men aged 26–35 was housed in some type of prison or jail when the 2000 census was taken. (Neal, 2006, shows that this rate is much higher among less-educated black men and dramatically lower among black college graduates.) Taken as a whole, Tables 1 and 2 suggest that black economic progress relative to whites has been anaemic at best since 1980.

Neal (2006) points out that, around 1990, black–white gaps in both educational attainment and achievement stopped closing among young adults and youth respectively. Thus, roughly since the mid-1980s, black youth and young adults have either barely kept pace or fallen farther behind their white peers with respect to numerous measures of human capital, such as achievement scores, total grade attainment, college graduation rates, and work experience. The National Assessment of Educational Progress, 2004, Long Term Trend scores provides some suggestive evidence that since 1999 black children have again begun to close the black–white gap in reading scores, but there is at best weak evidence of renewed progress in math. Overall, black–white math and reading gaps in 2004 among 9- and 13- year-olds are quite similar to the gaps observed in the late 1980s (NCES, 2005).

The recent stability of black–white gaps in educational attainment and measured cognitive skills is an alarming development because the black–white skill gap is an important source of economic inequality between blacks and whites. Neal and Johnson (1996) and Johnson and Neal (1998) show that a large portion of black–white differences in earnings and wages can be accounted for by differences in basic reading and math skills among teenagers that pre-date labour market entry. Black–white skill gaps are a driving force behind black–white differences in labour market outcomes among adults for several reasons. First, the black–white skill gap among the current generation of adults is quite large. For example, respondents in the National Longitudinal Survey of Youth (NLSY), 1979, are in their forties now, and the black–white gap in Armed Forces Qualifying Test (AFQT) scores for this sample was over one standard deviation. (The black–white AFQT gap is smaller among youth tested as part of the NLSY, 1997, but the gap remains close to one standard deviation.) Second, measured labour market returns to skill are now at historical highs in the United States. Third, the current market gradients between labour market outcomes and various measures of human capital are even steeper for blacks than for whites. Black and white high-school dropouts, on average, experience markedly different labour market outcomes but, among persons with a college degree and strong reading and math skills, race is much less salient as a predictor of labour market outcomes (Neal, 2006). Because the black–white skill gap is so costly to the current generation of black adults, economists are hard-pressed to explain the recent stability of the black–white skill gap. The 20th century saw several generations of black children make important human capital gains relative to their white peers during times when public expenditures on schooling and pre-school programmes available to black communities were not nearly as high as they are now relative to comparable spending in white communities and when government did much less to ensure that skilled blacks would be treated fairly in the labour market as adults.

What went wrong?

This record of progress is a key starting place for discussing black–white inequality. The logic of basic models of the intergenerational transmission of human capital suggests that one should expect black–white skill convergence. Because the time and attention of each child is a fixed factor in the production of the child's human capital, there are decreasing returns to investments in any child. Thus, in the absence of spillover effects, any group of parents who are more skilled than some other group of parents by a factor k must invest more than *k* times as much in their children to maintain the same inter-group skill gap in the next generation. In many models, diminishing returns forces skill convergence between two groups unless there is a barrier that hinders investment among one group. The challenge for economists is to understand what barriers are present now in the black community that were not present during 1940–90.

Economists have put forth several theories concerning potential obstacles to skill investment by blacks. None fits all the facts. Coate and Loury (1993) described a model of statistical discrimination in which blacks do not invest because they expect employers to be less likely to reward them for investing. Employers do not see investment levels but rather a noisy signal of worker skill. Because employers believe that black workers are less likely to invest, they screen black workers more stringently, thus lowering the returns to black skill investments, as black workers anticipated. Further, the rational reluctance of black youth to invest confirms the beliefs of employers concerning black investment

behaviour.

The Coate and Loury model has been quite influential because it provides an elegant theory of endogenous racial differences in human capital and labour earnings. However, the model is squarely at odds with a key feature of data on skills and labour market outcomes. As I note above, gradients between earnings and wages on the one hand and measures of achievement and attainment on the other are almost always as steep among blacks as among whites, and often steeper. This directly contradicts the scenario described in Coate and Loury (1993), and one cannot rescue their approach by arguing that the gradients observed in the data do not necessarily answer counterfactual questions concerning what less-skilled blacks would have earned if they had invested in skills. This model and others that explain statistical discrimination as a coordination failure are describing a market equilibrium and the resulting market gradients between skill and earnings in that equilibrium. However, no study has yet shown that there exists a gradient between any measure of labour market success and some dimension of worker skill that is systematically steeper among whites than among blacks in the post-civil rights era. (Precise tests of the model are difficult because the skill in question should be observed by the econometrician but not by employers. Nonetheless, blacks do enjoy equal or greater measured returns to the measures of skill and attainment available in current data sets; see Neal, 2006; Levy, Murnane and Willett, 1995.)

A satisfactory explanation of the recent stagnation of black–white skill gaps must begin on the supply side by describing the factors that raise the cost of investing in skills within the black community. Recent work by Austen-Smith and Fryer (2005) provides a model of ‘acting white’. In their model, loss of social cooperation constitutes an additional cost of human capital investment in the black community, and only the most gifted in the community actually invest. This model can produce the steep gradients that we observe between skills and both earnings and wages in the black community because blacks who invest in market skills enjoy expected returns from these investments that are high enough to offset any social sanctions they may suffer. However, the basic argument advanced by Austen-Smith and Fryer (2005) cannot account for all we know about black–white skill differences. Their model is presented as a description of peer pressure, but black–white skill gaps are quite large when children begin school, widen during elementary school, and do not increase much if at all after students enter high school (Neal, 2006). The gaps that exist prior to school entry are more likely to be connected to black–white differences in home environment than black–white differences in peer interactions. Further, it is not obvious why fears of being sanctioned for ‘acting white’ should have a more deleterious effect on black achievement during elementary school than during the teen years. Finally, if the social stigma of ‘acting white’ is sustaining the large black–white gaps in achievement and attainment that remain in 2005, we may need to think more carefully about potential sources of change in black culture during recent decades. It is logically possible but hard to imagine that the dramatic black progress observed during the 1940–90 period could have taken place in black communities where achievement and attainment were accompanied by sanctions for ‘acting white’.

Because black–white skill gaps are quite large even among young children, it is natural to examine the roles of parents and families when trying to understand why recent cohorts of black children have failed to continue closing the black–white skill gap. Neal (2006) discusses changes in the wage structure and contemporaneous changes in family structure within the black community since 1980 that have reduced the resources available to children in black families. These changes may have adversely affected

investment in black children, and if this is the case, the recent stability of the black-white skill gap will be temporary. Negative shocks to black wealth should only slow the process of black–white skill convergence. Even in models with imperfect credit markets, the standard expectation is that pure wealth effects will not persist indefinitely over generations. (See Loury, 1981, and Mulligan, 1997. Neal, 2006, provides a detailed discussion of factors that influence black–white skill convergence.)

Recent studies of parenting behaviours do indicate that there are important black–white differences in ways that parents interact with children and that these differences contribute to black–white differences in cognitive development at an early age (see Brooks-Gunn, Duncan and Klebanov, 1996; Brooks-Gunn et al., 1998), but it is not clear whether these parenting differences should be understood as differences in culture or differences in parenting practices that are driven by differences in family resources.

Conclusion

In closing, I must note that black workers may well face problems other than skill deficits. In particular, the extremely low earnings and employment levels currently observed among- less-skilled black men may be more than the results of an interaction between low skill levels and economy-wide shifts in labour demand that favour skilled labour. Mailath, Samuelson and Shaked (2000) construct an informative model of discrimination against minority groups based on search behaviour, and in their model equilibria exist in which members of minority groups suffer wage discrimination and higher rates of unemployment because employers direct search effort to networks populated by majority group members. Because minority workers and firms know that employers are not searching in minority networks, minority workers have little bargaining power when they do create an encounter with an employer through their own search efforts. In this model, affirmative action policies that mandate colour-blind search eliminate inter-group wage differences because they give all workers the same bargaining power.

In light of the Mailath, Samuelson and Shaked model, consider the real possibility that skilled labour markets may be more heavily influenced by government anti-discrimination efforts. (There is suggestive evidence that this is the case; see Smith and Welch, 1984; Leonard, 1990. Further, Holzer, 1998, provides evidence that large firms, which tend to hire more skilled workers and use formal hiring methods, are significantly more likely to hire black workers than small firms.) If so, the forces identified by Mailath, Samuelson and Shaked are a potential reason that less-skilled blacks fare so much worse relative to their white peers than highly skilled blacks. Further, the Mailath, Samuelson and Shaked reasoning helps us understand why gradients between skill and labour market outcomes have been relatively steep in the black community following the Civil Rights Act, but not before. (Welch, 1973, was the first to note this reversal; see Neal, 2006, for later results.)

Current black–white inequality is much less extreme than the inequality Myrdal observed, but the black–white inequality that remains is more ominous in some respects. The destitution of Southern blacks that Myrdal wrote about was clearly related to direct and oppressive action on the part of state and local governments that intentionally limited the educational and economic opportunities available to black citizens. Nonetheless, blacks made substantial economic and educational progress in the 1940s, and a combination of legal challenges and legislative efforts gradually began to undercut the systems of school financing and Jim Crow employment practices that afflicted blacks so greatly. In contrast, at the

beginning of the 21st century blacks no longer face overt government oppression. Yet, since the mid-1980s, black–white differences in potential wages and earnings have remained roughly constant or grown slightly, incarceration rates among black men have exploded, and black–white skill gaps have remained large and roughly constant. We still face *An American Dilemma*, but the primary causes of our current dilemma and the policy changes necessary to foster further progress are less clear than in Myrdal's day. The current experiences of blacks in the United States present a challenge for economists who wish to understand the dynamics of group outcomes within developed economies.

See Also

- affirmative action
- inequality (global)
- wage inequality, changes in

I thank Ana Rocca, Marie Tomarelli, and Richard Olson for research assistance. I thank Gadi Barlevy, Roland Fryer, and Harry Holzer for useful comments. I thank the Searle Freedom Trust for research support.

Bibliography

Austen-Smith, D. and Fryer, R., Jr. 2005. An economic analysis of acting white. *Quarterly Journal of Economics* 120, 551–83.

Brooks-Gunn, J., Duncan, G. and Klebanov, P. 1996. Ethnic differences in children's intelligence test scores: role of economics deprivation, home environment and maternal characteristics. *Child Development* 67, 396–408.

Brooks-Gunn, J., Phillips, M., Duncan, G., Klebanov, P. and Crane, J. 1998. Family background, parenting practices, and the black–white test score gap. In *The Black–White Test Score Gap*, ed. C. Jencks and M. Philips. Washington, DC: Brookings Institution.

Card, D. and Kreuger, A. 1992. School quality and black–white relative earnings: a direct assessment. *Quarterly Journal of Economics* 107, 151–200.

Card, D. and Krueger, A. 1996. School resources and student outcomes: an overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspectives* 10(4), 31–50.

Chandra, A. 2000. Labor-market dropouts and the racial wage gap: 1940–1990. *American Economic Review* 90, 333–8.

Coate, S. and Loury, G. 1993. Will affirmative-action policies eliminate negative stereotypes? *American*

Economic Review 83, 1220–40.

Donohue, J. III. and Heckman, J. 1991. Continuous versus episodic change: the impact of civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29, 1603–43.

Holzer, H. 1998. Why do small establishments hire fewer blacks than large ones? *Journal of Human Resources* 33, 896–914.

IPUMS (Integrated Public Use Microdata Series). Online. Available at http://www.ipums.org/usa, accessed 3 February 2006.

Johnson, W. and Neal, D. 1998. Basic skills and the black–white earnings gap. In *The Black–White Test Score Gap*, ed. C. Jencks and M. Philips. Washington, DC: Brookings Institution.

Leonard, J. 1990. The impact of affirmative action regulation and equal employment law on black employment. *Journal of Economic Perspectives* 4(4), 47–63.

Levy, F., Murnane, R. and Willett, J. 1995. The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* 77, 251–66.

Loury, G. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49, 843–67.

Mailath, G., Samuelson, L. and Shaked, A. 2000. Endogenous inequality in integrated labor markets with two–sided search. *American Economic Review* 90, 46–72.

Mulligan, C. 1997. *Parental Priorities and Economics Inequality*. Chicago: University of Chicago Press.

Myrdal, G. 1944. *An American Dilemma*. New York: Harper and Brothers.

NCES (National Center for Education Statistics). 2005. Long-term trend major results. Online. Available at http://nces.ed.gov/nationsreportcard/ltt/results2004/, accessed 19 February 2006.

Neal, D. 2004. The measured black–white wage gap among women is too small. *Journal of Political Economy* 112, S1–S28.

Neal, D. 2006. Why has black–white skill convergence stopped? In *The Handbook of Economics of Education*, vol. 1, ed. E. Hanushek and F. Welch. Amsterdam: North-Holland.

Neal, D. and Johnson, W. 1996. The role of pre–market factors in black–white wage differences. *Journal of Political Economy* 104, 869–95.

Smith, J.P. and Welch, F. 1984. Affirmative action in labor markets. *Journal of Labor Economics*, 2, 269–301.

Smith, J. and Welch, F. 1989. Black economic progress after Myrdal. *Journal of Economic Literature* 27, 519–64.

Welch, F. 1973. Black–white differences in returns to schooling. *American Economic Review* 63, 893–907.

How to cite this article

Neal, Derek. "black–white labour market inequality in the United States." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000306>
doi:10.1057/9780230226203.0139

The New Palgrave Dictionary of Economics Online

Blanqui, J é rôme-Adolphe (1798 – 1854)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Blanqui, J.-A.; economic history; history of economic thought; Say, J.-B.

Article

French labour economist, economic historian and first major historian of economic thought, Blanqui was born in Nice and educated both there and in Paris, subsequently teaching humanities at the Institution Massin. His teaching brought him into contact with J.B. Say, who ‘wished him for a disciple’ (Blanqui, 1880, p. ix) and to whose chair of political and industrial economy at the Conservatoire des Arts and des Métiers he succeeded in 1833. In addition, he was head of the Ecole Speciale du Commerce from 1830 to 1854, first editor of the *Journal des économistes* and from 1846 to 1848 served as member for Bordeaux in the Chamber of Deputies. In 1838 he was elected to the Académie des Sciences Morales et Politiques. He died in 1854 in Paris, more than a quarter of a century before his notorious younger brother, Louis Auguste, the revolutionary and member of the Paris Commune, with whom he is often confused.

Blanqui was a prolific writer but is now mainly remembered for his *Histoire de l’économie politique en Europe* (1837) which went through five editions. This is generally regarded as the first major history of political economy. In addition to doctrinal history it covered an enormous amount of economic history from the ancient world to the early 1840s. McCulloch (1845, p. 25) states that Blanqui's ancient economic history is ‘brief and superficial; but his accounts of the political economy of the middle ages and modern times are more carefully elaborated, interesting and valuable.’ Blanqui's treatment of history reflects his support of free trade and sympathy for the working class. Schumpeter (1954, p. 498, n.18) praises Blanqui's 1826 *Resumé de l'histoire du commerce et de l'industrie* as a valuable historical monograph, while his *Précis élémentaire d’économie politique* is also worthy of notice.

Selected works

1837. Blanqui, J.A. *Histoire de l'économie politique en Europe depuis les anciens jusqu'á nos jours*. English trans. by E.J. Leonard as *History of Political Economy in Europe*. New York: G.P. Putnam's sons, 1880.

Bibliography

McCulloch, J.R. 1845. *The Literature of Political Economy*. London: LSE Reprint, 1938.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen & Unwin, 1959.

How to cite this article

Groenewegen, Peter. "Blanqui, Jérôme-Adolphe (1798–1854)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000149>
doi:10.1057/9780230226203.0140

The New Palgrave Dictionary of Economics Online

Bodin, Jean (1530 – 1596)

D.P. O'Brien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bodin, J.; inflation; just price; monetary economics, history of;; quantity theory of money; Salamanca School; scholastic economics

Article

Jean Bodin was born at Anger, France, in 1530 and died of plague at Laôn in 1596.
Bodin is chiefly famous to a wider public for works in history and philosophy. His first work to attract widespread attention has become known in English as *Method for the Easy Comprehension of History* (1566). But his *Republic* (1576), which deals with sovereignty as well as social justice (including proportional taxation), is generally regarded as his masterpiece. However, it is Bodin's work on inflation which is the most important part of his output for economists.
In developing this part of his work, Bodin had as background two key elements. The first was the 16th-century European inflation, triggered by imports of silver from the New World. Remarkable work by the American economic historian Earl J. Hamilton indicates something like a fourfold rise in prices in Spain during the 16th century (Hamilton, 1934, 390–1, 493; see also Hauser, 1932, xi–xix, xlvii–xlix). The Spanish inflation necessarily spread to Spain's immediate trading partner France, through official channels, informal ones (including smuggling), and piracy (Hauser, 1932, xix–xxiv).
The second factor underlying Bodin's work was the contribution of Scholastic writers, stemming initially from an analysis of the effects of debasement, itself building upon the doctrine of the Just Price as founded on relative scarcity in a competitive market. If debasement of the currency increased its nominal amount, its relative scarcity would decrease accordingly. A leading member of the School of Salamanca, Martín de Azpilcueta Navarro (1493–1586), applied this to money in general, whether debased or not, arguing that the purchasing power of money was inversely related to its quantity (Grice-Hutchinson, 1952, 94–5).
Following Scholastic procedures, Bodin developed his own monetary analysis in the form of a critique of *Paradoxes* put forward by a writer called Malestroit. Malestroit's basic thesis was that, while prices

had risen in terms of currency units as a result of debasement, they had not risen in terms of the precious metals. Utilizing data on changes in the price of land, Bodin's estimate of monetary inflation arising from depreciation of precious metals was in excess of 2.5 times, which is remarkably close to the level of 3.0 calculated by 20th-century economic historians.

Bodin's analysis of this inflation involved a treatment of the demand for money (he argued that this depended on the stage of economic development); of the importance of changes in the supply of money; of the idea that the money market clears; of disturbances to either demand for or supply of money producing price and/or income changes; and of the direction of causality running clearly from monetary disturbances to the price level. All of these elements can be found in Bodin's response to Malestroit.

He had thus arrived at an important statement of the quantity theory. He did not claim that the fall in the value of silver was the sole cause of inflation; he certainly recognized the importance of debasement, and mentioned also monopolies, scarcity due to exports, and fashionable demand. But the increased supply of precious metals in France was of key importance.

Finally, Bodin recognized that inflation created economic uncertainty and interfered with economic activity. While changes in the supply of precious metals had to be treated as exogenous disturbances, inflation resulting from debasement should be checked, and he put forward a detailed case for currency reform.

See Also

- inflation
- just price
- monetary economics, history of
- scholastic economics

Selected works

1566. *Methodus, ad facilem historiarum cognitionem*. Paris: M. Iuuenem. Trans. B. Reynolds as *Method for the Easy Comprehension of History*. New York: Columbia University Press, 1945.

1568. *La Response de Maistre Jean Bodin Advocat en la cour au paradoxes de Monsieur de Malestroict, touchant l'enrichissement de toutes choses, & le moyen d'y remédier*. Paris: Martin le Jeune. Trans by H. Tudor and R. Dyson as *Responses to the Paradoxes of Malestroit*, with an introduction by D. O'Brien. Bristol: Thoemmes, 1997.

1576. *Les Six Livres de la Republique de I. Bodin*. Paris: I. du Puys. Trans. R. Knolles as *Six Books of a Commonwealth*. London: Bishop, 1606.

Bibliography

Grice-Hutchinson, M. 1952. *The School of Salamanca: Readings in Spanish Monetary Theory 1544–*

1605. Oxford: Clarendon.

Hamilton, E. 1934. *American Treasure and the Price Revolution in Spain*. Cambridge, MA: Harvard University Press.

Hauser, H. 1932. *La Response de Jean Bodin*. Paris: Armand Colin.

O'Brien, D. 2000. Bodin's analysis of inflation. *History of Political Economy* 32, 267–92.

How to cite this article

O'Brien, D. P. "Bodin, Jean (1530–1596)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000155> doi:10.1057/9780230226203.0141

The New Palgrave Dictionary of Economics Online

Böhm-Bawerk, Eugen von (1851 – 1914)

K.H. Hennings
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Austrian School; Austrian theory of capital; Böhm-Bawerk, E. von; capital deepening; Capital theory; classical economics; distribution theory; exploitation; Fisher, I.; goods theory; imputation; interest theory; intertemporal exchange; intertemporal preferences; intertemporal theory of value; inverse demand schedules; Knies, K. G. A.; labour theory of value; marginal productivity theory; market power; market rate of interest; Marxian economics; Menger, C.; neoclassical economics; price formation; roundabout methods of production; Schäffle, A. E. F.; shadow pricing; stationary state; subjective rates of interest; subsistence fund; time; time preference; Wicksell effects; Wicksell, J. G. K.

Article

As civil servant and economic theorist, Böhm-Bawerk was one of the most influential economists of his generation. A leading member of the Austrian School, he was one of the main propagators of neoclassical economic theory and did much to help it attain its dominance over classical economic theory. His name is primarily associated with the Austrian theory of capital and a particular theory of interest. But his prime achievement is the formulation of an intertemporal theory of value which, when applied to an exchange economy with production using durable capital goods, yields a theory of capital, a theory of interest, and indeed a theory of distribution in which the time element plays a crucial role. Both this construction and his equally famous critique of Marx's economics strongly influenced the development of economic theory from the 1880s until well into the 1930s.
Eugen Böhm Ritter von Bawerk was born in Brünn (now Brno) in Moravia on 12 February 1851, the youngest son of a distinguished civil servant who had been ennobled for his part in quelling unrest in Galicia in 1848, and who died in 1856 as deputy governor and head of the Imperial Austrian administration in Moravia. After reading law at the University of Vienna, Böhm-Bawerk entered the prestigious fiscal administration in 1872. In 1875, however, after taking his doctorate in law, Böhm-Bawerk obtained a government grant to do graduate work abroad and prepare himself for a teaching

position in economics at an Austrian university, as did his classmate and future brother-in-law Friedrich von Wieser. He worked for a year at Heidelberg with Karl Knies, and spent a term each at Leipzig, where Roscher taught, and at Jena, where Hildebrand taught. After working for another three years in the fiscal administration and the ministry of finance, he obtained his *Habilitation* (licence to teach) in 1880, and was immediately afterwards appointed to a professorship in economics at the University of Innsbruck, which he held until 1889. From a scholarly point of view, Böhm-Bawerk's years in Innsbruck were the most fruitful of his life. A book on the theory of goods, based on his *Habilitation* thesis, appeared in 1881, the first volume of *Kapital und Kapitalzins* in 1884. In 1886 he published a monograph on the theory of value in the most influential German language journal in economics, and in 1889 the second volume of *Kapital und Kapitalzins*. These publications established him as one of the leading members of the group of economists around Carl Menger who came to be known as the ‘Austrian School’. In 1889 Böhm-Bawerk preferred an appointment in the Austrian ministry of finance to a chair at the University of Vienna because it carried the assignment to work out a reform of the Austrian income tax. He distinguished himself in the execution of this task, and rapidly rose in rank, obtaining the position of a permanent secretary in 1891, and in 1892 also the vice-presidency of a commission to assess the proposal of a return to the gold standard. Having been appointed minister of finance in a caretaker government in 1893, Böhm-Bawerk was considered to have risen too high to return to his former position when it was replaced by a parliamentary post after a few months, and he was made president of one of the three senates of the Verwaltungsgerichtshof, the highest court of appeal in administrative matters. In 1896 he was again made minister of finance in a caretaker government, but returned once more to the Verwaltungsgerichtshof in 1897. He was yet again appointed minister of finance in 1900, this time in a civil servants’ government which fell when he resigned in 1904 after large increases in military expenditure had been voted which he deemed threatened financial stability. This time he was offered, among other positions, the post of governor of the central bank, the most lucrative position in the monarchy. Yet he turned it down in favour of a chair at the University of Vienna which was especially created for him. Alongside Friedrich von Wieser (who had succeeded Menger in 1902) and Eugen von Philippovich, Böhm-Bawerk lectured on economic theory and conducted a seminar that soon attracted many able students, among them Joseph Schumpeter, Rudolf Hilferding, Otto Bauer, Ludwig von Mises, Emil Lederer and Richard von Strigl. He did not, however, return to the quiet life of a scholar. Having been elected a member of the Austrian Academy of Sciences in 1902, he was elected its vice president in 1907, and its president in 1911. He had also been made a Geheimrat (privy councillor) in 1895, had been appointed to a seat in the upper house of the Austrian parliament in 1899, and was from time to time given various other official assignments. Böhm-Bawerk died on 27 August 1914 at Rattenberg-Kramsach in Tyrol where he had tried to restore his health after having fallen ill on his way to a congress of the Carnegie Foundation in Switzerland as the official Austrian representative.

Böhm-Bawerk was as much a civil servant as a scholar, and in his later years an elder statesman in academic affairs as much as in the public realm of what was still a great power. He was extremely successful as an administrator and economic policymaker. But it is for his contributions to economic theory that he is chiefly remembered today. *Kapital und Kapitalzins* has become an economic classic even though it is defective in both construction and exposition. The first edition was written in great haste, and although Böhm-Bawerk responded over-conscientiously and meticulously to almost every criticism in the two further editions which appeared in his lifetime, adding so much material that two

slim volumes grew into three massive tomes, he never found the time to rethink the structure as a whole. This absorptive attention to criticism was due to temperament as well as to circumstances. Böhm-Bawerk had a lawyer's mind and found it difficult to think in terms other than disjunct categories or ‘cases’ which needed to be distinguished sharply and did not fit into a continuum in which things shade into one another. Moreover, writing in a thoroughly anti-theoretical environment dominated by the German Historical School, he felt obliged to take issue and to sharpen differences for the sake of discussion. As a result, Böhm-Bawerk acquired an undeserved reputation as a casuistic and ungenerous controversialist which did much to place his (admittedly in some respects imperfect) contributions in a more critical light than they merit.

The core of Böhm-Bawerk's theoretical endeavours is the development of an intertemporal theory of value, capital and interest. This attempt owes much to his teachers in economics. A.E.F. Schäffle, Menger's predecessor in Vienna, seems to have convinced him that it was necessary to respond on a theoretical plane to the social question, the most pressing economic policy problem of the day, by developing a satisfactory theory of distribution (see Schäffle, 1870). Karl Knies (1873–79) drew his attention to the problems of capital theory and the work of Marx. Carl Menger, finally, provided the starting point for his own theory.

In his *Grundsätze der Volkswirthschaftslehre* (1871), Menger had developed an atemporal theory of value, allocation and exchange. In his exposition and elaboration of that theory, Böhm-Bawerk (1886) strongly emphasized two of its aspects. Firstly, consumer behaviour is sharply distinguished from producer behaviour because only the former can evaluate goods directly; producers can do so only indirectly on the basis of their expectations of consumers’ evaluations because production, being roundabout production, is necessarily time-consuming. Secondly, in both cases the evaluation of a commodity involves both the marginal utility of the commodity to the evaluating agent, and the marginal utility of the income available to him. In Böhm-Bawerk's usage, therefore, evaluations are shadow prices, or inverse demand schedules which imply an optimal allocation of commodities in the light of an agent's preferences as well as his income.

On the basis of such inverse demand schedules it was easy to show that the market price of a commodity could not be lower than the lowest price the ‘last’ buyer is prepared to offer, nor higher than the highest price the ‘last’ seller demands; here the ‘last’ seller is defined as the seller whose asking price is low enough to prevent any other seller from selling to the ‘last’ buyer: and the ‘last’ buyer as that buyer whose price offer is high enough to prevent any other buyer from buying from the ‘last’ seller. This definition, complicated as it is, is adapted to include the case of indivisible commodities which Böhm-Bawerk for one reason or another considered relevant.

Böhm-Bawerk also elaborated on Menger's seminal contribution by refining the analysis of distribution: he showed how inputs are evaluated by imputation, that is, by imputing to them their proper share of the value of the output they help to produce. In essence this amounted to a marginal productivity theory along lines laid down by J.H. von Thünen, but again adapted to his peculiarly Austrian assumptions of limited substitutability and finite divisibility of inputs.

Böhm-Bawerk generalized (in 1889) this theory of price formation in atemporal exchange to include intertemporal exchange by assuming that agents evaluate and trade not only currently available commodities, but also subjectively certain prospects of commodities available in the future. In his theory of goods, Böhm-Bawerk (1881) had shown in a surprisingly modern manner that such prospects exist, and how they can be evaluated. Assuming further that a market exists on which currently available

commodities can be exchanged for subjectively certain prospects of commodities available in the future, the same argument can be applied to intertemporal exchange as was applied to atemporal exchange. Böhm-Bawerk did so in two stages, first considering a pure exchange economy without production, and then analysing an exchange economy with production.

In a pure exchange economy, all agents are consumers. Their inverse demand schedules, Böhm-Bawerk argued, involve for each agent a subjective rate of interest at which he is prepared, given his preferences over time and his (expected) income over time, to exchange subjectively certain prospects of commodities available in the future for the same amount of commodities available in the present. They also, Böhm-Bawerk maintained, typically exhibit positive time preference: commodities available in the present are typically evaluated at higher prices than subjectively certain prospects of the same commodities available in the future. This assertion is contained in the first two of three reasons he adduced for the positivity of the rate of interest. The first reason postulates that the marginal utility of income will decline over the planning horizon because of higher expected incomes in the future. The second reason postulates that for psychological reasons such as the finiteness of life, the marginal utility of a commodity declines as a rule with the length of time that elapses before it becomes available. As both these postulates have been much disputed it should be added immediately that Böhm-Bawerk regarded them as no more than testable assumptions which he deemed realistic but which admit exceptions. If these postulates are granted for all agents, their subjective rates of interest will always be positive, so that the market rate of interest will always be positive. The same will hold true if only the majority of agents behave according to these postulates. Böhm-Bawerk admitted that not all agents will always behave as postulated by him: but argued that as an empirical regularity they almost always did, and that his theory was applicable also when they did not. All that follows in the latter case is that the rate of interest is not positive. Note, therefore, that Böhm-Bawerk's argument establishes at one and the same time the existence of a (market) rate of interest in a pure intertemporal exchange economy, and identifies as the determinants of its height the relative intensities of the demand for, and supply of, commodities in the present and in the future, as expressed in agents inverse demand schedules. Of course, these are commodity rates of interest which do not necessarily exhibit any particular term structure, nor uniformity across different types of commodities. Both these properties need the further assumption that intertemporal markets exist for all commodities, and that at least some agents are prepared to engage in arbitrage operations (see Nuti, 1974), Böhm-Bawerk did not explicitly make these assumptions, but he argued as if these properties were assured. Note also that Böhm-Bawerk conceived in this model of a pure exchange economy of the rate of interest as a property of an intertemporal price structure, and not as the specific price for something, be it abstinence, the productivity of money, waiting, or whatever.

In order to extend the model just considered to include production Böhm-Bawerk argued that producers can be shown to have intertemporal inverse demand schedules like consumers, and postulated in his third reason that producers under-evaluate commodities available in the future on technical grounds. These assertions he derived from his analysis of the nature of production, and the role of capital in it. Production is assumed to be roundabout. It transforms non-produced or ‘original’ factors of production into consumable output with the help of capital goods which are internal to the production process. Because some capital goods are durable, production takes time. Böhm-Bawerk emphasized strongly the heterogeneity and specificity of capital goods. He also denied that they can be aggregated into some physical measure for the capital stock; aggregation is in his view possible only by valuing capital goods.

He employed a forward-looking measure of capital value in which durable capital goods are valued by the present value of their services, and indeed generalized this procedure to all durable goods by showing that their valuation involves a subjective rate of interest which is equalized when durable goods are traded on markets.

The view of production as roundabout led Böhm-Bawerk to postulate a correspondence between the amounts of different capital goods used in production and the time which elapses before a particular dose of non-produced inputs has matured in the form of consumable output. This correspondence he formalized in the concept of a period of production which is defined as the average period for which the various doses of non-produced inputs required for the production of a unit output remain ‘locked up’ in the production process. This definition was a mistake which got him into more than one difficulty, and provided material for heated debates. To get round all the difficulties raised in these debates, assume that it is possible to define a period of production as a technical property of a particular production system which does not depend on factor prices; and assume further (with Böhm-Bawerk) that it can be used to order different methods of production in such a way that methods with a longer period of production can be said to be more capital intensive. More specifically, assume a temporal production function which (for a unit output) has only the period of production as argument, and which exhibits diminishing returns but is not homogeneous.

On this basis Böhm-Bawerk formulated a theory of producer behaviour in which competition forces producers to choose production methods that generate just enough output to pay the costs of production. As Böhm-Bawerk showed, this implied a discounted marginal productivity doctrine of (original) factor pricing, and hence the existence of positive quasi-rents at the margin. He also showed that this construction involved inverse demand schedules for capital goods which for each period of production define a profit maximizing rate of interest for given factor prices. At this point in his analysis, Böhm-Bawerk assumed the capital stock of an economy as given, and argued that the profit maximizing rate of profit can be determined with the help of that assumption. While that is correct it was another mistake which was duly seized upon (see for example Garegnani, 1960) and which led to many debates. For the value of the capital stock associated with any method of production is an endogenous variable in his construction, as Böhm-Bawerk realized in other contexts. Nor was it necessary to make this assumption. It is sufficient to note that a single producer is forced by competition to pay neither less nor more than the discounted marginal value for the inputs he uses, if a time-consuming roundabout method of production is in operation. Translated into output prices this implies that he under-evaluates output available in the future. This is what Böhm-Bawerk asserted in the third reason; the technical ground being the method of production in operation. Note that this is not so much a postulate or empirical regularity as it is an equilibrium condition.

Having thus established that producer behaviour can be characterized by derived inverse demand schedules for output which involve positive time preference, Böhm-Bawerk goes on to determine the market rate of interest in what is in effect a macroeconomic general equilibrium model. Attention is centred on the market for output available in the present, and the markets for claims to output available in the future. Supply on the market for output available in the present is fixed by decisions taken in the past; so is the supply available at all future dates whose production has already begun. Demand for output available in the present comes from consumers but will not exhaust supply if they save. Part of these savings will be taken up by other consumers in exchange for claims to output available in the future; transactions are consumption loans, and are likely, on Böhm-Bawerk's assumptions, to imply a

positive rate of interest. Another part of savings will be taken up by producers, again in exchange for claims of future output, who use it to bid for more non-produced inputs in an attempt to expand the scale of production. As Böhm-Bawerk assumed that the amount of non-produced original factors is fixed, this results in higher factors prices and a change in the method of production (because higher factor prices can only be sustained if more output is produced). Net savings in the form of loans for productive purposes therefore imply a change in the method of production which, on Böhm-Bawerk's assumptions, implies capital deepening. Both kinds of transactions together determine the market rate of interest, which is thus seen to be determined by intertemporal consumer behaviour as summarized in the notion of positive time preference, and based on intertemporal preferences and the (expected) intertemporal distribution of incomes, on the one hand; and intertemporal producer behaviour as summarized in the period of production and the marginal product of extending it, and based on the intertemporal structure of roundabout methods of production on the other hand. Or, as Böhm-Bawerk put it, the rate of interest is determined by the relative evaluation of (output available in) the present and the future on the part of both consumers and producers. On his assumptions, this rate of interest is positive.

In some passages Böhm-Bawerk suggested that the rate of interest determined in his model is equal to the marginal product of an extension of the period of production. That created the impression that he had done no more than to establish, in a more roundabout way, what Jevons (1871, ch. 7) had already demonstrated. In other passages, however, Böhm-Bawerk seems to be aware that a change in the method of production involves a change in the value of the capital goods it requires, and that these Wicksell (or revaluation) effects imply that the rate of interest is less than the marginal product of an extension of the period of production. Böhm-Bawerk also obscured his argument by introducing the concept of a subsistence fund, thereby suggesting that his theory was no more than a revamped wages fund theory. Neither these nor other infelicities in his exposition should obscure the fact, however, that the hard core of his argument is the determination of the rate of interest as the property of an intertemporal price structure which in turn is determined by an intertemporal theory of value and allocation in consumption and production.

Böhm-Bawerk's model consciously referred to a stationary state as he wished to show that the rate of interest has something to do with the efficient allocation of resources in stationary as well as in non-stationary states. This comes out most clearly when he considers a socialist economy and demonstrates that it would require a positive rate of interest as does a capitalist economy. He did, however, consider non-stationary states in an interesting comparative static analysis of the effects of an increase in savings, and of technical progress. That he obtained a positive rate of interest in a stationary state is of course due to his assumptions, and no contradiction to Schumpeter's argument (1912) which is based on a somewhat different model (see Böhm-Bawerk, 1913, for a discussion of these differences).

The argument sketched on the preceding pages is expounded in Böhm-Bawerk's *Positive Theory* (1889) which he prefaced by a ‘History and Critique of Interest Theories’ (1884) in which he critically examined earlier (and in later editions also contemporary) attempts to explain the rate of interest. The purpose of this volume has often been misunderstood. It is not a history of the subject which generously corrects mistakes, nor an attempt to differentiate his own product. Rather it is a ‘negative theory’ (Edgeworth): an attempt to survey the building blocks for his own theory and to pinpoint the pitfalls a satisfactory theory should avoid. Yet it cannot be denied that it is often overcritical. Thus Böhm-Bawerk shows again and again that the rate of interest cannot be said to be *determined by* marginal productivity considerations, but does not add that these nevertheless have a role to play in a

more complete explanation. A similar omission occurs when he discusses abstinence or more generally intertemporal preferences.

One of the conclusions Böhm-Bawerk drew from his demonstration is that the existence of the rate of interest is not due to exploitation. It is obvious that on his argument workers can get the whole product of labour only if production is instantaneous. As long as production is roundabout, the present value of the workers' share in the value of the output they have helped to produce is necessarily less than what it would be if production were instantaneous. This is due, of course, to the existence of capital; but Böhm-Bawerk argued that interest would have to be paid irrespective of who owns such capital goods. That was also the gist of his critique of Marx's economics (1896), in which he singled out the labour theory of value as the basis of all errors. Böhm-Bawerk was (apart from Schäffle and Knies) one of the first economists to discuss Marx's economics on a scholarly plane; but he remained curiously blind to Marx's critique of the social institutions of a capitalist society. Although his critique drew a long reply from one of his students (Hilferding, 1904) it was very influential and remained the best analytical performance of its kind until well into the 1950s (see Sweezy, 1949).

Böhm-Bawerk's single-minded concentration on economic phenomena is also evident in his discussion of the role of economic power on markets (1914): in the short run, he argued, economic power may cause deviations from the state of affairs as defined by economic forces; in the long run, however, the latter will prevail. Again he was blind to any changes economic power may cause to the environment in which economic forces operate.

The impact of Böhm-Bawerk's work was immense, but its reception was made difficult by its prolixity and its technical defects, which offered many openings to critics. In essence, Böhm-Bawerk combined elements of neoclassical economic theory with elements of classical economic theory. He was neoclassical in his concern with rational economic behaviour and its consequences for the demand and supply of commodities, their pricing on markets, the forces which bring about equilibrium on markets, and the interaction of different markets. By contrast, classical lines of thought predominate in Böhm-Bawerk's analysis of production. However much he denied any adherence to classical cost theories of value, his view of production and the role of capital and time in it bear the mark of the Ricardian tradition.

The neoclassical part of his argument, in particular his analysis of intertemporal consumer behaviour, was taken up by Irving Fisher (1907; 1930) and developed into a theory of interest which is based on the notion of time preference (which Fisher transformed into a property of utility functions) and the concept of investment opportunities; these Fisher assumed rather than derived, thus cutting away Böhm-Bawerk's analysis of production and the role of capital in it. In this form, which admittedly offers insights into the problem of intertemporal allocation Böhm-Bawerk did not offer, Böhm-Bawerk's intertemporal theory of exchange became part of the heritage of orthodox neoclassical economic theory.

The more classical part of Böhm-Bawerk's model was taken up and elaborated by Wicksell (1893; 1901). In an attempt to free it of its classical garb, Wicksell turned it into a marginal productivity theory of the rate of interest. He ran into difficulties, however, not only over the proper definition of the period of production, but also because his neglect of what Böhm-Bawerk had to say about intertemporal consumer behaviour forced him to assume a given capital stock in order to close his model. Wicksell used what had by then become the standard neoclassical concept of capital as a value sum, as proposed by J.B. Clark (1899), and (with good reason) combatted by Böhm-Bawerk. The shortcomings of such an

argument, which was before long imputed to Böhm-Bawerk himself, were soon pointed out (see Cassel, 1903, and Garegnani, 1960). Nevertheless Wicksell's interpretation became the standard portrayal of the ‘Austrian’ theory of capital and interest (see for example Lutz, 1956; Dorfman, 1959a; 1959b; Hirshleifer, 1967).

In the 1930s various attempts were made to reformulate Böhm-Bawerk's theory in such a way that it could be used as the basis of a theory of the short-run behaviour of an economy, particularly by Hayek (1931; 1939; and see Hicks, 1967), but also by Hicks (1939, parts III and IV). This led to an intensive debate in which especially the capital theoretic foundations of his argument were examined, and found wanting (see Kaldor, 1937, and Reetz, 1971, for a survey). There were some attempts at reconstruction (Eucken, 1934; and Strigl, 1934), but the definition of the period of production provided a major stumbling block. At the same time, Hayek and Knight repeated the debate between Böhm-Bawerk and Clark about the concept of capital on a somewhat different level. Finally Hayek (1941) made a major attempt to get round the difficulties the debate had shown up, and achieved some advances: but in the end his contribution turned out to be the final word that did not persuade anybody. The major difficulty which he did not manage to overcome was the fact that Böhm-Bawerk's construction does not lend itself to dynamic analysis precisely because his classical, macroeconomic approach to production and the role of capital requires an equilibrium approach, and does not provide a suitable basis for a discussion of producer behaviour out of equilibrium, and its dynamics.

More recent restatements of Böhm-Bawerk's argument consequently emphasize its static nature (Weizsäcker, 1971; Faber, 1979), but do not really go beyond an exact formulation, in terms of modern capital theory, of some aspects of his theory. By contrast, Hicks (1973) is an innovative attempt to salvage some of the salient features of Böhm-Bawerk's view of production and capital, especially his emphasis on the role of time in production processes, in a modern framework which once more attempts to formulate a dynamic analysis (see also Belloc, 1980; or Magnan de Bornier, 1980). It centres on the concept of a ‘transition’ from one steady state to another, that is, a more long-term kind of economic dynamics than was considered in the 1930s; this is a promising approach which proves the vitality of Böhm-Bawerk's ideas.

Böhm-Bawerk posed a problem which had not been seen before in its full importance: the role of the rate of interest in the choice of an optimal method of production when production is roundabout, and its determination in a theory which takes seriously the impossibility of aggregating capital goods in physical terms. The solution he proposed is not without problems. But however much economic theory has progressed, some parts of his argument stand out as landmarks in the development of economic thought. Among them are his discussion of price formation on markets, especially those on which indivisible or finitely divisible commodities are traded, his analysis of time preferences, his analysis of intertemporal exchange, and his demonstration that the rate of interest is no more than a property of intertemporal price structures. His definition of the period of production turned out to be a cul-de-sac, but the possibilities his analysis of the role of time in production offers do not yet seem to have been exhausted.

Finally, the importance of his emphasis on the value aspect of the notion of aggregate capital and its implications has only recently been recognized as a seminal contribution. He can perhaps no longer be accorded the stature of a Ricardo or Marx. But the vitality of his ideas still ranks him among the great economists.

See Also

- Austrian economics
- period of production

Selected works

1881. *Rechte und Verhältnisse vom Standpunkte der volkswirtschaftlichen Güterlehre*. Innsbruck: Wagner. Trans. as ‘Whether Legal Rights and Relationships are Economic Goods’ in Böhm-Bawerk (1962).

1884. *Kapital und Kapitalzins. Erste Abteilung: Geschichte und Kritik der Kapitalzins-Theorien*. Innsbruck: Wagner. 2nd edn, 1900; 3rd edn, 1914; 4th edn, Jena: Fischer, 1921. Translation of 1st edn as *Capital and Interest*, London: Macmillan, 1890. Translation of 4th edn as *Capital and Interest*, vol. 1. South Holland, IL: Libertarian Press, 1959.

1886. Grundzüge der Theorie des wirthschaftlichen Güterwerthes. *Jahrbücher für Nationalökonomie und Statistik* 13, 1–82 and 477–541. Reprinted separately, London: London School of Economics, 1932.

1889. *Kapital und Kapitalzins. Zweite Abteilung: Positive Theorie des Kapitales*. Innsbruck: Wagner. 2nd edn, 1902; 3rd edn in two volumes, 1909 and 1912; 4th edn in two volumes, 1921, Jena: Fischer. Translation of 1st edn as *The Positive Theory of Capital*, London: Macmillan, 1891. Translation of 4th edn as *Capital and Interest*, vols. 2 and 3, South Holland, IL: Libertarian Press.

1896. Zum Abschluss des Marxschen Systems. In *Staatswissenschaftliche Arbeiten, Festgaben für Karl Knies*, ed. O. von Boenigk. Berlin: Haering. Trans. as *Karl Marx and the Close of his System*, London: Fisher Unwin, 1898. Reprinted in Sweezy (1949). Also trans. as ‘Unresolved Contradictions in the Marxian Economic System’ in Böhm-Bawerk (1962).

1913. Eine ‘dynamische’ Theorie des Kapitalzinses. *Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung* 22, 520–85 and 640–57.

1914. Macht oder ökonomisches Gesetz? *Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung* 23, 205–271.

1924. *Gesammelte Schriften*, ed. F.X. Weisz. Vienna and Leipzig: Hölder-Pichler-Tempsky.

1926. *Kleinere Abhandlungen über Kaptial und Zins*. Vienna and Leipzig: Hölder-Pichler-Tempsky.

1962. *Shorter Classics*. South Holland, IL: Libertarian Press.

Bibliography

Belloc, B. 1980. *Croissance économique et adaption du capital productif*. Paris: Economica.

Cassel, G. 1903. *The Nature and Necessity of Interest*. London: Macmillan.

Clark, J.B. 1899. *The Distribution of Wealth*. New York: Macmillan.

Dorfman, R. 1959a. A graphical exposition of Böhm-Bawerk's interest theory. *Review of Economic Studies* 26, 153–8.

Dorfman, R. 1959b. Waiting and the period of production. *Quarterly Journal of Economics* 73, 351–72.

Eucken, W. 1934. *Kapitaltheoretische Untersuchungen*. 2nd edn. Tübingen: Mohr.

Faber, M. 1979. *Introduction to Modern Austrian Capital Theory*. Berlin: Springer.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Garegnani, P. 1960. *Il capitale nelle teorie della distribuzione*. Milan: Giuffrè.

Hayek, F.A. von. 1931. *Preise und Produktion*. Vienna: Springer. Trans. as *Prices and Production*, London: Routledge.

Hayek, F.A. von. 1939. *Profits, Interest and Investment*. London: Routledge.

Hayek, F.A. von. 1941. *The Pure Theory of Capital*. London: Routledge.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hicks, J.R. 1967. The Hayek story. In *Critical Essays in Monetary Theory*, ed. J.R. Hicks. Oxford: Clarendon Press.

Hicks, J.R. 1973. *Capital and Time*. Oxford: Clarendon Press.

Hilferding, R. 1904. Böhm-Bawerks Marx-Kritik. In *Marx Studien* 1, ed. M. Adler and R. Hilferding. Trans. as ‘Böhm-Bawerk's Criticism of Marx’ in Sweezy (1949).

Hirshleifer, J. 1967. A note on the Böhm-Bawerk/Wicksell theory of interest. *Review of Economic Studies* 34, 191–9.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan.

Kaldor, N. 1937. Annual survey of economic theory: the recent controversies on the theory of capital. *Econometrica* 5, 201–33.

Knies, K. 1873–9. *Geld und Credit*. 3 vols. Berlin: Weidemann'sche Buchhandlung.

Kuenne, R.E. 1971. *Eugen von Böhm-Bawerk*. New York and London: Columbia University Press.

Lutz, F.A. 1956. *Zinstheorie*. Tübingen: Mohr. Trans. as *The Theory of Interest*, Dordrecht: Reidel, 1967.

Magnan de Bornier, J. 1980. *Economie de la traverse*. Paris: Economica.

Menger, C. 1871. *Grundsätze der Volkswirthschaftslehre*. Vienna: Braumüller. Trans. as *Principles of Economics*, Glencoe, IL: Free Press, 1951.

Nuti, D.M. 1974. On the rates of return of investment. *Kyklos* 27, 345–69.

Reetz, N. 1971. *Produktionsfunktion und Produktionsperiode*. Göttingen: Schwartz.

Schäffle, A.E.F. 1870. *Kapitalismus und Sozialismus*. Tübingen: Laupp.

Schumpeter, J.A. 1912. *Theorie der wirtschaftlichen Entwicklung*. Leipzig: Duncker & Humblot. Trans. as *The Theory of Economic Development*, Cambridge, MA: Harvard University Press, 1934.

Schumpeter, J.A. 1914. Das wissenschaftliche Lebenswerk Eugen von Böhm-Bawerks. *Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung* 23, 454–528 Trans. as ch. 6 in *Ten Great Economists*, ed. J.A. Schumpeter, London: Allen & Unwin, 1952.

Schumpeter, J.A. 1925. Eugen von Böhm-Bawerk. *Neue österreichische Biographie 1815–1918*, 2. Vienna: Amalthea.

Stigler, G.J. 1941. *Production and Distribution Theories*. New York: Macmillan.

Strigl, R. von. 1934. *Kapital und Produktion*. Vienna: Springer.

Sweezy, P.M., ed. 1949. *Karl Marx and the Close of his System by Eugen von Böhm-Bawerk and Böhm-Bawerk's Criticism of Marx*. New York: Kelley.

Weizsäcker, C.C. von. 1971. *Steady State Capital Theory*. Berlin: Springer.

Wicksell, K. 1893. *Über Wert Kapital und Rente nach neueren nationalökonomischen Theorien*. Jena: Fischer. Trans. as *Value, Capital and Rent*, London: Allen & Unwin, 1954.

Wicksell, K. 1901. *Föreläsningar i Nationalekonomi*. Vol. I: *Teoretisk Nationalekonomi*. Lund: Berlinska Boktryckeriet. Trans. as *Lectures on Political Economy*. Vol. 1, *General Theory*, London: Routledge, 1934.

Wicksell, K. 1911. Böhm-Bawerks kapitalteori och kritiken därav. *Ekonomisk Tidskrift* 13, 39–49 Trans. in K. Wicksell, *Selected Papers on Economic Theory*, London: Allen & Unwin, 1958.

Wicksell, K. 1914. Lexis och Böhm-Bawerk. *Ekonomisk Tidskrift* 16, 294–300 and 322–34.

Bibliographic addendum

Hennings, K., *The Austrian Theory of Value and Capital: Studies in the Life and Work of Eugen von Böhm-Bawerk*, Cheltenham: Edward Elgar, 1997, is a detailed account.

How to cite this article

Hennings, K. H. "Böhm-Bawerk, Eugen von (1851–1914)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000157>
doi:10.1057/9780230226203.0142

The New Palgrave Dictionary of Economics Online

Boisguilbert, Pierre le Pesant, Sieur de (1645 – 1714)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Boisguilbert, P.; Cantillon, R.; circular flow; Colbert, J.-B.; equilibrium trade models; hoarding; Physiocracy; proportional income tax; tax farming; underconsumptionism

Article

French economist and lawyer. Born at Rouen into a *noblesse de robe* family, Boisguilbert was educated at a Jesuit college in Rouen, the city where he spent most of his life and where he died in 1714. The famous Port Royal and the Paris law school trained him as an *avocat* but initially inspired a literary career. This produced translations from the Greek (Dion Cassius and Herodotus) and some historical novels, one of which, *Marie Stuart, Reyne d'Ecosse* (1675) went through three editions. Marriage to a rich heiress in 1677 allowed him to pursue profitable activities in trade and agriculture for several years and enter the magistrature of Normandy. Such experiences brought home to him the deteriorating French economic position and the need to reverse this through fiscal and economic reform. His first economic work, *Le détail de la France* (1695) reflects these concerns. For the remainder of his life he unsuccessfully pressed plans for fiscal reform on various finance ministers, ultimately republishing his ideas, including the new *Factum de la France*, in various collected editions from 1707 (a detailed biography and bibliography is in Boisguilbert, 1966).
Boisguilbert is largely remembered as a precursor of the Physiocrats and as the economist whom Marx (1859, p. 52) linked with Petty as marking the start of classical political economy. His influence was undoubtedly more extensive: much of Cantillon's (1755) circular flow analysis appears inspired by his work; while Roberts (1935, pp. 273–320) argues for considerable similarity between his fundamental economic ideas and some of Adam Smith's. A wealth of embryonic tools and concepts can be found in his work and include:

> division of labour, circular flow, velocity of money, hoarding, confidence, the multiplier, and variability of employment, supply and demand, diminishing utility, elasticity of

> demand, natural and market price, price variability, price flexibility, cobweb price-model, cost of production, diminishing returns, labour supply curve, bargaining range, impulse propagation, economic equilibrium, optimum and suboptimum price structures, and competition. (Spengler, 1984, p. 77)

Tax criteria and class analysis need to be added to this list.
Boisguilbert's economic analysis ascribes France's economic distress to agricultural ruin from Colbert's edict prohibiting corn exports; excessive taxation worsened by tax farming; and financiers' power transforming money from a servant of trade into its tyrant. Underlying this diagnosis are models of equilibrium trade demonstrating the interdependence of the 200 occupations and professions exchanging their products at prices proportioned to necessary costs of production including a just profit. Hence buying, as the essential counterpart of selling and consumption, stimulates production. Disruptions to consumption prevent prices from covering costs, thereby initiating a downward spiral which ends in economic stagnation. Three causes for such disruptions are identified: low agricultural prices which lower rent and hence landlords' consumption demand; second, concentration of money among rich financiers leading to hoarding; third, lower consumption potential from excessive taxation. Since the livelihood of the poor depends on the consumption of the rich, unemployment and misery follow. Boisguilbert's remedy follows from his identification of these causes of underconsumption. Free trade and encouragement of agriculture lead to a 'proper' corn price, conducive to high rents and consumption spending. Tax reform achieved by introducing a general proportional income tax removes the problem of excessive taxation and eliminates hoarding and leakages from the circular flow because the abolition of tax farming ends concentrated financier power. Subsequent encouragement of consumption allows prosperity to return and creates wealth for both the state and its citizens. Basic model, diagnosis and remedy are present with varying degrees of sophistication in Boisguilbert's major works, including *Traité de la nature, culture, commerce et intérêts des grains* (1704a) and *Dissertation de la nature des richesses, de l'argent et des tributs* (1704b), to name those not so far mentioned.

Selected works

1695. *Le détail de la France*. Reprinted in Boisguilbert (1966), 581–662.

1704a. *Traité de la nature, culture, commerce et intérêt des grains, tant par rapport au public, qu'à toutes les conditions d'un état*. Reprinted in Boisguilbert (1966), 827–78.

1704b. *Dissertation de la nature des richesses, de l'argent et des tributs, ou l'on découvre la fausse idée qui règne dans le monde è l'égard de ces trois articles*. Reprinted in Boisguilbert (1966), 973–1012.

1707. *Factum de la France*. Reprinted in Boisguilbert (1966), 879–956.

1966. *Pierre de Boisguilbert ou la Naissance de l'économie politique*. Paris: Institut National d'Etudes Demographiques.

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. London: Fletcher Gyles. Reprinted with an English translation and other material by Henry Higgs, London, for the Royal Economic Society, 1931.

Marx, K. 1859. *Contribution to the Critique of Political Economy*. London: Lawrence & Wishart, 1971.

Roberts, H. Van Dyke. 1935. *Boisguilbert, Economist of the Reign of Louis XIV*. New York: Columbia University Press.

Spengler, J.J. 1984. Boisguilbert's economic views vis-à-vis those of contemporary *réformateurs. History of Political Economy* 16, 69–88.

How to cite this article

Groenewegen, Peter. "Boisguilbert, Pierre le Pesant, Sieur de (1645–1714)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000158> doi:10.1057/9780230226203.0143

The New Palgrave Dictionary of Economics Online

Bondareva, Olga (1937 – 1991)

Myrna Wooders
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balancedness; approximate balancedness; Bondareva, O.; clubs; coalitions; cooperative games; core equivalence; cores; duality; essential superadditivity; game theory; linear programming; local public goods; market games; price-taking equilibrium

Article

Olga Nikolajevna Bondareva was born in St Petersburg on 27 April 1937. She joined the Mathematical Faculty of the Leningrad State University in 1954, and completed a Ph.D. in mathematics at the Leningrad State University in 1963, in part under the supervision of Nicolaj Vorobiev. Her thesis was entitled ‘The Theory of the Core in an n-Person Game’. Bondareva rose through the ranks at Leningrad State University to become a senior research fellow in 1972 and a leading research fellow in 1989. Because she sympathized with a student who wished to emigrate to Israel, however, she was prohibited from teaching from 1973 until 1989. With *perestroika* and increased freedom to travel outside the Soviet Union, she became an active and energetic international figure in game theory. She died as a result of a traffic accident on 9 December 1991.

Bondareva published over 70 works on game theory and mathematics, supervised seven Ph.D. students, and was a member of the editorial board of *Games and Economic Behavior*. Her work on the core of a cooperative game plays a central role in game theory, and her insights can be seen underlying recent work on the theory of price-taking equilibrium and the core.

The following is a brief description of Bondareva's celebrated result. To allow us to see the relationship of this result to more recent research on games and economies with many players, it is stated for games with player types and requiring only ‘essential superadditivity’ in the definition of feasible payoffs. Define a (pre)game with *T* types of players as a function ψ from vectors of non-negative integers $s \in \mathbb{Z}_+^T, \ s \neq 0$, called *profiles* of coalitions, into the non-negative real numbers $\mathbb{R}_+$. Given a vector $m \in \mathbb{Z}_+^T$, representing the total player set of the game and $s \in \mathbb{Z}_+^T, \ s \leq m, \ \psi(s)$ is interpreted as the

total payoff to a coalition of players consisting of s_t identical players of type t, $t = 1, \ldots, T$. Let $(s^\ell : \ell = 1, \ldots, L)$ denote the collection of all profiles $s^\ell \leq m$. A *partition of a profile s* is determined by a collection of non-negative integers $(n_1, \ldots, n_L)$ satisfying the condition that $\Sigma n_\ell s^\ell = s$. With the domain of ψ restricted to profiles $s^\ell \leq m$, the pair (m, ψ) determines a cooperative game. Let $\psi^*(m)$ denote the maximum, over all partitions of *m*, of $\Sigma n_\ell \psi(s^\ell)$. A payoff vector $\bar{x} \in \mathbb{R}^T$ is in the (equal treatment) *core* if and only it holds that. $\bar{x} \cdot m \leq \psi^*(m)$ ($\bar{x}$ is *feasible*) and for each ℓ, $\psi(s^\ell) \leq \bar{x} \cdot s^\ell$. Now consider the following linear programming (LP) problem:

$$\min_{\bar{x}} \bar{x} \cdot m \text{ subject to } \psi(s^\ell) \leq \bar{x} \cdot s^\ell \text{ for all profiles } s^\ell \leq m.$$

A vector $\bar{x}^*$ is in the core if it is a solution to the above LP problem *and* $\bar{x}^* \cdot m = \psi^*(m)$. The dual LP problem is:

$$\max_{\omega_1, \ldots, \omega_L} \sum_\ell \omega_\ell \psi(s^\ell) \text{ subject to } \sum_\ell \omega_\ell s^\ell = m \text{ and } \omega_\ell \geq 0 \text{for all } \ell$$

From the fundamental duality theorem of linear programming, there is a solution to the first LP problem if and only if there is a solution to the second, and, in this case, it holds that the optimal values of the objective functions in the two LP problems are the same.

For the second LP problem, let (ω_ℓ^*) denote the solution for the ‘balancing weights’ (ω_ℓ). The game is *balanced* if and only if $\Sigma_\ell \omega_\ell^* \psi(s^\ell) = \psi^*(m)$. It follows that a game is balanced if and only if it has a non-empty core, Bondareva's result. (See also Shapley, 1967.)

Numerous applications of game theory to economics have employed the concept of balancedness. An outstanding contribution is Shapley and Shubik (1969), who show an equivalence between the set of totally balanced games (balanced games with the property that every subgame also has a non-empty core) and market games (cooperative games derived from economies where all players have concave utility functions). Bondareva's result as formulated above is a key ingredient in Wooders (1994), showing that under mild conditions games with many players are market games. Scarf (1967) demonstrates non-emptiness of the core of a balanced game without side payments (where the payoff set for a coalition *S* is a subset of R^S rather than a real number). Bondareva's result also underlies the approximate balancedness of economies with clubs or relatively small effective or nearly effective coalitions. While this result has been demonstrated in much generality, the key is simple. Since the coefficients of the dual LP problem are integers, when the total player set is replicated (becomes *rm*, $r = 1, 2, \ldots$) and no new effective coalitions are permitted (that is, if $\psi(s) \geq 0$ then $s \leq m$) then there is

an integer *k* such that all replicated games with total player profiles given by *rkm* are balanced (Wooders, 1994, and references therein). The integer *k* clears the denominators of the (rational) extreme points of the convex set of balancing weight vectors of the dual LP problem. In recent works on the theory of clubs and local public goods, balancedness plays a crucial role; see Demange and Wooders (2005) for several recent examples and additional references.

We refer the reader to Rosenmueller (1992) for some additional details of Olga Bondareva's life. See also Kannai (1992) for an excellent review of research on the core and balancedness.

See Also

- game theory

Selected works

1963. Some applications of linear programming to the theory of cooperative games. *Problemy Kibernetiki* 10, 119–39 [in Russian]. English translation in *Selected Russian Papers in Game Theory 1959–1965*. Princeton: Princeton University Press, 1968.

1969. Solution for a class of games with empty core. *Soviet Doklady* 185(2).

1975. Acyclic games. *Vestnik of Leningrad University* No. 7.

1978. Convergence of spaces with a relation and game-theoretical consequences. *Zhurnal Vych. Math. i Math, Phys*. 18(1).

1979. Development of game theoretical methods of optimization in cooperative games and their applications to multi criterial problems. In *Sovremennoe sostojanie terorii issledovaniji operatiskil* [State of the art in the theory of operations research], ed. N. Moiseev. Moscow: Nauka.

1983. Extensive coverings and some necessary conditions of existence of solutions in cooperative games. *Vestnik of Leningrad University* No. 19.

1987. Finite approximations of choice on infinite sets. *Izvestija AN SSSR, Tekhnicheskaja kibernetika* No. 1.

1989. Domination, core and solution (a short survey of Russian results). Discussion Paper No. 185. IMW, University of Bielefeld.

1990. Game theoretical analysis of one product market with similar utilities. *Kibernetika* No. 5.

1990. Revealed fuzzy preferences. In *Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory*, ed. J. Kacprzyk and M. Fedrizzi. Dordrecht: Kluwer Academic Publishers.

1994. (With T. Driessen.) Extensive coverings and exact core bounds. *Games and Economic Behavior* 6, 212–19.

Bibliography

Demange, G. and Wooders, M., ed. 2005. *Group Formation in Economics; Networks, Clubs and Coalitions*. Cambridge: Cambridge University Press.

Kannai, Y. 1992. The core and balancedness. In *Handbook of Game Theory with Economic Applications*, ed. R. Aumann and S. Hart. Amsterdam: North–Holland.

Rosenmueller, J. 1992. Olga Nikolajevna Bondareva: 1937–1991. *International Journal of Game Theory* 20, 309–12.

Scarf, H. 1967. The core of an *n*-person game. *Econometrica* 35, 50–69.

Shapley, L. 1967. On balanced sets and cores. *Navel Research Logistics Quarterly* 9, 45–8.

Shapley, L. and Shubik, M. 1969. On market games. *Journal of Economic Theory* 1, 9–25.

Wooders, M. 1994. Equivalence of games and markets. *Econometrica* 62, 1141–60.

How to cite this article

Wooders, Myrna. "Bondareva, Olga (1937–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000323> doi:10.1057/9780230226203.0144

The New Palgrave Dictionary of Economics Online

bonds

Donald D. Hester

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A bond is commonly understood to be a debt instrument in which a borrower receives an advance of funds and contracts to make future payments of interest and principal according to an explicit schedule. The nominal return from holding a bond is the sum of its interest payments and the change in its price over an arbitrary holding period. Bonds differ in terms of face value, maturity, callability, seniority, convertibility, risk of default, and size, frequency and taxability of interest payments. Since 1970 bond markets have experienced a number of major institutional changes with enduring consequences for capital markets.

Keywords

asset-backed debt; bankruptcy; bonds; capital gains and losses; central banks; default; defeasance; derivative securities; discount bonds; Eurobonds; Interest rates; junk bonds; medium term notes; Miller, M.; Modigliani, F.; options; sovereign debt; stripped bonds; Tobin, J.

Article

A bond is a contract in which an issuer undertakes to make payments to an owner or beneficiary when certain events or dates specified in the contract occur. The term has medieval origins in a system where an individual was bound over to another or to land. Subsequently, goods were put in a bonded warehouse until certain conditions (for example, payments of taxes or tariffs) were satisfied; individuals were released from jail when a bail bond guaranteeing their appearance in court was supplied; and individuals were allowed to perform certain tasks when a surety or performance bond guaranteeing satisfaction was provided. Governments and individuals have borrowed from others since earliest recorded history, as Sumerian documents attest. Perhaps public bonds first appeared in modern form with the establishment of the Monte in Florence in 1345. Monte shares were interest bearing, negotiable and funded by the Commune.

In contemporary economic discourse, a bond is commonly understood to be a debt instrument in which a borrower, typically a government or corporation, receives an advance of funds and contracts to make future payments of interest and principal according to an explicit schedule. The remainder of this entry focuses almost exclusively on these debt instruments. Terms of bonds are designed to protect the rights of borrowers and creditors; they are heterogeneous and their interpretations and enforceability vary across legal jurisdictions.

Bond heterogeneity

The distinction between bonds and other evidences of debt such as loans or notes is inherently arbitrary and imprecise. Bonds tend to have long specified maturities when issued, or none at all in the case of consols. However, issuers may reserve the right to call them after they have been outstanding for a specified time interval. Other things being equal, bonds that are callable have higher rates of return than those with no call provision, because issuers have an incentive to call them whenever market rates fall below rates that existed when the bonds were offered. While bonds ordinarily convey no equity stake in an enterprise, some corporate bonds are convertible; they include a clause that gives bondholders an option to convert bonds to shares of the issuer's common stock at a specified conversion value in some time interval. Other things being equal, convertible bonds have lower interest rates than bonds with no conversion rights, because the option to convert is valuable. Formulas for determining the values of options are discussed by Black and Scholes (1973) and Zhang (1997).

Bonds tend to be negotiable and can usually be traded on an established secondary market. Once bonds are issued, bondholders are strategically vulnerable to actions of a firm's management, equity holders, and short-term lenders, as has been argued by Bulow and Shoven (1978), especially if an issuer's financial condition deteriorates. Default occurs if a bond issuer fails to make scheduled payments of interest or principal or violates other covenants of a contract. A bondholder's rights in a default situation are circumscribed by the terms of the contract and by judicial authority.

In the event of a default by a corporation, bondholders or other interested parties may petition for protection under bankruptcy statutes. In some circumstances a bankruptcy court appoints a receiver to conserve the value of a firm's assets so as to protect creditors. The fraction of a creditor's claims that is paid is determined in part by their seniority (or priority) relative to other claims. Bonds may be either unsubordinated or subordinated to other debt. A bankrupt firm may be liquidated in favour of its creditors or be reorganized and allowed to continue with partial payouts to creditors.

In the United States, bonds issued by corporations and state and local governments are assigned credit ratings by firms such as Moody's Investors Services and Standard and Poor, Inc. Bonds with lower credit ratings are predicted to have a higher rate of default; they tend to have higher *ex ante* rates of return to compensate holders for higher expected default losses and risk of default. Bonds of state and local governments fall into two broad classes: (*a*) bonds which are general obligations of the issuing government and (*b*) revenue bonds, where interest and principal payments are dependent on income from some specific project. Because general obligation bonds are funded from taxes of the issuer, they tend to have higher ratings and lower rates of interest than revenue bonds. Corporate bonds with poor credit ratings are called ‘junk bonds’. Before 1980, most bonds had been issued with good ratings and were suitable for the portfolio of a prudent investor. If an issuer's condition subsequently deteriorated, its bonds were downgraded and possibly became junk bonds. Beginning in about 1982, this practice

changed and large amounts of funds were raised by issuing bonds that had low ratings when first offered. The reasons for offering junk bonds are incompletely understood but include avoidance of corporate income taxes, as was predicted by Modigliani and Miller (1963). Coinciding with the issuance of junk bonds were a substantial increase in leverage (the ratio of a firm's debt to net worth) and a wave of leveraged buyouts in which publicly traded corporations were reorganized into enterprises that were narrowly held by management and a few outside investors.

The significance of these changes in imperfect capital markets is controversial; in traditional financial theory it is often argued that high leverage makes a firm vulnerable to financial shocks and recessions. High leverage is believed to reduce the probability of a firm being taken over or bought up. Leverage on the books of a firm, however, can be misleading without knowledge of the contractual rate of interest on a firm's bonds. For example, when interest rates rise a firm may call its existing low-interest rate bonds which have a low market price and finance them with a smaller quantity of new bonds that bear the new high rates. This action, ‘defeasance’, reduces the ratio of debt to equity on a firm's books without reducing its interest costs.

Bonds issued by autonomous nation states are ‘sovereign’ debt. Defaults by issuers of sovereign debt do not result in bankruptcy proceedings, because there is no world bankruptcy court and applicable code. Moreover, as Bulow and Rogoff (1988) have argued, there is no credible basis for establishing seniority among sovereign debt issues in the event of a default. Sovereign bonds that default are traded at deep discounts for indefinitely long periods. While bankruptcy is impossible, negotiations leading to the restructuring of a country's debt obligations do occur, and sanctions against a defaulting country have been imposed by other countries where bondholders are concentrated. Credit ratings of sovereign debt vary widely across countries and, in part, are a function of the bond repayment history of a country.

Bond yields and rates of return

The ‘yield’ on a bond is the flow of interest income to its holders. Apart from defaults, bonds traditionally pay interest in fixed amounts on specified dates that are indicated by coupons on the bond. Coupon-bearing bonds may allow investors to choose portfolios that match interest and amortization streams with their own nominal future requirements for funds. A portfolio is said to be perfectly ‘immunized’ against interest rate fluctuations if such matching is achieved. Bonds that have no coupons are called ‘discount bonds’; they provide no interim cash flow and are retired at maturity with a payment equal to their face or par value, which is higher than the issue price. Default-free discount bonds thus afford nominal *income certainty* to investors, as was explained by Robinson (1951), but do not guarantee that an investor's spending goals can be achieved when inflation is unpredictable. Some protection against inflation is afforded by inflation-indexed bonds that first appeared in Israel in 1955, the United Kingdom in 1981 and the United States in 1997, when US Treasury Inflation-Protected Securities (TIPS) were first offered. With TIPS, protection takes the form of a percentage increase of the bond's principal that equals the rate of inflation. Because the increase is taxable and inflation is based on the rate of change of the consumer price index, TIPS only incompletely protect a representative investor against inflation. For a discussion, see Wrase (1997).

The nominal return from holding a bond is the sum of its interest payments and the change in its price over an arbitrary holding period. For example, if there are no transactions costs and taxes, the return

from holding a multi-year bond for two years is:

$$\text{return} = y_1 + y_2 - P_p + P_s \quad (1)$$

where P_p and P_s are respectively the purchase and selling price and y_1 and y_2 are annual interest payments. If interest payments are assumed to be paid at year end, the nominal annual rate of return, r, from this two-year investment is obtained by solving the polynomial:

$$P_p = y_1 / (1 + r) + (y_2 + P_s) / (1 + r)^2 \quad (2)$$

If the bond is bought at P_p and sold at P_s, a bond trader is said to ‘realize’ a capital gain (loss) if P_p is less (more) than P_s.

A condition for equilibrium in a bond market is that expected rates of return from holding similar bonds are similar. If this condition were not satisfied, bond traders could improve portfolio earnings through arbitrage, by selling the bond with the lower rate of return and buying the bond with the higher rate of return, so long as the difference exceeds transactions costs. When transactions costs are zero, bonds are perfectly ‘reversible’. When market rates of return rise, prices on outstanding bonds fall and rates of return experienced by existing bondholders fall; capital losses are sustained by holders of all but maturing bonds. Bond traders attempt to buy bonds immediately before market rates of return fall so that they may realize capital gains by buying at a low price and selling at a high price. Similarly, speculative traders of bonds seek to sell bonds immediately before market rates of return rise. While bonds that do not default mature at par, the prices of outstanding bonds are incompletely predictable; generally bonds with more years to maturity have more price volatility.

Bond issuance considerations

Bonds are issued by governments and corporations to finance deficits and acquire assets. While neither issuer can afford to ignore imminent movements in interest rates, their time schedules of outlays are somewhat inflexible. Deficits must be financed, and it is short-sighted to delay purchasing high rate-of-return assets to take advantage of transient interest rate movements. Firms needing funds may choose to finance a long-term asset with short-term borrowings from banks, with a long-term bond whose interest rate varies (or ‘floats’) over time in a fixed relation to short-term rates, or with a long-term fixed coupon bond. Bank borrowing to finance long-term assets exposes firms to the risk that banks may unilaterally alter loan terms or refuse to renew maturing loans. Firms avoid non-renewal risk by borrowing with

bonds. A firm's choice between issuing conventional fixed-rate bonds and floating rate bonds to finance an asset depends in part on the correlation between returns from the asset being acquired and short-term interest rates for reasons that are developed by Cox, Ingersoll and Ross (1981). Other things being equal, a floating rate bond exposes a firm to less risk when the short-term rate and the rate of return on the acquired asset are positively correlated.

Government deficits are financed by issuing short-term bills, notes, bonds and ‘outside’ or fiat money. Central banks control the ratio of outside money to interest-bearing government debt when conducting monetary policy. Central bank sales (purchases) of bonds decrease (increase) bond prices and increase (decrease) bond interest rates in the market. Other things being equal, an increase in bond interest rates increases the cost of financing new capital equipment and causes marginal investment projects to become unprofitable. Control of bond and other market interest rates by central banks is one handle through which monetary policy affects the level of macroeconomic activity. It has also been argued by Tobin (1963) that the composition of outstanding interest-bearing government debt can importantly influence the level of macroeconomic activity. If bonds are closer substitutes for physical capital in investors’ portfolios than are treasury bills, a debt management policy of selling bonds and buying an equivalent amount of bills discourages private sector capital formation.

Recent innovations in bond markets

Since 1970 capital markets have experienced a number of major institutional changes and innovations that have had enduring consequences for bond markets. Arguably the most important was the introduction of securitized debt by the US government-sponsored enterprises, Federal National Mortgage Association and Federal Home Loan Mortgage Corporation, and by the Government National Mortgage Association. While they could issue conventional bonds, they also could issue what amounts to second-order bonds, such as pass-through securities or collateralized mortgage obligations. Instead of an issuer being responsible for paying interest and retiring principal, securitized debt replaces the issuer with a constructed package of mortgage loans that generates a stream of interest and principal payments to holders of the securities. Initially, the underlying loans were insured against default, but they differed from traditional bonds because mortgage loans could be paid off before their contractual maturity. Thus, these securities were bonds with discrete, stochastic call provisions. The underlying stochastic process is in part a function of past and current market interest rates, because homeowners tend to refinance their houses when market interest rates fall.

In 1985, securitized debt evolved into generalized asset-backed debt, which serves to finance a package of self-liquidating financial assets. Like bonds, some of this debt is publicly rated for safety by investment services, but much of it is privately placed and not traded on a secondary market where ratings are important. The value of the assets underlying a debt issue typically exceeds the face value of the issue by an amount called a ‘haircut’, which serves as a partial safeguard against default. Asset-backed debt is heterogeneous; interest rates may be fixed or indexed to some market rate, amortization schedules vary, and the qualities of underlying assets differ. In 2004, new issues of asset-backed debt exceeded new issues of conventional bonds by corporations and governments for the first time. Asset-backed debt tends to be less costly to issue and to service, which largely accounts for its rapid growth. It is often issued by a ‘special purpose vehicle’, a legal entity which is intended to be bankruptcy-remote and whose sole function is to service a set of debt issues. Unlike conventional bonds, such debt usually

does not appear on government or corporate balance sheets, which partly explains its appeal in a world where leverage has been rising. However, especially in Europe there is a hybrid ‘covered bond’, which is a securitized bond that remains an obligation of the issuer and continues on balance sheets. Because it is collateralized, it retains value even when the issuer fails.

Another innovation that has partly displaced bonds are medium term notes (MTNs), which US corporations began to issue in the early 1970s. In recent years outstanding corporate MTNs have averaged about 14 per cent of corporate bonds. They tend to be issued by highly rated corporations and are distinctive in being issued through ‘shelf registrations’ rather than having a formal offering with the assistance of underwriters. In a shelf registration an issuer presents a menu of securities that it may choose to issue in a specified period, which allows it to have a closer correspondence between the time funds are needed and the time when securities are issued. MTNs range in maturity from nine months to 30 years.

A large off-shore ‘Eurobond’ market exists where governments and corporations issue bonds denominated in currencies that differ from the currency in the country where the security is issued. While recent data are unavailable, there was also a rapidly growing outstanding stock of EuroMTNs in the early 1990s. These large and expanding markets complicate the implementation of monetary policy in a country, because information about Euromarkets must be taken into account. International financial statistics often do not reveal the nationality of individuals issuing or holding securities in different countries.

The establishment of financial instrument futures markets in 1975 also modified the demand for bonds in investor portfolios. Short-term hedging and speculative positions are more inexpensively achieved in a futures market than they are by constructing forward cash flows through the assumption of long and/or short positions in a bond market.

A market for ‘stripped’ bonds, where all a bond's coupons are separated from the body of a bond and each coupon and the body (or principal) are traded as separate entities, emerged in 1982. The body of the bond and each coupon are traded as discount bonds. The market for stripped securities greatly expanded in February 1985 when the US Treasury adopted this private sector innovation by offering its own stripped securities in book entry form and was willing to reconstruct stripped securities beginning in May 1987. These innovations increased the attractiveness of Treasury securities and arguably lowered the cost of government borrowing. The innovation is important because discount bonds are especially convenient for matching expected cash flows from other assets and liabilities and thus hedging against fluctuations in interest rates. Because discount bonds make no interest payments they are sometimes called ‘zeros’ in the financial press.

During the 1980s, a new technique emerged that broke the linkage between the choice of fixed or floating interest rates paid by a bond issuer and the form in which interest is received by a bondholder. A simple (plain vanilla) bond ‘swap’ is a transaction in which the holder of a bond trades a fixed interest-rate stream for a floating interest-rate stream. Thus, a borrower can issue a fixed-rate bond to an investor who prefers floating-rate securities, because the latter can simultaneously execute a swap with a third party. Such transactions facilitate marketing of securities in imperfectly competitive markets. Swaps also allow investors to change the currency unit in which an interest stream is denominated from, for example, euros to US dollars. They can also be used to change the base of a floating interest rate bond from, say, the US Treasury bill rate to dollar-denominated Libor, the London interbank offer rate. Swaps and put and call options are early forms of ‘derivative’ securities, which allow investors to create

synthetic bonds that effectively increase the stock of conventional corporate bonds, as can be inferred from Stoll (1969). A derivative security's value is conditional on the price or price trajectory over time of another asset. In recent decades an enormous variety of ‘structured’ assets has been and continues to be created by combining derivatives and conventional assets such as bonds and MTNs. For a discussion see Zhang (1997).

Finally, automation in bond markets has reduced the costs of trading bonds and made them more convenient to hold. Most government bonds in the United States are no longer issued in certificate form; they are issued in book form and exist only as computer entries. They are readily transferable in a computer and can be lent or sold at low cost whenever a borrower requires cash. By making bonds more reversible, automation has reduced the distinction between bonds and outside money, a distinction that is crucial for the success of central-bank open market operations.

See Also

- fiat money
- Miller, Merton
- Modigliani, Franco
- monetary economics, history of
- residential real estate and finance
- sovereign debt
- third world debt
- Tobin, James

Bibliography

Becketti, S. 1988. The role of stripped securities in portfolio management. *Federal Reserve Bank of Kansas City Economic Review* 73(May), 20–31.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Bulow, J. and Rogoff, K. 1988. The buyback boondoggle. *Brookings Papers on Economic Activity* (2), 675–704.

Bulow, J. and Shoven, J. 1978. The bankruptcy decision. *Bell Journal of Economics* 9, 437–56.

Cox, J., Ingersoll, Jr., J. and Ross, S. 1981. The relation between forward prices and futures prices. *Journal of Financial Economics* 9, 321–46.

Crabbe, L. 1993. Anatomy of the medium-term note market. *Federal Reserve Bulletin* 79, 751–68.

Modigliani, F. and Miller, M. 1963. Corporate income taxes and the cost of capital: a correction. *American Economic Review* 53, 433–43.

Robinson, J. 1951. The rate of interest. *Econometrica* 19, 92–111.

Smith, W. 1970. The role of government-sponsored intermediaries. In *Housing and Monetary Policy*. Conference Series No. 4. Boston: Federal Reserve Bank of Boston, 86–101.

Stoll, H. 1969. The relationship between put and call option prices. *Journal of Finance* 24, 801–24.

Tobin, J. 1963. An essay on the principles of debt management. In *Fiscal and Debt Management Policies*. Prepared for the Commission on Money and Credit. Englewood Cliffs, NJ: Prentice-Hall, 143–218.

Wrase, J. 1997. Inflation-indexed bonds: how do they work? *Federal Reserve Bank of Philadelphia Business Review* (July/August) 3–16.

Zhang, P. 1997. *Exotic Options: A Guide to Second Generation Options*. Singapore: World Scientific Publishing Co.

How to cite this article

Hester, Donald D. "bonds." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000165> doi:10.1057/9780230226203.0145

The New Palgrave Dictionary of Economics Online

books, economics of

Frederick van der Ploeg, Marcel Canoy and Jan van Ours

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The tensions between books as expressions of culture and books as profitable products are analysed using insights from the theory of industrial organization. To stimulate the diversity of books on offer, maintain the density of bookshops and to promote reading, governments grant fixed price monopolies, subsidize authors, levy a lower consumption tax on books, and provide public libraries and education. Market structures and government policies vary widely and there is no case for harmonizing European book policies. The book market is innovative in solving its problems. The main task of the government is to promote reading.

Keywords

books, economics of; business stealing; cross-subsidy; cultural policy; experience goods; market failure; monopolistic competition; non-price competition; payola; product differentiation; product life cycle; retail price maintenance

Article

The market for books is characterized by the laws of demand and supply. However, the availability of a diverse supply of quality books is also an objective of cultural policy. This, combined with market failures, may provide grounds for government intervention as discussed for the arts in general in van der Ploeg (2006). Here we focus mainly on the market for general books, paying special attention to cultural books, leaving aside educational and scientific books. Governments influence book markets through subsidies for libraries, authors and publishers, tax concessions on the sale of books, and laws concerning the pricing of books. Apart from stimulating reading, it is not clear what role there is for government intervention. After all, the book market invents solutions to specific problems (contracts for authors, literary agents, gatekeeping by publishers, joint distribution by wholesalers cooperating on distribution, agreements concerning stocks between retailers and publishers, joint publicity, best-seller lists, reviews,

and so on). The book market flourishes in production of book titles, but not in reading.

Stylized facts

About half of Portuguese adults never read a book. This is in sharp contrast with the 20 per cent of readers in Belgium, Denmark, Italy and Norway who similarly do not read books. Reading is popular in Finland, Sweden and Switzerland where about 90 per cent of adults read. Nevertheless, even in Sweden almost 30 per cent failed to read a book during 2003. Although in most countries a majority of adults read, there are large numbers of people who never read a book.

At the low end of the distribution of book titles across countries is the United States, with 24 titles per 100,000 inhabitants, and only six of which concern arts and culture. At the high end, Denmark produces 275 titles per 100,000 inhabitants, of which 80 are devoted to the arts and culture. Most titles per inhabitant are produced in Scandinavia, in Switzerland, and in the United Kingdom. Relatively few titles are produced in Italy, Japan, Greece and Australia.

The typical average annual number of books sold per inhabitant is about five to six. The exceptions at the lower end are Portugal and Sweden with 2.6 and 3.6 books per inhabitant, while at the high end France has 6.9. Publishers' revenues from sales vary from 20 euros per inhabitant in Greece to 115 euros in Finland. In most countries the revenue from book selling is 40–60 euros per inhabitant. The largest industries are located in the United States, Germany, the United Kingdom, France and Italy. In 2001, total value added of the book publishing industry was about 0.11 per cent of GDP with some 140,000 employees in the EU-15. The industry is stable in terms of turnover and per capita sales.

The number of books available through public libraries is low in Greece, Italy, Portugal and Spain, but much larger in Denmark, Finland and Sweden. The number of loans per inhabitant correlates nicely with the number of books available. It ranges from less than one in Greece, Portugal, Spain and Switzerland to at least ten in Denmark, Finland and the Netherlands. Differences in book-reading frequency are large. Reading a book daily varies from about a quarter of all adult males in Australia, Canada, Ireland, Sweden, Switzerland, the United Kingdom and the United States to a mere five per cent for Portuguese male adults. In most countries 10–20 per cent of adult males read daily. Females read much more than males, less so in Belgium (Flanders) and Portugal and more so in Australia, Canada, Denmark and the Netherlands.

The level of education is an important determinant of reading habits but no systematic cross-country evidence is available. However, in France 62, 78 and 92 per cent of lower-, medium-, and higher-educated individuals, respectively, read at last one book during the year 2003. There is not much cross-country information concerning trends in reading. However, in the Netherlands there is a clear downward trend in book-reading. Furthermore, fewer people indicate that they read books, though the average time spent reading has hardly changed. All readers irrespective of gender or country spend on average 6.5–8 hours per week reading books. In Europe, people spend most of their leisure time watching television. In the United States trends suggest that Internet use is increasing, mainly at the expense of watching television rather than reading.

Finland, Denmark, Ireland and Switzerland produce more than 200, the United Kingdom almost 200, Spain about 150 and the United States circa 25 book titles per 100,000 inhabitants per year. Although the number of titles produced has increased steadily in most countries, the number of publishers is stable. The average size of a publishing enterprise in the EU is small. Most enterprises publish only

between 20 and 40 titles per year. The percentage of books published on arts and literature vary from 20 to 50 per cent across countries. Differences in the number of titles published may be related to economic prosperity, to the educational level of the population, or to population density. The empirical evidence for this is mixed. For example, a rich country like the United States publishes fewer titles per capita than some poorer southern European countries.

Total European book sales amounted to 27 billion euros in 2000. The biggest market is Germany with some 9.5 billion euros. Both Germany and the United Kingdom are strong exporters of books to countries that share their languages. Other large book markets are found in France, Spain and Italy. During the first two years of the 21st century, the United Kingdom book publishing industry has grown to be the largest in Europe. In contrast, there has been a decline in Germany. About half the revenues of publishers in most countries come from general books. Most sales are through retail channels (trade), except in the United States. In some countries there are strong retailers, but in others there are many independent bookshops. In France, the multimedia retailer Fnac accounts for around 15 per cent of sales. In Italy Feltrinelli commands 25 per cent of the retail market. However, in Germany, the largest bookseller, Thalia, has only three per cent of the market and there are many small independent bookshops. The largest retailers in the United Kingdom in 1998 were Waterstones and W.H. Smith with 20 per cent and 18 per cent of the market respectively. The United States book industry has limited opportunities for growth in a mature market, and competition is focused on growth through market shares. The United States has seen consolidation among retail chains. Barnes and Noble command 30 per cent of the market and independent booksellers struggle.

The share of book clubs is high in Australia (26 per cent), about 15–20 per cent in Denmark, Finland, France, and Sweden and low in Italy, the United Kingdom and the United States. Although Internet sales have grown in importance, they are still small. In the United Kingdom around 17 per cent of book sales go through Internet retailers, a percentage that is no longer thought to be growing very fast. For Germany estimates suggest between four and five per cent of sales are made through Internet retailers, although recent growth has been much faster than in the United Kingdom. Some reports have estimated Internet sales in France and Italy at 1–1.5 per cent. Spain has even lower Internet sales than France. Internet is mainly used as a channel for books and so far not for digital products. For example, E-books are not sold much in the European market. In the United States E-books are more important; over 7,000 titles were published in 2003 while over 1.3 million E-books were sold. Concentration of firms in the worldwide online book market is high, with 60 per cent for Amazon.com.

The book market functions well

According to Caves (2000) cultural goods are characterized by *nobody knows* (uncertain demand), *time flies* (short period of profitability), *infinite variety* (horizontal differentiation) and *A-list and B-list* (vertical differentiation). Beck (2003) adds spontaneous purchases of books, non-convexities in production with large fixed costs and small marginal costs, and free entry for the book trade. A book is a private good, since its consumption is rival and excludable. This suggests there is no fundamental market failure. Books can be borrowed by other people. However, if this yields utility to the owners, there is no market failure. The market for books has a traditional supply chain: production, wholesale, distribution and retail. In each part of the chain there is competition between private entrepreneurs. Government provision occurs only with libraries, but that does not exclude competition between private

firms in the rest of the chain. There is substantial product differentiation in each part of the chain, which generates niche markets. Branding is important. Making a new product successful often requires substantial investment and innovation. This includes accepting that some products will never make it. Most parts of the supply chain have a fairly large number of players. Consumers of books can easily switch from one product to the other. The book market knows relatively few consumer lock-ins, which helps the market to function properly. Transparency adds to that effect. Even though books are experience goods, author reputation, book reviews, book clubs and word-of-mouth ensure transparency. The book market is also dynamic: there is innovation, market shares fluctuate and there is entry and exit. All this suggests that the book market should not be exempted from competition law.

Books occupy niches, more so than publishers

The book market is characterized by monopolistic competition along the lines of Dixit and Stiglitz (1977), since (*a*) products are differentiated; (*b*) firms set the price of the goods; (*c*) the number of sellers is large and each firm disregards the effects of its price decisions on the actions of its competitors; (*d*) entry is unrestricted. There thus exists a trade-off between efficiency (exploiting scale economies by producing more of the same product type) and diversity. Consumers love variety, but variety comes at a cost and the market becomes less transparent. Since firms do not take the potential downside of the variety decisions of other firms into account (the business stealing effect), there could be a market failure and optimal product diversity is not guaranteed. But in the book market consumers do not engage in repeated purchases in the same way as they do for, say, cereals. This greatly reduces possibilities for exploiting economies of scale, especially in the light of *nobody knows*. This does not mean that the book market can never have too much variety, but the argument then rests on lack of transparency and not on economies of scale. The book market does not have repeated entry by publishers with each publisher filling a niche. It is books that occupy niches, not publishers. Publishers have a portfolio of authors and books that serve as a way of risk-smoothing. Some books make it while others do not, but publishers have difficulties either of forecasting the success or are happy to accept differences in success out of cultural motives. Additional complexities arise for two other reasons. First, the book market is characterized by the fact that a single product (a book) has a very short life cycle. This leads to high initial prices followed by discounts. Second, publishers face a trade-off between risk-smoothing and specialization. A science fiction publisher has a competitive edge over non-specialized publishers, but faces the risk that its clients might switch to video games.
A publisher thus has a quickly changing portfolio of books. Its strategy consists of deciding on the portfolio (trading off risks and specialization) and on the prices of the portfolio. Multi-product firms in a monopolistic competitive market face the decision whether to engage in new product lines (exploiting economies of scope) or not (reducing cannibalization). This is akin to the decision by a publisher whether to employ a new author in the same field as his current portfolio. This trade-off, combined with variety in publishers' 'love for culture', leads to a mix of publisher types. There are specialized publishers, small publishers and large publishers. This has been the case for many years in many countries.

The book market plays into special features of books

Books have some special features. First, books are experience goods as one only appreciates the value after reading the book. Second, books are characterized by high fixed and low marginal costs. Third, some books are extremely successful, while most are unsuccessful. Success is hard to forecast and sometimes leads to ‘winner takes all’ economics as developed by Rosen (1981). Booksellers and publishers thus cross-subsidize higher-risk books with profits on other books. These potentially welfare-enhancing cross-subsidies can be thwarted by non-branch shops (for example, supermarkets) which sell only the bestsellers. Fourth, the opportunity costs of reading a book (that is, time) typically outweigh the price of a book. This contributes to a low price elasticity compared with other goods. The evidence suggests that the market for books other than best-sellers is price-inelastic, probably because most readers have high incomes or buy books for study purposes. Fifth, reading a book can be viewed as a private investment in culture rather than consumption. Sixth, there is an (almost) free substitute for buying books, namely, libraries. However, the quality of the service in bookshops and libraries is not the same, which makes substitutability imperfect. Seventh, books have cultural value. Books may also have option, existence and bequest value, and contribute to national identity, social cohesion, national prestige and the development of criticism and experiments. None of these values is (fully) reflected in the price, so the total value of books is higher than what has been paid.

Still, the market need not fail, since publishers, booksellers and authors find solutions to cope with these special features. The book market is relatively simple compared with other cultural markets (Caves, 2000). First, there is the *motley crew* property. A play or movie involves a complex set of different professionals to interact. The success of the play or movie crucially depends on how these different professionals get along. Many parts of the chain have the possibility to break it and kill the project. This leads to a complex set of contracts and other institutions, largely unnecessary and therefore absent in the book industry. Second, the *nobody knows* and *time flies* principles are even more applicable to a play or movie than to a book. Third, the production costs of a play or movie are much higher than those of a book.

Authors and publishers share the risk associated with the *nobody knows* and *time flies* principles. Authors get a percentage of the sales (typically ten per cent) and a split of the gross profits (typically 58–42) between author and publisher. Only celebrity authors receive bigger advances. While celebrity authors do reduce the risk of publishers somewhat, there are also serious large-scale flops. Some 70 per cent of former US President Bill Clinton's *Between Hope and History* were returned from bookstores as unsold (Caves, 2000).

Changing the terms of the contract either in favour of the author or the publisher can lead to misallocations. A higher fee for the publisher leads to a higher number of published books, since it becomes more lucrative to publish books and there still exists a reservoir of authors wanting to accept lower fees (Caves, 2000, p. 57). However, there will be less commercial success per book on average and lower quality as good authors may spend their time on more profitable activities. This could be justified if the perception is that there is a lack of supply of books. There is no evidence of that, however (the contrary is more likely). A higher percentage for the authors implies higher risk for the publisher, fewer books and fewer possibilities for new authors.

Incentives differ between publishers and authors. Publishers want to maximize profits, while many authors want to maximize sales and impact as they can often supplement their royalties with other income from lectures, TV, film, and so on. With globalization and the Internet some authors obtain

superstar incomes by using the media to leverage their incomes. Many new authors find their way into the book market. In addition, sales of a novel increase the probability of future sales, a factor that influences an author more than the publisher. Authors may thus want to use agents. There is no marketplace for the literary reputations of new authors. The chance that a publisher accepts a manuscript is extremely low; Caves (2000) mentions one in 15,000 for novels. Agents reduce the cost of publishers by filtering out good and bad manuscripts. The publisher can then use the reputation of a good agent as a proxy for quality.

Nobody knows and *time flies* create problems with stocks in retail outlets. If a book does not perform, the retailer wants to dump stocks as shelf space is scarce and new potentially successful books are looming. Market solutions to this problem include second-hand sales shops, sales of remainders, pricing strategies and policies that aim at sharing risks between publishers and retailers. Book retailers also have a right to return books for full credit. They can further reduce risks by smart wholesaling agreements. There are distinct differences in market shares of wholesale firms in Europe. In France, Finland, Denmark and the Netherlands the wholesale market is concentrated, but in Anglo-Saxon countries wholesale is less concentrated. If publishers are larger, it is worthwhile for them to vertically integrate into distribution. In sum, the market seems fairly able to solve the coordination problems needed to sort out the economies of scale.

There also exists a trade-off between exploiting economies of scale in retail and other policy goals. Examples are the reduction of transport costs for consumers or equity ‘universal service’ type of arguments. Various trends such as the Internet tilt towards scale. Books are easy to transport and personal contact with the seller is not always needed. In fact, interactive service and personal advice from Internet bookstores is often excellent.

Books are experience goods, so consumers have difficulty in deciding which book to buy. Book reviews in newspapers and the Internet, best-seller lists, book clubs, prizes and awards, and word of mouth facilitate choices. The market for information does not seem to fail except perhaps for payola (Caves, 2000). Payola is a system where the author (or his agent) ‘bribes’ a gatekeeper to influence his choices (as with pop music on radio). For example, an author may buy many copies of his own book in order to be high on the best-sellers lists, or chain bookstores may offer deals to book publishers to selectively display books in eye-catching positions. The problem is that payola threatens the objectivity of gatekeepers.

Does the book market achieve cultural goals such as (i) a diverse supply of cultural book titles and genres; (ii) access of books for all in term of price and distance by having sufficient density and variety of (high-quality) retailers? Since books are rival and excludable, the book market should require less government interference. With the Internet one may expect a demand-driven growth in the sale of selected parts of handbooks and guidebooks. Because books are reproductive cultural goods, large-scale distribution of books is easier than for non-reproductive forms of art. The market thus produces a large variety of books, with prices that are low enough (with libraries as a fallback as well) to make books available to everybody interested. If retailers are unsuccessful in dealing with stock risks, there may be too few cultural books, too little reading or too many authors.

Should the government tolerate retail price maintenance?

One reason to intervene is to protect a dense network of well-stocked, high-quality bookshops and stimulate the publication of a large variety of books. Indeed, the number of high-quality bookshops is decreasing in many countries. This happens if it does not pay to invest too much in variety in low-selling books. Monopoly profits and cross-subsidies from profitable to less profitable books may allow bookshops to store a greater variety of books and publishers to take more risks. The current practice in many European countries of a fixed book price (FBP) in combination with a variety of subsidies handed out by literary funds is often motivated by these considerations. Critics argue that a FBP or subsidies for high-brow books may harm reading on the part of the general public, since monopoly prices and cross-subsidies for less popular books are paid for by ordinary people reading popular books. Furthermore, subsidies for authors, translators, bookshops and publishers are paid for by ordinary people who may not be interested in more culturally valuable books or high-quality bookshops.

When considering policy instruments for reaching cultural objectives, there are at least two trade-offs. The first is between efficiency and density and distance. Increasing the scale of booksellers can enhance efficiency, but leads to longer travelling time for consumers. The second trade-off is between efficiency and cultural goals. Diversity of books in a bookstore may conflict with productive efficiency. The optimal choice of policy instruments depends on culture-political preferences and on country-specific characteristics that determine the market outcome. For example, a large ‘language size’ generates market outcomes where cultural objectives are more easily achieved. This is why the United States, Australia and Canada do not have policies aimed at the book market. Harmonizing book policies in Europe is not necessarily a good idea. Governments may wish to stimulate reading of worthwhile books, production of a diverse menu of titles and/or an extensive network of high-quality bookshops.

The FBP involves retail price maintenance by which the publisher reserves the right to set the retail prices of books. Since the publisher also influences wholesale prices, he effectively sets gross margins for retail outlets. The cultural merits ascribed to such agreements have reached almost mythical proportions in Europe. Since monopoly profits are higher than profits in competitive equilibrium, more titles are profitable and are published or sold under the FBP than in competitive equilibrium. It is possible to print and sell extra books at low and almost non-increasing marginal cost, so the producer loss is likely to be small. Also, the price elasticity of the demand for books is small as a large part of the full cost of reading is the opportunity cost of time and thus monopoly profits are large. The FBP leads to more variety in book titles, but prices will be higher and sales of each title lower as discussed in van der Ploeg (2004). The welfare costs may in practice be much larger, since much of the profit is dissipated by unproductive rent-seeking along the lines of Tullock (1980).

The FBP also has dynamic costs. Price competition between retail outlets becomes impossible but it is also more difficult to vary prices in response to local conditions. A store on a remote island may want to charge more for the same book than a store in the capital, but under the FBP it cannot do so. Also, it is more difficult to vary prices for different types of customers or for different seasons. Some customers need no service and low prices, while others prefer service at a higher price. Most important is that the FBP discourages the development of innovative distribution channels, since realized cost savings cannot be passed on to customers. With the FBP, unconventional distribution channels (bookclubs, supermarkets, petrol stations, the Internet, and so on) have less of a chance. Against these costs there is the benefit that independent small bookshops may be able to recommend interesting books and order books from the publisher or distributor.

Potential gains from retail price maintenance

Even though the FBP eliminates price competition, non-price competition may intensify. For example, a bigger sale margin stimulates booksellers to give better service to customers (Holahan, 1979; Mathewson and Winter, 1998; Deneckere, Marvel and Peck, 1997). With a bigger profit margin, it pays to spend more effort on service in order to get extra customers. If the extra service (more attractive presentation in bookshops, better information to customers, more promotion, and so on) generates more sales than the fallback in sales due to higher monopoly prices, the FBP may be desirable. Otherwise, the market fails to deliver sufficient service, because bookshops have an incentive to operate as free riders by offering discounts and expecting their customers to get their information and service elsewhere. Bookshops hardly refuse service or charge for information provided to people who in the end may not buy a book. Still, most customers rarely engage in such a strategy, as the costs of roaming around various bookshops seem high in relation to the possible discount one might obtain. Much of this service is already made available through publishers' advertisements or book reviews in newspapers and other media or on the Internet. In any case, it is questionable whether the demand for books really depends on service. Better service does not seem a good argument for supporting a FBP.

The book trade also argues that a bigger margin provides incentives for better-stocked bookshops. Booksellers may take over some of the inventory risks from publishers, so that more titles will be published. At the margin it is more profitable for retail outlets with relatively high costs to open up. This argument works only if customers want to purchase their books at particular high-cost bookshops. The gain in sales from these outlets may then offset the drop in sales resulting from higher monopoly prices. Although a dense network of bookshops may be desirable from a cultural point of view, this argument for the FBP is difficult to justify on grounds of market failure. Another popular argument is that higher margins encourage more retail outlets to put new book titles with uncertain sales prospects on their shelves. Given that there seems to be no problem for new authors to get their first book published, this is not a strong argument either. Marvel and McCafferty (1984) suggest that resale price maintenance may sustain a luxury image, but that seems more relevant for the markets for perfumes and jewellery than for books.

Is the cross-subsidy argument really valid?

The novel *Endurance* by Ian McEwan is not a perfect substitute for *Il Nome della Rose* by Umberto Eco. They are different books, because the authors have different styles, the themes of the two novels are different, and last but not least the original languages in which the books are written are different. Still, Umberto Eco's books are closer substitutes for the novels of Ian McEwan than, say, a cookbook or a travel book. On the other hand, Martin Amis may be a closer substitute than Umberto Eco for Ian McEwan. One must therefore leave the realms of homogenous goods and adopt a framework of Chamberlinian monopolistic competition in which books are imperfect substitutes. Publishers and booksellers carve out a niche and make monopoly profits, which enable them to recoup fixed costs. It is thus profitable to publish books. In fact, an important argument of the lobby of booksellers and publishers rests on imperfect competition. They argue that the FBP allows for cross-subsidies from best-sellers to less popular books and leads to a more diverse supply of book titles and bookshops. In

addition, the book lobby suggests that publishing and stocking a large selection of books enhances reputation, yields economies of scope and satisfies the idiosyncratic taste of individual publishers and booksellers even though these arguments do not seem very strong (Canoy, van Ours and van der Ploeg, 2006).

The cross-subsidy argument seems at first blush irrelevant. In competitive markets with imperfect information about the success of a product, it is common to invest in many products and reap a success on only a few. Even without a fixed horse price agreement, horse owners purchase lots of yearlings, many of which are subsequently sold to the riding school or the butcher if they do not win races. Similarly, in a market without FBP, publishers invest in new authors, just as horse owners invest in yearlings. Indeed, the industry's rule of thumb formulated by Denis Diderot in 1767 suggests that one out of ten new editions is a profitable success, four cover costs, and five make losses (Beck, 2003). There are few barriers to new authors in the book market even though publishing is a risky business with only one-third of published books being profitable. The FBP then has all the welfare and political economy costs of a monopoly. This situation may arise if best-sellers are easily digestible, require little time to read and have high price elasticities of demand, while, say, poetry readings demand a lot of time and effort and have low price elasticities of demand. Indeed, anything worthwhile from a cultural point of view takes time and effort to appreciate and contributes to a low price elasticity of demand. Non-fiction books (dictionaries, cookbooks, travel guides, textbooks, and so on) are likely to be close substitutes within each genre and will thus have high price elasticities. Fiction books (children books, mysteries, and so on) often have close substitutes (perhaps with the exception of *Harry Potter*), especially for the pocketbook versions of old titles, and thus high price elasticities. We do not expect large monopoly profits on such titles, and there is little room for cross-subsidies to books with a special or unique character. Such books have low price elasticities and generate high monopoly profits. If this is the situation, the cross-subsidy argument is likely to be wrong. The problem with a FBP is that there is no guarantee that publishers and booksellers will use the monopoly profits to make sure that more esoteric titles will be published and stocked in the stores. Monopoly profits may well be directed towards unproductive managerial slack.

Summing up

In summary, a FBP may induce higher prices and fewer sales of any book title that is published. It may also hinder innovation and distribution, but more titles will be published and there will be more bookshops with a diverse assortment of titles. However, German data suggest that retail price maintenance does not facilitate above-average focal pricing where prices are bunched around focal points (Beck, 2004). The lowering of production costs due to technological progress will benefit the diversity of books being published. In any case, many FBPs are of limited duration and characterized by sensible exceptions. The welfare costs are probably not very large, but may be reduced a little by reducing the term and coverage of the agreement. It may also be helpful to abolish certification and exclusive trade arrangements, scrap the fixed discount for recognized booksellers, and move to individual rather than vertical price agreements (see also Appelman and van den Broek, 2002). Since educational and scientific books typically have relatively low price elasticities and are more susceptible to monopoly abuse, it helps to exclude them from the FBP. As a dogma, the FBP diverts attention and

energy away from making the book trade more innovative and customer-oriented. It may be more worthwhile to stimulate reading of a wide variety of books by investing in public libraries and education, subsidising authors to write books of high cultural value, translating the best books into other languages and promoting them abroad.

Other public policies

Stimulating demand: lower value-added tax

The general consumption of books can be increased by lowering the specific value-added tax (VAT) rate on books. This is a general instrument, which is not well suited to direct at special books of literary value. The lower VAT on books applies to cookbooks as well as to poetry. This instrument is therefore mainly used to stimulate the purchasing and, it is hoped, reading of books. Administrative costs are low, since no apparatus of literary experts has to be called upon. All countries of Europe, except Denmark, use a reduced VAT rate as instrument to stimulate book purchases. The United Kingdom and Ireland even abolished VAT on books altogether. The European Commission misguidedly attempts to harmonize VAT rates on books, making it difficult for other member states to abolish VAT on books. The Commission fails to take account of the subsidiarity principle. Since the book trade, especially between the non-English speaking countries, hardly distorts the intra-European book trade, there is no danger of tax competition and no harm in countries pursuing their VAT policies on books independently of each other.

Stimulating supply: prizes and grants for writers and subsidies for bookshops

Governments and commercial sponsors do many things to encourage writers. There are many prestigious and less prestigious prizes for the best novelist, the best detective writer, the best poet, the best translator, and so on. All these are meant to encourage quality. More important, they might guide the uninitiated reader to better books. Book clubs, best-seller lists and book programmes on television also help in this respect. They also probably increase sales. Literary funds help struggling authors to make a living if their project is deemed to be of literary interest. Since only best-seller authors can make a living on royalties and related incomes, others may need some help, especially if their output has cultural value but is perhaps of less general interest. These policies are designed to stimulate quality rather than quantity. Sometimes subsidies for publishers of high-quality books may help as well (witness Sweden).

Many politicians attach cultural importance to a dense network of retail outlets. We have already noted that density seems to be falling in some countries, perhaps more in countries without a FBP; and concentration is increasing as well. From a cultural point of view this is bad news. Consumers have to travel further and there is less variety of bookshops. If the main objective of cultural policies is to increase the density of high-quality outlets, subsidies for high-quality bookshops may be more effective than the FBP. If they act as cultural centres in less populated areas, they may deserve public support. Subsidizing in order to maintain well-stocked bookshops would probably prove an administrative nightmare, which may explain why there is not much experience of this. Subsidizing publishers to publish books of literary and cultural value would also seem to hinder the market mechanism and lead to

adverse effects. In Sweden the government subsidizes in this manner roughly one-third of all fiction and one-fifth of books for children. However, Swedish retailers do not stock all titles since the government, rather surprisingly, does not require subsidized books to be offered for sale.

Concluding remarks

The book market ensures reasonable cultural performance with little government intervention, especially in large language areas. Yet there are differences between countries in reading, retail outlets, wholesale and production. Due to lack of data and research it is not easy to explain these differences. They may be due to differences in preferences, logistics, population density or public policies or to being stuck in the wrong equilibrium. One important trend is that people seem to read fewer books over time. Perhaps they are reading on the Internet or spending time on other cultural leisure activities. Here are some important areas for further research: investigating the relationship between production of titles, books sold and prices; using survey data to study the effects of personal characteristics of readers on market outcomes; analysing empirically differences between book and other cultural markets; and using industrial organization to understand pricing and stocking behaviour of publishers and retailers.

The book industry is characterized by relatively few market failures and these can be relatively easily corrected with market instruments. The book industry can fend well for itself, in contrast to opera, film or theatre, characterized by high production costs, high risk and complex interactions between a large number of different professionals. Even though there are obvious returns to scale, production costs are low. Thresholds for new authors, publishers and retailers are small, contracts are relatively simple and fairly uniform. The market is quite capable of inventing solutions for specific problems and public policies are not always called for, except perhaps to stimulate reading.

Nevertheless, there is a strong lobby for government intervention. Prizes and grants for authors, translators, publishers, bookshops, special VAT regimes for books, stimulating reading through public libraries, and the FBP are possible policy instruments. The standard case against the FBP is that book prices are higher and sales lower than under perfect competition. This hurts the interests of buyers, particularly those with lower incomes, since prices will be higher. One possible argument in favour is that the FBP may induce more and better-stocked bookshops and lead to publication of more marginal book titles. The cross-subsidy argument of the lobby in favour of the FBP is not convincing, however. First, even without the FBP, the market cross-subsidizes new authors and other risky projects in the hope of a possible best-seller. Second, even if this policy ‘works’, there is no accounting for what is done with the cross-subsidies and no democratic checks. Third, there is no guarantee that profits on best-sellers will be used to cross-subsidize less popular books. In fact, publishers and booksellers have an incentive not to do this. Fourth, if less popular books are less price elastic than popular books (perhaps as they take more time to read), monopoly profits on less popular books are higher and the cross-subsidy argument does not work. Fifth, even if cross-subsidization does occur, one should evaluate whether its cultural gains outweigh the distortionary costs of the FBP. Arguments put forward to defend the FBP, stressing improved service, better distribution and retail networks, and other forms of increased non-price competition, do not stand up to scrutiny either. The book industry produces many titles and new authors do not experience severe problems. The FBP may slow down or even stop the declining number of well-stocked bookshops outside big cities, but hinders sales through the Internet and supermarkets.

A comparison of policies towards the book industry in different European countries teaches us that

harmonization is a bad idea. There is not much inter-European book trade, so that book policies hardly distort the single European market. Also, characteristics of book industry, cultural and social features and political preferences of the different countries of Europe differ substantially. It is therefore best to allow member states of the European Union to design their own book policies. For example, a FBP makes more sense for Greece than for the United Kingdom as it has a smaller ‘language size’ and fewer people have access to the Internet. Although there may be a problem of a ‘race to the bottom’ if VAT rates are not harmonized, tax competition seems pretty irrelevant for the book market. European countries should be free to lower or abolish VAT on books in order to promote reading.

Many of the privileges granted in the book industry will eventually be undermined by technical changes. Digital cameras, recording and editing equipment have made low budget radio and television as well as narrowcasting possible, thus undermining the monopoly power of public and other broadcasters. Similarly, the Internet has stimulated virtual book suppliers, printing and publishing on demand and E-books. Virtual dictionaries, encyclopedias and other handbooks have already overtaken, to a large extent, their physical counterparts. A dense network of well-stocked bookshops remains important. Some argue that the emergence of the Internet and the integration of books in smart product and digitized communication will lead to the disappearance of the printed book (Choi, Stahl and Whinston, 1998). While more retailing will take place through the Internet and new gadgets, for some people the physical bookshop, where one can feel the book and bump into surprise titles and people, will remain indispensable.

There are, however, trends that endanger books, the most important being that people read less and less. Some worry that the next generation will stop reading books altogether, but this may be too pessimistic. First, the population is aging so that more leisure time becomes available and the opportunity costs of reading decrease. Second, books are doing well. In 1947, some 85,000 books were in print in the United States, against 1.3 million in 1996. This is, in part, due to sharp reductions in production and printing costs. Third, there is no reason to believe that a cultural carrier as old as the book will suddenly disappear. Modern technology complements books rather than substitutes for them (Cowen, 1998). Each new development in the craft has led to outbursts of cultural pessimism that allegedly indicates the end of the book. Most of the developments have only improved the book business (Cowen, 1998). Also, prices fell considerably and steadily. The future of the book market may look very different. E-books will replace parts of the market where E-reading already outperforms traditional reading. As for novels, nobody knows. Perhaps our children will read their novels directly from the screen.

See Also

- art, economics of
- consumer expenditure (new developments and the state of research)
- internet, economics of the
- markets
- product differentiation

The authors would like to acknowledge that much of this article is based on Canoy, van Ours and van

der Ploeg (2006), which also contains more details on the stylized facts and references.

Bibliography

Allen, W. and Curwen, P. 1991. *Competition and Choice in the Publishing Industry*. London: Institute of Economic Affairs.

Appelman, M.D. and van den Broek, A. 2002. *Boek en Markt. Effectiviteit en Efficientië van de Vaste Boekenprijs*. The Hague: SCP/CPB.

Appelman, M.D. and Canoy, M.F.M. 2002. Horses for courses: why Europe should not harmonise its book policies. *De Economist* 150, 583–600.

Barro, R.J. and Lee, J.-W. 2000. International data on educational attainment: updates and implications. CID Working Paper No. 42. Cambridge, MA: Harvard University.

Beck, J. 2003. *Monopoly vs. Oligopoly in the Debate on Retail Price Maintenance for Books: A Preliminary Empirical Result*. Berlin: Diplomarbeit, Humboldt University.

Beck, J. 2004. Fixed, focal or fair? Book prices under optional retail price maintenance, SP II 2004–15. Berlin: Wissenschaftszentrum.

Bertarelli, S. and Censolo, R. 2000. Preference for novelty and price behaviour. Working Paper, University of Bologna.

Bittlingmayer, G. 1992. The elasticity of the demand for books, resale price maintenance and the Lerner index. *Journal of Institutional and Theoretical Economics* 148, 588–606.

Canoy, M., van Ours, J. and van der Ploeg, F. 2006. The economics of books. In *The Handbook of the Economics of Art and Culture*, ed. D. Ginsburgh and D. Throsby. Amsterdam: North-Holland.

Caves, R. 2000. *Creative Industries, Contracts between Art and Commerce*. Cambridge, MA: Harvard University Press.

Choi, S.-Y.C., Stahl, D.O. and Whinston, A.B. 1998. Gutenberg and the digital revolution: will printed books disappear? *Journal of Internet Banking and Commerce*, January. Online. Available at http://www.arraydev.com/commerce/jibc/9801-4.htm, accessed 26 May 2007.

Clerides, S.K. 2002. Book value: intertemporal pricing and quality discrimination in the US market for books. *International Journal of Industrial Organization* 20, 1385–408.

Coser, L.A., Kadushin, C. and Powell, W.W. 1982. *Books: The Culture and Commerce of Publishing*. Chicago: University of Chicago Press.

Cowen, T. 1998. *In Praise of Commercial Culture*. Cambridge, MA: Harvard University Press.

Creemers, M. 1999. In de electronische boekhandel is informatie belangrijker dan prijs. *Holland Management Review* 66, 58–63.

Cummings, M.C., Jr. and Katz, R.S. 1987. *The Patron State*. New York: Oxford University Press.

Deneckere, R., Marvel, H.P. and Peck, J. 1997. Demand uncertainty and price maintenance: markdowns as destructive competition. *American Economic Review* 87, 619–41.

Dixit, A. and Stiglitz, J.E. 1997. Monopolistic Competition and optimal product diversity. *American Economic Review* 67, 297–308.

Ecalle, F. 1988. Une évolution de la loi relative du 10 août au prix de livre. *Economie et prévision* 5/88, Paris.

Economisch Instituut voor het Midden- en Kleinbedrijf. 2001. *De Vaste Boekenprijs: Een Internationale 'Quick Scan*. Zoetermeer: EIM.

Epstein, J. 2001. *Book Business: Publishing Past, Present and Future*. New York: W.W. Norton.

European Commission. 2002. Commission accepts undertaking in competition proceedings regarding German book price fixing, Press release IP/02/461. Brussels: European Commission.

European Commission. 2004. Book publishing. Publishing Market Watch. Sectoral Report 2. Brussels: European Commission.

Eurostat. 2001. *Eurostat Yearbook 2002*. Luxemburg: European Commission.

Eurostat. 2002. Statistics in Focus, theme 4–24. Luxemburg: European Commission.

Eurostat. 2004. *How Europeans Spend their Time—Everyday Life of Women and Men*. Luxemburg: Office for Official Publications of the European Communities.

Fishwick, F. 1989. Les implications économiques du Net Book Agreement. *Cahiers de l'Économie du Livre* 2, 4–31.

Fishwick, F. and Fitzsimons, S. 1998. *Report into the Effects of the Abandonment of the Net Book*

Agreement. Cranfield: Cranfield School of Management.

Foster, J.E. and Horowitz, A.W. 1996. Complimentarily yours: free examination copies and textbook prices. *International Journal of Industrial Organization* 14, 85–99.

Goolsbee, A. and Chevalier, J. 2002. Measuring prices and competition online: Amazon and Barnes and Noble. Working Paper No. 9085. Cambridge, MA: NBER.

Grauwe, P. de. and Gielens, G. 1993. De prijs van het boek en de leescultuur. CES Working Paper No 12, University of Leuven.

Greco, A.N. 1999. The impact of horizontal mergers and acquisitions on corporate concentration in the U.S. book publishing industry: 1989–1994. *Journal of Media Economics* 12, 165–80.

Greco, A.N. 2000. Market concentration levels in the U.S. consumer book industry: 1995–1996. *Journal of Cultural Economics* 24, 321–36.

Hjorth-Andersen, Chr. 2000. A model of the Danish book market. *Journal of Cultural Economics* 24, 27–43.

Holahan, W.L. 1979. A theoretical analysis of resale price maintenance. *Journal of Economic Theory* 21, 411–20.

Klein, D.C. 2000. Web strategies for professional publishers: developing an information services portal. *Learned Publishing* 13, 83–94.

Landes, W.M. and Posner, R.A. 1989. An economic analysis of copyright law. *Journal of Legal Studies* 18, 325–63.

Latcovich, S. and Smith, H. 2001. Pricing, sunk costs, and market structure online: evidence from book retailing. *Oxford Review of Economic Policy* 17, 217–34.

Marvel, H.P. and McCafferty, S. 1984. Resale price maintenance and quality certification. *RAND Journal of Economics* 15, 346–59.

Mathewson, F. and Winter, R. 1998. The law and economics of resale price maintenance. *Review of Industrial Organisation* 13, 57–84.

O'Hagan, J.W. 1998. *The State and the Arts: An Analysis of Key Economic Policy Issues in Europe and the United States*. North Hampton, MA: Edward Elgar.

Ornstein, S.I. 1985. Retail price maintenance and cartels. *Antitrust Bulletin* 30, 401–32.

Ottavanio, G.I.P. and Thisse, J.F. 1999. Monopolistic competition, multiproduct firms and optimum product diversity. Discussion Paper No. 9919. Louvain: CORE.

Ours, J.C. van. 1990. De Nederlandse boekenmarkt tussen stabiliteit en verandering. *Massacommunicatie* 18, 22–35.

Plant, A. 1934. The economic aspects of copyright in books. *Economica* 1, 167–95.

van der Ploeg, F. 2004. Beyond the dogma of the fixed book price agreement. *Journal of Cultural Economics* 28, 1–20.

van der Ploeg, F. 2006. The making of cultural policy: a European perspective. In *The Handbook of the Economics of Art and Culture*, ed. D. Ginsburgh and D. Throsby. Amsterdam: North-Holland.

Ringstad, V. 2004. On the cultural blessings of fixed book prices: fact or fiction. *International Journal of Cultural Policy* 10, 351–65.

Rosen, S. 1981. The economics of superstars. *American Economic Review* 71, 848–58.

Rürup, B., Klopfleisch, R. and Stumpp, H. 1997. *Ökomomische Analyse der Buchpreisbindung*. Frankfurt: Hessischer Verleger- und Buchhändler-Verband, Büchhändler-Vereinigung.

Sutton, J. 2000. *Marshall's Tendencies. What can Economists Know?* Cambridge, MA: Leuven University Press and MIT Press.

Szenberg, M. and Youngkoo Lee, E. 1994. The structure of the American book publishing industry. *Journal of Cultural Economics* 18, 313–22.

Throsby, D.C. 2001. *Economics and Culture*. Cambridge: Cambridge University Press.

Tietzel, M. 1995. *Literaturökonomik*. Tübingen: Mohr.

Tirole, J. 1998. *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

Tullock, G. 1980. Efficient rent seeking. In *Towards a Theory of the Rent-Seeking Society*, ed. J.M. Buchanan, R.D. Tollison and G. Tullock. College Station, TX: Texas A&M Press.

Uitermark, P.J. 1986. Verticale prijsbinding van boeken; concurrentie en cultuur: een aanzet tot analyse. In *De Vaste Boekenprijs: Pro's en Contra's*. Leiden: Instituut voor Onderzoek van Overheidsuitgaven.

Whyte, J.L. 1994. Breaking the bookshop cartel. *Long Range Planning* 27, 75–87.

Yetkiner, I.H. and Horvth, C. 2000. *Macroeconomic Implications of Virtual Shopping: A Theoretical Approach*. Groningen: University of Groningen Press.

How to cite this article

van der Ploeg, Frederick, Marcel Canoy and Jan van Ours. "books, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000218> doi:10.1057/9780230226203.0146

The New Palgrave Dictionary of Economics Online

bootstrap

Joel L. Horowitz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data. It is often much more accurate in finite samples than ordinary asymptotic approximations are. This is important in applied research, because the familiar asymptotic normal and chi-square approximations can be very inaccurate. When this happens, the difference between the true and nominal coverage probability of a confidence interval or rejection probability of a test can be very large, and inference can be highly misleading. The bootstrap often greatly reduces errors in coverage and rejection probabilities, thereby making reliable inference possible.

Keywords

asymptotic distribution; asymptotic refinements; bias reduction; bootstrap; conditional Kolmogorov test statistic; Edgeworth approximations; maximum score estimator; Monte Carlo simulation; probability; probit models; statistical inference; statistics and economics; Subsampling; Tobit model

Article

The bootstrap is a method for estimating the sampling distribution of an estimator or test statistic by resampling one's data. It amounts to treating the data as if they were the population for the purpose of evaluating the distribution of interest. Under mild regularity conditions, the bootstrap yields an approximation to the distribution of an estimator or test statistic that is at least as accurate as the approximation obtained from ‘ordinary’ or first-order asymptotic theory. Thus, the bootstrap provides a way to substitute computation for mathematical analysis if calculating the asymptotic distribution of an estimator or statistic is difficult. Moreover, the bootstrap is often more accurate in finite samples than first-order asymptotic approximations are but does not entail the algebraic complexity of higher-order expansions. Thus, it can provide a practical method for improving upon first-order approximations. Such improvements are called ‘asymptotic refinements’.

The bootstrap is of considerable importance in applied research. Many important statistics in econometrics have complicated asymptotic distributions that depend on nuisance parameters and, therefore, cannot be tabulated. Examples include the conditional Kolmogorov test statistic of Andrews (1997) and Manski's (1975; 1985) maximum score estimator for a binary-response model. The bootstrap and related resampling techniques provide practical methods for estimating the distributions of such statistics. In other cases, the statistic of interest has a familiar distribution but with a complicated standard error that is difficult to work with analytically (for example, Horowitz and Manski, 2000). Again, the bootstrap provides a practical method for carrying out inference.

The bootstrap's ability to provide asymptotic refinements is especially important in applied research. First-order asymptotic approximations (for example, asymptotic normal and chi-square approximations) can be very inaccurate with the sample sizes that are found in applications. When this happens, the difference between the true and nominal coverage probability of a confidence interval (error in the coverage probability or ECP) can be very large. Similarly, the difference between the true and nominal probability that a test rejects a correct null hypothesis (error in the rejection probability or ERP) can be very large. Consequently, inference based on first-order asymptotic approximations can be highly misleading. White's (1982) information matrix test is a well-known example of this. There are many others. The bootstrap often greatly reduces the ECPs of confidence intervals and ERPs of tests, thereby making reliable inference possible.

Bias reduction is another use of the bootstrap's ability to provide asymptotic refinements. It is not unusual for an asymptotically unbiased estimator to have a large finite-sample bias. This may cause the estimator's finite-sample mean-square error to be very large. The bootstrap can be used to reduce the estimator's finite-sample bias and, thereby, its finite-sample mean-square error.

The bootstrap has been the object of research in statistics since its introduction by Efron (1979). The results of this research are synthesized in the books by Beran and Ducharme (1991), Davison and Hinkley (1997), Efron and Tibshirani (1993), Hall (1992), Mammen (1992), and Shao and Tu (1995). Hall (1994), Horowitz (1997, 2003), Maddala and Jeong (1993), and Vinod (1993) provide reviews with an econometric orientation. Horowitz (2001) provides a detailed discussion of the theory and use of the bootstrap in econometrics.

This article assumes that the data are an independent random sample from some distribution. Horowitz (2001) and Lahiri (2003) discuss bootstrap methods for time-series data.

1 How the bootstrap works

This section explains why the bootstrap works and how it is implemented in simple settings. The estimation problem to be solved may be stated as follows. Let the data, $\{X_i: i = 1, \ldots, n\}$, be a random sample of size *n* from a probability distribution whose cumulative distribution function (CDF) is *F*. Let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic (that is, a function of the data), possibly a test statistic. Let $G_n(\tau, F) \equiv P(T_n \leq \tau)$ denote the exact, finite-sample CDF of T_n. Usually, $G_n(\tau, F)$ is a different function of τ for different distributions *F*. An exception occurs if $G_n(\tau, F)$ does not depend on *F*, in which case T_n is said to be *pivotal*, but pivotal statistics are not available in most applications. Therefore, $G_n(\tau, F)$ cannot be calculated if, as is usually the case in applications, *F* is unknown. The bootstrap is a

method for estimating $G_n(\tau, F)$ or features of it such as its quantiles when F is unknown.
First-order asymptotic distribution theory is another method for estimating $G_n(\tau, F)$. The asymptotic distributions of many econometric statistics are standard normal or chi-square, possibly after centring and normalization, regardless of the distribution from which the data were sampled. Such statistics are called *asymptotically pivotal*, meaning that their asymptotic distributions do not depend on unknown population parameters. Let $G_\infty(\tau, F)$ denote the asymptotic distribution of T_n. If T_n is asymptotically pivotal, then $G_\infty(\cdot, F) = G_\infty(\cdot)$ does not depend on F. Therefore, if n is sufficiently large, $G_n(\cdot, F)$ can be estimated by $G_\infty(\cdot)$ without knowing F. This method for estimating $G_n(\cdot, F)$ is often easy to implement and is widely used. However, $G_\infty(\cdot)$ can be a poor approximation to $G_n(\cdot, F)$ with samples of the sizes encountered in applications.
The bootstrap provides an alternative approximation to $G_n(\cdot, F)$. Whereas first-order asymptotic approximations replace the unknown distribution function G_n with the known function G_∞, the bootstrap replaces the unknown distribution function F with a consistent estimator such as the empirical distribution function of the data. Let F_n denote the estimator of F. The bootstrap estimator of $G_n(\cdot, F)$ is $G_n(\cdot, F_n)$. Usually, $G_n(\cdot, F_n)$ cannot be evaluated analytically. It can, however, be estimated with arbitrary accuracy by carrying out a Monte Carlo simulation in which random samples are drawn from the data. Thus, the bootstrap is usually implemented by Monte Carlo simulation. The Monte Carlo procedure for estimating $G_n(\tau, F_n)$ is:

- *Step 1*: Generate a bootstrap sample, $\{X_i^* : i = 1, \ldots, n\}$ by sampling the estimation data randomly with replacement.
- *Step 2*: Compute $T_n^* = T_n(X_1^*, \ldots, X_n^*)$.
- *Step 3*: Use the results of many repetitions of steps 1 and 2 to compute the empirical probability of the event $T_n^* \le \tau$ (that is, the proportion of repetitions in which this event occurs).

If T_n is a test statistic, then the bootstrap can be used to estimate its critical value. Consider a test that rejects the null hypothesis, H_0, if $|T_n|$ is too large. The exact α-level critical value, $z_{n,\alpha/2}$, is the solution to $G_n(z_{n,\alpha/2}, F) - G_n(-z_{n,\alpha/2}, F) = 1 - \alpha$. Unless T_n is pivotal, however, this equation cannot be solved in an application because F is unknown. Therefore, the exact, finite-sample critical value cannot be obtained in an application if T_n is not pivotal. The bootstrap replaces F with F_n. Thus, the bootstrap critical value, $z_{n,\alpha/2}^*$, solves $G_n(z_{n,\alpha/2}^*, F_n) - G_n(-z_{n,\alpha/2}^*, F_n) = 1 - \alpha$. This equation usually cannot be solved analytically, but $z_{n,\alpha/2}^*$ can be estimated with any desired accuracy by Monte Carlo simulation. To illustrate, suppose, as often happens in applications, that T_n is an asymptotically standard normal, Studentized estimator of a parameter θ whose value under H_0 is θ_0. That is, $T_n = n^{1/2}(\theta_n - \theta_0) / s_n$, where θ_n is the estimator of θ, $n^{1/2}(\theta_n - \theta_0) \to^d N(0, \sigma^2)$ under H_0, and s_n^2 is a consistent estimator of σ^2. Then the Monte Carlo procedure for computing $z_{n,\alpha/2}^*$ is:

- *Step 1*. Use the estimation data to compute θ_n.
- *Step 2*. Generate a bootstrap sample of size *n* by sampling the data randomly with replacement. Compute the estimators of θ and σ from the bootstrap sample. Call the results θ_n^* and s_n^*. The bootstrap version of T_n is $T_n^* = n^{1/2}(\theta_n^* - \theta_n) / s_n^*$.
- *Step 3*. Use the results of many repetitions of step 2 to compute the empirical distribution of $|T_n^*|$. Set $z_{n,\alpha/2}^*$ equal to the $1-\alpha$ quantile of this distribution.

A test based on $|T_n|$ and the bootstrap critical value rejects H_0 at the α-level if $|T_n| > z_{n,\alpha/2}^*$. A symmetrical $1-\alpha$ confidence interval for θ based on the bootstrap critical value is $\theta_n - z_{n,\alpha/2}^* s_n \le \theta \le \theta_n + z_{n,\alpha/2}^* s_n$. For reasons that are explained in Section 2, use of the bootstrap critical value $z_{n,\alpha/2}^*$ instead of the critical value based on the asymptotic normal distribution can greatly reduce the ERP of a test of a hypothesis about θ and the ECP of a confidence interval for θ.
Since F_n and F are different functions, $G_n(\cdot, F_n)$ and $G_n(\cdot, F)$ are also different functions unless T_n is pivotal. Therefore, the bootstrap estimators $G_n(\cdot, F_n)$ and $z_{n,\alpha/2}^*$ are only approximations to the exact finite-sample CDF and critical value of T_n, $G_n(\cdot, F)$ and $z_{n,\alpha/2}$. However, F_n is close to F when n is large. Therefore, if G_n is a sufficiently smooth function, $G_n(\cdot, F_n)$ will be close to $G_n(\cdot, F)$. Moreover, we can expect $z_{n,\alpha/2}^*$ to approach $z_{n,\alpha/2}$ as $n \to \infty$. In other words, the bootstrap provides an approximation to the sampling distribution and critical value of T_n that becomes increasingly accurate as n increases. This property of the bootstrap is called *consistency*. Beran and Ducharme (1991) and Mammen (1992) give formal conditions under which the bootstrap is consistent. Horowitz (2001) gives some econometrically relevant examples in which the bootstrap is not consistent and, therefore, cannot be used to estimate the distribution of a statistic. These include Manski's maximum score estimator, the distribution of a parameter on the boundary of the parameter set, and estimation of the maximum of a sample.
When the bootstrap is inconsistent (that is, $G_n(\cdot, F_n) - G_n(\cdot, F)$ does not converge to 0), subsampling procedures can be used to estimate $G_n(\cdot, F)$. One approach to subsampling consists of drawing samples of size $m < n$ by sampling the data randomly *without* replacement. This produces random samples from the true population distribution of the data, F, not the empirical distribution, F_n, from which bootstrap samples are drawn. Consequently, subsampling yields a consistent estimator of $G_n(\cdot, F)$, even when the bootstrap does not. Politis, Romano and Wolf (1999) describe the theory of subsampling and methods for implementation. Subsampling is consistent in all known settings of practical importance, so it is much more widely applicable than the bootstrap. The price of this versatility, however, is reduced accuracy. The approximation provided by subsampling is typically less accurate than that provided by first-order asymptotic distribution theory, and subsampling can be much less accurate than the bootstrap when the bootstrap is consistent.

2 Asymptotic refinements

The bootstrap provides asymptotic refinements for statistics that are asymptotically pivotal. That is, the bootstrap provides a better approximation to the distribution of an asymptotically pivotal statistic than does ‘ordinary’ asymptotic distribution theory. A statistic is asymptotically pivotal if its asymptotic distribution does not depend on unknown population parameters. All the familiar test statistics whose asymptotic distributions are standard normal or chi-square are asymptotically pivotal. Estimates of regression coefficients, standard errors, and other population parameters typically are not asymptotically pivotal. The bootstrap does not provide asymptotic refinements for statistics that are not asymptotically pivotal. Whenever possible, the bootstrap should be applied to asymptotically pivotal statistics as opposed to statistics that are not asymptotically pivotal.

The bootstrap's ability to provide asymptotic refinements has important practical consequences. Specifically, the bootstrap can be used to obtain estimates of finite-sample critical values for test statistics that are more accurate than critical values obtained from the asymptotic normal or chi-square approximations. The use of bootstrap-based critical values can greatly reduce the ERP of a test and ECP of a confidence interval.

The bootstrap provides asymptotic refinements because it provides a higher-order asymptotic approximation, called an Edgeworth approximation, to $G_n(\tau, F)$. Suppose that T_n is asymptotically distributed as $N(0, 1)$, and let Φ denote the standard normal CDF. Then $G_n(\tau, F_n) - G_n(\tau, F) = O_p(n^{-1})$, whereas $G_n(\tau, F) - \Phi(\tau) = O(n^{-1/2})$. Thus, the error made by the bootstrap approximation to $G_n(\tau, F)$ converges to 0 more rapidly than does the error made by the asymptotic normal approximation. For $|T_n|$ or an asymptotic chi-square statistic, the error made by the bootstrap approximation is $O_p(n^{-3/2})$ whereas the error made by the asymptotic normal or chi-square approximation is $O(n^{-1})$. See Hall (1992) and Horowitz (2001) for details.

Rejection probabilities of tests and coverage probabilities of confidence intervals based on bootstrap critical values can be even more accurate. The ERPs of symmetrical tests and ECPs of symmetrical confidence intervals are $O(n^{-2})$ when the bootstrap is used to obtain the critical value, whereas they are $O(n^{-1})$ when the asymptotic normal or chi-square approximation is used. (A test based on an asymptotic chi-square statistic is symmetrical. So is a test that rejects the null hypothesis when $|T_n|$ exceeds the critical value, where T_n is asymptotically distributed as $N(0, 1)$.) Thus, the ERPs and ECPs of symmetrical tests and confidence intervals converge to 0 much more rapidly with bootstrap-based critical values than with critical values based on the asymptotic normal or chi-square approximations. The practical consequence of this is that the bootstrap often achieves spectacular reductions in the numerical values of ERPs and ECPs. Section 3 provides two examples of this. Horowitz (1997; 2001) provides others.

With one-sided tests and confidence intervals, the ERP and ECP are usually $O(n^{-1})$ with bootstrap critical values and $O(n^{-1/2})$ with asymptotic chi-square or normal critical values. However, there are cases in which the ERP of a bootstrap-based test is $O(n^{-3/2})$ (Hall, 1992; Davidson and MacKinnon, 1999).

3 Examples

This section presents two examples that illustrate the bootstrap's ability to reduce the ERP of a test or the ECP of a confidence interval.

3.1 White's (1982) information-matrix (IM) test

This is a specification test for parametric models estimated by maximum likelihood. The test statistic is asymptotically chi-square distributed, but the asymptotic distribution is a poor approximation to the finite-sample distribution.
Horowitz (1994) reports the results of Monte Carlo experiments that investigate the ERPs of the IM test with bootstrap critical values. Some of these results are summarized in Table 1, which gives the results of applying the Chesher (1983) and Lancaster (1984) form and White's (1982) original form of the test to Tobit and binary probit models. The results show that the ERPs are very large when critical values based on the asymptotic chi-square distribution are used. When bootstrap critical values are used, however, the ERPs are very small. The bootstrap essentially eliminates the differences between the true and nominal rejection probabilities of the two forms of the IM test.

Empirical rejection probabilities of nominal 0.05-level information-matrix tests of probit and tobit models

		Rejection probability using			
N	Distr. of X	Asymp. critical values		Bootstrap crit. values	
		White	Chesh.-Lan.	White	Chesh.-Lan.
Binary probit models					
50	N(0,1)	0.385	0.904	0.064	0.056
	U(−2,2)	0.498	0.920	0.066	0.036
100	N(0,1)	0.589	0.848	0.053	0.059
	U(−2,2)	0.632	0.875	0.058	0.056
Tobit models					
50	N(0,1)	0.112	0.575	0.083	0.047
	U(−2,2)	0.128	0.737	0.051	0.059
100	N(0,1)	0.065	0.470	0.038	0.039
	U(−2,2)	0.090	0.501	0.046	0.052

Source: Horowitz (1994).

3.2 Estimation of covariance structures

In estimation of covariance structures, the objective is to estimate the covariance matrix of a $k\times 1$ vector X subject to restrictions that reduce the number of unique, unknown elements to $r < k(k+1)/2$.

Estimates of the r unknown elements can be obtained by minimizing the weighted distance between sample moments and the estimated population moments. Weighting all sample moments equally produces the equally weighted minimum distance (EWMD) estimator, whereas choosing the weights to maximize asymptotic estimation efficiency produces the optimal minimum distance (OMD) estimator. The OMD estimator has poor finite-sample performance in applications (Abowd and Card, 1989). Horowitz (1998) reports the results of a Monte Carlo investigation of the ability of the bootstrap to reduce the ERPs of nominal 95 per cent symmetrical confidence intervals based on the OMD estimator. In each experiment, X has 10 components, and the sample size is n=500. The j'th component of X, X_j (j=1, …, 10) is generated by $X_j=(Z_j+\rho Z_{j+1})/(1+\rho^2)^{1/2}$, where $Z_1, \ldots, Z_{11}$ are i.i.d. random variables with means of 0 and variances of 1, and $\rho=0.5$. The Z's are sampled from five different distributions depending on the experiment. It is assumed that ρ is known and that the components of X are known to be identically distributed and to follow MA(1) processes. The estimation problem is to infer the scalar parameter θ that is identified by the moment conditions $Var(X_j)=\theta$ (j=1, …, 10) and $Cov(X_j, X_{j-1}) = \rho\theta/(1+\rho^2)$ (j=2, …, 10).

The results of the experiments are summarized in Table 2. The coverage probabilities of confidence intervals based on asymptotic critical values are far below the nominal value of 0.95 except in the experiment with uniform Z's. However, the use of bootstrap critical values greatly reduces the ERPs. In the experiments with normal, Student t, uniform, or exponential Z's, the bootstrap essentially eliminates the errors in the coverage probabilities of the confidence intervals.

Empirical coverage probabilities of nominal 95 per cent symmetrical confidence intervals based on the OMD estimator

Distr. of Z	Asymptotic critical value	Bootstrap critical value
Uniform	0.93	0.96
Normal	0.85	0.95
Student t with 10 d.f.	0.79	0.95
Exponential	0.54	0.96
Lognormal	0.03	0.91

Source: Horowitz (1998).

I thank Federico Bugni for helpful comments. The preparation of this article was supported in part by NSF Grant SES-0352675.

Bibliography

Abowd, J.M. and Card, D. 1989. On the covariance of earnings and hours changes. *Econometrica* 57, 411–45.

Andrews, D.W.K. 1997. A conditional Kolmogorov test. *Econometrica* 65, 1097–128.

Beran, R. and Ducharme, G.R. 1991. *Asymptotic theory for bootstrap methods in statistics*. Montréal:

Les Publications CRM, Centre de Recherches Mathematiques, Université de Montréal.

Chesher, A. 1983. The information matrix test. *Economics Letters* 13, 45–8.

Davidson, R. and MacKinnon, J.G. 1999. The size distortion of bootstrap tests. *Econometric Theory* 15, 361–76.

Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1–26.

Efron, B. and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer.

Hall, P. 1994. Methodology and theory for the bootstrap. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D.F. McFadden. Amsterdam: North-Holland.

Horowitz, J.L. 1994. Bootstrap-based critical values for the information matrix test. *Journal of Econometrics* 61, 395–411.

Horowitz, J.L. 1997. Bootstrap methods in econometrics: theory and numerical performance. In *Advances in Economics and Econometrics: Theory and Applications*, Seventh World Congress, vol. 3, ed. D.M. Kreps and K.F. Wallis. Cambridge: Cambridge University Press.

Horowitz, J.L. 1998. Bootstrap methods for covariance structures. *Journal of Human Resources* 33, 39–61.

Horowitz, J.L. 2001. The bootstrap in econometrics. In *Handbook of Econometrics*, vol. 5, ed. J.J. Heckman and E.E. Leamer. Amsterdam: North-Holland.

Horowitz, J.L. 2003. The bootstrap in econometrics. *Statistical Science* 18, 211–8.

Horowitz, J.L. and Manski, C.F. 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95, 77–84.

Lahiri, S.N. 2003. *Resampling Methods for Dependent Data*. New York: Springer.

Lancaster, T. 1984. The covariance matrix of the information matrix test. *Econometrica* 52, 1051–3.

Maddala, G.S. and Jeong, J. 1993. A perspective on application of bootstrap methods in econometrics. In *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao and H.D. Vinod. Amsterdam: North-Holland.

Mammen, E. 1992. *When Does Bootstrap Work? Asymptotic Results and Simulations*. New York: Springer.

Manski, C.F. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C.F. 1985. Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27, 313–34.

Politis, D.N, Romano, J.P. and Wolf, M. 1999. *Subsampling*. New York: Springer.

Shao, U. and Tu, D. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Vinod, H.D. 1993. Bootstrap methods: applications in econometrics. In *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao and H.D. Vinod. Amsterdam: North-Holland.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–26.

How to cite this article

Horowitz, Joel L. "bootstrap." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000307> doi:10.1057/9780230226203.0147

The New Palgrave Dictionary of Economics Online

Borch, Karl H. (1919 – 1986)

Knut K. Aase
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Borch, K.; game theory; general equilibrium; insurance; Norwegian School of Economics and Business Administration (NHH); reinsurance contracts; risk

Article

Karl Borch was born in Sarpsborg, Norway, on 13 March 1919. He graduated with an MSc in actuarial mathematics at the University of Oslo in 1947, and a Ph.D. in 1962.
From 1947 he worked for UNESCO and OECD until in 1959 he started his academic career at the Norwegian School of Economics and Business Administration (NHH) in Bergen, where he was appointed professor of insurance in 1963, a position he held until his untimely death on 2 December 1986, only just before his retirement was due.
In *Who's Who in Economics* (1986, p. 103) he wrote: ‘When in 1959 I got a research post which gave me almost complete freedom, as long as my work was relevant to insurance, I naturally set out to develop an economic theory of insurance.’ That within a year he should have made a decisive step in that direction is amazing. What he did during these first years of his research career was to write the first of a long series of seminal papers, which were to put him on the map as one of the world's leading scholars in his field.
One important contribution of his papers in *Skandinavisk Aktuarietidskrift* (1960a) and *Econometrica* (1962) was to derive testable implications from the abstract model of general equilibrium with markets for contingent claims. In this way, he brought economic theory to bear on insurance problems, thereby opening up that field considerably; and he brought the experience of reinsurance contracts to bear on the interpretation of economic theory, thereby considerably enlivening that theory.
Practically his entire production was centred on the topic of uncertainty in economics. Many of his thoughts were formulated in his successful book *The Economics of Uncertainty* (1968a), also available in Spanish, German and Japanese. He gave the first graduate lectures at NHH, where he supervised many Ph.D. students.
He had more than 150 publications, among them three books (published in 1968, 1974 and 1990). The

last one, *Economics of Insurance*, has also been translated into Chinese. Best known to actuaries is perhaps his pioneering work on Pareto-optimal risk exchanges in reinsurance (for example, 1960a). Borch also made many contributions to the application of game theory to insurance: in particular, he characterized the Nash bargaining solution of a reinsurance syndicate (1960b).
Borch served on many editorial boards, and he helped organize several key international conferences abroad and at NHH.

See Also

- general equilibrium
- insurance mathematics
- Pareto principle and competing principles
- risk
- risk aversion
- uncertainty

Selected works

1960a. The safety loading of reinsurance premiums. *Skandinavisk Aktuarietidsskrift* 43, 163–84.

1960b. Reciprocal reinsurance treaties seen as a two-person cooperative game. *Skandinavisk Aktuarietidsskrift* 43, 29–58.

1962. Equilibrium in a reinsurance market. *Econometrica* 30, 424–44.

1967. The theory of risk. *Journal of the Royal Statistical Society*, Series B 29, 423–67.

1968a. *The Economics of Uncertainty*. Princeton, NJ: Princeton University Press.

1968b. General equilibrium in the economics of uncertainty. In *Risk and Uncertainty*, ed. K. Borch and J. Mossin. London: Macmillan.

1969. A note on uncertainty and indifference curves. *Review of Economic Studies* 36, 1–4.

1974. *The Mathematical Theory of Insurance*. Lexington, MA: D.C. Heath.

1985. A theory of insurance premiums. *The Geneva Papers on Risk and Insurance* 10, 192–208.

1986. Entry in *Who's Who in Economics: A Biographical Dictionary of Major Economists, 1700–1986*, 2nd edn., ed. M. Blaug. Cambridge, MA: MIT Press.

1990. (With K. Aase and A. Sandmo, ed.) *Economics of Insurance*. Advanced Textbooks in Economics 29. Amsterdam, New York, Oxford, Tokyo: North-Holland.

How to cite this article

Aase, Knut K. "Borch, Karl H. (1919–1986)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000326> doi:10.1057/9780230226203.0148

The New Palgrave Dictionary of Economics Online

Borda, Jean-Charles de (1733 – 1799)

Charles K. Rowley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Black, D.; Borda, J.-C. de; Condorcet criterion; Condorcet, Marquis de; Laplace, Marquis de; mathematical theory of elections; method of marks; public choice; voting

Article

The second half of the 18th century in France was one of the outstanding epochs of scientific thought and witnessed significant attempts to carry the methods of rigorous and mathematical thought beyond the physical and into the realms of the human sciences. A brilliant start was made in political science by three French academicians, namely Borda, Condorcet and Laplace, with contributions which now play a central role in the literature of public choice. It is a salutary warning to those who view science as endlessly progressive to note that the contributions of these outstanding academicians were lost for two centuries until they were rediscovered in 1958 by Duncan Black.

Borda was the first of the three to develop a mathematical theory of elections shortly after becoming a member of the Academy of Sciences. Born in 1733 in Dax, near Bordeaux, Borda was successively an officer of cavalry, a naval captain, and a scholar of mathematical physics as well as an innovator in the field of scientific instruments. Newly elected to the Academy of Sciences, Borda read a paper entitled ‘Sur la forme des elections’ on 16 June 1770. Four members were charged to report on it, but failed to do so.

The Academy was not to consider elections again during the succeeding 14 years, until Borda again read a paper on elections in July 1784 following the favourable report by Bossut and Coulomb on Condorcet's manuscript, *Essai*. Borda's paper had been printed in the *Histoire de l’ Academie Royale des Sciences* in 1781, three years prior to this reading. It was finally published in 1784. In essence, it reflected the content of his 1770 paper. Condorcet had become acquainted with Borda's contribution prior to writing his *Essai*, as a consequence of the strong oral tradition of the Academy. He acknowledged the powerful influence of Borda's ideas upon his own writings.

Borda was concerned that the single vote system of elections might select the wrong candidate. He illustrated by reference to a situation in which eight electors had candidate A as first preference, seven

had candidate B, and six had candidate C. On the single vote, A would be elected, although the electors preferred B *or* C to A by a majority of 13 to 8. In essence, Borda was utilizing what later became known as the *Condorcet criterion*, though he failed to develop it himself. Instead, he attempted to remedy the defect of the single vote system by the method of marks, which he presented in two forms. Since one form is a special case of the other, only the more general form is here outlined.

The method of marks requires each elector to rank all the candidates by order of merit. The candidate is then allocated marks by reference to his ranking by each voter, for example, three marks for first place, two marks for second, and one mark for last in a three candidate election. The marks are then totalled across all elections. The candidate with the largest aggregate of marks is the winner.

To illustrate how the method of marks may provide a different result from that of the single vote, let us expand Borda's original example as outlined above into the form of Table 1.

Rank order of candidates by electors					
A	A	B	B	C	C
B	C	A	C	A	B
C	B	C	A	B	A
1	7	1	6	1	5

In the Table 1 example, Candidate A would receive an aggregate of 39 marks, Candidate B receives an aggregate of 41 marks, and Candidate C receives an aggregate of 46 marks. Candidate C is the winner, reversing the single vote outcome.

The method of marks allows a role for preference intensities, albeit only on a strictly linear scale, within the electoral process. For this reason, it has been called a ‘neo-utilitarian’ approach (Sugden, 1981). The method is not strategy proof, since voters will tend to lower the ranking of the candidate most threatening to their preferred candidate to the lowest level, irrespective of their actual preferences. Borda himself clearly recognized this danger, but, in an age more honourable than our own, was merely moved to comment: ‘My scheme is only intended for honest men.’

Borda's paper did not attempt to provide a comprehensive theory of elections. It failed to develop, though it implicitly embraced, the criterion of Condorcet. More important, it offered no real insight into the nature and/ or the objectives of group decisions. It was, however, a significant first step in both directions. The method of marks is extremely effective if each elector genuinely desires to secure the election of ‘that candidate who should be the most generally acceptable’ (Black, 1958). In reality, most electors desire to secure the election of their most favoured candidate. Herein lies the weakness of the method of marks.

Shortly after hearing Borda's paper in 1784, the Academy adopted his method in elections to its membership. The method of marks remained in use until 1800, when it was attacked by a new member, and soon afterwards, was modified. The new member in question was Napoleon Bonaparte.

See Also

- Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de

- social choice
- voting paradoxes

Selected work

1781. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royal des Sciences*, Paris.

Bibliography

Black, D. 1958. *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.

Condorcet, Marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris.

Lacroix, S.E. 1800. *Eloge historique de Jean-Charles Borda*. Paris.

Laplace, Marquis de. 1812. Leçons de mathématique données à I'Ecole Normale en 1885. *Journal de l'Ecole Polytechnique*.

Mascart, J. 1919. *La vie et les travaux du Chevalier Jean-Charles de Borda (1733–1799): épisodes de la vie scientifique au XVIII siécle*. Lyon.

Sudgen, R. 1981. *The Political Economy of Public Choice: An Introduction to Welfare Economics*. Oxford: Martin Robertson.

How to cite this article

Rowley, Charles K. "Borda, Jean-Charles de (1733–1799)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000169>
doi:10.1057/9780230226203.0149

The New Palgrave Dictionary of Economics Online

border effects

John H. Rogers
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International finance and trade economists have traditionally focused on the behaviour of cross-country prices and factor returns and the flow of goods and capital across nations. Studying these same variables across locations within countries provides a baseline for measuring the influence of the border. The ‘border effect’ is the difference between international and intra-national magnitudes. Large border effects were initially found in consumer goods prices and trade volumes. Subsequent studies have examined robustness and looked for explanations of the border effect.

Keywords

border effects; distance; exchange rate volatility; gravity models; internal trade; international trade barriers; market integration; price volatility; productivity shocks; purchasing power parity; sticky prices; trade flows; transaction costs; transportation costs; United States–Canada Free Trade Agreement

Article

International finance and trade economists have traditionally focused on the behaviour of cross-country prices and factor returns and the flow of goods and capital across nations. Studying these same variables across locations *within countries* provides a baseline for measuring the influence of the border. The ‘border effect’ is, to speak loosely, the difference between international and intra-national magnitudes. Large border effects were initially found in consumer goods prices and trade volumes (Engel and Rogers, 1996; McCallum, 1995) in data from the United States and Canada. Subsequent studies have examined robustness and looked for explanations of the border effect, often through extensions to other countries’ data-sets.

The starting point of Engel and Rogers (1996) is a fundamental proposition of economic theory: in the absence of transaction costs, identical goods must sell for the same price. Prices will fail to equalize when there are barriers, natural ones or man-made, to the free movement of goods. There are several

reasons to expect that national borders would give rise to such barriers.

Engel and Rogers (1996) examine the behaviour of prices of 14 categories of consumer goods and services in 14 US cities and 9 Canadian cities during the period 1978–94. They measure the border effect by comparing the extent to which prices of a particular category of goods fluctuate across cities intra-nationally with price fluctuations for city pairs that lie across the border. With q_{ij} defined as the log of the price of some good in city i relative to its price in city j, let $V(q_{ij})$ be a measure of relative price volatility over the sample time period. Engel and Rogers relate this to various explanatory variables including distance between cities and a ‘border dummy’ for whether the cities lie in different countries. They run regressions of the form:

$$V(q_{ij}) = \beta_1 d_{ij} + \beta_2 B_{ij} + \sum_{k=1..m} \lambda_k D_k, \quad (1)$$

where d_{ij} is the log of the distance between cities i and j; B_{ij} is a dummy variable equal to 1 if cities i and j are in different countries; and D_k are dummy variables for each city. Engel and Rogers (1996) consistently find that β_2 is positive, highly statistically significant, and large in magnitude. The coefficient on distance, β_1, is usually positive and significant.

McCallum (1995) estimates the effect of the border on trade flows between Canadian provinces and US states. McCallum's data-set includes imports and exports for all pairs of Canadian provinces, as well as imports and exports between each of the ten provinces and each of the 50 US states. The data are from 1988. McCallum uses a traditional gravity model, positing that trade is a function of the distance between trading partners and their individual economic sizes, measured by gross domestic product. (See Anderson, 1979, for model development, and Rose, 2000, for a noteworthy application.) McCallum augments the standard gravity model with a dummy variable equal to 1 for pairs of Canadian provinces. The coefficient on McCallum's inter-provincial trade dummy variable is estimated to be positive and highly statistically significant. The point estimate implies that, other things equal, trade between two Canadian provinces is more than 20 times larger than trade between a province and a US state.

Anderson and van Wincoop (2003) are critical of the gravity equations employed in the border effects papers on trade flows. They argue that these equations suffer from omitted variables bias (requiring that a ‘multi-lateral resistance’ term be added) and incorrect comparative statics analysis. Anderson and van Wincoop develop a methodology that allows them to get around these shortcomings. Taking up McCallum's exercise using data for 1993, these authors show that the border effect on trade flows is, although still large, considerably smaller than calculated by McCallum.

Engel and Rogers (1996) suggest several reasons why the border should matter. First, there might be direct costs to crossing the border such as tariffs and other trade restrictions. Alternatively, markups might differ across locations and vary with exchange rate changes. Markets for non-traded inputs (wages, marketing services) might be more highly integrated on a national basis than in two places separated by a border. Or productivity shocks might be more similar for city pairs that lie within a

country than for cross-border pairs. Finally, Engel and Rogers consider a sticky-price explanation. Goods sold in the United States may be sticky in US dollar terms while goods sold in Canada are sticky in terms of Canadian dollars. A highly variable nominal exchange rate could then give rise to a large, positive value of β_2 because cross-border relative prices would fluctuate along with the nominal exchange rate while relative prices within countries remained fairly stable. Although Engel and Rogers do not conduct an exhaustive examination of different factors, they conclude, ‘Sticky prices appear to be one explanation but probably do not explain most of the border effect’ (1996, p. 1112). (The Engel–Rogers work has an intellectual predecessor in Mussa, 1986, who noted that CPI-based real exchange rates are more variable for Toronto versus Chicago, Vancouver versus Chicago, Toronto versus Los Angeles, and Vancouver versus Los Angeles, than for Toronto versus Vancouver and Chicago versus Los Angeles under floating exchange rates. Mussa attributed this to sticky prices.)

Using updated data, Engel and Rogers (2000) examine the stability of the border effect around the United States–Canada Free Trade Agreement. They find little evidence of a change across several break dates corresponding with the signing or implementation of the agreement.

Subsequent studies have examined different data-sets and attempted to understand the dynamics of the border effect. Parsley and Wei (2001) examine data from 96 US and Japanese cities from 1976 to 1997. They ask two related questions. First, is there any evidence that the Japan–US ‘border’ narrows over time? Second, is there evidence linking the evolution of the border effect with plausible economic candidates (for example, the unit cost of international transportation)? They show that the simple average of good-level real exchange rates tracks the nominal exchange rate closely, providing strong evidence of sticky prices in local currencies. They find evidence that the border effect between Japan and the United States declines over time. Furthermore, distance, shipping costs, and exchange rate variability collectively explain a substantial portion of the border effect.

Engel and Rogers (2001) use consumer price data from European cities in 11 countries from 1981 to 1997 to explore deviations from short-run purchasing power parity (PPP) across several national borders. The European data-set has many advantages over that consisting of observations from US and Canadian cities only. In the latter, there is no distinction between the border dummy and a measure of nominal exchange rate variability, since all cross-border pairs have the same nominal exchange rate. With the European data-set, Engel and Rogers are able to include both a border dummy variable (unity for city pairs lying across the border) and a measure of nominal exchange rate variability. This allows a distinction between the role of sticky local currency pricing and the various other ‘real’ barriers to market integration. The authors find that, even with nominal exchange rate variability taken into account, distance between cities and the border continue to have positive and significant effects on real exchange rate variability. However, these effects are smaller than the local currency pricing effect.

Gorodnichenko and Tesar (2005) re-examine the Engel–Rogers and Parsley–Wei papers. They run the same regression as the earlier papers but propose a different measure of the border effect. To understand their measure, let γ_U be the average relative price variance for city pairs within the United States; γ_C be the average for pairs within Canada; and β be the average relative price variance for cross-border city pairs (after controlling for distance). Engel and Rogers (1996) measure the border effect as $\beta - 0.5(\gamma_U + \gamma_C)$. Gorodnichenko and Tesar propose the ‘conservative’ measure: $\beta - \max(\gamma_U, \gamma_C)$. Since γ_U is not very different from β (a feature of the data noted by Engel and Rogers), the border effect is small when measured in the conservative way. Under the Gorodnischenko–Tesar scheme there

are two border effects, one for a Canadian crossing into the US market and the other for an American crossing into Canada. In this case one is quite large, the other relatively small. Engel and Rogers measure the border effect as the average of the two (as do Parsley and Wei for the US–Japan data). Gorodnichenko and Tesar use the smaller of the two.

A large body of literature has expanded upon McCallum's (1995) findings. As with the literature that followed Engel and Rogers (1996), many have analysed different data-sets, especially from other countries. Examples include Helliwell (1996; 1998), Wei (1996), Anderson and Smith (1999), Yi (2003), Wolf (2000), Hillberry and Hummels (2003), and Evans (2003). One important issue highlighted by these papers is the need for accurate measures of ‘internal trade’, that is, the amount that countries trade with themselves. This literature is exhaustively surveyed by Anderson and van Wincoop (2004).

Progress in explaining the border effect on trade flows has been made by decomposing total international trade barriers into barriers associated with geographic factors such as distance and barriers due to national borders. According to Anderson and van Wincoop (2004, Table 7), estimates from several papers using different data-sets (Wei, 1996; Eaton and Kortum, 2002; Evans, 2003; Anderson and van Wincoop, 2003) put the tariff equivalent cost of total international trade barriers at around 40–80 per cent. Anderson and van Wincoop categorize further investigation of the trade barriers associated with national borders as attempts to quantify the effects due to (*a*) language barriers, (*b*) use of different currencies, (*c*) information barriers, (*d*) contracting costs and security, and (*e*) policy barriers. To summarize the results from this literature, these authors suggest very rough calculations of an eight per cent policy-related barrier, a seven per cent language barrier, a fourteen per cent currency barrier, a six per cent information cost barrier, and a three per cent security barrier, well within the range of 25–50 per cent for overall border barriers reported by different authors for OECD countries.

See Also

- exchange rate volatility
- gravity models
- price dispersion
- trade costs

Bibliography

Anderson, J. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69, 106–16.

Anderson, J. and van Wincoop, E. 2003. Gravity with gravitas: a solution to the border puzzle. *American Economic Review* 93, 170–92.

Anderson, J. and van Wincoop, E. 2004. Trade costs. *Journal of Economic Literature* 42, 691–751.

Anderson, M. and Smith, S. 1999. Do national borders really matter? Canada–U.S. regional trade reconsidered. *Review of International Economics* 7, 219–27.

Eaton, J. and Kortum, S. 2002. Technology, geography and trade. *Econometrica* 70, 1741–79.

Engel, C. and Rogers, J. 1996. How wide is the border? *American Economic Review* 86, 1112–25.

Engel, C. and Rogers 2000. Relative price volatility: what role does the border play? In *Intranational Macroeconomics*, ed. G. Hess and E. van Wincoop. Cambridge: Cambridge University Press.

Engel, C. and Rogers, J. 2001. Deviations from purchasing power parity: causes and welfare costs. *Journal of International Economics* 55, 29–57.

Evans, C. 2003. The economic significance of national border effects. *American Economic Review* 93, 1291–312.

Gorodnichenko, Y. and Tesar, L. 2005. A re-examination of the border effect. Working Paper No. 11706. Cambridge, MA: NBER.

Helliwell, J. 1996. Do national boundaries matter for Quebec's trade? *Canadian Journal of Economics* 29, 507–22.

Helliwell, J. 1998. *How Much Do National Borders Matter?* Washington, DC: Brookings Institution.

Hillberry, R. and Hummels, D. 2003. Intranational home bias: some explanations. *Review of Economics and Statistics* 85, 1089–92.

McCallum, J. 1995. National borders matter: Canada–U.S. regional trade patterns. *American Economic Review* 85, 615–23.

Mussa, M. 1986. Nominal exchange rate regimes and the behavior of real exchange rates: evidence and implications. *Carnegie-Rochester Conference Series on Public Policy* 25, 117–214.

Parsley, D. and Wei, S.J. 2001. Explaining the border effect: the role of exchange rate variability, shipping costs and geography. *Journal of International Economics* 55, 87–105.

Rose, A. 2000. One money, one market: estimating the effect of common currencies on trade. *Economic Policy* 30, 7–45.

Wei, S.-J. 1996. Intra-national versus inter–national trade: how stubborn are nations in global integration? Working Paper No. 5531. Cambridge, MA: NBER.

Wolf, H. 2000. (Why) do borders matter for trade? In *Intranational Macroeconomics*, ed. G. Hess and E. van Wincoop. Cambridge: Cambridge University Press.

Yi, K.-M. 2003. A simple explanation for the border effect. Working paper, Federal Reserve Bank of New York.

How to cite this article

Rogers, John H. "border effects." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000315> doi:10.1057/9780230226203.0150

The New Palgrave Dictionary of Economics Online

Bouniatian, Mentor (1877 – 1969)

Mauro Boianovsky

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bouniatian argued that productive forces cannot be transferred to the future just through the accumulation of capital goods; the choice of production methods determines an equilibrium relation between aggregate consumption and the capital stock. Economic fluctuations are explained by both the increase of the proportion of income saved when output is growing and the period of time necessary for the production of capital goods. The temporary separation between investment and consumption decisions is reflected in a more than proportional increase of capital goods. Changes in the marginal utility of consumption and capital goods and a generalization of the old King's law explain changes in the price level.

Keywords

acceleration principle; Aftalion, A.; aggregate consumption; Bouniatian, M.; business cycles; capital accumulation; depreciation; exogenous and endogenous business cycle theories; Great Depression; Hobson, J. A.; Keynes, J. M.; King's law; Lauderdale, Eighth Earl of; Mitchell, W. C.; Mummery, A. F.; overproduction; quantity theory of money; saving–investment equality; sticky wages; subjective theory of value; Tugan-Baranovsky, M. I.; underconsumptionism; Weber–Fechner law

Article

Bouniatian was born in Ervian (Armenia) on 22 January 1877, and died on 31 January 1969 in Montmorency (near Paris). He received a D.Sc. from the University of Munich in 1903, and then taught at the University of Moscow and at the Polytechnical Institute of Tiflis (Georgia). From 1916 to 1919 he was manager of the Merchants Bank of Tiflis. After emigrating to France in 1920 as a political refugee, Bouniatian served on the faculty of law of the University of Paris from 1925 to 1940. He later became director of the Office of Armenian Refugees (a public service of the French ministry of foreign affairs) from 1945 to 1952.

Bouniatian's main contribution to economics is contained in his *Studien zur Theorie und Geschichte der Wirtschaftskrisen*, published in two volumes dated 1908. (Bouniatian often pointed out that the book actually came out in October 1907; the date issue was important to his claim that many of his ideas were later incorporated in Albert Aftalion's better-known articles and books. In fact, the list of books received in the February 1908 issue of the *Journal of Political Economy* gives 1907 as the date of publication.) In the first volume Bouniatian put forward a theory of the business cycle based on an original combination of elements from the underconsumption tradition and the then new accelerator concept, plus a novel explanation of changes in the price level. The volume was later revised and translated into Russian (1915) and French (1922; 1930). English expositions can be found in two articles by Bouniatian (1928; 1934). The second volume of the 1908 set is a detailed historical investigation of economic crises in England in the two centuries from 1640 to 1840, which provided the empirical basis for the theoretical volume. It was written between 1899 and 1903, and then submitted as a dissertation to the University of Munich. Bouniatian's business cycle theory attracted some attention at the time (see, for example, W.C. Mitchell, 1913, pp. 9–10; J.M. Keynes, 1930, pp. 143–4) and his books were reviewed in the *Journal of the Royal Statistical Society* (June and September 1908), *American Economic Review* (June 1927, June 1936, December 1959), *Economic Journal* (September 1927, September 1932) and *Journal of Political Economy* (October 1934), among others.

Bouniatian's mix of theory and history in his *Studien* followed the pattern set by Mikhail Tugan-Baranovsky in his influential book about economic crises in England, published in Russian in 1894 and in a revised version in German in 1901. However, Bouniatian rejected the main elements of Tugan-Baranovsky's theory, that is, the compatibility between capital accumulation and decreasing consumption in the long run, and the notion that, in the depression, unused savings take the form of a fund of ‘free capital’ that is invested later in the upward period. It was not difficult for Bouniatian to show that actual saving and investment can never differ, although he did not consistently distinguish between desired and actual saving and investment – nor did Tugan-Baranovsky and most other contemporary economists for that matter. Concerning the first point, Bouniatian, building on J.M. Lauderdale (1804) and A.F. Mummery and J.A. Hobson (1889), carefully developed the view that there is in equilibrium a certain relation between aggregate consumption and the capital stock determined by the choice of production methods, which he called ‘degree of social capitalization’ (‘Grad der gesellschaftlichen Kapitalisierung’). This comes from Bouniatian's argument – against both Tugan-Baranovsky and the classical economists – that productive forces cannot be transferred to the future through the simple accumulation of capital goods, since these can be economically conserved only by being utilized in the process of production and sale of consumption goods.

According to Bouniatian, the evolution of the demand for investment through time is governed primarily by the evolution of consumers’ ‘new requirements’ as determined by population growth, changes in tastes and inventions. However, this cannot be a smooth process because of the characteristics of the saving function on one side and of the production process of capital goods on the other. From the savers’ side, whenever income grows there is a tendency – suggested by economic theory and confirmed by data – to increase the proportion of income saved. This ‘tendency toward excessive accumulation’ means that the demand for consumption goods tends to increase more slowly than production capacity, since saving is a ‘false demand’. Such a tendency is realized due to the existence of a period of time necessary for the production of capital goods, which allows for a temporary separation between investment and consumption decisions and a more than proportional increase of capital goods in relation to a given

intensification of ‘new requirements’, until the processes of production mature and consumers’ good start to pour out. This was an early formulation of an aspect of what would later become known as the acceleration principle. ‘Overcapitalization’ (‘Ueberkapitalisation’, a term apparently coined by Bouniatian) is the main feature of the boom, which is followed by ‘decapitalization’ in the depression period, when overproduction of consumers’ goods brings about a more than proportional fall in the value of capital goods. Equilibrium between production and consumption is restored through falling prices and depreciation of stocks and industrial plant, which transfer part of the capital to the consumption flow. However, equilibrium will not be attained if money wages are rigid downwards, as claimed by Bouniatian in his interpretation of the Great Depression of the early 1930s.

Apart from the saving function and the accelerator, another important element of Bouniatian's framework is his attempted application of the subjective theory of value to explain price level changes and, by that, the possibility of general overproduction. This was developed in detail in his 1927 book, where he used the Weber–Fechner law to generalize the old King's law – that the price of an important good varies inversely in geometrical progression as its quantity varies in arithmetical progression – to the economy as a whole. Bouniatian argued that, instead of the traditional quantity theory of money, price fluctuations should be explained by changes in the ‘absolute social value’ (marginal utility) of both consumption and capital goods, brought about by changes in their quantities throughout the business cycle. Such price changes are accompanied by changes in income distribution and, therefore, in the saving flow. This was used by Bouniatian (1908, vol. 1) to distinguish, for the first time in the literature, between ‘exogenous’ and ‘endogenous’ theories of the business cycle. In the latter, economic crises are explained as an organic part (the upper turning point) of the business cycle, not as accidents of economic history.

See Also

- acceleration principle
- Tugan-Baranovsky, Mikhail Ivanovich

Selected works

1908. *Studien zur Theorie und Geschichte der Wirtschaftskrisen*, 2 vols. Vol. 1: *Wirtschaftskrisen und Ueberkapitalisation – Eine Untersuchung über die Erscheinungsformen und Ursachen der periodischen Wirtschaftskrisen*. Vol. 2: *Geschichte der Handelskrisen in England im Zusammenhang mit der Entwicklung des englischen Wirtschaftslebens 1640–1840*. Munich: E. Reinhardt.

1915. *Economic Crises – Morphology and Theory of Periodic Economic Crises and Theory of the Economic Conjuncture* [in Russian]. Moscow: N.P. Mesnyankin and Co.

1922. *Les crises économiques – essai de morphologie et théorie des crises économiques périodiques et de la conjoncture économique*. Translated from Russian by J. Bernard and revised by the author. Paris: M. Giard.

1927. *La loi de variation de la valeur et les mouvements généraux des prix*. Paris: M. Giard.

1928. The theory of economic cycles based on the tendency towards excessive capitalization. *Review of Economic Statistics* 10, 67–79. Reprinted in *Business Cycle Theory – Selected Texts 1860–1939*, vol. 6, ed. M. Boianovsky. London: Pickering and Chatto, 2005.

1930. *Les crises économiques—essai de morphologie et théorie des crises économiques périodiques et de la conjoncture économique*. 2nd edn. Paris: M. Giard.

1933. *Crédit et conjoncture*. Paris: M. Giard.

1934. Economic depression and its causes. *International Labour Review* 30, 1–22.

1959. *Les fluctuations économiques – recueil d'études*. Paris: Pichon et Durand-Auzias.

1966. *Mes théories économiques et Albert Aftalion*. Paris: Pichon et Durand-Auzias.

Bibliography

Aftalion, A. 1908–9. La réalite des surproductions générales: essai d'une théorie des crises générales périodiques. *Revue d'Économie Politique* 22, 696–706; 23, 81–117, 201–29, 241–59. Trans. R. Leverdier in *Business Cycle Theory – Selected Texts 1860–1939*, vol. 6, ed. M. Boianovsky. London: Pickering and Chatto, 2005.

Keynes, J.M. 1930. *A Treatise on Money*, vol. 1. London: Macmillan.

Lauderdale, J. 1804. *An Inquiry into the Nature and Origin of Public Wealth, and into the Means and Causes of its Increases*. Edinburgh: Archibald Constable.

Mitchell, W. 1913. *Business Cycles*. Berkeley: University of California Press.

Mummery, A. and Hobson, J. 1889. *The Physiology of Industry*. London: J. Murray.

Tugan-Baranovsky, M. 1894. *Studien zur Theorie und Geschichte der Handelskrisen in England*. Jena: Fischer, 1901.

How to cite this article

Boianovsky, Mauro. "Bouniatian, Mentor (1877–1969)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_M000366> doi:10.1057/9780230226203.0151

The New Palgrave Dictionary of Economics Online

Bowley, Arthur Lyon (1869 – 1957)

J.R.N. Stone
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bowley, A. L.; Econometric Society; family budgets; Fisher, I.; ideal index numbers; index numbers; International Statistical Institute; Marshall, A.; mathematical economics; *Palgrave's Dictionary of Political Economy* ; Royal Statistical Society; sampling; statistics and economics

Article

Bowley was born on 6 November 1869 in Bristol, and died on 21 January 1957 at Haslemere. In 1922 he was made a Fellow of the British Academy and knighted in 1950. He was educated at Christ's Hospital from 1879 to 1888, and Trinity College, Cambridge, from 1888 to 1891 (10th Wrangler, 1891). He stayed on another two terms studying physics, chemistry and, under the influence of Alfred Marshall, who remained a lifelong friend, economics. After a period as a schoolmaster, he became lecturer in mathematics, and then professor of mathematics and economics at University College, Reading, from 1900 to 1919. He concurrently taught at the London School of Economics from its inception in 1895, first as lecturer, then reader, then professor, and finally, from 1919, as the first holder of the newly established Chair of Statistics at the University of London, becoming Emeritus Professor on his retirement in 1936.

Among his other activities, he was Acting Director of the Oxford University Institute of Statistics from 1940 to 1944; foundation member in 1933, and then President from 1938 to 1939, of the Econometric Society; President of the Royal Statistical Society from 1938 to 1940, and honorary President of the International Statistical Institute in 1949.

Bowley was an outstanding economic statistician who made substantial contributions to all areas in his field, from the theory of mathematical statistics to the methodology and practice of data collecting. His courses on statistics at the LSE formed the subject matter of two very successful textbooks (Bowley, 1901; 1910). He brought together and set out in a uniform way the developments of mathematical economics from Cournot to Pigou (Bowley, 1924). He wrote a detailed account of Edgeworth's contributions to mathematical statistics (Bowley, 1928). He collaborated with R.G.D. Allen on a masterly study of family budgets which deals with individual variation as well as average behaviour

(Allen and Bowley, 1935).

One of his early interests was the course of wages, on which he wrote several books and over 30 articles, many jointly with G.H. Wood; his first paper on the subject was Bowley (1895) and his first book Bowley (1900). This led him to write extensively on index-numbers of prices and it is interesting that in 1899, on p. 641 of vol. III of *Palgrave's Dictionary of Political Economy*, he gave the index-number formula later to become famous as Irving Fisher's ideal index-number. He followed this work with studies of the national income in Bowley (1919; 1920; 1937) and jointly with J.C. Stamp in Bowley and Stamp (1927).

Bowley was a pioneer in the development of sampling methods and spoke strongly in their favour in his presidential address to the British Association in 1906. In 1912 he carried out a well-designed sample survey of Reading and soon followed this with similar enquiries in Northampton, Warrington, Stanley and Bolton (Bowley and Burnett-Hurst, 1915). A second survey of the same towns was made after the war (Bowley and Hogg, 1925). In the same period he prepared a substantial report on the precision attained in sampling (Bowley, 1926). He played an important role in Llewellyn-Smith's new survey of London life and labour (Bowley, 1930–35).

Selected works

1895. Changes in average wages (nominal and real) in the United Kingdom between 1860 and 1891. *Journal of the Royal Statistical Society* 58, June, 223–78.

1900. *Wages in the United Kingdom in the Nineteenth Century*. Cambridge: Cambridge University Press.

1901. *Elements of Statistics*. London: P.S. King. 6th edn, 1937.

1910. *An Elementary Manual of Statistics*. London: P.S. King. 7th edn, London: Macdonald & Evans, 1951.

1915. (With A.R. Burnett-Hurst.) *Livelihood and Poverty: A Study in the Economic Conditions of Working-Class Households in Northampton, Warrington, Stanley and Reading*. London: G. Bell.

1919. *The Division of the Product of Industry: An Analysis of National Income before the War*. Oxford: Clarendon Press.

1920. *The Change in the Distribution of the National Income, 1880–1913*. Oxford: Clarendon Press.

1924. *The Mathematical Groundwork of Economics*. Oxford: Clarendon Press.

1925. (With M.H. Hogg.) *Has Poverty Diminished?* London: P.S. King.

1926. Measurement of the precision attained in sampling. *Bulletin de l'Institut International de*

Statistique 22, pt I(3), 1–62.

1927. (With J.C. Stamp.) *The National Income 1924*. Oxford: Clarendon Press.

1928. *F.Y. Edgeworth's Contributions to Mathematical Statistics*. London: Royal Statistical Society.

1930–35. Contributions to H. Llewellyn-Smith, *New Survey of London Life and Labour*. 9 vols, London: P.S. King.

1935. (With R.G.D. Allen.) *Family Expenditure*. London: P.S. King.

1937. *Wages and Income in the United Kingdom since 1860*. Cambridge: Cambridge University Press.

Bibliography

Allen, R.G.D. 1968. Bowley, Arthur Lyon. In *International Encyclopedia of the Social Sciences*, vol. 2. New York: Macmillan and Free Press.

Allen, R.G.D. 1971. Bowley, Sir Arthur Lyon. *Dictionary of National Biography, 1951–1960*. Oxford: Oxford University Press.

Allen, R.G.D. and George, R.F. 1957. Professor Sir Arthur Lyon Bowley (with bibliography). *Journal of the Royal Statistical Society*, Series A (General) 120, pt 2, 236–41.

Bowley, Agatha H. 1972. *A Memoir of Professor Sir Arthur Bowley* (1869–1957) *and his Family*. Privately printed.

Darnell, A. 1981. A.L. Bowley, 1869–1957. In *Pioneers of Modern Economics in Britain*, ed. D.P. O'Brien and J.R. Presley. London: Macmillan.

Maunder, W.F. 1977. Sir Arthur Lyon Bowley. In *Studies in the History of Statistics and Probability*, vol. 2, ed. M.G. Kendall and R.L. Plackett. London: Griffin.

How to cite this article

Stone, J. R. N. "Bowley, Arthur Lyon (1869–1957)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000181> doi:10.1057/9780230226203.0152

The New Palgrave Dictionary of Economics Online

Boyd, Walter (1754 – 1837)

David Laidler

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank of England; Boyd, W.; bullionist controversy; convertibility; Sinking Fund

Article

Before the French Revolution, Walter Boyd was engaged as a banker in France, but by the time his firm's property was confiscated by the French government in 1793, he was established in London as the leading member of the firm of Boyd Benfield & Co. At first this London venture was highly successful, and in 1797 Boyd entered Parliament as member for Shaftsbury, then a pocket borough owned by his partner. In this very year, however, Boyd Benfield & Co began to encounter the difficulties which were to culminate in its liquidation in 1800. The basic cause of Boyd's ruin was his having entered into engagements in the expectation that his French property would be restored to him, an expectation that was finally disappointed in September 1797, but the events which precipitated the final collapse of his firm were the government's refusal to employ it as a contractor for the loan of 1799 and the Bank of England's final refusal to grant assistance in early 1800.

When, in 1801, Boyd published his ‘Letter to William Pitt…’ attacking the Bank of England's policies since the suspension of specie convertibility of February 1797, he was hardly a disinterested observer. However, this pamphlet's appearance is widely regarded as marking the beginning of the ‘Bullionist Controversy’, and contains perhaps the first systematic, albeit crude, statement of what came to be known as the Bullionist position. It argued that exchange depreciation and food price increases since 1797 were the result of an overissue of paper money by the Bank of England; that though foreign transfers could depreciate the exchanges this factor had not been important since 1797; and that the Country Bank note issue could not affect prices independently of Bank of England policies.

Boyd's pamphlet drew a number of replies, some, as Fetter (1965) notes, aimed more at Boyd than at his case, but one by Sir Francis Baring (1801) prefigured subsequent anti-bullionist positions. Baring argued (with some justice) that food price behaviour had had more to do with bad harvests than the exchange rate (which had moved much less), and that the exchange rate's fall had been the result of British remittances to Continental allies and not of overexpansionary policy on the part of the Bank of England.

Boyd made no further contributions to wartime debates. After the Peace of Amiens (1802) he visited France, only to be trapped there until 1814 by the renewal of hostilities. Upon his return to England he re-established his fortunes sufficiently to be able to re-enter Parliament in 1823, as member of Lymington, which he represented until 1830. He published two further pamphlets, on the Sinking Fund (1815 and 1828), but neither of these has the historical significance of his 1801 contribution.

See Also

- bullionist controversies (empirical evidence)

Selected works

1801. *Letter to the Right Honourable William Pitt on the Influence of the Stoppage of Issues in Specie at the Bank of England on the Prices of Provisions and Other Commodities*. London. 2nd edn, 1811.

1815. *Reflections on the Financial System of Great Britain, and Particularly on the Sinking Fund*. London. 2nd edn, 1828.

1828. *Observations on Lord Greville's Essay on the Sinking Fund*. London.

Bibliography

Baring, Sir F. 1801. *Observations on the Publication of Walter Boyd*. London.

Fetter, F.W. 1965. *Development of British Monetary Orthodoxy: 1797–1875*. Cambridge, MA: Harvard University Press.

How to cite this article

Laidler, David. "Boyd, Walter (1754–1837)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000184> doi:10.1057/9780230226203.0153

The New Palgrave Dictionary of Economics Online

Bradford, David (1939 – 2005)

Daniel Shaviro

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bradford, D.; consumption taxation; flat tax; intertemporal preferences; local public finance; progressive and regressive taxation; public goods pricing; taxation of corporate profits; taxation of income

Article

David Bradford is best known for his work on fundamental tax reform, although his contributions to public economics were more wide-ranging. His early writings, after he joined the economics department at Princeton University in 1966, largely focused on municipal finance and public goods pricing. His interests took a dramatic turn, however, when he was named Deputy Assistant Secretary for Tax Policy at the US Treasury in 1975, and given lead responsibility for a Treasury study of comprehensive tax reform. The influential *Blueprints for Basic Tax Reform* (1977), which he co-authored with the US Treasury Tax Policy Staff, set forth models for comprehensive income and consumption taxes that remain influential to this day. The *Blueprints* cash flow consumption tax in particular influenced subsequent tax reform thinking by showing how a consumption tax, levied at the individual rather than the business level, could match the progressivity of an income tax and offer self-help income averaging through a mix of ‘prepaid’ and ‘postpaid’ (that is, yield-exempt and deductible) savings accounts.

His experiences at the Treasury made Bradford a lifelong advocate of consumption taxation, based on two main considerations. The first was that he considered it inequitable for people with the same lifetime earnings to face different tax burdens, as they would under an income tax, simply by reason of having different intertemporal consumption preferences. The second was that shifting to a consumption base might permit significant tax simplification, by eliminating the timing issues that bedevil a realization-based income tax. Bradford later developed a second consumption tax prototype, the X-tax, based on the Hall–Rabushka flat tax (Hall and Rabushka, 1995) but modified to permit greater progressivity and to address transition problems, which he recognized could arise not only upon initial enactment but whenever tax rates were changed.

Bradford also helped to pioneer the contemporary understanding that the only theoretical difference between pure income and consumption taxation lies in their treatment of the risk-free return to waiting,

which the former subjects to tax and the latter exempts. In addition, he advanced understanding of the economics of a transition from income to consumption taxation, showing that the ostensibly lump-sum revenue gain resulted from wiping out assets' income tax basis while solemnly pledging never to do so again. Bradford also helped develop the 'new view' of corporate taxation, which shows that a uniform tax on corporate distributions does not distort corporate decisions regarding when to pay out earnings.

See Also

- consumption taxation
- taxation of corporate profits
- taxation of income

Selected works

1970. (With W. Baumol.) Optimal departures from marginal cost pricing. *American Economic Review*60, 265–83.

1977. (With US Treasury Tax Policy Staff.) *Blueprints for Basic Tax Reform*. Arlington, VA: Tax Analysts, 1984.

1981. The incidence and allocation effects of a tax on corporate distributions. *Journal of Public Economics*15, 1–22.

1986. *Untangling the Income Tax*. Cambridge, MA: Harvard University Press.

1996. Consumption taxes: some fundamental transition issues. In *Frontiers of Tax Reform*, ed. M. Boskin. Stanford, CA: Hoover Institution Press.

2004. (With A. Auerbach.) Generalized cash-flow taxation. *Journal of Public Economics*88, 957–80.

Bibliography

Hall, R. and Rabushka, A. 1995. *The Flat Tax*. Stanford, CA: Hoover Institution Press.

How to cite this article

Shaviro, Daniel. "Bradford, David (1939–2005)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000325> doi:10.1057/9780230226203.0154

The New Palgrave Dictionary of Economics Online

Brady, Dorothy Stahl (1903 – 1977)

M. Reid

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Brady, D. S.; distribution of income; survey data

Article

A mathematician and statistician, Dorothy S. Brady combined in her professional life extended periods in both universities and US federal agencies. Most of her empirical work entailed the design and interpretation of survey data on household income and expenditures and critiques of applications of such data.

This began with analysis of data collected in the large 1935–6 survey of incomes and expenditures of rural households which together with its urban counterpart provided the basis for new tests of the validity of Commerce Department estimates of the size and distribution of national income, consumption and savings. At the Bureau of Labor Statistics (1943–8, 1951–6) she assessed consumption and price data in connection with efforts to control inflation, and she developed the statistical design for pricing the city workers' family budget which was used to estimate inter-area differences in the cost of living.

An active participant in the Conference on Income and Wealth of the National Bureau of Economic Research, Brady brought to its sessions a keen awareness of data limitations in the empirical identification of key elements in an analytical structure. Using statistical analysis to randomize effective unidentified factors, she found that the percentage of income saved by families tends to increase systematically with relative position in an income distribution, that the secular increase of income of a population tends to decrease the age at which children leave the family residence, often with financial help from parents, and that such leaving tends to increase the inequality of measured income distribution.

Selected works

1940. (With others.) *Family Income and Expenditures, 5 Regions*, Part 2. Miscellaneous Publication No. 396. Washington, DC: US Department of Agriculture.

1945. *Family Spending and Saving in Wartime*. US Bureau of Labour Statistics, Bulletin No. 872, Appraisal of survey data, p. 45.

1948. (With L.S. Kellog.) The city worker's family budget. *Monthly Labour Review* 66(25), 133–70.

1949. The use of statistical procedures in the derivation of family budgets. *Social Service Review* 23(2), 141–57.

1952. Family savings in relation to changes in the level and distribution of income. In *Studies in Income and Wealth*, vol. 15. New York: NBER.

1956. Family Savings, 1880 to 1950. Part II. In *A Study of Savings in the United States*, Vol. III: *Special studies*, ed. R.W. Goldsmith, D.S. Brady and H. Mendershausen. Princeton: Princeton University Press.

1957. Measurement and interpretation of the income distribution in the United States. *Income and Wealth*, Series VI. London: International Association for Research in Income and Wealth.

1958. Individual incomes and the structure of consumer units. *American Economic Review, Papers and Proceedings* 48, 269–78.

1965. *Age and the Income Distribution*. Research Report No. 8. Washington, DC: US Social Security Administration.

1966. Price deflators for final product estimates. In *Studies in Income and Wealth*, Vol. 30: *Output, Employment, and Productivity in the United States After 1800*. Conference on Research in Income and Wealth. New York: NBER.

1971. The statistical approach: the input–output system. In *Approaches to American Economic History*, ed. G.R. Taylor and L.F. Ellsworth. Charlottesville: University Press of Virginia.

How to cite this article

Reid, M. "Brady, Dorothy Stahl (1903–1977)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000185> doi:10.1057/9780230226203.0155

The New Palgrave Dictionary of Economics Online

brain drain

Frédéric Docquier and Hillel Rapoport
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘brain drain’ designates the international transfer of human resources and mainly applies to the migration of relatively highly educated individuals from developing to developed countries. While the brain drain has long been viewed as detrimental to poor countries’ growth potential, recent economic research emphasizes a number of positive feedback effects arising from skilled migrants’ participation in business networks, and suggests that under certain conditions the prospect of migration can positively affect human capital accumulation in the source countries.

Keywords

altruism; brain drain; diaspora networks; education; externality; foreign direct investment; human capital; information costs; international migration; remittances; transaction costs; transfer of technology; wage differentials

Article

The term ‘brain drain’ designates the international transfer of resources in the form of human capital and mainly applies to the migration of relatively highly educated individuals from developing to developed countries. In the non-academic literature, the term is generally used in a narrower sense and relates more specifically to the migration of engineers, physicians, scientists and other very highly skilled professionals with university training. The brain drain has long been viewed as a serious constraint on poor countries’ development and is also a matter of concern for many European countries such as the UK, Germany or France, which have recently seen a significant fraction of their talented workforce emigrate abroad. Recent comparative data reveal that by 2000 there were 20 million highly skilled immigrants (that is, foreign-born workers with a tertiary education) living in the Organisation for Economic Co-operation and Development (OECD) area, a 70 per cent increase in ten years against only a 30 per cent increase for unskilled immigrants. Skilled migrants now represent one-third of total immigration to the OECD countries, and most of this increase is due to immigration from developing and transition countries. The causes of this growing brain drain are well known. On the supply side, the globalization of the world economy has strengthened the tendency for human capital to agglomerate where it is already abundant and has contributed to increase positive self-selection among migrants. And on the demand side, host countries have gradually introduced quality-selective immigration policies and are now engaged in what appears as an international competition to attract global talents.

How big is the brain drain?

Extending and updating the work of Carrington and Detragiache (1998), Docquier and Marfouk (2006) recently collected OECD immigration data to construct estimates of emigration rates by educational attainment (primary, secondary and tertiary schooling) for all countries in 1990 and 2000. Their figures for the highest education level may be taken as a brain drain measure. This may seem too broad a definition for the most advanced countries where the highly educated typically represent about a third of the total workforce but seems appropriate in the case of developing countries, where this share is on average just about five per cent. Note that due to data constraints, South–South migration is not taken into account in the Docquier and Marfouk (2006) data-set; this can lead to potential underestimation of the brain drain for some countries for which other developing countries are

significant destinations. On the other hand, the very definition of immigrants as foreign-born workers does not account for whether education has been acquired in the home or in the host country; this can lead to potential overestimation of the brain drain as well as to possible spurious cross-country variation in skilled emigration rates (Rosenzweig, 2005). In an attempt to solve this problem, Beine, Docquier and Rapoport (2007a) used age of entry as a proxy for where education has been acquired and proposed alternative brain drain estimates excluding people who immigrated before a given age (12, 18 and 22); their results show country rankings by degree of brain drain intensity only mildly affected by the correction and extremely high correlations between corrected and uncorrected estimates.

Keeping this in mind, one can use a simple multiplicative decomposition of the brain drain: the skilled emigration rate is to equal to the average emigration rate times the schooling gap. The average emigration rate is the ratio of emigrants to natives (residents plus emigrants) and reflects the sending country's openness to emigration. The schooling gap is the ratio of skilled to average emigration rate which, by definition, is also the ratio of the proportion of educated among emigrants to the corresponding proportion among natives.

Table 1 summarizes the data for different country groups in 2000. Countries are grouped according to demographic size, income per capita (under the World Bank classification), and region. Unsurprisingly, we observe a decreasing relationship between emigration rates and country size, with average skilled emigration rates about seven times higher in small countries than in large countries. Regarding income groups, the highest emigration rates are observed in middle-income countries, where people have both the incentives and means to emigrate. Regarding the regional distribution of the brain drain, the most affected regions are the Caribbean and the Pacific islands, sub-Saharan Africa (where the schooling gap is exceptionally high), and Central America.

Data by country group in 2000

	Skilled emigration rate (%)	Average emigration rate (%)	Schooling gap
By population size (millions)			
Large countries (>25)	4.1	1.3	3.144
Upper-middle (>10–25)	8.8	3.1	2.839
Lower-middle (>2.5–10)	13.5	5.8	2.338
Small countries (<2.5)	27.5	10.3	2.666
By income group			
High-income countries	3.5	2.8	1.238
Upper-middle income countries	7.9	4.2	1.867
Lower-middle income countries	7.6	3.2	2.383
Low-income countries	6.1	0.5	12.120
By region			
AMERICA	3.3	3.3	1.002
•USA and Canada	0.9	0.8	1.127
•Caribbean	42.8	15.3	2.807
•Central America	16.9	11.9	1.418
•South America	5.1	1.6	3.219
EUROPE	7.0	4.1	1.717
•Eastern Europe	4.3	2.2	1.930
•Rest of Europe	8.6	5.2	1.637
••incl. EU15	8.1	4.8	1.685
AFRICA	10.4	1.5	7.031
•Northern Africa	7.3	2.9	2.489
•Sub-Saharan Africa	13.1	1.0	13.287
ASIA	5.5	0.8	7.123
•Eastern Asia	3.9	0.5	8.544

•South-central Asia	5.3	0.5	10.030
•South-eastern Asia	9.8	1.6	5.980
•Near and Middle East	6.9	3.5	1.937
OCEANIA	6.8	4.3	1.578
•Australia and New Zealand	5.4	3.7	1.479
•Other Pacific countries	48.7	7.6	6.391

Source: Docquier and Marfouk (2006).

It is clear that the *magnitude* of the brain drain has increased dramatically since 1980. However, in terms of *intensity* (or emigration rates), the picture is less clear as one must factor in the general progress in educational attainments observed across the world. Figure 1 presents skilled emigration rates by region computed by Defoort (2006) using a long-run perspective. Focusing on the six major destination countries (USA, Canada, Australia, Germany, UK and France), Defoort computed skilled emigration rates from 1975 to 2000 (one observation every five years). One can see that some regions experienced an increase in the intensity of the brain drain (especially Central America and sub-Saharan Africa) while significant decreases were observed in others (notably the Middle East and Northern Africa).

Figure 1

Long-run trends in skilled emigration, 1975–2000. *Source*: Defort (2006).

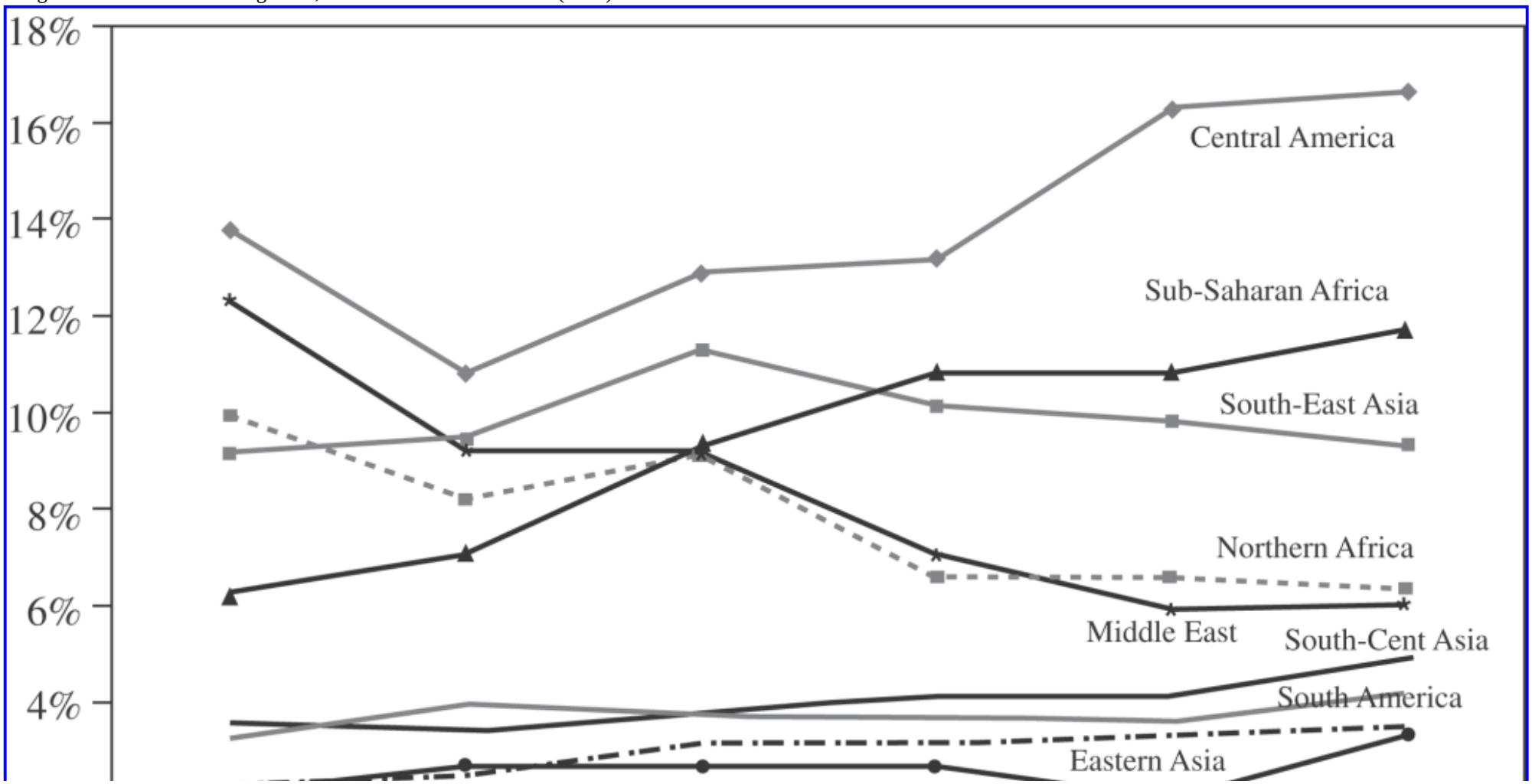

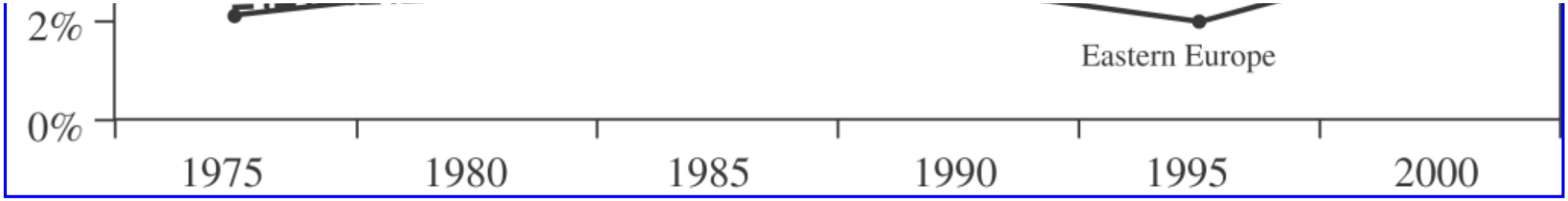

From brain drain to brain gain?

It is certainly a good thing for rich countries to host a skilled and talented workforce, and the move is also worthwhile (at least *ex ante*) from the perspective of the individual migrant. However, the social return to human capital is likely to exceed its private return given the many fiscal, technological, intra- and intergenerational (or Lucas-type) externalities involved. This externality argument is central in the early brain drain economic literature (Bhagwati and Hamada, 1974), which emphasized that the brain drain entails significant losses for those left behind and contributes to increased inequality at the world level. Another negative aspect of the brain drain is that it can induce shortages of manpower in certain activities, for example when engineers or health professionals emigrate in disproportionately large numbers, thus undermining the ability of the origin country to adopt new technologies or deal with health crises. This can be reinforced by governments distorting the provision of public education away from general (portable) skills when graduates leave the country, with the country ending up educating too few nurses, doctors or engineers, and too many lawyers (Poutvaara, 2004). The argument, however, can be reversed, since the prospect for migration may create a bias in the opposite direction (see Lucas, 2005, for an illuminating analysis of the Philippines higher-education market).
The prospect of migration can also impact on the very decision as to whether to study. When education is a passport to emigration, migration prospects create additional incentives to invest in human capital. If migration is probabilistic in that people are uncertain about their chances of future migration when they make education decisions, then the incentive effect just described may more than compensate the brain drain effect, resulting in a higher level of human capital in the source country. As demonstrated in a series of recent papers (for example, Mountford, 1997; Beine, Docquier and Rapoport, 2001), such a positive outcome is theoretically more likely when inter-country wage differentials are large enough to generate a high incentive effect and skilled emigration rates are sufficiently low. These theories have been confirmed empirically by Beine, Docquier and Rapoport (2007b), who found a positive and significant effect of migration prospects on human capital formation in a cross-section of 127 developing countries. From the latters' perspective, however, what matters is not how many of their native-born engage in higher education, but how many remain at home. To estimate the net effects country by country, Beine, Docquier and Rapoport (2007b) used counterfactual macro-simulations and found that countries combining relatively low levels of human capital and low skilled emigration rates are likely to experience a net gain. Their results show a positive effect on aggregate, but with more losers (which tend to lose a lot in relative terms) than winners. The situation of many small African and Central American countries appears extremely worrisome while the main globalizers (for example, India, China) all register moderate gains.

Feedback effects

Remittances

The literature on migrants' remittances shows that the two main motivations to remit are altruism, on the one hand, and exchange, on the other hand (Rapoport and Docquier, 2006). Altruism is primarily directed towards the immediate family, while remittances motivated by exchange pay for services such as care of the migrant's assets or relatives at home. Exchange-motivated transfers are typically observed in case of a temporary migration and signal the migrants' intention to return. It is therefore a priori unclear whether educated migrants remit more than their uneducated compatriots; the former may remit more to meet their implicit commitment to reimburse the family for funding of education investments (and, in addition, they have a higher income potential), but on the other hand, they tend to emigrate with their families, and on a more permanent basis. Indeed, at an aggregate level, Faini (2006) finds that brain drain migration (as measured by the proportion of skilled among emigrants) is associated with lower remittance inflows.

Return migration and brain circulation

Return migration is rare among the highly educated unless sustained growth precedes return. For example, less than one-fifth of Taiwanese and Korean Ph.D. students who graduated from US universities in the 1970s in the fields of science and engineering returned to Taiwan or Korea, a proportion that rose to two-thirds in the course of the 1990s, after two

decades of impressive growth in these countries. The figures for Chinese and Indian Ph.D. students graduating from US universities in the same fields during the 1990s are similar to those for Taiwan or Korea in the 1980s (OECD, 2002). These numbers suggest that return skilled migration is more a consequence than a trigger of growth. On a more reduced scale, however, there are many case-studies showing clear signs of brain circulation. For example, a recent survey conducted among 225 Indian software firms concluded that 30–40 per cent of the higher-level employees had previous work experience in similar occupations in a developed country (Commander et al., 2004).

Diaspora externalities

A large sociological literature emphasizes the potential for skilled migrants to reduce transaction and other types of information costs and thus facilitate trade, foreign direct investment (FDI) flows and technology transfers between their host and home countries. This has first been confirmed in the field of international trade (Gould, 1994; Head and Ries, 1998; Rauch and Casella, 2003). Regarding FDI, Kugler and Rapoport (2007) used US data on immigration and FDI outflows and found that past skilled immigration significantly increases a country's chances of attracting FDI in the subsequent period. These results complement recent case studies of the software industry showing that skilled migrants take an active part in the creation of business networks that lead to FDI deployment in their home country (Arora and Gambardella, 2005).

Conclusion

The number of skilled migrants from poor to rich countries has increased dramatically since the 1970s. In the face of rising wage differentials and of diverging demographic structures between rich and poor countries, this tendency is likely to be confirmed in the future. While the brain drain has long been viewed as detrimental to poor countries' growth potential, recent economic research has emphasized that, alongside positive feedback effects arising from skilled migrants' participation in business networks, one also has to consider the effect of migration prospects on human capital-building in source countries. This new literature suggests that a limited degree of skilled emigration could be beneficial for growth and development. Empirical research shows that this is indeed the case for a limited number of large, intermediate-income developing countries. For the vast majority of poor and small developing countries, however, current skilled emigration rates are most certainly well beyond any sustainable threshold level of brain drain.

Bibliography

Arora, A. and Gambardella, A. 2005. *From Underdogs to Tigers: The Rise and Growth of the Software Industry in Brazil, China, India, Ireland, and Israel*. Oxford and New York: Oxford University Press.

Beine, M., Docquier, F. and Rapoport, H. 2001. Brain drain and economic growth: theory and evidence. *Journal of Development Economics* 64, 275–89.

Beine, M., Docquier, F. and Rapoport, H. 2007a. Measuring international skilled migration: a new database controlling for age of entry. *World Bank Economic Review* (forthcoming).

Beine, M., Docquier, F. and Rapoport, H. 2007b. Brain drain and human capital formation in developing countries: winners and losers. *Economic Journal* (forthcoming).

Bhagwati, J.N. and Hamada, K. 1974. The brain drain: international integration of markets for professionals and unemployment. *Journal of Development Economics* 1, 19–42.

Carrington, W.J. and Detragiache, E. 1998. How big is the brain drain? Working Paper No. 98–102. Washington, DC: International Monetary Fund.

Commander, S., Chanda, R., Kangasniemi, M. and Winters, L.A. 2004. Must skilled migration be a brain drain? Evidence from the Indian software industry. Discussion Paper No. 1422, Institute for the Study of Labor (IZA), Bonn.

Defoort, C. 2006. Tendances de long terme en migrations internationales: analyse à partir de 6 pays receveurs. Mimeo, EQUIPPE, Universités de Lille, and IRES, Université Catholique de Louvain.

Docquier, F. and Marfouk, A. 2006. International migration by educational attainment (1990–2000). In *International Migration, Remittances and the Brain Drain*, ed. C. Ozden and

M. Schiff. Basingstoke: Palgrave Macmillan.

Faini, R. 2006. Remittances and the brain drain. Discussion Paper No. 2155, Institute for the Study of Labor (IZA), Bonn.

Gould, D.M. 1994. Immigrant links to the home country: empirical implications for U.S. bilateral trade flows. *Review of Economics and Statistics* 76, 302–16.

Head, K. and Ries, J. 1998. Immigration and trade creation: econometric evidence from Canada. *Canadian Journal of Economics* 31, 47–62.

Kugler, M. and Rapoport, H. 2007. International labor and capital flows: complements or substitutes? *Economics Letters* 94, 155–62.

Lucas, R.E.B. 2005. *International Migration and Economic Development: Lessons from Low-Income Countries*. Cheltenham: Edward Elgar.

Mountford, A. 1997. Can a brain drain be good for growth in the source economy? *Journal of Development Economics* 53, 287–303.

OECD (Organisation for Economic Co-operation and Development). 2002. *Trends in International Migration*. Paris: OECD Editions.

Poutvaara, P. 2004. Public education in an integrated Europe: studying to migrate and teaching to stay. Working Paper No. 1369, CESifo, Munich.

Rapoport, H. and Docquier, F. 2006. The economics of migrants' remittances. In *Handbook of the Economics of Giving, Altruism and Reciprocity*, vol. 2, ed. S.-C. Kolm and J. Mercier Ythier. Amsterdam: North-Holland.

Rauch, J.E. and Casella, A. 2003. Overcoming informational barriers to international resource allocation: prices and ties. *Economic Journal* 113, 21–42.

Rosenzweig, M.R. 2005. Consequences of migration for developing countries. Paper prepared for the UN conference on international migration and development, Population Division.

How to cite this article

Docquier, Frédéric and Hillel Rapoport. "brain drain." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000186> doi:10.1057/9780230226203.0156

The New Palgrave Dictionary of Economics Online

Braudel, Fernand (1902 – 1985)

R. Forster

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Annales School; Braudel, F.; demography; geography; technology

Article

One of the foremost social and economic historians of the 20th century, Fernand Braudel combined a perceptive grasp of historical interconnections, an exceptional skill of synthesis and an evocative, even ‘poetic’ style. Perception, scope and style were brought to successful fruition in Braudel's *La Méditerranée et le monde méditerranéen à l’époque de Phillipe II* (1949), which became a classic in historical literature and a model for a major school of French history known as the *Annales*. In this seminal volume and in many methodological articles that followed, Braudel proposed a triple notion of historical time – the long run (*longue durée*) over a millennium, trends (*conjonctures*) of a generation or more, and events (*événements*). According to Braudel, each of these notions or blocks of time involved unique historical problems, appropriate source materials, and even special approaches employing social-science disciplines neighbouring to history. Braudel's model emphasized the ‘constraints’ of human endeavour rather than the ‘permissive’ factors that had been so much a part of Whig history as practised by most early 20th-century historians. These constraints were imposed by geography, climate and soils, by demographic pressure, and by a static social structure held together by the bonds of custom. Braudel likened this ‘structure’ to a glacier or to the sea depths, imparting both a physical metaphor and a sense of timelessness or immobility. His second temporal level, the *conjoncture*, made some room for change as new technologies, new forms of economic organization (especially capitalism), and subtle shifts in social relations and customs altered the ‘structure’. Braudel likened these changes – he preferred the term ‘mutations’ – to the sea tides. Finally the ‘event’ was a kind of surface noise, an indication perhaps of deeper sea changes, but in itself of little significance for the historian. He likened these events to whitecaps on the vast ocean.

In addition to his emphasis on constraints and the obligation of historians to understand their deterministic effects on human behaviour, Braudel also stressed the cyclical nature of most of history – ‘le temps, quasi immobile, fait de repétitions, de retours insistants, de cycles sans cesse recommencés’.

There was about Braudel a strong sense of romantic conservatism that challenged Marxist and Whig historian alike. Braudel imparted to the Annales School a preference for metaphors taken from biology and anthropology (interconnection, *liens*, mutations, *glissements*) instead of the vocabulary, and indeed the goals, of physics or economics (parsimonious cause, leanness of argument, elegance of formula or theory). It is also clear that for Braudel geography and demography were basic objects of study, that technology and economic and social organization were important, but that political history, biography and the history of formal ideas were secondary and even trivial historical pursuits. In a direct attack on the kind of history taught at the Sorbonne, Braudel insisted that ‘events’ tell us little about the deeper and interlocking structures and their subtle mutations. Indeed, such surface history may suggest a misguided ‘voluntarism’ in human history. With such a perspective, it is understandable that Braudel was most comfortable in the thousands of years of pre-industrial history. The more recent 19th century and its urban-industrial dynamism were unsettling to his outlook, his methodology and even to his aesthetic sense. But, like a cultural anthropologist, Braudel never ceased to stress the fact that most of world history was pre-industrial.

Although Braudel was interested in quantification, he was never a model-builder, and in fact he used numbers illustratively rather than systematically. He had much to do with the Annales-style deployment of an array of graphic techniques – often very artfully designed – to demonstrate proportions and relationships, but as a descriptive technique in which the reader had to access the results by eye. Braudel did not use statistical measures, much less economic theory, perhaps because he considered them too abstract, and a threat to the living texture of social history that was his main concern. In the 1970s, like much of the Annales School, Braudel moved further towards cultural anthropology as reflected in his notion of ‘day-to-dayness’ (*la vie quotidienne*), in the cultural determinants of economic and social behaviour, in the values and attitudes (mentalities) of social groups, and in the *gestes* and *code* of an entire society or even a ‘civilization’. These features were already present in the *Méditerranée*, but they became even more pronounced in his more recent *Civilisation matérielle et capitalisme (XV-XVIIIe siècle*) (1967–79).

Fernand Braudel was also director of the Maison des Sciences de l'Homme in Paris, professor at the Ecole Pratique des Hautes Etudes and at the Collège de France, and co-editor of the *Annales: ESC*, one of the most prestigious journals of social and economic history in the Western world today. Braudel's seminal writings, his provocative teaching, his administrative and editorial talents, and, not least, his powerful personality made him an ‘animateur’ of the ‘School of the Annales’ for more than 30 years. Yet his work stands on its own as an appeal to approach history in its widest scope in time and place (*histoire totale*), in alliance with neighbouring disciplines, and presented with that special verve we call ‘Braudelian’.

Selected works

1949, 1966. *La Méditerranée et le monde méditerranéen à l’époque de Phillipe II*. Paris: Armand Colin. Trans. by S. Reynolds as *The Mediterranean and the Mediterranean World in the Age of Philip II*, London; Collins; New York: Harper & Row, 1972–3.

1967, 1979. *Civilization matérielle et capitalisme, XV–XVIIIe siècle*. 1967 edn, 1 vol.; 1979 edn, 3 vols;

Paris: Armand Colin. Trans. of 1st edn by M. Kochan as *Capitalism and Material Life, 1400–1800*, London: Weidenfeld & Nicolson, 1973. Trans. of 2nd edn by M. Kochan and S. Reynolds as *Civilization and Capitalism, 15th–18th Century*, London: Collins; New York: Harper & Row, 3 vols, 1981, 1982, 1984.

1969. *Ecrits sur l'histoire*. Paris: Flammarion.

1973. *Mélanges en l'honneur de F. Braudel*. 2 vols. Toulouse: Edouard Privat.

How to cite this article

Forster, R. "Braudel, Fernand (1902–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000189> doi:10.1057/9780230226203.0157

The New Palgrave Dictionary of Economics Online

Braverman, Harry (1920 – 1976)

David M. Gordon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Braverman, H.; capitalism; deskilling; division of labour; exploitation; Heilbroner, R.; labour power; Marx's analysis of capitalist production; reskilling; supervision

Article

Harry Braverman was born in 1920 in New York City and died on 2 August 1976 in Honesdale, Pennsylvania.

Born into a working-class family, he was able to spend only one year in college before financial problems forced him out of Brooklyn College and into the Brooklyn Navy Yard. He worked there for eight years primarily as a coppersmith and then moved around the United States, working in the steel industry and in a variety of skilled trades. He became deeply involved in the trade union and socialist political movements. He helped found *The American Socialist* in 1954 and worked as its co-editor for five years. After the journal ceased publication for practical reasons, he moved into publishing, working first at Grove Press as an editor and eventually as vice-president and general business manager. In 1967 he became Managing Director of Monthly Review Press, where he worked until his death.

Braverman is best known for his classic study of the labour process under capitalism, *Labor and Monopoly Capital* (1974), awarded the 1974 C. Wright Mills Award. ‘Until the appearance of Harry Braverman's remarkable book’, Robert L. Heilbroner wrote in the *New York Review of Books*, ‘there has been no broad view of the labour process as a whole…’ The book was all the more remarkable because of the void it filled in the Marxian analytic tradition – a literature ostensibly grounded in the analysis of the structural effects of class conflict but persistently reticent about the actual structure and experience of work in capitalist production.

Labour and Monopoly Capital advances three principal hypotheses about the labour process in capitalist societies.

First, Braverman helps formalize and extend Marx's resonant analysis, in Volume I of *Capital*, of the distinction between labour and labour power. Braverman highlights the essential importance and persistence of managerial efforts to gain increasing control over the labour process in order to rationalize

– to render more predictable – the extraction of labour activity from productive employees.
Second, Braverman argues that such managerial efforts lead inevitably to the homogenization of work tasks and the reduction of skill required in productive jobs. He concludes (p. 83) that ‘this might even be called the general law of the capitalist division of labor. It is not the sole force acting upon the organization of work, but it is certainly the most powerful and general.’
Third, as a corollary of the second hypothesis, Braverman argues both analytically and with rich empirical detail that this ‘general law of the capitalist division of labour’ applies just as clearly to later stages of capitalist development, with their proliferation of office jobs and white collars, as to the earlier stages of competitive capitalism and largely industrial work.
The first analytic strand of Braverman's work was both seminal and crucial in helping foster a renaissance of Marxian analyses of the labour process. The second and third hypotheses have proved more controversial. There are two grounds for concern. Braverman's analysis tends to reduce the character of the labour process to essentially one dimension – the level of skill required and control permitted by embodied skills – and therefore unnecessarily compresses the *many* essential dimensions of worker activity and effectiveness in production to a single monotonic index. At the same time, there is good reason for worrying about the simplicity of Braverman's argument of historically irreversible ‘deskilling’ for all segments of the productive working class; it is quite plausible to hypothesize that for some labour segments in recent phases of capitalist development there has been a ‘reskilling’, as many have since called it, which has not in any way liberated these workers from capitalist exploitation or intensive managerial supervision.

Selected works

1974. *Labor and Monopoly Capital*. New York: Monthly Review Press.

1976. Two comments. Special issue on ‘Technology, the Labor Process, and the Working Class’. *Monthly Review* 28(3), 119–24.

How to cite this article

Gordon, David M. "Braverman, Harry (1920–1976)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000190> doi:10.1057/9780230226203.0158

The New Palgrave Dictionary of Economics Online

Brentano, Lujo (Ludwig Josef) (1844 – 1931)

B. Schefold

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Brentano, L.; economic history; Engel, E.; Great Depression; Malthus's theory of population; Marx, K. H.; profit sharing; subjective theory of value; tariffs; trade unions

Article

Brentano was born in Aschaffenburg (Germany) into an old patrician family. Clemens Brentano, the poet, was his uncle; Bettina von Arnim, the writer, his aunt; and Franz Brentano, the philosopher, his brother. He was brought up in an atmosphere dominated by Catholicism (which he was later to abandon after the declaration of papal infallibility) and was particularly influenced by the anti-Prussian tradition of southern Germany. He studied law and economics in Heidelberg and Göttingen. From 1871 he taught political economy as professor in Berlin, Breslau, Strassburg, Vienna, Leipzig and Munich.

A decisive point for his later career was his participation in the Statistical Seminar connected with the Prussian Statistical Office. Its director was Ernst Engel (originator of Engel's law), whose strong interest in the social conditions of the working classes was to have a lasting influence on Brentano. Engel advocated profit-sharing schemes as a means to the solution of the social question. In 1868 Brentano accompanied him on a visit to England, where they studied the effects of such measures. His experiences in England convinced Brentano of the inadequacy of profit-sharing for the reform of capitalism, but suggested another approach, which was to remain the main topic of Brentano's intellectual work: the improvement of the worker's position in the labour market through the establishment of trade unions.

While the individual worker was forced to sell his labour power under any conditions, this would not be the case for an organized coalition of workers. Such a coalition would enable them to become as free and independent as the sellers of other commodities and would allow for an effective control of the labour supply (1871–2, vol. 2; 1877, ch. 2). It was Brentano's deep conviction that trade unions were the only means to secure an adequate participation of the working classes in the general increase of wealth. He was especially interested in the history of the trade unions, which he traced back to the medieval guilds (1871–2, vol. 1). Especially interesting – particularly for the current debate – was his discussion

of positive productivity effects of labour time reductions (1876).

He regarded the introduction of a general social security system as another important step for the reform of capitalism. He also favoured the cartelization of Germany industry. It was characteristic of him that he always intended to solve the social question within the framework of a capitalist economic system. He therefore rejected Marx and the Social Democrats of 19th-century Germany. Brentano emphasized that unequal conditions of material existence were absolutely necessary for the further cultural advancement of mankind (1877, pp. 303–4).

His concern for the social question shaped Brentano's attitude towards the classical economists: he opposed the classical notion of an abstract profit-maximizing individual as the central axiom of political economy, and found this particularly inadequate to describe working-class behaviour and the labour market (1923, ch. 1). It is in this context that his preoccupation with economic history (1916; 1927–9) must be seen. He intended to show that the relations between man and the economic system were changing through history, and that the individual of classical economics was not the starting-point, but the result of economic development (1927–9, vol. 1, pp. iii–iv).

Further fields of interest were Malthus's theory of population development (1924), the theory of value (where he favoured the subjective theory of value; 1924), the German corn tariffs (which he opposed), and different forms of the law of estate.

Throughout his life Brentano remained an open-minded and enlightened liberal of whom an English trade union leader once said: ‘He was our friend before it was fashionable to be our friend.’ Brentano was a founding member of the Verein für Socialpolitik, which he left in 1929, when he thought that it had become reactionary. He opposed Bismarck in the Kaiserreich, the extreme German annexationists during the First World War – although himself favouring limited territorial expansion – and the Socialist Revolutionaries in the post-war period. The republican government considered his appointment as first German post-war ambassador to Washington, but because of his advanced age he declined.

During the Weimar Republic Brentano was still concerned with social policy, mainly with the struggle for the eight-hour working day. He deplored the harsh austerity policy during the Great Depression. His memoirs, written in 1930, ended: ‘I do not understand this policy. Do they want a social revolution?’ (1931, p. 404).

Selected works

1870. *On the History and Development of Guilds, and the Origin of Trade Unions*. London: Trübner.

1871–2. *Die Arbeitergilden in der Gegenwart*. Vol. 1: *Zur Geschichte der englischen Gewerkvereine*, Leipzig: Duncker & Humblot, 1871; Vol. 2: *Zur Kritik der englischen Gewerkvereine*, Leipzig: Duncker & Humblot, 1872.

1876. *Über das Verhältnis von Arbeitslohn und Arbeitszeit zur Arbeitsleistung*. Leipzig: Duncker & Humblot. Trans. as *Hours and Wages in* Relation to *Production*, New York: Scribner; London: Sonnenschein, 1894.

1877. *Das Arbeitsverhältnis gemäss dem heutigen Recht*. Leipzig: Duncker & Humblot. Trans. as *The*

Relation of Labour to the Law of To-day, New York: Putnam, 1898.

1889. *Über die Ursachen der heutigen socialen Noth*. Leipzig: Duncker & Humblot.

1916. *Die Anfänge des modernen Kapitalismus*. Munich: Königlich Bayerische Akademie der Wissenschaften.

1923. *Der wirtschaftende Mensch in der Geschichte*. Leipzig: Felix Meiner.

1924. *Konkrete Grundbedingungen der Volkswirtschaft*. Leipzig: Felix Meiner.

1927–9. *Eine Geschichte der wirtschaftlichen Entwicklung Englands*. 3 Vols. Jena: Gustav Fischer.

1931. *Mein Leben im Kampf um die soziale Entwicklung Deutschlands*. Jena: Eugen Diederichs.

How to cite this article

Schefold, B. "Brentano, Lujo (Ludwig Josef) (1844–1931)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000195>
doi:10.1057/9780230226203.0159

The New Palgrave Dictionary of Economics Online

Bresciani-Turroni, Costantino (1882 – 1963)

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bortkiewicz, L. von; Bresciasni-Turroni, C.; *deflation* ; forecasting; foreign exchange equilibrium; Galiani, F.; German hyperinflation; Great Depression; Loria, A.; Pantaleoni, M.; Pareto's law of income distribution; productivity theory of interest; quantity theory of money; Ricca-Salerno, G.; Robinson, J. V.; Schmoller, G. von; speculation; stabilization policy; Wage indexation; Wagner, A.

Article

The last great exponent of old-time liberalism in Italian economics, Bresciani was an Italian counterpart of such distinguished libertarians as Robbins, Hayek or Friedman, a bit more moderate, perhaps, in his views and with a quantitative bent at least equal to Friedman's. Bresciani was born in Verona and his teachers in his homeland included Ricca-Salerno and Loria. After the completion of his studies at a number of universities in Italy, he went to the University of Berlin, at that time at the height of its prestige, to study with historical economists such as Adolf Wagner and Gustav Schmoller, and with L. von Bortkiewicz, the mathematical statistician and pioneer in Marxian econometrics.

Amidst the push and pull of these intellectual influences, Bresciani preserved an admirable independence of mind. Loria did not convert him to socialism and Schmoller did not turn him into an historical economist. More influenced by Pareto and Pantaleoni than by his great teachers, he became, first of all, an economic theorist, but again not a pure one but one looking for statistical verifications of theoretical propositions.

In his writings he would give a respectful hearing to the views of the classics and provide copious references to modern authorities, foreign languages and mathematical modes of expression constituting no barriers. As an Italian and libertarian, he was especially fond of citing Galiani. After the publication of Keynes's *General Theory* in 1936, Bresciani, like other contemporary economists, had to come to terms with the new economics. Again he showed his independence by continuing to adhere to such established doctrines as the quantity theory of money and the productivity theory of interest. This attitude, together with his insistence on the limitations rather than opportunities of public policies, gave an old-fashioned flavour to his later writings, published, as they were, at a time when Keynes's influence

reached its peak.

Bresciani's teaching career, which included chairs in statistics, led him eventually to the University of Milan (1926–57), but his work there was interrupted by various other activities. During the 1920s he served as an adviser to the Berlin office of the Allied Reparations Commission, and from 1927 to 1940 he lectured at the newly established Egyptian University of Cairo. This multiplication of jobs again confirmed his penchant for independence and gave him the opportunity to absent himself from fascist Italy. After the Second World War he served the new Republic of Italy as president of an important bank and for a brief period also as minister of foreign trade. In this capacity he again demonstrated his independence, this time from ideological preferences, by sponsoring a government organization for export credit and insurance.

As a writer Bresciani started out, at age 22, with a critical review of Pareto's law of income distribution, a subject to which he returned later more than once. Much of his work was devoted to the theory of prices, domestic and international, present and future, as well as the relation between prices and interest. Among other topics that he investigated were the influence of speculation on prices, which he recognized as not always beneficial, economic forecasting, the inductive verification of the theory of international payments, and the relation between the harvest and the price of cotton in Egypt. Late in life he wrote a number of broad syntheses of economics, including a two-volume *Corso* that went into many editions.

Bresciani's masterpiece, and the work for which he is best known, is *The Economics of Inflation*, published originally in Italian in 1931 and in a revised English translation in 1937. The Italian title of the book – *Le vicende del marco tedesco*, or the vicissitudes of the German mark – conveys the substance of the book better than the title of the English translation, which claims a level of abstraction far higher than that embodied in the work, and, correspondingly, a much wider applicability of the content. The subtitle of the English translation is also carelessly worded. The subject of the work is the great German inflation after the First World War, when prices had risen to astronomical heights and $1 in the end purchased 42 marks followed by 11 zeros. At that time this was considered a record, but the Hungarian inflation after the Second World War surpassed it, with the dollar then buying 145 pengö followed by 27 zeros.

Bresciani's book has been the standard work on the subject ever since. What was open to debate was never the completeness or reliability of the material that he presented but his interpretation. German students of the matter tended to adhere to the view that the rise in prices reflected the unfavourable rate of exchange, which in turn was ascribed, at least in part, to the burden of reparation payments that the Germans were eager to demonstrate as outrageously unreasonable. Bresciani opposed this interpretation. His principal argument was that foreign exchanges, by means of well-known mechanisms, will never fail to reach an equilibrium if only the external value of the currency falls deeply enough. Bresciani, instead of putting the blame on the foreign exchanges, placed it firmly on the German authorities which pursued policies of fiscal irresponsibility and unrestrained monetary expansion. Bresciani also discussed still other interpretations – conspiratorial or scandal theories – but found them unconvincing. One variant of these made the industrialists, who gained so much from the galloping inflation, responsible for it. Another one put the onus on the German authorities’ desire to prove the impossibility of reparation payments. It may be of some interest that the second variant of the scandal theory would constitute a corollary of the policy of *deflation* which Chancellor Brüning adopted a few years later during the Great Depression, a policy instrumental in helping Germany to rid herself of reparation payments.

Critics of the work brought still other points of view before the reader. Joan Robinson, to give an example, stressed the role of ever-rising money wages that became indexed and subject to automatic increases. This would seem to lend support to the view blaming the foreign exchanges, because the rise in money wages offset the forces making for equilibrium of the foreign-exchange rates. But Robinson does not fully endorse Bresciani's or the German interpretation. In her view the eventual stabilization of the mark in November 1923 does not support the conclusion that monetary stringency is necessary and sufficient to put an end to inflation. In Robinson's view the stabilization succeeded because by that time the old German mark had shed almost all the standard functions that money is to serve.

Selected works

1929. The movement of wages in Germany during the depreciation of the mark and after stabilization. *Journal of the Royal Statistical Society* 92, 374–414.

1931. *Le vicenda del marco tedesco*. Trans. by M. Sayers as *The Economics of Inflation: A Study of Currency Depreciation in Post-war Germany 1914–1923*, with a foreword by Lionel Robbins. London: Allen & Unwin; New York: Barnes & Noble, 1937.

1932. *Inductive Verification of the Theory of International Payments*. Cairo: Noury.

1934a. Egypt's balance of trade. *Journal of Political Economy* 42, 371–84.

1934b. The ‘purchasing power parity’ doctrine. *L'Egypte Contemporaine* 25, 433–64.

1936. The theory of saving. *Economica* 3, Part I, February, 1–23, Part II, May, 162–81.

1937. On Pareto's Law. *Journal of the Royal Statistical Society* 100, 421–32.

1938. The multiplier in practice: some results of recent German experience. *Review of Economics and Statistics* 20(May), 76–88.

1939. Annual survey of statistical data: Pareto's Law and the index of inequality of incomes. *Econometrica* 7(April), 107–33.

1942. *Economic Policy for the Thinking Man*. London: Hodge.

1964. Articles contributed by Costantino Bresciani-Turroni to the *Review of Economic Conditions in Italy* in the years from 1947 to 1962. Rome: Banco di Roma.

Bibliography

Anon. 1964. Obituary of C. Bresciani-Turroni, with Bibliography. *Review of Economic Conditions in Italy* 3, Rome: Banco di Roma.

Gambino, A. 1965. La politica monetaria e creditizia negli scritti di C. Bresciani-Turroni. *Studi economici* 20, 196–219.

Gambino, A. 1972. C. Bresciani-Turroni. *Dizionario biografico degli italiani*, vol. 14. Rome: Instituto della Enciclopedia Italiana.

Robinson, J. 1938. Review of The Economics of Inflation. *Economic Journal* 48, 507–13.

Tamagna, F.M. 1967. C. Bresciani-Turroni. *International Encyclopedia of the Social Sciences*, vol. 2. New York: Macmillan.

How to cite this article

Spiegel, Henry W. "Bresciani-Turroni, Costantino (1882–1963)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000196>
doi:10.1057/9780230226203.0160

The New Palgrave Dictionary of Economics Online

Bretton Woods system

Peter B. Kenen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Under the Bretton Woods System, created after the Second World War, each country had to peg its currency to gold or to the US dollar, but it could obtain temporary financing from the International Monetary Fund. In practice, countries pegged their currencies to the dollar and accumulated dollar reserves, which they could use to buy gold from the US Treasury. This regime served to finance US payments deficits but prevented the United States from changing its exchange rate. The system was undermined when other countries' dollar holdings came to exceed US gold holdings. It was abandoned in 1973, when the major industrial countries let their currencies float.

Keywords

bancor; Bank of England; Bretton Woods System; capital controls; convertibility; devaluation; exchange rate stability; fixed exchange rates; floating exchange rates; foreign-exchange market; gold exchange standard; gold standard; International Monetary Fund; Keynes, J. M.; Nixon, R.; Nurkse, R.; reserve currency; Special Drawing Rights; trade unions; Triffin, R.; White, H.

Article

The international monetary system established at the end of the Second World War is commonly known as the Bretton Woods System. It takes its name from the conference held at Bretton Woods, New Hampshire, USA, in 1944, which adopted the Articles of Agreement of the International Monetary Fund (IMF) and thus put in place the rules and arrangements that would govern international monetary relations in the post-war world.

A comprehensive history of the Bretton Woods System would have to review the monetary and fiscal policies of the major industrial countries, most notably those of the United States and United Kingdom, the key-currency countries, describe the evolution of monetary cooperation, and recite the history of the IMF itself. An analytic assessment would have to examine balance-of-payments adjustment under the

Bretton Woods System and compare the merits of pegged and floating exchange rates.
This account has narrower objectives. It reviews the origins of the system, the rules adopted at Bretton Woods, the differences between those rules and the way the system worked in practice, and the forces leading to the breakdown of the system in the early 1970s. Readers who want more detailed accounts may consult Cooper (1968), Solomon (1982), de Vries (1987), James (1996), Eichengreen (2006), and the official histories of the IMF (Horsefield, 1969; de Vries, 1976; 1986).

The origins of the system

The design of the Bretton Woods System cannot be understood without recalling the monetary history of the interwar period and the lessons drawn from it at the time. Recent writers have drawn somewhat different lessons. Thus, Eichengreen (1991) argues that the credibility of the gold standard in the decades before the First World War depended on close cooperation among central banks, not on the exercise of hegemonic influence by the Bank of England, and that the absence of comparable cooperation doomed the gold-standard arrangements of the interwar period; he also argues that fiscal rigidities greatly compounded the problems of monetary management. But these are lessons for our time, reflecting recent concerns, not those that influenced the design of the Bretton Woods System.
At the end of the First World War, governments were firmly committed to the restoration of the gold standard, and most of them returned to gold during the 1920s. They did so unilaterally and sequentially, however, by adopting gold values for their own currencies. Although some such as Keynes (1925) warned them of the risks they were running, they paid too little attention to the pattern of exchange rates established by their actions. Nor did they understand completely the new environment in which they would have to maintain the gold standard – how monetary and fiscal policies would be constrained by the transfer of financial activity and influence from London to New York, by the domestic and foreign debt-service burdens built up by wartime borrowing, and by the increased power of the trade unions and of the political parties affiliated with them.
The new gold standard collapsed in fewer than ten years, in the same sequential way that it was put together. Country after country let go of gold and allowed its exchange rate to float – to be determined by supply and demand in the foreign-exchange market – but they soon began to intervene in that market in order to influence the behaviour of exchange rates. Even at that point, moreover, they acted unilaterally, not cooperatively. Central banks began to cooperate in the late 1930s, but the process was halted by the outbreak of war and the imposition of currency controls.
What lessons were learned from this experience? Writing for the League of Nations (1944, p. 210), Ragnar Nurkse put them in terms that were widely endorsed at the time. The setting of exchange rates, he concluded, could not be left to market forces:

> A system of completely free and flexible exchange rates is conceivable and may have certain attractions in theory. … Yet nothing would be more at variance with the lessons of the past.
> Freely fluctuating exchanges involve three serious disadvantages. In the first place, they create an element of risk which tends to discourage international trade.…
> Secondly, as a means of adjusting the balance of payments, exchange fluctuations involve

> constant shifts of labour and other resources between production for the home market and production for export. Such shifts may be costly … and are obviously wasteful if the exchange-market conditions that call for them are temporary.…
> Thirdly, experience has shown that fluctuating exchanges cannot always be relied upon to promote adjustment. Any considerable or continuous movement of the exchange rate is liable to generate anticipations of a further movement in the same direction.

Yet the setting of exchange rates, Nurkse argued, cannot be left to individual governments:

> An exchange rate by definition concerns more currencies than one. Yet exchange stabilization [in the interwar period] was carried out as an act of national sovereignty in one country after another with little or no regard to the resulting interrelationship of currency values in comparison with cost and price levels. … The piecemeal and haphazard manner of international monetary reconstruction sowed the seeds of subsequent disintegration. (League of Nations, 1944, pp. 116–17)

Finally, governments should not be expected to sacrifice domestic economic stability merely to maintain exchange rate stability:

> Experience has shown that stability of exchange rates can no longer be achieved by domestic income adjustments if these involve depression and unemployment. Nor can it be achieved if such income adjustments involve a general inflation of prices which the country concerned is not prepared to endure. It is therefore only as a consequence of internal stability … that there can be any hope of securing a satisfactory degree of exchange stability as well. (League of Nations, 1944, p. 229)

The plans that governments drafted in anticipation of the Bretton Woods conference differed in many ways but did not disagree about these matters. A new international institution would be needed to supervise exchange rate policies, in order to promote exchange rate stability and prevent competitive devaluations, but it would also have to concern itself with ‘the promotion and maintenance of high levels of employment and real income’ (Articles of Agreement, Article I (ii)).

The design of the system

The design of the new monetary system was decided before the Bretton Woods conference, in talks between British and American negotiators. The British were led by John Maynard Keynes, the Americans by Harry Dexter White, and the two countries’ proposals are known as the Keynes and White plans. They differed mainly in the way that they would provide financing for external imbalances. (On the plans and subsequent negotiations, see Gardner, 1969; Horsefield, 1969; Dam, 1982.)

The Keynes plan was quite radical and reflected Keynes's concerns about the post-war situation. In the short run, Britain would need balance-of-payments financing; in the long run, the United States was likely to experience another depression, driving other countries into balance-of-payments deficit, and

forcing them to choose between domestic stability and exchange rate stability if they could not obtain adequate financing. Hence, Keynes sought to create a monetary institution able to issue a new international currency (which Keynes called ‘bancor’); it would be held and used by governments and central banks for settling external imbalances.

The White plan was more conservative and reflected White's concern that a large and elastic supply of international money would give other countries an open-ended claim on the real resources of the United States. (In other words, the United States would wind up holding all of Keynes's bancor.) Hence, White sought to limit the supply of reserve credit by providing the new monetary institution with a finite pool of national currencies and gold, rather than the power to issue a new money of its own. (Ironically, the White plan failed to anticipate the emergence of the US dollar as a reserve currency, which made the supply of reserves very elastic and helped to undermine the Bretton Woods System at the start of the 1970s, when it became apparent that the United States could not maintain convertibility between the dollar and gold.)

The plan adopted at Bretton Woods was much like the White plan, although it made concessions to Keynes's concern about the danger of a deep US depression. If a country's currency became ‘scarce’ in world trade and in the IMF itself, because the country was running a balance-of-payments surplus, the IMF could ration that currency and authorize its members to limit imports from the surplus country. (This clause was never invoked, however, even in the years of the so-called dollar shortage.)

The Bretton Woods System imposed two major obligations on national governments but gave them something in exchange.

First, every member of the IMF had to peg its currency to gold or the US dollar (which was, in turn, pegged to gold at $35 per ounce). The IMF had to approve the initial exchange rate and every significant change thereafter. Before it could change its exchange rate, moreover, a government would have to show that it faced a ‘fundamental disequilibrium’ in its external accounts. That term was not defined, however, and led to much debate. It came to be interpreted eventually as an unsustainable conflict between ‘external’ and ‘internal’ balance – a situation in which a country could not defend its exchange rate without suffering substantial unemployment or inflation; see Nurkse (1945) and Meade (1951). (The operational issues resemble those which still bedevil attempts to define a fundamental equilibrium exchange rate; see, for example, Williamson, 1983a, and International Monetary Fund 1984.)

With one notable exception, namely Canada, the major industrial countries did peg their exchange rates until the end of the 1960s and did not change them often. There was an extensive exchange rate realignment in 1949, triggered by a devaluation of sterling, but only a handful of changes thereafter. When they did change their rates, however, they did not let the IMF exercise effective supervision; it was informed at the very last minute, too late to offer advice or object. Developing countries, by contrast, adopted many exchange rate arrangements; a few had freely floating rates, and some had separate rates for different classes of transactions, with some rates pegged and others floating.

Second, every member of the IMF was expected to make its currency convertible as soon as possible. It could continue to control capital movements; recall the view expressed by Nurkse, that capital flows had been destabilizing in the interwar years. It could likewise continue to use tariffs and other trade controls for commercial-policy purposes. But it could not keep the resident of another country from using or converting domestic currency acquired from a current-account transaction. A Dane who earned French francs from exports to France was free to use them for another current-account transaction, sell them to someone else wanting to use them, or sell them to the Danish National Bank, which could then present

them to the Bank of France for conversion into Danish currency.

Britain made the pound *fully* convertible for foreigners in 1947, for capital as well as current-account purposes, but it had to retreat speedily when countries that had built up large sterling balances during the war rushed to exchange them for dollars and drained away a large part of a large US loan to Britain. Thereafter, most governments moved cautiously toward current-account convertibility. Western Europe did not reach it until 1958, and some European countries did not abolish all of their capital controls until 1990; see Triffin (1957) and Kaplan and Schleiminger (1989).

In exchange for these commitments, members of the IMF were entitled to draw on the Fund's holdings of currencies and gold when they ran balance-of-payments deficits and could not finance them by drawing down their own reserves. Each IMF member was given a *quota* that governed its subscription to the currency pool, how much it could draw from the pool, and its voting power in the IMF.

The Articles of Agreement, however, did not spell out the conditions under which countries could draw on the pool, and this became a contentious issue. The United States maintained that strict policy conditions would safeguard prospects for repayment and thus protect the drawing rights of other members. Other governments maintained that access should be automatic when a member needed short-term financing. The United States won this battle too, however, and access to most of the Fund's resources was (and remains) tightly linked to policy commitments made in advance by the government involved and monitored closely by the Fund. (On the origins and evolution of IMF conditionality, see Horsefield, 1969; for criticism from various perspectives, see Dell, 1981; Williamson, 1983b; Kenen, 1986.)

The functioning of the system

Under the Bretton Woods System, all governments had the same rights and obligations. But the monetary system did not function symmetrically. (For more on the asymmetries discussed below, see Cooper, 1972; Whitman, 1974.)

First, there was a basic asymmetry between the situations of surplus and deficit countries – an asymmetry typical of pegged-rate regimes. A country can run a balance-of-payments surplus forever, although it may become uncomfortable with the domestic monetary consequences. There is no upper limit to the stock of reserves that a surplus country can acquire when it intervenes in foreign-exchange markets to keep its currency from appreciating. But a country cannot run a deficit for ever. It will exhaust its reserves as it goes on intervening to keep its currency from depreciating. The speed at which it loses them, moreover, is likely to accelerate as its holdings fall; speculative pressures will build up as foreign-exchange markets become convinced that the country will have to devalue its currency. Therefore, pegged-rate regimes tend to display a devaluation bias.

This bias would not matter in a two-country world, where the devaluation of one currency is no different from a revaluation of the other. It matters importantly in a multi-country world, where devaluation by a deficit country revalues every other currency, not just the surplus countries’ currencies, and revaluation by a surplus country devalues every other currency, not just the deficit countries’ currencies. And the bias had significant effects on the viability of the Bretton Woods System. Devaluations by deficit countries were more frequent than revaluations by surplus countries, causing a gradual revaluation of the US dollar that weakened the competitive position of the United States.

This effect could have been offset by a devaluation of the dollar, but other asymmetries made that difficult. The dominance of the US economy and the key-currency role of the US dollar conferred important privileges on the United States but also limited its policy options.

The size and comparative stability of the US economy made for an asymmetry in policy determination. For most of the life of the Bretton Woods System, US monetary and fiscal policies were aimed exclusively at domestic targets – high employment, economic growth and price stability. There was no true policy coordination between the United States and the other industrial countries, although there were frequent consultations, especially in the 1960s. There were instead one-sided adaptations by the other countries, as they sought to keep their economies in line with the US economy; see, for example, Artis and Ostry (1986) and Kenen (1989).

Furthermore, the strength of the US economy permitted the United States to forgo an active exchange rate policy until the final years of the Bretton Woods System. It was the 'n^{th} country' in the system, whose exchange rate reflected the exchange rate policies of all other countries.

The passivity of the United States was helpful from one standpoint. In a world with n countries and currencies, there are only $n - 1$ independent exchange rates, which makes it impossible for all n countries to pursue independent exchange rate policies (Mundell, 1969). Therefore, the passivity of the United States helped to avoid policy conflict. Nevertheless, the arrangements supporting and promoting that passivity made the Bretton Woods System too brittle, forcing the United States to take very damaging measures in 1971, when it tried to achieve an exchange-rate realignment. Most countries defined their exchange rates with reference to the dollar, not gold, and stabilized those rates by buying and selling dollars. Hence, it was unnecessary for the United States to stabilize the dollar by buying and selling other countries' currencies. But it was also impossible for the United States to conduct an exchange rate policy of its own without other countries' tacit consent. It could change the gold price of the dollar, but it could not change the Deutschemark, franc and yen prices if Germany, France and Japan refused to change the dollar prices of their national currencies.

These asymmetries led to others. The US dollar was the only important convertible currency at the end of the Second World War, which caused it to become the key currency of the Bretton Woods System. It was used for official intervention in the foreign-exchange market and held along with gold as a reserve asset. There was, indeed, a neat division of labour under the Bretton Woods System. By buying and selling dollars in foreign-exchange markets, other governments stabilized the value of the dollar in terms of their national currencies. For its part, the United States stood ready to swap gold for dollars at $35 per ounce, making gold and dollars nearly perfect substitutes for the holders of reserves.

This arrangement imparted elasticity to the supply of reserves. Other governments wanting additional reserves could accumulate dollars, rather than compete for limited supplies of gold. But it had two serious defects.

First, it allowed the United States to run balance-of-payments deficits without necessarily suffering gold losses. When it started to lose gold in the 1960s, moreover, it negotiated ad hoc arrangements and agreements that encouraged other countries to hold dollars instead of buying gold; see Coombs (1976) and Solomon (1982). Accordingly, the United States was not obliged to deal quickly with its balance-of-payments problem. In the words of Charles de Gaulle, it enjoyed the 'exorbitant privilege' of using its domestic money to pay its foreign bills.

Second, the reserve-creating arrangements of the Bretton Woods System posed a basic threat to the

viability of the system – a point made emphatically by Robert Triffin (1960). Because the IMF could not create international money – the Keynes plan had been rejected – the United States *had* to run balance-of-payments deficits to supply reserves to the rest of the world. As it did so, moreover, its *net* reserve position was likely to deteriorate; its dollar liabilities were apt to grow faster than its gold stock. Any such deterioration, moreover, was bound to impair the credibility of the US promise to sell gold for dollars, reduce the attractiveness of the dollar as a reserve asset, and wreck the reserve-creating arrangement on which the system depended.

Triffin's critique of the gold–dollar standard and his own plan for reform produced a torrent of other proposals (see, for example, Grubel, 1963) and led eventually to a promising reform. In 1968, governments adopted the First Amendment to the Articles of Agreement of the IMF, allowing the Fund to create a new reserve asset, the Special Drawing Right (SDR), when and if this was required to meet the demand for reserves. The value of the SDR was defined initially in terms of gold (in a manner that priced it at one US dollar). In 1976, however, the Second Amendment to the Articles of Agreement took the IMF off gold by making the SDR the official standard of value, and the value of the SDR itself was redefined in terms of a basket of national currencies.

Small amounts of SDRs were actually created in 1970–72 and 1979–81. But the SDR arrived on the monetary scene too late to forestall the collapse of the Bretton Woods System, and has never acquired a major role in the international monetary system.

The collapse of the system

In 1960, when Triffin published his attack on the gold-exchange standard, the US reserve position was very strong; US gold holdings were far larger than US liabilities to foreign governments and central banks. But the balance-of-payments deficits of the 1960s eroded its reserve position, fulfilling Triffin's prophecy. The collapse of the Bretton Woods System, however, was not due to this development alone. It reflected the gradual deterioration in the competitive position of the United States, exacerbated by the economic consequences of the Vietnam War. By the late 1960s, the United States had ceased to be the stable centre of the monetary system; its inflation rate was rising, and its trade surplus was vanishing.

The first major break in the commitment to pegged exchange rates came in 1969. Rumours that the Deutschemark would be revalued vis-à-vis the dollar attracted huge amounts of speculative capital to Germany and caused the German authorities to let the Deutschemark float rather than accumulate more reserves and thus increase the German money supply. The Deutschemark appreciated by ten per cent during the next four weeks, after which the German authorities converted the appreciation into a revaluation by pegging the Deutschemark–dollar rate close to its new market level.

The fatal break came in 1971, when the US payments deficit widened suddenly. It ran at an annual rate of $20 billion during the first quarter of 1971, four times as large as it had been in any previous calendar year, producing new rumours that the Deutschemark would be revalued. On a single day in May, the German authorities had to buy more than $1 billion in the foreign-exchange market to keep the dollar from depreciating, and they had to buy a similar amount during the first hour of the next day's trading. Therefore, they quit and permitted the Deutschemark to float again.

An appreciation of the Deutschemark, however, could not solve the basic problem – the very large increase in the US payments deficit – and American officials began to look for the best way to achieve a general exchange rate realignment. They did not want to raise the dollar price of gold, the only option

open to them unilaterally. That would break faith with the governments that had held dollars rather than gold, and it might not work. A higher dollar price for gold would not devalue the dollar in a meaningful way unless other governments agreed to raise the dollar prices of their currencies. (On the discussions within the US government, see Solomon, 1982; Gowa, 1983; Leeson, 2003.)

The crisis came to a head in August, after France had bought gold from the United States to repay a drawing on the IMF, and there were rumours of a large gold purchase by the Bank of England. The rumours were inaccurate but influential. On 15 August 1971, President Richard Nixon announced major changes in US policies. He froze wages and prices temporarily to combat inflation and asked Congress to approve an investment tax credit to stimulate output and employment. He imposed a ten per cent tax on imports and instructed the Secretary of the Treasury to close the gold window – to suspend US purchases and sales of gold.

The last two measures were designed to achieve an exchange rate realignment. They imposed two penalties on any foreign government that refused to revalue its currency. Its exports would be penalized by the tariff, and it could no longer count on buying gold when it purchased dollars in the foreign-exchange market to keep its currency from appreciating. The United States was widely criticized for adopting ‘shock tactics’ and breaking the rules of the trading system as well as those of the monetary system. But the tactics worked. In the weeks following the President's speech, several governments joined Germany in letting their currencies float temporarily, and after three months’ bargaining a meeting at the Smithsonian Institution in Washington agreed to realign exchange rates formally. Most of the major industrial countries revalued their currencies against the dollar, and the United States devalued the dollar against gold. (It did not reopen the gold window, however, so that the new official price of gold was purely notional – the one at which the US Treasury would *not* buy or sell.)

The new pegged-rate regime, however, fell apart rapidly. The pound sterling was allowed to float in June 1972, and the end of the Bretton Woods System came early in 1973, after an attempt by the United States to negotiate a second exchange-rate realignment. Japan allowed the yen to float in February, and six members of the European Community agreed in March to allow their currencies to float jointly. These measures were seen to be temporary at the time, but governments soon came to believe that it would be impossible to return to pegged exchange rates, especially after the oil shock of 1973–4 and the economic problems it produced. In 1976, the Second Amendment to the Articles of Agreement of the IMF replaced the original commitment to pegged exchange rates with much looser obligations. Governments would be free to choose any exchange rate arrangement *except* a fixed gold price, and the IMF was told to ‘exercise firm surveillance over the exchange rate policies of members’ (Articles of Agreement, Art. IV (3)) but was not told how to do that.

Although the term ‘Bretton Woods System’ is usually used to characterize the monetary system that prevailed until the early 1970s, a few have used it to describe a far more recent regime, which they describe as Bretton Woods II (Dooley, Folkerts-Landau and Garber, 2003; 2004). What do they mean?

Throughout the 1960s, the United States ran balance-of-payments deficits because net capital outflows from the United States exceeded the US current-account surplus. In recent years, the United States has run balance-of-payments deficits because the US current-account deficit has exceeded net private capital inflows into the United States, and there has been as a result a huge accumulation of dollar reserves by countries that have been reluctant to let their currencies appreciate, most notably China, other East Asian countries, and the main oil-exporting countries. Many economists have warned that this payments pattern is unsustainable; see, for example, Obstfeld and Rogoff (2005) and Roubini and Setser (2004).

The dissenters, however, compare it to the payments pattern of the late 1950s and early 1960s, which lasted for a decade before the Bretton Woods System collapsed. They maintain that the surplus countries, especially those in Asia, have chosen deliberately to hold down the dollar values of their currencies and thereby accumulate dollar reserves because they count on export growth to foster rapid output growth and thus the transformation of their national economies. There is, of course, no way to resolve this controversy. Time alone can do that.

See Also

- international financial institutions (IFIs)
- International Monetary Fund
- World Bank

Bibliography

Artis, M. and Ostry, S. 1986. *International Economic Policy Coordination*. London: Royal Institute of International Affairs and Routledge and Kegan Paul.

Coombs, C. 1976. *The Arena of International Finance*. New York: Wiley.

Cooper, R. 1968. *The Economics of Interdependence*. New York: McGraw-Hill for the Council on Foreign Relations.

Cooper, R. 1972. Eurodollars, reserve dollars, and asymmetries in the international monetary system. *Journal of International Economics* 2, 325–44.

Dam, K. 1982. *The Rules of the Game*. Chicago: University of Chicago Press.

Dell, S. 1981. On being grandmotherly: the evolution of IMF conditionality. *Essays in International Finance* 144. Princeton: International Finance Section, Princeton University.

de Vries, M. 1976. *The International Monetary Fund, 1966–1971*. Washington: IMF.

de Vries, M. 1986. *The International Monetary Fund, 1972–1978*. Washington: IMF.

de Vries, M. 1987. *Balance of Payments Adjustment, 1945 to 1986: The IMF Experience*. Washington: IMF.

Dooley, M., Folkerts-Landau, D. and Garber, P. 2003. An essay on the revived Bretton Woods System. Working Paper No. 9971. Cambridge, MA: NBER.

Dooley, M., Folkerts-Landau, D. and Garber, P. 2004. Direct investment, rising real wages and the absorption of excess labour in the periphery. Working Paper No. 10626. Cambridge, MA: NBER.

Eichengreen, B. 1991. *Golden Fetters: The Gold Standard and the Great Depression, 1919–1939*. New York: Oxford University Press.

Eichengreen, B. 2006. *Global Imbalances and the Lessons of Bretton Woods*. Cambridge, MA: MIT Press.

Gardner, R. 1969. *Sterling–Dollar Diplomacy*. London: Oxford University Press.

Gowa, J. 1983. *Closing the Gold Window*. Ithaca, NY: Cornell University Press.

Grubel, H., ed. 1963. *World Monetary Reform*. Stanford: Stanford University Press.

Horsefield, J. 1969. *The International Monetary Fund, 1945–1965*. Washington: IMF.

International Monetary Fund. 1984. Issues in the assessment of the exchange rates of industrial countries. Occasional Paper No. 29. Washington, DC: IMF.

James, H. 1996. *International Monetary Cooperation since Bretton Woods*. Washington, DC: IMF; New York and London: Oxford University Press.

Kaplan, J. and Schleiminger, G. 1989. *The European Payments Union*. Oxford: Clarendon Press.

Kenen, P.B. 1986. Financing, adjustment and the International Monetary Fund. *Studies in International Economics*. Washington, DC: Brookings Institution.

Kenen, P.B. 1989. *Exchange Rates and Policy Coordination*. Manchester: Manchester University Press.

Keynes, J.M. 1925. *The Economic Consequences of Mr. Churchill*. In *The Collected Writings of John Maynard Keynes, vol. 9: Essays in Persuasion*. London: Macmillan.

League of Nations. 1944. *International Currency Experience*. Geneva: League of Nations.

Leeson, R. 2003. *Ideology and the International Economy: The Decline and Fall of Bretton Woods*. Basingstoke: Palgrave Macmillan.

Meade, J. 1951. *The Balance of Payments*. London: Oxford University Press.

Mundell, R. 1969. Problems of the international monetary system. In *Monetary Problems of the*

International Economy, ed. R. Mundell and A. Swoboda. Chicago: University of Chicago Press.

Nurkse, R. 1945. Conditions of international monetary equilibrium. *Essays in International Finance 4*. Princeton: International Finance Section, Princeton University.

Obstfeld, M. and Rogoff, K. 2005. Global current account imbalances and exchange rate adjustments. *Brookings Papers on Economic Activity* 2005(1), 67–123.

Roubini, N. and Setser, B. 2004. The U.S. as a net debtor: the sustainability of the US external balances. Mimeo, School of Business, New York University, and University College Oxford.

Solomon, R. 1982. *The International Monetary System, 1945–1981*. New York: Harper and Row.

Triffin, R. 1957. *Europe and the Money Muddle*. New Haven: Yale University Press.

Triffin, R. 1960. *Gold and the Dollar Crisis*. New Haven: Yale University Press.

Whitman, M. 1974. The current and future role of the dollar: how much symmetry? *Brookings Papers on Economic Activity* 1974(3), 539–83.

Williamson, J. 1983a. The exchange rate system. *Policy Analyses in International Economics 5*. Washington, DC: Institute for International Economics.

Williamson, J., ed. 1983b. *IMF Conditionality*. Washington, DC: Institute for International Economics.

How to cite this article

Kenen, Peter B. "Bretton Woods system." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000198> doi:10.1057/9780230226203.0161

The New Palgrave Dictionary of Economics Online

bribery

Susan Rose-Ackerman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bribery is a form of rent-seeking meant to induce officials to serve private interests. Principal–agent relations are at the heart of the economic analysis of the subject. Bribery undermines government functioning by influencing electoral outcomes, lowering the benefits from public contracts, distorting the allocation of public benefits and costs, and introducing delay and red tape. Empirical work documents the negative consequences of corruption, and economic theory helps one understand the underlying incentives for payoffs.

Keywords

bribery; bureaucracy; campaign finance, economics of; corruption; democracy; presidential democracy; principal and agent; privatization; procurement, government; proportional representation; public interest; public services; rent seeking

Article

Bribery and corruption are a form of rent seeking meant to induce official agents to serve the interests of those making payoffs.

Principal–agent relations are at the heart of the economic analysis of bribery. Payoffs induce agents to go against the interests of their principals, be they higher-level officials, politicians, or the citizenry in general. Bribery undermines the interests of principals by influencing electoral outcomes, lowering the benefits from public contracts, distorting the allocation of public benefits and costs, and introducing delay and red tape. The study of bribery thus highlights the conflict between the public interest and the market. Widespread bribery can transform government actions ostensibly based on democratic or meritocratic principles into ones based on willingness-to-pay.

The theory of perfect competition emphasizes the impersonality of all market dealings. A manufacturer will sell to all customers irrespective of their race, gender, or inherent charm. Similarly, the ideal official

makes decisions on the basis of objective, meritocratic criteria and is not influenced by personal, ethnic or family ties. Bribes can replace an impersonal meritocratic procedure with an impersonal willingness-to-pay procedure, or payoffs can support a system of personalized favours based on close personal relations. Alternatively, bribery can replace a personalized system based on family and ethnic ties with one based on financial capacity.

Early economic work on bribes concentrated on their role as prices and argued that they enhanced the efficiency of government (Leff, 1964). This perspective has been overtaken by both theoretical and empirical work arguing for and documenting the costs of systemic corruption. On the theory see, for example, Rose-Ackerman (1978), Shleifer and Vishny (1993), and the literature reviewed in Bardhan (1997) and Rose-Ackerman (1999). Cross-country empirical studies are reviewed in Graf Lambsdorff (2006) and Rose-Ackerman (2004, pp. 303–10). Kaufmann and Kraay (2002), part of a World Bank Institute governance team, deal with the issue of whether high corruption causes low growth or whether low growth generates corruption. They conclude that the causal arrow runs from high corruption to low growth, but the issue remains vexed and has led to a turn to history to seek independent causes. The problem with econometric studies that use historical data, however, is that they cannot be a guide to policy. If one is concerned with reform, it seems necessary to engage with the messy real world of feedback loops and multiple causes. History can then be put to different use as a source of case studies of successful and failed reform efforts (Glaeser and Goldin, 2006).

Corruption arises under many conditions in modern states. This article considers three variants: political corruption, kickbacks in major procurement and privatization contracts, and corruption in the allocation of benefits and burdens (for more details and references to the literature see Rose-Ackerman, 1978; 1999; 2004; 2006).

Political corruption

Non-democratic states tend to be more corrupt than democratic states, but democracies are clearly not immune from corruption. Obviously, corruption that arises from the competition for public office will be more prominent in democracies. The empirical results suggest that it is only long-established democracies that are less corrupt than other systems. As an example, the transition from socialism to market democracy in eastern Europe and central Asia has been fraught with corruption. During the transition, payoffs were a way to deal with an uncertain and rapidly changing environment just as, in the past, they had been a response to the excessive rigidities of a planned economy.

Furthermore, even within the universe of democracies, corruption levels vary with the constitutional structure of government. Kunicová and Rose-Ackerman (2005) find that presidential systems with legislatures selected by proportional representation are more subject to corruption than other democratic forms. Their explanation for this phenomenon is a bargaining situation in which a few strong party leaders negotiate with a powerful chief executive to share the spoils of office subject to relatively ineffective checks from voters, minority parties, and rank-and-file legislators.

At the individual level, the corruption of elected politicians depends upon the trade-off between their desire for re-election and their interest in monetary gain. Suppose voters are well-informed about politicians’ votes but cannot observe bribes directly. Assume that politicians run for re-election on their voting record and that no campaign spending is needed. Then a bribe designed to change a vote in the

legislature will cost the politician some constituency support. Bribes must be sufficient to compensate for the reduced chance of re-election. *Ceteris paribus*, politicians with the lowest reservation bribes are those who are either quite certain of being elected or quite sure of defeat; in each case a decline in electoral support has little impact on the ultimate outcome. The closer the race, the higher will be the politician's reservation bribe.

In this simple model there is no need for campaign contributions, so bribes are used only for personal gain, and there is a direct trade-off between bribes and the probability of re-election. If payoffs *can* be used either to support a re-election campaign or as personal income, then all politicians may be corruptible, depending on their moral scruples and the salience of the issues influenced by corruption (Rose-Ackerman, 1978, pp. 15–58). In electoral democracies, the control of corruption requires that re-election-seeking politicians feel insecure about their prospects but not too insecure. Too much security of tenure furthers corrupt arrangements. Too much insecurity can have the same effect.

Procurement and privatization

No bribes occur in a perfectly competitive market, where suppliers can sell and demanders can buy all they wish at the going price. If bribes are offered, there must be some prospective excess profits out of which to pay them, and, if bribes are accepted, it must be because the agent's superiors are either privy to the deal themselves or else cannot adequately monitor the agent's behaviour. Corruption requires market imperfections. These are widespread in government procurement, resource concessions, and the privatization of public firms. The government will often be a monopsony purchaser or a monopoly seller; and it may need products not available ‘off the shelf’ so that a negotiated contract is necessary. One might argue that corruption in procurement and the sale of assets furthers efficiency because the most efficient firm will have the highest prospective profits and so be willing to pay the highest bribe. This is simplistic. First, a winning firm in a procurement contract may gain advantage by lowering quality in subtle ways, not immediately obvious to government inspectors. Second, if managers of firms differ in respect for the law, the most unscrupulous have an advantage. Third, keeping payoffs secret both wastes resources and causes the market to operate poorly because of the low level of available information. Finally, the desire for payoffs may induce officials to contract for overly costly one-of-a-kind projects capable of hiding large kickbacks and to privatize firms on terms that favour corrupt bidders.

Mandating more effective competition is not always an option. In such situations one must consider the role of detection and punishment. Becker and Stigler (1974) first applied work on the economics of crime to corrupt payments. They stress the importance of giving each employee a stake in his or her job by, for example, providing non-vesting pensions. This will make workers less likely to take risks that could lead to their dismissal. More generally, the expected punishment for bribery should be tied to the marginal gain from marginal increases in the payoff (Rose-Ackerman, 1978, pp. 109–35; 1999, pp. 52–9). Otherwise only some bribes will be deterred. Thus the marginal expected penalty for the bribe-taker, that is, the probability of apprehension and conviction times the penalty if convicted, must rise by at least one dollar for every dollar increase in expected payoff. If it does not, then even if a large lump-sum penalty is levied, only relatively small bribes may be prevented. The bribe-payer's marginal penalty should be tied, not to the size of the bribe, but to the marginal increase in profit that a bribe makes

possible. Penalties set at a multiple of the bribe paid may have little deterrent effect on bribe-payers if the expected profits are many times larger.

Dispensers of benefits and burdens

Low-level officials frequently have considerable discretion to decide who should receive a scarce benefit such as a unit of public housing, expedited access to an important person, a liquor licence, or assignment to a particular judge. Others, such as health and safety inspectors, tax collectors, and the police, have the power to impose costs and the discretion to refuse to exercise that power. Although legal pricing systems can sometimes substitute for payoffs here, in many cases there is a strong public policy reason for opposing a market solution.

How then can corruption be controlled? There are many ways to limit the discretion of officials to extract payoffs (Rose-Ackerman, 1999, pp. 39–68). Consider just one option: the introduction of competitive pressures (Rose-Ackerman, 1978, pp. 137–66). If a bureaucracy dispenses a scarce benefit, competition can be introduced by permitting an applicant to reapply if he has been turned down by one official. Then if the cost of reapplication is small, the first official cannot demand a large bribe in return for approving the application; in fact the offered bribe may be forced down so low that the official may turn it down and instead behave honestly. A few honest officials in this system may produce honesty in the others. Notice, however, that unqualified applicants will still wish to make payoffs, and their willingness-to-pay increases if they expect that most other officials to whom they could apply are honest. The case for competition among inspectors or police is somewhat different and depends upon the feasibility and cost of overlapping authority. Thus, the operator of a gambling parlour will not pay much to a corrupt policeman if a second independent policeman is expected to come along shortly. The whole precinct must be on the take, that is, monopolized, to make high bribes worthwhile.

In short, the role of competitive pressures in preventing corruption may be an important aspect of a strategy to deter the bribery of low-level officials, but it requires a broad-based exploration of the impact of both organizational and market structure on the incentives for corruption facing both bureaucrats and their clients.

See Also

- directly unproductive profit-seeking (DUP) activities
- political institutions, economic approaches to
- principal and agent (i)
- principal and agent (ii)
- rent seeking

Bibliography

Bardhan, P. 1997. Corruption and development: A review of issues. *Journal of Economic Literature* 35, 1320–46.

Becker, G.S. and Stigler, G.J. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3, 1–18.

Glaeser, E.L. and Goldin, C. 2006. *Corruption and Reform: Lessons from America's Economic History*. Chicago: Chicago University Press.

Graf Lambsdorff, J. 2006. Causes and consequences of corruption: what do we know from a cross-section of countries? In Rose-Ackerman (2006).

Heidenheimer, A.J., Johnston, M. and LeVine, V.T., eds. 1980. *Political Corruption: A Handbook*. New Brunswick, NJ: Transaction Publishers.

Kaufmann, D. and Kraay, A. 2002. Growth without governance. *Economica* 3, 169–229.

Kunicová, J. and Rose-Ackerman, S. 2005. Electoral rules and constitutional structures as constraints on corruption. *British Journal of Political Science* 35, 573–606.

Leff, N. 1964. Economic development through bureaucratic corruption. *American Behavioral Scientist* 8 (3), 8–14.

Rose-Ackerman, S. 1978. *Corruption: A Study in Political Economy*. New York: Academic Press.

Rose-Ackerman, S. 1999. *Corruption and Government: Causes Consequences and Reform*. Cambridge UK and New York: Cambridge University Press.

Rose-Ackerman, S. 2004. Governance and corruption. In *Global Crises, Global Solutions*, ed. B. Lomborg. Cambridge: Cambridge University Press.

Rose-Ackerman, S., ed. 2006. *International Handbook on the Economics of Corruption*. Cheltenham, UK and Northampton, MA: Edward Elgar.

Shleifer, A. and Vishny, R. 1993. Corruption. *Quarterly Journal of Economics* 108, 599–617.

How to cite this article

Rose-Ackerman, Susan. "bribery." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000199> doi:10.1057/9780230226203.0162

The New Palgrave Dictionary of Economics Online

Bright, John (1811 – 89)

William D. Grampp
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cobden, R.; Corn Laws; female suffrage; Manchester School; Mill, J. S.

Article

John Bright, a Lancashire mill-owner, became a national figure in the campaign that repealed the Corn Laws in 1846 and that came to be known as the Manchester School.
Elected to the House of Commons in 1843, he continued to represent industrial constituencies most of his life and worked tirelessly for radical reform which to him meant reducing the scope of government, making it more representative and keeping its foreign policy peaceful. He was a man of strong views but not doctrinaire or unwilling to change them.
Believing in the market, he opposed factory legislation but not as it applied to children. At one time he supported John Stuart Mill's effort to give women the vote but later opposed the idea. He was against a state church, yet proposed its funds be distributed to all denominations as a once-and-never-again subsidy which recalls Smith's artful scheme. Although a Quaker, he never condemned war in principle and said that violence, while rarely called for, was sometimes necessary.
In his day Bright was said to be the pacifist who could have been a pugilist if he had not been a Quaker. He does evoke truculence but what stands out a century later is his honesty and fierce independence. He combined them with an extraordinary speaking ability – in turns eloquent, persuasive, charming, brutally frank, cogent, and clever – all of which he could be because he had a first-rate mind. Never quite the equal of his intimate friend and ally, Richard Cobden, he nevertheless was one of the great figures in the reform movements of the century.

See Also

- Manchester School

Selected works

1866. *Speeches on Political Reform*. London: Simpkin.

1878. *Speeches on Questions of Public Policy*, ed. J.E. Thorold Rogers. London: Macmillan.

1879. *Public Addresses of John Bright*, ed. J.E. Thorold Rogers. London: Macmillan.

How to cite this article

Grampp, William D. "Bright, John (1811–89)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000201> doi:10.1057/9780230226203.0163

The New Palgrave Dictionary of Economics Online

Britain, economics in (20th century)

K. Tribe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The foundations of the modern discipline of economics were Marshallian, symbolized by acknowledgement of his *Principles of Economics* (1890) as the leading English-language exposition, and the creation of the very first undergraduate teaching course in economics in Cambridge in 1903. The work of Maynard Keynes made a similar, lasting international impression, although in the second half of the century a neoclassical synthesis became internationally predominant, a tendency that fostered an interest in technique rather than economic problems that would have been quite alien to Marshall and Keynes.

Keywords

Ashley, W. J.; Boulding, K. E.; Britain, economics in (20th century); Cannan, E.; Chamberlin, E. H.; Chapman, S.; Coase, R. H.; Devons, E.; Dobb, M. H.; economics, teaching of; Edgeworth, F. Y.; Fawcett, H.; Flux, A.; free trade; Gorman, T.; Granger, C. W. J.; Hansen, A.; Harrod, R. F.; Hayek, F. A.; Henderson, H.; Hewins, W. A. S.; Hutchison, T.; innovation; Institute for Fiscal Studies (UK); invention; Jewkes, J.; Kahn, R. F.; Keynes, J. M.; Keynesianism; Klein, L. R.; Knight, F. H.; Lipsey, R.; Marschak, J.; Marshall, A.; Meade, J. E.; Meredith, H.; Monopoly; neoliberalism; neo-Ricardians; perfect competition; Phelps Brown, E. H.; Pigou, A. C.; Political Economy Club; Price, L.L.; Robbins, L. C.; Robertson, D.H.; Robinson, E. A. G.; Robinson, J. V.; Royal Economic Society; Samuelson. P. A.; Schneider, E.; Shackle, G. L. S.; Sraffa, P.; Stackelberg, H. von; Stone, J. R. N.; Walters, A.; Young, A. A.

Article

During the early 1900s, economics in Britain completed its transformation from a science accessible to a literate public to an academic discipline that required specific training; to be a student of economics henceforth implied that one was a college or university student. The literature of economics matched this

transition. It moved out of the sphere of public argument into the closed world of an increasingly specialized academic discipline. Although there was never a perfect match between the general development of economic thinking and the pool of thinkers, these thinkers were henceforth overwhelmingly employees of universities, paid to teach and think about modern economics. Consequently, the story of British economics in the 20th century is closely related to the advance of university institutions, and within these institutions, the formation of new departments of economics. Well into the 1960s, universities, colleges and schools remained the principal employers of ‘trained economists’, for there were very few alternative openings for ‘economists’ in business or public administration. In turn, the extension of opportunities for British university economists to develop their interest in the subject was for most of the 20th century conditional upon their ability to recruit undergraduate students; for taught graduate programmes were likewise a feature of the last third of the century.

In the 1990s, with the reclassification of virtually all higher education as university education and the general deterioration of student–staff ratios, the relationship between teaching and research that had prevailed through the greater part of the century broke down. Given the late appearance of graduate programmes, ‘teaching’ had meant lectures and classes to undergraduates, shared between the staff; while from the 1950s to the 1980s a ‘class’ was no more than a dozen students, in Oxford and Cambridge individual supervision being the norm. It was also usual for the more senior members of the department to present the more elementary lectures, but they, like their junior colleagues, pursued research projects alongside their other duties, supplemented by spells of departmental research leave. This arrangement did not survive into the 1990s. Those economists seeking to pursue a research career (and hence retain their reputation as economists) required a succession of external research grants to sustain any ambition of career development; they sometimes no longer taught at undergraduate level at all. The incentive to deploy senior economists in undergraduate teaching, and hence stimulate an interest in the subject among a younger generation, was seriously compromised. Meanwhile, employers specifically interested in economics graduates usually only required a first degree of their recruits. A Master's qualification was overqualification for anything other than appointment to a technical economic job, while an economics Ph.D. was serious overqualification for anything other than university employment. Given the unattractiveness of university employment to gifted young people, the number of British students studying at this level slumped. This evolutionary development in university institutions coincided with an unrelated transition in the discipline, from a focus on economic problems to an emphasis upon the elaboration of technique. In Britain, as elsewhere, mainstream training in economics had become instruction in a set of mathematical or statistical techniques that might, or might not, illuminate the kind of economic issues with which a wider public outside the university was concerned. Early in the century economics had been propelled into British universities by widespread belief in its public purpose and utility. By the end of the century, the discipline had become dominated by technicians for whom such beliefs were less important. As we shall see, this evolutionary progression was also related to the post-war internationalization of economics, so that by the end of the century the idea of a specifically ‘British’ economics had become an empty one.

Systematic tuition in economic principles originated in Britain. The first three-year university course was the Cambridge tripos, founded in 1903. The London BSc (Econ.), centred on the newly formed London School of Economics, had preceded this in 1901, but was structured in such a way that specialization in economics was only one of a number of social science options; and economics was

taught only during the first year, at a very elementary level, in the commerce degree initiated by Ashley in Birmingham in 1902. The Oxford PPE, linking the study of Philosophy, Politics and Economics, and in this particular order because it had first been proposed by philosophers and opposed by economists, was initiated in 1920 (Chester, 1986, 34 ff.). Ultimately, the London degree had the greatest influence in advancing the study of modern economics – not simply because of the success of the LSE in attracting both students and funding, but because the external London degree offered students resident outside London, and in the wider Empire, the opportunity of studying economics. The new University Colleges of Leicester, Nottingham, Exeter, Southampton, Reading, Hull and Bristol offered to their students of economics the external London BSc (Econ.); and a succession of London Professors, from Cannan through Benham, Stonier and Hague to Lipsey, wrote popular undergraduate textbooks which remained widely used until late in the century.

Alfred Marshall, arguing for his new Tripos, had appealed to the growing need of business and public administration for young recruits conversant with the new science; a plausible enough argument, but one that in practice took many years to realise (Groenewegen, 1995, pp. 556–7). William Ashley, generally unenthused by modern economics, sought a parallel development with his Birmingham commerce degree, intended to place appropriately trained recruits in the middle levels of management. The ambitions of both men were thwarted by a general lack of interest on the part of British business and public administration in ‘new men’. Business remained dominated by small- and medium-size family firms until the interwar years at the very least, and here a professional training in law or accountancy remained a more useful general qualification than a degree in economics or commerce. In the mid-1930s having a first class degree in economics from the University of Cambridge led nowhere in particular: Terence Hutchison, appointed in the 1950s to Birmingham's chair, worked as a *Lektor* at the University of Bonn before the war; Alexander Henderson, later Professor of Economic Theory at Manchester, took a year out but then replaced Kenneth Boulding as Assistant Lecturer in Edinburgh. Economics had become a university discipline, but a degree in economics was a qualification that had little cash value outside academia. Only with the general expansion of the university system in the 1950s did it become customary for bright undergraduates to become in turn graduate students and then junior members of staff – the path taken by Clive Granger at Nottingham, for example. This pattern of training and recruitment altered little until the 1970s when demand for trained economists on the part of financial institutions and public administration began to develop.

The institutions – Cambridge, Oxford, LSE and the provinces

The Cambridge Tripos was the first honours economics programme in the world because it was a key ambition of Alfred Marshall to establish the subject as a modern independent discipline, and he was in a position to realize this ambition. Appointed to the Cambridge Chair in 1884 in succession to Henry Fawcett, author of the Millian *Manual of Political Economy* (1863), Marshall published *Principles of Economics* in 1890, and in 1892 *Elements of Economics of Industry*, an abridged version of the *Principles* for use by students which proved extremely popular. Later in 1891, Marshall oversaw the founding of the British Economic Association (from 1902 the Royal Economic Society, RES) as a vehicle for the publication of the *Economic Journal* (*EJ*), the first number of which appeared in March 1891 (Tribe, 2001). In the United States, the *Quarterly Journal of Economics* had been founded in 1887

as the house journal of Harvard economists, while the *Journal of Political Economy*, founded in 1892, would be a house journal for Chicago economists. Marshall believed that the broad reception of new economics in Britain required a publication ‘open to all schools and parties’, and not therefore tied to any one institution. Following the publication of his textbook as a foundation for teaching, the *EJ* provided a platform for discussion among economic specialists while also keeping them informed of new publications, the current contents of foreign journals, and other relevant developments. The Tripos was the third of Marshall's stones in the new edifice.

Principles and *Elements* were a runaway success in the English-speaking world. The *EJ* in its early years indeed published a wide range of economic opinion – including, for example, the Erfurt Programme of the German Social Democratic Party, Vol. I September, 1891, pp. 531–3. But the Tripos remained merely a pedagogic monument for many years: during the 1930s, as many as 60 per cent of those taking the one-year Part I achieved modest Thirds (Tribe, 2000). Nonetheless, there were, during the 1930s, many graduates whose later reputation as economists began in Cambridge. In the 1880s and 1890s, economics had been taught as an option within the History and the Moral Sciences triposes at Cambridge; Marshall had made himself deeply unpopular among his colleagues with his persistence in seeking a separate existence for the teaching of economics, and having granted his wish in 1902 they proceeded to purge all economics from their own curricula. The tripos was certainly a model of a free-standing economics degree, but even in the boom years of the later 1940s the number of annual Firsts and Upper Seconds in Part II (the final examination) more or less matched the number of eminent economists in the faculty. The tripos, for the first 50 years of its existence, proved more successful in supporting the largest concentration of academic economists in Britain than teaching economics to receptive students.

On the other hand, many of Cambridge's economists turned to writing introductory textbooks under the auspices of the *Cambridge Economics Handbooks* series. The first of the handbooks was Hubert Henderson's *Supply and Demand*, published in 1921, followed by Dennis Robertson on money (1922), Maurice Dobb on wages (1928), and Austin Robinson on the structure of industry (1931) among many others. Maynard Keynes took over the series in the mid-1920s, and drafted a general introduction printed in all editions arguing that economics was a method, not a body of doctrine, ‘an apparatus of the mind, a technique of thinking, which helps its possessor to draw correct conclusions’. Keynes was here reiterating his belief in the *organon* as the core of the Marshallian legacy, ‘a machinery that we build up in our minds, a method, an organon of enquiry that can be turned to particular problems as they arise …’ (Pigou, 1925, pp. 86–7); to which Keynes added the republican principle that the purpose of the *Handbooks* was to expound the elements of economics ‘in a lucid, accurate and illuminating way, so that the number of those who can begin to think for themselves may be increased. It is intended to convey to the ordinary reader and to the uninitiated student some conception of the general principles of thought which economists now apply to economic problems’. Published in the United States and widely circulated in the Empire, some of the handbooks were also translated, emphasising the general absence at this time of similar short works suitable for students of economics, as well as the manner in which Cambridge economics, generally unreceptive to the development of economic thinking elsewhere in Britain and abroad, was nonetheless projected into a wider world.

Oxford economics followed a different path. It had been the centre of British economics in the 1880s, pursuing the development of extension teaching in many provincial centres and graduating among others Edwin Cannan, W.J. Ashley, L.L. Price and W.A.S. Hewins (Kadish, 1982, ch. 2). But Francis

Edgeworth, appointed to the Drummond Chair in 1892, entirely lacked Marshall's institutional ambition, and in any case did not share Marshall's view that an understanding of economics required three years of systematic tuition. During the early 1900s teaching in Oxford remained broadly Millian (Young and Lee, 1993, p. 7), with Marshall being reserved for the more advanced students. The background of those who taught was primarily in history – when Roy Harrod was elected fellow of Christ Church in 1922, it was to a fellowship in history, but he immediately took himself off to Cambridge to study with Keynes, and then on his return arranged for Edgeworth to provide informal graduate supervision. By the later 1920s, with student numbers growing, new appointments were predominantly PPE graduates, among them Henry Phelps-Brown and James Meade in 1930. John Hicks had graduated in 1926 from the PPE, but with a second-class degree and was very fortunate to get taken on at the London School of Economics (LSE), since that institution too was beginning to recruit staff from among the ranks of its own graduates. Oxford lacked the organizational thread that the tripos gave Cambridge economics, and had no central figure to match Keynes, but it was perhaps as a consequence more open to external developments. In 1935 Jacob Marschak, an Oxford lecturer since he had been stripped of his Heidelberg post in 1933, was appointed to a readership in statistics and was made founding Director of the Institute of Statistics. Although, the institute was not the first of such research bodies established in Britain – Manchester's Research Section under John Jewkes preceded it – its foundation predated any plans for Cambridge's own Department of Applied Economics which, delayed by the war, eventually began work in 1945. Also significant is that fact that the Institute was funded externally, by the Rockefeller Foundation, together with a number of new posts in the social sciences. Similarly, Lord Nuffield's benefaction of the later 1930s – he had approached the university with the idea of funding a new engineering college and was persuaded by the then Vice-Chancellor, A.D. Lindsay, of the need for a social science foundation – also provided a focus for collaborative research in economics that Cambridge lacked. In 1941, the Nuffield College Committee established a social reconstruction survey, while the Institute conducted studies on full employment. This complemented work that had been initiated in the mid-1930s by the Oxford Economists' Research Group, again funded with Rockefeller money, which conducted studies of business decision-making and the role of interest rates, this work being published in the first issue of *Oxford Economic Papers* in 1938.

By this time, the *EJ* was being edited from Cambridge by Keynes and Austin Robinson and was widely, and disparagingly, referred to as the *Cambridge Economic Journal*, while the RES had also become closely associated with Cambridge. The LSE had also founded its own journal, *Economica*, in 1920, and with the launch of the 'new series' in 1933 this became a dedicated economics journal. This coincided with the maturation of a style of work distinct from Cambridge, by the mid-1930s condensed into a general scepticism of the significance of Keynes's *General Theory* and what today would be recognized as a strong leaning to neoliberalism. The School had been established in 1895 with a legacy linked to the Fabian Society (Kadish, 1993, p. 230), the common denominator being Sidney Webb and his involvement with commercial education in London. Before the First World War its teaching staff had been predominantly part-time – Cannan, its first professor of economics, retained his part-time status until his retirement in 1926 – but teaching was reorganized during the 1920s, adding a commerce degree to the BSc (Econ.) and replacing part-time with permanent staff recruited from among its own students. Lionel Robbins, appointed to the chair of economics in 1929, and Arnold Plant, who became professor of commerce the following year, were both examples of this trend, Plant gaining a First in economics in 1923 having also been awarded a First in commerce the previous year (Plant read for the commerce

degree as an external student alongside his full-time study of economics). The arrival of Friedrich von Hayek in 1931 as visiting professor confirmed the neoliberal profile that LSE economics assumed from the 1930s to the 1950s, but also the openness of the institution. Cannan's successor as professor had been the Harvard economist Allyn Young, and there was widespread dismay when his early death from pneumonia in 1929 terminated a direct connection to American economists that had been expected to endure for many years.

Likewise, LSE was more catholic in its teaching and reading materials than any other British institution of the time – Frank Knight's *Risk, Uncertainty and Profit* was used as a central text and re-issued in 1933 as No. 16 in the School's reprint series. As a first-year undergraduate in 1948, Bernard Corry recalled being first given sections of Samuelson's *Foundations* to work through, followed by Erich Schneider on the theory of production, and Pallander on location theory (Corry, 1997, pp. 179–80). In a 1937 survey of the School's work, Plant and Robbins noted that Frank Taussig's *Principles of Economics* was a ‘good modern manual’ which, besides specialized sections on public finance, railways and social reorganization, covered much the same ground as the LSE course in economics. Marshall's *Principles* headed the list of works on general economics (Plant and Robbins, 1937, pp. 67, 69). At least part of the differences between Cambridge and LSE economists during the 1930s can be traced to this contrast between an LSE aggressively open to the international development of economics, and a Cambridge which simply assumed that it was in the van of such development and did not therefore need to take account of work elsewhere. Acknowledging her debts on the opening page of *The Economics of Imperfect Competition*, Joan Robinson referred exclusively to Cambridge colleagues – Marshall, Pigou, Sraffa, Kahn, Austin Robinson and Gerald Shove. She did note the contributions to competition theory of Erich Schneider and Heinrich von Stackelberg, but considered that ‘their work is marred by the use of unnecessarily complicated mathematical analysis where simple geometrical methods would serve’ (Robinson, 1933, p. vii).

By the 1930s, Cambridge was graduating 50–60 students from its Part II every year, and well over 100 students left the LSE annually with a BSc (Econ.) containing an increasingly variable amount of economics. The new universities founded from the turn of the century – Birmingham, Manchester, Liverpool and Sheffield – made little direct headway in finding a constituency of students eager to learn the new economics, but they did find a ready market for teaching in commerce, which contained some economics. In most cases this teaching was quite practical, covering law, banking, economic geography, history and languages; and railway management was often an important component, given the size of the railway companies and the numbers of their employees. For many students approaching economics for the first time, it was taught as part of a vocational course that had the support of significant local employers. This was especially true in Scotland, where the four ancient universities – Glasgow, Edinburgh, St. Andrews and Aberdeen – were closer to the Continental European model, law and medicine being a part of the university. Chartered accountants in Scotland took university courses in elementary economics, highlighting a natural link between the professions and the university absent in England.

Ashley returned to Britain from Harvard's new chair in economic history to found Birmingham's Faculty of Commerce in 1902, but although this has become the single most well-known example of commerce teaching in Britain, it was atypical in many ways. Ashley had ambitions for commerce analogous to Marshall's for economics, seeking to educate future management leaders rather than the future line managers and college teachers turned out in Liverpool and Manchester. He established an advisory

board with local business in a deliberate effort to recruit the sons of business families. But instead of drawing on the local business community for the teaching of accounts, commercial law and banking as Liverpool or Manchester had done for many years, Ashley made accounting a professorial position and in 1906 followed this with a chair in finance. These posts were not justified by the student numbers that he recruited. There were never more than 36 students registered for the commerce degree before 1914, and total registrations only averaged in the high fifties once the short-lived post-war boom had passed. Birmingham's later reputation was based not on its early commitment to commerce, but on the coincidence that Frank Hahn, Alan Walters and Terence Gorman all taught there in the mid-1950s. Birmingham, together with Nottingham, was the first British institution to make a significant effort to develop mathematical and statistical analysis in economics.

Manchester was another important centre: it was here that the first university-based research section was established under Jewkes in the early 1930s, and Manchester economists predominated among those recruited to government service during the Second World War. The Faculty of Commerce had been established by Sydney Chapman (a former student of Alfred Marshall) in the late 1903, building upon a solid foundation of teaching in political economy most recently developed by Alfred Flux, but reaching back to Jevons's classes in the 1870s. Degrees were offered in both commerce and honours economics, Chapman using part-time local professionals for the more specialized parts of the commercial curriculum and appointing young economists to do the non-specialized teaching. This strategy enabled him to develop the teaching of economics, and many of the pre-First World War junior staff went on to chair their own departments: Hugh Meredith taught in Manchester 1905–8, and then was professor at Queen's Belfast from 1911 to 1945; Robert Forrester taught in the Faculty 1910–13, went to Aberdeen, then the LSE, and was Professor at Aberystwyth from 1931 to 1951; Harold Hallsworth taught in Manchester during 1910, later becoming Professor at Newcastle; Douglas Knoop taught in 1909, became a lecturer in Sheffield in 1910 and was then later Professor from 1920 to 1948; A.N. Shimmin taught 1913–15, and was from 1945 professor of social science at Leeds. Clearly, Manchester became an important staging post in the development of careers which imposed a clear pattern on the development of the teaching of economics in provincial Britain, and hence by extension the propagation of economic understanding to a diverse range of students.

This pattern in the academic life cycle had important consequences for the advancement of economics in 20th-century Britain. Those appointed to junior posts in this initial phase of pre-First World War expansion quickly moved on to more senior posts as new departments were established, but they then stayed in them for many years. This blocked mobility during the later 1920s and 1930s. But many senior members in this first cohort retired together in mid-century, creating an opening for renewal in the organization of academic economics, reinforced by increased demand for the teaching of economics in the late 1940s. During the immediate post-war period departments expanded to meet this demand; new posts were created, and a fresh wave of young candidates filled senior appointments. These in turn dominated university departments during the 1950s and early 1960s, but reached retirement age at about the same time that new universities were being founded and the number of senior positions extended once more. The pace of development of research and teaching in economics that took place in Britain during the 1960s rested to a considerable degree on the fluidity and openness that this academic life cycle created.

But these two successive surges – in the 1940s and the 1960s – of mobility, expansion and disciplinary development faltered with the uncertainties of the 1970s, and then broke on the university cutbacks of

the 1980s. The mobility and advancement of younger staff trained in the later 1960s and early 1970s was blocked; this cohort grew old together in the same posts while bright young economists looked elsewhere for employment, and for which in any case they did not require to spend several years on a Ph. D. that an academic career now dictated. The average age of departments increased year by year, hollowing out the institutional hierarchy. By the 1990s, the pool of potential young British economists was severely depleted, given the small number of doctoral and postdoctoral students in the system; and with the slow resumption of recruitment the cycle simply skipped a generation expanded the pool from which it drew. Shortlists came to be dominated by applicants from the EU and beyond, attracted by the openness of the UK labour market and the experience of working in the English language. Graduate programmes likewise became dominated by foreign students. As with recruitment to medical staff in the National Health Service, British universities made good the manifest deficiencies of the British educational structure by turning for graduate students and faculty to those trained elsewhere.

The interwar years

The foregoing is not intended to substitute for a more orthodox ‘history of economic thought’ story. It instead demonstrates how the building of a discipline required a financial and institutional framework as a condition for the development of ‘economic careers’, which careers in turn provided the basis for the elaboration of economic argument as spoken, written and published discourse. The first movers in this latter process are indeed generally to be found in Oxbridge and London; but, for a discipline to flourish, followers are also needed, who in turn have access to a secure institutional structure. Hence, the importance of a national perspective upon the development of economics in Britain.

Cambridge did occupy centre stage in the first half of the century, partly as a consequence of the employment opportunities the new tripos presented: students had to be supervised and courses of lectures delivered, and this all added up to a significant number of college fellows and University lecturers. Marshall was also an important spiritual and pedagogic presence – after retirement in 1908, he continued his practice of open hours at home for students, lending them the books that would later form the core of the Marshall Library. His young protégé Arthur Pigou had marked himself out early on with a number of articles in the *EJ* notable for their brevity and formal exposition – anticipations of a style that had not then become customary. His *Wealth and Welfare* broke new ground in seeking to determine what ‘welfare’ might be, and noting that however defined, if the ‘National Dividend’ (as he termed GNP) increased, then welfare also increased. Redistribution of welfare through the population could also be brought about, but given the regressive nature of the contemporary taxation system he thought of this chiefly in terms of access to health and education services. He noted that monopoly tended to distort the distribution of welfare, so that this book also involved an extended treatment of duopoly and imperfect markets. This and the work of Alfred Marshall had a considerable contemporary impact upon American discussion of price and competition, forming a natural background to the later work of Frank Knight and Edward Chamberlin, especially in respect of Pigou's observations on the level of equilibrium output under monopolistic competition (Pigou, 1912, pp. 294, 356). The 1920 revision of this work into *Economics of Welfare* re-emphasized the social duties of the economist as outlined by Marshall in his inaugural lecture of 1887; and a new emphasis is laid upon the impact of taxation, commensurate with the consequences of the war for the post-war economy. The Marshallian cast of the work is highlighted by the following credo from the Preface:

> The complicated analyses which economists endeavour to carry through are not mere gymnastic. They are instruments for the bettering of human life. The misery and squalor that surrounds us, the dying fire of hope in many millions of European homes, the injurious luxury of some wealthy families, the terrible uncertainty overshadowing many families of the poor – these evils are too plain to be ignored. By the knowledge that our science seeks it is possible that they may be restrained. Out of the darkness light! To search for it is the task, to find it, perhaps, the prize, which the 'dismal science of Political Economy' offers to those who face its discipline. (Pigou, 1920, p. vi)

Keynes certainly shared this credo, as his introductory comments to the Cambridge Handbooks show, but his later characterization of Pigou as a 'classical' that is, superseded, economist has subsequently been too easily subsequently accepted at face value. Pigou, being the professor, was debarred from supervising undergraduates, so that his involvement in teaching was limited to lecturing, and this he generally did at an elementary level only. As with many of his generation – D.H. MacGregor in Oxford, Alec Macfie in Glasgow – he had been badly affected by his experiences in the First World War, and played little further part in the shaping of teaching and research in Cambridge. He has consequently, and unjustly, been excluded from 'Cambridge view' of the history of economics, which has come to be dominated instead by Sraffa, Kahn and the Robinsons, amongst others (Collard, 1981).

The *locus classicus* of this Cambridge 'insider story' is George Shackle's *The Years of High Theory*, although curiously Shackle was never a 'Cambridge man': he went to school there, but was never connected with the university. *The Years of High Theory* takes its departure from Sraffa's 1926 *EJ* article, and ascribes to contemporary non-Cambridge economists a dogmatic and universal belief in 'perfect competition'. Hence Sraffa's theoretical critique of perfect competition is presented as a radical, definitive, if unappreciated, settling of accounts, upon which new work can thereafter build. Here Shackle joins later neo-Ricardians, for whom likewise Sraffa is of decisive importance to the development of economic theory. 'Perfect competition' had however only just been systematically adumbrated, in Chapter 6 of Frank Knight's *Risk, Uncertainty and Profit* (1921), and by no means dogmatically; indeed, Shackle imputes to British economists of the 1920s views more common in the America of the later 1940s, and not before.

Dennis Robertson also fails to register in the Cambridge story, despite having Keynes as his Cambridge Director of Studies, and then spending almost his entire working life in Cambridge, retiring in 1957. This neglect can be attributed to his later criticism of Keynes, describing in 1948 the *General Theory* as 'a step backwards' which prematurely embraced 'stagnationism' 'on the strength of one bad depression' (Robertson, 1948, p. xvi). Remarks such as these make his relative neglect all too understandable, but this should not be allowed to obscure the larger significance of his early work. Hitherto studies of economic cycles had focused on the periodicity of price movements (Morgan, 1990, chs. 1, 2); the analysis of *Industrial Fluctuation* went behind price movements to the variations in output and employment that they represented. That bust follows boom was easily accepted; but why a slump should be followed by recovery was not so easy to explain. Robertson identified a number of causes, most important of which was invention and innovation, an emphasis which was new at the time in Britain, and which Robertson had arrived at without having read Joseph Schumpeter (Presley, 1981, pp.

178–9).

Robertson's *Banking Policy and the Price Level* (1926) was likewise an influential work, extending his study of fluctuations to cover monetary phenomena (Laidler, 1999, 93 ff.). Robertson's mannered writing style did not make this book any easier to read, but as Laidler points out, Pigou took over large sections of the argument in his own *Industrial Fluctuations* (1927), disseminating Robertson's ideas in more readable English. As with his first book, Robertson took his departure from observable facts – that the British banking system balanced deposit liabilities against short-term loans. The banking system was therefore charged with coordinating the public's short-term saving with firms requirements for working capital, and although he noted the forced saving involved in this, he also saw its potential as a stabilizing factor, moderate forced saving being therefore the price paid for progress.

Cambridge in the 1930s is however dominated by the figure of Keynes, and not only intellectually. He had resigned his University Lectureship in 1920, after which his formal connection to the university was solely as a college fellow. Nonetheless, he made up for Pigou's disengagement through his editorial work on the *EJ* with Austin Robinson, in the Political Economy Club, to which promising students were invited and required to ask questions of visiting speakers, through his work for the college, and through his engagement in the arts. In Cambridge lectures could be offered by any college fellow, and were not confined to faculty members. Keynes developed a practice of lecturing from the proofs of his next book, the experience obviously leading him to substantial revisions (Rymes, 1989). He found jobs for some bright graduates – while other bright graduates of whom he was unaware found that their Cambridge First might not necessarily lead anywhere in particular (Tribe, 1997, pp. 77, 129).

Keynes's reputation has long been overlaid with ‘Keynesianisms’ of various kinds. That his memorial service in 1946 was held in Westminster Abbey is indication enough that, whatever the nature of his reputation, it was a very great one. Much of his work in the 1920s took the form of superior economic journalism – from *The Economic Consequences of the Peace* (1919) that made his public reputation, through ‘The Economic Consequences of Mr. Churchill’ (1925) to ‘Can Lloyd George Do It?’ (1929). His rise to become the single most influential British economist of the century began in the early 1930s. Peter Clarke has provided a lucid account of the early part of this story: the nature of contemporary government policy, Keynes's evidence to the Macmillan Committee in 1930, its relation to the two volumes of the *Treatise on Money* published that year, the impact of the abandonment of the gold standard in September 1931 and of free trade over the winter of 1931–32, and the consequent genesis of a new *general* theory of employment, interest and money – there is little dispute about the main lines of these developments (Clarke, 1988).

Argument breaks out however over the substance and intentions of the *General Theory*, published in February 1936. David Bensusan-Butt captures precisely the sense of confusion a modern reader experiences coming to this work for the first time:

> Never did a book fall more quickly and more completely into the hands of summarisers, simplifiers, boilers-down, pedagogues and propagandists. To get at what it seemed like at the time (and perhaps what it really was and is) one has to fight one's way through a cloud of commentators, and try to see it in a more empty landscape. (Quoted in Skidelsky, 1992, p. 537)

Notoriously, Keynes was one of the earliest such commentators, reflecting on his intentions in an article in the *QJE* in February 1937. Although few would seriously dispute that the *General Theory* marks the inauguration of an integrated macroeconomics, it was built out of existing elements – and some at least of the disagreements engendered by the book can be related to incompleteness in the integration of these elements. David Laidler has also shown, for example, that one of the most general statements that can be made about the *General Theory* – that it provides a clear role for government not in substituting for market activity, but by influencing the expectations of investors and businessmen – adopts arguments already made in Lavington's *The English Capital Market* (1921) (Laidler, 1999, pp. 87–8).

The translation of Keynes's fluent prose into the diagrams and algebra better suited to an increasingly formalized style of economic argument followed publication very rapidly. Brian Reddaway, reading a review copy of the book on the way to a post at Melbourne University arranged by Keynes, sketched four equations relating savings, income, investment, the rate of interest and the supply of money and published these in the June 1936 issue of *Economic Record* (Reddaway, 1936). On 26 September 1936, at a meeting of the Econometric Society in Oxford, a session was devoted to the *General Theory*. Here Roy Harrod, James Meade and John Hicks made graphical and algebraic presentations, Hicks writing this up in his article ‘Mr. Keynes and the Classics’ published the following year (1937). Thus was born the classroom IS–LM presentation of Keynes's ideas (Young, 1987).

The transformation of the *General Theory* into a blueprint for managing the mixed economy was, however, effected along two separate paths. In the United States Lawrence Klein, Alvin Hansen and finally Paul Samuelson systematized Keynes's insights and rendered them consistent with the new neoclassical economics (Klein, 1948; Hansen, 1953; Samuelson, 1955). In Britain, the outbreak of war in 1939 and the entry of British economists, including Keynes, into government service provided a unique opportunity to deploy Keynes's insights in managing the wartime economy (Cairncross and Watts, 1989, chs. 2–7).

The basic framework had been laid down by Keynes in his ‘How to Pay for the War’, reversing the assumptions upon which the *General Theory* had been built. The basic task now was to run an economy at its maximum potential output for war production without generating inflationary pressures. Such diverse characters as Lionel Robbins, Ronald Coase, Brian Reddaway, John Jewkes, Ely Devons and James Meade were recruited into government service to facilitate the wartime management of the UK economy. Whereas financing the First World War had been primarily a matter of managing international money markets – a task in which Keynes had played a part – ‘paying for the war’ now meant management of the domestic economy. Inflation was to be avoided as a means of suppressing private consumption in favour of war production. Excess purchasing power was instead to be absorbed through additional taxation, which implied estimation of the actual level of excess. A thorough system of rationing was devised, and financial planning increasingly gave way to manpower planning. Allowance had to be made for the subsidies necessary to stabilize the cost of living, and, on the assumption that this stabilized gross incomes, total volume of money demand needed to be established. By subtracting the amount of goods and services coming on the market an ‘inflationary gap’ could be identified, representing the amount of excess demand that had to be siphoned off. As early as the winter of 1940 government treated pressures in the economy in terms of an ‘output gap’ separating the level of demand from the capacity of factors of production to meet these demands (Sayers, 1983, p. 106). The 1941 Budget broke new ground, presented in a national accounting framework that would enable such

estimations to be made (Kaldor, 1941, p. 181). Moreover, this approach implied that the primary economic aim of governments should be the stability and growth of national income, rather than the more narrowly financial considerations traditionally associated with reviews of government income and expenditure. This was underlined by the formulation of post-war plans such as William Beveridge's *Social Insurance and Allied Services* (1942), followed by the *Employment Policy* White Paper of June 1944, the month of the Normandy landings (Coats, 1993a, p. 558). It was this framework that wartime economists bequeathed to the peacetime civil servants who succeeded them, and which enabled them to manage the economy in terms of Keynesian aggregates. The Economic Section, the central body of economic advisors that had been led by Robbins for most of the war, survived the transition to peacetime, but with a much reduced role. Coats notes that fewer than 20 professional economists were employed by the government on matters relating to macroeconomic policy during the first two post-war decades (Coats, 1993b, p. 523).

There have been many versions of Keynesianism since (Backhouse, 2006), but the most misleading variant is that which links Keynes to the centralized management of peacetime mixed economies. Some sort of Keynesian consensus did prevail in the British academic establishment from the later 1940s until the early 1970s, but the overriding concern, which had brought its senior members into the discipline, was a belief that the depression of the 1930s should not be allowed to recur. ‘Keynesianism’ offered a route to a policy synthesis that could realize this, but this was not translated directly into the pursuit of ‘Keynesian’ economic policies on the part of post-war Labour and Conservative governments. The Economic Section was not ineffective in its advice, but it was very small; while academics outside Whitehall lacked direct influence on the formation and execution of policy, chiefly confined in their expression of opinion to the letters’ column of *The Times*. Hugh Gaitskell had been an economics lecturer at University College London and published on capital theory in the *Zeitschrift für Nationalökonomie*, but the Labour Party was never in power during his period of leadership. Harold Wilson likewise came from an Oxford economics background; his incoming Labour Government of 1964 did establish a Department of Economic Affairs, but its chief task was the drafting of a National Plan on the French model. The drafting and execution of legislation right up to the early 1980s was conducted by generalist civil servants with no special background in economics, directed for the most part by Ministers likewise lacking in formal economic training. The ‘Keynesian’ nature of their approach to government and the economy derived not from any particular theoretical beliefs, but chiefly from a generalized public expectation that it was the job of government to counter downturns, stabilize employment and promote growth. Until 1979, any party that denied its capacity to fulfil such electoral expectations stood no chance of gaining office. Harold Wilson observed acutely that ‘Whichever party is in office, the Treasury is in power’, but there is now an extensive literature which documents the essentially pragmatic, rather than dogmatic, nature of Treasury decision-making during the 1950s and 1960s, supposedly the heyday of Keynesianism (Peden, 1988).

The post-war legacy

During the 1930s a number of British economists made theoretical innovations of lasting significance. This was indeed the ‘decade of high theory’, to borrow from George Shackle, but it was certainly not, as he suggests in his book, an exclusively Cambridge preserve. Ronald Coase, who graduated with a

commerce degree from LSE in 1932, went that same year to his first appointment in Dundee, where he drafted his essay identifying a firm as a replacement for market transactions, eventually published in 1937. John Hicks, having published in 1934 an article in which consumer preferences displaced utility, went on in *Value and Capital* (1939) to create a neoclassical microeconomic synthesis. James Meade published in 1936 his *Introduction to Economic Analysis and Policy*, the first of many seminal works. All later gained the Swedish Riksbank Prize in Economic Sciences (in 1991, 1972 and 1977, respectively) for these and other works. But what is most notable about these annual awards, made since 1969 and beginning with Ragnar Frisch and Jan Tinbergen, is that they are dominated by American economists who began their careers in the 1940s and 1950s. For in this period American economics became international economics.

The war itself had turned out to be the apotheosis of British economics. US foreign policy sought to block any prospect that post-war Britain would resume its former world role, and assumed Britain's former international stance as model democracy and proponent of free trade and economic liberty. Teaching of economics in American universities expanded, and during the 1950s graduate programmes were developed on this foundation. There was a parallel expansion in demand for courses in undergraduate economics in Britain, but neither the will nor the money to develop graduate education. Increasingly, bright students and young economists looked to American connections to develop their careers. Coase was already there; Alexander Henderson went from Manchester to Carnegie Mellon in 1950, and became joint author of the first textbook on linear programming; Clive Granger had by the early 1970s gravitated to California. In turn, the teaching of economics in Britain became increasingly modelled upon American programmes, increasingly making use of American books and articles (Backhouse, 1996; 2000).

As already noted, with the end of the war the majority of economists had quickly left government employment and moved back into the university. Economics was widely regarded as a ‘modern’ subject in school and university (Coats, 1993c); educational opportunity was widely understood as the path to social mobility, a belief underwritten by Lionel Robbins's report to the government which argued that extension of university access would not compromise entry standards or teaching (Committee on Higher Education, 1963). This finding coincided with the opening of a number of new universities in which social sciences played a significant role. In 1964, Richard Lipsey moved from the chair at LSE to the founding chair at Essex, primarily because he saw the opportunity to develop the graduate economics programmes there that his colleagues at LSE had declined (Tribe, 1997, 217 ff.). Once established, this model rapidly spread, but then ran into the uncertainties of the 1970s. As economics became more technical, the capacity to train students in the new techniques remained very restricted. Generational succession, as outlined above, also played a role as a new generation, born into the certainties of the 1950s and 1960s, found themselves in an uncertain world.

As Roger Middleton has argued, financial pressure on universities in the later 1970s and 1980s was coupled with a collapse in the public authority of universities (Middleton, 1998, p. 312). Moreover, throughout the 1980s academic economists were, with a few notable exceptions, generally hostile to government policy. Notoriously, this was expressed in a letter to *The Times* in March 1981 where 364 economists signed up to the argument that government policy would deepen the current depression and slow recovery. This polarized politicians and economists, to the lasting cost of the latter (Backhouse, 2000, p. 31). University economists were consequently shut out of government decision-making while at

the same time a broader public found the increasingly technical preoccupations of economists of little relevance to an understanding of economic problems. The broad consensus that had in the 1950s and 1960s made economics the ‘modern’ discipline broke upon widespread popular disillusion with both modern economics and the universities within which it was practised.

The evolutionary development of the discipline was exacerbated by the process of research audit that began in the mid-1980s, ranking departments and their staff on the basis of research publications (the Research Assessment Exercise, RAE). Although this provides for a system of peer review and is not imposed by a separate educational bureaucracy, the resultant ranking was increasingly employed to determine the allocation of resources between and within universities. Furthermore, peer review has tended to sharpen the ‘scientization’ and public isolation of British economics, since ‘professional’ prestige and a high ranking comes only from publication in a very restricted number of international journals, not from an interest either in undergraduate education or in public issues (Middleton, 1998, 221 ff.). Each subject area draws up its own schedule of approved publication media, and in the case of economics this list has always been weighted towards ‘rigour’, which was what economists had come to pride themselves on as compared to the other social sciences. Since these other social sciences were less ‘rigorous’ in their judgement of what counted as worthwhile research outputs, median economics departments assessed in the 2001 RAE fared very badly within social science faculties, losing funding and strengthening the polarizing tendencies which concentrated ‘celebrity’ staff and resources in a handful of institutions.

The trend to internationalization in economics teaching and research was a general phenomenon during the last quarter of the century. The diversity, both between and within nations, with which the discipline had begun the century had, by the early post-war period, increasingly given way to homogenization of style and substance. This process accelerated in the 1980s as the personal computer offered every economist access to data and means for its processing without leaving the office. By contrast, most of Bill Phillips's work on inflation and unemployment in the 1950s had been done late at night on the National Physical Laboratory's computer in Teddington. Likewise, Richard Stone had during the 1940s done most of his own statistical work on a hand-cranked machine. The speed with which data could now be processed did away with the enforced lengthy periods during which one pondered the meaning of previous results and devised new strategies. But it also meant that such thinking was at a discount, given the range of data and software. The discipline of economics succumbed to a basic ‘law’ of markets: the larger the size, the less the diversity.

Nonetheless, public interest in economics survived, and economic careers developed that did not depend upon university status. This new trend originated in the 1980s. Nigel Lawson, Margaret Thatcher's Treasury minister, had a background in economic journalism, symbolizing the rise of a new source of authority independent of any academic institution. Many of the new breed of ‘City economist’ had no formal academic background in economics at all, but drew upon other technical skills. Independent ‘think tanks’ began making themselves heard, foremost among them the Institute for Fiscal Studies (IFS), which by the end of the century had grown into the leading non-government authority on domestic fiscal affairs. The rise of the IFS was accompanied by a number of similar organizations addressing the social, political and economic issues that university economics had for the most part left far behind. And finally, a new, non-academic popular literature of economics emerged, seeking to demonstrate the public utility of economic principles to an increasingly receptive readership.

See Also

- Keynesianism
- Marshall, Alfred

Bibliography

Backhouse, R.E. 1996. The changing character of British economics. In *The Post-1945 Internationalization of Economics*. Annual Supplement to *History of Political Economy*, vol. 28, ed. A. W. Coats. Durham, NC: Duke University Press.

Backhouse, R.E. 2000. Economics in mid-Atlantic British economics 1945–95. In *The Development of Economics in Western Europe since 1945*, ed. A.W. Coats. London: Routledge.

Backhouse, R.E. 2006. The Keynesian revolution. In *The Cambridge Companion to Keynes*, ed. R.E. Backhouse and B.W. Bateman. Cambridge: Cambridge University Press.

Cairncross, A. and Watts, N. 1989. *The Economic Section 1939–1961: A Study in Economic Advising*. London: Routledge.

Chester, N. 1986. *Economics, Politics and Social Studies in Oxford, 1900–85*. London: Macmillan.

Clarke, P. 1988. *The Keynesian Revolution in the Making 1924–1936*. Oxford: Oxford University Press.

Coase, R. 1937. The nature of the firm. *Economica* NS 4, 386–405.

Coats, A.W. 1993a. Some observations on the role of the economist in government. In *The Sociology and Professionalization of Economics. British and American Essays*, vol. 2. London: Routledge.

Coats, A.W. 1993b. Britain: the rise of the specialists. In *The Sociology and Professionalization of Economics. British and American essays*, vol. 2. London: Routledge.

Coats, A.W. 1993c. The market for economists in Britain, 1945–75: a preliminary survey. In *The Sociology and Professionalization of Economics. British and American Essays*, vol. 2, London: Routledge.

Collard, D. 1981. A.C. Pigou, 1877–1959. In *Pioneers of Modern Economics in Britain*, ed. D.P. O'Brien and J.R. Wesley. London: Macmillan.

Committee on Higher Education. 1963. *Higher Education. Report of the Committee appointed by the Prime Minister under the Chairmanship of Lord Robbins 1961–63*. Cmnd. 2154. London: HMSO.

Corry, B. 1997. Bernard Corry. In *Economic Careers. Economics and Economists in Britain 1930–1970*, ed. K. Tribe. London: Routledge.

Groenewegen, P. 1995. *A Soaring Eagle: Alfred Marshall 1842–1924*. Cheltenham: Edward Elgar.

Hansen, A. 1953. *A Guide to Keynes*. New York: McGraw Hill.

Hicks, J.R. 1937. Mr. Keynes and the classics. *Econometrica* 5, 147–59.

Hicks, J.R. 1939. *Value and Capital*. London: Oxford University Press.

Kadish, A. 1982. *The Oxford Economists in the Late Nineteenth Century*. Oxford: Oxford University Press.

Kadish, A. 1993. The city, the Fabians and the foundations of the London school of economics. In *The Market for Political Economy. The Advent of Economics in British University Culture, 1850–1905*, ed. A. Kadish and K. Tribe. London: Routledge.

Kaldor, N. 1941. The White Paper on national income and expenditure. *Economic Journal* 51, 181–91.

Klein, L. 1948. *The Keynesian Revolution*. New York: Macmillan.

Knight, F. 1921. *Risk, Uncertainty and Profit*. New York: Houghton Mifflin Company.

Laidler, D. 1999. *Fabricating the Keynesian Revolution: Studies of the Inter-war Literature on Money, the Cycle, and Unemployment*. Cambridge: Cambridge University Press.

Meade, J.E. 1936. *Economic Analysis and Policy*. London: Oxford University Press.

Middleton, R. 1998. *Charlatans or Saviours? Economists and the British Economy from Marshall to Meade*. Cheltenham: Edward Elgar.

Morgan, M. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Peden, G. 1988. *Keynes, The Treasury and British Economic Policy*. London: Macmillan.

Pigou, A.C. 1912. *Wealth and Welfare*. London: Macmillan.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Pigou, A.C. 1927. *Industrial Fluctuations*. London: Macmillan.

Pigou, A.C. 1925. In memoriam. In *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan.

Plant, A. and Robbins, L. 1937. La London School of Economics. In *L'Enseignement Économique en France et à l'étranger*, Special 50th Anniversary Issue of *Revue d'Économie Politique*, 66–78.

Presley, J.R. 1981. D.H. Robertson, 1890–1963. In *Pioneers of Modern Economics in Britain*, ed. D. O'Brien and J.R. Presley. London: Macmillan.

Reddaway, W.B. 1936. The general theory of employment, interest and money. *Economic Record* 12, 28–36.

Robertson, D. 1948. A study of industrial fluctuation, 1915: new introduction. In *A Study of Industrial Fluctuation*. LSE Reprint No. 8. London.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Rymes, T.K. 1989. *Keynes's Lectures, 1932–35*. Houndmills: Macmillan.

Samuelson, P. 1955. *Economics*, 5th edn. New York: McGraw Hill.

Sayers, R.S. 1983. 1941: the first Keynesian budget. In *The Managed Economy. Essays in British Economic Policy since 1929*, ed. C. Feinstein. Oxford: Oxford University Press.

Shackle, G. 1967. *The Years of High Theory: Invention and Tradition in Economic Thought 1926–1939*. London: Cambridge University Press.

Skidelsky, R. 1992. *John Maynard Keynes Vol. Two. The Economist as Saviour 1920–1937*. London: Macmillan Press.

Tribe, K. 1997. *Economic Careers: Economics and Economists in Britain 1930–1970*. London: Routledge.

Tribe, K. 2000. The Cambridge economics tripos 1903–55 and the training of economists. *Manchester School* 68, 222–48.

Tribe, K. 2001. Economic societies in great Britain and Ireland. In *The Spread of Political Economy and the Professionalisation of Economists*, ed. M. Augello and M. Guidi. London: Routledge.

Young, W. 1987. *Interpreting Mr. Keynes: The IS–LM Enigma*. Oxford: Polity Press.

Young, W. and Lee, F.S. 1993. *Oxford Economics and Oxford Economists*. Basingstoke: Macmillan.

How to cite this article

Tribe, K. "Britain, economics in (20th century)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000328> doi:10.1057/9780230226203.0164

The New Palgrave Dictionary of Economics Online

British classical economics

Mark Blaug

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Classical economics is not just a period in the history of economic thought immediately prior to the marginal revolution but involves a distinct approach to economic problems. But endless controversy surrounds the definition of that approach. Indeed, the scope of the science of political economy as conceived in Smith's *The Wealth of Nations* was sharply contracted in Ricardo's *Principles of Political Economy*. Some modern commentators characterize classical economics as surplus theory; others as general equilibrium theory. Economists who are divided in their views will always try to find those views embodied in the writings of the past.

Keywords

absolute advantage; Austrian economics; Babbage, C.; Bailey, S.; balance of payment; Blaug, M.; British classical economics; Cairnes, J.; capital accumulation; capital–labour ratio; Clark, J. B.; classical economics; comparative advantage; Corn Laws; corn model; Cournot, A.; distribution; Dupuis, A.-J.-E.; Eagly, R.; exchange value; factor endowments; free banking; free trade; general equilibrium theory; Gossen, H.; history of economic thought; Hollander, S.; increasing returns; international trade, theory of; invariable measure of value; Jevons, W.; joint production; Jones, R.; Keynes, J.M.; Knight, F.; labour theory of value; laissez faire; Lloyd, W.; Longfield, M.; Malthus, T.; marginal productivity theory; marginal revolution; Marshall, A.; Marx, K.; McCulloch, J.; Menger, C.; Mill, J.; Mill, J. S.; money; natural price; neoclassical economics; neo-Ricardian economics; normal price; O'Brien, D.; opportunity cost; organic composition of capital (Marx); Petty, W.; physiocracy; Pigou, A.; political economy; population growth; poverty alleviation; preferences; pressure groups; production techniques; productive vs unproductive labour; profit and profit theory; quantity theory of money; Quesnay, F.; Rae, J.; rate of profit; reciprocal demand; rent; reserve army of the unemployed; Ricardian revolution; Ricardo, D.; Say, J.-B.; Say's Law; Schumpeter, J.; Senior, N.; Smith, A.; Sowell, T.; Sraffa, P.; standard composite commodity (Sraffa); static equilibrium analysis; steady-state equilibria; subsistence and subsistence

wages; supply and demand; surplus; Thornton, H.; Thünen, G.; Tooke, T.; Torrens, R.; transformation problem; use value; utility theories of value; vent-for-surplus doctrine; wages fund doctrine; Wakefield, E.; Walrasian theory of general equilibrium; Wicksteed, P.

Article

The label ‘classical economics’ is sometimes employed to refer quite simply to an era in the history of economic thought from, say, 1750 to 1870, in which a group of predominantly British economists used Adam Smith's *Wealth of Nations* as a springboard for analysing the production, distribution and exchange of goods and services in a capitalist economy. So broad a definition of classical economics must include such contemporary Continental writers as Cournot, Dupuit, Thünen and Gossen, not to mention such British writers as Bailey, Lloyd and Longfield, who at first glance seem to stand outside the tradition founded by Adam Smith. It is difficult to resist the implication, therefore, that classical economics is more than a period in the history of economic thought: it seems to involve a definite approach to economic problems. The difficulty, however, is how to characterize this approach. Shrugging aside such tendentious definitions of classical economics as those of Marx and Keynes – for Marx (1867, pp. 174–5n) classical political economy begins with Petty in the 17th century and ends with Ricardo, and for Keynes (1936, p. 3n) the classical school begins with Ricardo and ends with Pigou – the first question is whether it was Adam Smith or David Ricardo who established the ‘essence’ or ‘core’ or classical economics. Of course, Adam Smith laid down the main issues that economists debated for a century after him, but there is also little doubt that the Smithian tradition was in some sense transformed with the appearance of Ricardo's *Principles of Political Economy and Taxation* in 1817. Some writers have nevertheless insisted that Smith and not Ricardo was the lasting influence on the character of classical economics, contending that the leading features of Ricardo's theoretical system were soon rejected even by his avowed followers in the decade after his death in 1823. Others, however, have insisted that, despite all the criticisms of Ricardo that no doubt appeared in the late 1820s and early 1830s, later writers like John Stuart Mill and John Elliott Cairnes continued to operate right up to the 1870s with the central Ricardian theorem that the rate of profit and hence the accumulation of capital depends critically on the marginal cost of production in agriculture; in that sense, they remained trapped in the Ricardian system. But even this assertion presupposes the notion that the Ricardian system is essentially characterized as a theory about the determination of the rate of profit, a proposition which is by no means accepted by all historians of economic thought.

It is only after clearing up this problem of the relative significance of Smith's and Ricardo's ideas in shaping the central current of classical economics that we can take up the question of where to place the utility theories of value put forward by such writers as Lloyd, Longfield, Senior, Dupuit and Gossen, the abstinence theories of interest of Bailey, Senior, Rae and John Stuart Mill, the use of both supply and demand forces in the determination of international prices by Mill, the theory of general gluts and the denial of Say's Law of Markets by Malthus, and the exploitation theory of profits by Marx – in short, all the elements of economic theorizing in the period 1770 to 1870 that so clearly do not belong to the corpus of doctrines bequeathed by Adam Smith and David Ricardo. Likewise, it is only then that we can start talking about the end of classical economics in the 1870s and the nature of the ‘marginal revolution’ that may or may not have marked a decisive break in the continuity of orthodox economics.

The endless debate on what was classical economics is neatly illustrated by the simultaneous appearance of three books on classical economics: *Classical Economics Reconsidered* by Thomas Sowell (1974), *The Structure of Classical Economic Theory* by Robert Eagly (1974) and *The Classical Economists* by Denis O'Brien (1975). Of the three, Eagly takes the widest view of the length of time over which something called ‘classical economic theory’ ruled the roost, beginning with the physiocrats in the 1750s and ending with the Walrasian theory of general equilibrium in the 1870s. His view is not only that the whole of classical economics can be defined in terms of a single conceptual framework but that this framework revolves essentially around a particular concept of capital as a stock of intermediate goods invested in staggered production periods, the question of the pricing of final goods always relegated to the next period after output has already been determined by the size of the labour force and the technology of the previous period; in short, the key to classical economics is to be found in the so-called ‘wages fund doctrine’. Whether this thesis is convincing or not, Eagly's book represents an extreme example of the tendency to define classical economics as one coherent body of ideas organised around a central unifying principle. The secondary literature is, of course, replete with other attempts to pin down once and for all the classical theory of economic growth (e.g. Lowe, 1954; Samuelson, 1978), but few allege, as Eagly does, that their modelling of classical economics captures all the essentials of the writings of Quesnay, Smith, Ricardo, Mill and Marx, as well as McCulloch, Torrens, Bailey, Jones, Senior, Longfield, Babbage, Tooke, Wakefield, etc.

Sowell, on the other hand, adopts the traditional definition of classical economics as in effect the School of Adam Smith, and he therefore excludes Marx and, more surprisingly, Malthus, Torrens and Senior at least in some respects from the mainstream of the tradition stemming from *The Wealth of Nations*. That tradition consisted, according to Sowell, of a common set of philosophical presuppositions, common methods of analysis and common conclusions regarding matters of substantive economic analysis: it comprised such major propositions as the labour theory of value, the Malthusian theory of population, Say's Law and the quantity theory of money and was predominantly oriented towards the issue of economic growth (although not in the modern sense of the term as a theory of the steady-state equilibrium growth path of an economy). However, Sowell admits that this picture has to be qualified after 1817 by such phrases as ‘classical economics in its Ricardian form’ because Ricardo worked a major change in Smith's eclectic mode of economic reasoning by adopting static equilibrium analysis as the only valid method of conducting an economic argument. At any rate, Sowell's treatment of classical economics leaves little doubt of the extensive and varied character of economics in the classical period, posing problems for anyone who seeks to define classical economics in one or two sentences.

Both Eagly's and Sowell's books are dwarfed by O'Brien's wide-ranging and comprehensive review of classical economics, which alone among the three begins with an incisive discussion of the extent to which the classical writers formed a ‘scientific community’. (O'Brien's book also contains excellent annotated bibliographical notes on classical economics; indeed, O'Brien, Blaug (1985) and Spiegel (1983) between them review the whole of the secondary literature.) O'Brien follows Schumpeter in arguing that the Ricardian system represented an analytical detour from the main line of advance running from Adam Smith to John Stuart Mill; it was not a fatal detour, however, because the full Ricardian apparatus attracted hardly any followers and in any case was more or less abandoned by the 1830s. As we noted earlier, this Schumpeter–O'Brien thesis has been questioned by some (e.g. Blaug, 1958; Hollander, 1977). The point is, however, that O'Brien's book perfectly illustrates our contention

that any stand taken on the nature of classical economics as a whole depends critically on the attitude adopted towards the Ricardian metamorphosis of Smithian economics.

The Sraffa interpretation of Ricardo

Still more recently a new note has been struck in the old argument about the essential meaning of classical economics. Inspired by the publication of Sraffa's *Production of Commodities by Means of Commodities* (1960), a number of commentators have argued that classical economics is in effect a Sraffa-system, that is, an analysis of the manner in which a capitalist economy invests its surplus of net output over consumption, which is to say an output in excess of that required to reproduce that level of output, subject to the condition that goods and services are so priced as to maintain a uniform rate of wages and a uniform rate of profit on capital in all lines of investment. This approach, they contend, was buried in the 1870s when the central object of economic analysis became that of investigating the optimum allocation of resources whose quantities are given at the outset of the analysis; in reviving classical surplus analysis, Sraffa not only provides a promising new way of studying economic problems but also illuminates precisely what it was that united Smith, Ricardo and Marx, thus licensing the use of a single label such as ‘classical economics’ to cover them all (see Meek, 1973, 1977, the originator of the argument; and Dobb, 1973; Roncaglia, 1978; Walsh and Gram, 1980; Bradley and Howard, 1982; Eatwell, 1982; Garegnani, 1984; Howard and King, 1985).

As is well known, a Sraffa-system consists of a set of linear production equations, one for each commodity in the economy, and is intended to demonstrate that these equations are sufficient to determine all relative prices in long-run equilibrium irrespective of the pattern of demand, provided that (1) the output of each commodity is given; (2) rate of profit on capital is uniform throughout the economy and (3) the real wage or (alternatively the rate of profit on capital) is somehow determined exogeneously. On the face of it, such a theory does indeed appear to be very much like ‘classical economics’. For example, after distinguishing between ‘natural’ and ‘market’ prices of commodities – or, as we would nowadays say, the long-run and short-run prices of commodities – Adam Smith focused much of his analysis on the determination of ‘natural’ prices, a tendency which became even stronger in the writings of Ricardo. Moreover, Smith and certainly Ricardo, not to mention Marx, always wrote as if demand played no role whatever in the determination of ‘natural’ price. We have all known ever since the work of Marshall that this neglect of demand can be justified if one assumes that commodities are produced under conditions of constant unit costs or constant returns to scale, the long-run supply curves of all industries being perfectly horizontal over the relevant range of output. Sraffa's production equations imply fixed coefficients of production and, again, we have known ever since the work of Leontief that fixed coefficients of production are sufficient (but not necessary) to produce constant costs. In short, Sraffa's demonstration that prices in his model are determined independently of demand is eminently ‘classical’.

Likewise, there is no doubt that the concept of a uniform rate of return on capital, or rather defining ‘natural’ prices to be those generated by a stationary equilibrium in which the rate of profit has become equalized by interindustry mobility of capital, is typical of all economic writing in the century between 1770 and 1870. Finally, the real wage rate in classical economics is determined by so-called ‘subsistence’ requirements and these were defined by Ricardo, Mill and Marx in historical rather than

physiological terms; in other words, it was assumed that the current ‘natural’ price of labour reflected the past history of the ‘market’ price of labour. The ‘natural’ price of labour was in effect determined by workers' attitudes to the size of their families but since the classical economists did little to analyse these attitudes, it is not too much to say that the so-called ‘subsistence *theory* of wages’ actually amounts to taking ‘subsistence’ as a datum (Schumpeter, 1954, p. 665). Once again, it can be argued that the Sraffian assumption of an exogeneous real wage is ‘classical’ in spirit.

There is no doubt that Sraffa's system captures many of the elements of ‘classical economics’. It provides a further bonus, however, in illuminating classical economics. Generations of critics have tried to make sense of Ricardo's lifelong quest for an ‘invariable measure of value’ and have given it up as a hopeless task. Ricardo was troubled by the fact that any change in money wages will alter the structure of relative prices owing to the fact that capital and labour are combined in different proportions in different industries. Thus, a rise in wages or a fall in the rate of profit raises the prices of labour-intensive goods relative to the price of capital-intensive goods. This violates the labour theory of value according to which relative prices are determined by the physical quantities of labour expended on production independently of the rate at which labour is rewarded. To remedy this difficulty, Ricardo struck upon the notion of expressing all prices in terms of a commodity produced by a ratio of capital to labour that is a weighted average of the entire spectrum of capital–labour ratios in the economy; such a commodity, he believed, constitutes an ‘invariable measure of value’ in the sense of providing a standard of measurement that is invariant to changes in the ratio of wages to profits. In the same way, Sraffa measures all prices in terms of a ‘standard composite commodity’ that consists only of outputs combined in the same proportions as the non-labour inputs that enter into all the successive layers of its manufacture. Moreover, in one of the many elegant demonstrations in his book, Sraffa succeeds in showing that such a ‘standard commodity’ is in fact embedded in any actual economic system and that the proportion of net output going to wages in that reduced-scale system determines the rate of profit in the economy as a whole.

The explanation of this result depends on Sraffa's distinction between ‘basic’ commodities which enter directly or indirectly into the production of every commodity in the economy, including themselves, and ‘non-basic’ commodities which enter only into final consumption. If we treat labour itself as a produced ‘means of production’ then wage goods constitute examples of ‘basic’ commodities, that is, they are technically required to cause households to produce the flow of labour services. Ricardo clearly believed that wheaten bread was ‘basic’ in this sense but Sraffa parts company with Ricardo in rejecting any and all versions of the subsistence theory of wages; workers in Sraffa are primary, non–reproducible inputs. Nevertheless, there are plenty of other basics besides wage goods in an actual economy and the upshot of Sraffa's distinction between basics and non-basics is that the ‘standard composite commodity’ consists only of basics and indeed of all the basics in the economy; this collection of basics enters into the production of the invariant yardstick in a ‘standard ratio’, that is, in the same proportion as they enter into their own production. It turns out that relative prices and either the rate of profit or the rate of wages (depending on which one is given exogeneously) depend only on the technical condition of producing the ‘standard commodity’ and are in no way affected by what happens to nonbasic commodities. In a way this is obvious: a change in the cost of producing a nonbasic no doubts alters its own price but, by the definition of a nonbasic commodity, the effect stops there since the product in question never becomes an input into any other technical process. It is also obvious, at least intuitively, that an exogenous change in wages unconnected with a change in productive techniques alters the rate of profit

but has no effect on relative prices measured in terms of the standard commodity for the simple reason that the change alters the measuring rod in the same way as it alters the pattern of prices being measured. The ‘standard commodity’ therefore provides an ‘invariable measure of value’, and Ricardo's old problem is at long last solved.

In developing his own ideas, Sraffa also advanced an entirely new interpretation of how Ricardo came to connect his theory of the determination of the rate of profit with the question of finding an invariable yardstick for measuring relative prices. In his early pamphlet *Essays on the Influence of a Low Price of Corn on the Profits of Stock* (1815), Ricardo wanted to show that the extension of cultivation to inferior soils depresses the rate of profit on capital throughout the economy by raising the marginal cost of producing ‘corn’, that is, wheat, the principal wage good consumed by workers. This is easy to demonstrate in a one-sector economy where the only output is wheat. However, from the beginning Ricardo operated with a two-sector economy in which an agricultural industry produces ‘corn’ and a manufacturing industry produces ‘cloth’. Of course, if wage goods consist entirely of corn and if cloth is always purchased out of profits and rents, it is still easy to show that the rate of profit on capital depends decisively on the action of diminishing returns in agriculture. In agriculture, wheat is the only output and it is also the input both in the form of wages ‘advanced’ to workers to tide them over the annual production cycle and seeds to plough back into the next agricultural cycle; hence, the ‘money’ rate of profit in agriculture cannot possibly diverge from the ‘wheat’ rate of profit because any change in the price of wheat affects inputs and output in the same degree. Manufacturing, however, only uses wheat as one of its inputs (namely, in the form of wage goods), and since the rate of profit earned on capital must be equal in between the two industries in equilibrium, the price of wheat determines a definite price for cloth. If, for example, the rate of profit in agriculture falls due to the operation of diminishing returns, the price of cloth in terms of wheat must likewise fall to prevent cloth from being more profitable to produce than wheat. To reiterate: measuring all prices in terms of wheat, the ‘money’ rate of profit in industry is governed by the ‘wheat’ rate of profit in agriculture, which, in turn, depends entirely on the technology of producing wheat, the unique wage good; in one of Ricardo's famous catch phrases: ‘it is the profits of the farmer which regulate the profits of all other trades’.

This ingenious argument, which appears to explain the determination of the rate of profit in purely physical terms without the use of a theory of value, is known in the literature as the ‘corn model’. In the preface to his edition of *The Works of David Ricardo* (1951), Sraffa argued that the corn model is implicit in Ricardo's 1815 *Essay*. To be sure, Ricardo never wrote it down in so many words because even in the *Essay* he could not swallow the assumption that wages are entirely spend on wheat, that all agricultural products are wage goods and that all manufactured products are luxuries which are never consumed by workers. Nevertheless, he did use wheat in the *Essay* as a measure for aggregating the heterogeneous inputs of agriculture on the assumption that all prices rise and fall with wheat prices, and he also employed arithmetical examples in which all inputs and outputs of both agriculture and manufacturing are expressed in terms of wheat. In the *Principles* he analysed an economy with many sectors in which a change in the terms of trade between wheat and cloth will alter real wages and hence the rate of profit on capital. Nevertheless, his preoccupation in this mature work with the ‘invariable measure of value’ may be read as an attempt to secure the same results obtained earlier with the aid of the corn model, that is, to tie the determination of the rate of profit directly to the production function of agriculture. Of course, if Ricardo could have ignored the varying proportions of labour and capital in different industries, he could have reached all his conclusions without the aid of an invariable yardstick

of value. He had placed so much emphasis, however, on what Marx was to call the unequal ‘organic composition of capital’ that this route was closed to him. Hence, the quest for an ‘invariable measure’ with which to recapture the simple truth of the corn model. Here then is a rational reconstruction of Ricardo's arguments that accounts neatly for both the form and the drift of his reasoning.

A general equilibrium interpretation of Ricardo

Sraffa's interpretation of Ricardo has won wide assent even among those who otherwise remain sceptical about Sraffa's system in its own right. However, Samuel Hollander's recent reexamination of the whole of Ricardo's writings has taken sharp exception to Sraffa's reading (Hollander, 1979, pp. 123–90, 684–9). Ricardo, according to Hollander, never entertained the corn model even implicitly, never assumed that corn alone enters the wage basket, never argued that the rate of profit in agriculture determines the general profit rate and, above all, never assumed that real wages remain constant either because they are determined by the subsistence requirements of workers or because they are determined exogenously. What Hollander really objects to is the notion that ‘distribution’, that is, the rate of wages and the rate of profit, are determined in Ricardo as in Sraffa's own model independently of and indeed prior to the value of commodities, so that the former causally determines the latter. This is to be contrasted with the approach of Walrasian general equilibrium theory in which the pricing of factor services is determined simultaneously with the pricing of final consumption goods. It is simply not true, argues Hollander, that the history of economic thought can be neatly divided into two great branches, a general equilibrium branch leading down from Walras and Marshall to Samuelson, Arrow and Debreu today, in which all relevant economic variables are mutually and simultaneously determined, and a completely different branch leading down from Ricardo and Marx to Sraffa in which distribution takes priority over pricing because economic variables are causally determined in a sequential chain starting from a predetermined real wage (Pasinetti, 1974, pp. 42–4, even enlists Keynes into the ranks of the Ricardo–Marx–Sraffa school). Ricardo, Hollander insists, was essentially a general equilibrium theorist – and so were Adam Smith, John Stuart Mill and even Karl Marx (Hollander, 1973, 1981, 1982).

Before passing judgement on this dispute, it is worth nothing that what has been called the ‘neo-Ricardian’ or ‘Cambridge’ interpretation of the history of economic thought claims superior merit for Ricardo because Ricardo divorced the question of distribution from the question of pricing. But this is precisely the grounds on which many pre-war historians of economic thought attacked Ricardo! Thus, Frank Knight in a famous essay on ‘The Ricardian Theory of Production and Distribution’ (1956) poured scorn on classical writers like Ricardo because they utterly failed to approach the problem of distribution as a problem of valuation and this despite the fact that the effective demand for any factor of production depends on the distribution of income, which in turn depends at least to some extent on the pricing of factor services; in short, ‘distribution theory has little meaning apart from a theory of general equilibrium’ (Knight, 1956, pp. 41, 63). Similarly, Schumpeter (1954, pp. 473, 568–9, 1171) spoke scathingly of the ‘Ricardian Vice’ whereby an already oversimplified economic model is further reduced by freezing one endogeneous variable after another by special *ad hoc* assumptions. First, rent in Ricardo is determined as an intra-marginal return to land treated as a factor in fixed supply; the location of the margin depends of course on the demand for agricultural produce, but this is in turn explained by the size of the population via the assumption of a perfectly inelastic demand for corn. Second, having

‘gotten rid of rent’ on the margins of cultivation, Ricardo then employed a subsistence theory of wages to determine the share of total-output-minus-rent that accrues to labour. Third, total profits in Ricardo are treated as a pure residual after the deduction of wages and rents, the rate of profit being determined as the quotient of total profits and the inherited stock of capital. In other words, the problem of distribution is explained by three totally different types of theories, which in turn are quite different from the principles employed to explain the pricing of goods and services, namely, the labour theory of value. How amazed Knight and Schumpeter would have been to see their critique stood on its head, so that what they regarded as vices are now viewed in certain quarters as virtues.

Ricardo versus Smith

Having expounded various interpretations of classical economics, it is time to attempt some sort of general assessment. To collect our thoughts, consider the number of problematic issues we have outlined above. Is the economics of Adam Smith something different from the economics of David Ricardo? Obviously there is no total break in the continuity of thinking, but nevertheless, is there a sufficient break to warrant the use of such dramatic language as the ‘Ricardian Revolution’? Was this ‘Ricardian Revolution’ the implicit resort to something like the ‘corn model’ to produce a clear-cut explanation of the determination of the rate of profit, or was it simply a change in the style of economic reasoning? Was Ricardo soon repudiated, so that the Smithian tradition survived right down to John Stuart Mill and beyond, or are the later phases of classical economics dominated by the ideas of Ricardo rather than those of Adam Smith? Is there sufficient coherence around a definite core of ideas to permit us to talk at all of ‘classical economics’? Is this core the notion of the origin and disposition of the ‘economic’ surplus and the proposition that distribution is independent of valuation? And, finally, is all of classical economics a primitive but prescient version of general equilibrium analysis?

We can deal quickly with the first question, the so-called ‘Ricardian Revolution’. With the exception of Hollander (1979, ch. 1), all modern commentators on classical economics agree that Ricardo altered the scope, method and focus of economics. Even if we take only *The Wealth of Nations* among Smith's books and essays, the scope of economics for Adam Smith is enormous and perhaps wider than that for any economist before or after him. The first two books of *The Wealth of Nations* consists largely of what later came to be regarded as the very hallmark of orthodox economics: the theory of value and the theory of production and distribution, employing in the main the method of comparative statics. But even the ‘Digression’ on the value of silver in chapter 11 of Book I takes up an unorthodox topic, namely, changes in the structure of prices over centuries with the aid of a method of analysis that might be called ‘inductive’ or ‘historical’. Moreover, here as elsewhere in *The Wealth of Nations* there is a remarkable emphasis on the notion of ‘increasing returns’ so widely defined as to include the effects of both increases in the scale of production and changes in the method of production or technical progress. Despite the flowering of a considerable literature in recent years purporting to model Smith's ‘theory of economic growth’, few have succeeded in capturing this vital element in Smith's thinking, which Kaldor (1972) has consistently emphasized (but see Eltis, 1984, ch. 3). Moreover, this notion of increasing returns soon dropped out of classical economics, coming back only ninety years later with the writings of Karl Marx.

Similarly, there is the famous distinction in Book III of *The Wealth of Nations* between productive and

unproductive labour which Ricardo and Mill accepted, which McCulloch and Senior denied, which Marx reinterpreted in a different way, but which nevertheless was never followed up and developed in any fruitful way. A simple explanation for this failure to elaborate Smith's distinction was that Smith made a mess of it, defining productive labour alternatively as labour which produces something tangible, produces a profit for its employer, and generates productive capacity that then creates a demand for additional employment. But another explanation is that the distinction between the employment of ‘manufacturers’ and ‘menial servants’, between wealth-creating and wealth-consuming activities, is only relevant in the context of long-run economic development, being partly a ‘positive’ account of different patterns of economic change in different nations and partly a ‘normative’ proposal for legislators seeking to maximize the rate of net investment in an economy. Although Mill was profoundly concerned with questions of economic development (see O'Brien, 1975, ch. 8), Ricardo had no real interest in the forces that govern the historical patterns of economic change, and for that reason alone the Smithian distinction between productive and unproductive labour, and the associated discussion of an optimum investment pattern between industries in chapter 5 of Book II of *The Wealth of Nations*, was effectively laid to rest all through the heyday of classical economics.

Smith's interest in ‘the different progress of opulence in different ages of nations’ totally dominates Book III of *The Wealth of Nations* and is at work even in Book IV on mercantilist theory and policy and Book V on public finance. In this latter half of *The Wealth of Nations* there is little appeal to the comparisons of steady-state equilibria, which was to figure so heavily in practically everything that Ricardo wrote. But there are two other elements in these pages that are totally missing in Ricardo and even in Mill, namely, a concern with the incentive effects of different institutional devices for rewarding self-employed professionals and individuals employed in the public sector (Rosenberg, 1960) and a keen sense of the role of pressure groups in the formulation of economic policies (Peacock, 1975; West, 1976; Winch, 1983). Thus, the modern theory of property rights as well as the economic theory of politics may properly claim Smith as a forerunner. At any rate, neither of these two aspects of *The Wealth of Nations* has any echoes in the writings of those that came immediately after Smith.

Consider next the theory of international trade. There is a static equilibrium theory of the gains of foreign trade in Smith based on the principle of absolute rather than comparative advantage, and here no doubt, Ricardo saw further than Smith. But there is also a dynamic theory of the gains of trade in Smith, the so-called ‘vent-for-surplus’ doctrine, according to which foreign trade widens the extent of the market and generates new wants; this view of foreign trade disappears in Ricardo and only comes back to classical economics with Mill (Bloomfield, 1975, 1978, 1981).

Smith's theory of money is also profoundly different from that of Ricardo, typically invoking the quantity theory of money in its dynamic 18th-century version in which the emphasis falls on the disequilibrium ‘transition period’ between an increase in the quantity of money and the rise in prices and not on the final equilibrium adjustment between money and prices (Laidler, 1981). In addition, Smith was an advocate of private, unregulated banking (qualified only by the prohibition of the issue of banknotes for small sums), reflecting the operation of Scottish banking, which was unregulated for over a century between 1716 and 1844. It was Henry Thornton who first rejected the Smithian tradition in his *Paper Credit of Great Britain* (1802), explicitly denying that the note issue in a free banking system would be self-regulating as Smith had argued. By the time of Ricardo it was orthodox to argue that the issue of banknotes was an obvious exception to the doctrine of laissez faire (White, 1984, ch. 3). Here

too, the gulf between Smith and Ricardo is almost total.

There is no need to underline Ricardo's differences with Adam Smith over the labour theory of value, since Ricardo set out explicitly to criticize Smith's failure to apply the labour theory of value to a modern economy rather than a purely conjectural ‘early and rude state of society’. But what is not so obvious is the fact that even in respect of labour as a measure of the ‘real price’ of commodities – Smith's tortured language in Book I, chapter 5, for the problem of specifying an index number of economic welfare – Smith's view of labour is profoundly subjective, whereas Ricardo in his comparable chapter 20 of the *Principles of Political Economy and Taxation* on ‘value and riches’ consistently treats labour as an objective, physical expenditure of energy. In the masterly tenth chapter of Book I of *The Wealth of Nations* on ‘relative wages’, Smith demonstrated that competition in labour markets equalize the net advantages of different occupations, that is, the monetary returns to units of disutility of labour. In other words, to the extent that labour is a ‘measure of value’ in Smith, it is labour conceived as ‘toil and trouble’ and reflects the preferences of workers as much as those of their employers. Although Ricardo, and for that matter Marx, never disputed this analysis of Smith, they ignored its implications and blithely treated labour as fundamentally homogeneous in quality, its role in the production of commodities being conceived as a brute reflection of purely technological data; in short, they took as given something like Sraffa's production equations. It is this and not the famous debate over whether the value of commodities in Smith is determined by the labour ‘commanded’ by goods or the labour ‘embodied’ in their production that represents the real watershed in the history of the labour theory of value (Robertson and Taylor, 1957; Gordon, 1959; Blaug, 1985, pp. 49–53).

But the most profound departure in Ricardo from the Smithian tradition is the notion that rent is in a class by itself as a source of income: it is ‘unearned income’, being an intramarginal return to purely natural differences in the quality of land which have nothing whatever to do with the activity of landlords. Despite Smith's references to landlords who ‘love to reap where they have never sowed’ and the ‘conspiracy’ of merchants, the Smithian world is one in which all economic interests are essentially harmonious or, at any rate, capable of being made harmonious by wise legislators. The Ricardian world, however, is one which conflicting class interests are unavoidable. It is this unique element in the Ricardian system, which gave classical economics its sharp political edge, an edge that clearly worries so many of the minor classical economists, such as Jones, Senior and Longfield.

Finally, the central and indeed sole focus of the Ricardian system is the question: what determines the rate of profit on capital, or rather, what governs its changes over time? This is a question which never really troubled Adam Smith. He made it clear that profit is equalized among industries in the long run, but he had no explanation of how the level of the rate of profit is determined. To be sure, Smith believed that the rate of profit was eventually doomed to fall because of the exhaustion of profitable investment outlets. But he never emphasized this proposition and on balance he took an extremely optimistic view of the feature prospectus for economic growth. Ricardo too was essentially an optimist about the long-run growth potential of the British economy but only if the Corn Laws were repealed; he was thus motivated to argue the strongest possible connection between the rate of profit on capital and the real cost of producing wheat exclusively with domestic resources. In consequence, Ricardo viewed absolutely every aspect of economic activity, including monetary forces, currency arrangements, taxation, the financing of the public debt, and of course foreign trade, through the lenses of his theory of profits. Many readers of Ricardo have been deceived by the preface to his *Principles* – ‘To determine the laws which regulate this distribution (of rent, profit, and wages), is the principal problem in Political

Economy' – into believing that the Ricardian system is largely devoted to an analysis of the determination of the relative shares of land, capital and labour. But while Ricardo certainly had much to say about the issue of relative shares, and indeed was responsible for introducing this theme into economics, his analysis is in fact concentrated on rents per acre, the rate of the profit per unit of capital and the rate of wages per man. It is, in a word, a book about the pricing of factor services and that is (surely?) much less than the subject-matter of *The Wealth of Nations*.

There is little doubt, therefore, that the scope of the science of political economy as conceived in *The Wealth of Nations* was sharply contracted in Ricardo's *Principles of Political Economy*. But, in addition, Adam Smith wrote much besides *The Wealth of Nations*. Quite apart from *The Theory of Moral Sentiments* and the remarkable essay on the *History of Astronomy*, the publication of the new University of Glasgow edition of the complete *Works and Correspondence of Adam Smith* (1976–83) strongly suggests that he intended to round off his contributions by a major work on the theory of jurisprudence which he never lived to write; nevertheless, even in *The Wealth of Nations* he never lost sight of the fact that political economy may be considered as 'a branch of the science of a statesman or legislator', the latter being therefore something more comprehensive than the former. A number of recent commentators (Cropsey, 1957; Lindgren, 1973; Winch, 1978; Skinner, 1979) have indeed insisted that all of Adam Smith's writings are held together by a unified vision of an all-embracing social science, which he unfortunately never succeeded in realizing to the full. Whether this thesis is persuasive or not, it certainly strengthens the contention that the economics of Adam Smith is conceived on grander lines than the economics of David Ricardo.

The corn model again

So there was what might be described in highly coloured language as a 'Ricardian Revolution': what began as a criticism of some of 'Professor Smith's opinions' ended up as a wholesale revision of the legacy of Adam Smith.

What was the cornerstone of this 'Revolution'? Was it the 'corn model'? It certainly was a denial of the Smithian cost-of-production theory according to which a rise in money wages would raise all prices, thus leaving the rate of profits unaffected. But that is not to say that Ricardo's fundamental theorem that 'profits vary inversely as wages' was based on an implicitly held corn model. It is true that the corn-model interpretation neatly rationalizes Ricardo's arguments in the early *Essay on Profits* in which the economy is conceived as consisting of two sectors but the rate of profit is determined exactly as it would be in a one-sector economy. In other words, Ricardo should have held the corn model for without it the *Essay* is simple logically inconsistent. Nevertheless, the corn-model version simply attributes far more rigour and consistency to Ricardo's analysis than is warranted (Peach, 1984). What Ricardo later put in place of the missing corn model was the 'invariable measure of value' which was designed to surmount two of his unresolved difficulties at one and the same time: (1) that workers consume both manufactured and agricultural goods, so that one can never be sure that the rising cost of producing wheat is directly transmitted to the rate of profit; and (2) that capital and labour combine in different proportions in different industries, so that a change in real wages for any reason whatsoever alters the structure of prices and, thus, affects the rate of profit even if nothing has happened to the technology of agriculture. We noted earlier that Sraffa's *Production of Commodities by Means of Commodities* may be said to have vindicated Ricardo's belief in the existence of an 'invariable measure of value', capable of separating

and measuring the effects of changes in technology from those due to changes in the rate of wage and profits. But doubts remain about the validity of this claim. In Ricardo, the divining rod of the invariable measure is supposed to be invariant (as Ricardo kept saying) not just to changes in wages in profits but also to changes in its own methods of production. Sraffa's ‘standard commodity’ fills the bill on the first score but fails on the second score: it is not invariant to changes in its own techniques of production and therefore falls short of solving Ricardo's problem of linking the determination of the rate of profit directly and unambiguously to the action of diminishing returns in agriculture. The truth is that there is no such thing as an ‘invariable’ yardstick that will satisfy all the requirements that Ricardo placed upon it (Ong, 1983). All of which is to say that, despite the fact that Ricardo was the first truly rigorous analytical economist, it is impossible to exonerate him from all analytical errors: he was at times inclined to square a circle using only a ruler and a compass!

Classical economics as surplus theory

We turn next to the thesis that classical economics is the economics of the creation and disposition of surplus output over consumption – a theory of the reproducibility of economic systems in the making – in sharp contrast to the later neoclassical theme of the allocation of *given* resources between competing ends, subject to the constraints of technology and existing property rights. Now, there can be little doubt that this is precisely the nature of the economics of physiocracy (Eltis, 1984, ch. 2), and it is little wonder that those who argue the surplus interpretation include the physiocrats in classical economics (Walsh and Gram, 1980, ch. 2). There is also little doubt that it captures much of the drift of *The Wealth of Nations* and turns up again in Mill's *Principles* and in Marx's *Capital*. On the other hand, it does not begin to do justice to dominant features of the Ricardian system and leaves out almost as much as it manages to include in the writings of the classical economists.

What does it tell us, for example, about the jewel in the crown of classical economics: Ricardo's law of comparative advantage as the foundation of the belief in free trade, which served throughout the whole of the 19th century as the litmus-paper test of an economic liberal? Ricardo treated foreign trade as a matter of moving along a static world production-transformation curve, constructed on the basis of *given* resources and the *given* techniques of production of the trading countries; the gains of foreign trade in his celebrated cloth–wine example show up in a global increase in physical output from given labour resources in Portugal and England. There is no hint here of ‘surplus theory’ and perhaps that is why the surplus interpretation of classical economics studiously avoids discussion of the theory of international trade.

It might be argued, however, that the subject of foreign trade lies outside the mainstream of classical economics because it violates the assumption of a uniform rate of profit on capital – if capital were mobile between countries, international trade would be based like intranational trade on absolute cost advantages. As a matter of fact, Thweatt (1976) has argued that Ricardo's view of foreign trade never went beyond the conception of absolute advantage and this despite the three-paragraph illustration of comparative advantage in his *Principles*, which may well have been written by James Mill rather than Ricardo. After all, free trade for Ricardo meant a policy appropriate to an advanced manufacturing nation in its relation with agrarian nations supplying it with food; the point of the chapter on foreign trade in the *Principles* is not to explain the gains of trade but to demonstrate that foreign trade only

affects the rate of profit insofar as it leads to the importation of cheaper wage goods.

Be that as it may, less than a decade after the death of Ricardo, the young Mill (1844, but written in 1829) completed Ricardo's argument by showing that the division of the overall gains from foreign trade in the two countries depends on ‘reciprocal demand’, thus putting another nail in the coffin of the labour theory of value: even when goods are produced by labour alone within countries, the barter terms of trade between countries depend on both demand and supply. Cairnes subsequently extended the reciprocal demand approach even to domestic trade at least in respect of exchange between ‘non-competing groups’. None of this has anything to do with the creation, accumulation and allocation of an economic surplus, and so the surplus interpretation must leave to one side the classical theory of international prices, the classical theory of balance of payment adjustments and with it the classical theory of monetary management.

But the shortcomings of the surplus interpretation extend even to classical theorizing about the operations of a closed economy. It can throw no light on the care with which Adam Smith spelt out the effects of a public mourning on the price of black cloth in Book I, chapter 7, of *The Wealth of Nations*, so as to demonstrate that ‘market’ prices cannot permanently diverge from ‘natural’ prices because they imply profit opportunities for producers that will sooner or later be exploited; all this is to say that the surplus interpretation has little time for those short-run adjustments that formed the staple of much of the practical wisdom of classical economists grappling with day-to-day economic problems. Similarly, the surplus interpretation must pass over the doctrine of opportunity costs that was part and parcel of the legacy of Adam Smith, namely, that effective costs to producers are not expenditures incurred in the past but present opportunities foregone. As Buchanan (1929) showed many years ago, Ricardo's characteristic doctrine of ‘getting rid of rent’ by concentrating attention on the rentless margin of production implies that land has no uses alternative to the growing of wheat; while this may at a pinch be justified at a macroeconomic level, Smith's theory of rent, which recognizes the fact that land employed in cultivation must compete with land for grazing or urban use, is thus more truly in the tradition of analysing allocation with given resources than is Ricardo's. This Smithian emphasis on the competing uses for land, so that ground rent does enter into the price of agricultural goods, was never lost sight of by classical writers between Ricardo and Mill and comes back into its own in Mill's *Principles*, notably in Book II, chapter 16, on rent theory.

The surplus interpretation is thus a limited view of classical economics, but it is not a misrepresentation. In one sense it is only fancy language for the old view that classical economics is essentially the economics of development, which starts from a fundamental contrast between augmentable labour and non-augmentable land given in quantity and asks how, under these circumstances, growth in the sense of per capita income can be maximized (Myint, 1948). Indeed, the notion that growth of population and the accumulation of capital are the great themes of classical economics in contrast to the question of efficient allocation of given supplies of the factors of production in neoclassical economics after 1870 is endorsed in many, if not in all, textbooks on the history of economic thought (e.g. Blaug, 1985, pp. 295–6). So why all the fuss? Why all this insistence on the surplus interpretation in recent years?

A close reading of those who have advocated a reading of classical economics in terms of surplus analysis suggests two rather different motivations for the ‘new’ interpretation: one is to provide Marx with a respectable pedigree, or at least to display Marx as the true heir of bourgeois economics in its days of glory, solving the riddles that that baffled Quesnay, Smith and Ricardo; the other is to reveal

Sraffa as the true heir of the classical tradition, demonstrating that there is an old and venerable tradition of explaining the determination of prices without resorting to the preferences and satisfactions of consumers and without relying on a market mechanism to price both capital and labour. Each of these two strands of the surplus interpretation produces its own special distortions of classical economics.

It is certainly true that Marx was in many ways a direct descendant of Smith and Ricardo, and particularly of Ricardo. He took over from Smith the distinction between use value and exchange value (as well as the denial that the former had anything to do with the determination of the latter), the distinction between market and natural prices, together with the notion that the business of the economist is to explain natural prices as terminal states of long-run equilibrium outcomes, the distinction between productive and unproductive labour, the conception of historically increasing returns as a major force in the process of development, the tripartite division of national revenue into wages, profits and rents as the incomes of three distinct social classes – and much else. But he learned even more from Ricardo, and particularly Ricardo's discovery that all the problems of the labour theory of value are reducible to the undeniable fact that capital and labour combine in different proportions in different industries, difficulties which may be resolved however by measuring all prices in terms of the price of a commodity produced by the ‘average’ industry. This was the key to Marx's ‘transformation problem’, which demonstrated that ‘prices of production’ must systematically diverge from labour ‘values’ if the rate of profit is to be uniform between industries, an insight which, Marx thought, had always eluded Ricardo. Marx hardly noticed that in correcting Ricardo's answer, he also corrected his question. Ricardo's problem had been: what determines the rate of profit? Marx's problem, however, was: what determines the rate of profit if profit is in the nature of unpaid labour, a mark-up on the outlays of wages disguised as a mark-up on all cost-outlays? But the nature of profit as ‘earned’ or ‘unearned’ income did not interest Ricardo: he devoted one sentence to this subject in the *Principles* and even this sentence was a throw-away remark.

Marx also learned from Ricardo how to reduce skilled labour to common labour by simply taking the structure of relatives wages as given, thus missing the thrust of Smith's theory of relative wages, namely, that wages are not determined solely by the demand side in labour markets. Marx discarded the Malthusian theory of population but retained the subsistence theory of wages relying on the ‘reserve army’ of the unemployed to keep wages fluctuating around subsistence levels. He failed to notice, however, that this made wages a function of the play of demand and supply in labour markets and not the labour-costs of producing wage goods; in short, the pricing of wage goods in Marx does not conform to the labour theory of value. Like Ricardo, Marx conceded that the level of ‘subsistence’ is itself historically conditioned: it is a standard of living that workers have become accustomed to expect by past experience. Thus, even the ‘natural’ price of labour in Marx is not entirely cost-determined but depends on the preferences of workers. Once again, the ‘value of labour-power’ in Marx does not conform to the labour theory of value.

Marx never paid much attention to Ricardo's doctrine of comparative advantage and apparently failed to notice that it too violates the labour theory of value. It is also doubtful whether he ever truly grasped the import of Ricardo's theory of differential rent and particularly its central implication that prices everywhere, and not just in agriculture, are determined by marginal rather than average costs of production.

Nevertheless, despite all the obvious differences between Smith and Ricardo on the one hand and Ricardo and Marx on the other in both analytical constructs and social vision, there are so many striking

similarities between them that Marxian economics is simply unimaginable without Smith, Ricardo and (although Marx did not like to admit it) John Stuart Mill. Marx went further than any of them in his grasp of business cycles, his treatment of technical change and the so-called ‘reproduction schema’ – the true starting point of the modern theory of steady-state growth – but he never emancipated himself from his starting point in classical economics with all its strengths and all its weaknesses.

There can be little quarrel, therefore, with a surplus interpretation of classical economics that treats Marx squarely as one of the last classical economists. However, it is when this Marxian strand in the surplus interpretation is combined with the Sraffian strand that we begin to encounter a mythical classical economics that never existed. We are told that the data for the analysis of prices in classical economics are the same as those for Sraffa, namely, (1) the size and composition of output, (2) the techniques of production in use, and (3) the real wage rate; these are contrasted with the data of neoclassical economics, namely, the preferences of individuals, the initial endowment of the factors of production among individuals and the existing techniques of production (e.g. Eatwell, 1977, p. 62). We are even told that long-run prices in classical theory are *not* the outcome of the opposing forces of demand and supply and that classical ‘natural’ prices are *not* what (ever since Marshall) are called long-run ‘normal’ prices (Harcourt, 1982, p. 265) or that, although classical ‘natural’ prices are indeed the same as neoclassical long-run ‘normal’ prices, the theories advanced by classical and neoclassical economists for the determination of these long-run equilibrium prices are quite different (Garegnani, 1976, pp. 28–9). But there is actually no warrant for any of these assertions.

The size and composition of output is certainly not treated as given in Smith and to say so is to make nonsense of Smith's emphasis on secular economic development and the optimum balance of manufacturing and agriculture in the course of secular growth. Ricardo, on the other hand, frequently but not invariably treats the output of agricultural produce as determined by the size of population via a perfectly inelastic demand for wheat (Barkai, 1965; Stigler, 1965). Thus, he does not assume the output of wheat (or any other product) to be a datum but to be an endogeneously determined variable, a function of population growth, which in turn is treated as an endogeneous variable. He never squarely faced up to all the difficulties created for his argument by commodity-substitution as the price of ‘corn’ rises relative to ‘cloth’, but he certainly recognized the problem. There is no support, therefore, for the contention that he took the composition of output to be a datum, except provisionally at certain points in his argument for the sake of producing what he called ‘strong results’. What we have said about Smith and Ricardo follows with double force for both Mill and Marx. So much then for this part of the attempt to bring the classical economists fully into the Sraffian fold.

We can agree that the classical economists took for granted an existing state of techniques – has there ever been an economist, apart possibly from Marx, who has not? – but the real question is whether they conceived of this state of techniques *à la* Sraffa as ruling out factor substitution. On balance, as we noted earlier, the answer to this question must be yes. Ricardo of course recognized the problem the moment he introduced the chapter on machinery in the third edition of the *Principles* (1821), but by then he was thoroughly committed to his invariable standard of value, which necessarily rules out factor substitution. On the other hand, a special kind of factor substitution was built into his theory of differential rent in which variable doses of capital-and-labour combined in fixed proportions are applied in increasing amounts to a fixed quantity of heterogeneous land; it is this idea which of course led John Bates Clark and Philip Wicksteed in later years to hail Ricardo as the ‘father’ of marginal productivity theory. When we consider that the theory of differential rent was the very cornerstone of the Ricardian system, we can

only gasp at Sraffa's bold declaration in the preface to his *Production of Commodities by Means of Commodities* (1960) that his own system, concerned as it is ‘exclusively with such properties of an economic system as do not depend on changes in the analysis of value production or in the proportions of “factors” is identical to the ‘standpoint … of the old classical economists from Adam Smith to Ricardo’.

Next, can it be argued that the classical economists took the real wage rate as a datum for their analysis of value and distribution? It is perfectly true that the much-maligned theory of subsistence wages in factor amounts to saying that the subsistence wage is whatever has been the real wage for a long time. How long is long? About a generation, Malthus said, and Ricardo agreed. But such assertions did not help much in specifying the subsistence wage, since annual population growth had been positive for as long as anyone could remember, and a positive rate of population growth implied that market wages exceed the natural subsistence wage rate. So, in effect, the classical economists regarded real wages as data but that is not what they thought they were doing; after all, the only reason that the Malthusian theory of population was so quickly incorporated into the mainstream of classical economics was that it appeared to provide a truly endogeneous explanation of the determination of real wages. The long-run equilibrium wage rate, Malthus had taught, was that wage rate, which, given the historically conditioned habits and customs of the working class, encouraged them to reproduce a family of given size. Some classical economists, like Senior and McCulloch, came to doubt the validity of the Malthusian theory but never managed to put any other theory of determination of long-run wages in its place. John Stuart Mill, on the other hand, found the Malthusian theory so suitable for his purpose of alleviating poverty through the self-help of the poor – birth control, education and the formation of consumer and producer cooperatives – that he espoused it more vehemently than even Malthus himself. All in all, there is simply no warrant for arguing that any classical economist (including Marx) *intended* to explain real wages by forces outside the purview of economic analysis.

Lastly, we come to the most grotesque distortion of all: the idea that any appeal to the forces of demand *and* supply in determining prices is necessarily alien to classical economics and that classical ‘natural’ prices have nothing whatsoever in common with Marshall's long-run ‘normal’ prices. Now, it is true that Ricardo (and Marx after him) propagated the misleading idea that demand-and-supply explanations only pertain to ‘market’ prices, whereas ‘natural’ prices are to be explained solely in terms of costs of production, as if costs can influence prices without acting through supply. Ricardo lacked the analytical apparatus to appreciate the fact that supply-side explanations of prices hold only if goods are produced under conditions of constant costs; this might well justify the neglect of demand in the case of the pricing of ‘cloth’ but certainly not on his own grounds in the case of the pricing of ‘corn’. This marvellous confusion of language, encouraged by Ricardo's tendency to think of demand and supply as quantities actually bought and sold and not as schedules of demand and supply prices, was almost entirely cleared up by Mill in his masterful treatment of value in Book III of his *Principles* in which he noted that an equilibrium price is one which equates demand and supply in the sense of a mathematical equation and concluded that ‘the law of demand and supply… is controlled but not set aside by the law of cost of production, since cost of production would have no effect on value if it could have none on supply’. In fact, this is not very different from what Ricardo (1952, Vol. IX, p. 172) once said in private to Jean Baptiste Say: ‘You say demand and supply regulates the price of bread; that is true, but what regulates supply? The cost of production.’

Marshall's schema of market-period, short-period and long-period prices, of constant-cost, increasing-

cost and decreasing-cost industries, and their accompanying diagrams of demand and supply, are indispensable aids to clear thinking about the determination of prices and imply nothing whatsoever about the truth or falsity of any particular theory of prices. To treat demand and supply as dirty words that classical economists would never have employed in the explanation of natural prices is to take their outmoded language at its face value and, indeed, to deny any analytical progress in the history of economics.

To reject Sraffian interpretations of classical economics is not to reject Sraffa's system on its own grounds. Whether or not it is faithful to both the spirit and the letter of classical economics, it is undeniably true that, like all advances in economic theory, it casts a new light on the ideas of the past. It has certainly made us think again about Ricardo's invariable measure of value and its intimate connection with Marx's transformation problem; it has illuminated the problem of joint production and the difficulties which this creates for the labour theory of value, however formulated; and it has highlighted the fact that any theory of prices necessarily involves some proposition about how total output is divided between wages and profits. Its impact on the ongoing debate about the great ideas of the past is perhaps best illustrated by the furore which it has created among Marxian economists, suggesting for example, that the labour theory of value in Marx is both unnecessary and incapable of producing Marx's results (Steedman, 1977, 1981). But to endorse Sraffa's system as a tool for historical exegesis is not to say that it successfully models the essence of classical economics. Smith, Ricardo, Mill and Marx are simply richer than anything captured in *Production of Commodities by Means of Commodities*.

Classical economics as general equilibrium theory

Every extreme reaction produces a counter-reaction. The surplus interpretation of classical economics is a reaction against Marshallian interpretation of classical economics in which Ricardo and Mill are viewed as neoclassical theorists in embryo; for Marshall there was one and only one thread of continuous thought from Adam Smith to his own times (e.g. Marshall, 1890, App. I). In reaction to the surplus interpretation, Hollander has argued that from Ricardo onwards, classical economics was, for all practical purposes, general equilibrium theory; there never was any ‘marginal revolution’. Since this assertion is, to say the least, surprising, let us quote his own words:

> Ricardian economics – the economics of Ricardo and J.S. Mill – in fact comprises in its essentials an exchange system fully consistent with the marginalist elaborations. In particular, their cost–price analysis is pre-eminently an analysis of the allocation of scarce resources, proceeding in terms of general equilibrium, with allowance for final demand, and the interdependence of factor and commodity markets. (Hollander, 1982, p. 590)

It is evident that by ‘general equilibrium theory’, Hollander means a number of interconnected propositions, such as efficient allocation of given resources among alternative uses subject to the principle of diminishing marginal returns, the simultaneous determination of both quantities and relative prices with the aid of the principle of equality between demand and supply, and the consequent interdependence between equilibrium in product and factor markets. Perhaps we have already said

enough to suggest that if this what is meant by general equilibrium theory, there is no sense in which we can subscribe to Hollander's interpretation of classical economics.

Hollander has spelled out his meaning in great detail in a major work on *The Economics of David Ricardo* (1979). In interpreting Ricardo as a general equilibrium theorist, Hollander found himself revising more or less the entire body of Ricardian scholarship, implying that absolutely everybody else before him had radically misinterpreted Ricardo. To convey the flavour of his iconoclasm, consider the following small sample of the extraordinary conclusions of this book (for a complete list, see O'Brien, 1981, pp. 354–5): (1) Ricardo's method of analysis was identical to that of Adam Smith; (2) Ricardo's theory of money was not very different from that of Smith; (3) Ricardo treated the pricing or products and the pricing of factors as fully interdependent; (4) Ricardo's profit theory did not originate in a concern over the Corn Laws, and Ricardo never believed, even in his early writings, that profits in agriculture determine the general rate of profit in the economy; (5) Ricardo's value theory was essentially the same as that of Marshall in that it paid as much attention to demand as to supply, and Ricardo never regarded the invariable measure of value as an important element in his theory; (6) Ricardo could have established his fundamental theorem of the inverse wage–profit relationship without his invariable yardstick and he frequently took the short-cut of assuming identical capital–labour ratios in all industries to give the answers he looked for; (7) wages in Ricardo are never conceived at any time as constant or fixed at subsistence levels; (8) Ricardo never assumed a zero price-elasticity of demand for corn, making the demand for agricultural produce a simple function of the size of population; (9) Ricardo did not predict a falling rate of profit or a rising rental share and never committed himself to any clear-cut predictions about any economic variable; and (10) Ricardo was never seriously concerned about the possibility of class conflict between landowners and everybody else or between workers and capitalists.

There must be something wrong with an interpretation of Ricardo that produces so many conclusions diametrically opposed to what every commentator has found in Ricardo, not only since his death but even while he was still alive. The distortions produced by the surplus interpretation of classical economics are therefore as nothing compared to those generated by Hollander's general equilibrium interpretation.

Walsh and Gram (1980) provide a more reasonable version of the general equilibrium characterization of classical economics: they take the view that general equilibrium analysis encompasses more or less the whole of the history of economic thought, but they distinguish between pre-Walrasian general equilibrium analysis of the allocation of the economic surplus over successive time periods and post-Walrasian general equilibrium analysis of the allocation of given resources within the same time period. One difficulty with their argument is that they never inform the reader what precisely is meant by ‘general equilibrium analysis’. If we mean a discussion of the determination of both product and factor prices which proceeds in terms of an explicit or implicit set of simultaneous equations in order to ensure that the number of unknowns to be determined are equal to the number of equations written down, then obviously classical economics is not general equilibrium analysis: factor pricing in classical economics is invariably explained on different principles from those governing the pricing of products. If we go further and demand that such a discussion must include not just a demonstration of the existence of a unique equilibrium solution for the vector of factor and product prices but also an analysis of the stability and determinacy of the set of equilibrium prices, such as Walras himself struggled to provide, then even more obviously classical economics is not general equilibrium analysis. But what Walsh and

Gram seem to mean by general equilibrium analysis is simply any analysis that involves the simultaneous determination of prices and one distribution variable on the assumption that other factor prices are given; in short, they define general equilibrium analysis to be nothing more nor less than Sraffian economics. Their book therefore collapses the general equilibrium interpretation of classical economics into the surplus interpretation, sharing the deficiencies of both in equal proportions.
Finally, Arrow and Hahn (1971, pp. 1–3) join the fray in the introduction to their textbook on general equilibrium theory. In contrast to Walsh and Gram, they are perfectly explicit about what is meant by general equilibrium theory: if it means anything it implies some notion of both determinateness and stability, that is, the relations describing the economic system are sufficient to determine the equilibrium values of its variables, and a violation of any one of these relations sets in motion forces to restore it. They go on to introduce a new note into the argument: general equilibrium theory is typically associated with the doctrine of unintended consequences – equilibrium outcomes may be and usually are different from those intended by individual actors – and the doctrine that competition is a social mechanism that is capable of achieving a determinate and stable set of equilibrium prices. In all these senses of the term, they count Adam Smith as a ‘creator’ of general equilibrium theory and Ricardo, Mill and Marx as early expositors. They add, however, that there is another sense in which none of the classical economists had a ‘true general equilibrium theory’: no classical economist gave explicit attention to demand as a coordinate element with supply in determining prices, and hence classical economics determined the prices but not the quantities of commodities, the only exception to this statement being their treatment of agricultural output; on the other hand, Mill's theory of foreign trade was ‘a genuine general equilibrium theory’.
To this brief but incisive discussion of the sense in which classical economics is or is not general equilibrium theory, one must add one word of caution: it is the subtle but nevertheless unmistakable difference in the conception of ‘competition’ before and after the ‘marginal revolution’. The modern concept of perfect competition, conceived as a market structure in which all producers are price-takers and face perfectly elastic sales curves for their outputs, was born with Cournot in 1838 and is foreign to the classical conception of competition as a process of rivalry in the search for unrealized profit opportunities, whose outcome is uniformity in both the rate of return on capital invested and the prices of identical goods and services but not because producers are incapable of making prices. In other words, despite a steady tendency throughout the history of economic thought to place the accent on the end-state of competitive equilibrium rather than the process of disequilibrium adjustments leading up to it, this emphasis became remorseless after 1870 or thereabouts, whereas the much looser conception of ‘free competition’ with free but not instantaneous entry to industries is in evidence in the work of Smith, Ricardo, Mill, Marx and of course Marshall and modern Austrians (Stigler, 1957; McNulty, 1967; Littlechild, 1982). For that reason, if for no other, it can be misleading to label classical economics as a species of general equilibrium theory except in the innocuous sense of an awareness that ‘everything depends on everything else’.

Summing up

We have reviewed the recent upswell of new and startling interpretations of classical economics in the light of developments in modern economics, such as the economics of development, growth theory, general equilibrium theory, and Sraffian analysis. In itself there is nothing surprising about this, nor is it

a new phenomenon: every turn and twist in the history of economic thought has always been attended by a fresh look at the past. Marx in propounding his own treatment of the ‘laws of motion’ of capitalism felt impelled to re-examine the ideas of his predecessors over more than a thousand pages. Jevons, Menger and Walras, the triumvirate that is said to have launched the ‘marginal revolution’, accompanied the exposition of their ‘new’ economics by scathing denunciations of the fallacies of classical political economy. Marshall, in seeking unsuccessfully to reconcile a static with a dynamic treatment of economic problems, naturally looked with sympathy at the work of his classical forebears and struggled to depict them as slightly exaggerating one side of the truth in contrast to Jevons, who exaggerated the other. Perhaps therefore the recent proliferation of definitely new but conflicting interpretations of the essential meaning of classical economics is simply an expression of the fact that modern economists are divided in their views and hence quite naturally seek comfort by finding (or pretending that they can find) these same views embodied in the writings of the past.

See Also

- classical growth model
- Marx, Karl Heinrich
- Mill, John Stuart
- Ricardo, David
- Smith, Adam

Bibliography

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Barkai, H. 1965. Ricardo's static equilibrium. *Economica* 32, 15–31.

Blaug, M. 1958. *Ricardian Economics: An Historical Study*. New Haven, CT: Yale University Press.

Blaug, M. 1985. *Economic Theory in Retrospect*. 4th edn. , Cambridge: Cambridge University Press.

Bloomfield, A.I. 1975. Adam Smith and the theory of international trade. In Skinner and Wilson (1975).

Bloomfield, A.I. 1978. The impact of growth and technology of trade in nineteenth-century economic thought. *History of Political Economy* 10, 608–35.

Bloomfield, A.I. 1981. British thought on the influence of foreign trade and investment of growth, 1800–1880. *History of Political Economy* 13, 95–120.

Bradley, I. and Howard, M.C. 1982. *Classical and Marxian Political Economy*. London: Macmillan.

Buchanan, D. 1929. The historical approach to rent and price theory. *Economica* 9(26), 123–55.

Cropsey, J. 1957. *Polity and Economy: An Interpretation of the Principles of Adam Smith*. The Hague: Nijhoff.

Dobb, M. 1973. *Theories of Value and Distribution since Adam Smith*. London: Cambridge University Press.

Eagly, R.V. 1974. *The Structure of Classical Economic Theory*. New York: Oxford University Press.

Eatwell, J. 1977. The irrelevance of returns to scale in Sraffa's analysis. *Journal of Economic Literature* 15(1), 61–8.

Eatwell, J. 1982. Competition. In Bradley and Howard (1982).

Eltis, W. 1984. *The Classical Theory of Economic Growth*. London: Macmillan.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in Modern Capital Theory*, ed. M. Brown, K. Sato and P. Zarembka. Amsterdam: North-Holland.

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 36, 291–325.

Gordon, D.F. 1959. What was the labour theory of value? *American Economic Review* 49, 462–72.

Harcourt, G.C. 1982. The Sraffian contribution: an evaluation. In Bradley and Howard (1982).

Hollander, S. 1973. *The Economics of Adam Smith*. Toronto: University of Toronto Press; London: Heinemann Educational Books.

Hollander, S. 1977. The reception of Ricardian economics. *Oxford Economic Papers* 29, 221–57.

Hollander, S. 1979. *The Economics of David Ricardo*. Toronto: University of Toronto Press; London: Heinemann Educational Books.

Hollander, S. 1981. Marxian economics as ‘general equilibrium’ theory. *History of Political Economy* 13, 121–55.

Hollander, S. 1982. On the substantive identity of the Ricardian and neoclassical conception of economic organization: the French connection in British classicism. *Canadian Journal of Economics* 15,

586–612.

Howard, M.C. and King, J.E. 1985. *The Political Economy of Marx*. 2nd edn., London: Longman.

Kaldor, N. 1972. The irrelevance of equilibrium economics. *Economic Journal* 82(December), 1237–55.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. In *The Collected Writings of John Maynard Keynes*, vol. 7. London: Macmillan, 1973.

Knight, F.H. 1956. *On the History and Method of Economics*. Chicago: University of Chicago Press.

Laidler, D.E.W. 1981. Adam Smith as a monetary economist. *Canadian Journal of Economics* 14(2), May, 187–200.

Lindgren, J.R. 1973. *The Social Philosophy of Adam Smith*. The Hague: Martinus Nijhoff.

Littlechild, S.C. 1982. Equilibrium and the market process. In *Method, Process, and Austrian Economics*, ed. I.M. Kirzner. Lexington, MA: D.C. Heath.

Lowe, A. 1954. The classical theory of economic growth. *Social Research* 21(Summer), 127–58.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Marx, K. 1867. *Capital: A Critique of Political Economy*. Trans. B. Fowkes, Harmondsworth: Penguin Books, 1976.

McNulty, P.J. 1967. A note on the history of perfect competition. *Journal of Political Economy* 75, 395–9.

Meek, R.L. 1967. *Economics and Ideology and Other Essays*. London: Chapman & Hall.

Meek, R.L. 1973. *Studies in the Labour Theory of Value*. 2nd edn. London: Lawrence & Wishart.

Meek, R.L. 1977. *Smith, Marx and After*. London: Chapman & Hall.

Mill, J.S. 1844. *Essays on Some Unsettled Questions Political Economy*. London: London School of Economics, 1948.

Myint, H. 1948. *Theories of Welfare Economics*. Cambridge, MA: Harvard University Press.

Myint, H. 1958. The ‘classical theory’ of international trade and underdeveloped countries. *Economic*

Journal 68, 317–37.

O'Brien, D.P. 1975. *The Classical Economists*. London: Oxford University Press.

O'Brien, D.P. 1981. Ricardian economics and the economics of David Ricardo. *Oxford Economic Papers* 33, 352–86.

Ong, N-P. 1983. Ricardo's invariable measure of value and Sraffa's ‘standard commodity’. *History of Political Economy* 15, 207–27.

Pasinetti, L.L. 1974. *Growth and Income Distribution: Essays in Economic Theory*. Cambridge: Cambridge University Press.

Peach, T. 1984. David Ricardo's early treatment of profitability: a new interpretation. *Economic Journal* 94, 733–51.

Peacock, A. 1975. The treatment of the principles of public finance in *The Wealth of Nations*. In Skinner and Wilson (1975).

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.

Robertson, H.M. and Taylor, W.L. 1957. Adam Smith's approach to the theory of value. *Economic Journal* 67(June), 181–98.

Roncaglia, A. 1978. *Sraffa and the Theory of Prices*. New York: Wiley.

Rosenberg, N. 1960. Some institutional aspects of the *Wealth of Nations*. *Journal of Political Economy* 68, 557–70.

Samuelson, P. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16, 1415–34.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Skinner, A.S. and Wilson, T. 1975. *Essays on Adam Smith*. Oxford: Clarendon Press.

Skinner, A.S. 1979. *A System of Social Science: Papers Relating to Adam Smith*. Oxford: Clarendon Press.

Sowell, T. 1974. *Classical Economics Reconsidered*. Princeton: Princeton University Press.

Spiegel, H.W. 1983. *The Growth of Economic Thought*. Durham, NC: Duke University Press.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

Steedman, I., ed. 1981. *The Value Controversy*. London: Verso Editions and New Left Books.

Stigler, G.J. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65, 1–17.

Stigler, G.J. 1965. Textual exegesis as a scientific problem. *Economica* 33, 447–50.

Thweatt, W.O. 1976. James Mill and the early development of comparative advantage. *History of Political Economy* 8, 207–34.

Walsh, V. and Gram, H. 1980. *Classical and Neoclassical Theories of General Equilibrium*. Oxford: Oxford University Press.

West, E.G. 1976. Adam Smith's economics of politics. *History of Political Economy* 8, 515–39.

White, L.H. 1984. *Free Banking in Britain: Theory, Experience and Debate, 1800–1845*. Cambridge: Cambridge University Press.

Winch, D. 1978. *Adam Smith's Politics: An Essay in Historiographic Revision*. Cambridge: Cambridge University Press.

Winch, D. 1983. Science and the legislator: Adam Smith and after. *Economic Journal* 93, 501–20.

How to cite this article

Blaug, Mark. "British classical economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000157> doi:10.1057/9780230226203.0165

The New Palgrave Dictionary of Economics Online

Brunner, Karl (1916 – 1989)

Allan H. Meltzer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand; Brunner, K.; information costs; interest rate; macroeconomic theory; monetarism; monetary theory; money; money supply; Patinkin, D.; shocks; stocks and flows; transaction costs; uncertainty

Article

Karl Brunner's scholarly contributions are in three areas, namely, monetary and macroeconomics, methodology and its application to cognitive science, and social, political, and institutional analysis. Brunner founded three major journals and organized many conferences, including the Konstanz Seminar in Germany and the Carnegie-Rochester Conference in the United States, which remain current in 2007. Laidler (1991) contains a more complete discussion of Brunner's contributions, and I have relied heavily on his paper. Brunner's own discussion of his intellectual and personal odyssey is in Brunner (1988). I was involved as co-author in much of the work on monetary economics, but I choose to use the pronouns ‘he’ and ‘his’ for this article.

Brunner was born in Zurich, Switzerland, in February 1916. His mother was from the French-speaking region, his father from the German-speaking. They met when both were in Russia working with Russian children. Later his father became the director of the Swiss Observatory. Karl received his doctorate in economics from the University of Zurich in 1943 after spending 1937–38 studying modern economics at the London School of Economics. He travelled to the United States as a Rockefeller Foundation Scholar at Harvard and the University of Chicago from 1949 to 1951. He served on the UCLA faculty from 1951 to 1966 when he left on visiting appointments at Wisconsin and Michigan State before becoming the Everett D. Reese Professor of Economics at Ohio State University. In 1966, he moved to the University of Rochester, where he remained until his death in 1989. From 1979 to 1989, he was the Fred H. Gowen Professor of Economics. During his years at Rochester he served also as Permanent Guest Professor at the University of Konstanz (Germany) from 1968 to 1973 and Professor Ordinarius at the University of Bern (Switzerland) from 1974 to 1985. He arranged for many of his doctoral students at Bern to study at the University of Rochester. This had a lasting influence on economics and finance in Switzerland and

Europe.

Brunner often commented on the gap, often a wide one, between economic policy and economic theory. Much of his research, his efforts to influence policy, his journals and conferences reflected his belief that this gap could be closed by substantive research. Much of his analysis of institutions and the policy process considered the incentives that produced these outcomes and the uncertainty under which policies are made. To properly analyse issues of this kind, he proposed (1987) replacing the ‘economic man’ of the textbooks with the more dynamic and uncertain REMM – resourceful, evaluating, maximizing man. He used REMM also to compare economists’, sociologists’, political scientists’ and psychologists’ ability to understand society's processes.

Macroeconomic theory and monetary theory were his major interests. His earliest work (1951) was a lasting contribution to the early post-war concern with the purely analytic issues raised by Don Patinkin and others as to the determinacy of equilibrium in classical macroeconomics. Brunner developed a stock–flow analysis and devised equilibrium conditions.

Purely formal analysis did not fit well with his developing ideas about methods and the means to scientific development and knowledge in economics. He saw economics as an empirical science that produced refutable hypotheses. He did not reject formal analysis; he no longer did it.

After a few years, he turned to money supply theory. The central idea was to go beyond the standard IS–LM framework in which typically bonds and real capital are perfect substitutes, so that a single interest rate could represent the panoply of relative prices that transmit monetary and other impulses through the economic system. Brunner began by making the interest rate and the money supply endogenous variables. This generation of models was used to reject reverse causation and to critique Federal Reserve policymaking in a study for the US Congress. He proposed an alternative (Brunner and Meltzer, 1964). Subsequent work (Brunner and Meltzer, 1989) introduced an output sector with endogenous prices and output. The complete static model had two endogenous relative prices, base money, bonds and real capital. Adding some institutional detail brought in the money stock and bank credit.

Although anticipated prices appear in these models, price expectations have a minimal role. Responding to the heightened emphasis in the 1970s on expectations and many discussions of stagflation, Brunner, Cukierman and Meltzer (1980; 1983) introduced transitory and persistent shocks into the analysis. This offered an explanation of asset markets requiring at least two relative prices to account for uncertainty of beliefs about the persistence of various impulses. It also offered an explanation of gradual adjustment of wages and employment in response to shocks of uncertain duration. The extended (1983) model introduced price setting and allowed inventories to absorb short-run shocks to aggregate demand.

The role of uncertainty and information was recognized early but took a central position in his monetary theory in ‘The Uses of Money’ (Brunner and Meltzer, 1971). The paper develops the reason that society adopts money, treats money's central role as a medium of exchange and explains why societies converge to a small number, often a single, money. The medium of exchange reduces transaction and information costs, thereby saving resources.

Karl Brunner is known as one of the founders of monetarism, a name he coined for the counter-revolution against Keynesian economics of the 1950s and 1960s.

See Also

- monetarism

Selected works

1951. Inconsistency and indeterminacy in classical economics. *Econometrica* 19, 152–73.

1964. (With A.H. Meltzer.) *The Federal Reserve's Attachment to the Free Reserve Concept*. Washington: House Committee on Banking and Currency.

1971. (With A.H. Meltzer.) The uses of money: money in the theory of an exchange economy. *American Economic Review* 71, 784–805.

1980. (With A. Cukierman and A.H. Meltzer.) Stagflation, persistent unemployment, and the permanence of economic shocks. *Journal of Monetary Economics* 6, 467–92.

1983. (With A. Cukierman and A.H. Meltzer.) Money and economic activity, inventories and business cycles. *Journal of Monetary Economics* 11, 281–319.

1987. The perception of man and the conception of society: two approaches to understanding society. *Economic Inquiry* 25, 367–88.

1988. My philosophy. *American Economist*.

1989. (With A.H. Meltzer.) *Money and the Economy: Issues in Monetary Analysis*. The 1987 Raffaele Mattioli Lectures. Cambridge: Cambridge University Press.

Bibliography

Laidler, D. 1991. Karl Brunner's monetary economics: an appreciation. *Journal of Money, Credit, and Banking* 23, 633–58.

How to cite this article

Meltzer, Allan H. "Brunner, Karl (1916–1989)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000320> doi:10.1057/9780230226203.0166

The New Palgrave Dictionary of Economics Online

Bruno, Michael (1932 – 1996)

Joseph Zeira
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bruno, M.; Falk Institute for Economic Research (Israel); inflation; intertemporal approach to the balance of payments; oil and the macroeconomy; Sachs, J.; stabilization policy; stagflation; World Bank

Article

Michael Bruno was born in Germany in 1932 and emigrated with his family to Israel in 1933. After military service he studied mathematics and economics at the Hebrew University of Jerusalem and at King's College, Cambridge. On returning to Israel, he worked at the research department of the Bank of Israel. In 1961 he was brought to Stanford University by Hollis Chenery and Kenneth Arrow, where he received his Ph.D. in 1962. He then returned to Israel and in 1963 joined the faculty of the Department of Economics at the Hebrew University of Jerusalem. Over the years he visited MIT, Harvard, the University of Stockholm, and the LSE. Many times during his academic career Michael Bruno was involved in economic policymaking. In the mid-1970s he participated in a tax reform in Israel and advised the government on economic policy. In 1985 he was chief advisor to the Israeli disinflation programme. From 1986 to 1991 he was Governor of the Bank of Israel, and between 1993 and 1996 he served as a Senior Vice-President and Chief Economist of the World Bank.

Michael Bruno's research covered many areas in macroeconomics, was both theoretical and empirical, but was always strongly related to the economic problems of the time. In the 1960s, living in a rapidly developing country, he studied economic growth and development, focusing on input–output analysis and on duality in growth theory. In the 1970s, following the oil shocks, he began to study the macroeconomics of open economies, especially their reaction to shocks. One outcome of this research contains a pioneering discussion of the important ‘intertemporal approach to the balance of payments’ (1976). Another outcome is the research conducted with Jeffrey Sachs on stagflation and supply shocks, which culminated in their important book on stagflation (1985). In the 1980s, influenced by high inflation in Israel and by his role in the Israeli stabilization programme of 1985, Bruno's attention turned to inflation and stabilization. His research then reflected his deep interest in issues of disinflation and of reform in general. He promoted the idea that creating consensus is important for the

success of reforms, and applied it also to the analysis of the post-Communist transition in eastern Europe. In the 1990s Michael Bruno served in the World Bank and there his focus returned to issues of development, which he had studied in the beginning of his career. Actually, he combined it with his deep understanding of inflation and studied how inflation affects economic growth. His main finding appears in a paper with Easterly (1998) that shows that high inflation has a strong negative effect on growth. Thus, his last period of life and of economic research saw a closing of a circle, where he synthesized knowledge that he had accumulated throughout his scientific career, to analyse this important issue.

In addition to his general research and to his effect on policymaking, Michael Bruno also contributed significantly to research on the Israeli economy, both through his research and through his roles as director of the research department in the Bank of Israel, as director of the Falk Institute for Economic Research in Israel, and as Governor of the Bank of Israel.

See Also

- factor price frontier
- growth models, multisector
- inflation
- oil and the macroeconomy

Selected works

1962. *Interdependence, Resource Use and Structural Change in Israel*. Jerusalem: Bank of Israel.

1962. (With H. Chenery.) Development alternatives in an open economy: the case of Israel. *Economic Journal*72, 79–103.

1966. (With E. Burmeister and E. Sheshinski.) The nature and significance of the reswitching of techniques. *Quarterly Journal of Economics*80, 526–53.

1969. Fundamental duality relations in the pure theory of capital and growth. *Review of Economic Studies*36, 49–62.

1976. The two-sector open economy and the real exchange rate. *American Economic Review*66, 566–77.

1977. (With Y. Ben-Porath.) The political economy of a tax reform. *Journal of Public Economics*7, 285–307.

1980. Import prices and stagflation in the industrial countries: a cross-section analysis. *Economic Journal*90, 479–92.

1984. Raw materials, profits and the productivity slowdown. *Quarterly Journal of Economics*99, 1–29.

1985. (With J. Sachs.) *Economics of Worldwide Stagflation*. Cambridge, MA: Harvard University Press.

1986. External shocks and domestic response: Israel's macroeconomic performance, 1965–1982. In *The Israeli Economy: Maturing through Crisis*, ed. Y. Ben-Porath. Cambridge, MA: Harvard University Press.

1986. Sharp disinflation strategy: Israel 1985. *Economic Policy*1(2), 379–407.

1989. Econometrics and the design of economic reform (Presidential address). *Econometrica*57, 275–306.

1990. (With S. Fischer.) Seigniorage, operating rules and the high inflation trap. *Quarterly Journal of Economics*105, 353–74.

1993. *Crisis, Stabilization and Economic Reform: Therapy by Consensus*. Clarendon Lectures in Economics. Oxford: Clarendon Press.

1998. (With W. Easterly.) Inflation crises and long-run growth. *Journal of Monetary Economics*41, 3–26.

1998. (With M. Ravallion and L. Squire.) Equity and growth in developing countries: old and new perspectives on the policy issues. In *Income Distribution and High-Quality Growth*, ed. V. Tanzi. Cambridge, MA: MIT Press.

How to cite this article

Zeira, Joseph. "Bruno, Michael (1932–1996)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000319> doi:10.1057/9780230226203.0167

The New Palgrave Dictionary of Economics Online

bubbles in history

Charles P. Kindleberger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bank lending; booms; bubbles; German inflation (1923); hedge finance; Kindleberger, C. P.; Law, J.; Mississippi bubble; Ponzi finance; railway mania; rational expectations; South Sea bubble; speculative finance; tulipmania

Article

A bubble may be defined loosely as a sharp rise in price of an asset or a range of assets in a continuous process, with the initial rise generating expectations of further rises and attracting new buyers – generally speculators interested in profits from trading in the asset rather than its use or earning capacity. The rise is usually followed by a reversal of expectations and a sharp decline in price often resulting in financial crisis. A boom is a more extended and gentler rise in prices, production and profits than a bubble, and may be followed by crisis, sometimes taking the form of a crash (or panic) or alternatively by a gentle subsidence of the boom without crisis.

Bubbles have existed historically, at least in the eyes of contemporary observers, as well as booms so intense and excited that they have been called ‘manias’. The most notable bubbles were the Mississippi bubble in Paris in 1719–20, set in motion by John Law, founder of the *Banque Générale* and the *Banque Royale*, and the contemporaneous and related South Sea bubble in London. Most famous of the manias were the Tulip mania in Holland in 1636, and the Railway mania in England in 1846–7. It is sometimes debated whether a particular sharp rise and fall in prices, such as the German hyperinflation from 1920 to 1923, or the rise and fall in commodity and share prices in London and New York in 1919–21, the rise of gold of $850 an ounce in 1982 and its subsequent fall to the $350 level, were or were not bubbles. Some theorists go further and question whether bubbles are possible with rational markets, which they assume exist (see e.g. Flood and Garber, 1980).

Rational expectations theory holds that prices are formed within the limits of available information by market participants using standard economic models appropriate to the circumstances. As such, it is claimed, market prices cannot diverge from fundamental values unless the information proves to have been widely wrong. The theoretical literature uses the assumption of the market having one mind and

one purpose, whereas it is observed historically that market participants are often moved by different purposes, operate with different wealth and information and calculate within different time horizons. In early railway investment, for example, initial investors were persons doing business along the rights of way who sought benefits from the railroad for their other concerns. They were followed by a second group of investors interested in the profits the railroad would earn, and by a third group, made up of speculators who, seeing the rise in the railroad's shares, borrowed money or paid for the initial instalments with no intention of completing the purchase, to make a profit on resale.

The objects of speculation resulting in bubbles or booms and ending in numerous cases, but not all, in financial crisis, change from time to time and include commodities, domestic bonds, domestic shares, foreign bonds, foreign shares, urban and suburban real estate, rural land, leisure homes, shopping centres, Real Estate Investment Trusts, 747 aircraft, supertankers, so-called ‘collectibles’ such as paintings, jewellery, stamps, coins, antiques etc. and, most recently, syndicated bank loans to developing countries. Within these relatively broad categories, speculation may fix on particular objects – insurance shares, South American mining stocks, cotton-growing land, Paris real estate, Post-Impressionist art, and the like.

At the time of writing, the theoretical literature has yet to converge on an agreed definition of bubbles, and on whether they are possible. Virtually the same authors who could not reject the no-bubbles hypothesis in the German inflation of 1923 one year, managed to do so a year later (Flood and Garber, 1980). Another pair of theorists has demonstrated mathematically that rational bubbles can exist after putting aside ‘irrational bubbles’ on the grounds not of their non-existence but of the difficulty of the mathematics involved (Blanchard and Watson, 1982).

Short of bubbles, manias and irrationality are periods of euphoria which produce positive feedback, price increases greater than justified by market fundamentals, and booms of such dimensions as to threaten financial crisis, with possibilities of a crash or panic. Minsky (1982a, 1982b) has discussed how after an exogenous change in economic circumstances has altered profit opportunities and expectations, bank lending can become increasingly lax by rigorous standards. Critical exception has been taken to his taxonomy dividing bank lending into hedge finance, to be repaid out of anticipated cash flows; speculative finance, requiring later refinancing because the term of the loan is less than the project's payoff; and Ponzi finance, in which the borrower expects to pay off his loan with the proceeds of sale of an asset. It is objected especially that Carlo Ponzi was a swindler and that many loans of the third type, for example those to finance construction, are entirely legitimate (Flemming, Goldsmith and Melitz, 1982). Nonetheless, the suggestion that lending standards grow more lax during a boom and that the banking system on that account becomes more fragile has strong historical support. It is attested, and the contrary rational-expectations view of financial markets is falsified, by the experience of such a money and capital market as London having successive booms, followed by crisis, the latter in 1810, 1819, 1825, 1836, 1847, 1857, 1866, 1890, 1900, 1921 – a powerful record of failing to learn from experience (Kindleberger, 1978).

See Also

- tulipmania

Bibliography

Blanchard, O. and Watson, M.W. 1982. Bubbles, rational expectations and financial markets. In *Crises in the Economic and Financial Structure*, ed. P. Wachtel. Lexington, Mass.: Heath.

Flemming, J.S., Goldsmith, R.W. and Melitz, J. 1982. Comment. In *Financial Crises: Theory, History and Policy*, ed. C.P. Kindleberger and J.-P. Laffargue. Cambridge: Cambridge University Press.

Flood, R.P. and Garber, P.M. 1980. Market fundamentals versus price-level bubbles: the first tests. *Journal of Political Economy* 88, 745–70.

Kindleberger, C.P. 1978. *Manias, Panics and Crashes: A History of Financial Crises*. New York: Basic Books.

Minsky, H.P. 1982a. *Can 'It' Happen Again?: Essays on Instability and Finance*. Armonk: Sharpe.

Minsky, H.P. 1982b. The financial instability hypothesis. In *Financial Crises: Theory, History and Policy*, ed. C.P. Kindleberger and J.-P. Laffargue. Cambridge: Cambridge University Press.

How to cite this article

Kindleberger, Charles P. "bubbles in history." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000212> doi:10.1057/9780230226203.0169

The New Palgrave Dictionary of Economics Online

bubbles

Markus K. Brunnermeier
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bubbles refer to asset prices that exceed an asset's fundamental value because current owners believe they can resell the asset at an even higher price. There are four main strands of models: (i) all investors have rational expectations and identical information, (ii) investors are asymmetrically informed and bubbles can emerge because their existence need not be commonly known, (iii) rational traders interact with behavioural traders and bubbles persist since limits to arbitrage prevent rational investors from eradicating the price impact of behavioural traders, (iv) investors hold heterogeneous beliefs, potentially due to psychological biases, and agree to disagree about the fundamental value.

Keywords

arbitrage; asset-pricing models; asymmetric information; autocorrelation; backward induction; bubbles; centipede game; central limit theorems; co-integration; efficient markets hypothesis; fiat money; gains from trade; hedge funds; limited liability; noise traders; overlapping generations model; rational expectations; risk aversion; transversality condition; unit roots

Article

Bubbles are typically associated with dramatic asset price increases followed by a collapse. Bubbles arise if the price exceeds the asset's fundamental value. This can occur if investors hold the asset because they believe that they can sell it at a higher price than some other investor even though the asset's price exceeds its fundamental value. Famous historical examples are the Dutch tulip mania (1634–7), the Mississippi Bubble (1719–20), the South Sea Bubble (1720), and the ‘Roaring ‘20s’ that preceded the 1929 crash. More recently, up to March 2000 Internet share prices (CBOE Internet Index) surged to astronomical heights before plummeting by more than 75 per cent by the end of 2000.
Since asset prices affect the real allocation of an economy, it is important to understand the circumstances under which these prices can deviate from their fundamental value. Bubbles have long

intrigued economists and led to several strands of models, empirical tests and experimental studies. We can broadly divide the literature into four groups. The first two groups of models analyse bubbles within the rational expectations paradigm, but differ in their assumption as to whether all investors have the same information or are asymmetrically informed. A third group of models focuses on the interaction between rational and non-rational (behavioural) investors. In the final group of models traders' prior beliefs are heterogeneous, possibly due to psychological biases, and consequently they agree to disagree about the fundamental value of the asset.

Rational bubbles under symmetric information

Rational bubbles under symmetric information are studied in settings in which all agents have rational expectations and share the same information. There are several theoretical arguments that allow us to rule out rational bubbles under certain conditions. Tirole (1982) uses a general equilibrium reasoning to argue that bubbles cannot exist if it is commonly known that the initial allocation is interim Pareto efficient. A bubble would make the seller of the 'bubble asset' better off, which – due to interim Pareto efficiency of the initial allocation – has to make the buyer of the asset worse off. Hence, no individual would be willing to buy the asset. Partial equilibrium arguments alone are also useful in ruling out bubbles. Simply rearranging the definition of (net) return, $r_{t+1,s} := (p_{t+1,s} + d_{t+1,s}) / p_t - 1$, where $p_{t,s}$ is the price and $d_{t,s}$ is the dividend payment at time t and state s, and taking rational expectations yields

$$p_t = E_t\left[\frac{p_{t+1} + d_{t+1}}{1 + r_{t+1}}\right]. \tag{1}$$

That is, the current price is just the discounted expected future price and dividend payment in the next period. For tractability assume that the expected return that the marginal rational trader requires in order to hold the asset is constant over time, $E_t[r_{t+1}] = r$, for all t. In solving the above difference equation forward, that is, in replacing p_{t+1} with $E_{t+1}[p_{t+2} + d_{t+2}] / (1 + r)$ in eq. (1) versus Equation (2) below and then p_{t+2} and so on, and using the law of iterated expectations, one obtains after $T - t - 1$ iterations

$$p_t = E_t\left[\sum_{\tau=1}^{T-t} \frac{1}{(1+r)^{\tau}} d_{t+\tau}\right] + E_t\left[\frac{1}{(1+r)^{T-t}} p_T\right].$$

The equilibrium price is given by the expected discounted value of the future dividend stream paid from $t+1$ to T plus the expected discounted value of the price at T. For *securities with finite maturity*, the price after maturity, say T, is zero, $p_T = 0$. Hence, the price of the asset, p_t, is unique and simply coincides with the expected future discounted dividend stream until maturity. Put differently, finite horizon bubbles cannot arise as long as rational investors are unconstrained from selling the desired number of shares in all future contingencies. For *securities with infinite maturity*, $T \to \infty$, the price p_t only coincides with the expected discounted value of the future dividend stream, call it fundamental value, v_t, if the so-called transversality condition, $\lim_{T \to \infty} E_t\left[\frac{1}{(1+r)^{T-t}} p_T\right] = 0$, holds. Without imposing the transversality condition, $p_t = v_t$ is only one of many possible prices that solve the above expectational difference equation. Any price $p_t = v_t + b_t$, decomposed in the fundamental value, v_t, and a bubble component, b_t, such that

$$b_t = E_t\left[\frac{1}{(1+r)} b_{t+1}\right], \tag{2}$$

is also a solution. Equation (2) versus eq. (1) needs to be made consistent. Equation (2) highlights that the bubble component b_t has to 'grow' in expectations exactly at a rate of r. A nice example of these 'rational bubbles' is provided in Blanchard and Watson (1982), where the bubble persists in each period only with probability π and bursts with probability $(1-\pi)$. If the bubble continues, it has to grow in expectation by a factor $(1+r)/\pi$. This faster bubble growth rate (conditional on not bursting) is necessary to achieve an expected growth rate of r. In general, the bubble component may be stochastic. A specific example of a stochastic bubble is an intrinsic bubble, where the bubble component is assumed to be deterministically related to a stochastic dividend process.

The fact that any bubble has to grow at an expected rate of r allows one to eliminate many potential rational bubbles. For example, a positive bubble cannot emerge if there is an upper limit on the size of the bubble. That is, for example, the case with potential bubbles on commodities with close substitutes. An ever-growing 'commodity bubble' would make the commodity so expensive that it would be substituted with some other good. Similarly, a bubble on a non-zero net supply asset cannot arise if the required return r exceeds the growth rate of the economy, since the bubble would outgrow the aggregate wealth in the economy. Hence, bubbles can only exist in a world in which the required return is lower than or equal to the growth rate of the economy. In addition, rational bubbles can persist if the pure existence of the bubble enables trading opportunities that lead to a different equilibrium allocation. Fiat money in an overlapping generations (OLG) model is probably the most famous example of such a bubble. The intrinsic value of fiat money is zero, yet it has a positive price. Moreover, only when the price is positive, does it allow wealth transfers across generations (that might not even be born yet). A negative bubble, $b_t<0$, on a limited-liability asset cannot arise since the bubble would imply that the asset price has to become negative in expectation at some point in time. This result, together with eq.

(2), implies that if the bubble vanishes at any point it has to remain zero from that point onwards. That is, rational bubbles can never emerge within an asset-pricing model; they must already be present when the asset starts trading.

Empirically testing for rational bubbles under symmetric information is a challenging task. The literature has developed three types of tests: regression analysis, variance bounds tests and experimental tests. Initial tests proposed by Flood and Garber (1980) exploit the fact that bubbles cannot start within a rational asset-pricing model and hence at any point in time the price must have a non-zero part that grows at an expected rate of r. However using this approach, inference is difficult due to an exploding regressor problem. That is, as time t increases, the regressor explodes and the coefficient estimate relies primarily on the most recent data points. More precisely, the ratio of the information content of the most recent data point to the information content of all previous observations never goes to zero. This implies that as time t increases, the time series sample remains essentially small and the central limit theorem does not apply. Diba and Grossman (1988) test for bubbles by checking whether the stock price is more explosive than the dividend process. Note that if the dividend process follows a linear unit-root process (for example, a random walk), then the price process has a unit root as well. However the change in price, Δp_t, and the spread between the price and the discounted expected dividend stream, $p_t - d_t/r$, are stationary under the no-bubbles hypothesis. That is, p_t and d_t/r are co-integrated. Diba and Grossman test this hypothesis using a series of unit root tests, autocorrelation patterns, and co-integration tests. They conclude that the no-bubble hypothesis cannot be rejected. However, Evans (1991) shows that these standard linear econometric methods may fail to detect the explosive nonlinear patterns of periodically collapsing bubbles. West (1987) proposes a different test that exploits the fact that one can estimate the parameters needed to calculate the expected discounted value of dividends in two different ways. One way of estimating them is not affected by the bubble, the other is. Note that the accounting identity (1) can be rewritten as

$p_t = \frac{1}{1+r}(p_{t+1} + d_{t+1}) - \frac{1}{1+r}(p_{t+1} + d_{t+1} - E_t[p_{t+1} + d_{t+1}])$. Hence, in an instrumental variables regression of p_t on $(p_{t+1} + d_{t+1})$ – using for example d_t as an instrument – one obtains an estimate for r that is independent of the existence of a rational bubble. Second, if, for example, the dividend process follows a stationary AR(1) process, $d_{t+1} = \varphi d_t + \eta_{t+1}$, with independent noise η_{t+1}, one can easily estimate ϕ. Furthermore, the expected discounted value of future dividends is $v_t = (\varphi / (1 + r - \varphi))d_t$. Hence, under the null-hypothesis of no bubble, that is $p_t=v_t$, the coefficient estimate of the regression of p_t on d_t provides a second estimate of $\varphi / (1 + r - \varphi)$. In a final step, West uses a Hausman specification test to test whether both estimates coincide. He finds that the US stock market data usually reject the null hypothesis of no bubble.

Excessive volatility in the stock market seems to provide further evidence in favour of stock market bubbles. LeRoy and Porter (1981) and Shiller (1981) introduced variance bounds that indicate that the stock market is too volatile to be justified by the volatility of the discounted dividend stream. However, the variance bounds test is controversial (see, for example, Kleidon, 1986). Also, this test, as well as all the aforementioned bubble tests, assumes that the required expected returns, r, are constant over time. In a setting in which the required expected returns can be time-varying, the empirical evidence favouring excess volatility is less clear-cut. Furthermore, time-varying expected returns can also rationalize the

long-horizon predictability of stock returns. For example, a high price–dividend ratio predicts low subsequent stock returns with a high R^2 (Campbell and Shiller, 1988).

Finally, it is important to recall that the theoretical arguments that rule out rational bubbles as well as several empirical bubble tests rely heavily on backward induction. Since a bubble cannot grow from time T onwards, there cannot be a bubble of this size at time $T-1$, which rules out this bubble at $T-2$, and so on. However, there is ample *experimental evidence* that individuals violate the backward induction principle. Most convincing are experiments on the centipede game (Rosenthal, 1981). In this simple game, two players alternatively decide whether to continue or stop the game for a finite number of periods. On any move, a player is better off stopping the game than continuing if the other player stops immediately afterwards, but is worse off stopping than continuing if the other player continues afterwards. This game has only a single subgame perfect equilibrium that follows directly from backward induction reasoning. Each player's strategy is to stop the game whenever it is his or her turn to move. Hence, the first player should immediately stop the game and the game should never get off the ground. However, in experiments players initially continue to play the game – a violation of the backward induction principle (see for example, McKelvey and Palfrey, 1992). These experimental findings question the theoretical reasoning used to rule out rational bubbles under symmetric information. More experimental evidence on bubbles in general is provided in the final section.

In a rational bubble setting an investor only holds a bubble asset if the bubble grows in expectations ad infinitum. In contrast, in the following models an investor might hold an overpriced asset if he thinks he can resell it in the future to a less informed trader or someone who holds biased beliefs. In Kindleberger's (2000) terms, the investor thinks he can sell the asset to a greater fool.

Asymmetric information bubbles

Asymmetric information bubbles can occur in a setting in which investors have different information, but still share a common prior distribution. In these models prices have a dual role: they are an index of scarcity and informative signals, since they aggregate and partially reveal other traders' aggregate information (see for example Brunnermeier, 2001 for an overview). In contrast to the symmetric information case, the presence of a bubble need not be commonly known. For example, it might be the case that everybody knows the price exceeds the value of any possible dividend stream, but it is not the case that everybody knows that all the other investors also know this fact. It is this lack of higher-order mutual knowledge that makes it possible for finite bubbles to exist under certain necessary conditions (Allen, Morris and Postlewaite, 1993). First, it is crucial that investors remain asymmetrically informed even after inferring information from prices and net trades. This implies that prices cannot be fully revealing. Second, investors must be constrained from (short) selling their desired number of shares in at least one future contingency for finite bubbles to persist. Third, it cannot be common knowledge that the initial allocation is interim Pareto efficient, since then it would be commonly known that there are no gains from trade and hence the buyer of an overpriced 'bubble asset' would be aware that the rational seller gains at his expense (Tirole, 1982). In other words, there have to be gains from trade or at least some investors have to think that there might be gains from trade. There are various mechanisms that lead to these. For example, fund managers who invest on behalf of their clients can gain from buying overpriced bubble assets, since trading allows them to fool their clients into believing that they have

superior trading information. A fund manager who does not trade would reveal that he does not have private information. Consequently, bad fund managers churn bubbles at the expense of their uninformed client investors (Allen and Gorton, 1993). Furthermore, fund managers with limited liability might trade bubble assets due to classic risk-shifting incentives, since they participate on the potential upside of a trade but not on the downside risk.

Bubbles due to limited arbitrage

Bubbles due to limited arbitrage arise in models in which rational, well-informed and sophisticated investors interact with behavioural market participants whose trading motives are influenced by psychological biases. Proponents of the ‘efficient markets hypothesis’ argue that bubbles cannot persist since well-informed sophisticated investors will undo the price impact of behavioural non-rational traders. Thus, rational investors should go against the bubble even before it emerges. The literature on limits to arbitrage challenges this view. It argues that bubbles can persist, and provides three channels that prevent rational arbitrageurs from fully correcting the mispricing. First, *fundamental risk* makes it risky to short a bubble asset since a subsequent positive shift in fundamentals might *ex post* undo the initial overpricing. Risk aversion limits the aggressiveness of rational traders if close substitutes and close hedges are unavailable. Second, rational traders also face *noise trader risk* (DeLong et al., 1990). Leaning against the bubble is risky even without fundamental risk, since irrational noise traders might push up the price even further in the future and temporarily widen the mispricing. Rational traders with short horizons care about prices in the near future in addition to the long-run fundamental value and only partially correct the mispricing. For example, in a world with delegated portfolio management, fund managers are often concerned about short-run price movements, because temporary losses instigate fund outflows (Shleifer and Vishny, 1997). A temporary widening of the mispricing and the subsequent outflow of funds force fund managers to unwind their positions exactly when the mispricing is the largest. Anticipating this possible scenario, mutual fund managers trade less aggressively against the mispricing. Similarly, hedge funds face a high flow-performance sensitivity, despite some arrangements designed to prevent outflows (for example, lock-up provisions). Third, rational traders face *synchronization risk* (Abreu and Brunnermeier, 2002, 2003). Since a single trader alone cannot typically bring the market down by himself, coordination among rational traders is required and a synchronization problem arises. Each rational trader faces the following trade-off: if he attacks the bubble too early, he forgoes profits from the subsequent run-up caused by behavioural momentum traders; if he attacks too late and remains invested in the bubble asset, he will suffer from the subsequent crash. Each trader tries to forecast when other rational traders will go against the bubble. Timing other traders’ moves is difficult because traders become sequentially aware of the bubble, and they do not know where in the queue they are. Because of this ‘sequential awareness’, it is never common knowledge that a bubble has emerged. It is precisely this lack of common knowledge that removes the bite of the standard backward induction argument. Since there is no commonly known point in time from which one could start backward induction, even finite horizon bubbles can persist. The other important message of the theoretical work on synchronization risk is that relatively insignificant news events can trigger large price movements, because even unimportant news events allow traders to synchronize their sell strategies. Unlike the earlier limits to arbitrage models, in which rational traders do not trade

aggressively enough to completely eradicate the bubble but still short an overpriced bubble asset, in Abreu and Brunnermeier (2003) rational traders prefer to ride the bubble rather than attack it. The incentive to ride the bubble stems from a predictable 'sentiment' in the form of continuing bubble growth.

Empirically, there is supportive evidence in favour of the 'bubble-riding hypothesis'. For example, between 1998 and 2000 hedge funds were heavily tilted towards highly priced technology stocks (Brunnermeier and Nagel, 2004). Contrary to the efficient markets hypothesis, hedge funds were not a price-correcting force even though they are among the most sophisticated investors and are arguably closer to the ideal of 'rational arbitrageurs' than any other class of investors. Similarly, Temin and Voth (2004) document that Hoares Bank was profitably riding the South Sea bubble in 1719–20, despite giving numerous indications that it believed the stock to be overvalued. Many other investors, including Isaac Newton, also tried to ride the South Sea bubble but with less success. Frustrated with his trading experience, Isaac Newton concluded 'I can calculate the motions of the heavenly bodies, but not the madness of people' (Kindleberger, 2005, p. 41).

Heterogeneous beliefs bubbles

Bubbles can also emerge when investors have heterogeneous beliefs and face short-sale constraints. Investors' beliefs are heterogeneous if they start with different prior belief distributions that can be due to psychological biases. For example, if investors are overconfident about their own signals, they have a different prior distribution (with lower variance) about the signals' noise term. Investors with non-common priors can agree to disagree even after they share all their information. Also, in contrast to an asymmetric information setting, investors do not try to infer other traders' information from prices. Combining heterogeneous beliefs with short-sale constraints can result in overpricing since optimists push up the asset price, while pessimists cannot counterbalance it since they face short-sale constraints (Miller, 1977). Ofek and Richardson (2003) link this argument to the Internet bubble of the late 1990s. In a dynamic model, the asset price can even exceed the valuation of the most optimistic investor in the economy. This is possible, since the currently optimistic investors – the current owners of the asset – have the option to resell the asset in the future at a high price whenever they become less optimistic. At that point other traders will be more optimistic, and hence be willing to buy the asset since optimism is assumed to oscillate across different investor groups (Harrison and Kreps, 1978). It is essential that less optimistic investors, who would like to short the asset, are prevented from doing so by the short-sale constraint. Heterogeneous belief bubbles are accompanied by large trading volume and high price volatility (Scheinkman and Xiong, 2003).

Experimental evidence

Many theoretical arguments in favour of or against bubbles are difficult to test with (confounded) field data. Laboratory experiments have the advantage that they allow the researcher to isolate and test specific mechanisms and theoretical arguments. For example, the aforementioned experimental evidence on centipede games questions the validity of backward induction. There is a large and growing literature that examines bubbles in a laboratory setting. For example, Smith, Suchanek and Williams (1988) study

a double-auction setting, in which a risky asset pays a uniformly distributed random dividend of $d \in \{0, d_1, d_2, d_3\}$ in each of the 15 periods. Hence, the fundamental value for a risk-neutral trader is initially $15 \sum_i \frac{1}{4} d_i$ and declines by $\sum_i \frac{1}{4} d_i$ in each period. Even though there is no asymmetric information and the probability distribution is commonly known, there is vigorous trading, and prices initially rise despite the fact that the fundamental value steadily declines. More specifically, the time-series of asset prices in the experiments are characterized by three phases. An initial boom phase is followed by a period during which the price exceeds the fundamental value, before the price collapses towards the end. These findings are in sharp contrast to any theoretical prediction and seem very robust across various treatments. A string of subsequent articles show that bubbles still emerge after allowing for short sales, after introducing trading fees, and when using professional business people as subjects. Only the introduction of futures markets and the repeated experience of a bubble reduce the size of the bubble. Researchers have speculated that bubbles emerge because each trader hopes to outwit others and to pass the asset on to some less rational trader in the final trading rounds. However, more recent research has revealed that the lack of common knowledge of rationality is not the cause of bubbles. Even when investors have no resale option and are forced to hold the asset until the end, bubbles still emerge (Lei, Noussair and Plott, 2001).

In summary, the literature on bubbles has taken giant strides since the 1970s that led to several classes of models with distinct empirical tests. However, many questions remain unresolved. For example, we do not have many convincing models that explain when and why bubbles start. Also, in most models bubbles burst, while in reality bubbles seem to deflate over several weeks or even months. While we have a much better idea of why rational traders are unable to eradicate the mispricing introduced by behavioural traders, our understanding of behavioural biases and belief distortions is less advanced. From a policy perspective, it is interesting to answer the question whether central banks actively try to burst bubbles. I suspect that future research will place greater emphasis on these open issues.

See Also

- behavioural finance
- Kindleberger, Charles P.
- South Sea bubble
- speculative bubbles
- tulipmania

Bibliography

Abreu, D. and Brunnermeier, M.K. 2002. Synchronization Risk and Delayed Arbitrage. *Journal of Financial Economics* 66, 341–60.

Abreu, D. and Brunnermeier, M.K. 2003. Bubbles and crashes. *Econometrica* 71, 173–204.

Allen, F. and Gorton, G. 1993. Churning bubbles. *Review of Economic Studies* 60, 813–36.

Allen, F., Morris, S. and Postlewaite, A. 1993. Finite bubbles with short sale constraints and asymmetric information. *Journal of Economic Theory* 61, 206–29.

Blanchard, O.J. and Watson, M.W. 1982. Bubbles, rational expectations, and financial markets. In *Crisis in the Economic and Financial Structure*, ed. P. Wachtel. Lexington, MA: Lexington.

Brunnermeier, M.K. 2001. *Asset Pricing under Asymmetric Information: Bubbles, Crashes, Technical Analysis and Herding*. Oxford: Oxford University Press.

Brunnermeier, M.K. and Nagel, S. 2004. Hedge funds and the technology bubble. *Journal of Finance* 59, 2013–40.

Campbell, J.Y. and Shiller, R.J. 1988. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195–228.

DeLong, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98, 703–38.

Diba, B.T. and Grossman, H.I. 1988. The theory of rational bubbles in stock prices. *Economic Journal* 98, 746–54.

Evans, G.W. 1991. Pitfalls in testing for explosive bubbles in asset prices. *American Economic Review* 81, 922–30.

Flood, R.P. and Garber, P.M. 1980. Market fundamentals versus price-level bubbles: the first tests. *Journal of Political Economy* 88, 745–70.

Harrison, J.M. and Kreps, D. 1978. Speculative investor behavior in a stock market with heterogeneous expectations. *Quarterly Journal of Economics* 89, 323–36.

Kindleberger, C.P. 2005. *Manias, Panics and Crashes: A History of Financial Crises*, 5th edn. New york: Wiley.

Kleidon, A.W. 1986. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94, 953–1001.

Lei, V., Noussair, C.N. and Plott, C.R. 2001. Nonspeculative bubbles in experimental asset markets: lack of common knowledge of rationality vs. actual irrationality. *Econometrica* 69, 831–59.

LeRoy, S.F. and Porter, R.D. 1981. The present value relation: tests based on implied variance bounds.

Econometrica 64, 555–74.

McKelvey, R.D. and Palfrey, T.R. 1992. An experimental study of the centipede game. *Econometrica* 60, 803–36.

Miller, E.M. 1977. Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32, 1151–68.

Ofek, E. and Richardson, M. 2003. DotCom mania: the rise and fall of Internet stocks. Working Paper No. FIN-01-037 58(3), 1113–38. New York University, Stern School.

Rosenthal, R. 1981. Games of perfect information, predatory pricing, and the chain-store paradox. *Journal of Economic Theory* 25, 92–100.

Scheinkman, J. and Xiong, W. 2003. Overconfidence and speculative bubbles. *Journal of Political Economy* 111, 1183–219.

Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71, 421–36.

Shleifer, A. and Vishny, R.W. 1997. The limits of arbitrage. *Journal of Finance* 52, 35–55.

Smith, V.L., Suchanek, G.L. and Williams, A.W. 1988. Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica* 56, 1119–51.

Temin, P. and Voth, H.-J. 2004. Riding the South Sea Bubble. *American Economic Review* 94, 1654–68.

Tirole, J. 1982. On the possibility of speculation under rational expectations. *Econometrica* 50, 1163–82.

West, K.D. 1987. A specification test for speculative bubbles. *Quarterly Journal of Economics* 102, 553–80.

How to cite this article

Brunnermeier, Markus K. "bubbles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000278> doi:10.1057/9780230226203.0168

The New Palgrave Dictionary of Economics Online

Buchanan, David (1779 – 1848)

Andrew Skinner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Buchanan, David; Smith, Adam; taxation

Article

Buchanan was born in Montrose, the eldest son of David Buchanan, the renowned printer, publisher and amateur literary scholar. Unlike his father, David the younger did not attend university, but entered the family business. Primarily interested in economics, geography and statistics, Buchanan is generally regarded as a journalist and writer, but also as a ‘Scottish economist’ (*Encyclopedia of the Social Sciences*, 1935, iii. 27). Buchanan's career amply justifies all of these claims.

Invited by Francis Horner and Francis Jeffrey to act as editor for the short-lived *Weekly Register* in 1808, Buchanan moved to the *Caledonian Mercury* two years later and remained in this post until 1827. In the same year he became editor of the *Edinburgh Courant*, a position he held until his sudden death in 1848.

In 1835 Buchanan helped to compile the *Edinburgh Geographical Atlas*, and made a number of contributions to the *Edinburgh Gazetteer*. He also contributed pieces on geography and statistics to the seventh edition of the *Encyclopaedia Britannica* (1842), which were acknowledged in the preface. But the bulk of Buchanan's output was on economics, with numerous articles appearing in Cobbett's *Political Register* and in the *Edinburgh Review*. The latter in particular carried pieces on ‘Lord Henry Petty's plan of Finance’ (1807), ‘Wheatley on money and finance’ (1807), ‘Spence on agriculture and commerce’ (with Francis Jeffrey, 1809), the Corn Laws (1815), and ‘Corn and money’ (1816).

This growing interest in economic subjects prepared Buchanan for his critical, annotated edition of the *Wealth of Nations* (1814), which in turn paved the way for his *Observations on the Subjects Treated of in Dr. Smith's Inquiry* published in the same year. In the Introduction to the latter work Buchanan set Smith's achievement in the context of the work done by Sir James Steuart and the physiocrats. While expressing qualified admiration for both, Buchanan noted that the *Wealth of Nations* ‘is a great display of reason on the business of the world; touching society in all its essential relations, containing lessons for government as well as for common life, and embracing subjects formerly placed without the limits of

philosophy' (1814, p. viii).
Yet Dr Smith had 'not published a perfect work'. The critical 'dissertations' which follow supplement the notes with the intention of correcting 'what is amiss' (p. xv).
Less successful in his treatment of Ricardo, Buchanan elaborated on the determinants of price and criticized Smith's theory of rent. Other subjects covered included metallic money and paper currency, wages, stock, productive and unproductive labour, the progress of opulence, the Corn Laws, commercial treaties, defence, public debt and the East India Company.
Buchanan included a section on taxation and went on to publish an *Inquiry into the Taxation and Commercial Policy of Great Britain* (1844) which subsequently attracted some critical acclaim.
In 1852 Buchanan was described as a man of 'unobtrusive habits, mild and gentle in his demeanour, and held in high respect by all who had an opportunity of forming an estimate of his character' (Anderson, 1863, p. 481).

Selected works

1814. (ed.) *Inquiry into the Nature and Causes of the Wealth of Nations*, in three volumes; to which is added *Observations on the Subjects Treated of in Dr. Smith's Inquiry* (1814), Edinburgh: Oliphant, Waugh & Innes; London: John Murray.

1844. *Inquiry into the Taxation and Commercial Policy of Great Britain, with Observation on the Principles of Currency and of Exchangeable Value*. Edinburgh: W. Tait.

Bibliography

Anderson, W. 1863. *The Scottish Nation*. Edinburgh: A. Fullarton.

David Buchanan. In *Dictionary of National Biography*, Vol. VII. London: Smith, Elder & Co., 1886.

David Buchanan. In *Encyclopedia of the Social Sciences*, Vol. III. New York: Macmillan, 1935.

David Buchanan. In *Dictionary of Political Economy*, ed, R.H. Inglis Palgrave, Vol. I. London: Macmillan, 1894.

Wellesley Index to Victorian Periodicals, 1824–1900. Ed. W.E. Houghton, Toronto: Toronto University Press 1966.

How to cite this article

Skinner, Andrew. "Buchanan, David (1779–1848)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_B000214> doi:10.1057/9780230226203.1854

The New Palgrave Dictionary of Economics Online

Buchanan, James M. (born 1919)

Richard E. Wagner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The customary Anglo-Saxon approach to public finance treats the state as exogenous to the economic process, which restricts public finance to the study of market-based reactions to exogenous fiscal impositions. In contrast, Buchanan has cultivated an approach to public finance that incorporates the state into the economic process. The domain of fiscal analysis is thus expanded in two directions. One direction, public choice, involves the study of the effect of political institutions on collective choices. The other direction, constitutional political economy, involves the emergence of and changes in political institutions.

Keywords

bequest motive; Buchanan, J.; constitutional political economy; constitutional vs. post-constitutional behaviour; decision costs; external costs; fiscal institututions; free rider problem; Knight, F.; majoritarianism; progressive and regressive taxation; public choice; public debt; Ricardian equivalence; state, size of; strategic behaviour; unanimity; voting rules; Wicksell, K.

Article

James M. Buchanan was awarded the 1986 Nobel Memorial Prize in Economic Science for his seminal role in developing ‘the contractual and constitutional bases for the theory of economic and political decision-making’.
Buchanan spent his boyhood in rural Tennessee near Murfreesboro. After receiving Bachelor's and Master's degrees from Middle Tennessee State College and the University of Tennessee respectively, he entered the US Navy in 1941. After completing his naval service in the Pacific, Buchanan enrolled at the University of Chicago in 1946, receiving his Ph.D. in 1948. He has spent the preponderance of his academic career at three Virginia universities: the University of Virginia (1956–68), Virginia Polytechnic Institute (1969–83), and George Mason University (since 1983). Buchanan has been a truly

prolific scholar throughout this period, as shown by the 20 volumes of his collected works published by Liberty Fund; moreover, he has continued his scholarly work at full speed since the completion of that collection in 2001.

The Nobel citation referred to above identifies two predominant strains within Buchanan's scholarly oeuvre. One of these is the theory of public choice, which entails the application of economic theorizing to politics. The other is constitutional political economy, which explores the relationship between constitutional rules and political outcomes. While Buchanan's body of work also contains numerous contributions to economic theory and methodology, which by themselves would have constituted a significant scholarly career, this short article focuses exclusively on Buchanan's approach to public choice and constitutional political economy.

Precursory influences

While Buchanan has been creative as well as prolific, he has nonetheless been inspired by, and has built upon, the contributions of others. Buchanan has acknowledged these precursory influences numerous times, particularly in his autobiographical *Better than Plowing*, where he identifies three sources of primary influence on his work.

The primary precursors to Buchanan's public choice theorizing were a set of Italian scholars, among them Antonio De Viti De Marco, Maffeo Pantaleoni, and Luigi Einaudi, who developed a unique orientation towards public finance between the 1880s and the 1930s. Where Anglo-Saxon scholars treated the state as outside the economy, the Italians sought to incorporate political outcomes into the economic process. For instance, much Anglo-Saxon fiscal scholarship sought to develop norms regarding the desirable degree of tax progressivity, as illustrated by various sacrifice theories of taxation. By contrast, the Italians sought to explain the actual structure of taxation independently of normative concern, and to do so with reference to the same categories of utility and cost as they invoked to explain market outcomes. This Italian orientation of sober realism towards political processes was central to the later development of public choice theorizing. For instance, in his foreword to the German translation of Amilcare Puviani's 1903 treatise on fiscal illusion, *Teoria della illusione finanziaria*, Gunter Schmölders observed that ‘over the last century Italian public finance has had an essentially political science character…. This work [Puviani] is a typical product of Italian public finance…. Above all, it is the science of public finance combined with fiscal politics, in many places giving a good fit with reality’ (Puviani, 1960). The Italians were thoroughgoing realists and not romantic idealists, and it was a short distance from their initial formulations to what subsequently became known as public choice.

The sober realism of the Italians implied, in keeping with the general equilibrium theorizing of the time, that actual fiscal outcomes were to be explained as equilibrium outcomes. If so, it might seem as though fiscal theorizing offered no coherent vantage point from which to pursue any programme of fiscal reform. Yet Buchanan has always sought to use fiscal knowledge as an instrument of fiscal reform. It was Knut Wicksell who provided Buchanan with the vehicle for combining his sober realism with his interest in reform. Buchanan's constitutional emphasis can be traced to the second of Wicksell's three essays in *Finanztheoretische Untersuchungen*, which Buchanan translated as ‘A New Theory of Just Taxation’, in *Classics in the Theory of Public Finance*, edited by Richard Musgrave and Alan Peacock. From Wicksell, Buchanan derived two themes that informed his work thereafter. One theme was the treatment of unanimous consent and not majority approval as the normative benchmark for appraising

political outcomes. The other theme was a distinction between constitutional politics, where institutional rules are selected, and post-constitutional politics, where particular outcomes emerge. Wicksell's treatment of two distinct levels of political activity led to Buchanan's articulation of a constitutional political economy, wherein political reform was a matter of changing the rules that govern the game, as distinct from changing the strategies of play within a game.

While Wicksell and the Italians cover the two themes mentioned in Buchanan's Nobel citation, any mention of precursory influences would be remiss without including Frank Knight, whom Buchanan initially encountered during his student days at the University of Chicago. Knight's influence on Buchanan is not so much one of particular ideas as of general attitude and orientation towards a scholar's life and work. From Knight, Buchanan carried forward the belief that no doctrine or authority should be treated as sacrosanct and above challenge. Everyone else may say that something is true, but this doesn't mean they are right; there may be many pretentious emperors walking around naked. Buchanan's work has also demonstrated the same multidisciplinary character that was prominent in Knight's work. For Buchanan, as for Knight, economic theorizing was not self-contained, but had points of contact throughout the humane studies, which led to a style of theorizing wherein Buchanan, like Knight, continually makes contact with such related fields of inquiry as law, ethics, history, philosophy, and politics.

From Italian public finance to public choice

The Italian approach to public finance treated the state as an entity whose actions conformed to the same principles of marginal utility as the actions of other economic participants. The Italians did not seek to advance statements concerning how large the state should be in order to promote some vision of social welfare. They sought instead to offer coherent explanations about the actual size of the state. At the level of formal analysis, this meant that the state would expand until the marginal utility from state-provided services equalled the marginal utility from market-supplied services. To be sure, the Italians recognized the numerous problems of aggregation that were involved in making such statements. In response, they developed a variety of models regarding just whose utility was driving the equilibrium. Where some models treated the state as a cooperative enterprise that worked to the benefit of all, others treated the state as an entity that promoted the advantage of ruling classes. In any case, it was a small step from the Italian fiscal theorizing to the public choice theorizing that began to take shape in the 1960s, as elaborated in Richard Wagner (2003).

Perhaps the best place to see the Italian influence on public choice is Buchanan's 1967 treatise *Public Finance in Democratic Process*, which was written at a time when ‘public choice’ was not yet a term of scholarly identification. Buchanan starts that book by noting the narrow and limited scope of Anglo-Saxon approaches to public finance, wherein public finance is concerned only with explaining market-based reactions to exogenously imposed taxes and expenditures. On the tax side of the budget, for instance, a progressive income tax with several brackets of rising marginal rates might be replaced by a degressive tax where a single marginal rate is imposed above some initial exemption. The task of the fiscal scholar would be to explain the impact of such an exogenous tax shift on such things as the amount of labour people supply, the amount of underground economic activity they undertake, and the amount of taxable income they earn. Alternatively, on the expenditure side of the budget, an appropriation might be made to finance a highway. The task of fiscal analysis would be to analyse the

market-based reactions to the highway. For instance, land rents near highway exits might rise due to the reduction in travel time that resulted. Whatever the particular topic examined, the analytical task of Anglo-Saxon public finance has everything to do with explaining market-based reactions to exogenously imposed fiscal measures and has nothing to do with explaining state budgets and fiscal institutions.

In treating state budgets as exogenous to fiscal inquiry, the Anglo-Saxon orientation towards public finance ignored two large areas of possible inquiry, both of which Buchanan explores in *Public Finance in Democratic Process*. One ignored area is the ability of fiscal institutions to influence budgetary outcomes and not just market outcomes. This topic occupies the first part of *Public Finance in Democratic Process*, and the analyses presented there were early illustrations of public choice theorizing. The second ignored area is the choice or emergence of fiscal institutions. This topic occupies the second part of *Public Finance in Democratic Process*, and the analyses presented there were harbingers of subsequent work in constitutional political economy.

Buchanan gives several illustrations in *Public Finance in Democratic Process* of how fiscal institutions and arrangements might influence fiscal outcomes, of which I mention three. First, Buchanan examines the possible budgetary consequences of a choice between general-fund financing and tax earmarking. Under the former practice, tax revenues accrue to a general fund from which various appropriations are made; under the latter practice, specific taxes are earmarked to finance particular services. Buchanan suggests that general-fund financing is a form of tie-in sale that might bring about a budgetary shift in favour of services in relatively elastic demand.

Second, Buchanan examines the possible budgetary consequences of the withholding of income taxes. His analysis in this case is related to claims about fiscal illusion or perception. Buchanan argues that individual perceptions about the costliness of public output depend on the manner in which tax extractions are made. Perhaps the most open and direct manner of paying for public output would be for people to write monthly checks to government, just as they pay their utility bills. Buchanan explores the possibility that withholding may create some tendency for individuals to perceive the cost of government to be less than it would otherwise be, which should in turn lead to some increase in the size of government.

Third, Buchanan examines the effect of public debt on budgetary outcomes, a topic that he initially explored in *Public Principles of Public Debt* and to which he returned in *Democracy in Deficit* (co-authored with Richard Wagner). The principle of Ricardian equivalence holds that tax finance and debt finance are identical. In the aggregate, this is true as a simple matter of double-entry accounting. If $1 million of tax revenue is replaced by public borrowing, the present value of the future payments necessary to service the debt will equal the tax reduction. However, the collectivity does not act as a unit, so a statement about aggregate equivalence is irrelevant to any effort to explain fiscal conduct. What matters for collective action is the direction of individual desires as these are mediated through political and fiscal institutions. For instance, people in higher age ranges will find debt to be less costly than taxation, increasingly so with age. Compare a tax of $1,000 now with a perpetual debt that entails payments of $100 when the appropriate discount rate is ten per cent. In terms of perpetuity, the debt and the tax are equivalent. For a younger person who might look forward to 50 taxpaying years, the present value of the debt is $991. For an older person who might only have ten years of tax-paying life expectancy left, the present value of the debt is but $614.

To be sure, it could be claimed that the older person has some bequest motivation towards heirs. If so, that older person would treat the debt obligation as continuing beyond his life. But not all older people

have heirs. And of those that do, not all of them seem to have the types of bequest motives that generate Ricardian equivalence. This point gets to another significant feature of Buchanan's thought: his unwillingness to make statements based on aggregates without exploring the underlying structural patterns to which those aggregates pertain. After all, aggregates are not entities that act, and in Buchanan's approach collective action must be generated out of choices by discernible, acting individuals, as these choices are mediated through institutional frameworks for making collective choices.

The literature on public choice has, of course, exploded since 1967, with entrées to this literature provided by such compendia as Mueller (1997), Rowley and Schneider (2004), and Shughart and Razzolini (2001). A good deal of that literature has carried forward the effort of Buchanan and his Italian forebears to articulate the impact of political institutions on collective outcomes.

From Wicksell to constitutional political economy

Where public choice examines the impact of political and fiscal institutions on collective outcomes, constitutional political economy examines the impact of constitutional rules on post-constitutional outcomes. The seminal statement of constitutional political economy is the *Calculus of Consent* (co-authored with Gordon Tullock), which the authors described as simply an elaboration with economic logic of the American constitutional framework of 1789. According to that framework, government is established by the consent of the governed, which provides unanimity as the conceptual starting point, just as it did for Wicksell (Wagner, 1988, explores the relationship between Wicksell and the *Calculus of Consent*). While unanimity is the conceptual starting point, any effort actually to implement unanimity will confront free riders and strategic hold-outs. If everyone's consent is required to undertake collective action, some people will be tempted to withhold their consent, not because they object to the action but because they are acting strategically to shift the fiscal terms of the action in their favour. Such strategic efforts at securing distributional gain can sabotage projects that are genuinely beneficial to all. Consequently, people may reasonably agree to be bound by something less than unanimous consent. Buchanan and Tullock conceptualized a trade-off between decision costs and external costs, as these are viewed from the perspective of participants in collective choice. Decision costs are the costs people bear in trying to reach a collective decision. The greater the degree of consent required, the higher will be those costs due to such things as free riding and strategic bargaining. External costs are the costs that individuals bear when collective choices run contrary to their desires. These costs will fall with increases in the degree of consent required to take to collective action, and will vanish when unanimity is required. An optimal voting rule, formally speaking, will result when the sum of those costs is minimized. With this analytical construction, Buchanan and Tullock provided a rationalization for Knut Wicksell's support for some super-majority rule within a parliamentary assembly, as illustrated by references to three-quarters and four-fifths consent.

A voting rule is a simple scalar. Actual constitutional frameworks for collective choice contain a vector of characteristics, and to some extent those other characteristics can substitute for greater inclusivity in the degree of consent required. For instance, a representative assembly that is bicameral can achieve a greater degree of consensus with a less inclusive voting rule than would be possible within a unicameral assembly. Legislative action, moreover, can be filtered in various fashions through different

parliamentary rules. There are many margins along which political and fiscal institutions can be modified, and with post-constitutional politics adapting to whatever constitutional framework is in place. There are two levels of analysis in Buchanan's analytical schema: constitutional and post-constitutional. Post-constitutional politics, public choice, represents the working out of interactions among political participants within the context of some particular institutional arrangement. Constitutional politics concerns the selection among possible institutional arrangements. Buchanan's distinction between constitutional and post-constitutional politics calls forth the distinction between choosing the rules of a game and choosing strategies by which to play a game. For Buchanan, reform is a constitutional and not a post-constitutional matter.

Consider, for instance, his approach to progressive income taxation. Where the Anglo-Saxon sacrifice theorists sought to specify the degree of progressivity that some exogenous authority should impose on a society, Buchanan sought to probe the circumstances under which people might choose to employ progressivity in taxing themselves. In several places, he explores the conditions under which people might support progressive income taxation as a form of income insurance. Progressive taxation, as compared with proportional taxation, allows people to achieve some smoothing of consumption in the presence of fluctuating income. The purchase of insurance, after all, is a constitutional and not a post-constitutional activity: people purchase insurance before they have had accidents and not after. To the extent that such formulations have merit, what appears to be redistribution when seen from an *ex post* perspective might represent mutual gains from trade when viewed from an *ex ante*, constitutional perspective.

Alternatively, consider the treatment of broad-based taxation in Buchanan and Congleton (1998). Without a constitutional requirement of uniformity in taxation, post-constitutional politics will generate increasingly complex revenue systems as tax favours are granted or removed within the political marketplace. While the resulting narrowing of the tax base imposes excess burdens on market participants, it also warps processes of collective choice. For instance, those who are favoured by the resulting fiscal discrimination will support more collective activity than they would otherwise. With the continual churning of the tax code that results, however, most participants may end up worse off than they would have been under a simple system of tax uniformity.

Buchanan's legacy

Until the late 1930s there was a flourishing Continental orientation towards public finance that stood in contrast to the Anglo-Saxon orientation, and pretty much along the lines articulated by Buchanan in *Public Finance in Democratic Process* (this thesis is elaborated in Backhaus and Wagner, 2005). Within this orientation, public finance was a multidisciplinary field of study, with a home in economics but with tentacles that reached out into such fields as politics, law, and public administration. Buchanan has carried forward the Continental approach to public finance, and has given it new life through his many creative works.

See Also

- constitutions, economic approach to

- sovereign debt

Selected works

1958. *Public Principles of Public Debt*. Homewood, IL: Richard D. Irwin.

1962. (With G. Tullock.) *The Calculus of Consent*. Ann Arbor: University of Michigan Press.

1967. *Public Finance in Democratic Process*. Chapel Hill: University of North Carolina Press.

1977. (With R. Wagner.) *Democracy in Deficit*. New York: Academic Press.

1992. *Better than Plowing*. Chicago: University of Chicago Press.

1998. (With R. Congleton.) *Politics by Principle, Not Interest*. Cambridge: Cambridge University Press.

1999–2002. *The Collected Works of James M. Buchanan*. 20 vols. Indianapolis: Liberty Fund. (This collection contains the preponderance of his academic work, both books and papers, through the late 1990s.).

Bibliography

Backhaus, J. and Wagner, R. 2005. From continental public finance to public choice: mapping continuity. *History of Political Economy* 37 (supplement), 314–32.

Mueller, D., ed. 1997. *Perspectives on Public Choice*. Cambridge: Cambridge University Press.

Musgrave, R. and Peacock, A., ed. 1958. *Classics in the Theory of Public Finance*. London: Macmillan.

Puviani, A. 1960. *Die Illusionen in der öffentlichen Finanzwirtschaft*. Berlin: Dunker & Humblot. [A German translation of A. Puviani, *Teoria della illusione finanziaria*, 1903.].

Rowley, C. and Schneider, F. 2004. *The Encyclopedia of Public Choice*. Dordrecht: Kluwer Academic Publishers.

Shughart, W. II and Razzolini, L., eds. 2001. *The Elgar Companion to Public Choice*. Cheltenham, UK: Edward Elgar.

Wagner, R. 2003. Public choice and the diffusion of classic Italian public finance. *Il pensiero economico* 11: 271–82.

Wagner, R. 1988. *The Calculus of Consent*: A Wicksellian retrospective. *Public Choice* 56, 153–66.

How to cite this article

Wagner, Richard E. "Buchanan, James M. (born 1919)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000303> doi:10.1057/9780230226203.0170

The New Palgrave Dictionary of Economics Online

budget deficits

William G. Gale
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes alternative measures of the federal budget deficit, discusses traditional and non-traditional channels through which deficits can affect the economy, and summarizes research on the effects of deficits on national saving and interest rates.

Keywords

aggregate consumption; budget deficits; consumption function; crowding out; Euler equations; financial market expectations; fiscal policy; generational accounting; interest rates; labour supply; national income; national saving; public debt; Ricardian equivalence theorem; Ricardo, D.; small open economy view

Article

Federal budget deficits reflect the extent to which current federal spending policies are not being financed with current federal tax policies, and can have significant effects on national saving and interest rates.

Economists have explored the effects of budget deficits extensively, and analysis of the aggregate effects of fiscal policy dates back at least to the work of David Ricardo. Modern academic interest was reinvigorated by the work of Barro (1974) and others, and by the large US federal budget deficits in the 1980s and early 1990s. These factors led to a substantial amount of research that is summarized in several excellent surveys (Barro, 1989; Barth et al., 1991; Bernheim, 1987; 1989; Elmendorf and Mankiw, 1999; Seater, 1993). The rapid but short-lived transition to unified budget surpluses in the late 1990s, followed by the sharp reversal in budget outcomes since 2000, has raised interest in this question again.

The budget deficit can be defined in many different ways, and the most appropriate measure is likely to depend on the particular model or application of interest. For any measure of the deficit, which is a flow

during a given time period, there is an analogous measure of the public debt, which is a stock at a given point in time and which represents the net accumulation of the associated deficits over all previous time periods.

The most widely used measure of the US federal deficit – the unified budget balance – is fundamentally (but not exactly) a cash-flow metric that includes both the Social Security and non-Social Security components of the federal budget. In a first approximation, the unified deficit shows the extent to which the government borrows or lends in credit markets. For some purposes, it is more informative to examine the primary budget, which excludes interest payments on the public debt (that is, it is equal to the unified budget balance minus net interest payments). The standardized budget balance adjusts the unified budget for the business cycle and special items. All these measures share a basic focus on cash flow.

Broader measures of the budget deficit look beyond cash flow and take into account the implicit or explicit promises embedded in current government policies, even if such promises do not result in current-period cash flow. Generational accounting, for example, aims to tally the net debt that each generation or birth cohort faces (see Auerbach, Gokhale, and Kotlikoff, 1991 for discussion of generational accounting and Auerbach et al., 2003 for discussion of alternative measures of the deficit). However, it is unclear how the market and households value implicit debts relative to the government's explicit debt. Thus, while the importance of the broader measures is clear conceptually, this article focuses mostly on the cash-flow related measures of the deficit.

In the fiscal year 2005, the unified US federal deficit was about 2.6 per cent of the GDP, and the standardized deficit was about 1.8 per cent (Congressional Budget Office, 2006). The current budget situation would largely not be a concern if future fiscal prospects were auspicious. Unfortunately, the longer-term budget outlook is dismal, primarily because of projected rising expenditures on health care and programmes for the elderly (Congressional Budget Office, 2005).

Economic effects of budget deficits: traditional channels

Economists tend to view the aggregate effects of tax cuts from one of three perspectives. To sharpen the distinctions, consider deficits induced by changes in the timing of lump-sum taxes, with the path of government purchases and marginal tax rates held constant. Under the Ricardian equivalence hypothesis, such deficits are fully offset by increases in private saving and have no effect on national saving, interest rates, exchange rates, future domestic production, or future national income. A second model, the small open economy view, suggests that budget deficits reduce national saving, but induce increased international capital inflows that finance the entire reduction in national saving. As a result, domestic production does not decline and interest rates do not rise, but future national income falls because of the burden of repaying the increased borrowing from abroad. A third model, which we call the conventional view, suggests that deficits reduce national saving and that the reduction in national saving is at least partly reflected in lower domestic investment. In this model, budget deficits partly crowd out private investment and partly increase borrowing from abroad; the combined effect reduces future national income and future domestic production. The reduction in domestic investment in this model is facilitated by an increase in interest rates, establishing a connection between deficits and interest rates.

It is worth emphasizing that the relationship between deficits and national saving is central to analysis of the economic effects of fiscal policy. National saving, which is the sum of private and government saving, finances national investment, which is the sum of domestic investment and net foreign investment. Higher national saving raises the capital stock owned by the nation's citizens and thus raises future national income.

An increase in the budget deficit reduces national saving unless it is fully offset by an increase in private saving. If national saving falls, national investment and future national income must fall as well, if all else remains equal. Therefore, to the extent that budget deficits reduce national saving, they reduce future national income. This reduction in future national income occurs even if there is no increase in domestic interest rates. In the case where there is no rise in domestic interest rates, the reduction in national saving associated with budget deficits would manifest itself solely in increased borrowing from abroad (as under the small open economy view). This is the sense in which the effect of deficits on interest rates and exchange rates (the distinction between the small open economy view and the conventional one) is subsidiary to the question of the effects on national saving (the Ricardian view versus the other two).

A key consideration is that the results above consider only the effects of increased budget deficits or debt per se. A full analysis of the effects of public policies on economic growth should take into account not only the effects of increased deficits and debt but also the direct effects of the spending programmes or tax reductions that cause them. The effects of fiscal policies on both economic performance and interest rates depend not only on the deficit but also on the specific elements of the policies generating that deficit. For example, spending one dollar on public investment projects would increase the unified budget deficit by one dollar, but the net effect on future income would depend on whether the return on the public investment project exceeded the return on the private capital that would have instead been financed by the national saving crowded out by the deficit. Similarly, a deficit of one per cent of GDP caused by reducing marginal tax rates will generally have different implications for both national income and interest rates from a deficit of one per cent of GDP caused by increasing government purchases of goods and services.

Economic effects of budget deficits: non-traditional channels

Beyond their direct effect on national saving, future national income and interest rates, deficits can affect the economy in other ways. For example, increased deficits may cause investors gradually to lose confidence in national economic stability and leadership. As Truman (2001) emphasizes, a substantial fiscal deterioration over the longer term may cause ‘a loss of confidence in the orientation of US economic policies’. Such a loss in confidence could then put upward pressure on domestic interest rates, as investors demand a higher risk premium on dollar-denominated assets. The costs of current account deficits – which are in part induced by large budget deficits – may even extend beyond narrow economic ones. More broadly, Friedman (1988, p, 76) notes that ‘World power and influence have historically accrued to creditor countries. It is not coincidental that America emerged as a world power simultaneously with our transition from a debtor nation … to a creditor supplying investment capital to the rest of the world.’

Both the traditional models and the non-traditional effects noted above focus on gradual negative effects from reduced national saving. This focus may be too limited, however, in that it ignores the possibility

of much more sudden and severe adverse consequences. In particular, the traditional analysis of budget deficits in advanced economies does not seriously entertain the possibility of explicit default or implicit default through high inflation. If market expectations regarding the probability of default were to change and investors had difficulty seeing how the policy process could avoid extreme steps, the consequences could be much more sudden and severe than traditional estimates suggest. The role of financial market expectations in this type of scenario is central. One of the key triggers would occur if investors begin to doubt whether the strong historical commitment to avoiding substantial inflation would be weakened in order to reduce the real value of the public debt (Ball and Mankiw, 1995; Rubin, Orszag and Sinai, 2004).

Although this article does not explicitly incorporate non-traditional effects into the discussion below, such effects serve as an important reminder of why budget deficits, especially chronic deficits, could exert large adverse effects on US economic performance. The focus on traditional effects is certainly justifiable in the context of historical analysis of post-war data from the United States. That does not imply, however, that to ignore such issues is appropriate when examining the likely impacts of future deficits. The nation has never before faced substantial deficits that are projected to be sustained and indeed to grow over many decades.

Deficits and consumption

Testing the effect of deficits on aggregate consumption, with government spending held constant, is an important focus of analysis for several reasons. First, these analyses provide a direct test of whether the timing of tax collections affects the economy, with other factors controlled for. Second, the aggregate time series tests measure the *magnitude* of the effects in question. This is particularly important because virtually no one claims that Ricardian equivalence is literally true. Rather, the controversy is over the extent to which Ricardian equivalence is a good approximation of the aggregate impact of fiscal policies. There is a wide variety of research findings from studies of aggregate consumption and fiscal policy, in part because of a variety of difficult econometric issues. Barro (1989) and Elmendorf and Mankiw (1999) conclude that the literature is inconclusive. Seater (1993) concludes that, once the studies are corrected for econometric problems, Ricardian equivalence is corroborated – or at least that it is not possible to reject Ricardian equivalence. Bernheim (1989) concludes that, once the studies are normalized appropriately, Ricardian equivalence should be rejected.

One strand of the literature specifies consumption functions and then tests for the effects of fiscal policy. Perhaps the best-known study in this area is Kormendi (1983), who finds no evidence of non-Ricardian effects. This work has spawned significant research, including three sets of exchanges in the *American Economic Review*. Recent research, however, has extended the Kormendi results in three ways: using more recent data, which captures significant variation in budget outcomes; controlling for measures of marginal tax rates; and (in the United States) allowing federal and state fiscal variables to have different effects on consumption. The last issue is particularly relevant because the states collect a significant share of their revenue through consumption taxes, which would be expected to vary *positively* with consumption, whereas other taxes would be expected, at least in non-Ricardian theory, to vary negatively. With these extensions, the results suggest that about 30 to 46 cents of every dollar in federal tax cuts is spent in the same year (Gale and Orszag, 2004). This is a rejection of the Ricardian view.

Another strand of the literature focuses on Euler equation tests (relating to the growth rate of consumption, as opposed to the tests above, which examine consumption levels), with mixed results. As Bernheim (1987) points out, Ricardian equivalence can fail even if the Euler equation does not, and vice versa. Nevertheless, some studies have found substantial effects of fiscal policy on consumption using the Euler framework, most recently Gale and Orszag (2004), who find that about 50 to 85 cents of every dollar in tax cuts is spent in the first year, with most of the effects measured precisely. This range is consistent with some previous assessments, but it is inconsistent with the Ricardian prediction of a full offset from private saving.

Deficits and interest rates

The effects of fiscal policy on interest rates have also proven difficult to pin down statistically. The issues include the appropriate definition of deficits and debt, whether deficits or debt should be the variable of interest, the difficulty of distinguishing expected and unexpected changes, and the potential endogeneity of many of the key explanatory variables (see Bernheim, 1987; Elmendorf and Mankiw, 1999; Seater, 1993).

In part because of these statistical issues, the evidence from the empirical literature as a whole is mixed. However, the key role of *expected* deficits rather than current deficits is sometimes overlooked. As Feldstein (1986, p. 14) has written, ‘it is wrong to relate the rate of interest to the concurrent budget deficit without taking into account the anticipated future deficits. It is significant that almost none of the past empirical analyses of the effect of deficits on interest rates makes any attempt to include a measure of expected future deficits.’ Since financial markets are forward-looking, to exclude expectations could bias the analysis towards finding no relationship between interest rates and deficits. In fact, studies that incorporate more accurate information on expectations of *future* sustained deficits tend to find economically and statistically significant connections between anticipated deficits and current interest rates. Gale and Orszag (2004) show that, of the 19 papers incorporating timely information on projected deficits, 13 find predominantly positive, significant effects between anticipated deficits and current interest rates, five find mixed effects, and only one finds no effects. The other studies in the literature that find no significant effect are disproportionately those that do not take expectations into account at all or do so only indirectly through a vector autoregression. Thus, while the literature as a whole, taken at face value, generates mixed results, analyses that focus on the effects of anticipated deficits tend to find a positive and significant impact on interest rates.

The challenge in incorporating market expectations about future deficits is that such expectations are not directly observable. An important caveat to the whole literature, then, is that, to the extent that proxies for expected deficits are imperfect reflections of current expectations, the coefficient on the projected deficit will tend to be biased towards zero because of classical measurement error, and the studies would tend to *underestimate* the effects of deficits on interest rates.

Even among studies that use expected deficits, one potential concern is that the business cycle could be affecting current yields. Laubach (2003) suggests a novel way to resolve this issue: he examines the relationship between projected deficits (or debt) and the level of real *forward* (five-year ahead) long-term interest rates. The underlying notion is that current business cycle conditions should not influence the long-term rates expected to prevail beginning five years ahead. Laubach uses projections of the US

Congressional Budget Office and Office of Management and Budget, and finds that a one percentage point increase in the five-year-ahead projected deficit-to-GDP ratio raises the five-year-ahead ten-year interest rate by between 24 and 40 basis points, and that a one percentage point in the projected debt-to-GDP ratio raises the long-term forward rate by between 3.5 and 5.5 basis points. The deficit-based results are not dissimilar from the debt-based results. Consider, for example, an increase in the budget deficit equal to one per cent of GDP in each year over the next ten years. After ten years, that would raise government debt by roughly ten per cent of GDP. The deficit-based results in Laubach would suggest about a 30 basis point increase in interest rates, whereas the debt-based results would suggest about a 45 basis point increase.

Using a similar framework, Engen and Hubbard (2004) obtain somewhat smaller effects while Gale and Orszag (2004) obtain somewhat larger effects. Indeed, despite a rancorous public debate, there appears to be a surprising degree of convergence in recent estimates of the effects of fiscal policy on interest rates, with a variety of econometric studies implying that a sustained one per cent of GDP increase in unified deficits over ten years would raise interest rates by 30 to 60 basis points. The relationship between deficits and interest rates not only provides further evidence against the Ricardian view, but also implies that the conventional view is a better description of reality for the United States than the small open economy view. Ardagna, Caselli and Lane (2004) find even stronger results in a panel of 16 Organisation for Economic Co-operation and Development (OECD) countries over several decades.

Conclusion

Sustained federal budget deficits have two sets of effects. The direct effect of the increase in government borrowing is to reduce national saving and raise long-term interest rates, often by empirically sizable amounts. The other set of effects depends on the specific tax or spending policies that were chosen to create the deficits. These findings have significant implications. First, both the consumption and the interest rate results reject the Ricardian view of the world. Second, the interest rate results reject the small open economy view, at least as it applies to the US economy. Third, the results suggest that the sustained deficits facing the nation will impose significant economic costs. Fourth, some tax-cut policies that have traditionally been considered growth-enhancing may actually backfire, because the generally positive effect of the tax rate cut on labour supply and investment, if interest rates are held constant, can be offset by the impact of the deficit on interest rates and on national saving. While it would be wrong to conclude that all these issues are decisively resolved in the economics literature, there is more than strong enough evidence to raise concerns about sustained projected future deficits.

See Also

- crowding out
- new open economy macroeconomics
- Ricardian equivalence theorem

Bibliography

Ardagna, S., Caselli, F. and Lane, T. 2004. Fiscal discipline and the cost of public debt service: some estimates for OECD countries. Working Paper No. 10788. Cambridge, MA: NBER.

Auerbach, A.J., Gale, W.G., Orszag, P.R. and Potter, S.R. 2003. Budget blues: the fiscal outlook and options for reform. In *Agenda for the Nation*, ed. H.J. Aaron, J.M. Lindsey and P.S. Nivola. Washington, DC: Brookings Institution.

Auerbach, A.J., Gokhale, J. and Kotlikoff, L.J. 1991. Generational accounts: a meaningful alternative to deficit accounting. In *Tax Policy and the Economy*, ed. D. Bradford. Cambridge, MA: NBER.

Ball, L. and Mankiw, N.G. 1995. What do budget deficits do? In *Budget Deficits and Debt: Issues and Options*. Kansas City: Federal Reserve Bank of Kansas City.

Barro, R.J. 1974. Are government bonds net worth? *Journal of Political Economy* 82, 1095–117.

Barro, R.J. 1989. The Ricardian approach to budget deficits. *Journal of Economic Perspectives* 3(2), 37–54.

Barth, J.R., Iden, G., Russek, F.S. and Wohar, M. 1991. The effects of federal budget deficits on interest rates and the composition of domestic output. In *The Great Fiscal Experiment*, ed. R.G. Penner. Washington, DC: Urban Institute Press.

Bernheim, B.D. 1987. Ricardian equivalence: an evaluation of theory and evidence. *NBER Macroeconomics Annual* 2, 263–304.

Bernheim, B.D. 1989. A neoclassical perspective on budget deficits. *Journal of Economic Perspectives* 3 (2), 55–72.

Congressional Budget Office. 2005. *The Long-Term Budget Outlook*. Washington, DC: Congressional Budget Office.

Congressional Budget Office. 2006. *The Budget and Economic Outlook: Fiscal Years 2007 to 2016*. Washington, DC: Congressional Budget Office.

Elmendorf, D.W. and Mankiw, N.G. 1999. Government debt. In *Handbook of Macroeconomics*, vol. 1C, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Engen, E.M. and Hubbard, R.G. 2004. Federal government debt and interest rates. In *NBER Macroeconomic Annual 2004*, ed. G. Mark and K.S. Rogoff. Cambridge, MA: MIT Press.

Feldstein, M.S. 1986. Budget deficits, tax rules, and real interest rates. Working Paper No. 1970.

Cambridge, MA: NBER.

Friedman, B. 1988. *Day of Reckoning: The Consequences of American Economic Policy under Reagan and After*. New York: Random House.

Gale, W.G. and Orszag, P.R. 2004. Budget deficits, national saving, and interest rates. *Brookings Papers on Economic Activity* 2004(2), 101–210.

Kormendi, R. 1983. Government debt, government spending, and private sector Behavior. *American Economic Review* 73, 994–1010.

Laubach, T. 2003. New evidence on the interest rate effects of budget deficits and debt. Finance and Economics Discussion Series No. 2003–12. Board of Governors of the Federal Reserve System.

Rubin, R.E., Orszag, P.R. and Sinai, A. 2004. Sustained budget deficits: longer-run U.S. economic performance and the risk of financial and fiscal disarray. Paper presented to the AEA-NAEFA Joint Session, Allied Social Science Associations Annual Meetings, 4 January, San Diego.

Seater, J.J. 1993. Ricardian Equivalence. *Journal of Economic Literature* 31, 142–190.

Truman, E.M. 2001. The international implications of paying down the debt. Policy Brief 01–7. Washington, DC: Institute for International Economics.

How to cite this article

Gale, William G. "budget deficits." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000332> doi:10.1057/9780230226203.0172

The New Palgrave Dictionary of Economics Online

budget projections

Douglas Holtz-Eakin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys different approaches to the construction of government budget projections, illustrated with procedures from the United States Office of Management and Budget (OMB) and Congressional Budget Office (CBO). It sets out the several distinct steps that are required in budget projections, from macroeconomic forecasting to comparing projections with outcomes and analysing the sources of deviations.

Keywords

aggregate demand; budget projections; budget window; business cycles; forecasting; intertemporal incentives; scoring; uncertainty

Article

Budget projections are central to governmental policymaking. In general, budgeting is the practice of devoting economic resources to policy objectives and providing specific means for raising these resources. A typical budget process includes budget proposals, review, adoption, and execution. Budget projections inform the process by providing estimated values for government revenues, government spending, and other budgetary concepts over a specific planning horizon (often referred to as the ‘budget window’). Budgetary projections are made under specific assumptions, for differing government programmes, using alternative approaches as part of the budgetary process. We discuss each in turn, with examples drawn from the United States federal government.
Threshold assumptions for budget projections fall along two dimensions: economic and policy.

Economic assumptions

One approach to developing a budget projection is based on a comprehensive economic forecast,

inclusive of any possible future business cycle fluctuations. In this instance, the result is a projection of the potential future outlays, receipts, and budget deficit or surplus. In the United States, both the White House Office of Management and Budget (OMB) and the Congressional Budget Office (CBO) adopt a variant of this approach in which the near-term forecast incorporates the state of the business cycle, while projections beyond the first two years assume an average of full employment.

Alternatively, it is sometimes assumed that the economy operates continuously at full resource utilization with no cyclical fluctuations. In this instance, the budget projections are often referred to as ‘cyclically adjusted’ or ‘full-employment’ projections of the budget and its balance.

Each approach serves distinct purposes. Budget projections are necessary, for example, to anticipate the cash-flow borrowing needs of the government on a year-by-year basis. In contrast, cyclically adjusted budget projections are useful for judging whether current deficits or surpluses are reflective of the state of the economy, and thus the degree to which fiscal policies are sustainable over the longer term.

Policy assumptions

The future path of the budget also depends on the evolution of tax and spending policies. In constructing the budget projection, one possible assumption is that current policies (or current laws) remain unchanged. Such a projection – known alternatively as a budget baseline projection or current services projection – provides a means by which to judge the future implications of current policies and a benchmark (or baseline) against which to measure the impact of policy changes.

Two issues arise in constructing and interpreting baseline budget projections. The first is the rules for anticipating any necessary future policy actions. For example, in the US federal budget a large fraction (roughly two-thirds in 2007) of spending results from ‘mandatory’ (or ‘direct’) spending programmes in which laws authorize automatic expenditures to eligible parties. Common examples are Social Security, Medicare, and farm support programmes. In these instances, projections of spending rely on combining rules of the programmes with projections of eligible populations and their relevant characteristics. An issue arises when the legal authorization for a programme expires during the projection period, requiring an assumption regarding whether spending will stop entirely or continue as if the current programme remains in place. (In the United States, ‘large’ programmes – spending in excess of $50 million – are assumed to continue.)

The remainder of spending (over one-third in 2007) is ‘discretionary’ and determined by the annual decisions of Congress. Consistent with the spirit of projecting current policy, baseline projections typically assume that this type of spending continues (in real, inflation-adjusted terms) exactly as in the most recently completed budget. An implication of this procedure is that baseline projections of discretionary spending may be heavily influenced by transitory policy events such as emergency spending.

These types of swings in projected spending are illustrative of the second key feature of baseline or current services projections. These projections are not forecasts of actual budget outcomes, but rather tools to inform the budgetary process.

A second approach is to embed in the budget projections a specific path for future policies, that is, to construct a policy-based budget projection. For example, the annual Presidential budget submitted to the US Congress is constructed under the assumption that all the proposed policies are adopted as requested. As with baseline budget projections, policy projections are not forecasts of actual budgetary outcomes.

Scoring

A topic closely related to budget projections is ‘scoring’ – the evaluation of the budgetary implications of policy proposals. Mechanically, scoring represents the difference between a policy-based projection and a baseline projection, thereby revealing the budgetary difference as a result of the specific policies. Scoring budgetary proposals permits comparisons of alternative proposals on a consistent basis. Traditionally, scores have been constructed under the assumption that overall macroeconomic performance is unchanged by the policy proposal (‘static scoring’). There are some proposals, however, of sufficient magnitude and impact on incentives (for example, tax reform) that it would be desirable to incorporate not only the direct budgetary impacts but also the budgetary feedbacks from changes in the overall levels of economic output and incomes (‘dynamic scoring’). Incorporating economic impacts, however, raises issues in maintaining consistency in scoring across proposals and details of executing the analysis (see Congressional Budget Office, 2002; Joint Committee on Taxation, 2006).

Steps for budgetary projections

Official governmental budget projections from, for example, the OMB and the CBO, are sophisticated, detailed exercises that require several distinct steps.

1. *Project macroeconomic performance*. The budget projection is built upon a macroeconomic forecast, including the path for real and nominal gross domestic product (GDP), the future rates of unemployment, the path for prices and inflation, and the path of future interest rates and exchange rates. As part of anticipating the near-term position in the business cycle, it is necessary to forecast the components of aggregate demand – consumption, residential and business investment, government spending, and net exports – as well as the determinants of the potential for overall output, such as capital stocks, labour force, and technological progress. Because of the importance of tax revenues to the budgetary projections, the projection of national income is more important than in other settings, imposing the requirement for projecting labour compensation, taxable versus non-taxable compensation, corporate profits, dividends, interest payments, and non-corporate business income.
2. *Impute a distribution to macroeconomic aggregates*. In the United States, personal income tax is progressive and heavily skewed towards the upper part of the income distribution (with the top one-half of households paying nearly all the income tax). Accordingly, the distribution of wage and salary earnings (as well as other components of household income) among households has a large impact on the overall level of tax receipts. In these circumstances, the macroeconomic forecast must be combined with microeconomic data drawn from tax returns and population surveys to provide accurate projections.
3. *Impose programme rules on the macroeconomic and microeconomic data to project spending and revenues*. For example, the projections for population, labour force, and the unemployment rate yield forecasts of the number of unemployed individuals. When combined with unemployment insurance programme rules, the unemployment forecast yields a projection of outlays for the unemployment insurance programme. Similarly, the projection of wage income, dividend payments, interest payments, and capital gains, along with distributional information on each, may be combined with parameters of the tax code to produce projections of individual income tax receipts.

An important aspect of this step is the sophistication of incorporating responses to incentives in the

projections. For example, if current law indicates that tax rates will rise in the next several years, it is likely that intertemporal incentives may shift forward some economic activity (for example, labour supply) and some tax-based planning behaviours (for example, realization of capital gains to obtain lower tax rates). It is desirable to incorporate these responses in the projection.

4. *Check for internal consistency*. In some circumstances, budget projections involve an element of simultaneity. For example, fiscal projections (spending and taxes) are necessary to forecast near-term aggregate demand, while actual outlays and tax receipts depend upon the employment and incomes generated by economic activity. Accordingly, it is desirable to check whether the budget totals are consistent with the economic projection.

5. *Compare projections with actual outcomes to improve projections*. The accuracy of budget projections is an obvious concern. Hence it is desirable to do a comparison of actual outcomes with past projections to identify systematic sources of error and opportunities for improvement. In addition, a second desirable attribute of projections is their credibility, which is aided by a transparent process for revealing differences between actual and projected outcomes, and a systematic analysis of the sources of deviation.

Uncertainty and valuation in budget projections

Uncertainty

Budgetary projections are fraught with uncertainty. At the most basic level, the future is literally unknowable, and budgetary projections will be affected by the future course of macroeconomic fluctuations, variations in inflation, the path of interest rates, and so forth. The degree to which projections are uncertain is important information to policymakers. One approach to revealing the scale of uncertainty is to undertake the budget projections in a series of scenarios (for example, ‘base case’, ‘faster growth and higher inflation’, and ‘slower growth and lower inflation’). The difficulty then becomes choosing scenarios that are representative of the likely fluctuations to be experienced.

A more complete and formal approach is to conduct the entire projection in the context of a stochastic simulation methodology. In this approach, historical joint distributions are constructed for the key inputs to the projection (GDP growth, inflation, interest rates, wages, and so forth). Undertaking a large number of projections, each based on a ‘draw’ from the joint distribution, permits policymakers to be presented with the full distribution of potential outcomes over the budget horizon.

A second type of uncertainty is important for individual programmes. In some cases, government budget flows are contingent upon uncertain outcomes. A prominent example is agriculture programmes that provide funds only in the event of poor harvests due to drought or other adverse events. How should budget projections be constructed for such programmes? Choosing a single scenario will probably yield a projection in which the programmes either have a budget impact *every* year or in no year – neither of which is a sensible projection. A simple solution is to use the average (perhaps over a historical period) as the projected value of the budget impact of the programme, with the logic being that the projection is never precisely correct, but on average informative. As above, however, an alternative is to undertake formal stochastic simulations of the programme in question and use the expected value of the programme as the budget projection.

Valuation

The practice of budgetary projections (and scoring) raises issues in the correct valuation of budgetary transactions. In the main, the goal is to value government purchases using market prices (and thereby adhering as closely as possible to private-sector measure of marginal cost and marginal benefit). Similarly, tax collections and transfers to individuals and governments are measured in dollar values. However, difficulties can arise in the consistent application of these principles.
A notable example is the provision of insurance and insurance-like programmes by the government. Adhering to the principles of taxes and transfers, the projections of these programmes consist of the future tax receipts by the government and payments to individuals. Put differently, the budget projection consists of the future cash flows, perhaps summarized in an expected value form. Note, however, that this budgetary treatment may complicate comparisons with an equivalent programme – the direct purchase of an equivalent private-sector insurance product, where the private-sector entity will charge a risk premium.

Bibliography

CBO (Congressional Budget Office). 2002. CBO Testimony: Federal Budget Estimating. Statement of Dan L. Crippen, Director, before the Committee on the Budget, U.S. House of Representatives, 2 May. Online. Available at http://www.cbo.gov/ftpdocs/33xx/doc3384/05-02-Testimony.pdf, accessed 14 February 2007.

Joint Committee on Taxation. 2006. *Exploring Issues in the Development of Macroeconomic Models for Use in Tax Policy Analysis* (JCX-19–06), 16 June. Online. Available at http://www.house.gov/jct/x-19-06.pdf, accessed 14 February 2007.

How to cite this article

Holtz-Eakin, Douglas. "budget projections." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000220> doi:10.1057/9780230226203.0173

close print preview

The New Palgrave Dictionary of Economics Online

Bücher, Karl Wilhelm (1847 – 1930)

B. Schefold
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bücher, K. W.; division of labour; exchange; German Historical School; gifts; industrial organization; law of mass production; Menger, C.; Methodenstreit; rhythm; Rodbertus, J. K.; Schmoller, G. von; stages theory of economic development

Article

Karl Bücher was born in Kirberg (Germany) into a poor family. He studied history and classical philology in Bonn and Göttigen. Bücher first worked as a journalist for the liberal *Frankfurter Zeitung*, and from 1881 taught political economy in Dorpat, Basle, Karlsruhe and Leipzig, where he retired in 1917.

Bücher is counted among the outstanding economists of the German ‘younger’ historical school. He remained, however, independent in his economic thinking. He did not adhere to the inductive method and in the Methodenstreit he sided with Menger against Schmoller. Although he advocated the adoption of social policy measures by the state, he confessed to being a liberal and did not follow the protectionist and state interventionist line of the ‘Kathedersozialisten’ (socialists of the chair). An important contribution to economics was Bücher's ‘law of mass production’, which described the relationship between production costs and output in industrial manufacturing. Moreover, Bücher carefully analysed the organization of the labour process and the division of labour (1893, pp. 261–334). His study on the importance of rhythm for the working process in pre-industrial societies is extremely interesting and may be regarded as his most original work (1896). He described how workers transformed monotonous physical labour through the adoption of rhythmic repetitions of their movements. By adjusting the work speed to this rhythm, the working process was both eased and intensified. Such a rhythm could be generated, for example, by singing. Bücher gave vivid examples of typical work songs and particularly described the role played by work songs in combining large masses of workers to carry out large-scale works. However, a precondition for all this was the worker controlling his individual work speed and dominating his working instruments. The fact that in modern industry this was no more the case led Bücher to interesting reflections on man and work in our industrial environment (1896, pp. 112–117). Bücher's historical research focused on primitive people, antiquity and the Middle Ages. His analysis of

primitive people (1893, pp. 1–82; 1918, pp. 1–26) was too generalized and did not grasp fully the extreme complexity of economic relations among these peoples. However, in his elaborate research on the distinction between exchange and gift he anticipated some of the problems which modern ethnology would later discuss. His studies on the economies of ancient Rome and Greece were important because they contributed to the refutation of authors who described these economies as simply capitalistic. Among his contributions on the Middle Ages were studies on the social situation of women and journeymen, and a demographic study on medieval Frankfurt, where Bücher applied statistical methods (1886; 1922).

Bücher developed a theory of stages of economic development (1893, esp. pp. 83–160), where he distinguished between the household economy (Hauswirtschaft) of classical antiquity (in accordance with J.K. Rodbertus' notion of the *oikos* economy), the town economy (Stadtwirtschaft) of the Middle Ages, and the national economy (Volkswirtschaft), that is, the extensive exchange economy of modern times. The role of exchange served as the central distinctive criterion: exchange was supposed to be virtually absent in the household economy, which is the reason why the characterization of antiquity (where trade had been more important than Bücher thought) as a household economy was inaccurate. Exchange was confined to locally produced commodities and local markets in the medieval town economy, and dominating every sphere of economic life in the 'national economy'.

Bücher may also be regarded as one of the founders of journalism as an academic discipline. He especially focused on the role of the press for public opinion and the problems raised by the capitalist and profit-oriented structure of the press.

Selected works

1886. *Die Bevölkerung von Frankfurt am Main im XIV. und XV. Jahrhundert: Socialstatistische Studien*. Tübingen: Laupp.

1893. *Die Entstehung der Volkswirtschaft*. 1. Sammlung. 16th edn, Tübingen: Laupp, 1922. Trans. as *Industrial Evolution*, Toronto: University of Toronto Press, 1901.

1896. *Arbeit und Rhythmus*. Abhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaft. Leipzig: Hirzel, 1897.

1918. *Die Entstehung der Volkswirtschaft*. 2. Sammlung. 8th edn, Tübingen: Laupp, 1925.

1922. *Beiträge zur Wirtschaftsgeschichte*. Tübingen: Laupp.

How to cite this article

Schefold, B. "Bücher, Karl Wilhelm (1847–1930)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 29 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000215> doi:10.1057/9780230226203.0171

The New Palgrave Dictionary of Economics Online

Bukharin, Nikolai Ivanovitch (1888 – 1938)

Donald J. Harris
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Bukharin, N. I.; capitalism; falling rate of profit; imperialism; Lenin, V. I.; Luxemburg, R.; marginal revolution; Marx, K. H.; monopoly; period of production; planning; socialism; underdevelopment; uneven development

Article

Nikolai Bukharin is commonly acknowledged to have been one of the most brilliant theoreticians in the Bolshevik movement and an outstanding figure in the history of Marxism. Born in Russia, he studied economics at Moscow University and (during four years of exile in Europe and America) at the Universities of Vienna and Lausanne (Switzerland), in Sweden and Norway and in the New York Public Library. While still a student, he joined the Bolshevik movement. Upon returning to Russia in April 1917, he worked closely with Lenin and participated in planning and carrying out the October Revolution. After the victory of the Bolsheviks he proceeded to assume many high offices in the Party (becoming a member of the Politbureau in 1919) and in other important organizations. In these various capacities he came to exercise great influence within both the Party and the Comintern. Under Stalin's regime, however, he lost most of his important positions. Eventually, he was among those who were arrested and brought to trial under charges of treason and was executed on 15 March 1938.

At the peak of his career Bukharin was regarded as the foremost authority on Marxism in the Party. He was a prolific writer: there are more than five hundred items of published work in his name, most of them written in the hectic 12-year period 1916–1928 (for a comprehensive bibliography, see Heitman, 1969). Only a few of these works have been translated into English and these are the works for which he is now most widely known. A brief description of the major items gives an indication of the scope and range of his intellectual interests.

The Economic Theory of the Leisure Class (1917) is a detailed and comprehensive critique of the ideas of the Austrian school of economic theory, as represented by the work of its chief spokesman Eugen von Böhm-Bawerk, but situated in the broader context of marginal theory as it had appeared up to that time. In *Imperialism and World Economy* (1918) he formulated a revision of Marx's theory of capitalist

development and set out his own theory of imperialism as an advanced stage of capitalism. This was written in 1914–15, a year before Lenin's *Imperialism*, and is credited with having been a major influence on Lenin's formulation. The theoretical structure of the argument is further elaborated in *Imperialism and the Accumulation of Capital* (1924) by way of a critique of the ideas of Rosa Luxemburg, another leading Marxist writer of that time. *The ABC of Communism* (1919), written jointly with Evgenii Preobrazhensky and used as a standard textbook in the 1920s, is a comprehensive restatement of the principles of Marxism as applied to analysis of the development of capitalism, the conditions for revolution, and the nature of the tasks of building socialism in the specific context of the Soviet experience. This book, taken with his *Economics of the Transition Period* (1920), constitutes a contribution to both the Marxist theory of capitalist breakdown and world revolution on the one hand and the theory of socialist construction on the other. *Historical Materialism: A System of Sociology* (1921), another popular textbook, combines a special interpretation of the philosophical basis of Marxism with what is perhaps the first systematic theoretical statement of Marxism as a system of sociological analysis. In style much of this work is highly polemical and geared to immediate political goals. But it reveals also a versatility of intellect, serious theoretical concern, and scholarly inclination. Arguably, his works represent in their entirety ‘a comprehensive reformulation of the classical Marxian theory of proletarian revolution’ (Heitman, 1962, p. 79). Viewed from the standpoint of their significance in terms of economic analysis, three major components stand out.

There is, first, the critique of ‘bourgeois economic theory’ in its Austrian version. Bukharin's approach follows that which Marx had adopted in *Theories of Surplus Value*, which is to give an ‘exhaustive criticism’ not only of the methodology and internal logic of the theory but also of the sociological and class basis which it reflects. He scores familiar points against particular elements of the theory, for instance, that utility is not measurable, that Böhm-Bawerk's concept of an ‘average period of production’ is ‘nonsensical’, that the theory is static. Such criticisms of the technical apparatus of the theory have since been developed in more refined and sophisticated form (see Harris, 1978; 1981; Dobb, 1969).

Moreover, certain weaknesses in Bukharin's presentation, such as an apparent confusion between marginal and total utility and misconception of the meaning of interdependent markets, can now be readily recognized. But these are matters that were not well understood at the time, even by exponents of the theory. Bukharin views them as matters of lesser importance. What is crucial for him is ‘the point of departure of the … theory, its ignoring the social-historical character of economic phenomena’ (1917, p. 73). This criticism is applied with particular force to the treatment of the problem of capital, the nature of consumer demand, and the process of economic evolution. As to the sociological criticism, his central thesis is that the theory is the ideological expression of the rentier class eliminated from the process of production and interested solely in disposing of their income through consumption. This thesis can be faulted for giving too mechanical and simplistic an interpretation of the relation between economic theory and ideology where a dialectical interpretation is called for (compare, for instance, Dobb, 1973, ch. 1, and Meek, 1967). But the issue of the social-ideological roots of the marginal revolution remains a problematic one, as yet unresolved, with direct relevance to current interest in the nature of scientific revolutions in the social sciences (see Kuhn, 1970; Latsis, 1976).

Secondly, Bukharin's work clearly articulates a conception of the development of capitalism as a world system to a more advanced stage than that of industrial capitalism which Marx had earlier analysed. This new stage is characterized by the rise of monopoly or ‘state trusts’ within advanced capitalist states,

intensified international competition among different national monopolies leading to a quest for economic, political and military control over ‘spheres of influence’, and breaking out into destructive wars between states. These conditions are seen as inevitable results deriving from inherent tendencies in the capitalist accumulation process, at the heart of which is a supposed falling tendency in the overall average rate of profit. Altogether they are viewed as an expression of the anarchic and contradictory character of capitalism. The formation of monopolies is supposed to take place through reorganization of production by finance capitalists as a way of finding new sources of profitable investment and of exercising centralized regulation and control of the national economy. This transformation succeeds for a time at the national level but only to raise the contradictions to the level of the world economy where they can be resolved only through revolutions breaking out at different ‘weak links’ of the world-capitalist system. The idea of a necessary long-term decline in the rate of profit, and also the specific role assigned to financial enterprises as such, can be disputed. A crucial ingredient of the argument is the idea of oligopolistic rivalry and international mobility of capital as essential factors governing international relations. In this respect the argument anticipates ideas that are only now being recognized and absorbed into the orthodox theory of international trade and which, in his own time, were conspicuously neglected within the entire corpus of existing economic theory. Much of the analysis as regards a necessary tendency to uneven development between an advanced *centre* and underdeveloped *periphery* of the world economy has also been absorbed into contemporary theories of underdevelopment. Underpinning the whole argument is a curious theory of ‘social equilibrium’ and of ‘crisis’ originating from a loss of equilibrium. ‘To find the law of this equilibrium’, he suggests (1920, p. 149), ‘is the basic problem of theoretical economics and theoretical economics as a scientific system is the result of an examination of the entire capitalist system in its state of equilibrium’.

The third component is a comprehensive conception of the process of socialist construction in a backward country. These ideas came out of the practical concerns and rich intellectual ferment associated with the early period of Soviet development but have a generality and relevance extending down to current debates both in the development literature and on problems of socialist planning. The overall framework is one that conceives of socialist development as a long-drawn-out process ‘embracing a whole enormous epoch’ and going through four revolutionary phases: ideological, political, economic and technical. The process is seen as occurring in the context of a kind of war economy involving highly centralized state control, though there is an optimistic prediction of an ultimate ‘dying off of the state power’. Room is allowed for preserving and maintaining small-scale private enterprise. The agricultural sector is seen as posing special problems, due to the assumed character of peasant production, which can only be overcome through transformation by stages to collectivized large-scale production. Even so, it is firmly held (in 1919) that ‘for a long time to come small-scale peasant farming will be the predominant form of Russian agriculture’, a view which Bukharin later abandoned in support of Stalin's collectivization drive. In industry, too, small-scale industry, handicraft, and home industry are to be supported, so that the all-round strategy is one that seems quite similar to that of ‘walking on two legs’ later propounded by Mao for China. An extensive discussion is presented of almost every detail of the economic programme, from technology to public health, but little or no attention is given to issues of incentives and organizational problems of centralization/decentralization which have emerged as crucial considerations in later work.

Cohen (1973) remains a classic biography; his widow's memoirs, Larina (1993) are also of interest.

Selected works

1917. *The Economic Theory of the Leisure Class*. New York: Monthly Review Press, 1972.

1918. *Imperialism and World Economy*. New York: Monthly Review Press, 1973.

1919. (With E. Preobrazhensky.) *The ABC of Communism*. Harmondsworth: Penguin Books, 1969.

1920. The economics of the transition period. In *The Politics and Economics of the Transition Period*, ed. K.J. Tarbuck. London: Routledge & Kegan Paul, 1979.

1921. *Historical Materialism, A System of Sociology*. Ann Arbor: University of Michigan Press, 1969.

1924. *Imperialism and the Accumulation of Capital*. New York: Monthly Review Press, 1972.

Bibliography

Cohen, S. 1973. *Bukharin and the Bolshevik Revolution: A Political Biography:1888–1938*. New York: Random House.

Dobb, M. 1969. *Welfare Economics and the Economics of Socialism*. Cambridge: Cambridge University Press.

Dobb, M. 1973. *Theories of Value and Distribution since Adam Smith*. Cambridge: Cambridge University Press.

Harris, D.J. 1978. *Capital Accumulation and Income Distribution*. Stanford: Stanford University Press.

Harris, D.J. 1981. Profits, productivity, and thrift: the neoclassical theory of capital and distribution revisited. *Journal of Post-Keynesian Economics* 3(3), 359–82.

Heitman, S. 1962. Between Lenin and Stalin: Nikolai Bukharin. In *Revisionism*, ed. LeopoldLabedz. Tarbuck New York: Praeger.

Heitman, S. 1969. *Nikolai I. Bukharin: A Bibliography*. Stanford: Hoover Institution.

Kuhn, T. 1970. *The Structure of Scientific Revolutions*. 2nd edn, enlarged. Chicago: University of Chicago Press.

Larina, A. 1993. *This I Cannot Forget: The Memoirs of Nikolai Bukharin's Widow*. New York: W. W. Norton.

Latsis, S., ed. 1976. *Method and Appraisal in Economics*. Cambridge: Cambridge University Press.

Lenin, V.I. 1917. *Imperialism, the Highest Stage of Capitalism*. New York: International Publishers, 1939.

Meek, R.L. 1967. *Economics and Ideology and Other Essays*. London: Chapman & Hall.

How to cite this article

Harris, Donald J. "Bukharin, Nikolai Ivanovitch (1888–1938)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000227>
doi:10.1057/9780230226203.0174

The New Palgrave Dictionary of Economics Online

bullionist controversies (empirical evidence)

Lawrence H. Officer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There are three historical episodes in which bullionist–anti-bullionist macroeconomic debates occurred: Sweden (1745–76), England (1797–1821), and Ireland (1797–1821). Expressing the bullionist and anti-bullionist models as chains of causation facilitates presentation of empirical studies of the experiences. As a group, the studies suggest that the anti-bullionist position is more supported by the empirical evidence.

Keywords

Bank of England; Bank or Ireland; Bank Restriction Period; bullionist controversies; cointegration; Continental System; copper standard; fixed exchange rates; floating exchange rates; gold standard; Granger causality; monetarism; monetary base; purchasing power parity; quantity theory of money; real bills doctrine; Riksbank; silver standard; spurious regressions; time series analysis

Article

The bullionist periods of Sweden, England, and Ireland involved bullionist–anti-bullionist macroeconomic debates, with empirical studies vindicating largely the anti-bullionist side.

History of bullionist periods

The bullionist controversy is a debate that can occur in monetary history when a paper currency and floating exchange rate interrupt a metallic standard. The three famous bullionist periods pertain to Sweden, England and Ireland. In 1745, the Riksbank made its notes inconvertible into copper bullion, resulting in the paper daler. It was not until 1776 that the Swedish bullionist period ended, with conversion to a new currency unit (the riksdaler) on a silver standard. The English, followed by the Irish, bullionist period began in 1797, each by government order requiring the Bank of England and Bank of Ireland to cease making gold payments for its notes. Legislation, periodically renewed, solidified the orders. In 1821 the Bank of England, followed by the Bank of Ireland, resumed payment in gold, and the countries were back on a gold standard. The English episode is called the ‘Bank Restriction Period’.

The three bullionist periods involved common elements: a prior metallic standard replaced by a paper standard, a fixed exchange rate (constrained within a band around an effective mint parity) giving way to a floating rate, unusually high inflation, depreciation of the currency in the foreign-exchange and bullion markets, a sub-period of deflation, and eventual return to a specie standard and fixed exchange rate. Also, periods of war occurred both before and during the bullionist periods.

Some characteristics were shared by only two of the periods. First, the proximate cause of the Swedish and English Restrictions was a tremendous loss of reserves on the part of the Riksbank and Bank of England. This was not the case for the Bank of Ireland; British pressure induced the Irish government to suspend convertibility of Bank of Ireland notes. Second, for Sweden and England, their main trading partners remained on a metallic standard. This was not so for Ireland, with England also on paper. Third, England and Ireland returned to a gold standard at the old parity; Sweden switched from an effective copper to an effective silver standard, and banknotes were depreciated by 50 per cent in terms of silver.

Two additional features characterize all three periods. First, the macroeconomic debate centred on determination of the exchange rate and price level, and their relationship to the balance of payments and note issues of the central bank. The bullionists adopted a monetarist approach, and the anti-bullionists a non-monetarist position. Second, Parliament played a key role in the controversy. In the case of Sweden, two political parties vied for control of Parliament. The 'Caps' had a bullionist agenda, and the 'Hats' an anti-bullionist policy. Both had intellectual supporters on the outside. The British House of Commons appointed committees, in 1804 and 1810, to investigate the depreciated Irish and English currencies. Each committee produced a highly bullionist report, important in the literature; but in neither case was the report favourably received by Parliament.

Bullionist, anti-bullionist, and country-bank models

To examine the empirical literature on the bullionist controversies, each side is represented by its mainstream model of chains of causality, sequential hypotheses. Notation is $X \rightarrow Y$ ('X causes Y, with $\partial Y / \partial X > 0$'). Multiple hypotheses are W, $X \rightarrow Y$ ('$W \rightarrow Y$ and $X \rightarrow Y$') and $X \rightarrow Y$, Z ('$X \rightarrow Y$ and $X \rightarrow Z$'). The subscript f designates a foreign variable. Variables are:

- BN: central-bank notes in circulation
- BP: balance-of-payments deficit
- CN: country banknotes in circulation
- ER: exchange rate, price of foreign currency
- FR: remittances to foreign countries
- HQ: quantity and quality of harvest
- MS: money supply (M1)
- PG: price of gold
- PL: price level
- PM: price of imports
- PW: price of wheat
- TR: foreign trade restrictions

The *bullionist model* is decidedly monetarist: only monetary variables affect only monetary variables. The English-bullionist chain of causation is: $BN \rightarrow MS \rightarrow PL \rightarrow ER$, PG.
$BN \rightarrow MS$ reflects the bullionist, and correct, perception that Bank of England notes constituted the monetary base during the Restriction Period. There was a hierarchy of banks: the Bank of England (central bank), London private banks, and country banks. Bank of England notes (held as reserves by the country banks and London private banks) were non-redeemable; deposits at the Bank (held as reserves only by the London private banks) were cashable only in Bank of England notes. The country banks – but not the London private banks – issued notes. There were no legal reserve requirements for any bank; but, like all companies, banks had to settle their debts (note and deposit liabilities) in cash. Reserves of the country banks were principally deposits at the London private banks, with Bank of England notes (and, in principle, gold) for vault cash. Bank of England notes circulated in and around London, as well as in Lancashire and Norwich; country banknotes circulated elsewhere in England and Wales. During the Bank Restriction Period, the English country banks and Scottish banks 'redeemed' their notes in Bank of England notes rather than gold. This was a matter of practice rather than law.
Strictly speaking, gold coin was a component of the monetary base, but the premium on gold bullion did not have a counterpart in the premium of gold coin over Bank of England notes. There was no legal market for domestic coin in terms of paper money, and an overwhelming proportion of the gold coin nominally in circulation or newly minted was in fact hoarded or exported.
For the bullionists (and anti-bullionists), the money supply had as components Bank of England notes, country banknotes, and coin. In excluding deposits from M1, the writers of the Restriction Period were not far off the mark. First, except in London, 'deposits' generally meant time or savings deposits rather than demand deposits. Second, if interbank transactions are excluded, demand deposits typically were exchanged for cash rather than transferred to another account.
$BN \rightarrow MS$ was also asserted by the Irish bullionists, even though the banking system was looser. In and around Dublin, notes of the Dublin private banks circulated along with notes of the Bank of Ireland. Gold did not circulate, except in the north until 1808–9, when it was replaced by the notes of newly established Belfast banks. Elsewhere, local private banknotes generally dominated, but in competition with Bank of Ireland notes and, to a lesser extent, Dublin private-bankers' notes. The private banks kept their reserves in Bank of Ireland notes (and gold), and by convention their notes were redeemed in Bank of Ireland notes.
In the Swedish bullionist period, BN=MS. With little coin circulating, no commercial banks in existence, and deposits at the Riksbank representing merely the right to make withdrawals in notes, Riksbank notes essentially equalled the money supply.
$MS \rightarrow PL$ pertains to the quantity theory of money. Underlying this theory is the bullionist view that the Bank of England effectively pegged the market interest rate at five per cent, by

standing ready to discount all ‘good’ commercial bills at that rate. Thus the monetary base is perfectly elastic at the constant discount rate of five per cent, a powerful impetus to the quantity theory.

There is good reason for this view: the usury laws set a five per cent limit on annual interest on bills of exchange, and the discount rate of the Bank of England was fixed at this rate. While bill brokers could charge a commission and private banks could require a minimum balance, the Bank did not use such devices. The market discount rate (for good bills) did not exceed five per cent during the Restriction. In fact, only for about a year (beginning July 1817) did the market rate even fall below five per cent. The situation was yet stronger regarding the Bank of Ireland. Its discount rate was limited to five per cent by charter.

However, the English and Irish bullionists were wrong in inferring that the monetary base (essentially BN) could rise without limit. First, there is evidence that in historical fact the monetary base was not perfectly elastic. Only ‘good’ bills—a minority of bills—were acceptable by the Banks. Also, the Bank of England effectively regulated discounts via a rationing system. These facts act against the quantity theory but support the concept of BN as an autonomous policy variable. Second, even if the *supply* of the monetary base (essentially BN) is perfectly elastic at the pegged market interest rate, BN is limited by the *demand* for the monetary base. The Bank of England and Bank of Ireland could not induce the private sector to hold more BN than demanded. BN was viewed by the bullionists as the first link in the causal chain; but it is an endogenous variable. A low level of economic activity could hold down the demand for BN.

$PL \rightarrow ER$ is the purchasing-power-parity theory (given PL_f), the causal nature of which is generally ignored in the modern literature. $PL \rightarrow PG$ involves a relatively unchanged PG_f, for, under perfect markets, PG is the product of ER and PG_f. PG was not as interesting to the Swedish and Irish bullionists as it was to the English. Sweden had been on a copper standard; the concern in Ireland was depreciation of the Irish currency against the British. For the Swedish and English protagonists, foreign exchange was Continental currencies.

For most Swedish and Irish bullionists, the latter part of the chain is merely $MS \rightarrow PL$, ER. The price level and exchange rate are co-determined by the money stock. Some Irish bullionists allowed for a changing foreign (English) price level, so the hypothesis becomes MS/MS_f (or BN/BN_f) $\rightarrow$ ER.

The English *anti-bullionist model* involves a balance-of-payments theory of the exchange rate, with demand for and supply of bills of exchange represented by the payments deficit (BP), yielding ER and PG. The state of the harvest, a real factor, determines the domestic price of grain, represented by the price of wheat (PW). The exchange rate is an ingredient in the price of imports, which, together with PW, determines PL. These anti-bullionists saw three principal determinants of BP, that is, of shifts in the demand for or supply of foreign exchange: PW, foreign trade restrictions (wartime restraints: the Continental System and the American embargo), and foreign remittances (external government payments: direct military expenditure and subsidies to allied countries). The English anti-bullionist causal chain is:

$$\begin{array}{ccccccc} 1/HQ & \rightarrow & PW & & \rightarrow & & PL \rightarrow BN \\ & & \downarrow & & & & \uparrow \\ TR, FR & \rightarrow & BP & \rightarrow & ER, PG & \rightarrow & PM \end{array}$$

In emphasizing the price of wheat, the anti-bullionists recognized the highly agrarian state of the British economy, notwithstanding the industrial revolution in progress. The emphasis on wartime interference with trade and on external military expenditure reflected the French Revolutionary and Napoleonic Wars, in which Britain was engaged for much of the Bank Restriction Period.

For the Irish anti-bullionists, concerned with the English exchange, TR and PG were unimportant. They did not make explicit the connection of PW and PM to PL, and FR took the form of payments to absentee landlords in England. Some consolidated the trade balance, interest payments, net capital exports, and FR, to compose (and presumably shift) BP in the causal chain. They left unclear the mechanism from BP to PL. The Swedish anti-bullionists had the chain: $BP \rightarrow ER \rightarrow PM \rightarrow PL$, allowing real shocks to operate on BP.

The anti-bullionists used the ‘real-bills’ doctrine to reverse the bullionist $BN \rightarrow PL$ causation. They accepted that the Bank behaved passively in its note issuance, but used the real-bills theory to demonstrate that excess issue (beyond the ‘needs of trade’) would be returned to the Bank instead of acting to increase the price level monetarily. Only non-monetary forces could cause real income and then the price level to increase, and would underlie the demand for discounting to finance a higher volume of transactions, whence $PL \rightarrow BN$. The Irish bullionists also propounded the real-bills doctrine (for the Bank of Ireland), although some saw ER playing the role of PL.

Bullionists in all three periods essentially inverted the real-bills theory by offering the *policy rule* that central-bank note issuance should be oriented to the exchange rate and (for the English bullionists) gold price: ER, $PG \rightarrow 1 / BN$.

Extension to country banks

A subsidiary part of the English and Irish bullionist controversies was the extent to which the country banks (in Ireland, including Dublin private banks) could affect the money supply independent of the central bank. Should the first hypothesis in the bullionist chain, $BN \rightarrow MS$, incorporate CN naturally as $BN \rightarrow CN \rightarrow MS$ (country banks unable to vary their note issues independent of the central bank)? Or should the hypothesis be $(BN + CN) \rightarrow MS$ (the central bank and country banks able either jointly or separately to change their issues)? Or should the hypothesis be $CN \rightarrow MS$ (only the country banks, not the central bank, having the power to change the money supply)? The question was answered differently by groups that cut across the bullionist–anti-bullionist line.

The correct hypothesis is not clear, because of the environment in which banks operated. Among the complicating, and largely unknown, elements are the extents to which (*a*) one-time replacement of gold by central-bank notes in reserves altered country-bank policy regarding reserve ratios, (*b*) country-bank reserve ratios varied over time, (*c*) public preference for central-bank over country-bank notes changed in particular geographic areas and over time, (*d*) circulation of counterfeit notes and unlicensed-bank notes affected the demand for and supply of country-bank and central-bank notes, and (*e*) London private banks were prepared to run down their reserve ratios to accommodate country-bank demand for additional reserves.

Empirical studies: visual comparison of movements of variables

The empirical studies examined here make use of quantitative information to test one or more component hypotheses of the bullionist or anti-bullionist models. It is logical to begin with contemporary studies, as it is the hypotheses of contemporary authors that are delineated in the previous sections.

All contemporary investigations use a simple technique: visual inspection of sets of figures, formal tables, or charts. The earliest such studies pertain to the Ireland bullionist period, with BN and BN_f the note circulations of the Bank of Ireland and Bank of England. Parnell (1804), Foster (1804) and the 1804 Currency Report (in Fetter, 1955) find that $BN \rightarrow ER$ is confirmed. Ó Gráda (1993) and Fetter (1955) criticize the Report for its small number of observations and selective observations. These criticisms can be extended to Parnell, but not to Foster. The report of 1804 and Parnell also claim successful testing of $BN / BN_f \rightarrow ER$. Ó Gráda (1991) finds this part of the Report misleading in several respects; but the Report is to be commended for making specific allowance for the replacement of gold coin by notes. The Report also claims to disprove $BP \rightarrow ER$, via computation of a net balance-of-payments surplus. However, this proves little, because there is no representation of shifts in the demand for or supply of bills on London.

Contemporary empirical work on the English bullionist period begins with Ricardo (1811), whose positive finding of $BN \rightarrow ER$ (Hamburg exchange) is reinforced by observation of a lagged effect and by accounting for replacement of gold coin by Bank of England notes. Galton (1813) confirms that $BN \rightarrow ER$, PG. Anonymous (1819) sees mixed evidence for that hypothesis, but observes that grain imports and FR (not precisely defined) affect the exchange rate – the first results in favour of anti-bullionism.

There is a hiatus of more than a century, but three groupings of subsequent work do not merit review. First is any investigation, such as Silberling (1924), involving the London price of the Spanish dollar to represent the exchange rate. That choice is methodologically unsound. Britain was on a suspended gold (not silver) standard, and the Spanish silver dollar was not a circulating coin in Hamburg, the main foreign-exchange market. Second are tests making use of Silberling-developed series of Bank of England total advances and their private versus public components. These series have been shown to be seriously inconsistent with the Bank's published data. Third, and most unfortunate, are all studies using ‘data’ on country banknote circulation. There exist no true data on country banknote circulation in England, or private banknote circulation in Ireland, during the bullionist period. Further, with no legal or fixed reserve ratio of note liabilities to cash, the circulation of the Bank of England, or Bank of Ireland, cannot be used to infer that of the private banks. Private banks were required to register at the Stamp Office and pay a stamp tax on notes prior to issuance. Some have used stamp-tax data to develop proxy CN series for England, based on the value of country banknotes stamped; but the series are based on assumptions so tenuous as to make the series unusable.

Silberling (1924) develops an annual series for FR (‘extraordinary foreign payments’), consisting of grain imports over a normal amount, Continental British war expenditures, and subsidies to foreign states. Using various definitions of FR, based largely on Silberling, Angell (1926) shows that $FR \rightarrow ER$, but can find no causal relationship between PL and ER. This result, favourable to anti-bullionism, is supported by Morgan (1939; 1943) and Viner (1937). Morgan rejects $BN \rightarrow PL$, but accepts $PL \rightarrow BN$. His only finding not supportive of anti-bullionism is the lack of a relationship between PW and PL or BN.

Gayer, Rostow, and Schwartz (1953, p. 932) support $BP \rightarrow ER$; but they represent BP by the balance of trade, the data of which are crude. For the Swedish period, Eagly (1971) and Bernholz (1982; 2003) support $BN \rightarrow PL$, ER, favourable to bullionism.

This entire body of literature must be viewed with caution. First, interpretation of relationships among variables is subjective when data are merely tabulated or plotted. Second, macroeconomic variables are generally non-stationary, leading to the possible outcome of ‘spurious regression’.

Empirical studies: time-series analysis

Myhrman (1976) computes annual growth rates of BN and PL, for Sweden and England, and argues that $BN \rightarrow PL$. Jonung (1976) does the same for Sweden alone. Transforming data to growth rates could yield stationarity. In a joint test of bullionist and anti-bullionist hypotheses, Arnon (1990) regresses PL on PW, BN, and a trend. He finds that BN contributes more to the regression than PW. The variables are transformed to correct for serial correlation, which could correct spurious regression.

Formal time-series analysis in the bullionist literature begins with Ó Gráda (1989; 1993). For England, he cannot reject a cointegration relationship between logPL and logBN. This means that there is no long-term equilibrium between the variables, a failure of support for either bullionism or anti-bullionism The same negative result holds for Ireland, with BN/BN_f used in place of BN.

Nachane and Hatekar (1995) use Granger causality and cointegration techniques for England. Their variables are PL, ER, PG, BP, and BN/Y (transformed to logarithms except for BP, the only non-stationary variable), where Y is real output. Their results are $ER \rightarrow PL$, $PL \rightarrow BN / Y$ (with PL and BN/Y the only cointegrated pair of variables), and $BP \rightarrow ER$, PG. The findings are strongly supportive of anti-bullionism; but measuring the money supply in relation to output is outside the mainstream controversy.

The analyses of Ó Gráda and Nachane–Hatekar are restricted to bivariate econometrics. Officer (2000) applies multivariate testing to PL, ER, BN, FR, and PW, for England. Non-stationarity cannot be rejected, but cointegration is rejected. The logarithmic variables are first-differenced (to achieve stationarity), and Granger causality testing along with innovation analysis is applied. Results are mixed for bullionism, but unambiguously favourable to anti-bullionism. For example, the real-bills doctrine, PL → BN, receives stronger support than does the quantity theory, BN → PL.

It is logical that the time period for testing hypotheses be strictly within the pertinent bullionist period, because the alternative (bullionist versus anti-bullionist) models are geared to a paper standard and floating exchange rate. As his sample, Officer uses the 96 quarters encompassed by the Bank Restriction Period (1797–2 to 1821–1). Nachane and Hatekar employ annual data, and extend the time period to 1838. Ó Gráda has quarterly observations, but begins his time periods prior to 1797.

Nachane and Hatekar can also be criticized for using the exchange rate on Paris rather than Hamburg to represent ER. There are no quotations on Paris until 1802 (whence they lose observations), and historians agree that the Hamburg exchange was more representative during wartime.

To conclude: certainly, at least for England, the anti-bullionist position receives greater support (or less contradiction) than the bullionist side of the controversy. This result is inconsistent with modern macroeconomics. The anti-bullionist approach to the exchange rate (a flow theory) and monetary policy (passive, accommodating the price level) has been superseded in modern theory. Also, modern monetarism emanates from bullionism.

See Also

- cointegration
- Granger–Sims causality
- monetarism
- purchasing power parity
- quantity theory of money
- real bills doctrine
- real bills doctrine versus the quantity theory
- spurious regressions

Bibliography

Angell, J.W. 1926. *The Theory of International Prices*. Cambridge, MA: Harvard University Press.

Anonymous. 1819. Two tables…showing the rates of exchange on Hamburgh compared with the amount of bank notes, and the price of gold, and with the foreign expenditure, and the value of grain imported from the year 1793 to 1819. *Pamphleteer* 15, 281–6.

Arnon, A. 1990. What Thomas Tooke (and Ricardo) could have known had they constructed price indices. In *Keynes, Macroeconomics and Method*, ed. D.E. Moggeridge. Aldershot:

Edward Elgar.

Bernholz, P. 1982. *Flexible Exchange Rates in Historical Perspective*. Princeton, NJ: International Finance Section, Princeton University.

Bernholz, P. 2003. *Monetary Regimes and Inflation*. Cheltenham: Edward Elgar.

Eagly, R.V. 1971. *The Swedish Bullionist Controversy*. Philadelphia: American Philosophical Society.

Fetter, F.W. 1955. *The Irish Pound, 1797–1826*. Evanston, IL: Northwestern University Press.

Foster, J.L. 1804. *An Essay on the Principles of Commercial Exchanges*. London: J. Hatchard.

Galton, S.T. 1813. *A Chart, Exhibiting the Relation Between the Amount of Bank of England Notes in Circulation, the Rate of Foreign Exchanges, and the Price of Gold and Silver Bullion and of Wheat*. London: J. Johnson.

Gayer, A.D., Rostow, W.W. and Schwartz, A.J. 1953. *The Growth and Fluctuation of the British Economy, 1790–1850*. Oxford: Clarendon Press.

Jonung, L. 1976. Money and prices in Sweden, 1732–1972. *Scandinavian Journal of Economics* 78, 40–58.

Morgan, E.V. 1939. Some aspects of the Bank Restriction Period, 1797–1821. *Economic History* 4, 205–21.

Morgan, E.V. 1943. *The Theory and Practice of Central Banking, 1797–1913*. Cambridge: Cambridge University Press.

Myhrman, J. 1976. Experiences of flexible exchange rates in earlier periods: theories, evidence, and a new view. *Scandinavian Journal of Economics* 76, 169–96.

Nachane, D.M. and Hatekar, N.R. 1995. The Bullionist Controversy: an empirical reappraisal. *Manchester School of Economic and Social Studies* 63, 412–25.

Officer, L.H. 2000. The Bullionist Controversy: a time-series analysis. *International Journal of Finance & Economics* 5, 197–209.

Ó Gráda, C. 1989. The paper pounds of 1797–1821: a co-integration analysis. Working paper. Dublin: Centre for Economic Research, University College.

Ó Gráda, C. 1991. Reassessing the Irish Pound Report of 1804. *Bulletin of Economic Research* 43, 5–19.

Ó Gráda, C. 1993. The Irish Paper Pound of 1797–1820: some cliometrics of the Bullionist Debate. *Oxford Economic Papers* 45, 148–56.

Parnell, H. 1804. *Observations upon the State of Currency in Ireland*. Dublin: M. N. Mahon.

Pressnell, L.S. 1956. *Country Banking in the Industrial Revolution*. Oxford: Clarendon Press.

Ricardo, D. 1811. The high price of bullion. In *The Works and Correspondence of David Ricardo*, vol. 3, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Silberling, N.J. 1924. Financial and monetary policy of Great Britain during the Napoleonic Wars. *Quarterly Journal of Economics* 38, 214–33.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

How to cite this article

Officer, Lawrence H. "bullionist controversies (empirical evidence)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000334> doi:10.1057/9780230226203.0175

The New Palgrave Dictionary of Economics Online

bundling and tying

Barry Nalebuff
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bundling can be thought of as akin to a volume discount, but one where the volume is based on aggregate sales across products. Instead of offering a discount for buying two apples rather than one, the customer is given a better price for buying an apple and an orange together. Bundling may be used to reduce cost and improve quality, and for price discrimination. While the Chicago School has argued that a monopolist cannot gain by bundling its monopoly good with a competitive product, recent work suggests that in a dynamic game bundling can help protect and leverage market power.

Keywords

antitrust policy; bundle discounting; bundling; Chicago School; complementarities; consumer surplus; Cournot, A.; envelope theorem; market power; Markup; metering; mixed bundling; one-monopoly profit argument; price discrimination; pure bundling; two-part tariffs; tying

Article

Bundling is a prevalent feature of pricing. It is akin to a volume discount, but where the volume is based on aggregate sales across products. Instead of offering a discount for buying two apples rather than one, the customer is given a better price for buying an apple and an orange together.

Under pure bundling, two goods A and B are sold together only as a package. Under mixed bundling, customers can also buy each good. Typically, the bundle is offered at a discount to the individual prices. Mixed bundling is the most general case. A pure bundle can be thought of as a case where the individual prices exceed the bundle price, so that no one has an incentive to purchase anything but the bundle. Tying can also be viewed as a special case of mixed bundling; customers are offered prices for A and B together or for B alone, but not A without B.

The first to study bundling was Cournot (1838), who showed how it solves a double markup problem for complementary products. Bundling may increase efficiency more directly by improving quality and

reducing cost (see Evans and Salinger, 2004; 2005). Manufacturers gain scale economies by standardizing the combination of goods and guaranteeing that the components work together.

The early bundling literature debated whether it was used to leverage market power or price discriminate. This debate was stimulated by a series of cases in which the courts viewed bundling (and tying) as anti-competitive. Director and Levi's (1956) and Stigler's (1963) influential Chicago School argument claimed that a monopolist cannot gain by leveraging its power from one market to another. Starting with Whinston (1990), the current literature suggests that, in a dynamic setting, bundling can profitably leverage market power by deterring entry, excluding one-good rivals, and amplifying existing market power. Three review articles provide a guide to the literature and antitrust cases: Kaplow (1985), Nalebuff (2003a), and Kobayashi (2005).

The Chicago School argument

In response to *United States v. Loew's* (1962), Stigler (1963) argued that block booking (selling movies bundled rather than individually) was best viewed as price discrimination. He argued that a monopolist in product A could not make more money by requiring buyers to take a product B that was competitively available – the alternative strategy of selling A alone for a price of p-c, where p is the bundle price and c the marginal cost of B, is more profitable. Any sale of A at price p-c would be just as profitable as selling the bundle at p. Yet anyone willing to buy the bundle at p would also be willing to buy A alone at p-c, as, by assumption, B is available at the competitive price of c. Bundling is weakly worse as it might cause the firm to lose sales to customers who value A at p-c but do not value B at c and thus do not buy the bundle. This has become known as the ‘one-monopoly profit’ or ‘Chicago School’ argument (see Director and Levi, 1956; Bork, 1978).

If leverage does not explain bundling, something else must. Stigler suggested price discrimination.

Price discrimination

The idea of bundling (and tying) as price discrimination dates from Bowman (1957) and Burstein (1960). As Burstein noted, a monopolist would generally like to employ a two-part tariff in pricing. Requiring customers to buy an overpriced B is an indirect way to charge a lump-sum fee.

If the monopolist starts from a profit-maximizing price, then (by the envelope theorem) profits lost from cutting price will be very small. In contrast, existing consumers will gain a great deal, and so will be willing to buy B at a inflated price in return for a lower-priced A.

The problem is that other producers of B end up excluded. Customers of A won't switch to lower-priced B goods because they don't want to lose the discount on A. While the monopolist could have used a two-part tariff directly, such pricing schedules seem rare in practice. Bundle pricing as a two-part tariff is explored in Mathewson and Winter (1997) and Nalebuff (2004b).

Two-part pricing becomes even more effective when B's demand is correlated with A's value. This leads to metering. For example, a firm selling printers would like to charge high-value customers more. But customer valuation may be unobservable. However, if value is correlated with usage, then a per-page charge would allow the seller to charge high-value customers more. A per-page charge could be levied

directly, although that would require monitoring usage. In practice, sellers patent the shape of their toner cartridge, thus requiring users to buy toner at a premium price.

These results rely on B's demand being either elastic or heterogeneous. Bundling permits price discrimination even when A and B are consumed in fixed amounts. Consider movies. If regional variation in the valuation of two movies is negatively correlated, a distributor can profit more by pricing the movies as a package than by selling them à la carte. Bundling reduces demand heterogeneity and thus captures more of consumer surplus (see Adams and Yellen, 1976).

The advantages of the bundle discount strategy are remarkably general. McAfee, McMillan and Whinston (1989) show that, for any two goods independent in value, a firm with market power will find an advantage offering them at a bundle discount (holding individual prices constant) – an impressive result, given the near endless opportunities for bundling products with independent values.

One intuition for their argument is that discounting via bundling leads to twice the demand expansion for the same price reduction. Consider the offer to lower A's price by one dollar if you buy B. The cost of the offer is one dollar to all customers who would have bought both A and B at the previous prices. The gain is the new demand from customers who were buying B but not A, as they now have an opportunity to get A at a dollar off. If A was priced optimally to begin with, then the incremental profit from increased demand should just offset the lost revenue on existing customers. (Here demand independence is critical, as it implies that customers buying B are representative of the entire A market.) So far, everything is a wash. However, the dollar off A if you buy B is the same as a dollar off B if you buy A. Thus there is a second set of incremental customers: those already buying A but on the margin on B. Demand for B expands without imposing any further cost in terms of lost revenue. The ability of the bundle to expand demand on two fronts for one discount is the ‘special sauce’ behind bundling.

Bundling to leverage monopoly

The recent re-examination of bundling as leveraging market power and foreclosing rivals uses dynamic reasoning, which is absent in the Chicago argument.

For example, a monopolist in A might bundle A with B to drive rivals out of the B market. The motivation could be to monopolize what was previously a competitive B market, or to protect the A monopoly. Eliminating firms in the B market protects A if being in the B market facilitates entry into A. The US Department of Justice (1998) argued thus in explaining Microsoft's motivation to bundle Explorer with Windows – defeating Netscape would prevent it from threatening Microsoft's operating system monopoly.

The first dynamic model appears in Whinston (1990), where the bundler has market power in both A and B and uses the bundle to deter potential entrants. The monopolist is concerned that there may be a rival who can produce B at a lower cost. In defence, it commits itself to sell A only along with B. Thus, if a rival were to create a lower-cost B, the monopolist would not concede, as that would cost it its profits in A sales. Since the monopolist is committed to selling A only along with B, it would have to subsidize B in order to sell A. Even more efficient B good rivals won't enter, realizing that they won't win; this preserves monopoly profits in B.

Nalebuff (2004a) offers a second perspective. Absent entry, the dual monopolist gains via price discrimination. With entry (and heterogeneous consumer preferences), the firm would rather respond

with a bundle than with head-to-head competition (and thereby lose all profits in the B market).
The incumbent's bundling reduces the potential market available to the entrant. The entrant is mostly limited to those customers who like B but not A. This market may not be large enough to cover costs of entry or to achieve a minimum efficient scale (Carlton and Waldman, 2002).
The bundling models illustrate the challenge for anyone contemplating entry against Microsoft Office. Given the large discount for buying the Office bundle, a firm that developed a better word-processing program (and nothing else) would find its market limited to those who value word processing, but not spreadsheets or presentations. The entrant could try to sell to those who already have Word, but that would limit the price to its product's incremental value over Word, which is much less than what it can charge customers who don't already have Word.
A firm could always develop a rival bundle of products. But this also discourages entry, as it is much harder to develop two better products than one. Furthermore, it turns out that bundle-against-bundle competition is particularly fierce (see Matutes and Regibeau, 1992).
These examples of bundling emphasized the use of pure bundles as a way of protecting and leveraging market power. Even with mixed bundling, firms can achieve similar results by keeping the component prices artificially high.
A bundle discount may be large due to a low bundle price or high individual prices, prices that might exceed monopoly levels. Although entry is blocked in both cases, the welfare implications are different, as discussed in Greenlee, Reitman and Sibley (2004). Bundling can be used to create a horizontal price squeeze, an issue considered by the Supreme Court in *Ortho Diagnostic v. Abbott Lab*. (1996) and developed in Nalebuff (2005).
A bundle discount leads to foreclosure if even the *monopolist* could not afford to sell B at a large enough discount to offset the loss of the bundle discount. Exclusionary bundling arises when the incremental price for an A–B bundle over A alone is less than the long-run average variable costs of B.

Bundling complements

An incentive to bundle arises when two products are perfect complements, so that customers care only about their combined price. Cournot (1838) considered copper and zinc, which combine to produce brass; a more modern example would be hardware and software.
Two monopolists selling A and B independently will charge inefficiently high prices. Were the two firms to merge or coordinate their pricing, they can lower prices and raise profits. The gain from bundling complements is the horizontal equivalent of vertical integration to avoid double marginalization. As consumers and firms are both better off, this is a Pareto improvement.
The situation is more complicated if there are multiple producers of A and B. Nalebuff (2000) and Choi (2001) consider the case where two firms are able to solve the coordination problem while their rivals are not. This issue arose in 2001 when the European Commission blocked the proposed US$42 billion merger between General Electric and Honeywell. The Commission was concerned that the merger would allow the combined firm to better coordinate the pricing of airplane engines and avionics, and give it an advantage over engine-only rivals such as Rolls Royce or avionics-only rivals such as Thales or Rockwell Collins (see Nalebuff, 2003b, for a cautionary note).
Bundling can change competition in two ways. When a bundle competes against components, the

bundled seller is better able to coordinate pricing and gains share against his rivals. Profits may not rise as rivals respond to their reduced market share with lower prices. When there is bundle-against-bundle competition, as shown by Matutes and Regibeau (1992), prices are the lowest of all, and profits fall substantially. Customers benefit from the lower prices but lose the ability to mix and match and thereby buy their ideal mix of product.

There may also be a combination of these two effects. With an imbalance between A and B producers, only some firms are able to offer bundles, and these firms compete aggressively. The left-out firms have only one good and end up disadvantaged; see Gans and King's (2004) analysis of bundle discounts offered by supermarket and gasoline retailers in Australia.

Conclusions

There is no grand unification theory of bundling. The decision to bundle is connected both to product design and to pricing. While price discounts are typically pro-competitive, in some cases bundling creates a cause for antitrust concern as it can be used to protect and leverage market power.

See Also

- price discrimination (theory)

Bibliography

Adams, W. and Yellen, J. 1976. Commodity bundling and the burden of monopoly. *Quarterly Journal of Economics* 90, 475–98.

Bork, R. 1978. *The Antitrust Paradox: A Policy at War with Itself*. New York: Basic Books.

Bowman, W. 1957. Tying arrangements and the leverage problem. *Yale Law Journal* 67, 19–36.

Burstein, M. 1960. The economics of tie-in sales. *Review of Economics and Statistics* 42, 68–73.

Carlton, D. and Waldman, M. 2002. The strategic use of tying to preserve and create market power in evolving industries. *Rand Journal of Economics* 33, 194–220.

Choi, J. 2001. A theory of mixed bundling applied to the GE-Honeywell merger. *Antitrust Magazine* 16, 32–3.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. Trans. N. Bacon as *Researches into the Mathematical Principles of the Theory of Wealth*. Mountain Center, CA: James and Gordon, 1995.

Director, A. and Levi, E. 1956. Law and the future: trade regulation. *Northwestern University Law Review* 51, 281–96.

Evans, D. and Salinger, M. 2004. An empirical analysis of bundling and tying: over-the-counter pain relief and cold medicines. Working Paper No. 1297. Munich: CESifo.

Evans, D. and Salinger, M. 2005. Why do firms bundle and tie? Evidence from competitive markets and implications for tying law. *Yale Journal on Regulation* 22, 37–89.

Gans, J. and King, S. 2004. Supermarkets and shopper dockets: the Australian experience. *Australian Economic Review* 37, 311–16.

Greenlee, P., Reitman, D. and Sibley, D. 2004. An antitrust analysis of bundled loyalty discounts. Discussion Paper No. 04–13. Economic Analysis Group, Antitrust Division, US Department of Justice.

Kaplow, L. 1985. Extension of monopoly power through leverage. *Columbia Law Review* 85, 515–54.

Kobayashi, B. 2005. Does economics provide a reliable guide to regulating commodity bundling by firms? A survey of the economic literature. *Journal of Competition Law and Economics* 1, 707–46.

Matutes, C. and Regibeau, P. 1992. Compatibility and bundling of complementary goods in a duopoly. *Journal of Industrial Economics* 40, 37–54.

Mathewson, F. and Winter, R. 1997. Tying as a response to demand uncertainty. *RAND Journal of Economics* 28, 566–83.

McAfee, R., McMillan, J. and Whinston, M. 1989. Multiproduct monopoly, commodity bundling, and correlation of values. *Quarterly Journal of Economics* 104, 371–84.

Nalebuff, B. 2000. Competing against bundles. In *Incentives, Organization, and Public Economics*, ed. P. Hammond and G. Myles. Oxford: Oxford University Press.

Nalebuff, B. 2003a. Bundling, tying, and portfolio effects. DTI Economics Papers No. 1. Online. Available at http://www.dti.gov.uk/files/file14774.pdf and http://www.dti.gov.uk/files/file14775.pdf h, accessed 4 July 2006.

Nalebuff, B. 2003b. Bundling and the GE–Honeywell merger. In *The Antitrust Revolution*, 4th edn. ed. J. Kwoka and L. White. New York: Oxford University Press.

Nalebuff, B. 2004a. Bundling as an entry barrier. *Quarterly Journal of Economics* 119, 159–87.

Nalebuff, B. 2004b. Bundling as a way to leverage monopoly. Working Paper No. ES-36, Yale School of Management.

Nalebuff, B. 2005. Exclusionary bundling. *Antitrust Bulletin* 50, 321–70.

Ortho Diagnostic Sys. v. Abbott Lab., 920 F. Supp. 455 (S.D.N.Y. 1996).

Stigler, G. 1963. US v. Loew's Inc.: a note on block booking. *1963 Supreme Court Review* (1964), 152–7. Reprinted in G. Stigler, ed., *The Organization of Industry*. Chicago: University of Chicago Press, 1968.

United States v. Loew's Inc. 371 U.S. 38 (1962).

US Department of Justice. 1998. *The United States v. Microsoft Corporation*. Civil Action No. 98–1232. Online. Available at http://www.usdoj.gov/atr/cases/f1700/1763.htm, accessed 4 July 2006.

Whinston, M. 1990. Tying foreclosure, and exclusion. *American Economic Review* 80, 837–59.

How to cite this article

Nalebuff, Barry. "bundling and tying." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000304> doi:10.1057/9780230226203.0176

The New Palgrave Dictionary of Economics Online

bureaucracy

Mancur Olson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bureaucracy in both businesses and governments continues to grow despite its unpopularity. Falling transport and communication costs have created global markets. The rising relative importance of firms with new technologies and methods often unsuited to market transfer via licensing of patents has given rise to multinational corporations with transnational bureaucracies. Government bureaucracies typically produce indivisible goods contributions to which by individual bureaucrats cannot be measured, giving rise to red tape and enabling bureaucracies to exploit society's demand for their products. Bureaucracies may not be highly efficient, but market failures that give rise to them also make them inevitable.

Keywords

asymmetrical information; bureaucracy; bureaucratic growth; business bureaucracy; Coase, R.; communication costs; economies of scale; firm size; indivisibilities; market failure; markets in the firm; multinational firms; Niskanen, W.; Olson, M.; output measurement; patents; portfolio investment; profit centres; red tape; size of government; social production function; transaction costs; transfer of technology; transport costs; Tullock, G.; Weber, M.; Williamson, O.

Article

The study of bureaucracy has to deal with an elemental paradox. The role of bureaucracy has obviously increased dramatically in modern times. This is true not only of government bureaucracies but business bureaucracies as well. Though there were a few bureaucracies of significant size in pre-industrial times, such as the hierarchy of the Roman Catholic Church and the civil services of various Chinese empires, they were clearly exceptional. By contrast, a very large proportion of the total resources in the developed nations are controlled by either governmental or private bureaucracies. The role of governmental bureaucracies, at least, has increased with some rapidity within the last few decades. The increase in the use of bureaucracies has occurred in so many countries that it could hardly be due entirely to chance,

and thus must be due to what are, in some sense, social choices to use more bureaucracy.
Normally, when there are great increases in the demand for or use of some product or instrumentality, this is accompanied by independent evidence of enthusiasm for the product or instrumentality in question. When a society experiences a great increase in the demand for automobiles or for personal computers, there is at the same time a considerable amount of favourable commentary about whatever product is experiencing the boom in demand. There is pride in automobile ownership or awe at the power or compactness of personal computers. Nothing is more natural than that people's choices should be influenced by enthusiasms.
But where is the enthusiasm for bureaucracy that might have been expected to accompany the dramatic increase in the use of bureaucratic mechanisms? Any such enthusiasm is difficult to discern, and there are many conspicuous examples of dislike (or even contempt) of bureaucracy. Some of this negativism may be traced to particular ideological traditions, but this is not sufficient to explain the negativism; the problem is not only that the prevalence of the relevant ideology needs to be explained, but also that the lack of enthusiasm for bureaucracy prevails in a wide variety of ideological and cultural contexts and tends to apply (at least to some extent) to business as well as to governmental bureaucracies. There is no doubt that ‘red tape’ is viewed negatively by almost everyone, and that it is associated with bureaucracy, and especially governmental bureaucracy; the phrase is derived from the colour of the ribbons that were once used to tie folders of papers in the British government.
Some strands of the literature on bureaucracy are called into question by the paradox. Much of the admiring literature on bureaucracy is difficult to reconcile with the negative popular image of bureaucracy, whereas much of the negative literature suffers from the lack of any explanation of why virtually all societies, at least implicitly, keep choosing to use the instrumentality that is alleged to be so faulty.
Perhaps the most influential scholarly analysis of bureaucracy is not by an economist, but rather by the sociologist and historian, Max Weber. According to Weber:

> … the fully developed bureaucratic mechanism compares with other organizations exactly as does the machine with the non-mechanical modes of production …
> Precision, speed, unambiguity, knowledge of the files, continuity, discretion, unity, strict subordination, reduction of friction and of material and personal costs – these are raised to the optimum point in the strictly bureaucratic administration (1946, p. 214).

Although also critical of ‘bureaucratic domination’, Weber's more positive view of bureaucracy has been influential in sociology and political science. Yet it does not appear to have generated systematic or quantitative empirical studies that have tended to provide any confirmation for it, and it surely is not in accord with the popular image of bureaucracy. Weber himself fails to identify any strong incentives in bureaucracies that would lead to efficient allocations of resources or to high levels of innovation.
Similarly, the popular pejorative view of bureaucracy is inadequate to the extent that it offers no explanation why modern societies choose or accept an increasing degree of bureaucratization. There is, admittedly, a rapidly growing economic literature on the growth of government that attempts to identify incentives that lead to a supra-optimal size of government. Examining this large literature would take us a long way from bureaucracy, and it has not in any case yet advanced to the point of generating a professional consensus on any incentive that would systematically bring about the overuse of

government and thus of governmental bureaucracy, though some contributions (e.g. Mueller and Murrell, 1985) are extremely promising. But even dramatic success in the literature on the growth of government would not be sufficient to solve the problem, as it would leave us with no explanation of the growth in modern times of business and other private bureaucracies.

Since an explanation of the growth of private bureaucracies is needed, and since an inquiry which begins with the growth of private bureaucracies may obtain some modest degree of detachment from the ideological controversies about the appropriate role of government, it may be best to consider private bureaucracies first. Here the basic question that must be answered is, ‘Why do firms with hierarchies of employees exist?’ Familiar economic theory explains that markets can under the appropriate conditions allocate resources efficiently, so we must ask why individuals in the business hierarchy, and owners of the buildings and equipment that a typical corporation uses, do not use the price signals of the market to coordinate their everyday interaction. As Ronald Coase pointed out in somewhat different language in his seminal article on ‘The Nature of the Firm’ (1937), the survival of firms with hierarchies of long-run employees and long-term ownership of complementary fixed capital can only be explained by a kind of market failure. The type of market failure that Coase, and Williamson (1964, 1975 and 1985) and the other economists that have developed the very important literature on private hierarchies have emphasized is ‘transactions costs’. It would cost too much to contract out each day each of the very many separate tasks that are usually needed in any complex productive process, so in many cases it pays to forego the use of the market and to make long-term deals with employees who will perform such tasks each day as their superiors instruct them to do and receive in turn a regular salary. Though most of the literature in this tradition emphasizes only transactions costs, it is important to note that any market failure, such as that arising from an externality, could provide the incentive for the establishment of a firm that would internalize the externality, and all but the smallest firms have bureaucracies.

Though the foregoing argument also applies to small firms of the kind that predominated in pre-industrial times, there have been some changes since the industrial revolution that, within this Coasian–Williamson framework, can provide important insights into the growth of business bureaucracies. One factor that made for larger and more bureaucratic firms was the discovery of technologies subject to indivisibilities that only a large enterprise can profitably exploit.

But the extraordinary improvement in the technologies of transportation and communication was probably far more important. Reductions in transportation and communication costs make it economic for firms to draw factors of production from farther away and also make it profitable for a firm to sell its output over a wider area. When transportation and communication technologies make it profitable for many firms to operate at a global rather than a village level, some very large firms can emerge. The improved transportation and communication also make it possible to coordinate the activities of a firm over a larger area. Superficial observers of the emergence of large firms have supposed that this growth of firm size entails a reduction in competition and a growth of monopoly. In fact, the dramatic reductions in transportation and communication costs have, of course, also increased the opportunities for market transactions over great distances, so the size of the market and the number of firms to which the typical consumer has access has (in the absence of extra trade barriers) also increased. At least in the Common Market or the United States, the average consumer, even if purchasing a product such as an automobile that is produced under greater-than-average economies of scale, has more firms competing for his business than did the average consumer in the typical rural village before the industrial revolution. Thus we see that the growth of business bureaucracy and the expansion of competitive

markets are by no means necessarily obverse tendencies, but rather the kinds of things that often happen together.

The technologies that facilitated larger markets and larger firms also gradually led to the discovery of better methods of governing large-scale business organizations, as the historian Alfred D. Chandler has shown in some seminal historical studies of what he has called *The Visible Hand* (1977; see also 1962 and 1980). Several of these innovations occurred in the unprecedently large and geographically scattered railroads in the 19th-century United States, and many involved the creation of separate ‘profit centres’ and other devices that enabled larger firms to use market mechanisms to fulfill some functions within the firm (Williamson, 1985). This suggests that the costs and control losses in bureaucracies are still very considerable, so that business bureaucracy can only be explained in terms of rather substantial costs of using markets. The same conclusion emerges from the observation that activities that are highly space-intensive, such as most types of agricultural production, are quite resistant to bureaucratization, even after the development of modern technologies of transportation and administration; the firms that succeed in surviving in most types of farming are normally too small to have bureaucracies (Olson, 1985).

By contrast, in activities in which the transfer of new technologies and other information is especially important, market failure is likely to be fairly extensive, mainly because new information would only be rationally purchased by those who did not already have this information, and from this it follows that the market for new information is particularly handicapped by the asymmetrical information of the parties to any transaction. Thus, as J.C. McManus (1972), Buckley and Casson (1976) and, especially, Hennart (1982) have shown, the emergence of the multinational firms with bureaucracies that transcend national borders can be explained in this framework; capital can cross national borders through portfolio investment (almost all British and other foreign investment in the 19th century was portfolio investment), but the rise in the relative importance of firms with new technologies and methods that were often not well suited to market transfer via licensing of patents, gave rise to the multinational corporation.

The foregoing emphasis on the business bureaucracies that are generally neglected in discussions of bureaucracy makes possible a brief and unified explanation of governmental bureaucracy as well. Governmental bureaucracies are similarly necessary only because markets fail, at least to some degree; the theory of market failure is readily capable of being generalized to include all functions for which governmental are an efficient response (Olson, 1986). Since governmental as well as market mechanisms are obviously imperfect, it does not follow from the presence of market failure that government intervention is normatively appropriate, since the government might fail even worse than the market, but market failures are nonetheless often important and always a necessary condition for optimal governmental intervention. Of course, it would be absurd to suppose that actual government intervention is always optimal or that governments always intervene when it is Pareto-efficient for them to do so. It is nonetheless instructive to look at the existence of government bureaucracy, as of business bureaucracy, in terms of market failure.

Among other reasons, it is instructive because the very conditions that give rise to market failure inevitably generate, in governments, and to a considerable degree also in firms, exactly those inefficiencies and rigidities that are popularly and correctly attributed to bureaucracies. Some of these inefficiencies also occur when either governmental or business bureaucracy is used inappropriately, but

the problem is most easily evident, and most serious, in precisely those cases where market failure makes bureaucratic mechanisms indispensable.

The reasons why the same conditions that make markets fail also generate difficulties and inefficiencies in bureaucracies unfortunately do not lend themselves to brief exposition. But perhaps a faint and intuitive sense of the matter will be evident from a moment's reflection about what could make a bureaucracy necessary. If, say, the fruits or vegetables grown on a farm are best picked by hand and the best way to pay each worker is by the number of bushels picked, there is no need to have any bureaucratic mechanism for getting the work done. When piece-rate or commission systems of reward work well, the market gives each worker a more or less optimal incentive to work and to be as efficient as the worker knows how to be. In essence, the reason is that the output is highly divisible into more or less homogeneous units or the revenue attributable to each worker is known, and so the output of different workers can be measured with reasonable accuracy.

Let us now shift to an opposite extreme. Consider a typical civil servant in the foreign ministry of a government. Even supposing that the only purpose of the foreign ministry in question was peacefully to maintain the country's independence, there would still be a stupendous difficulty in rewarding the civil servant on a piece-rate or commission basis, or in any way that is proportional to his productivity. The security of the country in question would normally depend in large part on what might loosely be described as the state of the international system – on world-wide indivisible or public good for which no one country could be entirely responsible. But even if the country in question were the only producer of this indivisible good, the foreign ministry would not be the only part of the government or the country that was relevant. Even in the foreign ministry, the typical civil servant is only one among thousands. How is his individual output to be measured, or even distinguished from that of his co-workers? The civil servant obviously cannot be paid in proportion to the revenue he generates, because if there really is market failure, the output cannot be sold in a market in the first place. Thus in practice, the remuneration of civil servants involved in producing public goods is not even a close approximation to each civil servant's true output; rewards in civil services will depend dramatically on proxy variables for performance such as seniority, education, and the fidelity of the employee to the interests of his superior and to the ‘culture’ or ideology of that bureaucracy. The peculiarities of civil service personnel systems, competitive bidding rules, and red tape are mainly explained by this logic (Olson, 1973, 1974).

The knowledge of the ‘social production function’ of a government bureaucracy producing public goods will also be limited by the same indivisibility that has been described; there are fewer countries, or even airsheds for pollution abatement, than there are farms (or experimental plots at agricultural experiment stations), so in general less is known about how to run countries or control pollution than about agriculture or about production processes in other competitive industries (Olson, 1982). The same indivisibility that obscures the social production function and the productivity of individual civil servants and other public inputs also insures that there cannot be even an imperfectly competitive market, so there is also no direct information on what an alternative bureaucracy could have achieved in the same circumstances.

In large part, it is the lack of information due to the indivisibilities described above that allows some of the bureaucratic pathologies described in Niskanen (1971) and Tullock (1965) to occur. In Niskanen's widely cited formal model, it is assumed that only the government bureaucrats know how many resources are required to produce a given public output. These bureaucrats are assumed to gain from growth of the bureaucracy, because an official's power, opportunities for promotion and other perquisites

are assumed to be an increasing function of the budget the bureaucrat administers. An agency faces the constraint, however, that the electorate will not sustain any government programme whose total costs exceed the total value of its output. The optimization of government bureaucrats therefore leads to a bureaucracy far larger than is Pareto-efficient; in essence the bureaucracy takes all of the surplus under the society's demand curve for the government output at issue. Critics of Niskanen's model have pointed out that it neglects the subordination of bureaucrats to politicians, and that politicians whose opportunities for re-election are positively correlated with the government's performance will endeavour to prevent bureaucracies from taking all of the surplus (see, for example, Breton and Wintrobe, 1975).

These criticisms have substantial empirical support, but it is also true that there are many known cases where officials who fear a lower budget allocation than anticipated for their agency will eliminate or threaten to eliminate their politically most cherished activity rather than a marginal activity; this is precisely what Niskanen's model predicts. Though any final conclusion must await further research, the evidence available so far appears to suggest that the lack of information due to the indivisibilities described above does often allow bureaucracies to appropriate some of the surplus that consumers might otherwise be expected to receive, but that the incentives faced by politicians tends to keep bureaucracies from getting anything resembling the whole of this surplus.

Bureaucracies operating in a market environment share some of the information problems that confront government agencies providing public goods, but not others. The divisions of a large corporation that handle personnel, accounting, finance or public relations for the entire corporation provide collective goods to the corporation as a whole. They are in many ways in a situation analogous to the foreign ministry described above when deciding how much of the total profits of the firm to attribute to a given corporate employee; this accounts for the many similarities of large corporate and civil-service bureaucracies. But the corporation as a whole, and even the nationalized firm producing private goods in a market, does not, when it sells its output, have as great a difficulty as the government agency that produces a collective of public output that is indivisible and unmarketable. The firm produces a good or service that is divisible in that it may be provided to purchasers and denied to non-purchasers. This means that the output is directly measurable in some physical units or at least that the revenue obtained from this output is measurable. Since consumers, even in the absence of any high degree of competition, will have alternative uses for their money, the private corporation or nationalized firm in a market economy will get some feedback about how much value it is providing. If there is no legal barrier to the operation of a competitive enterprise and the market is contestable, the society will also have at least potential information about what value an alternative organization could provide. An enterprise in the market produces an output from which non-purchasers may be excluded, and this also means there is normally better knowledge of the production functions for private goods than of production functions for public goods. All this implies that the problems of bureaucracy are less severe in private business than in government agencies producing public goods. Interestingly, they are also less severe in government enterprises that unnecessarily produce private goods that private firms would readily provide than they are in agencies that produce public goods that would not have been provided by the market. The more flexible personnel policies in some nationalized firms than in classical civil service contexts thus provides support for the conception offered here.

The paradox of a vast growth of both public and private bureaucracy at the same time that there is almost a consensus that bureaucracies are not very efficient or flexible, thus appears to have a resolution. There are fundamental reasons, arising from the inherent conditions causing market failure that make

both public and private bureaucracies inevitable. These same reasons also explain why bureaucracies lack the information needed for high levels of efficiency. But these same market failures show that (though the existing degree of bureaucracy may of course be far from optimal), it should not be surprising that societies choose to use more private and public bureaucracy even as they condemn such bureaucracy.

See Also

- hierarchy
- public choice
- socialism
- Weber, Max

Bibliography

Breton, A. and Wintrobe, R. 1975. The equilibrium size of a budget maximizing bureau. *Journal of Political Economy* 83, 195–207.

Buckley, P. and Casson, M. 1976. *The Future of the Multinational Enterprise*. London: Macmillan.

Chandler, A.D. 1962. *Strategy and Structure: Chapters in the History of American Industrial Enterprise*. Cambridge, MA: MIT Press.

Chandler, A.D. 1977. *The Visible Hand: The Managerial Revolution in American Business*. Cambridge, MA: Harvard University Press.

Chandler, A.D. and Daems, H., eds. 1980. *Managerial Hierarchies: Comparative Perspectives on the Rise of Modern Industrial Enterprise*. Cambridge, MA: Harvard University Press.

Coase, R.H. 1937. The nature of the firm. *Economica* 4, N.S., 386–405.

Hennart, J.-F. 1982. *A Theory of the Multinational Enterprise*. Ann Arbor: University of Michigan Press.

McManus, J.C. 1972. The theory of the international firm. In *The Multinational Firm and the Nation State*, ed. G. Paquet, D. Mills. Ontario: Collier Macmillan.

Mueller, D.C. and Murrell, P. 1985. Interest groups and the political economy of government size. In *Public Expenditure and Government Growth*, ed. F. Forte and A. Peacock. Oxford: Basil Blackwell.

Niskanen, W.A. 1971. *Bureaucracy and Representative Government*. Chicago: Aldine-Antherton.

Olson, M.L. 1973. Evaluating performance in the public sector. In *The Measurement of Economic and Social Performance*, ed. M. Moss, Studies in Income and Wealth, vol. 38, National Bureau of Economic Research, New York: Columbia University Press.

Olson, M.L. 1974. The priority of public problems. In *The Corporate Society*, ed. R. Marris. London: Macmillan.

Olson, M.L. 1982. Environmental indivisibilities and information costs: fanaticism, agnosticism, and intellectual progress. *American Economic Review, Papers and Proceedings* 72, 262–6.

Olson, M.L. 1985. Space, agriculture, and organization. *American Journal of Agricultural Economics* 67, 928–37.

Olson, M.L. 1986. Toward a more general theory of governmental structure. *American Economic Review, Papers and Proceedings* 76, 120–5.

Tullock, G. 1965. *The Politics of Bureaucracy*. Washington, DC: Public Affairs Press.

Weber, M. 1946. Bureaucracy. In *From Max Weber: Essays in Sociology*, ed. H. Gerth and C.W. Mills. New York: Oxford University Press.

Williamson, O.E. 1964. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and Anti-trust Implications*. New York: The Free Press.

Williamson, O.E. 1985. *The Economic Institutions of Capitalism*. New York: The Free Press.

How to cite this article

Olson, Mancur. "bureaucracy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000237> doi:10.1057/9780230226203.0177

The New Palgrave Dictionary of Economics Online

Burns, Arthur Frank (1904 – 1987)

G. Moore

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Burns, A. F.; business cycle measurement; business cycles; Council of Economic Advisers (USA); Economic Policy Advisory Board (USA); Federal Reserve System; Mitchell, W. C.; National Bureau of Economic Research

Article

Burns was born in Stanislau, Austria, on 27 April 1904. In 1914 his family emigrated to the United States, settling in Bayonne, New Jersey. Burns became a member of the economics faculty at Rutgers University in 1927, leaving in 1941 to accept an appointment at Columbia University, where he taught for many years and became John Bates Clark Professor of Economics Emeritus. He joined the staff of the National Bureau of Economic Research in New York in 1930, was director of research, 1945–53, and president 1957–67. In Washington Burns served as chairman of the Council of Economic Advisers, 1953–6; Counsellor to the President, 1969–70; chairman of the Federal Reserve System, 1970–78; and member of the President's Economic Policy Advisory Board since 1981. From 1981 to 1985 he was US Ambassador to the Federal Republic of Germany. In 1978–80 and again after 1985 he was distinguished scholar in residence at the American Enterprise Institute.

Burns's economic studies have been primarily concerned with economic growth, business cycles, inflation, and economic policies bearing upon these phenomena. In *Production Trends in the United States since 1870*, published in 1934, he examined growth rates in individual industries, noting the nearly universal tendency towards retardation. An initial stage of rapid growth in a new industry is usually followed by slower growth as it loses part of its market or its resources to still newer industries. Despite the tendency towards slower growth and eventual decline of most industries, Burns noted that this did not imply that growth in total output would slow. The underlying cause, that is the rise of new industries, would itself help to maintain rapid growth in total output.

Burns's collaboration with Wesley Mitchell in the study of business cycles led to many innovations in measurement technique and to a vast accumulation of knowledge about the characteristics of cycles and the economic interactions that generated them. It also led to a more realistic view of what business cycle

theory had to explain and what economic policy could be expected to accomplish. This in turn was useful to Burns in his later role as an economic policymaker, that is as a presidential adviser and as chairman of the Federal Reserve. Before taking on these responsibilities he wrote prophetically (1953): ‘It is reasonable to expect that contracyclical policy will moderate the amplitude and abbreviate the duration of business contractions in the future … But there are no adequate grounds, as yet, for believing that business cycles will soon disappear, or that the government will resist inflation with as much tenacity as depression …’ Burns's subsequent efforts were largely directed to improving the anti-recession, anti-inflation, and growth promoting policies of government.

Selected works

1930. *Stock Market Cycle Research*. New York: Twentieth Century Fund.

1934. *Production Trends in the United States since 1870*. New York: National Bureau of Economic Research.

1938. (With W.C. Mitchell.) *Statistical Indicators of Cyclical Revivals*. New York: National Bureau of Economic Research, Bulletin 69.

1946. (With W.C. Mitchell.) *Measuring Business Cycles*. New York: Columbia University Press.

1946. *Economic Research and the Keynesian Thinking of Our Times*. New York: National Bureau of Economic Research, 26th Annual Report.

1950. *New Facts on Business Cycles*. New York: National Bureau of Economic Research, 30th Annual Report. Reprinted in Burns (1969).

1952. *The Instability of Consumer Spending*. New York: National Bureau of Economic Research, 32nd Annual Report.

1952. (et al.) *Wesley Clair Mitchell: The Economic Scientist*. New York: Columbia University Press.

1953. *Business Cycle Research and the Needs of Our Times*. New York: National Bureau of Economic Research, 33rd Annual Report.

1954. *The Frontiers of Economic Knowledge*. Princeton: Princeton University Press.

1957. *Prosperity without Inflation*. New York: Fordham University Press.

1960. Progress towards economic stability. *American Economic Review* 50, 1–19.

1966. *The Management of Prosperity*. New York: Columbia University Press.

1967. (With P.A. Samuelson.) *Full Employment, Guideposts, and Economic Stability*. Washington, DC: American Enterprise Institute.

1968. (With J.K. Javits C.J. Hitch.) *The Defense Sector and the American Economy*. New York: New York University Press.

1968. Business cycles. *International Encyclopaedia of the Social Sciences*. New York: Macmillan and Free Press. Reprinted in Burns (1969).

1969. *The Business Cycle in a Changing World*. New York: Columbia University Press.

1978. *Reflections of an Economic Policy Maker*. Washington: American Enterprise Institute.

1984. The American trade deficit in perspective. *Foreign Affairs* 62(5), 1058–69.

1985. Interview: an economist's perspective over 60 years. *Challenge*, Jan.–Feb., 17–25.

How to cite this article

Moore, G. "Burns, Arthur Frank (1904–1987)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000240> doi:10.1057/9780230226203.0178

The New Palgrave Dictionary of Economics Online

Burns, Arthur Robert (1895 – 1981)

Henry W. Spiegel
From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Burns, Arthur R.; Burns, Eveline M.; competition theory

Article

A native of London, Burns was educated at the London School of Economics, where he was a pupil of Edwin Cannan. His doctoral dissertation, *Money and Monetary Policy in Early Times*, appeared in a prestigious series in 1927 and is still a standard work on the subject. After the completion of his studies Burns moved to the United States and taught at Columbia University from 1928 to 1963. His service there overlapped with that of his wife Eveline M. Burns, with whom he published an introductory economics text in 1928, and with that of Arthur F. Burns, another noted economist.

In 1936 Burns, still an assistant professor, published *The Decline of Competition*, the bulk of which consisted of chapters on trade associations, price leadership, market sharing, price stabilization, price discrimination, non-price competition and integration. The work formed part of a discussion that had been set in motion by the writings of Sraffa, Joan Robinson and Chamberlin and which explored the no man's-land between competition and monopoly. It was to serve as a bridge that linked the abstractions of the theories of imperfect or monopolistic competition with the world of reality. Standing between abstraction and description, Burns' work was in the main an attempt at classification. It holds middle ground between the soaring abstractions of pure theory and the industry studies published by Walton Hamilton and Associates under the title *Price and Price Policies* in 1938. Hamilton was a follower of Veblen. Burns shared a friendly disposition toward institutional economics with other Columbia economists.

The Decline of Competition constitutes Burns's main claim to fame. In later years he directed a Twentieth Century Fund study of electric power and government policy, and in 1955 he published *Comparative Economic Organization*. The former work was overtaken by the rise of atomic power as a source of electric energy, and the latter compared in isolation various factors affecting the national income.

Selected works

1927. *Money and Monetary Policy in Early Times*. London: Kegan Paul; New York: Knopf.

1928. (With Eveline M. Burns.) *Economic World: A Survey*. London: Oxford University Press.

1936. *The Decline of Competition*. New York: McGraw-Hill.

1948. (Director of Research.) *Electric Power and Government Policy*. New York: Twentieth Century Fund.

1955. *Comparative Economic Organization*. New York: Prentice-Hall.

Bibliography

Hamilton, Walton H. and Associates. 1938. *Price and Price Policies*. New York: McGraw Hill.

How to cite this article

Spiegel, Henry W. "Burns, Arthur Robert (1895–1981) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000111> doi:10.1057/9780230226203.1879

The New Palgrave Dictionary of Economics Online

business cycle measurement

Don Harding and Adrian Pagan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We describe different ways of measuring the business cycle. Institutions such as the NBER, OECD and IMF do this by locating the turning points in series taken to represent the aggregate level of economic activity. The turning points are determined according to rules that either come from a parametric model or are nonparametric. Once located, information can be extracted on cycle characteristics. We also distinguish between cases where single and multiple series are used to represent the level of activity.

Keywords

Burns, A.; business cycle; business cycle measurement; censoring operations; coincident indices; crossing points; data filters; fluctuations vs cycles; growth cycles; Markov switching (MS) processes; Mitchell, W.; periodic cycles; random variables; reference cycle; spectral analysis; turning points

Article

Measurement of business cycles provides a reference point against which macroeconomic theories and policy discussion can be assessed. The process requires an operational definition of a cycle, criteria to distinguish business cycles from other forms of fluctuation, procedures to detect the presence of a business cycle, and methods to measure its features. A central theme of this entry is that good measurement should not prejudge the nature of the phenomena under investigation. Moreover, it should produce statistics which are informative about features of interest and which can be formally analysed.

Defining and detecting cycles

In their classic work *Measuring Business Cycles*, Burns and Mitchell (BM) (1946) define *specific cycles* in a series y_t in terms of *turning points* in its sample path. This tradition has been central to work at the NBER and other institutions such as the IMF (2002) and the OECD (leading indicators). When it came to discussing the business cycle, BM simply referred to y_t as the *level* of aggregate economic activity, although in this article we will regard it as the *log* of economic activity, as the turning points in the level and the log of economic activity are the same. When Mintz (1969; 1972) had trouble finding turning points in the level of activity in surging economies such as West Germany's, this led her to first extract a permanent component p_t from y_t and to then study turning points in $z_t = y_t - p_t$. The resulting *growth cycle* in z_t has many forms depending on the method used to extract the permanent component. Others, such as the Economic Cycle Research Institute (ECRI) (growth rate cycle), have studied turning points in the differenced data Δy_t. A generalization of this, explored by Kedem (1980; 1994) and Harding (2003), is to study turning points in $\Delta^r y_t$.
At the time Mitchell began his work, the alternative way of thinking about cycles (or oscillations) was to view y_t as composed of periodic components represented by sine and cosine waves, that is

$$y_t = \sum_{j=1}^{m} \alpha_j \cos \lambda_j t + \beta_j \sin \lambda_j t, \quad (1)$$

where λ_j is the frequency of the *j'th* oscillation. If $m = 1$ there would be a single periodic cycle. The problem with this way of looking at cycles was that few economic time series showed evidence of periodicity. To overcome that problem α_j and β_j were allowed to vary stochastically over time. Specifically, they were treated as uncorrelated random variables with zero mean and variance σ_j^2. This formulation meant that y_t had to be a stationary random variable and so could not be applied to the levels of variables such as GDP (unlike turning point analysis). However, in this form one can measure the importance of the jth periodic cycle by looking at the ratio of σ_j^2 to the variance of y_t and it is the basis of spectral analysis. Such a perspective has increasingly been referred to as studying fluctuations rather than cycles, since the focus of attention is upon the variance of y_t.
To understand the difference between these alternative ways of measuring cycles, take the special case where $\lambda_1 = 0$ and there is another frequency λ_2. Then

$$y_t = \alpha_2 \cos \lambda_2 t + \beta_2 \sin \lambda_2 t + \alpha_{1t}, = y_t' + \alpha_{1t}. \quad (2)$$

Now there are certainly turning points in the series y_t' and the period between them is determined by λ_2. In contrast, the turning points in y_t will also be affected by the random variable α_{1t}, and thus may be very different to those in y_t'. Information about cycles gathered from spectral analysis concerns the nature of turning points in y_t' and not y_t. To give a more concrete illustration of this point, suppose that the model for y_t is of the form

$$y_t = 1.4 y_{t-1} - .53 y_{t-2} + e_t.$$

Then the periodic cycle in y_t can be isolated by setting $e_t = 0$ to get y_t'. To use the dating methods of an institution like NBER, the turning points in y_t' are 22 quarters apart, as could also be discovered by computing the roots of $(1 - 1.4L + .53L^2) = 0$. However, applying the same methods to y_t, one finds that the turning points in y_t will be on average 12 quarters apart. A further disadvantage of the periodic cycle approach is that the data needs to be filtered to render it stationary before analysis proceeds and, as Cogley observes elsewhere in this dictionary (data filters), the filters most commonly used by macroeconomists can introduce spurious periodic cycles, thereby blurring the picture.

Locating turning points

To locate turning points in a series it is necessary to define what these are and to provide some way of recognizing them in a given data-set. An obvious solution is to use the idea that peaks (troughs) are local maxima (minima) in the series y_t. Hence, if $\vee_t (\wedge_t)$ are binary variables taking the value of unity where there is a peak (trough) at *t* and zero otherwise, applying the proposed definition gives

$$\vee_t = \mathbf{1}(y_t < y_{t \pm j},\ 1 \le j \le k)$$

(3)

$$\wedge_t = \mathbf{1}(y_t > y_{tj\pm},\ 1 \le j \le k).$$
(4)

In eqs. (3) and (4) 1(*A*) is the indicator function taking the value 1 if the event *A* is true and zero otherwise. Of course, this still leaves one with the need to describe the interval over which the local maxima or minima are said to occur, that is, a choice needs to be made regarding *k*. To replicate the main features of Burns and Mitchell's specific cycle dating procedures, it is necessary to set $k = 5$ for monthly data or $k = 2$ for quarterly data.

This is not the last of the choices that need to be made when locating turning points, but the others do not relate to the location of local maxima and minima. Rather, they concern the question of whether one should eliminate some of the local turns in deciding on a final set of turning points. Mostly these extra restrictions are imposed as phase length constraints, where phases are the periods of expansions and contractions between turning points. Thus, NBER dating procedures require that completed phase and complete cycles durations last longer than 5 and 15 months respectively. These are generally referred to as censoring operations. Whether turning points should be censored depends on the objectives of the research. If the objective is to match NBER business cycle dates, then censoring is essential. But if the researcher is pursuing other objectives such censoring may not be necessary. Censoring turning points makes it much harder to formally analyse the statistics produced and this may provide an important reason for not imposing them.

BM acknowledged that the final set of dates they selected for turning points reflected considerable amounts of judgement and incorporated specific information about economic activity at particular dates. Today, academic economists are primarily interested in the average characteristics of the cycle, and so it may well be that automated methods of turning point detection become attractive. In the early post-Second World War period many of the procedures used by BM were codified, producing an expert system for locating turning points. Ultimately, Bry and Boschan (1971) produced an algorithm and FORTRAN program (called BB here) that largely replicated this expert system. Subsequently Mark Watson (1994) implemented this algorithm in the language GAUSS, and that code is available at (http://www.princeton.edu~mwatson).

There were three key components to the BB algorithm. The first was to engage in some smoothing of the series and to find an initial set of turning points using eqs. (3) and (4) with $k = 5$. The second was to eliminate enough of these turning points so as to ensure that expansion and contraction phases exceeded 5 months in duration, while completed cycles exceed 15 months in duration. The third component was to ensure that peaks and troughs alternated by deleting multiple sequential occurrences of these. That was done through the application of various rules, such as choosing between two peaks based on which had the higher value of y_t.

Although BB were interested in analysing monthly data, they suggested a method for working with quarterly data that involved treating the observations on each of the months in a quarter as one-third of the quarterly value. A variant of BB has been developed by Harding and Pagan (2002) and called BBQ. It omits the smoothing in the BB algorithm but retains the three key principles of the BB algorithm. It also sets $k = 2$ and makes the minimum phase and cycle lengths two and five quarters respectively. Faster recursive algorithms for locating turning points have been developed by Artis, Marcellino and Proietti (2004) and James Engel. Engel's computer programs are called MBBQ. They are written in MATLAB and GAUSS and are available at the National Centre for Econometric Research (MBBQ Code).

Model-based procedures for defining and locating turning points

The procedures above do not require any knowledge of the data-generating process for y_t. An alternative approach is to adopt a model of Δy_t and use this to locate turning points. To date the models used are *parametric* and generally feature two regimes. Perhaps the best known parametric model is that of Hamilton (1989), where the growth rate is treated as a Markov switching (MS) process of the form $\Delta y_t = \mu_0(1 - \xi_t) + \mu_1\xi_t + e_t$. Here μ_j are the growth rates in the two regimes, and these are indexed by a latent binary state, ξ_t, while e_t is a normally distributed zero mean error term. Here μ_0 is the growth rate of the low growth state and μ_1 is the high growth rate. Sometimes the restriction $\mu_0 < 0$ is also imposed. The model is completed by specifying the transition probabilities of moving from $\xi_{t-1} = 0$ or 1 to $\xi_t = 1$ or 0. The model can be made more complex with extra dynamics, different variances in each regime, allowing the transition probabilities to depend on some observable data, and so on. This parametric model is used to compute the conditional

probability, $\Pr[\xi_t = 1|A_t]$, where A_t is either all or a subset of the growth rates $\{\Delta y_j\}_{j=1}^{T}$. Thus the estimate of $\Pr[\xi_t = 1|A_t]$ is a function of whatever growth rates are in A_t. Generally this probability will be a nonlinear function of the elements in A_t although a linear function can be quite a good approximation – see Harding and Pagan (2003) for an example.

The cycle is then associated with a binary variable S_t that takes the value 1 in expansion and zero in contraction. A rule is used to construct S_t by comparing the estimated probability of being in the high growth state with some critical value. Hamilton chose .5, and most of those using the technique have followed suit. Consequently, if $\Pr[\xi_t = 1|A_t] > .5$, an expansion is signified and S_t is set to unity. If the criterion is not satisfied S_t is set to zero. Notice that the ξ_t are not the phase states; the latter are S_t. They are simply a device for producing some nonlinear structure in Δy_t, although often one can think of the outcomes for ξ_t as signifying a low or high growth period. The correlation between S_t and ξ_t may be very low. Many applications of this methodology have now been made and the MS model that one chooses seems to vary a lot with the series it is being applied to. The simple one described above rarely works satisfactorily.

In most instances a decision about the utility of the method is made by comparing the business cycle states produced by the rule based on the magnitude of $\Pr[z_t = 1|A_t] > .5$ with those found by turning point methods. Because of the latter comparison one has to ask what the advantages there are in using a model to locate turning points. Chauvet and Piger (2003) claim that an advantage of the model-based approach is that it allows an investigator to forecast turning points in real time. There is some truth to this but it is exaggerated. Since forecasts can be found for any such model, they could be passed through any chosen dating algorithm to determine the predicted phases.

Measuring cycle features

Turning points segment time series into phases. An expansion phase runs from the trough to the next peak. A contraction runs from a peak to the next trough. In what follows it is easiest to just describe the derivation of information on expansions.

The two most basic statistics related to phases are duration and amplitude. The *duration* of an expansion is the number of periods of time between the trough and next peak. The *amplitude* of an expansion measures the change in y_t from trough to the next peak. In many cases y_t is the log of some variable such as GDP or industrial production, that is, $y_t = \ln(Y_t)$, and the amplitude has a natural interpretation as the approximate percentage change in Y_t between trough and peak.

Duration and amplitude form two sides of a triangle. Connecting the trough and peak produces the hypotenuse. If $y_t = \ln(Y_t)$, then the hypotenuse represents the path followed by a variable that exhibits a constant growth rate during an expansion. With this in mind it is instructive to inspect the actual path followed by the data, and to compare that path with the constant growth path represented by the hypotenuse. Figure 1 shows how US expansion paths have deviated from the constant growth rate path in the post-Second World War period. The important feature evident in this figure is that the growth rate of GDP is not constant over the expansion phase and typically is highest in the first half of an expansion.

Figure 1

Deviation of sample path from hypotenuse: US GDP during expansions in the post-Second World War period. *Source*: Harding (2003).

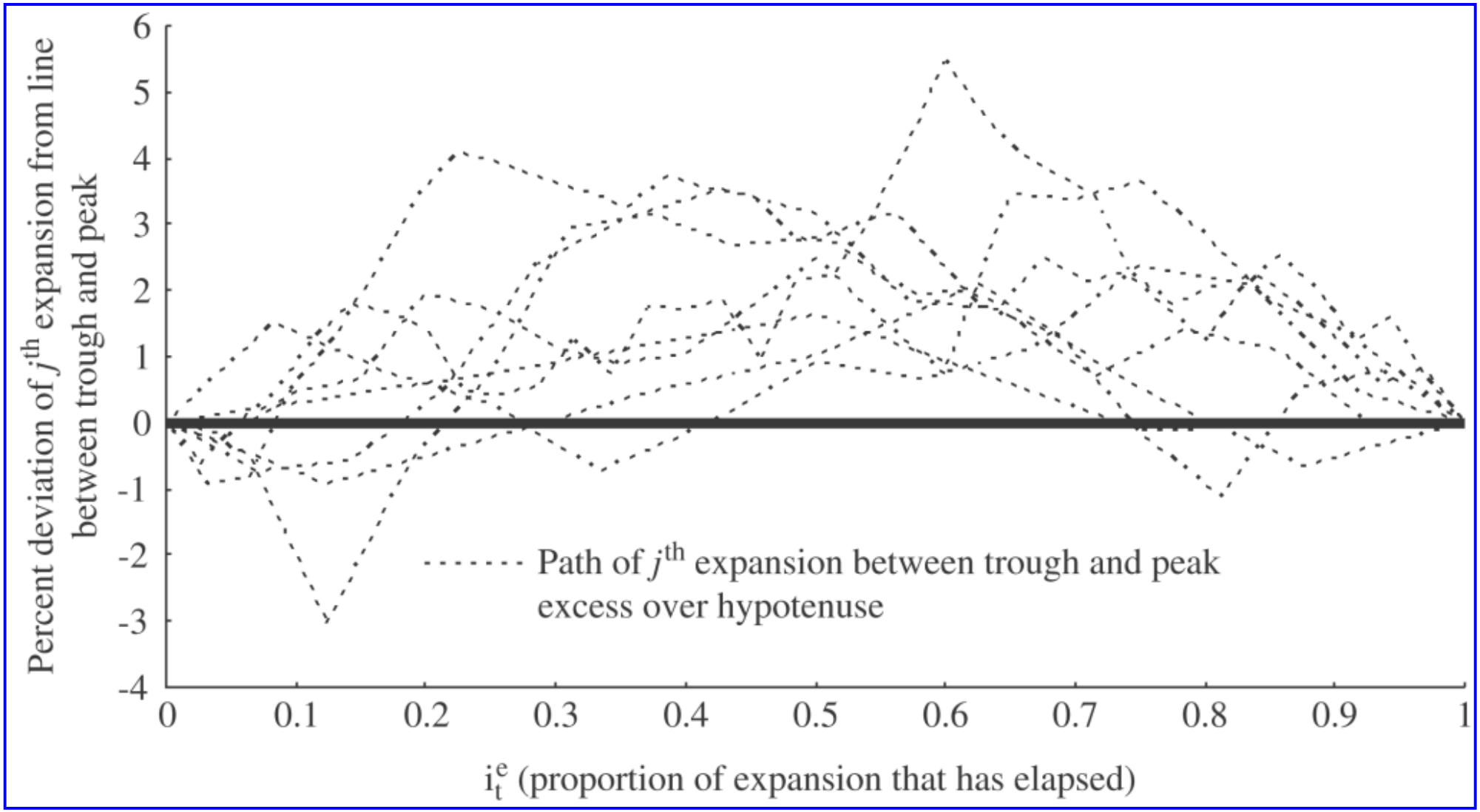

While comparisons such as that in Figure 1 are visually informative, there is also a need for statistics that summarize the average shape of phases. Sichel (1994) divides expansions into three stages, computes the average growth rate for each stage, and shows graphs of these, as well as providing formal statistical tests of equality of the growth rates in each stage. Harding and Pagan (2002) compare the cumulated gain in an expansion with what it would have been if growth had been constant throughout the phase. This comparison was motivated by the idea mentioned above, that a plot of y_t against t during an expansion would look like a triangle if growth had been constant. The area of such a triangle would be one-half the product of the amplitude and duration. If growth was not constant the area under the path actually followed by activity during the expansion would differ from the triangle. Thus, a comparison of the two areas provides a measure of the extent of departure from a constant growth scenario. The evidence seems to be that expansions do not feature constant growth in some countries like Australia, the United States and the UK, but do so in many Continental European countries. The shape analysis is interesting since a linear process for Δy_t will produce phases that, on average, have constant growth rates. So a failure to see this signals the need for a nonlinear process for Δy_t. The shape analysis also provides a useful tool for testing whether nonlinear models produce realistic business cycles.

All of the methods for summarizing business cycle information can be applied to growth cycles and to data that have undergone higher-order differencing. In addition, Sichel (1993) suggested tests for ‘deepness’ and ‘steepness’ in the growth cycle that were effectively tests for symmetry in the densities of z_t and Δz_t.

Using multivariate information in defining and detecting business cycles

Burns and Mitchell's famous definition of a business cycle – ‘Business cycles are a type of fluctuation found in the aggregate economic activity of nations…a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general … contractions…’ (1946, p. 1) – has two aspects. One points to the need to identify aggregate economic activity, and the other to the fact that there should be synchronization across many series during the phases of a business cycle. Burns and Mitchell commented that GDP was a suitable index of economic activity, although others, such as Moore and Zarnovitz (1986), have preferred a weighted average of several series rather than a single one. However, since data on GDP was not available to Burns and Mitchell, for either the time period or the frequency in which they were interested, it is natural that they placed more emphasis upon the second component of their definition when discussing the business cycle.

This second component emphasizes synchronization of the cycles in the specific series taken to represent economic activity. Burns and Mitchell took the turning points in many series and then extracted a *reference cycle* by determining those dates which peaks and troughs ‘clustered around’. So a primary task is to be able to measure the tightness of the clusters. At the end of the process one also wishes to know how synchronized each of the specific cycles is with the cycle in the aggregate.

Harding and Pagan (2006) develop procedures to measure the tightness of clusters of turning points and the degree of synchronization of cycles through concordance indices that measure the fraction of time spent in the same phase. They apply those procedures to the series referred to by the NBER when dating the business cycle, and find that the turning points in those series are tightly clustered together. Harding (2003) finds that between March 1949 and September 2001 there is a concordance of 0.96 between the NBER business cycle states and the cycle obtained by locating turning points in US GDP.

Automated construction of the reference cycle

To automate the calculation of the reference cycle requires some rules which will distill the specific cycle turning points into a single set of turning points. To determine what these rules might be, one could look at the NBER Business Cycle Dating Committee procedure. It has a similar modus operandi to that of Burns and Mitchell, as seen in its discussion about dating the 2001 recession (NBER, 2003). However, one rarely gets a precise description either of how its decisions are made or of the series used in that process. In addition, it seems as if the series which have been most influential in decisions may have been different at different periods in time. The clearest description of the procedures for aggregating turning points in a set of series to create a reference cycle is in Boehm and Moore (1984), who explain how NBER methods were used when establishing a reference cycle for Australia. Their description can be taken as authoritative because Moore was a pivotal figure in the NBER Business Cycle Dating Committee for many years. Moore and Zarnowitz (1986) also provide information on methods used by NBER in dating the business cycle.

Given that the process for establishing the reference cycle is a little vague, it should not be surprising that there have been few attempts at producing automated dating algorithms to establish it from multivariate series. Harding and Pagan (2006) construct an algorithm to replicate the NBER procedures described by Boehm and Moore (1984). They obtain the ‘clustering parameter’ which is essential to measuring the tightness of turning point clusters by looking at Boehm and Moore's spreadsheets. The resulting algorithm has produced a reference cycle that matches the Australian version established by Boehm and Moore quite well. Subsequently, it has been tested on US data, and is able to produce quite a good replication of the reference cycle for the United States, even though the clustering parameter had been calibrated with Australian data.

Model-based procedures for defining detecting and extracting a reference cycle

Recently, academic economists have used parametric models to construct a coincident index and the reference cycle from n multivariate series $\Delta y_{1t}, \ldots, \Delta y_{nt}$. A common element to all approaches is to write Δy_{jt} as a function of a common component Δf_t and idiosyncratic components $u_{jt} (j = 1, \ldots, n)$. Hence a simple representation would be $\Delta y_{jt} = a_j \Delta f_t + u_{jt}$. The f_t is often thought of as the coincident index of the business cycle. Of course, there may be more than one f_t but, ultimately, we can think of combining them to form a single variable. There are then many ways that models for Δf_t and u_{jt} might be specified, depending upon how strong the assumptions are that one wishes to make about the nature of f_t and u_{jt}. Often Δf_t is given an MS form (for example, Chauvet and Piger, 2003). Depending on what these assumptions are, they will determine how an estimate of f_t is to be made. Stock and Watson (1991) and Chauvet (1998) represent different approaches. In some instances one can avoid specifying precise parametric models for f_t and u_{jt}, restricting them only to be in a general class. Forni et al. (2001)'s dynamic factor approach is the main representative of this latter technique. The main issue with these approaches is that the coincident index and reference cycle obtained are conditioned on the assumptions made about the data-generating process. For that reason these approaches cannot provide a neutral measurement of the reference cycle.

Conclusion

Although widely used in official circles, Burns and Mitchell's methods of measuring cycles through turning points have been less popular in academia. But this has changed in recent years. There are a number of reasons why the methods have become increasingly attractive. First, information about the nature of the cycle phases can be generated, and this shape information proves important when one tries to construct models of economic activity. Second, the literature now contains expert systems for locating turning points, and these have been coded into various computer languages, thereby eliminating the judgmental aspect of the method. Nevertheless, the automatically generated turning points have been quite good approximations to those found via judgment. Third, the ability to produce simulated data from parametric models means that such information can be passed through the algorithms for locating turning points to produce simulated distributions for the statistics that summarize the features of the cycle. Fourth, the emerging mathematics literature on crossing points provides a natural foundation on which to build a distribution theory for Burns and Mitchell's methods. Fifth, there is now a large literature on parametric methods for locating turning points and measuring cycles. This latter literature can readily be linked to the nonparametric turning point approach of investigators such as Burns and Mitchell, as seen in Harding and Pagan (2003).

See Also

- Burns, Arthur Frank
- data filters
- international real business cycles
- Mitchell, Wesley Clair

Bibliography

Artis, M., Marcellino, M. and Proietti, T. 2004. Dating the Euro area business cycle: a methodological contribution with an application to the Euro area. *Oxford Bulletin of Economics and Statistics* 66, 537–65.

Boehm, E. and Moore, G. 1984. New economic indicators for Australia, 1949–84. *Australian Economic Review* 4th quarter, 34–56.

Bry, G. and Boschan, C. 1971. Cyclical analysis of time series: selected procedures and computer programs. Technical Paper No. 20. New York: NBER.

Burns, A. and Mitchell, W. 1946. *Measuring Business Cycles*. New York: National Bureau of Economic Research.

Chauvet, M. 1998. An econometric characterization of business cycle dynamics with factor structure and regime switches. *International Economic Review* 39, 969–96.

Chauvet, M. and Piger, J. 2003. Identifying business cycle turning points in real time. *Federal Reserve Bank of St. Louis Review* 85(2), 47–61.

ECRI (Economic Cycle Research Institute). Growth rate cycle. Online. Available at http://www.businesscycle.com, accessed 8 September 2006.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. 2001. Coincident and leading indicators for the EURO area. *Economic Journal* 101, 62–85.

Hamilton, J. 1989. A new approach to the economic analysis of non stationary time series and the business cycle, *Econometrica* 57, 357–84.

Harding, D. 2003. Essays on the business cycle. Ph.D. thesis, Yale University.

Harding, D. and Pagan, A. 2002. Dissecting the cycle: a methodological investigation. *Journal of Monetary Economics* 49, 365–81.

Harding, D. and Pagan, A. 2003. A comparison of two business cycle dating methods. *Journal of Economic Dynamics and Control* 27, 1681–90.

Harding, D. and Pagan, A. 2006. Synchronization of cycles. *Journal of Econometrics* 132, 59–79.

IMF (International Monetary Fund). 2002. Recessions and recoveries. *IMF World Economic Outlook*, April. Washington, DC: IMF.

Kedem, B. 1980. *Binary Time Series*. New York: M. Dekker.

Kedem, B. 1994. *Time Series Analysis by Higher Order Crossings*. Piscataway, NJ: IEEE Press.

Mintz, I. 1969. Dating postwar business cycles: methods and their application to Western Germany, 1950–1967. Occasional Paper No. 107. New York: NBER.

Mintz, I. 1972. Dating American growth cycles. In *The Business Cycle Today*, ed. V. Zarnowitz. New York: NBER.

Moore, G. and Zanrowitz, V. 1986. The development and role of the National Bureau of Economic Research's business cycle chronologies. In *The American Business Cycle: Continuity and Change*, ed. R. Gordon. Chicago: University of Chicago Press.

National Centre for Econometric Research. MBBQ Code. Online. Available at http://www.ncer.edu.au/data/, accessed 8 September 2006.

NBER (National Bureau of Economic Research). 2003. The NBER's recession dating procedure. Online. Available at http://www.nber.org/cycles/recessions.html, accessed 25 August 2006.

OECD (Organisation for Economic Co-operation and Development). Leading indicators. Online. Available at http://www.oecd.org/std/cli, accessed 25 August 2006.

Sichel, D. 1994. Inventories and the three phases of the business cycle. *Journal of Business and Economic Statistics* 12, 269–77.

Sichel, D. 1993. Business cycle asymmetry: a deeper look. *Economic Inquiry* 31, 224–36.

Stock, J. and Watson, M. 1991. A probability model of the coincident economic indicators. In *Leading Economic Indicators: New Approaches and Forecasting Records*, ed. K. Lahiri and G. Moore. Cambridge: Cambridge University Press.

Watson, M. 1994. Business cycle durations and postwar stabilization of the U.S. economy. *American Economic Review* 84, 24–46.

Watson, M.. Gauss code for Bry Boschan Algorithm. Online. Available at http://www.wws.princeton.edu/mwatson, accessed 8 September 2006.

How to cite this article

Harding, Don and Adrian Pagan. "business cycle measurement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000333> doi:10.1057/9780230226203.0179

The New Palgrave Dictionary of Economics Online

business networks

Francis Bloch
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Informal and formal business networks play an increasing role in economic activities. Business networks have been studied both by sociologists and economists in order to answer three questions: What is the influence of business networks on economic activities? What are the determinants of business networks? When and how are business networks alternatives to organized markets?

Keywords

business networks; class; commitment; corporate governance; credit markets in developing countries; diffusion of innovations; economic geography; enforceable international contracts; interlocking directorates; intermediate goods; international trade; knowledge spillovers; labour market search; market sharing; matching; spatial concentration; strategic alliances

Article

Informal and formal business networks play an increasing role in economic activities. A large literature in economics and sociology has focused attention on these business networks. Three sets of questions have been raised: What is the influence of business networks on economic activities? What are the determinants of business networks? When and how are business networks alternatives to organized markets?

The importance of social networks has been stressed in three spheres of economic activities: the job market, where personal referrals play an essential role; international trade, where the existence of networks helps explain the volume of trade across borders; and urban economics, where business relations are an important determinant of the degree of local knowledge spillovers.

Empirical studies show that as many as half of the jobs are found through personal contacts. Granovetter's landmark study (1974) of the importance of networks in the managerial and professional job market in a Boston suburb stresses the difference between ‘strong’ and ‘weak’ ties. According to his

study, weak ties – distant acquaintances to individuals who belong to different communities – play a much stronger role than ‘strong ties’ – close relations to individuals belonging to the same group – in helping business executives finding or changing jobs. Recent economic models of job networks emphasize the dynamic effect of networks on unemployment and inequality (Calvo-Armengol and Jackson, 2004).

In international trade, informal co-ethnic networks formed by migrants – like the Chinese trading network – and formal business networks like the Japanese *keiretsu* have a significant impact on the volume of trade across borders. Rauch's survey (2001) summarizes the empirical evidence and outlines different theoretical explanations of the effect of business networks on international trade. The existence of personal links allows traders to match opportunities better as the network provides an informational link across agents from different countries. Networks also allow traders to solve the problems of enforcement of international contracts – agents who do not meet their obligations may be expelled from the network.

Informal networks also play a fundamental role in the diffusion of innovations and the emergence of new ideas in local areas. In her celebrated comparison of business models in the Silicon Valley and on Route 128, Saxenian (1994) argues that the success of the Silicon Valley is in large part due to the flexible, informal organization of business relations in California. Economic geographers have long noted that these informal networks generate important knowledge spillovers, which help explain the concentration of industrial activities over space and justify the emergence of industrial districts.

The architecture of business networks has been extensively studied in two areas where precise data can be obtained: interlocking directorates and strategic alliances. Empirical studies of interlocking directorates – the exchange of directors across company boards – first show that networks of intercorporate relations are highly asymmetric: a small number of firms occupy a central position on the network, concentrating a large number of interlocks. Second, intercorporate links tend to be local, and interlocking occurs among firms in the same geographical area. Third, the number of interlocks increases with the firm's size.

To explain this pattern of interrelations, two competing theories have been proposed, resulting in a lively controversy in the sociological literature, reviewed by Mizruchi (1996). Proponents of the social class theory argue that interlocking reflects the dominance of the upper class, and that relations among firms are mostly explained by individual friendships and the desire to maintain hegemony over the corporate world. The resource dependence theory explains the existence of interlocks by the firms’ desire to access resources detained by other firms. According to this theory, industrial companies exchange directors with financial institutions in order to obtain easier access to credit and with their suppliers in order to guarantee access to intermediate goods needed in production.

Strategic alliances are bilateral agreements among firms in the same industry. Agreements to launch joint R&D projects have received special attention in the literature. On the empirical side, a large database of bilateral research agreements has been developed by the MERIT center in Maastricht (Hagedoorn, 2002). These data show a large increase in the number of partnerships in the 1990s, and demonstrate that firms increasingly use flexible contractual arrangements rather than joint-equity subsidiaries to launch new research programmes. Research partnerships are very unevenly distributed across industrial sectors, with high-tech industries (in particular information technology and the pharmaceutical industry) accounting for a very large share of agreements.

Goyal and Joshi (2003) propose a theoretical model to explain the formation of these collaborative networks. Their analysis explains the high density of the networks by showing that, in the absence of linking costs, firms always have an incentive to form strategic alliances. In the presence of linking costs, stable networks become asymmetric, with a small number of isolated firms facing a large group of interrelated firms. When firms choose their research investments after the network is formed, inefficiencies arise as firms have a tendency to fragment their investments over too many links. Belleflamme and Bloch (2004) study a different type of strategic alliance: reciprocal market-sharing agreements whereby firms divide markets geographically. They show that stable networks are typically asymmetric and contain complete components of different sizes.

Trade networks can provide a viable alternative to organized, anonymous, markets. Buyers and sellers establish personal links, and conduct trade on a bilateral basis rather than through a centralized market. Historically, business networks have played a fundamental role in the development of trade. Greif's celebrated study (1993) of the Maghribi network, formed by Jews in the western Mediterranean in early medieval Europe, points out that business networks were able to solve commitment problems in the absence of institutions enforcing contracts. Still in the western Mediterranean, but in modern times, Kirman's detailed study (2001) of the fish market in Marseille also shows that a larger volume of trade is conducted on a bilateral basis, with buyers and sellers linked through durable relations.

Casella and Rauch (2002) and Kranton (1996) propose alternative theoretical models to investigate the difference between anonymous markets and personalized networks. In Casella and Rauch (2002), business networks enable traders to overcome informational trade barriers, and to learn about matching opportunities in international markets. They show that agents who continue to conduct trade through organized markets suffer from the presence of the business network. Kranton's model (1996) is built around the issue of enforcement of contracts: agents can either choose to trade on the market at the risk of being cheated but benefiting from a wide variety of goods, or to use a personal network. Kranton shows that there exists a strong interaction between the two modes of exchange: the more people use networks, the lower their incentives to use markets; the larger the fraction of the population which uses markets, the lower are their incentives to engage in personal transactions.

In summary, the importance of business networks in economic activities, which has long been recognized by sociologists, is attracting increasing attention from economists. New theoretical and empirical methods enable researchers to revisit business networks. In this relatively new field of study, a number of problems remain open. For example, the theoretical corporate governance literature is still silent on the issue of interlocking directorates. The interaction between formal insurance and credit markets and informal network arrangements in developing countries also awaits further study.

See Also

- corporate governance
- network formation
- social networks in labour markets

Bibliography

Belleflamme, P. and Bloch, F. 2004. Market sharing agreements and stable collusive networks. *International Economic Review* 45, 387–411.

Calvo-Armengol, A. and Jackson, M.O. 2004. The effects of social networks on employment and inequality. *American Economic Review* 94, 426–54.

Casella, A. and Rauch, J. 2002. Anonymous markets and group ties in international trade. *Journal of International Economics* 58, 19–47.

Goyal, S. and Joshi, S. 2003. Networks of collaboration in oligopoly. *Games and Economic Behavior* 43, 57–85.

Granovetter, M.S. 1974. *Getting a Job*. Cambridge, MA: Harvard University Press.

Greif, A. 1993. Contract enforceability and economic institutions in early trade: the Maghribi trader's coalition. *American Economic Review* 83, 525–48.

Hagedoorn, J. 2002. Inter-firm R&D partnerships: an overview of major trends and patterns since 1960. *Research Policy* 4, 477–92.

Kirman, A. 2001. Market organization and individual behavior: evidence from fish markets. In *Networks and Markets*, ed. A. Casella and J. Rauch. New York: Russell Sage.

Kranton, R.E. 1996. Reciprocal exchange: a self-sustaining system. *American Economic Review* 86, 830–51.

Mizruchi, M.S. 1996. What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates. *Annual Review of Sociology* 22, 271–98.

Rauch, J. 2001. Business and social networks in international trade. *Journal of Economic Literature* 39, 1177–203.

Saxenian, A. 1994. *Regional Networks: Industrial Adaptation in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.

How to cite this article

Bloch, Francis. "business networks." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000310> doi:10.1057/9780230226203.0180

The New Palgrave Dictionary of Economics Online

Butlin, Noel George (1921 – 1991)

Graeme Donald Snooks

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Butlin, N. G.; economic development; national accounting; neoclassical growth theory; structural disequilibrium; technical change

Article

Noel George Butlin, one of Australia's leading historical economists, was born in Singleton, New South Wales on 19 December 1921. He was the sixth child and third son of Thomas Lyon Butlin, a railway porter, and Sara Mary Butlin (née Chantler). Butlin attended Maitland Boys High and studied economics at Sydney University. During his undergraduate years, Sydney had the nation's best economics department in terms of the professional qualifications of its teaching staff. Even so, Butlin claimed that, while his lecturers taught him how to deconstruct aspects of the economy, they were unable to show him how it all worked. He wanted to become a scholar to understand real-world economic processes.

Like many others of his generation, Butlin's career was disrupted by war. While he wanted to enter academia, the only avenue available on graduation was the Australian public service. Between 1942 and 1945 Butlin was mainly seconded to posts in the UK and USA. There he met J.M. Keynes, L. Robbins, A. Robertson, R. Stone, and H.J. Habakkuk. Back in Australia he participated in 1945 in making plans for Australia's post-war reconstruction, and in 1946 finally took up a lectureship at Sydney University. To further his research ambitions, Butlin accepted a Rockefeller Fellowship in 1949 to study for a Ph.D. at Harvard under Joseph Schumpeter. Unfortunately, the great man died a few months after Butlin's arrival, and he found himself in Harvard's Centre for Entrepreneurial Studies. He had little sympathy with their growing sociological interests and, after initial research on Canadian railways, decided in 1951 to return to Australia to work at the Australian National University (ANU). In 1963 he became Professor and Head of the Department of Economic History. Butlin's 40-year association with the ANU ended only with his death on 2 April 1991.

Back in Australia, Butlin was swept up in the post-war concern with economic development. On the theoretical side, the old influence of Schumpeter was joined by the new influences of Harrod and Solow–Swan, and, on the measurement side, the great statistician Coghlan was joined by Kuznets. Butlin

absorbed ideas from them all. He borrowed the ‘structural disequilibrium’ concept from Schumpeter, but ignored technological change in favour of the investment focus of the neoclassical growth model. Economic development in Butlin's analysis proceeded via long investment booms that created structural disequilibria and required depressions to reattain structural balance. The outcome of these influences, together with much hard work during the 1950s, was the publication of his two-volume magnum opus on Australian development (Butlin, 1962; 1964). This set the pattern for subsequent analysis by historians and economists in the 1970s and 1980s, and was only challenged in the 1990s (Snooks, 1994). Despite being an active researcher until his death, Butlin never surpassed this early work. His most interesting subsequent research focused on pushing his GDP estimates back to 1788 (Butlin, 1986), and on analysing the Aboriginal economy (Butlin, 1983; 1994).

What was the nature and importance of Butlin's contribution to economics and history? First and foremost, Butlin focused our attention on the *process* of Australian economic development, and showed that it was endogenously generated. This was an essential counterpoint to the traditional view that development was exogenously driven. Second, he demonstrated that real-world growth processes could not be encompassed by the simple neoclassical growth models that were fashionable among orthodox economists at the time. Unfortunately he was unable to fulfil his intention of writing a ‘strictly analytical’ volume to complete the 1960s trilogy. He failed, therefore, to develop a general dynamic theory that could displace these totally unrealistic growth models. That was left to others (Snooks, 1998). Third, while his hybrid national accounting techniques have been criticized, they have weathered the storm reasonably well. More than most Australian national accountants, Butlin had an impressive understanding of the history that generated the data he employed. When used for long-run rather than year-to-year analysis, the differences in alternative estimates are not significant (Snooks, 2007). In any case, it is Butlin's overarching interpretation, his realist vision, and his important example of what can be done with the available data that constitute his enduring contribution.

See Also

- Australasia, economics in

Selected works

1962. *Australian Domestic Product, Investment and Foreign Borrowing, 1861–1938/39*. Cambridge: Cambridge University Press.

1964. *Investment in Australian Economic Development, 1861–1900*. Cambridge: Cambridge University Press.

1983. *Our Original Aggression: The Aboriginal Populations of Southeastern Australia, 1788–1850*. Sydney: Allen & Unwin.

1986. Contours of the Australian economy, 1788–1860. *Australian Economic History Review* 26, 96–

125.

1994. *Economics and the Dreamtime*. Cambridge: Cambridge University Press.

Bibliography

Snooks, G.D. 1994. *Portrait of the Family within the Total Economy: A Study in Longrun Dynamics, Australia, 1788–1990*. Cambridge: Cambridge University Press.

Snooks, G.D. 1998. *Longrun Dynamics: A General Economic and Political Theory*. London: Macmillan.

Snooks, G.D. 2007. Dynamics downunder: Australian economic strategy and performance from the palaeolithic to the twenty-first century. In *World Economic Performance: Past, Present and Future*, ed. P. Rao and B. van Ark. Princeton: Princeton University Press.

How to cite this article

Snooks, Graeme Donald. "Butlin, Noel George (1921–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000337>
doi:10.1057/9780230226203.0181

The New Palgrave Dictionary of Economics Online

Cairnes, John Elliott (1823 – 1875)

R.D. Collison Black

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cairnes, J. E.; deductive method; Fawcett, H.; Jevons, W. S.; land tenure; Malthus's theory of population; Mill, J. S.; rent control; slavery; subjective theory of value; verificationist methodology; wages fund

Article

Cairnes was born at Castlebellingham, County Louth, Ireland. At the height of his career he was probably the best-known political economist in England after John Stuart Mill, whose friend and associate he was from 1859 onwards; but his interest in economic questions developed relatively late, after periods spent working in his family's brewing business and in journalism. In 1856 he competed in the examination by which the Whately professorship of political economy at Trinity College, Dublin, was then filled, and was appointed for a five-year term. In 1859 he was also appointed Professor of Political Economy and Jurisprudence at Queen's College, Galway, a post which he held until 1870. However, he employed a deputy to perform his duties in Galway after he himself moved to London in 1865. In 1866 he became Professor of Political Economy at University College, London, but was forced to resign in 1872 by the progress of the rheumatic disease which left him almost completely paralysed before his death in 1875.

Cairnes has often been described as ‘the last of the classical economists’. He always worked within the framework of the Ricardo–Mill tradition, devoting himself to refining and strengthening it and seeing no necessity for any radical reform or reconstruction. Within these self-imposed limits and in a career of less than 20 years as a professional economist, he succeeded in making contributions to both theoretical and applied economics which earned him a high reputation among his contemporaries and a definite place in the history of economic thought.

Cairnes's first work in economics proved to be one of his most enduring contributions to the subject. This was *The Character and Logical Method of Political Economy* (1857; 2nd edition, 1875) which is still regarded as one of the best statements of the verificationist methodology of the English classical school. Following the lines laid down by Senior and Mill, Cairnes stressed the neutrality of economic

science, emphasized the value of the deductive method and characterized the subject as a hypothetical science ‘asserting, not what *will* take place, but what *would* or what tends to take place’ ([1857] 1875, p. 55).

It was in the use of the deductive method to develop the central areas of economic theory that Cairnes's main interest came to lie. Yet it was through his work on applied economics and current issues of policy that he first came to be nationally and internationally known. In September 1859 Cairnes published the first of a series of ‘Essays towards a solution of the Gold Question’ in which he sought to ‘apply the principles of economic science’ in an attempt to ‘forecast the directions in which the course [of trade and prices] would be modified by the increased supplies of gold’. This a priori approach was almost precisely the opposite of that used by Jevons to deal with the same problem, but their results coincided remarkably.

It was another application of this approach which first made Cairnes's work known to a much wider audience. In *The Slave Power* (1862) he sought to explain on economic grounds the appearance of slavery in the southern parts of the United States, tracing out both the conditions for and the consequences of the operation of a slave economy. As an indictment of the political economy of the Confederate States it strongly influenced public opinion in Britain towards support of the Northern states in the American Civil War.

Between 1864 and 1870 Cairnes wrote a number of articles on the problems of land tenure in Ireland, in which he argued in favour of proposals to fix rent by law and contended that this was not inconsistent with classical rent theory. There is evidence that his views on this and other questions of the day, such as Irish university education, exerted considerable influence on (and through) Mill and Fawcett.

Cairnes's most important contribution to economic analysis, *Some Leading Principles of Political Economy Newly Expounded* (1874), was also to be his last work and that by which he came to be most widely known and judged. In it he restated, but with significant modifications, the essentials of classical doctrine on the central questions of value, distribution and international trade. His most important innovation was to show that the existence of ‘non-competing groups’ in labour markets implied that the cost of production theory must be supplemented by the analysis of reciprocal demand in the theory of domestic as well as international values.

Nevertheless his unsympathetic review of Jevons's *Theory of Political Economy* (*Fortnightly Review*, N. S., vol. 11, 1872) showed that he lacked interest in and understanding of the subjective approach to value theory which was then developing. Cairnes's treatment of distribution in the *Leading Principles* echoed Mill in showing sympathy for the position of the labourer combined with pessimism based on acceptance of Malthusian population theory; but it was chiefly notable for an elaborate but ultimately unsuccessful attempt to rehabilitate the wages-fund doctrine abandoned by Mill himself in 1869. The verdict of Schumpeter (1954, p. 533) still seems appropriate: Cairnes ‘expounded the old analytical economics and explicitly distanced himself from the new’.

Selected works

1857. *The Character and Logical Method of Political Economy*. London: Macmillan; 2nd edn, 1875; repr. 1888.

1862. *The Slave Power: Its Character, Career, and Probable Designs: Being an Attempt to Explain the Real Issues Involved in the American Contest*. London: Macmillan; 2nd edn, 1863.

1873. *Political Essays*. London: Macmillan.

1873. *Essays in Political Economy, Theoretical and Applied*. London: Macmillan.

1874. *Some Leading Principles of Political Economy Newly Expounded*. London: Macmillan.

Bibliography

Black, R.D.Collison. 1960. Jevons and Cairnes. *Economica*, n.s. 27, 214–32.

Boylan, T.A. and Foley, T.P. 1984. John Elliott Cairnes, John Stuart Mill and Ireland: some problems for political economy. In *Economists and the Irish Economy*, ed. A.E. Murphy. Dublin: Irish Academic Press.

O'Brien, G. 1943. J.S. Mill and J.E. Cairnes. *Economica*, n.s. 10, 273–85.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Weinberg, A. 1970. *John Elliott Cairnes and the American Civil War*. London: Kingswood Press.

How to cite this article

Black, R. D. Collison. "Cairnes, John Elliott (1823–1875)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000002> doi:10.1057/9780230226203.0182

The New Palgrave Dictionary of Economics Online

calculus of variations

Morton I. Kamien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The development of the calculus of variations is attributed to Euler and Lagrange, although some of it can be traced back to the Bernoullis. A history of the calculus of variations is provided by Goldstine (1980). The calculus of variations deals with the problem of determining a function that optimizes some criterion that is usually expressed as an integral. This problem is analogous to the differential calculus problem of finding a point at which a function is optimized, except that the point in the calculus of variations is a function rather than a number. The function over which the optimum is sought is usually restricted to the class of continuous and at least piecewise differentiable functions.
A typical calculus of variations problem is of the form

$$\max_{x(t)} \int_{t_0}^{t_1} F[t, x(t), x'(t)]\,dt \text{ . s . t . } x(t_0) = x_0, \tag{1}$$

where $x'(t)=dx/dt$, and t, $x(t)$, and $x'(t)$ are regarded as independent arguments of the function F. The necessary conditions for $x^*(t)$ to maximize (1) are the Euler equation

$$F_x = dF_{x'}/dt, \tag{2}$$

F the Legendre condition

$$F_{x'x'} \leq 0 \tag{3}$$

and the transversality conditions

$$F_{x'} = 0 \text{ at } t_1, \text{ if } x(t_1) \text{ is free,} \tag{4a}$$

$$F - x' F_{x'} = 0 \text{ at } t_1, \text{ if } t_1 \text{ is free,} \tag{4b}$$

where F_x and $F_{x'}$ refer to the partial derivatives of F with respect to x and x', respectively, and $F_{x'x'}$ is the second partial derivative of F with respect to x'. The Euler equation (2) is in general a nonlinear second order differential equation. The initial condition $x(t_0)=x_0$ and the transversality condition (4a) provide the means for determining the two constants of integration that arise in solving the Euler equation. The optimal value of the upper limit of integration, t_1, if it can be chosen, is determined by the transversality condition (4b). The problem posed in (1) can be extended to include additional arguments of the function F, to include a variety of additional constraints, and to involve double integrals (see Kamien and Schwartz, 1981). Concavity of F with respect to $x(t)$ and $x'(t)$ assures that the necessary conditions are also sufficient.

The earliest application of the calculus of variations to the analysis of an economic problem appears to have been attempted by Edgeworth (1881), who seems to have been greatly impressed by its successful employment in deriving some of the basic laws of physics. He sought to employ it to find a function for distributing income and assigning work among the members of society so as to maximize total social welfare. Many applications of the calculus of variations to economic problems have been conducted since then, a few of which will be described.

As the calculus of variations deals with the problem of finding a function or a path that maximizes some criterion, its major application in economics has been to problems involving optimal decision making through time where an entire course of actions is sought rather than a single action. One of the earliest and most influential applications along these lines is by Ramsey (1928). The question he addressed is how much should a nation save out of its national income through time so as to maximize its overall welfare over time. Ramsey argued that the discounting of future utilities was ‘ethically indefensible’ as

it means that we give less weight to the utility of future generations than to our own. He posited, therefore, a maximum level of net utility, the utility of consumption minus the disutility of work, that he called bliss. This bliss level of utility is the asymptotic limit of the achievable level of net utility. Ramsey then sought the savings rate through time that would minimize the integral over the indefinite future of the difference between the bliss level of utility and the actual net utility level at each point in time, subject to the constraint that savings plus consumption equal total output at each instant of time. The rule he derived for the optimal savings rate, through the Euler equations, is that the ‘rate of saving multiplied by the marginal utility of consumption should always equal bliss minus actual rate of utility enjoyed’. This is essentially a marginal sacrifice today equals marginal benefit tomorrow rule. The rationale for taking the upper limit of integration to be infinite in the objective function is that while individuals have finite lives, society as a whole goes on forever. Ramsey also took up the case where future utilities are discounted at a constant positive rate and derived what may be regarded as the fundamental equation of optimal consumption through time, namely that the proportionate rate of change of marginal utility of consumption should equal the difference between the marginal productivity of capital and the rate at which future utility is discounted. The Ramsey model became the basis for optimal growth theory that was intensely investigated in the late 1950s and 1960s.

Strotz (1956) addressed the question of the circumstances under which an individual would continue today to follow the optimal consumption plan through time that he had determined at an earlier date. In other words, he asked for the conditions under which an optimal consumption plan through time would be consistent. He found the necessary and sufficient conditions for consistency to be that ‘the logarithmic rate of change in the discount function must be constant’. Exponential discounting at a constant rate satisfies this criterion.

Yaari (1965) addressed the question of an individual's optimal consumption plan through time when his lifetime is uncertain. He also allowed for the possibility that the individual derives utility from a bequest to his heirs. Yaari found that a major effect of the presence of uncertainty about one's lifetime is the same as an increase in the rate at which future utilities are discounted. Thus, the ‘effective’ rate at which future utilities are discounted has a risk premium term added to the discount rate in the absence of uncertainty about one's lifetime. The risk premium term is the instantaneous conditional probability of dying in the next instant given survival to the present. The presence of the risk premium means that the rate of consumption at any point in time is higher than it is in its absence. Uncertainty about one's lifetime increases one's rate of current spending, if there is no bequest motive.

While Ramsey applied the calculus of variations to the problem of optimal savings through time, Evans (1924) appears to have been the first to have employed it for determining the optimal rate of output through time. Evans used, as his vehicle for making the problem of choosing the level of output so as to maximize a monopolist's profit over an interval of time nontrivial, i.e. just simple maximization of profit at each instant of time, the assumption that the demand function for a good depended both on its current price and the rate of change of price. In particular, he assumed that the demand function was linear in price and its first derivative, and that the cost of production was a quadratic function of the level of output. Under these assumptions Evans sought the level of production that would maximize the integral of profits over a finite horizon. He was able to characterize this path and to show that a particular solution to the second order differential equation stemming from the Euler equation was the static monopoly profit maximizing level of output. Indeed, it is not difficult to show that when the problem is posed as one of maximizing the present value of an infinite horizon profit stream that the static

monopoly profit maximizing level of output and the corresponding monopoly price constitute a steady-state towards which the output and price paths converge through time. This, of course, is intuitively plausible, as in the steady-state the rate of change of price with respect to time is zero, and so the demand function depends only on the current price level. Evans's work was extended by Roos (1925) to the case of duopolistic producers of a homogeneous product seeking to maximize their individual profits through time. The Roos paper may be regarded as the earliest analysis of what has come to be known as a differential game (see Fershtman and Kamien, 1987).

The last paper that deserves special mention because of its important application of the calculus of variations is Hotelling's (1931), dealing with the rate at which a mineral resource such as coal, copper or oil should be extracted from a mine and sold so as to maximize the present value of its profits. Hotelling derived the fundamental equation for optimal extraction, under competitive production of the resource, namely that the extraction rate be such as to equate the percent change in price through time with the rate of interest at each instant in time. The intuitive reason for this is that if the percent change in the price of the resource exceeds the interest rate then it pays to extract and sell more today, because the alternative of extracting less and earning the interest on the revenue from that level of extraction yields less. The increase in the current rate of extraction, however, causes price to decline until the percent change in the price through time is equalized with the rate of interest. A similar analysis yields that current extraction will decline if the percent change in price is below the interest rate, which in turn will cause price to rise until equality is achieved. Along the optimal extraction path the mine owner is just indifferent between extracting an extra unit of resource today and extracting it tomorrow. A similar analysis can be carried out for a monopolistic mine owner, with the percent change in marginal revenue through time being equated with the interest rate.

There have been a very large number of applications of the calculus of variations since these early ones. Many have employed optimal control methods and dynamic programming methods, both of which constitute generalizations of the calculus of variations. As long as decision making though time is regarded as an important subject of economic analysis, the calculus of variations will continue to find use in economics.

See Also

- Edgeworth, Francis Ysidro
- Evans, Griffith Conrad
- Ramsey model
- Roos, Charles Frederick

Bibliography

Edgeworth, F.Y. 1881. *Mathematical Psychics*. Reprinted, New York: Augustus M. Kelley, 1967.

Evans, G.C. 1924. The dynamics of monopoly. *American Mathematical Monthly* 31, 75–83.

Fershtman, C. and Kamien, M. 1987. Dynamic duopolistic competition with sticky prices. *Econometrica*.

Goldstine, H.H. 1980. *A History of the Calculus of Variations*. New York: Springer-Verlag.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Kamien, M.I. and Schwartz, N.L. 1981. *Dynamic Optimization*. New York: North-Holland.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Roos, C.F. 1925. A mathematical theory of competition. *American Journal of Mathematics* 47, 163–75.

Strotz, R.H. 1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23, 165–80.

Yaari, M.E. 1965. Uncertain lifetime, life insurance and the theory of the consumer. *Review of Economic Studies* 32, 137–50.

How to cite this article

Kamien, Morton I. "calculus of variations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000003> doi:10.1057/9780230226203.0183

The New Palgrave Dictionary of Economics Online

calibration

Edward C. Prescott and Graham V. Candler

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The methodologies used in aerospace engineering and macroeconomics to make quantitative predictions are remarkably similar now that macroeconomics has developed into a hard science. Theory provides engineers with the equations, with many constants that are not well measured. Theory provides macroeconomists with the structure of preference and technology and many parameters that are not well measured. The procedures that are used to select the parameters of the agreed upon structures are what have come to be called ‘calibration’ in macroeconomics.

Keywords

calibration; elasticity of intertemporal substitution; equity premium; impatience; Lucas critique; measurement; neoclassical growth theory; risk aversion; total factor productivity

Article

What is calibration? In the dictionary definition, calibration is the act of calibrating a measurement instrument so that it gives the correct measurement for some known conditions. When calibrating a thermometer that will be used to measure the air temperature, calibration would involve setting it to read 100 degrees Celsius when submerged in boiling water at sea level and zero degrees when submerged in ice water. Because the boiling point of water varies with altitude, the calibration would be different in Mexico City, which is more than a mile above sea level.

Sometimes macroeconomists calibrate a measurement instrument – that is, a model – in this narrow sense. But calibration has gained a broader meaning in economics and is what macroeconomists do when using theory to derive *quantitative theoretical inference*. Prescott emphasizes that calibration is not estimation. Calibration is a process that uses theory to construct a model – that is, an instrument – which will be used to provide a quantitative answer to a question.

Clearly, instruments are not measured; rather, they are calibrated so that they can be used to accurately answer quantitative questions. The nature of questions varies. Examples of questions are as follows:

what is the welfare benefit or cost of changing the currently employed policy arrangement to another one? What will happen to a spacecraft when it enters the atmosphere of Mars?

To predict the quantitative consequences of a particular policy, theory and observations are used to select a model economy, and the equilibrium behaviour of that economy is determined for the proposed policy. Theory provides a set of instructions for selecting the model economy. *This selection process is what calibration in economics has come to mean*. Needless to say, the nature of the application of theory and the availability of economic statistics dictate which model economy is selected.

Before proceeding, a little history of the development of macroeconomics is needed. The modern national accounts were developed by the NBER staff in the 1920s, with Simon Kuznets playing the leading role. In the 1950s and 1960s, macroeconomists searched for *the dynamic system* governing the behaviour of these accounts. The controls for this dynamic system were policy actions. Not having much theory, this activity was largely empirical. Macroeconomists would write down a parametric set of models and find the one that best *fitted* the national accounts, augmented with other statistics. This search for *the dynamic system* failed because, as established in the Lucas critique, the existence of such a policy invariant dynamic system is inconsistent with dynamic economic theory.

The failure of this search led to a vacuum in quantitative macroeconomics. The profession did not want to go back to conjecturing and story-telling that characterized pre-war business cycle theory. As a result, the 1970s was a frustrating decade for quantitative macroeconomists given the failure of the empirical approach and the lack of needed tools and theory to quantitatively study macroeconomic behaviour.

This vacuum was filled in the early 1980s when the extended neoclassical growth model was used to study business cycles. The national accounts had to be modified to be consistent with the model. The most important modification in the study of business cycles is treating consumer durable expenditures as an investment and imputing consumption services to the stock of consumer durables as is done for owner-occupied housing. The secular growth observations with constancy in shares of output led to a constant elasticity structure with share and elasticity parameters. The fact that capital share of income displayed no trend even though the relative price of labour increased secularly led to a unit elasticity of substitution aggregate production function with share parameters equal to income shares. The depreciation rate, for example, was calibrated to average depreciation share of product. The national accounts use prices of used capital goods to estimate depreciation.

This methodology is used in virtually all quantitative theoretical aggregate studies. We emphasize that quantitative theoretical research and empirical research are fundamentally different activities and fundamentally different tools are needed. If the objective of the research is to derive the quantitative implications of the neoclassical growth theory for business cycle fluctuations, the use of statistical tools to select the parameters that best fit the business cycle observations is not sound scientific practice.

In this short article macroeconomist Prescott will describe what he does when addressing macroeconomic issues and aerospace engineer Candler will describe what he does when addressing the problem of making predictions of what will happen when a capsule enters the atmosphere of Mars. These predictions are relevant to the design of the capsule. Prescott will conclude by comparing the approaches and argue that these scientific approaches are essentially the same. We begin with what aerospace engineers do so that comparison can be made with what they do and what macroeconomists do.

Candler: the aerospace engineer

I work in the field of aerodynamics, and specifically I try to predict what happens when a spacecraft enters the atmosphere of a planet. For example, one of my current projects involves predicting how the Mars Science Laboratory capsule will fly as it enters the Martian atmosphere. What is the peak heat transfer rate to the spacecraft? How much heat shield is required to protect it from the extremely high temperature gas that surrounds it during atmospheric entry? Will it produce enough lift so that it will fly along the planned trajectory? Will the uncertain state of the atmosphere cause the capsule to veer off course? These questions must be answered to a known level of accuracy before the spacecraft can be designed. Failure to predict heating levels or aerodynamic performance can result in a well-publicized and expensive loss of the mission. At the same time, excessive conservatism in the design reduces the useful payload of the spacecraft and increases the cost of the mission.

How do we go about modelling this complex problem? We cannot fly a statistical ensemble of missions and empirically extrapolate to the flight conditions of interest. Instead, we must rely on ground-based wind-tunnel testing and theory-based simulations. However, experiments have a number of limitations: it is impossible to test the full-scale capsule; it is usually impossible to produce the actual flight conditions; and we cannot produce the actual intense heating levels for realistic periods of time. On the other hand, we can use numerical simulations to predict the flow field around the full-scale spacecraft at critical points in the entry trajectory. In principle, these calculations can predict the heat transfer rates and aerodynamic forces, and provide accurate data for the spacecraft designers. Of course, these simulations are only as accurate as the underlying equations being solved, and herein lies the problem. We cannot rely on purely empirical measurements to test a spacecraft design, yet simulations require a set of governing equations that must be validated by realistic flight experiments.

Interestingly, the basic set of governing equations that describes the flow over a spacecraft entering a planetary atmosphere is well established. However, there are many parameters in these equations that are the subject of intense debate within my field. We do not have an accurate understanding of the chemical reaction rates in the flow field; we do not know how to model transition to turbulence in the flow near the surface; we cannot predict how much turbulent flow enhances the heat transfer rate; and we do not understand how the high-temperature gas interacts with the spacecraft surface. A complete model of the flow over a spacecraft entering the atmosphere of Mars has well over 100 model constants that must be determined before the equations are fully specified. Clearly, with our limited experience base and with the limitations of the ground-based testing facilities, it is fundamentally impossible to determine these model constants with the available data. Rather, we must impose a rigorous theoretical basis for the choice of these model parameters. Also, we must understand the sensitivity of the critical results (heat transfer rate and aerodynamic forces) to the choice of the parameters. For example, there is no sense in investing a lot of time and money to accurately determine a model parameter that has a one per cent effect on the lift at relevant conditions.

So what do we do? We attack the problem from two sides. First, we break the full problem into well-defined parts and use theory and experiment to determine specific parameters under controlled conditions. For example, we might be concerned with how high-temperature oxygen molecules attack a particular heat-shield material. We would commission experiments to address this specific issue at conditions that are as close as possible to the flight conditions. Typically, it is impossible to exactly reproduce the conditions, and we would then perform experiments in different test facilities to help bound the parameters. Theory would then be used to extrapolate from the test conditions to those

encountered in flight. We always try to use a theoretical basis to provide discipline to this process. We never perform atheoretic variations of parameters to try to match the data – if it is necessary to break the laws of physics, there is usually something wrong!

The second approach to modelling the flow field is to determine what parameters really matter to the design. A very useful approach is to use theory and experience to bound the range of all parameters in the model. Then a large number of simulations are performed, sampling from the distribution of each parameter. With enough simulations, it is possible to determine the sensitivity of the spacecraft design to each of the modelling parameters. Usually with this parametric uncertainty analysis it is possible to isolate several critical parameters that require particular attention. For example, Wright, Bose and Chen (2007) determined that eight modelling parameters out of several hundred were responsible for 90 per cent of the uncertainty in the design of a proposed spacecraft. New experiments were then designed and carried out to reduce the uncertainty in these critical parameters.

Another engineering perspective is worth noting. We fully recognize that our representation of the world will never be 100 per cent accurate. Rather, we must quantify the level of accuracy of a given model and determine if we can fly a mission with that implied level of risk. We must quantify levels of uncertainty in a design and recognize that a spacecraft that will never fail will be excessively expensive or will carry so little payload as to be worthless. Thus, there is a calculated risk associated with the uncertainty in our modelling parameters. Of course, we try to reduce this uncertainty, but ultimately we are always forced to live with some level of risk if we want to fly an interesting mission.

Prescott: the macroeconomist

The selection of parameters in quantitative theory is not measurement. However, quantitative theory is often useful in measurement. It is also useful in making predictions and in accounting for observations. Some examples of successful application are as follows.

The Lucas (1978) asset pricing model with the Markov process on the growth rate of endowments places restrictions on the joint behaviour of asset returns and consumption given two parameters that specify the stand-in household's preference ordering. The first parameter is the degree of risk aversion and the second parameter is the degree of impatience. These restrictions hold in worlds in which there are no transaction costs, no taxes, and no intermediation costs. Whether abstracting from certain factors is reasonable or not depends upon the question.

Mehra and Prescott (1985) used this asset-pricing model economy to estimate how much of the historical equity premium is a premium for bearing aggregate risk. We selected a Markov aggregate endowment growth-rate process whose first two moments matched the historical experience. We used observations and theory to restrict the values of the two preference parameters, including numerous observations on household behaviour. This process of restricting these parameters is part of the calibration process. We found that only a small part of the historical equity premium was a premium for bearing aggregate risk for *any* value of the parameters in the restricted range. This model economy is ill suited for measuring the curvature and impatience parameter of the stand-in household, but it was well suited for determining how much of the historical equity premium is for bearing aggregate risk.

I turn now to a case where a key economic parameter was estimated accurately using a calibrated *set* of model economies. The neoclassical growth model used to study business cycles was used to estimate the

leisure intertemporal elasticity of substitution parameter. This parameter is crucial for evaluating tax policies. Because the income and substitution effects roughly offset secularly, balanced growth observations say nothing about the magnitude of this elasticity parameter. If the neoclassical growth model is accepted as a good abstraction for studying business cycles, business cycle observations tie down this parameter. But the profession was reluctant to accept this theory as a useful one for studying business cycles and therefore did not accept the business cycle-based estimate of this elasticity.

This important parameter was tied down by cross-country and cross-time observations on tax rates and labour supply. Tax rates, broadly defined to be those features of policy that affect the households' budget constraint, account for virtually all the large differences in labour supply across the large advanced industrial countries and across time for France, Italy and Germany. That this estimate is the same one found in the study of business cycles gave confidence to the view that business cycles are in major part optimal responses to real shocks including productivity, taxes, and terms of trade. As established theory and measurement were used in this study, this is calibration.

I turn now to a specific application of the neoclassical growth model to the study of the aggregate value of the stock market, which also entailed calibration. The study that began in late 1999 was motivated by the question of whether the stock market was overvalued and about to crash. At that time people did not know how to use this theory to obtain an accurate answer to this question and relied on historical relations such as price–earnings ratios to answer the question.

To address this issue neoclassical growth theory as developed in the study of business cycles was used. The model economy had to be modified in three important ways. First, there had to be at least two production sectors, a corporate and a non-corporate sector. To have a reason for having two producing sectors, the outputs of the sectors must be different and must be aggregated in some way. McGrattan and Prescott (2005) use the standard procedure of introducing an aggregator of the sector outputs that produces a composite final output good. This aggregator has a share parameter that must be *calibrated* to some observation. The observation selected is the average relative outputs of these two sectors. This is a crucial dimension for the model to mimic reality, given the issue being addressed. The conclusion turned out to be insensitive to the elasticity of substitution between these inputs, which was fortunate given there is not good information on this elasticity. Second, the tax and regulatory system had to be modelled explicitly. For example, we set the model's tax rate on corporate distributions equal to the average marginal tax rates on distributions. This is calibration because in the model world this tax rate is the same for all individuals when in fact it is not. Third, we deal with the fact that corporations have large stocks of unmeasured productive assets and that these assets are an important part of the value of corporations, being stocks of knowledge resulting from investment in research and development, organization capital and brand capital. We figure out how to estimate this stock of unmeasured capital using national account data and the equilibrium conditions that the after-tax return on measured and unmeasured capital are equal.

A theory is tested through successful use. The theory correctly predicts the great variation in the value of the stock market in relation to GDP, which varied by a factor of 2.5 in the United States and by a factor of three in the United Kingdom in the 1960–2000 period. Little of this variation is accounted for by the obvious factors, namely after-tax earnings in relation to GDP and the debt–equity ratio, which varied little over time. The secular behaviour of the stock market value, with its large variation in relation to gross national income, turned out to be as predicted by theory and is not due to animal spirits.

Another example of successful calibration is Hayashi and Prescott (2002), who examined why Japan lost

a decade of growth. The neoclassical growth model used in their study is the one used in the study of business cycles. The exogenous parameter paths were working-age populations, capital income tax rates, and total factor productivity parameters (TFP). The TFP parameters were determined residually from the production function given the quantities of the factor inputs and the output. Given these exogenous elements the equilibrium path was computed. The finding is that the Japanese economy behaved as predicted by the theory. The reason for the lost decade of growth was the failure of TFP to grow. This led to the important question of why Japanese TFP failed to grow as it did in western Europe and North America in this period.

Similarities and differences between aerospace engineering and macroeconomics

Both Candler and Prescott study and model aggregate phenomena. Neither can find the answers empirically through trial and error and both must rely on theoretical computer simulations restricted by measurement. We both test for the robustness of our predictions when making predictions as to what will happen in situations never experienced. In one case the prediction is what will happen to a spacecraft that will be sent to Mars. In the other case it is what will be the consequences of implementing a proposed policy arrangement. Both rely on established theory and measurement to draw quantitative inference.
A difference is that the engineers have the equations, while macroeconomists have statements about preferences and technology. A consequence of this is that macroeconomists have the added step of determining the equilibrium equations of their model. Another minor difference is that computational intensity is much greater in aerospace engineering than in macroeconomics.

See Also

- financial market anomalies
- Kydland, Finn Erling
- Lucas critique
- real business cycles
- recursive competitive equilibrium

We thank Gary Hansen, Ellen McGrattan, Berthold Herrendorf, Lee Ohanian, and Bob Lucas for comments. We are responsible for all views expressed.

Bibliography

Hansen, G.D. 1985. Indivisibility and the business cycle. *Journal of Monetary Economics* 16, 309–27.

Hayashi, F. and Prescott, E.C. 2002. The 1990s in Japan: a lost decade. *Review of Economic Dynamics* 5, 206–35.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

McGrattan, E.R. and Prescott, E.C. 2005. Taxes, regulations, and the value of U.S. and U.K. Corporations. *Review of Economic Studies* 72, 767–96.

Mehra, R. and Prescott, E.C. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Wright, M.J., Bose, D. and Chen, Y.K. 2007. Probabilistic modeling of aerothermal and thermal protection material response uncertainties. *AIAA Journal* 45, 399–410.

How to cite this article

Prescott, Edward C. and Graham V. Candler. "calibration." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000571> doi:10.1057/9780230226203.0184

The New Palgrave Dictionary of Economics Online

cameralism

H.C. Recktenwald

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Becher, J. J.; cameralism; fiscal jurisprudence; Justi, J. H. G. von; mercantilism; Sonnenfels, J. von; utilitarianism

Article

Cameralism is the specific version of mercantilism taught and practised in the German principalities (*Kleinstaaten*) in the 17th and 18th centuries. Becher (1635–82), von Justi (1717–71) and von Sonnenfels (1732–1817) are the principal figures who contributed to a vast cameralist literature of about 14,000 titles (Humpert, 1935). The subject matter of *Kameralismus* reflected the political and economic phenomena and problems in the German territorial states. As a branch of ‘science’ it is a fiscal *Kunstlehre*, that is, the practical art of how to govern an autonomous territory efficiently and justly via financial measures designed to fill the state's treasury. Its subject matter includes economic policy, legislation, administration and public finance. While there is no unifying analytical foundation of cameralism, it did develop in two distinct phases (a younger and an older branch) with varied emphasis on its different elements, and since the rising state was, in theory and reality, the focus and *ultima ratio* of political, economic and ethical (occasionally promotive) speculation, cameralism takes on a unitary form (*Gestalt*) only when viewed in retrospect.

The term ‘cameralism’ itself originates in the management of the state's or prince's treasure (*Kammer*, *caisse*, *camera principis*), seen as the principal instrument of economic and political power. In the age of enlightened absolutism, German–Austrian cameralism, based on a somewhat obscure natural-law philosophy, emphasized the paternalistic character of the governments' centralized fiscal policy (not, as is sometimes mistakenly thought, a Keynesian short-run instrument but rather a regulator for development which was to serve the general happiness of the subjects (*Untertanen*), that is, an eudaemonistic utilitarianism). English and French mercantilism, on the other hand, stressed much more the wealth or ‘riches’ of the sovereign as an end.

The princely bureaucrats had been trained in their own universities (for example, Halle, Frankfurt/Oder, Vienna) in ‘fiscal jurisprudence’ (von Stein) – a mixture of both formal budget and tax ‘principles’ – and

a highly pedantic and descriptive systematization of facts and definitions. Analytical economics, insights into the laws of the market and the study of the interaction between market and state (or even of the bureaucratic and political mechanism) are relatively unknown in the simple textbooks of the cameralists, which show otherwise sound common sense. Statistics, important for census and grasping foreign trade, became a new discipline of the cameral curriculum.

The *practical* policy of cameralism concentrated on the development of a country which had been devastated and depopulated in the Thirty Years' War and impoverished by the discovery of the sea route to India and the fall of Constantinople. Under these abnormal circumstances a political and bureaucratic monopoly attempted to reconstruct the economic foundations of the country by an active population policy, the establishment of state manufactures and banks, the extension of infrastructure (canals, bridges, harbours and roads) and the promotion of modernization. It strictly regulated the still important agricultural sector, as well as trade and commerce.

The state protected the trades (*Gewerbe*) by means of high tariffs to restrict imports of unnecessary raw materials and it facilitated exports of manufactures and import substitution. On the other hand, the government removed internal trade barriers by abolishing the medieval guild organization and by unifying the law for municipalities. Mercantilist efforts to augment the state treasure via trade surplus and money policy were, of course, another main cameralistic aim. Finally, it is notable that its monetary policy was inconsistent, in so far as the hoarding of precious metals as opposed to their circulating function was not clearly distinguished.

To set cameralism in secular perspective, the famous arguments of Smith and the Physiocrats against the ‘mercantile system’ seem to be *mutatis mutandis* valid for neo-mercantilism, which also justifies both state intervention in the market and a greater GNP government share and often reverts to the regulatory rules and the principles of planning in this former epoch. However, neo-mercantilism fails to prove seriously both the state's competence to ensure efficiency and equity in the public sector and its ability to regulate the market reasonably. Some writers tend to overlook that in our times the basic conditions in the state and the economy are radically different from those of three centuries ago. For example, economic, political and administrative conditions in the German principalities differed strikingly from Ludwig Erhard's situation after the Second World War. And the wide gap between the Great Depression of the 1930s and the technologically influenced stagflation of the 1980s was obviously so fundamental that the regulatory Keynesian budget and employment theory, with its then unrealistic assumptions, became rather obsolete. Thus any attempt to revive the strict regulating prescriptions of all-embracing cameralism, which lacks sufficient analysis and empirical testing, would apparently be a violation of both reason and experience. In this case we would use analytically poor (and old) tools to repair the wrong (and modern) machine.

Bibliography

Becher, J.J. 1668. *Politischer Diskurs*. Frankfurt/Main: Bielcke.

Humpert, M. 1935–7. *Bibliographie der Kameralwissenschaften*. Cologne: Schroeder. (Includes nearly 14,000 items.)

Justi, J.H.G. von 1755. *Staatswirtschaft*. Leipzig: Breitkopf.

Mohl, R. 1855–8. *Geschichte und Literatur der Staatswissenschaften*, vols 1–3. Erlangen: Enke.

Recktenwald, H.C., ed. 1973. *Political Economy A Historical Perspective*. London: Collier-Macmillan.

Recktenwald, H.C. 1986. *Das Selbsinteresse – Zentrales Axiom der ökonomischen Wissenschaft*. Wiesbaden: Leibniz-Akademie der Wissenschaften.

Small, A.W. 1909. *The Cameralists*. New York: Franklin.

Sommer, L. 1920–5. *Die österreichischen Kameralisten in dogmengeschichtlicher Darstellung*, 2 vols. Vienna: Konegen.

Sonnenfels, J. von 1765–76. *Grundsätze der Polizei, Handlungs- und Finanzwissenschaft*. Vienna: Camesina.

How to cite this article

Recktenwald, H. C. "cameralism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000011> doi:10.1057/9780230226203.0185

The New Palgrave Dictionary of Economics Online

campaign finance, economics of

Scott Ashworth

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys recent work aimed at evaluating the welfare effects of campaign finance reform. The theoretical literature distinguishes two types of contributor: those who desire ideological policies and those who want personal favours. A series of models shows that these different types of contributor have different implications for campaign finance regulation. The models also give some suggestions about the sort of empirical evidence that would argue for or against certain campaign finance regulations. These suggestions have been followed up by recent empirical work.

Keywords

advertising; campaign finance reform; campaign finance, economics of; election; incumbency; matching funds; multiple equilibria; probabilistic voting; voting

Article

Campaign finance is a contentious issue in American politics. Reformers charge that a system in which interest groups provide the funds for campaigns creates opportunities for corruption, while others argue that restrictions on donations would limit the provision of information to voters. For an economist, the natural way to evaluate such arguments is to construct a model that explicitly treats the preferences and beliefs of the voters, to deduce the conditions under which the model predicts welfare improvements from regulation, and to check empirically if these conditions hold in actual elections. This article surveys a recent body of literature that does just that.

1 First-generation models

Early work on campaign finance took a reduced-form approach to the link between campaign activity and votes (Austen-Smith, 1987; Baron, 1989; 1994; Grossman and Helpman, 1996; Snyder, 1990). This

literature identified two ideal types of contributor: position-induced contributors, who help ideologically compatible candidates win office, and service-induced contributors, whose contributions are analogous to purchasing contingent claims on favours provided to the buyer at the expense of citizens in general. This literature yielded several important insights. For example, Baron (1989) finds that trades of contributions for promises of favours have interesting implications for the incumbency advantage (see, for example, Gelman and King, 1990, and Ansolabehere and Snyder, 2002, for empirical work on the incumbency advantage in US elections). A candidate with an exogenous advantage is more likely to be able to deliver the promised favours, making the promise more valuable. Thus an advantaged candidate can raise funds on more favourable terms, reinforcing the advantage. Morton and Myerson (1992) show that this mechanism can even lead to multiple equilibria, where predictions that one candidate will win become self-fulfilling because contributions flow to the presumptive winner.

As the comprehensive survey of this literature by Morton and Cameron (1992) emphasizes, this approach cannot address the welfare qsts raised by proposals for campaign finance reform. We now turn to more recent research that ‘opens up the black box’ and provides some welfare analysis.

2 Microfounded models

A bare-bones model illustrates the main points of the literature. The game has four players: two candidates, a voter, and an interest group.

Each candidate has some level of ‘quality’, which could be either ability or ideological similarity to the voter. The key is that quality is valued by the voter. Candidate *i*'s ability is θ_i. It is common knowledge that $\theta_1 = 1$, and that θ_2 is equally likely to be 0 or 2. Each candidate maximizes his probability of winning.

At the start of the game, the candidates learn θ_2, but the voter does not. At cost $c \in (0, 1)$, candidate 2 can truthfully reveal θ_2. Candidates have no funds of their own. The interest group has sufficient funds to pay for the information transmission, if it wants to.

Even without specifying the group's payoffs, we can derive two benchmarks.

The no-campaign solution. First, assume the interest group is prohibited from funding candidate 2's campaign. Then the voter goes to the polls not knowing θ_2. Thus she is indifferent between the two candidates, and gets expected payoff 1 no matter how she votes. The natural voting rule is to have her toss a fair coin. (This would be the outcome if there were a mean-zero popularity shock prior to the election.) In this case, each candidate gets payoff 1/2.

The voter's optimum. Second, assume there is a planner who can observe the true θ_2 and communicate it to the voter, paying for the communication with a lump-sum tax on the voter.

Announcing the true θ in only one of the states suffices for complete communication, and allows for a cost savings compared with always announcing the state. So the planner announces θ_2 if and only if $\theta_2 = 2$, and the voter votes for 2 if there is an announcement and for 1 if not. Her payoff is

$$\frac{1}{2} + \frac{1}{2}(2 - c) = \frac{3}{2} - \frac{1}{2}c > 1.$$

Thus the voter is better off than in the no-campaign solution. Furthermore, each candidate still wins with *ex ante* probability 1/2, so the policy represents an *ex ante* Pareto improvement over the no campaign solution.
This scheme would be hard to implement, because it is vulnerable to collusion between the regulator and candidate 1. Thus we are interested in whether or not interest-group finance can improve on the no-campaign benchmark.

2.1 Position-induced contributors

Now assume the interest group wants candidate 2 to win independent of θ, perhaps because it shares the candidate's ideology. Formally, the group's payoff is

$$bw - k,$$

where $b > 0$ is the payoff to the group from having 2 win, *w* is an indicator variable equalling 1 if and only if candidate 2 wins, and *k* is the contribution to candidate 2.
The timing is:

1. 1. The candidates and the group learn θ_2.
2. 2. The group chooses a contribution $k > 0$.
3. 3. If $k \geq c$, the candidate decides whether or not to advertise θ.
4. 4. The voter sees any ads purchased, and then selects the winner.

Proposition 1: If $b > c$, then there is a perfect Bayesian equilibrium (PBE) in which

- *the group contributes c if and only if* $\theta_2 = 2$ and
- *the voter chooses candidate 2 if and only if she sees an ad certifying that* $\theta_2 = 2$.

The idea is simple. The group is better off if 2 wins. If $\theta_2 = 2$, the group can ensure that 2 wins by funding a campaign informing the voter of her true preference for 2. And if the benefit from having 2 win (*b*) exceeds the cost (*c*), the group wants to do this. Finally, the group does not contribute to a low type of candidate 2 – this cannot help the group because the candidate cannot lie.
If there are contributions in equilibrium, then the voter gains over the no-campaign solution, having a payoff of $3/2 > 1$. Thus banning contributions reduces the voter's welfare. Furthermore, the equilibrium without contributions is Pareto dominated by the following matching fund policy. Fix γ strictly between 0 and *b*. If the group donates γ to candidate 2, then the regulator kicks in $c - \gamma$, paid for by a lump-sum tax on the voter. The group's *ex ante* payoff increases from 0 to $(b - \gamma)/2$ and the voter's payoff increases from 1 to $3/2 - (c - \gamma)/2 > 1$. The candidates are indifferent at the *ex ante*

stage.
Coate (2003) elaborates on this story in two ways. First, the voter is uncertain about both ideologies, and both candidates can receive contributions. Second, and more importantly, candidates are selected by the party's median member, who has different preferences from the median in the electorate. (Here quality is the inverse of distance from the median.) The interest group prefers less moderate candidates. However, the groups prefer to fund more moderate candidates – campaign ads are effective only when the ad reveals that the candidate is more moderate than a non-advertising candidate. This gives the party an additional incentive to choose moderate candidates, because moderate candidates can raise funds and thus do well in the election. In equilibrium, the party mixes between moderate and extremist candidates. In this environment, simply banning contributions creates both winners and losers. Moderate voters lose. First, they must make their choices with worse information, as in the bare-bones model above. Second, candidates are less likely to be moderate. Members of the interest groups, on the other hand, are better off. They save the cost of the contributions, and policy is no worse in expectation – the extra probability that policy is extreme in the wrong direction is exactly offset by the increased probability that policy is close to the group.

2.2 Service-induced contributors

Now assume the group does not care directly who wins the election. Instead, the group values transfers from the winner. The group and candidate 2 can sign a contract specifying that candidate 2 receives *c* from the group, and, if he wins, he transfers the amount *t* to the group. This transfer if financed by a tax on the voter of $(1+\lambda)t$, where λ represents the deadweight loss of the transfer.
The timing is:

1. 1. The candidates and the group learn θ_2.
2. 2. Candidate 2 makes a take it or leave it offer of a contract *t* to the group.
3. 3. The group accepts or not.
4. 4. If the contract is accepted, the candidate decides whether or not to advertise θ.
5. 5. The voter sees any ads purchased, and then selects the winner.

Proposition 2: If $(1+\lambda)c \leq 1$, *then there is a PBE in which the group funds the campaign if and only if* $\theta_2 = 2$ *and the voter selects candidate 2 if and only if she sees an ad certifying* $\theta_2 = 2$.
Again, the basic idea is simple. If the voter sees an ad, she learns two things. First, she learns that $\theta_2 = 2$, which improves her evaluation of candidate 2. Second, she learns that the group and the candidate have made a deal, so electing candidate 2 costs her $(1+\lambda)c$. This tradeoff is acceptable if $(1+\lambda)c \leq 1$.
In such an equilibrium, the voter's payoff is

$$\frac{1}{2} + \frac{1}{2}(2 - (1+\lambda)c) = \frac{3}{2} - \frac{(1+\lambda)c}{2}.$$

This payoff is lower than the voter-optimal benchmark payoff by $\lambda c / 2$.
Again, matching funds can help. Assume again that the regulator pays $c - \gamma$ of the cost. This policy reduces the welfare loss compared with the benchmark to $\lambda \gamma / 2 < \lambda c / 2$.
Most papers in the literature introduce some uncertainty in the voting stage. With this addition, Prat (2002), Coate (2004) and Ashworth (2006) show that the candidate might promise so much that the voter actually *loses* from the campaign. To see the intuition, consider the candidate's incentive to advertise. Without probabilistic voting, the incentive to expand transfers is limited – once the voter's cost of transfers passes 1, the probability of election changes discontinuously from 1 to 0. With probabilistic voting, by contrast, small changes in transfers have similarly small effects on the re-election probability. In this case, candidates have an incentive to expand transfers all the way to the point where the voter is indifferent between a high-quality candidate with transfers and a low-quality candidate with no transfers. In such a case, the voter actually loses from the possibility of a campaign, and would be better off if contributions were banned outright – the likelihood of getting a high-quality winner is no lower, and the voter escapes the cost of favours.
The key to the inefficiency here is that the voter's knowledge that ads imply favours to interest groups makes the ads less effective at ensuring a high-quality candidate is elected.
Again, matching funds might be a better solution. In Coate (2004), the scale of the campaign can vary continuously. Greater spending increases the fraction of the (large) electorate that is informed. Matching funds come into play if the benefit from winning is low enough that ads are not rendered totally ineffective. In that case, a limit on contributions reduces the amount of favours, preserving the effectiveness of the ads. And the matching funds allow the scale of the campaign to be unchanged from the unregulated case.
So far, matching funds have seemed like a great policy. But they have a cost in asymmetric contests. In Ashworth (2006), the scale of campaigns is fixed (as in the bare-bones model above), but candidate 2 has an advantage independent of advertising. For moderate levels of the advantage, the advantaged candidate mounts a costly campaign even though the value of the information to the voter is less than the cost the voter pays *ex post*. For greater values of the advantage, no campaign takes place in equilibrium – the possible increase in the voter's evaluation is too small to outweigh the promised favours. Matching funds can increase the likelihood of an active campaign in such cases, even though reducing their likelihood would be efficient.

2.3 Hard vs. soft information

The literature focuses on two mechanisms that make advertisements informative. The first is the one we have relied on above, namely, the candidate may have verifiable information, information that cannot be falsified. The second, studied by Gerber (1996), Prat (2002), and Potters, Slooof and van Winden (1997), is indirectly informative campaigns. Interest groups observe the quality of the candidates, but voters do not. If groups condition their contributions on quality, then voters can learn about quality by inverting the contribution schedule. Gerber and Prat show that equilibria with informative advertising exist, even thought the ads have no direct informational content. As in the case with hard information, service-induced contributions imply that a ban on contributions can benefit the median voter. On the

other hand, public financing would have no value with indirectly informative advertising – there's no signal if the election regulator hands out funds to everyone. Thus a non-trivial policy problem of public financing arises only with directly informative advertising.

3 Empirics

3.1 Do contributions buy favours?

Contributors' motivations played a key role in the welfare conclusions above. What do the data say about these motivations? The most direct approach to this question looks at correlations between donations from interest groups and votes that those groups care about. For example, we could regress votes in favour of increasing the minimum wage on contributions from unions. Of course, a positive correlation on its own does not discriminate between the theories – are the union contributions changing votes or do unions just contribute to exogenously union-friendly candidates? The many studies that try to disentangle these forces affecting roll-call votes find only weak evidence that contributions buy votes (Ansolabehere, de Figueiedo and Snyder, 2003). One interpretation is that contributions are position-induced rather than service-induced.

However, focusing on roll calls misses much Congressional activity (Hall, 1996). Thus researchers have also looked to more indirect evidence. For example, Gordon and Hafer (2005) find that firms making large donations are less monitored by agencies, suggesting that donations induce members of Congress to interfere in regulatory oversight. Many papers have shown that political action committees (PACs) direct their contributions in ways more consistent with service-induced motivations than with position-induced motivations (Kroszner and Stratmann, 1998; Romer and Snyder, 1994; Snyder, 1990). Perhaps the most convincing is McCarty and Rothenberg (1996), who document that individual PACs made significant shifts in donations from Democrats to Republicans after the Republicans took control of Congress in 1994, suggesting that the contributions were not ideological.

Attempts to directly estimate the impact of contributions on policy have not reached a consensus, except that the effects are smaller than public outcry might suggest (Ansolabehere, de Figueiedo and Snyder, 2003). The next subsection turns to a more theory-driven approach to evaluating the potential for welfare gains from regulation.

3.2 Spending and election outcomes

A substantial empirical literature has tried to estimate the effect of campaign spending on electoral outcomes. Cross-sectional analyses that do not condition on incumbent quality show that challenger spending is associated with better electoral performance, but incumbent spending is unrelated to success. (See the discussion in Jacobson, 2001, ch. 3, which summarizes the extensive empirical work initiated by Jacobson, 1978.) Of course, interpreting these correlations is difficult because of an endogeneity problem – candidates spend more when they expect the race to be competitive. Several researchers have tried to deal with this endogeneity issue (Green and Kranso, 1988; Levitt, 1994; Gerber, 1998; Erikson and Palfrey, 1998; 2000). These papers all find that spending is roughly equally effective for both

incumbents and challengers, but there is no consensus about the size of the effects. (Looking across several of the most prominent estimates, Gerber, 2004, calculates an implied cost for a House incumbent to get one additional vote ranging from $15 to $367.)

Prat (2000) points out that, even when one controls for candidate quality, there is an identification problem in these regressions. Simply put, the functional relationship between spending and election outcomes (with quality held fixed) depends on the way funds are raised. To see this, consider the models of service-induced contributions discussed previously. In all of the models, an exogenous increase in quality has two effects. First, the candidate raises more funds and informs the voters of his high quality, which helps his electoral chances. Second, the voter infers that the funds were given in exchange for promises of favours, which hurts his electoral chances. Thus the regressions estimate ‘the effect on electoral outcomes of an extra dollar of campaign spending net of the political cost of persuading lobbies to donate the extra dollar’ (Prat, 2006, p. 60).

In addition to providing an important critique of the standard inpts of the empirical evidence, the prediction that the effectiveness of advertising is decreasing in the degree of service-induced contributing provides a way to test empirically for the possibility of welfare-improving policy. In particular, the theoretical models suggest that limits on contributions and (perhaps) matching funds can improve welfare precisely when campaign spending is ineffective. Thus the prediction of reduced effectiveness speaks directly to the welfare implications of the models.

Stratmann and colleagues have been leaders in testing these implications. Houser and Stratmann (2006) carry out laboratory experiments modelled after the theoretical set-up of Coate (2004) and Ashworth (2006). High-quality candidates are more likely to win in a public financing treatment than in a privately financed treatment. They also find that margins of victory are greater in the public financing treatment. In a treatment with caps on contributions, they find that voter welfare goes up, but the probability of electing a high-quality incumbent does not. These experiments support the theoretical predictions, suggesting that voters are capable of inferring that interest-group financed ads imply that the candidate has promised favours.

Stratmann (2006) exploits state-level variation in campaign finance laws to see whether the theoretical predictions hold up in field data. He first estimates standard vote-share/spending regressions for each state's House elections. He then examines the relationship between the effectiveness of spending and the existence of limits on contributions. As predicted by the theory, he finds that effectiveness is lower when campaign finance regulations are more liberal. These results hold for all of incumbents, challengers, and open-seat candidates. Stratmann and Aparicio-Castillo (2006) show that states that limit giving subsequently have lower incumbent vote shares. This finding is consistent with Baron's (1989) and Ashworth's (2006) theoretical finding that the financing process can exaggerate incumbency advantages.

See Also

- political competition
- political institutions, economic approaches to
- rent seeking

Bibliography

Ansolabehere, S., de Figueiedo, J.M. and Snyder Jr., J.M. 2003. Why is there so little money in U.S. politics? *Journal of Economic Perspectives* 17(1), 105–30.

Ansolabehere, S. and Snyder Jr., J.M. 2002. The incumbency advantage in U.S. elections: an analysis of state and federal offices, 1942–2000. *Election Law Journal* 1, 315–38.

Ashworth, S. 2006. Campaign finance and voter welfare with entrenched incumbents. *American Political Science Review* 100, 55–68.

Austen-Smith, D. 1987. Interest groups, campaign contributions, and probabilistic voting. *Public Choice* 54, 123–39.

Baron, D.P. 1989. Service-induced campaign contributions and the electoral equilibrium. *Quarterly Journal of Economics* 104, 45–72.

Baron, D.P. 1994. Electoral competition with informed and uninformed voters. *American Political Science Review* 88, 33–47.

Coate, S. 2003. Political competition with campaign contributions and informative advertising. *Journal of the European Economic Association* 2, 772–804.

Coate, S. 2004. Pareto-improving campaign finance policy. *American Economic Review* 94, 628–55.

Erikson, R.S. and Palfrey, T.R. 1998. Campaign spending and incumbency: an alternative simultaneous equations approach. *Journal of Politics* 60, 355–73.

Erikson, R.S. and Palfrey, T.R. 2000. Equilibria in campaign spending games: theory and data. *American Political Science Review* 94, 595–609.

Gelman, A. and King, G. 1990. Estimating incumbency advantage without bias. *American Journal of Political Science* 34, 1142–64.

Gerber, A. 1996. Rational voters, candidate spending, and incomplete information: a theoretical analysis with implications for campaign finance reform. Working Paper No. 96–01.1, Institution for Social and Policy Studies, Yale University.

Gerber, A. 1998. Estimating the effect of campaign spending on senate election outcomes using instrumental variables. *American Political Science Review* 92, 401–11.

Gerber, A.S. 2004. Does campaign spending work?: Field experiments provide evidence and suggest new theory. *American Behavioral Scientist* 47, 541–74.

Gordon, S.C. and Hafer, C. 2005. Flexing muscle: corporate political expenditures as signals to the bureaucracy. *American Political Science Review* 99, 245–61.

Green, D.P. and Kranso, J.S. 1988. Salvation for the spendthrift incumbent: reestimating the effects of campaign spending in House elections. *American Journal of Political Science* 32, 884–907.

Grossman, G.M. and Helpman, E. 1996. Electoral competition and special interest politics. *Review of Economic Studies* 63, 265–86.

Hall, R.L. 1996. *Participation in Congress*. New Haven, CT: Yale University Press.

Houser, D. and Stratmann, T. 2006. Selling favors in the lab: experiments on campaign finance reform. Working Paper No. 1727, CESifo.

Jacobson, G.C. 1978. The effects of campaign spending in Congressional elections. *American Political Science Review* 72, 469–91.

Jacobson, G.C. 2001. *The Politics of Congressional Elections*, 5th edn. New York: Addison-Wesley Educational Publishers, Inc.

Kroszner, R.S. and Stratmann, T. 1998. Interest Group competition and the organization of congress: theory and evidence from financial services political action committees. *American Economic Review* 88, 1163–87.

Levitt, S.D. 1994. Using repeat challengers to estimate the effect of campaign spending on election outcomes in the U.S. House. *Journal of Political Economy* 102, 777–98.

McCarty, N. and Rothenberg, L.S. 1996. Investment in politicians? Evidence from the 1994 elections. Paper presented at the 1996 annual meeting of the American Political Science Association, 29 August–1 September.

Morton, R.B. and Myerson, R.B. 1992. Decisiveness of contributors' perceptions in elections. Working paper, New York University.

Morton, R. and Cameron, C. 1992. Elections and the theory of campaign contributions: a survey and critical analysis. *Economics and Politics* 4, 79–108.

Potters, J., Sloof, R. and van Winden, F. 1997. Campaign expenditures, contributions, and direct

endorsements: the strategic use of information to influence voter behavior. *European Journal of Political Economy* 13, 1–31.

Prat, A. 2000. Campaign spending with office-seeking politicians, rational voters, and multiple lobbies. *Journal of Economic Theory* 103, 162–89.

Prat, A. 2002. Campaign advertising and voter welfare. *Review of Economic Studies* 69, 997–1017.

Prat, A. 2006. Rational voters and political advertising. In *The Oxford Handbook of Political Economy*, ed. B.R. Weingast and D. Wittman. Oxford: Oxford University Press.

Romer, T. and Snyder Jr., J.M. 1994. An empirical investigation of the dynamics of PAC contributions. *American Journal of Political Science* 38, 745–69.

Snyder, J.M. 1990. Campaign contributions as investments: the U.S. House of Representatives, 1980–1986. *Journal of Political Economy* 98, 1195–227.

Stratmann, T. 2006. Contribution limits and the effectiveness of campaign spending. *Public Choice* 129, 461–74.

Stratmann, T. and Aparicio-Castillo, F.J. 2006. Competition policy for elections: do campaign contribution limits matter? *Public Choice* 127, 177–206.

How to cite this article

Ashworth, Scott. "campaign finance, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000255> doi:10.1057/9780230226203.0186

The New Palgrave Dictionary of Economics Online

Canada, economics in

Robert W. Dimand and Robin F. Neill

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the first half of the 20th century, economics in Canada was primarily economic history, and its contribution was the staple theory of Canadian economic development. After the Second World War Keynesian macroeconomics swept the nation and, despite its British origin, it indigenized into a theory of primary product export-based growth, and a Western Marxist theory of the staple trap. In the last quarter of the century, positivism, monetarism, and neoconservative new classical economics swept north from the United States, leaving only the specific domestic circumstances to which it was applied as the distinctive thing about economics in Canada.

Keywords

Ashley, W. J.; Bladen, V.; Brittnell, G.; Bryce, R.; Canada, economics in; conspicuous consumption; Cournot, A. A.; Dales, J.; dependency; Easterbrook, W.; Ely, R. T.; English Historical School; Fowke, V.; Galbraith, J. K.; German Historical School; Gourlay, R.; Hume, D.; infant-industry protection; Ingram, J.K.; Innis, H. A.; institutional economics; internal rate of return; Johnson, H. G.; Mackintosh, W. A.; mathematical methods in political economy; Mavor, J.; monetary policy rules; Mundell, R.; periphery; Plumptre, A. F. W.; political arithmetic; Rae, J. (1796–1872); Saunders, S.; Social Credit; staples thesis; technical change; time preference; Timlin, M. F.; Viner, J.; Wakefield, E. G.

Article

Beginnings

The pre-history of economics in Canada begins with the description of the society and products of New France by Pierre Boucher (1664), a former governor at Trois Rivières and the founding seigneur of Boucherville, writing in the political arithmetic tradition of Boisguilbert and Vauban. The most notable of such descriptive works was, after the British conquest, the vast, disorganized, but often incisive

Statistical Account of Upper Canada by the political dissident Robert Gourlay (1822), whose criticism of unrepresentative and corrupt government led to his exile as an undesirable alien – on the grounds of his birth in Scotland rather than England (Dimand, 1992). Although, apart from Boucher and Gourlay, early descriptive writings about settlement and economic conditions in Canada tended to have little economic analysis, Boucher displayed an intuitive sense of economies of scale, urging that policy should encourage concentration of settlement in small areas, where mutually beneficial exchange would lead to a surplus product. Independently, Gourlay later formulated a linear relationship between land values and the number of inhabitants per acre. He urged the government to borrow to fund increased immigration and settlement, paying off the loan by taxing the resulting increase in land value. The influence of Gourlay's theorizing about the appropriate structure of property rights to promote population density in a newly settled colony (such as limiting the size of land grants to avoid dispersion of settlers) was acknowledged by Edward Gibbon Wakefield, the English theorist of colonization who wrote Appendix B on land policy for Lord Durham's report on Canada after the 1837 rebellion and then served in the Canadian legislature before taking a leading role in the settlement of New Zealand (Wakefield, 1968; Goodwin, 1961, ch. 1; Neill, 1991, ch. 1). One Canadian topic, the playing card currency of New France, so often cited by later economic historians (for example, Shortt, 1987), attracted the attention of one of the great early economists, the philosopher David Hume, as British chargé d'affaires in Paris after the Seven Years War and then as Under-Secretary of State; Hume negotiated the settlement of the outstanding paper money of New France after the British Conquest (Dimand, 2005).

John Rae was an outstanding 19th-century economic theorist who wrote his *New Principles of Political Economy* (1834) while headmaster of the Gore District Grammar School in Upper Canada (now Ontario). Rae, although born and educated in medicine in Scotland, eventually became a district judge in the Kingdom of Hawaii before dying in Staten Island. For decades, he was known primarily through John Stuart Mill's citation of his statement of the infant industry argument for protection: although Sir John A. MacDonald, Canada's first prime minister, cited Rae in support of his national policy of tariff protection for manufacturing, he seems to have known of Rae only through Mill (MacDonald, quoted in Neill, 1991, pp. 85–91). C.W. Mixter's new, rearranged edition of Rae's book in 1905 revealed Rae's analysis of ‘effective desire of accumulation’ as a pioneering capital theory, and two years later Irving Fisher dedicated The Rate of Interest ‘to the memory of John Rae who laid the foundations upon which I have endeavored to build’, acknowledging Rae for foreshadowing both time preference and internal rate of return over costs. Rae has since been celebrated for his discussions of conspicuous consumption, more than six decades before Thorstein Veblen, and of endogenous technical change (James, 1965; Hamouda, Lee and Mair, 1998). University of Toronto mathematics professor John Bradford Cherriman (educated at St John's College, Cambridge, a few years before Alfred Marshall) made another striking, but isolated, contribution to economic theory: a ten-page review article and exposition of Cournot's essay in mathematical economics of 19 years before, endorsing the mathematical approach to political economy, hailing Cournot's work as more important than Ricardo, and long antedating Joseph Bertrand's 1883 article that was long thought to be the first review of Cournot's 1838 volume (Cherriman, 1857; Dimand, 1988; 1995). More characteristic of this period than the theorizing of Rae and Cherriman were the numerous practical and descriptive discussions of economic affairs, economics in the context of action (see Goodwin, 1961; Neill, 1991; Neill and Paquet, 1993).

The rise of academic economics in Canada

Although a few courses had been offered previously, economics in Canadian universities began in 1888 with the appointment of the English historical economist W.J. (later Sir William) Ashley as professor of political economy and constitutional history at the University of Toronto and of Adam Shortt (previously tutor in philosophy, instructor in botany, and demonstrator in chemistry) as lecturer in political economy at Queen's University, Kingston (promoted to Sir John A. Macdonald Professor of Political Science in 1891). Professorial appointments at the university were then made by Order in Council by the provincial government, and candidates were interviewed by the Premier of Ontario and by the chancellor of the University. No classical or neoclassical theorist would have been appointed, lest they promote free trade in their lectures, but the English Historical School was acceptable (Drummond, 1983). When Ashley departed in 1892 to become professor of economic history at Harvard (and later dean of commerce in Birmingham), he was succeeded by James Mavor, Scottish economic historian of Russia and friend to Tolstoy, Kropotkin, and the Doukhobors (see Mavor, 1923), and until 1970 the Department of Political Economy was led by a succession of distinguished economic historians (apart from one sociologist), notably Harold Innis and William Easterbrook, and the historian of economic thought Vincent Bladen (see Drummond, 1983; Bladen, 1978). Under Ashley's sponsorship, the University of Toronto published the first academic economic writing by a Canadian woman, Jean Scott Thomas (1889), ‘The conditions of female labour in Ontario’. As in other disciplines and elsewhere in the Dominions and the British Empire, several early professors of economics in English-speaking Canadian universities, notably Ashley and C.R. Fay in Toronto and A.W. Flux at McGill, were British scholars who had finished their careers in Britain, as was James Bonar, Deputy Master of the Mint in Ottawa and authority on Malthus. The British Association for the Advancement of Science met in Montreal in 1884; in other years it met in Dublin, Cape Town, or Sydney. The following year, the association commemorated its meeting with *Canadian Economics*, a volume of 27 papers by Canadian and American authors that, according to Goodwin (1961, p. 116), ‘marked the end of an era when description and analysis were carried out by interested persons in all walks of life and before there were any professional economists in government and the universities’. The Canadian Political Science Association met in September 1913, with Adam Shortt as president, and published a volume of proceedings, but the September 1914 meeting was cancelled when the First World War broke out, and the association lapsed until 1929.

Long after the social sciences separated in Britain and the United States, they remained institutionally linked in Canada, sharing a single Department of Political Economy at the University of Toronto until 1982 (the equivalent term at McGill and the University of Saskatchewan was Department of Economics and Political Science), a single Canadian Political Science Association and the *Canadian Journal of Economics and Political Science* (first published in 1935) until 1966 (the sociologists and anthropologists seceded in 1963), with the economists departing only much later from the joint annual conferences of the Learned Societies (now the Humanities and Social Science Congress). As Taylor (1960, p. 8) remarks, ‘Shortt, Skelton, Mavor, and Leacock throughout their careers could almost equally well be described as historians or political scientists.’ While the economic historian Harold Innis headed Toronto's Department of Political Economy during the 1930s and 1940s, scholars in the various disciplines there, not all of them within the department, were linked by their historical approach and by

Innis's influence, in historical sociology (S.D. Clark), history of political thought (C.B. Macpherson), history of economic thought (Vincent Bladen), economic history (John Dales, William Easterbrook), historical geography (Andrew Hill Clark), history of communications (Marshall McLuhan), Canadian history (Donald Creighton, Innis's biographer). Formal economic theory, in contrast, was conspicuously absent, except that A.F.W. Plumptre, before joining the public service, taught Keynes's *Treatise on Money*, having studied in Cambridge while that book was being written. When the University of Saskatchewan opened in 1910, economics was taught by the professor of history, using texts by Richard T. Ely, an American economist influenced by the German Historical School, and by Ashley, Archdeacon William Cunningham, and J. Kell Ingram of the English Historical School, but not Marshall or Jevons (Spafford, 2000). One consequence of multidisciplinary sharing of departments, association, and journal was that after the humorist Stephen Leacock, trained in political science and author of a successful textbook in that field, succeeded Flux as Dow Professor of Economics and Political Science at McGill in 1908, he acquired public credibility for his economic pronouncements, such as advocating a tariff union for the British Empire to end the Great Depression.

Growing numbers of academics, and the gains from division of labour in scholarly research and publication as in other activities, led the social sciences in Canada to become increasingly separate after the Second World War, well in advance of formal institutional separation. The British connection and the emphasis on a historical approach also faded in the same decades, as Canadian economics became more grounded in formal theory and quantitative methods and more attuned to intellectual developments in the United States.

The teaching of economics emerged later in French Canada. The journalist Etienne Parent (1846), an admirer of Adam Smith and Jean-Baptiste Say, was unusual in declaring political economy a science and urging the enlightened publication of the principles it taught, notably free trade and the respectability of commerce and industry as occupations. Although Parent became Under-Secretary of State when the Dominion of Canada was created in 1867, his views on the study of economics had little influence. Political economy was widely identified with doctrinaire free traders (such as Parent) and with the secular pursuit of material gain, and did not often find a place in the curriculum of the Jesuit classical colleges in Quebec, which steered promising students towards law, medicine and the Church. Attitudes toward social and economic research in Quebec changed following papal social encyclicals such as *Rerum Novarum* in 1891 (an influence that ceased to dominate intellectual life in Quebec after the 1960s). The École des Hautes Études Commerciales (HEC) was established in Montreal in 1911, and its journal *Actualité •conomique* began publication in 1925. Such HEC professors as Esdras Minville (1979), Edouard Montpetit (1939–42), and François-Albert Angers were concerned with the economic independence and distinctive cultural values of French Canadian society, beyond the technical aspects of the economics that Montpetit had studied under Charles Gide at the Sciences-Po in Paris, and the concerns of French Canadian economists were shaped by the uneasy relationship of their intellectual milieu and society with the rest of Canada and North America (see Falardeau, 1944; Angers, 1961; Parizeau 1968; and the extensive oral history in Paquet, 1989 on the emergence and evolution of francophone economics in Canada).

The staples thesis

The two outstanding figures of inter-war Canadian economics, William A. Mackintosh (1923; 1939), of Queen's University, and Harold A. Innis (1930; 1940; 1956), of the University of Toronto, developed a distinctive approach to understanding Canada's economic development, the staples thesis (see also Mary Quayle Innis, 1935; Creighton, 1937; Neill, 1972). Rejecting the universal applicability of neoclassical analysis of the market determination of relative prices, the staples thesis drew on a wide range of influences (including American institutionalists, notably Veblen) to argue that a newly settled, peripheral economy could not be studied in the same way as the core economies of the world economy. The keys to analysing Canadian economic development were the geographical setting (especially regional differences and the transport routes such as the St Lawrence Valley/Great Lakes) and the characteristics of the staple commodities such as cod, fur and wheat that successively dominated an export-oriented peripheral economy. The core–periphery distinction in the staples thesis was mirrored in the structure of interwar Canadian economics discipline: Mackintosh and Innis at the leading universities in the industrial and commercial heartland of Ontario developed the dominant interpretation of Canadian development as whole, while George Brittnell (1939) and Vernon Fowke (1946) at the University of Saskatchewan focused on the locally dominant staple, wheat, and maritime economists such as Stanley Saunders (1939) were concerned with the maritime provinces as an economically backward region within Confederation. This historical and institutional approach, which had parallels in later Latin American dependency theory, received considerable attention beyond Canada: at the time of his death in 1952, Innis had been elected president of the American Economic Association, the only foreigner or non-resident ever so honoured. Except for Creighton on the merchant class, the staple literature paid little attention to class until H. Clare Pentland's Toronto dissertation on the emergence of Canada's industrial working class, finished in 1961 and published posthumously 20 years later, but largely written at the University of Toronto before Innis's death (Pentland, 1950; 1981). Canadian political economy influenced by Innis and Pentland continues to flourish in the disciplines of political science and sociology (and Innis, 1951, is influential in communications studies in Canada), but has largely disappeared from economics departments, as Canadian economics has become part of an international mainstream in which the old (or original) institutional economics, widespread in the interwar United States, has been marginalized.

Economists in and on government in Canada

The Dominion Bureau of Statistics (now Statistics Canada) became a leading centre of quantitative research under Robert Coats, for 25 years the first Dominion Statistician, an achievement recognized internationally by the election of Coats as president of the American Statistical Association in 1938 (see Coats, 1932; Keyfritz and Greenway, 1961). Economists at Queen's and McMaster Universities produced two volumes of *Statistical Contributions to Canadian Economic History* in 1931. Economists became deeply involved in other areas of government, more so than in many other countries. After exploring Canada's monetary and banking history in a long series of articles in the *Journal of the Canadian Bankers Association* (reprinted as Shortt, 1987), Adam Shortt, the first economics professor at Queen's University, came to Ottawa to head the Civil Service Commission and then to superintend the publication of numerous documents on monetary history (see Shortt, 1976). His student and successor at Queen's, Oscar D. Skelton, winner of the Hart Shaffner & Marx Prize for a study of socialism (Skelton,

1911), was Under-Secretary of State for External Affairs from 1925 until his death in 1941, an especially important position because the External Affairs portfolio was held by the prime minister, so that Skelton was the prime minister's deputy minister. Skelton in turn recruited another Queen's economics professor, W. Clifford Clark, as Deputy Minister of Finance from 1932 until Clark's death in 1952. Noteworthy anniversary surveys of the progress of economic scholarship in Canada were written by the Under-Secretary of State for External Affairs (Skelton, 1932) and the Deputy Minister of Finance (Taylor, 1960), rather than by academics, and economic research in Quebec was surveyed by a future separatist Finance Minister and Premier of Quebec (Parizeau, 1968).

The Great Depression of the 1930s, which was especially severe in the Prairie provinces, and the Second World War expanded the role of the government in the economy, and of economists in government, notably with the creation of the Bank of Canada in 1934 and of a system of national accounts during the war. The extent of popular dissatisfaction with existing economic arrangements was shown in 1935 when Alberta gave 56 of the 63 seats in its provincial legislature (and, later that year, all 15 of its seats in the federal House of Commons) to Social Credit, a movement devoted to the heterodox monetary doctrines of Major C.H. Douglas (Ascah, 1999). Keynesian macroeconomic policy offered a way to stabilize the economy and avoid depressions without recourse to central planning or inflationary Social Credit (see Brecher, 1957, on interwar monetary and fiscal discussions in Canada). William A. Mackintosh of Queen's, nominally only a wartime special assistant to Clifford Clark but de facto head of the Economic Advisory Committee, drafted the federal government's 1945 White Paper on post-war employment policy. The White Paper made a commitment to macroeconomic demand management to maintain full employment that lasted in one form or another for three decades, until in 1975 Bank of Canada Governor, Gerald Bouey, announced the bank's conversion to targeting monetary aggregates to control inflation.

Keynesian ideas reached Canada through Keynes's wartime visits to Ottawa en route to and from the United States, and especially through a group of leading civil servants including some of his former students at Cambridge (Granatstein, 1982; Owram, 1986). A.F. Wynne Plumptre, who had studied with Keynes in the late 1920s, headed the economics division of the Department of External Affairs and then was Assistant Deputy Minister of Finance (1954–65) before returning to the University of Toronto. Robert Bryce, after attending Keynes's lectures for three years while Keynes was writing *The General Theory*, was secretary to the Economic Advisory Committee during the war, Secretary to the Cabinet and Clerk of the Privy Council (1954–63), and Deputy Minister of Finance (1963–70). Keynesian macroeconomics reached Canadian academic economists through Mabel Timlin's *Keynesian Economics* (1942). Timlin, a secretary at the University of Saskatchewan, began writing that remarkable book as a Ph.D. dissertation for the University of Washington as early as 1935, before the publication of Keynes's *General Theory*, when Benjamin Higgins arrived in Saskatoon with a copy of Robert Bryce's summary of Keynes's lectures, which Bryce had presented to Hayek's seminar at the London School of Economics, where Higgins was studying. Timlin's book, her first publication at the age of 50, led to a distinguished academic career at the University of Saskatchewan, the presidency of the Canadian Political Science Association, the executive committee of the American Economic Association, and being the first woman in the humanities or social sciences elected to the Royal Society of Canada (see Alexander, 1995, on the history of women in economics in Canada). After the war, Timlin wrote a series of review articles in the *Canadian Journal of Economics and Political Science* on welfare economics

and the applicability of general equilibrium methods to public policy analysis, helping introduce Canadian economists to advances in economic theory elsewhere.

Mabel Timlin was also an early academic critic of the Bank of Canada for permitting inflation during the Korean War by failing to pursue Keynesian stabilization policy. A few years later, many Canadian economists denounced the Bank of Canada Governor, James Coyne, for being more concerned about inflation than with expansionary Keynesian policy to end a recession (Gordon, 1961). Economists at the University of Western Ontario, notably David Laidler, Michael Parkin, and Thomas Courchene, later brought to Canada monetarist arguments that the Bank of Canada should adopt a monetary policy rule designed to combat inflation rather than pursuing Keynesian discretionary stabilization policy (Courchene, 1975–80).

After the Second World War

The Canadian economics profession expanded along with the great expansion of Canadian universities that began in the 1960s and also with the growing employment of economists in the business community (Parish, 1997). Along with the growth of numbers came specialization, first between the different Canadian social sciences (previously sharing departments, conferences and a journal), then between fields within economics. Canadian economics became increasingly theoretical and econometric, and decreasingly historical, in line with changes elsewhere, especially in the United States. Since the rise of academic economics in Canada, Canadian economists had studied in the United States (for example, Innis had taken his Ph.D. at the University of Chicago, with a thesis on the Canadian Pacific Railway) and taken part in American associations, but increasingly Canadian economics, like the rest of Canadian intellectual life, became more oriented towards the United States than to Britain (except that Quebec academics were very conscious of intellectual developments in France). Post-war Canadian economists made noteworthy contributions to economics, particularly the economics of natural resources (Gordon, 1954; Scott, 1955; Easterbrook, 1959; George, 1989) and international economics (for example, the effects of trade liberalization), but while Canada's position as a resource-based, small open economy guided the choice of topics, the analytical approaches taken were shared with the international community of economists. Many outstanding economics graduates of Canadian universities pursued careers outside the country, mostly in the United States, but among these, Jacob Viner, John Kenneth Galbraith, Harry Johnson, and Robert Mundell retained close ties to Canada, paid attention to Canada's distinctive economic experience (very large capital inflows relative to GDP before 1914, a floating exchange rate from 1950 to 1962), and took part both in Canadian policy debates and in influencing the development of the Canadian economics profession (for example, Viner, 1924; Johnson, 1963; 1968).

See Also

- Galbraith, John Kenneth
- historical economics, British
- Innis, Harold Adams
- Rae, John

- Viner, Jacob

Bibliography

Alexander, J.A. 1995. Our ancestors in their successive generations. *Canadian Journal of Economics* 28, 205–24.

Angers, F.-A. 1961. Naissance de la pensée économique au Canada français. *Revue de l'histoire de l'Amérique français* 15, 204–29.

Ascah, R.L. 1999. *Politics and Public Debt: The Dominion, the Banks and Alberta's Social Credit*. Edmonton: University of Alberta Press.

Bladen, V.W. 1978. *Bladen on Bladen: Memoirs of a Political Economist*. Toronto: Scarborough College.

Boucher, P. 1664. *Histoire véritable et naturelle des moeurs et productions de la Nouvelle France*. Paris. Repr. Boucherville, Québec: Société historique de Boucherville, 1964.

Brecher, I. 1957. *Monetary and Fiscal Thought in Canada, 1919–1939*. Toronto: University of Toronto Press.

Breckenridge, R.M. 1894. *The Canadian Banking System, 1777–1890*. New York: Macmillan.

Brittnell, G.E. 1939. *The Wheat Economy*. Toronto: University of Toronto Press.

Cherriman, J.B. 1857. Review of A.A. Cournot. *Recherches sur les principes mathématiques de la théorie des richesses. Canadian Journal of Industry, Science and Art* 2, 185–94. Repr. in Dimand (1995).

Coats, R.H. 1932. Fifty years of statistical progress. In *Fifty Years Retrospect: Canada 1882–1932*. Toronto: Royal Society of Canada.

Courchene, T. 1975–80. *Money, Inflation and the Bank of Canada*, 2 vols. Montreal: C.D. Howe Research Institute.

Creighton, D. 1937. *The Commercial Empire of the St. Lawrence 1760–1850*. Toronto: Ryerson.

Dimand, R.W. 1988. An early Canadian contribution to mathematical economics: J.B. Cherriman's 1857 review of Cournot. *Canadian Journal of Economics* 21, 610–16.

Dimand, R.W. 1992. Political protest and political arithmetic on the Niagara frontier: Robert Gourlay's

Statistical Account of Upper Canada. *Brock Review* 1, 52–63.

Dimand, R.W. 1995. Cournot, Bertrand, and Cherriman. *History of Political Economy* 27, 563–78.

Dimand, R.W. 2005. David Hume on Canadian paper money: a neglected contribution. *Journal of Money, Credit and Banking* 37, 783–87.

Drummond, I.M. 1983. *Political Economy at the University of Toronto: A History of the Department, 1888–1982*. Toronto: Faculty of Arts and Science, University of Toronto.

Easterbrook, W.T. 1959. Trends in Canadian economic thought. *South Atlantic Quarterly* 58, 91–107.

Falardeau, J.-C. 1944. Problems and first experiments of social research in Quebec. *Canadian Journal of Economics and Political Science* 30, 664–94.

Fowke, V.C. 1946. *Canadian Agricultural Policy: The Historical Pattern*. Toronto: University of Toronto Press.

George, P. 1989. Classical writings in Canadian economic history: natural resources in Canadian economic development. In *Research Tools in Canadian Studies*, ed. D.C. Poff. Montréal: Association for Canadian Studies.

Goodwin, C.D.W. 1961. *Canadian Economic Thought: The Political Economy of a Developing Nation 1814–1914*. Durham, NC: Duke University Press, and London: Cambridge University Press, for the Duke University Commonwealth Studies Center.

Gordon, H.S. 1954. The economics of a common property resource: the fishery. *Journal of Political Economy* 62, 124–42.

Gordon, H.S. 1961. *The Economists Versus the Bank of Canada*. Toronto: Ryerson.

Gourlay, R. 1822. *Statistical Account of Upper Canada Compiled with a View to a Grand System of Emigration, in Connexion with a Reform of the Poor Laws*. 3 vols. London: Simpkin and Marshall. Repr. New York: Johnson Reprint Corporation, 1966; abridged with introduction by S.R. Mealing, Toronto: McClelland & Stewart, Carleton Library, 1974.

Granatstein, J.L. 1982. *The Ottawa Men: The Civil Service Mandarins 1935–1957*. Toronto: Oxford University Press.

Hamouda, O.F., Lee, C. and Mair, D. 1998. *The Economics of John Rae*. London and New York: Routledge.

Helliwell, J. 1993. What have Canadian economists been doing for the past twenty-five years? *Canadian Journal of Economics* 26, 39–54.

Innis, H.A. 1930. *The Fur Trade in Canada*. New Haven, CT: Yale University Press.

Innis, H.A. 1940. *The Cod Fisheries*. New Haven, CT: Yale University Press.

Innis, H.A. 1951. *The Bias of Communication*. Toronto: University of Toronto Press.

Innis, H.A. 1956. *Essays in Canadian Economic History*, ed. M.Q. Innis. Toronto: University of Toronto Press.

Innis, M.Q. 1935. *An Economic History of Canada*. Toronto: Ryerson. 3rd edn, 1954.

James, R.W. 1965. *John Rae, Political Economist: An Account of his Life and a Compilation of his Main Writings*, 2 vols. Toronto: University of Toronto Press.

Johnson, H.G. 1963. *The Canadian Quandary*. New York: McGraw-Hill.

Johnson, H.G. 1968. Canadian contributions to the discipline of economic science since 1945. *Canadian Journal of Economics* 1, 129–46.

Keyfritz, N. and Greenway, H.F. 1961. Robert Coats and the organization of statistics. *Canadian Journal of Economics and Political Science* 27, 313–22.

Lower, A.R.M. 1943. The development of Canadian economic ideas. In *The Spirit of American Economics*, vol. 1, ed. J.F. Normano. New York: Day.

Mackintosh, W.A. 1923. Economic factors in Canadian history. *Canadian Historical Review* 4, 12–25.

Mackintosh, W.A. 1939. *The Economic Background of Dominion-Provincial Relations*. Ottawa: J.O. Patenaude (King's Printer), Royal Commission on Dominion-Provincial Relations (Rowell-Sirois Commission). Repr. Toronto: McClelland & Stewart, Carleton Library, 1964.

Mavor, J. 1923. *My Windows on the Street of the World*, 2 vols. New York: Dent.

Minville, E. 1979. *L'économie du Québec et la science économique, with a new introduction by F.-A. Angers*. Montréal: Fides.

Montpetit, E. 1939–42. *La conquête économique*, 3 vols. Montréal: Valiquette.

Neill, R. 1972. *A New Theory of Value: The Canadian Economics of H.A. Innis*. Toronto: University of Toronto Press.

Neil, R. 1991. *A History of Canadian Economic Thought*. London and New York: Routledge.

Neill, R. and Paquet, G. 1993. L'économie hérétique: Canadian economics before 1967. *Canadian Journal of Economics* 26, 3–13.

Owram, D. 1986. *The Government Generation: Canadian Intellectuals and the State 1900–1945*. Toronto: University of Toronto Press.

Paquet, G., ed. 1989. *La pensée économique au Québec français: Témoinages et perspectives*. Montréal: Association canadienne-française pour l'avancement des sciences, Les Cahiers scientifiques 67.

Parent, E. 1846. *Importance de l'étude de l'économie politique*. Montréal: Revue Canadienne.

Parish, J. 1997. A history of business economics in Canada. *Canadian Business Economics* 5, 1–169.

Parizeau, J. 1968. La recherche en sciences économiques. In *La recherché au Canada français*, ed. L. Beaudoin. Montréal: Presses de l'Université de Montréal.

Pentland, H.C. 1950. The role of capital in Canadian economic development before 1875. *Canadian Journal of Economics and Political Science* 15, 457–74.

Pentland, H.C. 1981. *Labour and Capital in Canada 1650–1860*, ed. with an introduction by P. Phillips. Toronto: James Lorimer & Company.

Rae, J. 1834. *A Statement of Some New Principles of Political Economy, Exposing the Fallacies of the System of Free Trade, and of Some Other Doctrines Maintained in the 'Wealth of Nations'*. Boston. Repr. in James (1965, vol. 2).

Saunders, S.A. 1939. *The Economic History of the Maritime Provinces*. Ottawa: J.O. Patenaude (King's Printer), Royal Commission on Dominion-Provincial Relations (Rowell-Sirois Commission).

Scott, A.D. 1955. *The Economics of Natural Resources*. Toronto: University of Toronto Press.

Scott Thomas, J. 1889. The conditions of female labour in Ontario. *University of Toronto Studies in Political Science*, series III, ed. W.J. Ashley. Repr. in *The Proper Sphere: Woman's Place in Canadian Society*, ed. R. Cook and W. Mitchinson. Toronto: Oxford University Press, 1976.

Shortt, A. 1987. *Adam Shortt's History of Canadian Money and Banking: 1600–1880*. Toronto:

Canadian Bankers Association.

Shortt, S.E.D. 1976. *The Search for an Ideal: Six Canadian Intellectuals and their Convictions in an Age of Transition 1890–1930*. Toronto: University of Toronto Press.

Skelton, O.D. 1911. *Socialism: A Critical Analysis*. Boston: Houghton, Mifflin.

Skelton, O.D. 1932. Fifty years of political and economic science in Canada. In *Fifty Years in Retrospect: Canada, 1882–1932*. Toronto: Royal Society of Canada.

Spafford, S. 2000. *No Ordinary Academics: A History of the Department of Economics and Political Science at the University of Saskatchewan*. Toronto: University of Toronto Press.

Taylor, K.W. 1960. Economic scholarship in Canada. *Canadian Journal of Economics and Political Science* 26, 6–18.

Timlin, M.F. 1942. *Keynesian Economics*. Toronto: University of Toronto Press. Repr. with introduction by L. Tarshis and biographical note by A.E. Safarian, Toronto: McClelland & Stewart, Carleton Library, 1977.

Viner, J. 1924. *Canada's Balance of International Indebtedness 1900–1913*. Cambridge, MA: Harvard University Press.

Wakefield, E.G. 1968. *Collected Works*, ed. M.F. Lloyd-Pritchard. Glasgow and London: Collins.

How to cite this article

Dimand, Robert W. and Robin F. Neill. "Canada, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000582>
doi:10.1057/9780230226203.0187

The New Palgrave Dictionary of Economics Online

Canard, Nicolas-François (c1750 – 1833)

R.F. Hébert
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Canard, N.-F.; Cantillon, R.; Cournot, A. A.; diffusion theory of taxation; elasticities of demand and supply; Fuoco, F.; intrinsic conception of price; labour theory of value; mathematics and economics; Physiocracy; Ricardian theory of land rent; tax incidence

Article

French mathematician and economist, Canard was born in Moulins, near Vichy, around 1750, and died there in 1833. Little is known about his life other than the fact that he taught mathematics at the Ecole Centrale de Moulins. His other interests included economics, jurisprudence and meteorology.
Canard's reputation as an economist rests on his *Principes d'économie politique* (1801), a study of the incidence of taxes, which, however, has drawn more attention for its use of mathematics in economic analysis. Written in the year of Cournot's birth, the *Principes* was honoured by the French Institute, the same body that refused to recognize the later efforts of Cournot and Walras. Cournot (1877, p. i) reviled Canard's work as ‘false’, even as he admitted that it provided him an important starting point for his own researches. Other harsh critics were Francis Horner, J.B. Say, Joseph Bertrand, W.S. Jevons, and Léon Walras. Despite this rejection by French and English economists, Canard had considerable influence in Italy, where a group of writers, led most conspicuously by Francesco Fuoco, defended his method and adopted some of his ideas. In the present century, Seligman (1927, pp. 159–62) has credited Canard with the diffusion theory of taxation, Schumpeter (1954) has discounted his contribution completely, while Theocharis (1983) has defended him.
The *Principes* was influenced by Cantillon and to a lesser extent by the Physiocrats, whose doctrine Canard sought to refute. Cantillon's influence is obvious in two major areas. First, without using the terms, Canard advanced both an ‘intrinsic’ and a ‘market’ conception of price. He held that everything derives its value from the quantity of labour bestowed upon it. Different (unmeasurable) qualities of labour, however, render labour quantity an unsatisfactory measure. Therefore, one must look to the market to discover the determinants of price. Canard developed an equilibrium theory based on the

relative bargaining power of buyer and seller, which he related to need and competition. (Clearly recognizing the forces of monopoly and monopsony, he nevertheless failed to develop a bilateral monopoly model.) Second, Canard revived Cantillon's ‘three rents’, and wove them into a general equilibrium conception of the economy, which he used to trace the effects of taxation (in the process, adumbrating the Ricardian theory of land rent).

Canard argued that the imposition of a new tax produces disequilibrium and sets in motion certain equilibrating adjustments which take time to work themselves through the economy. Each person who initially pays the new tax will attempt to pass it on to the purchaser of the good, but his success in doing so depends upon the ‘forces’ encountered; or as we would say today, the tax is shifted in proportion to the elasticities of demand and supply. Canard's maxim that ‘every old tax is good, every new tax is bad’, must be judged in this context.

Selected works

1801. *Principes d'économie politique*. Paris: F. Buisson.

1802. *Moyens de perfectionner le jury*. Moulins: P. Vidalin.

1808. *Traité élémentaire du calcul des inéquations*. Paris: Crapelet.

1824. *Eléments de météorologie*. Paris: P. Persan.

1826. *Mémoire sur les causes qui produisent la stagnation et le décroissement du commerce en France, et qui tendent à anéantir l'industrie commerciale*. Paris: Delaunay.

Bibliography

Allix, E. 1920. Un précurseur de l'école mathematique: Nicholas-François Canard. *Revue d'histoire économique et sociale* 13, 38–67.

Baumol, W.J. and Goldfeld, S.M. 1968. *Precursors in Mathematical Economics: An Anthology*. London: London School of Economics and Political Science.

Bousquet, G.H. 1957. N.F. Canard: Précurseur du marginalisme. *Revue d'économie politique* 50, 232–5.

Cournot, A.A. 1877. *Revue sommaire des doctrines économiques*. Paris: Hachette.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: George Allen & Unwin.

Seligman, E.R.A. 1927. *The Shifting and Incidence of Taxation*. 5th edn, New York: Columbia University Press.

Theocharis, R.D. 1983. *Early Developments in Mathematical Economics*. 2nd edn, London: Macmillan.

How to cite this article

Hébert, R. F. "Canard, Nicolas-François (c1750–1833)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000013> doi:10.1057/9780230226203.0188

The New Palgrave Dictionary of Economics Online

Cantillon, Richard (1697 – 1734)

Vivian Walsh

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agricultural surplus; allocation theory; Cantillon, R.; intrinsic value; Law, J.; luxury; Malthus's theory of population; market price; natural price; produced and non-produced means of production; Quesnay, F.; specie-flow mechanism; surplus

Article

One Richard Cantillon, son of Philip Cantillon of Ballyheigue, County Kerry, was born in Ireland. Joseph Hone argued convincingly that this was the economist, on the ground that this Richard married Mary Ann Mahony, daughter of Lady Clare, and had with her a daughter Henrietta, who married Lord Farnham (after the death of her first husband, the Earl of Stafford). Earlier writers had estimated Cantillon's birth to have been as many as 17 years earlier, but subsequent scholars have tended to accept Hone's evidence; for example, Joseph J. Spengler (1954, p. 283) and Anita Page (1952, p. xxiv).

Richard Cantillon's close association with France has often been noted, but certain facts about his family go far to explaining this connection. An Anglo-Irish county family whose establishment in Ireland was Elizabethan or later would of course be Protestant, and the term ‘Anglo-Irish Protestant ascendancy’ would then apply strictly. But those families which came to Ireland in Norman times were Catholics, and some of these remained so for hundreds of years, in spite of dungeon, fire and sword (to use an old phrase). They often became Jacobites, and in that case Europe was for them a place of refuge and support. These were the ‘Wild Geese’, who joined foreign flags after one or other Irish rebellion failed. Often educated in Europe, their ideas were cosmopolitan, their eyes on Paris and on Rome.

The Cantillons were established in Ireland in Norman times and remained Catholics, although not always very good ones. And in later centuries they became, and long remained, devoted to the Stuart cause. Roger Cantillon of Ballyheigue married Elizabeth Stuart in 1556, and his grandson Valentine fought for Charles I at Naseby, while his great-grandson Richard was wounded at the Boyne, went to France with James II and was made a chevalier for his pains. The chevalier, clearly more notable for gallantry than for worldiness, is said to have become banker to the Stuart Pretender in Paris (Spengler,

1954, p. 284) and died insolvent, a not unpredictable fate, in 1717. Our Richard appears to have come to the rescue of his uncle's honour, paying off most of the poor old Jacobite soldier's debts, many of which, indeed, were to him. This was not the end of the family's Stuart involvement; a James Cantillon, believed by Hone to be the young future economist's brother, followed King James to France and was decorated for valour, while a nephew, Thomas, mentioned in the economist's will, was with the Irish Brigade at Lauffelt. Migration to France and beyond was in the blood of these wild geese. It should cause no surprise that our Cantillon had houses in seven European cities, or that he lived much in Paris. He was there, active in banking, between 1716 and 1720. Brilliantly anticipating the fate of John Law's scheme, he was also daring enough to profit immensely by it and, if the sources consulted by W. Stanley Jevons can be believed, ‘made a fortune of several millions in a few days, but still, distrusting Law, prudently retired to Holland’ (Jevons, 1881, p. 336). He appears again in Paris between 1729 and 1732, and seems to have had to engage in litigation with people who had lost through the collapse of Law's scheme, and blamed Cantillon for his part in this. Henry Higgs, after surveying the evidence, commented that Cantillon appeared ‘to have triumphed in the Courts over all his opponents' (Higgs, 1931, p. 373). One gets the feeling as one reads of rather ordinary people playing a game for stakes they could not afford with a master they could not match. Bankers fell like autumn leaves in Paris between 1717 and 1720, and as Higgs remarks, ‘Their losses were probably very heavy in 1720 and much of them went into Cantillon's pocket’ (1931, p. 370).

Back in London in 1734, Cantillon's luck ran out. At the height of his success and his brilliance, he was robbed and murdered, left in the flames of his townhouse in Albermarle Street, Mayfair, during the early morning of 14 May. His precious manuscripts, the Marquis de Mirabeau tells us, perished with him (Higgs, 1931, p. 382). Lady Penelope Compton, who lived opposite, tells us that ‘it burnt very feirce two houses intirely down before they could get any water’ (1931, p. 374). Given this furious blaze, the really remarkable thing to the modern reader is that even despite the primitive state of the forensic science of the day, evidence of foul play was nevertheless found. Higgs, who read the account of the subsequent trial at the Old Bailey, observes that

> it was soon evident that he had been murdered before the house was set on fire. His body was burned to ashes. The Journals for 6 June 1734 say ‘Yesterday the refiners finished their search into the ashes of the late Mr Cantillon's house, when no plate, money, or jewels had been found; an undeniable circumstance of a robbery previous to the burning of the house’. (1931, p. 374)

Cantillon's servants were tried for murder, but quickly acquitted. Suspicion then fell on a Frenchman, Joseph Denier, alias Lebane, who, we are told by Higgs, had been Cantillon's cook for 11 years, but apparently had been dismissed a little more than a week before the murder. The French chef, whether in fact guilty or not, fled to Holland and thus evaded arrest.

So it came about that we possess only one work of Cantillon's, and that in what it has been claimed is a rough French translation. Even now its early publishing history is shrouded in mystery. The *Essay on the Nature of Trade in General* (1755) is thought to have been written between 1730 and Cantillon's death, but it was not published in a complete version until 1755, and then in the French translation, claiming on the title page to have been printed in London by Fletcher Gyles, a claim reasonably disputed by Jevons

(1881, p. 341). The Marquis de Mirabeau, who revealed that the French translation was in his possession for 16 years, insisted that Cantillon ‘never intended that the work should appear in French and only translated it for a friend’ (Higgs, 1931, p. 383).

Yet, as we have seen, there would be nothing odd in someone of Cantillon's family background and personal habits *writing* a book in French and publishing it in Paris. It would appear, however, that an English original must have existed, and had been in the hands of Malachy Postlethwayt, since the latter incorporated large parts of Cantillon's *Essay* in publications beginning in 1749. The first complete English translation from the French text, which was printed alongside it, was that of Higgs in 1931. Higgs, incidentally, collated his English translation with parallel passages from Postlethwayt. In addition we now have the scholarly French edition, edited by Alfred Sauvy (1952) with a number of studies and commentaries.

Since the ‘discovery’ of Cantillon by the English-speaking world following Jevons's enthusiastic article (1881), no less than justice has been done to the merits of the *Essay* on those topics treated by Cantillon whose significance can be expressed satisfactorily in broadly neoclassical terms. Over these topics we may pass quickly. Jevons himself noted that Cantillon had presented a treatment of currency, foreign exchanges, banking and credit which, judged against the work of its period, he felt to be ‘almost beyond praise’ (Jevons, 1881, p. 342). This enthusiasm has proved infectious, and we find Joseph Spengler, 73 years later, writing that Hume, assuming he knew Cantillon's work, missed ‘the import of Cantillon's brilliant analysis (which compares favourably with Keynes's) of the response of the price structure to changes in the quantity of money’ (Spengler, 1954, p. 283). Spengler was not quite as impressed by Cantillon's treatment of the international specie flow mechanism, but Joseph A. Schumpeter found it a brilliant performance and insisted that ‘the automatic mechanism that distributes the monetary metals internationally is … almost faultlessly described’ (1954, p. 223).

It was likewise recognized as early as Jevons that Cantillon had set out the leading ideas of Adam Smith's ‘important doctrine concerning wages in different employments’ (Jevons, 1881, p. 343), and that the *Essay* contained what Jevons (somewhat exaggeratedly) called ‘an almost complete anticipation of the Malthusian theory of population’ (p. 347). Jevons, with remarkable objectivity considering his own views on the formation of value, also singled out Cantillon's treatment of ‘the whole doctrine of market value as contrasted to cost value’ (1881, p. 345). It was also customarily recognized by neoclassical scholars later than Jevons that Cantillon made important contributions to the founding of allocation theory.

To intellectual historians approaching the *Essay* in terms of the neo-Walrasian class of models for general equilibrium theory, it became natural to construe Cantillon's land and labour as given resources. In the *Essay*, however, while land is a given non-produced input, labour is a *produced commodity* available in return for subsistence. A reproduction structure thus exists, and surplus may be defined. Cantillon is largely concerned with the allocation of surplus output. This was understood by the first classical theorist to read Cantillon, François Quesnay. For all his one-sided preoccupation with *agricultural* surplus, Cantillon's French successor picked up the importance of the role of surplus, embodied it in a formal model and passed it on to later classical economists.

From a *modern* classical point of view Cantillon made several important contributions, which are not always stressed by traditional scholars. For one thing, he offered an early analysis of the respective roles of produced and non-produced inputs in a more than minimally viable commodity reproduction structure. Developing Sir William Petty's concept of a ‘par’ between land and labour, Cantillon

investigated the assumptions upon which a reduction of labour to land is legitimate. But, of course, Cantillon was reducing labour to the *produce* of land; that is, to corn. He noted that ‘as those who labour must subsist on the produce of the Land it seems that some relation might be found between the value of labour and that *of the produce* of the Land’ (Cantillon 1755b, p. 31; emphasis added). Cantillon had entered an area which even today bristles with problems, which would nowadays be described as concerning the aggregation of heterogeneous objects. Cantillon was well aware of some of them. He used a concept of subsistence, that of the ‘meanest Peasant’ (p. 39), as his unit of labour, but he was well aware that this differed all over Europe, and had apparently offered statistical material on this in the lost supplement. It is then necessary to be able to express units of more skilled labour in terms of common labour. He argues that ‘it is easily seen that the difference of price paid for daily work is based upon natural and obvious reasons’ (p. 23). Even today not much progress has been made on this problem, and highly sophisticated models blithely assume it out of existence by using a single homogeneous labour input. Land is also heterogeneous, as Cantillon was well aware; furthermore, any given kind of land can be used to grow different crops. But the analysis of heterogeneous land in the case of a single crop was not developed until Ricardo's period, and the formal analysis of the case where *different crops* are grown had to wait for Piero Sraffa (1960, pp. 74–8), and more recent work on the relations between produced and non-produced means of production, such as that of Alberto Quadrio Curzio (1980, pp. 218–40).

Leaving aside the difficulties of heterogeneous labour and heterogeneous land with multiple uses, the par is the quantity of corn needed for the subsistence of a labourer and his family during a given period. To get a consistent model, corn must be treated as the only commodity strictly necessary to the reproduction system (the only ‘basic’ in the Sraffian sense). Other outputs have to be treated as luxury goods (non-basics), so that one can accommodate the changing modes and fashions of Cantillon's prince and landowners. Cantillon in fact allowed even his meanest peasant a number of commodities: ‘the married Labourer will content himself with Bread, Cheese, Vegetables, etc., will rarely eat meat, will drink little wine or beer’ (Cantillon, 1755b, p. 37).

To accept this and retain the par, only two options seem open. The poor peasant's commodities other than bread (or other things made in the household from corn, labour, and any free ingredients) could be regarded as non-basic. Or one could construct a composite commodity, containing bread, cheese, vegetables, and so on, in fixed proportions, and use this as the unit of measurement for the par. Then, if one is to avoid the problems of different crops, one must assume that any parcel of the uniform land can produce these commodities in the standard proportions. Cantillon stressed how much even peasant consumption varied from country to country in Europe in his day. But it was not absurd to suppose, as he did, that consumption habits were fixed and traditional among the peasants of a *particular* area. None of this is meant to deny the justice of Marian Bowley's claim that ‘the “par” between land and labour could only be found under special and unrealistic assumptions’ (1973, p. 105).

In a model where corn is the only basic, or where a unit of composite commodity is always consumed in fixed proportions, one can express the surplus as corn output minus necessary corn input (seed, subsistence, feed for animals), or alternatively one can express surplus as net output of the composite commodity. Passages such as the following are then consistent with the measurement of the surplus in terms of corn (or units of the composite commodity) as required for the par:

> The Farmers have generally two thirds of the Produce of the Land, one for their costs and the support of their Assistants the other for the Profit of their Undertaking … The

> Proprietor has usually one third of the produce of his Land and on this third he maintains all the Mechanicks and others whom he employs in the City as well, frequently, as the Carriers who bring the Produce of the Country to the City. (Cantillon, 1755b, pp. 43–5)

Cantillon's treatment of surplus strongly implies that it arises only in agriculture. All those in a state, we are told more than once, subsist at the expense of the proprietors of land. There are isolated passages where he seems to be recognizing that profits (in the sense in which these reflect the existence of surplus) can arise in manufacturing. Perhaps the classic case is the description of the master hatter, who, we are told, besides his upkeep, ought also to find ‘a profit like that of the Farmer who has his third part for himself’ (1755b, p. 203). Certainly Cantillon believed (unlike the Physiocrats) that farmers kept two-thirds of the total produce, one-third representing their *profit*. But Cantillon used his term ‘undertaker’ (entrepreneur) to cover chimneysweeps and water-carriers, and Samuel Hollander is probably correct in saying that, in Cantillon, ‘profits and wages were said to have a common source in, or to be dependent upon, the property of landowners’ (1973, p. 40, n. 48). The concept of surplus throughout *industry*, and the dual concept of a rate of profit tending to equality across all sectors, including industrial sectors, would not be clearly and systematically expressed until the mature work of Adam Smith (see Walsh and Gram, 1980, pp. 40–77).

Cantillon, however, did pioneering work in developing the theory of the allocation of surplus. His model is remarkably sophisticated. It is an isolated economy – one might think of it as an island – ruled by a prince or landowner. Cantillon is perfectly clear that the prince's significant freedom of choice concerns only that part of output which constitutes the surplus he receives after providing for necessary inputs. He remarks that the prince, deciding on the use of the estate, ‘will necessarily use part of it for corn to feed the Labourers, Mechanicks, and Overseers who work for him, another part to feed the Cattle, Sheep and other Animals’ (Cantillon, 1755b, p. 59). The consumption pattern of workers is fixed, just like fodder for the animals: ‘Labourers and Mechanicks who live from day to day change their mode of living only from necessity’ (p. 63).

Cantillon is far from assuming, however, that the composition of *surplus* output is unchanging. Indeed, changes in the allocation of surplus, dictated by changes in the demands of the prince and any other landowners, are his explanation of deviations of current market prices from natural prices, or intrinsic values. In the original classics, and indeed as late as Alfred Marshall (as Pierangelo Garegnani has noted), natural prices are centres of gravitation towards which market prices *tend* (Garegnani, 1976).

This idea is clearly present in Cantillon. The prince or landlord, who is assumed to have a third of the produce of each of the farms he owns, and is mainly responsible for luxury consumption, is ‘the principal Agent in the changes which may occur in demand’ (Cantillon, 1755b, p. 63). If a few prosperous farmers engage in some luxury consumption, they will imitate the tastes of the prince. Thus changes in fashion were the leading cause of ‘the variations of demand which cause the variations of Market prices’ (p. 65). Cantillon is well aware that good or bad harvests, extraordinary consumption resulting from foreign troops, and so on, can disturb the gravitation of market prices towards natural prices, but he eliminates such accidents ‘so as not to complicate my subject, considering only a State in its natural and uniform condition’ (p. 65). This is precisely the concept of a long-period position common to all the great classical economists.

Even more surprisingly, Cantillon shows that he is quite aware that a planned economy directed by the prince, and a system of prices, can each achieve the identical allocation of surplus output – a result

whose formal proof had to wait until the 20th century, and which lay fallow after Cantillon as classical political economy developed in other respects.

Cantillon, of course, was by no means the first to make *some* kind of distinction between market and natural prices. The Schoolmen had distinguished between the price ruling at a given moment on a market and the just price, sometimes relating the latter to costs. But in Cantillon the distinction between market and natural price is an integral part of a whole economic model. The natural price, or intrinsic value of a commodity ‘is the measure of the quantity of Land and of Labour entering into its production’ (1755b, p. 29). Labour is then reduced, through the par, to subsistence units, which, as we have seen, can either be measured in corn or in quantities of a composite commodity. These intrinsic values are assumed to be invariant (p. 31). Market prices may deviate from intrinsic values following a change in demand, as we have seen, but the actions of profit-maximizing capitalist farmers will then lead to supply changes, initiating the gravitation process. If the farmers ‘have too much Wool and too little Corn for the demand, they will not fail to change from year to year the use of the land till they arrive at proportioning their production pretty well to the consumption of Inhabitants’ (pp. 61–3). Notice that since we are considering a change in *demand* for corn and wool, these goods are here being used for *luxury* consumption. Corn can be fed to servants and musicians, and wool makes fine garments. What is more, Cantillon can allow for the existence of a number of agricultural sectors producing *only* luxuries: fine wines, silks, blood horses, and so on. His model clearly implies that there is a tendency towards a long-period position in which capitalist farmers in each of these sectors would receive profits at the uniform rate of one-third of the intrinsic value of their total output. Thus the extraction of surplus, and its reflection in a uniform intersectorial rate of profit, is certainly understood by Cantillon for those sectors where capitalist production relations were firmly established in his period. It remained for Adam Smith to extend this analysis to the newly widespread phenomenon of his time, capitalist production throughout industry.

Selected works

1755a. *Essai sur la nature du commerce en général*. Traduit de l'Anglois, à Londres, chez Fletcher Gyles, dans Holborn.

1755b. *Essai sur la nature du commerce en général*. Ed. with English trans. and other material by Henry Higgs, London: Macmillan (for the Royal Economic Society), 1931.

1952. *Essai sur la nature du commerce en général*. Ed. Alfred Sauvy, Paris: Institut National d'Etudes Demographiques.

Bibliography

Bowley, M. 1973. *Studies in the History of Economic Theory before 1870*. London: Macmillan.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in Modern Capital Theory*, ed. M. Brown, K. Sato and P. Zarembka. Amsterdam: North-

Holland.

Higgs, H. 1931. Life and work of Richard Cantillon. In Cantillon (1755b).

Hollander, S. 1973. *The Economics of Adam Smith*. Toronto: University of Toronto Press.

Hone, J. 1944. Richard Cantillon, economist: biographical note. *Economic Journal* 54(April), 96–100.

Jevons, W.S. 1881. Richard Cantillon and the nationality of political economy. *Contemporary Review* 39 (January). All citations from Higgs (1931).

Page, A. 1952. La vie et l'oeuvre de Richard Cantillon (1697–1734). In Cantillon (1952).

Quadrio Curzio, A. 1980. Rent, income distribution, and orders of efficiency and rentability. In *Essays on the Theory of Joint Production*, ed. L.L. Pasinetti. New York: Columbia University Press.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Spengler, J.J. 1954. Richard Cantillon: first of the moderns. *Journal of Political Economy* 62(Pt I), 281–95; Pt II, 406–24.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Walsh, V. and Gram, H. 1980. *Classical and Neoclassical Theories of General Equilibrium, Historical Origins and Mathematical Structure*. New York: Oxford University Press.

Bibliographic addendum

For an extended treatment of Cantillon's work, see A. Brewer, *Richard Cantillon*, London: Routledge, 1992.

How to cite this article

Walsh, Vivian. "Cantillon, Richard (1697–1734)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000016> doi:10.1057/9780230226203.0189

The New Palgrave Dictionary of Economics Online

capital asset pricing model

M.J. Brennan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Two general approaches to the problem of valuing assets under uncertainty may be distinguished. The first approach relies on arbitrage arguments of one kind or another, while under the second approach equilibrium asset prices are obtained by equating endogenously determined asset demands to asset supplies, which are typically taken as exogenous. The capital asset pricing model (CAPM) is an example of an equilibrium model in which asset prices are related to the exogenous data, the tastes and endowments of investors, although the CAPM is often presented as a relative pricing model.

Keywords

arbitrage pricing theory; asset price anomalies; capital asset pricing model; consumption capital asset pricing model; international asset pricing model; intertemporal capital asset pricing model; mean-variance analysis; portfolio theory; pricing kernels; probability distributions; relative pricing models; risk aversion; separation; Tobin separation theorem; von Neumann and Morgenstern

Article

If they are to be of practical use, equilibrium asset pricing models must be parsimonious in their parameterization of asset demands. To date this parsimony has been achieved only by a choice of assumptions which leads to universal portfolio separation: this is the property that the asset demand vector of every agent can be expressed as a linear combination of a set of basis vectors which may be thought of as portfolios or mutual funds. The distinguishing feature of the set of models which is collectively known as the capital asset pricing model (CAPM) is that each of these basis portfolios can be interpreted as the solution to a particular constrained portfolio variance minimization problem.

Historical perspective

The assumption that uncertainty about future asset returns can be described in terms of a probability distribution is at least as old as Irving Fisher (1906), although Hicks (1934b) appears to have been the first to suggest that preferences for investments could be represented as preferences for the moments of the probability distributions of their returns, and to propose that, as a first approximation, preferences could be represented by indifference curves in mean-variance space. Von Neumann and Morgenstern (1947) were the first to place the theory of choice under uncertainty on a rigorous axiomatic basis.

The story of modern portfolio theory really begins, however, with Markowitz (1952; 1958) who assumed explicitly that investor preferences were defined over the mean and variance of the aggregate portfolio return, related these parameters to the portfolio composition and the parameters of the joint distribution of security returns, and for the first time applied the principles of marginal analysis to the choice of optimal portfolios.

Both Markowitz and Tobin (1958) showed that mean-variance preferences can be reconciled with the von Neumann–Morgenstern axioms if the utility function is quadratic in return or wealth. This assumption is objectionable since it implies negative marginal utility at high wealth levels. Tobin also showed, however, that mean-variance preferences could be

derived by restricting the probability distributions over which choices are made to a two-parameter family. After some initial confusion it was recognized that, since portfolio returns are weighted sums of security returns, the two-parameter family must be stable under addition, and the only member of the stable class with a finite variance is the normal distribution. Subsequently Merton (1969) and Samuelson (1970) showed that mean-variance analysis is applicable for a broad class of continuous asset price processes if the trading interval is infinitesimal.

The major part of Tobin's analysis deals with the choice between a single risky asset and cash, but he demonstrated that nothing essential is changed if there are many risky assets, for they will always be held in the same proportions and can be treated as a single composite asset. This, the first separation theorem in portfolio theory, is illustrated in Figure 1, which plots mean returns, μ, against the standard deviation, σ. In this figure the curved locus AMOVB corresponds to the set of portfolios offering the lowest standard deviation for each level of mean return: the positively sloped segment is referred to as the efficient frontier, for points along it offer the highest μ for a given σ. In the absence of any riskless investment opportunities, risk-averse mean-variance investors will select portfolios corresponding to the points at which their indifference curves in (μ, σ) space are tangent to the efficient frontier (Tobin shows that the indifference curves of risk averters will have the requisite curvature). Point C represents cash which has zero risk and return. By combining cash with the portfolio of risky assets corresponding to the tangency portfolio O, investors are able to attain the (μ, σ) combinations along the line segment CO, and all investors who find it optimal to hold cash will find it optimal to combine their cash with the same risky portfolio O: their portfolio decisions can be *separated* into the choice of the optimal combination of risky asset (O) and the choice of the cash–risky asset ratio.

Figure 1

The efficient frontier and the CAPM

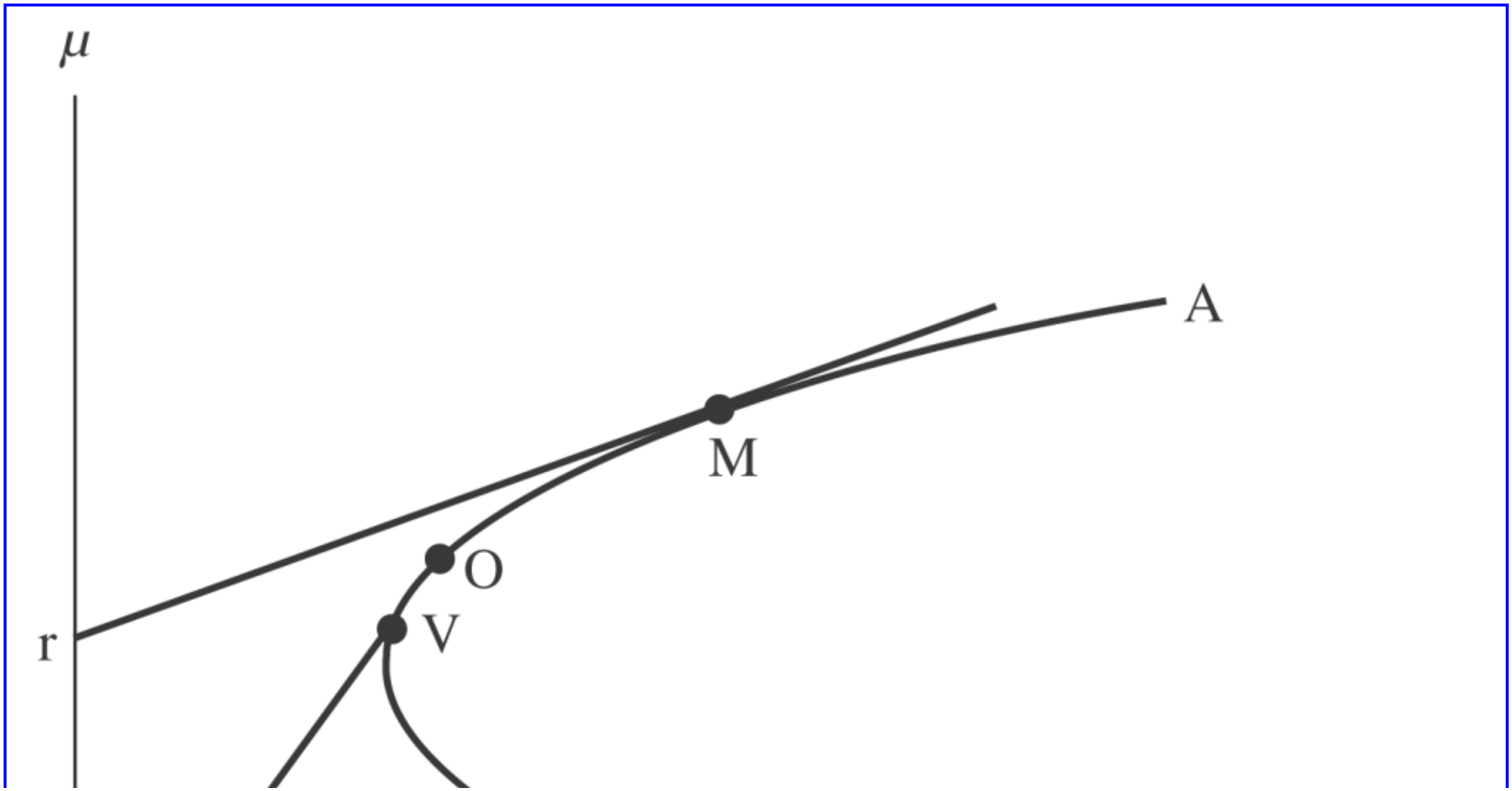

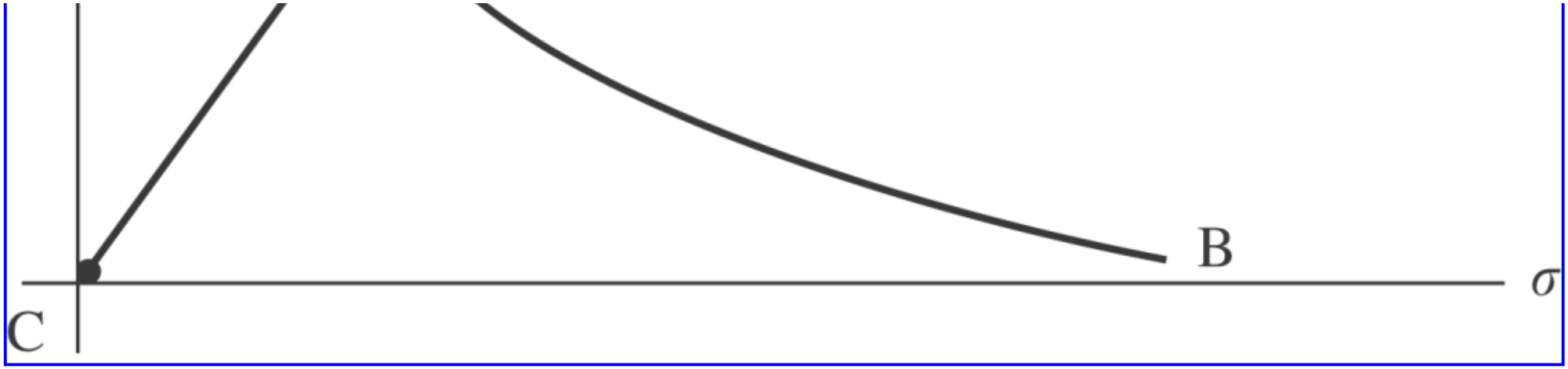

Six years elapsed before the equilibrium implications of the Tobin separation theorem were exploited by Sharpe (1964) and Lintner (1965). The reason for delay was undoubtedly the boldness of the assumption required for progress, namely, that all investors hold the same beliefs about the joint distribution of security returns. Nevertheless, this assumption of homogeneous beliefs, combined with the further assumption that all investors can borrow as well as lend at the riskless rate, *r*, leads to the powerful conclusion that all investors hold the same portfolio of risky assets, denoted by *M* in the figure. Then the only risky assets that will be held by investors in equilibrium are those contained in portfolio *M*, and *M* must be the market portfolio of all risky assets in the economy. This identification of the tangency portfolio *M* with the aggregate market portfolio is the essence of the Sharpe–Lintner CAPM.

The interest of this result derives from the restriction that it imposes on expected asset returns: the excess of μ_j, the expected return on any security *j*, over the risk-free rate *r*, must be proportional to the covariance of the security return with the return on the market portfolio, σ_{jM}:

$$\mu_j - r = \theta_M \sigma_{jM} \text{ for all } j \tag{1}$$

where θ_M is a measure of aggregate risk aversion. The intuition behind this important result is that if investors are content to hold portfolio *M*, the marginal rate of transformation between risk and return obtained by borrowing to invest in a risky security must be the same for all risky securities. Frequently the unknown risk aversion parameter, θ_M, is eliminated and the relative pricing result is obtained:

$$\mu_j - r = \beta_j(\mu_M - r) \text{ for all } j \tag{2}$$

where μ_M is the expected return on the market portfolio and $\beta_j \equiv \sigma_{jM} / \sigma_{MM}$ is the ‘beta’ coefficient, which corresponds to the slope of the regression line relating the return on the security to the return on the market portfolio.

During the first half of the 1970s extensive progress was made in relaxing the strong assumptions underlying the original model, and new separation theorems and models were obtained. At the same time, extensive empirical investigations made possible by the development of new stock-price databases found results which were interpreted as favourable to the model. The model also has an influence on practical investment management and corporate finance.

A turning point was reached with the publication of a paper by Roll (1977); this argued that the market portfolio of the theory, which includes all assets, could never be empirically

identified, and that therefore the CAPM, which simply asserts the efficiency properties of this portfolio, could never be empirically tested. This argument had substantial influence, and for some time played a major role in shifting attention away from the CAPM to the newly emerging arbitrage pricing theory (APT) of Ross (1976). However, since the early 1990s growing acceptance of the empirical importance of time variation in investment opportunities has led to a resurgence of interest in Merton's (1973) intertemporal version of the CAPM which is formally similar to the APT but is able to provide an economic interpretation of the return factors that are priced in equilibrium.

The CAPM is of great historical significance, not only because it was the first equilibrium model of asset pricing under uncertainty, but also because it showed the importance of portfolio separation for tractable equilibrium models; and, being derivable from assumptions of either quadratic utility or normal distributions, it revealed that the requisite separation properties could be obtained by restrictions either on preferences or on distributions. Cass and Stiglitz (1970) clarified the rather restrictive assumptions necessary for preference-based separation, and equilibrium models based on this have been constructed, for example, by Rubinstein (1976). Ross (1978) has identified the distributional assumptions required for separation in the absence of restrictions on preferences, and the arbitrage pricing theory is based on a generalization of his separating distributions. Chamberlain (1983) discusses spherical distributions, the subclass of separating distributions for which the expected utility is a function of the portfolio mean and variance. Both preference-based and distribution-based models of capital market equilibrium are lineal descendants of the CAPM.

A pricing kernel is a non-negative weighting function for asset returns under which the expected returns on all assets are equal to the risk-free interest rate; the kernel corresponds roughly to the marginal utility of a representative investor and the existence of a pricing kernel is a necessary and sufficient condition for arbitrage free security markets. Modern treatments of asset pricing such as Cochrane (2005) treat the general problem of asset pricing as that of specifying an appropriate pricing kernel: the CAPM specifies a class of pricing kernels that are linear in the aggregate market return.

An unfortunate consequence of the one-period nature of the CAPM was a concentration of attention on equilibrium rates of return, rather than on prices, which are the fundamental variables of interest. However, Merton (1973) placed the CAPM in an intertemporal context, and his necessary condition for equilibrium rates of return forms one cornerstone (the other being an assumption of rational expectations) for partial differential equations for asset prices which, following Cox, Ingersoll and Ross (1985), has tended to unify the pricing theories for bond and equity markets.

Formal models

While a complete asset pricing model endogenizes the riskless interest rate as well as the prices of risky securities, the CAPM adds nothing new to the theory of interest rate determination, and we shall simplify by taking the interest rate and current consumption decisions as given, concentrating our attention on portfolio decisions and the pricing of risky securities.

In considering the various versions of the CAPM we shall pay particular attention to the implied demands of investors. It will be seen that in all cases in which risks are freely traded asset demands exhibit the separation property, and even when there are restrictions on trading as in the Mayers (1972) asset pricing model, an approximate separation property obtains.

The Sharpe – Lintner model

Consider a setting in which each investor i(i=1,…,m) is endowed with a fraction $\bar{z}_{ij}$ of security j(j=1,…,n) and (a) investor utility is defined over the mean and variance of end of period wealth; (b) securities are traded in a competitive market with no taxes or transactions costs; (c) investors share homogeneous beliefs or assessments of the joint distribution of payoffs on the securities; there are no dividends; (d) there is an exogenously determined interest rate $r=R-1$ at which investors may borrow or lend without default; (e) there are no restrictions on short sales.

Then define:

- $\bar{p}_{j1}$ expected end of period value of security j;
- $\bar{p}_{j0}$ initial value of security j;
- ω_{jk} covariance between end of period value of j and k;
- $\bar{W}_i, s_i^2$ expectation and variance of end of period wealth of investor i;
- $V_i(\bar{W}_i, s_i^2)$ utility of investor i with
- $V_{i1} \equiv \partial V_i / \partial \bar{W}_i > 0,\ V_{i2} \equiv \partial V_i / \partial s_i^2 < 0$.

The investor's decision problem may be written as

$$\max_{z_{ij}} V_i(\bar{W}_i, S_i^2) \tag{3}$$

$$\text{s.t.} \quad \bar{W}_i = \sum_j z_{ij}\bar{P}_{j1} - R\sum_j (z_{ij} - \bar{z}_{ij})P_{j0} \tag{4}$$

$$S_i^2 = \sum_j \sum_k z_{ij} z_{ik} \omega_{jk}. \tag{5}$$

The first order conditions for an optimum are

$$V_{i1}(\bar{P}_{j1} - RP_{j0}) + 2V_{i2}\sum_k z_{ik}\omega_{jk} = 0, \ (j = 1, \ldots, n) \tag{6}$$

and the second conditions are satisfied by virtue of the assumption of risk aversion. Defining Ω^* as the variance covariance matrix $[\omega_{jk}]$ and using boldface type to denote vectors, the vector of fractional asset demands may be written

$$z_i = \theta_1^{-1}\mathbf{\Omega}^{*-1}(\bar{\mathbf{P}}_1 - R\mathbf{P}_0) \tag{7}$$

where $\theta_i^{-1} \equiv -V_{i1}/2V_{i2}$ is a measure of the investor's risk tolerance. Equation (7) is a statement of the Tobin separation theorem, that investor demands for risky assets differ only by a scalar multiple.

Market clearing requires that $\Sigma_i \mathbf{z}_i = \mathbf{1}$ where **1** is a vector of units. Then the equilibrium initial price vector is obtained by summing (7) over *i* and imposing the market clearing condition:

$$\mathbf{P}_0 = \frac{\mathbf{1}}{R}\left\{\bar{\mathbf{P}}_1 - \theta_m \mathbf{\Omega}^* 1\right\}$$
(8)

where $\theta_m \equiv (\Sigma_i \theta_i^{-1})^{-1}$. In this form the CAPM expresses equilibrium asset prices in terms of the exogenous variables, the distribution of end of period prices, investor risk aversion parameters and the interest rate, although it should be noted that in general the market risk aversion parameter θ_m will depend upon the endogenously determined distribution of wealth. This formulation corresponds to that of Lintner (1965) and emphasizes the one-period nature of the model and the exogeneity of the end of period prices. However, the CAPM is most often written as a necessary condition for the equilibrium rates of return, although this obscures the distinction between endogenous and exogenous variables.

In what follows we shall work with the rate of return formulation; thus define $x_{ij} \equiv z_{ij} P_{j0}$, the amount invested in security *j*; $\mu_j \equiv \bar{P}_{j1} / P_{j0} - 1$, the expected rate of return and $\sigma_{jk} \equiv \omega_{jk} / P_{j0} P_{k0}$, the covariance of the rates of return between securities *j* and *k*. Making these substitutions in (4) and (5), the first order conditions (6) become

$$V_{i1}(\mu_j - r) + 2V_{i2}\sum_k x_{ik}\sigma_{jk} = 0, \ (j = 1, \ldots, n) .$$
(9)

Then, defining Ω as the variance covariance matrix of rates of return, the vector of asset demands $\mathbf{x}_i$ may be expressed as

$$\mathbf{x}_i = \theta_i^{-1}\mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r\mathbf{1}).$$
(10)

This is an alternative statement of the Tobin separation theorem and the portfolio $\mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r\mathbf{1})$ corresponds to the point of tangency in Figure 1. This portfolio itself may be decomposed into the two portfolios $\Omega^{-1}\mu$ and $\Omega^{-1}\mathbf{1}$. The former is the solution to the problem of finding the minimum variance portfolio of risky assets with a given expected payoff, and the latter is the solution to the problem of finding the global minimum variance portfolio of risky assets; these two portfolios plot at points *O* and *V* in the figure. As Merton (1972) has shown, the whole locus may be constructed from just these two portfolios.

Let V_m denote the aggregate market value of all assets in the market portfolio and let $\mathbf{v}_m$ denote the vector of market proportions. Combining the market clearing condition $\Sigma_i \mathbf{x}_i = V_m \mathbf{v_m}$ with (10) yields

$$\boldsymbol{\mu} - r\mathbf{1} = \theta_m V_m \mathbf{\Omega} \mathbf{v}_m.$$
(11)

This form of the CAPM expresses asset risk premia as proportional to the covariances of their returns with the returns on the market portfolio; this of course is no more than the condition for the market portfolio to correspond to the tangency point in Figure 1. Equation (11) contains the market risk aversion parameter θ_m. This can be eliminated by pre-multiplying (11) by $\mathbf{v_m}$ and solving for $\theta_m = (\mu_m - r) / \sigma_m^2$, where μ_m and σ_m^2 are the expected return and variance of return on the market portfolio respectively. Then, substituting for θ_m in (11) we have the equation of the 'security market line':

$$\mu_j - r = \beta_j(\mu_m - r) \tag{12}$$

where $\beta_j \equiv \sigma_{jm} / \sigma_m^2$. In this form the CAPM is a relative pricing model which relates the risk premium on individual securities to the risk premium on the market portfolio. The proportionality factor, β_i, often referred to as the 'beta coefficient', is the coefficient from the regression of $\tilde{R}_j$, the return on security *j*, on $\tilde{R}_m$, the return on the market portfolio:

$$\tilde{R}_j = \alpha_j + \beta_j \tilde{R}_m + \tilde{e}_j \tag{13}$$

where $\tilde{e}_j$ is an orthogonal error term. Taking expectations in the market model equation (13), the asset pricing equation (12) is seen to imply the restriction $\alpha_j = (1 - \beta_j)r$. This restriction, and the existence of a positive risk premium on the market portfolio, are the major empirical predictions of the Sharpe–Lintner model. They have been the subject of extensive empirical tests.

Taxes and restrictions on riskless transactions

The absence of short sales restrictions is not critical to the Sharpe–Lintner model, since in equilibrium all investors hold the market portfolio, which does not involve short sales. The assumption is critical, however, for all the remaining models we shall consider which involve more than a single basis fund of risky securities.
Thus, following Black (1972) and Brennan (1970), assume that there are no opportunities for riskless borrowing or lending, and that each security pays predetermined dividends which are taxed in the hands of the investor at the rate $t_i (i = 1, \ldots, m)$. Denoting the dividend yield by δ_j, and assuming that investor preferences are defined over the moments of after tax wealth, the first order conditions corresponding to (9) are

$$V_{i1}(\mu_j - t_i\delta_j - \lambda_i) + 2V_{i2}\sum_k x_{ik}\sigma_{jk} = 0, \ (j = 1, \ldots, n) . \tag{14}$$

where λ_i is the Lagrange multiplier associated with the constraint that all wealth be invested in risky securities. The vector of asset demands may be written as

$$x_i = \theta_i^{-1}\mathbf{\Omega}^{-1}\boldsymbol{\mu} - (\theta_1^{-1}\lambda_1)\mathbf{\Omega}^{-1}\mathbf{1} - (\theta_1^{-1}\mathbf{t}_1)\mathbf{\Omega}^{-1}\delta.$$

(15)

Note first that if $t_i = 0$ the optimal portfolio for any preferences can be constructed from the two mutual funds $\Omega^{-1}\mu$ and $\Omega^{-1}\mathbf{1}$. Heterogeneous taxation of dividends introduces the third mutual fund, which can be interpreted as the solution to the problem of finding the minimum variance portfolio with a given total dividend. Aggregating the demand vectors, and imposing the market clearing conditions, yields an asset pricing equation which contains three utility dependent parameters, λ_m, θ_m and t_m, corresponding to the three funds in (15):

$$\boldsymbol{\mu} - \lambda_m \mathbf{1} = \theta_m V_m \boldsymbol{\Omega} \mathbf{v}_m + t_m \delta$$
(16)

t_m, the market tax rate, is a weighted average of the personal tax rates, and λ_m, the market shadow interest rate, is referred to for historical reasons as the zero beta return. When $t_m = 0$, (16) is just the condition for the market portfolio to be the tangency portfolio when the interest rate is λ_m. Thus the Black model, which does not include taxes, differs from the Sharpe–Lintner model only in leaving unspecified the relevant (shadow) riskless interest rate.

Non-marketable assets

Mayers (1972) has considered the effect of introducing an extreme form of market imperfection, namely, an absolute prohibition on trading certain assets. This is important, for a substantial part of total wealth is not held as part of well-diversified portfolios, on account either of prohibitions on trade (human capital), or of market imperfections such as transactions costs and information asymmetries. Thus let $\bar{h}_i$ denote the expected payoff on the non-marketable wealth (human capital) of investor *i*, and let σ^i_{jh} denote the covariance between the return on marketable security *j* and the human capital of investor *i*. Then the expression for $\bar{W}_i$ must be increased by $\bar{h}_i$ and the variance of end of period wealth becomes $S_i^2 = \Sigma_j \Sigma_k x_{ij} x_{ik} \sigma_{jk} + 2\Sigma_j x_{ij} \sigma^i_{jh} + \sigma$. The asset demand vector can then be written as

$$\mathbf{x}_i = \theta_i^{-1} \boldsymbol{\Omega}^{-1} (\boldsymbol{\mu} - r\mathbf{1}) - \mathbf{b}_i$$
(17)

where $\mathbf{b}_i = \boldsymbol{\Omega}^{-1} \sigma^i_h$ is the vector of coefficients from the regression of the return on human wealth on the marketable security returns. Defining $\mathbf{x}_i^e \equiv \mathbf{x}_i + \mathbf{b}_i$ as the vector of effective asset demands, we see from (17) that effective asset demands exhibit the standard separation property. This reflects the fact that, while the returns on human capital are not directly marketable, the component of the return which is linearly related to the returns on the marketable securities is indirectly marketable by appropriate offsetting positions in the marketable securities. The asset holdings of the individual may be represented as the sum of effective asset holdings $\mathbf{x}_i^e$ and an investment in the component of human wealth whose return is orthogonal to the returns on marketable assets. We refer to this as approximate portfolio separation since the first component exhibits portfolio separation, and the second component has no effect on the relative demands for marketable assets.
The Mayers model leads to an asset pricing equation which is identical to that of the Sharpe–Lintner model if the market portfolio is defined as the sum of the effective investment vectors $\mathbf{x}_i^e$.

Inflation and international asset pricing

Stochastic inflation has no effect on the foregoing results, provided that a common inflation rate can be defined for all investors and returns are restated in real terms. However, the international asset pricing models of Solnik (1974) and Stulz (1981) distinguish between nationalities precisely on the basis of their price indices, which may differ on account of either a violation of commodity price parity or differences in tastes and consumption baskets (see Adler and Dumas, 1983).

Define $\dot{\tilde{\pi}}_i$ as the inflation rate in the numeraire currency for investor *i*. Then, to a high order of approximation, which becomes exact as the time interval approaches zero, the mean and variance of real wealth can be written as

$$\bar{W}_i = \sum_j x_{ij}(\mu_j - r) + W_{0i}(1 + r - \tilde{\pi}_i + \sigma^i_{\pi\pi}) - \sum_j x_{ij}\sigma^i_{j\pi} \tag{18}$$

$$S_i^2 = \sum_j \sum_k x_{ij}x_{ik}\sigma_{jk} - 2W_{0i}\sum_j x_{ij}\sigma^i_{k\pi} + W^i_{0i}\sigma^i_{\pi\pi} \tag{19}$$

where W_{0i} is the investor's initial wealth.

The asset demand vector is then

$$x_i = \theta_i^{-1}\mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r\mathbf{1}) + \mathbf{b}_i \tag{20}$$

where $\mathbf{b}_i \equiv W_{0i}\mathbf{\Omega}^{-1}\sigma^i_x$ is the vector of coefficients from the regression at the individual's aggregate inflation risk, $W_{0i}\dot{\tilde{\pi}}_i$, on security returns. If we compare (20) with (17), it is apparent that this international asset pricing model is isomorphic to the Mayers' non-marketable wealth model with individual inflation risks playing the same role as human capital. Black (1974) has modelled segmentation in international capital markets by introducing a tax on foreign security holdings for residents of one country. This model is isomorphic to Brennan's (1970) tax model, if the foreign securities are thought of as paying dividends on which only domestic residents are taxable. Stulz (1981) extends Black's model by prohibiting negative taxes on short sales: as one might expect, this causes some indeterminacy in the pricing relations since the marginal conditions of portfolio optimality are no longer always satisfied.

Intertemporal models

Merton (1973) showed that the classical one-period CAPM can be extended to an intertemporal setting in which investors maximize the expected utility of lifetime consumption. With continuous trading and suitable restrictions on the stochastic process of asset prices, the essential mean-variance analysis is retained, the major innovation being that at each instant the individual may be represented as maximizing the expected utility of a derived utility function, defined over wealth and a set of *S* state variables describing the future investment and consumption opportunity sets. The state dependent derived utility function induces $(S + 1)$ fund separation in the risky asset portfolio, and the vector of risky asset demands may be written

$$x_1 = \theta_1^{-1}\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu} - r\mathbf{1}) - \sum_{s=1}^{S} \gamma_{is}\boldsymbol{\Omega}^{-1}\xi_s \tag{21}$$

where ξ_s is the vector of covariances of asset returns with the change in state variable S and γ_{is} depends on the utility function. Aggregation of asset demands and the imposition of the market clearing condition lead to an asset pricing equation in which asset risk premia are a linear function of covariances with aggregate wealth *and* covariances with changes in the state variables or factors that described the investment opportunity set. In the absence of prior information about the relevant state variables this model is empirically indistinguishable from the arbitrage pricing theory. Breeden (1979) showed that if consumption preferences are time separable this ‘multi-beta’ pricing model can be collapsed to a single beta measured with respect to changes in aggregate consumption, the ‘consumption’ CAPM (CCAPM), and much effort has been expended on testing this form of the model despite the difficulties of measuring consumption flows.

Campbell (1993) developed a model with recursive utility which, unlike the standard time-additive utility function defined over consumption, does not satisfy the von Neumann–Morgenstern axioms but does allow the intertemporal marginal rate of substitution to vary independently of risk aversion. This model contains elements of both the CAPM and the CCAPM in that expected returns depend on the covariances of asset returns with both consumption and the market return.

Recent empirical developments

During the 1990s renewed interest in Merton's (1973) ‘intertemporal’ CAPM (ICAPM) was generated by the empirical failures of both the CAPM and the CCAPM, the increasing evidence of time variation in investment opportunities, and the empirical success of an atheoretical three-factor model of security returns developed by Fama and French (FF) (1992; 1993) to account for high returns on small firms and the low returns on growth stocks relative to value stocks. The FF model could be interpreted as a version of either the APT or the ICAPM if no restrictions were placed on the types of factors that could enter these models. However, the factors that are important for pricing in the APT are those that explain the covariance of (one-period) returns, while the factors in the ICAPM are those that forecast future returns. Merton (1973) had suggested the interest rate as an example of an ICAPM state variable, and Nielsen and Vassalou (2006) showed formally that the only state variables that are relevant for the ICAPM are those with information about the current and future interest rate and the slope of the capital market line which is shown as *r*M in Figure 1. Brennan, Wang and Xia (2004) constructed a version of the ICAPM in which the interest rate and slope of the capital market line follow a joint Markov process, and showed that its empirical performance was at least as good as that of the FF model. Brennan and Xia (2006) used this framework to derive expressions for the prices of cash flow claims which depend explicitly on current capital market conditions as measured by the interest rate and the slope of the capital market line, as well as on the characteristics of the underlying cash flow. This implies that stock prices vary with discount rates as well as cash flow expectations, and Campbell and Vuolteenaho (2004) showed that, if market betas are decomposed into components due to changes in cash flow expectations and to changes in discount rates, then risk premia are associated primarily with the cash flow component of beta. These models attribute the low returns on growth stocks to the greater proportion of their risk arising from discount rate changes.

The classic CAPM may hold even with time variation in investment opportunities. Constantinides (1980; 1982) has identified two sets of sufficient conditions for the simple CAPM to hold with a time varying interest rate. In his models the social investment opportunity set is stationary and consists only of risky investments: stochastic variation in the interest rate then does not affect the CAPM relation if there is either demand aggregation or full Pareto efficiency of asset markets. Either condition is sufficient for prices to be determined as though there existed a single representative individual; for such an individual stochastic variation in the interest rate is irrelevant since the interest rate represents only a shadow price and not a real investment opportunity. Finally, the single period nature of the CAPM is retained if individuals behave myopically, ignoring stochastic variation in the investment opportunity set: this occurs if and only if the utility function is logarithmic.

Time variation in the distribution of asset returns can affect tests of asset pricing models even if the CAPM is true. For example, if betas and risk premia are time varying, then average returns need not be related to average betas as predicted by the CAPM even if period by period returns and betas are. Lettau and Ludvigson (2001) argued that the predictive power of the CCAPM is considerably enhanced by allowing the covariances of asset returns to depend on a measure of the aggregate consumption–wealth ratio. However, Lewellen and Nagel (2006) argued that time variation in risk premia is unlikely to be sufficient to account for the observed value anomaly.

See Also

- finance

Bibliography

Adler, M. and Dumas, B. 1983. International portfolio choice and corporation finance: a synthesis. *Journal of Finance* 38, 925–84.

Bergman, Y. 1985. Time preference and capital asset pricing models. *Journal of Financial Economics* 14, 145–59.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45, 444–55.

Black, F. 1974. International capital market equilibrium with investment barriers. *Journal of Financial Economics* 1, 337–52.

Breeden, D. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7, 265–96.

Brennan, M. 1970. Taxes, market valuation and corporate financial policy. *National Tax Journal* 23, 417–27.

Brennan, M., Wang, A. and Xia, Y. 2004. A simple model of intertemporal capital asset pricing and its implications for the Fama–French three factor model. *Journal of Finance* 59, 1743–76.

Brennan, M. and Xia, Y. 2006. Risk and valuation under an intertemporal capital asset pricing model. *Journal of Business* 79, 1–35.

Campbell, J. 1993. Intertemporal asset pricing with consumption data. *American Economic Review* 83, 487–512.

Campbell, J. and Vuolteenaho, T. 2004. Good beta, bad beta. *American Economic Review* 94, 1249–75.

Cass, D. and Stiglitz, J. 1970. The structure of investor preferences and asset returns, and separability in portfolio allocation: a contribution to the pure theory of mutual funds. *Journal of Economic Theory* 2, 122–60.

Chamberlain, G. 1983. A characterization of the distributions that imply mean-variance utility functions. *Journal of Economic Theory* 29, 185–201.

Cochrane, J. 2005. *Asset Pricing Theory*, revised edn. Princeton: Princeton University Press.

Constantinides, G. 1980. Admissible uncertainty in the intertemporal asset pricing model. *Journal of Financial Economics* 8, 71–86.

Constantinides, G. 1982. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *Journal of Business* 55, 253–67.

Cox, J., Ingersoll, J. and Ross, S. 1985. An intertemporal general equilibrium model of asset prices. *Econometrica* 5, 363–84.

Fama, E. and French, K. 1992. The cross-section of expected returns. *Journal of Finance* 47, 427–65.

Fama, E. and French, K. 1993. Common risk factors in the returns on stock and bonds. *Journal of Financial Economics* 33, 3–56.

Fisher, I. 1906. *The Nature of Capital and Income*. New York: Macmillan.

Hicks, J.R. 1934a. A note on the elasticity of supply. *Review of Economic Studies* 2(October), 31–7.

Hicks, J.R. 1934b. Application of mathematical methods to the theory of risk. *Econometrica* 2(April), 194–5.

Lettau, M. and Ludvigson, S. 2001. Resurrecting the (C)CAPM: a cross-sectional test when risk premia are time-varying. *Journal of Political Economy* 109, 1238–87.

Lewellen, J. and Nagel, S. 2006. The conditional CAPM does not explain asset pricing anomalies. *Journal of Financial Economics* 82, 289–314.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 13–37.

Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7, 77–91.

Markowitz, H. 1958. *Portfolio Selection: Efficient Diversification of Investments*. New York: Wiley.

Mayers, D. 1972. Non-marketable assets and capital market equilibrium under uncertainty. In *Studies in the Theory of Capital Markets*, ed. M. Jensen. New York: Praeger.

Merton, R. 1969. Lifetime portfolio selection under uncertainty: the continuous-time case. *Review of Economics and Statistics* 51(3), August, 247–57.

Merton, R. 1972. An analytic derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis* 7, 1851–72.

Merton, R. 1973. An intertemporal capital asset pricing model. *Econometrica* 41, 867–87.

Nielsen, L.T. and Vassalou, M. 2006. The instantaneous capital market line. *Economic Theory* 28, 651–4.

Roll, R. 1977. A critique of the asset pricing theory's test; Part I: On past and potential testability of the theory. *Journal of Financial Economics* 4, 129–76.

Ross, S. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13, 341–60.

Ross, S. 1978. Mutual fund separation in financial theory: the separating distributions. *Journal of Economic Theory* 17, 254–86.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7, 407–25.

Samuelson, P. 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances, and higher moments. *Review of Economic Studies* 37, 537–42.

Sharpe, W.F. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Solnik, B. 1974. An equilibrium model of international capital markets. *Journal of Economic Theory* 8, 500–24.

Stulz, R. 1981. A model of international asset pricing. *Journal of Financial Economics* 9, 383–406.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25(February), 65–86.

Von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd edn. Princeton: Princeton University Press.

How to cite this article

Brennan, M. J. "capital asset pricing model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000021> doi:10.1057/9780230226203.0190

The New Palgrave Dictionary of Economics Online

capital controls

Kristin J. Forbes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Capital controls can take many different forms and are broadly defined as any restrictions on the movement of capital across a country's borders. This article focuses on the debate on the merits of capital controls for emerging markets and developing economies. It describes the potential costs and benefits of capital controls, focusing on the recent empirical literature evaluating the impact of capital controls.

Keywords

Bretton Woods system; bubbles; capital account liberalization; capital controls; distortions; Dutch disease; *encaje* (Chile); fixed exchange rates; foreign direct investment; International Monetary Fund; Keynes, J. M.; Kindleberger, C.; monetary policy; Nurkse, R.; portfolio investment; risk diversification

Article

Capital controls are any restrictions on the movement of capital into or out of a country. Capital controls can take a wide variety of forms. For example, capital controls can be quantity-based or price-based, or apply to only capital inflows, only capital outflows, or all types of capital flows. Capital controls can also be directed at different types of capital flows (such as at bank loans, foreign direct investment or portfolio investment) or at different types of actors (such as at companies, banks, governments or individuals).
Most developed countries believe that the benefits from the free movement of capital across borders outweigh the costs, and therefore have very limited (if any) capital controls in place today. For emerging markets and developing economies, however, there has been a long-standing debate on the desirability of capital controls. Assessing the impact of capital controls is complicated due to a number of factors, including the various forms in which they can be structured. This article discusses the recent debate on capital controls, focusing on the theoretical arguments for and against controls and the existing empirical

evidence on their impact.

History of the debate

Throughout the 20th century, economists have regularly expressed concerns about international capital flows. For example, in the 1940s Ragnar Nurkse worried about ‘destabilizing capital flows’ and in the 1970s Charles Kindleberger described the role of capital in driving ‘manias, panics and crashes’ (see Nurkse, 1944; Kindleberger, 1978). When the world's leading economies met at Bretton Woods in 1944 to formulate rules governing the international financial system, John Maynard Keynes and other delegates debated the role of capital controls. The resulting compromise required that members of the International Monetary Fund (IMF), one of the newly created international monetary institutions, allow capital to be freely exchanged and convertible across countries for the purpose of all current account transactions, but permitted members to implement capital controls for financial account transactions. Most countries had capital controls in place at this time.

Over the following years, however, many developed countries gradually removed their capital controls, so that by the 1980s most had few controls in place. In the early and mid-1990s, many emerging markets and developing countries also began to lift their capital controls. The impact initially appeared to be positive – capital flowed into countries with liberalized capital accounts, investment and growth increased, and asset prices rose. In fact, support for lifting capital controls was so widespread that in 1996–7 leading policymakers discussed amending the rules agreed to at Bretton Woods to extend the IMF's jurisdiction to include capital movements and make capital account liberalization a goal of the IMF. In mid-1997, however, a series of financial crises started in Asia and spread across the world, appearing to disproportionately affect emerging markets that had recently liberalized their capital accounts. This series of crises sparked a reassessment of the desirability of capital controls for emerging markets and developing economies.

In a sharp sea change, many leading policymakers and economists began to support the use of capital controls for emerging markets in some circumstances, especially taxes on capital inflows. Much of this support was based on the belief that controls on capital inflows could reduce a country's vulnerability to financial crises. From 2002 to 2005, several emerging markets (such as Colombia, Russia and Venezuela) also implemented new controls on capital inflows, largely to reduce the appreciations of their currencies. Over the same period, however, several large emerging markets (such as India and China) moved in the opposite direction and lifted many of their existing controls.

Benefits and costs of capital controls

The free movement of capital across borders can have widespread benefits. Capital inflows can provide financing for high-return investment, thereby raising growth rates. Capital inflows – especially in the form of direct investment – often bring improved technology, management techniques, and access to international networks, all of which further raise productivity and growth. Capital outflows can allow domestic citizens and companies to earn higher returns and better diversify risk, thereby reducing volatility in consumption and income. Capital inflows and outflows can increase market discipline, thereby leading to a more efficient allocation of resources and higher productivity growth. Implementing

capital controls can reduce a country's ability to realise these multifaceted benefits.

On the other hand, the free movement of capital across borders can also have costs. Countries reliant on foreign financing will be more vulnerable to ‘sudden stops’ in capital inflows, which can cause financial crises and/or major currency depreciations. Large volumes of capital inflows can cause currencies to appreciate and undermine export competitiveness, causing what is often called the ‘Dutch disease’. The free movement of capital can also complicate a country's ability to pursue an independent monetary policy, especially when combined with a fixed exchange rate. Finally, capital inflows may be invested inefficiently due to a number of market distortions, thereby leading to overinvestment and bubbles that create additional challenges. Capital controls could potentially reduce these costs from the free movement of capital.

Empirical evidence on capital controls

Since capital controls can have costs and benefits, evaluating the desirability and aggregate impact of capital controls is largely an empirical question. (See Eichengreen, 2003, on the potential costs and benefits of capital controls.) Not surprisingly, an extensive literature has attempted to measure and assess the effects of capital controls.

The most studied experience with capital controls is the Chilean *encaje* – a market-based tax on capital inflows from 1991 to 1998 so structured that the magnitude of the tax decreased with the maturity of the capital flow. Chile's experience with capital controls is generally viewed positively, largely due to Chile's strong economic performance during the period the controls were in place. Empirical studies of the impact of Chile's capital controls, however, have reached several general conclusions. First, there is no evidence that the capital controls moderated the appreciation of Chile's currency (which was the primary purpose of the capital controls). Second, there is little evidence that the controls protected Chile from external shocks. Third, there is some evidence that the controls raised domestic interest rates (at least in the short term). Fourth, there is some evidence that the controls did not affect the volume of capital inflows, but did lengthen the maturity of capital inflows. Finally, the capital controls significantly raised the cost of financing for small and medium-sized firms and distorted the mechanisms by which Chilean companies procured financing. The general conclusion from this work is that Chile's strong economic performance during the 1990s resulted from sound macroeconomic and financial policies, not the capital controls, and that the capital controls had both costs and benefits. (See Forbes, 2007, for more information on this literature and the Chilean capital controls.)

A second major branch of literature examining the impact of capital controls focuses on the effects of lifting capital controls (that is, capital account liberalization). The majority of this work uses macroeconomic data, typically focusing on how capital account liberalization raises economic growth using cross-country growth regressions. Prasad et al. (2003) is a detailed survey of this literature and shows that, although several papers find a robust, positive effect of capital account liberalization on growth, other papers find no significant effect, and most papers find mixed evidence. This literature is generally read as showing weak evidence that lifting capital controls may have some positive effect on growth.

There are several explanations for the inconclusive results in this macroeconomic literature assessing the impact of capital controls. First, it is extremely difficult to measure capital account openness and to

capture the various types of capital controls in a simple measure that can be used for empirical analysis. Second, different types of capital flows and controls may have different effects on growth and other macroeconomic variables. For example, controls on portfolio investment may be more beneficial than other types of capital controls. Third, the impact of removing capital controls could depend on a range of other factors that are difficult to capture in cross-country regressions, such as a country's institutions, financial system, corporate governance or even the sequence in which different controls are removed. Fourth, capital controls can be very difficult to enforce (especially for countries with undeveloped financial markets) so the same capital control may have different degrees of effectiveness in different countries. Finally, most countries that remove their capital controls undertake simultaneously a range of reforms and undergo structural changes, so that it can be difficult to isolate the impact of removing the controls. (For additional details on the challenges in measuring the impact of capital controls, see Eichengreen, 2003; Forbes, 2006; Magud and Reinhart, 2006; and Prasad et al., 2003.)

Given these challenges in measuring the impact of capital controls, it is not surprising that the empirical literature has had difficulty documenting their effects on growth at the macroeconomic level. To put these results in perspective, however, the current status of this literature is similar to the literature in the 1980s and 1990s on how trade liberalization affects economic growth. Economists generally believe that trade openness raises growth, but most of the initial work on this topic also focused on cross-country, macroeconomic studies and reached inconclusive results. At a much earlier date, however, several papers using microeconomic data and case studies found compelling evidence that trade liberalization raises productivity and growth.

Similarly, recent work based on microeconomic data has been much more successful than the macroeconomic literature in documenting the effects of capital controls. Forbes (2006) surveys this new literature, which covers a variety of countries and periods, uses a range of approaches and methodologies, and builds on several different fields. This literature has, to date, reached five general results. First, capital controls reduce the supply of capital, raise the cost of financing, and increase financial constraints – especially for smaller firms and firms without access to international capital markets. Second, capital controls reduce market discipline in financial markets and the government, leading to a more inefficient allocation of capital and resources. Third, capital controls distort decision-making by firms and individuals as they attempt to minimize the costs of the controls, or even evade them outright. Fourth, the effects of capital controls vary across different types of firms and countries, reflecting different pre-existing economic distortions. Finally, capital controls can be difficult and costly to enforce, even in countries with sound institutions and low levels of corruption. Therefore, this series of microeconomic studies suggests that capital controls have widespread and pervasive costs, but has not yet provided significant evidence of the benefits of capital controls.

Conclusions

The debate on the effects and desirability of capital controls is likely to continue and to motivate new academic research. Most economists agree that countries should gradually lift their capital controls as they grow and develop, and that developed countries should have few (if any) capital controls in place. Most economists also believe that the free movement of capital can have widespread benefits, but that in countries with weak financial systems, poorly developed institutions, and vulnerable macroeconomies the free movement of capital can also generate distortions and increase a country's vulnerability. As a

result, emerging markets and developing countries that currently have capital controls should work to address the shortcomings in their economies as they liberalize their capital accounts. There continues to be widespread disagreement, however, on the exact sequencing of these reforms and the optimal pace of capital account liberalization for emerging markets and developing economies.

See Also

- international capital flows
- international monetary institutions
- Kindleberger, Charles P.
- Nurkse, Ragnar

Bibliography

Eichengreen, B. 2003. *Capital Flows and Crises*. Cambridge, MA: MIT Press.

Forbes, K. 2006. The microeconomic evidence on capital controls: no free lunch. In *Capital Controls and Capital Flows in Emerging Economies: Policies, Practices and Consequences*, ed. S. Edwards. Chicago: University of Chicago Press.

Forbes, K. 2007. One cost of the Chilean capital controls: increased financial constraints for smaller traded firms. *Journal of International Economics*.

Kindleberger, C. 1978. *Manias, Panics and Crashes: A History of Financial Crises*. New York: John Wiley & Sons.

Magud, N. and Reinhart, C. 2006. Capital controls: an evaluation. Working Paper No. 11973. Cambridge, MA: NBER.

Nurkse, R. 1944. *International Currency Experience: Lessons of the Interwar Experience*. Geneva: League of Nations.

Prasad, E., Rogoff, K., Wei, S.-J. and Kose, M. 2003. Effects of financial globalization on developing countries: some empirical evidence. Occasional Paper No. 220. Washington, DC: IMF.

How to cite this article

Forbes, Kristin J. "capital controls." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000025> doi:10.1057/9780230226203.0191

The New Palgrave Dictionary of Economics Online

capital gains and losses

E. Malinvaud
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

How capital gains and losses are distinct from income raises subtle and unresolved issues. Whereas national accountants measure income as the sum of the value of production and net current transfers, thus excluding stock revaluations that change the level of wealth, Hicks's definition implies that expected stock revaluations count as income. Such revaluations due to inflation benefit net debtors but mean losses for households. Irreversible environmental damage and depletion of non-renewable resources are often treated as capital loss, but great uncertainty affects the estimation of consequences, rendering the emergence of an objective methodology for economic decisions is particularly difficult.

Keywords

capital gains and losses; capital gains taxation; comprehensive definition of income; depreciation; exhaustible resources; expected and unexpected capital gains or losses; Fisher, I.; Haig–Simons definition of income; Hicks, J. R.; income, definition of; inflation; intergenerational equity; national accounting; residential real estate; uncertainty

Article

National accounting has made the definition of capital gains and losses rather precise in practice, but fundamentally their distinction from income raises quite subtle issues, about which great economists have long been wavering. Whenever it becomes important, inflation gives to some of these issues a fresh relevance. Much remains to be learned, moreover, on how capital gains affect economic behaviour and how the allocation of resources ought to deal with the capital losses resulting from current activity.

Definition

Although the reference books such as United Nations (1969) are not explicit enough about this basic

notion, national accounting systematically applies the following

$$\Delta W = Y + CT + CG - C \quad (1)$$

where ΔW is the variation of wealth between the beginning and end of the period under consideration, Y is income, CT the net capital transfer received (gifts, bequests, capital taxes and subsidies), CG the net capital gain and C consumption. The identity applies to any agent or group of agents. This identity may be taken as the *de facto* definition of net capital gains (that is, gains *minus* losses), to the extent that well-defined rules are used for the flows Y, C and CT, which appear in the current accounts, and to the extent that wealth is assumed to be unambiguously determined.

Looking carefully at the existing rules, one, however, realizes that the distinction between income and net capital gain is conventional to a large extent. It is precisely on the choice of this convention that some important questions about the definition of incomes lie.

Chapter 7 of Fisher (1906) shows that defining the concept of income was not an easy task for economists. Fisher's own preferred definition, ‘the services of capital’, may not seem quite clear, but it can be identified with consumption. This would make the whole of investment belong to capital gains, a solution that was seriously discussed by Samuelson (1961) but has hardly any advocate today. At the other extreme, the ‘comprehensive definition of income’, also called the Haig–Simons definition, was proposed by economists studying income taxes (Haig, 1921; Simons, 1938); income would be equal to the sum of consumption and wealth increase, thus leaving neither capital gains, nor capital transfers in eq. (1). One now most commonly refers to the definition introduced by Hicks (1939, p. 172), ‘A man's income is the maximum value which he can consume during a week, and still expect to be as well off at the end of the week as he was at the beginning’.

National accountants, however, measure income as the sum of the value of production and net current transfers. Production is essentially computed from physical outputs and inputs, valued at current prices and aggregated. This means that stock revaluations that explain part of the change of wealth are not incomes but capital gains or losses. Hicks's definition, on the contrary, implies that expected stock revaluations belong to income. In eq. (1) only windfalls would be true capital gains. But whether the change of value of an asset should be classified as expected or not is most often not clear. (How long in advance should it have been expected? Should an outside observer be able to make sure that the asset holder had expected the change?) The distinction between expected and unexpected capital gains or losses, however, remains essential in economic analysis.

Inflation

The most sizeable asset revaluations result from changes of the price level. When inflation is important, a good proportion of these revaluations are, moreover, expected by all agents. Their occurrence then plays a role in the determination of the equilibrium of all exchanges and economic operations, inducing

in particular high interest rates. On the other hand, the change of nominal wealth becomes of little interest in comparison with the change of real wealth; ‘real capital gains’ should then be distinguished from nominal ones. Hence, inflation perturbs the significance of normal accounting rules; new measurements are required for correct assessments of income flows (Jump, 1980).

This applies first to business accounting, in which reference to historical costs underestimates physical assets and depreciation of fixed capital, while it overestimates net returns from financial assets. This explains the search for new or alternative accounting rules that would be better suited in cases of fast inflation and would more correctly draw the line between income and capital gains or losses. This search went as far as the stage of implementation in the United Kingdom (see Walton, 1978).

At the level of the whole economy, when the rules of national accounting are applied, real capital gains and losses resulting from variations of the general level of prices are important. Typically they benefit enterprises and government, which are net debtors, whereas they mean large losses for households. When all these capital gains and losses are imputed to incomes, on the ground that they must have been expected, the current accounts of firms and government appear substantially more favourable, whereas sizeable redistribution is also found as between groups of households (see Bach and Stephenson, 1974; Babeau, 1978; Wolff, 1979).

The question has been considered whether national account practices should not be revised so as to better record true incomes in times of inflation (see Hibbert, 1982). A prerequisite is the regular production of national balance sheets. When this is done, important capital gains and losses, due for instance to booms in real estate or share prices, also appear beyond those due to changes of the general price level.

Capital gains in economic behaviour

Most econometric studies tend to neglect capital gains as flows, although wealth and indebtedness are often taken into account. The role of capital gains on the consumption behaviour of households has, however, been studied. Up to now the results have been rather inconclusive (Bhatia, 1972; Peek, 1983; Pesaran and Evans, 1984).

In all likelihood the difficulty comes from the fact that some capital gains are purely transitory, whereas most of them have some degree of permanence, but this degree varies widely from one to the other. A pure windfall is comparable to an exceptional gift; accidental losses or war damages occur once for all, whereas capital losses due to an inflation that is expected to last may appear to be as permanent as interest incomes, even sometimes as wage incomes. But to classify capital gains according to their supposed permanence is far from being an obvious operation.

Gains on the value of corporate shares have a permanent component following from the firms' policy of retaining part of their profits. This is why increases of retained earnings have been considered as likely to increase household consumption, but not as much as an increase of permanent income would, since the size of undistributed profits varies a good deal with business conditions (Feldstein and Fane, 1973; Malinvaud, 1986).

The problem becomes still more complex when capital gains are correlated with cost changes for items of household wealth. An extreme case occurs when prices of residential real estate increase: owners of houses make a capital gain, but simultaneously the cost of housing increases by the corresponding

amount; whether houses are let or used by their owners, a stimulating effect on real consumption is doubtful.

Capital losses, conservation and welfare

The existence of capital gains and losses raises a number of issues for the theory of allocation of resources, for instance what should be the taxation of capital gains (David, 1968; Green and Sheshinski, 1978), or how best to organize insurance against capital losses. But particular attention nowadays concerns the damages that economic activity causes to the environment and to reserves of exhaustible resources (Fisher, 1981).

Not all environmental effects mean capital losses; many of them are just externalities in the normal course of economic activity. But irreversible damages to the forests, the soil or even the climate must also be recognized and are usually not recorded as consumption or as inputs to production. Depletion of non-renewable reserves is similarly often treated as capital loss.

The detrimental effects of many of these losses will appear mainly in a rather distant future. Whether or not losses should be accepted – what for instance should be the optimal speed of depletion of natural resources – raises difficult questions of intergenerational equity, on which economists have uncomfortably to enter the field of social philosophy.

The problem cannot be discarded here on the ground that proper discounting makes the distant future negligible. Indeed, in the purest case, the shadow discounted price of an exhaustible resource is as high in the future as it is now, for as long as the resource will remain used (Hotelling, 1931). The remote future must then be taken into account for present decisions.

It is moreover notorious that enormous uncertainties affect the purely physical estimation of the consequences involved. Neither the effects of carbon dioxide emission on the climate, nor the existing reserves of fossil fuels, nor the future emergence of appropriate technologies for the wider use of renewable energy can be securely assessed. Under such circumstances, the emergence of an objective methodology for economic decisions is particularly difficult.

Bibliography

Babeau, A. 1978. The application of the constant price method for evaluating the transfer related to inflation: the case of French households. *Review of Income and Wealth* 24, 391–414.

Bach, G. and Stephenson, J. 1974. Inflation and the redistribution of wealth. *Review of Economics and Statistics* 56, 1–13.

Bhatia, K. 1972. Capital gains and the aggregate consumption function. *American Economic Review* 62, 866–79.

David, M. 1968. *Alternative Approaches to Capital Gains Taxation*. Washington, DC: Brookings Institution.

Feldstein, M. and Fane, G. 1973. Taxes, corporate dividend policy and personal savings: the British experience. *Review of Economics and Statistics* 55, 399–411.

Fisher, A. 1981. *Resource and Environmental Economics*. Cambridge: Cambridge University Press.

Fisher, I. 1906. *The Nature of Capital and Income*. New York: Macmillan.

Green, J. and Sheshinski, E. 1978. Optimal capital-gains taxation under limited information. *Journal of Political Economy* 86, 1143–58.

Haig, R. 1921. The concept of income: economic and legal aspects. In *The Federal Income Tax*, ed. R. Haig. New York: Columbia University Press.

Hibbert, J. 1982. *Measuring the Effects of Inflation on Income, Saving and Wealth*. Paris: OECD.

Hicks, J. 1939. *Value and Capital*. 2nd edn. Oxford: Oxford University Press, 1946.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Jump, G. 1980. Interest rates, inflation expectations, and spurious elements in measured real income and saving. *American Economic Review* 70, 990–1004.

Malinvaud, E. 1986. Pure profits as forced saving. *Scandinavian Journal of Economics* 88, 109–30.

Peek, J. 1983. Capital gains and personal saving behaviour. *Journal of Money, Credit and Banking* 15, 1–23.

Pesaran, M. and Evans, R. 1984. Inflation, capital gains and UK personal savings: 1953–81. *Economic Journal* 94, 237–57.

Samuelson, P. 1961. The evaluation of ‘social income’: capital formation and wealth. In *The Theory of Capital*, ed. F. Lutz and D. Hague. London: Macmillan.

Simons, H. 1938. *Personal Income Taxation*. Chicago: University of Chicago Press.

United Nations. 1969. *A System of National Accounts*. New York: United Nations.

Walton, J. 1978. Current cost accounting: implications for the definition and measurement of corporate income. *Review of Income and Wealth* 24, 357–90.

Wolff, E. 1979. The distributional effects of the 1969–75 inflation on holdings of household wealth in

the United States. *Review of Income and Wealth* 25, 195–208.

How to cite this article

Malinvaud, E. "capital gains and losses." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000031> doi:10.1057/9780230226203.0192

The New Palgrave Dictionary of Economics Online

capital gains taxation

William Gentry

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Capital gains taxation is the taxation of gains or losses from owning assets, usually as part of an income tax. Typically, tax systems measure capital gains or losses upon realization so that capital gains are taxed only when assets are sold. These realization-based tax rules create a number of behavioural incentives. Investors have an incentive to maximize the value of tax deferral by delaying the sale of assets. Capital gains taxes can also affect incentives for investing in risky assets. The realization-based tax rules also complicate the estimation of the revenue consequences of changing the tax rate on capital gains.

Keywords

capital gains and losses; capital gains taxation; inflation; tax base; tax incentives for saving; tax planning; taxation of corporate profits

Article

Capital gains taxation involves the taxation of changes in asset values, usually in the context of an income tax rather than as a separate tax. Under a pure income tax, these gains or losses would be measured on a periodic basis (for example, annually) and would be adjusted for inflation. However, actual tax systems tend to deviate in several important ways from this hypothetical treatment. The most important of these deviations is that capital gains are typically measured upon the realization of the gain or loss rather than under accrual accounting. The taxation of capital gains creates a wide variety of incentive issues, especially given the deviations between their tax treatment under a pure income tax and their treatment under actual tax rules.

Administrative issues

While the concept of a capital gain or loss from the ownership of an asset is straightforward, administering a tax on capital gains is a complicated part in the income tax codes of most countries. The primary difficulty arises from the challenge of measuring the size of a capital gain or loss over a specified period of time. This difficulty has led to most capital gains being taxed upon realization rather than as they accrue. The exceptions to this general rule tend to be for relatively sophisticated investors (for example, brokers) on assets that are relatively liquid and easily valued (for example, publicly traded equities). Realization-based taxation means that taxpayers keep records of the purchase price of assets, known as the basis in the asset, and calculate the gain or loss as the difference between the sales price and this basis when the asset is sold. The basis in an asset can be adjusted over time, with the most common type of adjustment being for the depreciation allowances accorded to depreciable assets.

An important issue in measuring capital gains is whether the gain is adjusted for changes in purchasing power created by inflation. Countries vary in their treatment of capital gains created by inflation. Most countries include the portion of the gain that is due to inflation in the tax base, but a few countries allow the asset's tax basis to be adjusted for inflation so that the tax base includes only the real portion of the capital gain. A pure income tax would allow for an adjustment for inflation, but such an adjustment would be part of a system that adjusted all forms of capital taxation for inflation.

In many countries, capital gains face lower marginal tax rates than other sources of income. Two rationales motivate these lower tax rates. First, policymakers may want to encourage investment in activities that generate capital gains. Second, the preferential tax rates provide an ad hoc method of adjusting tax burdens for inflation in tax systems that do not index the measurement of capital gains for inflation. These preferential rates, which can include the exemption of capital gains from income taxation, often depend on meeting a minimum holding period (for example, preferential rates apply to ‘long-term’ capital gains that are earned on assets held for longer than one year) and may apply only to specific types of assets (for example, gains on corporate stock qualify for preferential tax rates but gains on collectibles do not).

Another cumbersome feature of capital gains taxation is the specific rules dealing with how gains and losses offset each other. Typically, these loss-offset provisions limit a taxpayer's ability to use capital losses to offset other sources of income. The motivation for these limitations is that realization-based taxation provides taxpayers with the option of deferring the tax on gains but accelerating the deductions for losses through a strategy of holding on to appreciated assets but selling assets with losses.

In terms of administration, Auerbach (1991) and Auerbach and Bradford (2004) propose tax systems that allow for realization-based tax rules that would mimic the incentive and revenue effects of accrual taxation of capital gains. Such tax reforms would eliminate many of the complicated incentive effects created by current administrative rules for capital gains taxation.

Incentive effects

Taxing capital gains creates a variety of incentive, or disincentive, effects. Since taxing capital gains is typically part of a broader regime to tax capital income, the tax on capital gains can affect incentives to save. As a tax on capital income, the capital gains tax reduces the return to saving, which can have a theoretically ambiguous effect on the level of savings in the economy. Of course, since many countries provide preferential tax treatment for capital gains compared with other forms of capital income, tax policy towards capital gains often increases the return to saving by reducing the effective tax rate on

savings compared with a regime without preferential tax rates for capital gains.

Capital gains taxation can also affect incentives for taking risk. A tax on capital gains from risky investments reduces the expected return to these investments, which one might expect would discourage investment in risky assets. However, the tax on capital gains also reduces the variance in the payoffs to investing in risky assets and this reduction in variance may encourage investors to increase their investments in risky assets. The net effect of the reduction in both the expected return and the variance in returns may actually imply that the theoretical effect of a higher tax rate on capital gains is an increase in the amount of risk taking (see Domar and Musgrave, 1944). This result, however, rests on the symmetric tax treatment of gains and losses. When loss offset rules are imperfect, such that gains face a higher marginal tax rate than losses, then the theoretical predictions are much more complicated and it becomes more likely that the capital gains tax reduces the amount of risk taking in the economy because gains face a higher tax rate than losses.

The relative tax treatment of capital gains and other forms of capital income can also affect investors' portfolio choices (see Poterba, 2002; Poterba and Samwick, 2002). If capital gains face lower effective tax rates, due to either preferential tax rates or the ability to defer taxes by deferring realization of income, investors may prefer to invest in assets that are likely to generate capital gains rather than assets that generate interest or dividend income. In addition to affecting portfolio decisions, the relative tax treatment of different forms of capital income may also affect relative asset prices and expected returns (see Klein, 1999).

The realization-based feature of capital gains taxation creates several tax planning incentives (see Stiglitz, 1983). By not selling an appreciated asset, an investor can postpone paying the tax liability on the associated capital gain. This deferral of taxation reduces the discounted value of the tax (assuming that the statutory tax rate will remain constant in the future). This incentive to delay the realization of capital gains is known as the ‘lock-in’ effect since the tax liability that would be triggered by selling an asset reduces the incentive for investors to sell appreciated assets and locks them into holding assets. In the United States, the incentive to defer the realization of capital gains is compounded by tax rules that allow heirs to step-up the basis of appreciated assets that they inherit, which eliminates the income tax on capital gains on bequeathed assets.

In addition to incentives to delay the realization of capital gains, realization-based taxation also creates an incentive to accelerate the realization of capital losses since these losses can reduce taxation on other types of income (though this offset is possibly limited by loss offset rules) or capital gains on other assets (see Constantinides, 1983; Poterba, 1987; Auerbach, Burman and Siegel, 2000). This pattern of selective realization leads to the tax planning advice that taxpayers should sell their losers and hold their winners. In essence, realization-based taxation provides taxpayers with an option of whether to pay taxes, and it is typically more advantageous to exercise this option for assets that have lost value.

While most of the incentives discussed above deal with decisions made by investors, the tax treatment of capital gains can also affect the supply of different assets. For example, corporations may alter their payout policies in response to the relative tax treatment of dividends and capital gains. To the extent that capital gains face a lower effective tax rate than dividends at the investor level, corporations have an incentive to retain earnings so that investors can recognize income as capital gains rather than distribute earnings as dividends. Retaining earnings due to this tax rate differential does not necessarily imply that it leads to an increase in corporate investment. Instead of increasing investment, corporations that eschew dividends can repurchase shares as an alternative mechanism to distribute cash to shareholders

(see Green and Hollifield, 2003). These share repurchases allow investors to time their tax liabilities since the decision to sell shares back to the firm is discretionary and, for the shareholders who sell, the income associated with the transaction faces capital gains tax rates rather than dividend tax rates.

Revenue consequences

One of the more contentious issues surrounding capital gains taxation is the effects of capital gains taxes on government revenues. From the government's perspective, the incentive effects discussed above create opportunities for lost revenue. While the overall revenue effect of capital gains taxation depends on the whole myriad of incentives discussed above, much of the empirical literature on this issue has focused on the capital gains realization decisions of individuals. An important empirical issue has been separating how capital gains realizations respond to short-run fluctuations in the tax rate (or anticipated changes in tax rates) from how long-term realizations behaviour responds to the tax rate (or the ‘permanent’ response to tax changes). Auerbach (1988) examines the time series evidence in the United States and documents a large timing response of capital gains to anticipate tax rate changes but finds limited evidence of a permanent response of capital gains realizations to tax rates. Burman and Randolph (1994) examine a panel of US household taxpayers; their results also point towards a much larger transitory response than permanent response to changes in capital gains tax rates. Taken together, these studies cast doubt on the claim that reductions in capital gains tax rates can be self-financing.

See Also

- capital gains and losses
- individual retirement accounts
- taxation of corporate profits
- taxation of income

Bibliography

Auerbach, A.J. 1988. Capital gains taxation in the United States: realizations, revenue, and rhetoric. *Brookings Papers on Economic Activity* 1988(2), 595–631.

Auerbach, A.J. 1991. Retrospective capital gains taxation. *American Economic Review* 81, 167–78.

Auerbach, A.J. and Bradford, D.F. 2004. Generalized cash-flow taxation. *Journal of Public Economics* 88, 957–80.

Auerbach, A.J., Burman, L.E. and Siegel, J.M. 2000. Capital gains taxation and tax avoidance: new evidence from panel data. In *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, ed. J. Slemrod. New York: Russell Sage Foundation.

Burman, L.E. and Randolph, W.C. 1994. Measuring permanent responses to capital-gains tax changes in panel data. *American Economic Review* 84, 794–809.

Constantinides, G.M. 1983. Capital market equilibrium with personal tax. *Econometrica* 51, 611–36.

Domar, E.D. and Musgrave, R.A. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58, 387–422.

Green, R.C. and Hollifield, B. 2003. The personal-tax advantage of equity. *Journal of Financial Economics* 67, 175–216.

Klein, P. 1999. The capital gain lock-in effect and equilibrium returns. *Journal of Public Economics* 71, 355–78.

Poterba, J.M. 1987. How burdensome are capital gains taxes? Evidence from the United States. *Journal of Public Economics* 33, 157–72.

Poterba, J.M. 2002. Taxation, risk-taking, and household portfolio behavior. In *Handbook of Public Economics*, vol. 3, ed. A.J. Auerbach and M.S. Feldstein. Amsterdam: North-Holland.

Poterba, J.M. and Samwick, A.A. 2002. Taxation and household portfolio composition: U.S. evidence from the 1980s and 1990s. *Journal of Public Economics* 87, 5–38.

Stiglitz, J.E. 1983. Some aspects of the taxation of capital gains. *Journal of Public Economics* 21, 257–94.

How to cite this article

Gentry, William. "capital gains taxation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000033> doi:10.1057/9780230226203.0193

The New Palgrave Dictionary of Economics Online

capital measurement

W. Erwin Diewert and Paul Schreyer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Capital measures provide an indicator of wealth and of capital services, the contribution of assets to production. The wealth stock is the market value of assets, whereas capital services are measured in proportion to the quantity of past investment, adjusted for the relative efficiency of different vintages and capital goods in production. Although the two measures of capital are different, they are derived from a single theoretical framework whose centrepiece is a fundamental equilibrium relationship between stocks and flows of capital. Index number theory is used to guide the empirical implementation of stock and flow measures.

Keywords

aggregation; asset inflation; asset pricing; asset values; Böhm-Bawerk, E.; capital measurement; capital services; capital utilization; depreciation; fixed base (chain) indexes; geometric (declining balance) depreciation model; index numbers; interest, market rates; linear efficiency decline model; one-hoss shay efficiency model; price and quantity indexes; producer equilibrium; production function; rates of return; rental price; stocks and flows; straight line depreciation model; user cost; value ratios; Walras, L.

Article

Capital measures are constructed for two main purposes: (1) to measure wealth (the market value of assets) and (2) to analyse the role of capital in production. Because capital is durable, the value of using it in any given year is not the same as the value of owning it. There are thus different measures of capital depending on the purpose of accounting. However, these different measures should be consistently derived from a single framework.
The scope of the discussion below is restricted to fixed assets and land; we do not deal with financial or intangible assets, inventories or environmental assets.

Fundamental relations between stocks and flows of capital

In equilibrium, the *stock value* of an asset is equal to the discounted stream of future rental payments for *capital services* that the asset is expected to yield, an insight that goes at least back to Walras (1874) and Böhm-Bawerk (1888).

Let the price of an n-period old asset purchased at the beginning of period *t* be P_n^t. When prices change over time, it is necessary to distinguish between the observable rental prices for the asset at different ages in period *t* and future expected rental prices. Let f_n^t be the rental price of an n-period old asset at the beginning of period *t*. Then the fundamental equation relating the stock value of an asset, P_n^t, to the sequence of rental prices by age, $\left\{f_n^t : n = 0, 1, 2, \ldots\right\}$ is:

$$P_n^t = f_n^t + [(1+i^t)/(1+r^t)]\, f_{n+1}^t + [(1+i^t)/(1+r^t)]^2 f_{n+2}^t + [(1+i^t)/(1+r^t)]^3 f_{n+3}^t + \ldots n = 0, 1, 2, \ldots \quad (1)$$

where the i^t are expected rates of change of rental prices that are formed at the beginning of period *t*. For simplicity, it has been assumed that i^t does not depend on the asset's age. The term $1 + r^t$ is the discount factor that makes a dollar received at the beginning of period *t* equivalent to a dollar received at the beginning of period $t + 1$. Thus, the r_n^t are one-period *nominal interest rates* where the assumption has been made that the term structure of interest rates is constant. However, as the period *t* changes, r^t and i^t can change. The sequence of stock prices $\{P_n^t\}$ is not affected by general inflation provided that it affects the expected asset inflation rates i_n^t and the nominal interest rates r_n^t in a proportional manner.

The rental prices $\{f_n^t\}$ are potentially observable. In producer equilibrium, the ratio of any pair of rental prices equals the relative marginal productivity of the corresponding capital goods; see Hulten (1990).

By successive insertion for different P_n^t, (1) can be transformed into:

$$P_n^t = f_n^t + [(1+i^t)/(1+r^t)]\, P_{n+1}^t \quad (2)$$

or

$$f_n^t = (1+r^t)^{-1}[P_n^t(1+r^t) - (1+i^t)P_{n+1}^t] = [P_n^t - (1+i^t)P_{n+1}^t/(1+r^t)];\ n = 0, 1, 2, \ldots \quad (3)$$

Christensen and Jorgenson (1969) derived a version of (3) for the geometric depreciation model and end-of-period rental payments. Other variants are due to Christensen and Jorgenson (1973), Diewert (1980; 2005), Jorgenson (1989), Hulten (1990) and Diewert and Lawrence (2000).

(3) represents the *rental price* or *user cost* of an *n*-year old asset: the cost of using it during a period is given by the difference between the purchase price at the beginning of the period P_n^t and the value of the depreciated asset $(1+i^t)P_{n+1}^t = P_{n+1}^{t+1}$ at the end of period *t*. Since this offset to the initial expense will be received only by the end of the period, it must be divided by the discount factor $(1+r^t)$.

Depreciation, asset prices and user costs

Depreciation is typically defined as the decline in asset value as one goes from an asset of a particular age to the next oldest at the same point in time; see Hicks (1939), Hulten and Wykoff (1981a; 1981b), Hulten (1990), Jorgenson (1996) and Triplett (1996). Define the *depreciation rates* δ_n^t for an asset that is n periods old at the start of period t as:

$$\delta_n^t \equiv 1 - [P_{n+1}^t / P_n^t];\ n = 0, 1, 2, \ldots$$

(4)

Thus, given $\{P_n^t\}$, the period *t* sequence of $\{\delta_n^t\}$ is determined. Conversely, given $\{\delta_n^t\}$ and the price of a new asset in period *t*, $\{P_n^t\}$ is determined.

$$P_n^t = (1-\delta_0^t)(1-\delta_1^t)\ldots(1-\delta_{n-1}^t)P_0^t;\ n = 0, 1, 2, \ldots$$
(5)

With expressions (5) and (3), the sequence of user costs $\{f_n^t\}$ can be expressed in terms of the price of a new asset at the beginning of period *t*, P_0^t, and $\{\delta_n^t\}$:

$$f_n^t = (1+r^t)^{-1}(1-\delta_0^t)\ldots(1-\delta_{n-1}^t)[(1+r^t)-(1+i^t)(1-\delta_n^t)]P_0^t = (1+r^t)^{-1}[r^t+\delta_n^t(1+i^t)-i^t]P_n^t\ n = 0, 1, 2, \ldots$$
(6)

Thus, given any one of these sequences, all of the other sequences are completely determined. This means that assumptions about depreciation rates, the pattern of user costs by age or the pattern of asset prices by age cannot be made independently of each other. This point was first explicitly made by Jorgenson and Griliches (1967; 1972).

Aggregation

Asset prices are relevant for the construction of *wealth measures* of capital, and the user costs are relevant for the construction of *capital services measures*. Let there be *N* different types of assets and let the quantity of period *t* investment in asset *i* be I_i^t with a sequence of asset prices $\{P_{n,i}^t\}$. Then the value of the period *t* wealth stock is:

$$W_i^t \equiv P_{0,i}^t I_i^{t-1} + P_{1,i}^t I_i^{t-2} + P_{2,i}^t I_i^{t-3} + \ldots\ i = 1, 2, \ldots, N.$$
(7)

To turn to capital services (we set aside issues of capital utilization), the flow of services that an asset of a particular age delivers is proportional to the corresponding quantity of past investment. The value of capital services for all ages of a given asset class *i* during period *t* using the sequence of user costs $\{f_{n,i}^t\}$ is:

$$S_i^t \equiv f_{0,i}^t I_i^{t-1} + f_{1,i}^t I_i^{t-2} + f_{2,i}^t I_i^{t-3} + \ldots\ i = 1, 2, \ldots, N.$$
(8)

The value aggregates W_i^t and S_i^t can be decomposed into separate price and quantity components by standard index number methods, if each new unit of capital lasts only a finite number of periods, *L*. Define the period *t* price, user cost and quantity vectors, P_i^t, f_i^t and K_i^t respectively, as follows:

$$P_i^t \equiv [P_{0,i}^t, P_{1,i}^t, \ldots, P_{L-1,i}^t]; f_i^t \equiv [f_{0,i}^t, f_{1,i}^t, \ldots, f_{L-1,i}^t]; K_i^t \equiv [I_i^{t-1}, I_i^{t-2}, \ldots, I_i^{t-L-1}]; \; i = 1, 2, \ldots, N. \tag{9}$$

Fixed base or chain indexes may be used to decompose value ratios into price-change and quantity-change components. The values of W_i^t and S_i^t relative to their values in the preceding period, W_i^{t-1}, S_i^{t-1} have the following index number decomposition:

$$W_i^t / W_i^{t-1} = P_i^W(P_i^{t-1}, P_i^t, K_i^{t-1}, K_i^t) Q_i^W(P_i^{t-1}, P_i^t, K_i^{t-1}, K_i^t); \; i = 1, 2, \ldots, N; \tag{10}$$

$$S_i^t / S_i^{t-1} = P_i^S(f_i^{t-1}, f_i^t, K_i^{t-1}, K_i^t) Q_i^S(f_i^{t-1}, f_i^t, K_i^{t-1}, K_i^t); \; i = 1, 2, \ldots, N; \tag{11}$$

where P_i^W, P_i^S and Q_i^W, Q_i^S are bilateral price and quantity indexes respectively. In particular, Q_i^S measures the service flow of type *i* assets into production. It is thus an appropriate measure of capital input.

A functional form has to be chosen. For empirical work, Diewert (1976; 1992) has shown that the *Fisher* (1922) *ideal price and quantity indexes* appear to be ‘best’ from the axiomatic viewpoint, and can also be given strong economic justifications. The above index number approach to aggregating over vintages of capital was first suggested by Diewert and Lawrence (2000) and it is more general than the usual aggregation procedures for homogenous assets, which essentially assume that the different ages of the same capital good are perfectly substitutable so that linear aggregation techniques can be used. However, most researchers use an index number approach to form price and quantity aggregates across *different* types of assets. The overall values of the period *t* wealth stock and capital services are respectively:

$$W^t \equiv P_1^{W,t} Q_1^{W,t} + P_2^{W,t} Q_2^{W,t} + P_3^{W,t} Q_3^{W,t} + \ldots \tag{12}$$

$$S^t \equiv P_1^{S,t} Q_1^{S,t} + P_2^{S,t} Q_2^{S,t} + P_3^{S,t} Q_3^{S,t} + \ldots \tag{13}$$

Akin to (10)–(11), the value aggregates W^t and S^t can be decomposed into separate price and quantity components. Define the period t price and quantity vectors, $P^{W,t}$, $P^{S,t}$ and $K^{W,t}$, $K^{S,t}$ respectively, as follows:

$$P^{W,t} \equiv [P_1^{W,t}, P_2^{W,t}, \ldots, P_N^{W,t}]; P^{S,t} \equiv [P_1^{S,t}, P_2^{S,t}, \ldots, P_N^{S,t}]; K^{W,t} \equiv [K_1^{W,t}, K_2^{W,t}, \ldots, K_N^{W,t}]; K^{S,t} \equiv [K_1^{S,t}, K_2^{S,t}, \ldots, K_N^{S,t}]. \tag{14}$$

The values of W^t and S^t relative to their values in the preceding period, $W^{t-1}$and S^{t-1}, have the following index number decomposition:

$$W^t / W^{t-1} = P^W(P^{W,t-1}, P^{W,t}, K^{W,t-1}, K^{W,t}) Q^W(P^{W,t-1}, P^{W,t}, K^{W,t-1}, K^{W,t}); \tag{15}$$

$$S^t / S^{t-1} = P^S(P^{S,t-1}, P^{S,t}, K^{S,t-1}, K^{S,t}) Q^S(P^{S,t-1}, P^{S,t}, K^{S,t-1}, K^{S,t}) \tag{16}$$

where P^W, P^S and Q^W, Q^S are bilateral price and quantity indexes respectively. In particular, Q^S measures the overall service flow of capital into production.

Empirical determination of rates of return and asset price changes

Rates of return r^t can be based either on a balancing procedure or on market interest rates. The balancing procedure postulates that the value of capital services is equal to the value of gross operating surplus as shown by the national accounts plus the capital income of the self-employed. A rate of return is then chosen so that this equality holds. If market interest rates are used, there is still a choice between *ex ante* and *ex post* rates. Most empirical work on capital services has relied on an *ex post* balancing procedure based on Jorgenson and Griliches (1967; 1972) and Christensen and Jorgenson (1969). However, empirical problems arise when these methods yield highly volatile and sometimes negative user costs of capital. The debate has therefore continued – see Harper, Berndt and Wood (1989), Diewert (1980; 2005) and Schreyer (2006).

Possibilities for the choice of the asset inflation rates i^t include using the *ex post* asset price changes (consistent with the *ex post*, balancing procedure for rates of return), forecasting *ex ante* rates on the basis of *ex post* rates and assuming that expected asset price changes are equal to general inflation. The latter implies that the term $r^t - i^t$ in the user cost expression (6) becomes a real rate of return that is simple to measure and typically not too volatile. At the same time, the procedure may induce a bias in user costs and capital measures if the prices of different assets move with different trends and/or if asset prices move very differently from general inflation.

Empirical determination of rates of depreciation

Possibilities for determining depreciation rates include a number of approaches. First, information on market prices of assets of different age at the same point in time can be used to derive measures of depreciation. Empirical studies include Hall (1971), Beidelman (1973), Hulten and Wykoff (1981a; 1981b) and Oliner (1996). The literature has been reviewed by Hulten and Wykoff (1996) and Jorgenson (1996). The second approach uses rental prices for assets where they exist, along with information on the rate of return and on asset prices to solve the user cost eq. (6) for the rate of depreciation; for a review see Jorgenson (1996). The third approach is based on production function estimation where output is regressed on non-durable inputs and past investment. The estimated coefficients of the investment variable can be used to identify a constant rate of depreciation. Empirical studies using this approach include Epstein and Denny (1980), Pakes and Griliches (1984), Nadiri and Prucha (1996) and Doms (1996). The fourth method relies on insurance and other expert appraisals.

The fifth method makes assumptions about the relative efficiency sequence $\left\{f_n^t / f_0^t\right\}$ and the service life of assets, and then derives, via (1) and (5), a consistent measure of the rate of depreciation. For example, the *one-hoss shay* model of efficiency states that an asset yields a constant level of services throughout its useful life of *L* years: $f_n^t / f_0^t = 1$ for $n = 0, 1, 2, \ldots, L - 1$ and zero for $n = L, L + 1, L + 2, \ldots$. Another example is a *model of linear efficiency decline*, where the sequence $\left\{f_n^t / f_0^t\right\}$ is given by $f_n^t / f_0^t \equiv [L - n] / L$ for $n = 0, 1, 2, \ldots, L - 1$ and zero for $n = L, L + 1, L + 2, \ldots$.

The sixth method makes direct assumptions about the depreciation sequence $\left\{P_n^t / P_0^t\right\}$. The most frequent approaches are the *straight line depreciation model* and the *geometric* or *declining balance model*. Under the former, there is a constant amount of depreciation between every vintage: $P_n^t / P_0^t = [L - n] / L$ for $n = 0, 1, 2, \ldots, L$ and zero for $n > L$. Under the latter, which dates back to Matheson (1910), there is a constant rate of depreciation $\delta_n^t = \delta$ for $n = 0, 1, 2, \ldots$. The geometric model greatly simplifies the algebra of capital measurement and has been supported empirically through studies on used asset markets; see Hulten and Wykoff (1981a; 1981b). When there is only information on the average asset life *L*, the double declining balance method determines the rate of depreciation as $\delta = 2 / (L + 1)$.

See Also

- capital asset pricing model
- capital theory
- depreciation
- total factor productivity

Bibliography

Beidelman, C. 1973. *Valuation of Used Capital Assets*. Sarasota, FL: American Accounting Association.

Böhm-Bawerk, E. 1888. *The Positive Theory of Capital*, trans. W. Smart. New York: G.E. Stechert, 1891.

Christensen, L. and Jorgenson, D. 1969. The measurement of U.S. real capital input, 1929–1967. *Review of Income and Wealth* 15, 293–320.

Christensen, L. and Jorgenson, D. 1973. Measuring the performance of the private sector of the U.S. economy, 1929–1969. In *Measuring Economic and Social Performance*, ed. M. Moss. New York: Columbia University Press.

Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4, 115–45.

Diewert, W. 1980. Aggregation problems in the measurement of capital. In *The Measurement of Capital*, ed. D. Usher. Chicago: University of Chicago Press.

Diewert, W. 1992. Fisher ideal output, input and productivity indexes revisited. *Journal of Productivity Analysis* 3, 211–48.

Diewert, W. 2005. Issues in the measurement of capital services, depreciation, asset price changes and interest rates. In *Measuring Capital in the New Economy*, ed. C. Corrado, J. Haltiwanger and D. Sichel. Chicago: University of Chicago Press.

Diewert, W. and Lawrence, D. 2000. Progress in measuring the price and quantity of capital. In *Econometrics and the Cost of Capital: Essays in Honor of Dale W. Jorgenson*, ed. L. Lau. Cambridge, MA: MIT Press.

Doms, M. 1996. Estimating capital efficiency schedules within production functions. *Economic Inquiry* 34, 78–92.

Epstein, L. and Denny, M. 1980. Endogenous capital utilization in a short run production model: theory and empirical application. *Journal of Econometrics* 12, 189–207.

Fisher, I. 1922. *The Making of Index Numbers*. London: Macmillan.

Hall, R. 1971. The measurement of quality change from vintage price data. In *Price Indexes and Quality Change*, ed. Z. Griliches. Cambridge. MA: Harvard University Press.

Harper, M., Berndt, E. and Wood, D. 1989. Rates of return and capital aggregation using alternative rental prices. In *Technology and Capital Formation*, ed. D. Jorgenson and R. Landau. Cambridge, MA: MIT Press.

Hicks, J. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hulten, C. 1990. The measurement of capital. In *Fifty Years of Economic Measurement*, ed. E. Berndt and J. Triplett. *National Bureau of Economic Research Studies in Income and Wealth*, vol. 54. Chicago: University of Chicago Press.

Hulten, C. 1996. Capital and wealth in the revised SNA. In *The New System of National Accounts*, ed. J. Kendrick. New York: Kluwer Academic Publishers.

Hulten, C. and Wykoff, F. 1981a. The estimation of economic depreciation using vintage asset prices. *Journal of Econometrics* 15, 367–96.

Hulten, C. and Wykoff, F. 1981b. The measurement of economic depreciation. In *Depreciation, Inflation and the Taxation of Income from Capital*, ed. C. Hulten. Washington, DC: Urban Institute Press.

Hulten, C. and Wykoff, F. 1996. Issues in the measurement of economic depreciation: introductory remarks. *Economic Inquiry* 34, 10–23.

Jorgenson, D. 1989. Capital as a factor of production. In *Technology and Capital Formation*, ed. D. Jorgenson and R. Landau. Cambridge, MA: MIT Press.

Jorgenson, D. 1996. Empirical studies of depreciation. *Economic Inquiry* 34, 24–42.

Jorgenson, D. and Griliches, Z. 1967. The explanation of productivity change. *Review of Economic Studies* 34, 249–83.

Jorgenson, D. and Griliches, Z. 1972. Issues in growth accounting: a reply to Edward F. Denison. *Survey of Current Business* 52(5), Part II, 65–94.

Matheson, E. 1910. *Depreciation of Factories, Mines and Industrial Undertakings and their Valuations*, 4th edn. London: Spon.

Nadiri, M. and Prucha, I. 1996. Estimation of the depreciation rate of physical and R and D capital in the U.S. total manufacturing sector. *Economic Inquiry* 34, 43–56.

Oliner, S. 1996. New evidence on the retirement and depreciation of machine tools. *Economic Inquiry* 34, 57–77.

Pakes, A. and Griliches, Z. 1984. Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs. *Review of Economic Studies* 51, 243–62.

Schreyer, P. 2006. Measuring multi-factor productivity when rates of return are exogenous. In *Price and Productivity Measurement*, vols. 1 and 2, ed. W. Diewert et al. Vancouver: Trafford Press.

Triplett, J. 1996. Depreciation in production analysis and in income and wealth accounts: resolution of an old debate. *Economic Inquiry* 34, 93–115.

Walras, L. 1874. *Elements of Pure Economics*, trans. W. Jaffé. Homewood, IL: Richard D. Irwin, 1954.

How to cite this article

Diewert, W. E. and Paul Schreyer. "capital measurement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000583> doi:10.1057/9780230226203.0194

The New Palgrave Dictionary of Economics Online

capital theory (paradoxes)

Luigi L. Pasinetti and Roberto Scazzieri

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Capital theory has led economists to discover relationships that look ‘paradoxical’ or counter-intuitive, as they run counter widely accepted ‘parables’. The transformation of microeconomic diminishing returns relations into a macro-social law induced the mistaken belief of an inverse, monotonic relation between the interest rate (and profit rate, taken as the ‘price of capital’) and the quantity of capital per head. Subsequent work alerted economists to the difficulty of finding aggregate measures of heterogeneous capital goods, and to the possibility that a falling rate of interest (and of profit) may be associated with a decrease (not an increase) of the quantity of capital per head.

Keywords

aggregation (of capital); aggregation (production); Aghion, P.; Beccaria, C.; Bliss, C. J.; Böhm-Bawerk, E. von; Bruno, M.; Burmeister, E.; Cantillon, R.; capital as a fund; capital as a collection of physical assets; capital accumulation; capital intensity; capital measurement; capital theory; capital theory: parables; capital theory: paradoxes; Cas, A.; chain index method of measurement; Champernowne, D. G.; Clark, J. B.; classical capital theory; Cobb–Douglas functions; Cohen curiosum; Cohen, R.; construction period and utilization period, in Hicks's theory of capital; Denison, E.; diminishing returns; discontinuities in input use; distribution of income; duration parameters in production models; embodied technical change; financial and technical conceptions of capital; Fisher, F.; Garegnani, P.; Gordon, R.; Harcourt, G. C.; heterogeneity of capital goods; Hicks, J. R.; hidden reswitching; Howitt, P.; Hulton, C.; increasing returns; intertemporal equilibrium; Jevons, W. S.; Kaldor, N.; law of variable proportions; Levhari, D.; loanable funds; Locke, J.; Longfield, M.; marginal productivity theory; Morishima, M.; multifactor productivity; neoclassical production function; net capital stocks; non-monotonic relations in capital theory; Pasinetti, L.; Petty, W.; physiocracy; production matrix; production techniques; rate of interest; rate of profit; reswitching of technique; reverse capital deepening; Ricardo, D.; Robinson, J. V.; Rosser, J. B.; roundabout methods of production; Rymes, T.; Samuelson, P. A.; Say, J.-B.; Schefold, B.;

Senior, N.W.; Sheshinski, E.; Smith, A.; Solow, R.; Sraffa, P.; Starrett, D.; steady state; Stiglitz, J.; structural change; structural economic dynamics; surrogate production function; Sylos Labini, P.; temporary equilibrium; Thünen, J. H. von; variable proportions, law of; vertically integrated sectors; vertically integrated labour coefficients; vertically integrated units of productive capacity; Weitzman, M.; Wicksell, K.

Article

The idea that capital theory might lead economists to discover forms of ‘paradoxical’ behaviour emerged in the economic literature of the 1960s largely as an outcome of developments in the field of production theory (linear production models leading to enquiries into discrete and discontinuous relations). What happened in capital theory is in fact a special instance of a more general phenomenon. Economists sometimes tend to examine a large domain of economic phenomena by adapting theoretical concepts that had originally been devised for a much narrower range of special issues. The discoveries of ‘paradoxical’ relations derive from the fact that their process of generalization often turns out to be ill-conceived and misleading, if not entirely unwarranted.

For a long time, in capital theory it had been taken for granted that there is a unique, unambiguous profitability ranking of production techniques in terms of capital intensity, along the scale of variation of the rate of interest. The discovery that this is not necessarily true has induced many economists to speak of ‘paradoxes’ in the theory of capital. But the roots of apparently paradoxical behaviour are to be found, not in the economic phenomena themselves, but in the economists' tendency to rely on too simple ‘parables’ of economic behaviour.

Traditional beliefs about capital are deeply rooted in the history of economic analysis, and may be traced back to pre-classical literature. As will be shown in the next section, a long post-classical tradition was then developed on that basis. The length of ancestry might explain the survival of conventional beliefs.

The emergence of the conventional view

The notion of ‘capital’ was associated for a long time with investible wealth and its income generating power, and was largely independent of detailed consideration of the function of invested wealth in the production process. The earliest development of capital theory took place within the accounting framework of a pre-industrial economy (William Petty, John Locke, Richard Cantillon). Within this perspective, capital was often associated with purely financial transactions (lending and borrowing) and the relationship between capital and rate of interest came quite naturally to be conceived as the relationship between loanable funds and their price (see Cannan, 1929, pp. 122–53). The origin of the belief in an inverse monotonic relation between the demand for capital and the rate of interest may be traced back to this phase of the literature. The distinction between capital as a fund of purchasing power and capital as a ‘sum of values' embodied in physical assets remained in the background (see Hicks, 1977, p. 152), but was bound, in time, to generate tension ‘between the physical and financial conceptions of capital’ (Cohen and Harcourt, 2005, p. xli).

The association of capital with the process of production did not come to the fore until quite late, in spite

of certain isolated anticipations. (John Hicks, 1973, p. 12, even quotes Boccaccio's *Decameron* on the issue.) The description of capital as a stock of means of production became common with the Physiocrats and the classical economists. In this period, Cesare Beccaria (1804, ms 1771–72) presented what Jean-Baptiste Say considered to be the first analysis of ‘the true functions of productive capitals’ (Say, 1817, p. xliii). Soon after him, Adam Smith (1776) built upon the distinction between ‘productive’ capital and ‘unproductive’ consumption his theory of structural dynamics and economic growth. Finally, David Ricardo gave a definite shape to classical capital theory by examining the relationship between capital accumulation and diminishing returns and by considering in which way different proportions of capital in different industries might influence the relative exchange values of the corresponding commodities (Ricardo, 1817, ch. 1, sections 4 and 5).

Classical capital theory is characterized by lack of interest in the purely financial dimension of investment. As a result, the relation between capital accumulation and the rate of interest recedes into the background and is substituted by the relation between real capital accumulation and the rate of profit. In this way, the foundations of capital theory shifted from the exchange to the production sphere, and the demand-and-supply mechanism was confined to the process by which the rate of interest is maintained equal to the rate of profit in the long run. However, a number of economists (starting with Johann Heinrich von Thünen, Mountifort Longfield and Nassau William Senior) continued to be interested in the income-generating function of capital at the level of the individual investor, and tried to combine this approach with the emphasis on the productive function of capital that had emerged in the classical literature. The marginal productivity theory of capital and interest was developed as an answer to this conceptual problem. The essential features of that theory may be clearly seen in Thünen, who suggested a relationship between the rate of interest (i) and the rate of profit (r) quite different from the one found in Ricardo. The reason for this is that Ricardo had taken r to be fixed for the individual entrepreneur, so that equality between i and r was brought about by adjustment between the supply and demand for loans in the financial markets. Thünen suggested a different adjustment mechanism by taking r to be variable for the individual entrepreneur, so that the attainment of the long-run equality between the rate of profit and the rate of interest came to depend on the change in the physical productivity of capital as much as on adjustment in the financial markets (see Thünen, 1857).

This view is founded upon a thorough transformation of the Ricardian theory of diminishing returns and provided the logical starting point for the later marginalist theory of diminishing returns from aggregate capital. The analytical and historical process leading to this outcome is a rather complex one, and it is best understood by distinguishing two separate stages. In the first stage, the law of diminishing returns, which Ricardo considered to hold for the economy as a whole in the long run, was applied to the short-run behaviour of the individual entrepreneur. As result, the change in input proportions within any given productive unit is associated with the change in the physical productivity of capital. Here the variation of the capital stock is unlikely to influence the system of prices, so that the decrease (or increase) in the return from the last ‘increment of capital’ could be unambiguously associated with an increase (or decrease) in the physical capital stock. The second stage consisted in extending the above result to the variations in the aggregate quantity of capital available in the economic system as a whole.

The process which we have described made it possible to transform the classical conception of diminishing returns from a macro-social law into a microeconomic relation derived from the law of variable proportions. This new type of diminishing returns was then extended to the ‘macro-social’ sphere once again. As a result, it became possible to think that the rate of interest and the rate of profit

(tending to be equal to each other) are associated with the physical marginal productivity of aggregate capital: an increase in the relative quantity of capital with respect to the other inputs would be associated with lower marginal productivity of capital and thus with a lower equilibrium rate of interest and rate of profit. This inverse monotonic relation between the rate of interest (and the rate of profit) and the quantity of capital per head eventually became an established proposition of capital theory. The relevance of this relation can be seen from the attempts by William Stanley Jevons (1871), Eugen von Böhm-Bawerk (1889) and John Bates Clark (1899) to found on the theory of the marginal productivity of factors the explanation of the distribution of the social product among factors of production under competitive conditions.

Further light on the conceptual roots of the marginalist view of capital is shed by the contributions of Jevons and Böhm-Bawerk. In their theories, profit is considered as the remuneration due to the capitalist as a result of the higher productiveness of ‘indirect’ or ‘roundabout’ processes of production than of processes carried out by ‘direct’ labour only. The generalization of the marginal principles which they carried out is thus associated with the description of the production process as an essentially ‘financial’ phenomenon in which final output, like interest in financial transactions, could be considered as ‘some continuous function of the time elapsing between the expenditure of the labour and the enjoyment of the result’ (Jevons 1879, p. 266). The subsequent discovery of ‘anomalies’ in the field of capital accumulation was possible when economists started to question this extension of capital theory from the financial to the productive sphere, and when the technical structure of production was examined on its own grounds independently of the ‘financial’ aspect which might be considered to be characteristic of ‘the typical business man's viewpoint’ (Hicks, 1973, p. 12).

Anticipations of debate

It has just been shown that microeconomic diminishing returns provided the foundations for a theory of the diminishing marginal productivity of social capital, which was extended from the microeconomic sphere by way of logical analogy.

The pitfalls of this approach did not take long to emerge, as economic analysis came to grips with the full complexity of the production process. Knut Wicksell, discovered that, in the case of an economic system using heterogeneous capital goods, it might be impossible to describe diminishing returns from aggregate capital. The reason for this is that a variation in the capital stock might be associated with a change in the price system that would make it impossible to compare the quantities of capital before and after the change (see Wicksell, 1901–6, pp. 147 ff. and 180). Wicksell also recognized that this difficulty is characteristic of capital because ‘labour and land are measured each in terms of its own *technical* unit … capital, on the other hand, … is reckoned, in common parlance, as a sum of *exchange value*’ (1901–6, p. 149).

The special difficulty associated with heterogeneous capital goods is in fact an outcome of a particular procedure by which the fundamental theorems concerning capital and interest had been formulated with reference to the idealized setting of an isolated producer, and then extended by analogy to the case of the ‘social economy’. The drawbacks of this methodology were perspicaciously noted by Nicholas Kaldor in the late 1930s, when he complained that capital theory had been developed starting with ‘a … specialised set-up, with the picture of Robinson Crusoe engaged in net-making’ rather than with the

‘general case’ of ‘a society where *all* resources are produced and the services of all resources co-operate in producing further resources’ (Kaldor, 1937, p. 228.) Kaldor also noted that, had the analysis started with the ‘general case’, ‘a great deal of the controversies concerning the theory of capital might not have arisen’ (Kaldor, 1937, p. 228).

It is remarkable that so many ‘paradoxical’ results of modern capital theory were subsequently discovered precisely as an outcome of the procedure here described by Kaldor.

The stage of modern controversy was set by the consideration of two distinct problems: (*a*) the measurement of ‘aggregate capital’ in models with heterogeneous capital goods; and (*b*) the discovery that production techniques that had been excluded at lower levels of the rate of profit might ‘come back’ as the rate of profit is increased (this phenomenon is known as *reswitching of technique*).

Joan Robinson started the discussion by calling attention to the difficulties inherent in any physical measure of aggregate capital (Robinson, 1953–4). She also pointed out the ‘curiosum’ that the degree of mechanization associated with a higher wage rate and a lower rate of profit might be lower than the degree of mechanization associated with a lower wage rate and a higher rate of profit. (She attributed this ‘curiosum’ to Miss Ruth Cohen, but later on she attributed it to her reading of Sraffa's Introduction to Ricardo's *Principles*.)

Immediately afterwards, David Champernowne discovered that, in general, we must admit ‘the possibility of two stationary states each using the same items of equipment and labour force yet being shown as using different quantities of capital, merely on account of having different rates of interest and of food-wages’ (Champernowne, 1953–4, p. 119). Champernowne also admitted that the inverse monotonic relation between the rate of profit and the quantity of capital per head (as well as the inverse monotonic relation between the rate of profit and capital per unit of output) might not be generally true: ‘it is logically possible that over certain ranges of the rate of interest, a fall in interest rates and rise in food-wages will be accompanied by a *fall* in output per head and a *fall* in the quantity of capital per head’ (Champernowne, 1953–4, p. 118). Champernowne's explanation of what appeared to be perverse behaviour from the point of view of traditional theory was that changes in the interest rate can be associated with changes in the cost of capital equipment even if the physical capital stock is unchanged. As a result, perverse behaviour was attributed to pure ‘financial’ variations and a physical measure of capital was still thought to be possible. This Champernowne tried to obtain by introducing a chain index method for measuring capital (Champernowne, 1953–4, p. 125). A few years later, Joan Robinson again took up the same issue in her *Accumulation of Capital* (1956, pp. 109–10). The reason she gave for the ‘Ruth Cohen curiosum’ is quite different from the one proposed by Champernowne. She explicitly recognized that ‘financial’ factors such as a higher wage rate and a lower rate of interest would have ‘real’ consequences by influencing the actual choice of technique. (In the ‘perverse’ case a lower rate of interest would be associated with the choice of the less mechanized technique.)

When a few years later Michio Morishima attempted a multi-sectoral generalization of Joan Robinson's simple model he confirmed the possibility of a positive relationship between the rate of interest and the degree of mechanization of a technique (Morishima, 1964, p. 126). Finally John Hicks came up with the same problem when examining ‘the response of technique to price changes’ in the framework of a simple economy consisting of a consumption good ‘industry’ and a net investment good ‘industry’, and in which the same capital good is used in both industries (see Hicks, 1965, pp. 148–56).

But, in spite of all these anticipations, it must be admitted that the issue of technical reswitching was not

given an important place in economic theory before the publication of Piero Sraffa's *Production of Commodities by Means of Commodities* (1960). It is with Sraffa's work that the phenomenon took a prominent place. Sraffa was able to show that heterogeneity of capital goods and of ‘capital structures’ (different proportions between labour and intermediate inputs in the various processes of production) would normally give rise, with the variation of the rate of profit and of the unit wage, ‘to complicated patterns of price-movement with several ups and down’ (Sraffa, 1960, p. 37). This phenomenon would in turn bring about changes in the ‘quantity of capital’ that are not generally related to the rate of profit in a monotonic way. Reswitching of technique and reverse capital deepening are thus derived from a general property of production models with heterogeneous capital goods. (See reswitching of technique and reverse capital deepening.)

Neoclassical parables and the capital controversy

Following the publication of Sraffa's book, a lively debate on capital theory suddenly flared up in the 1960s, and the way it did is itself an interesting event.

It has already been pointed out that, when propositions derived from individual behaviour are applied to the more complex case of the ‘social economy’, the extension is admittedly possible on condition that the social economy has a number of special features making it identical, from the analytical point of view, to the case of the isolated individual. To test these features, the social economy is often described in terms of a ‘parable’ in which those particular conditions are satisfied. This ‘parable’, though unrealistic, is taken to be useful, from an heuristic or a persuasive point of view.

In this vein Paul Samuelson attempted to construct a ‘surrogate production function’ by analogy with microeconomic behaviour (Samuelson, 1962). His work can be considered as the first explicit attempt to get rid of the complexities of an economic system with heterogeneous capital goods by constructing a model in which that system is described in terms of an ‘aggregate parable’ with physically homogeneous capital. After introducing the assumption that ‘the same proportion of inputs is used in the consumption-goods and [capital-] goods industries’ (Samuelson, 1962, pp. 196–7), Samuelson was able to prove that ‘the Surrogate (Homogeneous) Capital … gives exactly the same result as does the shifting collection of diverse capital goods in our more realistic model’ (1962, p. 201). In particular, ‘the relations among w, r, and Q/L that prevail for [the] quasi-realistic complete system of heterogeneous capital goods’ could ‘be shown to have the same formal properties as does the parable system’ (1962, p. 203). This result was taken to be a justification for using the surrogate production function ‘as a useful summarizing device’ (1962, p. 203). In fact, Pierangelo Garegnani, who was present at a discussion of a draft of Samuelson's paper, did point out that Samuelson's result is crucially dependent on the assumption of equal proportions of inputs (see Garegnani, 1970). Samuelson acknowledged Garegnani's criticism in a footnote to his paper and admitted that it would be a ‘false conjecture’ to think that the ‘extreme assumption of equi-proportional inputs in the consumption and machine trades could be lightened and still leave one with many of the surrogate propositions’ (Samuelson, 1962, p. 202n). But Samuelson and various other economists continued to look for conditions that would ensure a monotonic relation between the rate of profit and the choice of technique even in presence of a nonlinear relation between w and r.

The outcome appeared a few years later. David Levhari, a Ph.D. student of Samuelson's, in his

dissertation and then in a paper for the *Quarterly Journal of Economics*, claimed he had proved that reswitching of the whole production matrix would be impossible if this matrix is of the ‘irreducible’ or ‘indecomposable’ type (Levhari, 1965). This property – Levhari claimed – would exclude reswitching and thus make it possible to extend the use of a ‘surrogate production function’ to the nonlinear case with production technologies for basic commodities.

However, Levhari's theorem was disproved by Luigi Pasinetti in a paper at the Rome First World Congress of the Econometric Society in 1965. Pasinetti's final draft of his paper was published in the November 1966 issue of the *Quarterly Journal of Economics* (Pasinetti, 1966) together with papers written in the meantime by David Levhari and Paul Samuelson (1966), Paul Samuelson (1966), Michio Morishima (1966), Michael Bruno, Edwin Burmeister and Eytan Sheshinski (1966) and Pierangelo Garegnani (1966). This set of papers was called by the journal editor ‘Paradoxes in Capital Theory: A Symposium’, thereby originating the term. Paul Samuelson concluded the discussion with a ‘Summing up’ in which he admitted that ‘the simple tale told by Jevons, Böhm-Bawerk, Wicksell, and other neoclassical writers’, according to which a falling rate of interest is unambiguously associated with the choice of more capital-intensive techniques, ‘cannot be universally valid’ (Samuelson, 1966, p. 568).

The various contributions to this discussion showed that reswitching might occur both with ‘decomposable’ and ‘indecomposable’ technology matrices. This result was proved in different ways by Pasinetti (1965; 1966), Morishima (1966), Bruno, Burmeister and Sheshinski (1966) and Garegnani (1966). Samuelson stated in his summing up that ‘reswitching is a logical possibility in any technology, indecomposable or decomposable’ (1966, p. 582). He then called attention to the associated phenomenon of reverse capital deepening and concluded that ‘there often turns out to be no unambiguous way of characterizing different processes as more “capital-intensive”, more “mechanized”, more “roundabout’” (1966, p. 582).

Although the logical possibility of reswitching was admitted by all participants in the discussion, Bruno, Burmeister and Sheshinski raised doubts as to its empirical relevance: ‘there is an open empirical question as to whether or not reswitching is likely to be observed in an actual economy for reasonable changes in the interest rate’ (Bruno, Burmeister and Sheshinski, 1966, p. 545n). The same doubt was expressed in Samuelson's summing up (Samuelson, 1966, p. 582). Bruno, Burmeister and Sheshinski also mentioned a theorem, which they attributed to Martin Weitzman and Robert Solow, according to which reswitching of technique may be excluded, in a model with heterogeneous capital goods, provided at least one capital good is produced by ‘a smooth neoclassical production function’, if ‘labour and each good are inputs in one or more of the goods produced neoclassically’ (Bruno, Burmeister and Sheshinski, 1966, p. 546). This theorem is based on the idea that ‘setting the various marginal productivity conditions and supposing that at two different rates of interest the same set of input–output coefficients holds, the proof follows by contradiction’ (Bruno, Burmeister and Sheshinski, 1966, p. 546).

It is worth noting that Weitzman–Solow's theorem is simply a consequence of the idea that, in the case of a commodity produced by a neoclassical production function, each set of input–output coefficients ought to be associated in equilibrium with a one-to-one correspondence between marginal productivity ratios and input price ratios. No ratio between marginal productivities would be associated with more than one set of input prices, and this is taken to exclude the possibility that the same technique be chosen at alternative rates of interest, and thus at different price systems. The Weitzman–Solow theorem is at the origin of a line of arguments that has been followed up by a number of other authors, such as David

Starrett (1969) and Joseph Stiglitz (1973). These authors have pursued the idea that ‘enough’ substitutability, by ensuring the smoothness of the production function, is sufficient to exclude reswitching of technique. However, non-reswitching theorems of this type involve that, for each technique of production, the capital stock may be measured either in physical terms or at given prices. For in a model with heterogeneous capital goods, if we allow prices to vary when the rate of interest or the unit wage are changed, there is no reason why the same physical set of input–output coefficients might not be associated with different price systems: even in the case of a continuously differentiable production function, the marginal product of ‘social’ capital cannot be a purely real magnitude independent of prices. Once it is admitted that ‘in general marginal products are in terms of net value at constant prices, and hence may well depend upon what those prices happen to be’ (Bliss, 1975, p. 195), it is natural to allow for different marginal productivities of the same capital stock at different price systems. It would thus appear that reswitching of technique does not carry with it any logical contradiction even in the case of a smoothly differentiable production function.

But Pasinetti also pointed out that the concept of neoclassical substitutability is itself a very restrictive concept indeed, as it requires the possibility of infinitesimal variations of each input at a time. In fact, Pasinetti noted that it is possible to have a continuous variation of techniques (that is, continuous substitutability) along the w–r relation and yet wide discontinuities in the variation of many inputs between one technique and another, thus making reswitching a quite normal phenomenon (see Pasinetti, 1969). Moreover, and even more significantly, a non-monotonic relation between the rate of profit and capital per man may well be obtained even in the absence of reswitching (Pasinetti, 1966; Bruno, Burmeister and Sheshinski, 1966). This last possibility calls attention to the phenomenon that lies at the root of the various ‘paradoxes’ in the theory of capital: the fact that, unless special assumptions are made, a change in the rate of profit and in the unit wage at given technical coefficients is associated with a change of relative prices.

This debate continued for a few years in the late 1960s and early 1970s, with a series of journal articles (see for example Robinson and Naqvi, 1967) and books (see for example Harcourt, 1972). In particular, John Hicks presented a ‘Neo-Austrian’ model in *Capital and Time* (1973), concluding that reswitching of technique can be excluded only in the special case in which all the techniques have the same ‘duration parameters’, which means the same ‘construction period’ and ‘utilization period’ (1973, pp. 41–4).

In the end, numerous details were added. Yet the basic essential results remained those that had come out of Sraffa's book and of the symposium on ‘Paradoxes in Capital Theory’. It is instructive to see that, in a recent exchange of views that has appeared in the *Journal of Economic Perspectives* (2003, Spring and Winter issues), Franklin Fisher (2003), Geoff Harcourt in Cohen and Harcourt (2003) and Luigi Pasinetti (2003), when asked to succinctly summarize the issues at stake, have essentially restated their original positions.

Aftermath and ways ahead

The discovery of paradoxes in capital theory has had a number of important repercussions, mostly beyond its original context. For it stimulated a large amount of analytical and empirical research on some of the issues that had been discussed in the controversy, without pressing the attention towards the fundamentals, as had been the case with the original debates. In many instances, the recent

developments have been motivated by the need to face the problem of measuring the stock of capital goods in economic systems subject to advances of technical knowledge and structural change, or some of the associated issues in the theory of economic dynamics. In this section we shall refer to some of these developments without pretending to give a complete picture, but with the purpose of identifying the main lines of inquiry.

A first area of research has been the analysis of the necessary conditions for the empirical measurement of aggregate capital. Franklin Fisher elaborated a research line he had himself started in an earlier contribution (Fisher, 1969) and called attention to the fact that the aggregation of outputs, as well as that of productive factors, ‘requires separability in each firm's production function’ (Fisher, 1987, p. 55). He also noted that, under constant returns, the two highly restrictive assumptions of no specialization and generalized capital augmentation are necessary, whereas, in most cases of non-constant returns, aggregation would not be allowed even when assuming the same production function for all firms (Fisher, 1987, p. 55). Robert Gordon proposed to measure collections of heterogeneous capital goods, under condition of embodied technical change, by considering the associated ‘net revenue at a given set of prices (w) of variable inputs’ (Gordon, 1993, p. 106; see also Gordon, 1990). Edward Denison did find Gordon's proposal objectionable and proposed instead to ‘equate’ new capital goods with the old ones by ‘what their relative costs *would* be if both *were* produced at a common date’ (Denison, 1993, pp. 89–90). An interesting link between this literature and the capital controversy debate has been suggested by Charles Hulton, who has called attention to the advantages of a ‘recursive description of the production possibility set’, in which the assumption of capital as an original input is dropped, and ‘capital and labour are assumed to produce gross output *and* capital which is one period older’ (Hulten, 1992, p. S15). Hulten's formulation highlights the central role of knowledge advances embodied in new capital goods and suggests the relevance, for distinct purposes, of gross outputs and net outputs ‘as indicators of capacity and economic welfare’ (Hulten, 1992, p. S11). Alexandra Cas and Thomas Rymes have specifically addressed the issue of whether ‘knowledge of the constant-price aggregate stock of capital would, *for the comparison of economies*, permit one to “predict” certain variables’ (Cas and Rymes, 1991, p. 7; emphasis added). In particular, they investigated capital measurement issues brought about by embodied technical change, and proposed a set of ‘new measures’ aimed at taking the fact into account that ‘the net capital stocks of each industry and at the aggregate are themselves being produced with increased efficiency when the capital goods industries are experiencing advances in technical knowledge’ (Cas and Rymes, 1991, p. 67). The same authors relate their measures of changing capital stocks under conditions of structural change to ‘Pasinetti's concepts of vertically integrated sectors and productivity aggregated by end use’ (Cas and Rymes, 1991, pp. 90–1). This point of view highlights the common ground behind recent attempts to measure stocks of heterogeneous capital goods in terms of an aggregate concept of productive capacity, be it Pasinetti's ‘unit of vertically integrated productive capacity’ (Pasinetti, 1973; 1981), Cas and Rymes' ‘new measures of multifactor productivity’ (Cas and Rymes, 1991), or Hulten's ‘accounting for capacity’ (Hulten, 1992). In all these cases, the producibility of capital goods is emphasized, as is the close connection between advances of technical knowledge and the reshuffling of inter-industry relationships (particularly those affecting intermediate goods). Philippe Aghion and Peter Howitt have commented on recent discussions about capital measurement for an economy subject to advances of knowledge by recalling Joan Robinson's view that the real issue is not

so much about the measurement of capital as rather about the *meaning* one wishes to assign to any given collection of capital goods (Aghion and Howitt, 1998, p. 435).

Another line of investigation has concerned the attempt to assess the empirical (or computational) relevance of capital paradoxes, as distinct from their theoretical possibility. In this connection, Stefano Zambelli has used computer simulations in order to investigate the ‘realism’ of capital paradoxes in artificial economies (Zambelli, 2004). This author has found a significantly higher likelihood that the capital–labour ratio be *positively* related to the rate of profit, contrary to the conventional belief of a negative relationship between these two variables. This result is consistent with the empirical investigation carried out by Zonghie Han and Bertram Schefold (2006). These authors have compared pairs of techniques from the OECD input–output database, and have found that ‘observed cases of reswitching and reverse capital deepening are more than flukes’ (Han and Schefold, 2006, p. 22), even if we are far from observing what has been called an ‘avalanche of switchpoints’ (Schefold, 1997, pp. 278–80).

A third line of research has carried the discussion of capital paradoxes into the field of dynamic economic theory. The literature relevant in this connection is itself quite differentiated. For example, Frank Hahn (1966) called attention to his earlier discovery of zones of instability in economies with heterogeneous capital goods, and pointed out that reswitching should be considered as one amongst the multiple causes of instability in capital markets (Hahn, 1982). It is interesting that this line of argument, while maintaining that reswitching is a special case of a larger class of phenomena, at the same time and rather surprisingly also makes reswitching to be *more general* than was the case with earlier treatments of the same phenomenon. For capital paradoxes are no longer mainly associated with an economy with heterogeneous capital goods and a uniform rate of profit, but are ‘extended’ to the case of multi-sectoral economies with many different capital goods and a *multiplicity* of rates of interest (and rates of profit). Luigi Pasinetti followed a different approach, and examined the analytical features of a dynamic economy in which market interactions are not explicitly examined (Pasinetti, 1981). In this case, too, there are reasons to think that reswitching and reverse capital deepening would not represent exceptional cases, and would not be limited to the institutional framework of a perfectly competitive economy. Other authors have examined the relationship between capital paradoxes and dynamic stability, and have argued that reswitching of technique and reverse capital deepening are neither necessary nor sufficient conditions for the economic system to show *lack* of stability and irregular behaviour (Mandler, 2005). It has also been emphasized that ‘reswitching’ adds an important element of instability, the importance of which depends on the process of adaptation, but also on the utility function’ (Schefold, 2005, p. 467).

More generally, the discovery of capital paradoxes has stimulated a deeper understanding of the features of continuity and discontinuity in the dynamics of economic systems. This line of research has its point of departure in a phenomenon detected by Luigi Pasinetti shortly after the climax of the controversy (Pasinetti, 1969). In Pasinetti's more recent words, ‘the vicinity, even the infinitesimal vicinity, of any two techniques on the scale of variation of the rate of profits does not entail at all vicinity of such techniques … discontinuities in input use.’ (Pasinetti, 2000, p. 409). John Barkley Rosser Jr. has picked up such suggestions and has investigated the discontinuities in order to identify the implications of capital paradoxes for the analysis of the optimal dynamic path followed by an economy characterized by ‘an infinite, differentiable technology’ (Rosser, 1983, p. 182). This author acknowledges that it may

sometimes be impossible to directly observe reswitching along optimal adjustment path (as maintained, for example in Burmeister and Hammond, 1977), but he notes that this would only happen ‘at the price of dynamic discontinuities’, that is, on the condition that the economic system be able to ‘jump over’ the zone associated with intermediate techniques. The above result has been interpreted as showing that ‘in a world of infinite and smooth technologies, reswitching is to be “observed” by observing discontinuities in optimal dynamic paths’ (Rosser, 1983, p. 183; see also Rosser, 2000, pp. 213–20). This point of view emphasizes the analytical importance of capital paradoxes as characteristic instances of the discontinuities that may be generated by the nonlinearity of certain structural relationships. In this way, the propositions discovered during the capital controversies of the mid-20th century are found to be consilient with much later developments in the economic analysis of nonlinear dynamic systems.

Synthesis

The source of most of the difficulties that have emerged in capital theory may be traced back to the fact that ‘capital’ may be conceived in two fundamentally different ways: (a) as a ‘free’ fund of resources, which can be switched from one use to another, without any significant difficulty: this is what may be called the ‘financial’ conception of capital; (b) as a set of productive factors that are embodied in the production process as it is carried out in a particular productive establishment: this is what may be called the ‘technical’ conception of capital.

The idea that there exists an inverse monotonic relation between the rate of interest and the demand for capital was born in the financial sphere. The parallel idea of an inverse monotonic relation between the rate of profit and the ‘quantity of capital’ employed in the production process is the outcome of a long intellectual process of extensions and generalizations reviewed earlier in this essay. But the recent debate on capital theory has conclusively proved that such extensions and generalizations are devoid of any foundation. It is logically impossible to make the ‘financial’ and the ‘technical’ conceptions of capital coincide, except under very restrictive conditions indeed. More precisely, there is no unambiguous way in which a decreasing rate of profit may be related to the choice of alternative techniques, in terms of monotonically increasing capital intensity, be this considered in terms of capital per unit of output or of capital per unit of labour.

These analytical results are hardly in dispute by now. But their ultimate significance and relevance for economic theory have been, and remain, controversial.

A group of economists have been so impressed by the new discoveries in capital theory, concerning the relations between rate of profit, capital per head, capital per output, and technical progress, as to become convinced that these discoveries are calling for a reconstruction of economic theory from its very foundations. It is stressed that the traditional beliefs are due to mistaken generalizations from the theory of short-run microeconomic behaviour, and it is argued that the economic theory (‘marginal economic theory’) that led to mistakes and inconsistencies should be abandoned. It is also pointed out that the obvious alternative is a resumption and development of the more comprehensive approach to value, distribution and growth of the classical economists (see Garegnani, 1970; 2005, and, in a different context, Pasinetti, 1981).

A second line of interpretation maintains that economic theorists should be prepared to give up the analytical tools of equilibrium analysis and concentrate much more on the actual historical dynamics of

economic systems. In this vein, reswitching of technique is acknowledged as a logical possibility but doubts are expressed on its importance in actual economic history (see Robinson, 1975, pp. 38–9; Hicks, 1979, p. 57).

A third line of interpretation is taken by more traditionally minded theoretical economists. It is argued that the discovery of ‘anomalies’ in the field of capital theory does point to an important deficiency in ‘marginal’ economic theory, which leads to the inevitable abandonment of the concept of ‘aggregate capital’. However, it is also argued that there is a way of overcoming this deficiency without giving up the basic premises of traditional theory, and in particular without rejecting the application of the demand-and-supply framework to the study of production. This way induces to concentrating the analysis either on the study of ‘short-run’ (‘temporary’) equilibria, in which the physical stocks of capital are given, or on the equilibrium of an intertemporal economy, in which goods are described by taking their dates of delivery into account. In either case, the logical possibility (or ‘existence’) of an equilibrium price vector is studied without explicitly considering the movement of ‘free’ capital from one use to another. In this approach, the importance of ‘capital paradoxes’ is explicitly recognized, but the associated difficulties are transferred either to the field of stability analysis or to the theory of the long-period supply of saving as financial capital (see, respectively, Hahn, 1982; Bliss, 2005).

A fourth line of interpretation has been pursued by many empirically oriented economists. It is acknowledged that the notion of ‘aggregate’ technical capital is untenable in terms of theory, but it is also argued that the utilization of aggregate production functions may be justified on pragmatic terms, due to supposedly satisfactory econometric fit (see, for example, Fisher, 1971; Fisher, Solow and Kearl, 1977). This view however, is by no means widely accepted. It has in fact been vigorously challenged by Paolo Sylos Labini (1995), who has reviewed the estimates that have emerged from using the Cobb–Douglas production function and has shown that such a ‘production function, when estimated econometrically, tends to yield, in general, poor results’ (Felipe and Fisher, 2003, p. 251; see also McCombie, 1998; and Felipe and Adams, 2005). In a recent evaluative essay on aggregation in production functions, Jesus Felipe and Franklin Fisher have sharply criticized the continued use of aggregate parables. In particular, they maintain that ‘the revival of growth theory during the last two decades no doubt has produced important discussions, and seemingly interesting empirical results’ but ‘authors do not realize that they are using a tool whose lack of legitimacy was demonstrated decades ago’ (Felipe and Fisher, 2003, pp. 250–1). The same economists emphasize that ‘the impossibility of testing empirically the aggregate production function’ is ‘substantially more serious than a mere anomaly’, and that ‘macroeconomists should pause before continuing to do applied work with no sound foundation and dedicate some time to studying other approaches to value, distribution, employment, growth, technical progress etc., in order to understand which questions can legitimately be posed to the empirical aggregate data’ (Felipe and Fisher, 2003, pp. 256–7). It is interesting that the theoretical and empirical researches that have taken up this challenge have devoted attention to the construction of a ‘capacity measure’ of the stock of technical capital that would allow comparisons across different states of technology without having recourse to the traditional ‘parables’ (see, for example, Pasinetti, 1973; 1981; Cas and Rymes, 1991; Hulten, 1992).

Finally, let us note how the discovery of ‘paradoxes’ in capital theory has contributed to stimulating research into the dynamic properties of economic systems outside the world of steady state comparisons. In particular, some economists have attempted the theoretical investigation of regularities in the long-run

dynamics of economic systems by suggesting a reformulation of the classical theory of structural change in a disaggregated framework (see Pasinetti, 1981; 1993; Hagemann, Landesmann and Scazzieri, 2003).

Others have investigated the complex interaction of behavioural patterns along a dynamic trajectory, and have called attention to increasing returns and other nonlinear phenomena in structurally adaptive economic systems (see Anderson, Arrow and Pines, 1988; Arthur, Durlauf and Lane, 1997).

Whatever the view that is taken, the major victim of the debate has been the Böhm-Bawerk–Clark–Wicksell theory of capital that was so patiently constructed towards the end of the 19th century. This theory relied on a conception of ‘aggregate capital’ that was taken as measurable independently of the rate of profit and of income distribution. Such a conception of ‘capital’ has had to be jettisoned, which has stimulated reformulations of the pure theory of capital. There has been on the one hand a return to the Walrasian general equilibrium theory in its intertemporal formulation, and on the other hand a remarkable revival of classical political economy. The controversy had also a number of less striking but perhaps longer-term consequences. The consideration of paradoxes has alerted economists to the richness and complexity of economic relationships, and to the need to avoid a process of generalization from the consideration of special cases. In any case the debate seems to have compelled theoretical economists to be more rigorous about the nature and limits of their assumptions. In many important cases, it has also brought about a change in the main focus of their analysis.

All this leads one reasonably to expect as unlikely that the next generation of economists will leave the issue of capital theory at rest.

See Also

- reswitching of technique
- reverse capital deepening

Bibliography

Aghion, P. and Howitt, P. 1998. *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Anderson, P., Arrow, K.J. and Pines, D., eds. 1988. *The Economy as an Evolving Complex System*. Redwood City, CA: Addison Wesley.

Arthur, W.B, Durlauf, S. and Lane, D., eds. 1997. *The Economy as an Evolving Complex System II*. Redwood City, CA: Addison-Wesley.

Baranzini, M. and Scazzieri, R. 1990. *The Economic Theory of Structure and Change*. Cambridge: Cambridge University Press.

Beccaria, C. 1804. *Elementi di economia pubblica* (ms 1771–72). In *Scrittori Classici Italiani di Economia Politica*, Vol. XVIII, 17–356 and Vol. XIX, 5–166, ed. P. Custodi. Milan: Destefanis.

Bliss, C.J. 1975. *Capital Theory and the Distribution of Income*. Amsterdam and Oxford: North-Holland; New York: American Elsevier.

Bliss, C.J. 2005. Introduction. The theory of capital: a personal overview. In *Capital Theory*, vol. 1, ed. J. Bliss, A.J. Cohen and G.C. Harcourt. Cheltenham, UK and Northampton, MA: Edward Elgar.

Böhm-Bawerk, E. von. 1889. *Positive Theorie des Kapitales* (*Kapital und Kapitalzins*, Zweite Ableitung). Trans. as The Positive Theory of Capital. London: Macmillan.

Bruno, M., Burmeister, E. and Sheshinski, E. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80, 526–53.

Burmeister, E. and Hammond, P. 1977. Maximin paths of heterogeneous capital accumulation and the instability of paradoxical steady states. *Econometrica* 45, 853–70.

Cannan, E. 1929. *A Review of Economic Theory*. London: P.S. King and Son Ltd.

Cas, A. and Rymes, T.K. 1991. *On Concepts and Measures of Multifactor Productivity in Canada, 1961–1980*. Cambridge and New York: Cambridge University Press.

Champernowne, D. 1953–4. The production function and the theory of capital: a comment. *Review of Economic Studies* 21, 112–35.

Clark, J.B. 1899. *The Distribution of Wealth*. New York: Macmillan.

Cohen, A.J. and Harcourt, G.C. 2003. Retrospectives. Whatever happened to the Cambridge capital theory controversies? *Journal of Economic Perspectives* 17(1), 199–214.

Cohen, A.J. and Harcourt, G.C. 2005. Introduction. Capital theory controversy: scarcity, production, equilibrium and time. In *Capital Theory*, vol. 1, ed. J. Bliss, A.J. Cohen and G.C. Harcourt. Cheltenham, UK and Northampton, MA: Edward Elgar.

Denison, E.F. 1993. Robert J. Gordon's concept of capital. *Review of Income and Wealth* 39, 89–102.

Felipe, J. and Adams, F.G. 2005. ‘A theory of production’: the estimation of the Cobb–Douglas function: a retrospective view. *Eastern Economic Journal* 31, 427–45.

Felipe, J. and Fisher, F.M. 2003. Aggregation in production functions: what applied economists should know. *Metroeconomica* 54, 208–62.

Fisher, F.M. 1969. The existence of aggregate production functions. *Econometrica* 37, 553–77.

Fisher, F.M. 1971. Aggregate production functions and the explanation of wages: a simulation experiment. *Review of Economics and Statistics* 53, 305–25.

Fisher, F.M. 1987. Aggregation problem. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Fisher, F.M. 2003. Cambridge capital controversies. *Journal of Economic Perspectives* 17(4), 228–9.

Fisher, F.M., Solow, R. and Kearl, J.M. 1977. Aggregate production functions: some CES experiments. *Review of Economic Studies* 44, 305–20.

Garegnani, P. 1966. Switching of techniques. *Quarterly Journal of Economics* 80, 554–87.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37, 407–36.

Garegnani, P. 2005. Capital and intertemporal equilibria: a reply to Mandler. *Metroeconomica* 56, 411–37.

Gordon, R.J. 1990. *The Measurement of Durable Goods Prices*. Chicago and London: University of Chicago Press.

Gordon, R.J. 1993. Reply: the concept of capital. *Review of Income and Wealth* 39, 103–10.

Hagemann, H., Landesmann, M. and Scazzieri, R., eds. 2003. *The Economics of Structural Change*. Cheltenham, UK and Northampton, MA: Edward Elgar.

Hahn, F. 1966. Equilibrium dynamics with heterogeneous capital goods. *Quarterly Journal of Economics* 53, 633–46.

Hahn, F. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6, 353–74.

Han, Z. and Schefold, B. 2006. An empirical investigation of paradoxes: reswitching and reverse capital deepening in capital theory. *Cambridge Journal of Economics* 30, 737–65.

Harcourt, G.C. 1972. *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.

Hicks, J. 1965. *Capital and Growth*. Oxford: Clarendon Press.

Hicks, J. 1973. *Capital and Time: A Neo-Austrian Theory*. Oxford: Clarendon Press.

Hicks, J. 1977. Capital controversies: ancient and modern. In *Economic Perspectives: Further Essays on Money and Growth*. Oxford: Clarendon Press.

Hicks, J. 1979. *Causality in Economics*. Oxford: Basil Blackwell.

Hulten, C.R. 1992. Accounting for the wealth of nations: the net versus gross output controversies and its ramifications. *Scandinavian Journal of Economics* 94(Supplement, September), 9–24.

Jevons, W.S. 1879. *The Theory of Political Economy*, 2nd edn. London: Macmillan and Co (1st edn 1871).

Kaldor, N. 1937. Annual survey of economic theory: the recent controversy on the theory of capital. *Econometrica* 5, 201–33.

Levhari, D. 1965. A nonsubstitution theorem and switching of techniques. *Quarterly Journal of Economics* 79, 98–105.

Levhari, D. and Samuelson, P. 1966. The nonswitching theorem is false. *Quarterly Journal of Economics* 80, 518–19.

McCombie, J.S.L. 1998. ‘Are there laws of production?’: an assessment of the early criticisms of the Cobb–Douglas production function. *Review of Political Economy* 10, 141–73.

Mandler, M.A. 2005. Well-behaved production economies. *Metroeconomica* 56, 477–94.

Morishima, M. 1964. *Equilibrium, Stability and Growth: A Multi-Sectoral Analysis*. Oxford: Clarendon Press.

Morishima, M. 1966. Refutation of the nonswitching theorem. *Quarterly Journal of Economics* 80, 520–25.

Pasinetti, L.L. 1965. Changes in the rate of profit and degree of mechanization: a controversial issue in capital theory. Unpublished paper presented at the First World Congress of the Econometric Society, Rome.

Pasinetti, L.L. 1966. Changes in the rate of profit and switches of techniques. *Quarterly Journal of Economics* 80, 503–17.

Pasinetti, L.L. 1969. Switches of technique and the ‘rate of return’ in capital theory. *Economic Journal* 79, 508–31.

Pasinetti, L.L. 1973. The notion of vertical integration in economic analysis. *Metroeconomica* 25, 1–29.

Pasinetti, L.L. 1981. *Structural Change and Economic Growth: A Theoretical Essay on the Dynamics of the Wealth of Nations*. Cambridge: Cambridge University Press.

Pasinetti, L.L. 1993. *Structural Economic Dynamics: A Theory of the Economic Consequences of Human Learning*. Cambridge: Cambridge University Press.

Pasinetti, L.L. 2000. Critique of the neoclassical theory of growth and distribution. *Banca Nazionale del Lavoro Quarterly Review* 53, 383–431.

Pasinetti, L.L. 2003. Cambridge capital controversies. *Journal of Economic Perspectives* 17(4), 227–8.

Ricardo, D. 1817. *On the Principles of Political Economy and Taxation*. Vol. I of *The Works and Correspondence of David Ricardo*, ed. P. Sraffa with the collaboration of M.H. Dobb, Cambridge: Cambridge University Press, 1951.

Robinson, J. 1953–4. The production function and the theory of capital. *Review of Economic Studies* 21 (2), 81–106.

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Robinson, J. 1975. The unimportance of reswitching. *Quarterly Journal of Economics* 89, 32–9. (Reprinted in J. Robinson, *Collected Economic Papers*, vol. 5. Oxford: Basil Blackwell, 1979.

Robinson, J. and Naqvi, K.A. 1967. The badly behaved production function. *Quarterly Journal of Economics* 81, 579–91.

Rosser, J.B., Jr. 1983. Reswitching as a cusp catastrophe. *Journal of Economic Theory* 31, 182–93.

Rosser, J.B. 2000. *From Catastrophe to Chaos: A General Theory of Economic Discontinuities. Vol. 1: Mathematics, Microeconomics, Macroeconomics, and Finance*, 2nd edn. Boston, Dordrecht and London: Kluwer Academic Publishers.

Samuelson, P.A. 1962. Parable and realism in capital theory: the surrogate production function. *Review of Economic Studies* 29, 193–206.

Samuelson, P.A. 1966. A summing up. *Quarterly Journal of Economics* 80, 568–83.

Say, J.B. 1817. *Traité d'économie politique*, 3rd edn. Paris: Déterville (1st edn 1803).

Schefold, B. 1997. *Normal Prices, Technical Change and Accumulation*. London: Palgrave Macmillan.

Schefold, B. 2005. Reswitching as a case of instability of intertemporal equilibria. *Metroeconomica* 56, 438–76.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Vol. II of *The Glasgow Edition of the Works and Correspondence of Adam Smith*. Oxford: Clarendon Press. General editors R. H. Campbell and A.S. Skinner; textual editor W.B. Todd.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Starrett, D. 1969. Switching and reswitching in a general production model. *Quarterly Journal of Economics* 83, 673–87.

Stiglitz, J. 1973. The badly behaved economy with the well-behaved production function. In *Models of Economic Growth*, ed. J.A. Mirrlees and N.H. Stern. London: Macmillan.

Sylos Labini, P. 1995. Why the interpretation of the Cobb–Douglas production function must be radically changed. *Structural Change and Economic Dynamics* 6, 485–504.

Thünen, J.-H. 1857. *Le salaire naturel et son rapport au taux de l'intérêt*. Trans. M. Wolkoff, Paris: Guillaumin et Cie. German original published in 1850.

Wicksell, K. 1901–6. *Lectures on Political Economy*. vol. 1, *General Theory*. London: Routledge, 1934.

Zambelli, S. 2004. The 40% neoclassical aggregate theory of production. *Cambridge Journal of Economics* 28, 99–120.

How to cite this article

Pasinetti, Luigi L. and Roberto Scazzieri. "capital theory (paradoxes)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000042>
doi:10.1057/9780230226203.0196

The New Palgrave Dictionary of Economics Online

capital theory

Robert A. Becker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Capital theory examines the special role played by time in resource allocation studies. The determination of the interest rate and functional distribution of income as well as how rational agents invest are analysed within single- and multi-sector general equilibrium frameworks. Here, agents exercise perfect foresight over alternative consumption and capital accumulation programs. Efficient programs are characterized. Representative and multi-agent infinitely lived households are studied. Equivalence principles link the equilibrium programs and optimal paths. Heterogeneous agent models with borrowing constraints are reviewed. A behavioural model of intertemporal choice is also compared to its constant discounting counterpart.

Keywords

aggregate capital; aggregation; Allais paradox; altruism; arbitrage; arbitrage pricing theory; balanced growth; behavioural economics; Cambridge controversies; capital accumulation; capital deepening; capital theory; capital value; Cobb–Douglas functions; commitment; comparative dynamics; continuous and discrete time models; deterministic models; discounting; dynamic non-substitution theorems; dynamic programming; economic growth; efficient allocation; elasticity of substitution; Epstein–Hynes utility functions; equivalence principle; Euler equations; exhaustible resources; existence of general equilibrium; expected utility; functional distribution of income; futures markets; general equilibrium; Hahn problem; Hotelling, H.; hyperbolic discounting; impatience; incomplete markets; infinite horizons; interest rates; interest rate determination; intertemporal choices; intertemporal utility functions; investment criteria; Fisher, I.; Kuhn–Tucker conditions; long run and short run; Markov perfect equilibrium; multi-capital goods models; named goods; neoclassical capital theory; no-arbitrage conditions; non-classical production functions; optimal growth; orthodox vision of capital theory; perfect foresight; portfolio equilibrium; present value investment criteria; probability; quasi-geometric discounting; Rae, J.; Ramsey, F. P.; Ramsey model; rational expectations; rationality; recursive utility functions; reduced form model; renewable resources; representative agent; risk; saving and investment; shadow pricing; spot markets; Sraffa, P.; stationary state; stylized facts; technical change; time; time consistency; time preference; transversality

condition; turnpike theorems; uncertainty; von Neumann, John

Article

1 Introduction

Capital theory examines the special role played by time in resource allocation studies. The determination of the rate of interest and the functional distribution of income are considered along with the development of criteria for evaluating investment decisions. Contemporary capital theory focuses on the intertemporal choices undertaken by rational actors within a general equilibrium setting where all prices and allocations are determined by market clearing. The central role played by time is that producing goods and services to supply future consumption requires withdrawing some output from current consumption in order to create the produced means of production, or capital goods, which enable future production to be undertaken in conjunction with other factors such as labour and land. That agents seek to make their investment decisions rationally is taken as a fundamental premise of capital theoretic models. The rationality hypothesis is implemented by assuming that agents maximize a utility function over paths of future consumption and that producers maximize the present discounted value of their profits. A specification of the degree of foresight must be postulated together with an assumption on which spot and futures markets are open for trade. Consumption and investment decisions are realized in a market equilibrium.

2 Dated commodities and prices

The classical general equilibrium model developed over the last half of the 20th century by Arrow, Debreu, McKenzie and their followers was sufficiently abstract that it could model any number of different economic activities by the device of *named goods*: a commodity was specified by its physical characteristics, date of availability, contingent events upon which its availability depended, as well as its location. For example, a consumption good available now was differentiated from the same physical commodity available at a different date even if the location or contingent events were the same at both dates. Capital theoretic models focused on the pure role of time assume certainty (no contingent events) and the same location. The simplest models assume that there is just one consumption good and that its characteristics are the same at each point of time. Only the date of its availability differentiates goods. These are the *deterministic models*. Agents are supposed to exercise perfect foresight over the paths of all relevant variables in this case. Other models treat both time and uncertainty by way of dated goods and contingent events. Rational expectations about the future probability distributions of variables are assumed to describe agents' behaviour. The basic principles and issues in capital theory are most easily reviewed in the deterministic setting with risk and uncertainty treated as a non-trivial extension of the basic theory.
The classical general equilibrium model assumes a finite number of commodities. In the deterministic intertemporal setting this means there are a finite number of dated commodities. Consumers have a finite planning horizon; time unfolds in discrete periods, $t = 1, 2, \ldots, T$. A finite number of goods are available at each date, indexed by $i = 1, 2, \ldots, N$. This makes for *NT* commodities. Consumers' preferences are defined over a commodity space contained in an *NT*-dimensional Euclidean space. Similarly, producers' technology sets were defined in the same commodity space. Competitive prices are established through a market mechanism on the presupposition that markets operate for all *NT* commodities. The classic existence of

equilibrium and welfare theorems apply under appropriate assumptions on the consumption and production sectors as well as the relations between them. This formal connection between intertemporal and atemporal static general equilibrium theory offers little that is new or special to capital theory. It is the recognition that time places restrictions on preferences and technologies that specialize the abstract Walrasian model to the type more suited to answering capital theoretic questions about interest rate determination and the corresponding division of the model's output among its participating consumers and resource owners.

The distinguishing feature of capital theoretic models is their focus on infinite horizon decision problems. The motivation for this lies in the open-ended nature of the economic problem. Economies do not have foreseeable ends and the problem of saving and investing for future consumption seemingly goes on for ever, even though all the decision makers know that our planet's time is limited. But that terminal date is so far in the future that we might as well act today as if an infinite horizon is a good approximation to a very long but finite horizon. The theoretical advantage of the infinite horizon is that it allows us to draw a sharp formal distinction between the short and the long runs. The short run represents the transitional time that model solutions follow, whereas the long run constitutes the solutions' properties as time runs towards infinity. The classical focus on the stationary state, or ‘long period’, presumes there is a long run and that the economy evolves towards it.

Frank Ramsey (1928) modelled infinite horizons in a seminal article on optimal growth. He argued that discounting by the planner was ethically indefensible. Ramsey's modern followers from Paul Samuelson to the present day have studied both undiscounted and discounted models. Von Neumann's (1937) celebrated model of capital accumulation at a maximum balanced growth rate implicitly assumed an infinite horizon. A balanced program occurs when each type of capital good grows from one period to the next at the same constant rate. By focusing attention on balanced growth paths, it would seem reasonable that von Neumann understood those programs might correspond to that model economy's long-run position. The infinite horizon assumption has a long tradition in capital theory and finance (for example, the consol bonds issued by the United Kingdom; see Goetzmann and Rouwenhorst, 2005, for other examples).

This article concentrates entirely on the discounted case and its connection to general competitive analysis. The primary focus is taken to be the one-sector discounted Ramsey model. Capital theory is viewed as a branch of general equilibrium theory. The masterful surveys by McKenzie (1986; 1987) lay out the undiscounted as well as discounted models for many capital goods and multiple sectors in great generality. His surveys also provide details on how those models can evolve over time (the so-called turnpike theorems) as well as general comparative dynamics results.

Ramsey (1928) formulated his seminal model in continuous time. The models presented here are cast in discrete time with periods $t = 1, 2, \ldots$ This turns out to have some technical advantages over continuous time modelling as well as expositional advantages as economic concepts are more readily grasped by readers unschooled in the calculus of variations and its modern development, optimal control theory.

3 Neoclassical capital theory: the one-sector model

3.1 The discounted Ramsey optimal growth model

Neoclassical capital theory is illustrated by the properties exhibited in the discrete time one-sector discounted Ramsey optimal growth model (Ramsey, 1928). This model encapsulates the fundamental consumption–investment trade-offs that a decision maker considers when choosing a consumption plan

over time to achieve a maximum lifetime utility. The model is simplified in many ways. There is a single decision maker, or planner, acting over an infinite horizon. There is no uncertainty or shocks that would make output available in the future look like a random variable when viewed from the present. The model examines an aggregated economy. There is a single all-purpose consumption good produced using capital goods (carried over from the previous period) and fixed labour. The capital and consumption goods available at each time are physically identical and can be costlessly converted from consumption to capital (and vice versa) at a one-to-one rate. The planner decides how much to consume in the current period and how much to save for next period's production. Capital depreciates entirely within the period. It is *circulating* as it is used up within the production period. Extensions to include durable capital that depreciates at a fixed rate are straightforward. The planner's exogenously given initial stock of capital produces goods available in the first period. The planner obtains utility from consumption at each time and maximizes the discounted sum of future utilities. The discount factor on future utility is a given constant. The planner's intertemporal optimization problem is:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u(c_t) \text{ by choice of } \left\{c_t, k_{t-1}\right\}_{t=1}^{\infty}, \tag{1}$$

subject to:

$$c_t + k_t \leq f(k_{t-1}) \text{ for } t = 1, 2, \ldots; c_t \geq 0, k_{t-1} \geq 0 \text{ all } t; \ k_0 \leq k, \text{ where } k > 0 \text{ is given.} \tag{2}$$

Feasible programs are sequences $\{c_t, k_{t-1}\}_{t=1}^{\infty}$ which satisfy (2). Assume $u: [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $u(0) = 0$, and satisfies the Inada condition: $\lim_{c \to 0+} u'(c) = \infty$. The production function $f: [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $f(0) = 0$, satisfies $\lim_{k \to 0+} f'(k) = \infty$, and $\lim_{k \to \infty} f'(k) < 1$ (also called Inada conditions). There is a maximum sustainable stock, $b > 0$, with $f(b) = b$ and $0<k<b$. The discount factor, δ, satisfies $0< \delta <1$; $\delta = 1/(1+\rho)$, where $\rho > 0$ is the pure rate of time preference (or rate of impatience). There is a unique optimal program, $\{\bar{c}_t, \bar{k}_{t-1}\}_{t=1}^{\infty}$. Its discounted utility sums, $\sum_{t=1}^{\infty} \delta^{t-1} u(\bar{c}_t) < \infty$. The optimal growth problem has a *time consistency property*: The optimal sequence $\{\bar{c}_t, \bar{k}_{t-1}\}_{t=1}^{\infty}$ has the property that $\{\bar{c}_{t+\tau}, \bar{k}_{t-1+\tau}\}_{t=1}^{\infty}$ solves the optimization problem with objective starting at time τ, $\sum_{t=1}^{\infty} \delta^{t-1+\tau} u(c_{t+\tau})$, subject to $c_{t+\tau} + k_{t+\tau} \leq f(k_{t-1+\tau})$ for $t = 1, 2, \ldots$ and $k_\tau = k$. Calender time is irrelevant: if the planner's objective is moved forward τ periods and the initial capital stock is maintained at the new starting time, then the optimal capital and consumption sequence are

identical to the ones initiated at time $\tau = 0$. The reason for this is $\Sigma_{t=1}^{\infty} \delta^{t-1+\tau} u(c_{t+\tau}) = \delta^{\tau} \Sigma_{t=1}^{\infty} \delta^{t-1} u(c_{t+\tau})$, which is multiple of (1) and the set of feasible programs is unchanged. Hence, the optimal solution is unchanged from the same initial condition even though time has simply been reset to start at τ.

The optimal program satisfies $(\bar{c}_t, \bar{k}_{t-1}) > 0$ for each *t*. The Kuhn–Tucker necessary conditions for an optimum, known as the *Euler*, or *no-arbitrage conditions*, are:

$$\delta f'(\bar{k}_t) u'(\bar{c}_{t+1}) = u'(\bar{c}_t), \text{ for each } t. \tag{3}$$

If the planner's horizon is a finite period, *T*, then (3) and the complementary slackness condition $\delta^{T-1} u'(\bar{c}_T) \bar{k}_T = 0$ obtain. The latter condition states capital's terminal value is zero. For the infinite horizon case of interest, it is natural to conjecture the *transversality condition* holds as a necessary condition for optimality:

$$\lim_{T \to \infty} \delta^{T-1} u'(\bar{c}_T) \bar{k}_T = 0. \tag{4}$$

This condition's necessity can be formally demonstrated in many problems. The conditions (3) and (4) are also sufficient conditions for optimality under the maintained hypotheses governing the concavity of the single period return function, *u*, and the production function, *f*.

Equation (3) expresses the unprofitability of the *one-period reversed arbitrages* developed below. An *arbitrage* represents a feasible change in the optimal path. Reversed arbitrages perturb the optimum for finitely many consecutive periods. Unreversed arbitrages change the optimal path permanently from some given time on to infinity. A necessary condition for an optimal path is that no arbitrage increase the discounted sum of future utilities above the optimal discounted utility. The necessity of the transversality condition can be interpreted as a type of no-arbitrage condition for *unreversed arbitrages* which never return to the optimal path.

Suppose that the consumption and capital sequences $(\bar{c}_t, \bar{k}_{t-1}) > 0$ (for each *t*) are optimal for the given initial capital stock. Then, the planner cannot increase utility by undertaking the following activity: at time *t* marginally increase the capital stock to be carried to time *t*+1. This costs the planner $u'(\bar{c}_t)$ utils on the margin. Now invest this extra capital to obtain $f'(\bar{k}_t)$ additional units of goods in period $t+1$ from the production sector. Convert this additional income into consumption at $t+1$ worth $u'(\bar{c}_{t+1})$ utils on the margin. This implies the marginal benefit of this incremental investment measured at $t+1$ is

$f'(\bar{k}_t)u'(\bar{c}_{t+1})$. Now discount this by the utility discount factor δ to place the marginal benefit at time $t+1$ and marginal cost at time *t* in comparable utility units. The marginal benefit cannot exceed the marginal cost along an optimal solution to the household's problem. This is formally expressed by the inequality $\delta f'(\bar{k}_t)u'(\bar{c}_{t+1}) \leq u'(\bar{c}_t)$, for each *t*. Since the capital stock at time *t* is positive, then this arbitrage calculation can be repeated for an increase in consumption at time *t* paid for by lower consumption at time *t*+1. In this case, the inequality is reversed and (3) holds.

This model has one special solution: it is the stationary optimal program (c^*, k^*), with $c^* = f(k^*) - k^*$ and $\delta f'(k^*) = 1$. By concavity of *f*, this program has the property that k^* solves the problem $\max_{k \geq 0}[\delta f(k) - k]$. This is a form of the dynamic non-substitution theorem: the stationary optimal capital stock is independent of the planner's felicity function and depends only on technology and the planner's discount factor. The equation $\delta f'(k^*) = 1$ is also the Euler equation for the program $c_t^* \equiv c^*$ and $k_{t-1}^* \equiv k^*$ for each $t \geq 1$. That is, if the initial capital stock is k^*, then it is optimal to maintain that capital stock for ever. The program $\{c_t^*, k_{t-1}^*\}_{t=1}^{\infty}$ is constant, or stationary, over time. Hence the name: the stationary optimal program (also called the *steady state*). In the case $\delta = 1$ the steady state maximizes stationary consumption over all feasible stationary consumption levels (it is the *optimal stationary consumption path*) and is called the *golden-rule consumption level* while the corresponding stationary capital stock is the *golden-rule capital stock*. For the discounted case, $0 < \delta < 1$, the steady states are also known as the *modified golden-rule consumption* and *capital stock*.

The optimal path of the infinite horizon problem with initial stocks $k \neq k^*$ converges monotonically to the stationary optimal program (c^*, k^*), with $c^* = f(k^*) - k^*$ and $\delta f'(k^*) = 1$. For example, if $0<k<k^*$, then the optimal capital sequence, $\{\bar{k}_{t-1}\}_{t=1}^{\infty} \nearrow k^*$. Moreover, paths *do not cross*: if $0<k<k'<k^*$, then $\bar{k}_t < \bar{k}_t'$, where $\{\bar{k}_{t-1}'\}_{t=1}^{\infty}$ is optimal from initial stocks, k'. The convergence of the optimal path implies it is bounded, and the transversality condition holds as a necessary condition for optimality in this model. Conversely, a feasible program satisfying the Euler equations and transversality condition is an optimal program. The convergence property of the optimal capital sequences is also known as the *turnpike theorem*: the optimal capital sequence from any initial starting stock converges to the modified golden-rule capital stock. The corresponding consumption sequences likewise converge (monotonically) to the golden-rule consumption level. The turnpike theorem's conclusion suggests that there is a distinction between the economy's *long-run steady state* and the *short-run transitional dynamics* that describe how the economy approaches that stationary optimal program. One consequence of the turnpike theorem is that optimal programs spend infinitely many periods in any neighborhood of the steady state. In that sense, the steady state is a good approximation for the transitional dynamics over long periods of time. The choice of the analyst lies in determining how small that neighbourhood is, and hence how many periods the economy is not ‘sufficiently close’ to the model's long-run solution.

•The canonical example

The *logarithmic utility, Cobb–Douglas production economy* is an important example of Ramsey's optimal

growth problem. Many writers refer to it as the *canonical example* of the one-sector model since its solution is explicitly found. The planner's single period utility function is $u(c_t) = \ln c_t$ and the production function has the Cobb–Douglas form $f(x) = x^{\rho}$ where 0< ρ <1 is a technology parameter (it is capital's constant share of total income in a competitive equilibrium setting). The Ramsey optimal growth problem for this specification (and no depreciation) can be solved explicitly by a variety of techniques (see Becker and Boyd, 1997, for one such approach based on symmetry techniques). The solution is described by the *consumption policy function* $g(k) = (1 - \delta\rho)k^{\rho}$ and the *capital policy function* $h(k) = \delta\rho k^{\rho}$. At each date, the policy functions tell the decision maker how much to consume and how much to save given the current level of the capital stock, *k*. The optimal capital and consumption sequences are given by iterating the policy functions. Carrying out that iteration for example leads to the explicit solution for the capital sequence:

$$x_t(k) = (\delta\rho)^{\rho^{t-1}+\cdots+1} k^{\rho^t} \tag{5}$$

The capital and consumption policy functions in this example have constant marginal propensities to save and consume, respectively. Solow's (1956) growth model postulated savings and consumption functions of this type within a one-sector framework with a Cobb–Douglas production function in order to model the process of economic growth. Solow also assumed exogenous technological progress in the form of labour augmenting technical change, whereby each worker becomes more productive at an exponentially growing rate. Solow aimed his model at describing stylized facts of economic growth. The model was not formally set up to reflect microeconomic based optimizing behaviour at the level of individual consumption–saving decisions. The canonical version of Ramsey's discounted model provides such a microfoundation for Solow's descriptive theory in case there is no exogenous technical progress.

Let $k_t = x_t(k)$. The policy functions satisfy the no arbitrage condition. Let $c_t = (1 - \delta\rho)k_{t-1}^{\rho}$ and $c_{t+1} = (1 - \delta\rho)k_t^{\rho}$, where k_t is the capital stock at time *t*. The no arbitrage condition is:

$$\frac{c_t}{\delta c_{t-1}} = \frac{(1 - \delta\rho)k_t^{\rho}}{\delta(1 - \delta\rho)k_{t-1}^{\rho}} = \rho k_t^{\rho-1} = f'(k_t).$$

This solution can also be shown to satisfy the transversality condition, which takes the form here:

$$\lim_{t \to \infty} \frac{\rho k_{t-1}^{\rho-1}\delta^{t-1}}{c_t} = 0.$$

Therefore, the policy functions tell us how to find the optimal solution to this optimal growth problem. The optimal policy functions have the time consistency property as well.
The qualitative features of the optimal solution also follow from the policy functions. The most important observation is that the optimal capital sequence is monotonic as can be shown by iterating the capital policy function. Notice that each optimal path converges to the unique positive fixed point of the capital policy function, k^*, where $h(k^*) = k^*$, which implies that:

$$k^* = (\delta\rho)^{\frac{1}{1-\rho}}.$$

This is the model's modified golden-rule capital stock. If the positive initial capital is below the modified golden rule, then the economy accumulates capital and the sequence of optimal capital stocks increases and converges to the modified golden-rule capital stock. Similarly, the optimal capital stocks decrease and converge to the modified golden rule when the starting stock is larger than the positive fixed point. If the initial capital happens to equal the modified golden-rule stocks, then it will be optimal to maintain those stocks in every period. Thus, the modified golden rule is a steady state of the dynamical system:

$$k_{t+1} = h(k_t) = \delta\rho k_t^{\rho}.$$

The corresponding consumption sequence is also monotonic since the consumption policy function is increasing in capital. The resulting consumption sequence converges to the *modified golden-rule consumption* level defined by:

$$c^* = (1 - \delta\rho)(k^*)^{\rho}.$$

The convergence of the optimal capital and consumption sequences illustrates the *turnpike theorem*. The monotonicity property for optimal capital sequences can also be viewed as a *non-crossing property*: if $k<k'$ are two different starting stocks, then $h(k) = k_1 < k_1' = h(k')$. Continuing in this way we see that, when two starting stocks are compared, the lower one always provides less capital than the higher one at any time along the optimal program.•
The steady state's sensitivity to the discount factor is readily shown for $0< \delta <1$ for the general discounted one-sector model. Let $k^* = k^*(\delta)$ denote the steady state capital stock as a function of the discount factor.

The condition $\delta f'(k^*(\delta)) = 1$ implies upon differentiation that $dk^* / d\delta > 0$. This *comparative steady state* result means that a more patient planner (there is a marginal increase in discount factor) produces a larger stationary optimal capital stock. Some writers on capital theory call this the *capital deepening* response to a change in the discount factor. The corresponding result for the consumption path $c^*(\delta) = f(k^*(\delta)) - k^*(\delta)$ states $dc^* / d\delta > 0$ as well. This is called *non-paradoxical consumption behaviour*. Note that this comparative steady state exercise does not compare the optimal program starting from k^* given the new discount factor to the optimal stationary plan k^* for the old discount factor. Comparative steady state exercises merely compare the steady states before and after a parameter change without evaluating the economy's transition path from one steady state to another.

Comparative dynamics results are available for the one-sector model which include studying the transition from one steady state to another in response to a parameter change. The planner considers all feasible plans in response to a change in one of the economy's deep taste or technology parameters. In particular, it is possible to compare the optimal programs before and after the parameter changes. For example, if the planner's discount factor increases (or, equivalently, the pure rate of time preference declines), then the planner becomes more patient. If the planner's discount factor increases from δ to δ', with $0 < \delta < \delta' < 1$, then the optimal capital paths starting from the same initial capital stock satisfy the conditions $\bar{k}'_t > \bar{k}_t$ for each time – there is a *generalized capital deepening response* because the economy's capital stock is increased at each time. Indeed, the discount factor's *initial impact* is to increase the first period's capital stocks at the expense of first period consumption since the initial capital stocks and first period output are unchanged after the discount factor increases. As the new consumption program converges monotonically to a larger modified golden-rule consumption level, $c^*(\delta')$, it follows that *eventually* (that is, in finite time) $\bar{c}_t(\delta') > \bar{c}_t(\delta)$ must obtain. These comparative dynamics results are easily verified for the canonical example with log utility and Cobb–Douglas production.

It is interesting to note that the monotonicity and non-crossing properties of the one-sector model are robust. For example, the concavity of the production function can be relaxed while preserving these qualitative properties. The production function is *non-classical* provided there is an inflection point, $0<k_I<b$ such that $f''(k) > 0$ for $k<k_I$ and $f''(k)<0$ for $k > k_I$. Non-classical production functions can arise in fishery models when representing the production of a new generation of fish from the existing population. See Becker and Boyd (1997, ch. 5) for details on the non-classical production extensions.

Generalizations of the one-sector model's turnpike property (the convergence of optimal capital sequences to the modified golden-rule stock) are also available for some multi-capital goods models, as found in McKenzie's surveys. The original turnpike theorem for many capital goods models was conjectured by Dorfman, Samuelson and Solow (1958) in the von Neumann model framework without an explicit consumption criterion. Radner (1961) provides the first rigorous proof of a turnpike theorem for a von Neumann style model with a unique maximum balanced growth path and a finite planning horizon. Radner's theory evaluated alternative programs from a given initial vector of capital stocks according to a criterion based on the value of those stocks in the program's final time period. As with Dorfman, Samuelson and Solow's model, Radner's theorem did not apply to a Ramsey-style planner with an objective based on discounted utility. Radner's value loss technique for demonstrating the turnpike theorem did turn out to apply to undiscounted Ramsey models as well as some forms of the discounted model, as summarized in McKenzie's survey articles.

Another generalization focuses on the representation of the intertemporal utility function. Some recursive

utility functions, which generalize the time consistency property of the time additive utility function, can be specified for concave production models while retaining the qualitative properties of optimal paths, such as capital monotonicity. The basic notion of a recursive utility function is illustrated below. The general theory of recursive utility functions is exposited by Becker and Boyd (1997).

Flexible time preference underlies many classic writings on capital theory – the agents discount factor depends on the underlying consumption stream. Recursive utility functions are one family of utilities that allow the steady state consumption stream to influence the corresponding discount factor. The brief development of recursive utility theory given here is grounded in a re-examination of the time consistency property of the planner's optimal choice in the one-sector discounted Ramsey model.

The discounted additive utility function, *U*, over infinite consumption streams $\mathbf{c} = \{c_1, c_2, \ldots\}$ is defined by the formula:

$$U(\mathbf{c}) = \sum_{t=1}^{\infty} \delta^{t-1} u(c_t)$$

where *u* is a bounded, strictly increasing, and strictly concave function on [0,∞) with 0< δ <1 as before. The time consistency property discussed above reflects the property that *U* is *recursive*: the behaviour embodied in this additive representation of utility has a self-referential property, that is, the behaviour of the planner over the infinite time horizon $t = 1, 2, \ldots$ is guided by the behaviour of that agent over the tail horizon $t = T, T+1, T+2, \ldots$ (for each *T*) hidden inside the original horizon. For this additive utility function, recursivity means the objective from time $T+1$ to+∞ has the same form as the objective starting at time $T = 0$ (except for some time shifts in consumption dates). Formally, *U* may be rewritten as:

$$\sum_{t=1}^{\infty} \delta^{t-1} u(c_t) = \sum_{t=1}^{T} \delta^{t-1} u(c_t) + \delta^T \sum_{t=T+1}^{\infty} \delta^{t-1} u(c_{t+T}),$$

where the last sum gives the utility of the stream $\{c_{T+1}, c_{T+2}, \ldots\}$. The utility of the consumption stream **c** can be written as the function:

$$U(\mathbf{c}) = u(c_1) + \delta U(S\mathbf{c}),$$

where *S* is the *shift operator*: $S\mathbf{c} = \{c_2, c_3, \ldots\}$. Let the *projection operator*, π, be defined by the formula $\pi\mathbf{c} = c_1$. The general notion of a recursive utility function is that the utility function *U* can be written in the form:

$$U(\mathbf{c}) = W(u(\pi\mathbf{c}), U(S\mathbf{c}))$$

for an appropriate real-valued function *W* defined on $[0, \infty) \times \mathcal{U}$, where $\mathcal{U}$ is the range of *U*. *W* is called the *aggregator function*. For the additive function, $W(c, v) = u(c) + \delta v$ for $v \in \mathcal{U}$. There are other examples of recursive utility functions. The Epstein–Hynes utility function developed below is generated by the *EH aggregator* $W(c, v) = (-1 + v)\exp(-v(c))$, where *v* is a strictly concave, increasing function of *c* with $v(0) > 0$.

The general theory of recursive utility functions provides a way to recover the utility function *U* from specification of the aggregator. Intuitively, *U* can be found by recursively substituting it into the equation $U(\mathbf{c}) = W(u(\pi\mathbf{c}), U(S\mathbf{c}))$. This substitution is performed by the *recursive operator* T_W defined by:

$$(T_W(U^0)(\mathbf{c}) = W(u(\pi\mathbf{c}), U^0(S\mathbf{c})),$$

where U^0 is considered the initial seed in this recursive substitution. For example, if $U^0 = 0$, the zero function that annihilates all consumption streams, then the Nth*–iterate* of T_W is:

$$(T_W^N 0) = W(c_1, W(c_2, \ldots, W(c_N, 0)\cdots)).$$

The recursive utility function is the unique fixed point of the operator T_W. The general theory provides conditions under which T_W has a unique fixed point and the successive iterates T_W^N converge to that fixed point independently of the choice of the initial seed function, U^0. Lucas and Stokey (1984) first proposed the specification of utility functions via aggregators and provided the basic theory of the recursion operator for bounded aggregators when consumption streams were elements of the set of all real-valued non-negative bounded sequences.

The basic ideas in recursive utility theory are readily illustrated for the case of the EH aggregator. This yields an example where the planner's utility function has flexible time preference and a recursive structure. A planner whose preferences over consumption streams is defined by the EH aggregator can be shown by recursive substitution to have the utility function *U*, which takes the form:

$$U(\mathbf{c}) = -\sum_{t=1}^{\infty} \exp\left(-\sum_{s=1}^{t} v(c_s)\right) \tag{6}$$

where $v: \mathbb{R}_+ \to \mathbb{R}_+$ is strictly concave, increasing, and satisfies $v(0) > 0$. Equation (6) is known as the *Epstein–Hynes (EH) utility function* after the continuous time analogue from Epstein and Hynes (1983); (6) was also studied in Epstein (1983). The EH utility from the consumption sequence's tail, $(c_{T+1}, c_{T+2}, \ldots)$, appears in the last term of the following expression breaking down the utility over the entire consumption path into segments for the first *T* periods and the subsequent periods:

$$-\sum_{t=1}^{\infty} \exp\left(-\sum_{s=1}^{t} v(c_s)\right) = -\sum_{t=1}^{T} \exp\left(-\sum_{s=1}^{t} v(c_s)\right) + \exp\left(-\sum_{\tau=1}^{T} v(c_s)\right) \times \left[-\sum_{t=T+1}^{\infty} \exp\left(-\sum_{\tau=T+1}^{t} v(c_s)\right)\right]$$

Hence, the utility of the tail of the program is just a time-shifted form of the utility of the original program – this is the identifying characteristic of a recursive utility function based on stationary preferences.
The steady state conditions for this economy are found by working out the no arbitrage conditions for the optimal growth problem which maximizes (6) subject to (2) and letting the consumption and capital sequences be constant sequences. Then the steady state conditions become:

$$f'(k^*) = 1 / \exp(v(c^*)), \tag{7}$$

where k^* is the aggregate steady state capital stock. Since $\exp(v(0)) > 1$ and $c^* + k^* = f(k^*)$, one can solve (7) for a unique long-run capital and consumption level. The capital monotonicity property holds for the optimal solution to the problem of maximizing (6) subject to (2) when the neoclassical production function satisfies the concavity and Inada conditions for the discounted Ramsey model (see Becker and Boyd, 1997, ch. 5, for a detailed proof and Beals and Koopmans, 1969, for the seminal article on recursive utility in optimal growth theory). In particular, if the initial capital stock is smaller than the steady state stock, then the economy's capital stock increases at each time and converges to the steady state; likewise, an initial capital stock above the steady state leads to a declining capital stock over time which converges to the steady state stock. The non-crossing property also obtains.

3.2 Equilibrium equivalence principles

The optimal growth model connects to the central questions of the determination of prices, including the rate of interest, and the functional distribution of income, by way of reinterpreting the optimal program as a competitive equilibrium for a fully specified dynamic general equilibrium model. This relationship is

obtained by proving a version of the fundamental welfare theorems for this economy. The traditional welfare theorems based on finitely many goods must be adapted to the case of infinitely many dated commodities. There is more than one way to interpret the equilibrium model. The first interpretation is one with perfect foresight and a sequence of budget constraints, one for each time. Prices are reckoned in units of current consumption. The second interpretation links the neoclassical model with Irving Fisher's theory of interest rate determination and emphasizes his famous *separation principle*. The Fisherian equilibrium model is also one where agents act with perfect foresight.

At the core of either equilibrium model's interpretation is what Christopher Bliss (1975) called the *orthodox vision of capital theory*: an economy accumulating capital will generate rising wages and a falling rate of interest. Since capital increases over time, labour–capital complementarity implies workers are more productive and their wage rises. Diminishing returns set in and the rental rate falls as so many early writers on capital theory hypothesized in their verbal models. One of Ramsey's great contributions was to provide a consistent mathematical model of this story.

3.2.1 The PFCE equivalence principle

The competitive economy consists of an infinitely lived representative *household*, or *consumer sector*, and a *production sector*. The representative household's preferences coincide with the Ramsey style planner introduced above. The representative household is derived for an economy with a continuum of identical infinitely lived households whose preferences coincide with the Ramsey style planner. These households' preferences and endowments are identical. The total labour supply of all households has unit mass. In a *symmetric equilibrium* each household will take the same action given the same endowment, so it is sufficient to examine the decisions undertaken by a representative household who is also taken as supplying the economy's labour services to the production sector. The production sector's production function is the same as the one in the corresponding optimal growth model.

The representative consumer forecasts sequences of rental and wage rates to maximize lifetime utility subject to a sequence of budget constraints, one for each period. Formally, the household sector solves for given $\{r_t, w_t\}_{t=1}^{\infty}$ the problem:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u(c_t)$$

by choice of the non-negative sequences $\{k_{t-1}, c_t\}_{t=1}^{\infty}$ subject to:

$$c_t + k_t = w_t + (1 + r_t)k_{t-1} \text{ for } t = 1, 2, \ldots \tag{8}$$

and $k_0 \leq k$. Here k is the initial capital stock (the same one as in the Ramsey optimal growth problem), r_t is the one-period rental rate on capital, and w_t is the wage rate earned by inelastically supplying one unit of labour in each time period. The prices r_t and w_t are reckoned in units of consumption available at time t. The consumer's problem has a no arbitrage condition analogous to the one obtained in the optimal growth problem:

$$\delta(1 + r_t)u'(c_{t+1}) = u'(c_t) \text{ for each } t.$$

The transversality condition is necessary for equilibrium programs as defined below. The combination of the transversality condition and the no arbitrage equation is also sufficient for a consumption–capital sequence to solve the consumer's problem for a given profile of wages and rental factors.
Producers take the rental rate as given and solve the following myopic maximization problem for the production sector's capital demand at each time period:

$$\sup_{x \geq 0} f(x) - (1 + r_t)x.$$

Here, x denotes a level of aggregate capital; the profit maximizing solution is denoted k_{t-1}, the planned capital demand at time t. It only depends on the current rental rate, r_t. The problem's point input–point output structure reflects the absence of adjustment costs or other structural production lags and the fact that all forward-looking consumption–investment decisions reside in the household sector. The necessary and sufficient condition for a positive capital stock to solve the production sector's optimization problem at time t is:

$$f'(k_{t-1}) = 1 + r_t,$$

which uniquely determines k_{t-1} in terms of $1 + r_t$. The *total capital income* is $(1 + r_t)k_{t-1} = f'(k_{t-1})k_{t-1}$.
The *wage bill* is the residual ‘profit’ given by:

$$w_t = f(k_{t-1}) - (1 + r_t)k_{t-1}.$$

Notice that $w_t = f(k_{t-1}) - f'(k_{t-1})k_{t-1}$. In the Cobb–Douglas case with $f(k) = k^{\rho}$, then this economy labour's share of the total output or national product, k^{ρ}_{t-1}, is $1-\rho$ and capital's share is ρ. The *total supply of goods in period t* is $f(k_{t-1})$ as a result of one-period profit maximization.

Sequences $\{1 + r_t, w_t, c_t, k_{t-1}\}_{t=1}^{\infty}$ constitute a *perfect foresight competitive equilibrium (PFCE)* provided that:

- **(PFCE-1)** $\{c_t, k_{t-1}\}_{t=1}^{\infty}$ solve the consumer's problem given $\{1 + r_t, w_t\}_{t=1}^{\infty}$;
- **(PFCE-2)** $f'(k_{t-1}) = 1 + r_t$, and
- **(PFCE-3)** $w_t = f(k_{t-1}) - (1 + r_t)k_{t-1}$ for each time *t*.

These three conditions yield via Walras's Law the materials balance condition, $c_t + k_t = f(k_{t-1})$ for each *t* and $k_0 = k$.

The *equivalence principle* tells us that for this dynamic economy the *PFCE* allocation is the same as the Ramsey planner's solution. Hence, a *PFCE* allocation is an optimum and vice versa. The argument is the no arbitrage conditions for the equilibrium and optimal growth problems coincide, and the respective transversality conditions hold as necessary conditions in their respective problems. The sufficiency of these conditions is used to finish the proof.

A PFCE determines the *functional distribution of income* as the payments to each productive factor at each point in time. Labour receives its wage and capital is paid its capital income. The share of income received by each factor is a constant and time independent when production is Cobb–Douglas. The functional distribution of income at each time also yields the representative agent's personal income by adding the two source's income at each time. Multi-agent models differentiate the personal income an agent enjoys at each time from the corresponding functional distribution of income.

3.2.2 The Fisher competitive equilibrium equivalence principle

The capital theoretic foundation for the present value investment criterion is the *Fisher separation principle* derived from Fisher's ‘second approximation’, which portrays the intertemporal consumption–investment decision of agents as a two-stage process. In the first stage, investment opportunities are exploited to realize a maximum value of initial wealth. The solution to the first-stage problem is found by maximizing the net present value over all feasible projects. Given competitive prices (and implicit discount rates), all agents whose intertemporal utility functions satisfy a mild non-satiation requirement will be led to choose the same wealth maximizing investment projects. In the second stage, those agents take their maximized wealth and access perfect capital markets to borrow and lend in order to obtain the most preferred lifetime consumption pattern.

The Fisher competitive equilibrium is the infinite horizon analogue of the Fisher separation principle. There is a single lifetime budget constraint; the savings–investment decision is separated from the consumption decision. Consumers maximize utility given their maximized wealth obtained as residual claimants to the production sector's discounted profit streams. Discounted profits are maximized within that sector. Letting $\{r_t\}$ be the sequence of interest rates and $q_t = \Pi_{\tau=1}^{t}(1 + r_\tau)^{-1}$ the discounted price of time *t* consumption, define the profit function by $\pi(k, \{r_t\}) = \max\{\Sigma_{t=1}^{\infty} q_t[f(k_{t-1}) - (1 + r_t)k_{t-1}] : k_0 = k\}$.

A sequence $\{r_t,c_t,k_t\}$ forms a Fisher competitive equilibrium (FCE) if:

- **(FCE-1)** $\pi(k, \{r_t\}) = \max\{\sum_{t=1}^{\infty} q_t[f(k_{t-1}) - (1+r_t)k_{t-1}] : k_0 = k\}$;
- **(FCE-2)** Consumers maximize $\sum_{t=1}^{\infty} \delta^{t-1} u(c_t)$ subject to the budget constraint $\sum_{t=1}^{\infty} q_t c_t = \pi(k, \{r_t\}) + k$;
- **(FCE-3)** The market clearing condition $c_t = f(k_{t-1}) - k_{t-1}$ holds.

Once again, by matching first-order conditions and transversality conditions the sufficiency conditions for the agents' optimization problems imply that the allocation $\{c_t,\bullet k_t\}$ in a FCE $\{r_t,\bullet c_t,\bullet k_t\}$ is an optimum, and vice versa: given the optimal allocation $\{c_t,\bullet k_t\}$, there is a sequence of interest rates such that the triple $\{r_t,\bullet c_t,\bullet k_t\}$ forms a FCE. The result is the Fisher equivalence theorem.

The twin equivalence theorems for the PFCE and FCE models connect Ramsey's theory of optimal growth in an aggregate economy to Fisher's theory of consumption and investment in an intertemporal choice market model as well as to Solow's descriptive growth theory (the logarithmic utility, Cobb–Douglas production function example has a constant marginal propensity to save, as assumed in Solow's growth model). The qualitative properties of the optimal growth model carry over to the two formulations of dynamic competitive economies. In the case where the initial capital stocks are smaller than the modified golden-rule stocks, the capital monotonicity property of the optimal program implies that the consumption sequence increases, the sequence of wage rates is increasing, and the sequence of interest rates/rental rates is decreasing. The orthodox vision of capital theory holds for the one-sector optimal growth model once the dynamic equilibrium is interpreted by way of the PFCE and FCE equivalence principles.

3.3 Many agents

The equivalence principles for the discounted Ramsey model postulate a representative agent. The orthodox vision of capital theory carries over to some forms of neoclassical capital theory when many distinct agents replace the assumption of a representative infinitely lived household. The introduction of many distinct consumers raises interesting questions concerning the determination of equilibrium prices and the distribution of personal (and factor) income both in short and long runs.

Frank Ramsey's seminal contribution to optimal growth also addressed the long-run, or steady state, distribution in a competitive economy. He conjectured that, with households having different rates of impatience, the steady state equilibrium would have very unequal income and wealth distributions. The most patient household would enjoy the maximum sustainable consumption (‘bliss’ in his conception) and all other households would consume at a minimal level necessary to sustain their lives. This was not a particularly new idea at the time his paper was published. The notion that time preference differences operating in a market economy might promote long-run differences in income and wealth can be found in the writings of such eminent economists as John Rae in 1834 and in several books by Irving Fisher beginning with his great work on the rate of interest first published in 1907. The Ramsey conjecture can be examined in two distinct neoclassical settings. The first deals with a natural extension of the optimal growth model to one of Pareto optimal growth. Agents are allowed to borrow and lend. The equilibrium version is analogous to the FCE set-up. Households have a single budget constraint expressed in present value terms. Here, long-run income distribution can be extreme if individuals have different discount factors – the

relatively impatient ones receive NO income. The second formulation is one of temporary equilibrium where markets are incomplete – households are forbidden to borrow against their future labour income (each person's capital stock is constrained to be non-negative at each time) and face a sequence of budget constraints, as in the PFCE model. In this setting, the relatively impatient households consume their wage income and the most patient household consumes wage and capital income – a modern formulation of Ramsey's two-class society.

3.3.1 Pareto optimal growth with many agents

Suppose there are *H* households $(h = 1, 2, \ldots, H)$ with one-period return functions u_h of the type met in the optimal growth setting. Let c_t^h denote agent *h*'s consumption at time *t* and suppose that each agent's discount factor is the same $\delta = \delta_h$ with $0 < \delta < 1$. Introduce *welfare weights* $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_H) \geq 0$ and $\sum_{h=1}^{H} \lambda_h = 1$. Given a weight vector λ, the Pareto optimal growth problem is to solve:

$$\sup \sum_{t=1}^{\infty} \sum_{h=1}^{H} \lambda_h [\delta^{t-1} u_h(c_t^h)] \tag{9}$$

$$\text{subject to } \left(\sum_{h=1}^{H} c_t^h\right) + k_t \leq f(k_{t-1}),\ t = 1, 2, \ldots, c_t^h,\ k_{t-1} \geq 0,\ k_0 \leq k,\ \ h = 1, 2, \ldots$$

The planner seeks a path of consumption for each person and an aggregate capital path satisfying the constraints with the maximum weighted discounted future utility. This problem can be rewritten in an interesting manner.

Given a weight vector λ, define on $\mathbb{R}_+$ the real-valued function u_λ^* as the following program's optimal value function:

$$u_\lambda^*(c) = \sup\left\{\sum_{h=1}^{H} \lambda_h u_h(c^h) : \sum_{h=1}^{H} c^h = c,\ c^h \geq 0\right\}. \tag{10}$$

If u_h is a concave, continuous, increasing function on $[0,\infty)$, and twice continuously differentiable function on $(0,\infty)$, then u_λ^* is concave, increasing in c, and continuously differentiable. Note that the Inada condition $u_h'(0) = +\infty$ and $\lambda_h>0$ imply $c^h>0$ in the solution to (10) whenever $c>0$. This also implies $u_\lambda^{*\prime}(0) = +\infty$ holds. Of course, if $\lambda_h=0$, then $c^h=0$ in the solution to (10).

The Pareto optimal growth model is then given by the classic discounted Ramsey model:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u_\lambda^*(c_t) \tag{11}$$

$$\text{subject to } c_t + k_t \le f(k_{t-1}),\ t = 1, 2, \ldots, c_t, k_{t-1} \ge 0,\ k_0 \le k.$$

This problem has a unique solution under our basic assumptions. The neoclassical optimal growth model's properties obtain for this Pareto optimal growth model: the optimal aggregate consumption and capital sequences are monotonic and converge to the modified golden-rule consumption, c^*, and capital, k^*. Notice that the steady state capital stock and aggregate consumption levels are independent of the welfare weights. However, given c^*, the steady state allocations to the various households do depend on those weights by way of the solution to (10) with $c=c^*$. Different weights will distribute the steady state aggregate consumption differently. Consumption is equally distributed in the steady state if and only if the welfare weights are equal with $\lambda h=1/H$. Along dynamic equilibrium paths aggregate consumption growth also implies each household's consumption grows provided that agent's welfare weight is positive.

The preservation of the capital monotonicity property in this Pareto optimal growth problem suggests that the orthodox vision applies to its equilibrium counterpart. It turns out that with many agents the form of the equivalence principle is more subtle than with a single, representative, agent. The essential issue is the same problem that arises with the classical welfare theorems in finite dimensional commodity spaces – a Pareto optimum may only be a competitive equilibrium with transfer payments. Once this problem is handled, the basic equivalence principles carry over to the many agent case provided all households discount future utility at the same rate. The orthodox vision prevails.

The orthodox vision's realization in the Pareto optimal growth problem with equal discount factors does not extend to a model with heterogeneous agents and distinct discount factors. In this case, the household with the largest discount factor is the most patient one. The modified golden-rule capital stock, k^*, is still well-defined. However, Le Van and Vailakis (2003) prove the Pareto optimal capital sequence initiated at k^* converges to it in the long-run — but it is not a constant sequence: if the economy starts with the stocks k^*, then it is optimal for the planner to deviate from those stocks and only return to them asymptotically. The

resulting optimal capital sequence cannot be monotonic, although the authors show it can be eventually monotonic. In part, this reflects the fact that the households enjoy time-varying consumption along their optimal path. The aggregate consumption levels change over time, but the first household emerges as the dominant consumer in the limit. The heterogeneous agent extension of the neoclassical representative agent theory does not exhibit the orthodox vision.

3.3.2 The Ramsey equilibrium model

The Ramsey equilibrium developed in Becker (1980) and reviewed in Becker (2006) interprets Ramsey's original long-run steady state conjecture with heterogeneous agents in a modern fashion. The basic model is developed for the case of agents with time additively separable utility functions with fixed discount factors. Each agent has a different discount factor, so one household is more patient than all the others. The technology is specified by a one-sector model with a single all-purpose consumption–capital good as before. The general complete market competitive one-sector model treats budget constraints as restricting the present value of an agent's consumption to be smaller than or equal to the agent's initial wealth defined as the capitalized wage income plus the present value of that person's initial capital. This allows us to interpret the choice of a consumption stream as if the agent were allowed to borrow and lend at market-determined present value prices subject to repaying all loans. Markets are complete – any intertemporal trade satisfying the present value budget constraint is admissible at the individual level. The Ramsey equilibrium model changes the budget constraint from a single one reckoned as a present value to a sequence, one for each period. Agents are forbidden to borrow against their future labour income, so they cannot capitalize the future wage stream into a present value. Markets are incomplete. It becomes crucial to track the evolution of each person's capital stock. This is unnecessary in the complete market models when all values entering the budget constraint are present values.

The incomplete market structure shows itself in an individual's budget constraint. At each time, a household's available income is derived from rental returns on its capital stocks, and its wage rate (all labour is alike and inelastically supplied). Expenditure at each time is for consumption goods and for capital goods to be carried over to the next period in order to earn rental income. The borrowing constraint takes the form of a non-negativity constraint on the capital stock holdings in each time period. The formal constraint is analogous to (8) with superscripts attached to individual consumption and capital holdings.

The heterogeneous discount factor, incomplete market economy, differs in another important respect: the operation of a borrowing constraint in the individual household problems also breaks the possibility of an equilibrium allocation arising as the economy's optimal allocation. The welfare maximization approach favoured in the complete market theory is inapplicable.

The Ramsey model has a unique stationary equilibrium in which only the most patient household has capital. That agent also enjoys a labour income. All other households consume their wages and own no capital. The model's dynamics have some distinctive features when compared with the capital and consumption monotonicity characteristic of the representative agent neoclassical model. The main results for the Ramsey equilibrium model appear in a series of papers beginning with Becker and Foias (1987). The survey article by Becker (2006) reviews those results as well as others in detail. Here, it is enough to note that the Ramsey equilibrium aggregate capital starting from an arbitrary distribution of initial capital stocks *eventually* has the capital monotonicity property in the case where the production function's elasticity of substitution is greater than or equal to 1, a condition satisfied by the Cobb–Douglas production function. In this case, the orthodox vision of capital eventually holds. If that elasticity of substitution condition fails,

then Becker and Foias showed it was possible for a two-period equilibrium cycle to exist; the orthodox vision necessarily fails.

3.4 Behavioural economics and quasi-geometric discounting

The discounted Ramsey model where the planner discounts future utilities at a constant rate is the fundamental dynamic model in macrodynamics and economic growth theory. The time consistency of the optimal plan, based on the stationarity of the planner's utility function (even in the general recursive case) has been questioned by behavioural economics researchers on the basis of experiments and empirical evidence. For example, Ainslie (1991, p. 334) states that a majority of adults report they would rather have $50 immediately rather than $100 in two years, but almost no one chooses $50 in four years instead of receiving $100 in six years. If these individuals have stationary preferences, the mere passage of four years calendar time should not change the ranking of $50 in year four to $100 in year six if $50 was preferred in the present to $100 in two years. Thus, Ainslie concludes these individuals are time-inconsistent in their intertemporal preference ranking. Ainslie, as well as many others (notably Laibson, 1997; also see the survey by Frederick, Loewenstein and O'Donoghue, 2002, for detailed summaries of the evidence and related references based on works by psychologists and economists) argue a different discounting function that describes real human behaviour better than the constant discounting model. The quasi-geometric discounting model developed below illustrates the simplest form of an alternative discounting function that these researchers argue better describes real human intertemporal choices. The quasi-geometric discounting function is an important example of the *hyperbolic discounting functions* appearing in behavioural discussions of time preference. The time preference reversals reported by Ainslie can be thought of as a criticism of standard discounted utility models in much the same way as the Allais paradox in risky choice experiments provides evidence against the expected utility model.

The standard constant discounting model's discounting function is $D(t)=\delta^{t-1}$, where $0<\delta<1$ is the discount factor and $t\geq 1$. The function D is also called the *exponential discount function*. The *quasi-geometric discounting* model posits a discounting function of the form $d(t)=\beta\,\delta^{t-1}$, where $\beta>0$ is a parameter. The case $\beta=1$ corresponds to the exponential discount function. If $\beta<1$, there is short-run impatience – the decision maker is willing to save in the future, just not in the present. If $\beta>1$, then there is short-run patience – the decision maker is more willing to consume in the future rather than the present. It is known from the fundamental paper by Strotz (1955) that, if a dynamic optimizing planner's discount factor does not have an exponential form, then the resulting optimal solution found from maximizing utility discounted to the present date will be time inconsistent. Thus, a planner solving the problem of maximizing the quasi-geometric utility function:

$$U(\mathbf{c}) = u(c_1) + \beta[\delta u(c_2) + \delta^2 u(c_3) + \cdots] \tag{12}$$

subject to (2) will exhibit time inconsistency. The solution $\{\bar{c}_t, \bar{k}_{t-1}\}_{t=1}^{\infty}$ so found will change if the

planner is able to re-optimize at time 2. That new solution $\left\{\bar{c}_t^{\#}, \bar{k}_{t-1}^{\#}\right\}_{t=2}^{\infty}$ will have the property that $\bar{k}_2 \neq \bar{k}_2^{\#}$ when $\bar{k}_1 = \bar{k}_1^{\#}$ expresses the initial condition for the second period's optimization problem. Put differently, unless the planner can credibly commit to implementing the solution found in the first period, the planner will make another choice of optimal plans once period 2 is attained than the one originally found at time 1. The time inconsistent solution found in period 1 is really not an optmium as the planner would not implement it when called on to do so in the absence of a credible commitment to that plan.

Phelps and Pollak (1968) proposed a different way to arrive at a solution to the problem of maximizing (12) subject to (2). Their approach recognizes the planner must correctly anticipate future actions. The choice of c_t at some future date t alters the planner's capital stock and impacts the choices of consumption levels for all times past t. These impacts must be somehow considered by the planner in the present when the optimal plan is determined.

Phelps and Pollak imagined the planner as really infinitely many planners, each a *generation* that lives, saves and consumes over just one period. The discount factor, δ, measures impatience; the parameter β reflects the degree to which the current generation values future generations' utility relative to their own utility. Perfect altruism corresponds to the case $\beta=1$ whereas imperfect altruism arises whenever $\beta<1$. Later writers, following Laibson (1997), interpreted the generations as different *selves*, one for each time period. In either interpretation, the planner acts as if there are really infinitely many selves in the infinite-horizon Ramsey-styled optimization problem. Phelps and Pollak go on to argue the Ramsey optimal growth problem should be considered as a game with the many selves as the players. A Nash equilibrium of this game constitutes a solution to the planner's problem in the sense that no self (or generation) can improve its payoff given the actions taken by future selves (generations). Modern game theory research published after Phelps and Pollak's article suggests that such a game might have many equilibrium points. One possibility is the Markov perfect equilibrium concept. A Markov perfect equilibrium is time consistent. At time t, no histories of past choices or measurement of the capital stock are assumed to matter for outcomes beyond the current value of the aggregate capital stock that is presented to the self active at that moment. Other equilibrium notions can be formulated to reflect the game's history as play unfolds over time. Trigger strategies provide one way to do this. Of course, a fundamental equilibrium existence question arises for Markov perfect equilibrium as well as those equilibrium concepts derived from the selves adopting trigger strategies.

A Markov perfect equilibrium is represented by a time independent capital policy function, $g(k)$, that the current self expects to govern all future selves' saving and capital accumulation decisions. In this way, the aggregate capital stock is expected to evolve according to the dynamical system $k_t=g(k_{t-1})$ with $k_0=k$, the capital stock endowment available at time 0. Note that this function depends only on the currently available capital stock. To solve the planner's quasi-geometric utility optimization problem is to find such a policy function. Recall that a policy function of this type characterized the solution to the canonical version of the discounted Ramsey model and reflected the underlying time consistency property of the planner's stationary utility function. It is also a Markov perfect equilibrium in the quasi-geometric case where $\beta=1$ and $u(c)=\ln c$ with $f(k) = k^{\alpha}$. Of course, a major technical problem is to show a Markov perfect equilibrium exists in models where $\beta \neq 1$. For the log utility, Cobb–Douglas production model, a Markov perfect equilibrium has been constructed in the quasi-geometric case with $\beta<1$ by Krusell, Kuruscu and Smith (2002). They showed that there is a Markov perfect equilibrium with policy function:

$$k' = g(k) = \frac{\beta\delta\alpha}{1 - \alpha\delta(1 - \beta)} k^{\alpha}, \tag{13}$$

a functional form that agrees with the canonical example's capital policy function when $\beta = 1$. Iteration of this capital policy function (13) from the given initial capital stock produces a monotonic aggregate capital sequence. The qualitative properties of this particular Markov perfect equilibrium in this parameterized quasi-geometric model is the same as the qualitative properties of the canonical discounted Ramsey model, even though the two models' quantitative properties differ. For example, the two models have different steady states. The similarity was noted in Barro's (1999) continuous time model; he dubbed this similarity an observational equivalence result as the two models could not be distinguished empirically on the basis of their qualitative features alone.

3.5 Efficient programs

Programs which are optimal for the discounted Ramsey model as well as its more general recursive utility formulations have an important efficiency property: there is no other feasible consumption sequence that provides more consumption in at least one period and as much in any other when compared with the optimal consumption path. This efficiency property can be studied in capital accumulation models in its own right as a minimal requirement for any reasonable objective function. Considered on its own, the efficiency criterion does not do much to single out a specific course of action for the planner. However, it can be used to eliminate some candidate optima without further reference to a specific welfare function. Moreover, examining efficient programs of consumption and capital accumulation can be undertaken in models with infinitely lived agents as well as models with finitely lived, overlapping generations where the economy evolves over an infinite horizon.

The interest in intertemporal efficiency stems from Malinvaud's (1953) seminal paper. He presented the first extension of Koopmans' activity analysis of efficient allocations in a static production world to an open-ended economy with a recursive technological structure, such as the aggregative one-sector model. He was also the first to recognize that the analog of Koopmans' profit conditions for characterizing an efficient program had to be supplemented in an infinite framework. This new terminal condition, the *transversality condition* (seen in the above discussion of the optimal growth model) was shown to be sufficient for an efficient program satisfying the profit conditions for an appropriate set of shadow prices.

Efficient programs are discussed below for the aggregative one-sector model. A sequence $\{c_t\}$ satisfying (2) for some capital stock sequence is *inefficient* if there is an alternative consumption program $\{c_t'\}$ satisfying (2) for some capital stock sequence that offers at least as much consumption in every period and more consumption in at least one period. A sequence $\{c_t\}$ satisfying (2) for some capital stock sequence is *efficient* if it is not inefficient. The *efficiency criterion* ranks programs as either efficient or inefficient. The planner's objective is to select an efficient program. The efficiency criterion presumes that consumption may never be satiated in any period. An infinite number of efficient programs exists in the discounted Ramsey model – for a fixed, finite, time period *T*, define a feasible program by consuming nothing for

periods $t = 0, 1, \ldots, T-1$, and letting the capital stock accumulate according to the difference equation $k_t = f(k_{t-1})$, with $k_0=k$. At time T, consume the resulting $f(k_{T-1})$ and set $k_T=0$. For each time after T, consume zero and accumulate no capital. Such a path is efficient. Since T is arbitrary, there are infinitely many efficient paths.

Efficient programs providing consumption in every period also exist. One important example is the path found by first solving for the combination of consumption and capital stock which maximizes stationary (or, sustainable) consumption. This program solves the problem $\max\{f(x) - x : x \in [0, b]\}$. The solution, denoted k^g, satisfies $f'(k^g) = 1$ and called the *golden-rule capital stock*; the corresponding *golden-rule consumption*, c^g, is defined by the relation $c^g = f(k^g) - k^g$. The interpretation is that if the economy's initial capital stock happens to equal the golden-rule stock, then it is efficient for the planner to choose this stock for all time and maintain the largest possible stationary consumption.

The golden-rule pair (c^g,•k^g) has an important relationship to the problem of characterizing efficient programs. The specific result is called the Phelps theorem (see Phelps, 1966, p. 59). It is a sufficient condition for an attainable path to be inefficient. A $\{c_t,k_t\}$ satisfying (2) also satisfies the *Phelps condition* if there is an $\varepsilon > 0$ and a natural number $T(\varepsilon)$ such that for all $t \geq T(\varepsilon)$, $k_t \geq k^g + \varepsilon$. The Phelps condition is equivalent to $\liminf_{t\to\infty} k_t > k^g$. Phelps' theorem states that a feasible program satisfying the Phelps condition is inefficient. In particular, the path of pure accumulation found by iterating $k_t = f(k_{t-1})$ for all t with $k_0=k$ is inefficient as this program converges to the maximum sustainable capital stock. Any feasible program for which the capital stocks converge to a stock larger than the golden-rule stock is also inefficient. Note that such a program would have the own rate of return, $f'(k_{t-1}) - 1 < 0$ for all t sufficiently large. In particular, this would imply $\prod_{t=1}^{T} f'(k_{t-1}) \to 0$ as $T\to\infty$. It turns out that this is a general property of inefficient programs, as shown by Cass (1972). Intuitively, these inefficient programs have shadow interest rates, $r_t = f'(k_{t-1}) - 1$ that are negative (no market mechanism is identified in this discussion, so the interpretation of $f'(k_{t-1}) - 1$ is provisionally made as a shadow price). It is reasonable then to presume that programs with positive shadow interest rates for all time are efficient. The precise criterion that is necessary and sufficient to characterize inefficient programs was identified by Cass (1972). He proved his result with additional curvature assumptions on the production function (which restrict the rate of change of capital's marginal product as capital accumulates, or decumulates) as well as assumed $f'(0) < \infty$. His theorem states that a feasible path is inefficient if and only if:

$$\sum_{t=1}^{\infty}\left[\prod_{s=1}^{t} f'(k_{s-1})\right] < \infty .$$

Notice that if a path satisfies this Cass condition, then $\prod_{s=1}^{t} f'(k_{s-1}) \to 0$ as $t\to\infty$, which is the Phelps sufficient criterion for inefficiency. Cass interpreted his condition as saying that the term $\prod_{s=1}^{t} f'(k_{s-1})$ goes to zero ‘sufficiently fast’. The term $\prod_{s=1}^{t} f'(k_{s-1})$ represents the shadow future value of a marginal

unit of capital in period 0. The Cass criterion's necessity then asserts that for an inefficient program, the future value of a marginal unit of capital at time 0 is bounded from above. This implies that the terms of trade from present to future never become very favorable (Cass, 1972, p. 207). General forms of the Cass criterion for one-sector models are discussed in the survey by Becker and Majumdar (1989) as well as additional applications to overlapping generations models and interpretations of these conditions for decentralized planning mechanisms. The survey by Tirole (1990) focuses on the connection between the Cass criterion for inefficient programs and the potential for the shadow prices associated with efficient programs to exhibit a type of bubble whereby the shadow market price of a unit of capital differs from its present discounted value of future shadow rental returns.

4 Controversies and critiques

Neoclassical capital theory has long been controversial. The famous *Cambridge Controversies* about whether or not the one-sector neoclassical model's properties were either sensible, or could be generalized, produced a substantial literature. See Birner (2002) for a thorough review of both sides' positions. Earlier references include Harcourt (1972), Bliss (1975), and Burmeister (1980). A few key points are noted here. The debates centred on whether or not there really is something called aggregate capital, whether or not it could be measured independently of the establishment of an equilibrium interest rate, and whether or not an increase in the steady state interest rate necessarily reduced steady state capital.

Bliss (1975) argued that aggregating capital was not more difficult than aggregating any other collection of commodities. It was enough to place a partial order on a vector of capital goods defining one vector of capital goods to be at least as much as another vector. Standard utility function existence theorems would imply the existence of a continuous, real-valued, order preserving functional representation that could be interpreted as an aggregate capital good. Burmeister (1980) gave conditions under which a generalized steady state regularity condition applied to a many capital goods model permitted theorists to construct an aggregate capital stock and aggregate production function with the desired neoclassical properties (at least across steady states). It should also be noted that there are models where there is a natural measure of an aggregate capital stock in physical terms. For example, the capital stock in renewable resource theories such as ones arising in fishery models measures the fish population as a *biomass*: the mass of living organisms present in a population at a particular point of time. Biomass can be measured as either a weight or as so many calories. Its measurement does not depend on any prices or other quantities that might be established only in an equilibrium. Of course, this is a special situation.

One practical way of arriving at a measure of aggregate capital is to compute its capital value. This can be done by multiplying the prices of the various underlying capital goods times their respective quantities. Presumably, these prices represent these capital goods' discounted future returns (for example, monetary or cash flows). Capitalization of future payments requires an interest rate (or a term structure of interest rates in case the rate of interest varies over time). It follows that capital value cannot be computed independently of the determination of prices. Critics of neoclassical theory stressed this issue. Modern equilibrium models establish the determination of capital goods prices and interest rates in an equilibrium configuration, for both the short and the long runs (this is one task solved by equivalence principles in many capital goods models, when those results are available).

The comparative steady state result for the one-sector neoclassical model is that the steady state capital stock, $k(\delta)$, viewed as a function of the discount (long-run interest) factor δ^{-1}, has the property $dk/d\delta > 0$.

The famous reswitching controversy attacks the generality of this result. In multi-sectoral models (even with aggregate capital) the choice of steady state production techniques can give rise to a particular capital–labour ratio arising from two different long-run interest rates.

The Cambridge controversies highlight the special features of the one-sector neoclassical theory. Those arguments concentrated on comparing steady states and either ignored or downplayed the role for transitions from one steady state to another in response to an exogenous change in an economy's deep taste or technology parameters. The debate also largely ignored the accumulation programs that flowed from the planner's decision when starting with initial capital other than the steady state level. The more dynamic view of modern capital theorists emphasizes the full dynamic possibilities open to the planner.

The orthodox vision applied to an aggregative economy portrays saving and consumption activities undertaken within the private sector as promoting a path of accumulation tending towards a steady state. When the economy's capital stock is initially smaller than its stationary level there is growth, and the rate of return on capital falls over time. This portrait of capital accumulation is consistent with the dynamics of the one-sector Ramsey optimal growth – perfect foresight equilibrium model provided there is a representative household whose preferences are taken as the planner's objective.

Bliss (1975) criticized the orthodox vision for models with many distinct capital goods as a single rate of interest could not be defined, and therefore the idea that growth accompanied a declining rate of interest made no sense. Subsequent research has shown that, even in aggregate capital Ramsey optimal growth models with a well-defined interest rate, the economy might not follow the orthodox vision provided there were at least two sectors producing a consumption good distinct from the capital good. The problem was that optimal cycles or even chaotic trajectories could emerge with a sufficiently impatient planner (see Boldrin and Woodford, 1990). Heterogeneous discount factor models also turn out to differ fundamentally from the representative agent theory, even in the classical one-sector case. The orthodox vision will only apply to *some* economies when there are heterogeneous discount factors.

The Cambridge controversy focuses on the difficulties of aggregating different types of capital and consumption goods. There are also difficulties inherent in interpreting results obtained for representative agent economies. The failure of the orthodox vision noted above is one such example. There is another, perhaps more fundamental, criticism of representative agent-based capital theories. The conditions under which the many different individuals populating a model economy's preferences might be aggregated so that the economic theorist can study the model as if there is a single, stand-in, representative agent are so restrictive as to make conclusions drawn from single agent models flawed on logical grounds alone. See Hartley (1997) for a detailed discussion of the representative agent controversy.

The idea of a representative agent economy such as the Ramsey model is that the aggregate activity in the economy generated by many different consuming and producing actors can be understood as the activity of a single entity, the representative agent, which acts exactly like each of the consuming and producing actors. By studying the microeconomic behaviour of those individuals we can also find the behaviour of the representative agent, and vice versa. However, the argument is made that, even if the microfoundations of each agent are well understood, it does not follow that their aggregate behaviour is explained by the representative agent that behaves exactly like them. Micro-behaviour need not translate into macro-behaviour of the same type. For example, the representative agent Ramsey model's capital monotonicity property holds up in the welfare optimum version of the many agent theory when agents have the same discount factors, but different one-period utility functions and possibly different initial capital stocks. The planner whose preferences are represented by the welfare function (11) does not give rise to the exact same behaviour as that of each of the individual agents' preferences underlying it – individual consumption

sequences differ from the aggregate, although they behave qualitatively the same (for example, they are monotonic). This distinction is even more pronounced in case agents also have different discount factors – the impatient agents' consumption tends to zero while the most patient one's consumption remains positive for all time. The aggregate consumption evolves over time in a very different manner from that of individual consumption streams.

5 Capital theory with many sectors and capital goods

Controversies surrounding the neoclassical capital theory of the one-sector model are partly attenuated by studying models with many sectors and types of capital goods. This general form of the theory emphasizes a disaggregated viewpoint, although it also applies to aggregative models. It should also be noted that specifying a multisector model need not be the same as formulating a many capital good model. There are two-sector models with aggregative capital and single-sector models with joint production of many distinct capital and consumption goods.

5.1 Pricing and the portfolio equilibrium condition

The major conceptual difference between the one-sector and multisector perfect foresight equilibrium models lies in the form taken by the no-arbitrage condition. This is readily seen in the two-sector model. Suppose there are two sectors consisting of a consumption goods sector and a capital goods sector. The capital and consumption goods are aggregate commodities, as in the one-sector model, but are conceived as distinct goods in the two-sector framework. Suppose that i_{t+1} is the one-period interest rate measured in units of a numeraire commodity, r_{t+1} is the rental rate on a unit of capital measured in the numeraire's units, and q_{t+1} is the unit purchase price for a unit of capital as measured in the numeraire's units. Suppose that the purchase of a unit of capital at time t entitles its owner to receive the rental flows from the next period on as long as the unit remains in service. Assume further that capital does not depreciate. One requirement for a perfect foresight equilibrium is that there are no one-period reversed arbitrage opportunities. Let an equilibrium path obtain with the prices $\{i_{t+1},r_{t+1},q_{t+1}\}$. Suppose the household decision maker acquires another unit of capital at time t. This costs the household q_t units of the numeraire. The opportunity cost of this action in the numeraire's units is $i_{t+1}q_t$, the interest charge that could have been earned otherwise. To reverse this capital acquisition at time t+1 the household will sell that unit of capital for q_{t+1} units of the numeraire. This gives the capital gain (loss) equal to $q_{t+1}-q_t$. The household also gets to keep the one-period rental, r_{t+1}. This one-period reversed arbitrage is unprofitable if the marginal revenue equals the marginal cost reckoned in units of the numeraire. That is,

$$i_{t+1}q_t = r_{t+1} + q_{t+1} - q_t. \tag{14}$$

This equation reflects the absence of arbitrage opportunities in a perfect foresight competitive equilibrium.

This *perfect foresight equation* is also called the *portfolio equilibrium condition* because it expresses the absence of arbitrage opportunities in the manner in which the agent's wealth is held. Rearranging this equation yields

$$i_{t+1} = \frac{r_{t+1}}{q_t} + \frac{q_{t+1} - q_t}{q_t}, \tag{15}$$

which says that the one-period interest rate, i_{t+1}, equals the capital good's *own rate of return*, r_{t+1}/q_t, plus the *capital gain yield*, $(q_{t+1}-q_t)/q_t$.

Note that $q_t=1$ holds in the one-sector model. This is the price of the consumption good in units of the numeraire commodity (chosen to be current consumption) since the capital and consumption goods are identical. Hence, there is no capital gain yield in that case and

$$i_{t+1} = r_{t+1}. \tag{16}$$

The interest rate equals the rental rate for capital goods. Thus, even if there is a single capital good, the portfolio equilibrium condition differs when the one-sector and two-sector models are compared.

Next, consider an aggregate model with an exhaustible resource. Suppose there are neither extraction nor storage costs. The aggregate capital stock at the end of time period t that is available for consumption at time $t+1$ is denoted by k_t and is interpreted as the amount of the resource remaining at the end of time t. Consumption at time t, c_t, represents a withdrawal from the stock k_{t-1}. Then the materials balance condition is $c_t+k_t=k_{t-1}$. The initial size of the resource stock is k. There is no rental return in this model; the resource owner's returns are entirely capital gain yields. The perfect foresight equation takes the form

$$i_{t+1} = \frac{q_{t+1} - q_t}{q_t}. \tag{17}$$

If the rate of interest is a constant: $i_{t+1}=r>0$, then (17) is a linear difference equation with solution $q_{t+1}=(1+r)^t q_0$, where q_0 is the resource's initial price. This implies *Hotelling's r-per cent rule* (Hotelling, 1931) holds in a perfect foresight equilibrium – the equilibrium (current) price of the resource, q_t, increases over time at rate of interest, r.

In models with several distinct capital goods the portfolio equilibrium condition applies to each capital good separately. If there are *m* capital goods, then the portfolio equilibrium condition takes the form:

$$i_{t+1} = \frac{r^j_{t+1}}{q^j_t} + \frac{q^j_{t+1} - q^j_t}{q^j_t}, \text{ for } j = 1, 2, \ldots, m. \tag{18}$$

Here, the superscript *j* labels capital good *j*. With many capital goods households have a variety of options for holding their wealth. The rates of return on any portfolio of capital stocks must be equalized or there will be a one-period reversed arbitrage opportunity. Hence, eq. (18) is the equilibrium condition expressing the absence of such arbitrage opportunities.

The major pricing differences between the one-sector and multisector models concern the form of the portfolio equilibrium condition. It is possible to develop equivalence principles for multisector models along the lines of the one-sector theory by making appropriate adjustments in the pricing of capital goods to reflect their multiplicity in the budget constraints and production sector while also recognizing the portfolio equilibrium form of the no-arbitrage conditions in the PFCE and FCE settings.

Establishing the formal equivalence between optimal accumulation models and their equilibrium counterparts in many capital good models requires the equilibrium economy to impose a transversality condition on itself, just as in the one-sector case. The general question is how is the initial price determined so that the equilibrium price profile satisfies the conditions for achievement of a Ramsey-styled central planning solution. This is the crux of the *Hahn problem*. The modern perfect foresight interpretation is that this problem is solved whenever a transversality condition obtains as necessary for an equilibrium. This requires the household sector to be forward looking over the infinite horizon, and markets to operate on all dates and for all commodities. Some writers on capital theory take a critical view of these conditions and argue that markets cannot be relied on to set the correct initial prices, and so the resulting equilibrium path is inefficient. On the other hand, a comparison of idealized markets with idealized planning, as embodied in the equivalence principles, suggests that at the most theoretical level the Hahn problem is resolved when rational, forward-looking agents conduct their economic activities in a complete market setting over an infinite horizon.

6 Final comments

The constraints of the neoclassical one-sector model can be used to substitute for consumption in the felicity function by noting $u(c_t) = u(f(k_{t-1}) - k_t)$, where $c_t \geq 0$ if and only if $f(k_{t-1}) - k_t \geq 0$. The current period's payoff depends only on the stocks of capital at the beginning and end of the period. This observation results in a reformulation of the one-sector model focused on the capital stock sequences. Let *u* (0)=0 to simplify the exposition. Let $D = \{(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+ : f(x) - y \geq 0\}$. Note that (0,0)$\in D$. The felicity function $v(x, y) \equiv u(f(k_{t-1}) - k_t)$ has domain *D* and v(0,0)=0. The properties of *u* and *f* imply that *v* is increasing in its first argument and decreasing in its second argument. The concavity of *u* and *f* also

imply that *v* is a concave function defined on the convex set *D*. The planner continues to discount future utility by the factor δ, $0<\delta<1$. This alternative representation of the neoclassical model, called the *reduced form model*, gives rise to an optimal growth problem with the planner choosing the sequence $\{k_t\}_{t=0}^{\infty}$ to achieve

$$\sup_{\{k_t\}_{t=0}^{\infty}} \sum_{t=1}^{\infty} \delta^{t-1} v(k_{t-1}, k_t);\ (k_{t-1}, k_t) \in D \text{ for each } t, \text{ and } 0 \le k_0 \le k. \tag{19}$$

This form of the one-sector model is just one realization of the general reduced form model. A complete exposition of this general structure's properties is found in McKenzie's surveys. The reduced form model can accommodate many varieties of capital theoretic problems including multisector and multi-capital good models, von Neumann's model of economic growth, exhaustible and renewable resource models, as well as individual firm investment theory when there are costs of adjusting the firm's capital stocks. The capital stocks of the one-sector model are replaced by a vector of capital stocks where each component represents a particular capital good; the set *D* is then contained in a multi-dimensional Euclidean space. Schefold (1997) is a recent treatment of multisector models derived from Sraffa's (1960) perspective on capital accumulation models that also revisits the reswitching controversy in a dynamic equilibrium setting. Also see Burmeister (1980) for a critical exposition of Sraffa's contribution. Burgstaller (1995) reviews models from the Sraffa tradition as well as neoclassical models in continuous time in order to find their common ground and connections to earlier capital theories.

The full scope of capital theoretic problems in deterministic, continuous time can be found in Weitzman (2003). The monograph by Becker and Boyd (1997) addresses the analogous problems in discrete time. Conrad and Clark (1987) covers natural resource models from a dynamic perspective. Stokey and Lucas (1989) provide an excellent introduction to stochastic dynamic models along with development of the discrete time theory using dynamic programming techniques. Chang (2004) presents basic continuous time stochastic calculus and optimal control theory with economic applications including the classical tree-rotation problem.

See Also

- capital theory (paradoxes)
- dynamic programming
- intertemporal equilibrium and efficiency
- neoclassical growth theory (new perspectives)
- present value
- Ramsey model

Bibliography

Ainslie, G. 1991. Derivation of ‘rational’ economic behavior from hyperbolic discount curves. *American Economic Review* 81, 334–40.

Barro, R.J. 1999. Ramsey meets Laibson in the neoclassical growth model. *Quarterly Journal of Economics* 114, 1125–52.

Beals, R. and Koopmans, T.C. 1969. Maximizing stationary utility in a constant technology. *SIAM Journal of Applied Mathematics* 17, 1001–15.

Becker, R.A. 1980. On the long–run steady state in a simple dynamic model of equilibrium with heterogeneous households. *Quarterly Journal of Economics* 95, 375–82.

Becker, R.A. 2006. Equilibrium dynamics with many agents. In *Handbook of Optimal Growth Theory, Volume I: Discrete Time Theory*, ed. C. Le Van, R.A. Dana, T. Mitra and K. Nishimura. New York: Springer-Verlag.

Becker, R.A. and Boyd, J.H. 1997. *Capital Theory, Equilibrium Analysis, and Recursive Utility*. Malden, MA: Blackwell Publishers.

Becker, R.A. and Foias, C. 1987. A characterization of Ramsey equilibrium. *Journal of Economic Theory* 41, 173–84.

Becker, R.A. and Majumdar, M. 1989. Optimality and decentralization in infinite horizon economies. In *Joan Robinson and Modern Economic Theory*, ed. G.R. Feiwel. London: Macmillan.

Birner, J. 2002. *The Cambridge Controversies in Capital Theory: A Study in the Logic of Theory Development*. London: Routledge.

Bliss, C.J. 1975. *Capital Theory and the Distribution of Income*. Amsterdam: North-Holland/America Elsevier.

Boldrin, M. and Woodford, M. 1990. Equilibrium in models displaying fluctuations and chaos: a survey. *Journal of Monetary Economics* 25, 189–222.

Brock, W.A. and Mirman, L.J. 1972. Optimal growth and uncertainty: the discounted case. *Journal of Economic Theory* 4, 479–513.

Burgstaller, A. 1995. *Property and Prices: Toward a Unified Theory of Value*. Cambridge: Cambridge University Press.

Burmeister, E. 1980. *Capital Theory and Dynamics*. Cambridge: Cambridge University Press.

Cass, D. 1972. On capital overaccumulation in the aggregative model of economic growth: a complete characterization. *Journal of Economic Theory* 4, 200–23.

Chang, F.R. 2004. *Stochastic Optimization in Continuous Time*. Cambridge: Cambridge University Press.

Conrad, J.M. and Clark, C.W. 1987. *Natural Resource Economics: Notes and Problems*. Cambridge: Cambridge University Press.

Dixit, A.K. 1976. *The Theory of Equilibrium Growth*. Oxford: Oxford University Press.

Dorfman, R., Samuelson, P.A. and Solow, R.M. 1958. *Linear Programs and Economic Analysis*. New York: McGraw-Hill.

Epstein, L.G. 1983. Stationary cardinal utility and optimal growth under uncertainty. *Journal of Economic Theory* 31, 133–52.

Epstein, L.G. and Hynes, J.A. 1983. The rate of time preference and dynamic economic analysis. *Journal of Political Economy* 91, 611–35.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Frederick, S., Loewenstein, G. and O'Donoghue, T. 2002. Time discounting and time preference: a critical review. *Journal of Economic Literature* 40, 351–401.

Goetzmann, W.N. and Rouwenhorst, K.G. 2005. *The Origins of Value: The Financial Innovations That Created Modern Capital Markets*. Oxford: Oxford University Press.

Harcourt, G.C. 1972. *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.

Hartley, J.E. 1997. *The Representative Agent in Macroeconomics*. London: Routledge.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–175.

Koopmans, T.C. 1958. *Three Essays on the State of Economic Science*. New York: McGraw–Hill.

Krusell, P., Kuruscu, B. and Smith, A.A. 2002. Equilibrium welfare and government policy with quasigeometric discounting. *Journal of Economic Theory* 105, 42–72.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112, 443–77.

Le Van, C. and Vailakis, Y. 2003. Existence of a competitive equilibrium in a one–sector growth model with heterogeneous agents and irreversible investment. *Economic Theory* 22, 743–71.

Lucas, R.E. and Stokey, N.L. 1984. Optimal growth with many consumers. *Journal of Economic Theory* 32, 139–71.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21, 233–68.

McKenzie, L.W. 1986. Optimal economic growth, turnpike theorems and comparative dynamics. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

McKenzie, L.W. 1987. Turnpike theory. In *The New Palgrave: A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Mirman, L.J. and Zilcha, I. 1975. Optimal growth under uncertainty. *Journal of Economic Theory* 11, 329–39.

Phelps, E.S. 1966. *Golden Rules of Economic Growth*. New York: Norton.

Phelps, E.S. and Pollak, R. 1968. On second-best national saving and game-equilibrium growth. *Review of Economic Studies* 35, 185–99.

Radner, R. 1961. Paths of economic growth that are optimal with regard only to final states. *Review of Economic Studies* 28, 98–104.

Rae, J. 1834. *Statements of Some New Principles on the Subject of Political Economy*. Reprinted 1964. New York: Augustus M. Kelley.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Schefold, B. 1997. *Normal Prices, Technical Change and Accumulation*. London: Macmillan.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Stokey, N.L. and Lucas, R.E., with Prescott, E. 1989. *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

Strotz, R.H. 1955. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 11, 165–80.

Tirole, J. 1990. Intertemporal efficiency, intergenerational transfers, and asset pricing: an introduction. In *Essays in Honor of Edmond Malinvaud, Volume I: Microeconomics*, ed. P. Champsaur et al. Cambridge, MA: MIT Press.

von Neumann, J. 1937. Über ein Ökonomisches Gleichungssystem une eine Verallgemeinerung des Brouwerschen Fixpunksatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, 73–83. Trans. in *Review of Economic Studies* 13(1945), 1–9.

Weitzman, M.L. 2003. *Income, Wealth, and the Maximum Principle*. Cambridge, MA: Harvard University Press.

How to cite this article

Becker, Robert A. "capital theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000575> doi:10.1057/9780230226203.0195

The New Palgrave Dictionary of Economics Online

capital utilization

Roger Betancourt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Utilization of capital can take place through variations in the duration of working time, given intensity, or through variations in the intensity of working time, given duration, or both. This article focuses on the economic factors determining duration and discusses the issues affecting and affected by variations in intensity. The latter can take the form of variations in speed or in the use of inputs that are complements to capital relative to some maximum or optimum. We provide a historical perspective, discuss modern theory, its main applications and links to the issues of speed and capacity, and identify important implications.

Keywords

agency costs; business cycles; capacity utilization; capital utilization; compensating differentials; depreciation; duality; duration; dynamic factor demand models; elasticity of substitution; exploitation; general equilibrium; incomplete contracts; Leontief production function; market imperfections; Marshall, A.; Marx, K. H.; monitoring; non-convexity; obsolescence; overtime; partial equilibrium; production function; shift work; shift differential; Solow residual; speed; specific-factors models; taxation of corporate profits; technical change; transaction costs; two-sector models; wear and tear; work day of capital; work day of labour

Article

Capital utilization is given different interpretations in the economic literature. If a machine is available for use during, say, a day, then various levels of utilization can be obtained by varying the duration of operations within the day. For any fixed duration within the day, however, it is also possible to vary the machine's rate of utilization by varying its speed. In each case there is variation in capital utilization, but both physical and economic characteristics differ widely in the two cases. Moreover, even with duration and speed constant within the day, some writers define variations in capacity utilization via variations in

the variable inputs employed with a given machine per day relative to some maximum or optimum daily output. Unfortunately, these as well as other writers frequently use the terms ‘capital utilization’ and ‘capacity utilization’ interchangeably.

The discussion here will focus on the analysis of variations in the duration of operations. A brief historical perspective sets the stage for a presentation of modern theory and applications, including links to the issues of speed and capacity. A succinct conclusion provides implications for closely related economic issues.

Historical perspective

Concern with the duration of operations dates to the late 18th century and the spread of the factory system in England. Early writing emphasized the appropriate length of the working day relative to its social consequence for workers and its economic consequence for capitalists. Positions on these issues were developed in the context of debates over the various Factory Acts in England. These discussions usually assumed the length of the working day to be the same for capital and labour.

Marx provides a most interesting example of the development of economic thinking on duration up to his time. The length of the working day is given substantial attention in his work (1867, ch. 10); indeed, it provides the cornerstone for his theory of exploitation (see, for example, Morishima, 1973, ch. 5); yet Marx pays only minor attention to the separation of capital's work day from labour's work day which is at the centre of modern analysis.

Marshall, like his predecessors, was interested in duration because of its implications for the well-being of workers and the viability of the economic system. But he saw the separation of the work day of labour from the work day of capital inherent in shift-work systems as an opportunity for resolving the conflicting interests of workers and capitalists with respect to the length of the work day. Thus he becomes an advocate of the adoption of multiple shifts early in his professional career (1873) and maintained his interest in the topic throughout his career (see, for example, 1923, p. 650).

Marshall's emphasis became the basis for the work of Robin Marris (1964), who treats capital utilization as a synonym for shift-work. Interestingly enough, the other modern pioneer, Georgescu-Roegen (for example, 1972), stresses the choice of the daily duration of operations, acknowledges Marx's emphasis on the topic, but overlooks Marshall as well as Marris. Both view the choice of duration at the plant level, either directly or through the selection of a shift-work system, as a long-run or *ex ante* decision, that is, before the plant is built. Moreover, both assume the *ex post* elasticity of substitution to be zero, that is, within the day no variations in choice of technique are allowed once the factory is built. However, while Marris uses discrete techniques of production and discrete systems of utilization to describe the structure of the firm's optimization problem, Georgescu-Roegen uses a continuous production function and a continuous index of the daily duration of operations; these differences of method do not generate substantial differences in results.

Both economists use their analyses to argue against anachronistic social legislation and draw implications from their work for an important contemporary economic problem, namely, the improvement of economic conditions in developing countries.

Before presenting the modern theory and its applications it is useful to note a few salient facts. Thanks to Foss's efforts (1981) there are reliable estimates of the average workweek of capital (plant hours) in US

manufacturing for 1929 and 1976 – 67 and 82 hours, respectively. These estimates can be compared to an average workweek for labour of 50 hours in 1929 and 40 hours in 1976. Furthermore, Foss views the rise in capital's workweek between 1929 and 1976 as an underestimate of the increase in shift-work, because of the decrease in the number of days worked per week during this same period. The most thorough update of this data work is Beaulieu and Mattey (1998). It generates an average workweek of capital for manufacturing during the period 1974–92 of 97 hours per week. These ‘facts’ underlie interest in the topic and the frequent identification of capital utilization with shift-work.

Modern theory and applications

A number of contributions have incorporated the choice of duration into the neoclassical theory of the firm. This work is most concisely exposited using a model which relies on duality theory to generate the main results available in this literature (see Betancourt, 1986).

The firm's optimization problem is viewed as a two-stage procedure. In the first stage the decision-maker generates a cost function for each given level of duration; in the second stage the decision-maker selects from these cost functions that one which leads to least total cost. The end result in the two-input case is:

$$C^* = dC(w^*, r^*, x^*). \tag{1}$$

For a given reference unit of duration, w^* represents the average wage rate, r^* the price of capital services, x^* the level of output, while d represents an index of duration of operations, C is a classical cost function, and C^* represents the total cost of operations at the optimal level of duration.

For example, if an eight-hour shift starting during normal hours is the reference unit of duration, as duration increases beyond this reference period: the average wage rate (w^*) increases because of shift differentials due to workers' preferences for normal hours or social legislation; and the price of capital services per eight-hour shift decreases, although there will be two opposite tendencies in this case. The daily price of a unit of capital increases due to the additional wear and tear created by the longer duration, but this price is now spread over a greater number of hours, and the price of capital services per eight-hour shift (r^*) decreases. Betancourt and Clague (1981, ch. 2, sect. 2) provide a detailed discussion of why the second effect predominates. Finally, as duration increases, the same daily output is spread over a greater number of hours, and the level of output per eight-hour shift (x^*) decreases.

The formulation in (1) yields the main insights about capital utilization or shift-work at the plant level offered by the early literature that followed Georgescu- Roegen and Marris. A brief listing of these results is as follows: (i) high shift differentials or overtime rates discourage capital utilization by increasing w^*; (ii) technologies with high degrees of returns to scale discourage utilization by raising the costs of operating at low levels of output (x^*); (iii) technologies with high degrees of capital intensity encourage capital utilization because the consequent fall in the relevant cost of capital (r^*) affects a

higher percentage of costs; and (iv) technologies with abundant *ex ante* substitution possibilities encourage utilization because they lower the costs of taking advantage of the consequent fall in the cost of capital (r^*) through the building of a more capital intensive factory. These four factors are the main long-run determinants of optimal duration on the cost side.

In addition, two other characteristics of the utilization decision are worth stating. First, factories built to operate at high levels of utilization will be designed to use capital-intensive techniques. Second, how exogenous changes in input costs affect duration depends critically on the *ex ante* elasticity of substitution. For instance, if this elasticity is greater than unity, under constant returns to scale an exogenous fall in the price of capital lowers the costs of building the plant to operate longer hours.

One application of the model is as the theoretical basis for empirical studies of the choice of duration at the plant level. The model's implications were consistent with several different bodies of plant level data (see Betancourt and Clague, 1981, chs 4–8) across non-continuous process industries. Recent work using more detailed plant level data for specific industries, for example automobiles, confirms the role of the number of shifts as a long-run margin of adjustment and it stresses the importance of changes in duration through overtime and daily closings as short-run margins of adjustment in the United States (Bresnahan and Ramey, 1994). Detailed studies of the auto industry for Europe and Japan (Anxo et al., 1995, chs 12 and 13, respectively) are also consistent with this long-run role for the number of shifts. Mayshar and Halevy (1997) develop a model that allows for *ex pos*t substitution possibilities as a short-run margin of adjustment. The above studies imply that there is a choice of duration, even in the short run, but in some industries continuous processes dominate and the choice is really to operate or not operate the process. A major extension of the model that captures this feature is provided by Das (1992), who develops and estimates a discrete dynamic programming model for the cement industry at the kiln level. In this context a plant is basically an additive collection of kilns and Das allows for three decisions, namely, operate, retire or keep idle a kiln in any plant.

Alternative approaches to the non-convexities that arise at the plant level have been developed by looking at the industry as the unit of analysis. Prucha and Nadiri (1996) provide an insightful and sophisticated example of this option applied to the US electrical machinery industry by making endogenous the capital utilization decision in the context of dynamic factor demand models. In a similar industry setting, Cardellichio (1990) uses the assumption of Leontief production functions at the mill level to analyse utilization for the lumber industry as a whole.

From a theoretical perspective an application of the model in (1) has been as the basis for the choice of duration in standard two-sector general equilibrium models. In the context of the international trade literature, Betancourt, Clague and Panagariya (1985), for example, use the specific-factors model with variable utilization to reconcile the dual scarcity explanation of Anglo-American trade in the 19th century with the empirical evidence on observed utilization levels. In the context of the public finance literature Coates (1991) generalizes the standard analysis of the incidence of the corporate profits tax by allowing for variable utilization. He concludes that overestimates of the burden of the tax in the order of 10–60 per cent are most probable as a result of ignoring this long-run margin of adjustment in a general equilibrium context. A more abstract general equilibrium approach allowing for firm's decisions over duration and starting times as well as for worker's preferences over these work schedules has been developed recently by Garcia Sanchez and Vazquez Mendez (2005). Its main substantive result replicates one partial equilibrium result noted above, namely, that high capital intensity in the form of a

high capital–labour ratio leads to an increase in utilization.

A short-run perspective has played an important role in dramatizing the policy implications of high levels of utilization for employment and output, since in this perspective a doubling of utilization implies a doubling of employment and output. Nevertheless a long-run perspective (see Betancourt and Clague, 1981, chs 9–11) provides a far less optimistic view about the likelihood of these outcomes. Ironically the evaluation of a shorter workweek for labour in Europe, which is analytically similar, has been carried out primarily from a short-run perspective (for example, Anxo et al, 1995, ch. 14). Garcia Sanchez and Vazquez Mendez (2005), however, suggest this topic as one for potential application of their long-run model.

Related issues: speed and capacity

The relations between duration, speed and capacity are difficult to analyse and provide an opportunity for confusion. To start, consider a dual representation of the cost function in (1). Namely,

$$x = dF(K, L) \tag{2}$$

where x is the level of daily output, that is, $x=dx^*=dF$; F is a neoclassical production function defined over the reference period of duration; K represents both the level of the capital stock employed and the rate of capital services, which implies that the speed of operations (v) is constant and set at unity; and L represents labour services per reference period of duration. Alternatively, those who analyse variations in utilization through choice of speed represent the productive process as follows:

$$x = F(vK, L) \tag{3}$$

where all variables have been previously defined. In (3) duration is set at unity.

Writers who employ (3) assume that the price of the capital stock is an increasing function of speed or utilization (for example, Smith, 1970). Since costs are defined as $C=r(v)K+wL$, where $r'(v)>0$, the cost of a unit of capital services obtained by increasing speed is an increasing function of v. While in the duration model the price of the capital stock $r(d)$ is an increasing function of duration ($r'(d)>0$), the cost of a unit of capital services obtained by increasing duration is a decreasing function of duration, that is, $r^*=r(d)/d$ and $r^{*\prime}(d)<0$. This difference implies that models with one utilization variable to describe the productive process can generate nonsensical economic results if this variable is interpreted as representing either duration or speed, because the behaviour of

costs can represent only one of the two interpretations. To illustrate, a recent body of literature relates capital utilization, economic growth and the speed of convergence (for example, Chatterjee, 2005), by assuming depreciation to increase with utilization at an increasing rate. This makes sense if one justifies increases in utilization as a result of increases in speed. Yet this literature justifies increases in utilization as a result of increases in duration through increases in the average workweek of capital.

Another interesting feature of the ‘speed’ model stems from the first-order conditions for cost minimization, which can be used to show that, if v, K and L are treated as choice variables, at the optimum, $r(v)=r'(v)v$. When duration and speed are endogenous this characteristic generalizes to $r(v, d) = r_v(v, d)\, v$ and optimal speed is determined by optimal duration (Madan, 1987). This is consistent with the finding by Bresnahan and Ramey (1994) for the auto industry that line speed and the number of shifts are long-run margins of adjustment.

Consider now the representation of the productive process underlying the typical definitions of capacity utilization. Namely,

$$x = F(K, L) \tag{4}$$

where all variables are defined as before and speed and duration set at unity. Using (4), Panzar's (1976) definition of capacity becomes:

$$h(K) = \max_{L} F(K, L) \tag{5}$$

where $h(K)$ is an increasing function of K. This definition leads to an output-based definition of short-run capacity utilization; that is:

$$CU = x / x\max \tag{6}$$

where x max is given by (5).

When capital equipment is capacity-rated in terms of output units, as in electricity generation, one can measure directly the denominator of (6) and short-run capital and capacity utilization coincide (cf. Winston, 1982, ch. 5). In general, however, the denominator in (6) is not well defined. An alternative

procedure is to define the denominator in (6) as the optimal level of output, x^0. For instance, in the literature on dynamic factor demand models x^0 is defined as the optimal level of output when the capital stock is endogenous (for example, Morrison, 1985; also see Prucha and Nadiri, 1996, for a generalization). Since 'optimal' output varies with the specification of the optimization problem, one can generate a variety of reasonable definitions of capacity utilization which measure different concepts. Not surprisingly, the corresponding empirical definitions fail to move together (de Leeuw, 1979) or with the average workweek of capital (Beaulieu and Mattey, 1998).

Implications

Perhaps the most important economic implication of the analysis of capital utilization above is for our understanding of technical change at the aggregate level. Ignoring increases in duration understates the contribution of capital services to output growth and, thus, overstates the estimates of technical change or the Solow residual in standard sources of growth analysis. Beaulieu and Mattey's estimate of the annual rate of growth in the average workweek of capital for manufacturing over the 1974–91 period is 0.17. They use employment per shift as weights, which are the appropriate ones, and find that only 25 per cent of the variation in growth can be accounted for by overtime.

Macroeconomists have pursued this issue but emphasized its business cycle implications. That is, when the Solow residual is adjusted for the workweek of capital it ceases to be pro-cyclical. For instance, Shapiro (1993) made this point in a widely cited paper. His results continued to hold in Beaulieu and Mattey's more recent data and they have given rise to a substantial literature that we will not explore here. One implication of this finding noted by Shapiro is that it casts doubts on alternative explanations of the behaviour of the residual stressing market power when there are substantial costs to adjusting the workweek of capital, for example through the shift differential.

There is an early literature on the human costs of shift-work which may be captured through the shift differential. Betancourt and Clague (1981, ch. 12) conclude from their review of this literature that observed shift differentials of four to five per cent in the United States substantially underestimate the human costs of shift-work. This conclusion is consistent with estimates in an unpublished paper by Shapiro (1995) that the marginal shift premium is 25 per cent. A strand of literature in labour economics on compensating differentials has considered shift-work. Kostiuk (1990) obtains estimates of the shift differential of well above ten per cent in the unionized sector for both 1979 and 1985. He relies on Census of Population Survey data for his analysis.

An issue neglected in the recent literature is the role of obsolescence in capital utilization. Marris (1964) argued that an increase in the rate of obsolescence should strengthen the economic incentive for shift-work, since it ameliorated disincentive effects of wear and tear depreciation. In the last few decades we have observed systematic shifts from mechanical technologies to electronic technologies, which diminish wear and tear costs and increase the rate of obsolescence. This shift should, thus, have provided an incentive for increased capital utilization. Yet, to my knowledge, the economic literature has not addressed this issue explicitly.

Finally, an important reason for interest in capital utilization as an economic variable is the existence of transaction costs and market imperfections. These frictions make ownership of capital equipment and structures attractive relative to rentals for instantaneous capital services. Of course these rental markets

do not exist in most cases. A substantial recent literature in industrial organization investigates the effect of transaction costs, including incompleteness of contracts and agency costs, on incentives and the evolution of institutions. With one exception, it has not addressed the impact of changes in transaction costs and market imperfections on capital utilization. The exception is the work of Hubbard (2003) on the trucking industry. He shows that improvements in monitoring technology in the form of on board computers increase capacity utilization, which in this industry coincides with short-run capital utilization just as in the electricity generation industry. Issues of long-run capital utilization and relevance for other industries, however, remain unexplored in this context.

See Also

- adjustment costs
- fixed factors
- labour market institutions

Bibliography

Anxo, D., Bosch, G., Bosworth, D., Cette, G., Sterner, T. and Taddei, D. 1995. *Work Patterns and Capital Utilization: An International Comparative Study*. London: Kluwer Academic Publications.

Beaulieu, J. and Mattey, J. 1998. The workweek of capital and capital utilization in manufacturing. *Journal of Productivity Analysis* 10, 199–203.

Betancourt, R. 1986. A generalization of modern production theory. *Applied Economics* 18, 915–28.

Betancourt, R. and Clague, C. 1981. *Capital Utilization: A Theoretical and Empirical Analysis*. New York: Cambridge University Press.

Betancourt, R., Clague, C. and Panagariya, A. 1985. Capital utilization and factor specificity. *Review of Economic Studies* 52, 311–29.

Bresnahan, T. and Ramey, V. 1994. Output fluctuations at the plant level. *Quarterly Journal of Economics* 108, 593–624.

Cardellichio, P. 1990. Estimation of production behavior using pooled micro data. *The Review of Economic and Statistics* 72, 11–18.

Chatterjee, S. 2005. Capital utilization, economic growth and convergence. *Journal of Economic Dynamics and Control* 29, 2093–124.

Coates, D. 1991. Endogenous capital utilization and taxation of corporate capital. *National Tax Journal*

44, 79–91.

Das, S. 1992. A micro-econometric model of capital utilization and retirement: the case of the US cement industry. *Review of Economic Studies* 59, 277–97.

de Leeuw, F. 1979. Why capacity utilization rates differ. In *Measures of Capacity Utilization: Problems and Tasks*, ed. F. de Leeuw et al. Staff Studies No. 105. Washington, DC: Board of Governors of the Federal Reserve System.

Foss, M. 1981. Long-run changes in the workweek of fixed capital. *American Economic Review, Papers and Proceedings* 71, 58–63.

Garcia Sanchez, A. and Vazquez Mendez, M. 2005. The timing of work in a general equilibrium model with shiftwork. *Investigaciones Economicas* 29, 149–79.

Georgescu-Roegen, N. 1972. Process analysis and the neoclassical theory of the firm. *American Journal of Agricultural Economics* 54, 279–94.

Hubbard, T. 2003. Information, decisions, and productivity: on-board computers and capacity utilization in trucking. *American Economic Review* 93, 1328–53.

Kostiuk, P. 1990. Compensating differentials for shift-work. *Journal of Political Economy* 98, 1054–75.

Madan, D. 1987. Optimal duration and speed in the long run. *Review of Economic Studies* 54, 695–700.

Marris, R. 1964. *The Economics of Capital Utilization*. Cambridge: Cambridge University Press.

Marshall, A. 1873. The future of the working classes. In *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan, 1925.

Marshall, A. 1923. *Industry and Trade*. 4th edn. Reprints of Economic Classics, New York: Augustus M. Kelley, 1970.

Marx, K. 1867. *Capital*, vol. 1, New York: Vintage Books; Random House, 1979.

Mayshar, J. and Halevy, Y. 1997. Shiftwork. *Journal of Labor Economics* 15, S198–S222.

Morishima, M. 1973. *Marx's Economics*. Cambridge: Cambridge University Press.

Morrison, C. 1985. On the economic interpretation and measurement of optimal capacity utilization with anticipatory expectations. *Review of Economic Studies* 52, 295–310.

Panzar, J. 1976. A neoclassical approach to peak load pricing. *The Bell Journal of Economics* 7, 521–30.

Prucha, I. and Nadiri, M. 1996. Endogenous capital utilization and productivity measurement in dynamic factor demand models. *Journal of Econometrics* 71, 343–79.

Shapiro, M. 1993. Cyclical productivity and the workweek of capital. *American Economic Review* 83, 229–33.

Shapiro, M. 1995. Capital utilization and the marginal premium for work at night. Mimeo, University of Michigan.

Smith, K. 1970. Risk and the optimal utilization of capital. *Review of Economic Studies* 37, 253–9.

Winston, G. 1982. *The Timing of Economic Activity*. Cambridge: Cambridge University Press.

How to cite this article

Betancourt, Roger. "capital utilization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000043> doi:10.1057/9780230226203.0197

The New Palgrave Dictionary of Economics Online

capitalism

Robert L. Heilbroner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Capitalism is a unique historical formation with core institutions and distinct movements. It involves the rise of a mercantile class, the separation of production from the state, and a mentality of rational calculation. Its characteristic logic revolving around the accumulation of capital reflects the omnipresence of competition. It displays broad tendencies to unprecedented wealth creation, skewed size distributions of enterprise, large public sectors, and cycles of activity. Whereas students of capitalism traditionally envisaged an end to the capitalist period of history, modern economists show little interest in historical projection.

Keywords

Baran, P.; Bentham, J.; Braudel, F.; capital; capital accumulation; capitalism; commodities; comparative systems approach; competition; economic freedom; employment contract; Engels, F.; English Revolution; exchange value; feudalism; French Revolution; Hirschman, A.; ideology; inequality; Keynes, J.; labour; labour power; Locke, J.; Marshall, T.; Marx, K.; mercantilism; Mill, J. S.; morality; Myrdal, G.; Polanyi, K.; political economy; political freedom; pre-capitalist social formations; private property; profit; property rights; public expenditure; public sector; quasi-rent; religion; Ricardo, D.; role of the state; Schumpeter, J.; Smith, A.; socialism; surplus value; use value; Veblen, T.; Weber, M.

Article

Capitalism is often called *market society* by economists, and the *free enterprise system* by business and government spokesmen. But these terms, which emphasize certain economic or political characteristics, do not suffice to describe either the complexity or the crucial identificatory elements of the system. Capitalism is better viewed as a historical ‘formation’, distinguishable from formations that have preceded it, or that today parallel it, both by a core of central institutions and by the motion these institutions impart to the whole. Although capitalism assumes a wide variety of appearances from period

to period and place to place – one need only compare Dickensian England and 20th-century Sweden or Japan – these core institutions and distinctive movements are discoverable in all of them, and allow us to speak of capitalism as a historical entity, comparable to ancient imperial kingdoms or to the feudal system.

The most widely acknowledged achievement of capitalist societies is their capacity to amass wealth on an unprecedented scale, a capacity to which Marx and Engels paid unstinting tribute in *The Communist Manifesto*. It is important to understand, however, that the wealth amassed by capitalism differs in quality as well as quantity from that accumulated in precapitalist societies. Many ancient kingdoms, such as Egypt, displayed remarkable capacities to gather a surplus of production above that needed for the maintenance of the existing level of material life, applying the surplus to the creation of massive religious or public monuments, military works or luxury consumption. What is characteristic of these forms of wealth is that their desirable attributes lay in the specific use-values – war, worship, adornment – to which their physical embodiments directly gave rise. By way of decisive contrast, the wealth amassed under capitalism is valued not for its specific use-values but for its generalized exchange-value. Wealth under capitalism is therefore typically accumulated as *commodities* – objects produced for sale rather than for direct use or enjoyment by their owners; and the extraordinary success of capitalism in amassing wealth means that the production of commodities makes possible a far greater expansion of wealth than its accumulation as use-values for the rulers of earlier historical formations.

Both Smith and Marx stressed the importance of the expansion of the commodity form of wealth. For example, Smith considered labour to be ‘productive’ only if it created goods whose sale could replenish and enlarge the national fund of capital, not when its product was intrinsically useful or meritorious. In the same fashion, Marx described the accumulation of wealth under capitalism as a circuit in which money capital (M) was exchanged for commodities (C), to be sold for a larger money sum (M′), in a never-ending metamorphosis of M–C–M′.

Although the dynamics of the M–C–M′ process vary greatly depending on whether the commodities are trading goods or labour power and fixed capital equipment, the presence of this imperious internal circuit of capital constitutes a prime identificatory element for capitalism as a historical genus. As such, it focuses attention on two important aspects of capitalism. One of these concerns the motives that impel capitalists on their insatiable pursuit. For modern economists the answer to this question lies in ‘utility maximization’, an answer that generally refers to the same presumed attribute of human nature as that which Smith called the ‘desire of bettering our condition’. The unappeasable character of the expansive drive for capital suggests, however, that its roots lie not so much in these conscious motivations as in the gratification of unconscious drives, specifically the universal infantile need for affect and experience of frustrated aggression. Such needs and drives surface in all societies as the desires for prestige and for personal domination. From this point of view, capitalism appears not merely as an ‘economic system’ knit by the appeals of mutually advantageous exchange, but as a larger cultural setting in which the pursuit of wealth fulfils the same unconscious purposes as did the pursuit of military glory or the celebration of personal majesty in earlier epochs. Such a description conveys the force of the ‘animal spirits’ (as Keynes referred to them) that both set into motion, and are appeased by, the M–C–M′ circuit. (Heilbroner, 1985, ch. 2; Sagan, 1985, chs 5, 6).

A second general question raised by the centrality of the M–C–M′ circuit concerns the manner in which the process of capital accumulation organizes and disciplines the social activity that surrounds it. Here analysis focuses on the institutions necessary for the circuit to be maintained. The crucial capitalist

institution is generally agreed to be private property in the means of production (not in personal chattels, which is found in all societies). The ability of private property to organize and discipline social activity does not however lie, as is often supposed, in the right of its owners to do with their property whatever they want. Such a dangerous social licence has never existed. It inheres, rather, in the right accorded its owners to withhold their property from the use of society if they so wish.

This negative form of power contrasts sharply with that of the privileged elites in precapitalist social formations. In these imperial kingdoms or feudal holdings, disciplinary power is exercised by the direct use or display of coercive force, so that the bailiff or the seneschal are the agencies through which economic order is directly obtained. The social power of capital is of a different kind – a power of refusal, not of assertion. The capitalist may deny others access to his resources, but he may not force them to work with them. Clearly, such power requires circumstances that make the withholding of access an act of critical consequence. These circumstances can only arise if the general populace is unable to secure a living unless it can gain access to privately owned resources or wealth. Capital thus becomes an instrument of power because its owners can establish claims on output as their *quid pro quo* for permitting access to their property.

Access to property is normally attained by the relationship of ‘employment’ under which a labourer enters into a contract with an owner of capital, usually selling a fixed number of working hours in exchange for a fixed wage payment. At the conclusion of this ‘wage-labour’ contract both parties are quit of further obligation to one another, *and the product of the contractual labour becomes* the property of the employer. From this product the employer will pay out his wage obligations and compensate his other suppliers, retaining as a profit any residual that remains.

In detail, forms of profit vary widely, and not all forms are specific to capitalism – trading gains, for example, long predate its rise. Explanations of profit vary as a consequence, but as a general case it can be said that all profits depend ultimately on inequality of economic position. When the inequality arises from wide disparities of knowledge or access to alternative supplies, profits typically emerge as the mercantile gains that were so important in the eyes of medieval commentators, or as the depredations of monopolistic companies against which Adam Smith inveighed. When the inequality stems from differentials in the productivity of resources or productive capability we have the quasi-rents to which such otherwise different observers as Marshall and Schumpeter attribute the source of capitalist gain. And when the inequality is located in the market relationship between employer and worker it appears as the surplus value central to Marxian and, under a different vocabulary, to classical political economy. As Smith put it, ‘Many a workman could not subsist a week, few could subsist a month, and scarce any a year without employment. In the long-run the workman may be as necessary to his master as his master is to him; but the need is not so immediate’ (Smith [1776] 1976, p. 84).

This is not the place to enter into a discussion of these forms of profit, all which can be discerned in modern capitalist society. What is of the essence under capitalism is that gains from whatever origin are assigned to the owners of capital, not to workers, managers or government officials. This is a clear indication both of the difference of capitalism from, and its resemblance to earlier social formations. The difference is that product itself now flows to owners of property who have already remunerated its producers, not to its producers – usually peasants in precapitalist societies – who must then ‘remunerate’ their lords. The resemblance is that both arrangements channel a social surplus into the hands of a superior class, a fact that again reveals the nature of capitalism as a system of social domination, not merely of rational exchange.

Thus we can see that the successful completion of the circuit of accumulation represents a political as well as an economic challenge. The attainment of profit is necessary for the continuance of capitalism not alone because it replenishes the wherewithal of each individual capitalist (or firm) but because it also demonstrates the continuing validity and vitality of the principle of M–C–M′ as the basis on which the formation can be structured. Profit is for capitalism what victory is for a regime organized on military principles, or an increase in the number of adherents for one built on a proselytizing religion.

The evolution of capitalism

Capitalism as a ‘regime’ whose organizing principle is the ceaseless accumulation of capital cannot be understood without some appreciation of the historic changes that bring about its appearance. In this complicated narrative it is useful to distinguish three major themes. The first concerns the transfer of the organization and control of production from the imperial and aristocratic strata of precapitalist states into the hands of mercantile elements. This momentous change originates in the political rubble that followed the fall of the Roman empire. There merchant traders established trading niches that gradually became loci of strategic influence, so that a merchantdom very much at the mercy of feudal lords in the 9th and 10th centuries became by the 12th and 13th centuries an estate with a considerable measure of political influence and social status. The feudal lord continued to oversee the production of the peasantry on his manorial estate, but the merchant, and his descendant the guild master, were organizers of production in the towns, of trade between the towns and of finance for the feudal aristocracy itself.

The transformation of a merchant estate into a capitalist class capable of imagining itself as a political and not just an economic force required centuries to complete and was not, in fact legitimated until the English revolution of the 17th and the French revolution of the 18th centuries. The elements making for this revolutionary transformation can only be alluded to here in passing. A central factor was the gradual remonetization of medieval European life that accompanied its political reconstitution. The replacement of feudal social relationships, mediated through custom and tradition, by market relationships knit by exchange worked steadily to improve the wealth and social importance of the merchant against the aristocrat. This enhancement was accelerated by many related developments – the inflationary consequence of the importation of Spanish gold in the 16th century, which further undermined the rentier position of feudal lords; the steady stream of runaway serfs who left the land for the precarious freedom of the towns and cities, placing further economic pressure on their former masters; the growth of national power that encouraged alliances between monarchs and merchants for their mutual advantage; and yet other social changes (see Pirenne, 1936; Hilton, 1978).

The overall transfer of power from aristocratic to bourgeois auspices is often subsumed under the theme of the rise of market society; that is, as the increasingly *economic* organization of production and distribution through purchase and sale rather than by command or tradition. This economic revolution, from which emerge the ‘factors of production’ that characterize market society, must however be understood as the end product of a *political* convulsion in which one social order is destroyed to make way for a new one. Thus the creation of a propertyless waged labour force – the prerequisite for the appearance of labour-power as a commodity that would become enmeshed in the M–C–M′ circuit – is a disruptive social change that begins in England in the late 16th century with the dispossession of peasant occupants from communal land and does not run its course until well into the 19th century. In similar fashion, the transformation of feudal manors from centres of social and juridical life into real

estate, or the destruction of the protected guilds before the unconstrained expansion of nascent capitalist enterprises, embody wrenching socio-political dislocations, not merely the smooth diffusion of pre-existing economic relations throughout society. It is such painful rearrangements of power and status that underlay the ‘great transformation’ out of which capitalist market relationships finally arise (Polanyi, 1957, Part II).

A second theme in the historical evolution of capital emphasizes a related but distinct aspect of political change. Here the main emphasis lies not so much in the functional organization of production as in the separation of a traditionally seamless web of rulership, extending over all activities within the historical formation, into two realms, each concerned with a differentiated part of the whole. One of these realms involved the exercise of the traditional political tasks of rulership – mainly the formation and enforcement of law and the declaration and conduct of war. These undertakings continued to be entrusted to the existing state apparatus which retained (or regained) the monopoly of legal violence and remained the centre of authority and ceremony. The other realm was limited to the production and distribution of goods and services; that is, to the direction of the material affairs of society, from the marshalling of the workforce to the amassing and use of the social surplus. In the fulfilment of this task, the second realm also extended its reach beyond the boundaries of the territorial state, insofar as commodities were sold to and procured from outlying regions and countries that became enmeshed in the circuit of capital.

The formation of these two realms was of epoch-making importance for the constitution of capitalism. The creation of a broad sphere of social activity from which the exercise of traditional command was excluded bestowed on capitalism another unmistakable badge of historic specificity; namely, the creation of an ‘economy’, a semi-independent state within a state and also extending beyond its borders. This in turn brought two remarkable consequences. One of these was the establishment of a political agenda unique to capitalism, in which the relationship of the two realms became a central question around which political discussion revolved, and indeed continues to revolve. In this discussion the overarching unity and mutual dependency of the two realms tends to be overlooked. The organization of production is generally regarded as a wholly ‘economic’ activity, ignoring the political function performed by the wage–labour relationship in disciplining the workforce in lieu of bailiffs and seneschals. In like fashion, the discharge of political authority is regarded as essentially separable from the operation of the economic realm, ignoring the provision of the legal, military and material contributions without which the private sphere could not function properly or even exist. In this way, the presence of two realms, each responsible for part of the activities necessary for the maintenance of the social formation, not only gives to capitalism a structure entirely different from that of any precapitalist society but also establishes the basis for a problem that uniquely preoccupies capitalism; namely, the appropriate role of the state vis-à-vis the sphere of production and distribution.

More widely recognized is the second major effect of the division of realms in encouraging economic and political freedom. Here the capitalist institution of private property again takes centre stage, this time not as a means of arranging production or allocating surplus, but as the shield behind which designated personal rights can be protected. Originally conceived as a means for securing the accumulations of merchants from the seizures of kings, the rights of property were generalized through the market into a general protection accorded to all property, including not least the right of the worker to the ownership of his or her own labour-power.

Now the wage–labour relationship appears not as means for the subordination of labour but for its

emancipation, for the crucial advance of wage-labour over enslaved or enserfed labour lies in the right of the working person to deny the capitalist access to labour-power on exactly the same legal basis as that which enables the capitalist to deny the worker access to property. There is, therefore, an institutional basis for the claim that the two realms of capitalism are conducive to certain important kinds of freedom, and that a sphere of market ties may be necessary for the prevention of excessive state power. This is surely an important part of Smith's celebration of the society of ‘natural liberty’, and has been the basis of the general conservative endorsement of capitalism. Unquestionably, the greatest achievements of human liberty thus far attained in organized society have been achieved in certain advanced capitalist societies. One cannot, however, make the wider claim that capitalism is a sufficient condition for freedom, as the most cursory survey of modern history will confirm.

A third theme in the evolution of capitalism calls attention to the cultural changes that have accompanied and shaped its institutional framework. Much emphasis has been given to this theme in the work of Weber and Schumpeter, both of whom stress the historic distinctions between the essentially rational – that is, means-ends calculating – culture of capitalist civilization compared with the ‘irrational’ cultures of previous social formations. Here it is important to recognize that rationality does not refer to the *principle* of capitalism, for we have seen that the impetus to amass wealth is only a sublimation of deeper-lying non-rational drives and needs, but to the behavioural paths followed in the pursuit of that principle. The drive to amass capital can be analysed in terms of a calculus that is less readily apparent, if indeed present at all, in the search for other forms of prestige and power. This pervasive calculating mind-set is itself the outcome both of the abstract nature of exchange-value, which makes possible commensurations that cannot be carried out in terms of glory or sheer display, and of the pressures exerted by the marketplace, which penalize economic actors who fail to follow the arrow of economic advantage. Capitalism is therefore distinguishable in history by the predominance of a prudent, accountant-like comparison of costs and benefits, a perspective discoverable in the mercantile pockets of earlier formations but highly uncharacteristic of the tempers of their ruling elites (see Weber, 1930; Schumpeter, 1942, ch. XI).

The cultural change associated with capitalism goes further, however, than the rationalization of its general outlook. Indeed, when we examine the general culture of capitalist life we are most forcibly struck by an aspect that precedes and underlies that highlighted above. This is the presence of an ideological framework that contrasts sharply with that of pre-capitalist formations. I do not use the word *ideology* in a pejorative sense, as denoting a set of ideas foisted on the populace by a ruling order in order to manipulate it, but rather as a set of belief systems to which the ruling elements of the society themselves turn for self-clarification and explanation. In this sense, ideology expresses what the dominant class in a society sincerely believes to be the true explanations of the questions it faces.

That which is characteristic of the ideologies of earlier formations is their unified and monolithic character. In the ancient civilizations of which we know, an all-embracing world view, usually religious in nature, explicates every aspect of life, from the workings of the physical universe, through the justification of rulership, down to the smallest details of social routines and attitudes. By way of contrast, the ideology that emerges within capitalism is made up of diverse strands, more of them secular than religious and many of them in some degree of conflict with other strands. By the end of the 18th century, and to some degree before, the explanation system to which capitalist societies turn with respect to the workings of the universe is science, not religious cosmology. In the same manner, rulership is no longer regarded as the natural prerogative of a divinely chosen elite but perceived as ‘government’; that

is, as the manner in which ‘individuals’ create an organization for their mutual protection and advancement. Not least, the panorama of work and the patterns of material life are perceived not as the natural order of things but as a complex web of interactions that can be made comprehensible through the teachings of political economy, later economics. The individual threads of these separate scientific, political-individualist and economic belief systems originate in many cases before the unmistakable emergence of capitalism in the 18th century, but their incorporation into a skein of culture provides yet another identifying theme of the history of capitalist development.

Within this skein, the ideology of economics is obviously of central interest for economists. A crucial element of this belief system involves changes in the attitude towards acquisitiveness itself, above all the disappearance of the ancient concern with good and evil as the most immediate and inescapable consequence of wealth-gathering. As Hirschman has shown, this change was accomplished in part by the gradual reinterpretation of the dangerous ‘passion’ of avarice as a benign ‘interest’, capable of steadying and domesticating social intercourse rather than disrupting and demoralizing it (Hirschman, 1977). Other crucial elements of understanding were provided by Locke's brilliant demonstration in *The Second Treatise on Government* (1690) that unlimited acquisition did not contravene the dictates of reason or Scripture, and by the full pardon granted to wealth-seeking by Bentham, who demonstrated that the happiness of all was the natural outcome of the self-regarding pursuit of the happiness of each. The problem of good and evil was thus removed from the concerns of political economy and relegated to those of morality; and economics as an inquiry into the workings of daily life was thereby differentiated from earlier inquiries, such as the reflections of Aristotle or Aquinas, by its explicit disregard of their central search for moral understanding. Perhaps more accurately, the constitution of a ‘science’ of economics as the most important form of social self-scrutiny of capitalist societies could not be attempted until moral issues, which defied the calculus of the market, were effectively excluded from the field of its investigations.

The logic of the system

This conception of capitalism as a historical formation with distinctive political and cultural as well as economic properties derives from the work of those relatively few economists interested in capitalism as a ‘stage’ of social evolution. In addition to the seminal work of Marx and the literature that his work has inspired, the conception draws on the writings of Smith, Mill, Veblen, Schumpeter and a number of sociologists and historians, notable among them Weber and Braudel. The majority of present-day economists do not use so broad a canvas, concentrating on capitalism as a market system, with the consequence of emphasizing its functional rather than its institutional or constitutive aspects.

In addition to the characteristic features of its institutional ‘nature’, capitalism can also be identified by its changing configurations and profiles as it moves through time. Insofar as these movements are rooted in the behaviour-shaping properties of its nature, we can speak of them as expressing the logic of the system, much as conquest or dynastic alliance express the logic of systems built on the principle of imperial rule, or the relatively changeless selfvreproduction of primitive societies expresses the logic of societies ordered on the basis on kinship, reciprocity and adaptation to the givens of the physical environment.

The logic of capitalism ultimately derives from the pressure exerted by the expansive M–C–M′ process, but it is useful to divide this overall force into two categories. The first of these concerns the

‘internal’ changes impressed upon the formation by virtue of its necessity to accumulate capital – its metabolic processes, so to speak. The second deals with its larger ‘external’ motions – changes in its institutional structure or in important indicia of performance as the system evolves through history.

The internal dynamics of capitalism spring from the continuous exposure of individual capitals to capture by other capitalists. This is the consequence of the disbursement of capital-as-money into the hands of the public in the form of wages and other costs. Each capitalist must then seek to win back his expended capital by selling commodities to the public, against the efforts of other capitalists to do the same. This process of the enforced dissolution and uncertain recapture of money capital in the circuit of accumulation is, of course, the pressure of competition that is the social outcome of generalized profit-seeking. We can see, however, that competition cannot be adequately described merely as the vying of suppliers in the marketplace. As both Marx and Schumpeter recognized, competition is at bottom a consequence of the mutual encroachments bred by the capitalist drive for expansion, not of the numbers of firms contending in a given market.

The process of the inescapable dissolution and problematical recapture of individual capitals now gives rise to the activities designed to protect these capitals from seizure. The most readily available means of self-defence is the search for new processes or products that will yield a competitive advantage – the same search that also serves to facilitate the expansion of capital through the development of new markets. Competition thus reinforces the introduction of technological and organizational change into the heart of the accumulation process, usually in two forms: attempts to cheapen the cost of production by displacements of labour by machinery (or of one form of fixed capital by another); or attempts to gain the public's purchasing power by the design of wholly new forms of commodities. As a consequence, one of the most recognizable attributes of capitalist ‘internal’ dynamics has been its constant revolutionizing of the techniques of production and its continuous commodification of material life, the sources of its vaunted capacity to change and elevate living standards.

A further internal change also arises from the expansive pressures of the core process of capital accumulation. This is a threat to the capacity as a whole to extract a profit from the production of commodities. This tendency arises from the long-run effect of rising living standards in strengthening the bargaining power of labour versus capital. There is no way in which individual enterprises can ward off this threat by cutting wages, for in a competitive market system they would thereupon lose their ability to marshall a workforce. Their only protection against a rising tendency of the wage level is to substitute capital for labour where that is possible. For the system as a whole, the need to hold down the bargaining power of labour must therefore hinge on a generalization of individual cost-reducing efforts, through the system-wide displacement of labour by machinery, or by the direct use of government policies to maintain a profit-yielding balance between labour and capital, or by systemic failures – ‘crises’ – that create generalized unemployment. Whether attempted by deliberate policy or brought about by the outcome of spontaneous market forces, the pressure to secure a profit-compatible level of wages thus becomes a key aspect in the internal dynamics of the system.

A final attribute of the internal logic of capitalism must also be traced to its core process of accumulation. This is the achievement of a highly adaptive method of matching supplies against demands without the necessity of political intervention. This cybernetic capacity is surely one of the historical hallmarks of capitalism, and is regularly emphasized in the ‘comparative systems approach’ in which the responsive capacities of the market mechanism are compared with the inertias and rigidities of systems in which tradition or command (planning) must fulfil the allocational task. A critique of the

successes and failures of the market system cannot be attempted here. Let us only emphasize that the workings of the system itself derive from institutional attributes whose genesis we have already observed – namely, the establishment of free contractual relations as the means for social coordination; the establishment of a social realm of production and distribution from which government intervention is largely excluded; the legitimation of acquisitive behaviour as the social norm; and activating the whole, the imperious search for the enlargement of exchange-value as the active principle of the historical formation itself.

Large-scale tendencies

From the metabolism of capitalism also emerges its larger ‘external’ motions – the overall trajectory often described as its macroeconomic movement, and the configurational changes that are the main concern of institutional economics. It may be possible to convey some sense of these general movements if we note three general aspects characteristic of them.

We have already paid heed to the first of these, the tendency of the capitalist system to accumulate wealth on an unparalleled scale. Some indication of the magnitude of this process emerges in the contrast between the increase in per capital GNP of developed (capitalist) and less-developed (noncapitalist) countries:

Table 1 GNP per capita (1960 dollars and prices)

	Presently developed countries	Presently less-developed countries
Around 1750	$180	$180–90
Around 1930	780	190
Around 1980	3,000	410

Source: Paul Bairoch in Faaland (1982), p. 162.

After our lengthy discussion of the central role of accumulation within capitalism it does not seem necessary to relate this historic trend to its institutional base. Two somewhat neglected aspects of the overall increase in wealth seem worth mentioning, however. The first is that the increase in per capita GNP includes both augmentations in the volume of output and an extension of the M–C–M′ process itself within the social world. This is manifested in a continuous implosion of the accumulation process within capitalist societies – the process of the commodification of material life to which we earlier referred – and its explosion into neighbouring noncapitalist societies.

This explosive thrust calls attention to the second attribute of the overall expansion of wealth. It is that capital, as such, knows no national limits. From its earliest historic appearance, capital has been driven to link its ‘domestic’ base with foreign regions or countries, using the latter as suppliers of cheap labour-power or cheap raw materials or as markets for the output of the domestic economy. The consequence has been the emergence of self-reinforcing and cumulative tendencies towards strength at the centre, to which surplus is siphoned, and weakness in the periphery, from which it is extracted. The economic dimensions of this global drift are immediately visible in the previous table. This is the basis for what has been called the ‘development of underdevelopment’ as the manner in which ancient patterns of international hegemony are expressed in the context of capitalist relationships (Myrdal, 1957, Part I: Baran, 1957, chs V–VII).

We turn next to a different overall manifestation of the larger logic of capitalist development – its changes in institutional texture. There have been, of course, many such changes in the long span of Western capitalist experience – indeed, it is the very diversity of the faces of capitalism that prompted our search for its deep-lying identifying elements. Nonetheless, two changes deserve to be singled out, not only because of their sweeping magnitude and transnational occurrence, but because they have deeply altered the evolutionary logic of the system itself. These have been the emergence within all modern capitalisms of highly skewed size distributions of enterprise, and of very large and powerful public sectors.

The general extent of these transformations is sufficiently well known not to require detailed exposition here. Suffice it to illustrate the trend by contrasting the largely atomistic composition of manufacturing enterprise in the United States at the middle of the 19th century with the situation in the 1980s, when seven-eighths of all industrial sales were produced by 0.1 per cent of the population of industrial firms. The enlargement of the public sector is not so dramatic but is equally unmistakable. During the present century in the United States, its size (measured by all government purchases of output plus transfer payments) has increased from perhaps 7.5 per cent of GNP to over 35 per cent, a trend that is considerably outpaced by a number of European capitalisms.

The first of these two large-scale shifts in the configuration can be directly traced to the pressures generated by the M–C–M′ circuit. The change from a relatively homogeneous texture of enterprise to one of extreme disparities of size is the consequence not only of differential rates of growth of different units of capital, but of defensive business strategies of trustification and merger, and the winnowing effect of economic disruptions on smaller and weaker units of capital. There is little disagreement as to the endemic source of this transformation in the dynamics of the marketplace and the imperative of business expansion.

The growth of large public sectors is not so immediately attributable to the accumulation process proper but rather results from changes in the logic of capitalist movements after the concentration of industry has taken place. Here the crucial change lies in the increasing instability of the market mechanism, as its constituent parts cease to resemble a honeycomb of small units, individually weak but collectively resilient, and take on the character of a structure of beams and girders, each very strong but collectively rigid and interlocked. It seems plausible that this rigidification was the underlying cause of the increasingly disruptive nature of the crises that appeared first in the late 19th century and climaxed in the great depression of the 1930s; and it is widely accepted that the growth of the public sector mainly owes its origins to efforts to mitigate the effects of that instability or to prevent its recurrence.

This brings us to the last general aspect of capitalist development; namely, the tendency for interruptions and failures to break the general momentum of capital accumulation. Perhaps no aspect of the logic of capitalism has been more intensively studied than these recurrent failures in the accumulation process. In the name of stagnation, gluts, panics, cycles, crises and long waves a vast literature has emerged to explain the causes and effects of intermittent systematic difficulties in successfully negotiating the passage from M to M′. The variables chosen to play strategic roles in the explanation of the phenomenon are also widely diverse: the saturation of markets; the undertow of insufficient consumption; the technological displacement of labour; the pressure of wages against profit margins; various monetary disorders; the general ‘anarchy’ of production; the effect of ill-considered government policy, and still others.

Despite the variety of elements to which various theorists have turned, a common thread unites most of

their investigations. This is the premise that the instabilities of capitalist growth originate in the process of accumulation itself. Even theorists who have the greatest confidence in the inherent tendency of the system to seek a steady growth path, or who look to government intervention (in modern capitalism) as the main instability-generating force, recognize that economic expansion tends to generate fluctuations in the rate of growth, whether from the ‘lumpy’ character of investment, volatile expectations, or other causes. In similar fashion, economists who stress instability rather than stability as the intrinsic tendency of the system do not deny the possibility of renewed accumulation once the decline has performed its surgical work; indeed, Marx, the most powerful proponent of the inherently unstable character of the M–C–M′ process, was the first to assert that the function of crisis was to prepare the way for a renewal of accumulation.

In a sense, then, the point at issue is not whether economic growth is inherently unstable, but the speed and efficacy of the unaided market mechanism in correcting its instability. This ongoing debate mainly takes the form of sharp disagreements with respect to the effects of government policy in supplementing or undermining the corrective powers of the market. The failure to reach accord on this issue reflects more than differences of informed opinion with regard to the consequences of sticky wages or prices, or ill-timed government interventions, and the like. It should not be forgotten that, from the viewpoint of capitalism as a regime, interruptions pose the same threats as did hiatuses in dynastic succession or breakdowns of imperial hegemony in earlier formations. It is not surprising, then, that the philosophic predilections of theorists play a significant role in their diagnoses of the problem, inclining economists to one side or the other of the debate on the basis of their general political sympathies with the regime, rather than on the basis of purely analytic considerations.

Periodization and prospects

All the foregoing aspects of the system can be traced to its inner metabolism, the money–commodity–money circuit. This is much less the case when we now consider the overarching pattern of change described by the configuration of the social formation as a whole as it moves from one historic ‘period’ to another.

Traditionally these periods have been identified as early and late mercantilism; pre-industrial, and early and late industrial capitalism; and modern (or late, or state) capitalism. These designations can be made more specific by adumbrating the kinds of institutional change that separate one period from another. These include the size and character of firms (trading companies, putting-out establishments, manufactories, industrial enterprises of increasing complexity); methods of engaging and supervising labour (cottage industry through mass production); the appearance and consolidation of labour unions within various sectors of the economy; technological progress (tools, machines, concatenations of equipment, scientific apparatus); organizational evolution (proprietorships, family corporations, managerial bureaucracies, state participation). David Gordon has coined the term ‘social structure of accumulation’ to call attention to the changing framework of technical, organizational and ideological conditions within which the accumulation process must take place. Gordon's concept, applied to the general problem of periodization, emphasizes the manner in which the accumulation process first exploits the possibilities of a ‘stage’ of capitalism, only to confront in time the limitations of that stage which must be transcended by more or less radical institutional alterations (Gordon, 1980).

The idea of an accumulation process alternately stimulated and blocked by its institutional constraints

provides an illumining heuristic on the intraperiod dynamics of the system, but not a theory of its long-run evolutionary path. This is because not all national capitalisms make the transitions with equal ease or speed from one social structure to another, and because it is not apparent that the pressures of the M–C–M′ process push the overall structure in any clearly defined direction. Thus Holland at the end of the 17th century failed to make the leap beyond mercantilism, and England in turn in the second half of the 19th century failed to create a successful late industrial capitalism. In this regard it is interesting that the explanatory narratives of the great economists apply with far greater cogency to the evolutionary trends within periods than across them – Smith's scenario of growth in *The Wealth of Nations*, for instance, containing no suggestion that the system would move into an industrial phase with quite different dynamics, or Marx's depiction of the laws of motion of the industrialized system containing no hint of its worldwide evolution towards a state-underwritten structure. Although the inner characteristics of the M–C–M′ process enable us to apply the same generic designation of capitalism to its successive species-forms, it does not seem to be possible to demonstrate, even after the fact, that the transition from one stage to another had to be made, or to predict before the fact what the direction of institutional adjustment will be.

These cautions apply to the prospectus confronting capitalism in our day. Its long post World War II boom seems to have been based on three attributes of the social structure of accumulation of that time. One of these was the increasing interconnection between the political and the economic realms, not merely to provide a public base for mass consumption but to utilize the state's power of finance and international leadership to promote foreign private trade and production. Japanese capitalism has been the much cited case in point for the latter development. A second characteristic of the boom was the extraordinary development of technology, based on the close integration of scientific research and technical application. A third was the pronounced bourgeoisification of working-class life, especially in Europe and Japan, greatly reducing the spectre of class conflict in capitalist politics.

On the basis of these developments capitalism enjoyed the longest uninterrupted period of accumulation in its history, from the early 1950s to the mid-1970s. Not only was the boom uninterrupted save for minor and shortlived recessions, but on the wings of its new technological breakthroughs, and under the auspices of its active state cooperation, capitalism made extraordinary advances in introducing its core institutions into many areas of the underdeveloped world.

This halcyon period came to a sharp end in 1980 when growth rates in the United States and Europe fell precipitously. Some, although not all of the causes of this depression can be ascribed to an exhaustion of the expansionary possibilities within the postwar social structure of accumulation. The effect of enlarged and sustained public expenditure gradually shifted from the encouragement of production to the inducement of inflation, thus setting the stage for the adoption of the tight money policies that finally broke the back of the boom. As markets became saturated, the advances in technology lost their capacity to stimulate capital expansion and attention was increasingly directed to their system-threatening aspects – ecologically dangerous products, employment-eroding processes and sovereignty-defying enhancements of the international mobility of money capital and commodities. The international character of capital acquired extraordinary importance, as multinational corporations transplanted fixed capital into underdeveloped regions, from which it launched artillery barrages of commodities back on its domestic territory. And not least, the bourgeoisification of labour may have removed a traditional source of adaptational pressure from capitalism.

It is not possible to foretell how these challenges will be met, or what institutional changes will be

forced upon the capitalist world as their consequence, or which capitalist nations will find the institutional and organizational means best suited to continue the accumulation process in this newly emerging milieu. Thus there is no basis for predicting the longevity of the social formation, either in its national instantiations or as a formational whole.

But while history forces on us a salutary agnosticism with regard to the long-term prospects for capitalism, it is interesting to note that all the great economists have envisaged an eventual end to the capitalist period of history. Smith describes the accumulation process as ultimately reaching a plateau when the attainment of riches will be ‘complete’, followed by a lengthy and deep decline. Ricardo and Mill anticipate the arrival of a ‘stationary state’, which Mill foresees as the staging ground for a kind of associationist socialism. Marx anticipates a series of worsening crisis, each crises serving a temporary rejuvenating function but bringing closer the day when the system will no longer be able to manage its internal contradictions. Keynes foresees ‘a somewhat comprehensive socialization of investment’; Schumpeter, an evolution into a kind of bureaucratic socialism. By way of contrast, contemporary mainstream economists are largely uninterested in questions of historic projection, regarding capitalism as a system whose formal properties can be modelled, whether along general equilibrium or more dynamic lines, without any need to attribute to these models the properties that would enable them to be perceived as historic regimes and without pronouncements as to the likely structural or political destinations towards which they incline. At a time when the need for institutional adaptation seems pressing, such an historical indifference to the fate of capitalism, on the part of those who are professionally charged with its self-clarification, does not augur well for the future.

See Also

- socialism

Bibliography

Baran, P. 1957. *The Political Economy of Growth*. New York: Monthly Review Press.

Faaland, J. 1982. *Population and the World Economy in the 21st Century*. Oxford: Blackwell.

Gordon, D. 1980. Stages of accumulation and long economic cycles. In *Processes of the World System*, ed. T. Hopkins and I. Wallerstein. Beverly Hills: Sage.

Heilbroner, R.L. 1985. *The Nature and Logic of Capitalism*. New York: W.W. Norton.

Hilton, R., ed. 1978. *The Transition from Feudalism to Capitalism*. London: Verso.

Hirschman, A. 1977. *The Passions and the Interests*. Princeton: Princeton University Press.

Myrdal, G. 1957. *Rich Lands and Poor*. New York: Harper & Bros.

Pirenne, H. 1936. *Economic and Social History of Medieval Europe*. London: K. Paul, Trench, Trubner & Co.

Polanyi, K. 1957. *The Great Transformation*. Boston: Beacon Press.

Sagan, E. 1985. *At the Dawn of Tyranny: The Origins of Individualism, Political Oppression, and the State*. New York: Knopf.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. New York: Harper & Bros.

Smith, A. 1776. *The Wealth of Nations*. Oxford: Clarendon Press, 1976.

Weber, M. 1930. *The Protestant Ethic and the Spirit of Capitalism*. London: Allen & Unwin.

How to cite this article

Heilbroner, Robert L. "capitalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000053> doi:10.1057/9780230226203.0198

The New Palgrave Dictionary of Economics Online

Carey, Henry Charles (1793 – 1879)

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Carey, Henry Charles; protectionism; tariffs; wage rates

Article

American social scientist. Born in Philadelphia, the son of Mathew Carey, he was a prolific author, and his influence, though short-lived, spread from Pennsylvania throughout the nation and to Europe. Carey's economic views were sharply at variance with those of Ricardo and Malthus, and reflect the optimism characteristic of American conditions favourable to economic expansion, conditions from which Carey himself benefited as a successful entrepreneur and promoter. The two leading themes of his writings were protectionism and harmony of interests. In his first book, *Essay on the Rate of Wages* (1835), he opposed trade restrictions as running counter to the providential order. But in *The Past, the Present, and the Future* (1848) and in later writings, he vigorously appealed for tariff protection as fulfilling his law of association, a law that called for diversified and balanced regional development. Narrow specialization and foreign trade would violate this law. In *The Slave Trade* (1853) Carey suggested protectionism for the South, where it would foster industrial development.

The scope of Carey's optimistic belief in a harmonious order gradually widened. In his first book he postulated harmony between capitalists and workers, the former benefiting from rising profits and the latter from wages that rose as a result of the accumulation of capital. In his *Principles of Political Economy* (1837–40) the landowner becomes part of the harmonious order, with his earnings depicted as a return on his capital rather than a gift of nature. Population growth does not disturb the harmony as it is restrained by social conditioning. There are further attacks against the Ricardian rent theory in *The Past, the Present, and the Future*, where cultivation is said to move from inferior to superior land, not vice versa as Ricardo had taught, and with returns increasing rather than decreasing. In the *Principles of Social Science* (1858–9) Carey expands his vision of a harmonious order to apply to the universe, and in *The Unity of Law* (1872) he maintains that cosmic and social laws are identical. Carey has been characterized as ‘easily the most perverse and the most original American political economist before Veblen’ (Conkin, 1980, p. 261).

Selected works

1835. *Essay on the Rate of Wages*. Philadelphia; Carey, Lea & Blanchard.

1837–40. *Principles of Political Economy*, 3 vols. Pennsylvania: Carey, Lea & Blanchard.

1838. *The Credit Systems in France, Great Britain and the United States*. Philadelphia: Carey, Lea & Blanchard.

1848. *The Past, the Present, and the Future*. Philadelphia: Carey & Hart; London: Longman, Brown, Green, and Longmans.

1851. *The Harmony of Interests, Agricultural, Manufacturing and Commercial*. Philadelphia: J.S. Skinner; 2nd edn, New York: M. Finch, 1852.

1853. *The Slave Trade, Domestic and Foreign: Why it Exists, and How it may be Extinguished*. Philadelphia: A. Hart.

1858–59. *Principles of Social Science*, 3 vols. Philadelphia: J.B. Lippincott & Co; London: Trübner & Co.

1863. *Financial Crises: Their Causes and Effects*. Philadelphia: Baird.

1867. *Reconstruction: Industrial, Financial & Political*. Philadelphia: Collins.

1872. *The Unity of Law; As Exhibited in the Relations of Physical, Social, Mental and Moral Science*. Philadelphia: H.C. Baird.

Bibliography

Conkin, P.K. 1980. *Prophets of Prosperity: America's First Political Economists*. Bloomington: Indiana University Press.

Dorfman, J. 1946. *The Economic Mind in American Civilization 1606–1865*, Vol. 2. New York: Viking.

Green, A.W. 1951. *Henry Charles Carey: Nineteenth-Century Sociologist*. Philadelphia: University of Philadelphia Press.

Kaplan, A.D.H. 1931. *Henry Charles Carey: A Study in American Economic Thought*. Baltimore: Johns Hopkins Press.

How to cite this article

Spiegel, Henry W. "Carey, Henry Charles (1793–1879)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000059> doi:10.1057/9780230226203.1855

The New Palgrave Dictionary of Economics Online

Carlyle, Thomas (1795 – 1881)

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Carlyle, T.; cash nexus; democracy; Engels, F.; McCullough, J. R.; Marx, K. H.; Ruskin, J.; utilitarianism

Article

The eldest of nine children of Margaret Aitkin and James Carlyle, Thomas Carlyle was born at Ecclefechan in Scotland on 4 December 1795. While Carlyle's contributions ranged over many fields (including history, literary and social criticism, biography, translation and political commentary), in economics he is remembered chiefly as the originator of the epithet ‘the dismal science’ (‘The Nigger Question’, 1849; in *Miscellaneous Essays*, vol. 7, p. 84). Among ‘the professors of the dismal science’, one M'Croudy (J.R. McCulloch) is a principal target of Carlyle's criticism. Yet Carlyle's writings on economics are more extensive than this small measure of recognition might suggest, and his key criticisms of the economic and political tendencies of the ‘present times’ (as he called them) are contained essentially in three works: *Chartism*, (1840), *Past and Present* (1843) and *Latter-Day Pamphlets* (1850). Almost inevitably, Carlyle's characteristically romantic reaction to the decline of authority and the rise of utilitarian individualism led him into head-on collision with the prevailing economic doctrines of the day. Since, for Carlyle, the challenge of democracy to the *ancien régime* had been carried forward under the mistaken banner ‘Abolish it, let there henceforth be no relation at all’ (1850, p. 21), it was natural for him to hold that laissez-faire, free competition, the law of supply and demand, and the ‘cash nexus’ were no more than ‘superficial speculations … to persuade ourselves … to dispense with governing’ (1850, p. 20). Although Carlyle's account of the ‘cash-nexus’ was adopted verbatim by Marx and Engels in the opening pages of *The Communist Manifesto*, in the latter sections of that document his overall position is roundly attacked (see there the reference to the ‘Young England’, of which Carlyle was a prominent member).
There is also a thinly veiled attack on Carlyle's ‘dissatisfaction with the Present … and affection and regret towards the Past’ in John Stuart Mill's *Political Economy* (1848, pp. 753–4). However, at Carlyle's hands the utilitarian calculus of pleasure and pain fared little better. It was charged with

ignoring all those sentiments, aspirations and interests which distinguished the human from other animals and was dubbed by Carlyle ‘the Pig Philosophy’ (1850, p. 268). Though Carlyle had few if any followers among economists, he exerted a profound impact upon the thinking of John Ruskin, and he may correctly be regarded as a principal exemplar in England of that reactionary or feudal brand of ‘socialism’ criticized by Marx and Engels in the *Communist Manifesto*. Carlyle died in Chelsea on 5 February 1881 and was buried in Ecclefechan.

Selected works

1888–9. *Works*. 37 vols. London: Chapman & Hall. (Page references above are from this edition.)

1896–9. *Works*. The centenary edition in 30 vols. London: Chapman & Hall.

Bibliography

Mill, J.S. 1848. *Principles of Political Economy*, ed. W.J. Ashley from the 7th edn (1871). London: Longmans, 1909.

How to cite this article

Milgate, Murray. "Carlyle, Thomas (1795–1881)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000061> doi:10.1057/9780230226203.0200

The New Palgrave Dictionary of Economics Online

cartels

Margaret C. Levenstein and Valerie Y. Suslow
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Cartels are associations of firms that restrict output or set prices. They may divide markets geographically, allocate customers, rig bids at auctions, or restrict non-price terms. They have often been formed with the participation or support of state actors. In contrast to the pre-Second World War period, today most cartels are illegal in most jurisdictions. The average duration of cartels is between five and seven years, but the distribution of duration is skewed: a large number of cartels break down within a year but a sizable proportion last for over a decade.

Keywords

antitrust enforcement; barriers to entry; cartels; cheating; collusion; communication; concentration; coordination; entry; innovation; price fixing; price wars; productivity growth; trust

Article

Producers form cartels with the goal of limiting competition to increase profits.
Cartels are associations of independent firms that restrict output or set prices. They may divide markets geographically, allocate customers to specific producers, rig bids at auctions, or restrict non-price terms offered to customers. They have often been formed with the active participation or support of state actors. In contrast to the pre-Second World War period, today most cartels are illegal in most jurisdictions.
Upon its creation a cartel immediately faces three key problems: coordination, cheating and entry. In a dynamic economy, the solution to these problems will change over time, so successful cartels must develop an organizational structure that allows them to re-solve these problems continuously.
Stigler's (1964) classic article highlights the incentive to cheat as the most important source of instability undermining cartels. In a repeated setting, a firm weighs the expected gain from cheating today (the benefit from cheating) with the expected reduction in future discounted profits that follows cheating (the cost of cheating). In order for firms to be willing to refrain from cheating, the following must hold:

$$\sum_{t=0}^{T} \delta^{t} \frac{\Pi^{m}}{n} \succ \Pi^{m}$$

where Π^{m} is the one-period cartel profit, *n* is the number of firms in the industry, and δ^{t} is the discount rate. Thus, collusion is easier to achieve the larger the difference between cartel and non-cartel profits, the smaller the number of firms, and the more patient these firms are (Tirole, 1988).

Friedman (1971) demonstrates that firms may use ‘off the equilibrium path’ threats of price wars in retaliation for cheating to provide firms with the incentive not to cheat. However, because in his model any cheating would be observed immediately and therefore subject to swift retaliation, firms do not cheat and price wars are not observed. In the Green and Porter class of models (Green and Porter, 1984; Abreu, Pearce and Stacchetti, 1986), firms cannot observe one another's output (or pricing) actions nor infer them with certainty from public information. Economic fluctuations require that firms revert to equilibrium ‘punishment’ or ‘price war’ behaviour at times in order to maintain the incentives necessary to achieve collusion. Thus, the appearance of on-and-off collusion does not represent inherent cartel instability, but rather a mechanism that cartels use to stabilize themselves.

This theoretical perspective also implies a second mechanism for increasing cartel stability: a cartel may invest in information collection in order to better monitor individual firm activities. Improved monitoring both deters cheating and allows cartels to avoid costly price wars that arise from the inability to distinguish cheating from external shocks.

The most successful cartels actively work to create barriers to entry. Sometimes this is done through collective predation, as in Scott Morton (1997) in which incumbent cartel members successfully deterred entry by financially weaker and smaller firms. In other cases, cartels have turned to the state to create regulations, tariffs, or provide anti-dumping protection with the goal of excluding outsiders. Cartels sometimes use vertical exclusion (for example, a joint sales agency) or restrict access to technology (for example, via a patent pool) to limit entry.

Cartels use direct and repeated communication to overcome obstacles to coordination. Cartel negotiations often begin with discussions of prices and market shares, but expand over time to restrict cheating in non-price dimensions, such as terms of sale, advertising, transport costs, and production capacities. Firm asymmetries and changes in firms' costs can make these negotiations challenging. Slade (1989) suggests that price wars arise from changes in firm or industry characteristics. These price wars then facilitate the learning necessary for firms to re-establish collusion. Cartels also learn how to structure incentives so that collusion is more profitable in the long run than cheating. For example, successful cartels often fashion self-imposed penalties or other compensation schemes for firms that exceed cartel quotas. Cartels sometimes develop elaborate internal hierarchies allowing for communication at various levels of management. A hierarchical cartel structure allows for high-level information exchange and bargaining activities to be separated from regional or local information exchange and monitoring efforts. When trust is particularly difficult to establish and firms doubt the accuracy of communication or data exchanges, cartels often turn to a third party – such as a trade

association – to facilitate information sharing.

The average duration of cartels measured over a range of countries and time periods is between five and seven years (Levenstein and Suslow, 2006). There is considerable dispersion in cartel duration: the standard deviation of duration is almost as high as the average. Observed cartel duration is very skewed, with a large number of cartels lasting less than a year or two and a long tail of cartels that endure for a decade or more.

Predictable fluctuations in product or industry demand do not generally undermine effective cartels, but rapid industry growth and unexpected shocks do. Macroeconomic fluctuations, which are close to common knowledge, have little impact on cartel stability. Many successful cartels develop an organizational structure that allows them to weather cyclical fluctuations. Cartels that are disrupted by observable cyclical fluctuations may be inherently fragile.

Large customers can undermine cartel stability by increasing the incentive to cheat, as posited by Stigler (1964) and tested by Dick (1996). On the other hand, large customers sometimes benefit from the existence of a cartel if they receive preferential pricing compared with that received by their smaller competitors, and can even contribute to its stability.

Although posited by theory, there is no simple empirical relationship between industry concentration and the likelihood of collusion. This may reflect sampling bias in studies that focus on prosecuted cartels, since cartels with many firms or with the involvement of an industry association may be easier to detect. Or it may be that industries with a small number of firms are able to collude tacitly without resorting to explicit cartels. Finally, it may reflect the endogeneity of concentration: collusion may allow more firms to survive and remain in the market (Sutton, 1991; Symeonidis, 2002).

Analyses of the impact of cartels on prices and profits generally use one of three approaches: changes in price following cartel formation, comparison between ‘good times’ and ‘price war’ periods, and, comparison between the cartel price and a counterfactual or ‘but-for’ price that would have prevailed in the absence of collusion. Connor and Lande (2005) provide an exhaustive survey of studies of cartel price effects. They conclude that the median overcharge resulting from cartels is approximately 25 per cent.

Cartels can also affect investment and productivity. Cartel participants have often argued that cartels increase investment and productivity growth by allowing firms to smooth production over time. Others have argued that, by removing the pressure of competition, cartels reduce innovation and productivity growth. Theoretical models have suggested that cartels lead to increased investment in capacity either because excess capacity can deter entry and provide enforcement (Dixit, 1980) or because, when price competition is suppressed, firms compete in other dimensions (Feuerstein and Gersbach, 2003). In some cases, cartels explicitly restrict investment in new capacity. Where there are not such explicit restrictions, empirical studies have found cartels are associated with increases in investment. On the other hand, no consistent relationship between cartels and productivity growth or innovation has been established empirically (Symeonidis, 2002).

As firms have become increasingly global, international antitrust law and policy has faced new challenges. Competition authorities have increased enforcement, attempted to harmonize practices and procedures, and increased cooperation across jurisdictions. The United States is the country with the longest history of prosecuting explicit collusion, with state laws antedating the national ban on price fixing enacted with the passage of the Sherman Act of 1890. Many Western European countries adopted

laws against price fixing following the Second World War, but also allowed a large number of exemptions. Since the mid-1990s these exemptions have been sharply reduced, and dozens of other countries have banned price fixing for the first time. Enforcement activities against cartels, and international cartels in particular, rose sharply in the United States in the late 1990s. European countries, including the newest members of the European Union, have also increased their enforcement activities against cartels, as have countries in Asia, Africa and Latin America. Price fixing – long a criminal offence in the United States – has now been criminalized in several other countries, including the United Kingdom and Ireland. This increased enforcement has demonstrated that cartels continue to be active in a wide range of industries in the 21st century.

See Also

- antitrust enforcement
- cooperation
- market structure
- Organization of the Petroleum Exporting Countries (OPEC)
- Stigler, George Joseph

Bibliography

Abreu, D., Pearce, D.G. and Stacchetti, E. 1986. Optimal cartel equilibria with imperfect monitoring. *Journal of Economic Theory* 39, 251–69.

Connor, J.M. and Lande, R.H. 2005. How high do cartels raise prices? Implications for optimal cartel fines. *Tulane Law Review* 80, 513–70.

Dick, A.R. 1996. When are cartels stable contracts? *Journal of Law and Economics* 39, 241–83.

Dixit, A. 1980. The role of investment in entry-deterrence. *Economic Journal* 90, 95–106.

Feuerstein, S. and Gersbach, H. 2003. Is capital a collusion device? *Economic Theory* 21, 133–54.

Friedman, J.W. 1971. A non-cooperative equilibrium for supergames. *Review of Economic Studies* 38, 1–12.

Green, E.J. and Porter, R.H. 1984. Noncooperative collusion under imperfect price information. *Econometrica* 52, 87–100.

Levenstein, M.C. and Suslow, V.Y. 2006. What determines cartel success? *Journal of Economic Literature* 64, 43–95.

Martin, S. 2002. *Advanced Industrial Economics*, 2nd edn. Oxford: Blackwell.

Posner, R.A. 1970. A statistical study of antitrust enforcement. *Journal of Law and Economics* 13, 365–419.

Scott Morton, F. 1997. Entry and predation: British shipping cartels 1879–1929. *Journal of Economics and Management Strategy* 6, 679–724.

Slade, M.E. 1989. Price wars in price-setting supergames. *Economica* 56, 295–310.

Stigler, G. 1964. A theory of oligopoly. *Journal of Political Economy* 72, 44–61.

Stocking, G.W. and Watkins, M.W. 1949. *Cartels in Action*. New York: The Twentieth Century Fund.

Sutton, J. 1991. *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration*. Cambridge, MA: MIT Press.

Symeonidis, G. 2002. *The Effects of Competition: Cartel Policy and the Evolution of Strategy and Structure in British Industry*. Cambridge, MA: MIT Press.

Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

How to cite this article

Levenstein, Margaret C. and Valerie Y. Suslow. "cartels." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000552> doi:10.1057/9780230226203.0201

The New Palgrave Dictionary of Economics Online

Cassel, Gustav (1866 – 1944)

Bo Gustafsson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

acceleration principle; business cycle theory; Cassel, G.; Davidson, D.; deflation; Hadley, A. T.; Harrod growth formula; inflation; Kitchin, J.; labour theory of value; labour–capital ratio; marginal cost pricing; marginal utility of money; marginal utility theory of value; Myrdal, G.; Ohlin, B. G.; purchasing power parity; quantity theory of money; Say's Law; Spiethoff, A. A. K.; Stockholm School; trade unions; Tugan-Baranowsky, M. I.; value; Walras, L.; Wicksell, J. G. K.; Woytinski, W. S.

Article

Along with Knut Wicksell and David Davidson, Gustaf Cassel was the founder of modern economics in Sweden. He started as a mathematician and began his career as an economist by treating problems of railway rates and progressive taxation from a mathematical point of view. In order to deepen his understanding of economics he went to Germany, where he attended the seminars of Schönberg, Cohn and other traditional representatives of the economic profession. After visits to England, where he made the acquaintance of Marshall and of Sidney and Beatrice Webb, and a short period of lecturing at the university of Copenhagen, in 1902 Cassel took up a position as associate professor in economics at the university of Stockholm. In 1904 he was appointed a professor in economics and public finance. As holder of the chair he acquired a series of gifted pupils, Gunnar Myrdal and Bertil Ohlin among others, who, although they developed the theoretical heritage of Wicksell rather than that of Cassel, became the founders of the Stockholm School of economics. Before the First World War Cassel frequently served as a government expert on problems of railway rates, taxation, state budgets and banking and his involvement in problems of economic policy increased with the post-war economic problems. During the 1920s he became an adviser to the League of Nations on monetary problems and was commonly regarded as a leading international authority in this field, lecturing and publishing widely. All his life he worked also as a columnist for the Swedish daily paper *Svenska Dagbladet*. Although Cassel was originally liberal, he progressively turned more and more conservative denouncing the labour movement, the welfare state and Keynesianism in the name of ‘Modern Scientific Principles’.
It is no easy task to evaluate the contributions of Gustav Cassel to economics. He never cared much

about paying homage to his predecessors, from whom he sometimes took over fruitful ideas, while at the same time being unjustifiably critical towards other theorists. His expositions are not seldom marred by contradictions and a vagueness in expression, only scantily veiled by his mastery in round and polished sentences. At the same time Cassel took a keen interest in very many fields of economic theory and practice, he had a firm grip on empirical economics and his gifts in tracking down the relevant and essential aspects of economic problems were unusual. These qualities, in combination with a forceful and pedagogical exposition and, on the top of this, an imperturbable conviction of being the chosen spokesman for progress and the principles of science, made him influential not only among men of practical matters but also among fellow economists.

Cassel's main work is his *Theoretische Sozialökonomie* (1918) but his most important theoretical ideas were in fact conceived already around the turn of the century. In his essay ‘Grundsätze für die Bildung der Personentarife auf den Eisenbahnen’ (1900b), he criticized the idea of calculating railway rates on the basis of average costs and instead advocated marginal cost pricing. For a railway enterprise as a monopolistic business unit, rates which equalized marginal costs and marginal revenues were the optimal ones, though this might imply that some rates were lower than average costs. Even if the principle had been advocated already in 1885 by the American railway economist A.T. Hadley, it was succinctly formulated by Cassel.

Venturing into general economic theory, Cassel in these years also criticized Ricardo's labour theory of value in the essay ‘Die Produktionskostentheorie Ricardos und die ersten Aufgaben der theoretischen Volkswirtschaftslehre’ (1901), presented an outline of his own theory of price, ‘Grundriss einer elementaren Preislehre’ (1899) and developed a theory of interest in *The Nature and Necessity of Interest* (1903). The Ricardian labour theory of value was, according to Cassel, untenable because it assumed that the labour–capital ratio was equal in different enterprises and industries, that labour was homogeneous and that the marginal land did not pay any rent. He did not care to take issue with the Marxian development of the labour theory of value. The labour theory of value belonged to the so-called one-sided value theories. But so did the marginal utility theory of value, which was deficient primarily because it lacked a clearly conceptualized unit of measurement for utility but also because goods, according to Cassel, are not generally divisible and the valuations of goods are not continuous functions of the supply. Therefore, Cassel suggested that one should do away with all conceptions of value and rest content with money prices and not bother with what might lie behind money prices. Thus Cassel did not consider the fact that money itself may vary in value, nor that the marginal utility of money certainly varies between individuals. Following Marshall, Cassel explained prices by reference to supply and demand and, following Walras, he devised a general equilibrium model for market prices in the form of a system of simultaneous equations. In fact, Cassel's price theory is a simplified version of the theory of Walras, who was characterized as ‘in a sense one of my precursors’. However, by popularizing Walras, Cassel contributed much towards the understanding of the mutual interdependencies in a market economy. It was quite logical that the theory of interest that Cassel devised also should be based upon supply and demand, viz. supply of waiting and demand for the use of capital, as a special case of the general theory of price, and he boldly asserted that waiting and use denoted the same thing. Although his theory of interest, showing a close resemblance to that of Senior, was not original, it still merits our attention because of its vivid illustrations and some striking applications. This is particularly the case for Cassel's argument against the idea of a continually falling rate of interest. Given that most saving is made in order to safeguard a permanent future level of income, the shortness of life puts a ceiling under

the rate of interest. This was the necessary and sufficient condition for the necessity of interest.
The year after the publication of *The Nature and Necessity of Interest*, Cassel also published his theory of the business cycle and his theory of the secular development of the general level of prices in two articles in the Swedish journal *Ekonomisk tidskrift*, ‘Om kriser och däliga tider’ (1904a) and ‘Om förändringar i den allmänna prisnivän’ (1904b). Both these theories were later incorporated and somewhat elaborated in his *Theoretische Sozialökonomie* (1918). In his theory of the business cycle Cassel was evidently influenced by Spiethoff and Tugan-Baranowsky, who recently had made public their theories explaining the business cycle with reference to the variations in investment of fixed capital and of loanable funds. What is really new in Cassel's treatment is his precise formulation of the accelerator principle, which he expounds with reference to the relationship between the demand for freights and the output of ships. The treatment of growth theory had to await the publication of his *Theoretische Sozialökonomie* and also on this point Cassel was wholly original, in fact foreshadowing the Harrod growth formula by his own formula for ‘the uniformly progressing economy’, the only difference being that Cassel worked with an average instead of a marginal capital coefficient.
Cassel's theory of the secular development of the general level of prices also demands our attention as a piece of brilliant imagination and was as late as 1930, after Kitchin's refinements, accepted as the theoretical basis for the first interim report of the gold delegation of the League of Nations. Cassel's theory was a straightforward quantity theory of money. By calculating the relative variations of gold output in relationship to a calculated normal need of gold for preserving a constant general level of prices, Cassel showed that there was a very good correlation between the relative variations of gold output and the corresponding variations in the general level of prices. Cassel's theory met with all the objections the quantity theory of money usually meets and in addition a series of more specific criticism: that it presupposes a constant ratio between velocity (V) and transactions (T), which is difficult to believe; that it overlooks the important role of silver in the 19th century as well as the varying proportions of the more relevant variable monetary gold; and that a case as good as Cassel's could be made, and in fact was made by Warren and Pearson, by making the gold price rather than gold output the effective cause of price changes. But since Kitchin's (and Woytinski's) calculations, taking only monetary gold in regard, showed a still better fit between the variations of gold output and prices, Cassel's theory is still a serious candidate.
After this first period of theoretical activity around the turn of the century, Cassel mainly devoted his energy to synthesizing and propagating his ideas on the national and the international scene. The only really new element in his theoretical set-up was the famous purchasing power parity theory of the exchange rates, according to which the international rates of exchanges are determined by the purchasing power of the national currencies. It is easy to show that this is a rather poor general theory for the explanation of the exchange rates. But it contained a pragmatic truth during and after the First World War, when trade balances and, hence, the supply and demand of currencies, to a great extent, were determined by the course of rapid inflation in different countries. It is precisely this instinct for pragmatic truths that explains Cassel's success and influence in the international community of bankers and politicians during the 1920s. In his memoranda to the international conferences of the League of Nations Cassel first and foremost advocated stability of monetary affairs by means of control of the quantity of money, increased interest rates and cut-downs of state expenditures. But he was also critical towards the subsequent ruthless policy of deflation creating widespread unemployment and new disequilibria in world trade as well as intolerable debt burdens. Together with Keynes he criticized the

unwillingness of the claimants to the German war debt to receive German goods as payment. When confronted by the permanent unemployment of the 1920s, Cassel concentrated his attacks on trade unions and the level of wages and untiringly explained the gospel contained in Say's Law. During the course of the 1930s it became all too clear that Gustav Cassel had been left behind by the march of events and of economic theory. It was his tragedy that he himself, who once waved his magic wand over international economic affairs, could not bear the truth. After some years of protracted rearguard skirmishes he devoted himself to more philosophical problems and wrote up a voluminous autobiography characteristically entitled ‘In the Service of Reason’ (*I förnuftets tjänst*, 1940–41). His last words on his death-bed were ‘A world currency!’

See Also

- purchasing power parity
- Wicksell, Johan Gustav Knut

Selected works

1899. Grundriss einer elementaren Preislehre. *Zeitschrift für die gesamte Staatwissenschaften* 55, 395–458.

1900a. *Das Recht auf den vollen Arbeitsertrag*. Göttingen: Vandenhoeck & Ruprecht.

1900b. Grundsätze für die Bildung der Personentarife auf den Eisenbahnen. *Archiv für Eisenbahnwesen*. Trans. as ‘The principles of railway rates for passengers’, *International Economic Papers* 6 (1956), 126–47.

1901. Die Produktionskostentheorie Ricardos und die ersten Aufgaben der theoretischen Volkswirtschaftslehre. *Zeitschrift für die gesamte Staatswissenschaft* 57, 68–100.

1902a. Der Ausgangspunkt der theoretischen Ökonomie. *Zeitschrift für die gesamte Staatswissenschaft* 58, 668–98.

1902b. *Sozialpolitik*. Stockholm: H. Geber.

1903. *The Nature and Necessity of Interest*. London, New York: Macmillan.

1904a. Om kriser och dåliga tider. *Ekonomisk Tidskrift* 6, 21–35 and 51–81.

1904b. Om förändringar i den allmänna prisnivån. *Ekonomisk Tidskrift* 6, 311–31.

1917. *Dyrtid och sedelöverflöd*. Stockholm: P.A. Norstedt & Söner.

1918. *Theoretische Sozialökonomie*. Leipzig: C.F. Winter. (In manuscript in 1914. There are five German editions.) Trans. into English as *Theory of Social Economy*, London: T.F. Unwin, 1923; new revised edn., London: E. Benn, 1932. French, Japanese and Swedish editions also available.

1921. *The World's Monetary Problems. Two Memoranda Presented to the International Financial Conference of the League of Nations in Brussels in 1920 and to the Financial Committee of the League of Nations in September 1921*. London: Constable & Co.

1922. *Penningväsendet efter 1914*. Stockholm: P.A. Norstedt. Trans. into English as *Money and Foreign Exchange after 1914*, London: Constable & Co., 1922. German, French, Spanish and American editions also available.

1925. *Fundamental Thoughts in Economics*. London: T.F. Unwin.

1927. *Memorandum till den Internationella ekonomiska Konferensen i Genève 1927*. Stockholm: P.A. Norstedt.

1928. *Socialism eller framätskridande*. Stockholm: P.A. Norstedt & Söners.

1935. *On Quantitative Thinking in Economics*. Oxford: Clarendon Press.

1936. *The Downfall of the Gold Standard*. Oxford: Clarendon Press.

1940–41. *I förnuftets tjänst, en ekonomisk självbiograft*. 2 vols, Stockholm: Bokfölaget Natur och Kultur.

1942. *Vår bildnings fåfänglighet*. Stockholm: Bonniers.

1944. *Den odelbara människan*. Stockholm: Natur och Kultur.

Bibliography

A good biography of Cassel has yet to appear. The biography by his secretary, Ingrid Giobel-Lilja (1948), may be consulted for biographical details, especially as regards his family life. Böhm-Bawerk (1914) made a critical examination of Cassel's theory of interest, and Wicksell (1919) published a very critical review of *Theoretische Sozialökonomie*. A more descriptive account of Cassel's main work is to be found in Mitchell (1969); Gunnar Myrdal (1945) wrote an obituary; and Gustafsson (1964) offers another overall evaluation with emphasis on Cassel's theory of long-run prices.

Böhm-Bawerk, E. von. 1914. Exkurs XIII, Kritische Glossen zur Zinstheorie Cassels, Exkurse zur ‘Positiven Theorie des Kapitals’.

Giobel-Lilja, I. 1948. *Gustav Cassel, En livsskildring*. Stockholm: Natur och Kultur.

Gustafsson, B. 1964. Gustav Cassel (1866–1944). *Vär Tid* No. 3.

Mitchell, W.C. 1969. Chapter 16 in *Types of Economic Theory*, vol. 2. New York: A.M. Kelley.

Myrdal, G. 1945. Gustav Cassel in memoriam. *Ekonomisk Revy*. Republished (in Swedish) in J. Schumpeter, *Stora nationalekonomer*. Stockholm: Natur och Kultur, 1953.

Wicksell, K. 1919. Professor Cassels nationalekonomiska system. *Ekonomisk Tidskrift* 21, 195–226. Republished in *Schmollers Jahrbuch* 52(5), 1928, and in K. Wicksell, Appendix I in *Lectures on Political Economy*, vol. I. London: G. Routledge & Sons, 1934.

How to cite this article

Gustafsson, Bo. "Cassel, Gustav (1866–1944)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000071> doi:10.1057/9780230226203.0202

The New Palgrave Dictionary of Economics Online

caste system

Susan Wolcott

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

India's caste system performed two fundamental functions: insurance through transfers between caste members and, in villages, insurance through protected job assignments across castes. In most of India the landlord had a social responsibility to maintain his lower caste workers in lean periods. This division of labour has been viewed as coercive and exploitative. Yet many groups changed their caste occupation, both upward and downward in ritual ranking. During industrialization, traditional occupational categories did not restrict occupational choices in new industries, but caste continued to play a role in recruitment and support during work stoppages.

Keywords

Akerlof, G.; caste system; division of labour; India; industrialization; insurance

Article

The caste system in India is a division of society into ranked, hereditary, endogamous occupational groups. It is loosely based on the four *varnas* of Brahmanas (priests), Kshatriyas (warriors and aristocracy), Vaishyas (merchants) and Shudras (the servants of the others). Castes either belonged to one of these four, or were below them in the hierarchy; these latter are the so-called untouchables. In practice, the *varnas* are less important than were the relationships among and between the numerous sub-castes, or *jatis*. The sub-castes were specific to each region and were the true functional unit of the caste system. They were, for example, the endogamous unit. And obligations of *jati* members to each other were much stronger than were obligations of caste members more generally. Below the terms '*jati*' and 'caste' are used interchangeably.

Caste was not a monolithic institution. Reviewing the historical literature on caste, Rudner (1994, p. 25) notes that it is impossible for any one description to capture the 'on-the-ground diversity of India's caste systems'. He suggests as a definition: 'complex, multilayered, multifunctional corporate kin groups with enduring identities, a variety of rights over property, and crucial economic roles, often within large

regions’.

Because of this diversity, caste's role in the Indian economy varied across regions and across groups. But two functions were fundamental: insurance through transfers between caste members and, in village India, insurance through protected job assignments across castes. On the first of these, Srinivas (1962, p. 70) writes, ‘joint family and caste provide for an individual in our society some of the benefits which a welfare state provides for him in the industrially advanced countries of the West’. Economists have completely ignored this aspect of caste. But in the modern period it seems to be economically significant: financial transfers among rural villagers are common in developing countries. However, this practice is much more common in India than in any other country yet studied (Cox and Jimenez, 1990, Table 1). As caste ties are weakening over time and as income rises, it is likely that such transfers were even more prevalent historically.

And across castes, because each *jati* was, at least in theory, occupationally segregated in the villages of colonial India, it played a protected role in the economic order and had a claim on the wealth produced by the village. This relationship is called the *jajmani* system in much of India, and the *baluta* system in Maharashtra (Kolenda, 1978).

A particular division of responsibilities is that between landlords and agricultural labourers. Especially in north, south and east India, the landlord had a social responsibility to maintain his workers in lean periods. Platteau (1995) reviews the literature on this topic and presents a mathematical formalization of this relationship. Greenough (1982) gives an account of the strains on this system and its ultimate collapse in an extreme crisis.

This division of labour has also been viewed as coercive and exploitative. Akerlof (1976) models a situation in which groups can be confined to inferior occupations by social opprobrium. Maddison (1971, p. 28) argues that these occupational divisions were not only coercive but also foolish: ‘One might think that some of the lowest productivity occupations were invented simply to provide everyone with a job in a surplus labor situation, but there was no shortage of land and the productivity of the economy would have been higher if there had been greater job mobility.’

But these authors exaggerate the rigidity of the caste system in regard to occupational segregation. Mukerjee (1937) provides a long list of groups which had changed their caste occupation, both upward and downward in ritual ranking, as well as lists of splitting and merging sub-castes. He argues that, although there was rigid social control within the caste, the system revealed ‘plasticity’ in regard to economic incentives. As an example of this, Commander (1983) notes that historical sources imply that the Chamars of the United Provinces – hereditarily leather workers – were for much of the 19th century largely agricultural labourers. He argues cogently that, although ritual and custom were important in determining economic rewards and relative position in the *jajmani* system, so were land availability and labour scarcity.

Did caste have a role in modern industrialization? The best survey on this subject remains that of Morris (1960). One point is obvious. Traditional occupational categories did not restrict occupational choices in new industries. Whether or not caste affected the economic lives of the workforce in other ways is less clear. Morris (1960, p. 128) writes that he ‘is inclined to the view that *jati* relationships ultimately are irrelevant in the factory’. Most analysts argue, however, that, because of the economically supportive links between *jati* members, caste did have a role in recruitment and support during work stoppages

(Chandavarker, 1994; Klass, 1978). The differentiated and fluid nature of caste makes a general statement impossible.

See Also

- Akerlof, George Arthur
- labour market institutions
- peasant economy
- risk-coping strategies

Bibliography

Akerlof, G. 1976. The economics of caste and of the rat race and other woeful tales. *Quarterly Journal of Economics* 90, 599–617.

Chandavarker, R. 1994. *The Origins of Industrial Capitalism in India: Business Strategies and the Working Classes in Bombay, 1900–1940*. Cambridge: Cambridge University Press.

Commander, S. 1983. The jajmani system in north India: an examination of its logic and status across two centuries. *Modern Asian Studies* 17, 283–311.

Cox, D. and Jimenez, E. 1990. Achieving social objectives through private transfers: a review. *World Bank Research Observer* 5, 205–18.

Greenough, P. 1982. *Prosperity and Misery in Modern Bengal: The Famine of 1943–1944*. Oxford: Oxford University Press.

Klass, M. 1978. *From Field to Factory: Community Structure and Industrialization in West Bengal*. Philadelphia, PA: Institute for the Study of Human Issues.

Kolenda, P. 1978. *Caste in Contemporary India*. Menlo Park, CA: The Benjamin/Cummings Publishing Co.

Maddison, A. 1971. *Class Structure and Economic Growth. India and Pakistan since the Moghuls*. New York: W.W. Norton and Co.

Morris, M. 1960. Caste and the evolution of the industrial workforce. *Proceedings of the American Philosophical Society* 104, 124–33.

Mukerjee, R. 1937. Caste and social change in India. *American Journal of Sociology* 43, 377–90.

Platteau, J. 1995. An Indian model of aristocratic patronage. *Oxford Economic Papers* 47, 636–62.

Rudner, D. 1994. *Caste and Capitalism in Colonial India: The Nattukottai Chettiars*. Berkeley: University of California Press.

Srinivas, M. 1962. *Caste in Modern India and Other Essays*. London: Asia Publishing House.

How to cite this article

Wolcott, Susan. "caste system." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000198> doi:10.1057/9780230226203.0203

The New Palgrave Dictionary of Economics Online

catallactics

Murray N. Rothbard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

catallactics; Condillac, E. B. de; Crusoe economics; exchange; Macleod, H. D.; Mises, L. E. von; Perry, A. L.; political economy; praxeology; property; Schumpeter, J. A.; Smith, A.; subjective theory of value; wealth

Article

The term, meaning ‘the science of exchanges’, was proposed as a replacement for the name ‘political economy’ by the Rev. Richard Whately in his 1831 Drummond Lectures at Oxford on political economy (Whately, 1831). As the leader of the group of embattled religious and economic liberals at Oriel College, Oxford, during the 1820s, Whately, a distinguished logician, had become tutor and lifelong friend of the economist Nassau W. Senior. In his Drummond Lectures, Whately was concerned to refute the dominant Oxford view that political economy, being concerned with wealth, was materialistic and opposed to Christianity. In focusing on exchanges, Whately denounced Adam Smith's definition of the scope of political economy as the science of wealth.

Whately defined man as ‘an animal that makes exchanges’, pointing out that even the animals nearest to rationality have not ‘to all appearance, the least notion of bartering, or in any way exchanging one thing for another’ (Whately, 1831, p. 7). Focusing on human acts of exchange rather than on the *things* being exchanged, Whately was led almost immediately to a subjective theory of value, since he saw that ‘the same thing is different to different persons’ (p. 8) and that differences in subjective value are the foundation of all exchanges.

In 1831 Whately was named Archbishop of Dublin, where he promptly used his influence to create and financially support a permanent five-year Whately Chair of Political Economy at Trinity College. For the rest of his life Whately personally selected the holders of the chair; as a result, the Whately professors carried on their mentor's tradition of catallactics and subjective utility theory. In contrast to John Stuart Mill's development of economics as a science of the abstraction ‘economic man’, man engaged only in avaricious pursuit of wealth, the third holder of the Whately Chair, James Anthony Lawson (1817–87), developed the idea of economics as catallactics, as studying exchanging man.

Lawson, holder of the chair in his twenties (1841–6), and later to become an MP and Attorney-General for Ireland, stated in his first lecture that economics views man ‘in connection with his fellow-man, having reference solely to those relations which are the consequences of a particular act, to which his nature leads him, namely, the act of making exchange’ (Lawson, 1844, pp. 12–13). Yet, Lawson himself fell back on discussions of wealth in his second lecture, demonstrating that, in their specific exposition, the catallacticians had not yet fully emancipated themselves from the older definitions of the scope and nature of political economy (Kirzner, 1960).

One pseudonymous English writer who adopted catallactics in this period was Patrick Plough, who included and explained the term in the title of his tract, *Letters on the Rudiments of a Science, called, formerly, improperly, Political Economy, recently more pertinently, Catallactics* (London, 1842).

Catallactics reached the status of a self-conscious school of thought in the writings of the zealous and indefatigable Scottish lawyer and economist Henry Dunning Macleod. Stressing value as the result of a subjective desire of the mind, Macleod furthered the emancipation of economics from material wealth by showing that immaterial goods or services are also subjects of exchange. Macleod insisted that catallactics was the only correct school of economic thought and traced back the origins of the school beyond Whately to the late 18th-century French philosopher Etienne Bonnot de Condillac. While Condillac, in his *Le commerce et le gouvernement* (1776), did not actually use the term catallactics, he defined economics as the philosophy of commerce, or the science of exchanges. Condillac also noted that value stems only from mental desires, and hence demand, for exchangeable goods, and proclaimed that men engage in exchange precisely because each man values what he gains in exchange more than what he gives up. Hence both parties to an exchange gain in value (Macleod, 1863, pp. 530–5).

The catallactic school found its culmination in the United States, in Arthur Latham Perry (1830–1905), for half a century a highly influential professor of political economy at Williams College. Perry endorsed the Macleod view of the history of economic thought, the sound catallactic school descending from Condillac through Whately and Macleod. He went beyond the inconsistencies of his forerunners, however, by purging the word ‘wealth’ from economics altogether, and proposing the ‘property’ – that which can be bought and sold – be used as a term denoting valuable things not yet sold and therefore in need of an estimate of their value (Perry, 1865).

While interest in the catallactic approach faded after the work of Perry, a variant appeared in the early work of Schumpeter (1908). In this manifesto for the reconstruction of economic theory, Schumpeter wished to purge economics of all concern about purposeful human motives or actions and replace it with exclusive concentration on mechanistic alterations of economic quantities. Exchanges then become ‘purely formal’ variations in economic quantities of goods (Schumpeter, 1908, pp. 49–55, 86, 582; Machlup, 1951; Kirzner, 1960).

Schumpeter did, however, manage to contribute positively to the catallactic approach. Whately and his followers had strongly rejected any element of Crusoe economics, since for them economic analysis had to be confined to interpersonal exchange. In Schumpeter's formalistic approach, actions of Crusoe could alter the placement of quantities of economic goods and therefore could be considered ‘exchanges’.

It remained for Ludwig von Mises (1949) to bring back the term catallactics in his treatise on economics, and to broaden it by embedding its analysis of the market, or the science of exchanges, in the wider discipline of ‘praxeology’, the science of human action. Crusoe economics then becomes

vindicated in the broader sense of analysing Crusoe's actions and his use of resources to achieve his values and goals, as well as in the sense of exchanging his present state for a more satisfying one.

See Also

- Macleod, Henry Dunning

Bibliography

Condillac, E.B. de. 1776. Le commerce et le gouvernement considérés relativement l'un à l'autre. In *Oeuvres philosophiques de Condillac*, vol. 2, ed. George LeRoy, Paris: Presses Universitaires de France, 1947–51.

Kirzner, I.M. 1960. *The Economic Point of View: An Essay in the History of Economic Thought*. Princeton: Van Nostrand.

Lawson, J.A. 1844. *Five Lectures on Political Economy*. Delivered before the University of Dublin, 1843. London and Dublin.

Machlup, F. 1951. Schumpeter's economic methodology. *Review of Economics and Statistics* 33(May), 145–51.

Macleod, H.D. 1863. A Dictionary of Political Economy, vol. 1. London.

Mises, L. von. 1949. *Human Action: A Treatise on Economics*. New Haven: Yale University Press.

Perry, A.L. 1865. *Political Economy*. 21st edn, New York: Scribners, 1892.

Plough, P. 1842. *Letters on the Rudiments of a Science, called, formerly, improperly, Political Economy, recently more pertinently, Catallactics*. London.

Schumpeter, J.A. 1908. *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie*. Leipzig: Duncker & Humblot.

Whately, R. 1831. *Introductory Lectures on Political Economy*. 2nd edn, London, 1832.

How to cite this article

Rothbard, Murray N. "catallactics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/

article?id=pde2008_C000072> doi:10.1057/9780230226203.0204

The New Palgrave Dictionary of Economics Online

catastrophic risk

Richard A. Posner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Catastrophic risks are defined here as events of low or unknown probability that if they occur inflict enormous losses often having a large non-monetary component. The Indian Ocean tsunami of 2004 is at the lower level of the catastrophic-risk scale of destruction; examples from higher levels including large asteroid strikes, pandemics and global warming. The challenge is to modify the principles of cost–benefit analysis to deal with serious problems caused by uncertainty (as distinct from risk), nonlinearity in value-of-life estimates, the need to project social discount rates into the distant future, and the difficulty of devising suitable policy instruments.

Keywords

catastrophic risk; climate change, economics of; cost–benefit analysis; discount rate; carbon emission tax; global warming; inverse cost–benefit analysis; Kyoto Protocol; research and development (R&D); risk; uncertainty; value of life

Article

The Indian Ocean tsunami of December 2004 and, less than a year later, the flooding of New Orleans as a result of Hurricane Katrina focused attention on a type of disaster to which policymakers pay too little attention – a disaster that has a low or unknown probability of occurring but that, if it does occur, creates enormous losses. Great as were the death toll, the physical and emotional suffering of survivors, and property damage caused by the tsunami, and the even greater property damage caused by the flooding of New Orleans, even greater losses could be inflicted by other disasters of low (but not negligible) or unknown probability. The asteroid that exploded above Siberia in 1908 with the force of a hydrogen bomb might have killed millions of people had it exploded above a major city. Yet that asteroid was only about 200 feet in diameter, and a much larger one (among the thousands of dangerously large asteroids in orbits that intersect the earth's orbit) could strike the earth and, wherever it struck, cause the

total extinction of the human race through a combination of shock waves, fire, tsunamis, and blockage of sunlight. Other catastrophic risks include, besides earthquakes such as the one that caused the 2004 tsunami, natural epidemics (the 1918–19 Spanish influenza epidemic killed between 20 million and 40 million people), nuclear or biological attacks by terrorists, certain types of lab accident (one discussed later in this article), and abrupt global warming. The probability of catastrophes resulting, whether or not intentionally, from human activity appears to be increasing because of the rapidity and direction of technological advances.

The economic approach to catastrophe

It is generally believed that the prediction, assessment, prevention, and mitigation of catastrophes is the province of science. However, economic analysis has an important role to play, as well. Able scientists can commit analytical errors when discussing policy that economists would easily avoid. Thus, Barry Bloom, dean of the Harvard School of Public Health, has criticized the editors of leading scientific journals for having taken the position that ‘an editor may conclude that the potential harm of publication outweighs the potential societal benefits’ (Bloom, 2003, pp. 48, 51). (The specific reference is to publications from which terrorists could learn how to create lethal bioweapons.) Bloom calls this ‘a chilling example of the impact of terrorism on the freedom of inquiry and dissemination of knowledge that today challenges every research university’ (Bloom, 2003, p. 51). The implication – that freedom of scientific research should enjoy absolute priority over every other social value – neglects the need to weigh costs and benefits in order to determine the best balance between public safety and scientific progress.

To illustrate the economic approach to catastrophe, suppose that a tsunami as destructive as the Indian Ocean tsunami occurs on average once a century and kills 250,000 people. That is an average of 2,500 deaths per year. Even without attempting a sophisticated estimate of the value of life to the people exposed to the risk, one can say with some confidence that, if an annual death toll of 2,500 could be substantially reduced at moderate cost, the investment would be worthwhile. A combination of educating the residents of low-lying coastal areas about the warning signs of a tsunami (tremors and a sudden recession in the ocean), establishing a warning system involving emergency broadcasts, telephoned warnings, and air-raid-type sirens, and improving emergency response systems would have saved many of the people killed by the Indian Ocean tsunami, probably at a total cost below any reasonable estimate of the average losses that can be expected from tsunamis. Relocating people away from coasts would be even more efficacious, but, except in the most vulnerable areas or in areas in which residential or commercial uses have only marginal value, the costs would probably exceed the benefits. For annual costs of protection must be matched with annual, not total, expected costs of tsunamis.

As another example, consider the question of optimal precautions against the type of flood that inundated New Orleans. In 1998 it was estimated that it would cost $14 billion to prevent such a flood; the estimated ‘economic’ cost (which ignores the loss of life and physical and emotional suffering) of the recent flood is $100 billion to $200 billion; and the Corps of Engineers estimated the annual probability of such a flood at 1 in 300. If we take the lower cost and assume that the $14 billion investment would eliminate the probability of a flood within 30 years, a period in which the probability of a flood if the measures were not taken would be a shade under ten per cent, yielding an expected

benefit from the flood-control measures of $10 billion, the measures would flunk a cost–benefit test. Note that the calculation does not include discounting future benefits to present value; the reason is that the benefits are likely to grow – a flood that occurred 30 years hence would be likely to do more damage because property values would increase.

Value of life estimates

What might tip the balance in favour of the flood-control measures would be monetizing the expected loss of life and other human suffering. There is now a substantial economic literature inferring the value of life from the costs people are willing to incur to avoid small risks of death; if from behaviour toward risk one infers that a person would pay $70 to avoid a 1 in 100,000 risk of death, his value of life would be estimated at $7 million ($70/.00001), which is in fact the median estimate of the value of life of an American (Viscusi and Aldy, 2003, pp. 5, 18, 63). The value of this transformation is simply that, once a risk is calculated, its expected cost is instantly derived simply by multiplying the risk by the value of life. But there is significant nonlinearity to be considered at both ends of the risk spectrum. At the high end, if one is asked what he would demand to play one round of Russian roulette, the typical answer will be a good deal more than 1/6 of $7 million. At the low-probability end of the risk spectrum, there is a tendency to write the cost of the risk down to or near zero (see, for example, Kunreuther and Pauly, 2004; Viscusi, 1997). In other words, the studies from which the $7 million figure is derived may not be robust with respect to risks of death either much larger or much smaller than the 1 in 10,000 to 1 in 100,000 range of most of the studies – and we do not know what the risk of death from a tsunami was to the people killed, though it was probably towards the low end of the range.

Even if we disregard this issue, because value of life is positively correlated with income, the $7 million figure cannot be used to estimate the value of life of the people killed by the Indian Ocean tsunami, or at least most of them (and perhaps likewise the people killed in the New Orleans flood, most of whom were poor). Additional complications arise from the fact that the deaths were only a part of the cost inflicted by the disaster – the injuries, the suffering, and the property damage that also resulted from the tsunami have to be estimated along with the efficacy and expense of precautionary measures that would have been feasible. The risks of smaller but still destructive tsunamis that such measures might protect against must also be factored in; nor is the ‘once a century’ risk estimate much better than a guess. Nevertheless, it seems apparent that the total cost of the tsunami was high enough to indicate that precautionary measures would have been cost-justified.

The tsunami, unlike the New Orleans flood, could not have been prevented. The only possible precautionary measures would have been either a warning system to enable prompt evacuation or permanently relocating population away from the coastline. Similar measures would have been possible alternatives to preventive measures for New Orleans as well, especially a system for prompt evacuation; but such a system would not have prevented either property damage or massive if temporary population relocation, both of which were huge costs of the flood.

The political economy of catastrophe prevention and response

Since precautionary measures of some kind taken in anticipation of a tsunami on the scale that occurred

would clearly have been cost-justified, why were they not taken? Tsunamis are a common consequence of earthquakes, which themselves are common; and tsunamis can have other causes besides earthquakes – a major asteroid strike in an ocean would create a tsunami that would dwarf the Indian Ocean one. The answer, or answers, may be economic in character.

First, although a once-in-a-century event is as likely to occur at the beginning of the century as at any other time, it is much less likely to occur some time in the first decade of the century than some time in the last nine decades of the century. (The point is simply that the probability is greater the longer the interval being considered: one is more likely to catch a cold in the next year than in the next 48 hours.) Politicians with limited terms of office and thus foreshortened political horizons are likely to discount low-risk disaster possibilities steeply because the risk of damage to their careers from failing to take precautionary measures is truncated.

Second, to the extent that effective precautions require governmental action, the fact that government is a centralized system of control makes it difficult for officials to respond to the full spectrum of possible risks against which cost-justified measures might be taken. Given the variety of matters to which they must attend, officials are likely to have a high threshold of attention below which risks are simply ignored. The US government, preoccupied with terrorist threats, paid insufficient attention to the risk of a disastrous flood of New Orleans, though the risk was understood to be significant.

Third, where risks are regional or global rather than local, many national governments, especially in the poorer and smaller countries, may drag their heels in the hope of taking a free ride on the larger and richer countries. Knowing this, the latter countries may be reluctant to take precautionary measures and by doing so reward and thus encourage free riding. Again, there is a US parallel: state and local government may stint on devoting resources to emergency response, expecting aid from other state and local governments and the federal government.

Fourth, countries are poor often because of weak, inefficient, or corrupt government, characteristics that may disable poor nations from taking cost-justified precautions. Again there is a US parallel: Louisiana is a poor state and New Orleans, which has a very large poor population, has a reputation for having an inefficient and even corrupt government.

And fifth, the positive correlation of per capita income with value of life suggests that it is quite rational for even a well-governed poor country to devote proportionately fewer resources to averting calamities than rich countries do. This would also be true of a poor state or city of the United States.

The failure to act in accordance with cost–benefit principles is dominant characteristic of public policy towards catastrophic risk. An example is the asteroid menace, which is analytically similar to the menace of tsunamis. The National Aeronautics and Space Administration, with an annual budget of more than $10 billion, spends only $4 million a year on mapping dangerously close large asteroids, and at that rate may not complete the task for another decade, even though such mapping is the key to an asteroid defence because it may provide many years of advance warning. Deflecting an asteroid from its orbit when it is still hundreds of millions of miles away from hitting the earth appears to be a feasible undertaking. Although asteroid strikes are less frequent than tsunamis, there have been enough of them to enable the annual probabilities of various magnitudes of such strikes to be estimated, and from these estimates an expected cost of asteroid damage can be calculated. As in the case of tsunamis, if there are measures, beyond those being taken already, that can reduce the expected cost of asteroid damage at a lower cost, thus yielding a net benefit, the measures should be taken, or at least seriously considered.

Cost – benefit analysis under uncertainty

Often it is not possible to estimate the probability or magnitude of a possible catastrophe; the situation is one of uncertainty rather than of risk; how then can cost–benefit analysis, or other techniques of economic analysis, help us in devising responses to such a possibility? The probability of bioterrorism or nuclear terrorism, for example, cannot be quantified; nevertheless, there is rough sense of the range of possible losses that such terrorism would inflict – a range that has no upper limit short of the extinction of the human race – and from this it can be inferred that, even if the probability of such a terrorist attack is small, the expected cost – the product of the probability of the attack and of the consequences if the attack occurs – probably is quite high.

An example of how economic analysis can produce insights even when catastrophic risks are non-quantifiable involves the Relativistic Heavy Ion Collider (RHIC) that went into operation at Brookhaven National Laboratory in Long Island in 2000. As explained by the distinguished English physicist Sir Martin Rees, the collisions in RHIC might conceivably produce a shower of quarks that would ‘reassemble themselves into a very compressed object called a strangelet…. A strangelet could, by contagion, convert anything else it encountered into a strange new form of matter…. A hypothetical strangelet disaster could transform the entire planet Earth into an inert hyperdense sphere about one hundred metres across’ (Rees, 2003, pp. 120–1). Rees considers this ‘hypothetical scenario’ exceedingly unlikely, yet points out that even an annual probability of 1 in 500 million is not wholly negligible when the result, should the improbable materialize, would be so total a disaster.

Concern with such a possibility led John Marburger, the director of the Brookhaven National Laboratory, to commission a risk assessment by a committee of distinguished physicists before authorizing RHIC to begin operating. The committee concluded that the risk of a strangelet disaster was negligible. No cost–benefit analysis of RHIC was conducted, with or without including the risk of a strangelet disaster on the cost side. RHIC cost $600 million to build, and its annual operating costs were expected to be $130 million. No attempt was made to monetize the benefits that the experiments conducted in it were expected to yield; because the experiments are designed to satisfy scientific curiosity rather than to create knowledge that is likely to lead to the invention of useful products, estimation of the benefits is impossible. They may be slight.

The probability of a strangelet disaster in the course of RHIC's planned ten-year life cannot actually be quantified, though there have been attempts. One team of physicists estimated the probability of a strangelet disaster as no more than 1 in 50 million. The official risk-assessment team offered a series of upper-bound estimates, including a 1 in 500,000 probability of a strangelet disaster over the ten-year period, which is 100 times greater than the other's team's estimate. These really are wild, as well as wildly divergent, guesses. Still another uncertainty is what dollar figure to place on the destruction of the earth and all its human and other inhabitants, given the nonlinearity of value of life estimates. Yet, given these uncertainties, the fact that the benefits of RHIC may be quite small suggests that the possibility, remote as it may seem, of a strangelet disaster would weigh heavily, in an economic analysis, against the project. There are more than six billion people on Earth – not to mention unborn future generations – and if their average value of life is estimated at a modest $1 million, the cost of extinction would be $6 quadrillion, and a 1 in 100 million annual risk of a strangelet disaster would yield an annual expected extinction cost of $60 million for ten years to add to the $130 million in annual operating costs and the initial investment of $600 million – roughly a one-third increase in total cost. This could well be

decisive against the project, given its entirely conjectural benefits.

Global warming: risk and response

Another, more familiar, example of the difficulty of quantifying catastrophic risk is the problem of global warming. The Kyoto Protocol, which came into effect by its terms when Russia signed it although the United States has not done so, requires the signatory nations to reduce their carbon dioxide emissions to a level seven to ten per cent below what they were in the late 1990s, but exempts developing countries, such as China, a large and growing emitter, and Brazil, which is destroying large reaches of the Amazon rainforest, much of it by burning. The effect of carbon dioxide emissions on the atmospheric concentration of the gas is cumulative, because carbon dioxide leaves the atmosphere (by being absorbed into the oceans) at a much lower rate than it enters it, and therefore the concentration will continue to grow even if the annual rate of emission is cut down substantially. Between this phenomenon and the exemptions, there is a widespread belief that the Kyoto Protocol will have only a slight effect in arresting global warming; yet the tax or other regulatory measures required to reduce emissions below their level of six years ago will be very costly.

The Protocol's supporters generally are content to slow the rate of global warming by encouraging – by means of heavy taxes (for example, on gasoline or coal) or other measures (such as quotas) that will make fossil fuels more expensive to consumers –conservation measures such as driving less or driving more fuel-efficient cars that will reduce the consumption of these fuels. But from an economic standpoint that is probably either too much or too little. It is too much if, as most scientists believe, global warming will continue to be a gradual process, producing really serious effects – the destruction of tropical agriculture, the spread of tropical diseases such as malaria to currently temperate zones, dramatic increases in violent storm activity (increased atmospheric temperatures, by increasing the amount of water vapour in the atmosphere, increase precipitation), and a rise in sea levels (eventually to the point of inundating most coastal cities) – only toward the end of the 21st century. By that time science, without prodding by governments, is likely to have developed economical ‘clean’ substitutes for fossil fuels (there already *is* a clean substitute – nuclear power) and even economical technologies for either preventing carbon dioxide from being emitted into the atmosphere by the burning of fossil fuels, or removing it from the atmosphere.

But the Protocol is too little and too late, as a response to the costs of global warming, if the focus is changed from gradual to abrupt global warming. At various times in the Earth's history, drastic temperature changes have occurred in the course of just a few years. During the Younger Dryas epoch of about 11,000 years ago, shortly after the end of the last ice age, global temperatures soared by about 14 degrees Fahrenheit in the course of a decade. Because the earth was still cool from the ice age, the effect of the increased warmth on the human population was positive. But a similar increase in a modern decade would have devastating effects on agriculture and on coastal cities, and might even cause a shift in the Gulf Stream that would result in giving all of Europe a Siberian climate.

Because of the enormous complexity of the forces that determine climate, and the historically unprecedented magnitude of human effects on the concentration of greenhouse gases, the possibility that continued growth in that concentration could precipitate – and within the near rather than the distant future – a sudden warming similar to that of the Younger Dryas cannot be excluded. Indeed, no probability, high or low, can be assigned to such a catastrophe. But it may be significant that, while

dissent continues, many climate scientists are now predicting dramatic effects from global warming within the next 20 to 40 years, rather than just by the end of the century (Lempinen, 2005). It may be prudent, therefore, to try to stimulate an increase in the rate at which economical substitutes for fossil fuels, and technology both for limiting the emission of carbon dioxide by those fuels when they are burned in internal-combustion engines or electrical generating plants, and for removing carbon dioxide from the atmosphere, are developed. This can be done by stiff taxes on carbon dioxide emissions. Such taxes give the energy industries, along with customers of theirs such as airlines and manufacturers of motor vehicles, a strong incentive to finance R&D designed to create economical clean substitutes for such fuels and devices to 'trap' emissions at the source before they enter the atmosphere. Given the technological predominance of the United States, it is important that these taxes be imposed on US firms, which they would be if the United States ratified the Kyoto Protocol.

One advantage of the technology-forcing tax approach over public subsidies for R&D is that the government would not be in the business of picking winners – the affected industries would decide what R&D to support – and another is that the brunt of the taxes could be partly offset by reducing other taxes, since emission taxes would raise revenue as well as inducing greater R&D expenditures.

It might seem that subsidies would be necessary for technologies that would have no market, such as technologies for removing carbon dioxide from the atmosphere. There would be no private demand for such technologies because, in contrast to ones that reduce emissions, technologies that remove already emitted carbon dioxide from the atmosphere would not reduce any emitter's tax burden. But this problem is easily solved by making the tax a tax on *net* emissions. Then an electrical generating plant or other emitter could reduce its tax burden by removing carbon dioxide from the atmosphere as well as by reducing its own emissions of carbon dioxide into the atmosphere.

It might seem that, because the demand for conventional fuel sources is inelastic in the short run, the imposition of stiff taxes or quotas required by the Kyoto Protocol would have little effect on the level of emissions. But the significance of the taxes, which actually depends on the inelasticity of demand, is that it would create both pressures and resources for finding a technological fix that would counter the cumulative effect of emissions on the atmospheric concentration of carbon dioxide by driving annual emissions to zero or even below.

Global warming: the discounting problem

A further advantage of focusing on the risk of abrupt rather than gradual global warming is that it allows the vexing problem of discount rate to be elided. The problem is acute when concern focuses on gradual global warming. Suppose that a $10 billion expenditure on capping emissions today would have no effect on human welfare during this century but, by slowing global warming, would produce a savings in social costs of $100 billion in 2100. At a discount rate of three per cent, the present value of $100 billion a century from now is only $5 billion. That would make the expenditure of $10 billion today seem a very poor investment. (For the sake of simplicity, benefits that are expected to accrue after 2100 are ignored in this analysis.) The same amount of money invested in financial instruments could be expected to grow to $192 billion by 2100, on the assumption of a three per cent real interest rate for the next 100 years (though in fact interest rates cannot be forecast over such a long period). If the fund were then disbursed to the victims of global warming, they would be better off than if the $100 billion cost of global warming assumed to be incurred in that year had been averted. Less conservative investments,

moreover, would yield larger expected returns – ten per cent or more rather than three per cent.

But it is not a real alternative to spending $10 billion now to invest it in a fund for future victims of global warming. No such fund will be created, and so they will not be compensated. In circumstances such as this, discounting future to present values is not a method of helping people to decide how to manage their affairs in the way most conducive to maximizing their welfare. Rather, it is a method of maximizing global wealth without regard to its distribution among persons. In the case of gradual global warming, the victims are likely to be concentrated in poor countries, so that basing policy on the discounted costs of global warming would further immiserate the future inhabitants of those countries by increasing the authorized level of emissions harmful to them.

A discount rate based on market interest rates tends to obliterate the interests of remote future generations. The implications are drastic. ‘At a discount rate of five per cent, one death next year counts for more than a billion deaths in 500 years. On this view, catastrophes in the further future can now be regarded as morally trivial’ (Parfit, 1984, p. 357). (What right would the Romans have had to regard our lives as worthless in deciding whether to conduct dangerous experiments?) The trade-off is only slightly less extreme if one substitutes 100 years for 500. At a five per cent discount rate, the present value of one dollar to be received in 100 years is only three-quarters of a cent – and if for money we substitute lives, then to save one life this year we should be willing to sacrifice almost 150 lives a century hence.

And yet not to discount future costs at all would be absurd, certainly as a practical political matter. For then the present value of benefits conferred on our remote descendants would approach infinity. Measures taken today to arrest global warming would confer benefits not only in 2100 but in every subsequent year, perhaps for millions of years. The present value of $100 billion received every year for a million years at a discount rate of zero per cent is $100 quadrillion.

But the vexing problem of how much weight to give to the welfare of remote future generations can be finessed, at least to some extent, if not solved. A discounted present value can be equated to an undiscounted present value simply by shortening the time horizon for the consideration of costs and benefits. For example, the present value of an infinite stream of costs discounted at four per cent a year is equal to the undiscounted sum of those costs for 25 years, while the present value of an infinite stream of costs discounted at one per cent a year is equal to the undiscounted sum of those costs for 100 years. The formula for the present value of one dollar per year forever is $1/*r*, where *r* is the discount rate. So if *r* is four per cent, the present value is $25, and this is equal to an undiscounted stream of one dollar per year for 25 years. If *r* is one per cent, the undiscounted equivalent is 100 years.

One way to argue for the four per cent rate (that is, for truncating our concern for future welfare at 25 years) is to say that we're willing to weight the welfare of the next generation as heavily as our own welfare but that's the extent of our regard for the future. One way to argue for the one per cent rate is to say that we are willing to give equal weight to the welfare of everyone living in this century, which will include us, our children, and our grandchildren, but beyond that we don't care. Looking at future welfare in this way, we may be inclined towards the lower rate, which would have dramatic implications for willingness to invest today in limiting global warming. The lower rate could even be regarded as a ceiling. Most people have some regard for human welfare, or at least the survival of some human civilization, in future centuries. We are grateful that the Romans didn't exterminate the human race in chagrin at the impending collapse of their empire.

Another way to bring future consequences into focus without conventional discounting is by aggregating risks over time rather than expressing them in annualized terms. If we are concerned about what may

happen over the next century, then instead of asking what the annual probability of a collision with a ten-kilometre-wide asteroid is, we might ask what the probability is that such a collision will occur within the next 100 years. An annual probability of 1 in 75 million translates into a century probability of roughly 1 in 750,000. That may be high enough – in view of the consequences if the risk materializes – to justify spending several hundred million, perhaps even several billion, dollars to avert it.

Inverse cost – benefit analysis

A helpful approach to cost–benefit analysis under conditions of extreme uncertainty is what can be called ‘inverse cost–benefit analysis’ (Posner, 2004, pp. 176–84). Analogous to extracting probability estimates from insurance premiums, it involves dividing what the government is spending to prevent a particular catastrophic risk from materializing by what the social cost of the catastrophe would be if it did materialize. The result is an approximation to the implied probability of the catastrophe. Expected cost is the product of probability and consequence (loss): $C = PL$. If P and L are known, C can be calculated. If instead C and L are known, P can be calculated: if $1 billion ($C$) is being spent to avert a disaster that if it occurs will impose a loss (L) of $100 billion, then $P = C / L = .01$.

If P so calculated diverges sharply from independent estimates of it, this is a clue that society may be spending too much or too little on avoiding L. It is just a clue, because of the distinction between marginal and total costs and benefits. The optimal expenditure on a measure is the expenditure that equates marginal cost to marginal benefit. Suppose we happen to know that P is not .01 but .1, so that the expected cost of the catastrophe is not $1 billion but $10 billion. It doesn't follow that we should be spending $10 billion, or indeed anything more than $1 billion, to avert the catastrophe. Perhaps spending just $1 billion would reduce the expected cost of catastrophe from $10 billion all the way down to $500 million and no further expenditure would bring about a further reduction, or at least a cost-justified reduction. For example, if spending another $1 billion would reduce the expected cost from $500 million to zero, that would be a bad investment, at least if risk aversion is ignored.

The federal government is spending about $2 billion a year to prevent a bioterrorist attack (increased to $2.5 billion for 2005 under the rubric of ‘Project BioShield’) (Office of Management and Budget, 2003, pp. 37–8; US Department of Homeland Security, 2004). The goal is to protect Americans, so in assessing the benefits of this expenditure casualties in other countries can be ignored. Suppose the most destructive biological attack that seems reasonably possible on the basis of what little we now know about terrorist intentions and capabilities would kill 100 million Americans. We know that value-of-life estimates may have to be radically discounted when the probability of death is exceedingly slight. But there is no convincing reason for supposing the probability of such an attack less than, say, 1 in 100,000; and the value of life that is derived by dividing the cost that Americans will incur to avoid a risk of death of that magnitude by the risk is about $7 million. Then, if the attack occurred, the total costs would be $700 trillion – and that is actually too low an estimate because the death of a third of the population would have all sorts of collateral consequences, mainly negative. Let us, still conservatively however, refigure the total costs as $1 quadrillion. The result of dividing the money being spent to prevent such an attack, $2 billion, by $1 quadrillion is 1/500,000. Is there only a 1 in 500,000 probability of a bioterrorist attack of that magnitude in the next year? One doesn't know, but the figure seems too low.

It doesn't follow that $2 billion a year is too little to be spending to prevent a bioterrorist attack; one

must not forget the distinction between total and marginal costs. Suppose that the $2 billion expenditure reduces the probability of such an attack from .01 to .0001. The expected cost of the attack would still be very high – $1 quadrillion multiplied by .0001 is $100 billion – but spending more than $2 billion might not reduce the residual probability of .0001 at all. For there might be no feasible further measures to take to combat bioterrorism, especially when we remember that increasing the number of people involved in defending against bioterrorism, including not only scientific and technical personnel but also security guards in laboratories where lethal pathogens are stored, also increases the number of people capable, alone or in conjunction with others, of mounting biological attacks. But there *are* other response measures that should be considered seriously. And one must also bear in mind that expenditures on combating bioterrorism do more than prevent mega-attacks; the lesser attacks, which would still be very costly both singly and cumulatively, would also be prevented.

Costs, moreover, tend to be inverse to time. It would cost a great deal more to build an asteroid defence in one year than in ten years because of the extra costs that would be required for a hasty reallocation of the required labour and capital from the current projects in which they are employed. And so would other crash efforts to prevent catastrophes. Placing a lid on current expenditures would have the incidental benefit of enabling additional expenditures to be deferred to a time when, because more will be known about both the catastrophic risks and the optimal responses to them, considerable cost savings may be possible. The case for such a ceiling derives from comparing marginal benefits to marginal costs; the latter may be sharply increasing in the short run.

See Also

- climate change, economics of
- cost–benefit analysis
- environmental economics
- risk
- social discount rate
- value of life

Bibliography

Bloom, B. 2003. Bioterrorism and the university: the threats to security – and to openness. *Harvard Magazine* 106(2), 48–52.

Kunreuther, H. and Pauly, M. 2004. Neglecting disaster: why don't people insure against large losses? *Journal of Risk and Uncertainty* 28, 5–21.

Lempinen, E. 2005. Scientists on AAAS panel warn that ocean warming is having dramatic impact. AAAS News Release, 17 February. Online. Available at http://www.aaas.org/new/releases/2005/0217warmingwarning.shtml, accessed 21 December 2005.

Office of Management and Budget. 2003. *2003 Report to Congress on Combating Terrorism*. Washington, DC: Executive Office of the President. Online. Available at http://www.whitehouse.gov/omb/inforeg/2003_combat_terr.pdf, accessed 22 December 2005.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Posner, R. 2004. *Catastrophe: Risk and Response*. New York: Oxford University Press.

Rees, M. 2003. *Our Final Hour: A Scientist's Warning; How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century—On Earth and Beyond*. New York: Basic Books.

US Department of Homeland Security. 2004. Fact Sheet: Department of Homeland Security Appropriations Act of 2005. 18 October. Online. Available at http://www.dhs.gov/dhspublic/interapp/press_release/press_release_0541.xml, accessed 22 December 2005.

Viscusi, W. 1997. Economic and psychological aspects of valuing risk reduction. In *Determining the Value of Non-Marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods*, ed. R. Kopp, W. Pommerehne and N. Schwarz. Boston: Kluwer Academic Publishers.

Viscusi, W. and Aldy, Jo. 2003. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27, 5–76.

How to cite this article

Posner, Richard A. "catastrophic risk." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000562> doi:10.1057/9780230226203.0205

The New Palgrave Dictionary of Economics Online

Catchings, Waddill (1879 – 1967)

Robert W. Dimand

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycle; Catchings, W.; Federal Reserve System; Foster, W.; Harrod–Domar growth theory; Keynesianism; short run and long run; monetarism; Pollak Foundation for Economic Research

Article

An investment banker and heterodox monetary economist, Waddill Catchings was born in Sewanee, Tennessee, on 6 September 1879, and died in Pompano Beach, Florida, on 31 December 1967. He graduated from Harvard College in 1901 and Harvard Law School in 1904. Joining the New York City law firm Sullivan & Cromwell on a salary of ten dollars a week, Catchings proved skilful in managing the affairs of companies that went into receivership during the financial panic of 1907, and became president of three ironworks. During the First World War, Catchings worked in the export department of J. P. Morgan & Company, then the US purchasing agent for the British and French governments. A Harvard classmate of Arthur Sachs, Catchings joined Goldman, Sachs & Company in 1918 as partner in charge of underwriting, helping to organize General Foods and National Dairy Products (later Kraft). Catchings complained that his Harvard professors ‘casually explained that their theories would hold true in the long run. But what people are interested in is the short, not the long, run. So I made up my mind that as soon as I had enough money I would set about reconciling these two phases of business – theory and practice’ (quoted in his obituary in the *New York Times*, 1 January 1968). In 1920, Catchings and his Harvard classmate William Trufant Foster (a rhetoric professor and college administrator) established the Pollak Foundation for Economic Research, directed by Foster, funded by Catchings, and dedicated to promoting their belief that, in Catchings's words, ‘If business is to continue zooming, production must be kept at high speed, whatever the circumstances’ (*New York Times* obituary). High and growing levels of production could be maintained by high and growing levels of consumer spending, and the business cycle could be eliminated by appropriate Federal Reserve policy and by keeping public works projects in reserve for economic downturns. In addition to a syndicated newspaper column, Foster and Catchings wrote *Money* (1923), *Profits* (1925), *Business without a Buyer* (1927), *The Road to Plenty* (1928), and *Progress and Plenty* (1930), all Pollak Foundation Studies. Gleason (1959) and Carlson (1962) consider

Foster and Catchings as possible precursors of Keynesian macroeconomics and Harrod–Domar growth theory. The four per cent annual increase in currency and credit endorsed by Foster and Catchings is a possible forerunner of monetarism, but they opposed any mandating of a price level rule, preferring a goal of maintaining prosperity (Tavlas, 1976).

In December 1928, Catchings launched the Goldman Sachs Trading Corporation (GSTC), a closed-end investment trust (ten per cent owned by Goldman, Sachs & Company) which in July 1929 launched the Shenandoah Corporation, another closed-end investment trust, 40 per cent owned by GSTC, followed in August by the Blue Ridge Corporation, with Shenandoah owning a majority of Blue Ridge's common shares. At their peak, this highly leveraged pyramid controlled $500 million of investments, but it was swept away in the stock market crash. GSTC shares, which were initially sold to the public at $104, reached $326 (thanks in part to $57 million that GSTC spent buying its own shares by March 1929, and more purchases later) before falling to $1.75. Catchings had launched Shenandoah and Blue Ridge without consulting the Sachs brothers (who were in Europe in the summer of 1929), and in May 1930 his partners forced his resignation, paying him $250,000 despite his capital account's deficit. Catchings withdrew from the Pollak Foundation (whose endowment disappeared in the crash) to concentrate on his own finances, and moved to California. In the 1950s Catchings was a director of Chrysler, Standard Packaging, and Warner Brothers. After Foster died in 1950, Catchings collaborated with Charles F. Roos (a co-founder of the Econometric Society) on *Money, Men and Machines* (1953). Denouncing Keynesian economics, Catchings and Roos accused the Federal Reserve System of interfering with economic freedom and destabilizing the economy through roller-coaster monetary policies in futile attempts to keep higher wages from causing higher prices. Their book won the Freedoms Foundation's George Washington Honor Medal. Catchings's last books were *Do Economists Understand Business?* (1955), *Bias Against Business* (1956), and *Are We Mismanaging Money?* (1960).

See Also

- Foster, William Trufant
- monetary cranks
- underconsumptionism

Selected works

1923. (With W. Foster.) *Money*. Boston: Houghton Mifflin.

1925. (With W. Foster.) *Profits*. Boston: Houghton Mifflin.

1927. (With W. Foster.) *Business without a Buyer*. Boston: Houghton Mifflin.

1928. (With W. Foster). *The Road to Plenty*. Boston: Houghton Mifflin.

1930. (With W. Foster.) *Progress and Plenty*. Boston: Houghton Mifflin.

1953. (With C. Roos.) *Money, Men and Machines*. 2nd edn. New York: Econometric Institute, 1958.

1955. *Do Economists Understand Business?* New York (privately printed).

1956. *Bias Against Business*. New York (privately printed).

1960. *Are We Mismanaging Money?* New York (privately printed).

Bibliography

Carlson, J. 1962. Foster and Catchings: a mathematical reappraisal. *Journal of Political Economy* 70, 400–2.

Dorfman, J. 1959. *The Economic Mind in American Civilization. Volumes 4 and 5: 1918–1933*. New York: Viking.

Endlich, L. 1999. *Goldman Sachs: The Culture of Success*. New York: Knopf.

Gleason, A. 1959. Foster and Catchings: a reappraisal. *Journal of Political Economy* 67, 156–72.

Tavlas, G. 1976. Some further observations on the monetary economics of Chicagoans and non-Chicagoans. *Southern Economic Journal* 42, 685–92, with comment by J. Davis and reply by Tavlas, 45 (1979), 919–31.

How to cite this article

Dimand, Robert W. "Catchings, Waddill (1879–1967)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_W000132> doi:10.1057/9780230226203.0206

The New Palgrave Dictionary of Economics Online

categorical data

A. Colin Cameron

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Categorical outcome (or discrete outcome or qualitative response) regression models are models for a discrete dependent variable recording in which of two or more categories an outcome of interest lies. For binary data (two categories) probit and logit models or semiparametric methods are used. For multinomial data (more than two categories) that are unordered, common models are multinomial and conditional logit, nested logit, multinomial probit, and random parameters logit. The last two models are estimated using simulation or Bayesian methods. For ordered data, standard multinomial models are ordered logit and probit, or count models are used if ordered discrete data are actually a count.

Keywords

additive random utility model (ARUM); binary outcomes; categorical data; categorical outcome models; choice-based sampling; cumulative distribution function (CDF); discrete outcome models: *see* categorical outcome models; heteroskedasticity; limited dependent variable models; logit models; maximum likelihood; maximum score methods; multinomial models; probit models; qualitative response models: *see* categorical outcome models; random parameters logit model; semiparametric estimation; simulation-based estimation; Tobit models

Article

Categorical outcome models are regression models for a dependent variable that is a discrete variable recording in which of two or more categories, usually mutually exclusive, an outcome of interest lies. Categorical outcome models are also called discrete outcome models or qualitative response models, and are examples of a limited dependent variable model. Different models specify different functional forms for the probabilities of each category. These models are binomial or multinomial models, usually estimated by maximum likelihood.
Key early econometrics references include McFadden (1974), Amemiya (1981), Manski and McFadden

(1981) and Maddala (1983). For textbook treatments see Amemiya (1985), Wooldridge (2002), Greene (2003) and Cameron and Trivedi (2005). The recent econometrics literature has focused on semiparametric estimation (see Pagan and Ullah, 1999) and on simulation-based estimation of multinomial models (see Train, 2003).

Binary outcomes: logit and probit models

Binary outcomes provide the simplest case of categorical data, with just two possible outcomes. An example is whether or not an individual is employed and whether or not a consumer makes a purchase. For binary outcomes the dependent variable *y* takes one of two values, for simplicity coded as 0 or 1. If $y_i = 1$ with probability p_i, then necessarily $y_i = 0$ with probability $1 - p_i$, where *i* denotes the ith of *N* observations. Regressors $\mathbf{x}_i$ are introduced by parameterizing the probability p_i, with

$$p_i = \Pr[y_i = 1|\mathbf{x}_i] = F(\mathbf{x}_i'\beta),$$

where $F(\cdot)$ is a specified function and a single-index form is assumed.
The obvious choice of $F(\cdot)$ is a cumulative distribution function (CDF) since this ensures that $0 < p_i < 1$. The two standard models are the logit model with $p_i = \Lambda(\mathbf{x}_i'\beta) = e^{\mathbf{x}_i'\beta} / (1 + e^{\mathbf{x}_i'\beta})$, where $\Lambda(z) = e^z / (1 + e^z)$ is the logistic CDF, and the probit model with $p_i = \Phi(\mathbf{x}_i'\beta)$, where $\Phi(\cdot)$ is the standard normal CDF.
Interest usually lies in the marginal effect of a change in regressor on the probability that $y = 1$. For the rth regressor, $\partial p_i / \partial x_{ir} = F'(\mathbf{x}_i'\beta)\beta_r$ where F' denotes the derivative of *F*. The sign of β_r gives the sign of the marginal effect, if *F* is a continuous CDF since then $F' > 0$, though the magnitude depends on the point of evaluation $\mathbf{x}_i$. Common methods are to report the average marginal effect over all observations or to report the marginal effect evaluated at $\bar{\mathbf{x}}$.
Parameter estimates are usually obtained by maximum likelihood (ML) estimation. Given p_i, the density can be conveniently expressed as $f(y_i) = p_i^{y_i}(1 - p_i)^{1 - y_i}$. On the assumption of independence over *i*, the resulting log-likelihood function is

$$\ln L(\beta) = \sum_{i=1}^{N} \left\{ y_i \ln F(\mathbf{x}_i'\beta) + (1 - y_i)\ln(1 - F(\mathbf{x}_i'\beta)) \right\}.$$

It can be shown that consistency of the ML estimator requires only that $p_i = F(\mathbf{x}_i'\beta)$, that is, that the functional form for the conditional probability is correctly specified.

There is usually little difference between the predicted probabilities obtained by probit or logit, except for very low and high probability events. For the logit model $\ln[p_i/(1-p_i)] = \mathbf{x}_i'\beta$, so that β_r gives the marginal effect of a change in x_{ir} on the log-odds ratio, a popular interpretation in the biostatistics literature.

A simpler method for binary data is OLS regression of y_i on $\mathbf{x}_i$, with White heteroskedastic robust standard errors used to control for the intrinsic heteroskedasticity in binary data. A serious defect is that OLS permits predicted probabilities to lie outside the (0,1) interval. But it can be useful for exploratory analysis, as OLS coefficients can be directly interpreted as marginal effects and standard methods then exist for complications such as endogenous regressors.

When one of the outcomes is uncommon, surveys may over-sample that outcome. For example, a survey of transit use may be taken at bus stops to over-sample bus riders. This is a leading example of choice-based sampling. Standard ML estimators are inconsistent and instead one must use alternative estimators such as appropriately weighted ML.

The preceding discussion presumes knowledge of F. A considerable number of semiparametric estimators that provide consistent estimates of β given unknown F have been proposed. Manski's (1975) smooth maximum score estimator was a very early example of semiparametric estimation.

Index models

Define a latent (or unobserved) variable y_i^* that measures the propensity for the event of interest to occur. If y_i^* crosses a threshold, normalized to be zero, then the event occurs and we observe $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$. If $y_i^* = \mathbf{x}_i'\beta + u_i$, then

$$p_i = \Pr[y_i^* > 0] = \Pr[-u_i < \mathbf{x}_i'\beta] = F(\mathbf{x}_i'\beta),$$

where $F(\cdot)$ is the CDF of $-u_i$.

The logit model arises if u_i has the logistic distribution. The probit model arises if u_i has the more obvious standard normal distribution, where imposing a unit error variance ensures model identification. The probit model ties in nicely with the Tobit model, where more data are available and we actually observe $y_i = y_i^*$ when $y_i^* > 0$. And it extends naturally to ordered multinomial data.

Random utility models

In many economics applications the binary outcome is determined by individual choice, such as whether

or not to work. Then the outcome should be the alternative with highest utility. The additive random utility model (ARUM) specifies the utility for individual *i* of alternative *j* to be $U_{ij} = \mathbf{x}_{ij}'\beta_j + \varepsilon_{ij}$, $j = 0, 1$, where the error term captures factors known by the decision-maker but not the econometrician. Then

$$p_i = \Pr[U_{i1} > U_{i0}] = \Pr[(\varepsilon_{i0} - \varepsilon_{i1}) \leq \mathbf{x}_{i1}'\beta_1 - \mathbf{x}_{i0}'\beta_0] = F(\mathbf{x}_{i1}'\beta_1 - \mathbf{x}_{i0}'\beta_0)$$

where *F* is the CDF of $(\varepsilon_{i0} - \varepsilon_{i1})$. For components x_{ir} of $\mathbf{x}_i$ that vary across alternatives (so $x_{i0r} \neq x_{i1r}$) it is common to restrict $\beta_{0r} = \beta_{1r} = \beta_r$. For components x_{ir} of $\mathbf{x}_i$ that are invariant across alternatives (so $x_{i0r} = x_{i1r}$) only the difference $\beta_{1r} - \beta_{0r}$ is identified.
The probit model arises, after rescaling, if ε_{i0} and ε_{i1} are i.i.d. standard normal. The logit model arises if ε_{i0} and ε_{i1} are i.i.d. type 1 extreme value distributed with density $f(\varepsilon) = e^{-\varepsilon}\exp(-e^{-\varepsilon})$. The latter less familiar distribution provides more tractable results when extended to multinomial models.

Multinomial outcomes

Multinomial outcomes occur when there are more than two categorical outcomes. With *m* outcomes the dependent variable *y* takes one of *m* mutually exclusive values, for simplicity coded as $1, \ldots, m$. Let p_j denote the probability that the jth outcome occurs. The multinomial density for *y* can be written as $f(y) = \Pi_{j=1}^{m} p_j^{y_j}$ where y_j, $j = 1, \ldots, m$, are *m* indicator variables equal to 1 if $y = j$ and equal to 0 if $y \neq j$. Introducing a further subscript for the ith individual and assuming independence over *i* yields log-likelihood

$$\ln L(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij},$$

where the probabilities p_{ij} are modelled to depend on regressors and unknown parameters β.
There are many different multinomial models, corresponding to different parameterizations of p_{ij}.

Unordered multinomial models

Usually the outcomes are unordered, such as in choice of transit mode to work. The benchmark model

for unordered outcomes is the multinomial logit model. When regressors vary across alternatives (such as prices), the conditional logit (CL) model specifies $p_{ij} = e^{\mathbf{x}_{ij}'\beta} / \sum_{k=1}^{m} e^{\mathbf{x}_{ij}'\beta}$. If regressors are invariant across alternatives (such as gender), the multinomial logit (MNL) model specifies $p_{ij} = e^{\mathbf{x}_i'\beta_j} / \sum_{k=1}^{m} e^{\mathbf{x}_i'\beta_k}$, with a normalization such as $\beta_1 = 0$ to ensure identification. In practice some regressors may be a mix of invariant and varying across alternatives; such cases can be re-expressed as either a CL or MNL model.

The CL and MNL models reduce to a series of pairwise choices that do not depend on the other choices available. For example, the choice between use of car or red bus is not affected by whether another alternative is a blue bus (essentially the same as the red bus). This restriction, called the assumption of independence of irrelevant alternatives, has led to a number of alternative models.

These models are based on the ARUM. Suppose the jth alternative has utility $U_{ij} = \mathbf{x}_{ij}'\beta + \varepsilon_{ij}, \quad j = 1, \ldots, m$. Then

$$p_{ij} = \Pr[U_{ij} \geq U_{ik} \text{ for all } k] = \Pr[(\varepsilon_{ik} - \varepsilon_{ij}) \leq (\mathbf{x}_{ij}'\beta - \mathbf{x}_{ik}'\beta) \ \forall \ k].$$

The CL and MNL models arise if the errors ε_{ij} are i.i.d. type 1 extreme value distributed. More general models permit correlation across alternatives j in the errors ε_{ij}.

The most tractable model with error correlation is a nested logit model. This arises if the errors are generalized extreme value distributed. This model is simple to estimate but suffers from the need to specify a particular nesting structure.

The richer multinomial probit model specifies the errors to be m–dimensional multivariate normal with $(m + 1)$ restrictions on the covariances to ensure identification. In practice it has proved difficult to jointly estimate both β and the covariance parameters in this model. A recent popular model is the random parameters logit model. This begins with a multinomial logit model but permits the parameters β to be normally distributed. For these two models there is no closed form expression for the probabilities and estimation is usually by simulation methods or Bayesian methods.

Ordered multinomial models

In some cases the outcomes can be ordered, such as health status being excellent, good, fair or poor. The starting point is an index model, with single latent variable, $y_i^* = \mathbf{x}_i'\beta + u_i$. As y^* crosses a series of increasing unknown thresholds we move up the ordering of alternatives. For example, for $y^* > \alpha_1$ health status improves from poor to fair, for $y^* > \alpha_2$ it improves further to good, and so on. For the ordered logit (probit) model the error u is logistic (standard normal) distributed.

An alternative model is a sequential model. For example, one may first decide whether or not to go to

college $(y = 1)$ and if chose college then choose either two-year college $(y = 2)$ or four-year college $(y = 3)$. The two decisions may be modelled as separate logit or probit models.

A special case of ordered categorical data is a count, such as number of visits to a doctor taking values 0, 1, 2,•…•. An ordered model can be applied to these data, but it is better to use count models. The simplest count model is Poisson regression with exponential conditional mean $E[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\beta)$. Common procedures are to use the Poisson but obtain standard errors that relax the Poisson restriction of variance-mean equality, to estimate the richer negative binomial model, or to estimate hurdle or two-part models or with-zeroes models that permit the process determining zero counts to differ from that for positive counts.

Multivariate outcomes and panel data

Multivariate discrete data arise when more than one discrete outcome is modelled. The simplest example is bivariate binary outcome data. For example, we may seek to explain both employment status (work or not work) and family status (children or no children). The standard model is a bivariate probit model that specifies an index model for each dependent variable with normal errors that are correlated. Such models can be extended to permit simultaneity.

For panel binary data the standard model is an individual specific effects model with $p_{it} = F(\alpha_i + \mathbf{x}_{it}'\beta)$ where α_i is an individual specific effect. The random effects model usually specifies $\alpha_i \sim N[0, \sigma_\alpha^2]$ and is estimated by numerically integrating out α_i using Gaussian quadrature. The fixed effects model treats α_i as a fixed parameter. In short panels with few time periods consistent estimation of β is possible in the fixed effects logit but not the fixed effects probit model. If $\mathbf{x}_{it}$ includes $y_{i,t-1}$, a dynamic model, fixed effects logit is again possible but requires four periods of data.

See Also

- contingent valuation
- hierarchical Bayes models
- logit models of individual choice
- maximum score methods
- semiparametric estimation
- simulation-based estimation

Bibliography

Amemiya, T. 1981. Qualitative response models: a survey. *Journal of Economic Literature* 19, 1483–536.

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Cameron, A. and Trivedi, P. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Greene, W. 2003. *Econometric Analysis*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

Maddala, G. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Manski, C. 1975. The maximum score estimator of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C. and McFadden, D., eds. 1981. *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

Pagan, A. and Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Wooldridge, J. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

How to cite this article

Cameron, A. Colin. "categorical data." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000561> doi:10.1057/9780230226203.0207

The New Palgrave Dictionary of Economics Online

Catholic economic thought

Pedro Teixeira and António Almodovar

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Although Catholic economics' roots date back to the beginnings of Christianity, its emergence as a structured discourse developed later and slowly. The establishment of a distinctive Catholic approach to modern social and economic problems had to await a more extensive development of the market system and the emergence of political economy. The most prolific period for Catholic economic thought began in 1891 and continued until the end of the Second World War. In the second half of the 20th century the church's interest focused on the analysis of such themes as development, international aid and cooperation.

Keywords

capitalism; Catholic economic thought; charitable contributions; class; concentration of wealth; economic freedom; exploitation; fair trade; foreign aid; just wage; natural law; poverty; redistribution of wealth; *Rerum Novarum* ; Salamanca School; social justice; social responsibility; socialism; solidarity; subsidiarity; usury

Article

Catholic economic thought is the outcome of a series of efforts to evaluate the workings of economic life according to a definite set of religious principles. In its more evolved forms, these efforts have inevitably led to include the findings of political economy, and later of economics, in its assessment of economic life, but also to assess the findings of economic analysis itself. According to a strict ecclesiological perspective, only the hierarchy of the Church is authorized to identify the appropriate religious principles that are to be applied to the analysis of the livelihood of man. Therefore, some of the assessments made by Catholics may be considered by the Church's hierarchy as inappropriate. Catholic economic thought is not to be confused with the social doctrine of the Catholic Church. Since 1891, the most relevant religious principles for the appraisal of social questions from a theological

perspective are gathered in the social doctrine of the Church, which is essentially based in the so-called social encyclicals, which are official documents written by several popes, often based on documents prepared by other high-ranking Church officials. These documents emerged as attempts to offer a better moral and philosophical framework for the workings of a modern society, not as in-depth and systematic discussions of man's economic life or as blueprints for a thorough discussion of economic concepts and theories. By being focused on the material aspects of life, Catholic economic thought is prone to give more emphasis to particular problems – such as usury and finance, social and labour questions or, later, the outline of an alternative economic and social system. However, Catholic economic literature has as a rule been less focused than political economy on technical aspects.

Catholic economic thought has an inescapable doctrinal and normative accent. Its '*ought*' sentences are considered as quasi-*positive* ones, in the sense that they were allegedly meant by God to become factual statements in a society functioning in accordance with natural law (see Barrera, 2001, pp. 117–31). This normative stance acts as an explicit incentive for social action, in order both to amend the workings of existing institutions and to establish new ones – such as charitable institutions, cooperatives, institutions of mutual assistance, and particular ways of labour–capital association. In certain periods, when Catholics were more openly engaged in the revision of economic life, their thought went as far as to suggest the establishment of a specific economic system, which was a third way between the liberal and the socialist ones. But, even when they were more focused on the implementation of particular social and economic measures, people engaged in these initiatives also left some thoughts that are of more general interest.

Early attempts to formulate Catholic economic thought

Although the roots of Catholic economics date back to the beginnings of Christianity, the emergence of a structured discourse developed later and slowly. Thus, even if some of the basic Catholic principles for social and ethical teaching were already present in the gospels and in the patristic literature, the systematic theology of Aquinas was instrumental in the move towards a more organized approach to economic problems. The earliest scholarly attempt to produce an explicit and meaningful set of theological principles applied to economic problems was performed in the 16th century by authors belonging to the Salamanca School. Under the philosophical umbrella provided by Thomism, Dominicans like Vitoria, Soto, and Mercado, and Jesuits like Molina, Mariana and Lugo addressed the problems of usury, prices, and justice in wages. Although these ideas were not formally adopted by the Church, this literature was widely used by confessors in search of appropriate answers for the moral questions raised by the development of economic activity (on the economic thought of the school of Salamanca, see Grice-Hutchinson, 1978; 1993, and Camacho, 1998; these authors are also relevant as examples of a revival of Thomist moral theology, which they applied to international law: see Curran, 2002).

The 19th century

The establishment of a distinctive and clear-cut Catholic approach to modern social and economic problems had nevertheless to wait for a more extensive development of the market system and the

emergence of political economy. By the late 1830s, the first Catholic political economists were already trying to infuse some basic Christian values into the teachings of classical political economy. Together with the socialists, they were concerned about the consequences of unbridled competition, the concentration of riches in the hands of the few, the exploitation of the poor and weak, and the existence of pervasive unemployment. However, contrary to socialists, Catholics thought that those evils, together with excessive materialism and burgeoning social and political unrest, were to be curbed by individuals renouncing material goods and by extended charity, not by abolishing private property or an expansion of the state. Their criticism voiced the fundamental Christian values of universal fraternity and respect for human dignity, as expressed in the Gospels and in the Apostolic letters.

It is important to note that in the mid-19th century there was a series of authors who wrote on economic subjects from a Catholic perspective before the *Rerum Novarum*, the encyclical of Pope Leo XIII on capital and labour, promulgated 15 May 1891. Among these we find the names of Charles de Coux, Alban de Villeneuve-Bargemont, Joseph Droz, Charles Périn, and Matteo Liberatore. The first four authors are representative of the Catholic perspectives that emerged gradually in the context of 19th-century France and Belgium. Three of them – Coux, Périn and Droz – were openly against any solution for economic problems that would require increased state intervention, and they asked the rich to voluntarily avoid all extreme forms of exploitation and competition; as a rule, they were reasonably sympathetic towards political economy, and may be considered as the forerunners of the conservative tendency that was later organized around the Angers school. Villeneuve-Bargemont had less confidence in voluntary individual action as a remedy for the emerging poor question. Contrary to the Catholic *conservative* approach, Villeneuve-Bargemont thought that the scale of the problem was so serious that the state should intervene in favour of the labouring masses before they fell irrevocably under the spell of socialism. Thus he may be considered as a precursor of the so-called *progressive* tendency, later developed by the Fribourg Union and the Liège school. Matteo Liberatore deserves mention, since he was one of the persons involved in the drafting of Leo XIII's *Rerum Novarum* (1891). His views were closer to Villeneuve-Bargemont than to Charles Périn, since he believed that modern poverty was a phenomenon that could not be solved by traditional means (charity), because its causes were embedded in modern social and economic organization. Modern exploitation and modern social unrest were seen not only as consequences of the acceptance of a social and economic model based on the erroneous philosophical notions underlying political economy, but also because the spread of the latter stimulated people to act in a way that damaged social cohesion. Contrary to materialist and utilitarian views, wealth should considered as a means, not an end, and should be distributed according to justice; and the human person should always be respected – meaning that in no circumstance should labour be considered as a mere commodity to be bought and sold in the market. Once individualism and competition were once again checked by an attention to mutual needs, modern phenomena such as the class struggle (between labour and capital) would vanish and a sense of mutually beneficial collaboration would take its place. Measures such as the re-establishment of updated medieval corporations – which had to be adapted to the new realities and not just re-established, were a possible institutional solution for bringing peace and harmony to the relations between producers, namely because they would help to resume natural social relations and reduce the moral, social, and professional void in which liberalism had placed the individuals. Efforts to promote the modern resurgence of these institutions were at the origins of what later became corporatism (see below).

The Golden Age, 1891 – 1940s

The most prolific period for Catholic economic thought began in 1891 and continued up to the end of the Second World War. Stirred by Leo XIII's *Rerum Novarum*, the willingness to address social and economic questions gave rise to extensive debate and to intense publishing activity (see De Rosa 2004; Hobgood, 1991, p. 112).

The central issue of *Rerum Novarum* is the condition of workers, especially industrial workers, and the moral and material risks arising from what was seen as their degrading situation. Leo XIII made clear from the outset that he considered the major cause to be the political and economic transformations of the previous hundred years. This had destroyed or seriously damaged valuable traditional social structures such as medieval corporations. It had also launched a process of secularization of the legal and political framework, which had greatly diminished the moral influence of the Church. Liberalism had created a social vacuum in which unregulated competition, greed and usury had prospered, resulting in a substantial concentration of wealth and power. The latter eventually created an unbalanced distribution of privileges that made possible the exploitation of the workers by the all-mighty owners of capital. Leo XIII also asserted that the supposed remedies offered by socialists were inadequate to the task. In addition to the obvious problem of atheism, the crucial issue in the Church's critique of socialism was the former's concern with private property. Although the Church criticized the extreme capitalist/individualist/utilitarian uses of private property, these criticisms did not question its fundamental existence.

The Church proposed a new relationship between workers and capitalists. Workers should opt for non-violent ways of solving labour disputes, and should perform faithfully and completely the tasks that were allocated to them. In return, paramount among the duties of capitalists was the acknowledgment of and the respect for the human dignity of the workers. This meant respecting the workers' physical and intellectual health, and the payment of a fair family wage that would put a stop to the need for female and youth labour. Capitalists' social responsibility was central to the way Christians should relate to wealth. Leo XIII underlined the ephemeral and secondary nature of earthly wealth and success. If the Church accepted the inequality of property, it also cared for the poorest members of society, knowing that, unless these were actively supported, they would fall into a state of quasi-serfdom, which would lead to social disruption.

One of the outcomes of the *Rerum Novarum* was the development of an array of books, typically bearing the title of *Principles or Courses on Social Economics*. Often written by Jesuits for the use of both clergy and active Catholic laity, this peculiar type of book tried to re-embed the political economy into a social philosophy so as to secure a coherent and global society based on Christian values (Galindo, 1996, p. 143). The authors of such works had to perform complex scholarly work if they were to fulfil their aim. First, they had to explain classical political economy to their readers; then they had to introduce and explain the Pope's criticisms of the philosophical tenets underlying economic liberalism; next they had to deal with socialism, in order to make sure that this doctrine would not be seen as a possible alternative to the shortcomings of economic liberalism; and finally, they had to highlight the proper course of Catholic thought and action that was to be followed in order to put right contemporary evils. Authors that engaged in this type of work include Charles Antoine, S.J. (1896), Giuseppe Toniolo (1907–9) and Heinrich Pesch, S.J. (1905–26), the latter being considered by Schumpeter as the best

example of neo-scholasticism (1954, p. 765). Another set of books focused on the outlines of a specifically Catholic system – a third, neo-corporative, way between liberalism and socialism. This system had its roots both in France (Mun, La Tour du Pin) and Germany (Vogelsang, Kettler), and was further developed under the auspices of the Liège School. It was eventually accepted, if not warmly supported, by the encyclical *Quadragesimo Anno*.

Quadragesimo Anno appeared in 1931, when Pius XI took the initiative in clarifying and updating the position of the Catholic Church on the economic and social condition of the contemporary world. His view was that, although capitalism per se was not an evil system, there was a problem with the way it had developed, for it had led to economic despotism, namely, a concentration of wealth which gave to a few members of society huge power, which was often used to influence and subjugate governments and countries. The subjugation of the state to the interests of a wealthy minority, whose power was nurtured by ambition, greed and speculative behaviour, fostered social disorder and could lead to the collapse of essential social bonds.

The Church supported the existence of private property, but it also underlined its dual nature (individual and social) and the difference between property ownership and property usage. Hence, the relations between capitalists and workers in the capitalistic system should be reorganized according to this view. According to Pius XI, labour and capital did have common interests, and this communality of efforts and purposes called for a sharing of both the responsibility for the productive process and of the wealth created, including the profits resulting from the productive activity. Commutative justice would be insufficient, and should be complemented by social justice.

Pius XI also emphasized the principle of subsidiarity. According to this principle, the state should not intervene when intermediate levels of society (associations, local community, and family) could act effectively. Social harmony ought therefore to be built upon the contribution of intermediate communities and groups, these taking multiple forms. However, the reconstruction of the social fabric, which had been ruined by unlimited competition and the concentration of wealth, required the state to regulate competition, subordinating it to the higher values of justice and charity. To accomplish the necessary rebalance of social power in order to promote the common good, Pius XI made explicit references to the advantages and risks of the emerging corporatist organization (in Italy and elsewhere). Overall, he thought that the advantages (pacifying society, curbing the insidious influence of socialist organizations, and bringing together workers and capitalists in the search for the common good) could outweigh the possible risks of bureaucratization and state dirigisme. Pius XI saw the establishment of the corporative system as a step in the right direction, towards a Christian social-economic order, through its contribution to a harmonious society and its emphasis on the pursuit of the common good.

The post-war period

At the beginning of the 1960s, the Catholic Church underwent profound institutional and theological changes. With the Second Vatican Council (1962–5), Thomism, the theological and philosophical basis of the earlier social and economic doctrine of the Church, lost its unique status (see Nichols, 2002, pp. 139–43). Vatican II also marked a change in the role of the laity, and opened the dialogue between different churches. Although until the early 1960s, socialism and communism stood at the forefront of Church's criticisms, some bridges were later to be established with Marxist sociology (see Curran, 2002,

pp. 201, 203).

The Catholic Church's approach to economic problems also took a different direction in the second half of the 20th century, now focusing in the analysis of themes like development and North–South relationships, international aid and cooperation. This is particularly visible in John XXIII's *Mater et Magistra* (1961) and Paul VI's *Populorum Progressio* (1967). In the latter, Paul VI considered that the wealthiest nations had the duties of solidarity, justice and charity towards less developed ones, and that these duties should be addressed through international aid, fair trade and a framework conducive to mutual progress. He was particularly critical of free trade, since he regarded any exchange between unequals as potentially unjust. Hence, he called for fair and just competition between nations.

The social question was nevertheless not forgotten. In the *Mater et Magistra*, John XXIII stated that wages should not be left to market forces alone, for they should be determined by the laws of justice and equity. Private property was not to be considered solely as a right that should be protected, but also as an obligation to practise solidarity among human beings. John XXIII also gave explicit support to the political organization of workers in order to promote their legitimate rights. This text was also the first to address, and largely support, the so-called welfare state and its associated system of social insurance and social security, on the grounds of its contribution to the desirable redistribution of wealth. Although the Church kept its distance vis-à-vis socialism, Paul VI considered that there were some possibilities for cooperation between Catholics and socialist movements insofar as this contributed to a more just society (see his apostolic letter *Octogesima Adveniens* on the occasion of the 80th anniversary of *Rerum Novarum*, 1971).

The dialogue between economic analysis and theology was, if not on hold, at least withdrawn to the backstage. This is likely to have been for several different reasons, ranging from the changing priorities in theology and a new emergence of ecclesiological concerns with the inner life of the Church, to the growing professionalization of economics, which made it ever more difficult to acquire the desirable proficiency in both fields (see Wilson, 1997, pp. 88–9 and 113). In the late 1950s, Catholic writers like Achille Dauphin-Meunier and Jean-Yves Calvez had already begun to assert that the Church had no other wish than to present its own social doctrine. To these authors, the Church was not to offer or to support ‘an economic theory’ but only a ‘philosophical and religious clarification of the fundamental aspects of human existence within economic relationships’ (Calvez and Perrin, 1958, p. 11).

The contemporary situation

Catholic social doctrine received a significant stimulus in the 1980s and 1990s with John Paul II. He used the 90th and 100th anniversaries of *Rerum Novarum* to express views on the economic realm. In *Laborem exercens* (1981) he focused on the role of work as a central feature of all human activity and therefore of all economic activity. He considered that contemporary developments in technological, economic and political conditions had reinforced the pastoral care that the Church should associate with all issues related to work, such as unemployment and lifelong learning. He criticized what he considered the error of considering human labour solely according to its economic purpose, and underlined the principle of priority of human labour over capital, which should not be attained through class or social warfare but by peaceful struggle for social justice. Likewise, in *Centesimus Annus* (1991) he focused on the harshness of the modern conditions of the working class and pointed out how erroneous the

collectivist and totalitarian solution was. Thus he insisted on the idea of redistribution of wealth in order to fulfil ‘the universal destination of material goods’. John Paul II also devoted special attention to economic and social development, with particular attention being paid to issues such as international division of labour, international debt and poverty. In the encyclical *Sollicitudo Rei Socialis* (1987) he criticized both ‘liberal capitalism’ and ‘Marxism collectivism’ and proposed a view of ‘authentic human development’ which was not only economic but also social and spiritual. Thus, underdevelopment had not only social and economic causes, but also moral ones, not the least being the lack of international solidarity that denied human interdependence beyond national or political borders. His position vis-à-vis social warfare and any possible analytical or political convergence with Marxism is vividly illustrated by the reaction of the Church's hierarchy to Liberation Theology, whose main proponents were either silenced or led to abandon the Catholic Church because of the restrictions imposed on them regarding teaching, preaching and writing.

Modern Catholic theology has focused on achieving a comprehensive and coherent presentation of social ethics (see Curran, 2002). Those who give a certain emphasis to economic aspects (see Barrera, 2001; Hobgood, 1991), always take care to reiterate ‘the caveat that [the Church's social teachings do] not offer an alternative school of thought between classical laissez-faire capitalism and socialist centralized planning’ (Barrera, 2001, p. viii). Notwithstanding this change of focus, the modern effort to systematize the teachings of the encyclicals has led in some cases to the identification of six basic principles: universal access, the primacy of labour, subsidiarity, socialization, solidarity, and stewardship (2001, p. 1, and table on p. 258). By means of these principles, the criticisms addressed to economics continue to stress its defective philosophical base and go on emphasizing the collective risks that are incurred by a society unwilling to restrain excessively individualist, materialist, and utilitarian behaviour. The claims of contemporary Catholic economic thought therefore continue to emphasize the need for justice and equity, something that can be achieved only through the establishment of corrective measures to the workings of the market in order to prevent its deleterious action on the social fabric. The basic appeal therefore remains, that economics should not refuse the normative approach provided by the Catholic view of mankind.

See Also

- Aquinas, St Thomas
- ethics and economics
- religion and economic development
- scholastic economics

Bibliography

Barrera, A.O.P. 2001. *Modern Catholic Social Documents and Political Economy*. Washington, DC: Georgetown University Press.

Calvez, J.-Y. and Perrin, J. 1958. *Église et Société Économique. L'énseignement social des Papes de*

Léon XIII à Pie XII. Paris: Éditions Montaigne.

Camacho, F.G. 1998. *Economia y Filosofia Moral. La formación del pensamiento económico europeo en la Escolástica española*. Madrid: Editorial Sintesis.

Curran, C.E. 2002. *Catholic Social Teaching, 1891–present: A Historical, Theological and Ethical Analysis*. Washington, DC: Georgetown University Press.

De Rosa, G. 2004. *I tempi della 'Rerum Novarum'*. Rome: Rubbettino/Istituto Luigi Sturzo.

Galindo, A. 1996. *Moral Socioeconómica*. Madrid: Biblioteca de Autores Cristianos.

Grice-Hutchinson, M. 1978. *Early Economic Thought in Spain, 1177–1740*. London: George Allen & Unwin.

Grice-Hutchinson, M. 1993. *Ensayos sobre el pensamiento económico en España*. Madrid: Alianza Universidad.

Hobgood, M.E. 1991. *Catholic Social Teaching and Economic Theory: Paradigms in Conflict*. Philadelphia: Temple University Press.

Nichols, A.O.P. 2002. *Discovering Aquinas: An Introduction to his Life, Work and Influence*. London: Darton, Longman & Todd.

Schumpeter, J.A. 1954. *A History of Economic Analysis*. London: Routledge, 1994.

Wilson, R. 1997. *Economics, Ethics and Religion: Jewish, Christian and Muslim Economic Thought*. London: Macmillan.

How to cite this article

Teixeira, Pedro and António Almodovar. "Catholic economic thought." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000075>
doi:10.1057/9780230226203.0208

The New Palgrave Dictionary of Economics Online

causality in economics and econometrics

Kevin D. Hoover
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economics was conceived as early as the classical period as a science of causes. The philosopher–economists David Hume and J. S. Mill developed the conceptions of causality that remain implicit in economics today. This article traces the history of causality in economics and econometrics, showing that different approaches can be classified on two dimensions: process versus structural approaches, and a priori versus inferential approaches. The variety of modern approaches to causal inference is explained and related to this classification. Causality is also examined in relationship to exogeneity and identification.

Keywords

Aristotle; causal inference; causality in economics and econometrics; correlation; Cowles Commission; econometrics; Edgeworth, F. Y.; endogeneity and exogeneity; Granger–Sims causality; graph theory; Hume, D.; identification; index numbers; Induction; instrumental variables; Jevons, W. S.; microfoundations; Mill, J. S.; Mises, L. von; natural experiments; observational equivalence; process analysis; Quetelet, A.; rational expectations; regression; Robbins, L. C.; Simon, H.; Smith, A.; statistical inference; structural vector autoregressions; Tinbergen, J.; vector autoregressions; Wold, H. O. A.

Article

1 Philosophers of economics and causality

The full title of Adam Smith's great foundational work, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), illustrates the centrality of causality to economics. The connection between causality and economics predates Smith. Starting with Aristotle, the great economists are frequently also the great philosophers of causality. Aristotle's contributions to economics are found principally in the *Topics*, the *Politics*, and the *Nicomachean Ethics*, while he lays out his famous four causes (material, formal, final and efficient) in the *Physics*. Material and formal causes are among the concerns of economic ontology, a subject addressed by philosophers of economics (see, for example, Mäki, 2001) albeit rarely by practicing economists. Sometimes, as for example in Karl Marx's grand theory of capitalist development, economists have appealed to final causes or teleological explanation (for a defence, see Cohen, 1978; for a general discussion, see Kincaid, 1996). But, for the most part, taking physical sciences as a model, economics deals with efficient causes. What is it that makes things happen? What explains change? (See Bunge, 1963, for a broad account of the history and philosophy of causal analysis.)

The greatest of the philosopher/economists, David Hume, set the tone for much of the later development of causality in economics. On the one hand, economists inherited from Hume the sense that practical economics was essentially a causal science. In ‘On Interest’, Hume (1742, p. 304) writes:

> it is of consequence to know the principle whence any phenomenon arises, and to distinguish between a cause and a concomitant effect. Besides that the speculation is curious, it may frequently be of use in the conduct of public affairs. At least, it must be owned, that nothing can be of more use than to improve, by practice, the method of reasoning on these subjects, which of all others are the most important; though they are commonly treated in the loosest and most careless manner.

On the other hand, Hume doubted whether we could ever know the essential nature of causation ‘in the objects’ (Hume, 1739, p. 165). Coupled with a formidable critique of inductive inference more generally, Hume's scepticism has contributed to a wariness about causal analysis in many sciences, including economics (1739; 1777). The tension between the epistemological status of causal relations and their role in practical policy runs through the history of economic analysis since Hume.

2 History

2.1 Hume's foundational analysis

Although Hume's dominant concerns are moral, historical, political, and social (including economic), physical illustrations serve as his paradigm causal relationships. *A* (say, a billiard ball) strikes *B* (another ball) and causes it to move. Any analysis must address two key features of causality: first, causes are asymmetrical (in general, if *A* causes *B*, *B* does not cause *A*). Hume sees temporal succession (the movement of *A* precedes the movement of *B*) as accounting for asymmetry. Second, causes are effective. A cause must be distinguished from an accidental correlation and must bring about its effect. Hume sees spatial contiguity (the balls touch) and necessary connection (the movement of *B* follows of necessity from the movement of *A*) as distinguishing causes from accidents and establishing their effectiveness.

Hume was famously sceptical of any idea that could not be traced either to logical or mathematical deduction or to direct sense experience. He asks, whence comes the idea of the necessary connection of cause and effect? It cannot be deduced from first principles. So, he argues that our idea of necessary connection, which he concedes is the most characteristic element of causality, can arise only from our experience of the constant conjunction of particular temporal sequences. But this then implies that causality stands on a very weak foundation. For one corollary of Hume's belief that all ideas are based either in logic or sense experience was that we do not have any secure warrant for inductive inference. Neither logic nor experience (unless we beg the question by implicitly assuming the truth of induction) gives us secure grounds from observing instances to inferring a general rule. Therefore, what we regard as necessary connection in causal inference is really more of habit of mind without clear warrant. Causes may be necessarily connected to effects; but, for Hume, we shall never know in what that necessary connection consists.

While later philosophers have differed with Hume on the analysis of causality, his views were instrumental in setting the agenda, not only for philosophical discussions, but for practical causal analysis as well.

2.2 The 19th century: logic and statistics

Even more influential than Hume in shaping economics, John Stuart Mill, another philosopher/economist, was less sceptical about causal inference in general, but more sceptical about its application to economics. In his *System of Logic* (1851), Mill advanced his famous canons of induction: the methods of (*a*) agreement, (*b*) difference, (*c*) joint (or double) agreement and difference, (*d*) residues, and (*e*) concomitant variations. For example, according to the method of difference, if we have two sets of circumstances, one in which a phenomenon occurs and one in which it does not, and the circumstances agree in all but one respect, that respect is the cause of the phenomenon. Mill's canons are essentially abstractions from the manner in which causes are inferred in controlled experiments. As such, Mill doubted that the canons could be easily applied to social or economic situations, in which a wide variety of uncontrolled factors are obviously relevant. Mill argued that economics was what Daniel M. Hausman (1992) has called an ‘inexact and separate science’, whose general principles were essentially known a priori and which held only subject to *ceteris paribus* clauses. Mill's apriorism proved to be hugely influential in later economics. Lionel Robbins (1935) expressed considerable scepticism about the place of empirical studies within economic science. Some Austrian economists, such as Ludwig von Mises (1966), went so far as to deny that economics could be an empirical discipline at all. Mill's apriorism also influenced those economists who see economic theory as similar to physical theory as a domain of universal laws.

Other 19th-century economists were less sceptical about the application of causal reasoning to economic data. For instance, W. Stanley Jevons (1863) pioneered the construction of index numbers as the core element of an attempt to prove the causal connection between inflation and the increase in worldwide gold stocks after 1849. Jevons's investigation can be interpreted as an application of Mill's method of residues (see Hoover and Dowell, 2001). He saw the various idiosyncratic relative price movements, owing to supply and demand for particular commodities, as cancelling out to leave the common factor that could only be the effect of changes in the money stock.

The 19th century witnessed extensive development in the theory and practice of statistics (Stigler, 1986). Inference based on statistical distributions and correlation measures was closely connected to causality. Adolphe Quetelet envisaged the inferential problem in statistics as one of distinguishing among constant, variable, and accidental causes (Stigler, 1999, p. 52). The economist Francis Ysidro Edgeworth pioneered tests of statistical significance (in fact Edgeworth may have been the first to use this phrase). He glossed the finding of a statistically significant result as one that ‘comes by cause’ (Edgeworth, 1885, pp. 187–8).

2.3 The 20th century: causality and identification

Further developments of statistical techniques, such as multiple correlation and regression, in the 20th century were frequently associated with causal inference. It was fairly quickly understood that, unlike correlation, regression has a natural direction: the regression of *Y* on *X* does not produce coefficient estimates that are the algebraic inverse of those from the regression of *X* on *Y*. The direction of regression should respect the direction of causation.

By the early 20th century, however, the dominant vision of economics was one in which prices and quantities are determined simultaneously. This is as much true for Alfred Marshall (1930), who is often described (not perfectly accurately) as an advocate of partial equilibrium analysis, as it is for Léon Walras (1954), the principal font of modern general equilibrium analysis. Simultaneity does not necessarily rule out causal order, though it does complicate causal inference. Although regressions may have a natural causal direction, there is nothing in the data on their own that reveal which direction is the correct one – each is an equally eligible rescaling of a symmetrical and non-causal correlation. This is a problem of observational equivalence. And it is the obverse side of the now familiar problem of econometric identification: in this case, how can we distinguish a supply curve from a demand curve? The problem of identification was pursued throughout most of the first half of the 20th century until the fairly complete treatment by the Cowles Commission at mid-century (Koopmans, 1950; Hood and Koopmans, 1953; see Morgan, 1990, for a thorough treatment of the history of the identification problem).

The standard solution to the identification problem is to look for additional causal determinants that discriminate between otherwise simultaneous relationships. Both the supply of milk and demand for milk depend on the price of milk. If, however, the supply also depends on the price of alfalfa used to feed the cows and the demand also on the daily high temperature (which affects the demand for milk to make ice cream), then supply and demand curves can be identified separately. Identification can be viewed through the glasses of simultaneous equations, pushing causality into the background, or it can be viewed as a problem in causal articulation. In the first case, economists frequently use the language of exogenous variables (the price of alfalfa, the temperature) and endogenous variables (the price and quantity of milk). Exogenous variables can also be regarded as the causes of the endogenous variables. From the 1920s to the 1950s, different economists placed different emphasis on the causal aspects of identification (Morgan, 1990) and the various papers reprinted in Hendry and Morgan (1995).

Modern econometrics can be dated from the development of structural econometric models following the pioneering work in the 1930s of Jan Tinbergen, the conceptual foundations of probabilistic econometrics in Trgyve Haavelmo's (1944) ‘Probability approach to econometrics’, and the technical elaboration of the identification problem in the two Cowles Commission volumes. Structural models did not in themselves necessarily favor the language of identification over the language of causality. Indeed, in Tinbergen's (1951) textbook, dynamic, structural models are explicated with a diagram that uses arrows to indicate causal connections among time-dated variables. Nevertheless, after the econometric work of the Cowles Commission, two approaches can be clearly distinguished.

One approach, associated with Hermann Wold and known as *process analysis*, emphasized the asymmetry of causality, typically grounded it in Hume's criterion of temporal precedence (Morgan, 1991). Wold's process analysis belongs to the time-series tradition that ultimately produced Granger causality and the vector autoregression (see Section 3).

The other approach, associated with the Cowles Commission, related causality to the invariance properties of the structural econometric model. This approach emphasized the distinction between endogenous and exogenous variables and the identification and estimation of structural parameters. Implicitly, structural modellers accepted Mill's a priori approach to economics. While they differed from Mill in their willingness to conduct empirical investigations, the selection of exogenous (or *instrumental*) variables was seen to be the province of a priori economic theory – a maintained assumption rather than something to be learned from data itself.

In his contribution to one of the Cowles Commission volumes, Herbert Simon (1953) showed that causality could be defined in a structural econometric model, not only between exogenous and endogenous variables, but also among the endogenous variables themselves. And he showed that the conditions for a well-defined causal order are equivalent to the well-known conditions for identification. Despite the equivalence, with the demise of process analysis and the ascendancy of structural econometrics – aided indirectly perhaps by a revival of Humean causal scepticism among the logical-positivist philosophers of science – causal language in economics virtually collapsed between 1950 and about 1990 (Hoover, 2004).

3 Alternative approaches to causality in economics

Different approaches to causality can be classified along two lines as shown in Figure 1. One the one hand, approaches may emphasize structure or process. On the other hand, approaches may rely on a priori identifying assumptions or they may seek to infer causes from data. The upper left cell, the a priori structural approach, represented by the Cowles Commission, dominated economics for most of the postwar period. But since we already discussed it at some length in Section 2, and since it was largely responsible for turning the economics profession away from *explicit* causal analysis, we add nothing more about it here and instead turn to the other cells in Figure 1.

Figure 1
Classification of approaches to causality in economics

	Structural	*Process*
A Priori	Cowles Commission: Koopmans (1953); Hood and Koopmans (1953)	Zellner (1979)
Inferential	Simon (1953) Hoover (1990; 2001) Favero and Hendry (1992) Natural experiments: Angrist and Krueger (1999; 2001)	Granger (1969) Vector autoregressions: Sims (1980)

3.1 The inferential structural approach

The most important of the inferential structural approaches is due to Simon (1953). Simon eschews temporal order as a basis for causal asymmetry and, instead, looks to recursive structure. As we observed in Section 2, Simon's account is closely related to the Cowles Commission's structural approach. Consider the bivariate system:

$$Y_t = \theta X_t + \varepsilon_{1t}, \quad (1)$$

$$X_t = \varepsilon_{2t},$$

(2)

where the random error terms ε_{it} are independent, identically distributed and θ is a parameter. Simon says that X_t causes Y_t, because X_t is recursively ordered ahead of Y_t. One knows all about X_t without knowing about Y_t, but one must know the value of X_t to determine the value of Y_t. Equations (1) and (2) also appear to show that any intervention in (2), say a change in the variance of ε_{2t}, would transmit to (1); while any intervention in (1), say a change in θ or the variance of ε_{1t}, would not transmit to (2). Apparently, X_t could then be used to control Y_t.

Unfortunately, merely being able to write an accurate description of the two variables in the form of (1) and (2) does not guarantee either the apparent asymmetry of information or control. The same data can be repackaged into a statistically identical form with an apparently different causal order. For example, consider the following related system:

$$Y_t = \omega_{1t}, \tag{3}$$

$$X_t = \delta Y_t + \omega_{2t}, \tag{4}$$

$$\text{where } \delta = \frac{\theta \operatorname{var}(\varepsilon_2)}{\theta^2 \operatorname{var}(\varepsilon_2) + \operatorname{var}(\varepsilon_1)}, \ \omega_{1t} = \varepsilon_{1t} + \theta\varepsilon_{2t}, \text{and} \ \omega_{2t} = (1 - \delta\theta)\varepsilon_{2t} - \delta\varepsilon_{1t}.$$

Equations (3) and (4) are derived from eqs. (1) and (2). The details of the algebra are not important. Essentially, (3) and (4) are linear combinations of (1) and (2) with multiplicative factors carefully chosen, so that the error terms ω_{1t} and ω_{2t} are uncorrelated. Such linear combinations preserve the values of X_t and Y_t and their statistical likelihood (that is, the two systems of equations have the same reduced form) and, so, describe the data equally well. Equations (3) and (4) have a form analogous to (1) and (2); but, on Simon's criterion, it appears that Y_t causes X_t on Simon's criterion. While it looks like the key parameters for (3) and (4) are derived from those of (1) and (2), we could have taken (3) and (4) as the starting point and derived (1) and (2) symmetrically. What we would like to do is to replace the equal signs with arrows that show that the causal direction runs from the right-hand to the left-hand sides in the regression equations in one of the systems, but not in the other. Unfortunately, there is no way to do this, no choosing between the systems, on the basis of a single set of data by itself. This is the problem of observational equivalence again.

The a priori approach of the Cowles Commission relies on economic theory to provide appropriate identifying assumptions to resolve the observational equivalence. Christopher Sims (1980) attacked the typical application of the Cowles Commission's approach to structural macroeconometric models as relying on ‘incredible’ identifying assumptions: economic theory was simply not informative enough to do the job. But Simon, who was otherwise supportive of the conception of causality in the Cowles Commission, took a different tack. Simon sees the problem as choosing between two alternative sets of parameters: which set contains the structural parameters, $\{\theta$ and the variances of the $\varepsilon_{it}\}$ or $\{\delta$ and the variances of the $\omega_{it}\}$? Simon suggested that experiments – either controlled or natural – could help to decide. If, for example, an experiment could alter the conditional distribution of X_t without altering the marginal distribution of Y_t, then it must be that Y_t causes X_t, because this would be possible only if a structure like (3) and (4) characterized the data. If it did, a

change in the conditional distribution would involve either δ or the variance of ω_{2t}, neither of which would affect the variance of ω_{1t}. In contrast, if (1) and (2) truly characterized the causal structure of the data, a change to the conditional distribution of X_t would, in fact, involve a change to the variance of ε_{2t}, which, according to the equivalences above, would alter either δ or the variance of ω_{2t}. Similar relationships of stability and instability in the face of changes to the marginal distribution can also be demonstrated (Hoover, 2001, ch. 7). The appeal to experimental evidence is what marks Simon's approach out as inferential rather than a priori.

Hoover (1990; 2001) generalizes Simon's approach to the type of nonlinear systems of equations found in modern rational-expectations models. He shows that Simon's idea of natural experiments can be operationalized by coordinating historical, institutional, or other non-statistical information with information from structural break tests on what, in effect, amounts to the four regressions corresponding to (1)–(4) above generalized to include lagged dynamics. With allowances for complications introduced by rational expectations, the key idea is that, in the true causal order, interventions that alter the parameters governing the true marginal distribution do not transmit forward to the conditional distribution (characterized by (1) or (4)) nor do interventions in the true conditional distribution transmit backward to the marginal distribution (characterized by (2) or (3)). Since the true structural parameters are not known a priori, non-statistical information is important in identifying an intervention as belonging to the process governing one variable or another.

Although avoiding the term ‘causality’, Favero and Hendry's (1992) analysis of the Lucas critique in terms of ‘super-exogeneity’ is also a variant on Simon's causal analysis (Ericsson and Irons, 1995; Hoover, 2001, ch. 7). Super-exogeneity is essentially an invariance concept (Engle, Hendry and Richard, 1983). Favero and Hendry find evidence against the Lucas critique (non-invariance in the face of changes in policy regime) in the super-exogeneity of conditional probability distributions in the face of structural breaks in marginal distributions – the same sort of evidence that Hoover cites as helping to identify causal direction.

The recent revival of causal analysis in microeconomics in the guise of ‘natural experiments’, although apparently developed independently of Simon, nonetheless proceeds in much the same spirit as Hoover's version of Simon's approach (Angrist and Krueger, 1999; 2001). This literature typically employs the language of instrumental variables. A natural experiment is a change in a policy or a relevant environmental factor that can be identified non-statistically. Packaged as an econometric instrument, the experiment can be used – in much the same way that variations in alfalfa prices and temperature were used in the example in Section 2 – to identify the underlying relationships and to measure the causally relevant parameters.

While the development of structural approaches in econometrics has largely been independent, there is some cross-fertilization between economists and philosophers (for example, Simon and Rescher, 1966); and recently philosophers of causality have looked to economics for inspiration and examples (for example, Cartwright, 1989; Woodward, 2003).

3.2 The inferential process approach

Perhaps the most influential explicit approach to causality in economics is due to Clive W. J. Granger (1969). Granger causality is an inferential approach, in that it is data-based without direct reference to background economic theory; and it is a process approach, in that it was developed to apply to dynamic time-series models (see Granger–Sims causality in this dictionary for technical details). Granger–Sims causality is an example of the modern probabilistic approach to causality, which is a natural successor to Hume (for example, Suppes, 1970). Where Hume required constant conjunction of cause and effect, probabilistic approaches are content to identify cause with a factor that raises the probability of the effect: *A* causes *B* if $P(B|A) > P(B)$, where the vertical ‘|’ indicates ‘conditional on’. The asymmetry of causality is secured by requiring the cause (*A*) to occur before the effect (*B*) (but the probability criterion is not enough on its own to produce asymmetry since $P(B|A) > P(B)$ implies $P(A|B) > P(A)$).

Granger's (1980) definition is more explicit about temporal dynamics than is the generic probabilistic account, and it is cast in terms of the incremental predictability of one variable conditional on another:

> X_t Granger-causes Y_{t+1} if P(Y_{t+1}| all information dated *t* and earlier)
> $\neq$ P(Y_{t+1}| all information dated *t* and earlier omitting information about *X*).

This definition is conceptual, as it is impracticable to condition on *all* past information.

In practice, Granger causality tests are typically implemented through bivariate regressions. As an illustration, consider the regression equations:

$$Y_t = \Pi_{11} Y_{t-1} + \Pi_{12} X_{t-1} + v_{1t}, \quad (5)$$

$$X_t = \Pi_{21} Y_{t-1} + \Pi_{22} X_{t-1} + \upsilon_{2t}, \quad (6)$$

where the Π_{ij} are parameters, and the υ_{it} are random error terms. In practice, lag lengths may be larger than one, but far less than the infinity implicit in the general definition. X_t Granger-causes Y_{t+1} if $\Pi_{12} \neq 0$, and Y_t Granger-causes X_{t+1} if $\Pi_{21} \neq 0$.

Sims (1972) famously used Granger causality to demonstrate the causal priority of money over nominal income. Later, as part of a generalized critique of structural econometric models, Sims (1980) advocated vector autoregressions (VARs) – atheoretical time-series regressions analogous to eqs (1) and (2), but generally including more variables with lagged values of each appearing in each equation. In the VAR context, Granger causality generalizes to the multivariate case.

While Granger causality has something useful to say about incremental predictability, there is no close mapping between Granger causality and structural notions of causality on either the Cowles Commission's or Simon's accounts (Jacobs, Leamer and Ward, 1979). Consider a structural model:

$$Y_t = \theta X_t + \beta_{11} Y_{t-1} + \beta_{12} X_{t-1} + \varepsilon_{1t}, \quad (7)$$

$$X_t = \gamma Y_t + \beta_{21} Y_{t-1} + \beta_{22} X_{t-1} + \varepsilon_{2t}, \quad (8)$$

where ε_{1t} and ε_{2t} are identically distributed, independent random errors and θ, γ, and the β_{ij}s are structural parameters. The independence of the parameters and the error terms implies that causality runs from the right-hand to the left-hand sides of each equation. Equations (5) and (6) can be seen as the reduced forms of (7) and (8).

We focus on *X* causing *Y*. *X* structurally causes *Y* if either θ or $\beta_{12} \neq 0$. And *X* Granger causes *Y* if $\Pi_{12} = \frac{\beta_{12} + \theta\beta_{22}}{1 - \theta\gamma} \neq 0$. Thus, if *X* Granger causes *Y*, then *X* structurally causes *Y*. Note, however, that this result is particular to the case in which (7) and (8) represents the universe, so that (5) and (6) represent the complete conditioning on past histories of relevant variables. If the universe is more complex and the estimated VAR does not capture the true reduced forms of the structural system, which in practice they may not, then the strong connection suggested here does not follow.

More interestingly, even if (5)–(8) are complete, structural causality does not necessarily imply Granger causality. Suppose that $\beta_{12} = \beta_{22} = 0$, but $\theta \neq 0$, then *X* structurally causes *Y*, but since $\Pi_{12} = 0$, *X* does not Granger cause *Y*.

Now suppose that *X* does not Granger cause *Y*. It does not necessarily follow that *X* does not structurally cause *Y*, since if θ, β_{12}, and $\beta_{22} \neq 0$, and $-\beta_{12} / \beta_{22} = \theta$, then it will still be true that $\Pi_{12} = 0$. This may appear to be an odd special case, but in fact conditions such as $-\beta_{12} / \beta_{22} = \theta$ arise commonly in optimal control problems in economics.

A simple physical example makes it clear what is happening. Suppose that *X* measures the direction of the rudder on a ship and *Y* the direction of the ship. The ship is pummeled by heavy seas. If the helmsman is able to steer on a straight course, effectively moving the rudder to exactly cancel the shocks from the waves, the direction of the rudder (in ignorance of the true values of the shocks) will not predict the course of the ship. The rudder would be structurally effective in causing the ship to turn, but it would not Granger-cause the ship's course.

3.3 The a priori process approach

The upper right-hand cell of Figure 1 is represented by Arnold Zellner's (1979) account of causality (cf. Keuzenkamp, 2000, ch. 4, s. 4). Zellner's notion of causality is borrowed from the philosopher Herbert Feigl (1953, p. 408), who defines causation '… in terms of predictability according to law (or more adequately, according to a set of laws)'. On the one hand, Zellner opposes Simon and sides with Granger: predictability is a central feature of causal attribution, which is why his is a process account. On the other hand, he opposes Granger and sides with Simon: an underlying structure (a set of laws) is a crucial presupposition of causal analysis, which is why his is an a priori account.

Much obviously depends on what a law is. Zellner's own view is that a law is a (probabilistic) description of a succession of states of the world that holds for many possible boundary conditions and covers many possible circumstances. He couches his position in an explicitly Bayesian theory of inference. Feigl identifies causality with lawlikeness or predictability. It is the fact that formulae fit previously unexamined cases, as well as examined ones, which constitutes their lawlikeness. This is close to Simon's invariance criterion (the true causal order is the one that is invariant under the right sort of intervention).

The central problem, then, is how to distinguish laws from false generalizations or accidental regularities – that is, how to distinguish conditional relations invariant to interventions from regularities that are either not invariant or are altogether adventitious. Zellner believes that a theory serves as the basis for discriminating between laws and casual generalizations. Although Zellner's approach permits us to learn some things from the data, in keeping with the spirit of Bayesian inference, it does so within a narrowly defined framework (cf. Savage's, 1954, pp. 82–91, 'small world' assumption). Economic theory in Zellner's account restricts the scope of an investigation a priori.

Zellner objects to Granger causality for two reasons. First, it is not satisfactory to identify cause with temporal ordering, as temporal ordering is not the ordinary, scientific or philosophical foundation of the causal relationship. Second, Granger's approach is atheoretical. In order to implement it practically, an investigator must impose restrictions – limit the information set to a manageable number of variables, consider only a few moments of the probability distribution (in our exposition, just the mean), and so forth. For Zellner, if these restrictions cannot be explained theoretically, Granger's methods will discover only accidental regularities.

Zellner explicitly criticizes Granger for ignoring the need for theoretical basis for empirical investigation – implicitly focusing on only one side of a process in which theory informs empirics and empirics inform theory. He criticizes Simon for defining cause to be a formal property of a model (recursive order) without making essential reference to empirical reality. Zellner's criticism is, however, more aptly directed at the Cowles Commission's approach, since (as we saw in Section 3.1) Simon distinguishes himself through tying causal order to empirical inference.

3.4 Structural vector autoregressions

Not all approaches to causality fall quite neatly into the cells of Figure 1; or, more to the point, an approach that falls into one cell may morph into one that falls into another cell. The history of Sims's VAR program is an important case.

Sims (1980) advocated VARs as a reaction to the manner in which the Cowles Commission programme, which identified structural models through a priori theory, had been implemented (see Section 3.2). From a causal perspective, it was closely related to Granger's analysis. Starting with VAR such as eqs (5) and (6), Sims wished to work out how various 'shocks' would affect the variables of the system. This is complicated by the fact that the error terms in (5) and (6), which might be taken to represent the shocks, are not in general independent, so that a shock to one is a shock to both, depending on how correlated they are. Sims's initial solution was to impose an arbitrary orthogonalization of the shocks (a Choleski decomposition). In effect, this meant transforming (5) and (6) into a system like (6) and (7) and setting either θ or γ to zero. This amounts to imposing a recursive order on X_t and Y_t, such that the covariance matrix of the error terms is diagonal (that is, ε_{1t} and ε_{2t} are uncorrelated). A shock to X can then be represented by a realization of ε_{1t} and a shock to Y by a realization of ε_{2t}.

Initially, Sims treated the choice of recursive order as a matter of indifference. Criticizing the VAR program from the point of view of structural models, Leamer (1985) and Cooley and LeRoy (1985) pointed out that the substantive results (for instance, impulse-response functions and innovation accounts) depend on which recursive order is chosen. Sims (1982; 1986) accepted the point and henceforth advocated Structural vector autoregressions (SVARs). SVARs can be identified through the contemporaneous causal order only. So, for example, to identify (5) and (6), it is enough to assume that either θ or γ in (7) or (8) is zero; one need not make any assumptions about the β_{ij}s. Ironically, since the initial impulse behind the VAR programme was to avoid theoretically tenuous identifying assumptions, the choice of restrictions on contemporaneous variables used to transform the VAR into the SVAR are typically only weakly supported by economic theory.

Nevertheless, the move from the VAR to the SVAR is a move from an inferential to an a priori approach. It is also a move from a fully non-structural, process approach to a partially structural approach, since the structure of the contemporaneous variables, though not of the lagged variables, is fully specified. The SVAR approach can, therefore, be seen as straddling the cells on the first line of Figure 1.

3.5 The graph-theoretic approach to causal inference

A final approach to causality in economics sometimes provides another example of an inferential structural approach, and sometimes straddles the cells on the second line of Table 1. Graph-theoretic approaches to causality were first developed outside of economics by computer scientists (for example, Pearl, 2000) and philosophers (for example, Spirtes, Glymour and Scheines, 2000), but have recently been applied within economics (Swanson and Granger, 1997; Akleman, Bessler and Burton, 1999; Bessler and Lee, 2002; Demiralp and Hoover, 2003).

The key ideas of the graph-theoretic approach are simple (see Demiralp and Hoover, 2003 or Hoover, 2005 for a detailed discussion). Any structural model can be represented by a graph in which arrows indicate the causal order. Equations (1) and (2) are represented by $X \rightarrow Y$ and eqs (3) and (4) by $Y \rightarrow X$. More complicated structures can be represented by more complicated graphs. Simultaneity, for instance, can be represented by double-headed arrows. The graphs allow us easily to see the dependence or independence among variables. Pearl (2000) and Sprites, Glymour and Scheines (2000) demonstrate the isomorphism between causal graphs and the independence relationships encoded in probability distributions. This isomorphism allows conclusions about probability distributions to be derived from theorems proven using the mathematical techniques of graph theory.

Many of the results of graph-theoretic analysis are straightforward. Suppose that $A \rightarrow B \rightarrow C$ (that is, *A* causes *B* causes *C*). *A* and *C* would be probabilistically dependent; but, conditional on *B*, they would be independent. Similarly for $A \leftarrow B \leftarrow C$. In each case, *B* is said to *screen A* from *C*. Suppose that $A \leftarrow B \rightarrow C$. Then, once again *A* and *C* would be dependent, but conditional on *B*, they would be independent. *B* is said to be the *common cause* of *A* and *C*. Now suppose that *A* and *B* are independent conditional on sets of variables that exclude *C* or its descendants, and $A \rightarrow C \leftarrow B$, and none of the variables that cause *A* or *B* directly causes *C*. Then, conditional on *C*, *A* and *B* are dependent. *C* is called an *unshielded collider* on the path *ACB*. (A *shielded* collider would have a direct link between *A* and *B*.) These are the simplest relationships of probabilistic dependence and independence. More complex ones may also obtain in which *A* is independent of *B* only conditional on more than one other variable (say, *C* and *D*).

A number of causal search algorithms have been developed (Sprites, Glymour and Scheines, 2000). These start with information about correlations (or other tests of unconditional and conditional statistical independence) among variables. The most common of these, the PC algorithm, assumes that graphs are strictly recursive (known in the literature as *acyclical*) and starts with a graph in which all variables are causally connected with an unknown causal direction (represented by the headless arrow, '—'). It then tests for independence among pairs of variables, conditioning on sets of zero variables, then one, then two, and so forth until the set of variables is exhausted. Whenever it finds independence, it removes the causal connection between the variables in the graph. Once the graph is pared down as far as it can be, it considers triples of variables in which two are conditionally independent but are connected through a third. If conditioning on that third variable renders the variables conditionally dependent, then that variable is an unshielded collider and it is connected to the other two variables with causal arrows running toward it. After all the unshielded colliders have been identified, further logical analysis can be used to orient additional causal arrows. For example, we might reason as follows: suppose we have a triple $A \rightarrow C - B$; unless the causal arrow runs away from *C* toward *B*, *C* would be identified as an unshielded collider; but *C* was not identified as an unshielded collider earlier in the search; therefore, the causal arrow must run away from *C* towards *B*, so that the graph becomes $A \rightarrow C \rightarrow B$.

Sometimes the data allow the complete orientation of a causal graph, but sometimes some causal connections are left undirected. In this case, the graph marks out an equivalence class, and the algorithm has identified 2! causal graphs consistent with the empirical probability distribution, where n=the number of undirected causal connections.

While most applications of graph-theoretic methods assume that the true causal structures are recursive (that is, strictly acylical), economics frequently treats variables that are cyclical or simultaneously determined. Although the recursiveness assumption is restrictive, it is an assumption that is also frequently made in the SVAR literature. Some progress has been made in developing graph-theoretic search algorithms for cyclical or simultaneous causal systems (Pearl, 2000, pp. 95–6, 142–3; Richardson, 1996; Richardson and Spirtes, 1999).

Swanson and Granger (1997) showed that estimates of the error terms of the VAR (the υ_{it} in eqs (5) and (6)) can be treated as the original time-series variables purged of their dynamics. A causal order identified on such variables corresponds to the causal order necessary to convert a VAR into an SVAR. Demiralp and Hoover (2003) present Monte Carlo evidence that the PC algorithm is effective at selecting the true causal connections among variables and, when signal strengths are high enough, moderately effective at directing them correctly. Search algorithms can, therefore, reduce or even eliminate the need to appeal to a priori theory when identifying the causal order of an SVAR.

Where Simon's approach looked for relatively important interventions as a basis for causal inference to a structure, the graph-theoretic approach uses relatively routine random variations to identify patterns of conditional independence that map out causal structures. The two approaches are complementary: Simon's approach may be used to resolve the observational equivalence reflected in causal connections that remain undirected after the application of a causal search algorithm.

4 From metaphysics to econometric practice

The analysis of causation was originally a branch of metaphysics. In moving from the scholastic to the practical, two deep divisions appeared among economists.

The first is the divide between those who believed that causality in economics could be characterized by relatively simple uniformities (the process approaches) and those who believed that it must be characterized by a rich understanding of the underlying mechanisms (the structural approaches). Economists debate the appropriate level at which to

characterize either the uniformities or the mechanisms – individual or aggregate. But this debate over the microfoundations of macroeconomics is another story. The second divide is between those who believe that economic logic itself gives privileged insight into economic behaviour (a priori approaches) and those who believe that we must learn about economic behaviour principally through observation and induction (the inferential approaches).

These are old debates – unlikely to be resolved decisively to the satisfaction of all economists in the near future. How one aligns oneself in them largely determines which particular approaches to causality appear to be compelling in practical economic research.

See Also

- endogeneity and exogeneity
- Granger–Sims causality
- graph theory
- Hume, David
- identification
- Mill, John Stuart
- Simon, Herbert A.
- structural vector autoregressions
- vector autoregressions

Bibliography

Akleman, D.G., Bessler, D.A. and Burton, D.M. 1999. Modeling corn exports and exchange rates with directed graphs and statistical loss functions. In *Computation, Causation, and Discovery*, ed. C. Glymour and G.F. Cooper. Menlo Park, CA, and Cambridge, MA: MIT Press and American Association for Artificial Intelligence.

Angrist, J.D. and Krueger, A.B. 1999. Empirical strategies in labor economics. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Angrist, J.D. and Krueger, A.B. 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives* 15 (4), 69–85.

Bessler, D.A. and Lee, S. 2002. Money and prices: U.S. data 1869–1914 (a study with directed graphs). *Empirical Economics* 27, 427–46.

Bunge, M. 1963. *The Place of the Causal Principle in Modern Science*. Cleveland: Meridian Books.

Cartwright, N. 1989. *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Cohen, G.A. 1978. *Karl Marx's Theory of History: A Defense*. Princeton: Princeton University Press.

Cooley, T.F. and LeRoy, S.F. 1985. Atheoretical macroeconomics: a critique. *Journal of Monetary Economics* 16, 283–308.

Demiralp, S. and Hoover, K.D. 2003. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics* 65(Supplement), 745–67.

Edgeworth, F.Y. 1885. Methods of statistics. In Royal Statistical Society of Britain, *Jubilee Volume of the Statistical Society*. London: E. Stanford.

Engle, R.F., Hendry, D.F. and Richard, J.-F. 1983. Exogeneity. *Econometrica* 51, 277–304.

Ericsson, N. and Irons, J. 1995. The Lucas critique in practice: theory without measurement. In *Macroeconometrics: Developments, Tensions and Prospects*, ed. K.D. Hoover. Boston: Kluwer.

Favero, C. and Hendry, D.F. 1992. Testing the Lucas critique: a review. *Econometric Reviews* 11, 265–306.

Feigl, H. 1953. Notes on causality. In *Readings in the Philosophy of Science*, ed. H. Feigl and M. Brodbeck. New York: Appleton-Century-Crofts.

Granger, C.W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.

Granger, C.W.J. 1980. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control* 2, 329–52.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12 (Supplement), iii–vi, 1–115.

Hausman, D.M. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Hendry, D.F. and Morgan, M.S., eds. 1995. *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press.

Hood, W. and Koopmans, T., eds. 1953. *Studies in Econometric Method*. Cowles Commission Monograph No. 14. New Haven: Yale University Press.

Hoover, K.D. 1990. The logic of causal inference: econometrics and the conditional analysis of causality. *Economics and Philosophy* 6, 207–34.

Hoover, K.D. 2001. *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K.D. 2004. Lost causes. *Journal of the History of Economic Thought* 26, 149–64.

Hoover, K.D. 2005. Automatic inference of the contemporaneous causal order of a system of equations. *Econometric Theory* 21, 69–77.

Hoover, K.D. and Dowell, M.E. 2001. Measuring causes: episodes in the quantitative assessment of the value of money. *History of Political Economy* 33, 137–61.

Hume, D. 1739. *A Treatise of Human Nature*. Ed. L.A. Selby-Bigge, Oxford: Clarendon Press, 1888.

Hume, D. 1742. Of interest. In *Essays: Moral, Political, and Literary*, ed. E.F. Miller, Indianapolis: Liberty Classics, 1985.

Hume, D. 1777. An Enquiry Concerning Human Understanding. In *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 2nd edn. ed. L.A. Selby-Bigge. Oxford: Clarendon Press, 1902.

Jacobs, R.L., Leamer, E.E. and Ward, M.P. 1979. Difficulties in testing for causation. *Economic Inquiry* 17, 401–13.

Jevons, W.S. 1863. A serious fall in the value of gold ascertained, and its social effects set forth. In *Investigations in Currency and Finance*, 1884, Reprinted New York: Augustus M. Kelley, 1964.

Keuzenkamp, H.A. 2000. *Probability, Econometrics, and Truth*. Cambridge: Cambridge University Press.

Kincaid, H. 1996. *Philosophical Foundations of the Social Sciences: Analyzing Controversies in Social Research*. Cambridge: Cambridge University Press.

Koopmans, T., ed. 1950. *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph No. 10. New York: Wiley.

Leamer, E.E. 1985. Vector autoregressions for causal inference. In *Understanding Monetary Regimes*, ed. K. Brunner and A.H. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 22, Spring. Amsterdam: North-Holland.

Mäki, U. 2001. *The Economic World View: Studies in the Ontology of Economics*. Cambridge: Cambridge University Press.

Marshall, A. 1930. *Principles of Economics: An Introductory Volume*, 8th edn. London: Macmillan.

Mill, J.S. 1848. *Principles of Political Economy with Some of Their Applications to Social Philosophy*, ed. W.J. Ashley. London: Longman, Green, 1909.

Mill, J.S. 1851. *A System of Logic, Ratiocinative and Deductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, vol. 1, 3rd edn. London: John W. Parker.

Mises, L. von. 1966. *Human Action: A Treatise on Economics*, 3rd edn. Chicago: Henry Regnery.

Morgan, M.S. 1991. The stamping out of process analysis in econometrics. In *Appraising Economic Theories: Studies in the Methodology of Research Programs*, ed. N. De Marchi and M. Blaug. Aldershot: Edward Elgar.

Morgan, M.S. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Richardson, T. 1996. A discovery algorithm for directed cyclical graphs. In *Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Congress*, ed. F. Jensen and E. Horwitz. San Francisco: Mogan Kaufman.

Richardson, T. and Spirtes, P. 1999. Automated discovery of linear feedback models. In *Computation, Causation and Discovery*, ed. C. Glymour and G.F. Cooper. Menlo Park, CA: AAAI Press.

Robbins, L. 1935. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Dover, 1972.

Simon, H.A. 1953. Causal order and identifiability. In Hood and Koopmans (1953).

Simon, H.A. and Rescher, N. 1966. Causes and counterfactuals. *Philosophy of Science* 33, 323–40.

Sims, C.A. 1972. Money, income and causality. *American Economic Review* 62, 540–52.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Sims, C.A. 1982. Policy analysis with econometric models. *Brookings Papers on Economic Activity* 1982(1), 107–52.

Sims, C.A. 1986. Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review* 10(1), 2–15.

Spirtes, P., Glymour, C. and Scheines, R. 2000. *Causation, Prediction, and Search*, 2nd edn. Cambridge, MA: MIT Press.

Stigler, S.M. 1986. *The History of Statistics: Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap Press.

Stigler, S.M. 1999. *Statistics on the Table*. Cambridge, MA: Harvard University Press.

Suppes, P. 1970. A probabilistic theory of causality. *Acta Philosophica Fennica* 24. Amsterdam: North Holland.

Swanson, N.R. and Granger, C.W.J. 1997. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association* 92, 357–67.

Tinbergen, J. 1951. *Econometrics*. New York: Blakiston Company.

Walras, L. 1954. *Elements of Pure Economics*. London: Allen and Unwin.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Zellner, A.A. 1979. Causality and econometrics. In *Three Aspects of Policy Making: Knowledge, Data and Institutions*, ed. K. Brunner and A.H. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 10. Amsterdam: North-Holland.

How to cite this article

Hoover, Kevin D. "causality in economics and econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000569> doi:10.1057/9780230226203.0209

The New Palgrave Dictionary of Eco

central bank communication

Michael Ehrmann and Marcel Fratzscher
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since the early 1990s, communication has become a primary tool for monetary authorities in managing expectations, both of financial markets and of the wider public, and an important ingredient in making the central bank accountable. The rapidly growing literature on central bank communication clearly confirms the importance of communication in managing expectations, thereby enhancing the effectiveness of monetary policy. Yet there is a large degree of heterogeneity in communication practices across monetary authorities in the world, and there continues to be a lively and controversial debate about what constitutes an optimal communication strategy.

Keywords

accountability; central bank; communication; expectations; financial markets; independence; inflation; monetary policy; objectives; policy decisions; predictability; transparency

Article

Central bank communication refers to the process by which monetary authorities convey information regarding their objectives, strategies and tools, as well as about their current assessment of the economic situation and the monetary stance. Such communication typically serves two purposes: making the central bank accountable, and enhancing the effectiveness of central bank policies.
Whereas only a few decades ago, transparency was usually seen as

counterproductive to an effective conduct of monetary policy, it is nowadays considered best practice in central banking. One trigger for this has been the move to grant independence to more and more central banks, which in turn entails an obligation of those central banks to be more accountable.

Central banking laws often specify a number of obligations to ensure a central bank's accountability, which at the same time shape their communication policies. Testimonies to parliament and the requirement to deliver annual reports are examples. In several cases the relevant central banking acts also prescribe the targets for the monetary authority; their communication is therefore automatic, and not at the discretion of the central bank. At the same time, most central banks communicate substantially more often, and in much greater detail, than required by law.

As to the second purpose, it became increasingly clear throughout the 1990s that managing expectations is a central part of monetary policy, and that transparency is vital for that purpose. Given that communication is essential for accountability and transparency, central banks are now putting considerable effort into designing and conducting their communication policies.

Blinder et al. (2008), in their survey of the literature on central bank communication, derive the conditions under which central bank communication may matter for the conduct of monetary policy. A crucial issue in this regard is that a central bank usually has direct control only over a short-term interest rate, yet needs to influence interest rates at all maturities and to affect market expectations not just about current levels but about the future path of interest rates.

If the economic environment were constant, if the central bank was credibly committed to an unchanging policy rule, and if private agents had full information and rational expectations, the path of monetary policy could be inferred correctly from the central bank's observed actions (Woodford, 2005). In reality, of course, none of these conditions are likely to hold. In particular, in a changing environment economic agents are subject to a continuous learning process. The central banks' views are also of interest to the public in a world of uncertainty and asymmetric information, especially given the complexity and extent of the information that feeds into monetary policy decisions, which often require judgment and the use of heuristics (Svensson, 2003; King, 2005).

The revolution in thinking and practice that has taken place over the recent decades can be exemplified with the case of the US Federal Reserve, which

over the last 15 years has gone a long way towards greater transparency. For instance, it has started to issue statements instantaneously after each monetary policy decision, where the decision is not only announced, but also briefly explained, and to provide qualitative forward guidance about future monetary policy decisions. It has also expedited the release of the minutes of Federal Open Market Committee (FOMC) meetings, and it has increased the frequency and expanded the content of the publicly released forecasts for several economic variables made by FOMC members.

The announcement of a central bank's objective

If a central bank is granted independence from its government, it must be given a clearly defined mandate. This is generally done by defining central bank objectives, often in a quantified fashion. But even if a central bank is not given a quantitative objective, it often decides to provide its own quantification, or is required to do so. The potential effects of such a clarification and quantification are substantial. Not only do they make an independent central bank more (easily) accountable, since its actions can be assessed by cross-checking actual economic outcomes with those mandated; furthermore, the announcement of an objective, and in particular its quantification, provides a yardstick for the expectations of economic agents. The available empirical evidence demonstrates that increased transparency about central banks' strategies, and in particular the announcement of an explicit inflation objective, has fostered central bank credibility as well as the predictability of the path of monetary policy. Moreover, the recent trend towards more transparent central banking practices has certainly played a considerable part in improving the short-term predictability of policy decisions by many central banks over recent decades (BIS, 2004, pp. 73–80). The announcement of central bank objectives may also have a direct bearing on economic outcomes. For instance, Benati (2008) finds that inflation persistence is considerably lower in countries with explicit inflation targets. Levin, Natalucci and Piger (2004) furthermore show that in inflation-targeting countries, private sector inflation expectations are not correlated with lagged inflation, indicating that inflation expectations are better anchored. This evidence has been corroborated by Gürkaynak, Levin and Swanson (2009), who show that in some advanced inflation-targeting countries, long-term inflation expectations are less responsive to macroeconomic data releases than in the United States, where no explicit inflation objective has been announced.

However, the fact that the announcement of a central bank's objective has an effect on inflation expectations need not automatically imply that there will be an effect on the ultimate objective. This question needs to be settled empirically. The available evidence is rather inconclusive at this stage, with for example Kuttner and Posen (1999) arguing that inflation is lower in inflation-targeting countries, whereas Ball and Sheridan (2005) cannot find any such evidence, given that also the countries in their control groups managed to achieve low inflation rates.
In sum, the empirical evidence suggests that the announcement of a central bank's objective is beneficial, since it eases the conduct of monetary policy through its effect on agents' expectations, and because it helps to achieve sound macroeconomic outcomes. At the same time, it does not seem to be the only means to achieve such outcomes.

The announcement of policy decisions

It is common practice nowadays among central banks to inform the public about monetary policy decisions as soon as the decision has been taken. There is substantial evidence that this practice improves the markets' understanding of monetary policy considerably. For example, Lange, Sack and Whitesell (2003) observe that the announcement of FOMC policy decisions since 1994 has enabled markets to improve their forecasts of monetary policy decisions. Furthermore, Demiralp and Jorda (2002) provide evidence that, by announcing changes to the intended federal funds rate in real time, it has been possible to move the federal funds rate with a smaller volume of open market operations, which indicates that the announcement of policy decisions can make policy implementation more efficient.

The communication of the current assessment of the economic situation and the monetary policy stance

By announcing an objective, and possibly releasing information about its monetary policy strategy, about the models used and about the variables considered in the economic analysis, a central bank aims to help the public better understand its broader framework and the way in which it reacts to different circumstances and contingencies. However, even if the broader framework is generally well understood, it will be impossible to communicate *ex ante* all contingencies in such a way that the public can always deduce perfectly the central bank's assessment, just by interpreting the incoming

macroeconomic data. Regular communication of the central bank's assessment of the current economic situation and the monetary policy stance does therefore remain important. Accordingly, central banks often release statements that provide explanations for a given policy decision, publish inflation and growth forecasts, and deliver speeches in the inter-meeting period.

As argued above, central banks have recently achieved a high degree of short-term predictability. Accordingly, markets react predominantly not to the announcement of a decision, but to the communication surrounding it, such as any explanation of the underlying reasons and any forward-looking component. Gürkaynak, Sack and Swanson (2005) find that longer-term maturities in the yield curve react in particular to the forward-looking component of the communication.

In line with this, Reeves and Sawicki (2007) show that the collective forms of Bank of England communication have a rather strong market impact, such as the minutes of the committee meetings and the Inflation Report.

Communication by individual committee members, such as speeches or interviews, has nonetheless also been shown to be important. Kohn and Sack (2004) find that the testimonies by the FOMC chairman have substantial effects on financial markets, throughout the entire maturity spectrum. Financial market responsiveness to committee members' speeches have been identified, for example in Ehrmann and Fratzscher (2007) for the Federal Reserve, the Bank of England and the European Central Bank (ECB).

Potential risks

The evidence suggests that central bank communication is an important policy tool, with substantial effects on financial markets, and the potential to enhance the effectiveness of monetary policy making. However, as any effective tool, it needs to be properly utilized; otherwise, it can lead to undesired outcomes. Communicating too frequently, or providing too much information, can be damaging if there is a limit to how much information can be digested effectively (Kahneman, 2003). A widespread example where central banks limit their transparency relates to the blackout periods, whereby committee members would typically not make public statements about monetary policy-related issues just before policy meetings. As shown by Ehrmann and Fratzscher (2009), there are good reasons for adhering to such a rule, because communication during the blackout period leads to excessive market

volatility.
A debate has also centred around how central banks should communicate if they receive noisy signals themselves. In coordination games in the vein of Morris and Shin (2002), or in learning models like Dale, Orphanides and Osterholm (2008), whether or not a central bank should communicate depends crucially on the relative precision of the central bank's and the private sector's information. However, it has been debated to what extent a case for limiting central bank communication can arise in these models. With regard to the coordination game literature, Svensson (2006) argues that central bank communication would need a much (and implausibly) lower signal-to-noise ratio than that of private information.
Clarity is essential for good communication. A possible risk to clarity can arise because monetary policy is typically set by committees rather than by single individuals. This can give rise to a ‘cacophony problem’ (Blinder, 2004, ch. 2) if too many disparate voices on a topic confuse rather than clarify the message. Central banks take different approaches in that regard. Whereas FOMC members communicate their individual views to the public (Bernanke, 2004), this is not so for the ECB, which now adheres to a one-voice policy (Jansen and De Haan, 2006). Importantly, however, markets adapt to such differences in communication style, for example by reacting more strongly to statements by the chairperson of committees with dispersed communication, and more equally to statements by all committee members if these communicate in a collegial fashion (Ehrmann and Fratzscher, 2007).

Open issues

The recent research on central bank communication, surveyed in Blinder et al. (2008), has provided a large number of relevant insights. Central bank communication is an important policy tool, with substantial effects on financial markets, and the potential to enhance the efficiency of monetary policy making. However, what constitutes an optimal communication strategy remains an unsettled issue. There is a large diversity in the communication policies of central banks, because the design of communication policies must take into account the cultural and institutional environment in which a central bank operates. Accordingly, it is evident that ‘one size does not fit all’.
Other issues and debates remain unresolved at the time of writing. For instance, there are different ways of providing forward guidance. It is difficult to evaluate the recent approach of publishing projected paths for the central

bank's policy rate, given that we have gained only very limited experience to date. Another open issue relates to the transmission of central bank communication. The role of the media has barely been studied. Finally, while much of the empirical research has focused on the effects of communication on financial markets, a better understanding of the communication with the general public is required, since it is the general public whose inflation expectations eventually feed into the actual evolution of inflation – for example, through corresponding wage claims and savings, investment and consumption decisions.

See Also

- Bank of England
- European Central Bank
- central bank independence
- federal reserve system
- inflation
- inflation targeting
- Taylor rules

The views expressed in this article do not necessarily coincide with those of the European Central Bank or the Eurosystem.

Bibliography

Ball, L. and Sheridan, N. 2005. Does inflation targeting matter? In *The Inflation-Targeting Debate*, ed. B. Bernanke and M. Woodford. Chicago: University of Chicago Press, 249–76.

Benati, L. 2008. Investigating inflation persistence across monetary regimes. *Quarterly Journal of Economics* 123, 1005–60.

Bernanke, B. 2004. *Fedspeak*. Available at http://www.federalreserve.gov/boarddocs/speeches/2004/200401032/default.ht accessed 3 January 2004.

Bank for International Settlements (BIS). 2004. *74th Annual Report*. Basel: BIS.

Blinder, A. 2004. *The Quiet Revolution: Central Banking Goes Modern*. New

Haven, Conn.: Yale University Press.

Blinder, A., Ehrmann, M., Fratzscher, M., de Haan, J. and Jansen, D.-J. 2008. Central bank communication and monetary policy: a survey of the evidence. *Journal of Economic Literature*, XLIV(4), 910–45.

Dale, S., Orphanides, A. and Osterholm, P. 2008. Imperfect central bank communication: information versus distraction. IMF Working Paper 08/60.

Demiralp, S. and Jorda, O. 2002. The announcement effect: evidence from open market desk data. *Federal Reserve Bank of New York Economic Policy Review* 8, 29–48.

Ehrmann, M. and Fratzscher, M. 2007. Communication by central bank committee members: different strategies, same effectiveness? *Journal of Money, Credit, and Banking* 39, 509–41.

Ehrmann, M. and Fratzscher, M. 2009. Purdah: on the rationale for central bank silence around policy meetings. *Journal of Money, Credit, and Banking*, 41 (2–3), 517–27.

Gürkaynak, R., Levin, A. and Swanson, E. 2009. Does inflation targeting anchor long-run inflation expectations? Evidence from long-term bond yields in the US, UK and Sweden. *Journal of the European Economic Association* (forthcoming).

Gürkaynak, R., Sack, B. and Swanson, E. 2005. Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *International Journal of Central Banking* 1, 55–93.

Jansen, D.-J. and De Haan, J. 2006. Look who's talking: ECB communication during the first years of EMU. *International Journal of Finance and Economics* 11, 219–28.

Kahneman, D. 2003. Maps of bounded rationality: psychology for behavioral economics. *American Economic Review* 93, 1449–75.

King, M. 2005. Monetary policy: practice ahead of theory. Available at http://www.bankofengland.co.uk/publications/speeches/2005/speech245.pdf, accessed17 May 2005.

Kohn, D. and Sack, B. 2004. Central bank talk: does it matter and why? In *Macroeconomics, Monetary Policy, and Financial Stability*, ed. Bank of Canada, 175–206.

Kuttner, K. and Posen, A. 1999. Does talk matter after all? Inflation targeting and central bank behavior. Federal Reserve Bank of New York Staff Report 88.

Lange, J., Sack, B. and Whitesell, W. 2003. Anticipations of monetary policy in financial markets. *Journal of Money, Credit, and Banking* 35, 889–909.

Levin, A., Natalucci, F. and Piger, J. 2004. The macroeconomic effects of inflation targeting. *Federal Reserve Bank of St. Louis Review* 86, 51–80.

Morris, S. and Shin, H.S. 2002. Social value of public information. *American Economic Review* 92, 1521–34.

Reeves, R. and Sawicki, M. 2007. Do financial markets react to Bank of England communication? *European Journal of Political Economy* 23, 207–27.

Svensson, L.E.O. 2003. What is wrong with Taylor rules? Using judgment in monetary policy through targeting rules. *Journal of Economic Literature* 41, 427–77.

Svensson, L.E.O. 2006. Social value of public information: Morris and Shin (2002) is actually pro transparency, not con. *American Economic Review* 96, 448–51.

Woodford, M. 2005. Central-bank communication and policy effectiveness. In *The Greenspan Era: Lessons for the Future*, ed. Federal Reserve Bank of Kansas City, 399–474.

How to cite this article

Ehrmann, Michael and Marcel Fratzscher. "central bank communication." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_C000611> doi:10.1057/9780230226203.1900

The New Palgrave Dictionary of Economics Online

central bank independence

Carl E. Walsh

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Many countries have implemented reforms designed to grant their monetary authorities greater independence from direct political influence. These reforms were justified by research showing central bank independence was negatively correlated with average inflation among developed economies. An important line of research developed measures of central bank independence and studied their relationship with inflation and real economic activity. Different theoretical approaches have been used to model central bank independence. Critics of the reform movements towards central bank independence have expressed concerns that independence can weaken the accountability of central banks.

Keywords

Bank of England; central bank independence; central banks; European Central Bank (ECB); Federal Reserve System; inflation; inflation targeting; monetary policy; price stability; rational expectations; Reserve Bank of New Zealand; Rogoff-conservative central bank

Article

Central bank independence refers to the freedom of monetary policymakers from direct political or governmental influence in the conduct of policy.
During the 1970s and early 1980s, major industrialized economies experienced sustained periods of high inflation. To explain these periods of inflation, one must account for why central banks allowed them to happen. One influential line of argument pointed to the inflation bias inherent in discretionary monetary policy if the central bank's objective for real output (unemployment) is above (below) the economy's natural equilibrium level or if policymakers simply prefer higher output levels (Barro and Gordon, 1983). Under rational expectations, the public anticipates that the central bank will attempt to expand the economy; as a consequence, real output is not systematically affected but average inflation is left inefficiently high.

This explanation for inflation raises the question why central banks might prefer economic expansions or have unrealistic output goals. Economists have frequently pointed to political pressures as the answer. Elected officials may be motivated by short-run electoral considerations, or may value short-run economic expansions highly while discounting the longer-run inflationary consequences of expansionary policies. If the ability of elected officials to distort monetary policy results in excessive inflation, then countries whose central banks are independent of such pressure should experience lower rates of inflation. Beginning with Bade and Parkin (1988), an important line of research focused on the relationship between the central bank and the elected government as a key determinant of inflation. This empirical research found that average inflation was negatively related to measures of central bank independence. Cukierman (1992) provides an excellent summary of the empirical work; references to the more recent literature can be found in Eijffinger and de Haan (1996) and Walsh (2003, ch. 8). The empirical findings led to a significant body of work addressing the following questions: what do we mean by central bank independence? How should central bank independence be measured? What causal interpretation should be placed on the empirical correlations between central bank independence and macroeconomic outcomes discovered in the data? What is the theoretical explanation for these correlations?

The meaning of independence

The historical, legal and de facto relationships between a country's government and its central bank are very complex, involving many difference aspects. These include, but are not limited to, the role of the government in appointing (and dismissing) members of the central bank governing board, the voting power (if any) of the government on the board, the degree to which the central bank is subject to budgetary control by the government, the extent to which the central bank must lend to the government, and whether there are clearly defined policy goals established in the central bank's charter.

Most discussions have focused on two key dimensions of independence. The first dimension encompasses those institutional characteristics that insulate the central bank from political influence in defining its policy objectives. The second dimension encompasses those aspects that allow the central bank to freely implement policy in pursuit of monetary policy goals. Grilli, Masciandaro and Tabellini (1991) called these two dimensions ‘political independence’ and ‘economic independence’. The more common terminology, however, is due to Debelle and Fischer (1994), who called these two aspects ‘goal independence’ and ‘instrument independence’. Goal independence refers to the central bank's ability to determine the goals of policy without the direct influence of the fiscal authority. In the United Kingdom, the Bank of England lacks goal independence since its inflation target is set by the government. In the United States, the Federal Reserve's goals are set in its legal charter, but these goals are described in vague terms (for example, maximum employment), leaving it to the Fed to translate these into operational goals. Thus, the Fed has a high level of goal independence. Price stability is mandated as the goal of the European Central Bank (ECB), but the ECB can choose how to interpret this goal in terms of a specific price index and definition of price stability.

Instrument independence refers only to the central bank's ability to freely adjust its policy tools in pursuit of the goals of monetary policy. The Bank of England, while lacking goal independence, has instrument independence; given the inflation goal mandated by the government, it is able to set its

instruments without influence from the government. Similarly, the inflation target range for the Reserve Bank of New Zealand is set in its Policy Targets Agreement (PTA) with the government, but, given the PTA, the Reserve Bank has the authority to sets its instruments without interference. The Federal Reserve and the ECB have complete instrument independence.

Measuring independence

The most widely employed index of central bank independence is due to Cukierman, Webb and Neyapti (1991), although alternative measures were developed by Bade and Parkin (1988) and Alesina, Masciandaro and Tabellini (1991), among others.

The Cukierman, Webb and Neyapti index is based on four legal characteristics as described in a central bank's charter. First, a bank is viewed as more independent if the chief executive is appointed by the central bank board rather than by the government, is not subject to dismissal, and has a long term of office. These aspects help insulate the central bank from political pressures. Second, independence is greater the more policy decisions are made independently of government involvement. Third, a central bank is more independent if its charter states that price stability is the sole or primary goal of monetary policy. Fourth, independence is greater if there are limitations on the government's ability to borrow from the central bank.

Cukierman, Webb and Neyapti combine these four aspects into a single measure of legal independence. Based on data from the 1980s, they found Switzerland to have the highest degree of central bank independence at the time, closely followed by Germany. At the other end of the scale, the central banks of Poland and the former Yugoslavia were found to have the least independence.

Legal measures of central bank independence may not reflect the actual relationship between the central bank and the government. In countries where the rule of law is less strongly embedded in the political culture, there can be wide gaps between the formal, legal institutional arrangements and their practical impact. This is particularly likely to be the case in many developing economies. Thus, for developing economies, it is common to supplement or even replace measures of central bank independence based on legal definitions with measures that reflect the degree to which legally established independence is honoured in practice. Based on work by Cukierman, measures of actual central bank governor turnover, or turnover relative to the formally specified term length, are often used to measure independence. High actual turnover is interpreted as indicating political interference in the conduct of monetary policy.

Empirical evidence

The 1990s saw many countries, both developed and developing, adopt reforms that increased central bank independence. This trend was strongly influenced by empirical analysis of the relationship between central bank independence and macroeconomic performance. Among developed economies, central bank independence was found to be negatively correlated with average inflation. The estimated effect of independence on inflation was statistically and economically significant. Based on data from the high inflation years of the 1970s, for example, moving from the status of the Bank of England prior to the 1997 reforms that increased its independence to the level of independence then enjoyed by the Bundesbank would be associated with a drop in annual average inflation of four percentage points.

The form of independence may also matter for inflation. Debelle and Fischer (1994) report evidence that it is the combination of goal *dependence* and instrument *independence* that produces low average inflation, although their empirical results were weak.

Even if central bank independence leads to lower inflation, the case for independence would be greatly weakened if it also leads to greater real economic instability. However, little relationship was found between measures of real economic activity and central bank independence (Alesina and Summers, 1993). In other words, countries with more independent central banks enjoyed lower average inflation rates yet suffered no cost in terms of more volatile real economic activity. Central bank independence appeared to be a free lunch.

While standard indices of central bank independence were negatively associated with inflation among developed economies, this was not the case among developing economies. Developing countries that experienced rapid turnover among their central bank heads tended to experience high rates of inflation. This is a case, however, in which causality is difficult to establish; is inflation high because of political interference that leads to rapid turnover of central bank officials? Or are central bank officials tossed out because they can't keep inflation down?

The empirical work attributing low inflation to central bank independence has been criticized along two dimensions. First, studies of central bank independence and inflation often failed to control adequately for other factors that might account for cross-country differences in inflation experiences. Countries with independent central banks may differ in ways that are systematically related to average inflation. After controlling for other potential determinants of inflation, Campillo and Miron (1997) found little additional role for central bank independence.

Second, treating a country's level of central bank independence as exogenous may be problematic. Posen (1993) has argued strongly that both low inflation and central bank independence reflect the presence of a strong constituency for low inflation. Average inflation and the degree of central bank independence are jointly determined by the strength of political constituencies opposed to inflation; in the absence of these constituencies, simply increasing a central bank's independence may not cause average inflation to fall.

Theoretical models of independence

Central bank independence has often been represented in theoretical models by the weight placed on inflation objectives. When the central bank's weight on inflation exceeds that of the elected government, the central bank is described as a Rogoff-conservative central bank (Rogoff, 1985). This type of conservatism accorded with the notion that independent central banks are more concerned than the elected government with maintaining low and stable inflation. Rogoff's formulation reflects a form of both goal independence – the central bank's goals differ from those of the government – and instrument independence – the central bank is assumed to be free to set policy to achieve its own objectives. Because the central bank cares more about achieving its inflation goal, the marginal cost of inflation is higher for the central bank than it would be for the government. As a consequence, equilibrium inflation is lower.

One problem with interpreting independence in terms of Rogoff-conservatism is that Rogoff's model implies that a conservative central bank will allow output to be more volatile in order to keep inflation

stable. Yet the empirical research finds no relationship between real fluctuations and measures of central bank independence.

An alternative way to model central bank independence is to view the central bank as having its own objectives, but the central bank must also take into account the government's objectives when deciding on policy. The central bank might have either a lower desired inflation target than the government or an output target that, unlike the government's target, is consistent with the economy's natural rate of output. If actual policy is set to maximize a weighted average of the central bank's and the government's objectives, the relative weight on the central bank's own objectives provides a measure of central bank independence. With complete independence, no weight is placed on the government's objectives; with no independence, all weight is placed on the government's objectives. If the objectives of the central bank and the government differ only in their desired inflation target, then the degree of central bank independence affects average inflation but not the volatility of either output or inflation. Such a formulation is consistent with the empirical evidence discussed above.

Often, theoretical approaches have not distinguished clearly between goal and instrument independence. Suppose independence is measured by the relative weight on the government's and the central bank's objectives. This can be interpreted as reflecting either goal dependence – the objectives of the central bank must put some weight on the goals of the government – or instrument dependence – the actual instrument setting diverges from what would be optimal from the central bank's perspective in order to reflect the government's concerns.

Independence and accountability

While many countries have granted their central banks more independence, the idea that central banks should be completely independent has come under criticism. This criticism focuses on the danger that a central bank that is independent will not be accountable. Although maintaining low and stable inflation is an important societal goal, it is not the only macroeconomic goal; monetary policy may have no long-run effect on real economic variables, but it can affect the real economy in the short run. In a democracy, delegating policy to an independent agency requires some mechanism to ensure accountability. For this reason, reforms have often granted central banks instrument independence while preserving a role for the elected government in establishing the goals of policy and in monitoring the central bank's performance in achieving these goals.

See Also

- inflation
- inflation targeting
- optimal fiscal and monetary policy (without commitment)

Bibliography

Alesina, A. and Summers, L. 1993. Central bank independence and macroeconomic performance. *Journal of Money, Credit, and Banking* 25, 157–62.

Bade, R. and Parkin, M. 1988. Central bank laws and monetary policy. Working paper. Department of Economics, University of Western Ontario.

Barro, R. and Gordon, D. 1983. A positive theory of monetary policy in a natural-rate model. *Journal of Political Economy* 91, 589–610.

Campillo, M. and Miron, J. 1997. Why does inflation differ across countries? In *Reducing Inflation: Motivation and Strategy*, ed. C. Romer and D. Romer. Chicago: University of Chicago Press.

Cukierman, A. 1992. *Central Bank Strategy, Credibility, and Independence: Theory and Evidence*. Cambridge, MA: MIT Press.

Cukierman, A., Webb, S. and Neyapti, B. 1992. Measuring the independence of central banks and its effects on policy outcomes. *World Bank Economic Review* 6, 353–98.

Debelle, G. and Fischer, S. 1994. How independent should a central bank be? In *Goals, Guidelines and Constraints Facing Monetary Policymakers*, ed. J. Fuhrer. Boston: Federal Reserve Bank of Boston.

Eijffinger, S. and de Haan, J. 1996. The political economy of central-bank independence. Special Papers in International Economics, No. 19. Princeton University.

Grilli, V., Masciandaro, D. and Tabellini, G. 1991. Political and monetary institutions and public financial policies in the industrial countries. *Economic Policy* 6, 341–92.

Posen, A. 1993. Why central bank independence does not cause low inflation: there is no institutional fix for politics. In *Finance and the International Economy*, vol. 7, ed. R. O'Brien. Oxford: Oxford University Press.

Rogoff, K. 1985. The optimal commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100, 1169–89.

Walsh, C. 2003. *Monetary Theory and Policy*, 2nd edn. Cambridge, MA: MIT Press.

How to cite this article

Walsh, Carl E. "central bank independence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000081> doi:10.1057/9780230226203.0210

The New Palgrave Dictionary of Economics Online

central limit theorems

Werner Ploberger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Central limit theorems describe the behaviour of distributions of sums of random variables. We start with the classical result of distributions of sums of independent random variables converging to the Gaussian (bell-curve) distribution. We describe the most important cases of convergence to Gaussian distributions (sums of martingale differences) as well as convergence to other distributions.

Keywords

central limit theorems; convergence; Edgeworth expansions; Feller condition; Laplace, P. S.; Lindeberg condition; long-term variance; Lyapunov condition; martingale differences; maximum likelihood; Monte Carlo simulation

Article

At the end of the 17th century, the mathematician Abraham de Moivre first used the normal distribution as an approximation for the percentage of successes in a large number of experiments. Later on, Laplace generalized his results, but it took 20th century mathematics to give an exact and complete description of this subject. So let me now describe the modern approach. We assume that for each n we have given a sequence $X_{1,n}$•,…,•$X_{n,n}$ of random variables, which we assume to be independent. Then we want to ‘approximate’ the distribution of

$$S_n = \sum_{i=1}^{n} X_{i,n}$$

by a standard normal distribution, whose density equals

$$\frac{1}{\sqrt{2\pi}}\int_A \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x.$$

Let us denote by $P(B)$ the probability of an event B. If X is a random variable, than let us denote by $E(X)$ its expectation. For $A \subseteq \mathbf{R}$ let $[X \in A]$ be the event that X takes a value into A. Written in formal terms, we want to establish that

$$\lim_{n\to\infty} P([S_n \in A]) = \frac{1}{\sqrt{2\pi}}\int_A \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \tag{1}$$

or

$$\lim_{n\to\infty} Ef(S_n) = \frac{1}{\sqrt{2\pi}}\int f(x)\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x. \tag{2}$$

The first question we have to ask ourselves is the nature of the approximation. Clearly it is impossible to approximate the distribution of S_n for *all* sets. Consider the binomial distribution discussed above. In this case, each S_n can only take a finite number of values. Therefore the possible values for all S_n lie for all n in a countable set, which has zero probability under the normal distribution.
So we have to aim at a compromise: the smaller the class of sets A or functions f, the more ‘convergent’ sequences S_n we have. The most successful compromise is the convergence in distribution of the random variables (or the weak convergence of the probability distributions). We postulate that (2) holds for all bounded, continuous functions f. This requirement can be shown to be equivalent to postulating that (1) holds for all sets A so that the boundary of A (that is, the difference between closure of A and inner points of A) has zero probability under the limiting measure. So in our case, where the limiting distribution is normal, (1) holds if A is an interval (a, b): the boundary consists of two points, namely a and b. Equation (1) does not hold if, for example, A is the set of all rational numbers in (0, 1): then the boundary equals [0, 1], which obviously has non-zero probability under the normal distribution (see

Billingsley, 1999).

It is noteworthy that there are many more equivalent ways to define convergence in distribution for unidimensional random variables; for example, convergence in distribution is equivalent to the convergence of the cumulative distribution functions to the cumulative distribution function of the limiting distribution in all points where the latter is continuous. Another well-known criterion is the convergence of the characteristic functions.

Now we are in a position to formulate our first main theorem, the central limit theorem (CLT) of Lindeberg and Feller (see Billingsley, 1995).

Suppose we have given a triangle array of random variables $X_{i,n}$, so that for each n the $X_{i,n}$ are independent, not necessarily identically distributed. We furthermore have

$$EX_{i,n} = 0,$$

$$\sum_{i=1}^{n} \mathrm{Var}(X_{i,n}) = 1.$$

Then the following two propositions are equivalent:

- The ‘Lindeberg’ condition: For all $\delta > 0$

$$\sum_{i=1}^{n} E\left(X_{i,n}^2 I\left[|X_{i,n}| > \delta\right]\right) \tag{L}$$

  converges to zero.
- Our sums

$$S_n = \sum_{i=1}^{n} X_{i,n}$$

converge in distribution to a standard normal and the ‘Feller’ condition is satisfied:

$$\max_{1 \le i \le n} \operatorname{Var}(X_{i,n}) \to 0. \tag{F}$$

It seems plausible to assume the Feller condition (F). It simply states that the maximal contribution of an individual $X_{i,n}$ to the variance of the sum gets arbitrarily small. This seems reasonable. The Lindeberg condition (L) which is necessary for our theorem is a little stronger. Not only the maximum, but the *total* contribution of the $X_{i,n}$ taking ‘large’ values to the variance of the sum, must vanish asymptotically!

It is quite easy to establish that (L) is fulfilled if

$$X_{i,n} = \frac{1}{\sqrt{n}} X_i, \tag{3}$$

where the X_i are independent and identically distributed. In the general case, a sufficient condition is the ‘Lyapunov condition’: for some fixed $\varepsilon > 0$ we have

$$\sum_{i=1}^{n} E\left(\left|X_{i,n}\right|^{2+\varepsilon}\right) \to 0.$$

So we need a little more than second moments to establish convergence to a standard normal. Practitioners often assume that the requirements of the theorems are fulfilled automatically. This assumption is quite dangerous. We need a little more than lack of outliers; the contribution to the variance of the largest values must be negligible.

This relation between higher moments and goodness of the approximation with a standard normal is extensive. Under the assumption of at least three absolute moments, the theorem of Berry–Esseen shows that in the case (3) of independent, identically distributed X_i the maximal difference between the cumulative distribution functions of S_n and the standard normal is $1/\sqrt{n}$. Related are ‘coupling’ results.

One can show that – possibly on a richer probability space – there exist *exactly* normally distributed random variables U_n. In particular, if the X_i have a Laplace transform, then the 'Hungarian construction' allows one to construct U_n so that the difference to S_n is $O(\log(n) / \sqrt{n})$. If the X_i 'only' have fourth moments, then it is easy (for the insider: use Skorohod embedding) to construct U_n so that the difference to S_n is of the order of $1 / \sqrt[4]{n}$.

All these bounds are very interesting from the theoretical point of view. Playing around with numbers for n with realistic sample sizes, one can easily see that the bounds found that way are unrealistic. Although these bounds cannot be improved, they are a little pessimistic. Nevertheless, they indicate when we venture into dangerous territory: a lack of fourth moments indicates a 'slow' convergence.

So the normal approximation is a useful first-order approximation of the distributions of sums of random variables. To improve this approximation, various techniques are used. Since the 19th century, Edgeworth expansions have proved useful. Nowadays, however, cheap computing makes direct calculation of distributions by Monte Carlo simulation possible.

Independent, non-normal limit theorems

Let us define $X_{i,n}$ to be independent, identically distributed and taking the value of zero with probability $1-\lambda/n$ and one with probability λ/n with some $\lambda > 0$. Now one has an easy example where the Lindeberg condition is not fulfilled. (For $\delta < 1$, $\Sigma_{i=1}^{n} E(X_{i,n}^2 I(|X_{i,n}| > \delta) = \lambda$, since $X_{i,n}$ can take only the values 0 and 1). Nevertheless, it is well known that $\Sigma_{i=1}^{n} X_{i,n}$ converges in distribution to a Poisson distribution with intensity λ. So the normal distribution is not the only limiting distribution of sums of independent random variables. One can, however, show that the normal and the Poisson distribution and mixtures (with possibly an infinite number of components) of these distributions are the only possible limits of sums S_n of independent, identically distributed random variables $X_{i,n}$. These limiting distributions are called 'infinitely divisible'. A precise formula for the logarithm of the characteristic function is given by the formula of Levy–Khinchin.

We even have some analogon, some generalization of the normal distribution. The properly normalized sum of normally distributed random variables is normal again. Can we generalize this property? Let us assume that

$$X_{i,n} = a_n(X_i - b_n), \tag{4}$$

where the X_i are independent and identically distributed, and the a_n are scale factors, and let us assume that the distribution of the S_n is identical to the distribution of the X_i. These distributions are called the 'stable' distributions. Their density is determined essentially by two parameters, traditionally called α and β. α determines the 'tail behaviour' and varies between 0 and 2, and β determines the symmetry.

For $\alpha=2$, we have the normal distribution, for $\alpha<2$ the distributions are more heavily tailed: in general, one has only moments of order smaller than α. There is no closed form for their densities in the general case, only the characteristic functions can be expressed by elementary functions. One special case ($\alpha=1$) is the Cauchy distribution with density

$$\frac{1}{\pi(1+x^2)}.$$

The index α determines the scale factors a_n: in general, one has $a_n = n^{\frac{1}{\alpha}}$.
Convergence of sums to stable distributions can be achieved in more general circumstances. In general, under certain conditions on the 'tail' of the X_i (the probabilities exceeding 'large' values have to obey certain regularity conditions) the sums of the $X_{i,n}$ defined by (4) one can ensure convergence (see Ibragimov and Linnik, 1971).

Central limit theorems for dependent random variables

Many econometric applications involve sums of dependent random variables. Hence it is important to remove the requirement of independence. Traditionally, one tried to replace independence by some form of 'mixing'.
Independence of two σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$ can be defined in various ways. Usually one defines $\mathfrak{A}$ and $\mathfrak{B}$ to be independent if for all $A\in\mathfrak{A}$ and $B\in\mathfrak{B}$

$$P(A\cap B) = P(A)P(B).$$

Another usual definition is that for all $A\in\mathfrak{A}$

$$P(A/\mathfrak{B}) = P(A),$$

where $P(\cdot/\cdot)$ should denote the conditional probability. Consequently, one can measure the 'degree of dependence' of σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$ by

$$\alpha(\mathfrak{A}, \mathfrak{B}) = \sup_{A \in \mathfrak{A}, B \in \mathfrak{B}} |P(A \cap B) - P(A)P(B)|$$

or

$$\psi(\mathfrak{A}, \mathfrak{B}) = \sup_{A \in \mathfrak{A}} |P(A/\mathfrak{B}) - P(A)|.$$

Suppose one has give a process X_t. Then one defines the ‘mixing coefficients’

$$\alpha_k = \sup_t \alpha\left(\mathfrak{A}_\sigma\left(X_t, X_{t+1}, \ldots\right), \mathfrak{A}_\sigma\left(X_{t-k}, X_{t-1-k}, \ldots\right)\right)$$

or

$$\psi_k = \sup_t \psi\left(\mathfrak{A}_\sigma\left(X_t, X_{t+1}, \ldots\right), \mathfrak{A}_\sigma\left(X_{t-k}, X_{t-1-k}, \ldots\right)\right).$$

Typically, conditions like

$$\sum \sqrt{\alpha_k} < \infty$$

or

$$\psi_k \to 0$$

are sufficient conditions for a CLT. So the CLT remains valid for stationary processes if the random variables in questions get less and less dependent if the time difference gets larger and larger (Ibragimov

and Linnik, 1971; Davidson, 1994).

CLT for martingale differences

One of the most important applications is the CLT for martingale differences. A process X_t is a 'martingale difference' if for all t

$$E(X_t / \mathfrak{F}_{t-1}) = 0,$$

where $\mathfrak{F}_{t-1}$ is an increasing sequence of σ-algebras which contain at least X_{t-1}, X_{t-2},•…•. Then we have a result perfectly analogous to the case of independent random variables.
Suppose we have given a triangle array $X_{t,T}$, t=1, …, T, of martingale differences with σ-algebras $\mathfrak{F}_{t-1,T}$ and the following two conditions are satisfied:

- the conditional Lindeberg condition

$$\sum_{t=1}^{T} E(X_{t,T}^2 I_{[|X_{t,T}| \geq \varepsilon]} / \mathfrak{F}_{t-1,T}) \to 0,$$

- the norming condition

$$\sum_{t=1}^{T} E(X_{t,T}^2 / \mathfrak{F}_{t-1,T}) \to 1,$$

where the convergence should be understood to be in probability. Then

$$S_n = \sum_{i=1}^{n} X_{i,n}$$

converges in distribution to a standard normal distribution (Davidson, 1994; Hall and Heyde,

1980).

This limit theorem is one of the most important ones for applications in econometrics. It is relatively easily seen that derivatives of log-likelihood functions are martingale differences. Hence this theorem is instrumental in establishing the limit theorems for maximum likelihood estimators.
An easy consequence of the theorem is that for every (strictly) stationary, ergodic martingale difference X with $\sigma^2 = E(X_t^2) < \infty$ we have an almost classical CLT:

$$\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n} X_i$$

which converges in distribution to a standard normal.

Gordin's theorem

Martingale differences form a large class of processes. Unfortunately, however, this class is not sufficiently large for many important applications (martingale differences must be, for example, uncorrelated). As an alternative, one might use mixing conditions. These conditions are, however, hard to verify. They usually involve inequalities involving *all* events from the σ-algebras involved. Hence a theorem allowing for general, autocorrelated processes with conditions which are easy to verify is an important tool in theoretical econometrics. Such a result was found by Gordin in 1969. Hayashi (2000) demonstrates the versatility of the theorem.
Suppose we have a stationary, ergodic process X_i, $i \in Z$ so that $EX_i^2 < \infty$. Assume that $\mathfrak{F}_i$ are adapted σ-algebras (that is, X_i are $\mathfrak{F}_i$-measurable), and let

$$\varepsilon_i = E(X_i / \mathfrak{F}_1) - E(X_i / \mathfrak{F}_0).$$

Then let us assume that

$$\sum_{i=1}^{\infty} \sqrt{E\varepsilon_i^2} < \infty .$$

Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$$

converges in distribution to a normal distribution with zero mean and variance σ_{LT}^2, where

$$\sigma_{LT}^2 = E\left(\sum_{i=1}^{\infty} \epsilon_i\right)^{2'}.$$

σ_{LT}^2 is usually called the ‘long-term variance’.

Conclusion

Almost all theorems about limit distributions of estimators and test statistics depend on central limit theorems. So it should not be surprising that central limit theorems and their generalizations are an active field of research. Especially, generalizations of the concept of convergence in distribution to more general spaces generate theorems, which are important from the theoretical as well as the practical point of view. Billingsley (1999) and Davidson (1994) give an introduction to these ‘functional limit theorems’.

See Also

- functional central limit theorems

Bibliography

Billingsley, P. 1995. *Probability and Measure*, 3rd edn. New York: Wiley.

Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd edn. New York: Wiley-Interscience.

Davidson, J. 1994. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.

Hall, P. and Heyde, C.C. 1980. *Martingale Limit Theory and its Application*. New York: Academic.

Hayashi, F. 2000. *Econometrics*. Princeton, NJ: Princeton University Press.

Ibragimov, I.A. and Linnik, Yu.V. 1971. *Independent and Stationary Sequences of Random Variables*. Groningen: Wolters-Noordhoff.

How to cite this article

Ploberger, Werner. "central limit theorems." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000578> doi:10.1057/9780230226203.0211

The New Palgrave Dictionary of Economics Online

central place theory

Marcus Berliant

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Central place theory is a descriptive theory of market area in a spatial context. Its main assumptions are that consumer population is distributed uniformly while firms locate in cities; the latter form a hierarchy with overlapping market areas. But central place theory runs afoul of Starrett's spatial impossibility theorem. Not grounded in the analytical tools of modern economics, central place theory does not have firm foundations. Thus, it is difficult to build on central place theory, either theoretically or empirically.

Keywords

central place theory; city hierarchy; increasing returns to scale; Krugman, P.; spatial impossibility theorem; urban agglomeration

Article

Central place theory is a descriptive theory of market area in a spatial context. Its definition, history, and relation to modern microeconomic theory are set out in this article.

Central place theory is a collection of loosely related, informal, descriptive models of city size, city location, and market area based on the trade-off between increasing returns to scale in production and the cost of transport of goods from firm to home. Land markets are often absent. At its core, central place theory is an empirically motivated description of production in southern Germany. It is a remarkable empirical regularity in search of a formal theory; a better name would be ‘central place regularity’.

The beginnings of the theory are attributed to Christaller (1933), who first made detailed observations of urban hierarchies and then attempted to model them. The basic ideas put forward are that consumer population is distributed uniformly, while firms locate in cities. Cities form a hierarchy in that cities higher in the hierarchy produce all the goods that cities one level lower in the hierarchy produce, and one more. The ratio of market areas of a commodity produced only at a given level of the hierarchy (and

above) to the market area of a commodity produced at the next lower level of the hierarchy (and above) is assumed to be constant, independent of the level in the hierarchy considered. Thus, the cities in a given area form a hierarchy where the size of a city's market area and the variety of commodities it offers are perfectly correlated. In graphical terms, the result is a collection of hierarchically ordered cities with the market areas of cities not at the same level of the hierarchy overlapping, but market areas of cities at the same level disjoint. Commodities characterized by low transport cost but high returns to scale are provided by a few cities high in the hierarchy. Commodities characterized by high transport cost but low returns to scale are provided by most cities.

Lösch (1944) expanded on this theory. He postulated a homogeneous agricultural plane with farmers. Some turn to beer production, and face linear, downward sloping demand curves with choke prices, that is, prices above which the demand is for beer is zero. For a given price at the brewery, total delivered price increases with distance from the plant due to transport cost. In the plane with a uniform distribution of inebriated consumers or farmers, demand for a firm's beer is given by the volume of a cone centred at the brewery, with height given by the brewery's mill price and the slope of its sides determined by the demand curve and the cost of beer transport. With a marginal cost curve, equilibrium can be found. Unfortunately, the collection of bases of cones, namely, disks, does not partition the plane. So hexagons are used, forming a Teutonic triangulation of hierarchical hexagons. In this theory, the central places are the breweries. (St. Louis is a prime example.)

One can view the theory as producing a complex of overlapping, ordered layers of hexagonal partitions of the plane corresponding to the market areas of cities in a hierarchy. Agriculture is the basis for and genesis of this structure.

The theory has developed beyond these basic descriptive models; see McCann (2001, ch. 2.7) for a nice summary and cites. Hartwick (2004) is the culmination of a line of research more in accord with optimizing behaviour, pricing, and trade theory that also relates the models to the rank-size rule.

The reader should be cautious in interpreting this entire literature because equilibrium and efficiency are often confused, while the models tend to be mechanistic in nature as opposed to allowing agents to optimize in equilibrium. To the general economist, the theory will appear to be informal and imprecise. Paul Krugman (1995, pp. 38–41) criticizes central place theory, or ‘Germanic geometry’, for its lack of formal foundations, particularly regarding market structure and firm behaviour. This criticism applies even to the contemporary literature. (Paul Krugman is also credited with the first alliteration in this literature. This article only builds on the original contribution.)

Even if one is willing to overlook these defects, there is one further important flaw. Central place theory generally runs afoul of Starrett's spatial impossibility theorem; see Starrett (1978), Fujita (1986), and Fujita and Thisse (2002, ch. 2.3) for discussion. In essence, the impossibility theorem says that, in a closed economy with perfect and complete markets at all locations, location-independent utility and production functions, and no relocation cost, there is no competitive equilibrium where commodities are transported. Thus, if the assumptions are satisfied, either there is no equilibrium or in equilibrium agents and commodities are distributed uniformly among inhabited locations, and locations are autarkic. Central place theory apparently makes these assumptions, though due to its imprecision perhaps it doesn't. Naturally, although the literature considers consumer migration at times, the assumption of a uniform distribution of consumers could render the theorem inapplicable. I conjecture that it simply makes the existence of an (autarkic) equilibrium more likely. But this is probably not worth pursuing, as

location models that fix consumer locations in a uniform distribution can generate only cities without people.

So where does this leave us? The modern theory of agglomeration, and thus the modern theory of central places, begins with the impossibility theorem. Its contrapositive tells us that, to generate models with non-trivial agglomeration at equilibrium, at least one of the hypotheses must be violated. Even then, equilibrium might not exist, or in equilibrium cities could collapse to a point or have agents spread uniformly. Models of non-trivial cities involve a very delicate balancing act between forces pulling agents together and forces pushing them apart. The New Economic Geography has provided one of several possible types of models capable of producing cities and even hierarchies of cities. Fujita and Mori (1997) and Fujita, Krugman and Mori (1999) generate a form of central place theory in a general equilibrium framework by employing imperfect competition and increasing returns at the firm level. Unfortunately, this type of model has many defects, as detailed in Berliant (2006), including a reliance on specific functional forms and indeterminacy: one equilibrium is selected from a continuum.

Central place theory is not grounded in the analytical tools of modern economics, so it does not have firm foundations. Thus, it is difficult to build on central place theory, either theoretically or empirically. In my view, the future of central place theory is as a stylized fact to be explained by our models, much like the rank-size rule.

See Also

- spatial economics
- systems of cities
- urban agglomeration
- urban economics

Bibliography

Berliant, M. 2006. Well isn't that spatial?! Handbook of Regional and Urban Economics, vol. 4: a view from economic theory. *Journal of Economic Geography* 6, 107–10.

Christaller, W. 1933. *Central Places in Southern Germany*. Englewood Cliffs: Prentice-Hall, 1966.

Fujita, M. 1986. Urban land use theory. In *Location Theory*, ed. J. Lesourne and H. Sonnenschein. New York: Harwood Academic Publishers.

Fujita, M., Krugman, P. and Mori, T. 1999. On the evolution of hierarchical urban systems. *European Economic Review* 43, 209–51.

Fujita, M. and Mori, T. 1997. Structural stability and evolution of urban systems. *Regional Science and Urban Economics* 27, 399–442.

Fujita, M. and Thisse, J.-F. 2002. *Economics of Agglomeration*. Cambridge: Cambridge University Press.

Hartwick, J. 2004. Trade in a hierarchical system of cities. Mimeo. Kingston, ON: Department of Economics, Queen's University.

Krugman, P. 1995. *Development, Geography, and Economic Theory*. Cambridge, MA: MIT Press.

Lösch, A. 1944. *The Economics of Location*. New Haven: Yale University Press, 1954.

McCann, P. 2001. *Urban and Regional Economics*. Oxford: Oxford University Press.

Starrett, D. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 17, 21–37.

How to cite this article

Berliant, Marcus. "central place theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000087> doi:10.1057/9780230226203.0212

The New Palgrave Dictionary of Economics Online

certainty equivalence

Xavier Freixas

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

In order to take a decision in an uncertainty context, it is necessary, from a theoretical point of view, to build a model and specify all the consequences in every possible state of the world. In applied work this method is much too involved. Consequently, for applied purposes, it would be interesting to have a model where uncertainty is treated in such a way that the decision problems are as simple as the equivalent ones in a certainty framework. The identification of the conditions under which such an isomorphism between the optimal decisions under uncertainty and the optimal decisions in an equivalent certainty context holds is called the certainty equivalent problem.

Theil (1954) has been the first to point out the problem and to suggest a specific model in which the certainty equivalent property holds.

Theil imposes the following two assumptions: (i) the vector *x* of instruments and the vectory *y* of result variables are related by a simple equation

$$y = g(x) + S \tag{1}$$

where *S* is a vector of random variables, that we can take to have a zero expected value without loss of generality. (ii) The decision-maker's objective function is quadratic and can be written as

$$u(x, y) = A(x) + \sum_{i=1}^{m} A_i(x)\, y_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{y=1}^{m} A_{ij} y_i y_j \tag{2}$$

Using such a model it is straightforward to show that whenever the optimal solution to the problem of maximizing the expected utility under the constraint (1) exists, it is the same as the optimal solution to the equivalent certain problem:

$$\begin{cases} \text{Max } u(x, y) \\ y = g(x) \end{cases}$$

This result is extended not only to the multiperiod problem but also to the case where the decision-maker receives more and more information as time elapses. The resulting stochastic problem is then more involved, but it is simply solved by use of dynamic programming, the optimal strategy in period *t* being a function of the previously observed signals η_{t-k}

$$X_t^* = x_t^*(\eta_1, \eta_2, \ldots, \eta_{t-1})$$

Again, the conditions for the first period solution to this problem to be the solution of the equivalent certain problem are very strong. As before, it has to be the case that the objective function is quadratic, but in addition the constraint relating instruments to results is restricted to be of the following type:

$$y = RX + S$$

where *R* is a matrix with some required specifications (namely, the value of the instrument variables of one period have no effect on the result variables of the preceding periods).
The conditions that guarantee the equivalence between the uncertainty problem and the certainty problem are so restrictive, that an alternative view of the problem has been suggested. Instead of setting restrictions on the parameters of the model, the uncertainty itself is restricted to be ‘small’. Formally, this is equivalent to consider an entire class of problems that can be ranked in their uncertainty as measured by a parameter ε and whose limit is the certain problem. The question is then to know under what conditions the solution to the limit of the random problems, that is equal to the one of the certain problem, is independent of ε to the first order, so that

$$\frac{dE\left[x_t^*(\eta_1, \ldots, \eta_t)\varepsilon\right]}{d\varepsilon} = 0 \quad \text{for } \varepsilon = 0.$$

This slightly different point of view is called the ‘first order certainty equivalence’ problem and has been dealt with by Theil (1957) and Malinvaud (1969).

The very general conditions obtained by Malinvaud for the first order certainty equivalent to hold are (i) that the objective function is twice differentiable and (ii) that the optimal strategy is continuous with respect to the degree of uncertainty. If this condition holds, the optimal values of the instruments at time 1 are, to the first order approximation, independent of the degree of uncertainty.

It is clear that this condition cannot be met if there are constraints on the future instrument variables, since this will bring in a kink. A particular and natural example of a framework where the first order certainty equivalence does not hold is when decisions are irreversible. As pointed out in Henry (1974), it is then the case that the value of the decision in the first period will affect the decision set in the following periods, and consequently, the use of the certainty equivalent would generate a systematic error.

See Also

- risk

Bibliography

Henry, C. 1974. Investment decisions under uncertainty: the irreversible effect. *American Economic Review* 64, 1996–2012.

Malinvaud, E. 1969. First order certainty equivalence. *Econometrica* 37, 706–18.

Simon, H. 1956. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica* 24, 74–81.

Theil, H. 1954. Econometric models and welfare maximization. *Weltwirtschaftigsches Archiv* 72, 60–83.

Theil, H. 1957. A note on certainty equivalence in dynamic planning. *Econometrica* 25, 346–9.

How to cite this article

Freixas, Xavier. "certainty equivalence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_C000092> doi:10.1057/9780230226203.0213

The New Palgrave Dictionary of Economics Online

CES production function

Ryuzo Sato

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bergson family of utility functions; CES production function; Cobb–Douglas functions; elasticity of substitution; Jensen inequalities

Article

The CES (constant elasticity of substitution) production function, including its special case the Cobb–Douglas form, is perhaps the most frequently employed function in modern economic analysis. Not only is the CES function used for the formal depiction of production technology, it is used as a convenient tool for empirical analysis as well. In addition to production theory, the CES function, more commonly known as the Bergson family of utility functions, is employed in utility theory.

Ordinary CES production functions

The simplest form of CES function utilized in production theory is the constant returns to scale type (Arrow et al., 1961):

$$Y = T[\alpha K^{-\rho} + (1-\alpha)L^{-\rho}]^{-(1/\rho)} \tag{1}$$

where *Y*=output, *K*=capital, *L*=labour, and the parameters *T*, α and ρ satisfy the conditions: $T \geq 0$, $0 \leq \alpha \leq 1$ and $\rho \leq -1$. As is implied by its name, the elasticity of factor substitution between capital and labour for production function (1) is expressed as some constant value.

For any neoclassical production function $Y = f(K, L)$, the elasticity of factor substitution between capital

and labour is defined as the proportionate change in the *K/L* ratio (*k*) relative to the proportionate change in the marginal rate of factor substitution $r = f_L / f_K$ along a given isoquant curve, where $f_L = \partial Y / \partial L$ and $f_K = \partial Y / \partial K$ are the respective marginal products. That is,

$$\sigma = \frac{d \log k}{d \log r} = \frac{f_K f_L (f_K K + f_L L)}{KL(2 f_{KL} f_K f_L - f_K^2 f_{LL} - f_K^2 f_{KK})}, \tag{2}$$

where σ represents the elasticity of substitution and f_{KL}, f_{KK} and represent the cross and own derivatives of the respective marginal products.
Applying definition (2) to production function (1) we obtain:

$$\sigma = \frac{1}{1+\rho} \quad \text{or} \quad \rho = \frac{1-\sigma}{\sigma}. \tag{3}$$

Consequently, it is easy to see why ρ is often referred to as the ‘substitution’ parameter. The α parameter in production function (1) is the ‘distribution’ parameter that permits the relative importance of capital and labour to vary in production. In the extreme case where $\rho \to 0$ or $\sigma = 1$, the CES function (1) converges to the Cobb–Douglas form:

$$Y = T K^{\alpha} L^{1-\alpha}. \tag{4}$$

In this form, it is evident that α and $1-\alpha$ are the production elasticities of capital and labour respectively. Under conditions of perfect competition, α and $1-\alpha$ will also equal the respective relative income shares (or income distribution). The *T* parameter in both production functions (1) and (4) is the ‘efficiency’ (or technical progress) parameter.
With the exception of its special case the Cobb–Douglas form, the ordinary CES production function is cumbersome and difficult to manipulate. However, the underlying expression for the marginal rate of factor (technical) substitution has a simple form and this is the primary reason for the popularity and wide use of this production function.

Homothetic and non-homothetic CES production functions

Any monotonic transformation of the ordinary CES production functions (1) belongs to a class of CES production functions called the homothetic class, that is,

$$Y = F(f),\ F' > 0,$$

where

$$f = T[\beta K^{-\rho} + (1-\beta)L^{-\rho}]^{-(1/\rho)}. \tag{5}$$

In addition to the class of homothetic CES production functions, there is a more general, and perhaps more meaningful, class of non-homothetic CES production functions. One can refer to the class of non-homothetic CES functions as the ‘general class’ of CES production functions as it contains the homothetic class as a special case.

The class of non-homothetic CES production functions is derived as a solution to the differential equation that defines a constant elasticity of factor substitution. However, unlike the case of the homothetic CES production functions where the marginal rate of factor substitution is (implicitly) assumed to be independent of either the output level and the process of technical change, the family of non-homothetic CES production functions explicitly assumes that output level and technical change will have some kind of impact on the factor input ratio.

The class of non-homothetic CES production functions can be expressed as follows (Sato, 1975):

$$C_1(Y)K^{-\rho} + C_2(Y)L^{-\rho} = 1,\ \rho = \frac{1-\sigma}{\sigma},\ \sigma \neq 1, \tag{6a}$$

$$C_1(Y)\log K + C_2(Y)\log L = 1,\ \sigma = 1, \tag{6b}$$

where C_1 and C_2 are functions of the output level Y. When $C_1=aC_2$, where a is a constant, we can express (6a) as

$$K^{-\rho} + aL^{-\rho} = \frac{1}{C_1(Y)} = B(Y)$$

or

$$Y = B^{-1}(K^{-\rho} + aL^{-\rho}).$$

Note that with the appropriate choice of B and a, we can always express the above in the form of the ordinary CES production function. In general, the non-homothetic CES production functions are in an implicit form and can never be expressed in an explicit form.

Classification of non-homothetic CES production functions

The general class of non-homothetic CES production functions can be classified in a number of ways, depending on the specific purpose in mind. For example, it is well known that the ordinary CES production function belongs to the explicit and separable class of homothetic CES functions. In a similar fashion, we can derive an explicit and separable class of non-homothetic CES functions (Sato, 1974). Another way of classifying non-homothetic CES production functions is to consider the form of the underlying marginal rate of factor substitution function. However, the most precise way of classifying the family of non-homothetic CES production functions is to utilize Lie group theory.

A historical note

It was Arrow et al. (1961) who first utilized the ordinary CES production function expressed in (1) for the estimation of constant returns to scale aggregate production functions using cross-country data. Since then, the ordinary CES function and its variants have been widely applied in both theoretical and empirical work involving production behaviour.
Prior to its application to production analysis, the ordinary CES function, was utilized in the study of demand as the Bergson family of utility functions (Samuelson, 1965). Earlier writers in growth economics, such as Dickinson (1955) and Solow (1956), used special cases of the CES function, such as $\sigma = 2$. In the field of mathematics, Courant (1959, vol. 1, pp. 557, 601) has used the explicit form of the ordinary CES function in conjunction with the so-called Jensen inequalities.

A published note by McElroy (1967) contains the first reference to the non-homothetic CES production family. However, it was not until later that Sato (1974) derived an explicit form of the non-homothetic CES production function. The application of Lie group theory to CES production functions was first presented in 1975. This work demonstrated that the ‘projective’ type of technical change with eight essential parameters can be used most effectively to classify the general non-homothetic CES family of production functions. This work is summarized in Sato (1981, ch. 5).

See Also

- Cobb–Douglas functions
- elasticity of substitution

Bibliography

Arrow, K., Chenery, H., Minhas, B. and Solow, R. 1961. Capital–labour substitution and economic efficiency. *Review of Economics and Statistics* 43, 225–50.

Courant, R. 1959. *Differential and Integral Calculus*. 2 vols. New York: Wiley.

Dickinson, H. 1955. A note on dynamic economics. *Review of Economic Studies* 22(3), 169–79.

McElroy, F. 1967. Note on the CES production function. *Econometrica* 35, 154–6.

Samuelson, P. 1965. Using full duality to show that simultaneously additive direct and indirect utilities implies unitary price elasticity of demand. *Econometrica* 33, 781–96.

Sato, R. 1974. On the class of separable non-homothetic CES production functions. *Economic Studies Quarterly* 25(1), 42–55.

Sato, R. 1975. The most general class of CES functions. *Econometrica* 43, 999–1003.

Sato, R. 1981. *Theory of Technical Change and Economic Invariance*. New York: Academic Press.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

How to cite this article

Sato, Ryuzo. "CES production function." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000095> doi:10.1057/9780230226203.0214

The New Palgrave Dictionary of Economics Online

ceteris paribus

J.K. Whitaker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

ceteris paribus; endogeneity and exogeneity; partial equilibrium; time-period analysis

Article

The Latin phrase ‘ceteris paribus’, which translates as ‘other things the same’, is much invoked by economists. Its popularity stems from its prominent use by Alfred Marshall (1920, pp. xiv–xv, 366–70), who invented the metaphor of ‘the pound called Coeteris Paribus’ – pound being used here in the same sense as in impoundment – in which are imprisoned ‘those disturbing causes, whose wanderings happen to be inconvenient’ (1920, p. 366).
The term ‘ceteris paribus’ has no clearly settled technical meaning among economists, so that an attempt to chronicle its usage would be both difficult and unrewarding. Instead, it seems preferable to distinguish the most important alternative ways in which the phrase might be employed, alluding only briefly to the pertinent literature. It is important to distinguish at the outset three broad ways in which the phrase might be used. These are:

- as a reminder that any practicable theory must take for granted the stability and continuance of certain background circumstances;
- • as a warning, when using a theory predictively, that certain variations in circumstances admitted by the theory have been assumed not to occur;
- as an instruction to hold hypothetically constant some members of a set of *necessarily* covarying variables while changes in the others are contemplated.

For example, an analysis of the movement of a group of adjacent cooling towers during gales might (i) abstract from earthquakes, or (ii) hold constant ambient temperature while considering the effects of varying wind speed, or (iii) analyse the swaying of one tower in a high wind on the assumption that the other towers are perfectly rigid, even though they too must actually sway in a way that subtly alters the wind currents buffeting the first tower. In the language of econometric models, these three usages of

‘ceteris paribus’ can be characterized as (i) a reminder that the model's structure is assumed not to change, or (ii) a warning that certain *exogenous* variables are presumed to remain constant when others change, or (iii) an instruction to hold constant certain *endogenous* variables while varying others, even though this is not justified by any separability properties of the model's structure.

The first two usages pose no difficulties. In each, the invocation of ceteris paribus merely serves as a reminder that a more comprehensive or elaborate analysis might have been attempted. The risk of earthquakes could have been incorporated into the analysis of cooling-tower stability at the price of added complexity. But a failure to do so is without methodological significance. The incidence of earthquakes is unlikely to be affected by any movement of the towers, so that the exclusion merely singles out a convenient stopping place on the inevitable trade-off between comprehensiveness and complexity. Analogously, in predicting with an econometric model it would be possible to make careful predictions of the changes in all exogenous variables that accompany a tax cut. But a failure to do so involves no logical inconsistency, and the resulting ceteris-paribus prediction of the tax cut's effects will still have substantive interest.

It is the third usage alone, with its implied logical inconsistency, which poses distinct difficulties of interpretation and methodological justification. To start with, the assertion that certain variables are mutually interdependent presumes knowledge, at least in principle, of a correct comprehensive theory in which these variables are endogenous. For economists, the requisite background theory has usually been that of Walrasian competitive general equilibrium. In such a context, the invocation of ceteris paribus in its third sense to freeze hypothetically certain endogenous variables (or, more generally, to treat them as if exogenous) can itself be given at least three alternative rationalizations.

1 Partial equilibrium analysis as an approximation

The focus here is on the demand–supply interactions in one market or a few closely interrelated markets as exogenous shifts occur, prices in all other markets being treated as hypothetically constant (or perhaps in some cases varied exogenously). Such a procedure is inconsistent with the supposed background general-equilibrium theory which implies that all prices vary interdependently. But it may give an adequate *approximate* representation of the particular markets being examined (see Viner, 1953, p. 199). This is more likely the weaker and more diffuse are connections to, and feedbacks from, markets outside the examined set. Smallness relative to the entire economy is usually helpful in this regard, but such questions have received surprisingly little detailed analysis.

2 Approach by successive approximation

Here the use of ceteris paribus restrictions is viewed as a necessary transitional step towards the evolution or understanding of a fully-comprehensive general-equilibrium theory. The limitations of human comprehension, its need to understand and test only one link of a complete chain at a time, calls for a piecemeal step-by-step progression from the crude but simple to the sophisticated but more complex, even though such a proceeding would appear illogical to an all-comprehending Cartesian intelligence. It should, however, be observed that this progression could well take place by starting with a highly aggregated general equilibrium model and successively reducing the degree of aggregation,

instead of by starting with a simple partial-equilibrium model and gradually expanding its coverage until general equilibrium is reached – as is Marshall's clearly stated strategy (1920, pp. xiv–xv).

3 Illuminating thought experiment

Conceptual experiments which hold constant certain endogenous variables, or vary them arbitrarily, may perform a valuable heuristic role in aiding comprehension of the attainment and character of general equilibrium, even though they are not part of the theory's logical structure. Thus, the construction of Walrasian market excess demand functions, by the mental experiment of facing each individual with the same arbitrary price vector and then aggregating, is heuristically valuable despite the fact that all market excess demands must be zero in equilibrium. In part this heuristic value comes from pertinence to the disequilibrium meta theory in which any equilibrium theory must be embedded, a meta theory which might be visualized only vaguely and informally. Mental experiments of this type have been termed ‘individual’ or ‘ceteris paribus’ experiments by Patinkin, who contrasts them with ‘market’ or ‘mutatis mutandis’ experiments in which endogenous variables are always constrained to satisfy the requirements of the underlying general equilibrium structure (1965, pp. 11–12).

These three different ways of invoking ceteris paribus to freeze or ‘exogenize’ some endogenous variables may be contrasted briefly by saying that the first views partial-equilibrium theory as sometimes preferable to general-equilibrium theory, the second regards partial-equilibrium theory as an interim step towards general-equilibrium theory, and the third interprets ceteris-paribus experiments as heuristic aids sustaining general-equilibrium theory.

The partial-equilibrium approach is closely associated with Marshall, who popularized its use, although Cournot (1838) among others had employed it previously. But Marshall's *methodological* discussion of the use of ceteris paribus restrictions arose in the narrower context of his time-period analysis, which is conducted within a framework already partial-equilibrium in character (1920, pp. 366–80). Considering a single industry (his example is fishing), he imprisons in the pound of ceteris paribus those variables, exogenous or endogenous, whose movement is very rapid or very slow compared with those whose equilibrium and comparative-static properties he wishes to explore. The aim is to gain rough insight into likely time paths, given that explicit dynamic analysis is not feasible (see Viner, 1953, p. 206).

The use, other than for frank approximations, of ceteris paribus assumptions which conflict with underlying general-equilibrium requirements (that is, the use of individual rather than market experiments) has been attacked as illogical or misleading by Friedman (1949) and Bailey (1954) in the context of demand functions, and by Buchanan (1958) more generally. A judicious assessment and summing up is provided by Yeager (1960).

Applications of ceteris paribus ideas to growth paths rather than stationary equilibria have been pioneered by Fisher and Ando (1962).

In closing, mention might be made of the classical notion of ‘disturbing causes’ as set out by J.S. Mill (1844, Essay V). Any deductive theorist who regards his assumptions as true, rather than mere means for generating refutable statements, must view his (valid) deductions as also true in the absence of disturbing causes not allowed for in his assumptions (see Keynes, 1891, pp. 204–13). Are such disturbing causes to be viewed as ruled out by a ceteris paribus assumption? According to Mill, they are in the statement of general economic theory (when, for example, other motives than the pursuit of

wealth are excluded) but not in its specific applications, when due allowance must be made *ex ante* for all likely disturbing causes. Thus, the ruling out of disturbing causes is meant as nothing but a device to permit statement and development of a common theoretical skeleton which must be fleshed out whenever specific use is made of it.

See Also

- Marshall, Alfred

Bibliography

Bailey, M.J. 1954. The Marshallian demand curve. *Journal of Political Economy* 62, 255–61.

Buchanan, J.M. 1958. Ceteris paribus: some notes on methodology. *Southern Economic Journal* 24, January, 259–70.

Cournot, A.A. 1838. *Mathematical Principles of the Theory of Wealth*. Trans., New York: Macmillan, 1897.

Fisher, F.M. and Ando, A.K. 1962. Two theorems on ceteris paribus in the analysis of dynamic systems. *American Political Science Review* 56, 108–13.

Friedman, M. 1949. The Marshallian demand curve. *Journal of Political Economy* 57, 463–95. Repr. in M. Friedman, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Keynes, J.N. 1891. *The Scope and Method of Political Economy*. London: Macmillan.

Marshall, A. 1920. *Principles of Economics*, vol. 1, 8th edn. London: Macmillan.

Mill, J.S. 1844. *Essays on Some Unsettled Questions of Political Economy*. London: Parker.

Patinkin, D. 1965. *Money, Interest and Prices*. 2nd edn, New York: Harper & Row.

Viner, J. 1953. Cost curves and supply curves. In American Economic Association, *Readings in Price Theory*. Homewood, IL: Irwin. First published in *Zeitschrift für Nationalökonomie* 3 (September 1931), 23–46.

Yeager, L.B. 1960. Methodenstreit over demand curves. *Journal of Political Economy* 68, 53–64.

How to cite this article

Whitaker, J. K. "ceteris paribus." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000096> doi:10.1057/9780230226203.0215

The New Palgrave Dictionary of Economics Online

Ceva, Giovanni (1647/48 – 1734)

Giancarlo Gandolfo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Ceva, G.; Gresham's Law; mathematical economics; quantity theory of money

Article

Mathematician, hydraulic engineer and mathematical economist, Ceva was born in Milan in 1647 or 1648 and died in Mantua in 1734. He studied at the University of Pisa; later he obtained a post at Gonzaga's court in Mantua, where he became the chief technician and applied his mathematical skill to technical and administrative problems.

As a mathematician he is known for the theorem (1678) concerning the concurrency of the transverse lines from the vertices of a triangle, which is named after him; his work on fluvial hydraulics is summed up in *Opus hydrostaticum* (1728). His studies in economics are contained in a work of 1711, where he studied monetary problems. Here we find a statement of the quantity theory of money: *ceteris paribus*, the value of money varies inversely with its quantity and directly with the number of people. The latter assertion may seem odd, but it is not if we interpret ‘number of people’ as a proxy for the transaction variable in the quantity theory equation (as is implicit in Ceva's Postulate II). We also find an independent statement of Gresham's Law and a study of the problems of a plurimetallic standard.

The interest of this work, however, does not lie in its economics, where no objectively new contributions are made, but in its methodological content and message. Ceva was the first to conceive, to state lucidly and to apply unhesitatingly the idea of *systematically* employing the mathematical method in economics as an indispensable tool with which to reason rigorously, to understand difficult and otherwise obscure phenomena and to put them in order. His analytico-deductive treatment, which proceeds by definitions, postulates, remarks, propositions, theorems and corollaries, is indeed the first example of mathematical economics as we now understand it.

Selected works

1678. *De lineis rectis se invicem secantibus statica constructio*. Mediolani. (A static construction

concerning straight lines which intersect one another. Milan.)

1711. *De re numaria quoad fieri potuit geometrice tractata*. Mantuae. (On money, treated mathematically as far as has been possible. Mantua.) Reprinted, with editor's Preface by E. Masè-Dari, as *Un precursore della econometria. Il saggio di Giovanni Ceva 'De re numaria' edito in Mantova nel 1711*, Modena: Pubblicazioni della Facoltá di Giurisprudenza, 1935. French translation, with translators' Introduction and notes by G.H. Bousquet and J. Roussier, in *Revue d'histoire économique et sociale*, 1958, No. 2, 129–69.

1728. *Opus hydrostaticum*. (A work on hydrostatics.) Mantua.

Bibliography

Dictionary of Scientific Biography, vol. 3. 1971. New York: Scribner's Sons. (On Ceva's contribution to mathematics.)

Dizionario biografico degli italiani, vol. 24. 1980. Rome: Istituto dell'Enciclopedia Italiana. (On Ceva's life and works.)

Nicolini, F. 1878. Un antico economista matematico. *Giornale degli Economisti*, October, 11–23.

How to cite this article

Gandolfo, Giancarlo. "Ceva, Giovanni (1647/48–1734)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000097> doi:10.1057/9780230226203.0216

The New Palgrave Dictionary of Economics Online

Chalmers, Thomas (1780 – 1847)

D.P. O'Brien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chalmers, T.; Malthus's theory of population; overproduction; over-saving; Poor Law

Article

Chalmers was born in Anstruther, Fife, and died in Edinburgh. Though he was strongly attracted to mathematics and physics in his youth, he is famous as a theologian and economist and as an active worker in the field of poor relief. Appointed to a parish in 1803, he later moved to Glasgow, where he began a famous and influential experiment in the administration of poor relief through dividing up the large parish of St John into small units and relying on a large number of voluntary helpers. He left Glasgow to become Professor of Moral Philosophy at St Andrews in 1823; in 1828 he became Professor of Divinity at Edinburgh and in 1843 he was centrally involved in the famous ecclesiastical divisions which produced the Free Church.

Endorsing Malthus's theory of population, he argued fervently (and repetitively) that the answer to the problem lay in moral education which would, in turn, lead to moral restraint. He opposed the Poor Law: it stimulated population, and interfered with private charity, which, his Glasgow experience had convinced him, was more effective. His work on aggregate demand and gluts – he argued that there could be both overproduction and over-saving since aggregate demand could be diminished not increased in proportion to both production and saving – is generally regarded as following the work of Malthus; but the essence of the argument, in terms of his aggregate demand and employment-creating analysis of trade, is present in his 1808 pamphlet, and thus precedes Malthus's own concern with aggregate demand.

Selected works

1808. *An Inquiry into the Extent and Stability of National Resources*. Edinburgh: Oliphant & Brown.

1821–26. *The Christian and Civic Economy of Large Towns*. 3 vols. Glasgow: Chalmers & Collins.

1832. *On Political Economy, in Connexion with the Moral State and Moral Prospects of Society*. Glasgow: Collins.

Bibliography

Blaikie, W.G. 1887. Chalmers, Thomas. In *Dictionary of National Biography*, vol. 3. Oxford: Oxford University Press, 1973.

Bonar, J. 1894. Chalmers, Thomas. In *Palgrave's Dictionary of Political Economy*, ed. H. Higgs. London: Macmillan, 1925.

How to cite this article

O'Brien, D. P. "Chalmers, Thomas (1780–1847)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000099> doi:10.1057/9780230226203.0217

The New Palgrave Dictionary of Economics Online

Chamberlin, Edward Hastings (1899 – 1967)

Robert E. Kuenne
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chamberlin, E. H.; collusion; excess capacity; exploitation; firm, theory of; Harrod, R. F.; imperfect competition; industrial organization; large-group case; monopolistic competition; monopoly; oligopoly; product differentiation; Robinson, J. V.; Schlesinger, K.; specialization

Article

A major innovator in modern microeconomic theory, Chamberlin was born in La Conner, Washington, on 18 May 1899, and died in Cambridge, Massachusetts, on 16 July 1967. He received his Ph.D. from Harvard in 1927, became a full professor there in 1937, and occupied the David A. Wells chair from 1951 until his retirement in 1966. He edited the *Quarterly Journal of Economics* from 1948 to 1958.

Chamberlin's career exhibits a unity of professional purpose and thematic dedication over its more than 40-year length that is rare for modern theorists. Beginning with the start of his thesis research in 1925, its publication in 1933 as the seminal *Theory of Monopolistic Competition*, and continuing through eight editions, Chamberlin devoted his life to his vision of realistic market structures as mixtures of monopoly and competition.

He opposed the alternative polar frameworks of pure competition and monopoly of the 1920s as unrealistic; proselytized for his merger of them at the level of the firm in both broad and narrow contexts; strove tirelessly (and rather stridently) to distinguished his concepts from Joan Robinson's similar constructs; and manned the academic ramparts in full echelon against all who sought either to criticize the concepts or, alternatively, take credit for their genesis.

In so doing, Chamberlin's broad contributions to microeconomic analysis were of fundamental and insufficiently acknowledged importance. His ‘large group case’ and revival of interest in oligopoly theory created the notion of market structure as a continuum between pure competition and monopoly with location dictated by numbers of firms and product differentiation. With his work he fathered modern industrial organization analysis by giving a theoretical core to what was previously institutional and anecdotal. He reoriented the interest of microeconomics from the industry to the firm, revealing the latter's target variables to include selling cost and product variation as well as price. And his frameworks led economists to comprehend the importance of differentiated oligopoly in developed economies through his emphasis upon product differentiation, his formalization of monopoly power as control over price, and his perception of the core feature of oligopolistic market structure as perceived mutual interdependence of decision making.

Monopolistic competition theory

In its generic sense, which Chamberlin stressed increasingly in his later career, monopolistically competitive market structures are those in which the firm feels the external compulsions of competitive forces tempered in varying degrees by a monopolistic power to price its product. Central to monopolistic competition in this wider sense is *product differentiation*, or the ability of the firm to distinguish its product in the preferences of consumers, where product is defined to include a complex of qualities in addition to those inherent in the physical good (for example, location, repair services, ambience and so on). The existence of differentiation (*a*) implies the possibility of *selling costs*, or costs aimed at adapting demand to the product (advertising, catalogues, discounts, and so on) as distinguished from *production costs*, or expenditures that adapt the product to demand, and (*b*) *product variation*, or the variability of the complex of qualities and attributes that characterize the firm's output in the mind of the consumer.

In his original presentation of monopolistic competition and into the 1940s, Chamberlin tended to identify it more narrowly with a specific market structure that isolated product

differentiation as its distinctive component. This was the *large-group case* with the 'tangency solution' as the firm's long-run equilibrium position, as shown in Figure 1. Each firm produces a slightly differentiated product which may be closely approximated by competing firms. Hence, a large number of close substitutes ensure that the firm's demand curve is only slightly tilted from the pure competitor's horizontal position. If, for simplicity, all firms are assumed to have identical cost functions and to share sales equally (the *symmetry* assumption) then competition will reduce profit to zero by equating average cost and price at a tangency of the demand curve dd' and the average cost function *AC*. Where the tangency occurs marginal revenue *MR* will equal marginal cost *MC*. Hence, at price $p°$ and sales $x°$ each firm will be maximizing its profits at zero and neither entry into nor exit from the industry will occur: no internal or external force will exist to upset the long-run status quo.

Figure 1

The firm's optimal solution in the large-group case

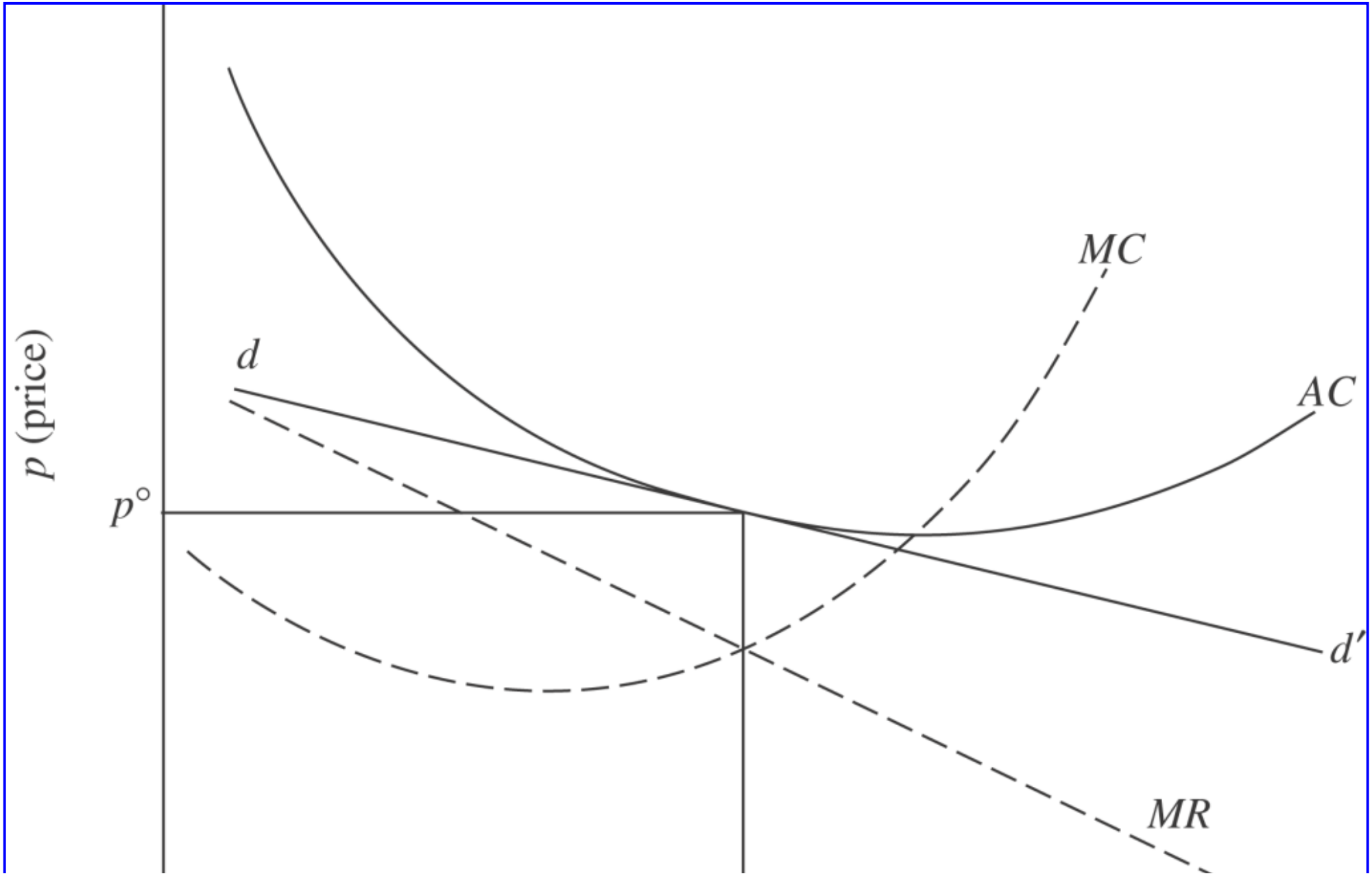

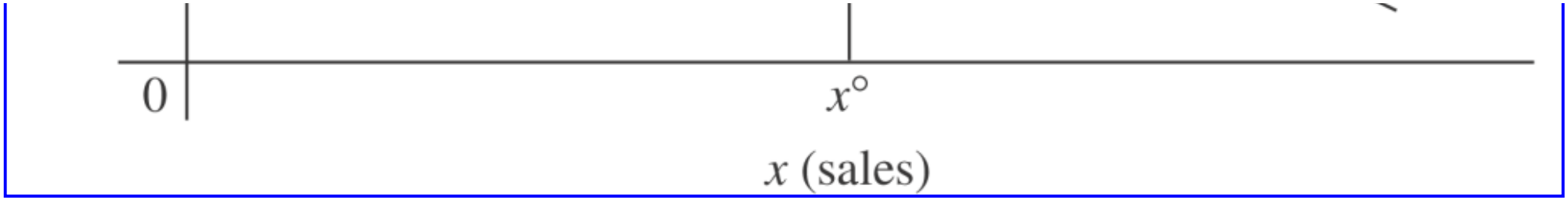

Despite Chamberlin's later disclaimers, there is little doubt that the large-group case was featured as the novel contribution of his theory, and it became identified with monopolistic competition theory. But from the beginning, Chamberlin did identify a second species in the generic theory: monopolistic competition caused by fewness of sellers of a homogeneous product. In the preface to the first edition of the *Theory* he included oligopoly in the concept of monopolistic competition. Oligopoly – he coined the word independently but later recognized its prior usage in 1914 by Karl Schlesinger – in the pure (that is, undifferentiated product) case formed the mirror image of the large-group case, with small rather than large numbers of sellers and undifferentiated rather than differentiated products. Surprisingly, given the centrality of product differentiation in his thought, he had little to say about differentiated oligopoly as a composite of the two purer cases of monopolistic competition – as late as 1948 the sixth edition of the *Theory* devoted only five pages to informal discussion of it – although he realized increasingly in his later work the prominent position it held in realistic market structures.

Chamberlin's contributions to the theory of pure oligopoly were noted above in listing his broader impacts on the field. More narrowly, they were not great advances. He ignored formal treatment of collusion and tended to urge that tacit collusion would lead to joint profit maximization for pure oligopoly and to a price solution intermediate between joint profit maximization and the large-group case for differentiated oligopoly. In his later, more informal, treatment of oligopoly, however, he asserted a general tendency toward ‘live-and-let-live’ limitations on oligopolistic rivalry.

But from the 1950s on, Chamberlin moved away from the large-group case as the featured form of monopolistic competition theory and shifted emphasis to oligopoly in its differentiated form. In part this was an aspect of his continuing desire to distance his theory from Joan Robinson's imperfect competition, in which she had independently developed the large-group case complete with tangency solution in the symmetry case. But, more importantly, the evolution of his thought reflected his increasing awareness that few market structures contained the uniform product competition implied by that solution. Rather, closer investigation of most realistic market structures with large numbers of sellers of slightly differentiated products revealed hierarchical clusters of oligopolistically competing firms. His book of essays (Chamberlin, 1957) reveals clearly his attempt to prevent monopolistic competition theory from being too closely identified with the large-group case.

Another aspect of this later effort was the playing down of his pioneering use of marginal revenue and marginal cost curves. In denying P.W.S. Andrews's assertion that full cost pricing was antithetical to monopolistic competition, Chamberlin asserted that it was integral to that body of analysis from the beginning, since profit maximization was never an exclusive motivation of the firm – as it was in Robinson's imperfect competition.

Other microeconomic contributions

An implication of the large-group equilibrium illustrated in Figure 1 is that firms would have long-run *excess capacity* in the sense that they would be operating at a production rate less than the rate associated with minimum average cost. This led to a dispute with Sir Roy Harrod, who seemed to believe that Chamberlin's results occurred because he was using short-run demand and cost curves in the large-group analysis. Harrod argued that businessmen would follow their long-run revenue and cost prospects and that excess capacity would not result. Chamberlin properly pointed out that his functions were long-run functions and that the long-run demand in Harrod's case did not attain the horizontality needed to eliminate excess capacity (Harrod, 1952, Essays 7, 8; Chamberlin, 1957, pp. 280–95; Kuenne, 1967, pp. 67–70). Later, Chamberlin argued that excess capacity also occurred in an industry when entrants flooded in irrationally even when profits disappeared (whose counter-argument was probably what Harrod had in mind) (Chamberlin, 1957, p. 290).

Chamberlin devoted a large portion of his writing to rationalizing the U-shaped average cost curve that was so fundamental to his market structures. Building upon the notion of the long-run average cost curve as the envelope of short-run average cost curves with fixed plants, he distinguished between using a fixed plant curve optimally in the short-run at its minimum-cost rate and producing a given rate of output optimally in the long-run by building an over-sized plant and using it at less than minimum cost capacity. Also, he denied that the rising portion of the long-run average cost curve was caused solely by management complexity or lumpy factors at higher output rates. In so doing, Chamberlin challenged the assertions of Knight (1921, pp. 98–9), Lerner (1944, pp. 165–7, 174–5), Stigler (1952, pp. 133, 202n.), and Kaldor (1934, p. 65n; 1935, p. 42) that, if all factors could be reduced to finely divisible units with (explicitly or implicitly assumed) constant efficiency, the average total cost would be constant as all product would be produced with optimal factor proportions. He argued that such factors would experience economies of scale as a function of factor-complex size owing to the ability to exploit specialization possibilities. These possibilities – $100 in capital might be concretized in ten shovels but $10,000 in capital might materialize as one back-hoe – permitted resource aggregates to become qualitatively different complexes with increased scale, rendering the notion of factor units with unchanged efficiency meaningless. The argument turns upon the semantics of constant efficiency units and the usefulness of the assumption, however, and was seen by most theorists to be non-illuminating and, as Chamberlin emphasized, tautological.

Two other contributions by Chamberlin are worthy of brief note. One was his destruction of Joan Robinson's notion of worker ‘exploitation’, because in non-purely competitive industries workers received marginal revenue product rather than marginal value product. Chamberlin demonstrated conclusively that the difference between the two was not received by any other factor, including the entrepreneur, but was experienced as an external revenue constraint by the firm. The second, quite different, contribution was Chamberlin's role as a founder of modern experimental market research by his publication of the results of mock market operations with his students.

The debate with Robinson

Chamberlin, like most microeconomic theorists of his generation, was thoroughly Marshallian in vision and methodology, and his innovations integrated neatly into the concerns of the post-Marshallian school. It was somewhat ironic, therefore, that Chamberlin found his major (and reluctant) opponent in Joan Robinson, as thoroughly Marshallian as himself. Chamberlin spent much of his professional life urging the fundamental divisions between his theory of monopolistic competition and Robinson's theory of imperfect competition.
The basis of the distinction changed fundamentally over his career. In the earlier objections, Chamberlin perceived correctly that Robinson's aim was to implement Sraffa's suggestion that microeconomic theory be rewritten in terms of a general theory of monopoly (Robinson, 1933, p. v). In so doing, he urged, Robinson failed to achieve the true blending of monopoly and competition that his theory achieved. Robinson evolved the large-group case in every detail, but passed quickly over it in pressing on to her larger goal of creating a general theory of ‘monopoly’ in industries with more than one firm. To Chamberlin, who in this early period stressed the large-group case, her emphasis upon near-homogeneous commodities with some differentiation of sellers in the consumers' minds slighted the competition among differentiated products and resulted in an analysis of industry ‘monopoly’, very close to the one-firm monopoly of standard theory.
There was some truth in this, although Chamberlin was ungenerous to Robinson in interpreting her achievements, for in addition to her large-group case development she paralleled him in isolating selling costs and in defining two types of imperfect markets: (*a*) firms which were not alike in customers' preferences, and (*b*) oligopoly. But she saw the threat to the existence of the ‘industry’ that non-homogeneous products posed, and her overall goal needed that solid Marshallian construct. Chamberlin from the beginning was willing to abandon the concept and speak of ‘product groups’.
However, as the large-group case came under criticism as incorporating too much of the purely competitive, and as oligopolistic structures received more attention in the literature, Chamberlin, as we have seen, shifted his ground and began to criticize Robinson for the opposite fault. The problem was, he now said, that imperfect competition failed to achieve the union of the competitive and the monopolistic because there was not enough monopoly content at the level of the firm. Implicitly, Robinson's large-group case was now focused upon for this fault, in comparison with his increasingly emphasized generic concepts that stressed oligopolistic elements.
The profession has ignored Chamberlin's strictures as distinctions without meaningful differences, and quite properly rewarded both theorists for their innovations. But the goals of the theorists were different, and, in most instances, Chamberlin's greater stress upon product differentiation and variation, selling cost and oligopoly proved to be more seminal in their professional impact.

Selected works

1933. *The Theory of Monopolistic Competition*. 6th edn., Cambridge, MA: Harvard University Press, 1948.

1954. (ed.) *Monopoly and Competition and Their Regulation*. London: Macmillan.

1957. *Towards a More General Theory of Value*. New York: Oxford University Press.

Bibliography

Harrod, R.F. 1952. *Economic Essays*. London: Macmillan.

Kaldor, N. 1934. The equilibrium of the firm. *Economic Journal* 44, March, 60–76.

Kaldor, N. 1935. Market imperfection and excess capacity. *Economica* 2, February, 33–50.

Knight, F.H. 1921. *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin.

Kuenne, R.E., ed. 1967. *Monopolistic Competition Theory: Studies in Impact*. New York: Wiley.

Lerner, A.P. 1944. *The Economics of Control*. New York: Macmillan.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Stigler, G.J. 1952. *The Theory of Price*. New York: Macmillan.

How to cite this article

Kuenne, Robert E. "Chamberlin, Edward Hastings (1899–1967)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000100> doi:10.1057/9780230226203.0218

The New Palgrave Dictionary of Economics Online

Champernowne, David Gawen (1912 – 2000)

F.A. Cowell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bayesian analysis; Champernowne, D. G.; fat-tailed distributions; income distribution; scaling laws; stochastic process models

Article

It was fortunate for the economics profession that the schoolboy Champernowne, a keen and able mathematician, was advised to read something in the school library to broaden his horizons: he chose Marshall's *Principles*.

David Champernowne was born on 9 July 1912 into an Oxford academic family. He was sent to school at Winchester and went from there as a scholar to King's College, Cambridge. While still an undergraduate he published his first paper (on ‘normal numbers’). Early contact with Dennis Robertson confirmed his previous interest in economics, and he was advised by J.M. Keynes to abandon his thoughts of becoming an actuary and switch to the Economics Tripos by taking his Part II Mathematics in one year rather the normal two. He obtained firsts throughout in both subjects.

His academic career spanned the London School of Economics (1936–8) Oxford (1945–59), and Cambridge (1938–40 and 1959–78). During the war period he served with Lindemann as Assistant in the Prime Minister's Statistical Section (1940–1) and worked with Jewkes at the Ministry of Aircraft Production's Department of Statistics and Programming.

He proved to be a genuine pioneer both in economic theory and statistics. His King's fellowship dissertation (submitted in 1936, but published 27 years later in the *Economic Journal*) laid the foundations for the application of stochastic process models to the analysis of income distributions; this work has been of importance in recent economic research on fat-tailed distributions and scaling laws. His pre-war interest in Frank Ramsey's theory of probability led on to work at Oxford on the application of Bayesian analysis to autoregressive series (at a time when the Bayesian approach was decidedly unfashionable), and culminated in his major trilogy on *Uncertainty and Estimation* (1969). However although he is thought of today primarily as a theoretician, his flashes of technical insight were always been tempered with healthy doses of practical scepticism. This is evident in his early work with

Beveridge on the regional and industrial distribution of employment and unemployment.
Champernowne acted as midwife to a number of major theoretical contributions over and above his own work. He provided an invaluable ‘translation’ to von Neumann's seminal paper on multisector growth. His role as behind-the-scenes expert at Cambridge over many theoretical issues is legendary: Joan Robinson acknowledged the assistance of his ‘heavy artillery’ in underpinning, and extending, her major work on capital and growth: A.C. Pigou's later writings on output and employment, Nicholas Kaldor's work on savings and economic growth models, and Dennis Robertson's *Principles* were all indebted to his intellectual influence.
He held Chairs at both Oxford and Cambridge, was director of the Oxford Institute of Statistics and was editor of the *Economic Journal*. He was elected Fellow of the British Academy in 1970.

Selected works

1935. A mathematical note on substitution. *Economic Journal* 15, 246–58.

1936. Unemployment, basic and monetary: the classical analysis and the Keynesian. *Review of Economic Studies* 3, 201–216.

1938. The uneven distribution of unemployment in the United Kingdom. 1929–36, I. *Review of Economic Studies* 5, 93–106.

1939. The uneven distribution of unemployment in the United Kingdom. 1929–36, II. *Review of Economic Studies* 6, 111–24.

1945. A note on J. von Neumann's article on ‘A Model of General Economic Equilibrium’. *Review of Economic Studies* 13, 10–18.

1948. Sampling theory applied to autoregressive sequences. *Journal of the Royal Statistical Society, Series B* 10, 204–42.

1952. The graduation of income distributions. *Econometrica* 20, 591–615.

1953. A model of income distribution. *Economic Journal* 63, 318–51.

1954. The production function and the theory of capital: a comment. *Review of Economic Studies* 21, 112–35.

1958. Capital accumulation and the maintenance of full employment. *Economic Journal* 68, 211–44.

1969. *Uncertainty and Estimation in Economics*. 3 vols. Edinburgh: Oliver and Boyd.

1971. The stability of Kaldor's 1957 model. *Review of Economic Studies* 38, 47–62.

1973. *The Distribution of Income Between Persons*. Cambridge: Cambridge University Press.

1974. A comparison of measures of income distribution. *Economic Journal* 84, 787–816.

1998. (With F. Cowall) *Economic Inequality and Income Distribution*. Cambridge University Press.

How to cite this article

Cowell, F. A. "Champernowne, David Gawen (1912–2000)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000101>
doi:10.1057/9780230226203.0219

The New Palgrave Dictionary of Economics Online

chaotic dynamics in economics

Jess Benhabib

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A new literature in the 1980s studied the possibility that endogenous cycles and irregular chaotic dynamics resembling stochastic fluctuations could be generated by deterministic, equilibrium models of the economy, in particular in overlapping generations models and in models with infinitely lived representative agents. Other empirical studies attempted to identify whether various economic time series were generated by deterministic chaotic dynamics or stochastic fluctuations. While dynamic equilibrium models calibrated to standard parameter values can generate chaotic dynamics and endogenous cycles even under intertemporal arbitrage and without market frictions, definitive empirical evidence for chaos in economics has not yet been produced.

Keywords

arbitrage; chaos; chaotic dynamics in economics; endogenous cycles; equilibrium cycles; ergodic chaos; intertemporal arbitrage; overlapping generations model

Article

When a new literature in the 1980s showed that endogenous cycles and chaos can arise in equilibrium models in economics, it came as a surprise. The possibility of deterministic fluctuations, as opposed to fluctuations driven by exogenous stochastic shocks, had been noted in an earlier literature on business cycles, for example in the well-known multiplier-accelerator models, but not in equilibrium models of the economy with complete markets and no frictions (see for example Frisch, 1933, or Samuelson, 1939). Yet deterministic fluctuations in equilibrium models with predictable relative price changes should be ruled out by intertemporal arbitrage. Such considerations led to the rejection of regular endogenous cycles in favour of models whose fluctuations are driven by stochastic shocks.
The new literature on chaotic dynamics showed that deterministic cycles and chaos were indeed possible under complete intertemporal arbitrage and without any market frictions, both in standard models of

overlapping generations and in calibrated models of infinitely lived representative agents (see for example Benhabib and Day, 1980; 1982; Benhabib and Nishimura, 1979; Grandmont, 1985; and Boldrin and Montrucchio, 1986). Of course, relative price fluctuations in such models had to be within the bounds allowed by the discount factor in order to be compatible with intertemporal arbitrage. (For an exploration of the relation between equilibrium cycles, chaos and discount rates in models with infinitely lived agents, see Benhabib and Rustichini, 1990; Sorger, 1992; Mitra, 1996; and Nishimura and Yano, 1996.) Furthermore, chaotic dynamics could exhibit not only deterministic endogenous cycles, but generate trajectories that are irregular, and that are statistically indistinguishable from stable linear stochastic AR1 processes (see Sakai and Tokumaru (1980)).

We can usually describe a dynamical system in discrete time as chaotic if it can generate cycles of every periodicity, where a sequence $\{x_j\}$ is of period n if $x_j = x_{j+n}$ but $x_j \neq x_i$, for $j < i < n - 1$. In addition, this simple definition of chaos requires the existence of an uncountable number of initial x which give rise to bounded but aperiodic (not even asymptotically) sequences. For example the well-known hump-shaped function, $4x(1 - x)$, when iterated, generates such chaotic dynamics. The kind of chaotic dynamics described above is usually referred to as ‘topological chaos’. If in addition we require that the set of initial conditions giving rise to aperiodic sequences are not simply uncountable but also have a positive (Lebesgue) measure, then we also have ergodic chaos. A useful sufficient condition to obtain topological chaos with a simple difference equation $x_{t+1} = f(x_t)$, with f continuous and mapping a closed interval into itself, is the existence of some x such that $f(f(f(x))) \leq x < f(x) < f(f(x))$. (See Li and Yorke, 1975; for simple sufficient conditions for chaos in higher dimensions, see Diamond, 1976, or Marotto, 2005.) Note that this condition will be satisfied if the difference equation has a solution of period three. A particularly interesting feature of some dynamic systems that are chaotic is their sensitive dependence on initial conditions: initial conditions that are arbitrarily close can generate sequences that tend to diverge over time. Thus, small measurement errors in initial conditions may cause large forecasting errors, which may explain some of the difficulties associated with business-cycle forecasting.

The aperiodic but bounded trajectories that characterize chaos and exhibit sensitive dependence on initial conditions cannot continue to diverge for ever. They converge not to a point or a periodic cycle but to a bounded chaotic or ‘strange’ attractor. The dynamical system which induces the local separation and instability of the trajectories must eventually bend them back. The combination of local stretching and global folding generates the complex nature of the dynamics. Such dynamic behaviour is in fact a familiar theme in economics that highlights the self-correcting nature of the economic system. Shortages create incentives for increased supply; dire necessities give rise to inventions as the invisible hand guides the allocation of resources. An equally familiar theme is that of instability: the multiplier interacts with the accelerator, leading to explosive or implosive investment expenditures; self-fulfilling expectations give rise to bubbles and crashes. In combination, these two themes suggest a nonlinear system, somewhat unstable at the core, but effectively contained further out. The contribution of the new literature on chaotic dynamics starting in the early 1980s has been to demonstrate the compatibility of endogenous irregular fluctuations with equilibrium dynamics in economics.

For a very simple example of chaotic dynamics, consider a simple overlapping generations model where each generation lives two periods. The utility function of a generation born at t is $U(c_0(t), c_1(t + 1))$,

where $c_0(t)$ is consumption when young and $c_1(t+1)$ is consumption when old. This generation faces a budget constraint $c_1(t+1) = w_1 + r(t)(w_0 - c_0(t))$, where w_0 is the endowment when young, w_1 is the endowment when old, and $r(t)$ is the rate of return on savings. The first order condition to the problem of maximizing utility subject to the budget constraint, on the assumption of interiority, yields $r(t) = \frac{U_1(c_0(t), c_1(t+1))}{U_2(c_0(t), c_1(t+1))}$. Here U_1 and U_2 denote the derivatives of the utility function U with respect to the first and second arguments. During each period t, market clearing requires that the sum of the endowments of the young and the old add up to the sum of their consumptions: $w_1 + w_0 = c_1(t) + c_0(0)$. Now consider the quadratic utility function $U(c, (t), c, (t+1)) = ac_0(t) - 0.5b(c_0(t))^2 + c_1(t)$, $0 \le c_0 \le a/b$, and $a, b > 0$. If we substitute the first order condition into the budget constraint, and use the market clearing condition, the difference equation describing the dynamics is given by $c_1(t+1) = ac_0(t)(1 - (b/a)c_0(t))$. Note that $c_0(t) \in (0, a/b)$ for all $c_0(0) \in (0, a/b)$, provided $a \le 4$. This difference equation will exhibit chaotic dynamics in c_0 for $a \in [3.53, 4]$, $b = a$. For example, if $a = 3.83$, the difference equation has a three-period cycle for $c_0(t) = 0.1561$, where $c_0(t+1) = 0.5096$ and $c_0(t+2) = 0.9579$. In this simple example utility saturates at $c_0 = a/b$, but the chaotic trajectories and those with a period greater than one never attain b/a, since if $c_0(t) = b/a$, $c(t+i) = 0$ for all $i = 1, 2, \ldots$ Another simple example of an exponential utility function that will generate chaotic dynamics in this simple overlapping generations model, for $a > 2.692$ and $w_1 > e^{a-1}$, is $U(c, (t), c, (t+1)) = A - e^{a + w_0 - c_0(t)} + c_1(t)$. (See Benhabib and Day, 1982, s. 3.4.)

Techniques to empirically distinguish between data generated by non-chaotic stochastic systems and deterministic chaotic systems have been developed by physicists and mathematicians (see for example Eckmann and Ruelle, 1985). These techniques have been further refined into statistical tests for applications to economic data by Brock (1986) and Brock et al. (1996), among others. Very roughly, these methods exploit the idea that deterministic systems will generate trajectories that are of lower dimension than those generated by stochastic systems, which have more scattered trajectories. For example, if we consider a one-dimensional difference equation that generates chaotic dynamics, say $x_{t+1} = 4x_t(1 - x_t)$ for initial $x_0 \in (0, 1)$, plotting x_{t+1} against x_t will yield a curve. By contrast, if the dynamics were generated by a linear or nonlinear stochastic system with noise, the same plot would produce a scatter of points, which could not be captured by a ‘relatively smooth’, one-dimensional line. By formalizing this idea, we may attempt to distinguish data generated by deterministic chaotic systems and by non-chaotic stochastic systems, even without explicit knowledge of the underlying economic system generating the data. In general, however, such a method is hard to apply because, unlike data generated by scientific experiments, economic time series are often not long enough. If the order of underlying dynamical system generating the data is high-dimensional, say of the order of five or higher, or alternatively if we can only observe the realizations of a subset of the variables of the underlying economic model, distinguishing between stochastically and chaotically generated data becomes very difficult. The difficulty of empirically identifying chaos in high dimensional economic systems may be particularly important if chaotic dynamics is more likely to be manifested in disaggregated sectoral or industry data whose components, because of resource constraints or other scarcities, can move in ways that partially offset one another's cyclic or irregular movements. It would therefore be fair to say that at

this point, while we know that standard dynamic equilibrium models with parameters calibrated to values often used in the literature may well generate chaotic dynamics, more definitive empirical evidence for chaos in economics has not yet been produced.

While it may be instructive to set the theories of endogenous economic fluctuations in opposition to the theories of fluctuations driven by stochastic shocks, in practice it is more helpful to consider endogenously oscillatory dynamics as complementary to stochastic fluctuations. In certain environments it may make little difference if endogenous mechanisms by themselves generate regular and irregular persistent fluctuations, or whether they give rise to damped oscillations that are sustained by stochastic shocks. On the other hand, if the underlying equilibrium system is subject to distortions and there is room for stabilization policy, correctly identifying the source of the fluctuations becomes much more important. (See for example Benhabib, Schmitt-Grohe and Uribe, 2002). Furthermore, recognizing the role of oscillatory dynamics may diminish our reliance on unrealistically large shocks to explain economic data, for example, in real business cycle theory.

See Also

- economy as a complex system

Bibliography

Benhabib, J. and Day, R. 1980. Erratic accumulation. *Economic Letters* 6, 113–17.

Benhabib, J. and Day, R. 1982. A characterization of erratic dynamics in the overlapping generations model. *Journal of Economic Dynamics and Control* 4, 37–55.

Benhabib, J. and Nishimura, K. 1979. The Hopf bifurcation and the existence and stability of closed orbits in multisector models of optimal economic growth. *Journal of Economic Theory* 21, 421–44.

Benhabib, J. and Rustichini, A. 1990. Equilibrium cycling with small discounting. *Journal of Economic Theory* 52, 423–32.

Benhabib, J., Schmitt-Grohe, S. and Uribe, M. 2002. Chaotic interest rate rules. *American Economic Review Papers and Proceedings* 92, 72–8.

Boldrin, M. and Montrucchio, L. 1986. On the indeterminacy of capital accumulation paths. *Journal of Economic Theory* 40, 24–39.

Brock, W. 1986. Distinguishing random and deterministic systems. *Journal of Economic Theory* 40, 68–195.

Brock, W., Dechert, W., LeBaron, B. and Scheinkman, J. 1996. A test for independence based upon the

correlation dimension. *Econometric Reviews* 15, 197–235.

Diamond, P. 1976. Chaotic behavior of systems of difference equations. *International Journal of Systems Science* 7, 953–56.

Eckmann, J. -P. and Ruelle, D. 1985. Ergodic theory of chaos and strange attractors. *Review of Modern Physics* 57, 617–56.

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic Essays in Honor of Gustav Cassel*. London: Allen and Unwin. Reprinted in *Readings in Business Cycles*, ed. R. Gordon and L. Klein. Homewood, IL: Richard D. Irwin, 1965.

Grandmont, J. 1985. On endogenous competitive business cycles. *Econometrica* 53, 995–1045.

Li, T. and Yorke, J. 1975. Period three implies chaos. *American Mathematical Monthly* 82, 985–92.

Marotto, F. 2005. On redefining a snap-back repeller. *Chaos, Solitons and Fractals* 25, 25–28.

Mitra, T. 1996. An exact discount factor restriction for period three cycles in dynamic optimization models. *Journal of Economic Theory* 69, 281–305.

Nishimura, K. and Yano, M. 1996. On the least upper bound of discount factors that are compatible with optimal period-three cycles. *Journal of Economic Theory* 69, 306–33.

Sakai, H. and Tokumaru, H. 1980. Autocorrelations of a certain chaos. *IEEE Transactions in Acoustic Speech Signal Process* 28, 588–90.

Samuelson, P. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economic Statistics* 21, 75–8.

Sarkovskii, A. 1964. Coexistence of cycles of a continuous map of the line into itself. *Ukrains'kyi Matematychnyi Zhurnal* 16, 61–71.

Sorger, G. 1992. On the minimum rate of impatience for complicated optimal growth paths. *Journal of Economic Theory* 56, 160–79.

How to cite this article

Benhabib, Jess. "chaotic dynamics in economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_C000104> doi:10.1057/9780230226203.0220

The New Palgrave Dictionary of Economics Online

charitable giving

James Andreoni
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Charitable giving is a significant vehicle for providing needed goods and services around the world. In the United States, for example, charitable giving accounts for nearly two per cent of income. Moreover, the tax deduction for charitable giving is one of the oldest and most widely used tax policies in the US tax code. This article describes the known facts on charitable giving, how and why people give, and discusses the impacts of government policies on giving.

Keywords

altruism; charitable giving; charitable organizations; crowding out; estate tax; free rider problem; fund-raising; permanent income hypothesis; philanthropy; public goods; self-interest; tax deductibility; two-stage least squares; warm glow

Article

In 2005 charitable giving in the United States totalled over 260 billon dollars, or around 1.9 per cent of personal income, making it a significant fraction of the economy. Individual giving accounted for 77 per cent of this total, while foundations accounted for 12 per cent, bequests for 7 per cent, and corporations for 5 per cent (Giving USA, 2006). Almost 70 per cent of US households report giving to charity. While the United States typically has one of the largest and most extensively studied charitable sectors, other countries around the world also have significant philanthropy (Andreoni, 2001; 2006).

There are three sets of actors in markets for charitable giving, and understanding each and their relationships to each other is essential to an understanding of charity. The first set is the donors who supply the dollars and volunteer hours to charities. The second is the charitable organizations, that is, the demand side of the market. They organize donors with fund-raising strategies, and produce the charitable goods and services with the money and time donated. The third player is the government. Governments are involved in charities in a number of ways. In many countries, including the United

States, individual taxpayers may be able to deduct charitable donations from their taxable income. Governments also give directly to charities in the form of grants.
The following highlights the most important and fundamental aspects of research on charitable giving.

What motivates giving?

Why would a self-interested agent give away a considerable fraction of his income, often for the benefit of complete strangers? Obviously, acting unselfishly must be in his self-interest. One model of this is that the public benefits of the charity enter directly into a giver's utility function, that is, charity is a privately provided public good. This approach is advanced by Warr (1982) and Roberts (1984), who show theoretically that, if giving is a pure public good, then we would predict that government grants to charities will perfectly crowd out private donations, meaning government spending is largely ineffective. Bergstrom, Blume and Varian (1986) develop this model further to provide a series of elegant derivations, including the (unrealistic) prediction that redistributions of income will be ‘undone’ if everyone gives to a public good. Andreoni (1988) pushes this model to its natural limits and shows that in large economies we would predict a vanishingly small fraction of people who will give to a public good, which is clearly contradicted by the statistics presented above.
For this reason, economists have felt more comfortable assuming that, in addition to caring about the total supply of charity, what could be called pure altruism, people also experience some direct private utility from the act of giving. While there are numerous models and justifications for such an assumption, they have often been gathered under the general (and slightly pejorative) term, the ‘warm glow’ of giving (Andreoni, 1989; 1990). In large economies, in fact, it is easy to show that this motive must dominate at the margin (Ribar and Wilhelm, 2002). The intuition is clear. If large numbers of others are collectively providing a substantial amount of charity, the incentive to free ride must be so overwhelming that the only remaining justification for giving is that there is some direct benefit to the act of giving.
The consequence of assuming a warm-glow motive is that we can treat individual donations as having the properties of a private good. When income is higher or when the price of giving is lower, we predict that individuals will give more.

What is the impact of the tax deduction for charitable giving?

Studies of the charitable deduction are aimed at understanding just how individual giving is responsive to changes in income and price. If t is the marginal tax rate faced by a giver, and if (in the United States) the person itemizes deductions, then the charitable deduction makes the effective price of a dollar of donations $1 - t$. The policy questions are how responsive is giving to the price, and is the policy successful in promoting additional giving.
Let g be the giving of the household. If the policy is effective, then the new giving received by the charity should exceed the lost revenue of the government, that is, total spending on giving will rise with the deduction. This means $d(1-t)g/dt>0$, which holds if $\varepsilon=[dg/d(1-t)]/[(1-t)/g]<-1$. This means that the policy is effective if giving is price-elastic, $\varepsilon<-1$. Since the first studies on giving (Feldstein and Clotfelter, 1976), researchers have debated whether this ‘gold standard’ has been met.

Dozens of studies of this question have been undertaken. Most employ cross-sectional data, either from surveys about giving or from tax returns. Each of these data sources has advantages and weaknesses, and each presents special challenges for identification and estimation (see Triest, 1998, for a careful discussion). These studies are summarized by Clotfelter (1985), Steinberg (1990), and Andreoni (2006). Prior to 1995, a consensus had formed that the income elasticity was below 1, typically in the range of 0.4 to 0.8, and that the price elasticity was below minus 1, generally in the range minus 1.1 to minus 1.3, thus meeting the gold standard. Only a few studies found giving was price-inelastic.

This consensus was upset by an important study of Randolph (1995). There are two important features of his analysis. First, he uses a panel tax returns rather than a cross section. Second, the period of his sample, 1979–89, spans two tax reforms. These reforms provide independent variation in price that can be helpful in identifying elasticities. Moreover, his instrumental variables analysis allows him to separate short-run and long-run elasticities. Contrary to the prior literature, he estimates a long-run price elasticity of only minus 0.51, meaning that the policy no longer satisfies the gold standard. Short-run elasticities, by contrast, are high, at minus 1.55. This means that givers are sophisticated at substituting giving from years of low marginal tax rates to years with high marginal tax rates. His analysis suggests that cross-sectional studies conflate short- and long-run elasticities and thus mislead policy analysts.

Auten, Sieg and Clotfelter (2002) challenged Randolph's results. They use a similar (although longer) panel of tax payers, but employ a different estimation technique. Their analysis capitalizes on restrictions placed on the covariance matrices of income and price by assumptions of the permanent income hypothesis. Their analysis again returns estimates to the consensus values, with a permanent price elasticity of minus 1.26. The sensitivity of the estimates to the estimation technique and the identification strategy has left the literature unsettled as to the true values of price and income elasticities.

Giving by the very wealthy

Most of the data available, for reasons of confidentiality, exclude the very wealthy. Yet, the richest 400 US tax filers in the year 2000 accounted for about seven per cent of all individual giving in that year. Auten, Clotfelter and Schmalbeck (2000) provide a fascinating analysis of wealthy givers drawn from income tax filings at the Internal Revenue Service. Among the most interesting findings is that giving as a percentage of income rises only modestly with income, up to about four per cent for those earning over 2.5 million dollars. However, the variance in giving rises sharply. The inference is that wealthy givers are ‘saving up’ for larger gifts. These larger gifts may allow them to exert some control over the charity, such as providing a seat on the board of directors, or may garner a monument, such as naming a university building after the donor.

In discussing the wealthy, one must also address the effects of the estate tax on giving. Bakija, Gale and Slemrod (2003) use 39 years' worth of federal estate tax filings to study the sensitivity of estate giving to the estate tax. They rely on variation in estate tax rates across states for identification and find that charitable giving from estates is extremely sensitive to the tax. They measure the price elasticity of estate giving to be around minus 2.0, while the ‘wealth elasticity’ is about 1.5. This indicates that the 2001 changes in US estate tax laws, which greatly reduce (and eventually eliminate) estate tax rates, can have huge impacts on giving.

Do government grants crowd out individual giving?

There are many studies on crowding out, and most show that crowding is quite small, often near zero, and sometime even negative (Kingma, 1989; Okten and Weisbrod, 2000; Khanna, Posnett and Sandler, 1995; Manzoor and Straub, 2005; and Hungerman, 2005). Payne (1998), however, noted that the government officials who approve the grants are elected by the same people who make donations to charities. Hence, positive feelings toward a charity will be represented in the preferences of both givers and the government. This positive relationship between public and private donations means that some of the prior estimates could be biased against finding crowding out.

Payne (1998) turns to two-stage least squares analysis to address this endogeneity. As an instrument for government grants she uses aggregate government transfers to individuals in the state, and finds that estimates of crowding out rise to around 50 per cent, which is significantly above the zero per cent crowing that comes when she applies prior techniques to her data. This is a significant new finding. None of this analysis, however, has accounted for the fact that government grants may also have an impact on the fund-raising of charities. Andreoni and Payne (2003) ask what happens to a charity's fund-raising expenses when it gets a government grant. Does it fall, and by how much? They look at 14-year panel charitable organizations and find there are significant reductions in fund-raising efforts by charities after receiving government grants. This raises the possibility, therefore, that grants crowd out fund-raising, which then indirectly reduces giving, and that this may be the actual channel through which ‘crowding out’ occurs.

Incorporating fund-raising into research on charitable giving

One of the exciting new challenges for research on charitable giving is accounting for the strategic actions of charities in the analysis. This typically means understanding how charities choose fund-raising strategies, and how givers respond. A theoretical literature has emerged to provide a framework for analysing fund-raising (see Andreoni, 2006, for a review). At the same time researchers have begun considering field and laboratory experiments on charitable giving. These studies look at the effectiveness of ideas proposed by the theoretical literature, and evaluate some of the standard practices of charities.

Rege and Telle (2004) and Andreoni and Petrie (2004) show in laboratory studies that the common practice of revealing the identities of givers, and reporting amounts given in categories (Harbaugh, 1998), can have positive impacts on donations. Soetevent (2005) shows similar social effects in a field experiment.

List and Lucking-Reiley (2002) use a field experiment to establish that when charities require a minimum amount of contributions before a new initiate can be pursued, having a ‘seed grant’ can be greatly effective (Andreoni, 1998), as can be guarantees of refunds in the event that the threshold of donations is not met (Bagnoli and Lipman, 1989).

Landry et al. (2006), explore the use of lotteries in raising money for charities (Morgan, 2000) in an actual door-to-door fundraising campaign. They find that lotteries increase giving, as expected. Perhaps surprisingly, however, they find that the physical attractiveness of the fundraiser has a significant affect

on the amounts raised, and that this was at least as important as any economic incentives offered.

Conclusion

Charitable giving has been one of the perennial topics for economists. It presents challenges for the theorists to understand the motives and institutions for giving, for policy analysts to measure and identify the effects of price and income, and for experimenters to explore innovations in the market for giving. As governments become increasingly reliant on private organizations to provide public services, and as charities become increasingly sophisticated at raising money and delivering needed services, understanding the relationships among the suppliers and demanders of charity will become essential for calculating the social costs and benefits of charitable institutions.

See Also

- altruism in experiments
- altruism, history of the concept
- crowding out
- externalities
- public finance
- public goods
- tax expenditures

Bibliography

Andreoni, J. 1988. Privately provided public-goods in a large economy the limits of altruism. *Journal of Public economics* 83, 57–73.

Andreoni, J. 1989. Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy* 97, 1447–58.

Andreoni, J. 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. *Economic Journal* 100, 464–77.

Andreoni, J. 1998. Toward a theory of charitable fund-raising. *Journal of Political Economy* 106, 1186–213.

Andreoni, J. 2001. The economics of philanthropy. In *International Encyclopedia of Social and Behavioral Sciences*, ed. N. Smeltser and P. Baltes. Oxford: Elsevier.

Andreoni, J. 2006. Philanthropy. In *Handbook of Giving, Reciprocity, and Altruism*, ed. S.-C. Kolm and J. Mercier Ythier. Amsterdam: North-Holland.

Andreoni, J. and Payne, A.A. 2003. Do government grants to private charities crowd out giving or fundraising? *American Economic Review* 93, 792–812.

Andreoni, J. and Petrie, R. 2004. Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of Public Economics* 88, 1605–23.

Auten, G.E., Clotfelter, C.T. and Schmalbeck, R.L. 2000. Taxes and philanthropy among the wealthy. In *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, ed. J.B. Slemrod. New York: Russell Sage.

Auten, G., Sieg, H. and Clotfelter, C.T. 2002. Charitable giving, income, and taxes: an analysis of panel data. *American Economic Review* 92, 371–82.

Bagnoli, M. and Lipman, B.L. 1989. Provision of public goods: fully implementing the core through private contributions. *Review of Economic Studies* 56, 583–601.

Bakija, J.M., Gale, W.G. and Slemrod, J.B. 2003. Charitable bequests and taxes on inheritances and estates: aggregate evidence from across states and time. *American Economic Review* 93, 366–70.

Bergstrom, T.C., Blume, L.E. and Varian, H.R. 1986. On the private provision of public goods. *Journal of Public Economics* 29, 25–49.

Clotfelter, C.T. 1985. *Federal Tax Policy and Charitable Giving*. Chicago: University of Chicago Press.

Feldstein, M. and Clotfelter, C.T. 1976. Tax incentives and charitable contributions in the United States: a microeconometric analysis. *Journal of Public Economics* 5, 1–26.

Giving USA. 2006. *The Annual Report on Philanthropy for the Year 2005*. New York: AAFRC Trust for Philanthropy.

Harbaugh, W.T. 1998. What do donations buy? A model of philanthropy based on prestige and warm glow. *Journal of Public Economics* 67, 269–84.

Hungerman, D.M. 2005. Are church and state substitutes? Evidence from the 1996 welfare reform. *Journal of Public Economics* 89, 2245–67.

Khanna, J., Posnett, J. and Sandler, T. 1995. Charity donations in the UK: new evidence based on panel data. *Journal of Public Economics* 56, 257–72.

Kingma, B.R. 1989. An accurate measurement of the crowd-out effect, income effect, and price effect

for charitable contributions. *Journal of Political Economy* 97, 1197–207.

Landry, C., Lange, A., List, J.A., Price, M.K. and Rupp, N.G. 2006. Toward an understanding of the economics of charity: evidence from a field experiment. *Quarterly Journal of Economics* 121, 747–82.

List, J.A. and Lucking-Reiley, D. 2002. The effects of seed money and refunds on charitable giving: experimental evidence from a university capital campaign. *Journal of Political Economy* 110, 215–33.

Manzoor, S. and Straub, J. 2005. The robustness of Kingma's crowd-out estimate: evidence from new data on contributions to public radio. *Public Choice* 123, 463–76.

Morgan, J. 2000. Financing public goods by means of lotteries. *Review of Economic Studies* 67, 761–84.

Okten, C. and Weisbrod, B.A. 2000. Determinants of donations markets. *Journal of Public Economics* 75, 255–72.

Payne, A.A. 1998. Does the government crowd-out private donations? New evidence from a sample of non-profit firms. *Journal of Public Economics* 69, 323–45.

Randolph, W.C. 1995. Dynamic income, progressive taxes, and the timing of charitable contributions. *Journal of Political Economy* 103, 709–38.

Rege, M. and Telle, K. 2004. The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics* 88, 1625–44.

Ribar, D.C. and Wilhelm, M.O. 2002. Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy* 110, 425–57.

Roberts, R.D. 1984. A positive model of private charity and public transfers. *Journal of Political Economy* 92, 136–48.

Soetevent, A.R. 2005. Anonymity in giving in a natural context: a field experiment in 30 churches. *Journal of Public Economics* 89, 2301–23.

Steinberg, R. 1990. Taxes and giving: new findings. *Voluntas* 1, 61–79.

Triest, R.K. 1998. Econometric issues in estimating the behavioral response to taxation: a nontechnical introduction. *National Tax Journal* 51, 761–72.

Warr, P.G. 1982. Pareto optimal redistribution and private charity. *Journal of Public Economics* 19, 131–8.

How to cite this article

Andreoni, James. "charitable giving." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000590> doi:10.1057/9780230226203.0221

The New Palgrave Dictionary of Economics Online

cheap talk

Vijay Krishna and John Morgan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Cheap-talk models address the question of how much information can be credibly transmitted when communication is direct and costless. When a single informed expert, who is biased, gives advice to a decision maker, only noisy information can be credibly transmitted. The more biased the expert is, the noisier the information. The decision maker can improve information transmission by: (*a*) more extensive communication, (*b*) soliciting advice from additional experts, or (*c*) writing contracts with the expert.

Keywords

cheap talk; communication equilibria; delegation principle; games with incomplete information; incentive contracts; revelation principle; signalling

Article

In the context of games of incomplete information, the term ‘cheap talk’ refers to direct and costless communication among players. Cheap-talk models should be contrasted with more standard signalling models. In the latter, informed agents communicate private information indirectly via their choices – concerning, say, levels of education attained – and these choices are costly. Indeed, signalling is credible precisely because choices are differentially costly – for instance, high-productivity workers may distinguish themselves from low-productivity workers by acquiring levels of education that would be too costly for the latter.

The central question addressed in cheap-talk models is the following. How much information, if any, can be credibly transmitted when communication is direct and costless? Interest in this question stems from the fact that with cheap talk there is always a ‘babbling’ equilibrium in which the participants deem all communication to be meaningless – after all, it has no direct payoff consequences – and as a result no one has any incentive to communicate anything meaningful. It is then natural to ask whether there are also equilibria in which communication is meaningful and informative.

We begin by examining the question posed above in the simplest possible setting: there is a single informed party – an expert – who offers information to a single uninformed decision maker. This simple model forms the basis of much work on cheap talk and was introduced in a now classic paper by Crawford and Sobel (1982). In what follows, we first outline the main finding of this paper, namely, that while there are informative equilibria, these entail a significant loss of information. We then examine various remedies that have been proposed to solve (or at least alleviate) the ‘information problem’.

The information problem

We begin by considering the leading case in the model of Crawford and Sobel (henceforth CS). A decision maker must choose some decision *y*. Her payoff depends on *y* and on an unknown state of the world θ, which is distributed uniformly on the unit interval. The decision maker can base her decision on the costless message *m* sent by an expert who knows the precise value of θ. The decision maker's payoff is $U(y, \theta) = -(y-\theta)^2$, and the expert's payoff is $V(y, \theta, b) = -(y-(\theta+b))^2$, where $b \geq 0$ is a ‘bias’ parameter that measures how closely aligned the preferences of the two are. Because of the tractability of the ‘uniform-quadratic’ specification, this paper, and indeed much of the cheap talk literature, restricts attention to this case.

The sequence of play is as follows:

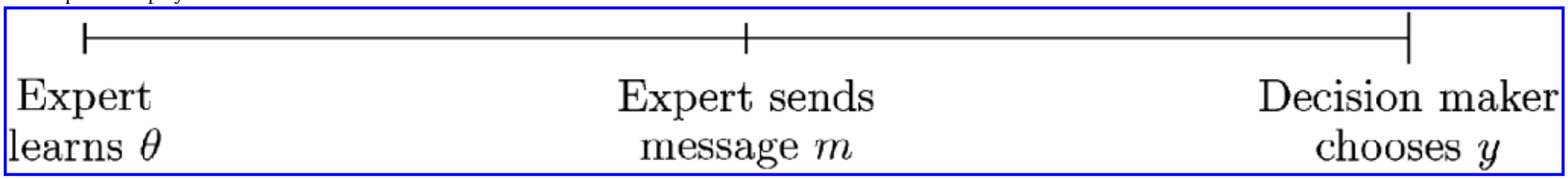

What can be said about (Bayesian-perfect) equilibria of this game? As noted above, there is always an equilibrium in which no information is conveyed, even in the case where preferences are perfectly aligned (that is, $b = 0$). In such a 'babbling' equilibrium, the decision maker believes (correctly it turns out) that there is no information content in the expert's message and hence chooses her decision only on the basis of her prior information. Given this, the expert has no incentive to convey any information – he may as well send random, uninformative messages – and hence the expert indeed 'babbles'. This reasoning is independent of any of the details of the model other than the fact that the expert's message is 'cheap talk'.

Are there equilibria in which all information is conveyed? When there is any misalignment of preferences, the answer turns out to be no. Specifically,

Proposition 1: If the expert is even slightly biased, all equilibria entail some information loss.

The proposition follows from the fact that, if the expert's message always revealed the true state and the decision maker believed him, then the expert would have the incentive to exaggerate the state – in some states θ, he would report $\theta + b$.

Are there equilibria in which some but not all information is shared? Suppose that, following message *m*, the decision maker holds posterior beliefs given by distribution function *G*. The action *y* is chosen to maximize her payoffs given *G*. Because payoffs are quadratic, this amounts to choosing a *y* satisfying:

$$y(m) = E[\theta|m] \tag{1}$$

Suppose that the expert faces a choice between sending a message *m* that induces action *y* or an alternative message, m', that induces an action $y' > y$. Suppose further that in state θ' the expert prefers y' to *y* and vice versa in state $\theta < \theta'$. Since the preferences satisfy the *single-crossing* condition, $V_{y\theta} > 0$, the expert would prefer y' to *y* in all states higher than θ'. This implies that there is a unique state *a*, satisfying $\theta < a < \theta'$, in which the expert is indifferent between the two actions. Equivalently, the distance between *y* and the expert's 'bliss' (ideal) action in state *a* is equal to the distance between action y' and the expert's bliss action in state *a*. Hence,

$$a + b - y = y' - (a + b) \tag{2}$$

Thus, message *m* is sent for all states $\theta < a$ and message m' for all states $\theta > a$.

To comprise an equilibrium where exactly two actions are induced, one would need to find values for *a*, *y*, and y' that simultaneously satisfy eqs. (1) and (2). Since *m* is sent in all states $\theta < a$, from eq. (1), $y = \frac{a}{2}$, Similarly, $y' = \frac{1+a}{2}$. Inserting these expression into eq. (2) yields

$$a = \frac{1}{2} - 2b \tag{3}$$

Equation (3) has several interesting properties. First, notice that *a* is uniquely determined for a given bias. Second, notice that, when the bias gets large $(b \geq \frac{1}{4})$, there is no feasible value of *a*, so no information is conveyed in any equilibrium. Finally, notice that, when the expert is unbiased $(b = 0)$, there exists an equilibrium where the state space is equally divided into 'high' $(\theta > \frac{1}{2})$ and 'low' $(\theta < \frac{1}{2})$ regions and the optimal actions respond accordingly. As the bias increases, the low region shrinks in size while the high region grows; thus, the higher the bias is, the less the information conveyed.

For all $b < \frac{1}{4}$, we constructed an equilibrium that partitions the state space into two intervals. As the bias decreases, equilibria exist that partition the state space into more than two intervals. Indeed, Crawford and Sobel (1982) showed that:

Proposition 2: All equilibria partition the state space into a finite number of intervals. The information conveyed in the most informative equilibrium is decreasing in the bias of the expert.

If the expert were able to commit to fully reveal what he knows, *both* parties would be better off than in any equilibrium of the game described above. With full revelation, the decision maker would choose $y = \theta$ and earn a payoff of zero, while the expert would earn a payoff of $-b^2$. It is easily verified that in any equilibrium the payoffs of both parties are lower than this. The overall message of the CS model is that, absent any commitment possibilities, cheap talk inevitably leads to information loss, which is increasing in the bias of the expert. The remainder of the article studies various 'remedies' for the information loss problem: more extensive communication, delegation, contracts, and multiple experts.

Remedies

Extensive communication

In the CS model, the form of the communication between the two parties was one-sided – the expert simply offered a report to the decision maker, who then acted on it. Of course, communication can be much richer than this, and it is natural to ask whether its form affects information transmission. One might think that it would not. First, one-sided communication where the expert speaks two or more times is no better than having him speak once, since any information the expert might convey in many messages can be encoded in a single message. Now, suppose the communication is two-sided – it is a conversation – so the decision maker also speaks. Since she has no information of her own to contribute, all she can do is to send random messages, and at first glance this seems to add little. As we will show, however, random messages improve information transmission by acting as *coordinating devices*.

To see this, suppose the expert has bias $b = \frac{1}{12}$. As we previously showed, when only he speaks, the best equilibrium is where the expert reveals whether the state is above or below $\frac{1}{3}$. Suppose instead that we allow for *face-to-face* conversation – a simultaneous exchange of messages – and that the sequence of play is:

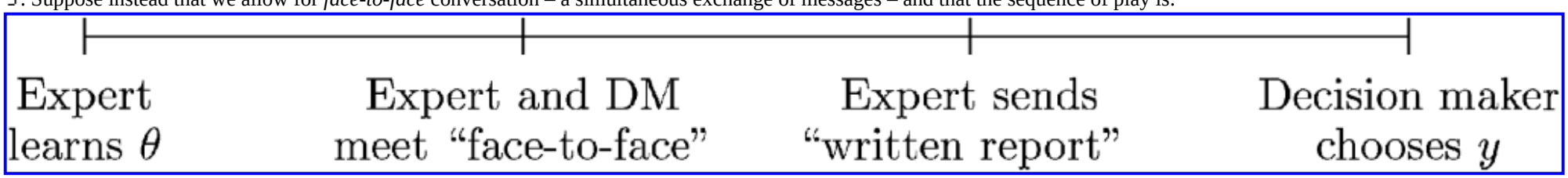

>The following strategies constitute an equilibrium. The expert reveals some information at the face-to-face meeting, but there is also some randomness in what transpires. Depending on how the conversation goes, the meeting is deemed by both parties to be a 'success' or a 'failure'. After the meeting, and depending on its outcome, the expert may send an additional 'written report' to the decision maker.

During the meeting, the expert reveals whether θ is above or below $\frac{1}{6}$; he also sends some additional messages that affect the success or failure of the meeting. If he reveals that $\theta \leq \frac{1}{6}$, the meeting is adjourned, no more communication takes place, and the decision maker chooses a low action $y_L = \frac{1}{12}$ that is optimal given the information that $\theta \leq \frac{1}{6}$.

If, however, he reveals that $\theta > \frac{1}{6}$, then the written report depends on whether the meeting was a success or a failure. If the meeting is a failure, no more communication takes place, and the decision maker chooses the 'pooling' action $y_P = \frac{7}{12}$ that is optimal given that $\theta > \frac{1}{6}$. If the meeting is a success, however, the written report further divides the interval

$[\frac{1}{6}, 1]$ into $[\frac{1}{6}, \frac{5}{12}]$ and $[\frac{5}{12}, 1]$. In the first sub-interval, the medium action $y_M = \frac{7}{24}$ is taken and in the second sub-interval the high action $y_H = \frac{17}{24}$ is taken. The actions taken in different states are depicted in Figure 1. The dotted line depicts the actions, $\theta + \frac{1}{12}$, that are 'ideal' for the expert.

Figure 1
Equilibrium with face-to-face meeting

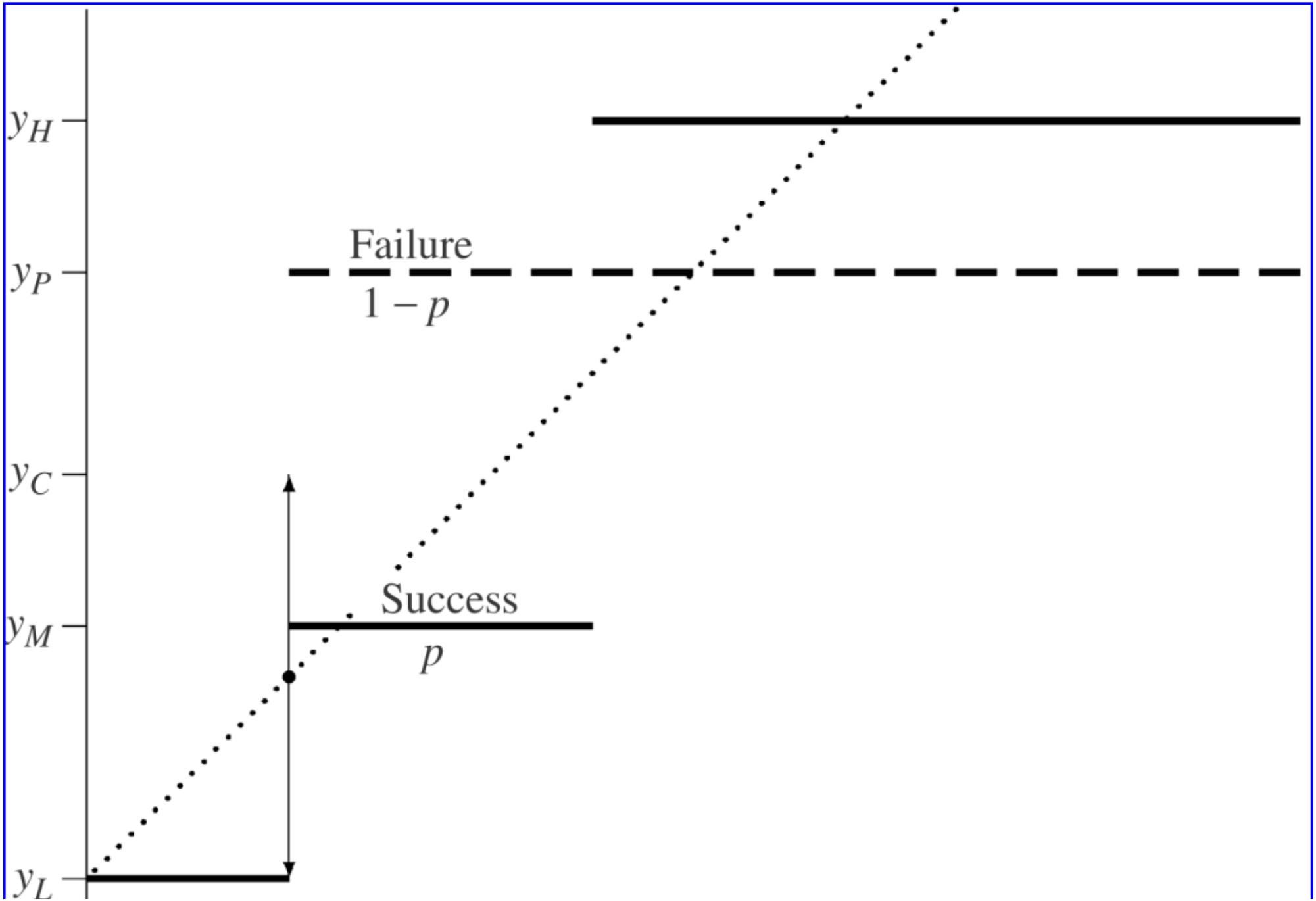

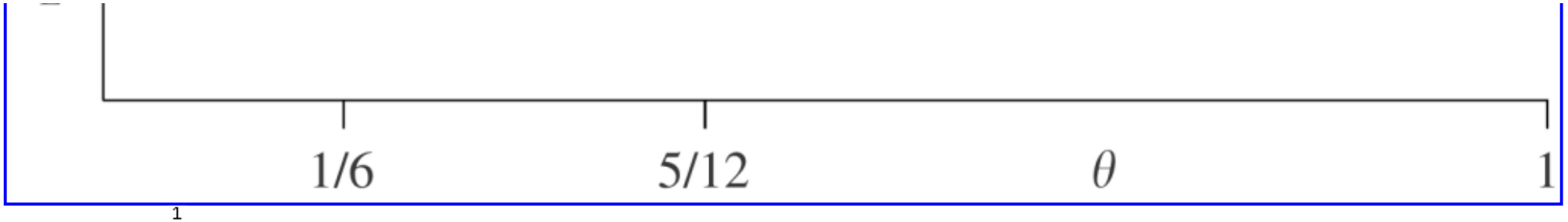

Notice that in state $\frac{1}{6}$, the expert prefers y_L to y_P (y_L is closer to the dotted line than is y_P) and prefers y_M to y_L. Thus, if there were no uncertainty about the outcome of the meeting – for instance, if all meetings were 'successes' – then the expert would not be willing to reveal whether the state is above or below $\frac{1}{6}$; for states $\theta = \frac{1}{6} - \varepsilon$, the expert would say $\theta \in [\frac{1}{6}, \frac{5}{12}]$, thereby inducing y_M instead of y_L. If all meetings were failures, then for states $\theta = \frac{1}{6} + \varepsilon$, the expert would say $\theta < \frac{1}{6}$, thereby inducing y_L instead of y_P.

There exists a probability $p = \frac{16}{21}$ such that when $\theta = \frac{1}{6}$ the expert is indifferent between y_L and a $(p, 1-p)$ lottery between y_M and y_P (whose certainty equivalent is labelled y_C in the figure). Also, when $\theta < \frac{1}{6}$, the expert prefers y_L to a $(p, 1-p)$ lottery between y_M and y_P, and when $\theta > \frac{1}{6}$, the expert prefers a $(p, 1-p)$ lottery between y_M and y_P to y_L.

It remains to specify a conversation such that the meeting is successful with probability $p = \frac{16}{21}$. Suppose the expert sends a message (*Low*, A_i) or (*High*, A_i) and the decision maker sends a message A_j, where $i, j \in \{1, 2, \ldots, 21\}$. These messages are interpreted as follows. *Low* signals that $\theta \leq \frac{1}{6}$ and *High* signals that $\theta > \frac{1}{6}$. The A_i and A_j messages play the role of a coordinating device and determine whether the meeting is successful. The expert chooses A_i at random and each A_i is equally likely. Similarly, the decision maker chooses A_j at random. Given these choices, the meeting is a

$$\begin{array}{lll} \textit{Success} & \text{if} & 0 \leq i - j < 16 \text{ or } j - i > 5 \\ \textit{Failure} & \text{otherwise} & \end{array}$$

For example, if the messages of the expert and the decision maker are (*High*, A_{17}) and A_5, respectively, then it is inferred that $\theta > \frac{1}{6}$ and, since $i - j = 12 < 16$, the meeting is a success. Observe that with these strategies, given any A_i or A_j, the probability that the meeting is a success is exactly $\frac{16}{21}$.

The equilibrium constructed above conveys more information than any equilibria of the CS game. The remarkable fact about the equilibrium is that this improvement in information transmission is achieved by adding a stage in which the *uninformed* decision maker also participates. While the analysis above concerns itself with the case where $b = \frac{1}{12}$, informational improvement through a 'conversation' is a general phenomenon (Krishna and Morgan, 2004a):

Proposition 3: Multiple stages of communication together with active participation by the decision maker always improve information transmission.

What happens if the two parties converse more than once? Does every additional stage of communication lead to more information transmission? In a closely related setting, Aumann and Hart (2003) obtain a precise but abstract characterization of the set of equilibrium payoffs that emerge in sender–receiver games with a finite number of states and actions when the number of stages of communication is infinite. Because the CS model has a continuum of states and actions, their characterization does not directly apply. Nevertheless, it can be shown that, even with an unlimited conversation, full revelation is impossible. A full characterization of the set of equilibrium payoffs with multiple stages remains an open qst.

Delegation

A key tenet of organizational theory is the 'delegation principle', which says that the power to make decisions should reside in the hands of those with the relevant information (Milgrom and Roberts, 1992). Thus, one approach to solving the information problem is simply to delegate the decision to the expert. However, the expert's bias will distort the

chosen action from the decision maker's perspective. Delegation this leads to a trade-off between an optimal decision by an uninformed party and a biased decision by an informed party.

Is delegation worthwhile? Consider again an expert with bias $b = \frac{1}{12}$. The decision maker's payoff from the most informative partition equilibrium is $-\frac{1}{36}$. Under delegation, the action chosen is $y = \theta + b$ and the payoff is $-b^2 = -\frac{1}{144}$. Thus delegation is preferred. Dessein (2002) shows that this is always true:

Proposition 4: If the expert's bias is not too large $(b \leq \frac{1}{4})$, delegation is better than all equilibria of the CS model.

In fact, by exerting only slightly more control, the decision maker can do even better. As first pointed out by Holmström (1984), the optimal delegation scheme involves limiting the scope of actions from which the expert can choose. Under the uniform-quadratic specification, the decision maker should optimally limit the expert's choice of actions to $y \in [0, 1 - b]$. When $b = \frac{1}{12}$, limiting actions in this way raises the decision maker's payoff from $-\frac{1}{144}$ to $-\frac{1}{162}$.

Optimal delegation still leads to information loss. When the expert's choice is ‘capped’, in high states the action is unresponsive to the state.

An application of the delegation principle arises in the US House of Representatives. Typically a specialized committee – analogous to an informed expert – sends a bill to the floor of the House – the decision maker. How it may then be amended depends on the legislative rule under effect. Under the so-called *closed rule* the floor is limited in its ability to amend the bill, while under the *open rule* the floor may freely amend the bill. Thus, operating under a closed rule is similar to delegation, while an open rule is similar to the CS model. The proposition above suggests, and Gilligan and Krehbiel (1987; 1989) have shown, that in some circumstances the floor may benefit by adopting a closed rule.

Contracts

Up until now we have assumed that the decision maker did not compensate the expert for his advice. Can compensation, via an incentive contract, solve the information problem? To examine this, we amend the model to allow for compensation and use mechanism design to find the optimal contract. Suppose that the payoffs are now given by

$$U(y, \theta, t) = -(y - \theta)^2 - t V(y, \theta, b, t) = -(y - \theta - b)^2 + t$$

where $t \geq 0$ is the amount of compensation.

Using the revelation principle, we can restrict attention to a direct mechanism where both *t* and *y* depend on the state θ reported by the expert. Notice that such mechanisms directly link the expert's reports to payoffs – talk is no longer cheap.

Contracts are powerful instruments. A contract that leads to full information revelation and first-best actions is:

$$t(\hat{\theta}) = 2b(1 - \hat{\theta})\, y(\hat{\theta}) = \hat{\theta}$$

where $\hat{\theta}$ is the state reported by the expert. Under this contract, the expert can do no better than to tell the truth, that is, to set $\hat{\theta} = \theta$, and, as a consequence, the action undertaken in this scheme is the ‘bliss’ action for the decision maker. Full revelation is expensive, however. When $b = \frac{1}{12}$, the decision maker's payoff from this scheme is $-\frac{1}{12}$. Notice that this is worse than the payoff of $-\frac{1}{36}$ in the best CS equilibrium, which can be obtained with no contract at all. The costs of implementing the fully revealing contract outweigh the benefits.

In general, Krishna and Morgan (2004b) show:

Proposition 5: With contracts, full revelation is always feasible but never optimal.

The proposition above shows that full revelation is never optimal. No contract at all is also not optimal – delegation is preferable. What is the structure of the optimal contract? A typical optimal contract is depicted as the dark line in Figure 2. First, notice that, even though the decision maker could induce his bliss action for some states, it is never optimal to do so. Instead, for low states $(\theta < b)$ the decision maker implements a ‘compromise’ action – an action that lies between θ and $\theta + b$. When $\theta > b$, the optimal contract simply consists

of capped delegation.

Figure 2

An optimal contract, $b \leq \frac{1}{3}$

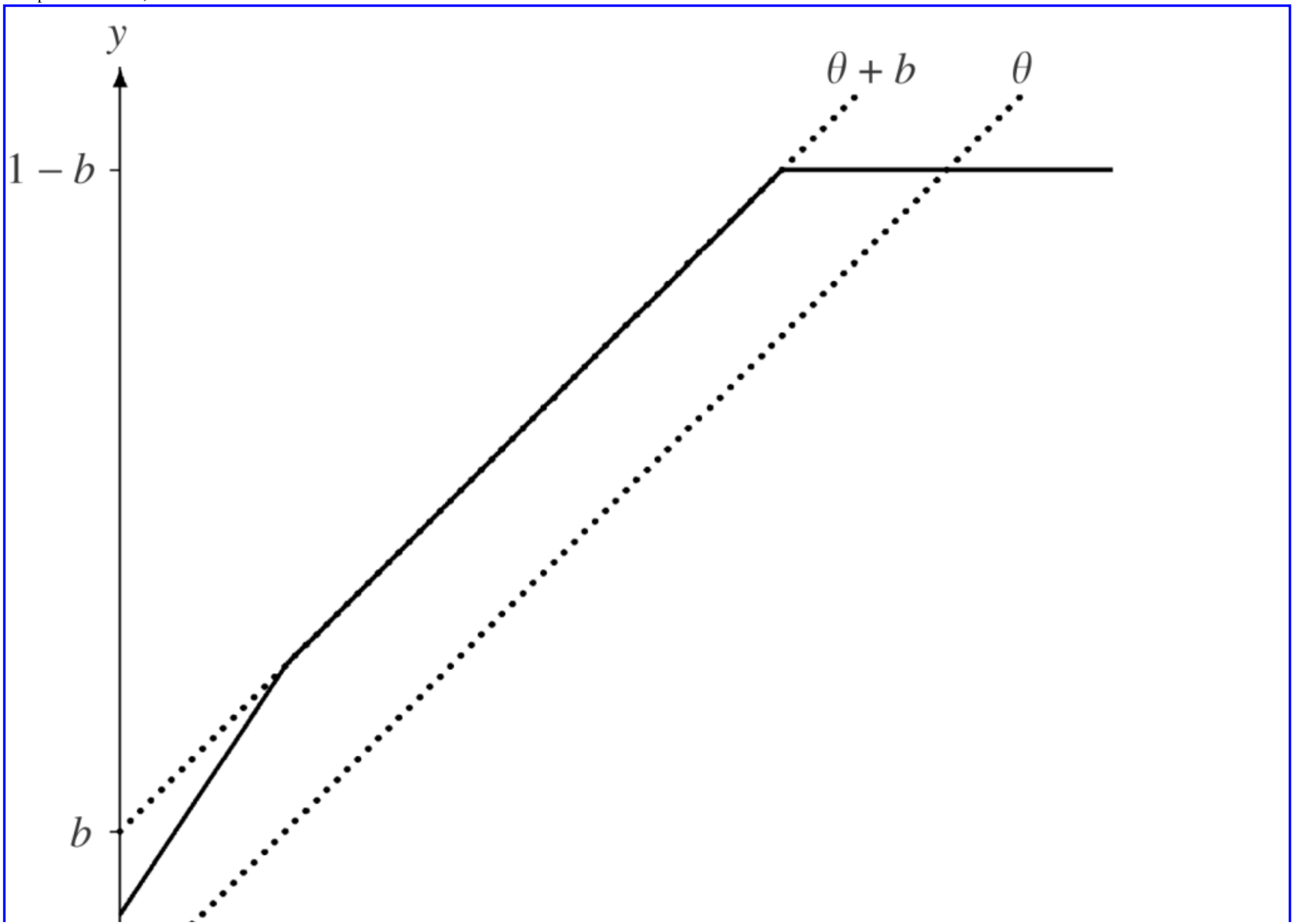

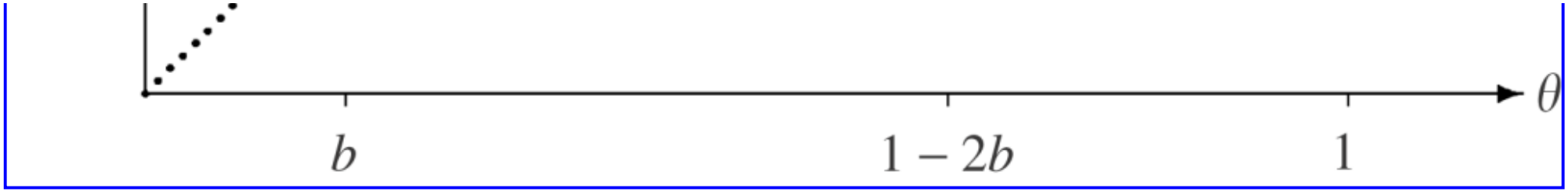

Multiple senders

Thus far we have focused attention on how a decision maker should consult a single expert. In many instances, decision makers consult multiple experts – often with similar information but differing ideologies (biases). Political leaders often form cabinets of advisors with overlapping expertise. How should a cabinet be constituted? Is a balanced cabinet – one with advisors with opposing ideologies – helpful? How should the decision maker structure the 'debate' among her advisors?

To study these issues, we add a second expert having identical information to the CS model. To incorporate ideological differences, suppose the experts have differing biases. When both b_1 and b_2 are positive, the experts have *like bias* – both prefer higher actions than does the decision maker. In contrast, if $b_1 > 0$ and $b_2 < 0$, then the experts have *opposing bias* – expert 1 prefers a higher action and expert 2 a lower action than does the decision maker.

Simultaneous talk

When both experts report to the decision maker simultaneously, the information problem is apparently solved – full revelation is now an equilibrium. To see this, suppose the experts have like bias and consider the following strategy for the decision maker: choose the action that is the more 'conservative' of the two recommendations. Precisely, if $m_1 < m_2$, choose action m_1 and vice versa if $m_2 < m_1$. Under this strategy, each expert can do no better than to report θ honestly if the other does likewise. If expert 2 reports $m_2 = \theta$, then a report $m_1 > \theta$ has no effect on the action. However, reporting $m_1 < \theta$ changes the action to $y = m_1$, but this is worse for expert 1. Thus, expert 1 is content to simply tell the truth. Opposing bias requires a more complicated construction, but the effect is the same: full revelation is an equilibrium (see Krishna and Morgan, 2001b).

Notice that the above construction is fragile because truth-telling is a weakly dominated strategy. Each expert is at least as well off by reporting $m_i = \theta + b_i$ and strictly better off in some cases. Battaglini (2002) defines an equilibrium refinement for such games which, like the notion of perfect equilibrium in finite games, incorporates the usual idea that players may make mistakes. He then shows that such a refinement rules out all equilibria with full revelation regardless of the direction of the biases. While the set of equilibria satisfying the refinement is unknown, the fact that full revelation is ruled out means that simply adding a second expert does not solve the information problem satisfactorily.

Sequential talk

Finally, we turn to the case where the experts offer advice in sequence:

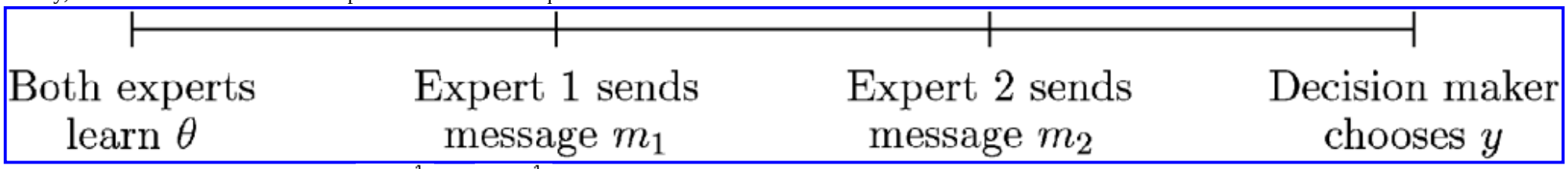

Suppose that the two experts have biases $b_1 = \frac{1}{18}$ and $b_2 = \frac{1}{12}$, respectively. It is easy to verify (with the use of (2)) that, if only expert 1 were consulted, then the most informative equilibrium entails his revealing that the state is below $\frac{1}{9}$, or between $\frac{1}{9}$ and $\frac{4}{9}$, or above $\frac{4}{9}$. If only expert 2 were consulted, then the most informative equilibrium is where he reveals whether the state is below or above $\frac{1}{3}$. If the decision maker were able to consult only one of the two experts, she would be better off consulting the more loyal expert 1.

But what happens if she consults both? It turns out that, if both experts actively contribute information, then the decision maker can do no better than the following equilibrium.

Expert 1 speaks first and reveals whether or not the state is above or below $\frac{11}{27}$. If expert 1 reveals that the state is above $\frac{11}{27}$, expert 2 reveals nothing further. If, however, expert 1 reveals that the state is below $\frac{11}{27}$, then expert 2 reveals further whether or not it is above or below $\frac{1}{27}$. That this is an equilibrium may be verified again by using (2) and recognizing that, in state $\frac{1}{27}$, expert 2 must be indifferent between the optimal action in the interval $[0, \frac{1}{27}]$ and the optimal action in $[\frac{1}{27}, \frac{11}{27}]$. In state $\frac{11}{27}$, expert 1 must be indifferent between the optimal action in $[\frac{1}{27}, \frac{11}{27}]$ and the optimal action in $[\frac{11}{27}, 1]$.

Sadly, by actively consulting both experts, the decision maker is worse off than if she simply ignored expert 2 and consulted only her more loyal advisor, expert 1. This result is quite general, as shown by Krishna and Morgan (2001a):

Proposition 6: When experts have like biases, actively consulting the less loyal expert never helps the decision maker.

The situation is quite different when experts have opposing biases, that is, when the cabinet is balanced. To see this, suppose that the cabinet is comprised of two equally loyal experts biases $b_1 = \frac{1}{12}$ and $b_2 = -\frac{1}{12}$. Consulting expert 1 alone leads to a partition $[0, \frac{1}{3}]$, $[\frac{1}{3}, 1]$ while consulting expert 2 alone leads to the partition $[0, \frac{2}{3}]$, $[\frac{2}{3}, 1]$. If instead the decision maker asked both experts for advice, the following is an equilibrium: expert 1 reveals whether θ is above or below $\frac{2}{9}$. If he reveals that the state is below $\frac{2}{9}$, the discussion ends. If, however, expert 1 indicates that the state is above $\frac{2}{9}$, expert 2 is actively consulted and reveals further whether the state is above or below $\frac{7}{9}$. Based on this, the decision maker takes the appropriate action. One may readily verify that this is an improvement over consulting either expert alone. Once again the example readily generalizes:

Proposition 7: When experts have opposing biases, actively consulting both experts always helps the decision maker.

Indeed, the decision maker can be more clever than this. One can show that, with experts of opposing bias, there exist equilibria where a portion of the state space is *fully revealed*. By allowing for a ‘rebuttal’ stage in the debate, there exists an equilibrium where *all* information is fully revealed.

See Also

- agency problems
- signalling and screening

Bibliography

Aumann, R. and Hart, S. 2003. Long cheap talk. *Econometrica* 71, 1619–60.

Austen-Smith, D. 1993. Interested experts and policy advice: multiple referrals under open rule. *Games and Economic Behavior* 5, 3–43.

Baron, D. 2000. Legislative organization with informational committees. *American Journal of Political Science* 44, 485–505.

Battaglini, M. 2002. Multiple referrals and multidimensional cheap talk. *Econometrica* 70, 1379–401.

Crawford, V. and Sobel, J. 1982. Strategic information transmission. *Econometrica* 50, 1431–51.

Dessein, W. 2002. Authority and communication in organizations. *Review of Economic Studies* 69, 811–38.

Gilligan, T. and Krehbiel, K. 1987. Collective decision-making and standing committees: an informational rationale for restrictive amendment procedures. *Journal of Law, Economics, and Organization* 3, 287–335.

Gilligan, T. and Krehbiel, K. 1989. Asymmetric information and legislative rules with a heterogeneous committee. *American Journal of Political Science* 33, 459–90.

Grossman, G. and Helpman, E. 2001. *Special Interest Politics*. Cambridge, MA: MIT Press.

Holmström, B. 1984. On the theory of delegation. In *Bayesian Models in Economic Theory*, ed. M. Boyer and R. Kihlstrom. Amsterdam: North-Holland.

Krishna, V. and Morgan, J. 2001a. A model of expertise. *Quarterly Journal of Economics* 116, 747–75.

Krishna, V. and Morgan, J. 2001b. Asymmetric information and legislative rules: some amendments. *American Political Science Review* 95, 435–52.

Krishna, V. and Morgan, J. 2004a. The art of conversation: eliciting information from experts through multi-stage communication. *Journal of Economic Theory* 117, 147–79.

Krishna, V. and Morgan, J. 2004b. Contracting for information under imperfect commitment. Paper CPC05-051. Competition Policy Center, University of California, Berkeley.

Milgrom, P. and Roberts, J. 1992. *Economics, Organization and Management*. Englewood Cliffs, NJ: Prentice Hall.

Wolinsky, A. 2002. Eliciting information from multiple experts. *Games and Economic Behavior* 41, 141–60.

How to cite this article

Krishna, Vijay and John Morgan. "cheap talk." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000545> doi:10.1057/9780230226203.0222

The New Palgrave Dictionary of Economics Online

chemical industry

Ashish Arora and Alfonso Gambardella

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

chemical industry; complementarities; innovation; licensing of technology; patents; petrochemical industry; technical change; vertical integration

Article

The chemical industry is among the largest manufacturing industries; its products range from acids to intermediate chemicals such as synthetic fibres and plastics, and to final products such as soaps, cosmetics, paints and fertilizers. Perhaps as a result, the chemical industry is under-studied by economists, though not by economic and business historians (e.g., Hounshell and Smith, 1988).

The modern chemical industry has its origins in the discovery of synthetic dyes in Britain in the 1850s. German chemical firms such as BASF, Bayer and Hoechst soon dominated the production of synthetic dyestuffs and related organic compounds. The American chemical industry grew by exploiting the rich American natural resource endowments, initially using European technology.

After the First World War, American firms, especially Du Pont, invested in R&D. The inter-war period saw rapid product innovation in synthetic fibres, plastics, resins, adhesives, paints, and coatings, based on polymer science. To succeed commercially, these products had to be produced cheaply, which meant large-scale production and, in turn, the development of chemical engineering. The Second World War marked a watershed. The chemical industry became closely linked with the oil industry, as many chemicals used petroleum-based inputs instead of coal by-products. The United States was the first country to develop a petrochemicals industry, mainly due to its abundant oil reserves, as well as wartime government programmes for aviation fuel and synthetic rubber.

The early advantage of the US chemical industry in petrochemicals was eroded as technologies diffused widely, first to Europe and Japan; and in the 1970s China, Taiwan and S. Korea emerged as leading producers. Increased competition, the oil shocks of the 1970s, and waning possibilities for product innovation together resulted in exit: larger, multi-product firms exited earlier, but larger plants closed later (Lieberman, 1990). In addition, firms reshuffled product portfolios so as to focus on fewer products but in more geographical markets (Arora and Gambardella, 1998). The restructuring took a heavy toll of

incumbents; and many familiar names such as Hoechst, Union Carbide, Ciba-Geigy, Sandoz, and American Cynamid have vanished.

A number of interesting themes emerge, some of which have been studied by economists. Others remain as potentially rich veins to be mined.

International competition: Why did British firms fail to exploit the rich potential of organic chemistry despite a head start, access to cheap inputs (coal tar) and to the British textile industry, and a well-functioning capital market? Many explanations, none entirely persuasive, have been offered, including the alleged bias of the British financial system towards low risk-projects (Da Rin, 1998), the weak links between English universities and industry (Murmann and Landau, 1998), and inferior management (Chandler, 2005).

Patents: Overenthusiastic patent protection in the 1870s nearly killed the French dyestuff industry, while German firms strategically used patent protection (Arora, 1997). The confiscation of German patents and industrial property in Britain, France and the United States after both world wars was a setback to German firms but proved insufficient for the Americans and British to catch up. Systematic analysis of this natural experiment can shed light on the role of patents in shaping oligopolistic competition.

Markets for technology: Arrow (1962) observed that Du Pont appeared to have profited as much from innovations it had licensed from others as from its own products, perhaps reflecting imperfections in the market for technology. Yet technology licensing has been extensive in chemicals (Arora, Fosfuri and Gambardella, 2001). The market for technology dramatically changed industry structure, with accumulated production experience of incumbents insufficient to deter successful entry (Lieberman, 1989).

Complementarities and industrial convergence: After the Second World War, oil refining and the production of synthetic fibres and plastics came to share a common technical base. The convergence led to vertical integration by oil firms into chemicals and chemical firms into petrochemicals (Lieberman, 1991). Thanks to a market for petrochemical technology, the European chemical industry was able to switch to petrochemicals very rapidly, despite very substantial investments in coal-based technologies.

Division of labour and vertical industry structure: Specialized engineering firms, which arose to provide plant construction and design services to chemical firms, led the way in diffusing petrochemical technologies worldwide (Freeman, 1968). This competition prodded even large chemical firms such as Union Carbide to give licences to others, further diffusing technology and promoting entry (Arora, Fosfuri and Gambardella, 2001). The chemical industry thus provides a clear example of the benefits of vertically disintegrated industry structures in promoting entry and competition.

The enduring lesson of the history of the chemical industry for economists is the important role of firms – their history and their capabilities – which largely explains why some countries dominated the industry for such long periods. But that history is also a strong reminder to that, in the end, even the mightiest firms must eventually bow to market forces.

See Also

- intellectual property, history of
- patents

- technical change
- vertical integration

Bibliography

Arora, A. 1997. Patent, licensing and market structure in the chemical industry. *Research Policy* 26, 391–403.

Arora, A., Fosfuri, A. and Gambardella, A. 2001. *Market for Technology: The Economics of Innovation and Corporate Strategy*. Cambridge, MA: MIT Press.

Arora, A. and Gambardella, A. 1998. Evolution of industry structure in the chemical industry. In *Chemicals and Long-Term Economic Growth*, ed. A. Arora, R. Landau and N. Rosenberg. New York: John Wiley and Sons.

Arrow, K. 1962. Comments on case studies. In *The Rate and the Direction of Inventive Activity: Economic and Social Factors*, ed. R. Nelson. Princeton, NJ: Princeton University Press.

Chandler, A. 2005. *Shaping the Industrial Century: The Remarkable Story of the Evolution of the Modern Chemical and Pharmaceutical Industries*. Cambridge, MA: Harvard University Press.

Da Rin, M. 1998. Finance and the chemical industry. In *Chemicals and Long-Term Economic Growth*, ed. A. Arora, R. Landau and N. Rosenberg. New York: John Wiley and Sons.

Freeman, C. 1968. Chemical process plant: innovation and the world market. *National Institute Economic Review* 45(August), 29–51.

Hounshell, D. and Smith, J. 1988. *Science and Strategy: Du Pont R&D, 1902–1980*. Cambridge: Cambridge University Press.

Lieberman, M. 1989. The learning curve, technology barriers to entry, and competitive survival in the chemical processing industries. *Strategic Management Journal* 10, 431–47.

Lieberman, M. 1990. Exit from declining industries: ‘shakeout’ or ‘stakeout’? *Rand Journal of Economics* 21, 538–54.

Lieberman, M. 1991. Determinants of Vertical integration: an empirical test. *Journal of Industrial Economics* 39, 451–66.

Murmann, P. and Landau, R. 1998. On the making of comparative advantage: the development of the chemical industries in Britain and Germany since 1850. In *Chemicals and Long-Term Economic*

Growth, ed. A. Arora, R. Landau and N. Rosenberg. New York: John Wiley and Sons.

How to cite this article

Arora, Ashish and Alfonso Gambardella. "chemical industry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000544>
doi:10.1057/9780230226203.0223

The New Palgrave Dictionary of Economics Online

Chenery, Hollis B. (1918 – 1994)

Shantayanan Devarajan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Hollis Burnley Chenery was born in Richmond, Virginia, in 1918. He received his Ph.D. at Harvard University, worked for the Marshall Plan in Europe, taught at Stanford University, served as Assistant Administrator of the US Agency for International Development before joining the World Bank in 1970 for a distinguished, 13-year career there. He returned to Harvard as a professor in 1983. He died in 1994.

Keywords

Chenery, H. B.; comparative advantage; computable general-equilibrium models; development economics; development strategy; economic development; foreign aid; poverty alleviation; total factor productivity; World Bank

Article

Hollis Burnley Chenery was the consummate development economist. He defined the contours of the field with his ground-breaking research on patterns of development and development strategy. He developed tools that helped translate research into policy, and, as Vice-President for Development Policy at the World Bank, he helped shift the focus of development economics from a narrow one of economic growth to the alleviation of poverty.

Patterns of development

In the tradition of Kuznets and Denison, Chenery was interested in how economies grow, whether there were systematic patterns in the process of development. His 1960 paper in the *American Economic Review*, ‘Patterns of Industrial Growth’, grew into a decade-long research project with Moshe Syrquin culminating in their 1975 book, *Patterns of Development, 1950–1970*. Many of the patterns that Chenery and Syrquin found are received wisdom today: as countries grow, the share of agriculture in GDP

declines, and the shares of industry and services increase; and overall GDP growth is typically accompanied by an increase in total factor productivity (TFP) growth. Chenery and Syrquin were the first to document these patterns, using the statistical techniques available at the time, for a large number of countries in the modern era. Their work has led to Chenery–Syrquin ‘norms’ (interestingly, a word they never used) whereby countries could benchmark their progress in the development process. They were also aware of the limitations of this approach, identifying for example the differences between large countries and small ones, work that has been extended by Perkins and Syrquin (1989). The observed pattern of TFP growth has been questioned by, among others, Young (1995) and is still a topic of vigorous debate.

Development strategy

In contrast with the recent work on cross-country growth (see Barro, 1991), the *Patterns* work was silent on what countries could do to grow faster. Chenery answered this question in a series of major pieces on development strategy. He entered the debate between outward- and inward-looking development strategies in his 1961 *American Economic Review* paper, ‘Comparative Advantage and Development Policy’. While countries should only produce those goods in which they have a comparative advantage, Chenery conjectured that comparative advantage in certain goods could be developed through careful investment policies. Chenery's notions saw a resurgence in the 1980s in the Brander–Spencer (1985) and other models of policy-induced comparative advantage. Of course, policies to create comparative advantage have to be carefully designed, especially because public investment has economy-wide impacts, as Chenery showed in his 1959 book with Peter Clark, *Interindustry Economics*.

Chenery's thinking on development strategy evolved over time. He became convinced that a country's underlying economic structure – the functioning of its labour and capital markets, its resource endowments – influenced the choices it could make in trying to create ‘dynamic comparative advantage’. Using case studies, cross-country analysis and model-based analysis, he distilled this work in his 1984 book with Sherman Robinson and Syrquin, *Industrialization and Growth: A Comparative Study*.

Structure also determines how foreign aid affects the economy, as Chenery showed in his ‘two-gap’ model (see Chenery and Strout, 1966; Chenery and Bruno, 1962). *Ex ante*, an economy may be foreign-exchange-constrained or fiscally constrained. Since foreign aid is both foreign exchange and resources to the government, its impact depends on which constraint is binding. An extended version of this simple model became the workhorse model of aid agencies such as the World Bank. It saw a resurgence during the debt crises of the 1980s. It has also been criticized for neglecting the role of prices and incentives (see Easterly, 1999), although it can be shown that, as long as domestic and foreign capital are imperfect substitutes, most of the results of the two-gap model survive in a fully specified, intertemporal, general-equilibrium model.

Tools

Building on his work on the interdependence of investment decisions, Chenery and his collaborators pioneered the development of multisectoral models for investment planning, collected in his co-authored

book, *Studies in Development Planning*. This work saw applications in various planning agencies, notably in India. Recognizing the limitations of linear programming approaches, Chenery encouraged the development of computable general-equilibrium (CGE) models at the World Bank and in universities. Today, CGE models are commonly used to inform policy in developing and developed countries, although they too have their limits (see Devarajan and Robinson, 2005).

Redistribution with growth

Arriving at the World Bank in 1970, Chenery proceeded to establish the first, and eventually one of the most influential, research programmes in economic development. In addition to producing academic-quality research, Chenery's group helped shape Bank policies. In 1974, Chenery and his associates published *Redistribution with Growth*, a seminal book that, while recognizing the need for direct action to alleviate poverty (especially since the high growth of the 1960s had not significantly reduced poverty), showed that wealth redistribution can and should be consistent with the promotion of economic growth. Chenery's approach has been the leitmotif of the World Bank's (and indeed most development agencies') strategy since then.

See Also

- development economics
- economic growth, empirical regularities in
- foreign aid
- redistribution of income and wealth
- structural change
- World Bank

Selected works

1959. (With P. G. Clark.) *Interindustry Economics*, New York: John Wiley & Sons.

1960. Patterns of industrial growth. *American Economic Review* 50, 624–54.

1961. Comparative advantage and development policy. *American Economic Review* 51, 18–51.

1962. (With M. Bruno.) Development alternatives in an open economy: the case of Israel. *Economic Journal* 72, 79–103.

1966. (With A. Strout.) Foreign assistance and economic development. *American Economic Review* 56, 679–733.

1971. *Studies in Development Planning*. Cambridge, MA: Harvard University Press.

1974. (With M. Syrquin.) *Patterns of Development, 1950–1970*. Oxford: Oxford University Press.

1974. *Redistribution with Growth*. London and New York: Oxford University Press.

1984. (With S. Robinson and M. Syrquin.) *Industrialization and Growth: A Comparative Study*. New York: Oxford University Press.

Bibliography

Barro, R. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics* 106, 407–43.

Brander, J.A. and Spencer, B.J. 1985. Export subsidies and international market share rivalry. *Journal of International Economics* 18, 83–100.

Devarajan, S. and Robinson, S. 2005. The influence of computable general equilibrium models on policy. In *Frontiers in Applied General Equilibrium Modeling*, ed. T. Kehoe, T.N. Srinivasan and J. Whalley. Cambridge: Cambridge University Press.

Easterly, W.R. 1999. The ghost of financing gap: testing the growth model of international financial institutions. *Journal of Development Economics* 60, 423–38.

Perkins, D.H. and Syrquin, M. 1989. Large countries: the influence of size. In *Handbook of Development Economics*, ed. H.B. Chenery and T.N. Srinivasan. Amsterdam: North-Holland.

Young, A. 1995. The tyranny of numbers: confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 110, 641–80.

How to cite this article

Devarajan, Shantayanan. "Chenery, Hollis B. (1918–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000599> doi:10.1057/9780230226203.0224

The New Palgrave Dictionary of Economics Online

Chevalier, Michel (1806 – 1879)

P. Bridel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chevalier, M.; free trade; Leroy-Beaulieu, P. P.; liberalism; Rossi, P.L. E.; Saint-Simon, C. H.; Walras, L.

Article

Born in Limoges, 13 January 1806; died in Paris, November 1879. Undoubtedly one of the most eminent 19th-century French economists, Chevalier belongs to that most typical brand of engineer-economists. First in his class (*major*) at the Ecole Polytechnique in 1830 and member of the Corps des Mines as an economist, Chevalier came very early under the spell of Saint-Simon's utopian doctrine. From his early editorship of the Saint-Simonian newspaper *Le Globe* (1830–2) and his subsequent sentence to a year in jail (for ‘outrage to morals’ for publishing advanced ideas on the liberation of women, sexual liberty and the need for communal life) to a made-to-measure niche as economic adviser to Napoleon III and ‘éminence grise’ to the Second Empire business and banking establishment, Chevalier applied his brilliant mind to various current problems and policy issues without managing, however, to escape completely from the Saint-Simonian mystique. His main claim to fame, the Anglo-French Treaty of 1860 (the Cobden–Chevalier Treaty), an important if short-lived interruption in the general protectionist policy of France, is one of the best illustrations of these twin components of Chevalier's approach to economics and economic policy: weak on the analytics and very strong on the factual analysis with a touch of Saint-Simonian idealism.

Together with public works, cheap bank credit and education, free trade is one of the articles of faith he took over from the Saint-Simonian doctrine. Chevalier returned to these issues throughout his life (notably in his penetrating analysis of the American economy and banking system in the early 1830s which earned him later the nickname of ‘Economic Tocqueville’). Binding these various elements with a quasi-philosophical concept of association (as the cornerstone of social order), Chevalier suggests a broad theory of economic growth which he considered flexible enough to be applied to different times and countries.

His Saint-Simonian antecedents and his extensive travelling (to England, Egypt and foremost to the

United States) rendered Chevalier suspicious of all ‘absolutist’ economic theory. In fact, in his most technical chapters (particularly on money) Chevalier never digs beneath the surface of things and contributes very little, if anything, to analytic economics. His only systematic work, his *Cours* (1843; 1844; 1850) delivered at the Collège de France offers little more in the field of theory than a lengthy (and flat) apology for Say's brand of ‘vulgar’ liberalism. With Rossi, his predecessor, and Leroy-Beaulieu, his successor at the Collège de France, Chevalier was in fact largely responsible for introducing and perpetuating in academic circles the liberal orthodoxy that was to bar Walras from getting an appointment in the 1860s and that dominated French economics for so long that as late as 1939 Keynes could still quip about its lack of ‘deep roots in systematic thought’ (1939, p. xxxii).

Selected works

1843, 1844, 1850. *Cours d'économie politique*. 3 vols. Paris: Capelle.

Bibliography

Keynes, J.M. 1939. French Preface to *The General Theory of Employment, Interest and Money*. In *Collected Writings*, vol. 7. London: Macmillan, 1973, 31–5.

Walch, J. 1975. *Michel Chevalier, économiste, saint-simonien*. 1806–1879. Paris: Vrin (with extensive bibliography).

How to cite this article

Bridel, P. "Chevalier, Michel (1806–1879)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000122> doi:10.1057/9780230226203.0225

The New Palgrave Dictionary of Economics Online

Chicago School (new perspectives)

Ross B. Emmett

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

M. W. Reder's entry on the Chicago School closed with the claim that the final chapter of the School's history was about to end. Chicago economics has changed, but it has also stayed the same. Each of the four movements of recent Chicago economics are rooted in common themes of the tradition. As well, our interpretation of economics at Chicago has evidenced both continuity and change. Historians are examining the history of the institutional structure of Chicago economics, as well as the histories of specific fields at Chicago (labour, economic history, quantitative analysis) and finding both change and continuity in the tradition.

Keywords

American Economic Association; Becker, G.; Chicago model; Chicago School; Coase, R. H.; Cowles Commission; Director, A.; Friedman, F.; Griliches, Z.; Hansen, L.; Harberger, A.; Heckman, J.; Johnson, D. G.; Johnson, H. G.; Knight, F.; Lange, O.; Levitt, S.; Lewis, H. G.; Lucas, R.; Markowitz, H.; Miller, M.; Mundell, R.; Murphy, K.; National Bureau of Economic research; positive economics; Rees, A.; Sargent, T.; Schultz, G.; Schultz, T. W.; Stigler, G.; Viner, J.

Article

The history of Chicago economics remains a story of continuity and change.
M.W. Reder closed the entry on the Chicago School in the first edition of *The New Palgrave* (and reproduced in this edition) with the claim that the final chapter of the School's history was about to end. Perhaps he was right: the apex of the School's influence on public policy – the presidency of Ronald Reagan – ended in 1988. By that time key figures in the School's history had retired, become inactive, left the University of Chicago, or died. Milton Friedman retired in 1977 and moved to the Hoover Institution at Stanford University, where he was eventually joined by Aaron Director (linchpin of the early Chicago law and economics movement) and George Schultz (former dean of the University of

Chicago's Graduate School of Business and Secretary of State under President Reagan); he died in late 2006. Arnold Harberger stepped down as chair of the Economics Department in the early 1980s and moved to UCLA shortly thereafter, following the previous departure of long-time graduate advisor H. Gregg Lewis to Duke in 1977. T.W. Schultz, former department chair, was largely inactive as a scholar by the late 1970s; his student and collaborator for many years, D. Gale Johnson, retired in the early 1980s. In international economics, Robert Mundell left the university in the early 1970s and Harry Johnson died in 1979. Of the early leaders, only George Stigler (industrial organization) and Ronald Coase (law and economics) remained active at Chicago, although both were retired.

But it would be a mistake to see the 1980s as the final chapter of the Chicago School. Four major movements in Chicago economics since 1980 are captured in the awarding of more recent Nobel Prizes. Gary Becker was awarded the prize for his work in the new home and social economics. Robert Lucas won for developments in empirical macroeconomics. Merton Miller was joined by former Chicago researcher Harry Markowitz for their development of finance theory. And James Heckman won the prize for the development of microeconometrics. Alongside these scholars (Miller died in 2000, but the others remain active), the next generation of Chicago economists is making a place for itself. Both Thomas Sargent and Lee Hansen have won the new Erwin Plein Nemmers Prize for significant contributions to new modes of analysis in economics, and Kevin Murphy and Steven Levitt have won the coveted John Bates Clark medal from the American Economic Association.

Each of the four recent movements within Chicago economics – finance, empirical macroeconomics, the new home economics, and microeconometrics – are rooted in common Chicago themes: the application of price theory, the development of methods for the quantitative analysis of social problems, and the notion that economics is an applied policy science. The Chicago approach rests on a three-legged stool which combines an appreciation for the ‘simple’ analytics of Marshallian price theory (as Reder observes, a constant at Chicago since the early 1930s), the development of quantitative tools as expressed in Friedman's classic article (1953) on ‘positive economics,’ and the Becker–Stigler prescription to focus attention on the elements of the constraint set, rather than changes in values and preferences, in the explanation of human behaviour (see Becker, 1976; Stigler and Becker, 1977). Once combined, this three-legged methodological stool provided a stable foundation for the continued expansion of the scope of social scientific problems that Chicago economists have addressed (Becker, 1981; Becker and Murphy, 2000; Levitt and Dubner, 2005). Economic imperialism it may be, but Chicago economists argue that it is the only basis upon which a true social science can be built (see Lazear, 2000).

Yet Reder's claim that the book on Chicago economics was about to close was right at least in one regard. Up to the mid-1970s, Chicago economists were an embattled minority (albeit growing in numbers and influence) of the economics profession. After the early 1980s, Chicago was no longer embattled, or even a minority. Its central ideas are still alive, but they are no longer the notions of a contrary-minded small group of scholars; in antitrust, law and economics, monetary theory, labour, finance and applied microeconomics, they comprise a position that has been widely adopted. Chicago economics today is part of the discipline's mainstream; indeed, in some sub-fields it has defined the mainstream. Success outside the confines of Chicago has also changed the School itself: since 1980, Chicago economics has gradually accommodated itself to the common standards of the discipline. Finally, the role of the Chicago School themes within the university has also been rendered more complicated by the remarkable expansion of the Graduate School of Business and the Law School as

centres of Chicago-style economic, legal and public policy analysis.

Change and continuity in Chicago economics

The 1980s were a period of transition in Chicago economics, in several regards. For most of the period from the late 1940s until the early 1980s, the department of economics was chaired by either T. W. Schultz, D. Gale Johnson or Arnold Harberger; the required price theory course (ECON 301) was taught by either Milton Friedman, Harberger or Gary Becker, and required thorough familiarity with the canon of Chicago price theory – the theory texts of Knight (1933), Friedman (1962), Stigler (1966), Becker (1971), and Alchian and Allen (1969); and the other required first-year course was titled ‘money’ (not macroeconomics). The continuity in leadership was disrupted in the early 1980s (just as it had been 30 to 40 years earlier by the departures of Jacob Viner, Oskar Lange and the Cowles Commission, and the retirement of Frank Knight), as the early luminaries retired and passed responsibility on to the next generation (although Becker still shares some of the teaching in ECON 301). But a successful programme is not built around individual scholars, even if they are luminaries like Friedman, Stigler, and Becker. Chicago's success, even in the period from the 1940s to the 1980s, is misunderstood if it is interpreted simply as the product of the unique cluster of scholars that it managed to attract (compare Van Overtveldt, 2007, with Emmett 1998). In the early 1950s, the economics department replaced the traditional lone-scholar model of graduate education and faculty research with a workshop model that created an educational environment for graduate students and faculty members more closely akin to a scientific laboratory within which students and faculty pursued a collaborative intellectual project. While the Chicago model is reasonably well-known today and emulated, it was quite unique in the post-war period, and is central to Chicago's success. After passing the core examinations in price theory and money at the end of the first year, students not only continued to take courses but also associated themselves with a workshop (most workshops were open, so students often attended more than one; but each student was primarily associated with one workshop). Faculty were also associated with at least one workshop, and frequently defined the workshop's style: Friedman's money workshop; Stigler's industrial organization workshop; Fogel and McCloskey's economic history workshop; Harberger's Latin American finance workshop; and Coase's law and economics workshop. In the early years no common model had been established, and the workshops varied significant. Eventually, most workshops adopted the ‘Chicago rules’: the workshop met once per week, papers were distributed beforehand and therefore assumed to have been read, and presenters knew that discussion of the paper might begin as soon as five minutes into their presentation. Most of the workshop time was spent dissecting the paper's thesis, method, and data. Because the pattern of discussion was repeated every week in a dozen or more workshops, students and faculty became quite adept at working within Chicago's rules, applying Marshallian price theory to a wide range of policy-relevant topics. By the early 1980s, the number of economics workshops in the department, the Graduate School of Business, and the Law School was approaching 20. Today, in 2006, it still numbers in the teens.

The transition of key personnel in the early 1980s, therefore, did not affect the structure of the research and educational enterprise which supports the Chicago School. However, it did have an impact on the nature of the research and education of Chicago economists. By the end of the 1980s, the texts which comprised the canon of Chicago price theory lost their pride of place in the reading lists for ECON 301.

At about the same time, the ‘money’ course (ECON 302) became a study of ‘income, employment and the price level’ built around standard Walrasian general equilibrium models that characterize macroeconomic analysis in most economics programmes. As well, the development of more sophisticated econometric models and techniques came to play a larger role in economic research at Chicago. ‘Quantitative methods’ was added as a core examination that all students had to pass in order to continue beyond the first year. In short, Chicago economics today looks a lot like economics everywhere else (in part, of course, because Chicago's approach is taught elsewhere and other programmes have created collaborative research environments like the Chicago workshops), although there remains a distinct Chicago ‘flavour’ that distinguishes it from MIT, Harvard, Berkeley and Yale, if not from Stanford, UCLA and Washington.

Change and continuity in the interpretation of Chicago economics

Even as the contemporary evolution of Chicago economics continues to involve both continuity and change, our understanding of the history of Chicago economics has also evidenced both continuity and change. Reder's original essay was constructed on a model of Chicago economics which placed a small group of key individuals and their ideas at the centre of the School; one could envision his essay as an examination of concentric circles emanating out from the inner circle that started with Viner and Knight and then included Friedman, Stigler and Becker. While not rejecting Reder's model entirely, historians have begun to construct a story of the development of Chicago economics that complicates the model significantly. Three aspects of Chicago School historiography can be highlighted to illustrate the direction of contemporary historical research on the School, and indicate the potential for further research. First, the transition from the Chicago economics of the inter-war period to the Chicago School of the 1950s and 1960s involved several significant changes. The elements of continuity that Reder emphasized remain – the pre-eminent role of price theory, for example – but discontinuities have crept in. Daniel Hammond's recent work on Milton Friedman's early career provides a glimpse into how that transition influenced even one of the mainstays of Chicago economics. Arguing against the continuity thesis about Chicago price theory articulated by Philip Mirowski and Wade Hands (1998), Hammond shows that Friedman had as much in common with NBER-style statistical work as he did with Knight's Chicago approach (Hammond, 2005; see also Hammond, 2008, and Rutherford, 2008) In fact, even Friedman's famous methodological essay may be more a statement of his experiences with the NBER and the Statistical Research Group at Columbia University (associated with Harold Hotelling) than any earlier Chicago economist. In more recent work, Mirowski and Rob van Horn (2008) argue that, whatever the continuities of Chicago's price theoretic tradition are, the Chicago School of the 1950s and 1960s was shaped more by new research projects initiated in the effort to define a new liberalism to in the Cold War period than it was by the classical liberalism of the Knight–Simons agenda in the 1930s and 1940s (see also Amadae, 2003). Thus, while the Chicago School of Friedman and company should not be seen as a totally new tradition, historical reconstructions of their work have opened the door to further exploration of continuities and potential discontinuities between ‘old’ and ‘new’ Chicago.
We have already seen the second aspect of contemporary historical reconstruction in the earlier discussion of the institutional framework of the Chicago School. Rather than seeing individual scholars and their ideas transforming modern economics (as suggested even recently by Van Overveldt, 2007),

contemporary historiography suggests that the intellectual success of the School was built upon a unique research infrastructure, focused in the workshops. Constructing the history of the workshops involves investigating the support network they developed, ranging from private foundation funding to international connections for research and students. Mirowski and van Horn (2008) focus on the role of the Volcker Fund, but other foundations and external research organizations like the Ford Foundation, Rockefeller Foundation (which funded many activities across the University of Chicago from its inception), Earhart Foundation, and the RAND Corporation participated in supporting Chicago's research infrastructure. In terms of international connections, much has been said of the role of the ‘Chicago boys’ in Chile, who set the groundwork for economic liberalization in Latin America and elsewhere, but were appointed to their positions by General Pinochet (Valdez, 1995; Barber, 1995). However, the institutional history of the Chile connection, which goes back to the early 1950s with an educational exchange between the University of Chicago and the Catholic University in Chile, has yet to be completely told. And we also do not have any histories of Chicago's other international research and student connections, including the equally unique relationship with the Hebrew University in Jerusalem and the University of Tel Aviv, despite the fact that Chicago was one of the few American academic institutions that welcomed Jewish scholars.

The third aspect of the Chicago School points toward two potential areas of research which would deepen the type of historical work illustrated above, while also providing insight into the degree of continuity and change within the School. Neither of these areas of research has made significant inroads into contemporary research. The first is the story of the integration of econometric developments at Chicago into the story of Chicago economics (as opposed to their place in the econometric literature). How did we go from Friedman and Stigler to Heckman, Hansen, and Levitt? Was it just Chicago accommodating itself to the mainstream of the discipline, as is often suggested? Did Zvi Griliches and the development of quantitative analysis in agricultural economics play a role? Or Gregg Lewis and Albert Rees and labour economics? Did a quiet revolution go on at Chicago in the fields outside the core exams that gradually changed the School as a whole? These stories need to be examined in greater detail (see Kaufman 2008, for the history of Chicago labour economics in this regard). Second, the Chicago School's laissez-faire reputation is offset by the fact that a large portion of its graduates have gone into public service both in the United States and elsewhere. Harberger alone can count approximately 20 former students who have become central bank governors and ministers of finance. And countless Chicago students staff national and international economic ministries, commissions, and other organizations. If Chicago economists do believe that economics is a policy science, then the history of their interaction with policy, both as policy advocates and as policymakers, needs to be incorporated into our history. Again, what we do know about this history is piecemeal or quite general (for a start in the right direction; see Banzhaf, 2008).

The new perspectives on Chicago economics open the door to both reconstructing the story of the Chicago School and to extending that story forward to the present. While Reder may have been premature to suggest the School's demise, both the reconstruction of its history and the story of its recent developments suggest both continuity and change.

See Also

- Becker, Gary S.
- Buchanan, James M.
- Chicago School
- Coase, Ronald Harry
- Director, Aaron
- economic history
- finance (new developments)
- Friedman, Milton
- Griliches, Zvi
- Heckman, James
- human capital
- Johnson, D. Gale
- Johnson, Harry Gordon
- Knight, Frank Hyneman
- labour economics
- laissez-faire, economists and
- law, economic analysis of
- Lucas, Robert
- macroeconomics, origins and history of
- methodology of economics
- Miller, Merton
- monetary economics, history of
- new classical macroeconomics
- Patinkin, Don
- Reid, Margaret Gilpin
- Rosen, Sherwin
- Scholes, Myron
- Schultz, Henry
- Schultz, T.W.
- Simons, Henry Calvert
- Stigler, George Joseph
- Viner, Jacob
- Yntema, Theodore O.

Bibliography

Alchian, A.A. and Allen, W.R. 1969. *Exchange and Production: Theory in Use*. Belmont, CA: Wadsworth.

Amadae, S. 2003. *Rationalizing Capitalist Democracy: The Cold War Origins of Rational Choice Liberalism*. Chicago: University of Chicago Press.

Banzhaf, S. 2008. The Chicago School of welfare economics. In *The Elgar Companion to the Chicago School*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.

Barber, W.J. 1995. Chile con Chicago: a review essay. *Journal of Economic Literature* 33, 1941–9.

Becker, G.S. 1971. *Economic Theory*. New York: Knopf.

Becker, G.S. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Becker, G.S. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Becker, G.S. and Murphy, K.M. 2000. *Social Economics: Market Behavior in a Social Environment*. Cambridge: Belknap Press.

Emmett, R.B. 1998. Entrenching disciplinary competence: the role of general education and graduate study in Chicago economics. In *From Interwar Pluralism to Postwar Neoclassicism*, ed. M. Rutherford and M. Morgan. Durham, NC: Duke University Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays on Positive Economics*. Chicago: University of Chicago Press.

Friedman, M. 1962. *Price Theory: A Provisional Text*. Chicago: Aldine.

Hammond, D.J. 2008. The development of postwar Chicago price theory. In *The Elgar Companion to the Chicago School of Economics*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.

Hammond, D.J. 2005. More fiber than thread? Evidence on the Mirowski–Hands yarn. In *Agreement on Demand: Consumer Choice Theory in the 20th Century, Annual Supplement to History of Political Economy*, ed. P.E. Mirowski and D.W. Hands. Durham, NC: Duke University Press.

Kaufman, B.E. 2008. Labor in Chicago economics. In *The Elgar Companion to the Chicago School of Economics*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.

Knight, F.H. 1933. *The Economic Organization*. Chicago: University of Chicago Press.

Lazear, E. 2000. Economic imperialism. *Quarterly Journal of Economics* 115, 99–146.

Levitt, S.D. and Dubner, S.J. 2005. *Freakonomics: A Rogue Economist Explores the Hidden Side of*

Everything. New York: William Morrow.

Mirowski, P.E. and Hands, D.W. 1998. A paradox of budgets: the postwar stabilization of neoclassical demand theory. In *From Interwar Pluralism to Postwar Neoclassicism, Annual Supplement to History of Political Economy*, ed. M.S. Morgan and M. Rutherford. Durham, NC: Duke University Press.

Mirowski, P.E. and van Horn, R. 2008. Neoliberalism and Chicago. In *The Elgar Companion to the Chicago School of Economics*, ed. R.B. Emmett. Cheltenham: Edward Elgar.

Rosen, S. 1974. Hedonic prices and implicit markets: product differentiation in perfect competition. *Journal of Political Economy* 82, 34–55.

Rosen, S. 1981. The economics of superstars. *American Economic Review* 71, 845–58.

Rosen, S. 2004. *Markets and Diversity*. Cambridge, MA: Harvard University Press.

Rutherford, M. 2008. Chicago economics and institutionalism. In *The Elgar Companion to the Chicago School of Economics*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.

Stigler, G.J. 1966. *The Theory of Price*, 3rd edn. New York: Macmillan.

Stigler, G.J. and Becker, G.S. 1977. De gustibus non est disputandum. *American Economic Review* 67, 76–90.

Valez, J.G. 1995. *Pinochet's Economists: The Chicago School in Chile*. Cambridge: Cambridge University Press.

Van Overtveldt, J. 2007. *The Chicago School: How the University of Chicago Assembled the Thinkers Who Revolutionized Economics and Business*. Chicago: Agate.

How to cite this article

Emmett, Ross B. "Chicago School (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000576> doi:10.1057/9780230226203.0227

The New Palgrave Dictionary of Economics Online

Chicago School

M.W. Reder

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article deals with the history and main protagonists of the Chicago School from c. 1930 to 1985. The two main beliefs of members of the School are (*a*) that neoclassical price theory can explain observed economic behaviour, and (*b*) that free markets efficiently allocate resources and distribute income, implying a minimal role for the state in economic activity. Chicagoans maintain that no opportunity for arbitrage gains goes unexploited, and subscribe to the efficient markets hypothesis. Their ‘disciplinary imperialism’ leads them frequently to challenge conventional wisdom by applying price theory to seemingly non-economic topics.

Keywords

aggregate demand; arbitrage; axiomatic theories; Becker, G.; Buchanan, J.; capital asset pricing model; capital theory; Chicago school; Clark, J.; Coase theorem; Coase, R.; competition; cost of time; Cowles Commission; deregulation; Director, A.; dividends; Douglas, P.; education; efficient markets hypothesis; egalitarianism; family, economic analysis of; finance; free markets; Friedman, M.; Friedman, R.; general equilibrium theory; Hardy, C.; history of economic thought; human capital; imperfect competition; income distribution; industrial organization; institutionalism; international trade; Keynesianism; Knight, F.; laissez faire; Lange, O.; Laughlin, J.; law, economic analysis of; Leland, S.; Lerner, A.; Lewis, H.; Lucas, R.; market prices; market socialism; Marschak, J.; mathematical economics; Metzler, L.; Miller, M.; Millis, H.; Mitchell, W.; Modigliani, F.; monetarism; monetary theory; money supply; monopolistic competition; Muth, J.; natural monopolies; Nef, J.; neoclassical monetary theory; neoclassical price theory; New Deal; Pareto optimality; political economy; political economy of policy reform; political markets; positive economics; Posner, R.; price theory; progressive and regressive taxation; public ownership; quantitative techniques; racial discrimination; rational expectations; Reder, M. W.; regulation; reserve/deposit ratio; resource allocation; risk; Schultz, H.; Schultz, T.; Simons, H.; social costs; statistical demand curves; statistics and economics; Stigler, G.; stock prices; testing; trade unions;

transaction costs; uncertainty; Veblen, T.; Viner, J.; Wallis, W.; Wright, C.

Article

To identify a Chicago School of economics requires some demarcations, both of ideas and persons, that may not be universally accepted. Justification for these decisions must be heuristic; that is, they facilitate the story to be told. But it is not denied that there may be alternative accounts that would entail different demarcations. In this account, the ‘Chicago School’ is and has been centred in the University of Chicago's Economics Department from about 1930 to the present (1985). However, it is convenient to define the School so as to include many members of the large contingent of economists in the Graduate School of Business and the group of economists and lawyer-economists in the Law School. Largely because of the intellectual loyalty of former students, the influence of the Chicago School extends far beyond the University of Chicago to the faculties of other universities, the civil service, the judiciary and private business. Moreover, this influence is not confined to the United States.

To restrict the retrospective horizon of the School to 1930 implies exclusion of a number of famous economists who had been on the University of Chicago faculty before that time; for example, Thorstein Veblen, Wesley C. Mitchell, J.M. Clark, J. Laurence Laughlin, C.O. Hardy. However, none of these shared the intellectual characteristics that have typified members of the Chicago School as defined here. In a nutshell, the two main characteristics of Chicago School adherents are: (1) belief in the power of neoclassical price theory to explain observed economic behaviour; and (2) belief in the efficacy of free markets to allocate resources and distribute income. Correlative with (2) is a tropism for minimizing the role of the state in economic activity.

Before discussing these characteristics in detail, let me give a brief historical account in which it is convenient to divide the history of the School into three periods: (1) a founding period, in the 1930s; (2) an interregnum, from the early 1940s to the early 1950s; and (3) a modern period, from the 1950s to the present.

During the founding period, the Chicago Economics Department contained a wide diversity of views both on methodology and public policy. Institutionalist views were well represented among the senior faculty, and institutionally oriented students constituted a large part of the graduate student population. Among the prominent Institutionalists were the labour economists H.A. Millis and (one side of) Paul H. Douglas; the economic historians John U. Nef and C.W. Wright, and Simeon E. Leland, a Public Finance specialist and long-time department chairman.

Like other social science departments at Chicago, economics was actively engaged in developing the (then) embryonic ‘quantitative techniques’. The leading figures in quantitative methods were Henry Schultz, a pioneer student of statistical demand curves, who taught the graduate courses in mathematical economics and mathematical statistics, and Paul Douglas who was (during the 1920s and 1930s) a leader in the estimation of and the measurement of real wages and living costs.

However, it is generally agreed that the progenitors of the Chicago School were Frank H. Knight and Jacob Viner. These two scholars shared an intense interest in the history of economic thought and both were, broadly speaking, devotees of neoclassical price theory. However, their intellectual styles and temperaments were quite different, and their personal relations were not close. Apart from his interest in the history of thought, Viner was primarily an applied theorist working on problems in international trade and related issues in monetary theory. Knight's work was focused on the conceptual underpinnings

of neoclassical price theory, and his main concerns were to clarify and improve its logical structure. Temperament and intellectual focus combined to make Knight a formidable critic, both of ideas and their protagonists. This led to a good deal of friction between him and both Douglas and Schultz. Personalities aside, Knight was strongly averse to the quantification of economics and was very outspoken on this, as on most other matters. (For further details, see Reder, 1982, pp. 362–5.)

By contrast, Viner was rather sympathetic to the aspirations of ‘quantifiers’, though sceptical of their prospects for success, at least in the near future. Viner's sympathy for quantitative work was prompted by the strong empirical bent of his own research, although friendship for Douglas and Schultz may also have been involved. On the other hand, Knight's purely theoretical studies of capital theory, risk, uncertainty, social costs, and so on, generated neither need for empirical verification nor exposure to research that might have offered it. As a result, Knight's relations with Douglas and Schultz were ridden with conflict, and theoretical disagreements with Viner spilled over into barbed comments to graduate students and kept personal relations (between Knight and Viner) from becoming more than merely correct (Reder, 1982, p. 365).

What Knight and Viner had in common was a continuing adherence to the main tenets of neoclassical price theory and resistance to the theoretical innovations of the 1930s, Monopolistic Competition and Keynes's *General Theory*. This theoretical posture paralleled an antipathy to the interventionist aspects of the New Deal and the full employment Keynesianism of its later years. Viner, who was actively consulting the government throughout the period, was much less averse to New Deal reforms than Knight and his protégés. However, there was a sharp contrast between the views of Knight and Viner, on the one hand, and those of avowed New Deal supporters such as Douglas, Schultz and some of the Institutionalists.

As a result of the division of faculty views, on both economic methodology and public policy, the graduate student body was exposed to a diversity of thought patterns and did not exhibit a great degree of conformity to any particular one. But despite their many disagreements, an effective majority of the Chicago faculty concurred in a set of degree requirements (for the PhD) that stressed competence in the application of price theory. These requirements were quite unusual in the 1930s and the process of satisfying them exercised a great influence in forming a (common) view of the subject among the students, in which price theory was of major importance.

The most important of the requirements was that all PhD candidates, without exception, pass preliminary examinations in both price theory and monetary theory. These examinations were difficult and attended with an appreciable failure rate. Even on second and third trials, there was a non-negligible probability of failure, with the result that some students were (and are) unable to qualify for the doctorate. For most students, the key to successful performance on the examinations was mastery of the material presented in relevant courses, especially the basic price theory course (301) and study of previous examinations.

For over half a century, the need to prepare for course and preliminary examinations, especially in price theory, has provided a disciplinary–cultural matrix for Chicago students. Examination questions serve as paradigmatic examples of research problems and ‘A’ answers exemplify successful scientific performance. The message implicit in the process is that successful research involves identifying elements of a problem with prices, quantities, and functional relations among them as these occur in price theory, and obtaining a solution as an application of the theory.

Although the specific content of examination questions has evolved with the development of the science,

the basic paradigm remains substantially unchanged: economic phenomena are to be explained primarily as the outcome of decisions about quantities made by optimizing individuals who take market prices as data with the (quantity) decisions being coordinated through markets in which prices are determined so as to make aggregate quantities demanded equal to aggregate quantities supplied.

Of course, students vary in the degree to which they assimilate price theoretic ideas to their thought processes, and resistance to these ideas was probably greater in the 1930s than later. Nevertheless, regardless of their special field of interest, all students were compelled to absorb and learn to use a considerable body of economic theory. In the 1980s these skills are very widespread, but in the 1930s they were rarely found and served to distinguish Chicago-trained PhD's – especially in applied fields – from other economists.

Despite the common elements of their training, as in other institutions, doctoral students tended to identify themselves with one or another particular faculty member, usually their dissertation supervisor. Thus each of the major figures in the department was associated with a cluster of advanced students. One such cluster, associated with Knight in the mid-1930s, became of very great importance in the history of the Chicago School. Key members of this cluster were Milton Friedman, George Stigler and W. Allen Wallis. The group established close personal relations with two junior faculty members, Henry Simons and Aaron Director, who were also protégés of Knight. Another member of the group was Director's sister, Rose, who later married Milton Friedman.

It was this group that provided the multigenerational linkage in intellectual tradition that is suggested by the term ‘Chicago School’. Although they admired Knight, and were devoted to him, the intellectual style of Friedman, Stigler, et al. was very different from Knight's. They were thoroughgoing empiricists with a distinct bias toward application of quantitative techniques to the testing of theoretical propositions. In their empirical bent and concern with ‘real world’ problems, they were much closer to Viner than to Knight, but, whatever the reason, they identified with the latter.

Partly because of his important role in the teaching of theory to undergraduates and (less well-prepared) beginning graduates, in the 1930s and until his untimely death in 1946, Henry Simons exercised an important influence on Chicago students. But he is remembered mainly for his essays on economic policy (collected in Simons, 1948) which constituted the principal statement of Chicago laissez-faire views during this period.

Simons's view had a distinctly populist flavour that is absent from those more recently associated with Chicago economics. For example, he favoured use of government power to reduce the size of large firms and labour unions. Where such policies would lead to unacceptable losses of efficiency (e.g. ‘natural monopolies’), Simons favoured outright public ownership. In sharp contrast to more recent Chicago statements on the matter, Simons emphatically supported progressive income taxation to promote a more egalitarian distribution of income (Simons, 1938).

Finally, Simons proposed a requirement of 100 per cent reserves against demand deposits and restriction of Federal Reserve discretion in monetary policy in favour of fixed rules designed to stabilize the price level (Simons, 1948). In this he was the direct forbear of Chicago monetarism, as later developed by Friedman and Friedman's students.

Historically, Friedman, Stigler and Wallis were both the intellectual and the institutional heirs of Knight and Viner. The story of Chicago economics would be less convoluted if the succession had been a matter of the older generation appointing their best students to succeed them. But it was not that simple. On the eve of World War II there was great concern, within the Economics Department and (probably)

in the central administration as well, that Chicago had none of the leading figures in the new theoretical developments of the period; that is, in nonperfect competition and Keynesian macroeconomics.

To rectify this, in 1938, they appointed Oscar Lange as assistant professor. In addition to his credentials as a contributor to the literature of Keynes's *General Theory*, especially its relation to general equilibrium theory, Lange was a leading participant in the current debate on the possibility of market socialism and its (alleged) advantages relative to laissez-faire capitalism in terms of efficiency. Further, he had made a number of contributions to mathematical economics and was able to provide backup support for Henry Schultz in that subject area, and in mathematical statistics as well.

As an outspoken and politically active socialist, Lange's views were diametrically opposed to laissez faire. That he managed to stay on friendly terms with virtually all of his colleagues was a testimonial both to his own tact and to their tolerance of dissent. Of course, it was no accident that the principal socialist in the Chicago tradition should have been a *market* socialist.

Within a few months of Lange's appointment, Henry Schultz was killed in an automobile accident and Lange became the sole mathematical economist in the Chicago department. Within a year the loss of Schultz was compounded by the partial withdrawal of Douglas from academic life to pursue a political career. Still further, with the outbreak of World War II, Viner became increasingly involved in Washington and, ultimately, in 1945, he resigned to accept an appointment at Princeton.

As a result of these losses, the Department had to be rebuilt. The process of reconstruction began during the war years, with Lange taking a leading role. He was very anxious to recruit colleagues who were leaders in current theoretical developments, especially in mathematical economics. Failing to obtain his first choice, Abba Lerner, he readily accepted Jacob Marschak and, for a short period, collaborated with the latter in making further appointments both to the Department and to the Cowles Commission, which had located at the University of Chicago in 1938. The collaboration ended abruptly in 1945 when Lange resumed Polish citizenship to become ambassador to the United States and, subsequently, to fill many other high positions in the socialist government of Poland.

During the war years, T.W. Schultz was attracted from Iowa State. A leading figure in agricultural economics, Schultz soon became chairman, a position from which he exercised much influence for over two decades. In addition to Schultz, in 1946 the Department acquired Lloyd Metzler to teach international trade and a number of younger theorists and econometricians associated mainly with the Cowles Commission. Whatever was the intention, these appointments served as a counterweight to the more or less contemporaneous appointments of Friedman (to the Economics Department) and Wallis (to the Business School).

There then ensued a struggle for intellectual pre-eminence and institutional control between Friedman, Wallis and their adherents on one side, and the Cowles Commission and its supporters on the other. The struggle persisted into the early 1950s, ending only with the partial retirement of Lloyd Metzler (due to ill health) and the departure of the Cowles Commission (for Yale) in 1953. While not monolithic, the Chicago economics department that emerged from this conflict had a distinctive intellectual style that set it apart from most others.

In positive economics, this style involves de-emphasizing the role of aggregate effective demand as an explanatory variable and stressing the importance of relative prices and ‘distortions’ thereof. In economic policy, it involves stressing the beneficial effects of allowing prices to be set by market forces rather than by government regulation. In an important sense, ‘Chicago economics’ in the 1950s and 1960s was simply an extension of the ideas of the Knight coterie of the 1930s. Indeed, some of the key

figures – notably Friedman, Stigler and Wallis – of that group were leading Chicago economists in the later period as well. Moreover, they were consciously concerned with explicating the continuity of the tradition and preserving it (see below).

The close personal relations of the members of the Knight coterie, maintained for over a half century, has reinforced the strong common elements in their idea-systems and made it easy to ignore the (important) points of disagreement, both among themselves and with others. As already mentioned, Friedman, Stigler and Wallis, like most Chicago economists of their own and subsequent cohorts, believe strongly in use of statistical data and techniques for testing economic theories. In this they differ from Knight, Simons, James Buchanan, Ronald Coase (1981) and a significant minority of other economists associated with Chicago, either as graduate students or faculty, who believe (on various grounds) that the validity of an economic theory lies in its intuitive appeal and/or its compatibility with a set of axioms, rather than in the conformity of its implications with empirical observation.

A second disagreement concerns the consistency of policy advocacy in any form, with the methodology applied in positive economics. (The most influential general description of this methodology is chapter 1 of Friedman, 1953.) This methodology recommends that explanations of economic behaviour be based on a model of (individual) decisions of resource allocation (among alternative uses) designed to maximize utility subject to the constraints of market prices and endowments of wealth. Market prices are presumed to be set so as to equate quantities supplied with those demanded, for all entities traded.

As traditionally applied by neoclassical economists with a predilection for laissez faire, this methodology coexists with advocacy of government policies designed to promote that objective. But in the late 1960s one group of Chicago economists led by Stigler (who had returned to Chicago in 1958 as Walgreen Professor in both the Economics Department and the Business School) began to apply the tools of economic analysis to the investigation of the determinants of political activity, especially government intervention in resource allocation. Thus study of the regulatory and taxing activities of the state became directed not simply at demonstrating their adverse effects upon economic efficiency, but primarily to explaining their occurrence as an outcome of the operation of ‘political markets’ for such activities.

So analysed, interventions traditionally viewed as efficiency impairing, such as tariffs, require reinterpretation. An individual's resources include not only his command over goods and services acquired through conventional markets, but also his political influence (however measured). Government interventions are considered to be endogenous outcomes of a political-economic process, reflecting the political as well as the economic wealth of decision making units, and not as aberrations of an exogenous state (e.g. see Stigler, 1982). So viewed, criticism of political outcomes is no more warranted than criticism of the expenditure behaviour of sovereign consumers; both are outcomes of the free choice of resource owners.

This is not to suggest that the ‘political economy’ wing among Chicago economists has become indifferent to laissez faire. On the contrary, opposition to government intervention (e.g. regulation) among Stigler and his allies is quite as strong as it ever has been. During the past decade many economists and lawyers at some time affiliated with the Law and Economics group at Chicago have been prominent advocates of deregulation. However, tension between advocacy of reform, and positive analysis of the political process through which reform must be achieved, presents a continuing existential problem to the heirs of the Chicago tradition. Although they are well aware of the problem, thus far they have refrained from divisive dispute and treat exercises in political advocacy as a

consumption activity by those engaged.

Political science is only one of the fields into which Chicago economics has expanded during the past quarter century. Beginning in the early 1940s and accelerating in the last two decades under Richard Posner's leadership, the economic analysis of legal institutions has become an important area of research both for economists and for legal scholars. Further, using the theory of labour supply as a point of departure, the economic analysis of the family has become an important part of the study of population, marriage, divorce and family structure. This development has challenged sociological and psychological modes of explanation in fields that had long been considered provinces of these other disciplines. Still further, the theory of human capital has had a major impact on the study of education.

It is convenient to date the ‘disciplinary imperialist’ phase of the Chicago School as beginning in the early 1960s and continuing to the present. However, its roots go back into the 1930s; since that time there has been, at least in the oral tradition, a tropism for application of the tools and concepts of price theory to (seemingly) alien situations, and for taking delight in confronting conventional wisdom with the results. Correlatively, there has been a strong tendency to resist explanations of behaviour that do not run in terms of utility maximization by individual decision-makers coordinated by market clearing prices. However, until well into the 1950s, the disciplinary imperialist aspect of the Chicago paradigm was overshadowed by the struggle to defend the integrity of neoclassical price theory from the attacks of Keynesians at the macro level and the attempts of various theorists of nonperfect competition to provide alternatives at the micro level. The counterattack on the *General Theory* produced a revival of neoclassical monetary theory in a refined and empirically implemented form; this revival is associated with the work of Milton Friedman (1956).

The struggle to re-establish the competitive industry as the dominant model for explaining relative prices was led by Stigler (1968, 1970), and generated much of the theoretical and empirical literature of the field of Industrial Organization. Both in Industrial Organization and Money-Macro, the earlier debates continue, with Chicago-based participants being identifiable as partisans of the standpoints of Friedman and Stigler a quarter of a century ago. However, in the 1970s and 1980s the topics related to these debates have been forced to share centre stage with newer subjects.

The expansion of Chicago economics beyond the traditional boundaries of the discipline began in the middle and late 1950s; two early examples were H.G. Lewis's application of price theory to the ‘demand and supply of unionism’ (Lewis, 1959) and Gary Becker's dissertation on racial discrimination (Becker, 1957). These were followed in the 1960s and 1970s by a number of others, as already mentioned. Many of these are more or less straightforward applications of conventional price theory to new problems. However, the analysis of time as an economic resource (Becker, 1965) has led to important improvements in the theory of household behaviour.

The analysis of time is also related to a methodological tendency to reject differences in tastes (including attitudes, opinions and beliefs in ‘tastes’) as a source for explanations of cross-individual differences in behaviour (Stigler and Becker, 1977; Becker, 1976). The rejection is based on the contention that (1) seeming differences of taste are usually reducible to differences of cost and (2) statements about cost differences are much more amenable to empirical test. While this methodological principle has met with resistance, at Chicago as elsewhere, it is reflected in a great deal of ongoing research, especially where cost of time is an important variable.

A separate path of disciplinary expansion has arisen in the field of Finance. Whether, prior to the 1960s,

this field was a province of Economics, is a point that it is convenient to bypass. But unquestionably, prior to the theoretical developments initiated by Modigliani and Miller's famous paper (1958) on the (non) relation of stock prices and dividends, the theory of price. Subsequent developments have completely reversed that situation, so that in the mid-1980s, the ‘capital asset pricing model’ has become an integrating matrix for the theories of security prices, asset structure of the firm, and, via the study of executive compensation, wages.

The dominant idea underlying these developments is that, save for transaction costs, *on average* no opportunity for arbitrage gains goes unexploited. One implication of this is the proposition that there is ‘no free lunch’; another implication is that no specifiable algorithm can be found that will enable a resource owner to utilize publicly available information to predict movements of asset prices well enough to gain by trading. The latter implication is tantamount to the ‘hypothesis of efficient markets’. While not formally identical with rational expectations, efficient markets will support any behaviour conforming to rational expectations, but will be compatible with other models of expectations only where one or another set of correlated forecast errors (across individuals) is assumed. Moreover, so long as expectations are rational, and regardless of how they are generated, there is no way in which variables operating through expectations can improve upon the neoclassical explanation of relative prices and quantities. This obviates any need for augmenting economic theory by variables reflecting psychological or sociological factors that operate upon individual decision-making via expectations. Obviously, such a theory of expectations is strongly supportive of the claims of economic theory in interdisciplinary competition.

The interrelated ideas of rational expectations and efficient markets originated at Carnegie-Mellon in the work of Muth (1961) and Modigliani and Miller (1958) rather than at Chicago. However, their consonance with the Chicago paradigm is such that they have found a home in the Chicago Business School under the leadership of Miller and his students, and (since the mid-1970s) in the Economics Department under Robert Lucas, rather than in their place of origin. While the claim of Chicago to be the primary locus for research in these fields is a strong one, it is a claim more subject to challenge than analogous claims in some other fields.

Yet a third Chicago innovation of the late 1950s is the ‘Coase Theorem’ (Coase, 1960). In essence this theorem states that, ignoring transaction costs, if there is any reallocation of goods, claims, rights (especially property) or alteration of institutions that – after making compensating side payments to losers – increases the utility of everyone, said reallocation will occur. If rationality is a maintained hypothesis and transaction costs are negligible, the theorem becomes a tautology. Thus the empirical content of the theorem will vary inversely with the importance attributed to transaction costs, which serve as a conceptual receptacle for all forces bearing upon decision-making other than those explicitly incorporated in the theory of price. To consider the Coase Theorem empirically important is to believe that transaction costs and departures from rationality are unimportant.

Put differently, the Coase Theorem suggests that the real world tends towards a position of Pareto optimality. Of course, for given tastes and technology, there may be a different Pareto optimum for each distribution of wealth. Therefore, to the extent that the distribution of wealth is exogenous and has important behavioural consequences, the predictive implications of both Pareto optimality and the Coase Theorem are less salient. Thus the rise in influence of the Coase Theorem at Chicago has more or less paralleled a decline in the marked concern with income distribution that existed in the 1930s and 1940s, especially in the work of Henry Simons (Reder, 1982, p. 389).

When objects of exchange are taken to include legislation and other political variables, the Coase Theorem strongly suggests that the forces of decentralized decision-making that govern production and exchange also control changes in laws and institutions. Thus belief in the Coase Theorem is – or should be – conducive to political passivity. Nevertheless, not all Chicago economists are politically quiescent. But with few exceptions, they are generally conservative, though with considerable differences of shading and intensity of belief, and in taste for political controversy. Probably these differences parallel differences in the degree to which they accept economic explanations of political behaviour. Perhaps the most common characteristic of Chicago economists is distrust of the state. This distrust, together with the belief that, given time, voluntary exchange will usually generate truly desirable reforms, acts as a powerful brake on wayward impulses to improve society through political action.

The saga of the Chicago School is at once the story of the evolution of a set of ideas – a paradigm – and of a particular institution with which its leading protagonists have been associated. In this essay I have emphasized certain central theoretical ideas and historical events to the exclusion of detailed coverage of applied work and mention of the individuals responsible for it. However, it is the association of these central ideas with an identifiable, multigenerational group of individuals located at a particular institution that justifies the title of this article. Many of the key individuals in this history – Director, Friedman, Stigler, Wallis – are still alive, intellectually active and in close touch with their successors on the Chicago faculty. This continuity, both of personalities and ideas, is a distinctive feature of the intellectual tradition called the Chicago School.

In the mid-1980s the vitality of this tradition is threatened more by the growing acceptance of many of its key ideas than by resistance to them. A quarter century ago, Chicago economics was distinguished by its emphasis on the importance of competition and money supply. Arguably, in 1985, these views and their extensions have become mainstream economics, leaving the story of the Chicago School as a nearly closed episode in the history of economic thought. While such an argument may prove valid, it is too soon to tell.

See Also

- Chicago school (new perspectives)
- Coase, Ronald Harry
- Douglas, Paul Howard
- Friedman, Milton
- Knight, Frank Hyneman
- Lange, Oskar Ryszard
- Laughlin, James Laurence
- Metzler, Lloyd Appleton
- Schultz, Henry
- Schultz, T.W.
- Simons, Henry Calvert
- Stigler, George Joseph
- Viner, Jacob

Bibliography

Becker, G.S. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–517.

Becker, G.S. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Coase, R.H. 1981. *How should economists choose?* Washington, DC: American Enterprise Institute for Public Policy Research.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Friedman, M. 1956. *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Lewis, H.Gregg 1959. Competitive and monopoly unionism. In *The Public Stake in Union Power*. ed. P. D. Bradely. Charlottesville: University of Virginia Press.

Modigliani, F. and Miller, M.H. 1958. The cost of capital, corporation finance and the theory of investment. *American Economic Review* 48, 261–97.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Reder, M.W. 1982. Chicago economics: permanence and change. *Journal of Economic Literature* 20(1), 1–38.

Simons, H.C. 1938. *Personal Income Taxation*. Chicago: University of Chicago Press.

Simons, H.C. 1948. *Economic Policy for a Free Society*. Chicago: University of Chicago Press.

Stigler, G.J. 1968. *The Organization of Industry*. Homewood, III: Richard D. Irwin.

Stigler, G.J. 1982. *Economists and public policy*. Washington, DC: American Enterprise Institute for Public Policy Research.

Stigler, G.J. and Becker, G.S. 1977. De gustibus non est disputandum. *American Economic Review* 67 (2), 76–90.

Stigler, G.J. and Kindahl, J.K. 1970. *The Behavior of Industrial Prices*. New York: Columbia University Press for the National Bureau of Economic Research.

How to cite this article

Reder, M. W. "Chicago School." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000127> doi:10.1057/9780230226203.0226

The New Palgrave Dictionary of Economics Online

child health and mortality

Janet Currie
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Child health is a major indicator of the direction and well-being of society. It is a significant factor predicting health and productivity in adult life, and the health of adults in turn affects the well-being of the next generation of children. The most important outstanding issues include determining the most cost-effective investments in child health, explaining the relationship between health and socio-economic status over the life course, and finding the interventions that are most effective in breaking the inter-generational cycle of ill health and poverty. As children are economic actors in their own right, their well-being is worthy of study.

Keywords

anthropometrics; asymmetric information; child health and mortality; education; family economics; fertility in developed countries; fertility in developing countries; health care; health economics; health insurance; human capital; mortality; pollution; poverty alleviation programmes; productivity; public health; risk

Article

Child health and mortality are of interest to economists for three reasons. First, they are important indicators of the success or failure of government policy. Second, children's health has long-term impacts on their health and productivity as adults. Third, there is increasing recognition that children are economic actors in their own right. Hence, their well-being is worthy of study.
The most common model of child health is one in which health is ‘produced’ by families using health ‘inputs’ (Grossman, 2000). Examples of inputs include the goods and services families buy to improve child health. Families maximize an inter-temporal utility function subject to the production function, prices, and budget constraints. Inputs are valued only because of their effect on health. Children start with a ‘health endowment’ that depreciates over time in the absence of health inputs. Public policy

affects either the price of inputs or the form of the production function. The model predicts that child health will be influenced by the price of health inputs. The inter-temporal nature of the model highlights the idea that health inputs are investments with long-term payoffs.

Studies of children in developing countries often focus on the ‘production’ of mortality rates, nutrient intakes, height, weight and other objective measures. In contrast, studies of children in richer countries often focus on the utilization of medical care. But health care is only one input into the production of child health, and it is not the most important. Improvements in standards of living, advances in knowledge about disease and hygiene, and public health measures such as improved sanitation have done more to improve child health in the past 150 years than even the most spectacular advances in personal medical care (Preston, 1977). Today, accidents and violence, rather than disease, are the major killers of young children in wealthy countries after the first year of life (UNICEF, 2001).

Measures of child health

Health is multidimensional and difficult to measure. Mortality and parent-reported health fall at two ends of a spectrum. Mortality is an objective but narrow measure. In countries with high death rates, child mortality is a relatively sensitive indicator of economic and social conditions. For example, in Zimbabwe mortality among children under five years old increased from 80 to 126 per 1,000 live births between 1990 and 2003 as the economic crisis deepened (United Nations Common Database). In countries with lower child mortality rates, the relationship between economic conditions and mortality may be masked by the effects of economic cycles on fertility. For example, some recent papers demonstrate that in developed countries poorer people have fewer children during economic downturns so that the average health of infants increases (see, for example, Lleras-Muney and Dehejia, 2004). The relationship between mortality and economic conditions is also masked by strong underlying downward trends in mortality due to technological advances.

A typical survey question eliciting parent reports about child health asks respondents to rate child health on a scale of 1 to 5. An advantage of this measure is that it applies to all children. A disadvantage is that parent reports may be biased. For example, sick parents are more likely to report sick children. Parents are also often asked about limitations on children's activities (for example: Did a health problem prevent school attendance?) and about the presence of chronic conditions. These questions have the advantage of being more specific, but capture only one dimension of health and also suffer from potential biases (Baker, Deri and Stabile, 2004; Strauss and Thomas, 1996).

In between are anthropometric measures such as birthweight, height, weight, height for age, and body mass index (Martorell and Habicht, 1986). Anthropometrics are objective measures that apply to large numbers of children. But, like mortality, they may not be sensitive measures in healthy populations. For example, American children are unlikely to be stunted (low height for age) and are increasingly likely to survive low birthweight (less than 2,500 grams) without significant impairments. American children are increasingly likely to be obese, however, suggesting that body mass index is likely to become a more important health indicator in the future.

A fourth class of measures involve ‘risky behaviours’ such as precocious or dangerous sexual activity, involvement in crime or victimization, use of handguns, and use of alcohol, tobacco, and illegal drugs. Given the importance of accidents and violence among children, these are important questions. But the

stigma associated with these activities makes it likely that they will be under-reported. Also, risky behaviours may or may not lead to poorer health. Unfortunately, the actual health effects of many behaviours are very poorly reported. For example, there is little information available about injuries that do not lead to deaths.

Some surveys include clinical assessments of children's health by doctors or other trained professionals in addition to some of the information about economic status that is usually collected in social surveys. Examples include the British birth cohort studies, the American National Health and Nutrition Examination Surveys, the World Bank's Living Standards and Measurement Surveys and the Indonesia Family Life Survey. Some of the most interesting work being done in this area involves measures of children's genetic make-up. Caspi et al. (2002) show, for example, that New Zealand men with a specific genetic marker were more likely to be violent adults, but only if they had been maltreated as children. Given the broad range of health outcomes, researchers should look at a range of outcomes and carefully consider whether the chosen ones are likely to be affected by the phenomena under study.

Health care utilization

The human capital model makes a clear distinction between health and health inputs. In the model, parents care about health rather than health inputs. Yet this distinction is often blurred. Williams and Miller (1992, p. 991) state that ‘One of the most impressive aspects of health policy implementation [in Europe is] that the programs were put in place not because of extensive documentation on cost effectiveness, but out of a value system that cherishes equity in health care.’ The underlying assumption is that all health care produces health. Yet the market for health care is plagued with imperfections. Some care is likely to be superfluous, for consumption rather than investment purposes, or even injurious.

Models of physician-induced demand show that asymmetric information can lead to excessive consumption of medical services if physicians take advantage of their superior information to ‘sell’ services that patients do not need (Pauly, 1980; Dranove, 1988). There may be considerable scope for inducement in the market for children's health care. Many child treatments are inexpensive but have a high clinical value when they are warranted, so parents perceive a low cost set against a potentially high benefit. The availability of insurance compounds the problem by further reducing costs to parents. Researchers should focus on measures of utilization that have a clear benefit. Whether or not a child visited a doctor in a year and whether a child is immunized are good examples. Measures such as the number of hospitalizations are problematic since many hospitalizations could be prevented with appropriate outpatient care. Some recent work focuses on ‘preventable hospitalizations’ as a measure of inadequate utilization of care (Casanova and Starfield, 1995).

Health as an investment

Child health affects adult health. Poor health in childhood also lowers future utility through its effects on future wages and labour force participation (Currie and Madrian, 1999) and through its effects on schooling. Currie (2005) provides a survey of literature linking several specific health conditions to cognitive outcomes and schooling achievement.

Using data from the 1999 Panel Study of Income Dynamics, James Smith (2005) shows that a retrospective self-reported question about health during childhood is remarkably predictive of future outcomes. Comparing siblings, he finds that those who were in excellent or very good health earn 25 per cent more as adults. Currie (2000) surveys some of the many studies that find positive associations between cognitive test scores and anthropometric measures of health such as birthweight, weight, height, head circumference, and the absence of abnormalities in children of various ages. More recently Currie and Moretti (2005) have shown that differences in birthweight between sisters are predictive of differences in education and median income in the zip code of residence at the time the sisters deliver their own children many years later.

But low birthweight is only one of a number of health shocks that low-income children are more likely to experience (Newacheck, Hughes and Stoddard, 1996). Case, Lubotsky and Paxson (2002) show that the gap in health status between rich and poor US children widens as children age. Currie and Stabile (2003) replicate this finding using Canadian data, and argue that the widening gap reflects the greater frequency of negative health shocks among poor children. The comparison between the United States and Canada suggests that public health insurance is not sufficient to shield children from the negative health consequences of poverty (since Canada has universal insurance). However, in Britain the gap between rich and poor children is smaller than in North America and does not widen as children age (Currie, Shields and Price, 2004). This suggests that some other aspect of the social safety net may be responsible for protecting child health in Britain.

Poor children are more likely than rich children to suffer from mental health problems (Currie, 2005, 2002). Mental health problems account for the largest share of days lost due to health problems in the United States. Many mental health conditions have their roots in childhood, but the relationship between mental health and child outcomes has been largely ignored in economics. Currie and Stabile (2005) investigate the relationship between symptoms of Attention Deficit Hyperactivity Disorder Activity disorder (ADHD) and educational attainment using US and Canadian panel data. We find large negative effects even in rich sibling-fixed effects models. Other research has shown that childhood behaviour problems predict negative future outcomes (cf. Gregg and Machin, 1998). The prevalence and potential economic importance of child mental health problems suggest that more work is warranted.

Policy and child health

It is easy to justify government intervention in the market for health care. In addition to asymmetric information between patients and providers, there are other informational problems. For example, imperfect information in the market for insurance can lead to market failure. And although parents make most decisions about child health inputs, these decisions have consequences for society. Parents who do not take account of externalities may not provide the optimal level of care for their children (cf. Kremer and Miguel, 2004). Finally, the health sector accounts for a large and growing share of the economy, and the government is already the major player in the health care markets in most countries, including the United States.

Policies can be divided into those that intervene in the market for health care and those that affect health through other means. Public health insurance is the most prominent example in the first category. It is

difficult to study the impact of universal health insurance because there is only a single ‘before/after’ comparison. But over the late 1980s and early 1990s, the United States greatly expanded its public health insurance coverage of pregnant women and children. Forty per cent of US births are now covered by public insurance. The expansion took place at an uneven rate across states, yielding a potential source of identification.

The effects of this expansion of insurance coverage are surveyed in Gruber (2003). It reduced infant and child mortality, increased utilization of preventive care, and reduced preventable hospitalizations among children. But increases in coverage also increased the inappropriate use of care (for example, increased rates of Caesarean section). And some who took up public health insurance would have had private health insurance in the absence of the expansions. Hence, public health insurance improves child health, but does not necessarily result in efficient service delivery.

Health care utilization is only one input into health production. Other inputs such as a healthy lifestyle and the avoidance of injury are arguably much more important. Government policy has a large role to play in affecting many health inputs beyond health care. A few examples follow.

Pollution is likely to be more harmful to children than to adults both because they are still developing and because of their small size. Hence, any policy that affects the environment may affect child health. For example, Chay and Greenstone (2003) show that the recession of the early 1980s reduced infant mortality. Currie and Neidell (2005) show that reductions in carbon monoxide pollution in California over the 1990s (largely due to cleaner vehicles) saved at least 1,000 infant lives.

Child obesity is a growing problem that threatens future health. The potential role for government ranges from the provision of information (for example, revising the ‘healthy food pyramid’ to reflect the most recent nutritional knowledge) to regulation (for example, eliminating Coke machines in schools). The government plays a similar role with respect to discouraging children from using alcohol and tobacco, though in these examples government also directly controls the price of the products through taxation. A good deal of research documents the relationships between prices, advertising, and youth consumption of tobacco and alcohol. But we know much less about the effectiveness of newer policies aimed at curbing obesity (see Gruber, 2001).

Although injuries remain a major cause of death, the incidence of accidental death has declined dramatically since the 1970s, especially in the United States (UNICEF, 2001). Glied (2001) argues that the decline is due to improvements in education resulting in increased use of, for example, bicycle helmets and seat belts. But many products, including cars, cribs, and medicine bottles, are much safer than they used to be. Is this a result of random technical innovation, government mandates, or fear of lawsuits? Similarly, trauma care has improved greatly. So there are many possible explanations for the reduction in mortality.

While health affects education, maternal education affects child health. Currie and Moretti (2003) find that increases in the availability of colleges increased women's education, leading to better infant health outcomes. Hence, there is an inter-generational payoff to government investments in education that leads to ‘increasing returns’ to investments in education (Rosensweig and Wolpin, 1994).

Finally, as discussed above, poor children are more likely than rich children to suffer virtually all forms of health insult. Hence, improving health is a goal of general poverty alleviation programmes such as public housing and income maintenance.

Summary

Child health is an important indicator of the direction and well-being of society. Health in childhood is one of the more important factors predicting health and productivity in adult life, and the health of adults will in turn affect the well-being of the next generation of children.
Many policies have impacts on child health. Some simple improvements in data collection efforts could have a large research payoff in terms of identifying these impacts. These include: allowing the release of geographical identifiers so that health data can be merged to other data; the inclusion of family income and demographics in health data-sets; and the collection of more objective measures of child health.
What are the most interesting outstanding questions? First, what are the most cost-effective investments in child health? Second, what explains the relationship between health and socio-economic status over the life course? And third, what interventions are most effective in breaking the inter-generational cycle of ill health and poverty?

See Also

- family economics
- fertility in developed countries
- fertility in developing countries
- health economics
- household production and public goods
- human capital

Bibliography

Baker, M., Deri, C. and Stabile, M. 2004. What do self-reported, objective measures of health measure? *Journal of Human Resources* 39, 1067–93.

Casanova, C. and Starfield, B. 1995. Hospitalizations of children and access to primary care: a cross-national comparison. *International Journal of Health Services* 25, 283–94.

Case, A., Lubotsky, D. and Paxson, C. 2002. Economic status and health in childhood: the origins of the gradient. *American Economic Review* 92, 1308–34.

Caspi, A. et al. 2002. Role of genotype in the cycle of violence in maltreated children. *Science* 297, 851–4.

Chay, K. and Greenstone, M. 2003. The impact of airpollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. *Quarterly Journal of Economics* 118, 1121–67.

Currie, A., Shields, M. and Price, S. 2004. Is the child health/family income gradient universal? Evidence from England. Discussion Paper No. 1328. Bonn: IZA.

Currie, J. 2002. Child health in developed countries. *Handbook of Health Economics*, ed. J. Newhouse and A. Culyer. Amsterdam: North-Holland.

Currie, J. 2005. Health disparities and gaps in school readiness. *The Future of Children* 15(1), 117–38 (issue on *School Readiness: Closing Racial and Ethnic Gaps*).

Currie, J. and Madrian, B. 1999. Health, health insurance and the labor market. In *Handbook of Labor Economics*, ed. D. Card and O. Ashenfelter. New York: North-Holland.

Currie, J. and Moretti, E. 2003. Mother's education and the intergenerational transmission of human capital: evidence from college openings. *Quarterly Journal of Economics* 118, 1495–532.

Currie, J. and Moretti, E. 2005. *Biology as destiny: short and long-run determinants of intergenerational correlations in birth weight*. Working paper No. 11567. Cambridge, MA: NBER.

Currie, J. and Neidell, M. 2005. Air pollution and infant health: what can we learn from California's recent experience? *Quarterly Journal of Economics* 120, 1003–30.

Currie, J. and Stabile, M. 2003. Socioeconomic status and health: why is the relationship stronger for older children? *American Economic Review* 93, 1813–23.

Currie, J. and Stabile, M. 2005. Child mental health and human capital accumulation: the case of ADHD. Working Paper No. 10435. Cambridge, MA: NBER, 2004, updated August 2005.

Dranove, D. 1988. Demand inducement and the physician/patient relationship. *Economic Inquiry* 26, 281–99.

Glied, S. 2001. The value of reductions in child injury mortality in the US. In *Medical Care Output and Productivity*, ed. D. Cutler and E. Berndt. Chicago: University of Chicago Press.

Gregg, P. and Machin, S. 1998. *Child development and success or failure in the youth labour market*. Discussion Paper No. 0397. London: London School of Economics.

Grossman, M. 2000. The human capital model. In *Handbook of Health Economics*, vol. 1a, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Gruber, J. 2001. *Risky Behaviors Among Youths*. Chicago: University of Chicago Press.

Gruber, J. 2003. Medicaid. In *Means-Tested Transfer Programs in the United States*, ed. R. Moffitt. Chicago: University of Chicago Press.

Kremer, M. and Miguel, E. 2004. Worms: identifying impacts on education and health in the presence of externalities. *Econometrica* 72, 159–217.

Lleras-Muney, A. and Dehejia, R. 2004. Booms, busts, and baby's health. *Quarterly Journal of Economics* 119, 1091–130.

Martorell, R. and Habicht, J.-P. 1986. Growth in early childhood in developing countries. In *Human Growth: A Comprehensive Treatise*, ed. F. Faulkner and J. Tanner. New York: Plenum Press.

Newacheck, P., Hughes, D. and Stoddard, J. 1996. Children's access to primary care: differences by race, income, and insurance status. *Pediatrics* 97(1), 26–32.

Pauly, M. 1980. *Doctors and their Workshops*. Chicago: University of Chicago Press.

Preston, S. 1977. Mortality trends. *Annual Review of Sociology* 3, 163–78.

Rosenzweig, M. and Wolpin, K. 1994. Are there increasing returns to the intergenerational production of human capital? Maternal schooling and child intellectual development. *Journal of Human Resources* 29, 670–93.

Smith, J. 2005. *The Impact of SES on Health over the Life-Course*. Santa Monica, CA: RAND.

Strauss, J. and Thomas, D. 1996. Measurement and mismeasurement of social indicators. *American Economic Review* 86(2)30–4.

UNICEF. 2001. *A League Table of Child Deaths by Injury in Rich Nations*. Innocenti Report Card, Issue No. 2. Florence: Innocenti Research Center.

United Nations Common Database. Online. Available at http://unstats.un.org/unsd/cdb/cdb_help/cdb_quick_start.asp, accessed March 2005.

Williams, B. and Miller, A. 1992. Preventive health care for young children: findings from a 10-country study and directions for United States policy. *Pediatrics* 89, S983–S998.

How to cite this article

Currie, Janet. "child health and mortality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000553> doi:10.1057/9780230226203.0228

The New Palgrave Dictionary of Economics Online

child labour

Kaushik Basu
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

According to latest available estimates, somewhere between 14 to 18 per cent of all children between the ages of 5 and 14 years in the world are labourers. The causes of child labour are many but the primary one is poverty, since for most parents sending children to work is an act of desperation. The availability of decent schools and the provision of small incentives, such as school meals, can help limit child labour. Hence, the best policy response is to improve conditions on the adult labour market, provide better schooling and, on rarer occasions, use legal interventions.

Keywords

child labour; education; household production; Industrial Revolution; poverty

Article

According to the International Labour Organization (ILO, 2002) there were 186 million child labourers in the world in 2000, that is, children between the ages of 5 and 14 years doing regular economic work. This implies a ‘participation rate’ (the number of labouring children as percentage of all children of that age group) of 15.5 per cent. Of these, 111 million were engaged in ‘hazardous work’. But by 2004 the number was down to 166 million – a participation rate of 13.7 per cent – and the number of children in hazardous work was down to 74 million. Some details and regional distribution estimates are available in Hagemann et al. (2006), but (at the time of writing) these new numbers are yet to be absorbed and analysed.

It is a truism that the incidence of child labour is hard to estimate, both because it is often illegal and so respondents would not proffer information too readily and because the work is usually in the informal sector where record keeping is weak. Not surprisingly, there are other estimates of child labour, higher and lower. According to the UNICEF (2006), which collates data from different sources from 1998 to 2004, the participation rate is 18 per cent.

These data sources have both upward and downward biases along different dimensions. Domestic work that is done in one's own household is usually recorded very poorly or not at all. But we have micro evidence that in poor regions children, especially girls, do huge amounts of work in their homes, ranging from fetching wood to hazardous work like cooking over open fires. Indirect evidence for this comes from the gender breakdown of child labour. According to ILO data, boys do more labour than girls; their participation rates are respectively 15.9 per cent and 15.2 per cent. But detailed micro studies that try to include heavy domestic work, such as that by Cigno and Rosati (2005, ch. 5), show that girls tend to do 30 per cent more work than boys. Hence, there is a downward bias in the macro numbers mentioned above.

On the other hand, one source of upward bias comes from ‘work’ being equated with doing more than one hour of work in the ‘reference period’, and from the fact that for most studies the reference period is one week. It is arguable that children who answer ‘yes’ because they barely satisfy that cut-off ought not to be classified as child labourers.

The reason for not becoming too weighed down by these statistical debates is that, no matter how one measures it and, as a consequence, whether the participation rate turns out to be 14 per cent or 18 per cent, it is easy to agree that the incidence of child labour is unacceptably high. In a world with as much opulence as ours there should not be so many children working and that too in grinding poverty and in intolerable working conditions.

This raises the question of the causes of child labour and the appropriate policy response. The primary cause is poverty. Well-off parents living in the same nation and under the same laws as poor ones almost never send their children to work. Hence, a child's non-work (whether this be leisure or schooling) is a luxury good. Sufficiently poor parents cannot afford this. This was called the ‘luxury axiom’ in Basu and Van (1998), and there is ample empirical evidence for it (see discussion in Ray, 2000; Basu and Tzannatos, 2003; Edmonds and Pavnick, 2005). But there are other causes as well. There are parents on the borderline of poverty, who, if they knew that there were decent schools in the area and/or that their children would get a square meal in school, would take the children out of labour and send them to school. Hence, the provision of schooling and, ideally, having some added incentives for sending children to school can make a large difference to the incidence of child labour (Ravallion and Wodon, 2000; Bourguignon, Ferreira and Leite, 2003).

The presence of other determinants is also evident from the fact that the location of a child in the rural–urban spectrum affects the probability of the kind and amount of work the child is likely to do. This was always believed to be true. There were commentators at the time of the Industrial Revolution in Britain who argued that the alleged increase in child labour was really not an increase but a shift of child labour from agriculture to industry and a dramatic change in the *nature* of work (see Horrell and Humphries, 1995, for discussion). Contemporary, casual evidence seems to support this. And a recent empirical study of child labour in Nepal (Fafchamps and Wahba, 2006) formally confirms for the first time that urban proximity matters in a significant way. Children who live in or close to cities participate significantly less in labour and have a higher incidence of schooling than their rural counterparts. The health effects of these two kinds of child labour – agrarian and industrial – remain to be investigated systematically. Work in factories can be in dark and dank settings; on the other hand, agricultural work can mean exposure to not just the elements but also to pesticides and fertilizers. The net effects of these deserve investigation.

Given the multiplicity of causes, one has to be careful about the policy response to child labour. It is no surprise that, despite attempts by the British government from 1802 till the mid-19th century to deter child labour through a series of Factory Acts, the participation rate remained consistently and intolerably high. Indeed, the participation rates in Britain in the first half of the 19th century were higher than those found in today's China or India. Likewise in the USA, despite a variety of legislative measures starting in 1837 in Massachusetts, the incidence of child labour remained high and in fact continued to rise till the end of the 19th century.

While there is no final word on policy, we know that some measures are likely to be more effective than others. Ameliorating poverty, improving adult labour market conditions and providing better schooling, as already discussed, can have a significant effect. The law – bans and fines – can also play a role but should be used with caution and after empirical tests of whether the context deserves such measures. It has been argued (see Basu and Van, 1998; Dessy and Pallage, 2001; Emerson and Souza, 2003) that the labour market can in different ways (such as the general equilibrium impact on market wages, coordination with technology and intergenerational dynamics) give rise to multiple equilibria. That is, the market, left to itself, can settle into different grooves; for instance, one with no child labour and another with a high participation rate. In such a case, if the economy settles into the latter equilibrium, a ban can be an effective tool. Otherwise a ban can lead children labouring in factories to worse outcomes, such as starvation or prostitution. Minimally, in such situations the law has to be combined with complementary interventions to ward off the extreme poverty and deprivation that can arise as a side effect of its implementation.

See Also

- childcare
- education in developing countries
- labour economics
- poverty

Bibliography

Basu, K. and Tzannatos, Z. 2003. The global child labor problem: what do we know and what can we do? *World Bank Economic Review* 17, 147–74.

Basu, K. and Van, P. 1998. The economics of child labor. *American Economic Review* 88, 412–27.

Bourguignon, F. Ferreira, F. and Leite, P. 2003. Conditional cash transfers, schooling and child labor. *World Bank Economic Review* 17, 229–54.

Cigno, A. and Rosati, F. 2005. *The Economics of Child Labor*. Oxford: Oxford University Press.

Dessy, S. and Pallage, S. 2001. Child labor and coordination failures. *Journal of Development*

Economics 65, 469–76.

Edmonds, E. and Pavnick, N. 2005. Child labor in the global economy. *Journal of Economic Perspectives* 19(1), 199–220.

Emerson, P. and Souza, A. 2003. Is there a child labor trap? *EconomicDevelopment and Cultural Change* 51, 375–98.

Fafchamps, M. and Wahba, J. 2006. Child labor, urban proximity, and householdcomposition. *Journal of Development Economics* 79, 374–97.

Hagemann, F., Diallo, Y., Etienne, A. and Mehran, F. 2006. *Child Labour Trends: 2000 to 2004*. Geneva: ILO.

Horrell, S. and Humphries, J. 1995. ‘The exploitation of little children’: child labour and the family economy in the Industrial Revolution. *Explorations in Economic History* 32, 485–516.

ILO (International Labour Organization). 2002. *Every Child Counts: New Global Estimates on Child Labor*. Geneva: ILO.

Ravallion, M. and Wodon, Q. 2000. Does child labor displace schooling? *Economic Journal* 110, 158–75.

Ray, R. 2000. Child labor, child schooling and their interaction with adult labor. *World Bank Economic Review* 14, 347–67.

UNICEF. 2006. *State of the World's Children 2006*. New York: UNICEF.

How to cite this article

Basu, Kaushik. "child labour." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000532> doi:10.1057/9780230226203.0229

The New Palgrave Dictionary of Economics Online

Child, Josiah (1630 – 1699)

Douglas Vickers
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bullion; Child, J.; East India Company; Interest rates; mercantilism; transferable bills of exchange

Article

The second son of Richard Child, a London merchant, Sir Josiah Child was born in 1630 and enjoyed a highly successful merchant career during which he amassed a considerable fortune. His business ventures, which included the provisioning of Navy ships, led to his appointment as Deputy to the Navy's Treasurer at Portsmouth in 1655 and he became Mayor of that city in 1658. He was appointed a director of the East India Company in 1674, and with the exception of 1676 he was re-elected to a directorship in every subsequent year until his death. In 1681 he was elected governor of the company and established a close relationship with the Crown. Following the Revolution of 1688, and in response to mounting attacks on his conduct of company affairs, he relinquished some of his active management responsibilities.

Child's claim to recognition as an economist rests on his *Brief Observations concerning Trade and Interest of Money*, first published in 1668 and reissued (anonymously) in expanded form as *A Discourse about Trade* in 1690 and again as *A New Discourse of Trade* (with Child's name on the title page) in 1693. The work summarizes the views he presented to the Council of Trade appointed by the King in 1668 (following the appointment of a Select Committee on the State of Trade by the House of Commons in the preceding year) and to a similar House of Lords Committee in 1669.

Among the reasons for the mercantile supremacy of the Dutch, he cites the establishment of banks and the widespread use of transferable bills of exchange, which he strongly argued should be adopted in England. He argued for a reduction of the legal maximum rate of interest from six to four per cent (referring to this as ‘my old theme’), claiming that the lower rate of interest in the Netherlands was ‘the *causa causans* of all the other causes of the riches of that people’. He saw the beneficial effects on trade of a lower cost of money capital, but he did not discuss, as did John Locke at the same time, the relation between a legally established rate of interest and the rate established by natural market forces.

Child's argument that the beneficial effect of lower interest rates would cause ‘all sorts of labouring

people that depend on trade (to be) more constantly and fully employed' took up the then widespread concern with the employment problem and he concluded: 'it is our duty to God and nature so to provide for and employ the poor'. A significant discussion of the question of the poor and a scheme for their relief and employment is included in Chapter II of the *Discourse of Trade*.

Notwithstanding his scattered observations that appear to support free trade principles and his assertion of the principles of competitive markets, Child was an exponent of monopoly when it suited his and the East India Company's advantage. He recognized the need to export bullion if that gave rise to further export trade opportunities. But his work abounds in arguments for trade restrictions in specific cases, such as those requiring the transportation of traded commodities in English vessels and requiring that colonial trade should be conducted only with England, thereby emphasizing the domestic employment-creating effects of the colonies. He stands as a latter-day mercantilist rather than an analytical anticipator of the laissez-faire doctrines of genuine and generalized freedom of trade.

Selected works

1668. *Brief Observations concerning Trade and Interest of Money*. London.

1693. *A New Discourse of Trade*. London.

Bibliography

Letwin, W. 1959. *Sir Josiah Child, Merchant Economist*. Boston: Harvard Graduate School of Business.

Macauley, T. (Lord) 1848, 1855, 1861. *The History of England from the Accession to James the Second*. New York: AMS Press, 1968.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

How to cite this article

Vickers, Douglas. "Child, Josiah (1630–1699)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000130> doi:10.1057/9780230226203.0230

The New Palgrave Dictionary of Economics Online

childcare

David M. Blau

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The market for childcare and the role of the government in the childcare market have grown enormously as mothers of young children have entered the labour force in very large numbers. Economists have made important contributions to understanding many aspects of childcare. This article focuses on (*a*) the effect of the price of childcare on labour force participation of mothers of young children, (*b*) the effect of childcare and early childhood interventions on children, and (*c*) the rationale for and effects of government involvement in childcare. Fruitful avenues of additional research are suggested.

Keywords

adverse selection; child development; childcare; childcare subsidies; education production function; Head Start; human capital; imperfect information; labour force participation; market imperfections; moral hazard; poverty; random assignment; self-sufficiency; women

Article

The market for childcare in advanced economies has grown enormously in response to the dramatic increase in labour force participation by mothers of young children. In 1950 12 per cent of married women in the United States with children under age six were in the labour force, compared to 63 per cent in 2000. Labour force participation of single mothers with children under six has also increased rapidly, reaching 65 per cent in 2000. As the market has grown, the role of the public sector in subsidizing, regulating, and providing childcare has increased substantially. One-third of all expenditure on childcare and preschool in the United States is financed by government subsidies or by direct provision of services. The public sector plays an even larger role in childcare in many European countries. Three aspects of childcare have received the most attention from economists: (*a*) the effect of the price of childcare on labour force participation of mothers, (*b*) the effect of childcare and early childhood interventions on child development, and (*c*) the rationale for and effects of government

involvement in childcare. Childcare is interpreted broadly here to include care provided by someone other than a child's parent either to facilitate employment of parents or to enhance child development. Blau and Currie (2006) summarize the findings of 20 studies that estimate the elasticity of maternal labour force participation with respect to the price of childcare. The estimates vary widely across studies, but studies that account for the availability of informal unpaid childcare options usually estimate relatively small elasticities, in the range of –.09 to –.20. These studies use a multinomial choice framework that allows for the possibility that a mother can work without using paid childcare. Use of unpaid childcare by family members, relatives, and others is very common. The relatively small elasticity estimates suggest that a price increase induces substitution of informal unpaid childcare for paid care, dampening the sensitivity of maternal employment to the price of childcare. Some evidence suggests that the price elasticity is larger in absolute value for lower-wage women. This evidence confirms that childcare costs are a significant but not major barrier to employment of mothers. The evidence also implies that childcare subsidies increase work incentives of mothers, a finding confirmed by a small number of studies that directly analyse the impact of subsidy programmes on employment.

An important concern about childcare is that low-quality care could be harmful to the development of young children. Conversely, high-quality care may help compensate children from low-income families for the disadvantages of growing up in poverty. The effect of childcare on child development has traditionally been the domain of developmental psychology, but in recent years economists have contributed to this literature, noting its similarities to the ‘education production function’ literature for school-age children.

The quality of childcare can be characterized by ‘structural’ features such as the size of the group in which care is provided, the ratio of adult caregivers to children, and the education and specialized training of providers. Alternatively, direct observation of the developmental appropriateness of the care received by children can be made by trained observers using standardized instruments. These ‘process’ measures of quality are more proximate determinants of child development than are the structural features.

The small amount of evidence available suggests that higher-income parents do *not* choose higher-quality childcare on average: among users of day-care centres, there is no systematic relationship between family income and the quality of childcare used, if other factors are controlled for (Blau, 2001). This is true whether the quality of care is measured by structural characteristics or process measures. This suggests that parents are either unable to discern the quality of care, or unwilling to pay the additional cost associated with higher-quality care, or both.

Several random assignment demonstration projects have evaluated the impact of high-quality preschool programmes for disadvantaged children (see reviews in Blau, 2001, and Blau and Currie, 2006). The results show that such programmes have delivered substantial long-run benefits to the participants and society: lower school dropout rates, higher earnings, fewer out-of-wedlock births, and lower public expenditures on welfare, criminal justice, and special education. Benefit-cost calculations show that these interventions have a very high social rate of return. This evidence is compelling, but it is based on very intensive and costly programmes that are of exceptionally high quality and are targeted at highly disadvantaged children. It is unclear whether childcare of moderately high quality provides positive but proportionately smaller developmental benefits, or whether there exists a threshold of quality below which benefits are negligible. It is also unclear how the quality of childcare affects children who are not highly disadvantaged. In non-experimental studies that follow children over time, higher-quality

childcare is associated with better developmental outcomes in the short run (one to three years). However, it remains uncertain to what extent this is a causal impact. Recent studies that control for many other potentially confounding factors find that the quality-development association is smaller than in models with fewer controls, but remains significantly different from zero.

Two main arguments have been used to rationalize a role for government in the childcare market. The arguments are based on attaining economic self-sufficiency, and childcare market imperfections. On self-sufficiency, childcare subsidies might help low-income families achieve economic self-sufficiency, defined as being employed and not enrolled in welfare programmes. Self-sufficiency is a desirable goal because it may inculcate a work ethic and generate human capital through on-the-job training and experience. These arguments explain why many childcare subsidies require employment or work-related activities such as education and training. Subsidies for childcare and other work-related expenses paid to employed low-income parents may cost the government more today than would cash assistance. But these subsidies could result in increased future wages and hours worked and lower lifetime government support than the alternative of cash assistance both today and in the future. This argument has nothing to do with the effects of childcare on children, and there are few restrictions on the quality of childcare that can be purchased with employment-related childcare subsidies. However, evidence on wage growth of low-skill workers suggests that wages grow only modestly with experience, too slowly to lift low-skill workers out of poverty (Gladden and Taber, 2000). Middle and upper-income families are generally not at risk of going on welfare, so it is not obvious that there is an economic rationale for subsidies for their employment-related childcare expenses.

As for market imperfections, the imperfections that are often cited are imperfect information available to parents about the quality of childcare, and positive external benefits to society generated by high-quality childcare (Walker, 1991). Imperfect information exists because consumers do not know the identity of all potential suppliers, and the quality of care offered by any particular supplier is not fully known. A potential remedy for the first problem is government subsidies to resource and referral agencies to maintain comprehensive and accurate lists of suppliers. The second information problem arises because consumers know less about product quality than does the provider, and monitoring the provider is costly to the consumer. This can lead to moral hazard and/or adverse selection. The limited evidence available suggests that parents are not well-informed about the quality of care in the arrangements used by their children. Childcare subsidies targeted at high-quality providers could induce parents to use higher-quality care.

The externality argument is a standard one that closely parallels the reasoning applied to education. High-quality childcare leads to improved intellectual and social development, which in turn increases school readiness and completion. This reduces the cost to society of problems associated with low education: low earnings, unstable employment, crime, drugs, teenage childbearing, and so forth. If parents are not fully aware of these benefits, or account for only the private and not the social benefits, then they may choose childcare of less than socially optimal quality. This argument could rationalize subsidies targeted to high-quality providers, such as Head Start, a US programme aimed at enhancing cognitive and social development of low-income children.

As this discussion implies, childcare policy can be used to facilitate employment of mothers and enhance development of young children. There is likely to be a trade-off between these goals because higher-quality care is more expensive. There is not a political agreement in the United States to spend enough to achieve both goals, or on which goal should have the highest priority. This is due in part to

conflicting views on the proper role of the government in a domain that was mainly left to families until the last quarter of the twentieth century. But it also reflects lack of knowledge about the magnitudes of important parameters that affect the costs and benefits of alternative policies. Economists could make significant contributions to knowledge by careful empirical studies that produce reliable estimates of such parameters. The following issues seem important and well-suited to analysis by economists. Despite a large number of studies, there is considerable uncertainty about the magnitude of the elasticity of maternal employment with respect to the price of childcare. A careful sensitivity analysis could help resolve this uncertainty. Research on the price-responsiveness of low-income mothers would be especially useful. Consumer demand for quality in childcare is not well-understood, and new research could be valuable. Research on the take-up decisions of families eligible for childcare subsidies would be useful in order to determine the likely effectiveness of different forms of subsidies. New research on the supply of childcare would be useful. Subsidies to consumers may bid up the price of childcare, and it is important to be able to quantify such effects. It would also be useful to examine the quality supply decisions of providers, in order to determine how responsive the supply of high-quality care might be to subsidies.

See Also

- education production functions
- family economics
- labour supply
- women's work and wages

Bibliography

Blau, D. 2001. *The Child Care Problem: An Economic Analysis*. New York: Russell Sage Foundation.

Blau, D. and Currie, J. 2006. Preschool, day care, and after school care: who's minding the kids? In *Handbook on the Economics of Education*, ed. E. Hanushek and F. Welch. Amsterdam: North-Holland.

Gladden, T. and Taber, C. 2000. Wage progression among less skilled workers. In *Finding Jobs: Work and Welfare Reform*, ed. R. Blank and D. Card. New York: Russell Sage Foundation.

Walker, J. 1991. Public policy and the supply of child care services. In *The Economics of Child Care*, ed. D. Blau. New York: Russell Sage Foundation.

How to cite this article

Blau, David M. "childcare." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/

article?id=pde2008_C000547> doi:10.1057/9780230226203.0231

The New Palgrave Dictionary of Economics Online

China, economics in

Kiichiro Yagi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Although pre-modern China possessed rich ideas pertaining to economic matters, they were not separated from the discourse of morality and politics. Even in the late 20th century, Chinese thinking, often unconsciously, reflected traditional ideas. Liberal economics missed the chance to guide the modernization of China. Marxian economics established its monopoly under the reign of the Communists. However, China had Marxian economists who supported its gradualist transition to a market economy. In the 1990s, the task of guiding economic reforms in China was handed over to a new generation of economists who absorbed ‘Western’ (non-Marxian) economics.

Keywords

land nationalization; Asiatic mode of production; central planning; China, economics in; Confucianism; Datong (Great Harmony); equality; George, H.; institutional economics; law of value; Legalist School (China); management buyouts; Marxism; modernization; Mohist School (China); political economy; population growth; public choice; socialist commodity economy; socialist market economy

Article

Economics in China has not been able to disassociate itself from politics. The Chinese word for economy or economics, *Jingji*, is the abbreviation of *Jingshi* (or *Jingguo*) *Jimin*, which means ‘ruling the society or state and saving the people’. In traditional Chinese learning, this is a generalized concept that covers almost the entire range of a state's administrative activity. However, the viewpoint implied in this word is that of the rulers or administrators and not that of individuals engaging in economic activity on their own account.

The quest for wealth and the control of morality

Policies oriented towards the attainment of ‘wealth and power’ had appeared already in ancient China, the Eastern Chou Period (722–256 bc), when the rule of Chou dynasty became in title alone and powerful vassal lords struggled with each other for leadership, which was based on the power of their feudal states. A crucial insight pertaining to economic growth that emerged during this period was that fostering the material welfare of the people was a precondition for a strong state. The famous saying ‘Man will care about honour and disgrace only when he has enough clothing and food’ is attributed to Guan Zhong (730–645 bc), the prime minister of a ducal state. He implemented policies that would bring stability to people's lives; these policies included the promotion of agriculture, monopolizing salt and iron, state intervention in the public distribution system, maintenance of a balanced budget and the consolidation of taxation and military services. Practical policies were further developed by many politicians in the Warring States Period (475 –221 bc). These became part of the arsenal of policy measures adopted by the administrators of the unified state of successive dynasties from the Qin (221–206 bc) to the Qing (ad 1644–1911). A text named after Guan, *Guan Zi*, was compiled in the Western Han Period (206 bc–ad 8). This contains detailed discussions of the practical economic policies of ancient China.

In ancient China, before the unification by the Qin, political control over merchants was not strict. Wealthy merchants in the pre-Qin period were vividly described in *Records of the Historian* (‘*Shiji*’). The editor–historian, Sima Qian (145–87 bc) clearly favoured a liberal economic policy that permitted the innovative activities of talented merchants.

Competitive schools in ancient China

Confucius (551–479 bc) also recognized the quest for wealth as a natural human trait. However, he stressed that the teachings of morality (*Ren*) should control the quest for wealth. According to him, superior men can understand and adhere to the virtues of righteousness and benevolence in their deeds, while inferior men (common people) cannot. The former belong to the ruling class and the latter are the ones who are ruled, who must be guided by the former. Confucius stressed the educational effect of a ruler on the people's perception of societal order. He was opposed to the levying of heavy taxes and unnecessary state intervention, since that might jeopardize the common man's standard of living. He maintained that a peaceful and fair reign of a virtuous ruler fosters allegiance. As long as people follow the basic order of society, the wealth of the state emerges as a spontaneous result of the growth in the population.

Meng Ke (c. 390–c. 305 bc), whose name is often mentioned together with Confucius, strictly excluded the consideration of material benefits from the political discourse of superior men. During his first meeting with the king of Liang, Meng declared that he spoke only of ‘righteousness’ (*Yi*) and not of ‘benefits’ (*Li*). However, he also stated that the dominance of ‘righteousness’ presupposes the maintenance of a ‘permanent property’ of the people in order to secure the morality of the people (*Mencius*).

Mo Di (c. 468–c. 367 bc) and his School (Mohists) grounded their altruistic teaching on the extended approval of ‘benefits’. They believed that economic transactions are acts of ‘mutual benefit’, which will eventually support the doctrine of ‘universal love’. From a utilitarian viewpoint, they regarded righteousness as a material benefit; this is in clear contradiction with the Confucians. Mohists further advocated a ban on war and simple burial. Apparently, this School originated from the craftsmen who

were not entirely integrated into the social hierarchy existing in the pre-Qin period.
Legalists such as Shang Yang (c. 390–338 bc) and Han Fei (c. 310–238 bc) differed from the Confucians with respect to the measures to be adopted for guiding people. They stressed the effective control of people by the strict enforcement of punishment. They prioritized agricultural production and considered manufacturing and commerce as tertiary activities. The Legalists were prepared to collaborate with princes and politicians who sought to enhance the wealth and power of their states.

Omnipotent state vs. virtuous reign

Ancient China was unified by the Qin dynasty, which had adopted the policies of Legalists. The first emperor of the Qin (221–206 bc) suppressed Confucians who criticized his reign as measured against the criterion of virtuous ruler. However, under the following dynasty, the Han, Confucianism established its position as the state orthodoxy, which continued until the end of the Qing dynasty. Still, a Legalist direction survived in the pragmatic mentality and policies of administrative bureaucracy. Thus, Chinese political history witnessed repetitive conflicts between the moralistic direction of Confucianism and the bureaucratic administration in the direction of Legalists.
One of the most noteworthy debates was the dispute on salt and iron (81 bc), in which San Hongyoung (152–80 bc) – the finance minister of the Western Han dynasty – had to defend his policy against the criticism of Confucian scholars. In order to compensate for the deficit in the state finance caused by an expansionary policy, San extended the state monopoly of salt and iron and introduced a state-managed storage and distribution system. Such a system could be legitimized if it was successful in guaranteeing the nationwide provision of necessaries and a stabilization of their prices. However, coupled with a heavy tax burden, San's system made a devastating impact on the nation. Confucian scholars voiced the dissatisfaction of the people and pressed for the abolition of San's system.
A similar constellation appeared in the dispute around the economic reforms of Wang Anshi (1021–86). Wang's attempt to consolidate public finances by suppressing the annexation of lands by rich families and establishing a strict taxation system was opposed by traditional scholars, who were in alliance with the richer families.
Apart from the taxation and market control, Chinese administrators showed their expertise in the area of currency. They are the first to have issued paper money (*Jiao Zi*) in the 11th century. The Yuan dynasty (1271–1368) adopted the idea of inconvertible notes in its monetary system. The paper currency ordinance of 1287 drafted by Ye Li (1242–92) contained sound measures to maintain the value of paper money in relation to the regularly inspected silver reserve fund. This paper currency system of the Yuan dynasty exerted a certain influence over the currency system of other countries through the commercial networks under the grand Mongolian rule.

The demand for equalization

Support for equality is another persistent trait of traditional Chinese economic thought. The equalization of land and wealth was a typical demand raised by numerous peasant rebellions. The Taiping Rebellion (1851–64) put into effect an equal distribution of land, and the rural revolution under Mao Zedong's (1893–1976) directive displayed a similar kind of egalitarianism. However, the ideal of equality in the distribution of wealth can also be found in Confucian classics. Confucius himself remarked that rulers

must worry ‘not about the scantiness of wealth but its inequality of distribution’ since ‘there will be no feeling of poverty under equal distribution’ (*Analects*). Here, equality is appreciated with respect to its ability to maintain harmony and tranquillity among the ruled. Meng Ke also proposed an egalitarian *Jin* land system, in which peasants, who were allotted equal amounts of land, jointly cultivated public land for the sake of generating public finance. This proposal was revived several times by reformist politicians as well as by egalitarian rebels.

A vision of an egalitarian ideal society, the Great Harmony (*Datong*), where neither private property nor egoistic interests exist, is mentioned in the Confucian classic *Li Ki*. *Xiaokang*, a society in which the people are guided by order and institutions is not an ideal but a second best, suited to the age of a civilized society. However, towards the end of the Qing dynasty, Kang Youwei (1858–1927), a reformist politician and scholar, revived the ideal of the Great Harmony to regenerate the whole nation.

Preconditions for Chinese modernization

The nationwide examination system for the recruitment of government officials was established under the Sui dynasty (581–618) and continued until 1905. Based on the Confucian orthodoxy, it moulded the thought of Chinese intellectuals over a millennium. However, Confucian orthodoxy was not totally exempt from change. In addition to the ideas that had emerged in the ancient period, it absorbed heterogeneous ideas from other intellectual schools of thought, such as Buddhism and Taoism. The effect of the development of a rationalistic Neo-Confucianism guided by Zhu Xi (1130–1200) and the emergence of the countervailing school of Wang Yangmin (1472–1528), which introduced an inner integrity to Confucianism, are interesting issues that need to be further researched. Towards the end of the Ming dynasty (1368–1644), these developments promoted a critical attitude towards the traditional order of the empire. Huang Zongxi (1610–95) and Wang Fuzhi (1619–95) developed a utilitarian concept of hierarchy based on the private property and self-interest of the people. Further, the diffusion of the teaching of Wang Yangmin (*Xinxue*) that stressed purity of mind nourished the morality of the merchants (Yu, 1987). However, these developments were not sufficient to modernize the Chinese intellectual tradition from within. The landlord class that recruited state officials through a nationwide examination formed the ruling alliance of the society. Merchants had no other option but to join this alliance as subordinate participants. However, the intellectual legacies of old China were preconditions for the Chinese to cope with the modernization that was initially forced on them by external forces.

Introduction of Western economics

It was the publication by Wei Yuan (1794–1857) of the *Geography of the Maritime Countries* (1843) that initiated the movement among Chinese intellectuals of learning from the West. However, Western economics was not introduced until two decades later. Using H. Fawcett's *A Manual of Political Economy* as a textbook, W.A.P. Martin, an American Christian missionary, began a course on policies for the wealth of nations at a government school in Beijing in 1867. Later, in 1883, this course was translated and published in Shanghai under the same title. A second significant contribution pertaining to the translation of Western economics was that of a British missionary, J. Edkins, who translated W.S. Jevons's *Primer of Political Economy* into Chinese. This translation was published in 1886 with the

Chinese title, *Policies for the Wealth of Nations and Support of People*. Fawcett and Jevons were neither mercantilists nor interventionists. However, both Chinese titles suggest that the Chinese people of this period regarded Western economics as a policy measure to strengthen the state.

Between 1901 and 1902, Yan Fu (1853–1921) published the translation of Adam Smith's *Wealth of Nations* in Shanghai under the title *Elements of Wealth* (‘*Yuan Fu*’). In his commentary on this translation, Yan clearly stated that the principles of economics advocate free competition, are against state intervention and limit the scope of state involvement in those tasks that are not suited for the private sector. However, most Chinese intellectuals, including Yan himself, accepted the theory of liberal economics because of its contribution to the recovery of the power of the nation (Schwartz, 1964).

However, the principles of liberal economics do not appear to have contributed much to the modernization of China. Late 19th-century reformers had to fight against the obsolete bureaucracy of the Qing dynasty. As was typical of revolutions in the 20th century, the social dimension of the Chinese revolution increased in significance with the passage of time. Democrats and liberals worked together on the cultural front of the 4 May Movement (1919). However, this collaboration soon broke down, since democrats shifted their position to that of Communist revolutionaries and began to attack liberals as ‘bourgeois intellectuals’.

The ideology of Western socialists and social reformers was introduced by Sun Yatsen (1866–1925) through his ‘Three People's Doctrines’. Sun regarded Western capitalism as the root cause of the social problems in the West and searched for an alternative route towards economic development for China. He recognized Henry George's idea of land nationalization and the German socialist idea of capital regulation. After experiencing the state of anarchy that followed the Xinhai Revolution (1911), he sympathized with the Russian Revolution and led his Nationalist Party, the Guomingdang, in cooperation with the Communists.

Period of the Republic of China

Despite continued struggles among the warlords and an unstable security environment in both domestic as well as external affairs, the period of the Republic of China (1912–49) marked the emergence of economic academism in China. Most of the renowned universities of today originated in this period, and specialized economists, some of whom were educated in the United States, Europe and Japan, began to teach there. There were 16 Chinese publications on economics in the decade following Yan's translation of Adam Smith's *Wealth of Nations*; this number increased to 20 between the 1911 revolution and the 4 May Movement. It further increased to 228 in the decade following 1919 and to 1,116 after 1929 (Shanghaishi, 2005, pp. 114–15).

The Chinese Economic Society was established in 1923, and after a decade its membership amounted to c. 600. In 1930, it launched the quarterly journal *Jinngjixue Jikan* in Shanghai. Ma Yinchu (1882–1982), a Ph.D. holder from Columbia University who had taught economics at Beijing University since 1915, was its president. He served the Guomingdan government as its economic advisor and published his views on the currency problems, banking and public finance in China. The Chinese economists of this period actively participated in policy discussions, such as the currency reforms of 1936, financial problems and industrial development plans.

However, it was the problem of agriculture that most concerned Chinese economists. A large-scale

research project in rural economy headed by Chen Hansheng (1897–2004) gave birth in 1933 to the Research Forum in Chinese Rural Economy. This forum gathered a membership of about 500 members and trained economists who continued their research activity in the post-1949 period. The most prominent member among them was Xue Muqiao (1904–2005), who edited *Rural China* (‘*Zhongguo Nongcun*’) from 1934.

Social scientists influenced by Marxism eagerly discussed the nature of existing Chinese society (1929–1931). This debate contained a political element since those who supported the Chinese Communist Party (CCP), which was founded in 1921, regarded Chinese society to be a semi-feudal and semi-colonized society, whereas the Trotskyists emphasized the dominance of the capitalistic elements. Such debates on the nature of Chinese social history and its periodization (1931–3) and on the Asiatic mode of production continued in the field of economic history.

Marxist monopoly under the PRC

The People's Republic of China (PRC) started in 1949 with the programme of the ‘New Democracy’ that was to be based on the alliance between Communists and democrats from all sections of the society. The government requested non-resident Chinese scholars to participate in the reconstruction of China.

Ma Yinchu, who was exiled to Hong Kong as a result of a dispute with the Guomingdang government, returned to take over as the president of Beijing University. Initially, several of his colleagues were those who had been educated in American universities. Thus, in the beginning of the PRC, universities in China had non-Marxian economists on their staff. However, the socialist reconstruction of academic system based on the Soviet–Russian model, and the intensifying confrontation with the United States, soon deprived ‘bourgeois economists’ of freedom. Abridged translations of Russian textbooks pertaining to Marxian economics became the standard education materials. In 1957, when the CCP declared a liberal policy towards intellectuals with the appealing phrase ‘Let a hundred flowers blossom’, Ma proposed his idea of population restraint to the People's Congress of the PRC. This offended Mao Zedong's positive view of population growth. The ensuing continuous attacks on ‘Malthus in China’ signalled the expulsion of non-Marxian ideas from the academic world under the PRC.

According to the original concept of the New Democratic Economy, the development of capitalism in China, except for ‘monopoly capital’, was to be welcomed as the basis for initiating future socialist transformations. However, in 1953 the success of the agrarian reforms motivated Mao to practise ‘the solution to the problem of ownership’. Through the socialization of the ownership of the means of production, a Soviet Russian-type of planned economy was established in the sectors of industry and commerce during 1953–6. This was followed by the establishment of people's communes in the rural areas in 1958.

Reform economists in China

The first criticism levelled against a centrally planned economy also emerged in the years of ‘Let a hundred flowers blossom’. In 1956 and 1957, Sun Yufang (1908–83) proposed an economic model of decentralization with the use of profit targets in the management of manufacturing sector. Sun was a Marxian economist who had studied in Moscow. He grounded his proposal on the validity of the ‘law of value’ in a socialist economy, which is distinguished from the ‘law of market’. In this respect, the views

of Gu Zhen (1915–74) were more progressive, in that he openly criticized the abolition of the market mechanism under socialism. During the wave of the Anti-Rightist Struggle that occurred during the latter half of 1957, Sun and Gu were labelled ‘revisionist’ and ‘bourgeois rightist’ respectively.
Chinese economists were aware of the shortcomings of a Russian-type planned economy and the need for reform. However, the ideological rejection of the ‘material interest’ as a tool of ‘revisionists’ prevented the introduction of reforms in the management system of state-owned enterprises (SOEs). Ideological politicians stuck to the appeal to ‘spiritual incentive’. Reforms were then directed towards an administrative decentralization, in which powers and benefits were divided among various administrative organs.
It was only after the declaration of the end of the Great Cultural Revolution (1966–1976) and with Deng Xiaoping (1904–1997) taking over the leadership of China that the damage caused by excessive decentralization and the need for management reforms were seriously taken into consideration. After the strategic decision of the CCP for economic reforms and an ‘open door’ policy, China implemented various policies such as the creation of special economic zones and township and village enterprises as well as the approval of private enterprises and households contracting in agriculture. Under the concept of the ‘planned commodity economy’ (1984), the market economy was theoretically subordinate to the planned economy. The existence of private sectors was legitimized by the theory of the ‘early stage of socialism’ (1987). At last, in the 1990s, by the definition of the ‘socialist market economy’ (1992), the private sector was clearly approved as the main and normal element of Chinese socialist economy.
A group of veteran economists, namely, Xue Muqiao, Du Rensheng (born 1913), Yu Guangyuan (born 1915), Liao Jili (1915–93), Lieu Guogang (born 1923), and others contributed to the transformation of the concept of ‘socialist economy’. In the early 1980s, they re-examined the orthodox and heterodox texts of Marxism, studied reform economics of former socialist eastern European countries, and endeavoured to draw conclusions from the empirical research on agriculture and manufacture sectors. They formed the ‘theory of the socialist commodity economy’.
After Mao's death and the end of the Great Cultural Revolution, academic economists soon regained their energy. The Chinese Academy of Social Sciences (CASS) was established in 1977. The oldest Shanghai Economics Society, whose origin can be traced back to 1950, resumed its activities in 1978. In the same year, the Chinese Research Forum of Overseas Economics was established and began to work for the diffusion of the ‘Western’ (non-Marxian) economics among Chinese economists.
In the 1980s, Chinese economists recovered their communications with the world community of economists. The government invited renowned Western economists to academic conferences pertaining to the economic reforms in China. It began to send young people to the graduate courses of top Western universities, and encouraged them to assimilate advanced analysis of modern economics. By the mid-1990s, China already had a group of talented economists who could analyse economic reforms in China in a manner similar to the Western (non-Marxian) economists. In the fields of research, economic teaching, and policymaking, the activities of non-Marxian economists became more significant with each passing year. Thus, the monopoly of the Marxian economists was broken.

Present situation of economics in China

The ideological/political control exercised by the CCP over Chinese intellectuals had been considerably reduced at the outset of the 21st century. Economists in China can now keep themselves abreast of the

latest developments in the field of economics. However, the following three features are noteworthy when compared with economics in other countries.

The first is the peculiar position of Marxian economics in China. At present, it is clear that Marxian economics is just a sub-area in the whole gamut of research activities undertaken by Chinese economists. It is therefore symbolic that the Marxian economists organized themselves into a society named the Chinese Forum for the Study of *Capital* (founded in 1981). However, Marxian economics still influences society by two privileged routes. One is that Marxian economics continues to be an obligatory course of political economy (*Zhengzhi Jingjixue*) in most Chinese universities. It is virtually a part of the political education imposed on academicians. The other route is the ideological function for the ruling CCP. The CCP needs Marxian economists to defend its policy on ideological grounds.

The second noteworthy feature of Chinese economics is the focus on institutional economics and political economy. Leading economists of the post-Great Cultural Revolution generation such as Lin Yifu (born 1952) and Fan Gang (born 1953) adopted the framework of institutional economics. Lin attributed the success of the Chinese economy after the implementation of the ‘open door’ policy to the switch of the development strategy and the institutional reforms accompanying it. Fan provided an analysis of the incremental reforms in China by applying the public choice approach. The theories pertaining to modern institutional economics – transaction cost theory, property rights theory, contract and corporate governance theory, and comparative institutional analysis – are widely accepted by Chinese economists.

Lastly, a new divide between the supporters of the prevailing liberal policy and its critics emerged in 2004, and a debate between these two groups has continued since then. First, Lang Xianping (born 1956), a professor at the Chinese University in Hong Kong, attacked managers of the firms whose stocks were newly listed on the stock market. They were charged with smuggling national property by the application of various techniques such as management buyouts. His attack on the privatization policy encouraged economists who were concerned about the increasing inequality in society and diminishing state intervention. They criticized over-hasty privatization and demanded a policy that would enhance the level of equality in society. They stressed the need to implement reforms in the field of social policy, and rejected the unconditional integration of the Chinese economy within the global market. Liberal economists, who stressed efficiency, rebutted them. Another group of economists declared themselves as taking a middle-of-the-road position. The government is said to have attentively followed the debate.

See Also

- Chinese economic reforms
- culture and economics

Bibliography

Fawcett, H. 1883. *Fuguoce* [Policies for the Wealth of Nations]. Translation by Wang Fengzao under the supervision of W.A.P. Marten of *A Manual of Political Economy*, London: Macmillan, 1863. Shanghai: Shanghai Meihua Shuguan.

Hu Jichuang. 1988. *A Concise History of Chinese Economic Thought*. Beijing: Foreign Language Press.

Jevons, W.S. 1886. *Fuguo Yangmince* [Policies for the Wealth of Nations and Support of People]. Translation by J. Edkins of *Primer of Political Economy*, London: Macmillan, 1878. Zongshuiwusishu.

Schwartz, B.I. 1964. *In Search of Wealth and Power: Yen Fu and the West*. Cambridge, MA: Harvard University Press.

Shanghaishi Shehuikexue Lianhehui., ed. 2005. *20 Shiji Zhongguo Shehuikexue Lilun Jingjixue* [Social Sciences in the 20th Century China, Economic Theory]. Shanghai: Shanghai Renmin Chubenshe.

Trescott, P.B. 2006. *Jingji Xue: History of Introduction of Western Economic Ideas into China 1850–1950*. Hong Kong: Chinese University Press.

Wei Yuan. 1843. *Haiguo Tuzhi* [Geography of the Maritime Countries]. Yangzhou.

Wu Jinglian. 2005. *Understanding and Interpreting Chinese Economic Reform*. Mason, OH: Thomson Higher Education.

Yu Yingshi. 1987. *Zhongguo Jinshi Zongjiao Lunli yu Shangren Jingshen* [Religious Ethics and Spirit of Merchants in Early Modern China]. Taipei: Lianjian Chuben.

Zhao Jing., ed. 1991–8. *Zhongguo Jingji Sixiang Tongshi* [Complete History of Chinese Economic Thought], vols. 4. Beijing: Beijing Daxue Chubanshe.

Zhao Jing., ed. 2004. *Zhongguo Jingji Sixiang Tongshi Xuji: Zhongguo Jingji Jindai Sixiangshi* [Complete History of Chinese Economic Thought Continued: History of Modern Chinese Economic Thought]. Beijing: Beijing Daxue Chubanshe.

Zhongguo Dabaike Quanshu Zongbianji Weiyuanhei Jingjixue Bianji Weiyuanhei, ed. 1998. *Zhongguo Dabaike Quanshu, Jingjixue 1* [Great Encyclopaedia of China, Economics 1]. Beijing: Zhongguo Dabaike Quanshu Chubenshe.

How to cite this article

Yagi, Kiichiro. "China, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000577> doi:10.1057/9780230226203.0232

The New Palgrave Dictionary of Economics Online

Chinese economic reforms

Loren Brandt and Thomas G. Rawski

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Why did China's modest reforms unleash an enormous boom? Three decades of socialist planning created vast untapped potential. China captured this potential by focusing on ‘big reforms’ linked to incentives, markets, prices, mobility, openness and competition. Advances in these areas created sufficient momentum to overcome the drag associated with remaining distortions and institutional shortcomings. China's political economy, which incorporates substantial local autonomy, facilitated experimentation that repeatedly identified feasible reform paths. Because China's political economy delivers undesirable outcomes along with rapid growth, and because China's success is linked to unique historical circumstances, the beneficial outcomes associated with Chinese policies and institutions may be limited in time and space.

Keywords

agricultural productivity; China, economics in; Chinese economic reform; collectivization; decentralization; dual price system; planning; rural–urban migration; socialist market economy (China); Cultural Revolution (China); Great Leap Forward (China)

Article

Since the late 1970s, China's economic performance has astonished the world. Official figures show that, after adjusting for inflation, China's GDP grew at an annual rate of 9.7 per cent between 1978 and 2006, and at a rate of 8.4 per cent in per capita terms (Yearbook, 2006, p. 60; National Bureau of Statistics, 2006). By 2006, the Chinese economy, measured in terms of purchasing power parity, was the world's second largest, behind only the United States: per capita incomes, measured on the same basis, rose from 324 dollars to 5,772 dollars between 1978 and 2004 (Heston, Summers and Aten, 2006). China's new dynamism includes a major shift towards intensive growth, with productivity change, which had contributed negatively to Chinese growth between 1957 and 1978, accounting for 40 per cent of overall

growth after 1978 (Perkins and Rawski, forthcoming).

Reform began in the late 1970s. The impetus for modifying the plan system came from two sources: general awareness that China's neighbours were running far ahead in the economic sphere, and stagnation of living standards, especially China's persistent problems with food supply. The initial objective was to improve economic results under the system of central planning.

Initial reform efforts

Not surprisingly, early reform efforts focused on agriculture. Starting in 1978, household cultivation swiftly replaced collective tillage as the norm in China's vast farm sector, as hundreds of millions voted with their feet to abandon collective farming, the central feature of the people's communes. Introduction of the household responsibility system meant that farmers could claim the fruits of extra effort for themselves. This brought an immediate multiplication of work effort, which was further encouraged by modest relaxation of restrictions on marketing and price flexibility, and by a considerable increase in procurement prices (Sicular, 1995). The result was a sudden upsurge of farm production and productivity (Lin, 1992). With the expansion of food supply, millions of farmers no longer needed to work the land and so began to move into non-farm employment. Improved diets raised the energy levels and hence the productivity of formerly undernourished villagers. Relaxation of efforts to enforce local self-sufficiency in favour of historic patterns of crop specialization, along with new opportunities to diversify into animal husbandry, horticulture, and aquaculture, also contributed to steep gains in farm output (Lardy, 1983).

The response to agricultural reform quickly spread beyond the farm sector. Rural factories, which had enjoyed a brief boom during the Great Leap Forward of 1958–60 (a massive and chaotic push to organize villagers into communes and to transfer rural labour into steel and other industries), suffered considerable retrenchment during the 1960s, and then expanded rapidly during the 1970s. Following the revival of agriculture, collectively owned rural industry, now fortified by greater access to the cities, rising rural incomes, increased supplies of agricultural inputs, and throngs of job-seekers, bounded ahead. In addition, new freedom encouraged a wide range of non-farm self-employment and family businesses. The resulting shift out of farming initiated what eventually became a massive exodus of labour from the countryside.

The explosive response to rural reform spurred officials to press forward with urban initiatives focused on ‘enlivening’ state-owned enterprises. While these early measures achieved only limited progress towards their main objective, they benefited rural and urban collective industry by opening new markets as well as new sources of materials, subcontracting opportunities, and technical expertise.

As the influence of markets, price flexibility, and mobility expanded, a separate strand of reform began to move China's isolated system towards greater participation in international trade and investment. China's leaders agreed to establish four tiny ‘special economic zones’ in the southern provinces of Guangdong and Fujian. Initial operations in these zones seemed directionless and inconsequential, but the arrival of ethnic Chinese entrepreneurs, most from Hong Kong and Taiwan, turned the zones into drivers of regional and eventually national growth. This novel combination of low-cost Chinese labour with the market knowledge and entrepreneurial capabilities of overseas Chinese businessmen gradually developed into an export bonanza that nudged China towards its subsequent embrace of economic globalization.

Although the limited extent of domestic reform restricted the initial response to growing openness, the buoyant prosperity of the new zones prompted cities along the coast, and eventually across the nation to clamour for access to the same tax, legal, and regulatory concessions that had powered their growth.
China's initial reforms focused on limited changes directed at specific sectors. These changes proved sufficient to accelerate growth despite the continued importance of state ownership, price controls, material-balance planning, and other key features of the socialist system. Early reform was particularly successful in removing long-standing constraints formerly imposed by limited availability of food and of foreign exchange.

Further reforms: expanding the cage

During this period, China's gathering boom encouraged a growing array of jurisdictions, constituencies, and interest groups to pursue the advantages enjoyed by reform participants, including expanded managerial autonomy and access to the special economic zones. The image of China's economy as a caged bird advanced by Chen Yun, an economic specialist within the leadership group, illustrates the underlying economic thinking (Lardy and Lieberthal, 1983). Chen argued that expanding the cage (reform) allows the bird to beneficially spread its wings; an overlarge cage threatens loss of control – thus the slogan ‘planned economy as the mainstream, market allocation as a supplement’.
Implementation of the dual price system, which partitioned allocation of most commodities into plan and market components and allowed the distribution of after-plan residuals at increasingly flexible prices, stands as the central policy achievement of this period. The expansion of market transactions began to whittle away at long-standing barriers to mobility, which had restricted the transfer of labour, capital, commodities and ideas across administrative boundaries, with negative consequences for growth of output and productivity.
Developments in the international sphere, including the continued growth of foreign trade, the northward spread of special zones, and the expansion of foreign direct investment, now involving multinational corporations as well as overseas Chinese entrepreneurs, extended the impact of market forces. The growth of cross-border transactions and the increased presence of foreign business operations on Chinese soil intensified pressures for contract arbitration, codification of urban land-use rights and other legal and institutional reforms needed to facilitate new activities.
The main impact of these reforms fell on flows – of labour, commodities, profits, and new investments. New entrants to the workforce, for example, including college graduates, were increasingly left to find their own positions, rather than receiving job assignments from local labour bureaus. Existing stocks, including assets or employees of extant firms, especially in the state sector, were not yet exposed to the full impact of market forces. Mergers appeared, but only on a microscopic scale. Despite the enactment of bankruptcy legislation, floundering companies rarely disappeared. Nor did redundant workers face the sack, although the ‘optimal labour programme’, which invited managers to identify essential and surplus workers, foreshadowed the mass layoffs of the late 1990s.

Economic reforms since 1992: towards a ‘socialist market economy’

The brief recession, triggered by efforts to quell inflation during the late 1980s, together with the anti-

reform backlash and pullback of foreign investment that followed the June 1989 suppression of popular unrest, slowed both growth and reform. The setback, however, was short. Deng Xiaoping's call for expanded reform during his southern tour of 1992, together with the Communist Party's 1992 decision to pursue a socialist market economy with Chinese characteristics gave fresh impetus and as well as new direction to economic reform.

The Party's 1992 decision replaced vague ideas of ‘doing better’ with a clear reform objective: a market economy in which the eventual role of the state will resemble the current circumstances of major economies such as those of France or Japan: macroeconomic management; regulation of health, environment, and so on; and strategic planning, with other functions explicitly assigned to the sphere of market determination.

Although the 1992 decision is a statement of principle rather than a description of reality, the ensuing 15 years witnessed decisive strides towards market outcomes, which we summarize in terms of four major shifts:

1. 1. *From plan to market*: price liberalization extended beyond the substantial achievements of the first reform decade: despite significant exceptions (energy, credit, foreign exchange) supply and demand now determine most prices (Li, 2006, pp. 104–7). The growing influence of market forces brought a considerable (but incomplete) hardening of budget constraints, even in the state sector. Market pressures compelled the dismissal of more than 50 million workers, most from state-owned factories. Mergers and acquisitions extended the reach of market pressures to much of China's capital stock. Barriers to the free flow of labour and goods continue to recede, and migrant workers have begun to attain normal citizenship rights in China's cities and towns. Growing expansion of wage differentials and of income inequality reflect the new prominence of market outcomes.
2. 2. *From village to town and city* and from agriculture to industry and services. The primary sector's GDP share dropped from 27.9 per cent in 1978 to 11.8 per cent in 2006. Following the departure of 150–200 million villagers from the land, survey data indicate that the primary sector's labour force share has declined from 69.2 per cent in 1978 to 31.8 per cent in 2004 (National Bureau of Statistics, 2006; Yearbook, 2006, p. 58; Brandt, Hsieh and Zhu, 2008).
3. 3. *From public to private ownership*. At the start of reform, the public sector (including collectives) held nearly all China's fixed capital. The growth of private business, while rapid in percentage terms, started from a tiny base. It was only from the late 1990s that the non-public sector, swollen by the privatization of rural collective enterprises, the transfer of (mostly small and medium) state-owned firms into private hands, and the rapid expansion of direct foreign investment, began to take on a prominent role in the national economy. The share of state-owned firms in industrial output fell from 81 per cent to 55 per cent between 1980 and 1990, and to 15–35 per cent in 2005/6 (depending on the treatment of state shareholdings; see National Bureau of Statistics, 2006; Perkins and Rawski, 2008). The pace of change has accelerated: by 2003, the private sector's GDP share had risen to 59.2 per cent (OECD, 2005, p. 125). The state sector's share in industrial output and non-farm employment during 2004/5 declined to 15.2 and 13.1 per cent (Yearbook, 2006, p. 505; Brandt, Hsieh and Zhu, 2008). Following lengthy reform efforts China's major banks and financial firms have begun to sell partial ownership stakes to overseas

financial companies.

4. 4. *From isolation to global engagement*. Beginning from near-autarchy during the 1960s and 1970s, China has gradually emerged as a leading participant in global trade. China's 2001 entry into the World Trade Organization (WTO) capped a gradual process of opening that has raised the ratio of combined imports and exports to GDP from under ten per cent prior to the reform to over 63 per cent in 2005 – surpassing comparable figures for all other large and populous nations (Lardy, 2002; Brandt, Rawski and Zhu, 2007). China has become the world's largest recipient of foreign direct investment, which initially clustered in manufacturing, but has recently extended into finance, property, retailing, logistics, infrastructure and R&D. Foreign firms have taken the lead in integrating China into multinational supply chains for manufacturing, research and design. Chinese firms have also begun to increase their own overseas investment in pursuit of raw materials, market access and knowledge.

Changes in institutions and public policies reflect these new economic realities. Administrative reforms have recast government ministries (of machinery, textiles and so on) as industry associations, which now engage in informal discussions and negotiations with official agencies, as do individual companies and interest groups (Kennedy, 2005). Fiscal reforms have sought to redress imbalance between central and local revenue shares and to enhance revenue buoyancy to keep pace with growing demands for spending on education, health care, pensions, infrastructure and environment.

Three decades of reform have reshaped China's economy into a hybrid that is increasingly responsive to domestic and international market forces even though some segments, for example, capital markets and investment spending, reflect the continued legacy of planning.

Key factors in China's reform success

Although the period since the late 1970s has brought huge increases in output, productivity, and incomes, China's reforms remain far from complete (Lardy, 1998). The costs and inefficiencies associated with unfinished or delayed reform are large. They include remnants of the plan era, for example the underpricing of energy, water, and bank loans, which exacerbates China's environmental and employment problems. Some stem from the reform itself, for instance the continuing epidemic of rent-seeking and graft. Others, including the consequences of weak systems of environmental management, law, public finance, banking, and investment allocation, reflect halfway houses that combine inherited political and economic structures with partial reform efforts (Pei, 2006).

How has China's reform achieved so much when its economic system contains so many weak links? China's recent experience encourages us to think of a hierarchy of desirable features that support growth or, if absent, hinder it. These growth-enhancing conditions are not equally important. In China, partial measures affecting incentives, prices, mobility, and competition – what we might term ‘big reforms’ – created a powerful momentum that overwhelmed the friction and drag arising from a host of ‘smaller’ inefficiencies associated with price distortions, imperfect markets, institutional shortcomings, and other defects that retarded growth and increased its cost but never threatened to stall the ongoing boom (Perkins, 1994).

In the presence of large gaps between current and potential output, and of neglected opportunities for

expanding the production frontier, limited reform that even partially ruptures the shackles surrounding incentives, marketing, mobility, competition, price flexibility and innovation may accelerate growth. Begin with an economy operating well below its potential, partly because its workers, perceiving that effort hardly affects their incomes, withhold much of their available energy (which itself is reduced by chronic undernutrition). Now restore the link between effort and reward, permit a partial market revival, and open the door to experimentation with international trade and investment. Without disruptive changes in trade flows and political structures that accompanied early reform efforts in the former Soviet Union and Eastern Europe, such simple initiatives – which approximate the circumstances of China's early reforms – can readily ignite a burst of growth, even if prices, financial institutions, judicial enforcement, policy transparency, corporate governance and many other features of the economy remain far from ideal.

A review of what we call ‘big reforms’ explains the unexpected coincidence of stunning growth with deeply flawed institutions.

Incentives. In China, restoring the link between effort and reward was hugely beneficial even with large price distortions and a limited market activity. The shift from collective to household farming produced an immediate surge in agricultural production even though the farm sector of the 1980s embodied fewer ‘free market’ characteristics than Chinese agriculture of the 1920s and 1930s, or even the early 1950s. The same observation applies to private business, which has expanded rapidly and become the largest source of new employment despite its limited access to official support, legal protection and formal credit markets.

Prices. The expansion of price flexibility, most notably through the dual price system, thrust market forces into the economic lives of all Chinese households and businesses. Participants in China's economy – including the large state-owned enterprises at the core of the plan system – suddenly faced a new world in which market prices governed the outcome of marginal decisions to sell above-plan output or to purchase materials and equipment. This partial and gradual liberalization of pricing opened the door to what Naughton (1995) has dubbed ‘growing out of the plan,’ in which directing incremental output towards market allocation gradually reduced the importance of the plan sector without a political struggle.

Mobility. As the reform progressed, rising urban incomes created new demands for labour in China's cities and towns, especially in construction, services and in new export industries. Responding to this demand, individual villagers began to circumvent regulations that had long barred rural workers from moving to the cities. With the assistance of would-be urban employers and of rural governments, the initial trickle of migration expanded into the largest internal migration in world history.

Partial liberalization of prices, which allowed cash markets to sell food and other necessities with no requirement for residence-based ration tickets, provided essential support for this growing flow of migrant labour. As with the earlier shift from collective to household farming, massive change responded to price signals that, however imprecise, indubitably reflected underlying resource scarcities. Villagers did not need an exact calculation to see that they could raise their incomes by taking up non-farm occupations; several hundred million recognized the opportunity and made the choice.

Competition. Planning attempts to reduce economic uncertainty by pairing suppliers with customers and by specifying the nature of future transactions. Planning also controls the entry of new firms and the exit of weak enterprises. In China, the expansion of incentives, mobility, and markets created unprecedented opportunities to rearrange supply links, to establish new enterprises and to develop existing firms (both

domestic and foreign) by commercializing new products and pursuing new markets. Entry squeezed profits (Naughton, 1992). The state, as the main owner of enterprise assets, suffered the financial consequences, as the GDP share of fiscal revenue suffered a long decline (Wong and Bird, 2008). The resulting fiscal pressures encouraged officials at all levels to respond to pleas from hard-pressed enterprises by allowing piecemeal expansion of reform (Jefferson and Rawski, 1994).

The scale of entry and exit is startling. The number of industrial firms rose from under 0.4 million in 1980 to nearly 8 million in 1990 and 1996; the 2004 economic census, which excluded enterprises with annual sales below RMB5 million, counted 1.33 million manufacturing firms (Jefferson and Singh, 1999, p. 25; Economic Census, 2004, pp. 1, 2, 23); in construction, the number jumped from 6,604 to 58,750 between 1980 and 2005, with the latter total excluding subcontractors (Yearbook, 2006, p. 579). On the exit side, bankruptcy and restructuring have eliminated many weak firms: between 2001 and 2004, for example, the number of state enterprises in all sectors declined by 177,700 (State Council, 2005). Employment in state-owned industry dropped from 45.2 to 8.9 million between 1992 and 2005 (Yearbook, 1996, p. 402; 2006, p. 505).

Although Young (2000) and others argue that internal trade barriers limit domestic competition by obstructing the flow of goods and funds across provincial and other administrative boundaries, we believe that the impact of such barriers has faded, allowing rapid expansion of road traffic, telecommunications, chain stores, supply networks and other new developments to push China's economy towards extraordinarily high levels of competition. Despite pockets of monopoly and episodic local trade barriers, intense competition now pervades everyday economic life. The auto sector provides a perfect illustration: two decades of competition have sucked a lethargic state-run oligopoly into a whirlwind of rivalries in which upstarts such as Chery and Geely wrestle for market share with state-sector heavyweights and global titans. The payoff – rapid expansion of production, quality, variety, and productivity, along with galloping price reductions – has injected a dynamic new sector (not just manufacture of vehicles, components and materials, but also auto dealers, service stations, parking facilities, car racing, publications, motels, tourism, and so on) into China's economy.

The auto sector also illustrates how economic opening has ratcheted up competition throughout China's economy. With few sectors sheltered from imports and with foreign-linked firms participating in a growing array of domestic activities, incumbent suppliers of soybeans, machine tools, retail services, and an endless array of other goods now face competition from rival producers in America, Japan or Brazil as well as Jilin, Zhejiang and Sichuan.

Price wars and advertising, two unmistakable signs of competition, have become commonplace. Chinese newspapers are filled with accounts of fierce price competition among producers of autos, televisions, microwaves, air conditioners, and many other products. Advertising expenditure in 2006 matches total urban retail sales for 1990 (Nielsen Media Research, 2006; Yearbook, 2006, p. 678). The decline of former industry leaders like Panda (televisions) and Kelon (home appliances) and the ascent of new pacesetters like Wahaha (beverages), Wanxiang (auto parts) and Haier (home appliances) from obscure beginnings show how competition has added new fluidity to Chinese market structures.

Innovation. Prior to reform, China experienced a general failure of dynamic efficiency. Under the plan system, apart from exceptional instances of direct high-level intervention (‘innovation by order’), producers neglected innovation in favour of pursuing short-term targets for physical output (‘fulfilling the plan’). As a result, the expansion of society's production frontier lagged behind the potential

embodied in available knowledge and resources. The consequences are readily visible: First Auto Works, one of China's premier manufacturers, found its ‘obsolescence of equipment and models worsening day by day’ following ‘30 years of standing still’ under the planned economy (Li Hong, 1993, p. 83).

Reform put an end to this stand-pat mentality by widening the gap between financial outcomes for strong and weak firms, their managers and their employees. The presence of price distortions, subsidies and official intervention could not obscure the central issue: do we pursue innovation in order to maintain and perhaps expand our sales, market share, profits, wages, and employment security, or do we sit tight and hope that current or potential rivals do not leave us behind? Especially since China's entry into WTO, the proportion of firms engaging in R&D has grown rapidly, as has the ratio of R&D spending to GDP (Hu and Jefferson, 2008).

On the supply side, efforts to upgrade the quality and variety of products benefited from rapid increases in China's supply of educated workers. China's growing engagement with the global economy created immense inflows of new technology, not just from imports of equipment and know-how, but from new links connecting millions of Chinese workers, engineers, and managers with the technical standards, engineering processes and management practices needed to compete in global markets.

Key elements in the political economy of Chinese reform

What of the policy process associated with these extraordinary changes? Despite the authoritarian nature of China's political system, pre-reform policy structures allowed widespread experimentation and regional variation within broad guidelines set at the centre. This encouraged local officials to develop strategies whose success might attract high-level attention and also allowed national leaders to ‘play to the provinces’ (Shirk, 1993) by assembling coalitions of like-minded officials to demonstrate the merits of their preferred policy options and to lobby for nationwide implementation of those policies.

This arrangement, under which national policies emphasized broad principles or parameters rather than specific instructions or regulations, continued into the reform period. What changed is the content of the directives articulated at the centre, formerly directed towards ideological matters, which now focused increasingly on issues surrounding economic growth.

Looking beyond the principles emanating from the top, we see three additional elements as completing the skeleton of China's reformist political economy. *Decentralization* endows provinces and localities with both the resources and the incentive to experiment with local approaches to specific policies (for example, rural industrialization) and difficulties (for example how to deal with redundant state-sector workers), providing they observe central guidelines. *Competition* within the political system is not new, but now focuses on economic outcomes, which exercise increasing leverage over the career paths of leaders at every level. Continued promotion and recruitment of leaders whose reputation and career prospects rest on past and future economic success has gradually created a large and expanding *coalition among growth-minded, market oriented individuals and groups* within China's policy elite, whose power and influence helps to shift the content of central guidelines towards market outcomes.

Broad guidelines – what they can and cannot do

Chinese tradition emphasizes the government of men (and, beginning in the late 20th century, some women) rather than laws. In the absence of detailed instructions, how do China's top leaders direct the behaviour of lower-level governments and individual officials? Functionaries at all levels study and discuss the speeches and writings of top leaders, which lay out the desired course of public policy and explain what lower levels of officialdom should and should not do. These guidelines become encapsulated in catchy slogans that gain wide currency. In turn, these slogans, and the policy guidelines that inform them, direct the flow of policy implementation at all levels.

From the start of China's reform in the late 1970s, these directives increasingly emphasized economic matters. Indeed, China's political economy has come to rest on a grand but unspoken bargain between the Communist Party and the Chinese public in which the party ensures economic growth and promotes China's global standing in return for public acquiescence to its autocratic rule and anachronistic ideology (Keller and Rawski, 2007b). As a result, the articulation and fulfilment of key economic objectives now constitute core ingredients in extending the political legitimacy of the Chinese state. Economic objectives embedded in documents, speeches, and slogans reverberate at every level of society, where they become benchmarks for evaluating current or proposed actions. Deng Xiaoping's praise of reform during his southern tour of 1992 was widely seen as a favourable signal for policy innovations, including many that received no specific mention from him. In similar fashion, emphasis (or omission) of praise for ‘small and medium enterprises’ will be interpreted as high-level encouragement of (or caution against) policies favouring private business.

Decentralized experimentation

The experience of the 20th century surely qualifies the Chinese as the world's leading practitioners of economic experimentation. China's reform economy amply displays this characteristic. We see the national government conducting trials of novel institutions, for example ‘special economic zones’, while provinces and localities develop their own variations of pension systems, industrial regulation, and so on. The decentralization of industry, which placed all but the largest enterprises under the control of lower-level governments, and of public finance, which, especially prior to the 1994 fiscal reforms, assigned major revenue streams to provincial and local administrations, provided regional and local governments with ample resources with which to pursue such experimentation.

Competition

Prior to the inception of reform, China developed a tradition of policy entrepreneurship in which local figures compete for high-level attention by demonstrating the beneficial implementation of the principles enshrined in broad central directives. This competition intensified under the reform, with GDP growth and other economic criteria replacing ideological benchmarks as the arbiters of success. Thus Li and Zhou (2005) find that promotion prospects for provincial leaders rise, and the likelihood of termination declines as provincial economic performance improves. Whiting (2001) makes similar observations about local officials.

Officials at all levels possess the authority as well as the resources needed to promote local growth. They also have strong incentives to do so, because their career prospects, as well as personal financial

opportunities for themselves and their families, are closely tied to the economic trajectory of the jurisdictions under their leadership. Growth expands the pools of public revenue and enterprise profits over which officials exercise varying degrees of control, enlarges business opportunities available to the families and associates of local leaders, and swells the flow of (legal and illicit) rents directed towards official agencies and their managers.

These circumstances have transformed China's local and provincial governments into eager champions of development, each striving to outdo its neighbours in expanding infrastructure and strengthening the foundations of ‘pillar industries’. This competition contributes mightily to the persistent ‘investment hunger’ visible in China's economy, as local administrations resist central calls for restraint in enlarging existing facilities and building new ones.

Pro-growth coalition

China's reform leaders, like politicians everywhere, endeavour to appoint and promote like-minded successors and subordinates. As Shirk (1993) and others have noted, the reform movement's initial successes acted as a powerful recruiting device, with the lure of rich payoffs adding many influential converts to the cause of reform. As the reform gained momentum, the circulation of elites, including the assignment of successful officials to lagging regions for the express purpose of jump-starting growth, created mentor-student relationships between growth-oriented officials and increasing numbers of would-be imitators. The widespread practice of sending study teams to absorb the ‘advanced experiences’ of dynamic localities further expanded the reform constituency among China's policy elites.

Of particular importance is the legacy of the Cultural Revolution, which truncated educational opportunities for whole cohorts of Chinese. This historical accident created a unique opportunity to advance the reform agenda. When the retirement of Deng Xiaoping and other ‘revolutionary elders’ focused attention on generational change, reformist leaders managed to bypass the customary emphasis on seniority, skipping over the ‘lost generation’ of Cultural Revolution victims to promote younger candidates. The increasing prominence of university graduates, including returnees from overseas study and young professionals with close ties to international business, accelerated the development of what became a loose and unorganized but increasingly potent coalition of like-minded officials whose objectives centred on growth-promoting and increasingly market-oriented reforms.

Despite these gains, the evolution of policy towards private business demonstrates the difficulty of translating power and influence into genuine institutional change. Legal documents confirm the painfully slow expansion of official protection. At the start of reform, private business operated in a legal limbo. Some entrepreneurs disguised their firms as collectives; others purchased informal protection from powerful individuals or agencies. A succession of amendments to China's 1982 constitution slowly expanded recognition of the non-public economy, first as a ‘complement’ to the state sector (1988), than as an ‘important component’ (1999) of the ‘socialist market economy’ (itself a new term dating from 1993). The ‘Law on Solely Funded Enterprises’, which took effect in 2000, guaranteed state protection for the ‘legitimate property’ of such firms, but without using the term ‘private’ or specifying any agency or process to implement this promise.

Further constitutional amendments adopted in 2004 breached the former taboo on the term ‘private’ by stating that ‘citizens’ lawful private property is inviolable’. The long march towards official recognition of private business came to an end only in 2007 when, following five years of fierce debate, China's

legislature enacted a landmark Property Rights Law which, for the first time, explicitly places privately held assets on an equal footing with state and collective property.

Conclusion

Reform has delivered enormous economic gains despite deep and potentially dangerous flaws in China's institutions and policy structures. The same framework of structures and incentives that spurs rapid economic advance also generates ambiguous and often disturbing consequences along other socioeconomic dimensions. Environment and inequality illustrate the range of outcomes.
Economy (2004) and others demonstrate how China's unbridled rush to maximize GDP growth, together with weak regulatory and legal structures, has produced environmental degradation on a scale that far exceeds internationally acceptable standards. Historical comparisons also show that improved technology and the spread of environmental consciousness among China's growing middle class are pushing China towards regulation and remediation of atmospheric and water pollution at an earlier stage of the development process than occurred in Japan, Korea, or the United States.
China's reforms have literally pulled hundreds of millions out of poverty, especially in the countryside. Reform has also increased China's income inequality to levels that now approach some of the highest in the developing world. Although attention focuses on income gaps between urban and rural areas and between coastal and interior provinces, growing income differences between neighbours within provinces and within the urban and rural sectors account for most of the increase in inequality (Benjamin et al., 2008). In rural areas, this increase is tied to the disequalizing role of some forms of non-agricultural income, and laggard growth of farming income, especially beginning in the mid-1990s. In urban areas, a decline in the role of subsidies and entitlements, increasing wage inequality related to labour market and enterprise reform, and the effect of SOE restructuring on some cohorts and households have enlarged the dispersion of incomes. Rising returns to human capital and differences in access to education have widened income differences in all sectors. Corruption, although difficult to quantify, may also have contributed to growing inequality of wealth and welfare.
Despite these and other difficulties, China's recent experience demonstrates that activating key economic drivers, including incentives, mobility, prices, competition, and innovation, can unleash sufficient momentum to overwhelm a variety of system costs. China's economic boom, in 2007 completing its third decade, rests on a unique set of historical circumstances, some favourable, others less so.
China's success cannot ensure the efficacy of ‘Chinese policies’ in other times and places. There is also no guarantee that the mechanism described in this article can enable China to extend its enviable record of high speed growth. Even so, China's continuing accumulation of physical resources and human capital, the intense focus of public policy on promoting growth, and the willingness of China's leaders to implement bold initiatives create a favourable climate for further reform and continued economic expansion.

See Also

- China, economics in
- dual track liberalization

- Maoist economics
- soft budget constraint

Bibliography

Benjamin, D., Brandt, L., Giles, J. and Wang, S. 2008. Income inequality during China's economic transition. In Brandt and Rawski (2008).

Brandt, L., Hsieh, C.-T. and Zhu, X. 2008. Growth and structural transformation in China. In Brandt and Rawski (2008).

Brandt, L. and Rawski, T.G., eds. 2008. *China's Great Economic Transformation*. New York: Cambridge University Press.

Brandt, L., Rawski, T.G. and Zhu, X. 2007. International dimensions of China's long boom: trends, prospects and implications. In *China and the Balance of Influence in Asia*, ed. W.W. Keller and T.G. Rawski. Pittsburgh: University of Pittsburgh Press.

Economic Census. 2004. *Zhongguo jingji pucha nianjian 2004* [Yearbook of China's 2004 Economic Census], 4 vols. Beijing: Zhongguo tongji chubanshe.

Economy, E. 2004. *The River Runs Black: The Environmental Challenge to China's Future*. Ithaca: Cornell University Press.

Heston, A., Summers, R. and Aten, B. 2006. Penn World Table Version 6.2, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September. Online. Available at http://pwt.econ.upenn.edu, accessed 15 June 2007.

Hu, A.G.Z. and Jefferson, G.H. 2008. Science and technology in China. In Brandt and Rawski (2008).

Jefferson, G.H. and Rawski, T.G. 1994. Enterprise reform in Chinese industry. *Journal of Economic Perspectives* 8(2), 47–70.

Jefferson, G.H. and Singh, I. 1999. *Enterprise Reform in China: Ownership, Transition, and Performance*. Oxford and New York: Oxford University Press.

Keller, W.W. and Rawski, T.G. 2007a. *China and the Balance of Influence in Asia*. Pittsburgh: University of Pittsburgh Press.

Keller, W.W. and Rawski, T.G. 2007b. China's peaceful rise: roadmap or fantasy? In Keller and Rawski (2007a).

Kennedy, S. 2005. *The Business of Lobbying in China*. Cambridge, MA: Harvard University Press.

Lardy, N.R. 1983. *Agriculture in China's Modern Economic Development*. Cambridge: Cambridge University Press.

Lardy, N.R. 1998. *China's Unfinished Economic Revolution*. Washington, DC: Brookings Institution.

Lardy, N. 2002. *Integrating China into the Global Economy*. Washington, DC: Brookings Institution.

Lardy, N.R. and Lieberthal, K. 1983. Ch'en Yün's Strategy for China's Development: A Non-Maoist Alternative. Armonk: M.E. Sharpe.

Li, H. 1993. *Zhongguo qiche gongye jingji fenxi* [Economic Analysis of China's Auto Industry]. Beijing: Zhongguo renmin daxue chubanshe.

Li, H. and Zhou, L.-A. 2005. Political turnover and economic performance: the incentive role of personnel control in China. *Journal of Public Economics* 89, 1743–62.

Li, X. 2006. *Assessing the Extent of China's Marketization*. Aldershot: Ashgate.

Lin, J.Y. 1992. Rural reforms and agricultural growth in China. *American Economic Review* 82, 34–51.

National Bureau of Statistics. 2006. National bureau of statistics of China. Statistical Communiqué of the People's Republic of China on the 2006 National Economic and Social Development. Beijing: China Statistics Press.

Naughton, B. 1992. Implications of the state monopoly over industry and its relaxation. *Modern China* 18, 14–41.

Naughton, B. 1995. *Growing out of the Plan: Chinese Economic Reform, 1978–1993*. Cambridge and New York: Cambridge University Press.

Nielsen Media Research. 2006. Review of the Chinese advertising market in 2006. Online. Available at http://www.pata.org/patasite/fileadmin/news_pata/2007/Nielsen_ad_market_review_06_eng.pdf, accessed 20 May 2007.

OECD (Organisation for Economic Co-operation and Development). 2005. *OECD Economic Surveys: China*. Paris: OECD.

Pei, M. 2006. *China's Trapped Transition: The Limits of Developmental Autocracy*. *Cambridge*, MA:

Harvard University Press.

Perkins, D.H. 1994. Completing China's move to the market. *Journal of Economic Perspectives* 8(2), 21–46.

Perkins, D.H. and Rawski, T.G. 2008. Forecasting China's economic growth to 2025. In Brandt and Rawski (2008).

Shirk, S.L. 1993. *The Political Logic of Economic Reform in China*. Berkeley: University of California Press.

Sicular, T. 1995. Redefining state, plan, and market: China's reforms in agricultural commerce. *China Quarterly* 144, 1020–46.

State Council. 2005. Report No. 1 of the main data from the first national economic census. Chinese document from the Office of the State Council Leading Small Group for the First National Economic Census and the National Bureau of Statistics, 6 December 2005. Online. Available at http://news.xinhuanet.com/fortune/2005-12/06/content_3883969.htm, accessed 20 May 2007.

Whiting, S.H. 2001. *Power and Wealth in Rural China: The Political Economy of Institutional Change*. Cambridge: Cambridge University Press.

Wong, C.P.W. and Bird, R. 2008. China's fiscal system: a work in progress. In Brandt and Rawski (2008).

Yearbook. 1996. *Zhongguo tongji nianjian 1996*. [China Statistical Yearbook 2006]. Beijing: China Statistics Press.

Yearbook. 2006. *Zhongguo tongji nianjian 2006*. [China Statistical Yearbook 1996]. Beijing: China Statistics Press.

Young, A. 2000. The razor's edge: distortions and incremental reform in the People's Republic of China. *Quarterly Journal of Economics* 115, 1091–136.

How to cite this article

Brandt, Loren and Thomas G. Rawski. "Chinese economic reforms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000600> doi:10.1057/9780230226203.0233

The New Palgrave Dictionary of Economics Online

Christaller, Walter (1894 – 1975)

Gordon C. Cameron

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

central place theory; Christaller, W.; market threshold; normal travelling distance

Article

Christaller, who never held an academic post but worked throughout his life in association with the University of Erlangen, is known for one seminal book *Die zentralen Orte in Süddeutschland* [Central Places in Southern Germany]. Published in Germany in 1933 it remained largely unnoticed by English-speaking scholars until a translation of August Lösch's *Economics of Location* (1954) brought it widespread attention. Later an accurate translation of Christaller's book by C.W. Baskin (in 1966) confirmed the elegance of his deductive theorizing.

Christaller sought to clarify and explain the laws which determine the number, sizes and distribution of towns. Drawing upon the work of von Thünen, Alfred Weber and Engländer, Christaller developed a general theory of why a *hierarchy* of villages and towns providing different services should appear and why this hierarchy should differ region by region. Making use of key concepts of market threshold, and normal travelling distance, he showed how the geographical extent of the trading areas for different goods and services vary and how low order centres provide limited ranges of goods to small trading areas whereas larger centres service much wider areas and contain all the goods of the lower centres as well as goods unique to their size.

Christaller's work has been criticized as ignoring the role of manufacturing in shaping the growth of towns and cities, of underplaying the effects of an unequal distribution of natural resources and of an all too rigid expression of the laws of market size and of the hierarchy of central places. Of the last point Christaller was fully aware and by 1950 he had modified his stance allowing for greater variability in the determinants of the hierarchy. And though his general theory of spatial relations is incomplete, all subsequent analysts of retail trade, of the location of services and of urban growth, recognize the rigour of his approach and the elegance of his attempt to provide the ‘economic theoretical foundations of town geography’.

See Also

- central place theory

Selected work

1933. *Die zentralen Orte in Süddeutschland: eine ökonomisch-geographische Untersuchung uber die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit stadtischen Funktionen*. Jena: G. Fischer Trans. C.W. Baskin as Central Places in Southern Germany, Englewood Cliffs, NJ: Prentice-Hall, 1966.

Bibliography

Berry, B.J.L. and Harris, C.D. 1970. Walter Christaller: an appreciation. *Geographical Review* 60, 116–19.

Lösch, A. 1954. *The Economics of Location*. Trans. from the 2nd (1944) edn by W.H. Woglom with the assistance of W.F. Stolper, New Haven: Yale University Press. 1st German edn, 1940.

How to cite this article

Cameron, Gordon C. "Christaller, Walter (1894–1975)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000136> doi:10.1057/9780230226203.0234

The New Palgrave Dictionary of Economics Online

circular flow

Giorgio Gilibert
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The analysis of the social process of production and consumption must start from some notion of commodity circulation. Consideration of the simple cycle of agricultural production suggests that production is an essentially circular process, in the sense that the same goods appear both among the products and among the means of production. From this viewpoint, commodity (as well as money) circulation is a triviality, whose discovery cannot really be attributed to any particular economist.

It has been suggested that the notion was originally developed by François Quesnay, a surgeon, by analogy with the circulation of the blood. However the popular analogy between money and blood is much older (see for instance ‘Money is for the state what blood is for the human body’, *Etats généraux*, 1484); and the process of money and commodity circulation among different classes (landlords, labourers, merchants) and areas (town and country) was clearly described by Boisguillebert and Cantillon several decades before the physiocrats.

What is truly novel with Quesnay is the idea that the essential task of economic science is the investigation of the technical and social conditions which allow the repetition of the circular process of production. This approach (at least in the extreme form given it by the physiocrats), and the peculiar model building activity that sprang from it, was later abandoned by economists. More than a century had to pass before the theme could be resumed, following the publication of Marx's own *tableaux* in the second volume of *Capital* (1885), but merely within the rather limited and isolated group of the German and Russian theoretical economists.

Tugan-Baranowsky considered circularity as the essential feature of capitalist economy, in which production was the end of consumption rather than the other way round; in his view, the economists were unable to understand this ‘paradox’ because (with the remarkable exception of Marx) they had strayed from the way opened up by Quesnay. The young Schumpeter, in a justly celebrated essay, dated the birth of economics as a science from the physiocratic analysis of the circular flow. And Leontief (1928) wrote in a similar vein, arguing in favour of the substitution of the principle of circular flow (the ‘reproducibility viewpoint’) for that of *homo oeconomicus* (the ‘scarcity viewpoint’) as the cornerstone of economic theory.

The reproducibility viewpoint is shared by the whole classical tradition of political economy. However, within this broad theoretical tradition, we can single out a radical strand which considers the economic behaviour of every individual as completely determined by the reproduction requirements of the system. This peculiar approach characterizes the pure theorists of the circular flow, with whom we will now briefly deal. Not surprisingly, this theoretical approach is often associated with a practical attitude in favour of some sort of central planning (as a consequence of the distrust for the ‘anarchy’ of the market).

The *Tableau Economique* depicts all the transactions taking place during the year among the three basic classes of society: the class of landowners (L), the ‘productive’ class of farmers (P_a), and the ‘sterile’ class of manufacturers (P_m). These transactions can be summarized by a graph, where three points – one for each class – are connected by lines, representing the transactions; the lines are oriented according to the direction of the money flows, whose value is shown by numbers (thousand millions of *livres*). Figure 1 is drawn on the data of Quesnay (1766); since the sum of the money flows leaving each point equals that of those coming in, the system is reproducible.

Figure 1

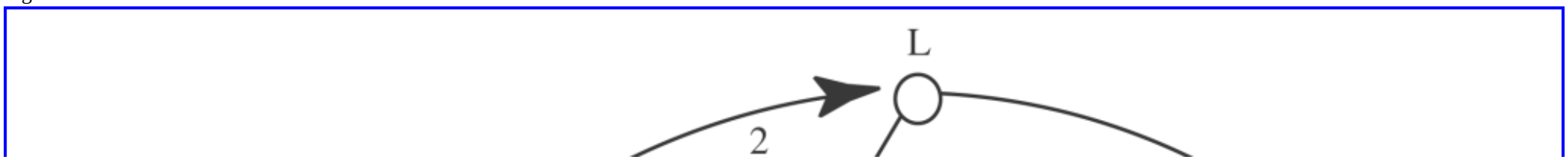

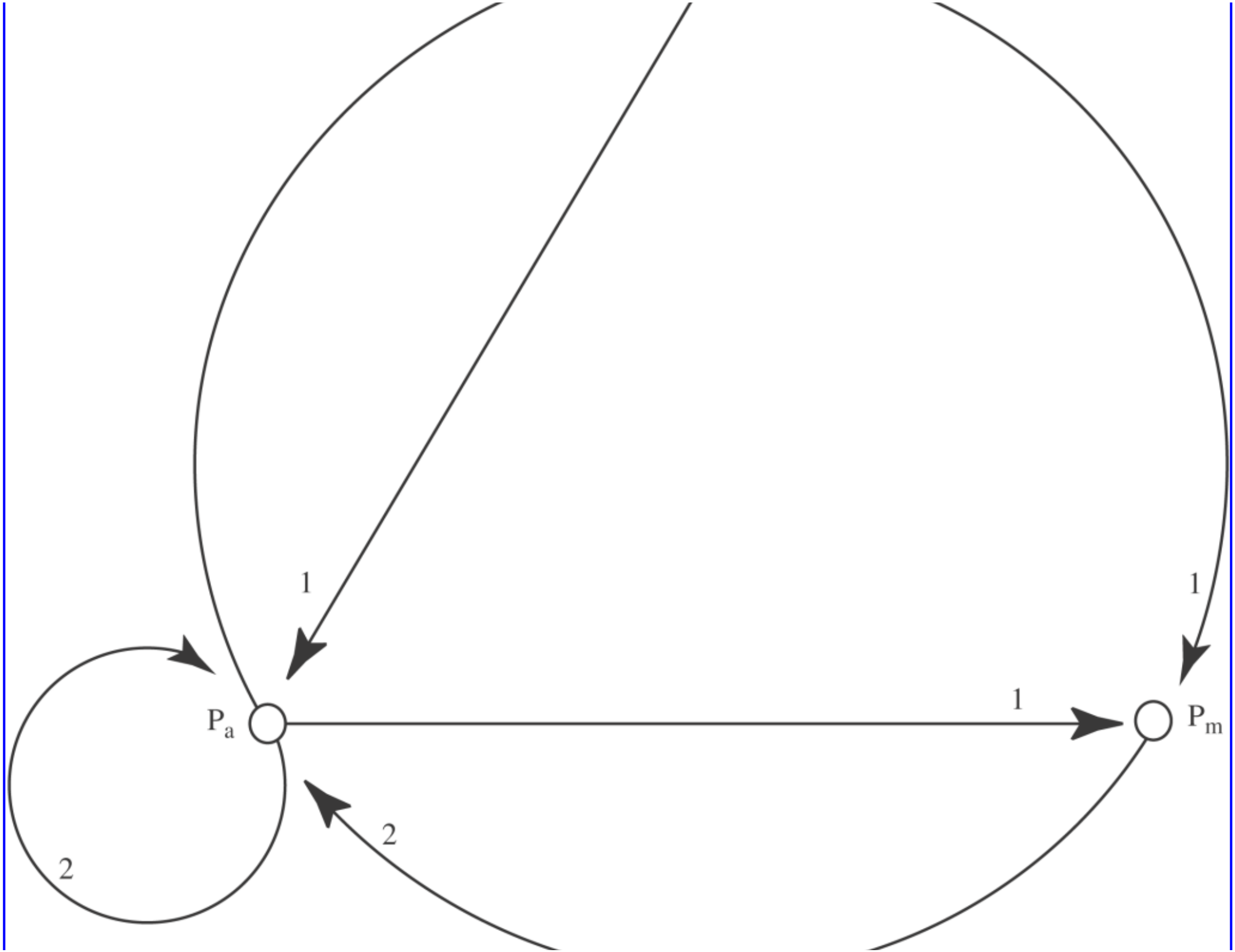
1
1
1
P_a
P_m
2
2

Marx's (simple) reproduction scheme can also be easily adapted to the same type of three-point graph, once capitalists are substituted for landowners, and the two industries producing intermediate goods (‘constant’ capital) and consumption goods (‘variable’ capital and luxuries) are substituted for the two classes of manufacturers and farmers respectively. It should be noted that, while Quesnay's *tableaux* are inherently static, Marx does also consider expanded reproduction: in his own words, the picture shifts from a circle to a spiral. A modern example of a circular representation of an expanding economy is the well-known von Neumann model, which, from this point of view, can be considered as the most sophisticated heir to the Marxian schemes.

Quesnay's and Marx's *tableaux* were offered in value terms; but there is no conceptual difficulty in imagining analogous schemes in physical terms. Now, if all the physical transactions taking place among all the agents of the economy are known, there is a unique set of relative prices which makes it possible for the process to be repeated.

Let us consider an economy in which n producers produce n goods. If we know all the physical amounts x_{ij} of the various goods consumed by the different producers, and if the economy is closed (i.e. production equals consumption for each good), relative prices p_i are determined by the following linear homogeneous equations:

$$\sum_i x_{ij} p_i = p_j \sum_h x_{jh}. \qquad (1)$$

This theory or prices has now come to be associated with the closed Leontief model (1941), but it was originally formulated in the late 18th century by Achille Isnard. He considered a simple example with three producers and consistently computed the corresponding prices.

His example is illustrated by the graph of Figure 2: three points, one for each producer, are connected by lines, corresponding to the physical amounts exchanged; the lines are now oriented according to the physical commodity flows. Relative prices have to be such as to equalize the value of the flows leaving each point with that of the flows coming in; the loops at the vertices (self-consumption) are not relevant to our problem.

Figure 2

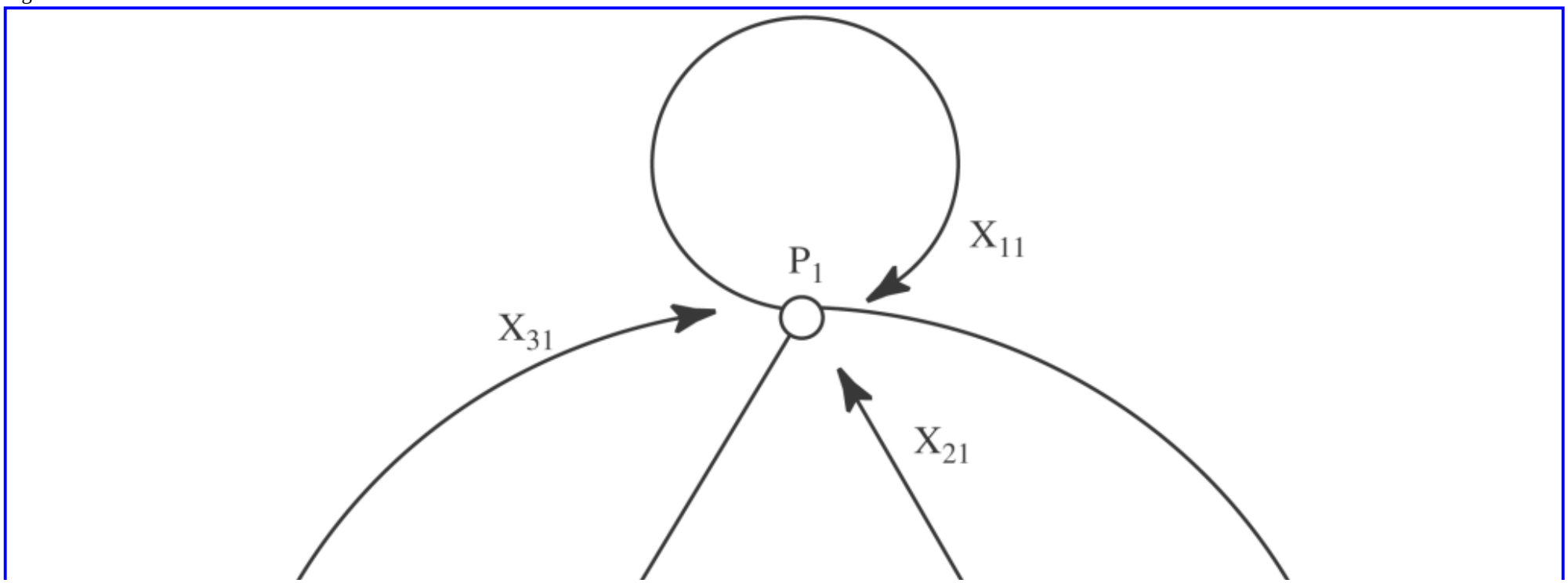

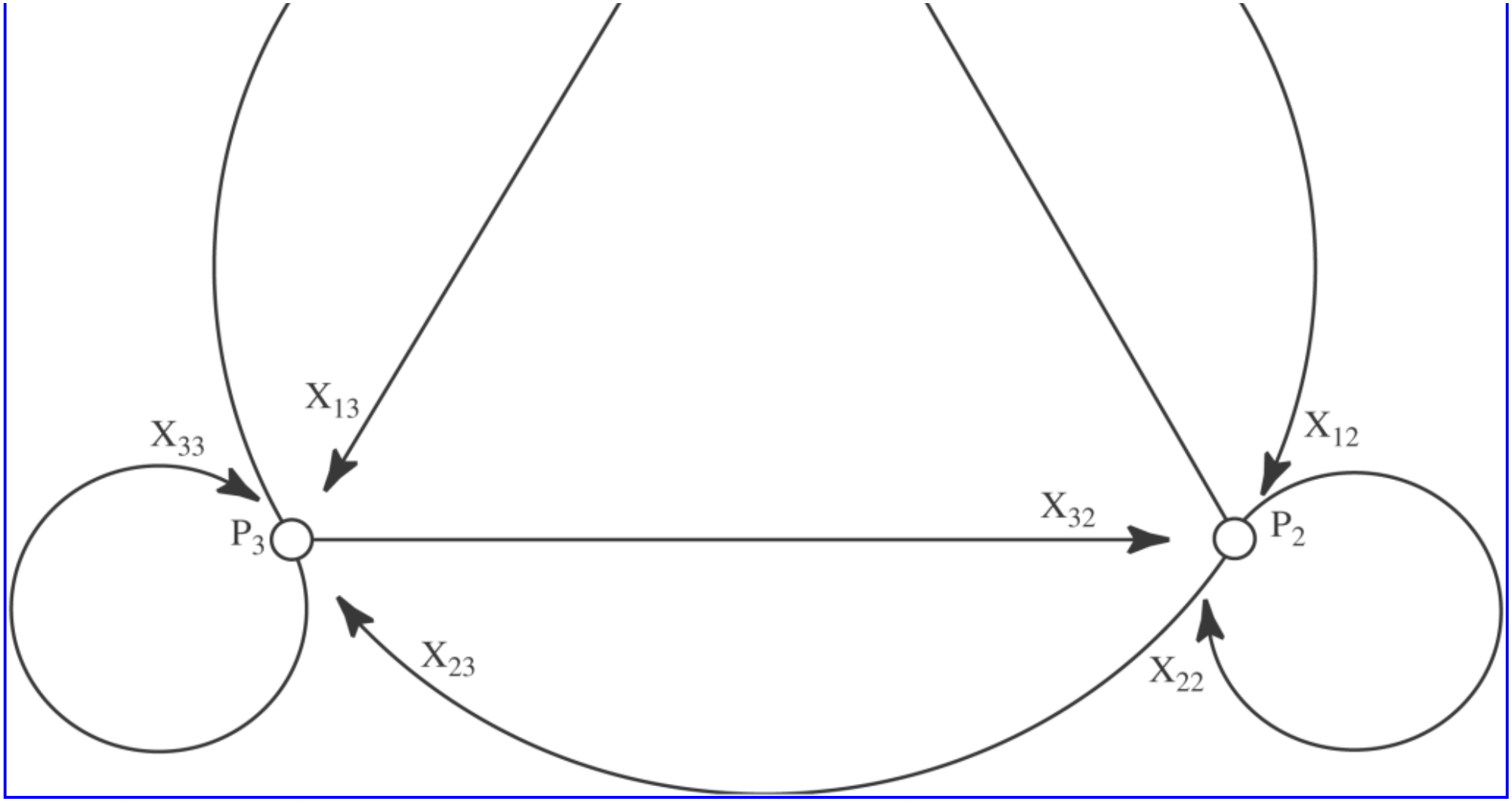

When Leontief, a century and a half later, rediscovered the theory, he recognized in it the ‘objective’ theory of value. One year later, the German mathematician Robert Remak interpreted system (1) as determining the rational prices for an economy in which the individual standards of living are fixed by a central authority. He showed that the system has in general meaningful solutions; and maintained that these prices could be practically computed and implemented.

Until now, we have considered only closed systems, in which all transactions are assumed as known irrespective of their nature (technical inputs or human ‘final’ uses). We can now open the model, by considering as given only those transactions which are dictated by the technology in use (including workers' subsistence) and leaving undetermined the final utilization of the surplus thus appearing.

There is now room for an additional relation, stating the way in which the surplus is distributed. If we assume that it is entirely appropriated by profit-earners in proportion to the capital advanced, we land on the familiar ground of the classical theory of production prices.

The case can be illustrated by a simple numerical example supplied by Sraffa: there are only two industries, producing wheat (P_a) and iron (P_m) respectively; the class of capitalists (C) gets the entire surplus, consisting only of wheat. In Figure 3 the numbers on the oriented graph refer to the physical quantities (quarters and tons) in the example.

Figure 3

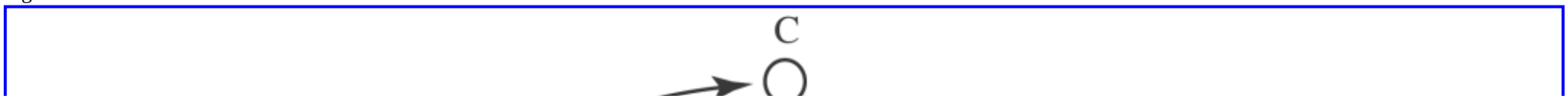

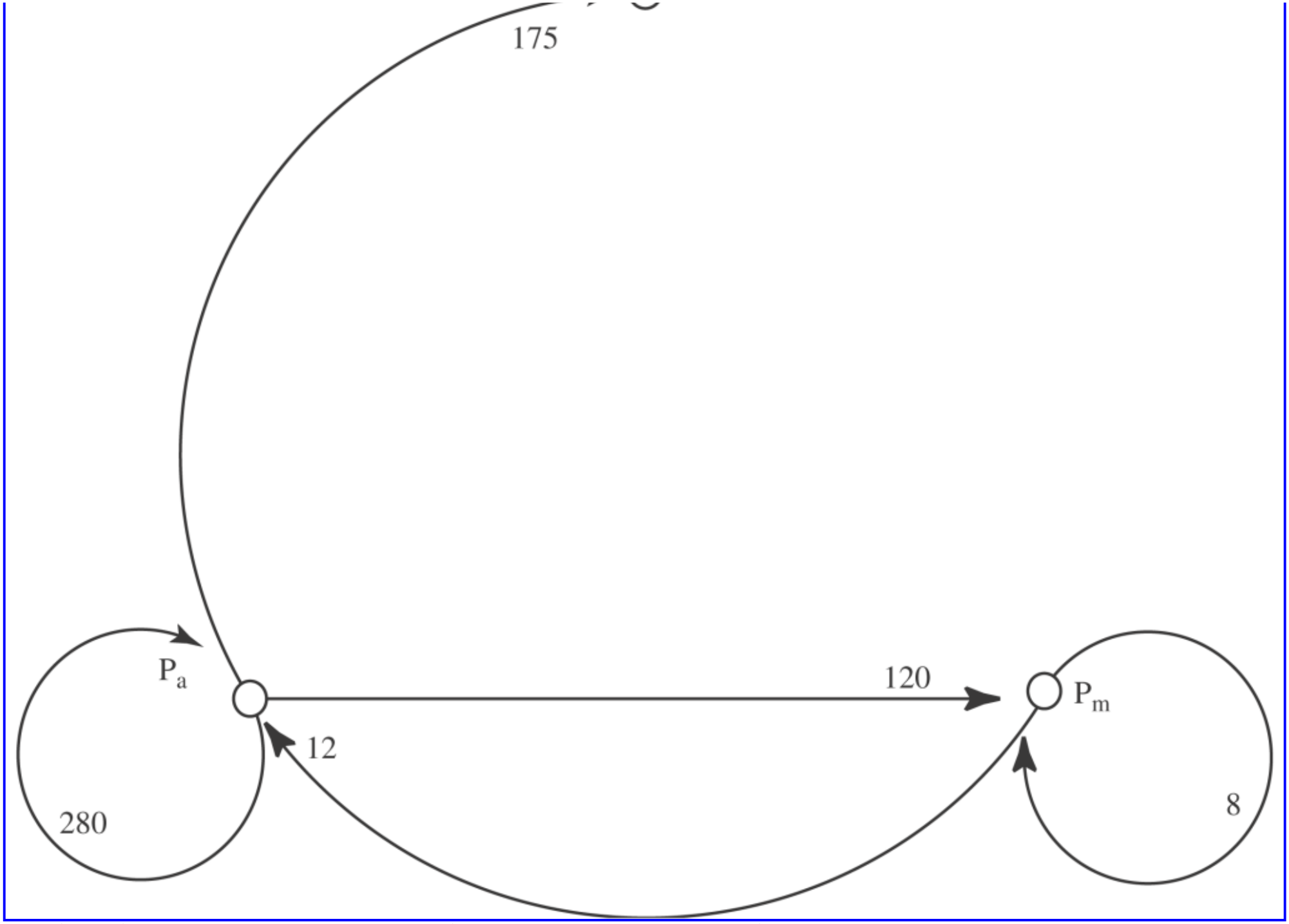

The uniform profit rate has to be such as to equalize the value of the surplus bought by capitalists to the profits accruing to them; and the exchange value between the two commodities has to be such as to enable each industry to replace its advances and to distribute profits in proportion to their value. Loops are now relevant.

The system is then reproducible when the money flows leaving each point are equal to those coming in; the situation is illustrated in Figure 4, and corresponds to a price of iron in

terms of wheat equal to 15 and to a common profit rate equal to 25 per cent.

Figure 4

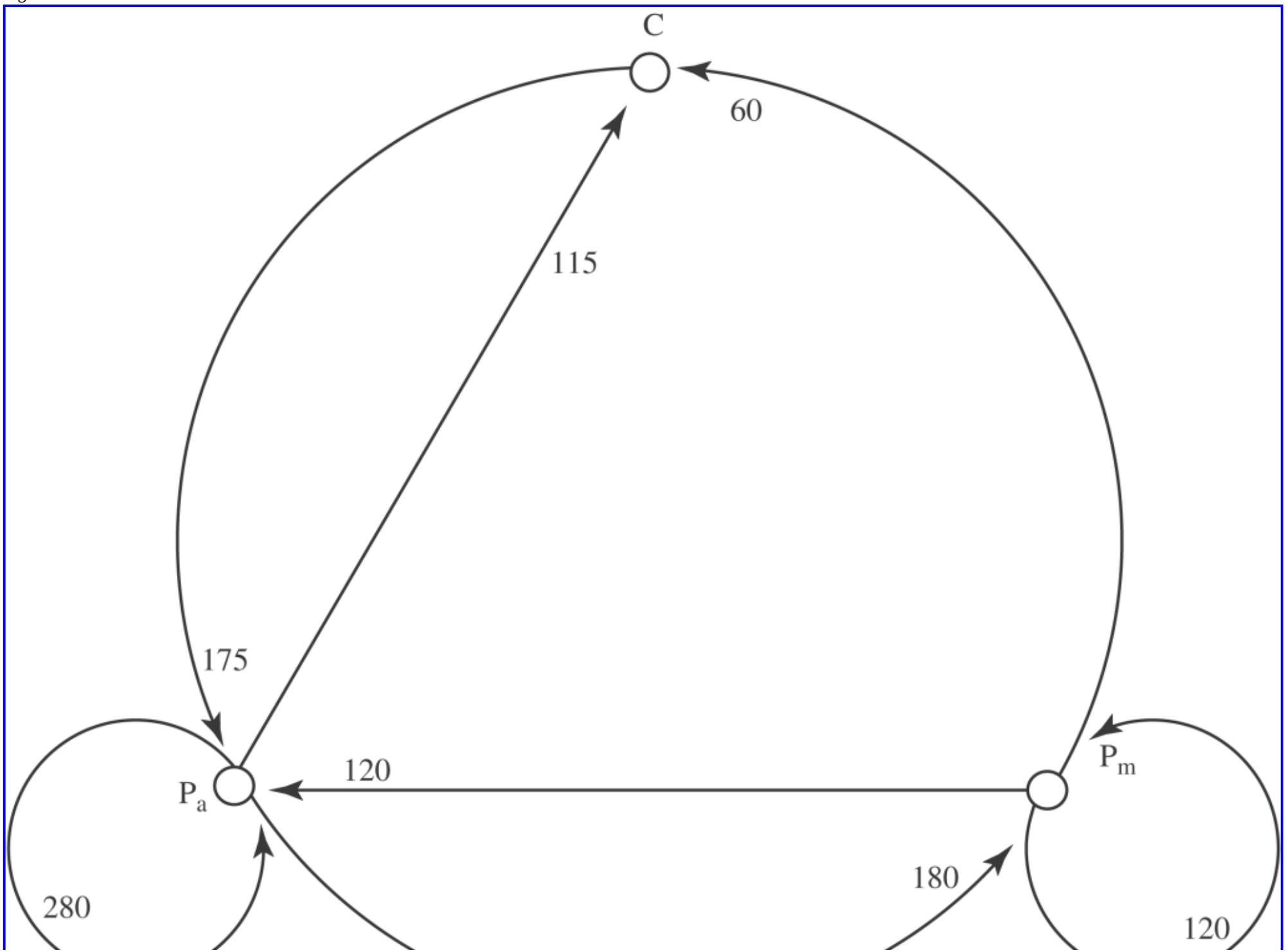

Finally, if we allow the wage earners to share the surplus with the capitalists, we generate the pure theory developed by Piero Sraffa (1960).

We are now able to interpret the abstract transition from our original circular theory to the classical theory of production prices, and eventually to its modern Sraffa version, as successive steps in a gradual opening of the model. From an initial system in which the economic behaviour of every individual is assumed to be rigidly determined by reproduction requirements, we have passed to a system in which capitalists (and rentiers) are assumed to be free in determining their final demand; and finally we have also granted some degree of freedom to the workers.

The term ‘free’ means here only that the composition of final demand is an issue which lies outside the domain of the pure theory of prices; of course, it can be the object of a distinct section of economic theory. In this perspective, we could say that the neoclassical theory of prices corresponds to a vision of the economy in which the individuals are supposed to be undifferentiated (i.e. there are no classes) and all equally free (the reproduction requirements do not play any essential role in determining prices).

See Also

- physiocracy
- Quesnay, François

Bibliography

Boisguillebert, P. de 1707. *Dissertation de la nature des richesses*. In *Oeuvres manuscrites et imprimées*, Paris: INED, 1966.

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Paris: INED, 1952.

Isnard, A.N. 1781. *Traité des richesses*. Lausanne: Grasset.

Leontief, W. 1928. Die Wirtschaft als Kreislauf. *Archiv für Sozialwissenschaft und Sozialpolitik 3*.

Leontief, W. 1941. *The Structure of American Economy*. Oxford and New York: Oxford University Press.

Marx, K. 1885. *Das Kapital*. Vol. II, Hamburg: Meissner.

Neumann, J. von. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums* VIII.

Peter, H. 1954. *Mathematische Strukturlehre des Wirtschafts-kreislaufes*. Göttingen: Schwartz.

Quesnay, F. 1766. Analyse de la formule arithmétique du Tableau économique. In Textes annotés, Paris: INED, 1958.

Remak, R. 1929. Kann die Volkswirtschaftslehre eine exakte Wissenschaft werden? *Jahrbücher für Nationalökonomie und Statistik* 76.

Remak, R. 1933. Können superponierte Preissysteme praktisch berechnet werden? *Jahrbücher fü Nationalökonomie und Statistik* 80.

Schumpeter, J. 1914. *Epochen der Dogmen und Methodengeschichte*. In *Grundriss der Sozialökonomik*, Tübingen: Mohr. Trans. by R. Aris as *Economic Doctrine and Method: an historical sketch*, Oxford: Philip Allen, 1954.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Tugan-Baranowsky, M. 1894. *Les crises industrielles en Angleterre*. French trans., Paris: Giard,1913.

How to cite this article

Gilibert, Giorgio. "circular flow." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000142>
doi:10.1057/9780230226203.0235

The New Palgrave Dictionary of Economics Online

circulating capital

Mark Blaug
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes the history of the distinction between circulating capital (whose full value returns to the capitalist from the sale of final goods) and fixed capital (whose value is never fully recovered in one production cycle) from its introduction by Smith and development by Ricardo to its treatment by Marx and the Austrian capital theorists. It gave rise to the wages fund doctrine, the problem of joint production, and the issue of the optimum rate of depreciation and replacement of old equipment.

Keywords

advances; Austrian capital theory; Böhm-Bawerk, E. von; capital theory; circulating capital; constant and variable capital; depreciation; fixed capital; Frisch, R. A. K.; joint production; marginal revolution; Marx, K. H.; Mill, J. S.; Quesnay, F.; Ricardo, D.; Smith, A.; Turgot, A. R. J.; wages fund; Wicksell, J. G. K.

Article

The explicit distinction between fixed and circulating capital first makes its appearance in Book II, chapter 1 of Adam Smith's *Wealth of Nations*, who derived it from ample hints in Quesnay and Turgot. Circulating capital goods, according to Smith, consist of those intermediate goods that embody a quantity of purchasing power that perpetually returns to the capitalist as he disposes of the final goods into the making of which they entered, in contrast to fixed capital goods, whose value is never fully recovered in one production cycle. The simplest example of circulating capital is raw materials, just as the simplest example of fixed capital is buildings and machines. However, all the classical economists, including Smith, included in circulating capital not just raw materials but also the consumer goods that support labour during the process of production; that is, wage goods.
This is the origin of the notorious ‘wages fund doctrine’, according to which wages are said to be ‘advanced’ to workers at the outset of a production period as a result of which they are determined by

the ratio between the volume of capital advanced and the size of the labour force. The notion arose out of a pronounced tendency in 18th-century economics to regard agriculture as an industry typical of production as a whole and to view wheat as both a representative output of agriculture and the staple article of consumption of workers. The fact that wheat only becomes available in the form of annual harvests, which must be willy-nilly stored as a ‘fund’ for future consumption if its actual use is to be more or less continuous throughout the year, made it possible to define capital simply as ‘advances’ to workers to support them from seed-time to harvest. Despite the fact that this agrarian model was gradually abandoned in the century after Smith, the wages fund doctrine lived on until J.S. Mill's recantation of the doctrine in 1867, and with it the definition of circulating capital as including all consumer goods that enter into the wage basket (Blaug, 1985, pp. 185–8). Surprisingly enough, this conception of capital as consisting largely if not solely of wage goods survived even beyond the ‘marginal revolution’: it lies at the heart of the theoretical schema adopted by Böhm-Bawerk in his *Positive Theory of Capital* (1887).

Adam Smith noted that fixed and circulating capital combine in different proportions in different industries, but it was Ricardo who converted this observation into one of the central facts of industrial life in a capitalist economy and a major problem for the theory of value. Ricardo wanted to argue that relative prices are determined by relative labour costs but, as he candidly admitted in the first chapter of the *Principles of Political Economy and Taxation*, this cannot be true, because not only does the ratio of fixed to circulating capital differ between industries but, in addition, the two kinds of capital may differ in durability between industries. Indeed, he added in a footnote, the distinction between fixed and circulating capital is not essential because any difference between them is solely a matter of degrees of durability; that is, the different time periods for which capital is locked up in the productive process: circulating capital is the sum of goods tied up in production for only as long as the period of production in question, whatever its length, whereas fixed capital is a joint output of this production period in the shape of a slightly older building or a slightly older machine. To put it in a nutshell: the distinction between fixed and circulating capital is not the difference in their absolute durability but rather the difference in their durability relative to the length of the production period in which they are employed.

Thus, despite the fact that Marx in *Capital* rejected the Smithian distinction between fixed and circulating capital and chose instead to distinguish ‘constant’ and ‘variable’ capital, confining the former to the wage bill and the latter to everything else on the grounds that wages might vary for a given production system even if all the technical input coefficients remained the same, he operated throughout the first volume of the book with a circulating capital model by virtue of the assumption that the capital stock of every industry in the economy turns over once a year: despite all the references to machinery in this first volume, all the analytical problems created by the use of fixed capital are eliminated by assuming that every industry operates with an annual production period. It is only in Volume 2 of *Capital*, and particularly chapters 8–14, that Marx takes account of differences in the durability or turnover rates of capital invested in different industries, and it is here that he begins to confront the problems created by the fact that fixed capital, unlike circulating capital, only transfers part of its value to the final product during each turnover of capital. This is the now famous problem of joint production, which, it has been argued (Steedman, 1977, ch. 10), may produce such anomalies as negative labour-costs for some products.

In the same way, all of the work of Böhm-Bawerk and most of that of Wicksell on the theory of capital is confined to the question of the optimum investment period of continuously applied circulating capital;

that is, to what Ragnar Frisch has called the ‘flow input–point output’ case. It is only when we take up the ‘point input–flow output’ or the even more typical case of ‘flow input–flow output’ that we confront the question of fixed capital, an issue that Böhm-Bawerk consistently avoided and that Wicksell only took up in one essay in later life (Blaug, 1985, pp. 563–4). The difficulty created by the use of fixed capital is simply that there is no obvious way of linking particular units of input embodied in fixed capital with particular units of finished output: all the inputs embodied in fixed equipment are jointly responsible for the whole stream of future outputs. Thus, by limiting itself to circulating capital, Austrian capital theory avoided such vexing questions as the optimum rate of depreciation and replacement of old equipment that are always linked with the decision to invest in new equipment, questions which perhaps are not completely resolved even to this day.

The increasing use of fixed capital is said to be one of the distinguishing characteristics of a capitalist system. If so, we might well expect capital theory to have been largely devoted to an analysis of fixed capital. It is one of the ironies of the history of economic thought, however, that capital theory from Turgot to the late Wicksell always treated circulating and not fixed capital as ‘capital’ *par excellence*.

Bibliography

Blaug, M. 1985. *Economic Theory in Retrospect*. 4th edn. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx After Sraffa*. London: New Left Books.

How to cite this article

Blaug, Mark. "circulating capital." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000143> doi:10.1057/9780230226203.0236

The New Palgrave Dictionary of Economics Online

city and economic development

J. Vernon Henderson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

As countries develop they urbanize, with resources shifting from labour-intensive agricultural production to manufacturing and services, which are located in cities because of agglomeration economies. This entry discusses the economic determinants of this process. But urbanization also moves populations from traditional rural environments with informal political and economic institutions to the relative anonymity and more formal institutions of urban settings. A major issue in the development process is development of institutions and national policies which allow cities to operate in markets that are well structured and conducive to good urban outcomes.

Keywords

city in economic development; congestion; core-periphery models; democracy; endogenous growth; human capital; Korea; land tax; regional models; rent seeking; rural-urban migration; spatial convergence; two-sector models; urban agglomeration; urbanization

Article

The city in economic development is fundamental to the urbanization process. Urbanization, or the shift of population from rural to urban environments, is a transitory process which is socially and culturally traumatic. As a country develops, it moves from labour-intensive agricultural production to labour being increasingly employed in industry and services. The latter are located in cities because of agglomeration economies. Thus, urbanization moves populations from traditional rural environments with informal political and economic institutions to the relative anonymity and more formal institutions of urban settings. That in itself requires institutional development within a country.
Once urbanization is complete, one might be tempted to simply move on to the traditional analysis of systems of cities, with the idea that the issues that face systems of cities in developed economies are the same as those that face cities in developing but fully urbanized economies (as in Latin America and the

Middle East). But in practice this is not the case; countries still face problems of developing institutions and national policies which allow cities to operate in markets that are well structured and conducive to good urban outcomes. Here, we discuss both the urbanization process and then the institutional-policy issues that face cities in developing countries.

The urbanization process

There are several models of the urbanization process. The traditional ones are two-sector models, where population moves from a rural sector to an all-purpose urban sector, due to exogenous factors such as unexplained shifts in technology (Lewis, 1954). Dual-sector models focus on the question of urban 'bias', or the effect of government policies on the urban–rural divide, and the efficient rural–urban allocation of population at a point in time. Generally, these models are static, and any urbanization is the result of exogenous forces – technological change favouring the urban sector or changes in the terms of trade favouring the urban sector. There is a new generation of two-sector models, namely, the core–periphery models, which have more of a spatial flavour (Krugman, 1991; Puga, 1999). Core–periphery models ask when in a two-region country industrialization, or 'urbanization', is spread over both regions rather than being concentrated in just one region. The models explore a key issue: the initial development of a core (say, coastal) region and a periphery (say, hinterland) region, as technology improves (transport costs fall) from a starting point with two identical regions. However core–periphery models have limited implications for urbanization per se. They are unidimensional in focus, asking what happens to core–periphery development as transport costs between regions decline; they are really regional models, with limited urban implications. Urban models are focused on the city formation process, where the urban sector is composed of numerous cities, endogenous in number and size. Efficient urbanization and growth require timely formation of cities and appropriate institutions. Henderson and Wang (2005) develop an endogenous growth model with accumulation of human capital, where there is a shift out of the rural sector into an urban sector as per capita human capital and income grow. The urban sector is composed of multiple cities which grow in size with knowledge accumulation and in numbers with national population growth and rural–urban migration. Urbanization occurs because demand for food products is postulated to be income inelastic, so as per capita incomes rise the relative demand for food products declines, while at the same time productivity in the rural sector is growing. That releases labour from the rural sector to migrate to the urban sector, where the relative national demand for urban products is rising overtime.

As the urban sector grows, new cities form in national land markets. Efficient city sizes are limited, reflecting a trade-off between marginal agglomeration economies as a city grows and steadily rising urban diseconomies in the form of commuting, congestion and other urban disamenities. Efficient city sizes are at or near the peak to each city's inverted-U shape relationship between real income per worker and city employment where, with economic growth, such peaks and efficient city sizes may be shifting out over time. With urbanization and national population growth, if existing cities are to stay near efficient sizes, new cities need to form in a timely fashion. That timely formation requires local governments to have the autonomy to tax land rents and exclude entrants through zoning provisions. Moreover, developers or local governments must have the autonomy to utilize land and undertake enormous urban infrastructure investments so as to form new large-scale settlements. Such institutions and market environments may not be in place or may be slow to develop, and national politics may

delay their evolution, especially in developing countries. These factors retard the timely formation of cities, forcing migrants into existing oversized cities. We discuss these issues below.

Empirics and policy issues

The policy and empirical literature on urbanization addresses three broad questions. These deal with the determinants of the rural–urban allocation of resources at any point in time, spatial convergence, and excessive urban concentration.

Rural – urban allocation of resources

Dual economy models in the traditional development literature ask whether market failures bias the allocation of resources between the urban and rural sectors or between bigger and smaller cities. Renaud (1981) makes the related point that it is not just market failures but explicit government policies that bias or influence urbanization through their effect on national sector composition. Policies affecting the terms of trade between agriculture and modern industry or between traditional small town industries (textiles, food processing) and high-tech large city industries affect the rural–urban or small–big city allocation of population. Such policies include import tariffs, price controls and product subsidies.

Spatial convergence

The issue of convergence across spatial units in a country was initially posed at the regional level. Williamson (1965) argued that national economic development is characterized by an initial phase of internal regional divergence of per capita incomes and the allocation of industrial resources, followed by a phase of later convergence. There is a related urban model of this divergence–convergence phenomenon, which looks at urban primacy and the quantity allocation of resources across cities. Following Ades and Glaeser (1995), conceptually the urban world is collapsed into two regions: the primate city versus the rest of the country, or at least the urban portion thereof. The question is: to what extent is urbanization concentrated in, or confined to, one (or a few) major metro areas, as opposed to being spread more evenly across a variety of cities? Primacy is commonly measured by the ratio of the population of the largest metro area to the entire urban population in the country. Ades and Glaeser (1995) and Davis and Henderson (2003) find that primacy first increases, peaks, and then declines with economic development, indicating a later spread of urban resources from the primate city to other cities over time.

As part of this spatial convergence process, Lee (1997) and Kolko (1999) explore the relationship between changes in urban concentration and industrial transformation for Korea since 1975 and for the USA since 1900. The idea is that manufacturing is first concentrated in primate cities at early stages of development, and then decentralizes to such an extent that at the other end of economic development it is relatively more concentrated in rural areas. Initial concentration fosters ‘incubation’ and adaptation of technologies from abroad in a concentrated urban environment. But once manufacturing has modernized with fairly standardized technologies, firms decentralize to hinterland locations where rent and wage

costs are cheaper. For example, in Korea Seoul's urban primacy peaked around 1970, when Seoul had a dominant share of national manufacturing. During the next ten or 15 years, manufacturing suburbanized from Seoul to nearby satellite cities, as well as to satellite cities surrounding the two other major metro areas, Pusan and Taegu. But then in the early 1980s manufacturing spread rapidly from the three major metro areas and their satellites to rural areas and other cities. The largest metro areas became business service-intensive, relying on economies of diversity in local business services, often purchased by headquarter units of firms as part of marketing, financing, and exporting activities for their goods produced by plants in hinterland locations. This spatial separation, with headquarters' activities of firms in large metro areas and production facilities in smaller specialized cities, is called 'functional specialization' by Duranton and Puga (2005).

Urban concentration

A third set of questions asks whether the degree of urban concentration in countries is too little or too much. Are there policies which bias development towards bigger, say, politically dominant coastal cities at the expense of smaller, say, hinterland cities? The basic idea is that the political system favours the national capital (or other seat of political elites such as São Paulo in Brazil). For example, direct restraints on trade for hinterland cities such as an inability to access capital markets or to get export or import licences favour firms in the national capital. Policymakers and bureaucrats may gain as shareholders in such firms, or they may gain rents from those seeking licences or other exemptions from trade restraints. Indirect trade protection for the primate city can also involve underinvestment in hinterland transport and communications infrastructure. Another strategy can be to retard development of institutions and national land markets that allow timely formation of large-scale, competitor hinterland cities. Whether as true beliefs or as a cover for rent-seeking behaviour, policymakers often articulate the view that large, favoured cities are more productive and thus should be the site for government-owned heavy industry (such as São Paulo or Beijing–Tianjin, historically). Unfortunately these heavy industries don't benefit sufficiently from the agglomeration economies in such large cities and can't afford their higher costs of land and labour, which is one reason why they lose money in such cities.

Favouritism of a primate city creates a non-level playing field in competition across cities. The favoured city draws in migrants and firms from hinterland areas, creating an extremely congested high-cost-of-living metro area. Local city planners can try to resist the migration response to primate city favouritism by, for example, refusing to provide legal housing development for immigrants or to provide basic public services in immigrant neighbourhoods. Hence squatter settlements, bustees, kampongs and so on may develop. But still, favoured cities tend to draw in enormous populations.

What is the econometric evidence indicating that politics plays a role in increasing sizes of primate cities? Based on cross-section analyses, Ades and Glaeser (1995) find that, if the primate city in a country is the national capital, it is 45 per cent larger. If the country is a dictatorship, or at the extreme of non-democracy, the primate city is 40–45 per cent larger. The idea is that representative democracy gives a political voice to hinterland regions, so limiting the ability of the capital city to favour itself; and fiscal decentralization helps level the playing field across cities, giving hinterland cities political autonomy to compete with the primate city. Davis and Henderson (2003) explore these ideas further, examining in a panel context the impact of democratization and fiscal decentralization upon primacy.

Examining democratization and fiscal decentralization together, they find moving from most to least democratic form of government reduces primacy by 8 per cent, and moving from most to least centralized government reduces primacy by 5 per cent. They also find transport infrastructure investment in hinterlands reduces primacy, a prediction of core–periphery models.

Given the urban primacy relationships, it is natural to ask whether urban concentration is important to growth. Is there an optimal degree of urban primacy with each level of development where significant deviations from this level detract from growth? Optimal primacy would involve a trade-off between the benefits of increasing primacy (enhanced local scale economies contributing to productivity growth) and the costs (more resources diverted away from productive and innovative activities to shoring up the quality of life in congested primate cities). Henderson (2003) examines this question with panel data methods and finds that there is an optimal degree of primacy at each level of development which maximizes national productivity growth. That optimal degree rises as country income declines: high relative agglomeration is important when countries have low knowledge accumulation, are importing technology, and have limited capital to invest in widespread hinterland development. There is an international tendency to excessive primacy, with effectively non-federated countries such as Argentina, Chile, Peru, Thailand, and Algeria having extremely high primacy.

While for countries where people are allowed to migrate freely across cities and from rural to urban areas the focus is on excessive urban concentration, in the former planned economy countries the concern goes the other way. Countries such as China have formal migration restrictions limiting the visas given to rural people to move to cities and limiting migrants' access to jobs, housing, medical care and schooling in destination cities to reduce the incentive to migrate. Other former planned economies primarily limited migration through restrictions on housing provision and land development in cities. Planned economies have much lower urban concentration than other large countries. The efficiency loss there derives from unexploited urban agglomeration economies.

See Also

- location theory
- spatial economics
- systems of cities
- urban agglomeration
- urbanization
- urban production externalities

Bibliography

Ades, A. and Glaeser, E. 1995. Trade and circuses: explaining urban giants. *Quarterly Journal of Economics* 110, 195–227.

Davis, J. and Henderson, J.V. 2003. Evidence on the political economy of the urbanization process. *Journal of Urban Economics* 53, 98–125.

Duranton, G. and Puga, D. 2005. From sectoral to functional urban specialization. *Journal of Urban Economics* 57, 343–70.

Henderson, J. 2003. The urbanization process and economic growth: the so-what question. *Journal of Economic Growth* 8, 47–71.

Henderson, J. and Wang, H. G. 2005. Urbanization and city growth. *Journal of Economic Geography* 5, 23–42.

Kolko, J. 1999. Can I get some service here? Information technology service industries, and the future of cities. Mimeo. Cambridge, MA: Harvard University.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99, 483–99.

Lee, T. C. 1997. Industry decentralization and regional specialization in Korean manufacturing. Ph.D. thesis. Providence, RI: Brown University.

Lewis, W. 1954. Economic development with unlimited supplies of labor. *Manchester School of Economic and Social Studies* 22, 139–91.

Puga, D. 1999. The rise and fall of regional inequalities. *European Economic Review* 43, 303–34.

Renaud, B. 1981. *National Urbanization Policy in Developing Countries*. New York: Oxford University Press.

Williamson, J. 1965. Regional inequality and the process of national development. *Economic Development and Cultural Change* 13(4), 3–45.

How to cite this article

Henderson, J. Vernon. "city and economic development." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000191> doi:10.1057/9780230226203.0237

The New Palgrave Dictionary of Economics Online

Clapham, John Harold (1873 – 1946)

Phyllis Deane

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Clapham, J. H.; economic development; economic history; Marshall, A.

Article

Sir John Clapham, who became in 1928 the first professor of economic history in the University of Cambridge, was born in Lancashire, the son of a prosperous jeweller. From the Cambridge boarding school (Leys) to which he was sent at the age of 14, he went up to King's College in 1892 to read history at a time when Acton, Maitland and Cunningham dominated the history school. It was as a graduate student at King's, researching into the French Revolution, that he attracted the attention of Alfred Marshall, who characteristically set about pressuring the promising young historian to devote his research efforts to filling the gaps in modern English economic history. There is an oft-quoted letter which Marshall wrote in 1897 to Acton saying:

> I feel that the absence of any tolerable account of the economic development of England during the last century and a half is a … grievous hindrance to the right understanding of the problems of our time … but till recently the man for the work had not yet appeared. But now I think the man is in sight. Clapham has more analytic faculty than any thorough historian whom I have ever taught: his future work is I think still uncertain: a little force would I think turn him this way or that. If you could turn him towards XVIII or XIX century economic history, economists would ever be grateful to you.

Unfortunately Marshall did not live to read Clapham's massive, three-volume *Economic History of Modern Britain*, the first volume of which appeared in 1926 (dedicated to Marshall and his old enemy William Cunningham), and the last in 1938. No doubt he approved of the scholarly monograph on *The Woollen and Worsted Industries* (1907), written when young Clapham was professor of economics at the University of Leeds – an appointment in which it is hard not to suspect that Marshall's influence was decisive. Nevertheless, when Clapham returned to a King's fellowship in 1908, he resumed his

researches in French political history and joined his fellow historians in criticizing the new Economics Tripos for being far too theoretical. It was not until after the First World War (during which he served in the Board of Trade and gained first-hand experience of the process of economic decision-making as a member of the Cabinet Committee on Priorities) that he in effect rejoined the path that Marshall had pointed out to him. His *Economic Development of France and Germany* (1921) was the first modern study in comparative economic development, but typically it involved juxtaposing his detailed analyses of two differing experiences of development, rather than relating them to a general theory of economic development, or even generalizing from these case histories.

The truth is that Clapham had no interest in theoretical economics except in so far as it supplied concepts and categories that would permit him to classify and analyse the empirical detail of economic history. He was repelled by the blatant unrealism of orthodox theorizing. His famous article ‘Of Empty Economic Boxes’, published in the September 1922 *Economic Journal*, accused the theorists of operating with concepts which were empty and irrelevant. ‘I think a great deal of harm has been done’, he complained, ‘through omission to make clear that the Laws of Return have never been attached to specific industries: that we do not, for instance, this moment *know* under what conditions of returns coals or boots are being produced’. But his complaints fell on deaf ears. The interwar theorists saw no point in relating the strategic concepts of their models to real-world constructs and were agreed that, as Keynes put in, Clapham was ‘barking up the wrong tree’.

What Clapham had learned from Marshall was that economics is the study of mutually interacting quantities and that it was the function of an economic historian to put the key quantitative questions to the historical record – for example, how large? how long? how often? how representative? – when spelling out the chains of cause and effect linking economic events. He made it his business to demolish, or qualify, facile generalizations that did not stand up to the available statistical evidence; for example, the Malthusian law of population, or the Marxian predictions of the pauperization of the masses. Though alive to the defects of historical statistics, he was bold enough to make the best of them, ‘to offer dimensions, in place of blurred masses of unspecified size’ and to analyse the bare aggregates into their strategic components. His training as a historian, however, kept a balance between quantitative and qualitative data, and his large-scale study of the economic development of modern Britain was diversified and illuminated by a continuous stream of vivid factual detail. His last book, *The Bank of England: A History, 1694–1914* (1944), commissioned by the Bank to commemorate its 250th anniversary, gave him access to the voluminous manuscript records of the first central bank. Writing its history and setting its operations and policies within its political and economic context was a task which by training and interests he was peculiarly well-equipped to perform. His intellectual energy seemed enhanced rather than diminished by his retirement from the Cambridge chair, and his sudden death in 1946 cut short a research programme which was still in full swing.

Selected works

1907. *The Woollen and Worsted Industries*. London: Methuen.

1921. *The Economic Development of France and Germany, 1815–1914*. Cambridge: Cambridge University Press.

1922. Of empty economic boxes. *Economic Journal* 32, 305–14.

1926–38. *An Economic History of Modern Britain*. 3 vols. Cambridge: Cambridge University Press. Vol. 1, *The Early Railway Age 1820–1850* (1926); vol. 2, *Free Trade and Steel* (1932); vol. 3, *Machines and National Rivalries (1887–1914) with an Epilogue (1914–1929)* (1938).

1944. *The Bank of England: A History, 1694–1944*. 2 vols. Cambridge: Cambridge University Press.

How to cite this article

Deane, Phyllis. "Clapham, John Harold (1873–1946)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000147> doi:10.1057/9780230226203.0238

The New Palgrave Dictionary of Economics Online

Clark, Colin Grant (1905 – 1989)

H.W. Arndt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand; agriculture and economic development; Clark, C. G.; economic growth; multiplier; national income; population growth

Article

Colin Clark, one of the most fertile minds in 20th-century applied economics, was born in London. After graduating in chemistry at Oxford University in 1924, he worked as assistant to W.H. Beveridge, Allyn Young and A.M. Carr-Saunders, stood unsuccessfully as a Labour candidate in the May 1929 general election, then joined the staff of the Economic Advisory Council, recently set up by Ramsay MacDonald, of which Keynes was a member. In 1931, rather than agree to write a protectionist manifesto for MacDonald, he accepted an appointment as lecturer in statistics at Cambridge, where he remained until, in 1937, he went to Melbourne University, initially as visiting lecturer. In Australia he occupied government posts, chiefly as economic adviser to the state government of Queensland, until 1952. After spells as visiting professor at the University of Chicago and as Director of the Oxford Institute of Agricultural Economics, he returned to Australia in 1968. He remained active as a research consultant at the University of Queensland.

In the first decade of an astonishingly prolific half-century of research and writing, Colin Clark established himself as one of the pioneers of national income estimates. He greatly improved existing estimates for the United Kingdom, and later for Australia and the Soviet Union, and in so doing made methodological contributions so fundamental that he has justly been described as co-author, with Simon Kuznets, of the ‘statistical revolution’ that accompanied the revolution in macroeconomics of the 1930s. He was the first to use the gross national product (GNP) and to present estimates in the framework of the main components of aggregate demand ($C+I+G$); he made some of the earliest estimates of Keynes's multiplier and, in an article published in 1937, one of the first international comparisons of the purchasing power of national currencies and thus of real national product. These were carried further in his monumental *Conditions of Economic Progress* (1940), which was important chiefly because it signalled the revival of interest among the profession in secular economic growth and development but

which also supplied the first substantial statistical evidence of the gulf in living standards between rich and poor countries (the ‘Gap’) and developed the thesis that, in the course of economic growth, the occupational structure shifts from primary to secondary and tertiary industries. During the Second World War, in *The Economics of 1960* (1942), Clark made one of the first ambitious attempts at a macroeconomic model of the world economy.

Recognized also as one of the ‘Pioneers in Development’, Colin Clark made significant contributions to empirical study of the relations between food supply and population growth, the economics of irrigation and subsistence agriculture, of determinants of economic growth and of productivity in agriculture in developing countries. At the same time, he was a gadfly in the political economy of developed countries, arguing against growthmanship, against high taxation and against welfarism long before it became fashionable to do so.

See Also

- Stone, John Richard Nicholas

Selected works

1932. *The National Income 1924–31*. London: Macmillan.

1937. *National Income and Outlay*. London: Macmillan.

1939. *A Critique of Russian Statistics*. London: Macmillan.

1940. *Conditions of Economic Progress*. London: Macmillan Revised edn, 1957.

1942. *The Economics of 1960*. London: Macmillan.

1961. *Growthmanship*. London: Institute of Economic Affairs.

1964(With M.R. Haswell.) *Economics of Subsistence Agriculture*. London: Macmillan.

1967. *Population Growth and Land Use*. London: Macmillan Revised edn, 1977.

1982. *Regional and Urban Location*. St Lucia: University of Queensland.

1984. Development economics: the early years. In *Pioneers in Development*, ed. G.M. Meier and D. Seers. New York: Oxford University Press.

How to cite this article

Arndt, H. W. "Clark, Colin Grant (1905–1989)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000148> doi:10.1057/9780230226203.0239

The New Palgrave Dictionary of Economics Online

Clark, John Bates (1847 – 1938)

Donald Dewey
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Adams, H. C.; American Economic Association; antitrust policy; arbitration; Böhm-Bawerk, E. von; capital accumulation; capital measurement; capital theory wages fund; Clark, J. B.; Clark, J. M.; Ely, R. T.; factor shares; George, H.; German Historical School; Jevons, W. S.; Johnson, A. S.; Knies, K. G. A.; marginal utility principle; natural and actual values; neoclassical theory of distribution; trade unions; rent; social Darwinism; Sraffa, P.; Sumner, W. G.; universal measure of value; Veblen, T.

Article

John Bates Clark, the first American economist to deserve and gain an international reputation, was born at Providence, Rhode Island, on 26 January 1847 into a modestly prosperous merchant family. His father's struggle with tuberculosis prompted a move to Minneapolis in search of a better climate and later required Clark to discontinue his studies at Amherst (he had transferred from Brown after two years) in order to run the family business. The business involved selling a line of ploughs to receptive but credit-needy country storekeepers throughout Minnesota. Following his father's death, the business was sold at a profit and Clark returned to Amherst, graduating with highest honours in 1872.

Clark's New England forebears had included many Congregational ministers and he seriously considered entering the Yale Divinity School. (He remained a communicant throughout his life and saw one son enter the ministry.) But encouraged by President Julius Seelye of Amherst, who had taught him political economy out of Amasa Walker's textbook, he chose instead the high-risk course of an academic career in a country still without universities. After Amherst, he went abroad, enrolling for two years at Heidelberg and six months at Zurich.

While Clark has left no detailed account of his European studies, his early work indicates that he was much influenced by the German Historical School, and especially by the lectures of Karl Knies. Whether the influence was for good or ill is not clear. It probably slowed his development as a theorist. (His formulation of the marginal utility principle was worked out before he had heard of Jevons.) But it also taught him that an economist needed a far more professional training than that provided by the thin textbook gruel offered in the American colleges of the day. Clark was one of three young

‘Germans’ (the other two being Richard Ely and Henry Carter Adams) who, at a meeting of the American Historical Society at Saratoga in 1885, issued the call that led to the formation of the American Economic Association. Their plainly avowed purpose was to encourage German-style empirical research and give a sympathetic hearing to the critics of laissez faire. The dogmatic social Darwinism of William Graham Sumner epitomized all that they disliked in American economics. Clark became the third president of the new group and his diplomacy and moderation are credited with making it more acceptable to the country's older economists, most of whom eventually joined (but not Sumner).

Shortly after going to his first professorship at Carleton College in Northfield, Minnesota, in 1876, Clark was incapacitated for two years by an illness that, according to his son, John Maurice, permanently lowered his energy level. Whatever its nature – the family memorial to Clark provides no details – the illness seems only to have strengthened his determination and powers of organization. Following his recovery, Clark worked steadily and with a notable economy of effort until shortly before his death at the age of 91. Most of his contributions to economic theory, however, were worked out in the first 15 years of his career though the most polished formulations did not come until *The Distribution of Wealth* (1899). Clark's need to choose his projects carefully may explain why, despite his admiration for the work of historians and institutionalists, he never tried to emulate them. All of his life Clark remained a theorist who often wrote on issues of the day.

Clark first gained recognition with a series of articles in *The New Englander* that, with revisions, were published in 1886 as *The Philosophy of Wealth*. Clark's admirers have found this first book something of an embarrassment, and not without reason. It is a young Victorian's book, full of grand historical generalizations and the elevated expressions of sentiment that have long been out of fashion. Still, on close reading, it reveals the qualities that were to make him a major figure in the history of economics – a superb command of language (Böhm-Bawerk, who debated capital theory with Clark, claimed that his literary elegance gave him an unfair advantage), a willingness to take a position on controversial issues, and, above all, a remarkable talent for economic theory.

The collection contains a totally original and quite sophisticated statement of the principle of marginal utility (‘effective utility’ in Clark's vocabulary), a reasoned rejection of Malthusian pessimism, and many perceptive comments on the rise of labour unions, cartels, and corporations. Even the main outlines of Clark's treatment of capital and interest are discernible in the *Philosophy*.

Clark's intellectual distinction was fully revealed two years later with the publication of his monograph, *Capital and its Earnings* (1888a) which has a good claim to stand as the foundation stone of modern capital theory. While the distinction between labour and capital is still accepted (though even here Clark wavers), all other things including land that directly or indirectly enter into the production of consumer goods are treated as capital. The existence of interest is firmly placed in the productivity of capital. The creation of income as a concomitant of the destruction of individual capital goods is emphasized. The irrelevance of the ‘period of production’ of individual capital goods to anything of importance is shown and the fallacy underlying the wages fund doctrine is exposed.

Clark has been criticized for introducing the ‘neoclassical fairy tale’ into capital theory – the notion that capital is some strange substance that, ‘transmutes itself from one machine form into another like a restless reincarnating soul’ (Samuelson, 1962). While the neoclassical fairy tale has its limitations as a construct for understanding capital accumulation in the real world, Samuelson's jibe is off target. Clark's view of the production process is perfectly correct. Machines do ‘transmute’ themselves into other machines in the course of wearing out.

A more serious challenge to capital theory in the Clark tradition goes back to Böhm-Bawerk. If there is such a thing as a quantity of capital ‘embodied’ at any given moment in a set of heterogeneous specialized capital goods, what is its unit of measure? Unlike Irving Fisher, Clark faced the question squarely and attempted an answer. Unfortunately, the effort led him to bring forth his ‘universal measure of value’ – the product of a strange and nearly unintelligible fusion of utility analysis and the labour theory of value. While Clark was inordinately proud of his measure (and credited its inspiration to some lectures of Knies) it quickly found a merciful oblivion.

Later writers in the Clark tradition – or, at any rate, those who have felt the need for an impeccably consistent set of assumptions – have curbed their ambitions and been content to solve (or evade) the measurement problem by positing a surrogate production function where all capital goods are moulded from some homogeneous putty-like substance. The limit case in the Clark tradition is the ‘Crusonia plant’ named by Frank Knight but first suggested by W.S. Jevons's ‘whole produce’. It supplies all human wants and, in the absence of consumption, grows at a constant geometric rate. Here the quantity of capital can be found either by measuring Crusonia directly or by dividing the plant's yield (income) in perpetuity by its natural growth rate, that is, the marginal (and average) productivity of investment.

Whether one prefers capital theory in Clark's tradition to its principal rival – capital theory in the Sraffa tradition – is ultimately a matter of personal taste. Both employ simplifications that take one far from reality. However, notwithstanding the measurement conundrum, to date capital theory in the Clark tradition has provided the basis for virtually all empirical work on wealth and income. This is not surprising. To statisticians, measuring changes in the quantity of capital (which they rename the real value of the stock of capital assets) is just another index number problem.

Very early in his career Clark began to work on the problem of factor shares (possibly because of his interest in Henry George) and concluded that the treatment of land rent as a surplus whose size is not determined by marginal productivity was gross error. The most complete statement of his views on distribution is in *The Distribution of Wealth* (1899) which drew heavily on his earlier articles and monographs. Despite its flaws (which include the universal measure of value) the *Distribution* is a remarkable book and, by any reasonable test, a landmark treatise in the development of economics.

The *Distribution* represents an advance on the prior art in two important respects. It offers a discussion of the relation of statics to dynamics – the terms were introduced into economics by Clark – superior to that of previous treatments. And it offers, for the first time, a complete and lucid exposition of the neoclassical theory of distribution. The *Distribution* also brought Clark's views on capital to a much wider audience.

Clark was as conscious of the rapid pace of economic change as any German or American institutionalist of his day, but he stressed that, at any given moment, there are ‘natural’ values in the marketplace and permanent pressures pushing actual values toward them.

> Reduce society to a stationary state, let industry go on with entire freedom, make labor and capital absolutely mobile – as free to move from employment to employment as they are supposed to be in the theoretical world that figures in Ricardo's studies – and you will have a regime of natural values. These are the values about which rates are forever fluctuating in the shops of commercial cities. You will also have a regime of natural wages and interest; and these are the standards about which the rates of pay for labor and capital are always hovering in actual mills, fields, mines, etc.

Only by a careful separation and delineation of static and dynamic forces, Clark believed, can the process of price formation in real-world markets be understood. His methodology is not as formal and austere as F.H. Knight's in *Risk, Uncertainty, and Profit* (1921), but it is essentially the same. (In the version of Knight's doctoral dissertation accepted at Cornell in 1916 his intellectual debts to Clark are gratefully and fully acknowledged; for reasons unknown, almost all of the favourable references to Clark are omitted in the rewritten version published five years later.)

To demonstrate that, in the static state, payments to the factors exhaust the product when each receives its marginal product, Clark devised a set of diagrams to show that, in a two-factor model, what is viewed as rent and what is viewed as a factor payment is a matter of perspective. One becomes the other by interchanging the fixed and variable factors in the diagrams. Clark's treatment of rent has been followed by an admiring Paul Samuelson in all of the many editions of his *Economics*.

Clark's approach to distribution is set forth in ‘words and pictures’ (his mathematical training did not include calculus) and so lacks the precision of the versions of Wicksell and Wicksteed. But, being more accessible to student readers, it was Clark's treatment that first gained widespread attention for the neoclassical theory of distribution.

Clark has often been reproved for implying both that factor payments ought to be according to marginal productivity and that in a real-world market economy most factor payments do closely approximate marginal productivity (see, for example, Stigler, 1941). A reading of the *Distribution* without reference to Clark's other writings would indicate that he did hold these views. Certainly his advocacy of compulsory arbitration to end long labour disputes assumed that economic justice consisted in giving striking workers the wages prevailing in comparable employments elsewhere. However, a brilliant essay, ‘The Theory of Economic Progress’ (1896), leaves no doubt that he placed a far higher value on economic growth than on short-run justice or efficiency.

Well before Schumpeter, Clark wrote:

> The picture of a stationary state presented by John Stuart Mill as the goal of competitive industry is the one thing needed to complete the impression of dismalness made by the political economy of the early period. A state could not be so good that that lack of progress would not blight it; nor could it be so bad that the fact of progress would not redeem it. … The decisive test of an economic system is the rate and direction of movement.

Clark was a leading participant in the trust controversy that occupied American politics in the 30 years before the First World War. His moral seriousness and literary ability (and, one suspects, his ability to meet deadlines) made him a favourite of magazine editors – he once described himself as ‘writing my trust article again’. Like all economists of that era he had to think through his attitude toward the many large firms with large market shares that had so suddenly appeared.

As recorded in the *Philosophy of Wealth*, Clark's first reaction to the American business scene on returning from Germany was one of fascinated revulsion joined to an expression of hope that businessmen could be led to behave in more acceptable ways by pressures from labour unions, Church, and State. As the years passed, his views of commerce became much more favourable and his policy recommendations more worldly and specific. He early pointed out that the conduct of most so-called

trusts was influenced by the fear of entry and he never depreciated the efficiency gains made possible by large-scale production. At first he urged only a modest amount of government intervention as in, *The Control of Trusts: An Argument in Favor of Curbing the Power of Monopoly by a Natural Method* (1901). Clark's ‘natural method’ was little more than the competition of the marketplace purged of its ‘destructive’ ingredients plus government regulation of railroad rates to prevent unjustified differentials. A much expanded version of *The Control of Trusts*, with John Maurice Clark, his son, as co-author and the subtitle omitted, appeared in 1912. The revisions were mostly the work of the son and contain a virtual blueprint for an antitrust policy. The Clayton and Federal Trade Commission Acts of 1914 which followed shortly received their enthusiastic approval.

By his writing Clark did more than any other economist to confer intellectual respectability on an antitrust policy that had had its origins in the populist discontent that produced the Sherman Act. In retrospect, this may seem to have been a dubious achievement. But in Clark's favour it can be said that he was dealing with new and difficult issues and approached them with more objectivity than most of his contemporaries, for example, W.Z. Ripley and F.A. Fetter.

Clark's life as a teacher was at Carleton, Smith, Amherst, and from 1895 to 1923 at Columbia. At Carleton his kindness helped Thorstein Veblen (a thoroughly unpopular undergraduate in that church college) to find his way. At Columbia it helped Alvin Johnson to gain the income needed to complete his doctoral programme. His encouragement led F.H. Giddings to leave provincial journalism for a seminal career in sociology. He was, of course, the omnipresent influence in the life of John Maurice Clark, who succeeded to his chair at Columbia. Still, Clark's direct influence through the classroom seems to have been surprisingly limited. His quiet and self-sufficient personality did not require disciples and his probing but loosely organized lectures appealed only to very able students. Then too, Clark was a theorist in an era when, in the United States, institutional economics, not theory, was the height of academic fashion.

From 1911 onward Clark's great concern became the contribution that social scientists could make to ending war. When the Carnegie Endowment for International Peace was formed in 1910, he became the first director of its economics and history section serving until 1923. There he took the initiative in obtaining support for the studies that became the *Social and Economic History of the World War*. The general editor was his friend and Columbia colleague in history, James T. Shotwell. The Carnegie *History* ultimately ran to over a hundred volumes and still stands as the most ambitious research project in the social sciences ever undertaken by a private foundation. Unfortunately, its initial promise was never realized. Shotwell sought to organize the Carnegie *History* on the strange principle that an accounting of the great war was too important to be left to historians. As a result, while the series contains a few memorable studies, for example, J.M. Clark, *The Costs of the World War to the American People* (1931), it served mainly to preserve the recollections of wartime ministers and civil servants that would otherwise have been lost. J.M. Keynes disdainfully withdrew from the *History* in the planning stage.

Clark's work for peace continued to the end of his life. His last small book was a moving plea for collective action to deter aggression, *A Tender of Peace: The Terms on Which Civilized Nations Can, if They Will, Avoid Warfare* (1935). Clark died in New York City on 21 March 1938.

An abundance of honours came to him in his lifetime both in the United States and abroad. They were all deserved.

See Also

- Clark, John Maurice
- Fisher, Irving
- marginal productivity theory
- ‘neoclassical’

Selected works

1886. *The Philosophy of Wealth: Economic Principles Newly Formulated*. Boston: Ginn & Co.

1888a. *Capital and Its Earnings*. Baltimore: American Economic Association.

1888b(With F.H. Giddings.) *The Modern Distributive Process*. Boston: Ginn & Co.

1893. The ultimate standard of value. *The Yale Review* 1, February–May, 252–74.

1896. The theory of economic progress. *American Economic Association: Economic Studies* 1, April, 1–22.

1899. *The Distribution of Wealth: A Theory of Wages, Interest and Profits*. New York: The Macmillan Co.

1901. *The Control of Trusts: An Argument in Favor of Curbing the Power of Monopoly by a Natural Method*. New York: Macmillan.

1904. *The Problem of Monopoly: A Study of a Grave Danger and of the Natural Mode of Averting It*. New York: Columbia University Press.

1907. *The Essentials of Economic Theory: As Applied to Modern Problems of Industry and Public Policy*. New York: Macmillan.

1912(With J.M. Clark.) *The Control of Trusts*. New York: Macmillan.

1914. *Social Justice without Socialism*. Boston: Houghton Mifflin.

1935. *A Tender of Peace: The Terms on which Civilized Nations Can, If They Will, Avoid Warfare*. New York: Columbia University Press.

A nearly complete listing of Clark's publications is in A Bibliography of the Faculty of Political Science,

Columbia University, 1880–1930, New York: Columbia University Press, 1931; also in Economic Essays Contributed in Honor of John Bates Clark, ed. J.H. Hollander, New York: Macmillan, 1927.

Bibliography

Böhm-Bawerk, E. 1906. Capital and interest once more. *Quarterly Journal of Economics* 21, 1–21, 247–82.

Clark, J.M. 1931. *The Costs of the World War to the American People*. New Haven, CT: Yale University Press.

John Bates Clark. 1938. A memorial volume prepared by his children. New York (privately printed).

Knight, F.H. 1921. *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin.

Samuelson, P. 1962. Parable and realism in capital theory: the surrogate production function. *Review of Economic Studies* 29, 193–206.

Stigler, G. 1941. *Production and Distribution Theories: The Formative Period*. New York: Macmillan.

How to cite this article

Dewey, Donald. "Clark, John Bates (1847–1938)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000149> doi:10.1057/9780230226203.0240

The New Palgrave Dictionary of Economics Online

Clark, John Maurice (1884 – 1963)

Warren J. Samuels

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; business cycles; Clark, J. B.; Clark, J. M.; Commons, J. R.; Dewey, J.; institutional economics; overhead costs; social economics; Veblen, T.; workable competition

Article

Clark was born on 30 November 1884 in Northampton, Massachusetts, and died on 27 June 1963 in Westport, Connecticut. Educated at Amherst College and Columbia University (Ph.D., 1910), he taught at Colorado College (1908–10), Amherst (1910–15), University of Chicago (1915–26) and Columbia University (1926–52), where he succeeded his father, John Bates Clark. He was president of the American Economic Association in 1935 and received its Francis A. Walker Medal in 1952. His dissertation, ‘Standards of Reasonableness in Local Freight Discrimination’, was written under the supervision of his father. He was associated with the National Bureau of Economic Research, the National Resources Planning Board, the Twentieth Century Fund, the Attorney General's National Committee to Study the Anti-Trust Laws, and other organizations.

Clark worked within both orthodox and heterodox economics, making important contributions to microeconomics, macroeconomics and institutional, or social, economics. Eclectic and open-minded, he was critical of the apologetic uses of economic theory, particularly of the drawing of narrow and misleading welfare implications. He emphasized the limits of economics as a science.

Clark's contributions within conventional theory dealt principally with economic dynamics. He developed and stressed the implications of overhead, fixed costs in capital intensive industry for competitive structure, business pricing policy, and economic stability. He was the principal of several discoverers of the acceleration principle, with its important implications for instability. His career-long concern with competitive structure and behaviour led to his formulation of the concept of ‘workable competition’, with a stress on potential competition and intercommodity substitution. The major result of his equally long work in macroeconomics was an exploration of the strategic factors in business cycles which effectively summarized, in a general theoretical context, the state of empirical knowledge at the time. He also wrote extensively on railroad and public utility rates, basing-point pricing, economic

planning, the economics of war and of peacetime conversion, wage-price (cost-push inflation) theory and policy, and related topics.

Clark departed from the conventional mainstream in his social economics, which was akin to the institutional economics of John R. Commons and Wesley C. Mitchell and which reflected the influence of Thorstein Veblen and John Dewey. Clark's work on the social control of business and the theory of regulation explored the fundamental legal-economic nexus of society in a non-ideological manner stressing the substance and inexorable presence of formal (legal) and informal controls in an economic system, even in a pluralistic and voluntaristic economy, controls typically obscured in conventional analysis of markets. Law was important to the structure of freedom, not something solely antagonistic to freedom. His work in welfare economics emphasized the role of institutions, the necessity of psychological realism, and the inexorable role of moral or ethical values. His concern with the costs of labour that are registered in neither the market nor by industry presaged later institutional work on externalities and social costs.

Selected works

1912. (With J.B. Clark.) *The Control of Trusts*. New York: Macmillan.

1923. *Studies in the Economics of Overhead Costs*. Chicago: University of Chicago Press.

1926. *Social Control of Business*. New York: McGraw-Hill.

1934. *Strategic Factors in Business Cycles*. New York: H. Wolff for the National Bureau of Economic Research.

1936. *Preface to Social Economics*. New York: Farrar and Rinehart.

1948. *Alternative to Serfdom*. New York: Knopf.

1949. *Guideposts in Time of Change*. New York: Harper.

1957. *Economic Institutions and Human Welfare*. New York: Knopf.

1961. *Competition as a Dynamic Process*. Washington, DC: Brookings Institution.

How to cite this article

Samuels, Warren J. "Clark, John Maurice (1884–1963)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000150> doi:10.1057/9780230226203.0241

The New Palgrave Dictionary of Economics Online

class

David B. Grusky

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The structure of inequality has historically been represented with an income paradigm that treats well-being as adequately indexed by income alone. By contrast, the class-analytic tradition treats inequality as fundamentally multidimensional, with such variables as health, education and social relations all deemed important non-income constituents of well-being. These variables may assume a class-based form in which social groups within the division of labour define characteristic constellations of scores. The class model is further supported in so far as class membership has true causal effects on behaviours that are not reducible to the effects of income or other correlates of class.

Keywords

class; compensating differentials; Human Development Index; identity; inequality, multidimensional; labour markets; Marx, K.; redistribution of income and wealth; Sen, A.; socio-economic index; underclass

Article

The labour market of contemporary societies is rife with various types of ‘classes’ that impede the free flow of labour by restricting entry to those who have the requisite degrees, certificates, memberships or capital. These classes take the form, for example, of occupations (such as economist, carpenter), aggregates of occupations (such as manager, farmer), or groups that represent competing factors of production (such as worker, capitalist). Although such classes are ubiquitous in contemporary labour markets, their effects on labour market processes are not always incorporated into formal economic models. The main type of class to which attention has historically been paid is that of industry. The bifurcation of labour markets into industry classes, while clearly a relevant and well-developed topic in the literature, is not covered here. For purely historical reasons, the term ‘class’ has been reserved for non-industrial forms of bifurcation, a usage that is adopted in the following discussion as well.

The descriptive rationale for a class model is usefully introduced in the context of a multidimensional representation of inequality. This representation, which is presented below, makes it possible to motivate the class concept, to consider how classes may be empirically revealed, and to assess whether the class concept is needed to represent the structure of labour markets.

The clustering rationale

It has become increasingly fashionable to claim that inequality is multidimensional, that income inequality is accordingly only one of many important forms of inequality, and that income redistribution in and of itself would not eliminate inequality (see, for example, Sen, 2006). If this line of argument is taken seriously, an obvious prescription is to examine separately each of the many variables that constitute the multidimensional space of interest. For example, one might usefully distinguish between the eight forms of inequality listed in Table 1, each such form pertaining to a type of good that is intrinsically valuable (as well as possibly an investment). The multidimensional space formed by these variables may be labelled the ‘inequality space’. The social location of an individual within this inequality space can then be characterized by specifying her or his constellation of scores on each of the eight types of variables in this table.

Types of valued goods and examples of advantaged and disadvantaged groups

Valued goods		Examples	
Type	Example	Advantaged	Disadvantaged
1. Economic	Wealth	Billionaire	Bankrupt worker
	Income	Professional	Laborer
	Ownership	Capitalist	Employee
2. Power	Political power	Prime minister	Disenfranchised person
	Workplace authority	Manager	Subordinate worker
	Household authority	‘Head of household’	Child
3. Cultural	Knowledge	Intelligentsia	Uneducated
	Popular culture	Movie star	High-culture ‘elitist’
	‘Good’ manners	Aristocracy	Commoner
4. Social	Social clubs	Country-club member	Non-member
	Workplace associations	Union member	Non-member
	Informal networks	Washington ‘A list’	Social unknown
5. Honorific	Occupational	Judge	Garbage collector
	Religious	Saint	Excommunicate
	Merit-based	Nobel Prize winner	Non-winner
6. Civil	Right to work	Citizen	Illegal immigrant
	Due process	Citizen	Suspected terrorist

	Franchise	Citizen	Felon
7. Human	On-the-job	Experienced worker	Inexperienced worker
	General schooling	College graduate	High-school dropout
	Vocational training	Law-school graduate	Unskilled worker
8. Physical	Mortality	Person with long life	A ‘premature’ death
	Physical disease	Healthy person	Person with AIDS, asthma
	Mental health	Healthy person	Depressed, alienated

At least implicitly, scholars of inequality long ago adopted precisely such a multidimensionalist approach, as revealed by the burgeoning research literatures that monitor not just income inequality but also inequality of health, social networks, education, computer usage and all manner of other valued goods. This line of research typically takes the form of an exposé of the extent to which seemingly basic human ‘entitlements’, such as living outside of prison, being gainfully employed, freely participating in digital culture, or living a reasonably long and healthy life, are unequally distributed in ways that may amplify or somehow complement well-known differentials of income or earnings.

Does the inequality space take on a simpler form than might be implied by the convention of analysing each of these variables separately and independently? Two possible simplifications may be considered here. First, scholars have frequently combined scores on the underlying variables to form indices, with sociologists often combining education and income into a socio-economic index (for example, Hauser and Warren, 2001) and development economists often combining measures of health, income, education and literacy into a ‘Human Development’ index (for example, UNDP, 2005). There is, however, growing concern that such standard multidimensional scales are excessively abstract and fail to capture the social organization of inequality, especially the emergence of social networks, norms, and adaptive preferences or tastes among individuals in similar life situations and circumstances. The socio-economic scale, for example, is a purely statistical tool that groups together individuals of similar income or education levels without any consideration of whether these individuals associate with one another or are co-members of some real group, such as a union or occupation.

This critique motivates a second, class-based approach to understanding the structure of the inequality space. The class model is defensible insofar as (*a*) individuals tend to cluster into a relatively small number of characteristic combinations or packages of scores on the underlying variables, and (*b*) the clusters are defined by such structural locations as detailed occupations (doctor, secretary, plumber), aggregates of detailed occupations (professional, manager, clerk, craft worker, labourer, farmer), or other types of ‘big classes’ (for example, capitalist, worker). These clusters generate a labour market that, instead of being a seamless distribution of incomes, is a lumpy entity with deeply institutionalized groups that constitute pre-packed combinations of valued goods.

The class of craft workers, for example, has historically comprised individuals with moderate educational investments (secondary-school credentials), considerable occupation-specific investments in human capital (vocational or on-the-job training), average income, relatively high job security, middling social honour and prestige, quite limited authority and autonomy, and comparatively good health outcomes (by virtue of union-sponsored health benefits and regulation of working conditions). By contrast, the underclass is characterized by a rather different package of scores, one that combines minimal educational investments, limited opportunities for on-the-job training, intermittent labour force

participation, low income, virtually no opportunities for authority or autonomy on the job (during brief bouts of employment), relatively poor health (by virtue of lifestyle choices and inadequate health care), and much social denigration and exclusion. The other classes appearing in class schemes (such as professionals, managers, clerks, labourers, farmers) may likewise be understood as particular combinations of scores on the variables of interest.

In a class-based society, the inequality space will accordingly have relatively low dimensionality, a dimensionality no more or less than the number of classes. This understanding of the class principle implies that the variables constituting the inequality space must be independent of one another *within* each class. If the independence assumption begins to break down within a postulated class, we can then speak of ‘subclasses’ forming by virtue of developing their own distinguishable packages of scores. It is useful in this context to distinguish between a big-class regime in which the dimensionality of the inequality space is small and a micro-class regime in which the dimensionality of the inequality space is large. Although Marx (1894) argued that the inequality space in the early industrial period was becoming increasingly consistent with a two-class solution (in which privileged capitalists were juxtaposed to disadvantaged workers), some contemporary class analysts emphasize, to the contrary, that the forces of market differentiation have generated a micro-class regime in which the independence assumption holds not at the big-class level but only within quite detailed occupations (for example, Weeden and Grusky, 2005). There is much ongoing debate among inequality scholars on the dimensionality of the contemporary inequality space and, in particular, on whether the dimensionality of that space has been increasing or diminishing.

The foregoing implies that one may usefully distinguish between big-class regimes with few classes and micro-class regimes with many classes. Additionally, one might distinguish inequality regimes not on the basis of how many classes there are but on the basis of how the classes differ from one another. In a purely ‘vertical’ class system, one can readily order classes on a single scale from ‘low’ to ‘high’, with low classes being systematically disadvantaged on all variables and high classes being systematically advantaged on all variables. This organization of the inequality space implies a stark form of inequality in which privilege on one dimension implies very reliably privilege on another. Alternatively, a class system that is (partly) horizontal will embody compensating forms of advantage and disadvantage, meaning that at least some classes are formed by combining high values on one dimension with low values on another. There is, again, much debate among class analysts as to whether the inequality space is becoming more or less vertically organized.

It is of course possible that the inequality space is organized in ways that are largely inconsistent with the class principle. Two types of non-class solutions, as reviewed below, may be usefully distinguished.

Extreme disorganization

First, one can imagine an inequality space in which the underlying variables don't covary at all, hence yielding a one-class solution or, equivalently, a non-class regime. To be sure, there would be much inequality under this hypothetical constellation of data, yet it would take a uniquely structureless form in which the independence assumption holds throughout the inequality space, not just within a given class. It is unlikely that such extreme disorganization would ever be realized, but some postmodernists (for example, Pakulski, 2005) have argued that we are moving gradually toward this form. If they are correct, it means that the growth in income inequality is at least counterbalanced by a decline in the

association between income and other valued goods. As with the horizontal class regime described above, here again we have a form of inequality that embodies much in the way of compensating differentials, although such differentials are not in this case packaged into institutionalized classes.

Individuals as classes

The second main type of non-class solution arises when the variables constituting the inequality space are related to one another in perfectly linear fashion. When the data are configured in this way, it is no longer possible to identify a set of classes within which independence holds, as the underlying inequality variables continue to covary with one another no matter how much one disaggregates. We are left with an extreme micro-class solution in which the data thin out to the point where each individual becomes a class unto himself or herself. This solution is consistent, for example, with the claim that income is a master variable, that it perfectly signals all other individual-level measures of inequality, and that no higher-level class organization therefore appears. Obviously, this ideal type would never be empirically realized in such extreme form, but it is nonetheless important to ask whether it comes closer to being realized in some societies or time periods than in others.

The ‘class effect’ rationale

We have to this point represented the class principle as a hypothesis about the clustering of observations in the inequality space. As an alternative motivation for the class hypothesis, it is sometimes claimed that classes are social contexts that affect attitudes, behaviour, and individual action of many kinds. When this motivation is adopted, classes are not typically construed as information-rich social containers that capture many life conditions of interest, but rather as analytic categories that single out a particular social context that is presumed to be very consequential in defining interests. Under such a formulation, a class analyst will therefore typically nominate a single variable (for example, authority, ownership) as especially useful in understanding the sources of social behaviour, with the class categories then defined so as to capture differences across workers on that underlying variable of presumed consequence. The Marxian model, for example, famously embodies the claim that classes are best defined in terms of employment status alone, with the rationale for this definition being that employment status putatively defines interests and hence attitudes and behaviour (Marx, 1894). In contemporary labour markets, the class of employed workers is of course very heterogeneous, thus motivating class analysts to introduce further distinctions within that class that are presumed to be consequential in defining interests and action. There is no shortage of such elaborated class models (Wright, 2005).

When a class model is motivated by presumed class effects, it is important to establish that such effects are indeed truly causal. If, for example, one finds that seeming differences in the politics of professionals, managers, craft workers and other social classes disappear when income is controlled, then presumably one can refer only to an income effect on politics, not a true class effect. Why might net effects of class be detected even with rigorous controls? In addressing this question, what must first be stressed is that, even when classes are *defined* in terms of a single analytic variable, the resulting classes are nonetheless often organic packages of conditions; and the constituents of these packages may

combine and interact in ways that lead to an emergent logic of the situation. The underclass, for instance, may be understood as a combination of negative conditions (intermittent labour force participation, limited education, low income) that, taken together, engender a sense of futility, despondency, or learned helplessness that is more profound than what would be expected from a model that simply allows for independent effects of each constituent class condition. To be sure, a committed reductionist might counter that, when modelling behaviour, one merely needs to include the appropriate set of interactions between the constituent variables. In so far as classes define the relevant packages of interacting conditions, such an approach just becomes an unduly complicated way of sidestepping the reality of classes.

This emergent logic of the situation may well be undergirded by a class culture. At one extreme, class cultures may be understood as nothing more than ‘rules of thumb’ that encode optimizing behavioural responses to prevailing environmental conditions, rules that allow class members to forgo optimizing calculations themselves and rely instead on cultural prescriptions that provide reliable short cuts to the right decision. In this vein, Goldthorpe (2000) argues that working-class culture is disparaging of educational investments not because of some maladaptive oppositional culture but because such investments expose the working class, more so than other classes, to a real risk of downward mobility. Typically, working-class children lack insurance in the form of substantial family income or wealth, meaning that they cannot easily recover from an educational investment gone awry (in the form of dropping out); and those who nonetheless undertake such an investment therefore face the real possibility of substantial downward mobility. The emergence, then, of a working-class culture that regards educational investments as frivolous may be understood as encoding that conclusion and thus allowing working-class children to undertake optimizing behaviours without explicitly engaging in decision-tree calculations. The behaviours that a rule-of-thumb culture encourages are, then, deeply adaptive because they take into account the endowments and institutional realities that class situations encompass.

The foregoing example may be understood as one in which a class-specific culture instructs recipients about the best means for achieving ends that are widely pursued by *all* classes. Indeed, the prior rule-of-thumb account assumes that members of the working class share the conventional interest in maximizing labour market outcomes, with their class-specific culture merely instructing them about the approach that is best pursued in achieving that conventional objective. At the other extreme, one finds class-analytic formulations that represent class cultures as more overarching world views, ones that instruct not merely about the proper means to achieve ends but additionally about the proper valuation of the ends themselves. For example, some class cultures (such as aristocratic ones) place an especially high valuation on leisure, with market work disparaged as ‘common’ or ‘polluting’. This orientation presumably translates into a high reservation wage within the aristocratic class. Similarly, oppositional cultures within the underclass may be understood as world views that place an especially high valuation on preserving respect and dignity for class members, with of course the further prescription that these ends are best achieved by (*a*) withdrawing from and opposing conventional aspirations, (*b*) representing conventional mobility mechanisms (for example, higher education) as tailor-made for the middle class and, by contrast, unworkable for the underclass, and (*c*) pursuing dignity and respect through other means, most notably total withdrawal from and disparagement of mainstream pursuits. This is a culture, then, that gives respect and dignity an especially prominent place in the utility function and that further specifies how respect and dignity might be achieved.

Whatever the mechanism that underlies class cultures and class effects, the common assumption is that classes are meaningful social contexts, just as neighbourhoods are likewise understood within the 'neighbourhood effects' literature as meaningful social contexts. These contexts are expected in both cases to have causal effects that are not reducible to mere selective processes. Again, we have to stress that such a 'class effects' rationale for class models is best treated as a hypothesis, as there is little in the way of substantiating evidence at this point (cf. Weeden and Grusky, 2005).

It is altogether possible that such class effects are weak or at least weakening. The relevant postmodernist position in this regard is that social class has lost much of the power it once had because (*a*) other cross-cutting social cleavages (such as race or gender) have squeezed out class-based identities and interests, (*b*) identity formation in the postmodern world is so atomized and individualized that all structural bases of social behaviour have become less relevant, (*c*) the institutions that once represented class interests (for example, political parties, unions) have developed into new forms that are less class-based, or (*d*) the forces of the market work to gradually eliminate pockets of rent-generating social action. Regardless of the particular form of the argument, the expectation in all cases is that emergent effects of classes have, during the last several decades, become less prominent.

Conclusions

It should by now be clear that sociologists operating within the class-analytic tradition have adopted very strong assumptions about how inequality and poverty are structured. As was noted, the class concept may be motivated in two main ways, by claiming either that the inequality space has a (low) dimensionality equalling the number of social classes, or that the class locations of individuals have a true causal effect on behaviours or attitudes of interest. The foregoing claims have been unstated articles of faith among class analysts in particular and sociologists more generally. In this sense, class analysts have behaved rather like stereotypical economists, the latter frequently being criticized (and parodied) for their willingness to assume almost anything provided that it leads to an elegant model.

This critique of class analysis is, however, increasingly less justifiable. Indeed, the class-analytic status quo has come under much criticism of late, with many scholars now feeling sufficiently emboldened to argue that the concept of class should be abandoned altogether (for example, Kingston, 2000; Pakulski, 2005). Although the resulting debate has sometimes been unproductive, it has clearly precipitated an increasing interest in assessing the empirical foundations of class models.

See Also

- economic sociology
- inequality (global)
- inequality (measurement)
- labour economics
- labour market institutions
- poverty
- social status, economics and

Bibliography

Goldthorpe, J. 2000. *On Sociology: Numbers, Narrative, and the Integration of Research and Theory*. New York: Oxford University Press.

Hauser, R. and Warren, J. 2001. Socioeconomic indexes of occupational status: a review, update, and critique. In *Social Stratification: Class, Race, and Gender in Sociological Perspective*, ed. D. Grusky. Boulder: Westview Press.

Kingston, P. 2000. *The Classless Society*. Stanford: Stanford University Press.

Marx, K. 1894. *Capital*, 3 vols. London: Lawrence and Wishart, 1972.

Pakulski, J. 2005. Foundations of a post-class analysis. In *Approaches to Class Analysis*, ed. E. Wright. Cambridge: Cambridge University Press.

Sen, A. 2006. Conceptualizing and measuring poverty. In *Inequality and Poverty*, ed. D. Grusky and R. Kanbur. Stanford: Stanford University Press.

UNDP (United Nations Development Programme). 2005. *Human Development Report 2005*. New York: UNDP.

Weeden, K. and Grusky, D. 2005. The case for a new class map. *American Journal of Sociology* 111, 141–212.

Wright, E., ed. 2005. *Approaches to Class Analysis*. Cambridge: Cambridge University Press.

How to cite this article

Grusky, David B. "class." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000151> doi:10.1057/9780230226203.0242

The New Palgrave Dictionary of Economics Online

classical distribution theories

Massimo Pivetti

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Classical distribution theories distinguish between that part of the annual product which is necessary for its reproduction (including necessary subsistence for workers and replacement of the means of production) and the remainder (the ‘surplus’), and seek to explain the size of the surplus and its distribution among classes. They do not view the real wage rate and the rate of profit as determined by the relative scarcity of labour and capital; rather, one of the two distributive variables is explained independently from both the social product and the other distributive variable, and the other is determined as a residual.

Keywords

classical distribution theories; classical economics; competition; German Historical School; interest rates; Keynesian distribution theory; Marx, K. H.; natural price; real wage rate; Ricardo, D.; Smith, A.; Sraffa, P.; surplus; subsistence; value theory; wage–profit relationship

Article

The terms ‘classical economists’ and ‘classical political economy’ were first used by Marx, whose monumental survey of economic theory from the middle of the 17th century up to the early 1860s was contained in the manuscript written between January 1862 and July 1863 which the author called *Theorien über den Mehrwert*. Marx used the terms to describe ‘the critical economists’, ‘the economic investigators … like the Physiocrats, Adam Smith and Ricardo’ whose ‘urge’ was ‘to grasp the inner connection of the phenomena’; he also referred to Ricardo as ‘the last great representative’ of classical political economy (Marx, 1862–3, vol. 3, pp. 453, 500 and 502; 1873, p. 24).

Marx's description implies that not only authors like Senior, Bastiat, Wilhelm Roscher and John Elliot Cairnes are extraneous to classical political economy, but also such faithful Ricardians as James Mill, McCulloch and John Stuart Mill do not properly fit into it. This can only be understood if one bears in

mind that the ranking of the various authors in *Theorien über den Mehrwert* is centred upon the nature of their contributions to the related subjects of distribution and value: the explanation of profit and the formation of a normal or general rate of profit; the relation between wages and profits, the difficulties in the theory of value that arise in connection with the wage–profit relationship and the formation of a general rate of profit are the chief theoretical questions in the light of which the various authors are surveyed.

Thus a first discriminating factor in Marx's critical survey is provided by each author's attitude towards the main analytical difficulties: whether this or that author shows himself to be aware of their presence and tries to solve them, albeit at the cost of falling into further difficulties and contradictions, or rather tends to present the theory as a fully satisfactory body of propositions by denying the difficulties and ‘immediately adapting the concrete to the abstract’ (Marx, 1862–3, vol. 3, p. 87). This factor explains why Marx is inclined to treat both Torrens (1815; 1821) and Malthus (in particular, 1827) as classical economists, while regarding James Mill as the beginner of the ‘disintegration’ of the Ricardian theory.

A second factor is the *weight* of the ‘vulgar’ element present in the contributions of the various authors – meaning by this the tendency to confine one's attention to the ‘superficial appearance of the phenomena’ *versus* ‘the urge to grasp [their] inner connection’. As an important example of this factor one may refer to the increasing tendency, after Ricardo, to explain distribution by competition and ‘the [changing] state of supply and demand’ (J. Mill, 1844, p. 42; see also J.S. Mill, 1848, p. 337, and Cairnes, 1874, pp. 168–74) – thereby gradually abandoning the classical conception according to which demand and supply can only determine the oscillations of distribution and prices either above or below their ‘natural’ values. A third discriminating factor is the ‘vulgar’ element represented by the mere apology for the existing state of affairs (Marx, 1962–63, vol. 3, p. 168), or, as Cannan was later to put it, by the ‘desire to strengthen the position of the capitalist against the labourer’ (Cannan, 1917, p. 206). Finally, a fourth factor may be indicated in the tendency to deny the existence of economic laws altogether, and to substitute shallow empiricism for theoretical analysis (think of the so-called Historical School of German political economy).

The theoretical approach to distribution and value ‘of the old classical economists from Adam Smith to Ricardo has been submerged and forgotten since the advent of the “marginal” method’ (Sraffa, 1960, p. v). A contribution to this effect certainly came from the fact that *Theorien über den Mehrwert* remained largely unknown among economists. (It was only in the early 1950s that some sections of the 1905–10 Kautsky edition were translated into English, whilst the complete English translation from the edition based on the original manuscript was made in 1963–71.) In what follows, we shall take ‘classical theory of distribution’ to mean the main elements which can be regarded as characterizing the approach to the problem of the division of the national product among classes followed by the English classical economists from Adam Smith to David Ricardo, later by Karl Marx, and, more recently, by Piero Sraffa – this century's greatest exponent of the ‘classical’ approach to distribution.

The classical method of approaching the problem of distribution is based upon a distinction between two parts in the annual product of society: that part which is necessary for its reproduction (which includes the necessary subsistence of the workers employed in the economy) and that part which can be ‘freely’ disposed of by the society and which constitutes its ‘net product’ or ‘surplus’ – what remains of the social product after deducting the necessary subsistence of the workers and the replacement of the means of production. It is the aim of the classical theory to explain the circumstances governing the size of the

surplus and its distribution among classes: ‘To determine the laws which regulate this distribution, is’, according to Ricardo, ‘the principal problem in Political Economy’ (Ricardo, 1821, p. 5). In the course of his work he succeeded in ‘getting rid of rent’, so as to concentrate on the problem of the distribution between capitalists and workers; in what follows rent will be left entirely out of account – one may suppose that fertile lands abound – and the essential features of the surplus approach to distribution will be illustrated with reference to the determination of wages and profits.

Contrary to the supply-and-demand approach, which has been the dominant method over the last hundred years, in the theoretical approach to distribution of the classical economists and of Marx, the real wage rate and the rate of profit are not symmetrically and simultaneously determined on the basis of the relative scarcity of labour and capital. Within the classical approach, one of the two distributive variables is explained independently from both the social product and the other distributive variable, and the other one is determined as a residual.

Both the classical economists and Marx considered the real wage as constituting the independent or ‘given magnitude’ in the relation between the two distributive variables, maintaining that its normal level is determined by ‘subsistence’. Normal profits, reckoned gross of interest, are determined as a residual, on the basis of the dominant techniques of production. Given the dominant techniques, the level of the wage rate is thus the only circumstance upon which the normal rate of profit depends and no increase in the latter can be conceived of but through a fall in the former.

Wages and profits

It is in the context of this relation between wages and profits that the problem of value arises within the classical theory. All the surplus product of the annual labour of the economy, exceeding the portion absorbed by labour itself in the form of wages, must be divided among the individual capitalists according to the capitals they have employed in production. It is the very task of relative prices (‘natural prices’ or ‘prices or production’) to ensure such proportional division of the profit share of the surplus, and in order to perform their task relative prices are bound to change in the face of any increase or fall in the quantities of the various commodities accruing to the labourers as wages. This change in relative prices, and in the value of the social product, which must necessarily take place whenever nothing changes but distribution, makes it difficult to determine the effect on profits of a rise and fall in wages; it obscures the inverse relationship between wages and profits which would be apparent if output and its means of production were the same in kind, or if their values remained unaffected by changes in the division of the product. Hence Ricardo's search for a measure of value which would be invariant to changes in wages (Ricardo, 1821, ch. I, sections IV, V and VI; Sraffa, 1951, pp. xlviii–xlix); hence also, Marx's determination of the general rate of profit *before and independently from* the ‘prices of production’, on the basis of magnitudes (the quantities of labour bestowed in the production of the relevant heterogeneous aggregates of commodities) invariant to changes in the division of the product (Marx, 1894, ch. 9).

Only recently was a solution provided (Sraffa, 1960) to the difficulties inherent in the theory of value that were left unresolved by Ricardo and Marx. The picture outlined above, however, points to a clear subordination of the problem of value to the determination of distribution. This contrasts sharply with the dominant supply-and-demand approach, where the theory of value – the conception of equilibrium

prices as allocators of given factor endowments and their determination simultaneously with normal outputs and the equilibrium prices of factor services (distribution) – comes almost to coincide with economics itself.

As mentioned above, the real wage rate is explained by the classical authors in terms of ‘subsistence’. They included in this notion ‘not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without’, and ‘the want of which would be supposed to denote that disgraceful degree of poverty, which no body can well fall into without extreme bad conduct’ (Smith, 1776, vol. 2, p. 399). Their conception, in other words, was that the normal wage rate ‘depends not merely upon the physical, but also upon the historically developed social needs, which become second nature. But in every country, at a given time, this regulating average wage is a given magnitude’ (Marx, 1894, p. 859; cf. also Torrens, 1815, pp. 62–3).

The classical authors also ascribed to the conditions of competition on the labour market the possibility of influencing real wages for fairly long periods of time, and hence of causing shifts away from the normal distribution of income between capitalists and workers. Smith referred to the possibility that under certain circumstances, connected with the pace of accumulation and the growth in productivity of labour, ‘the scarcity of hands’ or a ‘scarcity of employment’ may move the wage above or below the normal average level (Smith, 1776, vol. 1, pp. 77 and 80). Starting from Smith's analysis, Marx went on to consider the movements of wages in the periodic alternations of the industrial cycle as regulated ‘by the varying proportions in which the working-class is divided into active and reserve army, …, by the extent to which it is now absorbed, now set free’ (Marx, 1883, p. 596).

Normal wages having been explained in terms of subsistence, the normal rate of profit must be determined as a residual on the basis of the dominant techniques of production. Those firms which, within each sphere of production, employ more backward or more advanced techniques than the dominant ones, earn profits that are respectively smaller or greater than normal.

In this conception, the conditions of competition amongst capitalists do not have any role to play as regulator of the normal distribution of income between wages and profits. It is easy to see on the basis of Sraffa's price equations (Sraffa, 1960, paras 1–4) that, given the wage in terms of specified necessaries and the methods of production, if there is a surplus product in the economy then the system necessarily determines, together with prices, also a positive general rate of profit which no competition whatsoever among capitalists can eliminate or change. If real wages, in other words, determined by historical and social conditions independently from prices and from the rate of profit, absorb only a part of the net product of the economy, it is simply impossible for competition, however intense it may be, to determine prices such as to render nil or ‘as low as possible’ what remains of the value of the product after the means of production have been reintegrated and the wages paid.

It is true that the competition amongst the owners of capital plays an important role Smith's theory: he makes the level of the ‘natural’ rate of profit depend on it. But this is precisely where the basic contradiction in Smith's theory may be seen. On the one hand he considers the real wage to be determined by subsistence; on the other he maintains that the rate of profit is determined by competition amongst capitalists, which, by growing more intense as accumulation proceeds, would make ‘the ordinary rate of profit as low as possible’ (Smith, 1776, vol. 1, p. 106). In short, his reasoning proceeds as if *both* distributive variables could be determined independently from each other.

Leaving aside Smith's contradiction, it can be affirmed that in classical and Marxian theory competition is envisaged essentially as the mechanism whereby, in each sphere of production, a single price tends to be established: the price that enables the means of production to be reintegrated on the basis of the dominant production techniques, and wages and profits to be paid at their normal rates. These latter must be explained independently from competition, and, as Marx puts it, it is they that regulate competition, rather than being regulated by it (Marx, 1894, p. 865). The competition amongst firms within each sphere of production and the free transferability of capital from one sphere to another – hence the process whereby profit rates gravitate towards their respective normal levels – may be impeded by the presence of monopoly elements in this or that sphere of production. This however will affect the division of profits amongst the particular stocks making up social capital, but not the normal distribution of net output between wages and profits (Marx, 1894, p. 861).

Interest and profits

Profits on capital employed in production normally include, according to the classical economists, besides interest, also a remuneration for the ‘risk and trouble’ of productively employing it, or what may be termed a normal profit of enterprise. Production and accumulation would not continue, Ricardo argues, if the profits of the farmers and the manufacturers were ‘so low as not to afford an adequate compensation for their trouble and the risk which they must necessarily encounter in employing their capital productively’ (Ricardo, 1821, p. 122). Such ‘adequate compensation’ will be different in the various employments of capital, according to ‘any real or fancied advantage which one employment may posses over another’ (Ricardo, 1821, p. 90). On the basis of this conception, natural prices will have to be such as to ensure that, in each sphere of production, what remains of the value of the product after deducting wages and the replacement of the means of production, is sufficient to ‘adequately’ remunerate the ‘risk and trouble’ and pay interest at an uniform rate. It can thus be said that interest and profit of enterprise are conceived in the classical analysis as the two magnitudes into which normal profits – determined by real wages and production techniques – resolve themselves.

The money rate of interest emerges from this picture as a magnitude subordinate to the normal rate of profit, being ultimately determined by those real forces, the real wage rate and production techniques, which explain the course of the normal rate of profit. But what if actual experience did not validate the conception of the money rate of interest as a subordinate phenomenon? A few significant modifications would be called for within the classical–Marxian approach to distribution, if it had to be acknowledged that the level of the rate of interest in any one country is strongly influenced by circumstances which have nothing to do with the real forces regarded by the classical economists as governing the rate of profit. These modifications, as will be apparent from the determination of distribution outlined below, would lead to a view of the real wage as the residual rather than the independent or ‘given’ variable in the relation between profits and wages.

It is important to notice that the replacement of the wage by the rate of profit as the independent distributive variable is fully compatible with the surplus approach to distribution (cf. Garegnani, 1984, pp. 320–2). The concept of profits as surplus product is not under discussion when asking which of the two distributive variables should be regarded as ‘given’ in the present reality of the capitalist economy. The question is whether the relations that workers and capitalists establish with one another tend

primarily to act upon the real wage or upon the rate of profit, once the view is abandoned that real wages consist of the necessary subsistence of the workers and the possibility of variations in the division of the social surplus is admitted.

Actual experience seems in fact to validate the conception of an autonomous determination of the money rate of interest – autonomous in the sense that interest rates *do* experience lasting changes which are very reasonably explainable without any need to refer to a *primum movens* represented by changes in the normal profit rate. Interest rates in any one country depend directly on monetary policy; interest rate policy decisions, however, are taken under a wide range of constraints having different weights both amongst the various countries and for the same country at different times: external constraints, monetary and fiscal constraints, distributive constraints. The important point is that interest rate policies, both in the short and in the long run, do not appear to be constrained by a predetermined normal profitability of capital. Once this point is acknowledged, then, given the necessary (and generally admitted) long-run connection between the rate of interest and the rate of profit, it will also be acknowledged that it is the former which 'sets the pace' and that the latter will have to adapt itself. On this basis, one can proceed to discover the actual mechanism whereby the causation occurs and to study its implications (see Pivetti, 1985).

The actual mechanism whereby lasting changes in interest rates are susceptible of causing corresponding changes in normal profit rates, can be understood by following a three-stage line of reasoning. The first stage simply consists in regarding competition as the mechanism by which prices tend to be equated to normal costs. The second stage of the reasoning consists in looking at the rate of interest as a determinant of production costs, together with money wages and production techniques. Thus, lasting changes in interest rates *constitute* changes in normal costs, which, *ceteris paribus*, will result in corresponding changes of the price level. The third stage of the reasoning comes about as a consequence of the first two: by the competition amongst firms within each industry, a lasting change in interest rates causes a change in the same direction in the level of prices in relation to the level of money wages, thereby generating changes in income distribution.

The rate of interest thus emerges from our picture as the regulator of the ratio of prices to money wages. The reader will note the main difference between this view and the so-called post-Keynesian theory of distribution: whilst in that theory changes in the level of prices in relation to the level of money wages are determined by changes in aggregate demand, according to the present explanation of distribution they are determined by lasting changes in interest rates.

By taking into consideration also the excess of profit over interest, or profit of enterprise, our conception of the rate of interest as the regulator of the ratio of prices to money wages requires us to assume that lasting changes in the rate of interest do not tend, and are not likely, to be associated with opposite changes in the normal profit of enterprise. This assumption is largely consistent with classical conceptions as regards the normal excess of profit over interest: if profit *does* normally exceed interest (if competition, that is, does not tend to equalize profit and interest), then the excess of the former over the latter must cover objective elements of 'risk and trouble' or elements which are regarded as objective by the majority of the investing public. By taking into account all such elements, we can say that the normal rate of profit in each particular production sphere will be arrived at by adding up *two* autonomous components: the long-term rate of interest or 'pure' remuneration of capital, plus the normal profit of enterprise or the remuneration for the 'risk and trouble' of productively employing capital in that sphere of production. Provided this remuneration is a sufficiently stable magnitude, lasting changes

in the rate of interest will cause corresponding changes in profit rates, and inverse changes in the real wage.

Real wages as a residue

As we saw above, interest and profit of enterprise are conceived by the classical economists as the two magnitudes into which normal profits resolve themselves, whereas, according to our view, the same two magnitudes should rather be regarded as the *determinants* of the rate of profit. Given the money wage, the real wage appears here as a residue on the basis of the price level reflecting the dominant techniques in the different spheres of production and the normal profit rate determined in each sphere in the way we have just indicated. From this determination of distribution, quite different views from the classical ones may be developed concerning the role of competition amongst capitalists.

Since in our view the real wage constitutes the residual variable, the presence of monopoly elements in this or that sphere of production may affect not only the division of profits amongst the different employments of capital, but also the distribution between profits and wages. Given in fact the money wage, the possibility for some commodities to obtain a monopoly price which rises above the ‘price of production’ will translate into a ratio price-level/money wage which will be higher than it would be if there were no monopoly elements, and hence into a lower real wage. Assuming the long-term rate of interest to be unaffected by the presence of monopoly elements, it follows that lasting effects of the conditions of competition on distribution may only be obtained in one direction: higher profits than normal. For the long-term interest rate and the normal remuneration of ‘risk and trouble’ establish, in each sphere of production, the minimum or necessary level below which the profit rate cannot go, over the long run, however intense one may suppose the forces of competition to be.

The possibility must also be admitted that the conditions of competition influence the normal profit rate via the long-term interest rate. At the root of this possible influence of competition there is the fact that the level of the real wage constitutes *in any case* an important constraint on the freedom of monetary policy to establish the level of interest rates. To acknowledge that lasting variations in the rate of interest determine variations in the normal distribution between profits and wages is not to concede that the real wage may move to any level whatsoever. In each concrete situation, it would be hard to carry on the productive process in an orderly manner if the real wage were lower than certain levels reflecting institutional and historical as well as economic circumstances. Thus, if the conditions of competition have a negative effect on wages – via the levels of profits of enterprise or the methods of production adopted – then beyond certain limits, which will vary from one situation to another, a compensatory effect will have to be sought in the level of interest rates.

According to our view, then, the money rate of interest should be looked on as the magnitude on which the respective powers of capitalists and workers discharge themselves *in the first place*. Wage bargaining and monetary policy are regarded as the main channels through which class relations act in determining distribution, and those relations are seen as tending to primarily act upon the profit rate, via the monetary rate of interest, rather than upon the real wage rate as maintained by both the classical economists and Marx. The level of the real wage prevailing in any given situation is the *final* result of the whole process by which distribution of income between workers and capitalists is actually arrived at. It seems to us that in the conditions of modern capitalism it is difficult to conceive of the real wage rate as the independent or given variable in the relationship between wages and profits – the difficulty, as we

see it, arising from the fact that the direct outcome of wage bargaining is a certain level of the money wage, while the price level cannot be determined before and independently from money wages. Given distribution between profits and wages, and given the methods of production, the level of prices simply depends on the level of money wages. Thus, in our picture, the long-term rate of interest enters into the determination of the price level because it contributes to regulating the ratio of the latter to the money wage – that is, distribution between profits and wages.

If instead the real wage is taken as given, the ratio of prices to money wages will be determined by the condition that it must be such as to ensure the given level of the real wage; and on this basis wage bargaining, in determining money wages, can be thought of as determining also the price level. In such a picture monetary policy plays a purely passive role – the level of the rate of interest having to accommodate to lasting changes in the ratio of prices to money wages, rather than governing that ratio. Now what we are ultimately facing here is a conception of the ratio of prices to money wages as being determined by a magnitude, the real wage rate, which is not actually known before that ratio is known. This explains in our opinion why of the two alternative propositions – that the ratio of prices to money wages depends on the real wage rate, or that the real wage rate depends on the ratio of prices to the money wage – the latter is easier to digest: in actual fact, there are no circumstances determining real wages as distinct from those acting through money wages, the level of prices and the ratio of prices to money wages.

See Also

- surplus

Bibliography

Cairnes, J.E. 1874. *Some Leading Principles of Political Economy Newly Expounded*. New York: Harper & Brothers.

Cannan, E. 1917. *A History of the Theories of Production and Distribution in English Political Economy from 1776 to 1848*. 3rd edn, London: King.

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 36, 291–325.

Malthus, T.R. 1827. *Definitions in Political Economy, Preceded by an Inquiry into the Rules That Ought to Guide Political Economists in the Definition and Use of Their Terms; With Remarks on the Deviations from These Rules in Their Writings*. Reprinted, New York: Kelley, 1971.

Marx, K. 1862–63. *Theories of Surplus Value*, vols 1–3, Moscow: Progress Publishers, 1963–71.

Marx, K. 1873. Afterword to the 2nd German edition of *Capital*. In Marx (1883).

Marx, K. 1883. *Capital: A Critique of Political Economy*, vol. 1, 3rd edn, reprinted, London: Lawrence & Wishart, 1977.

Marx, K. 1894. *Capital: A Critique of Political Economy*, vol. 3, London: Lawrence & Wishart, 1977.

Mill, J. 1844. *Elements of Political Economy*. 3rd edn, New York: Kelley; 1965.

Mill, J.S. 1848. *Principles of Political Economy with Some of Their Applications to Social Philosophy*. Toronto: University of Toronto Press and London: Routledge & Kegan Paul, 1965.

Pivetti, M. 1985. On the monetary explanation of distribution. *Political Economy – Studies in the Surplus Approach* 1(2).

Ricardo, D. 1821. *Principles of Political Economy and Taxation*. 3rd edn, in *The Works and Correspondence of David Ricardo*, ed. P. Sraffa, vol. 1. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. E. Cannan, London: Methuen, 1961.

Sraffa, P. 1951. Introduction to Ricardo's *Principles*. In *Collected Works of David Ricardo*, vol. 1. Cambridge: Cambridge University Press.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Torrens, R. 1815. *An Essay on the External Corn Trade*. London: Hatchard.

Torrens, R. 1821. *An Essay on the Production of Wealth. With an Appendix in Which the Principles of Political Economy Are Applied to the Actual Circumstances of This Country*. Reprinted, New York: Kelley, 1965.

How to cite this article

Pivetti, Massimo. "classical distribution theories." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000156> doi:10.1057/9780230226203.0243

The New Palgrave Dictionary of Economics Online

classical economics and economic growth

Gavin Cameron

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The classical economists dealt with many of the issues now addressed by modern growth theories, albeit with different theoretical tools and with different perspectives. Classical analyses of the division of labour, population growth, and the difficulties when factors are in fixed supply, continue to have modern applications. However, the models they developed ran into difficulties after the ‘marginalist revolution’, when it became apparent that sustained technical change, abstinence and thrift by the labouring classes, and factor substitution might forestall the arrival of the stationary state.

Keywords

balanced growth; capital accumulation; classical economics; classical economics and economic growth; diminishing returns; distribution theory; division of labour; economic growth; falling rate of profit; fertility; human capital; Industrial Revolution; labour supply; luxury; Malthus, T. R.; marginalist revolution; Marx, K.H.; Mill, J. S.; mortality; population growth; productive and unproductive labour; Ricardo, D.; Say, J.-B.; Smith, A.; stationary state; technical change; value theory

Article

The analysis of economic growth was an important feature of the writings of the great classical economists, including Adam Smith, Thomas Malthus, David Ricardo, John Stuart Mill and Karl Marx. To place them in their historical context is straightforward if economic history is simplified into three distinct epochs. In the first, which spanned most of human history and still obtains in some unfortunate regions, Malthusian conditions prevailed: living standards were static even though there was some population growth. In the second, which began in the middle of the 18th century in England, living standards showed some upward tendency and there was a demographic change as fertility rates rose and mortality rates fell, resulting in a substantial rise in population. In the third epoch, characteristic of England from the 1820s perhaps, the move to sustained economic growth provoked a shift from quantity

to quality in child-rearing, and all the appurtenances of modern growth began to appear, such as human capital, professional R&D, and technical innovation.

There is much scope for discussion about what factors triggered, propagated, and enhanced such changes, and about when such changes began and whether they were smooth or discrete. For example, Mokyr (2005) argues that living standards in England rose gently between the 17th and 18th centuries due to the spread of world trade, commercialism and the rise of institutions less hostile to consumers and the industrious – nicely, this is sometimes called ‘Smithian’ growth. Somewhat in contrast, Allen (2001) argues that real wages did not rise significantly over that period in England, but that, since they were falling across most of Europe, the real question is what would have happened in the absence of the Industrial Revolution.

Mokyr also points out that many of the inventions associated with the 18th century Industrial Revolution were developed in north-west Europe, but successfully applied in England. It is not surprising that the classical economists were fascinated. Adam Smith was born in 1723, within the Malthusian growth regime, whereas Ricardo, Malthus and Jean-Baptiste Say were well placed to observe the demographic change in England and the beginnings of industry, even though England was still predominantly a rural society in the early 19th century. Unsurprisingly, Mill and Marx found it increasingly hard to defend Ricardian doctrines as the modern growth regime began to emerge across Europe and its offshoots in the middle of the 19th century.

Being products of the Enlightenment, the classical economists shared a concern for human progress that would do credit to a modern policymaker. One purpose of their analysis was to identify the forces in society that promoted or hindered progress and to provide a basis for policy and action in a time of considerable political innovation in England (including land enclosures, franchise reform, tariff reform, and the abolition of the slave trade) and revolution abroad (including land reform, the continental system, and the tumbrils). This background motivated Ricardo's campaign against the Corn Laws, as it did Malthus's concern with population growth, Smith's attacks on mercantilism, and Marx's analyses of social class.

The classical economists' work was grounded in the economic conditions of their times, and not in the abstract mathematical reasoning that appeared in economics during the marginalist revolution of the 1870s and after, popularized by Ysidro Edgeworth, William Stanley Jevons and Alfred Marshall. In contrast to more recent economic thought, the classical economists saw discussions of economic growth as being inextricably linked with discussions of the theory of value and the theory of distribution. Since their concerns were largely those of educated gentlemen of those times, they wanted to be able simultaneously to explain trade cycles, inflation and other short-run phenomena, as well as real wages and population growth and other long-run phenomena. While it is easy to see the current gap between short-run and long-run macroeconomic models as a lacuna (for example, see Solow, 2005), the classical economists tended to run into problems when treating both at the same time.

> The characteristic features of what is commonly meant by industrial progress, resolve themselves mainly into three, increase of capital, increase of population, and improvements in production; understanding the last expression in its widest sense, to include the process of procuring commodities from a distance, as well as producing them. (Mill, 1848, Book IV, ch. 3)

The classical economists also worried about the consumption of luxuries and the distinction between productive and unproductive labour. As Brewer (1997) discusses, this is particularly true of Adam Smith, who displays a good deal of ambivalence about luxuries:

> That portion of his revenue that a rich man annually spends is in most cases consumed by idle guests and menial servants, who leave nothing behind them in return for their consumption. That portion which he annually saves, for the sake of the profit it is immediately employed as a capital, is consumed in the same manner, and nearly the same time too, but by a different set of people, by labourers, manufacturers and artificers, who reproduce with a profit the value of their annual consumption. (Smith, 1776, Book II, ch. 3)

Smith's view contrasts somewhat with that of his predecessor David Hume, whose mild approval of luxuries was based on the notion that they might encourage economic and political development. Although such notions still figure in modern debates (Greenhalgh, 2005), this preoccupation with luxuries and unproductive labour turns out to be not very useful for modelling purposes, unless it is simply be taken to mean that different economic groups have different propensities to save, which is a truism. However, even if the classical economists did not always approve of certain kinds of consumption, Smith's contention that consumption is the sole end and purpose of all production was a vast improvement on the mercantilist doctrine.

Clearly, the classical economists cannot be written off as growth theorists manqué. The technical core of modern growth theory rests upon technical change, specialization, factor substitution, and factor accumulation, with various recent theorists emphasizing the effects on these of trade, institutions, inequality, political economy, geography and population size and growth. All these issues were concerns of the classical economists, even if they used a different vocabulary.

Nonetheless, it would be fair to say that the classical economists have had only a limited direct impact on recent growth theorists. Adam Smith receives seven references in the current two-volume *Handbook of Economic Growth* (Aghion and Durlauf, 2005). Malthus a very respectable 13, while of the other classical economists only Ricardo merits a single mention. Interestingly, an even older economist, William Petty from the 17th century, is often quoted in writing about the effect of population size on inventiveness in the scale effect literature (see Jones, 2005).

The stationary state

The classical economists saw all around them the effects of the development of the capitalist system, most importantly, of course, the accumulation of capital, but also the introduction of new techniques. Smith analysed in great detail the process of the division of labour, but more generally the classical economists did not attempt to deal with the relationship between capital accumulation and technical change (although Marx did highlight the issue). In addition to these basic forces of economic growth, they were also interested in the increase in the supply of labour through population growth. In the case of Thomas Malthus, this interest was quite morbid.

> The power of population is so superior to the power in the earth to produce subsistence for man that premature death must in some shape or other visit the human race. (Malthus, 1798)

The classical economists' analysis of the process by which capital, technology and labour grow over time led them to a common conclusion, motivated by different causes – that the process of economic growth was gradually self-attenuating and ended in a state of stagnation (the ‘stationary state’):

> When the stocks of many merchants are turned into the same trade, their mutual competition naturally tends to lower its profit; and when there is a like increase of stock in all the different trades carried on in the same society; the same competition must produce the same effect in them all. (Smith, 1776, Book I, ch. 4)

The principal way in which Smith envisaged a stationary state as obtaining was that the rate of profit would fall as capital accumulated in the long run due to increased competition. Smith associated this stationary state with the position of China, which he described as being one of the most fertile and industrious countries, but also as having low wages and having been long stationary. There is tension in the *Wealth of Nations* between three separate points: first, his worries about the falling rate of profit; second, his worries that wages could fall to a subsistence level; and third, his description of net saving creating higher levels of output. This shows that although the economic system he describes is very complex, it tends to neglect both the feedback between profits and saving, and substitution between capital and labour.

Some controversy exists about the origin of the idea of ‘diminishing returns’, although it certainly appears in the writings of Jacques Turgot in the 18th century. The early 19th-century English economists certainly saw the idea in action with the expansion of cultivated land in England during the Napoleonic Wars. Subsequently, the idea comes to life in Ricardo's ‘corn’ model. Modern presentations of this model are plentiful (see for example, Kaldor, 1956; Pasinetti, 1960; Samuelson, 1978; discussions in Glyn, 2004). The presentation here follows Bhaduri and Harris (1987).

Suppose that there is a single product, ‘corn’, produced in a capitalist agricultural economy. Land differs in its fertility and labour is applied in fixed proportions to land of diminishing fertility. The supply of labour is perfectly elastic at some fixed real wage equal to ‘subsistence’ (this is clearly an extreme form of the Malthusian hypothesis; see for example, in Samuelson, 1978, and discussion in Brezis and Young, 2003). Total output is distributed between rent paid to landlords, profits to capitalists, and wages. The level of land rent can then be shown to be determined by the difference between the average and marginal product of labour at the prevailing level of employment, and profits are the residual after rent and wages are paid (equal to the marginal product of labour minus the wage, times employment). Although there is a variety of Ricardian schemes for the determination of saving (and hence capital accumulation in a closed economy with no consumption loans), a typical presentation takes saving to be a constant proportion of profits, so the rate of accumulation is uniquely dependent upon the profit rate. However, as employment growth proceeds, the marginal product of labour falls and so must the profit rate. The system asymptotically approaches a stationary state when the profit rate is so low that accumulation ceases (the ‘minimum acceptable rate of profit’). What happens is that capitalists find

themselves squeezed between the diminishing product of labour and the need to pay the going wage rate, and paying out an increasing share of output as rent to landlords. There is thus a conflict between landlords and capitalists.

In the absence of technical change, the possibility that landlords or workers could themselves become savers, or substitution away from that resource, any other fixed resource would play the same role. Samuelson (1978) notes that neither Ricardo nor Marx was so naive as to believe literally in fixed proportions between capital goods and labour, but their models were unable fully to reflect this complexity.

Mill provides both a summary and a synthesis of previous writers, drawing particularly on Ricardo:

> On the whole, therefore, we may assume that in a country such as England, if the present annual amount of savings were to continue, without any of the counteracting circumstances which now keep in check the natural influences of those savings in reducing profit, the rate of profit would speedily attain the minimum, and all further accumulation of capital would for the present cease. (Mill, 1848, Book IV, ch. 4)

Mill contradicts Smith's assertion that competition is the cause of the falling profit rate and proposes instead a form of diminishing returns to capital, provided by limits to the ‘field of employment’ of capital. He then explicitly links capital accumulation with saving and notes that there is some minimum rate of profit, below which capital accumulation cannot take place. However, he does propose four mechanisms by which the stationary state may be overcome: first, that capital may be wasted during speculative booms; second, through improvements in production; third, through an expansion of foreign trade, and fourth, through the export of capital to other countries.

The second is the one that resonates with modern growth theory, although Mill muddies the waters with a contradictory passage about why an improvement in the production of luxuries (such as lace and velvet) will affect capital accumulation through a different mechanism.

Marx was also a firm believer in this movement towards a stationary state, exemplified by what he called the falling tendency of the rate of profit (FTRP). In the Marxian scheme, the FTRP is one of the main sources of crises under capitalism. Writers in this tradition usually understate the ability of technical progress to reliably prevent such crises and overstate the role of the business cycle in long-run development. Not every slump or financial crash heralds the end of capitalism. But on the former point, Marx was writing at an early stage of the sustained growth era, largely before the existence of large-scale industrial processes and certainly before professional R&D laboratories (see Glyn, 2006, for a discussion of whether the entry of China and India into the global economy might presage a return to a Marxian era of growth). In such an era, technical innovation may well have appeared more uncertain and less widespread than it would later appear, or, to use Harberger's analogy, more like mushrooms popping up here and there than like yeast leavening the entire economic process (Harley, 2003).

It can be seen that the classical economists were much more concerned about the stationary state than if it just represented an equilibrating tendency in a long-run growth model à la Solow where capital deepening slows in the absence of technical change (this is clear from Sweezy, 1942, ch. 9).

Nonetheless, in the idea of the stationary state (and from Mill's view that he was considering the ‘dynamics’ of the economy, having dealt with the ‘statics’), it is possible to see the seed-corn of the

Solow model, once economists such as Marshall, Frank Ramsey, Charles Cobb, and Paul Douglas had laid further foundations.

In contrast, classical theories of growth qua theories of growth became increasingly marginal as the 19th century wore on (although of course, Marxian and Marxist analysis remained influential for much longer). The Swedish unemployment of the early 1920s prompted Knut Wicksell to write three articles from a neo-Malthusian standpoint, one of which, entitled ‘Ricardo on Machinery and the Present Unemployment’, he submitted to the *Economic Journal*. John Maynard Keynes, the editor of the journal, rejected the paper, arguing ‘that any treatment of this topic at the present day ought to bring in various modern conceptions for handling the problem and the time has gone by for a criticism of Ricardo on purely Ricardian lines’ (J.M. Keynes, quoted in Jonung, 1981). In the end, even Piero Sraffa's remarkable work, *Production of Commodities by Means of Commodities* (1960), was not enough to revive Ricardian analysis, although some still see neoclassical economics as its direct descendant (Hollander, 1995).

Conclusion

Classical economists are often regarded as ‘pessimistic’ in their forecasts of the future development of the economy, and came in for heavy criticism from the unlikeliest of sources, the Romantic poets and literary critics such as Ruskin. This kind of *trahison des clercs* of poets and authors against a changing social order and increasing commercialization is familiar to a modern reader of tracts against global capitalism, and equally well grounded in theory and evidence.

The classical economists' search for a ‘theory of value’ and a ‘theory of distribution’ was an attempt to understand the significant economic, political, and social changes of their times, as well as an attempt to understand what would happen in the long run in those economies. There is much to be learnt from their analyses, both as an indicator of the conditions of the times (that is, the importance of land as a factor of production) and also as a precursor to the future development of the theory of economic growth. Without the analytical apparatus that arose during the marginalist revolution (such as production functions and utility functions), their analyses were hampered, but a number of the features that drive modern models of growth made their first appearance in the writings of the classical economists. For example, the importance of the division of labour, technical progress and the role of population growth, as well as the idea of diminishing returns, all feature prominently in modern models.

What is lacking from the classical accounts is the notion of a balanced growth path. The classical economists largely concluded that, in the long run, economies would tend towards a stationary, stagnant state. They emphasized the ability of population growth to keep wages at subsistence level, the notion that capital could only be accumulated out of profits, and the central role of land as a factor of production. In this sense, their analytical scheme is flawed. Economic progress has shown that the possibility of investment in human capital can lead to a demographic shift whereby households choose ‘quality’ over ‘quantity’ in their reproductive choices; that saving by workers can be an important source of capital accumulation; and that factor substitution tends to prevent the inexorable rise in the price of any factor, even if it is in fixed supply.

See Also

- balanced growth
- Malthus, Thomas Robert
- Marx, Karl Heinrich
- Mill, John Stuart
- Ricardo, David
- Smith, Adam

The author would like to thank Robert Allen, Julia Cartwright, Mary Dixon-Woods, Marcel Fafchamps, Nicholas Fawcett, Andrew Glyn, Mark Koyama, Silvia Palano, and Jonathan Temple for helpful comments on a preliminary draft.

Bibliography

Aghion, P. and Durlauf, S. 2005. *Handbook of Economic Growth*. Amsterdam: North-Holland.

Allen, R. 2001. The great divergence in European wages and prices from the Middle Ages to the First World War. *Explorations in Economic History* 38, 411–47.

Bhaduri, A. and Harris, D. 1987. The complex dynamics of the simple Ricardian system. *Quarterly Journal of Economics* 102, 893–902.

Brewer, A. 1997. Luxury and economic development: David Hume and Adam Smith. *Scottish Journal of Political Economy* 45, 78–98.

Brezis, E. and Young, W. 2003. The new views on demographic transition: a reassessment of Malthus's and Marx's approach to population. *European Journal of the History of Economic Thought* 10, 25–45.

Glyn, A. 2004. The corn model, gluts and surplus value. Working Paper No. 194, Department of Economics, Oxford University.

Glyn, A. 2006. Will Marx be proved right? *Oxonomics* 1, 13–6.

Greenhalgh, C. 2005. Why does market capitalism fail to deliver a sustainable environment and greater equality of incomes? *Cambridge Journal of Economics* 29, 1091–109.

Harley, K. 2003. Growth theory and industrial revolutions in Britain and America. *Canadian Journal of Economics* 36, 809–31.

Hollander, S. 1995. *Collected Essays I: Ricardo. The 'New View'*. London: Routledge.

Jones, C. 2005. Growth and ideas. In Aghion and Durlauf (2005).

Jonung, L. 1981. Ricardo on machinery and the present unemployment: an unpublished manuscript by Knut Wicksell. *Economic Journal* 91, 195–8.

Kaldor, N. 1956. Alternatives theories of distribution. *Review of Economic Studies* 28, 83–100.

Malthus, T.R. 1798. *An Essay on the Principle of Population*. Oxford: Oxford World's Classics, 1999.

Mill, J.S. 1848. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*. London: Penguin Classics, 1985.

Mokyr, J. 2005. Long-term economic growth and the history of technology. In Aghion and Durlauf (2005).

Pasinetti, L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies* 27, 78–98.

Samuelson, P. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16, 1415–34.

Samuelson, P. 1988. Mathematical vindication of Ricardo on machinery. *Journal of Political Economy* 96, 274–82.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, R.H. Campbell and A. S. Skinner, eds., 2 vols. Oxford: Oxford University Press, 1976.

Solow, R. 2005. Introduction: growth in retrospect and prospect. In Aghion and Durlauf (2005).

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. New York: Cambridge University Press.

Stigler, G. 1958. Ricardo and the 93% labor theory of value. *American Economic Review* 48, 357–67.

Sweezy, P. 1942. *The Theory of Capitalist Development*. New York: Monthly Review Press.

How to cite this article

Cameron, Gavin. "classical economics and economic growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008

<http://www.dictionaryofeconomics.com/article?id=pde2008_C000603>
doi:10.1057/9780230226203.0244

The New Palgrave Dictionary of Economics Online

classical growth model

Donald J. Harris

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The classical economists provided an account of the broad forces that influence economic growth and of the mechanisms underlying the growth process, stressing accumulation and productive investment of a part of the social surplus in the form of profits. Changes in the rate of profit were decisive for analysis of the long-term evolution of the economy. The analysis indicated that in a closed economy there is an inevitable tendency for the rate of profit to fall. In this article, the essential features of the classical analysis of the accumulation process are presented and formalized in terms of a simple model.

Keywords

capital accumulation; capitalism; class; classical economics; classical growth model; corn as basic commodity; Corn Laws; diminishing returns; dismal science; distribution theories, classical; division of labour; economic growth; falling tendency of the rate of profit; household production; labour productivity; labour supply; Malthus, T. R.; Malthus's theory of population; Marx, K. H.; physiocracy; population growth; progress; rate of profit; reserve army of labour; Ricardo, D.; Smith, A.; stationary state; subsistence; surplus; technical change; value; wages fund

Article

Analysis of the process of economic growth was a central feature of the work of the English classical economists, as represented chiefly by Adam Smith, Thomas Malthus and David Ricardo. Despite the speculations of others before them, they must be regarded as the main precursors of modern growth theory. The ideas of this school reached their highest level of development in the works of Ricardo. The interest of these economists in problems of economic growth was rooted in the concrete conditions of their time. Specifically, they were confronted with the facts of economic and social changes taking place in contemporary British society as well as in previous historical periods. Living in the 18th and 19th centuries, on the eve or in the full throes of the Industrial Revolution, they could hardly help but be

impressed by such changes. They undertook their investigations against the background of the emergence of what was to be regarded as a new economic system – the system of industrial capitalism. Political economy represented a conscious effort on their part to develop a scientific explanation of the forces governing the operation of the economic system, of the actual processes involved in the observed changes that were going on, and of the long-run tendencies and outcomes to which they were leading.

The interest of the classical economists in economic growth derived also from a philosophical concern with the possibilities of ‘progress’ an essential condition of which was seen to be the development of the material basis of society. Accordingly, it was felt that the purpose of analysis was to identify the forces in society that promoted or hindered this development, and hence progress, and consequently to provide a basis for policy and action to influence those forces. Ricardo's campaign against the Corn Laws must obviously be seen in this light, as also Malthus's concern with the problem of population growth and Smith's attacks against the monopoly privileges associated with mercantilism.

Of course, for these economists, Smith especially, progress was seen from the point of view of the growth of national wealth. Hence, the principle of national advantage was regarded as an essential criterion of economic policy. Progress was conceived also within the framework of a need to preserve private property and hence the interests of the property-owning class. From this perspective, they endeavoured to show that the exercise of individual initiative under freely competitive conditions to promote individual ends would produce results beneficial to society as a whole. Conflicting economic interests of different groups could be reconciled by the operation of competitive market forces and by the limited activity of ‘responsible’ government.

As a result of their work in economic analysis the classical economists were able to provide an account of the broad forces that influence economic growth and of the mechanisms underlying the growth process. An important achievement was their recognition that the accumulation and productive investment of a part of the social product is the main driving force behind economic growth and that, under capitalism, this takes the form mainly of the reinvestment of profits. Armed with this recognition, their critique of feudal society was based on the observation, among others, that a large part of the social product was not so invested but was consumed unproductively.

The explanation of the forces underlying the accumulation process was seen as the heart of the problem of economic growth. Associated with accumulation is technical change as expressed in the division of labour and changes in methods of production. Smith, in particular, placed heavy emphasis on the process of extension of division of labour, but there is, in general, no systematic treatment of the relation between capital accumulation and technical change in the work of the classical economists. It later becomes a pivotal theme in the work of Marx and is subjected there to detailed analysis (see, for instance, Marx, 1867, part 4). To these basic forces in economic growth they added the increase in the supply of labour available for production through growth of population. Their analysis of the operation of these forces led them to the common view, though they quite clearly differed about the particular causes, that the process of economic growth under the conditions they identified raises obstacles in its own path and is ultimately retarded, ending in a state of stagnation – the ‘stationary state’.

The conception of the stationary state as the ultimate end of the process of economic growth is often interpreted as a ‘prediction’ of the actual course of economic development in 19th-century England. There is no doubt that it was for a time so regarded by some, if not all, of the economists and their contemporaries, though the weight that was assigned to this particular aspect of the conception by Ricardo himself is a matter of some dispute. What is more significant, however, is that this conception

served to point to a particular social group, the landlord class, who benefited from the social product without contributing either to its formation or to ‘progress’ and who, by their support of the Corn Laws and associated restrictions on foreign trade, acted as an obstacle to the only effective escape from the path to a stationary state, that is, through foreign trade.

In examining the work of the classical economists we find also that problems of economic growth were analysed through the application of general economic principles, viewing the economic system as a whole, rather than in terms of a separate theory of economic growth as such. These principles were such as to recognize basic patterns of interdependence in the economic system and interrelatedness of the phenomena of production, exchange, distribution and accumulation. In sum, what we find in classical economic analysis is a necessary interconnection between the analysis of value, distribution and growth. Because of these interconnections it was by no means possible to draw a sharp dividing line between the inquiry into economic growth and that into other areas of political economy. As Meek (1967, p. 187) notes:

> To Smith and Ricardo, the macroeconomic problem of the ‘laws of motion’ of capitalism appeared as the primary problem on the agenda, and it seemed necessary that the whole of economic analysis – including the basic theories of value and distribution – should be deliberately oriented towards its solution.

Distribution of the social product was seen to be connected in a definite way with the performance of labour in production and with the pattern of ownership of the means of production. In this regard, labour, land, and capital were distinguished as social categories corresponding to the prevailing class relationships among individuals in contemporary society: the class of labourers consisted of those who performed labour services, landlords were those who owned titles or property in land, and capitalists were those who owned property in capital consisting of the sum of exchangeable value tied up in means of production and in the ‘advances’ which go to maintain the labourers during the production period. Each class received income or a share in the product according to specified rules: for the owners, the rule was based on the total amount of property which they owned – so much rent per unit of land, so much profit per unit of capital (and, for the class of finance capitalists or ‘rentiers’ who lent money at interest, so much interest per unit of money lent). For labourers it was based on the quantity of labour services performed: so much wages per hour.

Accumulation and distribution were seen to be interconnected through the use that was made by different social classes of their share in the product. Basic to this view was a conception, taken over from the Physiocrats, of the social surplus as that part of the social product which remained after deducting the ‘necessary costs’ of production consisting of the means of production used up and the wage goods required to sustain the labourers employed in producing the social product. This surplus was distributed as profits, interest and rent to the corresponding classes of property owners. For the classical economists, the possibility of accumulation was governed by the size and mode of utilization of this surplus. Accordingly, their analysis placed emphasis upon those aspects of distribution and of the associated class behaviour which had a direct connection with the disposal of the surplus and therefore with growth. In particular, it was assumed that, typically, workers consumed their wages for subsistence, capitalists reinvested their profits and landlords spent their rents on ‘riotous living’. On the other side, accumulation would also influence the distribution of income as the economy expanded over time.

It was this absolutely strategic role of the size and use of the surplus, viewed from the perspective of the economy as a whole and of its process of expansion, which dictated the significance of the distribution of income for classical economic analysis. Thus, for Ricardo especially, investigation of the laws governing distribution became the focus of analysis. In a letter to Malthus, Ricardo wrote (*Works*, VIII, pp. 278–9): ‘Political Economy you think is an inquiry into the nature and causes of wealth; I think it should rather be called an inquiry into the laws which determine the division of the produce of industry among the classes which occur in its formation.’ What was of crucial significance in this connection was the rate of profits because of its connection with accumulation, both as the source of investment funds and as the stimulus to further investment.

Having ‘got rid of rent’ as the difference between the product on marginal land and that on intra-marginal units, the Ricardian analysis focused on profits as the residual component of the surplus. Under the simplifying conditions on which the analysis was constructed, there emerged a very clear and simple relationship between the wage rate and the overall rate of profits, determined within a single sector of the economy – the corn-producing sector. The special feature of corn as a commodity was that it could serve both as capital good (seed corn) in its own production and as wage good to be advanced to the workers. With the wage rate fixed in terms of corn, the rate of profit in corn production is uniquely determined as the ratio of net output of corn per man minus the wage to the sum of capital per man consisting of seed corn and the fund of corn as wage good. Competition ensures that the same rate of profit enters into the price of all other commodities that are produced with indirect labour. The overall rate of profits, determined in this way, varies inversely with the corn wage. But, as soon as it is recognized that the wage and/or the capital goods employed in corn production consist of other commodities besides corn, the rate of profits can no longer be determined in this way. For the magnitude of the wage and of the total capital then depends on the prices of those commodities, and these prices incorporate the rate of profit. Attention then has to be directed to explaining the rate of profit by taking account of the whole system of prices. For this purpose the theory of value is called upon to provide a solution and Ricardo struggled with this problem until the end of his life. An elegant solution has been worked out by Sraffa (1960) which shows that, in a system of many produced commodities, with the real wage rate given at a specified level, the rate of profit is determined by the given wage and the conditions of production of the commodities that are ‘basics’. It so happens that Ricardo's case of corn is just such a ‘basic’ commodity in the strict sense that it enters directly and indirectly into the production of every commodity including itself.

The core idea that competition among firms under capitalist conditions tends to produce uniformity of profit rates across all markets remains problematical, especially in the dynamic real-world context of changing technology with various forms of factor immobility and barriers to entry (Harris, 1988).

Given the perceived centrality of the rate of profit in a capitalist economy, for classical political economy it becomes a crucial problem in the theory of economic growth to account for movements in the rate of profit associated with the process of capital accumulation and development of the economy. Such movements are a decisive reference point for understanding the long-term evolution of the economy. The classical answer to this problem, as worked out most coherently by Ricardo, is that in a closed economy there is an inevitable tendency for the rate of profit to fall in the course of the accumulation process and, hence, that the accumulation process itself is brought to a halt by its own logic.

Marx was later to propose this falling tendency of the rate of profit (FTRP) as a *law*. He considered it to

be ‘the most important law of modern political economy’ (1973, p. 748; 1894, part 3). He was, of course, following in the tradition of the classical economists in which the same idea had been firmly entrenched, though supported on different grounds. But, interestingly enough, it is also the case that there exists a distinct conception of a FTRP within neoclassical theory (Harris, 1978, ch. 9; 1981). In Keynes, as well, the idea is embodied in his projection of the long-term prospects for capitalism resulting in the ‘euthanasia of the rentier’ (1936, pp. 375–6). In Schumpeter (1934), it occurs in the form of the idea that the profitability of innovations tends inevitably to be eroded so that the economy settles back to the conditions of the ‘circular flow’ in the absence of new innovations. Though it is based in each case on quite different foundations, this conception is one of the most striking and persistent uniformities across different schools of economic thought. (For a discussion of the long history of the idea of a falling rate of profit, see Tucker, 1960.)

A model of accumulation

The essential features of the classical argument regarding the accumulation process can be exhibited with a simple model adapted from Kaldor (1956) and Pasinetti (1960). This model formalizes the Ricardian conception of an agricultural economy producing a single product, ‘corn’, under capitalist conditions. Land is of differing fertility and labour is applied in fixed proportion to less and less fertile land. Accordingly, the average and marginal product of labour falls as the margin of cultivation is extended through capital accumulation and increase of employment on the land. The system may indifferently be assumed to expand on the extensive or intensive margins of available land. Also, it does not matter for this analysis that there exists any production outside agriculture. It would turn out, in any case, that the overall average rate of profit for the economy as a whole is determined by the agricultural rate of profit or, in the general case, by the conditions of production of ‘basics’ (see Sraffa, 1960; Pasinetti, 1977). Of course, in a system with many produced commodities, it is not possible to define ‘less fertile land’ independently of the rate of profit (Sraffa, 1960). However, the problem does not arise in this simplified model of a corn-producing economy. We deliberately abstract from complications associated with the Malthusian population dynamics. This is perhaps the most problematic feature of the classical conception and we return to it below. Meanwhile, it is simply assumed, as in Lewis (1954), that a labour force is in perfectly elastic supply at some conventionally fixed real wage rate equal to ‘subsistence’.

Let the production function relating output Y to labour input L be

$$Y = F(L)\ F(0) \geqq 0\ F' > w^* > 0\ F'' < 0 \tag{1}$$

which satisfies the law of diminishing returns and allows for the existence of a surplus product above the ‘subsistence’ wage-rate w^*. Total capital K consists entirely of wages W (the ‘wage fund’) advanced at the beginning of the production period to hire labour. Thus

$$K = W = wL. \tag{2}$$

We are here, for simplicity, neglecting capital as seed-corn, and inputs of fixed capital are ignored. Total output is distributed between payment of rent *R* to landlords, profits *P* to capitalists, and replacement of the wage fund:

$$Y = R + P + W. \tag{3}$$

Given the margin of cultivation reached at any time, the level of land rent is determined as the difference between the average and marginal product of labour at the prevailing level of employment:

$$R = \left(\frac{F(L)}{L} - F'\right)L. \tag{4}$$

Profit emerges as the residual

$$P = \left(F' - w^*\right)L. \tag{5}$$

It follows that the rate of profit *r* is determined from

$$r = \frac{P}{W} = \frac{F'}{w^*} - 1. \tag{6}$$

It is the dynamics of the wage fund which represents the process of accumulation in this model. Accumulation of capital consists of the growth of the wage fund with a corresponding increase of employment. Additions to the wage fund come entirely from investment of capitalists' profits since the spendthrift landlords consume their share of the surplus. If the capitalists invest a proportion of profits equal to α, then

$$\Delta W = \alpha P Q < \alpha < 1. \tag{7}$$

The proportion α need not be a constant. It could vary in a manner dependent on the rate of profit as suggested by Ricardo's idea that

> [the capitalists'] motive for accumulation will diminish with every diminution of profit, and will cease altogether when their profits are so low as not to afford them an adequate compensation for their trouble and the risk which they must necessarily encounter in employing their capital productively. (*Works*, I, p. 122)

In that case we have

$$\alpha = \alpha(r) \quad \alpha' > 0 \alpha\left(r^*\right) = 0 \tag{8}$$

where r^* is the capitalists' minimum acceptable rate of profit. By definition the rate of capital accumulation is $g = \Delta W / W$, and from (6), (7), and (8) it follows that

$$g = \alpha(r) \cdot r. \tag{9}$$

Thus, the rate of accumulation is uniquely dependent on the profit rate.
The movement in the profit rate as accumulation proceeds can be derived from (6). Evidently, as employment increases the marginal product of labour falls. The rate of profit must therefore fall. It continues to fall as long as there is any increment to the wage fund so as to employ extra labour on the available land. The process comes to a halt when the profit rate is so low that accumulation ceases. The

economy is then at the stationary state.

In this model, the capitalists are caught between, on the one hand, the diminishing productivity of labour as the margin of cultivation is extended and, on the other, the need to pay the ongoing wage rate in order to secure labour for employment. As the productivity of labour falls on the marginal land the pressure of land rent increases for the existing intra-marginal units. The capitalists must therefore pay out an increasing share of the surplus to the landlords. In this way they gradually lose command over the investible surplus of the economy to the landlord class. This distributional conflict between the landlord class and the capitalists constitutes a central feature of the process that drives the economy towards its ultimate stationarity. The impenetrable barrier in the process is the diminishing fertility of the soil. More generally, it is the limitation of natural resources, in this case land, which brings the process to a halt. In this respect the classical model is a particular case of resource-limited growth. Any other limited resource would have the same effect, through increasing ‘rents’ for that resource. At the same time, this consequence is also the product of the capitalists' own actions in relentlessly seeking to expand the size of their capital.

The underlying dynamic process which expresses this conflictive evolution of capitalist accumulation has usually been assumed in the literature to converge towards the stationary state (see Pasinetti, 1960; Samuelson, 1978). Some reservation on this question of convergence was originally expressed by Hicks and Hollander (1977) and followed up by Gordon (1983). Subsequent discussion by Casarosa (1978), Caravale and Tosato (1980) and Caravale (1985) further emphasized the problematic nature of the convergence process. Much of the complexity of this process arises from the intertwined dynamics of distributional change and population growth typical of the Ricardian system. Day (1983) has shown that characterization of the population dynamics by itself may be sufficient to generate extremely erratic or ‘chaotic’ motions. Bhaduri and Harris (1987) analyse the essential dynamics of the Ricardian system as it is governed solely by the interplay of distribution and accumulation in a model similar to the present one. They find that the model can generate very complex ‘chaotic’ movements instead of any smooth and gradual convergence to the stationary state. The possibility of such behaviour is shown to depend uniquely on the initial configuration of parameters. This result should lead one to question the presumption that the Ricardian system necessarily converges to a stationary state.

The Malthusian population dynamics

A crucial role is played in the classical analysis by the population dynamics deriving from the Malthusian law of population growth. In particular this law requires that population grows in response to a rise of wages above subsistence. This response mechanism is supposed to provide the labour requirements for expansion and thereby hold wages in check. But this is evidently a highly implausible principle on which to base an account of the process of capitalist expansion. If capitalism had to depend for its labour supply entirely upon such a demographic–biological response, it seems doubtful that sustained high rates of accumulation could continue for long or even that accumulation could ever get started. This is because, first, there must exist a biological upper limit to population expansion. Accumulation at rates above this limit would drive up the wage to such a level as to reduce or perhaps choke off the possibility of continued accumulation. For the classical labour supply principle to work, it must be presumed arbitrarily that this limit is sufficiently far out or, equivalently, that the supply curve is

sufficiently elastic over a wide range.

Even if it is granted that population growth is significantly responsive to the level of wages, it is still the case that the adjustment of population is inherently a long drawn-out process having only a negligible effect on the actual labour supply in any short period of time. In the interim, any sizeable spurt of accumulation must then cause wages to be bid up, eat into profits, and bring accumulation itself, to a halt. From the start, therefore, accumulation could never get going in such a system. Even if it did, its continuation would always be in jeopardy because the mechanism of adjustment of labour supply is an inherently unreliable one, fraught with the possibility that at any time wages may rise to eat up the profits that are the well-spring of accumulation.

This feature of classical analysis was soundly criticized and rejected by Marx (1867, pp. 637–9). In its place, he sought to introduce a principle that was internal to the accumulation process, which would account for the continuing generation of a supply of labour to meet the needs of accumulation from within the accumulation process itself. This was the principle of the reserve army of labour or the ‘law of relative surplus population’ (1897, ch. 25, sections 3 and 4). The reserve army results from a process of ‘recycling’ of labour through its displacement from existing employment due to mechanization and structural changes in production. In addition to this pool of labour there are other possible sources of increased labour supply to feed the accumulation process. These originate, for instance, in increased labour force participation rates among existing workers, in labour migration, and in the erosion of household work and other forms of non-capitalist production. Capital export to other regions can play the same role. These sources have been observed historically to be more or less significant at various times and places. It appears, therefore, that there is considerable flexibility of labour supply, and hence of accumulation, even without taking account of population growth. The existence of population growth certainly adds to the pool of available labour, as is now widely recognized. But the singular and unique role attributed to it by the Malthusian theory has by now been discredited and abandoned.

Conclusion

The classical economists are often regarded as ‘pessimistic’ in their prognosis for economic growth. It is said that they constituted economics as the ‘dismal science’. Still, there is much to be learned, that is of contemporary relevance, from a close examination of their analytical system. What emerges from such an examination is a complex structure of ideas expressing a deep understanding of the nature of capitalism as an economic system, the sources of its expansionary drive, and the barriers or limits to its expansion. Their ideas were essentially limited, however, to the conditions of a predominantly agrarian economy, without significant change in methods of production, in which, because of the limited quantity and diminishing fertility of the soil, growth is arrested by increasing costs of production of agricultural commodities. Their analysis underestimated the far-reaching character of technological change as a powerful and continuing force in transforming the conditions of productivity both in agriculture and in industry. While they clearly perceived the possibilities opened up by international trade and foreign investment, they failed to incorporate these elements as integral components of a systematic theory of the growth process. It remained for Marx to pinpoint some of the major limitations and deficiencies of the classical analysis and to develop an analysis of the capitalist accumulation process that went beyond that of the classical economists in many respects while also leaving many unresolved questions. Subsequent work has continued to address the issues with limited success. Still today, the theory of

growth of capitalist economies continues to be one of the most fascinating and still unresolved areas of economic theory.

See Also

- development economics
- profit and profit theory
- Ricardo, David
- surplus

Bibliography

Bhaduri, A. and Harris, D.J. 1987. The complex dynamics of the simple Ricardian system. *Quarterly Journal of Economics* 102, 893–901.

Caravale, G.A., ed. 1985. *The Legacy of Ricardo*. Oxford: Blackwell.

Caravale, G.A. and Tosato, D.A. 1980. *Ricardo and the Theory of Value, Distribution and Growth*. London: Routledge & Kegan Paul.

Casarosa, C. 1978. A new formulation of the Ricardian system. *Oxford Economic Papers* 30, 38–63.

Day, R.H. 1983. The emergence of chaos from classical economic growth. *Quarterly Journal of Economics* 98, 201–13.

Gordon, K. 1983. Hicks and Hollander on Ricardo: a mathematical note. *Quarterly Journal of Economics* 98, 721–6.

Harris, D.J. 1978. *Capital Accumulation and Income Distribution*. Stanford: Stanford University Press.

Harris, D.J. 1981. Profits, productivity, and thrift: the neoclassical theory of capital and distribution revisited. *Journal of Post Keynesian Economics* 3, 359–82.

Harris, D.J. 1988. On the Classical theory of competition. *Cambridge Journal of Economics* 12, 139–67.

Hicks, J.R. and Hollander, S. 1977. Mr. Ricardo and the moderns. *Quarterly Journal of Economics* 91, 351–69.

Kaldor, N. 1956. Alternative theories of distribution. *Review of Economic Studies* 23, 83–100.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. New York: Harcourt, Brace.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *The Manchester School* 22 (2), 139–91.

Malthus, T.R. 1798. *Essay on the Principle of Population*, 1st edn. London: Macmillan, 1926.

Malthus, T.R. 1820. *Principles of Political Economy*. Reprinted in *The Works and Correspondence of David Ricardo*, ed. P. Sraffa and M. Dobb, vol. II, Cambridge: Cambridge University Press, 1951.

Marx, K. 1867. *Capital*, vol. 1. New York: International Publishers, 1967.

Marx, K. 1894. *Capital*, vol. 3. New York: International Publishers, 1967.

Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin Books.

Meek, R.L. 1967. *Economics and Ideology and Other Essays*. London: Chapman & Hall.

Pasinetti, L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies* 27 (2), 78–98.

Pasinetti, L. 1977. *Lectures on Theory of Production*. New York: Columbia University Press.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*, vols I–XI, ed. P. Sraffa with the collaboration of M.H. Dobb. Cambridge: Cambridge University Press.

Samuelson, P. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16, 1415–34.

Schumpeter, J. 1934. *The Theory of Economic Development*. New York: Oxford University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. New York: Modern Library, 1937.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Tucker, G. 1960. *Progress and Profits in British Economic Thought 1650–1850*. Cambridge: Cambridge University Press.

How to cite this article

Harris, Donald J. "classical growth model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000595> doi:10.1057/9780230226203.0245

The New Palgrave Dictionary of Economics Online

classical production theories

Giorgio Gilibert
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advances; agriculture; Babbage, C.; capital accumulation; circular flow; classical economics; classical production theories; division of labour; factory system; increasing returns; Industrial Revolution; marginal revolution; Marx, K. H.; net product; Physiocracy; production; productive and unproductive inputs; Quesnay, F.; rent; Ricardo, D.; Smith, A.; stationary state; surplus; technical change; Tugan-Baranovsky, M. I.; Ure, W.

Article

A theory of production cannot be said to have existed before the middle of the 18th century. The very word production was previously used in its narrow etymological sense (from the Latin *producere*, to bring forth) of giving birth to new material objects; and it was therefore normally confined to the fruits of the earth. ‘When we speak of it’, writes Daniel Defoe, ‘as the Effect of Nature ‘tis *Product* or Produce; when as the Effect of Labour ‘tis *Manufacture*’ (1728, p. 1).

It is with the writings of the French *économistes* that the term receives a precise technical meaning. At first sight, the Physiocratic terminology is not particularly novel: the words production, productivity, and so on are carefully reserved for agriculture; manufacture, as a mere transforming activity, is considered as eminently sterile. But Quesnay's fundamental innovation lies in the theory behind the terminology: it is not (or not so much) because of some physical property that agriculture is said to be productive, but because it is the only activity capable of generating a net revenue (rent). The way was, however, paved for the recognition of the productivity of non-agricultural activities, provided that the peculiar assumption of rent as the only possible net revenue was dropped (that is, that profit was accepted as a legitimate form of net revenue). This step was taken, a few years later, by Adam Smith.

In the following decades, production became one of the main topics of political economy; this was later sanctioned by the standard structure adopted by economic textbooks, whose first section typically came to be devoted to production. The first English example of this arrangement is given by the *Elements of Political Economy* published by James Mill in 1821 (following in this respect in Say's steps), the same

year in which Robert Torrens brought out his *Essay on the Production of Wealth*. Eventually, in Marxian economics, production analysis achieved the status of a cornerstone of the whole theory of social change.
In the second half of the 19th century, as a consequence of the so-called marginalist revolution, the focus of economic theory tended to shift from the sphere of production to that of exchange. Production theory was squeezed into the general framework of the optimal allocation of scarce resources: a framework originally developed to deal with the problem of pure exchange. The theory originally devised by Quesnay seemed, about one century after its birth, to conclude its own theoretical lifetime.
François Quesnay was the first to analyse the system of production and consumption as a single complex process. He looked for the ‘natural laws’ by which it was regulated, laws which were independent of the will of man but discoverable by the light of reason. The attempt to present the interplay of these laws in an abstract and manageable way originated the first theoretical model of the history of economic analysis.
The Physiocratic doctrine presents, though often under a misleading feudal disguise, most of the leading ideas of the classical theory of capitalist production. First and foremost, the picture of the system of production and consumption as a circular process: no one will ever deny that consumption is the ultimate end of production, but it is essential to bear in mind the simple fact that past production determines present consumption, and that consumption in turn is nothing but the necessary condition for future production.
The idea of production as a circular process immediately suggests the notion of surplus: if the economy produces more than the minimum necessary for the process to be repeated, then there is a surplus. Its value was called ‘net product’ by Quesnay: this is the strategic variable for economic activity. The nations’ prosperity can be assessed according to the size of their annual net product.
The answers given by the Physiocrats to the fundamental questions of the origins and destination of the net product account for their peculiar class analysis. They assumed that a net product was yielded exclusively by land-using activities; that is, that revenues could be higher than costs only in agriculture, and therefore rent was the only conceivable net revenue. The class of those engaged in agriculture (farmers, the labourers being equated to cattle) was thus called ‘productive’, in contrast to the ‘sterile’ class of those engaged in manufacture (artisans); the class of landowners got the whole net product, under the form of rent.
Since the process of production takes time (the agricultural year) it requires advances: for instance, the labourers’ subsistence must be available before the harvest. Quesnay distinguishes between annual advances (working capital: seed, subsistence), which are wholly used up in the course of the production process, and original advances (fixed capital, for which a depreciation is allowed), which are not. It is perhaps worth noticing that the word capital was commonly in use in the economic literature of the 18th century. Quesnay's unusual terminology was presumably due to his intention of stressing the physical nature of the advances required by the production process, as opposed to the current meaning of capital as a sum of money employed in trade.
The characteristic agricultural bias of the Physiocrats is shown not only by their doctrine of the sterility of manufacture, but also by the essentially static nature of their models. If the economy is organized according to the natural order, that is according to the ‘evident’ laws discovered by the economists, it will rapidly attain the maximum level of output consistent with the country's amount of arable land and with the state of technology. Indeed, the *Tableaux* depict this prosperous and stationary situation.
Both these aspects are definitely abandoned by Adam Smith. Precisely because production takes time,

and wages, materials and equipment have to be anticipated, the owners of these advances, the capitalists, are naturally entitled to a part of the net revenue, the profits. The advances are consumed by productive workers or used up as raw materials and wear and tear of equipment; the return, in manufacture as well as in agriculture, will normally cover their cost with an addition, which constitutes the profit.

The Smithian capitalist is thrifty and industrious; his profits are well above subsistence, and he will normally save most of them and employ these savings as capital, in order to obtain an additional profit in the future. As a result of these decisions, the capital of the nation as a whole, the fund that sets productive labour to work for the purpose of profit, naturally tends to increase each year in the course of economic progress.

In this way, Smith gave a clear-cut answer to an old dilemma. In his century, two traditional ideas coexisted unreconciled side by side: on the one hand, by analogy with the behaviour of a good husband, thriftiness was praised as a social virtue; on the other hand, it was maintained that a buoyant consumption stimulated trade. In Smith's view, every frugal man is a benefactor, every prodigal man a ‘public enemy’.

The progressive state of the economy – it is written in the *Wealth of Nations* – ‘is in reality the chearful and the hearty state to all the different orders of the society. The stationary is dull; the declining, melancholy’ (Smith, 1776, p. 99). The analysis is here primarily concerned with the process of capital accumulation and is therefore necessarily dynamic.

The analysis of the accumulation of wealth inevitably involved the question of the final outcome of the process. It was a common belief – among classical economists – that the economy would eventually tend towards a stationary state. This could be seen optimistically as ‘a full complement of riches’ (Smith) or, on the contrary, as a sad motionless state (Ricardo); still, it could be considered as relatively far ahead in the future (Smith and, with a suitable economic policy, Ricardo) or just round the corner (J.S. Mill).

An interesting technical feature of the theory of production can be introduced in connection with this question. The advances of every industry are normally composed of commodities that are not produced by that industry. In other words, each industry must exchange part of its output on the market with the necessary inputs to start the production process again. These transactions, dictated by the technology in use, were clearly described by the *Tableau:* in this highly aggregate picture, the two activities considered, productive and sterile, are both essential to reproduction. But, in a more detailed framework, we can distinguish between those commodities which play a productive role as inputs, and those which do not (‘luxuries’). The growth potential of the economy is affected only by the conditions of production of the first type of commodities (‘basics’ according to modern terminology).

Since every line of production requires labour, and workers consume food, foodstuffs are basics par excellence. Food production in turn requires land, a non-reproducible resource; the scarcity of land becomes therefore the limiting factor to accumulation. Land is essential, and is fixed in supply, so the eventual outcome of the growth process is the stationary state. (One might think that in this way we are back with the original Physiocratic perspective, but now attention is focused on the dynamic process rather than on its static outcome.)

David Ricardo presented a sophisticated version of this argument, in which the result that the growth process ends in a stationary state is analytically restated via his theory of profits. In evaluating this kind of argument, one must remember the vital *ceteris paribus* assumption, especially with regard to technology. Of course, the process of exhaustion of natural resources can be checked by improvements which affect agriculture. Ricardo has often been criticized for his allegedly cursory treatment of

technical progress: one instance can be found in *The Logic of Political Economy* written a quarter of a century later (1844) by his follower Thomas de Quincey.

With Karl Marx, the concept of production acquires new and wider meanings; in a sense it leaves the narrow field of economic theory to become the cornerstone of a general theory of social systems and of history (the development of material production, notes Marx in the first book of *Capital*, is) 'the basis of any social life and of any true history'). The starting point of the analysis is the notion of production in its elementary form: men produce the necessaries for their existence; their productive activity is labour, which materializes into products. In other words, men produce the conditions for their material life. What men are is then determined by production; more specifically, by what is produced and by the way in which it is produced.

Production is essentially a social process: there are no 'natural laws' to be investigated, but social relations which are historically determined. These relations constitute the structure of society and determine its material and intellectual way of life. The evolution of religion, ethics, art and government is an ultimate consequences of the evolution of the social relations of production.

In his justly famous preface to the *Critique of Political Economy*, Marx has left a very effective summary statement of this approach:

> In the social production which men carry on they enter into definite relations that are independent of their will; these relations of production correspond to a definite stage of development of their material powers of production. The sum total of these relations of production constitutes the economic structure of society – the real foundation on which rise the legal and political superstructures and to which correspond definite forms of social consciousness. The mode of production in material life determines the general character of the social, political, and spiritual processes of life. It is not the consciousness of men that determines their existence, but, on the contrary, their social existence determines their consciousness. (Marx, 1859, p. 100)

Production, distribution, exchange and consumption cannot be grasped in their essence but as successive moments of a unique circular process, thoroughly determined by the social conditions of production. Marx reproaches political economy for having arbitrarily separated the sphere of production, regulated by allegedly universal laws, from that of distribution, where we can take account of the social environment.

The search for universal laws of production has in turn led the economist to concentrate upon the trivial aspects of the phenomenon and to overlook the questions that are truly essential in investigating the present mode of production. For example, having defined as capital the set of the means of production, and having observed the obvious fact that men have always needed some kind of instrument to produce, the economists are ready to attribute a universal and ahistorical validity to the notion of capital. In this way, they have simply swept aside the key question: what is the socially determined relationship which turns an instrument used in production into 'capital'?

The formation of classical political economy historically coincided with the development of the factory system in manufacture. An early description of an integrated production process is offered by William Petty (1683) with reference to the watch trade. Another obvious reference is the famous pin factory described by Adam Smith in the first chapter of the *Wealth of Nations* (1776). In both cases, the division

of labour is presented as the main virtue of the new form of productive organization: provided that the extent of the market is sufficient, it is maintained that output can be expanded more than proportionately with the labour employed in manufacture (increasing returns to scale).

Marx used these two examples to draw a distinction between the ‘heterogeneous’ manufacture (exemplified by Petty's watch-making activity) in which the final output is obtained by simple assemblage of ‘partial and independent products’, and the more sophisticated ‘organic’ manufacture (exemplified by Smith's pin factory) in which a series of successive operations gradually transforms the original raw material into the finished product.

Smith referred to three arguments in favour of the technical superiority of an ever increasing division of labour:

> first, to the increase of dexterity in every particular workman; secondly, to the saving of the time which is commonly lost in passing from one species of work to another; and lastly, to the invention of a great number of machines which facilitate and abridge labour, and enable one man to do the work of many. (Smith, 1776, p. 17)

It has been observed that these arguments are not truly convincing. The importance attributed to increased dexterity conflicts with the relatively low level of skills required in contemporary factories (witness the common use of child labour). Time saving does not imply specialization by individuals: in principle, it could equally be attained by a suitable reorganization of the activity of a single artisan. And the introduction of machines does not seem to exhibit any necessary relation to the increasing division of tasks.

In fact the new organization of labour associated with the factory system did go along with the process of technical change associated with the Industrial Revolution. But its original role was primarily to discipline the manner in which the work was performed and to give the capitalist the power of controlling the production process in every single detail.

The introduction of machinery came after labour specialization and reinforced the need for a thorough organization of production. The effects of the introduction of the steam-engine and other complex machines were eventually studied by two scholars who possessed the necessary technical background, Charles Babbage (1832) and William Ure (1835); their tracts were very popular at the time and were widely used by the economists (for example, by John Stuart Mill and Marx). They conceived of the control and management of a factory as that of a single complex machine, under the full control of the capitalist and with manual work brought to a minimum.

It is worth noticing that these speculations about the rational management of a highly mechanized factory were easily extended to society as a whole. At the turn of the century, Mikhail Tugan-Baranovsky (1905) dreamed of an economy in which machines were automatically produced by machines, and where the labour force was paradoxically reduced to one worker alone. In a similar vein, especially in Germany after the First World War, we find many suggestions for a ‘rational’ organization of the economy as if it were a giant *Konzern* (as an extreme example, see the ‘natural economy’ proposed by Otto Neurath (1921) for the ephemeral Bavarian republic).

Bibliography

Babbage, C. 1832. *On the Economy of Machine and Manufactures*. London: Knight.

Defoe, D. 1728. *A Plan of the English Commerce*. Oxford: Blackwell, 1928.

De Quincey, T. 1844. The Logic of Political Economy. In *Collected Writings*, vol. 9, ed. D. Masson. London: Black, 1897.

Marx, K. 1859. Zur Kritik der politischen Ökonomie. In *Marx-Engels Gesamtausgabe* vol. 2, pt. II. Berlin: Dietz, 1980.

Marx, K. 1867. *Das Kapital*, vol. 1. In *Marx-Engels Gesamtausgabe*, vol. 2, pt V. Berlin: Dietz, 1983.

Mill, J. 1821. *Elements of Political Economy*. London: Baldwin.

Mill, J.S. 1848. *Principles of Political Economy*. Ed. J.M. Robson, Toronto: University of Toronto Press, 1965.

Neurath, O. 1921. *Durch die Kriegswirtschaft zur Naturalwirtschaft*. Munich: Callway.

Petty, W. 1683. Another Essay on Political Arithmetick. In *Economic Writings of Sir William Petty*, vol. 2, ed. C.H. Hull. Cambridge: Cambridge University Press, 1899.

Quesnay, F. 1759. *Tableau économique*. Ed. M. Kuczynski and R. Meek, London: Macmillan, 1972.

Ricardo, D. 1815. An Essay on the Influence of a Low Price of Corn. In *The Works and Correspondence of David Ricardo*, vol. 4, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1817. Principles of Political Economy. In *The Works and Correspondence of David Ricardo*, vol. 1, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. R.H. Campbell, A. S. Skinner and W.B. Todd, Oxford: Clarendon Press, 1976.

Torrens, R. 1821. *An Essay on the Production of Wealth*. London: Longman, Hurst, Rees, Orme & Brown.

Tugan-Baranovsky, M. 1905. *Theoretische Grundlagen des Marxismus*. German trans., Leipzig: Duncker & Humblot.

Ure, A. 1835. *The Philosophy of Manufactures*. London: Knight.

How to cite this article

Gilibert, Giorgio. "classical production theories." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000205> doi:10.1057/9780230226203.0246

The New Palgrave Dictionary of Economics Online

climate change, economics of

Lawrence H. Goulder and William A. Pizer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Climate-change economics attends to the various threats posed by global climate change by offering theoretical and empirical insights relevant to the design of policies to reduce, avoid, or adapt to such change. This economic analysis has yielded new estimates of mitigation benefits, improved assessments of policy costs in the presence of various market distortions or imperfections, better tools for making policy choices under uncertainty, and alternative mechanisms for allowing flexibility in policy responses. These contributions have influenced the formulation and implementation of a range of climate-change policies at domestic and international levels.

Keywords

carbon emissions tax; climate change, economics of; computable general equilibrium (CGE) models; contingent valuation; discount rate; global warming; hedonic approach; integrated assessment models; intergenerational equity; learning-by-doing; Monte Carlo methods; price-based vs. quantity-based policies; production function approach; technology policy; time preference; tradable emission permits; uncertainty

Article

The prospect of global climate change has emerged as a major scientific and public policy issue. Scientific studies indicate that human-caused increases in atmospheric concentrations of carbon dioxide (largely from fossil-fuel burning) and of other greenhouse gases are leading to warmer global surface temperatures. Possible current-century consequences of this temperature increase include increased frequency of extreme temperature events (such as heat waves), heightened storm intensity, altered precipitation patterns, sea-level rise, and reversal of ocean currents. These changes, in turn, can have significant impacts on the functioning of ecosystems, the viability of wildlife and the well-being of humans.

There is considerable disagreement within and among nations as to what policies, if any, should be introduced to mitigate and perhaps prevent climate change and its various impacts. Despite the disagreements, in recent years we have witnessed the gradual emergence of a range of international and domestic climate-change policies, including emission-trading programmes, emission taxes, performance standards, and technology-promoting programmes.

Beginning with William Nordhaus's ‘How fast should we graze the global commons?’ (1982), climate-change economics has focused on diagnosing the economic underpinnings of climate change and offering positive and normative analyses of policies to confront the problem. While overlapping with other areas of environmental economics, it has a unique focus because of distinctive features of the climate problem – including the long time-scale, the extent and nature of uncertainties, the international scope of the issue, and the uneven distribution of policy benefits and costs across space and time.

In our discussion of the economics of climate change, we begin with a brief account of alternative economic approaches to measuring the benefits and costs associated with reducing greenhouse gas emissions, followed by a discussion of uncertainties and their consequences. We then present issues related to policy design, including instrument choice, flexibility, and international coordination. The final section offers general conclusions.

Assessing the benefits and costs of climate change mitigation

Climate change damages and mitigation benefits

As noted, the potential consequences of climate change include increased average temperatures, greater frequency of extreme temperature events, altered precipitation patterns, and sea-level rise. These biophysical changes affect human welfare. While the distinction is imperfect, economists divide the (often negative) welfare impacts into two main categories: *market* and *non-market* damages.

Market damages. As the name suggests, market damages are the welfare impacts stemming from changes in prices or quantities of marketed goods. Changes in productivity typically underlie these impacts. Often researchers have employed climate-dependent production functions to model these changes, specifying wheat production, for example, as a function of climate variables such as temperature and precipitation. In addition to agriculture, this approach has been applied in other industries including forestry, energy services, water utilities and coastal flooding from sea-level rise (see, for example, Smith and Tirpak, 1989; Yohe et al., 1996; Mansur, Mendelsohn and Morrison, 2005).

The production function approach tends to ignore possibilities for substitution across products, which motivates an alternative, hedonic approach (see, for example, Mendelsohn, Nordhaus and Shaw, 1994; Schlenker, Fisher and Hanemann, 2005). Applied to agriculture, the hedonic approach aims to embrace a wider range of substitution options, employing cross-section data to examine how geographical, physical, and climate variables are related to the prices of agricultural land. On the assumption that crops are chosen to maximize rents, that rents reflect the productivity of a given plot of land relative to that of marginal land, and that land prices are the present value of land rents, the impact of climate variables on land prices is an indicator of their impact on productivity after crop-substitution is allowed for.

Non-market damages. Non-market damages include the direct utility loss stemming from a less hospitable climate, as well as welfare costs attributable to lost ecosystem services or lost biodiversity.

For these damages, revealed-preference methods face major challenges because non-market impacts may not leave a ‘behavioural trail’ of induced changes in prices or quantities that can be used to determine welfare changes. The loss of biodiversity, for example, does not have any obvious connection with price changes or observable demands. Partly because of the difficulties of revealed-preference approaches in this context, researchers often employ stated-preference or interview techniques – most notably the contingent valuation method – to assess the willingness to pay to avoid non-market damages (see, for example, Smith, 2004).

Cost assessment

The costs of avoiding emissions of carbon dioxide, the principal greenhouse gas, depend on substitution possibilities on several margins: the ability to substitute across different fuels (which release different amounts of carbon dioxide per unit of energy), to substitute away from energy in general in production, and to shift away from energy-intensive goods. The greater the potential for substitution, the lower are the costs of meeting a given emission-reduction target.

Applied models have taken two main approaches to assessing substitution options and costs. One approach employs ‘bottom-up’ energy technology models with considerable detail on the technologies of specific energy processes or products (for example, Barretto and Kypreos, 2004). The models tend to concentrate on one sector or a small group of sectors, and offer less information on abilities to substitute from energy in general or on how changes in the prices of energy-intensive goods affect intermediate and final demands for those goods.

The other approach employs ‘top down’ economy-wide models, which include, but are not limited to, computable general equilibrium (CGE) models (see, for example, Jorgenson and Wilcoxen, 1996; Conrad, 2002). An attraction of these models is their ability to trace relationships between fuel costs, production methods, and consumer choices throughout the economy in an internally consistent way. However, they tend to include much less detail on specific energy processes or products. Substitution across fuels is generally captured through smooth production functions rather than through explicit attention to alternative discrete processes. In recent years, attempts have been made to reduce the gap between the two types of models. Bottom-up models have gained scope, and top-down models have incorporated greater detail (see, for example, McFarland, Reilly and Herzog, 2004).

Because climate depends on the atmospheric stock of greenhouse gases, and because for most gases the residence times in the atmosphere are hundreds (and in some cases, thousands) of years, climate change is an inherently long-term problem and assumptions about technological change are particularly important. The modelling of technological change has advanced significantly beyond the early tradition that treated technological change as exogenous. Several recent models allow the rate or direction of technological progress to respond endogenously to policy interventions. Some models focus on *R&D-based* technological change, incorporating connections between policy interventions, incentives to research and development, and advances in knowledge (see, for example, Goulder and Schneider, 1999; Nordhaus, 2002; Buonanno, Carraro and Galeotti, 2003; Popp, 2004). Others emphasize *learning-by-doing-based* technological change where production cost falls with cumulative output, in keeping with the idea that cumulative output is associated with learning (for example, Manne and Richels, 2004). Allowing for policy-induced technological change tends to yield lower (and sometimes significantly

lower) assessments of the costs of reaching given emission-reduction targets than do models in which technological change is exogenous.

Integrated assessment

While the cost models described above are useful for evaluating the cost-effectiveness of alternative policies to achieve a given emissions target, the desire to relate costs to mitigation benefits (avoided damages) has spawned the development of *integrated assessment models*. These models link greenhouse gas emissions, greenhouse gas concentrations, and changes in temperature or precipitation, and they consider how these changes feed back on production and utility. Many of the integrated assessment models are optimization models that solve for the emissions time-path that maximizes net benefits, in some cases under constraints on temperature or concentration (see, for example, Nordhaus, 1994).

Dealing with uncertainty

The uncertainties about both the costs and the benefits from reduced climate change are vast. In a recent meta-analysis examining 28 studies' estimated benefits from reduced climate change (Tol, 2005), the 90 per cent confidence interval for the benefit estimates ranged from – $10 to +$350 per ton of carbon, with a mode of $1.50 per ton. On the cost side, a separate study found marginal costs of between $10 and $212 per ton of carbon for a ten per cent reduction in 2010 (Weyant and Hill, 1999).

Uncertainty and the stringency of climate policy

Increasingly sophisticated numerical models have attempted to deal explicitly with these substantial uncertainties regarding costs and benefits. Some provide an uncertainty analysis using Monte Carlo simulation, in which the model is solved repeatedly, each time using a different set of parameter values that are randomly drawn from pre-assigned probability distributions. This approach produces a probability distribution for policy outcomes that sheds light on appropriate policy design in the face of uncertainty. Other models incorporate uncertainty more directly by explicitly optimizing over uncertain outcomes. These models typically call for a more aggressive climate policy than would emerge from a deterministic analysis. Nordhaus (1994) employs an integrated climate-economy model to compare the optimal carbon tax in a framework with uncertain parameter values with the optimal tax when parameters are set at their central values. In this application, an uncertainty premium arises: the optimal tax is more than twice as high in the former case as in the latter, and the optimal amount of abatement is correspondingly much greater. The higher optimal tax could in principle be due to uncertainty about any parameter whose relationship with damages is convex, thus yielding large downside risks relative to upside risks. In the Nordhaus model, the higher optimal tax stems primarily from uncertainty about the discount rate (Pizer, 1999).

The choice of discount rate under uncertainty

The importance of the discount rate arises because greenhouse gases persist in the atmosphere for a

century or more, and therefore mitigation benefits must be measured on dramatically different timescales from those of ordinary environmental problems. A prescriptive approach links the discount rate to subjective judgements about intergenerational equity as indicated by a pure social rate of time preference (see, for example, Arrow et al., 1996). A descriptive approach relates the discount rate to future market interest rates. Under both approaches, significant uncertainties surround the discount rates. Recent work by Weitzman (1998) points out that a rate lower than the expected value should be employed in the presence of such uncertainty, a reflection of the relationships among the discount *factor*, the discount *rate*, and the time interval over which discounting applies. Put simply, the discount factor e^{-rt} is an increasingly convex function of the interest rate r as the period of discounting t increases. This implies that in the presence of uncertainty the certainty-equivalent discount rate is lower than the expected value of the discount rate: that is, $\ln(E[e^{-rt}])/t < E[r]$. The difference between the appropriate, certainty-equivalent rate and the expected value of the discount rate widens the longer the time horizon is. While Weitzman focuses on a single uncertain rate, Newell and Pizer (2003a) show that, under reasonable specifications of uncertainty about the evolution of future market rates, this approach doubles the expected marginal benefits from future climate change mitigation compared with the estimated benefits from an analysis that uses only the current rate.

Act today or wait for better information?

In addition to concerns about convexity and valuation, uncertainty raises important questions about whether and how much to embark on mitigation activities now as opposed to waiting until at least some uncertainty is resolved. Economic theory suggests that, in the absence of fixed costs and irreversibilities, society should mitigate (today) to the point where expected marginal costs and benefits are equal. Yet climate change inherently involves fixed costs and irreversible decisions both on the cost side, in terms of investments in carbon-free technologies, and on the benefit side, in terms of accumulated emissions. These features can lead to more intensive action or to inaction, depending on the magnitude of their respective sunk values (Pindyck, 2000). Despite the ambiguous theory, empirically calibrated analytical and numerical models tend to recommend initiating reductions in emissions in the present, reflecting initially negligible marginal cost and non-negligible environmental benefits (Manne and Richels, 2004; Kolstad, 1996).

The choice of instrument for climate-change policy

Policymakers can consider a range of potential instruments for promoting reductions in emissions of greenhouse gases. Alternatives include emissions taxes, abatement subsidies, emission quotas, tradable emission allowances, and performance standards. Policymakers also can choose whether to apply a given instrument to emissions directly (as with an emission-trading programme) or instead to pollution-related goods or services (as with a fuel tax or technology subsidy).
Initial economic analyses of climate-change policy tended to focus on a carbon tax because it was relatively easy to model and implement. This is a tax on fossil fuels – oil, coal, and natural gas – in proportion to their carbon content. Because combustion of fossil fuels or their refined fuel products leads

to carbon dioxide (CO_2) emissions proportional to carbon content, a carbon tax is effectively a tax on CO_2 emissions. In the simplest analysis, a carbon tax set equal to the marginal climate-related damage from carbon combustion would be efficiency-maximizing. However, in more complex analyses – where additional dimensions such as uncertainty, other market failures, and distributional impacts are taken into account – the superiority of such a carbon tax is no longer assured. We now consider these other dimensions and their implications for instrument choice.

Prices (taxes) vs. quantities (tradable allowances) in the presence of uncertainty

Theoretical and empirical work by Kolstad (1996) and Newell and Pizer (2003b) suggests that the marginal benefit (avoided damage) schedule for emissions reductions is relatively flat. Weitzman's (1974) seminal analysis indicates that under these circumstances expected welfare losses are smaller from a price-based instrument like a carbon tax than from a quantity-based instrument like emission quotas or a system of tradable emission allowances. That is, it is preferable to let levels of emissions remain uncertain (which is the result under a tax) than to let the marginal price of emission reductions remain uncertain (which is the result under a quota). Despite these economic welfare arguments, and recent work on hybrid approaches (Pizer, 2002), many environmental advocates prefer the quantity-based approach precisely because it removes uncertainty about the level of emissions.

Fiscal impacts and instrument choice

A second issue stems from the potential for policies such as carbon taxes and auctioned permits to generate revenues. A number of studies show that using such revenues to finance reductions in pre-existing distortionary taxes on income, sales, or payroll can achieve given environmental targets at lower cost – perhaps substantially lower cost – than other policies (see, for example, Goulder et al., 1999; Parry, Williams and Goulder, 1999; Parry and Oates, 2000). Therefore, carbon taxes and auctioned permit programmes that employ their revenues this way will lower the excess burden from prior taxes, giving them a significant cost-advantage. Correspondingly, subsidies to emission reductions or to new, ‘clean’ technologies will have a cost disadvantage associated with the need to raise distortionary taxes to finance these policies.

Distributional considerations

Despite these attractions of revenue-raising policies such as carbon taxes and auctioned tradable allowance systems, trading programmes with freely distributed permits have achieved greater popularity among policymakers. In New Zealand, for example, industry opposition led the government to drop its proposed carbon tax in 2005. At the same time, the European Union has, and Canada is planning, trading programmes where tradable permits are freely distributed, in line with virtually all conventional pollution trading programmes in the United States.

The politics may reflect differences between systems of freely allocated allowances and systems with auctioned allowances (or carbon taxes) in terms of the distribution of the regulatory burden. Under both types of emission-permit system, profit-maximizing firms will find it in their interest to raise output

prices based on the new, non-zero cost associated with carbon emissions. If the allowances are given out free, firms can retain rents associated with the higher output prices, and this offsets other compliance costs. In contrast, if the allowances are auctioned, firms do not capture these rents. Thus, firms bear a considerably smaller share of the regulatory burden in the case of freely allocated permits. Indeed, Bovenberg and Goulder (2001) show that freely allocating all carbon permits to US fossil fuel suppliers generally will cause those firms to enjoy *higher* profits than in the absence of a permit system; and freely allocating less than a fifth of the permits may be sufficient to keep profits from falling. These considerations reveal a potential trade-off between efficiency and political feasibility: the revenue-raising policies (taxes and auctioned permits) are the most cost-effective, while the non-revenue-raising policies (freely distributed permits) have distributional consequences that may reduce political resistance.

Emissions instruments vs. technology instruments

As noted in the cost discussion, the long-term nature of the climate-change problem makes technological change a central issue in policy considerations. Economic analysis suggests that both ‘direct emissions policies’ and ‘technology-push policies’ are justified on efficiency grounds to correct two distinct market failures. Direct emissions policies (emission trading or taxes) gain support from the fact that combustion of fossil fuels and other greenhouse-gas-producing activities generate negative externalities in the form of climate change-related damages. Technology-push policies (technology and R&D incentives) gain support from the fact that not all of the social benefits from the invention of a new technology can be appropriated by the inventor. The latter argument applies to research and development more generally, and is especially salient if the first market failure is not fully corrected (Fischer, 2004a). Numerical assessments reveal substantial cost-savings from combining the two types of policy (Fischer and Newell, 2005; Schneider and Goulder, 1997).

Policy designs to enhance flexibility

The previous discussion indicates that no single instrument is best along all important policy dimensions, including cost uncertainty, fiscal interactions, distribution and technology development. A further issue in policy choice is how to give regulated firms or nations the flexibility to seek out mitigation opportunities wherever and whenever they are cheapest. For both price- and quantity-based policies, flexibility is enhanced through broad coverage: specifically, by including in the programme as many emissions sources as possible and by providing opportunities for regulated sources to offset their obligations through relevant activities outside the programme. For quantity-based programmes, flexibility can also be promoted through provisions allowing trading of allowances across gases, time, and national boundaries. Such flexibility is automatically provided by price-based programmes simply because they involve no quantitative emissions limits. Importantly, as quantity-based programmes provide these additional dimensions of flexibility, they reduce the efficiency arguments for price-based policies in the face of uncertainty voiced in the preceding section by providing opportunities to adjust to idiosyncratic cost shocks across time, space and industry (Jacoby and Ellerman, 2004).

Flexibility over gases and sequestration

So far we have focused almost exclusively on emissions of carbon dioxide from the burning of fossil fuels as both the cause of human-induced climate change and the object of any mitigation policy. Yet emissions of a number of other gases (as well as non-energy-related emissions of carbon dioxide) contribute to the problem and possibly the solution, particularly in the short run. Models suggest that half of the reductions achievable at costs of $5–$10 per ton of carbon dioxide equivalent arise from gases other than carbon dioxide. In addition, carbon sequestration can be part of the solution. Biological sequestration (for example, through afforestation) has been cited as a particularly inexpensive response to climate change (Sedjo, 1995; Richards and Stavins, 2005). Geological sequestration (for example, injection into depleted oil or gas reservoirs) represents a very expensive proposition now, but could be an important component of a long-term policy solution if costs decline (Newell and Anderson, 2004). Four issues can complicate the inclusion of these activities: monitoring, baselines, comparability and, in some cases, liability. First, some of these sources are fugitive emissions that are difficult to monitor at any point in the product cycle. Second, some activities, especially those involving fugitive emissions, are often left unregulated but allowed to enter as ‘offsets’, requiring a counterfactual baseline against which actual emissions levels can be measured. Fischer (2004b) evaluates various approaches to defining project baselines. Third, a problem of comparability arises with non-CO_2 gases because it is necessary to determine relative prices among greenhouse gases in a market-based programme. As a theoretical matter, the ratio of prices of a ton of current emissions of two different gases should be the ratio of the present value of damages from these emissions (Schmalensee, 1993). In practice it is difficult to apply this formula because it requires a great deal of information about the damages and because it calls for time-varying trading ratios (Reilly, Babiker and Mayer, 2001), which implies significant administrative burdens. Under the Kyoto Protocol and the EU Emissions Trading Scheme, one set of trading ratios is used at all times, and the ratios are calculated by determining the ratio of warming impacts over a 100-year horizon beginning with the present time. Finally, a liability issue arises with regard to sequestration. For both biologically and geologically sequestered carbon, a key question is who should be held liable for carbon dioxide that is released accidentally or otherwise.

Flexibility over time

While price policies naturally allow emissions to rise and fall in response to shocks over time, quantity-based policies must explicitly address the question of whether regulated sources can bank unused allowances for future use or, in some cases, borrow them from future allocations. In the climate change context, merely shifting emissions across time, as opposed to allowing accumulated emissions to vary, holds the environment harmless because climate consequences are generally due to accumulated concentrations, not annual emissions (Roughgarden and Schneider, 1999, discuss the possibility of dependence on both accumulated concentrations and the rate of accumulation.) Such shifts across time might reflect either a more efficient choice of timing in response to capital turnover and technological progress (Wigley, Richels and Edmonds, 1996), or an attempt to ameliorate cost shocks (Williams, 2002; Jacoby and Ellerman, 2004). The rate of exchange between present and future emissions allowances need not be unity: Kling and Rubin (1997) show that the optimal rate at which banked allowances translate across periods should reflect the expected trend in marginal mitigation benefits, the

interest rate, and decay rate of the accumulated gas.

Flexibility over location

The defining feature of the climate-change problem may be its intrinsically global nature. Greenhouse gases tend to disperse themselves uniformly around the globe. As a result, the climate consequences of a ton of emissions of a given greenhouse gas do not depend on the location of the source, either within or across national borders, and shifts in emissions across locations do not change global climate impacts. Under these circumstances, economic efficiency calls for making market-based systems as geographically broad as possible. It supports federal over regional policies, and international coordination over idiosyncratic domestic responses.

International policy initiatives and coordination

International coordination is both crucial and exceptionally difficult to achieve. Studies indicate that the economic and social impacts of climate change would be distributed very unevenly across the globe, with the prospect of large damages to several nations in the tropics coupled with the potential for *benefits* to some countries in the temperate zones (see, for example, Tol, 2005; Mendelsohn, 2003). This uneven distribution makes achieving international coordination especially difficult.

The Kyoto Protocol is the first significant international effort to reduce greenhouse gas emissions. It assigns emission limits to participating industrialized countries for 2008–12, but offers flexibility in allowing these countries to alter their limits by buying or selling emission allowances from other industrialized countries or by investing in projects that lead to emission reductions in developing countries. The importance of these flexibility mechanisms for dramatically lowering compliance costs in this international setting is well documented (Weyant and Hill, 1999).

The Protocol has been criticized on the grounds that it imposes overly stringent emission-reduction targets and lacks a longer-term vision for action. In addition, a core feature of the Protocol – legally-binding emission limits – has been challenged on the grounds that such limits are not self-enforcing, an arguably necessary attribute in a world of sovereign nations (Barrett, 2003). Some argue that the Protocol's project-based mechanisms for encouraging (but not requiring) emission reductions in developing countries are highly bureaucratic and cumbersome, consistent with our earlier comments about project-based programmes more generally. These criticisms have led to considerable research considering the Kyoto structure and comparing it with various alternative international approaches. Aldy, Barrett and Stavins (2003) summarize more than a dozen alternatives, which include an international carbon tax and international technology standards.

A further major criticism is that the Protocol imposes no mandatory emissions limits on developing countries, which collectively are expected to match industrialized countries in emissions of greenhouse gases by 2035. The desire to promote greater participation by developing countries, as well as to involve the United States in the international effort, has motivated considerable research examining, within a game-theoretic framework, the requirements for broader participation and for stable international coalitions (see, for example, Carraro, 2003; Hoel and Schneider, 1997; Tulkens, 1998).

Conclusions

Climate-change economics has produced new methods for evaluating environmental benefits, for determining costs in the presence of various market distortions or imperfections, for making policy choices under uncertainty, and for allowing flexibility in policy responses. Although major uncertainties remain, it has helped generate important guidelines for policy choice that remain valid under a wide range of potential empirical conditions. It has also helped focus empirical work by making clear where better information about key parameters would be most valuable.

Clearly, many theoretical and empirical questions remain unanswered. We suggest (with some subjectivity) that there is a particularly strong need for advances in the integration of emissions policy and technology policy, in defining baselines that determine the extent of offset activities outside a regulated system, and in fostering international cooperation.

From 2003 until 2030 the world is poised to invest an estimated $16 trillion in energy infrastructure, with annual carbon dioxide emissions estimated to rise by 60 per cent. How well economists answer important remaining questions about climate change could have a profound impact on the nature and consequences of that investment.

See Also

- coalitions
- computation of general equilibria
- contingent valuation
- diffusion of technology
- environmental economics
- energy economics
- hedonic prices
- learning-by-doing
- options
- Pigouvian taxes
- second best
- social discount rate
- uncertainty

The authors gratefully acknowledge very helpful comments on earlier drafts by Kenneth Arrow, Steven Durlauf, Raymond Kopp, Richard Morgenstern, Robert Stavins and Roberton Williams III.

Bibliography

Aldy, J., Barrett, S. and Stavins, R. 2003. Thirteen plus one: a comparison of alternative climate policy architectures. *Climate Policy* 3, 373–97.

Arrow, K., Cline, W., Maler, K.-G., Munasinghe, M., Squitieri, R. and Stiglitz, J. 1996. Intertemporal equity, discounting and economic efficiency. In *Climate Change 1995 – Economic and Social Dimensions of Climate Change*, ed. J. Bruce, H. Lee and E. Haites. Cambridge: Cambridge University Press.

Barrett, S. 2003. *Environment and Statecraft*. New York: Oxford University Press.

Barretto, L. and Kypreos, S. 2004. Emissions trading and technology deployment in an energy-system 'bottom-up' model with technological learning. *European Journal of Operations Research* 158, 243–61.

Bovenberg, A. and Goulder, L. 2001. Neutralizing the adverse industry impacts of CO_2 abatement policies: what does it cost? In *Behavioral and Distributional Effects of Environmental Policies*, ed. C. Carraro and G. Metcalf. Chicago: University of Chicago Press.

Buonanno, P., Carraro, C. and Galeotti, E. 2003. Endogenous induced technical change and the costs of Kyoto. *Resource and Energy Economics* 25, 11–34.

Carraro, C., ed. 2003. *The Endogenous Formation of Economic Coalitions*. Northampton, MA: Edward Elgar.

Conrad, K. 2002. Computable general equilibrium models in environmental and resource economics. In *The International Yearbook of Environmental and Resource Economics 2002–2003*, ed. T. Tietenberg and H. Folmer. Cheltenham: Edward Elgar.

Fischer, C. 2004a. Emission pricing, spillovers, and public investment in environmentally friendly technologies. Discussion Paper 04–02. Washington, DC: Resources for the Future.

Fischer, C. 2004b. Project-based mechanisms for emissions reductions: balancing trade-offs with baselines. *Energy Policy* 33, 1807–23.

Fischer, C. and Newell, R. 2005. Environmental and technology policies for climate mitigation. Working paper. Washington, DC: Resources for the Future.

Goulder, L. and Schneider, S. 1999. Induced technological change and the attractiveness of CO_2 emissions abatement policies. *Resource and Energy Economics* 21, 211–53.

Goulder, L., Parry, I., Williams III, R. and Burtraw, D. 1999. The cost-effectiveness of alternative instruments for environmental protection in a second-best setting. *Journal of Public Economics* 72, 329–60.

Hoel, M. and Schneider, K. 1997. Incentives to participate in an international environmental agreement.

Environment and Resource Economics 9, 153–70.

Jacoby, H. and Ellerman, A. 2004. The safety valve and climate policy. *Energy Policy* 32, 481–91.

Jorgenson, D. and Wilcoxen, P. 1996. Reducing U.S. carbon emissions: an econometric general equilibrium assessment. In *Reducing Global Carbon Dioxide Emissions: Costs and Policy Options*, ed. D. Gaskins and J. Weyant. Stanford, CA: Energy Modeling Forum, Stanford University.

Kling, C. and Rubin, J. 1997. Bankable permits for the control of environmental pollution. *Journal of Public Economics* 64, 101–15.

Kolstad, C. 1996. Learning and stock effects in environmental regulation: the case of greenhouse gas emissions. *Journal of Environmental Economics and Management* 31, 1–18.

Manne, A. and Richels, R. 2004. The impacts of learning-by-doing on the timing and costs of CO_2 abatement. *Energy Economics* 26, 603–19.

Mansur, E., Mendelsohn, R. and Morrison, W. 2005. *A Discrete-Continuous Choice Model of Climate Change Impacts on Energy*. New Haven, CT: Yale School of Forestry and Environmental Studies.

McFarland, J., Reilly, J. and Herzog, H. 2004. Representing energy technologies in top-down economic models using bottom-up information. *Energy Economics* 26, 685–707.

Mendelsohn, R. 2003. Assessing the market damages from climate change. In *Global Climate Change: The Science, Economics, and Politics*, ed. J. Griffin. Cheltenham: Edward Elgar.

Mendelsohn, R., Nordhaus, W. and Shaw, D. 1994. The impact of global warming on agriculture: a Ricardian analysis. *American Economic Review* 84, 753–71.

Newell, R. and Anderson, S. 2004. Prospects for carbon capture and storage technology. *Annual Review of Environment and Resources* 29, 109–42.

Newell, R. and Pizer, W. 2003a. Discounting the distant future: how much do uncertain rates increase valuations? *Journal of Environmental Economics and Management* 46, 52–71.

Newell, R. and Pizer, W. 2003b. Regulating stock externalities under uncertainty. *Journal of Environmental Economics and Management* 45, 416–32.

Nordhaus, W. 1982. How fast should we graze the global commons? *American Economic Review* 72, 242–6.

Nordhaus, W. 1994. *Managing the Global Commons*. Cambridge, MA: MIT Press.

Nordhaus, W. 2002. Modeling induced innovation in climate-change policy. In *Technological Change and the Environment*, ed. A. Grübler, N. Nakicenovic and W.D. Nordhaus. Washington, DC: Resources for the Future.

Parry, I., Williams III, R. and Goulder, L. 1999. When can carbon abatement policies increase welfare? The fundamental role of distorted factor markets. *Journal of Environmental Economics and Management* 37, 52–84.

Parry, I. and Oates, W. 2000. Policy Analysis in the presence of distorting taxes. *Journal of Policy Analysis and Management* 19, 603–14.

Pindyck, R. 2000. Irreversiblities and the timing of environmental policy. *Resource and Energy Economics* 22, 233–59.

Pizer, W. 1999. Optimal choice of policy instrument and stringency under uncertainty: the case of climate change. *Resource and Energy Economics* 21, 255–87.

Pizer, W. 2002. Combining price and quantity controls to mitigate global climate change. *Journal of Public Economics* 85, 409–34.

Popp, D. 2004. ENTICE: endogenous technological change in the DICE model of global warming. *Journal of Environmental Economics and Management* 48, 742–68.

Reilly, J., Babiker, M. and Mayer, M. 2001. *Comparing Greenhouse Gases*. Cambridge, MA: MIT Joint Center for the Science and Policy of Global Change.

Richards, K. and Stavins, R. 2005. *The Cost of U.S. Forest-based Carbon Sequestration*. Arlington, VA: Pew Center on Global Climate Change.

Roughgarden, T. and Schneider, S. 1999. Climate change policy: quantifying uncertainties for damage and optimal carbon taxes. *Energy Policy* 27, 415–29.

Schlenker, W., Fisher, A. and Hanemann, M. 2005. Will U.S. agriculture really benefit from global warming? Accounting for irrigation in the hedonic approach. *American Economic Review* 95, 395–406.

Schmalensee, R. 1993. Comparing greenhouse gases for policy purposes. *Energy Journal* 14, 245–55.

Schneider, S. and Goulder, L. 1997. Achieving low-cost emissions targets. *Nature* 389, 13–14.

Sedjo, R. 1995. The economics of managing carbon via forestry: an assessment of existing studies. *Environment and Resource Economics* 6(2), 139–65.

Smith, J. and Tirpak, D. 1989. *The Potential Effects of Global Climate Change on the United States: Report to Congress*. Washington, DC: U.S. Environmental Protection Agency.

Smith, V. 2004. Fifty years of contingent valuation. In *The International Yearbook of Environmental and Resource Economics 2004–2005*, ed. T. Tietenberg and H. Folmer. Cheltenham: Edward Elgar.

Tol, R. 2005. The marginal damage costs of carbon dioxide emissions: an assessment of the uncertainties. *Energy Policy* 33, 2064–74.

Tulkens, H. 1998. Cooperation versus free riding in international environmental affairs: two approaches. In *Game Theory and the Environment*, ed. N. Hanley and H. Folmer. Cheltenham: Edward Elgar.

Weitzman, M. 1974. Prices vs. quantities. *Review of Economic Studies* 41, 477–491.

Weitzman, M. 1998. Why the far-distant future should be discounted at the lowest possible rate. *Journal of Environmental Economics and Management* 36, 201–8.

Weyant, J. and Hill, J. 1999. The costs of the Kyoto Protocol: a multi-model evaluation, introduction and overview. *Energy Journal*, Special Issue.

Wigley, T., Richels, R. and Edmonds, J. 1996. Economic and environmental choices in the stabilization of atmospheric CO_2 concentrations. *Nature* 379, 240–3.

Williams, R. 2002. Prices vs. quantities vs. tradable quantities. Working Paper No. 9283. Cambridge, MA: NBER.

Yohe, G., Ameden, H., Marshall, P. and Neumann, J. 1996. The economic cost of greenhouse-induced sea-level rise for developed property in the United States. *Climatic Change* 32, 387–410.

How to cite this article

Goulder, Lawrence H. and William A. Pizer. "climate change, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000229> doi:10.1057/9780230226203.0247

The New Palgrave Dictionary of Economics Online

cliometrics

Robert Whaples
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Cliometrics (from Clio, the ancient Greek muse of history) studies history by applying the rigour of economic theory and quantitative analysis while simultaneously using the historical record to evaluate and stimulate economic theory and to improve comprehension of long-run economic processes. It thus allows mainstream economists to study economic history using their familiar methods. Since the 1950s, when cliometrics demonstrated that antebellum slave-owning was profitable, it has grown to become the dominant approach to economic history. It is now addressing traditional economic historians' topics like non-market behaviour and embracing methods and findings from disciplines beyond economics.

Keywords

anthropometric history; cliometrics; computers in research; Domesday Book; economic history; Fogel, R.; neoclassical theory; North, D.; slavery

Article

Cliometrics aspires to enhance the study of the economic past by applying the rigour of economic theory and quantitative analysis, while simultaneously using the historical record to evaluate and stimulate economic theory and to improve comprehension of long-run economic processes (Greif, 1997). The term derives from Clio, the ancient Greek muse of history.

The methodology emerged in the United States in the late 1950s among a new generation of neoclassically trained economists who found that many historical writings contained analysis, frequently implicit, that did not conform to the minimum standards of economic literacy and so led to important misinterpretations of the historical record. Pioneering the use of computers in historical research, cliometricians were able to construct extensive macroeconomic time series and also to estimate economic relationships and marginal effects. Instead of imprecise qualitative statements such as ‘it is difficult to exaggerate the importance of this’, cliometrics tried to provide precise numerical estimates of

economic magnitudes and economic relationships.

The potential value of the new approach was convincingly displayed in one of the first cliometric papers, Alfred Conrad and John Meyer's ‘The economics of slavery in the Ante Bellum South’ (1958). Earlier historians had wanted to compare the profitability of owning slaves with that of other investments, but didn't know how. Conrad and Meyer derived the average capital cost per slave, including the average value of the land, animals and equipment used by a slave. Estimates of gross annual earnings were generated from data on the price of cotton and the physical productivity of slaves. Net earnings were then obtained by subtracting maintenance and supervisory costs. The average length of the stream of net earnings was determined from mortality tables. The computation for female slaves took account of the number and productivity of offspring, plus maternity and rearing costs. Conrad and Meyer's preliminary findings strongly refuted the dominant historical interpretation that slave owning wasn't profitable. Numerous subsequent refinements confirmed their conclusion, which is now almost universally accepted.

Among the early cliometric studies that transformed historical interpretation, several works stand out, including Douglass North's *The Economic Growth of the United States, 1790–1860* (1961), Robert Fogel's *Railroads and American Economic Growth* (1964), and Fogel and Stanley Engerman's *Time on the Cross: The Economics of American Negro Slavery* (1974). Indeed, Fogel and North's research was so influential that in 1993 the Royal Swedish Academy cited them ‘for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change’, and awarded them the Nobel Memorial Prize in Economics as ‘pioneers in … cliometrics’ (Royal Swedish Academy of Sciences, 1993).

One can gauge the rise of cliometrics by examining the *Journal of Economic History (JEH)*. In the early 1950s fewer than two per cent of the pages in the *JEH* were devoted to cliometric articles, that is, those using extensive quantification and explicit economic theory. This figure subsequently climbed to ten per cent in the late 1950s, 16 per cent in the early 1960s, 43 per cent in the late 1960s, and 72 per cent in the early 1970s (Whaples, 1991). In the late 1950s cliometrics was seen by some as a mere fad, but by the 1970s it was the standard approach for American economic historians. The cliometric tide has not ebbed; rather, the percentage of cliometric pages in the *JEH* rose to 83 per cent in the late 1980s and 90 per cent in 2004. Opening the pages of the *JEH*, *Explorations in Economic History* or the *European Review of Economic History* is very much like opening the pages of other empirically oriented economics journals, allowing mainstream economists to tackle historical research by familiarizing themselves with historical issues and applying the same methods they would use elsewhere. The overlap between cliometrics and economic history as practised by economists is now almost complete, as cliometrics has become dominant among economists doing historical research outside North America.

Cliometrics is not without critics. Traditional economic historians saw the young cliometricians as outsiders, as economists, not historians or economic historians; they claimed that these upstarts were theorists with little knowledge of the facts and with no sense of history, and that their findings were driven by restrictive theoretical assumptions (Goldin, 1995). The economic historian had always been a hybrid, like the mule able to work in a challenging environment because it shared its parents' best traits. The cliometrician, on the other hand, wasn't a hybrid but was akin to a horse (or, worse, a jackass) that was trying to plough a field for which it was unsuited. Many historians found cliometric methods, models and multivariate regressions incomprehensible and could no longer keep up with research in

economic history. Perhaps as a result many American history departments discontinued training and hiring specialists in economic history, and departments of economic history disappeared where they had been common outside the United States.

Many cliometricians, led by Douglass North, argued that most early cliometric research was too wedded to static neoclassical theory, which tends to focus analysis on historical episodes and topics for which markets were important but which severely limits the issues that can be examined. The neoclassical approach essentially assumes that the same preferences, technology and endowment lead to a unique economic outcome, implying that history does not affect equilibrium and that institutions other than the market don't matter. As the neoclassical grip was loosened in the 1980s, many cliometricians returned to studying issues traditional to economic historians, such as the nature and role of non-market institutions, culture, entrepreneurship, institutional innovation, politics, social factors, distributional conflicts, and the historical processes of economic growth and decline (Greif, 1997). The field has also stretched its boundaries by taking seriously findings and methods from disciplines outside economics, such as the use of anthropometrics (which measures human stature and even skeletal remains) and by reaching even further into the past, such as by analysing the efficiency of the English economy in the 11th century using data from the Domesday Book.

See Also

- anthropometric history
- economic history
- Fogel, Robert William
- North, Douglass Cecil

Bibliography

Conrad, A. and Meyer, J. 1958. The economics of slavery in the Ante Bellum South. *Journal of Political Economy* 66, 95–130.

Goldin, C. 1995. Cliometrics and the Nobel. *Journal of Economic Perspectives* 9, 191–208.

Fogel, R. 1964. *Railroads and American Economic Growth: Essays in Econometric History*. Baltimore, MD: Johns Hopkins University Press.

Fogel, R. and Engerman, S. 1974. *Time on the Cross: The Economics of American Negro Slavery*. Boston: Little, Brown and Co.

Greif, A. 1997. Cliometrics after 40 years. *American Economic Review* 87, 400–3.

North, D. 1961. *The Economic Growth of the United States, 1790–1860*. New York: W. W. Norton.

Royal Swedish Academy of Sciences. 1993. The Sveriges Riksbank (Bank of Sweden) prize in economic sciences in memory of Alfred Nobel for 1993. Press release, 12 October. Online. Available: http://nobelprize.org/economics/laureates/1993/press.html, accessed 27 June 2005.

Whaples, R. 1991. A quantitative history of the *Journal of Economic History* and the cliometric revolution. *Journal of Economic History* 51, 289–301.

How to cite this article

Whaples, Robert. "cliometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000173> doi:10.1057/9780230226203.0248

The New Palgrave Dictionary of Economics Online

clubs

Suzanne Scotchmer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The word ‘club’ has a deceptively frivolous connotation, as does the word ‘game’. But, like game theory, club theory has wide reach. By ‘club’ economists mean a small group of people sharing an activity, often in a context where they care about each other's characteristics. Such activities may include production of goods and services (firms), production of education (schools, academic departments), sharing of private goods in small groups, and community life (churches, charity organizations). The formation of firms, choice of schools, and choice of games to play are all covered by club theory, as are social arrangements like marriage.

Keywords

bundling; capitalization; club theory; clubs; competitive equilibrium; congestion; consumption; education; expected utility; externalities; games; group formation; group type; land markets; local public goods; lotteries; moral hazard; multiple equilibria; public goods; schools as clubs

Article

1 Origins

The word ‘club’ entered the economics literature with a seminal (1965) paper of James Buchanan, who used it to describe a group of people sharing a public good. The key idea he introduced was that public goods are often subject to congestion, and in that sense exhibit some of the rival aspect of private goods. As a consequence, it may be more efficient to replicate a public facility for different (small) groups of users rather than to bear the congestion cost imposed by many people using the same facility. As we will see, club theory has subsequently developed to focus more on interactions among the members of a group, in particular, firms, than on the facilities they share, but both aspects are important.
Buchanan's idea resonated with an idea of Tiebout (1956), who argued that ‘local public goods’ will be

provided optimally if agents are free to choose among jurisdictions. He argued that, if jurisdictions are relatively small, there should be enough jurisdictions and jurisdictional variety to satisfy most residents. These papers led to the conjecture, pursued by a long list of scholars (see Scotchmer, 2002), that competition should provide for optimal group formation. This was by analogy to other market contexts where demand and supply equilibrate at prices that support an efficient allocation, provided that all the actors, including firms, are small relative to the market. Allowing for group formation is a powerful extension of competitive theory, since groups have features that do not fit easily into the general equilibrium theory of Kenneth Arrow, Gerard Debreu and their successors. Such features include externalities among agents, learning of skills, and shared consumption of private goods, whether through rental markets or informal arrangements.

The research agenda surrounding clubs has only recently produced the modifications to general equilibrium theory that accommodate group formation. Along the way, it has been necessary to sort out competing equilibrium concepts, and the difference between models of pure group formation, for which I use the word ‘clubs’, and models of group formation where membership in the group is coupled to occupancy of land. For the latter I use the term ‘local public goods.’

The distinction between clubs and local public goods is the focus of Scotchmer (2002), so I will not focus on it here. Local public goods economies differ from club economies in that jurisdictions are defined by geographical boundaries, and access to local public goods is intermediated by a land market. The price of land serves two related purposes: it allocates land within each jurisdiction, and in conjunction with capitalization effects, allocates agents among jurisdictions. An important complexity is that land and local public amenities are not generally priced or consumed separately. Instead, they are bundled. Although there are two price systems, local taxes and land prices, these cannot generally be interpreted as separate prices for local public goods and land, due to the bundling and to capitalization. In this environment, there are many possibilities for how to define a commodity space and price system, none entirely satisfying. The possibilities are more limited in the club model, where there is no land market that intermediates access to groups. Nevertheless, there are many nuances in adapting general equilibrium theory to group formation, which I now explore.

2 Clubs (groups) in general equilibrium

There have been two approaches to putting clubs into general equilibrium theory, which I refer to as the EGSSZ approach and the CPPT approach. The EGSSZ approach follows Ellickson, Grodal, Scotchmer and Zame (1999; 2001; 2005; referred to here as EGSZ), Scotchmer (2005), Zame (2005), and Scotchmer and Shannon (2007). The CPPT approach follows Cole and Prescott (1997) and Prescott and Townsend (2006).

I begin with the EGSSZ model, and then discuss how it relates to the CPPT model. The commodity space begins with an exogenously given set of *group types*. In a state of the economy, there may be many copies of a given group type. A group type specifies a finite set of memberships, activities that the members engage in, and an input–output vector of private goods. Thus, group types may be interpreted as firms that produce private goods or use private goods as inputs to other activities. The memberships may have qualifications attached to them, such as to be smart or brawny, or to have skills such as the ability to write computer code. These qualifications are called *membership characteristics*. A given

membership may or may not be available to an agent in his consumption set and, if it is, his qualification for the membership may be innate or learned.

Using the notation of Scotchmer and Shannon (2007), let **G** be a finite set of group types, and for each $g \in \mathbf{G}$, let **M(*g*)** be a set of memberships. Each membership $m \in \mathbf{M}(g)$ has attached to it a membership characteristic. The definition of the group type also specifies the group's activities and an input–output vector, say $h(g) \in \mathbf{R}^N$. Some group types do not require inputs or produce outputs; some require only inputs; and some (firms) may require inputs to produce outputs. Labour in a firm is not modelled as an input but rather as a group membership for which skills or other characteristics may be required.

It is convenient to assume that a group's required input–output vector is distributed among members of the group. Thus, each group has associated to it an exogenously given transfer function $t_g: \mathbf{M}(g) \to \mathbf{R}^N$ such that $\sum_{m \in \mathbf{M}(g)} t_g(m) = h(g)$. The transfers specify each member's share of *h*, which may have positive and negative elements. Unless used for incentive purposes as in the papers referenced in Section 4 below, the transfer functions t_g can largely be arbitrary. Any maldistribution can be remedied through membership prices, discussed below, which are endogenous.

There is a continuum of agents, say, *A*=[0,1]. Each agent consumes a bundle of private goods $x \in R^N_+$ and a *list of memberships*, $\ell: \cup_{g \in G} \mathbf{M}(g) \to \{0, 1\}$. The value •(*m*)=1 means that the agent chooses membership *m*, hence belongs to a group of type *g* such that $m \in \mathbf{M}(g)$. A state of the economy is $(x_a, \bullet_a)$, $a \in A$, where $x_a \in \mathbf{R}^N_+$ is agent *a*'s consumption of private goods and $\bullet_a$ is a list of memberships. Each agent $a \in A$ has a utility function u_a, an endowment of private goods, $e_a \in \mathbf{R}^N_+$, and a consumption set.

The utility function takes values $u_a(x_a, \bullet_a)$, where $x_a \in \mathbf{R}^N_+$ is a consumption of private goods and $\bullet_a$ is a list of memberships.

An agent's consumption set determines which memberships are available to him. For example, an agent's consumption set would presumably not permit both a membership in a sumo wrestling club and a membership in a ballet club, since the qualifications for those memberships cannot coexist in the same agent. Consumption sets play a much larger role in club theory than in private-goods economies. Some memberships may not be available to a given agent at all, regardless of what other memberships he chooses or what private goods he invests.

A state of the economy is feasible if it satisfies material balance in private goods, and if, in addition, membership choices are consistent with each other. Membership choices are *consistent* if there exist non-negative real numbers $\alpha(g)$, $g \in \mathbf{G}$, such that the number (measure) of agents who choose each membership $m \in \mathbf{M}(g)$ is $\alpha(g)$. Thus, $\alpha(g)$ represents the number of type-*g* groups, and consistency implies that there are (almost) no groups that are only partially filled.

Consistency of membership choices presents the main technical difficulty in this model. The fixed point in the EGSZ (1999) proof of existence delivers prices such that membership choices are consistent. There is no analogous consistency condition for private-goods exchange economies, and consistency would typically be impossible if the club economy had a finite number of agents rather than a continuum.

To define equilibrium, we need two sets of prices: private-goods prices $p \in \mathbf{R}^N_+$ and membership prices

$q: \cup_{g \in G} \mathbf{M}(g) \to \mathbf{R}$. The membership prices can be positive or negative. An agent's budget is determined by the value of his endowment and the value of the transfers he receives (or is obligated for) in his memberships, evaluated at the equilibrium private goods prices p. These must generate enough income to pay for his memberships at prices q and for his private goods consumption at prices p. Stated informally, an equilibrium consists of private goods prices p, membership prices q, and an allocation $(x_a, \bullet_a)$, $a \in A$, such that each agent is optimizing in his budget set; supply equals demand for private goods; the membership choices are consistent; and the membership prices sum to zero in each group type. Thus, the profit in each group is shared among the members – there is no notion of ownership of groups or group types.

Since the membership prices sum to zero within each group, some members pay other members. Intuitively, some members are paid because they create positive externalities or production opportunities for the members who pay. If, for example, there is a membership that relatively few agents are qualified to fill, or if it is costly to acquire the qualification, then that membership may have a negative price – the member is paid to belong to the group.

All the technical difficulties of general equilibrium theory appear here, such as the distinction between quasi-equilibrium and equilibrium. The technical difficulties in going from quasi-equilibrium to equilibrium are exacerbated by group formation, since, for example, the inputs required for the group can exhaust the endowment of the members, who are then in the zero-wealth position. (See Gilles and Scotchmer, 1997, example 3.)

I now give two informal examples of how club theory expands the reach of general equilibrium theory. First, let the group type be a firm that uses inputs to produce outputs. The required labour, with its required skills, is modelled through group memberships. The required skills might be innate for some workers, but for others might have to be acquired through investments of private goods or memberships in other group types, such as schools or apprenticeships. The negative elements of the input–output vector $h(g)$ are inputs, and the positive elements are the firm's output. These inputs and outputs are divided up among the workers (members) according to the transfer function t_g, and ultimately bought or sold in the market. The transfers contribute to the members' incomes. However, the income from the firm is further redistibuted through the endogenous prices (wages) q.

Substitution in the production process is modelled by using different firm types. If it is possible, for example, to produce the same input/output vector with many unskilled workers or with fewer skilled workers, those options would be modelled as different firm types. Whether a given firm type is used in equilibrium depends on the prices of private goods and memberships, the opportunity costs of workers (reflected in membership prices), and ‘externalities’ created within the firm type. Agents might avoid a very profitable technology because they dislike the production process or because they dislike the characteristics required of the other workers. This feature of production economies is not otherwise accommodated in general equilibrium theory.

Firms are perfectly competitive because each firm of a given firm type has measure zero in the economy, and therefore has no market power. Each firm makes zero profit even though there is no concept of linearity in production. The only constant returns to scale is that many copies of a given firm type may form, each producing the same output from the same inputs. However, each copy of the group type is a separate zero-profit entity.

Second, let the group type be a school. Suppose for simplicity that there are no private goods inputs or

outputs, hence no internal transfers. Some of the memberships are called ‘teacher’ , and others are called ‘student’. The same person is typically not qualified for both roles. The student memberships may be further differentiated. Some student memberships may be called ‘advantaged student’ (and require the appropriate qualification) and others ‘disadvantaged student’. Which membership a student is qualified for is presumably constrained in his consumption set.

Since the membership fees sum to zero, the teacher will presumably be paid, and the students will pay. However, if advantaged students confer positive externalities on disadvantaged students, it might occur that both teachers and advantaged students are paid by disadvantaged students. Otherwise the advantaged students might prefer schools where all memberships are for advantaged students, where they themselves receive higher externalities.

The model I have described is a delicate amalgam of features inherited from the theory of general equilibrium for exchange economies and features of public goods economies, such as externalities and the sharing of private goods. In general equilibrium theory, the key features of a competitive equilibrium are that (*a*) the commodity space is defined independently of the set of agents, (*b*) the price system is complete with respect to the set of commodities, (*c*) prices are anonymous, and (*d*) agents optimize with respect to the price system, but not by observing other agents' preferences or endowments. Early discussions of price-taking equilibrium for club economies missed various of these requirements. For example, in analyses that use the ‘core’ equilibrium concept from game theory, following Pauly (1967), the commodity space has been defined as the set of groups (coalitions) that are feasible in the economy, even when the core is decentralized with prices. This idea departs from general equilibrium theory in that the available commodities (group types) depend on the set of agents. That model has other limitations as well. Since agents can only belong to a single club, it cannot accommodate the notion that an agent may want to belong to several groups, for example, a school where he acquires skills and a firm where he exercises the skills. Further, many of the earlier models also restricted to a single private good (often with transferable utility), and therefore did not allow the important interpretation that groups are firms in a production economy.

In the model I have described, following EGSZ (2005) and Scotchmer and Shannon (2007), characteristics are defined as part of the membership, rather than attached to the agent. An agent can only choose a given membership if he is innately endowed with the characteristic required for it or, alternatively, can acquire it. The earlier models of EGSZ (1999; 2001) made the more restrictive assumption that all characteristics are innate, but the same proofs of existence of equilibrium and related theorems apply to both cases.

3 Randomized memberships

In the model described above, agents choose memberships deterministically. However, the premise behind the CPPT branch of the clubs literature is that randomness can be utility enhancing, and randomness will therefore be created by the market. This depends on the premise that utility functions can be interpreted as von Neumann–Morgenstern utility functions (not assumed in the EGSZ model), and is illustrated by the following example.

Suppose there are two firm types, $g_1, g_2 \in \mathbf{G}$. The firm type g_1 has a single worker and g_2 has a worker and supervisor. The club memberships are $\mathbf{M}(g_1) = \{m_{w1}\}$, $\mathbf{M}(g_2) = \{m_s, m_{w2}\}$. Suppose that each

agent can choose a single membership, that a third of the agents have consumption sets that permit supervisor memberships, m_s, and two-thirds of the agents have consumption sets that permit worker memberships, m_{w1} or m_{w2}. There is a single private good, of which each agent has an endowment e. The utility of supervisors is equal to their consumption of the private good, regardless of memberships, and the utility of each worker is the following, where c is his consumption of the private good, and f is positive and increasing.

$$u(c, \ell) = \begin{cases} \frac{1}{2} f(c) & \text{if } \ell = 0 \\ f(c) & \text{if } \ell(m_{w1}) = 1 \\ f(c) + 1 & \text{if } \ell(m_{w2}) = 1 . \end{cases}$$

In an EGSZ-type equilibrium, the prices of memberships are $q(m_{w1}) = 0$ and $q(m_s) = -\hat{q}$, together with price p=1 for the private good, where $f(e - \hat{q}) + 1 = f(e)$. Workers receive utility $f(e) = f(e - \hat{q}) + 1$ and supervisors receive utility $e + \hat{q}$. The supervisors are paid by the workers because agents who are qualified to be supervisors are relatively scarce and therefore valuable. They facilitate the creation of high value in supervised firms.

The basic idea of the clubs model of Cole and Prescott (1997) can be seen in the example. If the workers' utility function can be interpreted as a von Neumann–Morgenstern utility function, and if f is concave, the EGSZ-type equilibrium is not efficient. The expected utility of workers can be increased without decreasing the utility of supervisors by equalizing the workers' consumption in the two memberships m_{w1}, m_{w2}, and letting them randomize on those two memberships. The equalized consumption is $\hat{c} = (1/2)(2e - \hat{q})$. Then the *ex post* utility of workers who end up in m_{w1} is less than the *ex post* utility of workers who end up with m_{w2}, but their *ex ante* expected utility is the same, namely, $(1/2)\ f(\hat{c}) + (1/2)(f(\hat{c}) + 1) = f(\hat{c}) + (1/2)$, and larger than $f(e)$.

Cole and Prescott argue that the randomized outcome can be achieved in two ways. The agents can buy lotteries on club memberships directly, or the agents can buy randomizations on wealth and then choose their club memberships deterministically as in the EGSZ model. In the first implementation, prices are on units of probability placed on different consumption bundles. In the example, consumption bundles would be elements of some finite set $\mathbf{L}=\{(c,m)\}$, where, for mathematical convenience, c is in a finite set of points in $\mathbf{R}_+$ and $m \in \{m_s, m_{w2}, m_{w1}, m_o\}$, where m_o is a null membership that means no group membership is chosen. The prices are $\{p(c, m) \in \mathbf{R}_+ : (c, m) \in \mathbf{L}\}$. If an agent chooses a consumption bundle (c,m) with probability one, he pays $p(c,m)$. More generally, an agent can choose probabilities (a 'lottery') $\{x(c, m) \in \mathbf{R}_+ : (c, m) \in \mathbf{L} \Sigma_{(c,m)\in\mathbf{L}} x(c, m) = 1\}$. It is then natural to define the utility function on the vectors x, so that the agent receives utility $u(x)$ and pays $p \cdot x$.

This transformation, also used by Prescott and Townsend (2006), gives the group-formation model a structure that is similar to an exchange economy. However, for analytical tractability some desirable

features are given up along the way, such as that the authors assume there is a finite set of preference types, and restrict each agent to a single membership.

Moreover, there is a single profit-maximizing ‘intermediary’ on the supply side, which offers a combination of lotteries that maximizes profit, and creates firms from the outcome of agents' (independent) randomizations. To do this, the intermediary must serve a continuum of agents. The intermediary is therefore a different type of firm than the group types in the EGSSZ model and the firms of the CPPT model, such as g_1, g_2.

An important role of the intermediary in the CPPT model is to make transfers of value among the groups over which lotteries are offered. In the randomization above, the single membership in the firm type g_1 is coupled with consumption $\hat{c} < e$. The value of the member's consumption in g_1 is less than the value of the endowment, while the value of the members' consumptions in g_2 is more than the value of their endowments. Since the value of consumption must equal the value of endowments in aggregate, there is a transfer of value from g_1 to g_2. The intermediary who creates the lottery absorbs both sides of this transfer.

Scotchmer and Shannon (2007) show how lotteries on memberships can be introduced to the EGSSZ model through *lottery group types*, which are finite and are formally treated the same as ordinary group types. There is no need for a distinguished firm (intermediary) that serves a continuum of agents. A lottery group type is composed of several constituent group types in **G**. A feasible lottery must have the same number of lottery memberships as there are memberships in the constituent group types, since the lottery members will be assigned to the memberships in the constituent group types. The probability distribution is uniform on all assignments that are consistent with the memberships.

In the example, a lottery group type is constructed from one copy of g_1 and one copy of g_2, and has three memberships. Worker memberships to the lottery group type are such that the member can be assigned to m_{w1} or m_{w2}, and a supervisor membership is such that the member can only be assigned to m_s. There are two ways to make this assignment, each with probability one-half. Each worker has probability one-half of being assigned to m_{w1} or m_{w2}, as required. If the lottery group type is defined such that the internal transfer of each worker to the supervisor is $e - \hat{c}$, the equilibrium membership prices q are zero. With this structure, each lottery is a group type with finite memberships, and, as such, fits directly into the EGSZ model with no modification. Each worker pays the same membership fee for a lottery membership, but receives different *ex post* utility, depending on the outcome of the internal lottery. There are no transfers of value among lottery groups, as required by the zero-profit condition, but there are transfers of value among groups within each lottery group type.

A caveat is that not all lotteries can be accommodated with a finite number of group types. Each lottery group defines fixed probabilities on wealths and memberships. Different probabilities are provided by different lottery groups. Since there are continuously many possible lotteries, a complete lottery space would require a continuum of lottery group types, some very large. Thus, as in the CPPT approach, there is some loss in the technical convenience of restricting to a finite number of group types.

4 Unverifiable characteristics and games

In game theory the game is primitive. An agent either finds himself in the game or he does not, but there

is generally no explanation for which game he finds himself in. Club theory allows agents to choose among games. However, to interpret a game as a finite group type, the theory must accommodate strategies and characteristics that are not verifiable. Such an extension is suggested by Prescott and Townsend (2006), who use the CPPT approach to discuss how the market chooses among firm types that are subject to moral hazard. Equilibrium will weed out the contractual arrangements that are inefficient, where that may depend on the prices at which private goods trade. The same idea is taken up and extended by Zame (2005) and Scotchmer and Shannon (2007). The latter two papers build closely on EGSZ (1999; 2005) but differ in emphasis and in the way group formation is formalized.

Some unverifiable characteristics are chosen, and some are innate. The natural word for an unverifiable characteristic that is chosen is ‘strategy’, while it is more natural to say ‘unverifiable characteristic’ when the characteristic is innate but unobservable. Both play the same role in the model. In a normal-form game, for example, the membership might indicate row player or column player, and the strategy might indicate the unverifiable play. In a group type that is a firm, the membership is a job, and the unverifiable job characteristic might be innate proficiency at writing computer code.

When strategies (characteristics) are unverifiable, the groups that materialize from a member's choices will have a random component, namely, the unverifiable characteristics of other members. For random realizations of groups, Scotchmer and Shannon (2007) use the term ‘augmented’ group types. The agents first choose their verifiable memberships and unverifiable strategies, and are then randomly matched into augmented groups consistent with their choices.

If the unverifiable characteristics can be distinguished according to something verifiable like output, then group types can be defined such that agents screen optimally into groups, just as if the characteristics themselves were verifiable (see example 2 in Scotchmer and Shannon, 2007.) No such ploy is available if the unverifiable characteristics affect utility directly.

After being matched into augmented groups, agents choose their consumptions of private goods. Each agent's income and demand for private goods may depend on the unverifiable characteristics in his groups. Since each agent's demand depends on the random matching, there is no conceptual reason to think that private-goods prices should be the same for all matchings, and Scotchmer and Shannon do not assume it. There may be two sources of uncertainty in an agent's consumption of private goods: uncertainty about the augmented groups and uncertainty about the prices of private goods. Both sources of uncertainty affect the *ex ante* demand for group memberships, and the optimizing choices of strategies.

If the set of agents were finite, the augmented groups realized by different agents could not be independent of each other. Duffie and Sun (2004a,b) show that the continuum remedies this problem. In the continuum, each agent's random match can be understood as independent of any other agent's random match, and a law of large numbers applies to demand. The law of large numbers provides an easy way to prove existence of equilibrium despite the randomness caused by unverifiable characteristics. If one assumes that the equilibrium prices must be the same at every random matching, aggregate demand can be treated as constant for all random matchings, and existence of equilibrium follows from EGSZ (1999). But this should not lead us to believe that constant prices are natural. There is no reason that the same equilibrium price vector should be selected at each random matching – constant prices are an assumption, not a conclusion. (This is an important difference between the treatments of Zame, 2005, who assumes constant prices, and Scotchmer and Shannon, 2007, who

explore the consequences when prices can depend on the random matching. Variation in prices may reduce welfare.)

Prescott and Townsend (2006) prove the first welfare theorem for a class of economies with moral hazard. In contrast, Zame (2005) and Scotchmer and Shannon (2007) show many senses in which equilibrium will be inefficient. The difference lies partly in the classes of economies considered, and partly in the definition of 'efficiency', which is only defined relative to the trading opportunities in the economy. For example, Scotchmer and Shannon point out that inefficiency in teams would vanish if agents could choose a game with a residual claimant. In the model of Prescott and Townsend, that is not an option.

These models have three broad classes of inefficiencies. First, the exogenous set of group types (games) in the economy may not be rich enough to achieve first-best efficiency, as in the teams example. Second, there are belief-driven coordination problems, well known in game theory, that are not solved by embedding games in general equilibrium. There may be multiple equilibria, including efficient ones and inefficient ones, each supported by beliefs that are correct in equilibrium. Third, there are inefficiencies in the trading of private goods. Trades in private goods are always efficient from an *ex post* point of view (conditional on the random matching) but not necessarily from an *ex ante* point of view. Depending on what is observable, the latter inefficiency may be remediable through insurance markets.

See Also

- consumption externalities
- externalities
- general equilibrium

Bibliography

Buchanan, J. 1965. An economic theory of clubs. *Economica* 33, 1–14.

Cole, H.L. and Prescott, E.S. 1997. Valuation equilibrium with clubs. *Journal of Economic Theory* 74, 19–39.

Duffie, D. and Sun, Y. 2004a. Existence of independent random matching. Working paper, Graduate School of Business, Stanford University.

Duffie, D. and Sun, Y. 2004b. The exact law of large numbers for independent random matching. Working paper, Graduate School of Business, Stanford University.

Ellickson, B., Grodal, B., Scotchmer, S. and Zame, W. 1999. Clubs and the market. *Econometrica* 67, 1185–217.

Ellickson, B., Grodal, B., Scotchmer, S. and Zame, W. 2001. Clubs and the market: large finite

economies. *Journal of Economic Theory* 101, 40–77.

Ellickson, B., Grodal, B., Scotchmer, S. and Zame, W. 2005. The organization of consumption, production and learning. In *The Birgit Grodal Symposium*, ed. K. Vind. Berlin: Springer-Verlag.

Gilles, R. and Scotchmer, S. 1997. On decentralization in replicated club economies with multiple private goods. *Journal of Economic Theory* 72, 363–87.

Holmstrom, B. 1984. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Pauly, M.V. 1967. Clubs, commonality and the core: an integration of game theory and the theory of public goods. *Economica* 34, 314–24.

Prescott, E.S. and Townsend, R.M. 2006. Firms as clubs. *Journal of Political Economy* 114, 644–71.

Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of Public Economics*, vol. 4, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Scotchmer, S. 2005. Consumption externalities, rental markets and purchase clubs. *Economic Theory* 25, 235–53.

Scotchmer, S. and Shannon, C. 2007. Verifiability and group formation in markets. Working Paper E07–347, Department of Economics, University of California, Berkeley.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Zame, W. 2005. Incentives, contracts and markets – a general equilibrium theory of firms. Working Paper No. 843, Department of Economics, University of California Los Angeles.

How to cite this article

Scotchmer, Suzanne. "clubs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000178> doi:10.1057/9780230226203.0249

The New Palgrave Dictionary of Economics Online

coalitions

Myrna Wooders and Frank H. Page, Jr.

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Coalitions appear in an incredible diversity of economic and game-theoretic situations, ranging from marriages, social coalitions and clubs to unions of nations. We discuss some of the major approaches to coalition theory, including models treating why and how coalitions form, equilibrium (or solution) concepts for predicting outcomes of models allowing coalition formation, and current trends in research on coalitions. We omit a number of related topics covered elsewhere in this dictionary, such as matching and bargaining.

Keywords

f-core; abstract games; admissible set; asymmetric information; bargaining; bargaining set; basins of attraction; clubs; coalitions; cooperative games; cores; differential information; domination; epsilon core; extensive form games; far-sighted stability; hedonic games; implicit coalitions; incomplete information; *information sharing* ; inner core; irreversibilities; kernel; law of demand; law of supply; linear programming; link formation; local public goods; Myerson value; Nash equilibrium; Nash program; network formation; non-cooperative games; non-transferable utility games; Owen equilibrium; Owen set; pairwise stability; partnered core; private information; public goods; Shapley value; small group effectiveness; solution concepts; strong stability; subgame perfection; superadditivity; supernetworks; tau value; Tiebout hypothesis; transferable utility games; von Neumann–Morgenstern stable set

Article

The traditional notion of a coalition is a group of players who can realize some set of outcomes for its own membership. How to define this set of outcomes is a fundamental question and its definition is typically either avoided, by assuming that the set of outcomes is given, or treated simultaneously with a solution concept. Alternatively, some process may be given that plays a role in determining the set of

outcomes that are achievable by each coalition.

How to define a coalition is an even more fundamental question. Typically a coalition is taken as a subset of players of a game. Yet we often perceive that individuals belong to overlapping coalitions. For example, an individual may belong to the Citizens Coalition for Responsible Media, Immunization Action Coalition and the Democratic Party. We also perceive that coalitions may be *temporary* alliances of groups of people, factions, parties, or nations. For most of this article, however, we view a coalition as simply a subset of players of a game.

When both the concepts of a coalition and its attainable set of outcomes have been defined, the question arises of how the gains from coalitional activities are to be allocated among the members of any coalition that might form, bringing us to the notion of a solution concept. A solution concept is a rule which must be satisfied by any allocation or attainable outcome that is viewed as stable or as an equilibrium. Given a description of the primitives of a situation (a game, economy, or social situation, for example) a solution concept may be viewed as predicting which outcome(s) will emerge. Implicitly, a solution concept involves assumptions about the behaviour of individuals or groups of individuals. Even in situations where a particular solution concept seems compelling, however, there may be no attainable outcomes satisfying the requirements of the solution concept. This problem, and the fact that no single solution concept seems to fit all situations, means that there are competing notions of solution concepts.

In this article we discuss issues of coalitions, the outcomes attainable by coalitions and the division of the benefits of coalition formation among the members of a coalition. Many of the fundamental questions that still intrigue researchers have their roots in the early literature of game theory. We will sketch some of the main concepts in the literature on coalitions, going back to von Neumann and Morgenstern's celebrated volume, with its notion of dominance, and also sketch some of the current approaches to questions of coalitional activities. We conclude by noting some new approaches to what a coalition might be and do and directions that research may be taking.

Domination

What a coalition can achieve, or, even more fundamentally, what a coalition can improve upon for its own membership is a fundamental question. This was realized already by von Neumann and Morgenstern (1953), who introduced the notion of domination. An imputation x (or payoff vector, listing a payoff for each participant in the society) *dominates* another imputation y with respect to a coalition S if the members of S are convinced or can be convinced that they have a positive motive for bringing about y and believe that they can do so. The coalition S is called *effective* (for x). Note that it is possible there is another payoff vector y', a coalition S' that is effective for y', and y' dominates y with respect to S' (but not with respect to S) and in general, the relation ‘dominates’ may not be transitive.

Solution concepts

A number of solution concepts based on notions of domination and effectiveness of coalitions have been defined. Three especially prominent concepts are the von Neumann–Morgenstern stable set, the Shapley value, and the core. A set V of payoff vectors, where each vector is a listing of payoffs to players in a

game, is a *von Neumann–Morgenstern stable set* if (*a*) no payoff vector in *V* is dominated by another payoff vector in *V* and (*b*) every payoff vector not in *V* is dominated by some vector in *V*. The *core*, introduced in Gillies and Shapley in 1953 (see the Logistics Research Project, 1957, which contains descriptions of the presentations of D. Gillies and L.S. Shapley, where the core was introduced), consists of those payoff vectors *x* that are feasible and undominated. The formulation of Gillies (1959) of the core of an abstract game can be widely applied. An *abstract game* consists of a set of alternatives for each coalition and a dominance relationship. The *Shapley value*, introduced in Shapley (1953), assigns to each player his *expected* marginal contribution to coalitions and is also used in numerous applications. Alternative notions of the core and of the value include the *Owen value* (Owen, 1977), the τ *-value* (Tijs, 1981), the *inner core* (Myerson, 1995; Qin, 1994; and references therein), and the *partnered core* (Albers, 1979; Bennett, 1983; Reny and Wooders, 1996a).

Let us consider a simple example. Let $N = \{1, 2, 3\}$ be the player set. Suppose that any one player can earn zero, any two players can earn one dollar and the three players together can earn $M \geq 0$ dollars. Suppose $M = 1$; then the von Neumann–Morgenstern stable set consists of the payoff vectors $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$, and $(0, \frac{1}{2}, \frac{1}{2})$. Any payoff vector ($z_1$, z_2, z_3) is in the core if $z_i \geq 0$ for all $i \in N$ and $z_i + z_j \geq 1$ for every pair *i*, *j*. This implies that, unless $M \geq \frac{3}{2}$, the core is empty. The Shapley value is defined for *superadditive games*, games with the property that the set of payoff vectors achievable by any union of disjoint coalitions is at least as large as the set of payoff vectors achievable by the coalitions independently. Superadditivity, for our example, implies that $M \geq 1$, in which case the Shapley value consists of the payoff vector $(\frac{M}{3}, \frac{M}{3}, \frac{M}{3})$.

The bargaining set, introduced by Aumann and Maschler (1964), is based on threats and counter-threats. A payoff vector *x* is in the *bargaining set* if for every credible objection there is a credible counter-objection. That is, if there is a payoff vector *y* that dominates *x* with respect to a coalition *S* then there is another payoff vector y' and coalition S' that is effective for y' and y' is at least as good as *x* for the members of S' who are not in *S* and at least as good as *y* for members of both *S* and S'. There are a number of related concepts. The *kernel*, introduced in Davis and Maschler (1965), requires that objections and counter-objections have equal strengths. For our example above, the point $(\frac{M}{3}, \frac{M}{3}, \frac{M}{3})$ is also in the bargaining set and in the kernel. Recent research on concepts of the bargaining set has been spurred by the *Mas-Colell bargaining set* (Mas-Colell, 1989) which adapts the bargaining set to economies with a continuum of agents and proves equivalence of the outcomes of the bargaining set and the core in an exchange economy.

Another interesting notion is the *admissible set*, introduced in Kalai and Schmeidler (1977). (See also references therein and Shenoy, 1980.) Take as given a set of feasible alternatives, denoted by *S*, a dominance relation M and the transitive closure of *M*, denoted by $\widehat{M}$. The *admissible set* is the set $A(S, M) = \{x \in S: y \in S \text{ and } y \widehat{M} x \text{ imply } x \widehat{M} y\}$. The admissible set describes those outcomes that are likely to be reached by any dynamic process that respects preferences. Note that the admissible set concept can be applied to a host of game-theoretic situations, ranging from non-cooperative games, where a coalition consists of an individual player, to fully cooperative games, where any coalition can be

allowed to form. As shown by Kalai and Schmeidler, under certain conditions the admissible set coincides with the set of Nash equilibria and, for cooperative games, the admissible set coincides with the core. More recently, it has been shown that the admissible set consists of the union of basins of attraction, and a von Neumann–Morgenstern set consists of one member of each basin (Page and Wooders, 2006).

Behaviour of coalition members

What a coalition can achieve also depends on the *behaviour of the members of the coalition*. For example, potential coalition members may bargain over the distribution of the gains to coalition formation and outcomes in the core may not be achievable as equilibria of non-cooperative bargaining processes (an important point made by John Nash, 1953, leading to the *Nash program*). Chatterjee et al. (1993) demonstrate this point very well for transferable utility (TU) games, which describe what a coalition can achieve by simply a number, in interpretation, an amount of money, for example.

As stressed by Xue (1998), it may matter whether players are farsighted or myopic in their thinking about forming coalitions. Myopic players take as given the actions of others and behave accordingly. In choosing their actions, farsighted players, in contrast, take into account the reactions of other players to their actions and thus the eventual consequences of their actions. See also Diamantoudi and Xue (2003) who study the far-sighted core of a hedonic game – a game where, instead of payoff sets for coalitions, preferences are given for each individual over all coalitions in which he is contained – and Mauleon and Vannetelbosch (2004) who both allow ‘spillovers’ between coalitions and farsightedness of players, and demonstrate sufficient conditions for there to exist stable outcomes. (Two important papers in the game theoretic literature studying farsightedness, but not coalition formation, are Chwe, 1994, and Harsanyi, 1974.)

Players may also take into account ‘asymmetric dependencies’ within coalitions. A solution displays an asymmetric dependency if one player needs the presence of a second player to realize his payoff in the solution, but the second player does not need the presence of the first. When a player i is dependent on another player j in this sense, but j is not dependent on i, then j is in a position to attempt to obtain a larger share of the surplus from i. Consider, for example, a two-person divide-the-dollar bargaining game. Any division giving the entire dollar to one participant displays an asymmetric dependency; the player receiving the dollar is dependent on the player receiving zero. The player receiving zero is not compelled to join the two-person coalition to receive his part of the payoff. In contrast, to achieve the payoff of 50 cents for each player the two-person coalition is compelled to form – the players are partnered. The partnered core, introduced in Albers (1979) and Bennett (1983) for TU games and in Reny and Wooders, (1996a) for non-transferable utility games (where the set of payoffs achievable by a coalition are described by vectors listing a payoff for each member of the coalition) consists of those outcomes in the core with the property that, to achieve his payoff, no individual needs another individual who does not need him. Even in well-behaved exchange economies there may be no outcomes in the core that are not partnered; that is, all outcomes in the core may be vulnerable to the threat of secession by some coalition of players. Page and Wooders (1996) provide an example.

Behaviour of non-coalition members

In many situations, what a coalition can achieve depends on assumptions about the *behaviour of non-coalition members* (sometimes called the 'complementary coalition', although there is no requirement that the complementary coalition actually forms an alliance); for example, individuals may steal, or drop garbage in the backyards of others, or there may be widespread pollution. Two alternative definitions of the core, from Aumann and Peleg (1960), highlight the dependence of the core on the assumptions made about what outcomes are perceived as feasible by coalitions: the α-core, consisting of those outcomes that a coalition can guarantee for its membership, and the β-core, consisting of those outcomes that a coalition cannot be prevented from achieving for its membership. In some situations, such as private goods economies without externalities or in some recent models of economies with clubs or local public goods, these two notions are equivalent, but, as noted by Shapley and Shubik (1969a), in the presence of externalities between coalitions these concepts may yield different outcomes.

Members of a coalition may also be directly affected by the structure of alliances among non-members of the coalition. This consideration underlies the Lucas and Thrall (1963) concept of *a partition function form game*, where the attainable total payoff to a coalition depend on the structure of coalitions formed by the complementary player set.

In the approach of Chander and Tulkens (1995; 1997), to predict the set of outcomes that it can achieve, a coalition presumes that the outside players will adopt their individually best reply strategies, leading to their notion of the gamma core. In the sense that the non-coalition members are treated as forming one-person coalitions, the Chander–Tulkens approach is more restrictive than that of Lucas and Thrall. When it is assumed that coalitions can freely merge or break apart and are farsighted, however, Chander (2007) demonstrates that, subsequent to a deviation by a coalition, the non-members will have incentives to break apart into singletons, thus providing a justification for the Chander–Tulkens approach.

Other approaches to the question of what a coalition can achieve for its membership have also appeared in the literature. Some recent contributions allow theft or pillage by non-coalition members; see, for example, Jordan (2006), where the payoffs attainable by a coalition are determined endogenously, and references therein.

In application, questions of the behaviour of the non-coalition members have been especially important in industrial organization and environmental economics; see, for example, Yi (1997) and Bloch (1996); see Bloch (2005) and Carraro (2005) for discussions of relevant literature.

Information sharing within coalitions

When players have private information new and difficult issues arise. Chief among these is the issue of *information sharing within coalitions*. How can members of a coalition be induced to share their private information truthfully? Or, if it is not shared truthfully, how much information will be shared and how much of it will be believed? In his seminal paper, Wilson (1978) introduced two notions of the core for situations with private information, namely, the coarse core and the fine core; later Yannelis (1991) introduced the private core. Each of these core notions corresponds to assumptions about the extent to which private information of individual players is shared within coalitions. These issues are further

addressed in Allen (2006), who treated core concepts in exchange economies, and Page (1997), who extended Allen's results to infinite dimensional commodity spaces. There is also the question of what informational time frame should be used in defining a solution concept. Following the informational distinctions introduced by Holmstrom and Myerson (1983) in extending the notion of Pareto efficiency to economies with private information, we can ask whether the solution concept should be *ex ante* (that is, defined relative to *ex ante* probability beliefs concerning the future information state of the economy – and therefore before players know their private information), whether it should be interim in nature (that is, defined relative to each possible profile of players' private information – and therefore after each player knows his private information but before players know the information of others), or whether it should be *ex post* (that is, defined relative to each possible information state of the economy – and therefore after each player knows the information state of the economy).

Following a mechanism design approach, Forges, Mertens and Vohra (2002) address the issue of honest information revelation within coalitions by focusing on coalitionally incentive-compatible direct mechanisms. A coalitional direct mechanism is a mapping from the set of information profiles of coalition members into coalitional allocations. A coalitional direct mechanism is *incentive compatible* if no coalition member has an incentive to lie about his private information –on the assumption that other coalition members report their private information truthfully (that is, truthful reporting is a Nash equilibrium of the coalitional revelation game induced by the mechanism). Formulating the coalitional mechanism design game as a TU game in characteristic function form, they demonstrate non-emptiness of the incentive compatible '*ex ante* core'. Other contributions which analyse interim core notions include Ichiishi and Idzik (1996), Hahn and Yannelis (1997), Vohra (1999), Volij (2000), Demange and Guesnerie (2001), Dutta and Vohra (2005) and Myerson (2007). See Forges, Minelli and Vohra (2002) for a survey.

The core with incomplete information is gaining prominence in applications, such as political economy (see, for example, Serrano and Vohra, 2006).

Coalition formation

Other important questions are *how coalitions form and how coalition structures influence the behaviour of individuals within coalitions*. Several approaches are possible. Coalition formation and individual behaviour can be viewed as outcomes of market mechanisms or as outcomes of assumed cooperation within groups that may form. Alternatively, coalition formation and individual behaviour can be viewed as outcomes induced from non-cooperative behaviour. More recently coalition formation and individual behaviour within coalitions have been modelled in network settings.

The market/cooperative game approach

As suggested by Tiebout (1956) and Buchanan (1965), individuals may take as given prices for membership in coalitions (clubs, firms, jurisdictions, and so on). Tiebout conjectured that if public goods are 'local' (that is, public goods are subject to congestion and individuals can be excluded from the public goods provided in jurisdictions in which they are non-members), then the possibility of individuals moving to the jurisdictions where their wants are best satisfied subject to their budget

constraints and to taxes creates a competitive ‘market-like’ outcome. A part of the outcome is a partition of individuals into jurisdictions. Buchanan (1965) stressed the importance of collective activities in a model of clubs with optimal club size; to illustrate, considering our example above where any two players can earn one dollar, if $M < \frac{3}{2}$, then two is the optimal club size. One way to formulate the Tiebout hypothesis (Pauly, 1970; Wooders, 1978; 1980) is to model the economy as one where individuals pay prices to join coalitions/clubs/jurisdictions and to demonstrate equivalence of the core and the set of outcomes of price-taking equilibrium. The results of these early papers have been greatly extended and refined; see, for example, Conley and Wooders (2001); Ellickson et al. (2001) and, for a survey, Conley and Smith (2005). The spirit of the main results is that, whenever *small group effectiveness* holds – that is, whenever all or almost all externalities can be internalized within relatively small groups of individuals (clubs, jurisdictions, firms, trading coalitions, and so on) or, in other words, whenever all or almost all gains to collective activities can be realized with some partition of the total player set into relatively small coalitions – then economies with many participants are ‘market like’ in the sense that price-taking economic equilibrium exists and the set of equilibrium outcomes is equivalent to the core of the economy.

The results for models of economies with local public goods and clubs suggest results for cooperative games with endogenous coalition structures. Under small group effectiveness, cooperative games with many players are ‘market games’ (as defined in Shapley and Shubik, 1969b) and thus can be represented as economies where all individuals have concave, continuous utility functions (Wooders, 1994a; 1994b). (That the conditions of Wooders, 1983, imply that games with many players are market games was first noted by Shubik and Wooders, 1982, and the concavity of the limiting per capita payoff function was first explicitly noted in 1987 by Robert Aumann in his entry game theory in the first edition of this dictionary, which is reproduced in the present edition).

A simple example may provide some intuition. Suppose any two players can earn $1.00, as in our earlier example, but now suppose that there are *n* players in total. If *n* is odd, then the core is empty, but for large *n* each player can receive nearly $0.50 so certain approximate cores are non-empty and the approximation is ‘close’. In defining an appropriate approximate core concept the modeller can either suppose that there are some costs to coalition formation, which can be allowed to go to zero as *n* becomes large, or that a relatively small set of players can be ignored. Now, more generally, suppose instead that the payoff to a coalition with *m* members is a real number *v*(*m*). Suppose the game is essentially superadditive – the total payoff achievable by $m + m'$ players is greater than or equal to $v(m) + v(m')$. Then the only condition required to ensure non-emptiness of approximate cores of games with many players is that there is a bound *K* such that $\frac{v(m)}{m} \leq K$ for all *m*, which implies small group effectiveness. The limiting concave utility function alluded to above is $u(n) = \sup \frac{v(m)}{m} n$. See also Robert Aumann's discussion of Wooders's (1983) result in game theory.

Some other market properties of a game with many participants are that: Outcomes in the core or approximate cores treat most similar players nearly equally (Wooders. 1983; Shubik and Wooders, 1982; and for the most recent results, Kovalenkov and Wooders 2001a). The Shapley value is in an approximate core (Wooders and Zame 1987). A ‘law of scarcity’ holds; that is, increasing the abundance

of one type of player leads to a decrease in the core payoffs to individual players of the same of similar types (Scotchmer and Wooders, 1988; and, for recent results and references, Kovalenkov and Wooders 2005b; 2006). The law of scarcity is in the spirit of the law of demand and law of supply of private goods economies but differs in that an additional player in a game creates both creates additional demand (for the cooperation of others) and additional supply (of players of the same type).

To illustrate further the intimate relationships between markets and economies with group activities such as clubs and/or local public goods, we will discuss Owen (1975), who treats a production economy where individuals are endowed with resources that may be used in production. Rather than selling their resources to firms, individuals form coalitions and use the resources owned by the coalition to produce output which can then be sold at given prices. Owen places conditions on the model – specifically linear production functions – that ensure non-emptiness of the core of the derived game, whose coalitions consist of owners of resources. From the fundamental theorem of linear programming, associated with any point in the core of the game there is a price vector for resources, which is analogous to a competitive equilibrium price vector for resources except that the budget constraint need not be satisfied by individuals but instead only by coalitions. Owen demonstrates that, when the economy is replicated, the core converges to the set of Owen equilibrium prices. The Owen set and the Owen equilibrium prices have been studied in a number of papers – for example, Kalai and Zemel (1982), Samet and Zemel (1984), Granot (1986) and Gellekom et al. (2000). (There is also some relationship to the literature on oligopoly and cost-sharing; see, for example, Sharkey, 1990, and Tauman, Urbano and Watanabe, 1997.)

It is easy to interpret the resources in Owen's model as attributes of individuals, such as their intelligence, skill level, wealth, ability to dance the tango, and so on. (Of course, labour is typically an input into a production process.) We can also easily interpret a coalition that forms as a club. For example, the club may be a dinner club, where each person brings himself – his personality, his gender, and so on – and also perhaps contributes a dish for the meal. The benefits to membership in a club depend on the attributes of its members – whether they are charming, whether they are good cooks. A difficulty in applying Owen's model to economies with clubs, jurisdictions, or any sort of essential group activity is that his results require linearity of the production function. However, as Owen remarks, concavity of preferences and production possibilities, as in Debreu and Scarf (1963), suffices for all his results except uniqueness of Owen equilibrium prices. But the concavity of limiting per capita payoff functions under the conditions of essential superadditivity and small group effectiveness of Wooders (1983; 1994a; 1994b) implies that in large games with clubs or coalitional activities the economy is representable as a market economy where individuals have concave preferences. Essential superadditivity simply allows a set of players to partition itself and achieve the outcomes achievable by the collective activities of the members of each element of the partition. Finiteness of the supremum of per capita payoffs (per capita boundedness) rules out average (per individual player) payoff from becoming infinitely large. Recent research investigates the relationship between club economies and games in more detail (see, for recent surveys, Wooders, 1994b; Kovalenkov and Wooders, 2005a; Conley and Smith, 2005).

Closely related in important ways to the market approach are approaches that assume cooperative behaviour on the part of members of the coalitions that form. As in the market approach, what a coalition can achieve is taken as defined, a solution concept assumed (which in some cases includes a

partition of the set of players into groups that can achieve their part of the outcome), and the existence and properties of outcomes satisfying the requirements of the solution concept are examined. Classic contributions to this literature, besides those mentioned above, include Aumann and Maschler (1964), Aumann and Shapley (1974), Shapley (1971), and Hart and Kurz (1983). More recent contributions include, among others, Demange (1994), Bogomolnaia and Jackson (2002), Banerjee, Konishi and Sonmez (2001), Le Breton, Ortuno-Ortin and Weber (2006), and Bogomolnaia et al. (2007). These interesting works deepen insight into the question of conditions on models ensuring there is some outcome satisfying the requirements of solutions having desirable properties, especially the core. Necessary and sufficient conditions for non-emptiness of cores are demonstrated by Bondareva (1963) and Shapley (1967) for games with transferable utility and, most recently, by Predtetchinski and Herings (2004) and Bonnisseau and Iehle (2007) for non-transferable utility games.

A small but growing literature, initiated by the assignment games of Gale and Shapley (1962), Shapley and Shubik (1972) and Aumann and Drèze (1974), addresses the question of what conditions on permissible coalition structures will ensure that a game has a non-empty cores, independently of the sets of attainable outcomes of the game. Early papers providing such conditions are Kaneko and Wooders (1982) and Le Breton, Owen and Weber (1992). Recent papers have treated sufficient conditions for non-emptiness of the core of a hedonic game, where preferences are defined directly over coalitions (Bogomolnaia and Jackson, 2002; Banerjee, Konishi and Sonmez, 2001; Papai, 2004) while Iehle (2006) provides necessary and sufficient conditions. Demange (2004) demonstrates that imposing a hierarchical structure on the set of players, limiting the coalitions that can form, will ensure existence of an efficient outcome that is stable in the sense that no admissible coalition, called a team, could improve upon the outcome for its members. A hierarchical structure is represented by a pyramidal network. A team is a group of individuals who can communicate through the channels created by the hierarchical structure.

A related branch of literature focuses on conditions ensuring that groups of agents do not break away from a coalition. Le Breton and Weber (2001), Haimanko, Le Breton and Weber (2004), and Drèze, Le Breton and Weber (2007) investigate models with heterogeneous individuals and conditions ensuring existence of secession-proof outcomes, that is, outcomes that are immune to breakaways by subgroups of individuals and are thus in the core. For a different approach motivated by the idea that if a group secedes from a larger group then it does not necessarily stand alone, see Reny and Wooders (1996b), who use the solution concept of the partnered core. See also Alesina and Spolaore (1997) who demonstrate that, in a model of public good provision with a continuum of consumers who are differentiated by their preferred location for a facility and voting within each community, in equilibrium there are too many coalitions (nations).

Non-cooperative game approach

Coalitions can arise as equilibrium outcomes of either static or dynamic non-cooperative games. In the non-cooperative literature on clubs or local public goods, it may be assumed that there is a fixed set of jurisdictions, each providing some level of a public good for its residents. Individuals who move to a jurisdiction pay the average cost of public good provision. Alternatively, individuals may be required to pay a proportion of their income towards financing the public good produced by the jurisdiction.

Individuals each chose a jurisdiction in which to live. The main questions are whether a non-cooperative equilibrium (Nash equilibrium in pure strategies) exists and its properties, such as whether, in equilibrium, members of the same jurisdiction have similar wealths. Contributions to this literature include Greenberg and Weber (1986), Demange (1994), Konishi, Le Breton and Weber (1997; 1998), Gravel and Thoron (2007). See also Demange (2005), who discusses literature involving both cooperative and non-cooperative approaches. Based on the concept of coalition-proofness (Bernheim, Peleg and Whinston, 1987) Conley and Konishi (2002) obtain existence of an efficient, migration-proof equilibrium for local public good (club) economies with many but a finite number of players. Casella (1992) and Casella and Feinstein (2002) consider the effects of the possibilities of trade in private goods in the formation of clubs/jurisdictions.

In a number of papers on dynamic games of coalition formation, a payoff set is given for each coalition. Suppose for simplicity that, for each coalition *S*, there is a unique attainable payoff vector $\{x^i(S) : i \in S\}$. If players are randomly ordered and if according to the ordering each player lists those players he would like as members of his coalition, then one possible solution to such a game of non-cooperative coalition formation would be a partition of the total player set into coalitions where for each coalition *S* in the partition the members of *S* all choose *S* and each player $i \in S$ receives the payoff $x^i(S)$. If player *i* belongs to no such coalition, then he receives some default payoff $x^i(\{i\})$. This sort of approach was introduced in Selten (1981). Perry and Reny (1994) provide a non-cooperative implementation of the core for TU games. In the Perry–Reny model proposed, time is continuous. This ensures that there is always time to reject a non-core proposal before it is consummated. Which coalitions will form typically depends crucially on the rules of the game. The Perry–Reny implementation is meant to reflect the standard motivation for the core as closely as possible. Hart and Mas-Colell (1996) implement the *consistent value* (Maschler and Owen, 1992) for NTU games, which, for TU games, is equivalent to the Shapley value. Bloch (1996) treats games where, as in the Lucas–Thrall model, the payoff achievable by a group of players may depend on the entire coalition structure of the remaining players. Ray and Vohra (1997; 1999) study coalitional agreements and coalitional bargaining in partition function games. See Bandyopadhyay and Chatterjee (2006) for a survey of coalition formation based on non-cooperative bargaining. See also Myerson (1995), Seidmann and Winter (1998), Mauleon and Vannetelbosch (2004), among others.

Networks and coalition formation

Because networks allow for a detailed specification of interactions between individuals and between coalitions, abstract games over networks have a greater potential to capture the subtleties of bargaining and negotiation than do the abstract coalitional form games of von Neumann–Morgenstern and Gillies and Shapley. A seminal contribution to this line of research is the paper by Myerson (1977). Myerson begins by assuming that the worth of each possible coalition depends on the structure of cooperation between individuals as given by a graph where nodes represent individuals and links between nodes represent interactions between individuals. As in much of the subsequent literature Myerson imposes an

allocation rule, a rule specifying how the worth of a coalition is to be shared among its members. The worth of any connected (linked) set of players is divided according to the rule. The specific rule chosen by Myerson is a variant of the Shapley value, now known as the Myerson value. As Myerson shows, this is the only rule satisfying both component efficiency (in sum, the members of each component of the network receive the worth of that component as a coalition) and a fairness property that requires any two players to benefit equally from the formation of a link. Aumann and Myerson (1988) work with extensive form games, where players choose links strategically and allow players to look ahead and to take into account the end effects of their actions. In their model, once a link is formed, it cannot be broken. The equilibrium concept is non-cooperative subgame perfection. Once players have formed links, the payoffs to players are determined by the Myerson value.

Jackson and Wolinsky (1996) also treat link formation between individual players. A network satisfies their pairwise stability condition if no two players could benefit by creating a link between them and no one player could benefit by cutting a link with another player. Based on the Jackson–Wolinsky model, numerous papers have now looked at costs and benefits of link formation between players and equilibrium outcomes; see Dutta, van den Nouweland, and Tijs (1998) for example, and van den Nouweland (2005) for some recent results and a review. Herings, Mauleon and Vannetelbosch (2006) introduce notions of pairwise farsighted stability. Jackson and van den Nouweland (2005) introduce the concept of a strongly stable network. A network is strongly stable if no coalition could benefit by making changes (additions or deletions) to the links of coalition members. As Jackson and van den Nouweland show, the existence of strongly stable networks is equivalent to non-emptiness of the core in a derived cooperative game. See also Jackson and Watts (2002), who use linking networks and stochastic dynamics to study the evolution of networks.

Other recent works addressing questions of coalition formation in networks make assumptions concerning what a coalition believes it can achieve. These contributions include Watts (2001), who assumes that dominance must be direct, in the sense that a coalition will act to change a network from g to g' only if it perceives an immediate gain. In contrast, Page, Wooders and Kamat (2005) consider indirect dominance where a network g dominates another network g' if there is a coalition S that believes it can trigger a series of changes beginning with the network g and ending with the network g' that is preferred by all members of S. Whether dominance is direct or indirect is of crucial importance, as illustrated in Diamantoudi and Xue (2003) and Page and Wooders (2007), among others. Consider, for example, a situation with two jurisdictions, say J_1 and J_2 and seven people. Each person would like to live in the jurisdiction with the fewest residents. With direct dominance, any partition of the people between the two jurisdictions with three people in one jurisdiction and four in the other is stable. In contrast, with indirect dominance, the situation changes; players can be more optimistic. Suppose that initially there are four people in jurisdiction J_1 and three in J_2. Two people in J_1 may move into J_2 in the belief that, since J_2 has become so crowded, three people will leave J_2 and move to J_1, with the result that the two initial movers will be better off.

Using supernetworks, introduced in Page, Wooders and Kamat (2005), where nodes represent networks and directed arcs represent coalitional moves and coalitional preferences, networks can also provide a simple representation of the rules of network formation and hence the rules of coalition formation. Network formation rules play a crucial role in determining coalitional outcomes. To illustrate, in the

literature on markets and on cooperative games, it is assumed that coalitions can exclude individuals. It may be, however, that groups (or coalitions) are subject to ‘free entry’ – any group of players can freely join another group without the consent of those being joined. This has long been important in the literature on economies with clubs/local public goods; compare, for example, the models of Konishi, Le Breton and Weber (1998) and Demange (1994) with that of Conley and Wooders (2001). As a special case, networks can also accommodate a systematic analysis of coalition formation and payoff division when there are potential irreversibilities. For example, given the informational environment, it may be that the only coalitions which can form are sub-coalitions of existing coalitions. Or the rules of network formation may not allow cycles.

How to define a coalition

The traditional approach of cooperative game theory models a coalition as an alliance of players who take as given a well-defined set of possible outcomes or payoffs. The alliance, when considering whether to ‘block’ a proposed outcome, is faced with the alternative of standing alone. In reality, however, we observe that individuals belong to multiple, possibly overlapping alliances. This fact has received remarkably little attention in the literature. Some papers in the club literature allow individuals to belong to multiple clubs for the purposes of local public good provision and private good production within each club, including Shubik and Wooders (1982), Ellickson et al. (2001) and Allouch and Wooders (2006). Roughly, if there is only a finite set of sorts of clubs, bounded in size, (Ellickson et al.) or if ‘per capita payoffs’ are bounded (Allouch and Wooders), then in large economies the core and the set of price taking equilibrium outcomes are equivalent. An interesting application of the idea of overlapping coalitions is developed in Conconi and Perroni (2002), who assume that a country can enter into different alliances, where each alliance to which it belongs is concerned with a different issue.
The definition of a coalition also becomes an issue when the total player set is an atomless continuum. There are two approaches. One approach, introduced in Aumann (1964), is to model a coalition as a subset of positive measure. Major theorems using this approach and relating to coalitions demonstrate equivalence of the core and outcomes of price-taking equilibrium of models of economies Another approach is to describe a coalition as a finite set of players, as in Keiding (1976). This has the advantage that individuals may interact with other individuals, and permits matching or marriage models, for example. An obvious difficulty with such an approach is that, at the heart of economics, is the problem of relative scarcities. Think of the diamond–water paradox; even though water is essential for life itself, it is abundant and thus inexpensive, while diamonds are relatively inessential but scarce and thus expensive.
To see the difficulty in retaining relative scarcities while allowing finite coalitions, suppose, for example, that the points in the interval [0,2] represent boys and the points in the interval [3,4] represent girls so that there are ‘twice’ as many boys as girls. Suppose the only effective coalitions consist of either boy, girl pairs (i, j) where $i \in [0, 2]$ and $j \in [3, 4]$, or singletons – a matching model. Consider the set of coalitions $\left\{(i, j) : j = 3 + \frac{1}{2} i\right\}$; this set describes a partition of the total player set and marries each boy to a girl; clearly this partition is not consistent with the relative scarcities given by Lebesgue measure. Indeed, since there are one-to-one mappings of a set of positive measure onto a set of measure

zero, it is even possible to have partitions of the total player set into boy–girl pairs and singletons that match each boy to a girl while leaving a set of girls of measure 1 unmatched! A solution to this problem was proposed by Kaneko and Wooders (1986) with the introduction of *measurement-consistent* partitions. A simple formulation of measurement consistency has recently been provided (Allouch, Conley and Wooders, 2006), and we use it here. Define an *index set* for a partition of a continuum of players as one member from each element of the partition. A partition of players into finite coalitions is 'measurement-consistent' if every index set for the partition has the same measure. The partition given above is not measurement-consistent while the partition $\{(i, j): j = 3 + i, \quad i \in [0, 1] \cup \{\{i\}: i \in (1, 2]\}\}$ is measurement-consistent. While in models of exchange economies, the core with finite coalitions (the *f*-core) and the Aumann core yield equivalent outcomes, in the presence of widespread externalities, such as global pollution, the *f*-core coincides with the set of competitive equilibrium prices while the Aumann core may be empty and, even if non-empty, may have an empty intersection with the set of equilibrium outcomes; the concepts of the Aumann core and the *f*-core are distinct with the *f*-core apparently most closely related to the set of competitive equilibrium prices (Kaneko and Wooders, 1986; Hammond, Kaneko and Wooders, 1989; Kaneko and Wooders 1994). Other works using the *f*-core approach include Berliant and Edwards (2004) and Legros and Newman (1996; 2002). These papers illustrate the advantage of the *f*-core approach in that it enables analysis of activities within groups (firms or clubs, or other organizations) that may contain any finite number of individuals but are negligible relative to the entire economy.

An interesting difference between the Aumann-core and the *f*-core is that, while the Aumann-core has been axiomatized by Dubey and Neyman (1984), the authors stress that the axiomatization is completely different than axiomatizations for the core in cooperative games with only a finite number of players. In contrast, Winter and Wooders (1994) provide an axiomatization for the core of a game with finite coalitions that applies whether the player set is finite or an atomless continuum.

Conclusions

This article began with some of the first works on coalitions in the literature of game theory and concluded with recent work on coalitions and networks. It becomes apparent that the concepts of early works underlie much of even the most recent research. We see at least a part of the future of coalition theory in network modelling of socio-economic coalitions and in more behavioural approaches to coalition theory, involving 'implicit' and 'tacit' coalitions. Language and the ability to communicate well are clearly involved; see multilingualism and references there. Instead of being bound together by commitments and contracts, members of an implicit coalition may be bound together by common language, culture, objectives or by common group memberships and, even though there may be no explicit agreement, members of an implicit coalition might act together, as if they were a coalition. This raises questions of to what extent individuals, who share common group memberships as in Durlauf (2002) for example, are an implicit coalition and whether such individuals have tendencies to form more explicit coalitions. While much has been done on coalitions, there remains much to do.

See Also

- bargaining
- core convergence
- game theory
- multilingualism
- network formation

We are grateful to Harold Kuhn for his generous assistance in tracking down the origins of the concept of the core.

Bibliography

Agastya, M. 1999. Perturbed adaptive dynamics in coalition form games. *Journal of Economic Theory* 89, 207–33.

Albers, W. 1979. Core-and-kernel variants based on imputations and demand profiles. In *Game Theory and Related Topics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.

Alesina, A. and Spolaore, E. 1997. On the number and size of nations. *Quarterly Journal of Economics* 112, 1027–56.

Allen, B. 1992. Incentives in market games with asymmetric information: the core. CORE Discussion Paper No. 9221, Université Catholique de Louvain.

Allen, B. 2006. Market games with asymmetric information: the core. *Economic Theory* 29, 465–87.

Allouch, N., Conley, J. and Wooders, M. 2006. Anonymous price taking equilibrium in Tiebout economies with a continuum of agents; existence and characterization. Working paper, Department of Economics, Vanderbilt University.

Allouch, N. and Wooders, M. 2006. Price taking equilibrium in club economies with multiple memberships and unbounded club sizes. Working paper, Department of Economics, Vanderbilt University.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Aumann, R.J. and Drèze, J. 1974. Cooperative games with coalition structures. *International Journal of Game Theory* 3, 217–37.

Aumann, R.J. and Maschler, M. 1964. The bargaining set for cooperative games. In *Advances in Game Theory, Annals of Mathematics Study* 52, ed. M. Dresher, L.S. Shapley and A.W. Tucker. Princeton:

Princeton University Press.

Aumann, R.J. and Myerson, R. 1988. Endogenous formation of links between players and of coalitions: an application of the Shapley value. In *The Shapley Value*, ed. A.E. Roth. Cambridge: Cambridge University Press.

Aumann, R.J. and Peleg, B. 1960. Von Neumann–Morgenstern solutions to cooperative games without side payments. *Bulletin of the American Mathematical Society* 66, 173–9.

Aumann, R.J. and Shapley, L.S. 1974. *Values of Non-Atomic Games*. Princeton: Princeton University Press.

Bag, P. and Winter, E. 1999. Simple subscription mechanisms for excludable public goods. *Journal of Economic Theory* 87, 72–94.

Bandyopadhyay, S. and Chatterjee, K. 2006. Coalition theory and applications: a survey. *Economic Journal* 116, 136–56.

Banerjee, S., Konishi, H. and Sonmez, T. 2001. Core in a simple coalition formation game. *Social Choice and Welfare* 18, 135–58.

Barberà, S. and Gerber, A. 2003. On coalition formation: durable coalition structures. *Mathematical Social Sciences* 45, 185–203.

Bennett, E. 1983. The aspiration approach to predicting coalition formation in side payments games. *International Journal of Game Theory* 12, 1–28.

Berliant, M. and Edwards, J.H.Y. 2004. Efficient allocations in club economies. *Journal of Public Economic Theory* 6, 43–63.

Bernheim, D., Peleg, B. and Whinston, M. 1987. Coalition proof Nash equilibria I: concepts. *Journal of Economic Theory* 42, 1–12.

Bloch, F. 1995. Endogenous structures of association in oligopolies. *RAND Journal of Economics* 26, 537–56.

Bloch, F. 1996. Sequential formation of coalitions in games with fixed payoff division and externalities. *Games and Economic Behavior* 14, 90–123.

Bloch, F. 2005. Group and network formation in industrial organization: a survey. In *Group Formation in Economics; Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge:

Cambridge University Press.

Bogomolnaia, A. and Jackson, M.O. 2002. The stability of hedonic coalition structures. *Games and Economic Behavior* 38, 201–30.

Bogomolnaia, A., Le Breton, M. Savvateev, A. and Weber, S. 2007. Stability of jurisdiction structures under the equal share and median rules. *Economic Theory*.

Bondareva, O. 1963. Some applications of linear programming to the theory of cooperative games [in Russian]. *Problemy kibernetiki* 10. English translation by K. Takeuchi and E. Wesley in *Selected Russian Papers on Game Theory 1959–1965*, ed. L. Billera, D. Cohen, and R. Cornwall. Princeton: Princeton University Press, 1968.

Bonnisseau, J-M. and Iehle, V. 2007. Payoff-dependent balancedness and core. *Games and Economic Behavior* 61, 1–26.

Bossert, W. and Sprumont, Y. 2002. Core rationalizability in two-agent exchange economies. *Economic Theory* 20, 777–91.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* 33, 1–14.

Burani, N. and Zwicker, W. 2003. Coalition formation games with separable preferences. *Mathematical Social Sciences* 45, 27–52.

Carraro, C. 2005. Institution design for managing global commons: lessons from coalition theory. In *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Casella, A. 1992. On markets and clubs: economic and political integration of regions with unequal productivity. *American Economic Review* 82, 115–21.

Casella, A. and Feinstein, J. 2002. Public goods in trade: on the formation of markets and jurisdictions. *International Economic Review* 43, 437–62.

Chander, P. 2007. The gamma-core and coalition formation. *International Journal of Game Theory* (forthcoming).

Chander, P. and Tulkens, H. 1995. A core-theoretic solution for the design of cooperative agreements on transfrontier pollution. *International Tax and Public Finance* 2, 279–93.

Chander, P. and Tulkens, H. 1997. The core of an economy with multilateral environmental

externalities. *International Journal of Game Theory* 26, 379–401.

Chatterjee, K., Dutta, B., Ray, D. and Sengupta, K. 1993. A non-cooperative theory of coalitional bargaining. *Review of Economic Studies* 60, 463–77.

Chwe, M.S-Y. 1994. Farsighted coalitional stability. *Journal of Economic Theory* 63, 299–325.

Conconi, P. and Perroni, C. 2002. Issue linkage and issue tie-in in multilateral negotiations. *Journal of International Economics* 57, 423–47.

Conley, J.P. and Konishi, H. 2002. Migration-proof Tiebout equilibrium: existence and asymptotic efficiency. *Journal of Public Economics* 86, 243–62.

Conley, J. and Smith, S. 2005. Coalitions and clubs: Tiebout equilibrium in large economies. In *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Conley, J. and Wooders, M. 2001. Tiebout economics with differential genetic types and endogenously chosen crowding characteristics. *Journal of Economic Theory* 98, 261–94.

Davis, M. and Maschler, M. 1963. Existence of stable payoff configurations for cooperative games. *Bulletin of the American Mathematical Society* 69, 106–8.

Davis, M. and Maschler, M. 1965. The kernel of a cooperative game. *Naval Research Logistics Quarterly* 12, 223–59.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

De Clippel, G. and Minelli, E. 2005. Two remarks on the inner core. *Games and Economic Behavior* 50, 143–54.

Demange, G. 1994. Intermediate preferences and stable coalition structures. *Journal of Mathematical Economics* 23, 45–8.

Demange, G. 2004. On group stability in hierarchies and networks. *Journal of Political Economy* 112, 754–78.

Demange, G. 2005. The interaction of increasing returns and preferences diversity. In *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Demange, G. and Guesnerie, R. 2001. On coalitional stability of anonymous interim mechanisms. *Economic Theory* 18, 367–89.

Diamantoudi, E. and Xue, L. 2003. Farsighted stability in hedonic games. *Social Choice and Welfare* 21, 39–61.

Drèze, J., Le Breton, M. and Weber, S. 2006. Rawlsian pricing of access to public facilities: a unidimensional illustration. *Journal of Economic Theory* 136, 759–66.

Dubey, P. and Neyman, A. 1984. Payoffs in nonatomic economies: an axiomatic approach. *Econometrica* 52, 1129–50.

Durlauf, S. 2002. The memberships theory of poverty: the role of group affiliations in determining socioeconomic outcomes. In *Understanding Poverty in America*, ed. S. Danziger and R. Haveman. Cambridge, MA: Harvard University Press.

Dutta, B., van den Nouweland, A. and Tijs, S. 1998. Link formation in cooperative situations. *International Journal of Game Theory* 27, 245–56.

Dutta, B. and Vohra, R. 2005. Incomplete information, credibility and the core. *Mathematical Social Sciences* 50, 148–65.

Echenique, F. and Oviedo, J. 2004. Core many-to-one matchings by fixed-point methods. *Journal of Economic Theory* 115, 358–76.

Economides, N.S. 1986. Non-cooperative equilibrium coalition structures. University Discussion Paper No. 273, Columbia University.

Einy, E., Moreno, D. and Monderer, D. 1999. On the least core and the Mas-Colell bargaining set. *Games and Economic Behavior* 28, 181–8.

Einy, E., Moreno, D. and Shitovitz, B. 2000. On the core of an economy with differential information. *Journal of Economic Theory* 94, 262–70.

Ellickson, B., Grodal, B. Scotchmer, S. and Zame, W. 2001. Clubs and the market: large finite economies. *Journal of Economic Theory* 101, 40–77.

Engelbrecht-Wiggans, R. and Granot, D. 1985. On market prices in linear production games. *Mathematical Programming* 32, 366–70.

Epstein, L. and Marinacci, M. 2001. The core of large differentiable TU games. *Journal of Economic*

Theory 100, 235–73.

Farrell, J. and Scotchmer, S. 1988. Partnerships. *Quarterly Journal of Economics* 103, 279–97.

Forges, F. Mertens, J.-F. and Vohra, R. 2002. The ex ante incentive compatible core in the absence of wealth effects. *Econometrica* 70, 1865–92.

Forges, F., Minelli, E. and Vohra, R. 2002. Incentives and the core of an exchange economy: a survey. *Journal of Mathematical Economics* 38, 1–41.

Gale, D. and Shapley, L.S. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69, 9–15.

Garratt, R. and Qin, C-Z. 1997. On a market for coalitions with indivisible agents and lotteries. *Journal of Economic Theory* 77, 81–101.

Gellekom, J.R.G., Potters, J.A.M., Reijnierse, J.H., Engel, M.C. and Tijs, S.H. 2000. Characterization of the Owen set of linear production processes. *Games and Economic Behaviour* 32, 139–56.

Gillies, D.B. 1959. Solutions to general non-zero-sum games. In *Contributions to the Theory of Games*, vol. 4, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

Granot, D. 1986. A generalized linear production model: a unifying model. *Mathematical Programming* 34, 212–22.

Gravel, N. and Thoron, S. 2007. Does endogenous formation of jurisdictions lead to wealth-stratification? *Journal of Economic Theory* 132, 569–83.

Greenberg, J. 1995. Coalition structures. In *Handbook of Game Theory with Economic Applications*, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Greenberg, J. and Weber, S. 1986. Strong Tiebout equilibrium under restricted preferences domain. *Journal of Economic Theory* 38, 101–17.

Greenberg, J. and Weber, S. 1993. Stable coalition structures with a unidimensional set of alternatives. *Journal of Economic Theory* 60, 62–82.

Hahn, G. and Yannelis, N. 1997. Efficiency and incentive compatibility in differential information economies. *Economic Theory* 10, 383–411.

Haimanko, O., Le Breton, M. and Weber, S. 2004. Voluntary formation of communities for the

provision of public projects. *Journal of Economic Theory* 115, 1–34.

Hammond, P., Kaneko, M. and Wooders, M. 1989. Continuum economies with finite coalitions: Core, equilibria, and widespread externalities. *Journal of Economic Theory* 49, 113–34.

Harsanyi, J. 1974. An equilibrium point interpretation of stable sets and a proposed alternative definition. *Management Science* 20, 1472–95.

Hart, S. and Kurz, M. 1983. Endogenous formation of coalitions. *Econometrica* 51, 1047–64.

Hart, S. and Mas-Colell, A. 1996. Bargaining and value. *Econometrica* 64, 357–80.

Herings, J-J., Mauleon, A. and Vannetelbosch, V. 2006. Farsightedly stable networks. RM/06/041, University of Maastricht.

Herings, J-J., van der Laan, G. and Talman, D. 2007. The socially stable core in structured transferable utility games. *Games and Economic Behavior* 59, 85–104.

Holmstrom, B. and Myerson, R. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51, 1799–819.

Ichiishi, T. and Idzik, A. 1996. Bayesian cooperative choice strategies. *International Journal of Game Theory* 25, 455–73.

Iehle, V. 2006. The core partition of a hedonic game. Working paper, Maison des Sciences Economiques, Université Panthéon-Sorbonne.

Izquierdo, J.M. and Rafels, C. 2001. Average monotonic cooperative games. *Games and Economic Behavior* 36, 174–92.

Jackson, M. and van den Nouweland, A. 2005. Strongly stable networks. *Games and Economic Behavior* 51, 420–44.

Jackson, A. and Watts, A. 2002. The evolution of social and economic networks. *Journal of Economic Theory* 106, 265–95.

Jackson, A. and Wolinsky, A. 1996. A strategic model of social and economic networks. *Journal of Economic Theory* 71, 44–74.

Jordan, J.S. 2006. Pillage and property. *Journal of Economic Theory* 131, 26–44.

Kajii, A. 1992. A generalization of Scarf's theorem: An α-core existence theorem without transitivity or completeness. *Journal of Economic Theory* 56, 194–205.

Kalai, E., Postlewaite, A. and Roberts, J. 1979. A group incentive compatible mechanism yielding core allocations. *Journal of Economic Theory* 20, 13–22.

Kalai, E. and Schmeidler, D. 1977. An admissible set occurring in various bargaining situations. *Journal of Economic Theory* 14, 402–11.

Kalai, E. and Zemel, E. 1982. Generalized network problems yielding totally balanced games. *Operations Research* 30, 998–1008.

Kaneko, M. and Wooders, M. 1982. Cores of partitioning games. *Mathematical Social Sciences* 3, 313–27.

Kaneko, M. and Wooders, M. 1986. The core of a game with a continuum of players and finite coalitions: the model and some results. *Mathematical Social Sciences* 12, 105–37.

Kaneko, M. and Wooders, M. 1989. The core of a continuum economy with widespread externalities and finite coalitions: from finite to continuum economics. *Journal of Economic Theory* 49, 135–68.

Kaneko, M. and Wooders, M. 1994. Widespread externalities and perfectly competitive markets. In *Imperfections and Behavior in Economic Organizations*, ed. R. Gilles and P. Ruys. Boston: Kluwer Academic Publishers.

Kaneko, M. and Wooders, M. 1996. The nonemptiness of the *f*-core of a game without side payments. *International Journal of Game Theory* 25, 245–58.

Keiding, H. 1976. Cores and equilibria in an infinite economy. In *Computing Equilibrium: How and Why*, ed. J. Los and M.W. Los. Amsterdam: North-Holland.

Kóczy, Á. and Lauwers, L. 2004. The coalition structure core is accessible. *Games and Economic Behavior* 48, 86–93.

Konishi, H., Le Breton, M. and Weber, S. 1997. Equilibrium in a model with partial rivalry. *Journal of Economic Theory* 72, 225–37.

Konishi, H., Le Breton, M. and Weber, S. 1998. Equilibrium in a finite local public goods economy. *Journal of Economic Theory* 79, 224–44.

Konishi, H. and Ray, D. 2003. Coalition formation as a dynamic process. *Journal of Economic Theory*

110, 1–41.

Kovalenkov, A. and Wooders, M. 2001a. Epsilon cores of games with limited side payments; Nonemptiness and equal treatment. *Games and Economic Behavior* 36, 193–218.

Kovalenkov, A. and Wooders, M. 2001b. An exact bound on epsilon for non-emptiness of the epsilon-cores of games. *Mathematics of Operations Research* 26, 654–78.

Kovalenkov, A. and Wooders, M. 2003. Approximate cores of games and economies with clubs. *Journal of Economic Theory* 110, 87–120.

Kovalenkov, A. and Wooders, M. 2005a. Many sided matchings, clubs and market games. In *Group Formation in Economics; Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Kovalenkov, A. and Wooders, M. 2005b. Laws of scarcity for a finite game – exact bounds on estimations. *Economic Theory* 26, 383–96.

Kovalenkov, A. and Wooders, M. 2006. Comparative statics and laws of scarcity for games. In *Rationality and Equilibrium; A Symposium in Honor of Marcel K. Richter*, ed. C.D. Aliprantis et al. Berlin: Springer-Verlag.

Kurz, M. 1989. Game theory and public economics. In *Handbook of Game Theory*, vol 2, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Le Breton, M., Ortuno-Ortin, I. and Weber, S. 2006. Gamson's Law and hedonic games. Working Paper No. 420, IDEI, Toulouse.

Le Breton, M., Owen, G. and Weber, S. 1992. Strongly balanced cooperative games. *International Journal of Game Theory* 20, 419–27.

Le Breton, M. and Weber, S. 2001. The art of making everybody happy: how to prevent a secession. Working Paper No. 01/176, International Monetary Fund.

Lee, D. and Volij, O. 2000. The core of economies with asymmetric information: an axiomatic approach. *Journal of Mathematical Economics* 38, 43–63.

Legros, P. and Newman, A.F. 2002. Monotone matching in perfect and imperfect worlds. *Review of Economic Studies* 69, 925–42.

Legros, P. and Newman, A.F. 1996. Wealth effects, distribution, and the theory of organization. *Journal*

of Economic Theory 70, 312–341.

Logistics Research Project. 1957. Reports of three informal conferences on the theory of games. Department of Mathematics, Princeton University. Accession No. (OCoLC)ocm39726894. Conferences held 20–1 March, 1953; 31 January–1 February 1955, and 11–12 March, 1957, Princeton.

Lucas, W. and Thrall, R. 1963. n-person games in partition function form. *Naval Research Logistics Quarterly* 10, 281–98.

Maschler, M. and Owen, G. 1992. The consistent Shapley value for games without side payments. In *Rational Interaction*, ed. R. Selten. Berlin: Springer-Verlag.

Maschler, M. and Peleg, B. 1966. A characterization, existence proof and dimensions bounds for the kernel of a game. *Pacific Journal of Mathematics* 18, 289–328.

Maschler, M. and Peleg, B. 1967. The structure of the kernel of a cooperative game. *SIAM Journal of Applied Mathematics* 15, 569–604.

Maschler, M., Peleg, B. and Shapley, L.S. 1971. The kernel and bargaining set for convex games. *International Journal of Game Theory* 1, 73–93.

Mas-Colell, A. 1989. An equivalence theorem for a bargaining set. *Journal of Mathematical Economics* 18, 129–39.

Mauleon, A. and Vannetelbosch, V.J. 2004. Farsightedness and cautiousness in coalition formation games with positive spillovers. *Theory and Decision* 56, 291–324.

McLean, R. and Postlewaite, A. 2005. Core convergence with asymmetric information. *Games and Economic Behavior* 50, 58–78.

Milchtaich, I. 1996. Congestion games with player-specific payoff functions. *Games and Economic Behavior* 13, 111–24.

Milchtaich, I. and Winter, E. 2002. Stability and segregation in group formation. *Games and Economic Behavior* 38, 318–46.

Milleron, J.C. 1972. Theory of value with public goods: a survey article. *Journal of Economic Theory* 5, 419–77.

Monderer, D. and Shapley, L.S. 1996. Potential games. *Games and Economic Behavior* 14, 124–43.

Muench, T. 1972. The core and the Lindahl equlibrium of an economy with public goods; an example. *Journal of Economic Theory* 4, 241–55.

Myerson, R.B. 1977. Graphs and cooperation in games. *Mathematics of Operations Research* 2, 225–9.

Myerson, R.B. 1995. Sustainable matching plans with adverse selection. *Games and Economic Behavior* 9, 35–65.

Myerson, R.B. 2007. Virtual utility and the core for games with incomplete information. *Journal of Economic Theory* 136, 260–85.

Nash, J. 1953. Two-person cooperative games. *Econometrica* 21, 12840.

Owen, G. 1975. On the core of linear production games. *Mathematical Programming* 9, 358–70.

Owen, G. 1977. Values of games with a priori unions. In *Essays in Mathematical Economics and Game Theory*, ed. R. Henn and O. Moeschlin. Berlin, Heidelberg and New York: Springer.

Page, F.H. Jr. 1997. Market games with differential information and infinite dimensional commodity spaces: the core. *Economic Theory* 9, 151–9.

Page, F.H. Jr. and Wooders, M. 1996. The partnered core and the partnered competitive equilibrium. *Economics Letters* 52, 143–52.

Page, F.H. Jr. and Wooders, M. 2006. Strategic basins of attraction, the path dominance core, and network formation games. Working Paper No. 06-W14 (revised), Department of Economics, Vanderbilt University.

Page, F.H. Jr. and Wooders, M. 2007. Networks and clubs. *Journal of Economic Behavior & Organization* (forthcoming).

Page, F.H. Jr., Wooders, M. and Kamat, S. 2005. Networks and farsighted stability. *Journal of Economic Theory* 120, 257–69.

Papai, S. 2004. Unique stability in simple coalition formation games. *Games and Economic Behavior* 48, 337–54.

Pauly, M. 1970. Cores and clubs. *Public Choice* 9, 53–65.

Peleg, B. 1985. The axiomatization of the core of cooperative games without side payments. *Journal of Mathematical Economics* 14, 203–14.

Perry, M. and Reny, P.J. 1994. A noncooperative view of coalition formation and the core. *Econometrica* 62, 795–817.

Predtetchinski, A. and Herings, J.-J. 2004. A necessary and sufficient condition for non-emptiness of the core of a non-transferable utility game. *Journal of Economic Theory* 116, 84–92.

Qin, C.-Z. 1994. The inner core of an n-person game. *Games and Economic Behavior* 6, 431–44.

Ray, D. and Vohra, R. 1997. Equilibrium binding agreements. *Journal of Economic Theory* 73, 30–78.

Ray, D. and Vohra, R. 1999. A theory of endogenous coalition structures. *Games and Economic Behavior* 26, 286–336.

Reny, P.J. and Wooders, M. 1996a. The partnered core of a game without side payments. *Journal of Economic Theory* 70, 298–311.

Reny, P.J. and Wooders, M. 1996b. Credible threats of secession, partnership, and commonwealths. In *Understanding Strategic Interaction: Essays in Honor of Reinhard Selten*, ed. W. Albers et al. Berlin: Springer-Verlag.

Roth, A.E. 1984. Stable coalition formation: aspects of a dynamic theory. In *Coalitions and Collective Action*, ed. M. Holler. Würzberg: Physica-Verlag.

Samet, D. and Zemel, E. 1984. On the core and dual set of linear programming games. *Mathematics of Operations Research* 9, 309–16.

Scotchmer, S. and Wooders, M. 1988. Monotonicity in games that exhaust gains to scale. Technical Report No. 525, IMSSS, Stanford University.

Seidmann, D.J. and Winter, E. 1998. Gradual coalition formation. *Review of Economic Studies* 65, 793–815.

Serrano, R. and Vohra, R. 2006. Information transmission in coalitional voting games. *Journal of Economic Theory* (forthcoming).

Selten, R. 1981. A non-cooperative model of characteristic function bargaining. In *Essays in Game Theory and Mathematical Economics in Honor of O. Morgenstern*, ed. V. Böhm and H. Nachtkamp. Mannheim: Bibliographisches Institut.

Shapley, L.S. 1953. A value for n-person games. In *Contributions to the Theory of Games* II, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

Shapley, L.S. 1967. On balanced sets and cores. *Navel Research Logistics Quarterly* 9, 45–8.

Shapley, L.S. 1971. Cores of convex games. *International Journal of Game Theory* 1, 11–26.

Shapley, L.S. and Shubik, M. 1966. Quasi-cores in a monetary economy with nonconvex preferences. *Econometrica* 34, 805–27.

Shapley, L.S. and Shubik, M. 1969a. On the core of an economic system with externalities. *American Economic Review* 59, 678–84.

Shapley, L.S. and Shubik, M. 1969b. On market games. *Journal of Economic Theory* 1, 9–25.

Shapley, L.S. and Shubik, M. 1972. The assignment game 1: the core. *International Journal of Game Theory* 1, 11–30.

Shapley, L.S. and Shubik, M. 1975. Competitive outcomes in the cores of market games. *International Journal of Game Theory* 4, 229–37.

Sharkey, W.W. 1990. Cores of games with fixed costs and shared facilities. *International Economic Review* 31, 245–62.

Shenoy, P.P. 1979. On coalition formation: a game-theoretic approach. *International Journal of Game Theory* 8, 133–64.

Shenoy, P.P. 1980. A dynamic solution concept for abstract games. *Journal of Optimization Theory and Applications* 32, 151–69.

Shubik, M. 1959. Edgeworth market games. In *Contributions to the Theory of Games IV, Annals of Mathematical Studies 40*, ed. F.R. Luce and A. Tucker. Princeton: Princeton University Press.

Shubik, M. 1971. The bridge game economy: an example of indivisibilities. *Journal of Political Economy* 79, 909–12.

Shubik, M. and Wooders, M. 1982. Near markets and market games. Cowles Foundation Discussion Paper No. 657, also published as ‘Clubs, near markets and market games’, *Topics in Mathematical Economics and Game Theory: Essays in Honor of Robert J. Aumann*, ed. M. Wooders. Fields Institute Communication 23. Providence, RI: American Mathematical Society, 2000.

Shubik, M. and Wooders, M. 1983. Approximate cores of replica games and economies: Part I. Replica games, externalities, and approximate cores. *Mathematical Social Sciences* 6, 27–48.

Shubik, M. and Wooders, M. 1983. Approximate cores of replica games and economies: Part II. Set-up costs and firm formation in coalition production economies. *Mathematical Social Sciences* 6, 285–306.

Tauman, Y., Urbano, A. and Watanabe, J. 1997. A model of multiproduct price competition. *Journal of Economic Theory* 77, 377–401.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Tijs, S. 1981. Bounds for the core and the tau-value. In *Game Theory and Mathematical Economics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.

van den Nouweland, A. 2005. Models of network formation in cooperative games. In *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

von Neumann, J. and Morgenstern, O. 1953. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Vohra, R. 1999. Incomplete information, incentive compatibility and the core. *Journal of Economic Theory* 86, 123–47.

Volij, O. 2000. Communication, credible improvements and the core of an economy with asymmetric information. *International Journal of Game Theory* 29, 63–79.

Watts, A. 2001. A dynamic model of network formation. *Games and Economic Behavior* 34, 331–41.

Weber, S. 1979. On ε -cores of balanced games. *International Journal of Game Theory* 8, 241–50.

Weber, S. 1981. Some results on the weak core of a non-sidepayment game with infinitely many players, *Journal of Mathematical Economics* 8, 101–11.

Weber, S. and Wiesmeth, H. 1991. The equivalence of core and cost share equilibria in an economy with a public good. *Journal of Economic Theory* 54, 180–97.

Wilson, R. 1978. Information, efficiency and the core of an economy. *Econometrica* 46, 807–16.

Winter, E. and Wooders, M. 1994. An axiomatization of the core for finite and continuum games. *Social Choice and Welfare* 11, 165–75.

Wooders, M. 1978. Equilibria, the core, and jurisdiction structures in economies with a local public good. *Journal of Economic Theory* 18, 328–48.

Wooders, M. 1980. The Tiebout hypothesis: near optimality in local public good economies. *Econometrica* 48, 1467–86.

Wooders, M. 1983. The epsilon core of a large replica game. *Journal of Mathematical Economics* 11, 277–300.

Wooders, M. 1994a. Equivalence of games and markets. *Econometrica* 62, 1141–60.

Wooders, M. 1994b. Large games and economies with effective small groups. In *Game-Theoretic Methods in General Equilibrium Analysis*, ed. J.-F. Mertens and S. Sorin. Dordrecht: Kluwer Academic Publishers.

Wooders, M. and Zame, W.R. 1987. Large games; fair and stable outcomes. *Journal of Economic Theory* 42, 59–93.

Xue, L. 1998. Coalitional stability under perfect foresight. *Economic Theory* 11, 603–27.

Yannelis, N. 1991. The core of an economy with differential information. *Economic Theory* 1, 183–98.

Yi, S.-S. 1997. Stable coalition structures with externalities. *Games and Economic Behavior* 20, 201–37.

How to cite this article

Wooders, Myrna and Frank H. Page, Jr. "coalitions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000179> doi:10.1057/9780230226203.0250

The New Palgrave Dictionary of Economics Online

Coase theorem

Francesco Parisi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Coase Theorem holds that, regardless of the initial allocation of property rights and choice of remedial protection, the market will determine ultimate allocations of legal entitlements, based on their relative value to different parties. Coase's assertion has occasioned intense debate. This article provides an intellectual history of Coase's fundamental theorem and surveys the legal and economic literature that has developed around it. It appraises the most notable attacks to the Coase Theorem, and examines its methodological implications and normative and practical significance in legal and policy settings.

Keywords

adverse selection; American Law and Economics Association; asymmetric information; bargaining; Coase Theorem; contract enforcement; efficient allocation; entropy; externalities; free rider problem; hold-up; inalienability; incentives; income effect; law, economic analysis of; liability rules; Pigou, A. C.; Pigouvian taxes; Plant, A.; private information; property fragmentation; property rights; public goods; Scarcity; social cost; Stigler, G.; strategic behaviour; tort; transaction costs; voluntary transfers

Article

Mutuality of advantage from voluntary exchange is one of the most fundamental concepts in economics. The well-known proposition of Ronald H. Coase (1960) – generally known as the Coase theorem – builds on this simple and yet fundamental insight. The law creates many rights and legal entitlements, establishing the initial allocation of rights and liabilities. Whenever there are no legal or factual impediments to exchange, the dynamic of the market will determine the final allocation of such rights. In this context, Coase suggests that the transferability of rights in a free economy leads toward their best use and an efficient final allocation. Whenever the initial allocation is not optimal, the owners of the rights will have an incentive to transfer them to other individuals who value them more. Such an exchange will continue until there is no further potential for reciprocal profit, which will not be

exhausted until each right is in the hands of the highest-valuing individual. The Coase theorem predicts that, in a competitive market environment without legal or factual impediments to exchange, the final allocation of rights will be efficient.

This article discusses the pervasive methodological implications of Ronald Coase's idea to the field of law.

1 A brief intellectual history

Coase's assertion that an initial assignment of property rights is often irrelevant to overall welfare has occasioned one of the most intense and fascinating debates in the history of legal and economic thought. Private property is often explained as the unavoidable by-product of scarcity in a world where common pool losses outweigh the sum of contracting costs and enforcement of exclusive property rights. At the turn of the 20th century, the underlying assumption in the economic literature was that private property emerged out of a spontaneous evolutionary process because of the desirable features of private property regimes in the creation of incentives for constrained optimization.

This understanding of the relationship between scarcity and emergence of legal entitlements characterized mainstream property right theory when Coase entered the academic world. Coase began his undergraduate studies at the London School of Economics in 1929, as a candidate for a Bachelor Degree in Commerce. In those years, one of Coase's teachers, Sir Arnold Plant, was re-examining the theme of property rights from a novel perspective. According to Plant, the traditional justification for private property – scarcity – was incapable of serving as the sole intellectual foundation for this institution. Plant showed that incentives, rather than scarcity, lay at the core of the property right problem (Plant, 1974).

Coase's use of legal rules as an object of economic research in his analysis of incentive structure and alternative final resource allocations reveals a remarkable technical affinity with the work of his undergraduate teacher. In his Nobel memorial lecture, Coase acknowledges the importance of his encounter with Plant as a ‘great stroke of luck’ that cultivated his interest in property rights theory (Coase, 1992, p. 715). For Coase, Plant's teaching that ‘[t]he normal economic system works itself’ (Salter, 1921, pp. 16–17) and that prices in a competitive market lead resources to their highest valuing uses was a revelation into the dynamic of the economic system: ‘I was then 21 years of age, and the sun never ceased to shine. I could never have imagined that these ideas would become some 60 years later a major justification for the award of a Nobel Prize. And it is a strange experience to be praised in my eighties for work I did in my twenties’ (Coase, 1992, p. 716).

The experience of the following years at the London School of Economics laid the methodological foundations of what would later become Coase's theorem on the problem of social costs. All the ingredients of his revolutionary analysis on the debated theme of social cost had been profiled during his LSE years (see Williamson and Winter, 1991, pp. 34–5). But it is not until the late 1950s that Coase verbalized such a simple and yet ingenious idea. He had first expounded the core of his later theorem in an article published in 1959. In those pages, one grasps what would later become the central theme of Coase's celebrated argument:

> Whether a newly discovered cave belongs to the man who discovered it, the man on

> whose land the entrance to the cave is located, or the man who owns the surface under which the cave is situated is no doubt dependent on the law of property. But the law merely determines the person with whom it is necessary to make a contract to obtain the use of the cave. Whether the cave is used for storing bank records, as a natural gas reservoir, or for growing mushrooms depends, not on the law of property, but on whether the bank, the natural gas corporation, or the mushroom concern will pay the most in order to be able to use the cave. (1959, p. 25)

The discussion of the rationale of property rights under Coase's highest bidder framework obviously contained an attack on the Pigouvian approach (Pigou, 1920) to the problem. The point was rather self-evident to Coase, but not so for some of the Chicago economists. George Stigler was among Coase's early critics:

> Ronald Coase criticized Pigou's theory rather casually, in the course of a masterly analysis of the regulatory philosophy underlying the Federal Communication Commission's [FCC] work. Chicago economists could not understand how so fine an economist as Coase would make so obvious a mistake. Since he persisted, we invited Coase (he was then at the University of Virginia) to come and give a talk on it. Some twenty economists from Chicago and Ronald Coase assembled one evening at the home of Aaron Director.•…•In the course of two hours of argument the vote went from twenty against and one for Coase to twenty-one for Coase. What an exhilarating event! (Stigler, 1988, pp. 75–6)

According to Coase, the objections to his FCC paper are at the origin of his later 1960 article on the problem of social costs. Coase recalls that he was urged to omit that section of his FCC article, something he refused to do. In retrospect, Coase believes that had it not been for the Chicago economists' attacks his full-fledged idea would have never been formulated (1993, p. 250).

2 The positive Coase Theorem

The arguments that were refined in the course of such debate were later put together in the form of an article for the *Journal of Law and Economics* in 1960, titled ‘The Problem of Social Cost’. This article – later known as the Coase theorem – soon became a milestone in legal and economic literature. In the course of his austere discussion, Coase does not reveal any sign of anticipated realization of the revolutionary power of his insight. Indeed, Coase insists that he never intended to convey his thoughts in the precise and analytical form of a theorem (1988, p. 157).

A few years after the publication of ‘The Problem of Social Cost’, a sizeable number of commentaries and theoretical elaborations were developed on Coase's newly presented theme. The unpretentious style of Coase's article had thus been crowned by a notoriety rarely attained by legal writings of any sort (Shapiro, 1985, p. 1540). Part of the uproar is explained by the fact that the article challenged an established principle of public finance (see Manne, 1975, pp. 123–6). Before ‘The Problem of Social Cost’, very little attention had been given to the possibility that the problem of externalities could be resolved through free market exchanges.

Coase boldly attacked the conclusions reached by the Pigouvian tradition by suggesting its influence was in part due to the lack of clarity in its exposition (1960, p. 39). Coase departs from the Pigouvian approach by demonstrating that, in the absence of transaction costs, generators and victims of externalities will negotiate an efficient allocation of resources, independent of the initial assignment of rights among them. In confuting the conclusions of the Pigouvian tradition, Coase gave life to a model with the potential for the evaluation of an unlimited number of legal and social issues.

George Stigler was the first scholar to restate Coase's model in the form of a theorem: ‘[U]nder perfect competition private and social costs will be equal’ (1966, p. 113). Demsetz (1967, p. 349) defined the theorem in the following terms: ‘There are two striking implications of this process that are true in a world of zero transaction costs. The output mix that results when the exchange of property rights is allowed is efficient and the mix is independent of who is assigned ownership (except that different wealth distributions may result in different demands)’. Soon thereafter, Guido Calabresi stated the same principle more descriptively: ‘Thus, if one assumes rationality, no transaction costs, and no legal impediments to bargaining, all misallocations of resources would be fully cured in the market by bargains’ (Calabresi, 1968, p. 68).

The implicit premise of Coase's analysis draws upon a fundamental postulate of microeconomic theory: the free exchange of goods in the market moves goods towards their optimal allocation. The voluntary transfer of individual rights in the marketplace, thus, will cure a non-optimal allocation of legal entitlements.

2.1 The Coasean methodological revolution

Coase's article constitutes, according to many commentators, the first example of an economic analysis of law in North American literature. The novelty of his approach inspired an entire generation of scholars – pioneers in this new branch of applied economics. Only a few months prior to receiving the Nobel Prize for economics, in occasion of the First Annual Meeting of the American Law and Economics Association, Ronald H. Coase was recognized, together with Guido Calabresi, Henry G. Manne and Richard A. Posner, as a founding father of Law and Economics. This recognition follows many years of challenging debate. Many of the writings that developed around ‘The Problem of Social Cost’ tested the premises of Coase's model, seeking to undermine the conditions of his model and stressing the lack of practical reach of his analysis.

Further criticisms pertained to three fundamental points. One group of critics observed that the Coase Theorem disregarded the inter-industrial long-term effects of the system (Calabresi, 1965; Wellisz, 1964). These critics argued that Coase ignored the possible disequilibria which may occur after the negotiation and the likely dynamic changes in the initial equilibrium. In the context of Coase's well-known example, if the right has been assigned to the ranchers, the farmer will have to pay local ranchers until they all relinquish their right of pasture. The entire cost will, thus, burden the farming industry. Farmers will either have to bear the burden of the injury caused by the livestock or agree to pay the price demanded by the ranchers, whichever is less, on the assumption that negotiation is costless. Under this liability rule, the cost of ranching will not reflect the cost imposed on the farmers. The transfer of rights and liability from one group to another will, therefore, result in a shift in the relative wealth and costs associated with the two industries. The criticism claims that, in the long run, every shift of wealth will

lead to an inter-industrial disequilibrium.

In 1968, Calabresi, one of the initial proponents of this criticism, reconsidered it, noting that in the presence of determined conditions the conclusions of Coase remain as true in the long run as in the short term (1968, p. 67). Calabresi's later analysis re-established the authority of the Coase Theorem, at least on this point. It became clear that Coase did not ignore the long-term effects of his model. Perhaps not explicitly, he had considered them to their logical extreme. Calabresi observes: ‘The reason is simply that (on the given assumptions) the same type of transactions which cured the short run misallocation would also occur to cure the long run ones.•…•This process would continue until no bargain could improve the allocation of resources’ (1967, pp. 67–8).

In 1972, Harold Demsetz joined this debate, demonstrating with a more systematic analysis that the conclusions reached by Coase are not corroded by the long-term effects of a change in the assignment of property rights. Demsetz's reasoning finds its basis in the principle that the process of allocation of scarce resources among alternative uses is analogous to the process of constrained optimization of the single owner of two conflicting activities.

An additional critique, formulated by Calabresi (1965) and Wellisz (1964), suggests that strategic behaviour in the bargaining process risks compromising Coase's results. These authors observe that the change in the rule of law creates the conditions for possible extortion on the part of the right holders against the other individuals who are bound by the rule. The argument is that individuals are likely to threaten the use of their own rights in a measure which exceeds the optimal level, in order to maximize the gain from the release of their own legal entitlements. By introducing the possibility of strategic behaviour in the negotiation, the result may differ from the optimal equilibrium. Demsetz (1972, p. 21) supplied a convincing answer to this criticism. According to Demsetz, the possibility of strategic behaviour in the negotiations does not alter the efficiency in the final allocation of resources between the two activities. Strategies will be capable of altering the internal distribution of the contractual surplus between the parties, but not the final outcome of the negotiation.

It should be noted, however, that the entire analysis presupposes that the so-called income effect can be ignored. In general, a different allocation of property rights implies a different distribution of wealth between the individuals involved. Different initial endowments generate different final allocations, notwithstanding an equal level of efficiency. In order for the final allocations to be identical, it is necessary that the utility functions of the individuals involved are almost linear. The absence of the income effects implies, in this sense, that the demand functions for the good are independent of the income level.

It should be further observed that the credibility of the threat made in the course of strategic bargaining finds its limits in the market structure in which the Coasean negotiation takes place. In general, the competitive structure of the market eliminates much of the advantage that can be obtained through strategic behavior in the negotiation process. Inasmuch as the market of resources is competitive, strategic bargaining is not capable of bringing about any abnormal return.

The criticism, however, appears to be on the mark when it argues that, in some marginal situations, the curing role of the free exchange may still be impeded. For example, consider reversing the assignment of property rights between the rancher and the farmer. In such a situation, the farmer is likely not to have an equally large number of alternatives. The transfer of a farm from one place to another is costly, and farming unavoidably requires the undertaking of location-specific investments. Since some capital investment is irreversibly locked in that specific location, the farmer has less opportunity to relocate than

the rancher. The rancher, consequently, finds himself in a position of local monopoly in the sale of his property right. Demsetz considers the monopoly that affects this feature of the Coasean exchange identical to the standard monopoly of microeconomic analysis (1972, p. 24). According to Demsetz, the concerns for possible monopolistic structures in the market of rights considered by Coase must not, however, be used to raise again the already resolved problem of the initial allocation of rights, since reversing the rule of liability would simply result in the farmer now having monopoly power (1972, pp. 24–5).

A second group of critics concentrated on the distributive effects of the model (Regan, 1972; Nutter, 1968). They argued that a final efficient allocation of resources requires transfers of wealth induced by the changed legal rule. Further, these critics observed that, even if one disregards the distributive effects of the rule, a different assignment of the right could in some cases create the conditions for strategic behaviour in negotiation capable of disturbing the efficiency of the final allocation.

A third group of authors focused on the scarce realism of the no-transaction-cost assumption (see Cooter, 1987, p. 457). According to this criticism, the true Achilles' heel of Coase's analysis was in the unrealistic assumption of absence of costs in the process of negotiation and transfer of the right. These authors observed that the idea of a transaction without cost is a logical fiction cloaking a mere tautology.

3 The normative Coase theorem

The utility of models predicting behaviour in a zero transaction-cost world is that they guide the law – whose object is to develop rules which approximate the zero transaction-cost world as closely as possible – in responding to legal problems arising in a positive transaction-cost environment (Epstein, 1993). The vast literature that developed around Coase's theorem formulated important normative corollaries of it, based on the evaluation of the relative costs of alternative assignments of rights. According to the positive Coase theorem, absent transaction costs, the final allocation of scarce resources would coincide with the use that an individual who is the single owner of different activities would make of his endowments, regardless of the initial assignment of rights and choice of remedial protection. When transaction costs are present, however, an exchange will be pursued only to the point at which its marginal benefit equals the marginal cost of the transaction. If transaction costs exceed the benefits of a contract, no exchange will take place in the market. For a right to be exchanged it is necessary that transaction costs be less than the difference between the demand and supply prices. If this condition is not met, the Coasean bargaining will not be carried out, and both initial assignment of rights and choice of remedies will affect final allocations.

3.1 The relevance of transaction costs and the simple normative Coase theorem

The notion of transaction costs has acquired particular importance in law and economics as the absence of transaction costs represents a fundamental condition for the applicability of the positive Coase theorem. Although at first impression transaction costs play a role analogous to transportation costs in international trade or, more generally, to the contracting costs in the economics of exchange (Demsetz, 1972, p. 20), in Coase's world the role of transaction costs has much greater normative implications. For purposes of the theorem, the notion of transactions costs should include not only bargaining costs

associated with the negotiation and conclusion of the contract but also all costs associated with the strategic behaviour of the parties and the execution and enforcement of the transaction. The notion of transaction costs should thus include *ex ante* costs due to asymmetric information, adverse selection, free riding, and hold-up strategies, as well as *ex post* costs associated with monitoring and enforcing the contracts.

Strategic behaviour may be an important source of transaction costs in a Coasean setting. In Coase's various examples, the property rights which are exchanged are private goods, characterized by their excludability. Difficulties arise when the object of the Coasean bargaining is an entitlement which has the nature of a public good (see Cheung, 1970, pp. 49–70). Due to the well-known problems associated with the supply of public goods, the Coasean bargaining solution may fail to cure a non-optimal allocation of rights that falls within this category. Consider a scenario in which the object of the Coasean negotiation consists of a non-excludable right (for example, the right to enjoy pollution-free air in a residential environment). As well known, individuals will not reveal their own preferences for public goods through the price system, placing public goods among those cases that are most resistant to the Coasean antidote.

A first simple normative reformulation of the Coase theorem focuses on transaction costs and the role that legal systems may play in reducing these impediments to voluntary bargaining. Legal rules can lower obstacles to private bargaining, such as by reducing transaction costs and minimizing other costs associated with transfer (strategic, legal, and so on). For this reason, transactional cost considerations should be fundamental to any analysis of legal regimes and the design of contracting processes, governance mechanisms and institutions.

3.2 The complex normative Coase theorem

The first original formulation of Coase's proposition can be restated as a normative theorem: in the presence of positive transaction costs, the efficiency of the final allocation is not independent of the choice of the legal rule, and that the preferable initial assignment of rights is that which minimizes the effects of such transaction costs. The various normative restatements of the Coase Theorem aim at identifying legal rules and remedies that replicate the outcomes of a hypothetical Coasean bargaining or to mimic the solution that would be chosen by the single owner of interfering resources.

Important normative reformulations of the Coase Theorem focus on two important elements: relevance of initial assignment of rights and relevance of remedial protection. Demsetz (1972) and Calabresi and Melamed (1972) were among the first to discuss systematically the problems resulting from lifting the assumption of zero transaction costs. Articulating the normative core of the Coase theorem, Demsetz observes that the introduction of significant transaction costs into the choice of liability rule analysis does affect resource allocation. One liability rule may be superior to another because the difficulty of avoiding costly interactions is usually different for the interacting parties. Accordingly, the normative predicament indicates that the rule of liability should be based on which party can avoid the costly interaction at the lowest cost.

When two or more parties have conflicting interests in the same resource, the law must decide which party shall prevail, that is, which party shall receive the entitlement. Once the entitlement decision is made, the law must decide how the entitlement is to be protected and whether it may be transferred.

Articulating a concept of entitlements protected by property, liability or inalienability rules, Calabresi and Melamed (1972) develop a framework that integrates the approaches of property and tort. Entitlements can be protected by property rules (transfer of the entitlement involves a voluntary sale by its holder), liability rules (the entitlement may be destroyed by another party if he is willing to pay an objectively determined value for it), or rules of inalienability (transfer of the entitlement is not permitted, even between a willing seller and a willing buyer). Calabresi and Melamed allow for a wide range of concerns to be balanced through the assignment of a particular entitlement. Calabresi and Melamed outline how, given the reality of transaction costs, an economic efficiency approach selects one allocation of entitlements over another. Entitlements cannot be enforced solely through property rules because, even if the transfer would benefit all parties, high transaction costs (especially the hold-up problem) may prevent an efficient reallocation. Calabresi and Melamed demonstrate how liability rules often achieve a combination of efficiency and distributive results that would be difficult to achieve under a property rule. Calabresi points out that Coase's analysis offers invaluable instruments for the identification of the areas in which public intervention becomes desirable (Calabresi, 1968, pp. 72–3). In its normative version, the theorem indicates that legal rules that minimize the effects of such costs are to be preferred for being relatively more efficient (Polinksy, 1989, p. 14). In its more complex formulation, the Coase theorem provides, indeed, a guide for such a choice.

The following is a classic illustration (Polinsky, 1989, pp. 11–14). The smoke of a factory soils laundry which is line drying on five neighbouring properties. The losses amount to $150 for each neighbour, for a total of $750. The damage could be eliminated through the installation of a purifying filter on the industrial smokestack or through the acquisition of electric dryers on the part of each one of the neighbouring owners. The cost of the filter would amount to $300, while the dryers would impose a cost of $100 per household, for a total of $500. The first solution is obviously more efficient, since the acquisition of five dryers would require a greater expenditure than the single filter. The Coase theorem predicts that in the absence of transaction costs the efficient solution will be chosen independently of the initial assignment of property rights. Even if we assume an initial allocation of polluting right to the industry (that is, fully legalizing industrial emissions), the landowners would jointly offer to buy the industrial filter at their expense. Sharing the cost of the filter in equal parts, each owner would face a cost of only $60, with a relative saving of $40 compared with the otherwise necessary acquisition of a personal dryer.

If we relax the initial assumption of no transaction costs, the initial allocation of property rights no longer is immaterial. Imagine that each owner has to face a cost of $120 in order to negotiate the contract with his neighbours and with the owner of the industrial plant. If the right is assigned to the industry, each landowner will have to choose whether to bear the loss of his soiled laundry for $150, to acquire the electric dryer for $100, or, finally, to undertake the negotiation process for a total pro-rata cost of $180. Considering these alternatives, each rational landowner would choose to acquire his own dryer, generating a socially non-optimal outcome. However, the assignment of property rights to the neighbouring residents rather than to the polluting industry would minimize the effect of positive transaction costs, since the industry would have incentives to install the filter, without any need for Coasean bargaining with the neighbours.

Two impediments to bargaining (that is, sources of transaction costs) take the form of externalities and hold-up, which Epstein (1993) shows stand in inverse relationship to each other. He defines the optimal legal rule as that which minimizes the sum of these externality and holdout costs in any particular

institutional setting. Epstein demonstrates, through examples in property, restitution and tort, how Coase's transaction costs model plays the central organizing role in developing legal responses to many private law problems. Notwithstanding the obvious measurement and information problems, Epstein (1993) stresses the importance of the ‘single owner test’: where resources are under the command of two or more persons, the legal arrangement should attempt to induce all the parties to behave in the same way that a single owner would. Epstein concludes that, where the single owner test yields a unique result, that result should be adopted as the legal rule. Where the single owner test does not yield clear results, however, no corollary principle will provide a decisive answer to the particular problem.

Further exploring the choice between property and liability rules suggested by Calabresi and Melamed, Kaplow and Shavell (1996) address several factors casting doubt on the equivalence of these alternatives in low transaction-cost environments. Their analysis considers several objections to Coasean costless bargaining, including the inability of a party to ascertain what the other is willing to pay or accept, victims' ability to mitigate harm, the problem posed by one party being judgment proof, and administrative costs. Kaplow and Shavell find a presumption in favour of liability rules over property rules in the context of harmful externalities, but that this may be overcome as a result of one of more of the factors they describe. After considering some of the proffered justifications for the use of property rules to protect possessory interests, the authors find a strong theoretical case for the protection of these interests using property rules. The normative Coase theorem thus underlies the choice of the optimal system to ensure the protection of various types of property rights.

Also bridging the gap between Coase, where liability rules and property rules are equally efficient, and Calabresi and Melamed, where high transaction costs lead to a preference for liability rules, is the work by Ayres and Talley (1995) on private information as a transaction cost. The inefficiency occurs when parties misrepresent their own valuations to gain strategic advantage in the bargaining process. Focusing on the effect of splitting an entitlement between two rivalrous users rather than among buyers or among sellers, these authors find that, when two parties have private information about how much they value an entitlement, endowing each party with a partial claim to the entitlement can reduce the incentive to behave strategically during bargaining by inducing greater disclosure. A bargainer has two Coasean alternatives: buy the other party's claim or sell one's own claim. The normative formulation of Ayres and Talley is that a liability rule regime is preferable because it allows a party's decision to pursue one of these alternative transactions to function as a credible signal of a low or high valuation, thereby encouraging more efficient trade.

Building upon the literature on property fragmentation (Heller, 1998; Buchanan and Yoon, 2000), Parisi (2002) and Schulz, Parisi and Depoorter (2002) suggest that property is subject to a fundamental law of entropy. In the property context, entropy induces a one-directional bias. This bias is driven by asymmetric transaction costs – it is often harder to reunite separated property bundles than to break them apart. Parisi hypothesizes that courts and legislators account for the presence of asymmetric transaction costs and correct for problem through the selective use of remedies and by selecting default rules designed to minimize the total deadweight losses of property fragmentation. Parisi (2006) offers a reformulation of the normative Coase theorem in situations characterized by asymmetric transaction and strategic costs, such as when complementary fragments of property are attributed to different owners. The asymmetry arises from the fact that it is often harder to reunite separated property bundles than to break them apart. This variant of the Coase theorem turns on (*a*) an initial allocation of entitlements that

minimizes the effects of the positive transaction costs, and (*b*) the selection of legal rules that reduce social welfare losses by facilitating optimal levels of reunification.

4 The Coase theorem and its legacy in law and economics

In 1960 Coase entrusted legal and economic scholars with the challenging task of deriving the implications of his theorem in their areas of research. Coase's invitation was taken up by a number of economists and lawyers who experimented with the unparalleled analytical potential of Coase's theorem in their research. According to Coase, economists in the Pigouvian tradition fail to consider the possible reciprocity of the effects of individual choices. By labelling one agent as injurer and the other as victim, the Pigouvian tradition presumes an initial allocation of rights (Cornes and Sandler, 1986, p. 59). In such a manner this approach falls into a serious methodological error, notwithstanding empirical psychological studies suggesting otherwise (see Kahneman, Knetsch and Thaler, 1990, pp. 1325–48). By taxing the generator of the externality in a measure corresponding to the difference between the private cost and the social cost of his own activity, the followers of Pigou fail to consider the effects of potential victims' behaviour. If the social cost of the industrial emissions is calculated by aggregating the economic disadvantages of the residents who are negatively affected by the smoke, the figure will vary with the number of individuals who fix their residence in that area. If the Pigouvian tax is imposed on the industrial activity only, there will be less incentive for each resident to consider moving into a different neighbourhood. New individuals may actually locate their residence in that area, without considering the potential increase in the costs imposed on the industrial activity.
Through these arguments, Coase's analysis demonstrates the incapacity of the Pigouvian approach to consider the interdependence of the harmful effects generated by individual choices. Coase's analysis occasioned a paradigmatic shift in legal and economic analysis, and, as Henry Manne once observed, ‘it is hard to imagine law ever again being free of the influence of the techniques and findings of objective economic analysis’ (1993, p. 4). His theorem, short of providing a simplistic formula for the social cost problem, suggests an alternative approach based on the evaluation of the relative costs of alternative assignments of rights and legal protection.

See Also

- hold-up problem
- property rights

Bibliography

Ayres, I. and Talley, E. 1995. Solomonic bargaining: dividing a legal entitlement to facilitate Coasean trade. *Yale Law Journal* 104, 1027–117.

Buchanan, J. and Yoon, Y.J. 2000. Symmetric tragedies: commons and anticommons property. *Journal of Law and Economics* 29, 1–13.

Calabresi, G. 1965. The decision for accidents: an approach to non-fault allocation of costs. *Harvard Law Review* 78, 713–45.

Calabresi, G. 1968. Transaction costs, resource allocation and liability rules: a comment. *Journal of Law and Economics* 11, 67–8.

Calabresi, G. and Melamed, A.D. 1972. Property rules, liability rules, and inalienability: one view of the cathedral. *Harvard Law Review* 85, 1089–128.

Cheung, S. 1970. The structure of a contract and the theory of a non-exclusive resource. *Journal of Law and Economics* 13, 49–70.

Coase, R.H. 1959. The Federal Communications Commission. *Journal of Law and Economics* 2, 1–40.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Coase, R.H. 1988. *The Firm, the Market, and the Law*. Chicago: University of Chicago Press.

Coase, R.H. 1992. The institutional structure of production. *American Economic Review* 82, 713–19.

Coase, R.H. 1993. Law and economics at Chicago. *Journal of Law & Economics* 36, 239–54.

Cooter, R. 1987. Coase Theorem. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Cornes, R. and Sandler, T. 1986. *The Theory of Externalities, Public Goods, and Club Goods*. Cambridge: Cambridge University Press.

Demsetz, H.M. 1967. Toward a theory of property rights. *American Economic Review* 57, 347–59.

Demsetz, H.M. 1972. When does the rule of liability matter? *Journal of Legal Studies* 1, 13–28.

Epstein, R.A. 1993. Holdouts, externalities, and the single owner: one more salute to Ronald Coase. *Journal of Law and Economics* 36, 553–86.

Heller, M.A. 1998. The tragedy of the anticommons: property in the transition from Marx to markets. *Harvard Law Review* 111, 621–88.

Kahneman, D., Knetsch, J.L. and Thaler, R.H. 1990. Experimental tests of the endowment effect and the Coase Theorem. *Journal of Political Economy* 98, 1325–48.

Kaplow, L. and Shavell, S. 1996. Property rules versus liability rules: an economic analysis. *Harvard Law Review* 109, 723–54.

Kuhn, T.S. 1970. *The Structure of Scientific Revolutions*, 2nd edn. Chicago: University of Chicago Press.

Manne, H.G. 1975. *The Economics of Legal Relationships*. St. Paul: West Publishing Co.

Manne, H.G. 1993. *The Intellectual History of George Mason University School of Law*. Arlington, VA: School of Law, George Mason University.

Nutter, G.W. 1968. The Coase Theorem on social cost: a footnote. *Journal of Law and Economics* 16, 503–7.

Parisi, F. 1995. Private property and social costs. *European Journal of Law and Economics* 2, 149–73.

Parisi, F. 2002. Entropy in property. *American Journal of Comparative Law* 50, 595–632.

Parisi, F. 2006. Entropy and the asymmetric Coase Theorem. In *Property Rights Dynamics: A Law and Economics Perspective*, ed. D. Porrini and G. Ramello. London: Routledge.

Polinsky, A.M. 1989. *An Introduction to Law and Economics*. Boston: Little, Brown and Company.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Plant, A. 1974. *Selected Economic Essays and Addresses*. London: Routledge and Kegan Paul.

Regan, D.H. 1972. The problem of social cost revisited. *Journal of Law and Economics* 15, 427–33.

Salter, J.A. 1921. *Allied Shipping Control*. Oxford: Clarendon Press.

Schulz, N., Parisi, F. and Depoorter, B. 2002. Fragmentation in property: towards a general model. *Journal of Institutional and Theoretical Economics* 158, 594613.

Shapiro, F.R. 1985. The most cited law review articles. *California Law Review* 73, 1540–54.

Stigler, G.J. 1966. *The Theory of Price*, 3rd edn. New York: Macmillan.

Stigler, G.J. 1988. *Memoirs of an Unregulated Economist*. New York: Basic Books.

Wellisz, S. 1964. On external diseconomies and the government assisted invisible hand. *Economica* 31, 345–62.

Williamson, W. and Winter, S.G. 1991. *The Nature of the Firm: Origins, Evolution, and Development*. New York: Oxford University Press.

How to cite this article

Parisi, Francesco. "Coase theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000589> doi:10.1057/9780230226203.0252

The New Palgrave Dictionary of Economics Online

Coase, Ronald Harry (born 1910)

Steven G. Medema

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Ronald Coase made seminal contributions to law and economics and to the theory of the firm, for which he received the 1991 Nobel Prize. The importance of understanding the role of transaction costs in economic activity and the influence of alternative institutional structures on economic performance are hallmarks of Coase's scholarship, and both the economic analysis of law and the new institutional economics are outgrowths of his work. Coase occupies a significant although somewhat controversial place in the history of the Chicago School of economics.

Keywords

accounting; average cost pricing; Bargaining; Chicago School; Coase conjecture; Coase theorem; Coase, R. H.; cobweb theorem; consumer theory; externalities; firm, organization of; government failure; imperfect competition; Knight, F.; law and economics; Lerner, A.; marginal cost pricing; market failure; mathematics and economics; monopoly; multi-part pricing; new institutional economics; opportunity cost; Posner, R.; public goods; public utilities; public utility pricing; rational expectations; social cost theory; special interests; transaction costs; Viner, J.

Article

Ronald Harry Coase was born on 29 December 1910 in the London suburb of Willesden. He received the BSc in Commerce from the London School of Economics in 1932 and while there was greatly influenced by Arnold Plant, who, as Coase has said, taught him many of the lessons that later came to be associated with the Chicago School. Interestingly, Coase did not take a single economics course while he was at the LSE, which he suggests gave him ‘a freedom in thinking about economic problems which [he] might not otherwise have had’ (1990, p. 3).
Upon completing his studies at the LSE, Coase took up a position at the Dundee School of Economics and Commerce, where he taught with his friend and public choice pioneer Duncan Black from 1932–34.

Coase moved on to the University of Liverpool in 1934–35 before returning to the LSE, where he remained from 1935 until 1951. His time at the LSE was interrupted by the Second World War, during which he served as a statistician at the Forestry Commission (1940–41) and in the Central Statistical Office, Offices of the War Cabinet (1941–46). Coase left the LSE for the US and the University of Buffalo in 1951, remaining there until 1958. Following a year spent at the Center for Advanced Study in the Behavioral Sciences at Stanford, he accepted an appointment at the University of Virginia in 1959. Although Coase is most closely associated with the Chicago School, his two most influential works – ‘The Nature of the Firm’ (1937) and ‘The Problem of Social Cost’ (1960) – were written before he arrived at Chicago, in 1964, to teach at the Law School and to join Aaron Director in editing the *Journal of Law and Economics*. Coase retired from the University of Chicago in 1981 and was awarded the Nobel Prize in Economics in 1991.

Scholarly work

While most economists identify Coase with his two classic articles on the firm and social costs, his published output is very extensive and ranges across topics such as accounting, advertising, public goods, consumer surplus, public utility pricing, monopoly theory, blackmail, the economic role of government, and the history of economic thought. Several themes appear throughout Coase's work: the important role played by institutions – in particular the firm, the market and the law – in determining economic structure and performance, the role of transaction costs in economic activity, the need for a comparative institutional approach to economic policy, and a distaste for abstract theorizing. These themes come through unmistakably in *The Firm, the Market and the Law* (1988) and *Essays on Economics and Economists* (1994), which, together, collect many of Coase's most significant writings.
The lion's share of Coase's work during the first part of his career dealt, in one way or another, with firm behaviour and organization. His earliest contributions analysed the formation of producers' expectations (for example, Coase and Fowler, 1935), using the pig cycle as the case study. The conventional cobweb theorem explanation for these cycles was that producers expected current prices and costs to continue into the future. The adjustments in supply that resulted then gave rise to disequilibrium cycles. Coase and Fowler found that this explanation was incorrect – that producers did in fact adjust their expectations of prices and costs very quickly, and that the prediction errors arose from the difficulty of predicting variations in demand and in foreign supply. This work was later cited by J.F. Muth (1961, p. 21) in one of his classic papers on rational expectations. Coase also collaborated with Fowler and Ronald Edwards on a series of pieces dealing with the interrelations between accounting and economics (for example, Coase 1938; Coase, Edwards, and Fowler, 1938). These writings, which were very much in the LSE cost tradition, demonstrated that traditional accounting practices do not adequately capture the true (opportunity) nature of costs and also pointed to the problematic nature of designing workable accounting methods to do so.
Coase also wrote a number of articles dealing with monopoly and imperfect competition, a few of which bear mention of here. Two of his theoretical pieces are of particular import. ‘Durability and Monopoly’ (1972) demonstrated that a monopoly firm which produces a good that is infinitely durable will be forced to sell the good at the competitive price, unless it can decrease the durability of the good or make contractual arrangements through which it promises to limit its production – a result which has come to be known as ‘the Coase conjecture’. ‘The Marginal Cost Controversy’ (1946) is Coase's most

significant work on monopoly and deals with public utility pricing and regulation. Abba Lerner and others had claimed that marginal cost pricing accompanied by a government subsidy is the efficient pricing policy for public utilities. Against this, Coase argued that marginal cost pricing is inferior to a system of multi-part pricing and may in fact be inferior to *average* cost pricing. This paper, and three related papers that followed it, are illustrative of one of the central themes in Coase's work – that, in assessing the efficiency of economic outcomes, one must focus broadly, rather than narrowly, on benefits, costs, and incentives.

Coase's work on public utilities also has an historical strand. Articles on the British Post Office discuss the rise of the penny postage in Great Britain under Rowland Hill and the attempts by the Post Office to enforce its monopoly against incursions by private entrepreneurs, including the messenger companies (for example, 1955). His study of British broadcasting analyses the development of wireless and wire radio broadcasting, as well as of television broadcasting and the rise of the BBC as the monopoly supplier of all of the above (1950; 1954). His interest in the government's role in broadcasting carried over to the United States and an analysis of the role of the Federal Communications Commission (1959; 1966) in the allocation of broadcast frequencies. In fact, it was from this study that ‘The Problem of Social Cost’ came to be written.

While the foregoing gives a sense for the breadth of Coase's contributions, it is unquestionable that his most influential work is contained in two papers – ‘The Nature of the Firm’ (1937) and ‘The Problem of Social Cost’ (1960), the two works cited by the Royal Swedish Academy in awarding Coase the Nobel Prize. In the former, Coase set out to explain why firms exist and what determines the extent of a firm's activities. He found the answer in a concept to which most economists had until recently paid scant attention – transaction costs. Coase suggested that we tend to see firms emerge when the cost of internal organization is lower than the cost of transacting in the market, and that the limit of a firm's activities (or, the extent of internal organization) comes at the point where the cost of organizing another transaction internally exceeds the cost of transacting through the market. Although published in 1937, ‘The Nature of the Firm’ attracted little attention until the early 1970s, when Oliver Williamson, Armen Alchian, Harold Demsetz and others began to build on or take off from Coase's contribution to bring transaction costs, the contracting process, and firm organization to the fore in economic analysis.

‘The Problem of Social Cost’ took the transaction-cost paradigm in a different direction – the legal-economic arena and situations of conflicts over rights. Although ‘The Problem of Social Cost’ is one of the most cited articles in all of the economics and legal literatures, it has also been widely misunderstood. From this paper comes the now-famous Coase theorem – actually codified as such by George Stigler (1966) – which says that when transaction costs are zero and rights are fully specified, parties to a dispute will bargain to an efficient outcome, regardless of the initial assignment of rights. But Coase recognized that the transaction costs are pervasive and will generally preclude the working of this bargaining mechanism. Coase thus concludes that legal decision-makers should assign rights so as to maximize the value of output in society – a concept that lies at the heart of the modern law and economics movement (Medema, 1999; Medema and Zerbe, 2000).

The crux of ‘The Problem of Social Cost,’ however, is Coase's attempt to demolish the Pigovian tradition of social cost theory (Pigou, 1932). The analysis that came to be known as the Coase theorem was used to demonstrate that, under standard neoclassical assumptions, Pigovian remedies for externalities are unnecessary: costlessly functioning markets, like the costlessly functioning governments of Pigovian welfare theory, will generate efficient outcomes. The problem, as Coase

pointed out, is that neither markets nor governments function costlessly, and thus neither will generate optimal solutions. This leaves policymakers with a choice among imperfect alternatives, and Coase advocates a close examination of the benefits and costs associated with the alternative policy options, in order to facilitate the adoption of policies (including doing nothing at all) which maximize the value of output.

That government failure is at least as pervasive as market failure, and that economists are too quick to advocate tax, subsidy, and regulatory solutions without a careful examination of the situation, are recurring themes in Coase's work. His analyses of social cost issues, public utility pricing, and his classic article on role of the lighthouse in public goods theory as against the actual history of private lighthouse provision in Great Britain (1974) are excellent examples of Coase's position here. When Coase looks at government, he sees agencies captured by special interests, making policies that usually make matters worse rather than better, and operating in virtual ignorance of the virtues of the market. Yet a careful reading of Coase suggests that he is not ‘anti-government’ but, rather, an advocate for economic theorizing and policymaking which recognizes that policy choices are always among imperfect alternatives.

These criticisms are part of Coase's more general concern about the way that economists practice their trade (1994). He is suspicious of consumer theory as a whole and of the way in which mathematical and quantitative techniques have been used in modern economics. His own writings evidence some graphs and some technical intuitive analysis, but, reflecting Coase's lifelong distaste for using mathematics in his work, there is not an equation to be found. Coase believes that economists are obsessed with what he calls ‘blackboard economics’, an economics where curves are shifted and equations are manipulated on the blackboard, with little attention to the correspondence (or lack thereof) between these models and the real-world economic system. This, he says, has manifested itself in economists' ignorance of the role played by transaction costs and economic institutions generally, and in an approach to public policy that fails to examine in any kind of depth the consequences of alternative policy actions.

Coase and Chicago

Coase's critical attitude toward the practice of economics does not stop at the doors of the University of Chicago. Indeed, his close association with the Chicago School belies a degree of tension in the relationship and highlights the risks involved in thinking in terms of a homogeneous Chicago school. In spite of his position as a founding father of law and economics and, by extension, the expansion of the boundaries of economics so closely associated with Chicago, Coase has been critical of economic imperialism generally and of the economic analysis of law in particular (Coase, 1977; 1993). Coase's interest is not the economic analysis of law, but rather the study of how the legal system impacts the economic system – old-style Chicago law and economics of the sort being published in the *Journal of Law and Economics* in the 1960s and 1970s. As such, his interest and intellectual commonalities lie much more with the older Chicago school of Frank Knight and Jacob Viner than with the Becker–Stigler–Posner generation, and he has a much greater interest in the new institutional economics (of which he is also regarded as a founding father) than in the modern economic analysis of law movement à la Richard Posner. Coase has been chastised by Posner (1993) on this and other counts, but he remains unapologetic. That Coase has a place within the Chicago tradition goes without saying, but he has also remained his own man – dissenting from the received doctrine when it did not fit with his views.

See Also

- Chicago School
- Chicago School (new perspectives)
- Coase theorem
- law, economic analysis of
- new institutional economics

Selected works

1935. (With R.F. Fowler.) Bacon production and the pig-cycle in Great Britain. *Economica* n.s. 2, 142–67.

1937. The nature of the firm, *Economica* n.s. 4, 386–405.

1938. Business organisation and the accountant. In *Studies in Costing*, ed. D. Solomons. London: Sweet and Maxwell, 1952.

1938. (With R.S. Edwards and R.F. Fowler.) *Published Balance Sheets as an Aid to Economic Investigation: Some Difficulties*. London: Accounting Research Association Publication No. 3.

1946. The marginal cost controversy. *Economica* n.s. 13, 169–82.

1950. *British Broadcasting: A Study in Monopoly*. London: Longmans, Green and Co.

1954. The development of the British television service. *Land Economics* 30, 207–22.

1955. The postal monopoly in Great Britain: an historical survey. In *Economic Essays in Commemoration of the Dundee School of Economics 1931–1955*, ed. J.K. Eastham. London: William Culcross and Sons.

1959. The Federal Communications Commission. *Journal of Law and Economics* 2, 1–40.

1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

1966. The economics of broadcasting and government policy. *American Economic Review* 56, 440–7.

1972. Durability and monopoly. *Journal of Law and Economics* 15, 143–9.

1974. The lighthouse in economics. *Journal of Law and Economics* 17, 357–76.

1977. Economics and contiguous disciplines. In *The Organization and Retrieval of Economic Knowledge*, ed. M. Perlman. Boulder, CO: Westview Press.

1988. *The Firm, the Market, and the Law*. Chicago: University of Chicago Press.

1990. Accounting and the theory of the firm. *Journal of Accounting and Economics* 12, 3–13.

1992. The institutional structure of production. *American Economic Review* 82, 713–19.

1993. Law and economics at Chicago. *Journal of Law and Economics* 36, 239–54.

1994. *Essays on Economics and Economists*. Chicago: University of Chicago Press.

Bibliography

Medema, S.G. 1994. *Ronald H. Coase*. London and New York: Macmillan and St. Martin's Press.

Medema, S.G. 1999. Legal fiction: the place of the Coase theorem in law and economics. *Economics and Philosophy* 15, 209–33.

Medema, S.G. and Zerbe, R.O., Jr. 2000. The Coase theorem. In *The Encyclopedia of Law and Economics*, ed. B. Bouckaert and G. De Geest. Aldershot: Edward Elgar Publishing.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Pigou, A.C. 1932. *The Economics of Welfare*, 4th edn. London: Macmillan.

Posner, R.A. 1993. Ronald Coase and methodology. *Journal of Economic Perspectives* 7(4), 195–210.

Stigler, G.J. 1966. *The Theory of Price*, 3rd edn. New York: Macmillan.

Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

How to cite this article

Medema, Steven G. "Coase, Ronald Harry (born 1910)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_C000498> doi:10.1057/9780230226203.0251

The New Palgrave Dictionary of Economics Online

Cobb – Douglas functions

Murray Brown
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Perhaps the most common form of production function in economics, the Cobb–Douglas function has a range of attractive properties. The input demand and supply of output functions have the property of continuous differentiability everywhere on their respective domains; and the form has a function coefficient that is identical to its degree of homogeneity, calculated by summing the factor production elasticities. Its restrictions have made it an object of disdain for some. But the Cobb–Douglas form is remarkably robust in a vast variety of applications and is therefore very likely to endure.

Keywords

aggregation (production); CES production function; Cobb, C.; Cobb–Douglas functions; Douglas, P. H.; elasticity of substitution; factor substitution; frontier production functions; production functions; technical change; Walras, L.; Wicksell, J. G. K.; Wicksteed, P. H.

Article

The Cobb–Douglas function is perhaps the most ubiquitous form in economics, owing its popularity to the exceptional ease with which it can be manipulated and to the fact that it possesses the minimal properties that economists consider desirable. It appeared early (at least by 1916; see Wicksell, 1958, p. 133), notably in the theory of distribution where it was used to prove the adding-up theorem of factor shares when the production elasticities sum to unity. It is the first form that many embryonic mathematical economists squeeze and buffet to obtain nice expressions for marginal products and utilities. It has been applied econometrically countless times, still surprising people that it can explain the data so well (Mairesse, 1974). It forces itself into relatively new areas such as frontier production functions (see Førsund, Lovell and Schmidt, 1980). And it has been used both as a utility and production function in analyses of growth, development, macroeconomics, public finance, labour and just about any other applied area in economics. Yet it possesses restrictive properties and perhaps for that reason it has

become for some an object of disdain, often regarded as a child's toy in the world of real economics. But for others, the Cobb–Douglas is at least a venerable form and, effectively, it and its putative inventor are regarded fondly.

In its unrestricted form, the Cobb–Douglas can be written as $f(x) = A\Pi_{i=1} x_i^{ai}$, where A is an efficiency parameter, a_i is the elasticity of $f(x)$ with respect to x_i, and **x** is confined to R^n_{++}. Defining the x_i as goods consumed, it has been used as a utility function; defining them as inputs in the production process, it is a production function; as normalized prices, it is an indirect utility function; and so on. We focus here on its use as a production function for a single output.

A large part of the appeal of the form stems basically from the fact that if $0 < a_i < 1$, $f(\mathbf{x})$ is strongly pseudo-concave on its domain. That entails that if the firm is a profit maximizer and factor supply and product demand functions are continuously differentiable on their domains, then the input demand and supply of output functions have the immensely useful property of continuous differentiability everywhere on their respective domains. Also, if $\Sigma_i a_i \leq 1$ and if factor supply and product demand functions are well-behaved, the input demand functions are downward sloping with respect to own price and the output supply function does not slope downward with regard to product price. What could be better and, moreover, it is all so simple to demonstrate.

Another attractive property of the form is that it has a function coefficient that is identical to its degree of homogeneity, calculated by summing the factor production elasticities. Thus, $\Sigma_i a_i \lesseqgtr 1$ for all i easily and succinctly characterizes decreasing, constant and increasing returns to scale, respectively. This characteristic also has important implications for the cost, profit and revenue duals of the production function. For example, the cost function of a price-taking firm which has a Cobb–Douglas technology decomposes into two parts, one a linear homogeneous function of factor prices and the other a function of output q, that is $C(q, \mathbf{w}) = B\Pi_{i=1} w_i^{ci} q^{co}$, where B is a positive constant, **w** is a (positive) price vector of the inputs, $c_i = a_i / \Sigma_i a_i$ and $c_0 = 1 / \Sigma_i a_i$.

The list of attractive properties extends to the aggregation problem since the Cobb–Douglas is homogeneous and weakly separable. First consider the question of aggregation across inputs. Suppose one can write a generalized Cobb–Douglas function as follows:

$$q = \prod_{s=1} \left(\prod_{j=1} x_{sj}^{b_{sj}} \right)^{\gamma_s},$$

where $b_{sj} = a_{sj} / \Sigma_j a_{sj}$, $\gamma_s = \Sigma_j a_{sj}$, J_s is the number of factors in the sth group, S is the number of groups, $s = 1, 2, \ldots, S$, and $j = 1, 2, \ldots, J_s$. Notice that $\Sigma_j b_{sj} = 1$. Since each expression in the parentheses is homogeneous of degree one for each s, the profit maximization procedure can be decomposed into two stages and there exist quantity and price indexes (call them x_s and W_s, respectively) such that the expenditure on the sth group is $W_s x_s$ for $s = 1, 2, \ldots, S$.

With respect to aggregation across firms, suppose the *r*th firm's production function were

$$q_r = A_r x_{1r}^{c_{1r}} x_{2r}^{c_{2r}}, \ldots, x_{nr}^{c_{nr}},$$

where $\Sigma_i c_{ir} = 1$ and $i = 1, 2, \ldots, R$. It is evident that the expansion paths for all firms are straight lines through their respective origins. Then under the extremely restrictive conditions that the expansion paths for each firm are parallel (i.e. if $c_{ir} = c_{it} = c_i$ for each *i* and for all *r* and *t*), and that the first order conditions are satisfied, the *R* functions consistently aggregate to

$$q = x_1^{c_1} x_2^{c_2}, \ldots, x_n^{c_n},$$

a nicely behaved aggregate production function.

There is another way to look at the aggregation-across-firms problem that involves the Cobb–Douglas function. Suppose that factors in each firm are used in fixed proportions with the Leontief coefficients being distributed across all firms according to a Pareto distribution. Then a surprising result by Houthakker (see Sato, 1975) is that the aggregate production function of the industry is a Cobb–Douglas form.

Of course, there is a price for these desirable implications and most of it is owing to the fact that the Cobb–Douglas technology entails that the elasticity of substitution takes on the knife-edge value of unity. If there is no technological change, a unit substitution elasticity implies that the income shares of all factors of production remain constant in the face of changes in things that are deemed germane such as saving, the rate of growth of the economy and relative factor supplies. Only the state of the technology matters in this instance, a highly disputable outcome. When technological change is allowed to proceed in a Cobb–Douglas world, it is a fact that Hicks-, Solow- and Harrod-neutral technological change are equivalent, thus blurring these distinctions. Another implication of the unit substitution elasticity of the (linear homogeneous) Cobb–Douglas form is that, used in growth models, it guarantees the existence and stability of equilibrium growth, again obscuring an important problem in economics. Furthermore, it is a fact that the Cobb–Douglas form requires that each factor of production be essential in the sense that no factor may be completely substituted for another. Hence the domain of the function must be confined to the set of strictly positive real numbers. This is not particularly disturbing for situations in which the factors can be taken to be large aggregates but it does limit the analysis in other contexts.

Technological change is represented in the Cobb–Douglas by changes in the efficiency parameter *A* which are Hicks neutral, by changes in the scale of the factor inputs which are factor augmenting and also Hicks neutral, and by changes in the elasticities of production which may be Hicks non-neutral. However, the unit elasticity of substitution is restrictive in still another way: it cannot represent a technological advance that results in a change in the ease of substitution among factors of production.

What is the form's provenance? It is generally attributed to Paul Douglas and although he gracefully acknowledged (Douglas, 1967) that Wicksteed and Walras were cognizant of it, he neglected to add Wicksell's name to the list. Be that as it may, Douglas relates in his gentle comments that in 1927 he asked a professor of mathematics, Charles Cobb, to devise a formula that could be used to measure the comparative effect of each of two factors of production upon the total product to satisfy a linear log–log relationship in his input and output data. His work encountered a host of theoretical concerns (see Brown, 1966 for a discussion) aside from the capital, output and labour measures for which he was faulted. But the production form remained in spite (or perhaps because) of its restrictive properties. Subsequent work has demonstrated that the Cobb–Douglas is a special case of a variety of forms and approaches. The constant elasticity of substitution (CES) production function is perhaps the most well known of the forms that yield the Cobb–Douglas as a special case, either by using L'Hôpital's rule when the elasticity of substitution goes to unity or it can be derived from certain expressions used in deriving the CES function (see Brown and De Cani, 1963). Parenthetically, the CES, itself, is known to mathematicians as a mean of order *t* [i.e. $(\Sigma_{i=1} a_i x_i^t)^{\frac{1}{t}}$ for $t \neq 0$] so that, if one takes the limit as $t \to 0$, of course, the Cobb–Douglas emerges. Also, it can be derived from the translog production form (Christensen, Jorgenson and Lau, 1973) and many others, besides, by judiciously restricting certain parameters. A different approach to the derivation of the Cobb–Douglas form has been taken by P. Zarembka (1987), who specifies each variable as $z(\lambda) = (z^{\lambda} - 1) / \lambda$ for $\lambda \neq 0$ and $z(\lambda) = \ln z$ for $\lambda = 0$. Then, applying this transformation to the production function, we would have $z_0 = q$ and $z_i = x_i$ for all *i*. Thus, if the $z_k (k = 0, 1, \ldots, n)$ are related linearly, the transformation turns out to be a useful procedure in econometrics to treat the general problem of functional form, an important special case of which is the Cobb–Douglas.

In sum, though it is restrictive and sometimes regarded as an economic toy, the Cobb–Douglas form is remarkably robust in a vast variety of applications and that it will endure is hardly in question.

See Also

- capital theory (paradoxes)
- CES production function
- Douglas, Paul Howard
- production functions

Bibliography

Brown, M. 1966. *On the Theory and Measurement of Technological Change*. Cambridge: Cambridge University Press.

Brown, M. and De Cani, J. 1963. Technological change and the distribution of income. *International Economic Review* 4, 289–309.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55, 28–45.

Douglas, P.H. 1948. Are there laws of production? *American Economic Review* 38, 1–41.

Douglas, P.H. 1967. Comments on the Cobb–Douglas production function. In *The Theory and Empirical Analysis of Production*, ed. M. Brown, National Bureau of Economic Research, Studies in Income and Wealth No. 31. New York: Columbia University Press.

Førsund, F.R., Lovell, C.A.K. and Schmidt, P. 1980. A survey of frontier production functions and of their relationship to efficiency measurement. *Journal of Econometrics* 13, 5–25.

Mairesse, J. 1974. *Comparison of Production Function Estimates*. Paris: Institut National de la Statistique et des Etudes Economiques.

Sato, K. 1975. *Production Functions and Aggregation*. Amsterdam: North-Holland.

Wicksell, K. 1958. *Selected Papers on Economic Theory*, ed. Erik Lindahl. Cambridge MA: Harvard.

Zarembka, P. 1987. Transformation of variables in econometrics. In *The New Palgrave: A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

How to cite this article

Brown, Murray. "Cobb–Douglas functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000183> doi:10.1057/9780230226203.0253

The New Palgrave Dictionary of Economics Online

Cobden, Richard (1804 – 1865)

William D. Grampp
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Anti-Corn Law League (UK); Bank of England; Bright, J.; Cobden, R.; colonialism; Corns Laws; Factory Acts; free trade; Manchester School; McCulloch, J. R.; New Poor Law; pacifism; Tooke, T.

Article

Cobden led the campaign that repealed the Corn Laws in 1846, after which there was free trade in grain. The son of a Middlesex farmer, he sought his fortune in Manchester, became an owner of a mill that employed 2,000 workers and was noted for excellence of its calicoes. At 35, he was a rich man.
His calling, however, was politics. After taking part in the successful effort to incorporate Manchester, he entered the movement against the Corn Laws in 1838. Until then it had been conducted by middle class radicals and various business interests, among them the Manchester Chamber of Commerce. Cobden, John Bright, and others like them wanted to enlarge the movement, make it bold and uncompromising. They were exasperated by the businessmen who so wanted to look respectable that they could not see where their interest lay. Thomas Tooke had said the same about the London merchants, when on their behalf he drafted the celebrated petition of 1820 for free trade and they were reluctant to sign it.
The militants of Manchester formed the National Anti-Corn Law League and agitated for free trade up and down the country. They become known as the Manchester School of Economics and were celebrated as arch advocates of laissez-faire. Actually they were a coalition of diverse interests that agreed on only one issue – repeal of the Corn Laws – and each did so for its particular reasons.
Cobden's reason was peace. He believed free trade would break down national barriers and give everyone a material interest in avoiding war. This was not an argument gotten up for the occasion but the expression of a view he had long held. When young he wrote two long tracts on foreign policy which denounced alliances among nations and political engagements of all kinds, decried the idea of a balance of power, was especially disapproving of colonies, then went on to extol free trade as the way to peace and its guarantor. Years later, after he and Bright had brought down the Corn Laws, he told him, ‘I have always had an instinctive monomania against the system of foreign interference, protocolling,

diplomatising, etc.’

That scarcely expressed the horror he had of violent action, even the suggestion of it. When the southern states of America seceded, he thought Lincoln was wrong in bringing the issue to battle although he had no sympathy with them (except their fondness for free trade). He was shocked by the massacres in India and was opposed to wars of independence and to revolution. He thought duelling was barbarous, was against capital punishment, objected to boxing, couldn't stand brass bands, and asked the Pope to prohibit bull fighting in Spain. He favoured free trade so long as its effect was peaceful, as he believed it usually was, but when he believed it was not he quickly put it aside. He opposed the sale of foreign bonds in the London market if the proceeds were to be used to buy arms. ‘No free trade in cutting throats’, he said.

Pacifism, not laissez-faire, was Cobden's guiding principle; and he applied laissez-faire less to domestic than to foreign markets. He did not care for the Factory Acts but only spoke, never voted, against them. He approved of increasing the monopoly powers of the Bank of England and of regulating aspects of railway construction. He had no use for the New Poor Law, of which most economists of the day approved, and spoke derisively of McCulloch's ‘usual dogmatism’. But he carefully read the latter's edition of the *Wealth of Nations* and wrote in the margins of the chapters that moved him. His notes are especially lively where Smith condemns the colonial policy of Great Britain. However, where he describes the operation of the invisible hand, the margins are quite untouched.

See Also

- Corn Laws, free trade and protectionism
- Manchester School

Selected works

1835. *England, Ireland and America, by a Manchester manufacturer*. London.

1836. *Russia, by a Manchester Manufacturer*. Edinburgh.

1849. *Speeches on Peace, Financial Reform and Other Subjects*. London.

1867. *Political Writings*. London: Ridgeway.

1868. *Speeches on Questions of Public Policy*. 2 vols, ed. J. Bright and J.E. Thorold Rogers. Oxford: Oxford University Press.

Bibliography

Ashworth, H. 1877. *Recollections of Richard Cobden, M.P., and the Anti-Corn Law League*. London: Cassell.

Morley, J. 1881. *The Life of Richard Cobden*. London: Fisher Unwin.

Ritchie, J.E. 1865. *The Life of Richard Cobden*. London: Ward and Lock.

How to cite this article

Grampp, William D. "Cobden, Richard (1804–1865)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000185> doi:10.1057/9780230226203.0254

The New Palgrave Dictionary of Economics Online

cobweb theorem

B. Peter Pashigian

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The cobweb theorem purports to explain persistent fluctuations of prices in selected agricultural markets. It was first developed in the 1930s under static price expectations where the predicted price equalled actual price in the last period. Muth's rational expectations hypothesis posited that forecast errors will not be serially correlated and the pattern of past forecast errors cannot be used to improve the accuracy of the forecasts. The fundamental question of whether observed price cycles are better explained by systematic errors in price forecasts or by the cumulative impact of unpredictable shocks has not as yet been definitively addressed.

Keywords

adaptive expectations; cobweb theorem; expectations; Kaldor, N.; price cycles; price expectations; rational expectations

Article

The persistent fluctuations of prices in selected agricultural markets have attracted the attention of economists from time to time, and the theory of the cobweb was developed to explain them. The theory is applicable to those markets where production takes time, where the quantity produced depends on the price anticipated at the time of sale, and where supply at time of sale determines the actual market price. One strand of the cobweb literature (the term was coined by Kaldor, 1934) concentrates on how expectations are formed and the effect of the price expectations mechanism on the stability of equilibrium. Cobweb theory was first developed under static price expectations where the predicted price equalled actual price in the last period. The cobweb theorem proved that the market price would (not) converge to (long-run) equilibrium price if the absolute value of the price elasticity of demand was greater (smaller) than the price elasticity of supply. This stability condition was modified later as more sophisticated expectations models were adopted. Early articles by Tinbergen, Ricci and Schultz

appeared in 1930 in German (see Waugh, 1964, for a review of this literature). Ezekiel's important article (1938) spells out in greater detail the conditions for convergence, divergence or perpetual oscillation and shows how cycles of different lengths could be generated under static expectations. Why the theory was developed in the 1930s and not earlier is a bit of mystery, for recurring price cycles for some agricultural products had been reported by agricultural economists for some time. Economists may have been attracted to the cobweb theory in the 1930s because of the events of the Depression. A theory that explained both oscillation and long departures from stationary equilibrium was more attractive after the events of the Depression. The fact that Ezekiel's paper was reprinted in the 1944 American Economic Association volume on business cycles lends credence to this view.

The impression left by Ezekiel and subsequent contributors is that the cobweb theory is a valuable tool for explaining price cycles. Ezekiel was aware of the simplicity of static expectations and not unmindful of the importance of shocks on the demand and the supply sides of the market in causing aberrant price fluctuations (for example, weather and the randomness of yields). Even so, agriculture economists, who were presumably more familiar with price fluctuations in agricultural markets, have been more prone to accept the theory, while other theorists have given the theory more of a mixed reception.

The price expectations mechanism has undergone many refinements over the years. In 1958 Nerlove proposed the use of adaptive expectations. This suggestion is motivated by the findings of econometric studies which showed the price elasticity of demand to be less than the price elasticity of supply for many agricultural goods. Under these conditions the static expectations version of the cobweb model predicts a price cycle of increasing amplitude. However, the observed price cycles in agricultural markets showed no sign of being explosive. Nerlove attempted to reconcile theory with evidence and to show that convergence is possible under a broader set of conditions provided expectations are adaptive. During the 1930s the attractiveness of the cobweb model seemed to be in its ability to explain persistent or even explosive price cycles. By the late 1950s these were no longer attractive features, and Nerlove felt compelled to offer an explanation of why price cycles of increasing amplitude are not observed even when demand elasticities are smaller than supply elasticities. Waugh (1964) took a different tack and attempted to reconcile the theory with the evidence of stable price cycles by suggesting that the price elasticity of supply becomes smaller (larger) than the price elasticity of demand at prices well above (below) the long-run equilibrium price. Under this assumption, a stable price cycle will eventually be reached.

The length of the cobweb price cycle is determined by the length of the production process. If it takes one year to bring a fattened hug to the market, then the complete price cycle should take two years. At first, little attention and superficial explanations were given to explain why the predicted length is often shorter than the actual length of the price cycle. It was left to the critics to point out these discrepancies. The critics are responsible for the other strand of the literature. They appeared early but were not very influential at first although their criticisms were ultimately given more weight. The critics questioned the rationality of using an arbitrary expectations mechanism by otherwise profit-maximizing agents, and pointed out that the theory implies that producers would expect to lose wealth if they entered and remained in an industry with a cobweb price cycle. In a perceptive article on the pig cycle in England, Coase and Fowler (1935) questioned the realism of static expectations. They showed that the price of a bacon (mature) pig less the cost of feeding for the next five months and less the cost of a feeder (young) pig, which would be stable in a competitive market if farmers had static expectations, fluctuated over time. Hence the empirical evidence contradicted the assumption of static expectations. They presented

evidence that pig breeders reacted quickly to a change in expected profits, and this implied that the pig price cycle should be only two years instead of the observed four-year period. The fluctuation in the profits per pig was attributed to the difficulty of predicting both demand and foreign imports. The Coase–Fowler paper advanced, if only in faint outline, the essence of the rational expectations hypothesis which was to blossom some 35 years later. They hinted that anticipated prices would not be formed in a mechanistic way because profits would be higher the more accurate are the forecasts. Prediction errors were due to the difficulty of predicting shifts in demand and in foreign supply.

Buchanan's paper (1939) criticized the cobweb model because it implied that producers suffer aggregate losses over the price cycle when output is determined by the long-run supply curve. He pointed out that the theory was based on the dubious assumption of a continued supply of entrepreneurs standing ready to dissipate their capital. The critics were also disturbed by the ambiguity of whether the supply curve is of the short- or long-run variety, and the failure to clarify how the adjustment from the short-run to the long-run supply curve is made. These early criticisms and ambiguities aside, references to the cobweb theory continued to appear in textbooks.

Nerlove's paper (1958) briefly rekindled the controversy. His purpose was to resurrect the theory and show that it could explain price behaviour if adaptive expectations were employed. Mills (1961) criticized the use of adaptive and other autoregressive expectations mechanisms in the deterministic model because they implied a simple pattern of forecast errors that producers could detect, incorporate into their forecasts and thereby improve the accuracy of their price forecasts. While Nerlove's suggestion did rectify one limitation of the cobweb theory, it did not address the critical issue of why producers relied on any particular forecasting mechanism. Muth (1961) developed the implications of rational expectations for cobweb theory in his now famous paper. Muth postulated that expectations were the predictions of the economic structure of the market and incorporated all available information. Under certain conditions the predicted price equals the conditional expectation of price, given currently available information. Adaptive expectations can be rational only under special conditions, and the coefficient of adaptation is determined by the values of the slopes of the demand and supply curves.

The rational expectations formulation has powerful implications for cobweb theory. If the price forecasts incorporate all available information and are on average correct, then forecast errors will not be serially correlated and the pattern of past forecast errors cannot be used to improve the accuracy of the forecasts. Moreover, what is then left of the supposed ability of the cobweb theory to explain the cyclical behaviour of prices? Price fluctuations would have to be explained either by the cyclical pattern of exogenous variables or by the summation of random shocks (Slutsky, 1937). Muth's paper represents a frontal attack on the traditional cobweb model. He notes that the traditional model tends to predict a shorter price cycle than is observed and indicates that the rational expectations version predicts a longer price cycle.

Interest in the cobweb model has ebbed in recent years and few articles on it have appeared in the major journals. Economists have found it more rewarding to apply the rational expectations hypothesis to areas like monetary or business-cycle theory than to the study of particular markets, even though the analysis of markets with inventories raises issues that are just as difficult and subtle. The question of whether the cobweb does or does not explain price cycles has not really been resolved. Freeman (1971) has suggested that the traditional cobweb model explains cycles in the markets for lawyers, physicists and engineers. Tests of the rational expectations hypothesis have been suggested by Pashigian (1970) when

expectations data are available and by Hoffman and Schmidt (1981) when expectations data are unavailable. So the methodology exists for distinguishing between the competing hypotheses. Few econometric tests have been made of the rational expectations hypothesis in markets where the assumptions of the cobweb model apply. The fundamental question of whether observed price cycles are better explained by systematic errors in price forecasts or by the cumulative impact of unpredictable shocks has not as yet been definitively addressed.

See Also

- adaptive expectations
- rational expectations

Bibliography

Buchanan, N. 1939. A reconsideration of the cobweb theorem. *Journal of Political Economy* 47, 67–81.

Coase, R.H. and Fowler, R.F. 1935. Bacon production and the pig-cycle in Great Britain. *Economica* 2 (May), 142–67.

Ezekiel, M. 1938. The cobweb theorem. *Quarterly Journal of Economics* 52(February), 225–80.

Freeman, R.H. 1971. *The Market for College-Trained Manpower*. Cambridge, MA: Harvard University Press.

Hoffman, D.L. and Schmidt, P. 1981. Testing the restrictions implied by the rational expectations hypothesis. *Journal of Econometrics* 15, 265–87.

Kaldor, N. 1934. A classificatory note on the determinateness of equilibrium. *Review of Economic Studies* 1(February), 122–36.

Mills, E.S. 1961. The use of adaptive expectations in stability analysis: comment. *Quarterly Journal of Economics* 75, 330–35.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrics* 29, 315–35.

Nerlove, M. 1958. Adaptive expectations and cobweb phenomena. *Quarterly Journal of Economics* 72, 227–40.

Pashigian, B.P. 1970. Rational expectations and the cobweb theory. *Journal of Political Economy* 78, 338–52.

Slutzky, E.E. 1937. The summation of random causes as the source of cyclical processes. *Econometrica* 5(April), 105–46.

Waugh, F.V. 1964. Cobweb models, *Journal of Farm Economics* 46, 732–50.

How to cite this article

Pashigian, B. Peter. "cobweb theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000186> doi:10.1057/9780230226203.0255

The New Palgrave Dictionary of Economics Online

cognitive ability

William T. Dickens

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Modern psychological theory views cognitive ability as multidimensional while acknowledging that the many different abilities are themselves positively correlated. This positive correlation across abilities has led most psychometricians to accept the reality of a general cognitive ability that is reflected in the full scale score on major tests of cognitive ability or IQ. This article provides an introduction to the history of cognitive testing and some of its major controversies. Evidence supporting the validity of measures of cognitive ability is presented and the nature and implications of group differences are discussed along with evidence on its malleability.

Keywords

ability tests; achievement tests; cognitive ability; cultural bias; external validity; factor analysis; heritability; IQ; intelligence; stereotype threat

Article

Some people are obviously and consistently quicker than others to understand new concepts; they solve unfamiliar problems faster, see relationships that others don't and are more knowledgeable about a wider range of topics. We call such people smart, bright, quick, or intelligent. Psychologists have developed tests to measure this trait. Originally called IQ tests (for Intelligence Quotient because the measures were constructed as the ratio of mental age to chronological age multiplied by 100), that name has fallen out of favour. Instead, such tests are now often referred to as tests of cognitive ability. Although the term IQ is still sometimes used to refer to what such tests measure, none constructs a ratio.

History

Spearman (1904) first popularized the observation that individuals who do well at one type of mental

task also tend to do well at many others. For example, people who are good at recognizing patterns in sequences of abstract drawings are also good at quickly arranging pictures in order to tell a story, telling what three-dimensional shapes drawn in two dimensions will look like when rotated, tend to have large vocabularies and good reading comprehension, and are quick at arithmetic. This pattern of moderate to strong positive correlations across the whole spectrum of mental abilities led Spearman to hypothesize the existence of a general mental ability similar to the common notion of intelligence. A person's ability with any particular type of task would be equal to the sum of that person's general ability plus considerations unique to that particular task. Thus general ability could be measured by constructing sub-tests of a number of similar items (individual tasks of the same type such as arithmetic problems) of differing complexity. Each sub-test would present items of a different type, and individual scores across sub-tests could be aggregated. Task specific factors would average out leaving the final score as mainly a measure of general ability or ‘g’. Using an approach like this Binet (1905) developed the first IQ test as a way of identifying students’ academic potential. That test was adapted for use in English by Terman and in 1916 became the Stanford–Binet IQ tests – still one of the most commonly administered tests of cognitive ability.

Spearman's hypothesis of a single general mental ability and many specific abilities was challenged by Thurstone (1935), who popularized the notion that people had a number of independent primary mental abilities rather than a single general mental ability. Both Spearman and Thurstone made contributions to the development of factor analysis as a way to identify the presence of unobserved variables (abilities) that affect a number of observable variables (sub-test or item scores). Today, the Spearman–Thurstone debate has been resolved with a compromise. The most common view among psychometricians who study cognitive ability is that there are a number of different abilities. Some people are better at solving problems verbally while others are good at solving problems that involve visualization. Some people who are good at both of these things may be only average at tasks that rely heavily on memory. However, there is a tendency for people who perform well in any of these broad areas to perform well in all others as well (Carroll, 1993). Most modern tests of cognitive ability provide both a full-scale score that is most reflective of general intelligence, and a number of special-ability specific sub-scores as well.

Validity

Binet's is considered the first successful test of cognitive ability in that it was able to accurately predict teachers’ assessments of their students on the basis of a relatively short verbally administered test. Scores on tests of cognitive ability correlate well with common perceptions of how bright or smart someone is. They are also strongly correlated with measures of academic achievement such as achievement test scores, grades and ultimate educational attainment (typically .5 or better). They are less highly correlated (.5 or less) with many important life outcomes including reported annual income and job status. Performance on a wide range of jobs and work tasks is positively related to cognitive test scores with performance on more demanding jobs having higher correlations. Some have claimed that general cognitive ability is responsible for most of this explanatory power (Ree and Earles, 1992; Ree, Earles and Teachout, 1994). This was a major theme of the controversial best-seller *The Bell Curve* (Herrnstein and Murray, 1994). Heckman (1995), in a review of that book, argues that even though g has significant explanatory power, many other factors, both cognitive and non-cognitive, matter as well.

Finally, test scores are correlated with a number of social behaviours including unwed motherhood, criminal activity, and welfare receipt (Jensen, 1998, ch. 9). While these correlations are substantial, and cognitive test scores are typically better predictors of most of these outcomes than any other single personal attribute, they still explain less than half the variance.

Individuals' scores on tests of cognitive ability also tend to be strongly correlated over time – much more so for adults than for children. A study of older adults found their full-scale IQ scores to be correlated .92 when tested at two points in time three years apart (Plomin et al., 1994). In contrast, a study of children tested at two points in time roughly two years apart found correlations of only .46 for those who were less than one year old at first testing and .76 for those who were one year old at first testing (Johnson and Bradley-Johnson, 2002).

It is common to draw a distinction between tests of achievement and tests of ability. Achievement tests measure how much knowledge the test taker has accumulated in a particular area while ability tests endeavour to measure how quickly a person can solve unfamiliar problems. Typically, scores on the two types of tests are highly correlated. In fact, all tests of ability are, to some degree, tests of achievement as it is impossible to measure ability without also measuring the test taker's reading or verbal comprehension at least. Further, to the extent that the task being tested relies on knowledge of geometry, arithmetic, general knowledge, and so on, the rolls of the achievement test and ability test are confounded.

Cultural bias has been a concern with knowledge-based tests. Some knowledge is more accessible to some people than others. For example, we would expect that a child growing up with upper middle-class parents in New York or Paris to find it easier to learn the distance between the two cities (a general knowledge question that was once on one of the popular IQ tests) than someone from the slums of St. Louis or a tribesman from the bush in Africa. For this reason a number of tests have been constructed that require a minimal amount of prior knowledge, such as Cattell's culture fair test (Cattell, 1960) or Raven's progressive matrices (Raven, 1941).

Group differences

No matter what test is administered, men and women of the same background tend to have very similar average scores on tests of cognitive ability, though they differ slightly in their performance on some sub-tests (Jensen, 1998, pp. 531–6). However, there are large differences across ethnic groups and geographic areas. The difference that has generated the most controversy is the difference in average scores of US blacks and whites, which is typically reported to be about one white standard deviation, though this gap has declined some in recent years (Dickens and Flynn, 2006). Do these represent real differences in cognitive ability or do they reflect cultural bias in the tests?

Defenders of the tests offer several pieces of evidence suggesting that they are unbiased. Foremost is the evidence of 'external validity' – that the same regression equation that predicts outcomes such as job performance, grades, or educational attainment for one group will typically do a similarly good job for any other group. Also, different groups find the same questions more or less difficult. Members of different groups with similar scores will have similar patterns of right and wrong answers. If some questions are more culturally biased than others, the disadvantaged group should find those items more difficult than the mainstream group does. But researchers looking for such cultural bias have found no

evidence of it (an exception occurs when one of the groups being compared is made up of non-native speakers of the language in which the test was administered, in which case scores on questions requiring a better knowledge of the language will be lower). Surprisingly, to the extent that there are black–white differences across test items, blacks do worse on what seem to be some of the least culturally dependent items – those involving abstract or symbolic problem-solving. Differences tend to be smaller on seemingly culturally rich items such as general knowledge. Herrnstein and Murray (1994, Appendix 5) provide a review of the evidence on bias.

The best evidence that tests can be biased in at least some circumstances comes from studies of a phenomenon called stereotype threat. It has been shown that reminding people of their group identity can cause them to perform in ways more consistent with stereotypes of the group's abilities. For example, blacks have been found to perform worse on some particularly difficult vocabulary items when given a questionnaire that asked them to state their race before taking the test or when the test was represented as a test of intelligence as opposed to a test of vocabulary. Women who were told that the difficult math test they were taking generally showed gender differences performed worse than those taking the same test who were told the test showed no differences. Men showed the opposite effect and performed better when told the test showed a gender difference (Steele, 1997). However, it has not been demonstrated that stereotype threat produces substantial bias on standard tests in standard test-taking circumstances.

While most evidence is consistent with the view that tests provide a fair measure of the underlying concept of cognitive ability across ethnic groups, it is not conclusive. For example, since tests rarely explain as much as half the variance in the outcomes in studies of external validity, there is always the possibility that the tests underestimate black cognitive ability but that other disadvantages pull down black performance. If true, the validity of the tests as predictors of practical outcomes is an artifact of offsetting biases. This could explain why it is that when regressions of white performance on white test scores fail to predict black performance they tend to predict better performance than is observed. Further, common-sense notions that people from different cultural backgrounds probably have less opportunity to acquire certain types of information or practise certain skills should be given some weight. If studies find that blacks do no worse than similarly scoring whites on highly culturally loaded items, that could indicate that the poor-scoring whites were similarly disadvantaged. If disadvantage is more common for blacks than whites due to discrimination, that disadvantage could still explain some of the score gap. However, the strong correlation of even the culturally reduced tests with performance, and the similar magnitude of the gap on those tests between groups, suggest that much of the measured gap in ability between groups reflects real differences in average developed ability. This conclusion naturally leads to the consideration of the sources of those differences.

The question of whether individual, and particularly group, differences in cognitive ability are due more to nature or to nurture has been enormously controversial for the last century. Dickens (2005) presents a summary of the evidence on the origin of black–white differences and concludes that they are most likely not substantially genetic in origin. Rushton and Jensen (2005) reach the opposite conclusion. Whatever the right answer, whether the black–white gap has genetic origins is probably the wrong question. It seems that people are concerned with the issue mainly because they confuse having a genetic cause with immutability. While genes almost certainly play a large role in explaining individual differences in cognitive ability within ethnic groups raised in similar circumstances, it also seems that developed cognitive ability is highly malleable.

Malleability

A large amount of evidence has accumulated on the role of genes in explaining individual differences in cognitive ability. Several reviews of this literature conclude that differences in genetic endowment explain somewhere between 60 per cent and 80 per cent of the variance in cognitive ability in representative samples of the adult population in developed countries. The percentage for children is lower than for adults, with most estimates placing it around 40 per cent for six-year-olds (Plomin et al., 2001; Neisser et al., 1996). The figure is also estimated to be lower among disadvantaged populations (Turkheimer et al., 2003) though not consistently (Asbury, Wachs and Plomin, 2005). This figure is referred to as the heritability of cognitive ability. It is estimated by contrasting people with different degrees of relatedness raised in the same home or people with similar relatedness raised in different homes. For example, the correlation of the cognitive ability of identical twins raised in completely independent environments will be equal to the heritability of cognitive ability under the assumptions typically employed to make such estimates. While this evidence establishes that genes play a large role in determining individual differences, little is known about which genes are involved or how they influence cognitive ability (Plomin et al., 2001).

The high heritability of cognitive ability has led some to conclude that people's environments play little role in shaping their ability and that, therefore, individual differences are largely immutable and group differences must be largely due to differences in average genetic endowment. It has been argued that, if all of the observable differences in environment between people produce only 40 per cent or less of the variance in cognitive ability, then the large differences between blacks and whites could not result from the relatively small differences in environment between the average white and the average black. Thus differences in genetic endowment must play a substantial role. A formal version of this argument was first presented by Jensen (1973, pp. 135–9). A similar argument was made by Herrnstein and Murray (1994, pp. 298–9).

Yet despite the high heritability of cognitive ability, it does seem to be quite sensitive to environmental changes. In a review of the effects of early education programmes, Lazar and Darlington (1982, p. 44) noted that ‘The conclusion that a well-run cognitively oriented early education program will increase the IQ sores of low-income children by the end of the programs is one of the least disputed results in educational evaluation’. The gains they surveyed were often quite large, though they also tended to decline substantially after children left the programmes. There is also evidence that being in a cognitively demanding environment can increase measured cognitive ability. Ceci (1991) surveys the evidence on the effects of school attendance on measured ability and finds it to be substantial.

Finally, the most profound changes in measured cognitive ability have taken place over time. James Flynn has documented huge gains in cognitive ability – as much as a standard deviation or more a generation – in more than 14 countries. Numerous other authors have found gains on other tests and in other countries (Flynn, 1987; 1998; 2006). This phenomenon of large and pervasive gains has been dubbed ‘the Flynn Effect’.

How is it that large gains are possible in the face of high heritability estimates? The chief flaw in the argument that high heritability implies a limited role for environment is that it misunderstands what heritability is measuring. It ignores the possibility that genetic and environmental influences might be

correlated. In particular, it ignores the possibility that genetic influences on ability are largely the work of environmental advantages that come about due to modest physiological advantages.

Consider a sports analogy. Identical twins raised apart have a shared genetic endowment that tends to make them notably taller than their peers. As such they are both better basketball players. Even though they are raised apart, both are likely to spend more time playing basketball than other children their age. They are good at it and thus enjoy it more than other activities in which they do not naturally excel. Consequently they both get more practice at basketball than their peers, and that makes them better at the game. Being better players than their peers, they are more likely to be picked by coaches for high-school teams and more likely to receive yet more practice and more intensive coaching. If this leads to them playing in college they will both be enormously better players than the average person. A small physiological difference, which would make only a very modest difference in their performance on the court if they were untrained and inexperienced, has mushroomed into a huge difference in performance because it has been reinforced by the environmental influences of practice and coaching.

It is not hard to imagine the same thing happening with cognitive ability. Someone who is slightly quicker or has an emotional disposition amenable to thought and contemplation will be more likely to spend more time in intellectual pursuits. Such a person will likely receive positive reinforcement from teachers and be more likely to be tracked into more demanding classes and to develop friendships with other similarly disposed children. Such a child will have much more opportunity to practise intellectual work and receive more ‘coaching’ in intellectual pursuits. A small initial physiological difference could mushroom into a large difference in ability through a process whereby the advantage leads to a better environment which improves ability and gives access to even better environments.

If such reciprocal causation is at work in the development of cognitive ability, then small persistent exogenous differences in environment could produce large differences in cognitive ability. Dickens and Flynn (2001) lay out a formal model of such a process. If in a cross section of people in the same ethnic group most exogenous environmental differences are transient, then they will not accumulate through reciprocal causation and will not explain much variance across individuals. However, small persistent differences between groups or generations could cause large differences if they drive the engine of reciprocal causation. Similarly, preschool programmes which enrich children's cognitive environment can have large effects, but once the children are removed from the programme the process can work in reverse and unravel the gains. The exogenous decline in the quality of the environment from the removal of the programme's stimulation sets off a downward spiral of poorer performance leading the child into poorer environments, yet poorer performance and so on.

Conclusion

Modern psychology views cognitive ability as having a number of dimensions, all of which seem to be correlated with one another. Many interpret this correlation as reflecting an underlying general cognitive ability, or g, that is measured by the full-scale scores on the major tests of cognitive ability or IQ. General cognitive ability is an important predictor of a wide range of economic and life outcomes, with similar predictive validity across groups with different average levels of ability. Still, cognitive test scores typically explain far less than half the variance in life outcomes, so cognitive ability is only one important factor among many that explain success.

Adult differences in cognitive ability within representative samples of ethnic groups raised in similar

circumstances are subject to substantial genetic influence, but this does not mean that group differences are genetic in origin. Despite the large role played by genetic differences in explaining adult variance in cognitive ability, there is considerable evidence that intelligence is highly malleable and the life outcomes influenced by intelligence even more so.

See Also

- behavioural genetics

Bibliography

Asbury, K., Wachs, T. and Plomin, R. 2005. Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. *Intelligence* 33, 643–61.

Binet, A. 1905. New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique* 12, 191–244. English translation by Elizabeth S. Kite, *The Development of Intelligence in Children*. Vineland, NJ: Publications of the Training School at Vineland, 1916.

Carroll, J. 1993. *Human Cognitive Abilities: A Survey of Factor-analytic Studies*. Cambridge: Cambridge University Press.

Cattell, R. 1960. *Measuring Intelligence with Culture Fair Tests*. Champaign, IL: Institute for Personality and Ability Testing.

Ceci, S. 1991. How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology* 27, 703–22.

Dickens, W. 2005. Genetic differences and school readiness. *The Future of Children* 15, 55–69.

Dickens, W. and Flynn, J. 2001. Heritability estimates versus large environmental effects: the IQ paradox resolved. *Psychological Review* 108, 346–69.

Dickens, W. and Flynn, J. 2006. Black Americans reduce the racial IQ gap: evidence from standardization samples. *Psychological Science* 17 (forthcoming in October).

Flynn, J. 1987. Massive gains in 14 nations: what IQ tests really measure. *Psychological Bulletin* 101, 171–91.

Flynn, J. 1998. IQ gains over time: toward finding the causes. In *The Rising Curve Long-term Gains in IQ and Related Measures*, ed. U. Neisser. Washington DC: American Psychological Association.

Flynn, J. 2006. Efeito Flynn: repensando a inteligência e seus efeitos [The Flynn effect: rethinking intelligence and what affects it]. In *Introdução à Psicologia das Diferenças Individuais* [Introduction to the psychology of individual differences], ed. C. Flores-Mendoza and R. Colom. Porto Alegre: ArtMed.

Heckman, J. 1995. Lessons from the bell curve. *Journal of Political Economy* 103, 1091–120.

Herrnstein, R. and Murray, C. 1994. *The Bell Curve*. New York: Free Press.

Jensen, A. 1973. *Educability and Group Differences*. New York: Harper and Row.

Jensen, A. 1998. *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.

Johnson, C. and Bradley-Johnson, S. 2002. Construct stability of the cognitive abilities scale-second edition for infants and toddlers. *Journal of Psychoeducational Assessment* 20, 144–51.

Lazar, I. and Darlington, R. 1982. *Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies*. Monographs of the Society for Research in Child Development 47. Chicago: Chicago University Press.

Neisser, U., Boodoo, G., Bouchard Jr., T., Boykin, A., Brody, N., Ceci, S., Halpern, D., Loehlin, J., Perloff, R., Sternberg, R. and Urbina, S. 1996. Intelligence: knowns and unknowns. *American Psychologist* 51, 77–101.

Plomin, R., Defries, J., McClearn, G. and McGuffin, P. 2001. *Behavioral Genetics*, 4th edn. New York: Worth Publishers.

Plomin, R., Pedersen, N., Lichtenstein, P. and McClearn, G. 1994. Variability and stability in cognitive abilities are largely genetic later in life. *Behavior Genetics* 24, 207–15.

Raven, J. 1941. Standardization of progressive matrices. *British Journal of Medical Psychology* 19, 137–50.

Ree, M. and Earles, J. 1992. Intelligence is the best predictor of job performance. *Current Directions in Psychological Science* 1, 86–89.

Ree, M., Earles, J. and Teachout, M. 1994. Predicting job performance: not much more than g. *Journal of Applied Psychology* 79, 518–24.

Rushton, J. and Jensen, A. 2005. Thirty years of research on race differences incognitive ability. *Psychology, Public Policy, and Law* 11, 235–94.

Spearman, C. 1904. General intelligence, objectively determined and measured. *American Journal of Psychology* 15, 201–93.

Steele, C. 1997. A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist* 52, 613–29.

Thurstone, L. 1935. *The Vectors of the Mind*. Chicago: University of Chicago Press.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B. and Gottesman, I. 2003. Socioeconomic status modifies heritability of IQ in young children. *Psychological Science* 14, 623–8.

How to cite this article

Dickens, William T. "cognitive ability." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000139> doi:10.1057/9780230226203.0256

The New Palgrave Dictionary of Economics Online

Cohen Stuart, Arnold Jacob (1855 – 1921)

Arnold Heertje
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cohen Stuart, A. J.; Cournot, A. A.; marginal utility of income; oligopoly; optimal taxation; Pierson, N. G.; progressive and regressive taxation

Article

Born in The Hague, Cohen Stuart was an engineer who took up the challenge put forward by the famous Dutch economist and politician N.G. Pierson to study the mathematical foundations of what we would call nowadays an optimal tax structure. His thesis (Cohen Stuart, 1889) has been reprinted in part (Musgrave and Peacock, 1958).

The international attention to Cohen Stuart's exposition is due to the thorough discussion by F.Y. Edgeworth in his article on the pure theory of taxation (Edgeworth, 1897). Following a lead by Pierson, Cohen Stuart studied the impact of the principle that each taxpayer should sacrifice an equal proportion of the total utility which he derives from material resources. He proved that it depends on the decrease of marginal utility of income, whether the income taxed above a certain minimum will be progressive, regressive or proportional in relation to the level of income. Cohen Stuart argues that in most practical cases a modest progressive tax rate will emerge.

Although based on old-fashioned concepts of measurable utility, Cohen Stuart's contribution to the analysis of the optimal income tax is part of the modern theory of optimal taxation (Mirrlees, 1971) and therefore comparable to Cournot's role in the development of the theory of oligopoly.

See Also

- Pierson, Nichlaas Gerard

Selected works

1889. *Bijdrage tot de theorie der progressieve inkomstenbelasting*. The Hague: Nijhoff.

Bibliography

Edgeworth, F.Y. 1897. *Papers Relating to Political Economy*, vol. 2, London: Macmillan, 1925.

Mirrlees, J.A. 1971. An exploration in the theory of optimum income Taxation. *Review of Economic Studies* 38, 175–208.

Musgrave, R.A. and Peacock, A.T., eds. 1958. *Classics in the Theory of Public Finance*. London: Macmillan.

How to cite this article

Heertje, Arnold. "Cohen Stuart, Arnold Jacob (1855–1921)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000523>
doi:10.1057/9780230226203.0257

The New Palgrave Dictionary of Economics Online

cointegration

Mark W. Watson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes the mathematical structure of cointegrated time series models and discusses econometric procedures commonly used to analyse cointegrated time series. This discussion is carried out in the context of stochastic trends that follow driftless I(1) or ‘unit root’ processes. The article concludes with a brief discussion of cointegration in the context of more general stochastic trends.

Keywords

Beveridge–Nelson decomposition; cointegration; Dickey–Fuller unit root tests; error-correction terms; Granger, C.W.J.; heteroskedasticity and autocorrelation corrections; maximum likelihood; trend/cycle decomposition; unit roots; variance; vector autoregressions; vector error correction model; vector moving average models

Article

Cointegration means that two or more time series share common stochastic trends. Thus, while each series exhibits smooth or trending behaviour, a linear combination of the series exhibits no trend. For example, short-term and long-term interest rates are highly serially correlated (so they are smooth and in this sense exhibit a stochastic trend), but the difference between long rates and short rates – the ‘term spread’ – is far less persistent and shows no evidence of a stochastic trend. Long rates and short rates are cointegrated.

The concept of cointegration was formalized by Clive W.J. Granger in a series of papers in the 1980s (Granger, 1981; Granger and Weiss, 1983; Granger, 1986; Engle and Granger, 1987), and in 2003 Granger received the Nobel Prize in Economics for this work. A flurry of research activity followed Granger's original contributions in this area and produced a practical set of econometric procedures for analysing cointegrated time series.

Mathematical structure of I(1) cointegrated models

Let X_t denote a scalar I(1) stochastic process, with moving average representation $X_t = c(L)\varepsilon_t$, where ε_t is a scalar white noise process, and $c(L) = \sum_{i=0}^{\infty} c_i L^i$ is a polynomial in the lag operator L, and where the moving average coefficients, c_i, decay sufficiently rapidly so that $\sum_{i=1}^{\infty} i|c_i| < \infty$. The Beveridge–Nelson decomposition (see trend/cycle decomposition) implies that X_t can be represented as $X_t = \tau_t + a_t$, where τ_t is a random walk, so that $\tau_t = \tau_{t-1} + e_t$, where e_t is white noise and a_t has a moving average representation $a_t = d(L)\varepsilon_t$, where $\sum_{i=1}^{\infty} |d_i| < \infty$. Thus, X_t can be expressed as the sum of a stochastic trend, τ_t, and an I(0) process, a_t.

When X_t is an $n \times 1$ vector of I(1) processes, a similar result implies that $X_t = A\tau_t + a_t$, where A is a matrix of constants, τ_t is a vector of random-walk stochastic trends, and a_t is a vector of I(0) processes. Because X_t contains n elements, the vector τ_t will generally contain n stochastic trends. However, when τ_t contains only $k < n$ stochastic trends, A is $n \times k$, so that $\alpha' A = 0$, for any vector α in the null space of the column space of A. This means that $\alpha' X_t = \alpha' a_t$, so that the linear combination $\alpha' X_t$ does not depend on the stochastic trends. In this case, the time series making up X_t are said to be cointegrated. Any non-zero vector α that satisfies $\alpha' A = 0$ will annihilate the stochastic trend in $\alpha' X_t$, and vectors with this property are called cointegrating vectors. When A has full column rank, the number of linearly independent cointegrating vectors is $r = n - k$, which is called the cointegrating rank of the process.

For example, suppose that X_t contains $n = 3$ series representing interest rates on one-month, three-month and six-month US treasury bills. Suppose that $X_{it} = \tau_t + a_{it}$, for $i = 1, 2, 3$, where τ_t is a common stochastic trend shared by the three interest rates. Then $X_t = A\tau_t + a_t$, where $k = 1$ (there is a single stochastic trend), $A = (1\ 1\ 1)'$ (the trend has an equal effect on each of the interest rates) and $\alpha_1 = (1\ 0\ -1)'$ and $\alpha_2 = (1\ 0\ -1)'$ are two linearly independent cointegrating vectors, so that $r = 2$ and $\alpha_1' X_t$ and $\alpha_2' X_t$ denote the interest rate term spreads.

Vector moving average models (VMAs) and vector autoregressions (VARs) are often used to represent the linear properties of vector stochastic processes. The Granger representation theorem (see Engle and Granger, 1987) shows that VMAs and VARs for cointegrated processes have special structures. In general, the VMA for an I(1) vector process is $\Delta X_t = D(L)\varepsilon_t$, where ε_t is white noise with full rank covariance matrix. When X_t is not cointegrated, the $n \times n$ matrix $D(1)$, which contains the sum of the moving average coefficients, has rank n. But, when X_t is cointegrated, $D(1)$ has rank $k < n$, where k denotes the number of stochastic trends. When X_t is not cointegrated, the VAR for X_t can be written in terms of ΔX_t and has the form $\Phi(L)X_t = \varepsilon_t$, where $\Phi(L)$ is a stable lag polynomial (so its roots are outside the unit circle) and ε_t is white noise. When X_t is cointegrated, the VAR has the form $\Phi(L)X_t = \beta\alpha' X_{t-1} + \varepsilon_t$, where α is an $n \times r$ matrix with columns that are the linearly independent cointegrating vectors. Thus, the cointegrated VAR expresses the elements of ΔX_t as functions of its own

lags, but also includes the r regressors $\alpha' X_{t-1}$ in each of the VAR's n equations. The variables $\alpha' X_{t-1}$ are called ‘error-correction terms’ and the cointegrated VAR is called a ‘vector error correction model’ (VECM).
Watson (1994) provides a summary of the algebra linking these various representations of the cointegrated model.

Testing for cointegration

The time series making up X_t are cointegrated if the linear combinations $\alpha' X_t$ are I(0) random variables. If X_t is not cointegrated, then $\alpha' X_t$ will be I(1) for any non-zero vector α. Tests of cointegration ask whether $\alpha' X_t$ is I(1) or I(0).
Consider the simple case in which there is only one potential cointegrating vector, so that $\alpha' X_t$ is a scalar. Cointegration can then be tested using a unit root test applied to $\alpha' X_t$. The straightforward application of a unit root test requires that α is known, so that the scalar variable $\alpha' X_t$ can be calculated directly from the data. This is possible in many empirical applications (such as the interest rate example described above) where the value of α can be pre-specified.
Thus, suppose that α is known, and consider the competing hypotheses $H_{I(1)}$: $\alpha' X_t$ is I(1) and $H_{I(0)}$: $\alpha' X_t$ is I(0). The hypothesis $H_{I(1)}$ means that the elements of X_t are not cointegrated and the hypothesis $H_{I(0)}$ means that the elements are cointegrated. Under $H_{I(1)}$ the autoregressive model for $\alpha' X_t$ contains a unit root, while under $H_{I(0)}$, the autoregressive model for $\alpha' X_t$ is stable.
The null $H_{I(1)}$ can be tested against the alternative $H_{I(0)}$ using an augmented Dickey–Fuller (ADF) unit root test or the modified ADF test developed in Elliott, Rothenberg and Stock (1996). The null $H_{I(0)}$ can be tested against $H_{I(1)}$ using the best local test proposed by Nyblom (1989), modified for serial correlation as described in Kwiatkowski et al. (1992), or a point-optimal test as discussed in Jansson (2004). (There are important practical considerations associated with the choice of the long-run-variance estimator (see heteroskedasticity and autocorrelation corrections) used in tests for the $H_{I(0)}$ null hypothesis because of the high degree of serial correlation under the alternative. See Müller (2005) for discussion.)

When α is not known, the unit root tests described in the last paragraph use $\hat{\alpha}' X_t$ in place of $\alpha' X_t$, where $\hat{\alpha}$ is an estimator of α. For example, Engle and Granger (1987) suggest estimating α by regressing the first element of X_t onto the other elements of X_t using OLS, and carrying out an ADF test using the residuals from this regression. Estimation of α changes the distribution of the ADF test statistic from what it is when α is known, so that critical values for the Engle–Granger test are different than the standard ADF critical values. As described in Phillips and Ouliaris (1990) and Hansen (1992) the correct critical values depend on the number of elements in X and on the properties of the deterministic trends in the model. Stock and Watson (2007) tabulate choices of critical values from the Phillips and Ouliaris (1990) and Hansen (1992) papers that are appropriate for data that follow I(1)

processes that may or may not contain drift, and thus serve as conservative critical values. Modifications for tests of the $H_{I(0)}$ null versus the $H_{I(1)}$ alternative are discussed in Shin (1994) and Jansson (2005). The tests outlined above are useful for testing whether a single series $\alpha' X_t$ is I(0) or I(1), but in many applications there may be more than one potential cointegrating relation ($r > 1$) so that it is useful to have tests for hypothesis that postulate different values of r. That is, it is useful to entertain hypotheses of the form $H_j: r = j$, for j=0, 1, … n. The hypothesis $r = 0$ means that there is no cointegration, $r = 1$ means that there is a single cointegrating vector, and so forth. As discussed in Johansen (1988), these tests are easily formulated and carried out using the VECM model. Recall that the VECM model has the form $\Phi(L)X_t = \beta\alpha' X_{t-1} + \varepsilon_t$. Consider the null and alternative hypotheses $H_o: r = r_o \; vs. \; H_a: r = r_a$ where $r_a > r_o$, and write the VECM as $\Phi(L)X_t = \beta_o\alpha_o' X_{t-1} + \tilde{\beta}\tilde{\alpha}' X_{t-1} + \varepsilon_t$, where α_o contains the r_o cointegrating vectors under the null and $\tilde{\alpha}$ contains the additional cointegrating vectors under the alternative. Under the null hypothesis, the variables $\tilde{\alpha}' X_{t-1}$ do not enter the VECM, while under the alternative these variables enter the VECM. Thus, the null and alternative can be written as $H_o: \tilde{\beta} = 0$ versus $H_a: \tilde{\beta} \neq 0$. As in the case of $r = 1$, the tests depend on whether the cointegrating vectors are known or unknown. When the cointegrating vectors are known, the regressors $\alpha_o' X_{t-1}$ and $\tilde{\alpha}' X_{t-1}$ can be constructed from the data, and the Wald test for $\tilde{\beta} = 0$ can be constructed using the usual regression formula. When the cointegrating vectors are unknown, the testing problem is more difficult, but Johansen (1988) provides an simple formula for the likelihood ratio test statistic. In either case, the critical values for the test are ‘non-standard’, that is, they are not based on the χ^2 or F distributions. Critical values for the tests depend on the values of $r_a - r_o$, the number of cointegrating vectors that are known and unknown, and the presence or absence of constants and time trends in the model. The various critical values are tabulated in Horvath and Watson (1995).

Estimating unknown cointegrating coefficients

Unknown coefficients in cointegrating vectors are typically estimated using least squares and Gaussian maximum likelihood estimators (MLEs). The properties of these estimators can be understood by considering a simple bivariate model

$$X_{1t} = \theta X_{2t} + \eta_{1t}$$

$$X_{2t} = X_{2t-1} + \eta_{2t}$$

where $\eta_t = [\eta_{1t}\ \eta_{2t}]' \sim \text{iid N}(0, \Sigma)$. In this example, there is one common trend that coincides with X_{2t}, the cointegrating vector is $\alpha = (1 - \theta)'$ where θ is an unknown parameter, the error correction term is $\alpha' X_t = \eta_{1t}$ which is potentially correlated with the innovation in the common trend, η_{2t}, and the assumption of normality is used to motivate the Gaussian MLE of θ.

The OLS estimator of θ has several interesting properties (Stock, 1987). Even though X_{2t} and η_{1t} are correlated, the OLS estimator is consistent; indeed it is 'super-consistent' in the sense that $\hat{\theta}^{OLS} - \theta \sim O_p(T^{-1})\hat{\theta}^{OLS}$, so that $\hat{\theta}^{OLS}$ converges to θ faster than the usual $\sqrt{T}$ rate familiar from regressions involving I(0) variables. These results follow because, in the cointegrated model, the regressor X_{2t} is I(1) and therefore is much more variable than an I(0) regressor $(\Sigma_{t=1}^{T} X_{2t}^2 \sim O_p(T^{-2})$ in this I(1) regression instead of $O_p(T^{-1})$ in the usual I(0) regression), and the correlation between X_{2t} and η_{1t} is non-zero, but vanishes as the sample size becomes large. (The covariance is constant, but the variance of X_{2t} increases linearly with t, so the correlation vanishes as t increases.)

Despite these intriguing and powerful features, the OLS estimator has two properties that make it unsatisfactory for many uses. First, while OLS is consistent, the correlation between the regressor and error term induces a bias in the large sample distribution of the estimator, and this bias can be severe in sample sizes typically encountered in applied work (Stock, 1987). Second, the large-sample distribution of the OLS estimator is non-normal, and this complicates statistical inference. For example, the standard interval $\hat{\theta}^{OLS} \pm 1.96SE(\hat{\theta}^{OLS})$ does not provide a 95 per cent confidence set even in large samples. Interestingly, Gaussian maximum likelihood estimators share the super consistency properties of OLS, but do not suffer from these unsatisfactory properties (Johansen, 1988; Phillips, 1991).

To construct the Gaussian MLE, factor the joint density of $\{X_t\}_{t=1}^{T}$ into the density of $\{X_{1t}|(X_{2t})_{t=1}^{T}\}_{t=1}^{T}$ and the density of $\{X_{2t}\}_{t=1}^{T}$. The density of $\{X_{2t}\}_{t=1}^{T}$ does not depend on θ, and the density of $X_{1t}|(X_{2t})_{t=1}^{T}$ is characterized by the Gaussian linear regression $X_{1t} = \theta X_{2t} + \beta X_{2t} + v_t$, where β is the regression coefficient from the regression of η_{1t} onto η_{2t} ($= \Delta X_{2t}$), ν_t is the error in this regression, and $v_t|(X_{2t})_{t=1}^{T} \sim \text{iid N}(0, \sigma_2)$. Simple calculations (Phillips, 1991) can then be used to show that $\hat{\theta}^{MLE} - \theta \sim O_p(T^{-1})$ and that $\hat{\theta}^{MLE}|(X_{2t})_{t=1}^{T} \sim \text{N}(\theta, V)$, where V depends on $(X_{2t})_{t=1}^{T}$. Thus, $\hat{\theta}^{MLE}$ is consistent, is conditionally normally distributed and unbiased, and $(\hat{\theta}^{MLE} - \theta) / V^{1/2} \sim \text{N}(0, 1)$, so that inference about θ can be carried out using standard methods associated with the Gaussian linear regression model. Thus, for example, $\hat{\theta}^{MLE} \pm 1.96SE(\hat{\theta}^{MLE})$ provides a valid 95 per cent confidence set for θ, where $SE(\hat{\theta}^{MLE})$ is computed using the usual regression formula.

While these results may appear quite special (X_t is bivariate and η_t is normally distributed and serially uncorrelated) they carry over to more general models with minor modifications. For example, X_{1t} and

X_{2t} may each be vectors and the regression $X_{1t} = \theta X_{2t} + \beta X_{2t} + v_t$ becomes a multivariate regression. Under weak assumptions on the distribution of η_t, there is sufficient averaging so that $V^{-1/2}(\hat{\theta}^{MLE} - \theta) \xrightarrow{d} N(0, I)$, meaning that the assumption of normality for η is not critical (although $\hat{\theta}^{MLE}$ still refers to the MLE computed by maximizing the Gaussian likelihood). Serial correlation in η_t can be handled in a variety of ways. For example, Saikkonen (1991) and Stock and Watson (1993) consider the 'dynamic OLS' (DOLS) regression $X_{1t} = \theta X_{2t} + \Sigma_{i=-k}^{k} \beta_i X_{2t-i} + v_t$, which includes enough leads and lags of ΔX_{2t} to insure that v_t is (linearly) independent of $(X_{2t})_{t=1}^{T}$. Phillips and Hansen (1990) and Park (1992) develop adjustments based on long-run covariance matrix estimators, and Johansen (1988) derives the exact Gaussian MLE based on the VECM. Under general assumptions, all of the estimators are asymptotically equivalent.

Alternative models for the common trends

The concept of cointegration involves variables that share common persistent 'trend' components. The statistical analysis outlined above utilized a particular model of the trend component, namely, the driftless unit root process $\tau_t = \tau_{t-1} + e_t$. Analysis of this model highlights many of the key features of cointegrated processes, but more general models are often needed for empirical analysis. For example, constant terms are often added to the model to capture non-zero means of error correction terms or drifts in the trend process. These constant terms change the distribution of test statistics for cointegration in ways familiar from the effect of constants and time trends in Dickey–Fuller unit root tests (see Hamilton, 1994). Hansen (1992) and Johansen (1994) contain useful discussion of the key issues. Higher-order integrated processes (for example, I(2) processes) are discussed in Johansen (1995), Granger and Lee (1990), and Stock and Watson (1993). Hylleberg et al. (1990) discuss cointegration at seasonal frequencies. Robinson and Hualde (2003) and the references cited therein discuss cointegration in fractionally integrated models.

Elliott (1998) discusses cointegrated models in which the trend follows a 'near-unit-root' process – an AR process with largest autoregressive root very close to 1.0. (Formally, the asymptotics use a local-to-unity nesting with largest root AR root equal to 1–*c/T*, where *c* is a constant.) Elliott shows that, while the basic cointegrated model remains unchanged in this case, the properties of Gaussian maximum likelihood estimators of unknown cointegrating coefficients change in important ways. In particular, the Gaussian MLEs are no longer conditionally unbiased, and confidence intervals constructed using Gaussian approximations (for example, $\hat{\theta}^{MLE} \pm 1.96SE(\hat{\theta}^{MLE})$) can be very misleading. Elliott's critique is important because small deviations from exact unit roots cannot be detected with high probability, and yet small deviations may undermine the validity of statistical inferences constructed using large-sample normal approximation applied to Gaussian MLEs.

Several papers have sought to address the Elliott critique by developing methods with good performance for a range of autoregressive roots close to, but not exactly equal to 1.0. For example, Wright (2000) argues that if θ_0 is the true value of a cointegrating coefficient, then $X_{1t} - \theta_0 X_{2t}$ will be I(0), but if

θ_0 is not the true value then $X_{1t} - \theta_0 X_{2t}$ will be highly persistent. He suggests testing that $\theta = \theta_0$ by testing the $H_{I(0)}$ null for the series $X_{1t} - \theta_0 X_{2t}$. Alternative testing procedures in this context are proposed in Stock and Watson (1996) and Jansson and Moreira (2006).

See Also

- heteroskedasticity and autocorrelation corrections
- trend/cycle decomposition
- unit roots

Bibliography

Elliott, G. 1998. The robustness of cointegration methods when regressors almost have unit roots. *Econometrica* 66, 149–58.

Elliott, G., Rothenberg, T.J. and Stock, J.H. 1996. Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–36.

Engle, R.F. and Granger, C.W.J. 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica* 55, 251–76.

Granger, C.W.J. 1981. Some properties of time series data and their use in econometric specification. *Journal of Econometrics* 16, 121–30.

Granger, C.W.J. 1986. Developments in the study of co-integrated economic variables. *Oxford Bulletin of Economics and Statistics* 48, 213–28.

Granger, C.W.J. and Lee, T.H. 1990. Multicointegration. *Advances in Econometrics* 8, 71–84.

Granger, C.W.J. and Weiss, A.A. 1983. Time series analysis of error-correction models. In *Studies in Econometrics, Time Series, and Multivariate Statistics, in Honor of T. W. Anderson*, ed. S. Karlin, T. Amemiya and L.A. Goodman. San Diego: Academic.

Hamilton, J.D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, B. 1992. Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends. *Journal of Econometrics* 53, 86–121.

Horvath, M.T.K. and Watson, M.W. 1995. Testing for cointegration when some of the cointegrating vectors are prespecified. *Econometric Theory* 11, 952–84.

Hylleberg, S., Engle, R.F., Granger, C.W.J. and Yoo, B.S. 1990. Seasonal integration and cointegration. *Journal of Econometrics* 44, 215–38.

Jansson, M. 2004. Stationarity testing with covariates. *Econometric Theory* 20, 56–94.

Jansson, M. 2005. Point optimal tests of the null of hypothesis of cointegration. *Journal of Econometrics* 124, 187–201.

Jansson, M. and Moreira, M. 2006. Optimal inference in regression models with integrated regressors. *Econometrica* 74, 681–714.

Johansen, S. 1988. Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control* 12, 231–54.

Johansen, S. 1994. The role of the constant and linear terms in cointegration analysis of non-stationary variables. *Econometric Reviews* 13, 205–29.

Johansen, S. 1995. A statistical analysis of cointegration for I(2) variables. *Econometric Theory* 11, 25–59.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* 54, 159–78.

Müller, U.K. 2005. Size and power of tests for stationarity in highly autocorrelated time series. *Journal of Econometrics* 128, 195–213.

Nyblom, J. 1989. Testing for the constancy of parameters over time. *Journal of the American Statistical Association* 84, 223–30.

Park, J.Y. 1992. Canonical cointergrating regressions. *Econometrica* 60, 119–43.

Phillips, P.C.B. 1991. Optimal inference in cointegrated systems. *Econometrica* 59, 283–306.

Phillips, P.C.B. and Hansen, B.E. 1990. Statistical inference on instrumental variables regression with I (1) processes. *Review of Economic Studies* 57, 99–124.

Phillips, P.C.B. and Ouliaris, S. 1990. Asymptotic properties of residual based test for cointegration. *Econometrica* 58, 165–93.

Robinson, P.M. and Hualde, J. 2003. Cointegration in fractional systems of unknown orders. *Econometrica* 71, 1727–66.

Saikkonen, P. 1991. Asymptotically efficient estimation of cointegrating regressions. *Econometric Theory* 7, 1–21.

Shin, Y. 1994. A residual-based test of the null of cointegration against the alternative of no cointegration. *Econometric Theory* 10, 91–115.

Stock, J.H. 1987. Asymptotic properties of least squares estimates of cointegrating vectors. *Econometrica* 55, 1035–56.

Stock, J.H. and Watson, M.W. 1993. A simple estimator of cointegrated vectors in higher-order integrated systems. *Econometrica* 61, 783–820.

Stock, J.H. and Watson, M.W. 1996. Confidence sets in regression with highly serially correlated regressors. Manuscript, Department of Economics, Princeton University.

Stock, J.H. and Watson, M.W. 2007. *Introduction to Econometrics*, 2nd edn. Boston: Pearson-Addison Wesley.

Watson, M.W. 1994. Vector autoregression and cointegration, vol. 4. *Handbook of Economics*, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Wright, J.H. 2000. Confidence sets for cointegrating coefficients based on stationarity tests. *Journal of Business and Economic Statistics* 18, 211–22.

How to cite this article

Watson, Mark W. "cointegration." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000579> doi:10.1057/9780230226203.0258

The New Palgrave Dictionary of Economics Online

Colbert, Jean-Baptiste (1619 – 1683)

D.C. Coleman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Colbert, J.-B.; Colbertism; mercantile system; mercantilism

Article

Colbert was born at Reims on 29 August 1619 and died on 6 September 1683. In no way at all could he be called an economist. He was, however, one of the most powerful administrators, known to history, of measures affecting the economic life of a nation, to such an extent and with such lasting influence that his name is preserved in the notion of Colbertism.

He came of a mercantile family which had acquired some public offices. He learned his job as economic administrator by entering the service, in 1651, of a man he was effectively to succeed, Cardinal Mazarin. Once successfully installed in the service of Louis XIV, after Mazarin's death in 1661 his climb to power was rapid. He soon came to hold numerous offices of state: finance, commerce, buildings, the navy, and more besides. His achievements rested in part upon his exercising virtually undisputed power for 22 years as the dominant minister of the grandest of absolute monarchs, and in part upon his own qualities of character which he brought to bear upon the economic problems of France as he perceived them. Those qualities included energy, tenacity, shrewdness, honesty, a notable ability to deploy the techniques of the courtier, and a wholly remarkable capacity for hard work. His hand was felt in every aspect of French economic life; and everywhere he exercised that passion for order which is so often the hallmark of the bureaucrat. Adam Smith sniffed at him as a ‘laborious and plodding man of business … accustomed to regulate the different departments of public offices’ (Smith, 1776, p. 627). But he was a lot more than that. Cold, humourless, and devoted, he was the super-servant of a super-king.

Those qualities did not, on the other hand, include any original economic ideas whatever. He had absorbed, with characteristic thoroughness, all the assumptions, maxims, dogmas, and assorted notions about economic matters which circulated in 16th- and 17th-century Europe, and to which the label of mercantilism has become attached. Consequently, by dint of his position and activities, and because a very large volume of his papers have survived for the historian, he has come down to posterity as the embodiment of conventional mercantilism in practice. Non-existent as a theoretical entity, mercantilism

has acquired the appearance of a coherent economic policy probably more from Colbert's activities than from any other single historical source. And because it appeared, and was continued after his death, in the grandeur which was France, it was copied or adapted in other aspiring monarchies. French mercantilism or Colbertism thus became a recognizable reality in a way that the English ‘mercantile system’ did not.

The nature of his economic ideas can often be gathered from the explanatory memoranda which he addressed to Louis XIV (who was not always as interested in such matters as Colbert thought he should be). They have a familiar ring. He wanted money circulating in the kingdom, not because he identified money with wealth, but because it facilitated the payment of taxes and helped to stimulate economic activity; those branches of overseas trade which brought in precious metals were therefore to be especially favoured. Manufacturing industry deserved encouragement because it lessened French dependence on imports, because it was the basis of an export trade which brought in wealth, and because it employed the idle (the Catholic Colbert had the zeal for work and the disapproval of idleness normally thought of as peculiar to Puritanism). In the interest of the economic unification of France, internal trade and transport needed improvement by the removal of tolls and the repair of roads and bridges. Royal support was needed, and was secured, for the construction of canals – of which the most spectacular achievement was the opening in 1681 of the Canal des Deux Mers, providing a waterway between the Atlantic and the Mediterranean.

Colbert shared the pervasive belief in a fixed cake of trade, so that, as he patiently explained to Louis in March 1669, the whole trade of Europe was carried in a fixed number of vessels and therefore ‘le commerce cause un combat perpétuel en paix et en guerre entre les nations de l'Europe, à qui on emportera la meilleure partie’. The Dutch, the English and the French were the ‘acteurs de ce combat’ (*Lettres* VI, p. 266). France's gain was to be secured by Holland's and/or England's loss. It followed that shipbuilding should be encouraged and the French navy and mercantile marine greatly enlarged. France should move in on trades hitherto dominated by her rivals. Hence his setting up in the 1660s of privileged trading companies: a French East India Company, a French West India Company to improve and exploit French colonies, and the Company of the North to tap the Baltic trade. Such views also provided an economic justification for the war which Louis launched against Holland in 1672.

Colbert had to find the revenue for these and others of his master's military activities. Consequently, he devoted much time to trying to reform the royal finances. Many of his measures – for example, to improve the collection of taxes or to unify the customs system – were thus again part of a policy designed to improve the performance of the economy so that it could in turn yield more wealth to the greater glory of *le roi soleil*.

How much success attended Colbert's policies has been a matter of debate. Laissez-faire economists and economic historians of similar views have inevitably disparaged them and stressed the rigidities which were built into the French economy in the 18th century. His efforts to unify the chaotic diversity of French fiscal and customs administration were only very partially successful; his overseas trading companies were inadequately financed and generally unprofitable; his comparative neglect of agriculture left the basis of the economy in a poor state. But his work did greatly improve the size and efficiency of the French navy and mercantile marine; stimulate – albeit at a high cost – certain areas of French manufacturing industry; and encourage French merchant enterprise in branches of trade hitherto the preserve of others. Not all of this was evident in his own lifetime. But one thing was: Colbert died a very rich man, ennobled as Marquis de Seignelay, his brothers and sisters and cousins amply provided with

lucrative sinecures, his sons as ministers or army officers, and his three daughters married off to dukes. Such were the 17th-century rewards of administering an economy.

Bibliography

Clément, P., ed. 1861–2. *Lettres, Instructions et Mémoires de Colbert*, 8 vols. Paris.

Cole, C.W. 1939. *Colbert and a Century of French Mercantilism*, 2 vols. New York: Columbia University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. E. Cannan, New York: Modern Library, 1937.

How to cite this article

Coleman, D. C. "Colbert, Jean-Baptiste (1619–1683)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000197> doi:10.1057/9780230226203.0259

The New Palgrave Dictionary of Economics Online

collective action (new perspectives)

David P. Myatt
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Olson's logic of collective action predicts that public-good provision is most likely to fail when the size of the consumer group is large; his public goods are partially rival, and so the private cost of provision is relatively high. With a pure public good, this logic no longer applies, and so attention turns to producer groups. When provision involves teamwork (so that the collective action succeeds when everyone works together) then coordination problems arise. Modern techniques suggest that ‘good’ equilibria in which provision is successful are robust only when the costs of provision fall below private rather than social benefits.

Keywords

chicken games; collective action; Cournot contributions games; critical mass theory; equilibrium-selection problem; externalities; global games; interdependent consumption; interdependent production; market failure; multiple equilibria; Olson, M.; provision games; public goods; risk dominance; Schelling, T.; selective incentives; strategic voting; strategy revision; teamwork dilemma; volunteer's dilemma

Article

In a review conducted on behalf of the UK Government, Stern (2007) concluded that ‘climate change is a serious global threat, and demands an urgent global response … the benefits of strong and early action far outweigh the economic costs of not acting’. The cuts in emissions that he suggested could generate global benefits. However, the costs would be borne individually by those making significant cuts (developed nations) or by those sacrificing future opportunities (rapidly developing nations).

A shared desire to cut greenhouse-gas emissions generates a classic problem of collective action: a group with common interests must rely on voluntary individual optimization for the pursuit of those interests. Stern's ‘urgent global response’ to a ‘serious global threat’ requires nations to act. Such sovereign states need respond only to their own incentives; any participation is voluntary. Within each state, the pursuit of national objectives is not automatic; environmental effects stem from the decisions of individual agents. Even if it were in a state's collective interest to support a collective action against climate change, it cannot be assumed that constituents of that state would individually offer their backing.

To economists, the collective-action problem boils down to the private provision of a public good or the private exploitation of a common resource. Law and order, military defence and pollution control are classic textbook examples of public goods: the benefits of provision are non-excludable, and so private providers fail to capture the full impact of their contributions. This market failure leads to inefficient under-provision. On the other hand, the commons exploitation of traffic congestion and commercial fishing yield negative externalities: market failure yields to inefficient overindulgence in these activities. In both cases, individuals fail to pursue efficiently their collective objectives.

The idea that group members will not always pursue their common interests was once not accepted widely. In his article in the first edition of the *New Palgrave*, Mancur Olson (1987) observed that ‘economists, like specialists in other fields, often took it for granted that groups of individuals with common interests tended to act to further those common interests, much as individuals might be expected to further their own interests’. He persuasively argued that ‘the existence of a common interest need not provide any incentive for an individual action in the group interest’. Hence consumers may fail to campaign for their collective protection, unions may fail to protect all their members, oligopolists may fail to maintain collusive prices, and nations of the world may fail to prevent further climate change.

Olson's point was simple and is now familiar: when contemplating choice, individuals consider only the private impact of their actions. For the classic case of a public good, an

individual faces the full marginal cost of provision but fails to account for the benefit spilling over to others; the presence of positive externalities leads to under-provision. If an individual could internalize these externalities, perhaps by excluding the consumption of others and charging them for it, then efficiency could be restored. Alas, pure public goods are non-excludable, and hence this route to efficiency is blocked.

Nevertheless, as long as individuals enjoy some private benefit from voluntary action then we can expect some, albeit too little, provision of public goods. The extent of any inefficiency depends upon the nature of the collective-action problem, the availability of mechanisms to restore efficiency, and the size and nature of the relevant group. Olson (1965) concluded that ‘unless the number of individuals in a group is quite small, or unless there is coercion or some other special device to make individuals act in their common interest, rational, self-interested individuals will not act to achieve their common or group interests’. In the context of small groups, when partial provision is deemed possible, he identified ‘a surprising tendency for the ‘exploitation’ of the great by the small’. These claims led to his theory of groups: (*a*) collective actions fail when the groups are large; (*b*) larger factions bear a disproportionate share of any provision; and (*c*) selective incentives are necessary if groups are to succeed. These three claims are considered in turn, before attention turns to a rather different perspective on collective action.

The first claim is Olson's ‘group size’ hypothesis: private provision should fall as a group grows larger. Olson (1965) painted a picture of a meeting at which too few people make careful contributions: ‘When the number of participants is large, the typical participant will know that his own efforts will probably not make much difference to the outcome, and that he will be affected by the meeting's decision in much the same way no matter how much or how little effort he puts into studying the issues.’ More directly, the claim is that the private benefit of any voluntary contribution falls with the group's size; equivalently, the private cost for any particular level of public provision rises with the group size. This claim leans on two implicit assumptions. First, an increase in the number consuming the good leads to an increase in the provision cost, and hence the good is (at least partially) rival; it is an impure public good. Second, the group size corresponds to the number of consumers, and not to the size of the contributor pool.

These two implicit assumptions that underpin the group-size hypothesis are often valid. For instance, the global climate change that worried Stern (2007) corresponds to a ‘large group’ global collective-action problem (Sandler, 2004). Nevertheless, the assumptions often exclude interesting collective-action problems. The first assumption rules out pure public goods. Consider, for instance, the contemporary voluntary provision of open-source software (Raymond, 1998; Johnson, 2002; Lerner and Tirole, 2002). The typical licence under which such software is distributed ‘requires that the source code … be made available to everyone, and that the modifications made by its users also be turned back to the community’ (Lerner and Tirole, 2001). This a modern instance of the ‘collective invention’ documented by Allen (1983). Open-source software is automatically non-excludable. Of course, software is a classic instance of a non-rival good: consumption by one individual does not hamper the consumption opportunities of others. Hence, an increase in the size of the group consuming the good, while fixing the size of the group able to provide it, has no direct impact on incentives.

Olson's second claim was that provision costs fall on larger members of a group. The idea is that such members consume large shares of the public good, and so face a relatively large private benefit. Once again, this builds upon the assumption that the collective output is rival; for a pure public good, the same logic would predict that those who care most contribute most, and such contributors need not be large in a conventional sense.

Olson's third claim concerned the possible response to the problem of collective action. Such a response requires, according to this claim, selective incentives that are ‘functionally equivalent to the taxes that enable governments to provide public goods … [they] either punish or reward individuals depending on whether or not they have borne a share of the costs of collective action, and thus give the individual an incentive to contribute …’ (Olson, 1987). The classic example of selective incentives is the ‘closed shop’ of labour unions; to enjoy the benefits of collective union bargaining power each worker must be a member, and hence pay the costs of any strike action. Interestingly, when the selective incentive is based on preventing a group member from enjoying the collective output then the implicit assumption is that the public good is at least partially excludable.

In summary, Olson (1965; 1987) forcefully clarified the inescapable logic of collective action: any theory of group behaviour must rely upon the incentives faced by individuals, and not simply assume that groups pursue their common interests. His theory of groups remains relevant for many contemporary problems. However, it steps outside the world of pure public goods by assuming the interdependent consumption of an impure public good, and does not allow for interdependence of production. Put more succinctly, Olson's groups consist of public good consumers rather than public good providers.

Attention now refocuses on collective-action problems in which economic players non-cooperatively choose whether to participate in the private production of a pure public good. Crucially, there can be interdependence of production: the incentive to participate in a collective action depends on the expected participation of others. Decisions become genuinely strategic, and this changes the nature of the collective-action problem.

A little notation proves helpful. Amongst n players, write x_i for the action of player i, and collect the actions of everyone together into a vector x. Payoffs satisfying

$$u_i(x) = G(x) - c_i(x_i) \qquad (\dagger)$$

comprise the value $G(x)$ of public good and the private cost $c_i(x_i)$ that player i incurs when contributing to it; the externality imposed on others is captured by $(n-1)G(x)$. The nature of the strategic interaction amongst players depends upon the form taken by $G(x)$. A simple specification is when x_i is a positive real number and $G(x) \equiv \sum_{i=1}^{n} x_i$. A player's decision is strategically independent of others' actions: he simply equates the private marginal benefit of the public good to its private marginal cost via the first-order condition $1=c'(x_i)$, yielding the usual under-provision problem (Cornes and Sandler, 1996).

A second natural specification to consider is where $G(x) \equiv F(\sum_{i=1}^{n} x_i)$ for some nicely behaved concave production function $F(\cdot)$. This falls within the class of Cournot contributions games (Chamberlin, 1974; McGuire, 1974; Young, 1982; Cornes and Sandler, 1985; Bergstrom, Blume and Varian, 1986; Bernheim, 1986). Here, strategic interaction is non-trivial since the marginal benefit of increased public good provision depends on the total contributions of all players. Nevertheless, a unique Nash equilibrium involves under-provision. The associated literature concerned itself with the comparative-static properties of such models, including the response of public good output and the burden of provision to the redistribution of income (Warr, 1983; Kemp, 1984).

These first two examples of equation (†) simply flesh out the implicit model of Olson (1965). The nature of the collective action problem changes significantly when $G(x)$ takes on more interesting and yet plausible shapes. For instance, $G(\cdot)$ might take a weakest link ($G(x)=\min\{x_i\}$) or best shot ($G(x)=\min\{x_i\}$) form (Hirshleifer, 1983; 1985); these are special cases of symmetric but non-additive specifications (Cornes, 1993).

Here, however, attention turns to situations in which the success of a collective action (that is, the successful provision of a public good) turns upon either the participation of a critical mass of players, or contributions that exceed a particular threshold. Returning once more to the economics of climate change, a plausible scenario is one in which the ice caps melt unless carbon emissions are pushed down below a critical level. Whereas in a Cournot contributions game the incentive to contribute decreases with the participation of others, here it may increase: a nation may find it worthwhile to chase environmental targets if and only if it expects others to play their part in international agreements.

A central feature of threshold-based scenarios is that an individual's decision depends on aggregate participation. This is easiest to explore in a binary-action game where $x_i \in \{0,1\}$ for each player i; hence $x_i=1$ can be interpreted as individual participation in a collective action. In many such situations, the incentive to participate depends on the number of others who do so. Hence, writing $\Delta u_i(x)$ for this incentive,

$$\Delta u_i(x) \equiv P(m) \text{ where } m = \sum_{j \neq i} x_j. \quad (‡)$$

When $P(m)<0$ for all m, no players participate; this is an n-player Prisoner's Dilemma. If $P(m)$ decreases with m, then the unique equilibrium entails the participation of m^* players, where $P(m^*-1) > 0 > P(m^*)$; for the Cournot games considered above the participation m^* might be socially suboptimal. If $P(m)$ increases with m, so that there is a threshold m^* satisfying $P(m^*-1) < 0 < P(m^*)$, then there are two pure-strategy Nash equilibria, one in which everyone participates, and one in which the collective action fails. This means that the problem of collective action becomes one of coordination.

Games satisfying equation (‡) drew the insightful attention of Schelling (1973; 1978). He opened his analysis by describing the use of protective helmets in ice hockey: players were willing to wear helmets only if others did so too. Other sociological examples are easy to find: members of a crowd will join a protest only if others do so (Berk, 1974; Granovetter, 1978) and successful consumer boycotts require a critical mass (Innes, 2006).

Political situations can also fit equation (‡). Consider a plurality rule election in which a group wishes to prevent the success of a disliked incumbent candidate. They can do so if and only if a critical number m^* abandon their first-preference candidate and vote for their second choice. Setting $P(m^*-1) > 0$ and $P(m)<0$ otherwise yields a strategic-voting model (Palfrey, 1989; Myerson and Weber, 1993; Cox, 1994; 1997; Myatt, 2007).

In sociology, collective-action games with threshold properties fall under the umbrella of the theory of critical mass (Oliver, Marwell and Teixeira, 1985; Oliver and Marwell, 1988; Marwell, Oliver and Prahl, 1988; Marwell and Oliver, 1993). Alas, these sociologists had no theoretical machinery for selecting between multiple equilibria. In economics, multiple equilibria arise in threshold-driven step-level public goods games (Palfrey and Rosenthal, 1984). Once again, the problem of coordination boils down to a need to choose amongst multiple equilibria. Fortunately, recent contributions to economics allow some progress to be made on the equilibrium-selection problem.

To explore further, it is instructive to consider a simple world: two individuals (A and B) either participate (Y) or not (N) in a collective action. Participation involves a private cost (either c_A or c_B), but may provide a public good to be enjoyed by both players. A natural representation is via a simple 2×2 strategic form game (Figure 1).

Figure 1
Public-good provision games

	Y	N
Y	$2v - c_B$ $2v - c_A$	v $v - c_A$
N	$v - c_B$ v	0 0

Provision game

	Y	N
Y	$v - c_B$ $v - c_A$	v $v - c_A$
N	$v - c_B$ v	0 0

Volunteer's dilemma

	Y	N
Y	$v - c_B$ $v - c_A$	0 $-c_A$
N	$-c_B$ 0	0 0

Teamwork dilemma

In the ‘provision game’ a participant produces a public good worth v to everyone. A player's marginal product is strategically independent: the incentive for player A to participate is always $v - c_A$, and hence he does so if and only if $v>c_A$. However, this generates a spillover of v for player B, and hence the social gain is $2v - c_A$. The parameter configuration $2v>c_A>v$ yields the classic under-provision of a public good.

But what if there is strategic interdependence? Suppose that only one player need provide, so that a second participant generates a cost but no additional benefit. This ‘volunteer's dilemma’ (Diekmann, 1985) is a textbook game of ‘chicken’ (Lipnowski and Maital, 1983). If $2v>c_A>v$ and $2v>c_B>v$ then neither player is willing to participate even though it is socially optimal for someone to do so. However, if $v>c_A$ then player A participates so long as player B does not. If $v>c_B$, then there are two pure-strategy Nash equilibria in which a single player provides the public good. But who provides?

One possibility is to use risk dominance (Harsanyi and Selten, 1988) as a selection criterion. The risk-dominant equilibrium is that which maximizes the product of players' incentives to remain at the equilibrium. So, in the volunteer's dilemma, the equilibrium in which A provides is risk-dominant if $(v - c_A)c_B > (v - c_B)c_A$, which holds if and only if $c_A<c_B$: the most efficient provider volunteers. Following Olson (1965), the strong (low-cost providers) bear the cost of provision to the benefit of the weak.

A coordination problem also arises in the ‘teamwork dilemma’ (Figure 1) where both players are needed for the collective action to succeed. This is an assurance or ‘stag hunt’ game: as long as $v>c_A$ and $v>c_B$ there is a pure-strategy equilibrium in which both players participate, and a second with no participation in which the collective action fails. The former equilibrium is risk dominant if and only if $(v - c_A)(v - c_B) > c_A c_B$, which boils down to $v>c_A+c_B$; this requires a single private (not social) benefit from the public good to exceed the total private cost of provision. If $2v>c_A+c_B>v$, then the collective action fails even though it would be socially optimal for it to succeed. Once again, this is a return to Olson (1965): success of the collective action relies on private incentives.

All well and good, but can the criterion of risk dominance be justified? In the recent literature two contrasting approaches lead to the same answer.

The theory of global games (Carlsson and Van Damme, 1993; Morris and Shin, 2003) supposes that players do not share common knowledge of the payoffs of games. Instead, players must rely upon privately observed signals of the game being played. For instance, players may be unsure of the true value v of the public good, and see an estimate of it. Crucially, this estimate allows them to infer not only this value but also the probable signals received by others, and hence their opponents' likely behaviour. When signals become very precise then the play of a simple 2×2 game almost always coincides with the risk-dominant Nash equilibrium (Carlsson and Van Damme, 1993).

Others have selected equilibria by studying the evolving play. Players (or populations from which players are drawn) may adjust their play over time in the direction of myopic best-reply, but occasionally ‘mutate’ to a different strategy (Kandori, Mailath and Rob, 1993; Young, 1993; 1998). As the probability of mutations vanishes, play in the long run focuses on a single stochastically stable equilibrium (Foster and Young, 1990). In a symmetric teamwork dilemma, it picks out the risk dominant equilbrium.

Can modern literature say anything about the general case of equation (†)? Players act as though they are attempting to maximize jointly the single real-valued function

$$\rho(x) \equiv G(x) - \sum_{i=1}^{n} c_i(x_i). \quad (\bigstar)$$

This is a potential function, and yields a potential game (Monderer and Shapley, 1996). This function has a natural interpretation: the private benefit that a single individual derives from a public good, minus the total private costs involved in its provision.

Clean results emerge when play of a potential game evolves via a payoff-responsive stochastic strategy-revision process (Blume, 1993; 1995; 1997; Brock and Durlauf, 2001; Blume and Durlauf, 2001; 2003). Over time, players occasionally revise their strategies. When a player does so, his decision is not a myopic best reply to the current play of others, but rather a quantal response (McKelvey and Palfrey, 1995): the log odds of choosing one action rather than another is determined by the difference in their payoffs, and so he is more likely to choose better performing strategies. An inspection of equation (★) reveals that the difference in a player's payoffs is equal to the difference in potential; the potential function captures the essential strategic interaction of the game.

Allowing play to evolve, the strategy-revision process is drawn towards the states-of-play with the highest potential. In the long run, when quantal responses approximate best replies, the process spends almost all time in the state that maximizes $\rho(x)$: evolution leads players to maximize the difference between a single private benefit and total private costs rather than social welfare which would incorporate the full social benefit of $nG(x)$.

This approach can be applied to the teamwork dilemma: the potential of the state-of-play in which neither player participates is zero, and the potential of the equilibrium in which the collective action succeeds is $v - (c_A+c_B)$. The latter equilibrium has positive potential if and only if $v>c_A+c_B$: only if a private individual would be willing to step forward and pay the full cost of provision himself will the collective action succeed. So, whereas it may at first appear that the success of a collective action (the coordinated play of $\{Y,Y\}$ in the teamwork dilemma) can follow from the interdependence of team members, evolving play results in failure (the play of $\{N,N\}$ in the teamwork dilemma) unless a private individual would be willing to fund the collective action himself.

On reflection, this should be unsurprising. Each step of evolving play (or each step of reasoning in a global-game argument) is driven by reference to private incentives. So what lesson should be taken away? Even when a group's problem is one of coordination, its members cannot escape Olson's (1965; 1987) fundamental logic of collective action.

See Also

- collective action
- externalities
- public goods

Bibliography

Allen, R.C. 1983. Collective invention. *Journal of Economic Behavior and Organization* 4, 1–24.

Bergstrom, T.C., Blume, L. and Varian, H.R. 1986. On the private provision of public goods. *Journal of Public Economics* 29, 25–49.

Berk, R. 1974. A gaming approach to crowd behavior. *American Sociological Review* 39, 355–73.

Bernheim, B.D. 1986. On the voluntary and involuntary provision of public goods. *American Economic Review* 76, 789–93.

Blume, L.E. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5, 387–424.

Blume, L.E. 1995. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior* 11, 111–45.

Blume, L.E. 1997. Population games. In *The Economy as an Evolving Complex System II*, ed. W.B. Arthur, S.N. Durlauf and D.A. Lane. Boulder, CO: Westview.

Blume, L.E. and Durlauf, S.N. 2001. The interactions-based approach to socioeconomic behaviour. In *Social Dynamics*, ed. S.N. Durlauf and H.P. Young. Cambridge, MA: MIT Press.

Blume, L.E. and Durlauf, S.N. 2003. Equilibrium concepts for social interaction models. *International Game Theory Review* 5, 193–209.

Brock, W.A. and Durlauf, S.N. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68, 235–60.

Carlsson, H. and Van Damme, E. 1993. Global games and equilibrium selection. *Econometrica* 61, 989–1018.

Chamberlin, J. 1974. Provision of collective goods as a function of group size. *American Political Science Review* 68, 707–16.

Cornes, R. 1993. Dyke maintenance and other stories: some neglected types of public goods. *Quarterly Journal of Economics* 108, 259–71.

Cornes, R. and Sandler, T. 1985. The simple analytics of pure public good provision. *Economica* 52, 103–16.

Cornes, R. and Sandler, T. 1996. *The Theory of Externalities, Public Goods and Club Goods*, 2nd edn. London: Cambridge University Press.

Cox, G.W. 1994. Strategic voting equilibria under the single nontransferable vote. *American Political Science Review* 88, 608–21.

Cox, G.W. 1997. *Making Votes Count*. Cambridge: Cambridge University Press.

Diekmann, A. 1985. Volunteer's dilemma. *Journal of Conflict Resolution* 29, 605–10.

Foster, D. and Young, H.P. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38, 219–32.

Granovetter, M. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83, 1420–43.

Harsanyi, J.C. and Selten, R. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press Classics.

Hirshleifer, J. 1983. From weakest-link to best-shot: the voluntary provision of public goods. *Public Choice* 41, 371–86.

Hirshleifer, J. 1985. From weakest-link to best-shot: correction. *Public Choice* 46, 21–3.

Innes, R. 2006. A theory of consumer boycotts under symmetric information and imperfect competition. *Economic Journal* 116, 355–81.

Johnson, J.P. 2002. Open source software: private provision of a public good. *Journal of Economics and Management Strategy* 11, 637–62.

Kandori, M., Mailath, G.J. and Rob, R. 1993. Learning, mutation and long-run equilibria in games. *Econometrica* 61, 29–56.

Kemp, M. 1984. A note on the theory of international transfers. *Economics Letters* 14, 259–62.

Lerner, J. and Tirole, J. 2001. The open source movement: key research questions. *European Economic Review* 45, 819–26.

Lerner, J. and Tirole, J. 2002. Some simple economics of open source. *Journal of Industrial Economics* 50, 197–234.

Lipnowski, I. and Maital, S. 1983. Voluntary provision of a pure public good as the game of chicken. *Journal of Public Economics* 20, 381–86.

Marwell, G. and Oliver, P.E. 1993. *The Critical Mass in Collective Action: A Micro-Social Theory*. Cambridge and New York: Cambridge University Press.

Marwell, G., Oliver, P.E. and Prahl, R. 1988. Social networks and collective action: a theory of the critical mass. III. *American Journal of Sociology* 94, 502–34.

McGuire, M. 1974. Group size, group homogeneity and aggregate provision of a pure public good under cournot behavior. *Public Choice* 18, 107–26.

McKelvey, R.D. and Palfrey, T.R. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10, 6–38.

Monderer, D. and Shapley, L.S. 1996. Potential games. *Games and Economic Behavior* 14, 124–43.

Morris, S. and Shin, H.S. 2003. Global games: theory and application. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, vol. 1, ed. M. Dewatripont, L.P. Hansen and S.J. Turnovsky. London: Cambridge University Press.

Myatt, D.P. 2007. On the theory of strategic voting. *Review of Economic Studies* 74, 255–81.

Myerson, R. and Weber, R. 1993. A theory of voting equilibria. *American Political Science Review* 87, 102–14.

Oliver, P.E. and Marwell, G. 1988. The paradox of group size in collective action: a theory of critical mass. II. *American Sociological Review* 53, 1–8.

Oliver, P.E., Marwell, G. and Teixeira, R. 1985. A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology* 91, 522–56.

Olson, M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Olson, M. 1987. Collective action. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Palfrey, T.R. 1989. A mathematical proof of Duverger's law. In *Models of Strategic Choice in Politics*, ed. R. Ordeshook. Ann Arbor: University of Michigan Press.

Palfrey, T.R. and Rosenthal, H. 1984. Participation and the provision of discrete public goods: a strategic analysis. *Journal of Public Economics* 24, 171–93.

Raymond, E.S. 1998. The cathedral and the bazaar. *First Monday* 3, 1–20. Online. Available at http://www.firstmonday.dk/issues/issue3_3/raymond/, accessed 19 April 2007.

Sandler, T. 2004. *Global Collective Action*. Cambridge: Cambridge University Press.

Schelling, T.C. 1973. Hockey helmets, concealed weapons, and daylight saving: a study of binary choices with externalities. *Journal of Conflict Resolution* 17, 381–428.

Schelling, T.C. 1978. *Micromotives and Macrobehavior*. New York: Norton.

Stern, N. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

Warr, P. 1983. The private provision of a public good is independent of the distribution of income. *Economics Letters* 13, 207–11.

Young, D.J. 1982. Voluntary purchase of public goods. *Public Choice* 38, 73–85.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

Young, H.P. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.

How to cite this article

Myatt, David P. "collective action (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000204> doi:10.1057/9780230226203.0261

The New Palgrave Dictionary of Economics Online

collective action

Mancur Olson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The logic of collective action undermines the assumption that common interests are always promoted by their beneficiaries. Where the number of beneficiaries is large, the benefits of collective action are a public good: beneficiaries will gain whether or not they participate in promoting them, while their individual efforts cannot secure them. Small groups can use selective incentives to ensure that their members contribute to promoting their common interests. This typically results in the paradoxical ‘exploitation of the great by the small’. The logic of collective action helps explain many notable examples of economic growth and stagnation since the Middle Ages.

Keywords

anarchy; bargaining; cartels; class conflict; collective action; collective bargaining; collusion; common interests; countervailing power; encompassing organizations; excess burden; exploitation; Galbraith, K.; group theory; industrial revolution; invisible hand; latent groups; lobbying; mercantilism; non-excludability; Olson, M.; patronage dividends; public choice; public goods; revelation of preferences; Samuelson, P.; selective incentives; Smith, A.; strategic behaviour; technical progress; trade unions; Wicksell, J.

Article

For a long while, economists, like specialists in other fields, often took it for granted that groups of individuals with common interests tended to act to further those common interests, much as individuals might be expected to further their own interests. If a group of rational and self-interested individuals realized that they would gain from political action of a particular kind, they could be expected to engage in such action; if a group of workers would gain from collective bargaining, they could be expected to organize a trade union; if a group of firms in an industry would profit by colluding to achieve a monopoly price, they would tend to do so; if the middle class or any other class in a country had the power to dominate, that class would strive to control the government and run the country in its own interest. The idea that there was some tendency for groups to act in their common interests was often merely taken for granted, but in some cases it played a central conceptual role, as in some early American theories of labour unions, in the ‘group theory’ of the ‘pluralists’ in political science, in J.K. Galbraith's concept of ‘countervailing power’, and in the Marxian theory of class conflict.

More recently, the explicit analysis of the logic of individual optimization in groups with common interests has led to a dramatically different view of collective action. If the individuals in some group really do share a common interest, the furtherance of that common interest will automatically benefit each individual in the group, whether or not he has borne any of the costs of collective action to further the common interest. Thus the existence of a common interest need not provide any incentive for individual action in the group interest. If the farmers who grow a given crop have a common interest in a tariff that limits the imports and raises the price of that commodity, it does not follow that it is rational for an individual farmer to pay dues to a farm organization working for such a tariff, for the farmer would get the benefit of such a tariff whether he had paid dues to the farm organization or not, and his dues alone would be most unlikely to determine whether or not the tariff passed. The higher price or wage that results from collective action to restrict the supply in a market is similarly available to any firm or worker that remains in that market, whether or not that firm or worker participated in the output restriction or other sacrifices that obtained the higher price or wage. Similarly, any gains to the capitalist class or to the working class from a government that runs a country in the interests of that class, will accrue to an individual in the class in question whether or not that individual has borne the costs of any collective action. This, in combination with the extreme improbability that a given individual's actions will determine whether his group or class wins or loses, entails that a typical individual, if rational and self-interested, would not engage in collective action

in the interest of any large group or class.

Analytically speaking, the benefits of collective action in the interest of a group with a common interest are a public or collective good to that group; they are like the public goods of law and order, defence, and pollution abatement in that voluntary and spontaneous market mechanisms will not provide them. The fundamental reality that unifies the theory of public goods with the more general logic of collective action is that ordinary market or voluntary action fails to obtain the objective in question. It fails because the benefits of collective or public goods, whether provided by governments or non-governmental associations, are not subject to exclusion; if they are received by one individual in some group, they automatically also go to the others in that group (Olson, 1965).

Since many groups with common interests obviously do not have the power to tax or any comparable resource, the foregoing logic leads to the prediction that many groups that would gain from collective action will not in fact be organized to act in their common interests. This prediction is widely supported. Consumers have a common interest in opposing the legislation that gives various producer groups supra-competitive prices, and they would sometimes also have a common interest in buyers' coalitions that would countervail producer monopolies, but there is no major country where most consumers are members of any organization that works predominantly in the interest of consumers. The unemployed similarly share a common interest, but they are nowhere organized for collective action. Neither do most taxpayers, nor most of the poor, belong to organizations that act in their common interest (Austen-Smith, 1981; Brock and Magee, 1978; Chubb, 1983; Hardin, 1982; Moe, 1980; Olson, 1965).

Though some groups can never act collectively in their common interest, certain other groups can, if they have ingenious leadership, overcome the difficulties of collective action, though this usually takes quite some time. There are two conditions either of which is ultimately sufficient to make collective action possible. One condition is that the number of individuals or firms that would need to act collectively to further the common interest is sufficiently small; the other is that the groups should have access to ‘selective incentives’.

The way that small numbers can make collective action possible at times is most easily evident on the assumption that the individuals in a group with a common interest are identical. Suppose there are only two large firms in an industry and that each of these firms will gain equally from any government subsidy or tax loophole for the industry, or from any supra-competitive price for its output. Clearly each firm will tend to get the benefit of any lobbying it does on behalf of the industry, and this can provide an incentive for some unilateral action on behalf of the industry. Since each firm's action will have an obvious impact on the profits of the other, the firms will have an incentive to interact strategically with and bargain with one another. There would be an incentive to continue this strategic interaction or bargaining until a joint maximization or ‘group optimal’ outcome had been achieved. This same logic obviously also applies to collective action in the form of collusion to obtain a supra-competitive price, and thus we obtain the well-known incentive for oligopolistic collusion in concentrated industries whenever there are significant obstacles to or costs of entry. As the number in a group increases, however, the incentive to act collectively diminishes; if there are ten identical members of a group with a common interest, each gets a tenth of the benefit of unilateral action in the common interest of the group, and if there are a million, each gets one millionth. In this last case, even if there were some incentive to act in the common interest, that incentive would cease long before a group-optimal amount of collective action had taken place. Strategic interaction or voluntary bargaining will not occur since no two individuals have an incentive to interact strategically or to bargain with one another. This is because the failure of one individual to support collective action will not them have any perceptible effect on the incentive any other individual faces so there is no incentive for strategic interaction or rational bargaining. Thus we obtain the result that, in time, sufficiently small groups can act collectively, but that this incentive for collective action decreases monotonically as the group gets larger and disappears entirely in sufficiently large or ‘latent’ groups.

When the parties that would profit from collective action have very different demand curves, the party with the highest absolute demand for collective action will have an incentive to engage in some amount of collective action when no other member of the group has such an interest. This leads to a paradoxical ‘exploitation of the great by the small’. This is true to a greater degree and is evident much more simply if income effects are ignored, as in the demand curves for a collective good depicted in the figure below. When the party with the highest demand curve for the collective good, D_h, has obtained the amount of the collective good, Q_1, that is in its interest unilaterally to provide, any and all parties with a lower demand curve, such as D_s, will automatically receive this same amount, and thus have no incentive to provide any amount at all! (Olson, 1965). When income effects and certain ‘private good’ aspects of some collective goods are taken into account the results are less extreme, but a distribution of burdens disproportionality unfavourable to the parties with the absolutely larger demands tends to remain. This disproportion has been evident, for example, in various military alliances and international organizations, in cartels, and in metropolitan areas in which metropolis-wide collective goods are provided by independent municipalities of greatly different size (Olson and Zeckhauser, 1966; Sandler, 1980)

Figure 1.

Figure 1

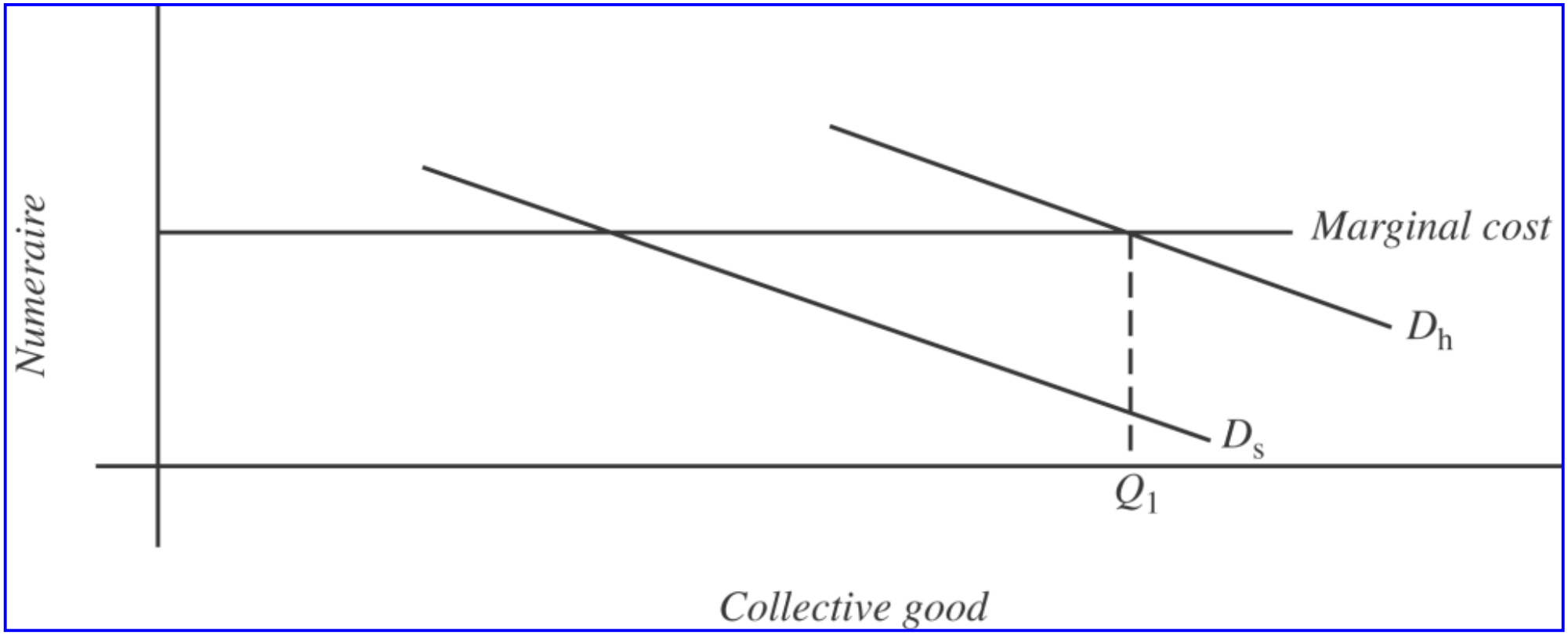

The other condition, besides small numbers, that can make collective action possible, is ‘selective incentives’. Those large groups that have been organized for collective action for any substantial period of time are regularly found to have worked out special devices, or selective incentives, that are functionally equivalent to the taxes that enable governments to provide public goods (Olson, 1965; Hardin, 1982). These selective incentives either punish or reward individuals depending on whether or not they have borne a share of the costs of collective action, and thus give the individual an incentive to contribute to collective action that no good that is or would be available to all could provide. The most obvious devices of this kind are the ‘closed shop’ and picket line arrangements of labour unions, which often make union membership a condition of employment and control the supply of labour during strikes (see, for example, McDonald, 1969; Gamson, 1975). Upon investigation it becomes clear that labour unions are not in this respect fundamentally different from other large organizations for collective action, which regularly have selective incentives that, though usually less conspicuous than the closed shop or the picket line, serve the same function.

Farm organizations in several countries, and quite notably in the United States, obtain most of their membership by deducting the dues in farm organizations from the ‘patronage dividends’ or rebates of farm cooperatives and insurance companies that are associated with the farm organizations. The professional associations representing such groups as physicians and lawyers characteristically have either relatively discreet forms of compulsion (such as the ‘closed bar’) or subtle individual rewards to association members, such as access to professional publications, certification, referrals, and insurance. In small groups, and sometimes in large ‘federal’ groups that are composed of many small groups, social pressure and social rewards are also important sources of selective incentives.

The selective incentives that are needed if large groups are to organize for collective action are less often available to potential entrants or those at the lower levels of the social order than to established and well-placed groups. The unemployed, for example, obviously do not have the option of making membership of an organization working in their interest a condition of employment, nor do they naturally congregate as the employed do at workplaces where picket lines may be established. Those who would profit from entering a cartelized industry or profession are similarly almost always without selective incentives. Experience in a variety of countries also confirms that those with higher levels of education and skill have better access to selective incentives than lower income workers; highly trained professionals such as physicians and attorneys usually come to be well organized before labour unions emerge, and the unions of skilled workers normally emerge before unions representing less skilled workers. The correlation between income and established status and access to selective incentives works in the same direction as the lesser difficulty of collective action of small groups of large firms in relatively concentrated industries explained above. Together these two factors generate a tendency for collective action to have, in the aggregate though not in all cases, a strong anti-egalitarian and pro-establishment impact (Olson, 1984).

The study of collective action goes back to the beginnings of economics, but then came to be strangely neglected during most of the rest of the history of the subject. Though this is not generally realized, the study of collective action, admittedly only in an inductive and intuitive way, was a crucial part of Adam Smith's analysis of the inefficiencies and inequities in the economies he observed (Smith, 1776). Smith even noted that the main beneficiaries of collective action in his time were by no means the poor or those of average means. He also emphasized the tendency for urban interests to profit from collective action at the expense of rural people, because the geographical dispersion of agricultural interests areas made it more difficult for them to combine to exert political influence or to fix prices; this emphasis presumably owed something to the poor transportation and communication systems in his day, which presumably obstructed the organization of rural interests more in his time than it does in developed countries now.

The label that Adam Smith gave to the set of public policies, monopolistic combinations, and ideas that he attacked was, after all, ‘mercantilism’, because the single most important source of the evils was the collective action of merchants, or merchants and ‘masters’, especially those organized into guilds or ‘corporations’. In his discussions of the ‘Inequalities Occasioned by the Policy of Europe’ and of ‘The Rent of Land’ (Bk. I, ch. 10, pt. ii and ch. 11), Smith emphasized that ‘whenever the legislature attempts to regulate the differences between masters and their workmen, its counsellors are always the Masters'. Similarly,

> it is everywhere much easier for a rich merchant to obtain the privilege of trading in a town corporate, than for a poor artificer to obtain that of working in it…. Though the interest of the labourer is strictly connected with that of the society … his voice is little heard and less regarded.

The rural interests are similarly at a disadvantage, according to Smith, especially as compared with those in ‘trade and manufacturers’:

> The inhabitants of a town, being collected into one place, can easily combine together. The most insignificant trades carried on in towns have accordingly, in some place or another, been incorporated … voluntary associations and agreements prevent that free competition which they cannot prohibit…. The trades which employ but a small number of hands run most easily into such combinations…. People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices.

By contrast, ‘the inhabitants of the country, dispersed in distant places, cannot easily combine together’.

These passages, though not in the order they appear in Smith, nonetheless correctly convey his alertness to collective action. Though the handicap that rural interests face in organizing for collective action is far less in developed countries today than it was in Smith's time, even this part of his argument still generally holds true in the developing countries, where transportation and communication in the rural areas are poor, peasants are generally unrepresented, and agricultural commodities normally underpriced (Anderson and Hayami, 1986; Schultz, 1978; Olson, 1985).

Adam Smith's insights into collective action and its consequences were ignored until recent times. Presumably one reason is that most economists in the 19th and early 20th centuries were mainly interested in the logic of the case for competitive markets. The logic of collective action, by contrast, is really a general statement of the logic of market failure; it embodies the central insight of the theories of public goods and externalities, that markets and voluntary market-type arrangements do not generally work in those cases where the beneficiaries of any collective good or benefit cannot be excluded because they have not paid any purchase price or dues (Baumol, 1952). It was not until Knut Wicksell's New Principle of Just Taxation’ was published in German in 1896 (Musgrave and Peacock, 1967) that any economist revealed a clear understanding of the nature of public goods, and only with the publication of Samuelson's articles in the 1950s (Samuelson, 1955) that this idea came to be generally understood in the English-speaking world.

A second obstacle to the development of the logic of collective action was that collective action by governments was normally taken for granted. Notwithstanding the difficulties of collective action, anarchy is relatively rare because a government that provides some sort of law and order quickly takes over. This in turn is due to conquerors and the gains they obtain in increased tax revenues from establishing some system of law and order and property rights. In the absence of the provision of these most elemental collective goods, there is not much for a conqueror to take, so the historic first movement of the invisible hand is evident in the incentive conquerors have to establish law and order. Those who lead the governments that succeed conquerors obviously must maintain a system of law and order if they are to continue collecting significant tax revenues. Since governments providing basic collective goods have been ubiquitous, the classic writers on public goods like Wicksell and Samuelson did not even ask how collective goods emerged in the first place. They focused instead on how to determine what was an appropriate sharing of the tax burdens and on the difficulty of determining what level of provision of public goods was Pareto-optimal. This in turn naturally led to Wicksell's recommendation that only those public expenditures that could, with an approximate allocation of the tax burdens, command approximate unanimity, should normally be permitted, and to Samuelson's and Musgrave's (1959) concern for the non-revelation of preferences for public goods. The difficulties of collective action and public good provision on a voluntary basis therefore naturally did not gain any theoretical attention.

When, as in the new political economy or public choice, the focus is also on the efforts of extra-governmental groups to obtain the gains from lobbying, cartelization, and collusion, and on private action to obtain collective benefits of other kinds, a more general conception becomes natural (Barry and Hardin, 1982; Olson, 1965; Taylor, 1976). It then becomes clear that the likelihood of voluntary collective action depends dramatically on the size of the group that would gain from collective action. When a group is sufficiently small and

there is time for the needed bargaining, the desired collective goods will normally be obtained through voluntary cooperation (Frohlich, Oppenheimer, and Young, 1971). If there are substantial differences in the demands for the collective good at issue, there will be the aforementioned paradoxical ‘exploitation of the great by the small’. When the number of beneficiaries of collective action is very large, voluntary and straightforward collective action is out of the question, and taxes or other selective incentives are indispensable. Selective incentives are available only to a subset of those extra-governmental groups that would gain from collective action. Even those extra-governmental groups that do have the potential of organizing through selective incentives will usually have great difficulty in working out these (often subtle) devices, and will normally succeed in overcoming the great difficulties of collective action only when they have relatively ingenious leadership and favourable circumstances.

If follows that it is only in long-stable societies that many extra-governmental organizations for collective action will exist. In societies where totalitarian repression, revolutionary upheavals, or unconditional defeat have lately destroyed organizations for collective action, few groups will have been able in the time available to have overcome the formidable difficulties of collective action. It has been shown elsewhere (Mueller, 1983; Olson, 1982), that (unless they are very ‘encompassing’) organizations for collective action have extraordinarily anti-social incentives; they engage in distributional struggles, even when the excess burden of such struggles is very great, rather than in production. They also will tend to make decisions slowly and thereby retard technological advance and adaptations to macroeconomic and monetary shocks. It follows that societies that have been through catastrophes that have destroyed organizations for collective action, such as Germany, Japan, and Italy, can be expected to enjoy ‘economic miracles’. An understanding of collective action also makes it possible to understand how Great Britain, the country that with industrial revolution discovered modern economic growth and had for nearly a century the world's fastest rate of economic growth, could by now have fallen victim to the ‘British disease’. The logic of collective action, in combination with other theories, also makes it possible to understand many of the other most notable examples of economic growth and stagnation since the Middle Ages, and also certain features of macroeconomic experience that contradict Keynesian, monetarist, and new classical macroeconomic theories (Balassa and Giersch, 1986).

See Also

- bargaining
- collective action (new perspectives)
- public choice
- social choice

Bibliography

Anderson, K. and Hayami, Y. 1986. *The Political Economy of Protection*. Sydney: George Allen & Unwin.

Austen-Smith, D. 1981. Voluntary pressure groups. *Economica* 48, 143–53.

Balassa, B. and Giersch, H., eds. 1986. *Economic Incentives*. Proceedings of the International Economic Association, London: Macmillan.

Barry, B. and Hardin, R., eds. 1982. *Rational Man and Irrational Society*. Beverly Hills: Sage.

Baumol, W.J. 1952. *Welfare Economics and the Theory of the State*. Cambridge, MA: Harvard University Press.

Brock, W. and Magee, S. 1978. The economics of special interest groups: the case of the tariff. *American Economic Review* 68, 246–50.

Chubb, J. 1983. *Interest Groups and Bureaucracy*. Stanford: Stanford University Press.

Frohlich, N., Oppenheimer, J. and Young, O. 1971. *Political Leadership and Collective Boards*. Princeton: Princeton University Press.

Gamson, W.A. 1975. *The Strategy of Social Protest*. Homewood, IL: Dorsey.

Hardin, R. 1982. *Collective Action*. Baltimore: Johns Hopkins University Press for Resources for the Future.

McDonald, D.J. 1969. *Union Man*. New York: Dutton.

Moe, T.M. 1980. *The Organization of Interests*. Chicago: University of Chicago Press.

Mueller, D.C., ed. 1983. *The Political Economy of Growth*. New Haven: Yale University Press.

Musgrave, R.A. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

Musgrave, R.A. and Peacock, A.T., eds. 1967. *Classics in the Theory of Public Finance*. 2nd edn, New York: McGraw-Hill.

Olson, M.L. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Olson, M.L. and Zeckhauser, R. 1966. An economic theory of alliances. *Review of Economics and Statistics* 48, 266–79.

Olson, M.L. 1982. *The Rise and Decline of Nations*. New Haven: Yale University Press.

Olson, M.L. 1984. Ideology and growth. In *The Legacy of Reaganomics*, ed. C.R. Hulten and I.V. Sawhill, Washington: Urban Institute Press, DC.

Olson, M.L. 1985. Space, organization, and agriculture. *American Journal of Agricultural Economics* 67, 928–37.

Samuelson, P.A. 1955. Diagrammatic exposition of a theory of public expenditure. *Review of Economics and Statistics* 37, 350–56.

Sandler, T., ed. 1980. *The Theory and Structure of International Political Economy*. Boulder, CO: Westview.

Schultz, T.W. 1978. *Distortion of Agricultural Incentives*. Bloomington: Indiana University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: J.M. Dent, 1910.

Taylor, M. 1976. *Anarchy and Cooperation*. London: John Wiley.

Wicksell, K. 1896. A new principle of just taxation. Trans. from the German by J.A. Buchanan, in Musgrave and Peacock (1967).

How to cite this article

Olson, Mancur. "collective action." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000602>
doi:10.1057/9780230226203.0260

The New Palgrave Dictionary of Economics Online

collective bargaining

William Brown
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

arbitration; bargaining; collective bargaining; collusion; consultation; employment contracts; industrial democracy; industrial relations; negotiation; trade unions; wage determination

Article

Collective bargaining is a term applied to a variety of methods of regulating relationship between employers and their employees. Its distinctive feature is that it clearly acknowledges a role for trade unions. In contrast with, for example, autocratic paternalism or producer cooperatives, the employer who engages in collective bargaining accepts the right of independent representatives of employees, acting as a collectivity, to argue their point of view on matters that affect their interests. Pay and working conditions are the most common subjects of collective bargaining, but it can encompass any aspect of management.
The impact of collective bargaining upon management, and its effectiveness from the point of trade union members, vary enormously between different employment circumstances. They depend ultimately upon the collective strength that can be mobilized by employees within the legislative constraints laid down by the state. Collective bargaining is thus best seen as a political institution. It provides a means of bringing at least temporary reconciliation of divergent interests between employers and employees in circumstances in which each side can, to a greater or lesser extent, inflict damage on the other. It is, however, a political institution that is intimately linked with economic processes. The relative power of the bargaining partners owes much to their respective labour and product markets. At the same time the outcome of their bargaining has a major impact upon both the wages and the productivity of labour.

Theoretical approaches

This view of collective bargaining as primarily a political rather than an economic institution is relatively recent. Beatrice Webb claimed, according to Marsh (1979), to have originated the expression in 1891 in her study *The Co-operative Movement of Great Britain*. She analysed it further with her

husband Sidney Webb in *Industrial Democracy* (1897). Although they did not define it, they saw it as an alternative to individual bargaining, so that the employer, instead of making separate deals with isolated individuals, ‘meets with a collective will and settles, in a single agreement, the principles on which, for the time being, all workmen of a particular group, or class, or grade, will be engaged’. They identified it as one of three methods used by trade unions to meet their objectives, the other two being to establish mutual assurance arrangements for their members and to press governments to enact favourable laws. For all the richness of the Webbs' analysis, collective bargaining remained for them essentially an economic institution, imposed upon the employer by a labour cartel whereby workers secured better terms of employment by controlling competition among themselves. A naive version of this view can be seen to underlie much formal analysis of collective bargaining by present-day labour economists.

For the next half century Marsh reports no substantial development of the concept apart from in Leiserson's *Constitutional Government in American Industries* (1922). Then in 1951 Chamberlain, in his book *Collective Bargaining*, argued that there were, in essence, three distinct theories. ‘They are that collective bargaining is (1) a means of contracting for the sale of labour, (2) a form of industrial government, and (3) a method of management.’ The first, ‘marketing’ theory was much the same as that of the Webbs. The second, ‘governmental’ aspect was concerned with the procedural needs of dispute resolution. The third ‘managerial’ theory referred to the way in which management and unions in practice combined ‘in reaching decisions on matters in which both have vital interests’; unions through collective bargaining become not the usurpers of management functions but ‘actually *de facto* managers’. At much the same time Harbison (1951) was stressing the very constructive social role that collective bargaining played in resolving industrial conflict and in pushing for the enhancement of the ‘dignity, worth and freedom of individuals in their capacity as workers’.

This more complex view of collective bargaining has been refined by Dunlop (1967) and Kochan (1980) in the United States, but probably the most influential discussion has been Flanders' attempt of 1968 to create a comprehensive theoretical analysis. He argued that the economic associations of the term ‘collective bargaining’ are misleading. The collective agreement commits no one to either buy or sell labour, but rather ensures that, when labour is bought or sold, the terms of the transaction will accord with the provisions of the agreement. Above all else, collective bargaining is a rule-making process covering many aspects of the employment relationship besides pay and conditions of work. The second characteristic feature of collective bargaining that Flanders stressed is that of the power relationship between the protagonists whose negotiations (‘the diplomatic use of power’) create the rules. Thus, while there are also technical rules and legal rules regulating work, what distinguishes the legitimacy of those that result from collective bargaining is their authorship. They are jointly determined by the accepted representatives of both employers and employees who consequently share responsibility for both the rules' contents and their observance.

Flanders' analysis has proved fertile in several respects. It has drawn attention to the extent to which collective bargaining is a positive management technique rather than just an impediment to effective management imposed by trade unions. As a result of this shift in emphasis, a major part of academic research into collective bargaining in the 1980s has explored managerial, as opposed to trade union, strategies, and has exposed the extent to which union behaviour is shaped by these management strategies. In addition, what could be seen as the Weberian undercurrent in Flanders' analysis has focused policymakers' attention upon the importance of procedural clarity in conflict resolution, and thereby upon the dangers of ambiguity in the legitimation of agreements. The most obvious example is

provided by the influential central recommendation of the British Royal Commission on Trade Unions and Employers' Associations of 1968. The emphasis it placed upon employer initiated procedural reform, rather than legislative constraints on trade unions, owed much to the evidence that Flanders had submitted. Finally, by conceptualizing wages as part of a broader package of regulations and as embodying strongly normative principles, the theory opened the way to a more fruitful understanding of wage determination than is offered simply by the market models of orthodox theory.

Two crucial features of the employment relationship ensure that the process of collective bargaining is fundamentally unlike that of non-labour commercial bargains. They are its open-endedness and its continuity. The labour contract is open-ended because the recruitment of an employee does not ensure the performance of work; the employee has to be motivated, by whatever means, to perform to the required standard. In all but highly oppressive societies such motivational techniques tend to be varied and complex, differing not least in the extent to which they place emphasis upon levels of pay and upon employee participation. Since social comparisons (and especially very local ones) play an important part in the motivation and demotivation of labour, the bureaucratic standardization of terms of employment, which is generally a characteristic of collective agreements, often fits in well with management's preferred personnel techniques. In this way, properly conducted collective bargaining can provide a socially stable working environment which facilitates the employer's prime aim of eliciting labour productivity. In short, the conduct of the bargain affects the quality of the labour bargained over.

The second distinctive feature of the employment relationship is its continuity. Employer and employees are bound together, for better or worse, for an indeterminate duration. Additions to and departures from the workforce generally occur in a piecemeal way. A host of potentially contentious issues feature in the relationship, only a small minority in contention at any one time, and many affecting only a minority of the workforce. Thus a bargain over a particular issue, such as a pay grievance, cannot be evaluated in isolation, but as one fibre in a thick rope of regulations, with many largely implicit trade-offs with respect to other issues, past, present and future.

Characteristics

The definition of collective bargaining as the joint regulation of the employment relationship by employer and employee representatives is one that covers a broad range of processes. It is helpful to analyse these further. An initial distinction has to be made between negotiation and consultation. In a negotiation the discussions are characterized, first, by the awareness of each side of the possibility of one inflicting costs on the other in the absence of an acceptable outcome. Second, a negotiation has to result in some sort of agreement, however informal, to which the two sides are, at least for the time being, committed. Consultation, by contrast, is unaccompanied by either the threat of sanctions or the need to reach binding agreement. Actions taken by management in the light of consultation result from a reappraisal of the facts of the case; those taken after negotiation reflect a compromise which has taken into account the threat (or experience) of sanctions inflicted by either or both sides. Under most collective bargaining arrangements it is felt advisable by both sides to distinguish as far as is possible between negotiations and consultations, at any rate in formal procedures. It is, for example, now normal in large unionized workplaces in Britain to deal with them in specifically different committees, even though the membership of those committees may be much the same.

In practice the distinction is far from clear-cut. The blend of approaches adopted in a particular

collective bargaining episode depends very much upon the issue in question and the relationship between the parties involved. In their study *A Behavioral Theory of Labor Negotiations* (1965), Walton and McKersie distinguished four classes of negotiation. First, there were ‘distributive’ bargains: zero–sum negotiations typified by annual wage bargains and characterized by very formal proceedings. Second were ‘integrative’ bargains: problem-solving discussions aiming at non-zero-sum gains for both sides and generally much more informal in procedure. Third, was ‘attitudinal structuring’, an almost didactic form of bargaining dialogue in which one side tries to alter the way in which their opponents perceive the problem and its context. Finally, ‘intra-organizational’ bargains were aimed at altering positions and attitudes, not on the other side, but within the negotiator's own side.

An important influence upon the way in which bargaining is conducted is the personal ‘bargaining relationship’ between the two individuals who have to take the lead in representing the two sides. This is a term given to the level of trust and facility of communication that exists between them. However acrimonious the collective dispute over which they are bargaining, the better the bargaining relationship between the individual negotiators, the more efficiently they will be able to assess each other's relative power position and the better the chance of the dispute being settled without recourse to expensive sanctions. In a mature bargaining relationship it is common for the negotiators to protect each other from their own sides by, for example, avoiding the humiliation of a bargaining opponent by helping him to gloss over the magnitude of a defeat and by manipulating public statements from one's own side so as to help in his intra-organizational bargaining with his own.

It is normal to draw a clear distinction between the substantive and procedural aspects of collective bargaining. A substantive agreement sets out the actual pay levels, working conditions, or whatever that have been agreed and will be worked to. A procedural agreement defines the way in which such substantive terms might be altered, added to, or interpreted. An effective procedure for negotiation or grievance settlement will state which agents on each side are entitled to be involved in negotiations, in what sequence different sets of negotiators are entitled to consider the matter, what their precedence is, and possibly also matters such as rights of appeal, time constraints, ratification methods and the form of the substantive outcome.

This distinction is particularly obvious in countries whose labour laws cause collective agreements to be tested in the courts; the substantive agreements tend to be written, detailed, formal, and established for specified duration. There are other countries where employer preference, or legal opportunity, makes it unusual for the bargaining opponents to use legal sanctions against each other. In these circumstances the great bulk of substantive regulation may be unwritten and in the form of verbal agreements, custom, and tacit understandings. Because of this a greater emphasis is placed upon the rectitude of the procedural agreements (which may still be very informal) whereby this amorphous body of substantive rules is interpreted and altered, not through comprehensive periodic negotiations, but by a constant incremental process of piecemeal adjustment. Although the United States might be described as exemplifying the legalistic extreme, and Great Britain the ‘voluntaristic’, most bargaining arrangements have elements of each, with the degree of legalism and formality varying by issue and industry, as well as by country.

Bargaining structure

The structure of bargaining in a country, industry, or enterprise, refers to several different characteristics

of collective bargaining. The two most important are the ‘bargaining units’ and ‘bargaining levels’ employed. A bargaining unit is a group of employees covered by a particular agreement. Within this basic territory of industrial government there is a coherence of terms of employment, procedures, and trade union representation that is not necessarily to be found between different bargaining units. The level of bargaining refers to the role played by the principal negotiators within their organizations; whether, for example, the employer representative responsible is a factory manager, a company director, or an employers' association representative.

These two characteristics are involved in the single most important decision in the shaping of any bargaining structure which is whether the employers confront the unions singly or in alliance. Single-employer bargaining, resulting in agreements at company-level or lower, is the majority practice in the United States and Japan and now in Britain. Multi-employer bargaining, in which associations of employers conclude industrywide agreements, remains the most important form in most of Continental Europe. In practice there is often some employer collusion in industries where single-employer bargaining dominates, and there is usually room for individual employer discretion in industries with strong employers' associations, but the distinction remains one of fundamental economic, political, and managerial significance.

Two other defining characteristics of bargaining structure are its ‘form’ and ‘scope’. The first refers to the extent to which proceedings and agreements are formalized and codified. As already mentioned, this depends in part upon the labour legislation of the country. The second matter, scope, refers to the range of issues covered by collective bargaining. At its narrowest it may include no more than pay and hours, while elsewhere it may take in issues as diverse as training policy, investment decisions and child-care facilities.

The most comprehensive theory seeking to explain industrial and national differences in bargaining structure is to be found in Clegg's *Trade Unionism under Collective Bargaining* (1976). This sees the strategy adopted by employers as the main determinant of bargaining structure, although changes in strategy may be slow to take effect. The legislative framework of a country is also of crucial importance. It defines the limits of rights to strike, the status of the employment contract, any guarantees of security for trade unions, and the legally responsible agents on each side.

Most countries acquired their principal labour legislation at some historic period of crisis – war, defeat, depression, or extreme industrial unrest – and the institutional arrangements that developed from that have become consolidated in subsequent, more peaceful times. This helps to account for the very great variations in collective bargaining practice to be found in different countries; they often owe their origin to a distant panic measure based upon a fashionable idea (such as, for example, compulsory arbitration in Australia or compulsory conciliation in Canada) to which employers and unions have adjusted so firmly that radical reformation is all but impossible. A recurring experience around the world is of legislatures finding extreme difficulty in reforming collective bargaining, other than in times of extreme crisis, because of the essential privacy of the bargaining relationship between employers and union.

Most industrialized countries publicly assert a commitment to collective bargaining as a necessary part of a democratic society, and for most it is the normal means of conducting industrial relations in the public sector. Convention 84 (1947) of the International Labour Organization asserts that ‘all practical measures shall be taken to assure to trade unions which are representative of the workers concerned the right to conclude collective agreements with employers and employers' associations’. In practice the freedom of collective bargaining in both public and private sectors varies substantially between

countries and over time.

No discussion of collective bargaining would be complete without a mention of the debate concerning its relationship with industrial democracy.

One view is that, because collective bargaining is essentially concerned with compromise, trade unions are sucked into collaborating with capitalism and thereby denied the opportunity of uniting the working class in overthrowing existing employers and then instituting true industrial democracy through workers' control. Opposing this is a view that deplores the fact that collective bargaining institutionalizes the opposition of capital and labour: them and us. It considers that the best form of industrial democracy is to be found where workers are brought to perceive an ultimate identity of interest with employers. Between these positions is that most clearly expressed by Clegg in *A New Approach to Industrial Democracy* (1960). This argues that there can never be complete identity of interest between employer and employee, and also that if employee representatives are given managerial responsibilities they will be forced to behave very similarly to the employers they have replaced. Consequently the role of the trade union is best seen as one of constant opposition, acting to modify management actions in the light of members' interests insofar as their organized power permits. Far from undermining the common interests of capital and labour, collective bargaining permits the joint regulation of aspects of employment which would otherwise generate greater disharmony and division.

See Also

- industrial relations

Bibliography

Chamberlain, N.W. 1951. *Collective Bargaining*. New York: McGraw-Hill.

Clegg, H.A. 1960. *A New Approach to Industrial Democracy*. Oxford: Blackwell.

Clegg, H.A. 1976. *Trade Unionism under Collective Bargaining*. Oxford: Blackwell.

Dunlop, J.T. 1967. The social utility of collective bargaining. In *Challenges to Collective Bargaining*, ed. L. Ulman. New York: Prentice-Hall.

Flanders, A. 1968. Collective bargaining: a theoretical analysis. *British Journal of Industrial Relations*. 6 (1), 1–26; reprinted in A. Flanders, *Management and Unions*, London: Faber & Faber, 1975.

Harbison, F.H. 1951. *Goals and Strategies in Collective Bargaining*. New York: Harper.

Kochan, T.A. 1980. *Collective Bargaining and Industrial Relations*. Homewood: Irwin.

Leiserson, W.M. 1922. Constitutional government in American industries. *American Economic Review*

12(Supp.), 56–79.

Marsh, A. 1979. *Concise Encyclopedia of Industrial Relations*. Farnborough: Gower.

Walton, R.E. and McKersie, R.B. 1965. *A Behavioral Theory of Labor Negotiations*. New York: McGraw-Hill.

Webb, S. and B. 1897. *Industrial Democracy*. London: Longmans Green.

How to cite this article

Brown, William. "collective bargaining." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000206> doi:10.1057/9780230226203.0262

The New Palgrave Dictionary of Economics Online

collective choice experiments

Rick K. Wilson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Collective choice experiments examine voting mechanisms that aggregate individual preferences. Two general topics have received the most attention. The first pertains to agents deciding on a single collective outcome or policy. The second topic covers election mechanisms that govern candidates and voters.

Keywords

agenda control; Arrows; theorem; collective choice experiments; electoral mechanisms; median voter; social choice; spatial committee experiments

Article

Duncan Black (1948) and Kenneth Arrow (1963) raised the key question of collective choice: if people have different preferences for policy outcomes are there general mechanisms that can (always) aggregate those preferences in consistent and coherent ways? The answer is ‘no’. Starting from simple premises involving individual transitivity, aggregate Pareto optimality and non-dictatorship there is no collective choice mechanism that yields a socially transitive outcome. Such a finding is startling given the confidence placed in democratic institutions that rely on voting mechanisms to choose a single outcome from many possible outcomes.

Experimentalists have thoroughly explored different institutions that can be used to aggregate preferences. Political economists who straddle both economics and political science have carried out much of this work. Their concern is with situations where actors who have opposed interests have to settle on a single outcome and with the properties of the institution used to produce an outcome. This article first turns to the institutional mechanisms by which individuals settle on a collective outcome. The second topic turns to electoral mechanisms used in representative democracies.

Spatial committee experiments

In the late 1960s theoretical papers by Davis and Hinich (1966) and Plott (1967) described a social choice environment for spatial committees. Those committees consist of a well-defined multidimensional policy space, with actors holding fixed preferences over the dimensions, and policies represented as points in the space. Using rules that mimic many parliamentary systems, these theoretical papers demonstrate that a Condorcet winner (a policy that can defeat all others under pairwise voting) exists only under rare distributions of voters' preferences. Plott (1967) establishes the conditions under which a Condorcet winner will exist and he makes the connection between this and a Nash equilibrium of a spatial committee game. Like others, he concludes that an equilibrium is rare in multidimensional spatial committee games.

Early spatial committee experiments by Berl et al. (1976) and Fiorina and Plott (1978) provide evidence that when a Condorcet winner exists, subjects choose it or outcomes that are close to it. In games where there is no such equilibrium (which is the most common case), subjects select outcomes that scatter in the policy space. These initial empirical findings, coupled with experiments by Laing and Olmsted (1978) and McKelvey, Ordeshook and Winer (1978), defined the standard for conducting spatial committee experiments. Subsequent experiments have adopted almost identical procedures.

The standard experimental design introduces a two-dimensional policy space. The orthogonal dimensions are arbitrary (X and Y in most settings) and typically range from zero to 200 or more units. Every point in the space characterizes a policy. Preferences over outcomes are induced by assigning each subject a payoff function mapping earnings in dollars to each point in the space. While many payoff functions have been tested, most experimenters have settled on a quadratic loss function, with monetary payoffs decreasing as a function of distance from a subject's ideal point. Usually five subjects are assigned different ideal points in the space, and it is the arrangement of these ideal points that allows the experimenter to manipulate, whether a Condorcet winner exists or not. Subjects are given an initial status quo and then allowed to introduce amendments. Voting takes place following an amendment, with the winner becoming (or remaining) the new status quo. Amending takes place in between votes. A motion to adjourn, passed under a voting rule, constitutes the stopping rule for the committee decision. This serves as the standard institution for subsequent spatial committee experiments. Changing these basic institutional rules became the way to test theories of collective choice.

Experimental results in the absence of equilibrium are both frustrating and profitable. Frustration arises over the fact that committee choices tend to be clustered in similar regions of the policy space. While there appears to be some pattern to the outcomes, the process by which these outcomes arise has not been fully characterized (but see the attempt by Bianco et al., 2006). Profitably, these empirical results led theorists and experimenters to add agenda control to the structure of the game. This led to a distinction between preference-induced and structure-induced equilibrium. For example, Plott and Levine (1978) showed the effectiveness of agenda control both in the laboratory and in a natural setting. Awarding agenda power created a structure-induced equilibrium and laboratory subjects converged to it. Recent experimental work by Frechette, Kagel and Lehrer (2003) illustrates that the equilibrium favours agenda setters.

Theoretical work by Buchanan and Tullock (1962) led experimentalists to examine whether changing the proportion of actors needed to pass a policy had any effect. Experiments by Laing and Slotznick

(1983) showed that moving from simple majority rule (50 per cent plus 1) to supermajority majority rule (67 per cent) resulted in many equilibria and that subjects chose them. Schofield (1985), among others, provided the theoretical basis for when an equilibrium exists as a function of the dimensionality of the policy space, the voting rule and the distribution of voters' preferences. These theoretical findings spurred experimentalists to examine other changes to the standard committee experiment. For example Wilson and Herzberg (1987) theoretically predicted and experimentally demonstrated that when a single player holds veto power, that player's ideal point is the equilibrium. Haney, Herzberg and Wilson (1992) empirically show committee choices converging to equilibrium when a weighted voting rule is used. Such a rule requires that a single player always be included in a coalition. These results are representative of the kind of work that has dominated the experimental spatial committee agenda. Experiments on spatial committees have added to a clearer understanding of institutional mechanisms. Experimental results demonstrate that changing who has the power to set the agenda, how the agenda is built, how many votes are needed and whether players enjoy veto powers, matters.

Electoral mechanisms

A second area of interest for collective choice experimentalists is with electoral mechanisms. Three broad directions have been taken that treat different aspects of representative democracies. The first is concerned with candidate behaviour. At the heart of this research is the question of whether candidate positions will converge to equilibrium when it exists. The second direction is concerned with voter behaviour, particularly how voters behave when they have little information about candidate positions. The final direction deals with the way in which electoral rules determine the likelihood that ‘types’ of candidates are elected, where types usually refer to racial and ethnic minority candidates.

The initial experimental work on candidate behaviour focused on candidates who cared only about winning and varied the information conditions that the candidates have about the preferences of voters. Most experiments use a unidimensional policy space that guarantees an equilibrium. This equilibrium is defined by the policy preference of the median voter. In the experiments elections are sequential, with two candidates announcing positions in the policy space and voters choosing between the candidates. Voters are assigned ideal points in the policy space, the winning candidate is required to implement the announced policy and voters are paid an amount that decreases with the distance of the winning position from their ideal point. Candidates are paid only if they win. Once the election is over another election is held with candidates free to change their previously announced policy. Not surprisingly, all candidates quickly adopt the position of the median voter when they are fully informed about voter preferences. Under incomplete information about voters, candidates also converge to the median voter's position, by responding to feedback about the vote share accruing to different policy positions, as in McKelvey and Ordeshook (1985). If candidates have policy preferences whereby their earnings depend not only on winning but also on implementing a policy close to their own preferred position, then the median voter result no longer holds (see the experimental results by Morton, 1993).

When voters are uninformed about candidate positions, are they able to cast accurate ballots? With minimal information, such as biased endorsements or polls, subjects do very well at inferring candidate positions. Lupia and McCubbins (1998) and Morton and Williams (2001) consider various aspects of voter information and show that voters are able to quickly determine the positions of candidates and cast

their vote accordingly.
Finally, several experiments have focused on differing electoral mechanisms and what they mean for the type of candidates that gain election. For example Gerber, Morton and Rietz (1998) compare two voting mechanisms in an experiment to test whether one or the other disadvantages a racial or ethnic minority candidate. A form of cumulative voting (in which voters can cast more than a single vote) leads to more minority candidates being elected. This should be no surprise to collective choice theorists who have long noted that different electoral mechanisms lead to predictable variation in outcomes. Cox (1997) offers an extended discussion of such mechanisms.

What we know

Collective choice experiments provide several insights. First, when a Nash equilibrium of the underlying game exists it is a strong predictor of the outcome of the experiment. The second finding is that when there is no Nash equilibrium for the underlying game, subjects choose outcomes that cluster in predictable areas of the policy space, but the process by which that occurs is not settled. At the same time, experimentalists have implemented institutional mechanisms altering such games, thereby producing an equilibrium that subjects choose. Often those institutional changes benefit one actor (for example, by assigning agenda control to a particular player). A third finding is that incomplete information does not prevent convergence to equilibrium for either candidate platform choice or voter behaviour. The fourth finding returns to Arrow's original insight: voting mechanisms can be manipulated to achieve predictable, but very different, outcomes. It all depends on the mechanism that is implemented.

See Also

- Arrow's theorem
- experimental economics
- political institutions, economic approaches to
- social choice
- strategic voting
- voting paradoxes

Bibliography

Arrow, K.J. 1963. *Social Choice and Individual Values*. New Haven: Yale University Press.

Berl, J.E., McKelvey, R.D., Ordeshook, P.C. and Winer, M.D. 1976. An experimental test of the core in a simple n-person cooperative nonsidepayment game. *Journal of Conflict Resolution* 20, 453–76.

Bianco, W.T., Lynch, M.S., Miller, G.J. and Sened, I. 2006. ‘A theory waiting to be discovered and used’: a reanalysis of canonical experiments on majority rule decision making. *Journal of Politics* 68,

837–50.

Black, D. 1948. On the rationale of group decision making. *Journal of Political Economy* 56, 22–34.

Buchanan, J. and Tullock, G. 1962. *Calculus of Consent*. Ann Arbor: University of Michigan Press.

Cox, G.W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. New York: Cambridge University Press.

Davis, O.A. and Hinich, M.J. 1966. A mathematical model of policy formation in a democratic society. In *Mathematical Applications in Political Science*, ed. J.L. Bernd. Dallas, TX: Southern Methodist University.

Fiorina, M.P. and Plott, C.R. 1978. Committee decisions under majority rule: an experimental study. *American Political Science Review* 72, 575–98.

Frechette, G., Kagel, J.H. and Lehrer, J.H. 2003. Bargaining in legislatures: an experimental investigation of open versus closed amendment rules. *American Political Science Review* 97, 221–32.

Gerber, E.R., Morton, R.B. and Rietz, T.A. 1998. Minority representation in multimember districts. *American Political Science Review* 92, 127–44.

Haney, P., Herzberg, R. and Wilson, R.K. 1992. Advice and consent: unitary actors, advisory models and experimental tests. *Journal of Conflict Resolution* 36, 603–33.

Laing, J.D. and Olmsted, S. 1978. An experimental and game theoretic study of committees. In *Game Theory and Political Science*, ed. P.C. Ordeshook. New York: New York University Press.

Laing, J.D. and Slotznick, B. 1983. Winners, blockers, and the status quo: simple collective decision games and the core. *Public Choice* 40, 263–79.

Lupia, A. and McCubbins, M.D. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need To Know*? Cambridge: Cambridge University Press.

McKelvey, R.D. and Ordeshook, P.C. 1985. Sequential elections with limited information. *American Journal of Political Science* 29, 480–512.

McKelvey, R.D., Ordeshook, P.C. and Winer, M.D. 1978. The competitive solution for n-person games without transferable utility with an application to competitive games. *American Political Science Review* 72, 599–615.

Morton, R.B. 1993. Incomplete information and ideological explanations of platform divergence. *American Political Science Review* 87, 382–92.

Morton, R.B. and Williams, K.C. 2001. *Learning by Voting: Sequential Choices in Presidential Primaries and Other Elections*. Ann Arbor, MI: University of Michigan Press.

Plott, C.R. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57, 787–806.

Plott, C.R. and Levine, M.E. 1978. A model of agenda influence on committee decisions. *American Economic Review* 68, 146–60.

Schofield, N. 1985. *Social Choice and Democracy*. Heidelberg: Springer.

Wilson, R.K. and Herzberg, R.Q. 1987. Negative decision powers and institutional equilibrium: experiments on blocking coalitions. *Western Political Quarterly* 40, 593–609.

How to cite this article

Wilson, Rick K. "collective choice experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000598> doi:10.1057/9780230226203.0263

The New Palgrave Dictionary of Economics Online

collective models of the household

Olivier Donni

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Collective models of the household are based on two fundamental assumptions: (*a*) each agent is characterized by specific preferences and (*b*) the decision process results in Pareto-efficient outcomes. The main results of the theory of collective models then refer to the empirical issue of deriving testable restrictions on household behaviour and recovering from this some information on the structural model that can be used to carry out welfare comparisons at the individual level.

Keywords

collective models of the household; exclusive goods; household behaviour; indirect utility function; Pareto efficiency

Article

Until recently ‘unitary’ models, which assume that household members act as if they maximize a unique utility function under a budget constraint, were largely predominant in the literature on household behaviour. There is increasing agreement, however, that economists cannot ignore the fact that most households are composed of several individuals who take part in the decision process. Consequently, the ‘collective’ models, which postulate that (*a*) each household member has specific, generally different preferences and (*b*) the decision process results in Pareto-efficient outcomes, have attracted considerable attention from the profession during recent years.

To examine the properties of collective models, let us consider a household consisting of two persons, *A* and *B*, who make decisions about consumption. These persons are characterized by well-behaved utility functions of the form: $u_i(\mathbf{x}_A, \mathbf{x}_B, \mathbf{X})$, where $\mathbf{x}_i$ denotes a vector of private goods consumed by member *i* and **X** a vector of public goods $(i = A, B)$. This specification of preferences is very general; it allows for altruism but also for externalities or any other preference interaction. We denote the vector of prices for private goods by **p**, the vector of prices for public goods by **P** and the household total expenditure by *y*.

Finally, we suppose that there exists a vector of distribution factors, that is, a set of exogenous variables which influence the intra-household allocation of resources without affecting preferences or the budget constraint. Examples are given by the respective contribution of each member to the exogenous household income, the state of the marriage market or divorce legislation. These variables, which are often assigned a crucial role in the derivation of the results, are denoted by **s**.

To simplify notation, let $\boldsymbol{\pi}' = (\mathbf{p}', \mathbf{P}')$ be the vector of prices. Then, efficiency essentially means that household behaviour can be described by the maximization of a utilitarian social welfare function, that is,

$$\max_{\mathbf{x}_A, \mathbf{x}_B, \mathbf{X}} \mu(\boldsymbol{\pi}, y, \mathbf{s}) u_A(\mathbf{x}_A, \mathbf{x}_B, \mathbf{X}) + (1 - \mu(\boldsymbol{\pi}, y, \mathbf{s})) u_B(\mathbf{x}_A, \mathbf{x}_B, \mathbf{X}) \tag{1}$$

subject to $\mathbf{p}'(\mathbf{x}_A + \mathbf{x}_B) + \mathbf{P}'\mathbf{X} = y$. In this programme, the function μ determines the location of the household equilibrium along the Pareto frontier. If $\mu = 1$, then the household behaves as though member *A* always gets her way whereas, if $\mu = 0$, it is as if member *B* is the effective dictator. We denote the solutions to (1) by $\mathbf{x}_A(\pi, y, \mathbf{s})$, $\mathbf{x}_B(\pi, y, \mathbf{s})$ and $\mathbf{X}(\pi, y, \mathbf{s})$.

Characterization

The first objective of the theory of collective models is to investigate the properties of the household demands derived from (1). These properties can either be tested statistically or be imposed a priori for simplifying the estimation task. From this perspective, one crucial point is that individual demands for private goods, $\mathbf{x}_A$ and $\mathbf{x}_B$, are generally unobservable by the outside econometrician; demands for these goods are observed only at the household level, $\mathbf{x} = \mathbf{x}_A + \mathbf{x}_B$. To be useful, the restrictions derived from the collective setting have thus to characterize household demands, **x** or **X**, instead of individual demands, $\mathbf{x}_A$ and $\mathbf{x}_B$.

Let $\boldsymbol{\xi} = (\mathbf{x}', \mathbf{X}')$ be the vector of household demands. We define the Pseudo–Slutsky matrix as follows:

$$\mathbf{S} = \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\pi}'} + \frac{\partial \boldsymbol{\xi}}{\partial y} \boldsymbol{\xi}'.$$

There exist at least three different sets of testable restrictions that characterize household behaviour.

SR1 condition

Browning and Chiappori (1998) and Chiappori and Ekeland (2006) show that household demands

compatible with (1) have to satisfy the following condition:

$$\mathbf{S} = \Sigma + \mathbf{R}_1,$$

where Σ is a symmetric, semi-definite matrix and $\mathbf{R}_1$ is a rank one matrix. The interpretation is the following. For any given pair of utility functions, (a) the budget constraint determines the Pareto frontier as a function of π and *y*, and (*b*) the value of μ determines the location of the household equilibrium on this frontier. Consequently, a change in π implies a shift of the Pareto frontier. The latter entails the modification of household demands described by Σ. However, the value of μ varies as well, hence the location of the equilibrium moves along the Pareto frontier. Since the frontier is of dimension one, this effect is very restricted and defined by $\mathbf{R}_1$.

Proportionality condition

The particular structure of (1) leads to further restrictions on behaviour. To make things simple, let us suppose that the vector of distribution factors is two-dimensional: $\mathbf{s}=(s_1, s_2)$. Then, Bourguignon et al. (1993) demonstrate the following result:

$$\frac{\partial \xi}{\partial s_1} = \theta \frac{\partial \xi}{\partial s_2},$$

where θ is a scalar. Thus, the response to different distribution factors is co-linear. The interpretation is that distribution factors can only change the location of the outcome on the frontier (through function μ), and the latter is of dimension one.

Specific conditions

The econometrician is often inclined to put more structure on preferences. For example, let us suppose that agents have utility functions of the form: $u_i(\mathbf{x}_i, \mathbf{X})$. In that case, we say that agents are ‘egoistic’ in the sense that the utility does not depend on the partner's consumption. This assumption implies, in particular, that the decision process can be decentralized. In a first step, household members agree on the level of public goods as well as on a particular distribution of the residual expenditure between them. In a second step, they maximize their utility function, taking into account the level of public goods and their own budget constraint. It means, formally, that there exists a pair of functions $(\rho_A(\mathbf{p}, \mathbf{X}, Y^*, \mathbf{s}), \rho_B(\mathbf{p}, \mathbf{X}, Y^*, \mathbf{s}))$, satisfying $\rho_A + \rho_B = Y^*$ where $Y^* = Y - \mathbf{P}'\mathbf{X}$, such that the demand for private goods by member *i* is the solution to

$$\max_{\mathbf{x}_i} u_i(\mathbf{x}_i, \mathbf{X}) \text{ subject to } \mathbf{p}'\mathbf{x}_i = \rho_i.$$

Hence, household demands for private goods, conditionally on the demands for public goods, can be written as:

$$\mathbf{x} = \mathbf{x}_A(\mathbf{p}, \mathbf{X}, \rho(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})) + \mathbf{x}_B(\mathbf{p}, \mathbf{X}, y^* - \rho(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})),$$

where $\rho = \rho_A$ and $y^* - \rho = \rho_B$. This structure generates strong testable restrictions because the same function $\rho(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})$ enters each demand for private goods. Bourguignon, Browning and Chiappori (1995) explicitly derive these restrictions under the form of partial differential equations, whereas Donni (2004) shows that the demands for public goods have a particular but different structure, which implies testable restrictions as well.

Welfare analyses – identification

One of the main sources of interest in collective models is to provide the theoretical background for performing welfare comparisons at the individual level. The key concept in that case is what Chiappori (1992) calls the ‘collective’ indirect utility function. Let us suppose again that agents are egoistic. If so, the collective indirect utility function is defined as follows:

$$v_i(\pi, y, \mathbf{s}) = u_i(\mathbf{x}_i(\pi, y, \mathbf{s}), \mathbf{X}(\pi, y, \mathbf{s})). \quad (2)$$

This expression describes the level of welfare that member *i* attains in the household when he or she faces the price-income bundle(π, *y*) and a set of distribution factors **s**. This representation of utility differs from the ‘unitary’ indirect utility function in that it implicitly includes the sharing function, and hence an outcome of the collective decision process. However, the knowledge of (2) is usually sufficient to evaluate the impact of economic policies on individual welfare.

In general, if agents are egoistic, the collective indirect utility functions can be retrieved. Nonetheless, the econometrician must observe the demand for some specific goods, referred to as ‘exclusive’, which benefit only one person in the household. More precisely, we say that good *X* (*x*) is exclusively

consumed by member *i* if $\partial u_j / \partial X = 0$ ($\partial u_j / \partial x_j = 0$) for $j \neq i$. The intuition is that the household demand for ‘exclusive’ goods can be used as an indicator of the distribution of bargaining power within the household. Donni (2006) considers the case of purely private consumption $(\mathbf{X} = 0)$ and shows that, if there is a single exclusive good, the collective indirect utility functions can be identified up to composition by an increasing transformation. Similarly, Chiappori and Ekeland (2003) consider the opposite case of purely public consumption $(\mathbf{x} = 0)$ and show that, if there are two exclusive goods (one for each member), the identification is still possible. However, the general case with both private and public consumption has not been completely treated until now; see Blundell, Chiappori and Meghir (2005) for a first investigation.

Bibliographical note

The main idea of collective models can be traced back to Leuthold (1968), who estimates a model of household labour supply based on non-cooperative game theory, where the individual is the basic decision-maker. However, this model differs from collective models in that the underlying decision process does not result in efficient outcomes. It actually belongs to the family of ‘strategic’ models (which are sometimes referred to as ‘collective’ models in a broad sense). Nevertheless, a significant advance towards the development of collective models is made by Manser and Brown (1980) and McElroy and Horney (1981) at the beginning of the 1980s. These authors study the properties of models based on bargaining theory, which implies Pareto-efficiency. In that case, the location along the Pareto frontier is determined by the Nash (or Kalai–Smorodinsky) solution. However, the first formal investigation of a model based on the sole efficiency assumption is due to Chiappori (1988; 1992) in the context of labour supply decisions. This model is not explicitly examined in this article because it can be seen as a particular case of the model of consumption. Note, however, that Apps and Rees (1997), Chiappori (1997), Donni (2003), and Fong and Zhang (2001) present theoretical extensions of Chiappori's initial model, whereas Chiappori, Fortin and Lacroix (2002) exhibit empirical results. Finally, we must mention that several authors have generalized collective models to inter-temporal decisions and uncertain environment. One of the most representative examples of these studies is given by Mazzocco (2005).

See Also

- family decision making
- gender roles and division of labour
- household production and public goods
- household surveys
- individualism versus holism
- integrability of demand
- intrahousehold welfare
- labour supply

- rotten kid theorem

Bibliography

Apps, P. and Rees, R. 1997. Collective labor supply and household production. *Journal of Political Economy* 105, 178–90.

Blundell, R., Chiappori, P.-A. and Meghir, C. 2005. Collective labor supply with children. *Journal of Political Economy* 113, 1277–306.

Bourguignon, F., Browning, M. and Chiappori, P.-A. 1995. The collective approach to household behaviour. Working Paper. Paris: DELTA.

Bourguignon, F., Browning, M., Chiappori, P.-A. and Lechene, V. 1993. Intrahousehold allocation of consumption: a model and some evidence from French data. *Annales d'économie et de statistique* 29, 137–56.

Browning, M. and Chiappori, P.-A. 1998. Efficient intrahousehold allocations: a general characterization and empirical tests. *Econometrica* 66, 1241–78.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56, 63–89.

Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100, 437–67.

Chiappori, P.-A. 1997. Introducing household production in collective models of labor supply. *Journal of Political Economy* 105, 191–209.

Chiappori, P.-A. and Ekeland, I. 2003. The micro economics of group behavior: identification. Working paper. Chicago: University of Chicago.

Chiappori, P.-A. and Ekeland, I. 2006. Characterizing group behavior. *Journal of Economic Theory* (forthcoming).

Chiappori, P.-A., Fortin, B. and Lacroix, G. 2002. Marriage market, divorce legislation, and household labor supply. *Journal of Political Economy* 110, 37–72.

Donni, O. 2003. Collective household labor supply: non-participation and income taxation. *Journal of Public Economics* 87, 1179–98.

Donni, O. 2004. The intrahousehold allocation of private and public consumption: theory and some evidence from U.S. data. Working paper. Cergy-Pontoise: University of Cergy-Pontoise.

Donni, O. 2006. Collective consumption and welfare. *Canadian Journal of Economics* 39, 124–44.

Fong, Y. and Zhang, J. 2001. The identification of unobservable independent and spousal leisure. *Journal of Political Economy* 109, 191–202.

Leuthold, J. 1968. An empirical study of formula transfers and the work decision of the poor. *Journal of Human Resources* 1, 312–23.

Manser, M. and Brown, M. 1980. Marriage and household decision making: a bargaining analysis. *International Economic Review* 21, 31–44.

Mazzocco, M. 2005. Household intertemporal behavior: a collective characterization and empirical tests. Working paper. Madison: University of Wisconsin.

McElroy, M. and Horney, M. 1981. Nash bargained household decisions. *International Economic Review* 22, 333–49.

How to cite this article

Donni, Olivier. "collective models of the household." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000542> doi:10.1057/9780230226203.0264

The New Palgrave Dictionary of Economics Online

collective rationality

Lu Hong

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the concepts of individual rationality and collective rationality as they appear in the economics literature. In particular, the existing literature on social choice and aggregate demand points to a fundamental disconnect between these two notions of rationality. A possible reconciliation of this disconnect is suggested.

Keywords

aggregate demand; Arrow's impossibility theorem; collective choice; collective rationality; Debreu–Mantel–Sonnenschein theorem; Gibbard–Satterthwaite impossibility theorem; individual rationality; prisoner's dilemma; rational choice; Sen, A.; social choice; social welfare function; strategic behaviour

Article

Since ancient times, men have argued that choice should be governed by ‘desire and reasoning directed to some end’ (Sen, 1995). Much modern economic theory is based on this rational choice principle paradigm. In an individual choice problem, the individual is assumed to have a preference ordering on the set of alternatives. The individual choice is rational if, for any given decision situation, the choice made is always the best among all feasible alternatives according to the preference ordering. In a collective choice problem, be it that of a society or a committee, the definition of this rational choice principle becomes problematic. As there is presumably a huge disparity among the desires and ends of the individuals within the collective, by whose desire and whose end should the collective choice be governed? Is it reasonable to expect the collective choice to be guided by a preference ordering? If so, how should it reflect individual preferences, as the choice made by the collective influences everyone in it?

Collective rationality and social choice

Of particular interest to the idea of collective rationality is the study of social choice. In a seminal work, Arrow (1951) connects collective rationality to social choice through the idea of the existence of a social welfare function. Formally, consider a large set of conceivable alternatives, *X*, that a society faces. A preference ordering *R* (weakly preferred) on *X* is a binary relation on *X* that is both complete and transitive. Its asymmetric and symmetric parts are denoted by *P* (strictly preferred) and *I* (indifferent) respectively. There are *n* number of individuals in the society. Each individual *i* has a preference ordering R_i on the set *X*. A *social welfare function* (SWF), *F*, maps a profile of individual preference orderings $(R_1, \ldots, R_n)$ to a preference ordering on *X*. The preference ordering $F(R_1, \ldots, R_n)$ is then interpreted as the *society's preference* on *X* for the society consisting of individuals with preference orderings $(R_1, \ldots, R_n)$. If such an SWF exists, then the social choice to be made from any set of feasible alternatives can be determined by comparing any pair of feasible alternatives according to the society's preference. The social choice thus made is guided by a preference ordering to reach the best among feasible alternatives – *collective rationality* is achieved. In other words, such an SWF, if it exists, is a preference aggregation procedure aggregating individual preference orderings into a society's preference ordering according to which a rational choice can be made.

In isolation, collective rationality is trivial to reach because an SWF always exists. For example, take any preference ordering *R* on *X*; the constant function that maps every possible profile of individual preference orderings to *R* is an SWF. Obviously, this SWF is not meaningful since no information about individual preferences is reflected by society's preferences. For an SWF to reasonably aggregate individual preferences, some minimal set of conditions should be imposed. Arrow (1951) considers four conditions: U (universal domain: an SWF's domain contains all possible individual preference orderings), P (Pareto principle: if all individuals strictly prefer one alternative to another, then the society strictly prefers the first alternative to the second), I (independence of irrelevant alternatives: the way the society ranks a pair of alternatives should depend only on the way individuals rank the same pair, not on how they rank any other alternatives), and D (non-dictatorship: no single individual always gets to determine the society's preference). He shows the famous *Arrow's Impossibility Theorem*: It is impossible to have a social welfare function satisfying U, P, I and D simultaneously. In other words, collective rationality is impossible to achieve universally if society is to take into account all individuals in a minimally reasonable way.

Arrow's Impossibility Theorem jump-started the modern day study of social choice. In the huge literature of social choice theory, two strands directly relate to collective rationality formulated in the context of Arrow's Impossibility Theorem. One strand focuses on identifying domain restrictions so that social welfare functions satisfying Arrow's three other conditions exist. For example, the SWF which derives society's preference from majority voting on each pair of alternatives (majority rule) with universal domain will lead to many cycles in society's preference, violating the transitivity requirement of a preference ordering. However, if individual preferences are restricted to those that are single-peaked when alternatives can be represented in one dimension, then majority rule will not generate cycles and satisfies all other requirements of Arrow's Theorem. In general, this strand of literature proves that collective rationality can be meaningfully restored for some restricted domains (Gaertner, 2002). However, domain restrictions are severe, and outside of them the problem of society's preference cycles is global (McKelvey, 1979).

The second strand of literature directly examines the formulation of collective rationality in the

definition of Arrow's social welfare function. Arrow's SWF requires society's preferences to be orderings, that is, binary relations that are complete and transitive. Suppose that we weaken collective rationality to requiring only that society's preferences be, say, acyclic as opposed to fully transitive. Can impossibility then be avoided? More generally, is the strong collective rationality formulated by requiring society's preferences to be orderings to blame for the impossibility? This line of research concludes that, even with a weakened notion of collective rationality, the impossibility remains (Sen, 1995). Therefore, social choice cannot be expected to be collectively rational, even weakly.

Collective rationality and strategic behaviour

The aforementioned work implicitly assumes that truthful individual preferences are aggregated. If, instead, strategic behaviour is allowed, then even if we require a social choice function to be only non-dictatorial (a social choice function maps a profile of individual preferences into an alternative – a choice of the society), every such social choice function can be manipulated. This is the Gibbard–Satterthwaite impossibility theorem (Gibbard, 1973; Satterthwaite, 1975). That is, even if the collective makes up its mind about what is good for the society in a given circumstance, as long as individuals are free to report their preferences and the collective does not always choose the top alternative of a given agent's reported preference, then the collective's goal cannot be achieved.

Collective rationality and aggregate demand

The disconnection between collective rationality and individual rationality exists in other areas of economics. In consumer demand theory, the Debreu–Mantel–Sonnenschein theorem (Debreu, 1974; Mantel, 1974; Sonnenschein, 1973) states that generally aggregate demand functions do not exhibit any regularity (such as being downward sloping regarding price) even when all individual demand functions are derived from rational decisions in the sense of preference maximization under budget constraints. More specifically, for any given shape of the aggregate demand function (not necessarily downward sloping), there exists a preference profile, one preference for each consumer, such that the aggregate demand function is generated by the individual demand functions derived from that preference profile. On the other hand, empirical evidence suggests that aggregate demand functions often exhibit some regularity even when individual demand functions do not exhibit regular properties from preference maximization under budget constraints (Kirman, 2004).

Possible reconciliation of individual rationality and collective rationality

The findings in social choice theory and demand theory suggest a fundamental separation between collective and individual rationality. On the one hand, if individuals in a collective are rational, the collective choice is responsive to individuals, and the collective power does not lie in some proper subset of the collective (democratic), then the collective choice cannot be ‘collectively rationalized’. On the other hand, in some situations, collective choices can be ‘rationalized’ even when individuals in the collective do not act as rational individuals. This separation between collective and individual rationality is not unlike Buchanan's critique of Arrow's formulation of collective rationality:

> We may adopt the philosophical bases of individualism in which the individual is the only entity possessing ends or values. In this case no question of social or collective rationality may be raised. A social value scale as such simply does not exist. Alternatively, we may adopt some variant of the organic philosophical assumptions in which the collectivity is an independent entity possessing its own value ordering. It is legitimate to test the rationality or irrationality of this entity only against this value ordering. (Buchanan, 1954, p. 116)

The philosophical bases of individualism have many followers in economics. Binmore (1994, p. 142) wrote: ‘Game theorists of the strict school believe that their prescriptions for rational play in games can be deduced, in principle, from one-person rationality considerations without the need to invent collective rationality criteria provided that sufficient information is assumed to be common knowledge.’ Under the standard assumptions of game theory accounting for individual interests, these game theorists will prescribe that players defect in the Prisoner's Dilemma game. Such play leads to a Pareto-inferior outcome and thus is in conflict with the collective interest. This is not problematic if game theory is a normative theory which prescribes what people should do rationally. However, as a predictive theory it fails to match what people actually play in the Prisoner's Dilemma game. Experimental evidence shows rampant cooperation among players of the Prisoner's Dilemma game (Rapoport and Chammah, 1965; Ledyard, 1995).

If we make the organic philosophical assumption that a collective is an independent entity, then do we arbitrarily assume a criterion of collective rationality? A more reasonable way of thinking about a collective being as organic is, perhaps, to consider that, in a collective, individuals become social creatures, not mere individuals, and as such their choices have social consequences that they take into account. This can be modelled as individuals' preferences over a given set of alternatives changing depending on whether they are individuals or members of a collective. How preferences are specifically influenced may reflect culture, social convention or custom, so that they are context-dependent. But whatever the cause, this may create sufficient restrictions on the preference domain that collective rationality results as a consequence of some aggregation procedure that is democratic.

See Also

- Arrow's theorem
- rational choice and sociology
- rationality, history of the concept
- social choice

Bibliography

Arrow, K. 1951. *Social Choice and Individual Values*, 2nd edn. New York: Wiley, 1963.

Binmore, K. 1994. *Playing Fair: Game Theory and the Social Contract, Volume I*. Cambridge, MA:

MIT Press.

Buchanan, J. 1954. Social choice, democracy, and free markets. *Journal of Political Economy* 62, 114–23.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–23.

Gaertner, W. 2002. Domain restrictions. In *Handbook of Social Choice and Welfare*, ed. K. Arrow, A. Sen and K. Suzumura. Amsterdam: North-Holland.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Kirman, A. 2004. The structure of economic interaction: individual and collective rationality. In *Cognitive Economics: An Interdisciplinary Approach*, ed. P. Bourgine and J. Nadal. Berlin: Springer Verlag.

Ledyard, J. 1995. Public goods: a survey of experimental research. In *Handbook of Experimental Economics*, ed. J. Kagel and A. Roth. Princeton: Princeton University Press.

Mantel, R. 1974. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7, 438–53.

McKelvey, R. 1979. General Conditions for Global Intransitivities in Formal Voting Models. *Econometrica* 47, 1085–112.

Rapoport, A. and Chammah, A. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor: University of Michigan Press.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Sen, A. 1995. Rationality and social choice. *American Economic Review* 85, 1–24.

Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6, 345–54.

How to cite this article

Hong, Lu. "collective rationality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/

article?id=pde2008_C000546> doi:10.1057/9780230226203.0265

The New Palgrave Dictionary of Economics Online

Collet, Clara Elizabeth (1860 – 1948)

R.D. Collison Black

From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Collet, Clara E.; women's employment

Article

After a period as a schoolteacher in Leicester, Clara Collet became one of Charles Booth's assistants on his Survey of London Life and Labour in 1886. In 1893 she entered the civil service as labour correspondent and later senior investigator for women's industries in the newly established Labour Department of the Board of Trade. The earnings and employment of women became and remained Clara Collet's main concern; her contemporaries recognized her as the principal authority on the subject in Britain. Articles on female labour and earnings were among her contributions to the first edition of Palgrave's *Dictionary of Political Economy* in 1894, and the thorough and lucid reports which she produced on women's industrial employment figured in Parliamentary Papers, contributing to the passing of the original Trade Boards Act of 1906. After her retirement in 1920 from what had by then become the Ministry of Labour, Collet herself served on a number of trade boards, and wrote the section on Domestic Service for the *New Survey of London Life and Labour* directed by her former chief, Sir H. Llewellyn Smith.

The first woman Fellow of University College, London, where she took her MA degree in 1885, Clara Collet was one of the founders in 1890, along with Henry Higgs and H.R. Beeton, of the Economic Club which met there monthly, and acted as its secretary from 1905 until 1922.

She was also a founder member of the British Economic Association, which later became the Royal Economic Society; she served on its Council from 1920 to 1941, and on that of the Royal Statistical Society from 1919 until 1935.

Selected works

1893–4. Royal Commission on Labour. Employment of Women. Miss C.E. Collet. Report. *Parliamentary Papers* 1893–4 [C.6894, XXIII] xxxvii, pt. I.

1894. Statistics of employment of women and girls. Miss Collet. Report. *Parliamentary Papers* 1894 [C.7564], 1xxxi, Pt. II.

1894. (With Dora M. Barton.) Female labour. In Palgrave's *Dictionary of Political Economy*, London: Macmillan & Co., Vol. II.

1894. Females and children, earnings of. In Palgrave's *Dictionary of Political Economy*, London: Macmillan & Co., Vol. II.

1898. The collection and utilization of official statistics bearing on the extent and effects of the industrial employment of women. *Journal of the Royal Statistical Society* 61, 219–60.

1899. Money wages of in-door domestic servants. Miss Collet. Report. *Parliamentary Papers* 1899 [C.9346] xcii.

1931. Domestic service. Chapter VIII in *New Survey of London Life and Labour*, ed. H. Llewellyn Smith, London: P. S. King, Vol. II.

1933. Appendix to *The Private Letter Books of Joseph Collet, sometime Governor of Fort St. George Madras*, ed. H. H. Dodwell. London: Longmans.

How to cite this article

Black, R.D. Collison. "Collet, Clara Elizabeth (1860–1948) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000612>
doi:10.1057/9780230226203.1880

The New Palgrave Dictionary of Economics Online

Colquhoun, Patrick (1745 – 1820)

D.P. O'Brien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Colquhoun, Patrick; education; unproductive labour

Article

Colquhoun was born in Dundee. A successful early career in business led to the position of Lord Provost of Glasgow in 1782 and 1783. In 1789 Colquhoun moved to London and became active as a magistrate. He worked on the provision of poor relief and put forward plans for the reform of London's police. He died in London in 1820.

Colquhoun's interest in poor relief led to his *New and Appropriate System of Education for the Labouring People* (1806), a pamphlet based on his own experience of running a school in Westminster. Like Thomas Chalmers later, he argued for the necessity of education to raise the standards and aspirations of the poor, though primarily in order to curb vice rather than population. This, he believed, was the most cost-effective way of tackling poverty. His *Wealth, Power and Resources of the British Empire* (1814), his last important work, is the one for which he is best known. This contained detailed figures on incomes and occupations and the relative importance of agriculture and manufacturing in Great Britain and Ireland. He also included a history of the public revenue, and descriptive material on the colonies. The work was not very securely based; McCulloch, who had first-hand experience of trying to construct large-scale statistical data for his *Commercial Dictionary*, was severely critical of it in the *Edinburgh Review* and in *Brande's Dictionary*. But it was followed by later writers and Colquhoun's estimate that unproductive labour, one fifth of the total, received one third of output was widely quoted.

Selected works

1806. *A New and Appropriate System of Education for the Labouring People*. London: J. Hatchard.

1814. *A Treatise on the Wealth, Power and Resources of the British Empire, In Every Quarter of the World, Including the East Indies; The Rise and Progress of the Funding System Explained; With*

Observations on the National Resources for the Beneficial Employment of a Redundant Population. London: J. Mawman.

Bibliography

Blaug, M. 1958. *Ricardian Economics*. New Haven: Yale University Press.

Deane, P. 1956. Contemporary estimates of national income in the first half of the nineteenth century. *Economic History Review* 8, 339–54.

Espinasse, F. 1887. Colquhoun, Patrick. In *Dictionary of National Biography*; reprinted, Oxford: Oxford University Press, 1973.

[McCulloch, J.R.] 1835. State and defects of British statistics. *Edinburgh Review* 61, 154–81.

McCulloch, J.R. 1837. *A Statistical Account of the British Empire; Exhibiting its Extent, Physical Capacities, Population, Industry, and Civil And Religious Institutions*. London: C. Knight.

How to cite this article

O'Brien, D.P. "Colquhoun, Patrick (1745–1820)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C0000214> doi:10.1057/9780230226203.1856

The New Palgrave Dictionary of Economics Online

command economy

Richard E. Ericson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The concept of a ‘command economy’, a construct in the theory of comparative economic systems, is defined, and its origins, characteristics, and consequences for any society in which it is implemented are explored. The impossibility of the absolute centralization which it requires generates compromises with the market forces it aspires to replace, fostering a symbiotic marketized ‘second economy’ which systematically undermines its foundations. Hence, although initially appearing to be a true alternative to the market economy, a command economy, most nearly realized in the Soviet Union (1930–87), proved to be ultimately non-viable, collapsing under reforms attempting to make it competitive with market systems.

Keywords

active vs passive money; aggregation; balance; bounded rationality; bureaucracy; central planning; centralization; command economy; command mechanism; command principle; communism; complex social economy; corruption; decentralization; discretion; Gorbachev, M. S.; gross output; Inca production system; incentive provision; industrialization; inequality of income; information; *khozraschet* ; labour discipline; micro-balance; market mechanism; market versus plan; moneyness; Mormon economic system; Neurath, O.; *perestroika* ; price control; principal and agent; rationing; resource allocation; second economy; sellers' market; shadow economy; soft-budget constraint; Soviet economic reform; Soviet Union; Stalin, J. V.; suboptimization; unit of measure; vested interests; war and economics; war communism

Article

A command economy is one in which the coordination of economic activity, essential to the viability and functioning of a complex social economy, is undertaken through administrative means – commands, directives, targets and regulations – rather than by a market mechanism. A complex social economy is

one involving multiple significant interdependencies among economic agents, including significant division of labour and exchange among production units, rendering the viability of any unit dependent on proper coordination with, and functioning of, many others.

Economic agents in a command economy, and in particular production organizations, operate primarily by virtue of specific directives from higher authority in an administrative/political hierarchy, that is, under the ‘command principle’. Thus the life cycle and activity of enterprises and firms, their production of output and employment of resources, adjustment to disturbances, and the coordination between them are primarily governed by decisions taken by superior organs responsible for managing those units' roles in the economic system. One of the most distinctive features of such an economy is the setting of the firm's production targets by higher directive, often in fine detail. The administrative means used include planning, material balances, quotas, rationing, technical coefficients, budgetary controls and limits, price and wage controls, and other techniques aimed at limiting the discretion of subordinate operational units/firms. The command principle strives to fully and effectively replace the operation of market forces in the key industrial and developmental sectors of the economy, and render the remaining (peripheral) markets manipulable and subordinate to political direction. Thus the command principle is likely to clash with the operation of market forces, yet a command economy may nonetheless contain and rely on the market mechanism in some of its sectors and areas, for example, influencing labour allocation, or stimulating small-scale private production of some consumables.

The term ‘command economy’ comes from the German *Befehlswirtschaft*, and was originally applied to the Nazi economy, which shared many formal similarities with that of the Soviet Union. It has received its fullest development in the analysis of the economic system of the Soviet Union, particularly under Stalin, although it has been applied to wartime administration of the US economy (1942–6; see Higgs, 1992), the Mormon economic system in mid-19th century Utah (Grossman, 2000), and the Inca production system in the 16th century Andes (La Lone and La Lone, 1987). Synonymous or near-synonymous terms include ‘centrally planned economy’, ‘centrally administered economy’, ‘administrative command economy’, ‘Soviet-type economy’, ‘bureaucratic economy’ and ‘Stalinist economy’.

The command economy's conceptual origins go back to the Viennese economist Otto Neurath, who in the years before and after the First World War developed an extreme version (to the point of moneylessness) based chiefly on prior experience with wartime economies (Raupach, 1966). The concept of the command economy has since become a central conceptual framework in the analysis of economic systems, as it captures a logically coherent alternative to ‘the market’ as a way of organizing socially complex economic activity and interaction. The Soviet Union provided the most complete, and for a while successful, example of a command economy as a working alternative to a market system. Indeed, apart from the relatively short-lived Nazi case, and even briefer ones under emergency conditions in some other countries, especially in wartime, actual instances of command economies are virtually limited to Communist-ruled countries, with the USSR as the prototype and prime exemplar. Thus, what follows is mainly inspired by the Soviet example (Ericson, 1991) as it existed, essentially little altered since its appearance in the 1930s, until its collapse in the aftermath of President Gorbachev's *perestroika*, begun in 1987.

Nature of the command economy

The seminal analysis of the nature, characteristics, and problems of a command economy is Grossman (1963).

Any complex social economy must, for its very survival, maintain at least a ‘tolerable’ micro-balance, ‘that minimal degree of coordination of the activities of the separate units (firms) which assures a tolerably good correspondence between the supply of individual producer and consumer goods and the effective demand for them’ (Grossman, 1963, p. 101). In such an economy, appropriate balance can be achieved through decentralized, market-based (monetized, price-mediated) interaction of autonomous units, or by virtue of explicit specific coordinating directives (commands, targets) from some higher authorities. While the former is characteristic of a market economic system, the latter is defining of a ‘command economy’. In the latter operational-level units (for example, firms) must merely ‘implement’ commands; they become ‘executants’ of plans and directives from above, plans which must insure balance through the coherence and consistency of the instructions they give. Thus the command mechanism requires relative centralization and severe restriction on the autonomy of subordinate operational units. It derives from the overwhelming priority of social goals, and requires the severe limitation, if not total destruction, of autonomous social and economic powers and the enforcement of strict obedience to directives.

A command economy is hence a creature of state authority, whose marks it bears and by whose hand it evolves, exists and survives. Command economies are imposed, whether through external duress or imitation, or indigenously in order to achieve specific purposes such as (*a*) maximum resource mobilization towards urgent and overriding national objectives, such as rapid industrialization or the prosecution of war; (*b*) radical transformation of the socio-economic system in a collectivist direction based on ideological tenets and power-political imperatives; and (*c*), not least, curing the disorganization of a market economy brought about by price control, possibly occasioned by inflationary pressure arising from (*a*) and/or (*b*).

The command economy therefore requires a formal, centralized, administrative hierarchy staffed by a bureaucracy, and it also needs to be embedded in (at least) an authoritarian, highly centralized polity if it is not to dissolve or degenerate into something else. And that bureaucracy, if it is to effectively implement the command principle, must exercise full control and discretion, if not necessarily formal ownership, with respect to the creation, use and disposal of all productive property and assets. At the same time, each office or firm and every economic actor within the command structure holds interests which, if only in part, do not coincide with those of superiors or of the overall leadership. This generates important problems of vested interests, principal–agent interaction, incentive provision, and general enforcement of the leadership's will, and calls for a variety of monitoring organizations (party, police, banks, and so on). The term ‘command’ must not be taken to preclude self-serving behaviour, bureaucratic politics, bargaining between superiors and subordinates, corruption, peculation and (dis) simulation. On the contrary, such behaviour tends to be widespread in a command economy; yet the concept of a ‘command economy’ remains valid so long as, in the main, authority relations and not a market mechanism govern the allocation of resources.

When not externally imposed, command economies typically arise from a millennialist elite, with unique access to ‘the truth’, achieving the political power to impose its will, while facing a crisis of apparently overwhelming proportions. The perception of a life-threatening crisis, driving the need for massive mobilization of all social resources and rendering potentially disastrous any hesitation or dissent, any

questioning of ways and means, naturally leads, pushed by the ‘logic of events’, to the usurpation of all power of discretion, all legitimate authority, by the ‘knowing’ elite, which then becomes responsible for all that is done or not done in the society and the economy. The crisis may be artificial or real (‘hostile encirclement’), externally or internally imposed (the need to industrialize, to ‘catch up’), but it requires moving resources rapidly and massively, forcing new activities and interactions in the face of severe scarcities, of shortage of competent personnel, of massive uncertainties, and of strongly held, stark priorities. Indeed, a sense of overwhelming urgency and need for haste drove the elite of the Soviet Union in the 1930s to test and establish, through trial and error over several decades, the institutional structure of a ‘command economy’, albeit *less than absolute* from both necessity and choice (for example, due to the ‘lessons’ of ‘War Communism’) (Grossman, 1962; Zaleski, 1968).

Consequences of command

Rational application of the command principle calls for planning, which is basically of two types. Longer-term, developmental planning expresses the leadership's politico-economic strategy (for example, five-year and ‘perspective’ plans); shorter-term, coordinative planning (annual, quarterly, monthly, ten-day) ideally translates the strategy into resource allocation while aiming to match resource requirements and availabilities for individual inputs, goods, and so on, in a sufficiently disaggregated way for given time periods and locations. The task of elemental coordination, of micro-balance, so effortlessly accomplished by any functioning (however poorly) market system, is overwhelmingly large, and grows rapidly with industrialization and economic development, both of which lead to exponential growth of the complexity of the economy, and hence of the planning problem. With centralization and the abandonment of markets comes the need for massive, detailed coordinative planning, for ‘making ends meet’ in the expanding web of interconnections that must be maintained for economic life to continue. Coordinative planning serves, therefore, as the basis for specific operational directives to producers and users, thereby implementing the command principle to achieve the prime imperative of a social economy – ‘balance’.

> It is *this* task that in fact consumes the largest part of the so-called planning in the command economy … Coordinative planning as it is conducted in the Soviet Union does little by way of consciously steering the economy's development or finding efficient patterns of resource allocation. Its overwhelming concern is simply to equate both sides of each ‘material balance’ by whatever procedure seems to be most expeditious. (Grossman, 1963, p. 108)

A major problem is that detailed planning and the corresponding directives are often late, are insufficiently detailed, may lack the requisite information, hence often cannot be effectively coordinated, and owing to their rigidity are peculiarly vulnerable to uncertainty (Ericson, 1983). Information in the command sector, by the logic of the system, tends to flow vertically up and down the administrative hierarchy rather than horizontally between buyer and seller, adding to difficulties of demand–supply coordination by informationally isolating operational units from their suppliers and users. In addition, problems of motivation, accountability (down as well as up), inappropriate decision-making parameters,

and divergent interests complicate the procedure. Even at best, this manner of resource allocation can hope to attain only internal consistency (in the sense of effectively matching partially disaggregated requirements and availabilities) but not a higher order of economic efficiency. Economic calculation in pursuit of efficiency enters, if at all, at the project-planning stage, and not short-term resource allocation and use.

These problems are aggravated by the logic of haste that drove imposition of the command economy – ‘the pressing contrast between urgent political goals and available resources’. (Grossman, 1963, p. 108)

The necessary attention to the growing problem of balance further militates against any effort to consider developmental objectives or efficiency in making allocative decisions, so that a further bias against allocative efficiency is built into the command economy. Coupled with limited ability to gather, filter, process, and communicate information, and to compute solutions to planning problems, this creates a fundamental and growing inability to acceptably solve the underlying coordination problem, and hence further undermines any consideration of efficiency.

The logic of ‘command’ has a number of other consequences reflected in the institutions of such an economy. Planning in a command economy must be largely in *physical terms* due to the crucial importance of balance. The bottom line of the planning process must be available physical units of required inputs, in appropriate assortment, quantity and timing, necessitating physical targets for production and input utilization. Thus tens of thousands of materials and equipment balances must be drawn up and coordinated for each plan period, and then broken down and allocated in directives to specific implementers. And, to be directly usable, these must be in physical or crypto-physical (constant price) units that directly relate to the production processes being coordinated. Using economic-value units requires flexible and changing, marginal scarcity-based prices for valuation, as well as giving significant autonomy to subordinate units that inevitably then will make the trade-offs in assortment, quantity and timing within planned constraints on values (that is, ‘budgets’). Hence, such valuations pose a fundamental challenge to the command economy.

Planning in physical terms, however, leads to ‘enormous waste and inefficiency, to production for waste as much as for use’ (Grossman, 1963, p. 110). There are at least three fundamental sources of this elemental waste: *grossness*, *aggregation*, and *unit of measure*. The need for these arises from the overwhelming complexity of the task of planning for, and directing the operation of, a complex social economy and the necessarily limited information gathering, processing, and dissemination capabilities of any economic agent or agency. However, the emphasis on *gross output* leads to ‘input intensiveness’, waste, and ignoring cost considerations. *Aggregation* leads to persistent subcategory imbalance in assortment, quality, type, timing, and so on, while *units of measurement* determine suboptimization objectives, distorting implementation decisions, particularly when they are, for material balance reasons, input oriented. Thus each of these is essential for the feasibility of directive central planning, of the command mechanism, yet each loses, or destroys, essential information for the ‘proper’ (in the eyes of the system directors) implementation of plans, and opens space for creative interpretation of instructions/commands, and hence for ‘suboptimization’ by implementing units whose interests are not perfectly aligned with those of the centre (Nove, 1977). While the command mechanism logically requires unauthorized initiative to be forbidden, and strictly punished when exercised, the size of the task it faces inevitably opens the opportunity, indeed often the need, for such unauthorized initiative. Thus the physical quantity planning required by the command economy to maintain minimal functional ‘balance’ contains its own antithesis, unleashing forces that undermine the consistency of the plan and the

coherence and balancedness of its realization. This fundamental contradiction lies behind most of the critical problems of the command economy in the Soviet Union and the myriad efforts to resolve them within the framework of the command mechanism that comprise the endless waves of reform following victory in the Great Fatherland War of 1941–5.

The ‘logic of command’ thus imposes a need to restrict autonomy, to restrict the capability of economic units to pursue any other than ‘planned’ or commanded purposes: economic agents must not have the capability to autonomously acquire and deploy resources for any purposes outside the plan. Comprehensive material balance planning and centralized materials and equipment allocation provide a necessary component, but one that is insufficient unless resources, including human, are denied the capability of autonomous movement and application. Severe restrictions on labour mobility, albeit not as severe as under Stalin, are required, as are comprehensive restrictions on the use of any ‘generalized command over goods and services’ – that is, money – that might be used to alter their patterns of allocation and use in the economy. The system must be substantially demonetized in order to ‘… constrict the … range of choice in the face of the state's demands’ (Grossman, 1966, p. 232).

Thus money must be deprived of ‘moneyness’ and prices must be kept ‘passive’, as mere accounting and measurement units. According to the logic of the command economy, the availability of money and the prices at which commodities and products are provided should have no essential impact on the allocation of goods and services, or on the nature and direction of economic/industrial development; all real activity is preordained in the plan and its subsequent implementing commands. The role of money is then to facilitate monitoring of commanded performance through the financial flows it generates. Thus monetary prices do not, and indeed should not, reflect to a substantial degree social goals and priorities; they merely reveal and measure the flow of commanded activity. Producer prices (and most retail prices), wages, prices of foreign currencies, and so on are generally centrally set and controlled, often remaining fixed for long periods of time. Micro-disequilibria naturally abound, while the widely perceived dubious meaningfulness of such prices and the administrative allocation of most producer goods in physical terms combine to sustain the system of detailed production plans and directives in terms of physical indicators – yet another bar to more efficient planning and management.

Finally, an absolutely essential, indeed defining, institution of the command economy is the *physical rationing* of resources and producers' goods. This is where the market is most fully and directly replaced, and where the central authorities have the ability to most directly influence and control the behaviour of subordinate operational units. It implements the centralized mobilization of resources to priorities, the most direct response to crises and challenges. And it most directly denies to subordinates the capability to produce, to develop, in ways outside those authorized in the plan. This makes the coexistence between the command principle and the market mechanism a source of continual conflict, as the market opens unauthorized opportunities to subordinates. In the Soviet Union the command principle, aided by the club of materials rationing, repeatedly pushed back and eliminated the market mechanism when (timidly) introduced in reforms, until the system collapsed in chaos, and the introduction of a full-fledged market economy was begun in 1992 (Schroeder, 1979; Aslund, 1995).

Thus the nature of the command system makes it fundamentally incompatible with real markets, although some market institutions can, and indeed must, be allowed to function both within the non-state sectors and as the interface between them and state economic institutions/sectors.

Inherent challenges to the command economy

As Grossman notes in his seminal article (1963, p. 107), ‘The chief persistent systemic problem of a command economy is the finding of the optimal degree of centralization (or decentralization) under given conditions and with reference to given social goals’. The fundamental dilemma is that full centralization poses an insoluble problem, while decentralization abandons the ability to direct, to control development, and to ensure the pursuit of social goals and priorities. With regard to the pure planning problem, a large body of theoretical literature arose in the late 1960s, and continued into the 1980s, on the problem of decentralizing the planning process to make its informational and computational burden manageable (Eckstein, 1971; Bornstein, 1973). But the problem is far greater, and less studied, with respect to implementation; rational planning is swamped by the struggle to maintain elemental coordination.

Decentralization versus priority

Looked at through the prism of relative advantages, operational decentralization shortens ‘lines of communication’, increasing flexibility, adaptation and responsiveness to a changing environment through local initiative and innovation, and vastly simplifying the decision problem of economic agents. But it does so at the cost of weakening or losing the ‘advantages of centralization’, including enforcement of regime values, capability for large-scale resource mobilization, concentration of scarce talent in central decision-making organs, and the maintenance of macro-balance. In particular, decentralization compromises the ability of the centre to directly manage the development and structure of the economy and to force the achievement of critical priorities regardless of cost. Furthermore, decentralization requires the introduction of the alternative coordination mechanism to insure tolerable micro-balance – the market – as decentralization undercuts the ability to directly coordinate, to balance from above. Thus, to prevent catastrophic imbalance, a more active money with economically flexible market prices must be allowed to function in a decentralized system.

The impossibility of planning and commanding the performance of all economic agents in full operational detail, however, forces some decentralization. This creates a chronic threat to balance which is thus a continuous argument for (re)centralization of planning and materials allocation. Furthermore, a partial decentralization of planning and management in a command economy may do more harm than good; it may impair balance without yielding sufficient benefit. Yet a complete decentralization, in the sense of a virtually full devolution of the major production decisions to the firm level, would be disastrous from the standpoint of balance, unless the price structure were properly altered to provide proper signals to firms *and* suitable behavioural rules were prescribed, that is, unless a market mechanism were introduced. Thus the logic of command predicts a ‘treadmill of reforms’ (Schroeder, 1979), an array of countervailing strengthenings of the oversight and control organs (in particular, the Party), and enhancements of their role in the economy, accompanying moves towards decentralization in the state sector. It also explains the Soviet institutional arrangement of inter-firm contracts as a decentralized implementation device. These are required to specify details of interaction within planned categories, and establish observable, and hence legally enforceable, commitments to planned implementation, constraining the autonomy necessarily granted through the minimal decentralization. And it explains the logic of the continuing restraints on the use of money and the continuing efforts at

effective price control to keep the autonomy of agents restricted to the minimum necessary for the continued functioning of the less-than-absolute command economy.

Even limited decentralization requires that money be used in the command sector (as well as in the household sector), but its role as a bearer of options and as the means of pecuniary calculation for decision-making is necessarily limited and deliberately subordinated to the planners' will and the administrators' power. Banks and the treasury accommodate the money needs of production, ensuring a soft budget constraint for the individual firm. At the same time, the ‘moneyness’ of money at the firm level is low, hemmed in as it is by administrative constraints and impediments, including the rationing of nearly all producer goods, and by the widespread ‘seller's market’ (shortages of goods and absolute lack of buyers' alternatives). This monetary ease, together with the sellers' market, plays an important role in ensuring individual workers' job security at the firm level and full employment in the large, while keeping the firm largely insensitive to money costs and/or benefits.

Within the command sector, money and prices have a necessary role in determining terms of alternate resource uses *only within* planned/commanded categories, and money has the role of limiting total claims to resources in areas, or at a level of detail, beyond the reach of plan directives. This requires ‘businesslike management’ within the firm – *khozraschet*, which is a ‘set of behavioral rules that is supposed to govern the actions of Soviet managers beyond their primary responsibility, the fulfillment of output targets’. It pushes the firm toward ‘technical efficiency’ and limitation of ‘claims on society's resources for productive use. … *khozraschet* is a system that is well devised to control the behavior of managers in a command economy where a certain amount of devolution of power to them is inevitable, and where, further, managers' goals and values do not necessarily coincide with the official ones’ (Grossman, 1963, p. 117). Thus money also has the role of facilitating the monitoring of performance in the command sector.

While administrative orders are the rule in a command economy, backed up by greater or lesser degree of state coercion (depending on country and period), any decentralization of implementation naturally relies on monetary (‘material’) incentives to elicit desired individual compliance and performance. Compounding the incentive problems arising from differences in information and interests between central authorities and implementing agents is the fact that the physical and other indicators to which the material incentives are linked may often be poor measures of social benefit (as seen by the leadership). Furthermore, resort to such rewards widens the distribution of official earnings and raises questions of permissible limits of income inequality. Yet there may be little choice in that the state must in effect compete with the much higher incomes from the *second economy*. Indeed, the Soviet Union during War Communism, Cuba in the 1960s, and the People's Republic of China during some periods before Mao's death in 1976 tended to downgrade material incentives in favour of normative controls, but never did quite abolish them.

The behaviour of the Soviet-type firm has been much studied (Granick, 1954; Berliner, 1957; Nove, 1977; Freris, 1984). Because its directives and the corresponding managerial incentives stress physical output, produced or shipped, and thanks to its low sensitivity to cost and the ambient sellers' market, the firm often sacrifices product cost, quality, variety, innovation and ancillary services to its customers to sheer product quantity. By the logic of command and the requirements of plan manageability, firms operate in an environment with sole suppliers and assigned users, reducing complexity by eliminating ‘wasteful’ redundancy in production and distribution. Thus firms in a command economy are largely insulated from any product competition, both from the outside world and from other domestic firms,

thanks to the climate of administrative controls and the prevalent excess demand for their output. Difficulties with supply, frequent revision of its plans, interference by Party and other authorities, and other systemic problems also stand in the way of its more efficient and effective operation. Indeed, to function at all, the firm's management is frequently forced to break rules and even resort to criminally punishable acts.

This compounds a further critical challenge posed by necessary decentralization – the conflict between the will, purposes, incentives and priorities of the higher authorities and those at lower levels, particularly of the firms and their managements. Even the best-motivated managers, following all official rules and incentives, will sometimes fail to replicate the decisions that their superiors would have made had they been in a position to make them. This problem is aggravated by the inevitable ambiguity, incompleteness and inconsistency of those rules, incentives and the information available on the spot. Only binding physical constraints and observable outcomes can be systematically enforced, making ‘centralized materials allocation the most powerful weapon at the disposal of the central authorities’ (Grossman, 1963, p. 118). Thus, where material inputs are less determinate of a unit's activities, this information and incentive problem is greater, and the defiance of central will relatively more widespread and successful. This observation explains the non-viability of any reform that fails to fundamentally alter the materials allocation system.

Under-planned, ill-commanded sectors

A major challenge to the command economy also arises from the existence of sectors outside, or only partially affected by, the command principle. In the Soviet Union these included most of agriculture, much of housing, the household sector and some consumer goods and services. ‘Markets’ were allowed to function for the distribution of final consumer goods and services, including agricultural produce, for much of the activity of the ‘collective’ sector in agriculture and for household labour supply. For transactions with ‘personal property’ within the household and collective sectors, money was active and agents responded to market prices, while in the quasi-markets interfacing with the state sector – for example, labour and consumer goods – money was relatively active but prices remained largely controlled and non-market. These are sectors where information on needs/preferences and capabilities proved too difficult to acquire reliably in real time for acceptable allocation and balance to be commanded, and so at least one side of a market was allowed to function with an active money. Here, the command mechanism proves too crude and clumsy, and hence politically counterproductive, to be used outside of pressing emergencies. Indeed, this might be considered a lesson of War Communism, the first experience with a command economy in Soviet Russia, 1918–21.

In view of the theoretical incompatibility of command and market, how could these ‘market’ sectors be successfully grafted on to the command mechanism? An explanation (Grossman, 1963) rests on the trade-offs between the authorities' limited capabilities, the complexity of those sectors, and their centrality to regime priorities. A sector which provides significant inputs to physical planning and plan fulfilment, where the unpredictability in the flow of goods is unacceptable, cannot be left to the market without seriously undermining command. However, if a ‘market’ sector can be treated as a residual for purposes of materials planning and allocation, a buffer for planning, then its coexistence is acceptable. Further, if its operation is characterized by rapid change and complexities rather outside the core interests of the regime, if without disrupting the industrial core greater incentives and risk can be placed

on those peripheral agents, and if non-market constraints can force the desired market response from it, then the centre will want to separate that sector from the command sphere, lowering its coordination burden by shifting it to the market.

These considerations were indeed active in the case of those sectors ‘left to the market’ in the Soviet Union: consumer goods retailing, the acquisition of labour services, the support of households in the countryside through a private agricultural sector, and a few peripheral and interstitial activities. Indeed, any attempt to truly ‘marketize’ any other sectors or activities in the command economy is doomed to fail unless the loss of fervour, of the sense of mission and urgency, leads to abandonment of the command mechanism. Yet even the existence of these limited market sectors, providing an outlet for incentive earnings and diverted resources, exerts a continuing corrosive pressure on the command economy and its control mechanisms.

The cancer of ‘money’

A truly monumental challenge to the command economy lies in the role of money in any *less-than-absolute* command economy. As the complete centralization of decisions in the production sector (let alone in the household sector) is an impossibility, something must be left to local initiative and dispersed decision making. Thus *khozraschet* is a logical necessity, ‘… an unfriendly bridgehead that threatens to seize ground whenever the planner fails or defaults’ (Grossman, 1966, p. 228). With the inevitable devolution of some decision making to firms and households, money acquires a necessary and critical role in the command economy, going well beyond that consistent with the logic of command. That role arises from the need to economize in making decentralized decisions, and as a medium of exchange and store of value in the decentralized interactions that relate to all decisions. In acquiring this role, this ‘moneyness’, it allows accumulations of power outside the control of the regime. Money is a ‘bearer of options’ whose power and influence must be restrained if the command mechanism is to operate properly – to determine priorities and to insure maximal commitment to their achievement. As Grossman (1962, p. 214) noted, ‘Money is a form of social power that may lead resources astray and is subject to only imperfect control by political authority.’

Thus the power of money has to be curbed in a command economy by limiting balances available to households and firms, by compartmentalizing money into cash and ‘firm’ circuits, and by erecting barriers and limits to the use of ‘monies’ in each category, although that undercuts the effectiveness of attempted decentralizations. Liquidity, ‘moneyness’, is constrained by the institutional structures and by all the characteristics and conditions of the ‘sellers' market’, rendering ‘money’ the only non-scarce commodity, in unusable excess to the extent the command mechanism is effective. Monetary policy in the properly functioning command economy reduces to limiting the volume of cash in the economy (‘macro-monetary’ control) through wage fund restrictions and cash control absorption plans of the retail sector, and the allocation of firm balances in restricted categories (‘micro-financial’ control) in just sufficient quantity to support the implementation of the plan, with confiscation of excess funds to prevent unauthorized activity by the firm (Garvy, 1977).

Similarly, the price system, expressed in terms of that money, must also be mobilized to the purpose of control. The inflexible, administratively segmented, average cost-based prices in command economies are a logical necessity of command- and haste-based shortage. For all the problems they cause, all the

unintended consequences and distortions in the behaviour of subordinates, such prices help to keep money largely passive, at least in the core state sectors, and allow both money and prices to remain instruments, rather than disrupters, of command. More than being ideologically justified, such prices are a response to the pragmatic and pressing requirements of running a shortage economy with a rapidly developing system of centralized direction of enterprises and of materials allocation.

Money, however, is not so easily contained. Once in unobserved hands, it exercises its ‘command over goods and services’ without reference to plans, commands or regime priorities. Hence, given any discretion, in any sphere of activity not directly monitored agents will naturally use money in ways they find desirable, placing new demands on a physical system otherwise tautly planned and characterized by general scarcity. This is facilitated by the existence of agents and spheres of activity outside the command system, providing ‘legitimate’ sources and uses for monies, however acquired or disposed. And the possibility of acquiring money provides incentives for unauthorized activities, incentives to undertake unplanned interactions and reallocations. An active money vastly expands the sphere of discretion of ‘subordinate’ agents beyond any authorized by a decentralizing reform, and calls for severe administrative restrictions, a reduction to passivity, if it is not to disrupt the planned activities and discretion of the central authorities.

Yet attempts to administratively constrain the influence, the ‘corruptive’ power, of money become increasingly futile once the ‘genie’ has been ‘let out of the bottle’. Even limited decentralizing reform, allowing money to influence some (subcategory) production and allocation decisions, inevitably lets loose more liquidity, more of a command over goods and services, than desired. This arises from a multitude of factors: errors in both physical and financial plans, inherent incompleteness of plans and commands due to limited information and time and to the necessity of aggregation, changing circumstances and shocks to the economy, mistakes in implementation and in responding to shocks, the irregularity and disruptions in the materials allocation system, the behavioural response of even the most enthusiastic and best-intentioned agents to these problems, and so on. All of these can lead to an unexpected lack of funds for doing what was commanded (if only implicitly), and hence disruption of commanded performance, unless additional liquidity is provided.

Thus monetary policy in a command economy, once money is allowed any room for activity, must be accommodating; a lack of funds can never be allowed to disrupt planned performance, just as an excess of funds cannot be allowed to facilitate unplanned or unauthorized activity. Thus the role, the influence, of money has a natural, inexorable tendency to grow: insufficient funds become an immediate problem generating new money through credits or additional allocations, while unused funds tend to stay hidden until ferreted out by inspection or accidental discovery. And as it grows, so does the challenge to the command principle. An increasing number of agents, in both the state and non-state sectors, has a growing ability to access resources, to divert them in the name, if not always the interest, of implementing decentralized plans, and thus to challenge the priorities of the political authorities. This growing challenge becomes a cancer in the system, a growth that undermines its health and feeds tendencies destructive of the priorities of the regime and its rulers.

The ‘second economy’

As the command economy matures, as the messianic fervour with which it was imposed wanes and the use of extraordinary force diminishes in ensuring compliance with commands, these challenges to

command metastasize into a competing yet symbiotically attached and dependent economic system: the *second economy*. This name highlights the distinction of this sphere of economic activity from the officially sanctioned, ‘first’, command economy. It is thus defined as ‘all production and exchange activity that meets at least one of two criteria: (a) being directly for private gain; (b) being in some significant respect in knowing contravention of existing law’ (Grossman, 1977, p. 25).

In the Soviet Union, attempts to strengthen ‘material incentives’ and activate ‘the profit motive’ in order to increase the effectiveness and technical efficiency of the implementation of central plans and directives and to stimulate technological progress and innovation, and the growing monetization of the agricultural sector, opened the door to massive expansion of money supply and eroded the barriers between the currency and the enterprise bank account monetary circuits. Collective farms and their subsidiary enterprises, owners of ‘small means of transport’, vodka manufacture and distribution, and the Caucasus republics (Georgia in particular) proved particularly rich sources of illicit (from the system's perspective) monetization and private ‘entrepreneurial’ activity. This both raised the spectre of inflation and opened the door to vastly increased opportunities for manipulation by self-interested subordinates in the command sector. Thus the use of ‘economic levers’ greatly increased the opportunity for and incidence of bribery, corruption, speculation, and even ‘honest’ private labour.

While the fundamental cause of the appearance and growth of the ‘second economy’ undoubtedly lies in the congenital institutional weaknesses of the command economy discussed above, there are a number of proximate sources that make it unsurprising. These include extensive price control, with consequent scarcity and misallocation, high taxes on non-state activities/incomes, prohibitions of private activity, unmet individual consumption needs, poorly protected impersonal (state) property, the personal power of bureaucrats and ‘gatekeepers’, and other historical factors, including the end of terror. These provide both motives and opportunities for officially illicit activity and for the authorities to overlook that activity. With the ageing of command and the decay in enthusiasm of its agents, the growth of such a second economy appears natural.

Thus growing ‘monetization’, the existence of ready and waiting market sectors, and the decline in the use of violent instruments of enforcement lead to a growing sphere and importance of activities outside the purview of ‘planning’ and ‘command’. These market-mediated activities are at times supportive, helping to achieve tolerable micro-balance in the increasingly complex economy, but often are in violation of planned implementation and regime values. Private interests, necessarily allowed some leeway, grow in significance, increasingly seizing ground from command. In the Soviet Union, the private agricultural sector, initially permitted only to secure survival of the peasantry under the extractive pressure of rapid industrialization, and the consumers' personal services sectors provided the basis for a ubiquitous, if still systemically marginal, second economy.

But then even the core industrial sectors under the command mechanism find their managers and activities increasingly influenced by this illegitimate, shadow market, system, as managers are often forced to break rules and undertake illegal acts in order to do their job. Such acts, together with ubiquitous and protean illegal activity on private account, add up to a large underground economy characteristic of every command economy, which together with legal private activity (allowed in varying degree in different countries) both supports and supplements the ‘first economy’ and is inimical to it. While the second economy significantly adds to the supply of goods and services, especially for consumption, it also redistributes private income and wealth, contributes to the widespread official corruption, and generally criminalizes the population. Virtually every area of economic life is touched

upon, and often entangled with, ‘second economy’ activities, while legal private activity naturally opens a loophole for illegal trading and entrepreneurship, generally below the purview of the authorities. And it goes hand in hand with the extension of corruption, ensuring that it remains outside of official notice. Those ‘violations’ of legality within the command sector, a ‘shadow economy’, build informal inter-enterprise relations which are generally beneficial to the operation of state enterprises. They work to substantially correct the allocative failures of the command mechanism, improving firm performance and hence benefiting its management, and also provide lucrative opportunities for managers to directly benefit through the activization of barter, personal connections, and bribery. However, they also spawn further distortions in economic behaviour, as managers seek to generate access to cash, the life blood of the ‘second economy’, to extract rents, and to hide their activity from supervisory and statistical organs.
Thus the second economy plays a dual and contradictory role in the command economic system. First, it addresses a number of the problems of coordination and balance endemic to the command mechanism, reallocating both producers' and consumers' goods, facilitating plan fulfilment and the use of financial incentives, and generating new incomes and ‘politically safe’ outlets for private initiative. Hence it becomes important for enhancing consumer welfare, for production stability, and even for social stability. The ‘second economy’, and in particular its ‘shadow’ side, plays an essential role in the first economy as a ‘pressure valve’, a release ‘fixing command’ by maintaining micro-balance and covering ‘holes’ in economic life left by the mistakes or oversight of the planners and central managers. And this role becomes increasingly important as the economy grows more complex and diversified, and hence becomes less susceptible to conscious oversight and direction.
As the central authorities struggle with their loss of control, searching for a solution through reform, decentralization and recentralization, monetization and administrative restriction, agents in the economy take advantage of gaps in control, of the autonomy and discretion offered by growing liquidity of the quasi-money in the system, to deal with problems of coordination and balance, inconsistency of plans and commands, and ubiquitous shortages and scarcities. Of course they operate in light of their own partial information, and in their own (private as well as official) interests, but in so doing save the system from collapsing under its own weight and rigidity (Powell, 1977). Thus the second/shadow economy provides a spontaneous surrogate economic reform that imparts a necessary modicum of flexibility, adaptability and responsiveness to a formal set-up that is too often paralyzing in its rigidity, slowness, and inefficiency. In doing so, the second economy also provides a valuable stabilizing influence on society and the polity, making life livable and the system humanly manipulable and responsive to private inducement. It makes everyone complicit in the way things work, equally ‘guilty’ before state and society, while providing an almost legitimate, and not politically dangerous or directly destructive, outlet for individual initiative and entrepreneurship. Finally, it relieves inflationary pressures (a ‘monetary overhang’) resulting from the command economy's necessary combination of monetary looseness and pricing rigidity.
Despite this positive functional role, the second economy also has a less positive *systemic* impact. It mocks the pretense of social direction and control, subverts its egalitarian impulse by accentuating differences in access and income, and gives the lie to the pretense of a ‘new’ ideologically correct (‘Soviet’) man. Its very existence and usefulness thus subvert the ideology of the regime, and it works against and undercuts regime priorities by exposing the incompetence and incapacity of the authorities. Its provision of alternatives weakens the ‘plan, production, and labour discipline’ so essential to the proper operation of the command mechanism. Indeed, it attacks the core of the command mechanism as

it '… elevates the power of money in society to rival that of the dictatorship itself, rendering the regimes implements of rule less effective and less certain' (Grossman, 1977, p. 36). In particular, it corrupts officialdom and distorts prices, adding a (positive or negative) 'second economy margin', both 'in kind' and in money, breaking prices as an effective instrument of control. This weakens monetized incentives for state activities by providing competing, and often better, alternatives to them. Hence the second economy, and in particular the 'shadow economy' in the state sector, completes the cancerous development of agent autonomy, of the ability to work outside the plan and its subsequent commands, by providing viable alternatives to the plan.

Other dysfunctional impacts, undermining the operation of the command system, arise from its diverting of resources and products to unplanned sectors and activities, including diversion from development/investment priorities to consumers. This naturally generates undesirable (from a system perspective) redistribution of incomes, although recipients, including many high-placed officials, find it very desirable. Indeed, it is further disruptive of command by creating a 'two-tiered' system of prices and incomes, of consumer goods and labour markets. One tier is comprised of the low-priced, scarcity-ridden quasi-markets of the 'less-than-absolute' command economy, where the unenterprising, the overly scrupulous, and the 'slow' can survive. The other tier consists of real, albeit highly distorted, markets in the generally high-priced, risky but well-endowed second economy where the enterprising, entrepreneurial, and criminal can thrive. In this high tier, substantial incomes are generated and allocated, although they largely accrue to corrupt officials and 'gatekeepers' of scarce materials or permissions who can extract rather phenomenal 'rents'. The inequities this generates further undermine the legitimacy of the regime and generate potentially explosive social pressures, only partially relieved by the second economy's 'pressure valve' aspects.

Finally, it is worth noting that the second/shadow economy, through its activity outside of the officially measured sphere, seriously distorts statistical data and the information available to planners and allocators in the official economy, and, due to its illegality, also hides necessary information from other agents in the shadow economy. This aggravates the economic problems that spawn 'second/shadow economy' activities, deepening the contradictions between the centre and decentralized agents, and further corroding the institutional structures of the command economy.

Performance and fate

Command economies have been instrumental in radically transforming societies more or less according to their drafters' intents, in mobilizing resources for rapid industrialization and modernization, at times on a vast scale, and in rapidly amassing industrial power and military strength. Indeed, they have shown themselves highly effective in rapidly implementing large-scale projects and achieving overriding social goals, albeit at great cost. It is this effectiveness, when cost is no object, which explains why the command principle is resorted to in times of emergency and war. Hence in the Soviet Union command facilitated defence during, and rapid recovery and rebuilding of the Stalinist economy after, the massive trauma of the Great Fatherland War. Economic growth has been especially marked (though not unparalleled by market economies) where large amounts of unemployed and underemployed labour and rich natural resources could be mobilized and combined with existing (advanced, Western) technology, and where the public's material improvement could be restrained, or even seriously depressed, under strong political control. As these possibilities waned, and as the economies grew in size and complexity

and thus became less amenable to centralized administrative management, rates of growth declined sharply. At the same time, the shortcomings of the command mechanism in adapting production to demand and its changes – providing consumer welfare, effecting innovation, serving export markets – became more apparent and less tolerable. This led to much discussion and repeated attempts at controlled institutional reform, at decentralizing and stimulating subordinate initiative without sacrificing ultimate control.

Some actual reforms in the externally imposed command economies of eastern Europe went so far as to introduce or extend the market mechanism to such a degree that one could no longer regard the system as a Soviet-type command economy, even if, before the 1990s, one could not speak of it as a full-fledged market economy either. Yugoslavia since the early 1950s, Hungary since 1968 and especially in the 1980s, and post-Mao China are the most important cases in point. Other actual reforms were of a minor or ‘within-system’ nature, aiming to decentralize certain types of decisions while eschewing the market mechanism and retaining the hierarchical form of organization and the command principle. In the hope of stimulating efficiency to revive growth rates, the decentralizing measures were accompanied by a number of other ‘reforms’ relating to organizational structure: prices (still controlled), incentives, indicators, materials rationing, and so on. The Soviet reforms of 1965, and those in the 1970s and 1980s prior to *perestroika*, were of that kind; many similar ones took place in other Communist countries after the mid-1950s and prior to the overthrow of Communism in 1989. On the whole, such reforms had little success in addressing the problems of the command economy. Bureaucratic and political obstacles apart, the attempt to decentralize economic decisions without bringing in a market mechanism almost inevitably leads to economic difficulties. The beneficiaries of devolution of decision-making lack the necessary information to produce just what the economy requires or to invest to meet prospective needs, and the coordination of plan-subsequent command is lost. Moreover, they may apply the additional power at their disposal to advance particularist causes or to divert resources into illegal channels. Microeconomic disequilibria mount, and soon superior authorities step in to recentralize on a case-by-case basis and the reform withers away (Grossman, 1963; Wiles, 1962, ch. 7; Kontorovich, 1988).

This failure of reform reflects the inherent contradictions of the command economy framed in the irreconcilable conflict between ‘command’ and ‘money’ discussed above (Ericson, 2005). The Soviet command economy, driven by the urgent need for and haste in industrialization and military development, initially relegated the influence of money and the market to the margins of the system, where they handled areas and activities in which command had been revealed as counterproductive during War Communism. That system, the ‘less than absolute command economy’, substantially industrialized, triumphed in the Great Fatherland War, and recovered to an almost perfect replica of its pre-war self by 1950. But by then the strains of its inherent inflexibility and the bounded rationality of the system's planners and managers began telling on continuing growth and the development of the economy. With economic growth came increasing complexity and growing intractability of the central planning and economic management problem. Some decentralization became essential, and increasingly so as time passed, opening the door once again to the rise of money as a significant influence on the operation and development of the economy. And that influence was only enhanced by the ageing and mellowing of the system. With the passing of ‘terror’ as an effective incentive mechanism, the stabilization of personnel and the regularization of procedures, it became ever harder to control agent behaviour, to contain the distractions of money and the self-interests it mobilized, and to uncover the rents that well-placed agents were able to extract, thus aggravating the inherent agency problems of the

command economy.

The remaining years of the Soviet system thus witnessed an epic struggle, barely perceptible at first, but increasingly evident as reforms, decentralizations, reorganizations and recentralizations cycled around each other in the search for a solution to the increasingly evident and destructive malperformance and waste, and aggravating behavioural distortions in response thereto, generated by the struggle between the ‘command principle’ and the weak, but inexorably emerging, ‘market’. Initially reflected in the dysfunctions of the marginal and quasi-markets of the command economy, and in the struggle to harness a ‘passive’ money to the purposes of command, the role of money grew along the ‘treadmill of reforms’ into the rival, if still largely subordinate and complementary, ‘second economy’, and in particular its ‘shadow’ component, on which the ‘command principle’ increasingly came to depend for its effectiveness. As long as the Soviet system remained a ‘command economy’, commands had to have last word, and money remained largely relegated to the sidelines, exercising its influence within the quasi-monetized instruments (‘economic levers’) of the command mechanism and the distorted markets of the second economy.

This inherent conflict, played out over Soviet history, revolves around a number of fundamental dualities, elemental oppositions which characterize these primary forces. The ‘command principle’ derives most basically from the urge, the will to control, to ‘rationally’ determine and direct the future, exercised by a ‘gnostic’ elite, immanent in the Party. It knows what needs to be done, by whom and how, and can tolerate no dissent or deviation. Juxtaposed to this ‘Will of Society’ stand the millions of independent ‘wills’, desires and objectives, anarchically coordinated through ‘the market’, whenever that set of institutions broke through the barriers and limits placed by ‘command’. This provides the foundation for the eternal struggle between ‘central priorities and control’ and ‘agent incentives and capabilities’.

This opposition is severely aggravated by urgency, by ‘virtuous haste’, in the pursuit of overriding social goals and central objectives. For the mobilization for, and focus of resources on, these priorities trample on the information, capabilities and goals of individual and organizational agents which must perforce implement that mobilization, implement those priorities. ‘Effectiveness’ in the pursuit of social objectives becomes opposed to ‘efficiency’ in the attainment of any objectives, denies trade-offs based on local information and incentives, and hence blocks flexibility in response to changing circumstances. Indeed, the single-minded pursuit of overriding objectives, of absolute priorities, naturally disrupts the fine coordination, the requirements of ‘balance’, necessary to consistently and efficiently pursue any objectives.

Throughout the history of the Soviet Union, the needs of centralization, given Soviet social goals, stood in fateful opposition to the necessity to decentralize in order to keep the system tolerably functioning. The latter necessity spawned repeated (partial) remonetizations and a ‘second economy’ that both shored up the operational foundations of the ‘first economy’ and undermined its long-term viability, corroding its ideological and systemic foundations. Money so unleashed intensified the dysfunctions and contradictions of the ‘command economy’, spurring further repeated ‘reforms’ and ‘experiments’ that merely further aggravated the inconsistencies, the ‘oppositions’ in the system, until the central leadership, largely unintentionally and out of ignorance, destroyed the ‘command economy’ in the radical systemic and economic ‘restructurings’ beginning with *perestroika* in 1987.

See Also

- agency problems
- decentralization
- informal economy
- second economy (unofficial economy)
- soft budget constraint
- Soviet economic reform
- Soviet growth record

Bibliography

Aslund, A. 1995. *How Russia Became a Market Economy*. Washington, DC: Brookings Institution.

Berliner, J. 1957. *Factory and Manager in the USSR*. Cambridge, MA: Harvard University Press.

Bornstein, M., ed. 1973. *Plan and Market*. New Haven, CT: Yale University Press.

Eckstein, A., ed. 1971. *Comparison of Economic Systems: Theoretical and Methodological Approaches*. Berkeley: University of California Press.

Ericson, R. 1983. A difficulty with the ‘command’ allocation mechanism. *Journal of Economic Theory* 31, 1–26.

Ericson, R. 1991. The classical Soviet-type economy: nature of the system and implications for reform. *Journal of Economic Perspectives* 5(4), 11–27.

Ericson, R. 2005. Command vs. shadow: the conflicted soul of the Soviet economy. *Comparative Economic Systems* 47(4), 1–27.

Freris, A. 1984. *The Soviet Industrial Enterprise*. New York: St Martin's Press.

Garvy, G. 1977. *Money, Financial Flows and Credit in the Soviet Union*. New York: Ballinger Publishing Company.

Granick, D. 1954. *Management of the Industrial Firm in the USSR*. New York: Columbia University Press.

Grossman, G. 1962. The structure and organization of the Soviet economy. *Slavic Review* 21, 203–22.

Grossman, G. 1963. Notes for a theory of the command economy. *Soviet Studies* 15(2), 101–23.

Grossman, G. 1966. Gold and the sword: money in the Soviet command economy. In *Industrialization in Two Systems*, ed. H. Rosovsky. New York: Wiley.

Grossman, G. 1977. The ‘second economy’ of the USSR. *Problems of Communism* 26(5), 25–40.

Grossman, G. 2000. Central planning and transition in the American desert: latter-day saints in present day sight. Online. Available as a preprocessed paper at http://www.econ.berkeley.edu/~grossman/mormons.pdf, accessed 23 June 2006.

Higgs, R. 1992. Wartime prosperity? A reassessment of the U.S. economy in the 1940s. *Journal of Economic History* 52, 41–60.

Kontorovich, V. 1988. Lessons of the 1965 Soviet economic reform. *Soviet Studies* 40, 308–16.

La Lone, M. and La Lone, D. 1987. The Inka state in the southern highlands: state administration and production enclaves. *Ethnohistory* 34, 47–62.

Nove, A. 1977. *The Soviet Economic System*. London: Allen and Unwin.

Powell, R. 1977. Plan execution and the workability of soviet planning. *Journal of Comparative Economics* 1, 57–76.

Raupach, H. 1966. Zur Entstehung des Begriffes Zentralverwaltungs-wirtschaft. *Jahrbuch für Sozialwissenschaft* 17(1), 86–101.

Schroeder, G. 1979. The Soviet economy on a treadmill of reforms. In *Soviet Economy in a Time of Change*, vol. 1, ed. Joint Economic Committee, US Congress. Washington, DC: Government Printing Office.

Wiles, P. 1962. *The Political Economy of Communism*. Cambridge, MA: Harvard University Press.

Zaleski, E. 1968. *Stalinist Planning for Economic Development*. Durham, NC: Duke University Press.

How to cite this article

Ericson, Richard E. "command economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000219> doi:10.1057/9780230226203.0266

The New Palgrave Dictionary of Economics Online

commodity fetishism

Andrew Levine
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An analysis of Marx's notion of ‘commodity fetishism’ – as a theory of the necessary (systemically induced) misperception of underlying production relations by participants in market exchanges. The appeal of the notion to the two main opposing tendencies of mid- and late 20th-century Marxism – Marxist humanism and structuralist Marxism – is discussed. Reasons are proposed to account for a recent decline of interest in the phenomenon among both economists and philosophers. It is suggested, however, that the concept remains viable.

Keywords

Althusser, L.; capitalism; capitalist social relations; commodity fetishism; exploitation; Feuerbach, L.; invisible hand; labour power; labour theory of value; laws of motion of capitalism; Lukacs, G.; market relations; Marx, K.; Marxist humanism; structural Marxism; surplus value; value controversy

Article

Since Plato, philosophy and then science have assumed first, that there is often a difference between appearance and reality; and, then, that it is sometimes possible to grasp what really is the case by investigating how things appear. Marx's account of commodity fetishism, a crucial step in his account of the capitalist mode of production, implements these assumptions explicitly. It describes how exchange relations appear to economic agents, where the appearance belies the reality at the same time that it provides cognitive access to it.

Market exchanges occur in all modes of production capable of sustaining an economic surplus. In capitalism, the process is generalized – not just in the sense that markets structure economic life but also, more importantly, because everything is commodified that can be. Universal commodification is the result of a protracted process that is definitively launched once labour – or, more precisely, labour power (labour time, adjusted for differences in intensity) – is commodified. The commodification of labour

power is pivotal because this commodity is the sole source of value and therefore, ultimately, of wages, profits and rents. The generation and distribution of surplus value, of what is produced in excess of what is needed to reproduce the labour power expended in production processes, is the invisible underlying reality upon which perceptions of exchange relations depend. To persons engaged in buying and selling labour power, what appears is just that, as in any other exchange, individuals aim to do as well for themselves as they can, given their resources, their preferences, and the production technologies available to them. But what is really going on is a struggle over the distribution of the economic surplus at the point of production. That reality is opaque. Economic agents are therefore governed by the appearanceof rational economic agents maximizing payoffs to themselves. In his account of commodity fetishism, Marx shows how this inevitable misperception helps to reproduce and sustain the underlying reality.

When Marx expressly addresses this phenomenon at the conclusion of the opening chapter of the first volume of *Capital* (1867), the economic agents he describes are property-holding individuals. Thus it is not exactly capitalism that he aims to model, but ‘simple commodity production’, an ahistorical idealization. However, the cogency of his account is unaffected as his analysis becomes more historical and concrete – to the point that the direct producers are, as in full-fledged capitalism, a propertyless proletariat with nothing to exchange except, of course, their own labour power. Commodity fetishism is therefore a general and pervasive fact wherever capitalist social relations hold sway. Thus the term denotes a systemic opacity at the level of appearance that helps to hold economic agents in thrall by masking the exploitation of labour. Because this misperception sustains the exploitation that engenders it, revolutionaries intent on overthrowing capitalism must tear away the veil of illusion by revealing the exploitation of workers that exchange relations conceal.

Marx does not directly address *how* commodity fetishism comes into being or how it is sustained. But he does provide fragments of an explanation when he focuses on the atomizing effects of market relations. All resource allocation mechanisms are social in the sense that they bring together a host of disparate and heterogeneous economic activities. However, where the commodity form prevails, the social character of market transactions is apparent only after goods and services are produced. The workers know that the corn they consume is produced by farmers, and the farmers know that the tools they use in growing corn are made by workers. Everyone also knows that, without food, workers would not be able to make the tools farmers use; and that, without tools, farmers would not be able to grow food for the workers. It is therefore evident in retrospect that workers and farmers are engaged in a collective endeavour. But it is not similarly evident prospectively. From that vantage point, it seems only that farmers and workers – and also the capitalists who provide them with means of production – are making individual choices aimed at bringing about the best outcomes for themselves, given the constraints they face. Even if they believe that these essentially egoistic activities are somehow socially beneficial, they can justify this belief by appealing to the workings of an ‘invisible hand’. Because there is no visible hand that directs the process, the terms of interaction appear *as if* they are forces of nature to which individuals must accommodate. Thus market relations appear as infrangible constraints that human beings are obliged to operate within, not as social constructions that human beings can change. In terms that Kant introduced and that Marx, following Hegel, effectively assumed, freedom (autonomy) is then forfeit. Wills are heteronomously determined, governed by laws of an (apparently) impersonal *other* (the market system itself). To be free, we must therefore take control of the aggregation mechanism we have concocted. We do so by putting reason in command – not just at the individual level of the rational

economic agent, but at the societal level as well.

What Marx says about commodity fetishism is concise and intriguing. For these reasons, and because it summarizes the very abstract analysis of the commodity form with which *Capital* begins, his account of the phenomenon has always been well known. ‘Commodity fetishism’ is one of those terms that everyone associates with Marx. But, even in what remains of Marxist circles, the basic tenets of Marx's account have faded from ongoing discussions. A number of factors have contributed to this turn of events: among them, the legacy of the so-called ‘value controversy’ of the 1970s; the efforts of mathematical economists in the 1970s and 1980s to put the categories of Marxist economic analysis on a sound, analytical footing; and attempts by analytical philosophers, working on Marxist themes, to reconstruct and, when possible, defend core Marxist positions. The conclusion that has emerged is that, *pace* Marx, there is nothing special about the commodification of labour power and therefore that the theory of surplus value cannot be sustained in the way that Marx believed. Nowadays, it is only the most doctrinaire Marxists who uphold the labour theory of value, the basis for Marx's account of commodity fetishism. This fact along with the decline of political movements that identify with the Marxist tradition and, its inevitable consequence, waning interest in Marx's work itself, has, for the time being, made commodity fetishism a matter of concern mainly to historians of economic thought.

Not long ago, the situation was quite the opposite. From roughly the 1950s through the 1970s, commodity fetishism played a central role in the two most important and innovative tendencies in Marxist theory: Marxist humanism and structuralist Marxism. These were opposing tendencies, politically and substantively. But they converged on according commodity fetishism centre stage. Marxist humanists sought to de-Stalinize Marxism by recovering its Left Hegelian roots. This meant reading Marx's work through the prism of his early writings, before he broke with his ‘erstwhile philosophical conscience’, as he proclaimed in 1845 in *The German Ideology*. For the Left Hegelians, Ludwig Feuerbach's philosophical anthropology, elaborated in *The Essence of Christianity* (1841), was fundamental. There Feuerbach ‘inverted’ the theological dogma that ‘God makes Man’ by showing how the God idea is an ‘objectification’ of essential human traits. Lacking materiality, God is purely an objectification, an ‘alienated’ expression of the human essence. In taking consciousness of this fact, one recovers essential humanity and becomes emancipated from the thrall of its systemic misrepresentation. In the *Paris Manuscripts* (1844), Marx applied the Feuerbachian programme to objects of labour; ‘objectifications’ too of essential humanity, but also material things and therefore not objectifications only. Feuerbach arrived at his conclusions by ‘interpreting’ the theology of Right Hegelian theologians. His working hypothesis was that they had gotten the concept of God right, but that they radically misconstrued what the concept *means*. In the *Paris Manuscripts* Marx treated (Smithian) political economy the same way. He assumed that it correctly describes ‘economic facts’. The task, then, was to interpret those facts – in order to reveal the alienation they express and, in so doing, to advance the emancipatory project of Left Hegelianism. How successful Marx was in implementing this programme is subject to debate. What is clear is that, as the focus of his theoretical work turned away from Hegelian philosophy towards political economy, history and politics, he became disabused of the idea that Adam Smith or any other classical economist had gotten political economy descriptively right. His life's project, thereafter, was to rework the conceptual apparatus of classical economics – more usually in its Ricardian, not Smithian, form – with a view to revealing the real ‘laws of motion’ of the capitalist mode of production. In this endeavour, Feuerbachian philosophical anthropology seemed to play no role. But, following the lead of Georg Lukacs (1923) several decades earlier, the Marxist humanists pointed out

that there was, in *Capital*, an explicit point of connection – in the text on commodity fetishism. It was there that Marx brought his analysis of the commodity form to completion. But it was also there that, in modelling the commodity form, Marx identified the objectification of essential human traits in the process of capital accumulation. In consequence, capital, becomes a ‘fetish’, a *god* in Feuerbach's sense – one who controls economic behaviour by force of (illusory) power.

Structuralist Marxists, like Louis Althusser, were intent on reading Left Hegelianism out of the Marxist canon. They therefore treated Marx's references to fetishes and gods as ironic figures of speech, even as they attempted to enlist the text on commodity fetishism in the service of opposition to Marxist humanism. Borrowing a concept from the French philosopher Gaston Bachelard (1884–1962), Althusser (1965) disparaged Marx's early work by asserting the existence of an ‘epistemological break’ within the Marxist corpus. What he had in mind was roughly a ‘paradigm shift’ – not, however, within an ongoing scientific practice but between pre-scientific modes of thought and the inception of a new science. In Althusser's account, two previously monumental epistemological breaks had occurred – one that established mathematics in ancient Greece, and one that established the sciences of nature in 17th-century Europe. Marx's achievement was supposedly on a par with these; he opened up a science of history. He did so by anticipating the structuralist turn the ‘human sciences’ (in France mainly) would later take – first in linguistics and psychology, later in anthropology and psychoanalysis. Specifically, in *Capital* and other writings of his maturity, Marx explained a range of diverse ‘surface’ phenomena by construing them as effects of the workings of a relatively small number of underlying, generally invariant ‘deep’ structures. The text on commodity fetishism lent itself to this construal of Marx's explanatory practice in as much as it depicted the perceptions of economic agents as effects of the unseen but causally efficacious process of surplus value extraction. Thus Marx's account can be seen as a theory of necessary (systemically induced) *misperception* – consonant with notions of explanation that contemporaneous structuralists endorsed. Perhaps the most innovative use Althusser made of commodity fetishism was in his theory of ideology, according to which modes of production constitute experiential subjectivity by ‘interpellating’ the human subjects who support or ‘bear’ them.

We now inhabit a different intellectual universe. In the past several decades, it has come to be widely believed, by erstwhile Marxists as much as by ‘bourgeois economists’, that Marx's focus on production rather than exchange inhibited the development of analytical economic tools. In so far as this belief is sound, the emphasis Marxists placed on commodity fetishism is partly to blame. The explanatory strategies of Marxist humanists and of structuralists have fallen into disrepute, too – largely because, in both cases, though for different reasons, the alleged connections between appearance and reality were never satisfactorily explained. No sustainable account was given either of how interpretation should proceed in the Marxist humanist case or, in the structuralist case, of how deep structures can be discerned in surface phenomena. Thus, commodity fetishism has fallen on hard times. However, we should not conclude that there is nothing viable in the concept or in the theoretical traditions that, until recently, magnified its importance. Hegelianism certainly, and structuralism possibly, still have much to teach us. The last word may not yet have been said on the theory of surplus value, either. If and when interest in Marx resumes, it will certainly be useful to revisit these issues. The notion of commodity fetishism played a key role in mid- and late 20th-century Marxism. The core idea it articulates – that necessary misperceptions sustain the capitalist order – can again provide useful insights. The concept may not be forever doomed to be of historical interest only.

See Also

- capitalism
- Marx's analysis of capitalist production

Bibliography

Althusser, L. 1965. *For Marx*. New York: Pantheon Press, 1969.

Althusser, L. 2002. Ideology and ideological state apparatus. In *Lenin and Philosophy and Other Essays*. New York: Monthly Review Press.

Althusser, L. and Baliber, E. 1968. *Reading Capital*. London: New Left Books, 1970.

Cohen, G. 1988. The labour theory of value and the concept of exploitation. In *History, Labour and Freedom: Themes from Marx*. Oxford: Clarendon Press.

Feuerbach, L. 1841. *The Essence of Christianity*. Amherst, NY: Prometheus, 1989.

Fromm, E. 1975. *Marx's Concept of Man*. New York: Continuum.

Lukacs, G. 1923. *History and Class Consciousness*. London: Merlin Press, 1967.

Marx, K. 1844. *Paris Manuscripts* (also called *Economic and Philosophical Manuscripts*). In Karl Marx and Frederick Engels, *Collected Works*, vol. 3. New York: International Publishers, 1975.

Marx, K. 1845. *The German Ideology*. In Karl Marx and Frederick Engels, *Collected Works*, vol. 5. New York: International Publishers, 1976.

Marx, K. 1867. *Capital*, vol. 1. Moscow: Progress Publishers, 1965.

Roemer, J. 1981. *Analytical Foundations of Marxist Economic Theory*. New York: Cambridge University Press.

Steedman, I. 1981. *Marx After Sraffa*, rev. edn. New York: Shocken.

How to cite this article

Levine, Andrew. "commodity fetishism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000231> doi:10.1057/9780230226203.0267

The New Palgrave Dictionary of Economics Online

commodity money

François R. Velde and Warren E. Weber

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Commodity money is a medium of exchange that may be transformed into a commodity, useful in production or consumption. Although commodity money is a thing of the past, it was the predominant medium of exchange for more than two millennia. Operating under a commodity money standard limits the scope for monetary policy, actions that alter the value of money. However, it does not eliminate monetary policy entirely. The value of money can be altered by changing the commodity content or legal tender quality of monetary objects, or by restricting the conversion of commodities into money or vice versa.

Keywords

Bank of England; banking; bimetallism; brassage; Bretton Woods system; coinage; commodity money; convertibility; debasement; discount rate; fiat money; foreign exchange markets; gold standard; Great Depression; inflation; monetary policy; money changers; seigniorage

Article

A commodity is an object that is intrinsically useful as an input to production or consumption. A medium of exchange is an object that is generally accepted as final payment during or after an exchange transaction, even though the agent accepting it (the seller) does not necessarily consume the object or any service flow from it. Money is the collection of objects that are used as media of exchange. Commodity money is a medium of exchange that may become (or be transformed into) a commodity, useful in production or consumption. This is in contrast to fiat money, which is intrinsically useless. Commodity money can also be thought of as a medium of exchange that contains an option to consume a predetermined service flow at little or no cost. The option can be exercised in various ways, depending on the object. Coins can be melted down (at little cost) and the metal applied to non-monetary uses. In the case of paper or token money under a commodity money standard, the medium of exchange itself is

intrinsically useless, but it is costlessly convertible into a specified quantity of the commodity on demand. Fiat money can also be converted into goods or services, but in quantities that will depend on market prices.

Commodity money is a thing of the past; countries worldwide now use fiat money standards. However, this is a relatively recent development. Commodity money, primarily in the form of coined metals, was the predominant medium of exchange for over two millennia. Although operating under a commodity money standard limits the scope for monetary policy, it does not eliminate it entirely. The history of commodity money is replete with numerous ways in which governments have altered the monetary system to achieve various goals.

From commodity money to fiat money

In early or primitive societies, it is often difficult to characterize the general patterns of trades and transactions, let alone determine how generally accepted a particular commodity might be. Nevertheless, a wide range of commodities have been reportedly used as money (cowry shells, wampum, salt, furs, cocoa beans, cigarettes and so on), perhaps the most exotic being the stone money of the island of Yap in Micronesia.

General acceptability of monetary objects is most clearly ascertained when the objects are standardized and exchanged repeatedly. With metallic commodities, the standardized objects are called coins. Coinage of metal began in the eastern Mediterranean region or the Middle East, India and China between the sixth and fourth centuries bc. Coinage has developed in parallel and broadly similar ways in these areas.

The metals most commonly used have been gold, silver and copper (in decreasing order of scarcity), in varying degrees of fineness (silver mixed with substantial amounts of copper, called billon). Lead, tin and various copper alloys (bronze, brass, potin) have also been used, although less frequently than the more common metals. The metal is either mined or acquired through trade. The most common method of coinage is striking with a die, although cast coins are also found. In many legal traditions the right of coinage is a prerogative of the public or central authority, although it may be delegated or leased to regional authorities or private parties. This prerogative may also extend to mining. In other words, the rules governing the supply of commodity money vary from government monopoly to minimal regulation.

In Europe and the Mediterranean, coinage – an invention mythically linked to Croesus, King of Lydia – began near the Aegean Sea in the sixth century bc. The use of money developed considerably in Greek and Roman times, leading to a three-tiered system of gold, silver, and copper denominations. In the Roman empire, the provision of coinage was a government monopoly. The collapse of the empire in the West led, after a long transition, to a purely silver-based monetary system, with a largely decentralized provision of minting. Uniformity of coinage was restored under Charlemagne but quickly disappeared along with political fragmentation. Gold returned in common use from the mid-13th century. By the 14th century, most mints in western Europe operated along similar lines, with more or less unrestricted coinage on demand provided by profit-making mints. A great multiplicity of monetary systems persisted, giving rise to both foreign exchange markets (the earliest financial markets) and money changers (the first financial intermediaries).

The first instances of token coinage (coins that are intrinsically useless but are claims to fixed amounts of the commodity) appeared in the 15th century in Catalonia. Notes convertible on demand appeared in

the 17th century, in Sweden and later in England. For a more complete discussion of medieval European coinage, see Spufford (1988).

Coins appear to have been used in India in the early fourth century bc and were probably used before then. The earliest coins were so-called punch-marked coins and were adaptations of Greek prototypes. Coins were first used in China and the Far East about the same time as in India. The distinctive bronze coinage with the square hole in the middle first appeared in the third century bc. Early coins in eastern Islamic lands were copies of Byzantine gold and bronze coins; those in the East were copies of Sassanian silver coins. For more on coinage in India and the Far East, see Williams (1997).

Until the 19th century, coins typically bore no indication of face value, and their market value could fluctuate even relative to one another. From the late Middle Ages, governments increasingly sought to regulate the value of coins in some manner, in particular assigning face value or legal tender value by decree. It became desirable to turn the collection of objects used as a medium of exchange into a stable system with fixed exchange rates between the objects. This was achieved to a large degree with bimetallism, a system in which gold and silver coins remained concurrently in circulation at a constant relative price. Its heyday was the mid-19th century, but beginning in 1873 the system was quickly abandoned, and by the First World War countries were using either gold only or (in Africa and eastern Asia) silver only. (Bimetallism is discussed in more detail in Redish, 2000, and Velde and Weber, 2000.)

The development of banking in the 19th century also led to increased use of (convertible) notes and other monetary instruments.

The First World War brought about the suspension of convertibility of the notes in many countries. Most countries returned to convertibility between 1926 and 1931, but the onset of the Great Depression reversed the movement. After the Second World War the only major country whose currency was in any way directly tied to a commodity was the United States under the Bretton Woods system: dollars were convertible by non-residents of the United States into gold on demand, while other currencies of the system were convertible into dollars. The link between gold and the dollar was severed in 1971. Fiat money standards are now universal.

The nature of commodity money

The definitions of commodity and fiat monies given above make it seem as if there is a clear distinction between the two. It is more helpful, however, to think of media of exchange along a continuum. An object serving a purpose as a medium of exchange has value above its intrinsic content, reflecting the value of the service as a medium of exchange.

Because the value of a commodity qua commodity and the value as a medium of exchange can differ, the value of all commodity monies has a fiat component. A pure fiat money is one for which this fiat component makes up its entire value. A nice theoretical discussion of commodity and fiat monies is given by Sargent and Wallace (1983).

Price-level determination

It is natural that the medium of exchange in an economy is what becomes the unit of account, the unit in which debt contracts and the prices of goods and services are expressed. It is natural because the money

appears on one side of virtually every transaction.

Because commodity money has an intrinsic value apart from that which it obtains by being a medium of exchange, its relative price will not be zero. Thus, in a commodity money economy, the value of money (the inverse of the price level) is bounded away from zero. Moreover, in a canonical commodity money system (see below) with unlimited minting at a set price, the value of money and its quantity tend to remain within a band. If the value of money falls far enough, it becomes preferable to exercise the option and convert some of it into other, non-monetary uses, thus reducing the quantity and preventing the value from falling further. Conversely, if the value of money rises high enough, it becomes worthwhile for agents to turn metal into coins at the mint at the set price, thus increasing the quantity of money. Such a self-regulating commodity money system provides an anchor to the price level. This has been touted as one of the advantages of a commodity money system, particularly in the case of the gold standard.

The question of price-level determination becomes more complicated when multiple commodity monies are made out of different commodities. An example is the circulation of full-bodied gold and silver coins. Should the unit of account be the gold coin or the silver coin? This matters because under a commodity money system a monetary authority does not have the ability to set the exchange rate between monies of different commodities forever. Thus, to the extent that the unit of account is used in contracts to determine the amount of future payments, the choice of the unit of account can affect the allocation of goods and services. This was one of the issues surrounding the possible adoption of a bimetallic standard mentioned above.

The inability of the monetary authority to set the exchange rate between different monies goes away under a pure fiat money system. Because fiat money is (virtually) costless to produce, the monetary authority can costlessly exchange one money for another to maintain whatever exchange rate is desired between different monies that it issues.

Monetary policy

The fact that a commodity is used as money alters its value. This is because part of the total quantity of the commodity – namely, the metal locked up in the form of coins, or the reserves held by the monetary authority – is not available for non-monetary uses. The allocation between monetary and non-monetary uses is determined in equilibrium. Restrictions on the ability to change this allocation, such as restrictions on melting or exporting coins, or limitations on the minting of metal, will have an effect on the equilibrium value of the money even if it has no immediate effect on the allocation itself. (Since money is an asset, its valuation is forward looking.) Thus, there is scope for monetary policy under a commodity money standard, although what constitutes monetary policy is different from and more limited in scope than what holds under a pure fiat money standard.

Monetary policy consists in actions that tend to alter the value of money. In a commodity money system, the value of money is the value of the option we have described. (The strike price of the option is zero, since the commodity is the money.) Most aspects of monetary policy with commodity money consist in modifying this option, typically by modifying the institutions governing the exercise of the option rather than by modifying the quantity of money, which the authority usually cannot control directly. When the monetary authority is directly involved in the provision of the money, it may directly profit from its actions. Potential profit is often an important consideration of monetary policy.

The canonical form of a commodity money standard comprises the following. One or more commodities are chosen to be the standard to which the monetary system will be anchored. The monetary authority defines the specifications of the monetary objects (weight, fineness) and defines the unit of account in terms of these monetary objects. The conversion of commodity into commodity money and vice versa is as costless as possible. In particular, the monetary authority provides for unlimited (and even costless) conversion of the commodity into monetary objects (coins or notes). Conversely, it places no hindrances on the conversion of monetary objects into commodities (coins can be melted, notes are convertible on demand), nor does it place limitations on the consumption of the commodity or its service flow (free possession, unrestricted import and export of the commodity). The monetary objects are unlimited legal tender.

One type of monetary policy modifies the specifications of monetary objects and units of account. An example is debasement, which is reducing the commodity content of a monetary object (and, frequently, of the corresponding unit of account). The result of debasement is inflation, since nominal prices will be adjusted to maintain the relative prices of goods and money. And, just as occurs with fiat money, inflation has the effect of transferring wealth from nominal creditors to nominal debtors. Since governments generally tended to be debtors, debasements were used to reduce the amount of their debts. Historically, debasements also had the secondary effect of increasing seigniorage revenue, since the quantity of coins minted tended to increase significantly after debasements that involved the introduction of new coins (see Rolnick, Velde and Weber, 1996; Sargent and Smith, 1997). Debasements were also used by governments to remedy malfunctions of a multiple-denomination commodity money system (see Sargent and Velde, 2002).

A second type of monetary policy adds or modifies restrictions on the conversion of commodity into money or money into commodity. For example, minting might be restricted by quantity, in which case the authority decides how much to mint. Minting might be unlimited but subject to a fee, called seigniorage. Governments typically charged such a fee, both to cover the actual costs of minting (called brassage) and as a tax (England was the first, in 1666, to provide minting at no cost). The rate of this tax or, equivalently, the price paid by the mint for bullion might be changed. These restrictions tended to alter the allocation of the commodity between monetary and non-monetary uses, and hence the value of the commodity and the money.

A third type of monetary policy sets limits to the legal tender quality of certain coins, or changes their legal tender value. Since coins did not have face values until the 19th century, it was up to monetary authorities to set, and from time to time alter, the legal tender values of coins. Frequently, foreign coins were authorized as legal tender at rates set for domestic coins. Countries attempting to maintain bimetallism in the face of fluctuations in the relative price of gold and silver often had to adjust the face value of either their gold or silver coins. Changes in the legal tender values could also be motivated by fiscal considerations or by attempts to target a particular price level or exchange rate.

The physical nature of the medium of exchange led to a particular set of concerns. Coins, like anything else, depreciate with use, through wear and tear. Since coins of different values have different usage rates, the depreciation rate varied by denomination. Also, being roughly constant over time, depreciation depended on the age of the coin. Finally, imperfect minting technology as well as actions by the public (clipping, sweating) aggravated the disparities between coins. This factor introduced heterogeneity among coins and hindered the achievement of a stable and uniform monetary system. Improvements in coin production partially remedied the problem, as did periodic recoinages.

When the monetary objects consist not only of coins but also of paper currency or tokens that are demand promises to the commodity, a fourth type of monetary policy is available: suspension of convertibility. The monetary authority can refuse to honour the promise of convertibility for some period of time. An example is the suspension of convertibility by the Bank of England between 1797 and 1819 during the wars with France. During the 19th century suspensions were not uncommon during financial or fiscal emergencies, with the understanding that the suspension would end after the emergency and convertibility would be restored at the pre-existing parity. This understanding has been described as a state-contingent gold standard (see Bordo and Kydland, 1996).

When there is a central bank, an additional monetary tool is to change the discount rate, the interest rate at which the central bank lends reserves to the banking system. During the gold standard period, this was the primary means by which central banks affected the exchange rate of their money against the monies of other countries.

Conclusion

Commodity money is a thing of the past; countries worldwide now use fiat money standards. This practice has led to an efficiency gain in the sense that resources that were once tied up in coins are now available for consumption and production (perhaps prompting John Maynard Keynes to refer to gold as the ‘barbarous relic’). It has also led to a greater scope for monetary policy because the supply of money can be changed almost costlessly. However, along with this greater scope has come the greater potential for governments to use inflation to collect seigniorage revenue or to reduce the real value of their debts. How to use the freedom that commodity money restricted is still a matter of debate.

See Also

- bimetallism
- Bretton Woods system
- fiat money
- gold standard

Bibliography

Bordo, M. and Kydland, F. 1996. The gold standard as a commitment mechanism. In *Modern Perspectives on the Gold Standard*, ed. T. Bayoumi, B. Eichengreen and M. Taylor. Cambridge: Cambridge University Press.

Kiyotaki, N. and Wright, R. 1989. On money as a medium of exchange. *Journal of Political Economy* 97, 927–54.

Luschin von Ebengreuth, A. 1926. *Allgemeine Münzkunde und Geldgeschichte des Mittelalters und der neueren Zeit*. Munich: R. Oldenbourg.

Redish, A. 2000. *Bimetallism: An Economic and Historical Analysis*. Cambridge: Cambridge University Press.

Rolnick, A., Velde, F. and Weber, W. 1996. The debasement puzzle: an essay on medieval monetary history. *Journal of Economic History* 56, 789–808.

Sargent, T. and Smith, B. 1997. Coinage, debasements, and Gresham's laws. *Economic Theory* 10, 197–226.

Sargent, T. and Velde, F. 2002. *The Big Problem of Small Change*. Princeton, NJ: Princeton University Press.

Sargent, T. and Wallace, N. 1983. A model of commodity money. *Journal of Monetary Economics* 12, 163–87.

Spufford, P. 1988. *Money and Its Use in Medieval Europe*. Cambridge: Cambridge University Press.

Sussman, N. and Zeira, J. 2003. Commodity money inflation: theory and evidence from France in 1350–1436. *Journal of Monetary Economics* 50, 1769–93.

Velde, F. and Weber, W. 2000. A model of bimetallism. *Journal of Political Economy* 108, 1210–34.

Williams, J., ed. 1997. *Money: A History*. New York: St Martin's Press.

How to cite this article

Velde, François R. and Warren E. Weber. "commodity money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000235>
doi:10.1057/9780230226203.0268

The New Palgrave Dictionary of Economics Online

common factors

Heather M. Anderson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article outlines and illustrates several types of common factor models that are found in the applied economics literature. These factor models include those based on principal components, classical factor analysis, dynamic factor analysis and common features, and the discussion addresses the identification and estimation of factors, as well as the use of common factor models.

Keywords

ARMA processes; autoregressive moving average (ARMA) processes; canonical correlations; capital asset pricing model; classical factor models; coincident indices; common factors; common feature models; common trend model; diffusion indexes; dynamic factor (or index) models; factor analysis; Kalman filter; principal component analysis; real business cycle models; reduced rank regressions; static factor models; Stone, J.; term structure of interest rates; time series analysis

Article

Economic analysis frequently involves the study of variables that exhibit similar behaviour, and it is often of interest to model this comovement. Well-known examples of comovement in multivariate data sets include business cycles in macroeconomic indicators and shifts in the entire term structure of interest rates, and researchers sometimes attribute this comovement to a small set of underlying forces or latent ‘factors’ that influence each variable in the system. It is then convenient to think of the variation in each variable in the system as the sum of two types of (unobserved) components, one of which captures variation that is due to ‘common factors’, while the other captures all other variation. Models that attribute comovement to common factors are called common factor models, and common factor analysis involves the identification and study of the common factors.
Common factor models are particularly popular in empirical settings because they offer parsimony, and simplify estimation by reducing the number of parameters that need to be estimated. Economists will

typically be interested in interpreting common factors so that they can explain why comovement occurs. Economic theory sometimes predicts common factors. Perhaps the best-known example of this is the capital asset pricing model, in which the (excess) return for the market portfolio is the common factor in the (excess) return for each individual stock. Another well-known example arises when the term structure of interest rates is modelled, because the no arbitrage condition implies that the entire term structure is determined by a single factor, which is the instantaneous interest rate.

A simple model that captures the concept of common factors in a set of *N* time-series in the (demeaned) vector $Y_t = (Y_{1t}, Y_{2t}, \ldots, Y_{Nt})'$ is given by

$$Y_t = AF_t + \varepsilon_t, \tag{1}$$

where F_t is an $r \times 1$ vector that contains *r* common factors, *A* is an $N \times r$ factor loading matrix (with $rank(A) = r < N$), and ε_t contains *N* idiosyncratic components. With the use of Σ_Y, Σ_F and Σ_ε to denote the variance covariance matrices of Y_t, F_t and ε_t, it is usual to assume that Σ_ε is diagonal, and it is also common to normalize the set of *r* factors in F_t by assuming that $\Sigma_F = I_r$.

Model (1) is similar to conventional factor models that are often used in cross-sectional settings, although the variables are specified here as time series, to facilitate discussion on dynamic factor models. If there is no serial correlation in Y_t or F_t, or if estimation is undertaken as if this is the case, then (1) is called a static factor model. It is usual to assume that F_t and ε_t are jointly stationary, that $E(\varepsilon_t) = 0$, $E(F_t\varepsilon_t') = 0$, and that ε_t contains no serial dependence, but these latter assumptions can be relaxed, depending on the type of factor model under consideration.

There are many ways to identify the factors in (1), and standard techniques include the use of principal component analysis, factor analysis and canonical correlations to estimate the parameters of various associated reduced rank regressions. More recently, researchers have focused on the time series properties of multivariate data-sets, and modern factor models include dynamic factor (or index) models, and models that incorporate common features. These latter models incorporate various ways of allowing the factors to follow specific dynamic processes, or to contain specific time-series properties.

Principal component models

Principal component analysis involves the intuition that most of the variance in Y_t will be attributable to variance in the *r* components in F_t. The factors F_t are modelled as linear combinations of Y_t, and their identification is based on finding the *r* (orthogonal) linear combinations of Y_t that have the most variance. In practice this involves finding the eigen values $\lambda_1 > \lambda_2 > \ldots > \lambda_N$ and associated eigenvectors $f_1, \ldots, f_N$ of the form $f_i = \beta_i' Y_t$ that are associated with the roots of the equation

$|\hat{\Sigma}_Y - \lambda I| = 0$, where $\hat{\Sigma}_Y$ is an estimate of Σ_Y. The β_i are picked so that $(\hat{\Sigma}_Y - \lambda_i I)\beta_i = 0$ and $\beta_i'\beta_i = 1$, and the factors F_t are then defined by $F_t = (f_1, \ldots, f_r)$. This decomposition ensures that $E(F_t\varepsilon_t') = 0$, but it implies that the ε_t (which are each linear combinations of $(f_{r+1}, \ldots, f_N)$) will be correlated with each other so that Σ_ε will not be diagonal. Principal components estimators of common factors and factor loadings are also the least squares estimators of the reduced rank regression given by

$$Y_t = \underset{(N\times r)}{A}\,\underset{(r\times N)}{B}\,Y_t + \varepsilon_t, \qquad (2)$$

where BY_t contains the r factors $F_t = (f_1, \ldots, f_r)$. Anderson (1984, ch. 11) provides a standard reference.

In practice, one needs to determine r before estimating common factor models, and this is often based on the ratio given by $\dfrac{\lambda_{r+1}^2 + \ldots + \lambda_N^2}{\lambda_1^2 + \ldots + \lambda_N^2}$. This ratio measures the loss of information in the reduced rank system relative to an unrestricted system, and typically investigators will choose r so that this ratio is kept small. Bai and Ng (2002) have developed model selection criteria that are consistent as $\{N, T\} \to \infty$.

Principal components are usually used for dimension reduction, and economic interpretation of the resulting factors is rarely straightforward. However, Stone (1947) has summarized a set of series from the US national accounts, associating the first three principal components with income, income growth and time and Chamberlain and Rothschild (1983) have promoted the use of principal components for estimating approximate factor models of asset prices. Stock and Watson (2002) have suggested the use of diffusion indexes (principal component factors associated with large macroeconomic data sets) for forecasting key macroeconomic variables, and the interest here centres on using information in the factors rather than interpreting the factors themselves.

Classical factor models

Classical factor models are closely related to principal components models, but the underlying intuition and assumptions are different. In this case the key assumption is that Σ_ε is diagonal so that the ε_t describe idiosyncratic effects that are unique to each variable in Y_t, while the factors describe joint effects in Y_t. The assumptions that $E(\varepsilon_t) = 0$ and $E(F_t\varepsilon_t') = 0$ still hold, and the ε_t are assumed to contain no serial dependence. Under these assumptions $\Sigma_Y = A\Sigma_F A' + \Sigma_\varepsilon$, and estimates for A, Σ_F and Σ_ε can be found by maximizing the function

$$L_T(A, \Sigma_\varepsilon) = -\frac{T}{2}\ln|\Sigma_Y| - \frac{1}{2}\sum_{t=1}^{t=T} Y_t'\Sigma_Y^{-1}Y_t$$

subject to the condition that $rank(A) = r$ and a set of normalization restrictions that will uniquely identify the $(r+1)(N+\frac{1}{2}r)$ parameters. Researchers often use the joint restrictions that $\Sigma_F = I_r$ and that $A'\Sigma_\varepsilon^{-1}A$ is diagonal for normalization, but other normalizations are common (see Anderson, 1984, ch. 14, for details). If Y_t and ε_t are normally distributed then $L_T(A, \Sigma_\varepsilon)$ is the log likelihood for Y_t (if we ignore the constant term), but, even when Y_t and ε_t are not normally distributed, the maximization of $L_T(A, \Sigma_\varepsilon)$ delivers quasi-maximum likelihood estimates. There are several ways of using the estimates of A and Σ_ε to obtain estimates of the factors in F_t, and perhaps the best-known of these is Bartlett's (1937; 1938) method based on generalized least squares given by

$$\hat{F}_t = (\hat{A}'\hat{\Sigma}_\varepsilon^{-1}\hat{A})^{-1}\hat{A}'\hat{\Sigma}_\varepsilon^{-1}Y_t.$$

As above, it is necessary to determine r prior to estimating the factors, and, on the assumption of normality, the likelihood ratio test statistic for testing $H_0: r = s$ versus $H_A: r > s$ is given by

$$-T\left[\ln\left|\hat{\Sigma}_Y\right| - \ln\left|\hat{A}\hat{A}' + \hat{\Sigma}_\varepsilon\right|\right] = -T\,\Sigma_{i=s+1}^{i=N}\ln(1-\hat{\lambda}_i^2),$$

where the $\hat{\lambda}_i$ are the characteristic roots of $\hat{A}'\hat{\Sigma}_\varepsilon^{-1}\hat{A}$ (in decreasing order) and $\hat{A}$ is estimated under the null. The test statistic is asymptotically distributed as a χ_q^2 with $q = [(N-s)^2 - N - s]/2$ degrees of freedom under the null.

There are numerous applications of classical factor analysis to economic problems, and an early example includes Stone's (1945) factor analysis of the demand for N commodities. Another example includes a factor model of returns by Deistler and Hamann (2005).

Dynamic factor models

Classical factor models are not well suited to multivariate analysis of time series because they assume no

serial correlation in ε_t, and, if there are any dynamics in F_t, then they are implicit and not explicitly modelled. Dynamic factor models address these concerns by treating the ε_t and F_t as autoregressive moving average (ARMA) processes. The innovations that underlie the N processes for ε_t are assumed to be mutually uncorrelated, and uncorrelated with the innovations that underlie the F_t at all leads and lags, but the factors themselves can be mutually correlated. Different variables in Y_t can then move together because they are functions of the same factor(s), or because they are functions of different factors that are themselves correlated.

The identification and estimation of small-scale dynamic factor models is sometimes based on spectral techniques (see Geweke, 1977; or Sargent and Sims, 1977), and use of the Kalman filter in the time domain (as in Engle and Watson, 1981, or Harvey and Koopman, 1997) provides an alternative approach. Dynamic factor models have been particularly popular for estimating factor models of business cycles (as in Geweke and Singleton, 1981), but they have also been used for studying the term structure (Singleton, 1980) and fluctuations in employment across different industrial sectors (Quah and Sargent, 1993).

Recent work has shifted towards the identification and estimation of common factors in large-scale models, relying on the use of large N to obtain consistent estimates of the factors. One strand of this literature adopts a static framework and standard principal components to estimate the factors, and then builds dynamic models of the factors. The resulting models are sometimes called approximate dynamic factor models. Applications of this approach include Stock and Watson's diffusion index (2002), and Bernanke and Boivin's (2003) estimation of a monetary policy reaction function. Another strand of this literature allows different variables to depend on different lags of common factors. These ‘generalized dynamic factor models’ are estimated using ‘dynamic principal components’, which are the principal components of spectral density matrices at different frequencies. Applications of this latter approach include a study of business cycle dynamics in the United States (Forni and Reichlin, 1998) and the development of a coincident index for Europe (Forni et al., 2000).

Canonical correlation-based models

Principal component and factor models assume that the factors are linear combinations of the N variables in Y_t, but sometimes it is useful to assume that the factors are linear combinations of M variables contained in another multivariate time series denoted by X_t. The variables in X_t will often include lags of the variables in Y_t, but X_t can also include variables that would be classified as explanatory variables in a regression context. The factors in (1) can now be written in the form $F_t = B'X_t$ (where $rank(B) = r < \min\{N, M\}$). In what follows, we assume that the ε_t in (1) are white noise and uncorrelated with X_t.

The main idea behind a canonical correlations approach is to find linear combinations of X_t that are strongly correlated with linear combinations of Y_t, and, as for principal component models, the estimators of common factors and factor loadings are the least squares estimates of a reduced rank regression. In this case the regression is

$$Y_t = \underset{(N \times r)}{A} \underset{(r \times M)}{B} X_t + \varepsilon_t, \tag{3}$$

and the factors and factor loadings are related to the r largest roots of $R = \Sigma_Y^{-\frac{1}{2}} \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}}$, which is the multivariate generalization of the (squared) correlation coefficient between two variables. If we order these roots (also called squared canonical correlations) so that $\lambda_1^2 > \lambda_2^2 > \ldots > \lambda_r^2$, and let $V_1, V_2, \ldots V_r$ be the r associated eigenvectors, then the factor loadings and factors are given by $A_i = \Sigma_Y^{-\frac{1}{2}} V_i$ and $B_i X_t = \Sigma_Y^{-\frac{1}{2}} \Sigma_{YX} \Sigma_X^{-1} V_i X_t$. Anderson (1984, ch. 12) provides a detailed discussion of canonical correlations, while Izenman (1980) discusses the associated reduced rank regressions. When the variables in X_t are simply lags of the variables in Y_t, then the first factor is the best predictor of Y_t based on past history, the second factor is the next best predictor, and so on, and the factors provide a set of leading indicators for Y_t. When X_t consists of explanatory variables for Y_t, then the factors are often called coincident indices. One can base a test of $H_0: r = s$ versus $H_A: r > s$ on the test statistic $-T\Sigma_{i=s+1}^{i=N} \ln(1 - \hat{\lambda}_i^2)$, which has a χ^2 distribution with $(m - s)(n - s)$ degrees of freedom under the null.

Common feature models

Common feature models are a special class of factor models in which the common factors have a statistical characteristic of interest, while the idiosyncratic components fail to have this characteristic. Common features were first introduced by Engle and Kozicki (1993) when they discussed serial correlation features – a situation in which each of N variables is serially correlated, but there are linear combinations that are white noise. Here, the presence of $N - r$ white noise linear combinations implies a factor model in which there are r serially correlated factors (which are sometimes called common cycles). An earlier example of a common feature model is Stock and Watson's (1988) common trend model which is valid when variables are cointegrated (as in Engle and Granger, 1987). In this case the common factors are integrated of order one but the remaining components (often called error correction terms) are stationary. Other examples of common features include Vahid and Engle's (1993) common trend–common cycle representation, and common nonlinearity (Anderson and Vahid, 1998).
The identification of common features involves finding linear combinations of the data that do not have the feature, and this can be done using a canonical correlations approach in which the variables in X_t model the characteristic of interest. To illustrate, lags of Y_t are put into X_t when testing for serial correlation features in Y_t, and lagged levels are included in X_t when testing for common trends. Factors associated with the lowest eigen values define linear combinations that do not contain the feature, while

factors associated with the highest eigen values are used to model the common features. Johansen's (1988) procedure provides a well-known example of this, although inference in this case is based on non-standard (rather than χ^2) distributions because the factors are non-stationary.

A well-known example of a common feature model is the real business cycle model of King, Plosser and Rebelo (1988), in which a common factor (productivity) generates the trend in output consumption and investment, and another factor (the deviation of capital stock from steady state) generates the common cycle.

See Also

- reduced rank regression
- time series analysis

Bibliography

Anderson, H. and Vahid, F. 1998. Testing multiple equation systems for common nonlinear components. *Journal of Econometrics* 84, 1–36.

Anderson, T. 1984. *An Introduction to Multivariate Statistical Analysis*, 2nd edn. New York: Wiley.

Bai, J. and Ng, S. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.

Bartlett, M. 1937. The statistical conception of mental factors. *British Journal of Psychology* 28, 97–104.

Bartlett, M. 1938. Methods of estimating mental factors. *Nature* 141, 609–10.

Bernanke, B. and Boivin, J. 2003. Monetary policy in a data rich environment. *Journal of Monetary Economics* 50, 525–46.

Chamberlain, G. and Rothschild, M. 1983. Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1305–24.

Deistler, M. and Hamann, E. 2005. Identification of factor models for forecasting returns. *Journal of Financial Econometrics* 3, 256–81.

Engle, R. and Granger, C. 1987. Cointegration and error correction representation, estimation and testing. *Econometrica* 55, 251–76.

Engle, R. and Kozicki, S. 1993. Testing for common features (with discussions). *Journal of Business*

and Economic Statistics 11, 369–95.

Engle, R. and Watson, M. 1981. A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association* 76, 774–81.

Forni, M. and Reichlin, L. 1998. Let's get real: a factor-analytic approach to disaggregated business cycle dynamics. *Review of Economic Studies* 65, 453–73.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. 2000. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* 82, 540–54.

Geweke, J. 1977. The dynamic factor analysis of economic times-series models. In *Latent Variables in Socioeconomic Models*, ed. D. Aigner and A. Goldberger. Amsterdam: North-Holland.

Geweke, J. and Singleton, K. 1981. Maximum likelihood ‘confirmatory’ factor analysis of economic time series. *International Economic Review* 22, 133–7.

Harvey, A. and Koopman, S. 1997. Multivariate structural time series models. In *Systematic Dynamics in Econometric and Financial Models*, ed. C. Heij, H. Schumacher and C. Praagman. Chichester: Wiley and Sons.

Izenman, A. 1980. Assessing dimensionality in multivariate regression. In *Handbook of Statistics*, vol. 1, ed. P. Krishnaiah. Amsterdam: North-Holland.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–54.

King, R., Plosser, C. and Rebelo, S. 1988. Production, growth and business cycles II. New directions. *Journal of Monetary Economics* 21, 309–41.

Quah, D. and Sargent, T. 1993. A dynamic index model for large cross sections. In *Business Cycles, Indicators and Forecasting*, ed. J. Stock and M. Watson. Chicago: University of Chicago Press.

Sargent, T. and Sims, C. 1977. Business cycle modeling without pretending to have too much a-priori economic theory. In *New Methods in Business Cycle Research*, ed. C. Sims et al. Minneapolis: Federal Reserve Bank of Minneapolis.

Singleton, K. 1980. A latent time series model of the cyclical behavior of interest rates. *International Economic Review* 21, 559–75.

Stock, J. and Watson, M. 1988. Testing for common trends. *Journal of the American Statistical*

Association 83, 1097–107.

Stock, J. and Watson, M. 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20, 147–62.

Stone, J. 1945. The analysis of market demand. *Journal of the Royal Statistical Society, Series A*, 108, 286–382.

Stone, J. 1947. On the interdependence of blocks of transactions. *Journal of the Royal Statistical Society, Series B*, Supplement, 1–45.

Vahid, F. and Engle, R. 1993. Common trends and common cycles. *Journal of Applied Econometrics* 8, 341–60.

How to cite this article

Anderson, Heather M. "common factors." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000565> doi:10.1057/9780230226203.0269

The New Palgrave Dictionary of Economics Online

common property resources

Jean-Philippe Platteau
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

common property resources; efficiency; market integration; open access; private property rights; tragedy of the commons

Article

The concept of common property has become famous in economics since Garett Hardin (1968) wrote his celebrated article on ‘The Tragedy of the Commons’. In this article, common property is taken to mean the absence of property rights in a resource, or what is equivalently known as a regime of ‘open access’. Under such a regime, where a right of inclusion is granted to anyone who wants to use the resource, Hardin argued, inefficiency inevitably arises in the form of over-exploitation of the resource accompanied by an over-application of the variable inputs. Open access leads to efficiency losses because ‘the *average product* of the variable input, not its *marginal product*, is equated to the input's rental rate when access is free and the number of exploiters is large’ (Cornes and Sandler, 1983, p. 787). The root of the problem lies in the fact that the average product rule does not enable the users to internalize the external cost which their decisions impose on the users already operating in the resource domain. Of course, the efficiency losses are conceivable only in a world of resource scarcity, implying that the variable input is subject to decreasing returns. Such losses are considerable since they amount to the dissipation of the whole resource rent. Here is the crucial intuition behind the open access regime: when no property right is attached to a resource, the value of this resource is zero in spite of its scarcity. Efficiency losses are to be measured not only in static but also in dynamic terms. Indeed, in an open access regime resource users are induced to compare average *instantaneous* returns with the input's rental price even though they may well be aware that they thereby contribute to reducing the future stock of the resource. The problem is simply that they are forced to follow a myopic rule because there is no way in which they can reap the future benefits of restraint in the present. Thus, for example, by refraining today from catching juvenile fish or from cutting down saplings in the forest, a villager can receive no assurance that he or she will be able in the next period to catch mature fish or to fell fully grown trees.

The main criticism levelled by numerous social scientists against the concept of open access is that the corresponding regime is rarely encountered on the ground. The typical regime, according to these critiques, is one under which a community possesses a collective ownership right over local natural resources. Under common property, therefore, a right of exclusion is assigned to a well-defined user group, and Hardin has created a lot of confusion by using the word ‘commons’ to refer to the alternative situation where no such right is granted to any agency. What is not always clear, however, is whether the ownership right involves only the ability to specify the rightful claimants to the resource, or whether it also involves the ability to define and enforce rules of use regarding that resource (for example, regulations about the harvesting season and production tools, allowed quotas of harvestable products of the resource, or taxes). Baland and Platteau (1996) have coined the term ‘unregulated common property’ to refer to the former situation, while the term ‘regulated common property’ is used for the latter.

Two polar situations can be considered on the basis of this analytically important distinction between two types of common property regimes. At one extreme, if common property is perfectly regulated, in the sense that the rules of use designed and enforced by the owner community allow a perfect internalization of the externalities, common property becomes equivalent to private property with a sole owner from an efficiency standpoint. This illustrates the general result that, absent transaction costs, institutions do not matter. At the other extreme, a strictly unregulated common property in the above sense implies that, as the number of users becomes quite large, over-exploitation of the resource becomes as important as under the open access regime: the rent attached to the resource is totally dissipated (see Platteau, 2000, ch. 3).

Between these two extremes we find the situations most typically observed on the ground and described in the numerous field studies devoted to this topic (see Ostrom, 1990; Baland and Platteau, 1996, for a review of such studies). In such instances, rules of use exist alongside membership rules, yet they tend to be imperfectly designed and imperfectly enforced by the village community. One key reason for these imperfections is the governance costs that unavoidably plague any collective decision-making process. Governance costs include all those costs incurred to reach a collective agreement and to organize a community of users. They are likely to be higher when the group is larger and when its membership is more heterogeneous (whether measured in terms of diversity of objectives or of wealth inequality). Moreover, governance costs are enhanced by the opportunistic tendencies of rights-holders not only to violate or circumvent collective rules but also to eschew efforts to create collective mechanisms of decision-making and enforcement. Costs arising from these proclivities are also dependent on the size of the user group: they are lower if the number of resource users is smaller and, at the limit, they are nil when there is a single user.

As a consequence of the aforementioned limitations, resources are less efficiently managed under a common property regime than they could be under a private ownership system. This is especially true if, owing to their scarcity, the resources carry high values which should be reflected in high rents. Population growth and market integration are thus two forces that tend to increase the monetary value of the efficiency losses arising from common property, that is, the forgone rents. This, at least, is the conclusion drawn by the so-called property rights school of Chicago economists (see, for example, Demsetz, 1967; Barzel, 1989). The advantages of private property appear all the more decisive as such a regime enables users to internalize externalities without incurring any governance costs. This is because it establishes a one-to-one relationship between individual actions and all their effects: ‘A primary function of property rights is that of guiding incentives to achieve a greater internalization of

externalities …' (Demsetz, 1967, p. 348).

Nevertheless, this ignores the costs of privatizing natural resources, which involve both directs costs and opportunity costs. Direct costs comprise transaction costs, such as the costs of negotiating, defining and enforcing private property rights. The usual argument is that such costs increase with the physical base of the resource. Thus, the wider the resource base (or the less concentrated the resource) the higher are the costs of delimiting and defending the resource 'territory' (Dasgupta, 1993, pp. 288–9). For many natural resources, the costs of dividing the resource domain appear prohibitive under the present state of technology. For example, the open sea – or, more exactly, the fish stock contained in it – presents insuperable difficulties for private appropriation. The enforcement of exclusive property rights to individual patches of the ocean would, indeed, be infinitely costly. This is especially evident when fish species are mobile and move within wide water spaces, since exclusive rights are too costly to establish and enforce whether over the resource or over the territory in which the resource moves.

The opportunity costs of privatization, for their part, correspond to the benefits that are lost when the common property regime is abandoned. Here, we can think of scale economies that may be present not only in the resource itself but also in complementary factors. The obvious advantage of coordinating the herding of animals so as to economize on shepherd labour in extensive grazing activities is probably the best illustration of the way scale economies in a complementary factor may prevent the division of a resource domain. Another important category of opportunity costs is the insurance benefits associated with common property. When returns to a resource are highly variable across time and space, the need to insure against such variability is yet another consideration that may militate against resource division. When a resource has a low predictability (that is, when the variance in its value per unit of time per unit area is high), users are generally reluctant to divide it into smaller portions because they would thereby lose the insurance benefits provided by keeping the resource whole.

For instance, herders (fishermen) may need to have access to a wide portfolio of pasture lands (fishing spots) in so far as, at any given time, wide spatial variations in yields result from climatic or other environmental factors. On the assumption that the probability distributions are not correlated too much across spatial groupings of land or water and that they are not overly correlated over time, a system offering access to a large area within which right-holding users can freely move appears highly desirable from a risk-reducing perspective.

The conclusion of the above discussion is, therefore, that the balance of the advantages and disadvantages of various property regimes is a priori undetermined. Economic theory, however, does provide useful guidance about which circumstances are more favourable to the persistence of common property or, conversely, to its demise and replacement by private property. Furthermore, instead of being fixed once for all, the balance sheet is susceptible to evolution depending on the transformation of the parameters on which the benefits and costs of privatization depend. Thus, the direct costs of resource division may fall with technological progress. For example, the introduction of modern borehole drilling facilitates the privatization of common grazing areas (Peters, 1994). It is therefore not only the factors which enhance resource value but also those which reduce the direct costs of partitioning that may favour the private appropriation of natural resources.

See Also

- access to land and development
- agriculture and economic development
- land markets
- population and agricultural growth
- property law, economics and
- tragedy of the commons

Bibliography

Baland, J. and Platteau, J.-P. 1996. *Halting Degradation of Natural Resources: Is There a Role for Rural Communities?* Oxford: Clarendon Press.

Barzel, Y. 1989. *Economic Analysis of Property Rights*. Cambridge: Cambridge University Press.

Cornes, R. and Sandler, T. 1983. On commons and tragedies. *American Economic Review* 83, 787–92.

Dasgupta, P. 1993. *An Inquiry into Well-Being and Destitution*. Oxford: Clarendon Press.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57, 347–59.

Hardin, G. 1968. The tragedy of the commons. *Science* 162, 1243–8.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Peters, P. 1994. *Dividing the Commons: Politics, Policy, and Culture in Botswana*. Charlottesville and London: University Press of Virginia.

Platteau, J.-P. 2000. *Institutions, Social Norms and Economic Development*. London: Harwood Academic Publishers.

How to cite this article

Platteau, Jean-Philippe. "common property resources." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000531> doi:10.1057/9780230226203.0270

The New Palgrave Dictionary of Economics Online

common rights in land

Leigh Shaw-Taylor
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Common rights’ and ‘common land’ refer to rights to use land in common in some way. Of several forms of common rights in pre-industrial Europe and elsewhere, only one – free access to land – involved what economists commonly think of as common rights. Common rights in Europe were largely swept away during the 18th and 19th centuries by a process termed ‘enclosure’. Some economic historians have reconsidered the inefficiency of open fields in an English context, but at present the data are too poor to allow a plausible rebuttal of the views of 18th-century critics of the open fields.

Keywords

access to land; common fields; common land; common property resources; common rights; common rights in land; common-pool resources; enclosure; open-access resources; tragedy of the commons

Article

Common rights are rights to use land in common. The most important of these rights was the right to graze livestock on common grassland. But rights to gather fuel (wood, peat, gorse and turves), fertilizers, timber for building and other natural resources were also important. Common land is land used by a number of distinct individuals or households whose rights over the land are known as common rights.

Today we are accustomed to think of land as private property with a clear owner and possibly a tenant. Although in some countries there may be legal rights of public access to certain types of wild or agricultural land, it is generally the case that the owner or tenant of the land has exclusive rights to use the land and, within the limits of planning or zoning laws, may use it as he or she wishes. But in Europe, for at least a thousand years and ending only in the 19th century, a high proportion of land was ‘common land’ which many individuals were entitled to use for a variety of purposes.

It cannot be overemphasized that common land was generally not open-access land – land which anyone could use. There were regulations governing who could use the land, what they could use it for and how

much they could use it. When economists think of common land and common rights they may have Garett Hardin's ‘tragedy of the commons in mind’ (Hardin, 1968). The principal subject of Hardin's article was in fact population growth, not historical common land or common rights. However, Hardin used a theoretical common land system as a model for the exploitation of open-access resources. In this system each herder could put as many animals as he wished on to the common pastures. Hardin argued that individual herders would choose to graze more and more animals on the common, thus inevitably leading to over-grazing and degradation of the resource. This model offers important insights into the destruction of, or damage to, unregulated open-access resources such as the atmosphere or fish stocks in the oceans. If common land and common rights had operated in this manner, it is unlikely that they would have remained a key part of European agriculture for so many centuries.

In the rest of this article the following questions are addressed: what were common rights? What was common land? Who had common rights? How was common land regulated? Was it efficient? How and why did it come to an end and with what consequences? The answers to these questions varied from one village to another across Europe and what follows is necessarily highly simplified (see de Moor, Shaw-Taylor and Warde, 2002, for a more detailed overview).

Common land

The types of common land and the terminology used to describe such land varied across Europe. Nevertheless, four major types of common land may be distinguished. First, the archetypical form of common land and the one with the widest geographical distribution is variously referred to as common waste, common pasture, waste, or common. This land was permanently common and most often grassland used for grazing animals. Usually such land was not suitable for arable cultivation typically because its natural fertility was low but sometimes for other reasons such a propensity to seasonal flooding. On some common wastes other resources were available, such as peat, turf, gorse or wood.

Second, in many parts of Europe much of the arable land (the land on which crops were grown) was also subject to common rights. Such land, known as open-fields, common fields, or common arable, was privately owned and cultivated but subject to common grazing. In its classic form each farmer held a number of long thin strips of land scattered over an extensive area and intermixed with the strips of other farmers. Each farmer cultivated his own crops on the arable. But when the harvest was over, or in years when the land was being fallowed, all those with common rights could turn their livestock into the fields to graze. Thus the open fields alternated between private and common land over the course of the agricultural cycle.

Third, common woodland for the production of fuel and timber was widespread on the European continent but unusual in England. This was similar to common waste in that it was permanently in common use.

Fourth, common meadows, which were permanent grasslands for the production of hay, were divided into separate blocks in private use but after the hay had been harvested were open to common grazing. Thus, like the open fields, common meadows alternated between private land and common land over the agricultural cycle.

Common rights

As private property, the right to cultivate the common arable or to harvest the hay in common meadows lay with the owner of the land or the owner's tenant. Access to the common rights was considerably more complex and took different forms in different places; but it is possible to distinguish four main forms of access. First, in England and some parts of the Continent, the ownership or tenancy of particular buildings or landholdings was a prerequisite. Second, in many parts of the Continent citizenship of (as distinct from residence in) a commune or a municipality which itself owned the common resource was necessary, sometimes in combination with a property qualification. Third, in other parts of the Continent membership of a cooperative association which owned the common resources was necessary. Membership of these institutions was sometimes inherited, but sometimes it was attached to buildings or land (as in the first case). Fourth, there were cases were all residents in an area had common rights. But outside largely uninhabited areas, such as northern Sweden, this situation was unusual.

In consequence by no means all individuals or households enjoyed common rights. The proportion of the population that enjoyed common rights varied considerably from one region to another and changed over time. Where individuals or households did have common rights, the kinds and levels of the rights they enjoyed were determined by local regulations.

Regulation

Common land was almost invariably regulated by local institutions, often at the level of the individual village or manor. The institutions varied but were usually manor or village courts or village assemblies or committees of some kind, with the decisions made by a group of jurors. These institutions normally issued sets of rules, ordinances or by-laws which governed the usage of the commons and set fines for the infringement of rules. Officials or monitors were appointed to police the by-laws. The degree to which these institutions and their by-laws were subject to the influence of feudal overlords and the state varied considerably across Europe.

The by-laws provided the basic regulatory framework for managing the commons (for examples of by-laws see Ault, 1972). Their most critical function was to restrict the usage of common land and thus prevent a ‘tragedy of the commons’ developing. This was done in two ways. First, the by-laws would normally serve to restrict common rights to well-defined groups of users. For example, in much of England only those holding land in the open fields or with certain recognized dwellings, known as common-right houses, were allowed to pasture animals in the open fields or on the common pasture, while on much of the Continent pasture rights were restricted to citizens of communes or the holders of ancient farmsteads. Second, by-laws defined the amount of resources to which each commoner was entitled. Thus, by-laws might specify the amount of peat or wood each commoner was entitled to dig or cut each year or the number and type of animals which could be kept, and for which months of the year they might be kept on the common pastures, open fields and common meadows.

The number of animals each commoner could put on the common land was generally controlled by one of two types of rules. One, known as ‘stinting’, simply specified the number and type of animals (the stint) which each commoner might keep on the common. Often the stint was proportional to the area or the value of land held. The other form of access, known as ‘levancy’ and ‘couchancy’, stated merely that each commoner could keep as many animals on the common as he could overwinter (that is, feed when

the common was closed) on his own holding. How this was policed in practice is a moot point, but it certainly served to limit numbers and may have differed little from stinting in practice.
One consequence of these types of rule is that some individuals had no common rights at all. Another is that different individuals who did have common rights could have very different levels of access. The situation varied too much to allow generalization, beyond the suggestion that the level of inequality in England was probably greater and had proceeded further at an early date than anywhere else.

Enclosure

The process by which common land and common rights were abolished and replaced by recognizably modern forms of private property was part and parcel of a broader reform of landholding known as ‘enclosure’ which could also entail the consolidation of scattered holdings and the wholesale reallocation of land to create ring-fenced farms. Enclosure in some form is probably as old as common land itself. In England significant enclosure took place in the medieval period and from the 17th to the early 19th centuries. In most of Europe the widespread attack on common land began in the late 18th century in the wake of Physiocratic critiques. The later Napoleonic reforms and a subsequent series of state-sponsored drives to modernize agriculture in the 19th century led to more sustained enclosure. Some common land survives to this day, generally in mountainous areas.

Efficiency

By the 18th century common rights and common land were being widely criticized by agricultural improvers and others for restricting agricultural productivity. Most agricultural writers have accepted this view of common land as inefficient, and associated enclosure with major increases in productivity (Ernle, 1936; Chambers and Mingay, 1966; Overton, 1996). Common rights and common land imposed two kinds of limitation on agricultural improvement. First, the communal regulation of common land made it more difficult to introduce new agricultural techniques and technologies or to respond to changes in market opportunities. Second, the sharing of the outputs from common land made individual investment less attractive. The spread of nitrogen fixing crops and new drainage technologies, which often allowed the cultivation of formerly uncultivable common land, together with better transport links made enclosure a steadily more pressing issue in the 18th and 19th centuries.
A number of economic historians have reconsidered the inefficiency of open fields in an English context. McCloskey (1976) has argued that the scattering of land in open fields in the medieval period was an efficient insurance against risk in a non-market economy. Allen (1992) has argued that enclosure did facilitate major technological changes obstructed by common land but that these innovations made only very marginal contributions to increased efficiency. Clark (1998) has argued that the inefficiencies imposed by common land were relatively modest and that, given the costs involved, enclosure was not economic until after 1750. However, the issue remains controversial essentially because it is inherently difficult to measure the agricultural productivity of farming in the 18th and 19th centuries with any degree of reliability. In other words, at present the data are too poor to allow an entirely plausible rebuttal of the views of 18th-century critics of the open fields. Moreover, much enclosure took place in the medieval period and in the 17th century (Wordie, 1983) and any fully satisfactory theory of the

efficiency or otherwise of open fields would need to be able to account for the longer-term chronology of enclosure. The persistence of open-field farming in France has been investigated by Grantham (1980) and Hoffman (1989).

Another controversial issue is the importance of common land to the poor. Many historians have argued that the poor derived considerable benefits from common land and that enclosure was socially damaging; but this remains controversial (see Neeson, 1993; Shaw-Taylor, 2001. The extent to which the poor benefited from common land and common rights is hard to reconstruct, poorly understood, and varied considerably across Europe.

Common-pool resources

This article has been concerned exclusively with common land and common rights as they existed in Europe before the 20th century. However, it should be noted that while open fields and common meadow may be peculiarly European forms, common waste and institutions for its management can be found all over the world. Analytically, these systems are part of a larger family of ‘common-pool-resource’ systems (Ostrom, 1990) which have been adopted in many parts of the world to manage not just land but water resources and fish stocks as well.

See Also

- access to land and development
- common property resources
- tragedy of the commons

Bibliography

Allen, R. 1992. *Enclosure and the Yeoman: The Agricultural Development of the South Midlands, 1450–1850*. Oxford: Oxford University Press.

Ault, W. 1972. *Open-field-farming in Medieval England: A Study of Village By-laws*. London: George Allen and Unwin.

Chambers, J. and Mingay, G. 1966. *The Agricultural Revolution 1750–1880*. London: Batsford.

Clark, G. 1998. Commons sense: common property rights, efficiency, and institutional change. *Journal of Economic History* 58, 73–102.

de Moor, M., Shaw-Taylor, L. and Warde, P., eds. 2002. *The Management of Common Land in North West Europe, c. 1500–1850*. Turnhout: Brepols.

Ernle, Lord. 1936. *English Farming Past and Present*, 5th edn. London: Longmans, Green and Co.

Grantham, G. 1980. The persistence of open-field farming in nineteenth century France. *Journal of Economic History* 40, 515–31.

Hardin, G. 1968. The tragedy of the commons. *Science* 162, 1243–8.

Hoffman, P. 1989. Institutions and agriculture in old-regime France. *Journal of Institutional and Theoretical Economics* 145, 166–81.

McCloskey, D. 1976. English open fields as behaviour towards risk. *Research in Economic History* 1, 124–70.

Neeson, J. 1993. *Commoners, Common-Right, Enclosure and Social Change in England 1700–1820*. Cambridge: Cambridge University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Overton, M. 1996. *Agricultural Revolution in England: The Transformation of the Agrarian Economy 1500–1850*. Cambridge: Cambridge University Press.

Shaw-Taylor, L. 2001. Parliamentary enclosure and the emergence of an English agricultural proletariat. *Journal of Economic History* 61, 640–62.

Wordie, J. 1983. The chronology of English enclosure, 1500–1914. *Economic History Review* 36, 483–505.

How to cite this article

Shaw-Taylor, Leigh. "common rights in land." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000541> doi:10.1057/9780230226203.0271

The New Palgrave Dictionary of Economics Online

Commons, John Rogers (1862 – 1945)

Warren J. Samuels
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Association for Labor Legislation; American Economic Association; capitalism; collective action; Commons, J. R.; Fisher, I.; institutional economics; methodological individualism; National Bureau of Economic Research; New Deal; trade unions; Veblen, T.

Article

Commons was born on 13 October 1862 in Hollandsburg, Ohio, and died on 11 May 1945 in Raleigh, North Carolina. He studied at Oberlin College (BA, 1888) and Johns Hopkins University (1888–90). He taught at Wesleyan, Oberlin, Indiana, Syracuse, and Wisconsin (1904–32).

The founder of the distinctive Wisconsin tradition of institutional economics, Commons derived his theoretical insights (generalized in his *Legal Foundations of Capitalism*, 1924, and *Institutional Economics*, 1934) from his practical, historical and empirical studies, particularly in the field of labour relations and in various areas of social reform. He drew insight not only from economics but also from the fields of political science, law, sociology and history. A principal adviser and architect of the Wisconsin progressive movement under Robert M. La Follette, Commons was active as an advisor to both state and federal governments. He was instrumental in drafting landmark legislation in the fields of industrial relations, civil service, public utility regulation, workmen's compensation and unemployment insurance. He served on federal and state industrial commissions, was a founder of the American Association for Labor Legislation, was active in the National Civic Federation, National Consumers' League (president, 1923–35), National Bureau of Economic Research (associate director, 1920–28), and the American Economic Association (president, 1917). He participated in antitrust litigation (especially the Pittsburgh Plus case) and in movements for reform of the monetary and banking system (often associated with Irving Fisher, who considered Commons one of the leading monetary economists of the period).

The critical thread uniting Commons's diverse writings was the development of institutions, especially within capitalism. He developed theories of the evolution of capitalism and of institutional change as a modifying force alleviating the major defects of capitalism. Commons came to recognize and stress that

individual economic behaviour took place within institutions, which he defined as collective action in control, liberation, and expansion of individual action. The traditional methodologically individualist focus on individual buying and selling was not capable, in his view, of penetrating the forces, working rules and institutions governing the structural features of the economic system within which individuals operated. Crucial to the evolution and operation of the economic system was government, which was a principal means through which collective action and change were undertaken.

Commons rejected both classical harmonism and radical revolutionism in favour of a conflict and negotiational view of economic process. He accepted the reality of conflicting interests and sought realistic, evolutionary modes of their attenuation and resolution. These modes focused on a negotiational psychology in the context of a pluralist structure of power. He sought to enlist the open-minded and progressive leaders of business, labour and government in arrangements through which they could identify problems and design solutions acceptable to all parties.

In other contexts, he sought to use government as an agency for working out new arrangements to solve problems, such as worker insecurity and hardship, rather than promote systemic restructuring, although to many conservatives his ventures were radical enough. To these ends Commons and small armies of associates engaged in fact finding – his look-and-see methodology – in a spirit of bringing all scientific knowledge to bear on problem solving. From these experiences, indeed already manifest in the underlying strategy, Commons developed a theory of government as alternately a mediator of conflicting interests and an arena in which conflicting interests bargained over their differences; a theory of the complex organization – in terms of freedom, power and coercion – and evolution of the legal foundations of capitalism, which centred in part on the composing of major structural conflicts through the mutual accommodation of interests; and a theory of institutions with an affirmative view of their roles in organizing individual activity and resolving conflict.

The institutions Commons studied most closely were trade unions and government, particularly the judiciary. He developed his theory of the economic role of government in part on the basis of his study of the efforts of workers to improve their market position and in part on the use of government by both enemies and friends of labour. Commons's was an interpretation of trade unions as a non-revolutionary development, as collective action seeking to do for workers what the organizations of business attempted to do for their owners and managers. His study of the reception given unions and reform legislation led him to recognize the critical role of the United States Supreme Court (and the courts generally), and its conception of what was reasonable in the development and application of the working rules which governed the acquisition and use of power in the market. Accordingly, Commons developed a theory of property which stressed its evolution and role in governing the structure of participation and relative withholding capacity in the market.

Commons also developed a theory of institutions which focused on their respective different mixtures of bargaining, rationing and managerial transactions, all taking place within a legal framework which was itself subject to change.

Although Commons's institutionalism had different emphases from that of Thorstein Veblen, for example, in that Commons stressed reform of the capitalist framework, they shared a view of economics as political economy and of the economy as comprising more than the market. Unlike Veblen, Commons was not antagonistic toward businessmen, and indeed accepted capitalism, though not necessarily on the terms given or preferred by the established power structure.

Commons was one of the few American economists to found a ‘school’, a tradition that was carried

forward by a corps of students, especially Selig Perlman, Edwin E. Witte, Martin Glaeser and Kenneth Parsons. Much mid-20th-century American social reform, the New Deal for example, drew on or reflected the work of Commons and his fellow workers and students.

Selected works

1893. *The Distribution of Wealth*. New York: Macmillan.

1905. *Trade Unionism and Labor Problems*. Boston: Ginn.

1910–11. (With others.) *Documentary History of American Industrial Society*, 10 vols. Cleveland: A.H. Clark.

1916. (With J.B. Andrews.) *Principles of Labor Legislation*. New York: Harper.

1919. *Industrial Goodwill*. New York: McGraw-Hill.

1919–1935. (With others.) *History of Labor in the United States*, 4 vols. New York: Macmillan.

1921. *Industrial Government*. New York: Macmillan.

1924. *The Legal Foundations of Capitalism*. New York: Macmillan.

1934. *Institutional Economics*. New York: Macmillan.

1934. *Myself*. New York: Macmillan.

1950. *The Economics of Collective Action*. Madison: University of Wisconsin Press.

How to cite this article

Samuels, Warren J. "Commons, John Rogers (1862–1945)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000247>
doi:10.1057/9780230226203.0272

The New Palgrave Dictionary of Economics Online

community indifference curves

Wayne Shafer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bergson–Samuelson social welfare functions; community indifference curves; comparative statics; Scitovsky, T.; utility possibility frontiers

Article

The idea of a community indifference curve, as the term is commonly used, is due to Scitovsky (1942). The genesis of the idea is the fact that comparative statics and welfare analysis in economic models is simplified considerably if there is a social preference ordering over aggregate commodity bundles which reflects the collective individual preferences of agents. Scitovksy's notion of a ‘community indifference curve’ essentially allows the analytical convenience of social indifference curves, in certain circumstances, without having to assume a specific Bergson–Samuelson social welfare function or having to assume the restrictive assumptions on agents' preferences needed to guarantee that agents act collectively as a single individual.

The definition of a community indifference curve is basically simple. Suppose there are m commodities and n agents. Let x denote a commodity vector (as m-vector with non-negative coordinates) and u_i a utility function representing agent i's preferences. We will assume that u_i is monotone increasing and quasi-concave. Given a vector $u' = (u'_1, \ldots, u'_n)$ of utility numbers, the community indifference curve at u', $CIC(u')$, is defined to be the set of all commodity vectors x such that there is a distribution $(x_1, \ldots, x_n)$ of commodity vectors satisfying $\sum_i x_i = x$ and $u_i(x_i) = u'_i$, $i = 1, \ldots, n$, and there is no $x' \leq x$, $x' \neq x$ which also has this property. Thus one can obtain any vector $x \in CIC(u')$ by fixing the quantities of all but one good and minimize the amount of the remaining good subject to achieving u'. As pointed out by Samuelson (1956), the community indifference curve can be interpreted as a ‘dual’ to the utility possibility frontier. The utility possibility frontier, for a given x, is the set of all vectors u' of utility numbers achievable by a Pareto efficient distribution of x to the agents. Let $U(x)$ denote the utility

possibility frontier for the commodity vector x. Then it is easy to see that $CIC(u') = \{x: u' \in U(x)\}$ and that $U(x) = \{u': x \in CIC(u')\}$.

We will now describe the most important properties of community indifference curves. First, each $CIC(u')$ looks like the indifference curve of a monotone quasi-concave utility function. That is, the set of vectors x such that $x \geq x^1$ for some $x^1 \in CIC(u')$ is a convex set. For example, when $m = 2$, CIC(u') is a curve with a diminishing marginal rate of substitution. Second, unlike the utility possibility frontier, the community indifference curve is essentially an ordinal concept, that is it does not depend on the choice of utility functions representing agents' preferences, in the following sense. Suppose, for each i, u_i and v_i are two utility functions representing agent i's preferences, and let $(x_1, \ldots, x_n)$ be a Pareto efficient allocation to the agents. Define $u' = [u_1(x_1), \ldots, u_n(x_n)]$ and $v' = [v_1(x_1), \ldots, v_n(x_n)]$. Then $CIC(u') \equiv CIC(v')$. Clearly, community indifference curves can be parameterized by a given Pareto efficient allocation of goods rather than a given vector of utilities. Third, assuming smooth utility functions, the marginal rate of substitution for any two commodities on a community indifference curve is equal to the common marginal rate of substitution of each agent. Specifically, pick an $x \in CIC(u')$, and let $(x_1, \ldots, x_n)$ be the Pareto efficient allocation of x such that $u_i(x_i) = u_i'$, $i = 1, \ldots, n$. Then for any two commodities h and h', the marginal rate of substitution of h and h' evaluated at $x \in CIC(u')$ is equal to the marginal rate of substitution of h for h' at x_i on agent i's indifference curve through x_i.

Fourth, and very important, community indifference curves are not, in general, ‘indifference’ curves in the sense of being level curves of some function. Pick any x, and $u', u'' \in U(x)$, such that $u' \neq u''$. Then by definition, $x \in CIC(u') \cap CIC(u'')$. Thus $CIC(u')$ must either coincide with $CIC(u'')$ or intersect properly. The condition for two community indifference curves never to intersect properly is then that $CIC(u') = CIC(u'')$ for all $u', u'' \in U(x)$, for all x. It turns out that this is true if and only if the agents have identical homothetic preferences, in which case the family of all community indifference curves will coincide with the family of indifference curves for the common preferences of the agents.

From the above definition and properties, the following observation constitutes the basic use of community indifference curves: if the economy is currently at a vector of utility numbers u', then x' is a commodity vector which lies above $CIC(u')$ if and only if there is some distribution of x' to the agents which will achieve a vector of utilities u'' such that $u'' > u'$. In this sense, x' is ‘better’ than any $x \in CIC(u')$. However, since from above community indifference curves can intersect properly, it may also be that there is a u''' such that $x' \in CIC(u''')$ and an $x' \in CIC(u')$ such that x lies above $CIC(u''')$, in which case x is also ‘better’ than x'. Thus it is important to realize that community indifference curves cannot be used to define a social ordering of aggregate output vectors. Nevertheless, community indifference curves can still be a useful analytical device. For example, consider a market economy with two produced goods. Consider an equilibrium in which all consumers face the same prices, in terms of the aggregate output vector x' and the vector of utilities u' obtained by the agents. Graphically this equilibrium can be represented by drawing the production possibility frontier and $CIC(u$

′), noting they meet at x'. The slope of the production possibility frontier at x' represents the price ratio faced by firms, and the slope of the $CIC(u')$ at x' the common price ratio faced by consumers. If firms and consumers face the same price ratio, then the $CIC(u')$ must be tangent to the production possibility frontier at x. Thus no feasible x can be produced which can make all agents better off, so the situation is Pareto optimal. If, however, firms face different prices than the agents because of, for example, taxes or tariffs, then the slope of the $CIC(u')$ will be different from the slope of the production possibility frontier, and thus the two curves will intersect properly. In this case there must exist an x' on the production possibility frontier which lies above $CIC(u')$, so the original situation is Pareto inefficient.

See Also

- Arrow's theorem
- optimality and efficiency
- social welfare function
- welfare economics

Bibliography

Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70(1), 1–22.

Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9(2), 89–110.

How to cite this article

Shafer, Wayne. "community indifference curves." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000251> doi:10.1057/9780230226203.0273

The New Palgrave Dictionary of Economics Online

comparative advantage

Ronald Findlay

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article traces the evolution of the theory of comparative advantage and the gains from trade from the pioneering work of David Ricardo to the factor proportions approach of Eli Heckscher and Bertil Ohlin. Extensions of the basic models to many goods, factors and countries, and to the long run are noted, as well as the attempts at empirical testing of the predictions derived from them.

Keywords

absolute advantage; comparative advantage; Corn Laws; distributive justice; division of labour; factor price equalization; factor proportions; fixed factors; free trade; gains from trade; Haberler, G.; Heckscher, E. F.; Heckscher–Ohlin trade theory; human capital; international trade; intra-industry' trade; labour mobility; Leontief, W.; Lerner, A. P.; monopolistic competition; neoclassical general equilibrium theory; Ohlin, B. G.; product differentiation; returns to scale; specialization; specific factors; stationary state; Stolper–Samuelson theorem; terms of trade; time preference; wages fund

Article

The modern economy, and the very world as we know it today, obviously depends fundamentally on specialization and the division of labour, between individuals, firms and nations. The principle of comparative advantage, first clearly stated and proved by David Ricardo in 1817, is the fundamental analytical explanation of the source of these enormous 'gains from trade'. Though an awareness of the benefits of specialization must go back to the dim mists of antiquity in all civilizations, it was not until Ricardo that this deepest and most beautiful result in all of economics was obtained. Though the logic applies equally to inter*personal*, inter*firm* and inter*regional* trade, it was in the context of inter*national* trade that the principle of comparative advantage was discovered and has been investigated ever since.

The basic Ricardian model

What constituted a ‘nation’ for Ricardo were two things – a ‘factor endowment’, of a specified number of units of labour in the simplest model, and a ‘technology’, the productivity of this labour in terms of different goods, such as cloth and wine in his example. Thus labour can move freely between the production of cloth and wine in England and in Portugal, but each labour force is trapped within its own borders. Suppose that a unit of labour in Portugal can produce one unit of cloth or one unit of wine, while in England a unit of labour can produce four units of cloth or two units of wine. Thus the opportunity cost of a unit of wine is one unit of cloth in Portugal while it is two units of cloth in England. On the assumption of competitive markets and free trade, it follows that *both* goods will never be produced in *both* countries since wine in England and cloth in Portugal could always be undermined by a simple arbitrage operation involving export of cloth from England and import of wine from Portugal. Thus wine in England or cloth in Portugal must contract until at least one of these industries produces zero output. If both goods are consumed in positive amounts, the ‘terms of trade’ in equilibrium must lie in the closed interval between one and two units of cloth per unit of wine. Which of the two countries specializes completely will depend upon the relative size of each country (as measured by the labour force *and* its productivity in each industry) and upon the extent to which each of the two goods is favoured by the pattern of world demand. Thus Portugal is more likely to specialize the smaller she is compared with England in the sense defined above and the more world demand is skewed towards the consumption of wine relative to the consumption of cloth.

The gains from trade

Viewed as a ‘positive’ theory, the principle of comparative advantage yields *predictions* about (*a*) the *direction* of trade: that each country exports the good in which it has the lower comparative opportunity cost ratio as defined by the technology in that country, and about (*b*) the *terms* of trade: that it is bounded above and below by these comparative cost ratios. From a ‘normative’ standpoint the principle implies that the citizens of each country become ‘better off’ as a result of trade, with the extent of the gains from trade depending upon the degree to which the terms of trade exceed the domestic comparative cost ratio. It is the ‘normative’ part of the doctrine that has always been the more controversial, and it is therefore necessary to evaluate it with the greatest care.

In Ricardo's example the total labour force in each country is presumably supplied by an aggregate of different households, each having the same *relative* productivity in the two sectors. Thus *all* households in *each* country must become better off as a result of trade if the terms of trade lie strictly in between the domestic comparative cost ratios. The import-competing sector in each country simply switches over instantaneously and costlessly to producing the export good (moving to the opposite corner of its linear production-possibilities frontier, in terms of the familiar geometry), obtaining the desired level of the other good by imports, raising utility in the process. When one country is incompletely specialized, then all households in that country remain at unchanged utility levels, all of the gain from trade going to the individuals in the ‘small’ country. Thus we have a situation in which *everybody gains*, in at least one country, while *nobody loses* in either country, as a result of trade.

This very strong result depends upon Ricardo's assumption of perfect occupational mobility in each country. Suppose we take the opposite extreme of completely *specific* labour in each sector, so that each country produces a fixed combination of cloth and wine, with no possibility of transformation. In this

case, labour in the import-competing sector in each country must necessarily *lose*, as a result of trade, while labour in each country's export sector must gain. It can be shown, however, that trade will improve *potential* welfare in each country in the Samuelson (1950) sense that the utility-possibility frontier with trade will dominate the corresponding frontier without trade, so that no one need be worse off, and at least some one better off, if lump-sum taxes and transfers are possible (Samuelson, 1962).

International factor mobility and world welfare

Another very important normative issue is the question of the relationship between the free-trade equilibrium and *world* efficiency and welfare. In the Ricardian model world welfare in general will *not* be maximized by free trade alone. In the numerical example considered here Ricardo stresses the fact that England can still gain from trade even though she has an *absolute* advantage in the production of *both* goods, her productivity being greater in both cloth and wine, though comparatively greater in cloth. Suppose that labour in Portugal could produce at English levels, *if it moved to England*; that is, the English superiority is based on climatic or other ‘environmental’ factors and not on differences in aptitude or skill. Then, if labour was free to move, and in the absence of ‘national’ sentiment, all production would be located in England, and Portugal would cease to exist. The former Portuguese labour would be better off than under free trade, since their real wage in terms of wine will now be two units instead of one. The English labour would be worse off, if the terms of trade were originally better than 0.5 wine per unit of cloth, but it is easy to show that they could be sufficiently compensated since the utility-possibility frontier for the world economy as a whole is moved out by the integration of the labour forces.

The case when each country has an absolute advantage in one good is more interesting. As is easy to see, from Findlay (1982), this case will involve a movement of labour to the country with the higher real wage under free trade, increasing the production of this country's exportable and reducing that of the lower-wage country under free trade. The terms of trade turn against the higher-wage country until eventually the real wage is equalized. The terms of trade that achieve this equality of real wages will be equal to the ratio of labour productivities in each country's export sector; that is, the ‘double factoral’ terms of trade will be unity. This solution of free trade *combined* with perfect labour mobility will achieve not only efficiency for the world economy as a whole but equity as well. ‘Unequal exchange’ in the sense of Emmanuel (1972) would not exist, while liberal, utilitarian and Rawlsian criteria of distributive justice would be satisfied as well, as pointed out in Findlay (1982). Despite all this, it still seems utopian to expect a policy of ‘open borders’, in *either* direction, for the contemporary world of nation-states.

Extensions of the basic Ricardian model

The two-country, two-good Ricardian model was extended to many goods and countries by a number of subsequent writers, whose efforts are described in detail by Haberler (1933) and Viner (1937). In the case of two countries and *n* goods the concept of a ‘chain of comparative advantage’ has been put forward, with the goods listed in descending order in terms of the *relative* efficiency of the two countries in producing them. It is readily shown that with a uniform wage in each country all goods from 1 to

some number *j* must be exported, while all goods from (*j*+1) to *n* must be imported. The number *j* itself will depend upon the relative sizes of the two countries and the composition of world demand. Dornbusch, Fischer and Samuelson (1977) generalize this result to a continuum of goods in an extremely elegant and powerful model that has been widely used in subsequent literature. An analogous chain concept applies to the case of two goods and *n* countries, this time ranking the countries in terms of the ratio of their productivities in the two goods, with country 1 having the greatest *relative* efficiency in cloth and country *n* in wine. World demand and the sizes of the labour forces will determine the 'marginal' country *j*, with countries 1 to *j* exporting cloth and (*j*+1) to *n* exporting wine.

The simultaneous consideration of comparative advantage with many goods and many countries presents severe analytical difficulties. Graham (1948) considered several elaborate numerical examples, his work inspiring the Rochester theorists McKenzie (1954) and Jones (1961) to apply the powerful tools of activity analysis to this particular case of a linear general equilibrium model. It is interesting to note in connection with mathematical programming and activity analysis that Kantorovich (1965) in his celebrated book on planning for the Soviet economy worked out an example of optimal specialization patterns for factories that corresponds *exactly* to the Ricardian model of trade between countries.

The three-factor Ricardian model

While most of the literature on the Ricardian trade model has concentrated on the model of Chapter 7 of the *Principles* in which it appears that labour is the sole scarce factor, his more extended model in the *Essay on Profits* has been curiously neglected, though the connections between trade, income distribution and growth which that analysis explores are quite fascinating. The formal structure of the model was laid out very thoroughly in Pasinetti (1960). The economy produces two goods, corn and manufactures, each of which has a one-period lag between the input of labour and the emergence of output. Labour thus has to be supported by a 'wage fund', an initially given stock that is accumulated over time by saving out of profits. Corn also requires land as an input, which is in fixed supply and yields diminishing returns to successive increments of labour. The wage-rate is given exogenously in terms of corn, and manufactures are a luxury good consumed only by the land-owning class, who obtain rents determined by the marginal product of land. Profits are the difference between the marginal product of labour and the given real wage, which is equal to the marginal product 'discounted' by the rate of interest, in this model equal to the rate of profit, defined as the ratio of profits to the real wage that has to be advanced a period before. Momentary equilibrium determines the relative price of corn and manufactures, the rent per acre and the rate of profit, as well as the output levels and allocation of the labour force between sectors. The growth of the system is at a rate equal to the product of the rate of profit and the propensity to save of the capitalist class. It is shown that the system approaches a stationary state, with a monotonically falling rate of profit and rising rents per acre.

The opportunity to import corn more cheaply from abroad will have significant distributional and growth consequences. Just as Ricardo argued in his case for the repeal of the Corn Laws, cheaper foreign corn will reduce domestic rents and raise the domestic rate of profit, and thus the rate of growth. The approach to the stationary state is postponed, though of course it cannot be ultimately averted, while the growth consequences for the corn exporter are definitely adverse. The main doctrinal significance of this wider Ricardian model, however, is to reveal the extent to which the subsequent 'general

equilibrium' or 'neoclassical' approach to international trade is already present within the Ricardian framework. For one thing, the pattern of comparative advantage itself depends upon the complex interaction of technology, factor proportions and tastes. In his Chapter 7 case the pattern of comparative advantage is *exogenous*, simply given by the four fixed technical coefficients indicating the productivity of labour in cloth and wine in England and Portugal. The production-possibility frontiers for each country are linear, and comparative advantage is simply determined by the relative magnitudes of the slopes. As demonstrated in Findlay (1974), however, the *Essay on Profits* model implies a concave production-possibilities frontier at any moment, since there are diminishing returns to labour in corn even though the marginal productivity of labour in manufactures is constant. With two countries the pattern of comparative advantage will depend upon the slopes of these curves at their autarky equilibria, which are *endogenous* variables depending upon the sizes of the 'wage fund' in relation to the supply of land and the consumption pattern of landowners, as well as the technology for the two goods.

As Burgstaller (1986) points out, however, the steady-state solution of the model restores the linear structure of the pattern of comparative advantage. The zero profit rate in the steady state requires the marginal product of labour to be equal to the given real wage, and this implies a fixed land–labour ratio and hence output per unit of labour in corn. We thus once again have two fixed technical coefficients, so that the slope of the linear production-possibilities frontier is once again an exogenous indicator of comparative advantage.

The 'neo-Ricardian' approach of Steedman (1979a; 1979b) considers more general time-phased structures of production. Technology alone determines negatively sloped wage–profit or factor-price frontiers, any point on which generates a set of relative product prices and hence a pattern of comparative advantage relative to another such economy.

Factor proportions and the Heckscher – Ohlin model

While J.S. Mill, Marshall and Edgeworth all made major contributions to trade theory, the concept of comparative advantage did not undergo any evolution in their work beyond the stage at which Ricardo had left it. They essentially concentrated on the determination of the terms of trade and on various comparative static exercises. The interwar years, however, brought fundamental advances, stemming in particular from the work of the Swedes Heckscher (1919) and Ohlin (1933). The development of a diagrammatic apparatus to handle general equilibrium interactions of tastes, technology and factor endowments by Haberler (1933), Leontief (1933), Lerner (1932) and others culminated in the rigorous establishment of trade theory and comparative advantage as a branch of neoclassical general equilibrium theory.

The essentials of this approach can be expounded in terms of the familiar two-country, two-good and two-factor model, on which see Jones (1965) for a detailed and lucid algebraic exposition. The given factor supplies and constant returns to scale technology define concave production-possibility frontiers, on the assumption that the goods differ in factor intensity. This determines the 'supply side' of the model, which is closed by the specification of consumer preferences. Economies that have identical technology, factor endowments and tastes will have the same autarky equilibrium price-ratio and so will have no incentive to engage in trade. Countries must therefore differ with respect to at least one of these characteristics for differences in comparative advantage to emerge. With identical technology and factor

endowments, a country will have a comparative advantage in the good its citizens prefer *less* in comparison to the foreign country, since then this good will be cheaper at home. Similarly, if factor endowments and tastes are identical, differences in comparative advantage will be governed by relative technological efficiency; that is, a country will have a comparative advantage in the good in which its relative technological efficiency is greater, just as in the Ricardian model. These differences in technological efficiency could be represented, for example, by the magnitude of multiplicative constants in the production functions; that is, 'Hicks-neutral' differences.

In keeping with the ideas of Heckscher and Ohlin, however, it is differences in factor proportions that have dominated the explanation of comparative advantage in the neoclassical literature. The Heckscher–Ohlin theorem, that each country will export the commodity that uses its relatively abundant factor most intensively, has been rigorously established and the necessary qualifications carefully specified, as in Jones (1956). Among the more important of these is the requirement that factor-intensity 'reversals' do not take place; that is, that one good is always more capital-intensive than the other at all wage-rental ratios or at least within the relevant range defined by the factor proportions of the trading countries.

The Stolper – Samuelson theorem

Associated with the Heckscher–Ohlin theorem is the Stolper–Samuelson theorem (1941), that trade benefits the abundant and harms the scarce factor while protection does the opposite, and the celebrated factor price equalization theorem of Lerner (1952, though written in 1932) and Samuelson (1948; 1949; 1953), which states that under certain conditions free trade will lead to complete equalization of factor rewards even though factors are not mobile internationally. The normative significance of this theorem is that free trade alone can achieve world efficiency in production and resource allocation, unlike the case of the Ricardian model as pointed out earlier. The requirements for the theorem to hold, however, are very stringent, such as that the number of tradable goods produced be equal to the number of factors. It also requires factor proportions to be sufficiently close to each other in the trading partners so that the production patterns are fairly similar. Thus it would be far-fetched to expect the price of unskilled labour to be equalized between Bangladesh and the United States, for example.

The specific-factors model

An important and popular variant of the factor proportions approach is what Jones (1971) calls the 'specific factors' and Samuelson (1971) the Viner–Ricardo model. In this model each production sector has its own unique fixed factor, while labour is used in all sectors and is perfectly mobile internally between them. Trade patterns still reflect factor endowments but factor price equalization does not hold in this model since the number of factors is always one greater than the number of goods. Gruen and Corden (1970) present an ingenious three-by-three extension of this approach, in which one sector uses land and labour, while the two others use capital and labour, thus neatly integrating the 'specific factors' model with the regular two-by-two Heckscher–Ohlin model. Findlay (1995, chs 4 and 6) uses adaptations and extensions of the Gruen–Corden specification to introduce human capital formation and the concept of a natural resource 'frontier' into the Heckscher–Ohlin framework.

Long-run extensions of the factor proportions model

One limitation of the Heckscher–Ohlin model was that the stock of ‘capital’, however conceived, should be an endogenous variable determined by the propensity to save or time preference of each trading community, rather than being taken as exogenously fixed. Oniki and Uzawa (1965) extended the model to a situation where the labour force is growing in each country at an exogenous rate and capital is accumulated in response to given propensities to save in each country. One of the goods is taken to be the ‘capital’ good, conceived of as a malleable ‘putty–putty’ instrument. They demonstrated that the system converges in the long run to a particular capital–labour ratio for each country, which will be higher for the country with the larger saving propensity. In Findlay (1970; 1984), it is shown that as the capital–labour ratio evolves the pattern of comparative advantage for a given ‘small’ country in an open trading world will also shift over time towards more capital-intensive goods, thus formalizing the concept of a ‘ladder of comparative advantage’ that countries ascend in the process of economic development. Thus comparative advantage should not be conceived as given and immutable, but evolving with capital accumulation and technological change. Much of the loose talk about ‘dynamic’ comparative advantage in the development literature, however, is misconceived since it attempts to change the pattern of production by protection *before* the necessary changes in the capacity to produce efficiently have taken place. Other models which endogenize the capital stocks of the trading countries are Stiglitz (1970) and Findlay (1978) which utilizes a variable rate of time preference and an ‘Austrian’ point-input/point-output technology, which implies a continuum of capital goods as represented by the ‘trees’ of different ages, and Findlay (1995, ch. 2), which addresses the question posed by Samuelson (1965) of whether trade equalizes not only the marginal product or rental of capital but the rate of interest itself.

Empirical testing

Empirical testing of the positive side of the theory of comparative advantage begins in a systematic way only with the work of MacDougall (1951) on the Ricardian theory and the celebrated article of Leontief (1954) that uncovered the apparent paradox that US exports were more labour-intensive than her imports. Leontief's dramatic finding spurred considerable further empirical research motivated by the desire to find a satisfactory explanation. The increasing scarcity of natural resources in the USA, by causing capital to be substituted for it in import-competing production, was stressed as an explanation for the paradox by Vanek (1963). The role of ‘human capital’ as an explanation was pointed to by Kenen (1965) and a number of empirical investigators, who found that US exports were considerably more skill-intensive than her imports, even though physical capital-intensity was only weakly correlated with exports and imports. This pointed to the need to reinterpret the simple Heckscher–Ohlin model in terms of skilled and unskilled labour as the two factors, rather than labour of uniform quality and physical capital. Since the formation of skill through education is an endogenous variable, a function of a wage differential that is itself a function of trade, we need a general equilibrium model that can simultaneously handle both these aspects, as in Findlay and Kierzkowski (1983) and Findlay (1995, ch. 4).
Many other extensions of the Heckscher–Ohlin theory are surveyed in Jones and Neary (1984) and

Ethier (1984), while Deardorf (1984) and Feenstra (2004) give very incisive accounts of the attempts at empirical testing of the theory of comparative advantage in its different manifestations. Further important progress in empirical testing of the Heckscher–Ohlin model has been achieved by the work of Leamer (1984), Trefler (1995), Harrigan (1997) and Davis and Weinstein (2001).

Increasing returns

Finally, the crucial role of increasing returns to scale in specialization and international trade has only recently been rigorously investigated, since it implies departures from perfect competition. Krugman (1979) and Lancaster (1980) introduced international trade into models of monopolistic competition with differentiated products, showing the possibility of gains from trade due to the provision of greater variety of similar goods rather than differences in comparative advantage, what is referred to as ‘intra-industry’ trade. Helpman and Krugman (1985) thoroughly examine and extend our knowledge in this area, while Grossman and Helpman (1991) expertly extend the monopolistic competition approach to deal with a number of issues involving endogenous technological progress and growth in the world economy.

See Also

- Heckscher, Eli Filip
- Heckscher–Ohlin trade theory
- international trade theory
- Leontief, Wassily
- Ohlin, Bertil Gotthard
- Ricardo, David
- terms of trade

Bibliography

Burgstaller, A. 1986. Unifying Ricardo's theories of growth and comparative advantage. *Economica* 53, 467–81.

Davis, D.R. and Weinstein, D.E. 2001. An account of global factor trade. *American Economic Review* 91, 1423–53.

Deardorf, A. 1984. Testing trade theories. In *Handbook of International Economics*, vol. 1, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.

Dornbusch, R., Fischer, S. and Samuelson, P.A. 1977. Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67, 823–39.

Emmanuel, A. 1972. *Unequal Exchange*. New York: Monthly Review Press.

Ethier, W. 1984. Higher dimensional issues in trade theory. In *Handbook of International Economics*, vol. 1, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.

Feenstra, R.C. 2004. *Advanced International Trade*. Princeton: Princeton University Press.

Findlay, R. 1970. Factor proportions and comparative advantage in the long run. *Journal of Political Economy* 78, 27–34.

Findlay, R. 1974. Relative prices, growth and trade in a simple Ricardian system. *Economica* 41, 1–13.

Findlay, R. 1978. An ‘Austrian’ model of international trade and interest equalization. *Journal of Political Economy* 86, 989–1007.

Findlay, R. 1982. International distributive justice. *Journal of International Economics* 13, 1–14.

Findlay, R. 1984. Growth and development in trade models. In *Handbook of International Economics*, vol. 1, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.

Findlay, R. 1995. *Factor Proportions, Trade, and Growth*. Cambridge, MA: MIT Press.

Findlay, R. and Kierzkowski, H. 1983. International trade and human capital: a simple general equilibrium model. *Journal of Political Economy* 91, 957–78.

Graham, F. 1948. *The Theory of International Values*. Princeton: Princeton University Press.

Grossman, G.M. and Helpman, E. 1991. *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.

Gruen, F. and Corden, W.M. 1970. A tariff that worsens the terms of trade. In *Studies in International Economics*, ed. I.A. MacDougall and R. Snape. Amsterdam: North-Holland.

Haberler, G. 1933. *The Theory of International Trade*, Trans. by A. Stonier and F. Benham, London: W. Hodge, 1936; revised edn, 1937.

Harrigan, J. 1997. Technology, factor supplies and international specialization: estimating the neoclassical model. *American Economic Review* 87, 475–94.

Heckscher, E. 1919. The effects of foreign trade on the distribution of income. In *Ekonomisk Tidskrift*.

English translation in *Readings in the Theory of International Trade*, ed. H.S. Ellis and L.A. Metzler. Philadelphia: Blakiston, 1949.

Helpman, E. and Krugman, P. 1985. *Market Structure and Foreign Trade*. Cambridge, MA: MIT Press.

Jones, R.W. 1956. Factor proportions and the Heckscher–Ohlin theorem. *Review of Economic Studies* 24, 1–10.

Jones, R.W. 1961. Comparative advantage and the theory of tariffs. *Review of Economic Studies* 28, 161–75.

Jones, R.W. 1965. The structure of simple general equilibrium models. *Journal of Political Economy* 73, 557–72.

Jones, R.W. 1971. A three-factor model in theory, trade and history. In *Trade, Balance of Payments, and Growth*, ed. J. Bhagwati et al. Amsterdam: North-Holland.

Jones, R.W. and Neary, P. 1984. Positive trade theory. In *Handbook of International Economics*, vol. 1, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.

Kantorovich, L. 1965. *The Best Use of Economic Resources*. Cambridge, MA: Harvard University Press.

Kenen, P.B. 1965. Nature, capital and trade. *Journal of Political Economy* 73, 437–60; Erratum, December, 658.

Krugman, P.R. 1979. Increasing returns, monopolistic competition and international trade. *Journal of International Economics* 9, 469–79.

Lancaster, K.J. 1980. Intra-industry trade under perfect monopolistic competition. *Journal of International Economics* 10, 151–75.

Leamer, E.P. 1984. *Sources of International Comparative Advantage: Theory and Evidence*. Cambridge, MA: MIT Press.

Leontief, W.W. 1933. The use of indifference curves in the analysis of foreign trade. *Quarterly Journal of Economics* 47, 493–503.

Leontief, W.W. 1954. Domestic production and foreign trade: the American capital position re-examined. *Economia Internazionale* 7, 9–38.

Lerner, A.P. 1932. The diagrammatic representation of cost conditions in international trade. *Economica*

12, 345–56.

Lerner, A.P. 1952. Factor prices and international trade. *Economica* 19, 1–15.

MacDougall, G.D.A. 1951. British and American exports. *Economic Journal* 61, 697–724.

McKenzie, L.W. 1954. Specialization and efficiency in world production. *Review of Economic Studies* 21(3), 165–80.

Ohlin, B. 1933. *Inter-regional and International Trade*. Cambridge, MA: Harvard University Press.

Oniki, H. and Uzawa, H. 1965. Patterns of trade and investment in a dynamic model of international trade. *Review of Economic Studies* 32, 15–38.

Pasinetti, L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies* 27, 78–98.

Ricardo, D. 1951. *The Works and Correspondence of David Ricardo*, vols. 1 and 4, ed. P. Sraffa. Cambridge: Cambridge University Press.

Samuelson, P.A. 1948. International trade and the equalization of factor prices. *Economic Journal* 58, 163–84.

Samuelson, P.A. 1949. International factor price equalization once again. *Economic Journal* 59, 181–97.

Samuelson, P.A. 1950. Evaluation of real national income. *Oxford Economic Papers* 2, 1–29.

Samuelson, P.A. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21, 1–20.

Samuelson, P.A. 1962. The gains from international trade once again. *Economic Journal* 72, 820–29.

Samuelson, P.A. 1965. Equalization by trade of the interest rate along with the real wage. In *Trade, Growth and the Balance of Payments*, ed. R. Baldwin et al. Chicago: Rand McNally.

Samuelson, P.A. 1971. Ohlin was right. *Swedish Journal of Economics* 73, 365–84.

Steedman, I. 1979a. *Trade Amongst Growing Economies*. Cambridge: Cambridge University Press.

Steedman, I., ed. 1979b. *Fundamental Issues in Trade Theory*. London: Macmillan.

Stiglitz, J. 1970. Factor–price equalization in a dynamic economy. *Journal of Political Economy* 78, 456–88.

Stolper, W. and Samuelson, P.A. 1941. Protection and real wages. *Review of Economic Studies* 9, 58–73.

Trefler, D. 1995. The case of missing trade and other mysteries. *American Economic Review* 85, 1029–46.

Vanek, J. 1963. *The Natural Resource Content of U.S. Foreign Trade 1870–1955*. Cambridge, MA: MIT Press.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

How to cite this article

Findlay, Ronald. "comparative advantage." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000254> doi:10.1057/9780230226203.0274

The New Palgrave Dictionary of Economics Online

comparative statics

John Nachbar
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Comparative statics in competitive general equilibrium (GE) environments provide insight into the operation of GE models and a means to confront GE models with data. This article focuses on a canonical comparative statics prediction: in an exchange economy, aggregate endowment changes are negatively related to equilibrium price changes. In particular, an increase in the aggregate endowment of a commodity lowers its equilibrium price.

Keywords

aggregate excess demand; asset pricing; comparative statics; competitive general equilibrium; Debreu–Mantel–Sonnenschein theorem; demand shocks; endowment changes; general equilibrium; gross substitutes; Hicks, J.; Partial equilibrium; production economies; Rybcyznski theorem; Stolper–Samuelson theorem; supply shocks; weak axiom of revealed preference

Article

Comparative statics in competitive general equilibrium (GE) environments provide insight into the operation of GE models and a means, at least in principle, to confront GE models with data.
For concreteness, I focus most of this article on what is arguably the canonical GE comparative statics conjecture: in finite exchange economies (that is, no production), equilibrium price changes are negatively related to endowment changes. In particular, if the endowment of good 1 increases and the endowments of other goods remain the same, then the price of good 1 falls. At the end of this article, I briefly survey other GE comparative statics results.
I break the analysis into three cases of increasing complexity.

Case I

There is a single consumer and two commodities. In an equilibrium of this trivial economy, the consumer eats her own endowment. Equilibrium relative prices (which are well defined even though there is no trade) are given by the slope of the consumer's indifference curve through her consumption/endowment point, ω^*. Let E denote her wealth expansion path through her initial endowment; E is the set of points where the slope of her indifference curve is the same as at ω^*.
If the new endowment, $\widehat{\omega}$, lies below E, as in Figure 1, then the equilibrium price ratio p_1/p_2 falls. If $\widehat{\omega}$ lies above E, then p_1/p_2 rises. If $\widehat{\omega}$ lies along E, then p_1/p_2 remains unchanged.
Figure 1
Comparative statics with one consumer and two commodities

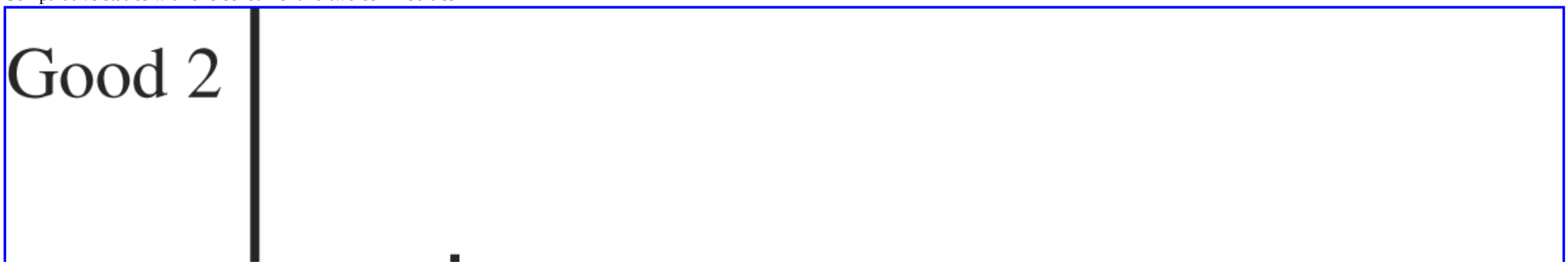

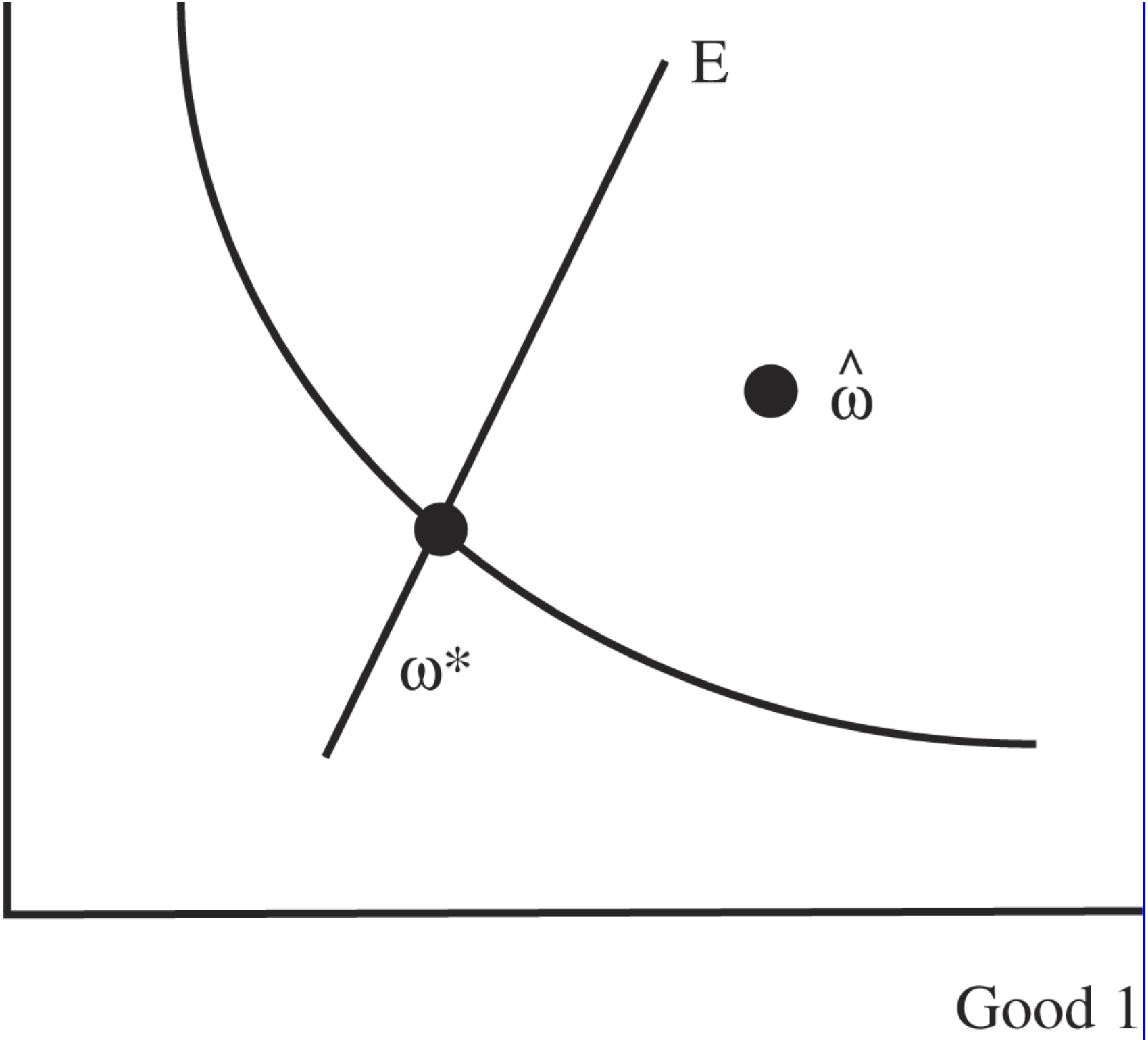
E
$\hat{\omega}$
ω^*
Good 1

The differential version of Figure 1 is given by Figure 2. The vector μ, the tangent to E, is the derivative with respect to nominal wealth of each good's demand (μ is mnemonic for marginal propensity to consume vector). To first order, the rule is that p_1/p_2 falls if and only if $\Delta\omega$, the vector of endowment changes, lies within 180°•clockwise from μ. The vector μ incorporates second order information from the utility function and is, in particular, typically not collinear with the utility gradient.

Figure 2

First order comparative statics with one consumer and two commodities

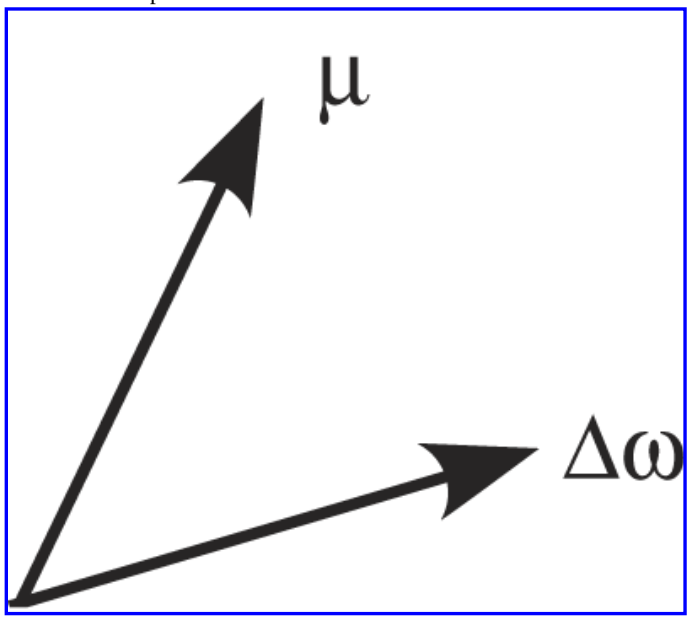

If the endowment of good 1 increases while the endowment of good 2 remains unchanged, then $\Delta\omega$ lies along the positive good 1 axis. Figure 2 implies that, in this case, p_1/p_2 falls if and only if good 2 is normal $(\mu_2 > 0)$; whether good 1 is normal or inferior (or even a Giffen good) is irrelevant.

Case II

There is again one consumer but L commodities. If $\Delta\omega$ lies along the positive good 1 axis, then a natural conjecture, to generalize the above observation for $L = 2$, is that p_1/p_ℓ falls for each good $\ell > 1$, provided each of these goods is normal. Hicks (1939) showed that this conjecture is false in general but that it is true if the gross substitute property holds (GS; the matrix of partial derivatives of excess demand with respect to price has positive off-diagonal entries). GS holds automatically in the one-consumer, $L = 2$, case because, at equilibrium, when $L = 2$, GS is equivalent to the weak axiom of revealed preference (WA). Matters are more complicated if two or more endowments are shifting at the same time. For multivariate endowment shocks, there appears to be little one can say in general about changes in the price ratio of any specific pair of commodities. Instead, the conjecture is that $\Delta\omega$ is negatively related to Δp, the vector of equilibrium price changes. Formally,

$$\Delta p \cdot \Delta\omega \leq 0.$$

Call this the comparative statics inequality, CS for short. Geometrically, CS says that the vectors Δp and $\Delta\omega$ lie at least 90° apart.

To interpret Δp as a change in relative prices, prices must be normalized. Consider linear price normalizations, in which all prices, in both the original economy and the perturbed economy, satisfy $p \cdot \lambda = 1$, where λ is an L vector. If all the coordinates of λ are positive, then a fall in the normalized price of good 1 means that the ratio

$$\frac{p_1}{p_{-1} \cdot \lambda_{-1}}$$

falls, where p_{-1} and λ_{-1} are subvectors corresponding to all goods other than the first: the price of good 1 falls relative to the value of a commodity bundle consisting of λ_ℓ units of each good $\ell > 1$. Standard choices of λ include $\lambda = (0, \ldots, 0, 1)$ (use the last commodity as numeraire) and $\lambda = \omega^*$ (normalize prices so that GNP remains constant; this is the Laspeyres normalization). Regardless of how, or whether, actual prices are normalized, one can re-normalize prices using whatever λ one chooses.

I can provide intuition for CS most easily by continuing to use figures for a two-good economy. Fix a normalizing vector λ. Since $p \cdot \lambda = 1$ for all p, $\Delta p \cdot \lambda = 0$. Therefore, Δp lies along the line that is at right angles to λ, labelled T_λ in Figure 3.

Figure 3
Condition CS with two goods

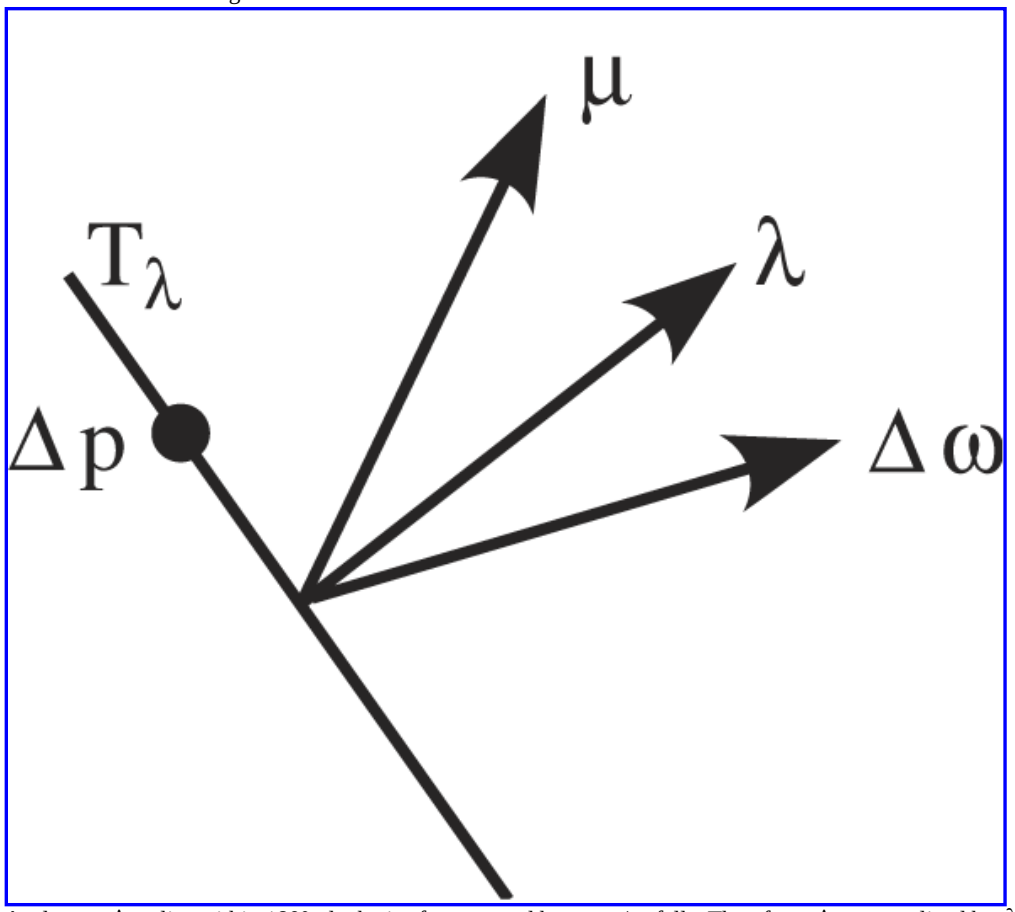

As drawn, $\Delta\omega$ lies within 180°•clockwise from μ and hence p_1/p_2 falls. Therefore, Δp, normalized by λ, lies on the upper left-hand branch of T_λ. As illustrated, $\Delta\omega$ and Δp are more than 90° apart; hence CS holds.

On the other hand, suppose that $\Delta\omega$ lies in the cone spanned by λ and μ. Since $\Delta\omega$ again lies within 180°•clockwise from μ, p_1/p_2 again falls and Δp again lies on the upper left-hand branch of T_λ. Now, however, $\Delta\omega$ and Δp are less than 90° apart. CS fails.

In general, in a one-consumer economy, for any number of commodities and for any preferences, CS fails whenever there is a gap between λ and μ and $\Delta\omega$ falls into this gap. Conversely, if $\lambda = \mu$ (or, more generally, if λ is a scalar multiple of μ) then CS holds for any endowment change: $\Delta p \cdot \Delta\omega \leq 0$ with $\Delta p \cdot \Delta\omega = 0$ if and only if $\Delta\omega$ is collinear with μ (which is the differential analog of $\Delta\omega$ landing on the wealth expansion path E in Figure 1). In one consumer economies, $\lambda = \mu$ is thus the unique (up to scalar multiplication) linear price normalization for which CS holds for all endowment changes.

If preferences are quasi-linear in good L, and consumption is interior, then $\lambda = \mu$ implies $\lambda = (0, \ldots, 0, 1)$; the last good is used as numeraire. If the preferences are homothetic then μ is a scalar multiple of the reference endowment, ω^*, and so one can set $\lambda = \omega^*$. Typically, however, $\lambda = \mu$ is different from price normalizations commonly used in economics.

The $\lambda = \mu$ normalization, although non-standard, does have a sensible interpretation, provided μ is positive (all goods are weakly normal). If μ is positive, then a decrease in p_1 means that $p_1 / (p_{-1} \cdot \mu_{-1})$ falls: the price of good 1 falls relative to the value of the consumer's *marginal* consumption of all other goods.

If $\Delta\omega$ lies along the positive good 1 axis and goods 2, …, L are normal, then a minor variation on CS implies that $p_1 / (p_{-1} \cdot \mu_{-1})$ falls, even if good 1 is inferior. This is a weaker conclusion than that of Hicks (1939) but it has a weaker hypothesis, since it does not assume the gross substitute property.

Case III

There are *I* consumers and *L* commodities. The generalization of CS is

$$\Delta p \cdot \Delta \overline{\omega} \leq 0,$$

where $\Delta \overline{\omega}$ denotes the change in the aggregate endowment. CS holds provided one uses an appropriate aggregate version of μ. Consider two alternatives. Each is a weighted sum of the individual marginal propensity to consume vectors, μ^i. In the first, $\overline{\mu}_{\Delta x}$, the weight on μ^i is consumer *i*'s share in the change in the value of consumption, evaluated at the prices of the reference equilibrium. In the second, $\overline{\mu}_{\Delta \omega}$, the weight on μ^i is *i*'s share in the change in the value of the endowment, again evaluated at reference equilibrium prices.

If one normalizes prices using $\lambda = \overline{\mu}_{\Delta x}$, then inequality CS holds provided *individual* excess demand satisfies the weak axiom (WA) at equilibrium. If $\lambda = \overline{\mu}_{\Delta \omega}$, then CS holds provided *aggregate* excess demand satisfies WA at equilibrium. See Nachbar (2002).

The hypothesis that aggregate excess demand satisfies WA is not implied by standard GE assumptions. One justification for nevertheless assuming WA is that it seems to be connected to the dynamic stability of the price adjustment process. WA holding at equilibrium is sufficient and almost necessary for local asymptotic stability under the Walrasian tâtonnement, for example. Comparative statics, by assuming that economies are at equilibrium, may therefore implicitly assume that aggregate excess demand satisfies WA. A second justification for assuming that aggregate excess demand satisfies WA is that this assumption, while strong, is not implausible in exchange economies. For some sufficient conditions for WA, see Becker (1962), Hildenbrand (1983), Grandmont (1992) and Quah (1997).

In the one-consumer case, the $\lambda = u$ price normalization was necessary as well as sufficient. There are analogous, but weaker, necessity results for $\overline{\mu}_{\Delta x}$ and $\overline{\mu}_{\Delta \omega}$. The important implication is that, because both $\overline{\mu}_{\Delta x}$ and $\overline{\mu}_{\Delta \omega}$ can vary with how the endowment changes are distributed across consumers, there may be *no* price normalization for which CS holds for all endowment changes. As the endowment distribution changes, the price normalization may have to change.

This illustrates an issue that has become a central theme in the recent literature on GE comparative statics. Given an arbitrary price normalization, standard GE assumptions impose no restrictions on the relationship between changes in equilibrium prices and changes in the aggregate endowment (see Chiappori et al., 2004). This negative result, a cousin of the Debreu–Mantel–Sonnenschein theorem (DMS), has a loophole: standard GE assumptions do provide comparative statics restrictions if one works with micro-level information (for example, on the endowment distribution) rather than exclusively with aggregates. In the CS results, micro-level data is used in the price normalization. Note that CS requires micro data even if one assumes that aggregate excess demand satisfies WA.

Relative to the objectives laid out in the first paragraph of this article, the results on CS comparative statics fare reasonably well in providing insight into the operation of GE models. The $\overline{\mu}_{\Delta \omega}$ result is much the easier to interpret, since it is computed with the use of endowment changes, which are exogenous. The $\overline{\mu}_{\Delta x}$ result, on the other hand, extends easily to production economies. In contrast, I do not know whether the $\overline{\mu}_{\Delta \omega}$ result has a useful analog for production economies.

The CS inequality fares less well as a tool for empirical work, because it requires a large amount of data just to estimate the normalization vector. The necessity results imply that this difficulty is intrinsic to CS.

Other comparative statics results

Brown and Matzkin (1996), a path-breaking paper that has heavily influenced subsequent work in this area, exploits the DMS loophole noted above to give testable restrictions linking equilibrium prices with individual endowments. For related work, see Snyder (1999), Williams (2002), Kübler (2003) and Chiappori et al. (2004). Relative to CS, the Brown–Matzkin restrictions are easier to implement empirically because they do not require estimating normalization vectors, but they are harder to interpret.

As already noted, CS-type reasoning can be extended to production economies (see Quah, 2003; Nachbar, 2004). CS-type reasoning can also be extended to asset pricing environments (see Quah, 2003).

For shocks to preferences rather than endowments or technologies, the analog of CS is

$$\Delta p \cdot \Delta \overline{x} \geq 0,$$

where $\Delta \overline{x}$ is the change in equilibrium consumption. Profit maximization implies that this inequality holds for any price normalization. In this respect, the analysis of demand shocks is trivial compared with the analysis of supply shocks.

Interest in comparative statics has helped motivate research on the uniqueness, regularity, and stability of equilibria (see Kehoe, 1987). Note that some of the comparative statics results cited above (for example, the Brown–Matzkin results and the $\overline{\mu}_{\Delta x}$ CS result) do not assume uniqueness or stability.

Finally, perhaps the most famous comparative statics results are the Stolper-Samuelson theorem and its dual, the Rybcyznski theorem (for a recent treatment, see Echenique and Manelli, 2005). Stolper–Samuelson links changes in factor prices with factor intensities and changes in output prices. Rybcyznski links changes in final goods production with factor intensities and changes in factor supplies. Although it is possible to embed these results within a highly restricted GE model, they are partial equilibrium in spirit; wealth effects play no role.

See Also

- general equilibrium
- general equilibrium (new developments)
- international trade theory

Bibliography

Becker, G. 1962. Irrational behavior and economic theory. *Journal of Political Economy* 70, 1–13.

Brown, D. and Matzkin, R. 1996. Testable restrictions on the equilibrium manifold. *Econometrica* 64, 1249–62.

Chiappori, P., Ekelund, I., Kübler, F. and Polemarchakis, H. 2004. Testable implications of general equilibrium theory: a differentiable approach. *Journal of Mathematical Economics* 40, 105–19.

Echenique, F. and Manelli, A. 2005. Comparative statics, English auctions, and the Stolper–Samuelson theorem. Mimeo. Tempe: Arizona State University.

Grandmont, J. 1992. Transformations of the commodity space, behavioral heterogeneity and the aggregation problem. *Journal of Economic Theory* 57, 1–35.

Hicks, J. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51, 997–1018.

Kehoe, T. 1987. Comparative statics. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. Basingstoke: Palgrave.

Kübler, F. 2003. Observable restrictions of general equilibrium with financial markets. *Journal of Economic Theory* 110, 137–53.

Nachbar, J. 2002. General equilibrium comparative statics. *Econometrica* 79, 2065–74.

Nachbar, J. 2004. General equilibrium comparative statics: the discrete case with production. *Journal of Mathematical Economics* 40, 153–63.

Quah, J. 1997. The law of demand when income is price dependent. *Econometrica* 65, 1421–42.

Quah, J. 2003. Market demand and comparative statics when goods are normal. *Journal of Mathematical Economics* 39, 317–33.

Snyder, S. 1999. Testable restrictions of Pareto optimal public good provision. *Journal of Public Economics* 71, 97–119.

Williams, S. 2002. Equations on the derivatives of an initial endowment-competitive equilibrium mapping for an exchange economy. Mimeo. Champaign-Urbana: University of Illinois.

How to cite this article

Nachbar, John. "comparative statics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000256> doi:10.1057/9780230226203.0275

The New Palgrave Dictionary of Economics Online

compensating differentials

Matthew E. Kahn
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Compensating differentials represent a wage premium for unpleasant aspects of a job. Jobs differ along several dimensions. Some jobs offer generous health insurance benefits. Others entail long hours or may expose workers to physical risks. Some are available only in polluted cities. In equilibrium, labour markets accommodate diversity by establishing wages that tend to make different jobs relatively close substitutes at the margin. Using hedonic wage regression techniques, researchers have estimated the equilibrium implicit market price that workers pay, through lower wages, for working in a more pleasant setting. This technique is widely used by labour and environmental economists.

Keywords

compensating differentials; hedonic wage function; labour economics; Rosen, S.; superstars, economics of; unobserved skill; urban economics; urban environment and quality of life; wage heterogeneity, sources of; wage premium; worker heterogeneity

Article

Compensating differentials represent a wage premium for unpleasant aspects of a job. Jobs differ along a number of dimensions. Some jobs offer generous health insurance benefits. Other jobs entail long hours or may expose workers to physical risks. Some jobs are only available in polluted cities. The theory of compensating differentials is based on the simple premise that there is ‘no free lunch’. In market equilibrium, more unpleasant jobs will offer a wage premium relative to other jobs. Similarly, homes in nicer communities or high-quality-of-life cities will sell for a premium. To quote Sherwin Rosen (2002, p. 2), ‘Markets accommodate diversity by establishing prices that tend to make different things relatively close substitutes at the margin. Adam Smith's insight that market prices tend to equalize their net advantages is fundamental to these problems. If one good has more desirable characteristics than another, the less preferred variety must compensate for its disadvantages by selling at a lower price.’

Defining compensating differentials

Jobs represent tied bundles of attributes. Suppose that a worker gains utility from earning a wage and from a job attribute. This attribute could represent job safety, or total days of vacation, or health insurance benefits. As shown in Figure 1, there are two jobs, A and B. Each job represents a different bundle of a wage and a non-market job-specific amenity level. The two jobs differ: job B is the more pleasant of the two. If all workers have the same utility function, then in equilibrium this representative worker must be indifferent between the two jobs. Thus, job A must pay a higher wage than job B to compensate this worker.
Figure 1

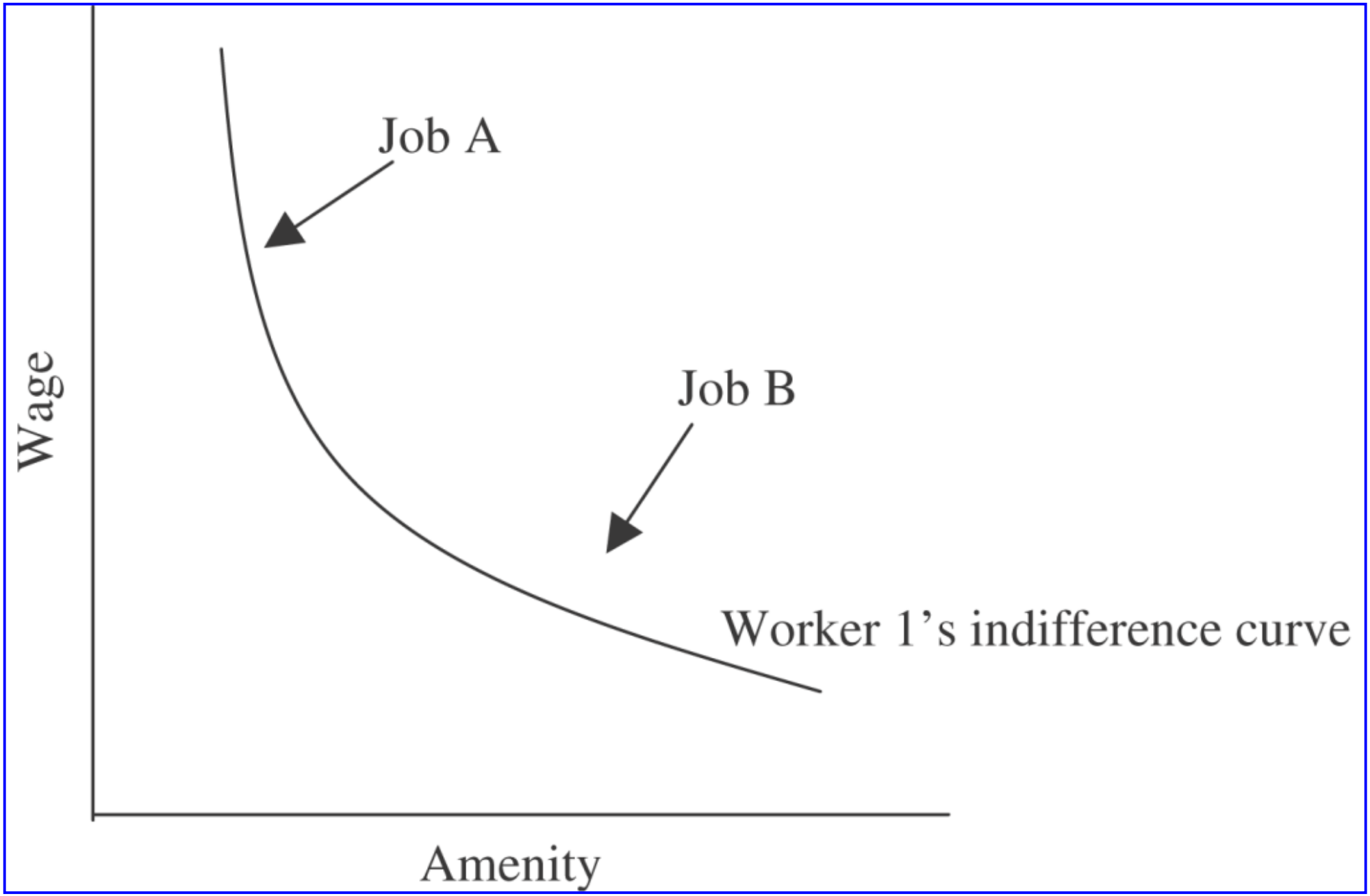

The econometrician can collect data on each job type's wage and amenities. In a more realistic economy where there are many types of jobs that differ with respect to the wage and their amenity level, the representative worker's indifference curve would be sketched out. The slope of the representative worker's indifference curve represents the compensating differential of how much lower a wage this worker would accept in return for a small increase in the job amenity.

To see how worker heterogeneity affects the interpretation about observed compensating differentials, consider the simple extension where we introduce two types of workers. These workers are equally productive but differ with respect to their demand for working in the more pleasant job. In Figure 2, worker 1 values the job amenity more than worker 2. In equilibrium, job A will pay a compensating differential to attract workers to be willing to work in this job. Worker 2 will choose to work in job A while worker 1 will choose to work in job B. Firm A will prefer to hire worker 2 rather than worker 1 because worker 1 requires a larger compensating differential for working in the more unpleasant job. The profit

maximizing firm seeks to minimize its costs of production.

Figure 2

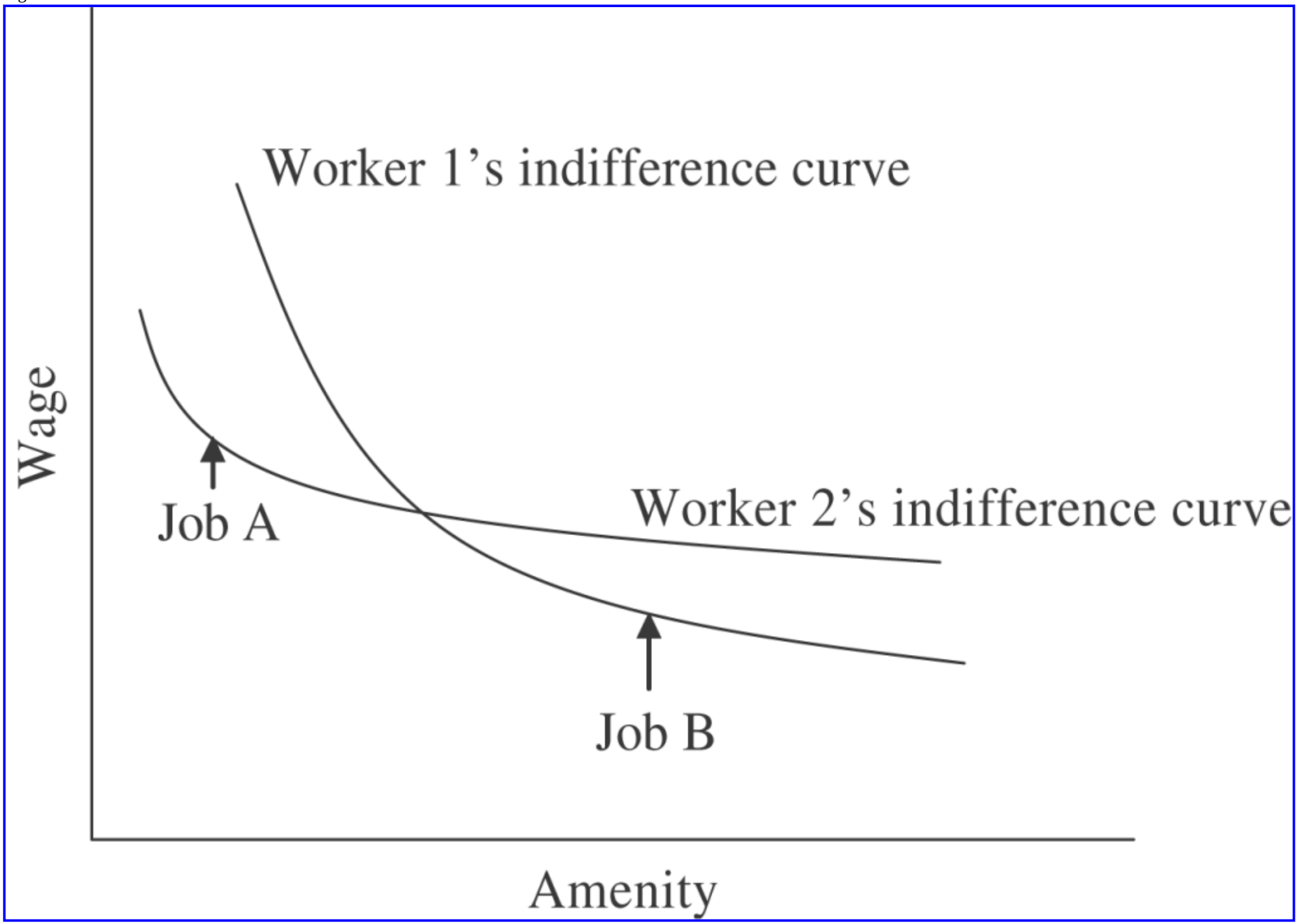

The econometrician will observe the equilibrium wage paid to workers in job A and B. As shown in Figure 2, this equilibrium wage–amenity relationship called the *hedonic wage*

function does not represent either worker 1's or worker 2's indifference curves. Instead, this hedonic wage function represents the envelope of the minimum wage that heterogeneous workers are willing to accept to do a job. This simple example highlights how introducing worker heterogeneity affects inference from observed data (see Rosen, 2002). Figure 2 focuses on just one dimension of worker heterogeneity. The recent compensating differentials literature has explored the consequences of other dimensions of worker heterogeneity such as unobserved skill (IQ, for example) and a worker's ability to self-protect against injury on the job (Hwang, Reed and Hubbard, 1992; Shogren and Stamland, 2002).

Labour econometric applications of compensating differentials theory

An enormous applied econometrics literature has estimated various versions of hedonic wage functions to recover estimates of the marginal valuation of non-market job attributes. One major focus of this research has been to estimate the value of life by measuring how much of a wage premium the marginal worker requires for working in a job with a higher probability of death (Viscusi and Aldy, 2003). Other studies have used hedonic methods to measure the compensating differential for mandated government health insurance benefits (Gruber, 1994).

The standard approach utilizes a large micro-data set. The dependent variable in such a study is a full-time worker's wage in a specific occupation, industry or city. For example, in equation (1) the dependent variable is the log of worker *i*'s wage in industry *j* in year *t*. In an urban application, *j* would refer to a city rather than an industry. The researcher will include a large number of demographic controls, such as age, ethnicity, or education, to ‘standardize’ the worker. If one controls for these factors, the key variables of interest are the *Z*'s in equation (1). In a labour economics application, the *Z* vector may represent a set of job specific attributes (length of day, job risk). In an urban economics application, the *Z* vector may represent attributes of the city where the job is located (climate, pollution, crime).

$$\mathrm{Log}(\mathrm{Wage}_{ijt}) = \gamma_0 + \gamma_1^{*} X_{it} + \gamma_2^{*} Z_{jt} + U_{ijt} \quad (1)$$

Ordinary least squares regression estimates of γ_2 are used to construct measures of the compensating differentials for job tasks and characteristics of employment locations. Estimates of such coefficients have been used to rank city quality of life (see Gyourko and Tracy, 1991) and represent the first stage of the hedonic two-step for recovering demand functions for non-market goods such as air quality or climate (Rosen, 1974; Ekeland, Heckman and Nesheim, 2004).

If the population differs with respect to its tastes for job attributes, then γ_2 can be used to construct a worker's budget constraint. For example, in a job-safety regression if γ_2 equals minus $100 then this means that a one-unit increase in job safety will cost the worker an extra $100 in wages. The rational worker facing this budget constraint will take this trade-off into account when choosing the job that maximizes her utility.

Hedonic estimates of compensating differentials can also be used to bound worker preferences. To return to Figure 2, a lower bound on worker 1's willingness to accept work in risky job A is the equilibrium wage paid to worker 2. Since we know that worker 1 chose the safe job and refused to work in job A at the wage that worker 2 accepted, worker 2's wage offer provides a lower bound (see Rosen, 2002).

The typical hedonic wage regression study estimates eq. (1) using ordinary least squares. This econometric approach will yield consistent estimates of γ_2 if the unobserved determinants of wages (that is, the error term) are uncorrelated with the explanatory variables. What is the error term in this hedonic pricing equation? While a researcher might hope that it represents measurement error in the dependent variable, it is more likely that the error term represents unobserved attributes of the worker and unobserved attributes of the geographical area where the worker lives and works.

Unfortunately for researchers, people self-select where to live and work. A researcher would like to know what wage the same worker would earn in every industry and in every city. In a cosmopolitan city such as New York, superstars of all fields, ranging from Don Trump in real estate to Derek Jeter in baseball, have all chosen to work there. A naive cross-city hedonic researcher would observe these stars living in New York City earning high wages *relative* to observationally identical people in Tulsa, and would conclude, based on the wage regression, that New York City's quality of life must be worse than Tulsa's. Clearly, the problem with this inference is the ‘apples to oranges’ comparison. New York City's amenities are a normal good. The high-skilled earn higher salaries and are attracted to living and working in this city.

Conclusion

A job's wage is not a sufficient statistic for its quality. Coal miners are paid a relatively high wage but the work is dangerous and unpleasant. A major research agenda in labour economics investigates how much people implicitly pay for non-market job attributes. Credible estimates of wage compensating differentials for living in less polluted cities or working in risky industries would greatly aid policy analysis that seeks to measure the benefits of environmental and safety regulation.

See Also

- Roy model
- wage inequality, changes in

Bibliography

Ekeland, I., Heckman, J. and Nesheim, L. 2004. Identification and estimation of hedonic models. *Journal of Political Economy* 112, S60–S109.

Gruber, J. 1994. The incidence of mandated maternity benefits. *American Economic Review* 84, 622–41.

Gyourko, J. and Tracy, J. 1991. The structure of local public finance and the quality of life. *Journal of Political Economy* 91, 774–806.

Hwang, H.-S., Reed, R. and Hubbard, C. 1992. Compensating wage differentials and unobserved productivity. *Journal of Political Economy* 100, 835–58.

Rosen, S. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* 82, 34–55.

Rosen, S. 2002. Markets and diversity. *American Economic Review* 92, 1–15.

Shogren, J. and Stamland, T. 2002. Skill and the value of life. *Journal of Political Economy* 110, 1168–73.

Viscusi, W. and Aldy, J. 2003. The value of statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27, 5–76.

How to cite this article

Kahn, Matthew E. "compensating differentials." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000539> doi:10.1057/9780230226203.0276

The New Palgrave Dictionary of Economics Online

compensation principle

John S. Chipman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The compensation principle holds that one of two possible states constitutes an improvement over the other if the gainers could compensate the losers for their losses and still be at least as well off as in the original state. The conflict between potentiality and actuality – one situation is judged better than another if everybody *could* be made better off in the new situation even though some in fact become worse off – ensures that the compensation principle does not allow for value-free policy decisions.

Keywords

autarky; Baldwin envelope; Barone, E.; Bergson–Samuelson social-welfare function; Bickerdike, C. F.; Cairnes, J. E.; compensation principle; competitive equilibrium; consumer surplus; cost–benefit analysis; deadweight loss; Debreu, G.; Dupuit, A.-J.-E. J.; exchange; excise taxes; free trade; fundamental theorem of welfare economics; gains from trade; game theory; general equilibrium; Hicks, J. R.; Hotelling, H.; ideal money; income-compensation function; interpersonal utility comparisons; Kaldor, N.; Kuznets, S.; Lerner, A. P.; Little, I. M. D.; lump sum taxes; Marshall, A.; Mishan, E. J.; monopoly; new welfare economics; optimal tariffs; Pantaleoni, M.; Pareto optimality; partial equilibrium; Pigou, A. C.; proportional income tax; revealed preference; Ricardo, D.; Robbins, L. C.; Samuelson, P. A.; Scitovsky indifference surface; Scitovsky, T.; taxation of income; transferable utility; utility-possibility frontier; value judgements; Viner, J.

Article

The term ‘compensation principle’ refers to the principle that, in comparing two alternative states in which a given community of persons might find itself, one of the states constitutes an improvement over the other (in the weak sense including equivalence) if it is possible for the gainers to compensate the losers for their losses and still be at least as well off as in the original state.
If the hypothetical compensation is actually carried out, the principle reduces to the Pareto criterion: all

are at least as well off, in one state compared to the other. There is no need to invoke the compensation principle in such a case. On the other hand, if the principle is used to compare two unique alternative states in which a community might find itself, neither of which is Pareto-superior to the other, the principle seems quite arbitrary unless interpreted in a broader context. There is a sense in which one person might be said to be basically healthier than another even though, at the particular moment, such a person might have a cold and the other one not. The compensation principle is usually used to make comparisons in this sense; one state of the economy is sounder, healthier, more robust, or has greater productive potential, than another. What this implies is that states under comparison are usually not unique, singleton states but composite ones, or sets of states. Formally, the objects being compared are usually sets of commodity bundles that could be made available to the aggregate of consumers, described in the literature as ‘situations’ in contrast to single ‘points’ in such sets (cf. Baldwin, 1954).

Examples of comparisons in which the compensation principle is typically used are those between (*a*) a perfectly competitive system of industrial organization and an imperfectly competitive one; (*b*) free trade and no trade (or restricted trade); (*c*) the state of an economy before and after a war, or depression, or change in productive techniques. Most but not all of these types of comparisons are relevant to policy decisions; and the policy decisions are usually not of an ad hoc type (for which the compensation principle would hardly be appropriate) but of a fundamental nature concerning the underlying system of industrial organization and trade.

Inasmuch as the principle can be applied without the need to make interpersonal comparisons, some of its more ardent proponents have maintained that it is ‘value-free’. However, there can be no doubt that it does require acceptance of some value judgements, since the Pareto criterion itself constitutes one – albeit a widely accepted one. Another value judgement implicit in the principle as it has usually been applied is that each individual is the best judge of his or her own well-being; while also quite widely accepted, this one is obviously controversial, and in fact government policy measures are often called for precisely in those instances where it is clearly an untenable assumption. But the most important and controversial way in which value judgements enter into the compensation principle is in the conflict between potentiality and actuality: one situation is judged better than another if everybody *could* be made better off in the new situation even though some in fact become worse off. This lacuna in the principle has led Little (1950) and Mishan (1969) to formulations in which compensation tests are combined with explicit distributional value judgements, and Samuelson (1947; 1956) into a full-fledged ethical system in which compensation is carried out to the extent that the ethical norms dictate.

In many applications the compensation principle is difficult to formulate in a precise manner unless one assumes absence of externalities in consumption, so it is usually formulated (but with some notable exceptions – for example, Coase, 1960) under the assumption that each person's welfare depends only on his or her own consumption of goods and services. In most applications, the data available for making comparisons are, almost inevitably, limited to aggregative information on the actual state of the economy in each situation; much of the work in applying the principle therefore consists in using economic theory to make inferences from the actual observations concerning underlying conditions in the economy. By its nature, the compensation principle is limited in its application to comparing alternative states (or sets of states) of a given community of individuals; thus, it cannot be applied (at least not literally) to historical comparisons of a country's condition over time (since the population has changed) or to comparisons of the living conditions of different countries (since the populations are different). However, extensions of the principle to cover such comparisons are possible provided suitable additional empirical assumptions

and value judgements are accepted; for example, if all individuals are assumed to have identical preferences, one could ask whether there exists a redistribution of income in each period (or country) such that each individual in the one situation would be better off than each individual in the other. This would obviously entail additional value judgements along with the additional empirical assumptions.

1 Historical development: from Dupuit to Hotelling

The compensation principle may be traced back to Dupuit (1844, pp. 359–60; Arrow and Scitovsky, 1969, p. 272) and Marshall (1890, p. 447; 1920, p. 467) who used the concept of consumers' surplus to compare the losses of consumers (say from a bridge toll or an excise tax) with the gains to the government. The demonstration that the former exceed the latter, so that consumers cannot be compensated for their losses out of the government revenues, provided a convincing case for the superiority of income tax to an excise tax (or for the superiority of government subsidization of bridge construction to its financing of it by tolls), and at the same time provided scientific prestige and great intuitive appeal to a method that was able to reach such a definitive conclusion and furnish a measure of the ‘deadweight loss’.

While Dupuit and Marshall used partial-equilibrium analysis, Pareto (1894, p. 58) was the first to introduce the concept into general-equilibrium theory, in the course of an article devoted to proving the optimality of competitive equilibrium. In the first part of this article (summarized by Sanger, 1895), Pareto used as his criterion of optimality the sum of individual utilities; in the second part, however – acknowledging the criticisms and suggestions of Pantaleoni and Barone (both admirers of Marshall, which Pareto was not) – he reformulated the problem so as to sum not the utilities of different consumers but the quantities they consume. His criterion of optimality (1894, p. 60) was that it should be impossible for one person to gain without another losing – ‘Pareto optimality’ – a criterion that had also been introduced by Marshall (1890, pp. 449–50; 1920, pp. 470–1). A more refined version of Pareto's argument later appeared in the *Cours* (Pareto, 1896–7, vol. 1, pp. 256–62; vol. 2, pp. 88–94).

The proposition formulated by Pareto (1894) anticipated what has now come to be known as the ‘fundamental theorem of welfare economics’, namely, that every competitive equilibrium is Pareto optimal and, conversely, every Pareto optimum can be sustained by a competitive equilibrium. Pareto considered the problem faced by a socialist state striving to attain an outcome in which it was impossible for one person to gain without another losing. The Ministry of Justice would concern itself with problems of income distribution, and the Ministry of Production with resource allocation and choice of production coefficients. A weakness of Pareto's argument was that he assumed a price system already to be established – perhaps our socialist state needs the prices of its capitalist neighbours to guide it. Pareto further assumed that each individual's budget constraint was adjusted by the addition of a parameter (a lump-sum subsidy or tax) controlled by the government. The government's objective was to maximize the sum of these parameters, which he showed was equal to aggregate profit – the value of commodities consumed less the value of factor services supplied, equal to the value of firms' output less the outlay on their factor inputs. If it were possible to increase all the parameters, the existing situation would not be Pareto optimal; if their sum were a maximum, it would not be possible to increase one of them without decreasing another, and the outcome would be Pareto optimal. Pareto showed that maximization of aggregate profit at the given prices, subject to the resource-allocation and production-function constraints, would lead to cost-minimization and zero profits. (For mathematical details of Pareto's arguments see Chipman, 1976, pp. 88–92.) Pareto summarized this result by stating (1896–7, vol. 2, p. 94):

> Free competition of entrepreneurs yields the same values for the production coefficients as would be obtained by determining them by the condition that commodity outputs should be chosen in such a way that, for some appropriate distribution, maximum ophelimity would be achieved for each individual in society.

The last clause was Pareto's unfortunately awkward way of stating the criterion of Pareto optimality. Barone (1908), who had originally spurred Pareto on to this line of argument, developed it further himself. He noted that a competitive equilibrium has the property that aggregate profit is at a maximum at the equilibrium prices, hence, for any feasible departure from this equilibrium, valuing consumption and factor services at the equilibrium prices, some individuals may gain and others will lose, the losses outweighing the gains so that, even if the gainers part with all their gains, the rest will still be worse off than originally. (Barone used what is now known as the criterion of revealed preference to make inferences concerning preferences from data on prices and incomes.) Such a state was described by Pareto and Barone as ‘destruction of wealth’, and its measure by aggregate income loss at the competitive-equilibrium prices provided an alternative to the deadweight loss considered by Dupuit and Marshall. Barone (1908) also related his arguments to those of Marshallian consumers'-surplus analysis.

Lerner (1934) invoked the compensation principle in his proposed method for measuring monopoly power, describing it as ‘a loss to the consumer which is not balanced by any gain reaped by the monopolist’. In this paper Lerner also formulated, apparently independently, the concept of Pareto optimality.

Hotelling (1938) made a noteworthy contribution by providing an alternative demonstration of the inferiority of excise taxes to income taxes, using the compensation principle directly. He considered a single individual consuming n commodities in amounts q_j and facing market prices p_j. Prior to the imposition of the excise taxes (or tolls), the individual consumes a bundle q^0 at prices p^0 and income (or fixed component of income) m^0, which maximizes a utility function $U(q)$ subject to the budget constraint $p^0 \cdot q = m^0$. Subsequent to the introduction of taxes, market (tax-inclusive) prices and after-tax income are p^1 and m^1 respectively, and a bundle q^1 is chosen which maximizes $U(q)$ subject to $p^1 \cdot q = m^1$. The government collects $r = \left(p^1 - p^0\right) \cdot q^1 - \left(m^1 - m^0\right)$ in revenues. Since the government is assumed to collect $\left(p_j^1 - p_j^0\right) \cdot q_j^1$ in taxes on commodity j, p_j^0 must be identified with the production cost after the tax (as well as with the market price=production cost before the tax); this is a fairly restrictive assumption, since it implies that the tax does not affect production costs. (In this respect Hotelling's treatment is less general than Dupuit's and Marshall's, involving infinite elasticities of supply.) We may denote the ad valorem excise-tax rate on commodity j by $t_j = p_j^1 / p_j^0 - 1$, and a proportional income-tax rate by $t_0 = 1 - m^1 / m^0$ (negative taxes are interpreted as subsidies). The government's revenues are

$$r = \sum_{j=1}^{n} t_j p_j^0 q_j^1 + t_0 m^0 = 0,$$

assumed zero since the government distributes the total proceeds of these excise taxes back to the consumer (or taxes the consumer if these are negative). The consumer's budget constraint after the imposition of the taxes is

$$\sum_{j=1}^{n}(1+t_j)p_j^0 q_j^1 = (1-t_0)m^0.$$

These two equations together imply that q^1 satisfies the budget constraint $p^0 \cdot q^1 = m^0$, hence q^1 was in the consumer's original budget set. Therefore, setting aside the ‘infinitely improbable …contingency’ that q^0 and q^1 lie on the same indifference surface, Hotelling concluded (1938, p. 252) that ‘if a person must pay a certain sum of money in taxes, his satisfaction will be greater if the levy is made directly on him as a fixed amount than if it is made through a system of excise taxes which he can to some extent avoid by rearranging his production and consumption’.

Unfortunately Hotelling overlooked the fact that if $t_j = t$ for all j then the government's budget constraint implies $p^0 \cdot q^1 = -m^0 t_0 / t$, whence $t_0 = -t$ and $q^1 = q^0$. That is, a system of uniform ad valorem excise taxes is equivalent to a proportional income tax. This was pointed out by Frisch (1939) and accepted by Hotelling (1939). As Frisch made clear, what Hotelling really proved was the non-optimality of a system of non-proportional excise taxes or subsidies when selling prices are given. If these selling prices are equal to marginal costs, Hotelling's theorem shows that market prices should be proportional to marginal costs. Since incomes are fixed in Hotelling's formulation, income taxes may be regarded as lump-sum taxes. If institutional consideration make excise taxes impossible for one commodity (say leisure), then they must be zero for all commodities and optimality requires that prices be equal to marginal costs. (For a less charitable interpretation of Hotelling's contribution see Silberberg, 1980.)

Hotelling went on to assert that his proposition could be extended to many consumers (though no details or proof were provided), and he proceeded to examine the consumers'-surplus measure of loss $\frac{1}{2}\left(p^1 - p^0\right)\cdot\left(q^1 - q^0\right) = \frac{1}{2}T\,p^0 \cdot \left(q^1 - q^0\right)$ (where T is a diagonal matrix of excise-tax rates t_j). He also made some general observations (1938, p. 267) that, to this day, constitute what is probably the best statement to be found of the philosophy underlying the compensation principle.

2 The years of the new welfare economics

In the cases to which the compensation principle was applied by Dupuit, Marshall, Lerner and Hotelling, compensation was made between the class of consumers on the one hand and a government or a monopolist on the other. While Pareto and Barone had discussed compensation between different classes of consumers (as had Hotelling in his general remarks) their work was unknown to English-speaking economists until the publication in 1935 of the English translation of Barone's 1908 work. Even this

seems not to have struck home, however, since Kaldor (1939) cited passages from Harrod (1938) and Robbins (1938) to the effect that, since movement towards free trade would affect different classes differently, no scientific statement could be made concerning the beneficial effect of free trade without making interpersonal comparisons of utility.

Kaldor (1939) proceeded to sketch an argument to the effect that removal of an import duty (using the classical example of repeal of the Corn Laws) would result in a situation in which the losses incurred by the landlords could be compensated by the gains (through lower import prices) obtained by the other consumers. Such an argument cannot be correct, however, since, as Kaldor (1940) pointed out only a year later, it follows from Bickerdike's theory of optimal tariffs that a country can gain from the imposition of a sufficiently small duty, and, as Graaff (1949) and others later demonstrated, the compensation principle can be used to show that, with suitable compensation, all persons can gain. Unless the rate of corn duty was above the optimal tariff rate, the opposite conclusion would follow to that indicated by Kaldor (1939).

A previous attempt by Pareto (1895) to show by means of the compensation principle that a tariff would lead to ‘destruction of wealth’ was defective, since he assumed trade to be balanced in domestic prices and thus he failed to take account of the improvement in the terms of trade and the beneficial effect of the tariff revenues.

Other attempts prior to 1939 to make the case for free trade suffered from vagueness both in specifying the criterion of gain and in specifying the alternative with which free trade was being compared. Ricardo (1815, p. 25) stated: ‘There are two ways in which a country may be benefited by trade – one by increase of the general rate of profits … the other by the abundance of commodities, and by a fall in their exchangeable value, in which the whole community participate’. According to Cairnes (1874, p. 418), ‘the true criterion of the gain on foreign trade [is] the degree in which it cheapens commodities, and renders them more abundant’. A hint of a compensation principle is found in Viner (1937, pp. 533–4):

> free trade … necessarily makes *available* to the community *as a whole* a greater physical real income in the form of more of *all* commodities, and … the state … can, by appropriate supplementary legislation, make certain that removal of duties shall result in more of *every* commodity for *every* class of the community.

Like Kaldor's statement, this is formally incorrect; but it was sufficiently suggestive to stimulate Samuelson (1939) into providing a formal proof of a gains-from-trade theorem, albeit under very restrictive assumptions.

Samuelson (1939) assumed that an open economy had a locus $\varphi(y, l) = 0$ of efficient combinations of outputs *y* and (variable) factor services *l*, and asserted that vectors of prices *p* and factor rentals *w* in competitive equilibrium would be such that aggregate profit $p \cdot y - w \cdot l$ is a maximum. This is the same as the proposition of Pareto (1894), and Barone (1908) referred to above. Letting *x* denote the bundle of commodities consumed, under both (balanced) free trade and autarky the budget equation $p \cdot x = p \cdot y$ holds. Letting superscripts 0 and 1 denote equilibrium values under autarky and free trade respectively, it follows that

$$p^1 \cdot x^1 - w^1 \cdot l^1 \geq p^1 \cdot x^0 - w^1 \cdot l^0.$$

Assuming all *N* individuals to be identical in their preferences and ownership of factors, and dividing this inequality through by the number of individuals, it states that each person chooses $\left(x^1 / N, l^1 / N\right)$ under free trade when $\left(x^0 / N, l^0 / N\right)$ is available, hence $\left(\text{if } p^1 \neq p^0\right)$ each person prefers $\left(x^1 / N, l^1 / N\right)$ to $\left(x^0 / N, l^0 / N\right)$. Therefore free trade is Pareto-superior to autarky.

Samuelson went on to assert (1939, p. 204) that, if the assumption of identical individuals is dropped, then, although it could no longer be said that each individual was better off under free trade, ‘it would always be possible for those who desired trade to buy off those opposed to trade, with the result that all could be made better off’. This argument went unchallenged until Olsen (1958) pointed out that, if compensation were paid from gainers to losers, a new equilibrium price constellation p^1 would result, and the argument no longer follows. For this reason Samuelson's 1939 results has come to be known as the gains-from-trade theorem for the ‘small-country case’, though this interpretation was not suggested by Samuelson at the time. But this description of Samuelson's result is inaccurate. Generalizing his argument we can say that if $\left(x_i^t, l_i^t\right)$ are the allocations of $\left(x^t, l^t\right)$ to individual *i*, where $\Sigma_{i=1}^N x_i^t = x^t$ and $\Sigma_{i=1}^N l_i^t = l^t$, the given the allocations $\left(x_i^1, l_i^1\right)$ of $\left(x^1, l^1\right)$ under free trade one can find Pareto-optimal allocations $\left(x_i^0, l_i^0\right)$ of $\left(y^0, l^0\right)$ under autarky such that

$$p^1 \cdot x_i^1 - w^1 \cdot l_i^1 \geq p^1 \cdot x_i^0 - w^1 \cdot l_i^0 \quad \text{for} \quad i = 1, 2, \ldots, N.$$

This proves that for *any* free-trade equilibrium it is possible to find a weakly Pareto-inferior Pareto-optimal autarky equilibrium. It does not prove the obverse proposition that for any autarky equilibrium it is possible to find a weakly Pareto-superior free-trade equilibrium. A general gains-from-trade theorem was therefore yet to be established, but Samuelson had provided an important first step.

Hicks (1939) ushered in the ‘new welfare economics’ with a synthesis building on Hotelling (1938) and Kaldor (1939) and based on the compensation principle, making it possible, according to him, to make policy proposals in favour of economic efficiency which were free of value judgements. Hicks's most original contribution (Hicks, 1940) was his attempt to apply the compensation principle to data on a country's real national income. This was a natural thing to try to do, since Pigou's (1920) main work was devoted to evaluating a country's welfare by national-income comparison, and it was largely Pigou's resort to interpersonal comparisons in order to justify this that was the object of Robbins's (1938) criticism. Hicks's (1940) basic tool was the ‘revealed-preference’ comparison which had been employed by Barone (1908) and Hotelling (1938). If observations are available at times 0 and 1 of a country's national income

in period-1 prices, and it is recorded that $p^1 \cdot y^1 \geq p^1 \cdot y^0$ (where p^t, y^t are vectors of prices and outputs at time t), what can be inferred? In the first place, to make any headway one must assume that the observed situations are competitive equilibria. Let us define an allocation of a commodity bundle x as an $N \times n$ matrix X whose ith row, x_i, is the bundle of n commodities allocated to individual i, and whose row sum $\sigma(X) = \Sigma_{i=1}^{N} x_i$ is equal to x. As between two bundles x_i^0, x_i^1 consumed by individual i, let us define $x_i^1 R_i x_i^0$ to mean that x_i^1 is preferred or indifferent to x_i^0 by individual i, where R_i is a continuous, convex, monotonic total order, with P_i denoting strict preference and I_i indifference. (This relation assumes the absence of externalities in consumption.) Finally, let $X^1 R X^0 \left(\text{resp. } X^1 P X^0\right)$ mean that X^1 is weakly (resp. strictly) Pareto-superior to X^0 (i.e. $x_i^1 R_i x_i^0$ for all i, resp. $x_i^1 R_i x_i^0$ for all i and $x_i^1 P_i x_i^0$ for some i). Then, from the real-income comparison $p^1 \cdot y^1 \geq p^1 \cdot y^0$, Hicks noted that there does not exist an allocation X of y^0 that is weakly Pareto-superior to the actual allocation X^1 of y^1. This follows from the same argument that establishes the Pareto optimality of the assumed competitive equilibrium in period 1. The non-existence of an allocation X of y^0 such that XRX^1, where $\sigma\left(X^1\right) = y^1$, constituted for Hicks the definition of an ‘increase in real social income’.

Kuznets (1948) pointed out by an example that, in the case considered by Hicks, it could also be true that there is no allocation X of y^1 which is weakly Pareto superior to the actual allocation X^0 of y^0. Accordingly he suggested that Hicks's criterion be supplemented by the condition that there should exist an allocation $\bar{X}$ of y^1 that is weakly Pareto superior to the actual allocation X^0 of y^0. But while the latter criterion implies $p^0 \cdot y^1 \geq p^0 \cdot y^0$, it is not implied by it, so a national-income comparison using current and base prices would still not yield Kuznets's criterion.

Kuznets's criticism of Hicks was similar to the objection raised by Scitovsky (1941) to the criterion proposed by Kaldor (1939). According to Scitovsky's interpretation of Kaldor, an allocation X^1 of y^1 is better than an allocation X^0 of y^0, if there exists a reallocation $\bar{X}^1$ of y^1, which is Pareto superior to X^0. Scitovsky objected that this gave preference to the *status quo ante*, and besides, he pointed out that the criterion was internally inconsistent in the sense that it allowed two such pairs $\left(X^t, y^t\right)$ to be superior to each other. He therefore proposed that Kaldor's test be supplemented by the criterion that there exist a reallocation $\bar{X}^0$ of y^0 that is Pareto inferior to X^1.

The literature on ‘compensation tests’ suffered from ambiguity as to the domain of definition of the relations and internal inconsistency of the relations. It was pointed out by Gorman (1955) that the relations were intransitive. It was shown in Chipman and Moore (1978) that the Hicks–Kuznets and Scitovsky double criteria, as well as the national-income comparisons in terms of base- and current-year prices, could lead to cycles of three competitive equilibria each superior to its successor.

The definitive contribution to the subject of national-income comparisons was that of Samuelson (1950) who introduced what Chipman and Moore (1971) described as the ‘Kaldor–Hicks–Samuelson (KHS) ordering’. The objects under comparison in this approach are sets Y of commodity bundles y, e.g. production-possibility sets. Letting $A(Y)$ denote the set of allocation matrices X such that $\sigma(X) \in Y$, this ordering is defined by

$$Y^1 >_R Y^0 \Leftrightarrow \left[\forall X^0 \in A\left(Y^0\right)\right]\left[\exists X^1 \in A\left(Y^1\right)\right] X^1 R X^0.$$

In words, Y^1, is potentially superior to Y^0 if, for all allocations of commodity bundles in Y^0, there exists a (weakly) Pareto superior allocation of a commodity bundle in Y^1. This is a reflexive and transitive relation; it also satisfies the condition that $Y^0 \subseteq Y^1$ implies $Y^1 >_R Y^0$. Samuelson also introduced the important concept of a utility-possibility frontier, which is the relative boundary of a utility-possibility set $U(Y, R; f)$; this in turn is a set of N-tuples of individual utilities, $u=f(X)$, for some $X \in A(Y)$, where f is an N-tuple of positive-valued utility functions representing R. If the sets Y are 'disposable' (that is, containing for every $y \in Y$ the bundles y' with $0 \leq y' \leq y$), and the R_i continuous and monotonic, then the utility-possibility sets are also disposable. If Y is non-empty, compact disposable, and convex, and the R_i are continuous, monotonic, and convex, then, provided the f_i are continuous and concave, $U(Y,R;f)$ is non-empty, compact, and convex (cf. Chipman and Moore, 1971, p. 24). If the f_i are only quasi-concave and not concave, $U(Y,R;f)$ need not be convex (cf. Kannai and Mantel, 1978). The KHS ordering among consumption-possibility sets translates into set-inclusion of the corresponding utility-possibility sets. Samuelson (1959, p. 10) gave an example of a case of crossing utility-possibility frontiers in which $X^2 \in A\left(Y^2\right)$ was Pareto superior to $X^1 \in A\left(Y^1\right)$ yet Y^1 would be ranked higher than Y^2 in terms of some value judgement. This established that the 'compensation tests' were not 'relatively *wertfrei*'.

Another approach was followed by Chipman and Moore (1973; 1976a), who asked the following question: if competitive equilibria $\left(X^t, y^t, p^t\right)$ are observed satisfying $p^1 \cdot y^1 \geq p^1 \cdot y^0$ and $p^0 \cdot y^1 \geq p^0 \cdot y^0$, where $y^t \in Y_t$ for $t = 0, 1,$ under what conditions on preferences must this imply that $Y^1 >_R Y^0$? For the case $Y^t = \left\{y^t\right\}$ they showed that the preference relations R_i must be identical and homothetic. This is a global result; with positive consumptions of all commodities the condition could no doubt be weakened to the aggregation criterion of Antonelli (1886), Gorman (1953), and Nataf (1953), namely, that consumer i's demand for commodity j have the form

$$x_{ij} = a_{ij}(p) + b_j(p) m_i$$

where m_i is consumer i's income.

Samuelson (1956) applied the compensation principle in a striking way in his proposed alternative to the new welfare economics. He discovered that, if a social-welfare function has the separable form $W[f(x)]$, then a social utility function $f_W(x) = \max\{W[f(x)] : X \in A(x)\}$ has the property that it can be achieved

in a decentralized manner by means of an income-distribution policy assigning individual shares of aggregate income as functions of prices and aggregate income. The first complete proof of this result was presented in Chipman and Moore (1972) (see also Chipman and Moore, 1979; Chipman, 1982). The main tool of analysis used was the concept of a Scitovsky indifference surface (Scitovsky, 1942) which is defined as the boundary of the set $\Sigma_{i=1}^{N} R_i x_i$, where $R_i x_i$ is the set of all commodity bundles preferred or indifferent to x_i by individual i. This set is necessarily a subset of the set $R_w x$ of aggregate bundles preferred or indifferent to x by the Samuelson social ordering. In a competitive equilibrium the aggregate consumption bundle minimizes aggregate expenditure at the equilibrium prices over both sets, hence the bundle x_i minimizes each individual's expenditure over $R_i x_i$ (cf. Koopmans, 1957, pp. 12–13).

3 Gains from trade and optimal tariffs

The new tools developed by Scitovsky (1942) and Samuelson (1950; 1956) made possible a rigorous proof of a gains-from-trade theorem, as well as of the proposition that a country could gain by a tariff. Kemp (1962) noted that Samuelson's 1939 theorem implied that for any point on the free-trade utility-possibility frontier, the autarky utility-possibility frontier must pass below it; he reasoned that, as a result, for any point on the autarky utility-possibility frontier, the free-trade utility-possibility frontier must pass above it. If this argument can be accepted, it follows that for every allocation $X^0 \in A(Y^0)$ where Y^0 is the autarkic production-possibility set, there exists a (weakly) Pareto-superior allocation $X^1 \in A(Y^1)$ where Y^1 is the free-trade consumption-possibility set. Then free trade is superior to autarky by Samuelson's 1950 criterion.

The trouble with this argument, however, is that it requires that one can define a free-trade utility-possibility frontier (or consumption-possibility frontier) with the strong topological property of homeomorphism to the (N–1)-dimensional unit simplex (intuitively, absence of ‘holes’). That this need not always be possible, was shown by Otani (1972, p. 149), and indeed admitted by Kemp and Wan (1972, p. 513). It is always possible if world prices are fixed, beyond our country's control. In that case the free-trade consumption-possibility set Y^1 is the budget set enclosing the production-possibility set Y^0 (cf. Samuelson, 1962, p. 821), and the gains-from-trade theorem follows immediately from the property $Y^1 \supseteq Y^0 \Rightarrow Y^1 >_R Y^0$. In similar fashion the famous ‘Baldwin envelope’ (Baldwin, 1948) defines a well-behaved consumption-possibility set containing the production-possibility set, from which one can prove the superiority of restricted trade (with an optimal tariff) to autarky (cf. Samuelson, 1962).

For the general case in which a country can influence world prices, a method was shown by Kenen (1957). If all but 1 of the N individuals are constrained to have the same level of satisfaction under trade as achieved under autarky, a net production-possibility set can be constructed which indicates the amount available for the Nth person. It remains only to show that the Nth person will gain from a movement from autarky to free trade. A similar approach was indicated by Vanek (1964).

Grandmont and McFadden (1972) and Chipman and Moore (1972) both used the concept of an income-distribution policy to establish the gains-from-trade theorem. In Chipman and Moore this policy was chosen to be one that maximizes a separable Bergson–Samuelson social-welfare function. A standard

argument is used to show that social utility is at least as high under free trade as under autarky. It remains to show that a function $W(u)$ can be chosen so that the corresponding distribution policy ensures that an increase in social utility implies an increase in each individual's utility. This is achieved by choice of $W(u) = \min_i \left(u_i - u_i^0\right) / c_i$ where $c_i > 0$ and u_i^0 is the level of utility achieved by individual *i* under autarky.

4 General-equilibrium theory

The compensation principle is used in the proof of the theorem that every competitive equilibrium is Pareto-optimal (Arrow, 1952, pp. 516, 519; Koopmans, 1957, p. 49; Debreu, 1959, pp. 94–5), in the sense that arbitrary allocations of feasible output bundles among consumers are assumed possible, regardless of resource-ownership constraints. A pair $\left(X^0, p^0\right)$ is a competitive equilibrium for the production-possibility set *Y* if X^0RX for all $X \in A(Y)$ satisfying $Xp^0 \leq X^0p^0$ and $Y^0p^0 \geq Yp^0$ for all $Y \in Y$, where $Y^0 = \sigma\left(X^0\right) \in Y$. Pareto-optimality means that one cannot find an $X \in A(Y)$ such that XPX^0. The proof is by contradiction: XPX^0 implies $Xp^0 \geq X^0p^0$ (the vector inequality being weak in all components and strict in at least one) hence taking column sums, $Yp^0 > Y^0p^0$.

The converse theorem, that every Pareto optimum can be sustained by a competitive equilibrium, requires stronger assumptions which are awkward to state (cf. Arrow, 1952, p. 518; Koopmans, 1957, p. 50; Debreu, 1959, p. 95). The basic idea of the proof (Koopmans, 1957, pp. 50–52; Debreu, 1959, p. 96) can be sketched in terms of the concept of a Scitovsky (1942) indifference surface. If X^0 is a Pareto-optimal allocation for a closed, convex production-possibility set *Y*, then the interior of the Scitovsky set of X^0 can be written $P_k x_k^0 + \sum_{i \neq k} R_i x_i^0$ for some *k*. Defining the allocation X^1 by $x_k^1 P_k x_k^0$ and $x_i^1 R_i x_i^0$ for $i \neq k$ we have X^1PX^0 hence $X^1 \notin A(Y)$. Therefore the interior of the Scitovsky set does not intersect *Y*, and these convex sets can be separated by a hyperplane defining the equilibrium prices. It is then verified that at these prices the properties of a competitive equilibrium are satisfied.

Debreu (1954, p. 590) introduced an alternative equilibrium concept according to which the condition that consumer preferences be maximized subject to their budget constraints was replaced by the condition that consumer expenditures be minimized subject to the constraints that the bundles considered be at least as desirable as the equilibrium bundles. (The second of the above theorems follows more easily under this alternative definition.) For a given set of positive-valued utility functions representing consumer preferences, Arrow and Hahn (1971, p. 108) called this a ‘compensated equilibrium’. As a means of proving existence of the latter they studied the utility-possibility frontier or ‘Pareto frontier’ (1971, p. 96), and obtained a new proof of the result of Chipman and Moore (1971) that the set of Pareto-optimal allocations *X* of *Y* (the ‘contract curve’) and the utility-possibility frontier are topologically homeomorphic to the unit simplex of dimension one less than the number of individuals. These results were further developed by Moore (1975).

5 Cost – benefit analysis

Hicks (1941, p. 112) made an interesting distinction between two tasks of welfare economics: (1) the study of (Pareto)-optimal organizations of the economy and (2) the study of deviations from such optima. More precisely, the first was concerned with when there was a deviation and the second with the size of the deviation. He also identified these two tasks with general- and partial-equilibrium analysis respectively, although there appears to be no justification for this other than the historical accident that consumers' surplus developed as a partial-equilibrium tool. He remarked that consumers' surplus is not needed for the first task, since lack of fulfillment of the proportionality between marginal utilities and marginal costs provides the needed information immediately. For the second task, he was not content with a ranking of the non-optimal states, but with measuring the size of their deviations from optimality, which of course would provide such a ranking. Thus, the staunch ordinalist in consumer theory became an equally ardent cardinalist in consumer theory.

Hicks's concepts of compensating and equivalent variation (Hicks, 1942) may most conveniently be defined in terms of the minimum-income or income-compensation functions of McKenzie (1957) and Hurwicz and Uzawa (1971). Denoting the *i*th consumer's demand function by $x_i = h_i(p, m_i)$ (where x_i and p are n-vectors), and defining the indirect preference relation R_i^* by $\left(p^0, m_i^0\right) R_i^* \left(p^1, m_i^1\right)$ if and only if $h_i\left(p^0, m_i^0\right) R_i h_i\left(p^1, m_i^1\right)$, the income-compensation function is defined by

$$\mu_i\left(p;\ p^0, m_i^0\right) = \inf\left\{m_i : (p, m_i) R_i^* \left(p^0, m_i^0\right)\right\}.$$

Following Chipman and Moore (1980b), the generalized compensating variation in going from $\left(p^0, m_i^0\right)$ to (p, m_i) is defined as

$$C_i\left(p, m_i;\ p^0, m_i^0\right) = m_i - \mu_i\left(p;\ p^0, m_i^0\right)$$

and the generalized equivalent variation by

$$E_i\left(p, m_i;\ p^0, m_i^0\right) = \mu_i\left(p^0;\ p, m_i\right) - m_i^0.$$

These reduce to Hicks's concepts when $m_i = m_i^0$.

The compensating variation expresses for each consumer the amount of money income he or she would be

willing to give up (or the negative of the amount by which he or she would have to be compensated), at the new prices, to make up for the change in prices and income. One of the reasons for the great appeal of the concept is that these are amounts that can be added up over the set of consumers. In Hicks's words (1942, p. 127):

> the general test for a particular reform being an *improvement* is that the gainers should gain sufficiently for them to be able to compensate the losers and still remain gainers on balance. This test would be carried out by striking the balance of the Compensating Variations.

Denoting by *m'* the vector of *N* incomes in state *t*, and by M^t their sum, we can define a dual potential-improvement ordering between pairs of price-income pairs (p^t, M^t) as follows. Let $A^*(p, M)$ be the set of (*n*+*N*)-tuples (*p*, *m*) such that $\Sigma_{i=1}^{N} m_i = M$, and let R^* be the relation such that $(p^0, m^0) R^* (p^1, m^1)$ if and only if $(p^0, m_i^0) R_i^* (p^1, m_i^1)$ for $i = 1, 2, \ldots, N$. Then we define the dual KHS relation $>_{R^*}$ by

$$\left(p^0, M^0\right) >_{R^*} \left(p^1, M^1\right) \Leftrightarrow \left[\forall\left(p', m'\right) \in A^*\left(p^1, M^1\right)(\exists (p, m) \in A^*\left(p^0, M^0\right)\right] : (p, m) R^* \left(p', m'\right).$$

Choosing price-income pairs (p^0, m^0) and (p^1, m^1) satisfying this definition, since $\mu_i(p^t; p, m_i)$ is an indirect utility function representing R_i^* for $t = 0$ or 1, we have $\mu_i(p^t; p^0, m_i^0) \geq \mu_i(p^t; p^1, m_i^1)$ for all individuals *i*, hence

$$M^0 = \sum_{i=1}^{N} m_i^0 = \sum_{i=1}^{N} \mu_i\left(p^0; p^0, m_i^0\right) \geqq \sum_{i=1}^{N} \mu_i\left(p^0; p^1, m_i^1\right)$$

so one obtains a multi-consumer analogue to the compensating variation from the formula

$$M^0 - \sum_{i=1}^{N} \mu_i\left(p^0; p^1, m_i^1\right) \geqq M^1 - \sum_{i=1}^{N} \mu_i\left(p^1; p^1, m_i^1\right) = 0$$

Likewise for the equivalent variation,

$$0 = \sum_{i=1}^{N} \mu_i\left(p^0;\ p^0, m_i^0\right) - M^0 \geqq \sum_{i=1}^{N} \mu_i\left(p^0;\ p^1, m_i^1\right) - M^0.$$

In the latter case the same indirect utility functions are summed on both sides of the inequality sign; it is a case where Benthamites and compensationists can find common ground.

Boadway (1974) considered the relationship between the condition of positive summed compensating variations and the fulfillment of compensation tests and came to the negative conclusion that the former was neither necessary nor sufficient for satisfaction of the latter in general, but was sufficient in the case of identical and homothetic preferences. Foster (1976) showed that, if there are no price distortions (but not otherwise), satisfaction of the compensation tests implies satisfaction of the ‘cost–benefit criterion’ (positive summed compensating variations). This conclusion is in accord with the above inequalities.

What about the Hicksian tenet that the size of the compensating variation is important so that one can compare two suboptimal states? This would require one to be able to conclude that, if the compensating variation from state 0 to state 2 is positive and greater than the compensating variation from state 0 to state 1, then state 2 should be superior to state 1 in terms of the dual KHS ordering. But this is not true even in the case of the single consumer. It was shown in Chipman and Moore (1980) that the function $c_i(p, m_i;\ p^0, m_i^0)$ cannot be an indirect utility function for unrestricted domain $(p, m_i) > 0,$ and can be if m_i is held constant if and only if preferences are homothetic, and if p_1 is held constant if and only if preferences are ‘parallel’ with respect to commodity 1. If preferences are identical and homothetic, since $\mu_i = \mu$ is homogeneous of degree 1 in m_i, $\Sigma_{i=1}^{N} \mu_i(p^0;\ p, m_i) = \mu(p^0;\ p, M),$ so exact aggregative analogues are obtained to both the compensating and equivalent variations. If the equivalent variation, which is an indirect utility function, is used, restrictions on consumer preferences are not needed, and the problem of finding an adequate indicator of the size of the deviation from a given Pareto optimum is satisfactorily resolved.

6 Game theory

One of the striking aspects of von Neumann and Morgenstern's theory of games (1947) was not only its postulate of measurability of utility but also that of its transferability between players. Since this was introduced as a positive rather than a normative assumption, it has met with even greater resistance of the part of economists than the hedonist calculus. Indeed, it was not until Debreu and Scarf (1963) showed how game theory could be liberated from this restriction with their development of the concept of the core of an economy that game theory began to be taken really seriously by economists. The replacement of transferability of utilities by transferability of commodities bears a striking resemblance to the replacement in welfare economics of the calculus of utilities by the principle of compensation.

In some branches of game theory the assumption of transferable utility is still retained, but it has been made somewhat more plausible, or at least interpretable, by means of the postulate that the utility functions of all individuals are linear in one distinguished commodity used for making side payments (cf.

Owen, 1982, p. 122). These utility functions have the form

$$U_i(x_{i1}, x_{i2}, \ldots, x_{in}) = c_i x_{i1} + V_i(x_{i2}, \ldots, x_{in}).$$

This form of the utility function goes back to Edgeworth (1891, p. 237n) and even earlier (though in garbled form) to Auspitz and Lieben (1889, p. 471). In Edgeworth it was used to illustrate the phenomenon of exchange when the marginal utility of one commodity serving as money was held constant, in accordance with one possible interpretation of Marshall's theory of consumers' surplus. (In the case $n = 2$ he showed that the exchange in commodity 2 would be constant, but in commodity 1 ‘indeterminate’; see the reply by Berry, 1891, on behalf of Marshall, and Marshall, 1891, p. 756; 1920, p. 845.) The above form for the utility function has been rediscovered many time, by Wilson (1939), Samuelson (1942), and others; cf. Chipman and Moore (1976b, p. 115). Barone (1894, p. 213n) gave the name ‘ideal money (numéraire)’ to a good with a constant marginal utility (commodity 1 in the above). For the case $c_i = c$ for all *i*, these ‘parallel’ preferences (cf. Boulding, 1945) yield a special case of the family of aggregable Antonelli–Gorman–Nataf demand functions referred to above.

7 Concluding observations

As Scitovsky (1941) pointed out, the compensation principle has been used in two quite different ways. Prior to Hicks (1940), it was used only to compare efficient with inefficient states of a given economy with a given technology or trading system. Starting with Hicks (1940), its use was extended to comparison of efficient states of an economy under different technologies. It has turned out that, in order for national-income comparisons to provide a correct indicator of potential-welfare improvement, very strong conditions are required concerning similarity of individual preferences: locally, the Antonelli–Gorman–Nataf conditions, and, globally, identical homothetic preferences. It is not even enough to assume that aggregate demand can be generated by an aggregate preference relation – for example, that preferences are homothetic and relative income-distribution constant (cf. Chipman and Moore, 1980a). Even in such cases, strong value judgements (such as acceptance of a particular Bergson–Samuelson social-welfare function) are required in order to draw welfare conclusions from national-income comparisons.

When attention is restricted to the efficient operation of an economy with a given technology, it turns out that, in most cases of interest, the ranking of consumption-possibility criterion sets according to the Kaldor–Hicks–Samuelson criterion follows from their ranking by set-inclusion. This does not mean, however, that the set-inclusion is always obvious or easy to prove.

The KHS ordering of consumption-possibility sets could be given simply a factual interpretation as indicating the ‘productive potential’ of an economy. But if it is given a normative interpretation then it obviously involves a value judgement, since a more efficient outcome, if it is not Pareto-superior, can obviously be judged worse in terms of some social-welfare function.

Samuelson's (1956) model of the ‘good society’, elegant though it is, is too sweeping for most economists to accept, and it begs the question of how the social-welfare function will be chosen. Little's (1950) and

Mishan's (1969) attempts to link plausible distributional value judgements with compensation criteria have encountered unresolvable logical difficulties (cf. Chipman and Moore, 1978). The hope that the compensation principle would allow policy decisions to be made free of value judgements has not been fulfilled. Nevertheless, much has been learned about the interrelationships among values, facts, and policies, and it can certainly be said that the development of the compensation principle has led to clearer thinking about economic policy issues.

See Also

- social welfare function
- welfare economics

Bibliography

Antonelli, G.B. 1886. *Sulla teoria matematica della economia politica*. Pisa: Folchetto. English translation: On the mathematical theory of political economy, in *Preferences, Utility, and Demand*, ed. J. S. Chipman, L. Hurwicz, M.K. Richter and H.F. Sonnenschein, New York: Harcourt Brace Jovanovich, 1971.

Arrow, K.J. 1952. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley and Los Angeles: University of California Press.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Arrow, K.J. and Scitovsky, T., eds. 1969. *Readings in Welfare Economics*. Homewood, IL: Irwin.

Auspitz, R. and Lieben, R. 1889. *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot.

Baldwin, R.E. 1948. Equilibrium in international trade: a diagrammatic analysis. *Quarterly Journal of Economics* 62, 748–62.

Baldwin, R.E. 1954. A comparison of welfare criteria. *Review of Economic Studies* 21, 154–61.

Barone, E. 1894. Sulla ‘consumers' rent’. *Giornale degli Economisti* Series 2, 9, September, 211–24.

Barone, E. 1908. Il Ministerio della produzione nello stato colletivista. *Giornale degli Economisti* Series 2, 37, August, 267–93; October, 391–414. English translation: The Ministry of Production in the collectivist state, in *Collectivist Economic Planning*, ed. F.A. Hayek. London: Routledge & Kegan Paul, 1935.

Berry, A. 1891. Alcune brevi parole sulla teoria del baratto di A. Marshall. *Giornale degli Economisti* Series 2, 2, June, 549–53.

Boadway, R.W. 1974. The welfare foundations of cost-benefit analysis. *Economic Journal* 84, 926–39. A reply, *Economic Journal* 86 (1976), 358–61.

Boulding, K.E. 1945. The concept of economic surplus. *American Economic Review* 35, 851–69.

Cairnes, J.E. 1874. *Some Leading Principles of Political Economy Newly Expounded*. New York: Harper & Brothers.

Chipman, J.S. 1976. The Paretian heritage. *Revue européenne des sciences sociales et Cahiers Vilfredo Pareto* 14(37), 65–171.

Chipman, J.S. 1982. Samuelson and welfare economics. In *Samuelson and Neoclassical Economics*, ed. G. R. Feiwel. Boston: Kluwer-Nijhoff Publishing.

Chipman, J.S. and Moore, J.C. 1971. The compensation principle in welfare economics. In *Papers in Quantitative Economics*, vol. 2, ed. A.M. Zarley. Lawrence: University Press of Kansas.

Chipman, J.S. and Moore, J.C. 1972. Social utility and the gains from trade. *Journal of International Economics* 2(May), 157–72.

Chipman, J.S. and Moore, J.C. 1973. Aggregate demand, real national income, and the compensation principle. *International Economic Review* 14, 152–81.

Chipman, J.S. and Moore, J.C. 1976a. Why an increase in GNP need not imply an improvement in potential welfare. *Kyklos* 29(3), 391–418.

Chipman, J.S. and Moore, J.C. 1976b. The scope of consumer's surplus arguments. In *Evolution, Welfare, and Time in Economics*, ed. A. Tang, F.M. Westfield and J.S. Worley. Lexington, MA: Heath.

Chipman, J.S. and Moore, J.C. 1978. The New Welfare Economics, 1939–1974. *International Economic Review* 19, 547–84.

Chipman, J.S. and Moore, J.C. 1979. On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21(August), 111–39.

Chipman, J.S. and Moore, J.C. 1980a. Real national income with homothetic preferences and a fixed distribution of income. *Econometrica* 48, 401–22.

Chipman, J.S. and Moore, J.C. 1980b. Compensating variation, consumer's surplus, and welfare.

American Economic Review 70, 933–49.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3(October), 1–44.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19, 273–92.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40, 588–92.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

Dupuit, J. 1844. De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées, Mémoires et documents relatifs à l'art des constructions et au service de l'ingénieur* Series 2,2, 2e semestre, 332–75, Pl. 75. English translation: On the measurement of the utility of public works, in Arrow and Scitovsky (1969).

Edgeworth, F.Y. 1891. Osservazioni sulla teoria matematica dell'economia politica con riguardo speciale ai principi di economia di Alfredo Marshall. *Giornale degli Economisti* Series 2, 2, March, 233–45. Ancora a proposito della teoria del baratto, *Giornale degli Economisti*> Series 2, 2, October, 316–18. Abridged English translation: On the determinateness of economic equilibrium, in F.Y. Edgeworth, *Papers Relating to Political Economy*, vol. 2, London: Macmillan, 1925.

Foster, E. 1976. The welfare foundations of cost-benefit analysis – a comment. *Economic Journal* 86, 353–8.

Frisch, R. 1939. The Dupuit taxation theorem. *Econometrica* 7, 145–50. A further note on the Dupuit taxation theorem, *Econometrica* 7, 156–7.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21, 63–80.

Gorman, W.M. 1955. The intransitivity of certain criteria used in welfare economics. *Oxford Economic Papers*, N.S. 7, February, 25–35.

Graaff, J.de V. 1949. On optimum tariff structures. *Review of Economic Studies* 17, 47–59. Reprinted in Arrow and Scitovsky (1969).

Graaff, J.de V. 1957. *Theoretical Welfare Economics*. Cambridge: Cambridge University Press.

Grandmont, J.M. and McFadden, D. 1972. A technical note on classical gains from trade. *Journal of*

International Economics 2(May), 109–125.

Harrod, R.F. 1938. Scope and method of economics. *Economic Journal* 48, 383–412.

Hicks, J.R. 1939. The foundations of welfare economics. *Economic Journal* 49, 696–712.

Hicks, J.R. 1940. The valuation of social income. *Economica*, N.S. 7, 105–24.

Hicks, J.R. 1941. The rehabilitation of consumers' surplus. *Review of Economic Studies* 8(February), 108–16. Reprinted in Arrow and Scitovsky (1969).

Hicks, J.R. 1942. Consumers' surplus and index-numbers. *Review of Economic Studies* 9(Summer), 126–37.

Hicks, J.R. 1957. *A Revision of Demand Theory*. Oxford: Clarendon Press.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6, 242–69. Reprinted in Arrow and Scitovsky (1969).

Hotelling, H. 1939. The relation of prices to marginal costs in an optimum system. *Econometrica* 7, 151–5. A final note, *Econometrica* 7, 158–9.

Hurwicz, L. and Uzawa, H. 1971. On the integrability of demand functions. In *Preferences, Utility, and Demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter and H.F. Sonnenschein. New York: Harcourt Brace Jovanovich.

Kaldor, N. 1939. Welfare propositions in economics and interpersonal comparisons of utility. *Economic Journal* 49, 549–52. Reprinted in Arrow and Scitovsky (1969).

Kaldor, N. 1940. A note on tariffs and the terms of trade. *Economica*, N.S. 7, 377–80.

Kannai, Y. and Mantel, R. 1978. Non-convexifiable Pareto sets. *Econometrica* 46, 571–5.

Kemp, M.C. 1962. The gains from international trade. *Economic Journal* 72, 803–19.

Kemp, M.C. and Wan, H.Y., Jr. 1972. The gains from free trade. *International Economic Review* 13, 509–22.

Kenen, P.B. 1957. On the geometry of welfare economics. *Quarterly Journal of Economics* 71, 426–47.

Koopmans, T.C. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Kuznets, S. 1948. On the valuation of social income – reflections on Professor Hicks' article. *Economica*, N.S. 15, February, 1–16, May, 116–31.

Lerner, A.P. 1934. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1(June), 157–75.

Little, I.M.D. 1950. *A Critique of Welfare Economics*. 2nd edn, London: Oxford University Press, 1957.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan. 2nd edn, 1891; 8th edn, 1920.

McKenzie, L.W. 1957. Demand theory without a utility index. *Review of Economic Studies* 24(June), 185–9.

Mishan, E.J. 1969. *Welfare Economics: An Assessment.* Amsterdam: North-Holland.

Moore, J.C. 1975. The existence of ‘compensated equilibrium’ and the structure of the Pareto efficiency frontier. *International Economic Review* 16, 267–300.

Nataf, A. 1953. Sur des questions d'agrégation en économétrie. *Publications de l'Institut de Statistique de l'Université de Paris* 2(4), 5–61.

von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd edn, Princeton: Princeton University Press.

Olsen, E. 1958. Udenrigshandelens gevinst [The gains of international trade]. *Nationaløkonomisk Tiddskrift* 98(1–2), 76–9.

Otani, Y. 1972. Gains from trade revisited. *Journal of International Economics* 2, 127–56.

Owen, G. 1982. *Game Theory*, 2nd edn. Orlando, FL: Academic Press.

Pareto, V. 1894. Il massimo di utilità dato dalla libera concorrenza. *Giornale degli Economisti* Series 2, 9 July, 48–66.

Pareto, V. 1895. Teoria matematica del commercio internazionale. *Giornale degli Economisti* Series 2, 10 April, 476–98.

Pareto, V. 1896, 1897. *Cours d'économie politique*, 2 vols. Lausanne: F. Rouge.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan. 4th edn, 1932.

Ricardo, D. 1815. *An Essay on the Influence of a Low Price of Corn on the Profits of Stock*. London: John Murray. In *The Works and Correspondence of David Ricardo*, vol. 4, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Robbins, L. 1938. Interpersonal comparisons of utility: a comment. *Economic Journal* 48, 635–41.

Samuelson, P.A. 1939. The gains from international trade. *Canadian Journal of Economics and Political Science* 5(May), 195–205.

Samuelson, P.A. 1942. Constancy of the marginal utility of income. In *Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz*, ed. O. Lange, F. McIntyre and T.O. Yntema. Chicago: University of Chicago Press.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. 1950. Evaluation of real national income. *Oxford Economic Papers*, N.S. 1, January, 1–29. Reprinted in Arrow and Scitovsky (1969).

Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70(February), 1–22.

Samuelson, P.A. 1962. The gains from international trade once again. *Economic Journal* 72, 820–29.

Sanger, C.P. 1895. Recent contributions to mathematical economics. *Economic Journal* 5(March), 113–28.

Scitovsky, T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* 9 (November), 77–88. Reprinted in Arrow and Scitovsky (1969).

Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9(Summer), 89–110.

Silberberg, E. 1980. Harold Hotelling and marginal cost pricing. *American Economic Review* 70, 1054–7.

Vanek, J. 1964. A rehabilitation of ‘well-behaved’ social indifference curves. *Review of Economic Studies* 31, 87–9.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper & Brothers.

Wilson, E.B. 1939. Pareto versus Marshall. *Quarterly Journal of Economics* 53, 645–50.

How to cite this article

Chipman, John S. "compensation principle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000260> doi:10.1057/9780230226203.0277

The New Palgrave Dictionary of Economics Online

competing risks model

Gerard J. van den Berg

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A competing risks model is a model for multiple durations that start at the same point in time for a given subject, where the subject is observed until the first duration is completed and one also observes which of the durations is completed first. This article gives an overview of the main issues in the empirical econometric analysis of competing risks models. The central problem is the non-identification of dependent competing risks models. Models with regressors can overcome this problem, but it is advisable to include additional data. Alternatively, effects of interest can be bounded.

Keywords

bounds; censoring; competing risks models; copulas; duration models; hazard rates; identification; latent durations; marriage and divorce; maximum likelihood; mixed proportional hazard models; multiple-spell competing risks data; multivariate duration models; nonparametric kernel estimators; regressors; Roy models; selection; semi-parametric models; stochastic dominance; unemployment durations; unobserved heterogeneity; Weibull distributions

Article

A competing risks model is a model for multiple durations that start at the same point in time for a given subject, where the subject is observed until the first duration is completed and one also observes which of the multiple durations is completed first.
The term ‘competing risks’ originates in the interpretation that a subject faces different risks *i* of leaving the state it is in, each risk giving rise to its own exit destination, which can also be denoted by *i*. One may then define random variables T_i describing the duration until risk *i* is materialized. Only the smallest of all these durations $Y = \min_i T_i$ and the corresponding actual exit destination, which can be expressed as $Z = \operatorname{argmin}_i T_i$, are observed. The other durations are censored in the sense that all is known is that their realizations exceed *Y*. Often those other durations are latent or counterfactual, for

example if T_i denotes the time until death due to cause i.

In economics, the most common application concerns individual unemployment durations. One may envisage two durations for each individual: one until a transition into employment occurs, and one until a transition into non-participation occurs. We observe only one transition, namely, the one that occurs first. Other applications include the duration of treatments, where the exit destinations are relapse and recovery, and the duration of marriage, where one risk is divorce and the other is death of one of the spouses. More generally, the duration until an event of interest may be right-censored due to the occurrence of another event, or due to the data sampling design. The duration until the censoring is then one of the variables T_i.

Sometimes one is interested only in the distribution of Y. For example, an unemployment insurance (UI) agency may be concerned only about the expenses on UI and not in the exit destinations of recipients. In such cases one may employ standard statistical duration analysis for empirical inference with register data on the duration of UI receipt. However, in studies on individual behaviour one is typically interested in one or more of the marginal distributions of the T_i. If these variables are known to be independent, then again one may employ standard duration analysis for each of the T_i separately, treating the other variables $T_j (j \neq i)$ as independent right-censoring variables. But often it is not clear whether the T_i are independent. Indeed, economic theory often predicts that they are dependent, in particular if they can be affected by the individual's behaviour and individuals are heterogeneous. It may even be sensible from the individual's point of view to use their privately observed exogenous exit rates into destinations j as inputs for the optimal strategy affecting the exit rate into destination $i (j \neq i)$ (see, for example, van den Berg, 1990). Erroneously assuming independence leads to incorrect inference, and in fact the issue of whether the durations T_i are related is often an important question in its own right.

Unfortunately, the joint distribution of all T_i is not identified from the joint distribution of Y, Z, a result that goes back to Cox (1959). In particular, given any specific joint distribution, there is a joint distribution with independent durations T_i that generates the same distribution of the observable variables Y, Z. In other words, without additional structure, each dependent competing risks model is observationally equivalent to an independent competing risks model. The marginal distributions in the latter can be very different from the true distributions.

Of course, some properties of the joint distribution are identified. To describe these it is useful to introduce the concept of the hazard rate of a continuous duration variable, say W. Formally, the hazard rate at time t is $\theta(t) := \lim_{dt \downarrow 0} \Pr(W \in [t, t + dt)) / dt$. Informally, this is the rate at which the duration W is completed at t given that it has not been completed before t. The hazard rate is the basic building block of duration analysis in social sciences because it can be directly related to individual behaviour at t. The data on Y, Z allow for identification of the hazard rates of T_i at t given that $T \geq t$. These are called the ‘crude’ hazard rates. If the T_i are independent, then these equal the ‘net’ hazard rates of the marginal distributions of the T_i.

We now turn to a number of approaches that overcome the general non-identification result for competing risks models. In econometrics, one is typically interested in covariate or regressor effects. The main approach has therefore been to specify semi-parametric models that include observed regressors X and unobserved heterogeneity terms V. With a single risk, the most popular duration model

is the mixed proportional hazard (MPH) model, which specifies that $\theta(t|X = x, V) = \psi(t)\exp(x'\beta)V$ for some function $\psi(.)$. V is unobserved, and the composition of the survivors changes selectively as time proceeds, so identification from the observable distributions of $T|X$ is non-trivial. However, it holds under the assumptions that $X \bullet V$ and $\mathrm{var}(X) > 0$ and some regularity assumptions (see van den Berg, 2001, for an overview of results). With competing risks, the analogue of the MPH model is the multivariate MPH (MMPH) model. With two risks,

$$\theta_1(t|x, V) = \psi_1(t)\exp(x'\beta_1)V_1 \text{and}$$
$$\theta_2(t|x, V) = \psi_2(t)\exp(x'\beta_2)V_2 .$$

where $T_1, T_2|X, V$ are assumed independent, so that a dependence of the durations given X is modelled by way of their unobserved determinants V_1 and V_2 being dependent. Many empirical studies have estimated parametric versions of this model, using maximum likelihood estimation.

The semi-parametric model has been shown to be identified, under only slightly stronger conditions than those for the MPH model (Abbring and van den Berg, 2003). Specifically, $\mathrm{Var}(X) > 0$ is strengthened to the condition that the vector X includes two continuous variables with the properties that (*a*) their joint support contains a non-empty open set in $\mathbb{R}^2$, and (*b*) the vectors $\ddot{\beta}_1, \ddot{\beta}_2$ of the corresponding elements of β_1 and β_2 form a matrix $(\ddot{\beta}_1, \ddot{\beta}_2)$ of full rank. Somewhat loosely, X has two continuous variables that are not perfectly collinear and that act differently on θ_1 and θ_2. Note that, with such regressors, one can manipulate $\exp(x'\beta_1)$ while keeping $\exp(x'\beta_2)$ constant. The two terms $\exp(x'\beta_i)$ are identified from the observable crude hazards at $t = 0$ because at $t = 0$ no dynamic selection due to the unobserved heterogeneity has taken place yet. Now suppose one manipulates x in the way described above. If $T_1, T_2|X$ are independent, then the observable crude hazard rate of T_2 at $t > 0$, given that $T_1 \geq t$, does not vary along. But, if $T_1, T_2|X$ are dependent, then this crude hazard rate does vary along, for the following reason. First, changes in $\exp(x'\beta_1)$ affect the distribution of unobserved heterogeneity V_1 among the survivors at t, due to the well-known fact that V_1 and X are dependent conditional on survival (i.e. conditional on $T_1 \geq t > 0$) even though they are independent unconditionally. Second, if V_1 and V_2 are dependent, this affects the distribution of V_2 among the survivors at t, which in turn affects the observable crude hazard of T_2 at t given that $T_1 \geq t$. In sum, the variation in this crude hazard with $\exp(x'\beta_1)$ for given $\exp(x'\beta_2)$ is informative on the dependence of the durations. An analogous argument holds for the crude hazard rate corresponding to cause $i = 1$. Note that identification is not based on exclusion restrictions of the sort encountered in instrumental variable analysis, which require a regressor that affects one endogenous variable but not the other. Here, all explanatory variables are allowed to affect both duration variables – they are just not allowed to affect the duration distributions in the same way. Identification with regressors was first established by Heckman and Honoré (1989), who considered a somewhat larger class of models than the MMPH model

and accordingly imposed stronger conditions on the support of *X*.

Although the MPH model is identified from single-risk duration data where we observe a single spell per subject, there is substantial evidence that estimates are sensitive to misspecification of functional forms of model elements (see van den Berg, 2001, for an overview). This implies that estimates of MMPH models using competing-risks data should also be viewed with caution. It is advisable to include additional data. For example, longitudinal survey data on unemployment durations subject to right-censoring can be augmented with register data or retrospective data not subject to censoring (see for example van den Berg, Lindeboom and Ridder, 1994). More in general, one may resort to 'multiple-spell competing risks' data, meaning data with multiple observations of *Y*, *Z* for each subject. For a given subject, such observations can be viewed as multiple independent draws from the subject-specific distribution of *Y*, *Z*, on the assumption that the unobserved heterogeneity terms V_1,V_2 are identical across the spells of the subject. Here, a subject can denote a single physical unit, like an individual, for which we observe two spells in exactly the same state, or it can denote a set of physical units for which we observe one spell each. Multiple-spell data allow for identification under less stringent conditions than single-spell data. Abbring and van den Berg (2003) showed that such data identify models that allow for full interactions between the elapsed durations *t* and *x* in $\theta_i(t|x, V)$, and, indeed, allow the corresponding effects to differ between the first and the second spell. The assumptions on the support of *X* are similar to above. Fermanian (2003) developed a nonparametric kernel estimator of the Heckman and Honoré (1989) model.

Another approach to deal with non-identification of dependent competing risks models is to determine bounds on the sets of marginal and joint distributions that are compatible with the observable data. Peterson (1976) derived sharp bounds in terms of observable quantities. They are often wide. In case of the marginal distributions of two sub-populations distinguished by a variable *X*, the bounds associated with the different *X* may overlap, whether or not *X* (monotonically) affects (one of) the marginal distributions. With overlap, the causal effects of *X* cannot even be signed.

Bond and Shaw (2006) combined bounds with regressors. In the case of a single binary regressor, the only substantive assumption made is that there exist increasing functions *g* and *h* such that $T_1, T_2|X = 0$ equals $g(T_1), h(T_2)|X = 1$ in distribution. In words, the dependence structure is invariant to the values of the regressors, so the latter affect only the marginal distributions. Specifically, the copula (and therefore Kendall's τ) of the joint distribution is invariant to the value of *X*. The assumption is satisfied by the aforementioned competing risks models with regressors. Clearly, by itself the assumption is insufficient for point identification. The bounds concern the regressor effects on the marginal distributions. If it is assumed that *X* affects the marginal distributions of T_i in terms of first-order stochastic dominance, the bounds are sufficient to sign the effect of *X* on at least one of the marginal distributions (so, in case of MMPH models, also on at least one of the individual marginal distributions conditional on *V*).

We end this article by noting some connections between competing risks models and other models. First, they are related to switching regression models, or Roy models. For example, if $T_i|X, V$ in the MMPH model have Weibull distributions, then we can write $\log T_i = x_i\alpha_i + \varepsilon_i (i = 1, 2)$ (for example, van den Berg, Lindeboom and Ridder, 1994), where we observe T_i iff $T_i < T_j (j \neq i)$. Second, competing risks models are building blocks of multivariate duration models, notably models where one of the durations

is always observed (for example, T_1 captures the moment of a treatment and T_2 is the observed duration outcome of interest).

We have considered only continuous-time duration variables T_i that have different realizations with probability 1. Recently, semi-parametric and nonparametric results have been derived for discrete-time or interval-censored competing risks models and models where different risks can be realized simultaneously (see for example Bedford and Meilijson, 1997; van den Berg, van Lomwel and van Ours, 2003; Honoré and Lleras-Muney, 2006). The biostatistical literature contains many studies in which specific assumptions are made on the dependence structure of the two durations T_i, enabling inference on the marginal distributions from data on Y, Z (see for example Moeschberger and Klein, 1995, for a survey).

See Also

- partial identification in econometrics
- proportional hazard model
- selection bias and self-selection

Bibliography

Abbring, J. and van den Berg, G. 2003. The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society, Series B* 65, 701–10.

Bedford, T. and Meilijson, I. 1997. A characterization of marginal distributions of (possibly dependent) lifetime variables which right censor each other. *Annals of Statistics* 25, 1622–45.

Bond, S. and Shaw, J. 2006. Bounds on the covariate-time transformation for competing-risks survival analysis. *Life time Data Analysis* 12, 285–303.

Cox, D. 1959. The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society, Series B* 21, 411–21.

Fermanian, J. 2003. Nonparametric estimation of competing risks models with covariates. *Journal of Multivariate Analysis* 85, 156–91.

Heckman, J. and Honoré, B. 1989. The identifiability of the competing risks model. *Biometrika* 76, 325–30.

Honoré, B. and Lleras-Muney, A. 2006. Bounds in competing risks models and the war on cancer. *Econometrica* 74, 1675–98.

Moeschberger, M. and Klein, J. 1995. Statistical methods for dependent competing risks. *Lifetime Data Analysis* 1, 195–204.

Peterson, A. 1976. Bounds for a joint distribution function with fixed sub-distribution functions: application to competing risks. *Proceedings of the National Academy of Sciences* 73, 11–13.

van den Berg, G. 1990. Search behaviour, transitions to nonparticipation and the duration of unemployment. *Economic Journal* 100, 842–65.

van den Berg, G. 2001. Duration models: specification, identification, and multiple durations. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

van den Berg, G., Lindeboom, M. and Ridder, G. 1994. Attrition in longitudinal panel data, and the empirical analysis of dynamic labour market behaviour. *Journal of Applied Econometrics* 9, 421–35.

van den Berg, G., van Lomwel, A. and van Ours, J. 2003. Nonparametric estimation of a dependent competing risks model for unemployment durations. Discussion Paper No. 898. Bonn: IZA.

How to cite this article

van den Berg, Gerard J. "competing risks model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000555> doi:10.1057/9780230226203.0278

The New Palgrave Dictionary of Economics Online

competition and selection

Sidney G. Winter

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The claim that a business firm must maximize profit if it is to survive serves as an informal statement of the common conclusion of a class of theorems characterizing explicit models of economic selection processes. Such models, by making explicit the strong assumptions needed to generate this sort of result, are the basis for a critique of standard economic theory which relies on competitive equilibrium. Models of Schumpeterian competition, emphasizing the centrality of innovation, plainly provide a much better description of the world we live in than do models of static equilibrium.

Keywords

adjustment, dynamic vs static; Alchian, A.; behavioural change; comparative statics; competition and selection; competitive equilibrium; constant returns to scale; decreasing returns to scale; diminishing returns; entrepreneurial rents; entrepreneurship; evolutionary economics; fixed factors; Friedman, M.; increasing returns to scale; innovation; latent productivity; market power; mimicry theorems; natural selection; neoclassical growth theory; patents; present value; profit maximization; research and development; rules of behaviour; satisficing; Schumpeter, J.; Schumpeterian competition; selection equilibrium; Simon, H.; static equilibrium; survival of the fittest; technological opportunity; Winter, S. G.

Article

Under competitive conditions, a business firm must maximize profit if it is to survive – or so it is often claimed. This purported analogue of biological natural selection has had substantial influence in economic thinking, and the proposition remains influential today. In general, its role has been to serve as an informal auxiliary defence, or crutch, for standard theoretical approaches based on optimization and equilibrium. It appeared explicitly in this role in a provocative passage in Milton Friedman's famous essay on methodology (Friedman, 1953, ch. 1), and it seems that many economists are familiar with it in

this context only.

There is, however, an alternative role that the proposition can and does play. It serves as an informal statement of the common conclusion of a class of theorems characterizing explicit models of economic selection processes. A model in this class posits, first, a range of possible behaviours for the firm. This range must obviously extend beyond the realm of profit maximization if the conclusion of the argument is to be non-trivial, and it must include behaviour that is appropriately termed ‘profit maximizing’ if the conclusion is to be logically attainable at all. The model must also characterize a particular dynamic process that in some way captures the general idea that profitable firms tend to survive and grow, while unprofitable ones tend to decline and fail. A stationary position of such a process is a ‘selection equilibrium’.

Models of this type occupy an important but non-central position in evolutionary economic theory (Nelson and Winter, 1982). They establish that the equilibria of standard competitive theory can indeed be ‘mimicked’ (in several different senses) by the equilibria of selection models. More importantly, by making explicit the strong assumptions that apparently are required to generate this sort of result, they are the basis for a critique of its generality and an appraisal of the strength of the crutch on which standard theory leans. They also provide a helpful entry-way to the much broader class of evolutionary models in which mimicry results fail to hold. This entry-way has the convenient feature that the return path to standard theory is well marked; the sense in which evolutionary theory subsumes portions of standard theory becomes clear.

The concept of competition need not, of course, be considered only in the context of perfectly competitive equilibrium. In a broader sense of the term, any non-trivial selection model in which the ‘fit’ prosper and the ‘unfit’ do not is a model of a ‘competitive’ process. The process need not have a static equilibrium, or any equilibrium, and it may easily lead to results that are clearly non-competitive by the standards of industrial organization economics.

The remainder of this essay first considers in more detail the theoretical links between selection processes and competitive equilibrium outcomes. It then examines a more interesting and less well-explored area that involves selection and, in a broad sense, competition; Schumpeterian competition.

Competitive equilibrium as a selection outcome

The intention here is to describe the heuristic basis of existing examples of this type of theorem, or, alternatively, to describe the basic recipe from which an obviously large class of broadly similar results could be produced. There may be other basic recipes, as yet unknown. There certainly are ways to ignore individual instructions of the recipe and yet preserve the result, though at the cost of delicately contrived adjustments in other assumptions.

(To avoid confusion, it should be noted at the outset that the word ‘equilibrium’ is used in two different senses in this discussion, the ‘no incentives to change behaviour’ sense employed in economic theory and the ‘stationary position of a dynamic process’ sense that is common outside of economics. The point of the discussion is, in fact, to relate these two equilibrium ideas in a particular way.)

(1) Constant returns to scale must prevail in the specific sense that the supply and demand functions of an individual firm at any particular time are expressible as the scale (or ‘capacity’) of that firm at that time multiplied by functions depending on prices, but not directly on scale or time. Increasing returns to scale must be excluded for familiar reasons. Decreasing returns must be excluded because they will in

general give rise to equilibrium ‘entrepreneurial rents’ which could be partially dissipated by departures from maximization without threatening the survival of the firm. Thus, for example, the U-shaped long run average cost curve of textbook competitive theory does not provide a context in which selection necessarily mimics standard theory if competitive equilibrium would require some firms to be on the upward sloping portion of the curve.

(2) Firms must increase scale when profitable and decrease scale (or go out of business entirely) when unprofitable. Alternatively, profitability of a particular firm must lead to entry by perfect imitators of that firm's actions. In the absence of such assumptions, it is plain that there will in general be equilibria with non-zero profit levels, which under assumption (1) cannot mimic the competitive result. While the ‘decline or fail’ assumption is a plausible reflection of long-run breakeven constraints characteristic of actual capitalist institutions, no such realistic force attaches to the requirement that profitability lead to expansion. If firms do not pursue profits in the long-run sense of expanding in response to positive profitability, stationary positions may involve positive profits. Such stationary positions fail to mimic competitive equilibria for that reason alone (given constant returns), but they also introduce once again the possibility that the short-run behavioural responses of surviving firms may dissipate some of the positive profit that is potentially achievable at selection equilibrium scale.

In standard theory, expansion in response to profitability may be seen as an aspect of the firm's profit-seeking on the assumption that it regards prices as unaffected by its capacity decisions. In turn, this ordinarily requires that the firm in question be but one of an indeterminately large number of firms that all have access to the same technological and organizational possibilities.

While the assumption that firms have identical production sets and behavioural rules is common and appears inoffensive in orthodox theorizing, it is very much at odds with evolutionary theory. The orthodox view comes down to the assertion that all productive knowledge is freely available to one and all – perhaps it is all in the public library. By contrast, evolutionary theory emphasizes the role of firms as highly individualized repositories of productive knowledge, not all of which is articulable. From the evolutionary perspective, the fact that mimicry theorems rely on assumptions of unimpaired access to a public knowledge pool is by itself sufficient to make it clear that the selection argument can provide only a weak and shaky crutch for standard competitive theory.

(3) A firm that is breaking even with a positive output at prevailing prices must not alter its behaviour; a potential entrant that would only break even at prevailing prices must not enter. This assumption is needed to assure that the competitive equilibrium position is in fact a stationary position of the selection process.

Models of natural selection in biology do not typically involve this sort of assumption, but neither do they conclude that only the fittest genotypes survive – the biological analogue of the proposition discussed here. Rather, they show how constant gene frequencies come to prevail as the selection forces that tend to eliminate diversity come into balance with mutation forces that constantly renew it. A strictly analogous treatment of economic selection would be much more appealing than the sort of result discussed here. It would admit that occasional disruptions may arise from random behavioural change, or from over-optimistic entrants. Thus, potentially at least, it could better serve the purpose of establishing the point that the results of standard competitive theory are in some sense robust with respect to its behavioural assumptions. Unfortunately, standard theory offers no clue as to what this sense might be. It is plain that the adjustment processes of the system are centrally involved, and there is no behaviourally plausible theory of adjustment that is the dynamic counterpart in the disciplinary

paradigm of static competitive equilibrium theory.

Within the limits defined by the requirement for a strictly static competitive outcome, the most plausible approach combines the idea of characterizing the firms in the selection process by their ‘rules of behaviour’ – an idea advanced in a seminal paper by Armen Alchian (1950) – with Herbert Simon's idea of satisficing (1955). In the simplest version, each firm simply adheres unswervingly to its own deterministic behavioural rule (or ‘routine’, in the language of Nelson and Winter, 1982). Such a rule subsumes or implies the firm's supply and demand functions, and given the conditions set forth in (1) and (2) above, a constant environment evokes a constant response. Satisficing may be introduced as a complication of this picture by an assumption that a firm that sustains losses over a period of time will search for a better behavioural rule; this adds behavioural plausibility to the adjustment process but does not introduce the possibility that random rule change might disrupt an otherwise stationary competitive equilibrium position.

(4) The final requirement can be succinctly but inadequately stated as ‘some firms must actually be profit maximizers’. Although this formulation does adequately cover some simple cases, it does not suggest the depth and subtlety of the issues involved.

Two points deserve particular emphasis here. The first is the distinction between profit maximizing *rules of behaviour* (functions) and profit maximizing *actions*. In general, a selection equilibrium that mimics a particular competitive equilibrium must clearly be one in which some firms take actions that are profit maximizing in that competitive equilibrium, and in this sense are profit maximizers. But this observation does not imply that the survivors in the selection equilibrium possess maximizing *rules*, and in general it is not necessary that survivors be maximizers in this stronger sense. (Proof: Consider a competitive equilibrium with constant returns to scale. Restrict the firms' supply and demand functions to be constant up to a scale factor at the values taken in the given equilibrium. Embed this static equilibrium in a dynamic adjustment system in which firms' scales of output respond to profitability in accordance with assumption (2). Then the given competitive equilibrium becomes a selection equilibrium – since the only techniques in use make zero profit – but the firms are not profit maximizers in the stronger sense.)

The second point extends the first. The notion of profit maximizing behavioural rules itself rests on the conceptual foundation of a production set or function that is regarded as a given. In evolutionary theory, however, it is the rules themselves that are regarded as data and as logically antecedent to the values (actions) they yield in particular environments. Thus, in this context, a problem arises in interpreting the basic idea of a selection equilibrium mimicking a standard competitive one: there is no obvious set of ‘possibilities’ to which one should have reference.

The most helpful approach here emphasizes internal consistency. Assumptions about the structure of what is ‘possible’ can be invoked without the additional assumption that there is a given set of possibilities – for example, additivity and divisibility may be assumed without implying that the set of techniques to which these axioms apply is a given datum of the system. Such an approach provides a basis for discussing whether a particular selection equilibrium is legitimately *interpretable* as a competitive equilibrium given the other assumptions in force. Along this path one can explore a rich variety of selection equilibrium situations that may be thought of as competitive equilibria. Precisely because the variety is so rich, to know only that an outcome is interpretable in this fashion is to know very little about it.

In the light of formal analysis of selection models of the sort described above, how strong is the crutch that selection provides to standard theory? For many analytical purposes, it is a crucial weakness that the

crutch relates only to equilibrium actions and not to behavioural rules; it is from the knowledge that the rules are maximizing that the results of comparative statics derive. A selection system disturbed by a parameter change from a ‘mimicking’ equilibrium does not necessarily go to a new ‘mimicking’ equilibrium, let alone to one that is consistent, in standard theoretical terms, with the information revealed in the original equilibrium. More fundamentally, selection considerations cannot compensate for the inadequacies of standard theory that arise from the basic assumption that production possibilities are given data of the system.

Schumpeterian competition

In two great works and in many other writings, Joseph Schumpeter proclaimed the central importance of innovative activity in the development of capitalism. His early book, *The Theory of Economic Development*, focused on the role and contribution of the individual entrepreneur. From today's perspective the work remains enormously insightful and provocative but may seem dated; the image of the late 19th-century captains of industry lurks implicitly in the abstract account of the entrepreneur. The late work, *Capitalism, Socialism and Democracy*, is likewise insightful, provocative and a bit anachronistic. In this case, the anachronism derives from the predictions of a future in which the innovative process is bureaucratized, the role of the individual entrepreneur is fully usurped by large organizations, and the sociopolitical foundations of capitalism are thereby undercut. Present reality does not correspond closely to Schumpeter's predictions, and it seems increasingly clear that he greatly underestimated the seriousness of the incentive problems that arise within large organizations, whether capitalist corporations or socialist states.

Substantial literatures have accumulated around a number of specific issues, hypotheses and predictions put forward in Schumpeter's various writings. Regardless of the verdicts ultimately rendered on particular points, everyday observation repeatedly confirms the appropriateness of his emphasis on the centrality of innovation in contemporary capitalism. It confirms, likewise, the inappropriateness of the continuing tendency of the economics discipline to sequester topics related to technological change in sub-sectors of various specialized fields, remote from the theoretical core.

The purpose of the present discussion is to assess the relationships of selection and competition from a Schumpeterian viewpoint, that is, to extend the discussion above by considering what difference it makes if firms are engaged in inventing, discovering and exploring new ways of doing things. Plainly, one difference it makes is that ‘competition’ must now be understood in the broad sense that admits a number of additional dimensions to the competitive process, along with price-guided output determination. In particular, costly efforts to innovate, to imitate the innovations of others, and to appropriate the gains from innovation are added to the firm's competitive repertoire.

Selection now operates at two related levels. The organizational routines governing the use made of existing products and processes in every firm interact through the market place, and the market distributes rewards and punishments to the contenders. These same rewards and punishments are also entries on the market's scorecard for the higher level routines from which new products and processes derive – routines involving, for example, expenditure levels on innovative and imitative R&D efforts. Over the longer term, selection forces favour the firms that achieve a favourable balance between the rents captured from successive rounds of innovation and the costs of the R&D efforts that yield these innovations.

In formal models constructed along these lines, it is easy to see how various extreme cases turn out. One class of cases formalizes the cautionary tale told by Schumpeter (1950, p. 105), in which competition that is ‘perfect – and perfectly prompt’ makes the innovative role non-viable. Sufficiently high costs of innovation and low costs of imitation (including costs of surmounting any institutional barriers such as patents) will lead to the eventual suppression of all firms that continue to attempt innovation, and the system will settle into a static equilibrium. (The character of this equilibrium may, however, depend on initial conditions and on random events along the evolutionary path; the production set ultimately arrived at is an endogenous feature of the process.) One can also construct model examples to illustrate the cautionary message ‘innovate or die’, the principal requirement being simply a reversal of the cost conditions stated above.

With the exception of some extreme or highly simplified cases, models of Schumpeterian competition describe complex stochastic processes that are not easily explored with analytical methods. Of course, the activity of writing down a specific formal model is often informative by itself in the sense that it illuminates basic conceptual issues and poses key questions about how complex features of economic reality can usefully be approximated by a model. Some additional insight can then be obtained using simulation methods to explore specific cases (Nelson and Winter, 1982, Part V; Winter, 1984). One of the most significant benefits from simulation is the occasional discovery of mechanisms at work that are retrospectively ‘obvious’ and general features of the model.

The discussion that follows pulls together a number of these different sorts of insights, emphasizing in particular some issues that do not arise in the related theoretical literature that explores various Schumpeterian themes using neoclassical techniques (For the most part these neoclassical studies explore stylized situations involving a single possible innovation, and thus do not address issues relating to the cumulative consequences of dynamic Schumpeterian competition. See Kamien and Schwartz (1981) and Dasgupta (1985) for references and perspectives on this literature.)

A fundamental constituent of any dynamic model of Schumpeterian competition is a model of technological opportunity. Such a model establishes the linkage between the resources that model firms apply to innovative effort and their innovative achievements. The long run behaviour of the model as a whole depends critically on the answers provided for a set of key questions relating to technological opportunity. Does the individual firm face diminishing returns in innovative achievement as it applies additional resources over a short period of time? If so, from what ‘fixed factors’ does the diminishing returns effect arise, and to what extent are these factors subject to change over time either by the firm's own efforts or by other mechanisms? Are selection forces to be studied in a context in which technological opportunity presents more or less the ‘same problem’ for R&D policy over an extended period, or is the evolutionary sorting out of different policies for the firm a process that proceeds concurrently with historical change in the criteria that govern the sorting?

Technological opportunity is said to be *constant* if R&D activity amounts to a search of an unchanging set of possibilities – in effect, there is a meta-production set or meta-production function that describes what is ultimately possible. *Increasing* technological opportunity means that possibilities are being expanded over time by causal factors exogenous to the R&D efforts in question – implying that, given a level of technological achievement and a level of R&D effort, the effort will be more productive of innovative results if applied later. With constant technological opportunity, returns to R&D effort must eventually be decreasing, approaching zero near the boundary of the fixed set of possibilities.

It is all too obvious that it may be very difficult to develop an empirical basis for modelling

technological opportunity in an applied analysis of a particular firm, industry or national economy. There is no easy escape from the conundrum that observed innovative performance reflects both opportunity and endogenously determined effort, not to mention the fact that neither performance nor effort is itself easily measured or the even more basic question of whether analysis of the past can illuminate the future. These difficulties in operationalizing the concept of technological opportunity do not, unfortunately, in any way diminish its critical role in Schumpeterian competition.

The evolutionary analysis of Schumpeterian competition has not, thus far, produced any counterpart for the sorts of mimicry theorems that can be proved for static equilibria. That is, there is no model in which it can be shown that selection forces, alone or in conjunction with adaptive behavioural rules, drive the system asymptotically to a path on which surviving firms might be said to have solved the remaining portion of the dynamic optimization problem with which the model situation confronts them – except in the cases where the asymptotic situation is a static equilibrium with zero R&D. The list of identified obstacles to a non-trivial positive result is sufficiently long, and the obstacles are sufficiently formidable, so as to constitute something akin to an impossibility theorem. It seems extremely unlikely that a positive result can be established within the confines of an evolutionary approach – that is, without endowing the model firms with a great deal of correct information about the structure of the total system in which they are embedded.

The most formidable obstacle of all derives from the direct clash between the future-oriented character of a dynamic optimization and the fact that selection and adaptation processes reflect the experience of the past. If firms cannot ‘see’ the path that technological opportunity will follow in the future, if their decisions can only reflect past experience and inferences drawn therefrom, then in general they cannot position themselves optimally for the future. They might conceivably do so if the development of technological opportunity were simple enough to validate simple inference schemes. Such simplicity does not seem descriptively plausible; who is to say that it is implausible that in a particular case technological opportunity might be constant, or exponentially increasing, or following a logistic, or some stochastic variant of any of these? And without some restriction on the structural possibilities, how are model firms to make inferences to guide their R&D policies?

This obstacle is not featured prominently in the simulations reported by Nelson and Winter, which are largely confined to very tame and stylized technological regimes in which opportunity is summarized by a single exponentially increasing variable, called ‘latent productivity’. Such an environment, reminiscent in some ways of neoclassical growth theory, seems at first glance to be a promising one for the derivation of a balanced growth outcome in which actual and latent productivity are rising at the same rate, the problem facing the firms is in a sense constant, and selection and adaptation might bring surviving firm R&D policies to optimal values.

In fact, such a result remains remote even under the very strong assumption just described. Demand conditions for the product of the industry (or the economy) affect the long run dynamics, and in this area also assumptions must be delicately contrived to avoid excluding a balanced growth outcome. For example, consider an industry model with constant demand in which demand is (plausibly) less than unit elastic at low prices. Then, cost reduction continued indefinitely would drive sales revenue to zero. Zero sales revenue will not cover the cost of continuing advance. What is involved here is a reflection of the basic economics of information; costs of discovery are independent of the size of the realm application, and on the assumption stated the economic significance of that realm is dwindling to nothing. The implication is that demand conditions may check progress even if technological opportunity is

continually expanding. Indeed, this may well be the pattern that is typically realistic for any narrowly defined sector.

This difficulty too can be dispatched by an appropriately chosen assumption. Beyond it lie some further problems. A model that acknowledges the partially stochastic nature of innovative success will display gradually increasing concentration (Phillips, 1971), unless some opposing tendency is present. A good candidate for an opposing tendency is the actual exercise of market power that has been acquired by chance (Nelson and Winter, 1982, ch. 13). But this market power can, presumably, also shelter various departures from present value maximization, including departures from dynamically optimal R&D policy.

To reiterate, the quest for mimicry theorems in the context of Schumpeterian competition seems foredoomed to failure. Since models of Schumpeterian competition plainly provide a much better description of the world we live in than do models of static equilibrium, the overall conclusion with regard to the strength of the selection crutch is distinctly more negative than the conclusion for static models alone. Assumptions that firms maximize profit or present value will have to stand on their own, at least until somebody invents a better crutch for them. In the meantime, it will continue to be the case that predictions based on these assumptions are sometimes sound and sometimes silly, and standard theory does not offer a means of discriminating between the cases. More direct attention should be paid to the mechanisms of selection, adaptation and learning, which among them probably account for as much sense as economists have actually observed in economic reality, and also leave room for a lot of readily observable nonsense.

Bibliography

Alchian, A.A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58, 211–21.

Dasgupta, P. 1985. The theory of technological competition. In *New Developments in the Analysis of Market Structure*, ed. J. Stiglitz and G.F. Mathewson. Cambridge, MA: MIT Press.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Kamien, M. and Schwartz, N. 1981. *Market Structure and Innovation*. Cambridge: Cambridge University Press.

Nelson, R. and Winter, S. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, Mass.: Belknap Press of The Harvard University Press.

Phillips, A. 1971. *Technology and Market Structure: A Study of the Aircraft Industry*. Lexington, Mass.: D.C. Heath.

Schumpeter, J.A. 1912. *The Theory of Economic Development*. Trans. Redvers Opie, Cambridge, MA: Harvard University Press, 1934.

Schumpeter, J.A. 1950. *Capitalism, Socialism and Democracy*. 3rd edn, New York: Harper.

Simon, H. 1955. A behavioral model of rational choice, *Quarterly Journal of Economics* 69, 99–118.

Winter, S.G. 1964. Economic ‘natural selection’ and the theory of the firm. *Yale Economic Essays* 4(1), 225–72.

Winter, S.G. 1971. Satisficing, selection and the innovating remnant. *Quarterly Journal of Economics* 85, 237–61.

Winter, S.G. 1984. Schumpeterian competition in alternative technological regimes. *Journal of Economic Behaviour and Organization* 5(3–4), 287–320.

How to cite this article

Winter, Sidney G. "competition and selection." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000264> doi:10.1057/9780230226203.0280

The New Palgrave Dictionary of Economics Online

competition, Austrian

Paul J. McNulty
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

allocative efficiency; Austrian economics; competition, Austrian; creative destruction; entrepreneurship; imperfect competition; innovation; market power; Mises, L.E. von; monopolistic competition; monopoly; perfect competition; Schumpeter, J.A.

Article

The essence of Austrian economics is its emphasis on the ongoing economic process as opposed to the equilibrium analysis of neoclassical theory. Austrian concepts of competition reflect this emphasis. Indeed, one of the central challenges by Austrians to the neoclassical model, and a common denominator of virtually all Austrian economics, is the rejection of the concept of perfect competition. In this respect, a number of economists who cannot be considered Austrian in all aspects of their work, share, nonetheless, the Austrian emphasis on actual market activities and processes – for example, Joseph Schumpeter (1942), J.M. Clark (1961), Fritz Machlup (1942) and others.

When the concept of competition entered economics at the hands of Adam Smith and his predecessors, it was not clearly defined, but it generally meant entry by firms into profitable industries (or exit from unprofitable ones) and the raising or lowering of price by existing firms according to market conditions. There was little recognition, and virtually no analysis, of entrepreneurship as it might be reflected in these and other forms of competition, but there was a recognition that business firms do in most situations have some control over market prices, with the degree of control varying inversely with the number of firms in the industry. These basic ideas, expanded and supplemented, are generally compatible with most modern Austrian analysis.

What is objectionable to Austrian economists is the neoclassical concept of perfect competition, developed during the 19th and early 20th centuries. The development began with Cournot (1838), whose concern it was to specify as rigorously as possible the *effects* of competition, after the *process* of competition had reached its limits. His conceptualization of this situation was a market structure in which the output of any one firm could be subtracted from total industry output with no discernible

effect on price. Later contributions by Jevons, Edgeworth, J.B. Clark and Frank Knight led to the model of perfect competition as we know it today (Stigler, 1957; McNulty, 1967).

The trouble with the concept from the Austrian point of view, as Hayek has emphasized, is that it describes an equilibrium situation but says nothing about the competitive process which led to that equilibrium. Indeed, it robs the firm of all business activities which might reasonably be associated with the verb ‘to compete’ (Hayek, 1948). Thus, firms in the perfectly competitive model do not raise or lower prices, differentiate their products, advertise, try to change their cost structures relative to their competitors, or do any of the other things done by business firms in a dynamic economic system. This was precisely the reason why Schumpeter insisted on the irrelevance of the concept of perfect competition to an understanding of the capitalist process.

For Schumpeter, any realistic analysis of competition would require a shift in analytical focus from the question of how the economy allocates resources efficiently to that of how it creates and destroys them. The entrepreneur, a neglected figure in classical and neoclassical economics, is the central figure in the Schumpeterian analytical framework. The entrepreneur plays a disequilibrating role in the market process by interrupting the ‘circular flow’ of economic life, that is, the ongoing production of existing goods and services under existing technologies and methods of production and organization. He does this by innovating – that is, by introducing the new product, the new market, the new technology, the new source of raw materials and other factor inputs, the new type of industrial organization, and so on. The result is a concept of competition grounded in cost and quality advantages which Schumpeter felt is much more important than the price competition of traditional theory and is the basis of the ‘creative destruction’ of the capitalist economic process. It produces an internal efficiency within the business firm, the importance of which for economic welfare is far greater, Schumpeter argued, than the allocative efficiency of traditional economic theory (Schumpeter, 1942).

His emphasis on the advantages of the firm's internal efficiency led Schumpeter to a greater tolerance for large-scale business organizations, even for those enjoying some degree of monopoly power, than was typical of many more traditional theorists of his time. This is a not uncommon characteristic of Austrian economics. Hayek, for example, makes the distinction between entrenched monopoly, with its probable higher-than-necessary costs, and a monopoly based on superior efficiency which does relatively little harm since in all probability it will disappear, or be forced to adjust to market conditions, as soon as another firm becomes more efficient in providing the same or a similar good or service (Hayek, 1948).

And that is precisely Schumpeter's point. The ground under even large-scale enterprise is constantly shaking as a result of the competitive threat from the new firm, the new management, or the new idea. Schumpeter's competitive analysis was less a defence of monopoly power than of certain business activities which were judged to be monopolistic only from the comparative standpoint of the model of perfect competition. He insisted that the quality of a firm's entrepreneurship was of far greater significance than its mere size.

The leading contemporary Austrian theorist of competition is Israel Kirzner (1973). Kirzner's approach draws on the analysis of market processes and the concept of ‘human action’ developed earlier by Ludwig von Mises. For von Mises, entrepreneurship is human action in the market which successfully directs the flow of resources toward the fulfillment of consumer wants (Mises, 1949). Kirzner's more fully developed theory of competition is based on the idea that the means – end nexus of economic life is not given but is itself subject to creative human action. This creative role Kirzner defines as

entrepreneurship, and it is essentially the ability to detect new but desired human wants, as well as new resources, techniques, or other ways through which to satisfy them. Whether he discovers new wants or new means of satisfying old ones, the Kirznerian entrepreneur is the one who sees and exploits what others fail to notice – the profit opportunities inherent in any situation in which the prices of factor inputs fall short of the price of the final product.

There is a difference between Kirzner's theory of entrepreneurship and that of Schumpeter. Schumpeter's entrepreneur is a disequilibrating force in the economic system; he initiates economic change. Kirzner's entrepreneur plays an equilibrating role; the changes he brings about are responses to the mistaken decisions and missed opportunities he detects in the market. Unlike Schumpeter's entrepreneur, he is not so much the creator of his own opportunities as a responder to the hitherto unnoticed opportunities that already exist in the market. Thus, in the competitive market process, the Schumpeterian and Kirznerian entrepreneurs may complement each other – the one creating change, the other responding to it.

Austrian dissatisfaction with the perfectly competitive model extends to the theories of imperfect and monopolistic competition. Hayek's and Kirzner's criticisms are the same as of perfect competition, namely, that the analysis is limited to an equilibrium situation in which the underlying data are assumed to be adjusted to each other, whereas the relevant problem is the process through which adjustment occurs. Schumpeter criticized monopolistic competition for its continued acceptance of an unvarying economic structure and forms of industrial organization. Nonetheless, the incorporation into economic theory of quality competition and sales efforts, complementing the traditional and limited focus on price competition, as well as the efforts on the part of some industrial organization specialists and institutional economists to analyse and explain actual market processes, are developments that are generally within the Austrian tradition.

See Also

- Austrian economics
- competition
- creative destruction

Bibliography

Clark, J.M. 1961. *Competition as a Dynamic Process*. Washington, DC: Brookings Institution.

Cournot, A.A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Hayek, F.A. von. 1948. The meaning of competition. In *Individualism and Economic Order*. ed. F.A. Hayek. London: Routledge.

Kirzner, I. 1973. *Competition and Entrepreneurship*. Chicago: University of Chicago Press.

Machlup, F. 1942. Competition, pliopoly, and profit. Pts. I–II. *Economica*, N.S. 9, Pt. I, 1–23, Pt. II, 153–73.

McNulty, P. 1967. A note on the history of perfect competition. *Journal of Political Economy* 75, 395–9.

Mises, L. von. 1949. *Human Action*. New Haven: Yale University Press.

Schumpeter, J. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper & Row, 1962.

Stigler, G. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65, 1–17.

How to cite this article

McNulty, Paul J. "competition, Austrian." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000268> doi:10.1057/9780230226203.0281

The New Palgrave Dictionary of Economics Online

competition, classical

John Eatwell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cantillon, R.; capital accumulation; classical theory of value and distribution; competition, classical; intrinsic value; market price; Marx, K. H.; natural price; neoclassical economics; perfect competition; perfect liberty; Petty, W.; Quesnay, F.; rate of profit; Ricardo, D.; Smith, A.; Turgot, A. R. J.

Article

Only through the principle of competition has political economy any pretension to the character of a science. So far as rents, profits, wages, prices, are determined by competition, laws may be assigned for them. Assume competition to be their exclusive regulator, principles of broad generality and scientific precision may be laid down, according to which they will be regulated. (Mill, 1848, p. 242)

In all versions of economic theory ‘competition’, variously defined, is a central organizing concept. Yet the relationship between different definitions of competition and differences in the theory of value have not been fully appreciated. In particular, the characteristics of ‘perfect’ competition (notably the conditions which ensure price-taking) are often read back, illegitimately, into classical discussions of competition.

The mechanisms which determine the economic behaviour of industrial capitalism are not self-evident. As a form of economy in which production and distribution proceed by means of a generalized process of exchange (in particular by the sale and purchase of labour), it possesses no obvious direct mechanisms of economic and social coordination. Yet, in so far as these operations constitute a system, they must be endowed with some degree of regularity, the causal foundations of which may be revealed by analysis. The first steps in economic investigation which accompanied the beginnings of industrial capitalism consisted of a variety of attempts to identify such regularities, often by means of detailed description and enumeration, as in the works of Sir William Petty, and hence to establish the dominant causes underlying the behaviour of markets. But what was required was not simply the description and classification which precedes analysis, but abstraction, the transcendence of political arithmetic (Smith, 1776, p. 501).

The culmination of the search for a coherent abstract characterization of markets, and hence the foundation of modern economic analysis, is to be found in Chapter 7 of Book I of Adam Smith's *Wealth of Nations* – ‘Of the Natural and Market Price of Commodities’. In this chapter Smith presented the first satisfactory formulation of the regularity inherent in price formation. The idea, partially developed earlier by Cantillon, and by Turgot in his discussion of the circulation of money, was that

> There is in every society … an ordinary or average rate of both wages and profits … When the price of any commodity is neither more nor less than what is sufficient to pay the rent of land, the wages of labour, and the profits of stock employed … according to their natural rates, the commodity is then sold for what may be called its natural price.

and that

> The natural price … is, as it were, the central price, to which the prices of all commodities are continually gravitating. Different accidents may sometimes keep them suspended a good deal above it, and sometimes force them down somewhat below it. But whatever may be the obstacles which hinder them from settling in this center of repose and continuance, they are continually tending towards it. (Smith, 1776, p. 65)

Thus the natural price encapsulates the persistent element in economic behaviour. And that persistence derives from the ubiquitous force of competition: or, as Smith put it, the condition of ‘perfect liberty’ in which ‘the whole of the advantages and disadvantages of the different employments of labour and stock must … be either perfectly equal or continually tending to equality’ (p. 111), for the natural price is ‘the price of free competition’ (p. 68).

The relationship between competition and the establishment of what Petty called ‘intrinsic value’ had been discussed in the works of Petty, Boisguillebert, Cantillon and Harris as the outcome of rival bargaining in price formation, competition being the greater when the number of bargainers was such that none has a direct influence on price. Quesnay expressed the formation of competitive prices as being ‘independent of mens’ will … far from being an arbitrary value or a value which is established by agreement between the contracting parties’ (in Meek, 1962, p. 90), but he did not relate the *organization of production* to the formation of prices in competitive markets. Consideration of that relationship required the development of a general conception of the role of capital, and with it the notion of a general rate of profit formed by the competitive disposition of capital between alternative investments (Vaggi, 1987).

A significant step in this direction was made by Turgot, who both conceived of the process of production as part of the circulation of money:

> We see … how the cultivation of land, manufactures of all kinds, and all branches of commerce depend upon a mass of capitals, or movable accumulated wealth, which, having been first advanced by the entrepreneurs in each of these different classes of work, must return to them every year with a regular profit … It is this continual advance and return of capitals which constitutes *what ought to be called the circulation of money*. (Turgot, 1973,

> p. 148)

and saw that the structure of investments would tend to be that which yielded a uniform rate of profit:

> It is obvious that the annual products which can be derived from capitals invested in these different employments are mutually limited by one another, and that all are relative to the existing rate of interest on money. (Turgot, 1973, p. 70)

However, Turgot neither related the determination of the rate of profit to production in general – he accepted the Physiocratic idea that the incomes of the industrial and commercial classes were ‘paid’ by agriculture – nor developed the conceptual framework which linked the formation of prices and of the rate of profit to the overall organization of the economy. These were to be Smith's achievements:

> If … the quantity brought to market should at any time fall short of the effectual demand, some of the component parts of its price must rise above their natural rate. If it is rent, the interest of all other landlords will naturally prompt them to prepare more land for the raising of this commodity; if it is wages or profit, the interest of all other labourers and dealers will soon prompt them to employ more labour and stock in preparing and bringing it to market. The quantity brought thither will soon be sufficient to supply the effectual demand. All the different parts of its price will soon sink to their natural rate, and the whole price to its natural price. (Smith, 1776, p. 65)

So in a competitive market there will be a tendency for the actual prices (or ‘market prices’ as Smith called them) to be relatively high when the quantity brought to market is less than the effectual demand (the quantity that would be bought at the natural price) and relatively low when the quantity brought to market exceeds the effectual demand. This working of competition was known as the ‘Law of Supply and Demand’. The working of competition which constitutes the ‘Law’ do not identify the phenomena which *determine* natural prices. The ‘Law’ of supply and demand should not be confused with supply and demand ‘theory’, that is, the neoclassical theory of price determination which was to be developed one hundred years later. Nor should Smith's discussion of the tendencies of concrete market prices be confused with supply and demand function, which are loci of equilibrium prices.
Adam Smith's conception of ‘perfect liberty’ consists of the mobility of labour and stock between different uses – the mobility that is necessary for the establishment of ‘an ordinary or average rate both of wages and profits’ and hence for the gravitation of market prices toward natural prices. Smith identifies four reasons why market prices may deviate ‘for a long time together’ above natural price, creating differentials in the rate of profit, all of which involve restriction of mobility:

1. (a) extra demand can be ‘concealed’, though ‘secrets of this kind … can seldom be long kept’;
2. (b) secret technical advantages;
3. (c) ‘a monopoly granted either to an individual or a trading company’;
4. (d) ‘exclusive privileges of corporation, statutes of apprenticeship, and all those laws which restrain, in particular employments, the competition to a smaller number than might otherwise go

> into them'.

For Smith there is some similarity in the forces acting on wages and profits which derives from his conceiving of the capitalist as personally involved in the prosecution of a particular trade or business. So the rate of profit, like the rate of wages, may be differentiated between sectors by 'the agreeableness of disagreeableness of the business', even though 'the average and ordinary rates of profit in the different employments of stock should be more nearly upon a level than the pecuniary wages of the different sorts of labour' (1776, p. 124). Landlords, capitalists and workers are all active agents of mobility. In Ricardo's discussion the emphasis shifted towards the distinctive role of capital:

> It is, then, the desire, which every capitalist has, of diverting his funds from a less to a more profitable employment, that prevents the market price of commodities from continuing for any length of time either much above, or much below their natural price. (Ricardo, 1817, p. 91)

Ricardo used the term 'monopoly price' to refer to commodities 'the value of which is determined by their scarcity alone', such as paintings, rare books and rare wines (1817 pp. 249–51) which have 'acquired a fanciful value', and he argued that for 'Commodities which are monopolised, either by an individual, or by a company … their price has no necessary connexion with their natural value' (p. 385). His analysis of value and distribution is accordingly confined to 'By far the greatest part of those goods which are the object of desire … such commodities only as can be increased in quantity by the exertion of human labour, and on the production of which competition operates without restraint' (p. 12).
For Marx competition is synonymous with the generalization of capitalist relations of production. Competition is thus related to the rise to dominance of the capitalist mode of production.

> While free competition has dissolved the barriers of earlier relations and modes of production, it is necessary to observe first of all that the things which were a barrier to it were the inherent limits of earlier modes of production, within which they spontaneously developed and moved. These limits became barriers only after the forces of production and the relations of intercourse had developed sufficiently to enable capital as such to emerge as the dominant principle of production. The limits which it tore down were barriers to its motion, its development and realization. It is by no means the case that it thereby suspended all limits, nor all barriers, but rather only the limits not corresponding to it … Free competition is the real development of capital. (Marx, 1973, pp. 649–50)

And as capitalism itself develops so does competition:

> On the one hand… [capital] creates means by which to overcome obstacles that spring from the nature of production itself, and on the other hand, with the development of the mode of production peculiar to itself, it eliminates all the legal and extra-economic impediments to its freedom of movement in the different spheres of production. Above all it overturns all the legal or traditional barriers that would prevent it from buying this or

> that kind of labour-power as it sees fit, or from appropriating this or that kind of labour. (Marx, 1867, p. 1013)

The concentration of capital (increasing unit size of firms) and, in particular, the centralization of capital (cohesion of existing capitals) destroys and *recreates* competition. Competition is one of the most powerful ‘levers of centralization’, and

> The centralization of capitals, or the process of their attraction, becomes more intense in proportion as the specifically capitalist mode of production develops along with accumulation. In its turn centralization becomes one of the greatest levers of its development. (Marx, 1867, p. 778n)

Like Smith and Ricardo, Marx, relates the development of competition to the establishment of the general rate of profit:

> What competition, first in a single sphere, achieves is a single market value and market price derived from the individual values of commodities. And it is competition of capitals in various spheres, which first brings out the price of production equalising the rates of profit in the different spheres. The latter process requires a higher stage of capitalist production than the previous one. (Marx, 1894, p. 180)

It is in his conception of the circuit of capital that Marx best portrays capitalist competition. The image is one of capital as a homogeneous mass of value (money) seeking its maximum return. Profits are created by embodying capital in the process of production, the commodity outputs of which must be realized, that is, returned to the homogeneous money form to be reinvested. Competition is thus characteristic of the capitalist mode of accumulation; mobility and restructuring are two aspects of the same phenomenon.

Marx's general conception of capital as a system corroborates Quesnay's notion of an economy operating ‘independent of men's will’. This does not mean that there may not be circumstances in which individual capitals exercise some control in particular markets – indeed, such limitations may be necessary for the accumulation process to proceed in certain lines. Capital removes only those barriers which *limit* its accumulation. The market control exercised in some lines of modern industry is not necessarily a limitation but may be a prerequisite of production on an extended scale. Aggregate capital flows discipline the actions of individual capitals, and hence endow the system with the regularity manifest in the perpetual tendency, successfully contradicted and recreated, towards a general rate of profit and associated prices.

Competition not only establishes the object of analysis, natural prices and the general rate of profit, but makes meaningful analysis possible, since it allows the operations of the capitalist economy to be characterized in a manner which permits theoretical statements of general validity to be made about them.

Theory proceeds by the extraction from reality of those forces which are believed to be dominant and persistent, and the formation of those elements into a formal system, the solution of which is to

determine the magnitude or state of the variables under consideration. It is obvious that the solution will not, except by a fluke, correspond to the actual magnitudes of the variables ruling at any one time, for these will be the outcome not solely of the elements grouped under the heading ‘dominant and persistent’, but also of the myriad of other forces excluded from the analysis as transitory, peculiar or specific (lacking general significance) which may, at any moment, exert a more or less powerful effect. Nonetheless, the practice of analysis embodies the assumption that the forces comprising the theory *are* dominant, and that the determined magnitudes will, on average, tend to be established. In any satisfactory analytical scheme these magnitudes must be centres of gravitation, capturing the essential character of the phenomena under consideration.

The importance of Smith's use of competition is now apparent. Theory cannot exist in a vacuum. Simply labelling forces dominant is not enough. These forces must operate through a process which establishes their dominance and through which the ‘law-governed’ nature of the system is manifest. That process is competition, which both enforces and expresses the attempt of individual capitals to maximize profits. Thus important aspects of the behaviour of a capitalist market economy may be captured at a sufficient level of generality to permit the formulation of general causal statements, that is, to permit analysis. Without this step, which constitutes the establishment of what was called above the *method* of analysis, it would have been impossible to develop any general form of economic *theory*.

The classical theory of value and distribution may be shown to provide a logically coherent explanation of the determination of the general rate of profit and hence of natural prices (prices of production) taking as data (see Sraffa, 1960):

1. (a) the size and composition of social output;
2. (b) the technique in use; and
3. (c) the real wage.

The classical achievement is thus composed of two independent elements: (*a*) the characterization of the object of the theory of value; and (*b*) the provision of a theory for the determination of that object. Underlying the former is the concept of gravitation imposed by competition, and underlying the latter the concept of gravitation inherent in theoretical abstraction. Any alternative system must not simply provide a different theory but also achieve a similar congruence with the traditional method.

The development in the final quarter of the 19th century of what was to become known as the neoclassical theory of value and distribution was an attempt to provide an alternative to a classical theory embroiled in the logical difficulties inherent in the labour theory of value and sullied by unsavoury associations with radicalism and Marxism. But despite the dramatic change in theory that was to be heralded by the works of Jevons, Menger and Walras, the method of analysis which characterized the object the theory was to explain stayed fundamentally the same; the new theory was an alternative explanation of the same phenomena. Marshall labelled natural prices ‘long-run normal prices’, and declared that, as far as his discussion of value was concerned ‘the present volume is chiefly concerned … with the normal relations of wages, profits, prices etc., for rather long periods’ (1920, p. 315). The same continuity of method may be found in the work of Walras (1874–7, pp. 224, 380), Jevons (1871, pp. 86, 135–6), Böhm-Bawerk (1899, p. 380) and Wicksell (1934, p. 97).

Nonetheless, the structure of neoclassical theory is such that a different notion of competition is

required. The classical emphasis on mobility must be supplemented by a precise definition of the relationships presumed to exist between individual agents. The fundamental concept of ‘perfect’ competition, for example, encompasses the idea that the influence of each individual participant in the economy is ‘negligible’, which in turn leads to the idea of an economy with infinitely many participants (Aumann, 1964). Such formulations are entirely absent from the classical conception of competition, since the classical theory is not constructed around individual constrained utility maximization.

See Also

- competition

Bibliography

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Böhm-Bawerk, E. von. 1899. *Capital and Interest*, vol. 2. South Holland, IL: Libertarian Press, 1959.

Jevons, W.S. 1871. *Theory of Political Economy*. Harmondsworth: Penguin, 1970.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Marx, K. 1867. *Capital*, vol. 1. Harmondsworth: Penguin, 1976.

Marx, K. 1894. *Capital*, vol. 3. New York: International Publishers, 1967.

Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin.

Meek, R.L. 1956. *Studies in the Labour Theory of Value*. London: Lawrence & Wishart.

Meek, R.L. 1962. *The Economics of Physiocracy*. London: Allen & Unwin.

Meek, R.L. 1973. Introduction to Turgot (1973).

Mill, J.S. 1848. *Principles of Political Economy*. London: Parker.

Ricardo, D. 1817. *Principles of Political Economy and Taxation*, ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Methuen, 1961.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Turgot, A.J.R. 1973. *Turgot on Progress, Sociology and Economics*, ed. R.L. Meek. Cambridge: Cambridge University Press.

Vaggi, G. 1987. *The Economics of François Quesnay*. London: Macmillan.

Walras, M.E.L. 1874–7. *Elements of Pure Economics*. Homewood, IL: Irwin, 1954.

Wicksell, K. 1934. *Lectures on Political Economy*, vol. 1. London: Routledge & Kegan Paul.

How to cite this article

Eatwell, John. "competition, classical." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000269> doi:10.1057/9780230226203.0282

The New Palgrave Dictionary of Economics Online

competition

George J. Stigler

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Competition arises whenever two or more parties strive for something that all cannot obtain. The classical economists felt no need for a very precise definition of competition because they viewed monopoly as highly exceptional. In the late 19th century competition became the subject of intense analysis; the concept of perfect competition emerged as the standard model of economic theory and as first approximation in the concrete studies of applied microeconomics. The limitations of the concept in dealing with conditions of persistent and imperfectly predicted change will be removed only when economics possesses a developed theory of change.

Keywords

bilateral monopoly; Clark, J.; Cliffe Leslie, T.; coalitions; comparative statics; competition; competitive equilibrium; contracting; Cournot, A.; creative destruction; Darwin, C.; Demsetz, H.; Edgeworth, F.; entrepreneurship; factor price equalization theorem; Fisher, I.; industrial organization; innovation; Jenkin, F.; Jevons, W.; joint action; Kirzner, I.; Knight, F.; labour markets; laissez faire; law of indifference; law of one price; law of supply and demand; long-run equilibrium; Malthus, T.; market price; markets; Marshall, A.; mathematics and economics; Mill, J. S.; monopoly; oligopoly; Pareto, V.; perfect competition; perfect information; perfect markets; Pigou, A.; profit-maximixing behaviour; rate of return; recontracting; resource mobility; Schumpeter, J.; second best; Senior, N.; Smith, A.; Stigler, G. J.; Thornton, H.; Walras, L.; workable competition

Article

Competition is a rivalry between individuals (or groups or nations), and it arises whenever two or more parties strive for something that all cannot obtain. Competition is therefore at least as old as man's history, and Darwin (who borrowed the concept from economist Malthus) applied it to species as economists had applied it to human behaviour.

A concept that is applicable to two cobblers or a thousand shipowners or to tribes and nations is necessarily loosely drawn. When Adam Smith launched economics as a comprehensive science in 1776, he followed this usage. He explained why a reduced supply of a good led to a higher price: the ‘competition [which] will immediately begin’ among buyers would bid up the price. Similarly if the supply become larger, the price would sink more, the greater ‘the competition of the sellers’ (Smith, [1776], 1976, pp. 73–4). Here competition was very much like a race: a race to obtain part of reduced supplies or to dispose of a part of increased supplies. Almost nothing except a number of buyers and sellers was necessary for competition to operate. And the greater the number of each, the greater the vigour of competition:

> If this capital [sufficient to trade in a town] is divided between two different grocers, their competition will tend to make both of them sell cheaper, than if it were in the hands of one only; and if it were divided among twenty, their competition would be just so much the greater, and the chance of their combining together, in order to raise the price, just so much the less (ibid., pp. 361–2).

With such a loose concept, there was little occasion to speak of one market as being more or less competitive than another, although this very passage presented the commonsense idea that larger numbers of rivals increased the intensity of competition.

The competition of grocers in a town pertained to competition *within* a market or an industry. Smith made much of the competition of different markets or industries for resources, and he developed what has always remained the main theorem on the allocation of resources in an economy composed of private, competing individuals or enterprises. The argument may be stated: Each owner of a productive resource will seek to employ it where it will yield the largest return. As a result, under competition each resource will be so distributed that it yields the same rate of return in every use. For if a resource were earning more in one use than another, it would be possible for its return in the lower-yielding use to be increased by reallocating it to the higher-yielding use. And this theorem led to what John Stuart Mill called the most frequently encountered proposition in economics: ‘There cannot be two prices in the same market’ (Mill, 1848, Book II, ch. IV, s. 3).

The competition of different markets or industries for the use of the same resources called attention to some problems which are less important within a single market such as the grocery trade in a town. One must possess knowledge of the investment opportunities in these different employments, and that knowledge is less commonly possessed than knowledge within one market. It often requires a good deal of time to disengage resources from one field and instal them elsewhere. Both of these conditions were recognized by Smith, who spoke of the difficulty of keeping secret the existence of extraordinary profits, and of the long run sometimes required for the attainment of equality of rates of return.

For the next three-quarters of a century the prevailing treatment of competition followed the practice of Smith. One can find occasional hints of a more precise definition of competition, well illustrated by Nassau W. Senior:

> But though, under free competition, cost of production is the regulator of price, its influence is subject to much occasional interruption. Its operation can be supposed to be

> perfect only if we suppose that there are no disturbing causes, that capital and labour can be at once transferred, and without loss, from one employment to another, and that every producer has full information of the profit to be derived from every mode of production. But it is obvious that these suppositions have no resemblance to the truth. A large portion of the capital essential to production consists of buildings, machinery, and other implements, the results of much time and labour, and of little service for any except their existing purposes … few capitalists can estimate, except upon an average of some years, the amount of their own profits, and still fewer can estimate those of their neighbours (1836, p. 102).

Senior is hinting at a concept of perfect competition, but the hint is not pursued.
The classical economists felt no need for a precise definition because they viewed monopoly as highly exceptional: Harold Demsetz has counted only one page in 90 devoted to monopoly in *The Wealth of Nations* and only one in 500 in Mill's *Principles of Political Economy*. Indeed the word ‘monopoly’ was usually restricted to grants by the sovereign of exclusive rights to manufacture, import or sell a commodity; witness the entry in the *Penny Cyclopedia* (1839):

> It seems then that the word monopoly was never used in English Law, except when there was a royal grant authorizing some one or more persons only to deal in or sell a certain commodity or article.
> If a number of individuals were to unite for the purpose of producing any particular article or commodity, and if they should succeed in selling such article very extensively, and almost solely, such individuals in popular language would be said to have a monopoly. Now, as these individuals have no advantages given them by the law over other persons, it is clear they can only sell more of their commodity than other persons by producing the commodity cheaper and better (XV, p. 341).

The ability of rivals to seek out and compete away supernormal profits, unless prevented by legal obstacles, was believed to be the basic reason for the pervasiveness of competition.
In the last third of the 19th century the concept of competition became the subject of intense study. The most popular reason given for this attention is that the growth of large-scale enterprises, including railroads, public utilities, and finally great manufacturing enterprises, made obvious the fact that a simple concept of competition no longer fit the economy of an industrial nation such as England.
A second source of misgiving with the broad definition of competition is that it might not lead to the uniformity of returns to a resource predicted by the theory. The Irish economist Cliffe Leslie repeatedly made this charge:

> Economists have been accustomed to assume that wages on the one hand and profits on the other are, allowing for differences in skill and so forth, equalized by competition, and that neither wages nor profits can anywhere rise above ‘the average rate’, without a consequent influx of labour or of capital bringing things to a level. Had economists, however, in place of reasoning from an assumption, examined the facts connected with the

> rate of wages, they would have found, from authentic statistics, the actual differences so great, even in the same occupation, that they are double in one place what they are in another. Statistics of profits are not, indeed, obtainable like statistics of wages; and the fact that they are not so, that the actual profits are kept a profound secret in some of the most prominent trades, is itself enough to deprive the theory of equal profits of its base (1888, pp. 158–9).

The easiest way to combat such criticisms was not to confront them with data – that path was not chosen for many years – but to define competition in such a way as to ensure the desired results such as uniformity of price.

The complications possible with competition were raised also on the theoretical side. William T. Thornton, in his book *On Labour* (1869), denied the fact that prices were determined by the ‘law of supply and demand’, particularly within labour markets. He employed bizarre examples, such as supply and demand curves which coincided over a vertical range, to show that price could be indeterminate or unresponsive to changes in supply or demand. These objections naturally called forth responses, from both J.S. Mill (*Collected Works*, V) and Fleeming Jenkin, a famous engineer.

The most persuasive reason for the increasing attention to the concepts of economics was the gradual move of economic studies to the universities, which proceeded rapidly in the last decades of the century. The expanding use of mathematics was one major symptom of the development of the formal and abstract theory of economics by Walras, Pareto, Irving Fisher and others. That formalization would scarcely be possible without a more precise specification of the nature of competition, and the precise specification of the nature of competition, and the replies to Thornton's criticisms were a precursor to this literature.

The groundwork for the development of the concept of perfect competition was laid by Augustin Cournot in 1838 in his *Mathematical Principles of the Theory of Wealth*. He made the first systematic use of the differential calculus to study the implications of profit-maximizing behaviour. Starting with the definition, $\text{Profits} = \text{Revenue} - \text{Costs}$, Cournot sought to maximize profits under various market conditions. He faced the question: How does revenue (say, pq) vary with output (q)? The natural answer is to *define* competition as that situation in which p does not vary with q – in which the demand curve facing the firm is horizontal. This is precisely what Cournot did:

> The effects of competition have reached their limit, when each of the partial productions D_k [the output of producer k] is *inappreciable*, not only with reference to the total production $D = F(p)$, but also with reference to the derivative $F'(p)$, so that the partial production D_k could be subtracted from D without any appreciable variation resulting in the price of the commodity (Cournot [1838] 1927, p. 90).

This definition of competition was especially appropriate in Cournot's system because, according to his theory of oligopoly, the excess of price over marginal cost approached zero as the number of like producers became large. The argument is as follows:

Let the revenue of the firm be $q_i p$, and let n identical firms have the same marginal costs, MC. Then the equation for maximum profits for one firm would be

$$p + q_i(\mathrm{d}p/\mathrm{d}q) = MC.$$

The sum of *n* such equations would be

$$np + q(\mathrm{d}p/\mathrm{d}q) = nMC,$$

for $nq_i = q$. This last equation may be written,

$$p = \mathrm{MC} - p/nE,$$

where *E* is the elasticity of market demand (Cournot, 1838, p. 84).
Cournot believed that this condition of competition was fulfilled ‘for a multitude of products, and, among them, for the most important products’.
Cournot's definition was enormously more precise and elegant than Smith's so far as the treatment of numbers was concerned. A market departed from unlimited competition to the extent that prices exceeded the marginal cost of the firm, and the difference approached zero as the number of rivals approached infinity. This definition, however, illuminated only the effect of number of rivals on the power of individual firms to influence the market price, on Cournot's special assumption that each rival believed that his output decisions did not affect the output decisions of his rivals. It therefore bore only on what we term market competition.
Cournot did not face the question of the role of information possessed by traders, and this question was taken up by William Stanley Jevons in 1871 in his *Theory of Political Economy*. He characterized a perfect *market* by two conditions:

> (1.) A market, then, is theoretically perfect only when all traders have perfect knowledge of the conditions of supply and demand, and the consequent ratio of exchange; … (2.) … there must be perfectly free competition, so that any one will exchange with any one else upon the slightest advantage appearing. There must be no conspiracies for absorbing and holding supplies to produce unnatural ratios of exchange (Jevons, 1871, pp. 86, 87).

By perfect knowledge Jevons meant only that each trader in a market knew the price bids of every other trader. The second condition ruled out any joint actions by two or more traders, without his noticing that with knowledge so perfect as to know the behaviour of rivals, there might appear the very conspiracies he ruled out. The two conditions dictated that ‘there cannot be two prices for the same kind of article’ in

a perfect market, which he called the ‘law of indifference’.

The merging of the concepts of competition and the market was unfortunate, for each deserved a full and separate treatment. A market is an institution for the consummation of transactions. It performs this function efficiently when every buyer who will pay more than the minimum realized price for any class of commodities succeeds in buying the commodity, and every seller who will sell for less than the maximum realized price succeeds in selling the commodity. A market performs these tasks more efficiently if the commodities are well specified and if buyers and sellers are fully informed of their properties and prices. Also a complete, perfect market allows buyers and seller to act on differing expectations of future prices. A market may be perfect and monopolistic or imperfect and competitive. Jevons's mixture of the two has been widely imitated by successors, of course, so that even today a market is commonly treated as a concept subsidiary to competition.

Edgeworth was the first economist to attempt a systematic and rigorous definition of perfect competition. His exposition deserves the closest scrutiny in spite of the fact that few economists of his time or ours have attempted to disentangle and uncover the theorems and conjectures of the *Mathematical Psychics* (1881), probably the most elusively written book of importance in the history of economics. His exposition was the most influential in the entire literature.

The conditions of perfect competition are stated as follows:

> The *field of competition* with reference to a contract, or contracts, under consideration consists of all individuals who are willing and able to recontract about the articles under consideration …
>
> There is free communication throughout a *normal* competitive field. You might suppose the constituent individuals collected at a point, or connected by telephones – an ideal supposition [1881], but sufficiently approximate to existence or tendency for the purposes of abstract science.
>
> A *perfect* field of competition professes in addition certain properties peculiarly favourable to mathematical calculation; … The conditions of a *perfect* field are four; the first pair referable to the heading *multiplicity* or continuity, the second to *dividedness* or fluidity.
>
> 1. I. An individual is free to *recontract* with any out of an indefinite number, …
> 2. II. Any individual is free to *contract* (at the same time) with an indefinite number; … This condition combined with the first appears to involve the indefinite divisibility of each *article* of contract (if any *X* deal with an indefinite number of *Y*s he must give each an indefinitely small portion of *x*); which might be erected into a separate condition.
> 3. III. Any individual is free to *recontract* with another independently of, *without the consent* being required of, any third party, …
> 4. IV. Any individual is free to *contract* with another independently of a third party; …

The failure of the first [condition] involves the failure of the second, but not vice versa; and the third and

fourth are similarly related (Edgeworth, 1881, pp. 17–19).
The essential elements of this formidable list of conditions are two:

1. (1) There are an indefinitely large number of independent traders on each side of a market (the Cournot condition).
2. (2) Each trader can costlessly make tentative contracts with everyone (hence the divisibility of commodities) and alter these contracts (recontract) so long as a more favourable contract can be made. The result is perfect knowledge (the Jevonian condition).

Edgeworth gave an intuitive argument for the need for an indefinitely large number of traders on both sides of a market. It proceeds as follows. Let there be one seller and two buyers, and let the seller gain all the benefits of the sale: each buyer is charged the maximum price he would pay rather than withdraw from the market. If now a second seller appears, he will find it advantageous to offer better terms to the two buyers: ‘It will in general be possible for *one* of the [sellers] (without the consent of the other), to *recontract* with the two [buyers], so that for all those three parties the recontract is more advantageous than the previously existing contract’ (ibid., p. 35). As the numbers of traders on each side increase, the price approaches the competitive equilibrium level where no individual trader can influence it.
A defect in this argument is that it ignores the fact that if the traders on one or both sides of the market, be they 2, or 2000 or 2,000,000, join together they can do better *individually* than by competing. If traders on each side join, however, there will be bilateral monopoly, not competition. Edgeworth gives no reason why the combination of traders fails to take place. Only in modern times has the reason for independent behaviour by rivals been established: the costs of reaching and enforcing agreements on joint action increase with both the number of rivals and the complexity of the transactions. At a certain level – quite possibly with only two traders under some conditions – the costs of joint action exceed the gain to at least some of the traders, and independent behaviour emerges.
Edgeworth's ‘conjecture’, as it is now often called, that a unique, competitive price would emerge when the number of traders became large, has given rise to a modern literature vast in scope and often highly advanced in its mathematical techniques (for references, see Hildenbrand, 1974). One result in this literature is that in the case of a large (infinite) number of traders, no coalition of a portion of the traders can exclude traders outside the coalition from trading at the price-taking equilibrium.
Edgeworth's introduction of the requirement that the commodity or service that is traded be highly divisible is a response to the following problem:

> Suppose a market, consisting of an equal number of masters and servants, offering respectively wages and service; subject to the condition that no man can serve two masters, no master employ more than one man; or suppose equilibrium already established between such parties to be disturbed by any sudden influx of wealth into the hands of the masters. Then there is no *determinate*, and very generally *unique*, arrangement towards which the system tends under the operation of, may we say, a law of Nature, and which would be predictable if we knew beforehand the real requirements of each, or of the average, dealer; … (Edgeworth, 1881, p. 46).

Consider the simple example: a thousand masters will each employ a man at any wage below 100; a thousand labourers will each work for any wage above 50. There will be a single wage rate: knowledge and numbers are sufficient to lead a worker to seek a master paying more than the going rate or a master to seek out a worker receiving less than the market rate. But any rate between 50 and 100 is a possible equilibrium. But if a single worker leaves the market, the wage will rise to 100, and if a single employer withdraws, the wage will fall to 50. This ability of a single trader to affect the price arises because of the lumpiness of the article traded (here a worker's labour for a given period). Once a worker can work for two masters, the withdrawal of one worker in a thousand will reduce the available hours of work per day to each employer by only 8/1000 hours or 4.8 minutes per day, with only negligible influence upon the wage rate. Alternatively, a distribution of wage offers and demands would also eliminate the indeterminacy and market power.

Edgeworth's analysis was limited to competition within a market, and it was left to John Bates Clark to emphasize the need for mobility of resources if the return on each resource was to be equalized in every use.

> …there is an ideal arrangement of the elements of society, to which the force of competition, acting on individual men, would make the society conform. The producing organism actually shapes itself about his model, and at no time does it vary greatly from it … We must use assumptions boldly and advisedly, make labour and capital absolutely mobile, and letting competition work in ideal perfection (Clark, 1899, pp. 68, 71).

Perfect and free mobility of resources is of course an even more extreme assumption than the other conditions required for perfect competition because there is less reason to believe that free movement of resources is even approached in the real economy. Nor is the assumption of perfect mobility necessary to eliminate monopoly power in a market: in the Victorian age, the price of wheat of Iowa was set in Liverpool even though transportation costs were substantial. The assumption is usually necessary to attain strict equality in the price of a good at every point (the law of one price), although even this is not strictly true (as in the factor price equalization theorem). Clark also demanded that the economy be stationary for perfect competition, a condition we shall return to later.

All the elements of a concept of perfect competition were in place by 1900, and this concept increasingly became the standard model of economic theory thereafter. The most influential statement of the conditions for perfect competition was made by Frank H. Knight in his doctoral dissertation, *Risk, Uncertainty and Profit* (1921). The conditions were stated in extreme form; for example, ‘There must be perfect, continuous, costless intercommunication between all individual members of the society’ (Knight, 1921, p. 78) – so Jones in Seattle would know the price of potatoes and be able costlessly to ship to Smith in Miami a bushel of potatoes at every moment of time.

Of course these conditions are not *necessary*, but only sufficient, to achieve the competitive equilibrium. For example, if even a considerable fraction of buyers knows that seller A is charging more than B for a given commodity, their patronage may be quite enough to force A to reduce his price to that of B. Nor are the various conditions independent of one another: for example, if it is very cheap for either a commodity or its buyers or sellers to move between two places, that will insure that the prices in the two places will be widely known.

Along with the development of the concept of competition as a standard component of the theory of

prices and the allocation of resources, it acquired a growing role as the criterion by which to judge the efficiency of actual markets. Adam Smith had already advanced the proposition that output was maximized in a private enterprise economy with competition. If each owner of a resource maximized the return from his resources, then (in the absence of ‘external’ effects of one person's actions on others) aggregate output would be maximized. This theorem (labelled ‘on maximum satisfaction’) was developed and qualified by Léon Walras (1874), Alfred Marshall (1980), Pareto (1895–6, 1907), Pigou (1912) and a host of modern economists.

Competition is much too central a concept in economics to remain unaffected when economists change their interests or analytical methods. We may illustrate this fact by the problem of economic change. In a regime of change, of growing population and capital or innovations or new consumer demands, the problem of defining competition is much more difficult than it is for the stationary economy. Unless the change is predictable with precision, knowledge must necessarily be incomplete and errors and lags in adaptation to new conditions can be large. For this reason, indeed, J.B. Clark believed that perfect competition was achievable only in the stationary economy.

Even short-run changes in market price raise the question: is the change in price initiated by a particular seller or buyer, and if so, is this trader not facing a negatively sloping demand curve or a positively sloping supply curve? The infinitely elastic supply and demand curves of perfectly competitive equilibrium seem inapplicable to periods of changing market conditions. Some economists nevertheless retain the condition that individual traders cannot influence price by introducing a hypothetical auctioneer who announces price changes.

A partial adaptation of the competitive concept to change is made by making it a long-run equilibrium concept. Even if resources are not costlessly mobile and even if entrepreneurs do not have perfect foresight, one can analyse the rate of approach of returns on resources to equality. If an industry experiences a once-for-all large change, it could be in competitive equilibrium before and after the change, and the equilibria could be studied by competitive theory (comparative statics).

This adaptation did not satisfy Joseph Schumpeter, who believed that incessant change in products and production methods was the very essence of competitive capitalism. He argued that the displacing of one product or method by another, a process which he called creative destruction, made the concept of perfect competition irrelevant to either positive analysis or welfare judgements. If the monopoly that reduced output, compared to competition, by 10 per cent in one year, increased output by 100 per cent over the next two decades, then monopoly might be preferred to stagnant competition.

It is crucial to this argument that monopoly provides large, though temporary, rewards to successful innovators but competition does not:

> But perfectly free entry into a *new* field may make it impossible to enter it at all. The introduction of new methods of production and new commodities is hardly conceivable with perfect – and perfectly prompt – competition from the start. And this means that the bulk of what we call economic progress is incompatible with it. As a matter of fact, perfect competition is and always has been temporarily suspended whenever anything new is being introduced – automatically or by measures devised for that purpose – even in otherwise perfectly competitive conditions. (Schumpeter, 1942, pp. 104–5)

Schumpeter relies on instantaneous rivalry to eliminate the incentives to innovation under competition,

and the conclusion would not hold if competition is defined in terms of long-run equilibrium. Nevertheless the issue is not disposed of so easily. If change is continuous rather than sporadic, long-run equilibria will never be fully achieved. Several economists have emphasized that alterations in the concept of competition are called for in periods of historical change. Kirzner has emphasized the role of entrepreneurial rivalry in competition, whereas such rivalry is nonexistent in a perfectly competitive equilibrium. Demsetz has proposed a concept of laissez-faire competition, in which freedom of resources to move into any use is the central element. Such realistic reversions to the competitive concept of the classical economists have not been systematically formalized into theoretical models.

The concept of perfect competition, or indeed any theoretically precise concept of competition, will not be met by the actual condition of competition in any industry. John Maurice Clark made the most influential effort to create a concept of ‘workable competition’ which would serve as a working rule for public policies which seek to preserve or increase competition.

Clark emphasized the fact that if one requisite of perfect competition is absent, it may be desirable that a second requisite also be unfulfilled. For example, with instantaneous mobility but imperfect knowledge, members of an occupation would keep shifting back and forth between two cities, always overshooting the amount of migration which would equalize wage rates. This propensity to overshoot equilibrium would be corrected with less mobility of labour. This problem was later formalized as the theory of the ‘second best’.

The essence of the concept of workable competition was the belief that ‘long-run curves, both of cost and of demand, are much flatter than short-run curves, and much flatter than the curves which are commonly used in the diagrams of theorists’ (J.M. Clark, 1940, p. 460). This correct and sensible view led to a proliferation of studies, usually in doctoral dissertations, of individual industries, in which the workableness of competition in each industry was appraised. Unfortunately there were no objective criteria to guide these judgements, and there was no evidence that the studies were accepted by the governmental agencies which administered competitive policies.

The popularity of the concept of perfect competition in theoretical economics is as great today as it has ever been. The concept is equally popular as first approximation in the more concrete studies of markets and industries that comprise the field of ‘industrial organization’ (applied microeconomics). The limitations of the concept in dealing with conditions of persistent and imperfectly predicted change will not be removed until economics possesses a developed theory of change. Even within a stationary economic setting the concept is being deepened by mathematical economists (see Mas-Colell, 1982). Meanwhile the central elements of competition – the freedom of traders to use their resources where they will, and exchange them at any price they wish – will continue to play a major role in the economics of an enterprise economy.

See Also

- exchange
- large economies
- perfect competition

Bibliography

Clark, J.B. 1899. *The Distribution of Wealth*. London: Macmillan.

Clark, J.M. 1940. Toward a concept of workable competition. *American Economic Review*, June; reprinted in *Readings in the Social Control of Industry*, Philadelphia: Blakiston, 1942.

Cournot, A. 1838. *Researches into the Mathematical Principles of the Theory of Wealth*. Reprinted New York: Macmillan, 1927.

Demsetz, H. 1982. *Economic, Legal and Political Dimensions of Competition*. Amsterdam: North-Holland.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul, 1932.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan.

Kirzner, I.M. 1973. *Competition and Entrepreneurship*. Chicago: University of Chicago Press.

Knight, F.H. 1921. *Risk, Uncertainty and Profit*, Part 2. Boston: Houghton Mifflin Co.

Cliffe Leslie, T.E. 1888. *Essays in Political Economy and Moral Philosophy*. London: Longmans Green.

McNulty, P.J. 1967. A note on the history of perfect competition. *Journal of Political Economy* 75, 395–9.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Mas-Colell, A., ed. 1982. *Noncooperative Approaches to the Theory of Perfect Competition*. New York: Academic Press.

Mill, J.S. 1848. *Principles of Political Economy*. In *Collected Works*, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Nutter, G.W. 1951. *The Extent of Enterprise Monopoly in the United States, 1899–1939*. Chicago: University of Chicago Press.

Penny Cyclopedia of the Society for the Diffusion of Useful Knowledge. 1839.

Schumpeter, J.A. 1942. *Capitalism, Socialism and Democracy*. New York: Harper & Bros.

Senior, N.W. 1836. *Political Economy*. London: W. Clowes.

Shepherd, W.G. 1982. Causes of increased competition in the US economy, 1939–1980. *Review of Economics and Statistics* 64, 613–26.

Smith, A. 1776. *The Wealth of Nations*. Glasgow edn, Oxford University Press, 1976.

Stigler, G.J. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65, 1–17.

Stigler, G.J. and Sherwin, R. 1985. The extent of the market. *Journal of Law and Economics* 28, 555–85.

Thornton, W.T. 1869. *On Labour*. London: Macmillan.

How to cite this article

Stigler, George J. "competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000261> doi:10.1057/9780230226203.0279

The New Palgrave Dictionary of Economics Online

computation of general equilibria (new developments)

Felix Kubler
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In this article, I review two recent developments in the theory of computation of general equilibria. First, following Brown, DeMarzo and Eaves (1996) several papers have developed globally convergent algorithms for the computation of general equilibria in models with incomplete asset markets. I review some of the developments in that area. Second, new developments in computational algebraic geometry lead to algorithms to compute effectively all equilibria of systems of polynomial equations. I point out some applications of these algorithms to general equilibrium theory.

Keywords

computation of general equilibria; Gröbner bases; homotopy algorithms; incomplete asset markets; Kuhn–Tucker conditions; multiple equilibria; Newton–Kantarovich conditions; real business cycles; semi-algebraic economies; Smale's alpha method; Tarski–Seidenberg th; uncertainty

Article

1 Introduction

After Scarf (1967) showed that there exist globally convergent (and effectively applicable) algorithms to compute economic equilibria, there is now a class of computable applied models which are routinely used to evaluate the economic consequences of different taxes and tariff structures (see, for example, Shoven and Whalley, 1992). Research on efficient algorithms for the computation of general equilibria in these models largely took place outside of economics.
A large literature in numerical analysis has developed algorithms that are much faster than Scarf's original method and that can be used for large-scale applications. Efficient iterative schemes, mostly based on global Newton methods, now allow applied researchers to solve for competitive equilibria in models with hundreds of commodities and agents (see, for example, Ferris and Pang, 1997).
Recently, there has been substantial research in theoretical computer science on the development of polynomial time algorithms for the computation of general equilibria. For most existing methods, the number of operations needed to approximate equilibria within a fixed precision ε grows exponentially in $1/\varepsilon$. Under restrictive assumptions on preferences, in models without production, researchers have developed algorithms to approximate equilibria ‘in polynomial time’, that is, the running time of the algorithm increases polynomially in the input parameters and in the precision with which equilibria are computed. Codenotti, Pemmaraju and Varadarajan (2004) give an overview on recent developments along this line.
In this article I will not discuss any of these practical aspects of the solution of large-scale models. I will instead focus on the following two unrelated developments in the computation of general equilibria in economics.

1. 1. The computation of equilibria in models with time, uncertainty and missing asset markets.
2. 2. The computation of all equilibria and the relationship between exact and approximate equilibria in the standard Arrow–Debreu model.

2 Models with asset markets

Due to their essential static nature, standard computable general equilibrium models suffer from an oversimplified treatment of uncertainty. Agents either solve a static problem or have myopic expectations, and the model can therefore not explicitly incorporate investment and saving decisions. The general equilibrium model with incomplete asset markets (GEI model) provides a basic framework with several agents and several commodities to incorporate uncertainty and financial markets. See, for example, Magill and Qunizii (1996) for an overview of the literature. The computation of equilibria in these models is challenging because in some specifications equilibria fail to exist while in others they are often numerically unstable.

Kehoe and Prescott (1995) argue that real business cycle models provide an alternative way to extend computable general equilibrium to models with time and uncertainty. There is now a large literature on the computation of equilibria in dynamic stochastic economies. This is reviewed elsewhere in this dictionary; see approximate solutions to dynamic models (linear methods); see also Judd (1998).

In the standard GEI model there are two time periods (Kubler and Schmedders, 2000, show how the problem of computation of equilibria in multi-period finance models can be essentially reduced to the two period case) and *S* possible states of the world in the second period. There are *L* perishable commodities available for trade at each state. There are *H* agents with endowments $e^h \in \mathbb{R}_+^{(S+1)L}$ and utility functions $u^h : \mathbb{R}_+^{(S+1)L} \to \mathbb{R}$. It is assumed throughout this article that utility functions are smooth in the sense of Debreu (1972) – that is, utility is C^2, strictly increasing, strictly quasi-concave, exhibits non-zero Gaussian curvature and indifference curves do not cut the axes.

There are *J* assets available for trade. In each state *s*, asset *j* pays a bundle of commodities $a_j(s) \in \mathbb{R}^L$. It is without loss of generality to assume that the $LS \times J$ matrix

$$A = \begin{pmatrix} a_1(1) & \dots & a_J(1) \\ \vdots & \ddots & \vdots \\ a_1(S) & \dots & a_J(S) \end{pmatrix}$$

has full rank *J*. Allowing assets to pay in different commodities is crucial when one wants to extend the model to several time periods and long-lived securities.

In the following, it will be useful to write commodity prices as

$$p = (p(0),\ p(1),\ \dots,\ p(S)) \in \Delta^{(S+1)L-1} = \left\{ p \in \mathbb{R}_+^{(S+1)L} : \sum_i p_i = 1 \right\},$$

and the $S \times J$ asset payoff matrix (as a function of spot prices $p(1) \dots p(S)$), $R(p)$, as

$$R(p) = \begin{pmatrix} p(1) \cdot a_1(1) & \dots & p(1) \cdot a_J(1) \\ \vdots & \ddots & \vdots \\ p(S) \cdot a_1(S) & \dots & p(S) \cdot a_J(S) \end{pmatrix}.$$

In part of the discussion we assume an exogenous short-sale constraint, that is, there is a number $0 < K \leq \infty$ such that the two-norm of an agent's portfolio must always be less than or equal to *K*. One can then write an agent's aggregate excess demand function as the solution of his maximization problem in the GEI economy.

$$(z^h(p),\ \varphi^h(p)) = \arg \max_{z \in \mathbb{R}^{L(S+1)},\ \varphi \in \mathbb{R}^J} u(e^h + z) \text{ s.t. } p \cdot z = 0 (p(1) \cdot z(1),\ \dots,\ p(S) \cdot z(S))^T = R(p) \cdot \varphi \| \varphi \| \leq K.$$

A GEI equilibrium is a collection of prices, portfolios and a consumption allocation such that markets clear and each agent maximizes her utility, i.e. equilibrium prices *p* are characterized by $\Sigma_{h=1}^{H} z^h(p) = 0$.
In a slight idealization (see also the more precise definition in the next section), we assume that the maximization problem can be solved exactly and we define an ε -equilibrium as a price $\bar{p}$ such that

$$\left\| \sum_{h=1}^{H} z^h(\bar{p}) \right\| < \varepsilon.$$

2.1 A general algorithm

Although generally *R*(*p*) will have full rank *J*, there will be so-called ‘bad prices’ at which the rank of *R*(*p*) drops. When there are no short sale constraints, that is, $K = \infty$, this leads to a discontinuity of excess demand. Scarf's algorithm fails: no matter how fine the simplicial subdivision, if the algorithm terminates at some $\bar{p}$, one cannot necessarily infer a bound on $\|z(\bar{p})\|$ and hence cannot find an ε -equilibrium.
Homotopy continuation methods (see Garcia and Zangwill, 1981; Eaves, 1972) turn out to be ideally suited for this numerical problem. In order to solve a system of equations $f(x) = 0$, $f: X \to Y$, the basic idea underlying homotopy methods is to find a smooth map $H: X \times [0, 1] \to Y$ with

$$H(x, 1) \equiv f(x) \text{ and } H(x, 0) \equiv g(x),$$

where $g: X \to Y$ has a known unique zero. The map *H* is called a smooth homotopy. In using homotopy methods it is crucial to set up the function, *H*, to ensure that there is a smooth path that connects (x^s,0) with $g(x^s) = 0$ to some $(\bar{x}, 1)$ with $f(\bar{x}) = 1$.
Brown, DeMarzo and Eaves (1996) develop a homotopy algorithm which can be shown to be globally convergent in that it finds an ε -equilibrium for any $\varepsilon > 0$ in a finite number of steps. Following the so-called Cass-trick, it is useful to introduce an unconstrained agent, that is, to define the first agent maximization problem as

$$z^u(p) = \arg\max_z u^1(e + z) \text{ s.t. } p \cdot z = 0,$$

and aggregate demand as $z(p) = z^u(p) + \Sigma_{h=2}^{H} z^h(p)$. Note that $\bar{p}$ is a GEI equilibrium (given that $K = \infty$) if and only if $z(p) = 0$. An ε -equilibrium is characterized by $\|z(p)\| < \varepsilon$.
Define the expenditure of the unconstrained agent y^u as

$$y^u = (p(1) \cdot z_1^u(p), \ldots, p(S) \cdot z_S^u(p)).$$

Define an extended payoff matrix $R^*(p)$ by

$$R^*(p) = [R(p), y^u(p)]$$

and let $R^*_{-i}(p)$ be $R^*(p)$ with the i'th column deleted. For the constrained agents $h = 2, \ldots, H$ define

$$z^h(p, R^*_{-i}(p)) = \arg\max_{z,\varphi} u^h(e^h + z) \text{ s.t. } p \cdot z = 0$$

$$(p(1) \cdot z(1), \ldots, p(S) \cdot z(S))^T = R^*_{-i}(p) \cdot \varphi.$$

Now consider a family of homotopies, indexed by i

$$H_i(p, t, \theta) = \begin{pmatrix} z^u(p) + t\sum_{h=2}^{H} z^h(p, R^*_{-i}(p)) \\ R^*(p)\theta \\ \theta \cdot \theta - 1 \end{pmatrix}.$$

To prove existence of a homotopy path, Brown, DeMarzo and Eaves (1996) show that $\cup_{i=1}^{J+1} H_i^{-1}(0)$ contains a smooth path connecting the starting point to a solution at $t = 1$.

While generically in endowments a homotopy path turns out to exist, the algorithm is hardly applicable in medium-sized problems, since the number of homotopies one has to consider can become quite large. An alternative is to focus on models with $K < \infty$ (or alternatively models with transaction costs) or to consider algorithms which might fail in a small class of problems but which are generally more efficient.

2.2 Short-sale constraints

In the presence of short-sale constraints, the excess demand function is continuous and equilibrium existence can be proven with Brouwer's theorem. Therefore, one could presumably use a version of Scarf's algorithm to compute equilibria in this case. However, while there are no new mathematical problems to be solved, the fact that the rank of the asset–payoff matrix can still collapse in equilibrium poses difficult numerical problems. Simple Newton method-based algorithms often do not work (see Kubler and Schmedders, 2000) unless one has a starting point very close to the actual solution. It turns out that, just as in the problem without short-sale constraints, homotopy continuation methods can provide a basis for reliable algorithms.

Schmedders (1998) develops a homotopy algorithm which can be used to solve models with a large number of heterogeneous households and goods. The basic idea of his algorithm is to modify the agents' problem by introducing a homotopy parameter $t \in [0, 1]$ as follows.

$$(z^h(p, t), \varphi^h(p, t)) = \arg \max_{z \in \mathbb{R}^{L(S+1)}, \varphi \in \mathbb{R}^J} u(e^h + z) - (1 - t)\frac{1}{2}\|\varphi\|^2 \text{ s.t. } p \cdot z = 0 (p(1) \cdot z(1), \ldots, p(S) \cdot z(S)) = R(p) \cdot \varphi \|\varphi\| \leq K.$$

Under the assumptions on utilities this is still a convex problem and the first order Kuhn–Tucker conditions are necessary and

sufficient. Schmedders provides various examples that show that even for $K = \infty$ his algorithm, although not guaranteed to converge, performs well in practice.

For $K < \infty$, the Kuhn–Tucker inequalities can be converted into a system of equalities via a change of variables (see Garcia and Zangwill, 1981, ch. 4). Kubler (2001), Herings and Schmedders (2006) and others subsequently used this idea to solve models with transaction costs, trading constraints and other market imperfections.

Of course, it is an important practical problem how to trace out a homotopy path numerically. See Watson (1979) for a theoretical algorithm. For a practical description of numerical homotopy path-following methods see Schmedders (2004).

3 Equilibria in semi-algebraic economies

While it is clear that sufficient assumptions for the global uniqueness of competitive equilibria are too restrictive to be applicable to models used in practice, it remains an open problem how serious a challenge the non-uniqueness of competitive equilibrium poses to applied equilibrium modelling. In the presence of multiple equilibria, comparative statics exercises become meaningless. Furthermore, even when for a given specification of the economy equilibrium is globally unique, as Richter and Wong (1999) point out, the possibility of multiple equilibria for close-by economies implies that it is generally impossible to compute prices and allocations that are close-by exact equilibrium prices and allocations (as opposed to computing prices at which aggregate excess demand is close to zero). In this section I argue that one can solve these problems by focusing on so-called ‘semi-algebraic’ economies.

While the arguments are also applicable to the GEI model, for simplicity, consider a standard Arrow–Debreu exchange economy, $(u^h, e^h)_{h=1}^H$. There are *H* agents trading *L* commodities. Each agent *h* has individual endowments $e^h \in \mathbb{R}_+^L$ and ‘smooth preferences’ characterized by an utility function $u^h: \mathbb{R}_+^L \to \mathbb{R}$.

A Walrasian equilibrium is a collection of consumption vectors $(x^h)_{h=1}^H$ and prices $p \in \Delta^{L-1}$ such that

$$x^h \in \arg\max_{x \in \mathbb{R}_+^L} u^h(x) \text{ s.t. } p \cdot x \leq p \cdot e^h \tag{1}$$

$$\sum_{h=1}^H (x^h - e^h) = 0. \tag{2}$$

An approximate (ε -) equilibrium consists of an allocation an prices such that

$$\| u^h(x^h) - [\max_{x \in \mathbb{R}_+^L} u^h(x) \text{ s.t. } p \cdot x \leq p \cdot e^h] \| < \varepsilon \tag{3}$$

$$\left\| \sum_{h=1}^H (x^h - e^h) \right\| < \varepsilon. \tag{4}$$

Given any $\varepsilon > 0$, Scarf's algorithm (as well as the more efficient algorithms used in practice) finds a p, x^h which constitute an ε -equilibrium.
This leaves open two important theoretical questions.

1. 1. Can one relate the approximate equilibrium prices and allocations, to exact equilibria, that is, given a computed ε -equilibrium $(\bar{p}, (\bar{x}^h))$, does there exist a Walrasian equilibrium $\tilde{p}, (\tilde{x}^h)$ with $\|(\bar{p}, (\bar{x}^h)) - (\tilde{p}, (\tilde{x}^h))\|$ small? Can one find good bounds on this distance which tend to zero as $\varepsilon \to 0$?
2. 2. Given an economy $(u^h, e^h)_{h=1}^H$ with N Walrasian equilibria $(p^n, (x^h)^n)_{n=1}^N$ and any $\delta > 0$, is it possible to approximate all N equilibria, that is, to find N ε -equilibria $(\tilde{p}^n, (\tilde{x}^h)^n)_{n=1}^N$ with $\|(p^n, (x^h)^n) - (\tilde{p}^n, (\tilde{x}^h)^n)\| < \delta$, for all $n = 1, \ldots, N$?

Clearly, the second problem is strictly more difficult to tackle than the first. Richter and Wong (1999) show that for general economies even the answer to the first question is negative. In order to obtain positive answers to both qsts, one needs to restrict possible preferences. One approach is to assume that better sets are semi-algebraic sets. I will make the slightly more useful assumption that marginal utilities are semi-algebraic functions.

3.1 Semi-algebraic economies

We assume that for each h, $D_x u^h(x)$ is a semi-algebraic function, that is, its graph $\left\{(x, y) \in \mathbb{R}_+^{2L} : y = D_x u^h(x)\right\}$ is a finite union and intersection of sets of the form

$$\left\{(x, y) \in \mathbb{R}^{2L} : g(x, y) > 0\right\} \text{ or } \left\{(x, y) \in \mathbb{R}^{2L} : f(x, y) = 0\right\}$$

for polynomials with real coefficients, f and g.
For practical purposes, the focus on semi-algebraic preferences is quite general. First note that Afriat's theorem implies that a finite set of observations on an individual's choices that can be rationalized by any utility function can also be rationalized by semi-algebraic preferences (in fact, Afriat's construction is piece-wise linear). Furthermore, note that the constant elasticity of substitution utility function which is often used in applied work is semi-algebraic if the elasticities of substitution are rational numbers.
It follows from the Tarski–Seidenberg theorem that for semi-algebraic economies the answers to both qsts above are positive, since the relevant statements can be written as first order sentences (see Basu, Pollack and Roy, 2003). However, algorithmic quantifier elimination which needs to be used to answer general qsts in this framework is so computationally inefficient that for practical purposes this does not help towards solving the above qsts for interesting specifications of economies.
Nevertheless, given a semi-algebraic economy it is possible to find a system of polynomial equations $f(x) = 0$, $f : \mathbb{R}^{H(L+1)+L-1} \to \mathbb{R}^{H(L+1)+L-1}$, and finitely many inequalities $g^i(x) \geq 0$, $g^i : \mathbb{R}^{H(L+1)+L-1} \to \mathbb{R}^M$, $i = 1, \ldots, N < \infty$ such that p, (x^h) is a Walrasian equilibrium for the economy (u^h, e^h) if and only if there exist $\lambda^h \in \mathbb{R}_{++}$, $h = 1, \ldots, H$ such that for some $i = 1, \ldots, N$,

$$f(p, (x^h, \lambda^h)) = 0, \; g^i(p, (x^h, \lambda^h)) \geq 0.$$

Therefore, the problem of finding Walrasian equilibria reduces to finding the real roots of polynomial systems of equations and verifying polynomial inequalities (see Kubler and Schmedders, 2006).
Having reduced the problem of finding Walrasian equilibria to finding roots of a polynomial system of equations, one can then answer the two qsts above affirmatively.

3.2 Question 1: Smale's alpha method

Smale's alpha method provides a simple sufficient conditions for approximate zeros to be close to exact zeros and can be viewed as an extension of the Newton–Kantarovich conditions. The following results are from Blum et al. (1998, ch. 8).

Let $D \subset \mathbb{R}^n$ be open and let $f : D \to \mathbb{R}^n$ be analytic. For $z \in D$, define $f^{(k)}(z)$ to be the k'th derivative of f at z. This is a multi-linear operator which maps k-tuples of vectors in D into $\mathbb{R}^n$. Define the norm of an operator A to be

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Suppose that the Jacobian of f at z, $f^{(1)}(z)$ is invertible and define

$$\gamma(z) = \sup_{k \geq 2} \left\| \frac{(f^{(1)}(z))^{-1} f^{(k)}(z)}{k!} \right\|^{\frac{1}{(k-1)}}$$

and

$$\beta(z) = \|(f^{(1)}(z))^{-1} f(z)\|.$$

Theorem 1: Given a $\bar{z} \in D$, suppose the ball of radius $\left(1 - \frac{\sqrt{2}}{2}\right) / \gamma(\bar{z})$ around $\bar{z}$ is contained in D and that

$$\beta(\bar{z})\gamma(\bar{z}) < 0.157.$$

Then there exists a $\hat{z} \in D$ with

$$f(\hat{z}) = 0 \text{ and } \|\bar{z} - \hat{z}\| \leq 2\beta(\bar{z}).$$

While the theorem applies to any locally analytic function, the bound $\gamma(z)$ can in general only be obtained if the system is in fact polynomial. For this case, the bound can be computed fairly easily. Given an ε-equilibrium the result gives an immediate bound on the distance between the approximation and an exact Walrasian equilibrium, hence answering Question 1 above.

3.3 Question 2: Polynomial system solving

In the following, I denote the collection of all polynomials in the variable $x_1, x_2, \ldots, x_n$ with coefficients in a field $\mathbb{K}$ by $\mathbb{K}[x_1, \ldots, x_n]$. The for this survey relevant examples of $\mathbb{K}$ are the field of rational numbers $\mathbb{Q}$, the field of real numbers $\mathbb{R}$, and the field of complex numbers $\mathbb{C}$. Polynomials over the field of rational numbers are computationally convenient since modern computer algebra systems perform exact computations over the field $\mathbb{Q}$. Economic parameters are typically real numbers, and equations characterizing equilibria lie in $\mathbb{R}[x]$. The algorithms to compute all solutions to polynomial systems always compute all solutions in an algebraically closed field, in this case $\mathbb{C}[x]$.

Given a polynomial system of equations $f : \mathbb{C}^M \to \mathbb{C}^M$ there is now a variety of algorithm to approximate numerically all complex and real zeros of *f*. Sturmfels's monograph (2002) provides an excellent overview. In this survey I briefly mention two possible approaches, homotopy continuation methods and solution methods based on Gröbner bases.
At the writing of this article, both approaches are too inefficient to be applicable to large economic models, but they can be used for models with four or five households and four or five commodities. To find all equilibria for a given economy, homotopy methods seem slightly more efficient, while Gröbner bases allow for statements about entire classes of economies.

3.3.1 All solution homotopies

Solving polynomial systems numerically means computing approximations to all isolated solutions. Homotopy continuation methods can provide paths to all approximate solutions. There are well-known bounds on the maximal number of complex solutions of a polynomial system. The basic idea is to start at a generic polynomial system $g(x)$ whose number of roots is at least as large as the maximal number of solutions to $f(x) = 0$ and whose roots are all known. Then one needs to trace out all paths (in complex space) of the homotopy $H(x, t) = tg(x) + (1 - t) f(x)$, which do not diverge to infinity. Smale's alpha method can be applied along the path to ensure that the approximate solutions are close to real exact solutions (see Blum et al., 1998). It can be shown that all solutions to $f(x) = 0$ can be found in this manner.
Sommese and Wampler (2005) provide a detailed overview. Applications of these methods in economics have so far been largely restricted to game theory, but the method is also applicable to Walrasian equilibria.

3.3.2 Gröbner basis

For given polynomials $f_1, \ldots, f_k$ in $\mathbb{Q}[x]$ the set

$$I = \left\{ \sum_{i=1}^{k} h_i f_i : h_i \in \mathbb{Q}[x] \right\} = (f_1, \ldots, f_k)$$

is called the ideal generated by $f_1, \ldots, f_k$. It turns out that under conditions which can often be shown to hold in practice, the so-called ‘reduced Gröbner basis’ of this ideal, *I*, in the lexicographic term order has the shape

$$G = \{x_1 - q_1(x_n), x_2 - q_2(x_n), \ldots, x_{n-1} - q_{n-1}(x_n), r(x_n)\}$$

where *r* is a polynomial of degree d and the q_i are polynomials of degree $d - 1$.
This basis can be computed exactly, using Buchberger's algorithm (recently, much more efficient versions of the basic algorithm have been developed; see for example Faugère, 1999). The number of real solutions to the original system then equals the number of real solutions of the univariate polynomial *r*(.) which can be determined exactly by Sturm's method (see Sturmfels, 2002, for details). The roots of $r(\cdot)$ can be approximated numerically with standard methods and the remaining solution to the original system is linear in these roots.
Kubler and Schmedders (2006) use the method to test for uniqueness of equilibria in semi-algebraic classes of economies.

See Also

- approximate solutions to dynamic models (linear methods)
- computation of general equilibria
- general equilibrium

Bibliography

Basu, S., Pollack, R. and Roy, M.-F. 2003. *Algorithms in Real Algebraic Geometry*. Springer Verlag: New York.

Blum, L., Cucker, F., Shub, M. and Smale, S. 1998. *Complexity and Real Computation*. Springer Verlag: New York.

Brown, D.J., DeMarzo, P.M. and Eaves, B.C. 1996. Computing equilibria when asset markets are incomplete. *Econometrica* 64, 1–27.

Codenotti, B., Pemmaraju, S. and Varadarajan, K. 2004. Algorithms column: the computation of market equilibria. *ACM SIGACT News* 35(4), 23–37.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40, 603–15.

Eaves, B.C. 1972. Homotopies for the computation of fixed points. *Mathematical Programming* 3, 1–22.

Faugère, J.C. 1999. A new efficient algorithm for computing Gröbner bases (f4). *Journal of Pure and Applied Algebra* 139, 61–88.

Ferris, M.C. and Pang, J.S. 1997. Engineering and economic applications of complementarity problems. *SIAM Review* 39, 669–713.

Garcia, C. and Zangwill, W. 1981. *Pathways to Solutions, Fixed Points, and Equilibria*. Englewood Cliffs, NJ: Prentice Hall.

Herings, P.J.J. and Schmedders, K. 2006. Computing equilibria in finance economies with incomplete markets and transaction costs. *Economic Theory* 27, 493–512.

Judd, K. 1998. *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Kehoe, T.J. and Prescott, E.C. 1995. Introduction to the symposium, the discipline of applied general equilibrium. *Economic Theory* 6, 1–11.

Kubler, F. 2001. Computable general equilibrium with financial markets. *Economic Theory* 18, 73–96.

Kubler, F. and Schmedders, K. 2000. Computing equilibria in stochastic finance economies. *Computational Economics* 15, 145–72.

Kubler, F. and Schmedders, K. 2006. Uniqueness of equilibria in semi-algebraic economies. Discussion paper, Northwestern University.

Magill, M.J.P. and Qunizii, M. 1996. *Theory of Incomplete Markets*. Cambridge, MA: MIT Press.

Richter, M.K. and Wong, K.-C. 1999. Non-computability of competitive equilibrium. *Economic Theory* 14, 1–27.

Scarf, H. 1967. On the computation of equilibrium prices. In *Ten Economic Studies in the Tradition of Irving Fisher*, ed. W.J. Fellner. New York: Wiley.

Schmedders, K. 1998. Computing equilibria in the general equilibrium model with incomplete asset markets. *Journal of Economic Dynamics and Control* 22, 1375–403.

Schmedders, K. 2004. Homotopy path-following with easyhomotopy: solving nonlinear equations for economic models. Working paper, Northwestern University.

Shoven, J.B. and Whalley, J. 1992. *Applying General Equilibrium*. Cambridge: Cambridge University Press.

Sommese, A.J. and Wampler, C.W. 2005. *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*. Singapore: World Scientific Press.

Sturmfels, B. 2002. *Solving Systems of Polynomial Equations*. CBMS Regional Conference Series in Mathematics No. 97. Providence, RI: American Mathematical Society.

Watson, L.T. 1979. A globally convergent algorithm for computing fixed points of C2 maps. *Applied Mathematics and Computation* 5, 297–311.

How to cite this article

Kubler, Felix. "computation of general equilibria (new developments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000564> doi:10.1057/9780230226203.0284

The New Palgrave Dictionary of Economics Online

computation of general equilibria

Herbert E. Scarf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Walrasian model of economic equilibrium is a generalization to the entire economy of the basic notion that prices move to levels that equilibrate supply and demand. Although the model avoids some factors of economic significance, it is extremely useful in helping us evaluate the effects of changes in economic policy or the economic environment. A moderately realistic model designed to illustrate a significant economic issue typically involves a large system of highly nonlinear equations and inequalities. Existence of a solution is demonstrated by non-constructive fixed point theorems. The explicit numerical solution of such a model requires sophisticated computational techniques.

Keywords

Arrow–Debreu model; Barone, E.; Brouwer's fixed-point theorem; Cobb–Douglas function; computation of general equilibria; general equilibrium; Harberger, A.; Johansen, L.; Kakutani's fixed-point theorem; Kuhn–Tucker Theorem; Lange, O.; non-convexity; Sperner's lemma; technical coefficients of production; uncertainty; Walras's Law; Walrasian model

Article

The general equilibrium model, as elaborated by Walras and his successors, is one of the most comprehensive and ambitious formulations in the current body of economic theory. The basic ingredients with which the Walrasian model is constructed are remarkably spare: a specification of the asset ownership and preferences for goods and services of the consuming units in the economy, and a description of the current state of productive knowledge possessed by each of the firms engaged in manufacturing or in the provision of services. The model then yields a complete determination of the course of prices and interest rates over time, levels of output and the choice of techniques by each firm, and the distribution of income and patterns of saving for each consumer.

The Walrasian model is essentially a generalization, to the entire economy and to all markets simultaneously, of the ancient and elementary notion that prices move to levels which equilibrate supply and demand. No intellectual construction of this scope, designed to address basic questions in a subject as complex and elusive as economics, can be described as simply true of false – in the sense in which these terms are used in mathematics or perhaps in the physical sciences. The assertions of economic theory are not susceptible to crisp and immediate experimental verification. Moreover, the Walrasian model disregards obvious aspects of human motivation which are of the greatest economic significance and which cannot be addressed in the language of our subject: economic theory is mute about our affective lives, about our opposing needs for community and individual assertion, and about the non-pecuniary determinants of entrepreneurial energy.

There are, in addition, aspects of economic reality which are capable of being described in the framework of the Walrasian model but which must be assumed away in order for the model to yield a determinate outcome. Uncertainty about the future is an ever-present fact of economic life, and yet the complete set of markets for contingent commodities required by the Arrow–Debreu treatment of uncertainty is not available in practice. Economies of scale in production are a central feature in the rise of the large manufacturing entities which dominate modern economic activity; their incorporation into the Walrasian model requires the introduction of non-convex production possibility sets for which the competitive equilibrium will typically fail to exist.

In spite of its many shortcomings, the Walrasian model – if used with tact and circumspection – is an important conceptual framework for evaluating the consequences of changes in economic policy or in the environment in which the economy finds itself. The effects of a major shock to the economy of the United States – such as the four-fold increase in the price

of imported oil which occurred in late 1973 – can be studied by contrasting equilibrium prices, real wages and the choice of productive techniques both before and after the event in question. Generations of economists have used the Walrasian model to analyse the terms of trade, the impact of customs unions, changes in tariffs and a variety of other issues in the theory of International Trade. And much of the literature in the field of Public Finance is based on the assumption that the competitive model is an adequate description of economic reality.

In these discussions the analysis is frequently conducted in terms of simple geometrical diagrams whose use places a severe restriction on the number of consumers, commodities and productive sectors that can be considered. This is in contrast to formal mathematical treatments of the Walrasian model, which permit an extraordinary generality in the elaboration of the model at the expense of immediate geometrical visualization. Unfortunately, however, it is only under the most severe assumptions that mathematical analysis will be capable of providing unambiguous answers concerning the direction and magnitude of the changes in significant economic variables, when the system is perturbed in a substantial fashion. In order for a comparative analysis to be carried out in a multi-sector framework it is necessary to employ computational techniques for the explicit numerical solution of the highly non-linear system of equations and inequalities which represent the general Walrasian model.

The use of fixed-point theorems in equilibrium analysis

One of the triumphs of mathematical reasoning in economic theory has been the demonstration of the existence of a solution for the general equilibrium model of an economy, under relatively mild assumptions on the preferences of consumers and the nature of production possibility sets (see Debreu, 1982). The arguments for the existence of equilibrium prices inevitably make use of Brouwer's fixed-point theorem, or one of its many variants, and any effective numerical procedure for the computation of equilibrium prices must therefore be capable of computing the fixed points whose existence is asserted by this mathematical statement.

Brouwer's fixed-point theorem, enunciated by the distinguished Dutch mathematician L.E.J. Brouwer in 1912, is the generalization to higher dimensions of the elementary observation that a continuous function of a single variable which has two distinct signs at the two endpoints of the unit interval, must vanish at some intermediary point. In Brouwer's Theorem the unit interval is replaced by an arbitrary closed, bounded convex set S in R^n, and the continuous function is replaced by a continuous mapping of the set S into itself: $x \to g(x)$. Brouwer's Theorem then asserts the existence of at least one point x which is mapped into itself under the mapping; that is, a point x for which $x = g(x)$. To see how this conclusion is used in solving the existence problem let us begin by specifying, in mathematical form, the basic ingredients of the Walrasian model. (Figure 1)

Figure 1

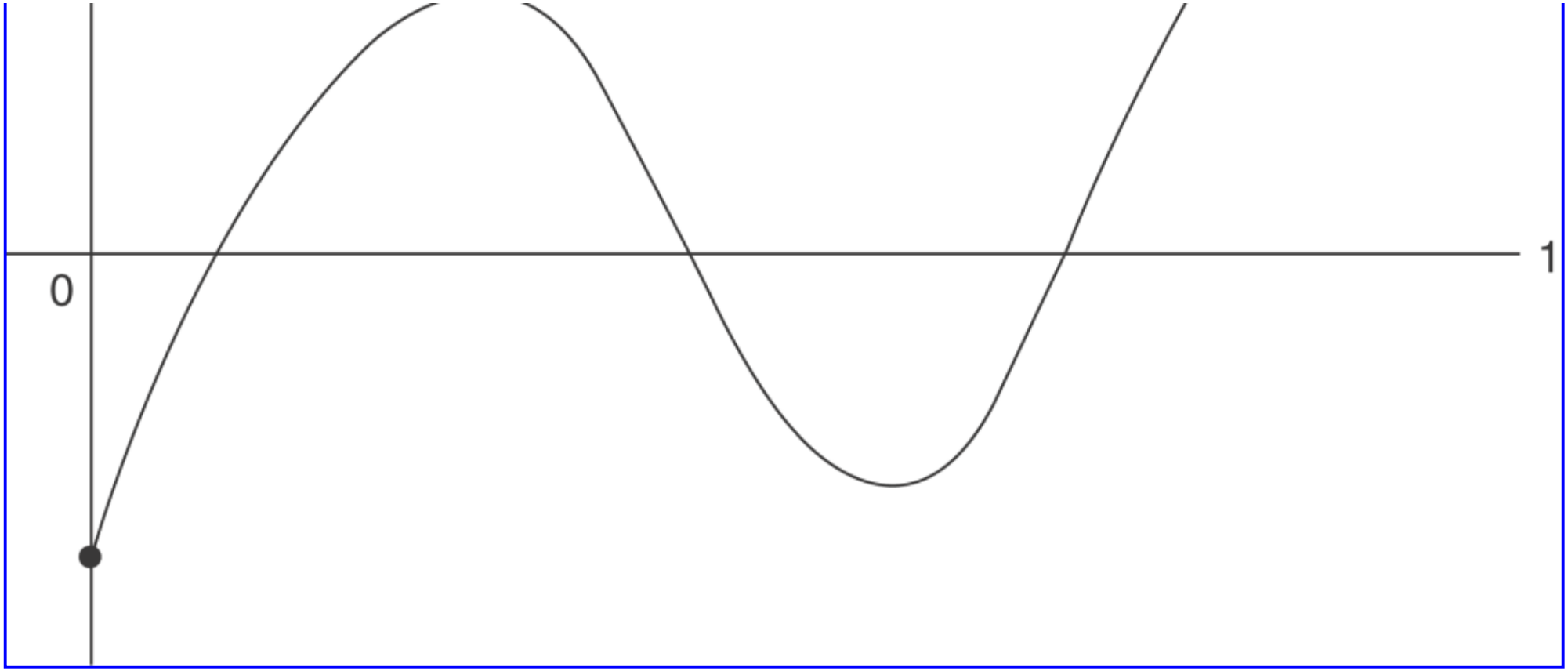

The typical consumer is assumed to have a preference order for, say, the non-negative commodity bundles $x = (x_1, x_2, \ldots, x_n)$ in R^n; the preference ordering is described either by a specific utility function $u(x_1, x_2, \ldots, x_n)$ or by means of an abstract representation of preferences. The consumer will also possess, prior to production and trade, a vector of initial assets $w = (w_1, w_2, \ldots, w_n)$. When a non-negative price vector $p = (p_1, p_2, \ldots, p_n)$ is announced the consumer's income will be $I = p \cdot w$ and his demands will be obtained by maximizing preferences subject to the budget constraint $p \cdot x \leq p \cdot w$. If the preferences satisfy sufficient regularity assumptions, the consumer's demand functions $x(p)$ will be single-valued functions of p, continuous (except possibly when some of the individual prices are zero), homogeneous of degree zero and will satisfy the budget constraint $p \cdot x(p) = p \cdot w$. (Figure 2)

Figure 2

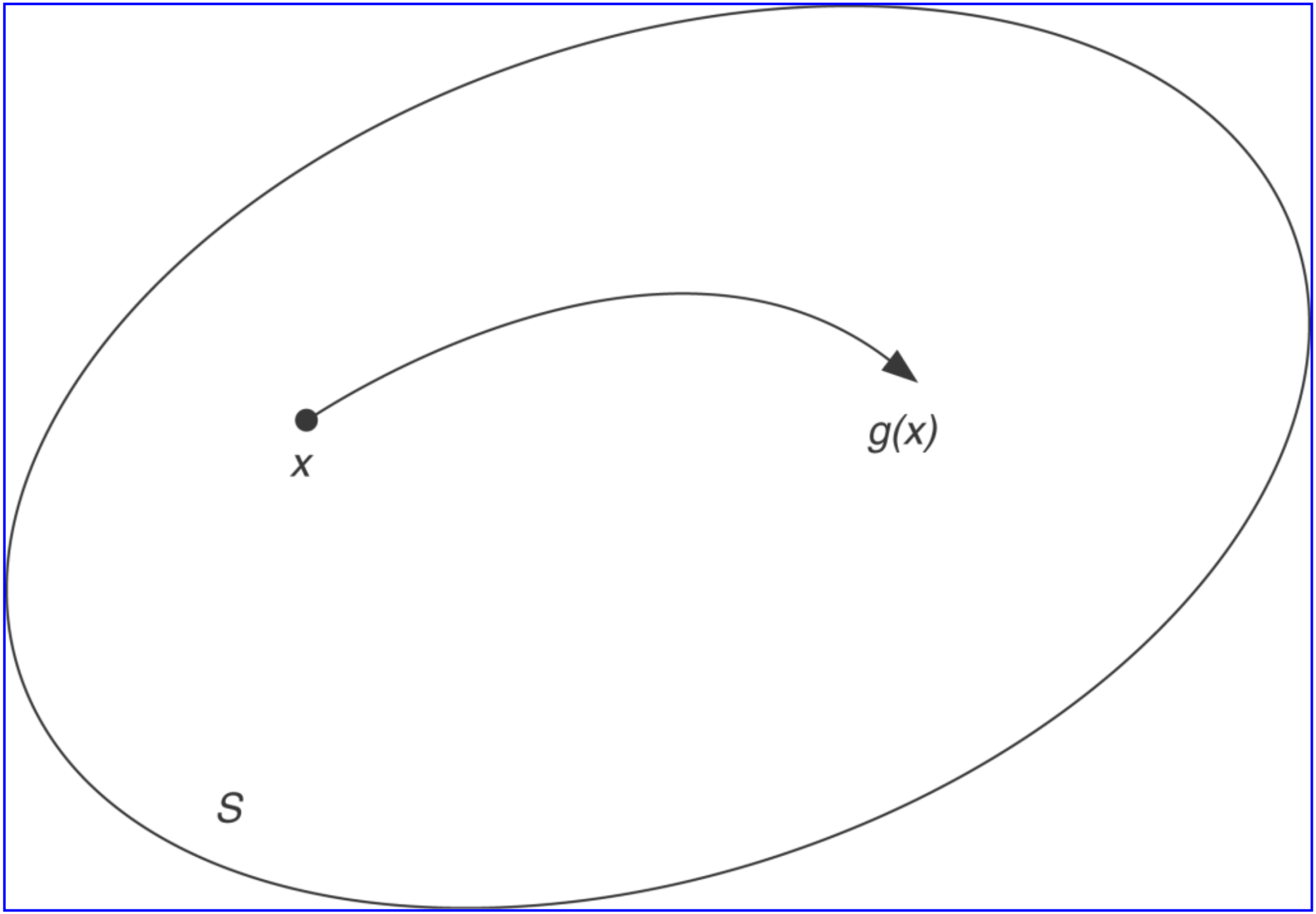

The market demands are obtained by aggregating over individual demand functions and, as such, will inherit the properties described above. The market *excess* demand functions, which I shall denote by $f(p)$, arise by subtracting the supply of assets owned by all consumers from the demand functions themselves. It is these functions which are required for a complete specification of the consumer side of the economy in the general equilibrium model: they may be obtained either by the aggregation of individual demand functions – as we have just described – or they may be directly estimated from econometric data. The following properties will hold, either as a logical conclusion or by assumption:

1. 1. *f*(*p*) is homogeneous of degree zero.
2. 2. *f*(*p*) is continuous in the interior of the positive orthant.
3. 3. *f*(*p*) satisfies the Walras Law $p \cdot f(p) = 0$.

The first of these properties permits us to normalize prices in any one of several ways; for example, $\Sigma p_j = 1$ or $\Sigma p_j^2 = 1$. Given either of these normalizations, I personally do not find it offensive to extend the property of continuity to the boundary, even though there are elementary examples of utility functions, such as the Cobb–Douglas function, for which this would not be correct.

The production side of the economy requires for its description a complete specification of the current state of technical knowledge about the methods of transforming inputs into outputs – with commodities differentiated according to their location and the time of their availability. This can be done by means of production functions, an input/output table with substitution possibilities and several scarce factors rather than labour alone, or by a general activity analysis model:

$$A = \begin{bmatrix} -1 & 0 & \cdots & 0 & a_{1,n+1} & a_{1,k} \\ 0 & -1 & & 0 & a_{2n+1} & a_{2,k} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & & -1 & a_{n,n+1} & a_{n,k} \end{bmatrix}$$

Each column of A describes a particular productive process, with inputs represented by non-negative entries and outputs by positive entries. The activities are assumed capable of being used simultaneously and at arbitrary non-negative levels $x = (x_1, x_2, \ldots, x_k)$; the net production plan is then $y = Ax$. (Figure 3)

Figure 3

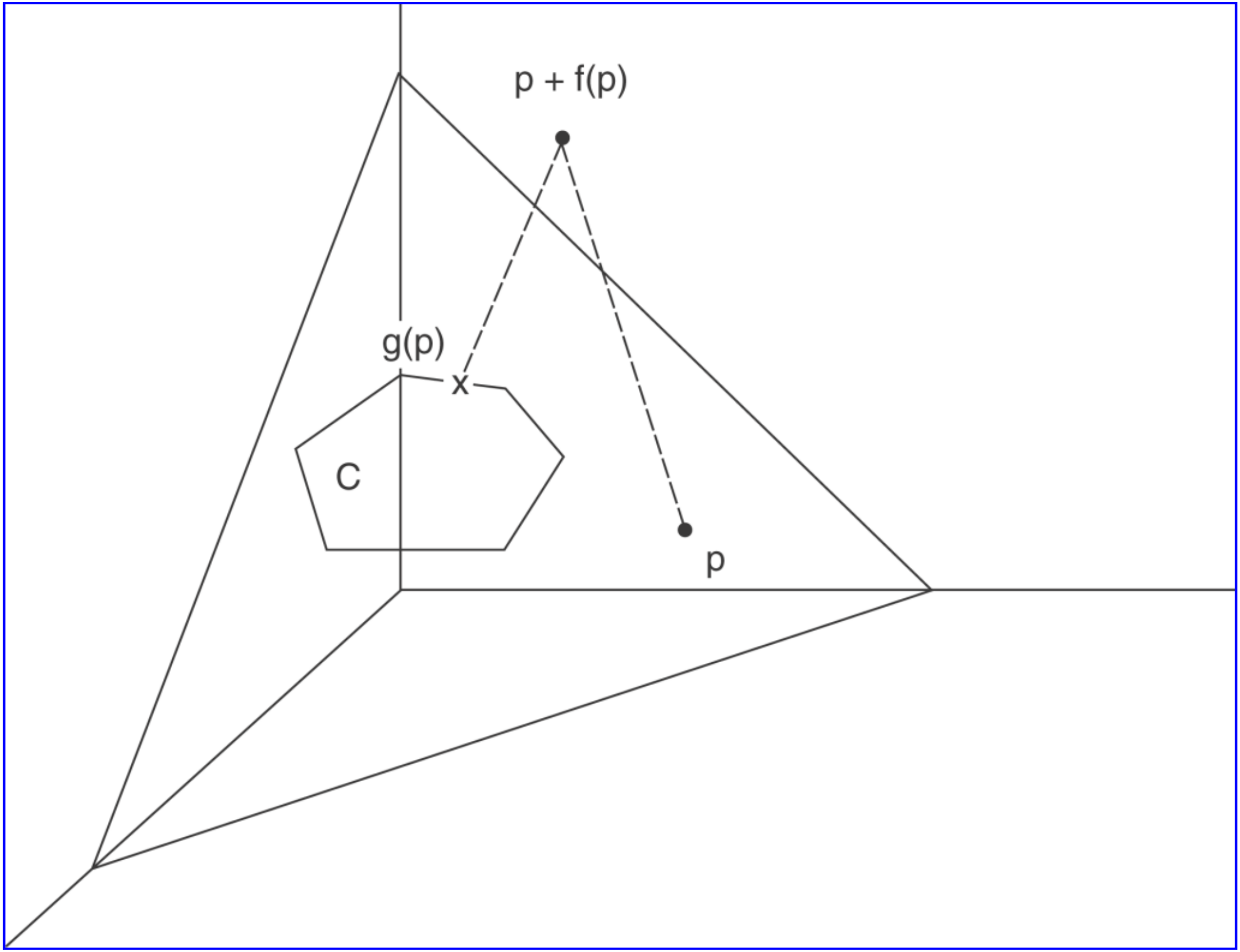
p + f(p)
g(p)
x
C
p

With this formulation, a competitive equilibrium is defined by a non-negative vector of prices $p = (p_1, p_2, \ldots, p_n)$ and a non-negative vector of activity levels $x = (x_1, x_2, \ldots, x_k)$ satisfying the following conditions:

1. 1. $f(p) = Ax$,
2. 2. $pA \leq 0$.

The first condition states that supply and demand are equal in all markets, and the second that there are not opportunities for positive profits when the profitability of each activity is evaluated at the equilibrium prices. Taken in conjunction with the Walras's Law, these conditions imply that those activities which are used at a positive level in the equilibrium solution make a profit of zero.

Given the assumption of continuous and single-valued excess demand functions and the description of the production possibility set by means of an activity analysis model, the following rather direct application of Brouwer's Theorem is sufficient to demonstrate the existence of a equilibrium solution. Under weaker assumptions on the model, variants such as Kakutani's Fixed-Point Theorem may be required.

Let prices be normalized so as to lie on the unit simplex $S = \{p = (p_1, p_2, \ldots, p_n) | p_i \geq 0, \Sigma p_i = 1\}$. The set of prices *p* for which $pA \leq 0$ is termed the *dual* cone of the production possibility set generated by the activity matrix A. Its intersection with the unit simplex is a convex polyhedron C consisting of those normalized prices which yield a profit less than or equal to zero for all activities.

We construct a continuous mapping of *S* into itself as follows: for each *p* in *S* consider the point $p + f(p)$; a point which is generally not on the unit simplex itself. We then define *g* (*p*) – the image of *p* under the mapping – to be that point in C which is closest, in the sense of Euclidean distance, to $p + f(p)$. It is then an elementary application of the Kuhn–Tucker Theorem to show that a fixed point of this mapping is, indeed, an equilibrium price vector.

The equilibrium model as a tool for policy evaluation

Brouwer's original proof of his theorem was not only difficult mathematically, but it was decidedly non-constructive; it offered no method for effectively computing a fixed point of the mapping. Brouwer did, in fact, reject his own argument during the later ‘intuitionist’ phase of his career, in which he proclaimed the acceptability of only those mathematical conclusions obtained by constructive procedures. In spite of the many simplifications in the proof of Brouwer's Theorem offered during the subsequent half-century, it was not until the mid-1960s that constructive methods for approximating fixed points of a continuous mapping finally made their appearance on the scene (Scarf, 1967) – aided by the development of the modern electronic computer and by the rapid methodological advances in the discipline of operations research.

In the early decades of this century, the question of the explicit numerical solution of the general equilibrium model was an active topic of discussion – not by numerical analysts – but rather by economists concerned with the techniques of economic planning in a socialist economy. The issue was raised in the remarkable paper published by Enrico Barone in 1908, entitled ‘The Ministry of Production in a Socialist Economy’. Barone, and subsequently Oskar Lange (1936), accepted the Walrasian model – with suitable transfers of income – as an adequate description of ideal economic activity in an economy in which the means of production were collectively owned. In the absence of markets, prices, levels of output and the choice of productive techniques were to be obtained by an explicit numerical solution of the Walrasian system. A key feature of Barone's analysis was the concept of the ‘technical coefficients of production’ – the input/output coefficients associated with those activities in use at equilibrium. Barone's contention was that the equilibrium could be found – by an extremely laborious calculation which might indeed claim a significant share of the national product – only if the correct activities were known in advance. For Barone, rational economic calculation in a socialist economy was defeated by the many opportunities for substitution in production: the particular activities in use at equilibrium would be impossible to determine by a prior computation. It is instructive to quote Barone on this point.

> The determination of the coefficients economically most advantageous can only be done in an *experimental* way: and not on a *small scale*, as could be done in a laboratory; but with experiments on a *very large scale*, because often the advantage of the variation has its origin precisely in a new and greater dimension of the undertaking. Experiments may be successful in the sense that they may lead to a lower cost combination of factors; or they may be unsuccessful, in which case the particular organization may not be copied and repeated and others will be preferred, which *experimentally* have given a better result.
>
> The Ministry of Production could not do without these experiments for the determination of the *economically* most advantageous technical coefficients if it would realize the condition of the minimum cost of production which is *essential* for the attainment of the maximum collective welfare.
>
> It is on this account that the equations of the equilibrium with the maximum collective welfare are not soluble *a priori*, on paper.

An elementary algorithm

Barone's negative conclusion is certainly valid if the full production possibility set, including all of the possibilities for substitution in production, is not known to the central planner. In this event, numerical calculation is impossible, and Lange's suggestion, made some 20 years later, may be appropriate: the problem can be turned on its head and the market, itself, can be used as a mechanism of discovery as well as a giant analogue computer. But if the production possibility set can be explicitly constructed, substitution – in and of itself – does not seem to me to be a severe impediment to numerical computation.

At the present moment, some 20 years after the introduction and continued refinement of fixed-point computational techniques, I have in my possession a small floppy disk with a computer program which will routinely solve – on a personal computer – for equilibrium prices and activity levels in a Walrasian model in which the number of variables is on the order of 100. (The authors of the program suggest that examples with 300 variables can be accommodated on a mainframe computer.) Substantial possibilities of substitution, if known in advance, offer no difficulty to the successful functioning of this algorithm. In my opinion, the modern restatement of Barone's problem is rather that even 300 variables are extremely small in number in contrast to the millions of prices and activity levels implicit in his account. The computer, while expanding our capabilities immeasurably, has taught us a severe lesson about the role of mathematical reasoning in economic practice and forced us to shift our point of view dramatically from that held by our predecessors. We realize that our preoccupations are not with universal laws which describe economic phenomena with full and complete generality, but rather with intellectual formulations which are an imperfect representation of a complex and elusive reality. The application of general equilibrium theory to economic planning, and more generally to the evaluation of the consequences of changes in economic policy, must be based on highly aggregated models whose conclusions are at best tentative guides to action.

An exercise in comparative statics is begun by constructing a general equilibrium model whose solution reflects the economic situation existing prior to the proposed policy change. The number of parameters required to describe demand functions, initial endowments and the production possibility set is considerable, and in practice the constraint of reproducing the current equilibrium must be augumented by a variety of additional statistical estimates in order to specify the model. The limitations of data in the form required by the Walrasian model inevitably make this estimation procedure less than fully satisfactory.

The second step in the exercise is to calculate the solution after the proposed policy changes are explicitly introduced into the model. In some cases the policy variables being studied can be directly incorporated as parameters in the equations whose solution yields the equilibrium values; if the changes are small, their effects on the solution may be obtained by differentiating these equations and solving the resulting linear system for the corresponding changes in the equilibrium values themselves. This approach was adopted by Leif Johansen (1960) and by Arnold Harberger (1962) in his study of the incidence of a tax on corporate profits. The use of this method in policy analysis continues in Norway, and it forms the basis of the amibitious programme carried out by Peter Dixon and his collaborators in Australia (1982). If, on the other hand, the policy changes are large, the equilibrium position may be shifted substantially, and its determination may require the use of more sophisticated computational methods.

Fixed-point algorithms can be divided into two major classes: those based on the elements of differential topology, surveyed by Smale (1981), and those which are combinatorial in nature. The most elementary of the combinatorial algorithms for approximating a fixed point of a continuous mapping of the unit simplex $S = \{(x = (x_1, x_2, \ldots, x_n) | x_i \geq 0, \Sigma x_i = 1\}$ begins by dividing the simplex into a large number of small subsimplices as illustrated in Figure 4. In our notation the simplex is of dimension $n - 1$ and has faces of dimension $n - 2, \ldots, 1$. It is a requirement of the subdivision that the intersection of any two of the subsimplices is either empty or a full lower dimensional face of both of them.

Figure 4

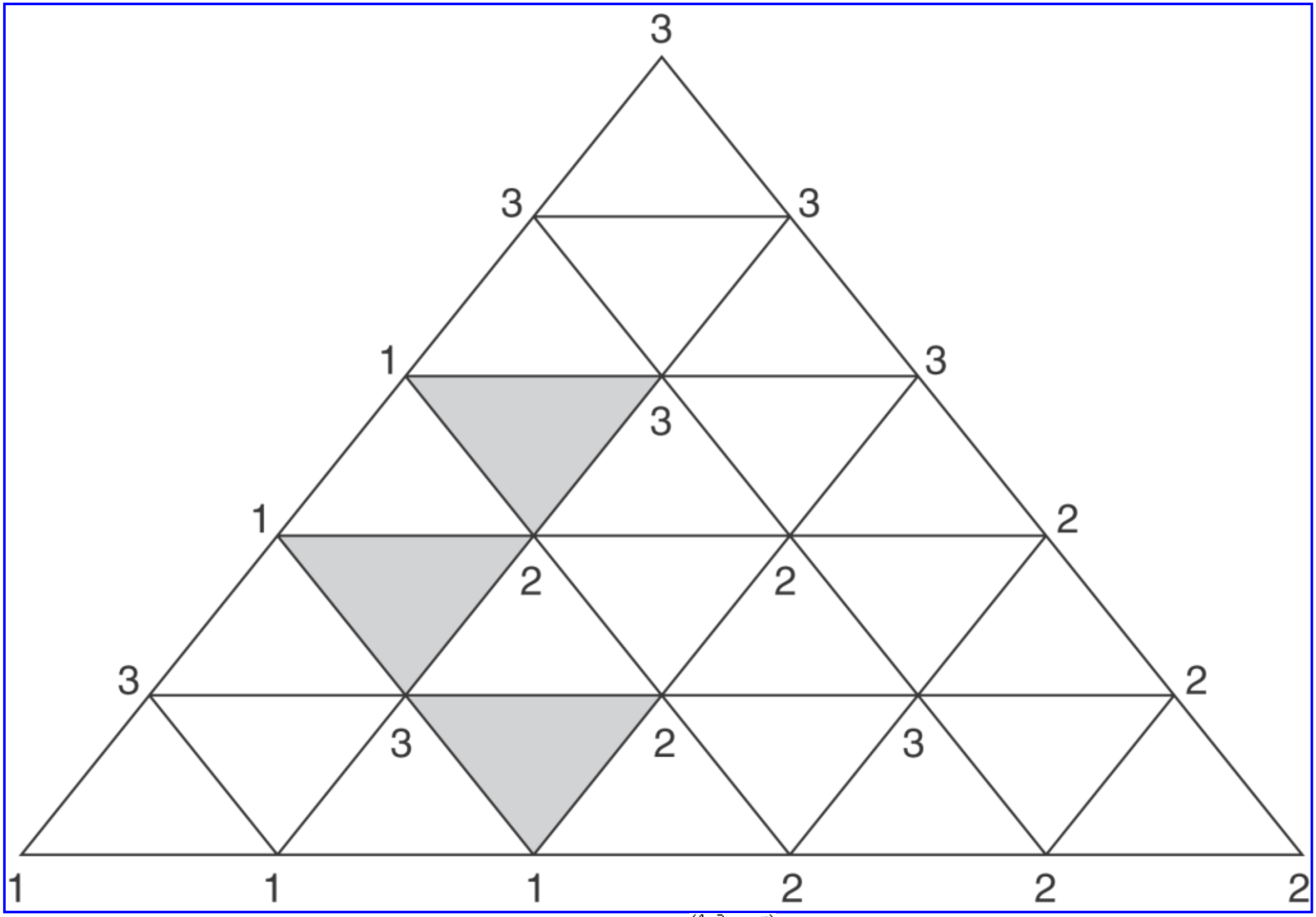

Each vertex of the subdivision will have associated with it an integer label selected from the set $(1, 2, \ldots, n)$. When the method is applied to the determination of a fixed point of a particular mapping, the labels associated with a vertex will depend on the mapping evaluated at that point. For the moment, however, the association will be arbitrary aside from the requirement that a vertex on the boundary of the simplex will have a label *i* only if the *i*th coordinate of that vertex is positive.

The remarkable combinatorial lemma demonstrated by Emanuel Sperner (1928) in his doctoral thesis is that at least one subsimplex must have all of its vertices differently labelled.

Assuming this result to be correct, let us consider a mapping of the simplex in which the image of the vector $x = (x_1, \ldots, x_n)$ is $f(x) = [f_1(x), \ldots, f_n(x)]$. The requirement that the image be on the simplex implies that $f_i(x) \geq 0$ and that $\Sigma f_i(x) = 1$. It follows that for every vertex of the subdivision *v*, unless *v* is a fixed point of the mapping, there will be a least one index *i* for which $f_i(v) - v_i < 0$. If we select such an index to be the label associated with the vertex *v*, then the assumptions of Sperner's Lemma are clearly satisfied, and the conclusion asserts the existence of a simplex whose vertices are distinctly labelled.

If the simplicial subdivision is very fine, the vertices of this sub-simplex are all close together; at each vertex a different coordinate is decreasing under the mapping, and by continuity every point in the small subsimplex will have the property that each coordinate is not increasing very much under the mapping. Since the sum of the coordinate changes is by definition zero, the image of any point in the completely labelled subsimplex will be close to itself, and such a point will therefore serve as an approximate fixed point of the mapping.

A formal proof of Brouwer's Theorem requires us to construct a sequence of finer and finer subdivisions, to find, for each subdivision, a completely labelled simplex, and to select a convergent sequence of these simplices tending to a fixed point of the mapping.

Sperner's Lemma may be applied to the equilibrium problem directly. For simplicity, consider the model of exchange in which the market excess demand functions are given by *g*(*p*), with *p* on the unit price simplex. As before, we subdivide the simplex and associate an integer label from the set (1,…,•*n*) with each vertex *v* of the subdivision, according to the following rule: the label *i* is to be selected from the set of those indices of which $g_i(p) \leq 0$. It is an elementary consequence of Walras's Law that a selection can be made which is consistent with the assumptions of Sperner's Lemma, and there will therefore be a subsimplex all of whose vertices bear distinct labels. By virtue of the particular labelling rule, any point in such a completely labelled simplex will be an approximate equilibrium price vector in the sense that all excess demands, at this price, will be either negative or, if positive, very small.

Sperner's original proof of his combinatorial lemma was not constructive; it was based on an inductive argument which required a complete enumeration of all completely labelled simplices for a series of lower dimensional problems. In order to develop an effective numerical algorithm for the determination of such a simplex let us begin by embedding the unit simplex, and its subsimplices, in a larger simplex *T*, as in Figure 5. The larger simplex is subdivided by joining its *n* new vertices to those vertices of the original subdivision lying on the boundary of the unit simplex. The assumptions of Sperner's Lemma permit the new vertices to be given distinct labels from the set (1,…,•*n*), in such a way that no additional completely labelled simplices are generated. For concreteness, let the new vertex receiving the label *i* be denoted by v^i.

Figure 5

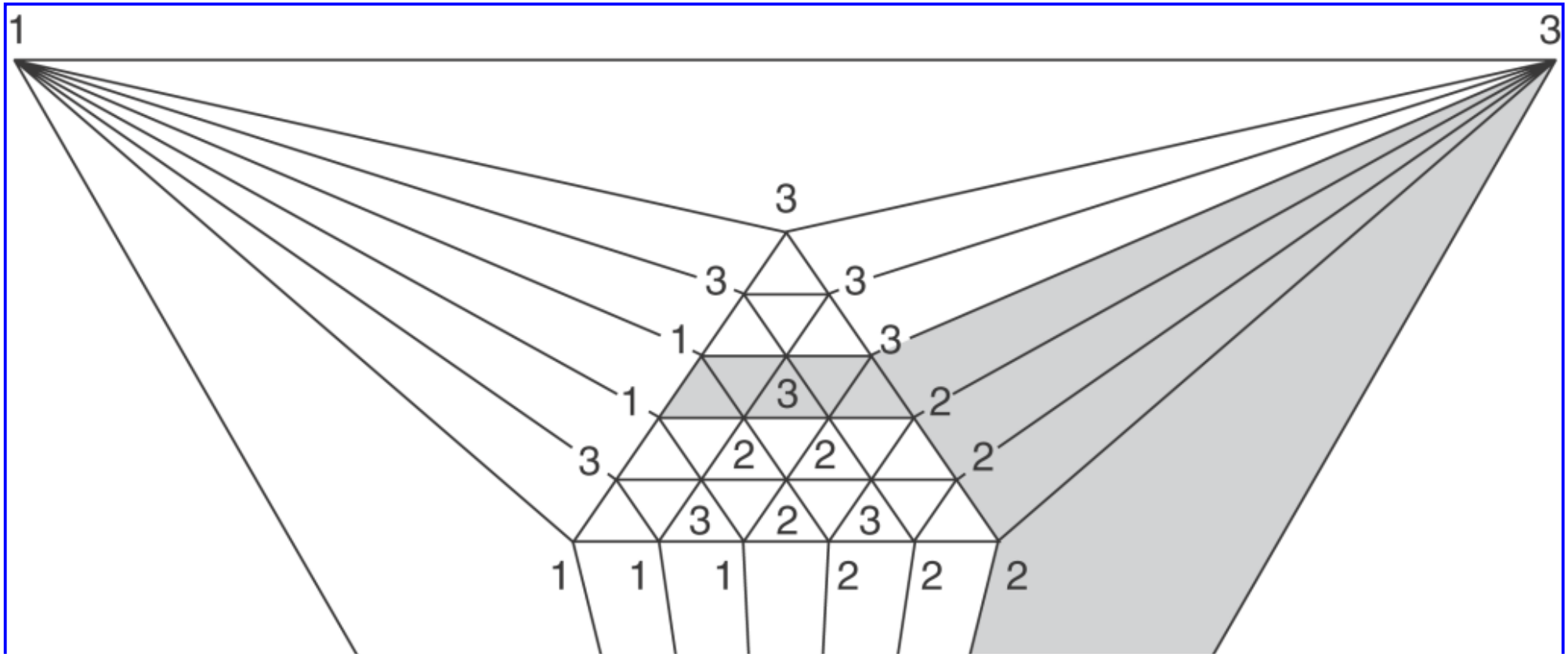

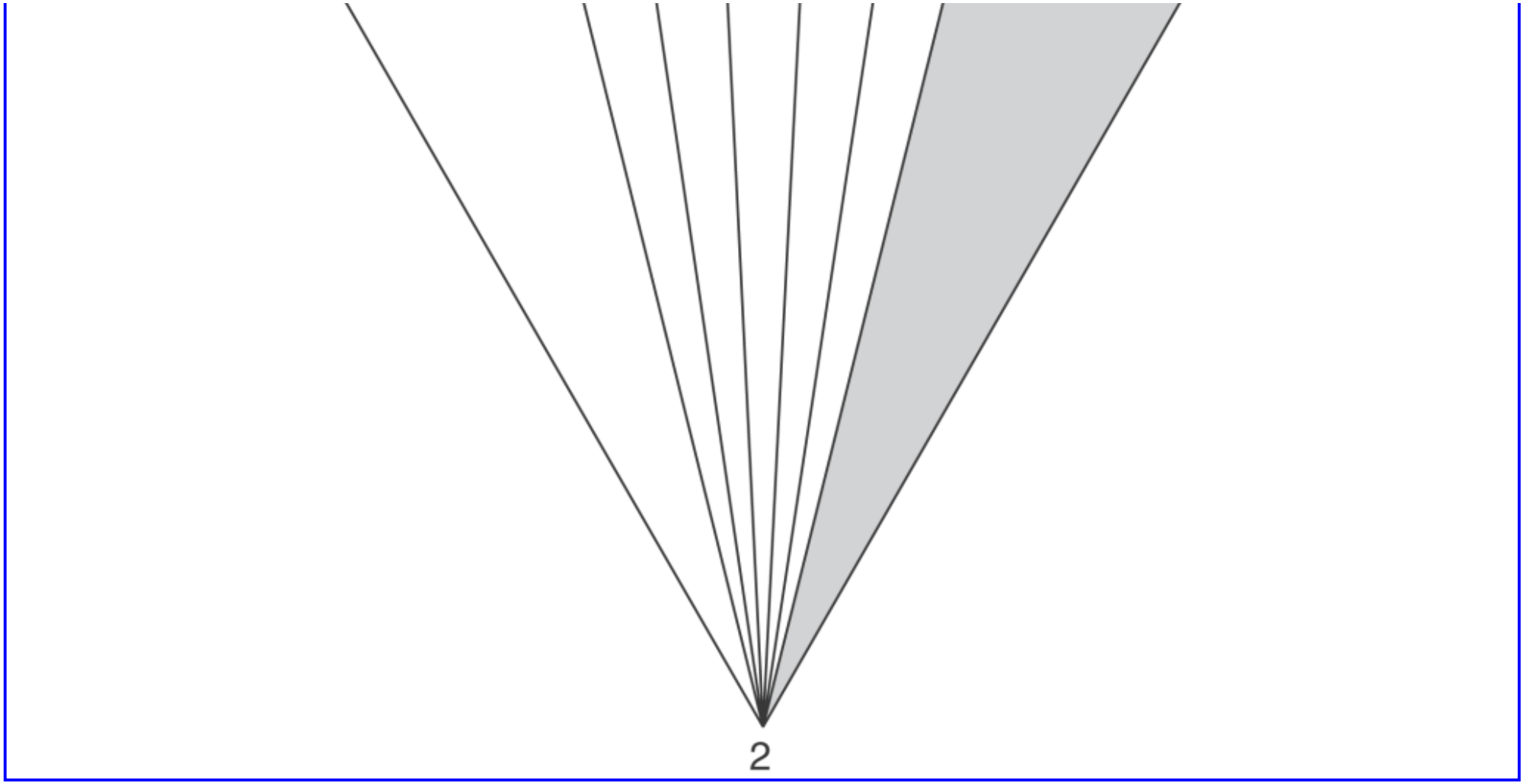

We begin our search for a completely labelled simplex by considering the simplex with vertices $v^2,\ldots,\bullet v^n$ and one additional vertex, say v^*. If v^* has the label 1, this simplex is completely labelled and our search terminates; otherwise we move to an adjacent simplex by removing the vertex whose label agrees with that of v^* and replacing it with that unique other vertex yielding a simplex in the subdivision. As the process continues, we are, at each step, at a simplex whose vertices bear the labels $2,\ldots,\bullet n$, with a single one of these labels appearing on a pair of vertices. Precisely two $n-2$ dimensional faces have a complete set of labels $2,\ldots,\bullet n$. The simplex has been entered through one of these faces; the algorithm proceeds by exiting through the other such face.

The argument first introduced by Lemke (1965) in his study of two person non-zero sum games was carried over by Scarf (1967) to show that the above algorithm never returns to a simplex previously visited and never requires a move outside of T. Since the number of simplices is finite, the algorithm must terminate, and termination can only occur when a completely labelled simplex is reached.

Improvements in the algorithm

The algorithm can easily be programmed for a computer, and it provides the most elementary numerical procedure for approximating fixed points of a continuous mapping and equilibrium prices for the Walrasian model. Since its introduction in 1967, the algorithm, in this particular form, has been applied to a great number of examples of moderate size, and it performs sufficiently well in practice to conclude that the numerical determination of equilibrium prices is a feasible undertaking. The algorithm does, however, have some obvious drawbacks which must be overcome to make it available for problems of significant size. For example, the information which yields the labelling of the vertices, and therefore the

path taken by the algorithm, is simply the index of a coordinate which happens to be decreasing when the mapping is evaluated at the vertex. More recent algorithms make use of the full set of coordinates of the image of the vertex instead of a single summary statistic.

Second, this primitive algorithm is always initiated at the boundary of the simplex. If the approximation is not sufficiently good, the grid size must be refined, and a recalculation, which makes no use of previous information, must be performed. It is of the greatest importance to be able to initiate the algorithm at an arbitrary interior point of the simplex selected as our best a priori estimate of the answer.

The following geometrical setting (Eaves and Scarf, 1976) for the elementary algorithm suggests the form these improvements can take. Let us construct a piecewise linear mapping, $h(x)$, of T into itself as follows: for each vertex v in the subdivision let $h(v) = v^i$, where i is the label associated with v. We then complete the mapping by requiring h to be linear in each simplex of the subdivision. The mapping is clearly continuous on T and maps every boundary point of T into itself. Moreover, every subsimplex in the subdivision whose vertices are not completely labelled is mapped, by h, into the boundary of T. If none of the simplices were completely labelled, this construction would yield a most improbable conclusion: a continuous mapping of T into itself which is the identity on the boundary and which maps the entire simplex into the boundary. That such a mapping cannot exist is known as the Non-Retraction Theorem, an assertion which is, in fact, equivalent to Brouwer's Theorem. The impossibility of such a mapping reinforces our conclusion that a completely labelled simplex does exist.

Select a point c interior to one of the boundary faces of T and consider the set of points which map into c; that is, the set of x for which $h(x) = c$. As Figure 6 indicates, this set contains a piecewise linear path beginning at the point c, and transversing precisely those simplices encountered in our elementary algorithm. There are however, other parts of the set $\{x | h(x) = c\}$: closed loops which do not touch the boundary of T and other piecewise linear paths connecting a pair of completely labelled simplices. Stated somewhat informally, the general conclusion, of which this is an example, is that the inverse image of a particular point, under a piecewise linear mapping from an n dimensional set to an $n - 1$ dimensional set, consists of a finite union of interior loops, and paths which join two boundary points (see Milnor, 1965, for the differentiable version).

Figure 6

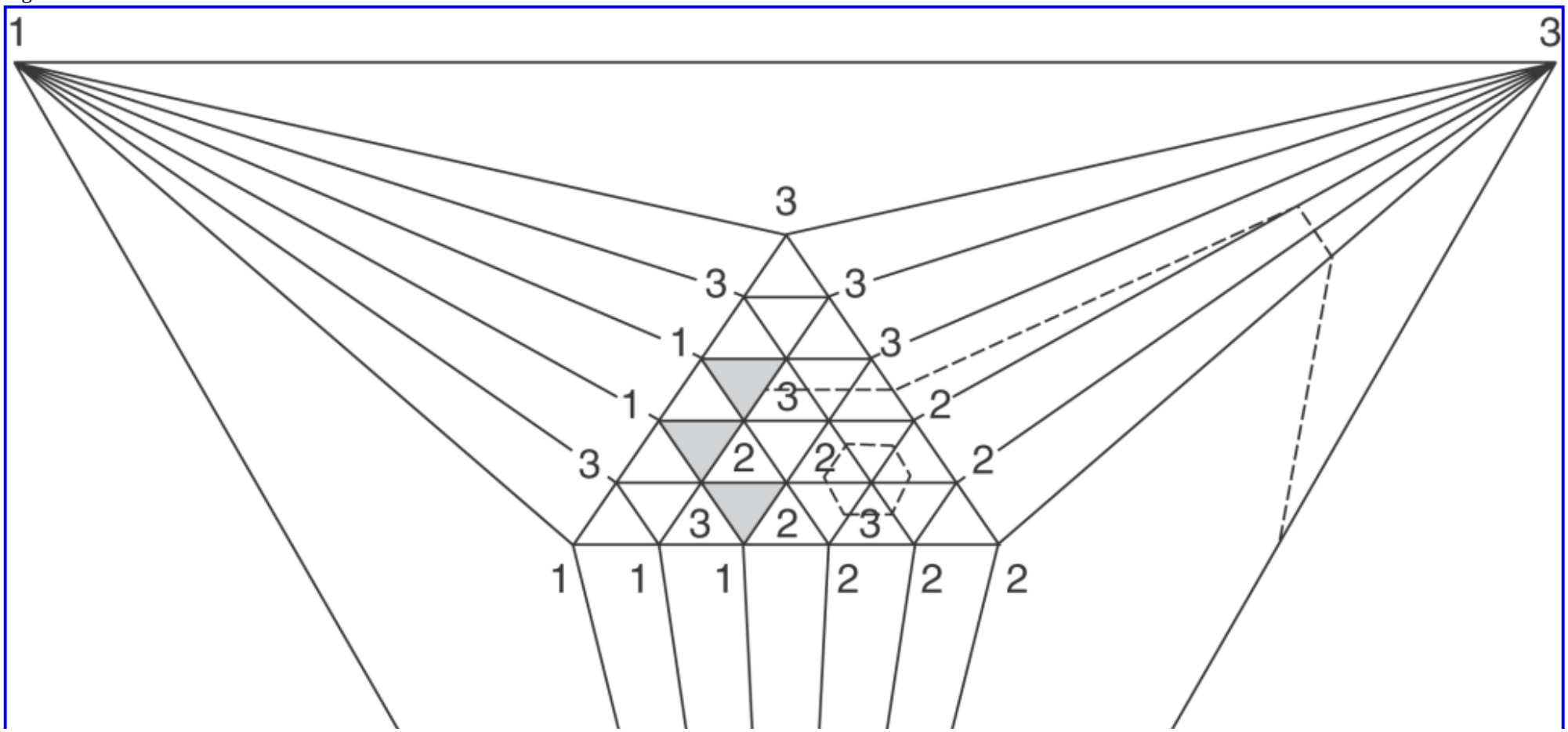

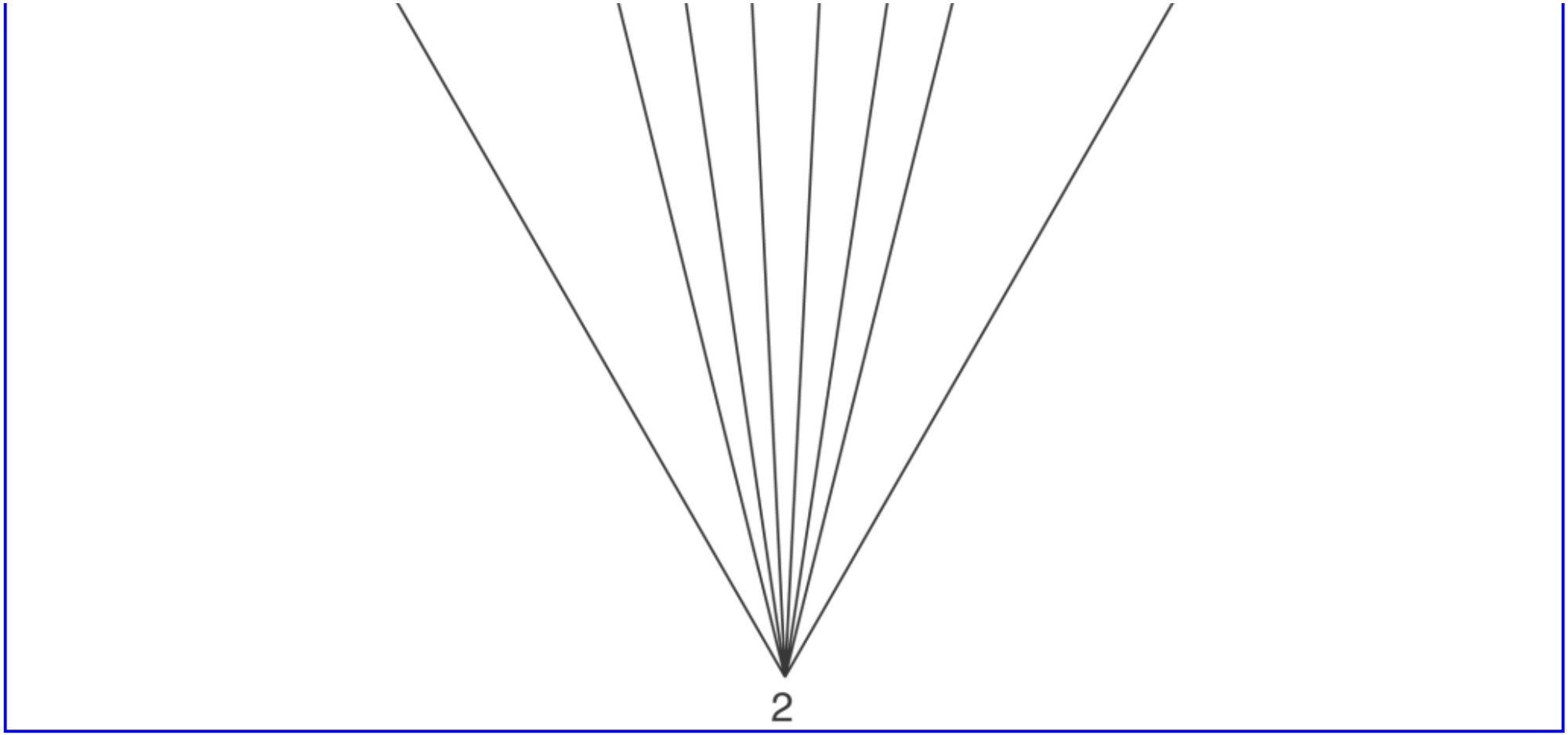

To see how this observation can be used, consider the product of the unit simplex S and the closed unit interval [0, 1]; that is, the set of points (x, t) with x in S and $0 \leq t \leq 1$, as in Figure 7. Extend the mapping from the unit simplex to this large set by defining $F(x, t) = (1 - t) f(x) + tx^*$, with x^* a preselected point on the simplex, taken to be an estimate of the true fixed point. The set of points for which $F(x, t) - x = 0$ is, by our general conclusion, a finite union of paths and loops. Precisely one of these paths intersects the upper boundary of the enlarged set. If the path is followed, its other endpoint must lie in the face $t = 0$ and yield a fixed point of the original mapping.

Figure 7

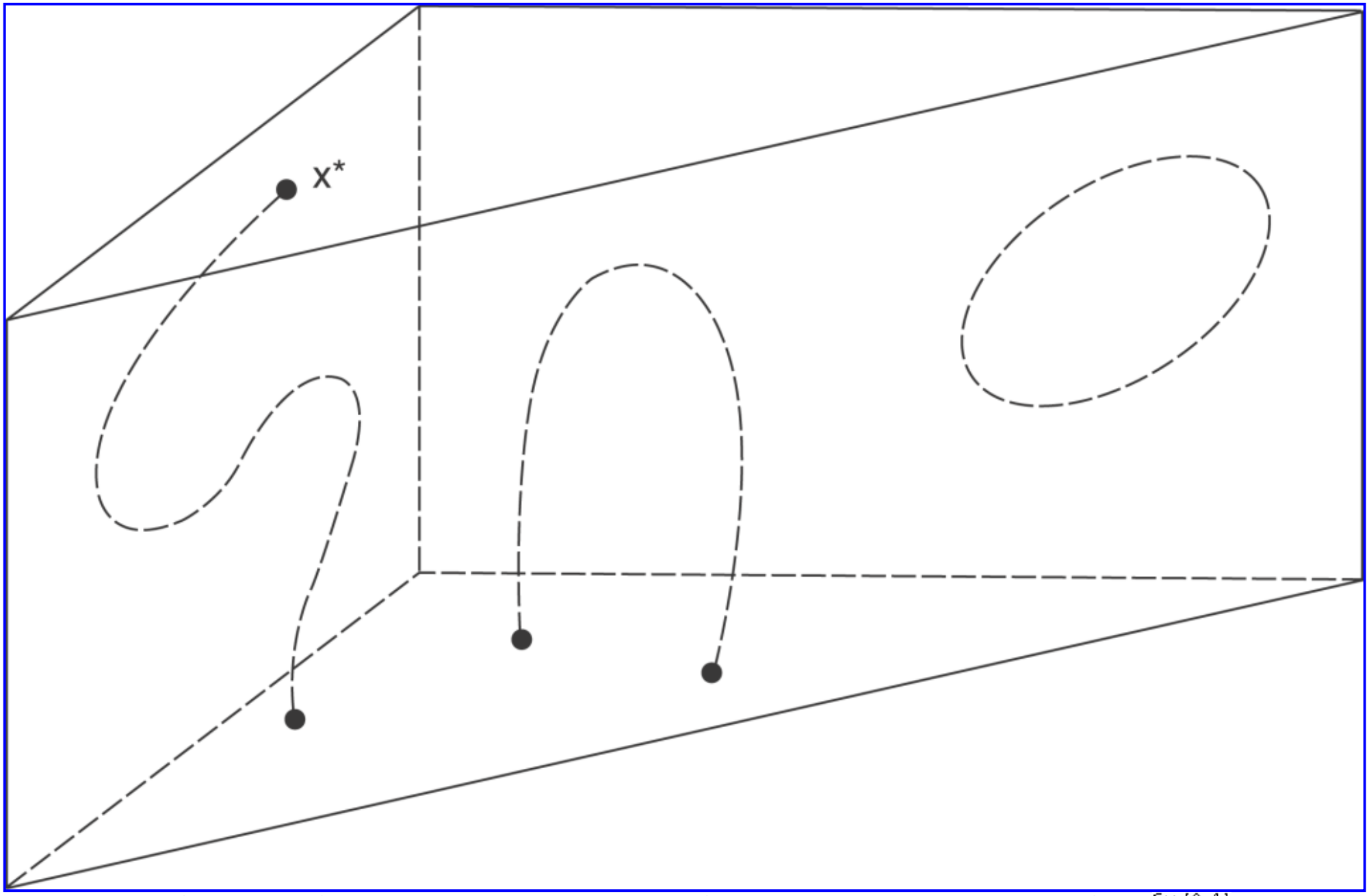

The path leading to the fixed point can be followed on the computer in several ways. We can, for example, introduce a simplicial decomposition of the set $S \times [0, 1]$ and approximate F by a piecewise linear mapping agreeing with F on the vertices of the subdivision. Following the path then involves the same type of calculation we have become accustomed to in carrying out linear programming pivot steps. There are a great many variations in the mode of simplicial subdivision leading to substantial improvements in the efficiency of our original fixed-point algorithm (Eaves, 1972; Merrill, 1971; van der Laan and Talman, 1979).

An alternative procedure, adopted by Kellogg, Li and Yorke (1976) and Smale (1976), is to impose sufficient regularity conditions on the underlying mapping so that differentiation of $F(x, t) - x = 0$ yields a set of differential equations for the path joining x^* to the fixed point on t=0. This leads to a variant of Newton's method which is global in the sense that it need not be initiated in the vicinity of the correct answer. But, whichever of these alternatives we select, the numerical difficulties in computing equilibrium prices can be overcome

for all problems of reasonable size.

Applied general equilibrium analysis

During the last 15 years, the field of Applied General Equilibrium Analysis has grown considerably; instead of the few tentative examples illustrating our ability to solve general equilibrium problems, we have seen the construction of a large number of models of substantial size designed to illuminate specific policy issues. The number of books and papers which have appeared in the field is far too large for a complete enumeration in this essay, and I shall mention only a few publications which may be consulted to obtain an indication of the diversity of this activity. The paper by Shoven and Whalley (1984) in the *Journal of Economic Literature* is a survey of applied general equilibrium models in the fields of taxation and international trade constructed by these authors and their colleagues. The volume by Adelman and Robinson (1978) is concerned with the application of general equilibrium analysis to problems of economic development. Whalley (1985) has written on trade liberalization, and Ballard, Fullerton, Shoven and Whalley (1985) on the evaluation of tax policy. Jorgenson (Hudson and Jorgenson, 1974) and Manne (1976) have made extensive applications of this methodology to energy policy, and Ginsburg and Waelbroeck (1981) provide a refreshing discussion of alternative computational procedures applied to a model of international trade involving over 200 commodities. The volume edited by Scarf and Shoven (1985) contains a collection of papers presented at one of an annual series of workshops in which both applied and theoretical topics of interest to researchers in the field of Applied General Equilibrium Analysis are discussed.

See Also

- general equilibrium
- computation of general equilibria (new developments)

Bibliography

Adelman, I. and Robinson, S. 1978. *Income Distribution Policy in Developing Countries: A Case Study of Korea*. Stanford: Stanford University Press.

Ballard, C.L., Fullerton, D., Shoven, J.B. and Whalley, J. 1985. *A General Equilibrium Model for Tax Policy Evaluation*. Chicago: University of Chicago Press.

Barone, E. 1908. Il Ministerio della Produzione nello stato colletivista. *Giornale degli Economisti e Revista di Statistica*. Trans. as ‘The Ministry of Production in the Collectivist State’, in *Collectivist Economic Planning*, ed. F.A. Hayek, London: G. Routledge & Sons, 1935.

Brouwer, L.E.J. 1912. Über Abbildungen von Mannigfaltigkeiten. *Mathematische Annalen* 71, 97–115.

Debreu, G. 1982. Existence of competitive equilibrium. In *Handbook of Mathematical Economics*, ed. K.J. Arrow and M. Intriligator, Amsterdam: North-Holland.

Dixon, P.B., Parmenter, B.R., Sutton, J. and Vincent, D.P. 1982. *ORANI: A Multisectoral Model of the Australian Economy*. Amsterdam: North-Holland.

Eaves, B.C. 1972. Homotopies for the computation of fixed points. *Mathematical Programming* 3, 1–22.

Eaves, B.C. and Scarf, H. 1976. The solution of systems of piecewise linear equations. *Mathematics of Operations Research* 1, 1–27.

Ginsburg, V.A. and Waelbroeck, J.L. 1981. *Activity Analysis and General Equilibrium Modelling*. Amsterdam: North-Holland.

Harberger, A. 1962. The incidence of the corporation income tax. *Journal of Political Economics* 70, 215–40.

Hudson, E.A. and Jorgenson, D.W. 1974. US Energy policy and economic growth. *Bell Journal of Economics and Management Science* 5, 461–514.

Johansen, L. 1960. *A Multi-Sectoral Study of Economic Growth*. Amsterdam: North-Holland.

Kellogg, R.B., Li, T.Y. and Yorke, J. 1976. A constructive proof of the Brouwer Fixed Point Theorem and computational results. *SIAM Journal of Numerical Analysis* 13, 473–83.

Kuhn, H.W. 1968. Simplicial approximation of fixed points. *Proceedings of the National Academy of Sciences* 61, 1238–42.

Lange, O. 1936. On the economic theory of socialism. *Review of Economic Studies* 4, 53–71, 123–42.

Lemke, C.E. 1965. Bimatrix equilibrium points and mathematical programming. *Management Science* 11, 681–9.

Manne, A.S. 1976. ETA: a model of energy technology assessment. *Bell Journal of Economics and Management Science* 7, 379–406.

Merrill, O.H. 1971. Applications and extensions of an algorithm that computers fixed points of certain non-empty convex upper semicontinuous point to set mappings. Technical Report 71–7, University of Michigan.

Milnor, J. 1965. *Topology from the Differentiable Viewpoint*. Charlottesville: University of Virginia Press.

Scarf, H.E. 1967. The approximation of fixed points of a continuous mapping. *SIAM Journal of Applied Mathematics* 15, 1328–43.

Scarf, H.E., with the collaboration of T. Hansen, 1973. *The Computation of Economic Equilibria*. London, New Haven: Yale University Press.

Scarf, H. and Shoven, J.B., eds. 1984. *Applied General Equilibrium Analysis*. Cambridge: Cambridge University Press.

Shoven, J.B. and Whalley, J. 1972. A general equilibrium calculation of the effects of differential taxation of income from capital in the U.S. *Journal of Public Economy* 1, 281–321.

Shoven, J.B. and Whalley, J. 1984. Applied general-equilibrium models of taxation and international trade. *Journal of Economic Literature* 22, 1007–51.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3, 107–20.

Smale, S. 1981. Global analysis and economics. In *Handbook of Mathematical Economics*, Vol. I, ed. K.J. Arrow and M. Intriligator, Amsterdam: North-Holland.

Sperner, E. 1928. Neur Beweis für die Invarianz der Dimensionszahl und des Gebietes. *Abhandlungen an den mathematischen Seminar der Universität Hamburg* 6, 265–72.

van der Laan, G. and Talman, A.J.J. 1979. A restart algorithm for computing fixed points without an extra dimension. *Mathematical Programming* 17, 74–84.

Whalley, J. 1985. *Trade Liberalization among Major World Trading Areas*. Cambridge, Mass.: MIT Press.

How to cite this article

Scarf, Herbert E. "computation of general equilibria." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000573> doi:10.1057/9780230226203.0283

The New Palgrave Dictionary of Economics Online

computational methods in econometrics

Vassilis A. Hajivassiliou

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The computational properties of an econometric method are fundamental determinants of its importance and practical usefulness, in conjunction with the method's statistical properties. Computational methods in econometrics are advanced through successfully combining ideas and methods in econometric theory, computer science, numerical analysis, and applied mathematics. The leading classes of computational methods particularly useful for econometrics are matrix computation, numerical optimization, sorting, numerical approximation and integration, and computer simulation. A computational approach that holds considerable promise for econometrics is parallel computation, either on a single computer with multiple processors, or on separate computers networked in an intranet or over the internet.

Keywords

Bayesian inference; bootstrap; classical inference; computational methods; generalized least squares; generalized method of moments; importance sampling simulation; jackknife; least absolute deviations; maximum likelihood; numerical integration; optimal control; ordinary least squares; random effects models; simulation-based estimation; Stone, J. R. N.; Markov chain Monte Carlo methods; parallel computation

Article

1 Introduction

In evaluating the importance and usefulness of particular econometric methods, it is customary to focus on the set of *statistical* properties that a method possesses – for example, unbiasedness, consistency, efficiency, asymptotic normality, and so on. It is crucial to stress, however, that meaningful comparisons cannot be completed without paying attention also to a method's *computational* properties. Indeed the practical value of an econometric method can be assessed only by examining the inevitable interplay

between the two classes of properties, since a method with excellent statistical properties may be computationally infeasible and vice versa. Computational methods in econometrics are evolving over time to reflect the current technological boundaries as defined by available computer hardware and software capabilities at a particular period, and hence are inextricably linked with determining what the state of the art is in econometric methodology.

To give a brief illustration, roughly from the late 1950s until the early 1960s we had the ‘Stone Age’ of econometrics, when the most sophisticated computational instrument was the slide rule, which used two rulers on a logarithmic scale, one sliding into the other, to execute approximate multiplication and division. In this Stone Age, suitably named in honour of Sir Richard Stone, winner of the 1984 Nobel Prize in Economics, the brightest Ph.D. students at the University of Cambridge were toiling for days and days in back rooms using slide rules to calculate ordinary linear regressions, a task which nowadays can be achieved in a split second on modern personal computers.

The classic linear regression problem serves to illustrate the crucial interaction between statistical and computational considerations in comparing competing econometric methods. Given data of size *S*, with observations on a dependent variable denoted by *S*×1 vector *y* and corresponding observations on *k* explanatory factors denoted by *S*×*k* matrix *X* (*k*<*X*), the linear plane fitting exercise is defined by Gauss's minimum quadratic distance problem:

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} (y - Xb)'(y - Xb) \equiv \underset{b}{\arg\min} \sum_{s=1}^{S} (y_s - x_s' b)^2 \tag{1}$$

where x_s' is the *s*th row of matrix *X* and *b* is a *k*×1 vector of real numbers defining the regression plane *Xb*. Under the assumption that *X* has full column rank *k*, the solution to this *ordinary least squares* minimization problem is the linear-in-*y* expression $\hat{\beta} = (X'X)^{-1}X'y$, which only requires the matrix operations of multiplication and inversion.

Suppose, however, that Gauss had chosen instead as his measure of distance the sum of absolute value of the deviations, and defined instead:

$$\tilde{\beta} = \underset{b}{\operatorname{argmin}} \sum_{s=1}^{S} \left| y_s - x_s' b \right|. \tag{2}$$

The vector $\tilde{\beta}$ that solves the second minimization is known as the *least absolute deviations* (LAD) estimator and has no closed-form matrix expression. In fact, calculation of $\tilde{\beta}$ requires highly nonlinear operations for which computationally efficient algorithms were developed only in the 1970s. To give a

concrete example, consider the *intercept-only* linear regression model where X is the $S\times 1$ vector of ones. Then the single $\hat{\beta}$ coefficient that solves (1) is the sample mean of y, while $\ddot{\beta}$ that solves (2) is the sample *median* of y. The latter is orders of magnitude more difficult to compute than the former since it involves sorting y and finding the value in the middle, while the former simply adds all elements of y and divides by the sample size. Clearly, it could be quite misleading if $\hat{\beta}$ and $\ddot{\beta}$ where compared solely in terms of statistical properties without any consideration of their substantially different computational requirements.

A second example in a similar vein is the following parametric estimation problem. Suppose a sample of size S is observed on a single variable y. It is believed that each observation y_s is drawn independently from the same uniform distribution on the interval $[\theta, c]$ where the lower value of the support is the single unknown parameter that needs to be estimated, while c is known. Two parametric estimation methods with particularly attractive statistical properties are the generalized method of moments (GMM) and the method of maximum likelihood (MLE). Indeed, for relatively large sample sizes these two methods are comparably attractive in terms of statistical properties, while they differ *drastically* in terms of computational requirements: the GMM solution is $\hat{\theta}_{gmm} = \frac{2}{S}\sum_{s=1}^{S} y_s - c$, thus requiring only the simple calculation of the sample mean $\bar{y}$, while the MLE involves the highly nonlinear operation of finding the minimum of the data vector y, $\hat{\theta}_{mle} = min(y_1, \ldots, y_S)$.

In the following section we discuss in turn the leading classes of methods that are of particular importance in modern econometrics, while Section 3 introduces the concept of parallel processing and describes its current value and future promise in aiding dramatically econometric computation.

2 Computational methods important for econometrics

The advancement of computational methods for econometrics relies on understanding the interplay between the disciplines of econometric theory, computer science, numerical analysis, and applied mathematics. In the five subsections below we discuss the leading classes of computational methods that have proven of great value to modern econometrics.

2.1 Matrix computation and specialized languages

To start with the fundamental econometric framework of linear regression, the *sine qua non* of econometric computation is the ability to program and perform efficiently matrix operations. To this end, specialized matrix computer languages have been developed which include Gauss and Matlab. Fundamental estimators of the linear regression coefficient vector β, like the OLS $(X'X)^{-1}X'y$ and its generalized least squares (GLS) variant $(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, are leading examples of the usefulness of such matrix languages, where the $S\times S$ matrix Ω is a positive definite, symmetric variance-covariance matrix of the disturbance vector $\varepsilon \equiv y - X\beta$. Matrix operations are useful even for nonlinear econometric methods discussed below, since a generally useful approach is to apply linearization approximations through the use of differentiation and Taylor's expansions.

In implementing econometric methods that involve matrix operations, special attention needs to be paid

to the dimensionality of the various matrices, as well as to any special properties a matrix may posses, which can affect very substantially the feasibility and performance of the computational method to be adopted. Looking at the OLS and GLS formulae, we see three different matrices that require inversion: $X'X$, Ω, and $X'\Omega^{-1}X$. The first and the third are of dimension $k \times k$, while the second is $S \times S$. Since the number of regressors k is typically considerably smaller than the sample size S, the inversion of these matrices can involve vastly different burden in terms of total number of computer operations required as well as memory locations necessary for holding the information during those calculations. (For example, in panel data settings where multiple observations are observed in different time-periods for a cross-section of economic agents, it is not uncommon to have total sample sizes of 300,000 or more.) To this end, econometric analysts have focused on importing from numerical analysis matrix algorithms that are particularly efficient in handling sparse as opposed to dense matrices. By their very nature, sparse matrices exhibit a very high degree of compressibility and concomitantly lower memory requirements. See Drud (1977) for the use of sparse matrix techniques in econometrics. A matrix is called sparse if it is primarily populated by zeros, for example, the variance-covariance matrix of a disturbance vector following the moving-average-of-order-1 model:

$$\Omega_{ma1} = \sigma^2 \begin{pmatrix} 1 & \frac{\lambda}{1+\lambda^2} & 0 & \cdots & 0 \\ \frac{\lambda}{1+\lambda^2} & 1 & \frac{\lambda}{1+\lambda^2} & \ddots & \vdots \\ 0 & \frac{\lambda}{1+\lambda^2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & \frac{\lambda}{1+\lambda^2} \\ 0 & \cdots & 0 & \frac{\lambda}{1+\lambda^2} & 1 \end{pmatrix}.$$

In contrast, a stationary autoregressive disturbance of order 1 has a dense variance-covariance matrix:

$$\Omega_{ar1} = \sigma^2 \begin{pmatrix} 1 & \gamma & \gamma^2 & \cdots & \gamma^{S-1} \\ \gamma & 1 & \gamma & \ddots & \vdots \\ \gamma^2 & \gamma & \ddots & \ddots & \gamma^2 \\ \vdots & \ddots & \ddots & 1 & \gamma \\ \gamma^{S-1} & \cdots & \gamma^2 & \gamma & 1 \end{pmatrix}.$$

Other matrix algebra methods especially important in econometrics are the Cholesky factorization (see Golub, 1969) of a positive definite matrix A into the product $A=R'R$ where R is an upper-triangular matrix, and the singular value decomposition that allows the calculation of pseudo-inverse of any matrix B which may be non-square, and if square, not positive definite (see Belsley, 1974).

It is important to note that on occasion a brilliant theoretical development can simplify enormously the computational burden of econometric methods that, though possessing attractive statistical properties, were thought to be infeasible with existing computation technology in the absence of the theoretical development. A case in point is the GLS/MLE estimator for the one-factor random effects model proposed by Balestra and Nerlove (1966), which is of great importance in the analysis of linear panel data models. The standard formulation gives rise to the GLS formula requiring the inversion of an equi-correlated variance covariance matrix Ω of dimension $S\times S$, where S is of the order of the product of the number of available observations in the cross-section dimension times the number available in the time dimension. For modern panel data-sets, this can exceed 300,000, thus making the calculation of Ω^{-1} infeasible even on today's super-computers, let alone with the slide rules available in 1966. Fuller and Battese (1973), however, showed that the equi-correlated nature of the one-factor random effects model made calculation of the GLS estimator equivalent to an OLS problem, where the dependent variable $\tilde{y}$ and the regressors $\tilde{X}$ are simple linear combinations of the original data $y_{it}, x_{1it},\ldots, x_{kit}$ and its time averages $\bar{y}_{i.}, \bar{x}_{1i.}, \ldots, \bar{x}_{ki.}$ defined by $\bar{y}_{i.} \equiv \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\tilde{y}_{it} \equiv y_{it} - \lambda\bar{y}_{i.}$, and analogously for the regressor variables. This realization allowed the calculation of the GLS estimator without the need for inverting the usually problematically large Ω matrix.

Another important case where a theoretical development in methodology led to a dramatic lowering of the computational burden and hence allowed the calculation of models that would otherwise have had to wait perhaps for decades for sufficient advancements in computer technology is the simulation-based inference for Limited Dependent Variable models, associated with the name of Daniel McFadden (1989). See Section 2.5 below, McFadden, Daniel and simulation-based estimation.

2.2 Optimization

Many econometric estimators with attractive statistical properties require the optimization of a (generally) nonlinear function of the form:

$$q \equiv \underset{\theta}{\operatorname{argmax}} F(\theta; data) \tag{3}$$

over a vector of unknown parameters θ of dimension p, typically considerably larger than 1. Examples are: the method of maximum likelihood, minimum-distance (OLS, LAD, GMM), and other extremum estimators. (The need to optimize functions numerically is also important for certain problems in

computational economics, for example, the problem of optimal control.) Algorithms for optimizing functions of many variables are a key component in the collection of tools for econometric computation. The suitability of a certain algorithm to a specific optimization econometric problem depends on the following classification:

1. 1. *Algorithms that require the calculation of first and possibly second derivatives* Versus *algorithms that do not*. Clearly, if the function to be optimized is not twice continuously differentiable (as is the case with LAD) or even discontinuous (as is the case with the maximum score estimator for the semiparametric analysis of the binary response model – see Manski, 1975), algorithms that require differentiability will not be suitable. The leading example of an algorithm not relying on derivatives is the nonlinear simplex method of Nelder and Meade (1965).
2. 2. *Local* Versus *global algorithms*. Optimization algorithms of the first type (for example, Gauss-Newton, Newton-Raphson, and Berndt et al. (1974)) search for an optimum in the vicinity of the starting values fed into the algorithm. This strategy may not necessarily lead to a global optimum over the full set of parameter space. This is of particular importance if the function to be optimized has multiple local optima, where typically the estimator with the desirable statistical properties corresponds to locating the overall optimum of the function. In such cases, global optimization algorithms (for example, simulated annealing and genetic optimization algorithm) should be employed instead.

Special methods are necessary for constrained optimization, where a function must be maximized or minimized subject to a set of equality or inequality constraints. These problems, in general considerably more demanding than unconstrained optimization, can be handled through three main alternative approaches: interior, exterior and re-parameterization methods.
Comprehensive reviews of optimization methods in econometrics can be found in Goldfeld and Quandt (1972), Quandt (1983), and Dennis and Schnabel (1984). These studies also discuss the related issue of the numerical approximation of derivatives and illustrate the fundamental link in terms of computation between optimization and the problem of solving linear and nonlinear equations. For similar methods used in economics, see numerical optimization methods in economics and nonlinear programming.

2.3 Sorting

Of special importance for computing the class of estimators known as robust or semiparametric methods is the ability to sort data rapidly and computationally efficiently. Such a need arises in the calculation of order statistics, for example, the sample median and sample minimum required by the first two estimation examples given above. The leading sorting algorithms, bubble-, heap- and quick-sort, have fundamentally different properties in terms of computation speed and memory requirements, in general depending on how close to being sorted the original data series happens to be. For a practical review of the leading sorting algorithms, see Press et al. (2001, ch. 8).

2.4 Numerical approximation and integration

Numerical approximation is necessary for any mathematical function that does not have a closed form solution, for example, exponential, natural logarithm and error functions. See Abramowitz and Stegun (1964) for an exhaustive study of mathematical functions and their efficient approximation. Judd (1996) focuses on numerical approximation methods particularly useful in economics and econometrics.
Numerical integration, also known as numerical quadrature, is a related approximation problem that is crucial to modern econometrics. There are two key fields of econometrics where integrals without a closed form must be evaluated numerically. The first is Bayesian inference where moments of posterior densities need to be evaluated, which take the form of high-dimensional integrals. See, inter alia, Zellner, Bauwens and VanDijk (1988). The second main class is classical inference in limited dependent variable (LDV) models; for example, Hajivassiliou and Ruud (1994). See Geweke (1996) for an exhaustive review of numerical integration methods in computational economics and econometrics, and Davis and Rabinowitz (1984) for earlier results.
It is important to highlight a crucial difference between the numerical integration problems in Bayesian inference and those in classical inference for LDV models, which makes various integration-by-simulation algorithms be useful to one field and not the other: in the Bayesian case, typically a single or a few high-dimensional integrals have to be evaluated accurately. In contrast, in the classical LDV inference case, quite frequently hundreds of thousands of such integrals need to be approximated.

2.5 Computer simulation

The need for efficient generation of pseudo-random numbers with good statistical properties on a computer appears very routinely in econometrics. Leading examples include:

- Statistical methods based on resampling, primarily the ‘jackknife’ and the ‘bootstrap’, as introduced by Efron (1982). These methods have proven of special value in improving the small sample properties of certain econometric estimators and test procedures, for example in reducing estimation bias. They are also used to approximate the small sample variance of estimators for which no closed form expressions can be derived.
- Evaluation of econometric estimators through Monte Carlo experiments, where hypothetical data-sets with certain characteristics are simulated repeatedly and the econometric estimators under study are calculated for each set. This allows the calculation of empirical (simulated) properties of the estimators, either to compare to theoretical mathematical calculations or because the latter are intractable.
- Calculation of frequency probabilities of possible outcomes in large-scale decision trees, for which the outcome probabilities are impossible to characterize theoretically.
- Sensitivity analyses and what-if studies, where an econometric model is ‘run’ on a computer under different scenarios of policy measures.
- Simulation-based Bayesian and classical inference, where integrals are approximated through computer simulation (known as Monte Carlo integration). Particularly important methods in this context are the following: frequency simulation; importance sampling; and Markov chain Monte Carlo methods (the leading exponents being Gibbs resampling and the Metropolis/Hastings

algorithm). A related class of methods, known as variance-reduction simulation techniques, includes control variates and antithetics. See Geweke (1988) and Hajivassiliou, McFadden and Ruud (1996) for reviews. See also simulation-based estimation.

3 Parallel computation

Parallel processing, where a computation task is broken up and distributed across different computers, is a technique that can afford huge savings in terms of total time required for solving particularly difficult econometric problems. For example, the simulation-based estimators mentioned in the previous section exhibit the potential of significant computational benefits by calculating them on computers with massively parallel architectures, because the necessary calculations can be organized in essentially an independent pattern. An example of such a computer is the Connection Machine CM-5 at the National Center for Supercomputing Applications in Illinois with 1,024 identical processors in a multiple-instruction/multiple-data (MIMDI) configuration. The benefits of such a parallel architecture on the problem of solving an econometric optimization classical estimator not involving simulation can also be substantial, since such estimators involve the evaluation of contributions to the criterion (for example, likelihood) function in the case of independently and identically distributed (i.i.d.) observations. Since typical applications in modern applied econometrics using cross-sectional and longitudinal data sets involve several thousands of i.i.d. observations, the potential benefits of parallel calculations of such estimators should be obvious. The benefits of a massively parallel computer architecture become even more pronounced in the case of simulation-based estimators. See Nagurney (1996) for a discussion of parallel computation in econometrics.

An alternative approach for parallel computation that does not involve a single computer with many processors has been developed recently and offers considerable promise for computational econometrics. Through the use of specialized computer languages, many separate computers are harnessed together over an organization's intranet or even over the internet, and an econometric computation task is distributed across them. The benefits of this approach depend critically on the relative burden of the overhead of communicating across the individual computers when organizing the splitting of the tasks and then collecting and processing the separate partial results. Such distributed parallel computation has the exciting potential of affording formidable super-computing powers to econometric researchers with only modest computer hardware.

See Also

- longitudinal data analysis
- McFadden, Daniel
- nonlinear programming
- numerical optimization methods in economics
- robust estimators in econometrics
- simulation-based estimation

Bibliography

Abramowitz, M. and Stegun, I. 1964. *Handbook of Mathematical Functions*. Washington, DC: National Bureau of Standards.

Balestra, P. and Nerlove, M. 1966. Pooling cross-section and time-series data in the estimation of a dynamic model. *Econometrica* 34, 585–612.

Belsley, D. 1974. Estimation of system of simultaneous equations and computational specifications of GREMLIN. *Annals of Economic and Social Measurement* 3, 551–614.

Berndt, E.K., Hall, B.H., Hall, R.E. and Hausman, J.A. 1974. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* 3, 653–66.

Davis, P.J. and Rabinovitz, P. 1984. *Methods of Numerical Integration*. New York: Academic Press.

Dennis, J.E. and Schnabel, R.B. 1984. *Unconstrained optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Drud, A. 1977. An optimization code for nonlinear econometric models based on sparse matrix techniques and reduced grades. *Annals of Economic and Social Measurement* 6, 563–80.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. CBMS-NSF Monographs No. 38. Philadelphia: SIAM.

Fuller, W.A. and Battese, G.E. 1973. Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association* 68, 626–32.

Geweke, J. 1988. Antithetic acceleration of Monte Carlo integration in Bayesian inference. *Journal of Econometrics* 38, 73–90.

Geweke, J. 1996. Monte Carlo simulation and numerical integration. In *Handbook of Computational Economics*, vol. 1, ed. H. Amman, D. Kendrik and J. Rust. Amsterdam: North-Holland.

Golub, G.H. 1969. Matrix decompositions and statistical calculations. In *Statistical Computation*, ed. R. C. Milton and J.A. Milder. New York: Academic Press.

Goldfeld, S. and Quandt, R. 1972. *Nonlinear Methods in Econometrics*. Amsterdam: North-Holland.

Hajivassiliou, V.A. and Ruud, P.A. 1994. Classical estimation methods using simulation. In *Handbook of Econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

Hajivassiliou, V.A., McFadden, D.L. and Ruud, P.A. 1996. Simulation of multivariate normal rectangle probabilities and derivatives: theoretical and computational results. *Journal of Econometrics* 72(1, 2), 85–134.

Judd, K. 1996. Approximation, perturbation, and projection methods in economic analysis. In *Handbook of Computational Economics*, vol. 1, ed. H. Amman, D. Kendrik and J. Rust. Amsterdam: North-Holland.

Manski, C. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

McFadden, D. 1989. A method of simulated moments for estimation of multinomial discrete response models. *Econometrica* 57, 995–1026.

Nagurney, A. 1996. Parallel computation. In *Handbook of Computational Economics*, vol.1, ed. H. Amman, D. Kendrik and J. Rust. Amsterdam: North-Holland.

Nelder, J.A. and Meade, R. 1965. A simplex method for function minimization. *Computer Journal* 7, 308–13.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. 2001. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Quandt, R. 1983. Computational problems and methods. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Zellner, A., Bauwens, L. and VanDijk, H. 1988. Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods. *Journal of Econometrics* 38, 73–90.

How to cite this article

Hajivassiliou, Vassilis A. "computational methods in econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000559>
doi:10.1057/9780230226203.0285

The New Palgrave Dictionary of Economics Online

computer industry

Shane Greenstein

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Commercial computing has grown to include an extraordinary range of economic undertakings. In any given era, computing markets are organized around platforms – a cluster of technically standardized components that buyers use together to make the wide range of applications. There has been an increasing secular trend in the number of firms that possess the necessary technical knowledge and commercial capabilities to bring to market some component or service. While general improvements in technical capabilities are readily apparent, it is quite difficult to calculate the productivity improvements arising from increased investment in and use of computing.

Keywords

computer industry; economic growth; information technology; innovation; Moore's law

Article

The commercial computing industry accounts for a large fraction of economic activity. From its military and research origins in the late 1940s, it spread into the commercial realm and has since grown to include an extraordinary range of economic undertakings. Many economists believe this expansion of applications for computing has been a driver of economic growth.

Computing aids the automated tracking of transactions, a function that finds use, for example, in automating billing, managing the pricing of inventories of airline seating, and restocking retail outlets in a geographically dispersed organization. It also facilitates the coordination of information-intensive tasks, such as the dispatching of time-sensitive deliveries or emergency services. Computing also enables performance of advanced mathematical calculations, useful in such diverse activities as calculating interest on loans and generating estimates of underground geological deposits. Computer-aided precision also improves the efficiency of processes such as manufacturing metal shapes or the automation of communication switches, to name just two.

In any given era, computing markets are organized around platforms – a cluster of technically

standardized components that buyers use together to make the aforementioned wide range of applications. Such platforms involve long-lived assets, both components sold in markets (that is, hardware and some software) and components made by buyers (that is, training and most software). Important computing platforms historically include the UNIVAC, the IBM 360 and its descendents, the Wang minicomputers, IBM AS/400, DEC VAX, Sun SPARC, Intel/Windows PC, Unix/Linux, and, after the mid-1990s, TCP/IP-based client-server platforms linked together.

Vendors tend to sell groups of compatible products under umbrella strategies aimed at the users of particular platforms. In the earliest eras of computing markets, the leading firms integrated all facets of computing and offered a supply of goods and services from a centralized source. In later eras, the largest and most popular platforms historically included many different computing, communications and peripheral equipment firms, software tool developers, application software writers, consultants, system integrators, distributors, user groups, news publications and service providers.

Until the early 1990s, most market segments were distinguished by the size of the tasks to be undertaken and by the technical sophistication of the typical user. Mainframes, minicomputers, workstations, and personal computers, in decreasing order, constituted different size-based market segments. Trained engineers or programmers made up the technical user base, while the commercial market was geared more towards administrators, secretaries and office assistants.

The most popular platform in the late 1980s and 1990s differed from the prominent platforms of earlier years. The personal computer (PC) began in the mid-1970s as an object of curiosity among technically skilled hobbyists, but became a common office tool after the entry of IBM's design. Unlike prior computing platforms, this one has diffused into both home and business use. From the beginning, this platform involved thousands of large and small software developers, third-party peripheral equipment and card developers, and a few major players. In more recent experience, control over the standard has completely passed from IBM to Microsoft and Intel. Microsoft produces the Windows operating system and Intel produces the most commonly used microprocessor. For this reason the platform is often called Wintel.

The networking and internet revolution in the late 1990s is responsible for blurring once-familiar distinctions. These new technologies have made it feasible to build client-server systems within large enterprises and across ownership boundaries. It employs internet-based computing systems networked across potentially vast geographic distances, supporting the emergence of a ‘network of networks’.

Despite frequent and sometimes dramatic technical improvements in specific areas of technology, many features of the most common platforms in use tend to persist or change very slowly. Many durable components make up platforms. And, though they lose their market value as they become obsolete in comparison with frontier products, they do not as quickly lose their ability to provide a flow of services to users. Consequently, new technology tends to be most successful when new components enhance and preserve the value of previous investments, a factor that creates demand for ‘backward compatible’ upgrades or improvements. It also creates a demand for support and service activities to reduce the costs of making the transition from old to new.

Control over changes to design and other aspects of technical standards shapes the backward compatibility for key components. Control of these decisions is coincident with platform leadership – determining the rate and direction of change in technical features of components around which other firms build their businesses. In each platform, it is very rare to observe more than a small number of firms acquiring leadership positions. Since such positions have been historically associated with high

firm profitability, firms compete fiercely for market dominance in component categories where standards are essential. Not surprisingly, competitive behaviour affiliated with obtaining and retaining market leadership does occasionally receive attention from antitrust authorities.

Though innovative change in computing began well prior to the invention of the integrated circuit, in popular discussion advances in computing have become almost synonymous with advances in microprocessors. This is due to an observation by Gordon Moore, who co-founded and became chairman at Intel. In 1965 he foresaw a doubling of circuits per chip every two years. This prediction about the rate of technical advance later became known as ‘Moore's law’. In fact, microprocessors and DRAMS have been doubling in capability every 18 months since the mid-1970s.

Moore's prediction pertained narrowly to integrated circuits. However, a similar pattern of improvement – though with variation in the rate – characterizes other electronic components that go into producing a computer or that are complementary with computing in many standard uses. This holds for disk drives, display screens, routing equipment, and data-transmission capacity, to name a few. Such widespread innovation creates opportunities for new entry and rearrangements in the conditions of supply.

Accordingly, there has been an increasing secular trend in the number of firms that possess the necessary technical knowledge and commercial capabilities to bring to market some component or service of value to computing users. This factor alone explains the increasing complexity of supply chains for the supply of most computing hardware and software products. It is also coincident with their increasing geographic reach. In addition, as in other manufacturing processes, the increasing use of sophisticated information technology helps coordinate design and production involving firms from many countries and continents.

While the spawning of new information technology businesses in North America has tended to be concentrated in a small number of locations, such as the Boston area and Santa Clara Valley (popularly known as Silicon Valley), every other facet of the supply chain for computing involves firms headquartered and operating in a much wider set of locations. In North America, these range from Seattle, Austin, Los Angeles, the greater New York area, Denver–Boulder, Washington DC, the North Carolina Research Triangle, Chicago, and virtually all major cities in the United States. The supply chain for many complementary components has also been associated with many firms in Western Europe and as well as in India, Israel, South Korea, Singapore, Taiwan and China. Even more widespread are computing service firms, which follow business and home users dispersed across the globe.

Despite this geographic dispersion since the 1950s, US companies have retained leadership in generating new platforms and commercializing frontier technologies in forms that most users find valuable. Part of this results from the persistence of platform leadership for a time within a segment. In addition, US firms have historically been ascendant whenever platform leadership has changed. However, this pattern seems likely to change in the 21st century, as non-US firms already have found leadership positions in producing components of many platforms and in related areas of electronics, such as consumer electronics, communication equipment and specialized software.

While general improvements in technical capabilities are readily apparent, it is quite difficult to calculate the productivity improvements arising from increased investment in and use of computing. There is no question that existing computing activities have become less expensive, while new capabilities have been achieved. This has allowed economic actors to attain previously unobtainable outcomes. This shift in economic possibilities has generated a restructuring of organizational routines, market relationships,

and other activities associated with the flow of goods, which inevitably improves the economy's ability to transform inputs into consumer welfare.

Yet altering the business use of computing can be slow. It often demands large adjustment costs and gradual learning about which organizational processes can best employ advances in computing. It can involve a reallocation of decision rights and discretion inside a large organization, especially when business units alter a wide array of intermediate routine processes (such as billing, account monitoring, and inventory management) or the coordination of services (such as the delivery of data for decision support). Moreover, the largest changes come from altering many complementary activities that respond to new and unanticipated opportunities, setting off new waves of invention. Each wave's productivity effect is interwoven with others.

Along with these improvements the boundaries of the ‘computing market’ have changed. A hardware-based definition for the computing market was barely adequate in the 1960s and is no longer adequate for economic analysis. However, there is no consensus about what alternative framing will be appropriate for understanding value creation, supplier behaviour, and user adoption in computing in the 21st century.

See Also

- diffusion of technology
- general purpose technologies
- information technology and the world economy
- Internet, economics of the
- technical change

Bibliography

Bresnahan, T. and Greenstein, S. 1999. Technological competition and the structure of the computer industry. *Journal of Industrial Economics* 47(1), 1–40.

Brensnahan, T. and Malerba, F. 1999. Industrial dynamics and the evolution of firm's and nations' competitive capabilities in the world computer industry. In *Sources of Industrial Leadership*, ed. D. Mowery and R. Nelson. Cambridge, UK: Cambridge University Press.

Brynjolfsson, E. and Lorin, H. 2000. Beyond computation: information technology, organizational transformation and business performance. *Journal of Economic Perspectives* 14(4), 23–48.

Dedrick, J. and Kraemer, K. 2005. The impacts of IT on firm and industry structure: the personal computer industry. *California Management Review* 47(3), 122–42.

Flamm, K. 2003. The new economy in historical perspective: evolution of digital electronics technology. In *New Economy Handbook*, ed. D. Jones. San Diego: Academic Press/Elsevier.

Greenstein, S., ed. 2006. *The Industrial Economics of Computing*. Northampton, MA: Edward Elgar.

Jorgenson, D. and Wessner, C., eds. 2005. *Deconstructing the Computer: Report of a Symposium*. Washington, DC: National Academies Press.

McKinsey Global Institute. 2001. *U.S. Productivity Growth, 1995–2000: Understanding the Contribution of Information Technology Relative to Other Factors*. Washington, DC: McKinsey and Co.

How to cite this article

Greenstein, Shane. "computer industry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000543> doi:10.1057/9780230226203.0286

The New Palgrave Dictionary of Economics Online

computer science and game theory

Joseph Y. Halpern

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Work at the intersection of computer science and game theory is briefly surveyed, with a focus on the work in computer science. In particular, the following topics are considered: various roles of computational complexity in game theory, including modelling bounded rationality, its role in mechanism design, and the problem of computing Nash equilibria; the *price of anarchy*, that is, the cost of using decentralizing solution to a problem; and interactions between distributed computing and game theory.

Keywords

algorithmic knowledge; algorithmic mechanism design; Bayesian networks; bounded rationality; Byzantine agreement; cheap talk; coalitions; combinatorial auctions; complexity theory; computational complexity; computer science and game theory; distributed computing; efficient representation of games; game theory; Gibbard–Satterthwaite th; implementing mediators; interactive epistemology; *k*-resilient equilibrium; (*k*,*t*)-robust equilibrium; learning; mechanism design; Markov networks; Nash equilibrium; price of anarchy; Prisoner's Dilemma; regret; strategic voting; tit for tat; voting

Article

1 Introduction

There has been a remarkable increase in work at the interface of computer science and game theory in the past decade. Game theory forms a significant component of some major computer science conferences (see, for example, Kearns and Reiter, 2005; Sandholm and Yokoo, 2003); leading computer scientists are often invited to speak at major game theory conferences, such as the World Congress on Game Theory 2000 and 2004. In this article I survey some of the main themes of work in the area, with a focus on the work in computer science. Given the length constraints, I make no attempt at being

comprehensive, especially since other surveys are also available, including Halpern (2003), Linial (1994), Papadimitriou (2001), and a comprehensive survey book (Nisan et al., 2007).

The survey is organized as follows. I look at the various roles of computational complexity in game theory in Section 2, including its use in modelling bounded rationality, its role in mechanism design, and the problem of computing Nash equilibria. In Section 3, I consider a game-theoretic problem that originated in the computer science literature, but should be of interest to the game theory community: computing the *price of anarchy*, that is, the cost of using a decentralizing solution to a problem. In Section 4 I consider interactions between distributed computing and game theory.In Section 5, I consider the problem of implementing mediators, which has been studied extensively in both computer science and game theory. I conclude in Section 6 with a discussion of a few other topics of interest.

2 Complexity considerations

The influence of computer science in game theory has perhaps been most strongly felt through complexity theory. I consider some of the strands of this research here. There are a numerous basic texts on complexity theory that the reader can consult for more background on notions like NP-completeness and finite automata, including Hopcroft and Ullman (1979) and Papadimitriou (1994a).

2.1 Bounded rationality

One way of capturing bounded rationality is in terms of agents who have limited computational power. In economics, this line of research goes back to the work of Neyman (1985) and Rubinstein (1986), who focused on finitely repeated Prisoner's Dilemma. In *n*-round finitely repeated Prisoner's Dilemma, there are 2^{2^n-1} strategies (since a strategy is a function from histories to {cooperate, defect}, and there are clearly 2^n–1 histories of length<*n*). Finding a best response to a particular move can thus potentially be difficult. Clearly people do not find best responses by doing extensive computation. Rather, they typically rely on simple heuristics, such as ‘tit for tat’ (Axelrod, 1984). Such heuristics can often be captured by finite automata; both Neyman and Rubinstein thus focus on finite automata playing repeated Prisoner's Dilemma. Two computer scientists, Papadimitriou and Yannakakis (1994), showed that if both players in an *n*-round Prisoner's Dilemma are finite automata with at least 2^n–1 states, then the only equilibrium is the one where they defect in every round. This result says that a finite automaton with exponentially many states can compute best responses in Prisoner's Dilemma.

We can then model bounded rationality by restricting the number of states of the automaton. Neyman (1985) showed, roughly speaking, that if the two players in *n*-round Prisoner's Dilemma are modelled by finite automata with a number of states in the interval [$n^{1/k}$, n^k] for some *k*, then collaboration can be approximated in equilibrium; more precisely, if the payoff for (cooperate, cooperate) is (3, 3) there is an equilibrium in the repeated game where the average payoff per round is greater than $3-\frac{1}{k}$ for each player. Papadimitriou and Yannakakis (1994) sharpen this result by showing that if at least one of the players has fewer than $2^{c_\epsilon n}$ states, where $c_\epsilon = \frac{\epsilon}{12(1+\epsilon)}$, then for sufficiently large *n*, there is an

equilibrium where each player's average payoff per round is greater than 3– ε . Thus, computational limitations can lead to cooperation in Prisoner's Dilemma.

There have been a number of other attempts to use complexity-theoretic ideas from computer science to model bounded rationality (see Rubinstein, 1998, for some exs). However, it seems that there is much more work to be done here.

2.2 Computing Nash equilibrium

Nash (1950) showed every finite game has a Nash equilibrium in mixed strategies. But how hard is it to actually find that equilibrium? On the positive side, there are well known algorithms for computing Nash equilibrium, going back to the classic Lemke–Howson (1964) algorithm, with a spate of recent improvements (see, for example, Govindan and Wilson, 2003; Blum, Shelton and Koller, 2003; Porter, Nudelman and Shoham, 2004). Moreover, for certain classes of games (for example, symmetric games, Papadimitriou and Roughgarden, 2005), there are known to be polynomial-time algorithms. On the negative side, many qsts about Nash equilibrium are known to be NP-hard. For example, Gilboa and Zemel (1989) showed that, for a game presented in normal form, deciding whether there exists a Nash equilibrium where each player gets a payoff of at least *r* is NP-complete. Interestingly, Gilboa and Zemel also show that computing whether there exists a *correlated* equilibrium (Aumann, 1987) where each player gets a payoff of at least *r* is computable in polynomial time. In general, qsts regarding correlated equilibrium seem easier than the analogous qsts for Nash equilibrium; see Papadimitriou (2005) and Papadimitriou and Roughgarden (2005) for further examples. Chu and Halpern (2001) prove similar NP-completeness results if the game is represented in extensive form, even if all players have the same payoffs (a situation that arises frequently in computer science applications, where we can view the players as agents of some designer, and take the payoffs to be the designer's payoffs). Conitzer and Sandholm (2003) give a compendium of hardness results for various qsts regarding Nash equilibria. Nevertheless, there is a sense in which it seems that the problem of finding a Nash equilibrium is easier than typical NP-complete problems, because every game is guaranteed to have a Nash equilibrium. By way of contrast, for a typical NP-complete problem like prptal satisfiability, whether or not a prptal formula is satisfiable is not known. Using this observation, it can be shown that if finding a Nash equilibrium is NP-complete, then NP=coNP. Recent work has in a sense completely characterized the complexity of finding a Nash equilibrium in normal-form games: it is a *PPAD-complete* problem (Chen and Deng, 2006; Daskalis, Goldberg and Papadimitriou, 2006). PPAD stands for ‘polynomial parity argument (directed case)’; see Papadimitriou (1994b) for a formal definition and examples of other PPAD problems. It is believed that PPAD-complete problems are not solvable in polynomial time, but are simpler than NP-complete problems, although this remains an open problem. See Papadimitriou (2007) for an overview of this work.

2.3 Algorithmic mechanism design

The problem of mechanism design is to design a game such that the agents playing the game, motivated only by self-interest, achieve the designer's goals. This problem has much in common with the standard

computer science problem of designing protocols that satisfy certain specifications (for example, designing a distributed protocol that achieves Byzantine agreement; see Section 4). Work on mechanism design has traditionally ignored computational concerns. But Kfir-Dahav, Monderer and Tennenholtz (2000) show that, even in simple settings, optimizing social welfare is NP-hard, so that perhaps the most common approach to designing mechanisms, applying the Vickrey–Groves–Clarke (VCG) procedure (Clarke, 1971; Groves, 1973; Vickrey, 1961), is not going to work in large systems. We might hope that, even if we cannot compute an optimal mechanism, we might be able to compute a reasonable approximation to it. However, as Nisan and Ronen (2000; 2001) show, in general, replacing a VCG mechanism by an approximation does not preserve truthfulness. That is, even though truthfully revealing one's type is an optimal strategy in a VCG mechanism, it may no longer be optimal in an approximation. Following Nisan and Ronen's work, there has been a spate of papers either describing computationally tractable mechanisms or showing that no computationally tractable mechanism exists for a number of problems, ranging from task allocation (Archer and Tardos, 2001; Nisan and Ronen, 2001) to cost-sharing for multicast trees (Feigenbaum, Papadimitriou and Shenker, 2000) (where the problem is to share the cost of sending, for example, a movie over a network among the agents who actually want the movie) to finding low-cost paths between nodes in a network (Archer and Tardos, 2002).

The problem that has attracted perhaps the most attention is *combinatorial auctions*, where bidders can bid on bundles of items. This becomes of particular interest in situations where the value to a bidder of a bundle of goods cannot be determined by simply summing the value of each good in isolation. To take a simple example, the value of a pair of shoes is much higher than that of the individual shoes; perhaps more interestingly, an owner of radio stations may value having a licence in two adjacent cities more than the sum of the individual licences. Combinatorial auctions are of great interest in a variety of settings including spectrum auctions, airport time slots (that is, take-off and landing slots), and industrial procurement. There are many complexity-theoretic issues related to combinatorial auctions. For a detailed discussion and references see Cramton, Shoham and Steinberg (2006); I briefly discuss a few of the issues involved here.

Suppose that there are n items being auctioned. Simply for a bidder to communicate her bids to the auctioneer can take, in general, exponential time, since there are 2^n bundles. In many cases, we can identify a bid on a bundle with the bidder's valuation of the bundle. Thus, we can try to carefully design a bidding language in which a bidder can communicate her valuations succinctly. Simple information-theoretic arguments can be used to show that, for every bidding language, there will be valuations that will require length at least 2^n to express in that language. Thus, the best we can hope for is to design a language that can represent the ‘interesting’ bids succinctly. See Nisan (2006) for an overview of various bidding languages and their expressive power.

Given bids from each of the bidders in a combinatorial auction, the auctioneer would like to then determine the winners. More precisely, the auctioneer would like to allocate the m items in an auction so as to maximize his revenue. This problem, called the *winner determination problem*, is NP-complete in general, even in relatively simple classes of combinatorial auctions with only two bidders making rather restricted bids. Moreover, it is not even polynomial-time approximable, in the sense that there is no constant d and polynomial-time algorithm such that the algorithm produces an allocation that gives revenue that is at least $1/d$ of optimal. On the other hand, there are algorithms that provably find a good solution, seem to work well in practice, and, if they seem to be taking too long, can be terminated early,

usually with a good feasible solution in hand. See Lehmann, Müller and Sandholm, (2006) for an overview of the results in this area.

In most mechanism design problems, computational complexity is seen as the enemy. There is one class of problems in which it may be a friend: voting. One problem with voting mechanisms is that of *manipulation* by voters. That is, voters may be tempted to vote strategically rather than ranking the candidates according to their true preferences, in the hope that the final outcome will be more favourable. This situation arises frequently in practice; in the 2000 US presidential election, American voters who preferred Nader to Gore to Bush were encouraged to vote for Gore, rather than ‘wasting’ a vote on Nader. The classic Gibbard–Satterthwaite theorem (Gibbard, 1973; Satterthwaite, 1975) shows that, if there are at least three alternatives, then in any nondictatorial voting scheme (that is, one where it is *not* the case that one particular voter dictates the final outcome, irrespective of how the others vote), there are preferences under which an agent is better off voting strategically. The hope is that, by constructing the voting mechanism appropriately, it may be computationally intractable to find a manipulation that will be beneficial. While finding manipulations for the plurality protocol (the candidate with the most votes wins) is easy, there are well-known voting protocols for which manipulation is hard in the presence of three or more candidates. See Conitzer, Sandholm and Lang (2007) for a summary of results and further pointers to the literature.

2.4 Communication complexity

Most mechanisms in the economics literature are designed so that agents truthfully reveal their preferences. However, in some settings, revealing one's full preferences can require a prohibitive amount of communication. For example, in a combinatorial auction of m items, revealing one's full preferences may require revealing what one would be willing to pay for each of the 2^m–1 possible bundles of items. Even if m=30, this requires revealing more than one billion numbers. This leads to an obvious qst: how much communication is required by various mechanisms? Formal work on this question in the economics community goes back to Hurwicz (1977) and Mount and Reiter (1974); their definitions focused on the dimension of the message space. Independently (and later), there was active work in computer science on *communication complexity*, the number of bits of communication needed for a set of n agents to compute the value of a function $f : \times_{i=1}^{n} \Theta_i \rightarrow X$, where each agent i knows $\theta_i \in \Theta_i$. (Think of θ_i as representing agent i's type.) Recently there has been an explosion of work, leading to a better understanding of the communication complexity for many important economic allocation problems; see Segal (2006) for an overview. Two important themes in this work are understanding the role of price-based market mechanisms in solving allocation problems with minimal communication, and designing mechanisms that provide agents with incentives to communicate truthfully while having low communication requirements.

3 The price of anarchy

In a computer system, there are situations where we may have a choice between a centralized and a decentralized solution to a problem. By ‘centralized’ here, I mean that each agent in the system is told

exactly what to do and must do so; in the decentralized solution, each agent tries to optimize his own selfish interests. Of course, centralization comes at a cost. For one thing, there is a problem of enforcement. For another, centralized solutions tend to be more vulnerable to failure. On the other hand, a centralized solution may be more socially beneficial. How much more beneficial can it be? Koutsoupias and Papadimitriou (1999) formalized this question by considering the ratio of the social welfare of the centralized solution to the social welfare of the Nash equilibrium with the worst social welfare (assuming that the social welfare function is always positive). They called this ratio the *price of anarchy*, and proved a number of results regarding the price of anarchy for a scheduling problem on parallel machines. Since the original paper, the price of anarchy has been studied in many settings, including traffic routing (Roughgarden and Tardos, 2002), facility location games (for example, where is the best place to put a factory) (Vetta, 2002), and spectrum sharing (how should channels in a WiFi network be assigned) (Halldórsson et al., 2004).

To give a sense of the results, consider the traffic-routing context of Roughgarden and Tardos (2002). Suppose that the travel time on a road increases in a known way with the congestion on the road. The goal is to minimize the average travel time for all drivers. Given a road network and a given traffic load, a centralized solution would tell each driver which road to take. For example, there could be a rule that cars with odd-numbered licence plates take road 1, while those with even-numbered plates take road 2, to minimize congestion on either road. Roughgarden and Tardos show that the price of anarchy is unbounded if the travel time can be a nonlinear function of the congestion. On the other hand, if it is linear, they show that the price of anarchy is at most 4/3.

The price of anarchy is but one way of computing the ‘cost’ of using a Nash equilibrium. Others have been considered in the computer science literature. For example, Tennenholtz (2002) compares the *safety level* of a game – the optimal amount that an agent can guarantee himself, independent of what the other agents do – to what the agent gets in a Nash equilibrium, and shows, for interesting classes of games, including load-balancing games and first-price auctions, that the ratio between the safety level and the Nash equilibrium is bounded. For example, in the case of first-price auctions, it is bounded by the constant *e*.

4 Game theory and distributed computing

Distributed computing and game theory are interested in much the same problems: dealing with systems where there are many agents, facing uncertainty and having possibly different goals. In practice, however, there has been a significant difference in emphasis between the two areas. In distributed computing, the focus has been on problems such as fault tolerance, asynchrony, scalability, and proving correctness of algorithms; in game theory, the focus has been on strategic concerns. I discuss here some issues of common interest. Most of the discussion in the remainder of this section is taken from Halpern (2003).

To understand the relevance of fault tolerance and asynchrony, consider the *Byzantine agreement* problem, a paradigmatic problem in the distributed systems literature. In this problem, there are assumed to be *n* soldiers, up to *t* of which may be faulty (the *t* stands for *traitor*); *n* and *t* are assumed to be common knowledge. Each soldier starts with an initial preference, to either attack or retreat. (More precisely, there are two types of nonfaulty agents – those that prefer to attack, and those that prefer to

retreat.) We want a protocol that guarantees that (1) all *nonfaulty* soldiers reach the same decision, and (2) if all the soldiers are nonfaulty and their initial preferences are identical, then the final decision agrees with their initial preferences. (The condition simply prevents the obvious trivial solutions, where the soldiers attack no matter what, or retreat no matter what.)

The problem was introduced by Pease, Shostak and Lamport (1980), and has been studied in detail since then; Chor and Dwork (1989), Fischer (1983), and Linial (1994) provide overviews. Whether the Byzantine agreement problem is solvable depends in part on what types of failures are considered, on whether the system is *synchronous* or *asynchronous*, and on the ratio of n to t. Roughly speaking, a system is synchronous if there is a global clock and agents move in lockstep; a 'step' in the system corresponds to a tick of the clock. In an asynchronous system, there is no global clock. The agents in the system can run at arbitrary rates relative to each other. One step for agent 1 can correspond to an arbitrary number of steps for agent 2 and vice versa. Synchrony is an implicit assumption in essentially all games. Although it is certainly possible to model games where player 2 has no idea how many moves player 1 has taken when player 2 is called upon to move, it is not typical to focus on the effects of synchrony (and its lack) in games. On the other hand, in distributed systems, it is typically a major focus.

Suppose for now that we restrict to *crash failures*, where a faulty agent behaves according to the protocol, except that it might crash at some point, after which it sends no messages. In the round in which an agent fails, the agent may send only a subset of the messages that it is supposed to send according to its protocol. Further suppose that the system is synchronous. In this case, the following rather simple protocol achieves Byzantine agreement:

- In the first round, each agent tells every other agent its initial preference.
- For rounds 2 to t+1, each agent tells every other agent everything it has heard in the previous round. Thus, for example, in round 3, agent 1 may tell agent 2 that it heard from agent 3 that its initial preference was to attack, and that it (agent 3) heard from agent 2 that its initial preference was to attack, and it heard from agent 4 that its initial preferences was to retreat, and so on. This means that messages get exponentially long, but it is not difficult to represent this information in a compact way so that the total communication is polynomial in n, the number of agents.
- At the end of round t+1, if an agent has heard from any other agent (including itself) that its initial preference was to attack, it decides to attack; otherwise, it decides to retreat.

Why is this correct? Clearly, if all agents are correct and want to retreat (resp., attack), then the final decision will be to retreat (resp., attack), since that is the only preference that agents hear about (recall that for now we are considering only crash failures). It remains to show that if some agents prefer to attack and others to retreat, then all the nonfaulty agents reach the same final decision. So suppose that i and j are nonfaulty and i decides to attack. That means that i heard that some agent's initial preference was to attack. If it heard this first at some round $t' < t+1$, then i will forward this message to j, who will receive it and thus also attack. On the other hand, suppose that i heard it first at round t+1 in a message from i_{t+1}. Thus, this message must be of the form 'i_t said at round t that … that i_2 said at round 2 that i_1 said at round 1 that its initial preference was to attack.' Moreover, the agents $i_1,\ldots, i_{t+1}$ must all be distinct. Indeed, it is easy to see that i_k must crash in round k before sending its message to i (but after sending its message to i_{k+1}), for k=1,…, t, for otherwise i must have gotten the message from i_k,

contradicting the assumption that i first heard at round t+1 that some agent's initial preference was to attack. Since at most t agents can crash, it follows that i_{t+1}, the agent that sent the message to i, is not faulty, and thus sends the message to j. Thus, j also decides to attack. A symmetric argument shows that if j decides to attack, then so does i.

It should be clear that the correctness of this protocol depends on both the assumptions made: crash failures and synchrony. Suppose instead that *Byzantine* failures are allowed, so that faulty agents can deviate in arbitrary ways from the protocol; they may ‘lie’, send deceiving messages, and collude to fool the nonfaulty agents in the most malicious ways. In this case, the protocol will not work at all. In fact, it is known that agreement can be reached in the presence of Byzantine failures iff $t<n/3$, that is, iff fewer than a third of the agents can be faulty (Pease, Shostak and Lamport, 1980). The effect of asynchrony is even more devastating: in an asynchronous system, it is impossible to reach agreement using a deterministic protocol even if t=1 (so that there is at most one failure) and only crash failures are allowed (Fischer, Lynch and Paterson, 1985). The problem in the asynchronous setting is that if none of the agents have heard from, say, agent 1, they have no way of knowing whether agent 1 is faulty or just slow. Interestingly, there are randomized algorithms (that is, behavioural strategies) that achieve agreement with arbitrarily high probability in an asynchronous setting [Ben-Or, 1983; Rabin, 1983].

Byzantine agreement can be viewed as a game where, at each step, an agent can either send a message or decide to attack or retreat. It is essentially a game between two teams, the nonfaulty agents and the faulty agents, whose composition is unknown (at least by the correct agents). To model it as a game in the more traditional sense, we could imagine that the nonfaulty agents are playing against a new player, the ‘adversary’. One of the adversary's moves is that of ‘corrupting’ an agent: changing its type from ‘nonfaulty’ to ‘faulty.’ Once an agent is corrupted, what the adversary can do depends on the failure type being considered. In the case of crash failures, the adversary can decide which of a corrupted agent's messages will be delivered in the round in which the agent is corrupted; however, it cannot modify the messages themselves. In the case of Byzantine failures, the adversary essentially gets to make the moves for agents that have been corrupted; in particular, it can send arbitrary messages.

Why has the distributed systems literature not considered strategic behaviour in this game? Crash failures are used to model hardware and software failures; Byzantine failures are used to model random behaviour on the part of a system (for example, messages getting garbled in transit), software errors, and malicious adversaries (for example, hackers). With crash failures, it does not make sense to view the adversary's behaviour as strategic, since the adversary is not really viewed as having strategic interests. While it would certainly make sense, at least in principle, to consider the probability of failure (that is, the probability that the adversary corrupts an agent), this approach has by and large been avoided in the literature because it has proved difficult to characterize the probability distribution of failures over time. Computer components can perhaps be characterized as failing according to an exponential distribution (see Babaoglu, 1987, for an analysis of Byzantine agreement in such a setting), but crash failures can be caused by things other than component failures (faulty software, for ex); these can be extremely difficult to characterize probabilistically. The problems are even worse when it comes to modelling random Byzantine behaviour.

With malicious Byzantine behaviour, it may well be reasonable to impute strategic behaviour to agents (or to an adversary controlling them). However, it is often difficult to characterize the payoffs of a malicious agent. The goals of the agents may vary from that of simply trying to delay a decision to that

of causing disagreement. It is not clear what the appropriate payoffs should be for attaining these goals. Thus, the distributed systems literature has chosen to focus instead on algorithms that are guaranteed to satisfy the specification without making assumptions about the adversary's payoffs (or nature's probabilities, in the case of crash failures).

Recently, there has been some work on adding strategic concerns to standard problems in distributed computing; see, for example, Alvisi et al. (2005) and Halpern and Teague (2004). Moving in the other direction, there has also been some work on adding concerns of fault tolerance and asynchrony to standard problems in game theory; see, for example, Eliaz (2002), Monderer and Tennenholtz (1999a; 1999b) and the definitions in the next section. This seems to be an area that is ripe for further developments. One such development is the subject of the next section.

5 Implementing mediators

The question of whether a problem in a multiagent system that can be solved with a trusted mediator can be solved by just the agents in the system, without the mediator, has attracted a great deal of attention in both computer science (particularly in the cryptography community) and game theory. In cryptography, the focus on the problem has been on *secure multiparty computation*. Here it is assumed that each agent i has some private information x_i. Fix functions $f_1,\ldots, f_n$. The goal is to have agent i learn $f_i(x_1,\ldots, x_n)$ without learning anything about x_j for $j \neq i$ beyond what is revealed by the value of $f_i(x_1,\ldots, x_n)$. With a trusted mediator, this is trivial: each agent i just gives the mediator its private value x_i; the mediator then sends each agent i the value $f_i(x_1,\ldots, x_n)$. Work on multiparty computation (Goldreich, Micali and Wigderson, 1987; Shamir, Rivest and Adelman, 1981; Yao, 1982) provides conditions under which this can be done. In game theory, the focus has been on whether an equilibrium in a game with a mediator can be implemented using what is called *cheap talk* – that is, just by players communicating among themselves (cf. Barany, 1992; Ben-Porath, 2003; Forges, 1990; Gerardi, 2004; Heller, 2005; Urbano and Vila, 2004). As suggested in the previous section, the focus in the computer science literature has been in doing multiparty computation in the presence of possibly malicious adversaries, who do everything they can to subvert the computation, while in the game theory literature the focus has been on strategic agents. In recent work, Abraham et al. (2006) and Abraham, Dolev and Halpern (2007) considered deviations by both rational players, who have preferences and try to maximize them, and players who can viewed as malicious, although it is perhaps better to think of them as rational players whose utilities are not known by the other players or mechanism designer. I briefly sketch their results here; the following discussion is taken from Abraham, Dolev and Halpern (2007).

The idea of tolerating deviations by coalitions of players goes back to Aumann (1959); more recent refinements have been considered by Moreno and Wooders (1996). Aumann's definition is essentially the following.

Definition 1: $\vec{\sigma}$ is a k-resilient $'$ equilibrium *if, for all sets C of players with* $|C| \leq k$, *it is not the case that there exists a strategy* $\vec{\tau}$ *such that* $u_i(\vec{\tau}_C, \vec{\sigma}_{-C}) > u_i(\vec{\sigma})$ *for all* $i \in C$.

As usual, the strategy $(\vec{\tau}_C, \vec{\sigma}_{-C})$ is the one where each player $i \in C$ plays τ_i and each player $i \notin C$

plays σ_i. As the prime notation suggests, this is not quite the definition we want to work with. The trouble with this definition is that it suggests that coalition members cannot communicate with each other during the game. Perhaps surprisingly, allowing communication can *prevent* certain equilibria (see Abraham, Dolev and Halpern, 2007, for an ex). Since we should expect coalition members to communicate, the following definition seems to capture a more reasonable notion of resilient equilibrium. Let the cheap-talk extension of a game Γ be, roughly speaking, the game where players are allowed to communicate among themselves in addition to performing the actions of Γ and the payoffs are just as in Γ.

Definition 2: $\vec{\sigma}$ is a *k-resilient equilibrium* in a game Γ if $\vec{\sigma}$ is a *k*-resilient $'$ equilibrium in the cheap-talk extension of Γ (where we identify the strategy σ_i in the game Γ with the strategy in the cheap-talk game where player *i* never sends any messages beyond those sent according to σ_i).

A standard assumption in game theory is that utilities are (commonly) known; when we are given a game we are also given each player's utility. When players make decisions, they can take other players' utilities into account. However, in large systems it seems almost invariably the case that there will be some fraction of users who do not respond to incentives the way we expect. For example, in a peer-to-peer network like Kazaa or Gnutella, it would seem that no rational agent should share files. Whether or not you can get a file depends only on whether other people share files. Moreover, there are disincentives for sharing (the possibility of lawsuits, use of bandwidth, and so on). Nevertheless, people do share files. However, studies of the Gnutella network have shown almost 70 per cent of users share no files and nearly 50 per cent of responses are from the top one per cent of sharing hosts (Adar and Huberman, 2000).

One reason that people might not respond as we expect is that they have utilities that are different from those we expect. Alternatively, the players may be irrational, or (if moves are made using a computer) they may be playing using a faulty computer and thus not able to make the move they would like, or they may not understand how to get the computer to make the move they would like. Whatever the reason, it seems important to design strategies that tolerate such unanticipated behaviours, so that the payoffs of the users with ‘standard’ utilities do not get affected by the nonstandard players using different strategies. This can be viewed as a way of adding fault tolerance to equilibrium notions.

Definition 3: A joint strategy $\vec{\sigma}$ is *t-immune* if, for all $T \subseteq N$ with $|T| \leq t$, all joint strategies $\vec{\tau}$, and all $i \notin T$, we have $u_i(\vec{\sigma}_{-T}, \vec{\tau}_T) \geq u_i(\vec{\sigma})$.

The notion of *t*-immunity and *k*-resilience address different concerns. For *t* immunity, we consider the payoffs of the players not in *T*, and require that they are not worse due to deviation; for resilience, we consider the payoffs of players in *C*, and require that they are not better due to deviation. It is natural to combine both notions. Given a game Γ, let $\Gamma^T_{\vec{\tau}}$ be the game that is identical to Γ except that the players in *T* are fixed to playing strategy $\vec{\tau}$.

Definition 4: $\vec{\sigma}$ is a (*k*, *t*) *-robust* equilibrium if $\vec{\sigma}$ is *t*-immune and, for all $T \subseteq N$ such that $|T| \leq t$ and

all joint strategies $\vec{\tau}$, $\vec{\sigma}_{-T}$ is a *k*-resilient strategy of $\Gamma_{\vec{\tau}_T}$.

To state the results of Abraham et al. (2006) and Abraham, Dolev and Halpern (2007) on implementing mediators, three games need to be considered: an *underlying game* Γ, an extension Γ_d of Γ with a mediator, and a cheap-talk extension Γ_{CT} of Γ. Assume that Γ is a *normal-form Bayesian game*: each player has a type from some type space with a known distribution over types, and the utilities of the agents depend on the types and actions taken. Roughly speaking, a cheap talk game *implements* a game with a mediator if it induces the same distribution over actions in the underlying game, for each type vector of the players. With this background, I can summarize the results of Abraham et al. (2006) and Abraham, Dolev and Halpern (2007).

- If $n>3k+3t$, a (k, t) -robust strategy $\vec{\sigma}$ with a mediator can be implemented using cheap talk (that is, there is a (k, t)-robust strategy $\vec{\sigma}'$ in a cheap talk game such that $\vec{\sigma}$ and $\vec{\sigma}'$ induce the same distribution over actions in the underlying game). Moreover, the implementation requires no knowledge of other agents' utilities, and the cheap talk protocol has bounded running time that does not depend on the utilities.
- If $n \leq 3k+3t$, then, in general, mediators cannot be implemented using cheap talk without knowledge of other agents' utilities. Moreover, even if other agents' utilities are known, mediators cannot, in general, be implemented without having a $(k+t)$-punishment strategy (that is, a strategy that, if used by all but at most $(k+t)$ players, guarantees that every player gets a worse outcome than they do with the equilibrium strategy) nor with bounded running time.
- If $n>2k+3t$, then mediators can be implemented using cheap talk if there is a punishment strategy (and utilities are known) in finite expected running time that does not depend on the utilities.
- If $n \leq 2k+3t$ then mediators cannot, in general, be implemented, even if there is a punishment strategy and utilities are known.
- If $n > 2k+2t$ and there are broadcast channels then, for all ε, mediators can be ε-implemented (intuitively, there is an implementation where players get utility within ε of what they could get by deviating) using cheap talk, with bounded expected running time that does not depend on the utilities.
- If $n \leq 2k+2t$, then mediators cannot, in general, be ε-implemented, even with broadcast channels. Moreover, even assuming cryptography and polynomially bounded players, the expected running time of an implementation depends on the utility functions of the players and ε.
- If $n > k+3t$, then, assuming cryptography and polynomially bounded players, mediators can be ε-implemented using cheap talk, but if $n \leq 2k+2t$, then the running time depends on the utilities in the game and ε.
- If $n \leq k+3t$, then even assuming cryptography, polynomially bounded players, and a $(k+t)$-punishment strategy, mediators cannot, in general, be ε-implemented using cheap talk.
- If $n > k+t$, then, assuming cryptography, polynomially bounded players, and a public-key infrastructure (PKI), we can ε-implement a mediator.

The proof of these results makes heavy use of techniques from computer science. All the possibility results showing that mediators can be implemented use techniques from secure multiparty computation. The results showing that if $n \leq 3k + 3t$, then we cannot implement a mediator without knowing utilities, and that, even if utilities are known, a punishment strategy is required, use the fact that Byzantine agreement cannot be reached if $t < n / 3$; the impossibility result for $n \leq 2k + 3t$ also uses a variant of Byzantine agreement.
A related line of work considers implementing mediators assuming stronger primitives (which cannot be implemented in computer networks); see Izmalkov, Micali and Lepinski (2005) and Lepinski et al. (2004) for details.

6 Other topics

There are many more areas of interaction between computer science than I have indicated in this brief survey. I briefly mention a few others here.

6.1 Interactive epistemology

Since the publication of Aumann's (1976) seminal paper, there has been a great deal of activity in trying to understand the role of knowledge in games, and providing epistemic analyses of solution concepts; see Battigalli and Bonanno (1999) for a survey. In computer science, there has been a parallel literature applying epistemic logic to reason about distributed computation. One focus of this work has been on characterizing the level of knowledge needed to solve certain problems. For example, to achieve Byzantine agreement common knowledge among the nonfaulty agents of an initial value is necessary and sufficient. More generally, in a precise sense, common knowledge is necessary and sufficient for coordination. Another focus has been on defining logics that capture the reasoning of resource-bounded agents. A number of approaches have been considered. Perhaps the most common considers logics for reasoning about *awareness*, where an agent may not be aware of certain concepts, and can know something only if he is aware of it. This topic has been explored in both computer science and game theory; see Dekel, Lipman and Rustichini (1998), Fagin and Halpern (1988), Halpern (2001), Halpern and Rêgo (2007), Heifetz, Meier and Schipper (2006), and Modica and Rustichini (1994; 1999) for some of the work in this active area. Another approach, so far considered only by computer scientists, involves *algorithmic knowledge*, which takes seriously the assumption that agents must explicitly compute what they know. See Fagin et al. (1995) for an overview of the work in epistemic logic in computer science.

6.2 Network growth

If we view networks as being built by selfish players (who decide whether or not to build links), what will the resulting network look like? How does the growth of the network affect its functionality? For example, how easily will influence spread through the network? How easy is it to route traffic? See Fabrikant et al. (2003) and Kempe, Kleinberg and Tardos (2003) for some recent computer science work in this burgeoning area.

6.3 Efficient representation of games

Game theory has typically focused on ‘small’ games, often two- or three-player games, that are easy to describe, such as Prisoner's Dilemma, in order to understand subtleties regarding basic issues such as rationality. To the extent that game theory is used to tackle larger, more practical problems, it will become important to find efficient techniques for describing and analysing games. By way of analogy, $2^n - 1$ numbers are needed to describe a probability distribution on a space characterized by *n* binary random variables. For *n*=100 (not an unreasonable number in practical situations), it is impossible to write down the probability distribution in the obvious way, let alone do computations with it. The same issues will surely arise in large games. Computer scientists use graphical approaches, such as *Bayesian networks* and *Markov networks* (Pearl, 1988), for representing and manipulating probability measures on large spaces. Similar techniques seem applicable to games; see, for example, Kearns, Littman and Singh (2001), Koller and Milch (2001), and La Mura (2000) for specific approaches, and Kearns (2007) for a recent overview. Note that representation is also an issue when we consider the complexity of problems such as computing Nash or correlated equilibria. The complexity of a problem is a function of the size of the input, and the size of the input (which in this case is a description of the game) depends on how the input is represented.

6.4 Learning in games

There has been a great deal of work in both computer science and game theory on learning to play well in different settings (see Fudenberg and Levine, 1998, for an overview of the work in game theory). One line of research in computer science has involved learning to play optimally in a reinforcement-learning setting, where an agent interacts with an unknown (but fixed) environment. The agent then faces a fundamental tradeoff between *exploration* and *exploitation*. The question is how long it takes to learn to play well (to get a reward within some fixed ε of optimal); see Brafman and Tennenholtz (2002) and Kearns and Singh (1998) for the current state of the art. A related question is efficiently finding a strategy minimizes *regret* – that is, finding a strategy that is guaranteed to do not much worse than the best strategy would have done in hindsight (that is, even knowing what the opponent would have done). See Blum and Mansour (2007) for a recent overview of work on this problem.

See Also

- computation of general equilibria
- computational methods in econometrics
- computing in mechanism design
- data mining
- electronic commerce
- epistemic game theory: an overview
- epistemic game theory: beliefs and types

- mathematics of networks
- mechanism design (new developments)
- rationality, bounded
- voting paradoxes

The work for this article was supported in part by NSF under grants CTC-0208535 and ITR-0325453, by ONR under grant N00014-02-1-0455, by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grants N00014-01-1-0795 and N00014-04-1-0725, and by AFOSR under grant F49620-02-1-0101. Thanks to Larry Blume, Christos Papadimitriou, Ilya Segal, Éva Tardos, and Moshe Tennenholtz for useful comments.

Bibliography

Abraham, I., Dolev, D., Gonen, R. and Halpern, J. 2006. Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation. In *Proceedings of the 25th ACM Symposium on Principles of Distributed Computing*, 53–62.

Abraham, I., Dolev, D. and Halpern, J. 2007. Lower bounds on implementing robust and resilient mediators. Available at http://arxiv.org/abs/0704.3646, accessed 19 June 2007.

Adar, E. and Huberman, B. 2000. Free riding on Gnutella. *First Monday* 5(10).

Alvisi, L., Ayer, A.S., Clement, A., Dahlin, M., Martin, J.P. and Porth, C. 2005. BAR fault tolerance for cooperative services. In *Proceedings of the 20th ACM Symposium on Operating Systems Principles (SOSP 2005)*, 45–58.

Archer, A. and Tardos, É. 2001. Truthful mechanisms for one-parameter agents. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, 482–91.

Archer, A. and Tardos, É. 2002. Frugal path mechanisms. In *Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms*, 991–9.

Aumann, R. 1959. Acceptable points in general cooperative *n*-person games. In *Contributions to the Theory of Games IV*, ed. A. Tucker and R. Luce. Princeton, NJ: Princeton University Press, pp. 287–324.

Aumann, R.J. 1976. Agreeing to disagree. *Annals of Statistics* 4, 1236–9.

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55, 1–18.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Babaoglu, O. 1987. On the reliability of consensus-based fault-tolerant distributed computing systems. *ACM Translation on Computer Systems* 5, 394–416.

Barany, I. 1992. Fair distribution protocols or how the players replace fortune. *Mathematics of Operations Research* 17, 327–40.

Battigalli, P. and Bonanno, G. 1999. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics* 53, 149–225.

Ben-Or, M. 1983. Another advantage of free choice: completely asynchronous agreement protocols. In *Proceedings of the Second ACM Symposium on Principles of Distributed Computing*, 27–30.

Ben-Porath, E. 2003. Cheap talk in games with incomplete information. *Journal of Economic Theory* 108, 45–71.

Blum, A. and Mansour, Y. 2007. Learning, regret minimization, and equilibria. In *Algorithmic Game Theory*, ed. N. Nisan, T. Roughgarden, É. Tardos, and V. Vazirani. Cambridge: Cambridge University Press.

Blum, B., Shelton, C.R. and Koller, D. 2003. A continuation method for Nash equilibria in structured games. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI '03)*, 757–64.

Brafman, R.I. and Tennenholtz, M. 2002. R-MAX: a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3, 213–31.

Chen, X. and Deng, X. 2006. Settling the complexity of 2-player Nash equilibrium. In *Proceedings of the 47th IEEE Symposium on Foundations of Computer Science*, 261–72.

Chor, B. and Dwork, C. 1989. Randomization in Byzantine agreement. In *Advances in Computing Research 5: Randomness and Computation*, ed. S. Micali. Greenwich, CT: JAI Press, pp. 443–97.

Chu, F. and Halpern, J.Y. 2001. A decision-theoretic approach to reliable message delivery. *Distributed Computing* 14, 359–89.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11, 17–33.

Conitzer, V. and Sandholm, T. 2003. Complexity results about Nash equilibria. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI '03)*, 765–71.

Conitzer, V., Sandholm, T. and Lang, J. 2007. When are elections with few candidates hard to manipulate? *Journal of the ACM* 54(3), Article 14.

Cramton, P., Shoham, Y. and Steinberg, R. 2006. *Combinatorial Auctions*. Cambridge, MA: MIT Press.

Daskalis, C., Goldberg, P.W. and Papadimitriou, C.H. 2006. The complexity of computing a Nash equilibrium. In *Proceedings of the 38th ACM Symposium on Theory of Computing*, 71–8.

Dekel, E., Lipman, B. and Rustichini, A. 1998. Standard state-space models preclude unawareness. *Econometrica* 66, 159–73.

Eliaz, K. 2002. Fault-tolerant implementation. *Review of Economic Studies* 69, 589–610.

Fabrikant, A., Luthra, A., Maneva, E., Papadimitriou, C.H. and Shenker, S. 2003. On a network creation game. In *Proceedings of the 22nd ACM Symposium on Principles of Distributed Computing*, 347–51.

Fagin, R. and Halpern, J.Y. 1988. Belief, awareness, and limited reasoning. *Artificial Intelligence* 34, 39–76.

Fagin, R., Halpern, J.Y., Moses, Y. and Vardi, M.Y. 1995. *Reasoning about Knowledge*. Cambridge, MA: MIT Press. A slightly revised paperback version was published in 2003.

Feigenbaum, J., Papadimitriou, C. and Shenker, S. 2000. Sharing the cost of muliticast transmissions (preliminary version). In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 218–27.

Fischer, M.J. 1983. The consensus problem in unreliable distributed systems. In *Foundations of Computation Theory*, vol. 185, ed. M. Karpinski. Lecture Notes in Computer Science. Berlin/New York: Springer, pp. 127–40.

Fischer, M.J., Lynch, N.A. and Paterson, M.S. 1985. Impossibility of distributed consensus with one faulty processor. *Journal of the ACM* 32, 374–82.

Forges, F. 1990. Universal mechanisms. *Econometrica* 58, 1341–64.

Fudenberg, D. and Levine, D. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Gerardi, D. 2004. Unmediated communication in games with complete and incomplete information. *Journal of Economic Theory* 114, 104–31.

Gibbard, A. 1973. Manipulation of voting schemes. *Econometrica* 41, 587–602.

Gilboa, I. and Zemel, E. 1989. Nash and correlated equilibrium: some complexity considerations. *Games and Economic Behavior* 1, 80–93.

Goldreich, O., Micali, S. and Wigderson, A. 1987. How to play any mental game. In *Proceedings of the 19th ACM Symposium on Theory of Computing*, 218–29.

Govindan, S. and Wilson, R. 2003. A global Newton method to compute Nash equilibria. *Journal of Economic Theory* 110, 65–86.

Groves, T. 1973. Incentives in teams. *Econometrica* 41, 617–31.

Halldórsson, M.M., Halpern, J.Y., Li, L. and Mirrokni, V. 2004. On spectrum sharing games. In *Proceedings of the 23rd ACM Symposium on Principles of Distributed Computing*, 107–14.

Halpern, J.Y. 2001. Alternative semantics for unawareness. *Games and Economic Behavior* 37, 321–39.

Halpern, J.Y. 2003. A computer scientist looks at game theory. *Games and Economic Behavior* 45, 114–32.

Halpern, J.Y. and Rêgo, L.C. 2007. Reasoning about knowledge of unawareness. *Games and Economic Behavior*. Also available at http://arxiv.org/abs/cs.LO/0603020, accessed 24 June 2007.

Halpern, J.Y. and Teague, V. 2004. Rational secret sharing and multiparty computation: extended abstract. In *Proceedings of the 36th ACM Symposium on Theory of Computing*, 623–32.

Heifetz, A., Meier, M. and Schipper, B. 2006. Interactive unawareness. *Journal of Economic Theory* 130, 78–94.

Heller, Y. 2005. A minority-proof cheap-talk protocol. Unpublished manuscript.

Hopcroft, J.E. and Ullman, J.D. 1979. *Introduction to Automata Theory, Languages and Computation*. New York: Addison-Wesley.

Hurwicz, L. 1977. On the dimensional requirements of informationally decentralized Pareto satisfactory processes. In *Studies in Resource AllocationProcesses*, ed. K.J. Arrow and L. Hurwicz. New York: Cambridge University Press, pp. 413–24.

Izmalkov, S., Micali, S. and Lepinski, M. 2005. Rational secure computation and ideal mechanism design. In *Proceedings of the 46th IEEE Symposium Foundations of Computer Science*, 585–95.

Kearns, M. 2007. Graphical games. In *Algorithmic Game Theory*, ed. N. Nisan, T. Roughgarden, É.

Tardos and V. Vazirani. Cambridge: Cambridge University Press.

Kearns, M., Littman, M.L. and Singh, S.P. 2001. Graphical models for game theory. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI 2001)*, 253–60.

Kearns, M.J. and Reiter, M.K. 2005. *Proceedings of the Sixth ACM Conference on Electronic Commerce* (*EC '05*). New York: ACM. Table of contents available at http://www.informatik.uni-trier.de/~ley/db/conf/sigecom/sigecom2005.html, accessed 19 June 2007.

Kearns, M. and Singh, S.P. 1998. Near-optimal reinforcement learning in polynomial time. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 260–8.

Kempe, D., Kleinberg, J. and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–46.

Kfir-Dahav, N.E., Monderer, D. and Tennenholtz, M. 2000. Mechanism design for resource-bounded agents. In *International Conference on Multiagent Systems*, 309–16.

Koller, D. and Milch, B. 2001. Structured models for multiagent interactions. In *Theoretical Aspects of Rationality and Knowledge: Proceedings of the Eighth Conference (TARK 2001)*, 233–48.

Koutsoupias, E. and Papadimitriou, C.H. 1999. Worst-case equilibria. In *Proceedings of the 16th Conference on Theoretical Aspects of Computer Science*, vol. 1563, Lecture Notes in Computer Science. Berlin: Springer, pp. 404–13.

Kushilevitz, E. and Nisan, N. 1997. *Communication Complexity*. Cambridge: Cambridge University Press.

La Mura, P. 2000. Game networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, 335–42.

Lehmann, D., Müller, R. and Sandholm, T. 2006. The winner determination problem. In *Combinatorial Auctions*, ed. P. Cramton, Y. Shoham, and R. Steinberg. Cambridge, MA: MIT Press.

Lemke, C.E. and Howson, J.J.T. 1964. Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics* 12, 413–23.

Lepinski, M., Micali, S., Peikert, C. and Shelat, A. 2004. Completely fair SFE and coalition-safe cheap talk. In *Proceedings of the 23rd ACM Symposium Principles of Distributed Computing*, 1–10.

Linial, N. 1994. Games computers play: Game-theoretic aspects of computing. In *Handbook of Game Theory*, vol. 2, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Modica, S. and Rustichini, A. 1994. Awareness and partitional information structures. *Theory and Decision* 37, 107–24.

Modica, S. and Rustichini, A. 1999. Unawareness and partitional information structures. *Games and Economic Behavior* 27, 265–98.

Monderer, D. and Tennenholtz, M. 1999a. Distributed games. *Games and Economic Behavior* 28, 55–72.

Monderer, D. and Tennenholtz, M. 1999b. Distributed Games: From Mechanisms to Protocols. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI '99)*, 32–7.

Moreno, D. and Wooders, J. 1996. Coalition-proof equilibrium. *Games and Economic Behavior* 17, 80–112.

Mount, K. and Reiter, S. 1974. The informational size of message spaces. *Journal of Economic Theory* 8, 161–92.

Nash, J. 1950. Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences* 36, 48–9.

Neyman, A. 1985. Bounded complexity justifies cooperation in finitely repeated Prisoner's Dilemma. *Economic Letters* 19, 227–9.

Nisan, N. 2006. Bidding languages for combinatorial auctions. In *Combinatorial Auctions*. Cambridge, MA: MIT Press.

Nisan, N. and Ronen, A. 2000. Computationally feasible VCG mechanisms. In *Proceedings of the Second ACM Conference on Electronic Commerce (EC '00)*, 242–52.

Nisan, N. and Ronen, A. 2001. Algorithmic mechanism design. *Games and Economic Behavior* 35, 166–96.

Nisan, N., Roughgarden, T., Tardos, É. and Vazirani, V. 2007. *Algorithmic Game Theory*. Cambridge, UK: Cambridge University Press.

Papadimitriou, C.H. 1994a. *Computational Complexity*. Reading, MA: Adison-Wesley.

Papadimitriou, C.H. 1994b. On the complexity of the parity argument and other inefficient proofs of

existence. *Journal of Computer and System Sciences* 48, 498–532.

Papadimitriou, C.H. 2001. Algorithms, games, and the internet. In *Proceedings of the 33rd ACM Symposium on Theory of Computing*. 749–53.

Papadimitriou, C.H. 2005. Computing correlated equilibria in multiplayer games. In *Proceedings of the 37th ACM Symposium on Theory of Computing*, 49–56.

Papadimitriou, C.H. 2007. The complexity of finding Nash equilibria. In *Algorithmic Game Theory*, ed. N. Nisan, T. Roughgarden, É. Tardos and V. Vazirani. Cambridge: Cambridge University Press.

Papadimitriou, C.H. and Roughgarden, T. 2005. Computing equilibria in multi-player games. In *Proceedings of the 16th ACM–SIAM Symposium on Discrete Algorithms*, 82–91.

Papadimitriou, C.H. and Yannakakis, M. 1994. On complexity as bounded rationality. In *Proceedings of the 26th ACM Symposium on Theory of Computing*, 726–33.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.

Pease, M., Shostak, R. and Lamport, L. 1980. Reaching agreement in the presence of faults. *Journal of the ACM* 27, 228–34.

Porter, R., Nudelman, E. and Shoham, Y. 2004. Simple search methods for finding a Nash equilibrium. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI '04)*, 664–9.

Rabin, M.O. 1983. Randomized Byzantine generals. In *Proceedings of the 24th IEEE Symposium on Foundations of Computer Science*, 403–9.

Roughgarden, T. and Tardos, É. 2002. How bad is selfish routing? *Journal of the ACM* 49, 236–59.

Rubinstein, A. 1986. Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory* 39, 83–96.

Rubinstein, A. 1998. *Modeling Bounded Rationality*. Cambridge, MA: MIT Press.

Sandholm, T. and Yokoo, M. 2003. *The Second International Joint Conference on Autonomous Agents and Multiagent Systems* (*AAMAS 2003*). Table of contents available at http://www.informatik.uni-trier.de/~ley/db/conf/atal/aamas2003.html, accessed 25 June 2007.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Segal, I. 2006. Communication in economic mechanisms. In *Advances in Economics and Econometrics: Theory and Application, Ninth World Congress (Econometric Society Monographs*, ed. R. Blundell, W. K. Newey and T. Persson. Cambridge: Cambridge University Press, pp. 222–68.

Shamir, A., Rivest, R.L. and Adelman, L. 1981. Mental poker. In *The Mathematical Gardner*, ed. D.A. Klarner. Boston, MA: Prindle, Weber, and Schmidt, pp. 37–43.

Tennenholtz, M. 2002. Competitive safety analysis: robust decision-making in multi-agent systems. *Journal of A.I. Research* 17, 363–78.

Urbano, A. and Vila, J.E. 2002. Computational complexity and communication: coordination in two-player games. *Econometrica* 70, 1893–927.

Urbano, A. and Vila, J.E. 2004. Computationally restricted unmediated talk under incomplete information. *Economic Theory* 23, 283–320.

Vetta, A. 2002. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science*, 416–25.

Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16, 8–37.

Yao, A. 1982. Protocols for secure computation (extended abstract). In *Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science*, 160–4.

How to cite this article

Halpern, Joseph Y. "computer science and game theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000566> doi:10.1057/9780230226203.0287

The New Palgrave Dictionary of Economics Online

computing in mechanism design

Tuomas Sandholm

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Computational issues are important in mechanism design, but have received insufficient research interest. This article briefly reviews some of the key ideas. I discuss computing by the *centre*, such as an auction server or vote aggregator, and computing by the *agents*, be they human or software. Limited computing hinders mechanism design in several ways, and presents deep strategic interactions between computing and incentives. On the bright side, novel algorithms and increasing computing power have enabled better mechanisms. Perhaps most interestingly, with computationally limited agents, one can implement mechanisms that would not be implementable among computationally unlimited agents.

Keywords

algorithmic mechanism design; automated mechanism design; Borda voting rule; bounded rationality; combinatorial auctions; complexity theory; computing by the centre; computing in mechanism design; deliberation equilibrium; elicitor; *ex post* equilibrium; expressive commerce; fielded combinatorial auctions; fielded expressive auctions; Gibbard–Satterthwaite th; incentive compatibility; maximin voting rule; non-truth-promoting mechanism; mechanism design; performance profile tree; plurality voting rule; preference determination; preference elicitation; pull–push mechanism; revelation principle; strategic computing; tree search; Vickrey auction; Vickrey–Clarke–Groves (VCG) mechanism

Article

1 Introduction

Computational issues in mechanism design are important, but have received insufficient research interest until recently. Limited computing hinders mechanism design in several ways, and presents deep strategic interactions between computing and incentives. On the bright side, novel algorithms and increasing computing power have enabled better mechanisms. Perhaps most interestingly, limited computing of the

agents can be used as a tool to implement mechanisms that would not be implementable among computationally unlimited agents. This article briefly reviews some of the key ideas, with the goal of alerting the reader to the importance of these issues and hopefully spurring future research.
I will discuss computing by the *centre*, such as an auction server or vote aggregator, in Section 2. Then, in Section 3, I will address the *agents*' computing, be they human or software.

2 Computing by the centre

Computing by the centre plays significant roles in mechanism design. In the following three subsections I will review three prominent directions.

2.1 Executing expressive mechanisms

As algorithms have advanced drastically and computing power has increased, it has become feasible to field mechanisms that were previously impractical. The most famous example is a *combinatorial auction* (*CA*). In a CA, there are multiple distinguishable items for sale, and the bidders can submit bids on self-selected packages of the items. (Sometimes each bidder is also allowed to submit exclusivity constraints of different forms among his bids.) This increase in the expressiveness of the bids drastically reduces the strategic complexity that bidders face. For one, it removes the exposure problems that bidders face when they have preferences over packages but in traditional auctions are allowed to submit bids on individual items only.
CAs shift the computational burden from the bidders to the centre. There is an associated gain because the centre has all the information in hand to optimize while in traditional auctions the bidders only have estimated projected (probabilistic) information about how others will bid. Thus CAs yield more efficient allocations.
On the downside, the centre's task of determining the winners in a CA (deciding which bids to accept so as to maximize the sum of the accepted bids' prices subject to not selling any item to more than one bid) is a complex combinatorial optimization problem, even without exclusivity constraints among bids. Three main approaches have been studied for solving it.

1. 1. *Optimal winner determination using some form of tree search*. For a review, see Sandholm (2006). The advantage is that the bidding language is not restricted and the optimal solution is found. The downside is that no optimal winner determination algorithm can run in polynomial time in the size of the problem instance in the worst case, because the problem is $\mathcal{NP}$-complete (Rothkopf, Pek ě c and Harstad, 1998). ($\mathcal{NP}$-complete problems are problems for which the fastest known algorithms take exponential time in the size of the problem instance in the worst case. $\mathcal{P}$ is the class of easy problems solvable in polynomial time. The statement of winner determination not being solvable in polynomial time in the worst case relies on the usual assumption $\mathcal{P} \neq \mathcal{NP}$. This is an open question in complexity theory, but is widely believed to be true. If false, that would have sweeping implications throughout computer science.)
2. 2. *Approximate winner determination*. The advantage is that many approximation algorithms run in polynomial time in the size of the instance even in the worst case. For reviews of such

algorithms, see Sandholm (2002a) and Lehmann, Müller and Sandholm (2006). (Other suboptimal algorithms do not have such time guarantees, such as local search, stochastic local search, simulated annealing, genetic algorithms and tabu search.) The downside is that the solution is sometimes far from optimal: no such algorithm can always find a solution that is within a factor

$$\min\left\{ \#\,bids^{1-\epsilon},\ \#\,items^{\frac{1}{2}-\epsilon} \right\} \tag{1}$$

of optimal (Sandholm, 2002a). (This assumes $ZPP \neq NP$. It is widely believed that these two complexity classes are indeed unequal.) For example, with just nine items for sale, no such algorithm can extract even 33 per cent of the available revenue from the bids in the worst case. With 81 items, that drops to 11 per cent.

3. 3. *Restricting the bidding language* so much that optimal (within the restricted language) winner determination can be conducted in worst-case polynomial time. For a review, see Müller (2006). For example, if each package bid is only allowed to include at most two items, then winners can be determined in worst-case polynomial time (Rothkopf, Pek ě c and Harstad, 1998). The downside is that bidders have to shoehorn their preferences into a restricted bidding language; this gives rise to similar problems as in non-combinatorial mechanisms for multi-item auctions: exposure problems, need to speculate how others will bid, inefficient allocation, and so on.

Truthful bidding can be made a dominant strategy by applying the *Vickrey–Clarke–Groves* (*VCG*) *mechanism* to a CA. Such incentive compatibility removes strategic complexity of the bidders. The mechanism works as follows. The optimal allocation is used, but the bidders do not pay their winning bids. Instead each bidder pays the amount of value he takes away from the others by taking some of the items. This value is measured as the difference between the others' winning bids' prices and what the others' winning bids' prices would have been had the agent not submitted any bids. This mechanism can be executed by determining the winners once overall, and once for each agent removed in turn. (This may be accomplishable with less computing. For example, in certain network auctions it can be done in the same asymptotic complexity as one winner determination – Hershberger and Suri, 2001.)

Very few canonical CAs have found their way to practice. However, auctions with richer bid expressiveness forms (that are more natural in the given application and more concise) and that support expressiveness also by the bid taker have made a major breakthrough into practice (Sandholm, 2007; Bichler et al., 2006). This is sometimes called *expressive commerce* to distinguish it from vanilla CAs. The widest area of application is currently industrial sourcing. Tens of billions of dollars worth of materials, transportation and services are being sourced annually using such mechanisms, yielding billions of dollars in efficiency improvements. The bidders' expressiveness forms include different forms of flexible package bids, conditional discounts, discount schedules, side constraints (such as capacity

constraints), and often hundreds of cost drivers (for example, fixed costs, variable costs, trans-shipment costs and costs associated with changes). The item specifications can also be left partially open, and the bidders can specify some of the item attributes (delivery date, insurance terms, and so on). in alternate ways. The bid taker also specifies preferences and constraints. Winner determination then not only decides who wins what, but also automatically configures the items. In some of these events it also configures the supply chain several levels deep as a side effect. On the high end, such an auction can have tens of thousands of items (multiple units of each), millions of bids, and hundreds of thousands of side constraints. Expressive mechanisms have also been designed for settings beyond auctions, such as combinatorial exchanges, charity donations and settings with externalities.

Basically all of the fielded expressive auctions use the simple pay-your-winning-bids pricing rule. There are numerous important reasons why few, if any, use the VCG mechanism. It can lead to low revenue. It is vulnerable to collusion. Bidders would not tell the truth because they do not want to reveal their cost structures, which the auctioneer could exploit the next time the auction is conducted, and so on (Sandholm, 2000; Rothkopf, 2007).

Basically all of the fielded expressive auctions use tree search for winner determination. In practice, modern tree search algorithms for the problem scale to the large and winners can be determined optimally. If winner determination were not done optimally in a CA, the VCG mechanism can lose its truth-dominance property (Sandholm, 2002b). In fact, any truthful suboptimal VCG-based mechanism for CAs is unreasonable in the sense that it sometimes does not allocate an item to a bidder even if he is the only bidder whose bids assign non-zero value to that item (Nisan and Ronen, 2000).

2.2 Algorithmic mechanism design

Motivated by the worry that some instances of *NP*-hard problems may not be solvable within reasonable time, a common research direction in theory of computing is approximation algorithms. They trade off solution quality for a guarantee that even in the worst case, the algorithm runs in polynomial time in the size of the input.

Analogously, Nisan and Ronen (2001) proposed *algorithmic mechanism design*: designing approximately optimal mechanisms that take the centre a polynomial number of computing steps even in the worst case. However, this is more difficult than designing approximately optimal algorithms because the mechanism has to motivate the agents to tell the truth.

Lehmann, O'Callaghan and Shoham (2002) studied this for CAs with single-minded bidders (each bidder being only interested in one specific package of items). They present a fast greedy algorithm that guarantees a solution within a factor $\sqrt{\#\ items}$ of optimal. They show that the algorithm is not incentive compatible with VCG pricing, but is with their custom pricing scheme. They also identify sufficient conditions for any (approximate) mechanism to be incentive compatible (see also Kfir-Dahav, Monderer and Tennenholtz, 2000). There has been substantial follow-on work on subclasses of single-minded CAs.

Lavi and Swamy (2005) developed a technique for a range of packing problems with which any *k*-approximation algorithm (that is, algorithm that guarantees that the solution is within a factor *k* of optimal) that also bounds the integrality gap of the linear programming (LP) relaxation of the problem by *k* can be used to construct a *k*-approximation mechanism. The LP solution, scaled down by *k*, can be represented as a convex combination of integer solutions, and viewing this convex combination as

specifying a probability distribution over integer solutions begets a VCG-based randomized mechanism that is truthful in expectation. For CAs with general valuations, this yields an $O(\sqrt{\#\,items})$-approximate mechanism.

In a different direction, several mechanisms have been proposed where the agents can help the centre find better outcomes. This is done either by giving the agents the information to do the centre's computing (Banks, Ledyard and Porter, 1989; Land, Powell and Steinberg, 2006; Parkes and Shneidman, 2004), or by allowing the agents to change what they told the mechanism based on the mechanism's output and potentially also based on what other agents told the mechanism (Nisan and Ronen, 2000). In VCG-based mechanisms, an agent benefits from lying only if the lie causes the mechanism to find an outcome that is better overall.

2.3 Automated mechanism design

Conitzer and Sandholm (2002) proposed the idea of *automated mechanism design*: having a computer, rather than a human, design the mechanism. Because human effort is eliminated, this enables custom design of mechanisms for every setting. The setting can be described by the agents' (discretized) type spaces, the designer's prior over types, the desired notion of incentive compatibility (for example, dominant strategies vs. Bayes–Nash implementation), the desired notion of participation constraints (for example, *ex interim*, *ex post* or none), whether payments are allowed, and whether the mechanism is allowed to use randomization.) This can yield better mechanisms for previously studied settings because the mechanism is designed for the specific setting rather than a class of settings. It can also be used for settings not previously studied in mechanism design.

For almost all natural (linear) objectives, all variants of the design problem are $\mathcal{NP}$-complete if the mechanism is not allowed to use randomization, but randomized mechanisms can be constructed for all these settings in polynomial time using linear programming. Custom algorithms have been developed for some problems in each of these two categories. (Even the latter category warrants research. While the linear programme is polynomial in the size of the input, the input itself can be exponential in the number of agents.) Structured representations of the problem can also make the design process drastically faster.

Beyond the general setting, automated mechanism design has been applied to specific settings, such as creating revenue-maximizing CAs (without the need to discretize types)–Likhodedov and Sandholm, 2005 (a recognized problem that eludes analytical characterization; even the two-item case is open), reputation systems (Jurca and Faltings, 2006), safe exchange mechanisms (Sandholm and Ferrandon, 2000), and supply chain settings (Vorobeychik, Kiekintveld and Wellman, 2006). Automated mechanism design software has recently also been adopted by several mechanism design theoreticians to speed up their research.

It turns out that even *multistage mechanisms* can be designed automatically (Sandholm, Conitzer and Boutilier, 2007). Furthermore, automated mechanism design has been applied to the design of *online mechanisms* (Hajiaghayi, Kleinberg and Sandholm, 2007), that is, mechanisms that execute while the world changes – for example, agents enter and exit the system.

3 Computing by the agents

I will now move to discussing computing by the agents.

3.1 Mechanisms that are hard to manipulate

This section demonstrates that one can use the fact that agents are computationally limited to achieve things that are not achievable via any mechanism among perfectly rational agents.

A seminal negative result, the *Gibbard–Satterthwaite th*, states that if there are three or more candidates, then in any non-dictatorial voting scheme there are candidate rankings of the other voters, and preferences of the agent, under which the agent is better off voting manipulatively than truthfully. One avenue around this impossibility is to construct desirable general non-dictatorial voting protocols under which *finding* a beneficial manipulation is prohibitively hard computationally.

There are two natural alternative goals of manipulation. In *constructive manipulation*, the manipulator tries to find an order of candidates that he can reveal so that his favourite candidate wins. In *destructive manipulation*, the manipulator tries to find an order of candidates that he can reveal so that his hated candidate does not win. These are special cases of the utility-theoretic notion of improving one's utility, so the hardness results, discussed below, carry over to the usual utility-theoretic setting.

Unfortunately, finding a constructive manipulation is easy (in $\mathcal{P}$) for the *plurality*, *Borda* and *maximin* voting rules (Bartholdi, Tovey and Trick, 1989), which are commonly used. On the bright side, constructive manipulation of the *single transferable vote* (*STV*) protocol is $\mathcal{NP}$-hard (Bartholdi and Orlin, 1991) (as is manipulation of the *second order Copeland* protocol (Bartholdi, Tovey and Trick, 1989), but that hardness is driven solely by the tie-breaking rule). Even better, there is a systematic methodology for slightly tweaking voting protocols that are easy to manipulate, so that they become hard to manipulate (Conitzer and Sandholm, 2003). Specifically, before the original protocol is executed, one pairwise elimination round is executed among the candidates, and only the winning candidates survive to the original protocol. This makes the protocols $\mathcal{NP}$-hard, $\#\mathcal{P}$-hard ($\#\mathcal{P}$-hard problems are at least as hard as counting the number of solutions to a problem in $\mathcal{P}$), or even $\mathcal{PSPACE}$-hard ($\mathcal{PSPACE}$-hard problems are at least as hard as any problem that can be solved using a polynomial amount of memory) to manipulate constructively, depending on whether the schedule of the pre-round is determined before the votes are collected, randomly after the votes are collected, or the scheduling and the vote collecting are carefully interleaved, respectively.

All of the hardness results of the previous paragraph rely on both the number of voters and the number of candidates growing. The number of candidates can be large in some domains, for example when voting over task or resource allocations. However, in other elections – such as presidential elections – the number of candidates is small. If the number of candidates is a constant, both constructive and destructive manipulation are easy (in $\mathcal{P}$), regardless of the number of voters (Conitzer, Sandholm and Lang, 2007). This holds even if the voters are weighted, or if a coalition of voters tries to manipulate. On the bright side, when a coalition of weighted voters tries to manipulate, complexity can arise even for a constant number of candidates: see Tables 1 and 2. Another lesson from that table is that randomizing over instantiations of the mechanism (such as schedules of a *cup*) can be used to make manipulation hard.

Complexity of constructive weighted coalitional manipulation

Number of candidates:	2	3	4, 5, 6	≥ 7
Borda	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Veto	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
STV	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Plurality with runoff	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Copeland	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Maximin	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Randomized cup	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{NP}$-complete
Cup	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$
Plurality	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$

Complexity of destructive weighted coalitional manipulation

Number of candidates:	2	≥ 3
STV	$\mathcal{P}$	$\mathcal{NP}$-complete
Plurality with runoff	$\mathcal{P}$	$\mathcal{NP}$-complete
Randomized cup	$\mathcal{P}$	?
Borda	$\mathcal{P}$	$\mathcal{P}$
Veto	$\mathcal{P}$	$\mathcal{P}$
Copeland	$\mathcal{P}$	$\mathcal{P}$
Maximin	$\mathcal{P}$	$\mathcal{P}$
Cup	$\mathcal{P}$	$\mathcal{P}$
Plurality	$\mathcal{P}$	$\mathcal{P}$

Source: Conitzer, Sandholm and Lang (2007).

As usual in computer science, all the results mentioned above are worst-case hardness. Unfortunately, under weak assumptions on the preference distribution and voting rule, most instances of any voting rule are easy to manipulate (Conitzer and Sandholm, 2006).

All of the hardness results discussed above hold even if the manipulators know the non-manipulators' votes exactly. Under weak assumptions, if weighted coalitional manipulation with complete information about the others' votes is hard in some voting protocol, then individual unweighted manipulation is hard when there is uncertainty about the others' votes (Conitzer, Sandholm and Lang, 2007).

3.2 Non-truth-promoting mechanisms

A challenging issue is that even if it is prohibitively hard to find a beneficial manipulation, the agents might not tell the truth. For example, an agent might take a chance that he will do better with a lie. The following result shows that, nevertheless, mechanism design can be improved by making the agents face

complexity. (This is one reason why computational issues can render the *revelation principle* inapplicable. One of the things the principle says is that for any non-truth-promoting mechanism it is possible to construct an incentive-compatible mechanism that is at least as good. The theorem below challenges this.)

Theorem 1: (Conitzer and Sandholm, 2004) *Suppose the centre is trying to maximize social welfare, and neither payments nor randomization is allowed. Then, even with just two agents (one of whom does not even report a type, so dominant strategy implementation and Bayes–Nash implementation coincide), there exists a family of preference aggregation settings such that:*

- *the execution of any optimal incentive-compatible mechanism is* $\mathcal{NP}$*-complete for the center, and*
- *there exists a non-incentive-compatible mechanism which* (*1*) *requires the centre to carry out only polynomial computation, and* (*2*) *makes finding any beneficial insincere revelation* $\mathcal{NP}$*-complete for the type-reporting agent. Additionally, if the type-reporting agent manages to find a beneficial insincere revelation, or no beneficial insincere revelation exists, the social welfare of the outcome is identical to the social welfare that would be produced by any optimal incentive-compatible mechanism. Finally, if the type-reporting agent does not manage to find a beneficial insincere revelation where one exists, the* ***social welfare of the outcome is strictly greater than the social welfare that would be produced by any optimal incentive-compatible mechanism****.*

An analogous theorem holds if, instead of counting computational steps, we count calls to a commonly accessible oracle which, when supplied with an agent, that agent's type, and an outcome, returns a utility value for that agent.

3.3 Preference (valuation) determination via computing or information acquisition

In many (auction) settings, even determining one's valuation for an item (or a bundle of items) is complex. For example, when bidding for trucking lanes (tasks), this involves solving two $\mathcal{NP}$-complete local planning problems: the vehicle routing problem with the new lanes of the bundle and the problem without them (Sandholm, 1993). The difference in the costs of those two local plans is the cost (valuation) of taking on the new lanes.

In these types of settings, the *revelation principle* applies only in a trivial way: the agents report their data and optimization models to the centre, and the centre does the computation for them. It stands to reason that in many applications the centre would not want to take on that burden, in which case such extreme direct mechanisms are not an option. Therefore, I will now focus on mechanisms where the agents report valuations to the centre, as in traditional auctions.

Bidders usually have limited computing and time, so they cannot exactly evaluate all (or even any) bundles – at least not without cost. This leads to a host of interesting issues where computing and incentives are intimately intertwined.

For example, in a one-object auction, should a bidder evaluate the object if there is a cost to doing so? It turns out that the Vickrey auction loses its dominant-strategy property: whether or not the bidder should pay the evaluation cost depends on the other bidders' valuations (Sandholm, 2000).

If a bidder has the opportunity to *approximate his valuation to different degrees*, how much computing

time should the bidder spend on refining its valuation? If there are multiple items for sale, how much computing time should the bidder allocate on different bundles? A bidder may even allocate some computing time to evaluate other bidders' valuations so as to be able to bid more strategically; this is called *strategic computing*.

To answer these qsts, Larson and Sandholm (2001) developed a deliberation control method called a *performance profile tree* for projecting how an anytime algorithm (that is, an algorithm that has an answer available at any time, but where the quality of the answer improves the more computing time is allocated to the algorithm) will change the valuation if additional computing is allocated toward refining (or improving) it. This deliberation control method applies to any anytime algorithm. Unlike earlier deliberation control methods for anytime algorithms, the performance profile tree is a *fully normative model of bounded rationality*: it takes into account all the information that an agent can use to make its deliberation control decisions. This is necessary in the game-theoretic context; otherwise a strategic agent could take into account some information that the model does not.

Using this deliberation control method, the auction can be modelled as a game where the agents' strategy spaces include computing actions. At every point, each agent can decide on which bundle to allocate its next step of computing as a function of the agent's computing results so far (and in open-cry auction format also the others' bids observed so far). At every point, the agent can also decide to submit bids. One can then solve this for equilibrium: each agent's (deliberation and bidding) strategy is a best-response to the others' strategies. This is called *deliberation equilibrium*.

This notion, and the performance profile tree, apply not only to computational actions but also to information gathering actions for determining valuations. (In contrast, most of the literature on information acquisition in auctions does not take into account that valuations can be determined to different degrees and that an agent may want to invest effort to determine others' valuations as well – even in private-value settings.)

Table 3 shows in which settings strategic computing can and cannot occur in deliberation equilibrium. This depends on the auction mechanism. Interestingly, it also depends on whether the agent has limited computing (for example, owning a desktop computer that the agent can use until the auction's deadline) or costly computing (for example, being able to buy any amount of supercomputer time where each cycle comes at a cost).

Can strategic computing occur in deliberation equilibrium? The most interesting results are in bold. As a benchmark from classical auction theory, the table also shows whether or not perfectly rational agents, that can determine their valuations instantly without cost, would benefit from considering each others' valuations when deciding how to bid

	Auction mechanism	Speculation by perfectly rational agents?	Strategic computing?
Limited computing	**Costly computing**		
Single item	First price	yes	yes	yes
Dutch	yes	yes	yes	
English	no	**no**	**yes**	
Vickrey	no	**no**	**yes**	
Multiple items	First price	yes	yes	yes

VCG	no	**yes**	**yes**

The notion of deliberation equilibrium can also be used as the basis for designing new mechanisms, which hopefully work well among agents whose computing is costly or limited. Unfortunately, there is an impossibility (Larson and Sandholm, 2005): there exists no mechanism that is *sensitive* (the outcome is affected by each agent's strategy), *preference formation independent* (does not do the computations for the agents; the agents report valuations), *non-misleading* (no agent acts in a way that causes others to believe his true type has zero probability), and *deliberation-proof* (no strategic computing occurs in equilibrium, that is, agents compute only on their own problems). Current work involves designing mechanisms that take part in preference formation in limited ways: for example, agents report their performance profile trees to the centre, which then coordinates the deliberations incrementally as agents report deliberation results. Current research also includes designing mechanisms where strategic computing occurs but its wastefulness is limited.

3.3.1 Preference elicitation by the centre

To reduce the agents' preference determination effort, Conen and Sandholm (2001) proposed a framework where the centre (also known as *elicitor*) explicitly elicits preference information from the agents incrementally on an as-needed basis by posing queries to the agents. The centre thereby builds a model of the agents' preferences, and decides what to ask, and from which agent, based on this model. Usually the process can be terminated with the provably correct outcome while requiring only a small portion of the agents' preferences to be determined. Multistage mechanisms can yield up to exponential savings in preference determination and communication effort the agents need to go through compared to single-stage mechanisms (Conitzer and Sandholm, 2004).

The explicit preference elicitation framework was originally proposed for CAs (but the approach has since been used for other settings as well, such as voting). For general valuations, an exponential number of bits in the number of items for sale has to be communicated in the worst case no matter what queries are used (Nisan and Segal, 2006). However, experimentally only a small fraction of the preference information needs to be elicited before the provably optimal solution is found. Furthermore, for valuations that have certain types of structure, even the worst-case number of queries needed is small. Research has also been done on the relative power of different query types.

If enough information is elicited to also determine the VCG payments, and these are the payments charged to the bidders, answering the elicitor's queries truthfully is an *ex post* equilibrium (a strengthening of Nash equilibrium that does not rely on priors). (This assumes there is no explicit cost or limit to valuation determination; mechanisms have also been designed for settings where there is an explicit cost (Larson, 2006).) This holds even if the agents are allowed to answer queries that the elicitor did not ask (for example, queries that are easy for the agent to answer and which the agent thinks will significantly advance the elicitation process). We thus have a *pull–push mechanism* where both the centre and the agents guide the preference revelation (and thus also the preference determination/refinement by the agents). For a review, see Sandholm and Boutilier (2006). Ascending (combinatorial) auctions are an earlier special case, and have limited power compared to the general framework (Blumrosen and Nisan, 2005).

Preference elicitation can sometimes be computationally complex for the centre. It can be complex to

intelligently decide what to ask next, and from whom. It can also be complex to determine whether enough information has been elicited to determine the optimal outcome. Even if the elicitor knows that enough has been elicited, it can be complex to determine the outcome – for example, allocation of items to bidders in some CAs.

3.4 Distributed (centre-free) mechanisms

Computer scientists often have a preference for distributed applications that do not have any centralized coordination point (centre). Depending on the application, the reasons for this preference may include avoiding a single vulnerable point of failure, distributing the computing effort (for computational efficiency or because the data is inherently distributed), and enhancing privacy. The preference carries over from traditional computer science applications to different forms of negotiation systems – for example, see Sandholm (1993) for an early distributed automated negotiation system for software agents. Feigenbaum et al. (2005) have studied lowest-cost inter-domain routing on the Internet, modifying a distributed protocol so that the agents (routing domains) are motivated to report their true costs and the solution is found with minimal message passing. For a review of some other research topics in this space, see Feigenbaum and Shenker (2002).

One can go further by taking into account the fact that agents might not choose to follow the prescribed protocol. They may cheat not only on information-revelation actions, but also on message-passing and computational actions. Despite computation actions not being observable by others, an agent can be motivated to compute as prescribed by tasking at least one other agent with the same computation, and comparing the results (Sandholm et al., 1999). Careful problem partitioning can also be used to achieve the same outcome without redundancy by only requiring agents to perform computing and message passing tasks that are in their own interest (Parkes and Shneidman, 2004). Shneidman and Parkes (2004) propose a general proof technique and instantiate it to provide a non-manipulable protocol for inter-domain routing. Monderer and Tennenholtz (1999) develop protocols for one-item auctions executed among agents on a communication network. The protocols motivate the agents to correctly reveal preferences and communicate. For the setting where agents with private utility functions have to agree on variable assignments subject to side constraints (for example, meeting scheduling), Petcu, Faltings and Parkes (2006) developed a VCG-based distributed optimization protocol that finds the social welfare maximizing allocation and each agent is motivated to follow the protocol in terms of all three types of action. The only centralized party needed is a bank that can extract payments from the agents.

Cryptography is a powerful tool for achieving privacy when trying to execute a mechanism in a distributed way without a centre, using private communication channels among the agents. Consider first the setting with passive adversaries, that is, agents that faithfully execute the specified distributed communication protocol, but who try to infer (at least something about) some agents' private information.

- If agents are computationally limited – for example, they are assumed to be unable to factor large numbers – then arbitrary functions can be computed while guaranteeing that each agent maintains his privacy (except, of course, to the extent that the answer of the computation says something about the inputs) (Goldreich, Micali and Wigderson, 1987). Thus the desire for privacy does not

constrain what social choice functions can be implemented.

- In contrast, only very limited social choice functions can be computed privately among computationally unlimited agents. For example, when there are just two alternatives, every monotonic, non-dictatorial social choice function that can be privately computed is constant (Brandt and Sandholm, 2005). With special structure in the preferences, this impossibility can sometimes be avoided. For example, with the standard model of quasi-linear utility, first-price auctions can be implemented privately; second-price (Vickrey) auctions with more than two bidders cannot (Brandt and Sandholm, 2004).

A more general model is that of active adversaries who can execute the distributed communication protocol unfaithfully in a coordinated way. A more game-theoretic model is that of rational adversaries that are not passive, but not malicious either. For a brief overview of such work, see computer science and game theory.

This work was funded by the National Science Foundation under ITR grant IIS-0427858, and a Sloan Foundation Fellowship. I thank Felix Brandt, Christina Fong, Joe Halpern, and David Parkes for helpful comments.

Bibliography

Banks, J.S., Ledyard, J. and Porter, D. 1989. Allocating uncertain and unresponsive resources: an experimental approach. *RAND Journal of Economics* 20, 1–25.

Bartholdi, J., III and Orlin, J. 1991. Single transferable vote resists strategic voting. *Social Choice and Welfare* 8, 341–54.

Bartholdi, J., III, Tovey, C. and Trick, M. 1989. The computational difficulty of manipulating an election. *Social Choice and Welfare* 6, 227–41.

Bichler, M., Davenport, A., Hohner, G. and Kalagnanam, J. 2006. Industrial procurement auctions. In Cramton, Shoham and Steinberg (2006).

Blumrosen, L. and Nisan, N. 2005. On the computational power of iterative auctions. *Proceedings of the ACM Conference on Electronic Commerce*, 29–43.

Brandt, F. and Sandholm, T. 2004. (Im)possibility of unconditionally privacy-preserving auctions. *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, 810–17.

Brandt, F. and Sandholm, T. 2005. Unconditional privacy in social choice. *Proceedings of the Conference on Theoretical Aspects of Rationality and Knowledge*, 207–18.

Conen, W. and Sandholm, T. 2001. Preference elicitation in combinatorial auctions: extended abstract.

Proceedings of the ACM Conference on Electronic Commerce, 256–9. More detailed description of algorithmic aspects in *Proceedings of the IJCAI-01 Workshop on Economic Agents, Models, and Mechanisms*, 71–80.

Conitzer, V. and Sandholm, T. 2002. Complexity of mechanism design. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 103–10.

Conitzer, V. and Sandholm, T. 2003. Universal voting protocol tweaks to make manipulation hard. *Proceedings of the International Joint Conference on Artificial Intelligence*, 781–8.

Conitzer, V. and Sandholm, T. 2004. Computational criticisms of the revelation principle. *Conference on Logic and the Foundations of Game and Decision Theory*. Earlier versions: AMEC-03, EC-04.

Conitzer, V. and Sandholm, T. 2006. Nonexistence of voting rules that are usually hard to manipulate. *Proceedings of the National Conference on Artificial Intelligence*.

Conitzer, V., Sandholm, T. and Lang, J. 2007. When are elections with few candidates hard to manipulate? *Journal of the ACM* 54(3), Article 14.

Cramton, P., Shoham, Y. and Steinberg, R., eds. 2006. *Combinatorial Auctions*. Cambridge, MA: MIT Press.

Feigenbaum, J., Papadimitriou, C., Sami, R. and Shenker, S. 2005. A BGP-based mechanism for lowest cost routing. *Distributed Computing* 18, 61–72.

Feigenbaum, J. and Shenker, S. 2002. Distributed algorithmic mechanism design: recent results and future directions. *Proceedings of the International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, 1–13.

Goldreich, O., Micali, S. and Wigderson, A. 1987. How to play any mental game or a completeness theorem for protocols with honest majority. *Proceedings of the Symposium on Theory of Computing*, 218–29.

Hajiaghayi, M.T., Kleinberg, R. and Sandholm, T. 2007. Automated online mechanism design and prophet inequalities. *Proceedings of the National Conference on Artificial Intelligence*.

Hershberger, J. and Suri, S. 2001. Vickrey prices and shortest paths: what is an edge worth? *Proceedings of the Symposium on Foundations of Computer Science*, 252–9.

Jurca, R. and Faltings, B. 2006. Minimum payments that reward honest reputation feedback. *Proceedings of the ACM Conference on Electronic Commerce*, 190–9.

Kfir-Dahav, N., Monderer, D. and Tennenholtz, M. 2000. Mechanism design for resource bounded agents. *Proceedings of the International Conference on Multi-Agent Systems*, 309–15.

Land, A., Powell, S. and Steinberg, R. 2006. PAUSE: a computationally tractable combinatorial auction. In Cramton, Shoham and Steinberg (2006).

Larson, K. 2006. Reducing costly information acquisition in auctions. *Proceedings of the Autonomous Agents and Multi-Agent Systems*, 1167–74.

Larson, K. and Sandholm, T. 2001. Costly valuation computation in auctions. *Proceedings of the Theoretical Aspects of Rationality and Knowledge*, 169–182.

Larson, K. and Sandholm, T. 2005. Mechanism design and deliberative agents. *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, 650–6.

Lavi, R. and Swamy, C. 2005. Truthful and near-optimal mechanism design via linear programming. *Proceedings of the Symposium on Foundations of Computer Science*, 595–604.

Lehmann, D., Müller, R. and Sandholm, T. 2006. The winner determination problem. In Cramton, Shoham and Steinberg (2006).

Lehmann, D., O'Callaghan, L.I. and Shoham, Y. 2002. Truth revelation in rapid, approximately efficient combinatorial auctions. *Journal of the ACM* 49, 577–602.

Likhodedov, A. and Sandholm, T. 2005. Approximating revenue-maximizing combinatorial auctions. *Proceedings of the National Conference on Artificial Intelligence*, 267–74.

Monderer, D. and Tennenholtz, M. 1999. Distributed games: from mechanisms to protocols. *Proceedings of the National Conference on Artificial Intelligence*, 32–7.

Müller, R. 2006. Tractable cases of the winner determination problem. In Cramton, Shoham and Steinberg (2006).

Nisan, N. and Ronen, A. 2000. Computationally feasible VCG mechanisms. *Proceedings of the ACM Conference on Electronic Commerce*, 242–52.

Nisan, N. and Ronen, A. 2001. Algorithmic mechanism design. *Games and Economic Behavior* 35, 166–96.

Nisan, N. and Segal, I. 2006. The communication requirements of efficient allocations and supporting prices. *Journal of Economic Theory* 129, 192–224.

Parkes, D. and Shneidman, J. 2004. Distributed implementations of generalized Vickrey–Clarke–Groves auctions. *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, 261–8.

Petcu, A., Faltings, B. and Parkes, D. 2006. MDPOP: faithful distributed implementation of efficient social choice problems. *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, 1397–404.

Rothkopf, M. 2007. Thirteen reasons why the Vickrey–Clarke–Groves process is not practical. *Operations Research* 55, 191–7.

Rothkopf, M., Pek ě c, A. and Harstad, R. 1998. Computationally manageable combinatorial auctions. *Management Science* 44, 1131–47.

Sandholm, T. 1993. An implementation of the contract net protocol based on marginal cost calculations. *Proceedings of the National Conference on Artificial Intelligence*, 256–62.

Sandholm, T. 2000. Issues in computational Vickrey auctions. *International Journal of Electronic Commerce* 4, 107–29. Early version in ICMAS-96.

Sandholm, T. 2002a. Algorithm for optimal winner determination in combinatorial auctions. *Artificial Intelligence* 135, 1–54. Earlier versions: ICE-98 keynote, Washington U. tech report WUCS-99-01 Jan. 1999, IJCAI-99.

Sandholm, T. 2002b. eMediator: a next generation electronic commerce server. *Computational Intelligence* 18, 656–76. Earlier versions: Washington U. tech report WU-CS-99-02 Jan. 1999, AAAI-99 Workshop on AI in Ecommerce, AGENTS-00.

Sandholm, T. 2006. Optimal winner determination algorithms. In Cramton, Shoham and Steinberg (2006).

Sandholm, T. 2007. Expressive commerce and its application to sourcing: how we conducted $35 billion of generalized combinatorial auctions. *AI Magazine* 28(3), 45–58.

Sandholm, T. and Boutilier, C. 2006. Preference elicitation in combinatorial auctions. In Cramton, Shoham and Steinberg (2006).

Sandholm, T., Conitzer, V. and Boutilier, C. 2007. Automated design of multistage mechanisms. *Proceedings of the International Joint Conference on Artificial Intelligence*, 1500–6.

Sandholm, T. and Ferrandon, V. 2000. Safe exchange planner. *Proceedings of the International*

Conference on Multi-Agent Systems, 255–62.

Sandholm, T., Larson, K., Andersson, M., Shehory, O. et al. 1999. Coalition structure generation with worst case guarantees. *Artificial Intelligence* 111, 209–38.

Shneidman, J. and Parkes, D.C. 2004. Specification faithfulness in networks with rational nodes. *Proceedings of the ACM Symposium on Principles of Distributed Computing*, 88–97.

Vorobeychik, Y., Kiekintveld, C. and Wellman, M. 2006. Empirical mechanism design: methods, with application to a supply chain scenario. *Proceedings of the ACM Conference on Electronic Commerce*, 306–15.

How to cite this article

Sandholm, Tuomas. "computing in mechanism design." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000563> doi:10.1057/9780230226203.0288

The New Palgrave Dictionary of Economics Online

concentration measures

Juan Esteban Carranza

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Concentration is a characterization of the size distribution and quantity of competing firms within a specific market or industry. The most common concentration measures are the Herfindahl index and the *n*-firm concentration rate. The Herfindahl index is the sum of the squared market shares of all the firms in a market, whereas the *n*-firm concentration rate is the sum of the market shares of the *n* biggest firms. These measures are a significant reflection of the underlying degree of competitiveness, but are sensitive to the adopted market definition, and must be interpreted carefully depending on the specifics of the case.

Keywords

competition; concentration; concentration measures; entropy coefficients; Gini coefficient; Herfindahl index; Lerner index; market definition; market structure

Article

The term *concentration* (also *firm concentration*, *industry concentration* or *market concentration*) refers to aspects of the distribution of firm size within a specific market or industry that have traditionally been used to characterize the degree of competitiveness in the market. Even though the size of firms can be measured using many different variables, such as employment or assets, the sales level is the most commonly used size measure. Accordingly, if very few firms serve a very large portion of the market, it is said that the given market is highly ‘concentrated’, whereas if no single firm has a large share of sales it is said that the market is not ‘concentrated’. Since concentration is an important reflection of the underlying market structure, its measurement is an important characterization of the interaction of firms within a specific market or industry.

The most common concentration measures are the ‘*n*-firm concentration rate’ and the ‘Herfindahl index’. Let S_i be the market share of firm *i*; the ‘*n*-firm concentration rate’ is the sum of the market shares of the *n* biggest firms within the market:

$$C(n) = \sum_{i=1}^{n} S_i.$$

As indicated, the summation above is taken over the set of *n* biggest firms in the market. So, for example, the two-firm concentration rate of a given market is the sum of the market shares of the two biggest firms in the market where size is measured according to observed sales. In order to fully characterize the concentration of any given market, though, a number of these rates must be used, since there is no agreed-on value for *n*. This complicates its use for comparing concentration over time and across sectors, and for its use in statistical analysis.

The Herfindahl index, first devised by Albert Hirschman to measure the concentration of trade across sectors (so that the index is also known as ‘Herfindahl–Hirschman index’; see Hirschman, 1964, for its history), is the sum of the squared market shares of all firms in the market:

$$H = \sum_{i=1}^{N} S_i^2.$$

The summation in this case is taken over the set of all *N* firms in the market. This index lies between zero and 1: if there is only one firm in the market, so that the market has the highest possible concentration, the index is 1. If, on the other hand, there are many equally sized firms in the market, the index will be close to zero. By squaring the individual market shares, this index gives relatively greater weight to the market shares of large firms. Conversely, the addition of one small firm to the market dilutes somewhat the market share of larger firms, and has a marginal negative effect on the index, which is consistent with any notion of market concentration. Any value of this index can correspond to multiple market configurations, being in that sense less illustrative of the actual concentration of a market than a set of *n*-firm concentration rates. On the other hand, this index can be easily correlated with other market characteristics and is therefore very useful for statistical analysis.

Other less commonly used concentration measures include entropy coefficients, the Gini coefficient and measures of the variance of market shares across firms within a market. The entropy coefficient is usually computed using the following formula:

$$E = \sum_{i=1}^{N} S_i \log_2\left(\frac{1}{S_i}\right).$$

This index takes value zero if there is only one firm in the market and grows as market concentration decreases. The interpretation of this coefficient is complicated, because its formula weights both large and small firms less heavily than mid-size firms and grows unboundedly as market concentration decreases. It is therefore less commonly used than the Herfindahl index.

The Gini coefficient is commonly used to characterize the income or wealth inequality within a society. Its drawback as a measure of market concentration is that it is useful only to measure the concentration of firms' sizes within a market, given a number of firms. So according to the Gini coefficient a duopolistic market with two firms of equal size is as concentrated as a market with 100 firms with identical size. The same drawback applies for the use of measures of the variance of firm size – whose definition is simply the sample variance of firm sizes.

All the concentration measures mentioned above are very sensitive to the actual market definition that is used. In markets for differentiated products, for example, products may face a continuum of similar products, and determining which similar products exactly constitute a market is not always easy. Take the specific example of the market faced by US mobile phone services: with just a handful of national providers it is concentrated given the standard concentration measures. These national firms, nevertheless, are also competing with local companies in various segments of the market and even with long-distance phone companies and Internet companies as providers of communication services. The Internet, on the other hand, competes in some instances with cable and satellite companies, radio stations and even newspapers as sources of news and entertainment. What exactly the relevant market faced by mobile phone companies is will depend on the type of issue being addressed. Accordingly, the concentration measures will change depending on the adopted market definition.

On the other hand, even if the market is well defined, computed concentration measures may not reflect at all the real competitive structure of the market. For example, even in markets as highly concentrated as the market for computer processors, the dominant firms have to account for the invisible competition of potential entrants. The same happens in regional markets where outside firms are kept at bay by few local firms with a combination of low prices and high transportation costs. Computed concentration measures for specific markets cannot account for this unobserved competition and may therefore lead to wrong conclusions regarding the underlying behaviour of firms. In these instances, a behavioural measure, such as the Lerner index, which measures the relative size of firms' markups, may be a better indicator of the competitive structure of the market.

There is a body of empirical literature that uses market concentration measures across industries to approximate the underlying differences in industries' competitiveness. They were then used to infer statistically the relationship between market ‘structure’ and market ‘performance’. For example, correlations of R&D expenditure and market concentration were computed to investigate whether firms in concentrated markets were more or less likely to innovate than firms in more competitive markets. The value of such correlations is limited because the observed concentration may be both a cause and an effect of individual firms' behaviour and the relationship is shaped by the specifics of the industry. In order to avoid the ambiguities of such an inter-industry approach, the more recent empirical microeconomic literature has generally focused instead on the understanding of firm behaviour within specific industries, for which the use of concentration measures is less relevant.

See Also

- competition
- Gini ratio
- market structure

Bibliography

Hirschman, A.O. 1964. The paternity of an index. *American Economic Review* 54, 761–2.

Ravenscraft, D. 1983. Structure-profit relationships at the line of business and industry level. *Review of Economics and Statistics* 65, 22–31.

Scherer, F.M. 1970. *Industrial Market Structure and Economic Performance*, 2nd edn. Boston: Houghton Mifflin.

Simon, H. and Bonini, C.P. 1958. The size distribution of business firms. *American Economic Review* 48, 607–17.

How to cite this article

Carranza, Juan Esteban. "concentration measures." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000592> doi:10.1057/9780230226203.0289

The New Palgrave Dictionary of Economics Online

Condillac, Etienne Bonnot de, Abb é de Mureau (1714 – 1780)

Peter Groenewegen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Baudeau, N.; Condillac, E. B. de; exchange; Galiani, F.; Le Trosne, G. F.; Physiocracy; price; Turgot, A. R. J.; value; Verri, P.

Article

Philosopher and economist. Born at Grenoble, the third son of a well-to-do aristocratic family, Condillac took his name from an estate purchased by his father in 1720. As a sickly child with poor eyesight he had little early education and was apparently still unable to read by the age of 12. After his father's death in 1727 he moved to Lyon to live with his oldest brother, continuing his education at its Jesuit college. Through this brother he may have first met Jean Jacques Rousseau, who was tutor to his nephews in 1740 and became a life-long friend. His second brother, l'Abbé de Mably, took Condillac to Paris in *c*. 1733 to study theology at Saint Sulpice and the Sorbonne. He was ordained in 1740 and for the rest of his life ‘ever faithful to the Christian church, would always wear his cassock, always remain l'Abbé’ (Lefèvre, 1966, p. 11).

For the next 15 years he lived the life of a Paris intellectual, studying the philosophy of Descartes, Malebranche, Leibniz and Spinoza, ‘to whose speculative systems he formed a life-long aversion, preferring the English philosophers Locke (who particularly influenced his thinking), Berkeley, Newton and rather belatedly, Bacon’ (Knight, 1968, pp. 8–9). In this period he published the works which made his philosophical reputation: the *Essay on the Origin of Human Knowledge* (1746), the *Traité des Systèmes* (1749), his most famous philosophical work *Treatise on the Sensations* (1754) described as the ‘most rigorous demonstration of the [18th-century] sensationalist psychology’ (Knight, 1968, p. 12) and his *Traité des Animaux* (1755).

Apart from giving him entry to the Paris salons, where at Mlle de Lespinasse's salon he is reputed to have first met Turgot, another life-long friend (Le Roy, 1947, p. ix), his intellectual reputation gained him the position of tutor to Louis XV's grandson, the Duke of Parma. From 1758 to 1767 he resided in Parma. Because of its prime minister's economic development policies, inspired by a mixture of

‘mercantilism, physiocracy and the ideas of Gournay’, Condillac developed an interest in economic matters, an interest ‘indirectly confirmed by his known contacts with the Italian political economists, Beccaria and Gherardo’ (Knight, 1968, pp. 231–2). In 1768 he returned to Paris, but by 1773 had retired to his estate of Flux near Beaugency, where he died in 1780. During the last decade of his life he published his *Cours d'Etudes* (1775), his work on economics (1776), a text on logic (1780) for use in Polish Palatinate schools, and commenced the unfinished *La Langue des Calculs* (1798). In 1752, he became a member of the Royal Prussian Academy; in 1768 after his return from Parma he was elected to the French Academy. His works have been frequently collected, most recently by Le Roy (1947–51).

The impetus for Condillac's writing *Le Commerce et le Gouvernement* has been ascribed to a desire to assist his friend Turgot in the difficulties he faced in 1775 as finance minister over the grain riots induced by his restoration of the free trade in grain (Le Roy, 1947, p. xxv; Knight, 1968, p. 232). This fits with the work's unqualified support for free trade in general and the grain trade in particular (1776, esp. pp. 344–5, which seems directly inspired by the Paris events of 1775). Writing the book may also be explained as a return favour for Turgot's assistance in getting Condillac (1775) published (cf. Knight, 1968, pp. 13, 232). Despite Condillac's strong support for this major part of Physiocratic policy and his close adherence to other aspects of Physiocracy, his argument that manufacturing was productive brought critical replies from Baudeau and Le Trosne (1777). In this context it may be noted that his work bears little direct Physiocratic influence, the major influence being Cantillon (1755), the only work directly cited apart from Plumard de Dangeul (1754). It is, however, possible to detect some influence from the economics of Turgot, Galiani and Verri on the theory of value, price and competition (cf. Spengler, 1968, p. 212).

As published, the work is divided into two parts. The first provides the elements of the science. Its starting point is the foundation of value, which Condillac finds in the usefulness of an object relative to subjective needs making relative scarcity the key variable determining value. Value is distinguished from price because price can only originate in exchange. It is determined by the competition between buyers and sellers guided by their subjective estimation of value. Gains from exchange arise from differences in value; for Condillac, value cannot exchange for equal value. Although Condillac did discuss the costs of acquiring commodities, his emphasis is on exchange, trade and price. Exchange presumes surplus production and a need for consumption. Hence trade inspires and animates production and is essential to increasing wealth. Only simple pictures of production are presented: farm labourers producing prime necessities of food and materials; artisans transforming raw materials into essentials and luxuries; traders who circulate these products at home and abroad. By this circulation trade distributes the annual product and under competitive conditions settles its true prices. Condillac is more concerned with developing the institutions associated with trade: growth of towns and villages, money, banking, credit, interest and the foreign exchanges, the defence of property by government and hence the need for taxation, and the effects of restraints on trade, including the grain trade. The second part is almost completely devoted to examining effects of specific obstacles to trade ranging from war, tariffs, taxes, excessive government borrowing to luxury spending in the capital city and exclusive trading privileges. Moderate wants combined with complete freedom constitute his recipe for the best form of economic development.

Condillac's economic work received a mixed reception from later economists. J.B. Say (1805, p. xxxv) described it as an attempt ‘to found a system of … a subject which [the author] did not understand’.

Jevons (1871, p. xviii) praised Condillac's ‘charming philosophic work [because] in the first few chapters … we meet perhaps the earliest distinct statement of the true connections between value and utility…’. Macleod (1896, p. 73) described it as a ‘remarkable work … utterly neglected but in scientific spirit … infinitely superior to Smith’. Since then, it has remained neglected even though as ‘a good if somewhat sketchy treatise on economic theory and policy [it was] much above the common run of its contemporaries’ (Schumpeter, 1954, pp. 175–6).

Selected works

1746. *An Essay on the Origin of Human Knowledge*, trans. Thomas Nugent. London: Nourse, 1756.

1749. *Traité des Systèmes, où l'on en démèle les inconvénine et les advantages*. Paris and Amsterdam.

1754. *Treatise on the Sensations*, trans. B.S. Geraldine Czar. London: Favill Press, 1930.

1755. *Traité des Animaux, où après avoir fain des observations critique sur le sentiment de Descartes et sur celui de M. Buffon on entreprend d'expliquer leurs principales facultés* …. Amsterdam.

1775. *Cours d'Etude pour l'instruction du Prince de Parma, Aujourd'hui Ferdinand, Duc de Parma*. Parma (and Paris).

1776. *Le Commerce et le gouvernement considerés relativement l'un à l'autre*. In *Oeuvres complètes de Condillac*, vol. 4. Paris: Brière, 1821.

1780. *The Logic*, trans. B.S. Joseph Neef. Philadelphia, 1809.

1798. *Le langue des Calculs, Ouvrage Posthume et élémentaire*. Paris.

Bibliography

Cantillon, R. 1755. *Essay on the Nature of Commerce in General*, trans. B.S.H. Higgs. London: Macmillan, 1931.

Jevons, W.S. 1871. *Theory of Political Economy*. 4th edn, London: Macmillan, 1911.

Knight, I.F. 1968. *The Geometric Spirit. The Abbé de Condillac and the French Enlightenment*. New Haven and London: Yale University Press.

Lefèvre, R. 1966. *Condillac ou la joie de vivre*. Paris: Editions Seghers.

Le Roy, G., ed. 1947–51. *Oeuvres philosophiques de Condillac*. Paris: Press Universitaires de France.

Le Trosne, G.F. 1777. *De L'intérêt Social, par rapport à la Valeur, à la circulation, à l'Industrie, & au commerce intérieur & extérieur: Ouvrage élémentaire dans lequel on discoute quelques Principes de M. l'Abbé de Condillac*. Paris.

Macleod, H.D. 1896. *The History of Economics*. London: Bliss, Sands & Co.

Plumard de Danguel, L.J. 1754. *Remarques sur les avantages et les désavantages de la France et de la Gr. Bretagne par rapport au commerce et aux autres sources de la puissance des états*. Leyden (and Paris).

Say, J.-B. 1805. *A Treatise on Political Economy of the Production, Distribution and Consumption of Wealth*, trans. C.R. Prinsep, New American Edition, Philadelphia, 1880; reissued New York: Kelley, 1963.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen & Unwin, 1959.

Spengler, J.J. 1968. Condillac, Etienne Bonnot de. In *Encyclopedia of the Social Sciences*, vol. 3, 2nd edn. Chicago: Chicago University Press.

How to cite this article

Groenewegen, Peter. "Condillac, Etienne Bonnot de, Abbé de Mureau (1714–1780)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000287> doi:10.1057/9780230226203.0290

The New Palgrave Dictionary of Economics Online

Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743 – 1794)

H. Moulin and H.P. Young

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Black, D.; Borda, J.-C. de; Condorcet, Marquis de; Condorcet's paradox; impossibility theorem; independence of irrelevant alternatives; maximum likelihood; strategic voting; voting rules

Article

Condorcet was a French mathematician and philosopher. With many of his fellow *encyclopédistes* he shared the conviction that social sciences are amenable to mathematical rigour. His pioneer work on elections, the *Essai sur l'application de l'analyse a la probabilité des decisions rendues a la pluralité des voix* (1785) is a major step in that direction.

The aim of the *Essai* is to ‘inquire by mere reasoning, what degree of confidence the judgement of assemblies deserves, whether large or small, subject to a high or low plurality, split into several different bodies or gathered in one only, composed by men more or less wise’ (*Discours préliminaire* to the *Essai*, p. iv).

In modern words, this is the jury problem: to decide whether the accused is guilty or not requires converting the opinions of several experts, with varying competence, into a single judgement. Systematic probabilistic computations for this problem occupy most of the *Essai*, often camouflaging the essential contributions. The opaqueness and technicality of the argument meant that a full recognition of its importance did not occur until more than 150 years later (Black, 1958). Since then Condorcet's findings have strongly influenced modern social choice theorists (for example, Arrow, Guilbaud and Black), and still play a central role in many of its recent developments.

The starting point is that majority voting is the unambiguously best voting rule when only two candidates are on stage. This fact, whose modern formulation is known as May's theorem (May, 1952) was clear enough to the encyclopedists, too. How, then, can we extend this rule to three candidates or more? The naive, yet widely used, answer is plurality voting (each voter casts a vote for one candidate; the candidate with most votes is elected). Both Condorcet and Borda (his colleague in the Academy of Sciences) raise the same objection against the plurality rule. Suppose, says Condorcet (*Discours préliminaire*, p. lviii) that 60 voters have the opinions shown in Table 1 about three candidates A, B, C.

	23	19	16	2
Top	A	B	C	C
	C	C	B	A
Bottom	B	A	A	B

In the illustration, candidate A wins by plurality. Yet if we oppose A against B only, A loses (25 to 35) and in A against C, A loses again (23 to 37). Thus the plurality rule does not convey accurately the opinion of the majority. From these identical premises, Borda proposes his well-known scoring method (each candidate receives 2 points from a voter who ranks him first, 1 point from one who ranks him second, and none from one who ranks him last; hence C is elected with score 78), whereas Condorcet opens a quite different route.

Condorcet posits a simple binomial model of voter error: in every binary comparison, each voter has a probability $1/2 < P < 1$ of ordering the candidates correctly. All voters are assumed to be equally able, and there is no correlation between judgements on different pairs. Thus for Condorcet the relevant data is contained in the ‘majority tournament’ that results from taking all pairwise votes:

B beats A, 35 to 25; C beats A, 37 to 23;

C beats B, 41 to 19.

Condorcet proposes that the candidates be ranked according to ‘the most probable combination of opinions’ (*Essai*, p. 125). In modern statistical terminology this is a maximum likelihood criterion (see Young, 1986).

In the above example the most probable combination is given by the ranking: CBA since the three statements C over B, C over A, B over A agree with the greatest total number of votes. Condorcet's ranking criterion implies that an alternative (such as C) that obtains a majority over every other alternative must be ranked first. Such an alternative, if one exists, is known as a ‘Condorcet winner’. As Condorcet points out, some configurations of opinions may not possess such a winner, because the majority tournament contains a cycle (a situation known as ‘Condorcet's paradox’). He exhibits the example shown in Table 2.

23	17	2	10	8
A	B	B	C	C
B	C	A	A	B
C	A	C	B	A

Here A beats B, 33 to 27; B beats C, 42 to 18; C beats A, 35 to 25. According to Condorcet's maximum likelihood criterion, this cycle should be broken at its weakest link (A over B), which yields the ranking B over C over A. Therefore in this case B is declared the winner.

Somewhat later in the *Essai* (pp. 125–6), Condorcet suggests that one may compute the maximum likelihood ranking of *n* candidates by, first, choosing the $n(n-1)/2$ binary propositions that have the majority in their favour; then, if there are cycles, *successively deleting* those with smallest majorities until a complete ordering of the candidates is obtained. Unfortunately, for $n > 3$ this heuristic algorithm does not necessarily yield the ranking that accords with the greatest number of votes. An axiomatic characterization of Condorcet's rule is given in Young and Levenglick (1978).

Condorcet's idea of reducing individual opinions to all pairwise comparisons between alternatives proved essential to the aggregation of preferences approach initiated by Arrow (1951). The key axiom independence of irrelevant alternatives (IIA) requires that voting on a pair of candidates be enough to determine the collective opinion on this pair: this generalizes majority tournaments by dropping the symmetry across voters and across candidates. In this sense Arrow's impossibility theorem means that the Condorcet paradox is inevitable in any non-dictatorial voting method satisfying IIA.

Many more useful insights can be discovered in the *Essai*. For instance the issue of strategic manipulations, which has played a central role in the theory of elections since the late 1960s, is suggested in places, although it is never systematically analysed. For example, on page clxxix of the *Discours Preliminaire*, Condorcet criticizes Borda's method as more vulnerable to a '*cabale*'. His argument is supported by the modern game theoretical approach: whenever the configurations of individual opinions guarantee existence of a Condorcet winner, it defines a strategy-proof voting rule. This is one of the principal arguments in favour of Condorcet consistent voting rules, namely, rules electing the Condorcet winner whenever it exists (see, for example, Moulin, 1983, ch. 4).

See Also

- Borda, Jean-Charles de
- social choice
- voting paradoxes

Selected works

1785. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité voix. Paris.

Bibliography

Arrow, K. 1951. *Social Choice and Individual Values*. New York: Wiley.

Black, D. 1958. *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.

May, K. 1952. A set of independent necessary and sufficient conditions for simple majority decision.

Econometrica 20, 680–84.

Moulin, H. 1983. *The Strategy of Social Choice*. Amsterdam: North-Holland.

Young, H.P. 1986. Optimal ranking and choice from pairwise comparisons. In *Information Pooling and Group Decisionmaking* ed. B. Grofman and G. Owen. Greenwich, CT: JAI Press.

Young, H.P. and Levenglick, A. 1978. A consistent extension of Condorcet's election principle. *SIAM Journal of Mathematics* 35, 285–300.

How to cite this article

Moulin, H. and H. Peyton Young. "Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743–1794)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000289> doi:10.1057/9780230226203.0291

The New Palgrave Dictionary of Economics Online

congestion

Richard Arnott and Marvin Kraus
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Congestion’ is the phenomenon whereby the quality of service provided by a *congestible facility* degrades as its aggregate usage increases, when its capacity is held fixed. Here, the economic theory of congestion is developed in the context of road traffic. The primary questions of interest are how the capacity of a congestible facility and its usage fee should be chosen. This leads naturally to the question of whether the usage fees collected will be sufficient to cover capacity costs at the optimum.

Keywords

clubs; congestion; externality cost; first-best pricing; local public goods; Ramsey pricing; second-best theory

Article

‘Congestion’ is the phenomenon whereby the quality of service provided by a *congestible facility* degrades as its aggregate usage increases, when its capacity is held fixed. We shall develop the economic theory of congestion in the context of road traffic, but congestion is pervasive: more telephone usage increases the probability of encountering a busy line; higher electricity demand may lead to voltage fluctuations, brownouts and eventually blackouts; more swimmers in a pool make comfortable swimming more difficult; more patients visiting a medical clinic results in longer waits and lower-quality care; in a more crowded classroom, students receive less individual attention, and more time is wasted on administration and discipline; and so on. The economic theory of congestion identifies how the capacity of a congestible facility and its usage fee should be chosen. Some degree of congestion is typically socially optimal.

The economic theory of congestion has much in common with the theory of clubs and local public goods (Scotchmer, 2002). The two literatures examine similar issues, but the economic theory of congestion has a policy perspective, while the theory of clubs and local public goods focuses on decentralized provision.

Formally, we may define congestion as follows. Consider a congestible facility in a steady state, that comprises *I* congestible elements. (Congestible elements for a sports stadium, for example, include nearby roads, parking facilities, the ticket office, washrooms, concessions, and seating.) Element *i* is characterized by a flow capacity, k_i and a stock capacity, K_i; the flow capacity is the maximum throughput per unit time, the stock capacity the maximum number of users at a point in time. Similarly, the level of usage is described in terms of the throughput of congestible element *i*, n_i and the number of users at a point in time, N_i. The congestible facility provides *J* dimensions of quality of service, with the level of dimension *j* indicated by s_j. Letting *k*, *K*, *n*, *N* and *s* denote the corresponding vectors,

$$s = S(k, K, n, N). \tag{1}$$

Congestion occurs when there is at least one combination of j and i for which s_j is monotone decreasing in n_i (flow congestion) or N_i (stock congestion), that is, when some dimension of quality of service falls as the throughput or stock of users of some congestible element of capacity increases. This is the static or steady-state definition of congestion. The dynamic definition of congestion adds time subscripts to s, k, K, n and N, and appends equations of motion relating stocks and flows for the various elements of capacity.

For some congestible elements, such as a turnstile, the bottleneck in the Vickrey (1969) bottleneck model of traffic congestion, or a switching circuit, the flow capacity constraint is the more important; for others, such as a telephone line, an elevator, a swimming pool, or seating at a football stadium, the stock capacity constraint is the more important. It should also be noted that a congestible facility can take the form of a network of congestible elements of capacity; a natural distinction is then between link congestion (for example, highway links) and nodal congestion (for example, traffic intersections).

To develop the theory, we consider a particular congestible facility having a single element of capacity and identical users, that is in a steady state: a road of uniform width connecting a single entry point A and a single exit point B, for which an increase in traffic flow increases travel time and an increase in road width reduces it. In this context, the deterioration of quality of service with an increase in usage is the increase in travel time from an increase in traffic flow.

We start with the short-run problem of determining optimal flow and its decentralized attainment, holding road width fixed. Let f denote flow, w road width, $t = t(f, w)$ the travel time function with (functional subscripts denote partial derivatives) $t_f > 0$ and $t_w < 0$ and ρ the value of time. Then the cost to an individual driver of travelling from A to B, the *user cost*, is $\rho t(f, w)$. Total user costs per unit time equal flow times user cost: $\rho f t(f, w)$. The social cost per unit time from increasing flow by one unit, with capacity held fixed, the *short-run marginal social cost*, is $\rho t(f, w) + \rho f t_f(f, w)$. The first term is the user cost of the extra driver; the second, the *congestion externality cost*. A driver imposes a congestion externality by slowing other drivers down; increasing steady-state flow by one car increases each car's travel time by $t_f(f, w)$ and social cost by $\rho f t_f(f, w)$.

Figure 1 displays short-run equilibrium. p denotes trip price, $D(p)$ the aggregate trip demand function, and $uc(f)$ and $srmsc(f)$ the user cost and short-run marginal social cost as a function of f, holding w fixed. With no toll, a user's trip price equals his user cost, and equilibrium occurs where the demand and user cost functions intersect, with flow f^e. Assuming that the marginal social benefit from a trip equals the corresponding marginal willingness to pay, the optimum occurs where the demand and short-run marginal social cost curves intersect, with flow f^*. Thus, with no toll, equilibrium flow is excessive. Efficiency obtains when economic agents face the social costs of their decisions and derive the social benefits from them. In the no-toll case, the price of a trip falls short of its marginal social cost since a driver does not pay for slowing down other drivers. Following Pigou (1947), the standard remedy for internalizing the congestion externality is to impose a toll equal to the congestion externality cost, evaluated at the social optimum: τ^* in Figure 1. This causes the trip price function to shift up from $uc(f)$ to $uc(f) + \tau^*$ and equilibrium flow to fall to the optimal level.

Figure 1

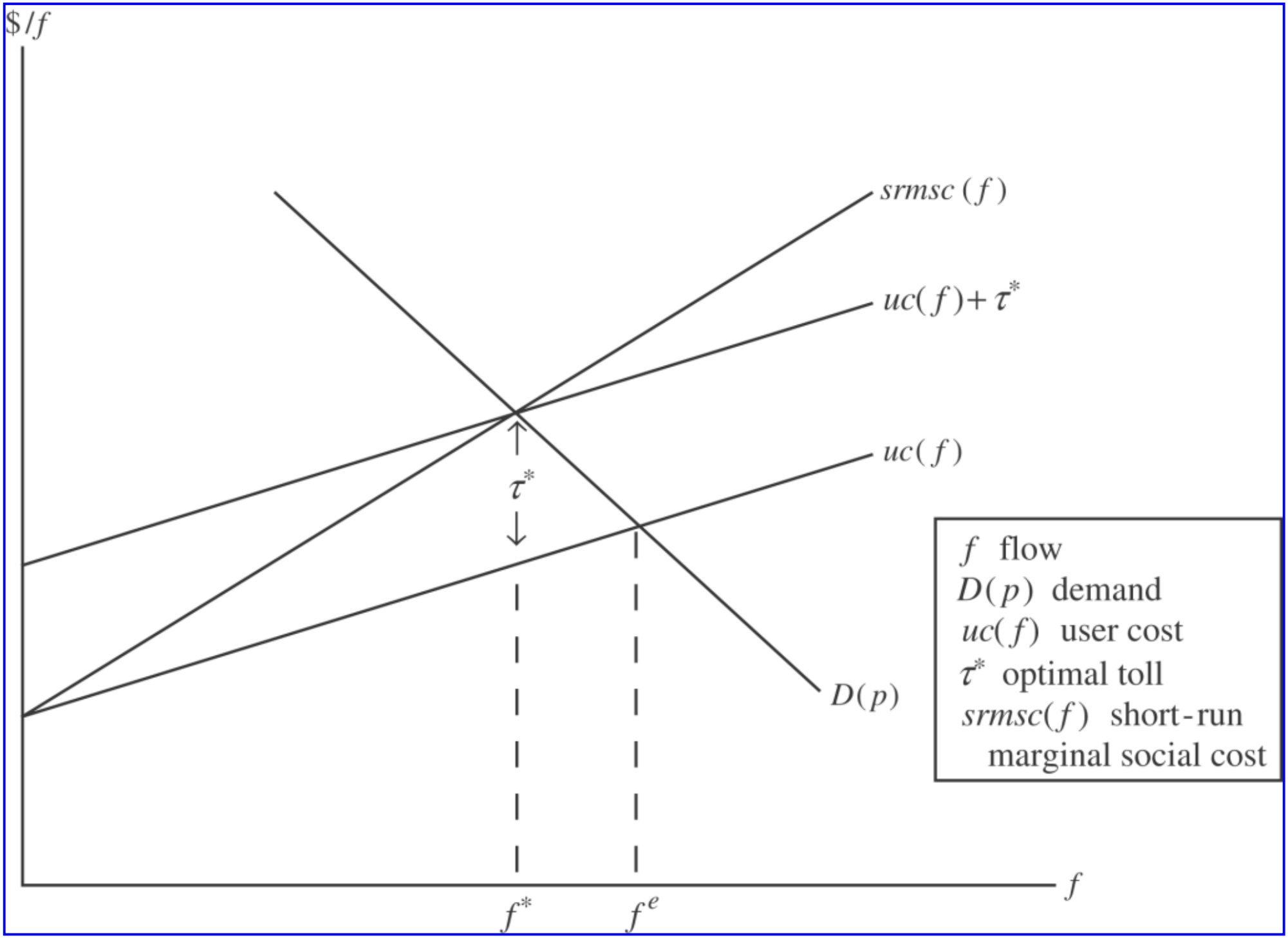

The above argument illustrates the general principle that efficient utilization of a congestible facility requires that the price equal short-run marginal social cost and the toll the congestion externality cost. Different user types – for example, cars and trucks – may impose different congestion externality costs. Efficiency then requires that the toll be

differentiated according to user type.

We now turn to the long-run planning problem in which both road width and flow are choice variables. We then consider decentralization of the optimum. Let $B(f)$ denote the social benefit per unit time from flow *f*, and *C*(*w*) the amortized capital cost of road width *w*. (We ignore the complications that arise when the congestible facility is sufficiently large that its construction alters factor prices.) The social surplus generated by the road (per unit time) equals social benefit minus social cost, and social cost equals total user cost plus amortized capital cost:

$$SS(f, w) = B(f) - p\,ft(f, w) - C(w). \tag{2}$$

It is easily seen from (2) that the road width that maximizes social surplus is that which minimizes social cost. This means that, when the long-run planning problem is solved, production is carried out according to the long-run cost structure, and the short-run marginal cost pricing (which is again) required for optimal flow is equivalent to long-run marginal cost pricing:

$$p = LRMC. \tag{3}$$

Now, recall the basic result of production theory that *LRMC* is equal to, less than or greater than *LRAC* (long-run average cost) according to whether *LRAC* is constant, decreasing or increasing. Combining this with (3), we have the result that, when *LRAC* is constant, *p=LRAC* holds at a long-run optimum. This is equivalent to equality between the total value of output and the total cost of output. Since total user cost is a component of both, this equality implies equality between toll receipts and amortized capital cost. Thus, *in the case of constant long-run average cost, the revenue raised from the optimal toll exactly covers the capital cost of providing a road of optimal width*. This is known as the ‘self-financing’ result. It was first derived by Mohring and Harwitz (1962) and subsequently generalized by Strotz (1965). (For a geometric derivation, see Arnott and Kraus, 2003.)

The self-financing result extends to congestible facilities with multiple elements of capacity, multiple dimensions of quality of service, and multiple user groups. If a congestible facility exhibits constant long-run average costs, provision of the facility can be decentralized via competing ‘clubs’; competition will result in each club charging each user a fee for use of its congestible facility equal to the congestion externality cost he imposes, and choosing optimal capacity.

The above theory was developed on the assumption of a steady state. In the extension to treat nonstationary dynamics, which is conceptually straightforward, the distinction between flow externalities and stock externalities becomes sharper.

The theory relates to *first-best* pricing and capacity choice when congestion is the only externality. When usage entails other externalities, such as pollution, first-best pricing should take these into account. In any policy context, additional practical constraints that rule out attainment of the full first-best allocation need to be considered. These are treated by applying second-best theory (Diamond and Mirrlees, 1971). Consider, for example, the pricing problem facing a public transit authority. The underpricing of urban auto travel may call for the underpricing of mass transit (Lévy-Lambert, 1968; Marchand, 1968); since optimal lump-sum redistribution is infeasible for informational reasons, the authority may choose to sacrifice some efficiency to improve equity by charging lower fares to needy groups (Atkinson and Stiglitz, 1980), rationing, or nonlinear pricing (Wilson, 1993); administrative costs may preclude fine-tuning the fare according to distance travelled or time of day, leading to variants of Ramsey pricing (Mohring, 1970); the authority may face a deficit constraint, requiring it to price above marginal social cost (Boiteux, 1956); with distortionary taxation, the social cost of financing an extra dollar of transit authority deficit may significantly exceed one dollar (Vickrey, 1959); and the government may choose to deviate from marginal social cost pricing to provide the public transit authority with higher-powered incentives (Laffont and Tirole, 1993) or to achieve political objectives. These considerations will also cause second-best capacity to deviate from first-best capacity.

See Also

- consumption externalities
- France, economics in (before 1870)

- network goods (empirical studies)
- network goods (theory)
- Pigouvian taxes
- urban economics
- value of time

Bibliography

Arnott, R. and Kraus, M. 2003. Principles of transport economics. In *Handbook of Transportation Science*, 2nd edn, ed. R. Hall. Boston: Kluwer.

Atkinson, A. and Stiglitz, J. 1980. *Lectures on Public Economics*. New York: McGraw-Hill.

Boiteux, M. 1956. Sur la gestion des monopoles publics astreints à l'équilibre budgétaire. *Econometrica* 24, 22–40.

Diamond, P. and Mirrlees, J. 1971. Optimal taxation and public production I: production efficiency and II: tax rules. *American Economic Review* 61, 8–27, 261–78.

Laffont, J.-J. and Tirole, J. 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

Lévy-Lambert, H. 1968. Tarification des services à qualité variable: Application aux péages de circulation. *Econometrica* 36, 564–74.

Marchand, M. 1968. A note on optimal tolls in an imperfect environment. *Econometrica* 36, 575–81.

Mohring, H. 1970. The peak-load problem with increasing returns and pricing constraints. *American Economic Review* 60, 693–705.

Mohring, H. and Harwitz, M. 1962. *Highway Benefits: An Analytical Framework*. Evanston, IL: Northwestern University Press.

Pigou, A. 1947. *A Study in Public Finance*, 3rd edn. London: Macmillan.

Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of Public Economics*, vol. 4, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Strotz, R. 1965. Urban transportation parables. In *The Public Economy of Urban Communities*, ed. J. Margolis. Washington, DC: Resources for the Future.

Vickrey, W. 1959. Statement on the pricing of urban street use. *Hearings: US Congress, Joint Committee on Metropolitan Washington Problems*, 11 November, 454–77.

Vickrey, W. 1969. Congestion theory and transport investment. *American Economic Review Proceedings* 59, 251–60.

Wilson, R. 1993. *Nonlinear pricing*. Oxford: Oxford University Press.

How to cite this article

Arnott, Richard and Marvin Kraus. "congestion." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000292> doi:10.1057/9780230226203.0292

The New Palgrave Dictionary of Economics Online

conjectural equilibria

F. Hahn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In imperfectly competitive economies, agents must take note of the effects of their decisions on the market environment. Such effects, being uncertain, are the subject of conjecture. Even if conjectures are not derivable from some first principles of rationality, conjectural theories are of interest because they attempt a general equilibrium analysis of non-perfect competition. The conjectural approach takes proper and explicit note of the perceptions by individuals of their market environment; it is possible that what is the case may depend on what agents believe to be the case.

Keywords

bootstrap equilibria; conjectural equilibria; duopoly; extensive form games; fixprice equilibria; fixprice models; game theory; general equilibrium; imperfect competition; no surplus condition; perfectly competitive equilibrium; rational conjectural equilibrium; reasonable conjectures; sequential equilibrium

Article

In an economy with very many agents the market environment of any one of these is independent of the market actions he decides upon. More generally one can characterize an economy as *perfectly competitive* if the removal of any one agent from the economy would leave the remaining agents just as well off as they were before his removal. (The economy is said to satisfy a ‘no surplus’ condition; see Makowski, 1980; and Ostroy, 1980.) When an economy is not perfectly competitive, an agent in making a decision must take note of its effect on his market environment, for example, the price at which he can sell. This effect may not be known (or known with certainty) and will therefore be the subject of *conjecture*. A conjecture differs from expectations concerning future market environments which may, say, be generated by some stochastic process. It is concerned with responses to the actions of the agent. In the first instance then the topic of conjectural equilibria is that of an economy which is not perfectly competitive by virtue of satisfying a no surplus condition. But, as we shall see, an economy could fail to

satisfy this condition and yet have a perfectly competitive equilibrium.

By an equilibrium in economics we usually mean an economic state which is a rest (critical) point of an (implicit) dynamic system. For instance, it is postulated in the textbooks that, when at going prices the amount agents wish to buy does not equal the amount they wish to sell, prices will change. Strictly this should mean that there would, in such a situation, be an incentive for some agent(s) to change prices. This causes difficulties when the economy is perfectly competitive (Arrow, 1959) since it implies that the agent can influence his market environment by his own actions. That is one reason why a fictitious auctioneer has been introduced to account for price changes.

When the economy is not perfectly competitive these difficulties are avoided. A price will be changed if some agent conjectures that such a change would be to his advantage. As a corollary then a conjectural equilibrium must be a state from which it is conjectured by each agent that it would be disadvantageous to depart by actions which are under the individual agent's control. (For a formal definition see below.)

But there are other difficulties. In particular, there is the question of the source of conjectures. If these are taken as given exogenously then there are many states which could be conjectural equilibria for *some* conjectures. It should be noted that a similar objection can be raised in conventional equilibrium analysis. There it is the preferences of agents which are taken as exogenous and there too there are many equilibria which are compatible with some (admissible) preferences. However, while conjectures may turn out to be false and this may occasion a change in conjectures, it is less easy to point to equally simple and convincing endogenous mechanisms of preference change. For that reason one may feel that conjectural equilibrium requires that conjectures are in some sense correct (‘rational’). For if they are not they will change in the light of experience. This argument is considered below.

The reason why the idea of conjectural equilibria is of interest is that economies which are not intrinsically perfectly competitive (for example, because of the large number of agents) are of interest and because it allows one to study price formation without an auctioneer.

An illustration

Consider two agents each of whom can chose an action a_i from a set of action A_i. Let $A = A_1 \times A_2$ with elements $a = (a_1, a_2)$. Then a conjecture c_i is a map from $A \times A_i$ to A_j written as

$$C_i = \theta_i(a, a_i').$$

Its interpretation is this: given the actions of the two agents (a), C_i is the action of *j* conjectured by *i* to be result from his choice of a_i'. (In a more general formulation the conjecture can be a probability distribution but that is not considered here.) We require conjectures to be *consistent*:

$$\theta_i(a, a_i) = a_j$$

(1)

This says that if agent *i* continues in his action a_i then he conjectures that *j* will do likewise. (This use of the word ‘consistent’ is *not* that of Bresnahan, 1981, and others who use it to mean ‘correct’.)
Suppose now that there is a function *v* from *A* to R^2, written as $v(a) = [v_1(a), v_2(a)]$, which gives the payoffs to the agents as a function of their joint action *a*. Consider a^* to be one such joint action. One says that a^* is a *conjectural equilibrium* for the two agents if

$$v_i\left[a_i, \theta_i(a^*, a_i)\right] \leq V_i\left[a_i^*, \theta_i(a^*)\right] \text{all } a_i \in A_i, i = 1, 2$$
(2)

That is, the joint action a^* is a conjectural equilibrium if no agent, given his conjecture, believes that he can improve his position by deviating to a different action.
It is not the case that conjectural equilibrium, as defined, always exists. For instance in the case of a duopoly in a homogeneous product where the action is ‘setting the price’, *v* may not be concave and a sensible conjecture may have discontinuities. One thus needs special assumptions to ensure existence or one must face the possibility that agents do not chose actions but probability distributions over actions (mixed strategies); for example, Kreps and Wilson (1982) in their work on sequential equilibrium employ conjectures which are probability distributions.
Supposing that a conjectural equilibrium exists, one may reasonably argue that until conjectures are less arbitrarily imposed on the theory not much has been gained – almost any pair of actions could be a conjectural equilibrium. A first attempt to remedy this is to ask that conjectures be correct (rational). If that is to succeed in any simple fashion it will be necessary to suppose that each agent has a unique best action under this conjecture. This is very limiting and it means that some of the classical duopoly problems cannot be resolved in this way.
Let the status quo again be a^*. Then if θ_1^* and θ_2^* are correct conjectures it must be that

$$v_2\left\{\theta_1^*(a^*, a_2), \theta_1^*\left[(a_1, a_1^*), \theta_2^*(a^*, a_2)\right]\right\} > v_1\left\{a_s', \theta_1^*\left[(a_1^*, a_2), a_1'\right]\right\} \text{all } a_1' = A_2.$$
(3)

$$v_1\left\{\theta_2^*(a^*, a_2), \theta_1^*\left[(a_1^*, a_2), \theta_2^*(a^*, a_2)\right]\right\} > v_1\left\{a_5', \theta_1^*\left[(a_1^*, a_2), a_1'\right]\right\} \text{all } a_1' = A_1. \tag{4}$$

A *rational conjectural equilibrium* is then a conjectural equilibrium a^* (with conjectures $\theta_1^*(\cdot), \theta_2^*(\cdot)$ which satisfy (3) and (4)). It must be re-emphasized that such an equilibrium may not exist for some *A* and *v* (see Gale, 1978; Hahn, 1978).

However, the idea is simple and, where applicable, coherent. It has however been criticized (in a somewhat intemperate and muddled paper) by Makowski (1983). This criticism appears to have had some appeal to some game theorists who like to think of games in extensive form (which they sometimes like to call dynamic). The criticism is this: when agent 1 deviates from a^* he is interested in the payoffs which he will get given this deviation and agent 2's response. This payoff Makowski thinks of as accruing in the ‘period’ after agent 1's deviation. But when agent 2 responds in that period he is interested in this payoff in the period following this response. So the agents expect ‘the game to end’ in different periods (Makowski, 1983, p. 8). Moreover, after agent 2 has responded, agent 1, in his turn, will again want to respond, that is, deviate from the deviation he started with. This criticism is then illustrated with an example in which one agent expects the other to return to the status quo *after* he has deviated from it.

All of this, however, is wrong. Firstly, if one wants to give a time interpretation to conjectures and so forth, then actions must be thought of as strategies. That is, the deviating agent deviates in one or more elements of his plan over the whole length of the game (perhaps infinite). Under correct conjectures responses and counter-responses are taken into account in evaluating the benefits of deviation. Hence, and secondly, a deviating agent is in this situation never surprised by the response of the other, which therefore does not lead him to further revise his deviation. On the definition, agent 1 expects the response to his deviation to be $\theta_1(a^*, a_1)$. Suppose this gives a_2 which is correct. Then that agent knows that the new status quo will be $(a_1, a_2) = a$ and if he has calculated benefits correctly he will not wish to deviate again.

However, there is the following to be said in favour of Makowski's criticism. Deviations in strategies may not be observable by the other agent. Therefore in traditional duopoly models with a sequential structure the re-interpretation of actions as strategies may be inappropriate. There is some evidence that in the duopoly literature with conjectures the consequent difficulties have not always been appreciated. It is also the case that too little attention has been paid to the assumption of a unique best response on which the above formulation depends.

An alternative to rational conjectures are *reasonable conjectures* (Hahn, 1978). A conjecture is reasonable if acting on any other conjecture would lower profits given the conjectures of other firms. Suppose that $\bar{\theta}$ is the set of all possible consistent conjectures. For any $\theta_i \in \bar{\theta}$, assume that there is a unique optimizing choice of output by firm *i* of $V_i(\theta_i)$. Then *i*'s conjecture $\theta_i^0 \in \theta$ is reasonable if given *j*th conjecture θ_j:

$$v_i\left[y_i(\theta_i^0),\ y_j(\theta_j)\right] \geq \hat{v}_i\left[y_i(\theta_i'),\ y_j(\theta_j)\right] \text{all } \theta_i' \in \theta. \quad (5)$$

But then a *reasonable conjectural equilibrium* is a pair $(\theta_1^0,\ \theta_2^0)$ each in θ such that

$$v_i\left[y_i(\theta_i^0),\ y_j(\theta_j^0)\right] \geq \hat{v}_i\left[y_i(\theta_i'),\ y_j(\theta_j^0)\right],\ i,\ j = 1,\ 2,\ \theta_i' \in \bar{\theta}. \quad (6)$$

This is just a Nash equilibrium where conjectures are interpreted as strategies (Hart, 1982).
While this is still quite demanding, it is significantly weaker than (3). If equilibria exist they may be ‘bootstrap equilibria’, that is, they will depend on beliefs about the actions of others, which beliefs may be incorrect. There is certainly no ground for believing that they will be efficient.
One can go one step further in the direction of plausibility by requiring that conjectures be reasonable only for small, or infinitesimal, deviations from the status quo. After all, large experiments are likely to be costlier than small ones. This will allow a larger class of reasonable conjectures and equilibria.

General conjectural equilibrium

It is fair to say that at present general equilibrium theory is in some way complete only for a perfectly competitive economy, that is, one where the returns to an individual agent are just equal to the contribution which he makes (Makowski, 1980; Ostroy, 1980). In general (although there are exceptions) such an economy exists when it is large (for example, it consists of a non-atomic continuum of agents). But there is now another possibility: an economy can be perfectly competitive if agents conjecture that their market actions will have no effect on the prices at which they can trade.
The following assertion will be clear from what has already been discussed. Let us say that an economy is *intrinsically* perfectly competitive if it satisfies the *no-surplus condition*. Then perfectly competitive conjectures are rational if an economy is intrinsically perfectly competitive. But perfectly competitive conjectures can be reasonable even when the economy is not intrinsically perfectly competitive. That is, conjectures may be such that, if an agent acts on any conjecture other than the perfectly competitive one, his profits will be lower. For instance, this may even be the case for two duopolists with constant marginal costs whose conjectures refer to the price charged by the rival firm. It will also be clear that if we do not require conjectures to be either reasonable or rational then, in general, conjectures can be found to support a competitive equilibrium in an economy which is not intrinsically perfectly competitive.
In a general equilibrium context it is not clear what it is that firms are supposed to conjecture. In some sense the conjecture must refer to the reaction of the whole economy to the action of the conjecturing

agent. In other words, it is not obvious how to define a game which adequately represents the economy. But in what sense?

Consider an economy with *n* produced goods and *m* non-produced goods. For simplicity suppose that all firms are single-product firms and that all firms producing the same good are alike, including their conjectures. There are very many households whose reasonable conjectures are always the competitive one. Households receive the profits of firms. Since the action of any one firm can affect the prices at which households can trade it is not at all clear what it is in the households' interest that the firms should maximize (Gabszewicz and Vial, 1972). If all households are alike it could be their common utility function, but that seems far removed from the world. I shall arbitrarily assume that firms maximize their profits in terms of one of the non-produced goods, say the first. This is arbitrary but it seems to me equally dubious to suppose that firms always choose in the ‘best interests of shareholders’, especially when that interest is often difficult and sometimes impossible to define.

Let $p \in R^n$, $w \in R_+^{m-1}$ be the price vectors in terms of good *m* of produced and non-produced goods respectively (so $w_m \equiv 1$). Let $y_j \in Y_j \subset R^{n+m}$ be the production of firm *j* where $y_{ij} > 0$ is its output of good j, $y_{ii} < 0$ is an input of good *i*, produced or non-produced. Let $y = \Sigma y_j$, where $y_j \in Y_j$ all *j*. Let $z \in R_+^m$ be the endowment of non-produced goods and

$$F = \{ y | y \geq (0, -z) \}$$

so that *F* is the set of feasible net production vectors *Y*. Let θ_{hj} be the share of household *h* in firm *j*. Given any $y \in F$ we think of each household as endowed with a certain strictly positive stock of non-produced goods and $\theta_{hj} y_j$ of the production of firm *j*. To avoid unnecessary complications assume θ_{hj} $(j = 1, \ldots, n)$ to be such that if z_h is the stock of non-produced goods owned by household *h*:

$$\text{For all } y \in F\text{: } z_h + \sum_j \theta_{hj} y_j \geq 0 \text{ all } h. \tag{7}$$

Households consume both types of goods. Hence for any $y \in F$ there is now an associated pure exchange economy where each household's endowment is given by (7). Making the usual assumptions there will exist at least one equilibrium $[p(y), w(y)]$. Suppose for the moment that there is only one for each $y \in F$. Now firm *j* in this equilibrium observes $[p(y), w(y)]$ and will deviate from y_j (if it deviates at all) if it can thereby increase its conjectured profits. Let

$$\hat{\pi}_j\left[p(y), w(y), y_j'\right]$$

be the conjectural profit function of firm *j*. Then y^0, $p(y^0)$, $w(y^0)$ is a conjectural equilibrium if for all $j = 1, \ldots, n$:

$$\hat{\pi}_j\left[p(y^0), w(y^0), y_j^0\right] \geq \hat{\pi}_j\left[p(y^0), w(y^0), y_j'\right] \text{all } y_j' \in Y. \tag{8}$$

Such a conjectural equilibrium will exist if all $\hat{\pi}_j(\cdot)$ are quasi-concave, an assumption for which there is scant justification (Hahn, 1978).

If we demand that conjectures be rational then conjectured and actual profit must coincide for all y_k' (the two coincide for $y_k' = y_k^0$ by the requirement that conjectures be consistent). One proceeds as follows. Let $y_k' = y_k^0$. Given the conjectures of the remaining firms find the conjectural equilibrium of the economy $p\left\{y^*(k), w\left[y^*(k), y^*(k)\right]\right\}$, where *y*(*k*) is the vector *y* with y_k' in the kth place and condition (8) is not imposed for firm *k*. One then requires that for all $y_k' > 0$

$$\hat{\pi}_k\left[p(y^0), w(y^0), y_k'\right] = \pi_k\left\{p\left[y^*(k)\right], w\left[y^*(k)\right], y_k'\right\} \tag{8a}$$

where $\pi_k(\cdot)$ is actual profit. For rational conjectures this should be true for all *k*.

It will be seen that rational conjectural equilibrium is very demanding. For a certain class of conjectures it will not even exist (Gale, 1978; Hahn, 1978). More importantly, the whole procedure breaks down if given a deviation by *k*, the conjectural equilibrium, is not unique. Lastly, even if by sufficient assumptions one overcomes these difficulties, it is not agreeable to common sense to suppose that firms can correctly calculate general equilibrium responses to their actions, nor is it obvious that they should always be concerned only with equilibrium states.

Reasonable conjectures do not fare much better, although a notable contribution to their study has recently been made by Hart (1982). Hart notices that conjectures of firms induce a supply correspondence (not generally convex) on their part. Here let us suppose that we can in fact speak of supply functions. These can be thought of as strategies in a manner already discussed. A reasonable

conjectural equilibrium then satisfies the condition that, given the supply functions of other firms, no deviation by firm *k* to another supply response can increase its profits. In (8) one then substitutes on the right-hand side for $y_j^i, \eta_j^i\left[p(y^0), w(y^0)\right]$, an admissible supply function (see Hart, 1982) of *j* and requires the inequality to hold for all such functions. Of course, one has

$$y_j^0 = \eta_j^0\left[p(y^0), w(y^0)\right]$$

for a reasonable conjectural equilibrium.

To show existence of such an equilibrium will require strong assumptions. The technicalities will be found in Hart (1982). However, one of the assumptions which he makes is not only technically useful but economically sensible since it leads firms to face a simpler task in forming conjectures. Hart supposes the economy to consist of a number of islands each of which has many consumers and one firm of each type $(j = 1, \ldots, n)$. The islands are small replicas of the whole economy. But households have shares in firms on all islands so that if there are enough islands their share in any firm on their own island is very small. That means that any firm can disregard the effect of a change in its own profits on the demand for the good it produces. To make this work one supposes that produced goods are totally immobile between islands while non-produced goods are totally mobile. By an appropriate assumption on consumers on each island one ensures that they all have the same demand. Lastly, since shares in a firm are held on many different islands the firm, in acting in the shareholder's interest is justified in neglecting the effect of its actions on relative prices on its own island and so is justified in maximizing profits.

From the point of view of conjectural equilibrium the island assumption allows firms (both reasonably and rationally) to ignore effects of their own actions on *w* – the price vector of non-produced goods. These will be determined by demand and supply over all islands and in this determination any one firm can be regarded as playing a negligible role. This is some gain in realism. But after all allowances have been made it is still true that (*a*) the assumptions required for the existence of reasonable conjectural equilibrium are uncomfortably strong and (*b*) even when that is neglected such an equilibrium seems to have small descriptive power.

Simpler approaches

Negishi (1960) made the first, justifiably celebrated, attempt to incorporate imperfect competition in general equilibrium analysis. He did this by letting single product firms have consistent inverse demand conjectures (the case he studies most thoroughly makes these linear). Consistency is all he asked for of conjectures but he also needed the uncomfortable postulate that the resulting conjectural profit functions be quasi-concave. Later Hahn (1978), Silvestre (1977) and others added the requirement that, besides being consistent, the conjectured demand functions have, if differentiable, the correct slope at equilibrium (that is, that the conjecture be *infinitesimally* or ‘first order’ *rational*). It turns out that this

extra requirement does not much restrict conjectures, nor thus the set of equilibria which can be generated by some conjectures. The reason roughly is this: in conjectural equilibrium, when conjectured profit functions are twice differentiable, the partial derivative of the conjectured profit function of firm *j* with respect to its own output much vanish. Suppose the economy to be in such an equilibrium and consider an infinitesimal output deviation by firm *k*. To find the equilibrium which ensues, differentiate all equilibrium relations, other than that for firm *k*, with respect to the output of firm *k*. Amongst these will be the condition that the marginal profit conjectured of every firm (other than *k*), be zero. Hence differentiation of that condition will yield second-order terms. But we can choose these arbitrarily since we are requiring only first-order rationality. One can show in fact that these second-order terms can be chosen so as to make the first-order conjectured change in profit of any firm *k* correspond to the actual change. (Details in Hahn, 1977.) Hence first-order rationality imposes few restrictions.

Both Hahn (1978) and Negishi (1979) have also considered kinked conjectures. The idea is this. If an agent can transact at the going price as much as he desires his conjectures are competitive. If he is quantity constrained (for example, if a firm cannot sell an amount determined by equality between marginal cost and price) his conjectures are non-competitive. That is, he considers that a price change is required to relax the quantity constraint. The fixprice methods of Drèze (1975) and others can be interpreted as an extreme form of such conjectures – for instance to relax a constraint on sales, price, it is conjectured, must be reduced to zero.

To such conjectures there have been two objections. Firstly, they assume that an agent's conjectures are not influenced by constraints on others. For instance, a firm which can hire as much labour as it wants at the going wage while workers cannot sell as much as they like does not conjecture that it could have the same amount of labour at a lower wage. To this one can answer that it is not easy for an agent to observe the quantity constraints on others. For instance, unemployment statistics do not tell us whether workers have chosen not to work or whether they are constrained in their sale of labour. None the less, this objection has some force and needs further study with proper attention to the information of agents.

The other objection is that these kinked conjectures are not explained. That is true if explanation turns on what an agent knows or can learn. None the less, the hypothesis seems to be to have psychological verisimilitude. If I can always sell my labour at the going wage there is little occasion for the difficult conjecturing of what would happen if I raised my wage. This is not so if I find that I cannot find employment at the going wage.

In any event these simpler approaches allow one to incorporate traditional monopolistic competition in a general equilibrium framework. Of course, some of the assumptions such as concave conjectured profit functions are strong. On the other hand, one can now allow for a certain amount of increasing returns (Silvestre, 1977).

Some conclusions

The conjectural approach has this merit: it takes proper and explicit note of the perceptions by individuals of their market environment. Economic theory perhaps too often neglects the possibility that what is the case may depend on what agents believe to be the case. Historians and others have long since studied the intimate mutual connection between beliefs and events but economists have not made much headway here. The conjectural approach is perhaps a small beginning. For it deals with the theories

agents hold and this must plainly enter into our theory of agents.

In particular one should not pay too much attention to the objection that conjectures may not be derivable from some first principles of rationality. It seems to me quite proper to find their description in history. Nor, as has been argued, will an appeal to learning render conjectures in some sense objectively justifiable. This is clear from the discussion of reasonable conjectures and from the costs of experimentation. For hundreds of years witches were burned in the light of a reasonable theory which few would now regard as having proper objective correlatives. There is no reason to suppose that it is possible for businesses or governments now to do better than some of the best minds of the past.

From a more immediately relevant standpoint, conjectural theories are of interest because they attempt a general equilibrium analysis of non-perfect competition. It is good to know that in a proper sense perfectly competitive economies can be viewed as limiting Cournot conjectural equilibrium economies (Novshek and Sonnenschein, 1978). But this knowledge does not contribute to the study of properly imperfectly competitive economies. Again the study of fixprice equilibria has borne some fruits, but not those which were first sought by Triffin (1940) when he proposed a framework for general equilibrium with monopolistic competition. If it is the case that actual economies are not perfectly competitive nor that they behave ‘as if’ they were, then the task set by Triffin requires serious attention, and it is likely that conjectural theories will have a role to play.

Recent developments in game theory (for example, Kreps and Wilson, 1982) suggest that these two conjectures will have to play a part. Indeed, quite generally in that theory players conjecture that their opponent is ‘rational’ in an appropriate sense. It is not the case that the conjectural equilibrium approach is an alternative to the game theoretic one.

See Also

- auctioneer

Bibliography

Arrow, K.J. 1959. Toward a theory of price adjustment. In M. Abramovitz et al., *The Allocation of Economic Resources*. Stanford: Stanford University Press.

Bresnahan, T.F. 1981. Duopoly models with consistent conjectures. *American Economic Review* 71, 934–45.

Drèze, J. 1975. Existence of equilibrium under price rigidity and quantity rationing. *International Economical Review* 16, 301–20.

Gabszewicz, J.J. and Vial, J.D. 1972. Oligopoly ‘à la Cournot’ in general equilibrium analysis. *Journal of Economy Theory* 4, 381–400.

Gale, D. 1978. A note on conjectural equilibria. *Review of Economic Studies* 45(1), 33–8.

Hahn, F.H. 1977. Exercise in conjectural equilibria. *Scandinavian Journal of Economics* 79, 210–26.

Hahn, F.H. 1978. On non-Walrasian equilibria. *Review of Economic Studies* 45, 1–17.

Hart, O. 1982. Reasonable conjectures. Theoretical Economics Paper No. 61. STICERD, London School of Economics.

Kreps, D.M. and Wilson, R.B. 1982. Sequential equilibria. *Econometrica* 50, 863–94.

Makowski, L. 1980. A characterisation of perfectly competitive economies with production. *Journal of Economic Theory* 22, 208–21.

Makowski, L. 1983. ‘Rational conjectures’ aren't rational and ‘reasonable conjectures’ aren't reasonable. Discussion Paper No. 60. SSRC Project on Risk, Information and Quantity Signals, Cambridge University.

Negishi, T. 1960. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28, 196–202.

Negishi, T. 1979. *Micro-Economic Foundations of Keynesian Macro-Economics*. Amsterdam: North-Holland.

Novshek, W. and Sonnenschein, H. 1978. Cournot and Walras equilibrium. *Journal of Economic Theory* 19, 223–66.

Ostroy, J. 1980. The no-surplus condition as a characterisation of perfectly competitive equilibrium. *Journal of Economic Theory* 22, 183–207.

Silvestre, J. 1977. A model of general equilibrium with monopolistic behavior. *Journal of Economic Theory* 16, 425–42.

Triffin, R. 1940. *Monopolistic Competition and General Equilibrium Theory*. Cambridge, MA: Harvard University Press.

Ulph, D. 1983. Rational conjectures in the theory of oligopoly. *International Journal of Industrial Organization* 1(2), 131–54.

How to cite this article

Hahn, F. "conjectural equilibria." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of

Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000296> doi:10.1057/9780230226203.0293

The New Palgrave Dictionary of Economics Online

conspicuous consumption

F. Stankovi•

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

conspicuous consumption; McCulloch, J. R.; Rae, J.; Say, J.-B.; Scarcity; Smith, A.; vanity; Veblen effect; Veblen, T.

Article

Conspicuous consumption means the use of consumer goods in such a way as to create a display for the purpose of impressing others rather than for the satisfaction of normal consumer demand. It is consumption intended chiefly as an ostentatious display of wealth. The concept of conspicuous consumption was introduced into economic theory by Thorstein Veblen (1899) in the context of his analysis of the latent functions of ‘conspicuous consumption’ and ‘conspicuous waste’ as symbols of upper-class status and as competitive methods of enhancing individual prestige.

Veblen argued that the leisure class is chiefly interested in this type of consumption, but that, to a certain degree, it exists in all classes. The leisure class undoubtedly has much more opportunity for this kind of consumption. The criterion as to whether a particular outlay fell under the heading of conspicuous consumption was whether, aside from acquired tastes and from the canons of usage and conventional decency, its result was a net gain in comfort or in fullness of life.

It is widely though that Veblen introduced the concept of conspicuous consumption into economic literature, but it was known much earlier. Adam Smith (1776, Book I, ch.11) wrote about people who like to possess those distinguishing marks of opulence that nobody but themselves can possess. In the eyes of such people the merit of an object that is in any degree either useful or beautiful is greatly enhanced by its scarcity, or by the great amount of labour required to accumulate any considerable quantity of it. This is the labour for which nobody but themselves can afford to pay. Smith concluded that this domain was ruled by fashion. J.-B. Say and McCulloch wrote about this issue in a similar way. But the author who first used the term ‘conspicuous consumption’ was the Canadian economist John Rae (1796–1872). His explanation of the nature and effects of luxury was based on the meaning of vanity in human life. He understood vanity to be the mere desire for superiority over others without any reference to merit. The aim is to have what others cannot have, whereas the stimulus to productivity in

economic life is the passion for effective accumulation: ‘Articles of which consumption is conspicuous, are incapable of gratifying this passion’ (Rae, 1834).

However, it was Veblen who introduced the concept of conspicuous consumption as a phenomenon important for the understanding of consumption as a whole. He gave Rae no reference at all. Veblen's historical and socio-economic explanation of this institution gave as a result the so-called ‘Veblen effect’. This is the phenomenon whereby as the price of an article falls some consumers construe this as a reduction in the quality of the good or loss of its ‘exclusiveness’ and cease to buy it.

See Also

- Rae, John
- Veblen, Thorstein Bunde

Bibliography

Mason, R.S. 1981. *Conspicuous Consumption: A Study of Exceptional Consumption Behaviour*. New York: St Martin's Press.

Rae, J. 1834. *The Sociological Theory of Capital*, ed. C. Mixter. New York: Macmillan.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. E. Cannan, London: Methuen, 1981.

Sweezy, P. 1952. Veblen and Marx. In *Socialism and American Life*, 2 vols. ed. D.D. Egbert and S. Persons. Princeton: Princeton University Press.

Veblen, T. 1899. *The Theory of the Leisure Class*. London: George Allen & Unwin.

How to cite this article

Stankovi•, F. "conspicuous consumption." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000301> doi:10.1057/9780230226203.0294

The New Palgrave Dictionary of Economics Online

constitutions, economic approach to

Dennis C. Mueller
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economic approach to constitutions applies the methodology of economics to the study of constitutions. This entry reviews the normative literature on constitutions, which assumes a two-stage collective decision process, and the positive literature that examines the decisions made by constitutional conventions and their economic consequences.

Keywords

Beard, C.; Bentham, J.; Buchanan, J.; budget deficits; collective choice; constitutionalism normative vs positive;; corruption; First Amendment; Harsanyi, J.; Landes, W.; negative externalities; party systems; Philadelphia Convention; Posner, R.; Rawls, J.; rent seeking; social contract theory; social welfare function; Tullock, G.

Article

The economic approach to constitutions applies the methodology of economics to the study of constitutions, just as public choice applies this methodology to the full range of topics of political science.
The economic approach to constitutions began with *The Calculus of Consent* by James Buchanan and Gordon Tullock (1962, hereafter B&T). Theirs was largely a *normative* analysis of what ought to go into a constitution. Their main findings and the literature that grew out of their work are reviewed first, after which the *positive* stream of the literature is discussed.

Normative research on constitutions

Arguably the most important contribution of *The Calculus* was to view democracy as a two-stage process. In stage one, institutions to make future collective decisions are placed into the constitution. In

stage two, collective decisions are made using these rules. The long-run nature of the choices at the first stage creates considerable uncertainty about the consequences of different voting rules. This uncertainty makes unanimous agreement on the rules of the political game likely, even though individuals would disagree in stage two about the outcomes of the game. This unanimity at the constitutional stage provides the normative underpinning for the constitution (B&T, p. 7). Harsanyi (1955) also used uncertainty over future positions to produce unanimity and to provide a normative argument for a Benthamite social welfare function (SWF), as did Rawls (1971) in his ethical theory of a social contract. Mueller (1973) discussed conditions under which a B&T constitution maximizes a Harsanyian SWF.

Another innovation in *The Calculus* was to introduce the *external costs of collective decisions* (B&T, pp. 63–8). When a collective choice is made without the consent of all members of the community, the decision can make some members worse-off. The votes of those favouring the decision thus impose a negative externality on those opposing it. The smaller the majority required to pass an issue, the more likely it is that an individual is on the losing side. However, the amount of time required to make a collective decision is also likely to increase with the required majority. The optimal majority minimizes the sum of collective decisions' external and decision-making costs.

There is nothing in B&T's costs-minimization-approach that implies that the optimal majority is likely to be a simple majority, and thus their approach does not account for this rule's ubiquitous use. The approach does imply the widespread use of the simple majority rule, if one of the two cost curves – most plausibly decision-making costs – has a sharp discontinuity at 50 per cent (Mueller, 2003, pp. 76–8).

Rae (1969) used the two-stage approach to provide a completely different normative justification for the simple majority rule. At the constitutional stage, each individual is uncertain of whether he will favour x or $\sim x$ in future votes on these binary issues. The expected gain if an individual favours x and x wins equals the expected loss if x wins and the individual favours $\sim x$. Rae further assumed that the probability of favouring x equals the probability of favouring $\sim x$. An egoist chooses the voting rule that minimizes the probability that she favours x in the future and $\sim x$ is imposed, or that she favours $\sim x$ and x is imposed. The simple majority is the only rule satisfying this condition. (For additional discussion and references see, Rae and Schickler, 1997.)

Mueller (2001) generalized the two-stage approach to show that the optimal majority for binary choices depends on the relative payoffs from the two issues. (Riley, 2001, presents a game theoretic analysis of a two-stage constitutional process.) As the loss to those favouring x rises relative to the gain to those favouring $\sim x$, higher required majorities become optimal to implement $\sim x$, with unanimity being optimal when the asymmetry in payoffs is very large. Mueller (1991; 1996, ch. 14) employed this analysis to explain why placing rights to act into a constitution would maximize the expected utilities of those writing it.

Positive research on constitutions

The positive literature of constitutions falls into two categories: studies of constitutional conventions and of the consequences of constitutions. The second category is obviously very large, and so I provide only the flavour of this type of work.

Charles Beard's work (1913) might well be regarded as the first *economic* analysis of the Philadelphia

Convention. Beard stressed the self-interest of the participants, and claimed that the final product reflected the interests of the landowning aristocracy. In an equally cynical analysis, Landes and Posner (1975, p. 893) claimed that the First Amendment was a result of pressure from ‘publishers, journalists, pamphleteers, and others who derive pecuniary and nonpecuniary income from publication and advocacy of various sorts’. Case studies of constitutional conventions confirm the importance of the self-interest of the participants in determining the constitution's content. For example, representatives from small parties favour rules that produce proportional representation and low percentage thresholds for taking seats in the parliament. Representatives from large parties favour the reverse. If delegates are selected geographically, the constitution protects geographic interests. (For further discussion and references to the literature, see Elster, 1991, and Mueller, 1996, ch. 21). Econometric analyses confirm these findings. McGuire and Ohlsfeldt (1986) and McGuire (1988) concluded that the votes of delegates to the Philadelphia convention reflected both their personal interests and those of their constituencies. Eavey and Miller (1989) reached the same conclusion from the voting patterns of those who ratified the Pennsylvania and Maryland constitutions.

A key decision facing any constitutional convention is whether to design institutions that will produce a two-party system or a multiparty system. In practice, this choice appears to rest upon the number of representatives elected from each electoral district (Taagepera and Shugart, 1989; Lijphart, 1990; Mueller, 1996, chs. 8–10). Recent theoretical and empirical work by Persson and Tabellini (1999; 2000; 2003; 2004a; 2004b) and Persson, Roland and Tabellini (2000) demonstrates the economic importance of this choice. They find more rent seeking, more corruption, more redistribution and larger deficits in multiparty systems. Presidential systems lead to smaller governmental sectors because they generally contain stronger checks and balances than parliamentary systems. (For a review and references to other contributions, see Persson and Tabellini, 2004a.)

Conclusions

There are two kinds of people in the world: those who believe that constitutions matter and those who do not. The contributors to the literature reviewed here fall into the former category. Their work helps illustrate *why and in what way* constitutions matter, and further illustrates the fruitfulness of undertaking an economic approach to the study of constitutions.

See Also

- Buchanan, James M.
- collective rationality

Bibliography

Beard, C. 1913. *An Economic Interpretation of the Constitution of the United States*. New York: Macmillan, 1941.

Buchanan, J. and Tullock, G. 1962. *The Calculus of Consent*. Ann Arbor: University of Michigan Press.

Eavey, C. and Miller, G. 1989. Constitutional conflict in state and nation. In *The Federalist Papers and the New Institutionalism*, ed. B. Grofman and D. Wittman. New York: Agathon Press.

Elster, J. 1991. Arguing and bargaining in two constituent assemblies. Mimeo. Storrs Lectures, Yale Law School.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economics* 63, 309–21.

Landes, W. and Posner, R. 1975. The independent judiciary in an interest-group perspective. *Journal of Law and Economics* 18, 875–901.

Lijphart, A. 1990. The political consequences of electoral laws, 1945–85. *American Political Science Review* 84, 481–96.

McGuire, R. 1988. Constitution making: a rational choice model of the Federal Convention of 1787. *American Journal of Political Science* 32, 483–522.

McGuire, R. and Ohlsfeldt, R. 1986. An economic model of voting behavior over specific issues at the Constitutional Convention of 1787. *Journal of Economic History* 46, 79–111.

Mueller, D. 1973. Constitutional democracy and social welfare. *Quarterly Journal of Economics* 87, 60–80.

Mueller, D. 1991. Constitutional rights. *Journal of Law, Economics, and Organization* 7, 313–33.

Mueller, D. 1996. *Constitutional Democracy*. Oxford: Oxford University Press.

Mueller, D. 2001. The importance of uncertainty in a two-stage theory of constitutions. *Public Choice* 108, 223–58.

Mueller, D. 2003. *Public Choice III*. Cambridge: Cambridge University Press.

Persson, T., Roland, G. and Tabellini, G. 2000. Comparative politics and public finance. *Journal of Political Economy* 108, 1121–61.

Persson, T. and Tabellini, G. 1999. The size and scope of government: comparative politics with rational politicians. *European Economic Review* 43, 699–735.

Persson, T. and Tabellini, G. 2000. *Political Economics – Explaining Economic Policy*. Cambridge, MA: MIT Press.

Persson, T. and Tabellini, G. 2003. *Economic Effects of Constitutions*. Cambridge, MA: MIT Press.

Persson, T. and Tabellini, G. 2004a. Constitutions and economic policy. *Journal of Economic Perspectives* 18, 75–98.

Persson, T. and Tabellini, G. 2004b. Constitutional rules and fiscal policy outcomes. *American Economic Review* 94, 25–45.

Rae, D. 1969. Decision-rules and individual values in constitutional choice. *American Political Science Review* 63, 40–56.

Rae, D. and Schickler, E. 1997. Majority rule. In *Perspectives on Public Choice*, ed. D. C. Mueller. Cambridge, MA: Cambridge University Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press.

Riley, J. 2001. Constitutional democracy as a two-stage game. In *Constitutional Culture and Democratic Rule*, ed. J. Ferejohn, J. Rakove and J. Riley. Cambridge: Cambridge University Press.

Taagepera, R. and Shugart, M. 1989. *Seats and Votes*. New Haven: Yale University Press.

How to cite this article

Mueller, Dennis C. "constitutions, economic approach to." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000212> doi:10.1057/9780230226203.0295

The New Palgrave Dictionary of Economics Online

consumer expenditure (new developments and the state of research)

Orazio P. Attanasio and Guglielmo Weber

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We provide an overview of recent developments in the life-cycle permanent income model under uncertainty, starting from the certainty equivalence case, and considering precautionary saving, the workings of insurance and credit markets, and non-standard preference structures.

Keywords

aggregate consumption; aggregate saving; asset pricing; asymmetric information; buffer stocks; certainty equivalence; commitment; consumer expenditure; consumer expenditure (new developments and the state of research); consumption function; durable goods; elasticity of intertemporal substitution; equity premium puzzle; Euler equation; excess sensitivity; excess smoothness of consumption; expected utility; financial asset accumulation; financial market imperfections; habits; impatience; interest rate; intertemporal optimization; intertemporal prices; labour supply; Lagrange multipliers; life cycle hypothesis; liquidity constraints; marginal utility of consumption; marginal utility of wealth; myopia; non-convexity; perfect insurance; permanent income life-cycle model; power utility; precautionary savings; preferences; private information; prudence; public finance; quadratic utility; quasi-hyperbolic discounting; real business cycles; reduced form equations; retirement; risk; risk aversion; saving; social insurance; Social Security in the United States; Stone–Geary utility; subjective discount rate; time preference; transaction costs; uncertainty

Article

The state of research on consumer expenditure up to the mid-1980s is described in consumer expenditure. Here, we provide an overview of recent developments on the intertemporal model of consumer behaviour under uncertainty. We organize our discussion around what has been the workhorse model for the analysis of dynamic consumption behaviour – the life-cycle permanent income model.

Although our discussion of the intertemporal model is self-contained, it is not meant to be an exhaustive survey of this large literature. We do not cover demand analysis, despite the many exciting developments that have occurred in recent years.

The permanent income life-cycle (PILC) model, introduced during the 1950s by Modigliani and Brumberg (1954) and Friedman (1957), still plays an important role in the consumption literature. The PILC model can be loosely defined as a framework where individuals maximize utility over time given a set of intertemporal trading opportunities. Consumption at different points in time is treated as different commodities, so that, given intertemporal trading opportunities, consumption in a given period depends on total (life-cycle) resources and (intertemporal) prices. Optimal consumption choices are such that the ratio of (expected) marginal utilities of consumption at different times equals the ratio of intertemporal prices. Therefore, the relationship between consumption and total resources is likely to depend on preferences (and in particular on the elasticity of intertemporal substitution and the rate at which the future is discounted) and on interest rates (as they represent intertemporal prices). If we allow for uncertainty, as we discuss below, risk will also enter as a potentially important determinant of consumption.

This model can generate implications and insights for many important questions not only in macroeconomics but also in public finance, and has therefore attracted much attention, both theoretically and empirically. Recent research has stressed the need to look at preferences on the one hand and markets on the other, as the policy implications are the result of both.

The permanent income life-cycle model

In its simplest incarnation the PILC model considers a finite horizon, no uncertainty and very simple preferences. In such a situation, it is simple to translate the basic intuition of the model, to which we referred above, into a closed form solution for consumption that depends not just on current income but on the total amount of resources available to an individual and intertemporal prices. The problem of this specification, of course, is its lack of realism. Not only do consumers in reality face much more complicated intertemporal environments, but it is likely that these complications have a first-order effect on consumption choices. Therefore, the simplest version of the model is a useful way to convey the main ideas behind PILC, but it needs to be complicated considerably to be of use for policy analysis.

The introduction of uncertainty in the model, which makes it much more realistic, complicates the problem enormously. The first formalizations of the life-cycle model under uncertainty date back to the 1970s (Bewley, 1977). Typically, one assumes that consumers maximize expected life-cycle utility choosing consumption and, in more general settings, leisure and financial asset holdings. Consumers are assumed to know the stochastic nature of their environment. Even with many simplifications on the nature of preferences, the model does not yield closed form solutions for consumption, except in the most special cases.

MaCurdy (1981; see also 1999) uses dynamic optimization techniques to derive necessary conditions for the optimal solution of the intertemporal optimization problem faced by consumers. The attractiveness of this approach lies in the fact that it cuts through the necessity of solving the model completely, which is a very hard task indeed, to focus on some useful implications of the model. In particular, these contributions focus on the basic first order condition, the so-called Euler equation, that equates the ratio

of marginal utilities to intertemporal prices.

The first macro paper to take this approach is Hall (1978): under strong assumptions on preferences and returns, (non-durable) consumption is a random walk, that is:

$$E(C_{t+1}|I_t) = C_t \tag{1}$$

where I_t denotes information available at time t. This remarkable proposition requires that utility be quadratic in consumption (and additively separable over time, states of nature and in its other arguments, notably male and female leisure and durable goods). It also requires that there is at least one financial asset with fixed real return, and that this equals the time-preference parameter. If consumers have rational expectations, then:

$$C_{t+1} = C_t + \varepsilon_{t+1} E(\varepsilon_{t+1}|Z_t) = 0 \tag{2}$$

for all variables Z known at time t. A notable feature of Hall's model is that the Euler equation for consumption aggregates perfectly, because it involves linear transformations of the data. Hall used the Euler equation to test for the prediction implied by (2): no variable known to the consumer at time t should help predict the change in consumption between t and (t+1).

Hall's paper was the first of many contributions that exploited the Euler equation and the fact that such an approach does not require the complete specification of the environment in which the consumer lives, or even the complete budget constraint. Moreover, the approach is robust to the presence of various imperfections in some intertemporal markets. And while the specification with quadratic utility yields a linear equation for consumption, alternative specifications, with more plausible preferences, are easily introduced. For instance, in the case of power utility, an expression similar to (2) can be obtained for the log of consumption.

The price that one pays in using the Euler equations approach, which we discuss below, is that one does not obtain a closed form solution for consumption. An approach that goes beyond the consideration of the Euler equations is taken up in an important paper by Flavin (1981).

Flavin (1981) adopts the same theoretical framework as Hall (1978), and assumes that no other asset is available to the consumer (as in Bewley, 1977). However, Flavin develops a solution for consumption. To do so, she has to specify completely the stochastic environment in which the consumer lives and use particularly simple preferences. In particular, Flavin (1981) assumes that the only stochastic variable is labour income, that preferences are quadratic and that the consumer can save or borrow in a single asset with a fixed rate of interest. Under these conditions, Flavin shows that consumption is set equal to

permanent income, and this is in turn defined as the present value of current and expected future incomes:

$$C_t = \frac{r}{1+r}A_t + \frac{r}{1+r}\sum_{k=0}^{\infty} E(y_{t+k}|I_t) \tag{3}$$

where *A* denotes financial wealth and *y* is labour income. In this model, the first difference in consumption equals the present value of income revisions, due to the accrual of new information between periods *t* and (*t*+1):

$$\Delta C_t = \frac{r}{1+r}\sum_{k=0}^{\infty}\frac{1}{(1+r)^k}\left[E(y_{t+k}|I_t) - E(y_{t+k}|I_{t-1})\right]. \tag{4}$$

Equation (3) makes clear the main implications of the model: consumption depends on the present discounted value of future expected income. The interest rate plays the important role of converting future resources to present ones and therefore constitutes an important determinant of consumption. Flavin (1981) noticed that eq. (3) imposes cross-equation restrictions on the joint time series process for income and consumption. A similar approach had been followed by Sargent (1978) and, subsequently, by Campbell (1987) who noticed that an implication of (4) is that saving predicts future changes in income, the so-called ‘saving for a rainy day’ motive.

One of the main implications of the PILC model, particularly evident in eq. (3), is that, in appraising the effects of a given policy, for instance a tax reform that affects disposable income, a distinction must be drawn between permanent and temporary changes (Blinder and Deaton, 1985; Poterba, 1988).

Another feature of Flavin's model is that the closed form solution for consumption is the same under certainty and uncertainty, as long as expected values of future incomes are taken. This is a direct consequence of the assumption of quadratic utility that makes the marginal utility linear in consumption. For this reason, it is often referred to as the certainty equivalent model.

Extensions of the simple certainty equivalent model

The certainty equivalent model is appealing for its simplicity, but its implications are typically rejected by the data: Hall and Mishkin (1982) were particularly influential in suggesting that some of the model implications were rejected in micro data. At the same time, the model with quadratic preferences was

perceived to be too restrictive in its treatment of financial decisions: quadratic preferences imply increasing absolute risk aversion in consumption (or wealth), something that is unappealing on theoretical grounds and strongly counterfactual (riskier portfolios are normally held by wealthier households). Quadratic preferences also imply that the willingness to substitute over time is a decreasing function of consumption: poor consumers should react much more to interest rate changes than rich consumers, after allowance has been made for the wealth/income effect.

The alternative adopted in much of the literature has been to assume power utility and to allow for the existence of a number of risky financial assets. Once one deviates from quadratic utility, however, and/or allows for stochastic interest rates, one loses the ability to obtain a closed form solution for consumption. Many of the studies that made this choice, therefore, focused on the study of the Euler equations derived from the maximization problem faced by the consumer. The basic first-order conditions used in this literature are two:

$$U_{ct} = \lambda_t, \tag{5}$$

$$\lambda_t = E(\lambda_{t+1} \frac{1 + r_{t+1}^k}{1 + \delta} | I_t). \tag{6}$$

Equation (5) says that, at each point in time, the marginal utility of consumption equals the Lagrange multiplier associated with the budget constraint relevant for that period, which is sometimes referred to as the marginal utility of wealth. The second condition, eq. (6), that is derived from intertemporal optimality, dictates the evolution of the marginal utility of wealth (δ is a subjective discount rate). An equation of this type has to hold for each asset *k* for which the consumer is not at a corner. This is because the consumer is exploiting that particular intertemporal margin.

The attractiveness of Euler equations is that one can be completely agnostic about the stochastic environment faced by the consumer, about the time horizon, about the presence of imperfections in financial markets (as long as there is at least one asset that the consumer can freely trade), about the presence of transaction costs in some component of consumption or labour supply. All relevant information is summarized in the level of the marginal utility of wealth. The approach is conceptually similar to the use of an (unobservable) fixed effect in econometrics. By taking first differences, one eliminates the unobservable marginal utility of wealth and is left only with the innovations to eq. (6).

Early papers along these lines were Hansen and Singleton (1982; 1983), who used power utility (also known as isoelastic, isocurvature or CRRA) as it has more appealing theoretical properties (relative risk

aversion is constant in wealth or consumption, the elasticity of intertemporal substitution is also a constant). If we substitute eq. (5) into (6) and using the properties of the CRRA utility function, the Euler equations for consumption corresponding to each asset (*k*) will be:

$$E\left\{\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\frac{1+r^k_{t+1}}{1+\delta}\right\}=1$$

(7)

where γ is a curvature parameter (equal to the relative risk aversion parameter and to the reciprocal of the elasticity of intertemporal substitution) and δ, the subjective discount rate, measures impatience. An equation such as (7) can be log-linearized to obtain (see Hansen and Singleton, 1983):

$$\Delta\log C_{t+1}=k+\frac{1}{\gamma}\log(1+r^k_{t+1})+\varepsilon_{t+1}.$$

(8)

Although consumption appears on the left-hand side of eq. (8), this equation is not a consumption function, but an equilibrium condition. It cannot explain or predict consumption levels: consumption is crucially determined by the residual term ε_{t+1} and there is nothing that tells us what determines such a term or how this term changes with news about income, interest rates or any other relevant variable.
The Euler equation for a single asset can identify the elasticity of intertemporal substitution, a key parameter for the evaluation of the welfare costs of interest taxation (Boskin, 1978; Summers, 1981) and for the analysis of real business cycles (King and Plosser, 1984). The joint estimation of several Euler equations can help identify the pure discount rate parameter (governing patience), but also shed light on risk aversion, given that different assets typically have different risk characteristics.
The derivation of a closed form solution for consumption when certainty equivalence does not hold was first successfully tackled by Caballero (1990; 1991). Caballero (1991) took the Flavin model with known finite life, and constant absolute risk aversion (CARA) preferences, and showed that, when the optimal consumption age profile is flat with no uncertainty, it is increasing with income uncertainty. This change in the slope of the consumption profile was described as precautionary saving, because early in life consumers save more if labour income is more uncertain. Later work by Gollier (1995) and Carroll and Kimball (1996) established that a similar result holds whenever the third derivative of the utility function is positive, a feature of preferences labelled prudence. Both CARA and power utility exhibit prudence. The presence and size of precautionary savings is a matter of great relevance for public policy, in so far as public insurance schemes covering such risks as unemployment, health and

longevity should reduce the need for consumers to accumulate assets.

The great merit of the model with prudence is that it highlights the need to save for rainy days even if sunny days are equally likely. An increased variance in the shocks to income reduces consumption even if expected income does not change. In the case of discrete variables, such as unemployment or illness, changes in first and second moments occur simultaneously, but this is not the case for continuous variables. The ability to distinguish between first and second moment effects is of crucial importance in the analysis of public policy, because of its social insurance characteristics.

The solution of the Bewley model with more general utility functions has to be computed numerically or rely on approximations. Several studies in the early 1990s took up the challenge of characterizing such solutions. Deaton (1991) studied a model with power utility and infinite life. Deaton considered the existence of liquidity or no-borrowing constraints, and showed that impatient consumers would hold limited assets to insure against low income draws. Carroll (1992) instead covered the case of finite lives, and showed that, if consumers are sufficiently impatient and their labour income is subject to both permanent and temporary shocks, they set consumption close to income. The model with impatient consumers under labour income uncertainty has been labelled ‘the buffer stock model’, because saving is kept to the lowest level compatible with the need to buffer negative income shocks. Later work by Attanasio et al. (1999) and Gourinchas and Parker (2002) clarifies the role played by age-related changes in demographics and the hump-shape age profile of labour income in generating income tracking for relatively young consumers (micro data show that financial asset accumulation starts around age 40). Hubbard, Skinner and Zeldes (1994; 1995) show instead how precautionary motives interact with the insurance properties of Social Security in the United States.

Many of the papers cited in the preceding paragraph consider relatively simple versions of the life-cycle model. In particular, a single non-durable commodity is assumed and preferences are assumed to be additively separable with leisure and over time. While this greatly simplifies the solution, the construction of a more realistic and complex model has become an important area of research. This development follows from the recognition that, for many purposes, and in particular for policy analysis, a model that delivers consumption as a function of exogenous variable is a very useful tool indeed.

This area of research has to deal with two important issues. First of all, the model can become very quickly, from a numerical point of view, very difficult to solve. The large number of state variables that characterize the solution of reasonably realistic models and the consideration of discrete choices and non-convexities linked to transaction costs can push the numerical capabilities of even very powerful computers. Second and even more importantly, if one wants to obtain solutions for consumption in a dynamic context, one has to characterize completely the stochastic environment in which the consumer lives. This contrast sharply with the Euler equation approach that allowed the researcher to be agnostic about most aspects of the environment and, under certain conditions, avoid solving difficult problems, such as labour supply, housing and other durable choices and so on. The Euler equation would hold regardless of the presence of non-convexities and other type of difficulties connected with these choices. These, instead have to be fully specified if one wants to work with a model that delivers a solution for consumption. These two difficulties constitute limits for the research in this area that, in all likelihood will not be overcome in the near future.

The empirical evidence on the PILC model

Since its introduction in the 1950s, there is no consensus about the empirical relevance of the PILC model. While the model it is one of the main tools in modern macroeconomics and public finance, its empirical performance is mixed. In this section, we discuss two branches of the literature.

The life-cycle model with various sources of uncertainty and generic preferences generates decision rules and behaviour of great complexity. Consumption and saving choices depend in an unknown fashion on every single aspect of the stochastic environment faced by the consumer, for instance on the entire distribution of future wages and earnings opportunities, on pension arrangements, on the asset markets the consumer can access, on mortality risks and so on and so forth. The Euler equation approach allows researchers to deal in a rigorous fashion with extremely rich models and yet derive relatively simple implications to test some aspects of the model and, with the help of additional assumptions, to identify some of the structural parameters that inform individual behaviour. We now understand that Euler equations can be used to determine what type of preferences fits the available data and can therefore provide one of the building blocks (preferences) in the study of the questions above. We also know that the presence of liquidity constraints does not necessarily produce violations of Euler equations because, even when liquidity constraints are present, they might be rarely binding.

The Euler equation is robust to a number of market imperfections, but is silent about how consumption or its growth reacts to specific news about shocks, changes in interest rates, taxation and so on. It is therefore useless for specific policy analysis. In other words, while the parameters of an Euler equation can be estimated in a wide set of circumstances, and one can use the equation to test the specification of the model, none of these results will provide an answer to questions like what is the effect of a change in taxation or interest rates on the level of consumption and saving?

This important shortcoming of the Euler equation approach explains why such an approach, which has informed and dominated the large empirical literature on the validity of the life-cycle permanent income model is virtually absent in the public economics literature on, say, the effect of pension reforms on saving or on the effect of changes in the taxation of interest on saving. And yet the conceptual framework that is behind the study of these issues is the same as that used to study consumption behaviour.

Policy analysis requires instead the availability of a consumption function, that is, a relation that explains consumption as a function of those variables that the consumer can take as exogenous at any given moment. Only in the simplest versions of the life-cycle model is it possible to derive an analytical expression for the consumption function. In general, given a set of assumptions on preference parameters and market and non-market opportunities, one has to rely on numerical solutions and/or approximations

A less ambitious but potentially profitable approach that does not require numerical methods or incredibly rich data-sets is the estimation of reduced form equations, whose specification is informed by the life-cycle model. These are particularly useful in situations in which one analyses large (and possibly exogenous) changes to some of the likely determinants of consumption or saving. Such studies can address substantive issues and even test some aspects of the life-cycle model. Examples of studies of this kind include the reaction of consumption (and saving) to changes in pension entitlements (Attanasio and Brugiavini, 2003; Attanasio and Rohwedder, 2003; Miniaci and Weber, 1999), to swings in the value of important wealth components (such as housing, Attanasio and Weber, 1994) and to changes in specific taxes (Parker, 1999; Souleles, 1999; Shapiro and Slemrod, 2003).

Below we review the empirical evidence on the PILC model, organizing it in two subsections. First we start with the empirical evidence derived from Euler equations. We then move on to evidence that considers the *levels* of consumptions, rather than its changes.

Evidence from Euler equations

Two important empirical issues can be addressed with the study of Euler equations:

- What is the empirical relevance of the model? Is there a sensible specification of preferences that fits the observed data?
- What is the magnitude of the relevant preference parameters?

Tests of the model

As mentioned above, a prediction of the model is that changes in consumption cannot be predicted by expected changes in income or any other variable known to the consumer at time $t - 1$. This is the essence of the Hall (1978) test and of many others. Evidence that consumption can be predicted by lagged variables has been interpreted as indicative of liquidity constraints, myopic behaviour, misspecification of preferences and so on. The relationship between consumption and income has received considerable attention. The first to observe that the life-cycle model predicts no relation between the life-cycle profile of income and consumption was Thurow (1969). Thurow argued that the fact that consumption tracked income over the life-cycle was a rejection of the main implications of the PILC model. To this argument, essentially identical to many others proposed subsequently, Heckman (1974) replied that non-separability between consumption and leisure could explain such a relationship. Despite this early exchange, after Hall (1978) a large fraction of the literature based on consumption Euler equations focused on the relationship between predictable changes in income and expected consumption growth. Hall and Mishkin (1982), as well as Campbell and Mankiw (1990; 1991) all report violations of this prediction, and label this finding ‘excess sensitivity’. Excess sensitivity can be explained by the presence of liquidity constrained consumers, or of rule-of-thumb consumers, that is, consumers who let their expenditure track their income as a way to avoiding the complexities of choosing the optimal consumption path. However, consistently with Heckman's (1974) argument, excess sensitivity can be reconciled with the intertemporal optimization model if more general, and sensible, utility functions are used. In particular, if one assumes that leisure affects utility in a non-additive way, consumption changes respond to predictable labour income changes, whether or not leisure is a freely chosen variable. Finally, and importantly, the aggregation issue proves to be important. Attanasio and Weber (1993) show that results obtained with improperly aggregated micro data are consistent with results obtained with aggregate data and indicate rejections of the model that instead disappear with properly aggregated data and rich enough preference structures.

To summarize the discussion so far, it seems that while simple tests of the life-cycle model seem to reject the implications from the model and in particular those derived from Euler equations, it is possible to find specification of preferences that do a good job at fitting the available data, especially for households that are headed by prime-aged individuals. Aspects that are crucial for fitting the data are the

use of household level data, allowing for changes in consumption needs induced by changes in family composition and the use of preferences specifications that allow for the marginal utility of consumption to depend on labour supply.

Estimation of preference parameters

Recent research on consumption and saving has singled out three preference parameters for attention: the elasticity of intertemporal substitution, the relative risk aversion parameter and the subjective discount rate. The size of these parameters has important implications in many applications of the model, ranging from macroeconomics to public finance to financial economics.

Perhaps surprisingly, not much evidence has been accumulated on the discount factor from the estimation of Euler equations. This can be explained by the fact that in log-linearized versions of the Euler equation, the parameter is not identified, while non-linear versions of the model are ridden by a number of econometric problems, particularly in relatively small samples of the type used in Euler equation estimation (see Attanasio and Low, 2004).

As for the distinction between the elasticity of intertemporal substitution (EIS) and the coefficient of risk aversion, it is absent in the most popular specifications used in the literature: a model where consumers maximize expected utility and preferences are iso-elastic and additively separable over time. In such a situation, the EIS is the reciprocal of the coefficient of relative risk aversion. Not many empirical papers have worked with preferences that allow for these two parameters to be disjoint.

An influential paper by Hall (1988) claimed that this parameter is close to zero. This finding has been challenged on various grounds. Attanasio and Weber (1993; 1995) point out that aggregation bias could be responsible for such a low estimate: they estimated a much higher elasticity (around 0.8) using UK and US cohort data (that is, data from repeated cross-sections, consistently aggregated over individuals born in the same years).

In the macro literature little attention has been paid to the possibility that the EIS may differ across consumers, particularly as a function of their consumption. A simple way to capture the notion that poor consumers may be less able to smooth consumption across periods and states of nature is to assume that the utility function does not depend on total (non-durable) consumption, but rather on the difference between consumption and needs. Thus we could retain the analytical attraction of power utility, but have $(C - C^*)$ as its argument, where C^* is an absolute minimum that the consumer must reach in each and every period. This functional form is known as Stone–Geary utility in demand analysis, and is the simplest way to introduce non-homotheticity in a demand system. One could interpret ‘external habits’ (Abel, 1990; Campbell and Cochrane, 1999) as a special way to parameterize C^* (by making it a fraction of past consumption of other consumers). Attanasio and Browning (1995), Blundell, Browning and Meghir (1994) and Atkenson and Ogaki (1996) are among the few examples of papers that explicitly allow for wealth-dependent EIS.

Demographics might also affect preferences, and might explain consumption changes and the shape of the consumption age profile, as argued by Attanasio et al. (1999) as well as Browning and Ejrnaes (2002).

Evidence from the levels of consumption

As stressed above, the Euler equation imposes some restrictions on the dynamics of consumption but, on its own, does not determine the level of consumption. If one neglects numerical complications, a solution for consumption can be obtained by considering jointly the Euler equation and the sequence of budget constraints faced by the consumer as well as his or her initial wealth and a terminal condition. As noted by Sargent (1978), Flavin (1981) and later by Campbell (1987), the Euler equation and the intertemporal budget constraint imply a number of cross-equation restrictions for the joint time series processes of consumption and income. When one is able to obtain a closed form solution for consumption, as is the case with quadratic utility, these restrictions can be easily expressed in terms of a linear time series model, and tested.

Some of these restrictions are also implied by the Euler equation, while others are not. In particular, the restrictions on the contemporaneous correlation between income and consumption are not implied: as we stressed above, the Euler equation is silent about how news about income is translated into news about consumption.

Campbell and Deaton (1989) and West (1988) proposed a test that links the innovation to permanent income to consumption and presented evidence that aggregate consumption seems to be ‘excessively smooth’ in that it does not react enough to news about income. Campbell and Deaton make a connection between excess sensitivity and excess smoothness. Within the certainty equivalent model, they jointly model the consumption and income processes as a vector autoregression, assuming that income has a unit root plus some persistence. In this context, consumption changes reflect the permanent income innovation more than one-to-one: not only is the income shock permanent, but it also predicts future, smaller shocks of the same sign. This implies that over the business cycle consumption should be more volatile than income. But in actual aggregate data consumption is smoother than income: this is labelled ‘excess smoothness’, and is shown to be exactly equivalent to excess sensitivity.

Clearly the implications of a given set of intertemporal preferences for policy relevant questions depend crucially on the markets individuals have access to, on their imperfections and on the nature of the equilibrium they give rise to. The implications of complete markets would be very different from those one would derive if liquidity constraints or other markets imperfections were prevalent.

Insurance and credit markets

So far we have taken the assets the consumer can use to move resources over time as given and, in the simplest versions of the model, we have made very strong assumptions on this crucial aspect. For instance, we have assumed that consumers can borrow and lend at a fixed interest rate. The reality is, obviously, much more complex and, from a theoretical point of view, very many different environments have been studied. In particular, the possibilities open to a consumer depend on the market arrangements available. Below we discuss several of these market arrangements and briefly mention their implications for the determination of consumption.

Perfect insurance markets

If markets are complete and consumers can trade a full set of contingent claims without cost, individual risk will be completely diversified. In such a situation, a number of results deliver very useful

predictions. In particular, it can be shown that a competitive equilibrium is symmetric and it is therefore possible to characterize the properties of competitive equilibria by considering the problem of a fictitious social planner, which, given a set of Pareto weights, maximizes social welfare. A strong implication of perfect markets is that the marginal utility of different consumers will move proportionally over time. The implication is very intuitive: the social planner faces a unique resource constraint, and marginal utility of all individuals, multiplied by the appropriate (and arbitrary) Pareto weight, will be equal to the multiplier associated to this unique constraint. As a consequence, marginal utility will move proportionally. If utility is isoelastic, consumption moves proportionally. These implications, stressed by Townsend (1994), have been tested in several papers (Cochrane, 1991; Attanasio and Davis, 1996; Hayashi, Altonji and Kotlikoff, 1996).

Many assets

When there are many assets, one can derive an Euler equation such as (7) for each of the assets for which the consumer is not at a corner. The Euler equations for consumption with different assets naturally ties up with asset pricing equations. This approach to asset pricing was developed by Breeden (1979) and Lucas (1978), and extended to the case of non-additive separability of consumption and leisure in an incomplete markets setting by Bodie, Merton and Samuelson (1992). The model we sketched above is quite restrictive: the relative risk aversion parameter is inversely related to the elasticity of intertemporal substitution: Epstein and Zin (1989) show how this restriction can be relaxed in a more general model with power utility where the timing of uncertainty resolution matters (see also Epstein and Zin, 1991; Attanasio and Weber, 1989).
Interestingly, an Euler equation for an asset holds even if there are important imperfections in some other assets. As long as the consumer is exploiting a given margin to move resources over time, an equation such as (7) will apply. If the interest rate for a given asset changes with the level of the asset, then the Euler equation (7) will have to be augmented with a term reflecting this effect (Pissarides, 1978).

Liquidity constraints

The Euler equation will be violated when the consumer is able, for some reason, to borrow against future income. In such a situation, eq. (7) will hold as an inequality and the marginal utility of current consumption will be higher than the present discounted value of future consumption. Consumers who are liquidity constrained will be very sensitive to changes in current income. This case has received a considerable amount of attention in the literature. Many of the tests of violation of the Euler equation, such as Zeldes (1989), have focused on the so-called ‘excess sensitivity’ of consumption changes to predictable changes in income. It should be mentioned that, in a model with finite lives and a non-zero probability that income would be zero in each time period, standard regularity conditions on the utility function imply that a consumer will never want to borrow. If income is bounded away from zero, then the maximum the consumer will want to borrow is the present discounted value of the minimum value of income repeated in the future. This type of constraint has been sometimes referred to as a ‘natural’

liquidity constraint. Notice that such a constraint does not imply a violation of the Euler equation. If the restriction to borrowing is tighter, the Euler equation will instead be occasionally violated. And, even in periods in which it is not violated, the level of consumption will be affected by the possibility that the constraint will be binding in the future. As Hayashi (1987) explains, the presence of an operative, albeit not binding, liquidity constraint is equivalent to a shortening of the planning horizon or an increase in the discount rate. Evidence can be obtained by noting that consumers who are liquidity constrained will not be sensitive to changes in the level of the interest rate. As they will be at a kink of an intertemporal budget constraint, the demand for loans will be inelastic to changes in the slope of such an intertemporal budget constraint: the interest rate.

Endogenous liquidity constraint

In recent years, several studies have tried to model the shortcomings of credit and insurance markets by allowing for specific imperfections and frictions explicitly. The two main causes of imperfections that have been considered are: (*a*) private and asymmetric information and (*b*) the inability to perfectly enforce contracts. Models of this type can be seen as ways to endogenize specific market structures (such as one where consumers have access to a single asset in which they cannot borrow). In an influential paper, for instance, Cole and Kocherlakota (2001) show that an economy where individuals have a single bond in which they can borrow can be derived as a constrained equilibrium outcome where individuals have private information both on their income and on their savings.

Further extensions and alternative models

While the evidence on the relevance of the life-cycle model is still inconclusive, a number of empirical puzzles have directed attention to more complex preference structures. In particular, the equity premium puzzle and the evolution of aggregate saving rates in high-growth economies (South East Asia) has led macroeconomists to incorporate habits into the model. However, there is still little formal evidence on the empirical relevance of habits in micro data. The widely documented retirement consumption puzzle (that is, a sudden drop of consumption at retirement) as well as a number of more or less anecdotal pieces of evidence on the inadequacy of saving for retirement and other forms of ‘irrational’ behaviour, have been interpreted as potentially supportive of time-inconsistent preferences. The most elegant way to introduce time-inconsistent preferences is provided by the hyperbolic discounting assumption (Laibson, 1997).

Habits

Habits cause consumers to adjust slowly to shocks to permanent income, thus potentially explaining the excess smoothness of aggregate consumption, but also increase the utility loss associated with consumption drops, and may therefore help explain the equity premium puzzle.
Habits can take various forms: today's marginal utility may depend on the consumer's own past consumption level (internal habits) or the past consumption level of other consumers (external habits). This latter model seems to work better on aggregate data (Campbell and Cochrane, 1999), even though a

recent survey by Chen and Ludvigson (2004) challenges this conclusion.

Empirical macro-evidence on the presence of habits is mixed, and this may be due to the very nature of aggregate consumption data, as stressed in Dynan (2000). The serial correlation of aggregate consumption growth is affected by time aggregation (Heaton, 1993), by aggregation over consumers, and by data construction methods (particularly for the services from durable goods). For this reason micro data seem preferable.

The simplest way to introduce habits (or durability) of consumption is to write the utility function as follows:

$$\Sigma_t u(x_t - \gamma' x_{t-1}; z_t) \tag{11}$$

where x is a vector of goods or services and z is any other variable that affects marginal utility (demographics, leisure, other goods that are not explicitly modelled). The γ parameters are positive for goods that provide services across periods (durability), negative for goods that are addictive (habit formation) or zero for goods that are fully non-durable, non-habit forming (Hayashi, 1985).

The Euler equations corresponding to (11) involves x at four different periods of time, and their estimation typically requires panel data. High-quality consumption panel data are rare, and this has limited the scope for empirical analysis. Meghir and Weber (1996) have used Consumer Expenditure Survey (CEX) quarterly data on food, transport and services (and a more flexible specification of intertemporal non-separabilities than is implied by eq. 11), and found no evidence of either durability or habits once leisure, stock of durables and cars as well as other conditioning variables are taken into consideration.

Similarly negative evidence on habits has been reported by Dynan (2000), using Panel Study of Income Dynamics (PSID) annual food at home data. Carrasco, Labeaga and López-Salido (2005) use Spanish panel data and find some evidence for habits.

The few studies that have used micro data on non-durable consumption items to investigate the issue find little or no evidence of habits, at least once preferences capture the presence of non-separabilities between goods and leisure.

Durable goods

The presence of durable goods has received less attention in the micro-based literature than in the macro-literature, which has stressed the importance of their high volatility to explain business cycle fluctuations (Mankiw, 1982; Chah, Ramey and Starr, 1995).

The simplest way to introduce durable goods into the analysis is to let the stock of durables affect utility (on the assumption that services are proportional to the stock), and to posit a relation between current stock, S_t, previous stock, S_{t-1}, and current purchases q_t (or maintenance and repairs) in physical terms

like:

$$S_t = (1 - \rho)S_{t-1} + q_t \tag{12}$$

where ρ is a constant depreciation rate. This leads to the standard first-order condition for the durable good, according to which the relevant price is the user cost.

Typically, durable goods are costly to adjust, because of transaction costs (resale markets are dominated by information problems, known as the ‘lemon’ problem, and search costs are non-negligible). Sometimes these costs are modelled as a convex, differentiable function (Bernanke, 1985), but the recent literature has stressed the need to take into account their non-differentiable nature (Grossman and Laroque, 1990; Eberly, 1994; Attanasio, 2000; Bertola, Guiso and Pistaferri, 2005). This generates infrequent adjustment: consumers do not adjust continuously in response to depreciation, or income and price shocks, but wait until the actual stock hits either a lower limit, s, or an upper limit, S, and then adjust it to a target level. An interesting feature of this literature is that aggregate behaviour reflects changes in both the number of consumers that adjust and in the target level.

Durable goods might also play an insurance role, because they can be used to sustain consumption when times are bad. Postponing the purchase of food, or clothing, is certainly harder than failing to replace an old refrigerator or car, and housing maintenance can be put off for very long periods before structural damage occurs (Browning and Crossley, 2000). Durable goods also play a more specific insurance role, against changes in the price of the corresponding services. This is particularly relevant in the case of housing, where owning your home may be the best way to hedge the risk of future increases in the market price of housing services (Sinai and Souleles, 2005). Durable goods can also play a liquidity role, if they can be used as collateral to obtain a loan that pays for current consumption (Alessie, Devereux and Weber, 1997). A typical example could be the ability to remortgage a house, or to borrow 100 per cent of the value of a newly purchased car.

Even if one is not interested in modelling durable goods, the existence of a stock of durables should not be neglected when estimating preference parameters if utility is not additive in non-durable goods and durable services. Significant effects of durable goods (cars) on the Euler equation for non-durables have been found in UK data (Alessie, Devereux and Weber, 1997), and US data (Padula, 1999).

Quasi-hyperbolic discounting

The widely documented consumption puzzle (that is the sudden drop of consumption at retirement, see Hamermesh, 1984; Banks, Blundell and Tanner, 1998; and Bernheim, Skinner and Weinberg, 2001), as well as a number of more or less anecdotal pieces of evidence on the inadequacy of saving for retirement and other forms of ‘irrational’ behaviour, have been interpreted as potentially supportive of time-inconsistent preferences. The most elegant way to introduce time-inconsistent preferences is provided by the quasi-hyperbolic discounting assumption (Laibson, 1997). Consumers maximize the expected value

of the following life-time utility index:

$$u(c_t) + \beta \sum_{\tau=1}^{T-t} \delta^{t} u(c_{t+\tau}) \tag{13}$$

This implies that a different, lower discount factor is used to choose between this period and the next (the product of β and δ) and between any two other periods (δ). This generates time-inconsistent plans, with too little saving for retirement. For this reason, consumers may choose to enter long-term commitment plans, such as 401(k)s in the United States.

The quasi-hyperbolic discounting model lends itself to estimation and testing, but requires solving for the consumption function numerically. Even though an Euler equation for this model has been derived, its empirical use is limited, because it involves the marginal propensity to consume out of wealth (Harris and Laibson, 2001). It also suffers from some potential difficulties related to the definition of the time period, which crucially affects the properties of the solution, the length of which is arbitrarily set by the researcher.

A more tractable specification of preferences that may be used to model quasi-rational impatience has been put forward by Gul and Pesendorfer (2001; 2004), who stress the importance of self-control problems leading to the postponement of saving.

Where do we stand?

Since the 1970s we have learned much about the empirical implications of the life-cycle model and about the details of the model that need to be modified to fit the available evidence. Much work, however, remains to be done. In particular, there is scope to develop more complex numerical models that incorporate several realistic features. The areas of labour supply and housing are, in our opinion, particularly important. We also need to develop our understanding of the empirical implications of alternative models, such as hyperbolic discounting and check the extent to which they are empirically distinguishable from more standard models with complex preferences. Finally, it is important to stress the need for more and better data. One of the lessons learned from the development of new surveys that have been used to measure household wealth is that with enough ingenuity and creativity one can measure several of the variables that are relevant for our understanding of consumption and saving behaviour.

Our analysis of consumption and saving requires that more comprehensive measures of consumption are included in existing surveys, and that we learn to make systematic use of records on expectations, perceived uncertainty and so on.

See Also

- consumer expenditure
- consumption-based asset pricing models (empirical performance)
- consumption-based asset pricing models (theory)
- elasticity of intertemporal substitution
- Engel curve
- Friedman, Milton
- Modigliani, Franco
- precautionary saving and precautionary wealth
- revealed preference theory

Bibliography

Abel, A.B. 1990. Asset prices under habit formation and catching up with the Joneses. *American Economic Review* 80, 38–42.

Alessie, R., Devereux, M.P. and Weber, G. 1997. Intertemporal consumption, durables and liquidity constraints: a cohort analysis. *European Economic Review* 41, 37–59.

Atkenson, A. and Ogaki, M. 1996. Wealth varying intertemporal elasticities of substitution: evidence from panel and aggregate data. *Journal of Monetary Economics* 38, 507–34.

Attanasio, O.P. 2000. Consumer durables and inertial behaviour: estimation and aggregation of (S, s) rules for automobile purchases. *Review of Economic Studies* 67, 667–96.

Attanasio, O.P., Banks, J., Meghir, C. and Weber, G. 1999. Humps and bumps in life-time consumption. *Journal of Business and Economic Statistics* 17, 22–35.

Attanasio, O.P. and Browning, M. 1995. Consumption over the life cycle and over the business cycle. *American Economic Review* 85, 1118–37.

Attanasio, O.P. and Brugiavini, A. 2003. Social security and households' saving. *Quarterly Journal of Economics* 118, 1075–119.

Attanasio, O.P. and Davis, S.J. 1996. Relative wage movements and the distribution of consumption. *Journal of Political Economy* 104, 1227–62.

Attanasio, O.P. and Low, H. 2004. Estimating Euler equations. *Review of Economic Dynamics* 7, 405–35.

Attanasio, O.P. and Rohwedder, S. 2003. Pension wealth and household saving: evidence from pension

reforms in the United Kingdom. *American Economic Review* 93, 1499–521.

Attanasio, O.P. and Weber, G. 1989. Intertemporal substitution, risk aversion and the Euler equation for consumption. *Economic Journal* 99, 59–73.

Attanasio, O.P. and Weber, G. 1993. Consumption growth, the interest rate and aggregation. *Review of Economic Studies* 60, 631–49.

Attanasio, O.P. and Weber, G. 1994. The UK consumption boom of the late 1980s: aggregate implications of microeconomic evidence. *Economic Journal* 104, 1269–1302.

Attanasio, O.P. and Weber, G. 1995. Is consumption growth consistent with intertemporal optimization? Evidence from the consumer expenditure survey. *Journal of Political Economy* 103, 1121–57.

Banks, J., Blundell, R. and Tanner, S. 1998. Is there a retirement-savings puzzle? *American Economic Review* 88, 769–88.

Bernanke, B. 1985. Adjustment costs, durables and aggregate consumption. *Journal of Monetary Economics* 15, 41–68.

Bernheim, B.D., Skinner, J. and Weinberg, S. 2001. What accounts for the variation in retirement wealth among U.S. households? *American Economic Review* 91, 832–57.

Bertola, G., Guiso, L. and Pistaferri, L. 2005. Uncertainty and consumer durables adjustment. *Review of Economic Studies* 72, 973–1007.

Bewley, T.F. 1977. The permanent income hypothesis: a theoretical formulation. *Journal of Economic Theory* 16, 252–59.

Blinder, A. and Deaton, A. 1985. The time series consumption function revisited. *Brookings Papers on Economic Activity* 1985(2), 465–521.

Blundell, R., Browning, M. and Meghir, C. 1994. Consumer demand and the life-cycle allocation of household expenditures. *Review of Economic Studies* 61, 57–80.

Bodie, Z., Merton, R.C. and Samuelson, W.F. 1992. Labor supply flexibility and portfolio choice in a life-cycle model. *Journal of Economic Dynamics and Control* 16, 427–49.

Boskin, M.J. 1978. Taxation, saving and the rate of interest. *Journal of Political Economy* S3–S28.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment

opportunities. *Journal of Financial Economics* 7, 265–96.

Browning, M. and Crossley, T.F. 2000. Luxuries are easier to postpone: a proof. *Journal of Political Economy* 108, 1022–6.

Browning, M. and Ejrnaes, M. 2002. Consumption and children. Working Paper No. 2002–6, Centre for Applied Microeconometrics, University of Copenhagen.

Caballero, R.J. 1990. Consumption puzzles and precautionary savings. *Journal of Monetary Economics* 25, 113–36.

Caballero, R.J. 1991. Earnings uncertainty and aggregate wealth accumulation. *American Economic Review* 81, 859–71.

Campbell, J.Y. 1987. Does saving anticipate declining labor income? An alternative test of the permanent income hypothesis. *Econometrica* 55, 1249–73.

Campbell, J.Y. and Cochrane, J. 1999. Force of habit: a consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107, 205–51.

Campbell, J.Y. and Deaton, A. 1989. Why is consumption so smooth? *Review of Economic Studies* 56, 357–73.

Campbell, J.Y. and Mankiw, N.G. 1990. Permanent income, current income, and consumption. *Journal of Business and Economic Statistics* 8, 265–79.

Campbell, J.Y. and Mankiw, N.G. 1991. The response of consumption to income: a cross-country investigation. *European Economic Review* 35, 723–56.

Carrasco, R., Labeaga, J.M. and López-Salido, J.D. 2005. Consumption and habits: evidence from panel data. *Economic Journal* 115, 144–65.

Carroll, C.D. 1992. The buffer-stock theory of saving: some macroeconomic evidence. *Brookings Papers on Economic Activity* 1992(2), 61–156.

Carroll, C.D. 1997. Buffer-stock saving and the life cycle/permanent income hypothesis. *Quarterly Journal of Economics* 112, 1–55.

Carroll, C.D. and Kimball, M.S. 1996. On the concavity of the consumption function. *Econometrica* 64, 981–92.

Chah, E.Y., Ramey, V.A. and Starr, R.M. 1995. Liquidity constraints and intertemporal consumer optimization: theory and evidence from durable goods. *Journal of Money, Credit and Banking* 27, 272–87.

Chen, X. and Ludvigson, S.C. 2004. *A Land of Addicts? An Empirical Investigation of Habits-Based Asset Pricing Models*. New York: New York University Press.

Choi, J.J., Laibson, D., Madrian, B.C. and Metrick, A. 2006. Saving for retirement on the path of least resistence. In *Behavioral Public Finance: Toward a New Agenda*, ed. E. McCaffrey and J. Slemrod. New York: Russell Sage Foundation.

Cochrane, J.H. 1991. A simple test of consumption insurance. *Journal of Political Economy* 99, 957–76.

Cole, H.L. and Kocherlakota, N.R. 2001. Efficient allocations with hidden income and hidden storage. *Review of Economic Studies* 68, 523–42.

Deaton, A. 1991. Saving and liquidity constraints. *Econometrica* 59, 1221–48.

Dynan, K. 2000. Habit formation in consumer preferences: evidence from panel data. *American Economic Review* 90, 391–406.

Eberly, J.C. 1994. Adjustment of consumers' durables stocks: evidence from automobile purchases. *Journal of Political Economy* 102, 403–36.

Epstein, L.G. and Zin, S.E. 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. *Econometrica* 57, 937–69.

Epstein, L.G. and Zin, S.E. 1991. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: an empirical analysis. *Journal of Political Economy* 99, 263–86.

Flavin, M.A. 1981. The adjustment of consumption to changing expectations about future income. *Journal of Political Economy* 89, 974–1009.

Friedman, M. 1957. *A Theory of the Consumption*. Function. Princeton, NJ: Princeton University Press.

Gollier, C. 1995. The comparative statics of changes in risk revisited. *Journal of Economic Theory* 66, 522–36.

Gourinchas, P.-O. and Parker, J.A. 2002. Consumption over the life cycle. *Econometrica* 70, 47–89.

Grossman, S.J. and Laroque, G. 1990. Asset pricing and optimal portfolio choice in the presence of

illiquid durable consumption goods. *Econometrica* 58, 25–51.

Gul, F. and Pesendorfer, W. 2001. Temptation and self-control. *Econometrica* 9, 1403–35.

Gul, F. and Pesendorfer, W. 2004. Self-control and the theory of consumption. *Econometrica* 72, 119–58.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86, 971–87.

Hall, R.E. 1988. Intertemporal substitution in consumption. *Journal of Political Economy* 96, 339–57.

Hall, R.E. and Mishkin, F.S. 1982. The sensitivity of consumption to transitory income: estimates from panel data on households. *Econometrica* 50, 461–81.

Hamermesh, D.S. 1984. Consumption during retirement: the missing link in the life cycle. *Review of Economics and Statistics* 66, 1–7.

Hansen, L.P. and Singleton, K.J. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–86.

Hansen, L.P. and Singleton, K.J. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249–65.

Harris, C. and Laibson, D. 2001. Dynamic choices of hyperbolic consumers. *Econometrica* 69, 935–57.

Hayashi, F. 1985. The permanent income hypothesis and consumption durability: analysis based on Japanese panel data. *Quarterly Journal of Economics* 100, 1083–113.

Hayashi, F. 1987. Tests for liquidity constraints: a critical survey. In *Advances in Econometrics II: Fifth World Congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

Hayashi, F., Altonji, J. and Kotlikoff, L. 1996. Risk-sharing between and within Families. *Econometrica* 64, 261–94.

Heaton, J. 1993. The interaction between time-nonseparable preferences and time aggregation. *Econometrica* 61, 353–85.

Heckman, J.J. 1974. Life cycle consumption and labor supply: an explanation of the relationship between income and consumption over the life cycle. *American Economic Review* 64, 188–94.

Hubbard, R.G., Skinner, J. and Zeldes, S.P. 1994. The importance of precautionary motives in explaining individual and aggregate saving. *Carnegie-Rochester Conference Series on Public Policy* 40, 59–125.

Hubbard, R.G., Skinner, J. and Zeldes, S.P. 1995. Precautionary saving and social insurance. *Journal of Political Economy* 103, 360–99.

King, R.G. and Plosser, C.I. 1984. Money, credit, and prices in a real business cycle. *American Economic Review* 74, 363–80.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 62, 443–77.

Lucas, R.E., Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 89, 1345–70.

MaCurdy, T.E. 1999. An essay on the life cycle: characterizing intertemporal behavior with uncertainty, taxes, human capital, durables, imperfect capital markets, and nonseparable preferences. *Research in Economics* 53, 5–46.

Mankiw, G.N. 1982. Hall's consumption hypothesis and durable goods. *Journal of Monetary Economics* 10, 417–25.

Meghir, C. and Weber, G. 1996. Intertemporal nonseparability or borrowing restrictions? A disaggregate analysis using a U.S. consumption panel. *Econometrica* 64, 1151–81.

Miniaci, R. and Weber, G. 1999. The Italian recession of 1993: aggregate implications of microeconomic evidence. *Review of Economics and Statistics* 81, 237–49.

Modigliani, F. and Brumberg, R. 1954. Utility analysis and the consumption function: an interpretation of cross-section data. In *Post Keynesian Economics*, ed. K. Kurihara. New Brunswick, NJ: Rutgers University Press.

Padula, M. 1999. Euler equations and durable goods. Working Paper No. 30, CSEF.

Parker, J.A. 1999. The reaction of household consumption to predictable changes in social security taxes. *American Economic Review* 89, 959–73.

Pissarides, C.A. 1978. Liquidity considerations in the theory of consumption. *Quarterly Journal of Economics* 92, 279–96.

Poterba, J. 1988. Are consumers forward looking? Evidence from fiscal experiments. *American Economic Review* 78, 413–18.

Sargent, T.J. 1978. Rational expectations, econometric exogeneity, and consumption. *Journal of Political Economy* 86, 673–700.

Shapiro, M.D. and Slemrod, J. 2003. Consumer response to tax rebates. *American Economic Review* 93, 381–96.

Sinai, T. and Souleles, N.S. 2005. Owner-occupied housing as a hedge against rent risk. *Quarterly Journal of Economics* 120, 763–89.

Souleles, N.S. 1999. The response of household consumption to income tax refunds. *American Economic Review* 89, 947–58.

Summers, L. 1981. Capital taxation and capital accumulation in a life-cycle growth model. *American Economic Review* 71, 533–44.

Thurow, L.C. 1969. The optimum lifetime distribution of consumption expenditures. *American Economic Review* 59, 324–30.

Townsend, R.M. 1994. Risk and insurance in village India. *Econometrica* 62, 539–91.

West, K.D. 1988. The insensitivity of consumption to news about income. *Journal of Monetary Economics* 21, 17–33.

Zeldes, S.P. 1989. Consumption and liquidity constraints: an empirical investigation. *Journal of Political Economy* 97, 305–46.

How to cite this article

Attanasio, Orazio P. and Guglielmo Weber. "consumer expenditure (new developments and the state of research)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000572> doi:10.1057/9780230226203.0297

The New Palgrave Dictionary of Economics Online

consumer expenditure

Angus Deaton

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Consumers' expenditure is a central concern of economics, both in microeconomic terms (the relationship between prices, expenditure and welfare) and in macroeconomic terms (the relationship between expenditure and income). This article examines the interplay between theory and evidence in the study of consumers' expenditure and its composition. Although models have been developed from the theory of consumption that illuminate much of the available data, many standard presumptions of economics lack substantial bodies of evidence such as central theories in the natural sciences enjoy.

Keywords

almost ideal demand system (AIDS); ARIMA process; Bernoulli utility functions; compensated demand; consumer expenditure; consumers' expenditure; consumption function; convexity; demand curve; demand functions; Denver Income Maintenance Experiment; econometrics; elasticity; Engel curves; Engel's Law; Euler equation; Fisher, I.; fixed needs model; flexible functional forms; Friedman, M.; generalized axiom of revealed preference (GARP); generalized Leontief system; Hicks, J.; household budgets; inflation; inter-temporal utility functions; Keynes, J.; King, G.; Kuznets, S.; law of demand; life cycle hypothesis; life-cycle income model; linear expenditure system; Lucas, R.; marginal rate of substitution; Michigan Panel Study of Income Dynamics (PSID); Modigliani, F.; non-linear optimization; optimal taxation; permanent income model; prices; Ramsey, F.; random walk model; revealed preference theory; risk aversion; Roy's identity; Samuelson, P.; saving ratio; Shephard's lemma; Slutsky matrix; Stigler, G.; Stone, J.; Stone-Geary utility function; substitution effect; tax reform analysis; unit root model; utility maximization

Article

The study of consumers' expenditure, both in total and in composition, has always been of major

concern to economists. Neoclassical economics sees the delivery of individual consumption as the main object of the economic system, so that the efficiency with which the economy achieves this goal is the criterion by which alternative systems, institutions and policies are to be judged. Within a capitalist economy, such considerations lead to an examination of the relationship between *prices* and consumption behaviour, and theoretical development and empirical analysis have been a major continuous activity since the middle of the last century. Even older is the tradition of using individual household budgets to dramatize poverty, and the relationship between household incomes and household expenditure patterns has occupied social reformers, statisticians and econometricians since at least the 18th century. In more modern times, it has been recognized that the study of public finance and of taxation depends on a knowledge of how price changes affect the welfare and behaviour of individuals, and the recent development of optimal tax theory and of tax reform analysis has placed additional demands on our understanding of the links between prices, expenditures and welfare.

In the last fifty years, aggregate consumption has become as much of an object of attention as has its composition, and in spite of a common theoretical structure, there has been a considerable division of labour between macro economists, interested in aggregate consumption and saving, and micro economists whose main concern has been with composition, and with the study of the effects of relative prices on demand. The interest of macroeconomics reflects both long-term and short-term interests. What is not consumed is saved, saving is thrift and the basis for capital formation, so that the determinants of saving are the determinants of future growth and prosperity. More immediately, aggregate consumption accounts for a large share of national income, typically more than three-quarters, so that fluctuations in behaviour or ‘consumption shocks’ have important consequences for output, employment, and the business cycle. Since Keynes's *General Theory*, the consumption function, the relationship between consumption and income, has played a central role in the study of the macroeconomy. Since the 1930s, there has been a continuous flow of theoretical and empirical developments in consumption function research, and some of the outstanding scientific achievements in economics have been in this field.

In this essay, the major themes will be the interplay between theory and evidence in the study of consumers' expenditure and its composition. If economists have any serious claim to being scientists, it should be clearly visible here. The best minds in the profession have worked on the theory of consumption and on its empirical implementation, and there have always been more data available than could possibly be examined. I hope to show that there have been some stunning successes, where elegant models have yielded far from obvious predictions that have been well vindicated by the evidence. But there is much that remains to be done, and much that needs to be put right. Many of the standard presumptions of economics remain just that, assumptions unsupported by evidence, and while modern price theory is logically consistent and theoretically well developed, it is far from having that solid body of empirical support and proven usefulness that characterizes similar central theories in the natural sciences.

1 A simple theoretical framework

Almost all discussions of consumer behaviour begin with a theory of *individual* behaviour. I follow neoclassical tradition by supposing that such behaviour can be described by the maximization of a utility function subject to suitable constraints. The axioms that justify utility maximization are mild, see any

microeconomic text such as Varian (1978/1984) or Deaton and Muellbauer (1980b), so that utility maximization should be seen as no more than a convenient framework that rules out the grossest kind of behavioural inconsistencies. The assumptions that have real force are those that detail the constraints facing individuals or else put specific structure on utility functions. Perhaps the most general specification of preferences that could be considered is one that is written

$$u_t = E_t\{f(q_1, q_2, \ldots, q_t, \ldots, q_T)\} \tag{1}$$

where u_t is utility at time t, E_t is the expectation operator for expectations formed at time t, q_1 to q_T are vectors of consumption in periods 1 to T, and $f(\cdot)$ is a quasi-concave function that is non-decreasing in each of its arguments. Several things about this formulation are worth brief discussion. The function $f(\cdot)$ yields the utility that would be obtained from the consumption vector under certainty, and it represents the utility from a *life-time* of consumption; the indices 1 to T therefore represent *age* with 1 the date of birth and T that of death. The expectation operator is required because choice is made subject to uncertainty, not about the choices themselves, which are under the consumer's control, but about the consequences of current choices for future opportunities. It is not possible to travel backward through time, so that choices once made cannot be undone, and yet the cost of current consumption in terms of future consumption foregone is uncertain, as is the amount of resources that may become available at future dates. The consumer must therefore travel through life, filling in the slots in (1) from left to right as best as he or she can, and at time (or age) t, everything to the left will be fixed and unchangeable, whether now seen to be optimal or not, while everything ahead of t is subject to the random buffeting of unexpected changes in interest rates, prices, and incomes. The solution to this sort of maximization problem has been elegantly characterized by Epstein (1975); here I shall work with something that is more restrictive but more useful and note in Section 3 below some phenomena that are better handled by the more general model.

Intertemporal utility functions are frequently assumed to be *intertemporally additive*, so that the preference rankings between consumption bundles in any two periods or ages are taken to be independent of consumption levels in any third period. If so, the utility function (1) takes the more mathematically convenient form

$$u_t = E_t \sum_{r=1}^{T} v_r(q_r). \tag{2}$$

Note that by writing utility in the form (2), since the expectation operator is additive over states of the

world preferences are in effect assumed to be *simultaneously* additive over both states and periods, an assumption that can be formally defended, see Gorman (1982) and Browning, Deaton and Irish (1985). It has the consequence that risk aversion and intertemporal substitutability become two aspects of the same phenomenon. Individuals that dislike risk, and will pay to avoid it, will also attempt to smooth their consumption over time and will require large incentives to alter their preferred consumption and saving profiles. Note also that the additive structure of (2) means that, unlike the case of (1), previous decisions are irrelevant for current ones. For decision-making at time *t*, bygones are bygones, and conditional on asset and income positions, future choices are unaffected by what has happened in the past. There can therefore be no attempt to make up for lost opportunities, nor can such phenomena as habit formation be easily modelled.

Because utility in (2) is intertemporally separable, maximization of life-time utility implies that, within each period, the period subutility function $v_t(\cdot)$ must be maximized subject to whatever total it is optimal to spend in that period. The period by period allocation of consumption expenditure to individual commodities need not, therefore, be planned in advance, but can be left to be determined when that period or age is reached, and period *t* allocation will follow according to the rule

$$\text{maximize } v_t(q_t) \text{ subject to } p_t \cdot q_t = x_t, \tag{3}$$

where p_t is the price vector corresponding to q_t and x_t is the total amount to be spent in *t*. Problem (3) is one of standard (static) utility maximization, though note that x_t is not given to the consumer, but is determined by the wider intertemporal choice problem. Nevertheless, not the least advantage of the intertemporally additive formulation is its implication that the composition of expenditure follows the standard utility maximization rule. It allows separate attention to be given to demand analysis on the one hand, i.e. to the problem (3), and to the consumption function on the other hand, this being understood to be the intertemporal allocation of resources, i.e. the determination of x_t.

Write the maximized value of utility from the period *t* problem as $\psi_t(x_t, p_t)$, where $\psi(\cdot)$ is a standard indirect utility function. The original intertemporal utility function then takes the form

$$u_t = E_t \sum_{r=0}^{T-t} \psi_r(x_{t+r}, p_{t+r}) . \tag{4}$$

The constraints under which this function is maximized are most conveniently analysed through the conditions governing the evolution of wealth from period to period. If A_t is the (ex-dividend) value of assets at the start of period *t*, N_{it} is the nominal holdings of asset *i* with price P_{it}, d_{it} is the dividend on *i*

paid immediately before the beginning of t, and y_t is income in period t, then

$$A_{t+1} = \sum_i N_{it}(P_{it+1} + d_{it+1})$$
(5)

$$\sum_i N_{it}P_{it} = A_t + y_t - x_t .$$
(6)

Conditions (5) and (6) determine how wealth evolves from period to period, and the picture is completed by requiring that the consumer's terminal assets be positive, i.e.

$$A_{T+1} \geq 0$$
(7)

To solve this problem, the technique of backward recursion is used. This rests on the observation that it is impossible to know what to do in period t without taking into account the problem in period $(t + 1)$, nor that in $(t + 1)$ without thinking about $(t + 2)$, and so on. However, in period T there is no future, so that looking ahead from date t, we can write subutility in period T in terms of that period's price and inherited assets, and we write this as v_T, i.e.

$$v_T = v_T(A_T) = \psi_T(A_T + y_T, p_T).$$
(8)

Given this, the consumer can look ahead from period t to period $(T - 1)$ and foresee that the problem then will be to choose the composition of assets N so as to maximize v_{T-1}, where

$$v_{t-1}(A_{T-1}) = \max_{N} \square\, \psi_{T-1}(A_{T-1} + y_{T-1} - N \cdot P_{T-1},\ p_{T-1}) + E_{T-1}\left\{v_T[N \cdot (P_T + d_T)]\right\} \square .$$
(9)

At the next stage, assets in $(T-2)$ will be allocated so as to trade off the benefits of consumption in $(T-2)$ versus the benefits of A_{T-1} in v_{T-1} in (9) above and again yielding a maximized value v_{T-2}. As we follow this back through time, the consumer finally reaches the current period *t*, where he or she faces an only slightly complicated version of the usual ‘today tomorrow’ trade-off; the asset vector *N* must be chosen to solve the problem,

$$u_t = \max_{N} \square\, \psi_t(A_t + y_t - N \cdot P_t,\ p_t) + E_t\left\{v_{t+1}[N \cdot (P_{t+1} + d_{t+1}]\right\} \square .$$
(10)

From this sequence of problems, several important results readily follow. First, consider the derivatives of each of the functions $v_r(A_r)$ which represent the marginal value of an extra unit of currency for the remaining segment of life time utility from *r* through to *T*. By the envelope theorem (see for example Dixit (1976) for a good exposition), it is legitimate to differentiate through the maximization problem, from which

$$v_r'(A_r) = \partial \psi_r / \partial x_r = \lambda_r, \text{ say,}$$
(11)

so that λ_r is the marginal utility of money in period *r*. Secondly, the maximization of (10) with respect to portfolio choice gives the relationship, for each asset *i*,

$$P_{it}\, \partial \psi_t / \partial x_t = E_t\left\{(P_{it+1} + d_{it+1})\, \partial \psi_{t+1} / \partial x_{t+1}\right\}$$
(12)

which, defining the asset *return* R_{it+1} as $(P_{it+1} + d_{it+1}) / P_{it}$, and using (11) can be rewritten in the simple form

$$\lambda_t = E_t(\lambda_{t+1} R_{it+1}). \tag{13}$$

This equation, in current parlance often referred to as the ‘Euler equation’, can be used to derive many of the implications of the theory of consumption. Note first that it is little more than the standard result that the marginal rate of substitution between today's and tomorrow's consumption should be equal to the relative price. However, the equation is set in a multiperiod framework, not a two-period one, and it explicitly recognizes the uncertainty in both asset returns and in the value of money in subsequent periods. The equation also holds for all *i*, i.e. for all assets, so that the result also has implications for asset pricing as well as for consumption and saving, and for this reason the model is often referred to as the consumption-asset pricing model. I shall return to these implications below.

The theory as presented above is the modern equivalent of the life-cycle theory of consumption that dates back to Irving Fisher (1930) and Frank Ramsey (1928), and that had its modern genesis in the papers by Modigliani and Brumberg (1954) and (1954, published 1979). Modigliani and Brumberg's treatment differs from the above only in not explicitly modelling uncertainty, and by including only a single asset. The modern version appears first in Breeden (1979) and in Hall (1978), see also Grossman and Shiller (1981).

2 Predictions and evidence

One of the most important implications of the theory above, and of equation (13) in particular, is that the evolution of consumption over the life-cycle is independent of the pattern of income over the life-cycle. The asset evolution equations (5) and (6) allow consumers to borrow and lend at will, so that the only ultimate constraint on their consumption is one of life-time solvency. In consequence, consumption patterns are free to follow tastes, the evolution of family structure, or the different needs that come with ageing, provided that in the end total life-time expenditure lies within (total) life-time resources, whether from inherited wealth or from labour income. It is often assumed that tastes are such that consumers prefer to have a relatively smooth consumption stream, and this can be illustrated from a special case of equation (13). Assume that the within-period utility function is homothetic so that $\psi(x, p)$ is $\phi(x/a(p))$ for some linearly homogeneous function $a(\cdot)$, and that $\varphi(\cdot)$ has the isoelastic from with elasticity $(1-\sigma)$. Life-time utility takes the form

$$u_t = \sum_{r=0}^{T-t} (1+\delta)^{-r} [x_{t+r} / a(p_{t+r})]^{1-\sigma} \tag{14}$$

where δ is the rate of pure time preference, and $\sigma \geq 0$ is the coefficient of relative risk aversion and the reciprocal of the intertemporal elasticity of substitution. Equation (14) can be used to evaluate (13), and gives immediately

$$E[\{(1 + r_{t+1}) / (1 + \delta)\}\{c_t / c_{t+1}\}^{\sigma}] = 1 \tag{15}$$

where r_{t+1} is the real after tax rate of interest from t to t+1 on any asset, and c_t is real consumption, x_t/a (p_t). Equation (15) shows that, if expectations are fulfilled, consumption will grow over the life-cycle if the real rate of interest is greater than the rate of pure time preference, and vice versa, while with $r_t = \delta$, consumption is constant with age. These results are of course an artefact of the specific assumptions about utility, and for any real household consumption can be expected to vary predictably with age according to patterns of family formation, growth, and ageing; Modigliani and Ando (1957) have suggested that consumption per ‘equivalent adult’ might be constant over the life-cycle. But whatever the shape of preferences, there need be no relationship between the profiles of consumption and of income; income can be saved until it is needed, or borrowed against if it is not yet available.

Independent of the life-time *pattern* of consumption is its level, which under the life-cycle model is determined by the level of total life-time resources, so that individuals with the same tastes but with higher incomes or higher inherited assets will have higher levels of consumption throughout their lives. If the future were entirely predictable, the consumption plan at any point in time could be decided with reference to the level of total wealth, this being the value of financial assets and the discounted present value of current and future incomes. In this sense, the life-cycle model is a permanent income theory of consumption, where permanent income is the annuity value of lifetime wealth, though the lifetime interpretation is only one of the many that are offered in Friedman's (1957) original statement. Whether life-cycle or not, linking consumption to *future* incomes has important consequences. First, consumption will respond only to ‘surprises’ or ‘shocks’ in income; changes in income that have been foreseen are already discounted in previous behaviour and should not induce any changes in plans. Of course, this does *not* mean that consumption will not change along with changes in income; a change may have been planned in any case, and some proportion of any actual change may well have been unforeseen. However, if a substantial fraction of the regular changes in income over the business cycle are foreseen by consumers, or if unanticipated fluctuations in income are regarded as only temporary with limited consequences for total life time resources, then consumption will not respond very much to cyclical fluctuations in income. Aggregate consumption is indeed much smoother than is aggregate income, and this has been traditionally accepted as an important piece of confirmatory evidence. I shall take up the matter again below when I deal with the recent econometric evidence.

The distinction between measured income and permanent income is also important for the interpretation of cross-sectional evidence. Since measured income can be regarded as an error-ridden proxy for permanent income, the regression of consumption on measured income will be biased downward (rotated clockwise) compared with the true regression of consumption on permanent income. Cross-

sectional regressions, or time-series regressions of simple Keynesian consumption functions will therefore tend to understate the long-run marginal propensity to consume. Well before the work on life-cycle models, Kuznets (1946) showed that the long-run saving ratio in the United States had been roughly constant in spite of repeated cross-sectional analyses showing that the saving ratio rose with income, and the life-cycle theory could also readily account for these findings. It is interesting to note that the constancy of the saving ratio is far from being well established as an empirical fact; the evidence for other countries with long-run data is very mixed, and even the United States saving ratio is clearly influenced in the long-run by technical change, migration patterns, and demographic shifts, see Kuznets (1962) and Deaton (1975). Life-cycle and permanent income theories also predict that households with atypically high income will tend to save a great deal of it, a prediction which explained the apparently anomalous finding that black households tend to save more than white households at the same level of measured income; since blacks typically have lower household income than whites, those with the same measured income can be expected to have a higher transitory component.

The Modigliani and Brumberg life-cycle story was also important because it offered a story of capital accumulation in society as a whole that relied on the way in which people made preparation for their own futures, particularly for their future retirement. In a stationary life-cycle economy, in which there is neither economic nor population growth, aggregate saving is zero, and the old, as they dissave, pass on the ownership of the capital stock to the next generation who are, in turn, saving for their own retirement. With either population or income growth, the aggregate scale of saving by the young would be greater than that of dissaving by the old, so that, to a first approximation, the aggregate saving ratio, while in the long run independent of the *level* of national income, would depend on the sum of its population and per capita real income growth rates. Modigliani (1986), in his Nobel address, has given an account of how very simple stylized models of saving and refinement yield quite accurate predictions of the saving ratio and of the ratio of wealth to national income, and the predictions about the growth effects have been repeatedly borne out in international comparisons of saving rates, see Modigliani (1970), Houthakker (1961, 1965), Leff (1969) and Surrey (1974). Perhaps the only problem with these interpretations is that there is little evidence that the old actually dissave, except by running down state social security or pension schemes; see for example Mirer (1979). Partly, this may be a rational response to uncertainty about the date of death and about possible medical expenses near the end of life (Davies, 1980), partly there may be statistical problems of measurement (Shorrocks, 1975), and partly consumers may wish to leave bequests. However, most countries' tax systems penalize donors who do not pass on assets prior to death, so the reason for the size of actual bequests remains something of a mystery. Bernheim, Schleiffer and Summers (1985) have gone so far as to suggest that parents retain their wealth until death in order to control their heirs and to solicit attention from them. They claim empirical support for a positive relationship between visits by children to their parents and parents' bequeathable assets; visits are apparently especially frequent to rich sick parents, but not at all frequent to poor sick parents. Related to the dispute about the reason for bequests is a parallel dispute on their importance in the transmission of the capital stock, see the original contribution by Kotlikoff and Summers (1981) and Modigliani's reply, summarized in his (1986) Nobel lecture.

The life-cycle and permanent income models also provided the econometric specifications for a generation of macroeconometric models. Ando and Modigliani (1963) suggested a simple form for the aggregate consumption function in which real aggregate consumption was a linear function of expected

real labour income, *YL*, and of the real value of financial assets, i.e.

$$c_t = \alpha E_t(YL) + \delta W_t. \tag{16}$$

In practical econometric work, the expectation was typically replaced by a linear function of current and past values of labour income, a procedure that can be formally justified by modelling labour income as a linear ARIMA process, a topic to which I shall return below. Wealth or a subset of wealth was included as data allowed, although sometimes the return to wealth was included with labour income which could then be replaced by total income, so that, with smoothing, (16) becomes a permanent (total) income model of consumption. A favourite variant, suggested in Friedman (1957), was to model permanent income as an infinite moving average of current income with geometrically declining weights,

$$y_t^p = (1-\lambda)\sum_{r=0}^{\infty}\lambda^r y_{t-r} \tag{17}$$

so that if current consumption is proportional to permanent income, substitution yields

$$c_t = kc_{t-1} + k(1-\lambda)y_t. \tag{18}$$

a formulation that is also easy to defend if consumers ‘partially adjust’ to changes in current income. Models like (18), possibly with additional lags, and with the occasional appearance of more or less ‘exotic’ regressors, such as wealth, interest rates, inflation rates, money supply, as well as various dummy variables for ‘problem’ observations, were the standard fare of macroeconometric models in their heyday, from the early sixties for about a decade and a half. They fit the data well, they accounted for the smoothness of consumption relative to income, and they accorded at least roughly with the general features of the life-cycle and permanent income formulations which provided them with pedigree and general theoretical legitimacy. Dozens of papers could be cited within this tradition; those by Stone (1964, 1966), Evans (1967), and Davidson et al. (1978) will perhaps stand as good examples.

3 Recent econometric experience

In the mid-1970s, the general state of complacency of macroeconomic modelling was rapidly eroded, largely by the apparent inability of the standard models to explain, let alone to predict, the coexistence of unemployment and inflation. The relationship between consumption and income did not escape some of the blame, although the main focus of attack was elsewhere. Standard consumption functions, which had worked well into the early seventies, seriously under-predicted aggregate saving during the period of (at least relatively) rapid inflation that characterized most Western economies in the middle of the decade. The implementation of the theory of the consumption function was also singled out for discussion in Lucas's famous (1976) essay that became known as the Lucas ‘critique’. As Lucas forcefully argued, if consumption is determined by the discounted present value of *expected* future incomes, the response of consumption to a change in income is not well-defined until we know how expectations of income are formed. Each observed realization will cause a re-evaluation of future prospects in accordance with formulae that depend on the nature of the stochastic process governing income. If the nature of the stochastic process is changed, for example by a fundamental change in the tax code, then the way in which information is processed will change, and new information about incomes will have different implications for future expectations and for future consumption. This insight is of great importance, although its implications for econometric modelling were initially taken much too negatively; if the rules keep changing, econometric models will be inherently unstable (as evidenced by their performance in the mid-seventies) and we should give up trying to find stable relationships. Instead, as events have shown, the introduction of rational expectations has given a whole new lease of life to the study of consumption, with developments as positive as anything that has happened since the life-cycle and permanent income models were the ‘new’ theories in the mid-fifties. Lucas's critique suggested at least two lines for research. First, could the failure of consumption functions, or indeed of macroeconometric models in general, really be traced to a change in the way expectations were formed? If so, it ought to be possible to detect changes in the stochastic process generating real income. Second, and more generally, if expectations are important, there ought to be high returns to the simultaneous modelling of consumption and income, so that knowledge of the structure of the latter can be used either to estimate the consumption function or to test for the validity of the expectations mechanism. My own reading of the evidence is that the Lucas critique is *not* capable of explaining the failure of the empirical consumption function, but that the under-prediction of saving resulted from ignorance of the fact that saving appears to respond positively to inflation, or at least to unanticipated inflation. There is overwhelming evidence from a large number of countries, see in particular Koskela and Viren (1982a, 1982b), that saving increased with inflation in the 1970s, even when we allow for real income and its various lags. Such a finding is also consistent with the life-cycle theory since unanticipated inflation imparts a negative shock to real assets, so that risk-averse, low inter-temporal elasticity consumers will save to replace the lost assets so as to avoid the chance of low consumption later. It is also possible to explain the relationship through the confusion between relative and absolute price changes that is engendered by unanticipated inflation in an environment in which goods are bought sequentially, see Deaton (1977), but it would be hard to devise a test that would separate this from the life-cycle explanation. But if inflation was indeed the cause of the failure of the empirical consumption functions, then it is a standard enough story. An important variable was omitted from the analysis, it had not been very variable in the past so that its omission was hard to detect, and economists had not been imaginative enough to perceive its importance in advance. The Lucas critique is only one of the many problems that can beset an econometric equation, and it does not seem to have been the fatal one in this

case.

The second research direction, the joint examination of income and consumption, has proved more productive. The first important step was taken by Hall (1978), who pointed out that equation (15) implies that, as an approximation consumption should follow a random walk with drift. To see why, assume that the real interest rate *r* is constant and known, and write (15) in the form

$$c_{t+1}^{-\sigma} = \{(1+\delta)/(1+r)\}c_t^{-\sigma} + \varepsilon_{t+1} \tag{19}$$

where the expectation at *t* of ε_{t+1} is zero. Equation (19) is exact, but a convenient expression can be reached by factoring c_t out of the right hand side, taking logarithms, and approximating. This gives

$$\ln c_{t+1} = \ln c_t + g + v_{t+1} \tag{20}$$

where *g* is positive or negative as *r* is greater than or less than δ, and the 'innovation' v_{t+1}, like ε_{t+}, has expectation zero at time *t*. Equation (20) shows that, in the absence of 'news', consumption will grow or decline at a steady rate *g*, so that nothing that is known by the consumer at time *t* or earlier should have any value for predicting the deviation of the rate of change of consumption from its constant mean. The result is often referred to as the 'random walk' property of consumption, though the theory does not predict that v_{t+1} has constant variance, so that, strictly speaking, the stochastic process is not a random walk.

For someone used to thinking about the consumption function as the relationship between consumption and income, Equation (2) is notable for the apparent absence of any reference to income. But of course income can appear through the stochastic term v_{t+1} if current income contains new information about its own value or about future values of income, and this will generally be the case. The random walk model does not predict that consumption should not respond to current income. It does however predict that, conditional on lagged consumption, past income or changes in income should not be correlated with the current change in consumption, and a considerable amount of effort has recently gone into testing this proposition. In Hall's (1978) original paper, to the surprise of the author and of much of the profession, the model worked well for an aggregate of United States consumption of non-durables and services. The level of consumption certainly depends on its own lagged value, but the addition of one or more lagged values of income or of further lagged values of consumption did not significantly add to the explanatory power of the model. Hall examined the role of the number of other lagged variables and discovered that lagged stockmarket prices had predictive power for the change in consumption, so that he concluded by

formally rejecting the model. However, the overwhelming impression was favourable, at least relative to expectations.

Hall's test procedures are attractive because they do not depend on the properties of the income process, and focus only on consumption and its lags. But robustness comes at the price of power, and later work has devoted considerable attention to the joint properties of consumption and real income. Perhaps the natural route to modelling is to find a representation of real income as a stochastic process, typically as some sort of ARIMA. Once this is known, changes in income can be decomposed into anticipated and unanticipated components using the standard forecasting formulae from statistical time series analysis, so that it becomes possible to test whether consumption responds to one but not to the other. The random walk model seemed not to survive these tests so well. Papers by Flavin (1981) and by Hayashi (1982) showed that, for United States data, consumption is sensitive to *anticipated* changes in income, something that should not be the case in a thoroughgoing life-cycle model in which consumers are efficiently looking into the future. The phenomenon became known as the ‘excess sensitivity’ result, and was typically ascribed to the existence of a substantial number of consumers who wish to borrow against future income but are unable to do so. Such liquidity constrained consumers can be expected to consume all their available income, so that their consumption will increase one for one with all income changes, whether anticipated or not.

However, it is not clear that the excess sensitivity finding is itself robust. First, it is becoming increasingly recognized that the problems of econometric testing in the time-series models are more severe than had been generally supposed. The time series of both consumption and income are non-stationary, and it sometimes seems as if hypothesis testing in models involving non-stationary variables is like building on shifting sands; see Mankiw and Shapiro (1985, 1986) and Durlauf and Phillips (1986) for some of the problems. Second, there are a large number of variables other than income which can affect consumption, so that, according to (20), surprises in wealth and in inflation should affect consumption, as should the level of real interest rates. Adding even a few of these variables reduces degrees of freedom and diminishes the probability of being able to reject the basic model. Both Bean (1985) and Blinder and Deaton (1985) find that time-series models of consumption with several variables are more easily reconciled with the theory than are the simple two variable models. Not all of this should be ascribed to lack of degrees of freedom; for example Blinder and Deaton consistently find that unanticipated changes in wealth affect consumption and that anticipated changes do not. Third, even in a bivariate income-consumption model, Campbell (1987) has found that the model is largely consistent with the time-series evidence. Campbell recognizes the possibility of time-series feedback from lagged consumption to income, and models saving and the change in income as a bivariate vector-autoregressive system in which each series is regressed on lagged values of both. The structure of this representation then turns out to be very close to what it would have to be if the life-cycle rational expectations model were correct. The conflict between Campbell's results and the excess sensitivity findings are presumably accounted for by the feedback from saving to changes in labour income, since his model is otherwise compatible with the earlier ones.

Similarly mixed findings are also being uncovered from longitudinal panels that follow individual households over time. In contrast to the situation with labour supply, there are few panel data in the United States that cover household consumption, and most work has used the data on expenditure on food that is contained in the Michigan Panel Study of Income Dynamics (PSID). In an elegant paper,

Hall and Mishkin (1982) found results that were in accord with the excess sensitivity results; there is a strong negative correlation in their data between changes in consumption and changes in lagged income that is inconsistent with the view that only surprises in income should matter. However, since in their data changes in income are negatively correlated over time, a negative correlation between the lagged income change and the change in consumption can be interpreted as a positive correlation between consumption changes and changes in actual income, as predicted by the model of liquidity constraints. Hall and Mishkin conclude that these results would be consistent with a model in which about one fifth of consumers were unable to borrow as much as they wished. Once again, these results were supported by other similar evidence, see in particular Zeldes (1985) and Bernanke (1984), also using the PSID, Runkle (1983), using data from the Denver Income Maintenance Experiment, and Hayashi (1985a) using panel data from Japan. However, one potential problem with the use of panels is the importance of errors of measurement in such data. There is a considerable body of evidence that PSID income changes are subject to very substantial reporting errors, see in particular Altonji (1986), Duncan and Hill (1985), and Abowd and Card (1985). Altonji and Siow (1985) have recently estimated a model similar to Hall and Mishkin's using the PSID but with allowance for measurement error, and they find little conflict with the view that consumption responds only to news. However, it is unclear, at least to this reader, whether the acceptance of the model represents low power once errors of measurement are allowed for, or whether such errors really offer a plausible explanation for Hall and Mishkin's findings.

A more formal line of research has attempted to estimate the Euler condition (15) directly, thus avoiding the approximations made by Hall and by others. Rewrite (15) once more, this time as

$$(1 + r_{t+1})(c_{t+1})^{-\sigma} - (1 + \delta)(c_t)^{-\sigma} = \varepsilon_{t+1} \tag{21}$$

where, as before ε_{t+1} is orthogonal to any variable known in period *t* or earlier. Hansen and Singleton (1982) proposed that the parameters in (21) be estimated by a generalized methods of moments scheme. Suppose that we have two variables or instruments z_{1t} and z_{2t}, each known at time *t*, so that we have $E_t(z_{it}\varepsilon_{t+1}) = 0$ for $i = 1, 2$. We can then estimate the two unknown parameters, σ and Δ, by equating sample and theoretical moments, and solving the two equations, $i = 1, 2$

$$T^{-1}\sum_{t=0}^{T-1}[z_{it}\{(1 + r_{t+1})(c_{t+1})^{-\sigma} - (1 + \delta)(c_t)^{-\sigma}\}] = 0. \tag{22}$$

If, as is typically the case, we have more than two *z*-variables, then it will not generally be possible to

choose the two parameters so that (22) is exactly zero. Instead, the vector can be made as small as possible, or more specifically, the parameters can be estimated by minimizing a quadratic form that can be thought of as a weighted sum of squares of the left-hand side of (22); see Hansen and Singleton for details. If the model were true, this minimized value ought to be small, so that with more instruments than parameters, the generalized method of moments procedure yields a test-statistic that is diagnostic for model adequacy.

Test procedures based directly on the Euler conditions have several notable advantages. As was the case for Hall's procedures, few assumptions have to be made about the structure of the income process, and the model satisfies the best professional standards of seeking a direct confrontation between theory and data with as few approximations and supplementary assumptions as possible. The model can also be readily extended to test the implications of the consumption asset pricing model by repeating the tests using the returns on a range of alternative assets, see (13) above. Hansen and Singleton's study, as well as several others, find that the test statistics are much too large to be consistent with the theory and so reject the intertemporal model implied by the Euler conditions. Given the apparent superiority of the tests, these results have been accorded a great deal of weight in the literature. However, while I believe that Hansen and Singleton's work represents a very important methodological advance, I think that there are good reasons for not treating their results as a definitive rejection of life-cycle theory. The high level of technique that is embodied in deriving the Euler equation, not to mention the complexity of generalized methods of moments estimation, should not blind us to the very simple, even simple-minded, economic story that underlies these models. Fundamentally, the Euler equation says that the marginal rate of substitution between today's and tomorrow's consumption should be equal to the rate of return on assets between today and tomorrow, so that estimation of the Euler equation, unlike the Hall or excess-sensitivity tests, focuses very directly on the relationship between real interest rates and changes in real consumption, and the model will not fit the data if there is no close association between the two. And it only takes a very cursory inspection of United States time-series data to see that there is no such association. Real consumption grew in all but one year between 1954 and 1984, while real after-tax interest rates were as often negative as positive, so that consistency with the theory would require that the pure rate of time preference be negative. Nor is there any association between the rate of growth of consumption and the level of real after-tax interest rates, see Deaton (1986b) for some data. But this in no way reflects badly on the life-cycle theory. As was made perfectly clear in the original Modigliani and Brumberg papers, and it is the *essence* of the life-cycle model, aggregate consumption cannot be expected to behave like individual consumption. Imagine a stationary economy with neither population nor real income growth, in which there is an excess of real interest rates over the rate of pure time preference, and in which all consumers have identical additive life-time preferences with isoelastic subutility functions. In such an economy, each individual has a consumption path that is growing over time, but aggregate consumption is constant, a result that is achieved by old people dying and being replaced by young people who have much lower consumption levels relative to their incomes. Unless we believe that there is some automatic and immediate relationship between real interest rates, time preference and growth, as would obtain for example along a ‘golden age’ growth path, or unless we believe that consumers have infinite lives, then there is no reason at all to suppose that aggregate consumption should look at all like the life-cycle path of a representative consumer. Representative agent models are frequently useful, and it is not very constructive to dismiss macroeconomics because it requires implausible aggregation assumptions. However, the life-cycle model provides a well-worked-

out account of individual and aggregate saving, an account that is consistent with a good deal of other evidence and theory, and it *does not* predict that aggregate consumption should be consistent with the intertemporal optimization conditions for a single individual. The general question of the effects of interest rates on consumption is something that has remained in dispute for a long time, and in spite of repeated attempts to isolate the effect, careful studies have tended to be unable to do so, or at least to find effects that are at all robust, or that can be replicated on even slightly different data sets or data periods. Economic theories or policy prescriptions that rely on intertemporal substitution of consumption in response to changes in real interest rates are not well-buttressed by any solid body of empirical evidence.

Another useful approach to testing the life-cycle model is to consider the stylized facts of the income and consumption processes, and to see whether consumption behaves in the way that is to be expected given the stochastic process of income. Most people who have studied the time series for quarterly real disposable income in the United States agree that, like GDP, the series can be parsimoniously described by a model that is linear in its first two lags, i.e. an autoregression of the form

$$y_t = \alpha_1 + \alpha_2 y_{t-1} + \alpha_3 y_{t-2} + u_t \tag{23}$$

where u_t is the income innovation, that part of current income that cannot be anticipated from previous observation of the series. Of course, real income is not a stationary series, but has a strong upward trend, and there is considerable disagreement about the nature of this trend, what is the economic story behind it, and how it should be modelled. One possibility is that real income contains a *deterministic* time trend, so that there is some sort of equilibrium growth path that cannot be altered by shocks to the economy. Shocks certainly exist, but they cause only short term temporary deviations from the path and have little or no long-term temporary deviations from the path and have little or no long-term significance. In this view, equation (23) applies to the *deviations* of income from trend, not to income itself; equivalently, (23) can be modified by including a linear or quadratic time trend. The alternative view is that there is no deterministic trend, but that the rate of change of income is a stationary stochastic series with constant mean. In practice, this can look very like the previous model, but there is the vital conceptual difference that in the second, non-deterministic model, there is nothing that will ever bring income back to any deterministic path. In consequence, shocks to current income have permanent and long-lasting effects. The version of (23) that corresponds to this view can be written.

$$\{(y_t - y_{t-1}) - \gamma\} = \rho\{(y_{t-1} - y_{t-2}) - \gamma\} + u_t \tag{24}$$

which can readily be seen to be a special case of (23), though note that it is the case where the time

series possesses a unit root, or is stationary in first differences. For (24) to be a valid specialization of (23), the quadratic equation with the α's of (23) as coefficients must have a unit root, hence the term. Equation (24) appears to fit the data well and the parameter ρ turns out to be around 0.4, so that (24) says that if the increase in real income in one quarter is greater than its long term mean, then the next quarter's increase is also likely to be above the mean, though by less. While the long-term mean of the rate of change of income is constant and equal to *y*, good fortune (positive *u*'s) and bad fortune (negative *u*'s) never have to be paid for (or made up), since shocks are immediately consolidated into the income level, and growth goes on in the same way as before, but from the new base. As Campbell and Mankiw (1986) have emphasized, the unit root model exhibits shock *persistence*, while the deterministic trend model does not; they suggest that shock persistence is what we should expect if supply shocks predominate over demand shocks, with the reverse in standard Keynesian models where shocks are typically attributed to fluctuations in aggregate demand.

It turns out that it is almost impossible to tell these two processes apart on United States time-series data. Processes with unit roots are inherently difficult to tell apart from processes that are stationary around deterministic trends, and the tests that are available, Dickey and Fuller (1981), Phillips and Perron (1986), certainly cannot reject the hypothesis that (24) is a valid specialization of (23). Nor would the tests convince a believer in the deterministic model that income does not have a deterministic trend, even though it will readily be recognized that the deviations from trend are themselves close to non-stationarity. Since both process are special cases of (23) with the inclusion of a trend, and since each assumes parameter values that are very close to one another, one might think (and hope) that the two models would have very similar implications. But it is easy to see this is not true. If permanent income is taken as the annuity value of discounted future incomes, then (24) implies that any innovation u_t to current income, because it will persist forever, and because it can be expected to be followed by another infinitely persistent innovation of the same sign, will change permanent income by more than the amount of the innovation. Equation (25) below gives the formula for the change in permanent income, if the real interest rate is *r*, and if real income follows (24), see Flavin (1981) or Deaton (1986b),

$$\Delta y_t^p = \frac{(1+r)^2}{r+1-\rho} u_t \tag{25}$$

so that the change in permanent income is between one and a half and twice as large as the innovation in current income. By contrast, fitting the deterministic model yields a much smaller effect, with the change in permanent income about one fifth of the shock in measured income. Since consumption should change by about the same amount as does permanent income, the life-cycle model, together with the unit root formulation, yields the uncomfortable prediction that consumption should be *more* variable than income over the business-cycle, not less. If the unit root model is correct, then the life-cycle and permanent income models can be rejected because they predict what they were designed to predict, that consumption is smooth relative to real income! The deterministic model gives no such problems, but as

yet we have no way of being sure that it is correct, unless, of course we assume from the start that the life-cycle story is true.

There is insufficient space in this essay to follow these issues further, or to discuss in detail the evidence for and against the two formulations of the stochastic process governing real income; the interested reader can refer to Deaton (1986b) and to the evidence on persistence in GDP presented by Campbell and Mankiw (1986) and by Cochrane (1986). There are a number of possible solutions to these puzzles, and a great deal of empirical work remains to be done, though I suspect that the time-series data on income are insufficiently long to allow the isolation of the very long-run properties on which the permanent income theory rests, see in particular the interesting paper by Watson (1986).

4 Variations on the basic theme

There exist many interesting developments of the basic life-cycle model, and I have space to discuss only a few. I have already mentioned the role of liquidity constraints, and many people would take it as transparent that many consumers do not have access to unlimited credit, or else face borrowing rates that are higher than the rates at which they can lend. Of course, many consumers may be able to smooth their consumption without recourse to borrowing, and the borrowing needs of many others may be met by the typically rather good markets in home mortgages. For consumers who nevertheless wish to borrow but cannot, their spending will be closely tied to their actual income. For some of the theoretical and empirical literature on this point see Flemming (1973), Dolde and Tobin (1971), and Hayashi (1985b).

The theoretical consequences of uncertainty about the date of death have been worked out by Yaari (1965), and as argued above, play a possibly important part in the explanation of the saving behaviour of the elderly.

Another line of research is the possible relaxation of the assumption that preferences are intertemporally additive. Allowing all periods (or ages) to interact with all other periods in an unrestricted way, as in equation (1), would be much too general to be useful, and the search has been for simple models that break the restriction in a natural and straightforward way. One useful analogy is with the theory of durable good purchases, where utility depends on the *stock* of assets possessed, the stock in turn being the integral of past purchases less depreciation. Purchases in one period therefore have consequences for utility in subsequent periods, something that will be taken into account by a forward looking consumer. In the case of durable goods, the assumption of perfect capital markets effectively converts durable into non-durable goods, with the price of a unit of stock for one period being the implicit rental or user cost, the latter being defined as the sum of interest cost, depreciation, and expected capital loss, see for example Diewert (1974) or Deaton and Muellbauer (1980b, ch. 13).

However, various authors, Houthakker and Taylor (1970) perhaps being the first, have extended the durable model to encompass ‘psychic’ stocks which, like physical stocks, are augmented by purchases and diminished by depreciation, but unlike physical stocks, can either increase or decrease utility. The latter case covers habit formation; consumption of an addictive good generates pleasure now, but engenders a hungry habit that is pleasureless but costly in the future. The model has been given an elegant formulation in two papers by Spinnewyn (1979a, 1979b). As an example, see also Muellbauer (1985), take the utility function

$$u_t = \sum_{k=0}^{T-t} (1+\delta)^{-k} v(c_{t+k} - \alpha c_{t+k-1})$$
(26)

where α is a measure of habit formation. Spinnewyn maximizes this function with respect not to c_t, but with respect to the 'net' quantities $z_t = c_t - \alpha c_{t-1}$, and shows how to rewrite the budget constraint so as to define corresponding prices of the *z*'s that reflect not only market prices of the goods, but also the costs of consumption now in terms of pleasure foregone later. Under certainty, and looking ahead from time *t*, the full shadow price of an additional unit of consumption now is

$$p_z = \sum_{k=0}^{T-t} [\alpha / (1+r)]^k p_{t+k}$$
(27)

because the habits that are built up now have to be paid for later. Note that this sort of formulation also predicts that it is $c_t - \alpha c_{t-1}$ not c_t, that is proportional to permanent income, so that consumption itself will adjust only sluggishly to changes in permanent income with habits causing a drag. Other formulations of non-separable preferences can be found in the papers by Kydland and Prescott (1982), and by Eichenbaum, Hansen, and Singleton (1984), both of which are concerned to reconcile fluctuations in the aggregate economy with the behaviour of a single representative agent.

Many of the models discussed so far assume that the consumption function actually exists, hence taking for granted the essentially keynesian assumption that income is given to the consumer, and is not chosen together with consumption. A considerable body of work has grown up in the last ten years that is concerned with the simultaneous choice of labour supply and consumption in a life-cycle setting. Heckman (1971) and Ghez and Becker (1975) are among the pioneers of this approach. Unlike the price of goods, the price of leisure tends to show a systematic pattern over the life-cycle, so that, if consumers are free to choose their hours, and if they can freely borrow and lend so as to transfer resources between periods, it will pay them to work hardest during those periods in their life-cycles when the rewards for doing so are highest, and to take their life-time leisure when wage rates are low and leisure is cheap. There is superficial evidence in favour of this story, and Ghez and Becker, followed by Smith (1977) and Browning, Deaton and Irish (1985), all find that workers tend to work longest hours in middle age when wage rates are high and the lowest number of hours at the beginning and end of the economically active life, when wage rates are relatively low. Consumption also tends to peak in the middle age, and this can be brought into the story by assuming that consumption and leisure are complements, so that the lack of leisure in middle age is partially compensated by high levels of expenditure. This elegant fable

has also been made much of in equilibrium theories of the business cycle, which accounts ‘unemployment’ as a voluntary vacation taken when the real wage is low and leisure is on sale, see in particular Lucas and Rapping (1969) and Lucas (1981).

There now exists a growing volume of literature that shows just how much violence to the facts is done by this story. All the evidence quoted above looks across different individuals at different points in their life-cycles, while the theory says that the same individual will change his or her hours of work along with changes in the real wage over the life-cycle. Time-series and panel data from the United States and time-series of cross-sections from the United Kingdom suggest that this is simply not the case, see for example Mankiw, Rotemberg and Summers (1985), Ashenfelter and Ham (1979), Ashenfelter (1984), and Browning, Deaton and Irish (1985). Even MaCurdy's (1981) more postive study provides only very weak evidence, see in particular Altonji (1986). The joint consumption and labour supply story fares even less well than the labour supply model alone, and there is clear evidence that the way in which consumption and hours fluctuate over the cycle (sometimes together and sometimes in opposite directions) is not consistent with the way in which they move together over the life-cycle. The attempt to provide a unified theory of business and life-cycles has been an interesting and important one, but it cannot be said to have been successful.

I have been somewhat cavalier in my treatment of aggregation issues, choosing to emphasize them when I believe them to be important, for example in the fitting of Euler conditions, and ignoring them when it has been convenient to do so. Attempts to do better than this have not been notably successful. Formal conditions that allow aggregation in consumption function models are typically too restrictive to be useful, so that, in theory, changes in the distribution of income should have detectable effects on aggregate consumption. However, attempts such as that by Blinder (1975) to link the distribution of income to consumption have not been notably successful, perhaps because the income distribution is not variable, or because it changes smoothly enough over time to preserve a stable relationship between average income and average consumption. There is also an issue of aggregation over goods in order to define real consumption at all, even at the level of the individual agent. In the derivation in section 1 above, I made the convenient assumption that within-period preferences were homothetic, so that an index number of real consumption could be formed. But homotheticity, although very convenient for studying the consumption function, is very inconvenient for studying the allocation of expenditure among goods since it implies that the within-period total expenditure elasticities of each good are all equal to unity. Fortunately, there are aggregation results of Gorman's (1959), see also Deaton and Muellbauer (1980b, ch. 5) for an exposition that allows us to have the best of both worlds, at least if we remain with intertemporally additive preferences. If the single-period indirect utility function $\psi(x, p)$ takes the form known as the ‘generalized Gorman polar form’

$$\psi(x, p) = F[x / a(p)] + b(p) \tag{28}$$

where $a(p)$ and $b(p)$ are linearly homogeneous functions of prices and $F(\cdot)$ is monotone increasing, then

the real expenditure index $x/a(p)$ can serve as an indicator of real consumption just as in the homothetic case. This happens because when the consumer chooses the allocation of life-time expenditure over periods so as to maximize the intertemporal sum of terms like (28), the $b(p)$ terms are irrelevant. However, the intra-period demand functions that correspond to (28) do not display unitary elasticities unless the $b(p)$ is identically equal to zero, and quite general functional forms are permitted. There is therefore no real conflict between the analysis of the consumption function on the one hand, and the analysis of demand on the other. It is to the latter that I now turn.

5 Theoretical and empirical demand functions

Demand functions are the relationships between the purchase of individual goods, income or total expenditure, prices, and a variety of other factors depending on the context. Economists have attempted to make empirical links between demand and price since Gregory King's famous demand curve for wheat, see Davenant (1699), and since the middle of the 19th century, there has been a great development in the theory of consumer behaviour. Much practical work continues in the tradition of King, paying little attention to formal theory, concerning itself instead with finding empirical regularities. For a firm studying the demand for its product, or for anyone interested in establishing a single price elasticity, this probably remains the best approach; the major developments in econometric technique and empirical formulation have not been much concerned with, or relevant to, these very practical questions. The pragmatic approach (the term comes from Goldberger's famous but unpublished (1967) study), probably reached its peak with the publication of Richard Stone's great monograph, (Stone 1954a), and much is still to be learned by a careful study of Stone's procedures for measuring income and price elasticities. However, in this essay, I shall follow the literature, and follow its more methodological approach.

The theory outlined in Section 1 above suggests that the demand functions of an individual consumer can be derived by maximizing a utility function $v(q)$ subject to a budget constraint $p \cdot q = x$, where x is total expenditure. In the analysis here, x is chosen at some previous level of decision making, but traditionally it is treated as if it were a datum by the consumer, the utility maximization yields a vector q that is some function $g(x,\bullet p)$, say, of total expenditure and prices. These demand functions cannot simply be any functions, but must have certain properties as a result of their origins in utility maximization. Obviously, the total value of the demands should be equal to total outlay x, the ‘adding-up’ property, and it must be true that proportional changes in x and in p do not have any effect on quantities demanded, the ‘homogeneity’ or ‘absence of money illusion’ property. Somewhat less obvious are the famous symmetry and negativity properties. These apply to the Slutsky (1915) matrix, S, the typical element of which is defined as

$$s_{ij} = \partial q_i / \partial p_j + q_i \partial q_j / \partial x. \tag{29}$$

As any intermediate text shows, see for example Deaton and Muellbauer (1980b, ch. 2), the Slutsky matrix must be symmetric and negative semi-definite. The symmetry property is not readily turned into simple intuition; negativity implies that the diagonal elements of the matrix are non-positive, a proposition often referred to as ‘the law of demand’. The four properties, adding-up, homogeneity, symmetry and negativity, essentially exhaust the implications of utility maximization, so that any empirical demand functions that satisfy them can be regarded as having been generated by utility maximization, or by rational choice, with ‘rational’ defined, following Gorman (1981), as ‘having smooth strictly quasi-concave preferences, and being greedy’.

Stone (1954b) was the first to attempt to use this theory directly to confront the data. He started from a (general) linear expenditure system of the form

$$p_i q_i = \sum_j a_{ij} p_j + b_i x \tag{30}$$

where a_{ij} and b_i are unknown parameters. Stone showed that, in general, the system (30) does not satisfy the four requirements, but will do so if, and only if, the parameters are restricted so that the model can be written in the form

$$p_i q_i = p_i \gamma_i + \beta_i (x - p \cdot \gamma) \tag{31}$$

with the β-parameters summing to unity. In this form the model is known as the linear expenditure system. As Samuelson (1947–8) and Geary (1949–50) had earlier shown, the utility function corresponding to (31) has the form

$$u = \sum_i \sum \beta_i \ln (q_i - \gamma_i), \tag{32}$$

sometimes referred to (somewhat inappropriately) as the Stone–Geary utility function. It can be thought of as a sum of Bernoulli utility functions of the quantity of each good above the minimal γ's.

Stone's achievement lay not in deriving the demand functions, but in thinking to estimate them. The demand functions (30), even if fitted to the data by least-squares, require non-linear optimization, and

Stone invented a simple and not very efficient scheme, but one that allowed him to obtain parameter estimates and a good fit to interwar British data for a six commodity disaggregation of expenditures. This was a major breakthrough, not only in demand analysis, but also in applied econometrics in general. Indeed, much of demand analysis for a decade or so after Stone's paper consisted of applying better algorithms and faster computers to the fitting of Stone's model to different data sets.

The linear expenditure system offers a demand model for a system of, say *n* goods, and requires only $2n - 1$ parameters, a degree of parsimony that was very important in allowing the model to be estimated on very short time-series data. However, such economy brings its own price, and the linear expenditure system is very restrictive in the sort of behaviour that it can allow. In particular, and pathological cases apart, the model cannot allow inferior goods (goods the demand for which falls as total outlay increases), nor can it allow goods to be complements rather than substitutes. (As defined by Hicks (1939) goods *i* and *j* are complements if the (*i*,•*j*)th term in the Slutsky matrix is negative, so that the utility compensated cross-price response of *i* to an increase in the price of *j* is positive.) Normal (non-inferior) goods that are substitutes for one another may be the most important case, but they do not encompass everything that we might want to study. The linear expenditure system also implies that the marginal propensity to consume each good is the same no matter what is the total to be spent, and many cross-section studies of household budgets have suggested that this is not in fact the case.

Unfortunately, it is quite difficult to write down utility functions that will lead to more general demand functions than those of the linear expenditure system, nor is there any obvious way of generalizing Stone's procedure of writing down functions and making them consistent with the theory. Progress was only really made once applied demand analysis started using ‘dual’ formulations of preferences to specify demands. In the demand context, duality refers to a switch of variables, from quantities to prices, so that utility becomes a function, not directly of quantities consumed, but indirectly of prices and total expenditure. This indirect utility formulation is given by the function $\psi(x, p)$, already used above, and this is simply the maximum attainable utility from total outlay *x* at prices *p*. Since $\psi(x, p) = u$, and the function is monotone increasing in *x*, it can be inverted to give $x = c(u, p)$, known as the ‘cost function’, since it gives the minimum necessary cost that is required to reach the utility level *u*. By a theorem usually attributed to Shephard (1953) and to Uzawa (1964), these two functions contain a complete representation of preferences; provided preferences are convex, and provided the functions satisfy homogeneity and convexity (or concavity) conditions, preferences can be reconstructed from knowledge of either of the two functions. It is also very easy to move from either cost or indirect utility functions to the demand functions. For the indirect utility function, we have Roy's identity (Roy, 1943).

$$q = -\nabla_p \psi(x, p) / \psi_x(x, p) \equiv g(x, p) \tag{33}$$

which immediately yields demand functions from preferences in a form that are suitable for estimation, while for the cost function, we have Shepard's Lemma (1953),

$$q = \nabla_p c(u, p) = \nabla_p c[\psi(x, p), p] \equiv g(x, p)$$
(34)

where, as in (33), the operator ∇ denotes a vector of partial derivatives.

Demand analysis now had a high road to specification. Think of some quasi-convex decreasing function of the ratios of price to total outlay and call it an indirect utility function, or think of some function of utility and prices that is increasing in its arguments and linearly homogeneous and concave in prices and call it a cost function. Either way, and with only simple differentiation, new (and sometimes) interesting demand functions will be generated. Alternatively, and even more importantly, it is possible to use theory to aid and check out empirical knowledge. If it is known that the marginal propensity to spend on food is a declining function of total expenditure, or if it is thought likely that some goods do not depend very directly on the prices of other goods, it is relatively straightforward to find out what preferences (if any) will yield the result. It becomes possible, not just to generate demand functions serendipitously, but to generate good and useful ones deliberately.

There are many examples that could be cited from the literature. One of the most widely used in the *translog* model which was first proposed in 1970 by Jorgenson and Lau, see Christensen, Jorgenson and Lau (1973) for a convenient reference. To derive the translog, write the indirect utility function in terms of the ratios of prices to outlay, $r = p / x$, and approximate the indirect utility function as a second order polynomial in the logarithms of *r*. Application of Roy's identity yields demand functions in which the budget share of each good is the ratio of two functions, each of which is linear in the logarithms of the price to outlay ratios. Estimation of these rational functions, like estimation of the linear expenditure system, requires the use of non-linear maximization techniques. A related model, the ‘almost ideal demand system’ (AIDS) has been proposed by Deaton and Muellbauer (1980a), and I use this to illustrate some of the issues that arise with the current generation of demand models. The AIDS is specified by the logarithm of its cost function which takes the form

$$\ln c(u, p) = \alpha_0 + \sum_k \alpha_k \ln p_k + 0.5 \sum_k \sum_m \gamma_{km} \ln p_k \ln p_m + u \exp\left\{\sum_k \beta_k \ln p_k\right\},$$
(35)

so that, applying Shephard's lemma and rearranging, we have demand functions

$$p_i q_i / x \equiv w_i = \alpha_i + \beta_i \ln(x / P) + \sum_j \gamma_{ij} \ln p_j$$
(36)

where P is a linearly homogeneous price index, the form of which can readily be inferred from (35). The parameters of the model must satisfy certain restrictions if (35) is to be a proper (log) cost function, and (36) a proper system of demand functions. The matrix of γ -parameters can be taken to be symmetric in (35), but must be so in (36), and its rows and columns must add to zero for the homogeneity and adding-up properties to be satisfied. The β -parameters can be positive or negative, with positive values indicating luxury goods, and negative values necessities. The main advantage of the AIDS model in time-series applications is that the price index P can typically be approximated by some known price index selected before estimation, so that the demand system is linear in its parameters. In consequence, it can be estimated by ordinary least squares on an equation by equation basis, at least if the symmetry of the γ -matrix is ignored. The homogeneity restrictions can be tested equation by equation using a t- or F-test, and while imposing or testing symmetry requires an iterative procedure, estimation can be done by straightforward iterated restricted generalized least-squares, see Barten (1979) or Deaton (1974a) for further discussion.

The results of estimating the AIDS model are sufficiently similar to those from other models and other studies, see e.g. Barten (1969), Deaton (1974a), Christensen, Jorgenson and Lau (1973), and many others, that perhaps they can be taken as representative. What typically seems to happen is that the homogeneity restrictions appear *not* to be satisfied, so that in the application of AIDS to British data, Deaton and Muellbauer found, for example, that the F-test for transport had a value of 172 compared with the 5 per cent critical value of 4.8. Results on symmetry from AIDS and other systems are more mixed, and it now seems clear that testing symmetry is not usually possible given the amount of data typically available in time series, or put more positively, that there is no convincing evidence against symmetry. The difficulty is that symmetry involves a set of restrictions *across* different equations, so that unlike homogeneity, which involves tests *within* each equation, exact, small sample tests are not available. Researchers have therefore fallen back on asymptotically valid tests, and it turns out that these work very badly for the usual sort of samples, especially when there are more than a very small number of goods in the demand system. The papers by Laitinen (1978) and Meisner (1979) first established the problem, see also Evans and Savin (1982) and Bera, Byron and Jarque (1981) for further evidence.

The AIDS model, like the translog and several others, e.g. Diewert's (1973) ‘generalized Leontief’ system, fall into the class of ‘flexible functional forms’. This criterion of flexibility, first proposed by Diewert (1971), is an important guarantee that the model is sufficiently richly parametrized so as to allow estimation of what are thought to be the main parameters of interest, typically the total expenditure elasticities, and the matrix of own and cross-price elasticities. A ‘second order’ flexible functional form is one that has sufficient parameters, so configured, that it is possible to set the value of the function, and of its first and second partial derivatives to any arbitrary set of (theoretically permissible) values. By applying Roy's identity or Shephard's lemma, it is clear that a cost or indirect utility function that is a second order flexible functional form will yield demand functions that are first-order flexible, so that it is possible for estimation to yield any set of price and expenditure elasticities that are consistent with utility theory. For empirical work, such a guarantee is important, because it ensures that the elasticities are being measured, not assumed. Contrast, for example, the linear expenditure system (31) with the AIDS model (36). Both could be fitted to the same set of data, and the parameter estimates of each could be used to generate a complete set of expenditure and price elasticities. But the linear expenditure

system is *not* a flexible functional form, and so its estimated elasticities are not independent of one another, as is apparent from the fact that there are $2n - 1$ parameters compared with the total number of potentially independent elasticities, which is $(n - 1)(1 + n/2)$. (There are $n - 1$ independent demand equations, each of which has an expenditure elasticity, and *n* price elasticities; however, one price elasticity per equation is lost to homogeneity, and symmetry imposes a further $(n - 1)(n - 2)/2$ constraints.) The linear expenditure system does not therefore *measure* all the price and income elasticities, but determines them by a mixture of measurement and assumption, the main assumption being that of additive preferences, see Deaton (1974b) for further details. The AIDS, by contrast, has exactly the right number of parameters to allow for intercepts and a full set of elasticities, so that when it (or the translog, or the generalized Leontief) is estimated, so is the full set of elasticities.

Being able to do this is a great step forward in methodology, but just as the linear expenditure system probably asks too little of modern data, (although not of the data available to Stone and the early pioneers of the systems approach), the second-order flexible functional forms probably ask too much, or equivalently, put too little structure on the problem. The consequences show up in large standard errors, a high frequency of apparently chance correlations, and a lack of robustness to functional form changes within the class of flexible functional forms, in other words, in all the standard symptoms of over-parametrization. These problems are particularly acute for the measurement of *price* elasticities, because in most time-series data, commodity prices tend to move together with relatively little variation in relative prices. And although the focus of most research on demand analysis over the last thirty years has been on the estimation and testing of price responses, there is certainly no consensus on what numbers, if any, are correct. Estimates obtained from the linear expenditure system are not credible because they are forced to satisfy an implausibly restrictive structure, while those from flexible functional forms are not credible because the data are not informative enough to supplement the lack of prior structure. Some intermediate forms are clearly required.

One of the attractions of flexible functional forms is their ability to approximate quite general forms for preferences. However, the models so far considered offer only approximations, and there is no guarantee that they have satisfactory *global* properties. Partly this is the standard problem that a fitted model will be forced to give a reasonable account of the data over the sample used for estimation, but may predict very badly elsewhere. But there are other deeper issues. Taking the AIDS as an example, estimation of (36) subject to symmetry and homogeneity will produce a system of estimated demand functions that will satisfy adding-up, homogeneity and symmetry for *all* values of *x* and *p*. However, there are two other important properties that are not assured. First, there is no guarantee that the predicted budget shares will necessarily lie between zero and one, so that there may be regions of price space in which the estimated model yields nonsensical predictions. Second, there is no way that the AIDS can be guaranteed to have a negative semi-definite Slutsky matrix for all prices, at least not without restricting parameters to the point where the model ceases to be a flexible functional form. The parameters could be chosen so as to satisfy negativity for some particular combination of prices and outlay, but there will be no guarantee that the law of demand will be satisfied elsewhere. In the translog model, it is possible to impose a restriction that guarantees negativity everywhere, but the model with the restriction has the property that all estimated own price elasticities must be less than minus one, independently of whether this is in fact true, and it almost certainly is not, see Diewert and Wales (1987). A demand system is described as ‘regular’ if it has a negative definite Slutsky matrix and predicts positive demands, and several empirical studies, see e.g. Wales (1977) for one of the first, found that estimated flexible

functional forms were not regular over disturbingly large regions of even the parameter space used to estimate them. Caves and Christensen (1980), and later Barnett and Lee (1985) and Barnett, Lee and Wolfe (1985), investigated the same problem theoretically by taking a known utility function, choosing the parameters of flexible functional forms to match its level and derivatives at a point, and then mapping out the regions of price space in which the systems remained regular. The results at least for the translog and the generalized Leontief model, were not good.

These regularity issues may seem of limited importance in practice, but this is far from being the case. One of the major reasons for being interested in complete empirical demand systems is to be able to examine the consequences of price changes, particularly of price changes that follow changes in government policy. The United States relies relatively little on indirect taxation as a source of public finance, but such is not the case in most of Europe, and the vast majority of developing countries maintain complex systems of price wedges, particularly for foods and for agricultural production. The effects of such systems cannot be predicted without good information on how demands respond to price changes, nor can reforms be intelligently discussed. However, estimated demand systems that are not regular are not a great deal of help. All of the theory of welfare economics, of consumer surplus, of optimal taxation and of tax reform, *assumes* that demand behaviour is generated by utility maximization at the individual level, and implementation without regularity risks internal contradiction. For example, if compensated demand functions slope *upwards*, the government can generate a dead-weight gain by imposing a distortionary tax. Of course, it may not be the empirical work that is wrong, but the theory that we used to try to model behaviour. If so, the estimated demand functions are still not useful, since we now have no idea what to do with them. But I doubt that evidence goes so far; it is not that behaviour itself is irregular, but that we have not yet found a good modelling strategy that contains a reasonable amount of prior information to supplement the paucity of data, and at the same time can deliver global regularity if it is warranted by the evidence.

A number of interesting experiments are currently under way that involve new modelling techniques. One possibility is that the Taylor series expansions that motivate most flexible functional forms are themselves inadequate to the task. In particular, Taylor approximations lose their ability to approximate if they are also asked to possess other properties of the functions that they are approximating. For example, we might want to test whether or not preferences are additively separable, as in the linear expenditure system. One strategy would be to write down some second-order approximation to preferences, estimate the resulting demand model, and then test whether or not the conditions imposed on the demands by additivity are satisfied. But this will not work in general, because there may be no additive system of demand equations that has the precise functional form demanded by the approximation. The same phenomenon is well illustrated by Stone's derivation of the linear expenditure system itself. The original general linear expenditure equations (29) can clearly be justified as a Taylor approximation to any set of homogeneous demand functions, and yet the imposition of only *symmetry* generates the demand system (30) which comes from the *additive* utility function (31). Additivity is not imposed, but linear expenditure systems are only symmetric if they are additive. Similarly many flexible function forms are only globally regular if they are homothetic, see for example, Blackorby, Primont and Russell (1977). Several recent studies have proposed alternative ways of making functional approximations. Gallant (1982) has proposed using Fourier series approximations while Barnett (1983) has suggested that Laurent series can be used to generate demand models with good properties. Gallant's

models are even more heavily parametrized than standard flexible functional forms, and there must be some question as to the suitability of trigonometrical functions for demand functions. Barnett's ‘miniflex Laurent’ model does not use the full flexibility of the Laurent series, but appears to have quite good approximation and regularity properties in practice, see Barnett and Lee (1985) and Barnett, Lee and Wolfe (1985); even so, its estimation is complex, and many of the parameters have to be estimated subject to inequality constraints.

A second line of current research has abandoned the standard approach of econometric analysis, taking instead a completely non-parametric approach. Since many of the difficulties discussed above arise from choice of functional form, it is useful to ask how far it is possible to go without assuming any functional form at all. We know from standard revealed preference theory that two observed vectors of prices and quantities can be inconsistent with utility maximization; if bundle one is chosen when bundle two is available, so that bundle one is revealed preferred to bundle two, then no subsequent choice should reveal bundle two to be preferred to bundle one. Before embarking on the exercise of fitting some specific utility function to any finite collection of price and quantity pairs, one might then ask whether the collection is conceivably consistent with any set of preferences. If it is, then contradictions between an estimated system and the theory must be a matter of inappropriate functional form. The conditions for utility consistency of a finite set of data were originally derived by Afriat (1967), who proposed a condition called cyclical consistency. Much later Varian (1982) not only provided an accessible and clear account of Afriat's results, but also recast the cyclical consistency condition into a ‘generalized axiom of revealed preference (GARP)’ that runs as follows. A bundle q^i is strictly directly revealed preferred to a bundle q if $p^i q^i > p^i q$, while q^i is revealed preferred to q, if there exists a sequence, $j, k, \ldots, m$ such that $p^i q^i \geq p^i q^j$, $p^j q^j \geq p^j q^k, \ldots, p^m q^m \geq p^m q$, so that q^i is directly or indirectly (weakly) revealed preferred to q. GARP is satisfied if for all q^i revealed preferred to q^j, it is not true that q^j is strictly directly revealed preferred to q^i, and given GARP the data can be rationalized by a continuous, strictly concave, and non-satiated utility function. Differentiability can also be ensured by a sight strengthening of GARP, see Chiappori and Rochet (1987). GARP is readily tested for any given set of data by checking the pairwise inequalities and using a simple algorithm provided by Varian to map out the patterns of indirect revealed preference. Repeated applications of the method to time-series data have nearly always confirmed the consistency of the data with the theory. In retrospect, it is clear that violations of GARP cannot occur unless some budget lines intersect, so that if, over time, economic growth has resulted in the aggregate budget line moving steadily outward with little change in slopes, GARP is bound to be satisfied. (However, post-war United States data budget planes do occasionally intersect, and Bronars (1987) has recently shown that hypothetical demands generated by selecting random points on the actual budget lines would more often than not fail GARP.)

The contradictions between the parametric and non-parametric approaches can perhaps be resolved by thinking of the latter as a modelling technique that uses a very large number of parameters, so that the failure of the parametric models to fit theory to data can be thought of as failure to parametrize the models sufficiently richly. But I have already argued that these models already have too many parameters, and adding more would only exacerbate the already serious problems of measurement. For many purposes, the theory is only useful if it is capable of delivering a description of the data that is reasonably parsimonious. There is also something rather simple minded about non-parametric

techniques that tends to be disguised by the sophisticated and elegant expositions that have been given them by Varian and others. Consider a very simple theory that says variable *x* should move directly with variable *y* as, for example, in the Euler equation (15) above which says that, under certainty consumption should grow from period *t* to *t*+1 if and only if the real interest rate from *t* to *t*+1 is greater than some fixed constant. A non-parametric test on a finite set of data would accept the theory if, in fact, *x*, and *y* always did move together, and reject it if *x* and *y* ever moved in opposite directions. That such testing procedures are widely employed in the press and by the uninformed public is no reason for treating them seriously in economics.

I have so far discussed the formulation and estimation of demand functions, meaning the relationships between quantities, outlay, and prices, and this has been the topic of most applied demand analysis over the last 30 years. However, there is an older tradition of demand analysis, in which the object of attention is household budget data, and this literature has recently been enjoying something of a revival. Since household budget data typically come from a cross-section of households over a short period of time, usually within a single year, prices are treated as common to all sample points, so that the focus of attention becomes the relationship between demand and outlay and the influence of household composition on the pattern of household expenditures. The oldest, and perhaps only law of economics, Engel's Law that the share of food in the budget declines as total outlay increases, comes from Engel's (1857, published 1895) study of Belgian working-class families, and early empirical studies of demand were almost inevitably based on household surveys (see Stigler (1954) for a masterly review). The modern study of Engel curves, the relationships between expenditure and total outlay, begins (and almost ended) with Prais and Houthakker (1955). Prais and Houthakker studied the shapes of Engel curves, the relationship between demand and households, particularly in relation to the choice of quality, a topic that has subsequently been unjustly neglected. The functional forms for Engel curves that Prais and Houthakker examined became the staple menu for most subsequent studies, even though only one of their forms, the linear Engel curve, is capable of satisfying adding-up, and the linear form typically performs very badly on the data. Since 1955 a number of other Engel curves have been proposed, notably the lognormal Engel curve of Aitchison and Brown (1957), and Leser's (1963) revival of the form suggested much earlier by Holbrook Working (1943). Working's form, which apparently escaped the attention of Prais and Houthakker, makes the budget share of each commodity a linear function of the logarithm of total outlay. The formulation is particularly useful, for not only is it capable of accounting for most of the curvature that is discovered in empirical Engel curves, but it is also consistent with utility theory, and corresponds to the case where the welfare elasticity of the cost of living is independent of income. Gorman (1981) has provided a general characterization theorem for Engel curves of the form

$$p_i q_i = \sum_k a_{ik}(p)\,\xi_k(x) \tag{37}$$

and has shown that the $\xi_k(\cdot)$ functions can be powers of x (polynomial Engel curves), or x multiplied by powers of log x (Engel curves relating budget shares to powers of the logarithm of outlay), or have trigonometric forms. This last form includes Fourier representations of Engel curves, while the first two allow Taylor or Laurent expansions for the expenditure/outlay and for the share/log-outlay forms. The Working–Engel curve is the first member of Gorman's ‘share to log’ class, and the theorem tells us that we may add quadratic or higher order terms to improve the fit. However, Gorman's paper contains a remarkable result; the matrix of the *a*-coefficients in (37) has rank at most equal to three. In consequence, the share to log and log-squared Engel curves are as general as any, as are the Engel curves of the quadratic expenditure system, see Howe, Pollak and Wales (1979). Given Gorman's results, and the empirical success of the Working form, it and its quadratic generalization deserve wide use in the analysis of budget studies. There is also accumulating evidence that such forms are indeed necessary. Thomas (1986), in a wide-ranging examination of household survey data from developing countries, has shown that Engel's Law itself does not appear to hold among the very poor, so that, in many cases, the share of the budget devoted to food at first rises with total outlay before falling in conformity with the Law.

Prais and Houthakker also proposed a much-used formulation for the effects of household composition on behaviour. It can be written

$$p_i q_i / m_i(a) = f_i\{x / m_0(a)\} \tag{38}$$

where a is a vector of household demographic characteristics (perhaps a list of numbers of people in each age and sex category) and m_i and m_0 are scalar valued functions known as the ‘specific’ and ‘general scales’ respectively. In this literature, scales are devices that convert family structure into numbers of equivalent adults, so that a family of two adults and two children might be two equivalent adults for theatre entertainment, three equivalent adults for food, and six equivalent adults for milk. The general scale is supposed to reflect the overall number of equivalent adults, so that the Prais and Houthakker model is a simple generalization of the idea that *per capita* demand should be a function of *per capita* outlay. Barten (1964), in a very important paper, took up the Prais–Houthakker idea of specific scales, but assumed that the arguments of the household utility function were the household consumption levels each deflated by the corresponding specific scale. The consequences of Barten's formulation are similar to those of Prais and Houthakker, but embody the additional insight that changes in family composition affect the effective shadow prices of goods, so that demographic changes will exercise, not only income, but also substitution effects on the pattern of demand. The story is often summarized by the phrase, ‘if you have a wife and child, a penny bun costs three-pence’, quoted in Gorman (1976), but the really far-reaching substitution effects of children are probably on time use and labour supply, particularly of women.

Since household surveys typically contain large samples of households, there is less need for theory to save degrees of freedom, and it is possible to estimate quite general functional forms that link expenditures to household composition patterns and then to interpret the results in terms of the various

models. In addition, neither the Prais–Houthakker nor the Barten model seem to yield easily implemented functional forms, e.g. linear ones, nor is it clear that either model is even identified on a single cross-sectional household survey in which all prices are constant, see for example Muellbauer (1980) and Deaton (1986a). However, some empirical results for the two models can be found in Muellbauer (1977, 1980) and in Pollak and Wales (1980, 1981) who also examine Gorman's (1976) extension of Barten's model in which additional people are supposed to bring with them fixed needs for particular commodities. The fixed needs model is close to the formulation proposed by Rothbarth (1943) for measuring the costs of children. Rothbarth pointed out that there are certain commodities, adult goods, that are not consumed by children, so that when children are added to a household, the only effects on the household's consumption of adult goods will be the income effects that reflect the fact that, with unchanged total resources, the household is now poorer. Deaton, Ruiz-Castillo and Thomas (1985) have recently attempted to test Rothbarth's contention, and in their Spanish data it seems possible to identify a sensible group of adult goods, the expenditure on each of which changes with additional children in the same way as they change in response to changes in outlay.

Studies of the effects of family composition on household expenditure patterns have frequently been concerned, not only with estimating demands, but also with attempts to measure the ‘cost’ of children. It would take me too far afield to do justice to this topic here. Readers interested in this controversial area should perhaps start with Rothbarth (1943), who in a few pages makes a very simple and quite convincing case, and look also at Nicholson (1976). Pollak and Wales (1979) weigh in on the opposite side, and claim that it is impossible to measure child costs from expenditure data. My own position is argued in Deaton and Muellbauer (1986); there are certainly grave problems to be overcome in moving from the analysis of household survey data to the measurement of the costs of children, and it is clear that identifying assumptions must be made that are more severe and more controversial than those required, for example, to go from demand functions to consumer surplus. But that does not mean that it is not possible for such assumptions to be proposed and to be sensibly discussed.

See Also

- bequests and the life cycle model
- consumer expenditure (new developments and the state of research)
- demand theory
- Euler equations
- Friedman, Milton
- rational expectations

Bibliography

Abowd, J.M. and Card, D. 1985. The covariance structure of earnings and hours changes in three panel data sets. Princeton University, mimeo.

Afriat, S.N. 1967. The construction of a utility function from expenditure data. *International Economic Review* 8, 67–77.

Aitchison, J. and Brown, J.A.C. 1957. *The Lognormal Distribution*. Cambridge: Cambridge University Press.

Altonji, J. 1986. Intertemporal substitution in labor supply: evidence from micro data. *Journal of Political Economy* 94, S176–215.

Altonji, J. and Siow, A. 1985. Testing the response of consumption to income changes with (noisy) panel data. Industrial Relations Section Working Paper No.186, Princeton University, mimeo.

Ando, A. and Modigliani, F. 1963. The life-cycle hypothesis of saving: aggregate implications and tests. *American Economic Review* 53, 55–84.

Ashenfelter, O. 1984. Macroeconomic analyses and microeconomic analyses of labor supply. Presented to Carnegie-Rochester Conference, Bal Harbor, Florida, November 1983.

Ashenfelter, O. and Ham, J. 1979. Education, unemployment, and earnings. *Journal of Political Economy* 87, S99–116.

Barnett, W.A. 1983. New indices of money supply and the flexible Laurent demand system. *Journal of Business and Economic Statistics* 1, 7–23.

Barnett, W.A. and Lee, Y.W. 1985. The global properties of the miniflex Laurent, generalized Leontief, and translog flexible functional forms. *Econometrica* 53, 1421–37.

Barnett, W.A., Lee, Y.W. and Wolfe, M. 1985. The three dimensional global properties of the miniflex Laurent, generalized Leontief, and translog flexible functional forms. *Journal of Econometrics*, 3–31.

Barten, A.P. 1964. Family composition, prices, and expenditure patterns. In *Econometric Analysis for National Economic Planning*, ed. P.E. Hart, G. Mills, and J.K. Whitaker. London: Butterworth.

Barten, A.P. 1969. Maximum likelihood estimation of a complete system of demand equations. *European Economic Review* 1, 7–23.

Bean, C.R. 1985. The estimation of surprise models and the surprise consumption function. Centre for Economic Policy Research (London), Discussion Paper No.54, mimeo.

Bera, A.K., Byron, R. and Jarque, C.M. 1981. Further evidence on asymptotic tests for homogeneity in large demand systems. *Economics Letters* 8, 101–5.

Bernanke, B.S. 1984. Permanent income, liquidity, and expenditure on automobiles: evidence from panel data. *Quarterly Journal of Economics* 99, 587–614.

Bernheim, B.D., Schleiffer, A. and Summers, L.H. 1985. Bequests as a means of payment. *Journal of Political Economy*, 1045–76.

Blackorby, C., Primont, D. and Russell, R.R. 1977. On testing separability restrictions with flexible functional forms. *Journal of Econometrics* 5, 195–209.

Blinder, A.S. 1975. Distribution effects and the aggregate consumption function. *Journal of Political Economy* 83, 447–75.

Blinder, A.S. and Deaton, A.S. 1985. The time series consumption function revisited. *Brookings Papers on Economic Activity* 2, 465–511.

Breeden, D. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7, 265–96.

Bronars, S.G. 1987. The power of non-parametric tests of preference maximization. *Econometrica* 55, 693–8.

Browning, M.J., Deaton, A.S. and Irish, M.J. 1985. A profitable approach to labor supply and commodity demands over the life cycle. *Econometrica* 53, 503–43.

Campbell, J.Y. 1987. Does saving anticipate declining labor income? An alternative test of the permanent income hypothesis. *Econometrica* 55, 1249–73.

Campbell, J.Y. and Mankiw, N.G. 1986. Are output fluctuations transitory? National Bureau of Economic Research Working Paper 1916, processed.

Caves, D.W. and Christensen, L.R. 1980. Global properties of flexible functional forms. *American Economic Review* 70, 422–32.

Chiappori, P.-A. and Rochet, J.-C. 1987. Revealed preferences and differentiable demand. *Econometrica* 55, 687–91.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55, 28–45.

Cochrane, J.H. 1986. How big is the random walk in GNP? Department of Economics, University of Chicago, processed.

Davenant, C. 1699. *Essay upon the Probable Methods of Making a People Gainers in the Balance of Trade*. London.

Davidson, J.E.H. et al. 1978. Econometric modelling of the aggregate time-series relationship between consumers expenditure and income in the United Kingdom. *Economic Journal* 88, 661–92.

Davies, J.B. 1980. Uncertain lifetime, consumption and dissaving in retirement. *Journal of Political Economy* 89, 561–77.

Deaton, A.S. 1974a. The analysis of consumer demand in the United Kingdom, 1900–1970. *Econometrica* 42, 341–67.

Deaton, A.S. 1974b. A reconsideration of the empirical implications of additive preferences. *Economic Journal* 84, 338–48.

Deaton, A.S. 1975. The structure of demand in Europe 1920–1970. In *The Fontana Economic History of Europe*, ed. C.M. Cippola. London: Collins-Fontana, Vol. 5.

Deaton, A.S. 1977. Involuntary saving through unanticipated inflation. *American Economic Review* 67, 899–910.

Deaton, A.S. 1986a. Demand analysis. In *Handbook of Econometrics*, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland, Vol. 3.

Deaton, A.S. 1986b. Life-cycle models of consumption: is the evidence consistent with the theory? NBER Working Paper No.1910, processed.

Deaton, A.S. and Muellbauer, J. 1980a. An almost ideal demand system. *American Economic Review* 70, 312–26.

Deaton, A.S. and Muellbauer, J. 1980b. *Economics and Consumer Behaviour*. New York: Cambridge University Press.

Deaton, A.S. and Muellbauer, J. 1986. On measuring child costs, with applications to poor countries. *Journal of Political Economy* 94, 720–44.

Deaton, A.S., Ruiz-Castillo, J. and Thomas, D. 1985. The influence of household composition on household expenditure patterns: theory and Spanish evidence. Woodrow Wilson School, Princeton University, processed.

Dickey, D.A. and Fuller, W.A. 1981. Likelihood ratio statistics for autoregressive time series with a unit

root. *Econometrica* 49, 1057–72.

Diewert, W.E. 1971. An application of the Shephard duality theorem: a generalized Leontief production function. *Journal of Political Economy* 79, 481–507.

Diewert, W.E. 1973. Functional forms for profit and transformation functions. *Journal of Economic Theory* 6, 284–316.

Diewert, W.E. 1974. Intertemporal consumer theory and the demand for durables. *Econometrica* 42, 497–516.

Diewert, W.E. and Wales, T.J. 1987. Flexible functional forms and global curvature conditions. *Econometrica* 55, 43–68.

Dixit, A.K. 1976. *Optimization in Economic Theory*. Oxford: Oxford University Press.

Dolde, W. and Tobin, J. 1971. Monetary and Fiscal Effects on Consumption in Consumer Spending and Monetary Policy: The Linkages. Boston: Federal Reserve Bank of Boston, Conference Series no. 5.

Duncan, G.J. and Hill, D.H. 1985. An investigation of the extent and consequences of measurement error in labor economic survey data. *Journal of Labor Economics*, 508–32.

Durlauf, S.N. and Phillips, P.C.B. 1986. Trends versus random walks in time-series analysis. Cowles Foundation Discussion Paper No.788. Yale University, New Haven, processed.

Eichenbaum, M.S., Hansen, L.P. and Singleton, K. 1984. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, mimeo.

Engel, E. 1895. Die lebenkosten Belgischer Arbeiter-Familien früher und jetzt. *International Statistical Institute Bulletin* 9, 1–74.

Epstein, L. 1975. A disaggregate analysis of consumer choice under uncertainty. *Econometrica* 43, 877–92.

Evans, G.B.A. and Savin, N.E. 1982. Conflict among the criteria revisited; the W, LR, and LM tests. *Econometrica* 50, 737–48.

Evans, M.K. 1967. The importance of wealth in the consumption function. *Journal of Political Economy* 75, 335–51.

Fisher, I. 1930. *The Theory of Interest*. New York: The Macmillan Company.

Flavin, M. 1981. The adjustment of consumption to changing expectations about future income. *Journal of Political Economy* 89, 974–1009.

Flemming, J.S. 1973. The consumption function when capital markets are imperfect: the permanent income hypothesis reconsidered. *Oxford Economic Papers* 25, 160–72.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Gallant, A.R. 1982. Unbiased determination of production technologies. *Journal of Econometrics* 20, 285–323.

Geary, R.C. 1949–50. A note on ‘A constant utility index of the cost of living’. *Review of Economic Studies* 18, 65–6.

Ghez, G. and Becker, G.S. 1975. *The Allocation of Time and Goods over the Life-cycle*. New York: Columbia University Press.

Goldberger, A.S. 1967. Functional form and utility: a review of consumer demand theory. Social Systems Research Institute, University of Wisconsin, processed.

Gorman, W.M. 1959. Separable utility and aggregation. *Econometrica* 27, 469–81.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis*, ed. M. Artis and A. R. Nobay. Cambridge: Cambridge University Press.

Gorman, W.M. 1981. Some Engel curves. In *Essays in the Theory and Measurement of Consumer Behaviour in honour of Sir Richard Stone*, ed. A.S. Deaton. Cambridge: Cambridge University Press.

Gorman, W.M. 1982. Facing an uncertain future. IMSS Technical Report No.359, Stanford University, processed.

Grossman, S.J. and Shiller, R.J. 1981. The determinants of the variability of stock market prices. *American Economic Review, Papers and Proceedings* 71, 222–7.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86, 971–87.

Hall, R.E. and Mishkin, F.S. 1982. The sensitivity of consumption to transitory income: estimates from panel data on households. *Econometrica* 50, 461–81.

Hansen, L.P. and Singleton, K.J. 1982. Generalized instrumental variables estimation of non-linear rational expectations models. *Econometrica* 50, 1269–86.

Hayashi, F. 1982. The permanent income hypothesis: estimation and testing by instrumental variables. *Journal of Political Economy* 90, 895–916.

Hayashi, F. 1985a. Permanent income hypothesis and consumption durability: analysis based on Japanese panel data. *Quarterly Journal of Economics*, 183–206c.

Hayashi, F. 1985b. Tests for liquidity constraints: a critical survey. Osaka University and NBER, processed. Presented at the Fifth World Congress of the Econometric Society, Cambridge, MA, August 1985.

Heckman, J.J. 1971. Three essays on the supply of labor and the demand for goods. Unpublished PhD thesis, Princeton University.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Oxford University Press.

Houthakker, H.S. 1961. An international comparison of personal saving. *Bulletin of the International Statistical Institute* 38, 55–70.

Houthakker, H.S. 1965. On some determinants of saving in developed and underdeveloped countries. In *Problems in Economic Development*, ed. A.G. Robinson. London: Macmillan.

Houthakker, H.S. and Taylor, L.D. 1970. *Consumer Demand in the United States: Analysis and Projections*. 2nd edn, Cambridge, MA: Harvard University Press.

Howe, H., Pollak, R.A. and Wales, T.J. 1979. Theory and time series estimation of the quadratic expenditure system. *Econometrica* 47, 1231–47.

Koskela, E. and Viren, M. 1982a. Saving and inflation: some international evidence. *Economics Letters* 9, 337–44.

Koskela, E. and Viren, M. 1982b. Inflation and savings: testing Deaton's hypothesis. *Applied Economics* 14, 579–90.

Kotlikoff, L.J. and Summers, L.H. 1981. The role of intergenerational transfers in aggregate capital accumulation. *Journal of Political Economy* 89, 706–32.

Kuznets, S. 1946. *National Income: A summary of findings*. National Bureau of Economic Research. New York: Arno Press.

Kuznets, S. 1962. Quantitative aspects of the economic growth of nations: VII: the share and structure of consumption. *Economic Development and Cultural Change* 10, 1–92.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Laitinen, K. 1978. Why is demand homogeneity so often rejected? *Economics Letters* 1, 187–91.

Leff, N. 1969. Dependency rates and saving rates. *Economic Journal* 59, 886–96.

Leser, C.E.V. 1963. Forms of Engel functions. *Econometrica* 31, 694–703.

Lucas, R.E. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A. Meltzer, Carnegie-Rochester Conference Series on Public Policy 1, Amsterdam: North-Holland.

Lucas, R.E. 1981. Introduction. In *Studies in Business Cycle Theory*, ed. R.E. Lucas. Cambridge, MA: MIT Press.

Lucas, R.E. and Rapping, L. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77, 721–54.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 89, 1059–85.

Mankiw, N.G. and Shapiro, M. 1985. Trends, random walks, and tests of the permanent income hypothesis. *Journal of Monetary Economics* 16, 165–74.

Mankiw, N.G. and Shapiro, M. 1986. Do we reject too often? Small sample properties of tests of rational expectations models. *Economics Letters* 20, 139–45.

Mankiw, N.G., Rotemberg, J.J. and Summers, L.H. 1985. Intertemporal substitution in macroeconomics. *Quarterly Journal of Economics* 100, 225–51.

Meisner, J.F. 1979. The sad fate of the asymptotic Slutsky symmetry test. *Economics Letters* 2, 231–3.

Mirer, T.W. 1979. The wealth–age relationship among the aged. *American Economic Review* 69, 435–43.

Modigliani, F. 1970. The life-cycle hypothesis of saving and inter-country differences in the saving ratio. In *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, ed. W.A. Eltis et al.,

Oxford: Clarendon Press.

Modigliani, F. 1986. Life cycle, individual thrift, and the wealth of nations. *American Economic Review* 76, 297–313.

Modigliani, F. and Ando, A. 1957. Tests of the life-cycle hypothesis of savings. *Bulletin of the Oxford Institute of Economics and Statistics* 19, 99–124.

Modigliani, F. and Brumberg, R. 1954. Utility analysis and the consumption function: an interpretation of cross-section data. In *Post-Keynesian Economics*, ed. K.K. Kurihara. New Brunswick: Rutgers University Press.

Modigliani, F. and Brumberg 1979. Utility analysis and aggregate consumption functions: an attempt at integration. In *The Collected Papers of Franco Modigliani*, ed. A. Abel, Vol. 2, Cambridge, MA: MIT Press.

Muellbauer, J. 1977. Testing the Barten model of household composition effects and the cost of children. *Economic Journal* 87, 460–87.

Muellbauer, J. 1980. The estimation of the Prais–Houthakker model of equivalence scales. *Econometrica* 48, 153–76.

Muellbauer, J. 1985. Habits, rationality and the life-cycle consumption function. Nuffield College, Oxford, mimeo.

Nicholson, J.L. 1976. Appraisal of different methods of estimating equivalence scales and their results. *Review of Income and Wealth* 22, 1–11.

Phillips, P.C.B. and Perron, P. 1986. Testing for a unit root in a time series regression. Cowles Foundation Discussion Paper No.795, Yale University, New Haven, processed.

Pollak, R.A. and Wales, T.J. 1979. Welfare comparisons and equivalent scales. *American Economic Review* 69, 216–21.

Pollak, R.A. and Wales, T.J. 1980. Comparisons of the quadratic expenditure system and translog demand system with alternative specifications of demographic effects. *Econometrica* 48, 595–612.

Pollak, R.A. and Wales, T.J. 1981. Demographic variables in demand analysis, *Econometrica* 49, 1533–51.

Prais, S.J. and Houthakker, H.S. 1955. *The Analysis of Family Budgets*. Cambridge: Cambridge

University Press.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Rothbarth, E. 1943. Note on a method of determining equivalent income for families of different composition. Appendix 4 in C. Madge ed., *War-time Patterns of Saving and Spending*, Occasional Paper 4, National Institute of Economic and Social Research, London.

Roy, R. 1943. *De l'utiliteé: contribution à la théorie des choix*. Paris: Herman.

Runkle, D.E. 1983. Liquidity constraints and the permanent income hypothesis: evidence from panel data. MIT, processed.

Samuelson, P.A. 1947–8. Some implications of linearity. *Review of Economic Studies* 15, 88–90.

Shephard, R. 1953. *Cost and Production* Functions. Princeton: Princeton University Press.

Shorrocks, A.F. 1975. The age-wealth relationship: a cross-section and cohort analysis. *Review of Economics and Statistics* 57, 155–63.

Slutsky, E. 1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti* 15, 1–26. English translation in *Readings in Price Theory*, ed. G.J. Stigler and K. Boulding, Chicago: Chicago University Press, 1952.

Smith, J.P. 1977. Family labor supply over the life cycle. *Explorations in Economic Research* 4, 205–76.

Spinnewyn, F. 1979a. Rational habit formation. *European Economic Review* 15, 91–109.

Spinnewyn, F. 1979b. The cost of consumption and wealth in a model with habit formation. *Economics Letters* 2, 145–8.

Stigler, G.J. 1954. The early history of empirical studies of consumer behavior. *Journal of Political Economy* 62, 95–113.

Stone, J.R.N. 1954a. *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920–1938*, Volume 1. Cambridge: Cambridge University Press.

Stone, J.R.N. 1954b. Linear expenditure systems and demand analysis: an application to the pattern of British demand. *Economic Journal* 64, 511–27.

Stone, J.R.N. 1964. Private saving in Britain, past, present, and future. *The Manchester School* 32, 79–

112.

Stone, J.R.N. 1966. Spending and saving in relation to income and wealth. *L'Industria*, 471–99.

Surrey, M.J.C. 1974. Saving, growth, and the consumption function. *Bulletin of the Oxford Institute of Statistics* 36, 125–42.

Thomas, D. 1986. Essays on the analysis of Engel curves in developing countries. PhD thesis, Princeton University.

Uzawa, H. 1964. Duality principles in the theory of cost and production. *International Economic Review* 5, 216–20.

Varian, H.R. 1978. *Microeconomic Analysis*. 2nd edn, New York: Norton, 1984.

Varian, H.R. 1982. The non-parametric approach to demand analysis. *Econometrica* 50, 945–73.

Wales, T.J. 1977. On the flexibility of flexible functional forms: an empirical approach. *Journal of Econometrics* 5, 183–93.

Watson, M.W. 1986. Univariate detrending method with stochastic trends. *Journal of Monetary Economics* 18, 49–75.

Working, H. 1943. Statistical laws of family expenditure. *Journal of the American Statistical Association* 38, 43–56.

Yaari, M.E. 1965. Uncertain lifetime, life insurance, and the theory of the consumer. *Review of Economic Studies* 32, 137–50.

Zeldes, S. 1985. Consumption and liquidity constraints: an empirical investigation. The Wharton School, University of Pennsylvania, processed.

How to cite this article

Deaton, Angus. "consumer expenditure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000313> doi:10.1057/9780230226203.0296

The New Palgrave Dictionary of Economics Online

consumer surplus

Daniel T. Slesnick

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Over the years, consumer surplus has been used to measure the welfare effects of price and income changes. Despite its widespread use, it provides a measure of well-being that is ordinally equivalent to the change in utility only under conditions that are inconsistent with long-standing empirical evidence. Hicksian surplus measures, such as the equivalent or compensating variations, provide exact indicators of the change in utility without such restrictions. Beginning in the early 1980s, empirical methods have been developed to estimate the equivalent variation that has the same data requirements as consumer surplus.

Keywords

aggregation; compensating variation; consumer surplus; equivalent variation; expenditure function; indirect utility function; integrability of demand; intertemporal welfare effects; linear expenditure system; marginal utility of income; representative agent; Roy's Identity; social choice; social expenditure function; well-being

Article

How does the market power exercised by firms influence consumer welfare? What is the effect of excise taxes on households with different levels of income? Does governmental regulation increase the welfare of consumers? Topical issues such as these indicate that the measurement of welfare is a fundamental element of public policy analysis. Indeed, a full consideration of taxes, subsidies, transfer programmes, health care reform, regulation, environmental policy, the social security system, and educational reform must ultimately address the question of how these policies affect individual well-being.
While centrally important to many problems of economic analysis, confusion persists concerning the relationship between commonly used indicators of welfare and well-established theoretical formulations. For more than 150 years, consumer surplus has been used to measure the welfare effects of changes in

prices and incomes. Its popularity can be ascribed to its intuitive appeal, the ease with which it is implemented, and its modest data requirements. Although it is generally accepted that Dupuit (1844) was the originator of the concept of consumer surplus, it is largely attributed to Marshall (1890). (Chipman and Moore, 1976, provide a brief survey of the history of the debate related to consumer surplus.) We begin with the following notation:

- $\mathbf{p}=(p_1, p_2,\bullet\ldots\bullet, p_n)$ – a vector of commodity prices.
- Y_k – the income of individual k.
- A_k – a vector of demographic characteristics of individual k.
- $x_{ik}=x_i(\mathbf{p},Y_k,\mathbf{A_k})$ is the demand for good i by individual k.

Suppose we are interested in the welfare impact of a change in the price of a single commodity from p_1^0 to p_1^1. The change in consumer surplus is given by:

$$\Delta CS_k = -\int_{p_1^0}^{p_1^1} x_1(t, p_2, \ldots p_n, Y_k, \mathbf{A_k})\, dt. \tag{1}$$

If ΔCS_k is positive (negative), the price change is judged to have increased (decreased) the welfare of individual k. Is it ordinally equivalent to the change in utility? A necessary condition is that the demand function is generated by a rational consumer who maximizes utility subject to a budget constraint. Unless consumers have optimized and are at the boundaries of their budget sets, it is impossible to assess the welfare effects of changes in prices and incomes. (That is, demands must be ‘integrable’ and consistent with a well-behaved utility function. Hurwicz and Uzawa, 1971, provide a formal statement of the integrability conditions.)
If demands are consistent with rational consumer behaviour, an indirect utility function $V(\mathbf{p}, Y_k, \mathbf{A_k})$ represents the maximum utility attained at prices $\mathbf{p}$ and income Y_k, and Roy's Identity provides the link between demands and utility:

$$x_1(\mathbf{p}, Y_k, \mathbf{A_k}) = -\frac{\partial V(\mathbf{p}, Y_k, \mathbf{A_k}) / \partial p_1}{\partial V(\mathbf{p}, Y_k, \mathbf{A_k}) / \partial Y_k}. \tag{2}$$

If the marginal utility of income is constant, substitution of (2) into (1) yields an explicit expression for the change in consumer surplus that is ordinally equivalent to the change in utility:

$$\Delta CS_k = \frac{V(\mathbf{p}^1, Y_k, \mathbf{A_k}) - V(\mathbf{p}^0, Y_k, \mathbf{A_k})}{\partial V / \partial Y_k}.$$

While constancy of the marginal utility of income is restrictive, Chipman and Moore (1976; 1980) have shown that application of consumer surplus is more problematical if there are changes in more than one price. In such circumstances, the change in consumer surplus must be evaluated using a line integral defined over the path of price changes from $\mathbf{p}^0$ to $\mathbf{p}^1$:

$$\Delta CS_k = \int_{\mathbf{p}^0}^{\mathbf{p}^1} \sum_{\mathrm{i}} \mathrm{x_i}(\mathbf{p}, Y_k, \mathbf{A_k})\, \mathbf{dp_i}. \tag{3}$$

Price paths are not observed so it is essential that (3) be path independent. This holds if the uncompensated price effects are symmetric (see, for example, Angus Taylor and Robert Mann, 1972, pp. 500–4):

$$\frac{\partial x_i}{\partial p_j} = \frac{\partial x_j}{\partial p_i} \text{for } all\ i \neq j.$$

This form of symmetry requires preferences to be homothetic, which is a restriction that is inconsistent with well-established empirical regularities.
In the most general circumstance of changes in prices and income, consumer surplus is defined as:

$$\Delta CS_k = -\int_Z \sum x_{\mathrm{i}}(\mathbf{p}, Y_k, \mathbf{A_k})\, \mathrm{d}\, p_i + (Y_k^1 - Y_k^0), \tag{4}$$

where **Z** is a path between $(\mathbf{p}^0, Y_k^0)$ and $(\mathbf{p}^1, Y_k^1)$. Chipman and Moore (1976) have demonstrated that there are no circumstances under which (4) is path independent and ordinally equivalent to the change in

utility of a rational consumer.

Hicksian surplus measures

Given the problems with consumer surplus, how should the welfare effects of price and income changes be measured? Hicks (1942) developed an approach that is exactly analogous to (4) once we substitute compensated for uncompensated demand functions:

$$\Delta HS_k = -\int_Z \sum x_i^c(\mathbf{p}, V, \mathbf{A_k}) \mathrm{d}\, p_i + (Y_k^1 - Y_k^0) \tag{5}$$

where $x_i^c(\mathbf{p}, V, \mathbf{A_k})$ is the compensated demand for the ith good evaluated at utility level V. Compensated price effects are symmetric, so the line integral in (5) is path independent and the surplus measure is single-valued.

For simple binary comparisons of policies, the utility level at which ΔHS_k is evaluated is often treated as a matter of little consequence. If it is calculated at the utility attained at prices $\mathbf{p}^1$ and income Y_k^1 (denoted V^1), a generalized version of the equivalent variation is obtained:

$$EV_k = E(\mathbf{p}^0, V^1, \mathbf{A_k}) - E(\mathbf{p}^1, V^1, \mathbf{A_k}) + (Y_k^1 - Y_k^0) \tag{6}$$

$$= E(\mathbf{p}^0, V^1, \mathbf{A_k}) - E(\mathbf{p}^0, V^0, \mathbf{A_k})$$

where $E(\mathbf{p}, V, \mathbf{A_k})$ is the expenditure function, defined as the minimum income needed for individual k to attain utility V at prices $\mathbf{p}$. Not only is the generalized equivalent variation single-valued, but it is ordinally equivalent to the change in utility. That is, EV_k is positive if and only if $V^1 > V^0$.

The utility level at which (5) is evaluated is important for multiple comparisons of price and income changes. The generalized equivalent variation will give an ordering of outcomes that is identical to that based on utility levels. If (5) is evaluated at $V^0 = V(\mathbf{p}^0, Y_k^0, \mathbf{A_k})$, we obtain the generalized

compensating variation:

$$CV_k = E(\mathbf{p}^0, V^0, \mathbf{A_k}) - E(\mathbf{p}^1, V^0, \mathbf{A_k}) + (Y_k^1 - Y_k^0) \quad (7)$$

$$= E(\mathbf{p}^1, V^1, \mathbf{A_k}) - E(\mathbf{p}^1, V^0, \mathbf{A_k}).$$

Because the utility levels are ‘cardinalized’ using different prices for each set of binary comparisons, the ordering of multiple outcomes based on (7) need not match the ordering based on utility levels. Chipman and Moore (1980) have shown that consistent rankings of outcomes require restrictions on preferences that are the same as for consumer surplus.

While the simple static formulation of consumer surplus is the most frequent application, the conceptual framework can be extended to analyse the effects of changes in utility in more general settings. For example, intertemporal welfare effects are often represented as the discounted sum of the within-period equivalent or compensating variations.

Keen (1990) has shown that this will differ from the lifetime equivalent variation to the extent that individuals are able to substitute intertemporally. As an alternative approach, he defines V_L to be the maximum level of lifetime utility of an individual who lives T periods when the profiles of prices and interest rates are $\{\mathbf{p_t}\}$ and $\{r_t\}$ respectively. If the (optimal) time path of utility corresponding to V_L at these prices and interest rates is $\{V_{kt}\}$, the lifetime expenditure function can be represented as:

$$\Omega_L(\{\mathbf{p_t}\}, \{r_t\}, V_L) = \sum_t g_t\, E(\mathbf{p_t}, V_{kt}, \mathbf{A_{kt}}),$$

where $g_t = \Pi_{s=0}^{t}(1 + r_s)^{-1}$.

As in the static framework, the lifetime expenditure function can be used to represent an exact measure of the change in lifetime welfare. Define V_L^1 to be the maximum level of lifetime welfare when the profile of prices and interest rates are $\{\mathbf{p_t^1}\}$ and $\{r_t^1\}$ and denote the corresponding time path of utility as $\{V_{kt}^1\}$. The reference prices and interest rates, $\{\mathbf{p_t^0}\}$ and $\{r_t^0\}$, yield a lifetime utility level of V_L^0 and

within-period utilities $\{V^0_{kt}\}$. Keen's exact measure of the change in lifetime welfare, evaluated at the reference prices, is exactly analogous to the generalized equivalent variation:

$$\Delta W_L = \Omega_L(\{\mathbf{p}^0_t\}, \{r^0_t\}, V^1_L) - \Omega_L(\{\mathbf{p}^0_t\}, \{r^0_t\}, V^0_L).$$

The concepts of the equivalent and compensating variation can also be extended to cases in which the choices made by consumers are discrete rather than continuous. Dagsvik and Karlstrom (2005) describe the compensating variation in the context of a random utility model defined as:

$$U_{jk} = V(\mathbf{p_j}, Y_k, \mathbf{A_k}) + \varepsilon_{jk} \ (j = 1, 2, \ldots, J),$$

where U_{jk} is the utility of individual k in alternative j, $V(.)$ is a deterministic indirect utility function, and ε_{jk} are random variables. There are a total of J choices available to the consumer and, for simplicity, it is assumed that only prices vary across alternatives.
Consider the welfare effect of a change in the set of prices and income facing individual k from $(\mathbf{p}^0_1, \mathbf{p}^0_2, \ldots, \mathbf{p}^0_\mathbf{I}, Y^0_k)$ to $(\mathbf{p}^1_1, \mathbf{p}^1_2, \ldots, \mathbf{p}^1_\mathbf{I}, Y^1_k)$. If the consumer chooses the alternative that maximizes U_{jk}, the compensating variation is defined implicitly as that value CV_k that satisfies the following equality:

$$\max_j V(\mathbf{p}^0_\mathbf{j}, Y^0_k, \mathbf{A_k}) + \varepsilon_{jk} = \max_j V(\mathbf{p}^1_\mathbf{j}, Y^1_k - CV_k, \mathbf{A_k}) + \varepsilon_{jk}.$$

Although conceptually analogous to the equivalent and compensating variation described previously, CV_k is now random and cannot, in general, be represented in closed form.

From demand functions to welfare measurement

While it was understood that the equivalent variation resolved the conceptual problem of welfare measurement, it had little influence on applied welfare economics because compensated demand functions were presumed to be unobservable. Willig (1976) made the first attempt to bridge the gap between theory and application by showing that, for a single price change, consumer surplus can provide an approximation to the equivalent or compensating variation. However, with multiple price and income

changes, consumer surplus is not single-valued and is of no use in approximating changes in economic welfare (McKenzie, 1979).

Shortly after the publication of Willig's paper, however, empirical procedures were developed to estimate the equivalent or compensating variation. Each method begins with the specification of a demand function and, under the assumption of integrability, is used to recover the utility or expenditure functions. The complexity of this procedure diminishes if demand functions are linear, and consideration is restricted to changes in the price of a single good.

Hausman (1981) provided an analytic solution to this problem for a demand function given by:

$$x_1 = \gamma_p p_1 + \gamma_Y Y_k + \gamma_{\mathbf{A}} \mathbf{A_k},$$

where γ_p, γ_Y, and $\gamma_{\mathbf{A}}$ are unknown parameters to be estimated econometrically. Roy's Identity provides a partial differential equation that can be solved to obtain an expenditure function of the form:

$$E(p_1, V, \mathbf{A_k}) = Ve^{\gamma_Y p_1} - (1/\gamma_Y)[\gamma_p p_1 + (\gamma_p/\gamma_Y) + \gamma_{\mathbf{A}} \mathbf{A_k}]. \quad (8)$$

The expenditure function allows the equivalent variation to be computed exactly as in (6) and Willig-type approximations are unnecessary. Hausman's method has the same data requirements as consumer surplus, and only linear regression methods are needed to estimate the unknown parameters.

Closed form solutions to the partial differential equation implied by Roy's Identity can be obtained for only a limited class of demand functions. An alternative approach is to begin with an assumed form of the indirect utility function and use Roy's Identity to obtain a system of demand equations. Since the form of the utility function is assumed from the outset, it is unnecessary to solve a complex system of partial differential equations.

Muellbauer (1974) provided an early example of this approach. He assumed that demands were consistent with a Stone–Geary utility function given by:

$$V(\mathbf{p}, Y_k) = \frac{(Y_k - \Sigma p_i \delta_i)}{\Pi p_i^{\alpha_i}} \quad (9)$$

where $\delta = (\delta_1, \delta_2, \ldots, \delta_n)$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ are unknown parameters. The corresponding

expenditure function is:

$$E(\mathbf{p}, V) = \Sigma p_i \delta_i + V(\Pi p_i^{\alpha_i}).$$

The unknown parameters can be estimated by fitting the linear expenditure system to household budget data:

$$p_i x_i = p_i \delta_i + \alpha_i (Y_k - \Sigma p_i \delta_i) \quad (i = 1, 2, \ldots, n). \tag{10}$$

Given estimates of α and δ, the expenditure function can be used to compute the equivalent or compensating variation as in (6) and (7).

While this is more general than Hausman's approach, it has its own disadvantages. For an assumed form of the utility function, the functional forms of the demands are the same for every good, which may hinder the ability of the model to fit the data. Is it possible to start with an arbitrary demand system (rather than a utility function) and measure the welfare effects of multiple price changes? Two elegant procedures were proposed that required more complicated calculations to recover the expenditure function, but did not impose restrictions on the form of the demand functions other than the standard integrability conditions.

The first method is based on an approximation to McKenzie's (1957) indirect money metric utility function defined as:

$$\mu(\mathbf{p}, Y_k, \mathbf{A_k}; \mathbf{p}^0) = E(\mathbf{p}^0, V(\mathbf{p}, Y_k, \mathbf{A_k}), \mathbf{A_k}).$$

McKenzie and Pearce (1976) showed that $\Delta\mu$ can be approximated by a Taylor's series expansion about the initial equilibrium:

$$\Delta\mu = \frac{\partial \mu}{\partial \mathbf{p}'} \Delta\mathbf{p}(1/2) + \Delta\mathbf{p}' \frac{\partial^2 \mu}{\partial p \partial p'} \Delta\mathbf{p} + \left(\frac{\partial \mu}{\partial Y} + \frac{\partial^2 \mu}{\partial \mathbf{p} \partial Y'} \Delta\mathbf{p} + 1/2 \frac{\partial^2 \mu}{\partial Y^2} \Delta Y \right) \Delta Y + R \tag{11}$$

where R represents higher order terms in the series.
The expression in (11) can be represented as a function of uncompensated demand functions when μ is evaluated at the reference prices (this follows from Roy's Identity and from the fact that at these prices the marginal utility of income is equal to one and all higher income derivatives are zero – see McKenzie and Pearce, 1976, for details):

$$\Delta\mu = -\mathbf{x}'\Delta\mathbf{p} - (1/2)\Delta\mathbf{p}'\left(\frac{\partial\mathbf{x}}{\partial\mathrm{p}} - \mathbf{x}\frac{\partial\mathbf{x}}{\partial\mathbf{Y}'}\right)\Delta\mathbf{p} + \left(1 - \frac{\partial\mathbf{x}}{\partial\mathbf{Y}'}\Delta\mathbf{p}\right)\Delta Y + R. \tag{12}$$

Given knowledge of the demand functions and the magnitudes of the price and income effects, one has all of the information necessary to get as accurate an estimate of the change in utility as desired. Vartia (1983) developed an algorithm that recovers the expenditure function numerically to any desired level of accuracy. Let $\mathbf{p}(t)$ and $Y_k(t)$ be the paths of price and income changes for $0 \leq t \leq 1$. As prices and income change, the movements of demands along an indifference curve can be represented implicitly by the differential equation:

$$\frac{dY_k(t)}{dt} = \sum_{i=1}^{n} x_i(\mathbf{p}(t), Y_k(t), \mathbf{A_k})\frac{dp_i(t)}{dt}.$$

Integrating over t yields an expression that can, in principle, be solved to obtain $E(\mathbf{p}(t), V^0, \mathbf{A_k})$ which is the centrepiece of the welfare calculations:

$$E(\mathbf{p}(t), V^0, \mathbf{A_k}) - E(\mathbf{p}^0, V^0, \mathbf{A_k}) = \sum_{i=1}^{n}\int_0^t x_i(\mathbf{p}(t), E(\mathbf{p}(t), V^0, \mathbf{A_k}), \mathbf{A_k})\frac{dp_i(t)}{dt}dt. \tag{13}$$

Vartia described several algorithms that can be used to solve this equation numerically over the price path $\mathbf{p}(t)$ so that, when evaluated at t=1, we obtain $E(\mathbf{p}^1, V^0, \mathbf{A_k})$. As long as the demands satisfy the integrability conditions, the solution to (13) will be independent of the price path used in the algorithm. This method is valid for multiple price and expenditure changes and, because a closed-form solution is unnecessary, facilitates flexibility in estimating demand patterns.

Aggregation

The methods described to this point provide estimates of the change in welfare for individuals. In practice, analysts are more concerned about the impact of policies on groups. Micro-level estimates are an essential first step, but, for welfare economics to be useful to practitioners, a method of aggregation is essential. The easiest approach is to assume that market demands are generated by a representative consumer. Under this condition, the methods described previously can be applied to aggregate demands and the utility function of the representative agent can be recovered.

While frequently applied, this is unsatisfactory for a number of reasons. Market demands need not be consistent with a rational representative consumer. Even if every individual has demands that are consistent with utility maximization, aggregate demands need not satisfy any of the integrability conditions other than homogeneity of degree zero in prices and income (Sonnenschein, 1972). Moreover, it is unclear what this utility function actually represents. Kirman (1992) presents an example in which the representative agent prefers (aggregate) market basket *A* to *B* even though all individuals prefer the reverse. This violation of the most basic principle of social choice suggests that the utility of the representative agent should not be used for policy analysis even in the unlikely event that aggregate demands are integrable.

An alternative approach is to define aggregate welfare to be a function of the individual surplus measures. Such an approach was advocated by Harberger (1971) in his effort to make consumer surplus the standard tool for applied welfare analysis. At a conceptual level, such an indicator of aggregate welfare appears to be a natural extension of the positive analysis of welfare measurement at the micro level. This is obviously not the case because aggregation necessitates normative judgements in which the gains to some must be weighed against the losses to others. Simply summing the surplus measures, for example, embodies a version of utilitarianism and ignores distributional concerns.

Since any method of measuring welfare for groups of individuals necessarily involves subjective judgements, it seems reasonable to state explicitly the underlying ethical basis for the method of ordering outcomes in the aggregate. The social choice theoretic framework used by Sen (1970) provides a reasonable way of presenting the normative assumptions related to the measurability and comparability of individual welfare levels that facilitate well-behaved social orderings of outcomes. Under conditions described by Sen and others, these orderings can be represented by a social welfare function:

$$W = W(V_1, V_2, \ldots, V_K)$$

where V_k is a welfare indicator of individual k.

A monetary measure of social welfare can be obtained using Pollak's (1981) concept of a social expenditure function:

$$M(\mathbf{p}, W) = \min \{Y: W(V_1, \ldots, V_K) \geq W, \Sigma Y_k = Y\}.$$

This function is exactly analogous to its micro-level counterpart and is the minimum level of aggregate income required to attain a specified social welfare contour. If W^0 is the social welfare under policy 0 and W^1 is the welfare under policy 1, the monetary measure of the change is social welfare is exactly analogous to the generalized equivalent variation:

$$\Delta W = M(\mathbf{p}, W^1) - M(\mathbf{p}, W^0).$$

ΔW is clearly ordinally equivalent to the changes in social welfare, and normative judgements are represented explicitly through the specification of the social welfare function.

See Also

- cost–benefit analysis
- cost minimization and utility maximization
- Hicksian and Marshallian demands
- indirect utility function
- social welfare function

Bibliography

Chipman, J.S. and Moore, J. 1976. The scope of consumer's surplus arguments. In *Evolution, Welfare and Time in Economics: Essays in Honor of Nicholas Georgescu-Roegen*, ed. A.M. Tang, F.M. Westfield, and J.S. Worley. Lexington, MA: Heath-Lexington Books.

Chipman, J.S. and Moore, J. 1980. Compensating variation, consumer's surplus, and welfare. *American Economic Review* 70, 933–49.

Dagsvik, J. and Karlstrom, A. 2005. Compensating variation and Hicksian choice probabilities in random utility models that are nonlinear in income. *Review of Economic Studies* 72, 57–76.

Dupuit, J. 1844. On the measurement of the utility of public works. Trans R.H. Barback. In *Readings in Welfare Economics*, ed. K.J. Arrow and T. Scitovsky. Homewood, IL: Richard D. Irwin.

Harberger, A.C. 1971. Three basic postulates for applied welfare economics. *Journal of Economic*

Literature 9, 785–97.

Hausman, J.A. 1981. Exact consumer's surplus and deadweight loss. *American Economic Review* 71, 662–76.

Hicks, J.R. 1942. Consumer's surplus and index numbers. *Review of Economic Studies* 9(2), 126–37.

Hurwicz, L. and Uzawa, H. 1971. On the integrability of demand functions. In. *Preferences, Utility and Demand*, ed. J. Chipman et al. New York: Harcourt, Brace and Jovanovich.

Keen, M. 1990. Welfare analysis and intertemporal substitution. *Journal of Public Economics* 42, 47–66.

Kirman, A.P. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6(2), 117–36.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

McKenzie, L. 1957. Demand theory without a utility index. *Review of Economic Studies* 24(65), 185–9.

McKenzie, G.W. 1979. Consumer's surplus without apology: comment. *American Economic Review* 69, 465–8.

McKenzie, G.W. and Pearce, I. 1976. Exact measures of welfare and the cost of living. *Review of Economic Studies* 43, 465–8.

Muellbauer, J. 1974. Prices and inequality: the U.K. experience. *Economic Journal* 84(333), 32–55.

Pollak, R.A. 1981. The social cost of living index. *Journal of Public Economics* 15, 311–36.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden Day.

Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40, 549–63.

Taylor, A.E. and Mann, R. 1972. *Advanced Calculus*, 2nd edn. Lexington: Xerox College Publishing.

Vartia, Y.O. 1983. Efficient methods of measuring welfare change and compensated income in terms of ordinary demand functions. *Econometrica* 51, 79–98.

Willig, R.E. 1976. Consumer's surplus without apology. *American Economic Review* 66, 589–97.

How to cite this article

Slesnick, Daniel T. "consumer surplus." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000312> doi:10.1057/9780230226203.0298

The New Palgrave Dictionary of Economics Online

consumption externalities

Robert H. Frank

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Consumption externalities occur when consumption by some creates costs or benefits for others. According to Duesenberry's ‘relative income hypothesis’, spending is influenced by the individual's own standard of living in the recent past and the living standards of others in the present. This hypothesis tracks observed behaviour more closely than Friedman's ‘permanent income hypothesis’, which assumes that context has no influence on spending. When context is more important for some goods (positional goods) than for others (non-positional goods), positional goods crowd out non-positional goods, causing welfare losses like those that occur when bombs crowd out consumption in military arms races.

Keywords

bequest motive; consumption externalities; Friedman, M.; Hirsch, F.; Marx, K.; permanent income hypothesis; positional goods; relative income hypothesis; revealed preference; savings; Smith, A.; Veblen, T.

Article

Consumption externalities occur when consumption by some creates external costs or benefits for others. Their recognition by economists dates at least as far back as Adam Smith's discussion of how local consumption standards influence the goods that people consider essential (or ‘necessaries’, as Smith called them). In the following passage, for example, he described the factors that influence the amount someone must spend on clothing in order to be able appear in public ‘without shame’:

> By necessaries I understand, not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without. A linen shirt, for example, is, strictly speaking, not a necessary of life. The Greeks and Romans lived, I suppose, very

> comfortably, though they had no linen. But in the present times, through the greater part of Europe, a creditable day-labourer would be ashamed to appear in publick without a linen shirt, the want of which would be supposed to denote that disgraceful degree of poverty which, it is presumed, no body can well fall into without extreme bad conduct. (Smith, 1776, pp. 869–70)

Consumption externalities received only limited attention in Smith's *Wealth of Nations* and only occasional mention by economists during the century that followed its publication. Karl Marx (1847), for example, noted that ‘A house may be large or small; as long as the neighboring houses are likewise small, it satisfies all social requirement for a residence. But let there arise next to the little house a palace, and the little house shrinks to a hut.’

It was not until Thorstein Veblen's *The Theory of the Leisure Class* appeared in 1899 that consumption externalities received their first serious, book-length treatment in economics. Veblen's thesis was that much of consumption is undertaken to signal social position. But although his book is still widely read and cited by scholars in numerous disciplines, its general theme was largely ignored by economists during the 50 years following its publication.

Duesenberry's relative income hypothesis

Interest in this theme was rekindled with the publication of James Duesenberry's *Income, Saving, and the Theory of Consumer Behavior* in 1949. In this volume, Duesenberry offered his ‘relative income hypothesis’, in which he argued that an individual's spending behaviour is influenced by two important frames of reference – the individual's own standard of living in the recent past and the living standards of others in the present. Thus, in Duesenberry's account, people are subject to both intrapersonal and interpersonal consumption externalities.

His theory attempted to explain three important empirical regularities: (*a*) long-run aggregate savings rates remain roughly constant over time, even in the face of substantial income growth; (*b*) aggregate consumption is much more stable than aggregate income in the short run; and (*c*) individual savings rates rise substantially with income in cross-section data. When Duesenberry's book was first published, individual consumption was generally modelled by economists as a linear function of income with a positive intercept term. This model could accommodate rising savings rates in cross-section data and the stability of consumption over the business cycle, but not the long-run stability of aggregate savings rates.

Duesenberry's hypothesis was hailed as an advance because of its ability to track all three stylized fact patterns. The poor save at lower rates, he argued, because they are more likely to encounter others with desirable goods that are difficult to afford. Moreover, since this will be true no matter how much national income grows, unfavorable comparisons will always occur more frequently for the poor – and hence the absence of any tendency for savings rates to rise with income in the long run.

To explain why consumption is more stable than income in the short run, Duesenberry argued that families compare their living standards not only to those of others around them but also to their own standards from the past. The high consumption level once enjoyed by a formerly prosperous family thus constitutes a frame of reference that makes cutbacks difficult when income falls.

Despite Duesenberry's success in tracking the data, many economists felt uncomfortable with his relative

income hypothesis, which to them seemed more like sociology or psychology than economics. The profession was therefore immediately receptive to alternative theories that purported to explain the data without reference to softer disciplines. The most important among these theories was Milton Friedman's permanent income hypothesis, variants of which still dominate today's research on spending.

In hindsight, however, there remain grounds for scepticism about whether Friedman's theory was a real step forward. For example, its fundamental premise – that savings rates are independent of permanent income – has been refuted by numerous careful studies (see, for example, Carroll, 1998). Some modern consumption theorists have responded by positing a bequest motive for rich consumers, a move that begs the question of why leaving bequests should entail greater satisfaction for the rich than for the poor. Another problem is that, contrary to Friedman's assertion that the marginal propensity to consume out of windfall income should be nearly zero, people actually consume such income at almost the same rate as permanent income (Bodkin, 1959). To this observation, Friedman (1963) himself responded that consumers appear to have unexpectedly short planning horizons. But if so, then consumption does not really depend primarily on permanent income.

Abundant evidence suggests that context influences evaluations of living standards (see, for example, Veenhoven, 1993; Easterlin, 1995; Luttmer, 2005). In the light of this evidence, it seems fair to say that Duesenberry's hypothesis not only has been more successful than Friedman's in tracking how people actually spend but also rests on a more realistic model of human nature. And yet the relative income hypothesis is no longer even mentioned in most leading economics textbooks. Its absence appears to signal the profession's continuing reluctance to acknowledge concerns about relative consumption.

Welfare implications

In traditional economic models, individual utility depends only on absolute consumption. These models lie at the heart of claims that pursuit of individual self-interest promotes aggregate welfare. In contrast, models that include concerns about relative consumption identify a fundamental conflict between individual and social welfare. This conflict stems from the fact that concerns about relative consumption are stronger in some domains than in others. The disparity gives rise to expenditure arms races focused on ‘positional goods’ – those for which relative position matters most. The result is to divert resources from ‘non-positional goods’, causing welfare losses. (The late Fred Hirsch, 1976, coined these terms.)

The nature of the misallocation can be made clear with the help of two simple thought experiments. In each, you must choose between two worlds that are identical in every respect except one. The first choice is between world A, in which you will live in a 4,000-square-foot house and others will live in 6,000-square-foot houses; and world B, in which you will live in a 3,000-square-foot house, others in 2,000-square-foot houses. Once you choose, your position on the local housing scale will persist.

If only absolute consumption mattered, A would be clearly better. Yet most people say they would pick B, where their absolute house size is smaller but their relative house size is larger. Even those who say they would pick A seem to recognize why someone might be more satisfied with a 3,000-square-foot house in B than with a substantially larger house in A.

In the second thought experiment, your choice is between world C, in which you would have four weeks a year of vacation time and others would have six weeks; and world D, in which you would have two weeks of vacation, others one week. This time most people pick C, choosing greater absolute vacation

time at the expense of lower relative vacation time.

The modal responses in these two thought experiments suggest that housing is a positional good and vacation time a non-positional good. The point is not that absolute house size and relative vacation time are of no concern. Rather, it is that positional concerns weigh more heavily in the first domain than in the second.

When the strength of positional concerns differs across domains, the resulting conflict between individual and social welfare is structurally identical to the one inherent in a military arms race. When deciding how to apportion available resources between domestic consumption and military armaments, each country's valuations are typically more context-dependent in the armaments domain than in the domain of domestic consumption. After all, being less well armed than a rival nation could spell the end of political independence. The familiar result is a mutual escalation of expenditure on armaments that does not enhance security for either nation. Because the extra spending comes at the expense of domestic consumption, its overall effect is to reduce welfare. Note, however, that if each country's valuations were equally context-sensitive in the two domains, there would be no arms race, for in that case the attraction of having more arms than one's rival would be exactly offset by the penalties of having lower relative consumption.

For parallel reasons, the modal responses to the two thought experiments suggest an equilibrium in which people consume too much housing and too little leisure (for a formal demonstration of this result, see Frank, 1985a). In contrast, conventional welfare theorems, which assume that individual valuations depend only on absolute consumption, imply optimal allocations of housing and leisure.

In addition to leisure, goods that have been classified as non-positional by various authors include workplace safety, workplace democracy, savings and insurance. And since public goods are, by definition, available in equal quantities to all consumers, they, too, are inherently non-positional. The general claim is that unregulated market exchange will tend to emphasize the production of positional goods at the expense of these and other non-positional goods (Frank, 1985b). Among the policies suggested as remedies for this imbalance have been income and consumption taxes, overtime laws, hours laws for commercial establishments, legal holidays, workplace safety and health regulation, non-waivable workers' rights, and tax-financed savings accounts.

Consumption externalities also have implications for the theory of revealed preference, which says that, if a well-informed individual chooses a risky job that pays $600 a week rather than a safer one that pays only $500, he reveals that the safety increment is worth less than $100 to him. If safety is a non-positional good, however, this inference does not follow, for it ignores the fact that, if all workers exchange safety for increased income, the anticipated increase in relative consumption does not occur. The value that workers assign to safety may thus be revealed as much in the patterns of safety regulation they favour as in the nature of the jobs they choose.

See Also

- leisure
- time use
- Veblen, Thorstein Bunde

Bibliography

Bodkin, R. 1959. Windfall income and consumption. *American Economic Review* 49, 602–14.

Carroll, C. 1998. Why do the rich save so much? In *Does Atlas Shrug: The Economic Consequences of Taxing the Rich*, ed. J. Slemrod. New York: Oxford University Press.

Duesenberry, J. 1949. *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.

Easterlin, R. 1995. Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior and Organization* 27, 35–47.

Frank, R. 1985a. The demand for unobservable and other nonpositional goods. *American Economic Review* 75, 101–16.

Frank, R. 1985b. *Choosing the Right Pond*. New York: Oxford University Press.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton, NJ: Princeton University Press.

Friedman, M. 1963. Windfalls, the horizon, and related concepts in the permanent income hypothesis. In *Measurement in Economics*, ed. C. Christ. Stanford: Stanford University Press.

Hemenway, D. and Solnick, S. 1998. Is more always better? *Journal of Economic Behavior and Organization* 37, 373–83.

Hirsch, F. 1976. *Social Limits to Growth*. Cambridge, MA: Harvard University Press.

Luttmer, E. 2005. Neighbors as negatives: relative earnings and well-being. *Quarterly Journal of Economics* 120, 936–1002.

Marx, K. 1847. Relation of wage-labour to capital. In *Wage Labour and Capital*. Marx/Engels Internet Archive, 1999. Online. Available at http://www.marxists.org/archive/marx/works/1847/wage-labour/ch06.htm, accessed 16 August 2005.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R. Campbell and A. Skinner. Oxford: Oxford University Press, 1976.

Veblen, T. 1899. *The Theory of the Leisure Class*. New York: Modern Library.

Veenhoven, R. 1993. *Happiness in Nations: Subjective Appreciation of Life in 56 Nations*. Rotterdam:

Erasmus University.

How to cite this article

Frank, Robert H. "consumption externalities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000548> doi:10.1057/9780230226203.0301

The New Palgrave Dictionary of Economics Online

consumption sets

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The idea of consumption sets was introduced into general equilibrium theory in July 1954 in Arrow and Debreu (1954, pp. 268–9) and Debreu (1954, p. 588), the name itself appearing only in the latter paper. Later expositions were given by Debreu (1959) and Arrow and Hahn (1971) and a more general discussion by Koopmans (1957, Essay 1). Although there have been several articles concerned with non-convex consumption sets (e.g. Yamazaki, 1978), in more recent years their role in general equilibrium theory has been muted, especially in approaches that use global analysis (see for example, Mas-Colell, 1985, p. 69). Such sets play no role in partial equilibrium theories of consumer's demand, even in such modern treatments as Deaton and Muellbauer (1980). Since general equilibrium theory prides itself on precision and rigour (e.g. Debreu, 1959, p. x), it is odd that on close examination the meaning of consumption sets becomes unclear. Indeed, three quite different meanings can be distinguished within the various definitions presented in the literature. These are given below (in each case the containing set is the commodity space, usually R^n): M1 The consumption set C1 is that subset on which the individual's preferences are defined. M2 The consumption set C2 is that subset delimited by a natural bound on the individual's supply of labour services, i.e. 24 hours a day. M3 The consumption set C3 is the subset of all those bundles, the consumption of any one of which would permit the individual to survive. Each definition in the literature can (but here will not) be classified according to which of these meanings it includes. In probably the best known of them (Debreu, 1959, ch. 4), the consumption set appears to be the intersection of all three subsets C1–C3. M1 is plain. After all, preferences have to be defined on *some* proper subset of the commodity space, since the whole space includes bundles with some inadmissibly negative coordinates. M2 is also reasonable, although a full treatment of heterogeneous labour services does raise problems for what is meant by an Arrow–Debreu ‘commodity’ (see for example, that of Arrow–Hahn, 1971, pp. 75–6). It is M3 that gives real difficulty, both in itself and in relation to the others.
First, there is little reason to expect either C1 or C3 to be a subset of the other, and so still less to expect M1 and M3 to define the same set. No individual would have any problem in preferring one bundle, the consumption of which would ensure her survival, to a second bundle, the consumption of which would

result in her death by starvation. However, she might well prefer the second bundle to a third, whose consumption would cause her to die from thirst (the representation of such preferences by a real-valued utility function might pose problems, but that is another matter). On the other hand, the same individual might not be able to rank in order of preference two bundles each of which contains exotic food and drink, even though fully assured that the consumption of either bundle would allow her to survive.

More importantly, M3 implicitly introduces *consumption* activities, the actual eating and drinking and sheltering that are essential to survival. Such activities constitute what are sometimes called, by analogy with production, the consumption technology. Some partial equilibrium models, such as ‘the new home economics’ and the theory of characteristics, have treated aspects of such technologies but so far general equilibrium theory has not. In particular, Arrow–Debreu theory has not done so. As a consequence (and unlike some forms of the classical ‘corn model’) it does not give a coherent account of the birth and death of individual persons, any more than it does of the birth and death of individual firms (*see* general equilibrium). Hence the third meaning M3, which in effect presumes that the model contains such an account when it does not, is hard to interpret. One major difficulty of interpretation arises with the Slater-like condition that each individual's endowment of goods and services, valued at the competitive prices p^*, should be strictly greater than inf $\{\langle x, p^* \rangle : x \in C\}$, where <.,.> denotes inner product and C is ‘the’ consumption set (see cost minimization and utility maximization). This condition is important in proofs of existence of competitive equilibrium, to ensure for example that the budget correspondence is continuous, or that a compensated equilibrium is a competitive equilibrium. It is itself guaranteed by assumptions (discussed by McKenzie, 1981, pp. 821–5) on the relations between ‘individual’ consumption sets and the aggregate production set.

If C is taken to contain C3 then the assumptions just referred to imply that every consumer survives in every competitive equilibrium, not merely for one period but over the whole (finite) Arrow–Debreu span. This is a breathtaking assertion of fact which recalls irresistibly Hicks's wry observation: ‘Pure economics has a remarkable way of producing rabbits out of a hat – apparently *a priori* propositions which apparently refer to reality. It is fascinating to try to discover how the rabbits got in’ (1939, p. 23). On the other hand if C is taken to be C1, then the assumptions take on a purely technical (and so less objectionable) aspect, whose role is essentially to ensure that the system stays within the (relative) interior of the sets concerned and so displays appropriate continuity. But then there is no presumption that individual agents survive in a competitive equilibrium, even for one period (cf. Robinson, 1962, p. 3). The multi-period versions of the Arrow–Debreu model are then at risk, since individuals disappear and take their labour service endowments with them. This should not come as a surprise – the problems of time in economics are really too complicated to be overcome simply by adding more dimensions to the one-period model.

Some models that include C3 in C attempt to justify Slater-like conditions directly, on the grounds that ‘Not many economies in the present day are so extremely laissez faire as to permit people to starve’ (Gale and Mas-Colell, 1975, p. 12). This justification clearly fails as long as the behaviour of the public agency whose actions allegedly prevent such starvation is not modelled *explicitly*, like that of the private agents.

It is usually assumed that consumption sets are bounded below, closed and convex. The first two assumptions are innocuous but the third poses issues of a conceptual kind, which spring from difficulties

in interpreting the idea of a convex combination $x^t = tx^1 + (1-t)x^2$ of two bundles x^1 and x^2, where $t \in [0, 1]$. Consider the example, sometimes used, in which x^1 is a house in London and x^2 a house in Paris. We cannot take seriously the claim that x^t is a house in the Channel, so t cannot refer to distance. An alternative claim that t refers to the proportion of the period that is spent in London could arise from many different finite partitions of the time interval, not all of which need to be ranked equally by the individual. In effect, convexity of the consumption set comes down to the divisibility of consumer goods, an assumption which in the past has proved not such a bad approximation if one is interested mainly in general equilibrium aspects of market demand, and representative rather than actual consumers. Indivisibilities of producer goods are of course much more serious.

See Also

- Arrow–Debreu model of general equilibrium
- cost minimization and utility maximization
- general equilibrium
- indivisibilities

Bibliography

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Deaton, A. and Muellbauer, J. 1980. *Economics and Consumer Behaviour*. Cambridge: Cambridge University Press.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40(7), 588–92.

Debreu, G. 1959. *Theory of Value*. Cowles Commission Monograph No.17. New York: Wiley.

Gale, D. and Mas-Colell, A. 1975. An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2, 9–15.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Koopmans, T.C. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

McKenzie, L.W. 1981. The classical theorem on existence of competitive equilibrium. *Econometrica* 49,

819–41.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium. A Differentiable Approach*. Cambridge: Cambridge University Press.

Robinson, J.V. 1962. The basic theory of normal prices. *Quarterly Journal of Economics* 76(1), 1–20.

Yamazaki, A. 1978. An equilibrium existence theorem without convexity assumptions. *Econometrica* 46, 541–55.

How to cite this article

Newman, Peter. "consumption sets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000319> doi:10.1057/9780230226203.0302

The New Palgrave Dictionary of Economics Online

consumption taxation

James M. Poterba

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Whether to tax households based on their income or on their consumption is one of the central and long-standing questions of tax design. Most developed nations rely on a combination of income and consumption taxes to raise revenue. The debate over alternative tax bases involves both philosophical arguments about what constitutes a fair measure of ability to pay and economic arguments about the relative efficiency of different tax bases. Consumption taxes can be implemented in a variety of ways, including value added taxes, retail sales taxes, and savings-exempt income taxes.

Keywords

capital gains taxation; consumption taxation; distortionary taxation; flat rate tax; Individual Retirement Accounts (USA); progressive and regressive taxation; redistribution; retail sales tax; savings-exempt income tax; tax compliance; taxation of income; value added tax

Article

Whether household income or household consumption constitutes a better measure of a household's ability to pay taxes, and whether there are substantial efficiency gains to choosing one tax base rather than the other, are two of the central questions of public finance. The debate between advocates of income taxes and advocates of consumption taxes has spanned several centuries. While income has often been viewed as the basis for taxation, and Adam Smith discusses taxation relative to household incomes, Thomas Hobbes, John Stuart Mill and Irving Fisher were all strong proponents of taxing consumption. Consumption tax supporters argue that the amount that an individual draws from the economy's resource pool should determine his or her tax burden. They also point out that an income tax levies a ‘double tax’ on saving, since saved income is taxed both when it is earned and when the savings yield a return to capital. Kaldor (1955) offers a broad review of the case for consumption taxation. Two notable reports in the late 1970s, one by the Meade Commission (Meade, 1978) in the United Kingdom and the other by

the staff of the US Treasury Department (1977), outlined the modern cases for consumption taxation and developed specific proposals.

Proponents of income taxation argue that the change in an individual's command over resources between one period and the next is an appropriate measure of ‘ability to pay’, even if those resources are not immediately consumed. This is the measure of taxable capacity suggested by Robert Murray Haig and Henry Simons: ‘Haig–Simons’ income. Moreover, they argue that changes in resources should be taxed regardless of whether they arise from labour income or from the returns to past saving.

Income taxes and consumption taxes exhibit different time profiles over the course of a lifetime. When individuals experience a period of retirement before they die, the time profile of tax payments under a consumption tax will fall later in the lifetime than the corresponding payments under an income tax. This is because individuals continue to consume after they stop earning labour income. Retirees under an income tax pay tax only on their capital income, while retirees under a consumption tax pay tax on their total outlays, which are likely to exceed their capital income.

The debate between proponents of consumption taxation and proponents of income taxation concerns whether or not capital income should be taxed. The foregoing philosophical issues notwithstanding, the efficiency cost of taxing capital income has been an active subject of economic research. Chamley (1986) and Judd (1985) argue that the effective distortions from capital taxes cumulate over time as the difference between discounting the future at before-tax and after-tax interest rates increases with the compounding horizon. They claim that the optimal steady-state capital income tax rate should be zero. However, they also point out that a one-time capital levy is an efficient device for raising revenue. A number of recent studies, described in Auerbach (2006), have examined the robustness of the theoretical claim that the optimal capital tax rate is zero.

Consumption tax proponents, such as Bradford (1980), claim not only that taxing consumption rather than income avoids intertemporal distortions, but also that it solves many of the most difficult measurement and accounting problems associated with income taxation. Under a consumption tax, for example, there would be no distinction between the tax burden on investment projects financed with debt and those financed with equity, or between realized and unrealized capital gains. There would be no need to measure the rate at which long-lived physical assets depreciate, as one must do under an income tax. Income tax proponents respond that some components of consumption may be difficult to measure, and that it is more difficult to tailor consumption taxes than income taxes to achieve redistributive goals.

Formalizing consumption taxation vs. income taxation

The essential difference between a consumption tax and an income tax can be illustrated by comparing the lifetime budget constraints that consumers would face under each tax system. An income tax is levied on both labour and capital income. When a household has assets of A_{t-1} at the beginning of period t, these assets earn a pre-tax return r and the household earns labour income of wL where w equals the real wage and L denotes labour supply, the income tax base is $wL+rA_{t-1}$. The income tax not only reduces the after-tax real wage but also lowers the after-tax return to saving. In a life-cycle model in which a household lives for T periods and in which there is no inflation, the life-cycle budget constraint with an income tax is

$$\sum_{t=1}^{T} C_t / (1 + r(1 - \tau))^t = \sum_{t=1}^{T} (1 - \tau) w_t L_t / (1 + r(1 - \tau))^t + A_0$$
(1)

In this expression, C denotes real consumption spending, and A_0 is the household's initial wealth endowment.
In contrast, the life-cycle budget constraint with a consumption tax levied at rate θ is

$$\sum_{t=1}^{T} (1 + \theta) C_t / (1 + r)^t = \sum_{t=1}^{T} w_t L_t / (1 + r)^t + A_0$$
(2)

The discount rate in this case is the pre-tax return. The consumption tax levied on outlays in each period is equivalent to a tax on labour income *and* the household's initial endowment. If $(1 - \nu) = 1/(1 + \theta)$, then eq. (2) can be rewritten as

$$\sum_{t=1}^{T} C_t / (1 + r)^t = \sum_{t=1}^{T} (1 - \nu) w_t L_t / (1 + r)^t + (1 - \nu) A_0$$
(3)

The timing of tax payments under the ‘wage-and-endowment tax’ in (3) is different from that under the consumption outlays tax in (2), but the present value of taxes and the effects on economic incentives are the same under the two systems. The tax on initial endowment is an essential component of this equivalence: a wage tax alone is *not* equivalent to a consumption tax because initial assets escape taxation when only wages are taxed.
The current tax system in most developed nations is a hybrid structure, reflecting some elements of income taxation but also embodying components of a consumption tax. This is most apparent in nations that rely on both an income tax and a consumption tax, such as a value added tax, for a substantial share of government revenue. Even within many income tax systems, however, there are provisions that move toward an income tax-consumption tax hybrid. In the United States, for example, capital income that accrues in employer-provided pension plans and in a variety of taxpayer-directed retirement saving accounts, such as Individual Retirement Accounts (IRAs), is excluded from income taxation. Some types of capital income are taxed at rates below the top statutory tax rates on wage income. Realized

capital gains have often been taxed at preferential rates, and in some cases dividend income to households is also subject to reduced rates of tax. There is substantial variation in tax structures across nations, but the principle of allowing some tax reduction on capital income is widespread. This makes it difficult to assess where any particular nation's tax system falls on the spectrum between an income tax and a consumption tax.

Types of consumption taxes

In practice, there are many ways to implement a consumption tax. Two, the retail sales tax and the value added tax, are widely used in practice. Both are examples of indirect consumption taxes, because they are levied without any reference to the consumer's identity. Direct consumption taxes, in contrast, are levied on households by computing their total consumption. In contrast to indirect consumption taxes, direct consumption taxes can be levied at progressive rates. While direct consumption taxes have never been used as the primary revenue source in any nation, they have been actively debated in the policy reform literature. Tax structures that closely resemble direct consumption taxes have been adopted as components of existing tax systems. The two most widely discussed direct consumption tax options are the savings-exempt income tax and the ‘X-tax,’ a combination of a cash-flow tax on business income and a household wage tax.

A *retail sales tax* (RST) is the simplest consumption tax. It is collected by retailers at the point of final sale, and it corresponds directly to the tax on consumption spending described in eq. (2) above. In 2006, 44 of the 50 US states levied some form of sales tax, with rates typically between four and seven per cent. There is little experience with RSTs above ten per cent. One unresolved question with regard to proposals that call for significantly higher RSTs is whether the difficulty of monitoring all points of purchase would lead to substantial problems of tax evasion.

A *value added tax* (VAT) is a very common form of consumption tax. Virtually all developed nations with the exception of the United States levy some form of VAT, with rates ranging up to 25 per cent in Denmark, Norway and Sweden. The VAT is collected from businesses on the difference between the gross value of their sales and the cost of any inputs that they purchase from other entities that have already paid VAT.

To illustrate the operation of VAT, consider a bakery that produces and sells bread for $100. The baker's input costs are $30 for flour and $65 for an employee. The bakery earns a $5 profit. If flour is purchased from another firm that has already paid VAT, then the bakery's VAT liability equals $70 times the VAT tax rate, since its value added equals its sales of $100•minus input purchases that have already paid VAT, or $30. Wages are *not* deducted from sales when computing value added. Although the VAT is collected in stages from all firms in a production chain, it is equivalent to an RST at the same rate. One attractive feature of the VAT is that downstream firms, such as the baker in this example, help ensure VAT compliance by upstream firms that supply intermediate goods. In this example if the flour seller cannot provide documentation for its VAT payment, the baker will face tax on value added of $100. Thus the baker has an incentive, all else equal, to purchase inputs from suppliers who pay VAT.

Ebrill et al. (2001) offer a comprehensive discussion of VAT implementation issues and summarize experience with the VAT in both developed and developing nations. The VAT accounts for a substantial share of revenue in most industrialized nations. The treatment of international transactions has proven a source of difficulty in some nations, since exporting firms are typically granted a rebate for their VAT

payments. Some tax evasion schemes involve exporting goods to qualify for the rebate and re-importing the same goods without paying VAT on the import. The taxation of financial services also proves challenging under the VAT.

A *savings-exempt income tax* (SEIT) is a consumption tax that is built on an income tax model. For those who are familiar with an income tax system, it provides a way of shifting to a consumption tax without drastic administrative changes in the tax system. The Nunn-Domenici 'USA Tax', introduced in the US Senate in the mid-1990s and analysed in Ginsburg (1995), was a based on this type of consumption tax.

Under the SEIT, the tax base is income less saving. To prevent taxpayers from simply claiming high levels of saving and thereby avoiding tax liability, saving must be documented in the form of a contribution to a 'qualified account'. Income earned on assets held in the qualified account is not taxed, but withdrawals from the qualified account are included in the tax base. Thus a taxpayer who earns $50,000 and contributes $5,000 would be taxed on $45,000 in the contribution period. If, some years later, when earnings equal $25,000, the taxpayer withdraws $10,000 from the qualified account, she would be taxed on $35,000.

Even though the SEIT taxes the earnings that have accrued on the contributions to the qualified account when the funds are withdrawn from this account, the return on capital is untaxed in this setting. Taxing accumulated capital income when the proceeds are withdrawn is *not* equivalent to taxing capital income as it accrues: this is the reason Individual Retirement Accounts, 401(k) plans and other tax-deferred saving programmes provide an incentive for personal saving. When capital income is taxed as it accrues, the value of earning one dollar, paying tax on it at rate τ, and then investing it for T periods at a pretax rate of return r but with an accrual tax rate τ, is $(1-\tau)(1+(1-\tau)r)^T$. In contrast, if the initial earnings are excluded from taxation, there is no taxation of accruing capital income, and withdrawals are taxed at 100 τ per cent, then the value after T periods is $(1-\tau)(1+r)^T$. The qualified account approach eliminates the tax burden on the 'inside build up' of capital assets.

One of the key challenges in implementing a SEIT is avoiding the wholesale reallocation of existing wealth into 'qualified accounts' at the time the SEIT is adopted. Such transfers could sharply reduce tax collections, but, since they involve previously accumulated assets, they would not translate into marginal incentives for new saving. If it were possible to inventory the assets of each taxpayer when the SEIT was implemented, this would make it possible to design regulations to limit the transfer problem. Absent such information on previously accumulated wealth, however, transfers of pre-existing wealth into qualified accounts are likely to prove a difficult implementation issue for the savings-exempt income tax.

An *X-tax* combines a cash flow tax on businesses, much like a VAT with a deduction for wages, with a household-level tax on wage income. The X-tax and its relatives are descended from proposals in the US Treasury Department's (1977) report on fundamental tax reform. Bradford (1986) discusses several plans of this type, and one widely discussed variant was developed by Hall and Rabushka (1995). The X-tax has greater flexibility than a VAT for achieving distributional goals, since the household level tax can include progressive rates or transfers to low-earning households. This illustrates the distributional flexibility of direct rather than indirect consumption taxes. If the household tax is a flat rate tax on wages at the same rate as the corporate cash flow tax, then the X-tax is equivalent to a VAT or an RST. When the rates are different, then the X-tax becomes a combination of a VAT and an additional tax or subsidy on labour income. The cash flow nature of the business tax eliminates the need to measure depreciation,

since firms can claim an immediate deduction – expensing – for purchases of capital goods.
In practice, neither the RST nor the VAT is implemented strictly along the principles described above. Proposals for both the SEIT and the X-tax also include additional features that often introduce efficiency costs that would not arise in ‘textbook’ versions of these taxes. The RST, for example, typically exempts some goods and services. Expenditures on food, medical care and clothing are often excluded from the tax base, thereby achieving a more progressive distribution of tax burdens while creating distortions between various classes of consumption goods. The VAT is often implemented at different rates on different goods, with exemptions for some goods, creating the same distortionary effects. Because both the savings-exempt income tax and the X-tax require households to file tax returns, they are prone to modification to allow deductions for some expenditure categories, such as mortgage interest or health insurance premiums. While neither of these consumption tax plans has been tried in practice, they probably would be influenced by the same political pressures that have generated a wide array of tax expenditures in the current income tax code.

Efficiency gains from replacing an income tax with a consumption tax

Income taxes create two distortions: one between the before-tax and the after-tax real product wage, which distorts the labour–leisure margin, and one between the before-tax and the after-tax real rate of return to saving. The latter distorts the lifetime allocation of consumption relative to the pattern that would be chosen if the return to delaying consumption equalled the economy's pre-tax marginal product of capital. Shifting from an income tax to a consumption tax eliminates the second distortion. The key analytical issue in evaluating the welfare consequences of replacing an income tax with a consumption tax is therefore measuring the efficiency costs associated with the taxation of saving and investment. This efficiency cost depends on the underlying structure of consumer preferences. The interest elasticity of saving is often invoked as a summary measure of the key preference parameters. When changes in after-tax returns induce only modest changes in household saving, the efficiency gain from switching from an income tax to a consumption tax will be smaller than when the interest elasticity of saving is large.
Auerbach and Kotlikoff (1987) use a dynamic general equilibrium model, including a realistic treatment of household life-cycle income and consumption streams, to evaluate the efficiency gains from replacing an income tax with a consumption tax. Their results suggest that for a given revenue requirement, the steady-state capital stock is larger with a consumption tax than with an income tax. This translates into higher steady-state per capita utility under the consumption tax than the income tax.
The steady-state comparison is not the only consideration when evaluating two alternative tax systems, however. It is possible to design tax reforms that raise steady-state welfare but cause welfare losses in the transition from an initial equilibrium to the new steady state. The trade-off between short-run and long-run policy effects depends on the policymaker's discount rate and in calibrated general equilibrium models it is possible to compute the present discounted value of the gains and losses to the cohorts alive at different dates.

Transition from one tax regime to another

Focusing on the present value of welfare gains and losses draws attention to the transitional rules that govern the switch from one tax system, say an income tax, to another, such as a consumption tax. These transition rules can determine whether a policy reform represents a net gain or a net loss relative to continuation of the initial income tax regime. Altig et al. (2001) illustrate this important point using a more elaborate version of the model developed in Auerbach and Kotlikoff (1987). They find that if the tax basis of existing assets is extinguished when the income tax is replaced by a consumption tax, so that depreciation allowances are no longer claimed after the reform, and if investors who accumulated savings under the income tax regime do not receive any relief from the consumption tax burden they will face when they draw down their assets, then the efficiency gains from adopting a consumption tax may be as large as five per cent of national income.

‘Grandfathering’ existing assets sharply reduces these efficiency gains, because it reduces the base of the consumption tax and requires higher tax rates to satisfy a given revenue constraint. This results in greater distortions on the labour–leisure margin. Designing transition relief that participants in the political process will view as fair, without forgoing most of the efficiency gains from a stark consumption tax transition, is likely to be one of the greatest challenges in any consumption-oriented tax reform.

See Also

- tax expenditures
- taxation of income
- value-added tax

Bibliography

Altig, D., Auerbach, A.J., Kotlikoff, L.J., Smetters, K.A. and Walliser, J. 2001. Simulating fundamental tax reform in the United States. *American Economic Review* 91, 574–95.

Auerbach, A.J. 2006. The choice between income and consumption taxes: a primer. Working Paper No. 12307. Cambridge, MA: NBER.

Auerbach, A.J. and Kotlikoff, L.J. 1987. *Dynamic Fiscal Policy*. Cambridge: Cambridge University Press.

Bradford, D. 1980. The case for a personal consumption tax. In *What Should be Taxed? Income or Expenditure*, ed. J. Pechman. Washington, DC: Brookings Institution.

Bradford, D. 1986. *Untangling the Income Tax*. Cambridge, MA: Harvard University Press.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54, 607–22.

Ebrill, L., Keen, M., Bodin, J.-P. and Summers, V. 2001. *The Modern VAT*. Washington, DC: International Monetary Fund.

Ginsburg, M. 1995. Some thoughts on working, saving, and consuming in Nunn-Domenici's tax world. *National Tax Journal* 48, 585–602.

Hall, R. and Rabushka, A. 1995. *The Flat Tax*, 2nd edn. Stanford, CA: Hoover Institution Press.

Judd, K.L. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 29, 59–83.

Kaldor, N. 1955. *An Expenditure Tax*. London: George Allen & Unwin.

Meade, J.E. 1978. *The Structure and Reform of Direct Taxation*. London: Allen & Unwin.

US Treasury Department. 1977. *Blueprints for Basic Tax Reform*. Washington, DC: US Treasury Department.

How to cite this article

Poterba, James M. "consumption taxation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000320> doi:10.1057/9780230226203.0303

The New Palgrave Dictionary of Economics Online

consumption-based asset pricing models (empirical performance)

Fatih Guvenen and Hanno Lustig

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Asset pricing is a branch of financial economics that is rich in puzzles and anomalies – that is, stylized empirical facts not easily explained by the canonical asset pricing models. These range from the equity premium puzzle and the risk-free rate puzzle to the fact that stock returns are highly predictable. This article discusses different consumption-based asset pricing models that have been developed to resolve these puzzles, and it evaluates their empirical performance.

Keywords

capital asset pricing model; consumption-based asset pricing models; elasticity of intertemporal substitution; equity premium puzzle; external habit; habit formation; heteroskedasticity; imperfect risk sharing; incomplete markets; precautionary savings; real business cycles; recursive preferences; representative agent; risk aversion; risk-free rate puzzle; stochastic discount factor

Article

The aim of consumption-based asset pricing models is to explain a number of important and puzzling features of asset returns using standard economic theory. Perhaps the best-known challenge for these models is the *equity premium puzzle*. Let us start from the Euler equations for stock and bond choice, and let us assume that both of these Euler equations hold with equality. If agents have constant relative risk aversion (CRRA) preferences and if returns and consumption growth are jointly log-normal, then the Sharpe ratio (that is, the equity premium per unit of risk) can be decomposed as:

$$\frac{E(R^e)}{std(R^e)} \approx \alpha \times std(\Delta c) \times corr(\Delta c, R^e), \tag{1}$$

where R^e is the excess return on stocks over bonds, α is the relative risk aversion (RRA) parameter, and Δc denotes log consumption growth. The equity premium is about six per cent per year in the US data with a standard deviation of 15 per cent, producing a Sharpe ratio ($E(R^e) / std(R^e)$) of 0.4. Mehra and Prescott (1985) used the construct of a representative agent who consumes the aggregate endowment stream. Constantinides (1982), Rubinstein (1974) and Wilson (1968) derived aggregation results that rely on either complete markets or the absence of idiosyncratic income risk. By appealing to these aggregation results, Mehra and Prescott could substitute *per-capita* consumption growth into (1). This series has a standard deviation of less than two per cent in the post-war US data, and a low correlation with stock returns – less than 0.25 by most estimates. Substituting these values into the expression above implies a lower bound for the relative risk aversion coefficient of 80, which is implausibly high judging by its implications for an individual's choices in other settings. In other words, we need extremely high risk aversion to rationalize the observed equity premium, and that is the puzzle. Furthermore, even if one is willing to accept such a high coefficient of risk aversion, this choice creates different puzzles itself – a point first noted by Weil (1989).

To understand Weil's ‘risk-free rate puzzle’, first note that the Euler equation for the risk-free asset choice can be linearized to obtain:

$$E\left[R^f\right] \approx -\ln\beta + \alpha E(\Delta c) - \frac{\alpha^2}{2} var(\Delta c). \tag{2}$$

Let us assume a positive time discount rate (β <1), and an average consumption growth rate of 1.5 per cent per year. Let us also abstract from uncertainty for the moment. Then a risk aversion of 40 would imply an implausibly high interest rate of nearly 60 per cent per year simply because these households are extremely unwilling to substitute consumption over time. As a result, they desire a flat consumption profile and, therefore, would like to transfer resources from the future to today. But since this is not feasible in an endowment economy, the equilibrium risk-free rate needs to be very high to discourage this type of consumption smoothing and make individuals willing to consume their endowment every period.

The last term in (2) captures the precautionary savings motive, which becomes active in the presence of uncertainty. For very high levels of risk aversion, this effect dominates the intertemporal substitution effect, and an increase in the RRA coefficient *reduces* the risk-free rate. Epstein and Zin (1989) developed a class of recursive preferences that disentangles the inverse of the elasticity of intertemporal substitution from the coefficient of risk aversion. As discussed below, these preferences allow one to make progress on the equity premium puzzle without running into the risk-free rate puzzle.

Against the backdrop of Mehra and Prescott's benchmark model, subsequent papers that attempt to resolve these puzzles can be categorized according to whether they modify (i) the preferences, (ii) the

endowment process, or (iii) the market and asset structure. We discuss each of these approaches in turn.

The utility function

Recursive preferences

In the case of CRRA utility, the stochastic discount factor (SDF) has the following form: $M_{t,t+1} = \beta(C_{t+1}/C_t)^{-\alpha}$, where *C* denotes the level of consumption. A drawback of this specification is that it restricts the elasticity of intertemporal substitution (EIS) to be the reciprocal of the RRA parameter when in fact these two parameters capture conceptually distinct aspects of individuals' preferences. Building on work by Kreps and Porteus (1978), Epstein and Zin (1989) and Weil (1989) introduced 'recursive preferences' (also called 'non-expected utility'):

$$U_t = \left[(1-\beta)C_t^{\rho} + \beta E_t(U_{t+1}^{1-\alpha})^{\rho/1-\alpha}\right]^{1/\rho}, \tag{3}$$

where α is still the RRA parameter, but now the EIS is captured by a separate parameter: $1/(1-\rho)$. In this case, the SDF is given by:

$$M_{t,t+1} = \left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{\rho-1}\right]^{\gamma}\left(\frac{1}{R_t^M}\right)^{1-\gamma},$$

where $\gamma = \alpha/\rho$, and R_t^M is the total return on the investors' wealth portfolio (including human capital which must be tradable for this representation to be derived; see Epstein and Zin, 1989, and Weil, 1989). An appealing feature of this SDF is that it combines two components that are each central to separate asset pricing theories: in particular, the SDF is a geometric average of consumption growth and the market return, where the latter is the relevant SDF in the standard capital asset pricing model (CAPM). Moreover, when $\alpha = 0$ (logarithmic risk preferences), then the CAPM emerges as a special case whereas $\alpha = \rho$ reduces it to the standard case of expected utility (see Epstein and Zin, 1989; Campbell, 2000). In addition, this preference specification is flexible enough to allow a choice of a coefficient of relative risk aversion that is high enough to match the equity premium without being forced to accept a very low EIS. The low EIS is responsible for the risk-free rate puzzle, as explained above. Bansal and Yaron (2004) exploit this agent's concern for long-run consumption risk by introducing a small predictable

component in consumption growth.

Habit formation and catching-up with the Joneses

Another approach, pioneered by Sundaresan (1989), Abel (1990) and Constantinides (1990), starts from the following specification of the investor's preferences over consumption streams C_t:

$$U_t = \frac{(C_t - X_t)^{1-\alpha}}{1-\alpha}$$

where X_t is some function of either (i) the individual's own past consumption or (ii) the past consumption of a reference group, such as an individual's peers, neighbours, or the population as a whole. Abel's specification features the ratio of C_t to X_t instead of the level difference. The first approach allows an individual's marginal utility to depend on her own past consumption history. This is commonly referred to as habit formation, endogenous habit, or internal habit. The second interpretation allows an individual's utility to depend on her status *relative* to her peers, neighbours or the population as a whole. This is referred to as catching-up with the Joneses or as external habit. These preference specifications amplify the effect of consumption growth shocks on the marginal utility growth of investors, in turn generating a high equity premium.
A particularly successful version of the catching-up-with-the-Joneses specification was developed by Campbell and Cochrane (1999) (henceforth CC) who choose the sensitivity of X to consumption growth shocks to match the conditional and unconditional moments of returns. In the baseline CC model, aggregate consumption and dividend growth are i.i.d. over time. Menzly, Santos and Veronesi (2004) introduce additional cash flow dynamics to explain the time series and cross-section of stock returns, while Santos and Veronesi (2005) emphasize the importance of labour income share variation to understand time variation in risk premia. Wachter (2002) applies a version of the CC model to the term structure, while Verdelhan (2004) uses the same model to explain the forward premium puzzle.

Looks like habit

Several recent papers have proposed models with standard preferences (such as CRRA) but consider economic environments that give rise to SDFs similar to those resulting from external habit preferences (such as the one used in CC). Examples include work by Piazzesi, Schneider and Tuzel (2007) who introduce housing services consumption into this framework, and by Yogo (2006) who considers durable consumption broadly defined, building on earlier work by Dunn and Singleton (1986) and Eichenbaum and Hansen (1990). Finally, Guvenen (2005) studies a model with limited stock market participation and shows that while the asset pricing implications of his model are similar to those in CC, the implications for macroeconomic questions (such as policy analysis, and so on) are quite different.

Additional arguments in the utility function

The models discussed so far assume that investors only derive utility from non-durable consumption. In exchange economy models (in which the consumption process is exogenous) this is equivalent to assuming that non-durable consumption enters the utility function in a separable manner. Some recent papers explicitly model the utility flow from housing consumption (in a non-separable manner), and find that such an extension improves the asset pricing performance (see Grossman and Laroque, 1990; Piazzesi, Schneider and Tuzel, 2007; Flavin and Yamashita, 2002). Similarly, a labour–leisure choice was introduced by Boldrin, Christiano and Fisher (2001) and Danthine and Donaldson (2002), in a representative agent framework, and by Uhlig (2006) in an incomplete markets framework. However, these authors find that this extension negatively affects the performance of asset pricing models, because it allows households to smooth their marginal utility by adjusting on the labour–leisure margin. As a result, one needs to introduce additional – typically labour market – frictions to counteract this new smoothing opportunity.

Consumption dynamics

In consumption-based asset pricing models, it is common to assume that aggregate consumption growth is i.i.d. over time, because the evidence for consumption growth predictability in the data is weak. In the i.i.d. case, the conditional market price of risk, which can be approximated by the conditional standard deviation of the log SDF, $\sigma_t(\log M_{t,t+1}) = \alpha \times \sigma_t(\Delta c)$, is constant. Therefore, these models cannot generate any time variation in risk premia on equity or any other asset.
In the context of a standard representative agent model, Kandel and Stambaugh (1990) generate time-variation in risk premia by introducing heteroskedasticity in aggregate consumption growth. Bansal and Yaron (2004) deviate from the i.i.d. assumption by introducing a small predictable component in consumption growth that is statistically hard to detect. This long-run component increases the market price of consumption risk. In addition, they add some time variation in the size of the long-run risk component. Colacito and Croce (2005) show these long-run risk models can reconcile the low volatility of exchange rate changes with the large market price of risk. Finally, Longstaff and Piazzesi (2002) argue that corporate earnings are much more risky than aggregate consumption growth, and that this can account for a large share of the equity premium puzzle.

Production economy models

These asset pricing puzzles have also attracted a lot attention from macroeconomists because the same basic framework used in Mehra and Prescott (1985) also forms the backbone of the Kydland and Prescott (1982) model and the subsequent real business cycle literature. Therefore, understanding why individuals dislike risk in financial markets could help shed light on individuals' perceptions of macro risk and consumption fluctuations, which are key issues for macroeconomic policy. However, macroeconomists are also interested in the determination of quantities, such as output, investment and

consumption, making the exchange economy framework unsuitable for their purposes. Therefore, macroeconomists replace the exogenous endowment stream with the endogenous equilibrium consumption process generated by a standard neoclassical production economy that faces technology shocks. One of the first findings of this approach, summarized in Rouwenhorst (1995), is that resolving the equity premium puzzle in a production economy is far more challenging than in an exchange economy, because this endogenous consumption process becomes too smooth if one increases risk aversion. As a result, one needs to resort to real frictions such as large adjustment costs in Jermann's (1998) model. Furthermore, and as noted above, allowing for an endogenous labour supply choice, as is common in macroeconomic analysis, gives consumers another margin to smooth marginal utility and further reduces the equity premium. Boldrin, Christiano and Fisher (2001) and Uhlig (2006) have successfully introduced labour market frictions to effectively shut down this channel.

Market and asset structure

The aggregation results we appeal to in order to use a representative agent in asset pricing depend on market completeness. A natural question is to ask what happens if some of these markets are shut down.

Incomplete markets

In an attempt to resolve the equity premium puzzle, *uninsurable* idiosyncratic income risk has been introduced into consumption-based asset pricing models by Aiyagari and Gertler (1991), Telmer (1993), Lucas (1994), Heaton and Lucas (1996), Krusell and Smith (1997) and Marcet and Singleton (1999), among others. Their main results, obtained numerically for a range of parameter values, suggest that the impact of uninsurable labour income risk on the equity premium is small, because agents manage to smooth consumption quite well by trading a risk-free bond. In fact, Levine and Zame (2002) show that under general conditions the equilibrium allocations and prices in incomplete market economies converge to the complete market counterparts as households become more patient, rendering the incompleteness moot.

So when does imperfect risk sharing matter? Mankiw (1986) derives a sufficient condition for imperfect risk sharing to increase the equity risk premium: the cross-sectional variance of consumption growth needs to increase when returns are low (that is, in recessions). Constantinides and Duffie (1996) embed this counter-cyclical cross-sectional variance mechanism in a general equilibrium model. Grossman and Shiller (1982) show that the Mankiw-Constantinides-Duffie (MCD) mechanism breaks down in continuous-time diffusion models, because the cross-sectional variance of consumption growth is deterministic.

Discussion of other models

Rietz (1988) was the first to argue that countries like the United States may simply have been very lucky. Hence, the observed history of the US economy may understate the actual probability of economic disasters, such as the Great Depression (at least as perceived by investors). In this case, the

volatility of the SDF may be significantly higher than the one estimated from historical time series. As a result, investors will shun stocks and demand a much higher equity premium to hold them. One difficulty with this explanation is that many economic disasters also result in governments reneging on their debt obligations. Barro (2006) extends Rietz's framework by distinguishing between two types of disasters – those that only affect the stock market and those that affect all asset markets – and explores the empirical implications of this mechanism in recent work.

See Also

- capital asset pricing model
- consumption-based asset pricing models (theory)
- elasticity of intertemporal substitution
- incomplete markets
- recursive preferences

Bibliography

Abel, A.B. 1990. Asset prices under habit formation and catching up with the Jones. *American Economic Review* 80, 38–42.

Aiyagari, S.R. and Gertler, M. 1991. Asset returns with transaction costs and uninsured individual risk. *Journal of Monetary Economics* 27, 311–31.

Bansal, R. and Yaron, A. 2004. Risks for the long run: a potential resolution of asset prizing puzzles. *Journal of Finance* 59, 1481–509.

Barro, R. 2006. Rare disasters and asset markets in the twentieth century. *Quarterly Journal. of Economics* 121, 823–66.

Boldrin, M., Christiano, L. and Fisher, J. 2001. Habit persistence, asset returns, and the business cycle. *American Economic Review* 91, 149–66.

Campbell, J.Y. 2000. Asset pricing at the millennium. *Journal of Finance* 55, 1515–67.

Campbell, J.Y. and Cochrane, J.H. 1999. By force of habit: a consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107, 205–51.

Colacito, R. and Croce, M. 2005. Risks for the long-run and the real exchange rate. Working paper, New York University.

Constantinides, G.M. 1982. Intertemporal asset pricing with heterogeneous consumers and without

demand aggregation. *Journal of Business* 55, 253–67.

Constantinides, G.M. 1990. Habit-formation: a resolution of the equity premium puzzle. *Journal of Political Economy* 98, 519–43.

Constantinides, G.M. and Duffie, D. 1996. Asset pricing with heterogeneous consumers. *Journal of Political Economy* 104, 219–40.

Danthine, J.-P. and Donaldson, J.B. 2002. Labour relations and asset returns. *Review of Economic Studies* 69, 41–64.

Dunn, K. and Singleton, K. 1986. Modeling the term structure of interest rates under nonseparable utility and durability of goods. *Journal of Financial Economics* 17, 769–99.

Eichenbaum, M. and Hansen, L.P. 1990. Estimating models with intertemporal substitution using aggregate time series data. *Journal of Business and Economic Statistics* 8, 53–69.

Epstein, L.G. and Zin, S. 1989. Substitution, risk aversion and the temporal behavior of consumption and asset returns: a theoretical framework. *Econometrica* 57, 937–69.

Flavin, M. and Yamashita, T. 2002. Owner-occupied housing and the composition of the house-hold portfolio. *American Economic Review* 79, 345–62.

Grossman, S. and Laroque, G. 1990. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. *Econometrica* 58, 25–51.

Grossman, S. and Shiller, R. 1982. Consumption correlatedness and risk measurement in economies with non-traded assets and heterogeneous information. *Journal of Financial Economics* 10, 195–210.

Guvenen, F. 2005. A parsimonious macroeconomic model for asset pricing: habit formation or cross-sectional heterogeneity? Working paper, University of Texas at Austin.

Hansen, L.P. and Singleton, K. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249–65.

Heaton, J. and Lucas, D. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104, 668–712.

Jermann, U. 1998. Asset pricing in production economies. *Journal of Monetary Economics* 257–75.

Kandel, S. and Stambaugh, R.F. 1990. Expectations and volatility of consumption and asset returns.

Review of Financial Studies 3, 207–32.

Kreps, D. and Porteus, E.L. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46, 185–200.

Krusell, P. and Smith, J.A. 1997. Income and wealth heterogeneity, portfolio selection, and equilibrium asset returns. *Macroeconomic Dynamics* 1, 387–422.

Kydland, Finn, E. and Prescott, Edward, C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1350–70.

Levine, D. and Zame, W. 2002. Does market incompleteness matter? *Econometrica* 70, 1805–39.

Longstaff, F. and Piazzesi, M. 2002. Corporate earnings and the equity premium. Working paper, UCLA Anderson School.

Lucas, D. 1994. Asset pricing with unidiversifiable income risk and short sales constraints: deepening the equity premium puzzle. *Journal of Monetary Economics* 34, 325–41.

Mankiw, G.N. 1986. The equity premium and the concentration of aggregate shocks. *Journal of Financial Economics* 17, 211–9.

Marcet, A. and Singleton, K. 1999. Equilibrium asset prices and savings of heterogeneous agents in the presence of incomplete markets and portfolio constraints. *Macroeconomic Dynamics* 3, 243–77.

Mehra, R. and Prescott, E. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Menzly, L., Santos, T. and Veronesi, P. 2004. Understanding predictability. *Journal of Political Economy* 112, 1–47.

Piazzesi, M., Schneider, M. and Tuzel, S. 2007. Housing, consumption, and asset pricing. *Journal of Financial Economics* 83, 531–69.

Rietz, T.A. 1988. The equity risk premium: a solution? *Journal of Monetary Economics* 22, 117–31.

Rouwenhorst, G. 1995. Asset pricing implications of equilibrium business cycle models. In *Frontiers of Business Cycle Research*, ed. T.F. Cooley. Princeton, NJ: Princeton University Press.

Rubinstein, M. 1974. An aggregation theorem for security markets. *Journal of Financial Economics* 1, 225–44.

Santos, J. and Veronesi, P. 2005. Labor income and predictable stock returns. *Review of Financial Studies* 19, 1–43.

Sundaresan, S. 1989. Intertemporally dependent preferences and the volatility of consumption and wealth. *Review of Financial Studies* 2, 73–89.

Telmer, C. 1993. Asset-pricing puzzles and incomplete markets. *Journal of Finance* 48, 1803–32.

Uhlig, H. 2006. Macroeconomics and asset prices: some mutual implications. Working paper, Humboldt University.

Verdelhan, A. 2004. Habit-based explanation of the exchange rate risk premium. Working paper, Boston University.

Wachter, J. 2002. Habit formation and returns on bonds and stocks. Unpublished paper, Stern School of Business, New York University.

Weil, P. 1989. The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics* 24, 401–24.

Wilson, R. 1968. The theory of syndicates. *Econometrica* 36, 119–32.

Yogo, M. 2006. A consumption-based explanation of the cross-section of expected stock returns. *Journal of Finance* 61, 539–80.

How to cite this article

Guvenen, Fatih and Hanno Lustig. "consumption-based asset pricing models (empirical performance)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000290> doi:10.1057/9780230226203.0299

The New Palgrave Dictionary of Economics Online

consumption-based asset pricing models (theory)

Fatih Guvenen and Hanno Lustig

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The essential element in modern asset pricing theory is a positive random variable called ‘the stochastic discount factor’ (SDF). This object allows one to price any payoff stream. Its existence is implied by the absence of arbitrage opportunities. Consumption-based asset pricing models link the SDF to the marginal utility growth of investors – and in turn to observable economic variables – and in doing so they provide empirical content to asset pricing theory. This article discusses this class of models.

Keywords

consumption-based asset pricing models; equity premium puzzle; Euler equations; Sharpe ratio; stochastic discount factor

Article

Consumption-based asset pricing models study the pricing of payoff streams using the covariance of these payoffs with the marginal utility growth of investors.
The central component of a consumption-based asset pricing model is the Euler equation, which imposes restrictions on the covariance between asset returns and the marginal utility growth of investors. An easy and intuitive way to derive this equation is by using a variational argument. Suppose that the optimal consumption path of investor *i* is given by $\left\{C_t^i\right\}_{t=0}^{T}$ where *T* is possibly infinite. Suppose further that an asset *j* is available with a return $R_{t,t+1}^j$ between periods *t* and *t*+1, and the investor is not facing a binding portfolio constraint with respect to this asset. Then a feasible strategy is to reduce consumption at time *t* by a small amount ε, invest it in asset *j*, and consume the proceeds, $C_{t+1}^i + \varepsilon R_{t+1}^j$, in the next period. Assuming a time-separable utility function, with the one-period felicity function denoted by *U* and a time discount factor of β, this strategy changes the investors' expected lifetime utility by

$-U_c(C_t^i, X_t)\varepsilon + E_t[\beta U_c(C_{t+1}^i, X_{t+1})\varepsilon R_{t+1}^j]$, where E_t is the mathematical conditional expectation operator; X represents the arguments of the utility function other than consumption; and U_c denotes the partial derivative with respect to consumption. The optimality of the original sequence implies that this strategy cannot be profitable for any amount ε and any asset available. Setting this gain to zero and rearranging yields the Euler equation:

$$E_t\left[M_{t,t+1}R_{t,t+1}^j\right] = 1 \quad \text{where} \quad M_{t,t+1} = \beta\frac{U_c\left(C_{t+1}^i, X_{t+1}\right)}{U_c\left(C_t^i, X_t\right)}. \tag{1}$$

This Euler equation was first derived by Rubinstein (1976) and Lucas (1978) in discrete time, and by Breeden (1979) in continuous time. While this class of models can in principle be used to study a broad variety of assets, this article will focus on stocks and short-term bonds, which have received the greatest attention in the consumption-based asset pricing literature.

In the case of a one-period discount bond with gross return $R_{t,t+1}^f = 1/P_t^f$ – a bond that costs P_t^f dollars today and pays off 1 dollar tomorrow – the Euler equation can be rewritten as

$$P_t^f = E_t\left[M_{t,t+1}\right]. \tag{2}$$

Similarly, when the asset is a stock with ex-dividend price P_t^S and dividend payment D_t, the Euler equation can be rearranged to read $P_t^S = E_t[M_{t,t+1}(P_{t+1}^S + D_{t+1})]$. By forward substitution this equation yields:

$$P_t^S = E_t\left[\sum_{s=1}^{\infty} M_{t,t+s}D_{t+s}\right], \tag{3}$$

which determines the price of a share of equity as the value of all future dividends it entitles discounted

by the SDF.

Lucas (1978) and Mehra and Prescott (1985) used a representative-agent endowment economy structure in which the dividend stream, $\{D_t\}_{t=1}^{\infty}$, is exogenously produced by a 'tree'. Furthermore, these dividends are assumed to be perishable ('fruit'), so in equilibrium the price of equity (in the tree) adjusts to the point where the representative agent is *willing* to consume all available dividends: $C_t=D_t$. Substituting this condition into the expression for M in eq. (1), and then using M in eqs (2) and (3) shows that the price of this stock and that of the one-period bond are entirely determined by the stochastic process for D_t together with the functional form for U (we ignore X_t for now).

Hansen and Singleton (1983) tested the representative agent's Euler equation on US consumption data, and found that the model was rejected. In a famous paper, Mehra and Prescott (1985) showed that when one chooses the properties of C_t to match the moments of aggregate consumption in the data ('calibrate the model to data'), the equity premium $E(R_{t+1}^s - R_t^f)$ generated by the model was about 60 *times* smaller than that observed in the historical US data. This 'equity premium puzzle' has generated enormous interest and led to the development of a wide range of consumption-based asset pricing models in an attempt to resolve it. For further discussion of the empirical performance of these models, see consumption-based asset pricing models (empirical performance).

An alternative way to explain the hurdles these models face is by deriving an empirical lower bound on the volatility of the stochastic discount factor (SDF). Subtracting the Euler equation for bond returns from the one for stock returns yields: $E[M_{t,t+1}(R_{t+1}^s - R_t^f)] = 0$. Noting that the left-hand side of this condition can be rewritten as $Cov(M_{t,t+1}(R_{t+1}^s - R_t^f)) + E(M_{t,t+1})E(R_{t+1}^s - R_t^f)$, some simple manipulations yield the following key decomposition:

$$\frac{E\left(R_{t+1}^s - R_t^f\right)}{\sigma\left(R_{t+1}^s - R_t^f\right)} = -\frac{\sigma\left(M_{t,t+1}\right)}{E\left(M_{t,t+1}\right)} corr\left(M_{t,t+1}, R_{t+1}^s - R_t^f\right), \tag{4}$$

where $\sigma(\cdot)$ denotes the standard deviation. Observing that the correlation term is bounded from above in absolute value by 1, we get

$$\frac{E\left(R_{t+1}^s - R_t^f\right)}{\sigma\left(R_{t+1}^s - R_t^f\right)} \leq \frac{\sigma\left(M_{t,t+1}\right)}{E\left(M_{t,t+1}\right)}.$$

(5)

The left-hand side of this inequality is the ‘Sharpe ratio’ – the (expected) excess return demanded by investors per unit (standard deviation) of risk they bear – which averages about 0.40 in annual US data. The right-hand side is called the ‘market price of risk’ or the ‘maximum Sharpe ratio’. This inequality bound implies that a consumption-based model must be able to generate an SDF with a coefficient of variation (standard deviation normalized by mean) of at least 40 per cent to be consistent with the Sharpe ratio observed in the data. This observation—developed by Shiller (1982) and further generalized by Hansen and Jagannathan (1991)—provides a ‘volatility bounds’ test for potential candidate models. As discussed in consumption-based asset pricing models (empirical performance), the majority of plausibly calibrated asset pricing models fail this test.

When the investor faces a binding borrowing constraint, she cannot increase her consumption today by reducing the holdings of asset j. As a result, her marginal utility today will remain higher than the value implied by the equality condition in (1), and the Euler condition for that asset will instead be an inequality: $E_t[M_{t,t+1}R^j_{t,t+1}] < 1$. This relaxes the lower bound on the volatility of the SDF derived in equation (5) (cf. Luttmer, 1996).

To develop further implications of consumption-based models it is necessary to impose additional structure on $M_{t,t+1}$, which requires being more specific about (i) the functional form and the arguments of the utility function; (ii) the stochastic properties of variables affecting marginal utility (that is, consumption, leisure, and so on); and (iii) the market structure. The latter determines whether an appropriate aggregation theorem holds (which happens for example when markets are complete), in which case C^i can be replaced with aggregate consumption. Therefore, consumption-based models can be broadly categorized based on the assumptions they make along these three dimensions. These different models are discussed in consumption-based asset pricing models (empirical performance).

Another feature of asset markets that has received much attention in the literature concerns the high volatility of stock prices. For example, the standard deviation of the log price/dividend (P/D) ratio of stocks is about 40 per cent per annum in the US data. In a world with a constant SDF (as would be the case with risk-neutral investors), it is impossible to rationalize this high volatility with the relatively low variability of the underlying dividend stream (LeRoy and Porter, 1981; Shiller, 1981). Let p_t denote the log price, d_t denote the log dividend, and r_t denote the log stock return. Using a first-order approximation, Campbell and Shiller (1988) show that the log P/D ratio can be decomposed as follows:

$$p_t - d_t = constant + E_t \sum_{j=1}^{\infty} \rho^{j-1} [\Delta d_{t+j} - r_{t+j}]$$

with $\rho = \exp(\overline{pd}) / (1 + \exp(\overline{pd}))$ and $\overline{pd}$ denotes the average log P/D ratio. The first term in the square brackets is referred to as the cash flow component, and the second part is referred to as the discount rate component. This decomposition implies that the variance of the log P/D ratio can be stated as:

$$var(p_t - d_t) = cov\left(p_t - d_t, \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}\right) - cov\left(p_t - d_t, \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}\right).$$

This expression shows that the P/D ratio moves only because it predicts future returns on stocks or because it predicts future dividend growth. In the data, most of the volatility in P/D ratio is due to news about future expected returns (‘discount rates’), not due to future dividend growth (‘cash flows’) (Campbell, 1991; Cochrane, 1991). There is a large literature that documents the predictability of stock returns over longer holding periods, starting with work by Campbell and Shiller (1988; 1998), Poterba and Summers (1986) and Fama and French (1988; 1989). Other variables that predict returns include the spread between long and short bonds (Fama and French, 1989) and the T-bill rate (Lamont, 1998). More recently, more attention has been paid to macroeconomic variables that predict returns, most notably in the work by Lettau and Ludvigson (2001a) who document that the consumption/wealth ratio is a powerful predictor of stock returns.

So, the volatility of P/D ratio implies that excess returns on stocks are highly predictable. In other words, expected excess returns change a lot over time, even per unit of risk. We use the conditional version of the expression in (4) to understand the implications of this finding:

$$\frac{E_t\left(R^s_{t+1} - R^f_t\right)}{\sigma_t\left(R^s_{t+1} - R^f_t\right)} = -\frac{\sigma_t(M_{t,t+1})}{E_t(M_{t,t+1})} corr_t\left(M_{t,t+1}, R^s_{t+1} - R^f_t\right), \tag{6}$$

where σ_t denotes the conditional standard deviation. Good models need to produce a lot of time variation in the right-hand side of (6) and this happens mostly through variation in the conditional market price of risk (first term). This is an upper bound on the conditional Sharpe ratio. (See also Lettau and Ludvigson, 2001b, on how to measure variation in the conditional Sharpe ratio.) Another test of consumption-based asset pricing models is whether they are able to generate as much predictability as found in the data. Examples of early models that match the variation in the conditional market price of risk include Kandel and Stambaugh (1990), Campbell and Cochrane (2000) and Barberis, Huang, and

Santos (2001). More recent work includes the work by Santos and Veronesi (2005), Menzly, Santos and Veronesi (2004), Piazzesi, Schneider and Tuzel (2007), Guvenen (2005), Lustig and Van Nieuwerburgh (2005; 2006) and Bansal and Yaron (2004). These models are discussed in detail in consumption-based asset pricing models (empirical performance).

See Also

- consumption-based asset pricing models (empirical performance)
- Euler equations

Bibliography

Bansal, R. and Yaron, A. 2004. Risks for the long-run: a potential resolution of asset pricing puzzles. *Journal of Finance* 59, 1481–509.

Barberis, N., Huang, M. and Santos, T. 2001. Prospect theory and asset prices. *Quarterly Journal of Economics* 116, 1–53.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7, 265–96.

Campbell, J.Y. 1991. A variance decomposition for stock returns. *Economic Journal* 101, 157–79.

Campbell, J.Y. and Cochrane, J.H. 2000. Explaining the poor performance of consumption-based asset pricing models. *Journal of Finance* 55, 2863–78.

Campbell, J.Y. and Shiller, R.J. 1988. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195–227.

Campbell, J.Y. and Shiller, R.J. 1998. Stock prices, earnings and expected dividends. *Journal of Finance* 43, 661–76.

Cochrane, J.H. 1991. Explaining the variance of price-dividend ratios. *Review of Financial Studies* 5, 243–80.

Fama, E.F. and French, K.R. 1988. Dividend yields and expected stock returns. *Journal of Financial Economics* 22, 3–27.

Fama, E.F. and French, K.R. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25, 23–49.

Guvenen, F. 2005. A parsimonious macroeconomic model for asset pricing: habit formation or cross-sectional heterogeneity? Working paper, University of Texas at Austin.

Hansen, L.P. and Jagannathan, R. 1991. Implications of security markets data for models of dynamic economies. *Journal of Political Economy* 99, 252–62.

Hansen, L.P. and Singleton, K. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249–65.

Kandel, S. and Stambaugh, R.F. 1990. Expectations and volatility of consumption and asset returns. *Review of Financial Studies* 3, 207–32.

Lamont, O. 1998. Earnings and expected returns. *Journal of Finance* 53, 1563–87.

LeRoy, S.F. and Porter, R.D. 1981. The present-value relation: tests based on implied variance bounds. *Econometrica* 49, 555–74.

Lettau, M. and Ludvigson, S.C. 2001a. Consumption, aggregate wealth and expected stock returns. *Journal of Finance* 56, 815–49.

Lettau, M. and Ludvigson, S.C. 2001b. Measuring and modeling variation in the risk-return tradeoff. In *Handbook of Financial Econometrics*, ed. Y. Ait-Sahalia and L.P. Hansen. Amsterdam: North-Holland.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–54.

Lustig, H. and van Nieuwerburgh, S. 2005. Housing collateral, consumption insurance and risk premia: an empirical perspective. *Journal of Finance* 60, 1167–219.

Lustig, H. and Nieuwerburgh, S.V. 2006. Can housing collateral explain long-run swings in asset returns? Working Paper NYU Stern and UCLA.

Luttmer, E. 1996. Asset pricing in economies with frictions. *Econometrica* 64, 1439–67.

Mehra, R. and Prescott, E. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Menzly, L., Santos, T. and Veronesi, P. 2004. Understanding predictability. *Journal of Political Economy* 112, 1–47.

Piazzesi, M., Schneider, M. and Tuzel, S. 2007. Housing, consumption, and asset pricing. *Journal of Financial Economics* 83, 531–69.

Poterba, J.M. and Summers, L.H. 1986. The persistence of volatility and stock market fluctuations. *American Economic Review* 76, 1142–51.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7, 407–25.

Santos, J. and Veronesi, P. 2005. Labor income and predictable stock returns. *Review of Financial Studies* 19, 1–43.

Shiller, R.J. 1981. The use of volatility measures in assessing market efficiency. *Journal of Finance* 36, 291–304.

Shiller, R.J. 1982. Consumption, assest markets and macroeconomic fluctuations, *Carnegie-Rochester Series on Public Policy*, Vol. 17, 203–238, North-Holland Publishing Company.

How to cite this article

Guvenen, Fatih and Hanno Lustig. "consumption-based asset pricing models (theory)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000322> doi:10.1057/9780230226203.0300

The New Palgrave Dictionary of Economics Online

contemporary capitalism

William Lazonick
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The key to understanding ‘capitalism’ as a mode of resource allocation that generates economic growth is the organization and performance of its most innovative business enterprises. The ‘Old Economy business model’ that made the United States the world's most powerful nation in the post-Second World War decades came under challenge in the 1970s and 1980s, and the ideology of ‘maximizing shareholder value’ arose to legitimize a redistribution of income from labour interests to financial interests. The ‘New Economy business model’ emerged in the 1980s and 1990s to drive the innovation process, contributing, however, to unstable and inequitable economic growth.

Keywords

acquisitions; business enterprises; capitalism; collective capitalism; competitive advantage; conglomerate movement; contemporary capitalism; corporations; creative destruction; defined-benefit pensions; developmental state; dividend yield.; division of labour; economic development; entrepreneurship; financial commitment; foreign direct investment; globalization; Great Depression; hostile takeovers; information and communications technology; innovation; Japan, economics in; junk bonds; leveraged buyouts; lifelong employment; mergers; Milken, M.; NASDAQ; New Deal; New Economy business model; New York Stock Exchange; Old Economy business model; organization man; organizational integration; outsourcing; patents; research and development; Schumpeter, J. A.; separation of ownership and control; shareholder value; social inclusion; stock options; stock price volatility; stock repurchases; strategic control; technology; venture capital

Article

1 What is ‘ capitalism ’ ?

At the beginning of the 21st century, ‘capitalism’ has triumphed as the dominant system for allocating a society's economic resources. The last time in history in which the persistence of capitalism in the world's most advanced economies was seriously called into question was the Great Depression of the 1930s – a decade during which the unemployment rate in the United States remained at 15 per cent or higher, notwithstanding unprecedented state intervention under the New Deal. It took the Second World War to pull the United States and the world economy out of depression, and in the subsequent decades it took substantial and sustained government spending in the rich economies of North America and western Europe to hold unemployment to acceptable levels.

In the post-war era, the Soviet Union's highly planned economy posed as a possible alternative to capitalism. The purported strength of the Soviet challenge, however, turned out to be based at least as much on Cold War ideology emanating from the United States as on the actual productive power of the Soviet Union and its satellites. By the 1990s the Soviet model had virtually vanished, as Russia itself sought to make the transition to a ‘market economy’, guided, tragically, by a mythical ideology of how capitalism is supposed to operate, imported from the United States.

Over the same period capitalism entrenched itself in East Asia. During the 1970s and 1980s Japan became a rich economy on the basis of a distinctive model of ‘collective capitalism’, and in the 1980s and 1990s the East Asian ‘Tigers’ – Hong Kong, Singapore, South Korea and Taiwan – closed the gap, each with its own variant of the Japanese model. More recently China and India, with one-third of the world's population, have experienced rapid economic growth, driven by what many would call ‘capitalist’ institutions. Yet, even as firms cross the globe to access Indian software engineers, and vice versa, India remains a nation with one-third of the world's illiterates. Meanwhile the fact that China, the world's second largest economy since the early 1990s, continues to be guided by an avowedly Communist government raises the question of what ‘capitalism’ really is.

Defining contemporary capitalism is not merely a question of semantics. If, as has been demonstrated since the mid-20th century, ‘capitalism’ is a powerful engine of economic growth, we want to know how it functions as a mode of resource allocation and the social conditions under which capitalist growth is not only strong but also stable, and equitable. We also want to know how the institutions of contemporary capitalism that generate growth might be transferred to those parts of the world – first and foremost Africa but also eastern Europe and Latin America as well as parts of the Middle East – that have economically been left behind. Given its pervasiveness and dominance, a depiction of the institutions that define contemporary capitalism is tantamount to a description of the economic world in which perhaps one-half of the world's population now lives and to which much of the other half now aspires.

There is no consensus among economists on the definition of contemporary capitalism. The dominant approach to analysing resource allocation and the economic performance of an advanced economy rests on the notion that a capitalist economy is essentially a market economy that allocates resources to their most productive uses. But what at any time and in any place, the student of economic development asks, explains how those most productive uses come to exist? And why in certain times and places? Fundamental to capitalist growth is ‘innovation’, the process that generates goods and services that, even with factor prices held constant, are of higher quality and lower cost than those previously available (Lazonick, 2006c). Can a theory of capitalism as a market economy comprehend the innovation process?

In the early 20th century a young Joseph Schumpeter asked this question. As a Viennese economics student, Schumpeter was versed in the relatively recent, and increasingly influential, Austrian and

Walrasian theories of how, through the equilibrating mechanism of the market, the economy could achieve an ‘optimal’ allocation of resources across productive uses. Schumpeter's insight was to recognize that such a view of the economic world could not explain economic development. In 1911 Schumpeter wrote *The Theory of Economic Development* (first translated into English in 1934) to argue that entrepreneurial activity that results in innovation – what he called the ‘Fundamental Phenomenon of Economic Development’ – can disrupt the ‘Circular Flow of Economic Life as Conditioned by Given Circumstances’ to change the ways in which the economy operates and performs. Without such disruption of equilibrium conditions, the economy would not develop. Over the next four decades Schumpeter sought to elaborate a theory of economic development informed by his own, evolving, understanding of the changing reality of the most advanced capitalist economies.

In particular, Schumpeter sought to understand the role of the business enterprise in advanced capitalist development. By the 1940s he had taken definitive leave of his youthful conceptions of the innovative entrepreneur as an individual actor and innovation as simply ‘new combinations’ of existing resources. Rather, he saw that powerful business organizations both developed and utilized productive resources to create new technologies and access new markets. The creation of new technologies, moreover, destroyed the commercial viability of old technologies. In *Capitalism, Socialism, and Democracy*, first published in 1942, Schumpeter argued that the process of ‘creative destruction’ had become embodied in established corporations as ‘technological “progress” tends, through systematization and rationalization of research and of management, to become more effective and sure-footed’, being ‘the business of teams of trained specialists who turn out what is required and make it work in predictable ways’ (Schumpeter, 1950, pp. 118, 132).

This article takes as its point of departure the proposition, suggested by Schumpeter, that the key to understanding ‘capitalism’ as a mode of resource allocation that generates economic growth is the organization and performance of its most innovative business enterprises. That is not to say that markets and states are unimportant to the operation, and hence definition, of capitalism. Historically, however, well-functioning markets are outcomes of successful capitalist development. For the individual, markets create the possibility of choosing what to consume and for whom to work, including the prospect of working for oneself. But markets cannot explain the development of the new products and processes that drive the growth of the capitalist economy. The innovation process is uncertain, collective and cumulative (see O'sullivan, 2000). The uncertain character of innovation means that investments in innovation require *strategic control* over resource allocation by individuals who have intimate knowledge of the technologies, markets and competitors that an innovative strategy must confront. The collective character of innovation means that the implementation of an innovation strategy requires the *organizational integration* of a hierarchical and functional division of labour into a process of organizational learning. The cumulative character of innovation means that the process requires *financial commitment* until it can generate financial returns. Enterprises, not markets, engage in strategic control, organizational integration and financial commitment (Lazonick, 2003).

Nor can one explain innovation by appealing to the notion of the developmental state as its driving force, as has often been done for the East Asian economies. Implicit, and at times explicit, in this view is an acceptance of the ideology that the economic development of the United States is an exemplar of the workings of the market economy. Yet from gun manufacture and interchangeable parts in the first half of the 19th century to the computer revolution and Internet in the late 20th century, as well as railroads, aviation and the life sciences in between, the history of US capitalism is replete with examples of the

critical role of the developmental state in allocating resources to the processes of knowledge creation that then provided the foundations for US industrial leadership. Yet, as important as the developmental state has been even in a so-called ‘market economy’ such as the United States, the allocation of resources to knowledge creation would have been wasted, and would probably never have been made, had it not been for the presence and influence of innovative enterprises that have made use of this knowledge to generate higher-quality, lower-cost products than had previously been available.

In this article I focus on the changing role of innovative enterprise in determining resource allocation and economic performance in contemporary capitalism. Space constraints dictate that I confine the analysis of contemporary capitalism to the case of the United States, with the caveat that, even in a highly globalized economy in which one might expect convergence to a common business model, there are almost as many distinctive ‘varieties of capitalism’ in terms of governance, employment and investment institutions, as there are advanced capitalist nations. The US economy is, however, the world's largest and richest economy. It is also the one in which market ideology is most virulent and the actual mode of resource allocation most misunderstood. Section 2 of this article provides historical background to understanding contemporary US capitalism by describing the key characteristics of the ‘Old Economy business model’ (OEBM) that made the United States the world's most powerful nation in the decades after the Second World War. Section 3 analyses the challenges that confronted OEBM in the 1970s and 1980s, and how the ideology of ‘maximizing shareholder value’ arose to legitimize a redistribution of income from labour interests to financial interests. Section 4 shows how the ‘New Economy business model’ (NEBM) emerged in the 1980s and 1990s to drive the innovation process, but in ways that have contributed to unstable and inequitable economic growth. Section 5 concludes with some questions about the future of the US model in a global economy in which many distinctive business models still compete.

2 The Old Economy business model

The United States emerged from the Second World War as the undisputed world leader in GDP per capita, a position that it still retains. With western Europe and Japan still in recovery from the war, the United States was at its peak of dominance in the 1950s on the basis of a highly collective model of capitalism embodied in the managerial corporation, and personified in the concept of the ‘organization man’ (Whyte, 1956). The stereotypical ‘organization man’ was white, Anglo-Saxon and Protestant, obtained a college education, got a well-paying job with an established company early in his career, and then worked his way up and around the corporate hierarchy over decades of employment, with a substantial ‘defined benefit’ pension, complete with highly subsidized medical coverage, awaiting him on retirement. The employment stability offered by an established corporation was highly valued, while inter-firm labour mobility was shunned.

‘Organization men’ rose to top executive positions where, as salaried managers rather than owners, they exercised strategic control. This separation of share ownership and managerial control, which continues to characterize the US industrial corporation, resulted from the widespread distribution among shareholders of the corporation's publicly traded stock. In principle, boards of directors representing the interests of shareholders monitor the decisions of these managers. In practice, incumbent top executives choose the outside directors and are themselves members of the board. Shareholders can challenge

management through proposals to the annual general meeting, but over the course of the 20th century a body of law evolved that enables management to exclude shareholder proposals that deal with normal business matters (for example, downsizings) as distinct from social issues (for example, sex discrimination).

The separation of ownership from control has worked effectively to generate innovation when the interests of salaried executives who exercise strategic control have been aligned with those of employees who engage in the development and ensure the utilization of the company's productive resources. In the post-Second World War decades the organizational integration of the capabilities of administrative and technical specialists enabled US firms to develop the world's most competitive systems of mass production. These personnel were products of the US system of higher education, which since the early decades of the century had prepared the labour force to enter employment in bureaucratic organizations. A distinctive feature of the US business model was the organizational segmentation between these salaried managers, in whose training and experience the corporation made substantial investments, and so-called ‘hourly’ workers. (Non-salaried employees were classified as ‘hourly’, or ‘non-exempt’, workers because of the stipulation of the National Labor Relations Act that emerged from the New Deal era that required employees who were paid an hourly wage receive 150 per cent of that wage if they worked longer than the normal working hours. The overtime work of salaried personnel is exempt from this provision.) The corporation viewed these operatives, who were typically high-school graduates, as interchangeable commodities in whose capabilities the company had no need to invest, notwithstanding the fact that they often spent their entire working lives with one company. At the same time, these industrial corporations needed reliable even if low-skill workers to tend mass production processes. The combination of dominant product-market positions and union power, which advanced the pay and protected the employment of senior workers, enabled the hourly worker to receive good pay and benefits, including a defined-benefit pension that assumed long-term employment with a single company.

The developmental state played an indispensable role in the innovation process by partially funding the system of higher education as well as, in the forms of research labs, subsidies and contracts, programmes for technology development in sectors such as aerospace, computers and life sciences. The development of the productive potential of these government investments relied in turn on corporate research capabilities. Retained earnings formed the foundation of committed finance for new corporate investments in innovation. When corporations needed additional investment financing, they issued corporate bonds at favourable rates that reflected the established position of the company as well as its conservative debt–equity ratios. Companies used bank loans almost exclusively for working capital, and made only limited use of the stock market as a source of investment funds.

These social conditions enabled US corporations to grow very large in the post-war decades. The 50 largest US industrial corporations by revenues on the Fortune 500 list averaged 87,070 employees in 1957, 117,393 in 1967, and 119,093 in 1977. These figures do not include employment at AT&T, the regulated telephone monopoly, which in 1971 employed 1,015,000 people, of whom 700,000 were union members with good wages, stable employment and excellent benefits. By the late 1960s and early 1970s increasing numbers of blacks were moving into union jobs in the steel, automobile, electrical equipment, consumer durable and telecommunications industries. The growth of established corporations in these industries in the three decades after the Second World War contributed to a more equal distribution of family income in the US economy.

3 ‘ Maximizing shareholder value ’

During the 1970s the US model faltered in the face of Japanese competition. Building on innovative capabilities developed for their home markets during the 1950s and 1960s, Japanese companies gained competitive advantage over US companies in industries such as steel, memory chips, machine tools, electrical machinery, consumer electronics and automobiles. US companies had entered the 1970s as world leaders in these industries. Many US observers attributed the rapid increase in Japanese exports to the United States in the 1970s to Japan's lower wages and longer working hours. By the early 1980s, however, with real wages in Japan continuing to rise, it became clear that Japanese advantage was based on the superior organization of their enterprises, and in particular on a more thoroughgoing integration of participants in the functional and hierarchical divisions of labour for the dual purposes of transforming technologies and accessing new markets. Indeed, during the 1980s Japan exported management practices as well as material goods to the West. From the second half of the 1980s, with the yen strengthening and trade surpluses generating political backlash, Japanese companies made a transition to direct investment in the United States and other advanced economies.

A growing financial orientation of US business that had surfaced in the conglomerate movement of the 1960s undermined the abilities and incentives of established US corporations to respond to the Japanese challenge. To some extent the growth of the US industrial corporation in the post-war decades had been based on strategic investments in new product lines and geographic areas that built on the corporation's existing productive capabilities, and yielded economies of scale and scope. The conglomerate movement, however, saw major corporations invest in scores of *unrelated* businesses, often through mergers and acquisitions, based on the prevailing, but erroneous, ideology that a good corporate executive could manage any type of business, and that conglomeration offered the synergies of superior corporate management. The conglomerate movement failed because it segmented top executives, in positions of strategic control, from the rest of the managerial organization that had to develop and utilize productive resources to sustain the firm's competitive advantage (Lazonick, 2004).

In the late 1970s and early 1980s the conglomerates unraveled. In the mid-1970s Michael Milken, a Drexel Burnham investment banker, had created the junk bond market by convincing institutional investors, in search of higher yields in an inflationary era, to hold downgraded corporate securities, many of them ‘fallen angels’ from unsuccessful conglomeration. By the late 1970s, with the junk-bond market well developed, it became possible to issue new junk bonds to finance leveraged buyouts (LBOs) in which the top managers of a conglomerate division turned it into an independent company to recapture strategic control over resource allocation. By the late 1980s, however, the junk bond had become an instrument for the hostile takeover of entire companies, with KKR's 1989 LBO of RJR Nabisco for $24.5 billion marking the height of what became known as ‘the deal decade’.

The ideology that justified hostile takeovers was that the corporation should be run to ‘maximize shareholder value’ (see Lazonick and O'sullivan, 2000). Proponents of shareholder value charged that, either because of opportunism or incompetence, many incumbent corporate managers were making poor allocative decisions. By exercising their influence through the market for corporate control, shareholders could force incumbents to alter their allocative decisions, replace them with those who would maximize shareholder value, or distribute cash to shareholders in the forms of dividends and stock repurchases so that shareholders themselves could, so the argument goes, reallocate the economy's resources to their best alternative uses.

While the hostile takeover movement did not directly threaten high-tech companies (in which the most valuable assets could walk out the door), by the end of the 1980s the top executives of virtually all US industrial corporations had embraced the ideology of maximizing shareholder value and made it their own. By the 1980s executive stock option compensation was a well-established practice. Since in the United States option awards did not require that the company's stock price outperform the stock market or even the stock prices of a group of competitors, those who received these awards could only gain from what, from July 1982 to August 2000, turned out to be the longest stock market boom in US history, with the Dow Jones Industrial Average and the S&P500 Index both rising about 1,300 per cent. As Table 1 shows, stock-price appreciation drove the extraordinary real stock yields that were sustained over the 1980s and 1990s. The relatively low dividend yields in the 1990s did not reflect stinginess on the part of US corporations; the US corporate payout ratio – the amount of dividends as a percentage of after-tax corporate profits (with inventory evaluation and capital consumption adjustments) – averaged 48 per cent in the 1980s and 57 per cent in the 1990s compared with 39 per cent in the 1960s and 41 per cent in the 1970s. It was just that the rate of increase of stock prices outstripped the rate of increase of dividend payments, thus depressing the dividend yield. The form that the stock yield takes is of significance because investors can capture the dividend yield by holding stocks, whereas they can capture the price yield only by buying and selling stocks. Inherent in high-price yields, therefore, is a volatile stock market.

US corporate stock and bond yields, 1960–2005. Average annual per cent change

	1960–9	1970–9	1980–9	1990–9	2000–5
Real stock yield	**6.63**	**–1.66**	**11.67**	**15.01**	**–1.87**
Price yield	5.80	1.35	12.91	15.54	–0.76
Dividend yield	3.19	4.08	4.32	2.47	1.58
Change in CPI	2.36	7.09	5.55	3.00	2.67
Real bond yield	**2.65**	**1.14**	**5.79**	**4.72**	**3.60**

Notes: Stock yields are for Standard and Poor's composite index of 500 US corporate stocks (424 of which are, as of 28 March 2006, NYSE). Bond yields are for Moody's Aaa-rated US corporate bonds.
Source: Council of Economic Advisers (2006, Tables B-62, B-73, B-95 and B-96).

A volatile stock market benefits those who are compensated in stock options on an annual basis, especially when, as is the case in the United States, options vest as quickly as one year from the date of grant and can be exercised for up to ten years. It has been estimated that, largely because of the gains from exercising stock options, on average the ratio of CEO pay of an S&P500 company to that of a production worker was 42 in 1985, 107 in 1990, 525 in 2000, and 411 in 2005. Top executives took a keen interest in their company's stock price, and in the 1980s and 1990s, in the name of ‘maximizing shareholder value’, they found ways in which they could use their positions of strategic control over corporate resource allocation to influence it. They could cook the corporate books to boost current earnings, a practice that became widespread in these decades and one for which a few executives have been fined or even jailed. The American Competitiveness and Corporate Accountability Act of 2002, better known as Sarbanes–Oxley, has sought to stem this practice. But quite apart from artificially inflating corporate earnings, top corporate executives also found that downsizing the labour force and repurchasing corporate stock helped to boost a company's stock price, even though these resource

allocations did not necessarily improve the company's competitive performance.

The era of corporate downsizing took hold in the recession of 1980–2 when hundreds of thousands of stable, well-paid blue-collar jobs were lost that were never subsequently restored (see Lazonick, 2004). It would appear that the blacks who had relatively recently moved into these types of jobs were particularly hard hit; last hired, they tended to be the first fired. The subsequent ‘boom’ years of the mid-1980s witnessed hundreds of plant closings. In the ‘white-collar’ recession of the early 1990s tens of thousands of professional, administrative and technical employees found that their jobs had been eliminated, although once again it was blue-collar workers who bore the brunt of the downturn. In 1980 manufacturing employment was 22 per cent of the labour force; by 1990 it had fallen to 17 per cent and by 2001 to 14 per cent. While the employment picture generally became much better during the Internet boom of the last half of the 1990s, job cutting remained a way of life for many major US corporations. According to data on layoff announcements by companies in the United States collected by the recruitment firm, Challenger, Gray and Christmas, announced job cuts averaged just under 550,000 per year for the period 1991–4, 450,000 per year in 1995–7, and 656,000 per year during the boom years 1998–2000.

Meanwhile, from the mid-1980s US corporations began to actively support their stock prices through large-scale stock repurchases. Companies included in the S&P500 in March 2006 distributed more cash to shareholders in repurchases than in dividends in 1997 through 2000 and again in 2004, and just slightly less in 2001 through 2003. Since 1978 net equity issues by US non-financial corporations has been positive in only six of 28 years (1980, 1982, 1983, 1991, 1992, 1993); since the early 1980s US industrial corporations have in aggregate been supplying capital to the stock market rather than vice versa. In 2005 the net flow of cash from non-financial corporations to the stock market was a record $366 billion, 1.42 times in real dollars the previous high in 1998 (Lazonick, 2006d).

4 The New Economy business model

On 29 December 1995, AT&T announced that, as part of the process of breaking itself up into three separate companies, it would be cutting 40,000 jobs. AT&T was a company that could trace its origins back to the 1870s, had created the world's most advanced telephone system, was the home of the famous Bell Labs that among many other accomplishments invented the transistor in 1947, and, despite having lost its status as a regulated monopoly in 1984, still employed 308,700 people. Now, however, AT&T became emblematic of the failure of US Old Economy corporations to continue to provide employment opportunities. With campaigning for the 1996 presidential election picking up steam, Patrick Buchanan, a right-wing politician, caught the attention of the media by denouncing the highly paid executives of AT&T and other downsizing corporations as ‘corporate hit men’. Fuel was added to the fire by the revelation that, in the name of ‘creating shareholder value’, Al Dunlap, whom the American public came to know as ‘Chainsaw Al’, had in 20 months as CEO of Scott Paper devastated the 115-year old company while putting an estimated $100 million in his own pocket. In March 1996, the *New York Times* ran a seven-part series, later released as a paperback, on ‘the downsizing of America’ (Lazonick, 2004).

By the spring of 1996, however, the furor over corporate downsizing had disappeared. In its place, Americans became enthralled by the prosperity promised by what in the second half of the 1990s came

to be called the ‘New Economy’. In the United States the previous half-century had seen a massive accumulation of information and communications technology (ICT) capabilities. The development of computer chips from the late 1950s had provided the technological foundation for the microcomputer revolution from the late 1970s, which in turn had provided the technological infrastructure for the Internet boom of the second half of the 1990s. The research funding for this accumulation of ICT capabilities had come mainly from the US government and the research laboratories of established Old Economy high-technology corporations. Each wave of technological innovation, however, created opportunities for the emergence of start-up companies that were to become central to the commercialization of the new technologies.

Although by the mid-1980s the Japanese had outcompeted even the leading US semiconductor firms in the memory chip market, US companies such as Intel, Motorola and Texas Instruments continued to dominate the microprocessor and logic chip markets that drove product innovation in the microelectronics industry (Lazonick, 2006a). While Silicon Valley was not the only US location for innovation in this industry, the concentration of semiconductor start-ups in the region from the late 1950s resulted in the emergence by the 1980s of a distinctive mode of combining strategy, finance and organization: the ‘New Economy business model’ (NEBM) (see Table 2). During the 1990s NEBM spread beyond Silicon Valley start-ups and was adopted successfully by leading Old Economy ICT companies such as Hewlett-Packard and IBM. In the Internet boom of the late 1990s elements of NEBM diffused to other ICT companies, including an Old Economy company such as Lucent Technologies, spun off from AT&T in its 1996 trivestiture, which almost destroyed itself in attempting to adopt the business model. In the 2000s NEBM characterizes the most innovative sectors of the US economy (for the case of biotechnology, see Pisano, 2006).

Comparing business models in ICT

	Old Economy business model (OEBM)	New Economy business model (NEBM)
Strategy, product	Firm growth based on multidivisional structure: multi-product firm	New firm entry into specialized ICT markets; accumulate new capabilities by acquiring (other) young technology firms
Strategy, process	Vertical integration of the value chain; in-house standards and proprietary R&D	Vertical specialization of the value chain; industry technology standards; R&D for cross-licensing and alliances; outsourcing routine work to specialist contract manufacturers and/or offshoring routine work to low-wage nations
Finance	Venture finance from savings, family and business associates; NYSE listing, growth finance from retentions, after dividends, and bond issues	Organized venture capital; early IPO on NASDAQ; retentions with zero dividends; use of own stock as a compensation and combination currency; systematic stock repurchases to support stock price

Organization	Secure employment; ‘organization man’ (career with one company), industrial union; defined-benefit pension, good medical coverage in employment and retirement	Insecure employment; interfirm mobility of labour, broad-based stock options, non-union, defined contribution pension, employees bear more burden of medical insurance

The founders of New Economy firms have typically been scientists and engineers who have gained specialized experience in existing firms, although in some cases they have been university faculty members intent on commercializing their academic knowledge. Some of these entrepreneurs have come from existing Old Economy companies, where it was often difficult for their new ideas to get internal backing. But New Economy companies themselves have become increasingly important as sources of new entrepreneurs who left their current employers to start new firms. Large numbers of high-tech entrepreneurs in the United States have been foreign-born, coming mainly from India and China (Saxenian, 2006).

Typically, the founding entrepreneurs of a New Economy start-up seek committed finance from venture capitalists with whom they share not only ownership of the company but also strategic control. In the 2000s Silicon Valley remains by far the leading location in the United States and the world for venture-backed high-tech start-ups. The region acquired this position from the 1960s as a distinctive venture capital industry emerged out of the opportunities for start-ups created by the microelectronics revolution. Besides sitting on the board of directors of the new company, the venture capitalist generally recruits professional managers, who are given company stock along with stock options, to lead the transformation of the firm from a new venture to a going concern. This stock-based compensation gives these managers a powerful financial incentive to develop the innovative capabilities of the company to the point where it can do an initial public offering (IPO) or private sale to a listed company, thus enabling the start-up's privately held shares to be transformed into publicly traded shares. Both before and after making this transition, their tenure with, and value to, the company depends on their managerial capabilities, not their fractional ownership stakes (Lazonick, 2006a).

The stock market speculation of the ‘dotcom’ era made it all too easy to cash out of a start-up, as many high-tech firms that had not engaged in innovation did IPOs or were sold to established companies. When start-ups do innovate, the key to making the transition from new venture to going concern has been the organizational integration of an expanding body of technical and administrative ‘talent’. As Silicon Valley developed from the 1960s, this educated and experienced labour had to be induced to trade secure employment with an Old Economy company for insecure employment with a start-up. To attract these highly mobile people and retain their services, Silicon Valley firms increasingly adopted ‘broad-based’ employee stock option plans that extended this form of compensation to a large proportion, sometimes all, of the firm's non-executive employees rather than just to top executives. In start-ups, stock options usually served as a partial substitute for cash salaries, and the eventual gains from exercising options were viewed as a substitute for a company-funded pension (Lazonick, 2006a).

Again, the underlying stock would become valuable if and when the start-up did an IPO or a private sale to a publicly listed company. Shortening the expected period between the launch of a company and an IPO was the practice of most venture-backed high-tech start-ups of going public on NASDAQ, created in 1971 as an electronic exchange for the over-the-counter markets with less stringent listing requirements than the ‘Old Economy’ New York Stock Exchange (NYSE). The 1978 cut in the capital

gains tax rate to 28 per cent, after it had been raised to 49 per cent just two years before, provided further encouragement to entrepreneurs and venture capitalists to found new companies, and for employees of these companies, rewarded with stock options, to provide the skills and efforts needed to transform new ventures into going concerns. In 1979 the clarification of the ‘prudent man’ rule as applied to the Employee Retirement Income Security Act (ERISA) of 1974 gave asset managers the green light to allocate a portion of their portfolios to riskier stocks and venture capital funds, and resulted in a flood of new money, especially from pension funds, into the venture capital industry. The American Electronics Association and the National Venture Capital Association, with their strongest and deepest roots in Silicon Valley, were the frontline Washington lobbyists for these regulatory changes.

While institutional money provided capital to NEBM, high-tech labour became more mobile from one firm to another than it had been in the Old Economy. Employee stock options induced this mobility, but what made it possible in terms of the knowledge bases that managers and engineers possessed were *industry* standards, as distinct from *in-house* standards, that emerged in the various sectors of ICT. In the Old Economy in-house standards promoted the growth of large vertically integrated firms on the basis of proprietary technologies, whereas in the New Economy industry standards encouraged new entry. Nevertheless, as demonstrated by the important cases of Intel and Microsoft in the development of the microcomputer industry, those New Economy firms that dominated in the setting of the industry standards could also grow very large (at the end of fiscal 2005 Intel employed 99,900 and Microsoft 61,000). By establishing industry standards, their growth encouraged rather than discouraged start-ups, which in turn depended on the availability of not only venture capital (which came from many sources besides the formal venture capital industry) but also mobile labour whose knowledge and experience could be easily integrated into the start-up's learning processes.

Of critical importance in setting industry standards in microelectronics was IBM's decision in 1980 to enter what became known as the personal computer (PC) industry with Intel supplying the microprocessor and Microsoft the operating system. At the time IBM controlled about 80 per cent of the computer market, had over 341,000 employees, and, with an explicit system of ‘lifelong employment’, trumpeted the fact that since 1921 it had not terminated an employee involuntarily. Yet between 1990 and 1994 IBM slashed its employment from 374,000 to 220,000. In 1991–3, the company had losses of $16 billion (including more than $8 billion in 1993, at the time the largest annual loss in US corporate history) on total revenues of $192 billion, and encouraged the media to believe that the mass layoffs were necessary to avoid bankruptcy. Yet virtually all of the losses came from ‘restructuring’ charges, that is, the cost of terminating employees (Lazonick, 2006a).

In retrospect, it is clear that these charges were the cost of ridding the company of its 70-year-old system of lifelong employment. The industry standards in ICT, which IBM had played a leading role in establishing, served to reduce the value to the company of older employees with experience accumulated at IBM over the course of their careers and to increase the value of younger employees who may have had experience working for other ICT companies. Explicitly reflecting this change in employment policy, in 1999 IBM announced that it would replace its traditional defined-benefit pension plan, which favoured long-term employees, with a portable ‘cash-balances’ plan that would be much more attractive to younger employees who did not envisage a lifelong career with IBM. In December 2004, as its employment reached 329,000, IBM announced that new employees would no longer be eligible for the cash-balances pension fund. Instead the company would offer them a defined-contribution pension, with the company matching the employee contribution up to six per cent of his or her compensation.

From the mid-1990s, with the Old Economy commitment to its employees out of the way, IBM adopted all of the elements of NEBM. It shifted out of hardware into services, and outsourced its manufacturing. It became by far the leading patenter in the United States, even as it cut R&D from the ten per cent of sales that prevailed in the 1980s to six per cent of sales since the mid-1990s, this change reflecting an expressed shift to product development and away from basic research. Since the early 1990s IBM has engaged in patenting much less to control proprietary technologies, as had been the case in the past, and much more to gain access through cross-licensing to technologies controlled by other companies and to generate intellectual property revenue ($1.3 billion per year in the 2000s).

As it rid itself of lifelong employment in the early 1990s IBM began to extend stock options, previously reserved for top executives, to a broad base of employees. In 1990 options outstanding were only four per cent of all shares outstanding; in 2005, 15.2 per cent. As for distributions to shareholders, in New Economy fashion, subsequent to its early 1990s restructuring IBM has favoured repurchases over dividends. In 1981–90 IBM's dividends were 48 per cent and repurchases 12 per cent of net income; in 1993–2005 dividends were 15 per cent and repurchases 91 per cent. In an effort to offset dilution of shareholdings as employees exercise stock options, and more generally to boost its stock price, in 1995–2005 IBM has spent $62.6 billion on stock repurchases. Over the same period the company has spent $56.6 billion on R&D.

As for a New Economy company that, unlike IBM, started out that way, Cisco Systems, which since the late 1990s has controlled about 75 per cent of the Internet router market, is a prime example of the importance, and implications, of broad-based stock options in NEBM compensation. Founded in Silicon Valley in 1984, Cisco grew from about 200 employees at the time of its IPO in 1990 to 40,000 employees during 2000. Throughout its history Cisco has awarded stock options to virtually all of its employees. By the end of fiscal 2000 stock options outstanding accounted for 14 per cent of the company's total stock outstanding; by 2005 that number was 23 per cent. In March 2000, at the peak of the Internet boom, Cisco had the highest market capitalization of any company in the world. Under such conditions its stock options were very lucrative. I have estimated that over the 11 years 1995–2005 (all years for which data are reported refer to fiscal year's end, the last week in July), Cisco employees, totaling about 256,000 employee-years, shared $21.5 billion in gains from exercising stock options, for an average of $84,000 per employee-year. The annual averages per employee ranged from less then $9,000 in 2003 to more than $281,000 in 2000. Of the total amount, the highest paid executives, totaling 57 executive-years, shared $893 million, for an average of $15.7 million per executive-year, with annual averages ranging from $1.3 million in 2003 to $51.3 million in 2000. The annual ratios of average top-executive to average employee gains from exercising stock options ranged from 36:1 in 1997 to 594:1 in 2005 (Lazonick, 2006d). Cisco employees have a clear financial interest in the company's stock price, and the company's top executives even more so.

Besides using their own stock as a compensation currency, during the 1990s some New Economy companies grew large by using their stock, instead of cash, to acquire other, smaller and typically younger, New Economy firms in order to gain access to new technologies and markets. Cisco mastered this growth-through-acquisition strategy. From 1993 through 2005 Cisco made 106 acquisitions valued in nominal terms at $46.9 billion, over 80 per cent of which was paid in the company's stock rather than cash. In 1999 and 2000 alone, Cisco did 41 acquisitions at a cost of $26.7 billion with over 99 per cent paid in stock (Carpenter, Lazonick and O'sullivan, 2003).

At the same time, like many if not most New Economy companies, Cisco conserved cash by paying no

dividends. Along with its use of stock as a combination currency, this payout policy enabled Cisco to become a giant company in the 1990s without taking on any long-term debt. Since the bursting of the Internet bubble from mid-2000 through 2005, however, Cisco has spent $27.2 billion repurchasing its own stock to support its sagging stock price. In 2004–5, as it spent $19.3 billion on stock repurchases, Cisco used $8.3 billion in cash – including $6.5 billion of it raised through its first-ever bond issue – to do 24 acquisitions rather than continue to use its stock as an acquisition currency that it would then feel compelled to offset with repurchases. (Cisco's decision to use cash rather than stock for acquisitions was helped by the Financial Accounting Standards Board's 2001 closing of the ‘pooling-of-interests’ loophole that enabled companies like Cisco that did all-stock acquisitions to record them on their balance sheets at book values, which were generally a small fraction of market values, and thus inflated future earnings. Nevertheless, in 2002 and 2003, with pooling-of-interests accounting outlawed, Cisco still used stock for payment of over 97 per cent of the price of its nine acquisitions.)

The corporate obsession with supporting its stock price through massive stock repurchases has therefore taken hold of companies in the most innovative sectors of the US economy. As further notable examples, for the years 1995–2005 Intel distributed $51.3 billion in repurchases along with $6.0 billion in dividends compared with R&D spending of $38.0 billion, while Microsoft distributed $45.4 billion in repurchases and $38.7 billion in dividends compared with R&D spending of $40.8 billion. Microsoft's dividends included a one-time payment of $36.1 billion in November 2004.

These companies would argue that R&D spending and stock repurchases are both working toward the same end: to enhance the company's innovative capabilities by, in the case of R&D, generating new knowledge, and in the case of repurchases, competing for high-tech labour capable of transforming that knowledge into innovative products and processes. By boosting stock prices, it is argued, repurchases help to attract, retain and motivate people who choose to work for companies in which they are partially compensated with stock options. In the case of Microsoft the argument has had less weight since July 2003 when the company ended its option programme (although many Microsoft employees still have unexercised options awarded prior to that date). In the 2000s, moreover, the extent and location of the talented labour supplies for which companies like Cisco, IBM, Intel and Microsoft compete have changed dramatically with the rise of India and China (Lazonick, 2006b). These dramatically changed labour market conditions for high-tech labour raise serious questions concerning which employees benefit from a company's stock price performance and for how long, and indeed whether established high-tech companies even need to use employee stock options to compete successfully for high-tech labour.

The offshoring to India and China in the 2000s of high value-added jobs of software engineers and computer programmers that it was previously thought could not go abroad represents the latest stage in four decades of the globalization of NEBM. Beginning in the early 1960s Silicon Valley semiconductor companies were among the first to offshore assembly to East Asia, and by the early 1970s virtually every US semiconductor manufacturer had followed suit. When these companies set up plants in places like South Korea, Taiwan, Hong Kong, Singapore and Malaysia, they employed, alongside unskilled and predominantly female assembly labour, indigenous university graduates as managers and engineers. Over time the US companies upgraded their facilities in these locations, and offered more and better employment opportunities for the indigenous well-educated labour force. As a striking example, in 1984 Intel claimed that, of its 8,500 employees outside of the United States (of 26,000 employees worldwide), only 60 were US citizens. This indigenous employment through foreign direct investment encouraged

the national governments to increase the level of investment in their already well-developed systems of higher education, thus augmenting the future high-tech labour supply (Lazonick, 2006b).

In the 1990s established US ICT companies, led by IBM and Hewlett-Packard, dramatically reduced their employment of production workers by outsourcing manufacturing operations to electronic manufacturing service providers, also known as contract manufacturers (Lazonick, 2006c). Indeed, younger companies like Cisco grew rapidly without doing any in-house manufacturing. Initially the contract manufacturers would set up operations or take over existing plants of their customers in the United States. But a key capability of the leading contract manufacturers is to shift production that has become more routine and cost-sensitive to lower wage areas of the world. In the late 1990s and early 2000s the leading contract manufacturers grew at a rapid pace; at the end of 2005 employment at the five largest – Flextronics, Solectron, Sanmina-SCI, Celestica and Jabil Circuit – totalled 260,000. While we do not know the global distribution of this labour force, North America accounts for only an estimated 25 per cent of the sales of these five companies.

Meanwhile, in the 1990s and 2000s hundreds of thousands of foreigners, especially Indians, with college degrees in science and engineering have migrated to the United States for graduate education and work experience (Lazonick, 2006b). Many acquired permanent resident (immigrant) status in the United States, as the US government expanded employment-based preferences in the issuance of immigrant visas. For access to US work experience, however, the most important mode of entry for high-tech employees has been on non-immigrant H-1B and L-1 visas. The H-1B programme enables non-immigrants, the vast majority of whom have at least a bachelor's degree and whose skills are purportedly unavailable in the United States, to work in the United States for up to six years. In the first half of the 2000s about 70 per cent of H-1B visa holders had science or technology degrees, and 40–50 per cent came from India (the next largest national group is from China, at less than ten per cent). The L-1 visa programme permits a company with operations in the United States to transfer foreign employees to the United States to acquire work experience, with no limitation of time. In 2001, there were an estimated 810,000 people on H-1B visas in the United States, and possibly as many on L-1 visas.

Many of these non-immigrant visa holders have continued to work in the United States by obtaining permanent resident status. But most have returned to their native countries with valuable industrial experience that can be used to start new firms and, more typically, to work as technical specialists for indigenous or foreign companies. As a result of both the migration of US companies abroad in search of high-tech labour as well as the migration of foreign high-tech labour to the United States, and then back to their home countries, in the 2000s, to an extent never before imagined, even the best-educated US high-tech employees compete with a truly global labour supply for jobs.

5 Stable and equitable growth?

On 16 March 2005 the Semiconductor Industry Association (SIA) organized a Washington, DC press conference in which it exhorted the US government to step up support for research in the physical sciences, including nanotechnology, to assure the continued technological leadership of the United States. Intel CEO Craig Barrett was there as a SIA spokesperson to warn: ‘U.S. leadership in technology is under assault’ (*Electronic News*, 2005):

> The challenge we face is global in nature and broader in scope than any we have faced in the past. The initial step in responding to this challenge is that America must decide to compete. If we don't compete and win, there will be very serious consequences for our standard of living and national security in the future…U.S. leadership in the nanoelectronics era is not guaranteed. It will take a massive, coordinated U.S. research effort involving academia, industry, and state and federal governments to ensure that America continues to be the world leader in information technology.

At the time Barrett was a member of the US National Academy of Sciences Committee on Prospering in the Global Economy in the 21st Century, which delved into deficiencies in the development of science and engineering capabilities in the United States. Notwithstanding his obvious concern about these problems from a public policy perspective, on a radio talk show in February 2006 Barrett (by this time Chairman of Intel) remarked: ‘Companies like Intel can do perfectly well in the global marketplace without hiring a single US employee’ (wbur.org, 2006).

The problem with this statement is not that US workers should have privileged access to jobs at a US-based company like Intel (which still employs half of its almost 100,000 employees in the United States). The problem is that, if a powerful company like Intel is not dependent on US high-tech employees for its future labour force, why should it be concerned about supporting the mass educational infrastructure in the United States needed to develop this future labour force? And what does it mean to say that ‘America must decide to compete’ if, as I would argue is the case (Lazonick, 2006b), the most innovative US corporations have more of an interest in the Malaysian or Indian system of mass education than in the US system?

Since the mid-1970s the US mass education system has been performing poorly in science and mathematics by the standards of both the advanced and many developing economies. Such was much less the case in the three decades or so after the Second World War, when the Old Economy corporation was more dependent upon a labour force that was well-educated at the primary and secondary levels in the United States. This shift in the performance of the mass education parallels the reversal of post-war progress towards a more equal distribution of income that began about three decades ago. The much less secure employment of most US corporate employees in the shift from OEBM to NEBM would seem to have contributed to this reversal.

Meanwhile in the 2000s the compensation of the CEOs of US corporations has long since passed levels that are at a minimum unseemly and some would say obscene. The ‘explosion in CEO pay’, which has been discussed in the United States since the mid-1980s, seems to have no limits, especially if, when the corporate stock price falls, it can be once again pumped up or boards of directors can replace the ‘lost’ stock option income by other forms of remuneration such as salaries, bonuses or restricted stock. The seemingly endless explosions in top-executive pay reflect the obsession of US corporate executives with ‘maximizing shareholder value’ and, cash flow permitting, disgorging billions upon billions of corporate cash to shareholders in the forms of repurchases and dividends to try to make it happen.

In terms of public policy initiatives, virtually nothing has been done to control top executive pay in the United States. One well-known attempt was misguided. In 1993 President Clinton carried out a campaign promise to control CEO pay by legislating a cap of $1 million on the amount of ‘non-performance-related’ top executive compensation – salary and bonus – that a corporation could claim as a tax deduction. One perverse result of this law was that companies that were paying CEOs less than $1

million in salary and bonus *raised* these components of CEO pay towards $1 million, which executives now viewed as the government-approved CEO 'minimum wage'. The other perverse result was that companies increased CEO option compensation, for which tax deductions were not in any case being claimed, as an alternative to exceeding the $1 million salary-and-bonus cap.

That having been said, the limits to the gains from stock options, not just for top executives but also for broad bases of the employees of US high-tech corporations, would long ago have been reached if not for the fact that many of these corporations have been in the forefront of innovation. Given the unchallenged sway that the ideology of 'maximizing shareholder value' has over the governance of these corporations, I have no doubt that instability, as reflected in the boom and bust of the stock market in the late 1990s and early 2000s, and inequity, as reflected in the worsening of the distribution of income, will continue to beset the US economy.

Whether US corporations will remain in the forefront of innovation that, by necessity, must underpin long-term economic growth is another matter. Notwithstanding globalization, the US model of contemporary capitalism is not a global model. No other contemporary capitalist economy has made the commitment to 'shareholder value' that is the most distinctive feature of the US model. Japan has come through the stagnation of the 1990s as a highly innovative economy, while eschewing shareholder value ideology and practices (Lazonick, 2005). In western European nations the ideology has been tempered by a commitment to 'social inclusion'; the question is whether the equity and stability that social inclusion brings can be harnessed to support innovative enterprise. In the emerging giants, India and China, the stock market has come to play a more important, and possibly dangerous, role. In all of these economies, the success of innovative companies has been based, however, not on the stock market, but on the principles of strategic control, organizational integration, and financial commitment. Historically these principles also underpinned innovative enterprise in the United States. Many corporate executives who exercise control over resource allocation in the US economy may, however, have forgotten these principles, or worse yet, while they have been busy enriching themselves, they may have never bothered to learn them.

Bibliography

To conserve both the word-count and flow of this essay, I have kept bibliographic references to a minimum, indicating instead works of mine in which these references can be found.

Carpenter, M., Lazonick, W. and O'sullivan, M. 2003. The stock market and innovative capability in the New Economy: the optical networking industry. *Industrial and Corporate Change* 12, 963–1034.

Council of Economic Advisers. 2006. *Economic Report of the President, 2006*. Washington, DC: Executive Office of the President.

Electronic News. 2005. US could lose race for nanotech leadership, SIA panel says. 16 March. Online. Available at http://www.reed-electronics.com/electronicnews/article/CA511197?nid=2019&rid=1344283927, accessed 5 September 2006.

Lazonick, W. 2003. The theory of the market economy and the social foundations of innovative enterprise. *Economic and Industrial Democracy* 24, 9–44.

Lazonick, W. 2004. Corporate restructuring. In *The Oxford Handbook of Work and Organization*, ed. S. Ackroyd et al. Oxford: Oxford University Press.

Lazonick, W. 2005. The institutional triad and Japanese development [translated into Japanese]. In *The Contemporary Japanese Enterprise*, vol. 1, ed. G. Hook and A. Kudo. Tokyo: Yukikaku Publishing.

Lazonick, W. 2006a. Evolution of the New Economy business model. In *Internet and Digital Economics*, ed. E. Brousseau and N. Curien. Cambridge: Cambridge University Press.

Lazonick, W. 2006b. Globalization of the ICT labor force. In *The Oxford Handbook on ICTs*, ed. R. Mansell et al. Oxford: Oxford University Press.

Lazonick, W. 2006c. Innovative enterprise and economic development. In *Business Performance in Twentieth Century: A Comparative Perspective*, ed. Y. Cassis and A. Colli. Cambridge: Cambridge University Press.

Lazonick, W. 2006d. The US stock market and the governance of innovative enterprise. Working paper, INSEAD.

Lazonick, W. and O'sullivan, M. 2000. Maximizing shareholder value: a new ideology for corporate governance. *Economy and Society* 29, 13–35.

New York Times. 1996. *The Downsizing of America*. New York: Times Books.

O'sullivan, M. 2000. The innovative enterprise and corporate governance. *Cambridge Journal of Economics* 24, 393–416.

Pisano, G. 2006. *The Science Business: Strategy, Organization, and Leadership in Biotechnology*. Cambridge, MA: Harvard Business School Press.

Saxenian, A. 2006. *The New Argonauts: Regional Advantage in a Global Economy*. Cambridge, MA: Harvard University Press.

Schumpeter, J. 1934. *The Theory of Economic Development*. Oxford: Oxford University Press.

Schumpeter, J. 1950. *Capitalism, Socialism and Democracy*, 3rd edn. New York: Harper.

Wbur.org. 2006. Sharpening the cutting edge. *On Point*. 9 February. Online. Available at http://www.

onpointradio.org/shows/2006/02/20060209_b_main.asp, accessed 5 September 2006.

Whyte, W. 1956. *The Organization Man*. New York: Simon and Schuster.

How to cite this article

Lazonick, William. "contemporary capitalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000570> doi:10.1057/9780230226203.0304

The New Palgrave Dictionary of Economics Online

contestable markets

Robert D. Willig
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

antitrust enforcement; barriers to entry; contestable markets; increasing returns; integer problem; long-run competitive equilibrium; marginal and average cost pricing; natural monopoly; oligopoly; partial equilibrium; perfect competition; perfectly contestable markets; Ramsey optimum; scale economies; weak invisible hand theorem of natural monopoly

Article

Contestable markets are those in which competitive pressures from potential entrants exercise strong constraints on the behaviour of incumbent suppliers. For a market to be contestable, there must be no significant entry barriers. Then, in order to offer no profitable opportunities for additional entry, an equilibrium configuration of the industry must entail no significant excess profits, and must be efficient in its pricing and in its allocation of production among incumbent suppliers. This is so of a contestable market whether it is populated with only a monopolist or with a large number of actively competing firms, because it is potential competition from potential entrants rather than competition among active suppliers that effectively constrains the equilibrium behaviour of the incumbents.

Perfectly contestable markets (PCMs) are a benchmark for the analysis of industry structure – a benchmark based on an idealized limiting case. Perfectly contestable markets are open to entry by entrepreneurs who face no disadvantages vis-à-vis incumbent firms and who can exit without loss of any costs that entry required to be sunk. The potential entrants have available the same best-practice production technology, the same input markets and the same input prices as those available to the incumbents. There are no legal restrictions on market entry and exit, and there are no special costs that must be borne by an entrant that do not fall on incumbent firms as well. Consumers have no preferences among firms except those arising directly from price or quality differences in firms' offerings.

Potential entrants into perfectly contestable markets are profit-seekers who respond with production to profitable opportunities for entry. They assess the profitability of their marketing plans by making use of the current prices of incumbent firms. Thus, for example, an entrepreneur will enter a market if he anticipates positive profit from undercutting the incumbent's price and serving the entire market demand

at the new lower price. Potential entrants are undeterred by prospects of retaliatory price cuts by incumbents and, instead, are deterred only when the existing market prices leave them no room for profitable entry.

These features of the behaviour of potential entrants are key to the workings of perfectly contestable markets, and they are fully rational only where entry faces no disadvantages and is costlessly reversible. Hence, the benchmark case of perfect contestability excludes the sunk costs, precommitments, asymmetric information and strategic behaviour that characterize many real markets and that are the focus of much current research attention in the field of industrial organization. With irreversibilities and the inducements for strategic behaviour absent, industry structure in PCMs is determined by the fundamental forces of demand and production technology.

Of course, this is also true of perfectly competitive markets. However, this most familiar idealized limiting case is not a satisfactory benchmark for the study of industry structure in general, because it is intrinsically inapplicable to a variety of significant cases. In particular, where increasing returns to scale are present, perfectly competitive behaviour is logically inconsistent with the long-run financial viability of unsubsidized firms.

Perfectly contestable markets can serve in place of perfectly competitive markets as the general standard of comparison for the organization of industry whether or not scale economies are prevalent. Where they are not, perfectly competitive behaviour is necessary for equilibrium in PCMs, and, where scale economies do prevail, equilibrium in PCMs entails behaviour different from that found in perfectly competitive markets but which none the less tends to exhibit desirable welfare properties. In other words, perfect contestability is a generalization of perfect competition that has strong implications in significant circumstances where the latter is inapplicable.

In order to clarify and expand on these ideas, subsequent sections offer analytic outlines of the theory of perfectly contestable markets and applications of the theory to single-product and multi-product industries. Finally, observations are offered on the implications of this theory for the formulation of government policy towards industry.

Perfectly contestable markets: definitions and basic properties

The theory presented here lies in the realm of partial equilibrium. It deals with the provision of the set of products $N=\{1, \ldots, n\}$, some of which may not actually be produced, and which is a proper subset of all the goods in the economy. The prices of these products are represented by vectors $p \in R^n_{++}$, and other prices are assumed to be exogenous and are suppressed in the notation. $Q(p) \in R^n_+$ is the vector-valued market demand function for the products in N, and it suppresses consumers' incomes which are assumed to be exogenous. For any output vector $y \in R^n_+$, $C(y)$ is the cost at exogenously fixed factor prices when production is efficient. The underlying technology is assumed to be freely available to all incumbent firms and all potential entrants. Where necessary, $C(y)$ and $Q(p)$ will be assumed to be differentiable.

Definition 1: A *feasible industry configuration* is composed of m firms producing output vectors $y^1, \ldots, y^m \in R^n_+$, at prices $p \in R^n_{++}$ such that the markets clear, $\Sigma^m_{i=1} y^i = Q(p)$, and that each firm at least breaks even, $p \cdot y^i - C(y^i) \geq 0, i = 1, \ldots, m$.

Thus, the industry configuration is taken to be comprised of *m* firms, where *m* can be any positive integer, so that the industry structure is monopolistic if *m*=1, competitive if *m* is sufficiently large, or oligopolistic for intermediate values of *m*. The term ‘feasibility’ refers to the requirements that each of the firms involved selects a non-negative output vector that permits its production costs, $C(y^i)$ to be covered at the market prices, *p*, and that the sum of the outputs of the *m* firms satisfies market demands at those prices.

Definition 2: A feasible industry configuration over *N*, with prices *p* and firms' outputs $y^1, \ldots, y^m$, is *sustainable* if $p^e \cdot y^e \leq C(y^e)$, for all $p^e \in R^n_{++}$, $y^e \in R^n_+$, $p^e \leq p$, and $y^e \leq Q(p^e)$.

The interpretation of this definition is that a sustainable configuration affords no profitable opportunities for entry by potential entrants who regard incumbents' prices as fixed (for a period sufficiently long to make $C(\cdot)$ the relevant flow cost function for an entrant). Here, a feasible marketing plan of a potential entrant is comprised of prices, p^e, that do not exceed the incumbents' quoted prices, *p*, and a quantity vector, y^e, that does not exceed market demand at the entrant's prices, $Q(p^e)$. The configuration is sustainable if no such marketing plan for an entrant offers a flow of profit, $p^e \cdot y^e - C(y^e)$, that is positive.

Definition 3: A *perfectly contestable market* (PCM) is one in which a necessary condition for an industry configuration to be in equilibrium is that it be sustainable.

A PCM so defined may be interpreted, heuristically, as a market subject to potential entry by firms that have no disadvantage relative to incumbents, and that assess the profitability of entry on the supposition that incumbents' prices are fixed for a sufficiently long period of time. Then, since one requirement for equilibrium is the absence of new entry, an equilibrium configuration in a PCM must offer no inducement for entry; that is, it must be sustainable.

Definition 4: A feasible industry configuration over $N, p, y^1, \ldots, y^m$, is a *long-run competitive equilibrium* if $p \cdot y \leq C(y)\ \forall y \in R^n_+$.

So defined, a long-run competitive equilibrium has precisely the characteristics usually ascribed to it. Together, $p \cdot y^j \geq C(y^j)$ and $p \cdot y \leq C(y),\ \forall y \in R^n_+$, imply that $p \cdot y^j = C(y^j)$ and that the $y^j \in \arg\max_y [p \cdot y - C(y)]$. Thus, each firm in the configuration takes prices as parametric, chooses output to maximize profits, earns zero profit, and equates marginal costs to prices of produced outputs. It is now easy to show

Proposition 1: A long-run competitive equilibrium is a sustainable configuration, so that a perfectly competitive market is a PCM.

Proposition 2: Sustainable configurations need not be long-run competitive equilibria, and a PCM need not be perfectly competitive.

The simplest example sufficient to prove this second proposition is an industry producing a single product with increasing returns to scale over the relevant range of output. Here, the only feasible configuration that is sustainable entails one firm producing the maximal output level y^* given by the intersection of the declining average cost curve with the industry demand curve, and selling at the price p^* given by the corresponding level of average cost. This configuration is sustainable because, at a price equal to or less than p^*, sale of any quantity on or inside the demand curve yields revenue no greater than production cost; in this range, price does not exceed average cost. Yet this configuration is not a

long-run competitive equilibrium, as defined above, because sale of a quantity greater than y^* would earn positive profit if the price could remain at p^*, and because at y^* price exceeds marginal cost which is less than average cost. In fact, in this example there is no possible long-run competitive equilibrium since marginal cost lies below average cost throughout the relevant range of output levels given by demand. In contrast, there is a sustainable configuration.

Hence, Propositions 1 and 2 show that the sustainable industry configuration is a substantive generalization of the long-run competitive equilibrium, and that the PCM is a substantive generalization of the perfectly competitive market. The following propositions summarize some characteristics of equilibria in PCMs.

Proposition 3: Let $p, y^1, \ldots, y^m$ be a sustainable industry configuration. Then each firm must (i) earn zero profit by operating efficiently, $p \cdot y^j - C(y^j) = 0$; (ii) avoid cross-subsidization, $p_S \cdot y^j_S \geq C(y^j) - C(y^j_{N-S}), \ \forall S \subset N$ (where the vector x_T agrees with the vector x in components $j \in T$ and has zeros for its other components); (iii) price at or above marginal cost, $p_j \geq \partial C(y^j) / \partial y_j$.

The interpretation of condition (ii) is that the revenues earned from the sales of any subset of the goods must not fall short of the incremental costs of producing that subset. Otherwise, in view of the equality of total revenues and costs, the revenues collected from the sales of the other goods must exceed their total stand-alone production cost. In PCMs, such pricing invites entry into the markets for the goods providing the subsidy.

Proposition 4: Let $p, y^1, \ldots, y^m$ be a sustainable configuration with $y^k_j < \sum_{h=1}^{m} y^h_j$. Then $p_j = \partial C(y^k) / \partial y_j$. That is, if two or more firms produce a given good in a PCM, they must select input–output vectors at which their marginal costs of producing it are equal to the good's market price.

The implications of this result are surprisingly strong. The discipline of sustainability in perfectly contestable markets forces firms to adopt prices just equal to marginal costs, provided only that they are not monopolists of the products in qst. Conventional wisdom implies that, generally, only perfect competition involving a multitude of firms, each small in its output markets, can be relied upon to provide marginal-cost prices. Here we see that potential competition by prospective entrants, rather than rivalry among incumbent firms, suffices to make marginal-cost pricing a requirement of equilibrium in PCMs, even those containing as few as two active producers of each product. The conventional view holds that the enforcement mechanism of full competitive equilibrium requires the smallness of each active firm in its product market, in addition to freedom of entry. We see that the smallness requirement can be dispensed with, almost entirely, with exclusive reliance on the freedom of entry that characterizes PCMs.

Proposition 5: Let $p, y^1, \ldots, y^m$ be a sustainable configuration. Then, for any $\hat{y}^1, \ldots, \hat{y}^k$ with

$$\sum_{j=1}^{k} \hat{y}^j = \sum_{j=1}^{m} y^j, \ \sum_{j=1}^{k} C(\hat{y}^j) \geq \sum_{j=1}^{m} C(y^j).$$

That is, a sustainable configuration minimizes the total cost to the industry of producing the total industry output.

This proposition is a generalization to PCMs of a well-known result for perfect competition. It can be interpreted as a manifestation of the power of unimpeded potential entry to impose efficiency upon the industry. For example, the proposition implies that if a monopoly occupies a PCM it must be a *natural* monopoly – production by a single firm must minimize industry cost for the given output vector. Thus, Propositions 3, 4 and 5 are powerful tools for the analysis of industry structure in PCMs. Proposition 5 permits information on the properties of production costs to be used to assess the scale and scope of firms' activities in PCMs. Then, Propositions 3 and 4 permit inferences to be drawn about the corresponding equilibrium prices.

PCMS with a single product

This analytic approach leads to very strong results in the single-product case. Propositions 3–5 show that there are only two possible types of sustainable configurations in single-product industries. The first type involves a single firm which charges the lowest price that is consistent with non-negative profit. The firm must be a natural monopoly when it produces the quantity that is demanded at this price. And, in this circumstance, the result maximizes welfare subject to the constraint that all firms in question be viable financially without subsidies. Such a second-best maximum is referred to as a ‘Ramsey optimum’. The second type of sustainable configuration involves production by one or more firms of outputs at which both marginal cost and average cost are equal to price. Here, in the long run, all active firms exhibit the behaviour that characterizes perfectly competitive equilibrium. And, of course, the result involves both (first-best) welfare optimality and financial viability. Hence, in this case, Ramsey optimality and the first-best coincide. This establishes the result that in a single-product industry any sustainable configuration is Ramsey optimal.

However, in general, because of the ‘integer problem’, sustainable configurations may generally not exist. This problem arises, for example, where there is only one output at which a firm's marginal and average costs coincide, and where the quantity of output demanded by the market at the competitive price is greater than this, but is not an integer multiple of that amount. Then, no sustainable configurations exist.

There is, however, a plausible assumption, supported by empirical evidence, at least to some degree, that eliminates the integer problem. Suppose that a firm's average cost curve has a flat-bottom rather than being ‘U’-shaped. In particular, suppose that the minimum level of average cost is attained not only at one output, but (at least) at all outputs between the minimum efficient scale, y_m, and twice the minimum efficient scale. Then any industry output, y^I, that is at least equal to y_m can be apportioned among an integer number of firms, each of which achieves minimum average cost. Specifically, y^1 can be divided evenly among $\lfloor y^1 / y_m \rfloor$ firms (where $\lfloor x \rfloor$ is the largest integer not greater than x) and each firm's output, $y^1 / \lfloor y^1 / y_m \rfloor$, must lie in the (half-open) interval between y_m and $2y_m$. Hence, in this case, the Ramsey optimum can either be a sustainable configuration of two or more firms performing competitively, or a sustainable natural monopoly. Such a monopoly may either produce an output at

which there are increasing returns to scale and it will then price at average cost, or it may produce an output between y_m and $2y_m$ with locally constant returns to scale and adopt a price equal both to average and marginal cost. This, together with the preceding argument, establishes the following result.

Proposition 6: In a single-product industry in which the firm's average cost curve has a flat-bottom between minimum efficient scale and twice minimum efficient scale, a configuration is sustainable if and only if it is Ramsey optimal.

This result shows that, under the conditions described, there is equivalence between welfare optimality and equilibrium in PCMs. This extends the corresponding result for perfectly competitive equilibria to cases of increasing returns to scale. Moreover, since the behavioural assumptions required for a PCM are weaker than those underlying perfectly competitive markets, the equivalence result is more sweeping. In particular, Proposition 6 implies that PCMs can be expected to perform well, whatever the number of firms participating in equilibrium. It is the potential competition of potential entrants, rather than the active competition of existing rivals, that drives equilibrium in PCMs with a single product to welfare optimality.

Multi-product perfectly contestable markets

In industries that produce two or more goods, a rich variety of industry structures become possible, even in PCMs. Here, while the constraints imposed upon incumbents by perfect contestability are not nearly as effective in limiting the range of possible outcomes as they are in single product industries, they nevertheless provide a helpful basis for analysis. In particular, Propositions 3–5 indicate connections among various qualitative properties of multi-product cost functions and various elements of industry structure in PCMs. These connections constitute one theme of this section. The other theme is the normative evaluation of the industry structures that arise in multi-product PCMs.

Before proceeding, it may be useful to provide definitions of some of the multi-product cost properties that are used in the analysis.

Definition 5: Let $P = \{T_1, \ldots, T_k\}$ be a non-trivial partition of $S \subseteq N$. There are (weak) *economies of scope* at y_S with respect to the partition P if $\sum_{i=1}^{k} C(y_{Ti}) > (\geq) C(y_S)$. If no partition is mentioned explicitly, then it is presumed that $T_i = \{i\}$.

Definition 6: The *degree of scale economies* defined over the entire product set, $N = \{1, \ldots, n\}$, at y, is given by $S_N(y) = C(y) / y \cdot \nabla C(y)$.

Returns to scale are said to be increasing, constant or decreasing as S_N is greater than, equal to or less than unity. This occurs as the elasticity of ray average cost with respect to t is negative, positive or zero; where *ray average cost* is $RAC(ty^0) \equiv C(ty^0) / t$.

Definition 7: The *incremental cost* of the product set $T \subseteq N$ at y is given by $IC_T(y) \equiv C(y) - C(y_{N-T})$. The *average incremental cost* of T is $AIC_T(y) \equiv IC_T(y) / \sum_{j \in T} y_j$.

The average incremental cost of T is decreasing, increasing or constant at y if $AIC_T(ty_T + y_{N-T})$ is a decreasing, increasing or locally constant function of t at t=1. These cases are labelled respectively, increasing, decreasing or constant *returns to the scale of the product line T*. The degree of scale economies specific to T is

$$IC_T(y) / \sum_{i \in T} y_i \frac{\partial C(y)}{\partial y_i}.$$

Definition 8: A cost function $C(y)$ is *trans-ray convex* through some point $y^* = (y_1^*, \ldots, y_n^*)$ if there exists at least one vector of positive constants $w_1, \ldots, w_n$ such that for every two output vectors $y^a = (y_1^a, \ldots, y_n^a)$ and $y^b = (y_1^b, \ldots, y_n^b)$ that lie on the hyperplane $\Sigma w_i y_i = w_0$ through point y^*, $C[ky^a + (1-k)y^b] \leq kC(y^a) + (1-k)C(y^b)$ for $k \in (0, 1)$.

In view of the general result that sustainable configurations minimize industry-wide costs (Proposition 5), these cost properties permit inferences to be drawn about industry structure in multi-product PCMs. The first issue that arises is when multi-commodity production is characteristic of equilibrium in a PCM.

Proposition 7: A multi-product firm in a PCM must enjoy (at least weak) economies of scope over the set of goods it produces. When strict economies of scope are present, there must be at least one multi-product firm in any PCM that supplies more than one good.

The second basic question that arises is whether there can be two or more firms actively producing a particular good in a PCM. If there are, then, by Proposition 4, marginal cost pricing must result. The answer depends upon the availability of product-specific scale economies.

Proposition 8: Any product with average incremental costs that decline throughout the relevant range (that is, that offers product-specific increasing returns to scale) must be produced by only a single firm (if it is produced at all) in a PCM. Further, such a product must be priced above marginal cost, unless the degree of product-specific scale economies is exactly one.

Thus, regardless of the presence or absence of economies of scope, globally declining average incremental costs imply that a product must be monopolized in a PCM. It is an immediate corollary that if all goods in the set N exhibit product-specific scale economies, and if there are economies of scope among them all, then the industry is a natural monopoly that must be monopolized in a PCM.

Another route to this result is provided by the ‘weak invisible hand theorem of natural monopoly’.

Proposition 9: Trans-ray convexity of costs together with global economies of scale imply natural monopoly. If, in addition certain other technical conditions are met, a monopoly charging Ramsey-optimal prices is a sustainable configuration.

In general, there may exist natural monopoly situations in which no sustainable prices are possible for the Ramsey optimal product set. Further, even where sustainable prices exist, the Ramsey optimal prices may not be among them. However, under the conditions of the weak invisible hand theorem, the Ramsey optimal prices for the Ramsey optimal product set are guaranteed to be sustainable, so that PCMs are consistent with (second-best) welfare optimal performance by a natural monopoly.

PCMs will yield first-best welfare optimality if there exist sustainable configurations with at least two firms actively producing each good. For in this case Propositions 4 and 5 guarantee industry-wide cost efficiency and marginal-cost pricing of all products. Here, two issues must be resolved: Does industry-wide cost minimization require at least two producers of each good? And if so, do sustainable configurations exist?

The existence problem can be solved in a manner analogous to its solution in the case of single-product industries: by assuming that ray average costs remain at their minimum levels for output vectors that lie (on each ray) between minimum efficient scale and twice minimum efficient scale. And the presence of at least two producers (or one operating in the region where constant returns prevail) of each good is assured if the quantities demanded by the market at the relevant marginal-cost prices are no smaller than minimum efficient scale (along the relevant ray) and if the cost function exhibits trans-ray convexity.

Policy implications of PCMS

One of the principal lessons of the analysis of PCMs is that monopoly does not necessarily entail welfare losses. Rather, the ‘weak invisible hand th’ shows that under certain conditions sustainability and Ramsey optimality are consistent, so that the total of consumers' and producers' surpluses may well be maximized (subject to the constraint that firms be self-supporting) in the equilibrium of a monopoly which operates in contestable markets.

Even stronger results follow from the discussed results that under certain conditions sustainability and a first-best solution are consistent in an oligopoly with a small number of firms. When minimization of industry cost requires that each good be produced by at least two firms, sustainability requires any equilibrium to satisfy the necessary conditions for a first-best allocation of resources. Thus, in these cases, the invisible hand has the same power over oligopoly in perfectly contestable markets that it exercises over a perfectly competitive industry.

This theory suggests that in a market that approximates perfect contestability, the general public interest is well-served by a policy of laissez-faire rather than active regulation by administrative or antitrust means. Small numbers of large firms, vertical and even horizontal mergers and other arrangements which have traditionally been objects of suspicion of monopolistic power, are rendered harmless and perhaps even beneficial by the presence of contestability.

On the other hand, contestability theory does not lend support to the proposition that the unrestrained market automatically solves all economic problems and that virtually all regulation and antitrust activity entails unwarranted and costly intervention. The economy of reality is composed of industries which vary widely in the degree to which they approximate the attributes of perfect contestability. Before the theory of contestability can be legitimately applied to reach a conclusion that intervention is unwarranted in a specific sector, it must first be shown that the sector lies unprotected by entry barriers and that the force of potential entry therefore actively constrains the behaviour of incumbent firms. This then becomes the appropriate first stage in an analysis of efficient government policy towards an industry. Only where the conditions of contestability are found to characterize the reality of an industry can there be validity in applying the normative conclusions of contestability theory concerning the power of potential competition actually to enforce efficient behaviour on incumbents.

Even where contestability is absent in reality, the formulation on efficient regulation can be usefully guided by the theory of contestable markets instead of the theory of perfectly competitive markets. The first-best lesson of the perfect competition model, calling for prices to be set equal to marginal costs, has no doubt contributed to the common regulatory ethos which *equates* price to *some* measure of cost. This doctrine has been used frequently where it is completely inappropriate and without logical foundation, that is, in cases where prices should be based on demand as well as cost considerations, because of the presence of economies of scale and scope. Such arbitrary measures as fully distributed costs cannot

substitute for marginal cost measures as decision rules for proper pricing, and the search for a substitute is a remnant of inappropriate reliance on the model of perfect competition for guidance in regulation. In contrast, contestability theory suggests cost measures that are appropriate guideposts for regulated pricing – incremental and stand-alone costs. The incremental cost of a given service is, of course, the increment in the total costs of the supplying firm when that service is added to its product line. In perfectly contestable markets, the price of a product will lie somewhere between its incremental and its stand-alone cost, just where it falls in that range depending on the state of demand. One cannot legitimately infer that monopoly power is exercised from data showing that prices do not exceed stand-alone costs, and stand-alone costs constitute the proper cost-based ceilings upon prices, preventing both cross-subsidization and the exercise of monopoly power. A simple example will show why this is so. First, suppose that a firm supplies two services, A and B, which *share no costs* and that each costs 10 units a year to supply. The availability of effective potential competition would force revenues from each service to equal 10 units a year. For higher earnings would attract (profitable) entry, and lower revenues would drive the supplier out of business. In this case, in which common costs are absent, incremental and stand-alone costs are equal to each other and to revenues, and the competitive and contestable benchmarks yield the same results.

Next, suppose instead that of the 20 unit total cost 4 are fixed and common to A and B, while 16 are variable, 8 of the 16 being attributable to A and 8 to B. If, because of demand conditions, at most only a bit more than 8 can be generated from consumers of A, then a firm operating and surviving in contestable markets will earn a bit less than 12 from B. These prices lie between incremental costs (8) and stand-alone costs (12), are mutually advantageous to consumers of both services, and will attract no entrants, even in the absence of any entry barriers. In contrast, should the firm attempt to raise the revenues obtained from B above the 12 unit stand-alone cost, it would lose its business to competitors willing to charge less. Similarly, the same fate would befall it in contestable markets if it priced B in a way that earned more than 8 plus the common cost of 4, less the contribution towards that common cost from service A.

Thus, the forces of idealized potential competition in perfectly contestable markets enforce cost constraints on prices, but prices remain sensitive to demands as well. Actual competition and potential competition are *effective* if they constrain rates in this way, and in such circumstances regulatory intervention is completely unwarranted. But if, in fact, market forces are not sufficiently strong, then there may be a proper role for regulation of natural monopoly, and the theoretical guidelines derived from the workings of contestable markets are the appropriate ones to apply. That is, prices must be constrained to lie between incremental and stand-alone costs. (This is the approach recently adopted by the Interstate Commerce Commission to determine maximum rates for US railroad services, and the method has already withstood appeals to the federal courts.)

See Also

- barriers to entry

Bibliography

Baumol, W.J., Panzar, J.C. and Willig, R.D. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich.

Baumol, W.J., Panzar, J.C. and Willig, R.D. 1985. On the theory of perfectly contestable markets. In *New Developments in the Analysis of Market Structure*, ed. J. Stiglitz and F. Mathewson. New York: Harcourt Brace.

Baumol, W.J. and Willig, R.D. 1986. Contestability: developments since the book. *Oxford Economic Papers* 38, 9–36.

How to cite this article

Willig, Robert D. "contestable markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000327> doi:10.1057/9780230226203.0305

The New Palgrave Dictionary of Economics Online

contingent commodities

Zvi Safra

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

contingent commodities; convexity; general equilibrium; risk; risk aversion; uncertainty

Article

The theory of general competitive equilibrium was originally developed for environments where no uncertainty prevailed. Everything was certain and phrases like ‘it might rain’ or ‘the weather might be hot’ were outside the scope of the theory. The idea of *contingent commodity*, that was introduced by Arrow (1953) and further developed by Debreu (1953), was an ingenious device that enabled the theory to be interpreted to cover the case of uncertainty about the availability of resources and about consumption and production possibilities. Basically, the idea of contingent commodity is to add the environmental event in which the commodity is made available to the other specifications of the commodity. With no uncertainty every commodity is specified by its physical characteristics and by the location and date of its availability. It is fairly clear, however, that such a commodity can be considered to be quite different where two different environmental events have been realized. The following examples clarify this: an umbrella at a particular location and at a given date in case of rain is clearly different from the same umbrella at the same location and date when there is no rain; some ice cream when the weather is hot is clearly different from the same ice cream (and at the same location and date) when the weather is cold; finally, the economic role of wheat with specified physical characteristics available at some location and date clearly depends on the precipitation during its growing season. Thus, specifying commodities by both the standard characteristics and the environmental events seems very natural, whereas the role of the adjective in ‘contingent commodities’ is simply to make it clear that one is dealing with commodities the availability of which is contingent on the occurrence of some environmental event. With this specification the model with contingent commodities is very similar to the classical model of general competitive equilibrium and thus questions like the existence of equilibrium and its optimality (with the additional aspect of efficient allocation of risk bearing) are answered in a similar way. Note that, although this model deals with uncertainty, no concept of probabilities is needed for its formal description.

To make things more explicit we look at a simple model with contingent commodities. Assume that, without referring to uncertain events, there are $k \geq 1$ commodities, indexed by *i*, and that there are $n > 1$ mutually exclusive and jointly exhaustive events (or states of nature), indexed by *s*, where *k* and *n* are finite. Thus a contingent commodity is denoted by x_{is} and the total number of these commodities is *kn*, which is greater than *k* but still finite. Consumption and production sets are thus defined as subsets of the *kn*-dimensional Euclidean space, and the economic behaviour of firms and consumers naturally follows from profit maximization (by firms) and utility maximization (by consumers). The price p_{is} of the contingent commodity x_{is} is the number of units of account that have to be paid in order to have the *i*th commodity being delivered at the *s*th event. It is assumed that the market is organized before the realization of the possible events. Thus payment for the contingent commodity x_{is} is done at the beginning while delivery takes place after the realization of events and only in case event *s* has occurred. Note that the price of the (certain) *i*th commodity, that is, the number of units of account that have to be paid in order to have the *i*th commodity *for sure*, is the sum over *s* of the prices p_{is}. For example, assume that the price of one quart of ice cream if the weather is hot is $2.00, the price of one quart of the same ice cream if the weather is cold is $1.00 and that $n = 2$ (either it is hot or cold). Thus the price of having one quart of that ice cream for sure is $\$2.00 + \$1.00 = \$3.00$.

It should be noted that, although the probabilities of the possible events do not explicitly enter the model, the attitude towards risk of both consumers and producers is of interest and does play a significant role in this framework. The preference relations of consumers defined on subsets of the *kn*-dimensional Euclidean space reflect not only their ‘tastes’ but also their subjective beliefs about the likelihoods of different events as well as their attitude towards risk. Convexity of consumers' preferences, for example, is interpreted as risk aversion while, in the same spirit, profit maximization of firms is interpreted as risk neutrality. It should be mentioned that both Arrow and Debreu basically assume expected utility maximizing behaviour, in the sense of the Savage (1954) framework. A more general approach to such preference relations can be found in Yaari (1969), where, again, convexity is taken to mean risk aversion.

A unified and more formal treatment of time and uncertainty using contingent commodities can be found in Debreu (1959, ch. 7). Radner (1968) presents an extension of the above model to the case in which different economic agents have different information.

See Also

- Arrow–Debreu model of general equilibrium
- uncertainty
- uncertainty and general equilibrium

Bibliography

Arrow, K.J. 1953. Le rôle de valeurs boursières pour la répartition la meilleure des risques. *Econométrie*. Paris: CNRS. English translation ‘The role of securities in the optimal allocation of risk-bearing’ in *Review of Economic Studies* (1964); reprinted in K.J. Arrow, *Essays in the Theory of Risk-Bearing*.

Chicago: Markham, 1971.

Debreu, G. 1953. Une économie de l'incertain. Mimeo, Paris: Electricité de France.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36, 31–58.

Savage, L.J. 1954. *Foundations of Statistics*. New York: Wiley.

Yaari, M.E. 1969. Some remarks on measures of risk aversion and their uses. *Journal of Economic Theory* 1, 315–29.

How to cite this article

Safra, Zvi. "contingent commodities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000329> doi:10.1057/9780230226203.0306

The New Palgrave Dictionary of Economics Online

contingent valuation

Trudy Ann Cameron

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Contingent valuation’ methods are used to generate demand data, usually from household surveys, when real markets do not supply reliable revealed preference data about demands for certain types of goods. A number of significant lawsuits have promoted their use in estimating demand for environmental goods. They are also used by transportation economists, health economists and market researchers. Although the degree of acceptance of these methods varies, many economists agree that a value based on stated preferences derived from carefully conceived and executed research is almost certainly preferable to no number at all.

Keywords

conjoint analysis; construct validity assessments; contingent valuation; discrete-choice models; dummy variables; environmental economics; expectations; health economics; household surveys; indirect utility functions; logit models; marginal utility of income; maximum likelihood; neuroeconomics; probit models; random utility models; revealed preference; scope tests; stated preference; willingness to pay

Article

Most economists would agree that no researcher should prefer demand data from hypothetical markets if data concerning the identical goods or services, based on real markets, are readily available. However, there are many situations when even the cleverest empirical economist cannot come up with revealed preference data from actual markets that can be relied upon for information about household demands for some types of goods. Environmental goods are one class of goods where real-market demands sometimes cannot be measured adequately. In the 1980s, environmental economists began in earnest to exploit stated-preference demand information, usually collected using household surveys. This demand information is used primarily to produce utility-theoretic measures of the social benefits of environmental protection measures for benefit–cost analyses.

Environmental economists called these methods ‘contingent valuation methods’ (CVM) because the valuations were elicited ‘contingent upon the conditions described in the survey’. Research that focused on the development and assessment of CVM in environmental economics was well under way by the mid-1980s. However, two events in 1989 thrust the method to the forefront of the field. First, the *Exxon Valdez*, an ocean tanker, ran aground in Prince William Sound in Alaska, spilling 11 million gallons of oil in an environmental disaster that attracted a huge amount of media attention worldwide. Second, just a few months later, the US Court of Appeals held that the economic damages assessed for spills of oil or other hazardous substances could include ‘lost passive use values’, and that these values could legally be measured via CVM.

Plaintiffs and defendants in the Exxon Valdez case thus had a big incentive to advocate and derogate CVM, respectively. For at least a time, the discussion in the literature teetered on the brink of losing its polite academic tone. Given the escalation of the controversy over CVM, the US National Oceanic and Atmospheric Administration (NOAA) convened a panel of experts (untainted by any active role in the Exxon Valdez litigation) to assess CVM. This exercise, by Arrow et al. (1993), produced a set of pronouncements concerning best practices for the conduct of CVM studies. While the 1993 NOAA Panel report cannot be considered the last word on CVM, it was very influential, and there has since been strong pressure on researchers either to conform to the NOAA best practices or to fully justify any departures from them.

As a result of the Exxon Valdez case, much doubt about the reliability of stated preference data led to numerous comparisons of the implications of stated and revealed preference data (for example, Carson et al., 1996). CVM works best when respondents have a clear sense of the consequences of their choices – in terms of both their own budgets and the exact nature of the good that they are being asked to consider paying for – and when they are reasonably familiar with market transactions involving that good. This means that CVM is, unfortunately, most successful when it is least needed. The challenge for researchers is to ensure that demand information gathered using CVM, in less-than-ideal contexts, is as valid and useful as possible.

Myriad biases and qualifications may afflict poorly executed CVM studies. A partial list includes incentive compatibility, hypothetical bias (if the choice is perceived to have absolutely no real consequences), strategic bias (when people try to manipulate the outcome by misrepresenting their preferences), non-response bias (since people cannot be compelled to participate), starting-point bias (for surveys with iterative bids), interviewer bias (for in-person surveys), and information bias (when some portion of the value is constructed during the survey where it did not exist before). Other problems include yea-saying, part–whole bias or embedding, scenario rejection, and the potential for respondents not to pay sufficient attention to their real budget constraints.

Choice formats have been an important issue in the development of CVM. For example, in some early applications of CVM survey respondents were asked directly to identify the single highest dollar amount that they would pay to obtain some change in conditions. These were called open-ended CV questions. Researchers quickly realized that such a task was difficult for consumers who were unfamiliar with naming their own price, especially for goods they may never before have thought much about having to pay for. CVM elicitation techniques evolved fairly quickly to a dichotomous-choice format, where respondents are given a choice between two states of the world. One state is typically the status quo, while the other involves a specified change or set of changes (such as an improvement in environmental quality, or some other rationed public good) that come at a price (typically a lump-sum payment). This

binary choice format was found to fit naturally into a random utility model (RUM) framework that had also become a popular approach to consumer choice problems, both real and hypothetical, in the transportation mode-choice literature and elsewhere in economics.

Respondents' preferences, based on their answers to dichotomous-choice CV questions, can be characterized either in an ad hoc fashion or in a more formal utility-theoretic framework. One standard approach is to specify an indirect utility function shared by all respondents. In its simplest form, the level of indirect utility is assumed to depend on the individual's net income under each of the two alternatives, and upon a discrete indicator of whether there is a change in the rationed public good, or no change, under each alternative. Respondents can choose the environmental improvement programme along with its associated cost (implying lower net income), or decline the environmental improvement programme in order to avoid the cost (preserving their net income). If a respondent prefers the programme with its associated cost, the researcher assumes that the respondent's utility level is higher under that alternative. Equivalently, this means that the *net* indirect utility associated with the programme alternative is positive.

A discrete-choice econometric model, typically involving a binary logit estimator, is used to estimate the sample average marginal utilities of (*a*) net income and (*b*) the discrete bundle of changes represented by the programme in question. It is of course possible to allow for heterogeneity across the sample in these marginal utilities. Most often, heterogeneity is introduced by allowing the otherwise scalar marginal utility associated with going from ‘no programme’ to ‘programme’ to become a systematically varying parameter. Of course, if the identical programme is offered to all respondents, it is not possible to allow this marginal utility to vary with attributes of the programme. However, it can easily be allowed to vary with characteristics of the respondent.

Less commonly, the marginal utility of income is also allowed to vary across respondents, either with the respondent's income (to allow for diminishing marginal utility of income) or with other respondent characteristics. However, there is a premium on simplicity for the marginal utility of income, stemming from the need to use the estimated marginal utility of income parameter(s) to recover demand information. For this reason, many researchers will, if it is justified by the data, prefer a choice model that is linear and additively separable in net income under each alternative.

Linearity and additive separability in income is convenient (when warranted) because the willingness to pay (WTP) function associated with the fitted model is given by the marginal rate of substitution (MRS) between the programme and income. This MRS is given by the ratio of the marginal utility of the programme to the marginal utility of income, producing a result that can be expressed in dollars per ‘unit’ of the programme, where the program indicator is either zero or 1 in the simple binary case. In the non-stochastic case, for a simple dichotomous choice CVM model, this is a single number – ‘WTP for the program’ – if the researcher has assumed homogeneous preferences throughout the sample.

Some extra empirical housekeeping is necessary when it is acknowledged that this point estimate is constructed as the ratio of two estimated quantities, each of which (due to the use of maximum likelihood estimation methods for the logit or probit model) is an asymptotically normally distributed random variable. In theoretical terms, the ratio of two normally distributed random variables has an undefined mean, because zero is a possible value for the denominator. As a practical matter, some researchers use simulation methods to build up a sampling distribution for the estimated WTP. It is possible to use packaged software to make a large number of random draws from the joint distribution of the logit or probit parameters (based on the estimated parameter point estimates and the parameter

variance–covariance matrix). One can then build up a sampling distribution for the needed ratio. Other strategies for dealing with this inconvenience involve estimating the model in ‘WTP-space’ rather than ‘utility-space’ or employing the newer mixed logit (random-parameters logit models) and stipulating that the marginal utility of income parameter be distributed lognormal (since it should be strictly positive, on average), rather than normal, so that the potential divide-by-zero problem goes away.

Over the 1990s contingent valuation researchers in environmental economics gradually made better contact with their counterparts working in other literatures who were confronted by similar problems where there is a lack of market data for products or public goods that need to be valued. In the transportation literature, researchers had grappled early on with the problem of forecasting demand for public transportation projects, or new types of vehicles, that did not yet exist. Researchers began to introduce hypothetical new transportation options which could be characterized in the same terms as existing options (in ‘attribute space’) but which had some attributes that lay well outside the set of existing options on some dimensions, or which involved attributes that were not relevant for existing options (such as travel range or recharge time for prospective electric vehicles). One key difference from contingent valuation was the practice of asking survey respondents to consider more than just ‘the status quo versus a single alternative’. Furthermore, the alternatives were more richly specified. Instead of using simply a dummy variable to indicate whether the policy, programme or public good was present or absent, each alternative was characterized in terms of an array of attributes.

Similar problems were also being addressed in the marketing literature, particularly in the context of ‘pre-test’ marketing. Companies considering whether to develop and introduce new products needed to know in advance about the likely demand for these products, perhaps as a function of alternative possible product configurations. Market researchers developed a set of techniques they called ‘conjoint analysis’. In the marketing literature, the specifications used for the choice models were initially very ad hoc. Little attention was paid to the interpretation of the estimated coefficients as marginal utilities, and simple linear and additively separable specifications were very common. The slope coefficients were known as ‘part-worths’ rather than marginal utilities. However, much was learned about the degree of consistency between planned purchase behaviour and actual purchase behaviour.

CVM has also recently grown in popularity in other sub-disciplines, notably health economics. However, Smith (2003) surveys that literature and suggests that researchers in that field have not yet developed a set of best practices for the use of CVM with the types of choices that are most common in health economics contexts.

In the transportation and marketing literatures, the desired demand information often spanned a number of possible alternative products or services. Stated preference studies were often conducted not just to determine respondents' willingness to pay for a single well-defined good but to understand how willingness to pay might be affected by variations in the mix of attributes making up a prospective good. It was often necessary to anticipate demands for differentiated products, where each product could be characterized as a bundle of attributes and the levels of these attributes differed across alternatives.

In contrast, more of the impetus for CVM non-market valuation research in environmental economics stemmed from a number of significant lawsuits. In the legal context, there is a premium on simplicity in economic modelling so as not to confuse the jury or the judge. It is often best to produce one value for one clearly defined commodity. (Providing a judge or jury with a function that describes demand, where WTP depends upon a wide array of attributes, conditions or respondent attributes, can actually be a liability when attorneys are trying to make a simple, clear and persuasive case. In a legal context, it is

most incisive to value one thing, and to value it as precisely as possible.) Eventually, however, environmental economists began to acknowledge the value of understanding the heterogeneity in demands for environmental goods, since this knowledge can be very helpful to policymakers who wish to consider how different versions of a policy might affect different constituencies.

There are many commonalities between the tasks faced by environmental economists and those faced by transportation economists and market researchers, but there is one key difference. In transportation economics and market research, it is often the case that the public transit system in question will actually be built, or the new product will actually be developed and put on the market. There is an opportunity to go back and see whether the level of demand predicted by the stated preference study actually materializes when there is a real market. In the environmental economics literature, there are typically fewer opportunities to ‘validate’ the stated preference demand information with revealed preferences for the same product.

One common expectation for a good CVM study is now a demonstration that the demand function that has been estimated should ‘walk and talk’ like a demand function. For example, is willingness to pay to preserve big-game hunting opportunities lower, on average, for elderly women than for middle-aged males? Is willingness to pay to preserve air quality higher for people with lung disease or asthma, or for people who have family members with these illnesses? These tests are commonly called ‘construct validity’ assessments. Contingent valuation studies that pass a battery of plausibility tests such as these can generally be viewed as more reliable.

Another common test of contingent valuation estimates that these stated preference demand functions are typically expected to satisfy is something called a ‘scope test’. This means that, on average, respondents' willingness to pay for an alternative that involves more of the ‘good’ in question should be greater than that for an alternative that involves less of the ‘good’. It is of course possible that marginal utility may be positive (as the scope test implies) at low levels of the good, but also that it may go to zero if the quantity of the good is high enough for satiation to set in, and there is no theoretical basis for expecting willingness to pay to be proportional to the amount of the good in question.

CVM data can also sometimes be pooled with actual choice data. This can allow portions of the underlying indirect utility function to be anchored upon real choices, even though the variability in attributes in the real alternatives may not span the full domain relevant to pending policy decisions. The stated choice questions can be used to extend the domain of the estimated demand function.

While economists will remain uncomfortable about reliance upon stated preference information, many now acknowledge that there are circumstances where stated preference data are all that can be collected. In fact, the need for economists to rely upon survey data (what people say as opposed to what they actually do in markets) is now being acknowledged in the other contexts in economics. For example, expectations about future income or life expectancy figure prominently in a number of economic theories. These expectations typically cannot be measured directly, but can sometimes be elicited using surveys and put to good use empirically (see Manski, 2004).

It is worth noting that not just stated preference data but also revealed preference data can be highly variable in its quality. Much revealed preference demand data is also drawn from consumer surveys. It is not always clear that the individual respondent sees the need for accuracy and completeness to be as critical as researchers using the data might hope. In consumer expenditure surveys, for example, interviewers prompt subjects for different types of expenditures, but the enthusiasm and engagement of the survey subject often determines the accuracy and completeness of the data. Rather than viewing

revealed preference data as of unambiguously high quality and stated preference data (such as that produced by CVM) as of unambiguously low quality, it may be prudent simply to acknowledge that both types of data can have their problems.

A partial list of current frontiers in CVM-related research is possible. These frontiers include continuing assessment of (*a*) alternative elicitation formats (there are many candidates beyond the simple NOAA-recommended binary choice format), (*b*) the choice contexts presented to subjects, (*c*) the effects of allowing subjects to express uncertainty about their choices, (*d*) the effects of practice and fatigue when several CVM questions are presented to each respondent, (*e*) integrating stated choices with additional types of real market information, (*f*) how the degree of complexity of the CVM choice scenarios interacts with the cognitive capacity of the subject and/or the subject's inclination to pay attention, and (*g*) the neuroeconomics of real as opposed to stated choices.

Two of the classic books on CVM are Cummings, Brookshire and Schulze (1986) and Mitchell and Carson (1989). Following the Exxon Valdez case, a provocative debate was featured in the *Journal of Economic Perspectives* (Diamond and Hausman, 1994; Hanemann, 1994; Portney, 1994). McFadden (1994) raised some specific concerns about the reliability of CVM data in the context of an empirical application to the existence value of wilderness areas in the western United States. In the intervening years, however, research concerning CVM has continued apace. Helpful expositions and discussions of recent innovations have made their way into textbook form, with one particularly useful summary being provided in Chapter 6 of Freeman (2003). A brief, accessible and very helpful introduction to CVM for non-specialists is contained in Carson (2000). Louviere, Hensher and Swait (2000) offer a comprehensive discussion of stated choice methods broadly defined, including experimental design, modelling, estimation and combining revealed and stated preference data, with illustrations in marketing, transportation and environmental economics. An inventory of the wide range of practical issues to consider in actually implementing a CVM study is provided by Boyle (2003), while Holmes and Adamowicz (2003) update the state of the art for attribute-based (conjoint choice) methods.

There is still considerable variation in individual researchers' levels of comfort with CVM and stated preference data more generally. We might reconsider the question posed by Diamond and Hausman (1994): ‘Contingent valuation – is some number better than no number?’ There are now many economists who would agree that a value based on stated preferences – from a study that is carefully conceived and executed, based on a sufficiently large sample that is representative of its intended population, that has been put through a battery of consistency and validity assessments, and that produces an implied demand function that behaves the way we would expect a ‘real’ demand function to behave – is almost certainly better than no number. This is especially true when ‘no number’ creates the risk that a value of zero would otherwise be imputed, by default, for use in policy decisions.

See Also

- environmental economics

Bibliography

Arrow, K., Solow, R., Portney, P.R., Leamer, E.E., Radner, R. and Schuman, H. 1993. Report of the NOAA panel on contingent valuation. *Federal Register* 58, 4601–14.

Boyle, K.J. 2003. Contingent valuation in practice. In *A Primer on Nonmarket Valuation*, ed. P.A. Champ, K.J. Boyle and T.C. Brown. Boston, MA: Kluwer Academic Publishers.

Carson, R.T. 2000. Contingent valuation: a user's guide. *Environmental Science & Technology* 34, 1413–18.

Carson, R.T., Flores, N.E., Martin, K.M. and Wright, J.L. 1996. Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. *Land Economics* 72, 80–99.

Cummings, R.G., Brookshire, D.S. and Schulze, W.D., eds. 1986. *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Totowa, NJ: Rowman and Allanheld.

Diamond, P.A. and Hausman, J.A. 1994. Contingent valuation – is some number better than no number? *Journal of Economic Perspectives* 8(4), 45–64.

Freeman, A.M.I. 2003. *The Measurement of Environmental and Resource Values: Theory and Methods*. Washington, DC: Resources for the Future.

Hanemann, W.M. 1994. Valuing the environment through contingent valuation. *Journal of Economic Perspectives* 8(4), 19–43.

Holmes, T.P. and Adamowicz, W.L. 2003. Attribute-based methods. In *A Primer on Nonmarket Valuation*, ed. P.A. Champ, K.J. Boyle and T.C. Brown. Boston, MA: Kluwer Academic Publishers.

Louviere, J.J., Hensher, D.A. and Swait, J.D. 2000. *Stated Choice Methods*. New York: Cambridge University Press.

Manski, C.F. 2004. Measuring expectations. *Econometrica* 72, 1329–76.

McFadden, D. 1994. Contingent valuation and social choice. *American Journal of Agricultural Economics* 76, 689–708.

Mitchell, R.C. and Carson, R.T. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington, DC: Resources for the Future.

Portney, P.R. 1994. The contingent valuation debate – why economists should care. *Journal of Economic Perspectives* 8(4), 3–17.

Smith, R.D. 2003. Construction of the contingent valuation market in health care: a critical assessment. *Health Economics* 12, 609–28.

How to cite this article

Cameron, Trudy Ann. "contingent valuation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000535> doi:10.1057/9780230226203.0307

The New Palgrave Dictionary of Economics Online

continuous and discrete time models

Christopher A. Sims
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Most modelling of economic time series works with discrete time, yet time is in fact continuous. While in many instances simple intuitive connections exist between results with discrete time data and the underlying continuous time dynamics, it is possible for discretization to create bias or have unintuitive effects. Some economics literature investigates such distortions. It is also possible to estimate explicitly continuous-time models, using discrete data. This approach raises its own difficulties, but has become more usable as computing power and the techniques to exploit it have improved.

Keywords

approximation theory; continuous and discrete time models; distributed lags; dynamic stochastic general equilibrium models; Granger causal priority; Markov chain Monte Carlo methods; martingales; no-arbitrage models; stochastic differential equations; vector autoregressions; Wiener process

Article

Discrete time models are generally only an approximation, and the error induced by this approximation can under some conditions be important. Most economists recognize that the use of discrete time is only as an approximation, but assume (usually implicitly) that the error of approximation involved is trivially small relative to the other sorts of simplification and approximation inherent in economic theorizing. We consider below first the conditions under which this convenient assumption may be seriously misleading. We discuss briefly how to proceed when the assumption fails and the state of continuous time economic theory.

Approximation theory

Some economic behaviour does involve discrete delays, and most calculated adjustments in individual patterns of behaviour seem to occur following isolated periods of reflection, rather than continually. These notions are sometimes invoked to justify economic theories built on a discrete time scale. But to say that there are elements of discrete delay or time discontinuity in behaviour does not imply that discrete time models are appropriate. A model built in continuous time can include discrete delays and discontinuities. Only if all delays were discrete multiples of a single underlying time unit, and synchronized across agents in the economy, would modelling with a discrete time unit be appropriate.

Nonetheless, sometimes discrete models can avoid extraneous mathematical complexity at little cost in approximation error. It is easy enough to argue that time is in fact continuous and to show that there are in principle cases where use of discrete time models can lead to error. But it is also true in practice that more often than not discrete time models, translated intuitively and informally to give implications for the real continuous time world, are not seriously misleading. The analytical task, still not fully executed in the literature, is to understand why discrete modelling usually is adequate and thereby to understand the special circumstances under which it can be misleading.

The basis for the usual presumption is that, when the time unit is small relative to the rate at which variables in a model vary, discrete time models can ordinarily provide good approximations to continuous time models. Consider the case, examined in detail in Geweke (1978), of a dynamic multivariate distributed leg regression model, in discrete time.

$$Y(t) = A^{*}X(t) + U(t), \tag{1}$$

where * stands for convolution, so that

$$A^{*}X(t) = \sum_{s=-\infty}^{\infty} A(s)X(t-s).$$

(2)

We specify that the disturbances are uncorrelated with the independent variable vector *X*, that is, $\mathrm{cov}[X(t), U(s)] = 0$, all *t*, *s*. The natural assumption is that, if approximation error from use of discrete time is to be small, *A*(*s*) must be smooth as a function of *s*, and that in this case (1) is a good approximation to a model of the form

$$y(t) = a^{*}x(t) + u(t)$$
(3)

where

$$a^{*}x(t) = \int_{-\infty}^{\infty} a(s)x(t-s)\,ds$$
(4)

and *y*, *a* and *x* are functions of a continuous time parameter and satisfy $y(t) = Y(t)$, $x(t) = X(t)$ and $a(t) = A(t)$ at integer *t*. In this continuous time model we specify, paralleling the stochastic identifying assumption in discrete time, $\mathrm{cov}[x(t), u(s)] = 0$, all *t*, *s*. If the discrete model (2) corresponds in this way to a continuous time model, the distributed lag coefficient matrices *A*(*s*) are uniquely determined by *a* and the serial correlation properties of *x*.

We should note here that, though this framework seems to apply only to the case where *X* is a simple discrete sampling of *x*, not to the time-averaged case where *X*(*t*) is the integral of *x*(*s*) from *t*–1 to *t*, in fact both cases are covered. We can simply redefine the *x* process to be the continuously unit-averaged version of the original *x* process. This redefinition does have some effect on the nature of limiting results as the time unit goes to zero (since the unit-averaging transformation is different at each time unit) but turns out to be qualitatively of minor importance. Roughly speaking, sampling a unit-averaged process is like sampling a process whose paths have derivatives of one higher order than the unaveraged process.

Geweke shows that under rather general conditions

$$\sum_{s=-\infty}^{\infty} \|A(s) - \tau a(s\tau)\|^{2} \to 0$$
(5)

as the time unit τ goes to zero, where $\|\ \|$ is the usual root-sum-of-squared-elements norm. In this result, the continuous time process *x* and lag distribution *a* are held fixed while the time interval corresponding to the unit in the discrete time model shrinks.

This is the precise sense in which the intuition that discrete approximation does not matter much is correct. But there are important limitations on the result. Most obviously, the result depends on *a* in (3) being an ordinary function. In continuous time, well-behaved distributed lag relations like (3) are not the only possible dynamic relation between two series. For example, if one replaces (3) by

$$y(t) = \alpha(\mathrm{d}/\mathrm{d}t)x(t) + u(t),$$
(6)

then the limit of *A* in (2) is different for different continuous *x* processes. In a univariate model with second-order Markov *x* (for example, one with $\mathrm{cov}[x(t), x(t-s)] = (1+\theta|s|)e^{-\theta|s|}\,\mathrm{var}[x(t)]$, the limiting discrete time model, as τ goes to zero, is

$$y(t) = a\{-0.02X(t+4) + 0.06X(t+3) - 0.22X(t+2) + 0.80X(t+1) - 0.80X(t-1) + 0.22X(t-2) - 0.06X(t-3) + 0.02X(t-4)\} + U(t)$$
(7)

(see Sims, 1971).

This result is not as strange as it may look. The coefficients on *X* sum to zero and are anti-symmetric about zero. Nonetheless, (7) is far from the naive approximation which simply replaces the derivative operator with the first difference operator. In fact, if the estimation equation were constrained to involve only positive lags of *X*, the limiting form would be

$$Y(t) = \alpha\{1.27X(t) - 1.161X(t-1) + 0.43X(t-2)\} - 0.12X(t-3) + 0.03X(t-4) - 0.01X(t-5) + U(t). \quad (8)$$

The naive approximation of (3) by $Y(t) = \alpha[X(t) - X(t-1)] + U(t)$ is valid only in the sense that, if this form is imposed on the discrete model a priori, the least squares estimate of α will converge to its true value. If the resulting estimated model is tested for fit against (8) or (7), it will be rejected.

Although the underlying model involves only the contemporaneous derivative of *x*, (8) and (7) both involve fairly long lags in *X*. If *x* paths have higher than first-order derivatives (for example, if they are generated by a third-order stochastic differential equation) the lag distributions in (8) and (7) are replaced by still higher-order limiting forms. Thus, different continuous time processes for *x* which all imply differentiable time paths produce different limiting discrete *A*. Here the fact that the time unit becomes small relative to the rate of variation in *x* does not justify the assumption that approximation of continuous by discrete models is innocuous. In particular, the notion that discrete differencing can approximate derivatives is potentially misleading.

It should not be surprising that the discrete time models may not do well in approximating a continuous time model in which derivatives appear. Nonetheless, empirical and theoretical work which ignores this point is surprisingly common.

If *a* is an ordinary function, there is still chance for error despite Geweke's result. His result implies only that the mean square deviation of *a* from *A* is small. This does not require that individual $A(t/\tau)$'s converge to the corresponding $a(t)$ values. For example, in a model where *x* is univariate and $a(t) = 0, t < 0, a(0) = 1$, $a(s)$ continuous on $[0,\infty]$, the limiting value for $A(0)$ is 0.5, not 1.0. Thus, if $a(t) = e^{-\beta t}$ on $[0,\infty)$, making *a* monotone decreasing over that range, $A(t)$ will not be monotone decreasing. It will instead rise between $t = 0$ and $t = 1$. This is not unreasonable on reflection: the discrete lag distribution gives a value at $t = 0$ which averages the continuous time distribution's behaviour on either side of $t = 0$. It should therefore not be surprising that monotonicity of *a* does not necessarily imply monotonicity of *A*, but the point is ignored in some economic research.

Another example of possible confusion arises from the fact that, if the *x* process has differentiable paths, $a(t) = 0$ for $t < 0$ does not imply $A(t) = 0$ for $t < 0$. The mean-square approximation result implies that when the time unit is small the sum of squares of coefficients on $X(t-s)$ for negative *s* must be small relative to the sum of squares on $X(t-s)$ for positive *s*, but the first few lead coefficients will generally be non-zero and will not go to zero as the time interval goes to zero. This would lead to mistaken conclusions about Granger causal priority in large samples, if significance tests were applied naively.

Geweke's exploration of multivariate models shows that the possibilities for confusing results are more numerous and subtle in that case. In particular, there are ways by which poor approximation of $\alpha_j(s)$ by $A_j(s/\tau)$ in some *s* interval (for example, around $s = 0$) can lead to contamination of the estimates of other elements of the *A* matrix, even though they correspond to x_j's and a_j's that in a univariate model would not raise difficulties.

In estimation of a dynamic prediction model for a single vector *y*, such as a vector autoregression (VAR) or dynamic stochastic general equilibrium model (DSGE), the question for approximation theory becomes whether the continuous time dynamics for *y*, summarized in a Wold moving average representation

$$y(t) = a^* u(t) \quad (9)$$

has an intuitively transparented connection to the corresponding discrete time Wold representation

$$Y(t) = A^* U(t). \quad (10)$$

In discrete time the $U(t)$ of the Wold representation is the one-step-ahead prediction error, and in continuous time $u(t)$ also represents new information about *y* arriving at *t*. There are two related sub-questions. Is the *A* function the same shape as the *a* function; and is the *U* vector related in a natural way to the *u* vector? The *u* vector is a continuous time white noise, so that *U* cannot possibly be a simple discrete sampling of *u*.

If *y* is stationary and has an autoregressive representation, then $U(t) = A^{-1*} a^* u_t$, with the expression interpreted as convolution in continuous time, but with A^{-1} putting discrete weight on integers. The operator connecting *U* and *u* is then $A^{-1*}a$. There are cases where the connection between

continuous and discrete time representations is intuitive. For example, if $a(s) = \exp(-Bs)$ (with the exponentiation interpreted as a matrix exponential in a multivariate case), then

$$U(t) = \int_0^1 e^{-Bs} u(t-s)\,ds \tag{11}$$

and $A(s) = a(s)$ at integers. This is a more intuitive and precise matching than in any case we examined above for projection of one variable on another. If $a(0)$ is full rank and right-continuous at zero and if $a(s)$ is differentiable at all $s > 0$, then a similar intuitively simple matching of A to a arises when the time unit is small enough.

However, non-singularity of $a(0)$ rules out differentiability of time paths for y. When time paths for y, or some elements of it, are differentiable, no simple intuitive matching between A and a arises as the time unit shrinks.

There is one clear pattern in the difference in shape between A and a that stands in contrast to the case of distributed lag projection considered above. If both the continuous time and the discrete time moving average representations are fundamental, then by definition the one-step-ahead prediction error in $y(t)$ based on $Y(t-s),\ s \geq 1$ is

$$\int_0^1 a(s)\,u(t-s)\,ds, \tag{12}$$

while the one-step-ahead prediction error in $Y(t)$ based on $Y(t-s),\ s = 1, 2, \ldots$ is $A_0 Y(t)$. Now the information set we use in forecasting based on the past of Y at integer values alone is smaller than the information set based on all past values of y, so the one-step-ahead error based on the discrete data alone must be larger. If we normalize in the natural way to give U an identity covariance matrix and to make $\mathrm{var}(g^* u(t)) = \int g(s) g'(s)$ (so u is a unit white noise vector), then it must emerge that

$$A_0 A_0 \geq \int_0^1 a(s)\,a(s)'\,ds, \tag{13}$$

where the inequality is interpreted as meaning that the left-hand-side matrix minus the right-hand-side matrix is positive semi-definite. In other words, the initial coefficient in the discrete MAR will always be as big or bigger than the average over (0,1) of the coefficients in the continuous MAR. This tendency of the discrete MAR to seem to have a bigger instant response to innovations is proportionately larger the smoother a is near zero.

More detailed discussion of these points, together with numerous examples, appears in Marcet (1991).

Estimation and continuous time modelling

How can one proceed if one has a model like, say, (6), to which a discrete time model is clearly not a good approximation? The only possibility is to introduce explicitly a model for how x behaves between discrete time intervals, estimating this jointly with (6) from the available data. Doing so converts (6) from a single-equation to a multiple-equation model. That is, the device of treating x as 'given' and non-stochastic cannot work because an important part of the error term in the discrete model arises from the error in approximating $a^* x$ by $A^* X$. Furthermore, because separating the approximation error component of U from the component due to u is essential, one would have to model serial correlation in u explicitly. The model could take the form

$$\begin{bmatrix} y(t) \\ x(t) \end{bmatrix} = \begin{bmatrix} c(s) & a^* b(s) \\ 0 & b(s) \end{bmatrix} * \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}, \tag{14}$$

where w and v are white noise processes fundamental (in the terminology of Rosanov, 1967), for y and x. To give b and c a convenient parametric

form, one might suppose them rational, so that (14) can be written as a differential equation, that is,

$$P(D)\,y(t) = P(D)\,a^{*}\,x(t) + w(t) \tag{15}$$

$$Q(D)\,x(t) = v(t), \tag{16}$$

where P and Q are finite-order polynomials in the derivative operator, $Q^{-1}(D)\,v = b^{*}v$, and $P^{-1}(D)\,w = c^{*}w$.

A discrete time model derived explicitly from a continuous time model is likely to be nonlinear at least in parameters and therefore to be more difficult to handle than a more naive discrete model. However with modern computing power, such models are usable. Bergstrom (1983) provides a discussion of estimating continuous time constant coefficient linear stochastic differential equation systems from discrete data, the papers in the book (1976) he edited provide related discussions, and Hansen and Sargent (1991), in some of their own chapters of that book, discuss estimation of continuous time rational expectations models from discrete data.

Estimating stochastic differential equation models from discrete data has recently become easier with the development of Bayesian Markov chain Monte Carlo (MCMC) methods. Though implementation details vary across models, the basic idea is to approximate the diffusion equation

$$dy_t = a(y_t)\,dt + b(y_t)\,dW_t, \tag{17}$$

where W_t is a Wiener process, by

$$y_t = e^{-a(y_{t-\delta})}\,y_{t-\delta} + b(y_{t-\delta})\,\varepsilon_t. \tag{18}$$

Such an approximation can be quite inaccurate unless δ is very small. But one can in fact choose δ very small, much smaller than the time interval at which data are observed. The values of y_t at times between observations are of course unknown, but if they are simply treated as unknown ‘parameters it may be straightforward to sample from the joint posterior distribution of the y's at non-observation times and the unknown parameters of the model. The Gibbs sampling version of MCMC samples alternately from conditional posterior distributions of blocks of parameters. Here, sampling from the distribution of y at non-observation dates conditioning on the values of model parameters is likely to be easy. If the model has a tractable form, it will also be easy to sample from the posterior distribution of the parameters conditional on all the y values, both observed and unobserved. Application of these general ideas to a variety of financial models is discussed in Johannes and Polson (2006).

Another approach that has become feasible with increased computing power is to develop numerical approximations to the distribution of $y_{t+\delta}$ conditional on data through time t. Aït-Sahalia (2007) surveys methods based on this approach.

Modelling in continuous time does not avoid the complexities of connecting discrete time data to continuous time reality – it only allows us to confront them directly. One reason this is so seldom done despite its technical feasibility is that it forces us to confront the weakness of economic theory in continuous time. A model like (15)–(16) makes an assertion about how many times y and x are differentiable, and a mistake in that assertion can result in error as bad as the mistake of ignoring the time aggregation problem. Economic theory does not have much to say about the degree of differentiability of most aggregate macroeconomic time series. When the theory underlying the model has no believable restrictions to place on fine-grained dynamics, it may be better to begin the modelling effort in discrete time. As is often true when models are in some respect under-identified, it is likely to be easier to begin from a normalized reduced form (in the case the discrete time model) in exploring the range of possible interpretations generated by different potential identifying assumptions.

Recent developments in financial economics have produced one area where there are continuous time economic theories with a solid foundation. Stochastic differential equations (SDEs) provide a convenient and practically useful framework for modelling asset prices. These SDE models imply non-differentiable time paths for prices, and it is known (Harrison, Pitbladdo and Schaefer, 1984) that differentiable time paths for asset prices would imply arbitrage opportunities, if there were no transactions costs or bounds on the frequency of transactions.

However, there are in fact transactions costs and bounds on transactions frequencies, and no-arbitrage models for asset prices break down at very fine,

minute-by-minute, time scales. Successful behavioural modelling of these fine time scales requires a good theory of micro-market structure, which is still work in progress.

It is worthwile noting that a process can have non-differentiable paths without producing white noise residuals at any integer order of differentiation: for example, a model satisfying (3) with $a(s) = s^{0.5}e^{-s}$. Such a process has continuous paths with unbounded variation and is not a semimartingale. That is, it is not the sum of a martingale and a process with bounded variation, and therefore cannot be generated from an integer-order SDE. Similarly, if $a(s) = s^{-0.5}e^{-s}$, the process has non-differentiable paths but is nonetheless not a semimartingale. The existence of such non-semimartingale processes and their possible applications to financial modelling is discussed in Sims and Maheswaran (1993).

See Also

- time series analysis

Bibliography

Aït-Sahalia, Y. 2007. Estimating continuous-time models using discretely sampled data. In *Advances in Economics and Econometrics, Theory and Applications. Ninth World Congress*, vol. 3, ed. R. Blundell, T. Persson and W.K. Newey. Cambridge: Cambridge University Press.

Bergstrom, A.R., ed. 1976. *Statistical Inference in Continuous Time Economic Models*. Amsterdam: North-Holland.

Bergstrom, A.R. 1983. Gaussian estimation of structural parameters in higher order continuous time dynamic models. *Econometrica* 51, 117–52.

Geweke, J. 1978. Temporal aggregation in the multiple regression model. *Econometrica* 46, 643–62.

Hansen, L.P. and Sargent, T.J., eds. 1991. *Rational Expectations Econometrics*. Boulder and Oxford: Westview Press.

Harrison, J.M., Pitbladdo, R. and Schaefer, S.M. 1984. Continuous price processes in frictionless markets have infinite variation. *Journal of Business* 57, 353–65.

Johannes, M. and Polson, N. 2006. MCMC methods for continuous-time financial econometrics. In *Handbook of Financial Econometrics*, ed. Y. Aït-Sahalia and L.P. Hansen. Amsterdam: North-Holland.

Marcet, A. 1991. Temporal aggregation of economic time series. In *Rational Expectations Econometrics*, ed. L.P. Hansen and T.J. Sargent. Boulder and Oxford: Westview Press.

Rozanov, Yu.A. 1967. *Stationary Random Processes*, trans A. Feinstein. San Francisco, Cambridge, London, Amsterdam: Holden-Day.

Sims, C.A. 1971. Approximate specifications in distributed lag models. In Proceedings of the 38th Session, *Bulletin of the International Statistical Institute* 44, Book 1.

Sims, C.A. and Maheswaran, S. 1993. Empirical implications of arbitrage-free asset markets. In *Models, Methods and Applications of Econometrics*, ed. P.C.B. Phillips. Oxford: Blackwell.

How to cite this article

Sims, Christopher A. "continuous and discrete time models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000332> doi:10.1057/9780230226203.0308

The New Palgrave Dictionary of Economics Online

contract theory

David Martimort
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article offers a brief overview of contract. It focuses on the theory of complete contracts and the three associated paradigms of adverse selection, moral hazard and non-verifiability. By showing difficulties in allocating resources between asymmetrically informed partners, contract theory has deeply changed our view of the functioning of organizations and markets.

Keywords

adverse selection; asymmetrical information; Bayesian-Nash equilibrium; collusion; contract theory; cost observability; free-rider problem; incentive compatibility; incentive constraints; incomplete contracts; informativeness principle; insurance; Laffont, J.-J.; limited liability; monotonicity; moral hazard; multi-agent organizations; non-verifiability; optimal contract; Pontryagyn principle of optimality; principal and agent; revelation principle; risk aversion; risk neutrality; sharecropping; Spence–Mirrlees condition; tournaments

Article

As with so many major concepts in economics, contract theory was introduced by Adam Smith who, in his monumental *Wealth of Nations* (1776, book III, ch. 2), considered the relationship between peasants and farmers through this lens. For instance, he pointed out the perverse incentives provided by sharecropping contracts, widespread in 18th-century Europe. However, it is fair to say that the issues of incentives and contract theory were largely ignored by economists until the end of the 20th century. By then, the focus of economic theory was on the working of markets and price formation. Firms were viewed only as production technologies, and the issue of the separation between ownership and control was most often put aside. This black-box approach was, of course, quite unsatisfactory. At the turn of the 1970s, with the methodological revolution of game theory, more emphasis was placed on strategic interactions between a small number of players in a world where informational problems matter. From

this new perspective, the allocation of resources is no longer ruled by the price system but by *contracts* between asymmetrically informed partners. Contract theory has deeply changed our view of the functioning of organizations and markets.

This article aims to provide a brief overview of contract theory, stressing a few major insights and illustrating them with useful applications. Due to space constraints, it does not do justice to several aspects of contract theory, and will mostly reflect my own tastes in the field. In particular, I focus on the so-called *theory of complete contracts*, leaving aside the burgeoning theory of incomplete contracts which is covered elsewhere in this dictionary. Successive sections deal respectively, with adverse selection, moral hazard and non-verifiability: the three different paradigms which have been used in the field of complete contract theory. Since the distinction between complete and incomplete contracts is easier to draw once these notions have already been explained, I will postpone such discussion to the end of the article.

Adverse selection

Consider the following buyer–seller relationship as the archetypical example of contractual relationship between a principal (the buyer) and his agent (the seller) who produces some good or service on his behalf. The mere delegation of this task to the agent gives the agent access to private information about the technology. This *adverse selection* environment is captured by assuming that a technological parameter θ is known only by the agent. It is drawn from a distribution in an exogenous type space Θ which is common knowledge. Neither the principal nor a court of law observes this parameter. Contracts cannot specify outputs and prices as a function of the realized state of nature.

The buyer enjoys a net benefit $S(\theta, q) - t$ when buying *q* units of output at a price *t*. The seller enjoys a profit $t - C(\theta, q)$ from producing that good. We will assume that these functions are concave in *q*. Notice that the state of nature θ might affect both the agent's and the principal's utility functions. This can, for instance, be the case if this parameter also determines the quality of the good to be traded. Under complete information, efficiency requires that the buyer and the seller trade the first-best quantity $q^*(\theta)$ such that the buyer's marginal benefit from consumption equals the seller's marginal cost of production:

$$\frac{\partial S}{\partial q}(\theta, q^*(\theta)) = \frac{\partial C}{\partial q}(\theta, q^*(\theta)). \tag{1}$$

Many mechanisms or institutions lead to this outcome. Both the price mechanism and a take-it-or-leave-it offer by one party to the other would achieve the same allocation, although with different distributions of the surplus between the traders. If the principal retains all bargaining power (for instance, because there is a competitive fringe of potential sellers), he could offer a forcing contract stipulating an output $q^*(\theta)$ and a transfer $t^*(\theta)$ which just covers the seller's cost. This forcing contract maximizes the

buyer's net gains from trade and leaves the seller just indifferent between participating or not.

In what follows, we mostly focus on the case where the uninformed principal has full bargaining power in contracting. In this framework, the *contract* between the buyer and the seller does not only have the allocative and distributive roles it has under complete information. It also has the role of *communicating information* from the informed party to the uninformed party. This communication role suggests that the informed party should be given a choice among different options and that this choice should reveal information about the adverse selection parameter.

A first step in the analysis consists of describing the set of allocations which are feasible under asymmetric information. The basic tool for doing so is the revelation principle (see Gibbard, 1973; Green and Laffont, 1977; Dasgupta, Hammond and Maskin, 1979; and Myerson, 1979, among others), which states that there is no loss of generality in restricting the analysis to *revelation mechanisms* that are *direct*, that is, of the form $\{t(\hat{\theta}), q(\hat{\theta})\}_{\hat{\theta}\in\Theta}$ with $\hat{\theta}$ a message ('report') sent by the informed seller to the uninformed buyer, and *truthful*, that is, such that the agent finds it optimal to report his true type. Therefore, incentive feasible contracts satisfy the following *incentive* constraints

$$t(\theta) - C(\theta, q(\theta)) \geq t(\hat{\theta}) - C(\theta, q(\hat{\theta})) \ \forall (\theta, \hat{\theta}) \in \Theta^2. \tag{2}$$

To be acceptable, a contract must also satisfy the seller's participation constraints

$$t(\theta) - C(\theta, q(\theta)) \geq 0 \qquad \forall \theta \in \Theta \tag{3}$$

which ensure that, irrespective of his type, the agent by contracting gets at least his reservation payoff (exogenously normalized to zero).

Once the set of incentive feasible allocations is described, the analysis may proceed further. Keeping in mind that the uninformed buyer designs his offer under asymmetric information, we might characterize an *optimal contract*. Such a contract maximizes the uninformed buyer's expected net surplus subject to the feasibility constraints (2) and (3).

Much of the theoretical literature developed over the 1980s and early 1990s has investigated the structure of the set of incentive feasible allocations and its consequences for optimal contracting. A key property is the so-called Spence–Mirrlees condition (see Spence, 1973; 1974; and Mirrlees, 1971) for early contributions which put forward that condition). This condition is satisfied when the slope of the agent's indifference curves can be ranked with respect to his type. In our example, this condition holds when $\frac{\partial^2 C}{\partial \theta \partial q} > 0$, that is, when higher types also have higher marginal costs and should thus produce

less. Therefore, the monotonicity condition

$$q(\theta) \geq q(\theta') \text{ for } \theta \leq \theta' \tag{4}$$

is a direct consequence of the incentive constraints. The Spence–Mirrlees condition can be viewed as a regularity assumption making the incentive problem well-behaved. It ensures that only incentive constraints between ‘nearby’ types matter in the optimization. Intuitively, this means that the seller with a given marginal cost may be tempted to overstate slightly its costs, receiving the higher transfer targeted to less efficient types but producing at a lower marginal cost. By so doing, this more efficient type receives an information rent. Once these local constraints are taken into account and when the Spence–Mirrlees condition holds, the incentives to mimic more distant types are no longer relevant. With this reduction of the set of relevant incentive constraints, the principal's optimization problem is significantly simplified.

The result of this optimization is straightforward. Inducing information revelation by the most efficient types requires giving up an *information rent* to those types. The basic intuition of most adverse-selection models is that reducing this rent requires production to be distorted. For instance, when efficient types want to mimic less efficient ones, the latter's allocation should be made less attractive. This is obtained by distorting their production downward and modifying transfers accordingly.

To see more formally the nature of the output distortion, consider the case where types are distributed over a compact set $[\underline{\theta}, \overline{\theta}]$ according to the cumulative distribution function $F(\cdot)$ (with a positive density $f(\cdot)$). The second-best optimal output $q^{SB}(\theta)$ under adverse selection is the solution to:

$$\frac{\partial S}{\partial q}(\theta, q^{SB}(\theta)) = \frac{\partial C}{\partial q}(\theta, q^{SB}(\theta)) + \frac{F(\theta)}{f(\theta)} \frac{\partial^2 C}{\partial q \partial \theta}(\theta, q^{SB}(\theta)). \tag{5}$$

Condition (5) states that, for any type θ, the buyer's marginal benefit must equal the seller's *marginal virtual cost* (see Laffont and Martimort, 2002, chs 2 and 3, for details). The virtual cost of a given type takes into account not only its cost of production but also the cost of deterring other types (here more efficient types) from mimicking that type. The allocation is no longer efficient, as under complete information, but *interim efficient* in the sense of Holmström and Myerson (1983).

Condition (5) is crucial, and is found in various forms in any adverse-selection model. It states that, under asymmetric information, there is a fundamental trade-off between implementing allocations close to efficiency and giving information rents to the most efficient types to induce information revelation. This trade-off calls for distortions away from efficiency.

Provided that the output schedule defined by (5) satisfies the monotonicity condition (4), this is the exact solution of our problem. To guarantee monotonicity, on top of assumptions on the concavity of $S(\cdot)$ and $\frac{\partial S}{\partial \theta}(\cdot)$, convexity of $C(\cdot)$ and $\frac{\partial C}{\partial \theta}(\cdot)$, $\frac{\partial^2 C}{\partial \theta \partial q}(\cdot) > 0$, $\frac{\partial^3 C}{\partial \theta^2 \partial q}(\cdot) > 0$ and $\frac{\partial^2 S}{\partial \theta \partial q}(\cdot) < 0$, one needs also to impose a property on the type distribution, the so-called *monotonicity of the hazard rate* $\frac{F(\theta)}{f(\theta)}$ (see Bagnoli and Bergstrom, 2005). Otherwise, the optimal contract may entail some area of pooling such that all types belonging to a set with positive measure produce the same amount and are paid the same price. The optimal solution may then be obtained using ‘ironing techniques’ (see for instance Guesnerie and Laffont, 1984).

Direct extensions

Adverse-selection methodology has been successfully extended in various directions allowing for multidimensional types (Armstrong and Rochet, 1999), and/or multiple outputs (Laffont and Tirole, 1993, ch. 3), and type-dependent reservation utilities (Lewis and Sappington, 1989; Jullien, 2000). There, the analysis is substantially more complex as types can no longer be ranked as easily as in the model sketched above. The Spence–Mirrlees condition might fail to hold and global incentive constraints may bind, leading to pooling allocations being optimal. Another interesting extension is the case of hidden knowledge, in which contracting takes place before the agent becomes informed. The logic of such models is very close to that we discuss below in the section on moral hazard. In a nutshell, the trade-off between allocative efficiency and rent extraction is now replaced by the trade-off between insuring the agent against shocks on costs and inducing him to reveal his cost once it is known. Output distortions still arise (see Laffont and Martimort, 2002, ch. 2, for details). Others have endogenized the asymmetric information structure and examined the incentives to learn about the unknown parameter (see, for instance, Crémer, Khalil and Rochet, 1998). Finally, there exists a literature that considers the case where the principal is the informed party (Maskin and Tirole, 1990; 1992). New difficulties arise from the fact that the mere offer of the contract may signal information.

Multi-agent organizations

The most important extensions of the adverse selection paradigm certainly concern multi-agent organizations. Such complex organizations emerge because of the need to share common resources, produce public goods, internalize production externalities or enjoy information economies of scale. Although any such reason calls for a specific analysis, a few common themes of the literature can be highlighted by remaining at a rather general level.

Regarding the implementation concept, different notions of incentive compatibility may be used depending on the context. First, agents may know each other's types and play a Nash equilibrium of the direct revelation mechanism offered by the principal (see Maskin, 1999, and the discussion of the non-verifiability paradigm below). Second, agents may only know their own type, form beliefs on each others' types and play a Bayesian–Nash equilibrium (see D'Aspremont and Gérard-Varet, 1979). Third,

one may also insist on dominant strategy implementation because it does not depend on the specification of beliefs (see Gibbard, 1973; Groves, 1973; Green and Laffont, 1977). To each implementation concept corresponds a notion of incentive feasibility. Once the set of incentive feasible contracts is defined, one can proceed to optimization. It is a trivial observation that, the more restrictive the implementation concept, the lower is the principal's payoff at the optimum.

In some cases, such as the provision of public goods within a society of privately informed agents or in bargaining models between a buyer and a seller with equal bargaining power, the goal is no longer to design a multilateral contract which would extract the rents of all agents but, instead, to maximize some *ex ante* efficiency criterion under incentive constraints. Groves (1973) showed that dominant strategy mechanisms suffice to implement the first-best decision in a public good context. One caveat is that the budget generally fails to be balanced. D'Aspremont and Gérard-Varet (1979) proposed a Bayesian incentive-compatible mechanism which implements the first-best and still satisfies budget balance. As argued by Laffont and Maskin (1979), such a mechanism may conflict with the agents' participation constraint. In a bargaining environment, Myerson and Satterthwaite (1983) showed in a similar vein that there exists no Bayesian bargaining mechanism that is efficient, budget-balance and individually rational.

The optimal multilateral contract can be very sensitive to the information structure. In environments where risk-neutral agents have correlated types but know only their own type, the principal can condition one agent's compensation on another's report. By doing so, the principal can fully extract the rent from both agents in a Bayesian-Nash equilibrium. One may view this result as a strong rationale for relative performance evaluation, yardstick competition, benchmarking and internalization of similar activities within the same organization. This puzzling insight of Crémer and McLean (1988) no longer holds when one introduces risk-aversion, *ex post* participation constraints or limited liability constraints. These assumptions reintroduce information rents in the multi-agent organization, and the standard trade-off between efficiency and rent extraction reappears.

When the agents' types are independently distributed, yardstick competition is ineffective and the agents derive information rents. However, the externality that one agent's task may exert on another can shape the distribution of these rents. In competitive environments, such as procurement auctions among sellers, it is no longer the distribution of the agents' marginal costs but the distribution of their virtual marginal costs (see Myerson, 1981) which determines who should produce and how much. Because virtual costs may be ranked differently from true costs, inefficiencies arise under asymmetric information. Moreover, competition may help reduce rents by putting each agent under the threat of being excluded from production if he overstates his cost too much. There is then a positive externality among competing agents.

Instead, more cooperative environments, such as public good problems or procurement of complementary inputs by several suppliers, involve negative externalities between agents. Given that each agent has a limited impact on the organization's overall production, the incentives to overstate costs and thereby receive greater transfers are exacerbated. ‘Free riding’ arises in such organizations (see Mailath and Postlewaite, 1990).

When competition between agents or between agents and the supervisors supposed to monitor them would benefit the principal, one must consider the possibility of collusion aimed at securing more rent. Reducing the scope for collusion requires using mechanisms that are less sensitive to information and reducing supervisory discretion. Incentive contracts look more like inflexible bureaucratic rules (see

Tirole, 1986; Laffont and Martimort, 2000). The optimal response to collusion may also entail more delegation to lower levels of the hierarchy, as in Laffont and Martimort (1998) and Faure-Grimaud, Laffont and Martimort (2003).

Dynamics

Different extensions of the static framework correspond to different abilities of the contractual partners to commit themselves inter-temporally and/or different ways for the cost parameters to vary over time. Under full commitment, the lessons of the static rent–efficiency trade-off can be easily extended, although the precise features of the optimal contract depend on how types evolve over time (see, for instance, Baron and Besanko, 1984, for the case of persistent types). The case of limited commitment is more interesting. Long-term contracts may either be renegotiated (Dewatripont, 1989; Hart and Tirole, 1988; Laffont and Tirole, 1990) or even are not feasible, in which cases the parties resort to spot contracts (Laffont and Tirole, 1988). The rent–efficiency trade-off must be adapted to take into account how information is revealed progressively over time. However, the basic idea still holds. As past performances reveal information about the agent's type, the optimal contract trades off *ex post* efficiency gains in contracting against the agent's desire to hide information in the earlier periods of the relationship so as to secure more rent in the later periods.

Applications

Since the mid-1980s, models of optimal contracting under adverse selection have spanned the economic literature. Let us quote only a few major applications. Mirrlees (1971) analysed optimal taxation schemes when the agent's productivity is privately observed. He introduced the Spence–Mirrlees condition and derived the implementability conditions. He also used optimal control techniques (Pontryagyn Principle) to compute the optimal taxation scheme. (The taxation problem differs from our buyer–seller example because participation in the mechanism is mandatory and the state's budget constraint must be added to the characterization of feasible allocations.)

Mussa and Rosen (1978) studied the problem of a monopolist selling one unit of a good to a continuum of consumers vertically differentiated with respect to their willingness to pay for the quality of this good. This was the first model using adverse selection techniques in a framework without income effect.

Maskin and Riley (1984) were interested in characterizing the optimal nonlinear price used by a monopolist in a second-degree price discrimination context.

Baron and Myerson (1982) applied the methodology to the regulation of natural monopolies privately informed about their marginal costs of production. Laffont and Tirole (1986) extended this analysis to allow for cost observability but also introduced moral hazard elements (the possibility for the regulated firm to reduce its costs by undertaking some non-observable effort). They derived cost-reimbursement rules and pricing policies. They showed that menus of linear contracts might implement the optimal contract.

Green and Kahn (1983) and Hart (1983) studied labour market contracts and discussed distortions towards overemployment or underemployment that may arise depending on the contractual environment

considered.
Finally, Townsend (1979) and Gale and Hellwig (1985) analysed optimal financial contracts in a framework where the borrower's income is observable only *ex post* and at a cost. Optimal contracts may look like debt in such environments.

Moral hazard

To return to our buyer–seller example, we now assume that there is only one unit of a good to be traded whose quality *q* is random and which yields a surplus $S(q)$ to the buyer. The distribution of quality is affected by an effort *e* undertaken by the agent at a cost $\psi(e)$ (where $\psi' > 0$ and $\psi'' > 0$). The cumulative distribution is $F(q|e)$ (with density $f(q|e)$) on a support $Q = [\underline{q}, \bar{q}]$ independent of the agent's effort. To simplify, the agent's preferences are separable in money and effort: $U = u(t) - \psi(e)$ where $u(\cdot)$ is increasing and concave $(u' > 0, u'' \leq 0)$. The agent's outside option is not to produce, which gives him a payoff normalized to zero.
The agent's effort is observable neither by the principal nor by a court of law. This is a *moral hazard* setting. Contracts stipulate the agent's payment as a function of the realized quality assumed to be observable and verifiable (contractible) by a court of law. Therefore, contracts are of the form $\{t(q)\}_{q \in Q}$.
If the effort were observable, its value could also be specified by contract. Therefore, the seller can at the same time be forced to exert the first-best level of effort and be fully insured against uncertainty on realized quality with a flat payment independent of his performance:

$$u(t^*) = \psi(e^*).$$

This is no longer the case when the agent's effort is non-verifiable. The first step of the analysis is to describe the set of feasible incentive contracts implementing a given level of effort *e*.
In a moral hazard setting, incentive constraints write as:

$$\int_{\underline{q}}^{\bar{q}} u(t(q))\, f(q|e)\, dq - \psi(e) \geq \int_{\underline{q}}^{\bar{q}} u(t(q))\, f(q|e')\, dq - \psi(e') \quad \forall (e, e'). \tag{6}$$

The agent's participation constraint is:

$$\int_{\underline{q}}^{\overline{q}} u(t(q))\, f(q|e)\, dq - \psi(e) \geq 0. \tag{7}$$

Risk neutrality

A first case of interest is when the agent is risk-neutral $(u(t) \equiv t)$. The simple ‘sell-out’ contract, $t(q) = S(q) - C$ where C is a constant, implements the first-best level of effort e^*. Provided that $C = \int_{\underline{q}}^{\overline{q}} S(q)\, f(q|e^*) - \psi(e^*)$, this scheme also extracts all the surplus from the agent who is just indifferent between producing or not.

Intuitively, with such a ‘sell-out’ contract, the agent's private incentives to exert effort are aligned with the social incentives. This efficient outcome is obtained by, first, having the agent pay a bond worth C for the right to serve the principal, and second, having the principal pay an amount $S(q)$ contingent on the quality realized.

Such a ‘sell-out’ contract requires that the agent bears the full consequences of a bad performance. It might not be feasible when the agent has limited liability and cannot be punished for bad performances. (For details, see Laffont and Martimort, 2002, ch. 4). The conjunction of moral hazard and limited liability allows the agent to derive a limited liability rent. Intuitively, only rewards, not punishments, can be used to provide incentives, and this restriction on instruments is costly for the principal. This rent creates a trade-off between efficiency and rent extraction, as in the adverse selection framework. Effort is distorted below the first-best level.

Risk aversion

Let us turn to the more complex case of risk aversion. A first concern of the literature has been to ‘simplify’ the set of incentive constraints (2) by replacing it with a first-order condition:

$$\int_{\underline{q}}^{\overline{q}} u(t(q))\, f_e(q|e)\, dq = \psi'(e). \tag{8}$$

Denoting by λ (resp. μ) the positive multiplier of the incentive (resp. participation) constraint (8) (resp. (7)), the optimal second-best schedule $t^{SB}(q)$ satisfies

$$\frac{1}{u'(t^{SB}(q))} = \mu + \lambda \frac{f_e(q|e)}{f(q|e)}. \tag{9}$$

This condition yields two important insights. First, the contract must simultaneously provide the risk-averse agent with insurance, which requires a fixed payment, and with incentives to exert effort, which requires that payments be linked to performance. There is now a trade-off between *insurance* and *incentives*.

Second, the monotonicity of the agent's compensation with respect to the quality level (a priori a quite intuitive property) is obtained only when the *monotone likelihood ratio property* holds, namely, when $\frac{\partial}{\partial q}\left(\frac{f_e(q|e)}{f(q|e)}\right) > 0$. This property means that higher levels of performance are more informative about the agent's effort.

Finally, the optimal contract must use all signals which are informative about the agent's effort but no uninformative signals. Using them would only let the agent bear more risk without any beneficial impact on incentives. This is the so-called informativeness principle of Holmström (1979).

Extensions

In a model with a finite number of quality and effort levels, Grossman and Hart (1983) offered a careful study of the set of incentive constraints and its consequences for the shape of optimal contracts. There is no general result on the ranking between the first-best and the second-best effort levels in such environments. The discrete version of the first-order approach requires that only nearby constraints matter in the agent's problem. This concavity of the agent's problem is ensured when $F(q|e)$ is itself convex in *q*. In models with a continuum of effort levels and outcomes, this first-order approach was suggested in Mirrlees (1999), more rigorously justified in Rogerson (1985) and Jewitt (1988) and applied in Holmström (1979) and Shavell (1979).

The moral hazard methodology has been used to justify the optimality of linear incentive schemes in well-structured environments (Holmström and Milgrom, 1987); an often found feature of real world contracts. Equipped with this tool, Holmström and Milgrom (1991; 1994) investigated how multiple tasks and jobs should be arranged in an organization.

To avoid the complexity of models with a continuum of effort levels, modellers have found it useful to focus on simplified environments with two levels of effort. This approach was instrumental in the work on corporate finance of Holmström and Tirole (1997).

Multi-agent organizations

When applied to multi-agent organizations, the ‘informativeness principle’ suggests that an agent's compensation should be linked to another's performance if it is informative about his own effort (see Mookherjee, 1984). Relative performance evaluation and benchmarking can help eliminate common

shocks affecting all agents' performances. Of particular importance in this respect are tournaments which use only the ranking of the agents' performances to determinate their compensations. Tournaments provide agents with insurance against common shocks, which has a positive incentive effect. More generally, the properties of tournaments and how they compare with (a priori suboptimal) linear schemes have been investigated in Nalebuff and Stiglitz (1983) and Green and Stokey (1983).

In more cooperative environments where different agents contribute to a joint project, the fundamental difficulty is how to share the proceeds of production among agents of the team and still provide some incentives. Since each agent enjoys only a fraction of those proceeds but bears the full cost of his effort, he reduces his effort supply. This leads to a free-rider problem within teams, which is analysed in Holmström (1982).

If we remain in cooperative environments but allow now for a principal acting as a budget breaker, this principal may find it worthwhile to reduce the agency cost of implementing a given effort profile by having agents behave cooperatively (Itoh, 1993). Even when agents do not cooperate, mutual observability of effort levels can also help to eliminate agency cost, as in Ma (1988). This last argument relies on the logic of non-verifiability models, developed below.

Dynamics

The basic issue investigated by dynamic models of moral hazard is the extent to which repeated relationships alleviate the moral hazard problem. The intuition is that the principal should filter out the agent's effort by looking at the whole history of his performances. This may eliminate any agency problem, at least when parties do not discount too much the future (see Laffont and Martimort, 2002, ch. 8, for an example). More generally, the insurance–incentives trade-off may be relaxed when the risk-averse agent's rewards and punishments can be smoothed over the whole relationship, as shown in Spear and Srivastava (1987). A direct consequence of inter-temporal smoothing is that the optimal dynamic contract exhibits *memory*; good (resp. bad) performance today will also affect positively (resp. negatively) future compensations. This insight has been used to formalize a theory of the wage dynamics inside the firm (Harris and Holmström, 1982).

Fama (1980) argued that reputation in the labour market exerts enough discipline on managers to alleviate moral hazard even in the absence of explicit contracts. Holmström (1999) built a model of career concerns where the manager's interest in influencing the labour market's beliefs concerning his or her quality provides incentives to exert effort. Career concerns are nevertheless in general not enough to induce first-best effort levels, and some inefficiencies remain.

Non-verifiability

Let us return to the buyer–seller model above. Although we now assume that it is observable by both the principal and the agent, the state of nature θ may still not be verifiable by a court of law, in which case it cannot be part of the contract. This shared knowledge stands in sharp contrast with the asymmetric information structures examined in previous sections.

The first difficulty consists of building a mechanism based only on verifiable variables (namely, the

quantities traded and corresponding payments) which implements the first-best quantity $q^*(\theta)$ and transfers $t^*(\theta)$. This problem was addressed by Maskin (1999). He demonstrated that the first-best quantities and transfers can easily be implemented with a direct revelation mechanism $\left\{t(\hat{\theta}_s, \hat{\theta}_b),\ q(\hat{\theta}_s, \hat{\theta}_b)\right\}_{(\hat{\theta}_s, \hat{\theta}_b)\in\Theta^2}$ where both the buyer and the seller report simultaneously the state of nature they commonly know. Truth-telling is obviously a Nash equilibrium of this mechanism provided that both traders are severely punished when making different reports, since such cases would be inconsistent with the underlying information structure.

A more subtle issue is how to design a mechanism such that this truthful Nash equilibrium is unique. Maskin (1999) proposed a condition for players' preferences such that this is the case. Moore and Repullo (1988) significantly extended the domain of preferences by hardening the implementation concept, replacing Nash behaviour by subgame-perfection in a sequential moves mechanism (see Laffont and Martimort, 2002, ch. 6, for an example, and Moore, 1992, for an exhaustive survey of the literature).

The basic thrust of the non-verifiability paradigm is that a court of law can get around non-verifiability by building such revelation mechanisms, at least as long as the non-verifiable state is payoff-relevant. If one sticks to that interpretation, non-verifiability does not present a significant limit on contracting.

A second issue of the literature is the impact of non-verifiability on the incentives of traders to perform specific and non-verifiable investments. Given our previous claim that non-verifiability is generally not a constraint, the model resembles the standard moral hazard model. Providing incentives for investments meets the same difficulties as in the previous section.

Extensions

In practice, revelation mechanisms have been criticized as overly complex, as relying on threats which may either be non-credible or violate limited liability constraints. The so-called incomplete contracts literature has thus focused on cases where such revelation mechanisms are not feasible. In such environments, either no contract at all or only a very rough one can be written *ex ante*. For instance, parties can agree *ex ante* on a simple fixed-price/fixed quantity contract which serves as a threat point for the bargaining which takes place *ex post* when the state of nature is realized (see Edlin and Reichelstein, 1996, among others).

Alternatively, this threat point may be determined by the allocation of ownership rights where such a right gives the owner the opportunity to use assets as he prefers in case bargaining fails (see Grossman and Hart, 1986; Hart and Moore, 1988). The issue is then to derive from those exogenous constraints distortions of investments and optimal organizations which may mitigate those distortions.

The incomplete contracts paradigm is similar to the complete contracts one (adverse selection, moral hazard and non-verifiability) in the sense that it also imposes limits on what a court may verify. It differs from it because it also imposes exogenous restrictions on the set of mechanisms available to the parties. The justification for these restrictions is found either in the bounded rationality of players or the difficulties in describing or foreseeing contingencies, all theoretical issues which remain high on the agenda of economic theorists and are still unsettled. The relevant literature on incomplete contracts is

too large to be summarized in this short article. The interested reader may refer to Tirole (1999) for an overview or to the entry for this term in this Dictionary.

See Also

- adverse selection
- agency problems
- incomplete contracts
- mechanism design
- mechanism design (new developments)
- moral hazard

I thank D. Gromb and J. Pouyet for helpful comments on an earlier version.

Bibliography

Armstrong, M. and Rochet, J. 1999. Multidimensional screening: a user's guide. *European Economic Review* 43, 959–79.

Bagnoli, M. and Bergstrom, T. 2005. Log-concave probability and its applications. *Economic Theory* 26, 445–69.

Baron, D. and Besanko, D. 1984. Regulation and information in a continuing relationship. *Information Economics and Policy* 1, 447–70.

Baron, D. and Myerson, R. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50, 911–30.

Crémer, J., Khalil, F. and Rochet, J. 1998. Strategic information gathering before a contract is offered. *Journal of Economic Theory* 81, 163–200.

Crémer, J. and McLean, R. 1988. Full extraction of surplus in Bayesian and dominant strategy auctions. *Econometrica* 56, 1247–57.

D'Aspremont, C. and Gérard-Varet, L. 1979. Incentives and incomplete information. *Journal of Public Economics* 11, 25–45.

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules. *Review of Economic Studies* 46, 185–216.

Dewatripont, M. 1989. Renegotiation and revelation information over time: the case of optimal labour contracts. *Quarterly Journal of Economics* 104, 489–520.

Edlin, A. and Reichelstein, S. 1996. Hold-ups, standard breach remedies, and optimal investments. *American Economic Review* 86, 478–501.

Fama, E. 1980. Agency problem and the theory of the firm. *Journal of Political Economy* 88, 288–307.

Faure-Grimaud, A., Laffont, J.-J. and Martimort, D. 2003. Collusion, delegation and supervision with soft information. *Review of Economic Studies* 70, 253–80.

Gale, D. and Hellwig, M. 1985. Incentive-compatible debt contracts: the one-period problem. *Review of Economic Studies* 52, 647–63.

Gibbard, A. 1973. Manipulations of voting schemes: a generalized result. *Econometrica* 41, 587–601.

Green, J. and Kahn, C. 1983. Wage-employment contracts. *Quarterly Journal of Economics* 98, 173–88.

Green, J. and Laffont, J.-J. 1977. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica* 45, 427–38.

Green, J. and Stokey, N. 1983. A comparison of tournaments and contracts. *Journal of Political Economy* 91, 349–64.

Grossman, S. and Hart, O. 1983. An analysis of the principal–agent problem. *Econometrica* 51, 7–45.

Grossman, S. and Hart, O. 1986. The costs and benefits of ownership: a theory of lateral and vertical integration. *Journal of Political Economy* 94, 691–719.

Groves, T. 1973. Incentives in teams. *Econometrica* 41, 617–31.

Guesnerie, R. and Laffont, J.-J. 1984. A complete solution to a class of principal–agent problems with an application to the control of a self–managed firm. *Journal of Public Economics* 25, 329–69.

Harris, M. and Holmström, B. 1982. A theory of wage dynamics. *Review of Economic Studies* 49, 315–33.

Hart, O. 1983. Optimal labour contracts under asymmetric information: an introduction. *Review of Economic Studies* 50, 3–35.

Hart, O. and Moore, J. 1988. Property rights and the nature of the firm. *Journal of Political Economy* 98,

1119–58.

Hart, O. and Tirole, J. 1988. Contract renegotiation and Coasian dynamics. *Review of Economic Studies* 55, 509–40.

Holmström, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Holmström, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13, 324–40.

Holmström, B. 1999. Managerial incentive problems: a dynamic perspective. *Review of Economic Studies* 66, 169–82.

Holmström, B. and Milgrom, P. 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55, 303–28.

Holmström, B. and Milgrom, P. 1991. Multi-task principal agent analysis. *Journal of Law, Economics, and Organization* 7, 24–52.

Holmström, B. and Milgrom, P. 1994. The firm as an incentive system. *American Economic Review* 84, 972–91.

Holmström, B. and Myerson, R. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51, 1799–819.

Holmström, B. and Tirole, J. 1997. Financial intermediation, loanable funds, and the real sector. *Quarterly Journal of Economics* 112, 663–91.

Itoh, H. 1993. Coalition incentives and risk-sharing. *Journal of Economic Theory* 60, 416–27.

Jewitt, I. 1988. Justifying the first-order approach to principal–agent problems. *Econometrica* 56, 1177–90.

Jullien, B. 2000. Participation constraints in adverse-selection models. *Journal of Economic Theory* 93, 1–47.

Laffont, J.-J. and Martimort, D. 1998. Collusion and delegation. *Rand Journal of Economics* 29, 280–305.

Laffont, J.-J. and Martimort, D. 2000. Mechanism design with collusion and correlation. *Econometrica* 68, 309–42.

Laffont, J.-J. and Martimort, D. 2002. *The Theory of Incentives: The Principal–Agent Model*. Princeton: Princeton University Press.

Laffont, J.-J. and Maskin, E. 1979. A differentiable approach to expected utility maximizing mechanisms. In *Aggregation and Revelation of Preferences*, ed. J.-J. Laffont. Amsterdam: North-Holland.

Laffont, J.-J. and Tirole, J. 1986. Using cost observation to regulate firms. *Journal of Political Economy* 94, 614–41.

Laffont, J.-J. and Tirole, J. 1988. The dynamics of incentive contracts. *Econometrica* 56, 1153–75.

Laffont, J.-J. and Tirole, J. 1990. Adverse selection and renegotiation in procurement. *Review of Economic Studies* 57, 597–626.

Laffont, J.-J. and Tirole, J. 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

Lewis, T. and Sappington, D. 1989. Countervailing incentives in agency problems. *Journal of Economic Theory* 49, 294–313.

Ma, C. 1988. Unique implementation of incentive contracts with many agents. *Review of Economic Studies* 55, 555–72.

Mailath, G. and Postlewaite, A. 1990. Asymmetric information bargaining problems with many agents. *Review of Economic Studies* 57, 351–638.

Maskin, E. 1999. Nash equilibrium and welfare optimality, *Review of Economic Studies* 66, 23–38.

Maskin, E. and Riley, J. 1984. Monopoly with incomplete information. *Rand Journal of Economics* 15, 171–96.

Maskin, E. and Tirole, J. 1990. The principal–agent relationship with an informed principal. I: Private values. *Econometrica* 58, 379–410.

Maskin, E. and Tirole, J. 1992. The principal–agent relationship with an informed principal. II: Common values. *Econometrica* 60, 1–42.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Mirrlees, J. 1999. The theory of moral hazard with unobservable behaviour. Part I. *Review of Economic Studies* 66, 3–22.

Mookherjee, D. 1984. Optimal incentive schemes with many agents. *Review of Economic Studies* 51, 433–46.

Moore, J. 1992. Implementation in environments with complete information. In *Advances in Economic Theory*, ed. J.-J. Laffont. Cambridge: Cambridge University Press.

Moore, J. and Repullo, R. 1988. Subgame-perfect implementation. *Econometrica* 56, 1191–20.

Mussa, M. and Rosen, S. 1978. Monopoly and product quality. *Journal of Economic Theory* 18, 301–17.

Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47, 61–73.

Myerson, R. 1981. Optimal auction design. *Mathematics of Operations Research* 6, 58–63.

Myerson, R. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 28, 61–73.

Nalebuff, B. and Stiglitz, J. 1983. Prizes and incentives: towards a general theory of compensation. *Bell Journal of Economics* 14, 21–43.

Rogerson, W. 1985. The first-order approach to principal–agent problems. *Econometrica* 53, 1357–68.

Shavell, S. 1979. On moral hazard and insurance. *Quarterly Journal of Economics* 93, 541–62.

Smith, A. 1776. *The Wealth of Nations*. New York: Prometheus Books, 1991.

Spear, S. and Srivastava, S. 1987. On repeated moral hazard with discounting. *Review of Economic Studies* 54, 599–617.

Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87, 355–74.

Spence, M. 1974. *Market Signalling: Informational Transfer in Hiring and Related Processes*. Cambridge, MA: Harvard University Press.

Tirole, J. 1986. Hierarchies and bureaucracies: on the role of collusion in organizations, *Journal of Law, Economics and Organization* 2, 181–214.

Tirole, J. 1999. Incomplete contracts: where do we stand? *Econometrica* 67, 741–82.

Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21, 417–25.

How to cite this article

Martimort, David. "contract theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000340> doi:10.1057/9780230226203.0309

The New Palgrave Dictionary of Economics Online

contracting in firms

Canice Prendergast
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides an overview of recent advances in theoretical and empirical work on incentive contracting in firms. The specific focus is on a variety of reasons why the prediction of the early literature on contracting –suggesting a strong relationship between performance and pay – has not been borne out.

Keywords

agency theory; contracting in firms; externalities; free rider problem; incentive contracts; input monitoring; measurement error; multitasking; performance-related pay; principal and agent; private information; risk

Article

In many realms of economic life, the actions of individuals affect the welfare of others. Nowhere is this more relevant than in firms, where employees act on behalf of owners or shareholders to provide services for customers and clients. This separation of the interests of employees from those whose actions they benefit has generated a large literature on incentive contracting, where the overarching objective is the alignment of such interests. The early literature on agency theory, described in the first edition of this volume by Lazear (1987), conceptually mimics that on externalities – the other area of economics that deals with welfare consequences of actions on others – by showing a variety of ways in which the compensation of agents can be constructed to internalize the effects on one's actions on others. There are two ways of doing this. First, one could simply tell employees what to do and to penalize them if they fail to do so. In the literature, this is referred to as input monitoring. While this can sometimes help, it is often hard to monitor either what workers do, or the intensity with which they do so – a salesman on the road would be a good example. Similarly, while overseers can sometimes identify what it is that agents are doing, they may not know what they *should* be doing – a board of directors

monitoring a CEO would be apposite here. Accordingly, the second solution to misaligned incentives is to design compensation plans such that the agent's pay depends on her contribution – ‘output’ – so that the concerns of other parties are internalized.

A simple model can illustrate this point, and is useful to describe other complications that can arise. The agent is assumed to take some action (‘effort’) $e \geq 0$, which is unobserved by the principal. She is averse to exerting effort. Consider a simple parameterization of the agent's utility function, where the agent cares about wages *w* and effort; assume that the agent has exponential utility $V = -\exp[-r(w - C(e))]$, where *w* is the worker's wage, $r \geq 0$ is the constant rate of absolute risk aversion, the worker's cost of supplying effort is $C(e) = \frac{ce^2}{2}$, and her reservation utility is U^*. To focus attention on the role of output contacting, assume that the principal cannot observe effort *e* (so monitoring of inputs is not possible), but instead only observes a signal on effort $y = e + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, with σ^2 representing measurement error. Assume also that the principal chooses to reward the agent in a linear fashion on output – a piece rate: $w = \beta_0 + \beta_y y$. (There is a large literature on the optimal shape of compensation contracts – see Prendergast, 1999; Gibbons, 1996, for an overview.) Then there is a simple solution to attaining efficient effort: choose the contact to internalize the benefit to others by setting $\beta_y = 1$. In words, efficiency arises when the agent is residual claimant on the benefit of others.

This solution, providing a simple prescription for how compensation contacts should be designed, is both simple and intuitive. And empirically false. There are, of course, some occupations where one can find evidence of such ‘high-powered incentives’, where agents are essentially residual claimants on output. Indeed, the literature on agency theory is replete with references to such occupations – taxi cab drivers, franchisees, sharecropper farmers and the self-employed. Yet these are exceptions; instead, ‘low-powered incentives’ in firms are more the norm (see Prendergast, 1999, for details). Consequently, one of the quandaries of the literature has become why so few workers seem to have contracts where their pay is strongly linked to their performance, and much of the subsequent literature to that outlined in the first edition of the *New Palgrave* has identified relevant constraints on incentive contracting.

The earliest candidate to explain why high-powered incentives are rare is that high-powered contracts impose *risk* on workers (Holmstrom, 1979). Consider the contract that induces efficient effort above: $\beta_y = 1$. The objective of the firm is to maximize profits subject to the worker's willingness to take the position. This implies that the fixed component, β_0, is changed to guarantee that agents earn their reservation utility, so the principal's objective becomes a surplus maximization exercise. When the worker is risk neutral, the fixed component is reduced sufficiently such that the total compensation cost is $U^* + \frac{c}{2}$. In words, the only cost that the employer incurs in addition to U^* is the effort cost. This is not true when the worker is risk averse. In the context of the preferences *V* above, compensation costs increase when incentive contracts are used for two reasons – the cost of increased effort as above, but also a risk cost imposed on workers. Both costs are increasing in β_y. With exponential preferences and linear contracts, this trade-off results in the optimal contract being $\beta_y^* = \frac{1}{1 + rc\sigma_y^2}$. This approach to studying incentive contracting has become knows as the ‘trade-off of risk and incentives’, where firms trade off the benefits of great effort with higher compensation costs induced by a risk premium, such that the chosen level of effort falls below the level that internalizes benefits to others. Only in the case

where there is either no measurement error ($\sigma_y^2 = 0$) or risk neutrality (r=0) does efficient effort arise. At its most general, this costliness of exposing a worker to large degrees of risk (or its analogue, liquidity constraints) surely explains some part of the absence of high-powered incentives. In much the same way as financial assets with higher undiversifiable risk require higher expected returns, so also are risky jobs likely to demand higher compensation. Despite this, the empirical literature on how compensation contracts trade off such risk issues against higher effort has shown little evidence in its favour. There are two principal empirical implications of the theory. First, riskier environments should have lower incentives – β_y^* declines with σ_y^2. There have been many studies of the relationship between risk and the strength of incentives in a variety of occupations. If anything, this literature suggests that the relationship between risk and the provision of incentives is positive rather than the negative relationship posited by this theory. See Prendergast (2002) for details and an explanation as to why this may be.

Second, the trade-off of risk and incentives implies that compensation should not depend on measures that workers cannot control. Again, this has found little support in the data. For example, Bertrand and Mullainathan (2001), have examined executive contracts in the United States, and found little evidence that contracts reward executives any less for measures that they cannot control (say, where an oil company's profits change simply because the price of crude changed) than for those that they can (such as a merger). More evidence on this failure to filter out uncontrollable factors concerns the infrequency of relative performance evaluation. Consider two sales-force workers (or executives) who carry out a similar job. If demand for the products that they sell varies for common reasons beyond their control, an efficient way of limiting risk exposure is to (at least partially) reward the workers on how well they do relative to each other. Yet empirically there is relatively little evidence of such benchmarking (for example, see Janakiraman, Lambert and Larker, 1992).

A second limitation on incentive contracting arises when measures do not reflect the objectives of the principal. Workers often carry out a host of activities in their jobs, yet measures of performance may not reflect all these aspects. A good example of this would be measuring the performance of a teacher. While measures may be available on some component of what they do – such as test scores for a teacher – many important aspects may remain unmeasured. When contracts are designed on the subset of things that can be measured, there is a danger that they ignore the unmeasured aspects. For instance, there is evidence of teachers ‘teaching for the test’ or cheating to achieve higher test scores (Jacob and Levitt, 2003). This phenomenon has become known as *multitasking* (Holmstrom and Milgrom, 1992), which becomes potentially important when there is no single measure that reflects the contribution of an agent. Accordingly, it is not surprising that a consistent empirical finding is that jobs which are described by firms as complex tend not to offer significant incentive pay (see Prendergast, 1999, for details).

Another limitation on the ability of firms to provide incentives to workers comes from team production. Measures of performance for most workers reflect not only what they do but also the contributions of others. In itself, this does not change the calculus above in any conceptual sense, other than that the measurement error now includes the actions of others. As an example, assume that two agents (1 and 2) work on a team and that output measures the true contributions of both plus an error term $y=e_1+e_2+\varepsilon$. Efficient effort arises as before by setting $\beta_y=1$ for each worker. However, there is now a potential problem of budget breaking, where marginal payments exceed marginal output. In this example, when total output rises by one dollar, compensation costs increase by two dollars. In many firms – for

instance, partnerships – such budget breaking is not possible. If instead the principal can pay out no more than one dollar for every dollar extra on output, this naturally places an upper bound of $\beta_y = \frac{1}{2}$ on average for the agents. Hence, budget balancing places a natural limitation on firm incentives. This also leads to a free rider problem in teams, where maximum incentive compensation in an *N* member team mechanically declines as *N* increases. (This is known as the '$\frac{1}{N}$ problem'.) There is also considerable empirical evidence (such as Gaynor and Pauly, 1990) on such free riding – mostly from legal and medical partnerships – illustrating how various measures of performance disimprove as the size of the team being rewarded increases.

Many measures of output are not denominated in dollar terms, but instead come in the form of evaluations by others. For instance, it would be difficult to measure the contribution of a social worker or a customer service representative without using feedback from supervisors or clients. Another limitation on contracts arises when such subjective measures can be corrupted by evaluators with vested interests. Two particular sources of such vested interests have been considered in the literature. First, information on performance often originates with clients as they are the only ones with first-hand experience of the agent's efforts. For instance, compensation for many customer service representatives depends on client evaluations. When clients have relatively similar preferences to the principal – such as that the agent should be courteous and efficient – contracts based on evaluations can mimic the objective contracts above. Yet in other instances, the vested interest of clients can render incentives difficult to implement. A good example of this arises in occupations such as police or immigration control, whose objective is not necessarily to make their clients happy. Making pay depend on evaluations in these instances can be harmful as it gives agents incentives to keep clients happy when they should not, such as a police officer not arresting a suspect. In these cases, incentive contracting on evaluations typically needs to be curtailed to avoid such incentives (see Prendergast, 2003, for details).

The second example of vested interests with subjective evaluations is where the principal has an incentive not to implement the (*ex ante*) efficient contract by reneging on a promised payment to save costs. Thus, even though an agent exerts effort and performs well, the supervisor claims otherwise to keep costs down. This can arise either by outright lying or perhaps by manipulating whatever measures are available. A relevant example here is the movie industry, where actors are sometimes paid on the 'net profits' of a film. As a result, there have been numerous court cases regarding firms using creative transfer pricing arrangements to reduce profits for very successful movies. See Cheatham, Davis and Cheatham (1996) for more details on this. Such incentives to renege are likely even worse when there are no objective measures of performance. Because the desire to renege is greater when discretionary incentive payments are higher, it follows that the only credible contracts often involve few incentives. (Clive Bull, 1987, considers a role for repeated interaction between the principal and agent as a means of reducing incentives to renege. While repeating the relationship can result in sufficient incentives for complete honesty by the principal, it remains the case that, if the relationship's value is not sufficiently great, incentives must be muted to reduce incentives for cheating.)

It is incorrect to assume that the ability to manipulate measures of performance always mute incentives – sometimes it can result in incentive pay being inefficiently high. Consider again two occupations where agents are typically residual claimants – taxi drivers and sharecroppers. At first blush, it would seem odd that they have such extreme incentives. Aren't these as likely candidates for trading off risk against

incentives as any? However, one characteristic of each of these occupations is that they have opportunities for hiding output, either by taking fares without using the metre (in the case of cab drivers) or selling crops privately (in the case of farmers). In both cases, the only outcome that makes this incentive irrelevant is to render them residual claimants, even if risk considerations would suggest otherwise.

Another issue which can constrain efficient incentives, yet which has received almost no attention in the empirical literature, is where agents hold *private information*. Take a specific instance – real estate agents. In Chicago, real estate contracts take a simple form – agents make three per cent of the sales price of the house. Assume that my home is worth $500. This linear contract not only offers only three per cent on the relevant margin for improving the selling price of the house, but predominantly rewards the agent for selling the house for say $450. Yet anyone could sell the house for $450 and it seems highly inefficient to reward in this way. So why not renegotiate to something better? An example of such an improvement (subject to risk issues) would be to offer nothing on the first $450, but to pay a piece rate of 30 per cent on anything over $450. In this way, the agent has more incentives on margin, yet breaks even relative to the original contract if the house sells for its original price.

One reason why such renegotiation does not arise is that the agent may privately know the true value of the home, while the owner believes it to be worth $500 on average. Consider a homeowner who offers the new contract above to the agent. It is clear that the agent rejects the new contract if it is truly worth less than expected, and accepts it if worth more. But this implies that the agent earns *information rents* on average. As a result, on average the homeowner loses money from the renegotiation unless effort increases enough. This option available to the agent limits the ability of contracts to attain efficiency. Instead, in the usual monopoly fashion, the homeowner would offer a contract to trade off the efficiency gains of increased effort with infra-marginal losses of the type described above, resulting in lower-powered incentives. (There is a large mechanism design literature on this topic that has largely been ignored by the empirical literature on incentives; see Laffont and Tirole, 1986, for example. This is surely partly because of the empirical conundrum as to why mechanisms are so rare in reality.)

Much of the recent literature has been focused on how incentive contracts can cause adverse behavioural responses. Another possible mechanism for such responses is where *intrinsic motivation* can be crowded out by the use of incentive contracts. The premise of this literature has been that in many occupations agents enjoy carrying out the activity or care about the outcomes of their actions. As a result, they will exert effort beyond that which they can get away with even in the absence of incentive contracts. This, in itself, is not enough to limit incentive contracting. However, there is some psychological evidence that agents enjoy their jobs less when incentive contracting is used. In effect, they feel that they are only doing it ‘for the money’ and hence lose interest. A commonly cited example is the willingness of people to donate blood, where the warm feeling from donating declines when payments are made. In some instances, this can imply that incentive contracting can *reduce* effort if these crowding out effects are strong enough. As a result, it can be optimal to provide no incentives even when effort is one-dimensional. This area of research, whose empirical testing has largely been restricted to the laboratory, is still in its early stages and is likely to see much refining over the coming years. See Frey and Jegen (2001) for a survey.

Another likely fruitful area of future research concerns non-monetary ways of motivating workers. This literature largely began as an exercise in how workers could be motivated to internalize the benefits of others, yet has almost exclusively become an exercise in how to motivate through monetary contracting.

Yet it is clear that there are a myriad of means of motivating workers – sense of achievement, 'doing good', status, and so on – that firms tap into. How such mechanisms operate, and the way in which they interact with monetary contracts, remain an unstudied topic of research, though see Besley and Ghatak (2005), for some theoretical work on this issue.

It is worthwhile to note a caveat before concluding. The discussion above concerns the absence of *observed* incentive contracts. Yet workers often have unobserved carrots and sticks that can motivate them. For instance, many workers exert effort in the hope of attaining a promotion (Lazear and Rosen, 1981), or a better job offer (Holmstrom, 1999). Many of these mechanisms for inducing desired behaviour are dynamic, where good performance today results in a greater likelihood of promotion, or better job offers in future. Such incentives are clearly important for workers. However, it remains the case that explicit incentive payments remain limited even in those cases where the above types of career concern are negligible. (For example, it is well known that promotion prospects become very limited for workers who remain in a job grade for a long period. Yet explicit incentives are no more common for those workers than for any other.) The interaction of unobserved (typically career) incentives with the more explicit set of piece rates and bonuses that have been considered above is surely of first-order importance to firms, though it remains surprisingly unexplored in the literature (see Baker, Gibbons and Murphy, 1994, for an exception).

To conclude, perhaps the central foundation of modern economics is the idea that appropriate prices guide behaviour in efficient ways. Despite this, one of the defining characteristics of the employment relationship in many firms is the absence of the kind of explicit prices whereby wages depend in a clear way on observed outcomes. The early incarnations of agency theory were concerned with designing prices in a way that could serve to fully internalize the effects of agents' actions on the welfare of their employers. Yet this initial optimism has now been tempered with a somewhat more nuanced view that shows trade-offs that will ultimately help in defining more precisely the nature of labour market relationships.

Bibliography

Baker, G., Gibbons, R. and Murphy, K.J. 1994. Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics* 109, 1125–56.

Bertrand, M. and Mullainathan, S. 2001. Are CEOs rewarded for luck? The ones without principals are. *Quarterly Journal of Economics* 116, 901–32.

Besley, T. and Ghatak, M. 2005. Competition and incentives with motivated agents. *American Economic Review* 95, 616–36.

Bull, C. 1987. The existence of self-enforcing wage contracts. *Quarterly Journal of Economics* 102, 147–59.

Cheatham, C., Davis, D. and Cheatham, L. 1996. Hollywood profits: gone with the wind? *CPA Journal*

12, 32–4.

Frey, B. and Jegen, R. 2001. Motivation crowding theory: a survey of empirical evidence. *Journal of Economic Surveys* 15, 589–611.

Gaynor, M. and Pauly, M. 1990. Compensation and productive efficiency in partnerships. Evidence from medical group practice. *Journal of Political Economy* 98, 544–74.

Gibbons, R. 1996. Incentives and careers in organizations. In *Advances in Economics and Econometrics: Theory and Applications*, ed. D. Kreps and K. Wallis. Cambridge: Cambridge University Press.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Holmstrom, B. 1999. Managerial incentive problems: a dynamic perspective. *Review of Economic Studies* 66, 169–82.

Holmstrom, B. and Milgrom, P. 1992. Multi-task principal agent analyses: linear contracts, asset ownership and job design. *Journal of Law, Economics, and Organization* 7, 24–52.

Jacob, B.A. and Levitt, S.D. 2003. Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118, 843–77.

Janakiraman, S.N., Lambert, R.A. and Larker, D.F. 1992. An empirical investigation of the relative performance evaluation hypothesis. *Journal of Accounting Research* 30, 53–69.

Laffont, J.-J. and Tirole, J. 1986. Using cost observation to regulate firms. *Journal of Political Economy* 94, 614–41.

Lazear, E.P. 1987. Incentive contracts. In *The New Palgrave: A Dictionary of Economics*, vol. 2., ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Lazear, E. and Rosen, S. 1981. Rank order tournaments as optimal labor contracts. *Journal of Political Economy* 89, 841–64.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37, 7–63.

Prendergast, C. 2002. The tenuous trade-off between risk and incentives? *Journal of Political Economy* 110, 1071–102.

Prendergast, C. 2003. The limits of bureaucratic efficiency. *Journal of Political Economy* 111, 929–59.

How to cite this article

Prendergast, Canice. "contracting in firms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000217> doi:10.1057/9780230226203.0310

The New Palgrave Dictionary of Economics Online

control functions

Salvador Navarro

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The control function approach is an econometric method used to correct for biases that arise as a consequence of selection and/or endogeneity. It is the leading approach for dealing with selection bias in the correlated random coefficients model. The basic idea of the method is to model the dependence between the variables not observed by the analyst on the observables in a way that allows us to construct a function *K* such that, conditional on the function, the endogeneity problem (relative to the object of interest) disappears.

Keywords

average treatment effect; control functions; endogeneity; identification; instrumental variables; Roy model; selection bias

Article

The control function approach is an econometric method used to correct for biases that arise as a consequence of selection and/or endogeneity. It is the leading approach for dealing with selection bias in the correlated random coefficients model (see Heckman and Robb, 1985; 1986; Heckman and Vytlacil, 1998; Wooldridge, 1997; 2003; Heckman and Navarro, 2004), but it can be applied in more general semiparametric settings (see Newey, Powell and Vella, 1999; Altonji and Matzkin, 2005; Chesher, 2003; Imbens and Newey, 2006; Florens et al., 2007).

The basic idea behind the control function methodology is to model the dependence between the variables not observed by the analyst on the observables in a way that allows us to construct a function *K* such that, conditional on the function, the endogeneity problem (relative to the object of interest) disappears.

In this article I deal exclusively with the problem of identification. That is, I assume access to data on an arbitrarily large population. As a consequence, I do not discuss estimation, standard errors or inference. In the examples, I analyse how to recover parameters in a way that, I hope, shows directly how to perform estimation via sample analogues.

The Set-up

The general set-up I consider is the following two-equation structural model; an outcome equation:

$$Y = g(X, D, \varepsilon),$$

(1)

and an equation describing the mechanism assigning values of *D* to individuals:

$$D = h(X, Z, v),$$
(2)

where *X* and *Z* are vectors of observed random variables, *D* is a (possibly vector valued) observed random variable, and ε and ν are general disturbance vectors not independent of each other but satisfying some form of independence of *X* and *Z*.

The problem of endogeneity arises because *D* is correlated with ε via the dependence between ε and ν. Because eq. (2) represents an assignment mechanism in many economic models, it is generically called the 'selection' or 'choice' equation. This set-up has been applied to problems like earnings and schooling (Willis and Rosen, 1979; Cunha, Heckman and Navarro, 2005), wages and sectoral choice (Heckman and Sedlacek, 1985) and production functions and productivity (Olley and Pakes, 1996), among others.

The goal of the analysis is to recover some functional of $g(X, D, \varepsilon)$ of interest

$$a(X, D)$$
(3)

that cannot be recovered in a straightforward way because of the endogeneity/selection problem. As an example, when *D* is binary interest sometimes centres on the effect of going from *D*=0 to *D*=1 for an individual chosen at random from the population, the so-called average treatment effect: $a(X, D) = E(g(X, 1, \varepsilon) - g(X, 0, \varepsilon))$.

The key behind the control function approach is to notice that (conditional on *X*, *Z*) the only source of dependence is given by the relation between ε and ν. If ν was known, we could condition on it and analyse eq. (1) without having to worry about endogeneity. The main idea behind the control function approach is to recover some function of ν via its relationship with the model observables so that we can now condition on it and solve the endogeneity problem.

Definition: The control function approach proposes a function *K* (the control function) that allows us to recover *a* (*X*, *D*) such that *K* satisfies

- **A-1**. *K* is a function of *X*, *Z*, *D*.
- **A-2**. ε satisfies some form of independence of *D* conditional on ρ (*X*, *K*), with ρ a knowable function.
- **A-3**. *K* is identified.

Assumption **A-2** is the key assumption of the approach. It states that, once we condition on *K*, the dependence between ε and *D* (that is, the endogeneity) is no longer a problem. To help fix ideas, consider the following example of a simple linear in parameters additively separable version of the model of eqs (1) and (2).

Example 1: Linear regression with constant effects. Write the outcome eq. (1) as

$$Y = X\beta + D\alpha + \varepsilon$$

and assume that our object of interest (3) is α. Assume that we can write eq. (2) as

$$D = X\rho + Z\pi + \nu \quad (4)$$

with ν, $\varepsilon \perp\perp X$, Z where $\perp\perp$ denotes statistical independence. Such a model arises, for example, if *Y* is logearnings and *D* is years of schooling as in Heckman, Lochner and Todd (2003). If ability is unobservable since high ability is associated with higher earnings but also with higher schooling, then ε and ν would be correlated.
If we let $K = \nu$ be the residual of the regression in (4), then we can recover α from the following regression

$$Y = X\beta + D\alpha + K\psi + \eta,$$

where it follows that $E(\eta|X, K) = 0$. It is easy to show that in this case the control function estimator and the two-stage least squares estimator are equivalent. (To my knowledge, although in a different context – a SUR model – Telser, 1964, was the first to use the residuals from other equations as regressors in the equation of interest.)
The previous case is a simple example of a control function where $K=D-E(D|X, Z)$. In this case, because of the constant effects assumption (that is, α is not random), standard instrumental variables methods and the control function approach coincide. In general, this is not the case.
In the next section I describe in detail the control function methodology for the binary choice case (Roy, 1951). This case is interesting both because it is the workhorse of the policy evaluation literature and because, by virtue of its nonlinearity, it highlights the implications of a nonlinear structure in a relatively simple context. I then briefly describe extensions to more general cases. For simplicity, I focus on the additively separable in unobservables case, but recent research provides generalizations to non-additive functions (see Blundell and Powell, 2003; Imbens and Newey, 2006, among others).

The case of a binary endogenous variable

In this section I describe how the control function approach solves the selection/endogeneity problem when the endogenous variable is binary. This problem has a long tradition in economics going back (at least) to Roy (1951). In Roy's original version of the model (see Roy model) an individual is deciding whether to become a fisherman (*D*=0) or a hunter (*D*=1).
Associated with each occupation is a payoff $Y_D = g_D(X) + \varepsilon_D$. Since we can only observe individuals in one sector at a time, the *observed* outcome for an individual is given by Y_1 if he becomes a hunter (*D*=1) and by Y_0 if he becomes a fisherman (*D*=0). That is, the observed outcome (*Y*) can be written as:

$$Y = DY_1 + (1 - D)Y_0 = g_0(X) + D(g_1(X) - g_0(X)) + \varepsilon_0 + D(\varepsilon_1 - \varepsilon_0). \quad (5)$$

The model is closed by assuming that individuals choose the occupation with the highest payoff. That is,

$$D = \mathbf{1}(Y_1 - Y_0 > 0) = \mathbf{1}(g_1(X) - g_0(X) + \varepsilon_1 - \varepsilon_0 > 0), \tag{6}$$

where **1**(*a*) is an indicator function that takes value 1 if *a* is true and 0 if it is false. Endogeneity arises because the error term in choice eq. (6) contains the same random variables as the outcome eq. (5). A generalized version of the model replaces the simple income maximization rule in (6) with a more general decision rule

$$D = \mathbf{1}(h(X, Z) - \nu > 0). \tag{7}$$

The model described by eqs (5) and (7) is general enough to be used in many different cases. Many qsts of interest in economics fit this framework if, instead of thinking of two sectors, fishing and hunting, we think of two generic potential states, the treated state (*D*=1) and the untreated state (*D*=0) with their associated potential outcomes. The decision rule in (7) is general enough to capture not only income maximization but also utility maximization and even a deciding actor different from the agent directly affected by the outcomes (parents deciding for their children, for example). The simple income maximization rule in (6) shows why, *in general* if $\varepsilon_1 \neq \varepsilon_0$, then $\varepsilon_1 - \varepsilon_0$ is likely to be correlated with *D*.

The correlated random coefficients model is a special case of the model described by (5) and (7) when $\varepsilon_1 - \varepsilon_0$ is not independent of *D* and $g_j(X) = \alpha_j + X\beta$ for *j*=0,1. (For simplicity I assume $\beta_1 = \beta_0 = \beta$. The case where $\beta_1 \neq \beta_0$ follows directly.) To see why simply rewrite (5) as

$$Y = \alpha_0 + X\beta + D(\alpha_1 - \alpha_0 + \varepsilon_1 - \varepsilon_0) + \varepsilon_0 \tag{8}$$

so that now the coefficient on *D* is (*a*) random and (*b*) correlated with *D*. In this case we have that the gains from treatment ($\alpha_1 - \alpha_0 + \varepsilon_1 - \varepsilon_0$) are heterogeneous (that is, they are not constant even after controlling for *X*) and they are correlated with *D*. I come back to this special linear in parameters case in ex 2.

Though other parameters of interest can be defined, I consider the case in which we are interested in the two particular functionals that receive the most attention in the evaluation literature – the average treatment effect and the average effect of treatment on the treated. I impose that ε_1, ε_0, ν are absolutely continuous with finite means, and that ε_1, ε_0, $\nu \perp\!\!\perp X, Z$. (One could weaken the assumption to be $\varepsilon_1, \varepsilon_0 \perp\!\!\perp X|Z$ and $\nu \perp\!\!\perp X, Z$.)

Under these assumptions the average treatment effect is given by

$$ATE(x) = E(Y_1 - Y_0|X = x) = g_1(x) - g_0(x) = x(\beta_1 - \beta_0)$$

where the last equality follows if eq. (8) applies. *ATE*(*X*) is of interest to answer qsts like the average effect of a policy that is mandatory, for example. When receipt of treatment is not mandatory or randomly assigned, the average effect of treatment among those individuals who are selected into treatment is commonly the functional of interest (see Heckman, 1997; Heckman and Smith, 1998). This effect is measured by the average effect of treatment on the treated:

$$TT(x) = E(Y_1 - Y_0|X = x, D = 1) = g_1(x) - g_0(x) + E(\varepsilon_1 - \varepsilon_0|X = x, D = 1) = \alpha_1 - \alpha_0 + E(\varepsilon_1 - \varepsilon_0|X = x, D = 1),$$

where the last equality follows for the linear in parameters case of eq. (8).
Now, suppose we ignored the endogeneity problem and attempted to recover either of these objects from the data on outcomes at hand. In particular, if we used the (observed) conditional means of the outcome

$$E(Y|X = x, D = 1) - E(Y|X = x, D = 0) = g_1(x) - g_0(x) + E(\varepsilon_1|X = x, D = 1) - E(\varepsilon_0|X = x, D = 0)$$

we would not recover either *ATE*(*X*) or *TT*(*x*). Notice too that, since the endogenous variable *D* is binary, we cannot directly recover ν and use it as a control as we did in the linear case of ex 1 above. Instead, we can recover a function of ν that satisfies the definition of a control function.
Let $F_\nu()$ denote the cumulative distribution function of ν. To form the control function in this case, first take eq. (7) and write the choice probability

$$P(x, z) \equiv \Pr(D = 1|X = x, Z = z) = \Pr(v < h(x, z)) = F_v(h(x, z)),$$

which under our assumptions implies

$$h(x, z) = F_v^{-1}(P(x, z)).$$

Following the analysis in Matzkin (1992), we can recover both *h*(*x*, *z*) and $F_\nu()$ nonparametrically up to normalization. Next, take the conditional (on *X*, *Z*) expectation of the outcome for the treated group

$$E(Y|X = x, Z = z, D = 1) = g_1(x) + E(\varepsilon_1|X = x, Z = z, D = 1).$$

We can write the last term as

$$E(\varepsilon_1|X = x, Z = z, D = 1) = E(\varepsilon_1|v < h(x, z)) = E\left(\varepsilon_1|v < F_v^{-1}(P(x, z))\right).$$

That is, we can write it as a function of the known $h(x, z)$ or, equivalently, as a function of the probability of selection $P(x, z)$,

$$E(Y|X = x, Z = z, D = 1) = g_1(x) + K_1(P(x, z)),$$

where $K_1(P(X, Z))$ satisfies our definition of a control function. So, provided that we can vary $K_1(P(X, Z))$ independently of $g_1(X)$, we can recover $g_1(X)$ up to a constant. We can identify the constant in a limit set such that $P \rightarrow 1$ since $\lim_{P\rightarrow 1} K_1(P)=0$. Provided that we have enough support in the probability of treatment – that is, provided that some people choose treatment with probability arbitrarily close to (1) –we can recover the constant. (See ex 2.) Using the same argument we can form

$$E(Y|X = x, Z = z, D = 0) = g_0(x) + K_0(P(x, z))$$

and identify $g_0(X)$ (up to a constant) and the control function $K_0(P(X, Z))$. As before, we can recover the constant in $g_0(X)$ by noting that $\lim_{P\rightarrow 0} K_0(P)=0$.

Intuitively, we need to be able to vary the $K_1(P(X, Z))$ function relative to the $g_1(X)$ function so that we can identify them from the observed variation in Y_1. One possibility is to impose that g_1 and K_1 are measurably separated functions. (That is, provided that, if $g_1(X)=K_1(P(X, Z))$ almost surely then $g_1(X)$ is a constant almost surely; see Florens, Mouchart and Rolin, 1990.) The simplest way to satisfy this restriction is by exclusion. That is, if $K_1(P(X, Z))$ is a nontrivial function of Z conditional on X and Z shows enough variation, we can vary the K_1 function by varying Z while keeping $g_1(X)$ constant. Another related possibility is to assume that g_1 and K_1 live in different function spaces. For example, g_1 a linear function and K_1 the nonlinear mills ratio term that results from assuming that $(\varepsilon_0, \varepsilon_1, v)$ are jointly normal as in the original Heckman (1979) selection correction model.

Once we have recovered $g_0(X), g_1(X), K_0(P(X, Z)), K_1(P(X, Z))$ we can now form our parameters of interest. Given $g_0(X)$ and $g_1(X)$, $ATE(X) = g_1(X) - g_0(X)$ immediately follows. To recover $TT(X)$, first notice that, by the law of iterated expectations

$$E(\varepsilon_0|X = x, Z = z) = E(\varepsilon_0|X = x, Z = z, D = 1)P(x, z) + E(\varepsilon_0|X = x, Z = z, D = 0)(1 - P(x, z)) = 0,$$

where $P(X, Z)$ is known from our analysis above and $E(\varepsilon_0|X = x, Z = z, D = 0) = K_0(P(x, z))$. Rewriting the

expression above we get $E(\varepsilon_0|X = x, Z = z, D = 1) = -\frac{K_0(P(x,z))(1-P(x,z))}{P(x,z)}$. With this expectation in hand we can recover $TT(X, Z) = g_1(X) - g_0(X) + K_1(P(X, Z)) + \frac{K_0(P(X,Z))(1-P(X,Z))}{P(X,Z)}$. By integrating against the appropriate distribution, we can recover $TT(X) = \int TT(X, z)dF_{Z|X,D=1}(z)$.

The following example shows how the control function methodology can be applied to recover average effects of treatment in a linear in parameters model with correlated random coefficients. This model arises when there are unobservable gains that vary over individuals and these gains are correlated with the choice of treatment (that is, when there is essential heterogeneity. See Heckman, Urzua and Vytlacil, 2006; Basu et al., 2006). The Roy model of eqs (5) and (6) in which the unobservable individual gains ($\varepsilon_1 - \varepsilon_0$) are correlated with the choice of sector is an example of this case.

Example 2: Correlated random coefficients with binary treatment. Assume we can write the outcome equations in linear in parameters form,

$$Y_j = \alpha_j + X\beta_j + \varepsilon_j \; j = 0, 1.$$

Let *D* be an indicator of whether an individual receives treatment (*D*=1) or not (*D*=0). We also write a linear in parameters decision rule:

$$D = 1(X\delta + Z\gamma - v > 0).$$

From the analysis in Manski (1988) we can recover δ, γ and F_v (up to scale). With $P(x, z) = \Pr(D = 1|X = x, Z = z)$ in hand, we then form

$$Y_j = \alpha_j + X\beta_j + K_j(P(X, Z)) + \eta_j$$

where $E(\eta_j|X = x, K_j(P(X, Z)) = k_j) = 0$. To emphasize the problem of identification of the constant α_j we can rewrite the outcome as

$$Y_j = \tau_j + X\beta_j + \ddot{K}_j(P(X, Z)) + \eta_j$$

where $K_j(P(X, Z)) = \kappa_j + \ddot{K}_j(P(X, Z))$ and $\tau_j = \alpha_j + \kappa_j$.

The elements of the outcome equations can be recovered by various methods. One could, for example, use Robinson (1988) and use residualized nonparametric regressions to recover β_j, τ_j and $K_j(P(X, Z))$. Alternatively, one could

approximate $K(P(X, Z))$ with a polynomial on $P(X, Z)$. In this case we would have

$$Y_j = \tau_j + X\beta_j + \pi_1 P(X, Z) + \pi_2 P(X, Z)^2 + \cdots + \pi_n P(X, Z)^n + \eta_j$$

where $\tilde{K}_j(P(X, Z)) = \Sigma_{i=1}^{n} \pi_{j1} P(X, Z)^i$. When j=0 then $\lim_{P\to 0} K_0(P)$=0 and it follows that $\tilde{K}_0(P) = K_0(P)$ and $\tau_0 = \alpha_0$. For the treated case (j=1) we have that $\lim_{P\to 1} K_1(P(X, Z)) = 0$. Since $\tilde{K}_1(1) = \Sigma_{i=1}^{n} \pi_{1i}$ it follows that $\kappa_1 = -\Sigma_{i=1}^{n} \pi_{1i}$ and $\alpha_1 = \tau_1 - \Sigma_{i=1}^{n} \pi_{1i}$.

Extensions for a continuous endogenous variable

In this section I briefly review the use of the control function approach for the case in which the endogenous variable D is continuous and we assume that $X, Z \perp\!\!\!\perp \varepsilon, \nu$. Following Blundell and Powell (2003) I assume that the object of interest is the average structural function

$$a(X, D) = \int g(X, D, \varepsilon)\, dF_\varepsilon(\varepsilon),$$

which, in the additively separable case $g(X, D, \varepsilon) = \mu(X, D) + \varepsilon$ is simply the regression function $\mu(X, D)$.
If we assume that the choice equation

$$D = h(X, Z, \nu)$$

is strictly monotonic in ν (which would follow automatically if it were additively separable in ν), we can recover $h()$ and F_ν from the analysis of Matzkin (2003) up to normalization. A convenient normalization is to assume that $\nu \sim \text{Uniform}(0, 1)$ in which case we can directly recover ν from the quantiles of F_ν, but other normalizations are possible. From the independence assumption it follows that $E(\varepsilon|X, D, Z) = E(\varepsilon|\nu)$, so we can write the outcome equation as

$$Y = \mu(X, D) + E(\varepsilon|\nu)$$

$$= \mu(X, D) + K(\nu)$$

which allows us to recover $\mu(X, D)$ directly (up to normalization). In the additively separable case we analyse, we can relax the full independence assumption and instead assume directly that the weaker mean independence assumption $E(\varepsilon|X, D, Z) = E(\varepsilon|v)$ holds.

See Also

- endogeneity and exogeneity
- identification
- Roy model
- selection bias and self-selection

Bibliography

Altonji, J.G. and Matzkin, R.L. 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73, 1053–102.

Basu, A., Heckman, J.J., Navarro, S. and Urzua, S. 2006. Use of instrumental variables in the presence of heterogeneity and self-selection: an application in breast cancer patients. Unpublished manuscript, Department of Medicine, University of Chicago.

Blundell, R. and Powell, J. 2003. Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, vol. 2, ed. L.P. Hansen, M. Dewatripont and S.J. Turnovsky. Cambridge: Cambridge University Press.

Chesher, A. 2003. Identification in nonseparable models. *Econometrica* 71, 1405–41.

Cunha, F., Heckman, J.J. and Navarro, S. 2005. Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers* 57, 191–261.

Florens, J.-P., Heckman, J.J., Meghir, C. and Vytlacil, E.J. 2007. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Unpublished manuscript, Columbia University.

Florens, J.-P., Mouchart, M. and Rolin, J.M. 1990. *Elements of Bayesian Statistics*. New York: M. Dekker.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–62.

Heckman, J.J. 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 32, 441–62. Addendum published in 33(1) (1998).

Heckman, J.J., Lochner, L.J. and Todd, P.E. 2003. Fifty years of Mincer earnings regressions. Technical Report No. 9732. Cambridge, MA: NBER.

Heckman, J.J. and Navarro, S. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* 86, 30–57.

Heckman, J.J. and Robb, R. 1985. Alternative methods for evaluating the impact of interventions: an overview. *Journal of Econometrics* 30, 239–67.

Heckman, J.J. and Robb, R. 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing Inferences from Self-Selected Samples*, ed. H. Wainer. New York: Springer. Repr. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

Heckman, J.J. and Sedlacek, G.L. 1985. Heterogeneity, aggregation, and market wage functions: an empirical model of self-selection in the labor market. *Journal of Political Economy* 93, 1077–125.

Heckman, J.J. and Smith, J.A. 1998. Evaluating the welfare state. In *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, ed. S. Strom. New York: Cambridge University Press.

Heckman, J.J., Urzua, S. and Vytlacil, E.J. 2006. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 88, 389–432.

Heckman, J.J. and Vytlacil, E.J. 1998. Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources* 33, 974–87.

Imbens, G.W. and Newey, W.K. 2006. Identification and estimation of triangular simultaneous equations models without additivity. Unpublished manuscript, Department of Economics, MIT.

Manski, C.F. 1988. Identification of binary response models. *Journal of the American Statistical Association* 83, 729–38.

Matzkin, R.L. 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica* 60, 239–70.

Matzkin, R.L. 2003. Nonparametric estimation of nonadditive random functions. *Econometrica* 71, 1393–75.

Newey, W.K., Powell, J.L. and Vella, F. 1999. Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67, 565–603.

Olley, G.S. and Pakes, A. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64, 1263–97.

Robinson, P.M. 1988. Root-n-consistent semiparametric regression. *Econometrica* 56, 931–54.

Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.

Telser, L.G. 1964. Iterative estimation of a set of linear regression equations. *Journal of the American Statistical Association* 59, 845–62.

Willis, R.J. and Rosen, S. 1979. Education and self-selection. *Journal of Political Economy* 87(5, Par 2), S7–S36.

Wooldridge, J.M. 1997. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters* 56, 129–33.

Wooldridge, J.M. 2003. Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters* 79, 185–91.

How to cite this article

Navarro, Salvador. "control functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000085> doi:10.1057/9780230226203.0311

The New Palgrave Dictionary of Economics Online

conventionalism

Lawrence A. Boland

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Conventionalism is the methodological doctrine that asserts that explanatory ideas should not be considered true or false but merely better or worse. The truth status of theories cannot be so easily dismissed. While a choice of language may be conventional, the truth status is not a matter of convenient choice. Among economists the most common practice is to avoid using the words ‘true’ (or ‘false’) when discussing models and assumptions and instead to invoke ‘best’ by using a conventionalist theory-choice truth-likeness criterion. The notion of a conventionalist theory-choice criterion presumes a philosophical necessity to choose one theory among competitors.

Keywords

Aumann, R.; conventionalism; conventions; Friedman, M.; Hume, D.; instrumentalism; Lucas, R.; mathematics and economics; methodological pluralism; methodology of economics; Popper, K.; probability calculus; problem of induction; Samuelson, P.; Simon, H.; subjective and objective probability; testing

Article

Conventionalism is the methodological doctrine that asserts that explanatory ideas should not be considered true or false but merely better or worse. At the beginning of the 20th century the status of the laws of physics was the burning issue. It was the famous philosopher Henri Poincaré who in 1902 asked whether the laws of physics were ‘only arbitrary conventions’. He answered ‘Conventions, yes; arbitrary, no’. Obviously, languages and measurement units are arbitrary conventions but nobody would seriously claim they were explanatory ideas. In Poincaré's day, the question bothering physicists who were dealing with Albert Einstein's new theory (namely, relativity) was whether the choice between Euclidian and non-Euclidian geometry was a matter of convention – that is, a matter of convenience. For everyday questions Euclidian geometry is convenient but perhaps for Einstein's physics non-Euclidian is

the better choice. For some matters, such as the choice of language to express an idea or of units to measure a distance, most people would allow that such a choice may be completely arbitrary.
Although few of them have ever heard of Poincaré, most economists will say almost the same thing whenever they make a methodological pronouncement concerning the truth status of economic theories, models or assumptions. Rarely, however, have economists been concerned with the questions raised about non-Euclidian geometry (except for John Maynard Keynes's metaphorical suggestion at the beginning of his *General Theory*). Of course, hardly any economist questions language being a matter of convenience; moreover, economists often justify the use of mathematics by claiming that its use is like that of language and thus should be judged by its convenience, not its truth status (Samuelson, 1952; 1954). But in the 1940s critics of Marshallian and Walrasian (that is, neoclassical) economics argued that the truth status of a theory's assumptions should matter. In his 1953 response to the critics of the realism of assuming perfect competition when explaining the economy, Milton Friedman advocated an alternative methodology: instrumentalism. Instrumentalism, unlike conventionalism, claims merely that the truth status of assumptions does not matter so long as the theory is useful. For those economists who still think the truth status of their theories matters, but realize that one can never prove a theory's truth status by induction, the most common response is something like Poincaré's conventionalism.
There are many examples of economists making methodological pronouncements that exhibit adherence to conventionalism. Paul Samuelson denied that any economic explanation was true, writing that ‘An explanation … is a better kind of description’ (1965, p. 1165). Obviously, some descriptions are better than others, and thus he claimed that we give the honorific title of ‘explanation’ to the best description. If one were to agree with Samuelson then one certainly could never claim that one's explanation was true. Herbert Simon chose to express this differently; he said all explanations are approximations. Specifically, he said (1963, p. 231) ‘Unreality of premises is not a virtue in scientific theory; it is a necessary evil – a concession to the finite computing capacity of the scientist that is made tolerable by the principle of continuity of approximation’. Robert Lucas agreed with that when he said ‘Any model that is well enough articulated to give clear answers to the questions we put to it will necessarily be artificial, abstract, patently “unreal”’ (1980, p. 696). Robert Aumann, the game theorist, has advocated an even more limited view for explanatory theories. As he put it ‘scientific theories are not to be considered “true” or “false”’. Going further, he said, ‘In constructing such a theory, we are not trying to get at the truth, or even to approximate to it: rather, we are trying to organize our thoughts and observations in a useful manner.’ In this regard, he argued that a theory is like ‘a filing system in an office operation, or to some kind of complex computer program’ (1985, pp. 31–2). Lucas and Aumann were merely restating Samuelson's 1965 position on methodology.

The philosophy of conventionalism

For followers of philosophers Willard Quine and Karl Popper, the truth status of explanations or theories cannot be so easily dismissed or limited. While any choice of language or units of measurement may be conventional, the truth status of theories is not a matter of choice, convenient or otherwise.
Unfortunately, the methodological doctrine of conventionalism is often confused with instrumentalism. As the philosopher Joseph Agassi (1966a) points out, they are responses to two different questions. One concerns the role of theories and the other the truth status of theories. Specifically, if we ask ‘What is the

role of a theory?', instrumentalism's answer is that theories are tools and should not be judged by epistemological standards of truth status or by conventionalist criteria of approximate truth or relative merit (except, perhaps, by simplicity or economy). Conventionalism's different answer is the one stated by Aumann: theories are filing systems or catalogues of observed data. Of course, every description is also an appeal to a filing system in that one depicts or locates it within a system by referring to other defined dimensions and concepts. If, instead, we chose the question, 'What is the *status* of a theory?', conventionalism's answer is that, of course, theories are approximations and thus should not be considered true or false but better or worse. Instrumentalism's position is simply that truth status does not matter. With this in mind, it is easy to find economists advocating both methodological positions depending on which question is asked. For example, after saying that a theory is like a filing system, Aumann goes on to say that 'We do not refer to such a system as being "true" or "untrue"; rather, we talk about whether it "works" or not, or, better yet, how well it works' (1985, p. 32).

From the perspective of the philosopher Karl Popper, the main question is: what problem is solved by the doctrine of conventionalism? Since the time when Adam Smith's friend David Hume observed that there was no logical justification for the common belief that much of our empirical knowledge was based on inductive proofs (see Russell, 1945), methodologists and philosophers have been plagued with what they call the 'problem of induction'. The paradigmatic instance of the problem of induction is the realization that we cannot provide an inductive proof that 'the sun will rise tomorrow'. This leads many of us to ask, 'So *how* do we know the sun will rise tomorrow?' If it is impossible to provide a proof, then presumably we would have to admit we do not know the answer to this burning question! Several writers have claimed to have solved this famous problem (for a discussion of such claims, see Miller, 2002). Such a claim is quite surprising since it is a problem that is impossible to solve. Nevertheless, what it is and how it is either 'solved' or circumvented is fundamental to understanding all contemporary methodological discussions.

Up to the time of Popper's entry into the discussion in the mid-1930s, most philosophers took it for granted that all claims to knowledge must be justified. Inductive arguments were seen to be the obvious method. But Popper acknowledged the problem that as a matter of simple logic an inductive argument is impossible. A logical argument is one in which, whenever all the premises are true, any logically derived statements must also be true. An inductive argument is one in which one would argue logically from the truth of particular statements (for example, observation statements such as 'the sun rose today at 7 a.m.') alone to prove the truth of a general statement (for example, the sun always rises). The 'problem of induction' would be solved if one could demonstrate the existence of such an inductive logic. The importance of this problem arises once one realizes that, without some premise of a general nature (such as we find in physics concerning the movement of the earth around the sun and earth's rotation), no finite set of observations could ever prove the non-existence of a counter-example (a refuting instance that would be denied by the general statement in question) somewhere or sometime in the future. For example, to prove that the statement 'All ravens are black' is true requires a proof that there does not exist anywhere in the universe a 'non-black raven'. Everyone agrees that one cannot provide such a literal negative proof. So, it has been argued (Boland, 1982; 2003), most discussions of methodology in economics are concerned with the problem *with* induction rather than the problem *of* induction. Conventionalism can be seen as a solution to the problem *with* induction. Conventionalism presumes that this problem can be solved even though the problem *of* induction cannot. That is, if there were an inductive logic, then the truth status of a true theory or model could in principle be provable since all

assumptions of a universal form could be inductively proven. Without such a logic, many think – still insisting that any claim to knowledge must be justified – that some other means must be found to sort through competing theories. That is, how can we choose the best from a set of competing theories? More specifically, by what criteria do we choose between competing theories? Obvious examples of such criteria are simplicity, generality, robustness, testability, falsifiability, verifiability, confirmability, operational meaningfulness, plausibility, probability, and so on. None of these criteria are considered substitutes for truth status (truth or falsity); they are only choice criteria (truthlikeness). If a criterion can be quantified, one could even see the choice as a matter of applying economics (see Boland, 1971). For example, one might choose the theory that is most confirmed – but it still must be remembered that today's most confirmed theory could be a false theory even today.

For many philosophers, such theory-choice criteria are just short-run solutions to the problem *with* induction. That is, in the short run we might be satisfied with invoking such criteria, so that we can choose between theories and thereby be able to push on, but it is hoped that in the long run someone can come up with a solution to the problem *of* induction.

Conventionalism as employed by economists

Among economists who openly practise conventionalism, it is a doctrine with many variants and relatives. The most common practice is the avoidance of using the words ‘true’ (or ‘false’) when discussing theories, models and assumptions. Instead, we see ‘best’ being invoked with the use of some conventionalist theory-choice truth-likeness criterion. Also common is the use of the word ‘valid’ to avoid saying ‘true’. Sometimes it is used to mean that a theory is valid if it is logically consistent with available data or evidence. The difficulty is that ‘valid’ is a question of the logicality of an argument (do the conclusions necessarily follow from the assumptions made?) A logically valid argument can still be false, so it is not always clear what is meant by a valid statement or a valid theory.

One weak form conventionalism is old-fashioned relativism. Another weak form is what the followers of McCloskey (1983) call modernism. In yet another weak form it can be seen to be the rationale for so-called methodological pluralism. The most common form is stronger in that it involves the notion that theories are to be evaluated or compared by means of some form of probability calculus.

Those adherents to conventionalism who advocate the objective form of probability calculus seem unaware of the logical difficulties involved. One might wish to use probability as the measure of confirmation of a theory so that one could use such a measure as the criterion for theory choice. The difficulty arises when one asks what constitutes positive evidence – namely, evidence to be used to calculate the probability measure that would serve as, say, the ‘degree of confirmation’. Of course, if one requires all observational evidence to be exactly true, then to be an actual confirmation the objective probability measure would have to be 1.00. That is, just one true observation that contradicts the theory in question requires the rejection of the theory. So it would seem that objective probability measures are inappropriate. But econometrics-based hypothesis testing is not as strict since it allows for errors in the observations of the variables. Hence, the objective probability measure can be of some value less that 1.00. Theory choice in this case would seem to be a simple matter of choosing the theory with the highest probability, that is, the highest degree of confirmation.

Among those who openly advocate a subjective form of probabilities, the most common view is based on Bayesian probabilities which provide a compromise by allowing for explicit roles for both

subjectivism in the form of prior probability assessments and objectivism in the form of adjustments based on new objective evidence. Again, the main question for using probabilities concerns what would count as confirming evidence or evidence that increases the subjective probability. Like all confirmation criteria, even if everyone attaches a high subjective probability to the theory in question being true it could still be false and perhaps refuted by the next observation report.

The common element underlying all probability measures to be used for theory choice is the notion that the number of confirming observations should somehow matter. Of course, such an expectation does not require the questionable use of probabilities as a measure of confirmation. But avoiding any reliance on probability will not circumvent the more well-known logical problems of confirmation. All conceptions of a logical connection between positive evidence and degrees of confirmation suffer from a profound logical problem called, by some philosophers, the ‘paradox of confirmation’ or the ‘paradox of the ravens’ (cf. Sainsbury, 1995; Agassi, 1966b).

The philosopher's paradox of confirmation merely points out that *any* evidence which does not refute a simple universal statement, say, ‘All ravens are black’ must increase the degree of confirmation. The paradox is based on the observation that, *in terms of what observable evidence would count*, this example of a simple universal statement is logically equivalent to its ‘contra-positive’ statement ‘All non-black things are non-ravens’. Any true observation that is consistent with one of the statements is consistent with the other (equivalent) statement. But in these terms it must be recognized that positive evidence consistent with the contra-positive statement includes red shoes as well as white swans – since in both cases we have non-black things which are not ravens. That is, the set of all confirming instances must include all things which are not non-black ravens. In other words, the more red shoes we observe, the more evidence there is in favour of the contra-positive statement – that is, a red shoe increases the universal statement's degree of confirmation – and, since the contra-positive statement is logically equivalent to the universal statement in question, the latter's degree of confirmation also increases. Obviously, this consideration merely divides the contents of the universe into non-black ravens and everything else (Hempel, 1966). This consideration calls into question all claims of confirmation.

Few economists who make pronouncements concerning the appropriate methodology to use in economics are aware of the philosophical problems involved. Almost all think we must have some criterion to choose between competing theories or models. All of them take for granted the necessity of justifying their choice. No recognition seems to be given to the simple fact that one's favourite theory can be true even though it cannot be proven true. That is, whether one's theory is true is a separate question from how one knows it to be true.

The notion of a conventionalist theory-choice criterion presumes that there is a philosophical necessity to choose one theory from among its competitors. But there is no such necessity, even though it will always be difficult to convince economists of this whenever they are naive concerning the philosophy of science. But, given that there are so many different criteria to use, one would think any theory that is best by all criteria should be the chosen theory. But it is doubtful that any theory could satisfy all criteria; so the question is begged as to which criterion is the best criterion. This question seems to put us on the road of an infinite regress: by what criterion do we choose the best criterion to choose between theories? Not a promising journey.

See Also

- assumptions controversy
- instrumentalism and operationalism
- pluralism in economics

Bibliography

Agassi, J. 1966a. Sensationalism. *Mind* 75, 1–24.

Agassi, J. 1966b. The mystery of the ravens: discussion. *Philosophy of Science* 33, 395–402.

Aumann, R. 1985. What is game theory trying to accomplish? In *Frontiers of Economics*, ed. K. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

Boland, L. 1971. Methodology as an exercise in economic analysis. *Philosophy of Science* 38, 105–17.

Boland, L. 1982. *The Foundations of Economic Method*. London: Allen & Unwin.

Boland, L. 2003. *The Foundations of Economic Method: A Popperian Perspective*. London: Routledge.

Friedman, M. 1953. Methodology of positive economics. In *Essays in Positive Economics*. Chicago: Univ. of Chicago Press.

Hempel, C. 1966. *Foundations of Natural Science*. Englewood Cliffs: Prentice-Hall.

Keynes, J.M. 1936. *General Theory of Employment, Interest and Money*. New York: Harcourt, Brace & World.

Lucas, R. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12, 696–715.

McCloskey, D. 1983. The rhetoric of economics. *Journal of Economic Literature* 21, 481–517.

Miller, D. 2002. Induction: a problem solved. In *Karl Poppers kritischer Rationalismus heute*, ed. J. Böhm, H. Holweg and C. Hoock. Tübingen: Mohr Siebeck.

Poincaré, H. 1902. *La science et l'hypothèse*. Paris: Flammarion.

Poincaré, H. 1905. *Science and Hypothesis*. London: Walter Scott Publishing Company.

Russell, B. 1945. *A History of Western Philosophy*. New York: Simon and Schuster.

Sainsbury, R. 1995. *Paradoxes*. Cambridge: Cambridge University Press.

Samuelson, P. 1952. Economic theory and mathematics: an appraisal. *American Economic Review* 42, 56–66.

Samuelson, P. 1954. Some psychological aspects of mathematics and economics. *Review of Economics and Statistics* 36, 380–2.

Samuelson, P. 1965. Professor Samuelson on theory and realism: reply. *American Economic Review* 55, 1164–72.

Simon, H. 1963. Problems of methodology: discussion. *American Economic Review, Proceedings* 53, 229–31.

How to cite this article

Boland, Lawrence A. "conventionalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000568> doi:10.1057/9780230226203.0312

The New Palgrave Dictionary of Economics Online

convergence

Steven N. Durlauf and Paul A. Johnson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

One of the most widely studied empirical questions in the new growth economics concerns the role of initial conditions in affecting long-run outcomes. The statistical formulation of this dependence is known as convergence. This article surveys empirical work on convergence, with emphasis on the relationships between conventional definitions of convergence, the main statistical frameworks of evaluating convergence, and various economic models.

Keywords

Cass–Koopmans growth model; Cobb–Douglas functions; cointegration; convergence; endogenous growth; Galton's fallacy; growth nonlinearities; identification; income distribution; literacy rates; neoclassical growth theory; production functions; Solow growth model; statistics and economics; technical change; time series analysis; regression tree

Article

The general question of convergence, understood as the tendency of differences between countries to disappear over time, is of long-standing interest to social scientists. In the 1950s and early 1960s, many analysts discussed whether capitalist and socialist economies would converge over time, in the sense that market institutions would begin to shape socialist economies just as government regulation and a range of social welfare policies grew in capitalist ones.

In modern economic parlance, convergence usually refers specifically to issues related to the persistence or transience of differences in per capita output between economic units, be they countries, regions or states. Most research has focused on convergence across countries, since the large contemporaneous differences between countries generally dwarf intra-country differences. In the context of economic growth, the convergence hypothesis arguably represents the most commonly studied aspect of growth, although the effort to identify growth determinants is arguably the main area of contemporary growth

research.

In this overview of convergence, our primary emphasis will be on the development of precise statistical definitions of convergence. This reflects an important virtue of the current literature, namely, the introduction of statistical methods to adjudicate whether convergence is present. At the same time, there is no single definition of convergence in the literature, which is one reason why empirical evidence on convergence is indecisive. Our discussion focuses on convergence across countries, which has dominated empirical studies, although there is reference to studies that focus on other units.

β-convergence

The primary definition of convergence used in the modern growth literature is based on the relationship between initial income and subsequent growth. The basic idea is that two countries exhibit convergence if the one with lower initial income grows faster than the other. The local (relative to steady state) dynamics of the neoclassical growth model in both its Solow and Cass–Koopmans variants imply that lower-income economies will grow faster than higher-income ones.

As a statistical question, this notion of convergence can be operationalized in the context of a cross-country regression. Let g_i denote real per capita growth of country i across some fixed time interval and $y_{i,0}$ denote the initial per capita income for country i. Then, unconditional β-convergence is said to hold if, in the regression

$$g_i = k + \log y_{i,0}\beta + \varepsilon_i,\ \beta < 0. \tag{1}$$

For cross-country regression analysis, one typically does not find unconditional β-convergence unless the sample is restricted to very similar countries, for example, members of the OECD. This finding is in some ways not surprising, since unconditional β-convergence is typically not a prediction of the existing body of growth theories. The reason for this is that growth theories universally imply that growth is determined by factors other than initial income. While different theories may propose different factors, they collectively imply that (1) is misspecified. As a result, most empirical work focuses on conditional β-convergence. Conditional β-convergence holds if $\beta < 0$ for the regression

$$g_i = k + \log y_{i,0}\beta + Z_i\gamma + \varepsilon_i \tag{2}$$

where Z_i is a set of those growth determinants that are assumed to affect growth in addition to a country's initial income. While many differences exist in the choice of controls, it is nearly universal to

include those determinants predicted by the Solow growth model, that is, population growth and human and physical capital accumulation rates.

Unlike unconditional β-convergence, evidence of conditional β-convergence has been found in many contexts. For the cross-country case, the basic finding is generally attributed to Barro (1991), Barro and Sala-i-Martin (1992) and Mankiw, Romer and Weil (1992). The Mankiw, Romer and Weil analysis is of particular interest as it is based on a regression suggested by the dynamics of the Solow growth model. Hence, their findings have been widely interpreted as evidence in favour of decreasing returns to scale in capital (the source of $\beta < 0$ in the Solow model), and therefore as evidence against the Lucas–Romer endogenous growth approach, which emphasizes increasing returns in capital accumulation (either human or physical) as a source of perpetual growth.

From the perspective of the neoclassical growth model, the term $-\beta$ also measures the rate at which an economy's convergence towards its steady-state growth rate, that is, the growth rate determined exclusively by the exogenous rate of technical change. The many findings in the cross-country literature are often summarized by the claim countries converge towards their steady-state growth rates at a rate of about two per cent per year, although individual studies produce different results. The convergence rate has received inadequate attention in the sense that a finding of convergence may have little consequence for questions such as policy interventions if it is sufficiently slow.

As is clear from (2), any claims about conditional convergence necessarily depend on the choice of control variables Z_i. This is a serious concern given the lack of consensus in growth economics on which growth determinants are empirically important. Doppelhofer, Miller and Sala-i-Martin (2004) and Fernandez, Ley and Steel (2001) use model averaging methods to show that the cross-country findings that have appeared for conditional β-convergence are robust to the choice of controls. A number of additional statistical issues such as the role of measurement error and endogeneity of regressors are surveyed and evaluated in Durlauf, Johnson and Temple (2005).

The assumption in cross-section growth regressions that the unobserved growth terms ε_i are uncorrelated with $\log y_{i,0}$ rules out the possibility that there are country-specific differences in output levels; if such effects were present, they would imply a link between the two. For this reason, a number of researchers have investigated convergence using panel data. This leads to models of the form

$$g_{i,t} = c_i + \log y_{i,t-1}\beta + Z_{i,t}\gamma + \varepsilon_{i,t} \tag{3}$$

where growth is now measured between $t-1$ and t. This approach not only can handle fixed effects, but can allow for instrumental variables to be used to address endogeneity issues. Panel analyses have been conducted by Caselli, Esquivel and Lefort (1996), Islam (1995) and Lee, Pesaran and Smith (1997). These studies have generally found convergence with rather higher rates than appear in the cross-section studies; for example, Caselli, Esquivel and Lefort (1996) report annual convergence rate estimates of ten per cent.

As discussed in Durlauf and Quah (1999) and Durlauf, Johnson and Temple (2005), panel data

approaches to convergence suffer from the problem that, once country specific effects are allowed, it becomes more difficult to interpret results in terms of the underlying economics. The problem is that, once one allows for fixed effects, then the question of convergence is changed, at least if the goal is to understand whether initial conditions matter; simply put, the country-specific effects are themselves a form of initial conditions. When studies such as Lee, Pesaran and Smith (1997) allow for rich forms of parameter heterogeneity across countries, β-convergence become equivalent to the question of whether there is some mean reversion in a country's output process, not whether certain types of contemporaneous inequalities diminish. This does not diminish the interest of these studies as statistical analyses, but means their economic import can be unclear.

σ-convergence and the cross-section distribution of income

A second common statistical measure of convergence focuses on the whether or not the cross-section variance of per capita output across countries is or is not shrinking. A reduction in this variance is interpreted as convergence. Letting $\sigma^2_{\log y,t}$ denote the variance across *i* of log•$y_{i,t}$, σ-convergence occurs between *t* and $t+T$ if

$$\sigma^2_{\log y,t} - \sigma^2_{\log y,t+T} > 0. \tag{4}$$

There is no necessary relationship between β- and σ-convergence. For example, if the first difference of output in each country obey $\log y_{i,t} - \log y_{i,t-1} = \beta \log y_{i,t-1} + \varepsilon_{i,t}$, then $\beta < 0$ is compatible with a constant cross-sectional variance (which in this example will equal the variance of log•$y_{i,t}$). The incorrect idea that mean reversion in time series implies that its variance is declining is known as Galton's fallacy; its relevance to understanding the relationship between convergence concepts in the growth literature was identified by Friedman (1992) and Quah (1993a). While it is possible to construct a cross-section regression to test for σ-convergence (cf. Cannon and Duck, 2000), they do not test β-convergence per se.

Work on β-convergence has led to general interest in the evolution of the cross-country income distribution. Quah (1993b; 1996) has been very influential in his modelling of a stochastic process for the distribution itself, with the conclusion that it is converging towards a bimodal steady-state distribution. Other studies of the evolution of the cross-section distribution include Anderson (2004) who uses nonparametric density methods to identify increasing polarization between rich and poor economies across time. Increasing divergence between OECD and non-OECD economies is shown in Maasoumi, Racine and Stengos (2007), working with residuals from linear growth regressions.

One difficulty with convergence approaches that emphasize changes in the shape of the cross-section distribution is that they may fail to address the original question of the persistence of contemporaneous

inequality. The reason for this is that it is possible, because of movements in relative position within the distribution, for the cross-section distribution to flatten out while at the same time differences at one point in time are reversed; similarly, the cross-section distribution can become less diffuse while gaps between rich and poor widen. That being said, an examination of the locations of individual countries in various distribution studies typically indicates that the increasing polarization of the world income distribution is mirrored by increasing gaps between rich and poor. A useful extension of this type of research would be to employ the dynamics of individual countries to provide additional information on how the cross-section distribution evolves.

Time series approaches to convergence

An alternative approach to convergence is focused on direct evaluation of the persistence of transitivity of per capita output differences between economies. This approach originates in Bernard and Durlauf (1995), who equate convergence with the statement that

$$\lim_{T \Rightarrow \infty} E(\log y_{i,t+T} - \log y_{j,t+T} | F_t) = 0 \tag{5}$$

where F_t denotes the history of the two output series up to time t. They find that convergence does not hold for OECD economies, although there is some cointegration in the individual output series. Hobijn and Franses (2000) find similar results for a large international data-set. Evans (1996) employs a clever analysis of the evolution of the cross-section variance to evaluate the presence of a common trend in OECD output, and finds one is present; his analysis allows for different deterministic trends in output and so in this sense is compatible with Bernard and Durlauf (1995).

The relationship between cross-section and time series convergence tests is complicated. Bernard and Durlauf (1996) argue that the two classes of tests are based on different assumptions about the data under study. Cross-section tests assume that countries are in transition to a steady state, so that the data for a given country at time t is drawn from a different stochastic process from the data at some future $t + T$. In contrast, time series tests assume that the underlying stochastic processes are time-invariant parameters, that is, that countries have transited to an invariant output process. They further indicate how convergence under a cross-section test can in fact imply a failure of convergence under a time series test, because of these different assumptions. For these reasons, time series tests of convergence seem appropriate for economies that are at similar stages and advanced stages of development.

From statistics to economics

The various concepts of convergence we have described are all purely statistical definitions. The economic questions that motivated these definitions are not, however, equivalent to these questions, so it is important to consider convergence as an economic concept in order to assess what is learned in the

statistical studies. As argued in Durlauf, Johnson and Temple (2005), the economic questions that underlie convergence study revolve around the respective roles of initial conditions versus structural heterogeneity in explaining differences in per capita output levels or growth rates. It is the permanent effect of initial conditions, not structural features that matters for convergence. If we define initial conditions as $\rho_{i,0}$ and the structural characteristics as $\theta_{i,0}$, convergence can be defined via

$$\lim_{t \to \infty} E(\log y_{i,t} - \log y_{j,t} | \rho_{i,0}, \theta_{i,0}, \rho_{j,0}, \theta_{j,0}) = 0 \text{ if } \theta_{i,0} = \theta_{j,0}. \quad (6)$$

The gap between the definition (6) and the statistical tests that have been employed is evident when one considers whether the statistical tests can differentiate between economically interesting growth models, some of which fulfil (6) and others of which do not. One such contrast is between the Solow growth model and the Azariadis and Drazen (1990) model of threshold externalities, in which countries will converge to one of several possible steady states, with initial conditions determining which one emerges. By definition (6), the Solow model produces convergence whereas the Azariadis–Drazen model does not. However, as shown by Bernard and Durlauf (1996) it is possible for data from the Azariadis–Drazen model to produce estimates that are consistent with a finding of β-convergence.

There is in fact a range of empirical findings of growth nonlinearities that are inconsistent with convergence in the sense of (6). Durlauf and Johnson (1995) is an early study of this type, which explicitly estimated a version of the Azariadis–Drazen model in which the Solow model, under the assumption of a Cobb–Douglas aggregate production function, is a special case. Durlauf and Johnson rejected the Solow model specification and found multiple growth regimes indexed by initial conditions. Their findings are consistent with the presence of convergence clubs in which different groups of countries are associated with one of several possible steady states. These results are confirmed by Papageorgiou and Masanjala (2004) using a CES production function specification.

The Durlauf and Johnson analysis uses a particular classification procedure, known as a regression tree, to identify groups of countries obeying a common linear model. Other statistical approaches have also identified convergence clubs. For example, Bloom, Canning and Sevilla (2003) use mixture distribution methods to model countries as associated with one of two possible output processes, and conclude that individual countries may be classified into high-output manufacturing- and service-based economies and low-output agriculture-based economies. Canova (2004) uses Bayesian methods to identify convergence clubs for European regions.

As discussed in Durlauf and Johnson (1995) and Durlauf, Johnson and Temple (2005), studies of nonlinearity also suffer from identification problems with respect to questions of convergence. One problem is that a given data-set cannot fully uncover the full nature of growth nonlinearities without strong additional assumptions. As a result, it becomes difficult to extrapolate those relationships between predetermined variables and growth to infer steady-state behaviour. Durlauf and Johnson give an example of a data pattern that is compatible with both a single steady and multiple steady states. A second problem concerns the interpretation of the conditioning variables in these exercises. Suppose one

finds, as do Durlauf and Johnson, that high- and low-literacy economies are associated with different aggregate production functions. One interpretation of this finding is that the literacy rate proxies for unobserved fixed factors, for example culture, so that these two sets of economies will never obey a common production function, and so will never exhibit convergence in the sense of (6). Alternatively, the aggregate production function could structurally depend on the literacy rate, so that, as literacy increases, the aggregate production functions of currently low-literacy economies will converge to those of the high-literacy ones. Data analyses of the type that have appeared cannot distinguish between these possibilities.

Conclusions

While the empirical convergence literature contains many interesting findings and has helped identify a number of important generalizations about cross-country growth behaviour, it has yet to reach any sort of consensus on the deep economic questions for which the statistical analyses were designed. The fundamentally nonlinear nature of endogenous growth theories renders the conventional cross-section convergence tests inadequate as ways to discriminate between the main classes of theories. Evidence of convergence clubs may simply be evidence of deep nonlinearities in the transitional dynamics towards a unique steady state. Cross-section and time series approaches to convergence not only yield different results but are predicated on different views of the nature of transitory versus steady-state behaviour of economies, differences that themselves have yet to be tested.

None of this is to say that convergence is an empirically meaningless question. Rather, progress requires continued attention to the appropriate statistical definition of convergence and the use of statistical procedures consistent with the definition. Further, it seems important to move beyond current ways of assessing convergence both in terms of better use of economic theory and by a broader view of appropriate data sources. Graham and Temple (2006) illustrate the potential for empirical analyses of convergence that employ well-delineated structural models. The research programme developed in Acemoglu, Johnson and Robinson (2001; 2002) provides a perspective on the micro-foundations of country-specific heterogeneity that speaks directly to the convergence question and which shows the power of empirical analysis based on careful attention to economic history. For these reasons, research on convergence should continue to be productive and important.

See Also

- economic growth, empirical regularities in
- endogenous growth theory
- neoclassical growth theory
- neoclassical growth theory (new perspectives)

Bibliography

Acemoglu, D., Johnson, S. and Robinson, J. 2001. The Colonial origins of comparative development: an

empirical investigation. *American Economic Review* 91, 1369–401.

Acemoglu, D., Johnson, S. and Robinson, J. 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 117, 1231–94.

Anderson, G. 2004. Making inferences about the polarization, welfare, and poverty of nations: a study of 101 countries 1970–1995. *Journal of Applied Econometrics* 19, 530–50.

Azariadis, C. and Drazen, A. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105, 501–26.

Barro, R. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics* 106, 407–43.

Barro, R. and Sala-i-Martin, X. 1992. Convergence. *Journal of Political Economy* 100, 223–51.

Bernard, A. and Durlauf, S. 1995. Convergence in international output. *Journal of Applied Econometrics* 10(2), 97–108.

Bernard, A. and Durlauf, S. 1996. Interpreting tests of the convergence hypothesis. *Journal of Econometrics* 71, 1–2, 161–73.

Bloom, D., Canning, D. and Sevilla, J. 2003. Geography and poverty traps. *Journal of Economic Growth* 8, 355–78.

Canova, F. 2004. Testing for convergence clubs in income per capita: a predictive density approach. *International Economic Review* 45, 49–77.

Cannon, E. and Duck, N. 2000. Galton's fallacy and economic convergence. *Oxford Economic Papers* 53, 415–19.

Caselli, F., Esquivel, G. and Lefort, F. 1996. Reopening the convergence debate: a new look at cross country growth empirics. *Journal of Economic Growth* 1, 363–89.

Doppelhofer, G., Miller, R. and Sala-i-Martin, X. 2004. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94, 813–35.

Durlauf, S. and Johnson, P. 1995. Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics* 10, 365–84.

Durlauf, S., Johnson, P. and Temple, J. 2005. Growth econometrics. In *Handbook of Economic Growth*,

ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Durlauf, S. and Quah, D. 1999. The new empirics of economic growth. In *Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Evans, P. 1996. Using cross-country variances to evaluate growth theories. *Journal of Economic Dynamics and Control* 20, 1027–49.

Fernandez, C., Ley, E. and Steel, M. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16, 563–76.

Friedman, M. 1992. Do old fallacies ever die? *Journal of Economic Literature* 30, 2129–32.

Graham, B. and Temple, J. 2006. Rich nations, poor nations: how much can multiple equilibria explain? *Journal of Economic Growth* 11, 5–41.

Hobijn, B. and Franses, P. 2000. Asymptotically perfect and relative convergence of productivity. *Journal of Applied Econometrics* 15, 59–81.

Islam, N. 1995. Growth empirics: a panel data approach. *Quarterly Journal of Economics* 110, 1127–70.

Lee, K., Pesaran, M. and Smith, R. 1997. Growth and Convergence in multi country empirical stochastic Solow model. *Journal of Applied Econometrics* 12, 357–92.

Maasoumi, E., Racine, J. and Stengos, T. 2007. Growth and convergence: a profile of distribution dynamics and mobility. *Jounal of Econometrics* 136(2) 483–508.

Mankiw, N., Romer, D. and Weil, D. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107, 407–37.

Papageorgiou, C. and Masanjala, W. 2004. The Solow model with CES technology: nonlinearities with parameter heterogeneity. *Journal of Applied Econometrics* 19, 171–201.

Quah, D. 1993a. Galton's fallacy and tests of the convergence hypothesis. *Scandinavian Journal of Economics* 95, 427–43.

Quah, D. 1993b. Empirical cross-section dynamics in economic growth. *European Economic Review* 37, 426–34.

Quah, D. 1996. Convergence empirics across economies with (some) capital mobility. *Journal of Economic Growth* 1, 95–124.

How to cite this article

Durlauf, Steven N. and Paul A. Johnson. "convergence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000534> doi:10.1057/9780230226203.0313

The New Palgrave Dictionary of Economics Online

convex programming

Lawrence E. Blume
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes the basic ideas of convex optimization in finite-dimensional vector spaces. Duality, the Fenchel transforms and the subdifferential are introduced and used to discuss Lagrangean duality and the Kuhn–Tucker theorem. Applications of these ideas can be found in duality.

Keywords

concave optimization; conjugate duality th; convex optimization; convex programming; convexity; duality; Fenchel transform; hyperplanes; Kuhn–Tucker th; Lagrange multipliers; monotonicity; quasi-concavity; saddlepoints; separation th

Article

1 Introduction

Firms maximize profits and consumers maximize preferences. This is the core of microeconomics, and under conventional assumptions about decreasing returns it is an application of convex programming. The paradigm of convex optimization, however, runs even deeper through economic analysis. The idea that competitive markets perform well, which dates back at least to Adam Smith, has been interpreted since the neoclassical revolution as a variety of conjugate duality for the primal optimization problem of finding Pareto-optimal allocations. The purpose of this article and the companion article duality is (in part) to explain this sentence. This article surveys without proof the basic mathematics of convex sets and convex optimization with an eye towards their application to microeconomic and general equilibrium theory, some of which can be found under duality.

Unfortunately there is no accessible discussion of concave and convex optimization outside textbooks and monographs of convex analysis such as Rockafellar (1970; 1974). Rather than just listing theorems, then, this article attempts to provide a sketch of the main ideas. It is certainly no substitute for the sources. This article covers only convex optimization in finite-dimensional vector spaces. While many of these ideas carry over to infinite-dimensional vector spaces and to important applications in infinite horizon economies and economies with non-trivial uncertainty, the mathematical subtleties of infinite-dimensional topological vector spaces raise issues which cannot reasonably be treated here. The reader looking only for a statement of the Kuhn–Tucker theorem is advised to read backwards from the end, to find the theorem and notation.

A word of warning. This article is written from the perspective of constrained maximization of concave functions because this is the canonical problem in microeconomics. Mathematics texts typically discuss the constrained minimization of convex functions, so textbook treatments will look slightly different.

2 Convex sets

A subset C of a Euclidean vector space V is convex if it contains the line segment connecting any two of its members. That is, if

x and y are vectors in C and t is a number between 0 and 1, the vector $tx+(1-t)\,y$ is also in C. A linear combination with non-negative weights which sum to 1 is a *convex combination* of elements of C; a set C is convex if it contains all convex combinations of its elements.

The key fact about convex sets is the famous *separation th*. A linear function p from the vector space V to $\mathbf{R}$ and a real number a define a *hyperplane*, the solutions to the equation $p\cdot x=a$. Every hyperplane divides V into two *half-spaces*; the upper (closed) half-space, containing those vectors x for which $p\cdot x\geq a$, and the lower (closed) half-space, containing those vectors x for which $p\cdot x\leq a$. The separation theorem uses linear functionals to describe closed convex sets. If a given vector is not in a closed convex set, then there is a hyperplane such that the set lies strictly inside the upper half-space while the vector lies strictly inside the lower half-space:

Separation theorem: If C is a closed convex set and x is not in C, then there is a linear functional p and a real number a such that $p\cdot y>a$ for all $y\in C$, and $p\cdot x<a$.

This theorem implies that every closed convex set is the intersection of the half-spaces containing it. This half-space description is a *dual* description of closed convex sets, since it describes them with linear functionals. From the separation theorem the existence of a *supporting hyperplane* can also be deduced. If x is on the boundary of a closed convex set C, then there is a (non-zero) linear functional p such that $p\cdot y\geq p\,x$ for all $y\in C$; p is the hyperplane that supports C at x.

The origin of the term ‘duality’ lies in the mathematical construct of the dual to a vector space. The *dual space* of a vector space V is the collection of all *linear functionals*, that is, real-valued linear functions, defined on V. The distinction between vector spaces and their duals is obscured in finite dimensional spaces because each such space is its own dual. If an n-dimensional Euclidean vector space is represented by column vectors of length n, the linear functionals are $1\times n$ matrices; that is the dual to $\mathbf{R^n}$ is $\mathbf{R^n}$. (This justifies the notation used above.) Self-duality (called reflexivity in the literature) is not generally true in infinite-dimensional spaces, which is reason enough to avoid discussing them here. Nonetheless, although V will be $\mathbf{R^n}$ throughout this article, the usual notation V^* will be used to refer to the dual space of V simply because it is important to know when we are discussing a vector in V and when we are discussing a member of its dual, a linear functional on V.

If the weights in a linear combination sum to 1 but are not constrained to be non-negative, then the linear combination is called an *affine combination*. Just as a convex set is a set which contains all convex combinations of its elements, an affine set in a vector space V is a set which contains all affine combinations of its elements. The set containing all affine combinations of elements in a given set C is an affine set, A(C). The purpose of all this is to define the *relative interior* of a convex set C, ri C. The relative interior of a convex set C is the interior of C relative to A(C). A line segment in $\mathbf{R}^2$ has no interior, but its relative interior is everything on the segment but its endpoints.

3 Concave functions

The neoclassical assumptions of producer theory imply that production functions are concave and cost functions are convex. The quasi-concave functions which arise in consumer theory share much in common with concave functions, and quasi-concave programming has a rich duality theory.

In convex programming it is convenient to allow concave functions to take on the value $-\infty$ and convex functions to take on the value $+\infty$. A function f defined on $\mathbf{R^n}$ with range $[-\infty,\infty)$ is concave if the set $\{(x,a): a\in, a\leq f(x)\}$ is convex. This set, a subset of $\mathbf{R^{n+1}}$, is called the *hypograph* of f and is denoted hypo f. Geometrically, it is the set of points in $\mathbf{R^{n+1}}$ that lie on or below the graph of f. Similarly, the *epigraph* of f is the set of points in $\mathbf{R^{n+1}}$ that lie on or above the graph of f: epi $f=\{(x,a): a\in, a\geq f(x)\}$. A function f with range $(-\infty,\infty]$ is convex $-f$ is concave, and convexity of f is equivalent to convexity of the set epi f. Finally, the *effective domain* of a concave function is the set dom $f=\{x\in\mathbf{R^n}: f(x)>-\infty\}$, and similarly for a convex function. Those familiar with the literature will note that attention here is restricted to *proper* concave and convex functions. Functions that are everywhere $+\infty$ will also be considered concave, and those everywhere $-\infty$ will be assumed convex when Lagrangeans are discussed below.

Convex optimization does not require that functions be differentiable or even continuous. Our main tool is the separation theorem, and for that closed convex sets are needed. A concave function f is *upper semi-continuous* (usc) if its hypograph is closed; a convex function is *lower semi-continuous* (lsc) if its epigraph is closed. Upper and lower semi-continuity apply to any functions, but these concepts interact nicely conveniently with convex and concave functions. In particular, usc concave and lsc convex functions are continuous on the relative interiors of their domain. A famous example of a usc concave function that fails to be continuous is $f(x,y)=-y^2/2x$ for $x>0$, 0 at the origin and $-\infty$ otherwise. Along the curve $y=\sqrt{\alpha x}$, $y\to 0$ as $x\to 0$, but f is constant at $-\alpha/2$, so f is not continuous at (0,0), but it is usc because the supremum of the limits at the origin is 0.

It is useful to know that, if f is concave and usc, then $f(x)=\inf q(x)=a\cdot x+b$ where the infimum is taken over all a and b such that $a\cdot x+b$ is everywhere at least as big as f. This is another way of saying that, since hypo f is closed, it is the intersection of all half-spaces containing it.

4 The Fenchel transform

The concave Fenchel transform associates with each usc function on a Euclidean space V, not necessarily concave, a usc concave function on its dual space V^* (which, we recall, happens to be V since its dimension is finite). The adjective 'concave' is applied because a similar transform is defined slightly differently for convex functions. The concave Fenchel transform of f is

$$f^*(p) = \inf_{x\in V}\left\{p\cdot x - f(x)\right\},$$

which is often called the *conjugate* of f. (From here on out we will drop the braces.) The conjugate f^* of f is concave because, for fixed x, $p\cdot x-f(x)$ is linear, hence concave, in p, and the pointwise infimum of concave functions is concave. The textbooks all prove that, if hypo f is closed, so is hypo f^*, that is, upper semi-continuity is preserved by conjugation. So what is this transformation doing, and why is it interesting?

The conjugate f^* of a concave function f describes all the non-vertical half-spaces containing hypo f. This should be checked. A half-space in $\mathbf{R}^{n+1}$ can be represented by the inequality $(p, q)\,(x, y)\geq a$ where q is a real number (as is a) and $p\in V^*$. The half-space is non-vertical if $p^*\neq 0$. In $\mathbf{R}^2$ this means geometrically that the line defining the boundary of the half-space is not vertical. So choose a linear functional $p\neq 0$ in V^*. For any $(x, z)\in$hypo f, and any $p\in V^*$,

$$p\cdot x - z \geq p\cdot x - f(x) \geq \inf_{x\in V} p\cdot x - f(x) = f^*(p).$$

In other words, the upper half-space $(p, -1)\cdot(x, z) \geq f^*(p)$ contains hypo f. It actually supports hypo f because of the infimum operation: If $a > f^*(p)$, there is an $(x, z)\in$hypo f such that $p\,x-z<a$, so the upper half-space fails to contain hypo f.

Before seeing what the Fenchel transform is good for, we must answer an obvious qst. If it is good to transform once, why not do it again? Define

$$f^{**}(x) = \inf_{p\in V^*} p\cdot x - f^*(p),$$

the *double dual* of f. The fundamental fact about the Fenchel transform is the following theorem, which is the function version of the dual descriptions of closed convex sets.

Conjugate duality theorem: If f is usc and concave, then $f^{**} = f$.

This is important enough to explain. Notice that just as p is a linear functional acting on x, so x is a linear functional acting on p. Suppose that f is concave and usc. For all x and p, $p\cdot x - f(x) \geq f^*(p)$, and so $p\cdot x - f^*(p) \geq f(x)$. Taking the infimum on the left, $f^{**}(x) \geq f(x)$.

On the other hand, take a $p\in V^*$ and a real number b such that the half-space $(p, -1)\cdot(x, z)\geq b$ in $\mathbf{R}^{n+1}$ contains the hypograph

of *f*. This is true if and only if $p \cdot x - b \geq f(x)$ for all *x* and because *f* is usc, $f(x)$ is the infimum of *p*·*x*–*b* over all such *p* and *b*. Since $p \cdot x - f(x) \geq b$ for all *x*, take the infimum on the left to conclude that $f^*(p) \geq b$. Thus $p \cdot x - b \geq p \cdot x - f^*(p)$, and taking the infimum now on the right, $p \cdot x - b \geq f^{**}(x)$. Taking the infimum on the left over all the *p* and *b* such that the half-space contains hypo *f*, $f(x) \geq f^{**}(x)$.

It is worthwhile to compute an example to get the feel of the concave Fenchel transform. If *C* is a closed convex set, the *concave indicator function* of *C* is $\phi(x)$ which is 0 for *x* in *C*, and $-\infty$ otherwise. This is a good example to see the value of allowing infinite values. The Fenchel transform of ϕ is $\varphi^*(p) = \inf_{x \in \mathbf{R}^n} p \cdot x - \varphi(x)$. Clearly the infimum cannot be reached at any $x \notin C$, for the value of ϕ at such an *x* is $-\infty$, and so the value of $p \cdot x - \phi(x)$ is $+\infty$. Consequently $\varphi^*(p) = \inf_{x \in C} p \cdot x$. This function has the enticing property of positive homogeneity: If *t* is a positive scalar, then $\varphi^*(tp) = t\varphi^*(p)$.

Compute the double dual, first for $x \notin C$. The separating hyperplane theorem claims the existence of some *p* in V^* and a real number *a* such that $p \cdot x < a \leq p \cdot y$ for all $y \in C$. Take the infimum on the right to conclude that $p \cdot x < \varphi^*(p)$, which is to say that $p \cdot x - \varphi^*(p) < 0$. Then, multiply both sides by an arbitrary positive scalar *t* to conclude that $tp \cdot x - \varphi^*(tp)$ can be made arbitrarily negative. Hence $\varphi^{**}(x) = -\infty$ if $x \notin C$. And if *x* is in *C*? Then $p \cdot x - 0 \geq \varphi^*(p)$ for all *p* (recall $\phi(x)=0$). So $p \cdot x - \varphi^*(p) \geq 0$. But $\phi^*(0)=0$, so $\varphi^{**}(x)$, the infimum of the left-hand side over all possible *p* functionals, is 0. Thus the Fenchel transform of ϕ^* recovers ϕ.

A particularly interesting version of this problem is to suppose that *C* is an 'at least as good as' set for level *u* of some upper semi-continuous and quasi-concave utility function (or, more generally, a convex preference relation with closed weak upper contour sets). Then $\phi^*(p)$ is just the minimum expenditure necessary to achieve utility *u* at price *p*. See duality for more discussion. Another interesting exercise is to apply the Fenchel transform to concave functions which are not usc, and to non-concave functions. These constructions have important applications in optimization theory which we will not pursue.

The theory of convex functions is exactly the same if, rather than the *concave Fenchel transform*, the *convex Fenchel transform* is employed: $\sup_{x \in \mathbf{R}^n} p \cdot x - f(x)$. This transform maps convex lsc functions on *V* into convex lsc functions on V^*. Both the concave and convex Fenchel transforms will be important in what follows.

5 The subdifferential

The separation theorem applied to hypo *f* implies that usc concave functions have tangent lines: For every $x \in$ ri dom *f* there is a linear functional p_x such that $f(y) \leq f(x) + p_x(y - x)$. This inequality is called the *subgradient inequality*, and p_x is a *subgradient* of *f*; p_x defines a tangent line for the graph of *f*, and the graph lies on or underneath it. The set of subgradients of *f* at $x \in$ dom *f* is denoted $\partial f(x)$, and is called the *subdifferential* of *f* at *x*. Subdifferentials share many of the derivative's properties. For instance, if $0 \in \partial f(x)$, then *x* is a global maximum of *f*. In fact, if $\partial f(x)$ contains only one subgradient p_x, then *f* is differentiable at *x* and $Df(x) = p_x$. The set $\partial f(x)$ need not be single-value, however, because *f* may have kinks. The graph of the function *f* defined on the real line such that $f(x) = -\infty$ for $x<0$ and $f(x) = \sqrt{x}$ for $x \geq 0$ illustrates why the subdifferential may be empty at the boundary of the effective domain. At 0, a subgradient would be infinitely steep.

There is a corresponding *subgradient inequality* for convex $f: f(y) \geq f(x) + p_x \cdot (y - x)$. With these definitions, $\partial(-f)(x) = -\partial f(x)$. Note that some texts refer to superdifferentials for concave functions and subdifferentials for convex functions. Others do not multiply the required terminology, and we follow them.

The multivalued map $x \mapsto \partial f(x)$, is called the *subdifferential correspondence* of *f*. An important property of subdifferential correspondences is monotonicity. From the subgradient inequality, if $p \in \partial f(x)$ and $q \in \partial f(y)$, then $f(y) \leq f(x) + p \cdot (y - x)$ and $f(x) \leq f(y) + q \cdot (x - y)$, and it follows that $(p-q) \cdot (x-y) \leq 0$. For convex *f* the inequality is reversed.

The Fenchel transforms establish a clear relationship between the subdifferential correspondences of concave functions and their duals. If *f* is concave, then the subdifferential inequality says that $p \in \partial f(x)$ if and only if for all $z \in X$, $p \cdot x - f(x) \leq p \cdot z - f(z)$. The map $z \mapsto p \cdot z - f(z)$ is minimized at $z=x$, and so *p* is in $\partial f(x)$ if and only if $f^*(p) = p \cdot x - f(x)$. If *f* is usc, then $f^{**} \equiv f$, and so $f^{**}(x) = f(x) = p \cdot x - f^*(x)$. That is, $p \in \partial f(x)$ if and only if

$x \in \partial f^*(p)$.

6 Optimization and duality

Economics most often presents us with constrained maximization problems. Within the class of problems with concave objective functions, there is no formal difference between constrained and unconstrained maximization. The constrained problem of maximizing concave and usc *f* on a closed convex set *C* is the same as the unconstrained problem of maximizing $f(x) + I_C(x)$ on $\mathbf{R}^n$, where $I_C(x)$ is the concave indicator function of *C*.

The general idea of duality schemes in optimization theory is to represent maximization (or minimization) problems as half of a minimax problem which has a saddle value. There are several reasons why such a seemingly odd construction can be useful. In economics it often turns out that the other half of the minimax problem, the dual problem, sheds additional light on properties and inpts of the primal problem. This is the source of the 'shadow price' concept: The shadow price is the value of relaxing a constraint. Perhaps the most famous example of this is the Second Theorem of Welfare Economics.

6.1 Lagrangeans

The *primal problem* (problem *P*) is to maximize *f*(*x*) on a Euclidean space *V*. Suppose there is a function $L: V \times V^* \to \mathbf{R}$ such that $f(x) = \inf_{p \in V^*} L(x, p)$. Define $g(p) = \sup_{x \in V} L(x, p)$, and consider the problems of maximizing *f*(*x*) on *V* and minimizing *g*(*p*) on V^*. The first problem is the primal problem, and the second is called the *dual problem*. For all *x* and *p* it is clear that $f(x) \le L(x, p) \le g(p)$, and thus that

$$\sup_x \inf_p L(x, p) = \sup_x f(x) \le \inf_p g(p) = \inf_p \sup_x L(x, p).$$

If the inequality is tight, that is, it holds with equality, then the common value is called a *saddle value* of *L*. In particular, a saddle value exists if there is a *saddle point* of *L*, a pair (x^*, p^*) such that for all $x \in V$ and $p \in V^*$, $L(x, p^*) \le L(x^*, p^*) \le L(x^*, p)$. A pair (x^*, p^*) is a saddlepoint if and only if x^* solves the primal, p^* solves the dual, and a saddle value exists. The function *L* is the *Lagrangean*, which is familiar from the analysis of smooth constrained optimization problems. Here it receives a different foundation.

The art of duality schemes is to identify an interesting *L*, and here is where the Fenchel transforms come in. Interesting Lagrangeans can be generated by embedding the problem max *f* in a parametric class of concave maximization problems. Suppose that there is a (Euclidean) parameter space *P*, and a usc and concave function $F: V \times Y \to$ such that $f(x) = F(x, 0)$, and consider all the problems $\max_{x \in V} F(x, y)$. A particularly interesting object of study is the *value function* $\varphi(y) = \sup_x F(x, y)$, which is the indirect utility function in consumer theory, and the cost function in the theory of the firm (with concave replaced by convex and max by min). The map $y \mapsto -F(x, y)$ is closed and convex for each *x*, so define on $V \times V^*$

$$L(x, p) = \sup_y p \cdot y + F(x, y),$$

its (convex) Fenchel transform. The map $p \mapsto L(x, p)$ is closed and convex on V^*. Transform again to see that $F(x, y) = \inf_{p \in V^*} L(x, p) - p \cdot y$. In particular, $f(x) = \inf_p L(x, p)$.

An example of this scheme is provided by the usual concave optimization problem given by a concave objective function *f*, *K* concave constraints $g_k(x) \ge 0$, and an implicit constraint $x \in C$: $\max_x f(x)$ subject to the constraints $g_k(x) \ge 0$ for all *k* and $x \in C$. Introduce parameters *y* so that $g_k(x) \ge y_k$, and define $F(x, y)$ to be $f(x)$ if all the constraints are satisfied and $-\infty$ otherwise. The supremum defining the Lagrangean cannot be realized for *y* such that *x* is infeasible, and so

$$L(x, p) = \begin{cases} f(x) + \sum_k p_k g_k(x) & \text{if } x \in C \text{ and } p \in \mathbf{R}_+^{\mathbf{K}}, \\ +\infty & x \in C \text{ and } p \notin \mathbf{R}_+^{\mathbf{K}}, \\ -\infty & \text{if } x \notin C \end{cases} \tag{1}$$

if there are feasible x, then $F(x,y)$ is everywhere $-\infty$, and so $L(x, p) \equiv -\infty$.
Here, in summary, are the properties of the Lagrangean for the problems discussed here:
Lagrangean theorem: If $F(x, y)$ is lsc and concave then (1) the Lagrangean L is lsc and convex in p for each $x \in V$, (2) L is concave in x for each $p \in V^*$, and (3) $f(x) = \inf_p L(x, p)$.
Following the original scheme, the objective for the dual problem is $g(p) = \sup_x L(x, p)$, and the *dual problem* (problem D) is to maximize g on V^*. Perhaps the central fact of this dual scheme is the relationship between the dual objective function g and the value function ϕ. The function ϕ is easily seen to be concave, and simply by writing out the definitions, one sees that $g(p) = \sup_y p \cdot y + \varphi(y)$, the convex Fenchel transform of the convex function $-\phi$. So $g(p)$ is lsc and convex, $g(p) = (-\varphi)^*(p)$ and whenever ϕ is usc, $\inf_p g(p) = \varphi(0)$.
To make the duality scheme complete, the min problem should be embedded in a parametric class of problems in a complementary way. Take $G(p, q) = \sup_{x \in V} L(x, p) - q \cdot x$, so that $g(p) = G(p, 0)$. With this definition, $-G(p, q) = \inf_{x \in V} q \cdot x - L(x, p)$, the concave Fenchel transform of $x \mapsto L(x, p)$. The value function for the parametric class of minimization problems is $\gamma(q) = \inf_p G(p, q)$. The relationship between F and G is computed by combining the definitions:

$$G(p, q) = \sup_{x, y} F(x, y) - q \cdot x + p \cdot y = -\inf_{x, y} q \cdot x - p \cdot y - F(x, y) = -F^*(q, -y) \text{ and so } F(x, y) = \inf_{p, q} G(p, q) + q \cdot x - p \cdot y$$

where the F^* is the concave Fenchel transform of the map $(x, y) \mapsto F(x, y)$. Computing from the definitions, $f(x) = \inf_{p, q} q \cdot x + G(p, q) = \inf_q q \cdot x + \gamma(q)$, so $f = (-\gamma)^*$, and whenever γ is lsc, $\sup_x f(x) = \gamma(0)$.
In summary, if $F(x, y)$ is concave in its arguments, and usc, then we have constructed a Lagrangean and a dual problem of minimizing a concave and lsc $G(p, q)$ over p. If the value functions $\varphi(y)$ and $\gamma(q)$ are usc and lsc, respectively, then $\sup_x F(x, 0) = \gamma(0)$ and $\inf_p G(p, 0) = \varphi(0)$, so a saddle value exists. Upper and lower semi-continuity of the value functions can be an issue. The hypograph of ϕ is the set of all pairs (y, a) such that $\sup_x F(x, y) \geq a$, and this is the projection onto y and a of the set of all triples (x, y, a) such that $F(x, y) \geq a$, that is, hypo F. Unfortunately, even if hypo F is closed, its projection may not be, so upper semi-continuity of ϕ does not follow from the upper semi-continuity of F.
In the constrained optimization problem with Lagrangean (1), the parametric class of dual minimization problems is to minimize $G(p, q) = \sup_{x \in C} f(x) + \Sigma_k p_k g_k(x) - q \cdot x$ if $y \in \mathbf{R}_+^{\mathbf{K}}$ and $+\infty$ otherwise. Specialize this still further by considering linear programming. The canonical linear program is to $\max a \cdot x$ subject to the explicit constraints $b_k \cdot x \leq c_k$ and the implicit constraint $x \geq 0$. Rewrite the constraints as $-b_k \cdot x + c_k \geq 0$ to be consistent with the formulation of (1). Then

$$G(p, q) = \sup_{x \geq 0} a \cdot x - \sum_k p_k (b_k \cdot x - c_k) - q \cdot x = \sum_k c_k p_k + \sup_{x \geq 0} \left(a - \sum_k p_k b_k - q \right) \cdot x$$

for $p \in \mathbf{R}^{\mathbf{K}}_{+}$ and +∞ otherwise. The dual problem is to minimize this over *p*. The sup term in *G* will be +∞ unless the vector in parentheses is non-positive, in which case the sup will be 0. So the dual problem, taking *q*=0, is to minimize $\Sigma_k c_k p_k$ over *p* subject to the constraints that $\Sigma_k p_k b_k \geq a$ and $p \in \mathbf{R}^{\mathbf{K}}_{+}$. If the primal constraints are infeasible, $\varphi(0) = -\infty$. If the dual is infeasible, $\gamma(0) = +\infty$, and this serves as an example of how the dual scheme can fail over lack of continuity. For linear programs there is no problem with the hypographs of ϕ and γ, because these are *polyhedral convex sets*, the intersection of a finite number of closed half-spaces, and projections of closed polyhedral convex sets are closed.

6.2 Solutions

Subdifferentials act like partial derivatives, particularly with respect to identifying maxima and minima: x^* in *V* solves problem *P* if and only if $0 \in \partial f(x^*)$. When *f* is identically $-\infty$, there are no solutions which satisfy the constraints. Thus dom *f* is the set of *feasible solutions* to the primal problem *P*. Similarly, $p^* \in V^*$ solves the dual problem *D* if and only if $\partial g(p^*) = 0$, and here dom *g* is the set of *dual-feasible solutions*. Saddlepoints of the Lagrangean also have a subdifferential characterization. Adapting the obvious partial differential notion and notation, (x^*, p^*) is a saddle point for L if and only if $0 \in \partial_x L(x^*, p^*)$ and $0 \in \partial_p L(x^*, p^*)$ (these are different 0's since they live in different spaces), which we write $(0, 0) \in \partial L(x^*, p^*)$. This condition is often called the *Kuhn–Tucker* condition. The discussion so far can be summarized in the following theorem, which is less general than can be found in the sources:
Kuhn–Tucker theorem: Suppose that *F*(*x*, *y*) is concave and usc. Then the following are equivalent:

1. 1. sup *f*=inf *g*,
2. 2. ϕ is usc and concave,
3. 3. the saddle value of the Lagrangean *L* exists,
4. 4. γ is lsc and convex.
   In addition, the following are equivalent:
5. 5. x^* solves *P*, p^* solves *D*, and the saddle value of the Lagrangean exists.
6. 6. (x^*, p^*) satisfy the Kuhn–Tucker condition.
   For economists, the most interesting feature of the dual is that it often describes how the value of the primal problem will vary with parameters. This follows from properties of the subdifferential and the key relation between the primal value function and the dual objective function, $g = (-\varphi)^*$: $-\partial\varphi(0) = \partial(-\varphi)(0)$, and this equals the set $\{p: (-\varphi)^*(p) = p \cdot 0 - (-\varphi)(0)\}$, and this is precisely the set $\{p: g(p) = \sup_x f(x)\}$. In words, if *p* is a solution to the dual problem *D*, then –*p* is in the subgradient of the primal value function. When $\partial\phi(0)$ is a singleton, there is a unique solution to the dual, and it is the derivative of the value function with respect to the parameters. More generally, from the subdifferential of a convex function one can construct directional derivatives for particular changes in parameter values. Similarly, $-\partial\gamma(0) = \{x: f(x) = \inf_p g(p)\}$, with an identical inpt. In summary, add to the Kuhn–Tucker theorem the following equivalence:
7. 7. $-p^* \in \partial\varphi(0)$ and $-x^* \in \partial\gamma(0)$

The remaining question is, when is any one of these conditions satisfied? A condition guaranteeing that the subdifferentials are non-empty is that 0 ∈ ri dom ϕ, since concave functions always have subdifferentials on the relative interior of their effective domain. In the constrained optimization problem whose Lagrangean is described in (1), an old condition guaranteeing the existence of saddlepoints is the *Slater condition*, that there is an *x*∈ri *C* such that for all *k*, $g_k(x)>0$. This condition implies that 0 ∈ ri dom ϕ, because there is an open neighbourhood around 0 such that for y in the neighbourhood and for all *k*, $g_k(x)>y_k$. Thus $\varphi(y) \geq F(x, y) > -\infty$ for all *y* in the neighbourhood. Conditions like this are called *constraint qualifications*. In the standard calculus approach to constrained optimization, they give conditions under which derivatives sufficiently characterize the constraint set for calculus approximations to work (see Arrow, Hurwicz and Uzawa, 1961).
Finally, it is worth noting that infinite dimensional constrained optimization problems, such as those arising in dynamic

economic models and the study of uncertainty, can be addressed with extensions of the methods discussed here. The main difficulty is that most infinite dimensional vector spaces are not like $\mathbf{R}^n$. There is no ‘natural’ vector space topology, and which topology one chooses has implications for demonstrating the existence of optima. The existence of separating hyperplanes is also a difficulty in infinite dimensional spaces. These and other problems are discussed in Mas-Colell and Zame (1991). Nonetheless, much of the preceding development does go through. See Rockafellar (1974).

See Also

- convexity
- duality
- Lagrange multipliers
- quasi-concavity

Bibliography

Arrow, K., Hurwicz, L. and Uzawa, H. 1961. Constraint qualifications in maximization problems. *Naval Logistics Research Quarterly* 8, 175–91.

Mas-Colell, A. and Zame, W. 1991. Equilibrium theory in infinite dimensional spaces. In *Handbook of Mathematical Economics*, vol. 4, ed. W. Hildenbrand and H. Sonnenschein. Amsterdam: North-Holland.

Rockafellar, R.T. 1970. *Convex Analysis*. Princeton, NJ: Princeton University Press.

Rockafellar, R.T. 1974. *Conjugate Duality and Opttimization*. CBMS Regional Conference Series No. 16. Philadelphia: SIAM.

How to cite this article

Blume, Lawrence E. "convex programming." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000348> doi:10.1057/9780230226203.0314

The New Palgrave Dictionary of Economics Online

convexity

Lawrence E. Blume

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

convexity; duality; existence of equilibrium; hyperplanes; large economies; Lyapunov's theorem; optimization; quasi-concavity; quasi-convexity; Shapley–Folkman theorem

Article

Convexity is the modern expression of the classical law of diminishing returns, which was prominent in political economy from Malthus and Ricardo through the neoclassical revolution. Its importance today rests less on any utilitarian or behavioural psychological rationale or physical principle than on its utility as a tool of mathematical analysis. In general equilibrium and game theory, proofs of the existence of equilibrium, competitive and Nash, respectively, rely on the application of a fixed-point theorem to a set-valued, convex-valued map from a convex set to itself. Welfare economics provides another example: The second theorem of welfare economics, which asserts that optimal allocations can be supported by competitive prices, relies on an application of the supporting hyperplane theorem to an appropriate convex set.

Convexity is a property of real vector spaces, and its domain of application in economic analysis is not just Euclidean spaces but also the infinite dimensional vector spaces which arise in the study of uncertainty and dynamics, where infinite numbers of goods are required. Nonetheless, this brief exposition will be confined to Euclidean spaces.

Definitions

A set $C \subset \mathbf{R}^n$ is *convex* if the line segment connecting any two points in C lies wholly within C. Formally put, C is convex if and only if for all points x and y in C and all scalars t in the unit interval [0, 1], the point $tx+(1-t)y$ is also in C. A ball is convex; a boomerang is not. An extended real-valued function f defined on a convex set $C \subset \mathbf{R}^n$ is *convex* if its *epigraph* or *supergraph*, $\{(x, \mu): x \in C, \mu \in \mathbf{R}, \ f(x) \leq \mu\}$ is convex. For real-valued functions, this is equivalent to the more familiar definition that for all x and y

in C and $t \in [0,1]$, $f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$. A function f is *concave* if $-f$ is convex.

Optimization

Students of economics first encounter convexity in the study of optimization. If $x^* \in \mathbf{R^n}$ is a critical point of a smooth function, and if x^* a local maximum, then the Hessian matrix at x^*, the matrix of second-order partial derivatives, must be negative semi-definite; that is, it is locally concave. Any critical point with a negative definite Hessian must be a local maximum. Negative definiteness of the Hessian implies but is not implied by strict (local) concavity. For Jevons, utility was additively separable, and so the principle of diminishing marginal utility itself was enough to derive concavity. Edgeworth, the first economist to consider non-separable utility functions, realized that diminishing marginal utility was not, in general, enough to guarantee convexity. His development of demand theory relied on a differential condition that can be shown to imply *quasi-concavity*. A real-valued function f with a convex domain C is *quasi-concave* if for each real number α, the set $\{x \in C: f(x) \geq \alpha\}$ is convex. To appreciate the difference between concavity and quasi-concavity, note that any strictly increasing function on the real line is quasi-concave. The differential description of convexity and its variants (quasi-convexity, pseudo-convexity) and the associated necessary and sufficient second-order conditions for constrained optimization problems has produced a volume of analysis, most of which is of second-order importance to contemporary economic theory. Exhaustive coverage can be found in Simon and Blume (1994).

Duality

The representation of consumers by expenditure functions and firms by profit functions is said to be ‘dual’ to the ‘primal’ representations by preferences and production sets, respectively. These representations rely on alternative ways of representing closed convex sets: The ‘primal’ description is a list of its elements, and the ‘dual’ description is the list of closed half-spaces containing it. The dual representation for closed convex sets is equivalent to the separating hyperplane theorem: If x in $\mathbf{R^n}$ is not in a closed convex set C, then there is a hyperplane $H \subset \mathbf{R^n}$ with x on one side and C on the other. That is, there is a $p \in \mathbf{R^n}$ and a number α such that $p \cdot x < \alpha$ and $p \cdot y > \alpha$ for all $y \in C$. (See convex programming and duality.)

Large numbers and convexity

Convexity is sometimes an inappropriate assumption. Half a box of two left shoes and half a box of two right shoes is surely preferred to either box, but the 50:50 mixture of a good burgundy and a good stout is only a headache. Fortunately, the analysis of perfectly competitive markets rests not on the preferences of any individual consumer, but on the average behaviour of a large number of consumers. A central insight behind much research of the 1970s and 1980s (and which was anticipated by Edgeworth, 1881, a century before) is that averaging is a convexifying operation. This is the content of the Shapley–Folkman theorem as applied to large finite economies, and Lyapunov's theorem in the analysis of economies with a continuum of agents. (See cores, large economies and perfect competition.)

For economies with large numbers of small consumers and small firms, the important analytical constructs are approximately convex. With respect to the existence of equilibrium and its welfare properties, large economies look like convex economies. Hildenbrand (1974) is an entry point to this important body of research.

See Also

- convex programming
- cores
- duality
- large economies
- perfect competition

Bibliography

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: C. Kegan Paul & Co.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Simon, C. and Blume, L. 1994. *Mathematics for Economists*. New York: W.W. Norton & Co.

How to cite this article

Blume, Lawrence E. "convexity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000508> doi:10.1057/9780230226203.0315

The New Palgrave Dictionary of Economics Online

convict labour

Farley Grubb
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

colonialism; convict labour; international migration; labour contracts; migration, international

Article

Some European countries banished convicts to labour in overseas colonies – sometimes using private markets to transport and employ this labour.
Punishing felons who did not warrant execution and were too poor to pay monetary fines posed a dilemma for early modern societies. The long-standing punishments of one-off physical chastisements, such as whippings, increasingly seemed too barbaric and returned malefactors to society too quickly. While long-term incarceration was more civilized and removed malefactors from society, penitentiaries were expensive to build and operate, and the criminal's labour was lost to society. Sentencing felons to labour in overseas colonies thus became an attractive solution.
Between 1854 and 1920 France sent between 20,000 and 30,000 convicts to French Guiana and New Caledonia. Spain sent convicts to North Africa, Cuba, and Puerto Rico. Britain, however, was the largest participant, sending 6,000–10,000 convicts to its colonies between 1614 and 1718 and another 50,000 mostly to its American colonies Virginia and Maryland between 1718 and 1775 (Coldham, 1992; Ekirch, 1987). After the United States closed its shores to British convicts, convict transportation was shifted to Australia where approximately 160,000 were landed between 1787 and 1868 (Nicholas, 1988). Another 18,000 were shipped to Bermuda and Gibraltar.
The Transportation Act of 1718 shifted the overseas banishment of British felons from a case-by-case petitioning of the Crown to a routine sentence imposed by courts. The sentences allowed were seven years, 14 years, or a lifetime of banishment –74 per cent, 24 per cent and two per cent, respectively, of those transported – which became the length of the convict's overseas labour contract. Most transported convicts were guilty of property crimes and were Englishmen. Between 13 per cent and 23 per cent were Irish, and between 10 per cent and 15 per cent were female (Ekirch, 1987). Sentences were not rigidly tied to crimes; for example, highway robbers received 7-year, 14-year, and lifetime sentences – 38 per

cent, 50 per cent, and 12 per cent, respectively (Grubb, 2000). Not until convicts had completed their sentences could they return to Britain without facing being hanged if caught.

The privatization of overseas convict disposal reached its zenith after the Transportation Act. The government minimized its cost of overseas convict disposal by channelling convicts through the existing competitive markets for voluntary servant labour, where emigrants traded forward-labour contracts to shippers for passage to America. Shippers recouped their cost by selling these contracts (emigrants) to private employers in America. Potential shipping profits related to labour heterogeneity were arbitraged away by bargaining over contract length. The typical voluntary servant negotiated a four-year labour contract and sold in America for eight and half pounds sterling. By contrast, courts fixed the length of convict sentences (labour contracts) independently of labour heterogeneity. Convicts were then transferred to private shippers for transportation overseas. Shippers sold their convicts as servant labour to private employers in America to recoup their shipping expense. The average convict sold for 11 pounds sterling (Grubb, 2000).

By fixing contract lengths – the parameter used to arbitraged shipping profits in the voluntary servant market – the courts altered the convict auction price distribution and profit arbitrage process from that which existed in the voluntary servant market. The distribution of convict contract prices had a higher mean, higher standard deviation, and lower kurtosis than that of voluntary servant contract prices. Shippers did not earn excess profits on convicts. The higher sale price was matched by the higher cost of chaining convicts during shipment and paying variable fees charged by county jailers. Jailers played shippers off against each other for access to convict cargo. The government subsidized one shipper in the London market who earned, net of political bribes, excess profits (Grubb, 2000).

Shippers carried both voluntary and convict servants concurrently. Potential employers were shown the conviction papers that stated each convict's sentence and crime. Post-auction convicts were largely indistinguishable from voluntary servants. Most were employed in agriculture and at iron forges alongside slaves and voluntary servants. They lived in their employer's house and ate at their employer's table. Criminal conviction, however, carried a stigma that led to price discounts. A year's worth of convict labour sold for a 21 per cent discount on average over that of comparable voluntary servant labour. Convicts guilty of more serious and professional crimes, such as arsonists and receivers of stolen goods, sold for even greater discounts. Convicts also ran away more often than did voluntary servants: 16 per cent versus six per cent, respectively (Grubb, 2001).

Per given crime, a 14-year versus a 7-year sentence signalled the courts' perception of the severity of harm inflicted by, and incorrigibility of, the convict. American employers responded to this information by demanding additional price discounts of 48 per cent and 68 per cent per year of labour for convicts sentenced to 14 years and to life, respectively, as opposed to seven years for the same crime. Employers also paid premiums and received discounts for certain convict attributes, other things equal. For example, taller convicts sold for a substantial premium, and female convicts with venereal disease sold for 19 per cent less than females without the disease (Grubb, 2001).

For underpopulated colonies lacking competitive labour markets, such as Australia, European governments typically had to transport convicts to the colonies themselves, directly employing them on government projects there (Nicholas, 1988). During the 19th century, European governments also became increasingly reluctant to use existing competitive markets to auction convict labour for fear that it would look like government-sanctioned slavery. Instead, convicts were transferred via bureaucratic

petition or assignment systems. Under these conditions, the system struggled to employ convict labour efficiently and to be a cost-effective punishment. Convict transportation waned as social reformers succeeded in replacing it with incarceration in newly built penitentiaries and as maturing colonies increasingly resisted being convict dumping-grounds.

See Also

- auctions (empirics)
- compensating differentials
- human capital, fertility and growth
- indentured servitude
- international migration
- labour market institutions

Bibliography

Coldham, P. 1992. *Emigrants in Chains*. Baltimore, MD: Genealogical Publishing.

Ekirch, A. 1987. *Bound for America: The Transportation of British Convicts to the Colonies, 1718–1775*. New York: Oxford University Press.

Grubb, F. 2000. The transatlantic market for British convict labor. *Journal of Economic History* 60, 94–122.

Grubb, F. 2001. The market evaluation of criminality: evidence from the auction of British convict labor in America, 1767–1775. *American Economic Review* 91, 295–304.

Nicholas, S., ed. 1988. *Convict Workers: Reinterpreting Australia's Past*. New York: Cambridge University Press.

How to cite this article

Grubb, Farley. "convict labour." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000558> doi:10.1057/9780230226203.0316

The New Palgrave Dictionary of Economics Online

cooperation

Samuel Bowles and Herbert Gintis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We review game-theoretic models of cooperation with self-regarding agents. We then study the folk theorem in large groups of self-regarding individuals with imperfect information. In contrast to the dyadic case with perfect information, the level of cooperation deteriorates with larger group size and higher error rates. Moreover, no plausible account exists of how the dynamic, out-of-equilibrium behaviour of these models would support cooperative outcomes. We then analyse cooperation with other-regarding preferences, finding that a high level of cooperation can be attained in large groups and with modest informational requirements, and that conditions allowing the evolution of such social preferences are plausible.

Keywords

cooperation; focal rules; folk theorem; game theory; general equilibrium; multiple equilibria; Prisoner's Dilemma; private information; public goods game; reciprocity, indirect; reciprocity, strong; repeated games; reputation maintenance; retaliation; social preferences; strategic interaction; subgame perfection; tit for tat

Article

Cooperation is said to occur when two or more individuals engage in joint actions that result in mutual benefits. Examples include the mutually beneficial exchange of goods, the payment of taxes to finance public goods, team production, common pool resource management, collusion among firms, voting for income redistribution to others, participating in collective actions such as demonstrations, and adhering to socially beneficial norms.

A major goal of economic theory has been to explain how wide-scale cooperation among self-regarding individuals occurs in a decentralized setting. The first thrust of this endeavour involved Walras's general equilibrium model, culminating in the celebrated ‘invisible hand’ theorem of Arrow and Debreu (Arrow and Debreu, 1954; Debreu, 1959; Arrow and Hahn, 1971). But, the assumption that contracts could completely specify all relevant aspects of all exchanges and could be enforced at zero cost to the exchanging parties is not applicable to many important forms of cooperation. Indeed, such economic institutions as firms, financial institutions, and state agencies depend on incentive mechanisms involving strategic interaction in addition to explicit contracts (Blau, 1964; Gintis, 1976; Stiglitz, 1987; Tirole, 1988; Laffont, 2000).

The second major thrust in explaining cooperation eschewed complete contracting and developed sophisticated repeated game-theoretic models of strategic interaction. These models are based on the insights of Shubik (1959), Taylor (1976), Axelrod and Hamilton (1981) and others that repetition of social interactions plus retaliation against defectors by withdrawal of cooperation may enforce cooperation among self-regarding individuals. A statement of this line of thinking, applied towards understanding the broad historical and anthropological sweep of human experience is the work of Ken Binmore (1993; 1998; 2005). For Binmore, a society's moral rules are instructions for behaviour in conformity with one of the myriad of Nash equilibria of a repeated *n*-player social interaction. Because the interactions are repeated, and these rules form a Nash equilibrium, the self-regarding individuals who comprise the social order will conform to the moral rules.

We begin by reviewing models of repeated dyadic interaction in which cooperation may occur among players who initially cooperate and in the next round adopt the action of the other player in the previous round, called *tit for tat*. These models show that as long as the probability of game repetition is sufficiently great and individuals are sufficiently patient, a cooperative equilibrium can be sustained once it is implemented. This reasoning applies to a wide range of similar strategies. We then analyse *reputation maintenance* models of dyadic interaction, which are relevant when individuals interact with many different individuals, and hence the number of periods before a repeat encounter with any given individual

may be too great to support the tit-for-tat strategy.

We then turn to models of cooperation in larger groups, arguably the most relevant case, given the scale on which cooperation frequently takes place. The folk theorem (Fudenberg and Maskin, 1986) shows that, in groups of any size, cooperation can be maintained on the assumption that the players are sufficiently future-oriented and termination of the interaction is sufficiently unlikely. We will see, however, that these models do not successfully extend the intuitions of the dyadic models to many-person interactions. The reason is that the level of cooperation that may be supported in this way deteriorates as group size increases and the probability of either behavioural or perceptual error rises, and because the theory lacks a plausible account of how individuals would discover and coordinate on the complicated strategies necessary for cooperation to be sustained in these models. This difficulty bids us investigate how other-regarding preferences, *strong reciprocity* in particular, may sustain a high level of cooperation, even with substantial errors and in large groups.

Repetition allows cooperation in groups of size two

Consider a pair of individuals who play the following *stage game* repeatedly: each can *cooperate* (that is, help the other) at a cost $c>0$ to himself, providing a benefit to the other of $b>c$. Alternatively, each player can *defect*, incurring no cost and providing no benefit. Clearly, both would gain by cooperating in the stage game, each receiving a net gain of $b-c>0$. However, the structure of the game is that of a *Prisoner's Dilemma*, in which a self-regarding player earns higher payoff by defecting, no matter what his partner does.

The behaviour whereby each individual provides aid as long as this aid has been reciprocated by the other in the previous encounter, is called *tit for tat*. Although termed ‘reciprocal altruism’ by biologists, this behaviour is self-regarding, because each individual's decisions depend only on the expected net benefit the individual enjoys from the long-term relationship.

On the assumption that after each round of play the interaction will be continued with probability δ, and that players have discount factor d (so $d=1/(1+r)$, where r is the rate of time preference), then provided

$$\delta d b > c, \tag{1}$$

each individual paired with a tit-for-tat player does better by cooperating (that is, playing tit for tat) rather than by defecting. Thus tit for tat is a best response to itself. To see this, let $\mathbf{v}$ be the present value of cooperating when paired with a tit-for-tat player. Then

$$\mathbf{v} = b - c + \delta d \mathbf{v}, \tag{2}$$

which gives

$$\mathbf{v} = \frac{b-c}{1-\delta d}. \tag{3}$$

The present value of defecting for ever on a tit-for-tat playing partner is b (the single period gain of b being followed by zero gains in every subsequent period as a result of the tit-for-tat player's defection), so playing tit-for-tat is a best response to itself if and only if $(b-c)(1-\delta d)>b$, which reduces to (1). Under these conditions unconditional defect is also a best response to itself, so either cooperation or defection can be sustained.

But suppose that, instead of defection for ever, the alternative to tit for tat is for a player to defect for a certain number of rounds, before returning to cooperation on round $k>0$. The

payoff to this strategy against tit for tat is $b-(\delta d)^k c+(\delta d)^{k+1}\mathbf{v}$. This payoff must not be greater than **v** if tit for tat is to be a best response to itself. It is an easy exercise in algebra to show that the inequality

$$\mathbf{v} \geq b-(\delta d)^k c+(\delta d)^{k+1}\mathbf{v}$$

simplifies to (1), no matter what the value of k. A similar argument shows that when (1) holds, defecting for ever (that is, $k=\infty$) does not have a higher payoff than cooperating.

Cooperation through reputation maintenance

Tit for tat takes the form of frequent repetition of the Prisoner's Dilemma stage game inducing a pair of self-regarding individuals to cooperate. In a sizable group, an individual may interact frequently with a large number of partners but infrequently with any single one, say on the average of once every k periods. Players then discount future gains so that a payoff of **v** in k periods from now is worth $d^k\mathbf{v}$ now. Then, an argument parallel to that of the previous section shows that cooperating is a best response if and only if

$$\frac{b-c}{1-\delta d^k} > b$$

which reduces to

$$\delta d^k b > c. \tag{4}$$

Note that this is the same equation as (1) except that the effective discount factor falls from d to d^k. For sufficiently large k, it will not pay to cooperate. Therefore, the conditions for tit-for-tat reciprocity will not obtain.

But cooperation may be sustained in this situation if each individual keeps a mental model of exactly which group members cooperated in the previous period and which did not. In this case, players may cooperate in order to cultivate a *reputation for cooperation*. When individuals tend to cooperate with others who have a reputation for cooperation, a process called *indirect reciprocity* can sustain cooperation. Let us say that an individual who cooperated in the previous period *in good standing*, and specify that the only way an individual can fall into *bad standing* is by defecting on a partner who is in good standing. Note that an individual can always defect when his partner is in bad standing without losing his good standing status. In this more general setting the tit-for-tat strategy is replaced by the following *standing strategy*: cooperate if and only if your current partner is in good standing, except that, if you accidentally defected the previous period, cooperate this period unconditionally, thereby restoring your status as a member in good standing. This *standing model* is due to Sugden (1986).

Panchanathan and Boyd (2004) have proposed an ingenious deployment of indirect reciprocity, assuming that there is an ongoing dyadic helping game in society based on the indirect reciprocity information and incentive structure, and there is also an n-player public goods game, played relatively infrequently by the same individuals. In the dyadic helping game, two individuals are paired and each member of the pair may confer a benefit b upon his partner at a cost c to himself, an individual remaining in good standing so long as he does not defect on a partner who is in good standing. This random pairing is repeated with probability δ and with discount factor d. In the public goods game, an individual produces a benefit b_g that is shared equally by all the other members, at a cost c_g to himself. The two games are linked by defectors in the public goods game being considered in bad standing at the start of the helping game that directly follows. Then, cooperation can be sustained in both the public goods game and in the dyadic helping game so long as

$$c_g \leq \frac{b(1-\varepsilon)-c}{1-\delta d}, \quad (5)$$

where ε is the rate at which cooperators unintentionally fail to produce the benefit. Parameters favouring this solution are that the cost c_g of cooperating in the public goods game be low, the factor δd is close to unity, and the net benefit $b(1-\varepsilon)-c$ of cooperating in the reputation-building reciprocity game be large.

The major weakness of the standing model is its demanding informational requirements. Each individual must know the current standing of each member of the group, the identity of each member's current partner, and whether each individual cooperated or defected against his current partner. Since dyadic interactions are generally private, and hence are unlikely to be observed by more than a small number of others, errors in determining the standing of individuals may be frequent. This contrasts sharply with the repeated game models of the previous section, which require only that an individual know how many of his current partners defected in the previous period. Especially serious is that warranted non-cooperation (because in one's own mental accounting one's partner is in bad standing) may be perceived to be unwarranted defection by some third parties but not by others. This will occur with high frequency if information partially private rather than public (not everyone has the same information). It has been proposed that gossip and other forms of communication can transform private into public information, but how this might occur among self-regarding individuals has not been (and probably cannot be) shown, because in any practical setting individuals may benefit by reporting dishonestly on what they have observed, and self-regarding individuals do not care about the harm to others induced by false information. Under such conditions, disagreements among individuals about who ought to be punished can reach extremely high levels, with the unravelling of cooperation as a result.

In response to this weakness of the standing model, Nowak and Sigmund (1998) developed an indirect reciprocity model which they term *image scoring*. Players in the image scoring need not know the standing of recipients of aid, so the informational requirements of indirect reciprocity are considerably reduced. Nowak and Sigmund show that the strategy of cooperating with others who have cooperated in the past, *independent of the reputation of the cooperator's partner*, is stable against invasion by defectors, and weakly stable against invasion by unconditional cooperators once defectors are eliminated from the population. Leimar and Hammerstein (2001), Panchanathan and Boyd (2003), and Brandt and Sigmund (2004; 2005), explore the applicability of image scoring.

Cooperation in large groups of self-regarding individuals

Repeated game theory has extended the above two-player results to a general n-player stage game, the so-called *public goods game*. In this game each player cooperates at cost $c>0$, contributing an amount $b>c$ that is shared equally among the other $n-1$ players. We define the *feasible payoff set* as the set of possible payoffs to the various players, assuming each cooperates with a certain probability, and each player does at least as well as the payoffs obtaining under mutual defection. The set of feasible payoffs for a two-player public goods game is given in Figure 1 by the four-sided figure ABCD. For the n-player game, the figure ABCD is replaced by a similar n-dimensional polytope.

Figure 1

Two-player public goods game

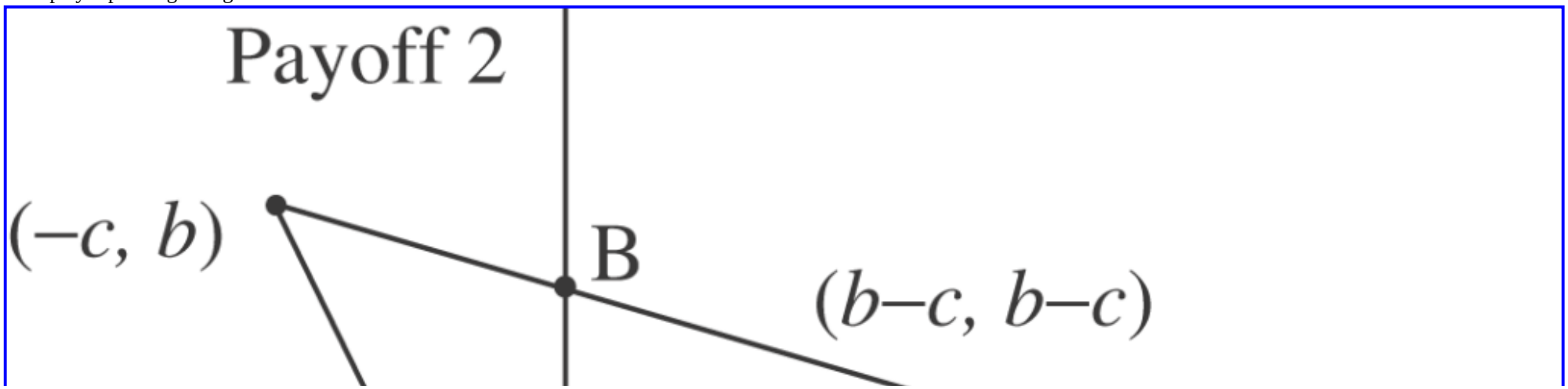

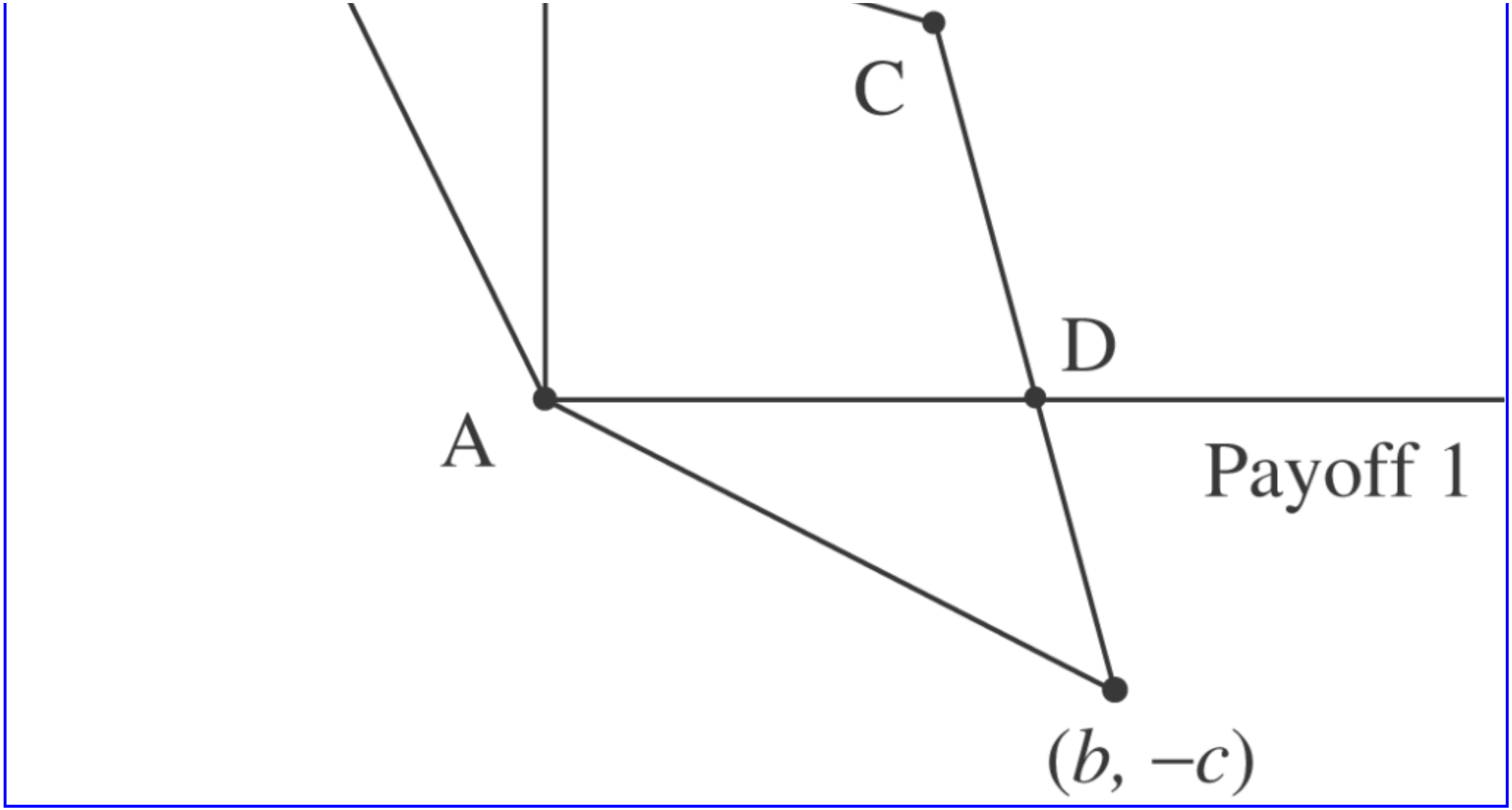

Repeated game models have demonstrated the so-called folk theorem, which asserts that any distribution of payoffs to the n players that lies in the feasible payoff set can be supported by an equilibrium in the repeated public goods game, provided the discount factor times the probability of continuation, δd, is sufficiently close to unity. The equilibrium concept employed is a refinement of subgame perfect equilibrium. Significant contributions to this literature include Fudenberg and Maskin (1986), assuming perfect information, Fudenberg, Levine and Maskin (1994), assuming imperfect information, so that cooperation is sometimes inaccurately reported as defection, and Sekiguchi (1997), Piccione (2002), Ely and Välimäki (2002), Bhaskar and Obara (2002) and Mailath and Morris (2006), who assume that different players receive different, possibly inaccurate, information concerning the behaviour of the other players.

The folk theorem is an *existence theorem* affirming that any outcome that is a Pareto improvement over universal defection may be supported by a Nash equilibrium, including point C (full cooperation) in the figure and outcomes barely superior to A (universal defection). The theorem is silent on which of this vast number of equilibria is more likely to be observed or how they might be attained. When these issues are addressed two problems are immediately apparent: first, equilibria in the public goods game supported in this manner exhibit very little cooperation if large numbers of individuals are involved or errors in execution and perception are large, and second, the equilibria are not robust because they require some mechanism allowing coordination on highly complex strategies. While such a mechanism could be provided by centralized authority, decentralized mechanisms, as we will see, are not sustainable in a plausible dynamic.

The dynamics of cooperation

The first difficulty, the inability to support high levels of cooperation in large groups or with significant behavioural or perceptual noise, stems from the fact that the only way players may punish defectors is *to withdraw their own cooperation*. In the two-person case, defectors are thus targeted for punishment. But for large *n*, withdrawal of cooperation to punish a single defector punishes all group members equally, most of whom, in the neighbourhood of a cooperative equilibrium, will be cooperators. Moreover, in large groups, the rate at which erroneous signals are propagated will generally increase with group size, and the larger the group, the larger the fraction of time group members will spend punishing (miscreants and fellow cooperators alike). For instance, suppose the rate at which cooperators accidentally fail to produce *b*, and hence signal defection, is five per cent. Then, in a group of size two, a perceived defection will occur in about ten per cent of all periods, while in a group of size 20, at least one perceived defection will occur in about 64 per cent of all periods.

As a result of these difficulties, the folk theorem assertion that we can approximate the per-period expected payoff as close to the efficient level (point C in Figure 1) as desired as long as the discount factor δ is sufficiently close to unity is of little practical relevance. The reason is that as $\delta \rightarrow 1$, the current payoff approximates zero, and the expected payoff is deferred to future periods at very little cost, since future returns are discounted at a very low rate. Indeed, with the discount factor δ held constant, the efficiency of cooperation in the Fudenberg, Levine and Maskin model declines at an exponential rate with increasing group size (Bowles and Gintis, 2007, ch. 13). Moreover, in an agent-based simulation of the public goods with punishment model, on the assumption of a benefit/cost ratio of $b/c=2$ (that is, contributing to the public good costs half of the benefit conferred on members of the group) and a discount factor times probability of repetition of $d\delta=0.96$, even for an error rate as low as $\varepsilon=0.04$, fewer than half of the members contribute to the public good in groups of size $n=4$, and less that 20 per cent contribute in groups of size $n=6$ (Bowles and Gintis, 2007, ch. 5).

The second limitation of the folk theorem analysis is that it has not been shown (and probably cannot be shown) that the equilibria supporting cooperation are dynamically robust, that is, asymptotically stable with a large basin of attraction in the relevant dynamic. Equilibria for which this is not the case will seldom be observed because they are unlikely to be attained and if attained unlikely to persist for long.

The Nash equilibrium concept applies when each individual expects all others to play their parts in the equilibrium. But, when there are multiple equilibria, as in the case of the folk theorem, where there are many possible patterns of response to given pattern of defection, each imposing distinct costs and requiring distinct, possibly stochastic, behaviours on the part of players, there is no single set of beliefs and expectations that group members can settle upon to coordinate their actions (Aumann and Brandenburger, 1995).

While game theory does not provide an analysis of how beliefs and expectations are aligned in a manner allowing cooperation to occur, sociologists (Durkheim, 1902; Parsons and Shils, 1951) and anthropologists (Benedict, 1934; Boyd and Richerson, 1985; Brown, 1991) have found that virtually every society has such processes, and that they are key to understanding strategic interaction. Borrowing a page from sociological theory, we posit that groups may have *focal rules* that are common knowledge among group members. Focal rules could suggest which of a countless number of strategies that could constitute a Nash equilibrium should all individuals adopt them, thereby providing the coordination necessary to support cooperation. These focal rules do not ensure equilibrium, because error, mutation, migration, and other dynamical forces ensure that on average not all individuals conform to the focal rules of the groups to which they belong. Moreover, a group's focal rules are themselves subject to dynamical forces, those producing better outcomes for their members displacing less effective focal rules.

In the case of the repeated public goods game, which is the appropriate model for many forms of large-scale cooperation, Gintis (2007) shows that focal rules capable of supporting the kinds of cooperative equilibria identified by the folk theorem are not evolutionarily stable, meaning that groups whose focal rules support highly cooperative equilibria do worse than groups with less stringent focal rules, and as a result the focal rules necessary for cooperation are eventually eliminated.

The mechanism behind this result can be easily explained. Suppose a large population consists of many smaller groups playing *n*-person public goods games, with considerable migration across groups, and with the focal rules of successful groups being copied by less successful groups. To maintain a high level of cooperation in a group, focal rules should foster punishing defectors by withdrawing cooperation. However, such punishment is both costly and provides an external benefit to other groups by reducing the frequency of defection-prone individuals who might migrate elsewhere. Hence, groups that ‘free ride’ by not punishing defectors harshly will support higher payoffs for its members than groups that punish assiduously. Such groups will then be copied by other groups, leading to a secular decline in the frequency of punishment suggested by focal rules in all groups. Thus, suppose that the groups in question were competitive firms whose profits depend on the degree of cooperation among firm members. If all adopted a zero-tolerance rule (all would defect if even a single defection was perceived), then a firm adopting a rule that tolerated a single defection would sustain higher profits and replace the zero-tolerance firms. But this firm would in turn be replaced by a firm adopting a rule that tolerates two defections.

These two problems – the inability to support efficient levels of cooperation in large groups with noisy information, and dynamic instability – have been shown for the case where information is public. Private information, in general the more relevant case, considerably exacerbates these problems.

Cooperation with other-regarding individuals

The models reviewed thus far have assumed that individuals are entirely self-regarding. But cooperation in sizable groups is possible if there exist other-regarding individuals in the form of *strong reciprocators*, who cooperate with one another and punish defectors, even if they sustain net costs. Strong reciprocators are altruistic in the standard sense that they

confer benefits on other members of their group (in this case, because their altruistic punishment of defectors sustains cooperation) but would increase their own payoffs by adopting self-regarding behaviours. A model with *social preferences* of this type can explain large-scale decentralized cooperation with noisy information as long as the information structure is such that defectors expect a level of punishment greater than costs of cooperating.

Cooperation is not a puzzle if a sufficient number of individuals with social preferences are involved. The puzzle that arises is how such altruistic behaviour could have become common, given that bearing costs to support the benefits of others reduces payoffs, and both cultural and genetic updating of behaviours is likely to favour traits with higher payoffs. This evolutionary puzzle applies to strong reciprocity. Since punishment is costly to the individual, and an individual could escape punishment by cooperating, while avoiding the costs of punishment by not punishing, we are obliged to exhibit a mechanism whereby strong reciprocators could proliferate when rare and be sustained in equilibrium, despite their altruistic behaviour.

This is carried out in Sethi and Somanathan (2001), Gintis (2000), Boyd et al. (2003), Gintis (2003) and Bowles and Gintis (2004). The evolutionary viability of other types of altruistic cooperation is demonstrated in Bowles, Jung-Kyoo and Hopfensitz (2003), Boyd et al. (2003), Bergstrom (1995) and Salomonsson and Weibull (2006). The critical condition allowing the evolution of strong reciprocity and other forms of altruistic social preferences is that individuals with social preferences are more likely than random to interact with others with social preferences. Positive assortment arises in these models due to deliberate exclusion of those who have defected in the past (by ostracism, for example), random differences in the composition of groups (due to small group size and limited between-group mobility), limited dispersion of close kin who share common genetic and cultural inheritance, and processes of social learning such as conformism or group level socialization contributing to homogeneity within groups. As in the repeated game models, smaller groups favour cooperation, but in this case for a different reason: positive assortment tends to decline with group size. But the group sizes that sustain the altruistic preferences that support cooperative outcomes in these models are at least an order of magnitude larger than those indicated for the repeated game models studied above.

In sum, we think that other-regarding preferences provide a compelling account of many forms of human cooperation that are not well explained by repeated game models with self-regarding preferences. Moreover, a number of studies have shown that strong reciprocity and other social preferences are a common human behaviour (Fehr and Gächter, 2000; Henrich et al., 2005) and could have emerged and been sustained in a gene-culture co-evolutionary dynamic under conditions experienced by ancestral humans (Bowles, 2006). The above models also show that strong reciprocity and other social preferences that support cooperation can evolve and persist even when there are many self-regarding players, where group sizes are substantial, and when behavioural or perception errors are significant.

Conclusion: economics and the missing choreographer

The shortcomings of the economic theory of cooperation based on repeated games strikingly replicate those of economists' other main contribution to the study of decentralized cooperation, namely, general equilibrium theory. Both prove the existence of equilibria with socially desirable properties, while leaving the question of how such equilibria are achieved as an afterthought, thereby exhibiting a curious lack of attention to dynamics and out-of equilibrium behaviour. Both purport to model decentralized interactions but on close inspection require a level of coordination that is not explained, but rather posited as a *deus ex machina*. To ensure that only equilibrium trades are executed, general equilibrium theory resorts to a fictive ‘auctioneer’. No counterpart to the auctioneer has been made explicit in the repeated-game approach to cooperation. Highly choreographed coordination on complex strategies capable of deterring defection are supposed to materialize quite without the need for a choreographer.

Humans are unique among living organisms in the degree and range of cooperation among large numbers of substantially unrelated individuals. The global division of labour and exchange, the modern democratic welfare state, and contemporary warfare alike evidence our distinctiveness. These forms of cooperation emerged historically and are today sustained as a result of the interplay of self-regarding and social preferences operating under the influence of group-level institutions of governance and socialization that favour cooperators, in part by protecting them from exploitation by defectors.

The norms and institutions that have accomplished this evolved over millennia through trial and error. Consider how real-world institutions addressed two of the shoals on which the economic models foundered. First, the private nature of information, as we have seen, makes it virtually impossible to coordinate the targeted punishment of miscreants. Converting private information about transgressions into public information that can provide the basis of punishment often involves civil or criminal trials, elaborate processes that rely on commonly agreed upon rules of evidence and ethical norms of appropriate behaviour. Even these complex institutions frequently fail to transform the private protestations of innocence and guilt into common knowledge. Second, as in the standing models with private information, cooperation often unravels when the withdrawal of cooperation by the civic-minded intending to punish a defector is interpreted by others as a violation of a cooperative norm, inviting further defections. In all successful modern societies, this problem was eventually addressed by the creation of a corps of specialists entrusted with carrying out the more severe of society's punishments, whose uniforms conveyed the civic purpose of the punishments they meted out, and whose professional norms, it was hoped, would ensure that the power to punish was not used for personal gain. Like court proceedings, this institution works imperfectly. It is hardly surprising then that economists have encountered difficulty in devising simple models of how large numbers of self-regarding individuals might sustain cooperation in a truly decentralized setting.

Modelling this complex process is a major challenge of contemporary science. Economic theory, favouring parsimony over realism, has instead sought to explain cooperation without reference to other-regarding preferences and with minimalist or fictive descriptions of social institutions.

See Also

- agent-based models
- behavioural economics and game theory
- behavioural game theory
- evolutionary economics
- game theory
- group selection
- public goods experiments
- repeated games
- social preferences

Bibliography

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Arrow, K.J. and Hahn, F. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Aumann, R.J. and Brandenburger, A. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 65, 1161–80.

Axelrod, R. and Hamilton, W.D. 1981. The evolution of cooperation. *Science* 211, 1390–6.

Benedict, R. 1934. *Patterns of Culture*. Boston: Houghton Mifflin.

Bergstrom, T.C. 1995. On the evolution of altruistic ethical rules for siblings. *American Economic Review* 85, 58–81.

Bhaskar, V. and Obara, I. 2002. Belief-based equilibria the repeated Prisoner's Dilemma with private monitoring. *Journal of Economic Theory* 102, 40–69.

Binmore, K. 1993. *Game Theory and the Social Contract: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, K. 1998. *Game Theory and the Social Contract: Just Playing*. Cambridge, MA: MIT Press.

Binmore, K.G. 2005. *Natural Justice*. Oxford: Oxford University Press.

Blau, P. 1964. *Exchange and Power in Social Life*. New York: Wiley.

Bowles, S. 2006. Group competition, reproductive leveling, and the evolution of human altruism. *Science* 314, 1669–72.

Bowles, S. and Gintis, H. 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology* 65, 17–28.

Bowles, S. and Gintis, H. 2007. A cooperative species: human reciprocity and its evolution, in preparation.

Bowles, S., Jung-Kyoo, C. and Hopfensitz, A. 2003. The co-evolution of individual behaviors and social institutions. *Journal of Theoretical Biology* 223, 135–47.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P.J. 2003. Evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* 100, 3531–5.

Boyd, R. and Richerson, P.J. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Brandt, H. and Sigmund, K. 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology* 231, 475–86.

Brandt, H. and Sigmund, K. 2005. Indirect reciprocity, image scoring, and moral hazard. *Proceeding of the National Academy of Sciences* 102, 2666–70.

Brown, D.E. 1991. *Human Universals*. New York: McGraw-Hill.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Durkheim, E. 1902. *De La Division du Travail Social*. Paris: Presses Universitaires de France, 1967.

Ely, J.C. and Välimäki, J. 2002. A robust folk theorem for the Prisoner's Dilemma. *Journal of Economic Theory* 102, 84–105.

Fehr, E. and Gächter, S. 2000. Cooperation and punishment. *American Economic Review* 90, 980–94.

Fudenberg, D., Levine, D.K. and Maskin, E. 1994. The Folk Theorem with imperfect public information. *Econometrica* 62, 997–1039.

Fudenberg, D. and Maskin, E. 1986. The Folk Theorem in repeated games with discounting or with incomplete information. *Econometrica* 54, 533–54.

Gintis, H. 1976. The nature of the labor exchange and the theory of capitalist production. *Review of Radical Political Economics* 8(2), 36–54.

Gintis, H. 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206, 169–79.

Gintis, H. 2003. The hitchhiker's guide to altruism: genes, culture, and the internalization of norms. *Journal of Theoretical Biology* 220, 407–18.

Gintis, H. 2007. *Modeling Cooperation with Self-regarding Agents*. Santa Fe: Santa Fe Institute.

Henrich, J., Boyd, R., Bowles, S., Camerer, C. et al. 2005. Economic man in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28, 795–815.

Laffont, J.J. 2000. *Incentives and Political Economy*. Oxford: Oxford University Press.

Leimar, O. and Hammerstein, P. 2001. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London, Series B* 268, 745–53.

Mailath, G.J. and Morris, S. 2006. Coordination failure in repeated games with almost-public monitoring. *Theoretical Economics* 1, 311–40.

Nowak, M.A. and Sigmund, K. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393, 573–7.

Panchanathan, K. and Boyd, R. 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology* 224, 115–26.

Panchanathan, K. and Boyd, R. 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432, 499–502.

Parsons, T. and Shils, E. 1951. *Toward a General Theory of Action*. Cambridge, MA: Harvard University Press.

Piccione, M. 2002. The repeated Prisoner's Dilemma with imperfect private monitoring. *Journal of Economic Theory* 102, 70–83.

Salomonsson, M. and Weibull, J. 2006. Natural selection and social preferences. *Journal of Theoretical Biology* 239, 79–92.

Sekiguchi, T. 1997. Efficiency in repeated Prisoner's Dilemma with private monitoring. *Journal of Economic Theory* 76, 345–61.

Sethi, R. and Somanathan, E. 2001. Preference evolution and reciprocity. *Journal of Economic Theory* 97, 273–97.

Shubik, M. 1959. *Strategy and Market Structure: Competition, Oligopoly, and the Theory of Games*. New York: Wiley.

Stiglitz, J. 1987. The causes and consequences of the dependence of quality on price. *Journal of Economic Literature* 25, 1–48.

Sugden, R. 1986. *The Economics of Rights, Co-operation and Welfare*. Oxford: Basil Blackwell.

Taylor, M. 1976. *Anarchy and Cooperation*. London: Wiley.

Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

How to cite this article

Bowles, Samuel and Herbert Gintis. "cooperation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000597> doi:10.1057/9780230226203.0317

The New Palgrave Dictionary of Economics Online

coordination problems and communication

Jack Ochs

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Coordination problems arise when a game has multiple Nash equilibria and all players have a common interest in avoiding a non-equilibrium state. To achieve an equilibrium state, agents must come to understand one another's intentions. Communication can facilitate this understanding under some, but not all, circumstances. In the absence of communication among agents, coordination may also sometimes be achieved with the aid of extrinsic signals that have come to be associated with the actions of others. In some settings, past actions themselves serve as precedents, without the benefit of any communication.

Keywords

cheap talk; communication; coordination equilibrium; coordination problems; extensive form games; Nash equilibrium; observability; Prisoners' Dilemma; signalling; sunspot equilibrium

Article

Lewis (1969) defined a *coordination equilibrium* as a Nash equilibrium in which no agent would be better off if any other agent had chosen a different action. When there are multiple coordination equilibria, agents face an obvious coordination problem. The resolution of coordination problems rests upon individuals coming to understand the intentions of one another. The most explicit way of developing this understanding is for the individuals to communicate with one another. Common knowledge of a language must precede communication. Even with common knowledge of a language, individuals may not be bound to do what they say they will do. In such circumstances, talk is ‘cheap’. When will the receiver, having received a message from a sender, behave differently from how the receiver would have behaved if no message had been sent? According to Farrell and Rabin (1996) *highly credible* messages will not be ignored. A message that signals an intention to take action X is highly credible if it satisfies two conditions: it is (*a*) *self-signalling* and (*b*) *self-committing*. A message that the

sender is taking action X is self-signalling if, and only if, it is both true and it is in the sender's interest to have it believed to be true. A message is self-committing if a belief by the receiver that the message is true creates an incentive for the sender to do what the sender said he or she would do. A message that is self-committing, if believed, will lead to an outcome that is a Nash equilibrium. A message can be self-committing without being self-signalling. For example, in the classic game of Chicken, if one player announces that he will be Passive, that message is self-committing since, if it is believed by the receiver then the receiver's best response is to be Aggressive, and the best response of the sender to the receiver's aggression is to be Passive. However, the sender would prefer to have the receiver believe that the sender will play Aggression. So the message, ‘I intend to play Passive’, is not self-signalling because it is not in the interest of the sender to have the receiver believe it is true.

A message is *cheap talk* if the sender is not bound to do what the message says. Crawford (1998) provides a survey of a number of cheap talk experiments. In experiments with structured communication, either only one player may send a message (one-sided communication) or more than one player can send a message. When the payoff functions of the players are symmetrical, one-sided communication breaks the symmetry of the game without communication. This is sufficient to allow a very high level of coordination. Indeed, in such games one-sided communication is much more effective in promoting coordination that is simultaneous, two-sided communication. This suggests that, when payoff functions are symmetric but players have different preference orderings over equilibria, as in the Battle of the Sexes, the principal impact of one-sided communication is to create an extensive form game in which the symmetry is broken by designating one player as the first mover. In games with Pareto-ordered equilibria communication is not needed to break symmetry, but may be effective in reducing uncertainty about the intentions or, in Crawford's terms, to give ‘reassurance’. Empirically this ‘reassurance’ appears to be most effective in achieving coordination on the Pareto-dominant equilibrium when communication is two-sided, but even one-sided communication has a positive effect on the likelihood of achieving the Pareto-optimal outcome. Furthermore, this effect has been found to be greater when a message was self-signalling than when such a message was only self-committing.

When there are multiple players each player must be interested in, and possibly condition his actions on, the entire message profile. Therefore, the concepts of self-signalling and self-committing messages may not have much meaning in this context. Nevertheless, there is some evidence that costless pre-play communication can help groups whose members repeatedly interact to achieve more efficient outcomes than is attainable without such communication (Blume and Ortmann, 2007).

A signal that is commonly observed may be used to coordinate actions even if the signal does not emanate from any of the players. Traffic signals play this role. We do as these signals say we should do because we believe that others will also do what the signals say they should do. This belief is reinforced by experience, so doing as the signals suggest has simply become a convention that is adopted by drivers. While this convention is backed by law, there is good reason to believe that it is so ingrained in people's expectations that they would continue to act as the signals suggest even in the absence of any law. Can signals be effective in coordinating actions when the signals are not sent by any of the players and do not themselves have any payoff consequences? Van Huyck, Gillette and Battalio (1992) found that, when a game has multiple coordination equilibria, all of which yield the same payoff, a signal from an outside ‘moderator’ that specifically says ‘play a particular equilibrium’ produces a very high degree of coordination on the suggested equilibrium, even though absent any signal there is a high frequency of coordination failure. However, in games where the equilibria are Pareto ordered the introduction of a

recommendation to play any equilibrium other than the payoff-dominant equilibrium significantly reduces the degree of coordination that is achieved. The authors also found that when there was an equilibrium that provided equal payoff a recommendation to play an equilibrium with unequal payoffs had little influence on how the game was played. Evidently some features, such as symmetry, may be sufficiently strong focal points that the introduction of extrinsic signals may have little influence. Similarly, some features of a game may make some coordination equilibria, once achieved by repeated interaction, exceedingly difficult to displace through the introduction of communication, even if everyone would gain by moving to another coordination equilibrium (Cooper, 2006).

A ‘sunspot’ is a commonly observable event that may have been correlated in the past with different outcomes. For example, published forecasts may have this property. When agents coordinate their actions on a ‘sunspot’ the resulting equilibrium is called a ‘sunspot equilibrium’. Marimon, Spear and Sunder (1993) devised an experiment to see whether they could generate a sunspot equilibrium where prices fluctuate with an extrinsic signal even though the fundamental parameter values remained fixed. During a ‘training interval’, the colour of a blinking light on a screen was perfectly correlated with a change in a parameter that induced changes in equilibrium prices. After this ‘training period’ the parameter value was fixed, but the signal continued to vary according to the same process. Prices continued to be volatile but there was little evidence that the variation in the sunspot variable had any effect on the observed price volatility. Duffy and Fisher (2005), using a quite different design, were able to induce sunspot equilibria under restricted conditions. They found that the semantics of the sunspot variable mattered. There were two fundamental equilibria in their design. One equilibrium had a high price, the other a low price. When the sunspot message was either ‘high’ or ‘low’ the outcomes of the actions were sometimes correlated with the message. But when the message was either ‘sunshine’ or ‘rain’ this correlation was never observed. Evidently, correlation of expectations with the signal depends upon how confident people are that everyone is interpreting the signal in the same way. They also found that information that is generated by observable actions subsequent to the observation of the signal itself tends to diminish the focal power of the signal.

Sometimes actions might ‘speak’ louder than words. In a Prisoners' Dilemma game the cooperative outcome is not a Nash equilibrium, but it does Pareto-dominate the Nash equilibrium. Since non-cooperation is a dominant strategy a message that one intends to play ‘Cooperate’ is neither self-committing nor self-signalling. Nevertheless, Duffy and Feltovich (2002) found that when this message was sent it tended to be truthful and also tended to induce a cooperative response. Similarly, when their past actions with other players were observable, subjects were more likely to cooperate than if neither communication nor observability was possible. Furthermore, observation increased the frequency of cooperative choices by more than cheap talk. This suggests that observability of past actions may sometimes be more effective than mere words in helping people achieve a good outcome.

See Also

- cheap talk
- experimental economics
- game theory

Bibliography

Blume, A. and Ortmann, A. 2007. The effects of costless pre-play communication: experimental evidence from games with Pareto-ranked equilibria. *Journal of Economic Theory* 132, 274–90.

Cooper, D. 2006. Are experienced managers experts at overcoming coordination failure? *Advances in Economic Analysis & Policy* 6(2), Article 6.

Crawford, V. 1998. A survey of experiments on communication via cheap talk. *Journal of Economic Theory* 78, 286–98.

Duffy, J. and Feltovich, N. 2002. Do actions speak louder than words? An experimental comparison of observation and cheap talk. *Games and Economic Behavior* 39, 1–27.

Duffy, J. and Fisher, E. 2005. Sunspots in the laboratory. *American Economic Review* 95, 510–29.

Farrell, J. and Rabin, M. 1996. Cheap talk. *Journal of Economic Perspectives* 10(3), 103–18.

Lewis, D. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Marimon, R., Spear, S. and Sunder, S. 1993. Expectationally driven market volatility: an experimental study. *Journal of Economic Theory* 61, 74–103.

Van Huyck, J., Gillette, A. and Battalio, R. 1992. Credible assignments in coordination games. *Games and Economic Behavior* 4, 606–26.

How to cite this article

Ochs, Jack. "coordination problems and communication." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000593> doi:10.1057/9780230226203.0318

The New Palgrave Dictionary of Economics Online

copulas

Pravin K. Trivedi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Copulas are functional forms that parameterize the joint distribution of random variables based on their stated marginal distributions and a dependence parameter. The approach is based on Sklar's theorem. Copulas provide a general method for modelling dependence between random variables that may exhibit asymmetric dependence, which is often inadequately captured by measures of linear dependence. Copulas are often generated by using mixtures and convex sums. Although a bivariate distribution is the most commonly encountered specification, higher dimensional joint distributions can also be generated.

Keywords

Clayton copula; copulas; cumulative distribution functions; GARCH effects; Gaussian copula; Gumbel copula; marginal distributions; selection models; Sklar, A.; Sklar's theorem; tail dependence

Article

Sklar introduced copulas in 1959 (Sklar, 1973; 1996). Concisely stated, copulas are functions that connect multivariate distributions to their one-dimensional margins. If *F* is an *m*-dimensional continuous cumulative distribution function (CDF) with one-dimensional margins $F_1, \ldots, F_m$, then there exists an *m*-dimensional unique copula *C* such that $F(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m))$. In general, marginal distributions alone cannot determine the joint distributions.
Copulas are useful because, first, they represent a method for deriving joint distributions given the fixed marginals, even when marginals belong to different parametric families of distributions; second, in a bivariate context copulas can be used to define nonparametric measures of dependence for pairs of random variables that can capture asymmetric (tail) dependence as well as correlation or linear association.

Copulas and dependence

We begin with Sklar's theorem. An *m*-copula is an *m*-dimensional CDF whose support is contained in $[0,1]^m$ and whose one-dimensional margins are uniform on [0,1]. In other words, an *m*-copula is an *m*-dimensional distribution function with all *m* univariate margins being *U*(0,1). To see the relationship between distribution functions and copulas, consider a continuous *m*-variate distribution function $F(y_1, \ldots y_m)$ with univariate marginal distributions $F_1(y_1), \ldots, F_m(y_m)$ and inverse probability transforms (quantile functions) $F_1^{-1}, \ldots, F_m^{-1}$. Then $y_1 = F_1^{-1}(u_1) \sim F_1, \ldots, y_m = F_m^{-1}(u_m) \sim F_m$ where $u_1, \ldots, u_m$ are uniformly distributed variates. Copulas are expressed in terms of marginal CDFs. The transforms of uniform variates are distributed as $F_i (i = 1, \ldots, m)$. Hence

$$F(y_1, \ldots, y_m) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)) = C(u_1, \ldots, u_m),$$

(1)

is the unique copula associated with the distribution function. The copula parameterizes a multivariate distribution in terms of its marginals. For an *m*-variate distribution *F*, the copula satisfies

$$F(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m); \theta),$$
(2)

where θ is usually a scalar-valued dependence parameter. For many empirical applications, the dependence parameter is the main focus of estimation. Because the marginal distributions may come from different families, copulas are a 'recipe' for generating joint distributions by combining given marginal distributions using a known copula. This construction allows researchers to consider marginal distributions and dependence as two separate, but related, issues.

The functional form of a copula places restrictions on the dependence structure; for example, it may support only positive dependence. Therefore, a pivotal modelling problem is to choose a copula that adequately captures dependence structures of the data without sacrificing attractive properties of the marginals. Copulas are multivariate distribution functions, hence Fr•chet bounds apply. A copula may impose restrictions such that the full coverage between the bounds is not attained.

An important advantage of copulas is that they generate more general measures of dependence than the correlation coefficient. Correlation is a symmetric measure of linear dependence, bounded between +1 and –1 and invariant with respect to only linear transformations of the variables. By contrast, copulas have an attractive invariance property: the dependence captured by a copula is invariant with respect to increasing and continuous transformations of the marginal distributions. The same copula may be used for, say, the joint distribution of (Y_1, Y_2) as $(ln \bullet Y_1, ln \bullet Y_2)$.

Measures of dependence based on concordance, such as Spearmans's rank correlation (ρ) and Kendall's τ , overcome limitations of the correlation coefficient. In some cases the concordance between extreme (tail) values of random variables is of interest. For example, one may be interested in the probability of the event that stock indexes in two countries exceed (or fall below) given levels. This requires a dependence measure for upper and lower tails of the distribution, rather than a linear correlation measure. Measures of lower and upper tail dependence can be readily derived for a stated copula. The copula dependence parameter θ can be converted to measures of concordance such as Spearman's ρ and Kendall's τ (Nelsen, 1999).

Examples

Nelsen (1999) and Joe (1997) catalogue many functional forms for copulas. Particularly important is the Archimedean class. Bivariate Archimedean copulas take the general symmetric form

$$C(u_1, u_2; \theta) = \phi^{-1}(\phi(u_1) + \phi(u_2)),$$
(3)

where the generator function $\varphi(\cdot)$ is a convex decreasing function; for example, $\phi(t) = -ln(t)$. The dependence parameter θ in imbedded in the functional form of the generator. The *Clayton copula*, a member of the Archimedean class, takes the form

$$C(u_1, u_2; \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$$
(4)

with the dependence parameter θ restricted on the region $(0,\infty)$. As θ approaches zero, the marginals become independent. The Clayton copula cannot account for negative dependence. It has been used to study correlated risks because it exhibits strong left tail dependence and relatively weak right tail dependence.

The *Gumbel copula* is another member of the Archimedean class and takes the form

$$C(u_1, u_2; \theta) = \exp(-(\tilde{u}_1^{\theta} + \tilde{u}_2^{\theta})^{1/\theta}) \quad (5)$$

where $\tilde{u}_j = -\log u_j$. The dependence parameter is restricted to the interval $[1,\infty)$. Like the Clayton copula, Gumbel does not allow negative dependence, but in contrast it exhibits strong right tail dependence and relatively weak left tail dependence. If outcomes are strongly correlated at high values but less correlated at low values, then the Gumbel copula is an appropriate choice.

The (non-Archimedean) *Gaussian copula* takes the form

$$C(u_1, u_2; \theta) = \Phi_G(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta), \quad (6)$$

where Φ is the CDF of the standard normal distribution, and $\Phi_G(u_1, u_2)$ is the standard bivariate normal distribution with correlation parameter θ restricted to the interval $(-1, 1)$. This copula allows equal degrees of positive and negative dependence.

Figures 1, 2, and 3 illustrate lower and upper tail dependence using three samples generated using Monte Carlo draws from the above three copulas. The samples have comparable degrees of linear dependence but different tail dependence properties.

Figure 1

Sample from Clayton copula, theta=4.67

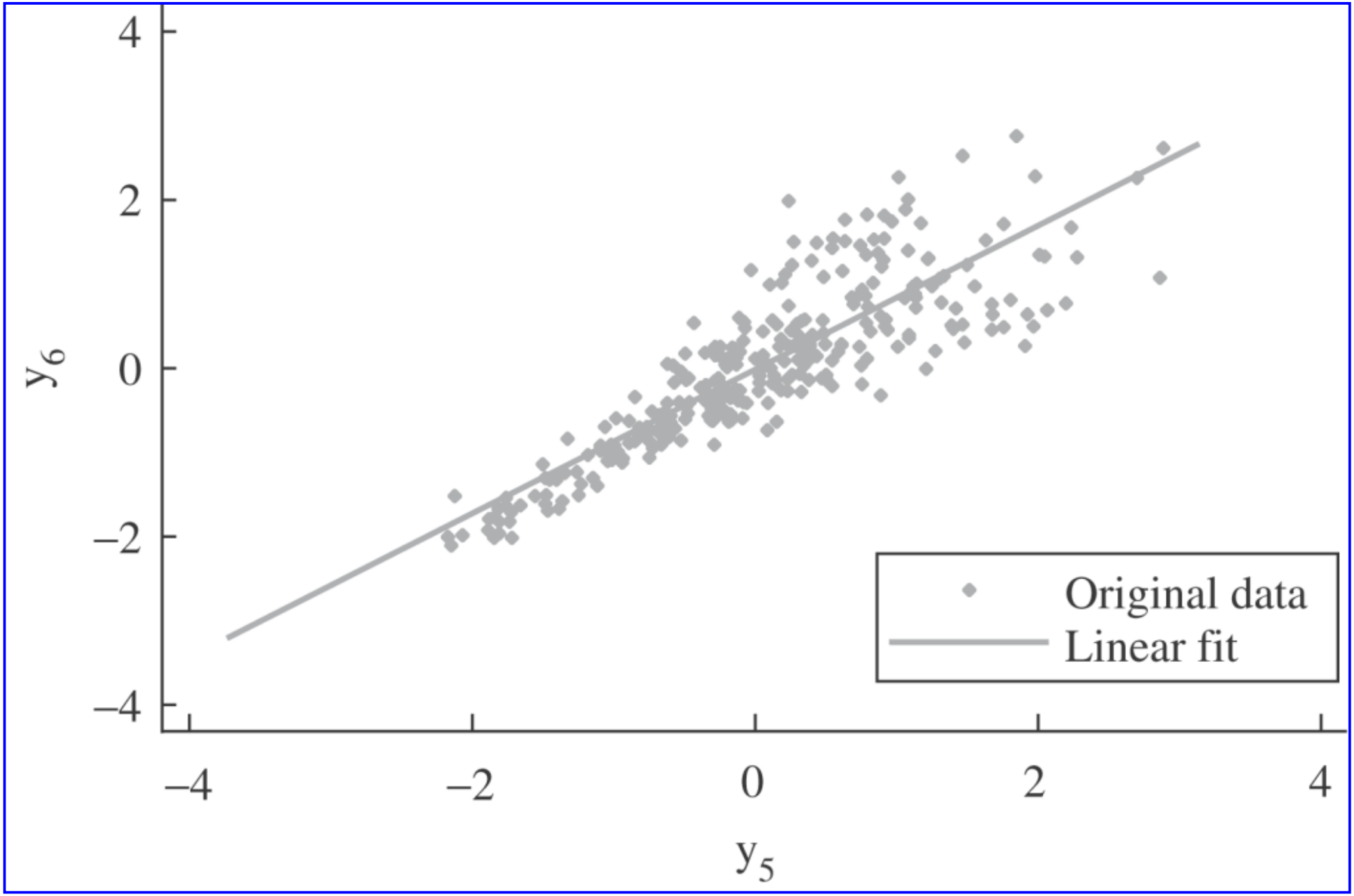

Figure 2
Sample from Gumbel copula, theta=3.3

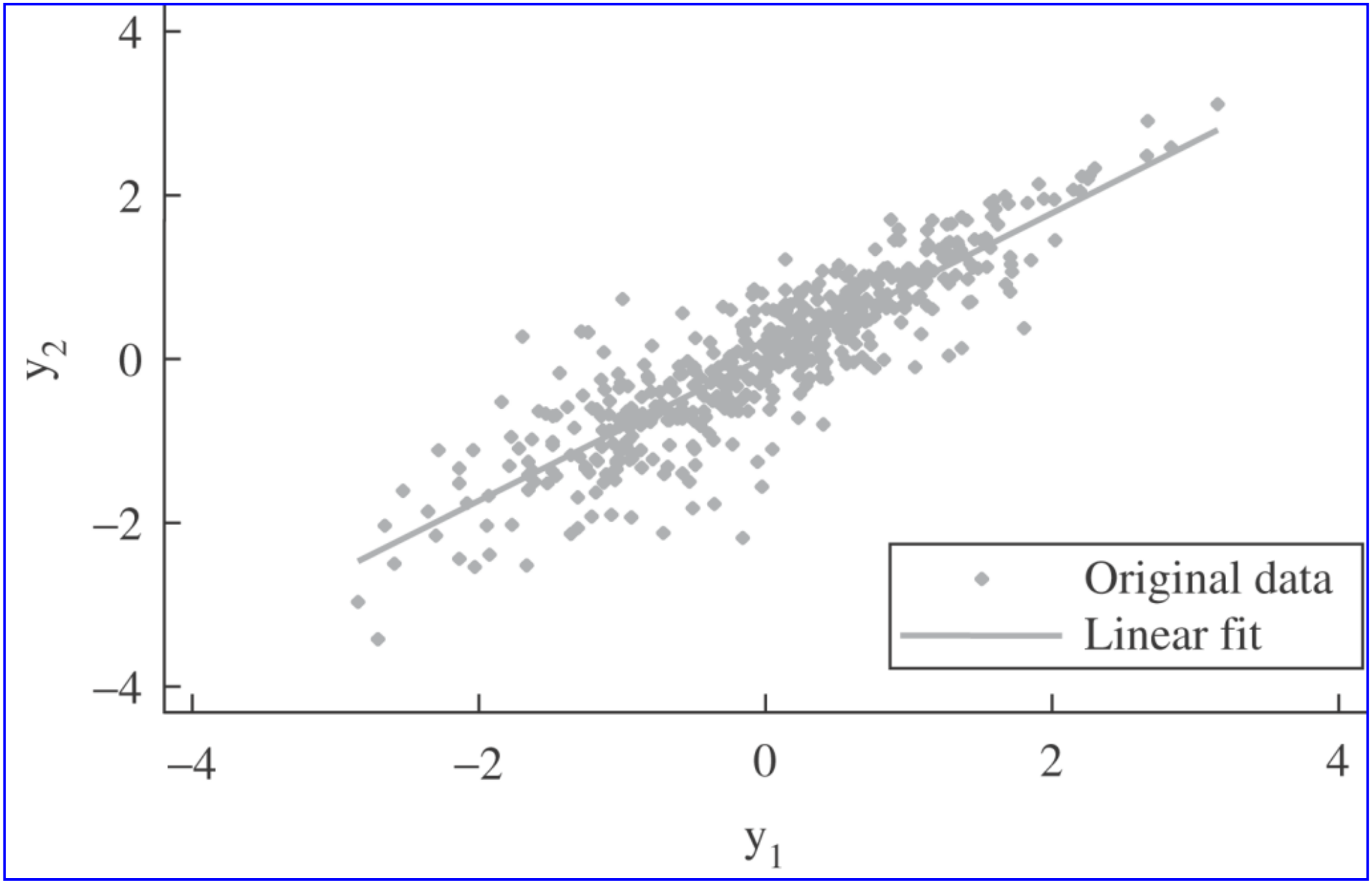

Figure 3
Sample from Gaussian copula, theta=.89

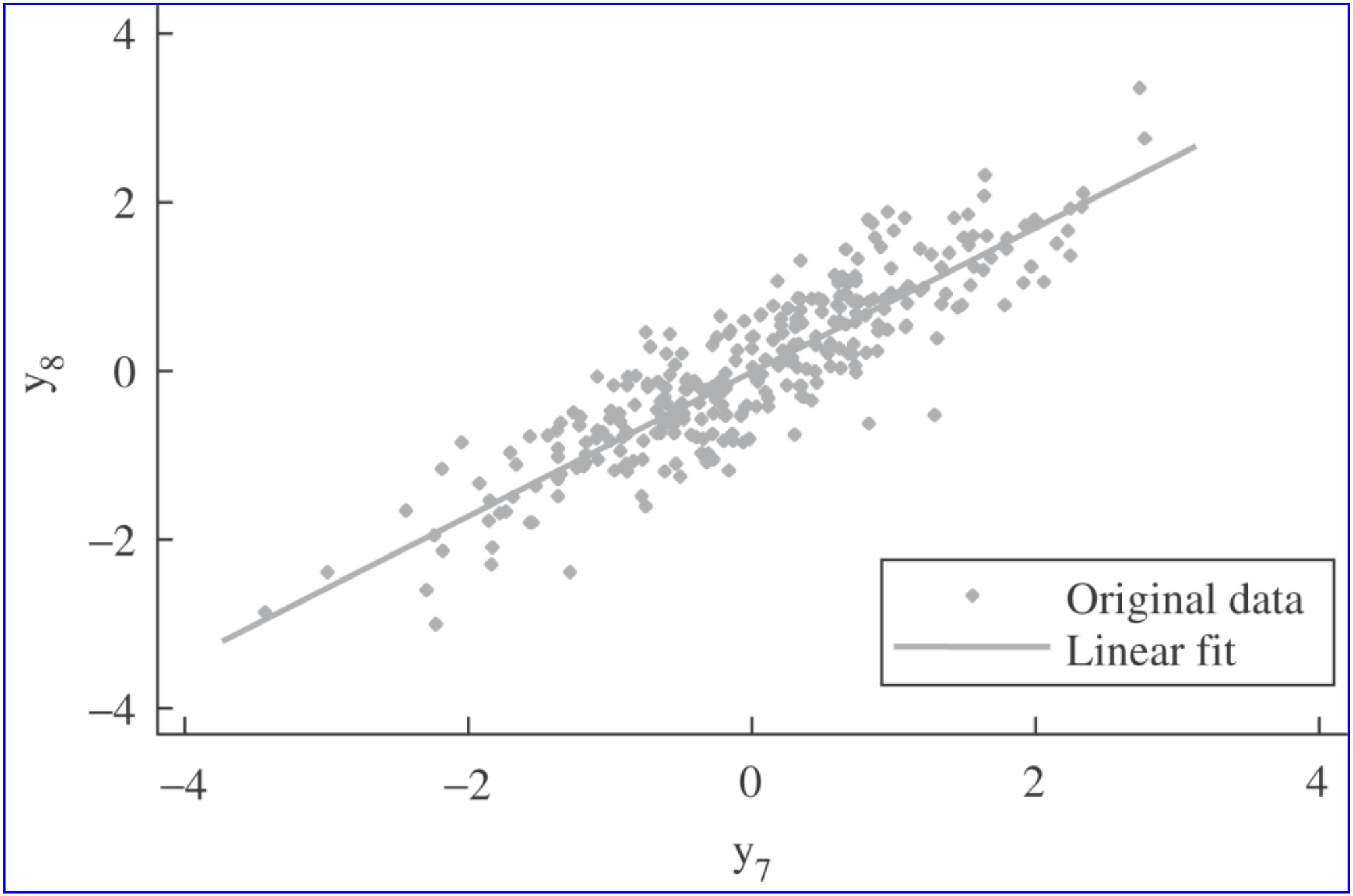

Estimation and applications

In some applications it would natural to parameterize the marginals in terms of a regression function with covariates z, that is, $u_j = F(y_j|z_j; \beta_j)$, where z_j is a vector of covariates. Then the bivariate copula takes the form $C(y_1, y_2|z_1, z_2; \beta_1, \beta_2, \theta) = C(F(y_1|z_1; \beta_1), F(y_2|z_2; \beta_2); \theta)$. The copula density, defined as

$$\frac{d}{dy_2 dy_1} C(F_1(\cdot), F_2(\cdot); \theta) = C_{12}(F_1(\cdot), F_2(\cdot)) f_1(\cdot) f_2(\cdot), \tag{7}$$

$f_j(\cdot) = \partial F_j(\cdot) / \partial y_j$, can be used to build the likelihood, which can be maximized simultaneously with respect to all unknown parameters. Alternatively, the marginal densities can be estimated first, either parametrically or nonparametrically, and then the likelihood can be maximized with respect to θ only at the second stage.
Multivariate models of survival data pioneered the application of copulas. Econometric applications are more recent, but growing rapidly. There are numerous time series and financial market applications of copulas (Cherubini, Luciano and Vecchiato, 2004). Few models in this literature include regressors. Other areas of applications include volatility and exchange rate modelling where GARCH effects and tail dependence are expected (Patton, 2006). Selection models provide leading examples of microeconometric applications of copulas (Smith, 2003; Zimmer and Trivedi, 2006).

See Also

- seemingly unrelated regressions
- simultaneous equations models

Bibliography

Cherubini, U., Luciano, E. and Vecchiato, W. 2004. *Copula Methods in Finance*. New York: John Wiley.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman and Hall.

Nelsen, R. 1999. *An Introduction to Copulas*. New York: Springer.

Patton, A. 2006. Estimation of multivariate models for time series of possibly different lengths. *Journal of Applied Econometrics* 21, 147–73.

Sklar, A. 1973. Random variables, joint distributions, and copulas. *Kybernetica* 9, 449–60.

Sklar, A. 1996. Random variables, distribution functions, and copulas – a personal look backward and forward. In *Distributions with Fixed Marginals and Related Topics*, ed. L. Ruschendorf, B. Schweizer and M. Taylor. Hayward, CA: Institute of Mathematic Statistics.

Smith, M. 2003. Modeling selectivity using Archimedean copulas. *Econometrics Journal* 6, 99–123.

Zimmer, D. and Trivedi, P. 2006. Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business and Economic Statistics* 24, 63–76.

How to cite this article

Trivedi, Pravin K. "copulas." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000580>
doi:10.1057/9780230226203.0319

The New Palgrave Dictionary of Economics Online

core convergence

Robert M. Anderson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The core of an economy is the set of all economic outcomes that cannot be ‘blocked’ by any group of individuals; it is an institution-free concept. A Walrasian equilibrium is an economic outcome based on the institution of market-clearing via prices: each individual consumes his or her demand, taking prices as given, and the demand for each good equals the supply of that good. Core convergence asserts that, for sufficiently large economies, every core allocation approximately satisfies the definition of Walrasian equilibrium; it is an important test of the price-taking assumption inherent in the definition of Walrasian equilibrium.

Keywords

convexity; cooperative game theory (core); core convergence; core; First Welfare Theorem; Edgeworth, F. Y.; Second Welfare Theorem; separating hyperplane th; Shapley–Folkman th; Walrasian equilibrium

Article

The core of an economy, first defined by Edgeworth (1881), is the set of all economic outcomes such that no group of individuals (‘coalition’) can make each of its members better off (‘improve on’ or ‘block’ the outcome), using only the resources available to the group. (A common mistake is to ask, in reference to a particular core allocation, ‘what coalition(s) have formed?’ An allocation is in the core precisely when no coalition can improve on it, and a core allocation does not identify an associated coalition or coalitions. It is when an allocation is *not* in the core that one can identify one or more coalitions that are associated with it, because they can improve on it and thus demonstrate that the coalition is not in the core.)

The most important reason for studying the core is the light it sheds on Walrasian equilibrium, introduced by Walras (1874). While the notion of Walrasian equilibrium is based entirely on the institution of trading via prices, and assumes that individuals take prices as given, the definition of the core is completely institution-free; this is one of its major virtues.

The core has both normative and positive significance apart from its relationship to Walrasian equilibrium. Normatively, if one accepts the distribution of the economy's initial resources as equitable, then any allocation outside the core is unfair to at least one coalition. Regardless of whether the distribution of initial resources is equitable, it would be surprising to find the economy settling on an allocation outside the core, since that would indicate there is a coalition which *could* have made each of its members better off, using only its own resources, but for some reason has failed to coalesce and do so; this is the positive significance.

While there has been much work on the cores of production economies, the bulk of the work on the core has been carried out in exchange economies, in which trading and consuming are the only economic activities. In part, this is because there are a number of competing definitions of the core in production economies, based on how the ownership of the production technology is assigned to individuals and groups. For simplicity, we shall focus our attention on exchange economies.

Walrasian equilibrium is an economic equilibrium notion based on market clearing, mediated by prices. Consumers choose the consumption vector which maximizes utility over their budget sets; firms choose production plans which maximize profit. Critically, it is assumed that individuals and firms take prices as given, without taking into account any ability they may have to influence those prices through their actions. A price vector is a Walrasian equilibrium price if the choices made by individuals and firms, taking prices as given, are consistent in the sense that market supply equals market demand. A Walrasian allocation is the vector of individual consumptions and firm productions generated by a Walrasian equilibrium price. A Walrasian equilibrium is a pair consisting of a Walrasian equilibrium price and its associated Walrasian allocation.

An income transfer is a vector which assigns to each agent a real number, and which satisfies budget balance: the sum of the numbers is zero. An allocation is a Walrasian equilibrium with transfers if there is an income transfer and a price vector such that the demand of each agent, given the prices and the budget of the agent, taking into account the agent's endowment of goods and income transfers, just equals the

individual's consumption at the allocation.

The First and Second Welfare Theorems are two of the most important results concerning Walrasian equilibrium. Recall that, in an exchange economy, an allocation is Pareto optimal if there is no reallocation of consumption which makes every agent better off. In other words, the coalition consisting of all agents (coalition of the whole) cannot improve upon the allocation. Thus, it is clear that every core allocation is Pareto optimal.

The First Welfare Theorem asserts that every Walrasian allocation with transfers is Pareto optimal. A slight modification of the proof suffices to show that every Walrasian allocation lies in the core. (Note that it is *not* true that every Walrasian allocation with transfers lies in the core. The income transfers allow us to move consumption among agents. For example, consider the allocation which gives the entire social consumption to a single agent. If we choose a price vector which supports that agent's preference at the social consumption, then there is an income transfer that makes this allocation a Walrasian allocation with transfers. But this allocation will rarely lie in the core, since the coalition consisting of all the other agents will generally be able to improve on it.) This is an important strengthening of the First Welfare Theorem, which has both positive and normative significance. On the positive side, it is a strong stability property of Walrasian equilibrium, since it asserts that no group of individuals would choose to upset the equilibrium by recontracting among themselves, making it more plausible that we will see Walrasian equilibrium arise in real economies. On the normative side, if we accept the distribution of initial endowments as equitable, it tells us that Walrasian allocations are fair to all groups in the economy.

The Second Welfare Theorem asserts that, in an exchange economy with standard assumptions on preferences (convexity is the crucial assumption), every Pareto optimal allocation is a Walrasian equilibrium with transfers. Note that while the definition of Pareto optimality makes no mention of prices, the Second Welfare Theorem asserts that every Pareto optimal allocation is closely associated to a price vector. The price vector appears magically; mathematically, this is a consequence of the separating hyperplane theorem, for which convexity is a critical assumption. As noted above, the most important use of the core is as a test of the price-taking assumption inherent in the definition of Walrasian equilibrium; a number of other tests have been proposed, but *core convergence* is the most commonly used. Core convergence is closely analogous to the conclusion of the Second Welfare Theorem. The definition of the core makes no mention of prices. However, if an exchange economy is sufficiently large, it is a remarkable fact that every core allocation is closely associated with a price vector that ‘approximately decentralizes’ it; in other words, every core allocation approximately satisfies the definition of Walrasian equilibrium, *without transfers*. This is an important strengthening of the Second Welfare Theorem. The notion of approximate decentralization depends to a considerable extent on the assumptions one is willing to make on the preferences and endowments of the individuals in the economy. (One version states that core allocations can be realized as *exact* Walrasian equilibrium with *small* income transfers.)

Core convergence has a number of implications, both normative and positive. The extent to which each of these implications is justified in a particular setting depends a great deal on the form of convergence, and thus on the assumptions one is willing to make on the economy. For an extensive survey focusing on the relationship between assumptions and the form of convergence, see Anderson (1992).

On the normative side, core convergence is a strong ‘unbiasedness’ property of Walrasian equilibrium, since it asserts that restricting attention to Walrasian allocations does not narrow the set of outcomes much beyond the narrowing that occurs in the core. Thus, Walrasian equilibrium has no hidden implications for the welfare of different groups, beyond whatever equity concerns one might have over the initial endowments. If one accepts the distribution of initial endowments as equitable, then any allocation that is far from Walrasian will not be in the core, and hence will treat some group of agents unfairly. On the positive side, if one accepts the core as a positive description of the allocations one is likely to see in practice in any economy, then core convergence tells us that the allocations we see will be nearly Walrasian.

However, the greatest significance of core convergence is as a test of the reasonableness of the price-taking assumption that is hidden in plain sight in the definition of Walrasian equilibrium. In real markets, we see prices used to equate supply and demand, but this does not guarantee Walrasian outcomes. Agents possessing market power may choose to demand quantities different from their price-taking demands at the prevailing price, thereby altering that price and leading to a non-Walrasian outcome. If the outcome is not at least approximately Walrasian, then the welfare theorems and the results on existence and generic determinacy of Walrasian allocations would have limited implications for real economies.

Core convergence and non-convergence allows us to identify situations in which price-taking is more or less reasonable. Core convergence implies that all trade takes place at almost a single price. An agent who tries to bargain cannot influence the prices much, so there is little incentive to be anything other than a price-taker. On the other hand, core non-convergence makes price-taking an implausible assumption.

Edgeworth (1881) doubted the positive significance of Walrasian equilibrium, and argued that the core, not the set of Walrasian equilibria, was the best positive description of the outcomes of a market mechanism. Moreover, while Edgeworth's name is closely associated with core convergence, and he did prove a core convergence theorem, he argued that in real economies, the presence of firms and syndicates which possess market power ensures that the core does *not* converge.

Edgeworth's argument about the effects of market power applies most strongly to the production side of the economy, where we do in fact see large firms, syndicates and labour unions. However, on the consumption side, the wealthiest individual in the world consumes a small part of the world's annual consumption. In exchange economies in which each consumer is small, core convergence holds. So core convergence provides a justification for the price-taking assumption on the consumption side, provided one views the world as an exchange economy in which the production decisions have been previously made by some exogenous process, outside the scope of the model, endowments include the income obtained by selling one's labour in the exogenous production process, and the only economic activity is trade and consumption of what has been produced.

The proof of the most basic core convergence theorem, which assumes very little about preferences and endowments, and establishes

approximate decentralization in a relatively weak sense, is closely analogous to the proof of the Second Welfare Theorem. The approximately decentralizing price vector appears magically, as a consequence of the separating hyperplane theorem and the Shapley–Folkman theorem, which asserts that the sum of a large number of sets is approximately convex. Convexity of preferences plays no role. Indeed, the definition of the core, because it allows for individuals to be included or excluded from potential coalitions, introduces a non-convexity which is not present in the Second Welfare Theorem, and the Shapley–Folkman theorem controls that non-convexity, whether or not preferences themselves are convex.

The definitions and results just described verbally are presented more formally below.

Many people have made important contributions to the study of core convergence. A survey of these contributions is given in Anderson (1992), and a list of some of the more important contributions is included in the bibliography.

Now, we turn to a more formal presentation

Definition 1: In an exchange economy with agents i=1,…, I having strict preferences $\succ_i$ and endowments $\omega_i \in \mathbf{R}^L_+$, a *coalition* is a set $S \subseteq$ {1,…, I }. An *exact allocation* is $x \in \left(\mathbf{R}^L_+\right)^I$ such that $\Sigma_{i=1}^I x_i = \Sigma_{i=1}^I \omega_i$. An exact allocation is *weakly Pareto optimal* if there is no other exact allocation x' satisfying $x_i' \succ_i x_i (i = 1, \ldots, I)$. A coalition S *blocks* or *improves on* an exact allocation x by x' if $\Sigma_{i \in S} x_i' = \Sigma_{i \in S} \omega_i$ and $\forall_{i \in S}\, x_i' \succ_i x_i$. The *core* is the set of all exact allocations which cannot be improved on by any nonempty coalition. The *price simplex* is $\Delta = \left\{p \in \mathbf{R}^k_+ : \Sigma_{\ell=1}^L \; p_\ell = 1\right\}$.

Theorem 2: In an exchange economy, every core allocation is weakly Pareto optimal.

Proof: If x is not weakly Pareto optimal, then there exists x' such that $\Sigma_{i=1}^I x_i' = \Sigma_{i=1}^I x_i$, $x_i' \succ_i x_i$. Then S={1,…, I } improves on x by x', so x is not in the core.

Theorem 3: (**Strong First Welfare Theorem**) *In an exchange economy, every Walrasian Equilibrium lies in the core.*

Proof: Suppose (p^*, x^*) is a Walrasian Equilibrium. If x^* is not in the core, there exists $S \subseteq I$, $S \neq \varnothing$ and x_i'($i \in S$) such that $\Sigma_{i \in S} x_i' = \Sigma_{i \in S} \omega_i$ and $x_i' \succ_i x_i^*$ for each $i \in S$. Since x_i^* lies in i's demand set at the price p^*, $p^* \cdot x_i > p^* \omega_i$, so $p^* \cdot \Sigma_{i \in S} x_i' = \Sigma_{i \in S} p^* \cdot x_i' > \Sigma_{i \in S} p^* \cdot \omega_i = p^* \cdot \Sigma_{i \in S} \omega_i$ but $\Sigma_{i \in S} x_i' = \Sigma_{i \in S} \omega_i$, a contradiction. Therefore, x^* is in the core.♦

Theorem 4: **(Core convergence, E. Dierker, 1975, and Anderson, 1978)** Suppose we are given an exchange economy with L commodities, I agents and preferences $\succ_1, \ldots, \succ_I$ satisfying weak monotonicity (if $x \gg y$, then $x \succ_i y$) and the following free disposal condition: $x \gg y,\; y \succ_i z \Rightarrow x \succ_i z$. If x is in the core, then there exists $p \in \Delta$ such that

$$\frac{1}{I}\sum_{i=1}^{I} |p \cdot (x_i - \omega_i)| \leq \frac{2L}{I} \max\left\{\|\omega_1\|_\infty, \ldots \|\omega_I\|_\infty\right\} \tag{1}$$

$$\frac{1}{I}\sum_{i=1}^{I} |\inf\left\{p \cdot (y - x_i) : y \succ_i x_i\right\}| \leq \frac{4L}{I} \max\left\{\|\omega_1\|_\infty, \ldots \|\omega_I\|_\infty\right\} \tag{2}$$

where $\|x\|_\infty = \max\{|x_1|, \ldots, |x_L|\}$.

If there are many more agents than goods, and the endowments are not too large, the bounds on the right-hand sides of eqs (1) and (2) will be small. In that case, eq. (1) says that trade occurs almost at the price p, and that each x_i is almost in the budget set, while eq. (2) says that the price p almost supports $\succ_i$ at x_i, in the sense that everything preferred to x_i costs almost as much as x_i. Taken together, eqs (1) and (2) say that the pair (p, x) satisfies a slightly perturbed version of the *def* of Walrasian equilibrium. Indeed, if we knew the left sides of eqs (1) and (2) were *zero*, then $p \cdot (x_i - w_i)$=0, so x_i lies in i's budget set, and $y \succ_i x_i \Rightarrow p \cdot y \geq p \cdot \omega_i$, so x would be a Walrasian quasi-equilibrium!

(A pair (p^*, x^*) is said to be a Walrasian quasi-equilibrium if it satisfies the definition of a Walrasian equilibrium except that instead of requiring that x_i^* lie in *i*'s demand set, we only require that x_i^* lie in *i*'s quasi-demand set, that is $p^* \cdot x_i^* \leq p^* \cdot \omega_i$ and every $y \succ_i x_i^*$ satisfies $p^* \cdot y \geq p^* \cdot \omega_i$.)

Outline of Proof: : Follow the proof of the Second Welfare Theorem.

- Suppose *x* is in the core. Define $B_i = \{y - \omega_i : y \succ_i x_i\} \cup \{0\} = (\{y : y \succ_i x_i\} \cup \{\omega_i\}) - \omega_i$ and $B = \sum_{i=1}^{I} B_i$. The first term in the definition of B_i corresponds to members of a potential improving coalition; for accounting purposes, we assign members outside the coalition their endowments. Note that B_i is *not* convex, even if $\succ_i$ is a convex preference.

  Claim: If *x* is in the core, then $B \cap \mathbf{R}_{--}^L = \varnothing$. Suppose $z \in B \cap \mathbf{R}_{--}^L$. Then there exists $z_i \in B_i$ such that $z = \sum_{i=1}^{I} z_i$. Let $S=\{i:z_i \neq 0\}$; since $z \ll 0$, $S \neq \varnothing$. For $i \in S$, let $x_i' = \omega_i + z_i - \frac{z}{|S|}$. Then $x_i' \gg \omega_i + z_i \succ_i x_i$ by the definition of B_i, $x_i' \succ_i x_i$ by free disposal, and $\sum_{i \in S} x_i' = \sum_{i \in S} \omega_i$, so *S* can improve on *x* by x', so *x* is not in the core.
- Let $v = -L(\max_{i=1,\&,I} \|\omega_i\|_\infty, \ldots, \max_{i=1,\&,I} \|\omega_i\|_\infty)$. *Claim*: $(\text{con } B) \cap \left(v + \mathbf{R}_{--}^L\right) = \varnothing$. If $z \in \text{con}B$, by the Shapley–Folkman theorem, and relabelling the agents, we may write

$$z = \sum_{i=1}^{I} z_i, \quad z_i \in \text{con } B_i \, (i = 1, \ldots, I), \quad z_i \in B_i \, (i \neq \{1, \ldots, L\})$$

  Choose

$$\hat{z}_i = \begin{cases} 0 & \text{if } i = 1, \ldots, L \\ z_i & \text{if } i = L+1, \ldots, I \end{cases}$$

  Then $\sum_{i=1}^{I} \hat{z}_i \in B$ so $\sum_{i=1}^{I} \hat{z}_i < I < 0$. If $z \ll v$, then $\sum_{i=1}^{I} \hat{z}_i = \sum_{i=1}^{L} 0 + \sum_{i=L+1}^{I} z_i \leq \sum_{i=1}^{L} (\omega_i + z_i) + \sum_{i=L+1}^{I} z_i = \sum_{i=1}^{L} \omega_i + \sum_{i=1}^{I} z_i = \sum_{i=1}^{L} \omega_i + z \ll \sum_{i=1}^{L} \omega_i + v \leq 0$, so $B \cap \mathbf{R}_{--}^L \neq \varnothing$, a contradiction which proves the claim.
- By the separating hyperplane theorem, there exists $p \neq 0$ such that $\sup p \cdot \left(v + \mathbf{R}_{--}^L\right) \leq \inf p \cdot (\text{con } B)$. If $p_\ell < 0$ for some ℓ, then $\sup p \cdot \left(v + \mathbf{R}_{--}^L\right) = +\infty$, while $\inf p \cdot (\text{con } B) \leq 0$, a contradiction, so p=0 and we can normalize $p \in \Delta$. Then $\inf p \cdot B \geq \inf p \cdot (\text{con } B) \geq p \cdot v = -L\max\{\|\omega_1\|_\infty, \ldots, \|\omega_I\|_\infty\}$.
- Adapt the remainder of the proof of the Second Welfare Theorem; this requires a few tricks.

See Also

- Arrow–Debreu model of general equilibrium
- cores
- Edgeworth, Francis Ysidro
- existence of general equilibrium
- general equilibrium
- general equilibrium (new developments)

Bibliography

Anderson, R.M. 1978. An elementary core equivalence th. *Econometrica* 46, 1483–7.

Anderson, R.M. 1981. Core theory with strongly convex preferences. *Econometrica* 49, 1457–68.

Anderson, R.M. 1985. Strong core theorems with nonconvex preferences. *Econometrica* 53, 1283–94.

Anderson, R.M. 1986. Core allocations and small income transfers. Working Paper No. 8621, Department of Economics, University of California at Berkeley.

Anderson, R.M. 1987. Gap-minimizing prices and quadratic core convergence. *Journal of Mathematical Economics* 16, 1–15. Correction, *Journal of Mathematical Economics* 20(1991), 599–601.

Anderson, R.M. 1992. The core in perfectly competitive economies. In *Handbook of Game Theory with Economic Applications*, vol. I, ed. R. J. Aumann and S. Hart. Amsterdam: North-Holland.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Bewley, T.F. 1973a. Edgeworth's conjecture. *Econometrica* 41, 425–54.

Brown, D.J. and Robinson, A. 1974. The cores of large standard exchange economies. *Journal of Economic Theory* 9, 245–54.

Brown, D.J. and Robinson, A. 1975. Nonstandard exchange economies. *Econometrica* 43, 41–55.

Debreu, G. 1975. The rate of convergence of the core of an economy. *Journal of Mathematical Economics* 2, 1–7.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 236–46.

Dierker, E. 1975. Gains and losses at core allocations. *Journal of Mathematical Economics* 2, 119–28.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Grodal, B. 1975. The rate of convergence of the core for a purely competitive sequence of economies. *Journal of Mathematical Economics* 2, 171–86.

Grodal, B. and Hildenbrand, W. 1974. Limit theorems for approximate cores. Working Paper IP-208, Center for Research in Management, University of California, Berkeley.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Kannai, Y. 1970. Continuity properties of the core of a market. *Econometrica* 38, 791–815.

Khan, M.A. 1974. Some equivalence theorems. *Review of Economic Studies* 41, 549–65.

Vind, K. 1964. Edgeworth allocations in an exchange economy with many traders. *International Economic Review* 5, 165–77.

Vind, K. 1965. A theorem on the core of an economy. *Review of Economic Studies* 32, 47–8.

Walras, L. 1874. *Eléments d’économie politique pure*. Lausanne: L. Corbaz.

How to cite this article

Anderson, Robert M. "core convergence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000554> doi:10.1057/9780230226203.0320

The New Palgrave Dictionary of Economics Online

cores

Werner Hildenbrand
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The *core* of an economy consists of those states of the economy which no group of agents can ‘improve upon’. The core is a rather theoretical fundamental equilibrium concept. Indeed, the core provides a theoretical foundation of a more operational equilibrium concept, namely, the competitive equilibrium, which is a very different notion of equilibrium.

Keywords

barter; coalitions; Cobb–Douglas functions; competitive equilibrium; continuum of agents; cooperative game theory; cooperative game theory (core); cores; Edgeworth, F.; Hildenbrand, W.; initial endowments; large economies; limit economy; limit theorems; Minkowski's separation th; Pareto efficiency; preference relations; replica economies; type economies

Article

The *core* of an economy consists of those states of the economy which no group of agents can ‘improve upon’. A group of agents can improve upon a state of the economy if, by using the means available to that group, each member can be made better off. Nothing is said in this definition of how a state in the core actually is reached. The actual process of economic transactions is not considered explicitly.
To keep the presentation as simple as possible, we shall consider only the core for exchange economies with an arbitrary number *l* of commodities, even though the core concept applies to more general situations.

Consider a finite set *A* of economic agents; each agent *a* in *A* is described by his *preference relation* $\precsim_a$ (defined on the positive orthant R^l_+) and his *initial endowments* e_a (a vector in R^l_+). The outcome of any exchange, that is to say, a state (x_a) of the exchange economy $\mathcal{E} = \{\precsim_a, e_a\}_{a \in A}$, is a *redistribution* of the total endowments, i.e.

$$\sum_{a \in A} x_a = \sum_{a \in A} e_a .$$

A *coalition* of agents, say $S \subset A$, can *improve upon* a redistribution (x_a), if that coalition *S*, by using the endowments available to it, can make each member of that coalition better off, that is to say, there is a redistribution, say $(y_a)_{a \in S}$, such that

$$y_a \succ_a x_a \text{ for every } a \in S \text{ and } \sum_{a \in S} y_a = \sum_{a \in S} e_a .$$

The set of redistributions for the exchange economy $\mathcal{E}$ that no coalition can improve upon is called the *core* of the economy $\mathcal{E}$, and is denoted by $C(\mathcal{E})$.
The core is a rather theoretical, however, fundamental equilibrium concept. Indeed, the core provides a theoretical foundation of a more operational equilibrium concept, the *competitive equilibrium* which, in fact, is a very different notion of equilibrium. The allocation process is organized through markets; there is a price for every commodity. All economic agents take the price system as given and make their decisions independently of each other. The equilibrium price system coordinates these independent decisions in such a way that all markets are simultaneously balanced.

More formally, an allocation (x_a^*) for the exchange economy $\mathcal{E} = \{\precsim_a, e_a\}_{a \in A}$ is a *competitive equilibrium* (or a *Walras allocation*) if there exists a price vector $p^* \in R^l_+$ such that for every $a \in A$, $x_a^* \in \varphi_a(p^*)$ and

$$\sum_{a \in A} x_a^* = \sum_{a \in A} e_a .$$

Here $\varphi_a(p^*)$ or more explicitly, $\varphi(p^*, e_a, \precsim_a)$ denotes the demand of agent *a* with preferences $\precsim_a$ and endowment e_a, i.e. the set of most desired commodity vectors (with respect to $\precsim_a$) in the budget-set $\left\{x \in R^l_+ \mid p^* \cdot x \leq p^* \cdot e_a\right\}$.
The set of all competitive equilibria for the economy $\mathcal{E}$ is denoted by $W(\mathcal{E})$.
The core and the set of competitive equilibria for an economy with two agents and two commodities can be represented geometrically by the well-known Edgeworth–Box (see Figure 1). The size of the box is determined by the total endowments e_1+e_2. Every point *P* in the box represents a redistribution; the first agent receives $x_1 = P$ and the second receives $x_2 = (e_1 + e_2) - P$.
Figure 1

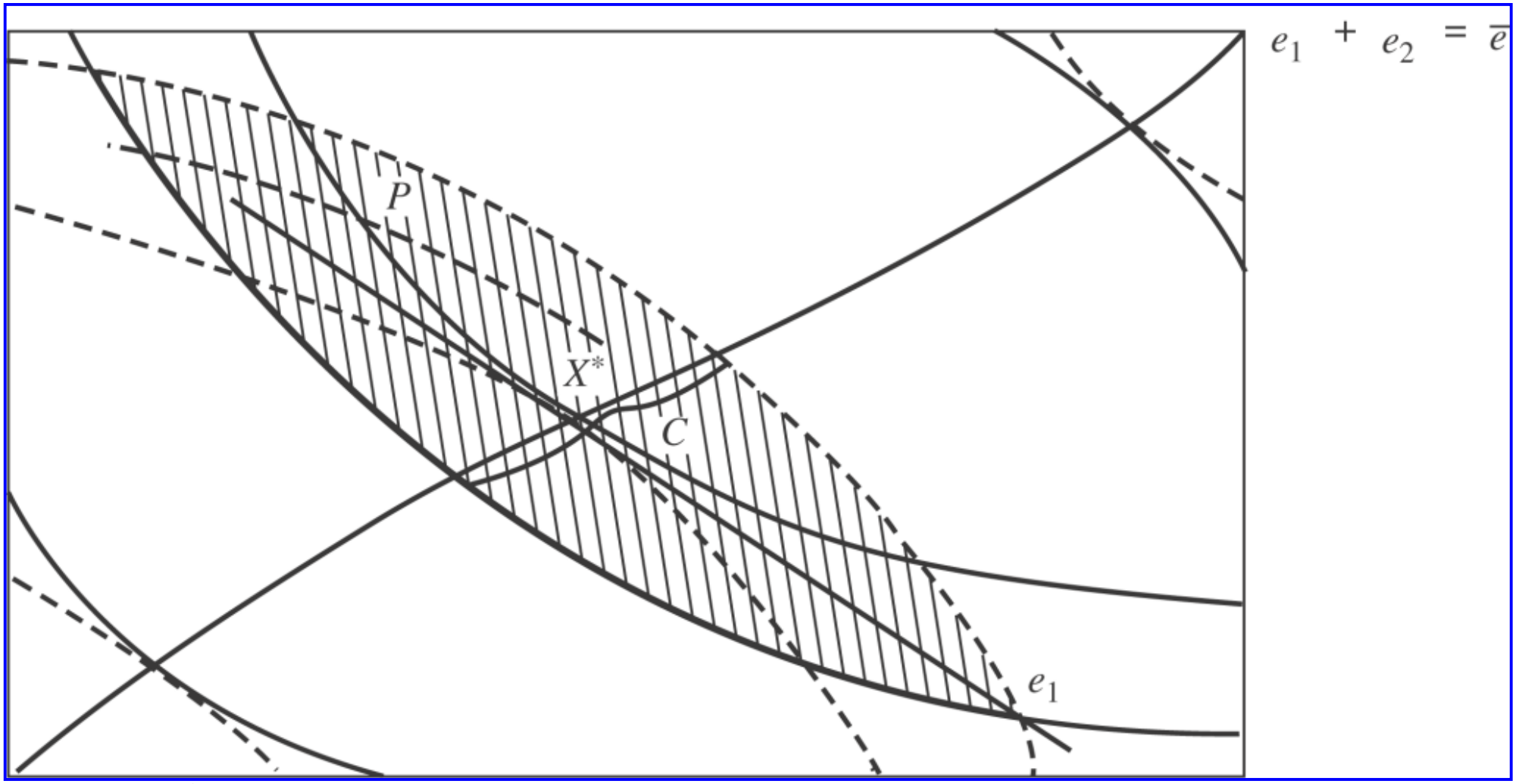

It is easy to show that for every exchange economy $\mathcal{E}$ a competitive equilibrium belongs to the core,

$$W(\mathcal{E}) \subset C(\mathcal{E}).$$

Thus, a state of the economy $\mathcal{E}$ which is decentralized by a price system cannot be improved upon by cooperation. This proposition strengthens a well-known result of Welfare Economics—every competitive equilibrium is Pareto-efficient.

The inclusion $W(\mathcal{E}) \subset C(\mathcal{E})$ is typically strict. Indeed, if the initial allocation of endowments is not Pareto-efficient, which is the typical case, then, if there are any allocations in the core at all, there are core-allocations which are not competitive equilibria.

This leads us to the *basic problem* in the theory of the core:

> For which kind of economies is the ‘difference’ between the core and the set of competitive equilibria small? Or in other words, under which circumstances do cooperative barter and competition through decentralized markets lead essentially to the same result?

Naturally, the answer depends on the way one measures the ‘difference’ between the two equilibrium concepts. However this is done one expects that the economy must have a large number of participants.

In answering the basic question we try to be comprehensible (for example by avoiding the use of measure-theoretic concepts) but not comprehensive. Therefore, if we refer in the remainder of this entry to an economy $\mathcal{E} = \{\precsim_a, e_a\}_{a \in A}$ we shall always assume that preference relations are continuous, complete, transitive, monotone and strictly convex. The total endowments $\sum_{a \in A} e_a$ of an economy are always assumed to be strictly positive. We shall not repeat these assumptions. Furthermore, if we call an economy smooth, then we assume in addition that preferences are smooth (hence representable by sufficiently differentiable utility functions) and individual endowments are strictly positive.

These assumptions simplify the presentation tremendously. For generalizations we refer to the extensive literature.

We remark that under the above assumptions there always exists a competitive equilibrium, and hence, the core is not empty.

Large economies

The simplest and most stringent measure of difference between the two equilibrium sets, $C(\mathcal{E})$, and $W(\mathcal{E})$, which we shall denote by $\delta(\mathcal{E})$, can be defined as follows.

Let $\delta(\mathcal{E})$ be the smallest number δ with the property: for every allocation $(x_a) \in C(\mathcal{E})$ there exists an allocation $(x_a^*) \in W(\mathcal{E})$ such that

$$\left| x_a - x_a^* \right| \leq \delta$$

for every agent *a* in the economy $\mathcal{E}$.

Thus, if $\delta(\mathcal{E})$ is small, then from every agent's view a core allocation is like a competitive equilibrium.

Unfortunately for this measure of difference, it is not true that $\delta(\mathcal{E})$ can be made arbitrarily small provided the number of agents in the economy $\mathcal{E}$ is sufficiently large (even if one restricts the agents' characteristics $(\precsim_a, e_a)$ to an a priori given finite set).

Consequently one considers also weaker measures for the ‘difference’ between the two equilibrium concepts $C(\mathcal{E})$ and $W(\mathcal{E})$. For example, define $\delta_1(\mathcal{E})$ and $\delta_2(\mathcal{E})$, respectively, as the smallest number δ with the property: for every $(x_a) \in C(\mathcal{E})$ there exists a price vector $p \in R^l_+$ such that

$$(\delta_1) |x_a - \varphi_a(p)| \leq \delta \text{for every agent } a \text{ in } \mathcal{E}$$

or

$$(\delta_2) \frac{1}{\# A} \sum_{a \in A} |x_a - \varphi_a(p)| \leq \delta .$$

Clearly, the measures δ_1 and δ_2 are weaker than δ since the price vector *p* is not required to be an equilibrium price vector for the economy $\mathcal{E}$. The number $\delta_1(\mathcal{E})$ (and, *a fortiori*, $\delta_2(\mathcal{E})$) does not measure the distance between the sets $C(\mathcal{E})$ and $W(\mathcal{E})$. but the degree by which an allocation in the core can be decentralized via a price system. Obviously one has $\delta_2(\mathcal{E}) \leq \delta_1(\mathcal{E}) \leq \delta(\mathcal{E})$.

One can show that $\delta_2(\mathcal{E})$ becomes arbitrarily small for sufficiently large economies. More precisely,

Theorem 1: Let *T* be a finite set of agents' characteristics $(\precsim, e)$ and let *b* be a strictly positive vector in R^l. Then for every $\varepsilon > 0$ there exists an integer *N* such that for every

economy $\mathcal{E} = \{ \precsim_a, e_a \}_{a \in A}$ with $\# A \geq N$,

$$\frac{1}{\# A} \sum_{a \in A} e_a \geq b$$

and $(\succsim_a, e_a) \in T$ one has

$$\delta_2(\mathcal{E}) \leq \epsilon.$$

(The finite set *T* in Theorem 1 can be replaced by a compact set with respect to a suitably chosen topology: see Hildenbrand, 1974.) We emphasize that this result does not imply that in large economies core-allocations are near to competitive equilibria. In fact, Theorem 1 does not hold if δ_2 is replaced by the measure of difference δ or even δ_1. Theorem 1 does imply, however, that for sufficiently large economies one can associate to every core-allocation a price vector which 'approximately decentralizes' the core-allocation. Some readers might consider this conclusion as a perfectly satisfactory answer to our basic problem. If one holds this view, then the rest of the paper is a superfluous intellectual pastime. We would like to emphasize, however, that the meaning of 'approximate decentralization' is not very strong. First, the demand $\varphi_a(p)$ is not necessarily near to x_a for every agent *a* in the economy; only the mean deviation

$$\frac{1}{\# A} \sum_{a \in A} |x_a - \varphi_a(p)|$$

becomes small. Second, total demand is not equal to total supply; only the mean excess demand

$$\frac{1}{\# A} \sum_{a \in A} [\varphi_a(p) - e_a]$$

becomes small.

There are alternative proofs in the literature, e.g. Bewley (1973), Hildenbrand (1974), Anderson (1981) or Hildenbrand (1982). These proofs are based either on a result by Vind (1965) or Anderson (1978).

Sharper conclusions than the one in Theorem 1 will be stated in the following sections. There we consider a sequence $(\mathcal{E}_n)_{n=1,\ldots}$ of economies and then study the asymptotic behaviour of $\delta(\mathcal{E}_n)$.

Before we present these limit theorems we should mention another approach of analysing the inclusion $W(\mathcal{E}) \subset C(\mathcal{E})$. Instead of analysing the asymptotic behaviour of the difference $\delta(\mathcal{E}_n)$ for a sequence of finite economies one can define a large economy where every agent has strictly no influence on collective actions. This leads to a *measure space without atoms* of economic agents (also called a *continuum of agents*). For such economies the two equilibrium concepts coincide. See Aumann (1964).

Replica economies

Let $\mathcal{E} = \{\precsim_i, e_i\}$ be an exchange economy with *m* agents. For every integer *n* we define the *n*-fold *replica economy* $\mathcal{E}_n$ of $\mathcal{E}$ as an economy with *n·m* agents; there are exactly *n* agents with characteristics $(\precsim_i, e_i)$ for every $i = 1, \ldots, m$.
More formally,

$$\mathcal{E}_n = \left\{\precsim_{(i,j)}, e_{(i,j)}\right\}_{\substack{1 \le i \le m \\ 1 \le j \le x}}$$

where $\precsim_{(i,j)} = \precsim_i$ and $e_{(i,j)} = e_i$, $1 \le i \le m$ and $1 \le j \le n$. Thus, an agent *a* in the economy $\mathcal{E}_n$ is denoted by a double index $a = (i, j)$. We shall refer to agent (*i*,•*j*) sometimes as the *j*th agent of type *i*.
Replica economies were first analysed by F. Edgeworth (1881) who proved a limit theorem for such sequences in the case of two commodities and two types of agents. A precise formulation of Edgeworth's analysis and the generalization to an arbitrary finite number of commodities and types of agents is due to Debreu and Scarf (1963).
Here is the basic result for replica economies.
Theorem 2: For every sequence $(\mathcal{E}_n)$ of replica economies the difference between the core and the set of competitive equilibria tends to zero, i.e.,

$$\lim_{n \to \infty} \delta(\mathcal{E}_n) = 0.$$

Furthermore, if $\mathcal{E}$ is a smooth and regular economy then $\delta(\mathcal{E}_n)$ converges to zero at least as fast as the inverse of the number of participants, i.e., there is a constant *K* such that

$$\delta(\mathcal{E}_n) \le \frac{K}{n}.$$

The proof of this rmkably neat result is based on the fact that a core–allocation (x_{ij}) assigns to every agent of the same type the same commodity bundle, i.e., $x_{ij} = x_{ik}$, This ‘equal treatment’ property simplifies the analysis of $\delta(\mathcal{E}_n)$ tremendously. Indeed, an allocation (x_{ij}) in $C(\mathcal{E}_n)$, which can be considered as a vector in $R^{l \cdot m \cdot n}$, is completely described by the commodity bundle of one agent in each type, thus by a vector $(x_{11}, x_{21}, \ldots, x_{m1})$ in $R^{l \cdot m}$, a space whose dimension is independent of *n*.
Thus, let

$$C_n = \left\{(x_{11}, x_{21}, \ldots, x_{m1}) \in R^{l \cdot m} \mid (x_{ij}) \in C(\mathcal{E}_n)\right\}.$$

One easily shows that $C_{n+1} \subset C_n$. It is not hard to see that Theorem 1 follows if

$$\cap_{n=1}^{\infty} C_n = W(\mathcal{E}_1).$$

But this is the well-known theorem of Debreu and Scarf (1963). The essential arguments in the proof go as follows. Let $(x_1, \ldots, x_m) \in \cap_{n=1}^{\infty} C_n$. One has to show that there is a price vector p^* such that $x \succ_i x_i$ implies $p^* \cdot x > p^* \cdot e_i$. For this it suffices to show that there is a p^* such that

$$p^* \cdot z \geq 0 \quad \textit{for every } z \in \cup_{i=1}^{m} \left(\left\{ x \in R^l_+ \mid x \succ_i x_i \right\} - e_i \right) = Z,$$

i.e., there is a hyperplane (whose normal is p^*) which supports the set z. One shows that the assumption $(x_1, \ldots, x_m) \in \cap_{n=1}^{\infty} C_n$ implies that 0 does not belong to the convex hull of z. Minkowski's Separation Theorem for convex sets then implies the existence of the desired vector p^*.
The second part of the conclusion of Theorem 2 is due to Debreu (1975).

Type economies

The limit theorem on the core for replica economies is not fully satisfactory since replication is a very rigid way of enlarging an economy. The conclusion '$\delta(\mathcal{E}_n) \to 0$' in Theorem 2, to be of general relevance, should be robust to small deviations from the strict replication procedure.
Consider a sequence $(\mathcal{E}_n)$ of economies where the characteristics of every agent belong to a given finite set of types $T = \{(\precsim_1, e_1), \ldots, (\precsim_m, e_m)\}$. We do not consider this as a restrictive assumption (considered as an approximation, one can always group agents' characteristics into a finite set of types). Let the economy $\mathcal{E}_n$ have N_n agents; $N_n(1)$ agents of the first type, $N_n(i)$ agents of type i. Of course the idea is that N_n tends to ∞ with increasing n. Consider the fraction $v_n(i)$ of agents in the economy $\mathcal{E}_n$ which are of type i, i.e.,

$$v_n(i) = \frac{N_n(i)}{N_n}.$$

The sequence $(\mathcal{E}_n)$ is a replica sequence of an economy $\mathcal{E}$ (not necessarily of $\mathcal{E}_1$) if and only if the fractions $v_n(i)$ are all independent of n. It is this rigidity which we want to weaken now.
A sequence $(\mathcal{E}_n)$ of economies with characteristics in a finite set T is called a *sequence of type economies* (over T) if

1. (i) the number N_n of agents in $\mathcal{E}_n$ tends to infinity and
2. (ii) $v_n(i) = \frac{N_n(i)}{N_n} \underset{(n \to \infty)}{\square} v(i) > 0$.

ex (random sampling of agents' characteristics):
Let π be a probability distribution over the finite set T. Define the economy $\mathcal{E}_n$ as a random sample of size n from this distribution $\pi(\cdot)$. The law of large numbers them implies property (ii): $v_n(i) \to \pi(i)$.
The step from replica economies to type economies – as small as it might appear to the reader – is conceptually very important. Yet with this 'small' generalization the analysis of the limit behaviour of $\delta(\mathcal{E}_n)$ or $\delta_1(\mathcal{E}_n)$ is made more difficult. Even worse, it is no longer true that for *every* sequence $(\mathcal{E}_n)$ of type economies one obtains $\delta(\mathcal{E}_n) \to 0$ — even if the preferences of all types are assumed to be very nice, say smooth. There are some 'exceptional cases' where the conclusion $\delta(\mathcal{E}_n) \to 0$ does not hold. But these are 'exceptional' cases

and the whole difficulty in the remainder of this section is to explain in which precise sense these cases are ‘exceptional’ and can therefore be ignored. We shall first exhibit the ‘cases’ where the conclusion fails to hold. Then we shall show that these cases are exceptional.

We denote by $\Pi(\mathcal{E})$ the set of normalized *equilibrium price vectors* for the economy $\mathcal{E} = \{ \precsim_a, e_a \}_{a \in A}$. Thus, for $p^* \in \Pi(\mathcal{E})$ the excess demand is zero, i.e.,

$$\sum_{a \in A} [\varphi_a(p^*) - e_a] = 0 .$$

To every sequence $(\mathcal{E}_n)$ of type economies we associate a ‘*limit economy*’ $\mathcal{E}_\infty$. This economy has an ‘indefinitely large’ number of agents of every type; the fraction of agents of type *i* is given by *v(i)*. The mean (per capita) excess demand of that limit economy $\mathcal{E}_\infty$ is defined by

$$z_v(p) = \sum_{i=1}^{m} v(i) [\varphi(p, e_i, \precsim_i) - e_i] .$$

An equilibrium price vector p^* of the limit economy $\mathcal{E}_\infty$ is defined by $z_v(p^*) = 0$. Let $\Pi(v)$ denote the set of normalized equilibrium price vectors for $\mathcal{E}_\infty$. Obviously for a replica sequence $(\mathcal{E}_n)$ we have $\Pi(\mathcal{E}_n) = \Pi(v)$ for all *n*. However, for a sequence of type economies the set $\Pi(\mathcal{E}_n)$ of equilibrium prices of the economy $\mathcal{E}_n$ depends on *n*, and it might happen that the set Π (*v*) is not similar to $\Pi(\mathcal{E}_n)$ even for arbitrarily large *n*. To fix ideas, it might happen that $\Pi(\mathcal{E}_n) = \{ p_n \}$ and $\Pi(v)$ contains not only $p = \lim p_n$ but also another equilibrium price vector. Such a situation has to be excluded.

We call a sequence of type economies *sleek* if $\Pi(\mathcal{E}_n)$ converges (in the Hausdorff-distance) to $\Pi(v)$.

It is known (Hildenbrand, 1974) that the sequence $(\Pi(\mathcal{E}_n))$ converges to $\Pi(v)$ if $\Pi(v)$ is a singleton (i.e., the limit economy has a unique equilibrium) or, in general, if (and only if) for every open set *O* in R^ℓ with $O \cap \Pi(v) \neq \varnothing$ it follows that $O \cap \Pi(\mathcal{E}_n) \neq \varnothing$ for all *n* sufficiently large.

We now have exhibited the cases where a limit theorem on the core holds true.

Theorem 3: For every sleek sequence $(\mathcal{E}_n)$ of type economies

$$\lim_{n \to \infty} \delta(\mathcal{E}_n) = 0.$$

Unfortunately there seems to be no short and easy proof. The main difficulty arises from the fact that for allocations in the core of a type economy the ‘equal treatment’ property, which made the replica case so manageable is no longer true. For a proof see Hildenbrand and Kirman (1976) or Hildenbrand (1982) and the references given there. The main step in the proof is based on a result of Bewley (1973).

It remains to show that non-sleek sequences of type economies are ‘exceptional cases’.

The strongest form of ‘exceptional’ is, of course, ‘never’. We mentioned already that a sequence $(\mathcal{E}_n)$ is sleek if its limit economy has a unique equilibrium. Unfortunately, however, only under very restrictive assumptions on the set *T* of agents' characteristics does uniqueness prevail; for example,

1. (1) if every preference relation leads to a demand function which satisfies gross-substitution (Cobb–Douglas utility functions are typical exs),
2. (2) if every preference relation is homothetic and the endowment vectors $e_i (i = 1, \ldots, m)$ are collinear.

Since there is no reasonable justification for restricting the set *T* to such special types of agents we have to formulate a model in which we allow non-sleek sequences to occur

provided, of course, this can be shown to be ‘exceptional cases’. Let S^{m-1} denote the open simplex in R^m, i.e.

$$S^{m-1} = \left\{ x \in R^m | x_i > 0, \sum_{i=1}^{m} x_i = 1 \right\}.$$

The limit distribution $v(i)$ of a sequence of type economies with m types is a point in S^{m-1}.
A closed subset C in S^{m-1} which has (m – 1 dimensional Lebesgue) measure zero is called *negligible*. Thus, if a distribution v is not in C then a sufficiently small change will not lead to C. Furthermore, given any arbitrary small positive number ε one can find a countable collection of balls in S^{m-1} such that their union covers C, and that the sum of the diameters of these balls is smaller than ε. Thus, in particular, if $v \in C$ then one can approximate v by points which do not belong to C. Clearly, a negligible set is a small set in S^{m-1}.
Theorem 4: Given a finite set T of m smooth types of agents, there exists a negligible subset C in S^{m-1} and a constant K such that for every sequence (ξ_n) of type economies over T whose limit distribution v does not belong to C one has $\delta(\xi_n) \leq K / \# A_n$, thus in particular, $\lim_{n \to \infty} \delta(\xi_n) = 0$.
The convergence of $\delta(\xi_n)$ follows from Theorem 3 and Theorems 5.4.3 and 5.8.15 in Mas–Colell (1985). For the rate of convergence see Grodal (1975).

See Also

- Edgeworth, Francis Ysidro
- existence of general equilibrium

Bibliography

There is an extensive literature on limit theorems on the core which contains important generalizations of the results given here. For a general reference we refer to Hildenbrand (1974) or (1982), Mas-Colell (1985), Anderson (1981) and the references given there.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46, 1483–7.

Anderson, R.M. 1981. Core theory with strongly convex preferences. *Econometrica* 49, 1457–68.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Bewley, T.F. 1973. Edgeworth's conjecture. *Econometrica* 41, 425–54.

Debreu, G. 1975. The rate of convergence of the core of an economy. *Journal of Mathematical Economics* 2, 1–8.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Grodal, B. 1975. The rate of convergence of the core for a purely competitive sequence of economies. *Journal of Mathematical Economics* 2, 171–86.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Hildenbrand, W. 1982. Core of an economy. In *Handbook of Mathematical Economics*, ed. K.J. Arrow and M.D. Intriligator. Vol. 2, Amsterdam: North-Holland.

Hildenbrand, W. and Kirman, A.P. 1976. *Introduction to Equilibrium Analysis*. Amsterdam: North-Holland.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium, A Differentiable Approach*. Cambridge: Cambridge University Press.

Vind, K. 1965. A theorem on the core of an economy. *Review of Economic Studies* 32, 47–8.

How to cite this article

Hildenbrand, Werner. "cores." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000361>
doi:10.1057/9780230226203.0321

The New Palgrave Dictionary of Economics Online

Corn Laws, free trade and protectionism

John Nye
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the 1840s, Britain repealed the export restrictions and import duties on wheat known as the Corn Laws. But the traditional story of British free trade was complicated by an unwillingness to eliminate the most binding tariffs on wine and other consumables. In contrast, Britain's avowedly protectionist rival France had a more liberal trade policy than did Britain for most of the 19th century. Only with the 1860 Anglo–French Treaty of Commerce did Britain and France both move to uniformly low tariffs on goods and services, ushering in a period of genuinely free trade throughout Europe.

Keywords

Anti-Corn Law League; Chevalier, M.; Cobden, R.; comparative advantage; Corn Laws; excise taxes; free trade; income tax; Industrial Revolution; infant-industry protection; mercantilism; Mun, T.; protection; Ricardo, D.; Scottish Enlightenment; Smith, A.; specie; Tariffs; trade deficit; World Trade Organization

Article

The Corn Laws were the parliamentary statutes that regulated the import and export of grain for the benefit of British producers in the early 19th century. Though these laws derived from legislation in the period 1804–15, they were but the extension or modification of a system that had been introduced in 1773 to prohibit exports of wheat when prices rose above a given level and that limited imports through a variety of duties based on a sliding scale. The goal of these laws was ostensibly the desire to stabilize the price of grain, which had been a regular goal of parliament since the late 17th century.
The debates about the abolition of the Corn Laws in the early to mid-1800s hold a special place in the economic history of Great Britain on account of their central role in shifting commercial policy to nearly free trade. Because of Britain's dominance of industrial trade in the 19th century and the leadership she exerted in international commerce, the struggles over the Corn Laws have been seen as emblematic of

all debates about the advisability of free trade or protectionism. Despite the symbolic importance of these events, it is easy to overlook the facts that Britain after the repeal of the Corn Laws did not immediately move to perfectly free trade and that the political struggle over their abolition had at least as much to do with domestic concerns over the importance of agriculture in a modern economy as with ideological questions about the advisability of free trade.

Mercantilism and the rise of British liberalism

The regulation or promotion of international trade has been perhaps the oldest policy issue in the political economy of international relations.

It is a common belief that trade is a primary source of a nation's wealth. But this has often been misunderstood to mean that exports enhance wealth while imports detract from it. This view, a central component of what is called mercantilism, stems from the mistaken belief that the benefits of trade flow only one way. One view was that a nation's wealth derived from the quantity of specie or gold and silver coin in the country. Therefore, exports contributed to this while imports detracted from it.

Some of this reasoning was theoretical, but more commonly mercantile theory was simply the evolution of a set of policies deriving from the fiscal needs of the newly emerging nation-states. Unsurprisingly, many states viewed the success of the state as synonymous with the success of the nation itself. Revenue was essential to the maintenance of the large armies that were a prerequisite for the nation-state. So trade was viewed as an essentially zero-sum game with both losers and winners. Moreover, this concern about revenue often translated into a concern for specie. Whereas modern economics treats specie as virtually irrelevant to the supply of money, contemporaries viewed coin itself as a necessary prerequisite of sound financial policy. Hence trade surpluses were preferred because they brought more precious metals in than they took out of the kingdom.

One of the earliest theoretical discussions of this view comes from Thomas Mun, who wrote ‘The ordinary means therefore to increase our wealth and treasure is by foreign trade, wherein we must ever observe this rule; to sell more to strangers yearly than we consume of theirs in value’ (1664, p. 11). Adam Smith, the founder of modern economics, was the most prominent critic of this view. Starting from the observation that voluntary trade was mutually beneficial, and noting that the wealth of a nation's inhabitants, not its quantity of coin, made for true wealth, Smith argued in the *Wealth of Nations* against what he labelled the ‘mercantile system’. He articulated the virtues of free and open trade, both in international and in home commerce. Indeed, the term ‘free trade’ was employed throughout the 18th and 19th centuries to refer to unregulated domestic trade as least as often as it referred to the free flow of goods from abroad.

These ideas were later modelled more systematically by the English economist David Ricardo, who formalized the analysis and showed that nations could maximize their welfare by specializing in the production of goods with the lowest opportunity cost and trading with other nations. This is the central idea behind the law of comparative advantage, usually attributed to David Ricardo, and developed more thoroughly by Paul Samuelson and others in the 20th century. Most important for this claim was the idea that a nation did not even have to be the ‘best’ producer of any product for there to be gains from trade. A nation that was more productive that another in all industries would still do better by specializing in some areas and trading for the other goods with another country. Thus, any claim that a nation could not benefit if it had no comparative advantage would be false. Every nation has a comparative advantage in

producing some product, even if it has an absolute advantage in none.

Smith's ideas and those of his successors provided the philosophical basis for the classical liberal movements of the late 18th and early 19th centuries. By the early 1800s, the idea of a limited state that minimized regulation and promoted welfare through the encouragement of open trade at home and abroad had emerged as an important ideological view, promoted by prominent intellectuals and supported by an influential subset of the British political class. Nonetheless, the strong interest in the liberal ideas derived from the Scottish Enlightenment persuaded states not to fully adopt a policy of free trade. This was often not so much the result of any ideological predisposition as a response to the state's desire for greater revenue. Taxing trade – both at home and from abroad – was one of the most common means of generating the income that supported the expanding bureaucracy of the modern state. Furthermore, special interests often worked to distort policy to favour of specific producers or economic sectors.

Since the late 17th century, Britain had been especially dependent on customs and excise taxes of various sorts. The rise of British liberalism had come in the same century (the 18th) that had seen the British state grow to an unprecedented size. Growth of government revenue had vastly outstripped the rate of overall economic growth and served to fund a professional bureaucracy at home and an expanding imperialist policy abroad. This enabled the British to either defeat or stalemate their traditional rival, France, in a series of military struggles that extended from the late 1600s to the era of Napoleon a century later. Moreover, this expansion of the central government came with little change in the revenues from land, the traditional source of income. Most of the gains came from steep increases in revenue from trade; and rising excises were some of the abuses cited by the American colonists as the basis for the independence movement.

However, changes in the landscape of the British economy – most notably the urban and industrial expansion that began in the late 1700s and is known as the Industrial Revolution – made Britain the premier industrial producer of the early 19th century and put pressure on the government to transform legislation that had kept agricultural prices high and had limited imports for the benefit of the farmers who were an increasingly small share of the economy.

The 19th-century Corn Law repeal: free trade rhetoric vs. protectionist reality

The interests of industrial producers who felt that workers would be better served by cheap bread and the ideas of liberal elites saw concrete expression in the creation of the Anti-Corn Law League beginning in the 1830s. Statesmen such as Richard Cobden explicitly saw the movement as the first step in an attempt to push the British government to adopt a general policy of free trade.

However, it is not clear that theoretical ideas played a large role in the actual dismantling of the Corn Laws. Furthermore, Smith had always held up the staple industries and national defence as areas that might be exceptions to the doctrine of pure laissez-faire. However, the end of the Napoleonic Wars in 1815 removed the basis for wartime support of the Corn Laws and pushed the government to consider modifying or abolishing the restrictions in a transition to a peacetime economy.

As early as 1821 the government of Lord Liverpool had begun to consider reforming a system that it regarded as temporary and motivated by a desire to secure stable prices during wartime with a mix of regulation and protection. The Corn Laws did not seem to be fulfilling that function and, in the absence of war, their maintenance seemed unnecessary for the public good. Of course, the farm interests that

gained from these rules would have fought for the continuation of these protections. Nonetheless, the increased voting power of urban workers empowered by the 1832 Reform Act reinforced Prime Minister Peel's conviction that support for industry was vital to the future development of Britain and led him to push for the abolition of all Corn Laws in the 1840s. The onset of the Irish potato famine in 1845 gave a special impetus to the desire to promote lower prices for basic staples and allowed Peel to push for the full abolition of the Corn Laws in 1846.

This legislation repealing the Corn Laws is often cited as the pivotal moment in the rise of free trade in Britain and in Europe because it was followed over the next decade with the reduction or removal of duties on hundreds of imports in Britain – hence the claim that henceforth Britain moved swiftly to full free trade. However, this accomplishment has been somewhat exaggerated in conventional history. Partly because of the need for continued revenue and partly because of pressure from special interests, a few large and important tariffs on coffee, tea, wine, spirits, sugar and tobacco continued up to the 1860s, tariffs which had a disproportionate impact on the trade of Britain.

The 1860 Anglo-French Trade Treaty and the true coming of free trade

The wine and spirit tariffs were especially important and had been mentioned prominently in Smith's criticism of the mercantile system in the *Wealth of Nations*. These tariffs had arisen from Britain's desire to punish her rival France and had developed as a means of protecting domestic beverage interests such as beer and gin at home, and colonial imports such as rum. Lacking an equivalent slogan to that of the cry for ‘cheap bread’, there was no great movement to reform these substantial duties.

Consequently, despite the British reputation as the leading free trader in the 19th century, Britain in fact had higher average tariffs than the more openly interventionist nation of France for the first three quarters of that century. The burden on the working classes from the combination of high tariffs on imported wine and liquor and the regulation and taxation of domestic production meant that consumption of alcohol was repressed throughout the 18th and early 19th centuries, despite all the income gains during the Industrial Revolution. Where basic alcoholic beverages had been seen as a necessary staple in the 17th century, they were more likely to be treated as luxuries in the 19th.

Full reform had to wait until 1860, when Britain and France concluded the Anglo-French Treaty of Commerce. This landmark treaty can be said to have truly ushered in the age of free trade in Europe. Brokered by Cobden in Britain and Michel Chevalier in France, the treaty had come after many years of negotiation. Early overtures to the French to sign such a treaty had been rebuffed in the 1840s because Britain had been unwilling to compromise their duties on wine – which had been the category of greatest concern to the French. However, changes in British fiscal structure arising from the imposition of an income tax in the 1850s made it easier for the British government to contemplate tariff cuts that might have compromised the budget in the short run. (British Liberals believed that given enough time, lower rates on imported wine would be offset by increased trade, a belief that proved accurate.) Moreover, the political considerations that led to wine duties being designed from the early 1700s to favour the products of friendly nations such as Portugal and Spain over that of France grew less important in the decades of peace following the defeat of Napoleon Bonaparte in 1815.

Thus, it became possible to conclude a treaty in 1860 in which Britain lowered and modified all its wine and spirit tariffs to remove any anti-French bias and caused France to lower tariffs and remove all prohibitions on goods – primarily textiles – imported from Britain. The 1860 Treaty was also significant

for being a Most Favoured Nation agreement in which any subsequent treaties with third countries negotiated by either party would cause concessions to be applied equally to the original signatories. Concern by other Western nations that they would be left out of a trading arrangement between the two leading European powers led to almost the whole of Europe concluding equivalent treaties with either Britain or France over the next decade. By the 1870s virtually the whole of Europe was an extremely open trading area with free movement of goods, capital, and labour that in some ways has never been matched even by today's European Union. And by the end of the 19th century Britain could be said to have genuinely become a free trader with few or modest tariffs on most items, and possibly the lowest average tariffs in all Europe.

It is also interesting to note that Britain provides something of a counter-example to the tendency of modern-day protectionists to fret about the trade balance. Britain was the undoubted leader in world trade throughout the 19th century yet she also ran a merchandise trade deficit for virtually the whole of that period up to the First World War.

The one major counter-example to the tendency in the West to move towards freer international commerce had been the United States. Whereas Europe was busy lowering or abolishing tariffs and trade restrictions after 1860, the USA raised tariffs substantially from the 1860s onwards. Tariffs were the major source of revenue for the federal government before the constitutional amendment that permitted an income tax. Furthermore, the civil war gave control of the government to the Republicans under Lincoln, who had made protection an important plank in the party's platform. To some extent the United States was fortunate in that many of the negative potential effects of the tariffs were somewhat offset by the free movement of capital, the large size of the internal US market, and the benefits of an extremely open immigration policy. Thus, while goods trade was restricted, capital and labour remained mostly mobile.

By the end of the 19th century, however, the free trade regime brought on by the 1860 Anglo-French Treaty began to unravel. As early as 1878 Germany began to modify her agricultural tariffs in response to pressure from farmers due to increased competition from Russia and the United States. French textile manufacturers pushed the government to abandon the treaty in 1882 and a new set of tariffs were put into place at the beginning of 1892. However, it is worth noting that in both cases the resulting tariff regimes were still relatively moderate and not comparable to the high protection of early Britain or mid-19th-century USA, and Europe still enjoyed vigorous exchange up to 1914, when the European system of open trade was effectively destroyed, first by the war and then by the high tariff walls that nations began to enact during the Great Depression.

The 19th-century trade debates have remained an important touchstone for both scholars and political elites. The same general issues persist to this day. How vigorously should a nation pursue free trade? Is it best to liberalize unilaterally or bilaterally with treaties or collectively through groups like the World Trade Organization? Today we continue to hear concerns about the importance of the trade deficit in hampering or restraining economic growth. Large and small nations often invoke the need to protect infant industries as a justification for tariffs, although it is interesting that in most cases throughout the world it is ageing and decaying industries that are likely to receive protection rather than the newer, more innovative sectors of the economy. And, as with Great Britain in the 19th century, the USA today is seen as the leader in world trade, with some of the same questions being asked about the extent to which trade is manipulated to improve world welfare or merely to enhance the narrow interests of the leading nations. And with the rise of treaties such as the North American Free Trade Agreement and the

Central American Free Trade Agreement, as well as the Eurozone, there remain questions as to the virtues of piecemeal reform or the extent to which these agreements are merely mechanisms for obstructing trade by parcelling out the world into separate trading blocs.

See Also

- globalization
- historical economics, British
- Smith, Adam

Bibliography

Hilton, B. 1977. *Corn, Cash and Commerce: The Economic Policies of the Tory Government 1815–1830*. Oxford: Oxford University Press.

Irwin, D.A. 1996. *Against the Tide: An Intellectual History of Free Trade*. Princeton: Princeton University Press.

Mun, T. 1664. *England's Treasure by Forraign Trade*. London: Printed by J.G. for Thomas Clark.

Nye, J.V.C. 1991. The myth of free trade in Britain and Fortress France: tariffs and trade in the nineteenth century. *Journal of Economic History* 51, 23–46.

Schonhardt-Bailey, C., ed. 1997. *The Rise of Free Trade. Vol. 1, Protectionism and its Critics, 1815–1837*. London: Routledge.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Clarendon Press, 1976.

How to cite this article

Nye, John. "Corn Laws, free trade and protectionism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000208> doi:10.1057/9780230226203.0322

The New Palgrave Dictionary of Economics Online

corporate governance

Luigi Zingales
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Introduced in the mid-1980s, the term ‘corporate governance’ can be defined as the set of conditions that shapes the *ex post* bargaining over the quasi-rents generated by a firm. The incomplete contracts approach has been very successful in explaining the corporate governance of entrepreneurial firms and also some important features of large corporations, such as allocation of ownership to the providers of capital who are dispersed, and the importance of internal organization. Aspects that remain to be investigated include the role of the board of directors, interaction between different mechanisms of corporate governance, and the normative implications of the approach.

Keywords

asymmetric information; bargaining; bilateral monopoly; Coase theorem; contractual governance; control rights; corporate governance; firm, theory of; free-rider problem; human capital; incomplete contracts; mechanism design; nexus of contracts view of the firm; property rights view of the firm; quasi-rent; risk aversion; takeovers

Article

While some of the questions have been around since Berle and Means (1932), the term ‘corporate governance’ did not exist in the English language until the mid-1980s. Since then, however, corporate governance issues have become important not only in the academic literature but also in public polity debates. During this period, corporate governance has been identified with takeovers, financial restructuring and institutional investors' activism. But what exactly is corporate governance? Why is there a corporate governance ‘problem’? Why does Adam Smith's invisible hand not automatically provide a solution? What role do takeovers, financial restructuring and institutional investors play in a corporate governance system?
In this article I will try to provide a systematic answer to these questions, making explicit the essential

link between corporate governance and the theory of the firm. My goal is to provide a common framework that helps to analyse the results obtained in these two fields and identify the questions left unanswered. This is not a survey, so I make no attempt to be comprehensive. For an excellent survey on the topic the reader is referred to Shleifer and Vishny (1997).

1 When do we need a governance system?

The word ‘governance’ is synonymous with the exercise of authority, direction, and control. These words, however, seem strange when used in the context of a free-market economy. Why do we need any form of authority? Isn't the market responsible for allocating all resources efficiently without the intervention of authority? The basic (neoclassical) undergraduate microeconomics courses rarely mention the words ‘authority’ and ‘control’.

In fact, neoclassical microeconomics describes well only one set of transactions, which Williamson (1985) calls ‘standardized’. Consider, for instance, the purchase of a commodity, like wheat. There are many producers of the same quality of wheat and many potential customers. In this context, Adam Smith's invisible hand ensures that the good is provided efficiently without the need of any form of authority.

Many daily transactions, however, do not fit this simple example. Consider, for instance, the purchase of a customized machine. The buyer must contact a manufacturer and agree upon the specifications and the final price. Unlike the case of wheat, the signing of the agreement does not represent the end of the relationship between the buyer and the seller. Producing the machine requires some time. During this time many events can occur, which alter the cost of producing the machine as well as the buyer's willingness to pay for it. More importantly, before the agreement was signed, the market for manufacturers was competitive. Once production has begun, though, the buyer and the seller are trapped in a situation of bilateral monopoly. The customized machine probably has a higher value to the buyer than to the market. On the other hand, the contracted manufacturer has probably the lowest cost, to finish the machine. The difference between what the two parties generate together and what they can obtain in the marketplace represents a quasi-rent, which needs to be divided *ex post*. In dividing this surplus Adam Smith's invisible hand is of no help, while authority does play a role.

In the spirit of Williamson (1985), I define a *governance system* as the complex set of constraints that shape the *ex post* bargaining over the quasi-rents generated in the course of a relationship. A main role in this system is certainly played by the initial contract. But the contract, most likely, will be incomplete, in the sense that it will not fully specify the division of surplus in every possible contingency (this might be too costly to do or outright impossible because the contingency was unanticipated). This creates an interesting distinction between decisions made *ex ante* (when the two parties entered a relationship and irreversible investments were sunk) and *ex post* (when the quasi-rents are divided). This contract incompleteness also creates room for bargaining.

The outcome of the bargaining will be affected by several factors besides the initial contract. First, which party has ownership of the machine while it is being produced. Second, the availability of alternatives: how costly is it for the buyer to delay receiving the new machine; how costly is it for the manufacturer to delay the receipt of the final payment; how much more costly is it to have the job finished by another manufacturer, and so on. Finally, a major role in shaping the bargaining outcome is

played by the institutional environment. For example: how effective and rapid is law enforcement; what are the professional norms; how quickly and reliably does information about the manufacturer's performance travel across potential clients, and so on. All these conditions constitute a governance system.

As illustrated by the machine example, there are two necessary conditions for a governance system to be needed. First, the relationship must generate some quasi-rents. In the absence of quasi-rents, the competitive nature of the market will eliminate any scope for bargaining. Second, the quasi-rents are not perfectly allocated *ex ante*. If they were, then there would be no scope for bargaining either.

2 Corporate governance

The above definition of governance is quite general. One can talk about the governance of a transaction, of a club, and, in general, of any economic organization. In a narrow sense, corporate governance is simply the governance of a particular organizational form – a corporation.

Yet the bargaining over the *ex post* rents, which I defined as the essence of governance, is influenced, but not uniquely affected, by the legal structure used. A corporation, in principle, is just an empty legal shell. What makes a corporation valuable is the claims the legal shell has on an underlying economic entity, which I shall refer to as the firm. While often the legal and the economic entity coincide, this is not always the case. For this reason, I define *corporate governance* as the complex set of constraints that shape the *ex post* bargaining over the quasi-rents generated by a firm.

To be sure, many problems that fall within the realm of corporate governance can be (and have been) profitably analysed without necessarily appealing to such a broad definition. Nevertheless, all the governance mechanisms discussed in the literature can be reinterpreted in light of this definition. Allocation of ownership, capital structure, managerial incentive schemes, takeovers, boards of directors, pressure from institutional investors, product market competition, labour market competition, organizational structure, and so on can all be thought of as institutions that affect the process through which quasi-rents are distributed. The contribution of this definition in simply to highlight the link between the way quasi-rents are distributed and the way they are generated. Only by focusing on this link can one answer fundamental questions such as who should control the firm.

Of course, this definition of corporate governance raises the age-old question of what a firm is. But this question should be central to corporate governance. Before we can discuss how a firm should be governed, we need to define the firm. This question is also important because it helps us identify to what extent, if any, corporate governance is different from the governance of a simple contractual relationship (such as in the machine example).

There are two main definitions of the firm available in the literature. The first, introduced by Alchian and Demsetz (1972), is that the firm is a nexus of contracts. According to this definition, there is nothing unique in corporate governance, which is simply a more complex version of standard contractual governance.

The second definition, due to Grossman and Hart (1986) and Hart and Moore (1990) (henceforth GMH), is that the firm is a collection of physical assets that are jointly owned. Ownership matters because it confers the right to make decisions in all the contingencies unspecified by the initial contract. On the one hand, this definition has the merit of differentiating between a simple contractual relationship and a firm.

Since the firm is defined by the non-contractual element (that is, the allocation of ownership), corporate governance (as opposed to contractual governance) is defined by the effect of this non-contractual element. Not surprisingly, the focus of the corporate governance literature since the mid-1990s has been the allocation of ownership (hence this literature is called the property rights view of the firm). On the other hand, this definition has the drawback of excluding any stakeholder other than the owner of physical assets from being important to our understanding of the firm.

More recently, Rajan and Zingales (2001; 1998) have proposed a broader definition. They define the firm as a nexus of specific investments: a combination of mutually specialized assets and people. Unlike the nexus of contracts approach, this definition explicity recognizes that a firm is a complex structure that cannot be instantaneously replicated. Unlike the property rights view, this definition recognizes that all the parties who are mutually specialized belong to the firm, be they workers, suppliers or customers. While this definition does not necessarily coincide with the legal definition, it does coincide with the economic essence of a firm: a network of specific investments that cannot be replicated by the market.

3 Incomplete contracts and governance

In an Arrow–Debreu economy it is assumed that agents can costlessly write all state-contingent contracts. As a result, all decisions are made *ex ante* and all quasi-rents are allocated *ex ante*. Thus, there is no room for governance. More surprisingly, even if we relax the assumption that every state-contingent contract can be written and admit that certain future contingencies are not observable (and thus not contractible), we still find no room for governance as long as one can costlessly write contracts on all future observable variables.

Recall the example of the customized machine, and assume that the manufacturer's effort is unobservable to others and is, therefore, not contractible. The neoclassical approach to this problem is to design a mechanism (hence the term ‘mechanism design’), contingent on all publicly observable variables, which provides the manufacturer with the best possible incentives to exert effort. Myerson (1979) shows that all optimal mechanisms are equivalent to a revelation (direct) mechanism in which the agent (manufacturer) publicly announces his information and receives compensation contingent on his announcement. An important consequence of this result is that, in the mechanism design approach, delegation (giving an agent discretion over certain decisions) is always weakly dominated by a fully centralized mechanism, where all decisions are made *ex ante* by the designer. The mechanism design approach reproduces several distinguishing features of an Arrow–Debreu economy: all decisions are made *ex ante* and executed only *ex post*; as a result, all conflicts are resolved and all rents are allocated *ex ante*. This leaves no room for *ex post* bargaining. All these features are incompatible not only with my definition of governance, but also with any meaningful (that is, related to the sense in which this term has been used) definition of governance. This is best illustrated with two examples.

One of the crucial questions in corporate governance concerns the party in whose interest corporate directors should act. In the mechanism design approach this question cannot even be raised. All possible future conflicts are resolved *ex ante* and the initial contract specifics how directors will behave in any observable state of the world. However, since this question is raised all the time, it must be that all possible conflicts are not resolved *ex ante*.

Second, the mechanism design approach avoids renegotiation: the initial contract is so designed that the

agents do not want to renegotiate. As a result, the designer wants to make renegotiation as inefficient as possible: this reduces the costs of providing incentives to the agents with no efficiency costs, since renegotiation never occurs in equilibrium (Aghion, Bolton and Felli, 1997). If this result were to be taken seriously, the optimal public policy approach would be to preserve any inefficiency in the system in order to avoid destroying its beneficial incentives *ex ante*. In reality, though, the jurisprudential approach is completely different. For example, courts do not support punitive damages that are considered excessive with respect to the issue at stake.

Only in a world where some contracts contingent on future observable variables are costly (or impossible) to write *ex ante* is there room for governance *ex post*. Only in such a world are there quasi-rents that must be divided *ex post* and real decisions that must be made. Finally, only in a world of incomplete contracts can we define what a firm is and discuss corporate governance as being different from contractual governance. Not surprisingly, the theory of governance is intimately related to the emergence and evolution of the incomplete contracts paradigm.

A fundamental milestone in this evolution is the residual rights of control concept introduced by Grossman and Hart (1986). In a world of incomplete contracts, it is necessary to allocate the right to make *ex post* decisions in unspecified contingencies. This residual right is both meaningful and valuable. It is meaningful because it confers the discretion to make decisions *ex post*. It is valuable because this discretion can be used strategically in bargaining over the surplus.

4 Why does corporate governance matter?

By definition, corporate governance matters for distribution of rents, but to what extent does it matter for economic efficiency? There are three main channels through which the conditions that affect the division of quasi-rents also affect the total surplus produced. In presenting these channels I make a sharp distinction between *ex ante* (when specific investments need to be sunk) and *ex post* (when quasi-rents are divided) effects, as though the firm lasted just one period. Of course, this is not true in reality because *ex post* considerations of one period are mixed with *ex ante* considerations for the next period.

4.1 Ex ante incentive effects

The process through which surplus is divided *ex post* affects the *ex ante* incentives to undertake some actions, which can create or destroy some value, in two main ways.

First, rational agents will not spend the optimal amount of resources in value-enhancing activities that are not properly rewarded by the governance system. In fact, one goal in designing a governance system is to motivate those investments that are not properly rewarded in the marketplace. The canonical example of how a change in the governance structure can change the incentives to make a value-enhancing relationship-specific investment is the Fisher Body case. In the early 1920s, Fisher Body (an auto body manufacturer) refused to locate its plants close to General Motors' plants in spite of the obvious efficiency improvement generated by such a move. Locating close to GM would have reduced Fisher Body's ability to supply other car manufacturers, which would have weakened its bargaining position *ex post* and possibly reduced its share of the quasi-rents generated by the relationship with GM (see Klein, Crawford and Alchian, 1978). A change in the governance system (the acquisition of Fisher

Body by GM) eventually led to the efficient plant location decision. Another famous illustration of the same phenomenon is managerial shirking. A manager will shirk if her *ex post* bargaining payoff does not increase sufficiently with her effort and, therefore, fails to compensate her for the cost of this effort.
Second, rational agents will spend resources in inefficient activities whose only (or main) purpose is to alter the outcome of the *ex post* bargaining in their favour. For example, a manager may specialize the firm in activities she is best at running because this increases her marginal contribution *ex post* and, thus, her share of the *ex post* rents (Shleifer and Vishny, 1989). Interestingly, this problem is not limited to the top of the hierarchy, but is present throughout. Subordinates, who do not have much decision power, will waste resources trying to capture the benevolence of their powerful superiors (Milgrom, 1988). Even the well-known tendency of managers to overinvest in growth can be interpreted as a manifestation of this problem. Managers like to expand the size of their business because this makes them more important to the value of the firm and, thus, increases the payoff they can extract in the *ex pest* bargaining.
Of course, a governance system might promote or discourage these activities. F or example, Chandler (1966) reports that, under the Durand reign, GM's capital allocation was highly politicized (‘a sort of horse trading’). The move to a multi-divisional structure, with the resulting increase in divisional managers' autonomy, reduced the managers' payoff from rent-seeking. Similarly, Milgrom and Roberts (1990) explain many organizational rules as a way to minimize influence costs. Finally, Rajan, Servaes and Zingales (2000) argue that inefficient ‘power-seeking’ is more severe the more investment opportunities a firm's divisions have. Consistent with this claim, they find that the value of a diversified firm is negatively related to the diversity of the investment opportunities of its divisions.
Thus, a governance system affects the incentives to invest or seek power, altering the marginal payoffs that these actions have in *ex post* bargaining. For instance, for an independent Fisher Body, the marginal effect on the bargaining payoff of localizing its plants close to GM is negative (it reduces the value of its outside options), but is positive for Fisher Body as a unit of GM, which does not have the authority to supply other manufacturers without GM's consent (see Rajan and Zingales, 1998). Thus, a different ownership structure alters the incentives to make specific investments.

4.2 Inefficient bargaining

A second channel through which a governance system affects total value is by altering *ex post* bargaining efficiency. This is tantamount to saying that the governance system affects the degree to which the assumptions of the Coase theorem are violated. A governance system, therefore, can affect the degree of information asymmetry between the parties, the level of coordination costs, or the extent to which a party is liquidity constrained.
For example, if control rights are assigned to a large and dispersed set of claimants (like the shareholders of most publicly traded companies), free-rider problems may prevent an efficient action from being undertaken even if property rights are well defined and perfectly tradeable (Grossman and Hart, 1980). Alternatively, the allocation of control rights can affect efficiency by determining the direction in which a compensating transfer must be made. The direction of the transfer matters when one of the parties to the *ex post* bargaining is liquidity constrained (Aghion and Bolton, 1992) or when it faces a different opportunity to invest these resources productively rather than in power-seeking activities (Rajan and

Zingales, 1996). In both cases an efficient transaction may not be agreed upon – in the first case because the party that should compensate does not have the resources, in the second case because the transaction (while efficient *per se*) may generate such an increase in wasteful power-seeking as to more than offset its benefits.

To this standard list of imperfections, Hansmann (1996) adds the divergence of interests among the parties who have control rights. Citing the political economy literature, Hansmann argues that *ex post* inefficiency is increasing in the divergence of interests among control holders. For example, he argues that allocating control to workers is more costly when they differ in their professional skills, hierarchical position and tenure. While Hansmann does not provide a formal model of why this relation occurs, he does provide very compelling evidence that in practice control rights are rarely allocated to parties with conflicting Interests. His conjecture is intriguing because there is no well-established general theory of how different governance systems lead to different levels of *ex post* inefficiency. There is little doubt, however, that these inefficiencies exist and are important. For example, Wiggins and Libecap (1985) document that an excessively dispersed initial allocation of drilling rights leads to an inefficient method of extracting oil, with estimated losses as big as 50 per cent of the total value of the reservoir.

4.3 Risk aversion

Finally, a governance system might affect the *ex ante* value of the total surplus by determining the level and the distribution of risk. If the different parties have different degrees of risk aversion (or different opportunities to diversify or hedge risk), then the efficiency of a governance system is also measured by how effectively it allocates risk to the most risk-tolerant party. This idea is the cornerstone of Fama and Jensen's (1983a; 1983b) analysis of organizational structure and corporate governance.

Different governance systems can also *generate* a different amount of risk. Suppose, for instance, that the total amount of surplus generated is constant. It is still possible that the payoff of each party is stochastic, if the governance structure generates a stochastic bargaining outcome. For example, a life insurance contract written in nominal dollars creates a pure gamble between the policy holders and the insurance company with respect to the future rate of inflation. This additional ‘governance’ risk (in this case created by the contract, in general created by the governance structure) reduces the value of the total surplus, if the parties are risk averse and cannot diversify away the risk.

In summary, the objectives of a corporate governance system should be: (*a*) to maximize the incentives for value-enhancing investments, while minimizing inefficient power seeking; (*b*) to minimize inefficiency in *ex post* bargaining; (*c*) to minimize any ‘governance’ risk and allocate the residual risk to the least risk averse parties.

5 Who should control the firm?

To show the utility of the framework developed thus far, I will use it to address one of the most controversial issues in corporate governance: who should control the firm? In particular, I will analyse it with regard to the first of the three above objectives of a corporate governance system. For an analysis focused on the second objective the reader is referred to Hansmann (1996), and for an analysis focused on the third to Fama and Jensen (1983a; 1983b).

As far as the first objective is concerned, the allocation of control is important because it affects the division of surplus. By controlling a firm's decisions, a party can ensure for itself of more and more valuable options without the collaboration of the other parties. This guarantees the controlling party a larger share of the surplus within the relationship. Thus, in the framework outlined above, the question of who should control the firm can be rephrased as: whose investments need more protection in the *ex post* bargaining? Again, the answer to this question is indissolubly linked to the underlying theory of the firm.

In the nexus of contracts view, the firm ‘is just a legal fiction which serves as a focus for the complex process in which the conflicting objectives of individuals … are brought in equilibrium within a framework of contractual relationship’ (Jensen and Meckling, 1976, p. 312). Thus, according to this view each party is fully protected by its contract with the exception of the shareholders, who accept a residual payoff because they possess a comparative advantage in diversifying risk. As a result, shareholders need the protection insured by control.

While widely popular, this explanation is unsatisfactory. The contractual protection provided to the parties involved in the nexus of contracts is complete only if contracts are complete. But if contracts are complete, then the statement that shareholders are in control is meaningless. In fact, in a world of complete contracts all the decisions are made *ex ante*, and thus shareholders are no more in control than are the workers: everything is contained in the initial grand contract. Furthermore, as I have already argued in Section 3, this conclusion is inconsistent with the existence of a debate on what a company should do.

Alternatively, if contracts are incomplete, then the argument that all other parties are fully protected by their contractual relationships does not automatically follow. In fact, in this context one should ask why shareholders need more protection than other parties to the nexus of contracts. I return to this issue below.

In the property rights view of the firm, the reason why shareholders should be in control is straightforward. Control is allocated so as to maximize the incentives to make human capital-specific investments. The owner of the firm will generally be the worker with the most expropriable investment. In other words, the property rights approach does not deal with outside shareholders and, thus, it applies only to entrepreneurial firms.

Outside of the GHM framework, the typical justification for why shareholders (or more generally the providers of finance) are in control is based on a combination of three arguments. Shareholders need more protection because: (*a*) their investment is more valuable; (*b*) other stakeholders can protect their investments better through contracts; (*c*) other stakeholders have other sources of power *ex post* that protect their investments. Of the three arguments, the first is clearly unfounded. Reviewing the empirical evidence on the return on specialized human capital, Blair (1995) estimates that the quasi-rents generated by specialized human capital are as big as accounting profits, which are likely to overestimate the quasi-rent generated by physical capital. Hence, there is no ground for dismissing human capital investments as second order to physical capital investments.

The second argument is harder to dismiss. Since we lack a fully satisfactory theory of why contracts are incomplete, we cannot easily argue which contracts are more incomplete. Nevertheless, it is hard to argue that human capital investments are easier to contract than physical capital investments. If there is one contingency that is easily verifiable, it is the provision of funds. Thus, it is not obvious why providers of funds are at a comparative disadvantage.

The most convincing argument is probably the third. As Williamson (1985) puts it,

> the suppliers of finance bear a unique relation to the firm: The whole of their investment in the firm is potentially placed at hazard. By contrast, the productive assets (plant and equipment; human capital) of suppliers of raw material, labor, intermediate product, electric power, and the like normally remains in the suppliers' possession.

Thus, the other stakeholders have a better outside option in the *ex post* bargaining, and they do not need the protection ensured via the residual rights of control.

Even this argument, however, is not fully satisfactory. In fact, it suggests only that the suppliers of finance should have some form of contractual protection – it does not necessarily imply that they should be protected via the residual rights of control.

A satisfactory explanation of why the residual right belongs to the shareholders can be obtained only in a theory of the firm that explicitly accounts for the existence of different stakeholders, and models the interaction between contractual (for example, ownership) and non-contractual (for example, unique human capital investments) sources of power. An attempt in this direction is made by Rajan and Zingales (1998).

To understand the argument, note that the residual right of control over an asset always increases the share of surplus captured by its owner (who has the opportunity to walk away with the asset), but it does not necessarily increase her marginal incentive to specialize. If, as is likely, a more specialized asset has less value outside the relationship for which it has been specialized, then specialization decreases the owner's outside opportunity and,; thus, her share of the quasi-rents. Owning a physical asset, then, makes an agent more reluctant to specialize it. As a result, the residual right of control is best allocated to a group of agents who need to protect their investment against *ex post* expropriation, but who have little control over how much the asset is specialized.

Consider now the different members of the specific investments nexus that makes up the firm. Most of the specific investments which form this nexus are in human capital and, therefore, can neither be contracted nor delegated *ex ante*. Granting the residual right to any of these members will have a negative effect on their incentive to specialize. By contrast, since the provision of funds is easily contractible, funds will be provided in the optimal amount as long as their providers receive sufficient surplus *ex post*. Thus, allocating the residual rights of control to the providers of funds has the positive effect of granting them enough surplus *ex post*, while avoiding the negative effect of reducing their marginal incentives.

Once they have provided funds, however, financial investors might be reluctant to use these funds for very specialized projects, for fear of seeing their share of the return fall. Thus, it is optimal that, while retaining a residual right of control over the assets, the providers of funds delegate the right to specialize the assets to a third party, who does not internalize the opportunity loss generated by this specialization. This third party, thus, should not be in the position of a mere agent, who owes a duty of obedience to the principal, but should be granted the independence to act in the interest of the firm (that is, the whole body of members of the nexus of specific investments), and not only of the shareholders. Blair and Stout (1997) claim that this is the role American corporate law attributes to the board of directors.

In sum, a broader definition of the firm allows us to understand why the residual right of control is

allocated to the providers of capital and why its use is mostly delegated to a board of directors.

6 Normative analysis

An interesting, and largely unexplored, application of the incomplete contract approach to corporate governance is the analysis of its normative implications. In a world of complete contracts, such analysis has limited scope. A benevolent social planner would be unable to improve the *ex ante* allocation reached by private contracting, because this will achieve the constrained-efficient outcome. *Ex post*, the outcome might be inefficient, but that inefficiency is always part of the written contract and needs to be preserved to maintain *ex ante* future efficiency. By contrast, in a world of incomplete contracts, there is ample scope to analyse both *ex ante* and *ex post* efficiency.

First, a privately optimal governance system may not be socially efficient. In fact, a world of incomplete contracts generates some incentives to 'arbitrage power' through time. Consider an entrepreneur, who has immense bargaining power today, but anticipates losing it in the near future. If she could write all the contracts she could succeed in extracting all the present and future surplus arising from a relationship without any distortion. But, if some contracts cannot be written, then the entrepreneur has an incentive to distort her choices so as to transfer some of her bargaining power today into the future, enabling her to capture some of the future surplus as well. This is the idea underlying the choice of ownership in Zingales (1995a) and Bebchuk and Zingales (1996), and of the hierarchical structure in Rajan and Zingales (2001). It can also be used to provide a rationale for the existing mandatory rules (see Bebchuk and Zingales, 1996).

Second, in a world of incomplete contracts one can discuss the welfare effects of different institutions. For example, in a world of complete contracts the type of legal system a country adopts is irrelevant, as long as private contracts are enforced. By contrast, it is at least conceivable that in an incomplete contract world it may have a significant effect. This is intriguing because empirically it has been shown that legal institutions have an effect on the appropriability of quasi-rents by outside investors (Zingales, 1995b), on the way corporate governance is structured (La Porta et al., 1996), and on the amount of external finance raised (La Porta et al., 1997).

Finally, the incomplete contract approach generates a potential role for government intervention *ex post*. Unlike in the mechanism design approach, in an incomplete contract world *ex post* inefficiency is not necessarily desirable *ex ante*. Thus, a selective intervention that eliminates *ex post* inefficiency, while preserving the distributional consequences sought *ex ante*, will improve welfare.

7 Limitations of the incomplete contract approach

While the incomplete contracts approach to corporate governance has brought tremendous insights to the corporate governance debate, it has two weaknesses.

First, its predictions for the optimal allocation of ownership are extremely sensitive to what contracts can be written. Consider, for instance, the plant localization problem discussed above. If no contracts can be written, then – according; to the property rights approach – Fisher Body (who makes the bigger specific investment) should own the asset. However, if General Motors could credibly commit through a contract to buy all its car bodies from Fisher Body (as it did), then giving ownership to Fisher Body will

confer too much power on it, and, thus, it is optimal for General Motors to own the asset (Hart, 1989), Thus, who should have the residual right of control depends crucially upon what the contractable rights are. But this is very difficult to argue on a priori grounds without a general theory of why contracts are incomplete (see Maskin and Tirole, 1997).

Second, this approach relies heavily (as does the complete contract approach) on the agents anticipating all future possible contingencies (Hellwig, 1997). This requirement can be reasonable when the subject of analysis is a small entrepreneurial firm, but it loses credibility when it is applied to large publicly held companies formed decades ago. Can we really interpret the capital structure of IBM today as the outcome of the design by Charles Flint (its founder) in 1911 attempting to allocate control optimally? Hart (1995) argues that the ‘founding father’ interpretation is simply a metaphor for the capital structure that a manager will choose under the pressure of the corporate control market. Yet Novaes and Zingales (1995) show that the two approaches lead to different predictions, not only about the level of debt but also about its sensitivity to the cost of financial distress and times. Thus, in the current state of knowledge, the *ex ante* approach to the capital structure of non-entrepreneurial companies lacks theoretical foundations.

8 Summary and conclusions

In this article I have tried to summarize the results obtained by applying the incomplete contracts approach to corporate governance. In a world where all future observable contingencies can be costlessly contracted upon *ex ante*, there is no room for governance. By contrast, in an incomplete contracts world, corporate governance can be defined as the set of conditions that shapes the *ex post* bargaining over the quasi-rents generated by a firm. A governance system has efficiency effects both *ex ante*, through its impact on the incentive to make relationship-specific investments, and *ex post*, by altering the conditions under which bargaining takes place. A governance system also affects the level and the distribution of risk.

The incomplete contracts approach has been very successful in explaining the corporate governance of entrepreneurial firms. It can explain how ownership is allocated and how capital structure is chosen. By contrast, it is difficult for this approach to cope with the complexity of large publicly traded complies. Nevertheless, recent contributions in the area are able to account for some important features of large corporations: allocation of ownership to the providers of capital who are dispersed, and the importance of internal organization.

Many aspects, however, remain to be investigated. First and foremost is the role of the board of directors. The second is the interaction between the different mechanisms of corporate governance. While we have many models that describe how each mechanism works in isolation, we know very little about how they interact. The effects are not obvious. For example, debt and takeovers are generally thought, in isolation, to be two instruments that reduce the amount of quasi-rents appropriated by management. But the use of debt may crowd out the effectiveness of takeovers, increasing rather than decreasing managerial rents (Novaes and Zingales, 1995). Third, the normative implications of this approach deserve more attention. In a world of incomplete contracts, privately optimal governance can be inefficient *ex ante* and *ex post*. Of course, this is only a theoretical possibility, whose relevance needs to be assessed in the data. The most important contribution, however, will arise from a development of

the underlying theory. Without a better understanding of why contracts are incomplete, all the results are merely provisional.

See Also

- hold-up problem
- incomplete contracts

Bibliography

Aghion, P. and Bolton, P. 1992. An incomplete contract approach to financial contracting. *Review of Economic Studies* 59, 473–94.

Aghion, P., Bolton, P. and Felli, L. 1997. Some issues on contract incompleteness. Working paper, London School of Economics.

Alchian, A. and Demsetz, H. 1972. Production, information costs and economic organization. *American Economic Review* 62, 777–95.

Bebchuk, L. and Zingales, L. 1996. Corporate ownership structures: private versus social optimality. Working Paper No. 5584. Cambridge, MA: NBER.

Berle, A. and Means, G. 1932. *The Modern Corporation and Private Property*. New York: Macmillan.

Blair, M.M. 1995. *Ownership and Control*. Washington, DC: Brookings Institution.

Blair, M.M. and Stout, L. 1997. A theory of corporation law as a response to contracting problems in team production. Working Paper, Brookings Institution.

Chandler, A. 1966. *Strategy and Structure*. Garden City, NY: Anchor Books.

Fama, F. and Jensen, M.C. 1983a. Separation of ownership and control. *Journal of Law and Economics* 26, 301–25.

Fama, E. and Jensen, M.C. 1983b. Agency problems and residual claims. *Journal of Law and Economics* 26, 327–49.

Grossman, S. and Hart, O. 1980. Takeover bids, the free rider problem and the theory of the corporation. *Bell Journal of Economics* 11, 42–69.

Grossman, S. and Hart, O. 1986. The costs and the benefits of ownership: a theory of vertical and lateral

integration. *Journal of Political Economy* 94, 691–719.

Hansmann, H. 1996. *The Ownership of Enterprise*. Cambridge, MA: Belknap Press of Harvard University Press.

Hart, O. 1989. An economist's perspective on the theory of the firm. *Columbia Law Review* 89, 1757–74.

Hart, O. 1995. *Firms, Contracts, and Financial Structure*. Oxford: Oxford University Press.

Hart, O. and Moore, J. 1990. Property rights and the nature of the firm. *Journal of Political Economy* 98, 1119–58.

Hellwig, M. 1997. Unternchmensfinanzierung, Unterichmenskontrolle und Ressourcenallokation: Was leister das Finanzsystem. Arbeitspapier Nr. 97/02, University of Mannheim.

Jensen, M.C. and Meckling, W. 1976. Theory of the firm; managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

Klein, H., Crawford, R. and Alchian, A. 1978. Vertical integration, appropriable rents and the competitive contracting process. *Journal of Law and Economics* 21, 297–326.

La Porta, R., Lopez de Silancs, F., Shleifer, A. and Vishny, R. 1996. Law and finance. Working Paper No. 5661. Cambridge, MA: NBER.

La Porta, R., Lopez de Silancs, F., Shleifer, A. and Vishny, R. 1997. Legal determinants of external finance. *Journal of Finance* 52, 1131–50.

Maskin, F. and Tirole, J. 1997. Unforeseen contingencies, property rights and incomplete contracts. Working Paper No. 1796, Institute of Economic Research, Harvard University.

Milgrom, P. 1988. Employment contracts, influence activities, and efficient organization design. *Journal of Political Economy* 96, 42–60.

Milgrom, P. and Roberts, J. 1990. Bargaining costs, influence costs and the organization of economics activity. In *Perspectives on Positive Political Economy*, ed. J. Alt and K. Shepsle. Cambridge: Cambridge University Press.

Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47, 61–73.

Novaes, W. and Zingales, L. 1995. Capital structure choice when managers are in control: entrenchment versus efficiency. Working Paper No. 5384. Cambridge, MA: NBER.

Rajan, R., Servaes, H. and Zingales, L. 2000. The cost of diversity: diversification discount and inefficient investment. *Journal of Finance* 55, 35–80.

Rajan, R. and Zingales, L. 1996. The tyranny of the inefficient: an enquiry into the adverse consequences of power struggles. Working Paper No. 5396. Cambridge, MA: NBER.

Rajan, R. and Zingales, L. 1998. Power in a theory of the firm. *Quarterly Journal of Economics* 113, 387–432.

Rajan, R. and Zingales, L. 2001. The firm as a dedicated hierarchy. *Quarterly Journal of Economics* 116, 805–51.

Shleifer, A. and Vishny, R. 1989. Management entrenchment: the case of manager-specific investments. *Journal of Financial Economics* 25, 123–40.

Shleifer, A. and Vishny, R. 1997. A survey of corporate governance. *Journal of Finance* 52, 737–83.

Wiggins, S.N. and Libecap, G.D. 1985. Oil field unitization: contractual failure in the presence of imperfect information. *American Economic Review* 75, 368–85.

Williamson, O. 1985. *The Economic Institutions of Capitalism*. New York: The Free Press.

Zingales, L. 1995a. Insider ownership and the decision to go public. *Review of Economic Studies* 62, 425–48.

Zingales, L. 1995b. What determines the value of corporate votes? *Quarterly Journal of Economics* 110, 1047–73.

How to cite this article

Zingales, Luigi. "corporate governance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000370> doi:10.1057/9780230226203.0323

The New Palgrave Dictionary of Economics Online

corporate law, economic analysis of

Robert Daines and Michael Klausner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic analysis of corporate law largely focuses on (*a*) the efficiency of legal rules and the proper role of the law, (*b*) the ways in which legal rules affect shareholders' ability to monitor managers, and (*c*) the effect of limited liability on the relationship between the corporation and third parties. This article reviews the literature in each of these areas.

Keywords

agency costs; collective action problem; contractarian conception of the corporation; corporate charters; corporate control; corporate governance; corporate law; corporations; federalism; hedge funds; herding; insider trading; law and economics; learning externalities; limited liability; monitoring; network externalities; ownership and control; poison pill; race to the top/bottom; shareholder activism; shareholder suits; state competition; takeover defence; takeovers

Article

The economic analysis of corporate law focuses primarily on publicly held corporations. Following Coase (1937), the corporation is conceptualized as a nexus of contracts. Because corporate law focuses primarily on the authority of management and its obligations to shareholders, the primary 'contract' of interest is that between management and shareholders. The content of the manager–shareholder contract is conceptualized in terms of the agency-cost model of Jensen and Meckling (1976), with management viewed collectively as agent, and shareholders viewed collectively as principal. Ideally, the terms of the manager–shareholder contract minimize agency costs and thereby maximize the value of the firm.
Most of the economics-oriented corporate law literature can be divided into three areas, all of which focus on the United States. First, there are papers that analyse the economic forces by which corporate law is created by states and adopted by firms, and the proper role of corporate law in light of those forces. Second, there are papers that analyse particular monitoring mechanisms that law creates –

shareholder voting, shareholder lawsuits, takeovers. A third set of papers analyses the basic features of a corporation, focusing on limited liability.

This review will discuss these three sets of papers. We do not address the substantial literature on law and finance that suggests that a country's corporate law rules may affect its financial markets and economic growth (see La Porta et al., 1997; 1998; and Rajan and Zingales, 1998). Nor do we address corporate governance strategies, such as CEO pay, that are largely independent of corporate law.

The role of corporate law

Economics-oriented scholarship on the role of corporate law can be roughly divided into two generations. The first generation, which spanned the period from the late 1970s to the mid-1990s, tended to reach the conclusion that market forces would yield socially optimal corporate governance outcomes. The second generation spans the period from the mid-1990s to the present. This generation, which includes more empirical work than the first, has painted a less perfect picture of the relationship between market forces and socially optimal corporate governance (see Klausner, 2006).

First-generation scholarship

The central insight of the first generation of economics-oriented corporate law scholarship was the conceptualization of the public corporation as a contractual arrangement between managers and shareholders. This insight has its origin in Coase (1937). It was developed within the agency cost framework in Jensen and Meckling (1976), and extended to the analysis of corporate law by Easterbrook and Fischel (1989; 1991). Although managers and shareholders do not negotiate governance arrangements, the price mechanism for a company's shares in an initial public offering (IPO) is expected to serve the same function, just as it does in other markets where buyers and sellers do not explicitly negotiate contracts. Consequently, the legally enforceable elements of the corporation's governance structure are viewed as the product of a market-mediated contracting process. Scholars writing in this framework therefore argue that firms’ governance structures tend to minimize the agency costs associated with the separation of ownership and control, and thereby maximize the value of the corporation.

Legally enforceable governance commitments can take either of two forms. First, firms select the corporate law rules that govern the rights of shareholders and the obligations of management. Each of the 50 US states has enacted corporate law rules. Firms are free to elect to be governed by any of these rules, regardless of where they do business. To be governed by any state's legal rules, a firm need only incorporate in that state at the time of its IPO. Subsequent disputes between managers and shareholders will then by the decided according to the corporate law of that state. A firm cannot change its state of incorporation unless its board of directors and shareholders holding at least a majority of its shares agree. Second, pre-IPO manager/shareholders must draft a charter that will govern the corporation once it goes public. A charter begins as a blank slate and can include any governance arrangements that a firm's pre-IPO shareholders choose to adopt. To a substantial degree, the law allows a firm's charter to override provisions of corporate law. Thus, corporate law rules are often simply default rules that can be superseded by a corporation's charter terms.

Thus, one insight of this first generation was that corporate law was a product that states produce and firms consume. Winter (1977) was the first to argue that states are engaged in a ‘race to the top’ to produce corporate law that would tend to minimize agency costs. In order to obtain revenues from franchise fees and to create business for their local lawyers, states were expected to offer corporate law (that is, default rules) that would maximize the value of many firms and thereby save firms the trouble of customizing their own charter terms. Romano (1985) provided empirical evidence consistent with the proposition that a race to the top was occurring. She found, however, that Delaware had already achieved a substantial lead and questioned whether the race would actually make it to the top. The argument that market forces would produce legally enforceable governance commitments that would minimize agency costs stood in contrast to an earlier claim by Cary (1974) that states were engaged in a ‘race to the bottom’ to create legal rules that appeal to management at the expense of shareholders.

Second-generation scholarship

The second generation of scholarship has cast both empirical and theoretical doubt on the contractarian claims described above.

Empirical findings

A central claim of the contractarian conception of the corporation and corporate law is that corporations are heterogeneous in their corporate governance needs – hence the value of atomistic contracting. Empirical studies have now shown, however, that there is a high degree of uniformity in firms’ governance commitments at the time they go public.

Daines (2002) found that, between 1978 and 2002, 50 per cent of firms incorporated in Delaware, and that during the second half of this period over 70 per cent of firms incorporated in Delaware. More importantly, however, Daines found that, among firms that did not incorporate in Delaware, nearly all incorporated in the state in which they were headquartered – whatever that state happened to be. Bebchuk and Cohen (2003) and Kahan (2006) confirmed Daines's findings.

These findings regarding incorporation decisions have three implications. First, the decision to incorporate in one's home state (when no out-of-state firms incorporate there) cannot be motivated primarily by the content of a state's laws. Something else must be at work. Daines's findings suggest that this choice may be made by the firm's local lawyer, hoping to keep the firm's business following the IPO, or by management wanting access to the state legislature if it needs a law passed. Romano (1987) found that most state anti-takeover legislation enacted in the 1980s was initiated by in-state management seeking protection from hostile bids. Bebchuk and Cohen (2003) found that states seem to retain more home-state incorporations if they already have state anti-takeover statutes on their books, but Kahan (2006) refuted this finding. Kahan did find, however, that states with very low-quality corporate law retained fewer home-state incorporations than did other states.

Second, these findings imply a high degree of uniformity in the governance commitments reflected in a firm's incorporation decision. Firms that focus on the quality of corporate law choose Delaware law. This uniformity casts some doubt on the contractarian assumption that firms are heterogeneous in their governance needs. Alternatively, the findings may suggest that there is value in uniformity itself, a point

addressed below. Either way, there is evidence that the choice of Delaware as a state of incorporation enhances firm value. Romano (1987) and Daines (2001) found evidence consistent with this conclusion. Subramanian (2004) argues that this is a small-firm effect.

Third, the findings on incorporation choices cast doubt on the proposition that states compete to attract incorporations – whether racing to the top or to the bottom, Delaware seems to be the only state competing. This is what Kahan and Kamar (2002) find in a study of states' efforts, or lack thereof, to attract incorporations and to earn revenues from them.

Empirical research has also revealed a high degree of uniformity in corporate charters. These supposed vehicles of customized contracting and innovation turn out to be fairly empty vessels. The only dimension on which they vary is in that of takeover defences (Klausner, 2006), and variability in that respect sits uneasily with the proposition that IPO charters maximize firm value. Three studies by Daines and Klausner (2001), Field and Karpoff (2002) and Coates (2001) have shown that firms commonly go public with charters providing for staggered boards, which are an effective anti-takeover defence that tends to reduce share value.

Theoretical challenges to the contractarian framework

It is possible that the contractarians overstated their premise that firms are heterogeneous in their governance needs. When it comes to legally enforceable governance commitments, perhaps one size fits all.

There are theoretical reasons, however, to doubt that homogeneous governance needs explain the uniformity described above. The essentially complete absence of customization or innovation in corporate charters suggests there are market imperfections in the contracting process. There has been no lack of innovation in corporate governance since the mid-1980s. None, however, originated in a corporate charter. Innovation at the firm level has taken the form of unilateral adoption of governance structures – for instance, an independent board or separation of CEO and chair – with no legally binding commitment to maintain those structures. The absence of legally binding commitments suggests that the cost of legal enforcement plays some role in the relative emptiness of corporate charters. While there have been innovations in legally enforceable governance mechanisms, they have not occurred at the level of the individual firm charter or even state law, as the contractarian thesis predicts. Instead, they have occurred, for better or worse, through Securities and Exchange Commission (SEC) regulation and federal statute (Sarbanes–Oxley Act, described below).

Klausner (1995; 2006) and Kahan and Klausner (1996) posit that there are learning and network externalities associated with state corporate law and corporate charter terms. As a result of these externalities, commonly used governance mechanisms have value independent of their intrinsic content; they tend to be better understood and less uncertain in their application than customized mechanisms. These externalities may thus explain the attraction of Delaware and the lack of customization or innovation in corporate charters.

In this context learning externalities take the form of judicial precedents interpreting and applying legal rules, and lawyers' familiarity with these precedents. Because many firms have been incorporated in Delaware, there is a large body of Delaware precedent. As a result, there is less uncertainty regarding how a legal rule will be applied. This may make Delaware valuable because firms have adopted it in the

past.
Future judicial interpretations are valuable as well. The larger the number of firms that use the same legal rule or charter term over time, the more the rule or term will be litigated in the future, and the more frequently it will be interpreted. As Hansmann (2006) explains, the alternative would be periodic charter amendments, which could be difficult to accomplish because of the need to have a majority of shareholders and the company's board agree. Consequently, the market dynamic by which firms choose a state of incorporation can be expected to mirror that of product markets in network industries. The equilibrium in those industries can be socially suboptimal uniformity, which may be what is reflected in the attraction of Delaware incorporation and the ‘plain vanilla’ charter – that is, a charter with no customization that adopts essentially all default rules.
Kahan and Klausner (1996) offer two additional explanations of uniformity in charter terms and incorporation choices. One is that lawyers who draft charters on behalf of their corporate clients may be exhibit the same sort of individually rational herd behaviour that Scharfstein and Stein (1990) and Zwiebel (1995) model for agents such as money managers. These models are based on reputational payoffs to winning or losing with or without the herd. The second explanation relies on results in psychological experiments that reveal a ‘status quo’ bias, an ‘anchoring’ bias and a ‘conformity’ bias in other settings.

Law-intensive monitoring mechanisms

Corporate law creates three monitoring mechanisms and influences a fourth. First, corporate law gives shareholders the right to vote for the board of directors and to approve certain major changes, such as a change to the firm's charter or a merger or sale of the firm. Second, corporate law specifies managers’ duties to shareholders and provides a way for shareholders to collectively sue management for its failure to fulfil these obligations. Third, corporate law regulates the takeover process, which allows a poorly run firm to be acquired by a third party. Finally, US federal securities law imposes mandatory disclosure obligations on publicly held firms, which facilitates each of these monitoring mechanisms and enables non-legal monitoring mechanisms (such as the press).

Shareholder voting

Corporate law gives control of the firm to the board of directors. Shareholder influence over managers comes from their right to elect the board and their implicit (or explicit) threat to vote out incumbent directors. Board elections are held annually and shareholders frequently have the ability to call interim elections. Today, voting is also the means by which control over firms changes hands in a takeover (Gilson and Schwartz, 2001).
Shareholders’ ability to oust directors is thus an important check on managerial misbehaviour. The primary limitation on the effectiveness of the shareholder vote is economic rather than legal: shareholders’ collective action problems. Individual shareholders with small stakes may not find it worthwhile to become informed and therefore typically either fail to vote or simply vote with management.
An important question is whether institutional investors, by virtue of their larger stakes, will solve the

collective action problem and monitor managers more effectively. Money managers, pension funds, mutual funds, banks, insurance companies, and hedge funds all aggregate large pools of equity capital and may be more effective monitors. Rock (1991) and Romano (1993) give some reason to be cautious about their impact, however. They point out that the interests of money managers sometimes conflict with those of other shareholders. Banks, pension funds and insurance companies may side with incumbent managers if doing so gives them other opportunities to profit by managing the firm's pension funds, making loans or selling other services. Index funds have different disincentives to monitor. They compete on cost, and activism would increase their costs. Public pension funds are frequently active in pressuring managers, but these funds are run by political appointees and may favour politically popular proposals unrelated to firm value. Thus, the empirical evidence suggests that institutional shareholder activism has had only weak effects on firm performance (see, for example, Romano, 2001).

Others, focusing on the rules that govern shareholder ownership and voting, are also cautious about the potential impact about institutional investors. Roe (1994) and Black (1992) argue that shareholder passivity and collective action problems are created not solely by economic forces but also by politically motivated legal constraints that limit the institutional shareholder's incentives and ability to check incumbent managers. In this political view of shareholder passivity, a variety of banking, insurance and financial regulations prevent institutional investors from owning larger stakes or from monitoring managers more closely.

More recently, hedge funds have begun to aggregate large blocks of stock and to use their voting power to influence firm policies. Some investigate whether hedge funds have interests that conflict with other shareholders, which would suggest that hedge fund activism should be regulated (see Kahan and Rock, 2007; and Hu and Black, 2006). The alternative view is that hedge funds’ large stakes and relative freedom from regulatory restrictions allows them to overcome collective action problems and to monitor managers.

Shareholder suits

The law provides mechanisms by which shareholders can collectively sue managers for mismanagement. As a means of controlling agency costs, however, shareholder suits are flawed. Because most shareholders will gain little from a successful lawsuit, shareholders often have no incentive to initiate or monitor these suits. Unless a major institutional shareholder is involved as lead plaintiff, lawyers initiate the suits, pay all costs, make litigation decisions, including settlement decisions, and collect a fee if the plaintiff class collects. To the extent the lawyer's interests diverge from the interests of the shareholders, agency costs are present on the plaintiffs’ side of these lawsuits.

On the defendants’ side, the familiar agency costs are present. Managers can use corporate funds to protect themselves – appropriately in some cases and inappropriately in others. They use corporate funds to purchase directors’ and officers’ liability insurance, which covers their personal liability and defence costs, unless they are proven to have engaged in deliberate fraud or the equivalent. Management can also use corporate funds to settle suits. Alexander (1991), Macey and Miller (1991), Coffee (1985; 1986), Romano (1991), Bohn and Choi (1996), among others, have argued that meritorious suits against management settle too easily, and that the prospect of settlement encourages frivolous suits.

The result of this battle of agents is nearly always a settlement in which the corporation and/or its

directors' and officers' liability (D&O) insurer are the sole sources of payments. Consequently, payments go from shareholder to shareholder either directly or via insurance companies through premiums. Unless these suits deter mismanagement, the net winners are the lawyers on both sides. The shareholders in the aggregate are net losers (see Arlen and Carney, 1992; Langevoort, 1996; Mahoney, 1992; Easterbrook and Fischel, 1985).

Without commenting on the merits of these suits, Black, Cheffins and Klausner (2006) found only 13 cases, out of several thousand filed since 1980, in which outside directors have made personal payments. Inside managers bear personal liability more often, but settlement dynamics leave their assets untouched in all but a handful of cases per year (Alexander, 1991). Consequently, there is a question whether these suits have a significant deterrent effect.

The Public Securities Litigation Reform Act of 1995 (PSLRA) created several mechanisms designed to deter the filing of non-meritorious suits and to deter early settlement of meritorious suits. For instance, the law empowered the courts to select a lead plaintiff to monitor the shareholders' lawyer, with a presumption favouring institutional shareholders with substantial shareholdings. The law also requires a court to dismiss a suit unless the plaintiffs have alleged particular facts that support a 'strong inference' that a violation of the securities laws was committed with the legally required intent. This requirement was directed at the reported practice by which lawyers would file suits simply because a company's shares took a sharp drop in price, and then force the company into an expensive discovery process.

Ever since its enactment, scholars have tried to assess the impact of the PSLRA on securities class actions. Event studies, on the whole, have indicated that the law had a positive impact on share prices (Spiess and Tkac, 1997; Johnson, Kasznik and Nelson, 2000; Johnson, Nelson and Pritchard, 2000.) However, Ali and Kallapur (2001) found that the legislation had a negative impact on share prices. Studies have also tended to show that the PSLRA reduced the filing of non-meritorious suits (Johnson, Nelson and Pritchard, 2000; Bajaj, Muzumdar and Sarin, 2003). Others suggest, however, that some meritorious suits are deterred as well (Choi, 2007; Sale, 1998).

Choi, Fisch and Pritchard (2005), Thomas and Cox (2006) and Perino (2006) have shown that, while private institutional shareholders have not assumed the role of lead plaintiff, public pension plans have assumed that role to some extent. Perino (2006) found evidence consistent with monitoring by public pension plans.

Market for corporate control

The market for corporate control in the United States is regulated by state and federal law and is an important check on agency costs. If a firm is run poorly, either because managers are inattentive, consume too many perks or miss profitable merger opportunities, it may become the target of a takeover and its managers replaced. An active market in corporate control thus gives managers incentive to increase firm value (Manne, 1965).

In a 'hostile takeover' a buyer attempts to purchase a large block of stock, use its voting power to oust incumbent managers, purchase the remaining shares, and replace management. Alternatively, in the shadow of a hostile takeover, managers can agree to a 'friendly merger'. Both are associated with large gains to target shareholders. The evidence generally suggests that the premium comes from

improvements in firm performance (see Andrade, Mitchell and Stafford, 2001; Romano, 1992).

The law and economics literature has focused on three questions. First, what should managers do when the firm becomes the target of a hostile takeover? Easterbrook and Fischel (1982; 1991) argue that target management should remain passive and that the law should prohibit them from resisting a takeover. They argue that resistance will reduce bidder returns and thus bidders' incentive to engage in takeovers. This will in turn reduce the disciplinary threat of takeovers and increase agency costs generally. Gilson (1982) and Bebchuk (1982) argue that managers should resist a takeover attempt to the extent necessary to hold an auction, which will assure that the assets of the firm end up in their highest valued use.

A second question involves whether managers' negotiating over a potential merger should be allowed to grant termination fees or 'lock-ups' to favoured bidders. Such measures may discourage competition and affect the outcome of an auction, raising the risk that managers will favour particular bidders in exchange for private benefits, such as job security. Ayres (1990) and Fraidin and Hanson (1994) argue that termination fees and lock-ups will often not change the outcome of the auction and should therefore not be disfavoured. Kahan and Klausner (1996) examine how termination fees and lock-ups affect bidders' incentives to make a bid in the first place and their impact on agency costs generally. They explain that there is no reason for a target to grant a termination fee greater than a bidder's cost of making a bid.

Finally, a large literature examines whether, on average, takeover defences help or harm shareholder wealth. The typical research strategy examines how a firm's stock price reacts to the adoption of a takeover defence (see, for example, Comment and Schwert, 1995). This strategy usually suffers from a fatal flaw: it ignores the fact that the most potent defence, the 'poison pill', is freely available to all firms even after a hostile bid is received. Therefore, in effect, all firms have a poison pill and most other takeover defences are relatively unimportant. To disable a poison pill, a hostile bidder must first wage a proxy fight to unseat incumbent managers, install new managers who can remove the poison pill, and then go forward with the merger. The only takeover defences that are relevant other than a poison pill are those that either prevent an acquirer from replacing a target board or delay an acquirer's effort to do so. The most common defence is a classified (or staggered) board, which prevents an acquirer from taking control of a target board for two annual election cycles (see, for example, Daines, 2006; Faleye, 2007; Coates and Subramanian, 2002). Dual class stock, which is rarely used, allows management to control the election of the board and can therefore prevent an acquisition altogether.

A related literature examines whether firm takeover defences and shareholder rights predict stock returns (see, for example, Gompers, Ishii and Metrick, 2003; Cremers and Nair, 2005).

Mandatory disclosure

The monitoring mechanisms described above all depend, in part, on informed shareholders. Shareholder monitoring (of the kind contemplated by voting, law suits and the market for corporate control) is more effective when investors are informed. Thus, in many ways, the central regulatory event in US financial history was probably the 1933 and 1934 Acts, which created the Securities and Exchange Commission and required that publicly traded firms disclose detailed information about their historical performance and financial condition. These rules force firms both to disclose what they would otherwise prefer to keep private and to keep private information they might otherwise wish to disclose.

It is easy to see why disclosure might be valuable to investors. Accurate information allows investors to price securities and to monitor managers' performance. It is less easy to see why disclosure rules must be mandatory. Firms that fail to disclose will find it hard to raise money, as investors may take silence for bad news and refuse to invest (Ross, 1979; Grossman, 1981; Milgrom, 1981). Therefore, firms and entrepreneurs may find it in their own interest to disclose information, whether or not it is required by law.

However, firms would not always find full disclosure to be in their interest. Disclosure imposes direct costs as firms produce and verify the information, as well as indirect costs if competitors, customers and others can use the information to the firm's disadvantage. Moreover, the costs and benefits of disclosure are likely to vary between firms. Left to their own devices, therefore, firms will commit to varying levels of disclosure. Some therefore argue that markets can sort out the costs and benefits of disclosure and believe that uniform and mandatory disclosure requirements are unnecessary and even harmful (Romano, 2005; Mahoney, 1997; Choi and Guzman, 1997). Others believe that there are externalities from a firm's disclosures and that a mandatory rule may therefore be socially beneficial (Easterbrook and Fischel, 1991; Coffee, 1984; Dye, 1990; Admati and Pfleiderer, 2000).

Empirical evidence has not conclusively resolved this debate. Stigler (1964), Benston (1969; 1973) and Simon (1989), report evidence that mandatory disclosure did not improve investor welfare, but may have changed the characteristics of firms that go public. Recent evidence examines the effect of mandatory disclosure on firm returns and on asymmetric information (see Greenstone, Oyer and Vissing-Jorgenson, 2006; Daines and Jones, 2007).

A related debate involves whether managers should be allowed to trade on non-public information. Some hold that trading by informed insiders reveals valuable information and reduces agency costs (Manne, 1966; Carlton and Fischel, 1983). Others argue that insider trading is inefficient (Cox, 1986; Kraakman, 1991) or reduces stock market liquidity (Goshen and Parchomovsky, 2000). Beny (2007) reviews international evidence.

Creditors and the corporation

Because the corporation is a legal entity, distinct from its shareholders and managers, shareholders in the firm have 'limited liability' in that they are generally not personally liable for the debts of the corporation. At worst, public shareholders can lose their equity in the firm if the firm becomes insolvent. This separate legal status gives rise to two issues. First, because shareholders will reap the upside of the firm's successes but will not bear the full downside of its failures, managers may promote the interests of shareholders at the expense of creditors (Jensen and Meckling, 1976). The legal rule of 'veil piercing' developed to respond to this problem, though to a very limited extent. Under extreme circumstances in which a corporation is undercapitalized and other conditions are met, a court may impose liability on the corporation's shareholders. As a practical matter, however, this rule is not applied to public companies' shareholders, and in the private company context the courts' application of this rule is notoriously unpredictable (Thompson, 1991).

The rule of limited liability makes sense for contract creditors, who can negotiate their own protection from default or charge and interest rate that compensates for the risk. Tort creditors, however, are

different. Those owed compensation for, say, a firm's pollution emissions, will not have had the opportunity to negotiate with the firm *ex ante* to address the possibility that it will not have sufficient net assets to pay them. Thus, to deter corporate management from externalizing costs in the form of accidents and other torts and to prevent excess investment in risky activities, Hansmann and Kraakman (1991) argue that it may be desirable, and practical, to hold public shareholders personally liable for a corporation's torts. Grundfest (1992) and Alexander (1992) disagree as to the practicality of this proposal. A second issue involving limited liability is the use of the corporate form to ‘partition’ assets to create separate pools of assets to bond separate debts and other contractual commitments. Hansmann and Kraakman (2000) explain how the partitioning of assets to separately bond the commitments of the corporate entity, individual shareholders and corporate entities within a group of affiliated corporations can promote efficiencies in creditor monitoring.

Sarbanes – Oxley Act of 2002

The Sarbanes–Oxley Act of 2002 (SOX) introduced sweeping corporate governance mandates on firms whose shares trade on US securities exchanges. Until this legislation, legal rules regarding substantive corporate governance were the province of US state law, and federal law was limited primarily to disclosure requirements. SOX imposed a series of federal requirements on the board operation and structure. Event studies of various legislative events leading to the enactment of SOX yielded mixed results. Li, Pincus and Rego (2004), Jain and Rezaee (2006), and Chhaochhaira and Grinstein (2004) show a positive reaction, but Zhang (2005) shows a negative reaction. Litvak (2007) finds a negative reaction by comparing foreign cross-listed firms subject to SOX with cross-listed firm not subject to SOX. Aggarwal and Williamson (2006) found that six of the governance structures mandated by SOX (all related to board independence) had a positive impact on share value when adopted by firms voluntarily prior to SOX. Romano (2005), on the other hand, looked at other SOX requirements (loans to officers, executive certification of financials, auditors’ provision of non-audit services, and audit committee independence) and reports that there is no evidence to support their value to shareholders. Linck, Netter and Yang (2006) find that whatever the benefit of SOX, it increased the cost of boards, especially for small firms.

See Also

- corporate governance

Bibliography

Admati, A.R. and Pfleiderer, P. 2000. Forcing firms to talk: disclosure regulation and externalities. *Review of Financial Studies* 13, 479–519.

Aggarwal, R. and Williamson, R. 2006. Did new regulations target the relevant corporate governance attributes? Working paper, McDonough School of Business, Georgetown University.

Alexander, J.C. 1991. Do the merits matter? A study of settlements in securities class actions. *Stanford Law Review* 43, 497–528.

Alexander, J.C. 1992. Unlimited shareholder liability through a procedural lens. *Harvard Law Review* 106, 387–445.

Ali, A. and Kallapur, S. 2001. Securities price consequences of the Private Securities Litigation Reform Act of 1995 and related events. *Accounting Review* 76, 431–60.

Andrade, G., Mitchell, M. and Stafford, E. 2001. New evidence and perspectives on mergers. *Journal of Economic Perspectives* 15(2), 103–20.

Arlen, J. and Carney, W. 1992. Vicarious liability for fraud on securities markets. *University of Illinois Law Review* 1992, 691–745.

Ayres, I. 1990. Analyzing stock lock-ups: do target treasury sales foreclose or facilitate takeover options? *Columbia Law Review* 90, 682–718.

Bajaj, M., Muzumdar, S. and Sarin, A. 2003. Securities class action settlements: an empirical analysis. *Santa Clara Law Review* 43, 1001–33.

Bebchuk, L. 1982. The case for facilitating competing tender offers. *Stanford Law Review* 35, 24–47.

Bebchuk, L.A. and Cohen, A. 2003. Firms' decisions where to incorporate. *Journal of Law and Economics* 46, 383–425.

Benston, G.J. and Cohen, A. 1969. The value of the SEC's accounting disclosure requirements. *Accounting Review* 44, 515–32.

Benston, G.J. 1973. Required disclosure and the stock market: an evaluation of the Securities Exchange Act of 1934. *American Economic Review* 63, 132–55.

Beny, L. 2007. Insider trading laws and stock markets around the world. *Journal of Corporate Law* 32, 237–300.

Black, B. 1992. Next steps in proxy reform. *Journal of Corporation Law* 18, 1–55.

Black, B., Cheffins, B. and Klausner, M. 2006. Outside director liability. *Stanford Law Review* 58, 1055–60.

Bohn, J. and Choi, S.J. 1996. Fraud in the new-issues market: empirical evidence on securities class actions. *University of Pennsylvania Law Review* 144, 903–64.

Carlton, D. and Fischel, D. 1983. The regulation of insider trading. *Stanford Law Review* 35, 857–95.

Cary, W. 1974. Federalism and corporate law: reflections upon Delaware. *Yale Law Journal* 83, 663–705.

Chhaochhaira, V. and Grinstein, Y. 2004. Corporate governance and firm value: the impact of the 2002 governance rules. Working paper, Cornell University.

Choi, S.J. 2004. The evidence on securities class actions. *Vanderbilt Law Review* 57, 1465–525.

Choi, S. 2007. Do the merits matter less after the Private Securities Litigation Reform Act? *Journal of Law, Economics & Organization*.

Choi, S.J., Fisch, J.E. and Pritchard, A.C. 2005. Do institutions matter? The impact of the Lead Plaintiff Provision of the Private Securities Litigation Reform Act. *Washington University Law Quarterly* 83, 869–905.

Choi, S. and Guzman, A. 1997. National laws, international money: regulation in a global capital market. *Fordham Law Review* 65, 1855–908.

Coase, R. 1937. The nature of the firm. *Economica* 4, 386–405.

Coates, J.C., IV. 2001. Explaining variation in takeover defenses: blame the lawyers. *California Law Review* 89, 1301–415.

Coates, J.C., IV. and Subramanian, G. 2002. The powerful antitakeover force of staggered boards. *Stanford Law Review* 54, 887–951.

Coffee, J.C., Jr. 1984. Market failure and the economic case for a mandatory disclosure system. *Virginia Law Review* 70, 717–53.

Coffee, J.C., Jr. 1985. The unfaithful champion: the plaintiff as monitor in shareholder litigation. *Law and Contemporary Problems* 48, 5–81.

Coffee, J.C., Jr. 1986. Understanding the plaintiff's attorney: the implications of economic theory for private enforcement of law through class and derivative actions. *Columbia Law Review* 86, 669–727.

Comment, R. and Schwert, G.W. 1995. Poison or placebo? Evidence on the deterrence and wealth

effects of modern antitakeover measures. *Journal of Financial Economics* 39, 3–43.

Cox, J.D. 1986. Insider trading regulation and the production of information: theory and evidence. *Washington University Law Quarterly* 64, 475–505.

Cremers, M. and Nair, V.B. 2005. Governance mechanisms and equity prices. *Journal of Finance* 60, 2859–94.

Daines, R. 2001. Does Delaware law improve firm value? *Journal of Finance & Economics* 62, 525–58.

Daines, R. 2006. Do classified boards affect firm value? Takeover defenses after the poison pill. Working paper, Stanford Law School.

Daines, R. and Jones, C. 2007. Mandatory disclosure, asymmetric information and liquidity: the impact of the 1934 Act. Working paper, Law School, Stanford University.

Daines, R. 2002. The incorporation choices of IPO firms. *New York University Law Review* 77, 1559–605.

Daines, R. and Klausner, M. 2001. Do IPO charters maximize firm value? Antitakeover protection in IPOs. *Journal of Law, Economics, & Organization* 17, 83–120.

Dye, R.A. 1990. Mandatory v. voluntary disclosures: the cases of financial and real externalities. *Accounting Review* 65, 1–24.

Easterbrook, F. and Fischel, D. 1982. Auctions and sunk costs in tender offers. *Stanford Law Review* 35, 1–19.

Easterbrook, F. and Fischel, D. 1985. Optimal damages in securities cases. *University of Chicago Law Review* 52, 611–42.

Easterbrook, F. and Fischel, D. 1989. The corporate contract. *Columbia Law Review* 89, 1416–48.

Easterbrook, F. and Fischel, D. 1991. *The Economic Structure of Corporate Law*. Cambridge, MA: Harvard University Press.

Faleye, O. 2007. Classified boards, firm value, managerial entrenchment. *Journal of Financial Economics* 83, 501–29.

Field, L.C. and Karpoff, J.M. 2002. Takeover defenses of IPO firms. *Journal of Finance* 57, 1857–89.

Fraidin, S. and Hanson, J. 1994. Toward unlocking lockups. *Yale Law Journal* 103, 1739–834.

Gilson, R. 1982. Seeking competitive bids versus pure passivity in tender offer defense. *Stanford Law Review* 35, 51–67.

Gilson, R.J. and Schwartz, A. 2001. Sales and elections as methods for transferring corporate control. *Theoretical Inquiries in Law* 2, 783–814.

Gompers, P.A., Ishii, J.L. and Metrick, A. 2003. Corporate governance and equity prices. *Quarterly Journal of Economics* 118, 107–55.

Goshen, G. and Parchomovsky, G. 2000. On insider trading, markets, and ‘negative’ property rights in information. Fordham Law & Economics Research Paper No. 06.

Greenstone, M., Oyer, P. and Vissing-Jorgensen, A. 2006. Mandated disclosure, stock returns, and the 1964 Securities Acts Amendments. *Quarterly Journal of Economics* 121, 399–460.

Grossman, S.J. 1981. The information role of warranties and private disclosure about product quality. *Journal of Law and Economics* 24, 461–83.

Grundfest, J.A. 1992. The limited future of unlimited liability: a capital markets perspective. *Yale Law Review* 102, 387–425.

Hansmann, H. 2006. Corporation and contract. *American Law and Economics Review* 8, 1–19.

Hansmann, H. and Kraakman, R. 1991. Toward unlimited shareholder liability for corporate torts. *Yale Law Review* 100, 1879–934.

Hansmann, H. and Kraakman, R. 2000. The essential role of organizational law. *Yale Law Journal* 110, 387–440.

Hu, H. and Black, B. 2006. The new vote buying: empty voting and hidden ownership. *Southern California Law Review* 79, 811–908.

Jain, P. and Rezaee, Z.J. 2006. The Sarbanes–Oxley Act of 2002 and capital-market behavior: early evidence. *Contemporary Accounting Research* 23, 629–54.

Jensen, M.C. and Meckling, W.H. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 303–60.

Johnson, M., Kasanik, R. and Nelson, K. 2000. Shareholder wealth effects of the Private Securities

Litigation Reform Act of 1995. *Review of Accounting Studies* 5, 217–33.

Johnson, M., Kasznik, R. and Nelson, K. 2001. The impact of securities litigation reform on the disclosure of forward-looking information. *Journal of Accounting Research* 39, 297–328.

Johnson, M., Nelson, K. and Pritchard, A. 2000. In re Silicon Graphics Securities Litigation: shareholder wealth effects of the interpretation of the Private Securities Litigation Reform Act's pleading standard. *Southern California Law Review* 73, 773–809.

Kahan, M. 2006. The demand for corporate law: statutory flexibility, judicial quality, or takeover protection. *Journal of Law, Economics, & Organization* 22, 340–65.

Kahan, M. and Kamar, E. 2002. The myth of state competition in corporate law. *Stanford Law Review* 55, 679–760.

Kahan, M. and Klausner, M. 1996. Path dependence in corporate contracting: increasing returns, herd behavior and cognitive biases. *Washington University Law Quarterly* 74, 347–66.

Kahan, M. and Klausner, M. 1997. Standardization and innovation in corporate contracting (or ‘The economics of boilerplate’). *Virginia Law Review* 83, 713–55.

Kahan, M. and Rock, E. 2007. Hedge funds in corporate governance and corporate control. *University of Pennsylvania Law Review* (forthcoming).

Klausner, M. 1995. Corporations, corporate law, and networks of contracts. *Virginia Law Review* 81, 757–833.

Klausner, M. 2006. The contractarian theory of corporate law: a generation later. *Journal of Corporate Law* 31, 779–97.

Kraakman, R. 1991. The legal theory of insider trading regulation in the United States. In *European Insider Dealing*, ed. K. Hopt and E. Wymeerisch. London: Butterworths.

Langevoort, D.C. 1996. Capping damages for open-market securities fraud. *Arizona Law Review* 38, 639–68.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R.W. 1997. Legal determinants of external finance. *Journal of Finance* 52, 1131–50.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R.W. 1998. Law and finance. *Journal of Political Economy* 106, 1113–55.

Li, H., Pincus, M. and Rego, S. 2004. Market reaction to events surrounding the Sarbanes–Oxley Act of 2002 and earnings management. Working paper, University of Iowa.

Linck, J., Netter, J. and Yang, T. 2006. Effects and unintended consequences of the Sarbanes–Oxley Act on corporate boards. Working paper, University of Georgia.

Litvak, K. 2007. The effect of the Sarbanes–Oxley Act on non-US companies cross-listed in the US. *Journal of Corporate Finance*.

Macey, J. and Miller, G. 1991. The plaintiffs' attorney's role in class action and derivative litigation: economic analysis and recommendations for reform. *Chicago Law Review* 58, 1–94.

Mahoney, P. 1992. Precaution costs and the law of fraud in impersonal markets. *Virginia Law Review* 78, 623–60.

Mahoney, P.G. 1997. The exchange as regulator. *Virginia Law Review* 83, 1453–500.

Manne, H. 1965. Mergers and the market for corporate control. *Journal of Political Economy* 73, 110–20.

Manne, H. 1966. *Insider Trading and the Stock Market*. New York: Free Press.

Milgrom, P. 1981. Good news and bad news: representation theorems and application. *Bell Journal of Economics* 12, 380–91.

Perino, M. 2006. Institutional activism through litigation: an empirical analysis of public pension fund participation in securities class actions. Legal Studies Research Paper No. 06–0055, St. John's University.

Pritchard, A.C. and Sale, H. 2005. What counts as fraud? An empirical study of motions to dismiss under the Private Securities Litigation Reform Act. *Journal of Empirical Legal Studies* 2, 125–49.

Rajan, R.G. and Zingales, L. 1998. Financial dependence and growth. *American Economic Review* 88, 559–86.

Rock, E. 1991. The logic and (uncertain) significance of institutional shareholder activism. *Georgetown Law Review* 79, 445–506.

Roe, M. 1994. *Strong Managers, Weak Owners*. Princeton: Princeton University Press.

Romano, R. 1985. Law as a product: some pieces of the incorporation puzzle. *Journal of Law*,

Economics, & Organization 1, 225–83.

Romano, R. 1987. The political economy of takeover statutes. *Virginia Law Review* 73, 111–98.

Romano, R. 1991. The shareholder suit: litigation without foundation? *Journal of Law, Economics, & Organization* 7, 55–87.

Romano, R. 1992. A guide to takeovers: theory, evidence and regulation. *Yale Journal on Regulation* 9, 119–80.

Romano, R. 1993. Public pension fund activism in corporate governance reconsidered. *Columbia Law Review* 93, 795–852.

Romano, R. 1998. Empowering investors: a market approach to securities regulation. *Yale Law Journal* 107, 2359–430.

Romano, R. 2001. Less is more: making institutional investor activism a valuable mechanism of corporate governance. *Yale Journal on Regulation* 18, 174–252.

Romano, R. 2005. The Sarbanes–Oxley Act and the making of quack corporate Governance. *Yale Law Journal* 114, 1521–611.

Ross, S.A. 1979. Disclosure regulation in financial markets: implications of modern finance theory and signaling theory. In *Issues in Financial Regulation: Regulation of American Business and Industry*, ed. F. R. Edwards. New York: McGraw Hill.

Sale, H. 1998. Heightened pleading and discovery stays: an analysis of the effect of the PSLRA's Internal-Information Standard on ’33 and ’34 Act claims. *Washington University Law Quarterly* 76, 537–95.

Scharfstein, D.S. and Stein, J. 1990. Herd behavior and investment. *American Economic Review* 80, 465–89.

Schwert, G.W. 1995. Comment: Poison or placebo: evidence on the deterrence and wealth effects of modern antitakeover measures. *Journal of Financial Economics* 39, 3–43.

Simon, C. 1989. The effect of the 1933 Securities Act on investor information and the performance of new issues. *American Economic Review* 79, 295–318.

Spiess, D.K. and Tkac, P. 1997. The Private Securities Litigation Reform Act of 1955: the stock market casts its vote. *Managerial and Decision Economics* 18, 545–61.

Subramanian, G. 2004. The disappearing Delaware effect. *Journal of Law, Economics, & Organization* 20, 32–59.

Stigler, G. 1964. Public regulation of the securities markets. *Journal of Business* 2, 117–42.

Thomas, R. and Cox, J. 2006. Does the plaintiff matter? An empirical analysis of lead plaintiffs in securities class actions. *Columbia Law Review* 100, 101–55.

Thompson, R. 1991. Piercing the corporate veil: an empirical study. *Cornell Law Review* 76, 1036–74.

Thompson, R. and Sale, H. 2003. Securities fraud as corporate governance: reflections upon federalism. *Vanderbilt Law Review* 56, 859–915.

Thompson, R. and Thomas, R. 2004. The public and private faces of derivative lawsuits. *Vanderbilt Law Review* 58, 1747–823.

Winter, R. 1977. State law, shareholder protection and the theory of the corporation. *Journal of Legal Studies* 6, 251–92.

Zhang, I. 2005. Economic consequences of the Sarbanes–Oxley Act of 2002. Related Publication 05–07. AEI–Brookings Joint Center for Regulatory Studies.

Zwiebel, J. 1995. Corporate conservatism, herd behavior and relative compensation. *Journal of Political Economy* 103, 1–25.

How to cite this article

Daines, Robert and Michael Klausner. "corporate law, economic analysis of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000285> doi:10.1057/9780230226203.0324

The New Palgrave Dictionary of Economics Online

corporations

Randall Morck

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A corporation is an artificial person with many of the rights of a biological one. The first business corporations pooled the savings of many individuals to permit ventures on a scale none could afford individually. Most large American and British corporations lack controlling shareholders; the consequent lack of monitoring and control gives rise to corporate governance problems reflecting the private benefits of control. The view that corporations should be run to maximize shareholder conflicts in many countries with the actual legal duties of corporate officers, and collides with evidence that stock prices are sometimes set by investors with incomplete information.

Keywords

agency costs; corporate governance; corporations; limited liability; monitoring; noise traders; private benefits of control; stock markets; time inconsistency

Article

A *corporation* is an artificial person, with many of the legal rights of a biological one. This modern legal and economic usage arose in the 16th century from the term's now archaic meaning of ‘a group acting as one body’ – encompassing municipal governments, businesses and other groups of individuals united towards a common goal. In that century and the next, trade with the Orient and the New World promised immense returns, but only after vast capital outlays on fleets of ships, networks of forts and private armies to defend them. The first business corporations, such as the Dutch East Indies Company, the British East India Company and the Hudson's Bay Company, were formed to pool the savings of many individuals and permit ventures on a scale none could afford individually. Each owner of a *share* of the corporation was periodically entitled to a *dividend* – a pro rata division of the corporation's free cash flow.

Polling all a corporation's shareholders for each business decision was impractical in an age of sailing ships and horse-drawn carriages. Instead, the shareholders elected *boards of governors* (later *directors*) –

reputable men trusted by the majority of shareholders to direct the corporation's affairs.

This did not prevent all dispute. The Dutch East Indies Company (*Vereenigde Oostindische Compagnie* in Dutch) was formed as a limited time venture. When that limit drew near, the board boldly announced that the corporation would persist indefinitely. The shareholders sued to force a liquidating dividend – and lost! Fortunately, they found they could sell their shares to other investors for the value of a liquidating dividend – or even more (Frentrop, 2002/3). Thus was born the first modern stock market, and the *alienability*, or unhindered sale, of shares became a defining characteristic of a corporation. Letting shareholders realize their investments by selling their shares, rather than liquidating the business, gave corporations a second defining characteristic: *indefinitely long lives*.

Boards occasionally betrayed their shareholders' trust and caused a corporation to contravene the law. Since individual shareholders were not consulted, holding them fully to account for the corporation's misdeeds seemed wrong. Since the corporation is a legal person, plaintiffs could sue it directly, and need not sue its shareholders personally. Thus, *limited liability* statutes came to shield individual shareholders from personal lawsuits for wrongs by corporations whose shares they own. Limited liability, a third defining characteristic of the modern business corporation, is an important innovation because it frees individuals to invest their savings in corporations run by strangers, undertaking risky ventures, or doing business in far off places. Vulnerability to personal lawsuits would otherwise make such investments seem indefensibly reckless to most savers.

Early corporations, like the Hudson's Bay Company, assigned one vote to each share in board elections. This essentially let the wealthiest shareholders appoint the board and, if they wished, run the corporation in their narrow interest rather than in the interests of all shareholders equally. For example, a large shareholder might force the corporation to do business with another corporation she controlled on disadvantageous terms. This sort of self-dealing, which Johnson et al. (2000) dub ‘tunneling’, remains a widespread corporate governance concern where firms typically have dominant shareholders. Or a dominant shareholder might simply relish the perks, power and prestige of running the corporation, and refuse to make way for more qualified managers – a corporate governance problem called ‘entrenchment’ (Morck, Shleifer and Vishny, 1988). Entrenchment and tunnelling provide controlling shareholders with *private benefits of control* – returns not shared with small shareholders (Dyck and Zingales, 2004). Distorted corporate governance associated with private benefits of control remains a first-order governance concern wherever corporations typically have a controlling shareholder. According to La Porta, Lopez-de-Silanes and Shleifer (1999), this includes the large corporate sectors of virtually all countries except Germany, Japan, the United Kingdom and the United States. Small and middle-sized corporations everywhere tend to have controlling shareholders.

In the 19th century, *democratic* corporate governance became associated with *one vote per shareholder*, rather than one vote per share (Dunlavy, 2004). Echoes of this remain in the *voting caps* of modern Canadian and European corporations, which limit any single shareholder's voting power regardless of shares owned. However, large shareholders in many countries later turned deviations from one vote per share to their advantage by granting themselves special classes of common stock with many votes per share. In most countries, such *dual class shares* now virtually always magnify, rather than limit, the voting power of large shareholders, and so amplify, rather than dampen, problems associated with private benefits of control (Nenova, 2003).

In the United States and the United Kingdom, however, one vote per share is the norm in shareholder

meetings. Disclosure rules, regulatory oversight, officer and director liability, and other restraints on private benefits of control also seem more effective in America and Britain than elsewhere in curtailing private benefits of control (LaPorta, Lopez-de-Silanes and Shleifer, 1999; Dyck and Zingales, 2004). This makes being a large shareholder less attractive, especially if holding a diversified portfolio of small stakes in many firms reduces risk (Burkart, Panunzi and Shleifer, 2003). Unsurprisingly, most large American and British corporations now lack controlling shareholders (LaPorta, Lopez-de-Silanes and Shleifer, 1999). They are run by professional managers who often own few shares (Morck, Shleifer and Vishny, 1988).

A small shareholder who monitored and controlled these corporate top managers would bear all the investigative, legal and administrative costs involved, but the benefits of better governance would be spread across all shareholders. The cost therefore typically exceeds the benefit for any small shareholder acting alone (Grossman and Hart, 1988). The consequent general lack of monitoring and control in corporations with no large shareholder gives rise to *other people's money* corporate governance problems. Adam Smith (1776) famously explains that, since corporate managers who own few or no shares are more

> the managers of other people's money than of their own, it cannot well be expected that they should watch over it with the same anxious vigilance with which partners in a private copartnery frequently watch over their own. Like the stewards of a rich man, they…. consider attention to small matters as not for their master's honour and very easily give themselves a dispensation from having it.

Unmonitored professional managers can thus enjoy the perks and privileges of running large corporations without any real concern for the returns they generate. Berle and Means (1932) argue that this sort of governance problem occurs in many large American corporations.

But in other countries, other people's money governance problems probably also afflict many corporations that, on first inspection, seem to have a controlling shareholder. This is because large corporations in most countries are not freestanding entities, but belong to *corporate groups* (LaPorta, Lopez-de-Silanes and Shleifer,1999). These are typically pyramidal structures, in which an apex shareholder, usually an extremely wealthy family, controls one or more listed corporations, which each control more listed corporations, which each control yet more listed corporations, *ad valorem et infinitum*. A family that controls 51 per cent of a listed corporation that controls 51 per cent of another that controls 51 per cent of yet another and so on actually owns only 0.51^n of the corporation n tiers down the in pyramid, with the remainder of each corporation financed by public or minority shareholders. Pyramids with a dozen or more layers are not uncommon, rendering the controlling shareholder's actual ownership of corporations at the pyramid's base negligible. Pyramidal business groups thus permit controlling shareholders to extract private benefits of control from corporate empires financed largely from other people's money (Morck, Stangeland and Yeung, 2000; Bebchuk, Kraakman and Triantis, 2000). Pyramids were common in the United States until the 1930s (Berle and Means, 1932; Bonbright and Means, 1932), but were eliminated by various New Deal initiatives, including the double and multiple taxation of inter-corporate dividends (Morck, 2005). British pyramids apparently

withered under sustained attacks from institutional investors (Franks, Mayer and Rossi, 2005). However, the relevant unit of economic analysis for many purposes elsewhere in the world should often be the *business group*, not the corporation.

Jensen and Meckling (1976) show that *agency costs*, the present value of the costs of expected future governance shortfalls of any sort, are born by the corporation's initial shareholders. A corporation's founders receive less per share when they first sell shares to outside investors if worse corporate governance problems seem likely.

This gives rise to a *time inconsistency* problem in securities and corporations law. Investors and entrepreneurs selling shares to the public benefit from credible guarantees of good governance because these limit agency costs and so raise share prices. But top corporate decision makers in firms that have already issued shares, who foresee issuing no more, wish to maximize their utility (Baumol, 1959; 1962; Williamson, 1964) and understandably value the freedom to spend public shareholders' money as they like and to capture such private benefits of control as they can. Actual public policy probably reflects these groups' relative political lobbying power, which can change over time (Morck, Stangeland and Yeung, 2000; Morck, Wolfenzon and Yeung, 2005).

The normative view that a corporation *should* be run to maximize shareholder value derives from economists' assumption that firms maximize profits. In neoclassical economic theory, a firm that maximizes the present value of all its expected future economic profits necessarily maximizes the market value of its shares. This follows from modelling the corporation as a *nexus* of contracts, with the shareholders the *residual claimants* to the firm's cash flows (Fama and Jensen, 1983a; 1983b).

Neoclassical theory further allows that profit maximization (value maximization in a multi-period setting) accords with economic efficiency under certain idealized conditions; see, for example, Varian (1992) and Malliaris and Brock (1983).

This normative view conflicts with the actual legal duties of corporate officers, directors and controlling shareholders in many countries. For example, many northern European countries and some US states impose a duty to balance shareholders' interests with those of *stakeholders*, especially employees. This is formalized in the German legal principle of *Mitbestimmung* (co-determination), which requires members of the *Aufsichtsrat* (supervisory board) of a large corporation to balance the interests of shareholders, employees and the state (Fohlin, 2005). Common law legal systems assign officers and directors a duty to act *for the corporation*. In Britain and the United States, this is often interpreted as a duty to act for the corporation's owners, its shareholders. A duty to maximize share value seems implicit (Jensen and Meckling, 1976; Black and Coffee, 1997). However, the Canadian Supreme Court holds in *Peoples v. Wise* that the duty of the officers and directors of a corporation is not to shareholders, nor to any other stakeholders, but to the corporation per se. The social welfare implications of assigning different legal duties to corporate top decision makers are incompletely understood. Giving labour a voice in corporate decision making seems to impede risk taking and hamper growth (Faleye, Mehrotra and Morck, 2006). Moreover, regardless of their assigned objective, if those entrusted to govern great corporations occasionally put their own interests ahead of their legal duties, agency costs must arise in some form.

The view that a corporation's top managers ought to maximize shareholder value also collides with evidence that stock prices are sometimes set by investors with incomplete information (Myers and Majluf, 1984) or behavioural biases (Shleifer, 2000). Coase (1937) argues that firms come about to

alleviate information asymmetries and other market imperfections, collectively denoted *transactions costs*, and that the boundaries of the firm correspond to an efficient solution to these problems. Alchian and Demsetz (1972) argue that the critical market imperfections arise from people working in teams. Williamson (1975) argues that interdependent assets are more generally important. Jensen (2004) calls for more research on normative theories about the boundaries of the corporation and the objective function of its top decision makers if stock prices are set by *noise traders*, that is, investors with behavioural biases. One approach holds that corporations actually exist primarily to lock the economy's capital into productive uses by isolating capital allocation decisions from maniac or panicked investors (Stout, 2004). This view long dominated discussions of corporate management in Japan (for example, Aoki and Dore, 1994) but appears to give rise to its own set of inefficiencies (see, for example, Morck and Nakamura, 1999).

See Also

- corporate governance
- firm, theory of the

Bibliography

Aoki, M. and Dore, R.P. 1994. *The Japanese Firm: The Sources of Competitive Strength*. New York: Oxford University Press.

Alchian, A. and Demsetz, H. 1972. Production, information costs and economic organization. *American Economic Review* 62, 777–95.

Baumol, W. 1959. *Business Behavior, Value and Growth*. New York: Macmillan.

Baumol, W. 1962. On the theory of expansion of the firm. *American Economic Review* 52, 1078–87.

Bebchuk, L., Kraakman, R. and Triantis, G. 2000. Stock pyramids, cross ownership and dual class equity: the mechanisms and agency costs of separating control from cash flow rights. In *Concentrated Corporate Ownership*, ed. R. Morck. Chicago: University of Chicago Press.

Berle, A. and Means, G. 1932. *The Modern Corporation and Private Property*. New York: Macmillan.

Black, B.S. and Coffee, J.C., Jr. 1997. Hail Britannia? Institutional investor behavior under limited regulation. *Michigan Law Review* 92, 1997–2087.

Bonbright, J. and Means, G. 1932. *The Holding Company – Its Public Significance and Its Regulation*. New York: McGraw-Hill.

Burkart, M., Panunzi, F. and Shleifer, A. 2003. Family firms. *Journal of Finance* 58, 2173–207.

Coase, R. 1937. The nature of the firm. *Economica* 4, 386–405.

Dunlavy, C. 2004. The unnatural origins of one vote per share – a chapter in the history of corporate governance. Working paper, Department of History, University of Wisconsin–Madison.

Dyck, A. and Zingales, L. 2004. Private benefits of control: an international comparison. *Journal of Finance* 59, 537–601.

Faleye, O., Mehrotra, V. and Morck, R. 2006. When labor has a voice in corporate governance. *Journal of Financial and Quantitative Analysis* 41, 489–510.

Fama, E. and Jensen, M. 1983a. Agency problems and residual claims. *Journal of Law and Economics* 26, 327–49.

Fama, E. and Jensen, M. 1983b. Separation of ownership and control. *Journal of Law and Economics* 26, 301–25.

Fohlin, C. 2005. The history of corporate ownership and control in Germany. In *The History of Corporate Governance around the World: Family Business Groups to Professional Managers*, ed. R. Morck. Chicago: University of Chicago Press.

Franks, J., Mayer, C. and Rossi, S. 2005. Spending less time with the family: the decline of family ownership in the UK. In *The History of Corporate Governance around the World: Family Business Groups to Professional Managers*, ed. R. Morck. Chicago: University of Chicago Press.

Frentrop, P. 2002/3. *A History of Corporate Governance*. Amsterdam: Deminor Press.

Grossman, S. and Hart, O. 1988. One share one vote and the market for corporate control. *Journal of Financial Economics* 20, 175–202.

Jensen, M. 2004. Agency costs of overvalued equity. Harvard NOM Research Paper No. 04-26, Harvard Business School.

Jensen, M. and Meckling, W. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

Johnson, S., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. 2000. Tunneling. *American Economic Review* 90, 22–7.

La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. 1999. Corporate ownership around the world. *Journal of Finance* 54, 471–517.

Malliaris, A.G. and Brock, W. 1983. *Stochastic Methods in Economics and Finance*. Amsterdam: North-Holland.

Morck, R. 2005. How to eliminate pyramidal business groups: the double-taxation of intercorporate dividends and other incisive uses of tax policy. *Tax Policy and the Economy* 19, 135–79.

Morck, R. and Nakamura, M. 1999. Banks and corporate control in Japan. *Journal of Finance* 54, 319–40.

Morck, R., Shleifer, A. and Vishny, R. 1988. Management ownership and market valuation: an empirical analysis. *Journal of Financial Economics* 20, 293–315.

Morck, R., Stangeland, D.A. and Yeung, B. 2000. Inherited wealth, corporate control, and economic growth: the Canadian disease. In *Concentrated Corporate Ownership*, ed. R. Morck. Chicago: University of Chicago Press.

Morck, R., Wolfenzon, D. and Yeung, B. 2005. Corporate governance, economic entrenchment, and growth. *Journal of Economic Literature* 43, 655–720.

Myers, S. and Majluf, N. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13, 187–222.

Nenova, Tatiana. 2003. The value of corporate voting rights and control: a cross-country analysis. *Journal of Financial Economics* 68, 325–51.

Shleifer, A. 2000. *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford: Oxford University Press.

Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Ward, Lock, and Tyler.

Stout, Lynn. 2004. On the nature of corporations. Law & Economics Research Paper No. 04-13, School of Law, UCLA.

Varian, H. 1992. *Microeconomic Analysis*. 3rd edn. New York: W. W. Norton.

Williamson, O. 1964. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs, NJ: Prentice Hall.

Williamson, O. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

How to cite this article

Morck, Randall. "corporations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000379> doi:10.1057/9780230226203.0325

The New Palgrave Dictionary of Economics Online

correspondence principle

Federico Echenique
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

comparative statics; correspondence principle; monotone models; strategic complementarities; tâtonnement; Walrasian general equilibrium

Article

The correspondence principle is the relation, which exists in certain economic models, between comparative statics of equilibria and the properties of out-of-equilibrium dynamics. The correspondence principle (CP) implies that one obtains unambiguous comparative statics by selecting equilibria with desirable dynamic properties. Generally, the CP determines comparative statics in models with a one-dimensional endogenous variable, and in monotone multidimensional models. It does not determine comparative statics in general multidimensional models, such as Walrasian general equilibrium models with more than two goods.

One-dimensional models

The CP holds quite generally in one-dimensional models. Consider, for example, a two-good economy with excess-demand function for good 1 given by z_1, shown in Figure 1. We fix the price of good 2; by Walras's Law the equilibrium prices are the zeroes of z_1: there are three equilibria, p_1^1, p_1^2 and p_1^3.

Figure 1
Two-good economy

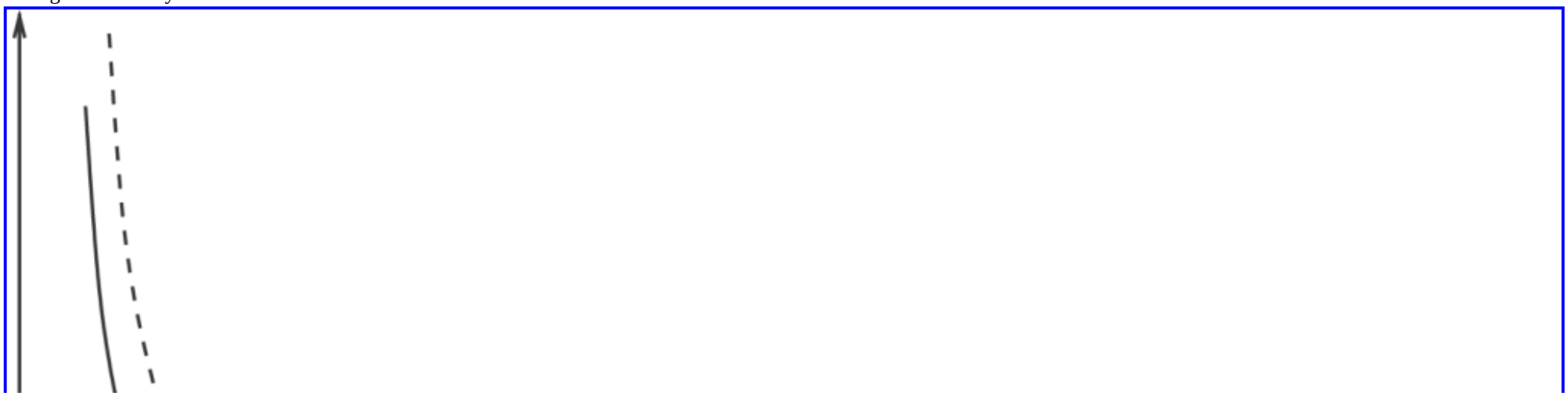

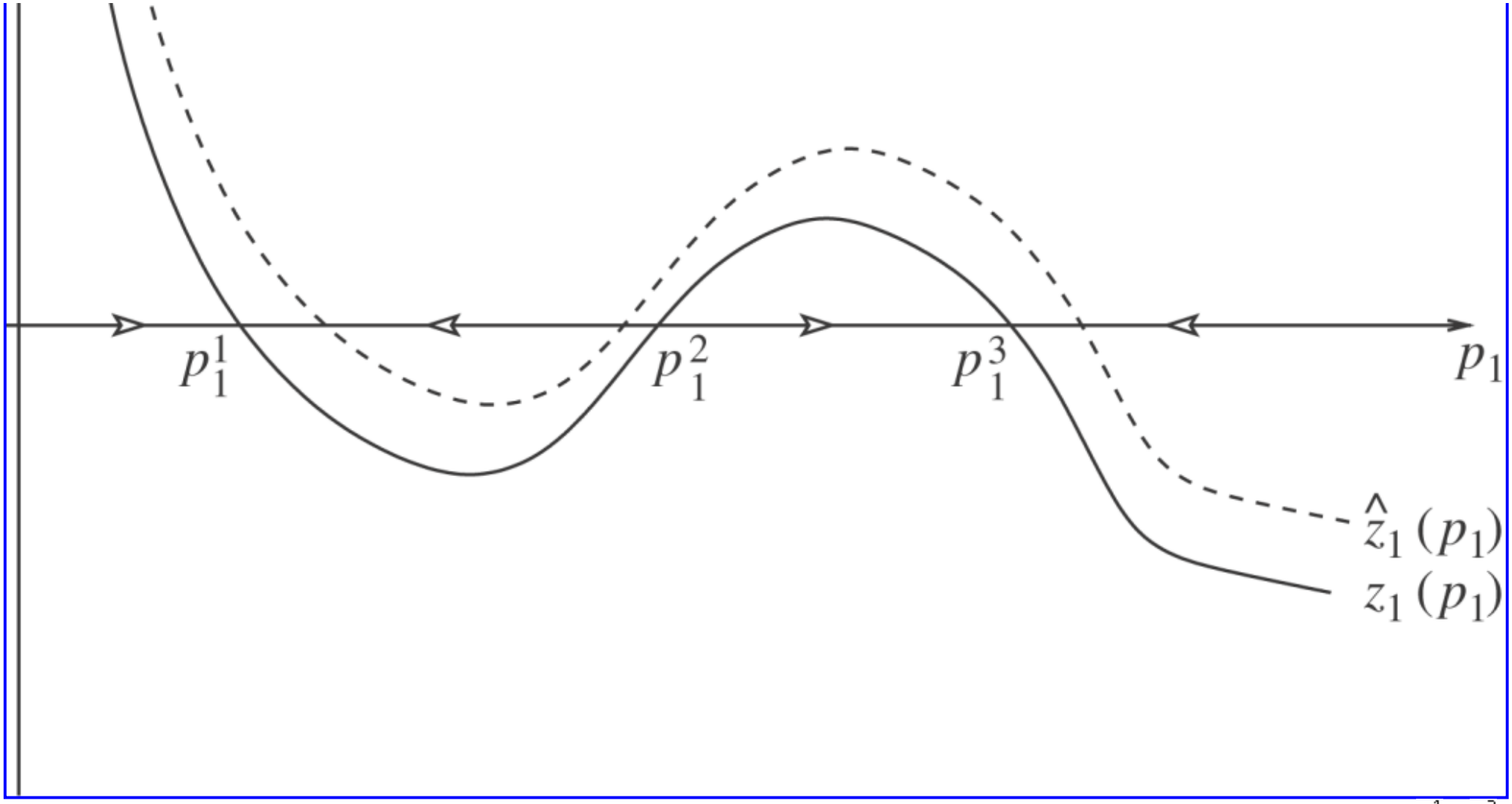

Now consider the comparative-statics exercise of shifting excess demand up to $\hat{z}_1$. What is the effect on equilibrium price? Locally, the price increases if the equilibrium is p_1^1 or p_1^3, but it decreases if it is p_1^2. The different comparative statics at p_1^1 and p_1^2 corresponds exactly to the different behavior of tâtonnement dynamics after a small perturbation: p_1^1 is stable while p_1^2 is unstable.

The difference between comparative statics at p_1^1 and at p_1^2 is easy to explain. The comparative statics at p_1^1 says: slightly larger prices than p_1^1 are reached by increasing excess demand, and smaller prices are reached by decreasing excess demand. Since excess demand is zero at p_1^1, there must be *positive* excess demand at slightly larger prices and *negative* excess demand at slightly smaller prices. Hence, tâtonnement dynamics, which respond to the sign of excess demand, converges to p_1^1 after a small perturbation from p_1^1. On the other hand, at p_1^2 larger prices result from a decrease in excess demand; hence excess demand is positive at larger prices. Similarly, excess demand is negative at smaller prices. As a result, tâtonnement dynamics will not approach p_1^2 after a small perturbation from p_1^2.

If the economy is subject to sporadic shocks, one should not observe p_1^2, the unstable equilibrium. Hence, as a consequence of the correspondence between comparative statics and dynamics, one should expect an increase in excess demand to produce an increase in equilibrium price.

I shall give a general statement of the correspondence principle for the one-dimensional case. Consider a model where the endogenous variable takes values in [0,1] and equilibria are determined as the fixed points of $f(\cdot, t): [0, 1] \to [0, 1]$; $t \in T \subseteq \mathbb{R}$ is an exogenous parameter. Assume that *T* is convex and that *f* is C^1.
A *selection of equilibria* is a function $e: T \to [0, 1]$ such that $e(t) = f(e(t), t)$ for all $t \in T$. Say that a fixed point $x \in [0, 1]$ is *stable* if there is a neighbourhood *V* of *x* such that any sequence x_n satisfying $x_0 \in V$ and $x_{n+1} = f(x_n)$ for $n \geq 1$, converges to *x*. Say that $x \in [0, 1]$ is *unstable* if, for any neighbourhood *V* of *x*, there is a neighbourhood *W* of *x* such that all sequences defined as above eventually lie in the complement of *W*.
Proposition 1: Let f be monotone increasing in t. If e is a continuous selection of equilibria that is strictly decreasing over some interval $[\underline{t}, \bar{t}]$, *then for all* $t \in (\underline{t}, \bar{t})$, $e(t)$ *is unstable*.

Multidimensional models

The one-dimensional CP is a relation between the sign of the comparative-statics change in prices, and the sign of excess demand for smaller and larger prices. When more than one price is determined, this relation does not need to exist. Still, the CP holds for monotone models – models where the different dimensions of the endogenous variables are in some sense complements. Monotone economic models stem mainly from game theoretic models with strategic complementarities.
I proceed to give a statement of the CP. Consider a model where the endogenous variable takes values in a compact rectangle $X \subseteq \mathbb{R}^n$, and equilibria are determined as the fixed points of $f(\cdot, t): X \to X$; $t \in T \subseteq \mathbb{R}$ is a parameter and *T* is convex.
Proposition 2: Let f be monotone increasing in (x, t) *and let e be a continuous selection of equilibria*.

- *If e is strictly decreasing over* $[\underline{t}, \bar{t}] \subseteq T$, *then for all* $t \in (\underline{t}, \bar{t})$, *e*(*t*) *is unstable*.
- *If e is strictly increasing over* $[\underline{t}, \bar{t}]$, *then for all* $t \in (\underline{t}, \bar{t})$, *if e*(*t*) *is locally isolated, it is stable*.

Literature

The CP was formulated by Paul Samuelson (1941; 1942; 1947), who also coined the term (though Hicks, 1939, stated the CP informally). Samuelson formulated the one-dimensional CP. The version in Proposition 1 is taken from Echenique (2000). Basset, Maybee and Quirk (1968) study the scope of the CP. Arrow and Hahn (1971) present a critical discussion of the CP, and, because it fails in economies with more than two goods, conclude that ‘very few useful propositions are derivable from this principle’. The monotone multidimensional CP is from Echenique (2002), who presents a general version of Proposition 2. Echenique (2004) presents a CP that does not rely on continuous selections of equilibria. The CP is also effective in dynamic optimization models (Brock, 1983; Burmeister and Long, 1977; Magill and Sheinkman, 1979) and in models of international trade (Bhagwati, Brecher and Hatta, 1987).

See Also

- comparative statics

Bibliography

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Bassett, L., Maybee, J. and Quirk, J. 1968. Qualitative economics and the scope of the correspondence principle. *Econometrica* 36, 544–63.

Bhagwati, J.N., Brecher, R.A. and Hatta, T. 1987. The global correspondence principle: a generalization. *American Economic Review* 77, 124–32.

Brock, W.A. 1983. A revised version of Samuelson's correspondence principle. In *Models of Economic Dynamics*, ed. H. Sonnenschein. New York: Springer-Verlag.

Burmeister, E. and Long, N.V. 1977. On some unresolved qsts in capital theory: an application of Samuelson's correspondence principle. *Quarterly Journal of Economics* 91, 289–

314.

Echenique, F. 2000. Comparative statics by adaptive dynamics and the correspondence principle. Working Paper No. E00-273, Department of Economics, University of California, Berkeley.

Echenique, F. 2002. Comparative statics by adaptive dynamics and the correspondence principle. *Econometrica* 70, 833–44.

Echenique, F. 2004. A weak correspondence principle for models with complementarities. *Journal of Mathematical Economics* 40, 145–52.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Magill, M.J.P. and Sheinkman, J.A. 1979. Stability of regular equilibria and the correspondence principle for symmetric variational problems. *International Economic Review* 20, 297–315.

Samuelson, P.A. 1941. The stability of equilibrium: comparative statics and dynamics. *Econometrica* 9, 97–120.

Samuelson, P.A. 1942. The stability of equilibrium: linear and nonlinear systems. *Econometrica* 10, 1–25.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

How to cite this article

Echenique, Federico. "correspondence principle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000383> doi:10.1057/9780230226203.0326

The New Palgrave Dictionary of Economics Online

correspondences

M. Ali Khan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Correspondences are versatile mathematical objects for which a rich theory can be developed. They arise naturally in many diverse areas of applied mathematics, including economic theory. For example, an individual consumer's demand correspondence associates with each price system the set of utility maximizing consumption plans. Similarly, an individual producer's supply correspondence associates with each price system the set of profit-maximizing production plans. These individual responses are correspondences rather than functions because of the constancy of marginal rates of substitution in consumption and in production over a range of commodity bundles.

Keywords

Berge's maximum theorem; Brouwer's fixed point theorem; correspondences; functions; Kakutani's fixed point theorem; Lyapunov's theorem

Article

A correspondence Q from a domain set X to a range set Y associates with each element x in X, a non-empty subset of Y, $Q(x)$. A *function* is a correspondence such that $Q(x)$ is a singleton for each x in X. It is for this reason that a correspondence is also termed a *multi-valued function* or, more simply, a *multi-function*. Another name for a correspondence is a *set-valued mapping*.

Correspondences arise naturally in economic theory. One may think of an individual consumer's demand correspondence, which associates with each price system the set of utility maximizing consumption plans; see, for example, Hildenbrand (1974, p. 92). An equally pervasive example is an individual producer's supply correspondence which associates with each price system the set of profit-maximizing production plans (see, for example, Arrow and Hahn, 1971, pp. 54–5). The fact that these individual responses are correspondences rather than functions is simply a consequence of ‘flats’ in the underlying indifference surfaces and isoquants or, more precisely, of the constancy of marginal rates of substitution

in consumption and in production over a range of commodity bundles. Indeed, the association of these marginal rates with the point at which they are evaluated is another example of a correspondence that arises naturally in economic theory, particularly in the study of marginal cost pricing equilibria in economies with increasing returns to scale (for example, Brown et al., 1986). The fact that there is no unique rate of substitution is simply a consequence of ‘kinks’ in the underlying function. In the case of a convex function, such a correspondence is termed the *subdifferential* correspondence, and, for more general functions, it is *Clarke's generalized derivative*.

If the domain and range of a correspondence are *topological spaces*, one can formulate various notions of continuity of a correspondence. Recall that (X, τ_X) is a topological space if X is a set and τ_X is a collection of subsets of X that contains X and the empty set ø and is closed under finite intersection and arbitrary union. We can now present one formalization of the intuitive idea of continuity of a correspondence. A correspondence $Q: X \to Y$, X, Y, both topological spaces, is said to be upper *semicontinuous* (u.s.c.) if for any V in τ_Y, the set $\{x \in X: Q(x) \subset V\}$ is in τ_X. Q is said to be lower *semicontinuous* (l.s.c.) if for any V in τ_Y, the set $\{x \in X: Q(x) \cap V \neq \emptyset\}$ is in τ_X. It is easy to convince oneself that a correspondence may be u.s.c. without being l.s.c. and vice versa. It is also easy to show that, if Y is a compact space, a correspondence Q is u.s.c. if and only if its graph, $\mathrm{Gr}Q$, $\mathrm{Gr}Q = \{(x, y) \in X \times Y: y \in Q(x)\}$, is such that its complement belongs to $\tau_X \times \tau_Y$. A correspondence is said to be *continuous* if it is both u.s.c. and l.s.c.

A very useful result for establishing u.s.c. of correspondences arising from maximization is Berge's *maximum theorem*. This states, in particular, that for any continuous correspondence Q from a topological space X to a topological space Y and any continuous function f from $X \times Y$ into the reals, the associated correspondence $\mu: X \to Y$ given by $\mu(x) = \left\{y \in Q(x): f(y, x) \geq f(y', x) \text{ for all } y' \in Q(x)\right\}$ is u.s.c. This theorem is used to show u.s.c. of the demand and supply correspondences in the theory of the consumer and of the producer.

A result which plays a significant role in the proof of the existence of a competitive equilibrium is Kakutani's *fixed point theorem* for convex valued, u.s.c. correspondences which take a non-empty convex compact subset of an Euclidean space to itself. The theorem states that such correspondences Q have a fixed point, that is, an element x such that $x \in Q(x)$. Kakutani's theorem yields as an immediate corollary Brouwer's fixed point theorem and generalizes, word for word, to locally convex spaces as has been shown by Glicksberg and Ky Fan (see, for example, Berge, 1963, p. 251).

It is of interest to know of conditions under which a correspondence $Q: X \to Y$ yields a *continuous selection*, that is, a continuous function $f: X \to Y$ such that $f(x) \in Q(x)$ for all x in X. The celebrated selection theorems of Michael (see, for example, Bessaga and Pelczynski, 1975, ch. II.7) give a variety of sufficient conditions for this. One of these requires X to be a paracompact topological space, Y to be a separable Banach space and Q to be convex valued and l.s.c. This theorem has been used by Gale and Mas-Colell (1974) to show the existence of competitive equilibrium for economies in which consumer preferences need neither be complete nor transitive. If Q is u.s.c. rather than l.s.c., recent work of Cellina gives sufficient conditions under which one may obtain an *approximate continuous selection*.

So far in this exposition we have been considering results on correspondences whose domain and range are both topological spaces. An alternative setting is one where the range is a topological space but the domain is a *measurable space*. (T, Σ) is a measurable space if T is a set and Σ is a family of subsets

that includes *T* and is closed under complementation and countable unions, that is, Σ is a σ*-algebra*. Such correspondences arise naturally in the study of economies in which the set of agents is modelled as a measurable space. An obvious example of such a correspondence is one which associates with every agent his/her set of utility maximizing consumption plans under a given price system.

One can develop concepts analogous to continuity for correspondences from a measurable space to a topological space. A correspondence $Q: T \to Y$ is said to be *measurable* if, for any set *V* in τ_Y, the set $\{t \in T: Q(t) \cap V \neq \emptyset\}$ is an element of Σ. Variants of this definition have been presented in the literature along with conditions under which these variants are all equivalent. One particularly fruitful variant requires the measure space to be *complete* and the correspondence to have a *measurable graph*, that is, Gr*Q* is a subset of $\Sigma \otimes \mathcal{B}(Y)$, the smallest σ-algebra generated by the sets in $\Sigma \times \mathcal{B}(Y)$ and where $\mathcal{B}(Y)$ is the smallest σ-algebra generated by sets in τ_Y.

We can now state a measure-theoretic analogue of Berge's theorem. Let *Q* be a correspondence with a measurable graph and *f* a $\Sigma \otimes \mathcal{B}(Y)$ measurable function from $T \times Y$ into the reals. Then a result due to the collective efforts of Debreu and Castaing–Valadier states that under a mild restriction on *Y*, namely Souslin, the correspondence $\mu: T \to X, \mu(t) = \left\{x \in Q(t): f(t, x) \geq f(t, x') \text{ for all } x' \text{ in } Q(t)\right\}$, has a measurable graph.

We have developed enough terminology to state a fundamental theorem due to the collective efforts of von Neumann, Aumann and St. Beuve. This states that under a restriction on the range space *Y*, namely, Souslin, every correspondence *Q* with a measurable graph yields a *measurable selection*, that is, a measurable function $f: T \to Y$ such that $f(t) \in Q(t)$ for all *t* in *T*.

Once we have a measurable selection theorem, we are in a position to formulate a satisfactory notion of an integral of a correspondence, a notion which may also be seen as a formalization of a sum of an infinite number of sets. However, one preliminary notion that still needs to be stated is that of a *measure* μ on (T, Σ). A measure μ is a set-valued function from Σ into (say) Euclidean space R^n such that

$$\mu(A) \geq 0, \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

for all *A*, A_i in Σ and such that A_i are mutually disjoint. Now let us assume we know how to integrate a function with respect to μ and can therefore specify a function $f: T \to R^n$ to be an *integrable function* if its integral (Lebesgue integral) is finite. Following Aumann, we can define the integral of a correspondence *Q*, $\int_T Q(t) d\mu$ to be the set $\{\int_T f(t) dt: f$, an integrable function which is a measurable selection from *Q*$\}$. It is now clear that $\int_T Q(t) d\mu$ is non-empty if *Q* has a measurable graph and if there exists an integrable function *g* with non-negative values and such that $|x| \leq g(t)$ for all $x \in Q(t)$ and for all $t \in T$.

Finally, we can state a consequence of Lyapunov's theorem on the range of an *atomless measure* that has played a fundamental role in the development of the theory of economies with a continuum of agents. A measure μ on a measurable space (T, Σ) is atomless if (T, Σ, μ) has no *atoms*, that is $A \in \Sigma$ such that

$\mu(A) > 0$ and $B \in \Sigma$, $B \subset A$ implies $\mu(B) = \mu(A)$ or $\mu(B) = 0$. The Lyapunov–Richter theorem states that the integral of a correspondence $Q: T \to R^n$ is convex if μ is an atomless measure on (T, Σ).
In summary, a correspondence is a versatile mathematical object for which a deep and rich theory can be developed and which arises naturally in many diverse areas of applied mathematics, including economic theory. For an introduction to this theory and to its applications, the reader is referred to the following references which also contain all the concepts and results not referenced in this entry.

See Also

- fixed point theorems
- Lyapunov functions

Bibliography

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Aubin, J.P. and Cellina, A. 1984. *Differential Inclusions*. New York: Springer-Verlag.

Berge, C. 1963. *Topological Spaces*. New York: Macmillan.

Bessaga, C. and Pelczynski, A. 1975. *Selected Topics in Infinite-Dimensional Topology*. Warsaw: Polish Scientific Publishers.

Brown, D., Heal, G., Ali Khan, M. and Vohra, R. 1986. On a general existence theorem for marginal cost pricing equilibria. *Journal of Economic Theory* 38, 111–19.

Castaing, C. and Valadier, M. 1977. *Convex Analysis and Measurable Multifunctions*. Lecture Notes in Mathematics No. 580. New York: Springer-Verlag.

Clarke, F.H. 1983. *Optimization and Nonsmooth Analysis*. New York: John Wiley.

Gale, D. and Mas-Colell, A. 1974. An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2, 9–15. Erratum in *Journal of Mathematical Economics* 6, 297–8.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton, NJ: Princeton University Press.

Klein, E. and Thompson, E.A. 1985. *Introduction to the Theory of Correspondences*. New York: John Wiley.

Rockafellar, R.T. 1970. *Convex Analysis*. Princeton, NJ: Princeton University Press.

How to cite this article

Khan, M. Ali. "correspondences." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000384> doi:10.1057/9780230226203.0327

The New Palgrave Dictionary of Economics Online

cost functions

W.E. Diewert
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost functions; duality; Euler's Theorem; expenditure functions; generalized Leontief cost function; indirect utility function; normalized quadratic cost function; production functions; Shephard's duality theorem; Shephard's lemma; substitutes and complements; translog cost function; unit cost function

Article

1 Introduction

Cost and expenditure functions are widely used in both theoretical and applied economics. Cost functions are often used in econometric studies which describe the technology of firms or industries while their consumer theory counterparts, expenditure functions, are frequently used to describe the preferences of consumers.
Cost and expenditure functions also play an important role in many theoretical investigations. This is due to the fact that a cost function embodies the consequences of cost minimizing behaviour on the part of a consumer or producer and so it is not necessary to spell out the details of the primal minimization problem that defined the cost function. This may seem like a very minor advantage, but when one is dealing with, say, the comparative statics of a general equilibrium problem, the use of cost functions leads to the analysis of a much smaller system of equations and hence the structure of the problem can be more easily understood.
Sections 2–5 below develop the theoretical properties of cost functions while Sections 6–8 are devoted to empirical applications of cost functions in the producer and consumer contexts.

2 Properties of cost functions

One of the fundamental paradigms in economics is the one which has a producer competitively minimizing costs subject to his technological constraints. Competitive means that the producer takes input prices as fixed during the given period of time irrespective of the producer's demand for those inputs.
We assume that only one output can be produced using *N* inputs and that the producer's technology can be summarized by a *production function F*: *y*=*F*(*x*) where $y \geq 0$ is the maximal amount of output that can be produced during a period, given the non-negative vector of inputs $x \equiv (x_1, \ldots, x_N) \geq 0_N$. We further assume that the cost of purchasing one unit of input *i* is $p_i > 0, i = 1, \ldots, N$ and that the positive vector of input prices that the producer faces is $p \equiv (p_1, \ldots, p_N) \gg 0_N$.
For $y \geq 0,\ p \gg 0_N$, the producer's *cost function C* is defined as the solution to the following constrained minimization problem:

$$C(y, p) \equiv \min_x \left\{ p \cdot x : F(x) \geq y,\ x \geq 0_N \right\} \tag{1}$$

where $p \cdot x \equiv \Sigma_{n=1}^{N} p_n x_n$. Thus *C*(*y*, *p*) is the minimum input cost of producing at least the output level *y*, given that the producer faces the input price vector *p*.
The minimization problem (1) can also be given a consumer theory interpretation: let *F* be a consumer's preference or *utility function*, let *y* be a utility or welfare level, let *x* be a vector of commodity purchases (rentals in the case of consumer durables), and let *p* be a vector of

commodity (rental) prices. In this case, the consumer attempts to minimize the cost of achieving at least the target welfare level indexed by *y*, and the solution to (1) defines the consumer's *expenditure function*.

Unfortunately, the minimum (1) may not exist in general. However, if we impose the following very weak regularity condition on *F*, it can be shown that *C* will be well defined as a minimum: *Assumption* 1 *on F: F* is continuous from above.

Assumption 1 means that for every *y* in the range of *F*, the *upper level set* $L(y) \equiv \{x: F(x) \geq y, x \geq 0_N\}$ is a closed set. The assumption is a technical one of minimal economic interest. It is also a very weak condition from an empirical point of view, since it cannot be contradicted by a finite set of data on the inputs and output of a producer.

If we assume that the production function *F* satisfies Assumption 1, it turns out that the cost function *C* has the following properties: *Property 1: C* is a non-negative function; that is, $C(y, p) \geq 0$; *Property 2: C* is *linearly homogeneous* in input prices *p* for each fixed output level *y*; that is, $C(y^1, p) \geq C(y^2, p)$ for $y^1 \geq y^2 \geq 0$ and $p \gg 0_N$; *Property 3: C* is *nondecreasing* in *p* for fixed y; that is, $C(y, p^1) \geq C(y, p^2)$ for $y \geq 0$, $p^1 \geq p^2 \gg 0_N$; *Property 4: C* is *concave* in *p* for fixed *y*; that is, $C(y, \lambda p^1 + (1-\lambda) p^2) \geq \lambda C(y, p^1) + (1-\lambda) C(y, p^2)$ for $y \geq 0$, $0 \leq \lambda \leq 1$, $p^1 \gg 0_N$ and $p^2 \gg 0_N$; *Property 5:C* is *nondecreasing* in *y* for fixed *p*; that is, *Property 6: C* is *continuous from below* in *y* for fixed *p*; that is $\{y: C(y, p) \leq \alpha\}$ is a closed set for every α and $p \gg 0_N$.

Properties 1–4 for *C* were derived by Shephard (1953) under stronger regularity conditions on *F* and Properties 4, 5, and 6 were obtained by McKenzie (1957), Uzawa (1964) and Shephard (1970) respectively.

From the viewpoint of economies, all of the properties of *C* are intuitively obvious except Properties 4 and 6. Property 6 on *C* is the technical counterpart to Assumption 1 on *F* and is of minimal economic interest. However, Property 4 has some significant economic implications as we shall see in Section 5 below.

We can already draw some useful empirical implications from the fact that a cost function must satisfy Properties 1–6 above. For example, in industrial organization and applied econometrics, it is quite common to assume that the true functional form for a firm's or industry's cost function has the following functional form:

$$C(y, p) \equiv \alpha + \beta \cdot p + \gamma y \tag{2}$$

where $\alpha, \beta \equiv (\beta_1, \ldots, \beta_N)$ and γ are unknown parameters. However, Property 2 implies that α and γ must be zero in order for the cost function to be linearly homogeneous in input prices. But then $C(y, p) = \beta \cdot p$ does not depend on the output level *y*, which is very implausible.

3 Duality between cost and production functions

It is easy to see that the family of upper level sets, $L(y) \equiv \{x: F(x) \geq y, x \geq 0_N\}$, completely determines the production function *F*. Furthermore, the cost function *C* may be defined in terms of the production function by (1) or equivalently, in terms of the family of upper level sets as follows:

$$C(y, p) = \min_x \left\{ p \cdot x : x \text{ belongs to } L(y) \right\}. \tag{3}$$

Thus given the production function *F* or the family of level sets *L*(*y*), the cost function *C* is determined.

We now ask the following question: given a cost function *C* which has Properties 1 to 6, can we use *C* to define the underlying production function *F*?

For a given output level *y* and input price vector $p \gg 0_N$, define the corresponding isocost plane by $\{x: p \cdot x = C(y, p)\}$. From the definitions of *C*(*y*, *p*) and *L*(*y*), it is obvious that the set *L*(*y*) must lie above this isocost plane and be tangent to it; that is, *L*(*y*) must be a subset of the set $\{x: p \cdot x \geq C(y, p)\}$ and this conclusion must be true for every positive input price vector *p*. Thus *L*(*y*) must be a subset of the intersection of all these sets which we denote by *M*(*y*):

$$M(y) \equiv \bigcap_{p \gg 0}^{N} \{x: p \cdot x \geq C(y, p)\}. \tag{4}$$

The set $M(y)$ is called the *disposal, convex hull of* $L(y)$; see McFadden (1966).
Each set $\{x: p \cdot x \geq C(y, p)\}$ is a halfspace and is a convex set. A set S *is convex* if and only if x^1 and x^2 belong to S and $0 \leq \lambda \leq 1$ implies $\lambda x^1 + (1 - \lambda) x^2$ also belongs to S. Since $M(y)$ is the intersection of a family of convex sets, $M(y)$ is also a convex set. $M(y)$ also has the following *free disposal* property:

$$x^1 \text{ belongs to } M(y),\ x^1 \leq x^2, \text{ then } x^2 \text{ belongs to } M(y). \tag{5}$$

We know $L(y)$ must be a subset of $M(y)$. If we want $L(y)$ to coincide with $M(y)$, then $L(y)$ must also be a convex set with the free disposal property. It can be shown that $L(y)$ will have these last two properties for every output level y if and only if the production function F has the following two properties: *Assumption 2 on F: F* is *quasiconcave* function: that is, for every y belonging to the range of F, $L(y) \equiv \{x: F(x) \geq y\}$ is a convex set. *Assumption 3 on F: F* is *nondecreasing*; that is, if $x^2 \geq x^1 \geq 0_N$, then $F(x^2) \geq F(x^1)$.
We may now answer our earlier question about whether a cost function C can completely determine the production function F: the answer is yes if the production satisfies Assumptions 1–3.
More precisely, we have the following result: given a cost function C which satisfies Properties 1–6, then the production function F defined by

$$F(x) \equiv \max_y \left\{ y: C(y, p) \leq p \cdot x \text{ for every } p \gg 0_N \right\},\ x \geq 0_N \tag{6}$$

satisfies Assumptions 1–3. Moreover, if we define the cost function C^* which corresponds to the F defined by (6) in the usual way [recall (1)], then $C^* = C$; that is, this derived cost function C^* coincides with the original cost function C. Thus there is a *duality* between production functions F satisfying Assumptions 1–3 and cost functions C having Properties 1–6: each function completely determines the other under these regularity conditions.
Duality theorems similar to the above results have been established under various regularity conditions by Shephard (1953; 1970), Uzawa (1964), McFadden (1966; 1978a) and Diewert (1971; 1982).

4 The derivative property of the cost function

The following result is the basis for most of the theoretical and empirical applications of cost functions.
Suppose the cost function C satisfies Properties 1–6 listed in Section 2 and in addition, C is differentiable with respect to the components of p at the point (y^*, p^*). Then the solution $x^* \equiv (x_1^*, \ldots, x_N^*)$ to the cost minimization problem $\min_y \left\{ p^* \cdot x: F(x) \geq y^*,\ x \geq 0_N \right\}$ is unique and

$$x_i^* = \partial C(y^*, p^*) / \partial p_i,\ i = 1, \ldots, N; \tag{7}$$

that is, the cost minimizing demand for the ith input is equal to the partial derivative of the cost function with respect to the ith input price.
The result (7) is known as the derivative property of the cost function (see McFadden, 1978a) or Shephard's Lemma, since Shephard (1953)

was the first to obtain the result. It should be noted that Hicks (1946) and Samuelson (1947) obtained the result (7) earlier, but under different hypotheses: they assumed the existence of a utility or production function *F* and deduced (7) by analysing the comparative statics properties of the cost minimization problem (1). On the other hand, Shephard (1953; 1970) assumed only the existence of a cost function satisfying the appropriate regularity conditions.

A very elegant proof of (7) using the hypotheses of Hicks and Samuelson is due to Karlin (1959) and Gorman (1976). Their proof proceeds as follows.

Let x^* be a solution to $\min_x \{p^* \cdot x : F(x) \geq y^*, x \geq 0_N\} = C(y^*, p^*)$. Then for every $p \gg 0_N$, x^* is feasible for the cost minimization problem defined by $C(y^*, p) = \min_x \{p \cdot x : F(x) \geq y^*, x \geq 0_N\}$ but it is not necessarily optimal. Thus for every $p \gg 0_N$, we have the following inequality:

$$p \cdot x^* \geq C(y^*, p). \tag{8}$$

We also have

$$p^* \cdot x^* = C(y^*, p^*). \tag{9}$$

For $p \gg 0_N$, define the function $g(p) \equiv p \cdot x^* - C(y^*, p)$. From (8), $g(p) \geq 0$ for all $p \gg 0_N$, and from (9), $g(p^*) = 0$. Thus $g(p)$ attains a global minimum at $p=p^*$. Since g is differentiable, the first-order necessary conditions for a minimum must be satisfied at p^*:

$$\nabla_p g(p^*) = x^* - \nabla_p C(y^*, p^*) = 0_N \tag{10}$$

where $\nabla_p g(p^*) \equiv [\partial g(p^*) / \partial p_1, \ldots, \partial g(p^*) / \partial p_N]$ denotes the vector of first-order partial derivatives of g with respect to the components of p evaluated at p^* and $\nabla_p C(y^*, p^*)$ denotes the vector of first-order partial derivatives of C with respect to the components of p evaluated at (y^*, p^*). The second set of equalities in (10) can be rearranged to yield (7).

From an econometric point of view, Shephard's Lemma is a very useful result. In order to obtain a valid system of cost minimizing input demand functions, $x(y, p) \equiv [x_1(y, p), \ldots, x_N(y, p)]$ all we have to do is postulate a functional form for C which satisfies Properties 1–6 and then differentiate C with respect to the components of the input price vector p; that is, $x(y, p) = \nabla_p C(y, p)$. It is not necessary to compute the production function F that corresponds to C via the Shephard Duality Theorem nor is it necessary to undertake the often complex algebra involved in deriving the input demand functions using the production function and Lagrangian techniques. In Section 6 below, we shall consider several functional forms for C that have been suggested for their econometric convenience.

5 The comparative statics properties of cost functions

Suppose that we are given a cost function C satisfying Properties 1–6 that is also twice continuously differentiable at (y^*, p^*) where $y^* > 0$ and $p^* \gg 0_N$. Applying Shephard's Lemma (7), the above differentiability assumption ensures that the cost minimizing input demand functions $x_i(y, p)$ exist and are once continuously differentiable at (y^*, p^*).

Define $[\partial x_i / \partial p_j] \equiv [\partial x_i(y^*, p^*) / \partial p_j]$ to be the N by N matrix of partial derivatives of the N demand functions $x_i(y^*, p^*)$ with respect to the N prices p_j, $i, j = 1, \ldots, N$. From (7), it follows that

$$[\partial x_i / \partial p_j] = \left[\partial^2 C(y^*, p^*) / \partial p_i \partial p_j\right] \equiv \nabla^2_{pp} C(y^*, p^*)$$
(11)

where $\nabla^2_{pp} C(y^*, p^*)$ is the matrix of second-order partial derivatives of the cost function with respect to the components of the input price vector evaluated at (y^*, p^*). The twice continuous differentiability property of C implies by Young's Theorem in calculus that $\nabla^2_{pp} C(y^*, p^*)$ is a symmetric N by N matrix. Thus using (11), we have

$$[\partial x_i / \partial p_j] = [\partial x_i / \partial p_j]^{\mathrm{T}} = [\partial x_j / \partial p_i]$$
(12)

where A^{T} denotes the transpose of the Matrix A. Thus we have established the Hicks (1946) and Samuelson (1947) *symmetry restrictions* on input-demand functions, $\partial x_i(y^*, p^*) / \partial p_j = \partial x_j(y^*, p^*) / \partial p_i$ for all i and j.
Since C is concave in p and is twice continuously differentiable with respect to the components of p at the point (y^*, p^*), it follows from a characterization of concave functions that $\nabla^2_{pp} C(y^*, p^*)$ is a negative semidefinite matrix. Thus by (11),

$$z^{\mathrm{T}}[\partial x_i / \partial p_j] z \leq 0 \text{for all vectors } z.$$
(13)

In particular, letting $z = e_i$, the ith unit vector, (13) implies

$$\partial x_i(y^*, p^*) / \partial p_i \leq 0 \text{for} i = 1, \ldots, N;$$
(14)

that is, the ith cost minimizing input demand function cannot slope upwards with respect to the ith input price for $i = 1, \ldots, N$.
Since C is linearly homogeneous in p, we have $C(y^*, \lambda p^*) = \lambda C(y^*, p^*)$ for all $\lambda > 0$. Partially differentiating this equation with respect to p_i for λ close to 1 yields the equation $C_i(y^*, \lambda p^*)\lambda = \lambda C_i(y^*, p^*)$ where $C_i(y^*, p^*) \equiv \partial C(y^*, p^*) / \partial p_i$. Thus $C_i(y^*, \lambda p^*) = C_i(y^*, p^*)$ and differentiation of this last equation with respect to λ yields when $\lambda = 1$:

$$\sum_{j=1}^{N} p_j^* \partial^2 C(y^*, p^*) / \partial p_i \partial p_j = 0 \text{for} i = 1, \ldots, N.$$
(15)

Equations (11) and (15) imply that the input-demand functions $x_i(y^*, p^*)$ satisfy the following N restrictions:

$$\sum_{j=1}^{N} p_j^* \partial x_i(y^*, p^*) / \partial p_j = 0 \text{for} i = 1, \ldots, N.$$

(16)

A final general restriction on the derivatives of the input-demand functions may be obtained as follows: for λ near 1 differentiate both sides of $C(y^*, \lambda p^*) = \lambda C(y^*, p^*)$ with respect to *y* and then differentiate the resulting equation with respect to λ. When λ=1, the last equation becomes:

$$\sum_{j=1}^{N} p_j^* \partial^2 C(y^*, p^*) / \partial y \partial p_j = \partial C(y^*, p^*) / \partial y.$$

(17)

The twice continuous differentiability of *C* at (y^*, p^*) and (7) imply:

$$\partial^2 C(y^*, p^*) / \partial y \partial p_j = \partial^2 C(y^*, p^*) / \partial p_j \partial y = \partial x_j(y^*, p^*) / \partial y.$$

(18)

Property 5 for cost functions implies that

$$\partial C(y^*, p^*) / \partial y^* \geq 0.$$

(19)

Using (18) and (19), (17) is equivalent to:

$$\sum_{j=1}^{N} p_j^* \partial x_j(y^*, p^*) / \partial y \geq 0 .$$

(20)

Thus for at least one *j*, we must have $\partial x_j(y^*, p^*) / \partial y \geq 0$; that is, as output increases, not every input demand can decrease.
We have shown that the assumption of cost minimizing behaviour implies a number of restrictions on input demand functions that are potentially testable. Hicks (1946) and Samuelson (1947) obtained the restrictions (12), (13), and (16) using the first-order conditions for the primal cost minimization problem (1) and the properties of determinants of bordered Hessian matrices; Samuelson also obtained (20). Our derivation of the restrictions on input-demand functions using the dual approach is due to McKenzie (1957), Karlin (1959) and McFadden (1978a).
Hicks (1946) also showed that when *N*=2, so that there are only two inputs, then (12), (13), and (16) imply that

$$\partial x_1(y^*, p_1^*, p_2^*) / \partial p_2 = \partial x_2(y^*, p_1^*, p_2^*) / \partial p_1 \geq 0.$$

(21)

Hicks (1946) called two distinct goods *i* and *j* *substitutes* if and only if $\partial x_i(y, p) / \partial p_j \geq 0$ and *complements* if and only if $\partial x_i(y, p) / \partial p_j < 0$. Thus in the two input case, the two goods must be substitutes. Hicks also showed that in the three input case, at least two of the three pairs of goods must be substitutes.
We turn now to empirical applications of cost functions.

6 Functional forms for cost functions

Shephard's Lemma (7) provides a convenient method for generating systems of cost minimizing input demand functions: simply postulate a functional form for $C(y, p)$ and then partially differentiate C with respect to each input price. Below, we present three examples to illustrate the technique.
Our first example is the *translog cost function* due to Christensen, Jorgenson and Lau (1971; 1973). The logarithm of the cost function is defined as follows:

$$\ln C(y, p) \equiv \alpha_0 + \sum_{i=1}^{N} a_i \ln p_i = (1/2) \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} \ln p_i \ln p_j + \sum_{i=1}^{N} a_{iy} \ln p_i \ln y + a_y \ln y + (1/2) a_{yy} \ln y \ln y \tag{22}$$

where the $a_i, a_{ij}, = a_{ji}, a_{iy}, a_y$ and a_{yy} are $1 + N + (1/2)N(N+1) + N + 2 = 3 + 2N + (1/2)N(N+1)$ parameters determined by the technology of the firm or industry. Differentiating both sides of (22) with respect to the logarithm of the *i*th input price, In p_i, for $i = 1, \ldots, N$ yields the following system of equations:

$$s_i = a_i + \sum_{j=1}^{N} a_{ij} \ln p_j + a_{iy} \ln y, \; i = 1, \ldots, N \tag{23}$$

where the *i*th input cost share is defined as $s_i \equiv [p_i \partial C(y, p) / \partial p_i] / C(y, p) = p_i x_i(y, p) / C(y, p)$ where the last equality follows using (7).
By Property 2 for cost functions, $C(y, p)$ must be linearly homogeneous in input prices. This property will be satisfied by the translog cost function cost function if and only if the following N+2 linear restrictions on the parameters hold:

$$\sum_{i=1}^{N} a_i = 1, \; \sum_{i=1}^{N} a_{iy} = 0 \text{ and } \sum_{j=1}^{N} a_{ij} = 0 \text{ for } i = 1, \ldots, N. \tag{24}$$

It is possible to append errors to equations (22) and N–1 of the equations (23) and econometrically estimate the unknown parameters, given data on inputs, input prices and output. The symmetry restrictions $a_{ij} = a_{ji}$ and the restrictions (24) may be imposed or one can test for their validity. If these restrictions are imposed, then the resulting translog cost function will have $1 + N + (1/2)N(N+1)$ free parameters.
What considerations are relevant in choosing a functional form for a cost function? The following four properties are desirable: (1) *flexibility*; that is, the functional form for C should have a sufficient number of free parameters to be able to provide a second-order approximation to an arbitrary twice continuously differentiable function with the appropriate theoretical properties, (ii) *parsimony*; that is, the functional form for C should have the minimal number of free parameters required to have the flexibility property, (iii) *linearity*; that is, the unknown parameters of C should appear in the system of estimating equations in a linear fashion in order to facilitate econometric estimation, and (iv) *consistency*; that is, the functional form for C should be consistent with Properties 1–6 for cost functions. These considerations were first suggested by Diewert (1971) in an informal manner; the term parsimony is due to Fuss, McFadden and Mundlak (1978) and the term flexible is due to Diewert (1974). The equivalence of various definitions of the flexibility property is discussed by

Barnett (1983).

How satisfactory is the translog cost function in the light of the above considerations? We consider the flexibility property first. In order to be able to approximate a function of 1+*N* variables to the second order, we require $1 + (1 + N) + (1 + N)^2$ free parameters. However, if we assume that the cost functions are twice continuously differentiable, then we can reduce the number by $N(N + 1) / 2$ due to the symmetry property of the second-order partial derivatives. The linear homogeneity property of the cost function, Property 2, yields an additional *N*+1 restriction on the first and second derivatives of *C*, (15) and (17), plus the following restriction (which follows using Euler's Theorem on homogeneous functions):

$$C(y, p) = \sum_{i=1}^{N} p_i \partial C(y, p) / \partial p_i. \tag{25}$$

Thus a flexible functional form for a cost function should have $1 + (1 + N) + (1 + N)^2 - [(1 / 2)N(N + 1) + N + 1 + 1] = 1 + N + (1 / 2)N(N + 1)$ free parameters, which is precisely the number the translog cost function has when the restrictions (24) are imposed. It can be shown that the translog cost function is indeed flexible and we have just shown that it is also parsimonious.

As can be seen by inspecting (22) and (23), the estimating equations are linear in the unknown parameters, so the linearity property is also satisfied.

If the restrictions (24) are imposed, Property 2 will be satisfied. In practice, the other properties that a cost function must have will be satisfied with the exception of Property 4, the concavity in prices property. If all of the a_{ij} and a_{iy} parameters are zero, then the translog cost function reduces to a Cobb–Douglas cost function which satisfies the concavity property globally. However, in the general case, the best we can hope for is that the concavity property is satisfied locally for a range of input prices.

If a production function is linearly homogeneous (that is, $F(\lambda x) = \lambda F(x)$ for $\lambda \geq 0$ and $x \geq 0_N$) so that the technology is subject to constant returns to scale, then the corresponding cost function has the following property:

$$C(y, p) = yC(1, p); \tag{26}$$

that is, total cost is equal to the output level *y* times the cost of producing one unit of output, $C(1, p) \equiv c(p)$, the *unit cost function*.

If *C* is twice continuously differentiable and satisfies (26), then one can show that the following 2+*N* restrictions on the first and second derivatives of *C* must hold:

$$C(y, p) = y\partial C(y, p) \partial y, \tag{27}$$

$$\partial^2 C(y, p) / \partial y^2 = 0; \tag{28}$$

$$\partial C(y, p) / \partial p_i = y\partial^2 C(y, p) / \partial y \partial p_i, \; i = 1, \ldots, N.$$

(29)

However, in view of (25), it can be seen that only N–1 of the restrictions (29) are new. Thus the assumption of a constant returns to scale technology imposes new restrictions on the derivatives of the cost function C.
It can be shown that necessary and sufficient conditions for the translog cost function defined by (22) and (24) to satisfy (26) are the following N+1 restrictions:

$$a_y = 1,\ a_{yy} = 1 \text{ and } a_{iy} = 0 \text{ for } i = 1, \ldots, N-1. \tag{30}$$

Of course (30) and (24) imply that a_{Ny}=0 as well.
It can be shown that if the restrictions (24) and (30) are imposed on the parameters of the translog cost function defined by (22), then the resulting functional form is flexible in the class of cost functions that satisfy the constant returns to scale property (26). Note that we can test for the validity of the constant returns to scale property by testing whether the N+1 linear restrictions (30) hold.
For our second example, consider the following functional form for a cost function:

$$C(y, p) \equiv c(p)y + \sum_{i=1}^{N} b_i p_i + b_{yy}\left(\sum_{i=1}^{N} \beta_i p_i\right)y^2; \tag{31}$$

$$c(p) \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij} p_i^{1/2} p_j^{1/2} \tag{32}$$

where the $b_{yy}, b_i, b_{ij} = b_{ji}$ and β_i are parameters which characterize the technology. If $b_i = 0$ for $i = 1, \ldots, N$ and $b_{yy} = 0$, then (31) reduces to the *Generalized Leontief cost function* defined by Diewert (1971). If in addition, $b_{ij} = 0$ for all $i \neq j$, then (31) reduces to the cost function $\Sigma_{i=1}^{N} b_{ii} p_i y$, which is dual to the Leontief (no substitution) production function, $F(x_1, \ldots, x_N) \equiv \min\{x_i / b_{ii}: i = 1, \ldots, N\}$.
In order for the cost function defined by (31) and (32) to satisfy the parsimony property, it is necessary for the empirical investigator to prespecify the β_i parameters; for example, one could set β_i equal to 1 or to the average input quantity x_i observed in the sample of data. Under these conditions, the Generalized Leontief cost function has $(1/2)N(N+1) + N + 1$ free parameters, which is just the required number for the flexibility property. In fact, Diewert and Wales (1987) show that this cost function is flexible and parsimonious when the β_i are predetermined.
Applying (7), the input-demand functions that correspond to (31) and (32) are:

$$x_i(y, p) = \sum_{j=1}^{N} b_{ij} p_i^{-1/2} p_j^{1/2} y + b_i + b_{yy}\beta_i y^2 .\ i = 1, \ldots, N. \tag{33}$$

For the purpose of econometric estimation, errors can be appended to the N equations (33). If the β_i are predetermined, it can be seen that the system of estimating equations is linear in the unknown parameters.

If we wish to test for a constant returns to scale technology, then the following 1+N linear restrictions on the parameters are necessary and sufficient for this property:

$$b_{yy} = 0 \text{ and } b_i = 0 \text{ for } i = 1, \ldots, N. \tag{34}$$

Note that the linear homogeneity in prices property is satisfied by the Generalized Leontief cost function. The other properties for cost functions will also be satisfied in practice with the exception of Property 4, the concavity in prices property. If all $b_{ij} \geq 0 \text{ for } i \neq j$, then the concavity property will be globally satisfied, but this assumption rules out complementary pairs of inputs (recall the discussion about substitutes and complements at the end of the previous section). Thus in general, one can only hope that the concavity property will be satisfied locally, as was the case with the translog cost function.

For our third and final example, consider the following *normalized quadratic cost function* defined by (31) but now $c(p)$ is defined as follows:

$$c(p) \equiv \sum_{i=1}^{N} b_{ii} p_i + (1/2) \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} p_i p_j \Big/ \left(\sum_{n=1}^{N} \alpha_n p_n \right) \tag{35}$$

where the N by N matrix $A \equiv [a_{ij}]$ is symmetric and satisfies the following restriction for some input price vector $p^* \gg 0_N$;

$$A p^* = 0_N. \tag{36}$$

This functional form is due to Diewert and Wales (1987); it generalizes some functional forms due to Fuss (1977) and McFadden (1978b). The functional form has N b_{ii} parameters, $(1/2)N(N-1)$ free a_{ij} parameters taking into consideration (36), N b_i parameters, 1 b_{yy}, N β_i and N α_n parameters or $1 + 3N + (1/2)N(N+1)$ free parameters in all. In order for this functional form to have the parsimony property, it is necessary for the empirical investigator to prespecify the β_i and α_n parameters; we assume that this has been done and these parameters are non-negative and not identically equal to zero. Under these conditions, Diewert and Wales (1987) show that this cost function is parsimonious and flexible at the point (y^*, p^*) where p^* is the price vector which appears in (36).

Applying (7), the system of input demand functions divided by the output y is:

$$x_i(y, p) / y = b_{ii} + \sum_{j=1}^{N} a_{ij} p_j \left(\sum_{n=1}^{N} \alpha_n p_n \right)^{-1} - \left(\sum_{j=1}^{N} \sum_{k=1}^{N} a_{jk} p_j p_k \right) \times \left(\sum_{n=1}^{N} \alpha_n p_n \right)^{-2} \alpha_i + b_i y^{-1} + b_{yy} \beta_i y, \; i = 1, \ldots, N. \tag{37}$$

Errors can be appended to (37) and we obtain a system of estimating equations which is linear in the unknown parameters, provided that the α_n and β_i are prespecified.

If we wish to test for a constant returns to scale technology, then again the 1+N linear restrictions (34) are necessary and sufficient for this property.

The normalized quadratic cost function with prespecified α_n and β_i is flexible, parsimonious and has linear estimating equations. As was

the case with our first two examples, our third example has no problem in satisfying Properties 1, 2, 3, 5 and 6 for cost functions. It also turns out that our third example has no problem in satisfying Property 4: Diewert and Wales (1986) using some results due to Lau (1978), show that the normalized quadratic cost function is globally concave if and only if the *A* matrix is negative semidefinite. They also indicate how this negative semidefiniteness property can be imposed if necessary without destroying the flexibility of the functional form; simply set $A = -SS^{T}$ where *S* is a lower triangular *N* by *N* matrix and S^{T} is its transpose. However, in this latter case, nonlinear regression techniques must be used in order to estimate the unknown parameters.
The extensive empirical literature on estimating cost functions is nicely reviewed by Jorgenson (1984).

7 Applications to the estimation of consumer preferences

The cost function techniques described in the previous section can be used to obtain empirical descriptions of technologies. Those techniques can also be adapted to obtain empirical descriptions of consumer preferences.
As was noted in Section 2, *y* may be interpreted as a household's welfare level, *F* as a utility or preference function, *p* as a vector of commodity prices and *C*(*y*, *p*) as the minimum cost of achieving at least the welfare level *y*.
However, the econometric techniques described in the previous section cannot be utilized immediately in the consumer context because utility cannot be observed whereas output can. We acknowledge this difference by using *u*, the consumer's utility or welfare level, in place of *y* in what follows.
The theory outlined in the previous sections is still valid: given a differentiable functional form for the cost function *C*(*u*, *p*) that satisfies Properties 1 to 6, we may form the consumer's system of *constant real income or Hicksian demand functions* $x(u, p) \equiv [x_1(u, p), \ldots, x_N(u, p)]$ by differentiating the cost function with respect to each commodity price p_i [recall (7)]:

$$x_i(u, p) = \partial C(u, p) / \partial p_i, \; i = 1, \ldots, N. \tag{38}$$

We determine *u* as a function of the prices *p* and the consumer's observed expenditure on commodities during the period *Y*, say, by equating the minimum cost of achieving the welfare level *u* to the observed expenditure; that is, we solve the following equation for *u*:

$$C(u, p) = Y. \tag{39}$$

The solution function *g* where *u*=*g*(*Y*, *p*) is known as the consumer's *indirect utility function*. Now replace *u* in the right-hand side of (38) by *g* (*Y*, *p*) and obtain the consumer's system of *market demand functions*:

$$x_i = \partial C(g(Y, p), p) / \partial p_i, \; i = 1, \ldots, N. \tag{40}$$

If we multiply equation *i* in (40) by p_i, sum the resulting equations and use (7), (25) and (39), then we obtain the identity $\Sigma_{i=1}^{N} p_i x_i = Y$, so only *N*–1 of the *N* equations in (40) are independent. Thus for econometric estimation purposes, we may add errors to *N*–1 of the equations in (40), and given a functional form for *C*, we may use these equations to estimate the unknown parameters in *C*. We shall discuss this technique in more detail shortly, but first, we must discuss the problems involved in cardinalizing utility.
The scaling of utility is irrelevant in describing a consumer's preferences. However, when we postulate a functional form for a cost function, we are implicitly imposing a cardinalization of the consumer's utility. Hence, we might as well impose a convenient cardinalization: *money metric scaling of utility* (the term is due to Samuelson, 1974). This involves setting utility *u* equal to ‘income’ *Y*, holding prices constant at some specified price vector p^*, that is, we have

$$Y = g(Y, p^*) \text{for all } Y > 0.$$

(41)

In terms of the cost function, (41) may be rewritten as

$$u = C(u,\ p^{*}) \text{for all } u > 0.$$
(42)

In examples 2 and 3 in the previous section, the cost function had the following form:

$$C(u,\ p) = c(p)u + \sum_{i=1}^{N} b_i p_i + b_{yy}\left(\sum_{i=1}^{N} \beta_i p_i\right)u^2 .$$
(43)

In order to make (43) consistent with money metric scaling, (42), the following three restrictions on the parameters of *C* must be satisfied:

$$c(p^{*}) = 1,\ \sum_{i=1}^{N} b_i p_i^{*} = 0 \text{ and } b_{yy} = 0.$$
(44)

Using b_{yy}=0 we find that the indirect utility function that corresponds to the *C* defined by (43) is

$$g(Y,\ p) = \left(Y - \sum_{i=1}^{N} b_i p_i\right) / c(p) .$$
(45)

Substitution of (45) into (40) yields the following system of consumer demand functions:

$$x_i = b_i + [\partial c(p) / \partial p_i]\left(Y - \sum_{i=1}^{N} b_i p_i\right) / c(p) .\ i = 1, \ldots, N.$$
(46)

Now add errors to *N*–1 of the equations (46), calculate the partial derivatives of the *c*(*p*) defined by (32) or (35), impose the normalizations (44) and we have a system of nonlinear estimating equations. An empirical example of this technique for estimating consumer preferences may be found in Diewert and Wales (1986).

Finally, we note that cost functions of the type defined by (43) with b_{yy}=0 have very convenient aggregation over consumers' properties; see Gorman (1953) and Deaton and Muellbauer (1980).

8 Cost functions and measures of welfare gain

Consider a consumer whose preferences can be represented by the differentiable cost function, $C(u, p)$. Suppose we can observe the consumer's choices x^1 and x^2 during periods 1 and 2 when prices p^1 and p^2 prevail. Let u^1 and u^2 be the welfare levels attained during those two periods. Then by (7),

$$x^i = \nabla_p C(u^i, p^i), i = 1, 2. \tag{47}$$

For many purposes in applied welfare economics, it is useful to evaluate the *ex post* welfare change of the consumer. Two natural measures, suggested originally by Hicks (1942), are his *equivalent* and *compensating variations* which we denote by $V(p^1)$ and $V(p^2)$:

$$V(p^1) \equiv C(u^2, p^1) - C(u^1, p^1); V(p^2) \equiv C(u^2, p^2) - C(u^1, p^2). \tag{48}$$

From (47) and (25), $C(u^1, p^1) = p^1 \cdot x^1$ and $C(u^2, p^2) = p^2 \cdot x^2$. However, the costs $C(u^2, p^1)$ and $C(u^1, p^2)$ are not observable. Hence the following question arises: can we form approximations to $V(p^i)$ that use only observable data?

Linear approximations to $C(u^i, p^j)$ may be obtained using Taylor's Theorem. Thus we have:

$$V\left(p^1\right) \cong \left[C\left(u^2, p^2\right) + \nabla_p C\left(u^2, p^2\right) \cdot \left(p^1 - p^2\right)\right] - C\left(u^1, p^1\right) = [p^2 \cdot x^2 + x^2 \cdot (p^1 - p^2)] - p^1 \cdot x^1 \text{ using } (47) = p^1 \cdot \left(x^2 - x^1\right) \tag{49}$$

and

$$V\left(p^2\right) \cong C\left(u^2, p^2\right) - \left[C\left(u^1, p^1\right) + \nabla_p C\left(u^1, p^1\right) \cdot \left(p^2 - p^1\right)\right] = p^2 \cdot x^2 - \left[p^1 \cdot x^1 + x^1 \cdot \left(p^2 - p^1\right)\right] \text{ using } (47) = p^2 \cdot \left(x^2 - x^1\right). \tag{50}$$

The first-order approximations (49) and (50) are essentially due to Hicks (1942; 1946).

To obtain a second-order approximation result, we proceed indirectly. Suppose the consumer's cost function is defined by

$$C(u, p) \equiv c(p) + \sum_{i=1}^{N} b_i p_i u \tag{51}$$

where $c(p)$ is the normalized quadrative unit cost function defined by (35) for some prespecified $\alpha \equiv (\alpha_1, \ldots, \alpha_n) > 0_N$.

It can be shown that the cost function defined by (51) can provide a second-order approximation to an arbitrary twice continuously differentiable cost function that satisfies the money metric scaling of utility property (42).

Now use the parameters vector α which occurred in the definition of $c(p)$ in order to define the *normalized prices* v^i

$$v^i \equiv p^i / (p^i \cdot \alpha), i = 1, 2.$$

(52)

Straightforward calculations show that if C is defined by (51), then the following identity holds *exactly*:

$$(1/2)V\left(v^1\right) + (1/2)V\left(v^2\right) = (1/2)\left(v^1 + v^2\right) \cdot \left(x^2 - x^1\right)$$
(53)

where $V(v^1)$ and $V(v^2)$ are equivalent and compensating variations evaluated using the normalized prices v^i in place of the commodity price vectors p^i. Thus (53) says that an average of the Hicksian variations using normalized prices is exactly equal to the average of the normalized prices inner producted with the vector of quantity differences, $x^2–x^1$, provided that preferences are defined by the cost function (51). Note that the right-hand side of (53) can be evaluated using observable price and quantity data. Since the formula on the right-hand side is exact for preferences which have a second-order approximation property, we could call it a *superlative welfare gain measure* in analogy to the terminology used in index number theory. The term gain measure is due to King (1983).

See Also

- duality
- production functions

Bibliography

Barnett, W.A. 1983. New indices of money supply and the flexible Laurent demand system. *Journal of Business and Economic Statistics* 1, 7–23.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1971. Conjugate duality and the transcendental logarithmic production function. *Econometrica* 39, 255–6.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55, 28–45.

Deaton, A. and Muellbauer, J. 1980. *Economics and Consumer Behaviour*. Cambridge: Cambridge University Press.

Diewert, W.E. 1971. An application of the Shephard duality theorem: a generalized Leontief production function. *Journal of Political Economy* 79, 481–507.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of Quantitative Economics*, vol. 2, ed. M.D. Intriligator and D.A. Kendrick. Amsterdam: North-Holland, 106–171.

Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of Mathematical Economics*, vol. 2, ed. K.J. Arrow and M. D. Intriligator. Amsterdam: North-Holland.

Diewert, W.E. and Wales, T.J. 1986. Normalized quadratic systems of consumer demand functions. Discussion Paper No. 86–16, Department of Economics, University of British Columbia, Vancouver, Canada, May.

Diewert, W.E. and Wales, T.J. 1987. Flexible functional forms and global curvature conditions. *Econometrica*. 55, 43–68.

Fuss, M.A. 1977. Dynamic factor demand systems with explicit costs of adjustment. In *Dynamic Models of the Industrial Demand for Energy*, ed. E.R. Berndt, M. Fuss and L. Waverman. Palo Alto, CA: Electric Power Research Institute.

Fuss, M., McFadden, D. and Mundlak, Y. 1978. A survey of functional forms in the economic analysis of production. In *Production Economics: A Dual Approach to Theory and Applications*, vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21, 63–80.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in Economic Analysis*, ed. M. Artis and R. Nobay. Cambridge: Cambridge University Press.

Hicks, J.R. 1942. Consumers' surplus and index-numbers. *Review of Economic Studies* 9, 126–37.

Hicks, J.R. 1946. *Value and Capital*. 2nd edn. Oxford: Clarendon Press.

Jorgenson, D.W. 1984. Econometric methods for modeling producer behaviour. Discussion Paper No. 1086, Harvard Institute for Economic Research, Harvard University, Cambridge, MA. In *Handbook of Econometrics*, vol. 3, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Karlin, S. 1959. *Mathematical Methods and Theory in Games, Programming and Economics*, vol. 1. Palo Alto, CA: Addison-Wesley.

King, M.A. 1983. Welfare analysis of tax returns using household data. *Journal of Public Economics* 21, 183–214.

Lau, L.J. 1978. Testing and imposing monotonicity, convexity and quasi-convexity constraints. In *Production Economics: A Dual Approach to Theory and Applications*, vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

McFadden, D. 1966. Cost, revenue and profit functions: a cursory review. Working Paper No. 86, IBER, University of California at Berkeley, March.

McFadden, D. 1978a. Cost, revenue and profit functions. In *Production Economics: A Dual Approach to Theory and Applications*, vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

McFadden, D. 1978b. The general linear profit function. In *Production Economics: A Dual Approach to Theory and Applications*, vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

McKenzie, L. 1957. Demand theory without a utility index. *Review of Economic Studies* 24, 185–9.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA.: Harvard University Press.

Samuelson, P.A. 1974. Complementarity – an essay on the 40th anniversary of the Hicks–Allen revolution in demand theory. *Journal of Economic Literature* 12, 1255–89.

Shephard, R.W. 1953. *Cost and Production Functions*. Princeton: Princeton University Press.

Shephard, R.W. 1970. *Theory of Cost and Production Functions*. Princeton: Princeton University Press.

Uzawa, H. 1964. Duality principles in the theory of cost and production. *International Economic Review* 5, 216–20.

How to cite this article

Diewert, W. E. "cost functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000390> doi:10.1057/9780230226203.0329

The New Palgrave Dictionary of Economics Online

cost minimization and utility maximization

Peter Newman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Arrow corner; compensated demand; cost function; cost minimization and utility; duality; indirect utility function; linear programming; maximization

Article

Consider the following standard problem in the theory of demand: Find $x^* \geq 0$ so as to max $u(x)$ subject to $(x, p) \leq \omega$ where (x, p) is the inner product of the n-dimensional commodity and price vectors, and $\omega > 0$ and u are the consumer's income and utility function respectively; this problem is here labelled $\max(p, \omega)$.

The functional dependence of the *value* $v^* [\equiv u(x^*)]$ of this nonlinear programming problem on its parameters (p, ω) is denoted by $v(p, \omega)$, where v is the *indirect* utility function. The similar dependence of the *solution* x^* of max (p, ω) is written $f(p, \omega)$, where f is the *ordinary* (or *Marshallian*) demand function (or correspondence). If v^* does not exist then neither do v, x^* or f. Important though they are such non-existence problems are irrelevant here, so without further ado assume that every optimization problem has a solution.

Consider next a problem whose form is similar to that of $\max(p, \omega)$ but whose objective is different, that is, cost minimization rather than utility maximization. Specifically, find $x^{**} \geq 0$ so as to min $\min(x, p)$ subject to $u(x) \geq \tau$ where x,•p and u are as before and τ is a *target* level of utility; this new problem is labelled min(p,• τ). The functional dependence of the value $\mu^{**} (\equiv (x^{**}, p))$ of min(p,• τ) on its parameters (p,• τ) is denoted c(p,• τ), where c is the *cost* (or *expenditure*) function. The similar dependence of x^{**} on (p,• τ) is written h(p,• τ), where h is the *compensated* (or *Hicksian*) demand function (or correspondence).

Suppose now that max(p,• ω) is solved and its value v^* is inserted into the second optimization problem, thus creating the problem min(p,•v^*). Is each solution x^* of max(p,• ω) necessarily also a solution of min (p,•v^*)? Call this Question I. A similar question can be asked of the reverse situation, which is: For

arbitrary $(p, \cdot\, \tau)$ solve $\min(p, \cdot\, \tau)$, obtain it value μ^{**} and then solve the resulting max problem, (p, μ^{**}). Question II is then: Is each solution x^{**} of $\min(p, \cdot\, \tau)$ necessarily also a solution of $\max(p, \cdot\, \mu^{**})$? Problem $\min(p, \cdot v^*)$ has often been called the *dual* of $\max(p, \cdot\, \omega)$, from as far back as Arrow and Debreu (1954, pp. 285–6) to Deaton and Muellbauer (1980, pp. 37 ff.) and beyond. Indeed, this usage is now so common that for most economists $\min(p, \cdot v^*)$ seems to be *the* leading species of the genus *dual problem*. One can see why. It appears to be quite analogous to dual problems in linear programming (lp), with max becoming min, and objective and constraint functions becoming constraint and objective functions, respectively. However, the analogy with duals in lp is misleading, for each solution x^{**} of the alleged 'dual' $\min(p, \cdot v^*)$ is located in the *same* space as each solution x^* of its 'primal' $\max(p, \cdot\, \omega)$, whereas in lp the solutions to the dual all lie in the *dual* space. As Deaton and Muellbauer justly remark: 'The essential feature of the duality approach is a *change of variables*' (1980, p. 47, their italics). So a new term for the relation that $\min(p, \cdot v^*)$ bears to $\max(p, \cdot\, \omega)$ is needed in order to distinguish it from genuine duality; the 'mirrored' (or 'reflected') problem is suggested in Newman (1982).

In demand theory it is sometimes recognized explicitly that Question I needs an answer (for example, Samuelson, 1947, p. 103; McKenzie, 1957, p. 186) but more often not, probably because the usual assumptions on preferences are quite sufficient for coincidence of x^{**} with x^*. An explicit treatment appears unnecessary: '…•clearly, the vector of commodities must in both cases be the same' (Deaton and Muellbauer, 1980, p. 37). In welfare economics, however, it has long been recognized that a suitably generalized form of Question I, simple as it is, has importance for the first fundamental theorem of welfare economics, namely that every competitive allocation is (strongly) Pareto-optimal.

Question II has always been considered more delicate than Question I. Indeed, it was not even put until Arrow (1951, pp. 527–8) exhibited his famous 'exceptional case' (now often known as the Arrow Corner) in which it receives a negative answer. Its relevance for proofs of existence of competitive equilibrium was fully grasped by Arrow and Debreu (1954, sections 4 and 5), and later Debreu (1959, pp. 67–71), for essentially this reason, devoted four pages of his terse classic to a detailed examination of both Questions.

It is interesting that although the second Question is economically more subtle than the first, from a sufficiently abstract point of view the two are logically isomorphic (see Newman, 1982, where in both Theorem (c) and Theorem (c′) the assertion 'iff' is wrong and should be replaced by 'if'). While such extreme abstraction is irrelevant here, both $\max(p, \cdot\, \omega)$ and $\min(p, \cdot\, \tau)$ do need to be put into a form suitable for general equilibrium theory.

The setting

The consumer is now endowed, not with an exogenous positive income, but with a nonzero bundle x^0, whose worth (x^0, p) may be zero. For simplicity (and only that), free disposal is assumed.

Assumptions about preferences

The consumer has two disjoint binary relations $\succ$ ('preference') and $\sim$ ('indifference') each defined on

some non-empty $S \subset R^n$; the union of $\succ$ and ~ is denoted $\succsim$. Indifference is reflexive and symmetric (so that preference is irreflexive) and the statements $x^1 \succ x^3$ and $x^2 \sim x^3$ together imply $x^1 \succ x^3$. Neither completeness nor transitivity of preference is assumed, so a utility function need not exist.

The generalized version of max(p,• ω) is then: Find $x^* \in S$ for which $(x^*, p) \leq (x^0, p)$ and such that $x \succ x^*$ implies $(x, p) > (x^0, p)$. In words, 'x^* is feasible and anything preferred to it is unaffordable'. This problem is labelled max(p,•x^0).

The generalized version of min(p,• τ) is: Find $x^{**} \in S$ for which $x^{**} \succsim t$ and such that $x \succsim t$ implies $(x, p) \geq (x^{**}, p)$. In words, 'anything at least as good as the target bundle $t \in S$ costs at least as much as x^{**}'. This problem is labelled min(p,•t).

Note that in the absence of a utility function max(p,•x^0) can have a solution but not a value, while min(p,• t) can have both value and solution, just as before.

Some definitions

Any bundle $x^\bullet$ to which no $x \in S$ is preferred is called *bliss*, while a bundle $x_\bullet$ for which at prices p there is no cheaper $x \in S$ is called *p-minimal*. Preferences are *locally nonsatiated* at x^1 if *any* neighbourhood N (x^1) contains $x \succ x^1$, while $x^2 \in S$ has *locally cheaper points* at p (a term apparently due to McKenzie, 1957) if *any* neighbourhood $N(x^2)$ contains a bundle $x \in S$ which at prices p is cheaper than x^2. If $x^\bullet$ is bliss it cannot be locally nonsatiated, and if $x_\bullet$. is p-minimal it has no locally cheaper points at p.

Following Bergstrom, Parks and Rader (1976), preferences are said to have *open upper sections* if $x' \succ x^2$ implies the existence of a neighbourhood $N(x^1) \subset S$ for which $x \succ x^2$ for every x in it.

The following simple result answers both Questions satisfactorily and generalizes easily to a wide class of infinite-dimensional commodity spaces.

Theorem: (i) Assume (a) that if $x \in S$ is not bliss it is locally nonsatiated, and (b) that the solution x^* of max(p,•x^0) is not bliss. Then x^* also solves min(p,•x^*). Moreover, the value μ^{**} of min(p,•x^{**}) equals (x^0, p).

(ii) Assume (c) that preferences have open upper sections, (d) that if $x \in S$ is not p-minimal it has locally cheaper points at p, and (e) that the solution x^{**} of min(p,•t) is not p-minimal. Then x^{**} also solves max (p,• μ^{**}), where $\mu^{**} = (x^{**}, p)$.

Proof: (i) Suppose the result false, so there exists $x^1 \bullet x^*$ such that $(x^1, p) < (x^*, p)$. Now $x^1 \succsim x^*$ cannot occur because if it did $(x^1, p) < (x^*, p) \leq (x^0, p)$ would imply that x^* does not solve max(p,• x^0), contrary to hypothesis. So $x^1 \sim x^*$.

Since the vector p represents a continuous linear function (al) there is a neighbourhood $N(x^1)$ all of whose points are cheaper at prices p than x^*. From (b) there exists $x \succ x^*$, and this and the symmetry of ~ imply that $x \succ x^1$ as well, so that x^1, is not bliss either. Hence from (a) at least one member of $N(x^1)$, say x^2, is such that $x^2 \succ x^1$. Because $x^1 \bullet x^*$ this leads to $x^2 \succ x^*$, which again contradicts the hypothesis. Thus x^* solves min(p,•x^*), which implies $\mu^{**} = (x^*, p)$.

Suppose $(x^*, p) < (x^0, p)$. By the continuity of p there exists $N(x^*)$ all of whose points are cheaper at prices p than is x^0, while from (b) and (a) at least one of them, say x^3, is such that $x^3 \succ x^*$. Yet again, this contradicts the hypothesis. So $(x^*, p) = (x^0, p)$.

(ii) By assumption $x^{**} \succsim t$. Suppose $x^{**} \succ t$. From (c) there exists $N(x^{**})$ such that $x \succ t$ for every x in it. From (e) and (d) x^{**} has locally cheaper points at p, so at least one x in $N(x^{**})$, say x^1, is cheaper than x^{**} at p. Since $x^1 \succ t$, this contradicts the hypothesis that x^{**} solves min(p,•t). Hence $x^{**} \sim t$.

Suppose now that x^{**} does not solve max(p,• μ^{**}), so there is an x^2 such that $x^2 \succ x^{**}$ and $(x^2, p) \leq (x^{**}, p)$. Hence $x^2 \bullet t$. If x^2 were cheaper at p than x^{**} that would again contradict the hypothesis. So $(x^2, p) \leq (x^{**}, p)$.

From (e) there is an $x \in S$ cheaper at p than x^{**}, hence cheaper than x^2, so x^2 is not p-minimal either. Since $x^2 \succ t$, from (c) there exists $N(x^2)$ such that $x \succ t$ for every x in it and from (d) at least one of these must be cheaper than x^2 at p, and so cheaper than x^{**}, which again contradicts the hypothesis. Q.E.D.

One sees just how few and how weak are the assumptions on preferences that enable Questions I and II to be answered, as distinct (for ex) from those needed to guarantee the *existence* of solutions x^* and x^{**}. Note that two assumptions are used for Question I and three for II, an inequality which occurs because the constraint in max(p,•x^0) is linear and hence continuous, whereas in the problem min(p,•t) some continuity in the (nonlinear) constraint has to be *imposed* by means of the 'extra' assumption (c). This asymmetry disappears in a more abstract treatment, with more general constraints.

The intuitions behind the proof help to see why Question II is a serious problem for general equilibrium theory. In the proof of (i) the bundle x^1 that is cheaper than x^* is made a little bigger, in effect increasing satisfaction by increasing expenditure, until a bundle is reached that is still affordable at income (x^0, p) but which is better than x^*; that expenditure can always be thus 'traded' for satisfaction is assured by local nonsatiation. In the proof of (ii) the bundle x^1 that is better than x^{**} is made a little smaller, lessening satisfaction in return for less cost, until a bundle is reached that is still as good as t but which costs less than x^{**}; such 'trading' of satisfaction for expenditure is guaranteed by the existence of locally cheaper points. However, if the expenditure on x^{**} at prices p is already least possible (that is, if x^{**} is p-minimal) then 'trading' in *that* direction cannot occur – one cannot go below least cost.

Of the five assumptions of the Theorem the only one whose meaning is not transparent and whose restriction is not 'reasonable' is (e), so that it comes as no surprise that the main thing wrong at the Arrow Corner is that (e) does not hold there. For further discussion of this Slater-like assumption and its role in general equilibrium theory, see consumption sets.

See Also

- duality

Bibliography

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of*

the Second Berkeley Symposium on Mathematical Statistics and Probability, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Bergstrom, T.C., Parks, R.P. and Rader, T. 1976. Preferences which have open graphs. *Journal of Mathematical Economics* 3, 265–8.

Deaton, A. and Muellbauer, J. 1980. *Economics and Consumer Behaviour*. Cambridge: Cambridge University Press.

Debreu, G. 1959. *Theory of Value*. Cowles Commission Monograph No. 17. New York: John Wiley.

McKenzie, L. 1957. Demand theory without a utility index. *Review of Economic Studies* 24, 185–9.

Newman, P. 1982. Mirrored pairs of optimization problems. *Economica* 49, 109–19.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA.: Harvard University Press.

How to cite this article

Newman, Peter. "cost minimization and utility maximization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000391>
doi:10.1057/9780230226203.0330

The New Palgrave Dictionary of Economics Online

cost – benefit analysis

David L. Weimer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Cost–benefit analysis (CBA) is a collection of methods and rules for assessing the social costs and benefits of alternative public policies. It promotes efficiency by identifying the set of feasible projects that would yield the largest positive net benefits to society. The willingness of people to pay to gain or avoid policy impacts is the guiding principle for measuring benefits. Opportunity cost is the guiding principle for measuring costs. CBA requires that appropriate shadow prices be derived when policies have effects beyond those that can be taken into account as changes of prices or quantities in undistorted markets.

Keywords

consumer surplus; contingent valuation; cost–benefit analysis; distortions; donor value; equivalent variation; Hicks compensation; Hicks, John R.; Kaldor, N.; Marshallian demand curves; opportunity cost; option price; present value; pure time preference; revealed preference; shadow prices; social choice; social surplus; substitutes and complements; travel-cost method; value of statistical life; willingness to pay

Article

Public policies, such as infrastructure projects, social welfare programmes, tax laws and regulations, typically have diverse effects in the sense that people would be willing to pay something to obtain effects they view as desirable and would require compensation to accept voluntarily effects they view as undesirable. If, across all members of society, the total amount willing to be paid by those who enjoy desirable effects (benefits) exceeds the total amount needed to compensate those who suffer undesirable effects (costs), then adopting the policy would make it potentially possible to achieve a Pareto improvement on the status quo. If the benefits do not exceed the costs, then adopting the policy does not offer a potential Pareto improvement. How should such costs and benefits be determined? Cost–benefit

analysis (CBA) is the collection of generally accepted methods and rules for assessing the social costs and benefits of alternative public policies.

The US Flood Control Act of 1936 appears to be the first call for CBA to be systematically used to inform public policy (Steiner, 1974); it became embedded within modern welfare economics with articles by John R. Hicks (1939) and Nicholas Kaldor (1940) that set out the efficiency rationale for requiring policies to have positive net benefits. Two forces have contributed to the increased use of CBA since the 1960s. First, budget pressures and the desire to avoid inefficient regulations have led many governments to promote, or even require, the subjection of certain types of policies to CBA. Its use in the United States, particularly in the area of economic regulation, has been mandated by a series of Executive Orders (Hahn and Sunstein, 2002). Her Majesty's Treasury in the United Kingdom publishes the *Green Book* to help public sector organizations apply CBA to ensure that ‘public funds are spent on activities that provide the greatest benefits to society, and that they are spent in the most efficient way’ (HM Treasury, 2002: v). Second, economists have shown ingenuity in finding ways to value goods not traded in efficient markets, thereby expanding the range of policies to which CBA can be reasonably applied. For example, the travel-cost method provides a way to value recreational facilities that charge an administratively determined entry fee (Clawson and Knetsch, 1966); hedonic pricing models facilitate valuation of spatially varying local public goods (Smith and Huang, 1995); and the development of the contingent valuation survey method, propelled by environmental damage assessment suits in US courts, permits the valuation of public goods, such as existence value, that lack readily observable behavioural traces needed for revealed preference estimation (David, 1963; Bateman and Willis, 2000).

CBA promotes efficiency by identifying the set of feasible projects that would yield the largest positive net benefits. Three conceptual criticisms can be made against this proposition. First, because those who suffer costs from a policy are almost never fully compensated, CBA in any particular application generally will not guarantee a Pareto improvement. The counter-argument is that, if CBA is consistently used to select policies offering the largest net benefits and there are no consistent losers, then it is likely that overall everyone will actually be made better off. Second, the CBA techniques for measuring net benefits cannot guarantee a coherent social ordering of policy alternatives. For example, it is possible to identify situations in which moving from one policy to another offers positive net benefits as does moving back to the original policy (Scitovsky, 1941; Blackorby and Donaldson, 1990). As no fair social choice rule can guarantee a transitive social ordering (Arrow, 1963), this result is not surprising and is of minor consequence compared with the practical difficulties encountered in applying CBA. Third, and most important, only a few economists argue that public policies should be selected solely to promote the goal of efficiency. Other goals, such as equity and preservation of human dignity, are often legitimately viewed as relevant to policy choice, so that CBA is inappropriate as a decision rule. Nonetheless, as efficiency is almost always one of the relevant goals of public policy, CBA remains useful as a method for assessing efficiency in the context of a broader multi-goal analysis.

Social perspective

CBA assesses social costs and benefits, which distinguishes it from the self-regarding calculus of individual economic actors. The meaning of ‘social’ in this context is twofold. First, it involves the definition of the relevant society; that is, it requires a determination of whose costs and benefits have

standing (Whittington and MacRae, 1986). Economists generally argue for national standing, recognizing that those in a particular country live under the same political contract, or constitution, and share a common economy with its own fiscal and monetary policy. In practice, however, sub-national governments often base their decisions only on their own costs and benefits and therefore demand CBA with standing restricted to those under their jurisdictions. Even when geographic standing is resolved, issues remain as to whether the costs and benefits of all residents – citizens, legal aliens, illegal aliens, those with legally proscribed preferences – should count (Zerbe, 1998).

Second, it requires comprehensive assessment of the valued effects of policies on those with standing. The effects are commonly divided into the categories of active and passive use. Policies affect active use by changing the observable quantities of goods consumed, such as day care or fishing. Passive use includes all those effects that cannot be readily identified with observable changes in behaviour: existence value, or the willingness to pay for some good, such as wilderness, that one never expects to consume actively (Krutilla, 1967); option value, or the willingness to pay for some good that one may wish to consume actively in the future (Weisbrod, 1964); donor value, or the willingness to pay for redistributions of goods to others (Hochman and Rogers, 1969). The absence of observable behaviour precludes valuation of passive use through the revealed preference methods most favoured by economists. Stated preference methods, such as contingent value surveys, are thus necessary for undertaking comprehensive assessments of policies with effects on passive use.

Social benefits: willingness to pay

A common metric for policy effects is required if these effects are to be aggregated across individuals within the relevant society. If more than one policy alternative is to be compared with the status quo, then this metric must have ordinal properties. Further, if it is to be compared with the resource costs of implementing the policy, then it must be measured in the monetary unit of the society. Equivalent variation (*EV*) satisfies these conditions (McKenzie, 1983). Consider the expenditure, or cost-utility, function C(U,P), where *C* is the amount of money needed to achieve utility *U* with price vector *P*. If U_1 is the person's utility under the price vector P_1 that would result from the policy change and P_0 is the price vector that would result under the status quo, then the equivalent variation of the policy change is given by

$$EV = C(U_1, P_0) - C(U_1, P_1)$$

the difference between the expenditure needed to achieve U_1 without the policy and with it. The EV is the amount of money that one would have to give to the person instead of implementing the policy so that the person is as well off as he or she would have been had the policy been implemented. A negative EV indicates that the person finds the net effects of the policy undesirable.

In its actual use, CBA almost always evaluates policy effects with willingness to pay, which differs conceptually from EV. Willingness to pay answers the question: how much money could be taken away

from a person in conjunction with the policy so that he or she has the same utility with the policy as without it? Rather than corresponding to EV, which holds utility constant at a level with the policy, willingness to pay corresponds to compensating variation, which holds utility constant at the pre-policy level. Although compensating variation is more intuitively appealing, it does not provide a fully satisfactory money metric like EV.

The equivalent or compensating variation of a price change in a single market can be calculated as the change in social surplus as measured under the appropriate Hicksian, or utility-compensated, demand schedule. In practice, however, analysts typically work with econometrically derived demand curves that do not hold utility constant. Changes in consumer surplus measured with these Marshallian demand curves only approximate the compensating variation, with differences driven by income effects that can be large for either large income elasticities or large price changes. Some progress has been made to put bounds on the differences between the Marshallian and Hicksian measures (Willig, 1976; Seade, 1978), but these bounds are rarely applied in practice.

The interpretation of Marshallian consumer surplus as willingness to pay becomes even more complicated when policies have secondary effects in the markets of complements and substitutes of the goods primarily affected by policies. Although a general equilibrium model would be most appropriate for taking account of these secondary market effects, common practice is to approximate the combined effect of the primary and secondary markets by measuring surplus changes with the use of an estimated demand schedule for the primary market that does not hold the prices of substitutes and complements constant (Sugden and Williams, 1978; Gramlich, 1990; Boardman et al., 2006). In such cases, analysts need not account for price changes in undistorted secondary markets. Indeed, doing so would likely result in double counting of benefits.

Social costs: opportunity costs

Public policies generally require the use of real resources to produce their effects. The guiding principle for monetizing the forgone value of these resources is opportunity cost: what is the value of the resources in their next-best use? That is, what is the value forgone by using the resources for the project? When factor markets are undistorted and the additional demand created by the project does not increase price, the opportunity cost of the resource just equals its market value, which, if the resource is obtained by purchase, just equals the expenditure on the resource. When factor markets are undistorted but the additional demand induced by the policy drives up price, then the opportunity cost of the resource equals the sum of expenditure and the change in social surplus, the algebraic sum of the change in consumer surplus and the change in rents usually measured as change in producer surplus based on the short-run supply schedule (Mishan, 1968). For example, if supply and demand curves are linear, then the opportunity cost equals the average of the pre- and post-purchase prices of the resource times the quantity purchased.

If markets are distorted, then even if price does not change the opportunity cost does not necessarily equal the expenditure required to secure supply. For example, a common factor-market distortion is involuntary unemployment resulting from minimum wages imposed by either law or custom. The expenditures needed to hire workers from a market with involuntary unemployment for a project clearly overestimate the opportunity cost of this labour. Nonetheless, the opportunity cost is almost certainly not zero, as sometimes argued by policy advocates, because the time of the workers hired by the project has

an opportunity cost in terms of forgone leisure and household production.

Accommodating uncertainty

CBA requires prediction of the effects of adopting a policy. Predictions are inherently uncertain. In addition to uncertainty about such parameters as price elasticities required for predictions of changes in social surplus, CBA often requires analysts to confront fundamental uncertainty about future states of the world. For example, preparing a vaccine to guard against a potential pandemic is costly but offers large benefits in the event that a pandemic actually materializes. CBA requires analysts to convert these uncertainties into risks by specifying representative states of the world and assigning probabilities to these states. Common practice is to model the policy choice as a decision analysis problem, or game against nature, and to choose the policy that maximizes the expected value of social surplus.

A more conceptually valid measure of the benefits of a project with certain costs in the face of risk about the future state of the world is option price (Graham, 1981). Option price answers that question: what is the maximum certain payment that an individual would be willing to make to obtain the project? The sum of these certain payments for all those with standing can then be compared with the certain cost of implementing the policy. In general, however, option price does not equal the expected value of an individual's surplus over the possible states of the world; it differs from expected surplus by the option value of the policy for the individual. Although contingent valuation surveys seek to elicit individuals' option prices directly, more commonly analysts estimate benefits as expected surpluses, and consider option value as an excluded value. Some progress has been made in signing option value (Larson and Flacco, 1992), but analysts rarely have enough information for confidently including it as a monetized correction to expected surplus.

Discounting for time

Policies typically have effects that extend far into the future. Infrastructure projects in particular are usually characterized by large initial investments followed by beneficial use over years or even scores of years. CBA requires that costs and benefits accruing in the future be converted into their present value equivalents. On the assumption that future costs and benefits are predicted in real dollars, then a dollar of cost or benefit occurring *t* periods beyond the present equals in present value terms

$$1/(1+d)^t$$

where *d* is the real discount rate for the period length. In practice, discounting is usually done on an annual basis. As valid comparison of projects requires that they be assessed over the same time horizon, it is often necessary to convert present values to equivalent perpetual streams of constant values through the use of an annuity factor.

The appropriate value for the real discount rate remains controversial. One approach is to set the

discount rate equal to the marginal rate of pure time preference, the rate at which consumers are indifferent between exchanging current for future consumption. Another approach is to set the discount rate equal to the opportunity cost of capital, the marginal rate of return on private investment. In an ideal capital market these two rates would be equal. In the presence of transaction costs and taxes, however, these rates differ substantially. For example, an estimate of the marginal rate of pure time preference based on the after-tax real rate of return on US treasury bonds is 1.5 per cent, while an estimate of the opportunity cost of capital based on the expected real yield on AAA corporate bonds is 4.5 per cent (Moore et al., 2004).

If all costs and benefits correspond to changes in consumption, then the marginal rate of pure time preference is the appropriate discount rate. Instead, if all costs and benefits correspond to changes in private investment, then the marginal rate of return on private investment is the appropriate discount rate. However, most projects involve changes in both consumption and investment. The shadow price of capital approach involves expressing all costs and benefits in terms of consumption changes so that the marginal rate of pure time preference can be applied (Bradford, 1975). In application, this means applying a shadow price to changes in private investment so that they are converted to the present values of their associated streams of consumption changes.

Shadow prices

Much of the challenge of CBA lies in deriving appropriate shadow prices when policies have effects beyond those that can be taken into account as changes of prices or quantities in undistorted markets. In developing countries, for example, import and export controls and the presence of subsistence agriculture often distort virtually all prices, necessitating the determination of a complete set of shadow prices based on prices in international markets (Little and Mirrlees, 1974; Squire and van der Tak, 1975; Dinwiddy and Teal, 1996). Economic research provides a number of shadow price estimates that can be used in conducting CBA. Indeed, were these shadow prices not readily available, the plausible range of application of CBA would be much narrower.

One of the most commonly needed shadow prices is the value of a statistical life. That is, what is the willingness of a representative member of a population to pay for reductions in mortality risk? Economists have used a variety of methods to estimate the value of a statistical life, most commonly taking advantage of differences in risks and wages across occupations or the purchases of safety devices. The number of studies is sufficiently large that a number of meta-analyses have been conducted to develop estimates of the value of a statistical life for the United States in the range of roughly $4 million to $6 million in 2002 dollars (Miller, 2000; Viscusi and Aldy, 2003). Tied to any estimate of the value of a statistical life is the value of a life year. Health economists have developed a number of methods for estimating the quality of life in various health states, so that, in conjunction with the value of a life year, they can monetize a quality-adjusted life year (QALY) for use in CBAs of health care interventions (Dolan, 2000). Estimates of shadow prices for injuries, noise, recreational activities, air pollutants, commuting time, and the marginal excess burden of taxation (for application to changes in government revenue) are also readily available (Boardman et al., 2006).

See Also

- consumer surplus
- contingent valuation
- hedonic prices
- Pareto principle and competing principles
- rent
- social discount rate
- value of life
- value of time

Bibliography

Arrow, K. 1963. *Social Choice and Individual Values*, 2nd edn. NewHaven: Yale University Press.

Bateman, I. and Willis, K., eds. 2000. *Valuing Environmental Preferences Theory and Practice of the Contingent Valuation Method in the US EC and Developing Countries*. Oxford: Oxford University Press.

Blackorby, C. and Donaldson, D. 1990. A review article: the case against the use of the sum of compensating variation in cost–benefit analysis. *Canadian Journal of Economics* 23, 471–94.

Boardman, A., Greenberg, D., Vining, A. and Weimer, D. 2006. *Cost–benefit Analysis: Concepts and Practice*. 3rd edn. Upper Saddle River, NJ: Prentice Hall.

Bradford, D. 1975. Constraints on government investment opportunities and the choice of discount rate. *American Economic Review* 65, 887–99.

Clawson, M. and Knetsch, J. 1966. *Economics of Outdoor Recreation*. Baltimore, MD: Johns Hopkins University Press.

David, R. 1963. Recreation planning as an economic problem. *Natural Resources Journal* 3, 239–49.

Dinwiddy, C. and Teal, F. 1996. *Principles of Cost-benefit Analysis for Developing Countries*. Cambridge: Cambridge University Press.

Dolan, P. 2000. The Measurement of health-related quality of life for use in resource allocation in health care. In *Handbook of Health Economics* 1B, ed. A. Culyer and J. Newhouse. Amsterdam: Elsevier.

Graham, D. 1981. Cost–benefit analysis under uncertainty. *American Economic Review* 71, 715–25.

Gramlich, E. 1990. *A Guide to Benefit-Cost Analysis*. 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Hahn, R. and Sunstein, C. 2002. A new executive order for improving federal regulation? Deeper and wider cost-benefit analysis. *University of Pennsylvania Law Review* 150, 1389–552.

Hicks, J.R. 1939. The Valuation of Social Income. *Economica* 7, 105–24.

HM Treasury. 2002. *The Green Book: Appraisal and Evaluation in Central Government*. London: The Stationary Office.

Hochman, H. and Rogers, J. 1969. Pareto optimal redistribution. *American Economic Review* 59, 542–57.

Kaldor, N. 1940. Welfare propositionsof economics and interpersonal comparisons of utility. *Economic Journal* 49, 549–52.

Krutilla, J. 1967. Conservation reconsidered. *American Economic Review* 57, 777–86.

Larson, D. and Flacco P. 1992. Measuring option prices from market behavior. *Journal of Environmental Economics and Management* 22, 178–98.

Little, I. and Mirlees, J. 1974. *Project Appraisal and Planning for Developing Countries*. London: Heinemann Educational.

McKenzie, G. 1983. *Measuring Economic Welfare*. Cambridge: Cambridge University Press.

Miller, T. 2000. Variations between countries in the values of statistical life. *Transport Economics and Policy* 34, 169–88.

Mishan, E. 1968. What is producer's surplus? *American Economic Review* 58, 1269–82.

Moore, M., Boardman, A., Vining, A., Weimer, D. and Greenberg, D. 2004. ‘Just give me a number!’ Practical values for the social discount rate. *Journal of Policy Analysis and Management* 23, 789–812.

Scitovsky, T. 1941. A note onwelfare propositions in economics. *Review of Economic Studies* 41, 77–88.

Seade, J. 1978. Consumer's surplus and linearity of Engel curves. *Review of Economic Studies* 9, 77–88.

Smith, V. and Huang, J. 1995. Can markets value air quality? A meta analysis of hedonic property value models. *Journal of Political Economy* 103, 209–77.

Squire, L. and van der Tak, H. 1975. *Economic Analysis of Projects*. Baltimore MD: Johns Hopkins University Press.

Steiner, P. 1974. Public expenditure budgeting. In *The Economics of Public Finance*, ed. A. Blinder et al. Washington, DC: The Brookings Institution.

Sugden, R. and Williams, A. 1978. *Principles of Practical Cost–benefit Analysis*. Oxford: Oxford University Press.

Viscusi, W. and Aldy, J. 2003. The value of statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27, 5–76.

Weisbrod, B. 1964. Collective consumption services of individual consumption goods. *Quarterly Journal of Economics* 78, 71–7.

Whittington, D. and MacRae, D. 1986. The issue of standing in cost–benefit analysis. *Journal of Policy Analysis and Management* 5, 665–82.

Willig, R. 1976. Consumer's surplus without apology. *American Economic Review* 66, 589–97.

Zerbe, R., Jr., 1998. Is cost–benefit analysis legal? Three rules. *Journal of Policy Analysis and Management* 17, 419–56.

How to cite this article

Weimer, David L. "cost–benefit analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000397> doi:10.1057/9780230226203.0328

The New Palgrave Dictionary of Economics Online

cost-push inflation

George L. Perry
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The concept of cost-push inflation emerged after the Second World War to describe the price increases arising from labour unions pushing up wages despite excessive unemployment. With the oil price shocks of the 1970s, it was used to describe any important shift up in supply schedules at given levels of aggregate demand. Most central banks differentiate between supply shock effects and demand effects by distinguishing between overall inflation and core inflation, the latter omitting the direct contribution of shocks to oil and food prices, the two most important sources of supply shock large enough to register on broad inflation measures.

Keywords

aggregate demand; business cycle; cost-push inflation; excess demand; expectations; full employment; incomes policies; inflation; labour supply; learning; market power; monetary policy; natural rate of unemployment; Organization of Petroleum Exporting Countries (OPEC); Phillips curve; stabilization policy; sticky prices; supply shocks; trade unions; unemployment; wage inflation; wage rigidity

Article

The concept of cost-push inflation emerged in the period after the Second World War. The Keynesian model of that time emphasized that the economy could operate with inefficiently low utilization of its capital and labour resources, and that expanding demand would employ those resources. Once full employment was achieved, further expansion of demand would only pull up nominal wages and prices. In contrast to this demand-pull inflation, cost-push described the price increases that came from labour unions pushing up wages despite the existence of excessive unemployment. Since the 1970s, when oil prices rose by many times in two abrupt steps, the idea of cost push has been extended to describe price increases arising from any important shift up in supply schedules at given levels of aggregate demand. The key distinction between price increases arising from monopoly power in wage settings or from any

other supply shock and price increases arising from an increase in aggregate demand along an unchanged supply curve is important both for empirical modelling of inflation and for stabilization policy. By the 1960s, the short-run Phillips curve had emerged as a description of the relation between inflation and unemployment over the business cycle. It described an empirical regularity according to which wages rose gradually faster as unemployment declined, with the relation becoming steeper the lower the unemployment rate. Subsequent amendments to this model took explicit account of learning and expectations and of the interrelation between wages and prices. In the dominant model that emerged, inflation will accelerate (decelerate) indefinitely if the economy operates persistently below (above) a natural rate of unemployment. And in models that stress the importance of expectations, the anticipation of faster or slower price increases speeds up this process of acceleration or deceleration.

While inflation is responsive to aggregate demand in all these models, its responsiveness to supply shocks is more nuanced. In models that stress inflationary expectations, shocks that are widely perceived as one-time shifts up nominal supply curves will lead only to one-time shocks to the price level. In models with institutions that partially or fully index wages to prices, or models with adaptive expectations of inflation, such shocks will have larger and more protracted effects.

Empirically, the distinction between cost-push and excess-demand effects is not always easily drawn. The inflation identified with unemployment below the natural rate or with the steep portion of the short-run Phillips curve is attributable to excess demand. The more modest variations in inflation that may occur as unemployment varies above the natural rate are not characterized so readily. A useful interpretation of these systematic cyclical tendencies is that they represent the normal operation of heterogeneous labour markets in response to cyclical variations in aggregate demand, with wages and prices in some sectors rising faster as their markets tighten while slack is still present in other sectors. On this interpretation, they neither signal that the economy is at a natural rate nor indicate the presence of exogenous cost-push effects on prices. However, these modest variations in inflation may also indicate cost-push effects in wage settings that interfere with achieving full employment, and at times past policymakers have interpreted them in this way and tried to suppress them.

The interdependence between prices and wages presents another difficulty in distinguishing endogenous from exogenous changes in wages. If labour supply depends on real wages – that is, wages relative to the average price level – then labour supply will not change if nominal wages change proportionally in response to disturbances to the cost of living. The narrowest concept of cost-push would, therefore, include only shifts up in labour supply schedules that raise wages relative both to their normal response to cyclical demand conditions and to their normal response to consumer prices.

Such complications obscure the possible presence of cost push from wages in typical circumstances. However, when the exercise of market power in wage setting is extreme, it becomes more apparent. In the United States, large wage increases in the early post-war years are examples of cost push. Coming after wartime controls, these did not raise the concerns that the abrupt acceleration of wages in many industrialized economies did in the late 1960s and early 1970s. For example, in Germany annual increases in hourly compensation jumped from 7.5 per cent in 1968 to 17.5 per cent in 1970, and in the United Kingdom the acceleration over the same period was from seven per cent to 15.5 per cent.

During the 1970s, supply shocks to important raw materials prices dominated world price developments in the decade, producing the second main type of cost-push inflation. These supply shocks included the historic increases in oil prices in 1973–4 and again in 1979, and the food price explosion of 1973. Although world aggregate demand was relatively strong in both 1973 and 1979, the magnitude of the

price increases that resulted would not be expected, and is better seen as a consequence of major shifts in world supplies. A succession of poor crops provoked the food price rise, while the successful organization of the OPEC oil cartel, aided by a levelling off in United States oil production, caused the oil price explosion.

Coincident with the 1973–5 supply shocks, further large jumps in wage inflation occurred in several countries. In both the United Kingdom and Japan, annual increases in hourly compensation rose to more than 30 per cent from less than half that rate in 1972. Most other major industrial countries experienced similar, though less dramatic, accelerations in wages. Although these wage developments were doubtless fuelled by the effects of the supply shocks on consumer prices, the differences across countries indicate another round of wage push in many, even when one allows for a normal response to price changes. The rapid wage increases in turn further boosted consumer prices. The eventual changes in real wages, as well as the eventual increase in price levels, varied significantly among the industrial countries during the mid-1970s. In the United States, the speed-ups and slowdowns in wage increases and in prices were far less dramatic than in Europe and Japan. However, over the entire decade of the 1970s wages in the highly unionized sectors of the US economy outpaced economy-wide wages substantially, indicating a moderate but persistent wage push from important major industries.

While this post-war record shows that both wage push and supply shocks have at times been important in pushing up price levels, several difficulties remain with the idea of cost push as a distinct source of inflation, and some analysts reject the idea altogether. First of all, inflation refers to an ongoing rate of increase in prices. A one-time rise in the average price level will translate into some rate of increase in prices over a period spanning the rise. Without quibbles over how long a time period is needed before a measurement qualifies as an ‘inflation rate’, the distinction between a one-time rise in the price level and an ongoing inflation rate is important. Second, inflation refers to the general price level, not to a subset of prices. A rise in oil prices is, first of all, a rise in the relative price of oil. If wages and prices were fully flexible and responded instantly to changes in the balance between demand and supply, then, in the presence of non-accommodating macroeconomic policies, cost-push shocks would indeed create only relative price changes; inflation, in the aggregate, would be impossible. Those who see monetary policy as able to control the overall price level, if not instantly at least over a relatively short period of time, see a cost push from some sectors as a relative price change that becomes a change in the overall price level only if accommodated by monetary policy. On this view, the accommodation rather than the cost-push causes the inflation.

However, such reasoning ignores the considerable downward rigidity in wages and stickiness in many prices as well as the interactions between prices and wages in modern economies, and thus loses the important role that cost-push shocks played in shaping economic performance in these inflationary periods. There are positive correlations among most prices and wages in the economy. In part these reflect common reactions to aggregate developments and in part they represent causal links among wages and prices throughout the economy.

When the links are strong, as they were in the inflationary periods of the late 1960s and 1970s, a cost-push supply shock will not only add directly to the average price level but will set in motion increases in other prices and wages strong enough to persist for some time, even in the face of slowing demand and increasing underutilization of resources. Consequently, an attempt by monetary policy to hold the overall price level unchanged in the face of such a cost-push shock will result mainly in reducing output and employment. Only gradually will the upward movement of prices originating from a supply shock

yield to restrictive monetary policies. On the other hand, because the initial shock induces positively correlated responses in wages and other prices, an accommodative policy that aims to maintain output and employment in the face of the shock will result in a rise in the overall price level that is substantially greater than the direct effect of the shock itself. The question confronting stabilization policy is thus how much to accommodate. And the best answer will differ with different institutions and at different times.
The idea that cost-push inflation originating in excessive union wage demands would interfere with the attainment of full employment prompted attempts in several countries to design incomes policies as part of the stabilization policy arsenal. The idea was that demand management by government would aim at keeping the economy around full employment, while understandings among government, labour and business would aim at heading off wage-push inflation that might otherwise arise before full employment was achieved. There was some evidence of success from incomes policies, known as wage-price guideposts in the United States, in the mid-1960s (Perry, 1967). But whatever chance such policies may have had in the longer run in a relatively benign environment, they were overwhelmed once economies were driven into the excess demand region during the Vietnam War, and the oil and food supply shocks of the early 1970s sharply raised average price levels everywhere. There has been little interest in incomes policies since that time.
By the 1990s, conventional stabilization policies had achieved low inflation rates throughout the industrial world, and the power of unions to originate more inflationary wage increases was very sharply reduced in almost all countries. Both these developments have lessened the problems of stabilization policy. There is evidence that the low-inflation environment has sharply reduced the links that formerly caused price shocks to spark a wage-price inflationary spiral, as they did in the 1970s (Brainard and Perry, 2000). Wages did not accelerate in response to the world oil price shocks of the mid-2000s, and monetary policymakers were able to focus largely on the core inflation rate – the aggregate inflation rate excluding food and energy prices – in setting policy. At least for now, inflation originating from cost push poses a much smaller risk for stabilization policies today than it has at times in the past.

See Also

- demand-pull inflation
- inflation
- inflation expectations

Bibliography

Akerlof, G., Dickens, W. and Perry, G. 1996. The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1996(1), 1–59.

Brainard, W. and Perry, G. 2000. Making policy in a changing world. In *Economic Events, Ideas, and Policies*, ed. and J. Tobin. Washington, DC: Brookings Institution.

Okun, A. 1981. *Prices and Quantities: A Macroeconomic Analysis*. Washington, DC: Brookings

Institution.

Okun, A. and Perry, G. 1978. *Curing Chronic Inflation*. Washington, DC: Brookings Institution. Also available as G. Perry, Slowing the wage-price spiral: the macroeconomic view. *Brookings Papers on Economic Activity* 1978(2), 259–99.

Perry, G. 1967. Wages and the guideposts. *American Economic Review* 57, 897–904.

Schultze, C. 1985. Microeconomic efficiency and nominal wage stickiness. *American Economic Review* 75, 1–15.

Tobin, J. 1972. Inflation and unemployment. *American Economic Review* 62, 1–18.

How to cite this article

Perry, George L. "cost-push inflation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000509> doi:10.1057/9780230226203.0331

The New Palgrave Dictionary of Economics Online

countertrade

Dalia Marin and Monika Schnitzer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International countertrade – tying an import to an export – emerged in the 1980s in response to the international debt crisis. Barter – the exchange of goods without using money – re-emerged in transition economies in the 1990s, in response to a domestic debt crisis. Both phenomena can be explained as institutional responses to contractual problems arising in imperfect capital and goods markets. Countertrade introduces a deal-specific collateral that improves the creditworthiness of countries and firms, and facilitates technology transfer to developing countries. Barter helps to overcome the lack of trust problem in the former Soviet Union.

Keywords

asymmetric information; barter; buyback; collateral, deal-specific; commitment; contract enforcement; counterpurchase; countertrade; creditworthiness; credit constraint; cross-subsidy; foreign direct investment; foreign exchange shortage; incentive contracts; international debt crisis; liquidity constraints; North–South economic relations; planning; price discrimination; reputation; social networks; social norms; soft budget constraint; transfer of technology; trust, lack of; virtual economy

Article

Countertrade is a commercial transaction in which a seller, typically from an industrialized country, supplies goods, services or technology to a buyer in a developing country or a formerly planned economy, and in which, in return, the seller purchases from the buyer an agreed amount of goods, services or technology. A distinctive feature of countertrade is the existence of a link between the two transactions, the original import in the developing country and the subsequent export.
Countertrade takes a variety of forms. The three most commonly distinguished are ‘barter’, ‘counterpurchase’ and ‘buyback’. Barter in the strict sense of the word refers to an import that is paid entirely or partly with an export from the importing country without using foreign exchange.

Counterpurchase refers to a transaction in which the import is paid with foreign exchange, but the industrialized country commits to buy export goods from the developing country in return. Buyback is a transaction in which the seller supplies a production facility and the parties agree that the supplier of the facility will buy goods produced with that production facility. All three forms of countertrade are frequently observed in international trade.

Under central planning, countertrade was especially observed in international trade among countries belonging to the Council for Mutual Economic Assistance (Comecon, an economic organization of communist states) as well as in East–West trade. In particular in the 1980s, in the aftermath of the international debt crisis, countertrade became prevalent in international trade with developing countries and Eastern Europe. Before 1989 countertrade accounted for up to 40 per cent of total trade between East and West. After 1989, with the domestic debt crisis in transition countries, barter became dominant in domestic trade in these countries. While countertrade continues to be significant in North–South trade, reliable estimates are not available.

Explanations for countertrade

One of the most frequently cited explanations of countertrade is that it allows countries to overcome a shortage of hard currency. The observation that countertrading countries are highly indebted is taken as evidence that these countries face a shortage of foreign exchange and that their low creditworthiness makes it impossible to finance imports with a simple loan from an international bank (for example, OECD, 1981; 1985). This interpretation is not fully plausible because countertrade uses export goods which otherwise could have been used to generate foreign exchange to pay for future imports. Furthermore, if the foreign-exchange shortage were the main explanation of countertrade we would expect barter to be the prevalent form of countertrade since only barter does in fact avoid the use of hard currency. However, barter accounts for only a small portion of total international countertrade (Marin and Schnitzer, 2002a). Mirus and Yeung (1987) find that countertrade in the form of simple barter or counterpurchase does not improve a country's foreign exchange position unless it improves economic efficiency in the sense that it leads to an increase in national income.

Empirical evidence points to another explanation, starting from the observation that in international trade contract enforcement is problematic and hence conventional contracts cannot be relied on as the main mechanism to sustain economic exchange. International countertrade (as well as domestic barter, as pointed out below) can be explained as an institutional response to such contractual problems arising in imperfect capital and goods markets. Difficulties in contract enforcement are an important impediment to international transactions in the world economy. In international trade, national sovereignty interferes with contract enforcement because national borders demarcate national jurisdictions. Such demarcations segment markets and impose severe transaction costs on exchanges across national jurisdictions. The hazards involved in international transactions are often disregarded, but they make headlines each time a sovereign debtor threatens to stop servicing its debt, as it happened in the international debt crisis in the 1980s or in the Russian financial crisis in 1998.

If contract enforcement is weak, problems may arise on both ends of a business transaction: the seller may fail to deliver the goods, and the buyer may fail to pay for them. If buyers have no cash to pay, and thus face liquidity constraints at the time of delivery, the business transaction can take place only if the seller can trust the buyer to pay in due course. On the other hand, the buyer is willing to engage in a

business transaction only if she can trust the seller to deliver the right goods. Both problems are prevalent in international trade. Enforcing the payment of goods can pose serious problems. In the aftermath of the debt crisis, highly indebted countries were liquidity constrained and could not finance necessary imports. Given their level of indebtedness, debt repayment could not be relied on. The debtor country could create more liquidity by not repaying its debt rather than by receiving a new loan. There are also important problems arising on the seller's side. In international trade, the most conspicuous example is the technology transfer problem. It is often reported that explicit contracts cannot be relied on to make sure that developing countries receive the advanced technology promised (Parsons, 1987; Kogut, 1986). These countries often complain that firms from industrialized countries sell inferior technology to them, technology that is outdated and cannot be sold on Western markets.

Solving the creditworthiness problem

Countertrade can be interpreted as the institutional response to the lack of creditworthiness of countries and firms. Countertrade introduces a deal-specific collateral for the credit granted for the original import. This collateral protects the interests of the creditor for one particular business transaction and thus mitigates the contractual hazards associated with indebtedness that would otherwise prevent the transaction from taking place.

The argument that payments in kind may have advantages over payments in cash contradicts the conventional wisdom in the theory of money. The common view is that barter is inefficient because it does not overcome the ‘double coincidence of wants problem’ (where each trading partner wants to buy exactly what the other partner wants to sell and vice versa) as money does. A seller may need to accept goods for which she has no use herself. The point, however, is that goods have superior credit enforcement properties to those of money. Money is an anonymous medium of exchange. This anonymity can prove disadvantageous in trade with countries which lack creditworthiness, since the debtor in the developing country or eastern Europe can use it for purposes other than repaying debt. Goods, by contrast, can more easily be earmarked as the property of the creditor and can thus serve as collateral. However, payment in goods is problematic if it is difficult to judge the quality of the goods offered as means of payment. Thus, it is important to choose goods that are very liquid and hardly anonymous, making it both easy to determine their value and easy to earmark. Goods can be ranked with respect to their liquidity and anonymity properties, providing an explanation for the export pattern of countertrade and barter (Marin and Schnitzer, 2002b).

Solving the technology transfer problem

Buyback contracts have been interpreted as incentive contracts that ensure the transfer of desirable quality technology and post-installation service performance if standard forms of internalization, like joint ventures or foreign direct investment, are not possible due to political and ownership constraints (Hennart, 1989; Chan and Hoy, 1991; Mirus and Yeung, 1993). But for the argument to work, it is essential that there be a technological relation between the two goods to be traded. However, buyback accounts for a surprisingly small fraction of all countertrade transactions. Thus, even though this explanation is theoretically appealing, it cannot explain the great majority of technology imports, which

take the form of counterpurchase.

Interestingly, the technology transfer problem can be solved with a simple counterpurchase transaction as well, despite the lack of a technological link (Marin and Schnitzer, 1995). Although the lack of liquidity makes it difficult to finance imports, it is this very lack of liquidity that can actually help when it comes to dealing with problems on the supplier's side. The idea is that the export from the developing country serves as a hostage that deters cheating on technology quality and defaulting on the payment of the original import from the industrialized country. For this mechanism to work, the export has to be profitable to both the industrialized country firm and developing country, and the contract is so designed that the export becomes sufficiently less profitable for either party that does not fulfil its obligations in the original import, be it technology transfer or payment of the import. The technology seller offers high-quality technology because otherwise she loses her collateral for the credit as the developing country firm would lack the revenues that are generated with the technology and that are necessary to produce the export goods. This contractual arrangement makes the technology supplier internalize the externality her technology imposes on the developing country. The developing country party will deliver the export goods because the terms of the contract are designed such that this is more profitable than selling them otherwise. So although the import and the export are not technologically related, the countertrade contract establishes a financial link that improves the incentives of the parties involved. Thus, countertrade is a first-best substitute for foreign direct investment when these countries are reluctant to give access to foreign ownership in their markets. This goes to prove that, in an imperfect world in which contract enforcement is weak, as in developing countries or imperfect capital markets, something that seems to be worse – that contractual problems arise on both sides of the business transaction rather than on only one side – can improve contract enforcement. In international trade, the liquidity constraint helps to solve the technology transfer problem.

Other explanations

Some other possible explanations of countertrade are that developing countries use countertrade transactions to promote the export of ‘new’ goods – goods they have not previously exported to industrialized countries – in order to gain access to new markets and to diversify their exports (OECD, 1981; 1985). The empirical evidence gives some support for the view that countertrade has helped to stimulate and diversify exports. Other studies confirm that the goods exported by developing countries through countertrade arrangements are often goods for which export markets have yet to be established. Readily marketable products, like raw materials, are usually not available for countertrade. It can also be observed that a country removes goods from the countertrade shopping list once it has gained some experience with exporting these particular goods (Banks, 1983). Furthermore, it has been argued that countertrade corrects distortions in non-competitive markets (Caves, 1974). Using barter may allow competing more aggressively without openly violating collusive agreements. It may also allow more effective price discrimination. There is indeed some evidence that barter is used as a vehicle to change the terms of trade to allow price discrimination by Western monopolists (Caves and Marin, 1992). Mandated countertrade has also been discussed as a policy response to contracting failures arising from asymmetric information about goods valuations (Ellingson and Stole, 1996).

Barter trade in transition economies

Barter trade has received renewed attention in the 1990s, when it became a dominant phenomenon in domestic trade in a number of transition countries, most notably in the successor states of the former Soviet Union. After 1989, domestic barter in Russia increased manifold after macroeconomic stabilization in 1994, from five per cent of GDP to 60 per cent in 1998. In Ukraine, the share of barter in industrial sales is estimated to have been more than 50 per cent in 1997. Only since the financial crisis in August 1998 have barter and the use of other money surrogates started to decline again.

A number of different explanations have been put forward for this phenomenon. Some experts have viewed it as a tax-avoidance mechanism because it allows a distortion of the true value of profits, and thus reduces tax liabilities. Furthermore, since the banking sector acts as a tax collection agency that transfers firms' cash income in bank accounts to the state to pay for outstanding tax arrears, barter allows tax avoidance because it avoids payments in cash. While there may be some truth in this kind of argument, few firms report tax advantages as a major reason for using barter (Marin and Schnitzer, 2002a).

A more popular explanation refers to soft budget constraints and the lack of market discipline. The absence of hard budget constraints, so the argument goes, leads managers and workers to avoid the costs arising from restructuring by maintaining production in inefficient activities. Barter would allow concealing the true market value of output. But the empirical evidence suggests that barter is not a phenomenon of state-owned enterprises. Newly established private firms display an exposure to barter that is similar to or greater than that of state-owned firms or cooperatives (Marin and Schnitzer, 2002a).

The ‘virtual economy’ argument of Gaddy and Ickes (1998) has been one of the most influential explanations of barter in Russia. The virtual economy argument claims that barter helps to create the image that the manufacturing sector in Russia is producing value while in fact it is not. This argument rests on the assumption that the manufacturing sector is value-subtracting, and most participants in the economy have an interest to pretend that it is not. Barter allows the parties to keep up this illusion by allowing the manufacturing sector to sell its output at a higher price than its market value and the value-adding natural resource sector (Gazprom) to accept this high price because of a lack of other sources. This way the manufacturing sector survives by drawing resources from the natural resource sector. According to the argument, keeping up the illusion of a value-adding manufacturing sector is highly costly for the Russian economy at large because this cross-subsidizing from the value-adding natural resource sector to the value-subtracting manufacturing sector prevents the manufacturing sector from moving into valuable activity.

This argument appeals to experts of central planning and policy observers in transition economies, because the practice of cross-subsidizing across different activities in the economy was a widespread feature of central planning. But it raises a number of questions. If the natural resource sector is producing valuable output, why does the sector not have other opportunities than to subsidize the manufacturing sector? In fact, the natural resource sector is supposed to have significant bargaining power in the interaction with other sectors when it is producing goods which the market values highly. Why then does the sector end up subsidizing the rest of the economy? And in fact, evidence from barter transactions in the Ukraine suggests that, in contrast to the assertions of the virtual economy proponents, the electricity and gas industries in the natural resource sector gained from barter transactions, instead of

losing (Marin, 2002).

A more plausible explanation refers to the similarities between barter in international trade and barter in transition economies, and links the surge of domestic barter to a ‘lack of trust’ problem (Marin and Schnitzer, 2005). In transition countries, poorly developed legal and financial institutions made contract enforcement unreliable and imposed severe transaction costs on any economic activity. These costs became prohibitively large in times of historic change and revolution. Unstable business partner relationships and rapidly changing social norms limited the extent to which economic exchanges could be sustained by reputation, by repeated interactions or by embedding them in social networks. This led to a lack of trust, meaning that reliable input supplies on the one hand and credit enforcement on the other hand were difficult to sustain, resulting in economic disorganization and a tremendous output fall. In such an environment, barter can be used as a commitment device to overcome the problems of unreliable input supplies and credit enforcement, by linking transactions and specifying terms of trade that give the right incentives to adhere to the terms of the barter contract.

See Also

- barter
- international trade theory
- planning
- third world debt
- transfer of technology

Bibliography

Banks, G. 1983. The economics and politics of countertrade. *World Economy* 6, 159–82.

Caves, R. 1974. The economics of reciprocity: theory and evidence on bilateral trading arrangements. In *International Trade and Finance: Essays in Honour of Jan Tinbergen*, ed. W. Sellekaerts. London: Macmillan.

Caves, R. and Marin, D. 1992. Countertrade transactions: theory and evidence. *Economic Journal* 102, 1171–83.

Chan, R. and Hoy, M. 1991. East–West joint ventures and buyback contracts. *Journal of International Economics* 30, 331–43.

Ellingson, T. and Stole, L.A. 1996. Mandated countertrade as a strategic commitment. *Journal of International Economics* 40, 67–84.

Gaddy, C.G. and Ickes, B.W. 1998. Russia's virtual economy. *Foreign Affairs* 77, 53–67.

Hennart, J.-F. 1989. The transaction-cost rationale for countertrade. *Journal of Law, Economics and Organization* 5, 127–53.

Kogut, B. 1986. On designing contracts to enforce contractibility: theory and evidence from East–West trade. *Journal of International Business Studies* 17, 47–61.

Marin, D. 2002. Trust versus illusion: what is driving demonetization in Russia? *Economics of Transition* 10, 173–200.

Marin, D. and Schnitzer, M. 1995. Tying trade flows: a theory of countertrade with evidence. *American Economic Review* 85, 1047–64.

Marin, D. and Schnitzer, M. 2002a. *Contracts in Trade and Transition: The Resurgence of Barter*. Cambridge: MIT Press.

Marin, D. and Schnitzer, M. 2002b. The economic institution of international barter. *Economic Journal* 112, 293–316.

Marin, D. and Schnitzer, M. 2005. Disorganization and financial collapse. *European Economic Review* 47, 387–408.

Mirus, R. and Yeung, B. 1987. Countertrade and foreign exchange shortages: a preliminary assessment. *Weltwirtschaftliches Archiv* 123, 535–44.

Mirus, R. and Yeung, B. 1993. Why countertrade? An economic perspective. *International Trade Journal* 7, 409–33.

OECD (Organization for Economic Cooperation and Development). 1981. *East–West Trade: Recent Developments in Countertrade*. Paris: OECD.

OECD. 1985. *Countertrade: Developing Countries Practices*. Paris: OECD.

Parsons, J.E. 1987. Forms of GDR economic cooperation with the nonsocialist countries. *Comparative Economic Studies* 29, 7–18.

How to cite this article

Marin, Dalia and Monika Schnitzer. "countertrade." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000407> doi:10.1057/9780230226203.0332

The New Palgrave Dictionary of Economics Online

countervailing power

Christopher M. Snyder
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

antitrust; countervailing power; mergers; monopoly; oligopoly; pharmaceutical industry; returns to scale

Article

‘Countervailing power’ is a term coined by J.K. Galbraith (1952) to describe the ability of large buyers in concentrated downstream markets to extract price concessions from suppliers. Galbraith saw countervailing power as an important force offsetting suppliers' increased market power arising from the general trend of increased concentration in US industries. He provided examples such as a nationwide grocery chain extracting wholesale price discounts from food producers, and large auto manufacturers extracting price discounts from steel producers.

The concept of countervailing power was controversial in Galbraith's day (see Stigler's, 1954, criticism), and continues to be so today. Formalizing the concept is difficult because it is difficult to model bilateral monopoly or oligopoly, and there exists no single canonical model. Whether and how wholesale discounts to large downstream firms are passed through to final-good consumers is unclear. The concept has the controversial antitrust implication that horizontal mergers between downstream firms may be pro-competitive.

There are a number of theories explaining why large buyers obtain price discounts from sellers. A simple theory is that the cost of serving large buyers is lower per unit than that of serving small buyers. Serving large buyers may involve lower distribution costs. For example, the supplier may be able to ship its product to a large buyer's central warehouse rather than having to ship it to the individual retail outlets owned by small buyers. Serving large buyers may also involve lower production costs. For example, if the supplier's production function exhibits increasing returns to scale and the supplier serves one buyer at a time each production period, per-unit production costs will be lower when serving a large buyer.

Other theories involve more subtle strategic effects. A literature including Horn and Wolinsky (1986), Stole and Zwiebel (1996), Chipty and Snyder (1999), Inderst and Wey (2003) and Raskovich (2003) considers a model in which a monopoly supplier bargains under symmetric information separately and

simultaneously with each of a number of buyers. Each buyer regards itself as marginal, conjecturing that all other buyers consummate their negotiations with the supplier efficiently. If aggregate surplus across all negotiations is concave in quantity, the marginal surplus from a transaction involving a large quantity is higher per unit than that from one involving a small quantity. This higher per-unit marginal surplus for large buyers translates into a lower per-unit price. The aggregate surplus function would be concave, for example, if the supplier has increasing marginal production costs. Even if the supplier's cost function were linear, the total surplus function effectively becomes concave if the supplier is assumed to be risk averse, as in Chae and Heidhues (2004) and DeGraba (2005).

Size discounts also emerge if large buyers' outside options are better. In Katz (1987) and Sheffman and Spiller (1992), for example, the larger the buyer, the more credible are its threat of integrating backward and producing the good itself. Size discounts also emerge if the supplier's outside option is worse when facing a large buyer. In Inderst and Wey (2007), for example, if bargaining with a large buyer breaks down, it is difficult for the supplier to unload this large quantity on the other buyers since this involves marching down these other buyers' declining marginal surplus functions.

Size discounts also emerge if one departs from the bargaining model with a monopoly supplier and instead considers competing suppliers. In Snyder (1998), collusion is difficult to sustain in the presence of a larger buyer because the benefit from undercutting and supplying the buyer is greater. To prevent undercutting in equilibrium, suppliers collude on a lower price for large buyers. In Dana (2004) and Inderst and Shaffer (2007), by pooling their demands and buying as a group from one supplier, buyers can increase the intensity of competition among suppliers of differentiated products.

Several papers have begun to examine the question of whether a downstream firm's countervailing power translates into lower final-good prices, using a model with competing downstream firms (Dobson and Waterson, 1997; von Ungern-Sternberg, 1996; Chen, 2003). This work suggests that an increase in countervailing power can have the opposite effect, raising consumer prices and/or lowering social welfare.

Early empirical studies of countervailing power (see Scherer and Ross, 1990, for a survey) took the standard structure–conduct–performance regressions (regressions of supplier profits or markups on supplier concentration using cross-sectional observations at the industry level) and added a buyer-concentration variable, often finding a significantly negative coefficient. Later intra-industry studies found more nuanced circumstances under which buyer-size discounts emerge. Ellison and Snyder (2002) and Sorensen (2003) observed size discounts in pharmaceutical and hospital-services markets only if there were competing, not monopoly, suppliers. In an experimental study, Normann, Ruffle and Snyder (2007) observed buyer-size discounts only when the total surplus function exhibited a certain curvature, consistent with theory.

See Also

- bargaining
- Galbraith, John Kenneth
- monopsony

- price discrimination (theory)

Bibliography

Chae, S. and Heidhues, P. 2004. Buyers' alliances for bargaining power. *Journal of Economics and Management Strategy* 13, 731–54.

Chen, Z. 2003. Dominant retailers and the countervailing power hypothesis. *RAND Journal of Economics* 34, 612–25.

Chipty, T. and Snyder, C.M. 1999. The role of firm size in bilateral bargaining: a study of the cable television industry. *Review of Economics and Statistics* 81, 326–40.

Dana, J. 2004. Buyer groups as strategic commitments. Mimeo, Northwestern University.

DeGraba, P. 2005. Quantity discounts from risk averse sellers. Working Paper No. 276, U.S. Federal Trade Commission.

Dobson, P.W. and Waterson, M. 1997. Countervailing power and consumer prices. *Economic Journal* 107, 418–30.

Ellison, S.F. and Snyder, C.M. 2002. Countervailing power in wholesale pharmaceuticals. Mimeo, MIT.

Galbraith, J.K. 1952. *American Capitalism: The Concept of Countervailing Power*. Boston: Houghton Mifflin.

Horn, H. and Wolinsky, A. 1986. Bilateral monopolies and incentive for merger. *RAND Journal of Economics* 19, 408–19.

Inderst, R. and Shaffer, G. 2007. Retail mergers, buyer power, and product variety. *Economic Journal* 117, 45–67.

Inderst, R. and Wey, C. 2003. Bargaining, mergers, and technology choice in bilaterally oligopolistic industries. *RAND Journal of Economics* 34, 1–19.

Inderst, R. and Wey, C. 2007. Buyer power and supplier incentives. *European Economic Review* 51, 647–67.

Katz, M.L. 1987. The welfare effects of third degree price discrimination in intermediate goods markets. *American Economic Review* 77, 154–67.

Normann, H.-T., Ruffle, B.J. and Snyder, C.M. 2007. Do buyer-size discounts depend on the curvature of the surplus function? Experimental tests of bargaining models. *RAND Journal of Economics*.

Raskovich, A. 2003. Pivotal buyers and bargaining position. *Journal of Industrial Economics* 51, 405–26.

Scherer, F.M. and Ross, D. 1990. *Industrial Market Structure and Economic Performance*. Boston: Houghton Mifflin.

Sheffman, D.T. and Spiller, P.T. 1992. Buyers' strategies, entry barriers, and competition. *Economic Inquiry* 30, 418–36.

Snyder, C.M. 1998. Why do large buyers pay lower prices? Intense supplier competition. *Economics Letters* 58, 205–9.

Sorensen, A. 2003. Insurer-hospital bargaining: negotiated discounts in post-deregulation Connecticut. *Journal of Industrial Economics* 51, 471–92.

Stigler, G.J. 1954. The economist plays with blocs. *American Economic Review* 44, 7–14.

Stole, L.A. and Zwiebel, J. 1996. Organizational design and technology choice under intrafirm bargaining. *American Economic Review* 86, 88–102.

von Ungern-Sternberg, T. 1996. Countervailing power revisited. *International Journal of Industrial Organization* 14, 507–20.

How to cite this article

Snyder, Christopher M. "countervailing power." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000538> doi:10.1057/9780230226203.0333

The New Palgrave Dictionary of Economics Online

Courcelle-Seneuil, Jean Gustave (1813 – 1892)

Albert O. Hirschman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advisers; convertibility; Courcelle-Seneuil, J. G.; deregulation of banks; tariffs

Article

French economist and economic adviser. Born in the Dordogne, he studied law in Paris, then returned to his native region to manage an industrial firm. At the same time, during the July monarchy, he wrote for Republican newspapers and economic periodicals. After the 1848 revolution, he held briefly a high position in the Ministry of Finance. In the following years he became a frequent contributor to the *Journal des économistes*, and published a successful textbook on banking in 1852. In 1853, the Chilean government contracted him to teach economics at the University of Chile in Santiago, and to be available as official economic adviser; he stayed for ten years, until 1863, when he returned to France. While in Chile he published his most ambitious work in economics, the *Traité théorique et pratique d'économie politique* (1858), which the Chilean government arranged to bring out in a Spanish translation. After his return to France, he resumed his activity as prolific writer of books and articles on economic affairs. He also published several works on political and historical topics and translated into French John Stuart Mill's *Principles of Political Economy*, Summer Maine's *Ancient Law* and William Graham Sumner's *What Social Classes Owe to Each Other*. He was appointed councillor of state in 1879, and three years later was elected member of the Académie des Sciences Morales et Politiques. Throughout his life, Courcelle-Seneuil was a stalwart defender of free trade and laissez-faire. Charles Gide, the co-author (with Charles Rist) of a well-known history of economic doctrines, wrote about him in rather sarcastic terms:

> He was virtually the *pontifex maximus* of the classical school; the holy doctrines were entrusted to him and it was his vocation to denounce and exterminate the heretics. During many years he fulfilled this mission through book reviews in the *Journal des économistes* with priestly dignity. Argus-eyed, he knew how to detect the slightest deviations from the liberal school. (Gide, 1895, p. 710)

Courcelle-Seneuil's special interest, starting with the publication of a small book on bank reform in 1840, was the introduction of more freedom into banking or, to use a modern term, the ‘deregulation’ of this industry. Above all, he advocated the abolition of the Bank of France's exclusive right of issue. According to Gide, Courcelle-Seneuil was more esteemed in England and the United States than in France. In any event, adoption of his monetary and banking proposals was never seriously considered in his own country.

Once in Chile, Courceile-Seneuil became a powerful policymaker and influential teacher. He arrived at a time when the international prestige of the laissez-faire doctrine was at its height and when gold booms and subsequent busts in California and Australia caused considerable fluctuations in Chile's agricultural exports to these areas, creating a need for flexible short- and long-term credit facilities. This combination of events, joined with the prestige emanating from the foreign savant, permitted him to obtain in Chile what he had failed to achieve in his own country: under his guidance, the administration of Manuel Montt (1851–61) promulgated a banking law that established total freedom for any solvent person to found a bank and permitted all banks to issue currency subject only to one limitation: the banknotes in circulation were not to exceed 150 per cent of the issuing bank's capital.

Courcelle-Seneuil's advice was also sought in connection with a new customs tariff and here again he achieved substantial change: the level of protection was severely cut back, although some tariffs were retained for revenue purposes.

But the principal influence exercised by Courcelle-Seneuil resided in his forceful teaching: as the University of Chile's first professor of economics, he was apparently successful in instilling doctrinaire zeal in his students, some of whom later became influential policymakers. Thus, Chilean historians have not only traced the abandonment of convertibility in 1878 to the permissiveness of the 1860 Banking Law and the lack of industrial development to the 1864 tariff; they also see Courcelle-Seneuil's indirect influence in the acquisition of the nitrate mines of Tarapacá by private foreign interests after Chile's victory over Peru in the War of the Pacific (1882) had given it title to the mines. Alienation of the mines was indeed recommended by a government committee dominated by Courcelle-Seneuil's disciples, who felt, like their teacher, that state ownership and management of business enterprises was to be strictly shunned. Secular inflation, industrial backwardness, domination of the country's principal natural resources by foreigners – all of these protracted ills of the Chilean economy have been attributed to the French expert.

Since the economically advanced countries were also those where economic science first flourished, they soon produced a peculiar export product: the foreign economic expert or adviser. Courcelle-Seneuil is probably the earliest prototype of the genre and his ironic career in Chile exhibits characteristics that were to remain typical of numerous later representatives. First, the adviser is deeply convinced that, thanks to the advances of economic science, he knows the correct solutions to economic problems no matter they may arise. Secondly, the country which invites the expert looks forward to his advice as to some magic medicine which will work even when (perhaps especially when) it hurts. Some countries seem particularly prone to this attitude. In Chile foreign or foreign-trained experts have played key roles at crisis junctures, from Courcelle-Seneuil in the mid-19th century to Edwin Kemmerer in the 1920s, the Klein–Saks Mission in the 1950s, and finally to the ‘Chicago boys’ in the 1970s. Thirdly, the influence of the adviser derives not only from the intrinsic value and persuasiveness of his message, but from the fact that he usually has good connections in his home country and can therefore facilitate access to its

capital market. Courcelle-Seneuil, for example, suspended his university courses in 1858–9 to accompany a Chilean financial mission that travelled to France in search of a railroad construction loan. Fourthly, the foreign adviser is often criticized for wishing to transplant the institutions of his own country to the country he advises, but his real ambition is more extravagant: it is to endow the country with those ideal institutions which exist in his mind only, for he has been unable to persuade his own countrymen to adopt them. Fifthly, history in general, and nationalist historiography in particular, is likely to be unkind to the foreign adviser. In retrospect he can easily become a universal scapegoat: whatever went wrong is attributed to his nefarious influence. This demonization is more damaging than the adviser himself could possible have been: it forestalls authentic learning from past experience.

Selected works

1840. *Le crédit et la Banque*. Paris.

1858. *Traité théorique et pratique d'économie politique*, 2 vols. Paris: Amyot.

1867. *La Banque libre*. Paris: Guillaumin.

Bibliography

Encina, F. 1951. *Historia de Chile*, vol. 18, ch. 58. Santiago: Nascimiento.

Fuentealba, H.L. 1946. *Courcelle-Seneuil en Chile: errores del liberalismo económico*. Santiago: Prensas de la Universidad de Chile.

Gide, C. 1895. Die neuere volkswirtschaftliche Litteratur Frankreichs. *Schmollers Jahrbuch*.

Hirschman, A.O. 1963. *Journeys toward Progress*. New York: Twentieth Century Fund, 163–8.

Journal des économistes. 1892. Obituary [of M.J.G. Courcelle-Seneuil]. July.

Juglar, C. 1895. Notice sur la vie et les travaux de M.J.G. Courcelle-Seneuil. Académie des Sciences Morales et Politiques, *Compte Rendu*, 850–82.

Pinto, S.C. 1959. *Chile, un caso de desarrollo frustrado*. Santiago: Edit. Universitaria.

Will, R.M. 1964. The introduction of classical economics into Chile. *Hispanic-American Historical Review* 44(1), 1–21.

How to cite this article

Hirschman, Albert O. "Courcelle-Seneuil, Jean Gustave (1813–1892)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000412>
doi:10.1057/9780230226203.0334

The New Palgrave Dictionary of Economics Online

Cournot competition

Andrew F. Daughety

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Cournot's 1838 model of strategic interaction between competing firms has become the primary workhorse for the analysis of imperfect competition, and shows up in a variety of fields, notably industrial organization and international trade. This article begins with a tour of the basic Cournot model and its properties, touching on existence, uniqueness, stability, and efficiency; this discussion especially emphasizes considerations involved in using the Cournot model in multi-stage applications. A discussion of recent applications is provided as well as a reference to an extended bibliography of approximately 125 selected publications from 2001 through 2005.

Keywords

auctions with competition; Bertrand competition; best-reply dynamics; best-response functions; complementarities; conjectural variations; Cournot competition; Cournot equilibrium; Cournot models; Cournot, A. A.; differentiated products; dynamic stability; efficiency; existence theorems; imperfect competition; information sharing among firms; licences; market power; mergers; multiple equilibria; multi-stage games; multi-stage models of competition; networks; oligopoly; patents; product differentiation; repeated games; research and development; signalling; strategic substitutes; subgame perfection; supermodular games; uniqueness; welfare

Article

The classic Cournot model is static in nature, with each (single-product) firm's strategy being the quantity of output it will produce in the market for a specific homogeneous good; as Kreps (1987) observed, Cournot's model was an early progenitor of Nash's famous paper. Many recent applications have involved multi-stage games; for example, each of *n* firms might first simultaneously choose investment levels (say, in cost-reducing R&D) and then simultaneously choose output levels in the second stage. Often now used in such a manner, we will see that the Cournot model is doing well,

contributing to a range of new research, as it moves towards the two-century mark.

1 The basic one-stage model and associated concepts

Consider an industry comprised of *n* firms, each firm choosing an amount of output to produce. Firm *i*'s *output level* is denoted as q_i, $i = 1, \ldots, n$; let the vector of firm outputs be denoted $\boldsymbol{q} \equiv (q_1, q_2, \ldots, q_n)$. The firms' products are assumed to be perfect substitutes (the *homogeneous-goods* case); let *Q* denote the aggregate industry output level (that is, $Q \equiv \sum_{i=1}^{n} q_i$). We will refer to the $(n-1)$ vector of output levels chosen by firm *i*'s rivals as $\boldsymbol{q}_{-i}$; so, let $(\boldsymbol{q}_{-i}, q_i)$ also be the *n*-vector $\boldsymbol{q}$. Market demand for the perfect-substitutes case is a function of aggregate output and its inverse is denoted as $p(Q)$; furthermore, let firm *i*'s cost of producing q_i be denoted as $c_i(q_i)$. Thus, firm *i*'s *profit* function is written as $\pi^i(\boldsymbol{q}) \equiv p(Q)q_i - c_i(q_i)$. All elements of the model are assumed to be commonly known by the firms, though extensions allowing incomplete information are not uncommon.

A *Cournot equilibrium* consists of a vector of output levels, $\boldsymbol{q}^{CE}$, such that no firm wishes to unilaterally change its output level when the other firms produce the output levels assigned to them in the (purported) equilibrium. Alternatively put (and reversing history), it is a Nash equilibrium of the normal-form game with quantities as strategies chosen from a compact space (for example, q_i in $[0, Q^*]$, for some appropriate Q^*, such as $p(Q^*) = 0$) and with the $\pi^i(\boldsymbol{q})$ as the payoff functions. Thus, $\boldsymbol{q}^{CE}$ is a Cournot equilibrium if the following n equations are satisfied: $\pi^i(\boldsymbol{q}^{CE}) \geq \pi^i(\boldsymbol{q}^{CE}_{-i}, q_i)$ for all values of q_i, for $i = 1, \ldots, n$.

In analysing his model applied to a duopoly (he also considered the *n*-firm version), Cournot provided the notion of *best-response functions*. In the duopoly case, this is a pair of functions, $\psi^1(q_2)$ and $\psi^2(q_1)$, which provide the profit-maximizing choice of output for firm 1 and 2 (respectively), given conjectures about the output level chosen by the rival firm (that is, each firm's choice of its output level reflects a *best-response property*). Hence, $\psi^i(q_j) = \arg\max_q \pi^i(q, q_j), \ i, j = 1, 2, \ i \neq j$. That is, we want $\psi^i(q_j)$ to be the solution to firm *i*'s first-order condition:

$p(\psi^i(q_j) + q_j) + p'(\psi^i(q_j) + q_j)\psi^i(q_j) - c_i'(\psi^i(q_j)) = 0, \ i, j = 1, 2, \ i \neq j$. We'll assume for now that the problem has a nice solution and that some sort of sufficiency condition holds (for example, strict quasi-concavity of profits), but the discussion below on existence and uniqueness of equilibrium shows that such classical assumptions are overly strong and are overly restrictive for some modern applications, such as those involving multi-stage games or discontinuous cost functions. More generally, $\psi^i(q_j)$ could be a correspondence (a point-to-set map); we generally restrict the discussion below to functions, and assume as much differentiability as needed.

If output-level choices are best responses to conjectures about each firm's rival's choice of output, *and* if these conjectures are correct in equilibrium, then the resulting vector of output levels provides a Cournot equilibrium: $q_i^{CE} = \psi^i(q_j^{CE})$ for $i, j = 1, 2,$ and $j \neq i$. In other words, the equilibrium occurs where the best-response functions cross when graphed in the space of output levels. Generalizing to *n* firms, this

condition can be written as $q_i^{CE} = \psi^i(\boldsymbol{q}_{-i}^{CE})$ for $i = 1, \ldots, n$: $\boldsymbol{q}^{CE}$ is a Cournot equilibrium if it consists of mutual best-responses for all the firms.

Some variations on the basic model are worth mentioning. If the cost function for a firm has both fixed and variable components, and if the fixed component is avoidable (that is, is zero at zero output), then the best-response function for the firm will be discontinuous at the positive output level where variable profits just cover the avoidable cost. This is important for two reasons. First, avoidable fixed costs are not unusual in many entry scenarios: think of an airline entering a market where there are already some competitors, with the avoidable cost being advertising. Second, this discontinuity could mean that the only equilibrium might involve some or all firms choosing to not enter (or to exit) the market, even if absent these avoidable costs $\boldsymbol{q}^{CE}$ would be strictly positive.

Another avenue for interaction would consider imperfect factor markets, so that instead of $c_i(q_i)$ the cost function for firm i would be written as $c_i(\boldsymbol{q}_{-i}, q_i)$; then strategic interaction occurs not only through revenue but also via factor markets. Finally, if the model is one of short-run competition, then the output level of the firm may be restricted to be less than some predetermined capacity level; a simple version is that there are parameters k_i, $i = 1, \ldots, n$, such that a constraint on firm i's quantity choice is $q_i \leq k_i$, $i = 1, \ldots, n$; this induces a vertical segment (at the capacity level) in a firm's best-response function. Such capacity levels might be choices made in an earlier stage.

Finally, a number of papers develop ‘non-Cournot’ models which generate Cournot-model results. Kreps and Scheinkman (1983) provide a two-stage model of capacity choice followed by price setting in a homogeneous-goods duopoly; the result is a unique subgame-perfect equilibrium with Cournot capacities and a market-clearing price consistent with the standard Cournot model (however, Davidson and Deneckere, 1986, show that this result is especially sensitive to the basis for rationing consumers over firms when out-of-equilibrium firm-level demand exceeds capacity). Klemperer and Meyer (1986) analyse a one-stage game wherein duopolists producing heterogeneous goods non-cooperatively choose either a price or a quantity as the firm's strategy; under either multiplicative or additive error in the demand function, if marginal costs are upward sloping, the outcome is that predicted by the Cournot model (applied to the heterogenous-goods case; see the discussion of this case in Section 2 below). The classic embedding of the Cournot model is that of Bowley (1924), the best-known developer of models with ‘conjectural variations’ (CV). This is a static story wherein the first-order conditions in the analysis include firm i's conjecture of each rival's reaction to a small change in firm i's quantity (for example, $\partial q_j / \partial q_i$ need not be zero for each $j \neq i$); different values of the CV generate competitive, collusive, or Cournot outcomes (among others). Such a handy static embedding of alternative degrees of competition has been employed in a number of theoretical applications, and in a variety of empirical analyses trying to estimate market power. However, Daughety (1985) shows that a basic rationality requirement (that each firm's CV be the same as the actual slope of the best-response function) leads to the Cournot outcome, so that alternative CV values violate this form of rational expectations. Furthermore, Korts (1999) shows that empirical analyses using the CV approach to assess market power will generally mis-measure the degree of competitiveness of the industry.

2 Properties of the Cournot equilibrium

For most of this section we emphasize results for an *n*-firm, homogeneous-goods, complete-information model, where a firm's cost function depends only on that firm's output level. As suggested earlier, possibly one of the most important reasons for the continuing interest in the properties of the Cournot equilibrium is that Cournot competition is frequently used as a final stage in a variety of models; analysis employing such refinements as subgame perfection rely on a well-behaved subgame.

Existence, uniqueness and stability

Novshek (1985) provides an existence theorem that has quite practical uses (for expository purposes we consider a slightly less general version). Besides continuity and twice differentiability of the inverse demand function, $p(Q)$, Novshek's existence theorem requires that: (1) $p(Q)$ crosses the quantity axis at a finite value and is strictly decreasing for quantities below that cut point; (2) the marginal revenue for each firm is decreasing in the aggregate output of its rivals; and (3) each firm's cost function is non-decreasing and lower semi-continuous. Requirement (2) is written formally as $p'(Q_{-i} + q_i) + p''(Q_{-i} + q_i)q_i < 0$, where $Q_{-i} \equiv Q - q_i$, for all *i*. This is equivalent to the assumption that $\partial^2 \pi^i(\mathbf{q}) / \partial Q_{-i} \partial q_i < 0$ for all *i*, that is, that Q_{-i} and q_i are *strategic substitutes*, which means that an expansion in Q_{-i} implies that the optimal q_i falls. The third requirement means that costs cannot fall as the output level is increased and that cost functions can have jumps (discontinuities), as long as the functions are continuous from the left. This was a substantial improvement over previous existence theorems and it allows for an important case: avoidable fixed costs, such as those in the airline-entry example mentioned earlier. Amir (1996) applies an ordinal version of the theory of *supermodular games* to the existence issue (see Vives, 2005, for a recent survey of supermodular games; see also Amir, 2005, for a comparison of *ordinal* and *cardinal complementarity* in this context); this change of techniques allows for weaker demand conditions (primarily that log $p(Q)$ is concave) but requires a slightly stronger condition on each firm's cost function (marginal costs are positive, so models wherein marginal costs might be zero – as might occur with capacity competition – are left out) in order to guarantee that a Cournot equilibrium exists. As an example of the advantages concerning demand analysis, let $p(Q) = (Q - \bar{Q})^2$ for $Q \leq \bar{Q}$, and zero otherwise. Such a function satisfies (1) above, is log-concave (actually, convex), but is excluded from consideration by Novshek's second condition.

Gaudet and Salant (1991) provide conditions for a Cournot equilibrium to be *unique* which address an important consideration when Cournot models are used in a subgame of a larger game: their theorem allows for degeneracy (one or more firms produce zero output but have marginal cost equal to the equilibrium price); thus, such firms are just at the shutdown point in the equilibrium. In a one-stage application this could be eliminated via a small perturbation in the parameters, but in a multi-stage application such an outcome need not be pathological, as some of the second-stage ‘parameters’ are strategic variables in the first-stage model (the authors provide a simple, full-information entry game to illustrate this). The sufficient conditions for uniqueness are (not surprisingly) more restrictive than those for existence (on the assumption that Novshek's conditions hold as well): (1) each firm's cost function must be twice continuously differentiable and strictly increasing; and (2) the slope of the marginal cost function is strictly bounded above the slope of the demand function. Thus, concave costs are allowed, to some degree, but the cost function cannot be ‘too concave’, even on subsets of its domain.

Cournot provided an explicit dynamic stability argument for his model by imagining sequential play by each agent (myopically best-responding in the current period to the existing output levels of all rivals); this is referred to as *best-reply dynamics* and when this process converges the solution is termed *stable*. Using best-reply dynamics to rationalize a static solution has, historically, been a source of substantial criticism, but nonetheless some papers use the requirement of Cournot stability to select an equilibrium when there are multiple equilibria (dynamic stability should not be confused with equilibrium refinement criteria in game theory such as strategic stability). A sufficient condition in the duopoly case is that $\left|\partial\psi^1(q_2)/\partial q_2 \| \partial\psi^2(q_1)/\partial q_1\right| < 1$ (see Fudenberg and Tirole, 1991); see Seade (1980) for more general conditions (and problems) for best-reply dynamics in the *n*-firm case. For an approach employing an explicit evolutionary process via replicator dynamics with noise, with firms able to choose ‘behavioural’ strategies (including, but not limited to, best-reply), see Droste, Hommes and Tunistra (2002).

Welfare

Two types of inefficiency can occur in a Cournot equilibrium: the equilibrium price exceeds the marginal cost of production, and aggregate output is inefficiently distributed over the firms. Compare the first-order conditions for firms in a duopoly, each producing under conditions of non-decreasing marginal costs (that is, $p(Q) + p'(Q)q_i = c_i'(q_i),\ i = 1, 2$) with those for a central planner choosing q_1 and q_2 so as to maximize total surplus: $p(Q) = c_i'(q_i),\ i = 1, 2$. Clearly, if demand is downward-sloping at the equilibrium, aggregate output in the Cournot equilibrium will be less than what the social planner would choose. However, a second distortion can be seen in this comparison: under the social planner, each firm's marginal costs are equalized with the others'. This will hold only in a symmetric Cournot equilibrium (where $q_1 = q_2$): production is, in general, inefficiently allocated across the firms.

The maldistribution of production implies that strategic interaction readily may yield counter-intuitive welfare results. As a simple example, consider a duopoly wherein (inverse) industry demand is $p = a - Q$ and firm *i*'s cost function is $c_i(q) = C_i q,\ i = 1, 2$, with $a > C_1 > C_2 > 0$; that is, the linear demand, constant-but-unequal-marginal-cost case. It is straightforward to find the equilibrium and show that it is interior and unique. Let *W* be the sum of producers' and consumers' surplus. Then a little work shows that $dW/dC_1 > 0$ if $11C_1 - 7C_2 - 4a > 0$; to see that these conditions are non-empty, consider the parameter specification ($a = 20,\ C_1 = 13,\ C_2 = 8$), which satisfies all the foregoing requirements. The point of the example is that a *reduction* of firm 1's marginal cost leads to a *decrease* in equilibrium welfare. Thus, strategic interaction by the firms in the marketplace can lead to reversals of the usual welfare intuition that cost-improving technological change is beneficial. The reason this occurs is that the cost reduction results in an increase in the high-cost firm's equilibrium output level and a (smaller) decrease in the low-cost firm's output level; this increased inefficiency in aggregate production can be sufficient to overwhelm other efficiency improvements (such as the increase in industry output). This is similarly true if in the above model firm 2 is an incumbent monopolist (using simple monopoly pricing) and firm 1 an entrant: welfare will fall due to entry.

In the *n*-firm version of the constant-marginal-cost model, changes in the distribution of production costs

(holding the mean fixed) do not affect industry output; this is seen by summing over the first-order conditions, whence $np(Q) + p'(Q)Q = \Sigma_{i=1}^{n} c_i$. Bergstrom and Varian (1985) showed that (on the assumption that the pre- and post-change equilibria are interior) such mean-preserving changes in the marginal costs strictly improve welfare if and only if the variance of the marginal costs strictly increases; the reason is that the aggregate cost of production has decreased if the variance increases. Salant and Shaffer (1999) extend this idea to consider the effects of changes in first-stage parameters (for example, cost-reducing R&D investments) on second-stage costs in models wherein Cournot competition is employed in the second stage. They argue that, since aggregate production costs are *maximized* when all firms have the same costs, it is the asymmetric equilibria in such games (which are often assumed away) which may yield the most important outcomes to examine, from both a social and a private perspective.

Does entry necessarily reduce the equilibrium price? A recent contribution provides a clean result if we restrict attention to the symmetric case wherein all firms have the same twice continuously differentiable and non-decreasing cost function, and demand is continuously differentiable and downward-sloping. Amir and Lambson (2000) show that the equilibrium price falls with an increase in the number of competitors if, for all Q, $p'(Q) < c''(q)$ for all q in [0, Q]. Thus, even with some degree of returns to scale (for example, as might occur with U-shaped average costs), entry will reduce price, at least with identical firms. However, Hoernig (2003) shows that, even if the equilibria are stable and there are no returns to scale, price can rise with entry if products are differentiated.

If the products of the firms are imperfect substitutes (that is, products are differentiated), then (in general) there is no aggregate demand function $p(Q)$; rather firm i's inverse demand function would be written as $p_i(\boldsymbol{q})$ and profits would be written as $\pi^i(\boldsymbol{q}) = p_i(\boldsymbol{q})q_i - c_i(q_i),\ i = 1, \ldots, n$. Welfare in this model can be contrasted with a reformulation of the model so that each firm chooses a price for its product; standard parlance is to call the price-strategy model the *(differentiated products) Bertrand model* (even though Bertrand's famous review of Cournot did not envision heterogeneity in products; see Friedman's 1988 translation of Bertrand's review). Without going into detail on the (differentiated products) Bertrand model, Singh and Vives (1984) have shown (for linear, symmetric demand and constant marginal costs in a duopoly setting) that, while profits under Cournot competition exceed those under Bertrand competition, total surplus is higher under Bertrand competition than under Cournot competition. Note that this result holds in the one-stage game. However, these results may be reversed in a two-stage application. For example, Symeonidis (2003) considers R&D investment with spillovers in a two-stage game, and shows that (at least for a portion of the parameter space) Cournot competition leads to higher welfare than Bertrand competition. The basic intuition is that, if profits are higher for second-stage Cournot competition than for second-stage Bertrand competition, and first-stage investment is inefficiently low in either case, then the increased second-stage profits may partly correct the inefficiently low first-stage investment, leading to an overall welfare gain for competition in quantities rather than prices.

Finally, convergence of a Cournot equilibrium to a competitive equilibrium, as the number of firms grows, was considered by Cournot in Chapter 8 of his book, and has been the subject of a number of papers; see Novshek and Sonnenschein (1978; 1987) for a general equilibrium treatment where appropriate replication of Cournot economies yields equilibria arbitrarily close to the Walrasian

equilibrium; see Alos-Ferrer (2004) for an evolutionary model (which allows for memory) at the level of an industry.

3 Applications

The literature exploring and applying the Cournot model is vast; an earlier extended bibliography can be found in Daughety (1988/2005). The more recent literature employing the Cournot model is already becoming significant in size: a survey of articles in 16 top mainline and field journals, for the period 2001–5, netted approximately 125 articles exploring or applying the Cournot model in one of its various common forms. An online Excel file of (abbreviated) citations and some characteristics of each article (number of firms, number of stages, welfare considerations, informational regime, and topic classification), as accessed on 21 November 2006, is available at http://www.vanderbilt.edu/Econ/faculty/Daughety/ExtendedCournotBib2001-2005.xls
However, some excellent papers have undoubtedly been missed (not to mention papers from the 1990s), and space limitations preclude anything beyond the briefest of tours and just a taste of the literature, so only a very few can be discussed below. This section addresses five topics which account for a significant portion of the literature, three areas that overlap other fields, and two (comparatively) new areas of research.

Delegation

Vickers (1985) uses an *n*-firm, two-stage model to examine performance measures for managers. Restricting the manager's performance measure to be a weighted average of profits and output, with the weights determined by the owner of each firm in the first stage, he shows that the weight on output is non-zero. This makes each manager more aggressive (each chooses to produce a higher output level), thereby leading to lower profits per firm. Sklivas (1987) considers the differentiated-products Bertrand version and shows that owners choose weights on revenue and profits so as to make managers more passive (they post higher prices), leading to increased profits. Miller and Pazgal (2001) have unified this literature, showing that incentive schemes based on own and rival's profits result in an equilibrium which is insensitive to whether the firm chooses price or quantity as its strategic variable.

Information transfer

Vives (1984), Gal-Or (1985), and Li (1985) all consider variants of ‘information transfer’ models to examine the possibility of information sharing, whereby firms may choose to pool information on either demand or cost parameters. These models are analysed as *Bayesian–Nash games*, so that, before seeing a private signal about the parameter of interest (for example, the demand intercept), each firm chooses whether or not to share the information with the other firms; then information is received and production (or pricing) occurs in the second stage. The nature of the good (substitutes or complements), the type of information (common or individual), and the strategy space (quantities or prices) all affect whether firms will share information. Ziv (1993) relaxes the verifiability of information and finds that firms will send

misleading information if they can; he then considers mechanisms for eliciting truthful messages.

Intellectual property

Katz and Shapiro (1985) and Kamien and Tauman (1986) consider the licensing of innovations in an oligopoly. Katz and Shapiro employ a three-stage duopoly game in which the innovation is developed, then a single license is auctioned, and then the firms compete. Kamien and Tauman use a two-stage, *n*-firm game with a posted price for the innovation (a fee or a royalty), followed by competition. More recently, Fauli-Oller and Sandonis (2003) consider optimal competition policy when considering licences as an alternative to merger. Anton and Yao (2004) allow for weak patent protection and consider how disclosure of information about an innovation (for example, through the patent application) can be a signalling device to influence competitors, but those same competitors may be able to employ the information to successfully use (infringe on) the patent; here small innovations are patented and substantial innovations are protected through secrecy.

Mergers

Salant, Switzer and Reynolds (1983) show that exogenously determined mergers of a subset of firms in the constant-marginal-cost set-up yields a problematical result: a sufficient condition for a merger to be unprofitable is that it involve less than 80 per cent of the industry, hardly a resounding endorsement of using such a model to analyse mergers. This result, however, is partly driven by the assumptions of homogeneous products, constant unit costs, and industry structure. Perry and Porter (1985) show that various mergers can be profitable if firms have sufficiently increasing marginal costs. Daughety (1990), using a two-tiered-industry, *n*-firm model, with *m* firms choosing output in the first stage (tier) and $n - m$ firms choosing output in the second stage, shows that if $1<m<n$, then, when *m* is comparatively small ($m<n/3$), mergers of two second-tier firms to make a first-tier firm can be both profitable and social-welfare-enhancing, even though such mergers increase concentration and have no cost synergies (all firms have identical constant unit costs). Recently, Pesendorfer (2005), using a repeated game model with entry, has found that merger to monopoly may not be profitable, but merger in a non-concentrated industry can be; these differences from the previous literature partly reflect long-run versus short-run profitability computations.

R&D

D'Aspremont and Jacquemin (1988) considered cost-reducing R&D in the presence of spillovers, and considered both non-cooperative and cooperative R&D decision-making; there have been a number of recent papers on cost-reducing spillovers (see, for example, Zhao, 2001, for more on the negative welfare effects of cost-reducing innovation, and Symeonidis, 2003, cited in Section 2 above, as well as the work discussed below under the subject of auctions with competition). Toshimitsu (2003) considers the incentive and welfare properties of quality-based R&D subsidies for firms in a model of endogenously determined product quality (and thus product differentiation); subsidizing high quality is

welfare-enhancing (independent of whether the Cournot or Bertrand model is employed).

Other fields

Areas of ongoing effort which extend into other fields include *experimental economics*, *the financial structure of the firm* (see, for example, Brander and Lewis, 1986, on determinate debt-equity due to imperfect competition, and see Povel and Raith, 2004, extending Brander and Lewis via endogenously determined debt contracts); and *international trade* (see, for example, Brander and Spencer, 1985, analysing the strategic use of subsidies in international competition; Mezzetti and Dinopoulos, 1991, discussing domestic firm–union bargaining and import competition; and Spencer and Qiu, 2001, concerning relationship-specific investments and trade).

New topics

Finally, a few examples of comparatively new topics. While auctions with private information has long been an area of interest, the developing literature on *auctions with competition* has started to take seriously the combination of incomplete information and post-auction competition. For example, see Das Varma (2003) or Goeree (2003), who find that signalling by winners of an auction causes bids to be biased when post-auction interaction between the auction's winner and losers can be influenced by the size of the bid. A nice example is when firms have private information about how acquiring a cost-reducing innovation might affect the firm's production costs, and bidding for a licence for the innovation precedes Cournot oligopoly interaction; here signalling with a high bid suggests that the winner will have low costs and will produce a high level of output.
A second new area is *networks*; one recent example is Goyal and Moraga-Gonzalez (2001), who model bilateral agreements to share knowledge, and allow for the possibility of partial collaboration, via considering possible networks of relationships. They examine how the nature of the firms' interaction in markets can contribute to the instability of certain types of strategic alliances and the stability of other ones.

4 A broader perspective on Cournot competition

If alive to critique this essay, Cournot might view the interpretation of the term ‘Cournot competition’ being limited merely to the legacy of his oligopoly analysis to be an overly restrictive interpretation of the assignment. And well he should. Hicks (1935; 1939) argues that Cournot was the first to present a modern model of monopoly as well as the precise conditions for perfect competition; furthermore, as noted earlier, Cournot's eighth chapter concerned ‘unlimited competition’. In the 1937 Cournot Memorial session of the Econometric Society, A. J. Nichol (1938) observed that, if ever there was an apt illustration of Carnegie's dictum that ‘It does not pay to pioneer’, then Cournot's life and work would be it. Cournot's oligopoly model was essentially ignored for many years, or was relegated to dusty corners of microeconomics texts, but over recent decades it has come to be an essential tool in many an economist's toolbox, and is likely to continue as such.

See Also

- Bertrand competition
- experimental economics

Bibliography

Alos-Ferrer, C. 2004. Cournot versus Walras in dynamic oligopolies with memory. *International Journal of Industrial Organization* 22, 193–217.

Amir, R. 1996. Cournot oligopoly and the theory of supermodular games. *Games and Economic Behavior* 15, 132–48.

Amir, R. 2005. Ordinal versus cardinal complementarity: the case of Cournot oligopoly. *Games and Economic Behavior* 53, 1–14.

Amir, R. and Lambson, V.E. 2000. On the effects of entry in Cournot markets. *Review of Economic Studies* 67, 235–54.

Anton, J.A. and Yao, D.A. 2004. Little patents and big secrets: managing intellectual property. *RAND Journal of Economics* 35, 1–22.

Bergstrom, T.C. and Varian, H.R. 1985. Two remarks on Cournot equilibria. *Economics Letters* 19, 5–8.

Bertrand, J. 1883. Review of Walras's *Théorie mathématique de la richesse social* and Cournot's *Recherches sur les principes mathématiques de la théorie des richesses*. Trans. J.W. Friedman, in A.F. Daughety (1988).

Bowley, A.L. 1924. *The Mathematical Groundwork of Economics*. Oxford: Oxford University Press.

Brander, J.A. and Lewis, T.R. 1986. Oligopoly and financial structure: the limited liability effect. *American Economic Review* 76, 956–70.

Brander, J.A. and Spencer, B. 1985. Export subsidies and international market share rivalry. *Journal of International Economics* 18, 83–100.

Cournot, A. 1838. *Researches into the Mathematical Principles of the Theory of Wealth*. Trans. N.T. Bacon, New York: Macmillan, 1929.

d'Aspremont, C. and Jacquemin, A. 1988. Cooperative and noncooperative R&D in duopoly with

spillovers. *American Economic Review* 78, 1133–7.

Das Varma, G. 2003. Bidding for a process innovation under alternative modes of competition. *International Journal of Industrial Organization* 21, 15–37.

Daughety, A.F. 1985. Reconsidering Cournot: the Cournot equilibrium is consistent. *RAND Journal of Economics* 16, 368–79.

Daughety, A.F. 1988. *Cournot Oligopoly – Characterization and Applications*. New York: Cambridge University Press (reprinted 2005).

Daughety, A.F. 1990. Beneficial concentration. *American Economic Review* 80, 1231–37.

Davidson, C. and Deneckere, R. 1986. Long-run competition in capacity, short-run competition in price, and the Cournot model. *RAND Journal of Economics* 17, 404–15.

Droste, E., Hommes, C. and Tunistra, J. 2002. Endogenous fluctuations under evolutionary pressure in Cournot competition. *Games and Economic Behavior* 40, 232–69.

Fauli-Oller, R. and Sandonis, J. 2003. To merge or to license: implications for competition policy. *International Journal of Industrial Organization* 21, 655–72.

Fudenberg, D. and Tirole, J. 1991. *Game Theory*. Cambridge, MA: MIT Press.

Gal-Or, E. 1985. Information transmission – Cournot and Bertrand. *Review of Economic Studies* 53, 85–92.

Gaudet, G. and Salant, S. 1991. Uniqueness of Cournot equilibrium: new results from old methods. *Review of Economic Studies* 58, 399–404.

Goeree, J.K. 2003. Bidding for the future: signaling in auctions with an aftermarket. *Journal of Economic Theory* 108, 345–64.

Goyal, S. and Moraga-Gonzalez, J.L. 2001. R&D networks. *RAND Journal of Economics* 32, 686–707.

Hicks, J.R. 1935. Annual survey of economic theory: the theory of monopoly. *Econometrica* 3, 1–12.

Hicks, J.R. 1939. *Value and Capital*, 2nd edn. London: Oxford University Press.

Hoernig, S.H. 2003. Existence of equilibrium and comparative statics in differentiated goods Cournot oligopolies. *International Journal of Industrial Organization* 21, 989–1019.

Kamien, M.I. and Tauman, Y. 1986. Fees versus royalties and the private value of a patent. *Quarterly Journal of Economics* 101, 471–92.

Katz, M.L. and Shapiro, C. 1985. On the licensing of innovations. *RAND Journal of Economics* 16, 504–20.

Klemperer, P. and Meyer, M. 1986. Price competition vs. quantity competition: the role of uncertainty. *RAND Journal of Economics* 17, 618–38.

Korts, K.S. 1999. Conduct parameters and the measurement of market power. *Journal of Econometrics* 88, 227–50.

Kreps, D.M. 1987. Nash equilibrium. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Kreps, D.M. and Scheinkman, J.A. 1983. Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell Journal of Economics* 14, 326–37.

Li, L. 1985. Cournot oligopoly with information sharing. *RAND Journal of Economics* 16, 521–36.

Mezzetti, C. and Dinopoulos, D. 1991. Domestic unionization and import competition. *Journal of International Economics* 31, 79–100.

Miller, N.H. and Pazgal, A.I. 2001. The equivalence of price and quantity competition with delegation. *RAND Journal of Economics* 32, 284–301.

Nichol, A.J. 1938. Tragedies in the life of Cournot. *Econometrica* 3, 193–7.

Novshek, W. 1985. On the existence of Cournot equilibrium. *Review of Economic Studies* 52, 85–98.

Novshek, W. and Sonnenschein, H. 1978. Cournot and Walras equilibrium. *Journal of Economic Theory* 19, 223–66.

Novshek, W. and Sonnenschein, H. 1987. General equilibrium with free entry. *Journal of Economic Literature* 25, 1281–306.

Perry, M.K. and Porter, R.H. 1985. Oligopoly and the incentive for horizontal merger. *American Economic Review* 75, 219–27.

Pesendorfer, M. 2005. Mergers under entry. *RAND Journal of Economics* 36, 661–79.

Povel, P. and Raith, M. 2004. Financial constraints and product market competition: ex ante vs. ex post incentives. *International Journal of Industrial Organization* 22, 917–49.

Salant, S.W. and Shaffer, G. 1999. Unequal treatment of identical agents in Cournot equilibrium. *American Economic Review* 89, 585–604.

Salant, S.W., Switzer, S. and Reynolds, R.J. 1983. Losses from horizontal merger: the effects of an exogenous change in industry structure on Cournot–Nash equilibrium. *Quarterly Journal of Economics* 98, 185–99.

Seade, J. 1980. The stability of Cournot revisited. *Journal of Economic Theory* 23, 15–27.

Singh, N. and Vives, X. 1984. Price and quantity competition in a differentiated duopoly. *RAND Journal of Economics* 15, 546–54.

Sklivas, S.D. 1987. The strategic choice of managerial incentives. *RAND Journal of Economics* 18, 452–58.

Spencer, B.J. and Qiu, L.D. 2001. Keiretsu and relationship-specific investment: a barrier to trade? *International Economic Review* 42, 871–901.

Symeonidis, G. 2003. Comparing Cournot and Bertrand equilibria in a differentiated duopoly with product R&D. *International Journal of Industrial Organization* 21, 39–53.

Toshimitsu, T. 2003. Optimal R&D policy and endogenous quality choice. *International Journal of Industrial Organization* 21, 1159–78.

Vickers, J. 1985. Delegation and the theory of the firm. *Economic Journal Supplement* 95, 138–47.

Vives, X. 1984. Duopoly information equilibrium: Cournot and Bertrand. *Journal of Economic Theory* 34, 71–94.

Vives, X. 2005. Complementarities and games: new developments. *Journal of Economic Literature* 43, 437–79.

Zhao, J. 2001. A characterization for the negative welfare effects of cost reduction in a Cournot oligopoly. *International Journal of Industrial Organization* 19, 455–69.

Ziv, A. 1993. Information-sharing in oligopoly: the truth-telling problem. *RAND Journal of Economics* 24, 455–65.

How to cite this article

Daughety, Andrew F. "Cournot competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000591> doi:10.1057/9780230226203.0336

The New Palgrave Dictionary of Economics Online

Cournot, Antoine Augustin (1801 – 1877)

Martin Shubik
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bertrand, J. L. F.; Chamberlin, E. H.; chance; Cournot, A. A.; Edgeworth cycle; entry; equation of exchange; Jevons, W. S.; large group equilibrium; law of demand; Marshall, A.; mathematical economics; monopolistic competition; monopoly; non-cooperative equilibrium; objective and subjective probability; oligopolistic competition; oligopoly; price as a strategic variable; probability; product differentiation; quantity as a strategic variable; silver standard; supply and demand; uncertainty; unlimited competition; value in exchange; Walras, L.; wealth

Article

Cournot was born at Gray (Haute-Saône) on 28 August 1801 and died in Paris on 30 March 1877. Until the age of 15 his education was at Gray. After studying at Besançon he was admitted to the Ecole Normale Supérieure in Paris in 1821. In 1823 he obtained his licentiate in sciences and in October of that year was employed by Marshal Gouvion-Saint-Cyr as literary adviser to the Marshal and tutor to his son. In 1829 he obtained his doctorate in science with a main thesis in mechanics and a secondary one in astronomy. Through the sponsorship of Poisson in 1834 he obtained the professorship in analysis and mechanics at Lyon.

After a year of teaching he became primarily involved in university administration. In 1835 he became rector of the Académie de Grenoble and subsequently became inspector general of education and from 1854 to 1862 was rector of the Académie de Dijon. He became a Knight of the Legion of Honour in 1838 and an Officer in 1845. He was afflicted with failing eyesight and in the last part of his life was nearly blind. In 1862 he retired from public life but continued his own researches in Paris until his death.

Cournot was a prolific writer. His writings can be broadly divided into three categories: (1) mathematics; (2) economics and (3) the philosophy of science and philosophy of history.

In considering Cournot as an economist it is necessary to place his major economic work, *Recherches sur les principes mathématiques de la théorie des richesses* (1838) in the context not only of *Principes de la théorie des richesses* (1863), which can be regarded as a literary version of his work of a quarter of

a century earlier, and his *Revue sommaire des doctrines économiques* (1877) which appeared in the last year of his life, but also of his writings on probability and the philosophy of science, in particular *Exposition de la théorie des chances et des probabilités* (1843) and *Matérialisme, vitalisme, rationalisme: Etudes des données de la science en philosophie* (1875).

It is possible to weave a broad cloth of interpretation taking into account not merely Cournot's other works but what appears to be known of his personality and the considerable social and political flux in France during the times in which he lived. Guitton (1968) has suggested that Cournot had a rather melancholic and solitary temperament and ‘did nothing to make his books attractive’. He notes that: ‘Cournot was a pioneer. He did nothing to court his contemporaries, and they, in turn, not only failed to appreciate him but ignored him.’ Palomba ([1981] 1984) provides a sketch of the historical background of his time, noting the growth of socialist ideas in Europe, the political actions and reactions to the French Revolution and the challenges to the concept of ownership. Rather than challenge or repeat the broad contextual interpretation of Cournot provided by Palomba, this article is confined primarily to the direct interpretation of his works in economics and supporting texts in the light of many of the developments in economics which are consistent with and may be indebted to his original ideas.

The texts followed here include the French given in the complete works of Cournot (1973) and the Nathaniel T. Bacon translation (1899) entitled *Researches into the Mathematical Principles of the Theory of Wealth*, which also contains an essay by Irving Fisher on Cournot and Mathematical Economics as well as a bibliography on Mathematical Economics from 1711 to 1897. The 1929 reprint of the 1897 edition was used.

The preface sets forth with great clarity Cournot's fundamental approach to political economy. He states:

> But the title of this work sets forth not only theoretical researches; it shows also that I intend to apply to them the forms and symbols of mathematical analysis. Most authors who have devoted themselves to political economy seem also to have had a wrong idea of the nature of the applications of mathematical analysis to the theory of wealth.
>
> But those skilled in mathematical analysis know that its object is not simply to calculate numbers, but that it is also employed to find the relations between magnitudes which cannot be expressed in numbers and between functions whose law is not capable of algebraic expression. Thus the theory of probabilities furnishes a demonstration of very important propositions, although without the help of experience it is impossible to give numerical values for contingent events, except in questions of mere curiosity, such as arise from certain games of chance. (p. 3)

Cournot continues in the preface to note that only the first principles of differential and integral calculus are required for his treatise. Professional mathematicians could be interested in it for the questions raised rather than the level of mathematics presented. He ends the preface with the caveat:

> I am far from having thought of writing in support of any system, and from joining the banners of any party; I believe that there is an immense step in passing from theory to governmental applications; I believe that theory loses none of its value in thus remaining preserved from contact with impassioned polemics; and I believe, if this essay is of any

> practical value, it will be chiefly in making clear how far we are from being able to solve, with full knowledge of the case, a multitude of questions which are boldly decided every day. (p. 5)

The first chapter, ‘Of Value in Exchange or of Wealth in General’, provides insight into the breadth of Cournot's concern for the social and historical context of wealth.

> Property, power, the distinctions between masters, servants and slaves, abundance, and poverty, rights and privileges, all these are found among the most savage tribes, and seem to flow necessarily from the natural laws which preside over aggregations of individuals and of families; but such an idea of wealth as we draw from our advanced state of civilization, and such as is necessary to give rise to a theory, can only be slowly developed as a consequence of the progress of commercial relations, and of the gradual reaction of those relations on civil institutions. (pp. 7–8)

He notes that: ‘it is a long step to the abstract idea of *value in exchange* which supposes that the objects to which such value is attributed *are in commercial circulation*.’
In order to illustrate the distinction between the word *wealth* in ordinary speech and value in exchange, he presents an example of a publisher who destroys two-thirds of his stock expecting to derive more profit from the remainder than the entire edition. The economics of elasticity is developed more formally in Chapter 4 on demand, but the concept is clear.
Chapter 2, ‘On Changes in Value, Absolute and Relative’, begins by noting that ‘we can only assign value to a commodity by reference to other commodities’. This leads to a discussion of the use of a corrected money which would serve as ‘the equivalent of the mean sun of the astronomers’.
Chapter 3, ‘Of the Exchanges’, is the first in which formal mathematical manipulation is employed. He considers a silver standard in which all currencies are fixed in ratio to a gram of fine silver. He observes that the ratios of exchange for the same weight of fine silver cannot differ by more than transportation and smuggling costs. Given the volume of trade measured in silver he considers the arbitrage conditions for the $m(m-1)/2$ ratios among m centres. Fisher (1892) notes, however, that Cournot did not appear to be acquainted with determinants as he did not attempt a general solution of the exchange equations he proposed, but limited his calculations to three centres of exchange.
It is in Chapter 4, ‘On the Law of Demand’, that the modernity of his approach stands out. He is interested in demand as it is revealed in sales at a given price. He represents the relationship between sales and price by the continuous function $D=F(p)$ and observes that this function generally increases in size with a fall in price and that the empirical problem is to determine the form of $F(p)$. He indicates an appreciation of the concept of elasticity of demand although he did not develop the formal measure.
Chapters 5 and 6 deal with monopoly without and with taxation; Chapter 7 is on the competition of producers and Chapter 8 on unlimited competition. The ninth chapter is on the mutual relations of producers and the tenth on the communication of markets. The final two chapters are somewhat macroeconomic in scope. Chapter 11 is entitled ‘Of the Social Income’ and 12 ‘Of Variations in the Social Income, Resulting from the Communication of Markets’.
As our commentary is primarily on Chapters 5–8, the order is reversed and 11 and 12 are dealt with first. Cournot explicitly avoids setting up the whole closed microeconomic system.

> It seems, therefore, as if, for a complete and rigorous solution of the problems relative to some parts of the economic system, it were indispensable to take the entire system into consideration. But this would surpass the powers of mathematical analysis and of our practical methods of calculation, even if the values of all the constants could be assigned to them numerically. The object of this chapter and of the following one is to show how far it is possible to avoid this difficulty, while maintaining a certain kind of approximation, and to carry on, by the aid of mathematical symbols, a useful analysis of the most general questions which this subject brings up.
>
> We will denote by *social income* the sum, not only of incomes properly so called, which belong to members of society in their quality of real estate owners or capitalists, but also the wages and annual profits which come to them in their capacity of workers and industrial agents. We will also include in it the annual amount of the stipends by means of which individuals or the state sustain those classes of men which economic writers have characterized as unproductive, because the product of their labour is not anything material or saleable. (pp. 127–8)

But, using a first order approximation, he studies the effect of a change in price and consumption of a good on social income as a whole under competition, under monopoly and when a new product is introduced.

> Finally, although we make continuous and almost exclusive use of the word *commodity*, it must not be lost sight of (Article 8) that in this work we assimilate to commodities the rendering of services which have for their object the satisfaction of wants or the procuring of enjoyment. Thus when we say that funds are diverted from the demand for commodity A to be applied to the demand for commodity B, it may be meant by this expression that the funds diverted from the demand for a commodity properly so called, are employed to pay for services or vice versa. When the population of a great city loses its taste for taverns and takes up that for theatrical representations, the funds which were used in the demand for alcoholic beverages go to pay actors, authors, and musicians, whose annual income, according to our definition, appears on the balance sheet of the social income, as well as the rent of the vineyard owner, the vine-dresser's wages, and the tavern-keeper's profits. (p. 149)

The last chapter considers international trade and national income and uses a first order approximation rather than a closed equilibrium system to study the benefits of opening up trade.

> Moreover (and this is the favourite argument of writers of the school of Adam Smith), it should be inferred from the asserted advantage assigned to the exporting market, and the asserted disadvantage suffered by the importing market, that a nation should so arrange as always to export and never to import, which is evidently absurd, as it can only export on

> condition of importing, and even the sum of the values exported, calculated at the moment of leaving the national market, must necessarily be equal to the sum of the values imported, calculated at the moment of arrival on the national market. (p. 161)

Cournot also notes the problem of analysing a tariff war:

> The question would no longer be the same if establishment of a barrier for the benefit of A producers might provoke, by way of retaliation, the establishment of another barrier for the benefit of B producers, against whom the first barrier was raised. The government of A would then have to weigh the advantage resulting from the first measure to the citizens of A against the drawbacks caused by the retaliation. The two markets A and B would thus again be placed in symmetrical conditions, and each should be considered as acting the double part of an exporting and importing market. (p. 164)

He closes his comments with:

> We have just laid a finger on the question which is at the bottom of all discussions on measures which prohibit or restrict freedom of trade. It is not enough to accurately analyse the influence of such measures on the national income; their tendency as to the distribution of the wealth of society should also be looked into. We have no intention of taking up here this delicate question, which would carry us too far away from the purely abstract discussions with which this essay has to do. If we have tried to overthrow the doctrine of Smith's school as to barriers, it was only from theoretical considerations, and not in the least to make ourselves the advocates of prohibitory and restrictive laws. Moreover, it must be recognized that such questions as that of commercial liberty are not settled either by the arguments of scientific men or even by the wisdom of statesmen. (p. 171)

He closes his work with the observation about theory that:

> By giving more light on a debated point, it soothes the passions which are aroused. Systems have their fanatics, but the science which succeeds to systems never has them. Finally, even if theories relating to social organization do not guide the doings of the day, they at least throw light on the history of accomplished facts. (p. 171)

Although the contribution of these last chapters is not as great as those to which we now turn, the spirit and style is that of a major theorist concerned deeply and objectively with application to practical affairs.
In Chapters 5–9 Cournot develops his theory of monopoly, oligopoly and unlimited competition. This can be contrasted with Ricardo (1817) before and Walras (1874) after, who concentrated on unlimited competition with no aim at producing a unified theory involving numbers.
In Chapter 5 Cournot deals with monopoly, considering increasing, decreasing and constant returns and in Chapter 6 the influence of taxation on a monopoly is considered. He notes direct taxes and indirect

taxes as well as bounties and their influences on both producers and consumers; and closes with an examination of two variations of taxation in kind.

Chapter 7 provides a smooth transformation from single person maximization to non-cooperative optimization where agents who mutually influence each other act without explicit cooperation.

> We say *each independently*, and this restriction is very important, as will soon appear; for if they should come to an agreement so as to obtain for each the greatest possible income, the results would be entirely different, and would not differ, so far as consumers are concerned, from those obtained in treating of a monopoly.
>
> Instead of adopting D=F(p) as before, in this case it will be convenient to adopt the inverse notation p=f(D); and then the profits of proprietors (1) and (2) will be respectively expressed by
>
> $$D_1 f(D_1 + D_2), \text{ and } D_2 f(D_1 + D_2),$$
>
> i.e. by functions into each of which enter two variables, D_1 and D_2. (p. 80)

It is at this point that Cournot switches from price to quantity of a homogeneous product as the strategic variable used by the competitors. His words and the mathematics do not quite match. He says, ‘This he will be able to accomplish by properly adjusting his price.’ The first order condition for the existence of a non-cooperative equilibrium with quantity as the strategic variable is given. A diagram showing a stable equilibrium and another with a non-stable equilibrium are presented. The analysis is generalized to *n* producers including the possibility of an extra group of producers beyond *n*, all of whom produce at capacity. He obtains *n* symmetric equations for the firms with interior production levels and sets the others at capacity.

When he introduces *n* different general cost functions for the *n* firms he handles the situation with all having an equilibrium defined by the simultaneous satisfaction of the equations arising from the first order conditions. But he does not deal with the possibility that costs could be such that different subsets of firms could be active in different equilibria.

The criticism levelled by Bertrand (1883) in his review written well after Cournot's death concerns the modelling rather than the mathematics. As Cournot considered competition without entry among firms selling an identical product it was fairly natural to avoid the discontinuity in the payoff function caused by selecting price as an independent variable. But the observation of Bertrand matters for markets with a finite number of firms. The choice of strategic variable causes not only mathematical difficulties but raises questions concerning economic realism and relevance. Quantity, price, quality, product differentiation and scope can all be considered as playing dominant roles in different markets. But the general explanation of price and quantity as strategic variables was and is critical to the development of economic theory. Cournot provided the foundations for the understanding of quantity. Bertrand, whose review of the books of Cournot and Walras was somewhat tangential to his professional interests offered

only an example rather than a developed theory of price competition. It remained for Edgeworth (1925, pp. 111–42) to explore the underlying difficulties with the payoff functions for duopoly with increasing marginal costs; and it has only been since the 1950s with the advent of the theory of games that there has been an adequate study of the properties of non-cooperative equilibria in games with price and quantity as strategic variables, without or with product differentiation.

The thesis of Nash (1951) on the existence of non-cooperative equilibria for a class of games in strategic form provided a broad general underpinning for the concept of non-cooperative equilibrium. It was then immediately observable that, although Cournot's work with equilibria of games with a continuum of strategies was not strictly covered by Nash's work, conceptually Cournot's solution could be viewed as an application of non-cooperative equilibrium theory to oligopoly (see Mayberry, Nash and Shubik, 1953). The broader investigation of the price model and the interpretation of the instability of the Edgeworth cycle in terms of mixed strategy equilibria has only taken place recently. This also includes a growing literature on how to embed both the Cournot and Bertrand–Edgeworth models into a closed economic system or Walrasian framework. A summary of much of this work is presented by Shubik (1984).

It is important to appreciate that the developments in the theory of monopolistic competition such as those of Hotelling (1929) and Chamberlin (1933) and J. Robinson (1933) were based upon the Cournot non-cooperative game model. Although it may be argued that Chamberlin's and Mrs Robinson's works possibly contained broader and richer models of competition among the few than that of Cournot, they represented a step backwards in their lack of mathematical sophistication and analysis. The Chamberlin discussion of large group equilibrium does have price as the strategic variable along with product differentiation and entry, but the solution concept is the non-cooperative equilibrium à la Cournot with the caveat that an attempt to produce a strict formal mathematical model of Chamberlin's large group equilibrium leads one to conclude that the game having price as a strategic variable is closer to Edgeworth's analysis than that of Cournot and a price strategy non-cooperative equilibrium may not exist.

In Chapter 8 Cournot shows his basic grasp of the important strategic difference between pure competition and oligopolistic competition. Using his own words, he states:

> The effects of competition have reached their limit, when each of the partial productions D_2 is *inappreciable*, not only with reference to the total production D=F(p), but also with reference to the derivative $F'(p)$, so that the partial production D_k could be subtracted from D without any appreciable variation resulting in the price of the commodity. This hypothesis is the one which is realized, in social economy, for a multitude of products, and, among them, for the most important products. It introduces a great simplification into the calculations, and this chapter is meant to develop the consequences of it. (p. 90)

In modern mathematical economics, in the linking of competition among the few and the Walrasian system into a logically consistent whole, two approaches to the study of large numbers have been adopted. The first is replication and has its roots in Cournot and, more formally, Edgeworth (1881). Following Edgeworth this method was used in cooperative core theory by Shubik (1959). The second

involves considering a continuum of economic agents where each agent can be regarded as a set of measure zero. Cournot clearly saw the need to consider a market in which each individual firm is too small to influence price. But it remained for Aumann (1964) to fully formalize the concept of an economic game with a continuum of agents.

After 25 years during which his seminal work in mathematical economics was essentially ignored, Cournot demonstrated his concern for his ideas by publishing *Principes de la théorie des richesses* (1863), where he offered a non-mathematical rendition of his early work. This book is of considerably greater length than its predecessor and is divided into four books: Book 1, Les Richesses (eight chapters); Book 2, Les Monnaies (seven chapters); Book 3, Le Système économique (ten chapters) and Book 4, L'Optimisme économique (seven chapters).

This book met with no more immediate success than his original work and is not as deep. For example the chapters on money, although they contain discursive and historical material of interest, have little material of analytic depth.

In spite of the indifference of the environment to his writings in economics, Cournot regarded his contribution as sufficiently important that some 14 years later, in the year of his death, he published his *Revue sommaire des doctrines économiques* (1877). This book was also longer, non-mathematical and of less significance than the work of almost 40 years earlier. But Cournot's own sense of having been at least partially vindicated after 40-odd years is indicated in his *avant-propos*:

> I was at that point in 1863, when I had the desire to find out whether I had sinned in the substance of ideas or only in their form. To that end, I went back to my work of 1838, expanding it where needed, and, most of all, removing entirely the algebraic apparatus which intimidates so much in these subjects. Whence the book entitled: ‘Principes de la théorie des richesses’. ‘Since it took me,’ I said in the preface, ‘twenty-five years to lodge an appeal of the first sentence, it goes without saying that I do not intend, whatever happens, to resort to any other means. If I lose my case a second time, I will be left only with the consolation which never abandons disgraced authors: that of thinking that the sentence that condemns them will one day be quashed in the interest of the law, that is of the truth.’
>
> When I took this engagement in 1863, I did not think that I would live long enough to see my 1838 case reviewed as a matter of course. Nevertheless, more than thirty years later, another generation of economists, to put it like Mr. the commander Boccardo, discovered that I opened up back then, though too timidly and too partially, a good path to be followed, on which I was even somewhat preceded by a man of merit, the doctor Whewell. While another Englishman, Mr. Jevons, was undertaking to enlarge this path, a young Frenchman, Mr. Leon Walras, professor of Political Economy at Lausanne, dared to maintain right in the Institute that it was wrong to pay so little attention to my method and my algorithm, which he used rightfully to expose a new theory, more amply developed.
>
> Now, look at my bad luck. If I won a little late, without any involvement, my 1838 case, I lost my 1863 case. If one wanted in retrospective to make a case for my algebra, my prose

> (I am ashamed of saying it) did not get better success from the publisher. The *Journal des Economistes* (August 1864) criticized me mainly ‘for not having moved on from Ricardo,’ for not having taken into account the discoveries that so many men of merit have made in twenty-five years in the field of political economy; thus the poor author that no one of the official world of French economists wanted to quote incurred the reproach of not having quoted others enough.

Cournot was central to the founding of modern mathematical economics. The average reader tends not to be aware that the textbook presentations of the ‘marginal cost equals marginal revenue’ optimizing condition for monopoly and ‘marginal cost equals price’ for the firm in pure competition come directly from the work of Cournot (including an investigation of the second order conditions).

He had to wait many years for recognition, but when it came in the works of Jevons, Marshall, Edgeworth, Walras and others, it moved the course of economic theory. Marshall notes (*Memorials of Alfred Marshall*, pp. 412–13, letter 2, July 1900) ‘I fancy I read Cournot in 1868’, this was when Marshall was 26, some 30 years after the book appeared. He acknowledges him both as a great master and as his source ‘as regards the form of thought’ for Marshall's theory of distribution. Jevons, in his preface to the second edition of *The Theory of Political Economy* records ‘I procured a copy of the work as far back as 1872’ and that it ‘contains a wonderful analysis of the laws of supply and demand, and of the relations of prices, production, consumption, expenses and profits’. He excuses himself for his lateness in coming to Cournot observing: ‘English economists can hardly be blamed for their ignorance of Cournot's economic works when we find French writers equally bad.’ Walras in the preface to the fourth edition of *Elements of Pure Economic* (Jaffé translation, p. 37) acknowledges his ‘father Auguste Walras, for the fundamental principles of my economic doctrine’; and ‘Augustin Cournot for the idea of using the calculus of functions in the elaboration of this doctrine’. His liberal references to Cournot include his discussion of monopoly and the description of supply and demand.

The art of formal modelling is different from but related to the use of mathematical analysis in economics. The clarity and parsimony of Cournot's modelling stand out and have served as beacons guiding the development of mathematical economics.

An important feature missing from Cournot's seminal work is the discussion of the role of chance and uncertainty in the economy. He stressed the importance of chance in both his book *Exposition de la théorie des chances et des probabilités* (1843) and in *Matérialisme, vitalisme, rationalisme* (1875). Although economics was the only social science he attempted to mathematize, he was well aware of the simplifications being made in cutting economic analysis from the context of history and society.

> The economist considers the body social in a state of division and so to say of extreme pulverization, where all the particularities of organization and of individual life offset each other and vanish. The laws that he discovers or believes to discover are those of a mechanism, not those of a living organism. For him, it is no longer a question of social physiology, but of what is rightfully called social physics (p. 56). We mention that these cases of regression which imply abstractions of the same kind, if not of the same type and of the same value, reappear in various stages of scientific construction.

Cournot's work on chance and probability does not appear to have provided any new mathematical

analysis, but he made three distinctions concerning the nature of probability. His book of 1843 was a text with the dual purpose of teaching the non-mathematician the rules of the calculus of probability and of dissipating the obscurities on the delicate subject of probability. He stressed the distinction between objective and subjective probability. His opening chapters provide a discussion of the appropriate combinatorics and frequency of occurrence interpretation of probability.

Cournot stressed the distinction between objective probability where frequencies are known and subjective probability. He noted:

> We could, since then, relying on the theorems of Jacques Bernoulli, who was already aware of their meaning and scope, pass immediately to the applications those theorems had in the sciences of facts and observations. However, a principle, first stated by the Englishman Bayes, and on which Condorcet, Laplace and their successors wanted to build the doctrine of ‘a posteriori’ probabilities, became the source of much ambiguity which must first be clarified, of serious mistakes which must be corrected and which are corrected as soon as one has in mind the fundamental distinction between probabilities which have an objective existence, which give a measure of the possibility of things, and subjective probabilities, relating partly to one's knowledge, partly to one's ignorance, depending on one's intelligence level and on the available data. (p. 155)

Subjective probability rests on the consideration of events which our ignorance calls for us to treat as equiprobable due to insufficient cause.

He added a third category which he entitled ‘philosophical probability’ (Chapter 17) ‘where probabilities are not reducible to an enumeration of chances’ but ‘which depend mainly on the idea that we have of the simplicity of the laws of nature’ (p. 440).

Cournot's views on probability appear to be intimately related to his concern for social statistics and economic modelling. Although he did not establish formal links between his mathematical economics models and chance he regarded history and the development of institutions as dependent on chance and economics as set in the context of institutions.

Cournot was at best an indifferent mathematician. Bertrand clearly dominated him in that profession. But from his own writings it is clear that Cournot was well aware of both his purpose in applying mathematics to economics and his limitations as a mathematician. At the age of 58 he wrote his *Souvenirs* which he finished in Dijon in October 1859. They were published many years later with an introduction by Botinelli (1913). In these writings Cournot provides his self-assessment as a mathematician.

> I was starting to be a little known in the academic world through a fairly large number of scientific articles. This was the basis of my fortune. Some of these articles ended up with Mr. Poisson, who was then the leader in Mathematics at the Institute, and mainly at the University, and he liked them particularly. He found in them philosophical insight, which I think was not all that wrong. Furthermore, he foresaw that I would go a long way in the field of pure mathematical speculation, which was (I always thought it and never hesitated to say it) one of his mistakes.

The general tenor of his *Souvenirs* is of a moderately conservative, quietly humourous, self-effacing man with considerable understanding of his environment and a broad belief in science and its value to society.

Regarding his work as a whole, his dedication and power as the founder of mathematical economics and the promoter of empirical numerical investigations emerges. He strove for around 40 years to have his ideas accepted. He did so with persistence and humour (referring to his major work as ‘mon opuscule’). He understood the need to wait for a generation to die. And before his death with the work and words of Jevons and Walras he saw the vindication of his approach.

See Also

- Bertrand, Joseph Louis François

Selected works

1838. *Researches into the Mathematical Principles of the Theory of Wealth*. Trans. N.T. Bacon, New York: Macmillan, 1929.

1841. *Traité élémentaire de la théorie des fonctions et du calcul infinitésimal*. 2nd edn, Paris: Hachette, 1857.

1843. *Exposition de la théorie des chances et des probabilités*. Paris: Hachette.

1861. *Traité de l'enchaînement des idées fondamentales dans les sciences et dans l'histoire*. New edn, Paris: Hachette, 1911.

1863. *Principes de la théorie des richesses*. Paris: Hachette.

1872. *Considérations sur la marche des idées et des évènements dans les temps modernes*. 2 vols, Paris: Boivin, 1934.

1875. *Matérialisme, vitalisme, rationalisme: Études des données de la science en philosophie*. Paris: Hachette, 1923.

1877. *Revue sommaire des doctrines économiques*. Paris: Hachette.

1913. *Souvenirs 1760–1860*. With an introduction by E.P. Bottinelli. Paris: Hachette. Published posthumously.

1973. *A.A. Cournot Oeuvres Complètes*, 5 vols, ed. André Robinet. Paris: Librairie Philosophique J. Vrin.

Bibliography

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Bertrand, J.L.F. 1883. (Book reviews of) *Théories Mathematique de la richesse sociale* par Léon Walras; *Recherches sur les principes mathématiques de la théorie de la richesse* par Augustin Cournot. *Journal des Savants* 67, 499–508.

Chamberlin, E.H. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Edgeworth, F.Y. 1925. *Papers Relating to Political Economy*, I. London: Macmillan.

Fisher, I. 1892. *Mathematical Investigations in the Theory of Value and Prices*. New Haven: Connecticut Academy of Arts and Sciences, *Transactions 9*. Reprinted, New York: Augustin M. Kerlley, 1961.

Guillebaud, C.W., ed. 1961. *Marshall's Principles of Economics*. Vol. 2, *Notes*. London: Macmillan.

Guitton, H. 1968. Antoine Augustin Cournot. In *The International Encyclopedia of the Social Sciences*, vol. 3. New York: Macmillan and Free Press.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 34, 41–57.

Jevons, W.S. 1911. *The Theory of Political Economy*. 4th edn, London: Macmillan, 1931.

Mayberry, J., Nash, J.F. and Shubik, M. 1953. A comparison of treatments of a duopoly situation. *Econometrica* 21, 141–55.

Nash, J.F., Jr. 1951. Noncooperative games. *Annals of Mathematics* 54, 289–95.

Palomba, G. 1984. Introduction à l'oeuvre de Cournot. *Economie Appliquée* 37, 7–97. Trans. from Italian, extracted from *Cournot Opere*, Turin: UTET (1981).

Ricardo, D. 1817. *The Principles of Political Economy and Taxation*. London: J.M. Dent, 1965.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Shubik, M. 1959. Edgeworth market games. In *Contributions to the Theory of Games IV*, ed. A.W. Tucker and R.D. Luce. Princeton, NJ: Princeton University Press.

Shubik, M. 1984. *A Game Theoretic Approach to Political Economy*. Cambridge, MA: MIT Press.

Walras, L. 1874–7. *Elements of Pure Economics*. Trans. W. Jaffée, London: George Allen & Unwin, 1954.

How to cite this article

Shubik, Martin. "Cournot, Antoine Augustin (1801–1877)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000413>
doi:10.1057/9780230226203.0335

The New Palgrave Dictionary of Economics Online

creative destruction (Schumpeterian conception)

Arnold Heertje

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

creative destruction; dynamic efficiency; information technology; innovation; restrictive practices; Schumpeter, J. A.

Article

Schumpeter invented the phrase ‘creative destruction’ in his famous book on the development of capitalism into socialism (Schumpeter, 1942). In his view the process of creative destruction is the essential fact about capitalism and refers to the incessant mutation of the economic structure from within, destroying the old and creating a new.

In the footsteps of Karl Marx, Schumpeter argues that in dealing with capitalism we are dealing with an evolutionary process. It is by nature a form or method of economic change and not only never is, but never can be, stationary. The fundamental impulse that sets and keeps the capitalist engine in motion comes from new goods and new methods of production and transportation, created by the Schumpeterian entrepreneur, who is always on the outlook for new combinations of the factors of production (Heertje, 2006).

The process of creative destruction takes time. For that reason there is no point in appraising its performance within a static framework. A system may produce an optimal allocation of resources at every point of time and may yet in the long run be inferior to a system without such optimal allocation, because the non-optimality may be a condition for the level and speed of long-run performances; in other words, for dynamic efficiency. Furthermore, the process of creative destruction in Schumpeter's vision must be seen as the background for individual decisions and strategies. Economic theory has a tendency to concentrate on decisions about prices by firms, which are assumed to maximize profits, within a given structure. Schumpeter argues that the relevant problem is how capitalism creates and destroys these structures (Metcalfe, 1998).

Schumpeter's conception of creative destruction overturns the idea that price competition is the only component of the market behaviour of entrepreneurs. In fact, it is not that kind of competition which counts, but the competition from the new commodity, the new technology, the new source of supply and

the new type of organization. Instead of marginal changes, fundamental upheavals are brought about by process and product innovations of existing firms and potential competitors.

Restrictive practices of monopolists and large firms are to be judged against the background of the perennial gale of creative destruction, rather than in the context of stationary development. The potential threat of process and product innovation reduces the scope and importance of restrictive practices that aim to guarantee the monopolist or big firm a quiet life. If however the profits are used to counterattack, restrictive practices may help to deepen the process of creative destruction and, therefore, the dynamic effects of capitalism (Reisman, 2004).

The process of creative destruction as described by Schumpeter has been experienced again since the 1980s in the United States, Japan and Western Europe and since the 1990s in China and India as well. On the basis of new technologies many old firms, structures and professions have been swept away and new industrial organizations and labour relations have emerged. In particular, the application of information technology and the Internet with the dramatic decrease in transaction costs of communication is leading to major changes of a quantitative and qualitative nature in both the private and public sector of the economy. On the one hand, ‘external’ growth of already large firms which take over others is a feature of modern capitalism; on the other hand, every day new small firms are established, often created by former executives of existing (and long-lived) companies.

This extensive discussion of the process of creative destruction illustrates Schumpeter's strong emphasis on the supply side of the economy. It would be an interesting question to study the impact of the process of creative destruction on employment. My guess would be that, on balance, the process of creative destruction is more creative than destructive, not only with regard to employment but also concerning broader perspectives of growth and welfare. This may be one of the reasons why Schumpeter's work has had a lasting and ever-increasing influence on economic theory.

See Also

- creative destruction
- market structure
- Schumpeter, Joseph Alois

Bibliography

Heertje, A. 2006. *Schumpeter on the Economics of Innovation and the Development of Capitalism*. Cheltenham: Edward Elgar.

Metcalfe, J.S. 1998. *Evolutionary Economics and Creative Destruction*. London: Routledge.

Reisman, D. 2004. *Schumpeter's Market*. Cheltenham: Edward Elgar.

Schumpeter, J. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper.

How to cite this article

Heertje, Arnold. "creative destruction (Schumpeterian conception)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000601>
doi:10.1057/9780230226203.0338

The New Palgrave Dictionary of Economics Online

creative destruction

Ricardo J. Caballero

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Creative destruction refers to the incessant product and process innovation mechanism by which new production units replace outdated ones. This restructuring process permeates major aspects of macroeconomic performance, not only long-run growth but also economic fluctuations, structural adjustment and the functioning of factor markets. Over the long run, the process of creative destruction accounts for over 50 per cent of productivity growth. At business cycle frequency, restructuring typically declines during recessions, and this add a significant cost to downturns. Obstacles to the process of creative destruction can have severe short- and long-run macroeconomic consequences.

Keywords

banking reform; business cycles; Canada–US Free Trade Agreement; creative destruction; deregulation of product markets; economic growth; factor reallocation; Hayek, F.; innovation; international competition; job flows; job security; labour market regulation; liquidationist thesis; productivity growth; recessions; Robbins, L.; Schumpeter, J.; structural change

Article

Creative destruction refers to the incessant product and process innovation mechanism by which new production units replace outdated ones. It was coined by Joseph Schumpeter (1942), who considered it ‘the essential fact about capitalism’.

The process of Schumpeterian creative destruction (restructuring) permeates major aspects of macroeconomic performance, not only long-run growth but also economic fluctuations, structural adjustment and the functioning of factor markets.

At the microeconomic level, restructuring is characterized by countless decisions to create and destroy production arrangements. These decisions are often complex, involving multiple parties as well as strategic and technological considerations. The efficiency of those decisions not only depends on

managerial talent but also hinges on the existence of sound institutions that provide a proper transactional framework. Failure along this dimension can have severe macroeconomic consequences once it interacts with the process of creative destruction (see Caballero and Hammour, 1994; 1996a; 1996b; 1996c; 1998a; 1998b; 2005). Some of these limitations are natural, as they derive from the sheer complexity of these transactions. Others are man-made, with their origins ranging from ill-conceived economic ideas to the achievement of higher human goals, such as the inalienability of human capital. In moderate amounts, these institutional limitations give rise to business cycle patterns such as those observed in the most developed and flexible economies. They can help explain perennial macroeconomic issues such as the cyclical behaviour of unemployment, investment and wages. In higher doses, by limiting the economy's ability to tap new technological opportunities and adapt to a changing environment, institutional failure can result in dysfunctional factor markets, resource misallocation, economic stagnation, and exposure to deep crises.

Given the nature of this short piece, I will skip any discussion of models, and refer the reader to Caballero (2006) for a review of the models behind the previous paragraph, and to Aghion and Howitt (1998) for an exhaustive survey of Schumpeterian growth models. Instead, I focus on reviewing recent empirical evidence on different aspects of the process of creative destruction.

Recent evidence on the pace of creative destruction

There is abundant recent empirical evidence supporting the Schumpeterian view that the process of creative destruction is a major phenomenon at the core of economic growth in market economies.

The most commonly used empirical proxies for the intensity of the process of creative destruction are those of factor reallocation and, in particular, job flows. Davis, Haltiwanger and Schuh (1996) (henceforth DHS) offered the clearest peek into this process by documenting and characterizing the large magnitude of job flows within US manufacturing. They defined job creation (destruction) as the positive (negative) net employment change at the establishment level from one period to the next. Using these definitions, they concluded that over ten per cent of the jobs that exist at any point in time did not exist a year before or will not exist a year later. That is, over ten per cent of existing jobs are destroyed each year and about the same amount is created within the same year. Following the work by DHS for the United States, many authors have constructed more or less comparable measures of job flows for a variety of countries and episodes. Although there are important differences across them, there are some common findings. In particular, job creation and destruction flows are large, ongoing and persistent. Moreover, most job flows take place within rather than between narrowly defined sectors of the economy.

Given the magnitude of these flows and that they take place mostly within narrowly defined sectors, the presumption is strong that they are an integral part of the process by which an economy upgrades its technology. Foster, Haltiwanger and Krizan (2001) provide empirical support for this presumption. They decompose changes in industry-level productivity into within-plant and reallocation (between-plant) components, and conclude that the latter – the most closely related to the creative destruction component – accounts for over 50 per cent of the ten-year productivity growth in the US manufacturing sector between 1977 and 1987. Moreover, in further decompositions they document that entry and exit account for half of this contribution: exiting plants have lower productivity than continuing plants. New plants,

on the other hand, experience a learning and selection period through which they gradually catch up with incumbents. Other studies of US manufacturing based on somewhat different methodologies (see Baily, Hulten and Campbell, 1992; Bartelsman and Dhrymes, 1994) concur with the conclusion that reallocation accounts for a major component of within-industry productivity growth. Bartelsman, Haltwanger and Scarpetta (2004) provide further evidence along these lines for a sample of 24 countries and two-digit industries over the 1990s.

Recent evidence on the cyclical features of creative destruction

At the business cycle frequency, sharp liquidations (rises in job destruction) constitute the most noted impact of contractions on creative destruction. In contrast, job creation is substantially less volatile and mildly pro-cyclical. There is an extensive literature that, extrapolating from the spikes in liquidations (recently measured in job flows but long noticed in other contexts), finds that recessions are times of increased reallocation. In fact, this has been a source of controversy among economists at least since the pre-Keynesian ‘liquidationist’ theses of such economists as Hayek, Schumpeter, and Robbins. These economists saw in the process of liquidation and reallocation of factors of production the main function of recessions. In the words of Schumpeter (1934, p. 16): ‘depressions are not simply evils, which we might attempt to suppress, but … forms of something which has to be done, namely, adjustment to … change.’

In Caballero and Hammour (2005) we turned the liquidationist view upside down. While we sided with Schumpeter and others on the view that increasing the pace of restructuring of the economy is likely to be beneficial, we provided evidence that, contrary to conventional wisdom, restructuring *falls* rather than rises during contractions.

Since the rise in liquidations during recessions is not accompanied by a contemporaneous increase in creation, implicit in the increased-reallocation view is the idea that increased destruction is followed by a surge in creation during the recovery phase of the cyclical downturn. This presumption is the only possible outcome in a representative firm economy, as the representative firm must replace each job it destroys during a recession by creating a new job during the ensuing recovery. However, once one considers a heterogeneous productive structure that experiences ongoing creative destruction, other scenarios are possible. The cumulative effect of a recession on overall restructuring may be positive, zero, or even negative, depending not only on how the economy contracts but also on how it recovers. Thus, the relation between recessions and economic restructuring requires one to examine the effect of a recession on aggregate separations not only at impact, but cumulatively throughout the recession-recovery episode. We explored this issue using quarterly US manufacturing gross job flows and employment data for the 1972–93 period, and found that, along the recovery path, job destruction declines and falls below average for a significant amount of time, more than offsetting its initial peak. On the other hand, job creation recovers, but it does not exceed its average level by any significant extent to offset its initial decline. As a result, our evidence indicates that, on average, recessions *depress* restructuring.

Similarly, in Caballero and Hammour (2001) we approached the question of the pace of restructuring over the cycle from the perspective of corporate assets. Studying the aggregate patterns of merger and acquisition (M&A) activity and its institutional underpinnings, we reached a conclusion that also

amounts to a rejection of the liquidationist perspective. Essentially, a liquidationist perspective in this context would consider fire sales during sharp liquidity contractions as the occasion for intense restructuring of corporate assets. The evidence points, on the contrary, to briskly expansionary periods characterized by high stock market valuations and abundant liquidity as the occasion for intense M&A activity.

Recent evidence on institutional impediments to creative destruction and their cost

For all practical purposes, some product or process innovation is taking place at every instant in time. Absent obstacles to adjustment, continuous innovation would entail infinite rates of restructuring. What are these obstacles to adjustment? The bulk of it is technological – adjustment consumes resources – but (over-?) regulation and other man-made institutional impediments are also a source of depressed restructuring.

While few economists would object to the hypothesis that labour market regulation hinders the process of creative destruction, its empirical support is limited. In Caballero et al. (2004) we revisited this hypothesis using a sectoral panel for 60 countries. We found that job security provisions – measured by variables such as grounds for dismissal protection, protection regarding dismissal procedures, notice and severance payments, and protection of employment in the constitution – hamper the creative destruction process, especially in countries where regulations are likely to be enforced. Moving from the 20th to the 80th percentile in job security cuts the annual speed of adjustment to shocks by a third. By impairing worker movements from less to more productive units, effective labour protection reduces aggregate output and slows down economic growth. We estimated that moving from the 20th to the 80th percentile of job security lowers annual productivity growth by as much as 1.7 per cent.

Similarly, the idea that well-functioning financial institutions and markets are important factors behind economic growth is an old one. The process of creative destruction is likely to be a chief factor behind this link. In Caballero, Hoshi and Kashyap (2006) we analysed the decade-long Japanese slowdown of the 1990s and early 2000s. The starting point of our analysis is the well-known observation that many large Japanese banks would have been out of business had regulators forced them to recognize all their loan losses. Because of this, the banks kept many zombie firms alive by rolling over loans that they knew would not be collected (evergreening). Thus, the normal competitive outcome whereby the zombies would shed workers and lose market share was thwarted. Using an extensive data-set, we documented that roughly 30 per cent of firms were on life support from the banks in 2002 and about 15 per cent of assets resided in these firms. The main idea in our article is that the counterpart to the congestion created by the zombies is a reduction in profits for potential and more productive entrants, which discourages their entry. We found clear evidence of such a pattern in firm-level data and of the corresponding reduced restructuring in sectoral data.

Bertrand, Schoar and Thesmar (2004) further drive home the point that problems in the banking sector can have grave consequences for the health of the restructuring process. They use a differences-in-differences approach on firm-level data for the period 1977–99 to analyse the impact of the banking reforms of the mid-1980s on firm and bank behaviour. These reforms eliminated government interference in bank lending decisions, eliminated subsidized bank loans, and allowed French banks to compete more freely in the credit market. They find that, after the reforms, firms' exit rates and asset reallocation rise, and are more correlated with performances.

International competition is an important source of creative destruction. Trefler (2004) concludes that there are significant productivity and reallocation effects from trade openness, even in industrialized economies. To reach this conclusion, Trefler takes advantage of the Canada–US Free Trade Agreement (FTA) to study the effects of a reciprocal trade agreement on Canada. He finds that, for industries that experienced the deepest Canadian tariff reductions, the contraction of low-productivity plants reduced employment by 12 per cent while raising industry-level labour productivity by 15 per cent. Moreover, he finds that at least half of this increase is related to exit and/or contraction of low-productivity plants. Finally, for industries that experienced the largest US tariff reductions, plant-level labour productivity soared by 14 per cent. Consistent with this evidence, Bernard, Jensen and Schott (2006) find that in the United States productivity growth is fastest in industries where trade costs (barriers) have declined the most.

Domestic deregulation of goods markets can have similar effects. For example, Olley and Pakes (1996) find that deregulation in the US telecommunications industry increased productivity predominantly through factor reallocation towards more productive plants rather than through intra-plant productivity gains. More broadly, Klapper, Laeven and Rajan (2004) study the effect of entry regulation on firm behaviour in a sample including firm-level data from countries of western and eastern Europe. Their findings support the notion that regulation affects entry: ‘naturally high-entry’ industries have relatively lower entry in countries that have higher entry regulations. Moreover, both the growth rate and share of high-entry industries are depressed in countries with more stringent barriers to entry. Finally, Fishman and Sarria-Allende (2004) extend the Klapper, Laeven and Rajan study to countries outside Europe and include both industry- and firm-level data from the UNIDO and WorldScope databases, and reach similar conclusions.

Final remarks

Evidence and models coincide in their conclusion that the process of creative destruction is an integral part of economic growth and fluctuations. Obstacles to this process can have severe short- and long-run macroeconomic consequences.

See Also

- Schumpeter, Joseph Alois
- structural change

Bibliography

Aghion, P. and Howitt, P. 1998. *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Baily, N., Hulten, C. and Campbell, D. 1992. Productivity dynamics in manufacturing establishments. In *Brookings Papers on Economic Activity: Microeconomics*, ed. M. Baily and C. Winston. Washington, DC: Brookings Institution.

Bartelsman, E. and Dhrymes, P. 1994. Productivity dynamics: US manufacturing plants, 1972–1986. Finance and Economics Discussion Series No. 94-1. Washington, DC: Board of Governors, Federal Reserve System.

Bartelsman, E., Haltiwanger, J. and Scarpetta, S. 2004. Microeconomic evidence of creative destruction in industrial and developing countries. Mimeo, University of Maryland.

Bernard, A., Jensen, J. and Schott, P. 2006. Survival of the best fit: exposure to low-wage countries and the (uneven) growth of US manufacturing plants. *Journal of International Economics* 68, 219–37.

Bertrand, M., Schoar, A. and Thesmar, D. 2004. Banking deregulation and industry structure: evidence from the French banking reforms of 1985. Discussion Paper No. 4488. London: Centre for Economic Policy Research.

Caballero, R. 2006. *Specificity and the Macroeconomics of Restructuring*. Yrjo Jahnsson Lectures. Cambridge, MA: MIT Press.

Caballero, R. and Hammour, M. 1994. The cleansing effect of recessions. *American Economic Review* 84, 1350–68.

Caballero, R. and Hammour, M. 1996a. The fundamental transformation in macroeconomics. *American Economic Review* 86(2), 181–6.

Caballero, R. and Hammour, M. 1996b. On the timing and efficiency of creative destruction. *Quarterly Journal of Economics* 111, 805–52.

Caballero, R. and Hammour, M. 1996c. On the ills of adjustment. *Journal of Development. Economics* 51, 161–92.

Caballero, R. and Hammour, M. 1998a. The macroeconomics of specificity. *Journal of Political Economy* 106, 724–67.

Caballero, R. and Hammour, M. 1998b. Jobless growth: appropriability, factor substitution and unemployment. *Carnegie-Rochester Conference Series on Public Policy* 48, 51–94.

Caballero, R. and Hammour, M. 2001. Institutions, restructuring, and macroeconomic performance. In *Advances in Macroeconomic Theory*, ed. J. Dreze. New York: Palgrave Macmillan.

Caballero, R. and Hammour, M. 2005. The cost of recessions revisited: a reverse-liquidationist view. *Review of Economic Studies* 72, 313–41.

Caballero, R., Hoshi, T. and Kashyap, A. 2006. Zombie lending and depressed restructuring in Japan. Working Paper No. 12129. Cambridge, MA: NBER.

Caballero, R., Cowan, K., Engel, E. and Micco, A. 2004. Effective labor regulation and microeconomic flexibility. Mimeo, MIT.

Davis, S., Haltiwanger, J. and Schuh, S. 1996. *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Fishman, R. and Sarria-Allende, V. 2004. Regulation of entry and the distortion of industrial organization. Working Paper No. 10929. Cambridge, MA: NBER.

Foster, L., Haltiwanger, J. and Krizan, C. 2001. Aggregate productivity growth: lessons from microeconomic evidence. In *New Developments in Productivity Analysis*, ed. E. Dean, M. Harper and C. Hulten. Chicago: University of Chicago Press.

Klapper, L., Laeven, L. and Rajan, R. 2004. Business environment and firm entry: evidence from international data. Working Paper No. 10380. Cambridge, MA: NBER.

Olley, S. and Pakes, A. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64, 1263–98.

Schumpeter, J. 1934. Depressions. In *Economics of the Recovery Program*, ed. D. Brown et al. New York: McGraw-Hill.

Schumpeter, J. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper & Bros.

Trefler, D. 2004. The long and short of the Canada–US Free Trade Agreement. *American Economic Review* 94, 870–95.

How to cite this article

Caballero, Ricardo J. "creative destruction." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000421> doi:10.1057/9780230226203.0337

The New Palgrave Dictionary of Economics Online

credit card industry

Victor Stango and Julian Wright

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

credit card industry; interchange fees; interest rates; stickiness of; networks; sticky prices; two-sided markets

Article

The concept of a general purpose credit card originated in 1949, when Frank McNamara dined in a New York restaurant and discovered that he could not pay for his meal (Evans and Schmalensee, 1999). By the 1980s credit cards had become ubiquitous, and they remain a popular form of payment in most economies. Banks offer cards, setting terms such as interest rates and annual fees. Transactions are handled by networks such as Visa and MasterCard, which emerged in the 1970s as joint member associations. Early research examining the market typically focused on the retail level, while more recent work has tended to focus on the network level, mirroring a shift in policy concerns in the 1980s.
In its early years the US retail credit card market was characterized by extreme interest rate ‘stickiness’ – credit card rates remained virtually constant over time, regardless of economy-wide changes in interest rates. Credit card issuers also appear to have earned super-normal profits during the same period. This presents a puzzle in an industry displaying many classic characteristics of a perfectly competitive market (Ausubel, 1991). Ausubel suggests a variety of explanations for this puzzle, including the possibility that credit card borrowers do not fully anticipate the degree to which they will use the cards.
Ausubel's research spurred a wave of subsequent work proposing explanations for interest rate stickiness. Mester (1994) and Brito and Hartley (1995) provide theoretical explanations for interest rate stickiness based on asymmetric information or consumer transaction costs. Calem and Mester (1995) provide empirical evidence that consumer search and switching costs might explain interest rate stickiness. A complementary explanation for interest rate stickiness is that state-level interest rate ceilings during the 1980s facilitated tacit collusion among card issuers, leading to greater-than-normal interest rate stability (Knittel and Stango, 2003).
By the early 1990s interest rates had become much more flexible as credit card issuers switched to

variable interest rates. By most accounts, the market also became more competitive during this time. One explanation for the change is technological progress that allowed more efficient credit scoring by large nationally marketed card issuers, creating a truly national market that fostered aggressive competition. Other explanations include the threat of interest rate regulation and the entry of new issuers.

At the network level, the key economic issue is that payment card systems like MasterCard and Visa are two-sided markets: they have to attract cardholders to get merchants and merchants to get cardholders. Diners Club did this in 1950 by initially giving away cards to consumers and charging merchants seven per cent of their bill. These days, consumers obtain rewards for using their cards. This structure of pricing has raised the concern of some policymakers. In their view, retailers pay too much to accept credit cards, costs that end up being covered by consumers who do not use credit cards (by way of higher retail prices). Card associations sustain such a price structure through the setting of an interchange fee, which determines how much the merchant's bank must pay the cardholder's bank for each card transaction. A high interchange fee results in a high merchant fee and a low (or negative) fee for cardholders.

The issue of how much to charge each type of user is a common one in other two-sided markets. Magazines and newspapers decide how much to charge readers versus advertisers, and shopping malls decide how much to charge shoppers versus shops. The interest of policymakers in credit cards has spurred research in two-sided markets more generally.

Baxter (1983) provides an early analysis of interchange fees (see Rochet, 2003, for a survey). His key insight is that efficiency calls for card transactions whenever the *joint* benefits to the consumer and merchant of using the card exceed the joint costs of doing so. In the absence of an interchange fee, each type of user will face only the private costs and benefits of cards. A payment from the merchant's bank (acquirer) to the cardholder's bank (issuer) via the interchange fee can align the private incentive to use cards with the social incentive. This provides a justification for setting an interchange fee, but does not imply that card associations will set it at the right level.

One reason a card association might set the interchange fee too high is that acquirers may pass through a larger proportion of interchange fees into merchant fees than issuers pass back to cardholders (in the form of lower fees or higher rewards). Then associations will want to pass revenues to the issuing side, via high interchange fees, where they are competed away less aggressively. A second possible reason is that, if merchants accept cards to attract customers from each other, their private willingness to accept cards includes the surplus their customers get from using cards. As a result, cardholder surplus is over-represented, and card associations tend to charge merchants too much and cardholders too little.

Although these theoretical possibilities highlight possible divergences between privately and socially optimal interchange fees, they provide no basis for the cost-based regulation of interchange fees.

See Also

- two-sided markets

Bibliography

Ausubel, L. 1991. The failure of competition in the credit card market. *American Economic Review* 81,

50–81.

Baxter, W. 1983. Bank interchange of transactional paper: legal perspectives. *Journal of Law and Economics* 26, 541–88.

Brito, D. and Hartley, P. 1995. Consumer rationality and credit cards. *Journal of Political Economy* 103, 400–33.

Calem, P. and Mester, L. 1995. Consumer behavior and the stickiness of credit-card interest rates. *American Economic Review* 85, 1327–36.

Evans, D. and Schmalensee, R. 1999. *Paying with Plastic*. Cambridge, MA: MIT Press.

Knittel, C. and Stango, V. 2003. Price ceilings as focal points for tacit collusion: evidence from credit cards. *American Economic Review* 93, 1703–29.

Mester, L. 1994. Why are credit card rates sticky? *Economic Theory* 4, 505–30.

Rochet, J.-C. 2003. The theory of interchange fees: a synthesis of recent contributions. *Review of Network Economics* 2(2), 97–124.

How to cite this article

Stango, Victor and Julian Wright. "credit card industry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000540> doi:10.1057/9780230226203.0339

The New Palgrave Dictionary of Economics Online

credit cycle

P. Bridel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

It was with the post-First World War attempts to integrate marginalist value and monetary theory that theorists started pondering the possible (in Hayek's words) ‘incorporation of cyclical phenomena into the system of economic equilibrium theory’. Hayek's own ‘intertemporal equilibrium’ approach overturned the traditional view of cycles as temporary deviations from long-period equilibrium conditions. But the publication of Keynes's *General Theory* redirected research efforts towards the determination of output at a point in time. Since the late 1960s, with the search for ‘microfoundations for macroeconomics’, this line of thought has been back on the theoretical agenda.

Keywords

bank rate; Cambridge School; capital theory; credit cycle; cumulative process; forced saving; full employment saving; Hawtrey, R. G.; Hayek, F. A. von; intertemporal equilibrium; Keynes, J. M.; marginal revolution; microfoundations; monetary theory of interest; natural rate and market rate of interest; quantity theory of money; rational expectations; Robertson, D.; saving and investment; temporary equilibrium

Article

Prior to Keynes's *General Theory*, the resolution of the question why, in capitalist economies, aggregate variables undergo repeated fluctuations about the trend was regarded by economists as a main challenge for the profession. What was then called business (or trade) cycle theory grew quite independently from the classical and subsequently neoclassical corpus of price theory. In fact, for all economists, a clear-cut distinction existed between the long-run forces at work in an economy – the subject of a rigorous value and distribution theory – and the more or less ad hoc explanations of the short-run oscillations around such an (equilibrium) centre of gravity. Of course, from Ricardo and Thornton down the 19th century to Overstone and Mill, money and credit played a substantial, but independent, part in these exogenous

explanations of the business cycle. Along the same line, the founding fathers of marginalism (in particular Walras, Marshall and Jevons) failed to coordinate, even in a remotely satisfactory way, money and trade cycle with their then novel price theory.

Following Wicksell's and Mises's lead, it is only with the post-First World War attempts to integrate marginalist value and monetary theory that theorists started pondering the possible ‘incorporation of cyclical phenomena into the system of economic equilibrium theory’ (Hayek, 1929, p. 33n.). The rediscovery of Tooke's (1844) income approach to the quantity theory of money is probably one of the earliest stepping-stones in the development of credit-cycle theories. This line of thought suggests that the explanation of money prices should start not from the quantity of money but from nominal income. Though another way of writing a Marshallian cash balance equation, Wicksell's (1898, p. 44) or Hawtrey's (1913, p. 6) emphasis on the ‘aggregate of money income’, on how it varies, is expanded or held, is a crucial turning-point on the road towards an analysis in terms of income, saving and investment. This shift of emphasis, together with the simultaneous progress in monetary theory proper (notably the development of a comprehensive and integrated monetary theory of interest), the 1914–18 inflationary episode and the post-war cyclical upheavals provided in the 1920s and 1930s the right intellectual stimulus for credit-cycle theories to grow and multiply.

Explicitly or implicitly, to tackle this issue, Continental economists (for example, Mises, Cassel, Hayek, Schumpeter and Aftalion), members of the Cambridge School then dominating in England (Keynes, Robertson, Pigou, Hawtrey), Fisher and Mitchell in the United States all used the common analytical framework established jointly by Walras, Menger, Marshall and Jevons. This is made up of two basic (though familiar) propositions: on the one hand, there is an inverse relation between the volume of investment and the rate of interest (that is, a downward-sloping investment demand curve) and, on the other, despite short-run ‘frictions’, the interest rate is assumed to be sensitive enough to divergences between investment decisions and full employment saving.

The central theme of this argument (first expressed with great clarity in Wicksell's cumulative process) is that the market rate of interest oscillates in the short run around a natural rate of interest determined in the long run by the supply of and the demand for capital as a stock, which, in turn, guarantees the equality between planned investment and full employment saving. Once this logic is understood, it then emerges that the entire development of interwar trade-cycle theories took place within the second proposition outlined above; namely, that, in the long run, the interest rate is assumed to be sensitive enough to divergences between investment decisions and full employment saving. Hence, since the twin concepts of an interest-elastic demand curve for investment and natural rate of interest were never called into question, the orgy of debates that took place in the 1920s and 1930s was conducted in terms of an analysis of various short-run forces which temporarily keep at bay the long-run forces of saving and investment.

These forces are, of course, of multiple nature. Of particular interest to interwar economists, and one of the essential features of business cycle, with its recurrence of upswings and downswings, is a *credit cycle*, an alternation of credit expansion and credit contraction. But it was assumed neither that an alternation of prosperity and depression would not exist in a barter economy (or in a purely specie system) nor that cycles could be viewed as functions of monetary factors only.

In fact, and thanks to their common capital theory, none of the leading interwar credit cycle theorists fell into either of these traps. Even Hawtrey who, with remarkable consistency kept claiming that business cycles are a purely monetary phenomenon, had clearly in mind a Wicksell-like cumulative process

derived from Marshall's oral tradition in monetary theory. This common theoretical background and a deep interest in a then fast-developing monetary theory make similarities between credit cycle theorists

> sufficiently pronounced to entitle us to speak of a single monetary theory [of the cycle], the votaries of which disagree on one issue only: whether bank-loan rates act primarily on ‘durable capital’ [Keynes, Robertson, Hayek] or via the stocks of wholesalers [Hawtrey]. (Schumpeter, 1954, p. 1121)

In 1913, Hawtrey was amongst the first to provide a detailed analysis of the financial working of the cumulative process in an Anglo-Saxon environment. However, even if his theory usefully describes the ways in which money and credit behave in the cycle, the main weakness of his contribution is, of course, its almost exclusive emphasis on dealers' stocks in the course of a credit cycle. If Hawtrey does not deny altogether that a credit expansion/contraction has an influence on the volume of investment, he holds it however to be unimportant when compared with the direct influence on the wholesalers' stocks. He then logically disputes the existence of forced saving on the very ground of this availability of stocks and fails completely to link his credit cycle theory with the dominant Marshallian capital theory. Such a model led Hawtrey not only to give Bank Rate the crucial part to play in any counter-cyclical policy but also to consider its fluctuations as the only explanation of cyclical fluctuations. To sketch British interwar depressions as almost exclusively functions of Bank Rate (itself a function of Britain's absorption of gold) is a rather bold simplification Hawtrey was never quite ready to abandon.

If the theoretical apparatus underlying the *Treatise on Money* proceeds from the same logic, Keynes's fundamental equations introduce, however, a number of very sophisticated and new variations on the basic credit-cycle theme. In particular, causes of credit cycles are of non-monetary nature (they result from fluctuations in the rate of investment relative to the rate of saving), the influence of Bank Rate on investment is not limited ‘to one particular kind of investments, namely, investments by dealers in liquid goods [stocks]’ (Keynes, 1930, vol. 1, p. 173), the cumulative process includes a theory of the demand for money beyond the traditional income motive (that is, an early version of liquidity preference), and, in the short run, there is no longer a direct relation between the quantity of money/credit and the price level: monetary or credit changes do not foster ipso facto a forced/abortive saving process. Despite the higher degree of sophistication shown in the *Treatise*, in a classic chapter on the modus operandi of the Bank Rate, Keynes displays bold confidence in this mechanism to smooth any credit cycle, to fill the gap between saving and investment and to correct all temporary monetary divergences from the long-run full employment equilibrium. However, Keynes's disaffection with the forced saving doctrine and the purely static nature of his fundamental equations drew sharp criticisms from Robertson and Hayek. Though from different standpoints, they both considered Keynes's credit cycle analysis as no more than an attempt to spell out the appropriate banking policy which could maintain a monetary equilibrium. In particular, Keynes's version of the credit cycle lacked, for the former, a proper sequential stability analysis and, for the latter, an explicit integration with capital theory.

Along lines very similar to Keynes's and, up to the late 1920s, in close cooperation with him, Robertson worked out a detailed sequential analysis of the interdependence of real and monetary magnitudes during the cycle. But clearly, for him, the cycle results from over-investment, this tendency to over-invest being a typical feature of decentralized economies stemming from the repercussions on the volume of investment of its gestation period. However, the largest part of Robertson's professional

output was devoted to studying the monetary or credit symptoms of such economic fluctuations, that is, how banks may respond to an increased demand for credit during expansion.

This led Robertson to a redefinition of the concept of saving in a monetary economy and to the rôle of this new concept in the cycle. This approach was linked with a sequential analysis of the lagged adjustments of output to monetary flows. In the ‘forced saving’ debate, central to all credit cycle theories, and contrary to Hayek who considered it as the villain of the piece, Robertson saw that phenomenon as only a relatively minor component of his theory, the factors at the root to his ‘credit inflation’ being the *real* cause of this expansion. Dragged among others by Keynes into endless discussions in the realm of monetary and interest theory, Robertson never managed however to offer an articulate and full-blown version of his theory of industrial fluctuations. In particular, the problem of the alteration in the structure of production, a question forming the core of Hayek's cycle theory, never received more than passing comment.

Grounded of course in the Austrian tradition and Wicksell's cumulative process (first extended by Mises, 1912, and Cassel, 1918), the distortion of the production time structure is absolutely central to Hayek's monetary cycle theory. The divergence between ‘natural’ and market rates of interest is linked by Hayek to the variability in forced saving and considered as the cause of cyclical fluctuations. Hayek's ‘additional credit’ theory places the cause of this gap between these two rates upon newly created money. The increase in loan capital resulting from a ‘trailing market rate’ makes investment surpass voluntary saving: a cumulative expansion results. Such an increase in investment alters the relative prices of capital and consumer goods in favour of the former. The increased output of capital goods distorts the production time structure. At a later stage, higher factor incomes drive up the demand for consumption goods, which through increased withdrawals from bank accounts will raise the market rate of interest and, finally, make some investment unprofitable. Then, the turnabout that takes place in the cycle brings a change in the other direction in the production structure, this time in favour of consumer goods. Clearly, crises are caused by over-investment, that is, by a decline in the desire to purchase the flow of capital goods coming on the market. The reversal of the process initiated by credit inflation does take place (as in most credit cycle theories) whenever the market rate catches up with prices; and since, sooner or later, banks run up against the limits set to their lending by their reserves, this process cannot be explosive (Fisher, 1911, also noticed, at least in his earliest writings, this stabilizing influence of the banking system).

Hayek's credit cycle theory thus marks a real break with what had come before. The theory of money is no longer a theory of the value of money ‘in general’ because relative prices may be changed by monetary influences and the Wicksellian full-employment assumption is dropped. The specific task of the trade cycle theorist is, for Hayek, to analyse short-period positions of the economy ‘in successive moments of time’ (1941, p. 23). The adoption of such an ‘intertemporal equilibrium’ approach to cycles (conceptually not different from modern temporary equilibrium) marks not only a crucial methodological turning point, but also the swan song of credit cycle theories.

On the one hand, this new method of ‘intertemporal equilibrium’ heralds the abandonment of the traditional framework in which cycles (defined as short-run disequilibria) are seen as temporary deviations from long-period equilibrium conditions determined by systematic and persistent forces at work in decentralized economies. In the present case, the ‘natural’ rate of interest determined in the long run by the supply of and the demand for capital is no longer the norm towards which the system is tending. It is in fact a property of such an ‘intertemporal equilibrium’ that not only will the price of the

same commodity be different at different points in time but also that the stock of capital will not yield a uniform ‘natural’ rate of interest on its supply-price.

On the other, the publication of Keynes's *General Theory* redirected research efforts away from this question into the problem of the determination of output at a point in time. It is only since the late 1960s, with the search for ‘microfoundations for macroeconomics’, and the subsequent advent of rational expectations and non-Walrasian equilibria, that this line of thought has been back on the theoretical agenda. However, given the extreme complexity of the problem and the relative crudeness of models still in their infancy, progress has so far been very modest.

See Also

- Hawtrey, Ralph George

Bibliography

Fisher, I. 1911. *The Purchasing Power of Money*. New York: Macmillan.

Hawtrey, R.G. 1913. *Good and Bad Trade*. London: Constable.

Hayek, F.A. von. 1929. *Monetary Theory and the Trade Cycle*. Trans. N. Kaldor and H.M. Crome, London: Jonathan Cape, 1933.

Hayek, F.A. von. 1941. *The Pure Theory of Capital*. London: Routledge.

Keynes, J.M. 1930. *A Treatise on Money*. Vol. 1, *The Pure Theory of Money*. As in *Collected Writings*, vol. 5, London: Macmillan, 1971.

Mises, L. von. 1912. *The Theory of Money and Credit*. Trans. H.E. Batson, London: Jonathan Cape, 1934.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Oxford University Press.

Tooke, T. 1844. *An Inquiry into the Currency Principle*. London: Longman, Brown, Green & Longmans.

Wicksell, K. 1898. *Interest and Prices*. Trans. R.F. Kahn, London: Macmillan, 1936.

How to cite this article

Bridel, P. "credit cycle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/

article?id=pde2008_C000429> doi:10.1057/9780230226203.0340

The New Palgrave Dictionary of Economics Online

credit rationing

Charles W. Calomiris and Stanley D. Longhofer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Credit rationing – a situation in which lenders are unwilling to advance additional funds to borrowers at the prevailing market interest rate – is now widely recognized as a problem arising because of information and control limitations in financial markets. This article reviews various motivations behind research on credit rationing, traces the history of theoretical efforts to explain how this phenomenon can persist in equilibrium, and reviews recent empirical research on its prevalence and effects. In the process, credit rationing is shown to be simply an extreme case of the more general problem of capital market misallocation.

Keywords

adverse selection; asset substitution; asymmetric information; availability doctrine; banking crises; bankruptcy; capital market misallocation; credit market discrimination; credit markets in developing countries; credit rationing; efficient allocation; financial intermediaries; financial repression; information economics; interest rate controls; limited liability; liquidity shocks; monetary policy transmission; moral hazard; rational expectations; separating contracts; usury

Article

Broadly speaking, ‘credit rationing’ refers to any situation in which lenders are unwilling to advance additional funds to a borrower even at a higher interest rate. In the words of Jaffee and Modigliani (1969, pp. 850–1), ‘credit rationing [is] a situation in which the demand for commercial loans exceeds the supply of these loans at the commercial loan rate quoted by the banks’. Key to this definition is that changes in the interest rate cannot be used to clear excess demand for loans in the market. In essence, this definition treats credit rationing as a supply side phenomenon, with the lender's supply function becoming perfectly price inelastic at some point.
If the projects that are being funded by the loan are not scalable, however, then a distinction must be

made between a situation in which a lender eventually restricts the size of loan it will provide to any individual borrower and one in which ‘rationed’ borrowers are denied credit altogether. This phenomenon arises in circumstances in which lending is not scalable. Stiglitz and Weiss (1981, pp. 394–5) therefore define credit rationing as follows:

> We reserve the term credit rationing for circumstances in which either (a) among loan applicants who appear to be identical some receive a loan and others do not, and the rejected applicants would not receive a loan even if they offered to pay a higher interest rate; or (b) there are identifiable groups of individuals in the population who, with a given supply of credit, are unable to obtain loans at any interest rate, even though with a larger supply of credit, they would.

According to this definition, lenders fully fund some borrowers but deny loans to others despite the fact that the latter are identical in the lender's eyes to those who receive loans.

Thus, there are two working definitions of credit rationing in the literature. The first focuses on situations in which increases in the interest rate cannot clear excess demand in the loan market, whether this excess demand reflects a single borrower (who would like a larger loan amount) or many. Under this definition, rationing would exist if every potential borrower received a loan but a smaller one than that desired at the equilibrium interest rate. The second definition – the Stiglitz–Weiss definition – restricts its attention to situations in which some borrowers are completely rationed out of the market, even though they would be willing to pay an interest rate higher than that prevailing in the market. Both of these definitions focus on the supply side of the market. One could argue, however, that it is useful to think of non-price rationing as any phenomenon that limits the amount of funding used by firms such that firms are not able to use the price mechanism to successfully bid for additional funds, whether this is caused by supply-side constraints (as under the narrow definitions of credit rationing described above) or by other distortions in credit markets (related, for example, to regulation). This would allow a broader definition of ‘credit rationing’ in which regulatory constraints, rather than just informational problems, lead to non-price allocations of credit.

Why care about credit rationing?

Early interest in credit rationing was driven in part by questions about the role that credit rationing might play in transmitting the macroeconomic effects of monetary policy, which was related to research on the so-called ‘availability doctrine’ in the 1950s and 60s (Scott, 1957). To the extent that monetary policy operates through a ‘credit channel’ (in which contractionary policy affects the economy through a decline in the supply of funds available for banks to lend), and to the extent that changes in the terms of lending include not only changes in loan pricing but also changes in the quantities of credit available to borrowers, credit rationing may play an important role in the transmission of monetary policy's effects on the economy (Blinder and Stiglitz, 1983).

In addition to the cyclical effects of rationing in credit markets related to monetary policy, development economists, especially Ronald McKinnon (1973), argued that a different credit rationing problem is more relevant for the long-term growth prospects of developing countries. High inflation, high zero-

interest reserve requirements, government-mandated loan allocations to favoured borrowers, and interest rate ceilings on loans or deposits in developing economies (a combination which McKinnon termed ‘financial repression’) subjected many developing countries' banking systems to an extreme form of regulation-induced credit rationing. High reserves, high inflation, and interest ceilings on deposits meant that banks were rationed in the deposit market, and thus had few funds to lend, while lending mandates and loan interest-rate ceilings meant that what funds were available to lend were often rationed by restrictions on who could bid for those funds.

Additionally, George Akerlof (1970), in his path-breaking article on the role of adverse selection in preventing market development, drew attention at an early date to the possible effects of information problems in retarding the development of lending markets, particularly in developing countries. In an ideal world, in the absence of any government policies limiting beneficial lending, all borrowers with positive net present value projects would be able to obtain outside funding (whether through debt or equity instruments, or bank or non-bank sources of funds). But Akerlof showed that, if markets were unable to distinguish good risks from bad ones, lending might not be feasible. The failure to develop institutions capable of producing credible information about borrowers and using that information to screen applicants could, according to Akerlof, play an important role in financial underdevelopment. Many development economists have come to recognize that the failure to properly allocate funds in the loan market – a broad phenomenon, within which credit rationing is a special and extreme case – can be an especially important potential impediment to growth in developing countries because of the relative absence of institutions in those countries that allow effective screening of borrowers (to mitigate adverse selection) or ongoing monitoring of borrowers' actions (to mitigate moral hazard).

An additional motivation for an interest in credit rationing comes from the literature on bank fragility. Credit rationing can also apply to the market in which financial intermediaries raise their funds. Financial institutions go to great pains to attract and maintain deposits through (*a*) the structure of their contracts (which typically afford withdrawal options to depositors), (*b*) their long-term relationships with market monitors who track their progress, and (*c*) their established reputations for good management. But sometimes the market suddenly decides to ration credit to a particular bank or to the whole banking system; and when this happens the affected banks find it hard to attract and maintain deposits at any price. Thus, the literature on ‘bank runs’ as an historical phenomenon can be thought of as a literature on credit rationing in the markets in which financial institutions raise their funds. Depositors that decide to participate in a bank run ration credit to their bank in the sense that the decision to withdraw is a quantity, not a price, decision. They are simply unwilling to leave their money in the bank.

Finally, much of the current research on discrimination in credit markets is driven by evidence that black and Hispanic minority loan applicants are denied more frequently than comparable whites (for example, Munnell et al., 1996; Cavalluzzo and Cavalluzzo, 1998; Cavalluzzo and Wolken, 2005). Of course, this begs the question of why borrowers are denied loans in the first place, rather than simply priced according to their risk. In other words, understanding why there are differences in denial rates across groups necessarily entails exploring why rationing (loan denial) occurs.

The development of credit rationing theory

Early views on credit rationing

The earliest discussions of credit rationing viewed it as a non-equilibrium phenomenon, arising either because of exogenous interest rate rigidities (for example, interest rate ceilings or usury laws) or because of a lack of competition in the loan market (Scott, 1957). Soon authors made a distinction between temporary credit rationing, in which market interest rates are slow to adjust to exogenous shocks such as changes in the lender's cost of funds or borrower demand, and ‘equilibrium’ credit rationing, which persists after the market has fully adjusted to these shocks. Clearly the more interesting and difficult to explain phenomenon is equilibrium credit rationing.

Hodgman (1960) was the first to try to explain how credit rationing can persist in a rational, equilibrium framework. In this model, lenders evaluate potential borrowers on the basis of the loan's expected return–expected loss ratio. In addition, it is assumes that there is a maximum repayment that the borrower can credibly promise, which effectively limits how much the lender will offer the borrower regardless of the interest rate: eventually the expected losses become too great relative to the expected return. This model was much debated in the ensuing years. In particular, Miller (1962) argued that Hodgman's analysis could be made consistent with rational expectations between the borrower and lender by incorporating bankruptcy costs that would be incurred by the lender upon the borrower's default. The real significance of the Hodgman article, however, was that it established as an important theoretical goal the objective of explaining how credit rationing could persist as an equilibrium phenomenon.

Freimer and Gordon (1965) resolved many of the issues regarding the structure of the Hodgman and Miller models by showing that credit rationing can occur with a risk-neutral lender if the borrower has a fixed-sized funding need. But this was done assuming an exogenous interest rate. Jaffee and Modigliani (1969) completed the picture by endogenizing the equilibrium interest rate by modelling both the supply and demand sides of the market. Credit rationing in their model, however, is the direct result of an exogenous assumption that borrowers within a given group must be charged the same interest rate, even though the lender can distinguish differences among them.

This early work was important in that it firmly established the idea that credit rationing could be a persistent equilibrium phenomenon. Ultimately, however, the solutions proposed relied on very restrictive assumptions about agent preferences or the contracts they could employ. More satisfactory explanations of credit rationing had to wait for the information economics revolution of the 1970s.

Modern credit rationing theory

Akerlof's (1970) pioneering article on adverse selection was motivated in part by the desire to explain extreme cases of credit rationing (the absence of a credit market), but Jaffee and Russell (1976) provide the first explicit asymmetric information rationale for credit rationing in the general sense. In their model, lenders cannot distinguish *ex ante* between high- and low-quality borrowers (that is, those who will repay their loans and those who will default). Contracts are written to determine the size of the loan offered and the interest rate. As in the Rothschild–Stiglitz (1976) insurance framework, low-quality borrowers must accept the contract that is preferred by the high-quality borrowers, lest they be identified as the deadbeats they are. Although a market-clearing interest rate/loan amount combination does exist, high-quality borrowers prefer a contract that entails a slightly lower interest rate with a reduced loan amount. As a result, the pooling outcome entails credit rationing. The primary problem with this model

is that the ‘equilibrium’ is not stable, in that unsustainable separating contracts dominate the pooling outcome.

In 1981, Joseph Stiglitz and Andrew Weiss published what has become the canonical model of credit rationing, because it was the first model that fully endogenized contract choices with a stable, rationing equilibrium. In the Stiglitz–Weiss framework, credit rationing occurs because the lender's expected return is not monotonically increasing in the interest rate. Instead, adverse selection or moral hazard problems eventually cause the lender's expected return to decline as the interest rate rises.

In the adverse selection version of the model, borrowers and lenders are both risk neutral. Borrowers are characterized by their projects, which are assumed to have the same expected returns but differ from one another in their risk. Specifically, borrower projects differ on the basis of mean-preserving spreads (Rothschild and Stiglitz, 1970). These projects are also assumed to require a fixed investment (that is, they are indivisible) and borrowers have a fixed amount of internal equity that they can invest in the project. Limited liability upon default means that the lender's payoff is a concave function of the project's return, while the borrower's profit function is convex.

These assumptions imply that, at any given interest rate, a subset of the least risky borrowers will drop out of the market, choosing instead to forgo their projects. In essence, the borrower's limited liability means that he reaps all of the project's gain (beyond the cost of debt service) when its return is high, but loses his collateral (his paid-in capital invested in the project, if any) only when the project's return is low. For low-risk projects, however, the potential upside gains are small. If those low-risk borrowers are pooled with high-risk borrowers, they will face higher than warranted interest rates. Low-risk borrowers will increasingly withdraw from the market as interest rates rise; as rates rise, borrowers with low-risk projects are better off withdrawing from the market and simply consuming their endowments rather than agreeing to invest and pay a high interest rate. As a result, increases in the interest rate cause more and more good borrowers to drop out of the market, lowering the average creditworthiness of the lender's remaining applicant pool. The size of the adverse selection premium faced by low-risk borrowers (the amount of interest low-risk borrowers have to pay in excess of what their project risks warrant) becomes larger with each interest rate rise because the interest rate must compensate for the default risk of an ever-worsening pool of borrowers.

Thus, increases in the interest rate affect lender returns in two ways. The first is the direct effect that a higher interest rate raises the lender's return (for a given pool of borrowers). Rising interest rates, however, also have the indirect effect of lowering the average quality of the lender's applicant pool, thereby lowering the lender's expected return from any given loan. Eventually, this secondary, adverse selection effect may outweigh the first interest rate effect, causing lender profits to decline as the interest rate rises.

Once the non-monotonicity of the lender's return in the interest rate is established, the possibility of credit rationing follows immediately. Profit-maximizing lenders will never voluntarily choose to raise the interest rate beyond where the adverse selection effect dominates. If excess demand exists in the market at this rate, credit rationing will be the equilibrium.

Paradoxically, in this model the very best credit risks do not seek funding because they do not find it worthwhile. This may seem odd, but it is important to remember that these borrowers are not rationed. Instead, they voluntarily drop out of the market because the cost of being pooled with higher-risk borrowers is too great. The rationed borrowers are the higher-risk borrowers who stay in the market and request funding.

Alternatively, Stigliz and Weiss show how changes in the interest rate may also affect the borrower's choice of project, so that moral hazard in project choice (sometimes referred to as ‘asset substitution’ in the finance literature) can be another reason that the lender's expected return is non-monotonic in the interest rate. Suppose that the borrower is able to choose among projects with different risk profiles. If, at a given interest rate, the borrower is indifferent between two projects, Stiglitz and Weiss show that an increase in the interest rate will cause the borrower to prefer the project that has the higher probability of default. Of course, the lender prefers the safer project. Thus (with slightly more restrictive distributional assumptions than in the adverse selection case), increases in the interest rate once again can eventually lower the lender's expected return, leading to credit rationing.

Models of credit rationing need not posit rationing for all borrowers. Realistically, some borrowers (certain firms for which information control problems are particularly acute) may be subject to rationing while other borrowers are not. Borrowers not subject to rationing may be able to avoid rationing because their prospects are more observable, or because their behaviour is more controllable.

Bank runs as credit rationing

The theoretical literature on credit rationing in the deposit market (bank runs) has some features that distinguish it from the literature on credit rationing in the loan market. The ultimate causes of deposit market rationing can be similar to, or very different from, the causes of loan market rationing. As discussed above, loan market rationing can reflect either information and incentive problems in the loan market or exogenous regulations. In the case of the deposit market, rationing can result either from incentive and information problems relating to the depositor–bank relationship or from exogenous liquidity needs of depositors.

With respect to the former, under some circumstances a bank run may reflect a loss of confidence in the market value of the bank's asset portfolio and changes in bank behaviour that attend such a loss. If the value of the portfolio falls sufficiently, and if the information and incentive problems are sufficiently severe, the perceived risk of losses in the bank can prompt depositors to ask for their money back because depositors have reason to be risk-intolerant (that is, to be unwilling to leave their money in a bank that has too high a level of risk). An example of such a model is Calomiris and Kahn (1991). Here the depositor withdraws funds in bad states of the world because doing so is necessary to prevent the banker from abusing his control over the bank's portfolio.

An alternative cause of credit rationing in the deposit market is a shock to the liquidity needs of depositors, which forces depositors to demand their funds from their banks irrespective of the portfolio performance of the banks. Diamond and Dybvig (1983) is an example of a model of this phenomenon.

Bank depositor runs are but one specific example of how financial intermediaries may be credit rationed due to creditor risk intolerance and/or liquidity shocks. During the 1998 Russian financial crisis, for example, it was widely reported that many emerging market hedge funds dumped their holdings of risky securities of all kinds in a scramble to reduce their risks and thus re-establish the high-quality credit ratings needed to retain their debtors. Intermediaries were also scrambling to accumulate liquidity, as many of their claimants needed to withdraw funds to meet other obligations related to the financial market upheaval.

The limits of credit rationing

Credit rationing as a problem of information and control (as it was modelled by Jaffee and Russell, 1976, and Stiglitz and Weiss, 1981) is properly seen as an extreme case of the more general phenomenon of capital market misallocation, which includes cases where capital is misallocated (due to adverse selection and moral hazard) without any rationing occurring. It is important to recognize that, from the standpoint of either cyclical concerns about the transmission of monetary policy or developmental concerns about the efficiency of the allocation of capital, the important phenomenon is not rationing per se but rather the extent to which the market fails to allocate resources efficiently. Even a market that never suffers from credit rationing can be highly inefficient in its allocation of capital. In that sense, credit rationing may be somewhat beside the point. Indeed, the corporate finance literature is full of examples of models of market imperfections involving moral hazard and adverse selection in which credit is misallocated, and in which positive net present-value projects are not funded or negative net present-value projects *are* funded.

In some cases, firms may even be priced out of the market for funds entirely, so that they avoid funding profitable investments. For example, Jensen and Meckling (1976) show that the potential for asset substitution at the expense of creditors can make it much more costly for firms to access debt markets. Indeed, asset substitution can make it prohibitively expensive to issue debt. Note that this is not a case of credit rationing as defined by Stiglitz and Weiss, since suppliers are not refusing credit. Rather, the high asset substitution premium that firms would be charged if they sought credit can result in a decision by the firm not to fund a positive net present-value investment. Similarly, Myers and Majluf (1984) show that because of adverse selection problems – which are particularly acute in the public equity market – some firms may decide to avoid issuing equity to fund a positive net present-value investment. Here, again, a firm is not being rationed by suppliers, but is unwilling to seek financing because of its prohibitive pricing.

As the literature on capital market misallocations and credit rationing developed in the late 1970s and early 1980s, critics pointed out some limiting circumstances in which capital markets did not have a tendency to underfund positive net present-value projects. For example, both adverse selection and moral hazard problems can be overcome by sufficient collateral. By placing collateral at risk a firm could signal its high quality, or commit itself not to abuse creditors by undertaking excessive risk (see Bester, 1985). Of course, collateral is not always available, nor is it costless to place collateral at risk. In the case of a limited liability enterprise, the firm's net worth limits its available collateral. Firms that can finance themselves from internal funds and limited amounts of low-risk debt can avoid the adverse selection and moral hazard costs associated with external finance, but young, growing firms tend to be in need of substantial amounts of external finance, far in excess of their accumulated net worth. If borrowers use all of their available ‘collateral’, then, on the margin, collateral cannot mitigate adverse selection or moral hazard problems.

In the consumer context, it is also important to recognize that the moral hazard and adverse selection problems that arise in corporate lending may differ in importance across the various areas of consumer lending. For example, moral hazard may be limited in the context of mortgage lending where actions destructive to the lender's interest are likely to harm the homeowner as well (consider inadequate protection against the risk of fire, for example). Furthermore, the modern use of credit scores and loan-

to-value ratios may make mortgage lenders more knowledgeable about an applicant's true credit risk than the applicant himself, particularly if that applicant has significant equity invested in the house and lacks experience in the credit market (Calomiris, Kahn and Longhofer, 1994). Under such circumstances, the implications of adverse selection models (which depend on the superiority of the information of the borrower about his type) may be irrelevant, or even reversed. On the other hand, in the context of uncollateralized credit card borrowing based only on past credit records, unobservably high-risk borrowers (those who know that they are about to have major medical costs, lose their job, or become divorced) may have strong incentives to borrow, implying the possibility for severe adverse selection.

How is credit rationing measured empirically?

Although credit rationing is a widely discussed phenomenon, there is a surprising paucity of evidence confirming its existence. The key problem is that, while the concept of a credit-rationed borrower is easy to understand in theory, under each of the various models of credit rationing discussed above it is extremely difficult to measure ‘excess demand’ of individual borrowers or the similitude of borrowers' creditworthiness.

Indirect methods

Jaffee and Modigliani (1969) attempt to infer the presence of credit rationing by measuring the proportion of new commercial loans originated at the prevailing prime rate and/or with very large loan sizes. The intuition they use is that prime and/or large borrowers have the lowest risk and are therefore the least likely to be rationed. As a result, a larger proportion of loans will go to these low-risk borrowers when credit rationing is severe. Jaffee and Modigliani use this proxy to see how market factors affect the prevalence of credit rationing. Of particular interest is their result that increases in the average commercial loan rate are associated with higher levels of rationing, which seems to confirm the appropriateness of their proxy for credit rationing.

Other authors have attempted to measure whether commercial loan rates are ‘sticky’ in response to changes in open-market interest rates. The idea here is that in most credit rationing models there is an implicit cap above which lenders will ration credit. As open-market rates rise, this cap is more likely to become binding, meaning that commercial loan rates will not fully respond to changes in open-market rates. Following this approach, a number of authors, including Goldfeld (1966) and Jaffee (1971), have found that commercial loan rates are, in fact, slow to adjust to changes in open-market rates, and offer this as evidence in support of credit rationing.

Berger and Udell (1992), however, provide convincing evidence that, although commercial-loan rate stickiness does occur, it does so in a fashion that is inconsistent with information-based credit rationing models. In particular, they find that nearly half of the observed loan rate stickiness occurs for loans made to borrowers who are exploiting a previously contracted bank loan commitment. Such borrowers are precluded from rationing by contract. Furthermore, they show that the fraction of loans made under commitment actually decreases during times of credit market tightness, exactly the opposite of what one would expect should credit rationing be an important phenomenon.

Direct methods

Other authors have attempted to directly measure credit rationing using survey data to identify ‘rationed’ borrowers. For example, Cox and Jappelli (1990) and Chakravarty and Scott (1999) use data from the Survey of Consumer Finances (SCF) in which households are directly asked whether they recently have been denied credit or been unable to obtain as much credit as they requested. Although these articles purport to measure how some outside factor affects the likelihood of being rationed, it is not clear that borrowers who self-report being denied credit have, in fact, been ‘rationed’ in the Stiglitz–Weiss meaning of the term. After all, their denial of credit could simply reflect a failure to properly select into the right risk class in order to be approved, or the fact that the borrower was simply uncreditworthy at any interest rate.

With regard to business lending, Cressy (1996) uses a sample of new businesses that opened accounts with a major British bank to ascertain whether credit rationing affects the likelihood of business survival. He concludes that firms self-select for finance based on the entrepreneur's human capital, implying that no credit rationing is occurring.

One strand of the empirical literature on credit rationing, broadly defined, focuses on whether differential mortgage loan denial rates between white and minority borrowers constitutes evidence of discrimination (a much cited reference is Munnell et al., 1996; Ross and Yinger, 2002, provide an excellent review of this literature). Although the discrimination literature does not specifically focus on the question of whether borrowers are credit rationed, any conclusion that one group is denied loans at a greater rate than others after creditworthiness is controlled for would imply that a form of credit rationing is occurring. This ‘rationing’, however, is distinct from that in Stiglitz–Weiss because the borrowers are not observably identical, and the underlying cause of ‘rationing’ is either lender preferences (Becker, 1971) or some form of statistical discrimination (Calomiris, Kahn and Longhofer, 1994; Longhofer and Peters, 2005).

Evidence on ‘intermediary rationing’

In contrast to the limited evidence of traditional borrower credit rationing, there is a significant body of evidence supporting the idea that financial institutions are rationed by their depositors. In recent years, a large literature has developed examining the determinants of deposit withdrawal from individual banks, and a parallel literature has developed on systemic banking panics. These articles find that in circumstances where the condition of banks is perceived to have deteriorated, depositors withdraw funds rather than simply demand a higher interest rate on deposits (Calomiris and Mason, 2003; Calomiris and Wilson, 2004). The links between bank characteristics and deposit withdrawals observed in these and other similar studies suggest that deposit rationing is related to information and incentive problems, rather than just liquidity shocks to depositors, although such shocks may still play a role.

Final thoughts

It is worth noting that improvements in underwriting processes may have dramatically altered the

practical impact of credit rationing in recent years. The use of risk-based pricing in consumer lending, including credit card loans and mortgages, has become widespread, reflecting the increased ability of lenders to distinguish between borrowers with different risk profiles (see, for example, Edelberg, 2003; Chomsisengphet and Pennington-Cross, 2006). The same is true for commercial credit markets, in which instruments such as junk bonds, senior-subordinated securitization issues, and the like serve to provide financial market access to broader classes of instruments, borrowers and risks. As a result, ‘sorting’ among borrowers overall has increased, and today there is likely much less diversity in pools of ‘observably identical’ borrowers than there was when Stiglitz and Weiss first developed their model. While this suggests that in some markets credit rationing is a very different and perhaps less important phenomenon today than it once was, an important potential role remains for credit rationing, particularly as it pertains to financial allocations in emerging markets, the pricing of particularly opaque segments of the lending markets of developed economies, and the ways in which financial institutions may be rationed in response to shocks to their portfolios.

See Also

- Akerlof, George Arthur
- banking crises
- micro-credit
- Stiglitz, Joseph E.

Bibliography

Akerlof, G.A. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Becker, G.S. 1971. *The Economics of Discrimination*, 2nd edn. Chicago: University of Chicago Press.

Berger, A.N. and Udell, G.F. 1992. Some evidence on the empirical significance of credit rationing. *Journal of Political Economy* 100, 1047–77.

Bester, H. 1985. Screening vs. rationing in credit markets with imperfect information. *American Economic Review* 75, 850–5.

Blinder, A.S. and Stiglitz, J.E. 1983. Money, credit constraints, and economic activity. *American Economic Review* 73, 297–302.

Calomiris, C.W. and Kahn, C.M. 1991. The role of demandable debt in structuring optimal banking arrangements. *American Economic Review* 81, 497–513.

Calomiris, C.W., Kahn, C.M. and Longhofer, S.D. 1994. Housing finance intervention and private

incentives: helping minorities and the poor. *Journal of Money, Credit and Banking* 26, 634–74.

Calomiris, C.W. and Mason, J.R. 2003. Fundamentals, panics, and bank distress during the depression. *American Economic Review* 93, 1615–47.

Calomiris, C.W. and Wilson, B. 2004. Bank capital and portfolio management: the 1930s ‘capital crunch’ and the scramble to shed risk. *Journal of Business* 77, 421–55.

Cavalluzzo, K.S. and Cavalluzzo, L.C. 1998. Market structure and discrimination: the case of small businesses. *Journal of Money, Credit, and Banking* 30, 771–92.

Cavalluzzo, K. and Wolken, J. 2005. Small business loan turndowns, personal wealth, and discrimination. *Journal of Business* 78, 2153–77.

Chakravarty, S. and Scott, J.S. 1999. Relationships and rationing in consumer loans. *Journal of Business* 72, 523–44.

Chomsisengphet, S. and Pennington-Cross, A. 2006. The evolution of the subprime mortgage market. *Federal Reserve Bank of St. Louis Review* 88(1), 31–56.

Cox, D. and Jappelli, T. 1990. Credit rationing and private transfers: evidence from survey data. *Review of Economic Statistics* 72, 445–54.

Cressy, R. 1996. Are business startups debt-rationed? *Economic Journal* 106, 1253–70.

Diamond, D.W. and Dybvig, P.H. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91, 401–19.

Edelberg, W. 2003. Risk-based pricing of interest rates in household loan markets. FEDS Working Paper No. 2003–62.

Ferguson, M.F. and Peters, S.R. 2000. Is lending discrimination always costly? *Journal of Real Estate Finance and Economics* 21, 23–44.

Freimer, M. and Gordon, M.J. 1965. Why bankers ration credit. *Quarterly Journal of Economics* 79, 397–416.

Goldfeld, S.M. 1966. *Commercial Bank Behavior and Economic Activity: A Structural Study of Monetary Policy in the Postwar United States*. Amsterdam: North-Holland.

Hodgman, D.R. 1960. Credit risk and credit rationing. *Quarterly Journal of Economics* 74, 258–78.

Jaffee, D.M. 1971. *Credit Rationing and the Commercial Loan Market*. New York: Wiley.

Jaffee, D.M. and Modigliani, F. 1969. A theory and test of credit rationing. *American Economic Review* 59, 850–72.

Jaffee, D.M. and Russell, T. 1976. Imperfect information, uncertainty, and credit rationing. *Quarterly Journal of Economics* 90, 651–66.

Jensen, M.C. and Meckling, W.H. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

Longhofer, S.D. and Peters, S.R. 2005. Self-selection and discrimination in credit markets. *Real Estate Economics* 33, 237–68.

McKinnon, R.I. 1973. *Money and Capital in Economic Development*. Washington, DC: Brookings Institution.

Miller, M.H. 1962. Credit risk and credit rationing: further comments. *Quarterly Journal of Economics* 76, 480–8.

Munnell, A.H., Tootell, G.M., Browne, L.E. and McEneaney, J. 1996. Mortgage lending in Boston: interpreting HMDA data. *American Economic Review* 86, 25–53.

Myers, S.C. and Majluf, N.S. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13, 187–221.

Ross, S.L. and Yinger, J. 2002. *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. Cambridge, MA: MIT Press.

Rothschild, M. and Stiglitz, J.E. 1970. Increasing risk I: a definition. *Journal of Economic Theory* 2, 225–43.

Rothschild, M. and Stiglitz, J.E. 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90, 630–49.

Scott, I.O. 1957. The availability doctrine: theoretical underpinnings. *Review of Economic Studies* 25, 41–8.

Stiglitz, J.E. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

How to cite this article

Calomiris, Charles W. and Stanley D. Longhofer. "credit rationing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000438>
doi:10.1057/9780230226203.0341

The New Palgrave Dictionary of Economics Online

crime and the city

Yves Zenou

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Crime is unevenly distributed across space and tends to be concentrated in poor areas. Recent theoretical advances show that social interactions and peer effects can explain this pattern because of contagion effects and social multipliers. An individual is more likely to commit crime if his or her peers commit crime than if they do not. Recent empirical findings suggest that, indeed, social interactions and networks are key to understand criminal behaviour in cities.

Keywords

Becker, G.; black–white wage differences; Bonacich centrality measure; contagion effects; cost–benefit analysis; crime and the city; crime rates; crime, economic theory of; social multiplier effects; social networks

Article

Crime is defined as an act committed in violation of a law forbidding it and for which punishment is imposed upon conviction. Crime is, however, not evenly distributed across space as it tends to be concentrated in specific areas where people are generally poor and uneducated. In both the United States and Europe, the typical urban pattern is that large cities have higher crime rates than smaller cities, and poor, largely minority neighbourhoods experience higher crime rates than more affluent white neighbourhoods (Raphael and Sills, 2005). According to the United Nations Interregional Crime and Justice Research Institute, (see Alvazzi del Frate, 1997), the percentage of population who are victims of burglary in urban areas with more than 100,000 inhabitants over a five-year period (between 1992 and 1996) is: 16 for Western Europe, 24 for North America, 20 for South America, 18 for Eastern Europe, 13 for Asia and 38 for Africa. Another typical pattern common to both the United States and Europe is that ethnic minorities are overrepresented in criminal activities. In the United States, the proportion of 20–29-year-old black men directly in trouble with the law (in jail or prison or on probation or parole)

reached 23 per cent in 1989 (Freeman, 1999). There is, however, one notable difference. Since the mid-1980s, crime has declined in the United States but increased in Europe, especially in large urban areas (Blumstein and Wallman, 2000).

Theories

In the standard crime model (Becker, 1968), each individual has to implement a cost–benefit analysis in order to choose between becoming a criminal and participating in the labour market. The cost is the severity of punishment, which obviously depends on the probability of being arrested. The benefit consists in the proceeds from crime. If crime is localized, then criminals will trade off a lower probability of being arrested (since, in some areas, a host of criminals are active and the number of policemen is not sufficient) against lower proceeds from crime (more criminals also imply less booty).

In this context, Sah (1991) examines the influence of the social environment on individuals' perceptions of the probability of arrest. Indeed, people develop their ideas about the relative benefits and costs of crime based on the observations they make every day. If a person lives in an area with a high crime rate, and particularly if the criminals are seen to be relatively successful, then that person is more likely to engage in criminal activity. The main result of this paper is that individuals in some areas tend to commit more crime than the Beckerian model would predict because of the gap between the perceived and the real cost of committing crime, which leads to a lower sense of impunity based on the information provided by their criminal friends.

Another approach (Verdier and Zenou, 2004) proposes that distance to jobs plays a role in crime behaviour and provides a unified explanation for why blacks commit more crime, are located in poorer neighbourhoods and receive lower wages than whites. The mechanism is as follows. If everybody believes that blacks are more prone to crime than whites, even if there is no basis for this, then blacks are offered lower wages and, as a result, locate further away from jobs. Because distant residence implies more tiredness and higher commuting costs, the black–white wage gap is widened further. Blacks have thus a lower opportunity cost of committing crime (lower outside option) and become indeed more criminal than whites. The loop is closed and the beliefs are self-fulfilling.

Whereas the standard Beckerian approach focuses on individual behaviour, Glaeser, Sacerdote, and Scheinkman (1996) stress the role of peers and social interactions in criminal activities, especially in urban areas because of the high variance in crime rates. Two types of individuals are assumed: those who, as in the standard model, base their crime decision on a cost–benefit analysis, and those who only imitate their neighbours. Because of these social interactions, the benefits from crime are greater than in the Beckerian model. Moreover, if these interactions are localized (as is usually the case), then it becomes easy to explain very high levels of crime in some areas of the city. Indeed, if there are already a lot of criminals in a particular location, then crime becomes 'contagious' by spreading like a virus and amplifies the number of criminals in this location. There are social *multiplier* effects through a feedback loop: negative social behaviour such as crime leads to more negative social behaviour.

Calvó-Armengol and Zenou (2004), and Ballester, Calvó-Armengol and Zenou (2004) propose a model along these lines but represent social interactions in terms of a social network of criminal friends. People in a network not only imitate but also influence each other. Here, the cost of committing crime is reduced thanks to the network of friends. Indeed, delinquents learn from other criminals belonging to the

same network how to commit crime in a more efficient way by sharing the know-how about the ‘technology’ of crime. They show that the influence of peers on the individual's criminal activity depends on his or her position in the network, and each agent's criminal effort is proportional to his or her Bonacich centrality measure (see Bonacich, 1987). For a given network, the Bonacich network centrality counts, for each agent, the total number of direct and indirect paths of any length in the network stemming from this agent. Such paths are weighted by a geometrically decaying factor (with path length). In other words, the ‘location’ of each individual in a network of friends, as measured by the Bonacich centrality measure, is a key determinant of his or her criminal activity.

As a result, in a spatial or social context, an efficient policy aiming at reducing crime would not be, as in the Beckerian model, to increase at random the cost of committing crime, but rather to target criminals according to their location in the urban or social space. Ballester, Calvó-Armengol and Zenou (2004) propose a policy that consists in finding and getting rid of the key player, that is, the criminal who, once removed, leads to the highest aggregate crime reduction. They show that the key player is not necessarily the most active criminal (that is, the one with the highest Bonacich centrality). Indeed, removing a criminal from a network has both a direct and an indirect effect. The direct effect is that fewer criminals contribute to the aggregate crime level. The indirect effect is that the network topology is modified, and the remaining criminals adopt different crime efforts. The key player is the one with the highest overall effect.

Empirical studies

One of the first tests of the Becker model was undertaken by Ehrlich (1973), who used as explanatory variables the imprisonment rate and the average sentence for the crime in question. More recently, the focus has been on urban or social problems because this is particularly fruitful for understanding personal and property crime as opposed to white-collar crime. Cullen and Levitt (1999), using data for 137 US cities from 1976 to 1993, explore the relationship between crime and urban flight (that is, the flight of the white population from city centres to suburbs). They find that each additional reported crime in city centre is associated with a net decline of about one resident. Causality runs from rising crime rates to city depopulation. Pursuing this area of research, Glaeser and Sacerdote (1999) provide three reasons for higher crime rates in big cities. They report that 27 per cent of the difference between urban and rural crime rates in the United States is due to higher pecuniary benefits for crime in cities, 20 per cent to a lower probability of arrest and recognition in cities, and the remaining 45–60 per cent to the observable characteristics of individuals. This last number can be explained by a positive covariance across agents’ decisions about crime, so that the variance of crime rate is higher than the variance predicted by local conditions. This implies that social interactions should matter, especially in cities. Case and Katz (1991) were among the first to investigate this last issue. Using data from the 1989 NBER survey of youths living in low-income Boston neighbourhoods, they find that the behaviours of neighbourhood peers appear to substantially affect youth behaviours in a manner suggestive of contagion models of neighbourhood effects. The direct effect of moving a youth with given family and personal characteristics to a neighbourhood where 10 per cent more of the youths are involved in crime than in his or her initial neighbourhood is to raise the probability the youth will become involved in crime by 2.3 per cent.

Glaeser, Sacerdote and Scheinkman (1996) find that, across crimes, crime committed by younger people has higher degrees of social interaction, while, across cities, for serious crimes in general and for larceny and auto theft in particular, the degree of social interactions is larger in those communities where families are less intact, that is, have more female-headed households. Ludwig, Duncan and Hirschfield (2001) and Kling, Ludwig and Katz (2005) explore this last result by using data from the Moving to Opportunity (MTO) experiment that assigned a total of 638 families from high-poverty Baltimore neighbourhoods into three 'treatment groups': (*a*) Experimental group families receive housing subsidies, counselling and search assistance to move to private-market housing in low-poverty census tracts; (*b*) Section 8-only comparison group families receive private-market housing subsidies with no programme constraints on relocation choices; and (*c*) a Control group receives no special assistance under MTO. They show that relocating families from high- to low-poverty neighbourhoods reduces juvenile arrests for violent offences by 30 to 50 per cent of the arrest rate for control groups. This also suggests very strong social interactions in crime behaviours.

Using a very detailed data-set of friendship networks in the United States from the National Longitudinal Survey of Adolescent Health (AddHealth), Calvó-Armengol, Patacchini and Zenou (2005) test the main results of Ballester, Calvó-Armengol and Zenou (2004). Contrary to the standard approach, here peer effects are conceived not as an average intra-group externality that affects identically all the members of a given group, but as a collection of dyadic bilateral relationships, which constitutes a social network. The position and thus the centrality of each individual are thus crucial to understand criminal behaviour. Calvó-Armengol, Patacchini and Zenou (2005) show that, after observable individual characteristics and unobservable network specific factors are controlled for, the individual's position in a network (as measured by his or her Bonacich centrality) is a key determinant of his or her level of criminal activity. A standard deviation increase in the Bonacich centrality increases the level of individual delinquency by 45 per cent of one standard deviation.

See Also

- law, economic analysis of
- neighbours and neighbourhoods
- racial profiling
- residential segregation
- social interactions (theory)
- social multipliers
- social networks in labour markets
- spatial mismatch hypothesis
- urban economics

Bibliography

Alvazzi del Frate, A. 1997. Preventing crime: citizens' experience across the world. Issues and Reports

No. 9. Rome: United Nations Interregional Crime and Justice Research Institute.

Ballester, C., Calvó-Armengol, A. and Zenou, Y. 2004. Who's who in crime networks. Wanted: the key player. Discussion Paper No. 4421. London: CEPR.

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Blumstein, A. and Wallman, J. 2000. *The Crime Drop in America*. New York: Cambridge University Press.

Bonacich, P. 1987. Power and centrality: a family of measures. *American Journal of Sociology* 92, 1170–82.

Calvó-Armengol, A., Patacchini, E. and Zenou, Y. 2005. Peer effects and social networks in education and crime. Discussion Paper No. 5244. London: CEPR.

Calvó-Armengol, A. and Zenou, Y. 2004. Social networks and crime decisions: the role of social structure in facilitating delinquent behavior. *International Economic Review* 45, 939–58.

Case, A. and Katz, L. 1991. The company you keep: the effects of family and neighborhood on disadvantaged youths. Working Paper No. 3705. Cambridge, MA: NBER.

Cullen, J. and Levitt, S. 1999. Crime, urban flight, and the consequences for cities. *Review of Economics and Statistics* 81, 159–69.

Ehrlich, I. 1973. Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of Political Economy* 81, 521–65.

Freeman, R. 1999. The economics of crime. In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Glaeser, E. and Sacerdote, B. 1999. Why is there more crime in cities? *Journal of Political Economy* 107, S225–S258.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111, 508–48.

Kling, J., Ludwig, J. and Katz, L. 2005. Neighborhood effects on crime for female and male youth: evidence from a randomized housing voucher experiment. *Quarterly Journal of Economics* 120, 87–130.

Ludwig, J., Duncan, G. and Hirschfield, P. 2001. Urban poverty and juvenile crime: evidence from a randomized housing-mobility experiment. *Quarterly Journal of Economics* 116, 655–79.

Raphael, S. and Sills, M. 2005. Urban crime in the United States. In *A Companion to Urban Economics*, ed. R. Arnott and D. McMillen. Boston: Blackwell.

Sah, R. 1991. Social osmosis and patternsof crime. *Journal of Political Economy* 99, 1272–95.

Verdier, T. and Zenou, Y. 2004. Racial beliefs, location and the causes of crime. *International Economic Review* 45, 731–60.

How to cite this article

Zenou, Yves. "crime and the city." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_U000068> doi:10.1057/9780230226203.0342

The New Palgrave Dictionary of Economics Online

cross-cultural experiments

Rob Boyd
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Experiments conducted in student populations suggest that people are not money maximizers, but also seem to have social preferences. To determine whether these social preferences are culturally variable, a group of economists and anthropologists undertook a series of economic experiments in a wide range of non-Western, small scale societies. Results in these societies were highly variable, and in some of them strikingly different from experiments in student populations. Variation in behaviour was correlated with societal characteristics, but not individual attributes. Finally, variation in punishment across societies predicted variation in cooperation across societies.

Keywords

altruism; cooperation; culture and economics; dictator game; economic experiments; experimental games; group characteristics; public goods games; reciprocity; social preferences; third-party punishment game; ultimatum game

Article

A large number of well-replicated results using a wide variety of experimental games are inconsistent with the assumption that people are money maximizers. Instead, people's behaviour is consistent with choices based on social preferences in which people place a positive value on fairness, reciprocity, or equity (see Camerer, 2003, for a review). For example, subjects typically make significant positive contributions in the public goods games, reject positive offers in the ultimatum game, and impose costly punishment in the third-party punishment game (see, Camerer, ch. 2, for descriptions of these games.) In some games these results are insensitive to framing and whether behaviour is anonymous to the experimenter (‘double blind’).

These experiments are open to two qualitatively different interpretations: It could be that pro-social behaviours like cooperation in the public goods game and punishment in the third-party punishment game reflect human nature. Cooperation in the public goods game could result from universal cognitive systems that cause people everywhere to behave as if all acts have reputational consequences, even when facts suggest no one will know what they have done. Punishment in the third-party punishment game could result from a pan-human motivational system that causes people to prefer outcomes that are fair or mutually beneficial, and to derive satisfaction from punishing unfair behaviour. However, with few exceptions experimental subjects have been university students in urbanized, industrial societies. Thus, it also could be that observed pro-social behaviour results from culturally evolved beliefs and values that are specific to such social environments. It is obviously of great importance to determine which of these two interpretations is correct.

To answer this question, a team of anthropologists and economists performed two rounds of experimental games in a wide range of cultural environments. The first round (Henrich et al., 2004; 2005) comprised a diverse group of 15 societies including peoples like the Aché and Hadza who live in nomadic foraging bands, the Achuar and Au who live in small villages and mix hunting and horticulture, Mongol and Sangu pastoralists, and sedentary Shona farmers in Zimbabwe. The ultimatum game was performed in all 15 societies, and the public goods game and the dictator game were performed in different subsets. The second round (Henrich et al., 2006) included a similar and overlapping range of 15 societies. Based on experience in the first round, experimental protocols were improved and standardized, and a greater effort was made to collect standardized data on individual characteristics. During the second round the ultimatum, dictator, and third-party punishment games were performed in all 15 societies. In addition complete strategies for second players in the ultimatum game and punishers in the third-party punishment game were elicited using the strategy method.

These experiments reveal a number of interesting results.

1. *Behaviour in non-Western populations can be quite different from that of Western university subjects*. Figure 1 shows the distribution of ultimatum game offers in the first round of experiments. The Pittsburgh data taken from Roth et al. (1991) are typical for university populations – the modal offer is 50 per cent but many subjects make somewhat lower offers. Behaviour in other populations can be very different. For example, modal offers are much lower among two lowland tropical forest groups; the Achuar and the Machiguenga are quite low. Interestingly, these very low offers were usually accepted, behaviour much closer to the predictions of money maximization than the behaviour of Western university subjects. Non-western populations also exhibited novel behaviours not seen in university populations. Figure 2 shows the rejection probabilities for different ultimatum game offers. Notice that in several populations increasing offer level above 50 per cent *increased* the rate of rejections, a phenomenon not observed among student subjects.

Figure 1

Ultimatum game offer. *Note*: A bubble plot showing the distribution of ultimatum game offers for each group. The diameter of the circle at each location along each row represents the proportion of the sample that made a particular offer. The right edge of the lightly shaded horizontal grey bar is the mean offer for that group. In the Machiguenga row, for example, the mode is 0.15, the secondary mode is 0.25, and the mean is 0.26. *Source*: Henrich et al. (2005).

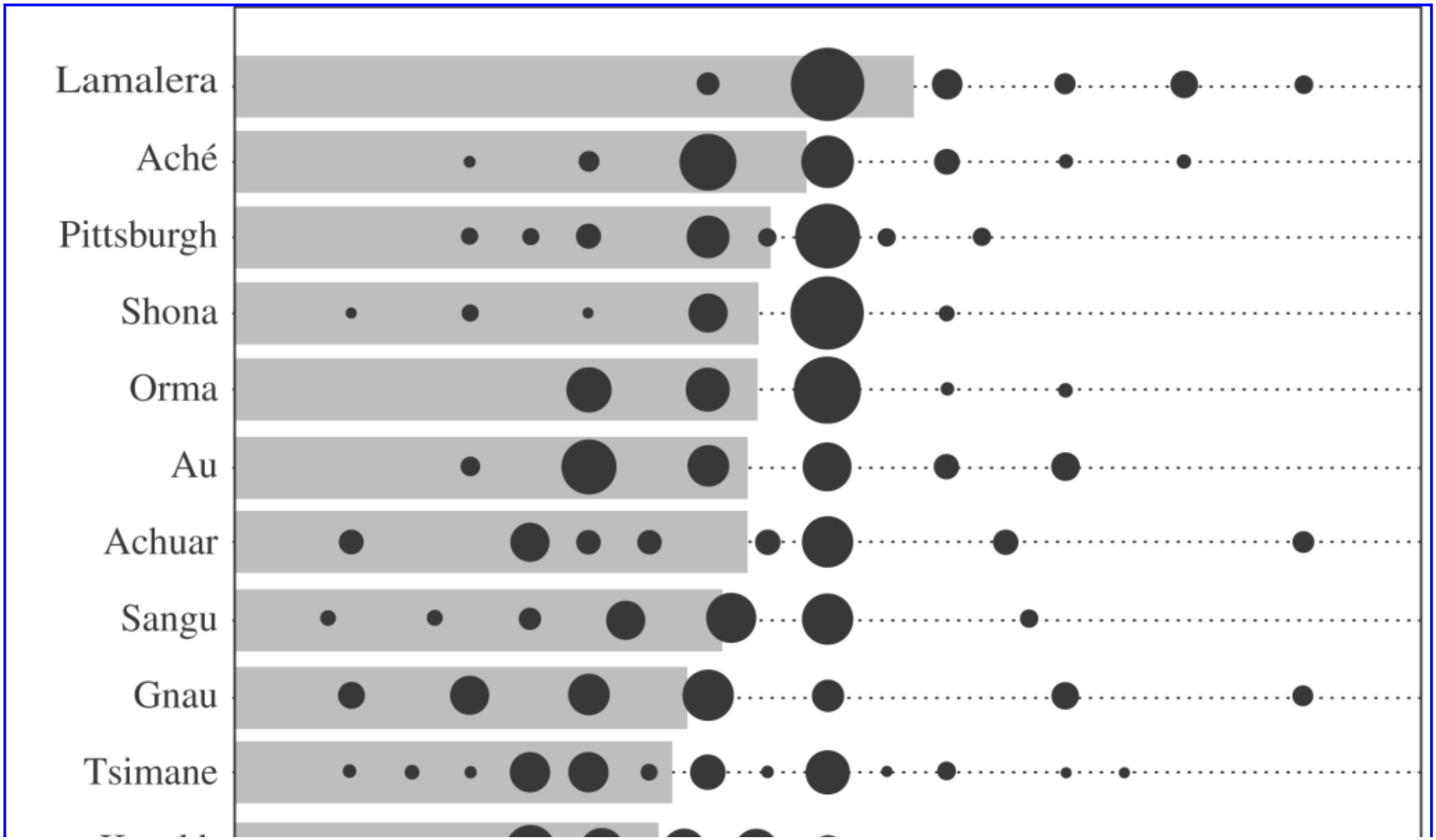

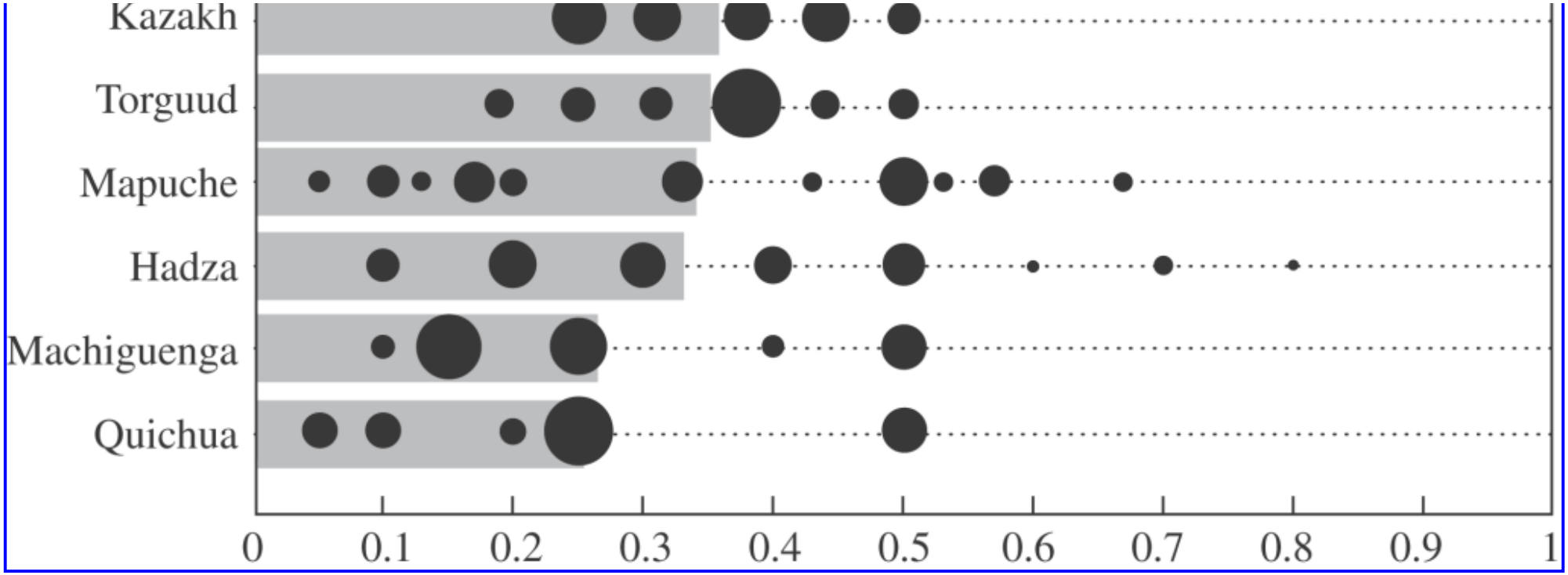

Figure 2

Ultimatum game rejection rates. *Note*: The diameter of the black circles is proportional to the fraction of offers that would have been rejected in the ultimatum game during the second round of experiments plotted as a function of the offer as a percentage of the maximum offer. For scale, note that the Gusii and Maragoli rejected all offers of zero. Notice that in all societies offering 50% of the stake minimizes the probability of rejection, but that in a number of societies increasing offers above 50% increases the rate of rejection. *Source*: Henrich et al. (2006).

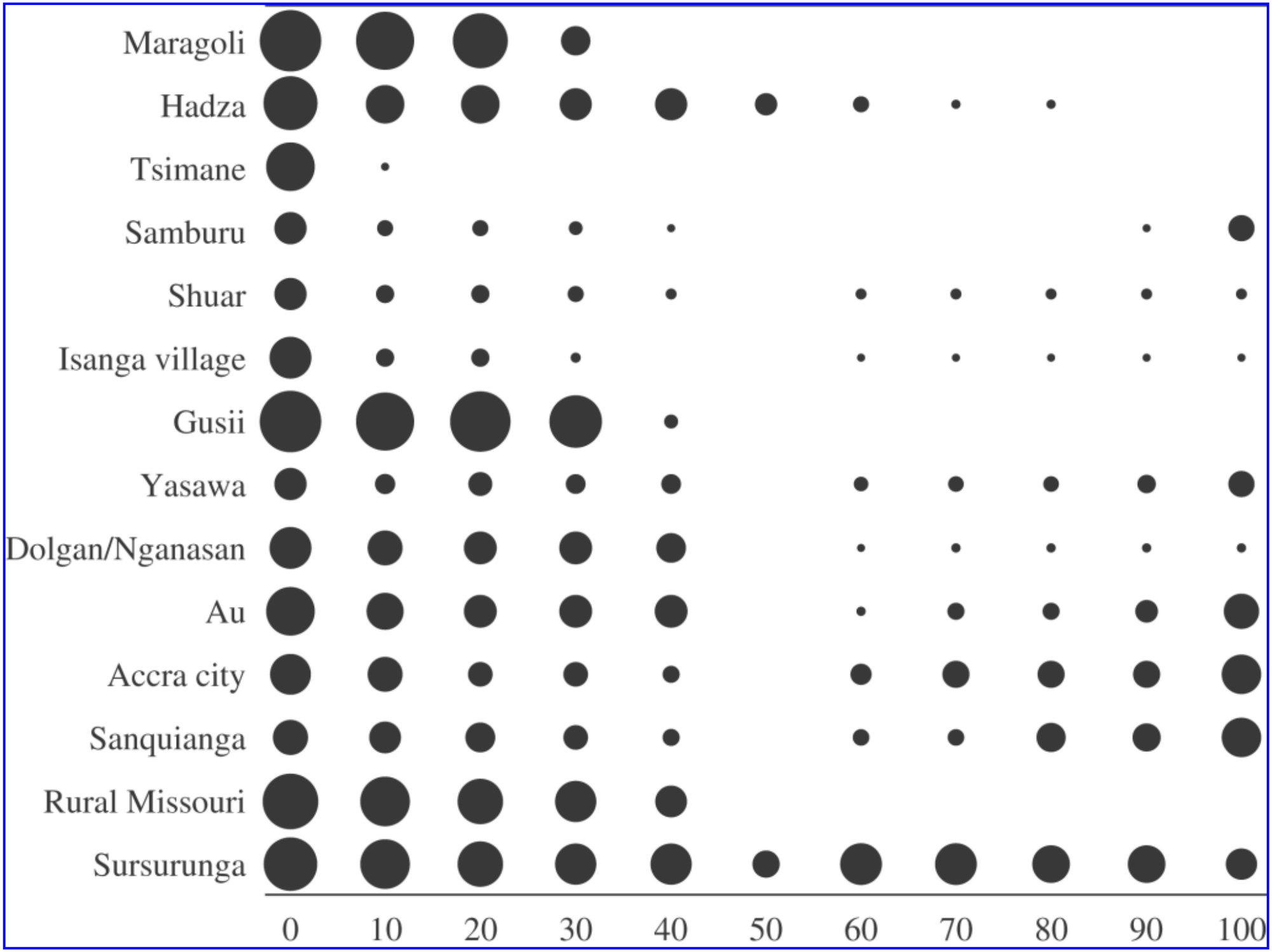

2. *Behavioural differences are correlated with group characteristics but not individual characteristics*. The ethnographers who performed most of these experiments have studied

these groups for many years and have detailed data on subjects about income, wealth, education, market contact, and a variety of other factors. None of these factors was significantly correlated with ultimatum game offers within social groups in first round, or offers or rejections in the second round. Because measures of wealth, income, and so on are not comparable across groups, these measures could not be aggregated to derive group characteristics. However, during the first round, ethnographers who were blind to the results ranked each of the groups along five dimensions: extent of cooperation in subsistence, degree of market contact, amount of privacy, amount of anonymity, and social complexity. We also had comparable data on settlement size. It turned out that market contact, settlement size, and social complexity were all highly correlated, so these were collapsed into a single variable labelled 'aggregate market contact'. Multiple linear regression showed that increasing aggregate market contact and cooperation in subsistence significantly predicted increased ultimatum game offers, and together the two variables accounted for more than half of the variance among groups in average offers.

3. *Variation in punishment predicts variation in altruism across societies*. In the third-party punishment game, an individual, the 'punisher' observes a dictator game and can punish the dictator at a cost to him or herself. The average minimum offer acceptable to the punisher in this game provides a measure of the level of punishment in that society. As is shown in Figure 3, this measure of punishment also predicts the level of altruism measured by dictator offers in the ordinary dictator game.

Figure 3

Mean minimum acceptable offer, third-party punishment game. *Note*: The mean offer in the dictator game for a society plotted against the mean value of the minimum acceptable offer in the third-party punishment game. The different symbols indicate continents. The size of each symbol is proportional to the number of DG pairs at each site. The dotted line gives the weighted regression line, with continental controls of mean dictator game offers against mean minimum acceptable offer in the third-party punishment game. *Source*: Henrich et al. (2006).

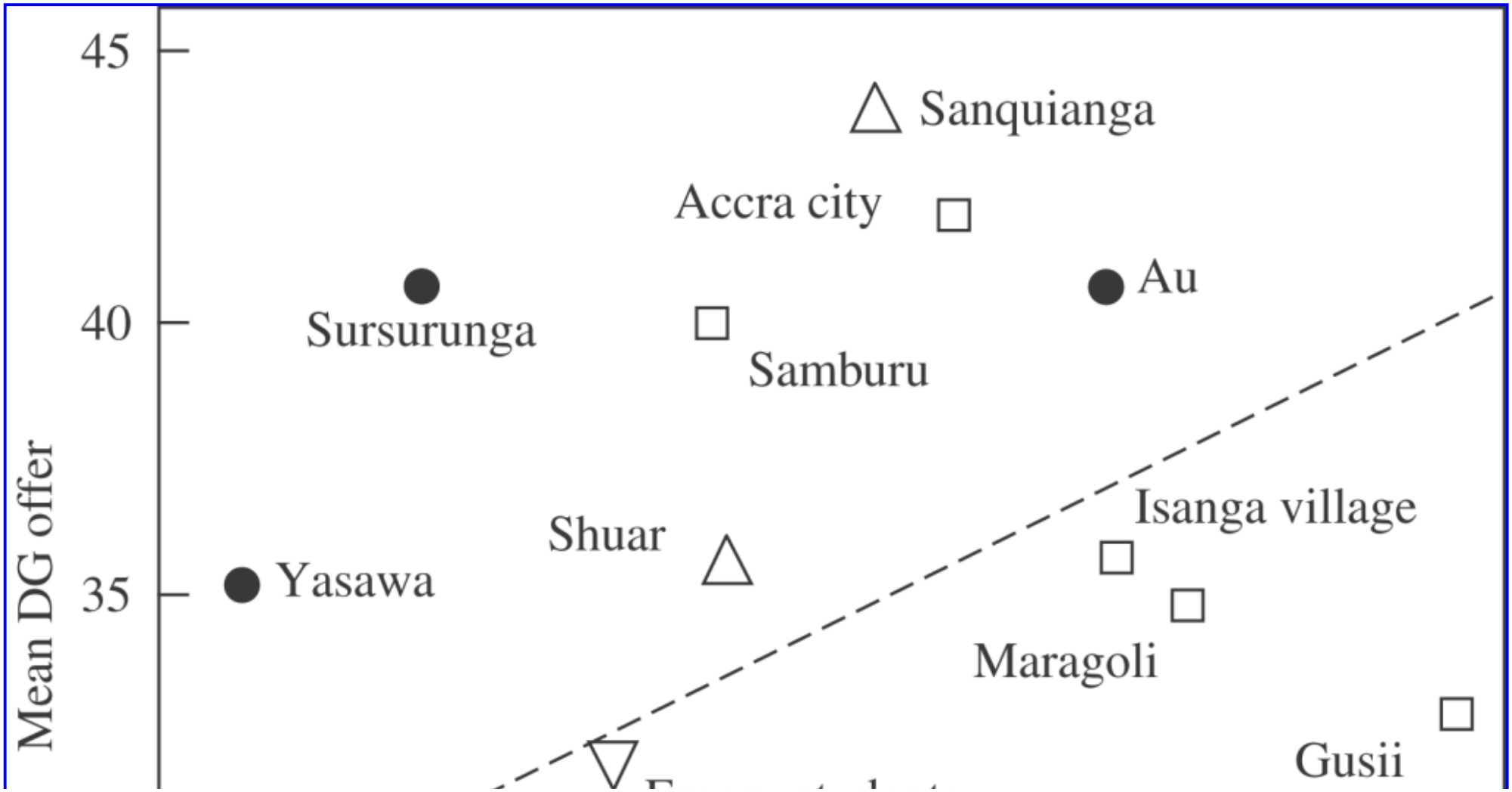

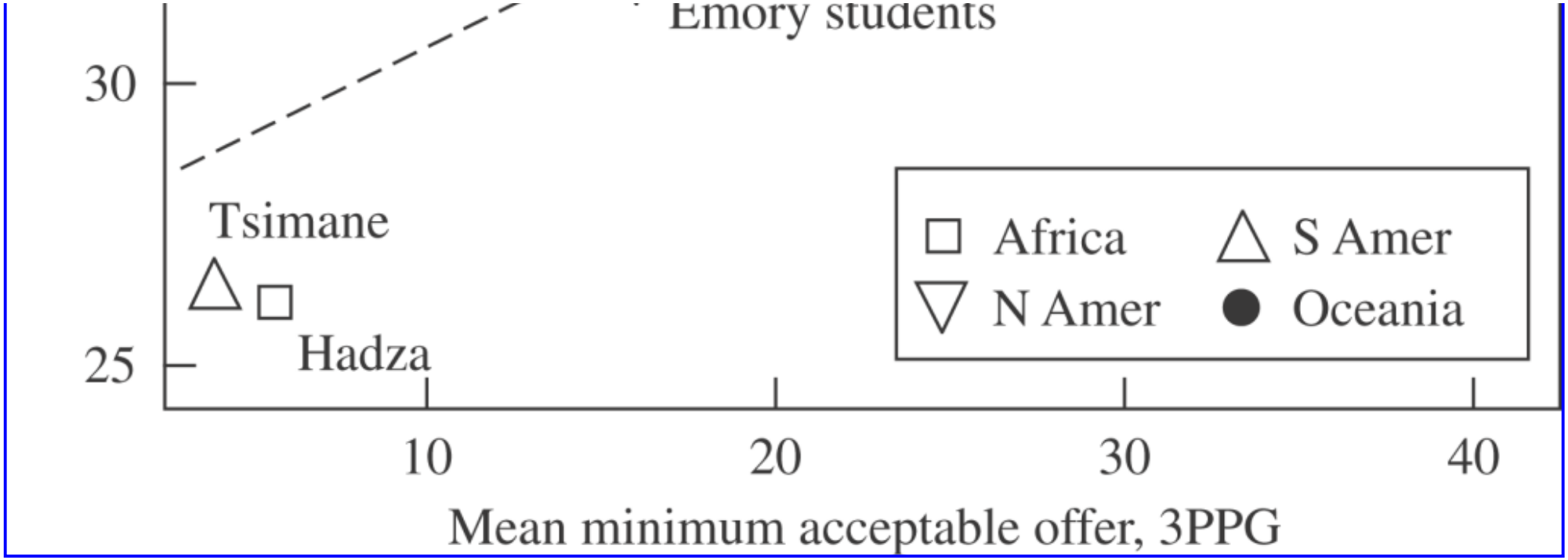

Taken together these results indicate that pro-social behaviour in economic experiments does not result from an invariant property of our species, and instead suggest that there are significant cultural differences between societies. The fact that ultimatum game behaviour is predicted by the average level of cooperation and average level of market contact further indicates that these cultural differences are not arbitrary, but may reflect economic, ecological and social differences between societies. However, the lack of correlation between individual characteristics and individual behaviour indicates that the differences between societies are not likely to be explained as the simple aggregation of individual experiences. Instead, it is more plausible that cultures evolve over time in response to the average conditions which they face, and that individual behaviour is, in turn, shaped by these cultural differences.

See Also

- experimental economics

Bibliography

Camerer, C. 2003. *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton: Princeton University Press.

Henrich, J.R.B., Bowles, S., Camerer, C., Fehr, E. and Gintis, H. 2004. *The Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. New York: Oxford University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Hill, K., Gil-White, F., Gurven, M., Marlowe, F., Patton, J.Q., Smith, N. and Tracer, D. 2005. ‘Economic man’ in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28, 795–855.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. and Ziker, J. 2006. Costly punishment across human societies. *Science* 312, 1767–70.

Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review* 81, 1068–95.

How to cite this article

Boyd, Rob. "cross-cultural experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000596> doi:10.1057/9780230226203.0343

The New Palgrave Dictionary of Economics Online

crowding out

Olivier J. Blanchard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Crowding out’ refers to all the things which can go wrong when debt-financed fiscal policy is used to affect output. While the initial focus was on the slope of the LM curve, ‘crowding out’ now refers to a multiplicity of channels through which expansionary fiscal policy may in the end have little, no or even negative effects on output.

Keywords

accumulation of capital; aggregate demand; budget deficits; crowding out; fiscal consolidation; fiscal expansion; fiscal policy; flexible exchange rates; full employment; inflation; interest rates; intertemporal taxation; investment tax credit; IS–LM model; labour supply; lump sum taxes; multiplier analysis; Mundell–Fleming model; natural rate of unemployment; private spending; public debt; public expenditure; Ricardian equivalence theorem; risk premium; taxation of income

Article

‘Crowding out’ refers to all the things which can go wrong when debt-financed fiscal policy is used to affect output.

A first line of argument questions whether fiscal policy has any effect at all on spending. Changes in the pattern of taxation which keep the pattern of spending unaffected do not affect the intertemporal budget constraint of the private economy and thus may have little effect on private spending. This argument, known as the ‘Ricardian equivalence’ of debt and taxation, holds only if taxes are lump sum (Barro, 1974). Some taxes which induce strong intertemporal substitution, such as an investment tax credit for firms, will have stronger effects if they are temporary; for most others, such as income taxes, changes in the intertemporal pattern may have only a small effect on the pattern of spending.

The Ricardian equivalence argument is not settled empirically and its validity surely depends on the circumstances. A change in the intertemporal taxation of assets such as land or housing, leaving the

present value of taxes the same, will have little effect on their market value, thus on private spending. An explicitly temporary income tax increase may have little effect on spending while the anticipation of prolonged deficits may lead taxpayers to ignore the eventual increase in tax liabilities. Evidence from specific episodes, such as the 1968 temporary tax surcharge in the United States, suggests partial offset at best.

Changes in the pattern of government spending obviously have real effects. But here again, various forms of direct crowding out may be at work. Public spending may substitute perfectly or imperfectly for private spending, so that changes in public spending may be directly offset, fully or partially, by consumers or firms. Even if public spending is on public goods, the effect will depend on whether the change in spending is thought to be permanent or transitory. Permanent changes, financed by a permanent increase in taxes, will, as a first approximation, lead to a proportional decrease in private spending, with no effect on total spending. Temporary changes in spending, associated with a temporary increase in taxes, lead to a smaller reduction in private spending and thus to an increase in total spending.

In summary, one should not expect any change in taxation or government spending to have a one-for-one effect on aggregate demand. An eclectic reading of the discussion above may be that only sustained decreases in income taxation, or the use of taxes that induce strong intertemporal substitution, or temporary increases in spending, can reliably be used to boost aggregate demand. The focus in what follows will be on these forms of fiscal expansion.

Crowding out at full employment

Not every increase in aggregate demand translates into an increase in output.

This is clearly the case if the economy is already at full employment (I use ‘full employment’ to mean employment when unemployment is equal to its natural rate). While tracing the effects of fiscal expansion at full employment is of limited empirical interest, except perhaps as a description of war efforts, it is useful for what follows. If labour supply is inelastic, output is fixed and any increase in aggregate demand must be offset by an increase in interest rates, leaving output unchanged. In the case of an increase in public spending, private spending will decrease; in the case of a decrease in income taxation, private spending will in the end be the same, but its composition will change as the share of interest sensitive components decreases. (If labour supply can vary, the story is more complicated. See, for example, Baxter and King, 1993, for an analysis of changes in government spending in an otherwise standard RBC model.)

This is just the beginning of the story, however. Over time, changes in capital and debt lead to further effects on output. The decrease in investment in response to higher interest rates leads to a decline in capital accumulation and output, reducing the supply of goods. If fiscal expansion is associated with sustained deficits, the increase in debt further increases private wealth and private spending at given interest rates, further increasing interest rates and accelerating the decline in capital accumulation (see, for example, Blanchard, 1985, for a characterization of these dynamic effects in an economy with finite horizon consumers). How strong is this negative effect of debt on capital accumulation likely to be? One of the crucial links in this mechanism is the effect of government debt on interest rates; empirical evidence, both across countries and from the last two centuries, shows surprisingly little relation between the two. This probably reflects, however, more the difficulty of identifying and controlling for other factors than the absence of an effect of debt and deficits on interest rates.

Worse can happen. It may be that the fiscal programme becomes unsustainable. There is no reason to worry about a fiscal programme in which debt grows temporarily faster than the interest rate. But there is reason to worry when there is a positive probability that, even under the most optimistic assumptions, debt will have to grow for ever faster than the interest rate. When this is the case, it implies that the government can meet its interest payments on existing debt only by borrowing more and more. What happens then may depend on the circumstances. Bond holders may start anticipating repudiation of government debt and require a risk premium on the debt, further accelerating deficits and the growth of the debt. If they instead anticipate repudiation through inflation, they will require a higher nominal rate and compensation for inflation risk in the form of a premium on all nominal debt, private and public. What is sure is that there will be increased uncertainty in financial markets and that this will further contribute to decreases in output and in welfare. The historical record suggests that it takes very large deficits and debt levels before the market perceives them as potentially unsustainable. England was able in the 19th century to build debt-to-GDP ratios close to 200 per cent without apparent trouble. Some European countries are currently running high deficits while already having debt-to-GDP ratios in excess of 100 per cent, without any evidence of a risk premium on government debt. The threshold seems lower for Latin American economies. But even if one excludes this worst-case scenario, fiscal expansion can clearly have adverse effects on output at full employment. The relevant issue, however, is whether the same dangers are present when fiscal expansion is implemented to reduce unemployment, which is presumably when it is most likely to be used.

Crowding out at less than full employment

The historical starting point of the crowding out discussion is the fixed price IS–LM model. In that model, a fiscal expansion raises aggregate demand and output. The pressure on interest rates does not come from the full employment constraint as before but from the increased demand for money from increased output. Thus the fiscal multiplier is smaller the lower the elasticity of money demand to interest rates, or the larger the elasticity of private spending to interest rates. Fiscal expansion crowds out the interest-sensitive components of private spending, but the multiplier effect on output is positive. As output and interest rates increase, it is quite possible for both investment and consumption to increase. But what happens when the model is extended to take into account dynamics, expectations and so on? Can one overturn the initial result and get full crowding out or even negative multipliers?

Even within the static IS–LM, one can in fact get zero or negative multipliers. This is the case, for example, if money demand from agents is higher than that from the government and the change in policy redistributes income from the government to agents. While this case is rather exotic, a much stronger case can be made if the economy is small, open, and with capital mobility and flexible exchange rates, as in the ‘Mundell–Fleming’ model. In this case, with the interest rate given from outside, and fixed money supply, money demand determines output; fiscal policy leads only to exchange rate appreciation. Exchange rate-sensitive components are now crowded out by fiscal expansion. The multiplier is equal to zero.

When dynamic effects are taken into account, other channels arise for crowding out. The analysis of these dynamic effects, with the dynamics of debt accumulation taken into account, was initially conducted under the maintained assumption of fixed prices and demand determination of output (Tobin and Buiter, 1976). Then, as debt was accumulating, private wealth and spending increased, leading to

even larger effects of fiscal policy on output in the long run than in the short run. But the assumption of fixed prices, while debt and capital accumulation are allowed to proceed, is surely misleading; when prices are also allowed to adjust, the effects of fiscal policy become more complex, and crowding out more likely. This is because some of the full employment effects come back into prominence: if fiscal expansion is maintained even after the economy has reached full employment, then the perverse effects of higher interest rates on capital accumulation and full employment output come again into play. This is true even if deficits disappear before the economy returns to full employment; the economy inherits a larger level of debt, and thus must have higher interest rates and lower capital accumulation than it would otherwise have had. The fiscal expansion trades off a faster return to full employment for lower full-employment output.

Anticipations of these full employment effects are likely to feed back and modify the effects of fiscal policy at the start, when the economy is still at less than full employment. Anticipations of higher interest rates, perhaps also of higher distortions due to the higher taxes needed to service the debt, may dominate the direct effects of higher government spending on demand, and lead to an initial decrease rather than an initial increase in demand and output. Symmetrically, fiscal consolidation, to the extent that it implies lower interest rates and lower distortions in the future, may be expansionary. This is even more likely to be the case if fiscal consolidation decreases the risk of default on government debt, and thus decreases the risk of major economic disruptions. There is indeed some evidence that, when initial fiscal conditions are very bad, and the fiscal consolidation is large and credible, the net effect of consolidation may be expansionary (Giavazzi and Pagano, 1990).

Crowding out: an assessment

Should one conclude from this that fiscal policy is an unreliable macroeconomic tool, with small and sometimes negative effects on output? The answer is ‘no’. Fiscal policy is likely to partly crowd out some components of private spending, even in the best circumstances, but there is little reason to doubt that it can help the economy return to full employment. Ricardian equivalence and direct crowding out warn us that not any tax cut or spending increase will increase aggregate demand. But there is little question that temporary spending or sustained income tax cuts will do so. Results of full crowding out at less than full employment, such as the Mundell–Fleming result, are simply a reminder that the monetary-fiscal policy mix is important.

In all cases, monetary accommodation of the increased demand for money removes the negative or the zero multipliers. That fiscal expansion affects capital accumulation, and output adversely at full employment, and that unsustainable fiscal programmes may lead to crises of confidence, is a reminder that fiscal expansion should not be synonymous with steady increases in the debt-to-GDP ratio even after the economy has returned to full employment. This shows one of the difficulties associated with fiscal expansion: if done through tax cuts, it has to be expected to last long enough to affect private spending, but not so long as to lead to expectations of runaway deficits in the long run. The room for manoeuvre is, however, substantial. Some taxes, such as the investment tax credit, work best when temporary. These can be used, as they work in the short run and have few adverse implications for the long run.

See Also

- budget deficits
- real business cycles
- Ricardian equivalence theorem
- Tobin, James

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1095–117.

Baxter, M. and King, R. 1993. Fiscal policy in general equilibrium. *American Economic Review* 83, 315–34.

Blanchard, O. 1985. Debt, deficits, and finite horizons. *Journal of Political Economy* 93, 223–47.

Giavazzi, F. and Pagano, M. 1990. Can severe fiscal contractions be expansionary? *NBER Macroeconomics Annual* 5, 75–122.

Tobin, J. and Buiter, W. 1976. Long-run effects of fiscal and monetary policy on aggregate demand. In *Monetarism*, ed. J. Stein. Amsterdam: North-Holland.

How to cite this article

Blanchard, Olivier J. "crowding out." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000452> doi:10.1057/9780230226203.0344

The New Palgrave Dictionary of Economics Online

cultural transmission

Alberto Bisin and Thierry Verdier

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economic literature analyses cultural transmission as the result of interactions between purposeful socialization decisions inside the family (‘direct vertical socialization’) and indirect socialization processes like social imitation and learning (‘oblique and horizontal socialization’). This article reviews the main contribution of these models from theoretical and empirical perspectives. It presents the implications regarding the long-run population dynamics of cultural traits, and discusses the links with other approaches to cultural evolution in the social sciences as well as in evolutionary biology. Applications to economic problems are also briefly surveyed.

Keywords

altruism; cooperation; cultural transmission; evolutionary biology; evolutionary economics; genetic evolution; identity; imperfect empathy; inter-generational altruism; nature–nurture debate; religion, economics of; social interaction; social norms; socialization

Article

Preferences, beliefs, and norms that govern human behaviour are partly formed as the result of genetic evolution, and partly transmitted through generations and acquired by learning and other forms of social interaction. The transmission of preferences, beliefs and norms of behaviour which is the result of social interactions across and within generations is called *cultural transmission*. Cultural transmission is therefore distinct from, but interacts with, genetic evolution.
Cultural transmission is an object of study of several social sciences, such as evolutionary anthropology, sociology, social psychology and economics, as well as of evolutionary biology. The theoretical contributions of Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985), who apply models of evolutionary biology to the transmission of cultural traits, as well as the empirical study of cultural socialization in American schools by Coleman (1988), had a great multidisciplinary impact. Recently,

economists have also studied the determination and the dynamics of preferences, beliefs, norms and, more generally, cultural and cognitive attitudes.

Cultural transmission arguably plays an important role in the determination of many fundamental preference traits, like discounting, risk aversion and altruism. It plays a central role in the formation of cultural traits and norms, like attitudes towards the family and fertility practices, and in the job market. It is, however, the pervasive evidence of the resilience of ethnic and religious traits across generations that motivates a large fraction of the theoretical and empirical literature on cultural transmission. For instance, the fast assimilation of immigrants into a ‘melting pot’, which many social scientists predicted until the 1960s (see, for example, Gleason, 1980, for a survey), simply did not materialize. Moreover, the persistence of ‘ethnic capital’ in second- and third-generation immigrants has been documented by Borjas (1992), and recently also by Fernandez and Fogli (2005) and Giuliano (2007) for norms of behaviour regarding, respectively, work and fertility practices and living arrangements. Orthodox Jewish communities in the United States constitute another example of the strong resilience of culture (see Mayer, 1979, and the discussion of a ‘cultural renaissance’ rather than the complete assimilation of Jewish communities in New York in the 1970s). Outside the United States, Basques, Catalans, Corsicans, and Irish Catholics in Europe, Quebecois in Canada, and Jews of the diaspora have all remained strongly attached to their languages and cultural traits even through the formation of political states which did not recognize their ethnic and religious diversity.

Models of cultural transmission have implications regarding the determinants of the persistence of cultural traits and more generally regarding the population dynamics of cultural traits. In the economic literature in particular, cultural transmission is modelled as the result of purposeful socialization decisions inside the family (‘direct vertical socialization’) as well as of indirect socialization processes like social imitation and learning (‘oblique and horizontal socialization’). Therefore, the persistence of cultural traits or, conversely, the cultural assimilation of minorities is determined by the costs and benefits of various family decisions pertaining to the socialization of children in specific socio-economic environments, which in turn determine the children's opportunities for social imitation and learning.

Evolutionary biology models

L. Cavalli-Sforza and M. Feldman are the first to formally study the transmission of cultural traits. Their formal models are adopted from evolutionary biology. In a baseline version of these models, they obtain a simple differential equation which describes the population dynamics of cultural traits. Consider the dynamics of a dichotomous cultural trait in the population; formally, a fraction q^i of the population has trait *i*, and a fraction $q^j = 1 - q^i$ has trait *j*. Families are composed of one parent and a child, and hence reproduction is asexual. All children are born without defined preferences or cultural traits, and are each first exposed to their parent's trait, which they adopt with probability d^i. If a child from a family with trait *i* is not directly socialized, which occurs with probability $1 - d^i$, he or she picks the trait of a role model chosen randomly in the population (that is, he or she picks trait *i* with probability q^i and trait *j* with probability $1 - q^i$). Therefore, the probability that the child of parents of trait *i* will also have trait *i* is $\Pi^{ii} = d^i + (1 - d^i)q^i$; while the probability that he or she will have trait *j* is $\Pi^{ij} = (1 - d^i)(1 - q^i)$. It follows that the dynamics of the fraction of the population with trait *i*, in the continuous time limit, are

characterized by:

$$\dot{q}^i = (d^i - d^j)q^i(1 - q^i) \quad (1)$$

The dynamics that equation (1) describes implies that the distribution of cultural traits in the population converges to a degenerate distribution concentrated on trait *i* whenever $d^i > d^j$ (and on trait *j* when $d^i < d^j$), while any initial distribution is stationary in the knife-edge case in which $d^i = d^j$. This model therefore predicts the complete assimilation of the trait with weaker direct vertical socialization. Moreover, it predicts faster assimilation for smaller minorities. Both predictions are at odds with the documented strong resilience of cultural traits discussed above. Cavalli-Sforza and Feldman show how these extreme predictions can be relaxed by considering other effects like mutations, migrations and horizontal cultural transmission among peers. Boyd and Richerson (1985) in turn extend the analysis of Cavalli-Sforza and Feldman (1981) by considering forms of direct vertical socialization called *frequency dependent* biased transmission, which depend on the distribution of the population by cultural trait. Formally, they allow d^i to be a function of q^i.

Bisin and Verdier (2001a) study the same differential equation for the population dynamics of cultural traits, with the objective of characterizing the conditions which give rise to culturally heterogeneous stationary distributions, that is, limit population with a positive fraction of either cultural trait, $0 < q^i < 1$. They show that the crucial determinant of the composition of the stationary distribution consists in whether the socio-economic environment (oblique socialization) acts as a substitute or as a complement to direct vertical socialization. More precisely, when direct vertical socialization and oblique transmission are *cultural substitutes*, parents by definition socialize their children less the more widely dominant are their cultural traits in the population. In such a case, $d^i(q^i)$ is a strictly decreasing function in q^i, and in the long run a non-degenerate stable stationary distribution exists. It is characterized by a q^i such that the direct vertical socialization of the two cultural types are equalized (that is, $d^i(q^i) = d^j(1 - q^i)$). Intuitively, when family and society are substitutes in the transmission mechanism, in fact families socialize children more intensely whenever the set of cultural traits they wish to transmit is common only to a minority of the population. Conversely, families which belong to a cultural majority spend fewer resources directly socializing their children, since their children adopt or imitate with high probability the predominant cultural trait in society at large, which is the one their parents desire for them. *Cultural substitutability* tends to preserve cultural heterogeneity in the population because in this case minorities directly socialize their children more than majorities. The other typical situation is the opposite one in which direct vertical transmission is a *cultural complement* to oblique transmission; that is, when parents socialize their children more intensely the more widely dominant their cultural trait is in the population. In such a case, $d^i(q^i)$ is a strictly increasing function in q^i and in the long run the dynamics converges to a culturally homogeneous cultural population (with

either $q^i = 0$ or $q^i = 1$ depending on the initial distribution).

Economic models of cultural transmission

Economic models of cultural transmission induce testable restrictions on the form of the function $d^i(q^i)$. In their baseline specification, for instance, Bisin and Verdier (2001a) assume that parents are altruistic towards their children and hence might want to socialize them to a specific cultural model if they think this will increase their children's welfare. If we let V^{ij} denote the utility to a type *i* parent of a type *j* child, $i, j \in \{a, b\}$, the formal assumption is

$$\text{for all } i, j \text{ with } i \neq j, \; V^{ii} > V^{ij}$$

This assumption, called *imperfect empathy*, can be interpreted as a form of myopic or paternalistic altruism. Parents are aware of the different traits children can adopt and are able to anticipate the socio-economic choices a child with trait *i* will make in his or her lifetime. However, parents can evaluate these choices only through the filter of their own subjective evaluations and cannot ‘perfectly empathize’ with their children. As a consequence of imperfect empathy, parents, while altruistic, tend to prefer children with their own cultural trait and hence attempt to socialize them to this trait. (Some justifications of imperfect empathy from an evolutionary perspective are provided by Bisin and Verdier, 2001b. The assumption can be relaxed, as for example in Sáez-Martí and Sjogren, 2005.) Assume socialization is costly and let costs be denoted by $C(d^i)$. Parents of type *i* then choose d^i to maximize:

$$-C(d^i) + (\Pi^{ii}V^{ii} + \Pi^{ij}V^{ij}) \tag{2}$$

$$s.t. \; \Pi^{ii} = d^i + (1 - d^i)q^i, \Pi^{ij} = (1 - d^i)(1 - q^i) \tag{3}$$

Under standard assumptions, the solution to this problem provides a continuous map $d^i = d(q^i, \Delta V^i)$, where $\Delta V^i = V^{ii} - V^{ij}$ is the subjective utility gain of having a child with trait *i*. It reflects the degree of

‘cultural intolerance’ of type *i*'s parents with respect to cultural deviations from their own trait. Given imperfect empathy on the part of parents, $\Delta V^i > 0$. The dynamics of the fraction of the population with cultural trait *i* is then determined by equation (1) evaluated at $d^i(q^i) = d(q^i, \Delta V^i)$. It is straightforward to demonstrate that this class of socialization mechanisms generates cultural substitutability and therefore the preservation of cultural heterogeneity. Other micro-founded specifications and examples are provided in Bisin and Verdier (2001a), some of which illustrate the contrary possibility of cultural complementarity and the tendency of cultural homogenization over time.

Direct socialization mechanisms and socio-economic interactions

Several specific choices contribute to direct family socialization and hence to cultural transmission. Prominent examples are education decision, family location decisions, and marriage choices While education choices have been studied by Cohen-Zada (2004), and marriage choices by Bisin and Verdier (2000), the literature has to date shown little interest in the socialization effects of location choices, for instance, the socialization effects of urban agglomeration by ethnic or religious trait.
The simple analysis of the economic model of cultural transmission of Bisin and Verdier depends crucially on the assumption that the utility to a type *i* parent of a type *j* child, V^{ij} is independent of the distribution of the population by cultural trait, that is, independent of q^i. Many interesting analyses of cultural transmission require this assumption to be relaxed. In many instances the adoption of the cultural trait of the majority in fact favours children, for example in the labour market; a typical example is language adoption. In this case altruistic parents, even if paternalistic, might favour (or discourage less intensely) the cultural assimilation of their children. If we allow for interesting socio-economic effects interacting with the socialization choices of parents, the basic cultural transmission model of Bisin and Verdier has been applied to several different environments and cultural traits and social norms of behaviour, from preferences for social status (Bisin and Verdier, 1998) to corruption (Hauk and Sáez-Martí, 2002), hold-up problems (Olcina and Penarrubia, 2004), development and social capital (François, 2002), inter-generational altruism (Jellal and Wolff, 2002), labour market discrimination (Sáez-Martí and Zenou, 2005), globalization and cultural identities (Olivier, Thoenig and Verdier, 2005), and work ethics (Bisin and Verdier, 2005).

Empirical analysis of cultural transmission models

While an interesting literature has documented the relevance of cultural factors in several socio-economic choices, much less is known about cultural transmission per se. Nonetheless, several important questions are beginning to be answered. First of all, several important correlations have been documented in sociology, in particular with regard to the role of marriage in socialization (see, for instance, Hayes and Pittelkow, 1993; Ozorak, 1989; Heaton, 1986). The literature in economics has instead concentrated more specifically on the direct empirical validation of the economic approach to cultural transmission surveyed above, thereby estimating the relative importance of direct and oblique socialization for different specific traits and the prevalence of cultural substitution or complementarity in specific socio-economic environments. Patacchini and Zenou (2004) find evidence of cultural

complementarity in education in the United Kingdom. Cohen-Zada (2004) finds instead for the United States that the demand for private religious schooling decreases with the share of the religious minority in the population, in accord with cultural substitution. Fernandez, Fogli and Olivetti (2004) find evidence of an important role for mothers in the transmission to their sons of attitudes favouring the participation of women in the labour force and acquisition of higher education. Finally, Bisin, Topa and Verdier (2004a), using the General Social Survey data for the United States over the period 1972–96, estimate for religious traits the structural parameters of the model of marriage and child socialization in Bisin and Verdier (2000). They find that observed intermarriage and socialization rates are consistent with Protestants, Catholics and Jews having a strong preference for children who identify with their own religious beliefs, and taking costly decisions to influence their children's religious beliefs. The estimated ‘relative intolerance’ parameters are high and asymmetric across religious traits, suggesting an interestingly rich representation of ‘cultural distance’.

Genetic and cultural evolution

Cultural transmission possibly has a role also in the determination of fundamental preference parameters, such as time discounting, risk aversion, altruism, and interdependent preferences. Purely evolutionary models have been complemented by alternative models of cultural transmission and genetic and cultural co-evolution. The wealth of different approaches proposed is best exemplified by the study of preferences for cooperation. The observation that humans often adhere to collectively beneficial actions which are not in their private interest (or which are not rationalizable as strategic equilibria) has led to a theoretical literature explaining how psychological ‘preferences for cooperation’ can be sustained in the context of genetic and/or cultural evolution (this is called the *puzzle of pro-sociality* by Gintis, 2003a). For instance, in the context of the Prisoner's Dilemma, Becker and Madrigal (1995) exploit the ability of habits to induce preferences; Guttman (2003), Stark (1995), and Bisin, Topa and Verdier (2004b) show how cooperation can be sustained by different modes of cultural evolution; Gintis (2003b) shows that a general capacity to internalize fitness-enhancing norms of behaviour can be genetically adaptive, and hence that cooperation can also be internalized by ‘hitchhiking’ on this general capacity.

The empirical evidence on the nature–nurture debate (see Ceci and Williams, 1999, for a review) has not yet been systematically taken to the point of distinguishing the genetic from the cultural factors in the determination of fundamental preference parameters. Similarly, the empirical evidence distinguishing the different cultural transmission models of fundamental preference traits is almost non-existent. The only exception is by Jellal and Wolff (2002), who study the implication of the pattern of *inter vivos* transfers within the family in France for the transmission of inter-generational altruism. They argue that the evidence is more consistent with a cultural transmission model such as that of Bisin and Verdier (2001a) rather than with a ‘demonstration effect’ model, as in Stark (1995), where parents take care of their elders in order to elicit similar behaviour in their children.

See Also

- culture and economics
- identity
- social interactions (empirics)

Bibliography

Becker, G. and Madrigal, V. 1995. On cooperation and addiction. Mimeo, University of Chicago.

Bisin, A., Topa, G. and Verdier, T. 2004a. Religious intermarriage and socialization in the United States. *Journal of Political Economy* 112, 615–64.

Bisin, A., Topa, G. and Verdier, T. 2004b. Cooperation as a transmitted cultural trait. *Rationality and Society* 16, 477–507.

Bisin, A. and Verdier, T. 1998. On the cultural transmission of preferences for social status. *Journal of Public Economics* 70, 75–97.

Bisin, A. and Verdier, T. 2000. Beyond the melting pot: cultural transmission, marriage and the evolution of ethnic and religious traits. *Quarterly Journal of Economics* 115, 955–88.

Bisin, A. and Verdier, T. 2001a. The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory* 97, 298–319.

Bisin, A. and Verdier, T. 2001b. Agents with imperfect empathy might survive natural selection. *Economics Letters* 2, 277–85.

Bisin, A. and Verdier, T. 2005. Work ethic and redistribution: a cultural transmission model of the welfare state. Mimeo, New York University.

Borjas, G. 1992. Ethnic capital and intergenerational income mobility. *Quarterly Journal of Economics* 57, 123–50.

Boyd, R. and Richerson, P. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Cavalli-Sforza, L. and Feldman, M. 1981. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.

Ceci, S. and Williams, W. 1999. *The Nature–Nurture Debate: The Essential Readings*. Oxford: Blackwell.

Cohen-Zada, D. 2004. Preserving religious identity through education: economic analysis and evidence from the US. Mimeo, Ben-Gurion University.

Coleman, J. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 94, S95–S120.

Fernandez, R. and Fogli, A. 2005. Culture: an empirical investigation of beliefs, work, and fertility. Working paper No. 11268. Cambridge, MA: NBER.

Fernandez, R., Fogli, A. and Olivetti, C. 2004. Mothers and sons: preference formation and female labor force dynamics. *Quarterly Journal of Economics* 119, 1249–99.

François, P. 2002. *Social Capital and Economic Development*. New York: Routledge.

Gintis, H. 2003a. Solving the puzzle of prosociality. *Rationality and Society* 15, 155–87.

Gintis, H. 2003b. The hitchhikers guide to altruism: genes, culture and the internalization of norms. *Journal of Theoretical Biology* 220, 407–18.

Giuliano, P. 2006. Living arrangements in Western Europe: does cultural origin matter? *Journal of the European Economic Association* 5 927–52.

Gleason, P. 1980. American identity and Americanization. In *Harvard Encyclopedia of American Ethnic Groups*, ed. T. Stephan, O. Ann and H. Oscar. Cambridge, MA: Harvard University Press.

Guttman, J. 2003. Repeated interaction and the evolution of preferences for reciprocity. *Economic Journal* 113, 631–56.

Hauk, E. and Sáez-Martí, M. 2002. On the cultural transmission of corruption. *Journal of Economic Theory* 107, 311–35.

Hayes, B. and Pittelkow, Y. 1993. Religious belief, transmission, and the family: an Australian study. *Journal of Marriage and the Family* 55, 755–66.

Heaton, T. 1986. How does religion influence fertility? The case of Mormons. *Journal for the Scientific Study of Religion* 28, 283–99.

Jellal, M. and Wolff, F. 2002. Cultural evolutionary altruism: theory and evidence. *European Journal of Political Economy* 18, 241–62.

Mayer, E. 1979. *From Suburb to Shetl: The Jews of Boro Park*. Philadelphia: Temple University Press.

Olcina, G. and Penarrubia, C. 2004. Hold-up and intergenerational transmission of preferences. *Journal of Economic Behavior and Organization* 54, 111–32.

Olivier, J., Thoenig, M. and Verdier, T. 2005. Globalization and the dynamics of cultural identity. Mimeo. Paris: Paris-Jourdan Sciences Économiques.

Ozorak, E. 1989. Social and cognitive influences on the development of religious beliefs and commitment in adolescence. *Journal for the Scientific Study of Religion* 28, 448–63.

Patacchini, E. and Zenou, Y. 2004. Intergenerational education transmission: neighborhood quality and/or parents' involvement? Working Paper No. 631. Stockholm: Research Institute of Industrial Economics.

Sáez-Martí, M. and Sjogren, A. 2005. Peers and culture. Working Paper No. 642. Stockholm: Research Institute of Industrial Economics.

Sáez-Martí, M. and Zenou, Y. 2005. Cultural transmission and discrimination. Mimeo. Stockholm: Research Institute of Industrial Economics.

Stark, O. 1995. *Altruism and Beyond: An Economic Analysis of Transfers and Exchanges within Families and Groups*. Cambridge: Cambridge University Press.

How to cite this article

Bisin, Alberto and Thierry Verdier. "cultural transmission." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000549>
doi:10.1057/9780230226203.0345

The New Palgrave Dictionary of Economics Online

culture and economics

Raquel Fernández
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Modern neoclassical economics has, until recently, ignored the potential role of culture in explaining variation in economic outcomes, largely because of the difficulty in rigorously separating the effects of culture from those of institutions and traditional economic variables. This article selectively reviews some recent attempts to empirically identify the effects of culture on economic outcomes and to answer the question, ‘does culture matter and, if so, how much?’ Open theoretical and empirical questions are discussed, including the relationship between culture and institutions.

Keywords

agency problems; assortative matching; cultural assimilation; culture and economics; endogenous preferences; epidemiology; fertility; human capital; institutions; instrumental variables; international migration; multiple equilibria; occupational selection; religion; social norms; technology; trust; women's work and wages; beliefs; identity; preference transmission

Article

Economic decisions are made within a social context; as Aristotle reminds us, man is a social animal. The relevance of this statement to economics, however, is far from clear. In what ways, if any, do we need to consider the social nature of man in order to study economic questions? This article attempts to provide a partial answer to this question.

Traditionally, economists seek to explain differences in economic outcomes by studying how agents, with given preferences and beliefs, react to changes in the policy environment, institutions and technology. At a deeper level than the taste for apples versus oranges, however, few would deny that preferences and beliefs must be, to some extent, endogenous. Our level of trust in others, the determinants of status in society, our beliefs about the correct trade-off between efficiency and equity, or the ‘proper’ roles for men and women, are all examples of beliefs or preferences that have differed

across societies and over time. These beliefs and preferences impact on individual behaviour and how society allocates scarce resources. At the individual level they help determine whether a woman participates in the formal labour market and the career she follows, the extent to which racism is tolerated, or the degree of assortative matching on wealth in marriages. At a collective level, they help determine, for example, the range and depth of the welfare state, the legality of slavery, or the proportion of the budget that is dedicated to foreign aid.

Although at some general level few may disagree that preferences, beliefs, or values of the type discussed above are endogenous (and may therefore differ across societies), whether they have a quantitatively significant impact on economic outcomes is another matter. Do differences in beliefs and preferences that vary systematically across groups of individuals separated by space (either geographic or social) or time – what I shall henceforth term *culture* – play an important role in explaining differences in outcomes? (For the purposes of this article, I will not give a more rigorous definition of culture than the abbreviated one here. See Elster, 1989, for a discussion of social norms and culture and Manski, 2000, for a discussion of peer effects and social interactions.) Modern economics (as opposed to sociology or anthropology) has largely been, until recently, reluctant to investigate this question. Although in principle there is nothing non-standard about positing preference/belief heterogeneity among individuals to explain differences in outcomes, the Stigler–Becker dictum *de gustibus non est disputandum* (Stigler and Becker, 1977) and its assertion that ‘no scientific behavior has been illuminated by assumptions of differences in taste’ has cast a long shadow in economics. Thus, the main challenge faced by those who believe that culture might matter has been to find a convincing way to show that culture can be studied rigorously and, in particular, that it is possible to separate the influence of culture from institutions and standard economic variables. In this sense, running, say, cross-country regressions on variables that one suspects reflect cultural attitudes (for example, different savings patterns may reflect attitudes towards thrift) to study the effect of culture has long (and correctly) been considered unsatisfactory. Despite one's best efforts to control for differences in countries’ economic environments, identifying the residual with culture is ultimately unconvincing. It is difficult, if not impossible, to summarize the economic environment faced by agents with a few aggregate variables. Thus, there are bound to be omitted variables and problems of endogeneity, which are all further confounded by mismeasurement.

Hence, despite a long history of writers on the relationship between culture and economics (which includes Marx, Weber, Gramsci, Polanyi, Banfield and, more recently, Putnam and Landes, among others), modern neoclassical economics has been by and large silent on the topic of culture and only in recent years have economists started to think seriously again about how culture may help explain economic phenomena. In this article I will selectively review some recent attempts to empirically identify the effects of culture on important economic outcomes and to answer the question, ‘does culture matter?’ Answering this question affirmatively naturally leads one to explore the propagation mechanisms of culture, to theorize about the relationship between institutions and culture, and to investigate the dynamic of culture – all topics that I will briefly touch upon at the end.

Empirical evidence on culture

In this section I examine some of the recent evidence on the importance of culture for economic

outcomes. For expository ease, I have divided the empirical evidence into that which uses survey data, evidence based on immigrants or their descendants (what I call the ‘epidemiological approach’), and historical case studies. There is also a small body of experimental work that, by showing that across societies there exist marked differences in how individuals play games such as the ultimatum, public good or dictator game, has also shed light on the relationship between culture and economics (see, for example, Henrich et al., 2001).

Survey-based evidence

Perhaps the most natural approach to doing empirical work on culture consists in using the beliefs expressed by individuals in surveys (for instance, the World Value Surveys) on a variety of issues as expressions of culture and correlating them with economic outcomes. This approach, however, must overcome the problem of reverse causality. That is, differences in beliefs may be solely a consequence of different economic and institutional environments. Hence, the use of instrumental variables is required in order to identify causality. Overall, this has been difficult to achieve.

As shown by Guiso, Sapienza and Zingales (2003), the intensity of religious beliefs and religious denomination are correlated with a variety of individual attitudes such as trust in others, government's role, views of working women and the importance of thrift. Guiso, Sapienza and Zingales (2006) show that these attitudes, aggregated at the country level, are correlated with cross-country aggregate outcomes (for example, savings, redistributive versus regressive taxation, and trade). In order to ensure that the reverse causality is not at play, the attitudes are instrumented, usually by the religious composition in the country. This work is suggestive but there are several concerns associated with it. In addition to questions about omitted variables, it is not clear that religious composition is a valid instrument since it may also help explain the aggregate outcome through other channels. (Indeed, the coefficients on the instrumental variable results tend to look very high relative to the ones obtained by ordinary least squares. Running regressions at the individual outcome level would be more convincing, but opinion surveys unfortunately tend not to have high-quality economic data (the World Value Survey, for example, classifies income levels into ten categories). Recent work by Guiso, Sapienza and Zingales (2005) on the relationship between trust and trade, instead instruments trust with the genetic distance between indigenous populations. This seems a promising avenue of research.

Tabellini (2005) takes a significant step towards overcoming some of the weaknesses discussed above. To study whether culture affects economic development across European regions, he also aggregates (at the regional level) individual responses from the World Value Surveys to questions about trust, respect and the link between individual effort and economic success. The scope for omitted variables is reduced by focusing on within-country variation in Europe (by including country fixed effects). The attitudes are then instrumented with historical variables, such as regional literacy rates at the end of the 19th century and indicators of political institutions in the period from 1600 to 1850. The author finds that the proxies for culture are quantitatively significant determinants of per capita GDP levels and growth rates across regions. It is possible of course that the instruments are not valid. For example, they could affect output directly via sectoral composition or public investment. The paper contains a good discussion of these and other alternative hypotheses.

The epidemiological approach

A very different approach to relying on opinion data is to examine the economic outcomes of immigrants or their descendants. This is reminiscent of the epidemiology literature that, in order to attempt to identify the contribution of the environment broadly defined (namely, physical and cultural) relative to genes in disease, studies various health outcomes for immigrants and compares them to outcomes for natives (see, for example, the classic study by Marmot et al., 1975).

To understand the strengths and weaknesses of such an approach, suppose that the level of, say, heart disease differs markedly between two countries (the source and host countries). If heart disease in immigrants converges to that of natives in the host country, the difference between the two countries is unlikely to be driven by genetics and instead results from the environment. Failure to find convergence, on the other hand, does not imply the opposite. There are many reasons why the environment may be solely responsible and still sustain differential levels of heart disease. For example, cultural assimilation may occur slowly (for instance, if immigrants maintain the same dietary patterns as in the source country), or living in the source country at a young age may confer some degree of immunity, or selection into immigration may be correlated with a particular health outcome.

The epidemiological strategy in economics has its own set of problems. In particular, it is important to recognize that immigrants may be subject to many shocks (language difficulties, worse employment opportunities, greater uncertainty and so forth) which cause them to deviate from their traditional behaviour. Culture, furthermore, is socially constructed: to be replicated, the behaviour may require the incentives—rewards and punishments – provided by a larger social body such as a neighbourhood, school, or ethnic network. Furthermore, immigrants are unlikely to be a representative sample of their home-country's population. Their beliefs, preferences, and unobserved differences in their economic circumstances may differ significantly from the country average. Lastly, the exposure of immigrants (or their descendants) to a different culture from the one prevalent in their country of heritage presumably weakens the latter's impact on their behaviour. Note that all the factors mentioned above introduce a bias towards finding culture to be insignificant. Thus, on the whole, comparisons of behaviour or outcomes across different immigrant groups are a very demanding test of the importance of culture. In epidemiology, when differences across groups remain, one must be careful not to conclude that genetics is determinative when the underlying cause may be cultural; in economics, when significant differences are not observed, one must be careful not to rule out cultural forces.

In economics, the paper by Carroll, Rhee and Rhee (1994) is the first that, to my knowledge, follows an approach similar to the one described above. The authors are interested in exploring whether cross-country differences in savings rates may be culturally driven. Using individual-level data on immigrants to Canada, they estimate individual consumption levels as a function of permanent income (as captured by labour and asset income), the interaction of this variable with demographic variables, some measures of wealth, and finally the interaction of a region of origin dummy (and years since arrival to Canada) with their measure of permanent income. If there exist different cultural attitudes towards savings, and if this attitude is maintained in immigrants, then one should observe different propensities across immigrants, by region of origin, to consume out of permanent income (that is, the regional dummies should be significantly different from one another). The authors find that the saving patterns of immigrants do not vary significantly by region of origin. Recent immigrants as a whole save less than native-born Canadians, but there is no statistically significant difference in behaviour across immigrant

groups.

There are several weaknesses in the data-set used in the study above that may bias it against finding results that show a significant impact of culture. Wealth, for example, is not well measured. In particular, as only South East Asia's saving rate differed markedly from those of other regions in the immigrant population (31 per cent relative to 18–20 per cent across the remaining regions), the small number of immigrants from this group in the sample limits the power of the test. Note also that, if the motivation to save more stems from the desire to provide one's child with greater status via a larger bequest, the incentive to do this may be much less marked in a society in which savings are generally low or in which status stems from consumption behaviour.

Fernández and Fogli (2005; 2006) use a similar, but arguably less problematic, methodology by studying second-generation Americans in order to investigate the quantitative importance of culture. Their research focuses on the fertility and work behaviour of married second-generation American women (that is, women who were born in the United States but whose parents were born elsewhere). The use of second-generation immigrants attenuates the problems associated with the first generation's adjustment to a foreign setting (for example, language difficulties) and even some selection problems are less likely to play a role for the second generation. On the other hand, second-generation individuals have been more exposed to the new culture, and that will tend to diminish the role of culture from the country of heritage. Our hypothesis is that attitudes towards woman's ‘proper’ role in society and towards ideal family size are culturally different across countries and that this culture is likely to be transmitted intergenerationally and show up in systematic differences in female labour force participation (LFP) and fertility, even if individuals were raised in the United States.

In our 2005 paper, the challenge was how to best capture the attitudes towards women and family size in the parents’ country of origin. We chose not to use country dummies (as in Carroll, Rhee and Rhee, 1994) but to instead examine whether past values of economic variables in the country of origin that should reflect this culture – in particular, past values of female LFP and total fertility rates (TFR) – are able to play a quantitatively significant role in explaining differences in outcomes across second-generation women in the United States. Our argument is that these economic variables reflect the institutions (for example, markets, legal framework, minimum wages and so on), the strictly economic environment (demand and supply, transportation costs, access to day care, for example), as well as the preferences and beliefs (that is, the culture) of individuals in the country making decisions at that time. If these variables are able to explain the behaviour of women who, by virtue of living in the USA and in a different time period, face different institutions and economic variables, then solely the cultural component of these variables should affect their choices. This is a more demanding test that is superior to the ‘black box’ approach of using country dummies which leaves open the question of what it is about the country that matters to outcomes.

In individual level regressions, we find that our cultural proxies – past values of female LFP and TFR – help explain both how much second-generation American women work and their fertility. As our data-set – the 1970 US Census – does not allow us to control for family factors such as parental wealth, income, and education, we include the woman's education, her spouse's education, and total personal income (as well as location, age, and so on) in our regressions. By including these variables, the coefficient on the cultural proxy only captures the direct effect of culture rather than its full direct and indirect effects (for example, a woman who wants engage in market work is more likely to invest in education and hence, by controlling for education, we are eliminating the effect of culture on this

variable), but this is preferable to not controlling for differences in parental background, other than culture, that may affect women's work and fertility outcomes. We find that the cultural proxies still matter even after including these additional variables. Furthermore, the cultural proxies are quantitatively significant: a one standard-deviation increase in the corresponding cultural proxy is associated with approximately an eight per cent increase in hours worked per week and about a 14 per cent increase in the number of children. The forces of assimilation means that these numbers should be taken, if anything, as a downward biased estimate of the true power of culture in the original setting (that is, in the country of ancestry).

We also examine the most compelling alternative economic explanation for our results, namely, the hypothesis that these are driven by unobserved human capital. We do this by showing that the results are robust to the inclusion of the country of ancestry's level of per capita GDP in various years and to the years of education of immigrants (by country of ancestry) in 1940 (this remains the case when Hanushek and Kimko's (2000) measures of education quality in the parents' country of origin are included). We also demonstrate that the work cultural proxy does not have explanatory power in a Mincer wage regression which it would be expected to have if it captured unobserved human capital. Lastly, we show that the work cultural proxy is insignificant in explaining how much married second-generation American men work whereas the fertility cultural proxy retains its explanatory power. (If the work cultural proxy had a negative effect on how much these men work, that might indicate a substitution effect. In our regressions, the coefficient is basically zero and insignificant.) This is important because it implies that there does not exist some omitted economic variable at the parental country-of-origin level that affects the productivity of both men and women and that helps explain how much they work.

The methods used in Fernández and Fogli (2005) could be profitably extended to examine other issues, such as entrepreneurship or savings behaviour. It might also be interesting to elaborate upon the recent approach by Algan and Cahuc (2006) that attempts to combine survey evidence with the epidemiological approach in order to study the effects of culture on cross-country labour market outcomes. Although this work is too preliminary to discuss in depth, using the attitudes of, say, second-generation Americans to instrument for the attitudes of individuals in the home country seems cleaner than relying on variation in religious denominations. As usual, the question will be whether there is some omitted background economic variable correlated with the country of origin (particularly given the quality of the survey data-sets) that could be driving the results, but it seems a promising avenue of research (see also the interesting work on culture and migrants within regions in Italy by Ichino and Maggi, 2000. As shown recently in Fernández (2007a) using the World Value Survey, the attitudes of individuals in the country of ancestry towards women's market work and housework have explanatory power for the work outcomes of second-generation American women in 1970.

Historical case studies

The analysis of historical episodes in which changes in either culture or environment yield 'natural experiments' is likely to add richness and depth to our understanding of culture and the economy. Greif's 1994 paper is probably the best-known work in economics that makes the link between culture and institutional development. In brief, Greif argues that cultural beliefs (collectivist versus individualist) are reflected in the different ways in which in the 11th century Genoese traders and Maghrebi traders set up

their trading institutions. Both groups of merchants required agents to conduct their business overseas, and in both cases there was an agency problem as the overseas agent might be tempted to cheat the merchant. Maghrebi traders set up ‘horizontal’ relations in which merchants served as agents for traders and vice versa. Information was shared among merchants/traders and an agent who was dishonest with one merchant could expect to be shunned by other merchants. The Genoese, on the other hand, set up ‘vertical’ relationships in which individuals specialized as merchants or agents. Information was not shared among merchants. This led the Genoese to set up more formal enforcement institutions. The two different responses, argues Greif, then had important consequences once trading opportunities were expanded in previously inaccessible areas. The Maghrebi expanded trade using other Maghrebi agents whereas the Genoese were able to establish agency relations with non-Genoese, leading to very different economic development paths thereafter (see also Greif's, 2005, recent book on the topic).

Another compelling example is provided by Botticini and Eckstein (2005) who present the thesis that an ‘exogenous’ cultural change gave rise to the pattern of Jewish occupational selection that we see to this day. They argue that with the destruction of the Temple in Jerusalem in 70 ce, the Pharisees became the dominant religious group and transformed Judaism from a religion based on sacrifices to one whose main rule required each male to read and to teach his sons the Torah. This reform was implemented in places where most Jews were farmers who would not gain anything from investing in education. When urbanization expanded many centuries later, Jews had a comparative advantage in the skilled occupations demanded in the new urban centres. Thus, culture – the religious requirement of reading skills for other than human capital reasons – gave rise to the pattern of Jewish occupational selection seen since the ninth century.

Theories of culture

Is it necessary to modify the standard economic model in order to incorporate culture? The answer definitely is ‘no’. What appear to be societal differences in preferences may only be choice of equilibrium strategies in a game with multiple equilibria and standard preferences. This is in fact the most common way to think about the role of culture in economics, and is fully in keeping with our working definition of culture as systematic differences (across groups) in preferences or beliefs. Here the heterogeneity lies in the expectations (beliefs) over the strategies that will be played in equilibrium. Hence differences in culture can be identified with, for example, which equilibrium we play in a static game (for example, do we drive on the right- or left-hand side of the road) or the degree of cooperation (‘trust’) sustained in a repeated Prisoner's Dilemma game.

Within the ‘culture as multiple equilibria’ literature, I find particularly interesting the research that attempts to generate behaviour that looks like social norms (such as determinants of status). Take, for example, a dynamic matching model in which individuals who differ in wealth choose a partner with whom to match and obtain utility from joint consumption and the utility of their child. As shown in Mailath and Postlewaite (2003), in addition to an equilibrium in which there is assortative matching on wealth, there may also be an equilibrium with imperfectly assortative matching that depends also on non-economic characteristics such as whether one has blue eyes. In this equilibrium, blue eyes matter not because of their intrinsic value, but simply because the matching rule allocates, for the same wealth level, a wealthier partner to individuals with blue eyes. Thus, a woman would be willing to match with a man with blue eyes and slightly lower wealth than another man without blue eyes, because although she

obtains lower joint consumption, there is a 50 per cent chance that her child would inherit blue eyes and hence a better match and higher consumption in the future. To an outside observer, it might therefore appear that in this society people had an intrinsic preference for blue eyes, although this inference would be incorrect.

Although the example above is interesting, its explanation for a particular social norm seems incomplete and intuitively less than compelling. The preference for blue eyes or light skin may perhaps initially come about as a choice among many equilibria and involve solely a calculation about the trade-off between one's own consumption and that of one's child (though that too seems doubtful and is more likely the result of a history in which these traits are correlated with higher status). Over the longer run, however, one may conjecture that what sustains these equilibria – what makes these cultural traits less fragile to perturbations – is that these calculations are embodied in the individual and in society as preferences and beliefs about the inherent superiority/desirability of such features. People come to prefer blue eyes; people become racist. Thus, what is missing more generally in the theory of culture is an analysis of how preferences and beliefs (about things other than equilibrium strategies) themselves evolve.

The hypothesis that certain features of culture (those that have greater depth than driving on the left or the right side of the street) become part of preferences and beliefs implies that they cannot be discarded easily simply because they are no longer useful or beneficial, though over time this will certainly lessen their appeal. In this way, the operation of culture may be clearest to perceive when it no longer serves any useful societal purpose or particular group interest but nonetheless, at least for some time, persists – for example, religious prohibition on eating pork. (One reason speculated for this prohibition is that consumption of undercooked pork is linked to trichinosis. It is now known that this problem can be eliminated, however, by thoroughly cooking the meat.) In the context of the matching example above, individuals may eventually be willing to match with lower wealth people with blue eyes because this matching rule is incorporated into preferences/beliefs over what type of mate is intrinsically better even if the benefit derived by passing this trait on to their offspring is no longer substantial (say, because family size falls and decreases the payoff from the inheritable trait relative to the decrease in immediate joint consumption).

So far, we have discussed differences in culture as systematic differences in preferences and beliefs without distinguishing much between the two. This is not accidental, since, in general, the distinction between preferences and beliefs for our purposes is rather fuzzy. Even for simple preferences such as the trade-off between apples and oranges, what one knows (or believes) about the nutritional contents of the two may affect how one ‘feels’ about them, as may any other mental associations (for example, whether one is considered more exotic, how they were grown and so forth). In general, there are few pure (or naive) preferences – what one thinks or believes influences how one feels (and the same may be true vice versa. See Damasio, 1995, for an interesting exposition of evidence in favour of the hypothesis that emotions affect – and in fact are necessary for – the ability to think well). This is not to deny that people have some inherent tastes (for example, it is believed that human beings have a taste for fat, probably because of the evolutionary advantage associated with an inclination to eat meat in an environment in which protein and iron were not easily obtained).

For more complex questions the above is even more likely to be true. Consider, for example, the large increase in female labour force participation in the 20th century. Is it that woman's disutility from market work decreased or that her beliefs about the meaning or consequences of her working that changed over

time? The dichotomy between the two alternatives does not seem very useful in this case. If the focus is on understanding why actions change over time, then using standard preferences and modelling the evolution of beliefs as giving rise to changes in expected payoffs may be the more useful strategy (the latter is the approach taken by Fernández, 2007b, who shows that a model of the evolution of female LFP as an intergenerational learning process does a good job of replicating a century of US female LFP data). If instead one wished to understand the utility from a given action, particularly one in which identity is concerned, then incorporating cultural beliefs into preferences may be a better route (see, for instance, Akerlof and Kranton, 2000). For example, wearing a dress or having a woman as a boss may decrease a man's utility, independently of any expectations of future consequences, simply because it makes him feel (culturally) less masculine.

Culture and institutions

As seen previously, the main challenge faced by most empirical work on culture is to convincingly isolate its effects from the incentives provided by traditional economic variables and institutions. This should not be taken to mean that culture and institutions are independent variables. Indeed, one way to think about institutions is as congealed culture: that is, which institutions are set up and how these evolve depends not only on the problems faced by society (or by a particular group in society) at a particular moment in time but also the beliefs/preferences – the culture – that are prevalent. As elaborated on in our earlier discussion of Greif (1994), cultural beliefs (collectivist versus individualist), for example, were reflected in the different ways in which in the 11th century Genoese traders and Maghrebi traders set up their trading institutions, leading to very different economic development paths thereafter. My hypothesis is that the reverse causality is also likely to hold: that is, not only does culture affect institutions but also institutions affect the dynamic evolution of culture. In this sense, work that attempts to establish whether institutions or culture are the most important determinants of economic development seems misconceived (see Fernández, 2007c, for a theoretical analysis of the dynamic dependency of culture and institutions; also Bowles, 1998, for a review of some of the theoretical and empirical evidence on the effect of markets on culture).

Concluding remarks

The rigorous study of culture and economics is in its infancy. We would like to understand, for example, how culture propagates and evolves. In particular, what is the relative importance of family versus other institutions as cultural transmission mechanisms for different beliefs or in different environments? To what extent is cultural transmission purposeful, that is, optimizing on the part of an individual or her parents (as in Bisin and Verdier, 2000) or for a social group, and to what extent is it involuntary? (Fernández, Fogli and Olivetti, 2004, show that whether a man's mother worked while he was growing up is correlated with whether his wife works, even after controlling for a whole series of socioeconomic variables. They interpret this as preference transmission, but whether it is voluntary – optimizing – or simply by example is an open question.) When and why does culture change abruptly whereas at other times it proceeds glacially?
The relationship between technology and culture also needs to be investigated. How does technology

influence culture and how does culture shape technological change? Some papers (for instance, Greenwood and Guner, 2005; Greenwood, Seshadri and Yorukoglu, 2002) argue that sexual norms and female LFP changed because of changes in technology. These papers ignore, among other things, the endogeneity of demand for new technology. Despite the convenient simplification of treating technology as a primitive, it too is endogenous. The extent to which societies put resources into developing technology that 'liberates' individuals from household work, for example, depends on things such as whether slavery is available or whether women expect to work in the market or at home. Put differently, both the relative price of market versus household labour and the elasticity of labour supply depend on the institutions (for example, slavery) and expected division of labour (for example, clearly differentiated gender roles) that are in place. The opposite is also true – the extent to which one can substitute capital for labour, whether at work or at home, helps determine which institutions are viable and may determine the pace and ease with which beliefs or preferences change.

From a theoretical perspective, the endogeneity of preferences and beliefs raises difficult questions for welfare. How should we evaluate policies once we recognize that preferences can change? While this is indeed a vexing and problematic question for welfare economics, recognizing that man is a social animal that is (perhaps uniquely) capable of reflecting upon, and hence changing, his preferences and beliefs greatly enriches our view of ourselves and the world and within it the potential role of economic discourse. In the words of A.O. Hirschman, '*de valoribus est disputandum*'.

See Also

- cultural transmission
- social norms

Bibliography

Algan, Y. and Cahuc, P. 2006. Minimum wage: the price of distrust. Mimeo, CREST-INSEE.

Akerlof, G. and Kranton, R.E. 2000. Economics and identity. *Quarterly Journal of Economics* 115, 715–33.

Bisin, A. and Verdier, T. 2000. Beyond the melting pot: cultural transmission, marriage, and the evolution of ethnic and religious traits. *Quarterly Journal of Economics* 115, 955–88.

Botticini, M. and Eckstein, Z. 2005. Jewish occupational selection: education, restrictions, or minorities? *Journal of Economic History* 65, 922–48.

Bowles, S. 1998. Endogenous preferences: the cultural consequences of markets and other economic institutions. *Journal of Economic Literature* 36, 75–111.

Carroll, C., Rhee, B. and Rhee, C. 1994. Are there cultural effects on saving? Some cross-sectional

evidence. *Quarterly Journal of Economics* 109, 685–99.

Damasio, A. 1995. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Harper Perennial.

Elster, J. 1989. Social norms and economic theory. *Journal of Economic Perspectives* 3(4), 99–117.

Fernández, R. 2007a. Women, work, and culture. *Journal of the European Economic Association* 5, 305–32.

Fernández, R. 2007b. Culture as learning: the evolution of female labor force participation over a century. Mimeo, New York University.

Fernández, R. 2007c. The co-evolution of culture and institutions. Mimeo, New York University.

Fernández, R. and Fogli, A. 2005. Culture: an empirical investigation of beliefs, work, and fertility. Working Paper No. 11268. Cambridge, MA: NBER.

Fernández, R. and Fogli, A. 2006. Fertility: the role of culture and family experience. *Journal of the European Economic Association* 4, 552–61.

Fernández, R., Fogli, A. and Olivetti, C. 2004. Mothers and sons: preference formation and female labor force dynamics. *Quarterly Journal of Economics* 119, 1249–99.

Greenwood, J. and Guner, N. 2005. *Social Change*. Economie d'Avant Garde Research Reports No. 9, Economie d'Avant Garde.

Greenwood, J., Seshadri, A. and Yorukoglu, M. 2005. Engines of liberation. *Review of Economic Studies* 72, 109–33.

Greif, A. 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy* 102, 912–50.

Greif, A. 2005. *Institutions: Theory and History. Comparative and Historical Institutional Analysis*. Cambridge: Cambridge University Press.

Guiso, L., Sapienza, P. and Zingales, L. 2003. People's opium? Religion and economic attitudes. *Journal of Monetary Economics* 50, 225–82.

Guiso, L., Sapienza, P. and Zingales, L. 2005. Cultural biases in economic exchange. Working Paper No. 11005. Cambridge, MA: NBER.

Guiso, L., Sapienza, P. and Zingales, L. 2006. Does culture affect economic outcomes? *Journal of Economic Perspectives* 20(2), 23–48.

Hanushek, E. and Kimko, D. 2000. Schooling, labor-force quality, and the growth of nations. *American Economic Review* 90, 1184–208.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. and McElreath, R. 2001. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review* 91 (2), 73–8.

Ichino, A. and Maggi, G. 2000. Work environment and individual background: explaining regional shirking differentials in a large Italian firm. *Quarterly Journal of Economics* 115, 1057–90.

Mailath, G. and Postlewaite, A. 2003. The social context of economic decisions. *Journal of the European Economic Association* 1, 354–62.

Manski, C. 2000. Economic analysis of social interactions. *Journal of Economic Perspectives* 14(3), 115–36.

Marmot, M.G., Syme, S.L., Kagan, A., Kato, H., Cohen, J.B. and Belsky, J. 1975. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: prevalence of coronary and hypertensive heart disease and associated risk factors. *American Journal of Epidemiology* 102, 514–25.

Stigler, G. and Becker, G. 1977. De gustibus non est disputandum. *American Economic Review* 67, 76–90.

Tabellini, G. 2005. Culture and institutions: economic development in the regions of Europe. Working Paper No. 1492, CESifo.

How to cite this article

Fernández, Raquel. "culture and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000282> doi:10.1057/9780230226203.0346

The New Palgrave Dictionary of Economics Online

Cunningham, William (1849 – 1919)

O. Kurer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cunningham, W.; economic history; English Historical School; imperialism; Marshall, A.; protectionism

Article

A member of the English Historical School, Cunningham was educated at the Universities of Edinburgh and Cambridge. He held various posts as lecturer at Cambridge and was elected Fellow of Trinity College in 1891. From 1891 to 1897 he was Tooke Professor of Statistics of King's College London. In addition, he pursued a religious career. He was ordained in 1874 and rose to be Archdeacon of Ely (1907–19).

Cunningham was one of the most important pioneers in economic history. His *Growth of English Industry and Commerce* (1882) was the first textbook in the field, widely used for several decades and an important foundation on which English economic history was to be constructed, and he relentlessly fought for the recognition and establishment of economic history as an independent discipline.

Cunningham became increasingly hostile towards economic theory. He felt that its assumptions about human behaviour and the institutional framework were leading to insufficiently complete analyses and were blatantly unrealistic for most periods in history. In 1892 he started the English Methodenstreit by attacking Marshall for constructing economic history from general principles instead of empirical data. The debate was partly the result of his personal and professional antagonism towards Marshall and his wish to apply economics to politics.

Cunningham shifted from an internationalist and free trader to a nationalist and protectionist, making the preservation and strengthening of the nation-state his most weighty political and economic objective. By the time of the fiscal controversy in 1903 he fully endorsed the tariff reform movement and subscribed to imperialism, with the great empire securing peace and order.

Selected works

1882. *The Growth of English Industry and Commerce*. Cambridge: Cambridge University Press.

1892a. The perversion of economic history. *Economic Journal* 2, 491–506.

1892b. The perversion of economic history. A reply to Professor Marshall. *Pall Mall Gazette*, 29 September, and *Academy*, 1 October, 288.

1904. *The Rise and Decline of the Free Trade Movement*. Cambridge: Cambridge University Press.

1911. *The Case Against Free Trade*. Preface by Joseph Chamberlain. London: John Murray.

Bibliography

Cunningham, A. 1950. *William Cunningham: Teacher and Priest*. Preface by F.R. Salter. London: Society for Promoting Christian Knowledge.

Foxwell, H.S. 1919. Archdeacon Cunningham (obituary). *Economic Journal* 29, 382–90.

Maloney, J. 1976. Marshall, Cunningham, and the emerging economics profession. *Economic History Review* 29, 440–51.

Maloney, J. 1985. *Marshall, Orthodoxy and the Professionalisation of Economics*. Cambridge: Cambridge University Press.

Scott, W.R. 1920. William Cunningham, 1849–1919. *British Academy, Proceedings* 9, 465–74.

Semmel, B. 1960. *Imperialism and Social Reform: English Imperial Thought 1895–1914*. London: George Allen & Unwin.

Wood, J.C. 1983. *British Economists and the Empire, 1860–1914*. Beckenham, Kent: Croom Helm.

How to cite this article

Kurer, O. "Cunningham, William (1849–1919)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000460> doi:10.1057/9780230226203.0347

The New Palgrave Dictionary of Economics Online

Cunynghame, Henry Hardinge (1848 – 1935)

J.K. Whitaker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cunynghame, Henry Hardinge; demand; supply curve

Article

Soldier, lawyer, civil servant, polymath and amateur economist, Sir Henry Cunynghame was born of distinguished forebears on 8 July 1848 at Penshurst. He died at Eastbourne on 3 May 1935, having been knighted in 1908. In 1870 he entered St John's College, Cambridge, to study law, throwing over a promising military career. There he became a favourite of Alfred Marshall and was infected by an enthusiasm for ‘geometrical political economy’, a topic on which he was eventually to publish one of his many books (1904). There too he invented for Marshall a machine (now lost) for drawing a grid of rectangular hyperbolae (Guillebaud, 1961, Vol. II, pp. 37–8).

Called to the Bar in 1875, Cunynghame had a varied career in law and government, but always retained his interest in economics. He occasionally lectured on the subject (his *Notes on Exchange Value* (1880) were printed for one such course) and in the later 1880s belonged to the economic discussion group which met at the Hampstead home of Henry Ramée Beeton. (P.H. Wicksteed, G.B. Shaw, H.S. Foxwell and F.Y. Edgeworth were among the regulars.) There he presented a paper (1888) defending Marshall's supply curve against Wicksteed's criticisms. The analysis of external effects in production and consumption, his most significant theoretical contribution, first appeared here, the arguments being amplified, but not much clarified, in Cunynghame (1892). His other notable contribution, the use of back-to-back demand–supply diagrams to analyse markets linked by trade, appeared in Cunynghame (1903). The 1904 book, although lively and praised by J.M. Keynes, added little and, indeed, rather compounded earlier ambiguities by a certain flabbiness of thought. Cunynghame's last economic publication (1912) was a valedictory address on methodology. For further biographical detail see Keynes (1935) and Ward and Spencer (1938). Consult also letters by Marshall (reproduced in Pigou, 1925, pp. 447–52; Guillebaud, 1961, Vol. II, pp. 809–13) and Edgeworth's review (1905) of Cunynghame (1904).

Selected works

1880. *Notes on Exchange Value*. London, privately printed.

1888. *Some remarks on Demand and Supply Curves, and their Interpretation*. London, privately printed.

1892. Some improvements in simple geometrical methods of treating exchange value, monopoly, and rent. *Economic Journal* 2(March), 35–52.

1903. The effect of export and import duties on price and production examined by the graphic method. *Economic Journal* 13(September), 313–23.

1904. *Geometrical Political Economy*. Oxford: Clarendon Press.

1912. Address to the Economic Science and Statistical Section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society* 76(December), 88–98.

Bibliography

Edgeworth, F.Y. 1905. Review of Cunynghame (1904). *Economic Journal* 15(March), 62–71. Reprinted in F.Y. Edgeworth, *Papers Relating to Political Economy*, Vol. 3, London, Macmillan, 1925.

Guillebaud, C.W. (ed.) 1961. Alfred Marshall, *Principles of Economics* Ninth (Variorum) Edition, London: Macmillan.

Keynes, J.M. 1935. Obituary: Sir Henry Cunynghame. *Economic Journal* 45(178), June, 398–406. Reproduced in J.M. Keynes, *Collected Writings*, Vol. X (Essays in Biography), London: Macmillan.

Pigou, A.C. (ed.) 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Ward, C.H.D. and Spencer, C.B. 1938. *The Unconventional Civil Servant: Sir Henry Cunynghame*. London: Michael Joseph.

How to cite this article

Whitaker, J.K. "Cunynghame, Henry Hardinge (1848–1935)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000461>
doi:10.1057/9780230226203.1857

The New Palgrave Dictionary of Economics Online

currency boards

Federico Sturzenegger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Currency boards are exchange rate arrangements in which the exchange rate is fixed to an anchor currency and central banks just buy and sell domestic currency at this exchange rate. We review the advantages and disadvantages of currency boards. While some of the alleged benefits of currency boards have diminished hand in hand with a reduction in inflation rates in most countries since the mid-1990s, currency boards may remain an attractive option for certain countries.

Keywords

bank crises; central bank independence; commitment; credibility; currency boards; currency unions; dollarization; exchange rate policy; foreign exchange markets; inflation; inflation targeting; inflation expectations; international reserves; lender of last resort; monetary base; monetary policy; money supply; optimal currency area; seigniorage

Article

A currency board is defined as an exchange rate arrangement in which the exchange rate is fixed to an anchor currency and the central bank operates with a simple rule that precludes the monetary authorities from issuing money unless they obtain an equivalent amount of international assets to back it. From a practical point of view this means that the central bank has no independent monetary policy and that it creates or contracts the money supply only as the result of its interventions in the foreign exchange market. If there is excess demand for domestic currency capital will flow in (probably in response to an increase in interest rates) and the central bank, by acquiring these flows, will expand the money supply. If there is excess supply of domestic currency, the central bank will take in this excess supply by giving away international assets, thus contracting the money supply. In some cases this rule is implemented by forcing the central bank to have full backing of domestic base money with international reserves. In some cases a currency board does not require a one-to-one backing of the monetary base, but it still

precludes the conduct on an independent monetary policy beyond very strict limits. In fact, a currency board also differs from a typical peg in its commitment to the system, which is usually enshrined in law and in the Central Bank charter.

As of July 2006 the exchange rate arrangement classification published by the International Monetary Fund (IMF) identifies 13 countries with currency boards. Of these, six correspond to countries in the Eastern Caribbean Currency Union (Antigua and Barbuda, Dominica, Grenada, St Kitts and Nevis, St Lucia and St Vincent and the Grenadines), plus seven others: Bosnia and Herzegovina, Brunei Darussalam, Bulgaria, China-Hong Kong SAR, Djibouti, Estonia and Lithuania. Because all these countries are relatively small, currency boards are placed in a relatively unpopular category amongst potential exchange rate regimes.

There are two main reasons why countries have typically used currency boards. In some cases the currency board is more attractive than a common currency. For example, for the Eastern Caribbean countries mentioned above it seems relatively obvious they should use the US dollar as currency to maximize the benefits from a stable exchange rate arrangement with their almost sole trading partner. However, the currency board allows them to keep the exchange rate credibly fixed without giving up the seigniorage revenue of domestic currency. In other cases countries have resorted to a currency board as a way out of monetary and inflation chaos. Argentina's currency board experience in the 1990s and Bulgaria's currency board are appropriate examples. Even though, as we will see below, the evidence points to large trade benefits of currency boards, it is typically assumed that the main benefit of currency boards is as a tool to fight inflation.

The interest in and excitement about currency boards reflects both the need that countries have faced to solve either of the two problems mentioned above – currency integration without seigniorage cost and exiting from a high inflation situation – and the assessment made at the time of whether a currency board is the most efficient way to reach those objectives. Recent years have been unkind to currency boards on both counts. While the use of a currency board as a replacement for a common currency remains a valid motive, its effect as an anti-inflation device has become less relevant as inflation rates fell throughout the 1990s. In 2007 most countries exhibit single-digit inflation rates, and only a handful of exotic cases appear to have a monetary policy that is out of control. The high-inflation history of yesteryear has been critical to this improvement by fostering much stronger fiscal policies and monetary policies that are much freer from political pressures (both when central banks are independent and when they are not) and increasingly within an inflation targeting framework. As inflation has decreased, so have the benefits of a currency board, thus making it a relatively less attractive proposition.

Furthermore, while before the demise of Argentina's currency board in early 2002 no currency board had been forced to end, the fact that Argentina's currency board came to an end in the midst of a major crisis (after enduring a long period of high interest rates) raised some questions as to how much credibility the regime actually bought. As a result, many countries have opted to jump directly all the way to dollarization (for example, El Salvador and Ecuador) or to pursue integration into a currency union (Slovenia) thus making currency boards lose ground even to alternative ‘harder’ exchange-rate commitments.

In spite of the recent drop in interest in this specific regime, nothing precludes a rise in interest again in the future, so a discussion of the specifics of currency boards remains useful. The best way to organize the discussion is to present the advantages of a currency board, then move to the disadvantages, and then attempt a synthesis.

Advantages of a currency board

The main advantage that is ascribed to a currency board is the credibility gains that it allows, helping deliver lower inflation and better fiscal results. The argument is simple: a currency board represents a strong commitment that if broken can have a large and costly effect on expectations. Because politicians fear this loss of credibility, the currency board, while in place, lowers inflation expectations and inflation itself and should provide the incentives for an improvement in fiscal behaviour.

These predictions have been broadly borne out. On the inflation front Ghosh, Gulde and Wolf (1998), drawing on a data-set for all IMF countries between 1970 and 1996, found that countries with currency boards delivered an inflation rate that was about four per cent lower, a sizable effect. This result has held up in later work (see for example Levy-Yeyati and Sturzenegger, 2001; and Kuttner and Posen, 2001).

The record on fiscal discipline is also relatively favourable. Ghosh, Gulde and Wolf (1998) and Culp, Hanke and Miller (1999) find that countries on currency boards tend to run tighter fiscal policies. Fatas and Rose (2000) also find that currency boards are associated with fiscal restraint (though, somewhat surprisingly, this restraint does not carry on to dollarized economies or those operating within the context of a common currency). Anecdotal evidence also seems to point in the same direction. In 2001, as Argentina's currency board was under fire, fiscal authorities implemented large budget adjustments in an attempt to strengthen the system.

Currency boards may also have an effect on trade as a result of the stability it induces on the exchange rate, an effect similar to the one that has been identified for countries that adopt a common currency with other countries. This exercise is specifically undertaken in Frankel and Rose (2002), who find that the effect of a currency board is a more than tripling of trade (in fact they find that the trade effects for currency boards and common currencies are statistically indistinguishable). Thus the trade motive for a currency board seems to be important. Added to the benefits of saving on seigniorage, it explains why currency boards may remain an attractive option for some small countries.

Disadvantages of a currency board

Four main arguments have been advanced against currency boards. First, the fact that it precludes monetary authorities from running an independent monetary policy and that the exchange rate cannot adjust in response to real shocks; second, that it may ‘hide’ underlying problems, leading to larger crises down the road; third, that it stimulates large currency mismatches in the portfolio structures of government and the private sector; and fourth, that it limits the ability of the central bank to act as a lender of last resort, thus hindering the possibility of developing a locally based financial sector.

The debate has focused mostly on whether alternative mechanisms and policies within the context of the currency board are available to deal with these problems. Let us review each of them briefly.

On the loss of monetary/exchange rate policy, the question is how relevant a loss this it. It can be argued that the idea of a currency board is indeed to limit the scope for an independent monetary policy, which had otherwise proven unable to contain high inflation. To the extent that inflation and fiscal policy improve, not much may be lost relative to the situation in which monetary policy merely induced inflation without any particular benefit in terms of macroeconomic stabilization. Thus, assessing

whether this is a cost requires us to evaluate what the counterfactual is. Proponents of currency boards could argue that only countries where monetary policy serves no purpose choose currency boards as a commitment device.

Of course, if monetary policy *were* possible, the costs of doing without it may turn out to be particularly costly for currency boards. The case of Argentina helps illustrate why this should be so. Argentina had established a currency board with the dollar to quell inflation expectations in the early 1990s. Like any other emerging country, it was hurt by the rush out of emerging markets following Russia's default in 1998. This rush strongly appreciated the dollar, making Argentina's currency stronger exactly when the country needed it to weaken. The fact that currency boards require a strong anchor currency and that capital flows may strengthen these currencies when there is turmoil in emerging countries – thus moving the exchange rate exactly in the opposite direction to the one the country would have otherwise chosen – poses a problem for currency boards during periods of high turbulence in international financial markets.

Of course, as much as in the optimal currency area debate, how costly the loss of the monetary instrument is depends on the availability of alternative adjustment mechanisms: fiscal transfers, remittances, labour market mobility, or internal price flexibility, which may all operate as substitutes for the loss of monetary policy (the effectiveness of these alternative mechanisms may explain the different fates of Hong Kong's and Argentina's currency boards). Fiscal policy can also be used as a stabilizer that may substitute for the lack of exchange or monetary policy, though the ability of countries to use it seems relatively limited, particularly for those countries that opted for a currency board as a result of their poor fiscal policies. Some evidence for the fact that the lack of monetary policy may hurt is provided by Levy-Yeyati and Sturzenegger (2001), who compare the growth performance of hard pegs generally (including currency boards) with other regimes. They find that hard pegs trail floating regimes in growth performance (though not by more than pegs or intermediate regimes). However, this allows us to conclude that, in the end, the lack of policy responses may have a detrimental effect on overall economic performance.

The fact that currency boards may delay an adjustment has also been a cause of concern. Aizenman and Glick (2005) and Kuttner and Posen (2001) have both found that the harder and longer the peg, the larger are the depreciations upon exiting. This is to be expected, because the stronger the commitment, the fixed exchange rate spell will be typically longer, and only under more unfavourable conditions will the peg be abandoned, suggesting that an earlier adjustment may have been beneficial. This conclusion, however, should be treated with care because it fails to take into account the fact that this stringency also helps avoid many exits that later on would have turned out to be unnecessary.

The same caution should be used when evaluating the tendency of currency boards to foster the evolution of mismatches in government and private sector debt structures. The basic idea is that as long as the currency board holds countries develop a tendency to ‘dollarize’ their financial sectors (see Catao and Terrones, 2000), with banks piling foreign currency deposits on their liability side, firms borrowing in dollars abroad and governments issuing debt in dollars. This is a problem because the asset side of these borrowers is in most cases linked mostly to the local economy, and thus, whether denominated in foreign currency or not, subject to currency risk in the event of a devaluation. This mismatch, however, is a double-edged sword. On the one hand it increases the commitment of the authorities to the peg (and this is why sometimes it is encouraged by the authorities as an additional credibility booster), but on the other it may also trigger large capital outflows in anticipation of a crisis. In the presence of large mismatches, agents would correctly anticipate a devaluation to produce a costly crisis, thus accelerating

the run and the likelihood that the currency will sink. How these two factors play out during a crisis depends on the specifics of each individual country.

Finally, a currency board limits the ability of the central bank to operate as lender of last resort, particularly in the event of a bank run. This has been suggested as an explanation of why countries with currency boards quickly develop an international based banking system (typically with local institutions bought by foreign banks) which is better insured against runs at any specific location. Proponents of currency boards have suggested several alternatives to replace the central bank's function as lender of last resort with other mechanisms. Among these are the possibility of the government operating as lender of last resort, potentially by borrowing in dollars in times of need; the setting up of insurance schemes by which financial institutions buy in advance the access to funds in the context of a systemic liquidity run (these schemes were implemented by Mexico and Argentina); tighter capital and liquidity requirements on the banking sector; and the piling up of ‘extra reserves’ as far as possible. The first of these mechanisms is doubtful, as the government may have limited access to financing when it faces a crisis, and the others entail a cost. However, it may be said that some of these schemes have been implemented and used successfully. Specifically, Argentina used its contingent credit line with private banks during its 2001 crisis and banks honoured their pledge at the time.

Where does this leave us?

The conclusion is then that, as much as with currency unions, there seems to be a strong trade motive to set up a currency board. In fact, for a fiscally sound small country with the ability to conduct fiscal policy with some flexibility a currency board may be superior to a common currency as it allows the country to retain the seigniorage on its money stock. For larger middle-income countries a currency board has been pursued more as a way of improving credibility than anything else. While currency boards seem to have delivered, the Argentina case also suggests that their role in improving credibility cannot be taken fully for granted. If a currency board is implemented in times of easy access to international financial markets, fiscal discipline may be sidestepped and a fiscal and currency crisis may still occur at the end of the day. Additionally, policymakers should ask themselves if it makes sense to buy the credibility through a peg, or to buy it the hard way, day by day, implementing reasonable fiscal policies while maintaining some degree of flexibility in monetary policy. The successful experience since the mid-1990s of many countries with managed floating regimes and inflation targeting seems to point to this direction. If this trend continues, currency boards may become even rarer in the future.

See Also

- currency unions
- dollarization

Bibliography

Aizenman, J. and Glick, R. 2005. Pegged exchange rate regimes – a trap? Working Paper No. 2006–07, Federal Reserve Bank of San Francisco.

Catao, L. and Terrones, M. 2000. Determinants of dollarization: the banking side. Working Paper No. 00/146, International Monetary Fund.

Culp, C., Hanke, S. and Miller, M. 1999. The case for an Indonesian currency board. *Journal of Applied Corporate Finance* 11, 57–65.

Fatas, A. and Rose, A.K. 2000. Do monetary handcuffs restrain Leviathan? Fiscal policy in extreme exchange rate regimes. Discussion Paper No. 2692, CEPR.

Frankel, J. and Rose, A. 2002. An estimate of the effect of common currencies on trade and income. *Quarterly Journal of Economics* 117, 437–66.

Ghosh A., Gulde, A.-M. and Wolf, H.C. 1998. Currency boards: the ultimate fix? Working Paper No. 98/8, International Monetary Fund.

Kuttner, K. and Posen, A. 2001. Beyond bipolar: a three-dimensional assessment of monetary frameworks. Working Paper No. 52, Oesterreichische Nationalbank.

Levy-Yeyati, E. and Sturzenegger, F. 2001. Exchange rate regimes and economic performance. *IMF Staff Papers* 47, 62–98.

How to cite this article

Sturzenegger, Federico. "currency boards." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000466> doi:10.1057/9780230226203.0348

The New Palgrave Dictionary of Economics Online

currency competition

Stacey L. Schreft
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Currency competition’ means the virtually free entry of private-sector firms into the issuance of a currency. Such competition no longer exists, but interest in it revived in the 1970s as high inflation was attributed by some to governments' incentives to overissue their currencies to generate additional seigniorage. Competition was advocated as a potential remedy because it was thought to give issuers an incentive to protect the value of their currencies by limiting issuance.

Keywords

Bank Act 1844 (UK); central banks; currency competition; Currency School; fiat money; free banking; Hayek, F.; inflation; inside and outside money; seigniorage; Suffolk Banking System

Article

‘Currency competition’ refers to the free, or virtually free, entry of private-sector firms into the issuance of a circulating medium of exchange in lieu of a government monopoly on currency issue. Although there is little analytical basis for focusing on the private issuance of securities that circulate at the expense of those that do not, that is exactly the approach of the literature on currency competition and thus of this article.

The best real-world examples of currency competition come from periods, some lasting more than a century, in which countries allowed banks to operate relatively free from regulation. This freedom allowed, among other things, banks to issue paper notes. Shuler (1992) identified 66 countries as having free banking for some period in the 19th and 20th centuries, and all of them reportedly had multiple private-sector note issuers.

Today, there is no true private note issuance. Any privately issued notes are issued by banks that operate as agents of their respective central banks. Shuler (1992) attributed the demise of privately issued notes to several factors. One factor was a shift in attitudes about the need for and proper role of central banks.

The view took hold that currency issuance could be destabilizing if left to the private sector, and governments nationalized currency issuance in their central banks. This was the case in England, for example, where the Currency School came to dominate and the Bank Act of 1844 eliminated private note issuance. Another major factor leading to government monopolies over currency issuance was the First World War and governments' need for additional sources of revenue. The ability to issue currency directly became very appealing.

By the 1970s, governments' monopoly on currency creation was raising its own concerns. These government issuers had an incentive to overissue to generate additional seigniorage revenue. When inflation began rising in the 1970s, some blamed this incentive to overproduce and called for denationalization of currency issuance. Friedrich Hayek (1990) was perhaps the most prominent proponent of a return to currency competition. Hayek argued, in the terms of today, that an equilibrium could exist with competitive issuance and that it would likely dominate the equilibrium arising when the government monopolizes currency issuance. The logic was that the demand for a privately issued currency depends in part on the currency's quality because such currencies are distinguishable. The more units of a currency supplied, the lower is the currency's value in exchange and thus its perceived quality and the public's demand for it. Competition would thus give issuers an incentive to protect the value of their currencies by not overissuing.

In considering what currency competition might look like, economists rediscovered the free banking periods, and a literature arose studying them. The first wave of that literature consisted of historical studies of free banking and private note issuance, although there were also a few theoretical models. Later, in the 1990s, the potential for new electronic means of payment, such as stored value and digital currencies for the Internet, led to another generation of research on currency competition, this time primarily theoretical.

Most discussions of currency competition, whether from a theoretical or an historical perspective, failed to distinguish inside money from outside money. Hellwig's work (1985) was an exception. Inside money is a claim that obligates its issuer to redeem or exchange the money for some specified monetary or non-monetary object. Failure to do so, perhaps because of insufficient reserves held against the money, can result in a failure to fulfil that obligation and ultimately bankruptcy. The value of a privately issued inside money depends in part, then, on the likelihood of the issuer fulfilling its claim, and only in part on the value of using the money in exchange. Outside money is not a claim against the issuer or anyone else. The issuer makes no promise to redeem its currency at any time for anything of value. The value of a privately issued outside money derives solely from its value in exchange.

The experience in the US free banking era (1837–63) is an example of the importance of the claim that backs an inside money. Bank notes issued in the free banking era were supposed to be fully backed to guarantee the issuer's ability to redeem them, but often they were not. In some cases, no backing was held. Bank note reporters kept track of the financial condition of issuers and of the prices at which notes were trading. Weber (2002) found that notes traded for one another at flexible exchange rates that often depended in part on the extent to which the notes were backed. When the public became aware that an issuer's notes lacked backing, the notes stopped circulating.

The distinction between inside and outside money is important for studying currency competition. Competition in outside note issuance is likely to divert fewer resources from consumption and production than competition in inside note issuance because there is no need to hold reserves against the outside notes. However, without reserves to back outside money, the money is likely to be overissued

because of its near-zero marginal cost of production. Thus, the welfare gain from avoiding overissuing with an inside money must be balanced against the welfare loss from holding full or fractional reserves against such money.

Historical experience with outside money has almost always involved a single, government-issued fiat currency. The existing theoretical literature suggests why privately issued outside money is virtually never observed. In many different economic environments, economists have shown that there can be no equilibrium with competitive issuance of outside money if issuers cannot make binding commitments about the volume of notes they will issue. Taub (1985) and Bryant (1981) showed this in an overlapping-generations model. Ritter (1995) did so in a search model of money. In all cases, the argument is as follows, and similar to Hayek's. If issuing new money is costless, issuers cannot make binding commitments, and money has some positive value, then any private agent that issues notes will issue an unlimited quantity, driving the inflation rate to infinity and the real value of the money to zero. Rational agents would anticipate this ultimate outcome and be unwilling to hold the money at any earlier date. The inability to make binding commitments, coupled with a time inconsistency problem, is a key feature of this argument because issuers always want to believe they will constrain their note issuance, but when they need to they never have the incentive to do so.

A few models have gotten around this result. Klein (1974), for example, provided an early argument based on reputation formation for the existence of equilibria with free entry into private issuance. He argues that the monies of different issuers can be distinguishable by quality, so they can circulate at flexible exchange rates with one another. His discussion, however, blurs the distinction between inside and outside money.

In another example, Martin and Schreft (2006) showed that privately issued outside money can be valued if agents believe that all notes issued up to some threshold will be valued, but additional notes will be worthless. These beliefs create a discontinuity in the value of the marginal unit of currency. Because the value of a marginal unit of currency reaches zero for some finite supply, the limit argument no longer applies. Martin and Schreft derived their existence result in both an overlapping generations and a search-theoretic environment, though it should hold in any environment in which fiat currency could be valued. Interestingly, welfare is not necessarily greater with competitive issuance than with monopoly issuance and depends on the environment considered. In the search environment, neither competitive issuers nor a monopolist achieve the efficient quantity of money in the long run. In the overlapping-generations environment, the efficient allocation is achieved in finitely many periods if agents incur a cost of becoming money issuers. A monopoly issuer might achieve as desirable an allocation, but only if its actions are sufficiently constrained by agents' beliefs.

In contrast, the historical experience with inside money has involved multiple inside monies that are all convertible into some single dominant outside money. A modern literature on privately issued inside notes, largely attributable to Wallace and others, has considered this case. Cavalcanti and Wallace (1999a; 1999b) studied a search-theoretic model with an exogenously given and indivisible outside money and inside money issued by private agents known as banks. To get the private money to be valued, they assumed that issuers who do not accept a note when presented with one face a stiff punishment: they lose the ability to issue notes and revert to autarky. This assumption is reminiscent of the redemption requirements of successful systems for private inside currency issuance, like the Suffolk Banking System that operated in New England in the early 1800s. The authors found that, if the stock of outside money is sufficiently small, then the optimal mechanism has private notes issued and also

redeemed on demand. Additionally, expected utility is greater in economies with inside money than only outside money because the set of implementable allocations is larger.

In the United States, at least, it is claimed that little currently prohibits private-sector issuance of outside currency in either paper or digital form. The laws prohibiting it have either expired or been repealed. It will be interesting to see if a resurgence of private issuance occurs.

See Also

- fiat money
- free banking era
- Hayek, Friedrich August von
- inflation
- inside and outside money

Bibliography

Bryant, J. 1981. The competitive provision of fiat money. *Journal of Banking and Finance* 5, 587–93.

Cavalcanti, R. de O. and Wallace, N. 1999a. A model of private bank-note issue. *Review of Economic Dynamic*s 2, 104–36.

Cavalcanti, R. de O. and Wallace, N. 1999b. Inside and outside money as alternative media of exchange. *Journal of Money, Credit, and Banking* 31, 443–57.

Hayek, F. 1990. *Denationalisation of Money: The Argument Refined*, 3rd edn. London: Institute of Economic Affairs.

Hellwig, M. 1985. What do we know about currency competition? *Zeitschrift für Wirtschafts- und sozialwissenchaften* 105, 565–88.

Klein, B. 1974. The competitive supply of money. *Journal of Money, Credit, and Banking* 6, 423–53.

Martin, A. and Schreft, S. 2006. Currency competition: a partial vindication of Hayek. *Journal of Monetary Economics* 53, 2085–111.

Ritter, J. 1995. The transition from barter to fiat money. *American Economic Review* 85, 134–49.

Schreft, S. 1997. Looking forward: the role for government in regulating electronic cash. *Federal Reserve Bank of Kansas City Economic Review* (Fourth Quarter), 59–84.

Shuler, K. 1992. The world history of free banking: an overview. In *The Experience of Free Banking*, ed. K. Dowd. New York: Routledge.

Taub, B. 1985. Private money with many suppliers. *Journal of Monetary Economics* 16, 195–208.

Weber, W. 2002. Banknote exchange rates in the antebellum United States. Working Paper No. 623. Research Department, Federal Reserve Bank of Minneapolis.

How to cite this article

Schreft, Stacey L. "currency competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000560> doi:10.1057/9780230226203.0349

The New Palgrave Dictionary of Economics Online

currency crises models

Craig Burnside, Martin Eichenbaum and Sergio Rebelo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Currency crises have occurred frequently in the post-war era. In this article we review the literature on the causes and consequences of currency crises. First-generation models attribute a central role to fiscal policy as a fundamental determinant of crises. Second-generation models emphasize the possibility of self-fulfilling speculative attacks and multiple equilibria. Third-generation models stress how financial fragility can lead to currency crises.

Keywords

Asian currency crisis (1997); bank runs; budget deficits; consumer optimization; credit risk; currency crises; currency crisis models; exchange rate regimes; financial crises; fiscal theory of the price level; fixed exchange rates; government budget constraint; government guarantees; intertemporal budget constraint; liquidity crises; money demand functions; money supply; purchasing power parity; seigniorage; speculative attacks

Article

There have been many currency crises during the post-war era (see Kaminsky and Reinhart, 1999). A currency crisis is an episode in which the exchange rate depreciates substantially during a short period of time. There is an extensive literature on the causes and consequences of a currency crisis in a country with a fixed or heavily managed exchange rate. The models in this literature are often categorized as first-, second- or third-generation.

In first-generation models the collapse of a fixed exchange rate regime is caused by unsustainable fiscal policy. The classic first-generation models are those of Krugman (1979) and Flood and Garber (1984). These models are related to earlier work by Henderson and Salant (1978) on speculative attacks in the gold market. Important extensions of these early models incorporate consumer optimization and the government's intertemporal budget constraint into the analysis (see Obstfeld, 1986; Calvo, 1987; Drazen

and Helpman, 1987; Wijnbergen, 1991). Flood and Marion (1999) provide a detailed review of first-generation models.

In a fixed exchange rate regime a government must fix the money supply in accordance with the fixed exchange rate. This requirement severely limits the government's ability to raise seigniorage revenue. A hallmark of first-generation models is that the government runs a persistent primary deficit. This deficit implies that the government must either deplete assets, such as foreign reserves, or borrow to finance the deficit. It is infeasible for the government to borrow or deplete reserves indefinitely. Therefore, in the absence of fiscal reforms, the government must eventually finance the deficit by printing money to raise seigniorage revenue. Since printing money is inconsistent with keeping the exchange rate fixed, first-generation models predict that the regime must collapse. The precise timing of its collapse depends on the details of the model.

The key ingredients of a first-generation model are its assumptions regarding purchasing power parity (PPP), the government budget constraint, the timing of deficits, the money demand function, the government's rule for abandoning the fixed exchange rate, and the post-crisis monetary policy. In the simplest first-generation models there is a single good whose domestic currency price is P_t and whose foreign currency price is 1. Let S_t denote the nominal exchange rate. PPP implies $P_t = S_t$. Suppose for simplicity that the government has a constant ongoing primary deficit, δ. It finances this deficit by reducing its stock of foreign reserves, f_t, which can either evolve as a smooth function of time or jump discontinuously. In the former case, f_t evolves according to $\dot{f}_t = rf_t - \delta + \dot{M}_t / S_t$, where r is the real interest rate, M_t is the monetary base, and a dot over a variable denotes its derivative with respect to time. When foreign reserves change discontinuously, $\Delta f_t = \Delta(M_t / S_t)$. When $\delta > rf_0$ interest income from foreign assets will not be sufficient to finance the deficit.

To illustrate the key properties of first-generation models, we make three simplifying assumptions. First, money demand takes the Cagan (1956) form, $M_t = \theta P_t \exp[-\eta(r + \pi_t)]$, where $\theta > 0$ and $\pi_t = \dot{P}_t / P_t$ is the inflation rate. Second, the government abandons the fixed exchange rate regime when its foreign reserves are exhausted. Third, as soon as foreign reserves are exhausted, the government prints money at a constant rate μ to fully finance its deficit.

These assumptions imply that after the crisis the level of real balances, $m_t = M_t / P_t$, is constant and equal to $\overline{m} = \theta \exp[-\eta(r + \mu)]$. The post-crisis government budget constraint reduces to $\delta = \mu\overline{m}$. This equation determines μ. Let t^* denote the date at which foreign reserves are exhausted and the government abandons the fixed exchange rate regime. PPP implies $S_{t^*} = P_{t^*} = \overline{M} / \overline{m}$, where $\overline{M}$ is the monetary base the instant after date t^*. Under perfect foresight the exchange rate cannot jump discontinuously at t^* since such a jump would imply the presence of arbitrage opportunities. Given that the exchange rate must be a continuous function of time at t^*, $S_{t^*} = S$ and $\overline{M} = \overline{m}S$.

Prior to the crisis real balances are given by $m = \theta \exp(-\eta r)$. Therefore, at date t^* there is a sudden drop in real money demand from m to $\overline{m}$ implying that reserves drop discontinuously to zero at time t^*: $\Delta f_{t^*} = \overline{m} - m$. This is why the literature refers to t^* as the date of the speculative attack. Prior to the crisis the government's reserves fall at the rate $\dot{f}_t = rf_t - \delta$. The budget constraint implies that $t^* = \ln\{[\delta - r(m - \overline{m})] / (\delta - rf_0)\} / r$. While the collapse of the fixed exchange rate regime is

inevitable, it does not generally occur at time zero unless $m - \overline{m} > f_0$.

A shortcoming of this type of first-generation model is that the timing of the speculative attack is deterministic and the exchange rate does not depreciate at the time of the attack. These shortcomings can be remedied by introducing shocks into the model, as in Flood and Garber (1984).

Early first-generation models predict that ongoing fiscal deficits, rising debt levels, or falling reserves precede the collapse of a fixed exchange rate regime. This prediction is inconsistent with the 1997 Asian currency crisis. This inconsistency led many observers to dismiss fiscal explanations of this crisis. However, Corsetti, Pesenti and Roubini (1999), Burnside, Eichenbaum and Rebelo (2001a), and Lahiri and Végh (2003) show that bad news about prospective deficits can trigger a currency crisis. Under these circumstances a currency crisis will not be preceded by persistent fiscal deficits, rising debt levels, or falling reserves. These models assume that agents receive news that the banking sector is failing and that banks will be bailed out by the government. The government plans to finance, at least in part, the bank bailout by printing money beginning at some time in future. Burnside, Eichenbaum and Rebelo (2001a) show that a currency crisis will occur before the government actually starts to print money. Therefore, in their model, a currency crisis is not preceded by movements in standard macroeconomic fundamentals, such as fiscal deficits and money growth. Burnside, Eichenbaum and Rebelo argue that their model accounts for the main characteristics of the Asian currency crisis.

This explanation of the Asian currency crisis stresses the link between future deficits and current movements in the exchange rate. This link is also stressed by Corsetti and Mackowiak (2006), Daniel (2001), and Dupor (2000), who use the fiscal theory of the price level to argue that prices and exchange rates jump in response to news about future deficits.

In first-generation models the government follows an exogenous rule to decide when to abandon the fixed exchange rate regime. In second-generation models the government maximizes an explicit objective function (see, for example, Obstfeld, 1994; 1996). This maximization problem dictates if and when the government will abandon the fixed exchange rate regime. Second-generation models generally exhibit multiple equilibria so that speculative attacks can occur because of self-fulfilling expectations. In Obstfeld's models (1994; 1996) the central bank minimizes a quadratic loss function that depends on inflation and on the deviation of output from its natural rate (see Barro and Gordon, 1983, for a discussion of this type of loss function). The level of output is determined by an expectations-augmented Phillips curve. The government decides whether to keep the exchange rate fixed or not. Suppose agents expect the currency to devalue and inflation to ensue. If the government does not devalue then inflation will be unexpectedly low. As a consequence output will be below its natural rate. Therefore the government pays a high price, in terms of lost output, in order to defend the currency. If the costs associated with devaluing (lost reputation or inflation volatility) are sufficiently low, the government will rationalize agents’ expectations. In contrast, if agents expect the exchange rate to remain fixed, it can be optimal for the government to validate agents’ expectations if the output gains from an unexpected devaluation are not too large. Depending on the costs and benefits of the government's actions, and on agents’ expectations, there can be more than one equilibrium. See Jeanne (2000) for a detailed survey of second-generation models.

Morris and Shin (1998) provide an important critique of models with self-fulfilling speculative attacks. They emphasize that standard second-generation models assume that fundamentals are common knowledge. Morris and Shin demonstrate that introducing a small amount of noise into agents’ signals

about fundamentals will lead to a unique equilibrium.

Many currency crises coincide with crises in the financial sector (Diaz-Alejandro, 1985; Kaminsky and Reinhart, 1999). This observation has motivated a literature that emphasizes the role of the financial sector in causing currency crises and propagating their effects. These third-generation models emphasize the balance-sheet effects associated with devaluations. The basic idea is that banks and firms in emerging market countries have explicit currency mismatches on their balance sheets because they borrow in foreign currency and lend in local currency. Banks and firms face credit risk because their income is related to the production of non-traded goods whose price, evaluated in foreign currency, falls after devaluations. Banks and firms are also exposed to liquidity shocks because they finance long-term projects with short-term borrowing. Eichengreen and Hausmann (1999) argue that currency mismatches are an inherent feature of emerging markets. In contrast, authors such as McKinnon and Pill (1996) and Burnside, Eichenbaum and Rebelo (2001b) argue that, in the presence of government guarantees, it is optimal for banks and firms to expose themselves to currency risk.

Different third-generation models explore various mechanisms through which balance-sheet exposures may lead to a currency and banking crisis. In Burnside, Eichenbaum and Rebelo (2004) government guarantees lead to the possibility of self-fulfilling speculative attacks. In Chang and Velasco (2001) liquidity exposure leads to the possibility of a Diamond and Dybvig (1983) style bank run. In Caballero and Krishnamurthy (2001) firms face a liquidity problem because they finance risky long-term projects with foreign loans but have access to limited amounts of internationally accepted collateral.

An important policy question is: what is the optimal nature of interest rate policy during and after a currency crisis? There has been relatively little formal work on this topic. Christiano, Braggion and Roldos (2006) take an important first step in this direction. They argue that it is optimal to raise interest rates during a currency crisis and to lower them immediately thereafter. Studying optimal monetary policy in different models of currency crises remains an important area for future research.

See Also

- currency crises
- fiscal theory of the price level

Bibliography

Barro, R. and Gordon, D. 1983. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91, 589–610.

Burnside, C., Eichenbaum, M. and Rebelo, S. 2001a. Prospective deficits and the Asian currency crisis. *Journal of Political Economy* 109, 1155–98.

Burnside, C., Eichenbaum, M. and Rebelo, S. 2001b. Hedging and financial fragility in fixed exchange rate regimes. *European Economic Review* 45, 1151–93.

Burnside, C., Eichenbaum, M. and Rebelo, S. 2004. Government guarantees and self-fulfilling speculative attacks. *Journal of Economic Theory* 119, 31–63.

Caballero, R. and Krishnamurthy, A. 2001. International and domestic collateral constraints in a model of emerging market crises. *Journal of Monetary Economics* 48, 513–48.

Cagan, P. 1956. Monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Calvo, G. 1987. Balance of payments crises in a cash-in-advance economy. *Journal of Money Credit and Banking* 19, 19–32.

Chang, R. and Velasco, A. 2001. A model of financial crises in emerging markets. *Quarterly Journal of Economics* 116, 489–517.

Christiano, L., Braggion, F. and Roldos, J. 2006. The optimal monetary response to a financial crisis. Mimeo, Northwestern University.

Corsetti, G. and Mackowiak, B. 2006. Fiscal imbalances and the dynamics of currency crises. *European Economic Review* 50, 1317–38.

Corsetti, G., Pesenti, P. and Roubini, N. 1999. What caused the Asian currency and financial crisis? *Japan and the World Economy* 11, 305–73.

Daniel, B. 2001. The fiscal theory of the price level in an open economy. *Journal of Monetary Economics* 48, 293–308.

Diamond, D. and Dybvig, P. 1983. Bank runs, deposit insurance and liquidity. *Journal of Political Economy* 91, 401–19.

Diaz-Alejandro, C. 1985. Good-bye financial repression, hello financial crash. *Journal of Development Economics* 19, 1–24.

Drazen, A. and Helpman, E. 1987. Stabilization with exchange rate management. *Quarterly Journal of Economics* 102, 835–55.

Dupor, W. 2000. Exchange rates and the fiscal theory of the price level. *Journal of Monetary Economics* 45, 613–30.

Eichengreen, B. and Hausmann, R. 1999. Exchange rates and financial fragility. In *New Challenges for Monetary Policy: A Symposium Sponsored by the Federal Reserve Bank of Kansas City*. Kansas City:

Federal Reserve Bank of Kansas City.

Flood, R. and Garber, P. 1984. Collapsing exchange rate regimes: some linear examples. *Journal of International Economics* 17, 1–13.

Flood, R. and Marion, N. 1999. Perspectives on the recent currency crisis literature. *International Journal of Finance and Economics* 4, 1–26.

Henderson, D. and Salant, S. 1978. Market anticipations of government policies and the price of gold. *Journal of Political Economy* 86, 627–48.

Jeanne, O. 2000. Currency crises: a perspective on recent theoretical developments. Special Papers in International Economics, No. 20, International Finance Section, Princeton University.

Kaminsky, G. and Reinhart, C. 1999. The twin crises: the causes of banking and balance-of-payments problems. *American Economic Review* 89, 473–500.

Krugman, P. 1979. A model of balance of payments crises. *Journal of Money, Credit and Banking* 11, 311–25.

Lahiri, A. and Végh, C. 2003. Delaying the inevitable: interest rate defense and BOP crises. *Journal of Political Economy* 111, 404–24.

McKinnon, R. and Pill, H. 1996. Credible liberalizations and international capital flows: the overborrowing syndrome. In *Financial Deregulation and Integration in East Asia*, ed. T. Ito and A. Krueger. Chicago: University of Chicago Press.

Morris, S. and Shin, H. 1998. Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review* 88, 587–97.

Obstfeld, M. 1986. Speculative attack and the external constraint in a maximizing model of the balance of payments. *Canadian Journal of Economics* 29, 1–20.

Obstfeld, M. 1994. The logic of currency crises. *Cahiers Economiques et Monétaires* 43, 189–213.

Obstfeld, M. 1996. Models of currency crises with self-fulfilling features. *European Economic Review* 40, 1037–47.

Wijnbergen, S. van. 1991. Fiscal deficits, exchange rate crises and inflation. *Review of Economic Studies* 58, 81–92.

How to cite this article

Burnside, Craig, Martin Eichenbaum and Sergio Rebelo. "currency crises models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000204> doi:10.1057/9780230226203.0351

The New Palgrave Dictionary of Economics Online

currency crises

Graciela Laura Kaminsky

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes models and empirical evidence on currency crises. The evidence from developed and developing countries indicates that crises are of different varieties. It also shows that crises do not occur in economies with sound fundamentals, with vulnerabilities far more widespread and profound in emerging economies. Vulnerabilities are associated with fiscal problems, loss of competitiveness and a deteriorating current account, external debt unsustainability, or problems in the financial sector – especially banks. Interestingly, those crises associated with bank fragility are the costliest in terms of output losses and loss of access to international capital markets.

Keywords

budget deficits; currency crises; currency crisis models; deposit insurance; European Monetary System; Exchange Rate Mechanism; fixed exchange rates; foreign-debt defaults; imperfect information; international capital flows; international capital markets; moral hazard; regression tree analysis; self-fulfilling currency crises; sovereign defaults; sticky prices; sudden-stop currency crises

Article

A currency crisis occurs when investors flee from a currency en masse out of fear that it might be devalued. Currency crises are episodes characterized by sudden depreciations of the domestic currency, large losses of foreign exchange reserves of the central bank, and (or) sharp hikes in domestic interest rates.

There have been numerous currency crises since 1980. The so-called debt crisis erupted in 1982 following Mexico's default and devaluation in August. This crisis spread rapidly to all Latin American countries, and by the time it was over, most Latin American countries had devalued their currencies and defaulted on their foreign debts. The debt crisis was followed by a decade of negative growth and isolation from international capital markets. The output costs of this crisis were so large that the 1980s

became known as the ‘lost decade’ for Latin America.

Crises are not just emerging-market phenomena. The 1990s opened with crises in industrial Europe – the European Monetary System (EMS) crises of 1992 and 1993. By the end of these crises, in the summer of 1993, the lira and the sterling had been driven from the Exchange Rate Mechanism (ERM); Finland, Norway, and Sweden had abandoned their unofficial peg to the European Currency Unit (ECU); the Spanish peseta, the Portuguese escudo and the Irish punt had devalued; and Europe's central bank governors and finance ministers had widened the ERM's intervention margins to ±15 per cent from±2.25 per cent. Only then did the currency market stabilize.

Crises are hardy perennials. Within one year of the EMS crises, a currency crisis exploded in Mexico, with currency jitters spreading around the Latin American region. In 1997, it was Asia's turn. A new episode of currency turbulences started in July of that year with the depreciation of the Thai baht. Within a few days the crisis had spread to Indonesia, Korea, Malaysia and the Philippines. Turmoil in the foreign exchange market heightened in 1998 with the Russian default and devaluation in August. The Russian crisis spread around the world with speculative attacks in economies as far apart as South Africa, Brazil and Hong Kong. Currency crises have continued to erupt in the new millennium, with Argentina's crisis in December 2001 including the largest foreign-debt default in history.

The numerous financial crises that have ravaged emerging markets as well as mature economies have fuelled a continuous interest in developing models to explain why speculative attacks occur. Models are even catalogued into three generations. The first-generation models focus on the fiscal and monetary causes of crises. These models were mostly developed to explain the crises in Latin America in the 1960s and 1970s. In these models, unsustainable money-financed fiscal deficits lead to a persistent loss of international reserves and ultimately to a currency crash (see, for example, Krugman, 1979).

The second-generation models aim at explaining the EMS crises of the early 1990s. These models focus on explaining why currency crises tend to happen in the midst of unemployment and loss of competitiveness. To explain these links, governments are modelled facing two targets: reducing inflation and keeping economic activity close to a given target. Fixed exchange rates may help in achieving the first goal but at the cost of a loss of competitiveness and a recession. With sticky prices, devaluations restore competitiveness and help in the elimination of unemployment, thus prompting the authorities to abandon the peg during recessions. Importantly, in this setting of counter-cyclical policies, the possibility of self-fulfilling crises becomes important, with even sustainable pegs being attacked and frequently broken (see, for example, Obstfeld, 1994).

The next wave of currency crises, the Mexican crisis in 1994 and the Asian crisis in 1997, fuelled a new variety of models – also known as third-generation models – which focus on moral hazard and imperfect information. The emphasis here has been on ‘excessive’ booms and busts in international lending and asset price bubbles. These models also link currency and banking crises, sometimes known as the ‘twin crises’ (Kaminsky and Reinhart, 1999). For example, Diaz-Alejandro (1985) and Velasco (1987) model difficulties in the banking sector as giving rise to a balance of payments crisis, arguing that, if central banks finance the bail-out of troubled financial institutions by printing money, we have the classical story of a currency crash prompted by excessive money creation. Within the same theme, McKinnon and Pill (1995) examines the role of capital flows in an economy with an unregulated banking sector with deposit insurance and moral hazard problems of the banks. Capital inflows in such an environment can lead to over-lending cycles with consumption booms, real exchange rate appreciations, exaggerated current account deficits, and booms (and later busts) in stocks and property markets. Importantly, the

excess lending during the boom makes banks more prone to a crisis when a recession unfolds. In turn, the fragile banking sector makes the task of defending the peg by hiking domestic interest rates more difficult and may lead to the eventual collapse of the domestic currency. Following the crisis in Argentina in 2001, the links between debt sustainability, sovereign defaults, and currency crises again attracted the attention of the economics profession. Finally, currency crises have also been linked to the erratic behaviour of international capital markets. For example, Calvo (1998) has brought to general attention the possibility of liquidity crises in emerging markets due to sudden reversals in capital flows, in large part triggered by developments in the world financial centres.

To summarize, all models suggest that currency crises erupt in fragile economies. Importantly, the three generations of models conclude that vulnerabilities come in different varieties. Still, the first attempts to study the vulnerabilities that precede crises have adopted ‘the one size fits all’ approach (see, for example, Frankel and Rose, 1996; and Kaminsky, 1998). That is, the regressions estimated to predict crises include all possible indicators of vulnerability. These indicators include those related to sovereign defaults, such as high foreign debt levels, or indicators related to fiscal crises, such as government deficits, or even indicators related to crises of financial excesses, such as stock and real estate market booms and busts. In all cases, researchers impose the same functional form on all observations. When some indicators are not robustly linked to all crises, they tend to be discarded even when they may be of key importance for a subgroup of crises. Naturally, these methods leave many crises unpredicted and, furthermore, cannot capture the evolving nature of currency crises.

The next step in the empirical analysis of crises should be centred on whether crises are of different varieties. The first attempt in this direction is in Kaminsky (2006). In this article, a different methodology is used to allow for *ex ante* unknown varieties of currency crises. To identify the possible multiple varieties of crises, regression tree analysis is applied. This technique allows us to search for an unknown number of varieties of crises and of tranquil times using multiple indicators. This technique was also applied to growth by Durlauf and Johnson (1995).

Interestingly, this method catalogues crises into six classes:

1. 1. *Crises with current account problems*. This variety is characterized by just one type of vulnerability, that of loss of competitiveness, that is, real exchange rate appreciations.
2. 2. *Crises of financial excesses*. The fragilities are associated with booms in financial markets. In particular, they are identified as crises that are preceded by the acceleration in the growth rate of domestic credit and other monetary aggregates.
3. 3. *Crises of sovereign debt problems*. These crises are characterized by fragilities associated with ‘unsustainable’ foreign debt.
4. 4. *Crises with fiscal deficits*. This variety is just related to expansionary fiscal policy.
5. 5. *Sudden-stop crises*. This type of crisis is only associated with reversals in capital flows triggered by sharp hikes in world interest rates, with no domestic vulnerabilities.
6. 6. *Self-fulfilling crises*. This class of crises is not associated with any evident vulnerability, domestic or external.

These estimations allow us to answer four important questions about crises.

1. 1. *Do crises occur in countries with sound fundamentals?* Even though this estimation allows for the identification of self-fulfilling crises (crises in economies with sound fundamentals), the results indicate that basically all crises are preceded by domestic or external vulnerabilities. Only four per cent of the crises are unrelated to economic fragilities.
2. 2. *How important are sudden reversals in capital flows in triggering crises?* While many have stressed that the erratic behaviour of international capital markets is the main culprit in emerging market currency crises, only two per cent of the crises in developing countries are just triggered by sudden-stop problems. While sudden-stop problems do occur, the reversals in capital flows mostly occur in the midst of multiple domestic vulnerabilities (see, Calvo, Izquierdo and Talvi, 2004).
3. 3. *Are crises different in emerging economies?* Crises in emerging markets are preceded by far more domestic vulnerabilities than those in industrial countries. Overall, 86 per cent of the crises in emerging economies are crises with multiple domestic vulnerabilities, while economic fragility characterizes only 50 per cent of the crises in mature markets.
4. 4. *Are some crises more costly than others?* It is a well-established fact that financial crises impose substantial costs on society. Many economists have emphasized the output losses associated with crises. But these are not the only costs of crises. In the aftermath of crises, most countries lose access to international capital markets, losing the ability to reduce the effect of adverse income shocks by borrowing in international capital markets. In most cases, countries have to run current account surpluses to pay back their debt. Finally, the magnitude of the speculative attack is itself important. For example, large depreciations may cause adverse balance sheet effects on firms and governments when their liabilities are denominated in foreign currencies. *Crises of financial excesses*, those also associated with banking crises – twin crisis episodes – are the costliest. Not only does the domestic currency depreciate the most, but also output losses are higher and the reversal of the current account deficit is attained via a dramatic fall in imports. In the aftermath of these crises, exports fail to grow even though the depreciations in this type of crises are massive. This evidence suggests that countries are even unable to attract trade credits to finance exports when their economies are mired in financial problems. In contrast, *self-fulfilling crises* and *sudden-stop crises* (but with no domestic vulnerabilities) have no adverse effects on the economies. Output (relative to trend) is unchanged or continues to grow in the aftermath of crises with no observed domestic fragility. In these crises, booming exports are at the heart of the recovery of the current account.

See Also

- currency crises models

Bibliography

Calvo, G. 1998. Capital flows and capital-market crises: the simple economics of sudden stops. *Journal of Applied Economics* 1, 35–54.

Calvo, G., Izquierdo, A. and Mejia, L. 2004. *On the empirics of sudden stops: the relevance of balance-sheet effects*. Working paper No. 10520. Cambridge, MA: NBER.

Diaz-Alejandro, C. 1985. Good-bye financial repression, hello financial crash. *Journal of Development Economics* 19, 1–24.

Durlauf, S. and Johnson, P. 1995. Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics* 10, 365–84.

Frankel, J. and Rose, A. 1996. Currency crises in emerging markets: an empirical treatment. *Journal of International Economics* 41, 351–66.

Kaminsky, G.L. 1998. Currency and banking crises: the early warnings of distress. International Finance Discussion Papers No. 629. Board of Governors of the Federal Reserve System.

Kaminsky, G.L. and Reinhart, C. 1999. The twin crises: the causes of banking and balance-of-payments problems. *American Economic Review* 89, 473–500.

Kaminsky, G.L. 2006. Currency crises: are they all the same? *Journal of International Money and Finance* 25, 503–27.

Krugman, P. 1979. A model of balance-of-payments crises. *Journal of Money, Credit, and Banking* 11, 311–25.

McKinnon, R.I. and Pill, H. 1995. Credible liberalizations and international capital flows: the ‘overborrowing syndrome’. In *Financial Deregulation and Integration in East Asia*, ed. T. Ito and A.O. Krueger. Chicago: University of Chicago Press.

Obstfeld, M. 1994. The logic of currency crises. *Cahiers Economiques et Monétaires* 43, 189–213.

Obstfeld, M. 1996. Models of currency crises with self-fulfilling features. *European Economic Review* 40, 1037–47.

Velasco, A. 1987. Financial and balance-of-payments crises. *Journal of Development Economics* 27, 263–83.

How to cite this article

Kaminsky, Graciela Laura. "currency crises." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_C000468> doi:10.1057/9780230226203.0350

The New Palgrave Dictionary of Economics Online

currency unions

Andrew K. Rose

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews currency unions, that is, groups of countries that use a common money. There are a large number of such monetary unions in both the industrial and the developing worlds. I review both the theoretical reasons why countries choose to belong to currency unions and the empirical performance of these unions.

Keywords

asymmetric shocks; automatic stabilizers; business cycles; capital mobility; counter-cyclical fiscal policy; currency unions; diversification; Economic and Monetary Union (EMU); exchange rate regimes; inflation; labour mobility; Latin Monetary Union; monetary unions *see* currency unions; Mundell, R.; optimal currency area; output volatility; progressive taxation; rigidities; Scandinavian Monetary Union; size of nations

Article

Currency unions (also known as monetary unions) are groups of countries that share a single money. Currency unions are unusual, since most countries have their own currency. For instance, the United States, Japan and the United Kingdom all have their own monies. But a reasonable number of countries participate in currency unions, and their importance is growing. In May 2005, 52 of the 184 IMF members participated in currency unions.

Currency unions present and past

Currency unions commonly come about when a small or poor country unilaterally adopts the money of a larger, richer ‘anchor’ country. For instance, a number of countries currently use the US dollar, including Panama, El Salvador, Ecuador, and a number of smaller countries and dependencies in the

Caribbean and Pacific. Swaziland, Lesotho and Namibia all use the South African rand. Both the Australian and New Zealand dollars are used by a number of countries in the Pacific; Liechtenstein uses the Swiss franc; and so forth. In the past, a number of countries have used the currency of their colonizer; over 50 countries and dependencies have used the British pound sterling at one time or another. Cases like this are known as official dollarization (unofficial dollarization occurs when the currency of a foreign country circulates widely but is not formally the national currency). In such cases, the small country essentially relinquishes its right to sovereign monetary policy. It loses its ability to independently influence its exchange and interest rates; these are determined by the anchor country, typically on the basis of the interests of the anchor.

There are also a number of multilateral currency unions between countries of more or less equal size and wealth. For instance, the East Caribbean dollar circulates in Anguilla, Antigua and Barbuda, Dominica, Grenada, Montserrat, Saint Kitts and Nevis, Saint Lucia, and Saint Vincent and the Grenadines. The Central Bank of the West African States circulates the Communauté française d'Afrique (CFA) franc in Benin, Burkina Faso, Côte d'Ivoire, Guinea-Bissau, Mali, Niger, Senegal, and Togo. The Bank of the Central African States circulates a slightly different CFA franc in Cameroon, the Central African Republic, Chad, Republic of Congo, Equatorial Guinea, and Gabon.

The largest and most important currency union is the Economic and Monetary Union of the European Union (EMU). EMU technically began on 1 January 1999, although the euro was physically introduced only three years later. Twelve countries are formally members of EMU: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Spain. (A number of smaller European territories and French dependencies also use the euro.) These countries jointly determine monetary policy for EMU through the international European Central Bank. The number of members in EMU is expected to grow with time, especially as countries that acceded to the European Union in 2004 become eligible for EMU entry. However, both Sweden and Denmark have rejected membership in referenda, and the euro remains unpopular in the UK.

While a number of currency unions currently exist, many have not survived. The Latin Monetary Union began in 1865 when France, Belgium, Italy and Switzerland (later joined by Greece, Romania, and others) adopted common regulations for their individual currencies to encourage the free international flow of money. This essentially amounted to a commitment to mint silver and gold coins to uniform specifications, but without other restrictions on monetary policy. The union effectively ended with the onset of the First World War. The war also ended the Scandinavian Monetary Union which Denmark, Norway, and Sweden began in 1873. The economic union between Belgium and Luxembourg that began in 1921 has been absorbed into EMU. Multilateral currency unions in East Africa, Central Africa, West Africa, South Asia, South-East Asia, and the Caribbean have also disappeared.

Theory: why should countries enter currency union?

Historically, most countries have had their own moneys. There seems to be a tight connection between national identity and national money; a country's money is a potent symbol of sovereignty. Still, some countries have entered into currency union. Why? Economists have theorized about the potential economic benefits of currency union which can, in certain circumstances, overwhelm the perceived political costs.

Like all other monetary regimes, currency unions are fully compatible with Robert Mundell's (1968)

celebrated ‘Trilemma’ or ‘Incompatible Trinity’. A country would like its monetary regime to deliver three desirable goals that turn out to be mutually exclusive: domestic monetary sovereignty, capital mobility, and exchange rate stability. Currently, large rich countries like the United States, Japan and the UK have domestic monetary sovereignty and open capital markets but have floating exchange rates. By way of contrast, members of a currency union essentially relinquish the first objective (monetary independence) in exchange for the latter benefits (capital mobility and stable exchange rates). Indeed, some economists think of currency unions as simply extreme forms of fixed exchange rates, with all the associated pros and cons. Countries inside currency union receive more microeconomic benefits than they would from a fixed exchange rate, since sharing a single money leads to deeper integration of real and financial markets. On the other hand, a country can devalue or float the exchange rate more easily than it can leave a currency union. Still, this is an unsatisfying theoretical approach the issue of currency unions. It does not address to the vital question: what is the optimal size of a currency union? If the right size for a currency union is not necessarily the country, how should we tackle the problem?

The theoretical analysis of currency unions began with a seminal paper by Mundell (1961). Mundell's analysis answered the question: what is the appropriate domain for a currency? Mundell briefly argued there are advantages to regions that use a common money. In particular, currency union facilitates international trade; a single medium of exchange reduces transactions costs, as does a common unit of account. However, a common currency can also cause problems in the dual presence of asymmetric shocks and nominal rigidities (in prices and wages). Suppose demand shifts from Western to Eastern goods. The increase in demand for Western output results in inflationary pressures there, while East goes into recession. Mundell argued that, if unemployed labour could move freely from East to relieve inflationary pressures in West, the two problems could be resolved simultaneously. However, in the absence of labour mobility, the asymmetric shock could be better handled by allowing the Western currency to appreciate. But in order for this to happen, both East and West must have their own monies! Mundell concluded that the optimal currency area was the area within which labour is mobile; regions of labour mobility should have their own currencies.

Two other classic contributions to the theory of optimal currency areas are worthy of note. McKinnon (1963) examined the effects of country size on currency unions; he concluded that smaller countries tend to be more open and have fewer nominal rigidities, making them better candidates for currency union. Kenen (1969) considered the effects of the economy's degree of diversification, and argued that more diversification resulted in fewer asymmetric shocks, and accordingly fewer benefits from national monetary policy.

The key focus of Mundell's theoretical optimum currency area framework – the adjustment to asymmetric shocks – has stood the test of time well. The ability of a region to respond to such shocks is viewed as a critical part of a sustainable and desirable currency union. Still, hardly anyone now takes the narrow specifics of Mundell's original article seriously. In particular, Mundell's conclusion that the optimum currency area is a region of labour mobility is no longer widely believed. The problem of asymmetric business cycles that Mundell described is intrinsically a problem of … business cycles. The costs of shifting labour are high almost everywhere in the world, which is why labour moves only slowly, even within countries with relatively flexible labour markets like the United States. Accordingly, most economists are uncomfortable thinking that labour could or should shift in response to the shocks and propagation mechanisms that cause business cycles. After all, the nominal rigidities that are responsible for business cycles do not last for ever. Thus, Mundell's idea of labour mobility is no longer

viewed as a viable adjustment mechanism. (This conclusion is tempered if one believes that real shocks cause business cycles without nominal rigidities.)

Still, there are other ways to share the risks of, or adjust to, asymmetric shocks, and much of the relevant work has incorporated these other mechanisms. Mundell originally ignored capital mobility. But private capital markets can, in principle, spread shocks internationally if investors diversify across regions or sectors. However, more attention has been paid to the public sector, since a federal system of taxes and transfers may be an efficient way to spread risks across regions. To continue with the East and West example, a progressive federal tax structure reduces inflationary Western pressures, and allows benefits to be paid to the unemployed in the East. Both regions suffer less macroeconomic volatility with such automatic stabilizers in place. The most controversial adjustment mechanism is counter-cyclical fiscal policy. In response to an asymmetric shock, regions that are free and capable of deploying discretionary fiscal policy can uses changes in taxes and government spending to respond to asymmetric shocks, even within the monetary confines of a currency union. More generally, mechanisms to handle asymmetric shocks are still an integral part of the theory of currency unions.

Mundell originally thought the great benefit of currency union was the facilitation of trade since money is a convenience that lowers transactions costs. But suppose that countries produce moneys of different qualities. Argentina has gone through five currencies since 1970; high Argentine inflation results in a low convenience value for Argentine money. Suppose Argentina decides to give up on a national money altogether and enter into a currency union with a foreign producer of higher-quality money: the United States, say. Argentina will surely experience different shocks from the United States, and these shocks have to be handled. Perhaps then Argentina should enter a currency union with a country with more similar shocks? The problem is that the most obvious contender, Brazil, also has a history of monetary incompetence. The larger point is that a low-quality domestic monetary authority increases a country's willingness to enter currency union, as does the availability of high-quality foreign money. Alesina and Barro (2002) provide an elegant model that incorporates such features. In their model, countries enter currency unions with neighbours in order to facilitate trade, so long as the neighbours possess monetary institutions of quality. Lower inflation and reduced transactions costs of trade provide gains, while the inability to respond to idiosyncratic asymmetric shocks generates losses.

Empirics: what do we know in practice about currency unions?

During the run-up to EMU, a considerable empirical literature developed that quantified different aspects of optimal currency areas. Much attention was paid to estimating the synchronization of business cycles for potential EMU candidates; Bayoumi and Eichengreen (1992) was the first important paper. The tradition has since been generalized to more countries by Alesina, Barro and Tenreyro (2002), who characterized co-movements in prices as well as output. Frankel and Rose (1998) showed that the intensity of trade had a strong positive effect on business cycle synchronization; that is, the optimum currency area criteria are jointly endogenous. If currency union lowers the transactions costs of trade and thus leads to an increase in trade, it may also thereby reduce the asymmetries in business cycles; areas that do not look like currency unions *ex ante* may do so *ex post*. Bayoumi and Eichengreen (1998) successfully link optimum currency area criteria (principally the asymmetry of business cycle shocks) to exchange rate volatility and intervention, and show that a number of features of the optimum currency

area theory appear in practice, even for countries not in currency unions.

Somewhat curiously, little work was done to analyse actual currency unions until around 2000. This is probably because the currency unions that preceded EMU consisted mostly of small or poor countries, which were viewed as irrelevant for EMU. But this gap in the literature implicitly allowed economists to focus their attention on the costs of currency union, which tend to be macroeconomic in nature (resulting from the absence of national monetary policy as a tool to stabilize business cycles). As Mundell clearly pointed out, there are also benefits from a currency union, mostly microeconomic in nature. Fewer monies mean lower transactions costs for trade, and thus higher welfare. An unresolved issue of importance is the size of the benefits that stem from currency union. There is evidence that currency unions have been associated with increased trade in goods, though its size is much disputed. Using data on pre-EMU currency unions (such as the CFA franc zone), Rose (2000) first estimated the effect of currency union on trade, and found it to result in an implausibly high tripling of trade. This finding and the intrinsic interest of EMU have resulted in a literature that has almost universally found smaller estimates, which are yet of considerable economic size. Rose and Stanley (2005) provide a quantitative survey that concludes that currency union increases trade by between 30 and 90 per cent. Engel and Rose (2002) examine other macroeconomic aspects of pre-EMU currency unions, and find that currency union members are more integrated than countries with their own monies, but less integrated than regions within a single country. Edwards and Magendzo (2003) compare inflation, output growth and output volatility in countries inside currency unions and those outside them, and find that currency unions have lower inflation and higher output volatility than countries with their own currencies.

Areas of ignorance

The impact of currency union on financial markets is not something that is currently well understood. Yet this is an area of great interest, since currency union might result in deeper financial integration – or it might not. It is clearly of concern to the British government, which has made the financial effects one of its five tests for EMU entry (see HM Treasury, 2003).

More generally, Europe's experiment with currency union is still young. It is simply too early to know whether EMU has resulted in substantial changes in the real economy, financial or labour markets, or political economy. As the data trickles in, most expect a continuing reassessment of currency unions in theory and especially practice.

See Also

- Mundell, Robert

I thank Steven Durlauf for helpful comments.

Bibliography

Alesina, A. and Barro, R. 2002. Currency unions. *Quarterly Journal of Economics* 117, 409–36.

Alesina, A., Barro, R. and Tenreyro, S. 2002. Optimal currency areas. In *NBER Macroeconomics Annual 2002*, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Bayoumi, T. and Eichengreen, B. 1992. Shocking aspects of European monetary unification. In *The Transition to Economic and Monetary Union in Europe*, ed. F. Torres and F. Giavazzi. New York: Cambridge University Press.

Bayoumi, T. and Eichengreen, B. 1998. Exchange rate volatility and intervention: implications of the theory of optimum currency areas. *Journal of International Economics* 45, 191–209.

Edwards, S. and Magendzo, I. 2003. A currency of one's own? An empirical investigation on dollarization and independent currency unions. Working Paper No. 9514. Cambridge, MA: NBER.

Engel, C. and Rose, A. 2002. Currency unions and international integration. *Journal of Money Credit and Banking* 34, 1067–89.

Frankel, J. and Rose, A. 1998. The endogeneity of the optimum currency area criteria. *Economic Journal* 108, 1009–25.

HM Treasury. 2003. *UK Membership of the Single Currency: EMU Studies*. Online. available at http://www.hm-treasury.gov.uk./documents/international_issues/the_euro/assessment/studies/euro_assess03_studindex.cfm, accessed 25 March 2006.

Kenen, P. 1969. The theory of optimum currency areas: an eclectic view. In *Monetary Problems of the International Economy*, ed. R. Mundell and A. Swoboda. Chicago: University of Chicago Press.

McKinnon, R. 1963. Optimum currency areas. *American Economic Review* 53, 717–24.

Mundell, R. 1961. A theory of optimum currency areas. *American Economic Review* 51, 657–65.

Mundell, R. 1968. *International Economics*. New York: Macmillan.

Rose, A. 2000. One money, one market: estimating the effect of common currencies on trade’ *Economic Policy* 30, 7–46.

Rose, A. and Stanley, T. 2005. A meta-analysis of the effect of common currencies on international trade. *Journal of Economic Surveys* 19, 347–65.

How to cite this article

Rose, Andrew K. "currency unions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000567> doi:10.1057/9780230226203.0352

The New Palgrave Dictionary of Economics Online

Currie, Lauchlin (1902 – 1993)

Roger Sandilands
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

At Harvard in the early 1930s Currie pioneered a monetary diagnosis of the 1929–32 collapse and placed blame on the Federal Reserve Board. As a prominent New Dealer at the Fed during 1934–9 he urged contra-cyclical monetary and fiscal activism. During 1939–45 he worked in Washington as President Roosevelt's economic adviser. After heading a World Bank mission to Colombia in 1949 he spent 40 years advising on national development there. He emphasized urban housing as a leading sector, based on an innovative housing finance system, and extended Allyn Young's ideas on macroeconomic increasing returns and endogenous growth.

Keywords

Bretton Woods conference; commercial loan theory of banking; credit; Currie, L.; endogenous growth; Federal Reserve System; Great Depression; Hansen, A.; income velocity of money; increasing returns; land tax; monetary and financial forces in the Great Depression; Plan of the Four Strategies (Colombia); quantity of money; reserve requirement; Rostow, W.; urban planning; Viner, J.; White, H.; Young, A.

Article

Lauchlin Currie was born on 8 October 1902 in West Dublin, Nova Scotia, and died in Bogotá, Colombia, on 23 December 1993 after an unusually long and varied career as an academic economist and top-level policy adviser. After two years at St Francis Xavier University, Nova Scotia, 1920–2, he moved to the London School of Economics (LSE), where his teachers included Edwin Cannan, Hugh Dalton, A. L. Bowley, R. H. Tawney and Harold Laski. In 1925 he obtained his BsC and moved to Harvard, where the chief inspiration for his Ph.D. thesis, ‘Bank Assets and Banking Theory’ (January 1931), was Allyn Abbott Young. However, when Young moved to the LSE in 1927 his formal supervisor was John H. Williams.
He remained at Harvard until 1934 as teaching assistant to Williams, Ralph Hawtrey and Joseph

Schumpeter. His Ph.D. thesis attacked the ‘commercial loan’ or ‘needs of trade’ theory of banking by showing that it was not only unsound in theory but had been more honoured in the breach than the observance – until its disastrous influence on monetary policy in the late 1920s and early 1930s.

In a January 1932 memorandum, Currie, Harry Dexter White and Paul Theodore Ellsworth presented a radical anti-depression programme (see Laidler and Sandilands, 2002). In keeping with their explanation of the contraction as due to a collapsing money supply, they urged vigorous open-market operations and deficit spending financed by money creation. This memorandum was part of an early Harvard influence (through Young, Hawtrey, Williams and Currie; see Laidler, 1999) on what had been claimed as a unique Chicago monetary tradition.

In Currie (1933a) he showed the hopeless confusion that resulted from the ambiguity of the word ‘credit’. He stressed control over the quantity of money (defined as cash plus demand deposits, for which there had been no estimates until Currie published a series in 1934) rather than the quantity or quality of credit or loans. He also computed the first estimate of the income velocity of money in the United States (Currie, 1933b), with an explanation of its cyclical variations.

His ‘The Failure of Monetary Policy to Prevent the Depression of 1929–32’ (1934a) fully anticipated Milton Friedman and Anna Schwartz's (1963) diagnosis of this period. He argued that apart from the stock market there were none of the traditional signs of a boom in the 1920s. Tight monetary policies had been ineffectual in checking the rise in stock prices but only too effective in contributing to the decline in building activity and the pressure on foreign countries that preceded the Depression.

He also demonstrated the perverse elasticity of money in the business cycle due to differences in reserve requirements for different classes of bank and bank deposit (1934b). In the face of the banks' reserve losses in 1929–32 and their abhorrence of heavy indebtedness to the reserve banks, the administration's policy was ‘one of almost complete passivity and quiescence’, so the self-generating forces of the Depression continued unchecked.

In 1934 Jacob Viner recruited him to the ‘freshman brain trust’ at the US Treasury where he developed a blueprint for a system of 100 per cent reserves against demand deposits, to break the link between the lending and the creation of money and to strengthen central bank control (see Phillips, 1995). Later that year Marriner Eccles, the new governor of the Federal Reserve Board, hired Currie as his top adviser, from 1934 to 1939. (Many of his memoranda to Eccles are published in Sandilands, 2004.)

At the Fed Currie drafted what became the 1935 Banking Act that gave the Fed increased powers to raise reserve requirements. In 1936–7 these powers were used, ‘as a precautionary measure’, to reduce the huge build-up of banks' excess reserves. This has been widely blamed for the sharp recession of 1937–8, a view Currie consistently rejected (1938). Instead, he invoked his newly constructed ‘net federal income-creating expenditure series’ (1935; and see Sweezy, 1972) to show the strategic role of fiscal policy in complementing monetary policy to revive an acutely depressed economy. In November 1937 he had a four-hour meeting with President Roosevelt to explain that the recession was due to sharp fiscal contraction and that balancing the budget was not the way to restore business confidence. He insisted on the need for better coordination of monetary and fiscal policy. In May 1939 the rationale for this was explained in theoretical and statistical detail by Currie and Alvin Hansen (respectively ‘Mr Inside’ and ‘Mr Outside’, according to Tobin, 1976), in joint testimony before the Temporary National Economic Committee.

From 1939 to 1945, Currie was President Roosevelt's special adviser on economic affairs in the White

House. He was also in charge of lend-lease to China, 1941–3, and ran the Foreign Economic Administration, 1943–4. In early 1945 he headed a tripartite (United States, British and French) mission to Bern to persuade the Swiss to freeze Nazi bank balances and stop shipments of German supplies through Switzerland to the Italian front. He was also closely involved in loan negotiations with British and Soviet allies and in preparations for the 1944 Bretton Woods conference (staged primarily by his friend Harry White).

After the war it was alleged by Elizabeth Bentley, an ex-Soviet agent, that Currie and White had participated in Soviet espionage. Though she had never met them herself, she claimed they had passed information to other Washington economists who were abetting her own espionage, and that they probably knew this. White and Currie were heavily involved in official wartime cooperation with the Soviets, but Bentley put a sinister interpretation on these activities. They appeared together before the House Committee on Un-American Activities in August 1948 to rebut Bentley's charges. Their testimony satisfied the Committee at that time, though the strain contributed to the fatal heart attack that White suffered three days after the hearing.

No charges were laid against Currie, and in 1949 he headed a major World Bank survey of Colombia. In 1950 the Colombians invited him to return to Bogotá, where he remained for most of the next 40 years as a top presidential adviser. He has been falsely accused of fleeing the United States to avoid charges of disloyalty. In fact in December 1952 he was a witness before a grand jury in New York investigating Owen Lattimore's role in the famous *Amerasia* case that involved the publication of secret State Department documents by that magazine, though his next visit to the United States was not until 1961 when he had a meeting in the White House with Walt Rostow, then President Kennedy's National Security Adviser, to discuss a development plan for Colombia.

By that time Currie had assumed Colombian citizenship (personally conferred on him by President Alberto Lleras in 1958), partly because in 1954 the US government had refused to renew his passport, ostensibly because he was only a naturalized US citizen and was now residing abroad. However, the reality was probably connected with the then secret ‘Venona’ project that had deciphered wartime Soviet cables that mentioned Currie. The related cases of Currie and White are discussed in Sandilands (2000) and Boughton and Sandilands (2003), where it is shown that the evidence against them is far from conclusive. After reading the latter paper, Major-General Julius Kobyakov, deputy director of the KGB's American desk in the late 1980s, wrote to the present writer on 22 December 2003 to confirm our conclusions. After extensive archival research on Soviet intelligence in the 1930s and 1940s he found that

> there was nothing in [Currie's] file to suggest that he had ever wittingly collaborated with the Soviet intelligence… However, in the spirit of machismo, many people claimed that we had an ‘agent’ in the White House. Among the members of my profession there is a sacramental question: ‘Does he know that he is our agent?’ There is very strong indication that neither Currie nor White knew that.

There were two breaks to Currie's advisory and academic work in Colombia: during a military dictatorship, 1953–8, he retired to develop a prize-winning herd of Holstein cattle; and from 1966 to 1971 he was a professor at Michigan State (1966), Simon Fraser (1967–8 and 1969–71), Glasgow (1968–9), and Oxford (1969) universities. He returned permanently to Colombia in 1971 at the behest of

President Misael Pastrana to prepare a national plan of development known as the Plan of the Four Strategies, with a focus on urban housing and export diversification. The plan was implemented and the institutions that were established in support of the plan played a major role in accelerating Colombia's urbanization.

He remained as chief economist at the National Planning Department for ten years, 1971–81, followed by 12 years at the Colombian Institute of Savings and Housing until his death in 1993. There he defended the unique index-linked housing finance system (based on ‘units of constant purchasing power’ for both savers and borrowers) that he had established in 1972. The system thus continued to boost Colombia's growth rate and urban employment opportunities year by year. Currie was also a top adviser on urban planning, and played a major part in the first United Nations Habitat conference in Vancouver in 1976. His ‘cities-within-the-city’ urban design and financing proposals (including the public recapture of land's socially created ‘valorización’, or ‘unearned land value increments’, as cities grow) were elaborated in *Taming the Megalopolis* (1976). To the time of his death he was a regular teacher at the National University of Colombia, Javeriana University, and the University of the Andes, and continued to publish widely (a comprehensive bibliography is in Sandilands, 1990, reviewed by Charles Kindleberger, 1991). His writings and policy advice were heavily influenced by his old Harvard mentor, Allyn Young. Notable is his posthumous (1997) paper that offers a unique macroeconomic interpretation of Youngian increasing returns and the endogenous nature of self-sustaining growth.

See Also

- development economics
- endogenous growth theory
- Federal Reserve System
- Great Depression, monetary and financial forces in
- Hansen, Alvin
- housing supply
- monetary policy, history of
- urbanization
- Viner, Jacob
- Young, Allyn Abbott

Selected works

1931. Bank assets and banking theory. Ph.D. thesis, Harvard University.

1932. (With P. Ellsworth and H. White.) Memorandum on anti-depression policy. *History of Political Economy* 34(2002), 533–52.

1933a. The treatment of credit in contemporary monetary theory. *Journal of Political Economy* 41, 509–25.

1933b. Money, gold and incomes in the United States, 1921–32. *Quarterly Journal of Economics* 48, 77–95.

1934a. The failure of monetary policy to prevent the Depression of 1929–32. *Journal of Political Economy* 42, 145–77. Reprinted in *Landmarks in Political Economy*, ed. E. Hamilton, H. Johnson and A. Rees. Chicago: University of Chicago Press, 1962.

1934b. *The Supply and Control of Money in the United States*. Cambridge, MA: Harvard University Press.

1935. Comments and observations on ‘Federal Income-Increasing Expenditures, 1933–35’. *History of Political Economy* 10(1978), 507–48.

1938. Causes of the recession. *History of Political Economy* 12(1980), 303–35.

1976. *Taming the Megalopolis: A Design for Urban Growth*. Oxford: Pergamon Press.

1997. Implications of an endogenous theory of growth in Allyn Young's macroeconomic concept of increasing returns. *History of Political Economy* 29, 414–43.

Bibliography

Boughton, J. and Sandilands, R. 2003. Politics and the attack on FDR's economists: from Grand Alliance to Cold War. *Intelligence and National Security* 18(3), 73–99.

Friedman, M. and Schwartz, A. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press.

Kindleberger, C. 1991. Review of Roger J. Sandilands, *The Life and Political Economy of Lauchlin Currie. Journal of Political Economy* 99, 1119–22.

Laidler, D. 1999. *Fabricating the Keynesian Revolution*. Cambridge: Cambridge University Press.

Laidler, D. and Sandilands, R. 2002. An early Harvard *Memorandum* on anti-depression policies. *History of Political Economy* 34, 515–32.

Phillips, J. 1995. *The Chicago Plan and New Deal Banking Reform*. Armonk, NY: M.E. Sharpe.

Sandilands, R. 1990. *The Life and Political Economy of Lauchlin Currie: New Dealer, Presidential Adviser, and Development Economist*. Durham, NC: Duke University Press.

Sandilands, R. 2000. Guilt by association? Lauchlin Currie's alleged involvement with Washington economists in Soviet espionage. *History of Political Economy* 32, 473–515.

Sandilands, R., ed. 2004. New light on Lauchlin Currie's monetary economics in the New Deal and beyond. *Special issue of Journal of Economic Studies* 31(3/4).

Sweezy, A. 1972. The Keynesians and government policy, 1933–1939. *American Economic Review* 62, 116–24.

Tobin, J. 1976. Hansen and public policy. *Quarterly Journal of Economics* 90, 32–7.

How to cite this article

Sandilands, Roger. "Currie, Lauchlin (1902–1993)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000530> doi:10.1057/9780230226203.0353

The New Palgrave Dictionary of Economics Online

cyclical markups

Julio J. Rotemberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article first shows that countercyclical variations in the ratios of prices to marginal cost (markups) can cause pro-cyclical fluctuations in the demand for labour at a given real wage and thus induce fluctuations in economic activity that look like business cycles. It then discusses methods for measuring cyclical movements in markups and shows that several types of evidence suggest that these are counter-cyclical. Lastly, it discusses economic mechanisms that can explain these counter-cyclical markup movements.

Keywords

Cobb–Douglas functions; cyclical markups; elasticity; imperfect competition; increasing returns; inventory investment; labour productivity; labour supply; leisure; limit pricing; real business cycles; sticky prices; wealth effects

Article

Firms that have increasing returns to scale, that produce differentiated products, or that are part of a small oligopoly can generally be expected to set a price above marginal cost. In so far as a firm's ratio of price to marginal cost is larger than one, there is no particular reason to suppose that this ratio, or markup, will stay constant when overall economic conditions change. Indeed, different models of imperfect competition have different predictions concerning how this markup should vary as aggregate income and activity expands and contracts. Thus, an analysis of whether markups rise when aggregate activity rises or whether they rise when aggregate activity declines provides a useful lens for determining which theories of firm behaviour have more validity.
Markup variations are also of central importance for macroeconomics. One of the central questions for macroeconomics is why the economy expands and contracts at cyclical frequencies in the first place, and cyclical movements in markups are potentially an important nexus that allows such fluctuations to occur.

When a single firm (or industry) raises the ratio of its price to its marginal cost, one expects its relative price to rise so that the quantity it sells falls. However, when every firm in the economy tries to raise its price relative to its marginal cost, relative prices need not be affected.

When every firm raises its markup two important consequences follow. The first is that real marginal cost, which can be defined as nominal marginal cost divided by the typical price charged by firms, must fall. Thus, the question of whether markups are countercyclical is the same as the question of whether real marginal costs are procyclical. The second consequence of all firms varying their markups at the same time is that the aggregate demand for labour changes. To see this, notice that nominal marginal cost is equal to the nominal wage divided by the marginal product of labour. Thus, a generalized increase in markups means that prices must rise relative to nominal wages if employment is to remain at a level that keeps the marginal product of labour constant. Alternatively, firm are willing to pay the same real wage only if the marginal product of labour rises, and this requires that employment fall if labour is subject to diminishing returns. In either way of seeing this change, the demand for labour at any given wage falls.

The role of markup changes in economic fluctuations

The capacity of markup changes to generate changes in aggregate labour demand is important because several pieces of evidence suggest that short-run business fluctuations are the result of changes in the demand for labour. That the willingness of firms to hire labour at any given real wage increases in economic expansions is suggested first of all by the tendency of real wages to increase when the economy expands. As shown by Bils (1985), this tendency is particularly strong when one looks at the wages of individuals (as opposed to looking at average wages paid to all workers). Moreover, as emphasized by Bils (1987), firms tend to use more overtime hours in economic booms, and firms are legally obliged to pay higher hourly wages for these overtime hours. When combined with the pro-cyclicality of real wages, other pieces of evidence also suggest that labour demand is higher in booms. In booms, both the unemployment rate and the fraction of the unemployed who have been unemployed for longer than five weeks tend to be lower (both of which suggest that finding jobs is easier) and that the number of help-wanted advertisements is larger (suggesting that it is more difficult for firms to find workers even as they pay them higher wages).

The real business cycle literature stresses a different source of labour demand movements: namely, exogenous changes in the productivity of the typical firm. This hypothesis has the advantage that it explains in a straightforward fashion why labour productivity is somewhat pro-cyclical. However, as discussed below, movements in markups lead to pro-cyclical productivity under a variety of plausible assumptions. In this regard, a clear advantage of the view that markup movements are responsible for important labour demand movements is that labour productivity and real wages rise together with output also when output increases appear to be due to non-technological factors such as increases in military spending, expansionary monetary policy or reductions in the price of oil. (Evidence of these conditional correlations of productivity and output can be found in Hall, 1988.)

Relative to markup variations, exogenous short-run changes in technical progress have another disadvantage as sources of cyclical fluctuations. This is that technical progress not only increases the willingness of firms to hire workers but also reduces the willingness of workers to work at any given wage. These contractionary movements in labour supply are the result of ‘wealth effects’: technical

progress makes people richer and thus induces them to consume both more goods and more leisure. These effects are particularly large if technical progress is somewhat permanent, as tends to be true with actual examples of such progress. These reductions in labour supply imply that shocks to technical progress have only small expansionary effects on employment. By contrast, reductions in markups induce only modest wealth effects, so employment responds more strongly to the resulting increases in labour demand.

These conceptual benefits of countercyclical markups raise the question of whether markups do indeed rise in economic contractions and fall in booms. To discuss this, it is worth starting with the case where the value added production function takes the Cobb–Douglas form. With capital essentially fixed in the short run, this implies that aggregate value added Y is equal to the labour input H to the power α. The marginal product of labour is then equal to α times the average product of labor *Y/H* The ratio of marginal cost to price is then the wage divided by both the marginal product of labour and the price, so that it is proportional to the labour share in value added (or unit labour cost) *WH/PY*.

Measuring markup variations

If aggregate data are used, the labour share in value added is not a very cyclical variable. Labour productivity *Y/H* tends to rise mildly in expansions, as does the average real wage – though the size of these effects depends on how one measures economic expansions. Because cyclical productivity changes are slightly larger than the corresponding average changes in real wages, the labour share has a modest tendency to fall in expansions. If the labour share were seen as equal to the inverse of the markup (as implied by the Cobb–Douglas assumptions), markups would be pro-cyclical and actually dampen cyclical fluctuations.

As suggested in the survey by Rotemberg and Woodford (1999), this Cobb–Douglas case is a good baseline, but a number of corrections to the resulting measure of the markup immediately suggest themselves, and these tend to make measured markups more counter-cyclical. The first of these is that, as already alluded to above, what matters for marginal cost is not the average wage but the marginal wage for an additional hour of work. The average wage is dragged down in booms by the absorption into employment of many relatively low-wage workers who are not employed in recessions. If these workers are less productive, their wage per effective unit of labour input may actually be relatively large. Whatever the case, individual workers who remain employed do see their wages rise more substantially, as emphasized by Bils (1987). Admittedly, these wage increases are concentrated among workers who change jobs, and the increases in the ‘straight-time’ wages of people who stay in the same job are more modest. The marginal hour of work, on the other hand, is more likely to be an overtime hour in booms, and this is probably the most important reason for believing that the marginal hour of labour is more expensive then.

It also seems important to correct the way the Cobb–Douglas approach measures the marginal product of labour. According to this functional form. the marginal product of labour is simply proportional to the average product of labour. Given that the average product of labour actually rises slightly in booms, this functional form essentially requires that the economy become ‘more productive’ in booms, perhaps as a result of increased technical progress.

The tendency of labour productivity to be pro-cyclical can be interpreted in two rather different ways,

both of which have a direct bearing on calculations of the cyclical properties of the marginal product of labour. The first is that firms are subject to increasing returns to scale. The simplest functional form that captures this supposes that there are fixed costs, that is, that some of their inputs are ‘overhead’ inputs that are required to produce even a minuscule positive quantity of output for sale. Suppose for example, that $\bar{H}$ units of labour are overhead units so that output continues to be given by the Cobb–Douglas form but is now proportional to $(H-\bar{H})$ to the power α. The marginal product of labour is then proportional to the ratio $Y/(H-\bar{H})$. In booms, the percentage increase in $H-\bar{H}$ obviously exceeds the percentage by which *H* rises so that $Y/(H-\bar{H})$ falls by more than *Y/H*. This means that for $\bar{H}$ sufficiently large, the marginal product of labour falls, marginal costs rise and measured markups fall. Assuming that some of the labour input takes this overhead form can thus easily lead to the inference that markups are indeed counter-cyclical.

A second possible reason for the observation that the average product of labour is pro-cyclical is that firms do not fully utilize all their labour in recessions. They ‘hoard’ labour to avoid having to incur hiring and training costs when economic activity recovers. This raises two important questions. The first is whether workers produce something else other than measured output when they are being hoarded. The second is whether the firm needs to pay them less when their GDP-producing effort is lower. Given that real wages are only slightly pro-cyclical, it is probably more realistic to suppose that the cost of an hour of labour services to the firm is the same whether the worker incurs effort (and produces) or not. Particularly if the workers are not producing much unmeasured output in recessions, this implies that marginal cost in recessions is considerable smaller than is implied by *H/Y*. Real marginal cost is more pro-cyclical than *WH/PY* and markups are more counter-cyclical. One attractive feature of this explanation for pro-cyclical labour productivity is that it is very compatible with the idea that markups are counter-cyclical. Firms are willing to keep workers idle in recessions even though marginal cost is extremely low precisely because they are keeping their prices high relative to marginal cost.

There are two additional types of evidence suggesting that markups are relatively low in booms and high in recessions. The first comes from the behaviour of intermediate inputs relative to final goods. A crude view of materials is that these are used in fixed proportions relative to the gross output of final goods. However, Basu (1995) shows that the ratio of materials to final goods tends to rise when the economy expands. If the material intensity of output is a choice variable, the ratio of marginal cost to price must also equal the real price of materials divided by the marginal product of materials. It is reasonable to suppose with Basu (1995) that the marginal product of materials diminishes as the level of materials inputs rises. With constant returns, the increase in the ratio of materials to output in booms thus implies that real marginal cost is pro-cyclical even if the price of materials relative to final output were constant. In fact, Murphy, Shleifer and Vishny (1989) show that that prices of more processed goods tend to fall relative to prices of less processed goods in economic expansions, and this too indicates a tendency of price to fall relative to marginal cost during booms.

The second additional source of evidence comes from the behaviour of inventories. Inventories rise in booms but, as stressed by Bils and Kahn (2000), they rise by less in percentage terms than sales. At the same time, long-run growth in sales does tend to be associated with equiproportonate increases in inventories in the industries they consider. In addition, they discuss cross-sectional evidence that shows that, within industries, the inventory–sales ratios of products with large sales are not smaller than the inventory-sales ratios for products with low sales. This suggests that there is something special about the

decline in inventory–sales ratios that is observed in booms. It suggests, in particular, that conditions in booms lead firms to economize on inventory holding. As Bils and Kahn (2000) argue, the evidence seems most consistent with the idea that firms keep their inventories relatively low in booms because real marginal cost is relatively high.

Theories of cyclical markup variations

A considerable body of evidence, then, seems consistent with counter-cyclical markups, and suggests that countercyclical markups might be central to aggregate fluctuations because they rationalize the changes in employment that characterize such fluctuations. The question that remains is why markups should vary cyclically. There are basically five types of models that explain these movements in markups. These are: models of variable demand elasticity, models of variable entry, models of sticky prices, models of investment in market share and models of implicit collusion.

In a monopolistically competitive setting, markups are equal to the elasticity of demand over the the elasticity of demand minus 1. Increases in the elasticity of demand thus lower markups (towards the competitive level of 1) and could thus be a source of business expansions. This still leaves the question of why the elasticity of demand facing the typical firm should vary over time. One possibility is that the proportion of demand that comes from highly elastic customers rises in booms. Gali (1994) obtains such composition effects under the supposition that investment is more price sensitive than consumption. Ravn, Schmidt-Rohe and Uribe (2004) obtain a related effect by supposing that people have formed a ‘habit’ for at least a fraction of past purchases, and the elasticity of demand for these habitual purchases is negligible relative to the elasticity of demand for non-habitual ones. As consumption rises in economic expansions, more of the purchases are non-habitual so that the elasticity of demand is higher and markups have to be correspondingly lower.

Devereux, Head and Lapham (1996) show that changes in demand induced, for example, by changes in government purchases lead new firms to enter existing industries. Entry of new firms is indeed quite pro-cyclical. Such entry can, in turn, make each firm's perceived elasticity of demand higher (because they fear more competitors). Thus variable entry can be seen as a reason for changes in elasticities that lead to counter-cyclical markups. Even if the expansion in the number of firms that takes place in booms is seen as too small for this effect to be large, the potential for increases in entry may lead incumbents to keep their prices low to avert the creation of an even larger number of new firms. This limit pricing might also be able to rationalize counter-cyclical markups.

Sticky prices, which are widely assumed in new Keynesian macroeconomics, probably provide the most straightforward model of counter-cyclical markups. Firms that keep their prices constant when demand increases (as a result of expansionary government policy, for example) will generally see their marginal costs rise both because of diminishing returns and because of increases in the costs of factor inputs. Thus, keeping their prices relatively constant will lead them to have lower markups. The argument that sticky prices derive their influence on the economy from their consequences for variable markups is presented in more detail in Kimball (1995).

If customers who have already purchased a good have relatively inelastic demand, keeping price low is like an investment activity for the firm. It encourages new customers (those whose demand is elastic) to become addicted. Changes in economic conditions can lead firms to desire to either increase or decrease

these investments. Increases in interest rates in particular might lead firms to wish to reduce these investments, at least temporarily. Chevalier and Scharfstein (1996) provide evidence that the cash condition of firms plays a large role in these investments as well. They show that recessions have a disproportionate effect on the pricing of cash-strapped firms, who turn out to be more eager to raise prices and thereby reduce their investment in market share.

Lastly, Rotemberg and Saloner (1986) have emphasized that high prices may be more difficult to sustain for implicitly collusive oligopolists in economic expansions. When current sales are high, each firm perceives a greater benefit from undercutting the implicit agreement because it can thereby secure even higher sales. To prevent this, the oligopolists must lower their markups of price relative to marginal cost. Some cross-sectional evidence suggests that markups are indeed more counter-cyclical in more concentrated sectors, as a theory that applies only to implicitly collusive oligopolists suggests. As shown by Rotemberg and Woodford (1992), the model can be embedded in a general equilibrium structure so that increases in government purchases raise output together with real wages. The increased rate of interest induced by additional government purchases lowers the present value of the future benefits from cooperation. It thus forces oligopolies to be less ambitious in the profits that they seek from current prices, so that markups fall and labour demand rises.

See Also

- microfoundations
- new Keynesian macroeconomics

Bibliography

Bils, M.J. 1985. Real wages over the business cycle: evidence from panel data. *Journal of Political Economy* 93, 666–89.

Bils, M.J. 1987. The cyclical behavior of marginal cost and price. *American Economic Review* 77, 838–57.

Bils, M. and Kahn, J.A. 2000. What inventory behavior tells us about business cycles. *American Economic Review* 90, 458–81.

Basu, S. 1995. Intermediate inputs and business cycles: implications for productivity and welfare. *American Economic Review* 85, 512–31.

Chevalier, J.A. and Scharfstein, D.S. 1996. Capital-market imperfections and countercyclical markups: theory and evidence. *American Economic Review* 86, 703–25.

Devereux, M.B., Head, A.C. and Lapham, B.J. 1996. Monopolistic competition, increasing returns, and the effects of government spending. *Journal of Money, Credit, and Banking* 28, 233–54.

Gali, J. 1994. Monopolistic competition, business cycles, and the competition of aggregate demand. *Journal of Economic Theory* 63, 73–96.

Hall, R.E. 1988. The relation between price and marginal cost in U.S. industry. *Journal of Political Economy* 96, 921–47.

Kimball, M.S. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27, 1241–77.

Murphy, K.M., Shleifer, A. and Vishny, R.W. 1989. Building blocks of market clearing business cycle models. In *NBER Macroeconomics Annual*, ed. O.J. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Ravn, M., Schmitt-Grohe, S. and Uribe, M. 2004. Deep habits. Working Paper No. 10261. Cambridge, MA: NBER.

Rotemberg, J.J. and Saloner, G. 1986. A supergame-theoretic model of price wars during booms. *American Economic Review* 76, 390–407.

Rotemberg, J.J. and Woodford, M. 1992. Oligopolistic pricing and the effects of aggregate demand on economic activity. *Journal of Political Economy* 100, 1153–207.

Rotemberg, J.J. and Woodford, M. 1999. The cyclical behavior of prices and costs. In *Handbook of Macroeconomics*, vol. 1B, ed. J.B. Taylor and M. Woodford. Amsterdam; New York and Oxford: North-Holland.

How to cite this article

Rotemberg, Julio J. "cyclical markups." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000585> doi:10.1057/9780230226203.0354

The New Palgrave Dictionary of Economics Online

Dalton, Edward Hugh John Neale (1887 – 1962)

Alan Peacock

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank of England; Bernoulli's formula; bonds; cheap money; Dalton, E. H. J.; public finance; redistribution of income and wealth; taxation

Article

British fiscal economist and prominent Labour politician, Hugh Dalton was a student of A.C. Pigou and J.M. Keynes. His main professional interest was in the use of taxation as an instrument for the redistribution of income and wealth, an interest inspired by Pigou's teaching and by his revulsion at the contrast between the sufferings inflicted on younger generations by the First World War and the material gains of those who financed or profited from the war itself. (Dalton spent four years on military service in France and Italy and lost several close friends, including the poet Rupert Brooke.) His main contribution was to investigate the properties of a modification of Bernoulli's formula $dw = dw / x$ where *w*=economic welfare and *x*=income but in which equal increases in welfare should correspond to more than proportionate increases in income, a condition satisfied by Dalton's formula $dw = dx / x^2$ so that $w = c - 1 / x$ where *c* is a constant. Using this formula he concluded that economic welfare would be improved by transfers from rich to poor (Dalton, 1935), a proposition that has excited the interest of ‘modern’ public finance theorists of the neo-utilitarian school (see Fishburn and Willig, 1984). He elaborated his ideas in several works including his highly successful standard text *Principles of Public Finance* and in his lectures as Reader in Economics at the London School of Economics (1923–36). There he was responsible for teaching and for recommending Lionel Robbins to be Professor of Economics, a typical example of his desire not only to ‘corrupt the young’ (as he termed it) but also to promote the interests even of those with whom he disagreed.

Dalton combined teaching with a political career throughout the 1920s and 1930s, rising to political eminence as a member of Churchill's coalition government during the Second World War. As Minister of Economic Warfare he was responsible for setting up the famous sabotage team, the Special Operations Executive (SOE). Later as President of the Board of Trade he formulated plans for post-war distribution of industry designed to prevent mass unemployment. In the Attlee Labour government of

1945 he reached the pinnacle of his political career as Chancellor of the Exchequer, one of his first acts being to nationalize the Bank of England. His famous attempt to drive down interest rates through a cheap money policy in order to float off an issue of Treasury stock at 2.5 per cent is a classic example of the failure of even an experienced and able economist to understand that, other than in the short run, governments can control either the price or the supply of bonds but not both.

Selected works

1923. *Principles of Public Finance*. London: George Routledge & Sons.

1935. *The Inequality of Incomes*. 4th Impression, London: George Routledge & Sons, especially the Appendix.

Bibliography

Davenport, N. 1961–70. Hugh Dalton. In *Dictionary of National Biography*, Oxford: Oxford University Press.

Fishburn, P.C. and Willig, R.D. 1984. Transfer principles in income redistribution. *Journal of Public Economics* 25, 323–8.

Pimlott, B. 1985. *Hugh Dalton*. London: Jonathan Cape.

How to cite this article

Peacock, Alan. "Dalton, Edward Hugh John Neale (1887–1962)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000003>
doi:10.1057/9780230226203.0355

The New Palgrave Dictionary of Economics Online

Dantzig, George B. (1914 – 2005)

Richard W. Cottle, B. Curtis Eaves and Mukund N. Thapa

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

George Dantzig is known as ‘father of linear programming’ and ‘inventor of the simplex method’. This biographical sketch traces the high points of George Dantzig's professional life and scholarly achievements. The discussion covers his graduate student years, his wartime service at the US Air Force's Statistical Control Division, his post-war creativity while serving as a mathematical advisor at the US Air Force Comptroller's Office and as a research mathematician at the RAND Corporation, his distinguished career in academia – at UC Berkeley and later at Stanford University – and finally as an emeritus professor of operations research.

Keywords

complementarity; computational complexity; convex programming; convexity; Dantzig, G. B.; decomposition principle; degeneracy; distributed computation; Hurwicz, L.; integer programming; interior point methods; Kantorovich, L.V.; Koopmans, T.C.; Lagrangian function; Leontief, W. W.; linear programming; logarithmic barrier method; mathematical programming; Neyman, J.; nonlinear programming; operations research; simplex method for solving linear programs; stochastic programming with recourse; von Neumann, J.; Wood, M.K.

Article

George Dantzig is known as the ‘father of linear programming’ and the ‘inventor of the simplex method’. Employed at the Pentagon (the US government's defence establishment) in 1947 and motivated to ‘mechanize’ programming in large time-staged planning problems, George Dantzig gave a general statement of what is now known as a linear program, and invented an algorithm, the simplex method, for solving such optimization problems. By the force of Dantzig's theory, algorithms, practice, and professional interaction, linear programming flourished. Linear programming has had an impact on economics, engineering, statistics, finance, transportation, manufacturing, management, and

mathematics and computer science, among other fields. The list of industrial activities whose practice is affected by linear programming is very long.

Over the subsequent half century, Dantzig remained a major contributor to the subject of linear programming as researcher, practitioner, teacher, mentor, and leader. The impact of linear programming and extensions on theory, business, medicine, government, the military, all in the broadest sense, is now hard to overstate. In the words of the editors of the Society for Industrial and Applied Mathematics: ‘In terms of widespread application, Dantzig's algorithm is one of the most successful of all time: linear programming dominates the world of industry, where economic survival depends on the ability to optimize within budgetary and other constraints …’ (quoted in Dongarra and Sullivan, 2000).

There were some significant contributions to what became linear programming prior to Dantzig's work. In their time, however, these results were not applied, linked together, or continued. In fact, they were nearly lost, perhaps because the prevailing historical setting was not favourable. As these contributions have been recognized, they have been drawn into the history of linear programming.

A linear program defined

In mathematical terms, a linear program is most simply stated as the problem of minimizing a multivariate linear function constrained by linear inequalities. Dantzig's first formulation of a linear program was the equivalent problem of minimizing a linear function over non-negative variables constrained by linear (material balance) equations. In matrix notation such a linear program is:

$$\underset{x,z}{\text{Minimize}}\ c^T x = zLP(A, b, c): \quad \text{subject to } Ax = b,\ x \geq 0.$$

Here the given data are the $m \times n$ matrix A and vectors b and c; the unknowns to be determined are the objective scalar value z and the decision-variable vector x. The simplex method solves a linear program in a comprehensive sense; in particular, no conditions are imposed on the data (A, b, c). Dantzig assessed a linear program as the simplest optimization model with broad applicability.

The study, solution, and application of linear programs constitute the subject of linear programming. The use of the words ‘programming’ and ‘program’ has changed somewhat over time. The original idea was that ‘programming’ is the activity of deciding now upon a plan, called a program, for some system that would be executed later in time. The same meaning was subsequently adopted in computer programming where the system is a computer. (See linear programming.)

Early life and education

George Bernard Dantzig was born to Tobias Dantzig and his wife, Anja Ourisson, in Portland Oregon, on 8 November 1914. Tobias, a housepainter and pedlar in his early years in the United States, later held professional positions at John Hopkins University (1919–20) and the University of Maryland (1927–46) where he chaired the mathematics department from 1930 to 1941. He is best known for his book

Number, The Language of Science, which is still in print (T. Dantzig, 1930).

In 1936, George Dantzig both received an A.B. in Mathematics and Physics at the University of Maryland and married Anne Shmuner (1917–2006), who at age 19 received an A.B. in French at Maryland. In 1938 Dantzig received an MA in Mathematics at the University of Michigan; he was a Horace Rackham Scholar. In 1937–39 Dantzig worked at the U.S. Bureau of Labor Statistics as a junior statistician. Inspired by a paper of J. Neyman on which he had been assigned to report, Dantzig wrote to Neyman, then at University College London, asking if he could study under his supervision. Neyman relocated to the University of California at Berkeley, and Dantzig became his student in 1939. As is now folklore, one day Dantzig arrived late for one of Neyman's theoretical statistics classes and proceeded to copy two problems from the blackboard. In a few weeks time, with some effort, Dantzig solved the problems and submitted his homework, whereupon it was tossed onto a large pile of papers on Neyman's desk. Early one Sunday morning, about two weeks later, George and Anne were awakened by a pounding on their apartment door. There was Neyman waving George's homework. As it turned out, the assumed homework problems were, in fact, important unsolved problems. Furthermore, Neyman continued, these solutions, suitably presented, would suffice for George's Ph.D. dissertation. A. Wald independently obtained one of the same results, and the work was eventually published jointly in Dantzig and Wald (1951). Before Dantzig could complete his degree, Pearl Harbor was attacked, and he took leave of absence to work at the U.S. Air Force Comptroller's Office.

Dantzig in Washington, DC, 1941 – 52

At the outbreak of the Second World War, Dantzig began working at the War Department, again as a junior statistician. By the war's end, he was in charge of the Combat Analysis Branch of the Statistical Control Division of the United States Air Force. His office collected and consolidated data with hand-operated mechanical desk calculators about sorties flown, tons of bombs dropped, planes lost, personnel attrition rates, and so on. By end of the war, Dantzig had a personnel force of 300 reporting to him.

In 1946 Dantzig returned to Berkeley for one semester to defend his thesis and complete his minor thesis in dimension theory. Throughout his life, Dantzig acknowledged a great debt to J. Neyman, his mentor. Dantzig nonetheless turned down a position in mathematics at UC Berkeley for the greater financial security of a position at the Pentagon. There he undertook the challenge to ‘mechanize’ the planning process. War planning required coordination of an entire nation and yet was executed with desk calculators; the need for mechanization was clear. To this end, a group in the Air Force was organized under the name Project SCOOP (Scientific Computation of Optimum Programs) and headed by M.K. Wood. Dantzig was a principal. Two movements suggested that progress was possible: Leontief's (1936) work and the emergence of the computer; indeed, Project SCOOP arranged for Pentagon support of computer development (see Dantzig, 1947).

In early 1947, Dantzig formulated the general statement of a linear program. In June of that year he learned from T.C. Koopmans, who had been studying transportation problems (Koopmans, 1947) that economists had no algorithm for solving a linear program. By July Dantzig had designed the simplex method, a name suggested by Leo Hurwicz (see simplex method for solving linear programs). Experiments with the simplex method in the following year at the Pentagon were encouraging. Linear programs were also solved with the simplex method at the National Bureau of Standards (NBS) in

coordination with SCOOP. At NBS a ‘large one’, the diet problem, was undertaken by J. Laderman. It had been studied earlier by Stigler (1945). The question was: what selection of 77 foods produces a diet meeting nine nutritional criteria at the least cost? The problem was solved by the simplex method with five statistical clerks using desk calculators. According to (Dantzig 1963), ‘approximately 120 man-days were required to obtain a solution’. The simplex method was gaining acceptance. Air Force applications of linear programming in years following included contract bidding, crew training, deployment scheduling, maintenance cycles, personnel assignments, and airlift logistics.

From special cases as a triangular model to the general algorithm, the simplex method was first implemented on a computer in 1949 by M. Mantalbano (NBS) on an IBM 602-A, in 1950 by C. Diehm on the SEAC, in 1951 by A. Orden (Air Force) and A. Hoffman (NBS) on the SEAC, and in 1952 by the Air Force for the Univac. The next generation of codes, circa 1952–56, which achieved commercial quality, was developed by W. Orchard-Hays at the RAND Corporation on a sequence of IBM machines. For the matrix A of $LP(A, b, c)$ of size 200 by 1000, linear programs could be solved in five hours (Orchard-Hays, 1954). In years following, there was a flood of computer implementations, both by commercial vendors and in research institutions. As of 2006, linear programs where both m and n exceed hundreds of thousands are routinely solved in hours by the simplex method on personal computers.

After describing and testing the simplex method, Dantzig had an audience with J. von Neumann at Princeton in 1947. Among world-class mathematicians, von Neumann had the broadest interests. Dantzig began his explanation of linear programming with the 30-minute version when von Neumann snapped ‘Get to the point’. Dantzig began again, this time with his one-minute version. Von Neumann responded, ‘Oh, that!’ He envisioned an analogy with matrix games as developed in von Neumann and Morgenstern (1944). Extrapolating from what he knew about duality in matrix games, von Neumann expounded on what was to become known as duality in linear programming. As a by-product of the meeting, it was evident that any matrix game problem could be solved by a linear program. Volume VI of John von Neumann: Collected Works contains his previously uncirculated manuscript dated 15–16 November 1947 on duality in linear programming (von Neumann, 1947). The following January, Dantzig (1948a) wrote ‘A Theorem on Linear Inequalities’. This memorandum clarified his understanding of von Neumann's duality monologue. Von Neumann's (1947) paper is regarded as the earliest on this subject; Dantzig's memorandum is the second. A.W. Tucker, also at Princeton, took an interest in the relationship of linear programming and game theory and involved his students, D. Gale and H.W. Kuhn. These three subsequently wrote the definitive account of duality in linear programming (Gale, Kuhn and Tucker, 1951).

First linear programming conference, 1949

Koopmans organized a conference on ‘linear programming’ and economics in Chicago at the Cowles Commission for Research in Economics in 1949. Koopmans and others (including Dantzig) edited the conference proceedings volume *Activity Analysis of Production and Allocation* (1951). Dantzig's work was the focus of the proceedings; of the 25 papers, Dantzig co-authored a paper with M.K. Wood and authored four others, including the two leading papers which developed linear programming for time-staged planning. Earlier versions of these two papers appeared in *Econometrica* (1949). Four of the 20

contributors to these proceedings – K.J. Arrow, T.C. Koopmans, P.A. Samuelson, and H.A. Simon – were later to win Nobel Prizes. Hundreds of books on, or inspired by, linear programming followed over the years. Four of note are Dorfman, Samuelson and Solow (1958), Arrow, Hurwicz and Uzawa (1958), Dantzig (1963), and Schrijver (1986). The terminology ‘linear programming’ was not in regular use at the time of this conference; Koopmans had suggested it to Dantzig (1948b) in lieu of expressions like ‘programming in a linear structure’. Even so, Koopmans (1951) observed, ‘To many economists the term linearity is associated with narrowness, restrictiveness, and inflexibility of hypotheses’. R. Dorfman, at the Pentagon with Dantzig, had suggested the broader expression of ‘mathematical programming’.

Nonlinear programming, 1950

Following the early successes of linear programming, there was a natural inclination to generalize the model, the algorithm, and duality to results beyond linear functions to a next layer of difficulty such as differentiable, convex, quadratic, or polynomial functions. This body of research has become known as ‘nonlinear programming’. As for optimality conditions and duality, the paper ‘Nonlinear Programming’ of Kuhn and Tucker (1951) was pivotal at the time: their investigation proceeded through the Lagrangian function and saddle points thereof with the duality in linear programming as a target. The Lagrangian had been used in equality-constrained optimization, and results obtained there were less general. Kuhn and Tucker cited the fundamental paper of John (1948), which includes inequality constraints. Some 25 years later, the master's thesis of Karush (1939) came to light in the mathematical programming community; Karush, as far as is known, was the first to lay down optimality conditions for a nonlinear (inequality constrained) program. Rockafellar (1970) carried the convex duality analysis to a new level. As for nonlinear programming algorithms, tens, and eventually hundreds, were forthcoming, many using ideas from the simplex method in one way or another.

Dantzig at RAND, 1952 – 60

Reorganization of the Air Force preceded Dantzig's taking a position at the RAND Corporation in Santa Monica, California, as a research mathematician. Awareness of the power of linear programming set the scene for a second growth. For the next few years most theoretical development of linear programming took place at RAND and Princeton. Dantzig's book *Linear Programming and Extensions* (1963) records his own (and collaborative) contributions during this period.

Transportation and network optimization problems

The war years has seen interest in optimal transportation research. Historically significant papers from this period include Hitchcock (1941), Kantorovich (1942), Koopmans (1947; 1949), Kantorovich and Gavurin (1949), and Flood (1956). Flood, through M. Shiffman, had come upon the Kantorovich papers on translocation and transportation; however, linear programming launched the general analysis of optimal transportation. Dantzig made several contributions here, starting with Dantzig (1951). Dantzig,

Fulkerson and Johnson (1954) is a seminal work on the travelling salesman problem. Others are Dantzig and Fulkerson (1954) on tanker routing, and Dantzig and Fulkerson (1955) on maximizing flow through a network. For networks with non-negative arc distances, Dantzig (1960a) stated an algorithm for shortest distances. Dijkstra (1959) produced similar results at about the same time. *Flows in Networks* by the RAND Corporation's Ford and Fulkerson (1962) was then the definitive work on the subject.

Large-scale methods and decomposition

Dantzig and Orchard-Hays (1954) described the ‘revised simplex method’ as a more efficient version of the simplex method. As linear programming was applied to more applications and with a broader scope, including time and alternate scenarios, the size of linear programs that needed to be solved continued to grow. Dantzig was among the first to observe that large linear programs typically had two convenient features: sparsity and structure. Sparsity refers to the fact that a very small percentage, often less than one hundredth of one per cent, of the *A* data matrix is non-zero. Structure refers to the fact that the non-zeros typically occur an orderly pattern of submatrices of *A*. Dantzig (1955a) wrote the first paper on methods for large-scale linear programs addressing upper bounds, block triangular systems, and secondary constraints. Building on the Dantzig, Orden and Wolfe (1955) paper on generalized linear program, Dantzig and Wolfe (1960) devised a generalization of the simplex method, called the *decomposition principle*, for certain structured large-scale linear programs, wherein the problem is decomposed allowing for use for what is now called distributed computation.

Quadratic programming

A most natural first extension of a linear program is a quadratic program, that is, a linear program except that the objective is a quadratic function such as $x^T Qx + q^T x$. A convex quadratic program is one with a convex objective function to be minimized. Following the success of linear programming, there was a proliferation of studies on convex quadratic programming and associated algorithms.

Convex programming

Convex programming is also a natural extension of linear programming. Here a convex function is minimized over a convex region; the latter is specified by convex inequality constraints. If the feasible region is bounded, the convex program can be approximated as close as desired by a linear program, and one can improve the approximation as the simplex method runs. A special case of a convex program is one having linear inequality constraints and a *separable* objective function, that is, a function that is the sum of univariate convex functions. Charnes and Lemke (1954) and Dantzig (1956) solved such problems with linear programming approximations.

Stochastic programming with recourse

Linear programming offered a breakthrough for mathematical approximation and solution of planning problems. Dantzig knew that to move to the next level of approximation of planning, an accommodation

of uncertainty and of discrete variables was needed; he made inroads on each. Linear programming has been extended in a number of directions to incorporate uncertainty. An elementary example is a linear program where the costs $c = (c_1, c_2, \ldots, c_n)$ are random variables and the desire is to minimize the expected value. In this case the problem is solved as the linear program where the costs are simply taken as their expected value. More interesting is the Markowitz (1956) portfolio selection where quadratic programming is used to obtain at least a desired level of expected return while minimizing risk. Dantzig's early work on stochastic programming was stimulated by his work with A.R. Ferguson on the assignment of aircraft to routes, where a deterministic formulation proved insufficient, and so uncertain demand needed to be considered (Ferguson and Dantzig, 1955). Subsequently, Dantzig (1955b) applied linear programming to solve multistage decision problems sequenced amidst uncertainty; this topic is often referred to as stochastic programming with recourse. Such a multistage problem concerns the optimization of a sequence of decisions in time where each decision depends on random events which in turn are dependent on previous decisions. The vision in this paper was truly extraordinary, and has been reprinted as one of the ten most influential papers in management science since the mid-1950s in Hopp (2004).

Integer programming and cuts

An integer program is a linear program except that some, or all, of the variables $x_1, x_2, \ldots, x_n$ are required to take on integer values, as in $x_i = 0, 1, 2, \ldots$. Dantzig, Fulkerson and Johnson (1954) took the first steps towards obtaining integer solutions for a large problem with the simplex method. They addressed an instance of the travelling salesman problem: find the shortest route, by car, through major cities of the 48 states and Washington, DC. Let a directed network represent the available roads and let costs represent distances. The variables are flows on each link of the network. Constrain for one unit of flow into each capital, constrain for one unit of flow out of each capital, constrain for conservation of flow at other nodes, and find the minimum cost flow. The linear programming solutions here, which yield flows of 0 or 1, are deficient as a solution for the travelling salesman problem in that isolated loops of flow may occur. To combat such loops, Dantzig et al. sequentially and dynamically (as the simplex method was stopped and continued) introduced additional constraints, called cuts, which would prohibit those loops which had occurred in a solution of the expanding linear program, without constraining out desired solutions. The concept of a cut or cutting planes was so conceived. In addition, this study revealed the inherent difficulty of the travelling salesman problem. Over the following decades, aspects of this matter would grow to become a major issue in applied mathematics. There is a vast difference between linear constraints and linear inequality constraints (both with unconstrained variables); there is an even larger difference between real variables and integer variables. Subsequently, Gomory (1958), at Princeton, began the design of several general purpose cutting plane algorithms for solving integer programs, and gave proofs for finite convergence. These algorithms did not work well for a reason not understood at the time namely, that general integer programs are inherently hard to solve.

Other edge path descent algorithms

By 1955 the simplex method was regarded as *the* algorithm for solving linear programs. Indeed, the

simplex method inspired dozens of related fundamental ideas for algorithms, and hundreds of variations. In particular, there was steady research on variations of edge path descent algorithms, that is, those which accept the simplex method strategy but strive to improve upon it. One target was to reduce computation time by reducing the number of pivots and the work per pivot. Example contributions include: the dual simplex method of Lemke (1954), the parametric method of Orchard-Hays (1954), the primal-dual method of Dantzig, Ford and Fulkerson (1956) and the parametric objective method of Gass and Saaty (1955). In a slightly different direction were the column generation and the decomposition method of Dantzig and Wolfe (1960; 1961). Essentially all of these variants of the simplex method have proved valuable for various specialized tasks related to linear programs, and sometimes nonlinear programs. For nonlinear programs the main ideas of the simplex method have been adopted; here one can think of solving linear or quadratic programs that are approaching the nonlinear program. It is interesting to note that as late as the early 1970s an eminent speaker of a plenary session of a national mathematical programming conference said that the simplex method was the best algorithm for linear programming and that it always would be; the statement was accepted, without objection.

Problem reduction

The mathematical subject of computational complexity aims to categorize problems by their solution difficulty. Several of Dantzig's papers (1957; 1960b; 1968) contributed to the foundation of this subject. A basic technique of computational complexity is the reduction of one class of problems to another. For the reduction of discrete problems, Dantzig focused on problems in mixed binary form, *MBP*, and the related relaxed form *RMBP* obtained by replacing binary constraints with corresponding interval constraints. *MBP*(*A*, *b*, *c*) is a linear program *LP*(*A*, *b*, *c*) plus the discrete constraints $x_i = 0, 1$ for $i = 1, \ldots, k$ for some $k \leq n$. *RMBP*(*A*, *b*, *c*) is the linear program *LP*(*A*, *b*, *c*) plus the linear inequalities $0 \leq x_i \leq 1$ for $i = 1, \ldots, k$. For emphasis, *MBP*(*A*, *b*, *c*) is not a linear program whereas *RMBP*(*A*, *b*, *c*) is. A few problem classes of form *MBP* can be solved as the corresponding linear program *RMBP*; that is if (*x*, *z*) is an extreme point solution, as the simplex method would generate, of *RMBP*, then (*x*, *z*) is a solution to *MBP*. Problem classes *MBP* which can be so solved by *RMBP* include the assignment problem, shortest route problems with non-negative distances, and the tanker scheduling problem. Other problems, such as the empty container problem, most scheduling problems, fixed charge problems, and travelling salesman problems, do not permit such solution; nevertheless, the corresponding *RMBP* can be most helpful in solving or approximately solving *MBP*. As time and theory have revealed, general problems of type *MBP* are difficult to solve.

Let C^* be the convex hull of all feasible solutions of *MBP* and let *C* be the set of all solutions of *RMBP*. Then C^* is a subset of *C*, and all extreme points of C^* are extreme points of *C*; the issue is, however, that there are extreme points of *C* that are not in C^*. Note that, if there is but one binary variable, then *MBP* can be solved as two linear programs, one with $x_1 = 0$ and one with $x_1 = 1$; but for general *k*, this scheme requires the solution of an exponential number 2^k of linear programs. For reducing problems to the *MBP* form, Dantzig (1960b) illustrated a number of examples such as: (*a*) dichotomies, (*b*) discrete variables, (*c*) piecewise linear objective functions, (*d*) conditional constraints, and (*e*) the fixed charge problem.

Recognition of earlier work, 1958 – 60

Towards the end of the 1950s, the mathematical programming community became aware of three relevant works from the past. The first two are pertinent to the simplex method and the third relevant to the formulation of real problems as linear programs. Fourier (1826) had also written on the idea of descending from vertex to adjacent vertex in the polyhedron defined by linear inequalities for minimizing a linear error over linear inequalities. De la Vallée Poussin (1911), independently of Fourier's work, made a similar suggestion and gave two examples. There appears to have been no follow-up on their suggestions. Also, neither Fourier nor de la Vallée Poussin described his ideas fully enough to reveal any awareness of degeneracy considerations and corresponding non-convergence possibilities, much less any procedures for coping with the matter. Made aware of Kantorovich's transportation papers by Flood (1956), Koopmans (1960) corresponded with Kantorovich. In due course, an English translation of Kantorovich's remarkable 1939 paper was made available to the West as ‘Mathematical Methods of Organizing and Planning Production’ (Kantorovich, 1960). Therein Kantorovich had formulated a collection of problems as what we now call linear programs. These problems were: machine utilization, production planning, scrap management, refinery scheduling, fuel utilization, construction planning, and arable land distribution. Using the Minkowski separation theorem, Kantorovich proved in this work that optimal multipliers exist. He suggested some ideas based on ‘resolving multipliers’ (essentially dual variables, or marginal costs) towards an algorithm, but none has emerged following this line of thought. According to Dantzig (1963), ‘Kantorovich should be credited with being the first to recognize that certain important broad classes of production problems had well-defined mathematical structures which, he believed, were amenable to practical numerical evaluation and could be numerically solved’. But although Koopmans (1960) argued that, with a suitable transformation, one of Kantorovich's problems had the generality of Dantzig's linear program, Koopmans's conclusion was not justified as the argument did not and could not cover the possibilities of infeasibility and an unbounded objective, a point made by Charnes and Cooper (1962). Koopmans's argument notwithstanding, the statement of a general linear program belongs to Dantzig.

Dantzig returns to UC Berkeley, 1960 – 66

Dantzig left RAND to become a professor in the industrial engineering department at the University of California at Berkeley. There, that year, he established the Operations Research Center. Operations research (OR) was a term that emerged in the Second World War to describe the activity of studying an operation (process, system, and so on) with mathematical methods with the intent of improving performance. In 1963 Dantzig completed his classic *Linear Programming and Extensions*. The book was based on his research which began at the Pentagon and continued through RAND and UC Berkeley. By the time Dantzig left UC Berkeley in 1966, he had produced 11 Ph.D. students, and written about 25 research papers on the theory and practice of linear programming and extensions (integer, nonlinear, stochastic, and so on). As a mentor of Ph.D. students, Dantzig was among the very best. Within a course or two he could bring students to the frontier on some aspect of linear programming. His new book offered a full perspective of linear programming right up to 1963. Dantzig supplied the time, inspiration, guidance, knowledge, and example that students needed. He lived and breathed research.

Interest in the study of linear and nonlinear complementarity problems, as such, began in the early 1960s. Dantzig's second student, Cottle (1964), wrote on this topic, and his work was extended in Cottle and Dantzig (1968). Problems in this category can be viewed as abstractions of optimality conditions or of (economic or physical) equilibrium conditions. In a complementarity problem, one has a mapping W of R^N into itself and seeks a solution z of the conditions $W(z) \geq 0$, $z \geq 0$, $z^T W(z) = 0$. In the linear complementarity problem, the mapping would be of the form $W(z) = Mz + q$. The linear complementarity problem is related to the minimization of $z^T(Mz + q)$ subject to the constraints $Mz + q \geq 0$ and $z \geq 0$. This would be easy enough to solve as a quadratic program, if the objective function were convex. However, the excitement arose from the fact that the problem could be solved, effectively, in the absence of convexity. From the classic paper of Lemke (1965) followed the computation of points in the core of a balanced game and the computation of economic equilibria (Scarf, 1967; 1973), the computation of fixed points with piecewise linear homotopies (Eaves, 1972), and the computation with differentiable functions (Smale, 1976).

Dantzig at Stanford University, 1966 – 96

Dantzig joined the Stanford faculty in 1966, half-time in the inter-departmental Operations Research Program and half-time in Computer Science. In 1967 the OR Program became the Department of Operations Research in the School of Engineering; this is where Dantzig conducted his work. He was away for two years: in 1973–74 at the International Institute for Applied Systems Analysis in Austria, and in 1978–79 at the Center for Advanced Study in the Behavioral Sciences on the Stanford campus. In 1973 he was appointed to the C.A. Criley Professorship in Transportation Science. While at Stanford, Dantzig produced 41 Ph.D. students and published about 115 research papers on the theory and applications of mathematical programming. Dantzig's Ph.D. progeny, if Berkeley and Stanford graduates and subsequent generations are counted, as of 2006 exceeded 200. Dantzig had long felt that the development of good software was key to widespread usage of linear programming in industry. This vision led him to create the Systems Optimization Laboratory (SOL) at Stanford for research and development of numerical algorithms for mathematical programming. Under the SOL banner were the PILOT and planning under uncertainty programs (see Dantzig et al., 1973; Gill et al., 2007).

Stochastic programming with recourse, continued, 1989 – 2005

Cognizant of the potentially enormous size of multi-stage stochastic linear programs, Dantzig and Madansky (1961) suggested the incorporation of statistical sampling of uncertainties together with approximating time-staged models to solve the full problem. Following this avenue some 30 years later, Dantzig and Glynn (1990) brought together decomposition, Monte Carlo sampling, and multiprocessing to solve time-staged linear programs (see also Infanger, 1991; Dantzig and Infanger, 1992). In a series of papers, importance sampling was used to estimate second-stage costs and Benders cuts. Portfolio optimization and electric power planning were among the applications envisioned; the latter problems, with 39 uncertain parameters leading to 15 million scenarios, were solved to high accuracy with a confidence level of 95 per cent; in equivalent deterministic form, such problems would have more than

four billion constraints. However, Dantzig, to the end, regarded stochastic linear programming as a major unresolved problem.

Computational complexity, 1972 – 2006

Since its inception, the question of the number of steps required by the simplex method for a given linear program has been of interest. In the 1970s the field of ‘computational complexity’ emerged; a theory of problem difficulty which draws a sharp distinction between categories of problems that could be solved in polynomial time (number of steps) in the size of their data, and those which could not. How did linear programming fit into this scheme? Klee and Minty (1972) produced a worst case example of a simple linear program on which the simplex method takes an exponential number of iterations. But the expected number of pivots of the simplex method over a random selection of problems was shown to be polynomial in (m, n) (Smale, 1983). This raised the question: could a linear program be solved in polynomial time? Khachiyan (1979) defined a polynomial time algorithm for linear programs based on a sequence of convergent ellipsoids; however, unexpectedly according to computational complexity, the algorithm was very slow, and certainly no competitor of the simplex method. Later, Karmarkar (1984) gave a polynomial time interior point algorithm for linear programs which was claimed to be superior to the simplex method in the sense of solving linear programs much faster on a computer; the method required the linear program to be expressed in a special form with an optimal objective value of zero and viewed each iterate as being at the centre of a polyhedron in a different coordinate system. The method typically required considerably fewer iterations than the simplex method, but each iteration required significantly more computations. The method was patented by AT&T and published as a theoretical result. There was considerable secrecy associated with the particulars of its implementation; and, thus, no independent verification was possible regarding its claimed superiority in computational speed over the simplex method. It was later shown to be equivalent under the same special form to the logarithmic barrier method, a method traceable back to Frisch (1955) and Fiacco and McCormick (1968). The logarithmic barrier method, however, could be applied to a linear program in standard form. The logarithmic barrier method was in the public domain and so allowed researchers to focus on computational improvements. Today, it is known that there are problems for which the logarithmic barrier method is superior to the simplex method; notable are those very large problems for which AA^T is sparse. For a survey of interior point methods, see Todd (1996). It is also interesting to note that most practical interior-point algorithms include an option to move the ε -optimal interior point solution to the nearest extreme point, a procedure requiring a significant number of simplex-type pivots. A technique to do this was proposed by Dantzig (1963, ch. 6, exercise 11). As of 2006, the simplex method (and various realizations thereof) remains the algorithm of choice for the majority of linear programs.

Dantzig in retirement, 1996 – 2005

Dantzig was retired from Stanford in stages, each firmly resisted. He was formally retired from the regular faculty at age 65 in 1980, but was recalled until age 82 in 1996. Until that year he remained as active as formal members of the faculty. After that he met at home with all who wished to consult him: students, colleagues, and strangers. Whenever presented with an idea, Dantzig would respond, as

always, with something of value. Until around 2001 he continued to travel and present papers. At his 90th birthday celebration, he attended a full day of presentations followed by a banquet and additional talks. He was full of energy, enthusiasm, keen observations, and wit. Dantzig's mind was razor-sharp up to the end.

In retirement, Dantzig's principal project was the writing of a multi-volume book on linear programming and extensions. Dantzig had always felt that software was a key element that would contribute to the success of linear programming usage. He wanted to write another book on linear programming that incorporated software to aid students in learning both the theory and the practice of linear programming, and in particular in learning how to implement the simplex method and other algorithms for commercial use. In 1985 he invited M.N. Thapa to coauthor such a book. As work on the book progressed, it became apparent to the authors that the amount of material required a really huge book. One volume became two, and two became four. In the end only two volumes were completed (Dantzig and Thapa, (1997; 2003). Dantzig continued to be fascinated by interior point methods; von Neumann's and Karmarkar's algorithms were reanalysed and included in the second volume. According to M. Thapa, Dantzig never tired of editing and re-editing to improve proofs and readability. He would say: ‘it is like polishing a stone; the more you polish it, the more it will shine.’ Dantzig also continued his work with G. Infanger on planning under uncertainty. In addition to their research together, Dantzig and Infanger consulted on financial portfolio design. They intended to edit a collection of papers (including work of their own) on planning under uncertainty. Dantzig was convinced that the way to get further exposure for, and research into, planning under uncertainty was to set up an institute; to no avail, he tried at Stanford, tried at EPRI, and finally tried to create a stand-alone non-profit organization. In addition to these projects, Dantzig reworked the text of a science fiction novel he had begun in 1980.

Dantzig's honours

In 1975 L.V. Kantorovich and T.C. Koopmans received the Nobel Prize in Economics for ‘their contributions to the theory of optimum allocation of resources’. Both mentioned Dantzig in their Nobel Lectures. That Dantzig did not participate in this prize came as a great shock and disappointment to those familiar with his contributions. Himself aside, Dantzig regarded Leontief, Kantorovich, von Neumann, and Koopmans as the principal early contributors to linear programming.

Dantzig, the man, and his contributions have nevertheless been honoured extensively. His honours include distinguished memberships, prizes, honorary doctorates, and dedications. He was elected to membership in the National Academy of Sciences, the National Academy of Arts and Sciences, and the National Academy of Engineering. He was a fellow of the Econometric Society, the Institute of Mathematical Statistics, the Association for the Advancement of Science, the Operations Research Society, IEEE, and the Omega Rho Society. He was awarded the War Department Exceptional Civilian Service Medal, the National Medal of Science, the John von Neumann Theory Prize, the NAS Award in Applied Mathematics and Numerical Analysis, the Harvey Prize (Technion), the Silver Medal of Operational Research Society (England), the Adolph Coors American Ingenuity Award, the Special Recognition Award of Mathematical Programming Society (MPS), the Harold Pender Award, and the Harold Lardner Memorial Prize (Canada). He received honorary doctorates from the Israel Institute of Technology (Technion), University of Linkøping (Sweden), University of Maryland, Yale University, Université Catholique de Louvain (Belgium), Columbia University, the University of Zurich, and

Carnegie-Mellon University. Dantzig was also honoured as the dedicatee of a symposium of MPS, in two volumes of *Mathematical Programming*, in the first issue of the *Journal of Optimization of the Society for Industrial and Applied Mathematics* (SIAM), with the joint MPS-SIAM Dantzig Prize, and with the INFORMS Dantzig Prize for students. In 2006, a fellowship in his name was established in the Department of Management Science and Engineering at Stanford University.

Acknowledgements

The authors are grateful to David Dantzig, Jessica Dantzig Klass, and many of Dantzig's friends and colleagues who have contributed to this biographical article. These include A.J. Hoffman, G. Infanger, E. Klotz, J.C. Stone, M.J. Todd, J.A. Tomlin and M.H. Wright. This article has also benefited from other writings on G.B. Dantzig's life, namely: Albers and Reid (1986), Albers, Alexanderson and Reid (1990), Cottle (2003; 2005; 2006), Cottle and Wright (2006), Dantzig (1982; 1991), Dorfman (1984), Gill et al. (2007), Kersey (1989), Lustig (2001), Gass (1989; 2002; 2005).

See Also

- Kantorovich, Leonid Vitalievich
- Koopmans, Tjalling Charles
- Leontief, Wassily
- linear programming
- simplex method for solving linear programs
- von Neumann, John

Selected works

1947. Prospectus for the AAF electronic computer. Unpublished manuscript.

1948a. A theorem on linear inequalities. Unpublished manuscript, 5 January.

1948b. Programming in a linear structure. Washington, DC: Comptroller, USAF. February. Abstract in *Econometrica* 17(1949), 73–4.

1951. Application of the simplex method to a transportation problem. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley & Sons.

1951. (With A. Wald.) On the fundamental lemma of Neyman and Pearson. *Annals of Mathematical Statistics* 22, 87–93.

1954. (With D.R. Fulkerson.) Minimizing the number of tankers to meet a fixed schedule. *Naval*

Research Logistics Quarterly 1, 217–22.

1954. (With D.R. Fulkerson and S.M. Johnson.) Solution of a large scale traveling-salesman problem. *Journal of Operations Research Society of America* 2, 393–410.

1954. (With W. Orchard-Hays.) The product form for the inverse in the simplex method. *Mathematical Tables and Other Aids to Computation* 8, 64–7.

1955a. Upper bounds, secondary constraints, and block triangularity in linear programming. *Econometrica* 23, 174–83.

1955b. Linear programming under uncertainty. *Management Science* 1, 197–206.

1955. (With A.R. Fergusson.) The problem of routing aircraft. *Aeronautical Engineering Review* 14(4), 51–5. RAND Research Memorandum RM1369, 1954.

1955. (With D.R. Fulkerson.) Computation of maximal flows in networks. *Naval Research Logistics Quarterly* 2, 277–83.

1955. (With A. Orden, A. and P. Wolfe.) The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics* 5(2), 183–195. RAND Research Memorandum RM-1264, 1954.

1956. Recent advances in linear programming. *Management Science* 2, 131–44.

1956. (With L.R. Ford, Jr. and D.R. Fulkerson.) A primal-dual algorithm for linear programs. In *Linear Inequalities and Related Systems*, Annals of Mathematics Study No. 38, ed. H.W. Kuhn and A.W. Tucker. Princeton, NJ: Princeton University Press.

1957. Discrete variable extremum problems. *Operations Research* 5, 226–77.

1960a. On the shortest route through a network. *Management Science* 6, 187–90. RAND memorandum P-1345, 1959.

1960b. On the significance of solving linear programming problems with some integer variables. *Econometrica* 28, 30–44.

1960. (With P. Wolfe.) Decomposition principle for linear programs. *Operations Research* 8, 101–11.

1961. (With A. Madansky.) On the solution of two-stage linear programs under uncertainty. In *Proceedings, Fourth Berkeley Symposium on Mathematical Statistics and Probability*, I, ed. J. Neyman.

Berkeley University of California Press. RAND memorandum P-2039, 1960.

1961. (With P. Wolfe.) The decomposition algorithm for linear programming. *Econometrica* 29, 767–78.

1963. *Linear Programming and Extensions*. Princeton, NJ: Princeton University Press.

1968. Large-scale linear planning. In *Mathematics of the Decision Sciences*, vol. 1, ed. G.B. Dantzig and A.F. Veinott, Jr. Providence, RI: American Mathematical Society.

1968. (With R.W. Cottle.) Complementary pivot theory of mathematical programming. *Linear Algebra and its Applications* 1, 103–25.

1973. (With others.) On the need for a System Optimization Laboratory. In *Mathematical Programming*, ed. T.C. Hu and S.M. Robinson. New York: Academic Press.

1982. Reminiscences about the origins of linear programming. *Operations Research Letters* 1(2), 43–48.

1990. (With P.W. Glynn.) Parallel processors for planning under uncertainty. *Annals of Operations Research* 22, 1–21.

1991. Linear programming. In *History of Mathematical Programming: A Collection of Personal Reminiscences*, ed. J.K. Lenstra, A.H.G. Rinnooy Kan and A. Schrijver. Amsterdam: North-Holland.

1992. (With G. Infanger.) Large-scale stochastic linear programs: importance sampling and Benders decomposition. In *Computational and Applied Mathematics I – Algorithms and Theory, Proceedings of the 13th IMACS World Congress*, ed. C. Brezinski and U. Kulisch. Amsterdam: North-Holland.

1997. (With M.N. Thapa.) *Linear Programming, 1. Introduction*. New York: Springer.

2003. (With M.N. Thapa.) *Linear Programming 2: Theory and Extensions*. New York: Springer.

Bibliography

Albers, D.J., Alexanderson, G.L. and Reid, C. 1990. *More Mathematical People*. New York: Harcourt Brace Jovanovich.

Albers, D.J. and Reid, C. 1986. An interview with George B. Dantzig: the father of linear programming. *College Mathematics Journal* 17, 292–314.

Arrow, K.J., Hurwicz, L. and Uzawa, H. 1958. *Studies in Linear and Non-Linear Programming*. Stanford, CA: Stanford University Press.

Benders, J.K. 1962. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik* 4, 238–52.

Charnes, A. and Cooper, W.W. 1962. On some works of Kantorovich, Koopmans and others. *Management Science* 8, 246–63.

Charnes, A. and Lemke, C.E. 1954. Minimization of non-linear separable convex functionals. *Naval Research Logistics Quarterly* 1, 301–12.

Cottle, R.W. 1964. Nonlinear programs with positively bounded Jacobians. Ph.D. thesis. Department of Mathematics. University of California at Berkeley.

Cottle, R.W. 2003. *The Basic George B. Dantzig*. Stanford, CA: Stanford University Press.

Cottle, R.W. 2005. George B. Dantzig: operations research icon. *Operations Research* 53, 892–8.

Cottle, R.W. 2006. George B. Dantzig: a life in mathematical programming. *Mathematical Programming* 105, 1–8.

Cottle, R.W. and Wright, M.H. 2006. Remembering George Dantzig. *SIAM News* 39(3), 2–3.

Dantzig, T. 1930. *Number, The Language of Science*. New York: Macmillan. 4th edition, ed. J. Mazur, republished New York: Pi Press, 2005.

de la Vallée Poussin, M.C.J. 1911. Sur la méthode de l'approximation minimum. *Annales de la Société Scientifique de Bruxelles* 35, 1–16.

Dijkstra, E. 1959. A note on two problems in connection with graphs. *Numerische Mathematik* 1, 269–71.

Dongarra, J. and Sullivan, F. 2000. The top 10 algorithms. *Computing in Science and Engineering* 2(1), 22–3.

Dorfman, R. 1984. The discovery of linear programming. *Annals of the History of Computing* 6, 283–95.

Dorfman, R., Samuelson, P.A. and Solow, R.M. 1958. *Linear Programming and Economic Analysis*. New York: McGraw-Hill.

Eaves, B.C. 1972. Homotopies for the computation of fixed points. *Mathematical Programming* 3, 1–22.

Fiacco, A.V. and McCormick, G.P. 1968. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: John Wiley & Sons.

Flood, M.M. 1956. The traveling-salesman problem. *Operations Research* 4, 61–75.

Ford, L.R., Jr. and Fulkerson, D.R. 1962. *Flows in Networks*. Princeton, NJ: Princeton University Press.

Fourier, J.B.J. 1826. Solution d'une question particulière du calcul des inégalités. *Nouveau Bulletin des Sciences par la Société Philomatique de Paris*, 99–100. Reprinted in *Oeuvres de Fourier, Tome II*, ed. G. Darboux. Paris: Gauthier, 1890.

Frisch, R.A.K. 1955. *The Logarithmic Potential Method of Convex Programs*. Oslo: University Institute of Economics.

Gale, D., Kuhn, H.W. and Tucker, A.W. 1951. Linear programming and the theory of games. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley & Sons.

Gass, S.I. 1989. Comments on the history of linear programming. *IEEE Annals of the History of Computing* 11(2), 147–51.

Gass, S.I. 2002. The first linear-programming shoppe. *Operations Research* 50, 61–8.

Gass, S.I. 2005. In Memoriam, George B. Dantzig. 2005. Online. Available at http://www.lionhrtpub.com/orms/orms-8-05/dantzig.html, accessed 12 January 2007.

Gass, S.I. and Saaty, T.L. 1955. The computational algorithm for the parametric objective function. *Naval Research Logistics Quarterly* 2, 39–45.

Gill, P.E., Murray, W., Saunders, M.A., Tomlin, J.A. and Wright, M.H. 2007. George B. Dantzig and systems optimization. Online. Available at http://www.stanford.edu/group/SOL/GBDandSOL.pdf, accessed 2 February 2007.

Gomory, R.E. 1958. Essentials of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society* 64, 256.

Hitchcock, F.L. 1941. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics* 20, 224–30.

Hoffman, A. 1953. *Cycling in the Simplex Algorithm*. Report No. 2974. Washington, DC: National Bureau of Standards.

Hopp, W.J., ed. 2004. Ten most influential papers of *Management Science*'s first fifty years. *Management Science* 50(12 Supplement), 1764–9.

Infanger, G. 1991. Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. *Annals of Operations Research* 39, 41–67.

John, F. 1948. Extremum problems with inequalities as side conditions. In *Studies and Essays, Courant Anniversary Volume*, ed. K.O. Friedrichs, O.E. Neugebauer and J.J. Stoker. New York: Wiley-Interscience.

Kantorovich, L.V. 1942. Translocation of masses. *Doklady Akademii Nauk SSSR* 37, 199–201. Reprinted in *Management Science* 5 (1958–59), 1–4.

Kantorovich, L.V. 1960. Mathematical methods of organizing and planning production *Management Science* 6, 363–422. English translation of original monograph published in 1939.

Kantorovich, L.V. and Gavurin, M.K. 1949. The application of mathematical methods to freight flow analysis. *Problems of Raising the Efficiency of Transport Performance*. Akademiia Nauk, USSR. (Kantorovich confirmed the paper was completed and submitted in 1940, but publication delayed by the Second World War.)

Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorica* 4, 373–95.

Karush, W. 1939. Minima of functions of several variables with inequalities as side constraints. MsC dissertation, Department of Mathematics, University of Chicago.

Kersey, C. 1989. *Unstoppable*. Naperville, IL: Sourcebooks, Inc.

Khachiyan, L.G. 1979. A polynomial algorithm in linear programming in Russian. *Doklady Akademii Nauk SSSR* 244, 1093–6. English translation: *Soviet Mathematics Doklady* 20 (1979), 191–4.

Klee, V. and Minty, G.J. 1972. How good is the simplex algorithm? In *Inequalities III*, ed. O. Shisha. New York: Academic Press.

Koopmans, T.C. 1947. Optimum utilization of the transportation system. *Proceedings of the International Statistical Conferences, 1947*, vol. 5. Washington D.C. Also in *Econometrica* 16 (1948), 66–8.

Koopmans, T.C. 1949. Optimum utilization of the transportation system. *Econometrica* 17(Supplement), 136–46.

Koopmans, T.C., ed. 1951. *Activity Analysis of Production and Allocation*. New York: John Wiley & Sons.

Koopmans, T.C. 1960. A note about Kantorovich's paper, ‘Mathematical Methods of Organizing and Planning Production.’ *Management Science* 6, 363–5.

Kuhn, H.W. and Tucker, A.W. 1951. Nonlinear Programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Lemke, C.E. 1954. The dual method of solving the linear programming problem. *Naval Research Logistics Quarterly* 1, 36–47.

Lemke, C.E. 1965. Bimatrix equilibrium points and mathematical programming. *Management Sciences* 11, 681–9.

Leontief, W.W. 1936. Quantitative input and output relations in the economic system of the United States. *Review of Economic Statistics* 18, 105–25.

Lustig, I. 2001. e-optimization.com. Interview with G.B. Dantzig. Online. Available at http://e-optimization.com/directory/trailblazers/dantzig, accessed 29 December 2006.

Markowitz, H.M. 1956. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly* 3, 111–33. RAND Research Memorandum RM-1438, 1955.

Orchard-Hays, W. 1954. A composite simplex algorithm-II. RAND Research Memorandum RM-1275. Santa Monica, CA: RAND Corporation.

Rockafellar, R.T. 1970. *Convex Analysis*. Princeton, NJ: Princeton Press.

Scarf, H. 1967. The core of an N-person game. *Econometrica* 35, 50–69.

Scarf, H. 1973. *The Computation of Economic Equilibria*. New Haven, CT: Yale University Press.

Schrijver, A. 1986. *Theory of Linear and Integer Programming*. Chichester: John Wiley & Sons.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3, 1–14.

Smale, S. 1983. The problem of the average speed of the simplex method. In *Mathematical Programming: The State of the Art*, ed. A. Bachem, M. Grötschel and B. Korte. Berlin: Springer-Verlag.

Stigler, G.J. 1945. The cost of subsistence. *Journal of Farm Economics* 27, 303–14.

Todd, M.J. 1996. Potential-reduction methods in mathematical programming. *Mathematical Programming* 76, 3–45.

von Neumann, J. 1947. Discussion of a maximum problem. Unpublished working paper dated 15–16 November. In *John von Neumann: Collected Works*, vol. 6, ed. A.H. Taub. Oxford: Pergamon Press, 1963.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

How to cite this article

Cottle, Richard W., B. Curtis Eaves and Mukund N. Thapa. "Dantzig, George B. (1914–2005)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000251> doi:10.1057/9780230226203.0356

The New Palgrave Dictionary of Economics Online

data filters

Timothy Cogley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Empirical economists often filter data prior to analysis to remove features that are a nuisance from the point of view of their theoretical models. Examples include trends and seasonals. This article describes how data filters work and the rationale that lies behind them. It focuses on the Baxter–King and Hodrick–Prescott filters, which are popular for measuring business cycles.

Keywords

ARIMA models; band-pass filters; Baxter–King filter; business cycle measurement; Cramer's representation theorem; data filters; deterministic linear trends; Gaussian log likelihood; generalized method of moments; Granger causation; high-pass filters; Hodrick–Prescott filter; impulse response function; rational-expectations business-cycle models; seasonal adjustment; seasonal fluctuations; shocks; spurious cycle problem; stochastic general equilibrium models; stochastic trends; trend reversion; vector autoregressions

Article

Economic models are by definition incomplete representations of reality. Modellers typically abstract from many features of the data in order to focus on one or more components of interest. Similarly, when confronting data, empirical economists must somehow isolate features of interest and eliminate elements that are a nuisance from the point of view of the theoretical models they are studying. Data filters are sometimes used to do that.

For example, Figure 1 portrays the natural logarithm of US GDP. Its dominant feature is sustained growth, but business cycle modellers often abstract from this feature in order to concentrate on the transient ups and downs. To relate business cycle models to data, empirical macroeconomists frequently filter the data prior to analysis to remove the growth component. Until the 1980s, the most common way to do that was to estimate and subtract a deterministic linear trend. Linear de-trending is conceptually unattractive, however, because it presupposes that all shocks are neutral in the long run. While some disturbances – such as those to monetary policy – probably are neutral in the long run, others probably are not. For instance, a technical innovation is likely to remain relevant for production until it is superseded by another, later technical innovation.

Figure 1
Real US GDP, 1947–2006. *Source*: Federal Reserve Economic Database.

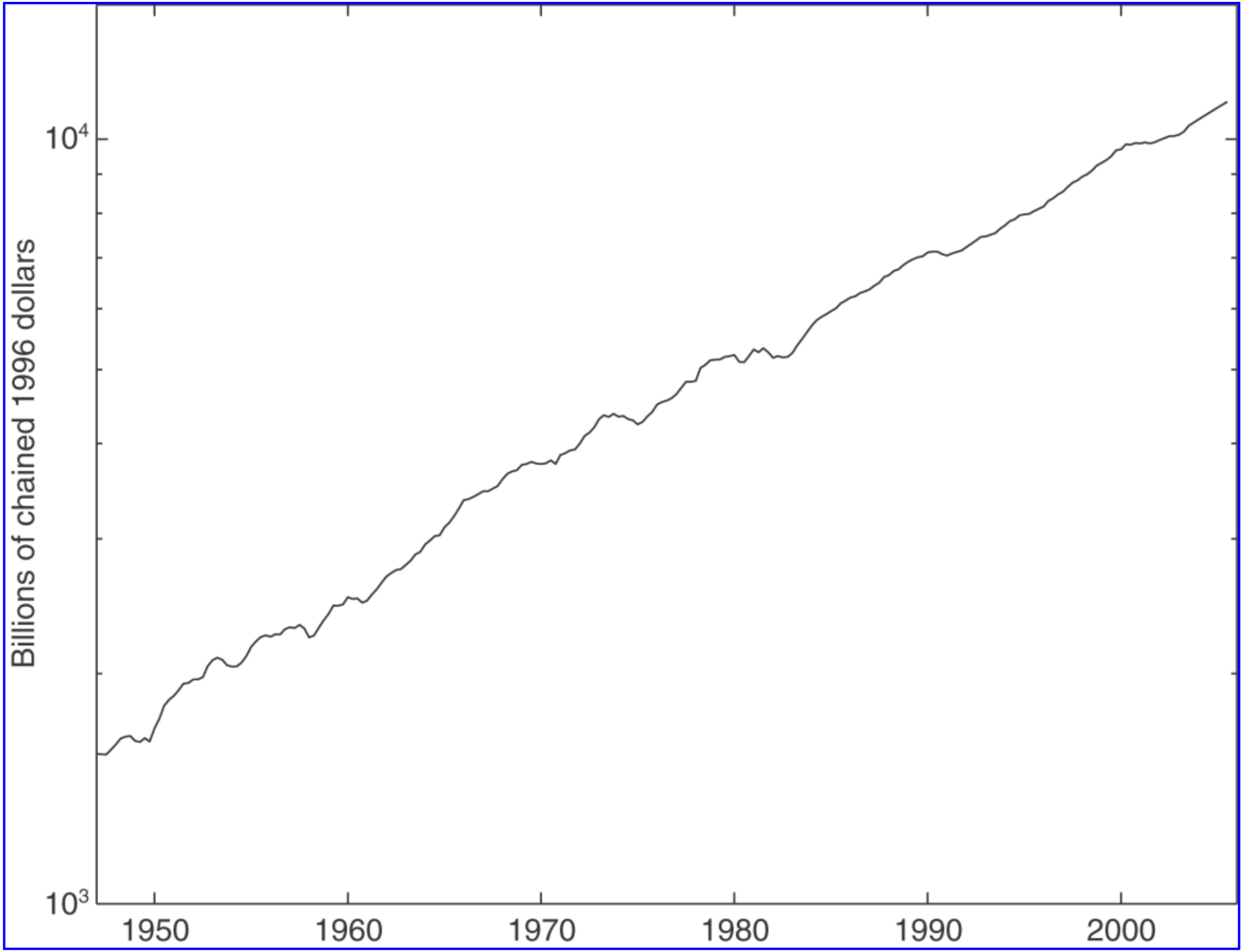
Billions of chained 1996 dollars
10^4
10^3
1950
1960
1970
1980
1990
2000

The desire to model permanent shocks in macroeconomic time series led to the development of a variety of stochastic de-trending methods. For example, Beveridge and Nelson (1981) define a stochastic trend in terms of the level to which a time series is expected to converge in the long run. Blanchard and Quah (1989) adopt a more structural approach, enforcing identifying restrictions in a vector autoregression that separate permanent shocks that drive long-run movements from the transitory disturbances which account for cyclical fluctuations.

Another popular way to measure business cycles involves application of band-pass and high-pass filters. Engle (1974) was one of the first to introduce band-pass filters to economics. In the business cycle literature, the work of Hodrick and Prescott (1997) and Baxter and King (1999) has been especially influential. Figure 2 illustrates measures of the business cycle that emerge from the Baxter–King and Hodrick–Prescott filters.

Figure 2

Filtered GDP, 1949–2003. *Sources*: Federal Reserve Economic Database and author's calculations

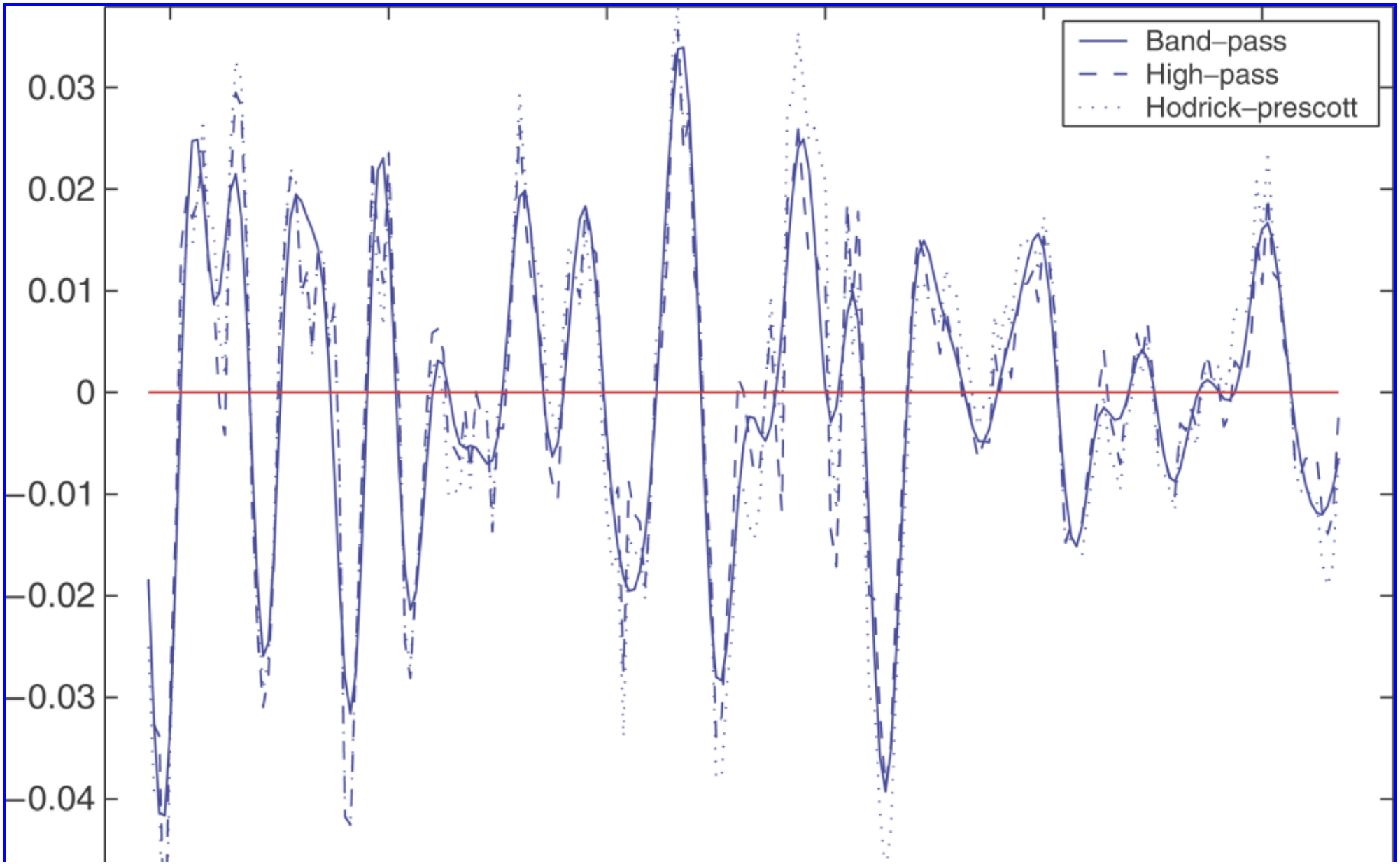

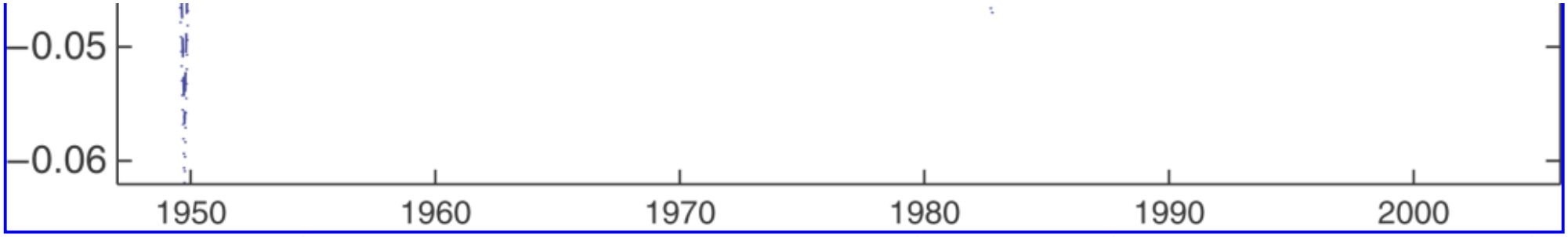

In this article, I describe how data filters work and explain the theoretical rationale that lies behind them. I focus on the problem of measuring business cycles because that is one of the principal areas of application. Many of the issues that arise in this context are also relevant for discussions of seasonal adjustment. For a review of that literature, see Fok, Franses and Paap (2006).

How data filters work

The starting point is the Cramer representation theorem. Cramer's theorem states that a covariance stationary random variable x_t can be expressed as

$$x_t - \mu_x = \int_{-\pi}^{\pi} \exp(i\omega t)\, dZ_x(\omega), \tag{1}$$

where μ_x is the mean, t indexes time, $i = \sqrt{-1}$, ω represents frequency, and $dZ_x(\omega)$ is a mean zero, complex-valued random variable that is continuous in ω. The complex variate $dZ_x(\omega)$ is uncorrelated across frequencies, and at a given frequency its variance is proportional to the spectral density $f_{xx}(\omega)$. If we integrate the spectrum across frequencies, we get the variance of x_t,

$$\sigma_x^2 = \int_{-\pi}^{\pi} f_{xx}(\omega)\, d\omega. \tag{2}$$

This theorem provides a basis for decomposing x_t and its variance by frequency. It is perfectly sensible to speak of long- and short-run variation by identifying the long run with low-frequency components and the short run with high-frequency oscillations. High frequency means that many complete cycles occur within a given time span, while low frequency means the opposite.

Baxter and King (1999) define a business cycle in terms of the periodic components $dZ_x(\omega)$. They partition x_t into three pieces: a trend, a cycle, and irregular fluctuations. Inspired by the NBER business cycle chronology, they say the business cycle consists of periodic components whose frequencies lie between 1.5 and 8 years per cycle. Those whose cycle length is longer than 8 years are identified with the trend, and the remainder are consigned to the irregular component.

The units for ω are radians per unit time. A more intuitive measure of frequency is units of time per cycle, which is given by the transformation $\lambda = 2\pi / \omega$. Often we work with quarterly data. To find the ω corresponding to a cycle length of 1.5 years, just set $\lambda_h = 6$ quarters per cycle and solve for $\omega_h = 2\pi / 6 = \pi / 3$. Similarly, the frequency corresponding to a cycle length of 8 years is $\omega_l = 2\pi / 32 = \pi / 16$. Baxter and King define the interval $[\pi/16, \pi/3]$ as ‘business cycle frequencies’. The interval $[0, \pi/16)$ corresponds to the trend, and $(\pi/3, \pi]$ defines irregular fluctuations. One nice feature of the Baxter–King filter is that it can be easily adjusted to accommodate data sampled

monthly or annually, just be resetting ω_l and ω_h.

To extract the business cycle component, we need to weigh the components $dZ_x(\omega)$ in accordance with Baxter and King's definition and integrate across frequencies,

$$x_t^B = \int_{-\pi}^{\pi} B(\omega)\exp(i\omega t)\,dZ_x(\omega), \tag{3}$$

where

$$B(\omega) = 1 \text{for } \omega \in [\pi/16, \pi/3] \text{ or } [-\pi/3, -\pi/16], = 0 \text{otherwise}. \tag{4}$$

In technical jargon, $B(\omega)$ is an example of a ‘band-pass’ filter: the filter passes periodic components that lie within a pre-specified frequency band and eliminates everything else. The Baxter–King filter suppresses all fluctuations that are too long or short to be classified as part of the business cycle and allows the remaining elements to pass through without alteration.

Many economists are more comfortable working in time domain, and for that purpose it is helpful to express the cyclical component as a two-sided moving average,

$$x_t^B = \sum_{j=-\infty}^{\infty} \beta_j (x_{t+j} - \mu_x). \tag{5}$$

The lag coefficients can be found by solving

$$\beta_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} B(\omega)\exp(i\omega j)\,d\omega. \tag{6}$$

The solution is

$$\beta_0 = \frac{\omega_h - \omega_l}{\pi},$$
$$\beta_j = \frac{\sin(\omega_h j) - \sin(\omega_l j)}{\pi j} \text{for } j \neq 0.$$

(7)

Notice that an ideal band-pass filter cannot be implemented in actual data samples because it involves infinitely many leads and lags. In practice, economists approximate x_t^B with finite-order moving averages,

$$x_t^B \doteq \sum_{j=-n}^{n} \tilde{\beta}_j (x_{t+j} - \mu_x).$$

(8)

Baxter and King (1999) and Christiano and Fitzgerald (2003) analyse how to choose the lag weights $\tilde{\beta}_j$ in order to best approximate the ideal measure for a given *n*.

For real-time applications, the two-sided nature of the filter is a drawback because the current output of the filter depends on future values of x_{t+j}, which are not yet available. Kaiser and Maravall (2001) address this problem by supplementing the filter with an auxiliary forecasting model such as a vector autoregression or univariate ARIMA model, replacing future x_{t+j} with forecasted values. This substantially reduces the approximation error near the end of samples.

That the filter is two-sided is also relevant for models that require careful attention to the timing of information. Economic hypotheses can often be formulated as a statement that some variable z_t should be uncorrelated with any variable known in period $t-1$ or earlier. These hypotheses can be examined by testing for absence of Granger causation from a collection of potential predictors to z_t. The output of a two-sided filter should never be included among those predictors, however, for that would put information about present and future conditions on the right-hand side of the regression and bias the test towards a false finding of Granger causation. Similar comments apply to the choice of instruments in generalized-method-of-moments problems. For applications like these, one-sided filters are needed in order to respect the integrity of the information flow.

While Baxter and King favour a three-part decomposition, other economists prefer a two-part classification in which the highest frequencies also count as part of the business cycle. The trend component is still defined in terms of fluctuations lasting more than eight years, but the cyclical component now consists of all oscillations lasting eight years or less. To construct this measure, we define a new filter $H(\omega)$such that

$$H(\omega) = 1 \text{for } \omega \in [\pi/16, \pi] \text{ or } [-\pi, -\pi/16], = 0 \text{otherwise}.$$

(9)

This is known as a ‘high-pass’ filter because it passes all components at frequencies higher than some pre-specified value and eliminates everything else. If we use this filter in the Cramer representation, we can extract a new measure of the business cycle by computing

$$x_t^H = \int_{-\pi}^{\pi} H(\omega) \exp(i\omega t)\, dZ_x(\omega).$$

(10)

Once again, this corresponds to a two-sided, infinite-order moving average of the original series x_t,

$$x_t^H = \int_{j=-\infty}^{\infty} \gamma_j (x_{t+j} - \mu_x), \tag{11}$$

with lag coefficients $\gamma_0 = 1 - \omega_l / \pi$ and $\gamma_j = -\sin(\omega_l j) / \pi j$. As before, this involves infinitely many leads and lags, so an approximation is needed to make it work. The approximation results of Baxter and King (1999) and Christiano and Fitzgerald (2003) apply here as well.

Hodrick and Prescott (1997) also seek a two-part decomposition of x_t. They proceed heuristically, identifying the trend τ_t and the cycle c_t by minimizing the variance of the cycle subject to a penalty for variation in the second difference of the trend,

$$\min_{\{\tau_t\}} \left\{ \sum_{t=-\infty}^{\infty} [(x_t - \tau_t)^2 + \phi(\tau_{t+1} - 2\tau_t + \tau_{t-1})^2] \right\}. \tag{12}$$

The Lagrange multiplier φ controls the smoothness of the trend component. After experimenting with US data, Hodrick and Prescott set $\phi = 1600$, a choice still used in most macroeconomic applications involving quarterly data. After differentiating (12) with respect to τ_t and rearranging the first-order condition, one finds that c_t can be expressed as an infinite-order, two-sided moving average of x_t,

$$c_t = HP(L)\, x_t = \frac{\phi(1-L)^2(1-L^{-1})^2}{1+\phi(1-L)^2(1-L^{-1})^2} x_t, \tag{13}$$

where L is the lag operator. Although Hodrick and Prescott's derivation is heuristic, King and Rebelo (1993) demonstrate that $HP(L)$ can be interpreted rigorously as an approximation to a high-pass filter with a cut-off frequency of eight years per cycle. The close connection between the two filters is also apparent in Figure 2, which shows that high-pass and Hodrick–Prescott filtered GDP are highly correlated.

Data filters for measuring of business cycles?

While data filters are very popular, there is some controversy about whether they represent appealing definitions of the business cycle. For one, there is a disconnect between the theory and macroeconomic applications, for the theory applies to stationary random processes and applications involve non-stationary variables. This is not critical, however, because the time-domain filters $\beta(L)$, $\gamma(L)$, and $HP(L)$ all embed difference operators, so business cycle components are stationary even if x_t has a unit root.

A more fundamental criticism concerns the fact that the Baxter–King definition represents a deterministic vision of the business cycle. According to a theorem of Szego, Kolmogorov, and Krein, the prediction error variance can be expressed as

$$\sigma_\varepsilon^2 = 2\pi \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f_{BC}(\omega)\, d\omega\right],$$

(14)

where $f_{BC}(\omega)$ is the spectrum for the business-cycle component (see Granger and Newbold 1986, pp. 135–6). For an ideal band-pass filter, the spectrum of x_t^B is

$$f_{BC}(\omega) = |B(\omega)|^2 f_{xx}(\omega). \tag{15}$$

Since $B(\omega) = 0$ outside of business cycle frequencies, it follows that $f_{BC}(\omega) = 0$ on a measurable interval of frequencies. But then eq. (14) implies $\sigma_\varepsilon^2 = 0$, which means that x_t^B is perfectly predictable from its own past. The same is true of measures based on ideal high-pass filters. A variable that is perfectly predictable based on its own history is said to be ‘linearly deterministic’. Thus, according to the Baxter–King definition, the business cycle is linearly deterministic.

In practice, of course, measured cycles are not perfectly predictable because actual filters only approximate the ideal. But this means that innovations in measured cycles are due solely to approximation errors in the filter, not to something intrinsic in the concept. The better the approximation, the closer the measures are to determinism.

How to square this deterministic vision with stochastic general equilibrium models is not obvious. Engle (1974), Sims (1993) and Hansen and Sargent (1993) suggest one rationale. They were interested in estimating models that are well specified at some frequencies but mis-specified at others. Engle studied linear regressions and showed how to estimate parameters by band-spectrum regression. This essentially amounts to running regressions involving band-pass filtered data, but band-pass filtering induces serial correlation in the residuals, and Engle showed how to adjust for this when calculating standard errors and other test statistics. He also developed methods for diagnosing mis-specification on particular frequency bands.

Sims (1993) and Hansen and Sargent (1993) are interested in fitting a rational-expectations model of the business cycle to data that contain seasonal fluctuations. They imagine that the model abstracted from seasonal features, as is common in practice, and they wonder whether estimates could be improved by filtering the data with a narrow band-pass filter centred on seasonal frequencies. They find that seasonal filtering does help, because otherwise parameters governing business cycle features would be distorted to fit unmodelled seasonal fluctuations. Filtering out the seasonals lets the business cycle parameters fit business cycle features.

Business cycle modellers also frequently abstract from trends, and Hansen and Sargent conjectured that the same rationale would apply to trend filtering. Cogley (2001) studies this conjecture but finds disappointing results. The double-filtering strategy common in business cycle research (which applies the filter to both the data and the model) has no effect on periodic terms in a Gaussian log likelihood, so it is irrelevant for estimation. The seasonal analogy (which filters the data but not the model) also fails, but for a different reason. The key assumption underlying the work of Engle, Sims, and Hansen and Sargent is that specification errors are confined to a narrow frequency band whose location is known a priori. That is true of the seasonal problem but not of the trend problem. Contrary to intuition, trend-specification errors spread throughout the frequency domain and are not quarantined to low frequencies. That difference explains why the promising results on seasonality do not carry over to trend filtering.

Finally, some economists question whether filter-based measures capture an important feature of business cycles. Beveridge and Nelson (1981) believe that trend reversion is a defining characteristic of the business cycle. They say that expected growth should be higher than average at the trough of a recession because agents can look forward to a period of catching up to compensate for past output losses. By the same token, expected growth should be lower than average at the peak of an expansion. Cochrane (1994) confirms that this is a feature of US business cycles by studying a vector autoregression for consumption and GDP.

Cogley and Nason (1995) consider what would happen if x_t were a random walk with drift. For a random walk, expected growth is constant regardless of whether the level is a local maximum or minimum. Because it lacks the catching-up feature, many economists would say that a random walk is acyclical. Nevertheless, when the Hodrick–Prescott filter is applied to a random walk, a large and persistent cycle emerges. Thus the Hodrick–Prescott filter can create a business cycle even if no trend reversion is present in the original data. Cogley and Nason call this a spurious cycle. Furthermore, the problem is not unique to the Hodrick–Prescott filter; Benati (2001), Murray (2003) and Osborn (1995) document similar results for band-pass filters and for other approximations to high-pass filters.

Conclusion

Christiano and Fitzgerald remark that data filters are not for everyone. They are certainly convenient for constructing rough and ready measures of the business cycle, and they produce nice pictures when applied to US data. But some economists worry about the spurious cycle problem, especially in applications to business cycle models where the existence and properties of business cycles are points to be established. In much of that literature, attention has shifted away from replicating properties of filtered data to matching the shape of impulse response functions.

See Also

- business cycle measurement
- seasonal adjustment
- spectral analysis
- structural vector autoregressions
- trend/cycle decomposition

Bibliography

Baxter, M. and King, R. 1999. Measuring business cycles: approximate band-pass filters for economic time series. *Review of Economics and Statistics* 81, 575–93.

Benati, L. 2001. Band-pass filtering, cointegration, and business cycle analysis. Working Paper No. 142, Bank of England.

Beveridge, S. and Nelson, C. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* 7, 151–74.

Blanchard, O. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.

Christiano, L. and Fitzgerald, T. 2003. The band pass filter. *International Economic Review* 44, 435–65.

Cochrane, J. 1994. Permanent and transitory components of GNP and stock prices. *Quarterly Journal of Economics* 109, 241–65.

Cogley, T. 2001. Estimating and testing rational expectations models when the trend specification is uncertain. *Journal of Economic Dynamics and Control* 25, 1485–525.

Cogley, T. and Nason, J. 1995. Effects of the Hodrick–Prescott filter on trend and difference stationary time series: implications for business cycle research. *Journal of Economic Dynamics and Control* 19, 253–78.

Engle, R. 1974. Band-spectrum regression. *International Economic Review* 15, 1–11.

Fok, D., Franses, P. and Paap, R. 2006. Comparing seasonal adjustment methods. In *Palgrave Handbook of Econometrics: Volume1; Econometric Theory*, ed. T. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Granger, C. and Newbold, P. 1986. *Forecasting Economic Time Series*. New York: Academic Press.

Hansen, L. and Sargent, T. 1993. Seasonality and approximation errors in rational expectations models. *Journal of Econometrics* 55, 21–55.

Hodrick, R. and Prescott, E. 1997. Postwar U.S. business cycles: an empirical investigation. *Journal of Money Credit and Banking* 29, 1–16.

Kaiser, R. and Maravall, A. 2001. *Measuring Business Cycles in Economic Time Series*. New York: Springer-Verlag.

King, R. and Rebelo, S. 1993. Low-frequency filtering and real business cycles. *Journal of Economic Dynamics and Control* 17, 207–31.

Murray, C. 2003. Cyclical properties of Baxter–King filtered time series. *Review of Economics and Statistics* 85, 472–76.

Osborn, D. 1995. Moving average detrending and the analysis of business cycles. *Oxford Bulletin of Economics and Statistics* 57, 547–58.

Sims, C. 1993. Rational expectations modeling with seasonally adjusted data. *Journal of Econometrics* 55, 9–19.

How to cite this article

Cogley, Timothy. "data filters." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000255>
doi:10.1057/9780230226203.0357

The New Palgrave Dictionary of Economics Online

data mining

Clinton A. Greene

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Data mining is defined by presenting an example contrasting the role of specification search in economics to its role in experimental science. Historical references are provided, along with a short review of contemporary proposals to remedy sources of and problems with data mining.

Keywords

data mining; model selection; regression analysis; specification problems in econometrics

Article

‘Data mining’ and the older word ‘fishing’ are pejorative terms for illusory or distorted statistical inference from an empirical regression model, where the distortion results from explorations of various models in a single sample of data. This process usually involves adding or dropping variables, but may involve exploring a variety of alternative nonlinear functional forms or data subsamples. Data mining properly applies as a derogatory term only when exploratory results are used for inference within the sample used in exploration. But the term is sometimes used to refer to the exploratory process itself, as economists emphasize inference over data exploration, and even use inference to discuss exploratory activities. Some take data mining to be a more serious offence when there is conscious effort to manipulate, although data mining will distort results regardless of intent.

Importance and history

Some economists consider data mining to be pervasive in applied work. But the portion subscribing to this view is unclear, since those who do so understandably retreat from applied work into economic or econometric theory. Leamer and Leonard (1983, p. 306) give voice to the view that collective data mining renders standard inference meaningless, and hence in general ‘statistical analyses are either

greatly discounted or completely ignored'. This stance may have reached a peak in the late 1970s, fuelled by an explosion in the volume of regression studies. But contemporary suspicion is still quite common. Kennedy (2003, pp. 82–3) characterizes the 'average economic regression' as perpetrating some of the worst data mining practices.

The issue was known to the originators of econometrics. Ragnar Frisch (1934) advocated methods to deal with the data mining issue which were applied into the 1950s, then neglected for two decades and reincarnated in modern form by Leamer (1983). Because Frisch found that differing but reasonable specifications could yield disparate results, he came to believe attempts at formal inference were illegitimate. Malinvaud (1966, chs. 1 and 2) provides a wonderful exposition of Frisch's methods and of why Frisch's stance was replaced by contemporary textbook assumptions. Even Haavelmo's (1944, ch. 7, sect. 17) founding statement of the contemporary inferential approach discusses data mining.

Econometrics textbooks quite properly warn against data mining, yet it is difficult to avoid and is pervasive in published work. This places the new practitioner in a difficult position. It is helpful to be armed with an understanding of the consequences of data mining and why data mining is difficult to avoid. Econometrics in the contemporary sense began when we decided that economic data could be treated as equivalent to sampling from an uncontrolled experiment (Haavelmo, 1944), borrowing from R. A. Fisher's methods for experimental data. The following illustration clarifies these issues.

An illustration

Suppose two students of the economy live in parallel universes. Both are interested in a variable y, believing the most important determinant of this variable y to be another variable x_1, but also supposing that variables x_2 and x_3 may be relevant. Their initial data-sets are identical, and they propose to model y via a linear regression model. Both start out assuming that the errors of the model (ε) are independent and normally distributed with constant variance. Thus they propose the model $y = b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon$, where the coefficients 'b_i' are to be estimated.

The first student lives in a universe in which he can generate more data via experiments. The second student must wait passively for the passage of time before she can see more data; data generated by events she does not control. Thus, the first student is confident of his science, while the second student is in the actual universe of economics.

Now suppose that in their initial regression results for the coefficient on x_1 they find the sign is the opposite of what they expected. As in standard practice they take this to imply that they have omitted an important variable. After fiddling with their specifications they find that adding a variable x_4 yields a more sensible coefficient estimate for the variable x_1. Suppose also they find that, for the coefficients on x_2 and x_3, the null hypothesis for coefficients of zero would be accepted individually (leaving the other variable coefficient unrestricted, as in a t-test). But suppose they find the joint hypothesis $(b_2 = b_3 = 0)$ would be rejected. They find the fit of the regression is penalized least by dropping the variable x_3 and do so. They have used a process of specification search to arrive at a model for y as a function of x_1, x_2 and x_4.

The first student takes the results to his professor. The professor commends the effort to learn from the

world, but corrects the student on one point. He notes that, although the estimated standard error for the coefficient on x_3 included zero, it also included (we will suppose) five, and if this coefficient is truly so far from zero then (given expected variation in x_3) the variable x_3 would have appreciable effects. So the professor tells him to run another experiment designed so that the resulting data-set is large enough (and so standard errors of coefficient estimates are small enough) to usefully distinguish between large and small values of b_3. The student does so, and publishes the results with the statistics and standard critical values treated as valid 'tests'. This is not data mining.

Now the second student takes her results to her professor. This professor says the first regression result (employing x_1, x_2 and x_3) can be treated as possibly generating test statistics drawn from standard distributions. However, in the final model (x_1, x_2 and x_4) some of the t-statistics were created by design. Since one 'fished' or fiddled with variables included in the model until the coefficient on x_1 had the correct sign, the t-statistic was drawn from a distribution such that there was 100 per cent probability it would have the 'correct' sign. Likewise the student explored specifications until the t-statistic for the coefficient on x_2 appeared to be significant. This implies for the final specification that within the interval bounded by the standard critical values (approximately plus or minus 2) the probability of the t-statistic for b_2 falling within this standard range must actually be zero, hardly a standard t-distribution. This process of modifying the model and re-estimating it using the same sample used to suggest those modifications will also affect in an unknown manner the distribution of other test statistics, even those that were not direct objects of exploration and design. These are data mined results.

Note that the two professors agree that something was potentially learned in the exploratory stage. Both students could use data exploration to reveal aspects of the first sample, but the results of exploration over this same sample could not then provide a formal test. As in any legitimate science, the first professor views taking inspiration from observation to be a process separate from confirmation or testing. The second student also hopes to have learned something from the sample, but her professor objects to treating the statistics resulting from this exploration as providing a test. The second student treated each regression as though it was a separate experiment, but regressions and their associated statistics are mere calculations that organize the data. Also note that, when these students took the initial estimate of b_1 as having the 'wrong sign', they were applying strong prior beliefs which led them to place little weight upon this empirical result. Bayesian inference provides a formal treatment of such priors.

The second student continues the consultation with her professor. The professor says these first results are not publishable because economists are interested in inference, and all she has shown is that the first model did not make sense. The professor may advise that she should first have chosen a successful regression model from the empirical literature, modifying it only slightly if at all. If the student is alert, she will notice the data available to her is identical to that in the literature, except for a few more recent observations.

So this alert student will go back to her professor and tell him she already knows the regression results will be the same as those already published, except to the extent the new data observations have some effect when averaged with the old. The test statistics will not have the usual distributions; instead, the distributions are a function of the previous results and the portion of new observations relative to those used in the previously published results. The student has discovered that, to the extent data-sets overlap, taking guidance from the regressions of other researchers is collective data mining, even if one runs only

one regression oneself. Thus collective data mining is pervasive, and the meaning of published test statistics is unclear. Only if each data-set is entirely distinct can one learn from the work of others while preserving known statistic distributions.

Contemporary practice and remedies

Three partial remedies for data mining are practised in the current literature. One is to insist upon seeing all the possible regression results a reasonable researcher might propose, supplementing imperfect ‘tests’ with a range of results. This is most associated with Leamer (1983), but we have already mentioned the earlier work of Frisch. Current practice is moving towards this approach, more often presenting multiple specifications.

A second remedy is inspired by noting that it is possible to calculate probabilities for statistics resulting from specification search, if the process begins with a model including a set of variables large enough that the true model is reasonably assumed nested within, and respecification deletes and does not add variables. An example is the general-to-specific approach. This approach is now common when specifying lag-lengths of time-series models, but in other contexts is controversial. The statistical consequences of such an approach fall under the heading of ‘pretest’ estimators discussed in most econometrics textbooks, but the best introductory discussion is found in Campos, Ericsson and Hendry (2005, Introduction, sects. 3.3–3.4). Interestingly, Hoover and Perez (1999) show that when pretest distributions are not accounted for this second remedy leads to an acceptable level of distortion.

A third remedy reserves some of the available data for ‘out-of-sample’ tests. Here one engages in specification search in one portion of the data and then tests in the reserved portion. We place ‘out-of-sample’ in quotes because this is not confirmation in a new sample. This response cannot avoid collective data mining because it is likely that among many projects the more satisfactory reserved-sample results will be selected for publication, if not by individual authors then through the collective filter of journal referees. But this remedy is useful to the individual researcher.

The first two remedies focus on data exploration, and only the third remedy adds the key scientific step of confirmation in separate data. Followers of the second remedy such as David Hendry and others of the ‘London School’ are often accused of data mining. Yet they have been the strongest proponents and practitioners of the third remedy, which provides the legitimate test in separate data, even inventing new out-of-sample tests such as for forecast encompassing. A good introduction to the second and third remedies is found in Charemza and Deadman (1997).

As noted in our discussion of the third remedy, universal adoption of these remedies cannot avoid collective data mining. Collective data mining would be avoided if upon accepting a paper the journal offered an explicit or implicit contract to accept a follow-up study. Formal and precise testing would be performed in the subsequent study employing only data not available for the initial paper. This is yet to be practised by any journal, so as a result the methodological issues remain troublesome, leaving room for vague and inconsistent norms across referees and journals. New practitioners must develop their own approaches to navigating these norms and practices, while deciding how to preserve their own sense of integrity.

See Also

- Bayesian statistics
- extreme bounds analysis
- model selection
- specification problems in econometrics
- spurious regressions

Bibliography

Campos, J., Ericsson, N. and Hendry, D., eds. 2005. *General-To-Specific Modeling*. Cheltenham, UK: Edward Elgar.

Charemza, W. and Deadman, D. 1997. *New Directions in Econometric Practice*. Cheltenham, UK: Edward Elgar.

Fisher, R. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Frisch, R. 1934. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: University Economics Institute.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(supplement), iii–115.

Hoover, K. and Perez, S. 1999. Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal* 2, 167–91.

Kennedy, P. 2003. *A Guide to Econometrics*. Cambridge, MA: MIT Press.

Leamer, E. 1983. Let's take the con out of econometrics. *American Economic Review* 73, 31–43.

Leamer, E. and Leonard, H. 1983. Reporting the fragility of regression estimates. *Review of Economics and Statistics* 65, 306–17.

Malinvaud, E. 1966. *Statistical Methods of Econometrics*. Chicago: Rand McNally and Company.

How to cite this article

Greene, Clinton A. "data mining." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000254> doi:10.1057/9780230226203.0358

The New Palgrave Dictionary of Economics Online

Davanzati, Bernardo (1529 – 1606)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

barter; Davanzati, B.; division of labour; foreign exchange markets; money; paradox of value; quantity theory of money; scarcity

Article

Merchant, classical scholar, translator and economist, Davanzati was born in Florence where, apart from a period of residence in Lyon as a merchant, he worked until his death. His contributions to economics are contained in *Notizia dei cambi* (1582) which explains the operation of the foreign exchanges, and *Lezione delle Monete* (1588), translated into English in 1696 as *A Discourse Upon Coin* presumably because of its relevance to the recoinage controversies. Besides these economic writings, Davanzati produced a history of the English Reformation (1602) and a translation of Tacitus (1637) frequently described as a masterpiece of Italian literature.

Davanzati's observations on the foreign exchanges present a detailed discussion of the origins and practice of this art classified by him as the third type of mercantile transaction, the others being barter (goods for goods) and trade (goods for money). The analysis demonstrates how exchange rates fluctuate between gold points according to the supply and demand of bills, the gold points being determined by a risk premium, transport costs and interest lost while the funds are in transit. His illustration of a foreign exchange transaction by bills of exchange involving six parties residing in Lyon and Florence (1582, pp. 62–8) has been argued by De Roover (1963, p. 113) to be so instructive that, had it been more thoroughly studied by historians and economists, ‘fewer blunders in the history of banking’ would have been made.

Davanzati's lecture on coin is one of the earliest presentations of the metallist view of the origin and nature of money. He stresses the advantages of money over barter in facilitating both the division of labour and trade of ‘superfluities’ between cities and nations. In the metallist tradition, money is defined as ‘Gold, Silver, or Copper, coin'd by Publick Authority at pleasure, and by the consent of Nations, made the Price and Measure of Things’ (1588, p. 12). Non-metallic and non-convertible money can only be made acceptable to the public through coercion. Money is therefore a human convention and its

intrinsic value is small relative to its value as means of exchange. To explain this value, Davanzati presents an early quantity theory which relates the value of stocks of commodities to the world's money stock. Although he is aware of the importance of monetary circulation (he compares it to the importance of the circulation of blood in the animal body), he does not develop a concept of its velocity. The lecture on money concludes with a forceful critique of the practice of debasing the coinage, based on analysing its consequences and illustrated with many examples of the practice. Davanzati argues that this ‘evil’ can be avoided only by making ‘Money pass according to its Intrinsick Value’ (1588, p. 24). Davanzati's lecture has also been noted because of its hints at the so-called ‘paradox of value’ and its references to elements of scarcity and usefulness in the determination of commodity prices. This and other aspects of his work were noted by Galiani (1750). Earlier his views appear to have been well received by Locke who owned, annotated and may even have inspired the Toland translation (Harrison and Laslett, 1965, p. 120).

Selected works

1582. *Notizia de'Cambi a M. Giulio del Caccia*. In *Scrittori classici Italiani di economia politica*. Parte Antica vol. 2, ed. Pietro Custodi, Milan: G.G. Destefanis, 1804.

1588. *Lezione delle Monete*. Trans. John Toland as *A Discourse Upon Coin*, London: Awnsham and Churchill, 1969.

1602. *Scisma d'Inghilterra sino alla morta della reina Maria ristretto in lingua propria Fiorentina*. Milan.

1637. *Gli Annali di C. Cornelio Tacito … con la traduzione in volgar Fiorentino*. Florence: Landini.

Bibliography

De Roover, R. 1963. *The Rise and Decline of the Medici Bank*. Cambridge, MA: Harvard University Press.

Galiani, F.,, ed. 1750. *Della Moneta*. In *Della Moneta e scritti inedite*, ed. A. Merola, Milan: Feltrinelli, 1963.

Harrison, J. and Laslett, P. 1965. *The Library of John Locke*. Oxford: Clarendon Press, 1971.

How to cite this article

Groenewegen, Peter. "Davanzati, Bernardo (1529–1606)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_D000006> doi:10.1057/9780230226203.0359

The New Palgrave Dictionary of Economics Online

Davenant, Charles (1656 – 1714)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Davenant, C.; King, G.; King–Davenant law of demand; mercantilism; political arithmetic; recoinage debates; tax administration

Article

Economist and administrator. Born in London, eldest son of William Davenant, the playwright and Poet Laureate, he was educated at Cheam School, Surrey, and entered Balliol College, Oxford, in 1671, going down in 1673 without a degree to take over the management of his father's theatre. In 1675 he wrote a tragedy, *Circe* (Davenant, 1677), but the theatre gained him little financial success. He also obtained an LL.D from Cambridge in 1675 and practised law for a short period. From 1678 to 1689 he was Commissioner of Excise. He sat as MP for St Ives from 1685 to 1688 and represented Great Bedwin in the Tory interest following the elections of 1698 and 1700. The financial consequences of his loss of office as Excise Commissioner in 1689 and unsuccessful attempts in 1692 and 1694 to obtain other positions in the revenue service appear to have inspired a career as pamphleteer, starting in 1695. Until 1702, when he again obtained preferment by being appointed Secretary to the Commission for negotiating the union between England and Scotland, he produced a steady flow of political and economic writings dealing with aspects of taxation, public debt, monetary and trade questions, foreign policy and criticisms of Whig policy in general. In June 1703 he obtained the post of Inspector-General of Exports and Imports in the Customs Office, a position he retained till his death in 1714. Most of his political and commercial writings were collected by C.E. Whitworth (1771) but two manuscript works on money and credit (Davenant, 1695b and 1696) were not published till 1942 (Evans, 1942).
Davenant's position in the history of economics rests on a variety of contributions. Initially, his work was largely depicted as typically that of an ‘adherent of the mercantile theory’ (Hughes, 1894, p. 483), but ‘Tory free trader’ (Ashley, 1900, p. 269) better describes his pronouncements on foreign trade policy as he particularly advocated the removal of trade restrictions, such as those affecting woollen exports, which benefited the landed interest by raising land values (Davenant, 1695a, pp. 16–17; 1697, pp. 98–

104). His free trade position is not unambiguous. Although Davenant's remark that ‘Trade is by its nature free, finds its own channel, and best directeth its own course .’ (1697, p. 98) is often quoted, the contradictory view that ‘it is the prudence of a state to see that [its] industry, and stock, be not diverted from things profitable to the whole, and turned upon objects unprofitable, and perhaps dangerous to the public’ (1697, p. 107) is less frequently noticed. Schumpeter's (1954, p. 196, n.4, and p. 242) depiction of Davenant's work as ‘comprehensive quasi-system’ emphasizing the interdependence of economic activity is also rather difficult to sustain, though it is possible to quote isolated remarks from Davenant's works in support. For example, Davenant's statement that ‘all trades have a mutual dependence one upon the other, and one begets another, and the loss of one frequently loses half the rest’ (1697, p. 97) cannot really be described as the general theoretical proposition it appears to be. Its only use is to provide a basis for some special pleading on behalf of the East India trade. Waddell's conclusion (1958, p. 288) that Davenant was a person neither of ‘exceptional ability, nor of any great strength of character’ and ‘a competent publicist’ rather than ‘an original thinker’ or ‘practical man of affairs’ seems a more appropriate assessment from an examination of his economic writings.

Davenant's plea for the importance of ‘political arithmetic’ or ‘the art of reasoning by figures, upon things relating to government’ (1698, p. 128) provides a further claim to fame, partly because it made more readily available the fairly sophisticated national income and expenditure estimates of his friend Gregory King (1696). Most of Davenant's political arithmetic application relates to taxation and estimating the gains from trade in terms of bullion, but he himself also made a useful contribution to the collection of international trade data as part of his duties as Inspector-General of Exports and Imports. The precise details of Davenant's association with Gregory King are not fully known, but their names are also linked in another famous ‘statistical’ exercise, the so-called King–Davenant law of demand, first noted by Thornton (1802) and Lauderdale (1804), and later extensively discussed by Jevons (1871, pp. 154–8), who on the evidence available to him cautiously attributed to Davenant the data on which the law is based (but see Barnett, 1936, pp. 6–7). However, apart from providing these data, Davenant himself characteristically drew no such analytical conclusions from this information (1698, Part II, pp. 224–5; see Creedy, 1986, for a detailed discussion).

Davenant's contributions to the recoinage debates (1695b; 1696) are less well known because they were not included in Whitworth (1771). Full recoinage was not necessary in Davenant's view when the inferior (because clipped or worn) coins were still usefully employed in small retail transactions. In addition, the detrimental effects on the exchange rate and commodity prices of the deteriorating currency were greatly exaggerated. The rise in prices, Davenant argues, could be attributed to a great many other causes; the depreciated exchange rate was more easily explained by the substantial overseas remittances induced by the European war and was therefore better remedied by floating a public loan in Holland. Although in these essays, Davenant's exposition is not always complete, Evans (1942, p. vi) regards them as containing ‘all the essential elements of the analysis of money and credit’ and integrating ‘the entire problem of currency and public finance’. Finally, Davenant's contributions to tax administration need to be recognized. They have been described as ‘translating into principles, and trying to provide a reasonable justification for the practices that the more methodical and innovating officials (such as Pepys at the Navy Office and Admiralty, and Downing and Lowndes at the Treasury …) were adopting and enforcing’ and that in these matters of administrative thinking, unlike his economics, ‘Davenant's viewpoint steadily became [dominant] in the course of the next century or so’ (Hume, 1974, p.477). His

writings also remain a useful source for much information on trade and finance over the final decades of the Stuart monarchy.

See Also

- King, Gregory

Selected works

1677. Circe, *A Tragedy. As it is acted at His Royal Highness the Duke of York's Theatre*. London: Richard Tonson.

1695a. *An Essay on Ways and Means of Supplying the War*. In Whitworth (1771, vol. 1).

1695b. *A Memorial Concerning the Coyn of England*. Reprinted in Evans (1942).

1696. *A Memoriall Concerning Creditt*. Reprinted in Evans (1942).

1697. *An Essay on the East-India Trade*. Reprinted in Whitworth (1771, vol. 1).

1698. *Discourses on the Public Revenues, and on the Trade of England in Two Parts*. Reprinted in Whitworth (1771, Part I in vol. 1; Part II in vol. 2).

Bibliography

Ashley, W.J. 1900. The Tory origin of free trade policy. In *Surveys Historic and Economic*, ed. W.J. Ashley. London: Longmans.

Barnett, G.E. 1936. *Two Tracts by Gregory King*. Baltimore: Johns Hopkins Press.

Creedy, J. 1986. On the King–Davenant ‘law’ of demand. *Scottish Journal of Political Economy* 33(3), 193–212.

Evans, G.H. 1942. *Two Manuscripts by Charles Davenant*. Baltimore: Johns Hopkins Reprints of Economic Tracts.

Hughes, D. 1894. Charles D'Avenant (1686–1714). In *Dictionary of Political Economy*, vol. 1, ed. R.H. I. Palgrave. London: Macmillan.

Hume, L.J. 1974. Charles Davenant on financial administration. *History of Political Economy* 6, 463–77.

Jevons, W.S. 1871. *Theory of Political Economy*. 4th edn, London: Macmillan, 1911.

King, G. 1696. *Natural and Political Observations and Conclusions upon the State and Condition of England*. Reprinted in Barnett (1936).

Lauderdale, J.M., Eighth Earl of. 1804. *An Inquiry into the Nature and Origin or Public Wealth*. Edinburgh. Repr. with an introduction and revisions from the 2nd edn, New York: Kelley, 1966.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen & Unwin, 1959.

Thornton, H. 1802. *An Inquiry into the Nature and Effects of the Paper Credit of Great Britain*. London.

Waddell, D. 1958. Charles Davenant (1656–1714), a biographical sketch. *Economic History Review* 11, 279–88.

Whitworth, Sir C.E. 1771. *The Political and Commercial Works of that Celebrated Writer Charles D'Avenant LL.D*. London.

How to cite this article

Groenewegen, Peter. "Davenant, Charles (1656–1714)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000007> doi:10.1057/9780230226203.0360

The New Palgrave Dictionary of Economics Online

Davenport, Herbert Joseph (1861 – 1931)

Warren J. Samuels
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

absolutist value theory; American Economic Association; Davenport, H. J.; positive economics; relativistic price theory; supply and demand; Veblen, T.

Article

Davenport was born on 10 August 1861, in Wilmington, Vermont, and died on 16 June 1931, in New York City. He commenced a professorial career at the age of 41 after having been a land speculator (initially successful, but wiped out in the Panic of 1893) and high school teacher and principal. His academic work was at the University of South Dakota, Harvard Law School, Leipzig, Paris and Chicago (Ph.D., 1898). He taught at Chicago (1902–8), Missouri (1908–16) and Cornell (1916–29). He was President of the American Economic Association in 1920.

A leading, albeit somewhat iconoclastic, economic theorist of his day, he contributed to the reformulation of microeconomics from absolutist value theory to relativistic price theory. He stressed that, while there were real forces at work in the economy, identifying them as human desires and productive capacities, price itself reflected nothing more fundamental than a temporary equation of demand and supply. Prices are not determined *by* the margins but *at* the margins. Recognizing the limits imposed by a resultant superficiality and simultaneity of determination, he felt that economists qua economists need not inquire into the formation of desires or institutions but should study the pecuniary logic of phenomena from the standpoint of price in a society dominated by the private and acquisitive point of view. His economics focused on entrepreneurial opportunity-cost adjustments and encompassed a non-normative distribution theory based directly on price theory.

While differing from his close friend Thorstein Veblen on certain substantive issues, Davenport's work nonetheless reflected the impact of Veblen's critiques of traditional theory and of the actual market economy. Emphasizing positive economics and rejecting apologetics (economic theory was not to be the monopoly of reactionaries), Davenport was willing to recognize that the search for private gain did not always conduce to social welfare, but this conclusion was not to be considered a part of economic science per se.

Selected works

1896. *Outlines of Economic Theory*. New York: Macmillan.

1897. *Outlines of Elementary Economics*. New York: Macmillan.

1908. *Value and Distribution*. Chicago: University of Chicago Press.

1913. *Economics of Enterprise*. New York: Macmillan.

1935. *The Economics of Alfred Marshall*. Ithaca: Cornell University Press.

How to cite this article

Samuels, Warren J. "Davenport, Herbert Joseph (1861–1931)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000008>
doi:10.1057/9780230226203.0361

The New Palgrave Dictionary of Economics Online

Davidson, David (1854 – 1942)

C.G. Uhr

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Böhm-Bawerk, E. von; capital gains taxation; capital theory; consumption taxation; Davidson, D.; progressive taxation; taxation of income; taxation of wealth; Wicksell, J. G. K.

Article

Born into a Jewish merchant family in Stockholm, Davidson studied law and economics at Uppsala University from 1871, became a docent in 1878, professor extraordinarius from 1880 to 1889, and then professor ordinarius for 30 years until he retired in 1919. Frequently called on to serve on parliamentary committees from 1891 to 1931, Davidson's influence was strongly felt on Sweden's monetary and tax policies, for instance the ‘gold exclusion policy’ of 1916–24.

In 1899 Davidson launched Sweden's first economic journal, *Ekonomisk Tidskrift*, to which he contributed almost all his work over 40 years as its owner and editor (in 1965 it was renamed *The Swedish Journal of Economics* and issued in English). This journal greatly stimulated economic research in Sweden with numerous contributions from, among others, Wicksell, Cassel, Lindahl, Myrdal and Ohlin.

Unlike Wicksell and Cassel, who published their works in German (later translated into English), all of Davidson's writings are in Swedish, none of them translated. This, and the fact that his work – five tracts 1878–89, over 200 articles in his journal on a variety of subjects, plus chapters in several government reports – was never systematized in treatise form, accounts for his contributions to economics having been known, until recently, only to Scandinavian academics.

In his dissertation, *Bidrag till läran om de ekonomiska lagarna för kapitalbildningen* (A Contribution to the Theory of Capital Formation), Davidson anticipated Böhm-Bawerk's *Positive Theory of Capital* (1884). To Davidson, capital was generated in the main by the unequal distribution of income. To the wealthy, increases in present goods have small and declining utility relative to that of future goods. The latter are obtained in greater quantity, variety and value by investing savings for a return – interest – in production of capital goods which, indirectly, increase productivity. This perspective inverts the first of Böhm-Bawerk's famous ‘three grounds’ for interest, and transforms the third to a marginal productivity

theory of waiting. In his later work Davidson adopted the substance of Wicksell's amendments and reconstruction of Böhm-Bawerk's capital theory.

Davidson's monetary theory is best understood from his response in articles of 1908–25 to his friend Wicksell's path-breaking work in this area. *Inter alia*, Davidson criticized Wicksell's monetary norm of price level stability as inappropriate in conditions of ‘commodity shortage’. Eventually, by 1925 Wicksell was moved to amend his norm to accommodate Davidson's critique (Uhr, 1960, chs 10 and 11).

In his early tract *Om beskattningsnormen vid inkomstskatten* (A Taxation Norm for the Income Tax, 1889), Davidson urged the replacement of Sweden's several property taxes and most of its excises by a progressive income tax with a uniquely broad base. It base was to include ‘the citizen's potential consumption power’ by levying the tax (*a*) on any increment in his net worth accrued (*whether realized or not*) between the end and the beginning of the tax year; and (*b*) also on his actual consumption spending during the year. Net worth increments accrue to a person as the value of his assets increases over that of his liabilities, due to savings, capital gains, bequests, and so on. Such gains confer potential consumption power, which should be taxed along with actual consumption spending out of income.

Over the years, aware of difficulties his proposed tax base would encounter as it called for annual balance sheet and income–consumption statements, Davidson conceded some simplifications on the tax declarations, and to taxing capital gains only when realized by the sale of value-appreciated assets. He also agreed that the tax rates levied on net worth increments would have to be lower than the rates levied on consumption expenditures.

These concessions notwithstanding, Sweden's parliament in its first comprehensive income tax of 1910 adopted only one part of Davidson's proposal. It passed a progressive tax on income as usually defined (rather than on consumption spending as such), and added to it a second title, a tax on net worth increments at rates substantially lower than on income. Largely due to Davidson, this combination of an income and a net worth increments tax has remained a standard feature in Sweden's tax system since 1910.

Selected works

1878. *Bidrag till läran om de ekonomiska lagarna för kapitalbildningen* [A Contribution to the Theory of Capital Formation]. Uppsala.

1889. *Om beskattningsnormen vid inkomstskatten* [A Taxation Norm for the Income Tax]. Uppsala.

Bibliography

Böhm-Bawerk, E. von. 1884. *The Positive Theory of Capital*. Trans. W. Smart, New York: G.E. Stechert & Co., 1930.

Heckscher, E.F. 1952. David Davidson. *International Economic Papers* 2, 111–35.

Uhr, C.G. 1960. *Economic Doctrines of Knut Wicksell*. Berkeley: University of California Press.

Uhr, C.G. 1975. *Economic Doctrines of David Davidson*. Uppsala: Studia Oeconomica Upsaliensis.

How to cite this article

Uhr, C. G. "Davidson, David (1854–1942)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000009> doi:10.1057/9780230226203.0362

The New Palgrave Dictionary of Economics Online

de Finetti, Bruno (1906 – 1985)

Giancarlo Gandolfo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

concavification; consumer theory; De Finetti, B.; exchangeability (in probability); International Institute of Mathematical Statistics; probability; statistical inference; subjective probability

Article

De Finetti was born in Innsbruck, Austria, and died in Rome. After a degree in mathematics at Milan University, he chose practical activities rather than an academic career, and worked at the Istituto Centrale di Statistica (1927–31) and then at the Assicurazioni Generali (1931–46). Only later did he turn to an academic career and win a chair in financial mathematics at Trieste University (1939); from 1954 to 1961 he held the chair in the same subject at the University of Rome and from 1961 to 1976 the chair of calculus of probabilities at the same university. He was a member of the Accademia Nazionale dei Lincei and Fellow of the International Institute of Mathematical Statistics.

De Finetti's fame rests on his contributions to probability and to decision theory, but he also worked in descriptive statistics, mathematics and economics.

Together with Ramsey and Savage, de Finetti is one of the founders of the subjectivist approach to probability theory. The first illustrations (in non-technical terms) of his conception are in (1930a) and (1931b). He considers probability as a purely subjective entity ‘as it is conceived by all of us in everyday life’. The probability that a person attributes to the occurrence of an event is nothing more or less than the measure of the person's degree of confidence (hope, fear, …) in this event actually taking place. This can be interpreted as the amount (say, 0.72) that the person deems it fair to pay (or receive) in order to receive (or pay) the amount 1 if the event in question occurs. The mathematical theory was presented in his 1935 lectures at the Institut Poincaré (1937); see also (1970) and (1972).

De Finetti also introduced the important concept of *exchangeability* in probability (1929; 1930b; 1937; 1938) and proved the theorem on exchangeable variables named after him. Exchangeability is a weaker concept than independence and has been receiving increasing attention in probability theory (in fact, the natural assumption for a Bayesian is not independence, but exchangeability). In his 1935 Poincaré lectures (1937) he also treated the relations between the subjectivist point of view and the concept of

exchangeability, which in his vision are at the basis of sound inductive reasoning and behaviour and, hence, of (statistical) decision theory (1959; 1961). It goes without saying that his position on the subject of statistical inference is fundamentally Bayesian.

In descriptive statistics he adhered to the functional concept according to which a statistic is an index selected on the basis of the single case (the aspects that one wants to stress, the aim of the statistical investigation, and so on); in (1931a) he stressed the importance of means which have the property of being associative.

Among his mathematical contributions the (1949) paper is especially interesting for economists. Here de Finetti investigates the conditions under which a concave function can be associated with a given ‘convex stratification’ (that is, a one-parameter family of convex sets, one interior to the other as the parameter varies). The author also discusses the conditions for a quasi-concave function to be transformed into a concave one by means of an increasing function. This paper started the literature on the ‘concavification’ of quasi-concave functions. As the author pointed out, these investigations also bear on consumer theory – where the convex stratification is the indifference map and the associated function is the utility function.

De Finetti also wrote on economic problems, where he stressed the importance of rigorous reasoning and verification, and emphasized the idea that the scope of economics, freed from the tangle of individual and corporative interests, should always and only be that of realizing a collective optimum (in Pareto's sense) inspired by criteria of equity (1969). An important initiative of his for the diffusion and correct application of mathematical and econometric methods in economics was the annual CIME (Centro Internazionale Matematico Estivo) seminar that he organized from 1965 to 1975; this enabled young Italian economists to benefit from courses given by Frisch, Koopmans, Malinvaud, Morishima, Zellner, to mention only a few of the lecturers.

See Also

- Bayesian statistics
- convexity
- Savage, Leonard J. (Jimmie)

Selected works

A full bibliography of de Finetti's works up to 1980 is contained in B. de Finetti, *Scritti (1926–1930)*, ed. L. Daboni et al., Padua: Cedam, 1981, with an autobiographical note.

1929. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici* (1928), Bologna: Zanichelli, 179–190.

1930a. Fondamenti logici del ragionamento probabilistico. *Bolettino dell'Unione Matematica Italiana* 9, December, 258–61.

1930b. Funzione caratteristica di un fenomeno aleatorio. *Memorie della Reale Academia dei Lincei*,

Classe di scienze fisiche, matematiche e naturali, vol. 4, fasc. 5.

1931a. Sul concetto di media. *Giornale dell'Istituto Italiano degli Attuari* 2, 369–96.

1931b. *Probabilismo. Saggio critico sulla teoria delle probabilità e sul valore della scienza*. Naples: Perrella; also in *Logos* (1931), 163–219.

1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, vol. 7, fasc. I. Trans. as ‘Foresight: its logical laws, its subjective sources’, in *Studies in Subjective Probability*, ed. H.E. Kyburg Jr. and H.E. Smokler, New York: Wiley, 1964.

1938. Sur la condition de ‘équivalence partielle’. (Conférence au Colloque consacré à la théorie des probabilités, University of Geneva, 1937.) In *Actualités Scientifiques et Industrielles*, No. 739. Paris: Herman.

1949. Sulle stratificazioni convesse. *Annali di matematica pura e applicata*, series 4, vol. 30, 173–83.

1959. La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista. In *Atti corso CIME su Induzione e Statistica* (Varenna), Rome: Cremonese.

1961. Dans quel sens la théorie de la décision est-elle et doit-elle être ‘normative’. In *Colloques internationaux du Centre National de la Récherche Scientifique*. Paris: CNRS.

1969. *Un matematico e l'economia*. Milan: F. Angeli (anthology of previously published papers).

1970. *Teoria delle probabilità. Sintesi introducttiva con appendice critica*, 2 vols. Turin: Einaudi. Trans. as *Theory of Probability*, 2 vols. New York: Wiley, 1974–5.

1972. *Probability, Induction and Statistics*. New York: Wiley (anthology of writings).

How to cite this article

Gandolfo, Giancarlo. "de Finetti, Bruno (1906–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000011> doi:10.1057/9780230226203.0368

The New Palgrave Dictionary of Economics Online

Debreu, Gerard (1921 – 2004)

Lawrence E. Blume
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys the life and work of Gerard Debreu. Although his research was largely confined to general equilibrium theory and welfare economics, the influence of his work can be seen throughout contemporary economics.

Keywords

Allais, M.; Arrow, K. J.; competitive equilibrium; contingent commodities; contract curve; convexity; core; deadweight loss; Debreu, G.; Edgeworth, F.Y.; excess demand; existence of competitive equilibrium; general equilibrium; incomplete markets; inverse function theorem; marginal rates of substitution; mathematical economics; models; multiple equilibria; natural preference models; optimality; risk aversion; Sard's theorem; Savage's subjective expected utility model; separating hyperplane theorem; social equilibrium; value theory; welfare economics

Article

Life

Gerard Debreu, the son of a Calais lace manufacturer, was born on 4 July 1921. He took his baccalauréat in 1939, just before the outbreak of the Second World War. Instead of entering university, he then began an improvised mathematics curriculum in Ambert and, later, in Grenoble. In 1941 he was admitted to the École normale supérieure, where he studied with Henri Cartan and the Bourbaki group. After D-Day he enlisted in the French Army, and served in Algeria and Germany. Returning to his studies, he completed the agrégation de mathématiques in early 1946. While pursuing his mathematical studies in Paris, he was captivated by Maurice Allais's (1943) exposition of the Walrasian general equilibrium analysis, which became the central pillar of his research programme. It was the flip of a coin which determined that he, rather than Edmond Malinvaud, would receive a travelling fellowship from the Rockefeller Foundation.

This funded a year at Harvard, Berkeley and the Cowles Commission at Chicago, followed by studies at Uppsala and, with Ragnar Frisch, in Oslo. Debreu returned to Chicago and the Cowles Commission, and moved with it to Yale in 1955 with his wife of ten years and his nine- and five-year-old daughters. A year at the Center for Advanced Study in the Behavioral Sciences at Stanford gave the Debreu family a taste for California, and in 1962 Debreu accepted a position at the University of California at Berkeley. There he remained until his retirement. Debreu became a US citizen in 1975, having been deeply moved by America's response to the Watergate affair.

Gerard Debreu received numerous honours and awards. He was a Fellow of the American Academy of Arts and Sciences (1970), vice president and president of the Econometric Society (1970, 1971), a Chevalier de la Légion d'honneur (1976), a member of the National Academy of Sciences (1977), a Distinguished Fellow of the American Economic Association (1982) and its president in 1990, a Foreign Associate of the French Académie des sciences (1984) and a Fellow of the American Association for the Advancement of Science (1984). He was awarded honorary degrees from, among many, the University of Bonn, Université de Lausanne, Northwestern University, Université des sciences sociales de Toulouse, and Yale University. Most prominent of all, in 1983 he was the recipient of the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel.

The elegance of Gerard Debreu's work was reflected in his personal style. He was also a competitive bridge player, and perhaps his first publication was a monograph on the game. In contrast to his revealed preference for the spare prose and clean, elegant arguments of the *Theory of Value* (1959) was his love of *A La Recherche du Temps Perdu*. ‘My appreciation of Proust’, he said in a 1983 *New York Times* interview, ‘is in his style, subtlety and taste. I prize conciseness very much, and that is certainly something that you cannot accuse Proust of. His compulsion, as you know, eventually killed him. I'll try to escape that fate.’ Debreu was reserved in person, but displayed a quick and subtle wit. I remember his beginning a lecture on the computation of economic equilibrium with the observation that the existence of equilibrium had been established and that now Herbert Scarf has taught us how to compute the zeros of the excess demand function. It only remains, he said, for the econometricians to estimate it, and we would be done. Gerard Debreu died in Paris on New Year's Eve 2004. His ashes were placed in a niche in the Père Lachaise cemetery, the final resting place of many of Frances's most eminent artists and intellectuals, including Marcel Proust.

Work

The influence of Gerard Debreu's work can be seen throughout contemporary economics, but his research output was largely confined to general equilibrium theory and its requirements.

The existence of competitive equilibrium

Gerard Debreu's broad fame in the economics community is due to his work on the existence of competitive equilibrium. The complexity of simultaneous price and quantity determination in multiple markets of related and unrelated goods stands in stark contrast to the cutting power of the simple Marshallian scissors of supply and demand in a market with a single good. It is certainly not obvious that a multi-market equilibrium should exist. The existence problem, open since the publication of Léon Walras's *Éléments d’économie politique pure* (1874), was first given a broad and general treatment by

Arrow and Debreu (1954a). As Arrow tells the story, in earlier work on the problem, he and Debreu had each made a mistake for which the other had a solution. It was suggested that they collude, and the outcome was displayed at the remarkable 1952 Winter Meeting of the Econometric Society in Chicago where both the Arrow and Debreu's paper (1954a) and McKenzie's (1954) paper were presented. The Arrow and Debreu ‘private ownership economy’ is today the standard reference for a general competitive model. McKenzie's treatment of technology is somewhat more special, although the two models are not directly comparable. The method of proof is to introduce a fictitious agent, a Walrasian auctioneer, whose role is to choose prices. Then the entire problem sets up like a non-cooperative game, with the added wrinkle that feasible strategies for one player may depend upon the choices of the others. Fortunately, Debreu (1952) had already established the existence of a kind of Nash equilibrium for these games, which he called a ‘social equilibrium’. This approach to the existence of equilibrium is quite different from the approach through the excess demand correspondence, which was already developed in 1954 and appears in Debreu's (1959) essential masterwork, the *Theory of Value*. The social equilibrium approach is particularly well-suited to economies in which it is difficult to get one's hands on excess demand directly, such as economies with externalities, public sector decision-making, non-convexities, and incomplete and intransitive preferences.

Welfare economics

The central question of economic analysis, the workings of the invisible hand, is formulated today as the achievement (or not) of an optimal allocation of resources. The characterization of optimality by means of marginal rates of substitution was first completed by Oscar Lange (1942). This characterization, however, is unsatisfactory for several reasons, including the facts that marginal rates of substitution may fail to exist for otherwise unremarkable preference orders, the treatment of corners is complicated, and the corresponding second-order conditions are sufficient only for local optimality. At about the same time on two different American coasts, Kenneth Arrow (1952) and Gerard Debreu (1951) proposed an alternative analysis of the relationship between equilibrium and optimality, making use of convexity assumptions and, in particular, the separating hyperplane theorem instead of the calculus. Debreu (1954b) extended his geometric analysis from finite dimensional vector spaces to linear topological vector spaces, that is, from finite to an infinite number of commodities. This advance is important for such diverse topics as financial markets, uncertainty, dynamic modelling and commodity differentiation. The first half of Debreu (1951) establishes the classical welfare theorems, relying only on convexity and topological assumptions on preferences. The second half of the paper introduces the coefficient of resource utilization, a measure of deadweight loss. Debreu (1954a) applied this measure to the deadweight loss associated with tax-subsidy schemes, a measure that has been implemented empirically by Farrell (1957) and Whalley (1976) to study productive efficiency and the deadweight loss of alternative tax schemes. A comparison of the Debreu coefficient with other measures of deadweight loss, including that of his contemporary M. Boiteux at the École normale supérieure, can be found in Diewert (1981).

The theory of value

Debreu's *Theory of Value* (1959) is not simply about the existence and optimality of equilibrium. It is a statement of method that has profoundly changed the way economics is practised. For this alone it is among the most original books of 20th-century economic thought. Most economists identify Debreu with mathematics, manipulating formulas and proving theorems. But for Debreu this, although pleasurable, was the easy part of economic theory. He once told me that it was harder to be an economist than a mathematician. A mathematician had to be correct and elegant; but an economist had to be all that and also interesting. The power of a model lies in the economist's ability to interpret with it, and this is the point of all the ‘elegance’ and clarity in Debreu's exposition. In the preface, he writes (1959, p. x), ‘Allegiance to rigor dictates the axiomatic form of the analysis where the theory, in the strict sense, is logically entirely disconnected from its interpretations. … Such a dichotomy reveals all the assumptions and the logical structure of the analysis.’ Debreu taught that the separation of logical analysis from interpretation is crucial to good theory. The logic of market equilibrium is independent of what commodities actually are, except in so far as what they are may suggest additional structure on the primitives of the equilibrium model. This is most clearly demonstrated in Chapter 7. Here Debreu reinterprets the model by appending to the description of commodities the state of nature in which it is available. The use of Arrow's (1953) contingent commodities ‘allows one to obtain a theory of uncertainty free from any probability concept and formally identical with the theory of certainty developed in the preceding chapters’ (1959, p. 98). Three pages later, Debreu observes that the convexity assumptions required by the theoretical analysis could be understood as risk aversion. And although Debreu stops here, it is not a big step to observe that natural preference models, like Savage's subjective expected utility model, lead to an additive structure for preferences that may have implications for the nature of equilibrium.

Large economies and the core

Competitive equilibrium requires prices, and prices in turn already require a sophisticated set of market institutions. Nonetheless, ‘general’ is a key word in the phrase general competitive equilibrium. The principle behind the abstract treatment of market equilibrium is that the workings of supply and demand are more or less the same whether the market under discussion is a modern financial market in London or New York or a village market of farmers and petty traders in India or East Africa. This is quite a claim. Support for this idea comes from the fact that the Walrasian outcome from markets with quoted prices can also be supported by a seemingly more fundamental equilibrium concept that makes no mention of prices at all: the core.

The core comes from F. Y. Edgeworth's *Mathematical Psychics* (1881), in which the contract curve is first introduced, and which, remarkably, undertakes a limit analysis of the economy with two types of traders and two goods. Edgeworth showed that the set of core allocations shrinks to the set of competitive equilibria as the number of agents becomes large. Debreu and Scarf (1963) pick up this question and quickly dispatch it for replica economies, which are generalizations of the large population structures Edgeworth studied. Immediately thereafter came Aumann's (1964) equivalence theorem for the core and equilibrium set of an economy with a continuum of agents, which, among other things, launched the subject of economies described by a measure space of agents. These developments are important because perfect competition is most naturally expressed as a large economy (large number of agents) phenomenon, and because empirical descriptions of large markets may be best described by

distributions on the space of agent characteristics.

Smooth economies

It is often said that Gerard Debreu took the calculus out of economics with his topological equilibrium analysis of the 1950s and early 1960s. If so, it returned with a vengeance in his 1970 and 1972 papers on economies with differentiable excess demand. It has been clear since the Edgeworth box that economies with multiple equilibria are inescapable, a fundamental indeterminacy of the analysis. One can easily construct exchange economies with a continuum of equilibria. But how far does it extend? Is this the norm or are these economies pathological? In a path-breaking series of papers Debreu drew the line between normal and bizarre. He demonstrated that if individual demand is differentiable, then the ‘generic’ case is one in which there are only a finite number of isolated equilibria; that is, equilibria are locally unique. ‘Economies with a Finite Set of Equilibria’, his 1970 paper, is particularly striking in its simplicity. Once it is determined that an economy is regular, the main result follows from the inverse function theorem – surely a result known to anyone who has taken a multivariate calculus course. Only the deeper fact that regularity is generic requires more advanced tools such as Sard's theorem. Again, Debreu's intuition was geometric. In lectures this was explained with a simple diagram. Subsequent work has used the tools of differential topology to uncover the deeper structure of the equilibrium manifold, the graph of the equilibrium correspondence. These tools are also of fundamental importance for economies with incomplete markets. With incomplete markets and financial assets rather than real assets, indeterminacy is no longer unusual, and this is of critical importance for applications to macroeconomics and finance. Some of this work is surveyed in the monographs of Balasko (1988) and Mas-Colell (1985).

Excess demand

It is important to ask of any theory, ‘what can it say?’ That is, what kinds of predictions will the theory make, and what patterns in data will contradict the theory? In general equilibrium theory this question was first asked by Sonnenschein (1972) in the following way: in exchange economies, the market excess demand function satisfies the restrictions of continuity, homogeneity and Walras's Law. This and a boundary condition is enough to prove the existence of equilibrium prices. Sonnenschein asked if excess demand functions had any additional structure beyond these three requirements. Sonnenschein (1972), Mantel (1974) and Debreu (1974), with an important extension by (Mas-Colell, 1977), showed that the answer is ‘no’. Any function defined for strictly positive prices and satisfying these three conditions is identical up to boundary behaviour with an excess demand function for an exchange economy containing no more agents than goods, each agent with continuous, strictly convex and monotonic preferences. Thus the hypothesis of utility maximization in exchange economies, with no additional assumptions about agents’ characteristics, will place few restrictions on comparative static results or on the nature of the equilibrium price set.

These results are often incorrectly interpreted to mean that general equilibrium theory is empty, that it predicts nothing. This is entirely incorrect. General equilibrium theory is not so much a theory as a theoretical framework within which theories can be built by making explicit assumptions about the

nature of tastes, technologies and endowments. To say that the framework does not limit market behaviour without any assumptions about its primitive objects is to say that the framework is maximally expressive. Its power to predict market behaviour comes from assumptions about the population of agents participating in the market. The so-called ‘anything goes’ theorems simply imply that more results will require more assumptions about the preferences and endowments of agents. It had been Debreu's hope that restrictions on the distributions of agents’ characteristics would lead to interesting conclusions: but progress has been slow.

Other contributions

Debreu has produced seminal papers in areas of economic theory other than general equilibrium analysis. Which preference orders have a continuous utility representation? This question is answered by (1954c). Which preferences have additive separable representations? Debreu's (1958) answer to this very difficult question is topological in nature, and quite distinct from the algebraic answers found in the mathematical psychology literature.
Debreu was exceptional in the classroom and in seminar. His lectures were crystalline, elegantly shaped, and parsimonious. Often they were too clear; we students left the class convinced we understood, only to discover on problem sets how subtle were the arguments that had seemed so obvious on the blackboard. Debreu's expository writings, especially his Nobel Address (1984), are required for everyone with a serious interest in contemporary economics.

Conclusion

It is impossible to imagine modern economics without the scholarship of Gerard Debreu. Debreu, Kenneth Arrow and a few others who solved the big open questions of general equilibrium theory in the 1950s had an impact that reached far beyond the confines of formal competitive analysis. They were responsible for making formal modelling a requirement for serious economic analysis of any kind. Formal modelling is not merely a theoretical discourse; the availability of formal models requires a means for the models to confront data. Modern econometrics is inconceivable without the idea of formal modelling as a strategy of enquiry. It is not by accident that, just as the general equilibrium theory was taking off at the Cowles Commission in the 1950s, so too was modern econometrics. The contributions of the ‘mathematical economists’ launched a revolution that has touched on every area of economic practice.

See Also

- core convergence
- cores
- general equilibrium
- welfare economics

Selected works

1951. The coefficient of resource utilization. *Econometrica* 19, 273–92.

1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences* 38, 597–607.

1954a. (With K.J. Arrow.) Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

1954b. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40, 584–92.

1954c. A classical tax-subsidy problem. *Econometrica* 22, 14–22.

1954d. Representation of a preference ordering by a numerical function. In *Decision Processes*, ed. R. M. Thrall, C.H. Coombs and R.L. Davis. New York: Wiley.

1958. Stochastic choice and cardinal utility. *Econometrica* 26, 440–4.

1959. *Theory of Value*. New York: Wiley. Repr. New Haven: Yale University Press, 1971.

1963. (With and H. Scarf.) A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

1972. Smooth preferences. *Econometrica* 40, 603–15.

1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–21.

1984. Economic theory in the mathematical mode. *American Economic Review* 74, 267–78.

Bibliography

Arrow, K.J. 1952. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie* 40, 41–8, Cahiers du CNRS.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Balasko, Y. 1988. *Foundations of the Theory of General Equilibrium*. Boston: Academic.

Diewert, W.E. 1981. The measurement of deadweight loss revisited. *Econometrica* 49, 1225–44.

Edgeworth, F.Y. 1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: C. Kegan Paul.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society* 120, 253–91.

Lange, O. 1942. The foundations of welfare economics. *Econometrica* 10, 215–28.

Mantel, R. 1974. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7, 348–53.

Mas-Colell, A. 1977. On the equilibrium price set of an exchange economy. *Journal of Mathematical Economics* 4, 117–26.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: A Differentiable Approach*. Cambridge: Cambridge University Press.

McKenzie, L. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22, 147–61.

Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40, 549–63.

Walras, M.E.L. 1974. *Éléments d'économie politique pure, ou théorie de la richesse sociale*. Lausanne: Corbaz.

Whalley, J. 1976. Some general equilibrium analysis applied to fiscal harmonization in the European community. *European Economic Review* 8, 290–312.

How to cite this article

Blume, Lawrence E. " Debreu, Gerard (1921–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000238> doi:10.1057/9780230226203.0363

The New Palgrave Dictionary of Economics Online

decentralization

E. Malinvaud
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Central planning is inefficient because it lacks incentives and is poorly informed. Complete decentralization risks being inequitable and also inefficient because markets are incomplete and public goods may be neglected. Intermediate systems can overcome these difficulties to the extent that planning mechanisms can mimic the market system while avoiding its deficiencies, public goods can be successfully delivered at the local level, and incentives to report and behave faithfully and to avoid free riding can be secured.

Keywords

autonomy; central planning; competitive equilibrium; decentralization; decomposability; free rider problem; Hayek, F.; incentive compatibility; incentives; information; non-economic motivation; Lange, O.; Malinvaud, E.; misreporting; Pareto efficiency; principal and agent; private information; prospective indices; public goods; resource allocation mechanisms; social choice; socialism; Mises, L. von

Article

The main question to be answered by the theory of resource allocation, or by the theory of economic organization, concerns the performances of alternative systems characterized by different degrees of centralization of decision taking. A fully centralized system runs the risk of being inefficient because it does not create proper economic incentives and the centre is poorly informed. A pure market system with its high degree of decentralization runs the risk of bringing inequitable results and being inefficient because markets can never be complete, externalities exist and public wants tend to be neglected. Can these risks be avoided within the two opposite extremes of pure centralization or full decentralization? Can intermediate systems better resolve the difficulties? And if so, how?
Basic to the discussion are two features: the nature of the *information* held by various agents, and the *incentives* that should lead them to behave in conformity with collective requirements. These features

and the issue of decentralization do not only appear for full economic systems, which this entry will consider, but also for the internal organization of firms or communities. They are stylized in the principal–agent problem: which rules should determine how to share the proceeds of an activity between the principal owner and his better-informed agent? (Ross, 1973; Grossman and Hart, 1983).

For the clarification of the complex issues involved, theory starts from a model of the conditions of economic activity. It makes assumptions such that, independently of economic organization, there exists a best outcome, or at least a set of 'optimal' outcomes. It then asks how well alternative forms of organization succeed in finding, implementing or at least approaching this best outcome or set of optimal outcomes.

By so doing, the theory discussed here neglects two related questions: how to determine what should be considered as 'the best' outcome in a society with many individuals, and which non-economic considerations interfere with the issue of decentralization? The theory of social choice shows the fundamental difficulty of the first question (Arrow, 1951), which is avoided when optimality is identified with Pareto efficiency. As for the second, philosophers may find in human nature or in the aims pursued by human societies reasons that favour some organization, beyond its economic performance; in particular, the right of individuals to autonomy appears fundamental in Western culture and is an important justification of decentralization, and even of the market system for such economists as Hayek (1944).

Formal concepts and preliminaries

The following conceptual apparatus, although not yet common, is well suited to the purpose (see Hurwicz, 1960; Mount and Reiter, 1974).

An *economic environment* is defined by a set of commodities and their possible uses, by a list of agents and their characteristics (technology, endowments, preferences, and so on), and by an initial information structure (what each agent knows). The feasible set of economic environments defines 'the economy'.

An important property of an economy is its higher or lower degree of *decomposability*, which concerns agents' characteristics and the information structure. The highest decomposability is assumed in competitive equilibrium theory, where all consumption is private, no external effect exists and a *private information structure* prevails (each agent perfectly knows its own characteristics and the situation on all markets, but nothing else). But models with public goods, for instance, usually admit some decomposability, which matters for the validity of the results.

An *optimality correspondence* P: $E \rightarrow A$ defines which vectors of actions simultaneously taken by the various agents are optimal when the economic environment is e, i.e. optimal vectors belong to $P(e)$ (clearly, E is the set of feasible e, that is 'the economy', while A is the set of feasible vectors a, each one of them defining the actions taken by all the agents). For instance $P(e)$ may be the set of Pareto efficient vectors. But in the theory discussed here, it is often more narrowly defined so as to take equity considerations into account: a social utility function may have to be maximized or a rule on the consumers 'income distribution' satisfied.

A *resource allocation mechanism* $f: E \rightarrow A$ should select one $a = f(e)$ for each environment e (in some cases f may be multivalued, i.e. become a correspondence). The best formalized mechanism is the competitive equilibrium of a 'private ownership economy'. A study of decentralization requires a careful

specification of the mechanism, which is typically viewed as operating in two stages: first, an iterative exchange of messages, usually between the agents and a centre, resulting in a message correspondence $g: E \to M$ (the message $m = g(e)$ specifies what information about *e* has been collected at the centre), second an outcome function $h: M \to \mathcal{A}$. For instance, the competitive mechanism is often specified as resulting from the tâtonnement process, in which an auctioneer learns which demands and supplies are announced at various proposed vectors of prices, and searches for the equilibrium prices; once these prices are found, the outcome function gives the equilibrium exchanges, hence productions and consumptions.

The performances of alternative mechanisms of course concern the final result: one must know whether the outcome $f(e)$ belongs to the optimal set $P(e)$ for all environments in E, or at least for a precise subset of E, and how close it is to $P(e)$ otherwise. But interesting performances also concern intermediate features of the mechanism, which usually is iterative. At step t the previously collected message m_{t-1} is enriched according to $m_t = g_t(m_{t-1}, e)$ and, if necessary, the process could end by $a = h_t(m_t)$. In a *finite* procedure it does end at T with $m = m_T$ and $h(m) = h_T(m_T)$; but most mechanisms assume an infinite sequence of m_t for $t = 1, 2 \ldots$ ad infinitum. One must then know whether and how $h_t(m_t)$ approaches $P(e)$, monotonically or otherwise. Since the transmission of information is costly, the nature and size of the message space M_t to which m_t belongs are also important characteristics (Mount and Reiter, 1974).

The planning problem

Early in this century many economists objected to socialist planning programmes that could not be implemented, because they unrealistically assumed that a central administration could have the knowledge and computing power required for an efficient control of economic activity. The leading figure was L. von Mises (1920 in particular); but Hayek (1935) was first to emphasize the problems raised by the decentralization of information. Socialist economists answered that decentralized mechanisms could operate, either mimicking the market system while being free of its deficiencies (Lange, 1936) or using different well conceived modes of information gathering (Taylor, 1929). The debate was, in the interwar years, the subject of the '*economic theory of socialism*'. (For a well-documented survey, see Bergson, 1948.)

The problem was again taken up during the 1960s, in particular because the logic of efficient planning was discussed in Eastern and Western Europe (Arrow and Hurwicz, 1960; Kornai, 1967; Malinvaud, 1967; Heal, 1973). Many planning procedures were rigorously studied as resource allocation mechanisms. Their definition implied an iterative exchange of information between a Central Planning Board and firms, sometimes also representative consumers. The additional messages provided by the function g_t at step t then consisted of *prospective indices* announced by the Board, for instance prices for the various commodities, and replies called *proposals* sent to the Board by firms and other agents, for instance preferred techniques of production and their input requirements, or supplies and demands.

In this discussion it is common to distinguish between price-guided procedures, in which the Board announces price vectors, and other procedures, in which quantity indices or targets worked out at the centre play a more or less important role. The nature and properties of the environment are then found to

be crucial for the determination of the relative performances of alternative procedures, in particular of price-guided against quantity-guided procedures (Weitzman, 1974).

The analytical study of various procedures usually assumes that decentralized agents exactly follow specified rules for the determination of their proposals and so faithfully reveal part of their private information. Some procedures are then found to be efficient and to permit achievement of distributive objectives. But efficiency is typically easier precisely in those environments that are also favourable to the efficiency of free competition. Besides the possibility of incorrect reporting, the main difficulty concerning the relevance of this literature is to know whether its models provide an approximate representation of procedures that are actually used, or at least administratively feasible. Manove (1976) has made this claim for his representation of Soviet planning.

The public good problem

The most relevant field of application may very well be the theory of public goods. Decisions concerning the provision of public services and their financing cannot be fully decentralized; but the knowledge required is dispersed and must be gathered in a proper way. Hence even the positive theory of public goods was often formulated along lines that look like those of planning procedures (Malinvaud, 1971). The same remark applies to decisions concerning public projects with large fixed costs, even if their output is privately consumed.

Considered as a planning procedure, the search for the best decision is often viewed as involving ‘prospective indices’ that define amounts of service to be provided, ask for corresponding individual marginal utilities and look whether the sum of the latter would cover the cost of additional service. This is compatible with the dual arrangement for private goods, prices being announced, supplies and demands being the replies. The procedure is then quantity-guided for public goods and price-guided for private goods (Drèze and Vallée Poussin, 1971).

The collective consumption of many types of public goods is not really national but limited to local communities (primary education, city transports, and so on). Administrative science sees the decentralization issue as being to know at which level should decisions be taken: at the national level, so as to distribute fairly these services among communities, or at the local level, so as to permit better adaptation to local needs and wishes. Economists do not seem to have contributed to this issue; their discussion of local public goods assumes full administrative decentralization (Tiebout, 1956).

Incentive compatibility

The study of a decentralized system has to consider whether the actual reports and behaviour of individual agents do not deviate from what they are supposed to report and do; in case of deviations, how are the performances of the system affected? The problem is serious: once the rules of organization and decisions are known, individual agents may benefit from misreporting their private information or from behaving in a way that, although deviant, does not clearly appear to be so. In other words, they may act as players in a game, rather than as members of a team, and this may be more or less detrimental for the optimality of the final result.

The problem has long been known for organizations in which some agents do not individually benefit

from what is achieved and therefore lack the incentive to do their best. Monopolistic or other non-competitive behaviour is often interpreted as a breach of the normal rules of resource allocation. In the theory of public good the ‘free rider problem’ occurs as soon as some individuals, having a high marginal utility for the public good, would benefit from hiding this fact so as to contribute little to the financing of the good.

Study of the problem has been active during the past two decades (Green and Laffont, 1979). The fundamental difficulty has been exhibited by such results as the following one: in the classical model of an exchange economy with a finite number of consumers, no procedure can be found that would necessarily lead to a Pareto efficient result in which individuals, acting as players in a non-cooperative game, would faithfully report (Hurwicz, 1972). However, misreporting may not prevent a procedure from eventually leading to an optimum, as was proved in a number of cases.

Experiments moreover show that the game-theoretic approach to the incentive problem may be misleading because it neglects non-economic motivations that individuals may find for accepting a team-like behaviour and therefore for faithfully reporting (Smith, 1980).

Bibliography

Arrow, K. 1951. *Social Choice and Individual Values*. New York: Wiley.

Arrow, K. and Hurwicz, L. 1960. Decentralization and computation in resource allocation. In *Essays in Economics and Econometrics in Honour of Harold Hotelling*, ed. R. Pfouts. Chapel Hill: University of North Carolina Press.

Bergson, A. 1948. Socialist economics. In *A Survey of Contemporary Economics*, ed. H. Ellis. Philadelphia: Blakiston.

Drèze, J. and Vallée Poussin, D. de la. 1971. A tâtonnement process for public goods. *Review of Economic Studies* 38, 133–50.

Green, J. and Laffont, J.-J. 1979. *Incentives in Public Decision-Making*. Amsterdam: North-Holland.

Grossman, S. and Hart, O. 1983. An analysis of the principal-agent problem. *Econometrica* 51(1), 7–45.

Hayek, F. 1935. Socialist calculation: the state of the debate. In *Collectivist Economic Planning*, ed. F. Hayek. London: G. Routledge & Sons.

Hayek, F. 1944. *The Road to Serfdom*. Chicago: University of Chicago Press.

Heal, G. 1973. *The Theory of Economic Planning*. Amsterdam: North-Holland.

Hurwicz, L. 1960. Optimality and information efficiency in resource allocation processes. In

Mathematical Methods in the Social Sciences, ed. K.J. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and Organization*, ed. R. Radner and C. McGuire. Amsterdam: North-Holland.

Kornai, J. 1967. *Mathematical Planning of Structural Decisions*. Amsterdam: North-Holland.

Lange, O. 1936. On the economic theory of socialism. *Review of Economic Studies* 4, 53–71, 123–42.

Malinvaud, E. 1967. Decentralized procedures for planning. In Malinvaud, E. and Bacharach, M., *Activity Analysis in the Theory of Growth and Planning*, Macmillan: London.

Malinvaud, E. 1971. A planning approach to the public good problem. *Swedish Journal of Economics* 11, 96–112.

Manove, M. 1976. Soviet pricing, profit and technological choice. *Review of Economic Studies* 43, 413–21.

Mises, L. von. 1920. Economic calculation in the socialist commonwealth. First published in German in *Archiv für Sozialwissenshaft*, April; English translation in *Collectivist Economic Planning*, ed. F. Hayek, London: G. Routledge & Sons, 1935.

Mount, K. and Reiter, S. 1974. The informational size of message spaces. *Journal of Economic Theory* 8 (2), 161–92.

Ross, S. 1973. The economic theory of agency: the principal's problem. *American Economic Review* 63 (2), 134–9.

Smith, V. 1980. Experiments with a decentralized mechanism for public good decisions. *American Economic Review* 70, 584–99.

Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19, 1–8.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Weitzman, M. 1974. Prices versus quantities. *Review of Economic Studies* 41, 477–91.

How to cite this article

Malinvaud, E. "decentralization." The New Palgrave Dictionary of Economics. Second Edition. Eds.

Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000040> doi:10.1057/9780230226203.0364

The New Palgrave Dictionary of Economics Online

decision theory in econometrics

Keisuke Hirano
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The decision-theoretic approach to statistics and econometrics explicitly specifies a set of models under consideration, a set of actions that can be taken, and a loss function that quantifies the value to the decision-maker of applying a particular action when a particular model holds. Decision rules, or procedures, map data into actions, and can be ordered according to their Bayes, minmax, or minmax regret risks. Large sample approximations can be used to approximate complicated decision problems with simpler ones that are easier to solve. Some examples of applications of decision theory in econometrics are discussed.

Keywords

admissibility criterion; auction models; Bayes risk; Bayes rule; computational methods; decision rules; decision theory in econometrics; instrumental variables; local asymptotic normality (LAN); Markov chain Monte Carlo methods; maximum likelihood; minmax principle; minmax-regret principle; nonparametric density estimation; nonparametric models; nonparametric regression; point estimators; portfolio choice; Savage, L. J.; search models; semiparametric models; statistical decision theory; time series models; treatment assignment; white noise

Article

The decision-theoretic approach to statistics and econometrics explicitly specifies a set of models under consideration, a set of actions available to the analyst, and a loss function (or, equivalently, a utility function) that quantifies the value to the decision-maker of applying a particular action when a particular model holds. Decision rules, or procedures, map data into actions, and can be evaluated on the basis of their expected loss.
Abraham Wald, in a series of papers beginning with Wald (1939) and culminating in the monograph (Wald, 1950), developed statistical decision theory as an extension of the Neyman–Pearson theory of

testing. It has since played a major role in statistical theory for point estimation, hypothesis testing, and forecasting, especially in the construction of ‘optimal’ procedures. Some textbooks such as Ferguson (1967) and Berger (1985) emphasize statistical decision theory as a foundation for statistics. But the decision theory framework is sufficiently flexible that it can be used for many empirical applications that do not fit neatly into the usual statistical set-ups. Some examples are discussed below.

Like the Neyman–Pearson theory, Wald's approach emphasizes evaluating the performance of a decision rule under various possible parameter values. There does not always exist a single rule that dominates all others uniformly over the parameter space, just as there does not always exist a uniformly most powerful test in the special case of hypothesis testing. Wald, who also made contributions to game theory, proposed to evaluate a procedure by its minmax risk – the worst-case expected loss over the parameter space. Savage (1951) discusses the minmax principle and suggests an alternative, the minmax-regret principle. Alternatively, one can place a probability measure on the parameter space, and evaluate rules by their weighted average (Bayes) risk.

Basic framework

In Wald's basic framework, we start with a set of actions $\mathcal{A}$, and a parameter space Θ, which characterizes the set of models under consideration. A loss function $L(\theta, a)$ gives the loss or disutility suffered from taking action $a \in \mathcal{A}$ when the parameter is $\theta \in \Theta$. The decision maker observes some random variable Z, distributed according to a probability measure P_θ when θ is the ‘true’ parameter. Here, the parameter space Θ could be finite-dimensional (corresponding to a parametric family of distributions) or infinite-dimensional (corresponding to semiparametric and nonparametric models). The observed random variable Z could be a vector, as for example in the situation of observing a random sample of size n from some distribution. Often, the set of possible probability measures $\{P_\theta : \theta \in \Theta\}$ is called a *statistical experiment*.

A decision rule or procedure $d(z)$ maps observations on Z into actions. In some cases, it is useful to allow for randomization over the actions. A randomized decision rule is a mapping from observations into probability measures over the action space. A simpler, usually equivalent formulation is to consider rules $\delta(z, u)$ which are allowed to depend on the observed value z and the value u of a random variable U, distributed standard uniform independently of Z. The *risk*, or expected loss, of a decision rule δ under θ is defined as

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(Z, U))] = \int_0^1 \int L(\theta, \delta(z, u))\, dP_\theta(z)\, du.$$

A rule δ is *admissible* if there exists no other rule δ' with

$$R(\theta, \delta') \leq R(\theta, \delta), \ \forall \theta \in \Theta,$$

and

$$R(\theta, \delta') < (\theta, \delta) \text{for some } \theta.$$

Ordering decision rules

In general, there are many admissible decision rules, which may do well in different parts of the parameter space. Thus, while the admissibility criterion eliminates obviously inferior rules, it may not provide concrete guidance on how to ‘solve’ the decision problem. Additional criteria can help by providing a sharper partial ordering of decision rules.

One way to rank decision rules is to average their risk over the parameter space. Let Π be a probability measure on Θ. The *Bayes risk* of a decision rule δ is

$$r(\Pi, \delta) = \int R(\theta, \delta)\, d\Pi(\theta) .$$

A rule is a *Bayes rule* if it minimizes this weighted average risk. Let the probabilities P_θ have densities p_θ with respect to some dominating measure, and let the prior Π have density π. Typically, a Bayes rule can be implemented by choosing, for any given observed data *z*, the action that minimizes the posterior expected loss

$$\int L(\theta, a)\, d\Pi(\theta|z),$$

where $\Pi(\theta|z)$ is the posterior distribution with density

$$\pi(\theta|z) = \frac{\pi(\theta)\, p_\theta(z)}{\int p_\theta(z)\, d\Pi(\theta)}.$$

There is a close connection between the admissible rules and the Bayes rules. If the parameter set is finite, a Bayes rule for a prior that places positive probability on every element of Θ is admissible. Furthermore, 'complete class theorems' give results in the opposite direction. In particular, if the parameter set is finite, any admissible rule is Bayes for some prior distribution. If Θ is not finite, some care needs to be taken to make a precise statement of the relationship between the admissible and Bayes rules; see for example Ferguson (1967).

An alternative ordering is based on the worst-case risk $\sup_{\theta\in\Theta} R(\theta, \delta)$. A *minmax* rule δ_m satisfies

$$\sup_{\theta\in\Theta} R(\theta, \delta_m) = \inf_{\delta} \sup_{\theta\in\Theta} R(\theta, \delta).$$

In general, a minmax rule need not be admissible.

A closely related criterion is the *minmax regret* criterion. The regret loss of a rule is the difference between its loss and the loss of the best possible action under θ:

$$L_r(\theta, a) = L(\theta, a) - \inf_{a\in A} L(\theta, a).$$

We can then define regret risk as $R_r(\theta, \delta)=E_\theta(L_r(\theta, \delta(Z, U))$. The *minmax regret rule* minimizes the worst-case regret risk. This rule was suggested by Savage (1951) as an alternative to the minmax criterion. He argued that in cases where the minmax criterion is unduly conservative, minmax regret rules can be reasonable.

Savage (1954) showed that a decision-maker who satisfied certain axioms of coherent behaviour would act as if she placed a prior on the parameter space and minimized posterior expected loss. Gilboa and Schmeidler (1989) showed that, under a different set of axioms, a decision-maker would follow the minmax principle.

Calculation of Bayes and minmax rules can be difficult in many applications. Bayesian posterior distributions can be calculated directly when the prior and likelihood have a conjugate form. One way to solve for a minmax rule is to guess the form of a 'least favourable' prior and solve for the associated Bayes rule. If the risk function of the Bayes rule is everywhere less than the Bayes risk, then the rule is minmax. A related method is to construct a least favourable sequence of prior distributions, and calculate the limit of the Bayes risks. If a particular rule has worst-case risk lower than the limit of Bayes risks, then the rule is minmax. Another useful technique for obtaining minmax rules makes use of invariance properties of the decision problem. If the model and loss are invariant with respect to a group of transformations, and that group satisfies a condition called amenability, then the best equivariant procedure is minmax by the Hunt–Stein theorem. These techniques are discussed in Ferguson (1967) and Berger (1985).

If Bayes and minmax rules cannot be obtained analytically, computational methods can sometimes be useful. Recently developed simulation methods such as Markov chain Monte Carlo have greatly expanded the range of settings where Bayes rules can be numerically computed. Chamberlain (2000) develops algorithms for computing minmax rules, and applies them to an estimation problem for a dynamic panel data model.

Asymptotic statistical decision theory

Despite advances in computational methods, many statistical decision problems remain intractable. In such cases, large-sample approximations may be used to show that certain rules are approximately optimal. Le Cam (1972; 1986) proposed to approximate complex statistical decision problems by simpler ones, in which optimal decision rules can be calculated relatively easily. One then finds sequences of rules in the original problem that approach the optimal rule in the limiting version of the problem.

As an example, suppose we observe n i.i.d. draws from a distribution P_θ, where $\theta \in \Theta \subset \mathbb{R}^k$ and the probability measures $\{P_\theta\}$ satisfy conventional regularity conditions with non-singular Fisher information I_θ. We can think of this as defining a sequence of experiments, where the nth experiment consists of observing an n-dimensional random vector distributed according to P_θ^n, the n-fold product of P_θ. Since, in the limit, θ can be determined exactly, we fix a centring value θ_0, and reparametrize the model in terms of local alternatives $\theta_0 + h/\sqrt{n}$, for $h \in \mathbb{R}^k$. This sequence of experiments has as its 'limit experiment' the experiment consisting of observing a single draw $Z \sim N(h, I_{\theta_0}^{-1})$, and we say that the original sequence of experiments satisfies local asymptotic normality (LAN). More precisely, according to an asymptotic representation theorem (see van der Vaart, 1991), for any sequence of procedures δ_n in the original experiments that converge in distribution under every local parameter h, these limit distributions are matched by the distributions associated with some randomized procedure $\delta(Z)$ in the limit experiment. Thus, the limit experiment characterizes the set of attainable limit distributions of procedures in the original sequence of experiments. Solving the decision problem in the limit experiment leads to bounds on the best possible asymptotic behaviour of procedures in the original problem, and often suggests the form of asymptotically optimal procedures.

Le Cam's theory underlies the classic result that in regular parametric models, Bayes and maximum likelihood point estimators of θ are 'asymptotically efficient'. In the LAN limit experiment $Z \sim N(h, I_{\theta_0}^{-1})$, a natural estimator for the parameter h is $\delta(Z)=Z$. This can be shown to be minmax and best equivariant for 'bowl-shaped' loss functions. Both the Bayes and MLE estimators in the original problem are matched asymptotically by this optimal estimator, so they are locally asymptotically minmax and best equivariant. The ideas have been extended to models with an infinite-dimensional parameter space (see Bickel et al. (1993) and van der Vaart, 1991, among others), to obtain semiparametric efficiency bounds for finite-dimensional sub-parameters. More recently, a body of work has developed limit experiment theory for nonparametric problems such as nonparametric regression and nonparametric density estimation (see Brown and Low, 1996, and Nussbaum, 1996, among others).

These results show that nonparametric regression and density estimation are asymptotically equivalent to a white-noise model with drift, for which a number of optimality results are available.

Applications in economics

Portfolio choice

A number of authors have used statistical decision theory to study portfolio allocation when the distribution of returns is uncertain. Some examples include Klein and Bawa (1976), Kandel and Stambaugh (1996), and Barberis (2000), who develop Bayes rules for portfolio choice problems.

Treatment choice

Another econometric application of statistical decision theory is to treatment assignment problems, in which a social planner wishes to assign individuals to different treatments (for example, different job training programmes) to maximize some measure of social welfare. Manski (2004) develops minmax-regret results for the treatment assignment problem, Dehejia (2005) develops Bayesian rules, and Hirano and Porter (2005) obtain asymptotic minmax regret-risk bounds and show that certain simple rules are optimal according to this criterion.

Model uncertainty and macroeconomic policy

Brainard (1967) studied a macroeconomic policy problem, in which a parameter describing the effect of a policy instrument on a macroeconomic outcome is not known with certainty but is given a distribution. The policymaker has a utility function over outcomes and chooses the policy that makes expected utility. More recently, a number of authors have continued this line of work, extending the analysis to more general forms of model uncertainty and developing both Bayesian and minmax solutions. Some examples include Hansen and Sargent (2001), Rudebusch (2001), Onatski and Stock (2002), Giannoni (2002), and Brock, Durlauf and West (2003).

Instrumental variables models

Decision-theoretic ideas underlie recent work on the linear instrumental variables model in econometrics. Chamberlain (2005) develops minmax optimal point estimators in the IV model using invariance arguments. Andrews, Moreira and Stock (2004) have developed tests in the IV model that are optimal under an invariance restriction, and Chioda and Jansson (2004) have developed optimal conditional tests.

Time series models

Asymptotic statistical decision theory has been useful in studying certain time series models which do

not satisfy standard regularity conditions. Jeganathan (1995) shows that a number of models for econometric time series have limit experiments that are not of the standard LAN form, but are locally asymptotically mixed normal (LAMN) or locally asymptotically quadratic (LAQ). Ploberger (2004) obtains a complete class theorem for hypothesis tests in the LAQ case, which nests the LAMN and LAN cases.

Auction and search models

Some parametric auction and search models, in which the support of the data depends on some of the model parameters, do not satisfy the LAN regularity conditions. For such models, Hirano and Porter (2003) showed that the maximum likelihood point estimator is not generally optimal in the local asymptotic minmax sense, but that Bayes estimators are asymptotically efficient.

See Also

- Bayesian econometrics
- Markov chain Monte Carlo methods
- maximum likelihood
- Savage, Leonard J. (Jimmie)
- Wald, Abraham

Bibliography

Andrews, D.W.K., Moreira, M.M. and Stock, J.H. 2004. Optimal invariant similar tests for instrumental variables regression. Discussion Paper No. 1476. Cowles Foundation, Yale University.

Barberis, N.C. 2000. Investing for the long run when returns are predictable. *Journal of Finance* 55, 225–64.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Bickel, P.J., Klaasen, C.A., Ritov, Y. and Wellner, J.A. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.

Brainard, W.C. 1967. Uncertainty and the effectiveness of policy. *American Economic Review* 57, 411–25.

Brock, W.A., Durlauf, S.N. and West, K.D. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 2003(1), 235–322.

Brown, L.D. and Low, M.G. 1996. Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics* 24, 2384–98.

Chamberlain, G. 2000. Econometric applications of maxmin expected utility. *Journal of Applied Econometrics* 15, 625–44.

Chamberlain, G. 2005. Decision theory applied to an instrumental variables model. Working paper, Harvard University.

Chioda, L. and Jansson, M. 2004. Optimal conditional inference for instrumental variables regression. Working paper, UC Berkeley.

Dehejia, R.H. 2005. Program evaluation as a decision problem. *Journal of Econometrics* 125, 141–73.

Ferguson, T.S. 1967. *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Giannoni, M.P. 2002. Does model uncertainty justify caution? Robust optimal monetary policy in a forward-looking model. *Macroeconomic Dynamics* 6(1), 111–44.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Hansen, L.P. and Sargent, T.J. 2001. Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics* 4, 519–35.

Hirano, K. and Porter, J. 2003. Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* 71, 1307–38.

Hirano, K. and Porter, J. 2005. Asymptotics for statistical treatment rules. Working paper, University of Arizona.

Jeganathan, P. 1995. Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11, 818–87.

Kandel, S. and Stambaugh, R.F. 1996. On the predictability of stock returns: an asset-allocation perspective. *Journal of Finance* 51, 385–424.

Klein, R.W. and Bawa, V.S. 1976. The effect of estimation risk on optimal portfolio choice. *Journal of Financial Economics* 3, 215–31.

Le Cam, L. 1972. Limits of experiments. *Proceedings of the Sixth Berkeley Symposium of Mathematical Statistics* 1, 245–61.

Le Cam, L. 1986. *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Manski, C.F. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72, 1221–46.

Nussbaum, M. 1996. Asymptotic equivalence of density estimation and Gaussian white noise. *Annals of Statistics* 24, 2399–430.

Onatski, A. and Stock, J.H. 2002. Robust monetary policy under model uncertainty in a small model of the U.S. economy. *Macroeconomic Dynamics* 6(1), 85–110.

Ploberger, W. 2004. A complete class of tests when the likelihood is locally asymptotically quadratic. *Journal of Econometrics* 118, 67–94.

Rudebusch, G.D. 2001. Is the fed too timid? Monetary policy in an uncertain world. *Review of Economics and Statistics* 83, 203–17.

Savage, L.J. 1951. The theory of statistical decision. *Journal of the American Statistical Association* 46, 55–67.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

van der Vaart, A.W. 1991. An asymptotic representation theorem. *International Statistical Review* 59, 97–121.

Wald, A. 1939. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics* 10, 299–326.

Wald, A. 1950. *Statistical Decision Functions*. New York: Wiley.

How to cite this article

Hirano, Keisuke. "decision theory in econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000244> doi:10.1057/9780230226203.0365

The New Palgrave Dictionary of Economics Online

default and enforcement constraints

Fabrizio Perri

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article illustrates when limited enforcement of contracts induces enforcement constraints (limits to intertemporal exchange) or default (the breaking of intertemporal promises with the associated punishment), and sheds light on how enforcement policies should be related to the observed frequency of default. When limited enforcement is the only friction equilibrium default is never observed, yet tightening enforcement of contracts is socially beneficial. When limited enforcement coexists with other frictions, default occurs in equilibrium but tightening enforcement might be socially undesirable. The reason is that equilibrium default, although detrimental to intertemporal exchange, might lead to improved allocation of resources across states.

Keywords

Arrow–Debreu promises; default; enforcement constraints; intertemporal exchange; Lagrange multipliers; limited enforcement of contracts with default; limited enforcement of contracts without default; risk sharing

Article

Intertemporal exchange, that is the exchange of resources today for a promise of resources at a later date in a given state, is key for promoting economic efficiency. For example, to finance an investment, a government borrows capital abroad in exchange for a promise of repayment once the investment has paid off. Or, to finance consumption, an individual who loses her job borrows resources in exchange for the promise of repayment once she gets a new job. If the enforcement of promises is limited, the extent of intertemporal exchange can be reduced by so-called *enforcement constraints* and, under some conditions, *default*, that is, the breaking of promises, can arise. This article presents a simple general equilibrium set-up to analyse these issues and provide some direction for the design of enforcement policies. Key references for the theory of limited enforcement without default are Kehoe and Levine

(1993), Kocherlakota (1996) and Alvarez and Jermann (2000), while for limited enforcement with default see Zame (1993) and Dubey, Geanakoplos and Shubik (2005).

The set-up

The goal of this set-up is to capture the need for intertemporal exchange, as described in the examples above. There are two agents which live for two periods and consume a single good. Agent 1, the borrower, owns a technology such that, if k units of the good are invested in period 1, AK^{α}, $0 < \alpha < 1$, units are produced in period 2, where A is a random variable realized in period 2, with positive support and distribution $F(A)$ known to both agents. Agent 2, the lender, is endowed with e units of the consumption good in period 1. Consumption allocations of agent i are consumption at date 1, c_{i1} and the function $c_{i2}(A)$ which assigns period 2 consumption for each possible realization of A. Borrower's utility is given by $u(c_{11}) + \int u(c_{12}(A))dF(A)$ where u is a concave utility function satisfying Inada conditions. The lender has linear utility given by $c_{21} + \int c_{22}(A)dF(A)$. Linear utility implies that lender's equilibrium utility is constant across different market structures so that borrower's utility is the only statistic needed to Pareto-rank equilibria. In all the economies described below the following resource constraints hold

$$c_{11} + c_{21} + k = e c_{12}(A) + c_{22}(A) = Ak^{\alpha} \text{for every } A$$

A frictionless benchmark

Assume agents can trade a complete set of Arrow–Debreu promises which are fully and costlessly enforceable. The budget constraints of the borrower are

$$c_{11} + k = \int p(A)dF(A) \tag{1}$$

$$c_{12}(A) = Ak^{\alpha} - p(A) \text{for every } A \tag{2}$$

where $p(A)$ denotes the amount that the borrower promises to repay in state A. Equilibrium allocations display complete risk sharing, that is, the ratio of marginal value of consumption of the two agents is constant across dates and states of the world. We denote with c^{AD} the constant, across dates and states, level of consumption of the borrower in this economy.

Limited enforcement

This section describes an economy denoted as ADLE (Arrow–Debreu Limited Enforcement) and shows that limited enforcement prevents full risk sharing, reduces investment and welfare. Assume that in period 2 the borrower can walk away from any promise made to the lender by suffering a default deadweight cost proportional to its output and equal to δAk^{α} where $\delta > 0$ is a parameter that measures the strength of enforcement. This implies that any Arrow–Debreu promise $p(A) > \delta Ak^{\alpha}$ will not be honoured by the borrower and thus will not be purchased by the lender. Also, promises satisfying $p(A) \leq \delta Ak^{\alpha}$ will be fully honoured and priced as in the frictionless economy. So limited enforcement limits the use of state-contingent promises but does not induce default. A convenient way of capturing this, following Alvarez and Jermann (2000), is to assume that the borrower faces constraints on the sales of each promise so as to guarantee no default. These *enforcement* constraints have the form

$$p(A) \leq \delta Ak^{\alpha} \text{for every } A \tag{3}$$

as the borrower can sell each promise only up to the point where the cost of keeping it is equal to the cost of defaulting on it. Equilibrium allocations can be characterized by substituting budget constraints (1) and (2) into the borrower's utility and taking first-order conditions with respect to k and $p(A)$ subject to constraints (3). This yields

$$u'(c_{11}) = \int [A\alpha k^{\alpha-1} u'(c_{12}(A)) + A\alpha k^{\alpha-1} \delta \mu(A)]\, dF(A) \tag{4}$$

where

$$\mu(A) = u'(c_{11}) - u'(c_{12}(A))$$

are the Lagrange multipliers on the enforcement constraints. If the cost of default δ is sufficiently small and the distribution of *A* is sufficiently spread out, $c_{11} = c(A) = c^{AD}$ is not a solution of (4) as enforcement constraints on the high *A* promises would be violated. The solution is then characterized by a level of productivity A^* such that for all $A > A^*$ enforcement constraints are binding and $c(A) = (1-\delta)Ak^{\alpha} > c_{11}$. For $A \leq A^*$ enforcement constraints are not binding and $c(A) = c_{11} < c^{AD}$. Complete risk sharing involves the borrower selling promises to repay in states with high *A*, in order to finance consumption today (when she has no output) and consumption tomorrow in states with low *A*. But if the distribution of *A* is spread out, complete risk sharing calls for promises of a large transfer of resources from the borrower to the lender in the states with high *A*. When enforcement is limited (δ is low) the lender, in period 1, correctly anticipates that these transfers will not be made and buys a smaller amount of the promises. So, relative to complete risk sharing, the borrower has fewer resources in period 1 and in the period 2 states with low *A*, but consumes more in period 2 states high *A*. This allocation of consumption increases the marginal value of resources in period 1 relative to the expected marginal value of resources in period 2 and thus reduces *k* relative to the full enforcement case. Finally, equilibria in economies with strong enforcement (high δ) Pareto-dominate equilibria with weak enforcement (low δ). To see this, note that, for the borrower, the equilibrium allocation in the weak enforcement economy is budget-feasible in the strong enforcement economy, so, if it is not chosen, it must yield her lower utility.

ADLE economies have been used extensively in a variety of applications such as asset pricing (Alvarez and Jermann, 2000), international business cycles (Kehoe and Perri, 2002) and consumption inequality (Krueger and Perri, 2006). All these studies show that limited enforcement prevents complete risk sharing, and for this reason it provides a much better fit with the data than standard Arrow–Debreu economies. This environment, though, cannot be used to understand equilibrium default (that is, the actual break of a promise and the suffering of the associated cost) as the trade in contingent promises makes incurring the default cost unnecessary. In order to understand when default arises and what its consequences are, the next section considers an economy in which contingent promises cannot be traded, either because markets are exogenously missing or because the borrower has private information about realizations of *A*.

Limited enforcement and non-contingent promises

The borrower finances consumption and investment only by selling a non-contingent promise *p* which can be defaulted on in state *A* by suffering the default cost δAk^{α}. Since the cost of repaying the promise does not vary with the state while the default cost is increasing with *A*, if there is equilibrium default it will happen in the low *A* states. In particular, if the borrower invests *k* and sells a promise *p*, she will default in all the states such that $A \leq \frac{p}{\delta k^{\alpha}}$.

As a consequence, the equilibrium price of the promise is given by

$$q(p, k) = 1 - F\left(\frac{p}{\delta k^{\alpha}}\right).$$
(5)

The problem of the borrower is then

$$\max_{p,k} u(q(p, k)\, p - k) + \int_0^{\frac{p}{\delta k^{\alpha}}} u((1-\delta)Ak^{\alpha})\, dF(A) + \int_{\frac{p}{\delta k^{\alpha}}}^{\infty} u(Ak^{\alpha} - p)\, dF(A).$$
(6)

The equilibrium is characterized by a couple *p*, *k* which solve (5) and (6). It can be immediately shown that equilibria in this economy are, generically, Pareto-inferior to equilibria in the corresponding ADLE economy. Also, for many parameter values equilibria in this set-up differ from those in the ADLE economy along two important dimensions: (*a*) there is a positive measure of states for which default occurs and (*b*) there is a positive measure of values for δ for which welfare is *decreasing* in the strength of enforcement. As a simple example, consider the case in which *A* can take only two values: a high value A_h with probability π and a low value A_l with probability $1 - \pi$, with $\pi > A_l / A_h$. In this case there is a range of values for δ for which the equilibrium promise and capital satisfy

$$\delta A_l k^{\alpha} < p < \delta A_h k^{\alpha},$$
(7)

so that default happens only when state A_l is realized and consequently $q(p, k) = \pi$. Now consider the effect of a marginal reduction in δ. Equation (7) shows that, if the borrower kept *k* and *p* unchanged in response to the change in δ, default patterns, and hence *q*(*p*, *k*), would not change; however reducing δ increases the returns of borrower in the default state so its utility would increase relative to the initial equilibrium. Here weakening enforcement allows the borrower to implicitly transfer, through default, more resources to the low *A* state and thus to achieve a better allocation of risk across states. In the ADLE economy this transfer was achieved through the Arrow–Debreu promises so default was not necessary. When promises cannot be made state-contingent, increasing payoffs in the default states is the only way of obtaining this transfer.

In this simple example weakening enforcement does not affect default frequency, but in more general set-ups it does and as a consequence increases equilibrium interest rates and hampers intertemporal

exchange. This effect is detrimental for welfare. But the example above suggests that the detrimental effect can be offset by the positive effect of the better risk allocation across states. Note that this result does not rely on the two-state assumption, and it can be shown to hold, for example, also when *A* is log-normally distributed.

Summary

Limiting contract enforcement in otherwise frictionless environments constrains intertemporal exchange and hampers risk sharing, investment and welfare, but does not induce default. When additional frictions, such as incomplete markets or private information, limit the span of tradable promises, then limited enforcement can play a positive role by inducing equilibrium default, which can be used as a (costly) way of providing better allocation of risk across states. The analysis sheds light on how enforcement policies should be related to the observed frequency of default.
When limited enforcement is the only friction, default is never observed, yet tightening enforcement is socially beneficial. When limited enforcement coexists with other frictions, default happens in equilibrium but this does not necessarily mean that enforcement should be tightened. Indeed, tightening enforcement without ameliorating the additional friction might reduce default but also risk sharing and welfare.

See Also

- risk sharing
- sovereign debt

Bibliography

Alvarez, F. and Jermann, U. 2000. Efficiency, equilibrium, and asset pricing with risk of default. *Econometrica* 68, 775–97.

Dubey, P., Geanakoplos, J. and Shubik, M. 2005. Default and punishment in general equilibrium. *Econometrica* 73, 1–37.

Kehoe, P. and Perri, F. 2002. International business cycles with endogenous incomplete markets. *Econometrica* 70, 907–28.

Kehoe, T. and Levine, D. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60, 865–88.

Kocherlakota, N. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63, 595–609.

Krueger, D. and Perri, F. 2006. Does income inequality lead to consumption inequality? Evidence and

theory. *Review of Economic Studies* 73, 163–93.

Zame, W. 1993. Efficiency and the role of default when securities markets are incomplete. *American Economic Review* 83, 1142–64.

How to cite this article

Perri, Fabrizio. "default and enforcement constraints." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000252> doi:10.1057/9780230226203.0366

The New Palgrave Dictionary of Economics Online

defence economics

Keith Hartley

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Defence economics is a new field of economics. Its development and research agenda have reflected current events. Examples include the superpower arms race of the cold war, disarmament following the end of the cold war, international terrorism, peacekeeping and conflict. A brief history is presented; the field is defined and the facts of world military spending are outlined; the defence economics problem, namely, the need for difficult choices, is considered; and conflict and terrorism are used to illustrate some of the new developments in the field.

Keywords

arms races; arms trade; cost-plus contracts; crowding out; defence economics; disarmament costs; economic theories of military alliances; ethnicity; European Security and Defence Policy (EU); fixed-price contracts; free rider problem; game theory; military employment contract; military outsourcing; military wage differential; military–industrial complex; nationalism; one-shot games; principal and agent; private finance initiatives; procurement; public–private partnerships; purchasing power parity; religion; repeated games; research and development; strategic behaviour; substitution effect; substitution principle; technology; terrorism, economics of; tit for tat; two world wars, economics of the; voting paradoxes; war and economics

Article

Defence economics is a relatively new part of the discipline of economics. One of the first specialist contributions in the field was by C. Hitch and R. McKean, *The Economics of Defense in the Nuclear Age* (1960). This book applied basic economic principles of scarcity and choice to national security. It focused on the quantity of resources available for defence and the efficiency with which such resources were used by the military. For example, defence consumes scarce resources that are therefore not available for social welfare spending (for example, missiles versus education and health trade-offs).

Once resources are allocated to defence, military commanders have to use them efficiently, combining their limited quantities of arms, personnel and bases to ‘produce’ security and protection. Within such a military production function, there are opportunities for substitution. For example, capital (weapons) can replace (and have replaced) military personnel; imported arms can replace nationally produced weapons; and nuclear forces have replaced large standing armies. Defence economics is about the application of economic theory to defence-related issues.

The development of defence economics and its research agenda reflected current events. For example, during the cold war there was a focus on the superpower arms races, alliances (NATO and the Warsaw Pact), nuclear weapons and ‘mutually assured destruction’. The end of the cold war resulted in research into disarmament, the challenges of conversion and the availability of a peace dividend. Since the end of the cold war, the world remains a dangerous place with regional and ethnic conflicts (for example, Bosnia, Kosovo, Iraq), threats from international terrorism (for example, terrorist attacks on USA on 11 September 2001), rogue states and weapons of mass destruction (that is, biological, chemical and nuclear weapons). NATO has accepted new members (for example, former Warsaw Pact states) and has developed new missions, and the European Union has developed a European Security and Defence Policy. Changing threats and new technology require the armed forces and defence industries to adjust to change and new challenges. Peacekeeping has become a major mission for armed forces and is an example of the trend towards globalization.

The modern era of globalization involves more international transactions in goods, services, technology and factors of production, which brings new security challenges for both nation states and the international community. Defence firms have become international companies with international supply networks. Globalization also highlights the importance of international collective action to respond to new threats such as international terrorism and to maintain world peace (for example, through international peacekeeping missions under UN, NATO or EU control). But international collective action experiences the standard problems of burden-sharing and free riding.

This article outlines the development of defence economics; it defines the field and describes the ‘stylized facts’ of world military expenditure; the defence economics problem is considered; and a case study of conflict and terrorism illustrates some of the new developments in the field.

A brief history

Defence issues have existed throughout history as nations have been involved in armed conflict of various forms and durations (for example, the Hundred Years War). Great powers have used military force to dominate regions and parts of the world (for example, Alexander the Great; Roman legions; Genghis Khan; Ottoman Turks; Nazi Germany), with such powers rising and falling (Kennedy, 1988). Conflict has also been characterized by major technical changes ranging from bows and arrows to cannons and machine guns, from sailing ships to iron and steel warships and nuclear-powered vessels, from horse cavalry to tanks, from flag communications to radios and satellite communications and from balloons to aircraft, missiles, nuclear weapons and space systems. Historically, the economic base for conflict was first an agricultural society, then an industrial society followed by a knowledge economy. Some of the classical economists studied war and conflict (for example, Smith, Ricardo, Malthus, J. S. Mill: see Goodwin, 1991, ch. 2). For these economists, war departed from much of their conventional thinking: it involved chaos and disorder rather than market equilibrium, and it required government

action rather than private market behaviour. Yet it remains surprising that, with a long history of wars, including two world wars and the superpower arms race of the Cold War, relatively few economists have been attracted to the field. A review of the economics literature on conflict concludes that ‘We were surprised at the relative absence of applied economics studies of actual conflicts’ (Sandler and Hartley, 2003, p. xl). There are various possible explanations for the relative absence of economists studying war and conflict. These include data and security problems, the difficulty of applying conventional market analysis to the chaos and disequilibrium of conflict, a traditional reluctance to analyse the public sector (with defence assumed to be exogenous), and the feeling that defence and security issues are not as important as other social welfare issues, with war viewed as an immoral and unethical subject. Furthermore, security issues have not been as an attractive career path for economists (compared with issues such as inflation, unemployment, growth and developing countries), and conflicts are usually of short duration so that they offer only limited research prospects before peace returns to remove war-related problems (Goodwin, 1991, pp. 1–2).

Definitions

Defence economics studies all aspects of war and peace and embraces defence, disarmament and conversion. This definition includes studies of both conventional and non-conventional conflict such as civil wars, revolutions and terrorism. It involves studies of the armed forces and defence industries and the efficiency with which these sectors use scarce resources in providing defence output in the form of peace, protection and security. Reductions in defence spending (such as those following the end of the cold war) result in disarmament, which involves reallocating resources from the defence to the civilian sector. This raises questions about the impact of disarmament on the employment and unemployment of both military personnel and defence industry workers; the possibilities for converting military bases and arms industries to civil uses (the Biblical swords to ploughshares); and the role of public policies in assisting the transition and reallocation of resources.

The coverage of the subject is extensive and involves economic theory, empirical testing and policy-related issues, including applications of public choice analysis. Both defence and peace have distinctive economic characteristics in that they are public goods which are non-rival and non-excludable. There are large literatures dealing with the determinants of military expenditure, including economic theories of military alliances and arms races (that is, threats) and the impact of defence spending on economic growth and development. Armed forces are major buyers of both equipment (arms/weapons) and military personnel, and such procurement choices affect defence industries and both local and national labour markets. For example, government procurement of weapons involves choices between competition and preferential purchasing and between various types of contracts (for example, fixed-price, cost-plus), each with different implications for contractor efficiency and profitability. There is a related literature on industrial and alliance policies comparing the economics of supporting a national defence industrial base with alternative industrial policies such as international collaboration, licensed production or importing foreign equipment. Imports also involve the international arms trade, its economic impacts on both buyers and suppliers, and policy initiatives to regulate such trade. More generally, there is an extensive literature on arms control and disarmament, the adjustment costs of disarmament, the economics of conversion and the contribution of public policy to minimizing such

adjustment costs. Finally, there have been some new developments involving the application of economics to the study of conflict, civil wars, revolutions and terrorism (Brauer, 2003; Hegre and Sandler, 2002; Sandler and Hartley, 1995; 2007).

Defence economics became established in the 1960s with the publication of a number of pioneering contributions, mostly by US economists. These contributions applied economics to some novel areas and included economic models of alliances (Olson and Zeckhauser, 1966), the economics of arms races (Richardson, 1960; Schelling, 1966), the procurement of weapons and military personnel (Peck and Scherer, 1962; Oi, 1967), and the impact of military spending on economic development (Benoit, 1973).

A further development confirming the emergence of defence economics as an accepted part of the discipline of economics was the launch in 1990 of a field journal, *Defence Economics*, later renamed *Defence and Peace Economics* (initially it was published four times per year, but in 2000 it was expanded to six issues per year).

Inevitably, defence economics generates controversy reflected in myths and emotion. Critics point to the ‘wastes’ of defence spending and its ‘crowding-out’ of ‘valuable’ civil expenditure. Classic examples include the sacrifice of schools and hospitals associated with major weapons projects such as modern combat aircraft and aircraft carriers (for example, the US F-22 aircraft and the European Typhoon). Peace economists are similarly critical of defence economics and military spending: they focus on peace topics such as disarmament and the maintenance of peace, arms control and international security, conflict analysis and management, and crises and war studies. Defence economists are not, however, ‘warmongers’: they are instead interested in understanding the economics of the military–industrial–political complex and all aspects of defence whereby a proper understanding of these issues will contribute to a more peaceful world. A starting point in showing how economists analyse defence is to review the ‘stylized facts’ of world military spending.

The stylized facts of world military spending

What is known about military spending, and where are the gaps in the data? Good quality data exist on world military spending, the world's armed forces and the arms trade. Cross-section and time-series data are available at the country level; some examples are shown in Table 1. The data on world military expenditure show aggregate spending by the USA accounting for 45 per cent of total world military spending and NATO accounting for some 70 per cent. Similarly, in 2004 the USA dominated defence R&D spending, accounting for some 75 per cent of the world total and 31 per cent of world arms exports.

World military spending and armed forces, various years

World military expenditure	*US$ billion, 2004*
NATO	722
USA	467
France	52
Germany	38
UK	54

China	37
Russia	23
World total	**1,035**
Defence share of GDP	%[a]
USA	3.9
France	2.6
Germany	1.4
UK	2.3
Eritrea	19.4
Burundi	5.9
Sudan	2.4
India	2.1
Pakistan	4.4
Israel	9.1
Jordan	8.9
Oman	12.2
Defence research and development[b]	*US$ billion, 2004* (2001 prices and PPP rates)
USA	67.5
Russia	6.1
UK	4.7
USA and EU total	80.9
Estimated world total of defence R&D	**90.0** +
World armed forces	*Number of military personnel, 1999 (’000s)*
Developed nations	6,550
Developing nations	14,700
NATO	4,580
USA	1,490
UK	218
Eritrea	215
China	2,400
World	**21,300**
World arms trade	*US$ million, 2000–04 (1990 prices)*
Major importers	
China	11,677

India	8,526
Greece	5,263
UK	3,395
Turkey	3,298
World total	**84,490**
Major exporters	
Russia	26,925
USA	25,930
France	6,358
Germany	4,878
UK	4,450
World total	**84,490**

Notes: [a]Defence share data for USA, France, Germany and UK are for 2004; all other data are for 2003. [b]Defence R&D data are for government-funded defence R&D. PPP=purchasing power parity.

Sources: US DoS (2002); NATO (2005); OECD (2004); SIPRI (2005).

Table 1 shows examples of defence shares of GDP to illustrate the burdens of defence spending, especially for developing nations such as Eritrea, India and Pakistan (an arms race situation) and for the Middle East (a conflict region). Burundi and Sudan have defence burdens similar to or greater than those of the UK and Germany. Table 1 also shows other measures of the economic burdens of defence for the world's poorer nations (that is, nations which cannot feed, house or educate their populations and which have poor health records). Developing nations accounted for 70 per cent of the world total of 21.3 million military personnel, and such totals further show the importance of military manpower economics. Similarly, developing nations are major importers of arms, while the developed nations are the major arms exporters. Such data provide an introduction to some of the major themes of defence economics, namely, the determinants of military expenditure, arms races, alliances, the relationship between defence spending and economic development, the arms trade and the economics of military personnel.

Micro-level data are more limited but there are some useful sources especially on defence contractors and defence industries. Table 2 provides examples of such micro-level data based on the 100 largest arms-producing companies (SIPRI, 2005) and employment in national defence industries (BICC, 2005). Again, these data are available on a cross-section and time-series basis, and the company data include total sales, total profits and aggregate employment. From Table 2 it can be seen that the USA has six of the world's top ten arms companies and that the American firms have a substantial scale advantage over their European rivals: the average size of a US firm from the top ten is almost twice the corresponding average of the European companies. These data are the basis for research questions about the determinants of firm size, the impact of economies of scale, scope and learning, and the determinants of performance in terms of labour productivity and profitability.

Defence companies and industries

Major defence companies	Arms sales, 2003 (US$ million)
Lockheed Martin (USA)	24,910
Boeing (USA)	24,370
Northrop Grumman (USA)	22,720
BAE Systems (UK)	15,760
Raytheon (USA)	15,450
General Dynamics (USA)	13,100
Thales (France)	8,350
EADS (Europe)	8,010
United Technologies (USA)	6,210
Finmeccanica (Italy)	5,290
Major defence industries	Employment numbers, 2003 (‘000s)
Industrialized countries	4,710
Developing countries	2,769
NATO	3,452
EU	645
USA	2,700
China	2,100
Russia	780
France	240
UK	200
World total	**7,479**

Sources: BICC (2005); SIPRI (2005).

Table 2 also shows data on defence industry employment. The industrialized nations accounted for 63 per cent of total employment in the world's defence industries, with the developing countries accounting for the remaining 37 per cent. The USA, China and Russia have the largest defence industries by employment, accounting for 75 per cent of the world total. Overall, the world military–industrial complex employed almost 29 million personnel in the armed forces and defence industries, reinforcing its role as a major employer of labour, including some highly qualified R&D staff and other highly skilled workers. Such scarce labour has alternative uses in the civilian sector, raising questions as to whether defence spending ‘crowds out’ valuable civil investment and diverts scientific manpower from civil research projects.

Despite the available data, there remain significant gaps in our knowledge of the world's military sector. Typically, new defence projects are surrounded by secrecy; there are problems in identifying some defence goods (for example, dual use goods, such as civil airliners which can be used as military transport aircraft); there is a lack of good-quality data on defence R&D, including employment in defence R&D; and little is known about China, especially its defence R&D programmes (Hartley,

2006a). International comparisons of military expenditure data are also sensitive to the choice of exchange rate adjustments, with country rankings sensitive to the use of market exchange rates or purchasing power parity rates (SIPRI, 2005). At the firm and industry levels, analysis of the military business in terms of defence output, employment and profitability is complicated because the typical output comprises a mix of military and civil components, making it difficult to compare the performance of defence contractors and civil firms. Further gaps exist in our knowledge of the world regional distribution of military bases and defence plants, so that it is difficult to assess the economic dependence of various regions on defence spending. Little is known about defence industry supply chains both within countries and within the global economy. Finally, there is a need for more reliable data on the international trade (including illegal transactions) in small arms (these are often the main weapons used in many regional conflicts, such as in Bosnia).

The defence economics problem

This is the standard choice problem of economics, but applied to defence. Typically, following the end of the cold war defence budgets have been either constant or falling in real terms; and these limited budgets are faced with rising input costs of both capital and labour. Equipment costs have been rising by some ten per cent per annum in real terms, which means a long-run reduction in the numbers of weapons acquired for the armed forces (for example, the US Air Force's original requirement for F-22 combat aircraft for 750 units was later reduced to some 180 aircraft). Similarly, with an all-volunteer force, the costs of military personnel have to rise faster than wage increases in the civil sector. This wage differential is required to attract and retain military personnel by compensating them for the net disadvantages of military life. Here, the military employment contract is unique in that armed forces personnel are subject to military discipline; they are required to deploy to any part of the world at short notice; they could remain overseas indefinitely; and some might never return (that is, death and injury are a feature of this contract). This combination of constant or falling defence budgets and rising input costs means that governments and defence policymakers cannot avoid the need for difficult choices in a world of uncertainty (that is, where the future is unknown and unknowable, and no one can accurately predict the future).

Faced with this defence choice problem, governments have adopted various solutions. They can adopt a policy of ‘equal misery’ whereby each of the services is subject to budget cuts (for example, reduced training, cancelling some new equipment projects and delaying others); or they can undertake a major revision of a nation's defence commitments (for example, a defence review such as the UK's 1998 Strategic Defence Review); or they can seek to improve efficiency in the armed forces and defence industries (for example, via a competitive equipment procurement policy and military outsourcing). Other policy options include joining a military alliance (such as NATO; EU) or avoiding the defence choice problem by increasing the defence budget (as in the USA since 11 September 2001); but then choices are needed between defence and social welfare spending.

Economics offers three broad policy principles for formulating an efficient defence policy, namely, final outputs, substitution and competition. Take first the principle of *final outputs*. Measuring defence output is notoriously difficult, but it can be expressed in such general terms as peace, security and threat reduction. The UK has solved the problem by committing (and funding) its armed forces to having the capacity to fight simultaneously three small to medium conflicts (for example, Bosnia, Kosovo, Sierra

Leone) or one large-scale conflict as part of an international coalition (for example, the Gulf War, Iraq). This approach is a departure from the traditional focus on measuring inputs in terms of the numbers of infantry regiments, warships, tanks and combat aircraft. Such a focus fails to address the key issue of the contribution of these inputs to final defence output in the form of peace and protection. A focus on inputs also fails to address the marginal contribution of each of the armed forces: what would be the implications for defence output if, say, the air force were expanded by five to ten per cent, or the navy was reduced by five to ten per cent?

The second economic principle is that of *substitution*. There are alternative methods of achieving protection, each with different cost implications. Possible examples of partial substitutes include reserves replacing regular personnel, civilians replacing regulars (for example, police in Northern Ireland replacing army personnel), attack helicopters replacing tanks, ballistic and cruise missiles and unmanned combat air vehicles replacing manned strike and bomber aircraft, air power replacing land forces, and imported equipment replacing nationally produced equipment. Some of these substitutions might alter the traditional monopoly property rights of each of the armed forces. For example, surface-to-air missiles operated by the army might replace manned fighter aircraft operated by the air force, and maritime anti-submarine aircraft operated by the air force might replace frigates supplied by the navy.

The third economic principle is that of *competition* as a means of achieving efficiency. Standard economic theory predicts that, compared with monopoly, competition results in lower prices, higher efficiency, and competitively determined profits and innovation in both products and industrial structure. For equipment procurement, competition means allowing foreign firms to bid for national defence contracts and awarding fixed-price contracts rather than cost-plus contracts; it also means ending any ‘cosy’ relationship between the defence ministry and its national champions and any preferential purchasing and guaranteed home markets.

Competition can be extended to activities undertaken by the armed forces. Here, there is a public sector monopoly problem whereby the armed forces have traditionally undertaken a range of activities ‘in house’ without being subject to any rivalry. Military outsourcing allows private contractors to bid for and undertake such activities. Examples include accommodation, catering, maintenance, repair, training, transport and management tasks (for example, managing stores or depots and firing ranges). In some cases, outsourcing involves private finance initiatives whereby the private sector finances the activity (for example, new buildings or an aircrew simulator training facility) and then enters into a long-term contract with the defence ministry to provide services to the armed forces in return for rental payments. Another variant is a public–private partnership whereby the private sector finances an activity or asset in return for rental payments from the defence ministry, but the contractor is allowed to sell any peacetime spare capacity to other users (for example, tanker aircraft capacity which when not needed in peacetime can be rented to other users).

Application of the policy guidelines to an efficient defence policy requires that individuals and groups in the military–industrial–political complex are provided with sufficient incentives to behave efficiently. There are the inevitable principal–agent problems where agents have considerable opportunities to pursue their own interests which may conflict with those of their principals (for example, leading a quiet life rather than bearing the costs of change). Individuals and groups in the armed forces and defence ministries will be reluctant to apply the substitution principle if there are no personal or group incentives and rewards for achieving efficient substitution (that is, interest groups can be barriers to change). Compare the private sector, where there are market and institutional arrangements promoting efficiency

in the form of rivalry between suppliers, the profit motive and the capital market as a ‘policing and monitoring’ mechanism through the threats of takeover and bankruptcy. Such market arrangements are absent in the armed forces (and elsewhere in the public sector).

There is also the challenge of achieving ‘top level’ efficiency in defence provision. Economic theory solves this challenge as a standard optimization problem involving the maximization of a social welfare function subject to resource or budget constraints (where welfare is dependent on civil goods and security, with security provided by defence). Operationalizing this apparently simple optimization rule is much more difficult. Individual preferences for defence are subject to its public good characteristics and free riding problems and the continued difficulty of defining defence output. In democracies, society's preferences are usually expressed through voting at elections. However, elections are limited as a means of obtaining an accurate indication of society's preferences for defence and its willingness to pay. Elections occur infrequently; they are usually for a range of policies of which defence is only one element in the package (which includes policies on, for example, education, health, transport, the environment, foreign policy and taxation); and the ‘voting paradox’ shows the difficulty of deriving a society's preferences using the voting system. Nor do voters have reliable information on the output of defence spending.

Defence economics explains military spending using a demand model of the form:

$$ME = M(P, Y, T, A, Pol, S, Z)$$

where *ME*=real military spending; *P*=relative prices of military and civil goods and services; *Y*=real national income; *T*=threats in the form of the military expenditure of a rival nation (arms race models); *A*=membership of a military alliance and the real military expenditure of the allies (such as NATO); *Pol*=variable for the political composition of the government (for example, left- or right-wing, with the latter favouring ‘strong defences’); *S*=a variable representing the security and strategic environment (such as the end of the cold war; conflicts such as Korea, Vietnam, the Gulf War and Iraq); and *Z*=other relevant influences (for example, land mass to be protected). Estimation of the demand model usually proceeds without a price variable, mainly because most nations do not provide relative price data. This omission can be justified if the price of military goods and services has inflated at the same rate as civil goods and services; but such an assumption is not always realistic. A survey of empirical results is presented in Sandler and Hartley (1995; 2007).

Conflict and terrorism

The demand model for military expenditure recognized the relevance of threats such as terrorism and conflict as determinants of defence spending. Traditionally, conflict and terrorism have been the preserve of disciplines other than economics. For example, debates and decisions about war involve political, military, moral and legal judgements. But conflict has an economic dimension, namely, its costs. Wars are not costless: they can involve massive costs (for example, the Second World War). Economics has also made further contributions in analysing the causes of conflict and in identifying

potential targets during conflict (for example, the Second World War selection of aircraft factories, dams, submarine yards and oil fields as targets for Allied bombing raids on Germany).

Economic models start by analysing conflict as the use of military force to achieve a reallocation of resources within and between nations (that is, civil wars and international conflict). Nations invade to capture or steal another nation's property rights over its resources (such as land, minerals, oil, population, water). Conflict has a distinctive feature: it destroys goods and factors of production, and it is easier to destroy than to create. In peacetime, civilian economies aim to create more goods and services through growth and expanding a nation's production possibility frontier. Conflict uses military force and destructive power to enable a nation to acquire resources from another state, so expanding its production boundary through military force (Vahabi, 2004).

Conflict and terrorism provide opportunities for applying game theory. They involve strategic behaviour, interactions and interdependence between adversaries ranging from small groups of terrorists, rebels and guerrillas to nation states. Strategic interaction means that conflict can be analysed as games of bluff, chicken and ‘tit-for-tat’ with first-mover advantage and possibilities of one-shot or repeated games. For example, first-mover advantage might indicate a pre-emptive strike (for example, Pearl Harbour in 1941; Kuwait in 1990). However, there are other, non-economic explanations of conflict. These include religion, ethnicity and grievance (for example, Germany after the First World War); the desire for a nation state (such as Palestine); the absence of democracy; and mistakes and misjudgement.

The costs of war are a relatively neglected dimension of conflict. War involves both one-off and continuing costs. One-off costs are those of the actual conflict, while continuing costs are any post-conflict costs including those of occupation and peacekeeping. A further distinction is necessary between military and civilian costs. In principle, the military costs of conflict are the marginal resource costs arising from the conflict (that is, those costs which would not otherwise have been incurred). Examples include the costs of preparation and deployment prior to a conflict; the costs of the conflict, including the costs of basing forces overseas and the use of ammunition, missiles and equipment, including human capital and equipment losses in combat; the post-conflict occupation and peacekeeping missions and the costs of returning armed forces to their home nation.

There are further costs of conflict in the form of impacts on the civilian economies of the nations involved in the war. For example, the US and UK involvement in the Iraq war that began in 2003 had possible short- and long-term impacts for both economies. There were possible impacts on oil prices, share prices, the airline business, tourism, defence industries, private contractors, aggregate demand and future public spending plans. Further substantial costs were imposed on the Iraq economy in the form of deaths and injuries of military and civilian personnel, together with the damage and destruction of physical assets. Table 3 shows some examples of the costs of various conflicts for the UK and USA. The general point remains that wars are costly and require scarce resources which have alternative uses (that is, wars involve the sacrifice of hospitals, schools and social welfare programmes). Questions also arise as to whether the benefits of conflict exceed its costs.

Costs of conflict

UK: Conflict	Military costs to UK (US$ billion, 2005 prices)
World War I	357
World War II	1,175

Gulf War	6.0
Bosnia	0.7
Kosovo	1.7
Iraq	6.0 +
USA: Conflict	Military costs to USA (US$ billions, 2005 prices)
World War I	208
World War II	3,148
Korea	365
Vietnam	537
Gulf War	83
Iraq	440
Estimated civilian costs:	Civilian costs (US$ billion, 2005 prices)
Iraq war	
Costs to US economy[a] *from Iraq war*	557
Costs to world economy[b] *from Iraq war*	1,183
Iraq war: costs to Iraq	US$ billion (2005 prices)
Reconstruction costs	20–60

Notes: [a]US civilian costs are of lost GDP for the period 2003–10.

[b]Cost to world economy is lost GDP for the period 2003–10.

Source: Hartley (2006b).

Defence economists have also contributed to the analysis of terrorism using both choice-theoretic and game-theoretic models. Terrorism shows that non-conventional conflict is also costly. The attacks of 11 September 2001 on the USA resulted in almost 3,000 deaths and economic losses of $80 billion–90 billion (Barros, Kollisa and Sandler, 2005). Other terrorist-related costs include nations spending on homeland security measures, on terrorist-related intelligence, on security measures in airports, the lost time waiting at airports to clear security, the losses of liberty and freedoms and the war on terror (for example, in Afghanistan and Iraq).

Choice-theoretic models of terrorism apply standard consumer choice theory with terrorists maximizing a utility function subject to budget constraints. The utility function can be specific, such as a choice between attack modes, say, skyjackings and bombings, or more generally involve a choice between terrorist and peaceful activities. The approach offers some valuable insights into terrorist behaviour and possible policy solutions. The model shows that terrorist behaviour and activities can be influenced by governments acting to reduce terrorist funds (that is, an income effect), by changing relative prices (that is, promoting a substitution effect), and by efforts to change terrorist preferences towards more peaceful activities (for example, Northern Ireland). The substitution effect is an especially powerful insight showing that policies which increase the relative price of one attack mode, such as skyjackings, will encourage terrorists to substitute an alternative and lower-cost method of attack (for example, assassinations, bombings, or kidnappings: Frey and Luechinger, 2003; Anderton and Carter, 2005).

Conclusion

Defence economics is now established as a reputable sub-discipline of economics. It shows how economic theory and methods can be applied to the defence sector embracing the armed forces, defence industries and the political–institutional arrangements for making defence choices. But this is only the beginning. Massive opportunities remain for further research in the field. Changes in threats, new technology and continued budget constraints will require further adjustments in the armed forces and defence industries, and will generate a new set of research problems. Examples include space warfare, the economics of nuclear weapons policy, assessing the efficiency of armed forces, improving the efficiency of military alliances and developing more efficient approaches to international governance and international collective action.

See Also

- arms races
- arms trade
- terrorism, economics of
- war and economics
- World Wars, economics of

Bibliography

Anderton, C. and Carter, J. 2005. On rational choice theory and the study of terrorism. *Defence and Peace Economics* 16, 275–82.

Barros, C., Kollisa, C. and Sandler, T. 2005. Security challenges and threats in a post-9/11 world. *Defence and Peace Economics* 16, 327–9.

Benoit, E. 1973. *Defense and Economic Growth in Developing Countries*. Boston: DC Heath.

BICC (Bonn International Centre for Conversion). 2005. *Conversion Survey 2005*. Baden-Baden: Nomos Verlagsgellschaft.

Brauer, J. 2003. Economics of conflict, war and peace in historical perspective, Special Issue, *Defence and Peace Economics* 14, 151–236.

Frey, B. and Luechinger, S. 2003. How to fight terrorism: alternatives to deterrence. *Defence and Peace Economics* 14, 237–49.

Goodwin, C., ed. 1991. *The Economics of National Security*. Durham and London: Duke University Press.

Hartley, K. 2006a. Defence R&D: data issues. *Defence and Peace Economics* 17(3), 1–10.

Hartley, K. 2006b. The economics of conflict. In *The Elgar Companion to Public Economics: Empirical Public Economics*, ed. A. Ott and R. Cebula. Cheltenham: Edward Elgar.

Hegre, H. and Sandler, T. 2002. Economic analysis of civil wars. Special Issue of *Defence and Peace Economics* 13, 429–96.

Hitch, C. and McKean, R. 1960. *The Economics of Defense in the Nuclear Age*. Cambridge, MA: Harvard University Press.

Kennedy, P. 1988. *The Rise and Fall of the Great Powers*. London: Fontana Press.

NATO. 2005. *NATO – Russia Compendium of Financial and Economic Data Relating to Defence*. Brussels: NATO.

OECD. 2004. *Main Science and Technology Indicators*. Paris: OECD.

Oi, W. 1967. The economic cost of the draft. *American Economic Review* 57(2), 39–62.

Olson, M. and Zeckhauser, R. 1966. An economic theory of alliances. *Review of Economics and Statistics* 48, 266–79.

Peck, M. and Scherer, F. 1962. *The Weapons Acquisition Process*. Boston: Harvard University Press.

Richardson, L. 1960. *Arms and Insecurity: A Mathematical Study of the Causes and Origins of War*. Pittsburgh, PA: Homewood.

Sandler, T. and Hartley, K. 1995. *The Economics of Defense*. Cambridge Surveys of Economic Literature. Cambridge: Cambridge University Press.

Sandler, T. and Hartley, K., eds. 2003. *The Economics of Conflict*, 3 vols. Cheltenham: Edward Elgar.

Sandler, T. and Hartley, K., eds. 2007. *Handbook of Defense Economics*, vol. 2, Amsterdam: North-Holland.

Schelling, T. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

SIPRI (Stockholm International Peace Research Institute). 2005. *SIPRI Yearbook 2005*. Oxford: Oxford University Press.

US DoS (US Department of State). 2002. *World Military Expenditures and Arms Transfers, 1999–2000*. Washington, DC: Bureau of Verification and Compliance, US Department of State.

Vahabi, M. 2004. *The Political Economy of Destructive Power*. Cheltenham: Edward Elgar.

How to cite this article

Hartley, Keith. "defence economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000247> doi:10.1057/9780230226203.0367

The New Palgrave Dictionary of Economics Online

de-industrialization, ‘ premature ’ de-industrialization and the Dutch Disease

José Gabriel Palma
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

OECD countries began de-industrializing in the late 1960s, while some high-income developing countries in East Asia entered this phase in the 1980s. Soon afterwards, some middle-income Latin American countries and South Africa also began to de-industrialize (‘prematurely’) after radical economic reforms, despite their level of income per capita being far lower than other countries which began to de-industrialize earlier. Since manufacturing is considered by many as the most effective engine of growth, it has been argued that de-industrialization could have significant negative long-term effects on growth, investment and employment, especially when done ‘prematurely’ or due to ‘Dutch disease’.

Keywords

activity-specific growth; de-industrialization; de-industrialization, ‘premature’ de-industrialization and the Dutch Disease; Dutch Disease; endogenous growth; growth policy; Hirschman, A. O.; import substitution; Kaldor, N.; Kalecki, M.; manufacturing; manufacturing employment; monetarism; of technical change; outsourcing; Post Keynesian economics; Prebisch, R.; premature de-industrialization; productivity growth; remittances; research and development; resource curse hypothesis; resource endowments; Schumpeterian growth; sector-specific growth; social capital; structuralism; terms of trade

Article

One of the most notable stylized facts of the world economy since the late 1960s is the rapid decline in manufacturing employment in industrialized countries (a fall of about 25 million jobs). Although the structure of employment has changed substantially over the long-term course of economic development, changes of the scale and speed during this period constitute an unprecedented phenomenon – manufacturing employment in the European Union, for example, has fallen by more than a third. Manufacturing is an activity considered by many as the most effective engine of growth – either because it is a crucial driver of outward shifts of the production frontier, or due to its capacity to set in motion processes of cumulative causation based on increasing returns (for example, in Post Keynesian, structuralist and Schumpeterian thought). It has therefore been argued that a process of de-industrialization on this scale is likely to have significant negative long-term effects on growth (on both its rate and its sustainability), investment, and employment.

This concern has been particularly pronounced in countries that experienced drastic de-industrialization following the discovery of mineral resources – a phenomenon that became known as the ‘Dutch Disease’. The key issue is the double-edged effect of a mineral discovery. On the one hand, it allows for an expansion of expenditure and employment; but on the other, it could easily lead to a contraction of the non-commodity tradable sector. This phenomenon was first analysed in the 1950s in relation to the mixed effects of sudden increases in mineral exports or in the price of wool for the Australian economy.

OECD countries began de-industrializing in the late 1960s, while some high-income developing countries in East Asia entered this phase in the late 1980s. Soon afterwards, some middle-income Latin American countries and South Africa also began to de-industrialize after radical economic reforms, despite their level of income per capita being far lower than other countries which began to de-industrialize earlier. This latter process has been labelled ‘premature’ de-industrialization (Palma, 2005), and should not be confused with the so-called ‘resource curse’ hypothesis, which refers to the poor macroeconomic performance of many mineral-exporting economies.

The following tables (Tables 1 and 2) show the above-mentioned regional trends in the share of manufacturing in total employment and the share of manufacturing value added in GDP.

Manufacturing employment (% of total), 1960–2003

Region	1960	1970	1980	1990	2003
Sub-Saharan Africa	4.4	4.8	6.2	5.5	5.5
•South Africa	11.3	12.8	18.2	15.7	14.1
Latin America	15.4	16.3	16.5	16.8	14.2
•Southern Cone and Brazil	17.2	17.5	17.3	17.9	13.1
Middle East and North Africa	7.9	10.7	12.9	15.1	15.3
South Asia	8.7	9.2	10.7	13.0	13.9
East Asia (excluding China)	10.0	10.4	15.8	16.6	14.9
•NICs	14.6	19.2	27.5	28.7	19.4
China	10.9	11.5	10.3	13.5	12.3
Third World	10.2	10.8	11.5	13.6	12.5
OECD	26.5	26.8	24.1	20.1	17.3

Notes: Averages are weighted by economically active population. Southern Cone: Argentina, Chile and Uruguay. NICs: Korea, Taiwan, Singapore and Hong Kong.

Sources: Calculations using International Labour Organization (ILO) databank; for Taiwan, *The Republic of China Yearbook of Statistics*.

Manufacturing value added (% of GDP), 1960–2003

Region	1960	1970	1980	1990	2003
Sub-Saharan Africa	15.3	17.8	17.4	14.9	13.8
•South Africa	21.0	23.9	22.5	25.5	18.1
Latin America	28.1	26.8	28.2	25.0	16.7
•Southern Cone and Brazil	32.2	29.8	31.7	27.7	16.9
Middle East and North Africa	10.9	12.2	10.1	15.6	14.2
South Asia	13.8	14.5	17.4	18.0	16.2
East Asia (excluding China)	14.0	19.2	23.3	25.5	27.6
•NICs	15.4	22.5	27.1	26.5	24.9
China	23.7	30.1	40.6	33.0	31.3
Third World	21.6	22.1	24.3	23.9	22.7
OECD	28.9	28.3	24.5	22.1	17.3

Note: NICs does not include Taiwan.

Source: Calculations using data (in real terms) from World Bank (1984; 2006).

The four sources of de-industrialization

The first source of de-industrialization: an ‘ inverted-U ’ relationship between manufacturing employment and income per capita.

The most commonly used concept of de-industrialization emerges from an understanding of the relationship between manufacturing employment and income per capita as an ‘inverted-U’. De-industrialization is simply the drop in manufacturing employment occurring when countries reach a certain level of income per capita – that is, mature economies switching employment to specialized services as part of their ‘normal’ process of development. As such, de-industrialization could well have positive long-term growth effects. According to Rowthorn (1994), using data for 1990, this drop begins at US$12,000 (Figure 1).

Figure 1
Rowthorn's regression: manufacturing employment and income per capita, 1990. *Source*: Rowthorn (1994).

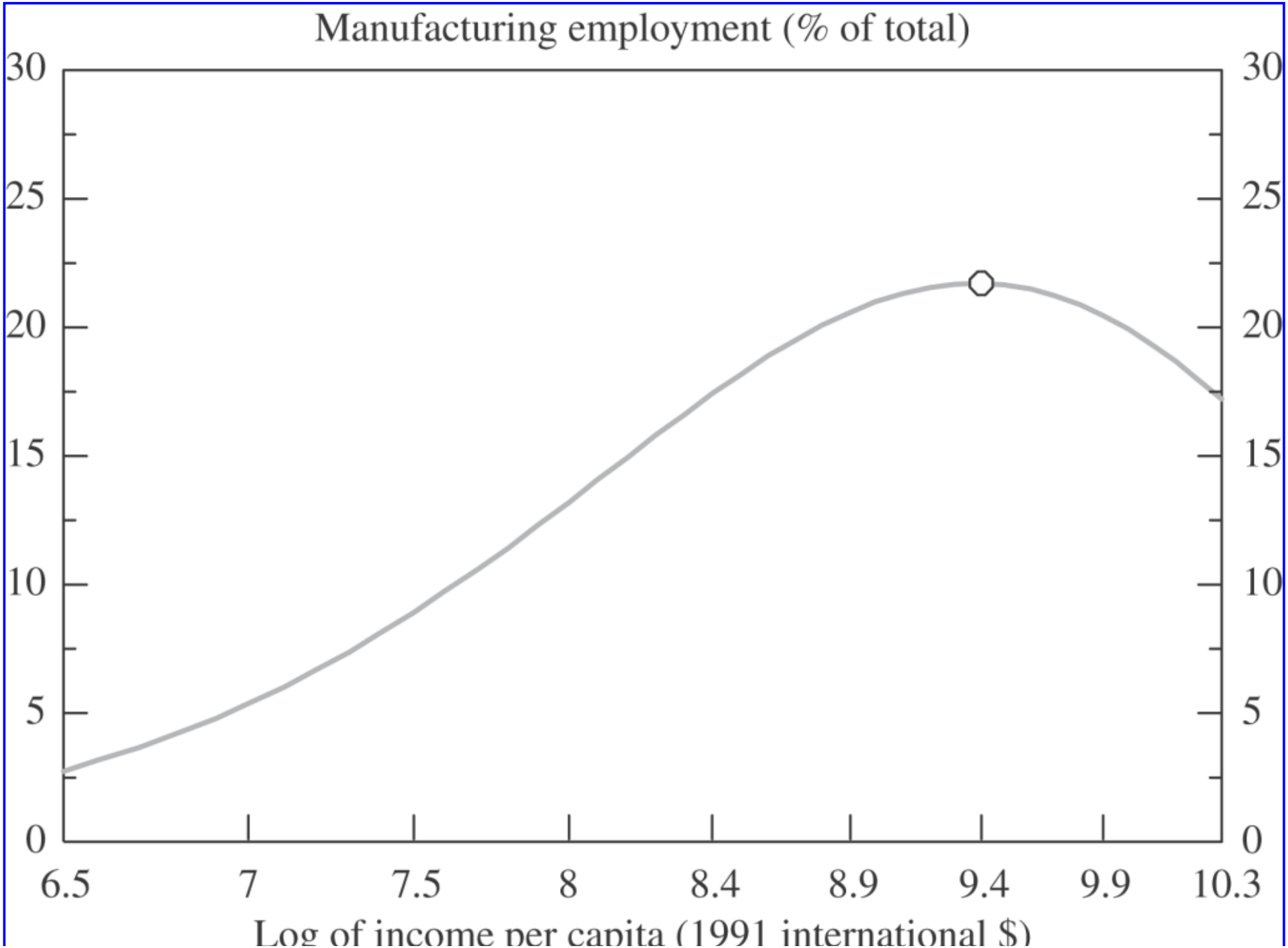

Although many other analyses are consistent with this hypothesis, Palma (2005) has suggested that de-industrialization has been a more complex phenomenon. He argues that, in addition to the ‘inverted-U’, there are three further processes at work.

ment

One additional source of de-industrialization has been the remarkable collapse of the ‘inverted-U’ relationship over time.
In essence, for high- and middle-income countries, the level of manufacturing employment associated with a given level of income per capita has been falling over time. In fact, the four better-known hypotheses originally developed to explain the ‘inverted-U’ relationship mentioned above are more relevant to this ‘second source’ of de-industrialization, as until the mid-1980s no country had reached the level of income corresponding to the turning point of the respective curve. These hypotheses are as follows:

- The fall in manufacturing employment is merely a statistical illusion caused primarily by the re-allocation of labour from manufacturing to services through contracting-out of activities such as transport, cleaning, design, security, catering, recruitment and data processing. This process could be the result of a further movement in a long line of progressive transformations aiming at enlarging the scope for specialization and the division of labour, or just a cost-cutting operation aimed, for example, at bypassing labour legislation.
- The fall in manufacturing employment results from a reduction in the income elasticity of demand for manufactures, particularly in high income countries.
- It is the consequence of higher productivity growth in manufacturing than in other sectors of the economy.
- It is the result of a new international division of labour (including ‘outsourcing’), which has a negative impact on manufacturing employment in industrialized countries, especially for non-skilled labour.

Although a detailed analysis of the role of each of these factors in de-industrialization is outside the scope of this article (see Rowthorn and Ramaswamy, 1999), it is important at least to add that the 1980s switch in ‘policy regime’ in OECD countries (broadly speaking, from post-war Keynesianism to demand-constraining monetarism) did also contribute to the huge 1980s drop in manufacturing employment. (For the 1980s debate on de-industrialization, see Singh, 1987.) The technological revolution that took off in the 1980s also played a major role (Pérez, 2002).

The third source of de-industrialization: changing income per capita corresponding to the turning point of the regression

This additional source of de-industrialization is also evident in Figure 2 (see also Figure 3). This concerns the remarkable leftwards movement in the turning point of the regressions during the 1980s. (Rowthorn and Wells, 1987, had discussed the possibility of the ‘inverted-U’ relationship peaking at a lower level of income per capita over time.) During the 1980s the income per capita at which the curve peaked fell by about half – from approximately $21,000 in 1980 to just over $10,000 in 1990 (in 1985 international US$). Until 1980 no country was located to the right of the turning point of the corresponding cross-section curve, but in the 1990 regression there were more than 30 countries beyond that critical point. However, during the 1990s this process was reversed, and by 2000 again no country was beyond that critical point.
Figure 2
Second source of de-industrialization: a declining relationship, 1960–2000. *Notes*: The range of the horizontal axis is the actual income range of the sample for 2000. The regressions are based on a sample of 105 countries. In all regressions in this and the following figures, all parameters are significant at the 1% level, and the adjusted R^2 are between 66% and 77%. All regressions pass the relevant diagnostic tests. Note that these regressions are simply a cross-sectional description of cross-country differences in manufacturing employment, when categorized by income per capita; hence they should not be interpreted in a ‘predicting’ way, because there are a number of difficulties with a curve estimated from a single cross-section – especially regarding the homogeneity restrictions that are required to hold. *Source*: Palma (2005), using ILO Databank and the Penn World Tables; this is the source for all figures other than Figure 8.

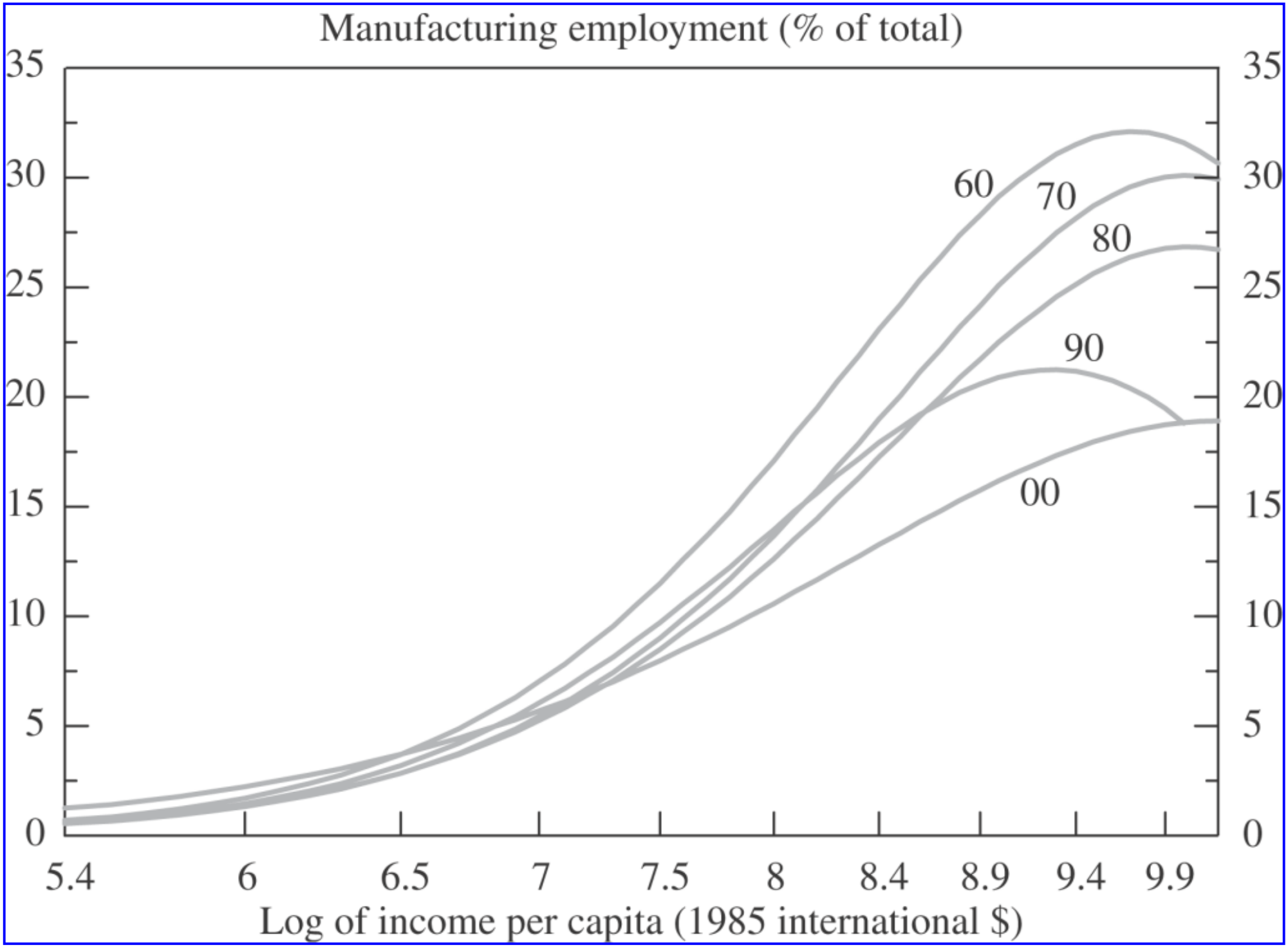

Figure 3

The changing nature of de-industrialization between the 1980s and the 1990s. *Note*: The range of the horizontal axis is the actual income range of the sample for 2000.

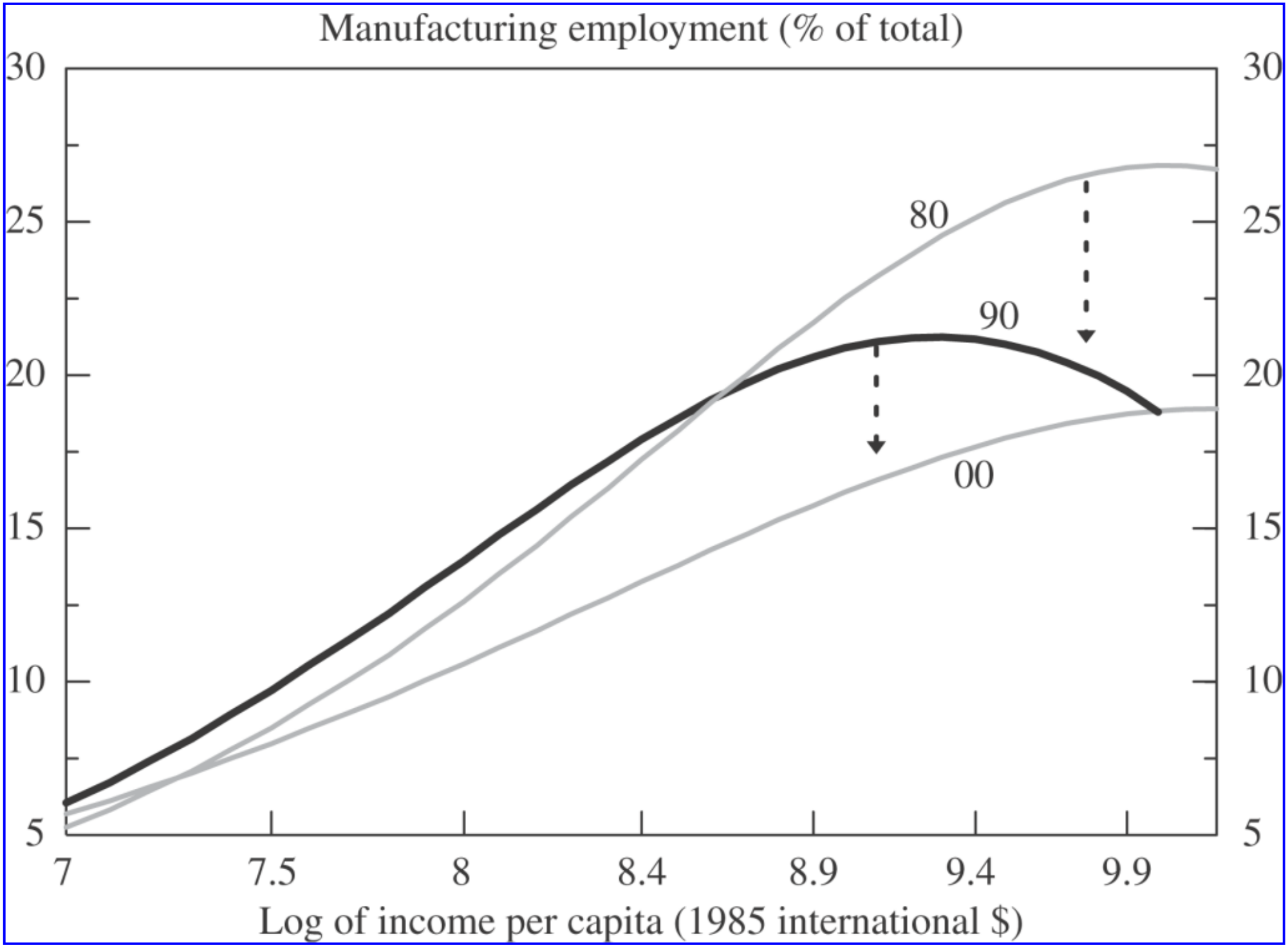

The changing shape of the curves is crucial to the understanding of the dynamic of the interrelationship between the three sources of de-industrialization discussed so far. Basically, as the arrows of Figure 3 indicate, during the 1980s there was a remarkable degree of de-industrialization in high-income countries; during the 1990s, by contrast, de-industrialization

affected mainly middle-income countries.

The fourth source of de-industrialization: the Dutch Disease

Finally, in several countries we can observe a further degree of de-industrialization. These countries experienced a fall in their manufacturing employment that was clearly greater than would have been expected, given the three sources of de-industrialization discussed above (Figure 4).

Figure 4
Fourth source of de-industrialization: cases of ‘overshooting’? *Notes*: Neth: The Netherlands; Bra: Brazil.

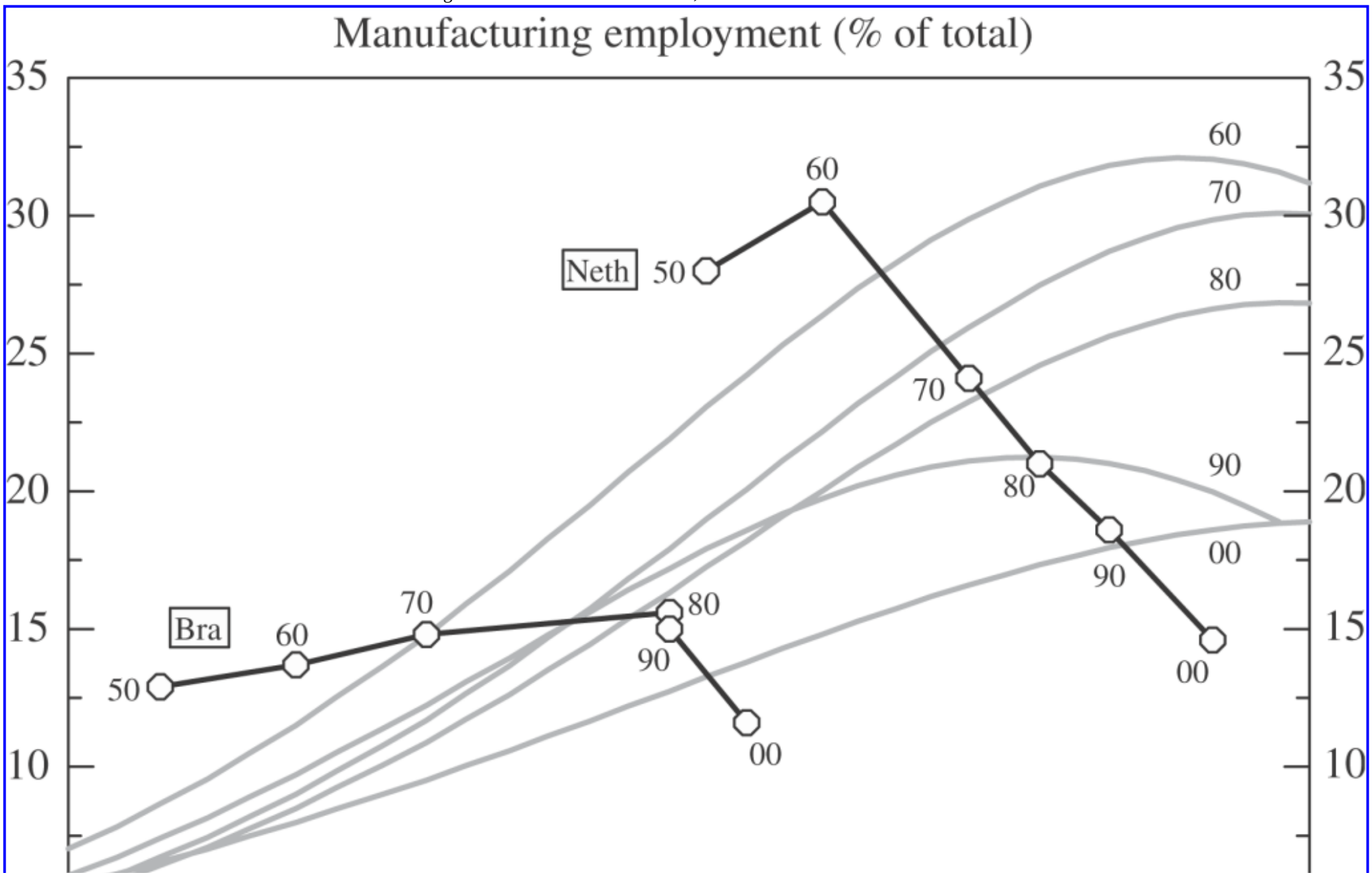

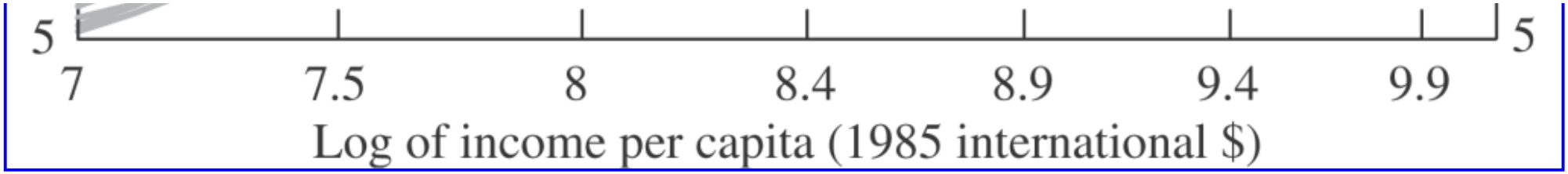

Rather than simple cases of ‘overshooting’, Palma (2005) identifies this phenomenon with a specific conceptualization of the Dutch Disease: in countries that have an export surge of commodities or services, or a major shift in economic policy, a unique *additional* degree of de-industrialization is typical (that is, additional to the three de-industrialization forces already discussed).

Originally, Dutch Disease had a narrow meaning – the appreciation of the real exchange rate resulting from a boom in commodity exports. (For an analysis of the macro-processes at work, see Corden and Neary, 1982; Ros, 2000.) Elsewhere, mostly in neoclassical models, it simply referred to the adverse terms of trade effect for tradables following a sudden shift in their production frontier. However, with time the meaning has widened to include all possible negative macroeconomic effects associated with the ‘resource curse’ hypothesis—for Woolcock, Pritchett and Isham (2001), for example, resource-rich countries are not very good at accumulating social capital. (See also Mehlum, Moene and Torvik, 2006; for a critical analysis of the ‘resource curse’ hypothesis, see DiJohn, 2007.)

The origins of this ‘disease’ lie in the fact that the relationship between manufacturing employment and per capita income tends to differ between those countries that generate a trade surplus in manufacturing and those that do not. Note that the ‘trade surplus in manufacturing’ group includes economies that find themselves in this position out of *necessity* as well as others due to *growth policy*. In the first case, given resource endowments force some countries to aim for a manufacturing surplus to finance inevitable trade deficits in commodities and/or services (for example, Japan and India). In the second, some resource-rich countries still try to achieve a trade surplus in manufacturing by implementing a growth policy based on a strong ‘industrialization agenda’ (for example, Finland, Malaysia and Vietnam).

Figure 5 shows the long-term changes between manufacturing employment and income per capita in the ‘trade surplus in manufacturing’ (mf) and in the ‘trade surplus in primary commodities or services’ (pc) groups of countries.

Figure 5

Changes in manufacturing employment and income per capita, 1960–2000. *Note*: An intercept dummy differentiates the two groups of countries.

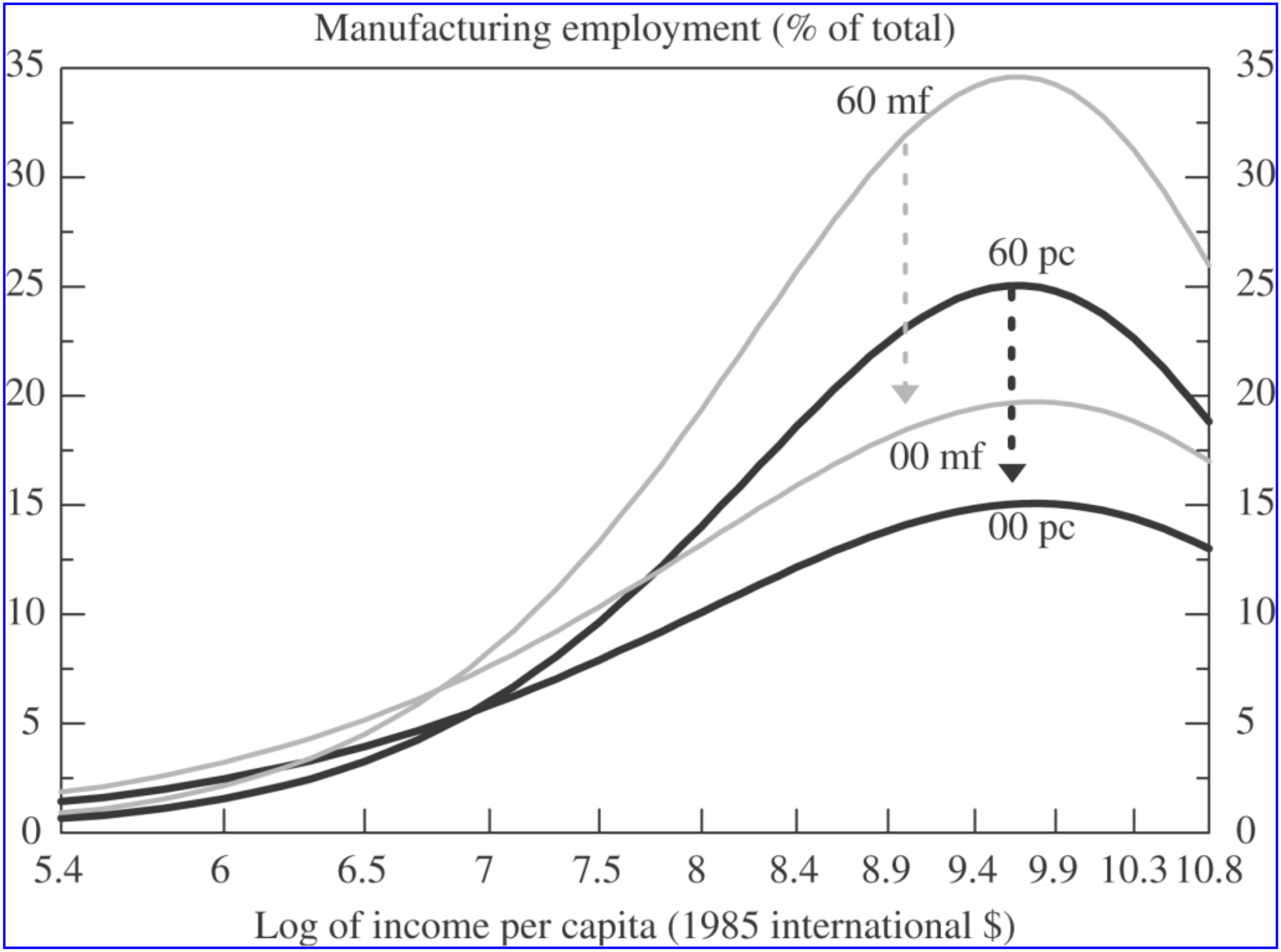

Although the ‘pc’ countries tend to reach a lower level of industrialization at any given point in time, the ‘pc effect’ per se has not led to a higher degree of de-industrialization. In

fact, if we take the highest point of the curves, in these four decades the share of manufacturing employment in both ‘mf’ and ‘pc’ countries dropped by about half.

After this introduction, it is now possible to explain the concept of Dutch Disease. There is a group of countries – both industrialized and developing – that exhibit a specific *additional* degree of de-industrialization. The Netherlands rightly gives its name to this phenomenon.

From this perspective, what happened in the Netherlands was a discovery of a natural resource (gas) leading manufacturing employment to switch from an ‘mf’ structure to a ‘pc’ one. When this occurs, as Figure 6A shows, the country experiencing this ‘disease’ moves along *two* different paths of de-industrialization. The first path consists of the three processes of de-industrialization discussed above (from ‘60-mf’ to ‘00-mf’). The second corresponds to a further component of de-industrialization resulting from the change in the reference group (from ‘00-mf’ to ‘00-pc’). In this context, the Dutch Disease should be regarded only as the *additional* level of de-industrialization associated with the latter movement. In the case of the Netherlands, then, it is the (five percentage points) difference between manufacturing employment falling by 10.9 percentage points between 1960 and 2000 (hypothetical non-Dutch Disease scenario), or by 15.9 percentage points (actual Dutch-Disease situation) – that is, manufacturing employment falling from 30.5 per cent of total employment in 1960, to 19.6 per cent in 2000 in the former scenario, or to 14.6 per cent in the latter one.

Figure 6

(A) The Netherlands: unravelling the Dutch Disease, 1960–2000. (B) The United Kingdom; catching the Dutch Disease, 1960–2000. *Notes*: Ne: The Netherlands; UK: United Kingdom. (C) Five countries of the European Union, 1960–2000. *Notes*: EU5: Germany, France, Italy, Belgium and Austria. (D) Four traditional primary commodity exporters, 1960–2000. *Notes*: UCAN: United States, Canada, Australia and New Zealand.

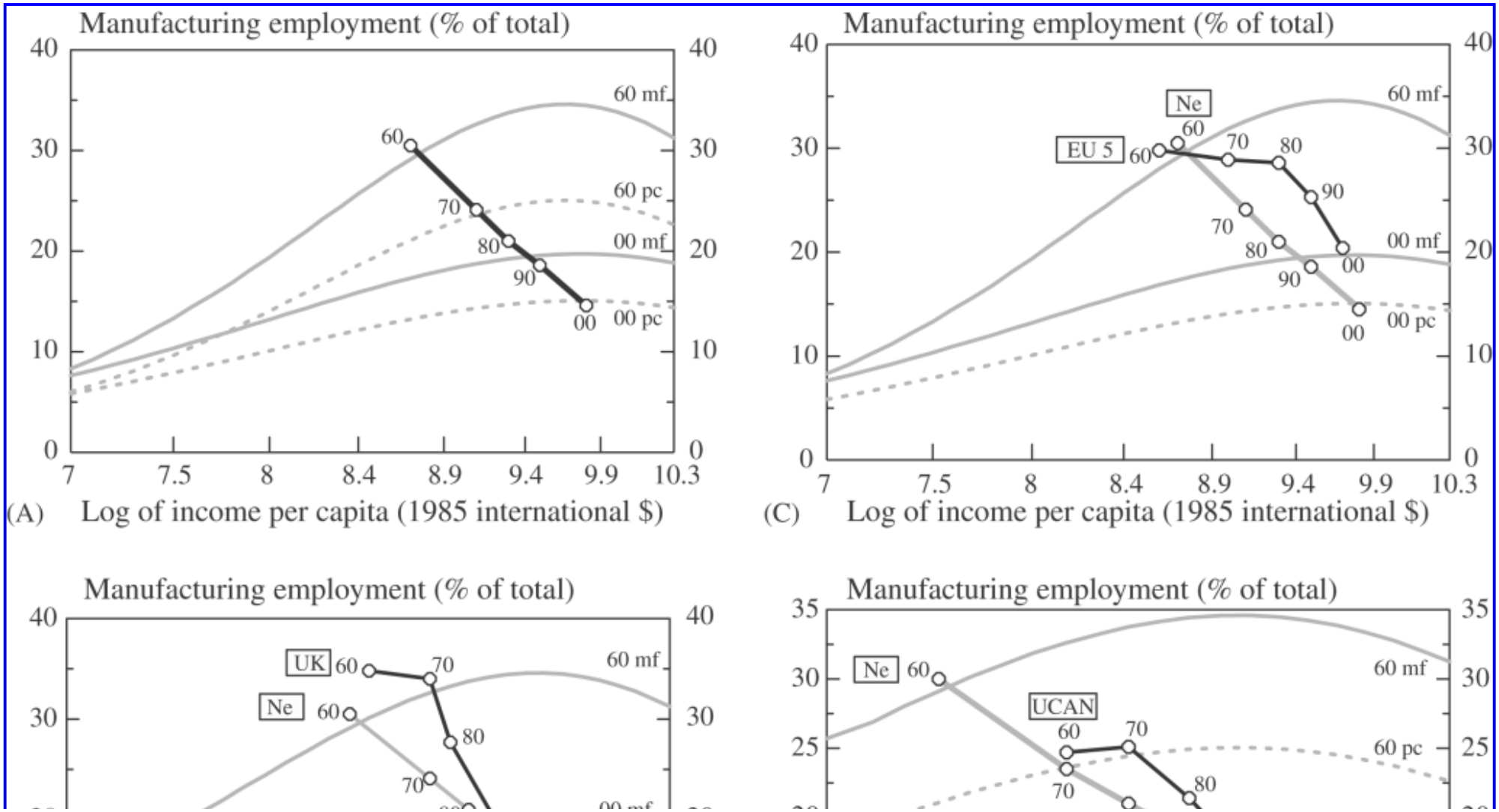

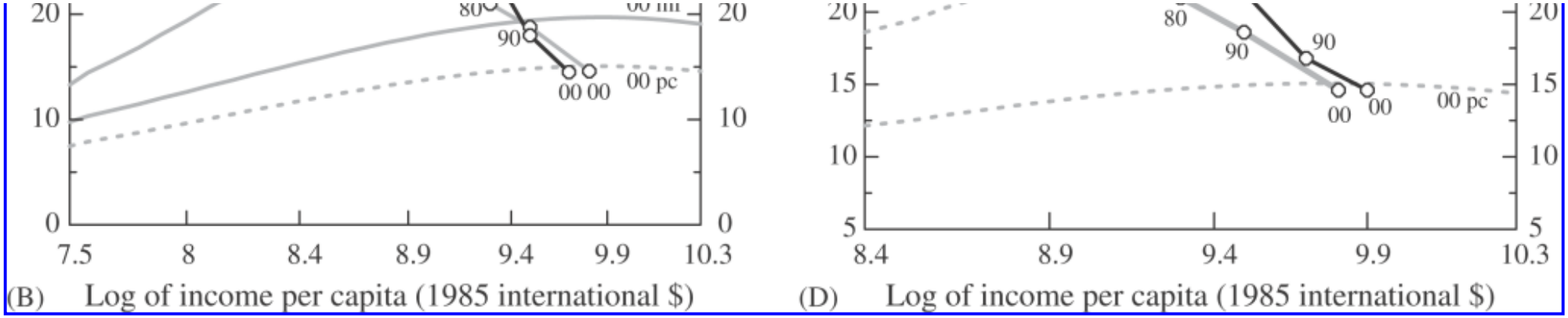

Dutch Disease is thus clearly not a phenomenon limited to the Netherlands or to the discovery of mineral resources; it has also occurred in other countries and for other reasons. One case is the United Kingdom, which had a boom in both oil and financial-services exports (see Figure 6B). As a result of this (and of Prime Minister Thatcher) the trade balance in manufacturing switched from a surplus of four per cent of GDP (late 1970s) to a deficit of four per cent (mid-2000s). Figure 6C shows that, by contrast, the share of manufacturing employment in other EU countries fell only according to the changes in the ‘mf’ scenario. In turn, Figure 6D shows that, although four other industrialized countries (major commodity exporters throughout the period) also found themselves in the ‘pc’ category in 2000, they did not suffer from the Dutch Disease simply because they were in the ‘pc’ category from the start. Although both the ‘EU-5’ and the ‘UCAN’ countries experienced a *similar* drop in the share of manufacturing employment (9.7 and 10.5 percentage points, respectively), neither switched from one reference group to another.

The phenomenon of the Dutch Disease also occurred in countries that developed flourishing service-exporting sectors, such as tourism (for example, Greece, Cyprus and Malta) and financial services (for example, Switzerland, Luxembourg and Hong Kong); see Palma (2005).

Finally, this ‘disease’ was also experienced after 1980s in some middle-income Latin American countries (and to some extent in South Africa) where state-led import-substituting industrialization (ISI) had achieved industrialization levels characteristic of the ‘mf’ group (despite the fact that these countries generated large trade surpluses in commodities). In these cases, radical change of the economic policy regime (from ISI to comprehensive trade and financial liberalization) resulted in the Dutch Disease; that is, the transformation of their employment structure from a policy-induced ‘mf’ to a more ‘Ricardian’ resource-rich ‘pc’.

Brazil and the three Southern Cone countries experienced the greatest de-industrialization following their economic reforms, while also being among the countries of the region that had previously industrialized the most and that had subsequently implemented the most drastic reforms (Figure 7).

Figure 7

Argentina, Brazil, Chile and Uruguay: catching the Dutch Disease, 1960–2000. *Notes*: Ar: Argentina; Br: Brazil; Cl: Chile; Ur: Uruguay; Ne: The Netherlands. The year 1990 has been omitted not to ‘congest’ the figure. South Africa's share of manufacturing employment also fell from an ‘mf’ level in 1980 to close to a ‘pc’ one in 2000.

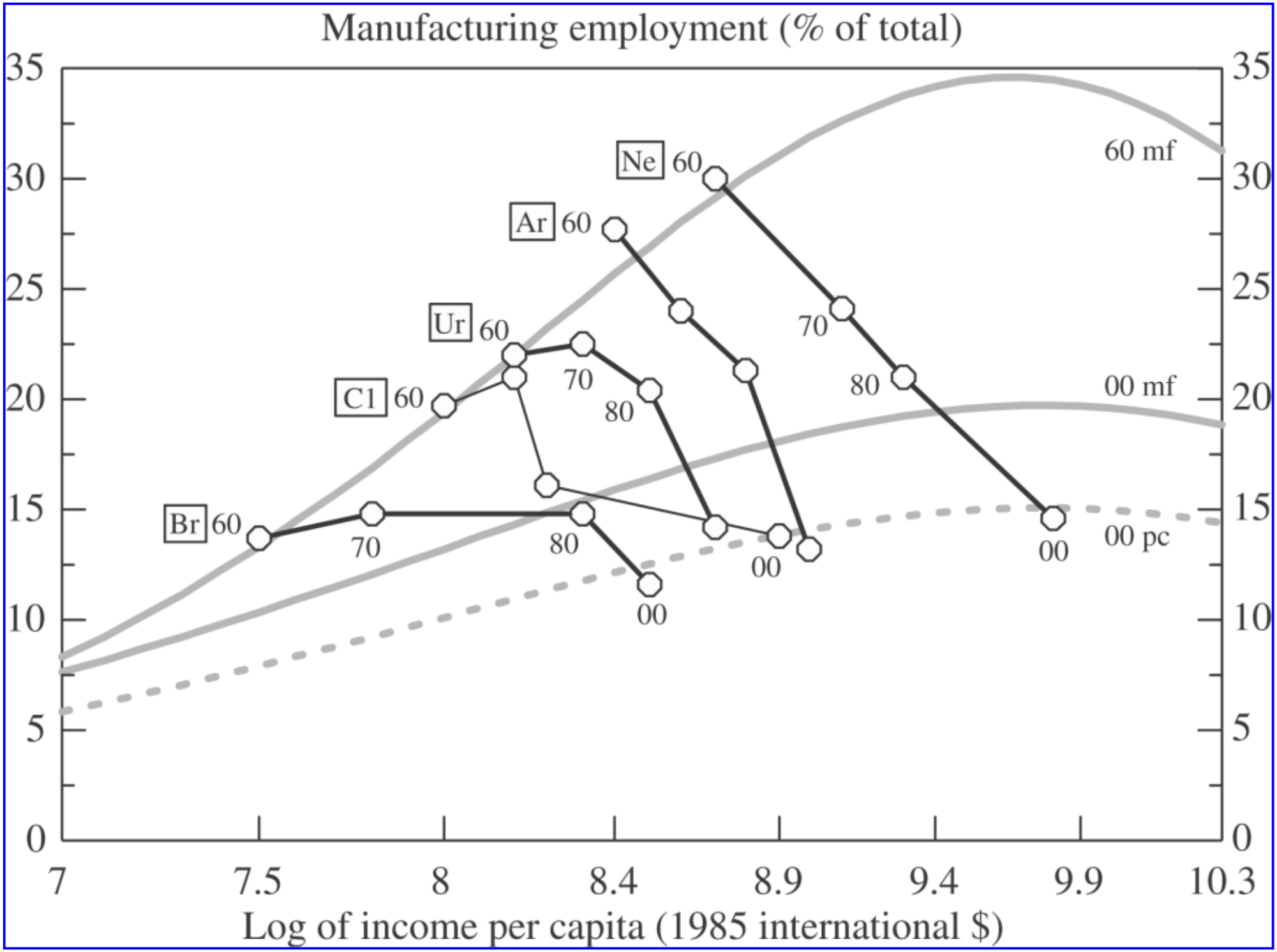

These four Latin American countries began this period – as did the Netherlands – with a level of manufacturing employment typical of countries aiming at a trade surplus in

manufacturing (‘60-mf’), even though these resulted from different causes. The case of the Netherlands is due to its (pre-natural gas) poor resource endowment, whereas in the four Latin American countries their position was the result of a ‘structuralist’ industrialization agenda (see structuralism). And if both reached 2000 with levels of manufacturing employment typical of the ‘pc’ group, this was once again for different reasons: in the Netherlands, the discovery of a natural resource at a ‘mature’ stage of industrialization was decisive, whereas in Latin America the sharp reversal of the ISI policies was responsible. Note that in the latter the ‘extra’ degree of de-industrialization (‘mf’ to ‘pc’) took place over and above the already mentioned huge collapse of the ‘mf’ path for middle-income countries during the 1990s (Figure 3).

From this perspective, the key difference between developing Asia and ‘premature de-industrializers’ in Latin America in terms of the implementation of economic reforms is that in the latter these reforms seem to have obstructed their transition towards a more mature – that is, self-sustaining – form of industrialization. (For the concept of ‘self-sustaining industrialization’, see Kaldor, 1967.) Resource-poor *and* resource-rich developing Asia, instead, succeeded in combining these reforms with a dynamic ‘mf’ path (Figure 8).

Figure 8

Brazil, Argentina and China: manufacturing production, 1965–2005. *Notes*: M&T: Malaysia and Thailand. Three-year moving averages. The relative decline of South Africa's manufacturing sector is even greater than Brazil's (though not as extreme as Argentina's). Manufacturing output measured in US$, 2000. *Source*: World Bank (2006).

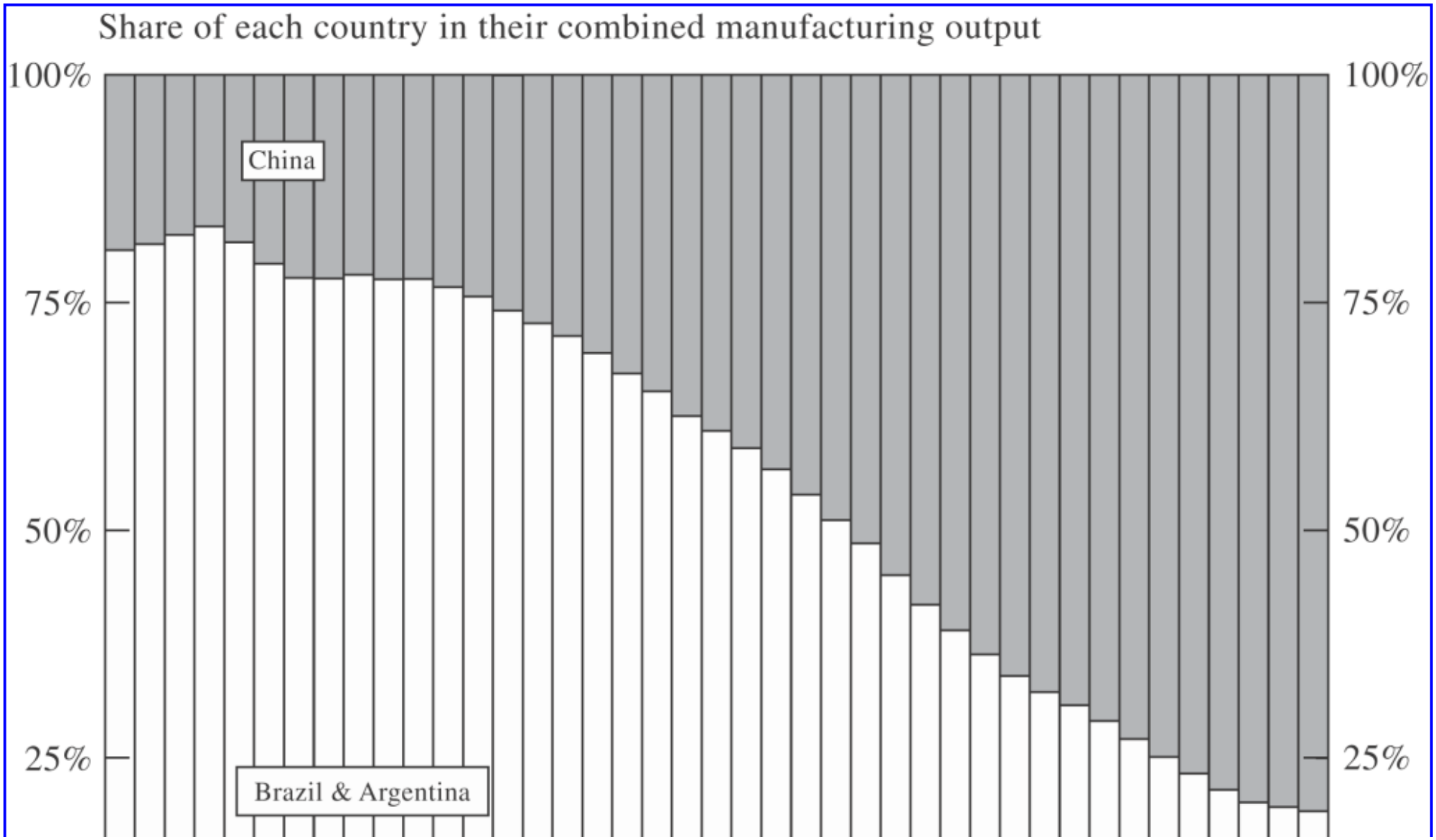

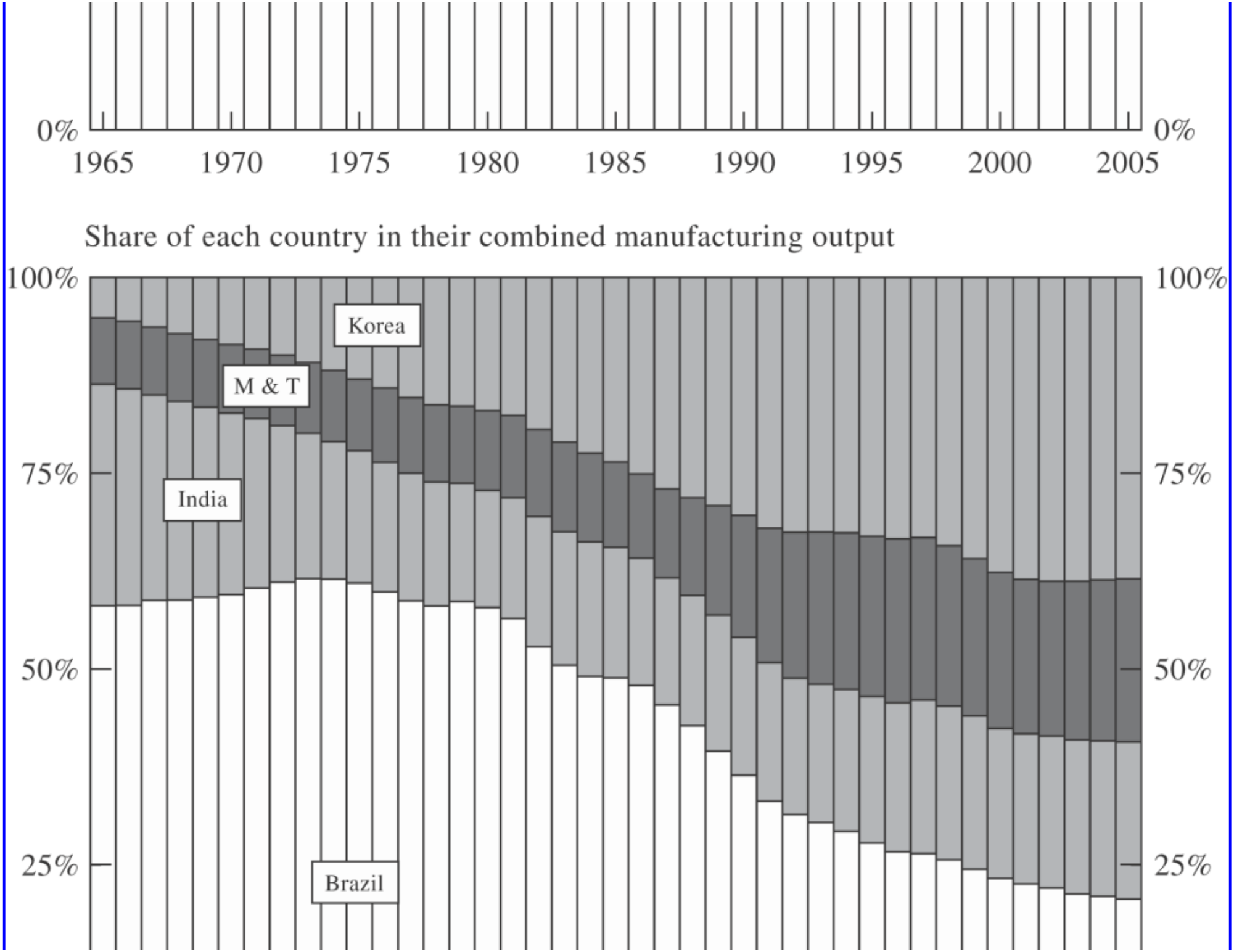
0%
0%
1965
1970
1975
1980
1985
1990
1995
2000
2005
Share of each country in their combined manufacturing output
100%
100%
75%
75%
50%
50%
25%
25%
Korea
M & T
India
Brazil

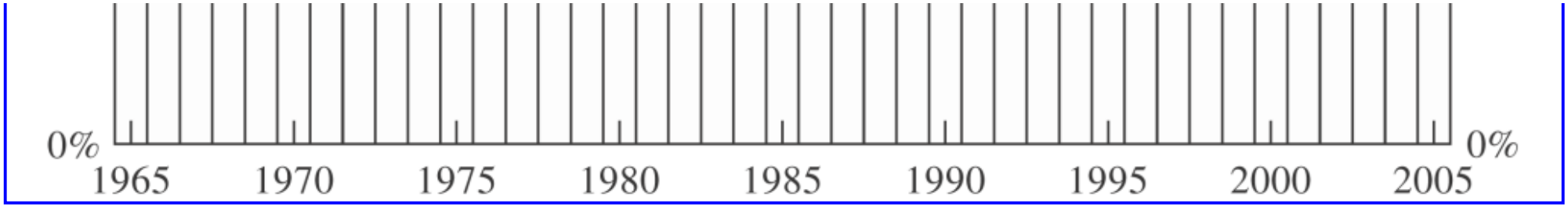

Perhaps Latin America is in desperate need of a touch of the so-called East Asian Confucianism; that is, once a new development path has been chosen, a significant degree of pragmatism, self-confidence, a progressive capitalist elite and an avant-garde political leadership can be of great assistance in policymaking success.

In sum, the Dutch Disease should not be seen as simple ‘overshooting’ of de-industrialization, but as a specific type of ‘additional’ de-industrialization. In general, this has taken place for one of three different reasons: the discovery of natural resources (for example, the Netherlands); the development of export-service activities, mainly tourism and finance (for example, Greece in the former, and Hong Kong in the latter); and finally, changes in economic policy (for example, Brazil and South Africa).

All the above types of de-industrialization should also be distinguished from those of the late 1980s and 1990s in many sub-Saharan economies and countries of the former Soviet Union and eastern Europe, which experienced a process of de-industrialization associated with a *fall* in income per capita: a case of de-industrialization ‘in reverse’.

Finally, Finland, Sweden, Denmark, Malaysia, Vietnam and, to a lesser extent, other south-east Asian resource-rich countries (such as Thailand and Indonesia) prove both that economic policies do exist to avoid the Dutch Disease in commodities- and export-services-booming economies (see Pesaran, 1984; Palma, 2000), and that there is no such thing as an unavoidable ‘curse of natural resources’. Countries with high potential for developing commodities and export-services activities have sufficient degrees of policy freedom to follow ambitious and successful ‘industrialization agendas’ (not least of the commodities themselves, as in the Nordic case and Malaysia). Also, export rents could be used effectively in that direction. However, as the Latin American experience in particular shows, it seems that, as globalization progresses, fewer and fewer countries are willing to take advantage of such degrees of policy freedom. This is not only because forces in the new international institutional and financial order are rapidly working to reduce these degrees of policy freedom, but also because of domestic changes in economic ideologies and the structure of property rights.

However, whether a process of structural change that includes ‘premature’ de-industrialization can deliver rapid and *sustainable* economic growth is another matter altogether; so is the issue of whether the current ‘premature’ de-industrialization occurring in Latin America and South Africa contains important components of policy-induced ‘uncreative destruction’.

De-industrialization: does it matter?

Rapid de-industrialization has reopened an age-old debate in economic theory: is a unit of value added in manufacturing equal to one in commodities or services, especially in terms of its growth-enhancing properties?

Although a detailed discussion of this debate is beyond the scope of this article, from the perspective of de-industrialization we may classify growth theories into three groups (in doing this, of course, we have to acknowledge the necessary degree of simplification which every classification of intellectual tendencies entails). This requires a distinction between two concepts: ‘activity’ and ‘sector’. Examples of the former include R&D and education, and of the latter manufacturing. The first camp of growth theories includes those (mainly neoclassical models) that treat growth as both ‘sector-*indifferent*’ and ‘activity-*indifferent*’. Examples are Solow-Swam-type models (both traditional and ‘augmented’ ones), and the branch of ‘endogenous’ theories that associates growth with increasing returns which are activity-indifferent. Examples are early ‘AK’ models and more recent ones in which changes in the rate of growth are the result of the cumulative effect of market imperfections arising in the process of technical change. However, these imperfections, and the associated increasing returns, are somehow seen as stemming directly from within the production function (rather than being based on the use of R&D or the production of human capital).

The second camp still regards growth as ‘sector-indifferent’, but models it as ‘activity-*specific*’ (for example, Romer's work and neo-Schumpeterian models). In these models, increasing returns, though generated by research-intensive activities, are explicitly not associated with manufacturing activities as such or with investment in manufacturing; nor do they allow for specific effects from manufacturing on R&D activities (except that investment in *any* sector could be ‘complementary’ to R&D through its effect on the profitability of research; see Aghion and Howitt, 1998). Therefore, in these models there is no room for Kaldorian-style effects concerning investment embedding or embodying technical change.

Finally, in the third camp are those (mainly Post Keynesian, Schumpeterian and structuralist theories) that argue that growth is both ‘sector-*specific*’ and ‘activity-*specific*’ (but the latter only in the sense that it is specific to the nature of the sector involved). For instance, the approaches to growth found in Hirschman, Kaldor, Kalecki, Prebisch, Furtado, Thirlwall and (arguably) Schumpeter follow this line of argument. What is common to these ‘sector-specific’ growth theories is that the pattern, the dynamic and the sustainability of growth are crucially dependent on the activities being developed. In particular, there are specific growth enhancing effects associated with manufacturing due to its capacity to set in motion processes of cumulative causation. This is because ‘learning by doing’, dynamic economies of scale, increasing returns, externalities and spillover effects are more prevalent

in manufacturing than elsewhere in the economy. Therefore, the crucial feature distinguishing this camp from the previous two is that issues such as technological change, synergies, balance-of-payments sustainability and the capacity of developing countries to ‘catch up’, are directly linked to the size, strength and depth of the manufacturing sector.

Then, in terms of the possible growth consequences of de-industrialization, the first growth camp does not regard de-industrialization as a particularly relevant growth issue per se. Even when it becomes a major growth or employment issue, this is only due to market imperfections. For example, Sachs and Warner (1997) argue that if neoclassical competitive conditions prevail, a declining manufacturing sector implies no hindrance to growth or full employment. Furthermore, for these growth theories, even if the discovery of natural gas did produce some structural changes in output and employment in the Dutch economy, labelling these transformations a ‘disease’ would be a misleading dramatization. Also, from this perspective, if ‘premature’ de-industrialization in resource-rich countries consists of the transformation of employment structures from an artificially policy-induced ‘mf’ to a more Ricardian ‘pc’ path, that can hardly be bad for growth!

From the point of view of the second camp, de-industrialization in ‘mature’ economies may or may not have an impact on growth per se; this would all depend on the specific form that the de-industrialization takes. For instance, it could actually result in a stimulus for growth if the ‘upward’ de-industrialization in mature economies is associated with the reallocation of resources within manufacturing into more R&D-intensive products. However, in the case of ‘premature’ de-industrialization in middle-income countries it is more difficult to argue from this approach that such a phenomenon could be positive for long-term growth.

Finally, except for normal (or ‘upward’) de-industrialization in properly mature economies, the third approach to economic growth understands de-industrialization and the Dutch Disease as unequivocally negative for growth and employment – especially if it involves ‘premature’ de-industrialization in developing countries. The same is true of the current narrowing-down of the policy space to fight them. For example, an interpretation from this perspective of the industrialized countries’ remarkable slowdown in productivity growth since the mid-1970s could be that this may well be the result of ‘wrong’ policies (such as monetarism) and ‘wrong’ structural change (such as ‘financialization’) excessively intensifying de-industrialization in the 1980s. (‘Financialization’ is the rise in size and dominance of the financial sector relative to the non-financial sector, as well as the diversification towards financial activities in non-financial corporations.) And one interpretation of the remarkably poor growth performance of most Latin American economies and South Africa since economic reform, especially Brazil, would be that this is the likely consequence of ‘premature’ de-industrialization – affecting not just the pace of their economic growth but its sustainability.

See Also

- Kaldor, Nicholas
- Kalecki, Michal
- Post Keynesian economics
- Prebisch, Raúl
- Robinson, Joan Violet
- Schumpeter, Joseph Alois
- structuralism
- unemployment
- Young, Allyn Abbott

I am extremely grateful to Fiona Tregenna for many constructive comments.

Bibliography

Aghion, P. and Howitt, P. 1998. *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Auty, R.M., ed. 2001. *Resource Abundance and Economic Development*. New York: Oxford University Press.

Corden, W.M. and Neary, J.P. 1982. Booming sector and Dutch disease economics. *Economic Journal* 92, 825–48.

DiJohn, J. 2007. *The Political Economy of Late Industrialization in Oil-Exporting Countries*. Pennsylvania: Penn State University Press.

Kaldor, N. 1967. *Problems of Industrialization in Underdeveloped Countries*. Ithaca: Cornell University Press.

Mehlum, H., Moene, K. and Torvik, R. 2006. Institutions and the resource curse. *Economic Journal* 116, 1–20.

Palma, J.G. 2000. Trying to ‘tax and spend’ oneself out of the Dutch-Disease: the Chilean economy from the war of the Pacific to the Great Depression. In *An Economic History of Latin America*, ed. E. Cárdenas, J.A. Ocampo and R. Thorp. Basingstoke: Palgrave.

Palma, J.G. 2005. Four sources of ‘de-industrialisation’ and a new concept of the ‘Dutch-disease’. In *Beyond Reforms: Structural Dynamic and Macroeconomic Vulnerability*, ed. J. A. Ocampo. Palo Alto: Stanford University Press and the World Bank.

Pérez, C. 2002. *Technological Revolutions and Financial Capital*. Cheltenham: Elgar.

Pesaran, H. 1984. Macroeconomic policy in an oil-exporting economy with foreign exchange controls. *Economica* 49, 253–70.

Pieper, U. 2000. De-industrialisation and the social and economic sustainability nexus in developing countries. *Journal of Development Studies* 36, 66–99.

Ros, J. 2000. *Development Theory and Economic Growth*. Ann Arbor: University of Michigan Press.

Rowthorn, R. 1994. Korea at the cross-roads. Working Paper No. 11, Centre for Business Research, Cambridge.

Rowthorn, R. and Ramaswamy, R. 1999. Growth, trade and deindustrialization. *IMF Staff Papers* 46(1), 18–41.

Rowthorn, R. and Wells, J. 1987. *De-Industrialisation and Foreign Trade*. Cambridge: Cambridge University Press.

Sachs, J.D. and Warner, A.M. 1997. Natural resource abundance and economic growth. Working paper, HIID, Harvard University.

Singh, A. 1987. Manufacturing and de-industrialization. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Thirlwall, A. 2002. *The Nature of Economic Growth*. Cheltenham: Elgar.

Woolcock, M., Pritchett, L. and Isham, J. 2001. The social foundations of poor economic growth in resource-rich countries. In *Resource Abundance and Economic Development*, ed. R.M. Auty. New York: Oxford University Press.

World Bank. 1984. *World Development Indicators 1984*. New York: Oxford University Press.

World Bank. 2006. *World Development Indicators 2006*. Washington, DC: World Bank.

How to cite this article

Palma, José Gabriel. "de-industrialization, ‘premature’ de-industrialization and the Dutch Disease." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000268> doi:10.1057/9780230226203.0369

The New Palgrave Dictionary of Economics Online

demand price

J.K. Whitaker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Earlier economic literature doubtless contains casual usages of the phrase ‘demand price’, but its appropriation as a technical term appears to date from Alfred Marshall's *Principles of Economics* (Marshall, 1890: see Marshall, 1920, pp. 95–101). Marshall applied the term in the contexts of both individual and market demand. Starting with a commodity (tea) purchasable in integral units of a pound's weight, an individual's demand price for the xth pound is the price he is just willing to pay for it given that he has already acquired $x - 1$ pounds. The basic assumption is that this demand price is lower the larger is x. A schedule of demand prices for all possible quantities (values of x) defines the consumer's demand schedule. Its graph is naturally drawn with quantity on the horizontal axis. In the case of a perfectly divisible commodity, the demand price of quantity x must be redefined as the price *per unit* which the consumer would be willing to pay for a tiny increment, given that he already possesses amount x. The demand schedule then graphs as a continuous negatively sloped demand curve showing demand price in this sense as a function of x.

If the individual is free to buy any quantity at a fixed price, his ‘marginal demand price’ is the demand price for that quantity ‘which lies at the margin or terminus or end of his purchases’ (Marshall, 1920, p. 95). For a perfectly divisible commodity, marginal demand price must equal market price. For a commodity purchasable in integral units only, market price may lie anywhere below marginal demand price, but not so low as to make the next unit marginal.

Marshall's discussion of consumer behaviour is based on two general assumptions, although these are informally relaxed at various points. The first is that the utility obtained from consuming a commodity depends only on the amount of that commodity. The second is that the marginal utility of ‘money’, or expenditure on all other goods, remains approximately constant with respect to variation in the expenditure on any particular commodity – the presumption being that the latter expenditure is only a small fraction of total expenditure. These assumptions have convenient consequences for the concept of demand price. If $u(x)$ denotes the utility a consumer obtains from consuming quantity x of a given good in a specified period, while λ is the constant marginal utility of money to him, then demand price for quantity x is $(du/dx)/\lambda$ in the case of divisible quantity and $[u(x) - u(x-1)]/\lambda$ in the case when only integral quantities are feasible. In either case, given the value of λ, demand price depends on x alone

and is proportional to marginal utility. The hypothesis of diminishing demand price is tantamount to that of diminishing marginal utility. A further advantage is that the demand price for quantity x is independent of the pecuniary terms on which the earlier units were, or are to be, acquired, as these terms will not change the marginal utility of money.

Although demand price is, on the above assumptions, proportional to marginal utility it has the great advantage of being measured in operational money units. This permits a monetary measure of the net benefit or consumer surplus obtained from the option of buying the commodity in question on specified monetary terms, rather than having to divert the expenditure to other goods. The distinction between demand price and market price is an operational version of the classical distinction between value in use and value in exchange.

The concept of demand price features prominently in Marshall's analysis of the market for a single commodity sold at a fixed price which is uniform to all buyers. Demand price is now interpreted as the maximum *uniform* price at which any specified aggregate quantity of the commodity can be sold on the market during a given period. The negatively sloped market demand curve is simply a lateral addition of the individual demand curves and expresses the common demand price as a function of the aggregated quantity. Marshall recognized (1920, p. 457n) that it would be more natural when dealing with market demand to view quantity as a function of price, as Cournot (1838, pp. 44–55) had done, but chose the converse approach to maintain symmetry with his treatment of supply. Believing in the importance of scale economies in production, he deemed it generally impossible to treat quantity supplied per unit of time as a single-valued function of market price. Instead, adopting what he took to be the businessman's perspective, he introduced the concept of ‘supply price’; the minimum uniform price at which any given quantity will be supplied to the market.

Market equilibrium occurs at any quantity whose demand price and supply price are equal, so that the market demand curve intersects the market supply curve – the latter the graph of supply price as a function of aggregate quantity supplied, a lateral sum of individual supply curves. Equilibrium is locally stable if the demand curve cuts the supply curve from above at the equilibrium quantity. This result is justified by the argument that the rate of supply will increase if the current market price (always determined by demand price) exceeds supply price at the current quantity, so that additional production offers excess profit, decreasing in the opposite case (Marshall, 1920, pp. 345–7). The resulting dynamic process is usually referred to as the Marshallian adjustment process.

It is probably due to Marshall's influence that English-speaking economists still graph demand and supply curves with quantity on the horizontal axis even though adopting a more Walrasian perspective which treats quantities demanded and supplied as functions of market price.

Marshall's conception of the demand price of a lone commodity, segregated from other commodities by an assumed constancy of the marginal utility of money, does not feature prominently in modern theoretical work. Instead, a multi-commodity formulation of utility and demand is typically adopted. Consider a consumer maximizing the utility function $u(x_1, x_2, \ldots, x_n)$ subject to the budget constraint $\sum p_i x_i = M$. (Here the x_i are quantities and the p_i prices of the n commodities and M is a preset total expenditure level. The utility function, u, is assumed strictly increasing, strictly quasi-concave, and differentiable.) This maximization implies the consumer's direct demand functions $x_i=d_i(p_1/M, p_2/M, \ldots, p_n/M)$, i=1, 2, …, n, sometimes (but with dubious justification) referred to as Marshallian demand functions to distinguish them from Hicksian compensated or constant-utility demand functions. These

demand functions can usually be inverted to yield the indirect or inverse demand functions $p_i/M=g_i(x_1, x_2, \ldots, x_n)$, $i=1, 2, \ldots, n$. However, these can be obtained more immediately from the budget constraint and the first-order conditions $\partial u / \partial x_i = \lambda p_i, i = 1, 2, \ldots, n$ (where λ is the Lagrange multiplier associated with the budget constraint). We have, for $i=1, 2, \ldots, n$,

$$\frac{p_i}{M} = \frac{\partial u / \partial x_i}{\lambda M} = \frac{\partial u / \partial x_i}{\Sigma \lambda p_j x_j} = \frac{\partial u / \partial x_i}{\Sigma x_j (\partial u / \partial x_j)} \equiv g_i(x_1, x_2 \ldots, x_n) \quad (1)$$

(The g_i are clearly unaffected by a monotone increasing transformation of u and reduce to $(\partial u / \partial x_i) / u$ if u is homogeneous of degree one.) The indirect demand functions (1) are the natural generalization of Marshall's demand-price concept at the individual level, defining an n-vector of normalized prices at which a given n-vector of commodities will be demanded.

Indirect demand functions may be useful in the contexts of central planning or rationing, where they can indicate the prices planners should choose to clear markets given the quantities available, or the notional prices at which ration allotments would just be freely purchases (see Pearce, 1964, pp. 57–64). But unfortunately, although indirect demand functions are readily obtained for the individual, they are not as easily aggregated to the market level as are direct demand functions. The asymmetry arises from the fact that individuals face identical prices but do not make identical quantity choices. Thus, market-level indirect demand functions must generally be obtained by first aggregating the individual direct demand functions and then inverting the resulting market functions.

The modern duality approach to consumer behaviour has revealed fundamental symmetries in the roles of prices and quantities. The alternatives of viewing quantity demanded as a function of price or demand price as a function of quantity can now be seen as only one of a variety of dual alternatives which considerably enrich theoretical and econometric analysis. (See Gorman, 1976, for a simple treatment.)

See Also

- Marshall, Alfred

Bibliography

Cournot, A.A. 1838. *Mathematical Principles of the Theory of Wealth*. New York: Macmillan, 1897.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in Economic Analysis*, ed. J.J. Artis and A. R. Nobay, London: Cambridge University Press.

Marshall, A. 1890. *Principles of Economics*, Vol. I. London: Macmillan.

Marshall, A. 1920. *Principles of Economics: An Introductory Volume*. London: Macmillan. Eighth edition of Marshall (1890).

Pearce, I.F. 1964. *A Contribution to Demand Analysis*. Oxford: Clarendon Press.

How to cite this article

Whitaker, J. K. "demand price." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000068> doi:10.1057/9780230226203.0370

The New Palgrave Dictionary of Economics Online

demand theory

Volker Böhm and Hans Haller
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Demand theory describes and explains individual choice of consumption bundles. Traditional theory considers optimizing behaviour when the consumer's choice is restricted to consumption bundles that satisfy a budget constraint. The budget constraint is determined by price–income pairs. A demand correspondence assigns to each price–income pair a non-empty set of optimal consumption bundles. A demand function assigns to each price–income pair a unique optimal consumption bundle. Optimality of consumption bundles is based on a preference relation. The theory derives existence and properties of demand correspondences (demand functions) from assumptions on preference relations and, if applicable, their utility representations.

Keywords

budget sets; cardinal utility; completeness; consumption plans; consumption sets; contingent commodities; continuity; continuous preference orders; convexity; demand correspondences; demand functions; demand sets; demand theory; expenditure functions; Giffen goods; Hicksian (income-compensated) demand function; inferior goods; integrability of demand; inverse demand function; Lebesgue measure approach; normal goods; ordinal utility; preference maximization; preference orders; quasi-concavity; reflexivity; representability of preferences; revealed preference theory; separability; Slutsky matrix; Slutsky, E.; transitivity; utility maximization; Walras, L.

Article

The main purpose of demand theory is to describe and explain observed consumer choices of commodity bundles. Market parameters, typically prices and income, determine constraints on commodity bundles. Given a combination of market parameters, a commodity bundle or a non-empty set of commodity bundles, which satisfies the corresponding constraints, is called a demand vector or a demand set. The mapping which assigns to every admissible combination of market parameters a unique demand vector (or a non-empty demand set) is called a demand function (or a demand correspondence, respectively). Traditional demand theory considers the demand function (or correspondence) as the outcome of some optimizing behaviour of the consumer. Its primary goal is to determine how alternative assumptions on the constraints, objectives and behavioural rules of the consumer affect his observed demands for commodities. The traditional model of the consumer postulates preferences over alternative commodity bundles to describe the objectives of the consumer. Its behavioural rule consists in maximizing these preferences on the set of feasible commodity bundles which satisfy the budget constraint imposed by the market parameters. If there is a unique preference maximizer under each budget constraint, then preference maximization determines a demand function. If there is at least one preference maximizer under each budget constraint, then preference maximization determines a demand correspondence.

Once the traditional view is adopted, the occurrence of demand correspondences cannot be avoided. Compatibility of observed demand, which is always unique, with some demand correspondence poses a minor problem in general. However, the correspondence should be obtained through preference maximization. The last requirement leads to the main issues of modern demand theory: Which demand correspondences are compatible with preference maximization? Given any conditions necessary for demand correspondences to be compatible with preference maximization, are they sufficient? Which demand correspondences are compatible with a special class of preferences? What type of preferences yields a particular class of demand correspondences? When addressing these issues, modern demand theory attempts to link two concepts: preferences and demand.

Historically, the important concept was utility rather than preference. Before Fisher (1892) and Pareto (1896), utility was conceived as cardinal: that is, it was assumed to be a

measurable scale for the degree of satisfaction of the consumer. Fisher and Pareto were the first to observe that an arbitrary increasing transformation of the utility function has no effect on demand. Edgeworth (1881) had already written utility as a general function of quantities of all commodities and had employed indifference curves. It is now widely accepted in demand theory that only ordinal utility matters. That is, a utility function serves merely as a convenient device to represent a preference relation, and any increasing transformation of the utility function will serve this purpose as well.

Representability by utility functions imposes some restrictions on preferences. The problem of representability of a preference relation by a numerical function was solved by Debreu (1954; 1959; 1964) based on work by Eilenberg (1941), and by Rader (1963) and Bowen (1968). While still assuming cardinal utility, Walras (1874) developed the first 'theory of demand'. His demand was a function of all prices and the endowment bundle, obtained through utility maximization. Slutsky (1915) finally assumed an ordinal utility function with enough restrictions to yield a maximum under any budget constraint and testable properties of the resulting demand functions. In particular, he obtained negativity of diagonal elements and symmetry of the 'Slutsky matrix'.

Antonelli (1886) was the first to go the opposite way: construct indifference curves and a utility function from the so-called inverse demand function. Pareto (1906) took the same route. Katzner (1970) reports on recent results in this direction. The construction of preference relations from demand functions was achieved in two ways:

- 1. Samuelson (1947) and Houthakker (1950) introduced the concept of revealed preference into demand theory. Considerable progress in relating utility and demand in terms of revealed preference was achieved by Uzawa (1960), further refinements being due to Richter (1966).
- 2. Hurwicz and Uzawa (1971) contributed to the following so-called integrability problem: construct a twice continuously differentiable utility representation from a continuously differentiable demand function which satisfies certain integrability conditions (including symmetry and negative semi-definiteness of the Slutsky matrix).

Kihlstrom, Mas-Colell and Sonnenschein (1976) unified the two approaches by relating the axioms of revealed preference to properties of the Slutsky matrix.

Since there exists a sizable literature on demand theory, many of the concepts and results are well established and well-known. These have become so much part of standard knowledge in economic theory that they are included in any contemporary microeconomic textbook and other surveys. It would substantially reduce the space available for a presentation of the new results of recent decades if an extended introductory account of demand theory were to be included here as well.

Commodities and prices

Consumers purchase or sell commodities, which can be divided into goods and services. Each commodity is specified by its physical quality, its location, and the date of its availability. In the case of uncertainty, the state of nature in which the commodity is available may be added to the specification of a commodity. This leads to the notion of a contingent commodity (see Arrow, 1953; Debreu, 1959). We assume as in traditional theory that there exists a finite number l of such commodities. Quantities of each commodity are measured in real numbers. A *commodity bundle* is an l-dimensional vector $x=(x_1,\ldots,x_l)$. The set of all l-dimensional vectors $x=(x_1,\ldots,x_l)$ is the l-dimensional Euclidean space $\mathbb{R}^l$ which we interpret as the *commodity space*. $|x_h|$ indicates the quantity of commodity $h=1,\ldots,l$. Commodities are assumed to be perfectly divisible, so that their quantity may be expressed as any (non-negative) real number. The standard sign convention for consumers assigns positive numbers for commodities made available to the consumer (inputs) and negative numbers for commodities made available by the consumer (outputs). Hence, a priori any commodity bundle $x \in \mathbb{R}^l$ is conceivable.

The price p_h of a commodity h, $h=1,\ldots,l$, is a real number which is the amount in units of account that has to be paid in exchange for one unit of the commodity. For the consumer, p_h is given and has to be paid now for the delivery of commodity h under the circumstances (location, date, state) specified for commodity h. A *price system* or *price vector* is a vector $p=(p_1,\ldots,p_l)$ in $\mathbb{R}^l$ and contains the prices for all commodities. The value of a commodity bundle x given the price vector p is $px = \sum_{h=1}^{l} p_h x_h$. This means that commodity bundles are *priced linearly*.

Consumption sets and budget sets

Typically, some commodity bundles cannot be consumed by a consumer for physical reasons. Those consumption bundles which can be consumed form the consumer's *consumption set*. This is a non-empty subset X of the commodity space $\mathbb{R}^l$. A consumer must choose a bundle x from his consumption set X in order to subsist. Traditionally, inputs in consumption are described by positive quantities and outputs by negative quantities. So in particular, the labour components of a consumption bundle x are all non-positive, unless labour is hired for a service. One usually assumes that the consumption set X is closed, convex, and bounded below. Vectors $x \in X$ are sometimes called *consumption plans*.

Given the sign convention on inputs and outputs and a price vector p, the value px of a consumption plan x defines the net outlay of x, that is the value of all purchases (inputs) minus

the value of all sales (outputs) for the bundle x. Trading the bundle x in a market at prices p implies payments and receipts for that bundle. Therefore, the value of the consumption plan should not exceed the initial wealth (or income) of the consumer which is a given real number w. If the consumer owns a vector of initial resources ω and the price vector p is given, then w may be determined by $w=p\omega$. The consumer may have other sources of wealth: savings and pensions, bequests, profit shares, taxes, or other liabilities. Given p and w, the set of possible consumption bundles whose value does not exceed the initial wealth of the consumer is called the *budget set* and is defined formally by

$$\beta(p, w) = \{x \in X | px \leq w\}.$$

The ultimate decision of a consumer is to choose a consumption plan from his budget set. Those vectors in $\beta(p,w)$ which the consumer eventually chooses form his *demand set* $\varphi(p, w)$.

Preferences and demand

The choice of the consumer depends on his tastes and desires. These are represented by his *preference relation* $\succsim$ which is a binary relation on X. For any two bundles $x, y \in X$, $x \succsim y$ means that x is at least as good as y. If the consumer always chooses a most preferred bundle in his budget set, then his demand set is defined by

$$\phi(p, w) = \left\{x \in \beta(p, w) | x' \in \beta(p, w) \text{ implies } x \succsim x' \text{ or not } x' \succsim x\right\}.$$

Three basic axioms are usually imposed on the preference relation $\succsim$ which are taken as a definition of a rational consumer:

- *Axiom 1* (reflexivity). If $x \in X$, then $x \succsim x$, that is, any bundle is as good as itself.
- *Axiom 2* (transitivity). If $x, y, z \in X$ such that $x \succsim y$ and $y \succsim z$, then $x \succsim z$.
- *Axiom 3* (completeness). If $x, y \in X$, then $x \succsim y$ or $y \succsim x$.

A preference relation $\succsim$ which satisfies these three axioms is a complete preordering or weak order on X and will be called a *preference order*. Already Axioms 2 and 3 define a preference order, since Axiom 3 implies Axiom 1. A preference relation $\succsim$ on X induces two other relations on X, the relation of strict preference, $\succ$, and the relation of indifference, ~.

Definition: Let $\succsim$ be a preference relation on the consumption set X. A bundle x is said to be *strictly preferred to* a bundle y, that is $x \succ y$, if and only if $x \succsim y$ and not $y \succsim x$. A bundle x is said to be *indifferent to* a bundle y, that is $x \sim y$, if and only if $x \succsim y$ and $y \succsim x$.

Lemma: Suppose $\succsim$ is reflexive and transitive. Then

1. (i) *$\succ$ is irreflexive, that is, not $x \succ x$, and transitive*;
2. (ii) *~ is an equivalence relation on X, which means that ~ is reflexive, transitive, and symmetric: that is, x~y if and only if y~x.*

For $Z \subseteq X$, $x \in Z$, x is called *maximal in Z*, if for all $z \in Z$: not $z \succ x$. x is called a *best element of Z* or *most preferred in Z*, if for all $z \in Z$: $x \succsim z$. Best elements are maximal; maximal elements are not necessarily best elements. If $\succsim$ is complete, then best and maximal elements coincide. Obviously for any price vector p and initial wealth w,

$$\phi(p, w) = \{x \in \beta(p, w) | x \text{ is maximal in } \beta(p, w)\}.$$

Axioms 1–3 are not qsted in most of consumer theory. However, transitivity and completeness may be violated by observed behaviour. Recent developments in the theory of consumer demand indicate that some weaker axioms suffice to describe and derive consistent demand behaviour (see, for example, Sonnenschein, 1971; Katzner, 1971; Shafer, 1974; Kihlstrom, Mas-Colell and Sonnenschein, 1976; Kim and Richter, 1986). In an alternative approach, one could start from a strict preference relation as the primitive concept. This may sometimes be convenient. However, the weak relation $\succsim$ seems to be the more natural concept. If the consumer chooses *x*, although *y* was a possible choice as well, then his choice can only be interpreted in the sense of $x \succsim y$, but not as $x \succ y$.

For the remainder of this section, let us fix a preference order $\succsim$ on *X* and a non-empty subset *B* of $\mathbb{R}^{l+1}$ such that for every $(p,w) \in B$, there is a unique $\succsim$-best element in $\beta(p,w)$: that is, maximization of $\succsim$ defines a *demand function* $f: B \to X$ such that $\varphi(p,w) = \{f(p,w)\}$ for all $(p,w) \in B$.

Let $x, x' \in X$, $x \neq x'$. We call *x revealed preferred to x′* and write xRx', if there is $(p,w) \in B$ such that $x = f(p,w)$ and $px' \leq px$. xRx' implies that both *x* and *x′* belong to the budget set $\beta(p,w)$ and *x* is chosen. Since *f* is derived from $\succsim$-maximization, xRx' implies $x \succ x'$. We call *x indirectly revealed preferred to x′* and write xR^*x', if there exists a finite sequence $x_0 = x, x_1, \ldots, x_n = x'$ in *X* such that $x_0 R x_1, \ldots, x_{n-1} R x'$. Obviously, R^* is transitive. Since $\succ$ is transitive, xR^*x' implies $x \succ x'$. Consequently, the following must hold (otherwise $x \succ x$!):

$$\text{(SARP)} \quad xR^*x' \Rightarrow \text{not } (x'R^*x) .$$

(SARP) implies

$$\text{(WARP)} \quad xRx' \Rightarrow \text{not } (x'Rx) .$$

(SARP) is the *strong axiom of revealed preference*; (WARP) is the *weak axiom*. Hence $\succsim$-maximization implies the strong axiom and a fortiori the weak axiom. For the inverse implication, see Chipman et al. (1971, chs. 1, 2, 3 and 5). For $l \geq 3$, there exist demand functions which satisfy (WARP) but not (SARP), whereas for $l = 2$, (WARP) and (SARP) are equivalent; see Section 3.J of Mas-Colell, Whinston and Green (1995) and Kihlstrom, Mas-Colell and Sonnenschein (1976, p. 977).

Continuous preference orders and utility functions

Axioms 1–3 have intuitive appeal. This is less so with the topological requirements of the following Axiom 4.

- *Axiom 4* (continuity). For every $x \in X$, the sets $\{y \in X \mid y \succsim x\}$ and $\{y \in X \mid x \succsim y\}$ are closed relative to *X*.

If $\succsim$ is a preference order, then Axiom 4 is equivalent to: For every $x \in X$, the sets $\{y \in X \mid y \succ x\}$ and $\{y \in X \mid x \succ y\}$ are open in *X*.

Closedness of $\{y \in X \mid y \succsim x\}$ requires that for any sequence y^n, $n \in \mathbb{N}$, in *X* such that y^n converges to $y \in X$ and $y^n \succsim x$ for all *n*, the limit *y* also satisfies $y \succsim x$. Openness of $\{y \in X \mid y \succ x\}$ means that if $y \succ x$, then $y' \succ x$ for any *y′* close enough to *y*.

The sets $\{y \in X \mid y \succsim x\}$ are called *upper contour sets* of the relation $\succsim$ and the sets $\{y \in X \mid x \succsim y\}$ are called *lower contour sets* of $\succsim$. For $x \in X$, the set $I(x) := \{y \in X \mid y \sim x\}$ is called the *indifference class* of *x* with respect to $\succsim$ or the $\succsim$*-indifference surface* through *x* or the $\succsim$*-indifference curve* through *x*. In the case $\succsim$ is reflexive and transitive, $I(x)$ is the equivalence class of *x* with respect to the equivalence relation ~.

There is a preference order $\succsim$ on $\mathbb{R}^l$, $l \geq 2$, which does not satisfy Axiom 4, namely the *lexicographic order* defined by $(x_1, \ldots, x_l) \succsim (y_1, \ldots, y_l)$ if and only if $x = y$ or there exists $k \in \{1,\ldots,l\}$ such that: $x_j = y_j$ for $j < k$ and $x_k > y_k$. Few studies of the relationship between the order properties of Axioms 1–3 and the topological property of Axiom 4 have been made. We emphasize the following result.

Theorem: (Schmeidler, 1971). *Let* $\succsim$ *denote a transitive binary relation on a connected topological space X. Assume that there exists at least one pair* $\bar{x}, \bar{y} \in X$ *such that* $\bar{x} \succ \bar{y}$. *If for every* $x \in X$, *(i)* $\{y \in X \mid y \succsim x\}$ *and* $\{y \in X \mid x \succsim y\}$ *are closed and (ii)* $\{y \in X \mid y \succ x\}$ *and* $\{y \in X \mid x \succ y\}$ *are open, then* $\succsim$ *is complete.*

Definition: Let *X* be a set and $\succsim$ be a preference relation on *X*. Then a function *u* from *X* into the real line $\mathbb{R}$ is a *(utility) representation* or a *utility function for* $\succsim$, if for all

$x, y \in X: u(x) \geq u(y)$ if and only if $x \succsim y$. Clearly, if *u* is a utility representation for $\succsim$ and $f: \mathbb{R} \to \mathbb{R}$ is an increasing transformation, then the composition $f \circ u$ is also a representation of $\succsim$. If $u: X \to \mathbb{R}$ is any function, then $\succsim$, defined by $x \succsim y$ if and only if $u(x) \geq u(y)$ for $x, y \in X$, is a preference order on *X* and *u* is a utility representation for $\succsim$. Most utility functions used in consumer theory are continuous. If *u* is continuous and $\succsim$ is represented by *u*, then by necessity $\succsim$ is a continuous preference order. In our case where $X \subseteq \mathbb{R}^l$, the opposite implication also holds: If $\succsim$ is a continuous preference order, then it has a continuous utility representation.

Theorem: (Debreu, Eilenberg, Rader). *Let X be a topological space with a countable base of open sets (or a connected, separable topological space) and* $\succsim$ *be a continuous preference order on X. Then* $\succsim$ *has a continuous utility representation*.

In our context of Euclidean commodity spaces, explicit constructions of continuous utility representations for continuous and monotonic preference orders are available. See Arrow and Hahn (1971) for the 'Euclidean distance approach' and Neuefeind (1972) for the 'Lebesgue measure approach'. For topological spaces *X* with a countable base of open sets, it has further been shown by Rader (1963) and Bosi and Mehta (2002) that an upper semi-continuous preference order on *X* has an upper semi-continuous utility representation.

As an immediate consequence of the representation theorem for preference relations, one obtains one of the standard results on the non-emptiness of the demand set $\varphi(p,w)$, since any continuous function attains its maximum on a compact set (Weierstrass's th), though a direct proof is also possible.

Corollary: Let $X \subseteq \mathbb{R}^l$ *be bounded below and closed,* $\succsim$ *be a continuous preference order on X,* $p \in \mathbb{R}^l_{++}$ *(that is,* $p \gg 0$*) and* $w \in \mathbb{R}$*. Then* $\beta(p, w) \neq \varnothing$ *implies* $\phi(p, w) \neq \varnothing$.

There has been a recent shift from proving existence to a more systematic study of the non-existence of utility representations. Needless to say that there are many preference orders on $\mathbb{R}^l$ or on subsets thereof with continuous utility representations. There are also total orders $\succsim$ (that is, preference orders $\succsim$ with $x \sim y \Leftrightarrow x = y$) on $\mathbb{R}^l$, $l \geq 2$, which admit utility representations, since there exist bijections $u: \mathbb{R}^l \to \mathbb{R}$. However, for $l \geq 2$, there is no total order on $\mathbb{R}^l$, $\mathbb{R}^l_+$ or $[0,1]^l$ which has a continuous utility representation; see Candeal and Induráin (1993). Moreover, a preference order $\succsim$ on *X*, which is not continuous, need not have a utility representation. For instance, the lexicographic order on $\mathbb{R}^l$, $l \geq 2$, a total order first discussed by Debreu (1954), does not have a utility representation, nor even a discontinuous one. Beardon et al. (2002) provide a classification of total orders which do not admit a utility representation. Estévez Toranzo and Hervés Beloso (1995) show that, if *X* is a non-separable metric space, then there exists a continuous preference order on *X* which cannot be represented by a utility function.

Some properties of preferences and utility functions

Some of the frequent assumptions on preference relations correspond almost by definition to analogous properties of utility functions, while other analogies need demonstration. We discuss the assumptions most commonly used.

Monotonicity: A preference order $\succsim$ on $X \subseteq \mathbb{R}^l$ is *monotonic*, if $x, y \in X, x \geq y, x \neq y$ implies $x \succ y$.

This property means desirability of all commodities. If a monotonic preference order has a utility representation *u*, then *u* is an increasing function (in all arguments). Inversely, if $\succsim$ is represented by an increasing function, then $\succsim$ is monotonic.

Non-satiation: Let $\succsim$ be the preference relation of a consumer over consumption bundles in *X* and let $x \in X$.

1. (i) *x* is a *satiation point for* $\succsim$ if $x \succsim y$ for all $y \in X$: that is, *x* is a best element in *X*.
2. (ii) The preference relation is *locally not satiated at x*, if for every neighbourhood *U* of *x* there exists $z \in U$ such that $z \succ x$.

Consider a utility representation *u* for $\succsim$. Then $x \in X$ is a satiation point if and only if *u* has a global maximum at *x*. $\succsim$ is locally not satiated at *x* if and only if *u* does not attain a local maximum at *x*. Local non-satiation rules out that *u* is constant in a neighbourhood of *x*. If $\succsim$ is locally not satiated at all *x*, then $\succsim$ cannot have thick indifference classes or satiation points.

Convexity: A preference relation $\succsim$ on $X \subseteq \mathbb{R}^l$ is called

1. (i) *convex*, if the set $\{y \in X | y \succsim x\}$ is convex for all $x \in X$;
2. (ii) *strictly convex*, if *X* is convex and $\lambda x + (1-\lambda)x' \succ x'$ for any two bundles $x, x' \in X$ such that $x \neq x', x \succsim x'$ and for any λ such that $0 < \lambda < 1$;
3. (iii) *strongly convex*, if *X* is convex and $\lambda x + (1-\lambda)x' \succ x''$ for any three bundles $x, x', x'' \in X$ such that $x \neq x', x \succsim x'', x' \succsim x''$ and for any λ such that $0 < \lambda < 1$.

Quasi-concavity: A function $u: X \to \mathbb{R}$ is called

1. (i) *quasi-concave*, if $u(\lambda x + (1-\lambda)y) \geq \min\{u(x), u(y)\}$ for all $x, y \in X$ and any $\lambda \in [0,1]$;
2. (ii) *strictly quasi-concave*, if $u(\lambda x + (1-\lambda)y) > \min\{u(x), u(y)\}$ for all $x, y \in X$ with $x \neq y$ and any $\lambda \in (0,1)$.

Let u be a representation of the preference order $\succsim$. Then u is (strictly) quasi-concave if and only if $\succsim$ is (strictly) convex. Quasi-concavity is preserved under increasing transformations: that is, it is an ordinal property. In contrast, concavity is a cardinal property which can be lost under increasing transformations. With respect to the difficult problem to characterize those preference orders which have a concave representation, we refer to Kannai (1977). Clearly, if $\succsim$ is locally not satiated at all x, then $\succsim$ does not have a satiation point. In general, the inverse implication is false. If, however, $\succsim$ is strictly convex and does not have a satiation point, then $\succsim$ is locally not satiated at all x. Moreover, if $\succsim$ is strictly convex, then it has at most one satiation point. An immediate implication is the following lemma.

Lemma: Let $X \subseteq \mathbb{R}^l$ be bounded below, convex, and closed. Let $\succsim$ be a strictly convex, continuous preference order on X, $p \in \mathbb{R}^l_{++}$, and $w \in \mathbb{R}$. Then $\beta(p, w) \neq \varnothing$ implies that φ(p, w) is a singleton.

Separability Separable utility functions were used in classical consumer theory long before associated properties of preferences had been defined. All early contributions to utility theory assumed without much discussion an additive form of the utility function over different commodities. It was not until Edgeworth (1881) that utility was written as a general function of a vector of commodities. The particular consequences of separability for demand theory were discussed well after the general non-separable case in demand theory had been treated and generally accepted. Among the many contributors are Sono (1945), Leontief (1947), Samuelson (1947), Houthakker (1960), Debreu (1960), and Koopmans (1972). We follow Katzner (1970) in our presentation.

Let $N = \{N_j\}_{j=1}^k$ be a partition of the set $\{1,\ldots,l\}$ and assume that $X = S_1 \times \cdots \times S_k$. Let $J=\{1,\ldots,k\}$ and for any $j \in J$, $y \in X$, $y = (y_1, \ldots, y_k) \in \Pi_{i \in I} S_i$ write $y_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_k)$ for the vector of components different from j. For any y_{-j}, a preference order $\succsim$ on X induces a preference order $\succsim_{y_{-j}}$ on S_j which is defined by $x_j \succsim_{y_{-j}} x_j'$ if and only if $(y_{-j}, x_j) \succsim (y_{-j}, x_j')$ for $x_j, x_j' \in S_j$. In general, the induced ordering $\succsim_{y_{-j}}$ will depend on y_{-j}. The first notion of separability states that for any j, the preference orders $\succsim_{y_{-j}}$ are independent of $y_{-j} \in \Pi_{i \neq j} S_i$. The second notion of separability states that for any proper subset I of J, the induced preference orders $\succsim_{y_{I \setminus I}}$ on $\Pi_{i \in I}$ are independent of $y_{I \setminus I} \in \Pi_{i \notin I} S_i$.

Definition: Let $\succsim$ be a preference order on $X = \Pi_{j \in J} S_j$.

1. (i) $\succsim$ is called *weakly separable* with respect to N if $\succsim_{y_{-j}} = \succsim_{z_{-j}}$ for each $j \in J$ and any $y_{-j}, z_{-j} \in \Pi_{i \neq j} S_i$.
2. (ii) $\succsim$ is called *strongly separable* with respect to N if $\succsim_{y_{I \setminus I}} = \succsim_{z_{I \setminus I}}$ for each $I \subseteq J$, $I \neq \varnothing$, $I \neq J$ and any $y_{I \setminus I}, z_{I \setminus I} \in \Pi_{i \notin I} S_i$.

Definition: Let $u: \Pi_{j \in J} S_j \to \mathbb{R}$. u is called

1. (i) *weakly separable* with respect to N, if there exist continuous functions $v_j: S_j \to \mathbb{R}$, $j \in J$, and $V: \mathbb{R}^k \to \mathbb{R}$ such that $u(x) = V(v_1(x_1), \ldots, v_k(x_k))$;
2. (ii) *strongly separable* with respect to N, if there exist continuous functions $v_j: S_j \to \mathbb{R}$, $j \in J$, and $V: \mathbb{R} \to \mathbb{R}$ such that $u(x) = V(\Sigma_{j \in J} v_j(x_j))$.

The two important equivalence results on separability are due to Debreu and Katzner. The version of Debreu's theorem given here is slightly weaker than his original result.

Theorem: (Katzner, 1970). *Let $\succsim$ be a continuous, monotonic preference order on $X = \Pi_{j \in J} S_j$ with $S_j = \mathbb{R}^{N_j}$ for all $j \in J$. Then $\succsim$ is weakly separable if and only if every continuous representation of $\succsim$ it is weakly separable.*

Theorem: (Debreu, 1960). *Let $\succsim$ be a continuous, monotonic preference order on $X = \Pi_{j \in J} S_j$ with $S_j = \mathbb{R}^{N_j}$ for all $j \in J = \{1, \ldots, k\}$ and $k \geq 3$. Then $\succsim$ is strongly separable if and only if every continuous representation is strongly separable.*

Under the assumptions of this theorem, if $\succsim$ is strongly separable with representation $u(x) = V(\Sigma_{j \in J} v_j(x_j))$, then V must be increasing or decreasing. Therefore,

$$v(x) = \begin{cases} \sum_{j \in J} v_j(x_j) & \text{for } V \text{ increasing} \\ -\sum_{j \in J} v_j(x_j) & \text{for } V \text{ decreasing} \end{cases}$$

is also a representation of $\succsim$. This is the additive form of separable utility used by early economists who thought that each commodity *h* had its own intrinsic utility representable by a scalar function u_h. The overall utility was then simply obtained as the sum of these functions, $u(x) = \sum_h u_h(x_h)$. Such a formulation is given by Jevons (1871) and Walras (1874) and implicitly contained in Gossen (1854).

In the case of uncertainty, with finitely many states of nature $j \in J = \{1, \ldots, k\}$, respective probabilities $\pi_j > 0$ and consumption $x_j \in S_j$ in state $j \in J$, an additively separable utility representation $u(x) = \sum_{j \in J} v_j(x_j)$ is tantamount to an expected utility representation $u(x) = \sum_{j \in J} \pi_j u_j(x_j)$ with $u_j = v_j / \pi_j$. Hence, an expected utility representation in the tradition of Savage (1954) implies separability with respect to states of nature. In contrast, the novel concept of Choquet expected utility à la Schmeidler (1986; 1989) typically violates separability with respect to states of nature.

For *k*=2, weak and strong separability of preferences coincide. But there are separable preferences which do not admit a strongly separable utility representation, for instance $X = \mathbb{R}^2_+$, $N_j = \{j\}$ for *j*=1, 2, $\succsim$ given by $u(x_1, x_2) = \sqrt{x_1} + \sqrt{x_1 + x_2}$. Separability of preferences imposes restrictions on demand correspondences and on demand functions (for details see Barten and Böhm, 1982, Sections 9, 14, and 15).

Continuous demand

Given any price–wealth pair $(p, w) \in \mathbb{R}^{l+1}$, the budget set of the consumer was defined as $\beta(p, w) = \{x \in X \mid px \leq w\}$. Let $S \subseteq \mathbb{R}^{l+1}$ denote the set of price–wealth pairs for which the budget set is non-empty. Then β describes a correspondence from *S* into *X*: that is, β associates to any $(p,w) \in S$ the non-empty subset $\beta(p,w)$ of *X*. There are two standard notions of continuity of correspondences, upper hemi-continuity and lower hemi-continuity (see Hildenbrand, 1974).

Definition: A compact-valued correspondence Ψ from *S* into an arbitrary subset *T* of $\mathbb{R}^l$ is *upper hemi-continuous* (u.h.c.) at a point $y \in S$, if for all sequences $(y^n, z^n) \in S \times T$ such that $y^n \to y$ and $z^n \in \Psi(y^n)$ for all *n*, there exist $z \in \Psi(y)$ and a subsequence z^{n_k} of z^n such that $z^{n_k} \to z$.

Definition: A correspondence Ψ from *S* into an arbitrary subset *T* of $\mathbb{R}^l$ is *lower hemi-continuous* (l.h.c.) at a point $y \in S$, if for any $z \in \Psi(y)$ and any sequence y^n in *S* with $y^n \to y$ there exists a sequence z^n in *T* such that $z^n \to z$ and $z^n \in \Psi(y^n)$ for all *n*.

Definition: A correspondence is *continuous* if it is both lower and upper hemi-continuous.

For single-valued correspondences, the notions of lower and upper hemi-continuity coincide with the usual notion of continuity for functions. For proofs of the following lemmas, see Debreu (1959) or Hildenbrand (1974).

Lemma: Let $X \subseteq \mathbb{R}^l$ be a convex set. Then the budget correspondence $\beta : S \to X$ has a closed graph and is lower hemi-continuous at every point $(p,w) \in S$ for which $w > \min\{px \mid x \in X\}$ holds.

Combining a previous corollary on the non-emptiness of the demand set and a fundamental theorem of Berge (1966) yields the next result.

Lemma: Let $X \subseteq \mathbb{R}^l$ be a convex set. If the preference relation has a continuous utility representation, then the demand correspondence is defined (that is, non-empty valued), compact-valued, and upper hemi-continuous at each $(p,w) \in S$ such that $\beta(p,w)$ is compact and $w > \min\{px \mid x \in X\}$.

It follows immediately from the definitions that $\varphi(\lambda p, \lambda w) = \varphi(p,w)$ for any $\lambda > 0$ and any price–wealth pair (p,w): that is, demand is homogeneous of degree zero in prices and wealth. For convex preference orders, the demand correspondence is convex-valued. For strictly convex preference orders, the demand correspondence is single-valued: that is, one obtains a demand function. The results of this section and of the section on continuous preference orders and utility functions are summarized in the following lemma, which uses the weakest assumptions of traditional demand theory to generate a continuous demand function.

Lemma: Let $S' := \{(p, w) \in S \mid \beta(p, w)$ is compact and $w > \min\{px \mid x \in X\}\}$. If $\succsim$ denotes a strictly convex and continuous preference order, then $\varphi(p,w)$ defines a continuous demand function which satisfies: (i) homogeneity of degree zero in prices and wealth and (ii) the strong axiom of revealed preference.

Continuous demand without transitivity

Transitivity is often violated in empirical studies. This excludes utility maximization, but not necessarily preference maximization. However, as the next theorem indicates, existence and continuity of demand do not depend on transitivity as crucially as one may expect. The theorem follows from a result by Sonnenschein (1971).

Theorem: Let $S^* = \{(p, w) \in S | \phi(p, w) \neq \varnothing\}$. *Suppose that X is compact and* $\succsim$ *is complete and has a closed graph.*

1. (i) *If* $\{x' \in X | x' \succ x\}$ *is convex for all* x∈X, *then* $\phi(p, w) \neq \varnothing$ *whenever* $\beta(p, w) \neq \varnothing$ *(that is,* $S^*=S$*).*
2. (ii) *If* $S^*=S$ *and* $(p^0,w^0)\in S$ *such that* β *is continuous at* (p^0,w^0), *then* φ *is u.h.c. at* (p^0,w^0).

The assumption that X is compact is not necessary. For case (i) it suffices that all budget sets $\beta(p,w)$ under consideration be compact. For case (ii) it is sufficient that there exist a compact subset X^0 of X and a neighbourhood S^0 of (p^0,w^0) such that $\varphi(S^0)\subseteq X^0$.

To complete this section we state a lemma on the properties of a demand function obtained under preference maximization without transitivity. This contrasts with the lemma at the end of the previous section. Intransitivity essentially implies that the strong axiom of revealed preference need not hold. The lemma follows from the theorem by Sonnenschein and from the result by Shafer (1974).

Lemma: Let $X = \mathbb{R}^l_+$, $B = \mathbb{R}^{l+1}_{++}$. *Suppose continuity and strong convexity of* $\succsim$ *(in addition to completeness). Then preference maximization yields a continuous demand function f:* $B\rightarrow X$ *which satisfies (i) homogeneity of degree zero in prices and wealth and (ii) the weak axiom of revealed preference.*

The converse statement of the lemma does not hold. For $l = 2$, $X = \mathbb{R}^2_+$, $B = \mathbb{R}^3_{++}$, there is a C^1-function $f:B\rightarrow X$ which fulfils (i), (ii), and (iii) $pf(p,w)=w$ for all $(p,w)\in B$, but which cannot be obtained as the demand function for a continuous, complete and strictly convex preference relation (John, 1984; Kim and Richter, 1986). In addition, John (1995) has shown that continuity of f, (ii) and (iii) imply (i).

Smooth preferences and differentiable utility functions

Owing to the representation theorem of Debreu, Eilenberg and Rader, continuity of a utility function and continuity of the represented preference order are identical under the perspective of demand theory. When continuous differentiability of demand is required, continuity of the preference relation will not suffice in general. The first rigorous attempt to study 'differentiable preference orders' goes back to Antonelli (1886). We follow the more direct approach of Debreu (1972) to characterize 'smooth preference orders'. Smoothness of preferences is closely related to sufficient differentiability of utility representations and the solution of the integrability problem (see Debreu, 1972; also Debreu, 1976; Hurwicz, 1971; and the section below on integrability). For the purpose of this and subsequent sections, let $P = \mathbb{R}^l_{++}$ denote the (relative) interior of $\mathbb{R}^l_+$ and assume that $X=P$. Let $\succsim$ be a continuous and monotonic preference order on P which we may consider as a subset of $P\times P$: that is, $(x, y) \in \succsim \Leftrightarrow x \succsim y$ for $(x,y)\in P\times P$. Also, the associated indifference relation $\sim$ will be considered as a subset of $P\times P$. To describe a smooth preference order, differentiability assumptions will be made on the (graph of the) indifference relation in $P\times P$.

For $k \geq 1$, let C^k denote the class of functions which have continuous partial derivatives up to order k, and consider two open sets X and Y in an Euclidean space $\mathbb{R}^n$. A bijection h: $X\rightarrow Y$ is a C^k-*diffeomorphism* if both h and h^{-1} are of class C^k. $M \subseteq \mathbb{R}^n$ is a C^k-*hypersurface*, if for every $z\in M$, there exist an open neighbourhood U of z, an open subset V of $\mathbb{R}^n$, a hyperplane $H\subset \mathbb{R}^n$ and a C^k-diffeomorphism $h:U\rightarrow V$ such that $h(M \cap U) = V \cap H$. A C^k-hypersurface has locally the structure of a hyperplane up to a C^k-diffeomorphism.

Considering the indifference relation $\sim$ as a subset of $P\times P$, the set $\tilde{I} = \{(x, y) \in P \times P | x \sim y\} \subset \mathbb{R}^{2l}$ constitutes the 'indifference surface' of the preference relation. Then $\succsim$ is called a C^2-*preference order* (or *smooth preference order*), if $\tilde{I}$ is a C^2-hypersurface.

Theorem: (Debreu, 1972). *Let* $\succsim$ *be a continuous and monotonic preference order on P and* $\tilde{I}$ *be its indifference surface. Then* $\succsim$ *is a* C^2*-preference order if and only if it has a monotonic utility representation of class* C^2 *with no critical point.*

Properties of differentiable utility functions

Utility functions of class C^2 provide the truly classical approach to demand theory (see, for example, Slutsky, 1915; Hicks, 1939; Samuelson, 1947).

Let $\succsim$ be a monotonic, strictly convex C^2-preference order on P and $u: P \to \mathbb{R}$ be a C^2-utility representation of $\succsim$ with no critical point. Then u is continuous, increasing in all arguments, and strictly quasi-concave. Moreover, all second-order partial derivatives $u_{ij}(x) = (\partial^2 u / \partial x_i \partial x_j)(x),\ i,\ j = 1, \ldots, l,\ x \in P$, exist, all u_{ij} are continuous functions of x and $u_{ij}=u_{ji}$ for $i,j=1,\ldots,l$. Let $D^2u=(u_{ij})$ denote the Hessian matrix of u. Then D^2u is symmetric. The first-order derivatives $u_i(x) = (\partial u / \partial x_i)(x),\ i = 1, \ldots, l$, are continuous functions of x. Assume that $u_i(x)>0$ for $i=1, \ldots,l$, $x \in P$ and define

$$Du(x) = \begin{bmatrix} u_1(x) \\ \vdots \\ u_l(x) \end{bmatrix}$$

as the gradient of u at x. For any $m \times n$-matrix M, let M' denote the transpose of M.

Theorem: If $u: P \to \mathbb{R}$ is a strictly quasi-concave utility function of class C^2, then $z' D^2 u(x) z \leq 0$ for all $x \in P$ and $z \in \{z \in \mathbb{R}^l \mid z Du(x) = 0\}$.
(For a proof, see Barten and Böhm, 1982.)
It will be shown in the next section that the conclusion of this theorem does not guarantee the existence of a differentiable demand function. The following definition strengthens the property of strict quasi-concavity.
Definition: u is called *strongly quasi-concave* if

$$z' D^2 u(x) z < 0 \text{ for all } x \in P,\ z \neq 0 \text{ and } z \in \{z \in \mathbb{R}^l \mid z Du(x) = 0\}.$$

Consider the bordered Hessian matrix

$$H(x) = \begin{bmatrix} D^2 u(x) & Du(x) \\ [Du(x)]' & 0 \end{bmatrix}.$$

Then u is strongly quasi-concave whenever u is strictly quasi-concave and $H(x)$ is non-singular. (For a proof, see Barten and Böhm, 1982).
The properties of strict and strong quasi-concavity are invariant under increasing C^2-transformations. For other results and consequences of differentiable utility functions the reader may consult Barten and Böhm (1982) and the references listed there, or Debreu (1972), Mas-Colell (1974).

Differentiable demand

The earlier section on continuous demand without transitivity provides sufficient conditions on preferences for the existence of a continuous demand function which is homogeneous of degree zero in prices and wealth and satisfies the strong axiom of revealed preference. In this section, the implications of smooth preferences for differentiability of demand will be studied. Consider an assumption (D), consisting of the following three parts:

- (D1) $X=P$.
- (D2) $\succsim$ is a monotonic, strictly convex C^2-preference order on X and the closure relative to $\mathbb{R}^l_+ \times \mathbb{R}^l_+$ of its indifference surface $\ddot{I}$ is contained in $P\times P$.
- (D3) The price-wealth space is $B = \mathbb{R}^{l+1}_{++}$.

Given (D), there exists a demand function $f:B\rightarrow X$ with $p\cdot f(p,w)=w$ for all $(p,w)\in B$. Let u be an increasing strictly quasi-concave C^2-utility representation for $\succsim$. The following key result on the differentiability of demand was first given by Katzner (1968). For a detailed proof see Barten and Böhm (1982).

Theorem: Let $(\bar{p}, \bar{w}) \in B$ *and* $\bar{x} = f(\bar{p}, \bar{w})$. *Then the following assertions are equivalent:*

1. (i) *f is* C^1 *in a neighbourhood of* $(\bar{p}, \bar{w})$.
2. (ii) $\begin{bmatrix} D^2u(\bar{x}) & \bar{p}' \\ \bar{p} & 0 \end{bmatrix}$ *is non-singular*.
3. (iii) $H(\bar{x})$ *is non-singular*.

Once the demand function f is continuously differentiable, it is straightforward to derive all of the well-known comparative statics properties, for the proof of which we refer again to Barten and Böhm (1982). Let $f = (f^1, \ldots, f^l)$ be a demand function of class C^1 and define the respective partial derivatives

$$
\begin{aligned}
f_w &= (f^1_w, \ldots, f^l_w) = \left(\frac{\partial f^1}{\partial w}, \ldots, \frac{\partial f^l}{\partial w}\right), \\
f^i_j &= \frac{\partial f^i}{\partial p_j}, \qquad i, j = 1, \ldots, l; \\
s^i_j &= f^i_j + f^i_w f^j, \qquad i, j = 1, \ldots, l.
\end{aligned}
$$

From these we obtain the Jacobian matrix of f with respect to prices, $J = (f^i_j)$, and the so-called Slutsky matrix $S = (s^i_j)$.

Theorem:

1. (i) $pf_w=1$, $pJ=-f$,
2. (ii) $Sp'=0$,
3. (iii) *S is symmetric*,
4. (iv) $ySy'<0$, *if* $y\in\mathbb{R}^l$, $y\neq\alpha p$ *for all* $\alpha\in\mathbb{R}$,
5. (v) rank $S=l-1$.

Property (iv) implies that all diagonal elements of S are strictly negative: that is, $s^i_i = f^i_i + f^i_w f^i < 0$. If $f^i_w > 0$, commodity i is called a *normal good* which implies that $f^i_i < 0$: that is, demand is downward sloping in its own price. On the other hand, a negative income effect $f^i_w < 0$, that is, when commodity i is an *inferior good*, is a necessary, but not a sufficient condition for a positive own price effect $f^i_i > 0$, that is, for commodity i to be a *Giffen good*.

Duality approach to demand theory

With the notion of an expenditure function, an alternative approach to demand analysis is possible which was suggested by Samuelson (1947). For the further development and details, we refer to Diewert (1974; 1982).

As a matter of convenience and for ease of presentation, assumption (D) will be imposed on the preference relation $\succsim$. Let u denote a strictly quasi-concave increasing C^2-utility representation for $\succsim$ and let f: $B \to X$ be the demand function derived from preference maximization. Let us further assume that $u(X) = \mathbb{R}$. (This requirement can always be fulfilled by means of an increasing transformation.) Define the *indirect utility function* v: $B \to \mathbb{R}$ associated with u by $v(p,w)=u(f(p,w))$ for $(p,w) \in B$.

Given a price system $p \in \mathbb{R}^l_{++}$ and a utility level $c \in \mathbb{R}$, let $e(p, c) = \min\{p \cdot x | x \in X, u(x) \geq c\}$. Since u is strictly quasi-concave and increasing, there exists a unique minimizer $h(p, c)$ of this problem such that $e(p,c)=ph(p,c)$. $h: \mathbb{R}^l_{++} \times \mathbb{R} \to \mathbb{R}^l_{++}$ is called the *Hicksian (income-compensated) demand function* and $e: \mathbb{R}^l_{++} \times \mathbb{R} \to \mathbb{R}_{++}$ is called the *expenditure function* for u. Since assumption (D) holds, preference maximization and expenditure minimization imply the following properties and relationships:

1. (i) $c=v[p,e(p,c)]$ for all (p,c).
2. (ii) $w=e[p,v(p,w)]$ for all (p,w).
3. (iii) $v(p,\cdot)$ and $e(p,\cdot)$ are inverse functions for any p.
4. (iv) $h(p,c)=f[p, e(p, c)]$ for all (p,c).
5. (v) $f(p,w)=h[p,v(p,w)]$ for all (p,w).
6. (vi) e is strictly increasing and continuous in c.
7. (vii) e is non-decreasing, positive linear homogeneous, and concave in prices.
8. (viii) v is strictly increasing in w, and continuous.
9. (ix) v is non-increasing in prices and homogeneous of degree zero in income and prices.

Moreover, some interesting and important consequences of these properties can be obtained if the functions are sufficiently differentiable.
Theorem:

1. (i) *e is C^k if and only if v is C^k. (k=1, 2).*
2. (ii) *If e is C^1, then $\partial e/\partial p=h$.*
3. (iii) *If f is C^1, then: v is C^2.*
4. (iv) $f=-(\partial v/\partial p)/(\partial v/\partial w)$ (*Roy's identity*).
5. (v) *h is C^1 and e is C^2.*
6. (vi) *$\partial h/\partial p=S$ (Slutsky equation) with $\partial h/\partial p$ evaluated at $[p,v(p,w)]$ and S evaluated at (p,w).*

Integrability:
A review of the previous discussions and analytical results involving the concepts of

- $\succsim$ preference
- u utility
- h income-compensated demand function
- e expenditure function
- v indirect utility
- f (direct) demand function

makes apparent their relationships which can be characterized schematically by the following diagram:

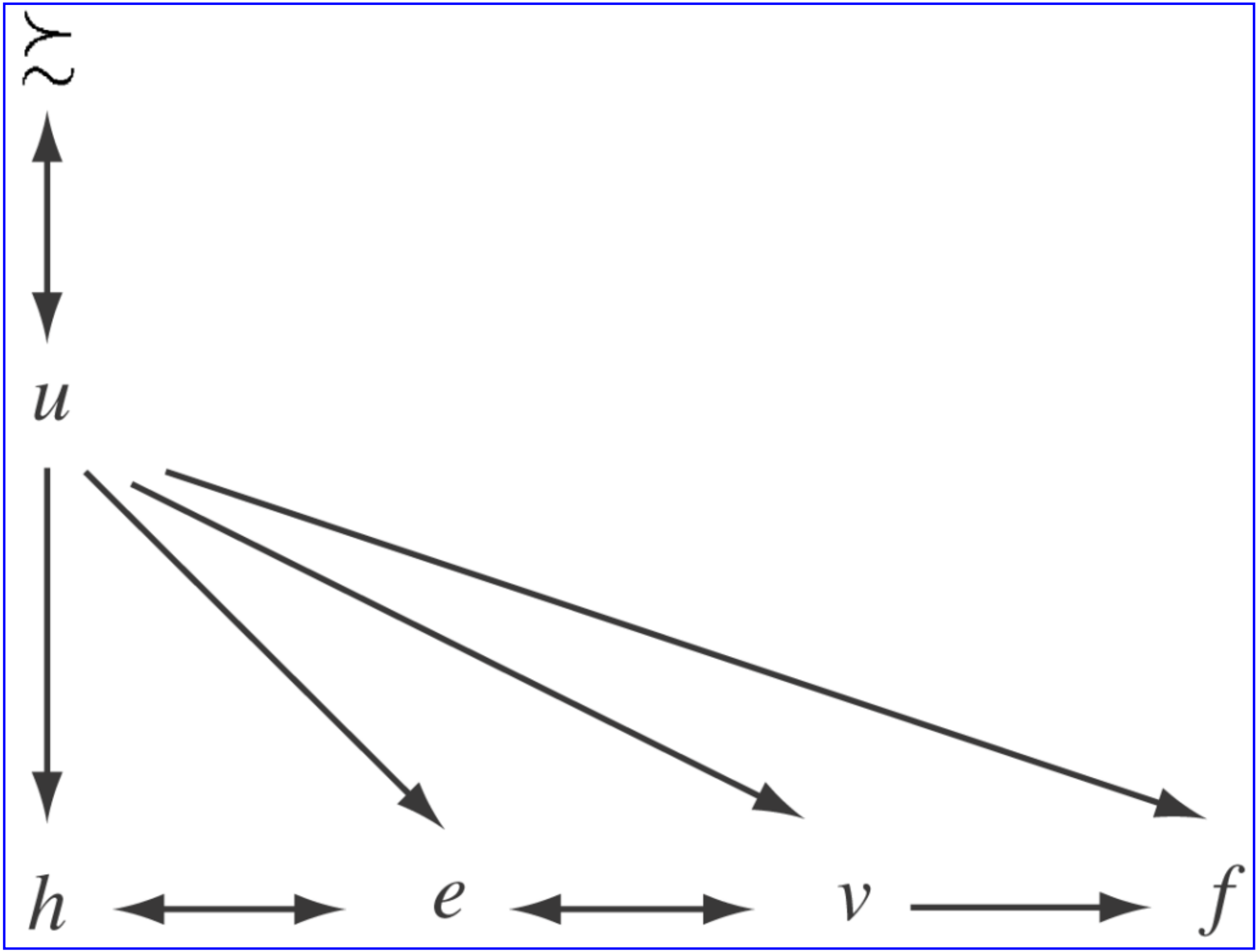

where $a \rightarrow b$ indicates that concept b can be derived from concept a under certain conditions. The integrability problem is to establish $f \rightarrow u$: that is, to recover the utility function from

the demand function *f*.

Two recent developments

Advanced microeconomic theory assumes a distribution of consumer characteristics to determine mean demand of a consumption sector. In accordance with traditional demand theory, the primitive characteristics of a consumer are his preference relation $\succsim$ and his wealth *w*, and possibly his consumption set *X*. If we disregard the latter, the corresponding distribution of consumer characteristics is a preference–wealth distribution (see Hildenbrand, 1974). This approach lends itself to both positive and normative analysis. In contrast, Hildenbrand (1994) and others adopt a purely positive point of view and take pairs (*f*,*w*) as the primitive concepts, where f is a demand function not necessarily derived from preference maximization of ‘rational’ consumers.
Like traditional demand theory, most of theoretical and empirical economics has not distinguished between households and individual consumers. Chiappori (1988; 1992) and others have developed models of collective rationality of multi-person households where each member has his or her own preferences.

See Also

- aggregation (theory)
- collective rationality
- correspondences
- Hicksian and Marshallian demands
- integrability of demand
- revealed preference theory
- separability

Bibliography

Antonelli, G.B. 1886. *Sulla Teoria Matematica della Economia Politica*. Pisa: Nella Tipografia del Folchetto. Trans. as ‘On the mathematical theory of political economy’, in Chipman et al. (1971).

Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie*, Colloques Internationaux du Centre National de la Recherche Scientifique, Paris, vol. 11, pp. 41–7.

Arrow, K.J. and Hahn, F. 1971. *General Competitive Analysis*. San Francisco: Holden-Day and Edinburgh: Oliver and Boyd.

Arrow, K.J. and Intriligator, M.D., eds. 1982. *Handbook of Mathematical Economics*, vol. 2. Amsterdam: North-Holland.

Arrow, K.J., Karlin, S. and Suppes, P. 1960. *Mathematical Methods in the Social Sciences*. Stanford: Stanford University Press.

Barten, A.P. and Böhm, V. 1982. Consumer theory. In Arrow and Intriligator (1982).

Beardon, A.F., Candeal, J.C., Herden, G., Induráin, E. and Mehta, G.B. 2002. The non-existence of a utility function and the structure of non-representable preference relations. *Journal of Mathematical Economics* 37, 17–38.

Berge, C. 1966. *Espaces Topologiques. Fonctions Multivoques*. Paris: Dunod. Trans. as ‘Topological Spaces’, Edinburgh: Oliver and Boyd, 1973.

Bosi, G. and Mehta, G.B. 2002. Existence of a semicontinuous or continuous utility function: a unified approach and an elementary proof. *Journal of Mathematical Economics* 38, 311–28.

Bowen, R. 1968. A new proof of a theorem in utility theory. *International Economic Review* 9, 374.

Candeal, J.C. and Induráin, E. 1993. Utility functions on chains. *Journal of Mathematical Economics* 22, 161–8.

Chipman, J.S., Hurwicz, L., Richter, M.K. and Sonnenschein, H.F. 1971. *Preferences, Utility, and Demand*. New York: Harcourt Brace Jovanovich.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56, 63–89.

Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100, 437–67.

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision Processes*, ed. R.M. Thrall et al. New York: Wiley.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Debreu, G. 1960. Topological methods in cardinal utility theory. In *Mathematical Methods in the Social Sciences*, ed. K.J. Arrow et al. Stanford: Stanford University Press.

Debreu, G. 1964. Continuity properties of Paretian utility. *International Economic Review* 5, 285–93.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40, 603–15.

Debreu, G. 1976. Smooth preferences. A corrigendum. *Econometrica* 44, 831–2.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of Quantitative Economics*, vol. 2, ed. M.D. Intriligator and D.A. Kendrick. Amsterdam: North-Holland.

Diewert, W.E. 1982. Duality approaches to microeconomic analysis. In Arrow and Intriligator (1982).

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Eilenberg, S. 1941. Ordered topological spaces. *American Journal of Mathematics* 63, 39–45.

Estévez Toranzo, M. and Hervés Beloso, C. 1995. On the existence of continuous preference orderings without utility representations. *Journal of Mathematical Economics* 24, 305–9.

Fisher, I. 1892. Mathematical investigations in the theory of value and prices. *Transactions of the Connecticut Academy of Arts and Sciences* 9, 1–124. Repr. in *The Works of Irving Fisher*, vol. 1, ed. W.J. Barber. London: Pickering and Chatto, 1997.

Gossen, H.H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs und der daraus fließenden Regeln für menschliches Handeln*. Braunschweig, 2nd edn. Berlin: Prager, 1889.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Hildenbrand, W. 1994. *Market Demand: Theory and Empirical Evidence*. Princeton: Princeton University Press.

Houthakker, H.S. 1950. Revealed preference and the utility function. *Economica* N.S. 17, 159–74.

Houthakker, H.S. 1960. Additive preferences. *Econometrica* 28, 244–57; Errata: *Econometrica* 30 (1962), 633.

Hurwicz, L. 1971. On the problem of integrability of demand functions. In Chipman et al. (1971).

Hurwicz, L. and Uzawa, H. 1971. On the integrability of demand functions. In Chipman et al. (1971).

Jevons, W.S. 1871. *Theory of Political Economy*. London: Macmillan.

John, R. 1984. A counterex to a conjecture concerning the nontransitive consumer. Discussion Paper No. 151. Sonderforschungsbereich 21, University of Bonn.

John, R. 1995. The weak axiom of revealed preference and homogeneity of demand functions. *Economics Letters* 47, 11–16.

Kannai, Y. 1977. Concavifiability and construction of concave utility functions. *Journal of Mathematical Economics* 4, 1–56.

Katzner, D.W. 1968. A note on the differentiability of consumer demand functions. *Econometrica* 36, 415–8.

Katzner, D.W. 1970. *Static Demand Theory*. New York: Macmillan.

Katzner, D.W. 1971. Demand and exchange analysis in the absence of integrability conditions. In Chipman et al. (1971).

Kihlstrom, R., Mas-Colell, A. and Sonnenschein, H. 1976. The demand theory of the weak axiom of revealed preference. *Econometrica* 44, 971–8.

Kim, T. and Richter, M.K. 1986. Nontransitive-nontotal consumer theory. *Journal of Economic Theory* 38, 324–63.

Koopmans, T. 1972. Representation of preference orderings with independent components of consumption. In *Decision and Organization*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Leontief, W. 1947. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15, 361–73. Repr. in *Selected Readings in Economic Theory*, ed. K.J. Arrow. Cambridge, MA: MIT Press, 1971.

Mas-Colell, A. 1974. Continuous and smooth consumers: approximation theorems. *Journal of Economic Theory* 8, 305–36.

Mas-Colell, A., Whinston, M.D. and Green, J. 1995. *Microeconomic Theory*. Oxford: Oxford University Press.

Neuefeind, W. 1972. On continuous utility. *Journal of Economic Theory* 5, 174–6.

Pareto, V. 1896. *Cours d'Économie Politique*. Lausanne: Rouge.

Pareto, V. 1906. L'ofelimità nei cicli non chiusi. *Giornale degli economisti* 33, 15–30. Trans. as 'Ophelimity in non-closed cycles', in Chipman et al. (1971).

Rader, T. 1963. The existence of a utility function to represent preferences. *Review of Economic Studies* 30, 229–32.

Richter, M.K. 1966. Revealed preference theory. *Econometrica* 34, 635–45.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. 1950. The problem of integrability in utility theory. *Economica* 17, 355–85.

Savage, L. 1954. *Foundations of Statistics*. New York: Wiley.

Schmeidler, D. 1971. A condition for the completeness of partial preference relations. *Econometrica* 39, 403–4.

Schmeidler, D. 1986. Integral representation without additivity. *Proceedings of the American Mathematical Society* 97, 255–61.

Schmeidler, D. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57, 571–87.

Shafer, W. 1974. The nontransitive consumer. *Econometrica* 42, 913–9.

Slutsky, E. 1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti e Rivista di Statistica* 51, 1–26. Trans. as ‘On the theory of the budget of the consumer’, in *Readings in Price Theory*, ed. G.J. Stigler and K.E. Boulding. Homewood, IL: Irwin, 1953.

Sonnenschein, H. 1971. Demand theory without transitive preferences, with applications to the theory of competitive equilibrium. In Chipman et al. (1971).

Sono, M. 1945. The effect of price changes on the demand and supply of separable goods. *Kokumni Keizai Zasski* 74, 1–51 [in Japanese]. English translation: *International Economic Review* 2 (1960), 239–71.

Uzawa, H. 1960. Preference and rational choice in the theory of consumption. In Chipman et al. (1971).

Walras, L. 1874. *Elements d’économie politique pure*. Lausanne: Corbaz. Trans. W. Jaffé as *Elements of Pure Economics*. London: Allen and Unwin, 1954.

How to cite this article

Böhm, Volker and Hans Haller. "demand theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000069> doi:10.1057/9780230226203.0372

The New Palgrave Dictionary of Economics Online

demand-pull inflation

George L. Perry
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘demand-pull inflation’ originated with the Keynesian macroeconomic model and was used to contrast price increases arising from excess demand with those arising from shocks to aggregate supply. Phillips curve models were initially amended by natural rate models and by models that appended rational expectations and flexible wages and prices to natural rate models. It is now recognized that the response of inflation and unemployment to shifts in aggregate demand itself depends on the inflation environment, and moderate inflation is the desired environment. Stabilization policy continues to distinguish between supply shocks affecting prices and the effects of aggregate demand.

Keywords

accelerationist inflation models; aggregate demand; aggregate supply; core inflation; cost-push inflation; demand-pull inflation; excess demand; Federal Reserve System; Friedman, M.; full employment; incomes policies; inflation; inflation targeting; inflationary expectations; Keynesianism; monetary policy; natural rate of unemployment; neo-Keynesian models; Organization of Petroleum Exporting Countries; Phelps, E.; Phillips curve; price control; rational expectations; stabilization policies; sticky prices; sticky wages; Tobin, J.; unemployment; Volcker, P.; wage-price spiral

Article

The term ‘demand-pull’ inflation originated with the simple Keynesian model of the macroeconomy and was used as a contrast to price increases arising from shocks to aggregate supply. In the Keynesian model, there is a well-defined level of potential GDP corresponding to full employment levels of employment and unemployment. Nominal wages are downwardly rigid, so that below full employment aggregate supply increases with prices while aggregate demand decreases. The difference between potential and actual GDP is the output gap, and there is an asymmetry in the economy's response to shifts in demand when output gaps are positive and when they are not. With a positive gap – that is, in

the operating region below full employment – an expansion of aggregate demand mainly raises employment and output and only moderately raises prices. But at full employment the aggregate supply curve is vertical, and an expansion of demand only pulls up wages and prices. Hence the term ‘demand-pull inflation’.

Macro models of fluctuations have evolved in important ways from this simple Keynesian case. The early empirical Phillips curves described an empirical relation between the level of unemployment and rates of change, rather than levels, of prices. Such relations were estimated from periods characterized by frequent cycles in activity. They did not control for expected or ongoing rates of inflation, so did not directly address the consequences of maintaining real aggregate demand at levels that raised prices. James Tobin (1972), among others, reasoned that the average wage and price increases associated with approaching full employment in the empirical Phillips curves came from the operation of a heterogeneous labour market in which demand constantly shifted among sectors. In his model, the short-run inflation that was observed in the typical cyclical episode reflected wage and price changes that reduced wasteful search unemployment, rather than a misguided attempt to sustain employment above the full employment level.

The first important departure came from theoretical models based on representative agents and firms that examined the consequences of permanently maintaining demand at levels that raised wages and prices in the short run. In the late 1960s Milton Friedman (1968) and Edmund Phelps (1969) independently formulated models of a natural rate of unemployment in which inflation fed back fully into wages and hence prices, so that an unemployment rate below the natural rate could be sustained only by ever-higher inflation rates. In effect, these accelerationist price models resurrected the vertical Keynesian supply curve at full employment for the long run, but allowed demand policies that raised the inflation rate in the short run to achieve lower levels of unemployment, but only temporarily. Since the higher employment associated with price increases could not be sustained, a corollary was that zero inflation was the appropriate target for policy. Tobin's model, with its heterogenous economy, denied that a natural rate identified by prices rising faster corresponded to full employment. However, the natural rate model became widely accepted as a theoretical construct, especially after the introduction of rational expectations models in which anticipation of faster or slower price increases would speed up the process of price acceleration or deceleration. Some theoretical models also assumed price and wage flexibility rather than stickiness. And some even rejected the idea that aggregate demand could leave the economy below full employment, modelling all cyclical variations in output and employment as shocks to aggregate supply. Modern neo-Keynesian models retain both the assumption of price and wage stickiness, which is supported by empirical research, and the implication that output can depart from its potential level. But they attach a more central role to expectations than do early Keynesian models.

All these models share the original idea of demand-pull inflation in that inflation arises when aggregate demand is excessive. They differ in their description of how the process works out over different time horizons and empirically in how the region of excess demand can be identified for informing forecasters and policymakers. Empirical implementation of rational expectations models continues to be elusive, and most empirical work has used adaptive expectations with accelerationist models to estimate the natural rate and the level of potential output. These estimates proved to be unreliable in the 1990s when economic expansion steadily reduced unemployment rates well below those predicted to cause accelerating inflation in those models. Some recent research has supported the idea that a modest rate of inflation, rather than complete price stability, is necessary to maintain the fullest utilization of resources.

This can be so for a variety of reasons. With downward wage rigidity, price stability will keep real wages above their efficient level in a noticeable fraction of firms. Moderate inflation will minimize this problem, permitting the economy to achieve optimal employment (Akerlof, Dickens and Perry, 1996). Furthermore, very low inflation rates will be ignored by many economic agents, leading firms to sustain output and employment at levels above those of a full expectational equilibrium (Akerlof, Dickens and Perry, 2000). And on the demand side, with very low or zero inflation, the zero floor on nominal interest rates may prevent monetary policy from getting real interest rates low enough to achieve full employment. The experience of Japan after its financial bubble burst is an example (Krugman, 1998).

Originally, the explicit modelling of demand-pull inflation was important because of the distinction it drew between price increases arising from excess demand and price increases originating in shifts up in the aggregate supply schedule, also referred to as cost-push. The sharp increases in wage costs that occurred in the heyday of union strength in industrialized economies are important historical examples of shifts in aggregate supply schedules. In the 1960s and 1970s, the experience with such cost-push shocks motivated the attempts to impose wage-price guideposts in the United States, and similar incomes policies in the United Kingdom and elsewhere. Such incomes policies were seen as a way to contain excessive wage and price increases that arose when the economy was operating below its full employment level.

Although there has been no recent interest in incomes policies, the distinction between price increases originating in excess aggregate demand and those originating from shifts in important supply schedules continues to be a feature of policy deliberations and of empirical work today. Core inflation rates, which omit the impact effect of energy and food prices on aggregate price indices, are routinely reported in monthly statistical releases, reflecting a distinction most analysts find useful. Core inflation rates are seen as more likely to feed back into wage increases, and are a better indicator of demand-pull effects on prices. And policymakers regularly make allowances for the effect of supply shocks in considering their stabilization response to changes in reported inflation rates.

History provides examples of significant inflation in which excess demand or major supply shocks or both were important. In the United States, during the Second World War and the Korean War maximizing output was the paramount goal of government even though it meant expanding demand well beyond the normal full-employment point. The potential inflation generated by operating in this excess-demand region was moderated, if not completely suppressed, by rationing and price controls. Demand-pull inflation was also a feature of the industrial economies in the late 1960s, when US military spending was greatly enlarged and labour and product markets became tight for an extended period throughout the industrial world. An abrupt explosion of wage increases at the end of the 1960s and in the early 1970s in most industrialized countries suggests that cost-push contributed importantly to the inflation of that period. The rise in food prices in 1973 and the oil supply shocks of 1973 and 1979 added further to the ongoing inflation of that decade and doubtless contributed to an increase in inflationary expectations and to the response of unions and firms to those expectations.

It was particularly striking that inflation was so little affected by the very deep recessions of the mid-1970s in the advanced economies. That episode convinced most economists of the shortcomings of the simple short-run Phillips curve model, which predicted that inflation would slow cyclically in the mid-1970s. But it was also not consistent with flexible price accelerationist models which predict that prices and wages will fall when the economy is operating below its natural rate. It did support the pessimistic verdict that a well-established inflation can persist long after the initiating shocks have disappeared and

long after a reduction of demand has eliminated any excess demand from the economy.
The stabilization challenge confronting policymakers in that period was seen not merely as avoiding excess aggregate demand, but also as choosing how much to accommodate inflation in order to maintain real growth and how much to give up in output and employment in order to suppress inflation. After the second OPEC oil price shock in 1979, Paul Volcker was appointed Chairman of the US Federal Reserve and, under his leadership, the Fed chose to strongly suppress demand until inflation receded sharply. The lower inflation that ensued is consistent with the predictions of some conventional cyclical models. The severity of the policy used, as reflected in the record high interest rates it produced and the very deep recession that policymakers tolerated, can also be interpreted as evidence that policymakers can shape expectations and that doing so affects how promptly the inflation rate changes.
In the United States, the period that began in the 1990s was a sharp contrast to the 1970s in that inflation had been moderate for many years. As noted above, by the end of the decade the unemployment rate had fallen well below existing empirical estimates from natural-rate models. Yet inflation remained very low, both before the modest recession of the early 2000s and in the several years after it, even after a new oil price shock. Most European economies experienced similarly low inflation in this period. However, several suffered from chronically high rates of unemployment. While considerable controversy surrounds the reasons for this persistence of unemployment, some analysts believe inadequate aggregate demand over an extended period is partly to blame. There are several implications for stabilization policies aimed at avoiding inflation from all this experience: While empirical estimates from the 1970s suggested inflation was prone to quicken through a wage-price spiral, the recent period suggests no such tendency so long as inflation rates are modest (Brainard and Perry, 2000). Furthermore, the economy's potential output and the attainable unemployment rate – the thresholds of the demand-pull region of resource utilization – cannot be adequately estimated using typical accelerationist models. Finally, the contrasting experiences across the United States and European economies show that policies targeting inflation alone are not sufficient to assure full employment.

See Also

- cost-push inflation
- inflation
- monetary business cycle models (sticky prices and wages)
- monetary and fiscal policy overview

Bibliography

Akerlof, G., Dickens, W. and Perry, G. 1996. The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1996(1), 1–59.

Akerlof, G., Dickens, W. and Perry, G. 2000. Near-rational wage and price setting and the long-run Phillips curve. *Brookings Papers on Economic Activity* 2000(1), 1–44.

Blanchard, O. and Summers, L. 1986. Hysteresis and the European unemployment problem. In *NBER Macroeconomics Annual*, ed. S. Fischer. Cambridge, MA: MIT Press.

Brainard, W. and Perry, G. 2000. Making policy in a changing world. In *Economic Events, Ideas, and Policies*, ed. G. Perry and J. Tobin. Washington, DC: Brookings Institution.

Calvo, G. 1983. Staggered prices in a utility maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Gordon, R. 1981. Output fluctuations and gradual price adjustment. *Journal of Economic Literature* 19, 493–530.

Hall, R., ed. 1982. *Inflation: Causes and Effects*. Chicago: University of Chicago Press for the NBER.

Krugman, P. 1998. It's baaack: Japan's slump and the return of the liquidity trap. *Brookings Papers on Economic Activity* 1998(2), 137–87.

Lucas, R. 1972. Econometric testing of the natural rate hypothesis. In *The Econometrics of Price Determination*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.

Nickell, S. 1997. Unemployment and labor market rigidities: Europe versus North America. *Journal of Economic Perspectives* 11(3), 55–74.

Nordhaus, W. 1981. Macroconfusion: the dilemma of economic policy. In *Macroeconomics, Prices, and Quantities*, ed. J. Tobin. Washington, DC: Brookings Institution.

Okun, A. 1975. Inflation: its mechanics and welfare costs. *Brookings Papers on Economic Activity* 1975 (2), 351–90.

Okun, A. 1981. *Prices and Quantities: A Macroeconomic Analysis*. Washington, DC: Brookings Institution.

Perry, G. 1980. Inflation in theory and practice. *Brookings Papers on Economic Activity* 1980(1), 207–41.

Phelps, E. 1969. The new microeconomics in inflation and employment theory. *American Economic Review* 59, 147–60.

Sargent, T. 1982. The ends of four big inflations. In *Inflation: Causes and Consequences*, ed. R. Hall.

Chicago: University of Chicago Press.

Schultze, C. 1981. Some macro foundations for micro theory. *Brookings Papers on Economic Activity* 1981(2), 521–76.

Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–24.

Tobin, J. 1972. Inflation and unemployment. *American Economic Review* 62, 1–18.

How to cite this article

Perry, George L. "demand-pull inflation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000218> doi:10.1057/9780230226203.0371

The New Palgrave Dictionary of Economics Online

democratic paradoxes

Norman Schofield
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Formal models of voting have emphasized the *mean voter theorem*, namely, that all parties should rationally adopt identical positions at the electoral mean. The lack of evidence for this assertion is a *parado*x which this article attempts to resolve by considering an electoral model that includes ‘valence’ or non-policy judgements by voters of party leaders. In a polity such as Israel, based on proportional electoral rule, low-valence parties would adopt positions far from the centre, making coalition formation unstable. In Britain, by contrast, a party with a low-valence leader would be subject to the demands of non-centrist activists.

Keywords

Condorcet, Marquis de; democratic paradoxes; Downs, A.; Hotelling, H.; local Nash equilibrium; Madison, J.; median voter theorem; mixed strategy Nash equilibrium; plurality electoral rule; political competition; proportional representation; pure strategy Nash equilibrium; valence; vote maximizing strategies; voting

Article

Models of elections tend to give two quite contradictory predictions about the result of political competition. In two-party competition, if the ‘policy space’ involves two or more independent issues, then ‘pure strategy Nash equilibria’ generally do not exist and instability or *chao*s may occur (see Plott, 1967; McKelvey, 1976; 1979; Schofield, 1978; 1983; 1985; McKelvey and Schofield, 1986; 1987; Saari, 1997; Austen-Smith and Banks, 1999). That is to say, whatever position is picked by one party, there always exists another policy point which will give the second party a majority over the other. Moreover, vote maximizing strategies could lead political candidates to wander all over the policy space.

On the other hand, the earlier electoral models based on the work of Hotelling (1929) and Downs (1957) suggest that parties will converge to an electoral centre (at the electoral *median*) when the policy space has a single dimension. (An equilibrium can also be guaranteed as long as the decision rule requires a sufficiently large majority – Schofield, 1984; Strnad, 1985; Caplin and Nalebuff, 1988 – or when the electoral distribution has a certain concavity property – Caplin and Nalebuff, 1991.) Although a pure strategy Nash equilibrium generically fails to exist in competition between two agents under majority rule in high enough dimension, there will exist mixed strategy Nash equilibria (Kramer, 1978) whose support lies within a subset of the policy space known as the ‘uncovered set’ (see McKelvey, 1986; Banks, Duggan and Le Breton, 2002). These various and contrasting theoretical results can be seen as a paradox: will democracy tend to generate centrist compromises, or can it lead to chaos? This question is of fundamental importance in a world in which many countries are experimenting with democracy for the first time.

Partly as a result of these theoretical difficulties with the ‘deterministic’ electoral model, and also because of the need to develop empirical models of voter choice (Poole and Rosenthal, 1984), attention has focused on ‘stochastic’ vote models. A formal basis for such models is provided by the notion of ‘quantal response equilibria’ (McKelvey and Palfrey, 1995). In such models, the behaviour of each voter is modelled by a vector of choice probabilities (Lin, Enelow and Dorussen, 1999). A standard result in this class of models is that all parties converge to the electoral origin when the parties are motivated to maximize vote share or plurality (in the two-party case) (see McKelvey and Patty, 2006; Banks and Duggan, 2005). The predictions concerning convergence are at odds with empirical evidence that parties appear to diverge from the electoral centre (Merrill and Grofman, 1999; Adams, 2001; Schofield and Sened, 2006).

The *paradox* that actual political systems display neither *chaos* nor *convergence* is the subject of this article. The key idea is that the convergence result need not hold if there is an asymmetry in the electoral perception of the 'quality' of party leaders (Stokes, 1992). The average weight given to the perceived quality of the leader of the jth party is called the party's 'valence'. In empirical models this valence is independent of the party's position, and adds to the statistical significance of the model. In general, valence reflects the overall degree to which the party is perceived to have shown itself able to govern effectively in the past, or is likely to be able to govern well in the future (Penn, 2003). The early empirical model of US presidential elections by Poole and Rosenthal (1984) included these valence terms. The authors noted that there was no evidence of candidate convergence.

Formal models of elections incorporating valence have been developed (Ansolabehere and Snyder, 2000; Groseclose, 2001; Aragones and Palfrey, 2002), but the theoretical results to date have been somewhat inconclusive. Extension to the multiparty case is of interest because of recent empirical models of voting in the Netherlands and Germany (Schofield et al., 1998; Quinn, Martin, and Whitford, 1999; Quinn, and Martin, 2002), Britain (Schofield, 2005a; 2005b), Israel (Schofield, Sened and Nixon, 1998; Schofield and Sened, 2002; 2005; 2006) and Italy (Giannetti and Sened, 2004). All these empirical models have suggested that divergence is generic. Most of these empirical models have been based on the 'multinomial logit' assumption that the stochastic errors had a 'Type I extreme value distribution' (Dow and Endersby, 2004).

Schofield (2007) provides a 'classification theorem' for the formal vote model based on the same stochastic distribution assumption. The 'policy space' is assumed to be of dimension w, and there is an arbitrary number, p, of parties. The party leaders exhibit differing valence. A 'convergence coefficient' incorporating all the parameters of the model can be defined. Instead of using the notion of a Nash equilibrium, the result is given in terms of the existence of a 'local Nash equilibrium'. It is shown that there are necessary and sufficient conditions for the existence of a 'pure strategy vote maximizing local Nash equilibrium' (LNE) at the mean of the voter distribution. When the necessary condition fails, then parties, in equilibrium, will adopt divergent positions. In general, parties whose leaders have the lowest valence will take up positions furthest from the electoral mean. Moreover, because a pure strategy Nash equilibrium (PNE) must be a local equilibrium, the failure of existence of the LNE at the electoral mean implies non-existence of such a centrist PNE. The failure of the necessary condition for convergence has a simple explanation: if the variance of the electoral distribution is sufficiently large in contrast to the expected vote share of the lowest-valence party at the electoral mean, then this party has an incentive to move away from the origin towards the electoral periphery. Other low-valence parties will follow suit, and the local equilibrium will result with parties distributed along a 'principal electoral axis'.

An empirical study of voter behaviour for Israel for the election of 1996 (based on Schofield and Sened, 2005) is used to show that the necessary condition for party convergence failed for this election. The equilibrium positions obtained from the formal result, under vote maximization, are in general comparable with, but not identical to, the estimated positions. The two highest-valence parties (Labour and Likud) were symmetrically located on either side of the electoral origin, while the lowest-valence parties were located far from the origin. In such a polity, based on a proportional electoral system, it is generally necessary to form coalition governments. The existence of small, low-valence, radical parties on the electoral periphery may create serious difficulties in the formation of majority government. It is possibly for this reason that Ariel Sharon, formerly leader of Likud, and Shimon Peres, formerly leader of Labour, in 2005 formed Kadima, a new centrist party.

This article also presents results from analysis of the 1997 election in Britain (Schofield, 2005a; 2005b). In this case the empirical estimates of the parameters of the model, taken together with the formal analysis, suggest that convergence should have occurred. Instead the Conservative Party was estimated to be at a position far from the electoral centre. It is suggested that the discrepancy between the formal and the empirical models can be accommodated by considering the effect of activists on the optimal party position. Since concerned activists will raise funds for the party, but only if the party adopts a policy position that accords with activists' concerns, there is a tension between activist demands and the electoral concerns of the party leadership. The model based on activist support estimates the marginal trade-off generated by opposed activist groups within a party. It is suggested that the low valence of recent Conservative leaders obliged them to seek support from activists supporting British sovereignty against the European Union, and thus to take up radical positions on the second, 'European' axis.

In contrast, the apparent move by the Labour Party towards the electoral centre between 1992 and 1997 was a consequence of the increase of the electoral valence of Tony Blair, the leader of the party, rather than a cause of this increase.

Recent work by Miller and Schofield (2003) using this model suggests that, in the United States, the movement of presidential candidates in a two-dimensional policy space generated by economic and social dimensions is the result of contending and opposed activist groups.

The underlying premise of the notion of the *local Nash equilibrium*, used in these models, is that party leaders will not consider 'global' changes in party policies, but will instead propose small changes in the party position in response to changes in beliefs about electoral response. These models regard elections as the aggregation of both electoral evaluation or 'valence' and electoral preferences. Valence can be regarded as that element of a voter's choice which is determined by judgement rather than preference. This accords well with the arguments of James Madison in *Federalist 10* of 1787 (Rakove, 1999) and of Condorcet (1785) in his treatise on social choice theory. Schofield (2005c; 2006) provides a discussion of the relevance of these valence models for the constitutional basis of the US polity.

Empirical analysis for Israel

Figure 1 shows the estimated positions of the parties at the time of the 1996 Israeli election. Figure 1 also gives the estimated distribution of voter ideal points for the 1996 election,

based on a factor analysis of the survey responses derived from the survey of Arian and Shamir (1999). The two dimensions of policy deal with attitudes to the Palestine Liberation Organization (PLO) (the horizontal axis) and religion (the vertical). The party positions were obtained from analysis of party manifestos (Schofield, Sened and Nixon, 1998; Schofield and Sened 2005; 2006). With the use of information on the individual voter intentions, it is possible to construct a multinomial logit model (based on the Type I extreme value distribution).

Figure 1

Voter distribution and estimated party positions in the Knesset at the 1996 election

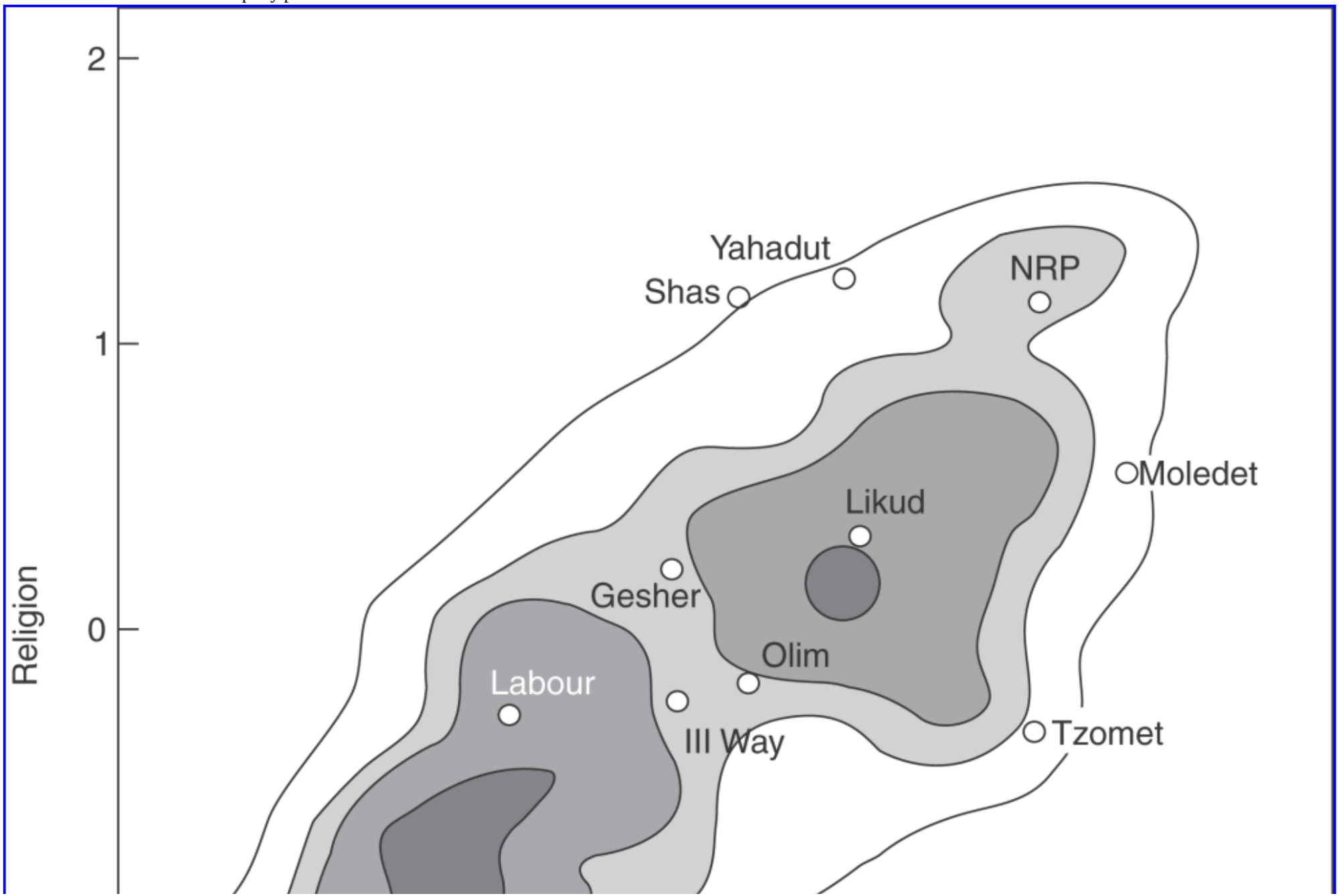

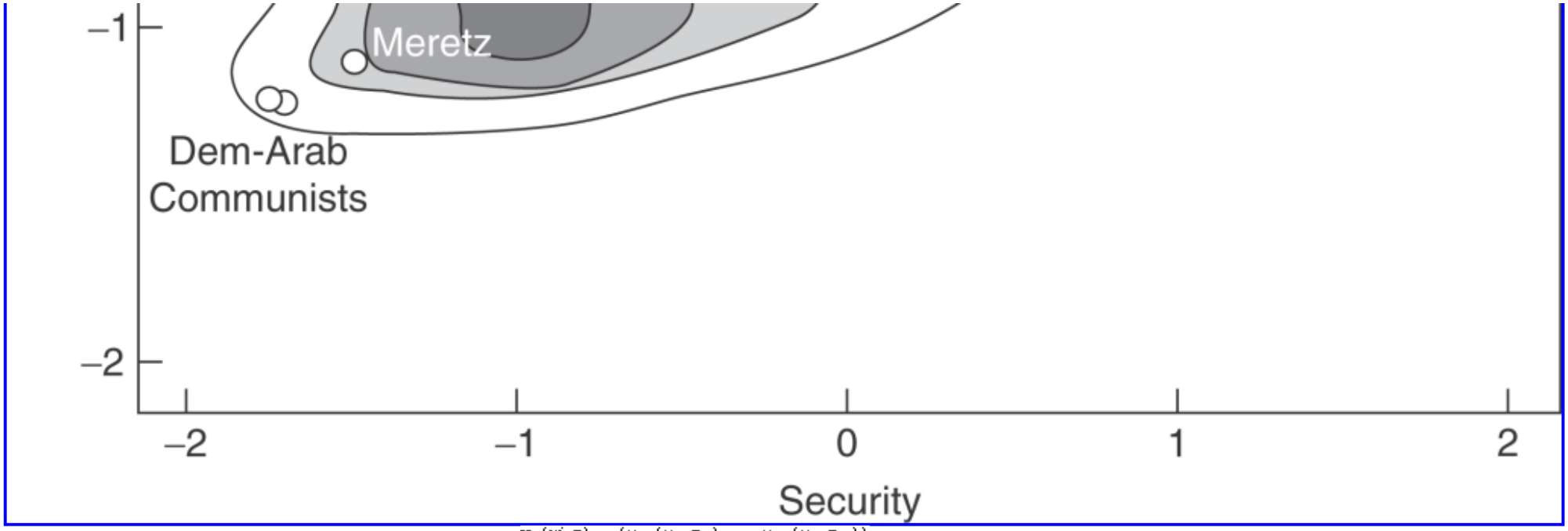

The model assumes that the voter utility vector has the form $\mathbf{u}_i(xi, \mathbf{z}) = (u_{i1}(x_i, z_1), \ldots, u_{ip}(x_i, z_p))$, where

$$u_{ij}(x_i, z_j) = u^*_{ij}(x_i, z_j) + \varepsilon_j$$
$$\text{and } u^*_{ij}(x_i, z_j) = \lambda_j - \beta \| x_i - z_j \|^2 .$$

Here the position of voter *i* is x_i while the position of party *j* is z_j. The term $\| x_i - z_j \|$ is the distance between these two points. The *valence*s of the *p* parties are given by the vector $\lambda = (\lambda_p, \lambda_{p-1}, \ldots, \lambda_2, \lambda_1)$ and are ranked

$$\lambda_p \geq \lambda_{p-1} \geq \ldots \geq \lambda_2 \geq \lambda_1.$$

The error terms $\{ \varepsilon_j \}$ have the *Type I extreme value distribution*, Ψ.
(The cumulative distribution, Ψ, takes the closed form $\Psi(h) = \exp[- \exp[- h]]$.)
The probability that a voter *i* chooses party *j* is

$$p_{ij}(\mathbf{z}) = \Pr[[u_{ij}(x_i, z_j) > u_{il}(x_i, z_l)], \text{ for all } l \neq j].$$

Here Pr stands for the probability operator associated with Ψ. The expected vote share of agent *j* is

$$V_j(\mathbf{z}) = \frac{1}{n}\sum_{i \in N} \rho_{ij}(\mathbf{z}).$$

This model is denoted $M(\lambda, \beta; \Psi)$. A *local pure strategy Nash equilibrium* (LNE) is simply a vector $\mathbf{z} = (z_1, \ldots, z_p)$ of party positions with the property that each z_j locally maximizes $V_j(\mathbf{z})$, taking the other party positions A necessary condition for $\mathbf{z}^* = (\mathbf{0}, \ldots, \mathbf{0})$ to be *pure strategy Nash equilibrium* (PNE) is that it be a LNE and thus that all Hessians have eigenvalues at $\mathbf{z}^*$ that are non-positive. This can be expressed as a single necessary condition on a 'convergence coefficient' defined terms of the Hessian of the vote share function of the party with the lowest valence (Schofield, 2006b). Since the lowest-valence party is the National Religious Party (NRP) (for the 1996 model for Israel), a *necessary* condition for the NRP vote share to be maximized at the origin is that *both* eigenvalues of this Hessian be non-positive. However, the calculation given below shows that that one of the eigenvalues was positive. It follows that the NRP position that maximizes its vote share is *not* at the origin. Thus the convergent position (**0**,…, **0**) cannot be a Nash equilibrium to the vote maximizing game.

Indeed it is obvious that there is a principal component of the electoral distribution, and this axis is the eigenspace of the positive eigenvalue. It follows that low-valence parties should then position themselves on this eigenspace, as illustrated in the simulation given in Figure 2.

Figure 2

A simulated local Nash equilibrium in the vote maximizing game in Israel in 1996. *Note*: 1: Shas; 2: Likud; 3: Labour; 4: NRP; 5: Molodet; 6: III Way; 7: Meretz.

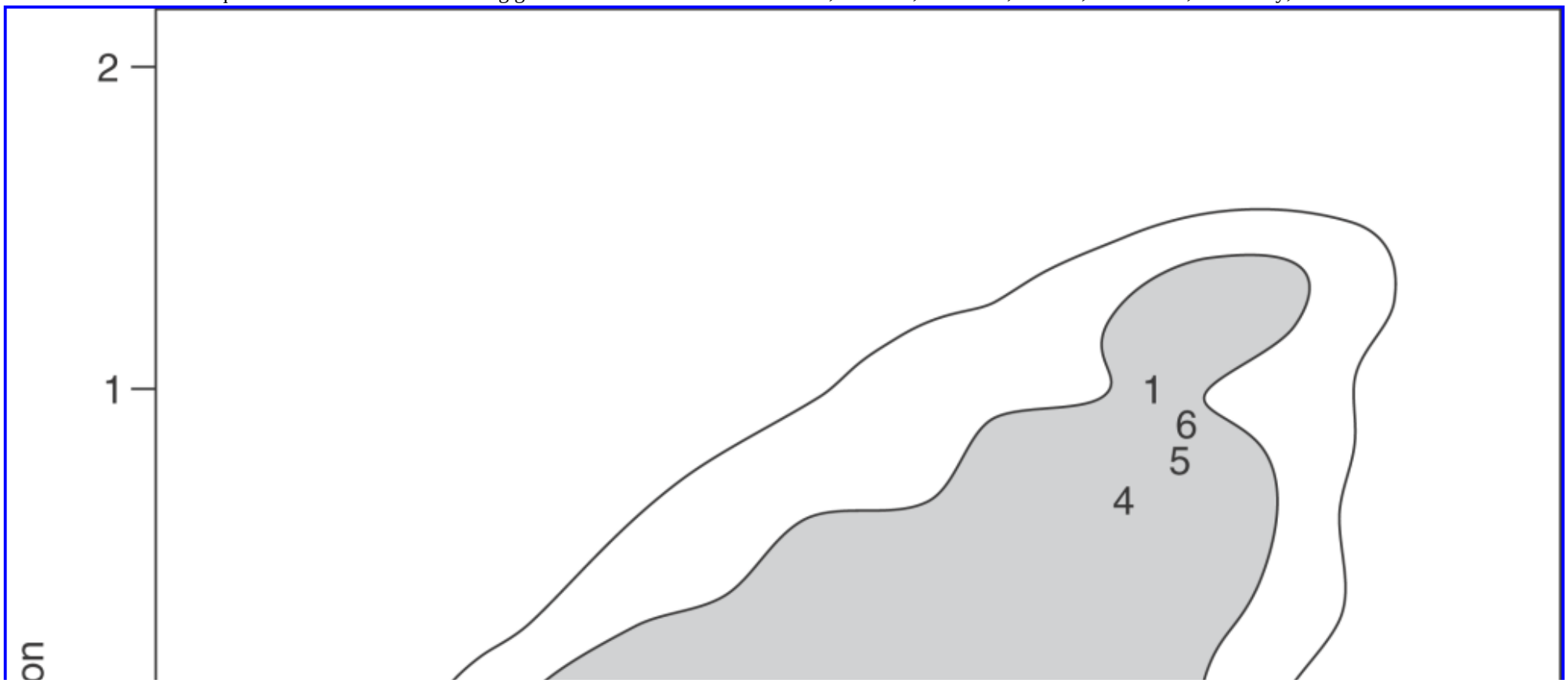

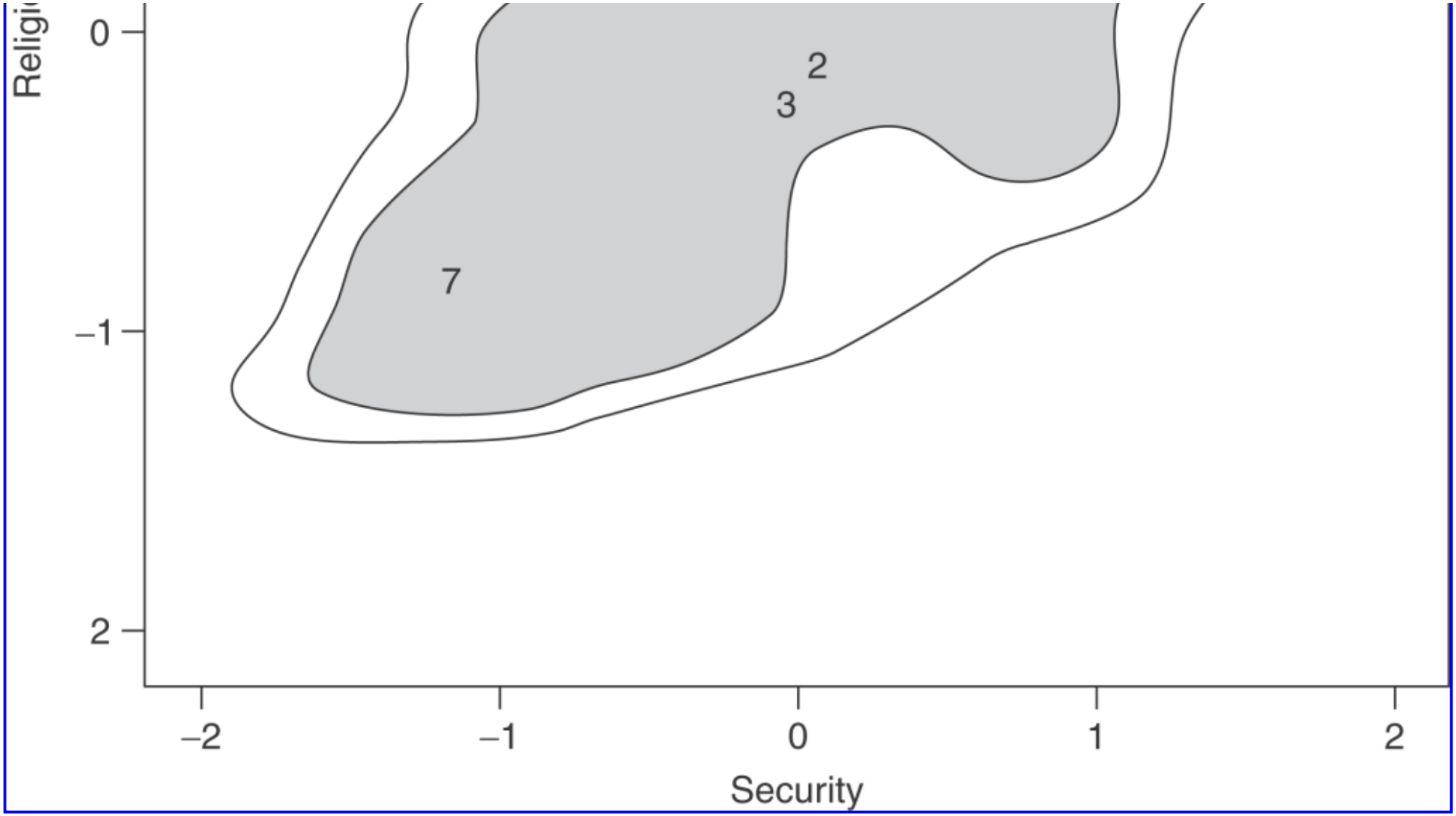

To present the calculation, we use the fact that the valence of the NRP was – 4.52. The spatial coefficient is $\beta = 1.12$. Because the valences of the major parties are 4.15 and 3.14, the formal analysis implies that, when all parties are at the origin, the vote share, ρ_{NRP}, can be computed to be

$$\rho_{NRP} \simeq \frac{1}{1 + e^{4.15+4.52} + e^{3.14+4.52}} \simeq 0.$$

Moreover, the Hessian of the NRP at the origin depends on the electoral variance and this is

$$C_{NRP} = 2(1.12)\begin{pmatrix} 1.0 & 0.591 \\ 0.591 & 0.732 \end{pmatrix} - I = \begin{pmatrix} 1.24 & 1.32 \\ 1.34 & 0.64 \end{pmatrix}.$$

The eigenvalues of the NRP Hessian at the origin are 2.28 and –0.40, giving a saddle point. Thus, the origin cannot be a Nash equilibrium. The ‘convergence coefficient’ can be calculated to be 3.88, larger than the necessary upper bound of 2.0. The major eigenvector for the NRP is (1.0,0.8), and along this axis the NRP vote share function increases as the party moves away from the origin. The minor, perpendicular axis is given by the vector (1,–1.25) and on this axis the NRP vote share decreases. Figure 2 gives one of the local equilibria in 1996, obtained by simulation of the model. The figure makes it clear that the vote maximizing positions lie on the principal axis through the origin and the point (1.0, 0.8). In all, five different LNE were located. However, in all the equilibria the two high-valence parties, Labour and Likud, were located at precisely the same positions, as shown in Figure 2. The only difference between the various equilibria was that the positions of the low-valence parties were perturbations of each other.

Figure 2 suggests that the simulation was compatible with the predictions of the formal model based on the extreme value distribution. All parties were able to increase vote shares by moving away from the origin, along the principal axis, as determined by the large, positive principal eigenvalue. In particular, the simulation confirms the logic of the above analysis. Low-valence parties, such as NRP and Shas, in order to maximize vote shares must move far from the electoral centre. Their optimal positions will lie in either the north-east quadrant or the south-west quadrant. The vote maximizing model, without any additional information, cannot determine which way the low-valence parties should move. As noted above, the simulations of the empirical models found multiple LNE essentially differing only in permutations of the low-valence party positions.

In contrast, since the valence difference between Labour and Likud was relatively low, their optimal positions would be relatively close to, but not identical to, the electoral mean. The simulation for the elections of 1988 and 1992 are also compatible with this theoretical inference. Figure 2 also suggests that every party, in local equilibrium, should adopt a position that maintains a minimum distance from every other party. The formal analysis, as well as the simulation exercise, suggests that this minimum distance depends on the valences of the neighbouring parties. Intuitively it is clear that, once the low-valence parties vacate the origin, then high-valence parties like Likud and Labour will position themselves almost symmetrically about the origin, and along the major axis.

Comparison between Figure 1, of the estimated party positions, and Figure 2, of simulated equilibrium positions, reveals a notable disparity particularly in the position of Shas. In 1996 Shas was pivotal between Labour and Likud, in the sense that, to form a winning coalition government, either of the two larger parties required the support of Shas. It is obvious that the location of Shas in Figure 1 suggests that it was able to bargain effectively over policy and, presumably, perquisites. Indeed, it is plausible that the leader of Shas was aware of this situation, and incorporated this awareness in the utility function of the party.

The close correspondence between the simulated LNE based on the empirical analysis and the estimated actual political consuggests that the true utility function for each party *j* has the form $U_j(\mathbf{z}) = V_j(\mathbf{z}) + \delta_j(\mathbf{z})$, where $\delta_j(\mathbf{z})$ may depend on the beliefs of party leaders about the post-election coalition possibilities, as well as the effect of activist support for the party.

This hypothesis leads to the further hypothesis that, for the set of feasible strategy profiles in the Israel polity, $\delta_j(\mathbf{z})$ is ‘small’ relative to $V_j(\mathbf{z})$. A formal model to this effect could indicate that the LNE for $\{U_j\}$ would be close to the LNE for $\{V_j\}$. Note, however, that this perturbation of the party utility function causes parties to leave the main electoral axis. It is possibly for this reason that coalition politics in Israel has been very complex.

The Likud Party, under Ariel Sharon, was constrained by the religious parties in its governing coalition. This apparently caused Sharon to leave Likud to set up a new centrist party, Kadima (‘Forward’) with Shimon Peres, previously leader of Labour. The reason for this reconfiguration was the victory on 10 November 2005 of Amir Peretz over Peres for leadership of the Labour Party, and Peretz's move to the left along the principal electoral axis.

Consistent with the model presented here, Sharon's intention was to position Kadima very near the electoral centre on both dimensions, to take advantage of his high valence among the electorate. Sharon's subsequent hospitalization had an adverse effect on the valence of Kadima, under its new leader, Ehud Olmert. Even so, in the election of 28 March 2006 Kadima took 29 seats, against 19 seats for Labour, and only 12 for Likud. One surprise was a new centrist pensioners’ party with 7 seats. Because Kadima with Labour and the other parties of the left had 70 seats, Olmert was able to put together a majority coalition on 28 April, including the Orthodox party Shas. As Figure 1 illustrates, Shas is centrist on the security dimension, indicating that this was the key issue of the election.

Empirical analysis for Britain

This section analyses the general election in Britain in 1997 in order to suggest how activists for the parties may influence party positioning. The analysis shows that the valence model as presented above cannot always explain divergence of party positions. For example, Figure 3 shows the estimated positions of the party leaders, based on a survey of party MPs in 1997 (Schofield, 2005a; 2005b). In addition to the Conservative Party, Labour Party, and Liberal Democrats, responses were obtained from Ulster Unionists, Scottish

Nationalists and Plaid Cymru (Welsh Nationalists). The axis is economic, the second pro or anti the European Union. The electoral model was estimated for the election in 1997, using only the economic dimension.

Figure 3

Estimated party positions in the British Parliament for a two-dimensional model. *Notes*: Highest-density contours of the voter sample distribution at the 95%, 75%, 50% and 10% levels. CONS: Conservative Party; LAB: Labour Party; LIB: Liberal Democrats; PC: Plaid Cymru (Welsh Nationalists); SNP: Scottish National Party; UU: Ulster Unionist Party. *Source*: MP survey data and a National Election Survey.

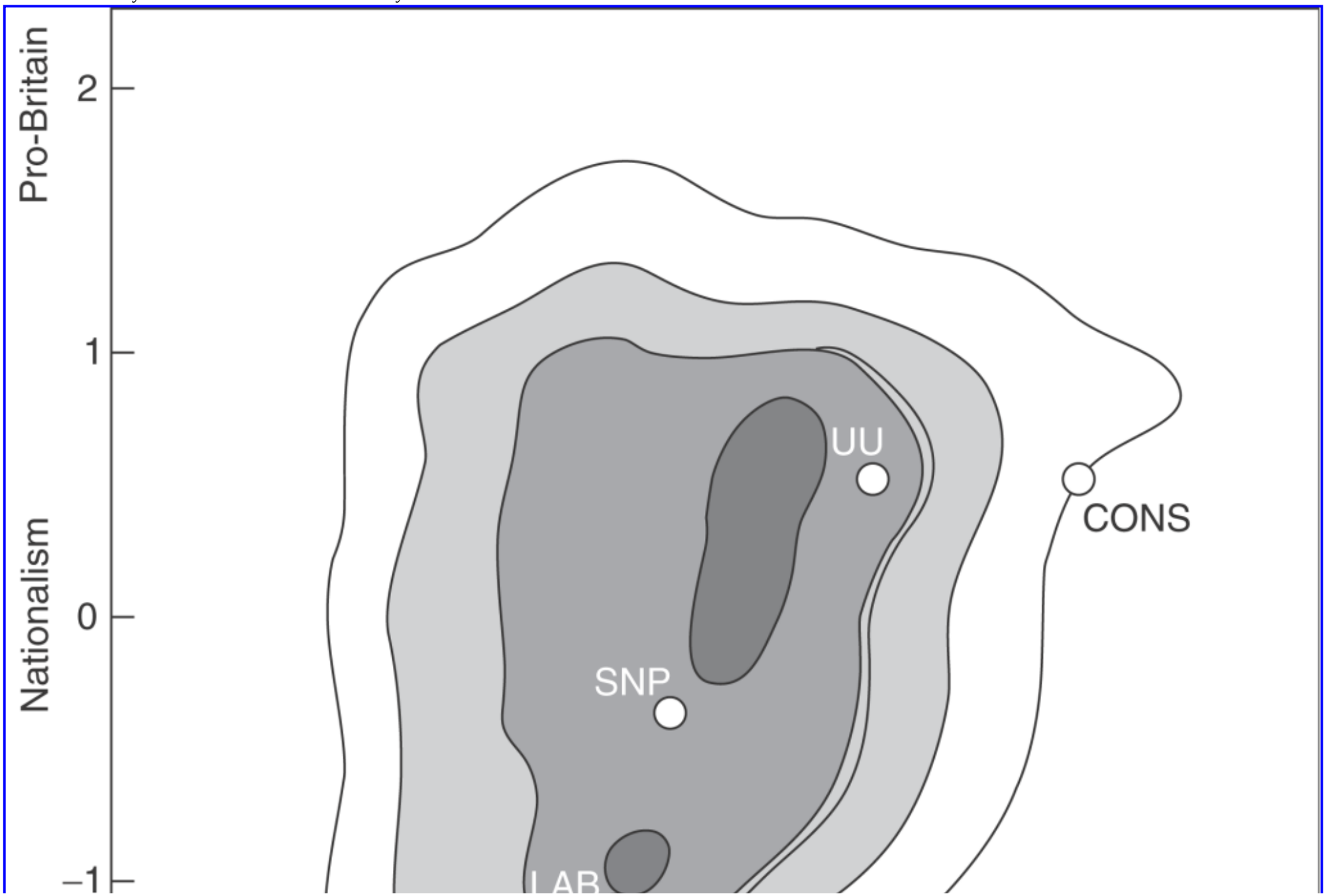

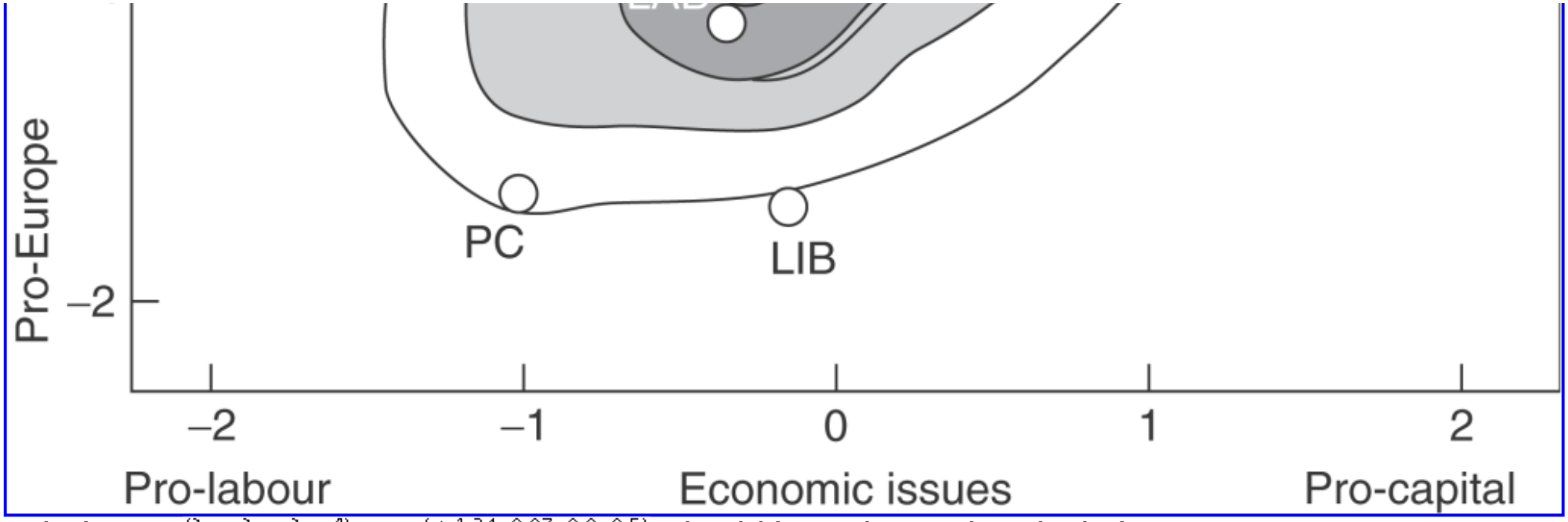

For this election, we $(\lambda_{con}, \lambda_{lab}, \lambda_{lib}, \beta)_{1997} = (+1.24, 0.97, 0.0, 0.5)$ so the probability ρ_{lib}, that a voter chooses the Liberal Democrats is

$$\rho_{lib} = \frac{e^0}{e^0 + e^{1.24} + e^{0.97}} = \frac{1}{7.08} = 0.14.$$

The Hessian for this party at the origin is $C_{lib} = -0.28$, which is compatible with a Nash equilibrium at the origin. Extending the model to two dimensions gives a Hessian

$$C_{lib} = (0.72)\begin{pmatrix} 1.0 & 0 \\ 0 & 1.5 \end{pmatrix} - I = \begin{pmatrix} -0.28 & 0 \\ 0 & +0.8 \end{pmatrix}.$$

According to the formal model, all parties should have converged to the origin on the first axis. Because the eigenvalue for the Liberal Democrats is positive on the second axis, we have an explanation for its position away from the origin on the Europe axis. However, there is no explanation for the location of the Conservative Party so far from the origin on both axes. Schofield (2005a; 2005b) adapts the activist model of Aldrich (1983a; 1983b) wherein the falling exogeneous valence of the Conservative Party leader increases the marginal importance of two opposed activist groups in the party: one group ‘pro-capital’ and one group ‘anti-Europe’, as in Figure 4.

Figure 4

Illustration of vote maximizing positions of Conservative and Labour Party leaders in a two-dimensional policy space

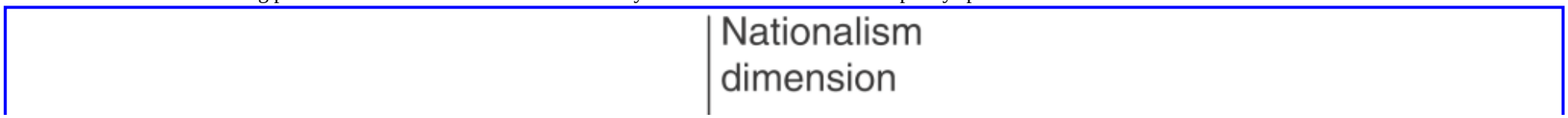

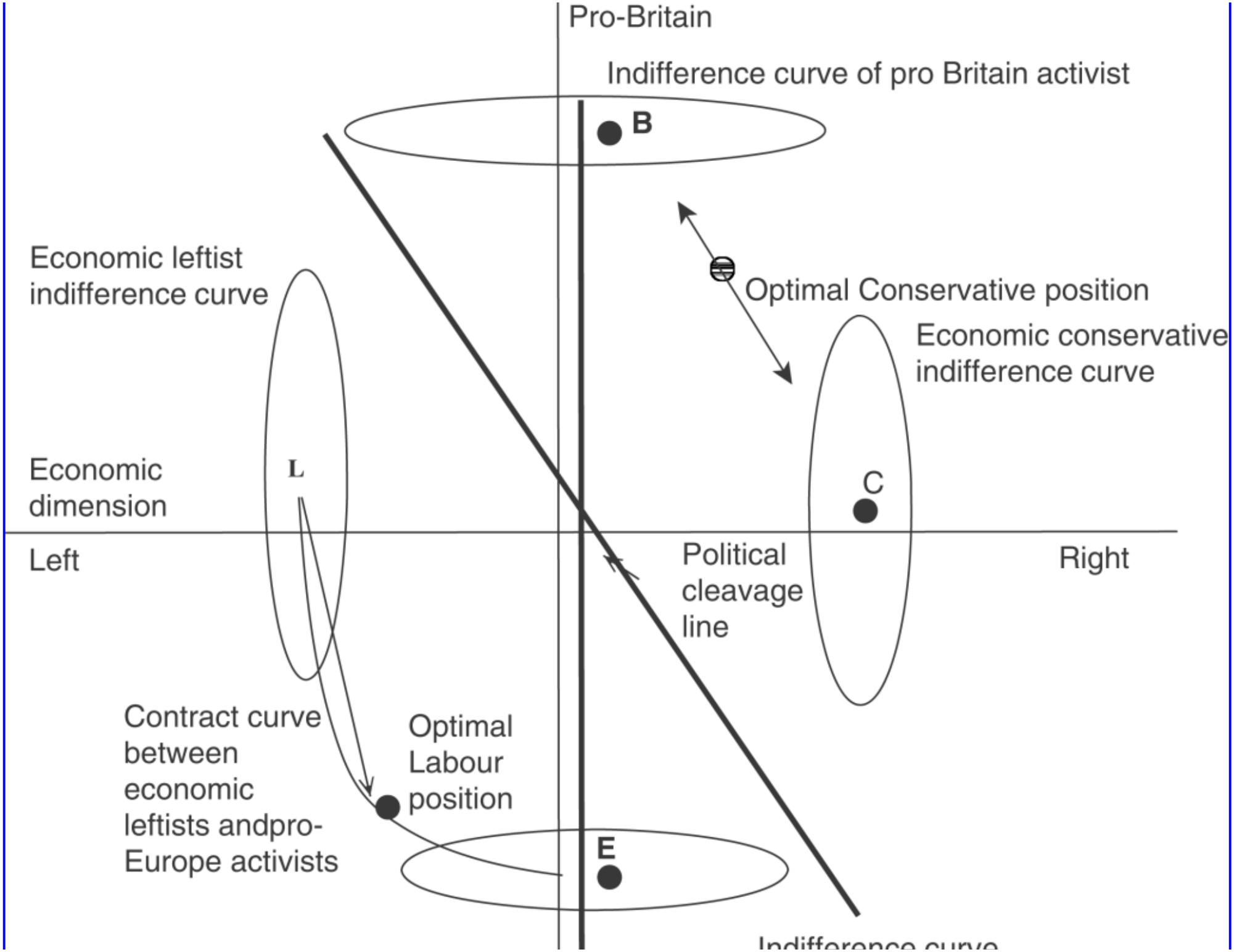
Pro-Britain
Indifference curve of pro Britain activist
B
Economic leftist indifference curve
Optimal Conservative position
Economic conservative indifference curve
Economic dimension
L
C
Left
Political cleavage line
Right
Contract curve between economic leftists andpro-Europe activists
Optimal Labour position
E

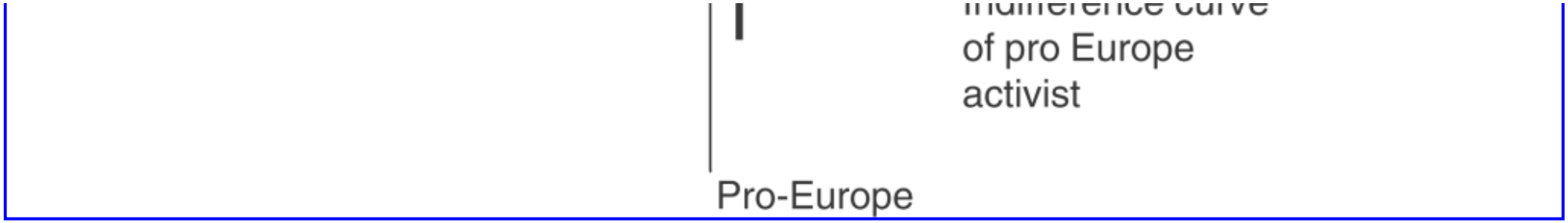

The optimal Conservative position will be determined by balancing the electoral effects of these two groups. The optimal position for this party will be one which is ‘closer’ to the locus of points that generates the greatest activist support. This locus is where the joint marginal activist pull is zero. This locus of points can be called the ‘activist contract curve’ for the Conservative Party.

Note that in Figure 4 the indifference curves of representative activists for the parties are described by ellipses. This is meant to indicate that preferences of different activists on the two dimensions may accord different saliences to the policy axes. The ‘activist contract curve’ given in the figure, for Labour, say, is the locus of points satisfying the first order. This curve represents the balance of power between Labour supporters more interested in economic issues (centred at L in the figure and those more interested in Europe (centred at E). The optimal positions for the two parties will be at appropriate positions that satisfy the optimality condition.

According to this model, a party's optimal position will tend to be nearer to the electoral origin when the valence of the party leader is higher. In contrast, a party whose leader has low valence will be more influenced by activist groups, and will tend to adopt a position further from the electoral centre and nearer to the position preferred by the dominant activist group. This model has been applied to the US polity by Miller and Schofield (2003) and Schofield, Miller and Martin (2003).

Proportional representation and plurality rule

Most of the early work in formal political theory focused on two-party competition, and generally concluded that there would be strong centripetal electoral forces causing parties to converge to the electoral centre. The extension of this theory to the multiparty context, common in European polities, has proved very difficult because of the necessity of dealing with coalition governments (Riker, 1962). However, the symmetry conditions developed by McKelvey and Schofield (1987) showed that a large, centrally located party could dominate policy if it occupied what is known as a ‘core position’. Thus, in situations where there is a stable policy core there would be certainty over the post-election policy outcome of coalition negotiation (Laver and Schofield, 1998). Absent a policy core, the post-election outcome will be a lottery across various possible coalitions, all of which are associated with differing policy outcomes and cabinet allocations. Modelling this post-election ‘committee game’ can be done with cooperative game theoretical concepts (Banks and Duggan, 2000).

Although the non-cooperative stochastic electoral model presented here can give insight into the relationship between electoral preferences and beliefs (regarding the valences of party leaders), it is still incomplete. The evidence suggests that party leaders pay attention not only to electoral responses but also to the post-election coalition consequences of their choices of policy positions. Nonetheless, the combination of the electoral model and post-election bargaining theory (Schofield and Sened, 2002) suggests the following.

In a polity based on a proportional electoral rule, the high-valence parties will be attracted towards the electoral centre. However, if there are two such competing parties of similar valence neither will locate quite at the centre. There may be many low-valence parties, whose equilibrium, vote maximizing positions will be far from the electoral centre. In order to construct winning coalitions, one or other of the high-valence parties must bargain with more ‘radical’ low-valence parties, and this could induce a degree of coalitional instability. However, it is possible that a charismatic leader, such as Sharon in Israel, can adopt a centrist position and dominate politics by controlling the policy core.

In a polity based on a plurality electoral rule, the disproportionality between votes and seats may increase the importance of activist groups. A party with a relatively low-valence leader will be forced to depend on activist support. Consequently, the party will be obliged to move to a more radical position so to attract activist support.

This may provide a reason why Britain's Labour Party appeared to acquiesce to the demands of its left-wing supporters during the leadership of Michael Foot in 1980–3 and of Neil Kinnock in 1983–92. This led to Labour defeats in the elections between 1983 and 1992. Tony Blair became Labour leader following the death of John Smith in 1994 and his high valence allowed him to overcome union opposition and to craft the centrist ‘New Labour’ policies that led to Labour victories in the elections of 1997, 2001 and 2005.

Concluding remarks

To sum up, these models suggest how the democrat paradox can be resolved: convergence to an electoral centre is not a generic phenomenon, but can occur when a party leader is generally regarded by the electorate to be of superior quality or valence. Chaos does not occur in these models, though a degree of coalitional instability is possible under proportional

electoral rule when there is no highly regarded political leader at the policy core.

See Also

- political competition
- rational behaviour
- rational choice and political science

This article is based on research supported by NSF Grant SES 024173. The table and figures are reproduced from Schofield and Sened (2006) by permission of Cambridge University Press.

Bibliography

Adams, J. 2001. *Party Competition and Responsible Party Government*. Ann Arbor: University of Michigan Press.

Adams, J. and Merrill III, S. 1999. Modeling party strategies and policy representation in multiparty elections: why are strategies so extreme? *American Journal of Political Science* 43, 765–91.

Aldrich, J. 1983a. A spatial model with party activists: implications for electoral dynamics. *Public Choice* 41, 63–100.

Aldrich, J. 1983b. A Downsian spatial model with party activists. *American Political Science Review* 77, 974–90.

Ansolabehere, S. and Snyder, J. 2000. Valence politics and equilibrium in spatial election models. *Public Choice* 103, 327–36.

Aragones, E. and Palfrey, T. 2002. Mixed equilibrium in a Downsian model with a favored candidate. *Journal of Economic Theory* 103, 131–61.

Aragones, E. and Palfrey, T. 2005. Spatial competition between two candidates of different quality: the effects of candidate ideology and private information. In *Social Choice and Strategic Decisions*, ed. D. Austen-Smith and J. Duggan. Heidelberg: Springer.

Arian, A. and Shamir, M. 1999. *The Election in Israel: 1996*. Albany: SUNY Press.

Austen-Smith, D. and Banks, J. 1999. *Positive Political Theory I*. Ann Arbor: University of Michigan Press.

Banks, J. and Duggan, J. 2000. A bargaining model of collective choice. *American Political Science Review* 94, 73–88.

Banks, J. and Duggan, J. 2005. *The theory of probabilistic voting in the spatial model of elections*. In *Social Choice and Strategic Decisions*, ed. D. Austen-Smith and J. Duggan. Heidelberg: Springer.

Banks, J., Duggan, J. and Le Breton, M. 2002. Bounds for mixed strategy equilibria and the spatial model of elections. *Journal of Economic Theory* 103, 88–105.

Caplin, A. and Nalebuff, B. 1988. On 64% majority rule. *Econometrica* 56, 787–814.

Caplin, A. and Nalebuff, B. 1991. Aggregation and social choice: a mean voter theorem. *Econometrica* 59, 1–23.

Condorcet, N. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale. Translated in part in I. McLean and F. Hewitt, *Condorcet: Foundations of Social Choice and Political Theory*. Aldershot: Edward Elgar, 1994.

Coughlin, P. 1992. *Probabilistic Voting Theory*. Cambridge: Cambridge University Press.

Dow, J. and Endersby, J. 2004. Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral Studies* 23, 107–22.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.

Enelow, J. and Hinich, M. 1984. *The Spatial Theory of Voting*. Cambridge: Cambridge University Press.

Giannetti, D. and Sened, I. 2004. Party competition and coalition formation: Italy 1994–1996. *Journal of Theoretical Politics* 16, 483–515.

Groseclose, T. 2001. A model of candidate location when one candidate has a valance advantage. *American Journal of Political Science* 45, 862–86.

Hinich, M. 1977. Equilibrium in spatial voting: the median voter result is an artifact. *Journal of Economic Theory* 16, 208–19.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39, 41–57.

Kramer, G. 1978. Existence of electoral equilibrium. In *Game Theory and Political Science*, ed. P. Ordeshook. New York: New York University Press.

Laver, M. and Schofield, N. 1998. *Multiparty Government: The Politics of Coalition in Europe*. Ann Arbor: Michigan University Press.

Lin, T.-M., Enelow, J. and Dorussen, H. 1999. Equilibrium in multicandidate probabilistic spatial voting. *Public Choice* 98, 59–82.

McKelvey, R. 1976. Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic Theory* 12, 472–82.

McKelvey, R. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47, 1085–111.

McKelvey, R. 1986. Covering, dominance and institution-free properties of social choice. *American Journal of Political Science* 30, 283–314.

McKelvey, R. and Palfrey, T. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10, 6–38.

McKelvey, R. and Patty, J. 2006. A theory of voting in large elections. *Games and Economic Behavior* 57, 155–80.

McKelvey, R. and Schofield, N. 1986. Structural instability of the core. *Journal of Mathematical Economics* 15, 179–98.

McKelvey, R. and Schofield, N. 1987. Generalized symmetry conditions at a core point. *Econometrica* 55, 923–33.

Merrill III, S. and Grofman, B. 1999. *A Unified Theory of Voting*. Cambridge: Cambridge University Press.

Miller, G. and Schofield, N. 2003. Activists and partisan realignment in the U.S. *American Political Science Review* 97, 245–60.

Penn, E. 2003. A model of far-sighted voting. Working paper, Institute of Quantitative Social Science, Harvard University.

Plott, C. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57, 787–806.

Poole, K. and Rosenthal, H. 1984. U.S. presidential elections 1968–1980: a spatial analysis. *American Journal of Political Science* 28, 283–312.

Quinn, K. and Martin, A. 2002. An integrated computational model of multiparty electoral competition. *Statistical Science* 17, 405–19.

Quinn, K., Martin, A. and Whitford, A. 1999. Voter choice in multiparty democracies. *American Journal of Political Science* 43, 1231–47.

Rakove, J., ed. 1999. *James Madison: Writings*. New York: Library of America.

Riker, W. 1962. *The Theory of Political Coalitions*. New Haven, CT: Yale University Press.

Saari, D. 1997. The generic existence of a core for q-rules. *Economic Theory* 9, 219–60.

Schofield, N. 1978. Instability of simple dynamic games. *Review of Economic Studies* 45, 575–94.

Schofield, N. 1983. Generic instability of majority rule. *Review of Economic Studies* 50, 695–705.

Schofield, N. 1984. Social equilibrium and cycles on compact sets. *Journal of Economic Theory* 33, 59–71.

Schofield, N. 1985. *Social Choice and Democracy*. Heidelberg: Springer.

Schofield, N. 2005a. A valence model of political competition in Britain: 1992–1997. *Electoral Studies* 24, 347–70.

Schofield, N. 2005b. Local political equilibria. In *Social Choice and Strategic Decisions: Essays in Honor of Jeffrey S. Banks*, ed. D. Austen-Smith and J. Duggan. Heidelberg: Springer.

Schofield, N. 2005c. The intellectual contribution of Condorcet to the founding of the US republic. *Social Choice and Welfare* 25, 303–18.

Schofield, N. 2006. *Architects of Political Change: Constitutional Quandaries and Social Choice Theory*. Cambridge: Cambridge University Press.

Schofield, N. 2007. The mean voter theorem: necessary and sufficient conditions for convergent equilibrium, *Review of Economic Studies* 74, 965–80.

Schofield, N., Martin, A., Quinn, K. and Whitford, A. 1998. Multiparty electoral competition in the Netherlands and Germany: a model based on multinomial probit. *Public Choice* 97, 257–93.

Schofield, N., Miller, G. and Martin, A. 2003. Critical elections and political realignment in the U.S.: 1860–2000. *Political Studies* 51, 217–40.

Schofield, N. and Sened, I. 2002. Local Nash equilibrium in multiparty politics. *Annals of Operations Research* 109, 193–211.

Schofield, N. and Sened, I. 2005. Multiparty competition in Israel: 1988–1996. *British Journal of Political Science* 35, 635–63.

Schofield, N. and Sened, I. 2006. *Multiparty Government: Elections and Legislative Politics*. Cambridge: Cambridge University Press.

Schofield, N., Sened, I. and Nixon, D. 1998. Nash equilibrium in multiparty competition with stochastic voters. *Annals of Operations Research* 84, 3–27.

Stokes, D. 1963. Spatial models and party competition. *American Political Science Review* 57, 368–77.

Stokes, D. 1992. Valence politics. In *Electoral Politics*, ed. D. Kavanagh. Oxford: Clarendon Press.

Strnad, J. 1985. The structure of continuous-valued neutral monotonic social functions. *Social Choice and Welfare* 2, 181–95.

Train, K. 2003. *Discrete Choice Methods for Simulation*. Cambridge: Cambridge University Press.

How to cite this article

Schofield, Norman. "democratic paradoxes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000250> doi:10.1057/9780230226203.0373

The New Palgrave Dictionary of Economics Online

demographic transition

Ronald D. Lee
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The ‘demographic transition’ refers to the fall of fertility and mortality from initially high to subsequent low levels and accompanying changes in the population. It began around 1800 with declining mortality in Europe, and is expected to be complete worldwide by 2100. In that time the global population will have risen tenfold, the ratio of elders to children will have risen by a factor of ten, longevity will have tripled, and fertility fallen from six births per woman to two. Individual and population ageing will pose many challenges, from life-cycle planning to the rising costs of health care and retirement.

Keywords

ageing populations; capital–labour ratio; demographic transition; dependency ratio; fertility in developed countries; fertility in developing countries; health care; international capital flows; international migration; life expectantcy; Malthus, T.; mortality; nutrition and development

Article

The demographic transition is the process whereby fertility and mortality move from initially high levels to subsequent low levels, with accompanying changes in the size, growth rate and age distribution of the population.

Before the start of the demographic transition, life was short, fertility was high, growth was slow, and the population was young. Declining mortality starts the typical transition, followed after a considerable lag by fertility decline (France and the United States were important exceptions to this ordering). This pattern of change causes growth rates first to accelerate and then to slow again, as population moves towards low fertility, long life and an old age structure.

The transition began around 1800 with declining mortality in Europe. It has now spread to all parts of the world and is projected to be completed by 2100. This global demographic transition has brought momentous changes, reshaping the economic and demographic life cycles of individuals and restructuring populations. Global population size increased by a factor of 6.5 between 1800 and 2000, and by 2100 will have risen by a factor of ten. There will then be 50 times as many elderly but only five times as many children: the ratio of elders to children will have risen by a factor of ten. The length of life, which has already more than doubled, will have tripled, while births per woman will have dropped from six to two. In 1800, women spent about 70 per cent of their adult years bearing and rearing young children, but that fraction has decreased in many parts of the world to only about 14 per cent due to lower fertility and longer life (Lee, 2003). These changes are sketched in Table 1.

Global population trends over the transition: estimates, guesstimates and forecasts, 1700–2100

Year	Life expectancy (years at birth)	Total fertility rate (births per woman)	Pop. Size (billions)	Pop. growth rate (%/year)	Pop. <15 (% total pop.)	Pop >65 (% total pop)
1700	27	6.0	.68	0.50	36	4
1800	27	6.0	.98	0.51	36	4
1900	30	5.2	1.65	0.56	35	4

1950 47	5.0	2.52	1.80	34	5
2000 65	2.7	6.07	1.22	30	7
2050 74	2.0	9.08	0.33	20	16
2100 81	2.0	9.46	0.04	18	21?

Sources: United Nations estimates and projections, 1900–2100; other sources for earlier years (see Lee, 2003, for details).

Before the demographic transition

According to Thomas Malthus (1798), slow population growth in the pre-industrial past was no accident. Faster population growth would depress wages, causing fertility to fall and mortality to rise due to famine, war or disease. Thus, population size was held in equilibrium with the slowly growing economy. The need to establish a separate household at marriage kept mean age at first marriage high, averaging around 25 years for women, and overall fertility low, at four to five births per woman. Mortality was moderately high, with life expectancy between 25 and 35 years. Outside of Europe and its offshoots, fertility and mortality were higher in the pre-transitional period. In India in the late 19th century, life expectancy was in the low twenties, while fertility was six or seven births per woman (Bhat, 1989). In Taiwan, the picture was similar around 1900. In the 1950s and 1960s, fertility in the less developed countries (LDCs, see UNPD, 2005, for definition) was typically six or higher.

Declining mortality

The demographic transition began first in north-west Europe, where mortality started its secular decline around 1800. In many low-income countries, the decline in mortality began in the early 20th century and then accelerated dramatically after the Second World War. The first stage of mortality decline is due to reductions in contagious and infectious diseases. Starting with the development of smallpox vaccine in the late 18th century, preventive medicine played a role in mortality decline in Europe. Public health measures were important from the late 19th century, and some quarantine measures may have been effective in earlier centuries. Improved personal hygiene also helped as the germ theory of disease became more widely known and accepted. Improving nutrition was also important in the early phases of mortality decline. Famine mortality was reduced by improvements in storage and transportation that permitted integration of regional and international food markets. Secular increases in incomes led to improved nutrition in childhood and throughout life. Better-nourished populations with stronger organ systems were better able to resist disease.

Today, the high-income countries have already largely achieved the potential mortality reductions through control of contagious disease and improved nutrition. For them, further reduction in mortality must continue to come from reductions in chronic and degenerative diseases, notably heart disease and cancer (Riley, 2001).

Most LDCs did not begin the mortality transition until the 20th century but then made rapid gains. Between 1950–4 and 2000–4, life expectancy in LDCs has increased from 41.1 years to 63.4, with average gains of 0.45 years of life per calendar year. Such rapid rates of increase in low-income countries will surely taper off as mortality levels approach those of the more developed countries (MDCs), whose gains have been less than half as rapid at 0.19 years per year. There are notable exceptions to this generally favourable picture. In sub-Saharan Africa, life expectancy has been declining since the early 1980s, largely due to HIV/AIDS. In the southern African region, life expectancy dropped from 62 to 48 between the early 1990s and the early 2000s. On average, eastern European (including the former USSR) life expectancy is lower now than it was in the late 1960s (UNPD, 2005).

How far and how fast will mortality fall and life expectancy rise in the 21st century? Methods that extrapolate historical trends in mortality by age suggest greater longevity gains than MDC government actuaries typically project, but past official projections have under-predicted subsequent gains, particularly at the older ages. Some experts argue that we are approaching biological limits and that these historical trends cannot be expected to continue; they foresee an upper limit of 85 years for life expectancy. Others, impressed by advances in genetic and stem cell research, foresee much more rapid gains for the future than occurred in the past.

Fertility transition

Most economic theories of fertility start with the idea that couples wish to have some number of surviving children rather than a number of births per se. On this assumption, once potential parents recognize an exogenous increase in child survival, fertility should decline. However, mortality and fertility interact in complicated ways. For example, increased survival raises the return on post-birth investments in children, while some of the improvement in child survival is itself a response to parental decisions to invest more in the health and welfare of a smaller number of children. Nonetheless, it is very likely that mortality decline has exerted an important independent influence on fertility decline.

Economic change also influences fertility by altering the costs and benefits of childbearing and rearing, which are time-intensive. Technological progress and increasing physical and human capital make labour more productive, raising the value of time in all activities and making children increasingly costly relative to consumption goods. Since women have had

primary responsibility for childbearing and rearing, variations in the productivity of women have been particularly important. For example, physical capital may substitute for human strength, reducing or eliminating the productivity differential between male and female labour, and thus raising the opportunity cost of children. Rising incomes have shifted consumption demand towards non-agricultural goods and services, for which educated labour is a more important input. A rise in the rate of return to education then leads to increased investments in education. Overall, these patterns have several effects: children become more expensive, their economic contributions are diminished by school time, and educated parents have higher value of time, which raises the opportunity costs of childrearing. Furthermore, parents with higher incomes choose to devote more resources to each child, and, since this raises the cost of each child, it also leads to fewer children. Developing markets and governments replace many economic functions of the traditional family and household, to which children contributed, further weakening the value of children.

The importance of contraceptive technology for fertility decline in the past and future is hotly debated, with many economists viewing it as of relatively little importance (Pritchett, 1994). The European fertility transition, for example, was achieved using *coitus interruptus*, a widely known traditional method requiring no modern technology.

Between 1890 and 1920, fertility within marriage began to decline in most European provinces, with a median decline of about 40 per cent from 1870 to 1930 (Coale and Watkins, 1986). The fertility transition in the MDCs largely occurred before the Second World War. After the war, many of these countries experienced baby booms and busts, followed by the 'second fertility transition' as fertility fell far below replacement level, marriage rates fell, and increasing proportions of births occurred outside marriage. Many LDCs began the fertility transition in the mid-1960s, and these later transitions have typically been more rapid than earlier ones, with fertility reaching replacement level (around 2.1 births per woman) within 20 to 30 years after onset. Fertility transitions in East Asia have been particularly early and rapid, while those in South Asia and Latin America have been slower (Bulatao and Casterline, 2001). The transition in sub-Saharan Africa started from a higher initial level of fertility and began later. By now, almost all countries have begun the fertility transition (UNPD, 2005; Bulatao and Casterline, 2001).

Currently, 66 countries with 44 per cent of the world's population have fertility at or below replacement level. Of these, 43 are MDCs, but 23 are LDCs. Average fertility in the MDCs is 1.56 births per woman, and in many it has fallen below 1.3. Many LDCs, particularly in East Asia, also have fertility far below replacement. It is not yet clear whether fertility will fall farther, rebound towards replacement, or stay at current levels.

Age at first marriage and first birth are generally moving to older ages throughout the industrial world and much of the developing world as well. This depresses the total fertility rate, which is a synthetic cohort measure, by 10–40 per cent below the underlying completed fertilities of generations. When the average age of childbearing stops rising, the total fertility rate should increase to this underlying level.

Population growth

A steady state population growth rate for a hypothetical zero migration population can be associated with each level of fertility and life expectancy, as depicted in Figure 1.

Figure 1

Life expectancy and total fertility rate with population growth isoquants: past and projected trajectories for more, less, and least developed countries. *Source*: Bhat (1989); UNPD (2005); see Lee (2003) for further details.

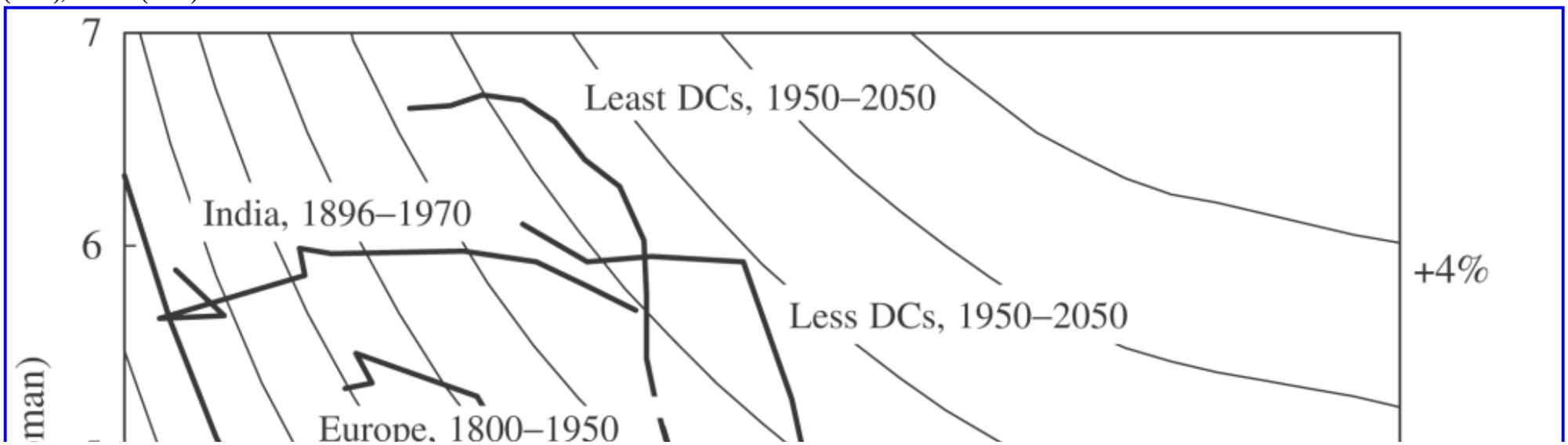

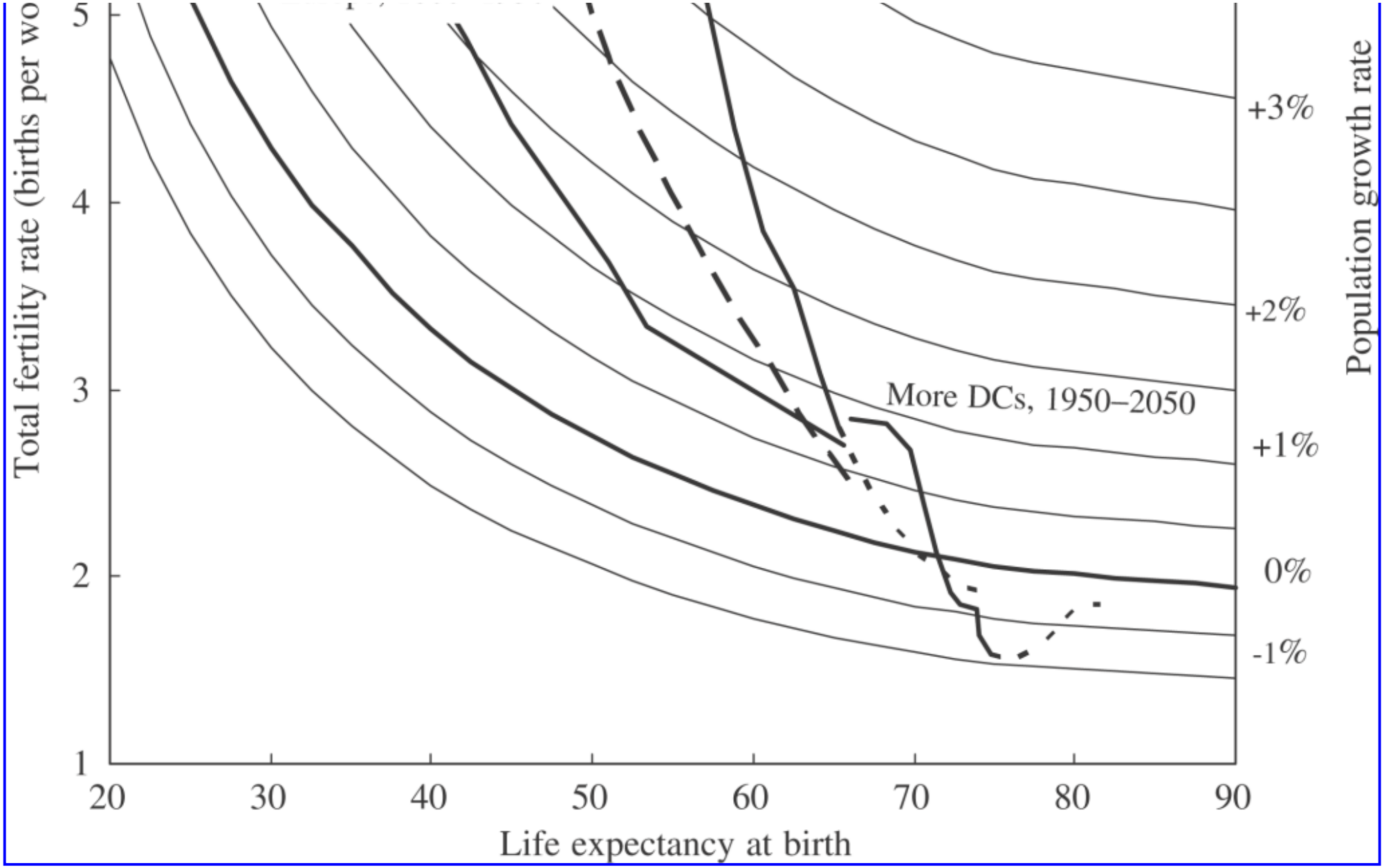

Figure 1 plots growth rate contours or isoquants. These differ from actual growth rates due to net migration and the transitory influence of age distribution. In this figure, a demographic transition begins as a move to the right, representing a gain in life expectancy with little change in fertility and therefore movement to a higher population growth contour. Next, a diagonal downward movement to the right reflects a simultaneous decline in fertility and mortality, recrossing contours towards lower rates of growth and perhaps going negative, as do the MDCs. Historical data are extended using UNPD (2005) data and projections, by development status.

India, shown separately, had higher initial fertility and mortality than Europe, as did the least developed countries relative to the LDCs in 1950, which in turn had far higher mortality and fertility than the MDCs in that year. In all cases, the initial path is horizontally to the right, indicating that mortality decline preceded fertility decline, causing accelerating population growth approaching three per cent for the LDCs and least developed countries. Europe briefly attains 1.5 per cent steady state population growth, but then fertility plunges, a decline picked up after 1950 by the group of LDCs, ending with population decline at 1 per cent annually (the actual European population growth rate is slightly higher than this

hypothetical steady state one due to age distribution and immigration). All three groups are projected by the UN to approach the zero-growth contour by 2050.

Historical and projected population growth rates, as opposed to hypothetical steady state ones, can be seen over a longer time period in Figure 2. Growth rates in the MDCs rose about a half of one per cent above those in the LDCs in the century before 1950. But after the Second World War, population growth surged in the LDCs, with the growth rate peaking at 2.5 per cent in the mid-1960s, then dropping rapidly. The population share of the MDCs is projected to drop from its current 20 per cent to only 13 per cent in 2050. Long-term United Nations projections suggest that global population growth will be close to zero by about 2100. The projection for the MDC population is nearly flat, with population decrease in Europe and Japan offset by population increase in the United States and other areas.

Figure 2

Population growth rates, 1750–2150. *Source*: UNPD (2005); see Lee (2003) for further details.

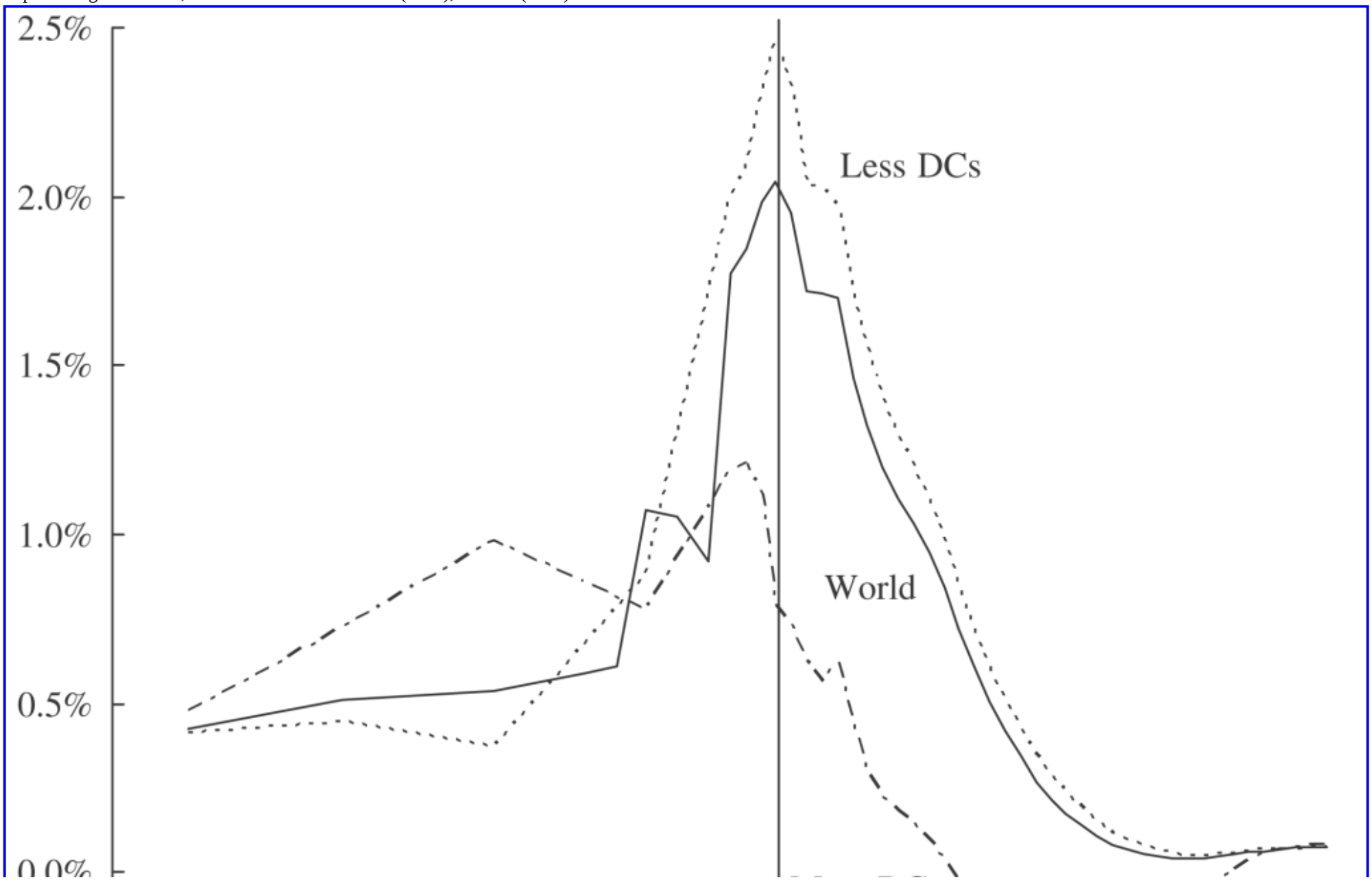

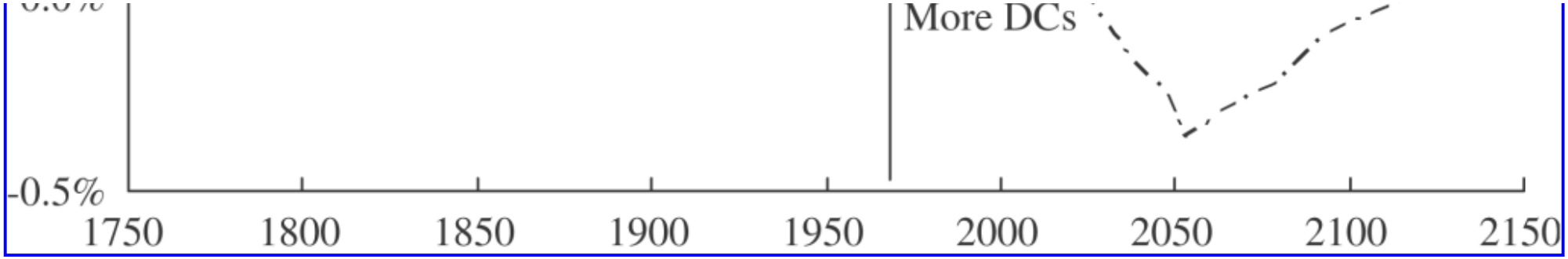

Changing age distribution over the demographic transition

Figure 3 plots the changes in age distribution that accompany a classic demographic transition, using historical data from India from 1896 to 2000 (stars) and United Nations projections through 2050 (hollow circles). These data are superimposed on a stylized transition that was simulated with the use of mathematical functions for the trajectories of fertility and life expectancy. Simulated fertility starts close to six births per woman and ends at 2.1. Life expectancy starts at 24 years and ends at 80. Mortality decline starts in 1900, 50 years before the fertility decline begins in 1950. The Indian fertility transition is slower than that of East Asia but similar to that of Latin America.

Figure 3

Changing age distribution over a classic demographic transition: actual and projected dependency ratios for India and simulations, 1900–2100. *Source*: Actual India data for the period 1891–1901 to 1941–51 are taken from Bhat (1989). Other actual and projected data are taken from UNPD (2003). *Note*: Lines indicate a simulated demographic transition superimposed over actual (*) and projected (o) dependency ratios for India.

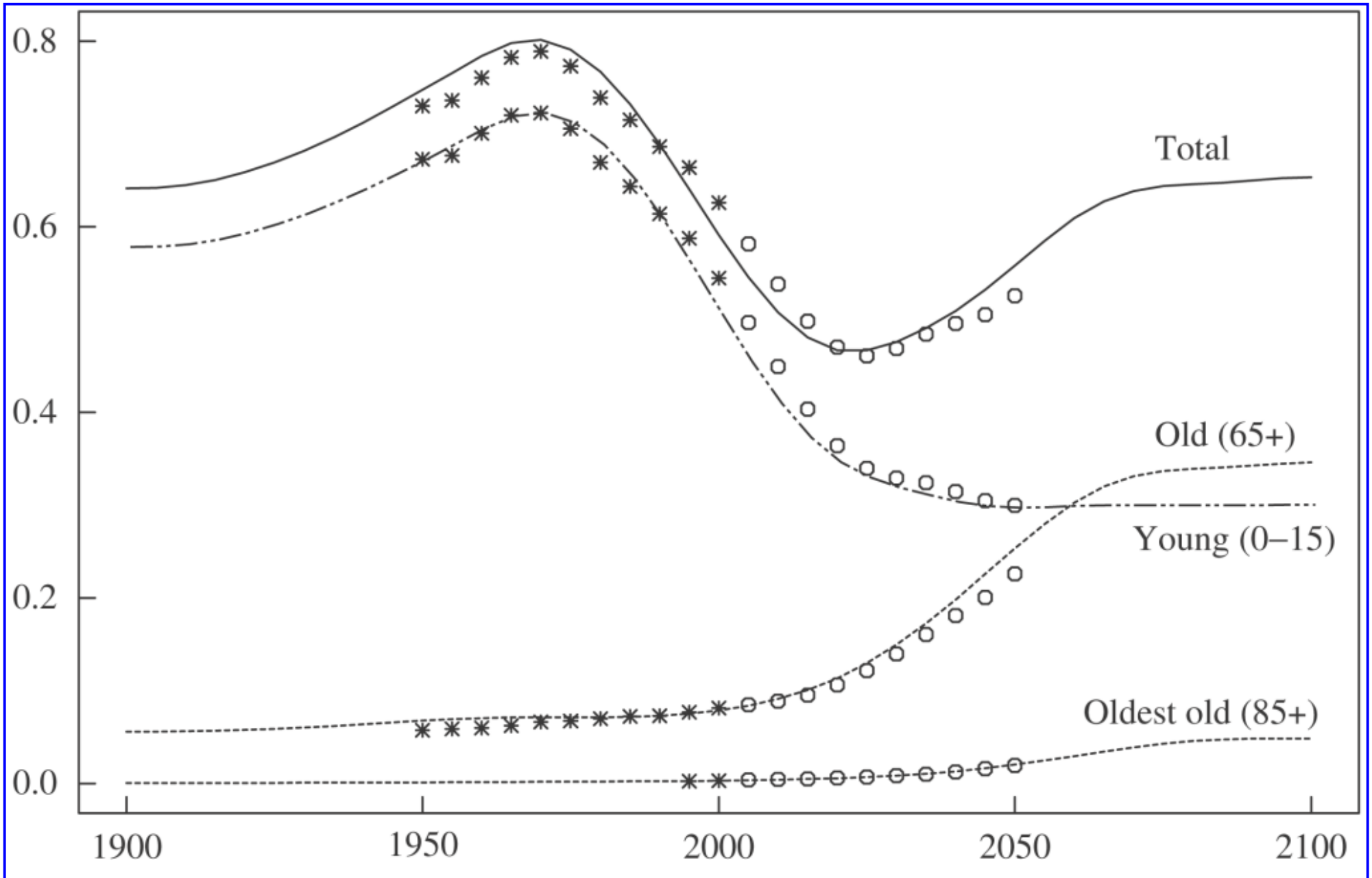

Note: Lines indicate a simulated demographic transition superimposed over actual (*) and projected (o) dependency ratios for India.

The distinctive changes in the age distribution can be seen in the ‘dependency ratios’, which take either the younger or the older population and divide it by the working-age population. The initial mortality decline, while fertility remains high, raises the proportion of surviving children in the population, as reflected in the increasing child dependency ratios. Counter-intuitively, mortality decline initially makes populations younger rather than older in a phase which here lasts 70 years. Families find themselves with increasing

numbers of surviving children, and both families and governments may struggle to achieve human capital investment goals for the unexpectedly high number of children.

Next, as fertility declines, child dependency ratios decline and soon fall below their pre-transition levels. The working age population grows faster than the population as a whole, so the total dependency ratio declines. This second phase may last 40 or 50 years. Some analysts have worried that the rapidly growing labour force in this phase might cause rising unemployment and falling capital–labour ratios, while others have stressed the advantages of this phase, calling these a demographic gift or bonus. Figure 3 shows that in India the bonus occurs between 1970 and 2015, when the total dependency rate is declining. The decline in dependents per worker would by itself raise per capita income by 22 per cent, other things equal, adding 0.5 per cent per year to per capita income growth over the 45-year span.

In a third phase, increasing longevity leads to a rapid increase in the elderly population while low fertility slows the growth of the working age population. The old-age dependency ratio rises rapidly, as does the total dependency ratio. In India, population ageing will occur between 2015 and 2060. If the elderly are supported by transfers, either from their adult children or a public-sector pension system supported by current tax revenues, then the higher total dependency ratio means a greater burden on the working-age population. However, to the extent that the elderly prepare for their retirement by saving and accumulating assets earlier in their lives and then dissave in retirement, population ageing may cause lower aggregate saving rates, as life-cycle savings models and some empirical analyses suggest. But even with lower savings rates the capital–labour ratio may rise, since the labour force is growing more slowly. The net effect would then be to stimulate growth in labour productivity due to capital deepening.

At the end of the full transitional process, the total dependency ratio is back near its level before the transition began, but now child dependency is low and old-age dependency is high. Presumably mortality will continue to decline in the 21st century, and the process of individual and population ageing will continue. No country in the world has yet completed this phase of population ageing, since even the industrial countries are projected to age rapidly over the next three or four decades. In this sense, no country has yet completed its demographic transition.

Population ageing is due both to low fertility and to long life. Low fertility raises the ratio of elderly to working-age people, with no corresponding improvement in health to facilitate a prolongation of working years. For this reason, it imposes important resource costs on the population, regardless of institutional arrangements for old-age support. Lower total expenditures on children and increased capital per worker will offset these costs.

By contrast, population ageing due to declining mortality is generally associated with increasing health and vitality of the elderly. Such ageing may put pressure on pension programmes that have rigid retirement ages, but this problem is a curable institutional one, since the ratio of healthy, vigorous years over the life cycle to frail or disabled years has not changed, and individuals can adjust by keeping the fractions of their adult life spent working and retired constant, for example.

Some consequences of the demographic transition

The three centuries of demographic transition from 1800 to 2100 will reshape the world's population in a number of ways. Population will rise from 1 billion in 1800 to 9.5 billion in 2100. The average length of life will increase by a factor of two or three, fertility will have declined by two-thirds, and the median age of the population will double from the low twenties to the low forties. The population of Europe will decline by ten per cent between 2005 and 2050, and its share of world population will have declined by two-thirds since 1950. But many other changes will also have been set in train in family structure, health, institutions for saving and supporting retirement, and even in international flows of people and capital.

At the level of families, as the number of children born declines sharply, childbearing becomes concentrated into only a few years of a woman's life; combined with greater longevity, this means that many more adult years are available for other activities. Parents with fewer children are able to invest more in each child, reflecting the quality–quantity trade-off, which may also be one of the reasons parents reduced their fertility.

The processes which lead to longer life also alter the health status of the surviving population. For the United States, it appears that years of healthy life are growing roughly as fast as total life expectancy. In other industrial populations the story is more mixed. Trends in health, vitality and disability are of enormous importance for human welfare.

The economic pressures on pension programmes caused by the increasing proportion of elderly are exacerbated in the MDCs by dramatic declines in the age at retirement, which for US men fell from 74 in 1910 to 63 in 2000. Population ageing will also generate intense financial pressures on publicly funded systems for health care and for long-term care.

At the international level, the flow of people and capital across borders may offset these demographic pressures. As population growth has slowed or even turned negative in the MDCs, it is not surprising that international migration from Third World countries has accelerated. Net international migration to the MDCs experienced a roughly linear increase from near-zero in the early 1950s to around 2.6 million per year in the 1990s. Of course, these net numbers for large population aggregates conceal a great deal of offsetting international gross migration flows within and between regions (UNPD, 2005). For example, prior to 1970 Europe was a net sending region, but from 1970 to 2000 it received 18 million net immigrants. During the 1990s, repatriation of African refugees reversed the net flows from the least developed countries. But overall, while MDCs may seek to alleviate their population ageing through immigration, United Nations simulations indicate that the effect will be only modest, since immigrants also grow old, and their fertility converges to levels in the receiving country.

Might international flows of capital cushion the financial effects of population ageing? Population ageing may cause declining aggregate saving rates, but, with slowing labour force growth, capital–labour ratios will probably rise and profit rates fall, particularly if there is a move towards funded pensions. Capital flows from the MDCs into the LDCs might help

keep the rate of return on investments from falling, but the possibilities are limited by the much smaller size of Third World economies.
Dramatic population ageing is the inevitable final stage of the global demographic transition, and it will bring serious economic and political challenges. Meeting these challenges will require flexible institutional structures, adjustments in life-cycle planning, and a willingness to pay for rising costs of health care and retirement.

See Also

- fertility in developed countries
- historical demography
- international migration
- mortality
- population ageing

Bibliography

Bhat, P. 1989. Mortality and fertility in India, 1881–1961: a reassessment. In *India's Historical Demography: Studies in Famine, Disease and Society*, ed. T. Dyson. London: Curzon Press.

Bulatao, R. and Casterline, J., eds. 2001. *Global Fertility Transition*, a supplement to *Population and Development Review* 27.

Coale, A. and Watkins, S., eds. 1986. *The Decline of Fertility in Europe*. Princeton: Princeton University Press.

Lee, R. 2003. The demographic transition: three centuries of fundamental change. *Journal of Economic Perspectives* 17(4), 167–90.

Malthus, T. 1798. *An Essay on the Principle of Population*, ed. D. Winch and P. James. Cambridge: Cambridge University Press, 1992.

Pritchett, L. 1994. Desired fertility and the impact of population policies. *Population and Development Review* 20, 1–55.

Riley, J. 2001. *Rising Life Expectancy: A Global History*. Cambridge: Cambridge University Press.

UNPD (United Nations Population Division). 2003. *World Population Prospects: The 2002 Revision*. New York: United Nations.

UNPD. 2005. *World Population Prospects: The 2004 Revision*. New York: United Nations.

How to cite this article

Lee, Ronald D. "demographic transition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000074>
doi:10.1057/9780230226203.0374

The New Palgrave Dictionary of Economics Online

demography of the ancient world

Walter Scheidel

From The New Palgrave Dictionary of Economics Online, ,
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This survey of demographic conditions in ancient Greek and Roman history discusses life expectancy and causes of death, reproduction and fertility control, marriage practices and household structure, population size and its change over time, and the relationship between demographic and economic development.

Keywords

ancient world; birth rates; death rates; demographics; family; Greece; life expectancy; marriage; population; Rome

Article

Ancient demography covers the population history of early civilizations from the third millennium BCE to the seventh century CE.
Due to the uneven distribution of relevant evidence, scholars have focused primarily on Middle Eastern, Greek and Roman populations. This survey deals primarily with the demography of the Greco-Roman world. Demographic conditions in antiquity are generally only dimly perceptible, and attempts to reconstruct them inevitably entail considerable uncertainty and conjecture. Information is provided by tombstone inscriptions, census documents on papyri, skeletal remains and literary accounts.
Ancient birth and death rates were extremely high by modern standards. Mean life expectancy at birth is commonly estimated to have been around 20 to 30 years. The distribution of ages recorded in census returns from Roman Egypt is consistent with model life tables that posit a mean life expectancy at birth of 22 to 25 years. This estimate receives additional support from a variety of other data samples including funerary inscriptions from Roman North Africa, a Roman schedule used to calculate annuities known as ‘Ulpian's Life Table’, and the age structure of a few cemetery populations. Roman emperors who died of natural causes had a similarly low life expectancy. This suggests that socio-economic standing had little effect on longevity. Mortality regimes were highly localized and determined primarily

by the prevalence of particular infectious diseases. In parts of the Roman empire, seasonal variation in mortality can sometimes be reconstructed with the help of dates of death reported in tombstone inscriptions. These seasonality patterns also allow inferences about the underlying disease environments. According to these datasets, seasonal spikes in adult death rates were much stronger than in the more recent past, suggesting that even the most resilient age groups were susceptible to fatal infections. The main causes of death can be inferred from ancient medical texts and literary sources. Gastro-intestinal diseases, malaria and tuberculosis were particularly important. Both malaria and leprosy expanded during the Greco-Roman period. Smallpox epidemics occurred, possibly in Athens in 430 bce and probably throughout the Roman empire in the 160/180•s ce. Plague spread from 540 to 750 ce in a pandemic that foreshadowed the medieval Black Death.

High levels of mortality required correspondingly high birth rates. The average woman surviving to menopause had to give birth to five or six children to ensure reproduction at replacement level. Birth rates within marriage were higher still: it has been calculated that in Roman Egypt, a woman who had been continuously married between menarche and menopause would on average have given birth eight or nine times. According to census records from the same region, 95 per cent of freeborn children were born to married parents. These documents also allow us to reconstruct the maternal age distribution of childbirths, which implies what is known as a ‘natural fertility’ regime in which fecundity was a direct function of a woman's age, peaking around age 20 and gradually declining over time. Signs of stopping behavior – that is, the cessation of reproduction in response to family size or composition – are absent from these data.

At the same time, early and near-universal marriage for women and high birth rates went hand in hand with fertility control within marriage. Census returns from Roman Egypt indicate mean birth intervals of three to four years. Birth-spacing may have been achieved by prolonged breastfeeding or by other means. Ancient medical texts discuss a variety of putative contraceptives and abortifacients. More drastic intervention in the form of child exposure and infanticide was often socially condoned, although the actual scale of these practices remains unknown. The extent to which parents discriminated against female offspring is particularly controversial. While Greek and Roman sources sometimes refer to femicide, and evidence of male-biased sex ratios has been taken to reflect this custom, we are usually unable to determine whether ‘missing’ females had been killed or exposed after birth or were merely omitted from written records.

Among Greeks and Romans, (serial) monogamy was the norm. Polygamous unions were largely confined to ruling and elite families in Middle Eastern societies. At the same time, sexual access to slave women facilitated resource polygyny even in formally monogamous settings. In ancient Greek culture, women often appear to have been married off in their mid-teens while men took wives considerably later, around the age of 30. Funerary inscriptions from the western half of the Roman empire point to typical marriage ages of about 20 years for women and 30 years for men. Roman aristocrats generally married at younger ages. For Roman Egypt, the census records reflect mean marriage ages of 17 or 18 years for women and 25 years for men. They also show that whereas almost all women had married by their late 20•s, it was only by age 50 that most men had married at least once. This pattern of moderately early female and late male marriage resembles the so-called ‘Mediterranean marriage pattern’ which prevailed in the more recent past, suggesting a measure of long-term continuity in that region. Divorce could be initiated by both husbands and wives, and commonly lacked strong stigma. Remarriage was much more common for men than for women, especially after age 30: according to the Roman Egyptian

census returns, two-thirds of men but only one-third of women were still married at the age of 50. In the pre-Christian period, celibacy was not normally considered desirable.

Marriages were mostly virilocal or neolocal. Bridal dowries were common but are best attested for elite circles. Slaves could not legally marry but were able to enter informal unions, primarily (but not only) with other slaves. Consanguineous unions were more widespread in the eastern Mediterranean, and especially in the Middle East, than in western Europe. Thus, first-cousin marriage occurred mostly in the East, and occasional half-sibling unions are known from the Greek world. Scholars still debate whether references to married couples of full siblings found in Roman Egyptian census documents reveal genuinely incestuous unions or record the unions of cousins who had legally become siblings through adoption. However, instances of brother–sister and parent–child marriage are credibly attested for ancient Middle Eastern rulers, and more generally for members of the Zoroastrian community in Mesopotamia and Iran.

The Greek and Latin languages lack specific terms for what we would call the nuclear family. Notions of family and household were more inclusive: next to parents and their offspring, the Greek *oikos* and the Roman *familia* or *domus* routinely encompassed co-resident kin and slaves. At the same time, Roman funerary inscriptions tend to privilege commemorative ties within the nuclear family, showing it to have been the principal locus of familial sentiment and obligation, of inheritance, and probably also of residence. More complex households were common in the eastern Mediterranean and the Middle East. In Roman Egypt, for example, the majority of the rural population belonged to households comprised of extended or multiple families. High death rates offset high fertility, thereby limiting family size, which averaged 4.3 in the same region.

Owing to unpredictable mortality and the desire to preserve male lineages, adoption of relatives appears to have been common. Partible inheritance rather than primogeniture was the norm. Daughters either received dowries as a substitute for an inheritance or inherited alongside their brothers. The social effects of high death rates undermined the formally patriarchal character of ancient households. A significant share of Greeks and Romans must have lost their fathers as minors and were assigned guardians, while many widows were unable to remarry. For these reasons, family units in which women and children were under the control of fathers and husbands were less common and more fragile than modern observers have often imagined.

Population numbers are very poorly known and continue to generate controversy. Statistical documents survive only from parts of Egypt, and literary references to population size are commonly vitiated by rhetorical stylization, ignorance or indifference. Archaeological data help to fill this gap but pose their own problems of interpretation. What we do know is that the Mediterranean regions and its hinterlands underwent significant population growth in the Greco-Roman period. In the Aegean, the collapse of late Bronze Age civilization around 1200 bce coincided with strong demographic contraction. Population recovered from the early first millennium bce onward and peaked in the classical period, in the fifth and fourth centuries bce, when Greece may have been more densely populated than at any other time prior to the 20th century.

During this growth phase, Greek settlers established hundreds of colonies in Sicily, South Italy and the Black Sea littoral. By the fourth century bce, up to 1,000 Greek city-states were inhabited by 7 million people or more. Most of these communities were very small. The conquests of Alexander the Great in the late fourth century bce triggered Greek emigration to Egypt, Syria, Mesopotamia, Iran and Central Asia. Large-scale state formation under his successors led to the creation of capital cities in excess of

100,000 residents, most notably Alexandria in Egypt. Meanwhile, populations expanded farther west in Italy, where this process drove the conflicts that eventually resulted in Roman regional hegemony, and more generally in western Europe and the Maghreb. A series of Roman census tallies from the last three centuries bce offers insight into demographic change on the Italian peninsula. Even so, the total size of its population cannot be established with precision: depending on different interpretations of the extant census counts, by the beginning of the Common Era Italy may have been inhabited by no more than 6 million people (including slaves) or by two or even three times as many. These uncertainties interfere with modern assessments of Roman economic performance.

For a variety of reasons, an estimate between the extreme ends of spectrum seems appropriate: with a peak population of perhaps 10 or 12 million people, Roman Italy may well have matched the population densities of the high medieval and early modern periods. The population of the Roman empire as a whole is necessarily even more difficult to ascertain: while a total of 60 to 80 million seems realistic, a higher figure cannot entirely be ruled out. Maybe 10 to 20 per cent of these individuals lived in some 2,000 cities. The capital city of Rome appears to have grown to a million residents, an urban population unparalleled in Europe prior to London around 1800. Starting in the late second century ce, epidemics reduced population numbers, although settlement densities remained high into late antiquity. Massive population losses finally accompanied the disintegration of the western half of the Roman empire in the fifth century ce and the onset of recurrent plague pandemics in the 540•s ce.

Despite its overall paucity and numerous shortcomings, demographic information from the Greco-Roman world is of considerable relevance to our understanding of ancient economic history. Centuries of continuous population growth, first in the eastern Mediterranean and later farther west, highlight the scale and persistence of an economic expansion which was driven by the spread of farming, technological innovation and gains from trade. Concurrent urbanization reinforces our impression of dynamic economic development. In the long run, however, ancient economies resembled other premodern economies in their inability to overcome Malthusian pressures through ongoing technological innovation. As population continued to expand, per capita economic growth eventually abated, first in Greece and later in the western Mediterranean. Judging by a variety of archaeological proxies of economic performance, by the time exogenous shocks in the form of plague and invasions began to affect the Roman empire in the second and third centuries ce the economy had already ceased to grow in real terms.

There is no indication that the Greco-Roman economic-demographic expansion significantly improved health or longevity: accretions to the stock of knowledge proved insufficient to mitigate the impact of the main causes of death, and potential gains from infrastructural provisions (such as aqueducts) may well have been offset by the demographic burden of urbanization and rising population densities which increased exposure to infection. In a number of skeletal samples, average body height was smaller in the Roman period than both immediately before and after, which likewise speaks against the notion of improvements in physiological well-being. Moreover, widespread skeletal evidence of deficiency diseases points to pervasive morbidity which would have curtailed productivity. High death rates discourage investment in education and impede human capital formation. Correspondingly high fertility depresses female labour participation and the status of women. In this environment, sustainable economic growth, let alone a fertility transition, was not feasible.

See Also

- economic history
- historical demography
- population dynamics
- population health, economic implications of

Bibliography

Bagnall, R.S. and Frier, B.W. 1994. *The Demography of Roman Egypt*. Cambridge: Cambridge University Press.

Brunt, P.A. 1971. *Italian Manpower 225bc–ad 14*. Oxford: Clarendon Press.

Frier, B.W. 1994. Natural fertility and family limitation in Roman marriage. *Classical Philology* 89, 318–33.

Frier, B.W. 2000. Demography. In *The Cambridge Ancient History Volume 11*, ed. A.K. Bowman, P. Garnsey and D. Rathbone. Cambridge: Cambridge University Press.

Hansen, M.H. 2006. *The Shotgun Method: The Demography of the Ancient Greek City-State Culture*. Columbia: University of Missouri Press.

Parkin, T.G. 1992. *Demography and Roman Society*. Baltimore: Johns Hopkins University Press.

Pomeroy, S.B. 1997. *Families in Classical and Hellenistic Greece: Representations and Realities*. Oxford: Clarendon Press.

Sallares, R. 1991. *The Ecology of the Ancient Greek World*. London: Duckworth.

Saller, R.P. 1994. *Patriarchy, Property and Death in the Roman Family*. Cambridge: Cambridge University Press.

Scheidel, W. 2001a. *Death on the Nile: Disease and the Demography of Roman Egypt*. Leiden: Brill.

Scheidel, W. 2001b. Progress and problems in Roman demography. In *Debating Roman Demography*, ed. W. Scheidel. Leiden: Brill.

Scheidel, W. 2007. Demography. In *The Cambridge Economic History of the Greco-Roman World*, ed. W. Scheidel, I. Morris and R. Saller. Cambridge: Cambridge University Press.

Scheidel, W. 2008. Roman population size: the logic of the debate. In *People, Land and Politics:*

Demographic Developments and the Transformation of Roman Italy 300 bc–ad 14, ed. L. De Ligt and S. Northwood. Leiden: Brill.

How to cite this article

Scheidel, Walter. "demography of the ancient world." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000270> doi:10.1057/9780230226203.1875

The New Palgrave Dictionary of Economics Online

Denison, Edward (1915 – 1992)

Barry Bosworth

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; capacity utilization; capital accumulation; Denison, E.; growth accounting; national income accounting; production theory; standardized system of national accounts (SNA); Stone, J.R.N.; total factor productivity

Article

Edward Denison was a major contributor to the development of the US national income accounts and one of the originators of growth accounting. He received a Ph.D. in economics from Brown University in 1941. Denison's early career (1941–56) was spent in the national income division of the US Commerce Department where he worked with Milton Gilbert, George Jaszi, and Charles Schwartz to develop the national accounts of the United States. The United States had published estimates of national income and its components in 1934; and Richard Stone and others developed both expenditure and income-side estimates of GNP for the United Kingdom that were published in 1941. The US expenditure-side estimates were first published in 1942.

Denison participated in a 1944 tripartite meeting with Canada, the United Kingdom, and the United States that worked to establish consensus on a set of concepts and methods for the national accounts. That meeting and subsequent work provided much of the basis for the standardized system of national accounts (SNA) that was adopted and expanded by the United Nations and the OECD. The United States did not initially adopt the SNA; but by 2000 it was following the SNA in all of its important respects.

Denison moved to the Committee on Economic Development (CED) in 1956 where his research focused on identifying the sources of economic growth. In expanding the framework of growth accounting, Denison sought to go beyond a simple partitioning of economic growth into the contributions of the factor inputs and a residual of total factor production. He incorporated changes in the quality of the inputs, such as job skills, economies of scale, and other contributors to the residual, such as research and development. His initial analysis was published by the CED in 1962 as *The Sources of Economic Growth in the United States and the Alternatives Before Us*. A distinctive feature of his approach was the extent to which he anchored it in the basic accounting framework of the national accounts rather than

the concepts of neoclassical production theory employed a few years later by Jorgenson and Griliches (1967). This aspect made it easy for other researchers to duplicate his methodology within their own countries.

Denison moved to the Brookings Institution in 1963 and extended his analysis to international comparisons with publication of *Why Growth Rates Differ* (1967). Two important later contributions were *How Japan's Economy Grew So Fast* (with W.K. Chung, 1976) and *Accounting for Slower Economic Growth: The United States in the 1970s* (1979). In *Accounting for Slower Growth*, he explored a wide range of popular explanations for the productivity slowdown, including higher energy prices, government regulation, and reduced R&D expenditures, and argued that their effects were too small to account for the magnitude and persistence of the slowdown. He received the Distinguished Fellow Award of the American Economic Association in 1981.

Denison's exchanges with Jorgenson and Griliches (1967; 1972a), while centred around differences in their approaches to measuring the contributions to growth, served to highlight an ongoing debate about the relative importance of capital accumulation and total factor productivity gains. Denison's approach, by minimizing several aspects of the measurement of capital's contribution, tended to support the conventional wisdom of the time that TFP accounted for a substantial portion of growth. Jorgenson and Griliches were attempting to argue that careful measurement of the factor inputs could drastically shrink the residual contribution of TFP. Denison won out on the issue of the relative importance of TFP by pointing to some problems with Jorgenson–Griliches adjustment for variations in capacity utilization; but the longer-term value of the debate was in showing that their approaches were quite similar. In subsequent years, the Jorgenson–Griliches approach, with its anchor in production theory, has dominated the conceptual discussion. However, many of the empirical studies continue to follow Denison's careful use of national income accounts data.

See Also

- economic growth
- growth accounting
- total factor productivity

Selected works

1962. *The Sources of Economic Growth in the United States and the Alternatives Before Us*. New York: Committee for Economic Development.

1967. *Why Growth Rates Differ*? Washington, DC: Brookings Institution.

1969. Some major issues in productivity analysis: an examination of estimates by Jorgenson and Griliches. *Survey of Current Business* 49(5, Part 2), 1–27.

1972. Final comments. *The Measurement of Productivity*. Washington, DC: Brookings Institution.

1979. *Accounting for Slower Economic Growth: The United States in the 1970s*. Washington, DC: Brookings Institution.

1976. (With W. K. Chung.) *How Japan's Economy Grew So Fast: The Sources of Postwar Expansion*. Washington, DC: Brookings Institution.

Bibliography

Jorgenson, D. and Griliches, Z. 1967. The explanation of productivity change. *Review of Economic Studies* 34, 249–83.

Jorgenson, D. and Griliches, Z. 1972a. Issues in growth accounting: a reply to Edward F. Denison. *The Measurement of Productivity*. Washington, DC: Brookings Institution.

Jorgenson, D. and Griliches, Z. 1972b. Final reply. *The Measurement of Productivity*. Washington, DC: Brookings Institution.

How to cite this article

Bosworth, Barry. "Denison, Edward (1915–1992)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000265> doi:10.1057/9780230226203.0375

The New Palgrave Dictionary of Economics Online

dependency

José Gabriel Palma

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The focus of all ‘dependency’ analyses is the development of peripheral capitalism (or lack of it). One approach, begun by Baran, Sweezy and Frank, attempted to construct a theory of the practical impossibility of capitalist development in the periphery. A second emerged from the Structuralist School, especially Furtado, Pinto and Sunkel, and tried to reformulate the classical ECLAC analysis from the perspective of the obstacles to ‘national’ development. A third, initiated by Cardoso and Faletto, concentrated on studying ‘concrete situations of dependency’ – how the specific dynamic of different peripheral societies emerges from the interaction between their internal and external structures.

Keywords

Baran, P.; capitalism; capitalist development; dependency; economic development; exploitation; Frankfurt School; Furtado, C.; Harrod–Domar theory; imperialism; industrialization; Lenin, V. I.; Marx, K. H.; Marx's analysis of capitalist production; monopoly capital; multinational corporations; periphery; socialism; structuralism; surplus; underdevelopment

Article

Dependency theories emerged in Latin America in the early 1960s as a challenge to traditional Marxist and structuralist thinking regarding whether capitalist development in the periphery was both still *viable* (given the transformations of the world economy after the Second World War), and still *necessary* (as an unavoidable transition step towards socialism).

There can be little doubt that the Cuban Revolution was a turning point in Marxist analysis of capitalist development in the periphery. The events in Cuba gave rise to a new approach, of which most of the ‘dependency analyses’ form part. This argued that capitalism had totally lost its historical ‘progressive’ role in the periphery (if it ever had one); that is, it was both no longer capable of developing the productive forces of backward societies, and (thus) no longer able to bring them closer towards

socialism. Consequently, this approach also argued against the politics of the popular fronts in the periphery and in favour of an immediate transition towards socialism.

Following traditional Marxist analysis, the pre-dependency, pre-Cuban Revolution approach saw capitalism as still historically progressive in the periphery; however, it argued that its key historical task – the ‘bourgeois-democratic’ revolution – was being inhibited by a new alliance between imperialist forces and the traditional oligarchies. The bourgeois-democratic revolution was the revolt of the emerging capitalist forces of production against the old pre-capitalist order. This revolution would be based on an alliance between the rising bourgeoisie and other progressive forces of society; the principal battle line would be between the new capitalist elites and the traditional oligarchies – between industry and land, capitalism versus pre-capitalist forms of monopoly and privilege. Because it would be the result of the pressure of a rising class whose path was being blocked in political, economic and social terms, this revolution would bring to the periphery (as it had done in the centre) not only political emancipation but economic progress as well.

One of the main analytical challenges facing the pre-dependency Marxist analysis was to explain why the ‘bourgeois–democratic’ revolution in the periphery was not really happening as expected (a phenomenon that was seriously hindering the process of capitalist development there). Since Lenin, this analysis had identified imperialism as the unmistakable main obstacle facing this revolution. The traditional oligarchies could not be the reason for this as on their own, they were not expected to prove any match for the new emerging capitalist classes. Therefore, the principal target in this struggle was unmistakable: North American imperialism. The allied camp for this fight, by the same reasoning, was also clear: everyone, except those (pre-capitalist) internal groups allied with imperialism. Thus, the anti-imperialist struggle was at the same time a struggle for domestic capitalist development and industrialization. The state and the ‘national’ bourgeoisie appeared as the potential leading agents for the development of the new capitalist economy, which in turn was viewed as a necessary stage towards socialism.

The Cuban Revolution questioned the very essence of this approach, insisting that the domestic bourgeoisies in the periphery no longer existed as a progressive social force but had become ‘lumpen’, ‘rent seekers’, incapable of rational accumulation and rational political activity, dilapidated by their consumerism, and blind to their ‘real’ long-term interests. It is within this framework, and with the explicit motive of developing theoretically and documenting historically this new approach that dependency analysis appeared on the scene. At the same time, both inside and outside the Economic Commission for Latin America (ECLAC), two other major Dependency Schools began to develop (see structuralism).

The general focus of all ‘dependency’ analyses is the development of peripheral capitalism (or, rather, the lack of it). More specifically, these studies attempted to analyse the obstacles to capitalist development in the periphery from the point of view of the new interplay between ‘internal’ and ‘external’ structures that had emerged after the Second World War. However, this interplay was analysed in several different ways.

With the necessary degree of simplification that every classification of intellectual tendencies entails, I distinguish between three major approaches – not mutually exclusive from the point of view of intellectual history – in ‘dependency’ analysis. First is the approach begun by Paul Baran, Paul Sweezy and Andre Gunder Frank; its essential characteristic is that it attempted to construct a comprehensive theory of the practical impossibility of capitalist development in the periphery. In these theories the

‘dependent’ character of peripheral economies is the crux on which the whole analysis of underdevelopment turns; that is, dependency is seen as causally linked to permanent capitalist underdevelopment.

The second approach is associated with the ECLAC Structuralist School, especially Celso Furtado, Aníbal Pinto and Osvaldo Sunkel. These writers sought to reformulate the classical ECLAC analysis of Latin American development from the perspective of a critique of the obstacles to ‘national’ development. This attempt at reformulation was not just process of adding new elements (mainly political and social) that were lacking in the original Prebisch–ECLAC analysis (see Prebisch, Raúl), but a thoroughgoing attempt to proceed beyond that analysis, adopting an increasingly different perspective.

Finally, the third approach, started by Fernando Henrique Cardoso and Enso Faletto, attempted to distance itself from the first by deliberately avoiding the formulation of a mechanico-formal theory of dependency and underdevelopment – specifically, by trying to avoid a mechanico-formal theory of the inevitability of underdevelopment in the capitalist periphery based on its dependent character. In turn, it concentrated on the study of what have been called ‘concrete situations of dependency’; that is to say, the precise forms in which the different economies and polities of the periphery have been articulated with those of the advanced nations at different times, and how their specific dynamics have thus been generated.

The first approach: dependency as a formal theory of the inevitability of capitalist underdevelopment: on cutting a knot that could not be unravelled

There is no doubt that the ‘father’ of this approach was Baran. His principal contribution (1957) took up the approach of the Sixth Congress of the COMINTERN regarding the supposedly irresolvable nature of the contradictions between the economic and political needs of imperialism and those of the processes of political transformation, economic development and industrialization of the periphery.

To defend its interests, international monopoly capital would not only form alliances with pre-capitalist domestic oligarchies intended to block progressive capitalist transformations in the periphery, but its activities would also have the effect of distorting the process of capitalist development in these countries. As a result, international monopoly capital would have easy access to peripheral resources and finance, and the traditional élites in the periphery would be able to maintain their monopoly on power and their traditional (mostly predatory and rent-seeking) modes of surplus extraction. Within this context the possibilities for any form of dynamic economic growth in dependent countries were extremely limited or non-existent; the surplus they were able to generate (mainly from primary commodity export activities) was largely appropriated by foreign capital, or otherwise squandered by traditional elites. Therefore, long-term economic stagnation and underdevelopment was inevitable. The only way out was political. At a very premature stage, capitalism had become a fetter on the development of the productive forces in the periphery and, consequently, its historical role had already come to an early end.

Baran developed his ideas influenced both by the Frankfurt School's general pessimism regarding the nature of capitalist development (see Jay, 1996) and by Sweezy's (1946) proposition that the rise of monopolies imparts to capitalism a tendency towards stagnation and decay (see monopoly capitalism). He also followed the main growth paradigm of his time, the Harrod–Domar model, which held that the size of the investable surplus was the crucial determinant of growth (together with the efficiency with which it was used: the incremental capital–output ratio).

Starting out with Baran's analysis, Frank (1967) attempted to prove the thesis that the only political and economic solution to capitalist underdevelopment was a radical transformation of an immediately socialist character. For our purposes we may identify three levels of analysis in Frank's model of the ‘development of underdevelopment’. In the first (arguing against ‘dualistic’ analyses), he attempted to demonstrate that the periphery had been incorporated and fully integrated into the world capitalist economy since the very early stages of colonial rule. In the second, he tried to show that such incorporation into the world capitalist economy had transformed the countries in question immediately into capitalist economies. Finally, in the third level, Frank attempted to prove that the integration of these supposedly capitalist economies into the world capitalist system was achieved through an interminable metropolis–satellite chain, through which the surplus generated at each stage was successfully siphoned off towards the centre. Therefore, for Frank the choice was clear: continue to endlessly underdevelop within capitalism, or socialist revolution.

In my opinion, the real value of Frank's analysis is his critique of the supposedly dual structure of peripheral societies. Frank argues convincingly that the different sectors of the economies in question are and have been, since very early in their colonial history, well integrated to the world economy. Moreover, he has correctly emphasised that this integration has not automatically brought about capitalistic economic development, such as ‘optimistic’ models (derived from Adam Smith) would have predicted, in which increased international trade and the division of labour would inevitably bring about economic growth and prosperity. Nevertheless, Frank's error lies in his attempt to explain this phenomenon by using the same economic deterministic framework of the model he purports to transcend. In fact, he merely turns it upside-down: integration into the world economy cannot possibly bring about capitalism development in the periphery because the development of the industrialised centre necessarily requires the underdevelopment of the periphery. Frank's error is characteristic of the whole tradition of which he is part, including Baran (1957), Sweezy (1946), Amin (1970) and Wallerstein (1974; 1980) among the better known. In their analysis, there is always a priority of external over internal structures; in order to do this, they have to separate almost metaphysically the two sides of the opposition (the internal and the external), losing in the process the notion of movement through the dynamic of the contradictions between these two structures. The analysis which emerges is one typified by ‘antecedent causation and inert consequences’.

It is not surprising that this type of analysis leads Frank to develop a circular concept of capitalism. Although it is evident that capitalism is a system where production for profits via exchange predominates, the opposite is not necessarily true: the existence of production for profits in the market is not necessarily an indication of capitalist relationship of production. For Frank, this is a sufficient condition for the existence of capitalist forms of surplus extraction (and for the periphery to have been ‘capitalist’ since the beginning of colonial rule).

Although Frank did not go very far in his analysis of the world capitalist system as a whole, of its origins and its development, Amin (1970) and Wallerstein (1974; 1980) tackled this tremendous challenge. The central concerns of Frank's theory of the ‘development of underdevelopment’ are also addressed by dos Santos (1970), Marini, Caputo, Pizarro, Hinkelammert, and continued later on by many non-Latin American social scientists. The most thoroughgoing critiques of these theories of underdevelopment have come from Brenner (1977), Cardoso (1972), Kay (1989), Laclau (1971), Lall (1975), Palma (1978), and Warren (1980).

I would argue that the theories of dependency examined here are mistaken not only because they do not ‘fit the facts’, but also – and equally important – because their mechanico-formal nature renders them both static and ahistorical. Their analytical focus has not been directed to the understanding of how new forms of capitalist development in the periphery have been marked by a series of specific economic, political, and social contradictions, instead only to assert the claim that capitalism had lost, or never had, a historically progressive role in the periphery.

Now, if the argument is that the progressiveness of capitalism has manifested itself in the periphery differently from in advanced capitalist countries, or in diverse ways in the different branches of the peripheral economies, or that it has generated inequality at regional levels and in the distribution of income, and has been accompanied by such phenomena as unemployment, and has benefited the elite almost exclusively, or again that it has taken on a cyclical nature, then this argument does no more than affirm that the development of capitalism in the periphery has been characterized by its contradictory and exploitative nature. The specificity of capitalist development in the Third World stems precisely from the particular ways in which these contradictions have been manifested, the different ways in which many of these countries have faced and temporarily overcome them, the ways in which this process has created further contradictions, and so on. It is through this process that the specific dynamic of capitalist development in different peripheral countries has been generated.

Reading their political analysis, one is left with the impression that the whole question of what course the revolution should take in the periphery revolves solely around the problem of whether or not capitalist development is viable. In other words, their conclusion seems to be that, if one accepts that capitalist development is feasible on its own terms, one is automatically bound to adopt the political strategy of waiting for and/or facilitating such development until its full productive powers have been exhausted, and only then to seek to move towards socialism. As it is precisely this option that these writers wish to reject, they have been obliged to make in their work a forced march back towards a pure ideological position to deny any possibility of capitalist development in the periphery.

The second approach: dependency as a reformulation of the ECLAC analysis of Latin American development

Towards the end of the 1960s the analysis of ECLAC regarding Latin American development suffered a gradual decline due to several key factors (see Furtado, Celso). The apparently gloomy panorama of capitalist development in Latin America in the 1960s led to substantial ideological changes in many influential ECLAC thinkers, and it strengthened the convictions of the Marxist ‘dependency’ writers reviewed earlier. The former were faced with the problem of trying to explain the apparent failure of their structuralist policies, particularly concerning import-substituting industrialization (see structuralism). The latter felt vindicated in their view of the unfeasibility of any form of ‘dependent capitalist development’.

Finally, by making a basically ethical distinction between ‘economic growth’ and ‘economic development’, most of the research done within the perspective of this second approach followed two separate lines, one concerned with the obstacles to economic growth (and in particular to manufacturing), the other with the apparently perverse character taken by capitalist development. The fragility of this formulation lies in its inability to distinguish between a socialist critique of capitalism and the analysis of the actual obstacles to capitalist development in the periphery.

The third approach: dependency as a methodology for the analysis of ‘concrete situations of development’

In my critique of the dependency studies reviewed so far, I have described the fundamental elements of what I understand to be the third of the three approaches within the dependency school. This approach is primarily associated with the work of Cardoso and Faletto, dating from the completion of their 1967 book.

Briefly, this third approach to the analysis of dependency can be summarized as follows.

1. In common with the two other approaches to ‘dependency’ discussed already, this third approach sees the Latin American economies as an integral part of the world capitalist system, in the context of increasing internationalization of the system as a whole. It also argues that the central dynamic of that system lies outside the peripheral economies and that, therefore, the options which are open to them are limited (but not determined) by the development of the system at the centre. In this way the ‘particular’ is in some way conditioned by the ‘general’. Therefore, a basic element for the analysis of these societies is given by the understanding of the general determinants of the world capitalist system, which is itself rapidly changing. However, the theory of imperialism, which was originally developed to provide an understanding of the dynamics of that system, has had enormous difficulty in keeping up with the significant and decisive changes in the capitalist system since the death of Lenin. During this period, capitalism underwent substantial changes, and the theory failed to keep up with them properly. One widely recognized characteristic of the third approach to dependency has been its effort to incorporate these transformations. For example, this approach was quick to grasp that the rise of the multinational corporations after the Second World War progressively transformed centre–periphery relationships, as well as relationships between the countries of the centre. As foreign capital became increasingly directed towards manufacturing industry in the periphery, the struggle for industrialization, which was previously seen as an anti-imperialist struggle, in some cases increasingly become the goal of foreign capital itself. Thus dependency and industrialization ceased to be necessarily contradictory processes, and a path of ‘dependent development’ for important parts of the periphery became possible.
2. The third approach has not only accepted but has also tried to enrich the analysis of how developing societies are structured through unequal and antagonistic patterns of social organization, showing the social asymmetries, the exploitative character of social organization and its relationship with the socio-economic base. This approach has also given considerable importance to the particular aspects of each economy like the effect of the diversity of natural resources, geographic location and so on, thus also extending the analysis of the ‘internal determinants’ of the development of peripheral economies.
3. However, while these improvements are important, the most significant feature of this approach is that it attempts to go beyond the analysis these internal and external elements, and insists that from the premises so far outlined one arrives at only a partial, abstract and indeterminate characterization of the historical process in the periphery, which can only be overcome by understanding how the ‘general’ and the ‘specific’ determinants interact in particular and concrete situations. It is only by understanding the specificity of ‘movement’ in the peripheral societies as a dialectical unity of both these internal and external factors that one can explain the particularity of social, political and economic processes in these societies.

Only in this way can one explain how, for example, the same process of mercantile expansion could simultaneously produce systems of slave labour, systems based on other forms of exploitation of

indigenous populations, and incipient forms of wage labour. What is important is not simply to show that mercantile expansion was the basis of the transformation of most of the periphery, and even less to deduce mechanically that that process made these countries immediately capitalist. Rather, this approach emphasizes the specificity of history and seeks to avoid vague, abstract concepts by demonstrating how, throughout the history of backward nations, different sectors of local classes allied or clashed with foreign interests, organized different forms of the state, sustained distinct ideologies or tried to implement various policies or defined alternative strategies to cope with a constantly changing imperialist challenge.

The study of the dynamic of dependent societies as a dialectical unity of internal and external factors implies that the conditioning effect of each on the development of these societies can be separated only by undertaking a static (and metaphysical) analysis. Equally, if the internal dynamic of the dependent society is a particular aspect of the general dynamic of the capitalist system, it does not imply that the latter produces concrete effects in the former, but only that it finds concrete expression in that internal dynamic.

The system of ‘external domination’ reappears as an internal phenomenon through the social practices of local groups and classes, who share the interests and values of external forces. Other internal groups and forces oppose this domination, and in the concrete development of these contradictions the specific dynamic of the society is generated. It is not a case of seeing one part of the world capitalist system as ‘developing’ and another as ‘underdeveloping’, or of seeing imperialism and dependency as two sides of the same coin, with the underdeveloped or dependent world reduced to a passive role determined by the other.

There are, of course, elements within the capitalist system that affect all developing economies, but it is precisely *the diversity within this unity* that characterizes historical processes. Thus the analytical focus should be oriented towards the elaboration of concepts capable of explaining how the general trends in capitalist expansion are transformed into specific relationships between individuals, classes and states, how these specific relations in turn react upon the general trends of the capitalist system, how internal and external processes of political domination reflect one another, both in their compatibilities and their contradictions, how the economies and polities of peripheral countries are articulated with those of the centre, and how their specific dynamics are thus generated.

However, as is obvious, this third approach to the analysis of peripheral capitalism is not unique to ‘dependency’ studies and as such, in time, has superseded them.

See Also

- Baran, Paul Alexander
- Engels, Friedrich
- Furtado, Celso
- Lenin, Vladimir Ilyich [Ulyanov]
- Marx, Karl Heinrich
- Marx's analysis of capitalist production
- monopoly capitalism
- structuralism

- Sweezy, Paul Marlor

I am extremely grateful to Fiona Tregenna for many constructive comments.

Bibliography

Amin, S. 1970. *L'accumulation à l'échelle mondiale: critique de la théorie du sous-développement*. Paris: Anthropos; New York: Monthly Review Press, 1975.

Baran, P. 1957. *The Political Economy of Growth*. New York: Monthly Review Press.

Brenner, R. 1977. The origins of capitalist development: a critique of neo–Smithian Marxism. *New Left Review* 104, 25–93.

Cardoso, F.H. 1972. Dependency and development in Latin America. *New Left Review* 74, 83–95.

Cardoso, F.H. and Faletto, E. 1967. *Dependencia y Desarrollo en América Latina*. Mexico, Siglo XXI; Berkeley: University of California Press, 1977.

Dos Santos, T. 1970. The structure of dependence. *American Economic Review* 60, 231–6.

Frank, A.G. 1967. *Capitalism and Underdevelopment in Latin America: Historical Studies of Chile and Brazil*. New York: Monthly Review Press.

Jay, M. 1996. *The Dialectical Imagination: A History of the Frankfurt School and the Institute for Social Research 1923–1950*. Berkeley: University of California Press.

Kay, C. 1989. *Latin American Theories of Development and Underdevelopment*. London: Routledge.

Laclau, E. 1971. Feudalism and capitalism in Latin America. *New Left Review* 67, 19–38.

Lall, S. 1975. Is dependence a useful concept in analysing underdevelopment? *World Development* 11, 799–810.

Owen, R. and Sutcliffe, B., eds. 1972. *Studies in the Theory of Imperialism*. London: Longman.

Palma, J.G. 1978. Dependency: a formal theory of underdevelopment or a methodology for the analysis of concrete situations of underdevelopment? *World Development* 6, 881–924.

Sweezy, P.M. 1946. *The Theory of Capitalist Development*, London: D. Dobson.

Wallerstein, I. 1974. *The Modern World System: Capitalist Agriculture and the Origins of the European World–Economy in the Sixteenth Century*. New York: Academic Press.

Wallerstein, I. 1980. *The Modern World System II: Mercantilism and the Consolidation of the European World-Economy, 1600–1750*. New York: Academic Press.

Warren, B. 1980. *Imperialism: Pioneer of Capitalism*. ed. J. Sender. London: Verso.

How to cite this article

Palma, José Gabriel. "dependency." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000077> doi:10.1057/9780230226203.0376

The New Palgrave Dictionary of Economics Online

deposit insurance

Stephen G. Cecchetti
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The purpose of deposit insurance is to ensure financial stability, as well as protect the interests of small investors. But with government guarantees in hand, bankers take excessive risks, driving up the chances of failure. Evidence suggests that these schemes increase rather than decrease the probability of financial crises. There is a good chance that deposit insurance does more harm than good. This article surveys the rationale for and history of deposit insurance, and discusses its consequences and possible alternatives.

Keywords

assets and liabilities; asymmetric information; Bagehot, W.; banking crises; banking industry; deposit insurance; excessive risk taking; Federal Reserve System; financial intermediaries; financial market contagion; Great Depression; lender of last resort; moral hazard; non-bank financing mechanisms; risk

Article

People living in countries where bank deposits are insured would never question the wisdom of an explicit insurance scheme. The idea that their savings are protected by a government-backed guarantee is something they simply take for granted. Only some crazy economist would ask whether deposit insurance makes sense. Well, does it? Surprisingly, the evidence is that it may not. Deposit insurance, which is supposed to stabilize the financial system, may do more harm than good.
This article examines the nature of deposit insurance by answering the following series of questions: (*a*) What do financial intermediaries do that warrants government intervention? (*b*) What is the history of deposit insurance? (*c*) Does deposit insurance do what it is designed to do? And (*d*), are there any alternatives?

Financial intermediaries, banks, and bank runs

The term ‘financial intermediaries’ encompasses a large set of institutions that include depository institutions as well as insurance companies, securities firms and pension funds. The first of these – what we all call ‘banks’ – are both the most commonly known to individuals and provide the broadest array of services. They pool savings, accepting resources from a large number of small savers in order to provide large loans to borrowers; provide access to the payments system, so that individuals can make and receive payments; provide liquidity, allowing depositors to transform their financial assets into money quickly and easily at low cost; and diversify risk, giving even the smallest saver a mechanism for diversification.

To appreciate the importance of financial intermediaries, consider what it would be like without them. If banks didn't exist, all finance would be direct, with borrowers obtaining funds straight from the lenders. Such a system would be costly and ultimately ineffective. It would be so difficult and expensive for borrowers and lenders to find each other, and then to come to agreement over the terms of a loan, that it is unlikely there would be any transactions at all. And without a financial system to transfer funds from savers to investors, there would be no economic development. The world would be a very different place.

Because of the services they provide, banks face a risk that other financial institutions (and industrial firms) do not. They are vulnerable to runs. Here's why. Banks issue liquid liabilities in the form of short-term demand deposits, and hold illiquid long-term assets, structured as securities and loans. The bank promises all its depositors that, if they want the entire balance of their checking account, they just have to come and ask. If a bank has insufficient funds to meet requests for withdrawal on demand, it will fail.

Banks not only guarantee their depositors immediate cash on demand; they promise to satisfy depositors’ withdrawal requests on a first-come, first-served basis – what is called a ‘sequential service constraint’. This commitment has important implications. Suppose depositors begin to lose confidence in a bank's ability to meet their withdrawal requests. True or not, reports that a bank has become *insolvent* can spread fear that it will run out of cash and close its doors. Mindful of the bank's first-come, first-served policy, panicked depositors rush to convert their account balances into cash before other customers arrive. Such a *bank run* can cause a bank to fail. Importantly, if people believe that a bank is in trouble, that belief alone can make it so.

While banking system panics and financial crises can result from false rumours, they can also come about for more concrete reasons. Widespread downturns in economic activity drive down the value of loans and securities, so bank capital (the difference between assets and liabilities) falls. If things get bad enough, banks become insolvent and fail. A big economic downturn can put the entire financial system at risk. Gorton (1988) reports that significant contractions are associated with all seven of the severe financial panics in the United States that occurred between 1871 and 1914.

In a market-based economy, the opportunity to succeed is also an opportunity to fail. It would be natural to dismiss bank failures as analogous to the closing of an unpopular restaurant. But, while individual banks should be, and are, allowed to fail, the fact that banks are dependent on one another (in a way that restaurants are not) means that when one bank fails it puts others at risk.

Banks are linked both on their balance sheets and in their customers’ minds. In recent years in the United States, inter-bank loans make up roughly four per cent of bank assets – an amount that represents almost half of bank capital. If one bank fails, it could put the system at risk. Information asymmetries are the reason that a depositor run on a single bank can turn into a bank panic that threatens the entire financial system. Most of us are not in a position to assess the quality of a bank's balance sheet. So, when rumours spread that a certain bank is in trouble, depositors everywhere begin to worry about their

own banks' financial condition. Concern about even one bank can create a panic that causes profitable banks to fail, leading to a complete collapse of a country's banking system. Bank failure is contagious. All of this leads to the following conclusions. Not only are individual banks fragile and vulnerable to runs, but the entire banking system is prone to panics. Contagion creates an externality that provides the economic justification for government intervention in the system.

Deposit insurance and the government safety net

Government officials intervene in the financial system both to protect small investors and to ensure financial stability. They do it with two related tools: the lender of last resort, where a central bank that can issue liabilities without limit provides loans to banks that are illiquid but not insolvent; and deposit insurance.

History reveals that the presence of a lender of last resort significantly reduces, but does not eliminate, bank panics. The series of three bank panics in the United States during the Great Depression of the 1930s, described in Friedman and Schwartz (1963), is one example of a failure of this sort. The Federal Reserve System was in place and had the capacity to operate as a lender, but did not.

The first national deposit insurance scheme was enacted by the US Congress in 1935 as a direct response to the bank panics in the 1930s. White (1995) sets out the history, noting that the debate was contentious, and that the stated purpose of deposit insurance was to stabilize the banking system. As surprising at it may seem from a modern perspective, investor protection per se was not the point.

When one thinks about deposit insurance, it is important to keep in mind that no private fund can be large enough to withstand a system-wide panic. Only the fiscal authority (possibly combined with the central bank) has the necessary resources.

For decades the US system was nearly unique. In 1974 only 12 countries had explicit national deposit insurance systems. Explicit deposit insurance is a phenomenon of the last quarter of the 20th century, when it became a part of the generally accepted best-practice advice international organizations gave to developing countries. Demirgüç-Kunt and Kane (2002) report that by 1999 the number of countries with deposit insurance had risen to 71 (with the insurable limits ranging up to more than eight times a country's per capita GDP). Prior to this, most systems were implicit, whereby depositors would exert their substantial political influence to force fiscal authorities to supply unlimited deposit guarantees in the event of a bank failure. This is all somewhat surprising, given the obvious political appeal of any system that has no immediate budgetary outlay associated with it. What politician wouldn't want to make an apparently costless promise to protect the bank deposits of his or her constituents?

Does deposit insurance work?

In their classic theoretical treatment of deposit insurance, Diamond and Dybvig (1983) show that, if self-fulfilling depositor runs result from information asymmetries, then government-supplied insurance can improve social welfare. But at what cost?

Insurance changes people's behaviour. Protected depositors have no incentive to monitor their bankers' behaviour. Knowing this, a bank's managers take on more risk than they would otherwise, since they get the benefit of risky bets that pay off while the government assumes the costs of the ones that don't. In

protecting depositors, then, the government creates moral hazard. This is not just a theory. In 1980, the deposit insurance limit in the United States was raised to $100,000, four times its earlier level. Over the following ten years, several thousand depository institutions (banks and savings and loans) failed. That was more than four times the number that failed in the first 46 years of explicit deposit insurance. While a vast majority of the institutions that failed in the 1980s were small, the cost of reimbursing depositors exceeded 3 percent of one year's GDP. The bill was ultimately paid by US taxpayers.

The problem of excessive risk taking did not stop with the resolution of the 1980s crisis. Today, the US banking system's assets are worth between 10 and 12 times their equity. In the 1920s, this same leverage ratio was closer to four. Industrial firms typically have leverage that is half that lower number. In other words, deposit insurance has driven up leverage in banking. And with the increase in leverage comes an equal increase in risk (as measured by the standard deviation of returns).

So, in an attempt to solve one problem, deposit insurance created another. And to combat bankers' excessive risk taking, governments were forced to set up regulatory and supervisory structures. Among other things, there are now constraints on the assets banks can hold, rules governing the minimum levels of capital that banks must maintain, and requirements that banks make public information about their balance sheets. Supervisors have to enforce the detailed web of regulations.

Does this complex mechanism actually work to stabilize the financial system? The evidence is not encouraging. Demirgüç-Kunt and Kane (2002) summarize international research and conclude that explicit deposit insurance actually makes financial crises more likely. When countries have either implemented a new scheme or expanded an existing one, the probability of crises has increased.

To make matters worse, the creation of deposit insurance retards the evolution of non-bank financing mechanisms. Cecchetti and Krause (2005) find that countries with more extensive deposit insurance schemes tend to have both smaller financial markets and a fewer publicly traded firms per capita. To put it bluntly, deposit insurance is bad for financial development, and may be bad for real economic growth.

Are there alternatives?

So, if deposit insurance schemes do more harm than good, what should we do to stabilize the financial system? The natural response of an economist is to use the price system. Measure how risky a bank's balance sheet is, and set its deposit insurance premiums accordingly. Beginning in 1991, the US Federal Deposit Insurance Corporation did implement a risk-based premium structure. But this is extremely difficult to do well. Banks can always find ways to evade detailed rules, exploiting the system to reduce the prices they pay. In the end, this is not a solution.

There are three other options. We could implement changes that further restrict the assets held by banks, eliminating their asset transformation function. We could increase our reliance on the central bank's lender-of-last-resort function. Or it may be possible to design a scheme to ensure that large depositors will impose discipline on the risk taking of bank managers.

Proposals for narrow banking are in the first category. A narrow bank is an institution that holds only a very limited set of very low-risk, highly liquid assets, such as short-term government securities. Since insolvency is impossible for such an institution, liability holders would not have to worry about the quality of the narrow bank's assets, and there would be no fear of a run. Deposit insurance would be unnecessary.

Second, it may be possible to address the potential for systemic bank panics by improving the

effectiveness of the lender of last resort. In 1873, Walter Bagehot suggested that, in order to prevent the failure of solvent but illiquid financial institutions, the central bank should lend freely on good collateral at a penalty rate. By lending freely, he meant providing liquidity on demand to any bank that asked. Good collateral would ensure that the borrowing bank was in fact solvent, and a high interest rate would penalize the bank for failing to manage its assets sufficiently cautiously. While such a system could work to stem financial contagion, it has a critical flaw. For it to work, central bank officials who approve the loan applications must be able to distinguish an illiquid from an insolvent institution. But during times of crisis computing the market value of a bank's asset is almost impossible, since there are no operating financial markets and no prices for financial instruments. Because a bank will go to the central bank for a direct loan only after having exhausted all opportunities to sell its assets and borrow from other banks without collateral, its illiquidity and its need to seek a loan from the government draw its solvency into question. Officials anxious to keep the crisis from deepening are likely to be generous in evaluating the bank's assets, and to grant a loan even if they suspect the bank might be insolvent. And, knowing this, bank managers will tend to take too many risks.

Finally, we could require that banks issue subordinated debt. These are unsecured bonds, with the lender being paid only after all other bondholders are paid. Someone who buys a bank's subordinate debt has a very strong incentive to monitor the risk-taking behaviour of the bank. The price of these publicly traded bonds then provides the market's evaluation of the quality of the bank's balance sheet and serves to discipline its management.

By eliminating the accountability of bank managers to their depositors, deposit insurance encourages risky behaviour. So, while financial stability is clearly in the public interest, deposit insurance may not be.

See Also

- Bagehot, Walter
- financial intermediation
- financial structure and economic development
- moral hazard
- risk

Bibliography

Bagehot, Walter. 1873. *Lombard Street: A Description of the Money Market*. London: Henry S. Kin & Co.

Cecchetti, S. 2006. *Money, Banking, and Financial Markets*. Boston, MA: McGraw-Hill Irwin.

Cecchetti, S. and Krause, S. 2005. Deposit insurance and external finance. *Economic Inquiry* 43, 531–41.

Demirgüç-Kunt, A. and Kane, E. 2002. Deposit insurance around the globe: where does it work?

Journal of Economic Perspectives 16(2), 175–95.

Diamond, D. and Dybvig, P. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91, 401–19.

Friedman, M. and Schwartz, A. 1963. *A Monetary History of the United States: 1867–1960*. Princeton, NJ: Princeton University Press.

Gorton, G. 1988. Banking panics and business cycles. *Oxford Economic Papers* 40, 751–88.

White, E. 1995. Deposit insurance. Policy Research Working Paper No. 1541. Washington, DC: World Bank.

How to cite this article

Cecchetti, Stephen G. "deposit insurance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000081> doi:10.1057/9780230226203.0377

The New Palgrave Dictionary of Economics Online

depreciation

Gautam Mathur

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Depreciation estimates the decline in the value of capital over time. It is highly important to capital accounting, since the rate of dividend is calculated as the ratio of the surplus to the current value of assets. The causes of the depreciation of equipment are twofold: its productivity may fall with age, and over time its expected remaining earning life is shorter. Depreciation is taken to be the difference between gross and net investment but total new employment is given by gross investment: the physical counterpart of replacement (of equipment or manpower) is needed give the full picture.

Keywords

Champernowne, D. G.; depreciation; Kahn, R. F.; national income statistics; one-hoss shay; Robinson, J. V.; Sraffa, P.; Torrens, R.; Von Neumann, J.

Article

Depreciation estimates the decline in the value of capital as a result of ageing, its maximum value being near its age of manufacture and its minimum value when it is dismantled and sold as scrap. It is of great importance to capital accounting, for the rate of dividend is calculated as the ratio of the surplus to the current value of assets. The reduction in value of equipment comes about from two causes – firstly that its productivity may fall with age; and secondly that, as time advances, the expected remaining earning life of the plant is shorter. Hence, the capitalized value of the present value of expected future stream of quasi-rents from an old piece of equipment is smaller for any given rate of interest than for a younger machine.

‘One-hoss shay’ assumption

The influence of declining productivity over time may be eliminated by assuming a ‘one-hoss shay’ type

of equipment, which keeps its efficiency constant over its service life and falls to pieces at the end. However, the product of a process is not only its current output but the stock of equipment which remains at the end of the production period – as stressed by von Neumann (1933) and by Sraffa (who in 1960 referred to Robert Torrens as having insisted in the years 1818 and 1821 on its being considered as a part of output).

On account of the shorter remaining service life of equipment at the end of a period (and the consequent smaller number of expected items of quasi-rent in its stream of earnings), there is lesser value of capital remaining at the end of a production period – a so-called ‘year’. This reduction in value of a stock output affects adversely the productivity in value terms (even with ‘one-hoss shay’ equipment) and it measures the depreciation. There is, therefore, an aggravated tendency of the value of capital embodied to fall as the plant is older.

Shape of decline in valuation curve

In a straight line approximation, depreciation is taken as constant in absolute amount per year. In a formula using the exponential concept depreciation is at a constant rate; hence, the fall in value is more when machines are younger and higher priced, than when they are older – as in radioactive decay, that is, it indicates a curve convex to the origin. But depreciation is at higher rates for older capital in service – not as would be given by an exponentially falling value of equipment at a constant rate with respect to time. When there is a rising rate of reduction of value, it makes the decline more than exponential as the machines are older, and yields a steeply falling value towards the end of the service life, that is, it yields a curve with respect to time which is concave to the origin. The straight line approximation of value of capital (with respect to its age) which is used in some calculations is thus wide of the mark; and even the exponentially falling value according to a constant rate of reduction does not make the value of old machines decline sufficiently markedly.

In a Sraffa or von Neumann valuation of capital (of different ages taken as different commodities) this decline is well brought out automatically, for differences between value of the commodity called ‘equipment *t* years old’ and the one called ‘$t + 1$ years old’, increases as *t* becomes larger.

This aspect of the Sraffa system (1960) was not known to Joan Robinson or to Professor Richard Kahn and D.G. Champernowne in 1954 when the text of *Accumulation of Capital* (1956) was being finalized – especially its Mathematical Appendix (to a part of which the latter two had contributed as authors, their names appearing in the original printed text). It is all the more remarkable that it was discovered that, in the measurement of value of ageing equipment, one could strike upon another useful device – of balanced age composition of capital.

Balanced age composition of capital

In demographic studies as part of the subject of manpower, it is well known that, for a population of human beings growing at *g* per cent, there are higher numbers of children of age *t* in comparison to those a year older (of age *t*+1) by the factor (1+*g*). The same principle can be applied to a population of plants, and we can derive a universe of plants ordered according to their ages in this particular manner. One can try to ascertain what the number of plants in a cohort of each age is and the value of capital embodied in

each cohort.
The value of plant at the centre of gravity of the age-composition pyramid may then be used as the standard unit of measurement of the value of a plant of any particular age. The result would be in agreement with the well-known, but rather mystifying, Kahn–Champernowne formula of the reciprocal value of a new plant in terms of value of the plant of average age. This reciprocal will be called a K–C unit in honour of those two authors who worked out the said formula.

Kahn – Champernowne units of measurement

In a generalized version of this concept, as the set of pieces of capital of constant physical productivity and of balanced age composition growing exponentially at a steady rate, keep the composition in terms of relative sizes of cohorts constant; hence the value of the average plant does not change. This is the justification of the K–C units.
In terms of a balanced age composition of equipment (with *T* years expected service life since its manufacture), a piece of equipment *t* 'years' old is replaced at the end of *t* years by a piece which was $t - 1$ years old in the beginning of the year. Except for this replacement by equipment which is now of the same age as the piece it substitutes, there is no depreciation visible in the physical system or its statistical depiction.

Redundancy of gross and net concepts

It is to be remembered that Joan Robinson had correctly realized that depreciation was not a physical phenomenon but a notional or value one. The implication of depreciation not being a physical phenomenon in terms of effect upon the concepts of gross and net investment had to wait until the von Neumann model was integrated (in 1960) with the Robinsonian golden-age system. In traditional analysis the system is depicted as *z* machines (newly produced and added) in a factory, and at the same time another *z* machine rendered inoperative (by completion of their natural life). But the net investment is not an act of accretion–depreciation in physical terms; for the machines added through current investment are new ones and the depletion is of old machines – and it makes no sense if value measurement were not resorted to for calculating the excess of accretion over depreciation.
The balanced age composition is a device by which one can realize that in a von Neumann system as a growing economy *m* machines of age *t* years exist and $m(1 + g)$ of age $t - 1$ years are automatically substituted a year later by $m(1 + g)$ machines – also now of *t* years age. The stock as well as each age cohort grows at rate *g*, and depreciation of value by ageing is exactly counterbalanced by that much capital of erstwhile younger age and erstwhile higher value (but now of the same age and the same value as the *m* plants at the beginning of the year) replacing it. In addition one has *mg* times more machine of age *t*. The total stock grows at a given rate of growth, and depreciation is also compensated for exactly, for $m(1 + g)$ is equal to *m* for replacement, and *mg* for accumulation for each age cohort.
In Sraffa–von Neumann analysis (as a simplified purposive model combining the two general constituents of those two models and integrating the resultant with the Robinsonian golden-age system), this fact was noticed in 1961, and it was discovered that in a state of steady growth and balanced age composition depreciation of a stock of inputs in terms of writing down of value of equipment (due to

ageing) is a dispensable concept (Mathur, 1965). Each age cohort is replenished exactly by an age cohort from within the system in physical numbers and value, and there is nothing to be written down of any piece of equipment by a chartered accountant at the end of the year. The pieces of equipment of each age are higher by the rate *g*, and valuation is required for finding out cumulative accumulation of equipment in each age cohort. In a body of equipment of balanced age composition as the value of capital of different ages differs by the amount of depreciation, the concept of depreciation is required for measuring aggregate accumulation due to ageing, not for decumulation due to ageing as was required in the traditional concept.

It is because of the total absence of writing down of value of stock of any age that it was realized that there is no concept of gross or net necessary in such a reckoning (Mathur, 1965), and depreciation is important not as the difference between gross and net investment, but as the difference of value of an older machine in relation to a younger one for purposes of measuring accumulation (of positive-age equipment from within the firm and of new equipment from the manufacturers). It is only when the age composition is grossly unbalanced – as for newly established firms – that it may be necessary to use depreciation in the traditional sense of writing down value of stocks. But in that case measurement of depreciation or of amount to be written off is itself a procedure not entirely free from logical doubts.

Depreciation and maintenance

In manpower-employment terms, total new employment is given by gross investment and not by net investment, because the amount spent on activities of maintaining capital intact (repairing, renovating) also creates employment, and not only the building of new capital. Hence, in national income statistics, it is gross investment which creates manpower employment and not net investment by itself. The difference between gross and net is taken to be depreciation, but in manpower terms it does not so follow – for employment created for maintenance of a machine (like a sealed unit) might be very low, and yet the reduction of its value year by year very high due to ageing. When viewing manpower statistics, the activity of operatives of a particular type ought to be supplemented by statistics of valuation (Mathur, 1983). While figures in terms of counting heads are important for a physical count, greater economic significance would be acquired if the productivity of each type of human equipment were determined and its true value calculated in K–C units with respect to a balanced age composition and age structure (Mathur, 1964). But depreciation in value terms alone without the physical counterpart of replacement (of equipment or manpower) also tells us an incomplete story, and only valuation and quantification (in physical terms) together give the full picture.

See Also

- amortization

Bibliography

Champernowne, D.G. and Kahn, R.F. 1956. The value of invested capital. Mathematical Note appended

at the end of Robinson (1956).

Mathur, G. 1964. The valuation of human capital for manpower planning. *Applied Economic Papers* (Hyderabad) 4(2), 14–35.

Mathur, G. 1965. *Planning for Steady Growth*. Oxford: Basil Blackwell; New York: Augustus Kelly.

Mathur, G. 1983. Web of inequity. Presidential Address to the Silver Jubilee Session of the Indian Labour Economic Association, Lucknow, 1982. *Indian Journal of Labour Economics*, Sec. VIII.

Neumann, J. von. 1933. A model of general economic equilibrium. *Review of Economic Studies* 13 (1945), 1–9.

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

How to cite this article

Mathur, Gautam. "depreciation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000083> doi:10.1057/9780230226203.0378

The New Palgrave Dictionary of Economics Online

derived demand

J.K. Whitaker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost functions; derived demand; elasticity of substitution; Hicks, J. R.; Marshall, A.; Robinson, J. V.

Article

The idea that the demand for intermediate goods is *derived* from the demand for the final goods they help produce is obvious and appealing. It was implied by Cournot (1838, pp. 99–116) and explicitly stated by Gossen (1854, pp. 31, 113) and Menger (1871, pp. 63–7). That the British classical school failed to make use of such a perspective – Mill's famous proposition that ‘demand for commodities is not demand for labour’ (1848, Book I, ch. 5) came close to denying it – was doubtless due to the strong emphasis placed on prior accumulation of capital as a prerequisite for production. But it was Alfred Marshall in his *Principles of Economics* (1890, pp. 381–93, 852–6) who introduced the term ‘derived demand’ and developed the concepts of the derived demand curve for an input and the elasticity of derived demand.

Marshall focused on a case in which a commodity is produced by the cooperation of several inputs, which are thus jointly demanded for the purpose, the demand for each being derived from the demand for the product. His formal analysis proceeded on the assumption that the inputs were all combined in fixed proportions (which might vary with the scale of output) although he suggested that the variable-proportions case would be similar.

A derived demand curve can be constructed for a selected input on the assumptions that production conditions, the demand curve for output, and the supply curves for all other inputs remain fixed, and that the competitive markets for output and all other inputs are always in equilibrium. The resulting derived demand curve can most easily be interpreted as the outcome of a hypothetical experiment. Make the selected input available, perfectly elastically, at an arbitrary price, y, per unit. Now ascertain, under the above conditions about the markets for output and other inputs, what quantity, x, of the selected input would be demanded. All other markets must be in equilibrium, and each seller or buyer must be optimally adjusted to the assumed terms of availability of the selected input. Repeating this experiment for different values of y would generate the inverse of the relationship between x and y, $y = f(x)$, whose

graphical representation is Marshall's derived demand curve for the selected input. Bringing this demand curve into conjunction with the actual supply curve of the selected input will determine the actual equilibrium price and quantity for this input and thereby implicitly determine the actual equilibrium prices and quantities of output and all other inputs. But the point of obtaining the derived demand curve is not to permit such a two-stage determination of the actual equilibrium. It is rather to permit a simplified analysis of the effect of *changes* in the supply conditions of the selected input when supply conditions of other inputs, as well as technology and the demand conditions for output, remain unaltered. Marshall invoked a simple example in which the final product, a knife, is obtained by joining costlessly a unit each of the two inputs, blades and handles. The derived demand curve for handles is then given by the rule that y, the derived demand price for x handles, is the demand price for x knives less the supply price for x blades.

Marshall analysed the conditions producing a low elasticity of derived demand for an input, a condition which would encourage supply restriction. The first condition, the lack of a good substitute, is already implied by the fixity of production coefficients. The second is that the demand for the final output be inelastic. The third, aptly described by Henderson (1922, p. 59) as ‘the importance of being unimportant’, is that expenditure on the input in question be only a small fraction of total production cost. The final condition is that cooperating inputs be in inelastic supply. These last three conditions ensure that a large rise in the price of the input will not raise product price much, that a rise in the product price will not reduce sales much, and that a reduction in sales and production will lower the cost of cooperating inputs substantially.

The next major contribution was that of Hicks (1932, pp. 241–6) who formally relaxed the assumption of fixed production coefficients. He analysed the consequences of input substitutability for a two input case with constant returns to scale in production, making use of his newly invented concept of the elasticity of substitution. His principal finding was that, to get a low elasticity of derived demand, ‘It is “important to be unimportant” only when the consumer can substitute more easily than the entrepreneur’, that is, only when the elasticity of demand for the product exceeds the elasticity of input substitution (1932, p. 246). This finding, which is not easily explained intuitively, has been the subject of intermittent controversy, aptly summarized and resolved in Maurice (1975). The extension of Hicks's analysis to the many-input case has been accomplished by Diewert (1971), using an elegant dual approach based on the cost function concept. However, modern theoretical work is more prone to work explicitly and symmetrically with complete systems of input demand equations for firm and industry.

More or less contemporaneously with Hicks, Joan Robinson (1933, chs. 23, 24) was studying the derived demand curve for an input in cases where the final product is sold by a monopolist, who might also acquire cooperating inputs monopsonistically. The question of when areas under a derived demand curve can be given a welfare interpretation, analogous to consumer surplus for a final demand curve, has been broached by Wisecarver (1974).

The concept of derived demand finds its main application in discussions of labour-market questions, and Marshall's tools still play a significant part in the teaching and writing in that area.

See Also

- acceleration principle

- Marshall, Alfred

Bibliography

Cournot, A.A. 1838. *Mathematical Principles of the Theory of Wealth*. Trans., New York: Macmillan, 1897.

Diewert, W.E. 1971. A note on the elasticity of derived demand in the n-factor case. *Economica* 38, 192–8.

Gossen, H.H. 1854. *The Laws of Human Relations*. Trans., Cambridge, MA: MIT Press, 1983.

Henderson, D.H. 1922. *Supply and Demand*. London: Nisbet.

Hicks, J.R. 1932. *The Theory of Wages*. London: Macmillan.

Marshall, A. 1890. *Principles of Economics*, vol. 1, 8th edn. London: Macmillan, 1920.

Maurice, S.C. 1975. On the importance of being unimportant: an analysis of the paradox in Marshall's third rule of derived demand. *Economica* 42, 385–93.

Menger, C. 1871. *Principles of Economics*. Trans., New York: New York University Press, 1981.

Mill, J.S. 1848. *Principles of Political Economy*. London: Parker.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Wisecarver, D. 1974. The social costs of input-market distortions. *American Economic Review* 64, 359–72.

How to cite this article

Whitaker, J. K. "derived demand." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000093> doi:10.1057/9780230226203.0379

The New Palgrave Dictionary of Economics Online

Destutt De Tracy, Antoine Louis Claude (1754 – 1836)

R.F. Hébert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalization theory of taxation; Destutt de Tracy, A. L. C.; ideology; labour theory of value; Physiocracy; productive and unproductive labour

Article

French philosopher and economist, Tracy was born into a noble family of the *ancien régime* at Paris on 20 July 1754 and died in the same city on 10 March 1836. His life spanned the most tumultuous period of French history, from the twilight of the Old Regime to the dawn of capitalism, romanticism and socialism. One of the last *philosophes*, Tracy began as an 18th-century classical metaphysician, preoccupied with the sensationalist doctrine of Locke and Condillac, and ended up, in the words of Auguste Comte, as the philosopher ‘who had come closest to the positive state’. In the interim he knelt at the feet of Voltaire; served alongside Lafayette in the Royal Cavalry, and as deputy to the French Estates General and the Constituent Assembly; was imprisoned during the Reign of Terror; released after Thermidor (escaping the guillotine by a mere two days); subsequently helped to establish his country's first successful national programme of public education; led the opposition to Napoleon from his seat in the French Senate; regained his title under the Bourbon Restoration; counted among his associates the likes of Mirabeau, Condorcet, Cabanis, DuPont de Nemours, Jefferson, Franklin, Lavoisier, Ricardo and Mill; and retained his early sympathies for liberty throughout.
Long before it took on its pejorative sense at the hands of Marx, Tracy coined the term ‘ideology’ (by which he meant the science of ideas) to describe his philosophy, which embraced and intertwined psychological, moral, economic and social phenomena, but which gave primacy to economics because he thought that the purpose of society was to satisfy man's material needs and multiply his enjoyments. Tracy rejected the Physiocratic notion of value, substituting a labour theory that Ricardo subsequently endorsed in his *Principles*. Like Say, he denied Smith's distinction between productive and unproductive labour. But unlike Smith or Say, he reduced all wealth, including land, to labour. On numerous other topics (that is, wages, profits, rents, exchange, price variations, international trade) he was far less thorough and rigorous than either Smith or Say, but his exposition of the capitalization theory of taxation

was superior to the rest. In the final analysis, his *Traité* was not properly a treatise on political economy so much as a part of a general study of the human will. Yet the resulting lack of depth did not impair his remarkable ability to allure great minds. Ricardo found him ‘a very agreeable old gentleman’, and Jefferson was influenced to the point of including ‘ideology’ among the ten projected departments in his plan for the University of Virginia.

Along with Say, Destutt de Tracy was one of the earliest members of the French liberal school. Patrician, philosopher and patriot, caught in the grips of major social and economic upheaval, he denounced the interests of his own class (the *rentiérs*) and became the spokesman of a nascent capitalism in which he had neither role nor vested interest.

Selected works

1804–18. *Eléments d'idéologie*, 5 vols. Paris: Courcier.

1811. *A Commentary and Review of Montesquieu's Spirit of Laws*. Trans, Philadelphia: William Duane.

1817. *A Treatise on Political Economy*. Ed. T. Jefferson. New York: Augustus M. Kelley, 1970.

Bibliography

Allix, E. 1912. Destutt de Tracy, économiste. *Revue d'Economie Politique* 26, 424–51.

Kennedy, E. 1978. *Destutt de Tracy and the Origins of ‘Ideology’*. Philadelphia: American Philosophical Society.

Picavet, F.J. 1918. *Les Idéologues*, chs 5 and 6. Paris: F. Alcan.

How to cite this article

Hébert, R. F. "Destutt De Tracy, Antoine Louis Claude (1754–1836)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000096>
doi:10.1057/9780230226203.0380

The New Palgrave Dictionary of Economics Online

determinacy and indeterminacy of equilibria

Chris Shannon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses work on the determinacy and indeterminacy of equilibria in models of competitive markets. Determinacy typically refers to situations in which equilibria are finite in number, and local comparative statics can be precisely described. The article describes basic results on generic determinacy for exchange economies and the general underlying principles, together with various applications and extensions including incomplete financial markets and markets with infinitely many commodities.

Keywords

ambiguity aversion; Arrow–Debreu model; comparative statics; concavity; continuous-time trading; continuum of equilibria; determinacy and indeterminacy of equilibria; Edgeworth box economy; excess demand functions; existence of equilibrium; implicit function theorems; incomplete markets; infinite horizons; infinite-dimensional economies; Lipschitz continuous functions; loss aversion; multiple equilibria; reference dependence; restricted participation; Sard's th; transversality th; uniqueness of equilibrium; Walras's Law

Article

1 Introduction

The Arrow–Debreu model of competitive markets is one of the cornerstones of economics. Part of the explanatory power of this model stems from its flexibility in capturing price-taking behaviour in many different markets, and from the predictive power arising from the great generality under which equilibrium can be shown to exist. This predictive power is significantly enhanced when equilibria are determinate, meaning that equilibria are locally unique and local comparative statics can be precisely described. Instead, when equilibria are indeterminate, even arbitrarily precise local bounds on variables might not suffice to give a unique equilibrium prediction, the model might exhibit infinitely many equilibria, and each might be infinitely sensitive to arbitrarily small changes in parameters.

Simple exchange economies cast in an Edgeworth box with two agents and two goods illustrate the possibility of indeterminacy in equilibrium. One easy example arises when agents view the goods as perfect substitutes. In this case, every profile of initial endowments leads to a continuum of equilibria. Another example comes from the opposite extreme, in which each agent views the goods as perfect complements. Every profile of initial endowments dividing equal social endowments of the two goods leads to a continuum of equilibria. These examples may seem degenerate, since they involve individual demand behaviour either extremely responsive to prices, or extremely unresponsive to prices. Similar examples can be constructed using preferences that are less extreme, however, and that can be chosen to satisfy a number of regularity conditions including strict concavity, strict monotonicity, and smoothness. Problems from standard graduate texts illustrate this possibility. In fact, indeterminacy is unavoidable, at least for some endowment profiles, in almost any model that may exhibit multiple equilibria for some choices of endowments. The conditions leading to unique equilibria or unambiguous global comparative statics are well-known to be very restrictive, suggesting that equilibrium indeterminacy may be a widespread phenomenon.

In a deeper sense, however, these examples of indeterminacy remain knife-edge. Under fairly mild conditions on primitives, if an initial endowment profile leads to indeterminacy in equilibrium, arbitrarily small perturbations in endowment profiles must restore the determinacy of equilibrium. More powerfully, the set of endowment profiles for which equilibria are determinate is generic, that is, an open set of full Lebesgue measure. Explaining this remarkable result – originally postulated and established by Debreu (1970) – and its many

extensions and generalizations is the focus of this article. Section 2 lays out the basic question of determinacy of equilibrium in finite exchange economies, and sketches the results. Section 3 describes the general underlying principles, together with various applications and extensions. Section 4 concludes by examining recent work on determinacy in markets with infinitely many commodities.

2 Determinacy in finite exchange economies

Imagine a family of exchange economies, each with a fixed set of *L* commodities and a fixed set of *m* agents, $i = 1, \ldots, m$, with given preferences $\{\succsim_i\}_{i=1,\ldots,m}$, indexed by varying individual endowments $(e_1, \ldots, e_m) \in \mathbf{R}^{mL}_{++}$. Denote the social endowment $\bar{e} := \sum_i e_i$ and a particular profile of individual endowments by $e := (e_1, \ldots, e_m) \in \mathbf{R}^{mL}_{++}$. An economy $E(e)$ then refers to the exchange economy with preferences $\{\succsim_i\}_{i=1,\ldots,m}$ and endowment profile *e*. For simplicity this article focuses on exchange economies. Mas-Colell (1985) is a comprehensive reference that includes discussion of extensions allowing for production.

The crucial departure in Debreu (1970) is to view each economy as a member of this parameterized family, and to ask whether perhaps almost no economies exhibit indeterminacy or pathological comparative statics when indexed this way. To formalize this, Debreu (1970) summarizes an agent's choice behaviour by a C^1 demand function $x_i: \mathbf{R}^L_{++} \times \mathbf{R}^L_{++} \to \mathbf{R}^L_+$ satisfying basic properties such as homogeneity of degree 0 in prices, Walras's Law, and boundary conditions as prices converge to zero. This leads to the familiar characterization of equilibria as zeros of excess demand:

$$0 = z(p, e) := \sum_i x_i(p, e_i) - \bar{e}.$$

Two simplifying normalizations are then commonly adopted. Demand functions derived from optimal choices of price-taking agents are homogeneous of degree zero in prices, so normalize by setting $p_1 \equiv 1$. Normalized prices thus can be taken to range over $\mathbf{R}^{L-1}_{++}$. Next, Walras's Law ensures that excess demand functions are not independent across markets, as $p \cdot z(p, e) = 0$ for each $p \in \mathbf{R}^{L-1}_{++}$. This renders one market clearing equation redundant, and leads to the characterization of equilibria by normalized price vectors $p \in \mathbf{R}^{L-1}_{++}$ such that

$$z_{-L}(p, e) = 0$$

where, adopting common conventions, the subscript –*L* refers to all goods except *L*, so $z_{-L}(p, e) = (z_1(p, e), \ldots, z_{L-1}(p, e))$. Using these normalizations, the equilibrium correspondence can be defined by

$$E(e) := \left\{(x, p) \in \mathbf{R}^{mL}_+ \times \mathbf{R}^{L-1}_{++} : z_{-L}(p, e) = 0, x_i = x_i(p, e) \text{ for } i = 1, \ldots, m\right\}.$$

Fix a particular equilibrium price vector p^* in the economy $E(e)$. One way to answer local comparative statics qsts at this equilibrium is to apply the classical implicit function theorem. If $D_p z_{-L}(p^*, e)$ is invertible, then the implicit function theorem provides several immediate predictions: the equilibrium price p^* is locally unique; locally, on neighbourhoods *W* of *e* and *V* of p^*, the equilibrium price set is described by the graph of a C^1 function $p: W \to \mathbf{R}^{L-1}_{++}$; and local comparative statics are given by the formula

$$Dp(e) = -[D_p z_{-L}(p^*, e)]^{-1} D_e z_{-L}(p^*, e).$$

If this analysis can be performed for each equilibrium, then there are only finitely many equilibria, because the equilibrium set is compact. Moreover, for each equilibrium $(x, p) \in E(e)$ there is a neighborhood U of $(x,\bullet p)$ for which $E(\cdot) \cap U$ has a unique, C^1 selection on a neighbourhood W of e, with the comparative statics derived from the preceding formula. Call such a correspondence *locally* C^1 *at e*. The following definition offers a convenient way to summarize these properties.

Definition 1: The economy $E(e)$ is 'regular' if it has finitely many equilibria, and E is locally C^1 at e.

An alternative way to describe the problem uses the language of differential topology. For a C^1 function $f: \mathbf{R}^m \to \mathbf{R}^n$, $y \in \mathbf{R}^n$ is a *regular value* of f if $Df(x)$ has full rank for every $x \in f^{-1}(y)$. Notice that this is precisely the condition identified above, for the case of equilibrium prices, under which local uniqueness and local comparative statics could be derived from the implicit function theorem. Whenever 0 is a regular value of $z_{-L}(\cdot,e)$, the corresponding economy $E(e)$ is regular. For a fixed function f, a given value y may fail to be a regular value, but almost every other value is regular: this is the conclusion of Sard's theorem. Dually, the fixed value y may fail to be a regular value for a particular function f, but is a regular value for almost every other function. When the set of functions is limited to those drawn from a particular parameterized family, the conclusion remains valid for almost all members of this family provided the parameterization is sufficiently rich. This idea of a rich parameterization can be expressed by requiring y to be a regular value of the parameterized family, and this parametric version of Sard's theorem is typically called the transversality theorem. Figure 1 depicts this idea for smooth excess demand functions.

Figure 1

Generic determinacy for smooth excess demand

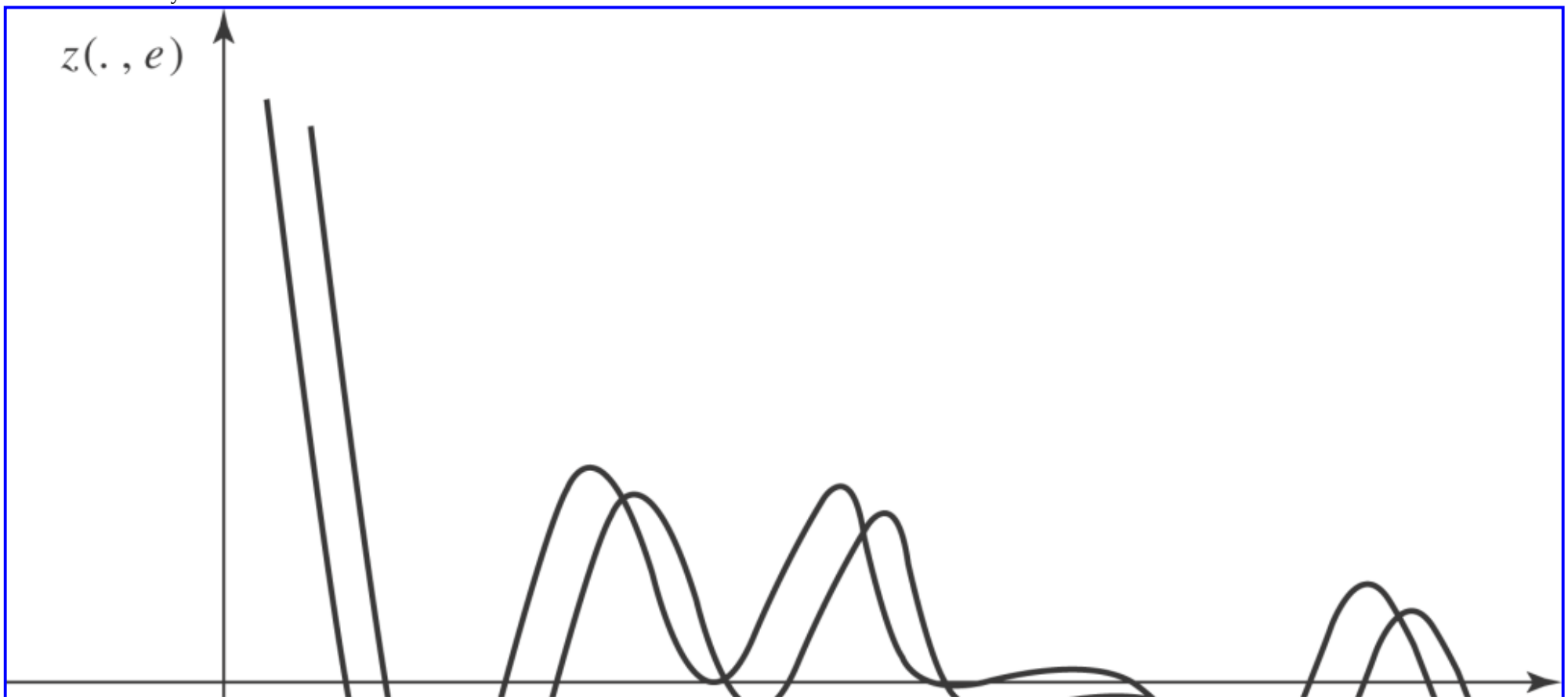

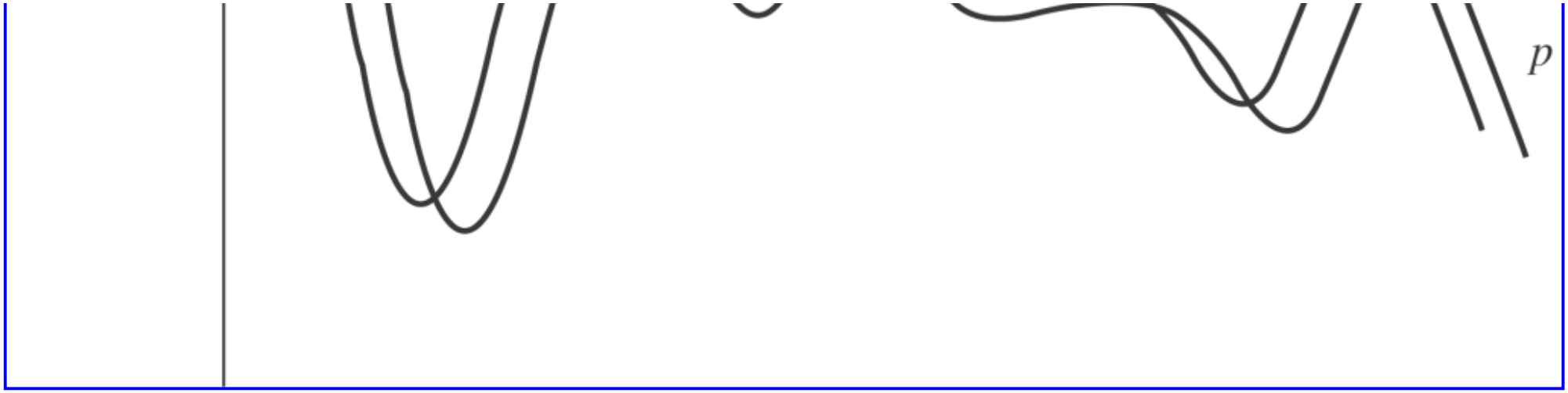

These observations suggest that, while extremely restrictive assumptions might be required to ensure that *every* economy is regular, generic regularity might follow simply from the differentiability of demand functions once the problem is framed this way. Straightforward calculations verify that 0 is a regular value of the excess demand function (viewed as a function of both prices and initial endowment parameters). From the transversality theorem we conclude that there is a subset $R^* \subset \mathbf{R}^{mL}_{++}$ of full Lebesgue measure such that for all $e \in R^*$, $E(e)$ is regular. With the use of additional properties of excess demand and equilibria, it is similarly straightforward to show that the set of regular economies is also open, giving a strong genericity result for regular economies.

This discussion follows Debreu's original development original development closely. This approach takes demand functions as primitives, and gives conditions on individual demand functions under which regularity is a generic feature of exchange economies. To take a step back and start with preferences as primitives, we seek conditions on preferences sufficient to guarantee that individual demand is suitably differentiable. Debreu (1972) addresses this point by introducing a class of ‘smooth preferences’, depicted in Figure 2.

Figure 2
Smooth preferences

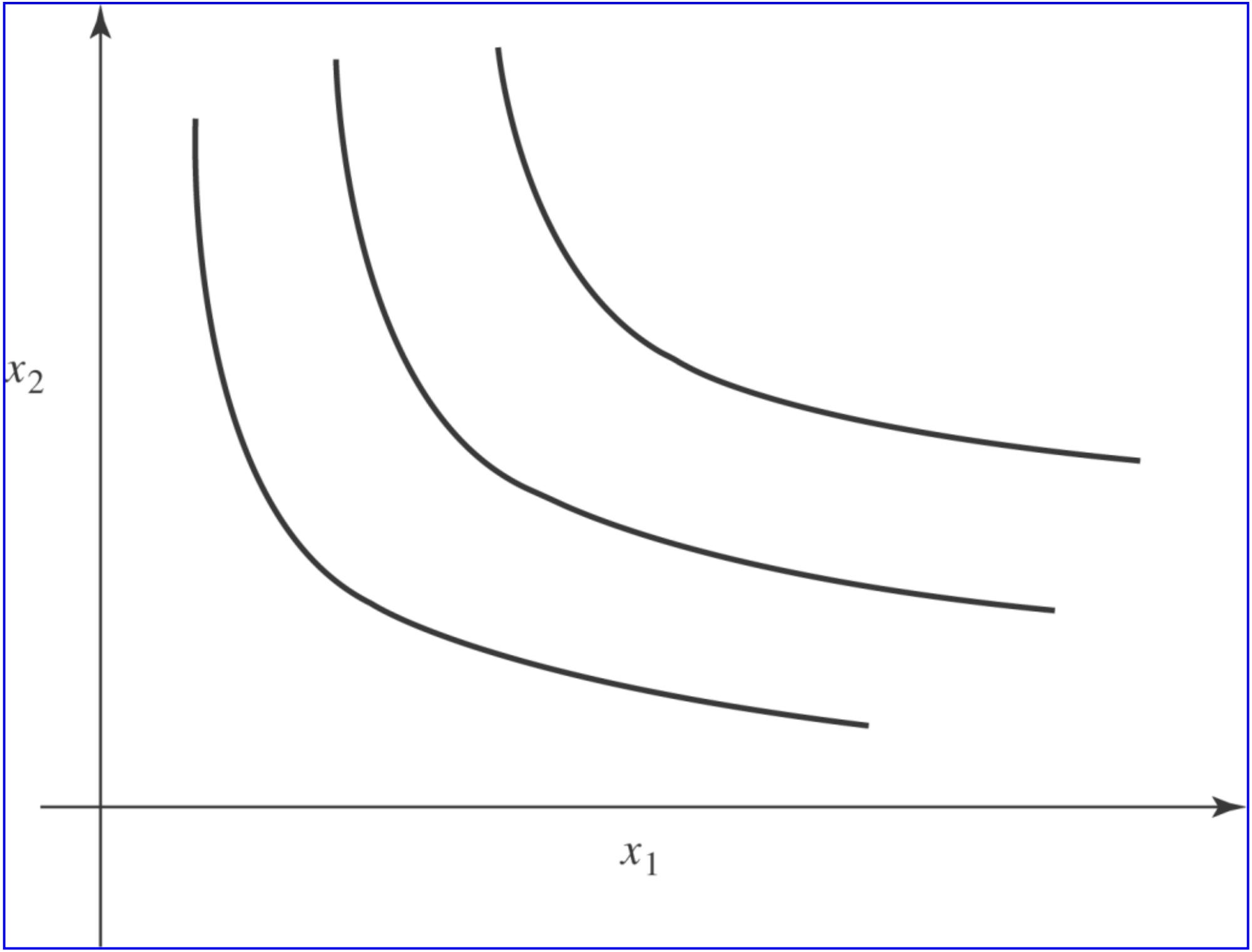
x_2
x_1

Definition 2: The preference order $\succsim$ *on* $\mathbf{R}^L_+$ *is 'smooth' if it is represented by a utility function U such that*

- $U: \mathbf{R}^L_+ \to \mathbf{R}$ *is* C^2 *on* $\mathbf{R}^L_{++}$
- *for each* $x \in \mathbf{R}^L_{++}$, $\{y \in \mathbf{R}^L_+ : y \sim x\} \subset \mathbf{R}^L_{++}$
- *for each* $x \in \mathbf{R}^L_{++}$, $DU(x) \gg 0$
- *for each* $x \in \mathbf{R}^L_{++}$, $D^2U(x)$ *is negative definite on ker* $DU(x) := \{z \in \mathbf{R}^L : DU(x) \cdot z = 0\}$

Fairly straightforward arguments, again using the implicit function theorem, establish that individual demand functions derived from smooth preferences are C^1. Putting all of these results together yields:

Theorem 1: Let $\succsim_i$ *be a smooth preference order on* $\mathbf{R}^L_+$ *for each* $i = 1, \ldots, m$. *There exists an open set* $R^* \subset \mathbf{R}^{mL}_{++}$ *of full Lebesgue measure such that for all* $e \in R^*$, $E(e)$ *is regular*.

3 Determinacy and indeterminacy: a new approach to many problems

Behind this result for equilibria in finite exchange economies is a broad, powerful, and simple principle that has found many important and ingenious applications in the 35 years since Debreu's original 1970 paper. To cast the problem more generally, take a parameterized family of equations, captured by a function $f: \mathbf{R}^m \times \mathbf{R}^k \to \mathbf{R}^n$. This describes a problem with *m* variables and *k* parameters simultaneously entering *n* different equations. Imagine that for each parameter value $r \in \mathbf{R}^k$,

$$E(r) := \{x \in \mathbf{R}^m : f(x, r) = 0\}$$

gives the set of objects of interest. Moreover, imagine that the equations are sufficiently independent in determining the solutions, in the sense that 0 is a regular value of *f*. Counting the number of equations and unknowns produces three distinct cases, corresponding in turn to three different sorts of applications.

In the canonical case exemplified by the simple exchange economy described above, the number of relevant endogenous variables, *m*, is equal to the number of equations, *n*. In this case, 0 being a regular value of *f* characterizes exactly the case in which the equations are sufficiently independent that the loose 'counting equations and unknowns' heuristic corresponds with the precise technical result of generic determinacy. One prominent illustration of this case is given by two-period incomplete markets models with real assets, that is, assets that pay off in bundles of commodities. In these models, there are as many distinct budget equations as there are states. If we let *S* denote the number of states, this means there are *S*+1 distinct Walras's Law statements, leading to *S*+1 redundant market clearing equations. Because asset payoffs are in real terms, all budget constraints are homogeneous of degree 0 in state prices. This generates $S + 1$ distinct normalizations of state prices, compensating exactly for the drop in independent market clearing equations determining equilibrium. Generic determinacy in this case is established by Geanakoplos and Polemarchakis (1987).

When $m < n$, there are fewer equations than unknowns, and the regularity of the system of equations means that it is generically overdetermined. In this case, generically it is impossible to satisfy the equations simultaneously, that is, generically *E*(*r*) is empty. As a simple example of this argument, consider the prevalence of trade at equilibrium in an Edgeworth box economy. One market-clearing condition in one (normalized) price characterizes equilibria, and standard arguments show that this excess demand function has 0 as a regular value. In fact, varying the endowment of the first agent alone is enough. How often does equilibrium involve trade in some goods? With only two agents, trade occurs in equilibrium if and only if $x_2 \neq e_2$, so the additional two equations $x_2(p, e) - e_2 = 0$ characterize endowment and price combinations for which there is no trade in equilibrium. A simple calculation shows that 0 is a regular value of $f(p, e) := (z_{-2}(p, e), x_2(p, e) - e_2)$. Fixing the endowment profile *e*, however, this is a problem with three equations in a single variable, so there must be a set $R^{**} \subset \mathbf{R}^{mL}_{++}$ of full Lebesgue measure such that for every $e \in R^{**}$, there are no solutions to the equation $f(p, e) = 0$. For every endowment profile $e \in R^{**}$, every equilibrium then must involve trade, as every equilibrium price solves the first equation $z_{-2}(p^*, e) = 0$, so cannot also involve no trade, $x^2(p^*, e) \neq e_2$. Similar logic but more involved calculations show that equilibrium allocations are generically inefficient in incomplete markets models, and generically constrained inefficient in

multi-good incomplete markets models. Geanakoplos and Polemarchakis (1987) pioneered this approach to efficiency with incomplete markets.

Finally, when $m > n$, generically indeterminacy arises, as generically the solution set $E(r)$ is an $(m - n)$-dimensional manifold. (A subset $M \subset \mathbf{R}^m$ is a d-dimensional C^ℓ manifold if for each $x \in M$ there exist open sets $V \subset \mathbf{R}^m$ and $W \subset \mathbf{R}^d$, where V is a neighbourhood of x, and a C^ℓ diffeomorphism $\varphi: V \to W$ such that $\varphi(V \cap M) = W$.) In this case, generically there is a continuum of solutions, and the set of solutions is locally, up to diffeomorphism, a set of dimension $m - n$. An important example of this case is provided by two-period incomplete financial markets models with nominal assets. Here, asset payoffs are in nominal terms, in some specified unit of account. As in the case of real assets described above, there are $S + 1$ independent budget constraints when there are S possible states of nature, so there are $S + 1$ redundant market clearing equations. Because asset payoffs are nominal, however, budget constraints are not all homogeneous of degree zero, and price levels matter. With only two homogeneity conditions, one for period one prices and one relating all commodity and asset prices, this leaves $S - 1$ dimensions of indeterminacy in equilibria generically. The detailed result is established by Geanakoplos and Mas-Colell (1989).

These three cases, and the generic properties of solution sets that follow, are collected below.

Theorem 2: Let $f: \mathbf{R}^m \times \mathbf{R}^k \to \mathbf{R}^n$ be a C^ℓ function, where $\ell > \max\{m - n, 0\}$, and suppose 0 is a regular value of f.

1. (a) *Suppose m=n. There exists a set $R^* \subset \mathbf{R}^k$ of full Lebesgue measure such that, for every $r \in R^*$, E(r) contains only isolated points, E(r) is finite when compact, and E is locally C^1 at r.*
2. (b) *Suppose m<n. There exists a set $R^* \subset \mathbf{R}^k$ of full Lebesgue measure such that for every $r \in R^*$, E(r) is empty.*
3. (c) *Suppose m>n. There exists a set $R^* \subset \mathbf{R}^k$ of full Lebesgue measure such that for every $r \in R^*$, E(r) is an (m–n)-dimensional C^ℓ manifold.*

The techniques pioneered by Debreu have found widespread applications, and have proven to be rmkably powerful. Nonetheless, the smoothness needed to study determinacy using the tools of differential topology does stem from assumptions that often carry real economic content. These assumptions restrict both the nature of admissible preferences and the nature of admissible constraints.

For example, to avoid problems arising when non-negativity constraints on consumption may become binding, these results rest on ‘boundary’ restrictions, both on endowments, because individual endowments are strictly positive, or on equilibrium consumption via boundary conditions on preferences that imply individual demands are strictly positive at all prices. Unless goods are aggregated extremely coarsely, neither pattern is supported by observations on consumer behaviour or characteristics. Relaxing the constraint on endowments turns out, perhaps surprisingly, to generate indeterminacy much more readily than relaxing the assumptions on positive consumptions, or incorporating other more general constraints on choices. Minehart (1997) shows by means of an example that for one natural case of restricted endowments, in which each agent is constrained to hold a single, individual-specific, good, an open subset of such parameters leads to indeterminacy in equilibrium. Highlighting the fact that the choice of parameterization can be important, Mas-Colell (1985) shows that this conclusion is not robust to perturbations in preferences; generic determinacy, in a topological sense, is restored by considering variations in preferences as well as constrained endowments. If the assumption that individual endowments of every good are positive is maintained, the restriction to positive individual demand for every good can be relaxed. For example, Mas-Colell (1985) provides generic determinacy results for exchange economies allowing for boundary consumptions; Figure 3 depicts such preferences.

Figure 3

Preferences allowing boundary consumption

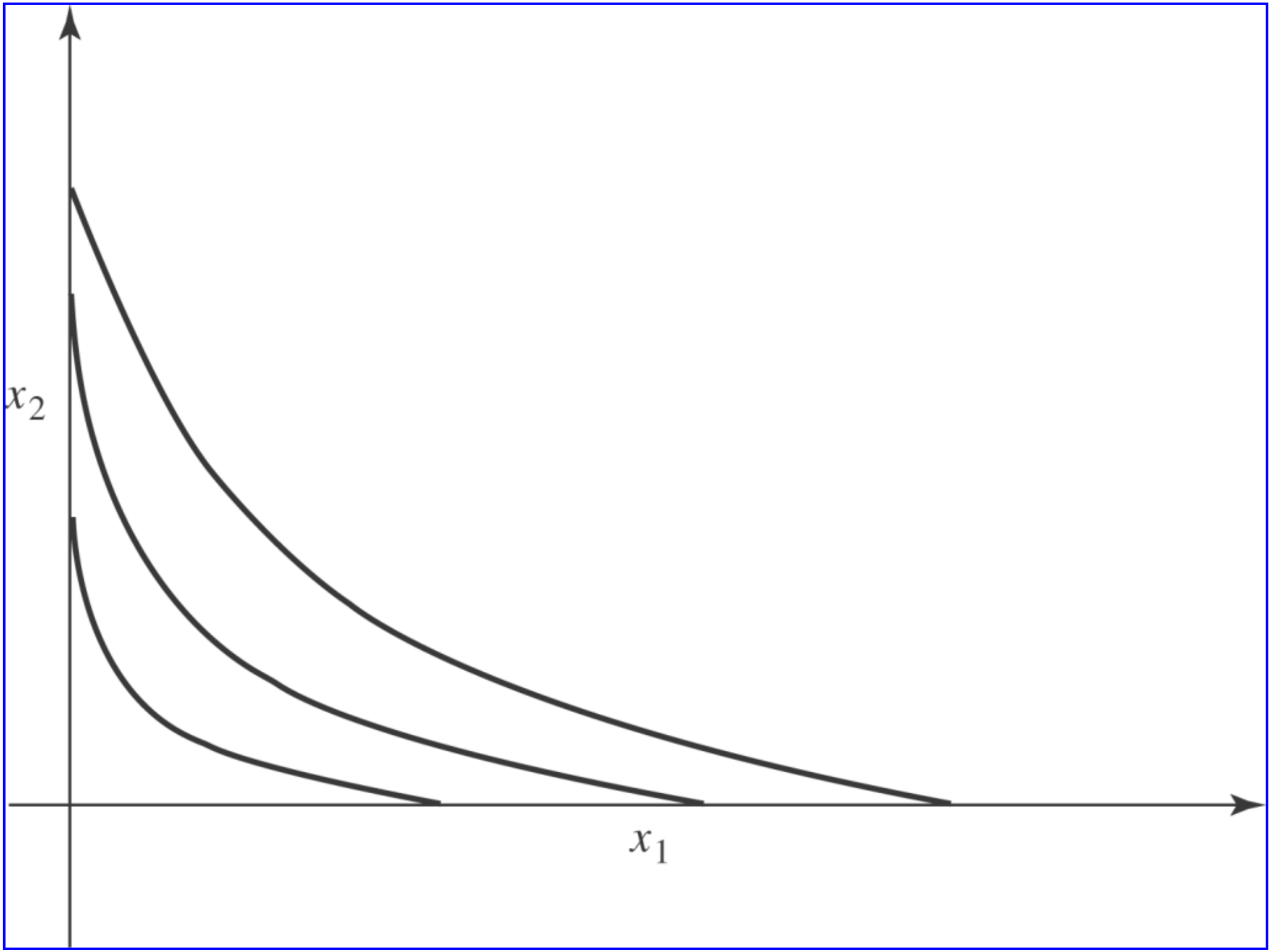

Smooth preferences, as defined by Definition 2 above, obviously rule out preferences with non-differentiabilities in level sets, a restriction that also has important behavioural content.

Kinks have arisen as central manifestations of various behavioural phenomena, including loss aversion, ambiguity aversion, and reference dependence; examples include Kahneman and Tversky (1979), Tversky and Kahneman (1991), Koszegi and Rabin (2006), Sagi (2006), and Gilboa and Schmeidler (1989). Such kinks typically lead to excess demand functions that fail to be differentiable for some prices. Rader (1973), Pascoa and Werlang (1999), Shannon (1994), and Blume and Zame (1993) all develop methods to address such cases. With the exception of Blume and Zame (1993), these techniques can be roughly understood as expanding differential notions by adding to ‘regularity’ the condition that the function (for example, excess demand) is differentiable at every solution, and establishing that analogues of implicit function theorems, Sard's theorem or the transversality theorem remain valid for sufficiently nice non-smooth functions, such as Lipschitz continuous functions; in particular, see Shannon (1994; 2006). Blume and Zame (1993) instead use results that exploit the structure of algebraic sets to establish generic determinacy for utilities that are, roughly, finitely piecewise analytic, and need not be strictly concave. Examples in which determinacy has been studied using techniques along these various lines include asset market models with restricted participation (for example, see Cass, Siconolfi, and Villanacci, 2001) and models of ambiguity aversion (for example, see Rigotti and Shannon, 2006). Figure 4.

Figure 4

Generic determinacy for non-smooth excess demand

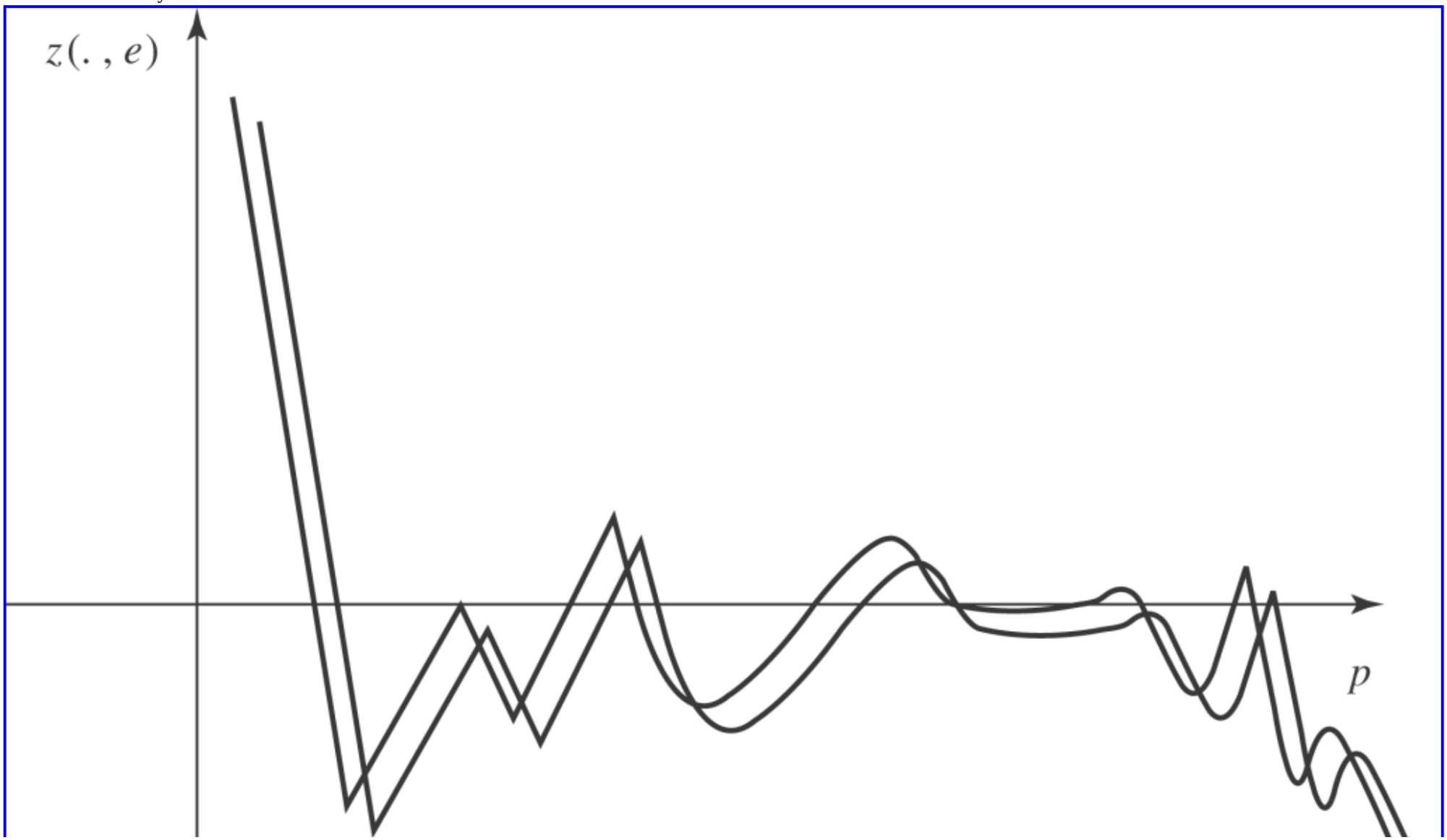

4 Determinacy in infinite-dimensional economies

Many economic models require an infinite number of marketed commodities. Important examples include dynamic infinite horizon economies, continuous-time trading in financial markets, and markets with differentiated commodities. Such infinite-dimensional models present big obstacles to studying determinacy, starting with the fact that individual demand is not defined for most prices, precluding any straightforward parallel of Debreu's arguments for finite economies. In addition, the positive cone in most infinite-dimensional spaces has empty interior in the relevant topologies, meaning individual consumption sets are ‘all boundaries’, and existence of equilibrium typically requires conditions, such as uniform properness or variants, that effectively bound marginal rates of substitution. Thus boundary conditions akin to those in Debreu's smooth preferences are likely either to be impossible to satisfy or to contradict equilibrium existence in many important applications.

Provided there are finitely many agents and no market distortions, using the welfare theorems and Negishi's argument provides an alternative characterization of equilibria, replacing excess demand with ‘excess savings’. Some version of this characterization of equilibria provides the framework for much of the existing equilibrium analysis in economies with infinitely many commodities, including the seminal work on existence of Mas-Colell (1986) and Aliprantis, Brown and Burkinshaw (1987), and the approach to determinacy for discrete-time infinite horizon models with time separability pioneered by Kehoe and Levine (1985). To explain this, let X denote the commodity space. The efficient allocations are the solutions to a social planner's problem of the following form: given $\lambda \in \Lambda := \left\{\lambda \in \mathbf{R}_+^m : \Sigma_{i=1}^m \lambda_i = 1\right\}$, choose a feasible allocation $x(\lambda)$ to solve:

$$\max \sum_{i=1}^{m} \lambda_i U_i(x_i) \; s.t. \sum_{i=1}^{m} x_i \leq \bar{e}.$$

Under standard assumptions, the solution $x(\lambda)$ to this problem is well-defined and unique for each $\lambda \in \Lambda$, and a unique price $p(\lambda)$ supporting $x(\lambda)$ can be characterized. Equilibria then correspond to the solutions λ to the budget equations

$$\begin{gathered} p(\lambda) \cdot (x_2(\lambda) - e_2) = 0 \\ \vdots \\ p(\lambda) \cdot (x_m(\lambda) - e_m) = 0 \end{gathered}$$

where Walras's Law accounts for the missing equation. In parallel with excess demand, define the *excess savings map* $s: \Lambda \times X_+^m \to \mathbf{R}^{m-1}$

$$s(\lambda, e) := (p(\lambda) \cdot (x_2(\lambda) - e_2), \ldots, p(\lambda) \cdot (x_m(\lambda) - e_m)).$$

Through this construction, the question of determinacy for infinite-dimensional economies can be cast in close parallel to finite economies, with the only change that the set of parameters is now infinite-dimensional. This raises several technical issues, most importantly the choice between topological and measure-theoretic notions of genericity due to the impossibility of defining a suitable analogue of Lebesgue measure in infinite-dimensional spaces (see Hunt, Sauer and Yorke, 1992, and Anderson and Zame, 2001, for a discussion of these issues). This construction also makes imperative the need to link conditions on excess savings used to imply determinacy with conditions on preferences since, in contrast with excess demand, excess savings depends on artificial and unobservable constructs. Somewhat surprisingly, Shannon (1999) and Shannon and Zame (2002) show that generic determinacy follows from conditions on preferences that closely resemble Debreu's (1972) smooth preferences, after suitable renormalization. As in the finite case, these conditions can roughly be understood as strengthened notions of concavity, requiring that near feasible bundles utility differs from a linear approximation by an amount quadratic in the distance to the given bundle. These notions of concavity thus rule out preferences displaying local or global substitutes. Shannon and Zame (2002) provide a simple geometric argument showing that the excess spending mapping is Lipschitz continuous. Generic determinacy then follows by arguments similar to those sketched above for other problems with non-differentiabilities, making use of Shannon (2006) on comparative statics and a version of the transversality theorem for this setting. The direct, geometric nature of these arguments render them applicable in a wide range of examples, including models of continuous-time trading, trading in differentiated commodities, and trading over an infinite horizon.

Bibliography

Aliprantis, C., Brown, D.J. and Burkinshaw, O. 1987. Edgeworth equilibria. *Econometrica* 55, 1108–38.

Anderson, R. and Zame, W.R. 2001. Genericity with infinitely many parameters. *Advances in Theoretical Economics* 1.

Blume, L. and Zame, W.R. 1993. The algebraic geometry of competitive equilibrium. In *Essays in General Equilibrium and International Trade: In Memoriam Trout Rader*, ed. W. Neuefeind. New York: Springer-Verlag.

Cass, D., Siconolfi, P. and Villanacci, A. 2001. Generic regularity of competitive equilibrium with restricted participation on financial markets. *Journal of Mathematical Economics* 36, 61–76.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40, 603–15.

Dierker, E. 1972. Two remarks on the number of equilibria of an economy. *Econometrica* 40, 951–53.

Geanakoplos, J. and Mas-Colell, A. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47, 22–38.

Geanakoplos, J. and Polemarchakis, H. 1987. Existence, regularity and constrained suboptimality of competitive portfolio allocations when the asset market is incomplete. In *Essays in Honor of Kenneth J. Arrow*, vol. 3, ed. W.P. Heller, R.M. Starr and D.M. Starrett. Cambridge: Cambridge University Press.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Hunt, B.R., Sauer, T. and Yorke, J.A. 1992. Prevalence: a translation invariant ‘almost every’ on infinite dimensional spaces. *Bulletin (New Series) of the American Mathematical Society* 27, 217–38.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Kehoe, T. and Levine, D. 1985. Comparative statics and perfect foresight in infinite horizon economies. *Econometrica* 53, 433–52.

Koszegi, B. and Rabin, M. 2006. A model of reference-dependent preferences. *Quarterly Journal of Economics* 121, 1133–65.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: A Differentiable Approach*. Cambridge: Cambridge University Press.

Mas-Colell, A. 1986. The price equilibrium existence problem in topological vector lattices. *Econometrica* 54, 1039–54.

Minehart, D. 1997. A note on the generic finiteness of the set of equilibria in an exchange economy with constrained endowments. *Mathematical Social Sciences* 34, 75–80.

Pascoa, M.R. and Werlang, S. 1999. Determinacy of equilibrium in nonsmooth economies. *Journal of Mathematical Economics* 32, 289–302.

Rader, J.T. 1973. Nice demand functions. *Econometrica* 41, 913–35.

Rigotti, L. and Shannon, C. 2006. Sharing risk and ambiguity. Discussion paper, UC Berkeley.

Sagi, J. 2006. Anchored preference relations. *Journal of Economic Theory* 130, 283–95.

Shannon, C. 1994. Regular nonsmooth equations. *Journal of Mathematical Economics* 23, 147–66.

Shannon, C. 1999. Determinacy of competitive equilibria in economies with many commodities. *Economic Theory* 14, 29–87.

Shannon, C. 2006. A prevalent transversality theorem for Lipschitz functions. *Proceedings of the American Mathematical Society* 134, 2755–65.

Shannon, C. and Zame, W.R. 2002. Quadratic concavity and determinacy of equilibrium. *Econometrica* 70, 631–62.

Tversky, A. and Kahneman, D. 1991. Loss aversion in riskless choice: a reference dependent model. *Quarterly Journal of Economics* 106, 1039–61.

How to cite this article

Shannon, Chris. "determinacy and indeterminacy of equilibria." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000248> doi:10.1057/9780230226203.0381

The New Palgrave Dictionary of Economics Online

deterministic evolutionary dynamics

William H. Sandholm
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We review the literature on deterministic evolutionary dynamics in game theory. We describe the micro-foundations of dynamic evolutionary models and offer some basic examples. We report on stability theory for evolutionary dynamics, and we discuss the senses in which evolutionary dynamics support and fail to support traditional game-theoretic solution concepts.

Keywords

asymptotic stability; convergence; deterministic evolutionary dynamics; evolutionarily stable strategies; evolutionary dynamics; extensive form games; game theory; imitative dynamics; index theory; Lyapunov functions; Markov processes; mean dynamic; Poincaré–Hopf theorem; population games; predictions; refinements of Nash equilibrium; replicator dynamic; revision protocols

Article

1 Introduction

Deterministic evolutionary dynamics for games first appeared in the mathematical biology literature, where Taylor and Jonker (1978) introduced the *replicator dynamic* to provide an explicitly dynamic foundation for the static evolutionary stability concept of Maynard Smith and Price (1973). But one can find precursors to this approach in the beginnings of game theory: Brown and von Neumann (1950) introduced differential equations as a tool for computing equilibria of zero-sum games. In fact, the replicator dynamic appeared in the mathematical biology literature long before game theory itself: while Maynard Smith and Price (1973) and Taylor and Jonker (1978) studied game theoretic models of animal conflict, the replicator equation is equivalent to much older models from population ecology and population genetics. These connections are explained by Schuster and Sigmund (1983), who also coined the name ‘replicator dynamic’, borrowing the word ‘replicator’ from Dawkins (1982).
In economics, the initial phase of research on deterministic evolutionary dynamics in the late 1980s and early 1990s focused on populations of agents who are randomly matched to play normal form games, with evolution

described by the replicator dynamic or other closely related dynamics. The motivation behind the dynamics continued to be essentially biological: individual agents are preprogrammed to play specific strategies, and the dynamics themselves are driven by differences in birth and death rates. Since that time the purview of the literature has broadened considerably, allowing more general sorts of large population interactions, and admitting dynamics derived from explicit models of active myopic decision making.

This article provides a brief overview of deterministic evolutionary dynamics in game theory. More detailed treatments of topics introduced here can be found in the recent survey article by Hofbauer and Sigmund (2003), and in books by Maynard Smith (1982), Hofbauer and Sigmund (1988; 1998), Weibull (1995), Vega-Redondo (1996), Samuelson (1997), Fudenberg and Levine (1998), Cressman (2003), and Sandholm (2007).

2 Population games

Population games provide a general model of strategic interactions among large numbers of anonymous agents. For simplicity, we focus on games played by a single population, in which agents are not differentiated by roles; allowing for multiple populations is mostly a matter of introducing more elaborate notation.

In a single-population game, each agent from a unit-mass population chooses a strategy from the finite set $S = \{1, \ldots, n\}$, with typical elements *i* and *j*. The distribution of strategy choices at a given moment in time is described by a *population state* $x \in X = \left\{x \in \mathbf{R}^n_+ : \Sigma_{i \in S} x_i = 1\right\}$. The *payoff* to strategy *i*, denoted $F_i: X \to \mathbf{R}$, is a continuous function of the population state; we use the notation $F: X \to \mathbf{R}^n$ to refer to all strategies' payoffs at once. By taking the set of strategies *S* as fixed, we can refer to *F* itself as a *population game*.

The simplest example of a population game is the most commonly studied one: random matching to play a symmetric normal form game $A \in \mathbf{R}^{n \times n}$, where A_{ij} is the payoff obtained by an agent choosing strategy *i* when his opponent chooses strategy *j*. When the population state is $x \in X$, the expected payoff to strategy *i* is simply the weighted average of the elements of the *i*th row of the payoff matrix: $F_i(x) = \Sigma_{j \in S} A_{ij} x_j = (Ax)_i$. Thus, the population game generated by random matching in *A* is the linear population game $F(x) = Ax$.

Many models of strategic interactions in large populations that arise in applications do not take this simple linear form. For example, in models of highway congestion, payoff functions are convex: increases in traffic when traffic levels are low have virtually no effect on delays, while increases in traffic when traffic levels are high increase delays substantially (see Beckmann, McGuire and Winsten, 1956; Sandholm, 2001). Happily, allowing nonlinear payoffs extends the range of possible applications of population games without making evolutionary dynamics especially more difficult to analyse, since the dynamics themselves are nonlinear even when the underlying payoffs are not.

3 Foundations of evolutionary dynamics

Formally, an *evolutionary dynamic* is a map that assigns to each population game *F* a differential equation $\dot{x} = V^F(x)$ on the state space *X*. While one can define evolutionary dynamics directly, it is preferable to derive them from explicit models of myopic individual choice.

We can accomplish this by introducing the notion of a *revision protocol* $\rho: \mathbf{R}^n \times X \to \mathbf{R}^{n \times n}_+$. Given a payoff vector *F*(*x*) and a population state *x*, a revision protocol specifies for each pair of strategies *i* and *j* a non-negative number ρ_{ij}•(*F*(*x*), *x*), representing the rate at which strategy *i* players who are considering switching strategies switch to strategy *j*. Revision protocols that are most consistent with the evolutionary paradigm require agents to possess only limited information: for example, a revising agent might know only the current

payoffs of his own strategy *i* and his candidate strategy *j*.

A given revision protocol can admit a variety of interpretations. For one all-purpose interpretation, suppose each agent is equipped with an exponential alarm clock. When the clock belonging to an agent playing strategy *i* rings, he selects a strategy $j \in S$ at random, and then switches to this strategy with probability proportional to $\rho_{ij} \bullet (F(x), x)$. While this interpretation is always available, others may be simpler in certain instances. For example, if the revision protocol is of the imitative form $\rho_{ij} = x_j \times \hat{\rho}_{ij}$, we can incorporate the x_j term into our story by supposing that the revising agent selects his candidate strategy *j* not by drawing a strategy at random, but by drawing an opponent at random and observing this opponent's strategy.

A population game *F* and a revision protocol ρ together generate an ordinary differential equation $\dot{x} = V^F(x)$ on the state space *X*. This equation, which captures the population's *expected* motion under *F* and ρ, is known as the *mean dynamic* or *mean field* for *F* and ρ:

$$\dot{x} = V_i^F(x) = \sum_{j \in S} x_j \rho_{ji}(F(x), x) - x_i \sum_{j \in S} \rho_{ij}(F(x), x). \qquad \text{(M)}$$

The form of the mean dynamic is easy to explain. The first term describes the ‘inflow’ into strategy *i* from other strategies; it is obtained by multiplying the mass of agents playing each strategy *j* by the rate at which such agents switch to strategy *i*, and then summing over *j*. Similarly, the second term describes the ‘outflow’ from strategy *i* to other strategies. The difference between these terms is the net rate of change in the use of strategy *i*.

To obtain a formal link between the mean dynamic (M) and our model of individual choice, imagine that the population game *F* is played not by a continuous mass of agents but rather by a large, finite population with *N* members. Then the model described above defines a Markov process $\{X_t^N\}$ on a fine but discrete grid in the state space *X*. The foundations for deterministic evolutionary dynamics are provided by the following finite horizon deterministic approximation theorem: Fix a time horizon $T < \infty$. Then the behaviour of the stochastic process $\{X_t^N\}$ through time *T* is approximated by a solution of the mean dynamic (M); the approximation is uniformly good with probability close to 1 once the population size *N* is large enough. (For a formal statement of this result, see Benaïm and Weibull, 2003.)

In cases where one is interested in phenomena that occur over very long time horizons, it may be more appropriate to consider the infinite horizon behaviour of the stochastic process $\{X_t^N\}$. Over this infinite time horizon, the deterministic approximation fails, as a correct analysis must explicitly account for the stochastic nature of the evolutionary process. For more on the distinction between the two time scales, see Benaïm and Weibull (2003).

4 Examples and families of evolutionary dynamics

We now describe revision protocols that generate some of the most commonly studied evolutionary dynamics. In the table below, $\bar{F}(x) = \sum_{i \in S} x_i F_i(x)$ represents the population's average payoff at state *x*, and

$B^F(x) = \operatorname{argmax}_{y \in X} y' F(x)$ is the best response correspondence for the game *F*.

A common critique of evolutionary analysis of games is that the choice of a specific revision protocol, and hence the evolutionary analysis that follows, is necessarily arbitrary. There is surely some truth to this criticism: to the extent that one's analysis is sensitive to the fine details of the choice of protocol, the conclusions of the analysis are cast into doubt. But much of the force of this critique is dispelled by this important observation: *evolutionary dynamics based on qualitatively similar revision protocols lead to qualitatively similar aggregate behaviour*. We call a collection of dynamics generated by similar revision protocols a ‘family’ of evolutionary dynamics.

To take one example, many properties that hold for the replicator dynamic also hold for dynamics based on revision protocols of the form $\rho_{ij} = x_j \hat{\rho}_{ij}$ where $\hat{\rho}_{ij}$ satisfies

$$\operatorname{sgn}\left((\hat{\rho}_{ki} - \hat{\rho}_{ik}) - (\hat{\rho}_{kj} - \hat{\rho}_{jk})\right) = \operatorname{sgn}(F_i - F_j) \text{for all } k \in S.$$

(In words: if *i* earns a higher payoff than *j*, then the net conditional switch rate from *k* to *i* is higher than that from *k* to *j* for all $k \in S$.) For reasons described in Section 3, dynamics generated in this way are called ‘imitative dynamics’. (See Björnerstedt and Weibull, 1996, for a related formulation.) For another example, most properties of the pairwise difference dynamic remain true for dynamics based on protocols of the form $\rho_{ij} = \phi(F_i - F_j)$, where $\phi : \mathbf{R} \to \mathbf{R}_+$ satisfies sign-preservation:

$$\operatorname{sgn}(\phi(d)) = \operatorname{sgn}([d]_+).$$

Dynamics in this family are called ‘pairwise comparison dynamics’. For more on these and other families of dynamics, see Sandholm (2007, ch. 5).

5 Rest points and local stability

Having introduced families of evolutionary dynamics, we now turn to questions of prediction: if agents playing game *F* follow the revision protocol ρ (or, more broadly, a revision protocol from a given family), what predictions can we make about how they will play the game? To what extent do these predictions accord with those provided by traditional game theory?

A natural first question to ask concerns the relationship between the rest points of an evolutionary dynamic V^F and the Nash equilibria of the underlying game *F*. In fact, one can prove for a very wide range of evolutionary dynamics that if a state $x^* \in X$ is a Nash equilibrium (that is, if $x \in B(x)$), then x^* is a rest point as well.

One way to show that $NE(F) \subseteq RP(V^F)$ is to first establish a *monotonicity* property for V^F: that is, a property that relates strategies’ growth rates under V^F with their payoffs in the underlying game (see, for example, Nachbar, 1990; Friedman, 1991; and Weibull, 1995). The most general such property, first studied by Friedman (1991) and Swinkels (1993), we call ‘positive correlation’:

$$\text{If } x \notin RP(V^F), \text{ then } F(x)'V^F(x) > 0. \quad \text{(PC)}$$

Property (PC) is equivalent to requiring a positive correlation between strategies' growth rates $V_i^F(x)$ and payoffs $F_i(x)$ (where the underlying probability measure is the uniform measure on the strategy set S). This property is satisfied by the first three dynamics in Table 1, and modifications of it hold for the remaining two as well. Moreover, it is not difficult to show that if V^F satisfies (PC), then all Nash equilibria of F are rest points of V^F: that is, $NE(F) \subseteq RP(V^F)$, as desired (see Sandholm, 2007, ch. 5).

Revision protocol	Evolutionary dynamic	Name	Origin
$\rho_{ij} = x_j(K - F_i)$, or $\rho_{ij} = x_j(K + F_j)$, or $\rho_{ij} = x_j[F_j - F_i]_+$	$\dot{x}_i = x_i(F_i(x) - \bar{F}(x))$	Replicator	Taylor and Jonker (1978)
$\rho_{ij} = [F_j - \bar{F}]_+$	$\dot{x}_i = [F_i(x) - \bar{F}(x)]_+ - x_i \Sigma_{j \in S}[F_j(x) - \bar{F}(x)]_+$	Brown–von Neumann–Nash (BNN)	Brown and von Neumann (1950)
$\rho_{ij} = [F_j - F_i]_+$	$\dot{x}_i = \Sigma_{j \in S} x_j [F_i(x) - F_j(x)]_+ - x_i \Sigma_{j \in S}[F_i(x) - F_j(x)]_+$	Pairwise difference (PD)	Smith (1984)
$\rho_{ij} = \dfrac{\exp(\eta^{-1}F_j)}{\Sigma_{k \in S}\exp(\eta^{-1}F_k)}$	$\dot{x}_i = \dfrac{\exp(\eta^{-1}F_i(x))}{\Sigma_{k \in S}\exp(\eta^{-1}F_k(x))} x_i$	Logit	Fudenberg and Levine (1998)
$\rho_{ij} = B_i^F(x)$	$\dot{x} = B_i^F(x) - x_i$	Best response	Gilboa and Matsui (1991)

In many cases, one can also prove that every rest point of V^F is a Nash equilibrium of F, and hence that $NE(F) = RP(V^F)$. In fact, versions of this statement are true for all of the dynamics introduced above, with the notable exception of the replicator dynamic and other imitative dynamics. The reason for this failure is easy to see: when revisions are based on imitation, unused strategies, even ones that are optimal, are never chosen. On the other hand, if we introduce a small number of agents playing an unused optimal strategy, then these agents will be imitated. Developing this logic, Bomze (1986) and Nachbar (1990) show that, under many imitative dynamics, every Lyapunov stable rest point is a Nash equilibrium.

As we noted at the onset, the original motivation for the replicator dynamic was to provide a foundation for Maynard Smith and Price's (1973) notion of an evolutionarily stable strategy (ESS). Hofbauer, Schuster and Sigmund (1979) and Zeeman (1980) show that an ESS is asymptotically stable under the replicator dynamic, but that an asymptotically state need not be an ESS.

More generally, when is a Nash equilibrium a dynamically stable rest point, and under which dynamics? Under differentiable dynamics, stability of isolated equilibria can often be determined by linearizing the dynamic around the equilibrium. In many cases, the question of the stability of the rest point x^* reduces to a question of the negativity of certain eigenvalues of the Jacobian matrix $DF(x^*)$ of the payoff vector field. In non-differentiable cases, and in cases where the equilibria in question form a connected component, stability can sometimes be established by using another standard approach: the construction of suitable Lyapunov functions. For an overview of work in these directions, see Sandholm (2007, ch. 6).

In the context of random matching in normal form games, it is natural to ask whether an equilibrium that is stable under an evolutionary dynamic also satisfies the restrictions proposed in the equilibrium refinements literature. Swinkels (1993) and Demichelis and Ritzberger (2003) show that this is true in great generality under even the most demanding refinements: in particular, any component of rest points that is asymptotically stable under a dynamic that respects condition (PC) contains a strategically stable set in the sense of Kohlberg and Mertens (1986). While proving this result is difficult, the idea behind the result is simple. If a component is asymptotically stable under an evolutionary dynamic, then this dynamic stability ought not to be affected by slight perturbations of the payoffs of the game. *A fortiori*, the *existence* of the component ought not to be affected by the payoff perturbations either. But this preservation of existence is precisely what strategic stability demands.

This argument also shows that asymptotic stability under evolutionary dynamics is a qualitatively stronger requirement than strategic stability: while strategic stability requires equilibria not to vanish after payoff perturbations, it does not demand that they be attracting under a disequilibrium adjustment process. For example, while all Nash equilibria of simple coordination games are strategically stable, only the pure Nash equilibria are stable under evolutionary dynamics.

Demichelis and Ritzberger (2003) establish their results using tools from index theory. Given an evolutionary dynamic V^F for a game F, one can assign each component of rest points an integer, called the *index*, that is determined by the behaviour of the dynamic in a neighbourhood of the rest point; for instance, regular, stable rest points are assigned an index of 1. The set of all indices for the dynamic V^F is constrained by the *Poincaré–Hopf theorem*, which tells us that the sum of the indices of the equilibrium components of V^F must equal 1. As a consequence of this deep topological result, one can sometimes determine the local stability of one component of rest points by evaluating the local stability of the others.

6 Global convergence: positive and negative results

To provide the most satisfying evolutionary justification for the prediction of Nash equilibrium play, it is not enough to link the rest points of a dynamic and the Nash equilibria of the underlying game, or to prove local stability results. Rather, one must establish convergence to Nash equilibrium from *arbitrary* initial conditions. One way to proceed is to focus on a class of games defined by some noteworthy payoff structure, and then to ask whether global convergence can be established for games in this class under certain families of evolutionary dynamics. As it turns out, general global convergence results can be proved for a number of classes of games. Among these classes are *potential games*, which include common interest games, congestion games, and games generated by externality pricing schemes; *stable games*, which include zero-sum games,

games with an interior ESS, and (perturbed) concave potential games; and *supermodular games*, which include models of Bertrand oligopoly, arms races, and macroeconomic search. A fundamental paper on global convergence of evolutionary dynamics is Hofbauer (2000); for a full treatment of these results, see Sandholm (2007).

Once we move beyond specific classes of games, global convergence to Nash equilibrium cannot be guaranteed; cycling and chaotic behaviour become possible. Indeed, Hofbauer and Swinkels (1996) and Hart and Mas-Colell (2003) construct examples of games in which all reasonable deterministic evolutionary dynamics fail to converge to Nash equilibrium from most initial conditions. These results tell us that general guarantees of convergence to Nash equilibrium are impossible to obtain.

In light of this fact, we might instead consider the extent to which solution concepts simpler than Nash equilibrium are supported by evolutionary dynamics. Cressman and Schlag (1998) and Cressman (2003) investigate whether imitative dynamics lead to subgame perfect equilibria in reduced normal forms of extensive form games – in particular, generic games of perfect information. In these games, interior solution trajectories do converge to Nash equilibrium components, and only subgame perfect components can be interior asymptotically stable. But even in very simple games interior asymptotically stable components need not exist, so the dynamic analysis may fail to select subgame perfect equilibria. For a full treatment of these issues, see Cressman (2003).

What about games with strictly dominated strategies? Early results on this question were positive: Akin (1980), Nachbar (1990), Samuelson and Zhang (1992), and Hofbauer and Weibull (1996) prove that dominated strategies are eliminated under certain classes of imitative dynamics. However, Berger and Hofbauer (2006) show that dominated strategies need not be eliminated under the BNN dynamic. Pushing this argument further, Hofbauer and Sandholm (2006) find that dominated strategies can survive under any continuous evolutionary dynamic that satisfies positive correlation and *innovation*; the latter condition requires that agents choose unused best responses with positive probability. Thus, whenever there is some probability that agents base their choices on direct evaluation of payoffs rather than imitation of successful opponents, evolutionary dynamics may violate even the mildest rationality criteria.

7 Conclusion

Because the literature on evolutionary dynamics came to prominence shortly after the literature on equilibrium refinements, it is tempting to view the former literature as a branch of the latter. But, while it is certainly true that evolutionary models have something to say about selection among multiple equilibria, viewing them simply as equilibrium selection devices can be misleading. As we have seen, evolutionary dynamics capture the behaviour of large numbers of myopic, imperfectly informed decision makers. Using evolutionary models to predict behaviour in interactions between, say, two well-informed players is daring at best.

The negative results described in Section 6 should be understood in this light. If we view evolutionary dynamics as an equilibrium selection device, the fact that they need not eliminate strictly dominated strategies might be viewed with disappointment. But, if we take the result at face value, it becomes far less surprising: if agents switch to strategies that perform reasonably well at the moment of choice, that a strategy is never optimal need not deter agents from choosing it.

A similar point can be made about failures of convergence to equilibrium. From a traditional point of view, persistence of disequilibrium behaviour might seem to undermine the very possibility of a satisfactory economic analysis. But the work described in this entry suggests that in large populations, this possibility is not only real but is also one that game theorists are well equipped to analyse.

See Also

- learning and evolution in games: adaptive heuristics
- learning and evolution in games: an overview
- learning and evolution in games: ESS
- Nash equilibrium, refinements of
- stochastic adaptive dynamics

I thank John Nachbar for helpful comments. Financial support from NSF Grant SES-0092145 is gratefully acknowledged.

Bibliography

Akin, E. 1980. Domination or equilibrium. *Mathematical Biosciences* 50, 239–50.

Beckmann, M., McGuire, C. and Winsten, C. 1956. *Studies in the Economics of Transportation*. New Haven, CT: Yale University Press.

Benaïm, M. and Weibull, J. 2003. Deterministic approximation of stochastic evolution in games. *Econometrica* 71, 873–903.

Berger, U. and Hofbauer, J. 2006. Irrational behavior in the Brown–von Neumann–Nash dynamics. *Games and Economic Behavior* 56, 1–6.

Björnerstedt, J. and Weibull, J. 1996. Nash equilibrium and evolution by imitation. In *The Rational Foundations of Economic Behavior*, ed. K. Arrow et al. New York: St Martin's Press.

Bomze, I. 1986. Non-cooperative two-person games in biology: a classification. *International Journal of Game Theory* 15, 31–57.

Brown, G. and von Neumann, J. 1950. Solutions of games by differential equations. In *Contributions to the Theory of Games I*, ed. H. Kuhn and A. Tucker. Annals of Mathematics Studies 24. Princeton: Princeton University Press.

Cressman, R. 2003. *Evolutionary Dynamics and Extensive Form Games*. Cambridge, MA: MIT Press.

Cressman, R. and Schlag, K. 1998. The dynamic (in)stability of backwards induction. *Journal of Economic Theory* 83, 260–85.

Dawkins, R. 1982. *The Extended Phenotype*. San Francisco: Freeman.

Demichelis, S. and Ritzberger, K. 2003. From evolutionary to strategic stability. *Journal of Economic Theory* 113, 51–75.

Friedman, D. 1991. Evolutionary games in economics. *Econometrica* 59, 637–66.

Fudenberg, D. and Levine, D. 1998. *Theory of Learning in Games*. Cambridge, MA: MIT Press.

Gilboa, I. and Matsui, A. 1991. Social stability and equilibrium. *Econometrica* 59, 859–67.

Hart, S. and Mas-Colell, A. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93, 1830–6.

Hofbauer, J. 2000. From Nash and Brown to Maynard Smith: equilibria, dynamics, and ESS. *Selection* 1, 81–8.

Hofbauer, J. and Sandholm, W. 2006. Survival of dominated strategies under evolutionary dynamics. Mimeo, University College London and University of Wisconsin.

Hofbauer, J., Schuster, P. and Sigmund, K. 1979. A note on evolutionarily stable strategies and game dynamics. *Journal of Theoretical Biology* 27, 537–48.

Hofbauer, J. and Sigmund, K. 1988. *Theory of Evolution and Dynamical Systems*. Cambridge: Cambridge University Press.

Hofbauer, J. and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Hofbauer, J. and Sigmund, K. 2003. Evolutionary game dynamics. *Bulletin of the American Mathematical Society N.S.* 40, 479–519.

Hofbauer, J. and Swinkels, J. 1996. A universal Shapley example. Mimeo, University of Vienna and Northwestern University.

Hofbauer, J. and Weibull, J. 1996. Evolutionary selection against dominated strategies. *Journal of Economic Theory* 71, 558–73.

Kohlberg, E. and Mertens, J.-F. 1986. On the strategic stability of equilibria. *Econometrica* 54, 1003–38.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, J. and Price, G. 1973. The logic of animal conflict. *Nature* 246, 15–18.

Nachbar, J. 1990. ‘Evolutionary’ selection dynamics in games: convergence and limit properties. *International Journal of Game Theory* 19, 59–89.

Samuelson, L. 1997. *Evolutionary Games and Equilibrium Selection*. Cambridge, MA: MIT Press.

Samuelson, L. and Zhang, J. 1992. Evolutionary stability in asymmetric games. *Journal of Economic Theory* 57, 363–91.

Sandholm, W. 2001. Potential games with continuous player sets. *Journal of Economic Theory* 97, 81–108.

Sandholm, W. 2007. *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press.

Schuster, P. and Sigmund, K. 1983. Replicator dynamics. *Journal of Theoretical Biology* 100, 533–8.

Smith, M. 1984. The stability of a dynamic model of traffic assignment – An application of a method of Lyapunov. *Transportation Science* 18, 245–52.

Swinkels, J. 1993. Adjustment dynamics and rational play in games. *Games and Economic Behavior* 5, 455–84.

Taylor, P. and Jonker, L. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–56.

Vega-Redondo, F. 1996. *Evolution, Games, and Economic Behavior*. Oxford: Oxford University Press.

Weibull, J. 1995. *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

Zeeman, E. 1980. Population dynamics from game theory. In *Global Theory of Dynamical Systems*, ed. Z. Nitecki and C. Robinson. Berlin: Springer.

How to cite this article

Sandholm, William H. "deterministic evolutionary dynamics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000249> doi:10.1057/9780230226203.0382

The New Palgrave Dictionary of Economics Online

deterrence (empirical), economic analyses of

Steven D. Levitt and Thomas J. Miles

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Empirical economic analyses of deterrence attempt to test the central prediction of Becker's (1968) rational-actor model of criminal behaviour: that less crime occurs when the expected penalties are greater. When economists have broken the simultaneity of crime rates and crime-control policies, they have generally concluded that policing levels and the scale of incarceration reduce crime rates. Economists have made less progress in determining whether these reductions in crime are due to deterrence or incapacitation, but the research suggests that both effects are likely present. Evidence on the deterrent effect of capital punishment and particular victim precautions is far less convincing.

Keywords

Becker, G.; crime and the city; crime, economic theory of; deterrence (empirical), economic analyses of; deterrence (theory), economics of; Granger–Sims causality; incapacitation vs deterrence; public enforcement of law; punishment

Article

Empirical economic analyses of deterrence seek to test the central prediction of the economic or rational-actor model of criminal behaviour that Becker (1968) pioneered. In the Beckerian model, a potential offender compares the expected costs and benefits of criminal activity, and when the expected utility of crime exceeds the expected utility loss of any punishment, the actor engages in the criminal activity. Economists have attempted to confirm or refute this model by relating geographic and temporal variation in punishment regimes, which proxy for the expected cost of offending, to aggregate crime rates, which measure the frequency of criminal activity. This approach poses two challenges. First, criminal justice policies are endogenous to crime rates, because jurisdictions often devote greater resources to crime control when the incidence of crime is higher. Second, even if the econometrician breaks the simultaneity of crime rates and crime-control policies, the estimates typically do not reveal

whether deterrence or incapacitation is the operative mechanism.

Estimates of the causal effect of policing levels on crime rates

The criminal justice policies that have most often received empirical evaluation are the scale of policing and imprisonment, which in the economic model of crime correspond to the probability of apprehension and the magnitude of the sanction, respectively. Early studies tried to infer the causal effect of police levels on crime rates by drawing cross-sectional comparisons across cities or states, but Fisher and Nagin (1978) showed that cross-sectional estimates suffer from simultaneity bias because jurisdictions with higher crime rates respond by employing more police. In the 1990s a second wave of literature emerged.

The new studies employed more sophisticated econometric strategies to break the simultaneity problem. For example, Marvell and Moody (1996) used Granger causality to identify the impact of policing levels on crime rates. A variable ‘Granger causes’ another when changes in the first variable generally precede changes in the second, and thus Granger causality refers to a temporal relationship between two variables rather than actual causation (Granger, 1969). Marvell and Moody (1996) applied this technique to more than 20 years of state and city data and found that police Granger-caused lower crime, or that increases in police were associated with future declines in crime.

Levitt (1997) employed a different econometric strategy: an instrumental variables or ‘natural experiment’ approach. He argued that mayoral and gubernatorial elections were valid instruments, because they correlate with police but do not correlate with crime, except through the other explanatory variables in the crime equation. He showed that sizable increases in the police forces in major cities were concentrated in election years, perhaps because greater police generate electoral benefits for politicians. His estimate, that a ten per cent increase in the police force produced at most a ten per cent reduction in crime rates, was comparable in magnitude to Marvell and Moody's (1996). McCrary (2002) argued that, when properly measured, electoral cycles induced insufficient variation in the size of police forces to measure the impact of crime. However, Levitt (2002) showed that an alternative instrumental variable, the number of firefighters, also produces negative and sizable estimates of the impact of police on crime. Recently, Evans and Owens (2005) demonstrated that the federal subsidies from the Clinton Crime Bill stimulated police hiring and produced similar reductions in crime rates.

Other authors used more finely disaggregated data to identify the effect of police on crime. In data with annual observations, any increase in crime and police occurring within a calendar year appears contemporaneous rather than sequential, and the short-term causal effect of police on crime is not observed. Corman and Mocan (2000) examined the short-term effect using nearly 30 years of monthly data from New York City and applying Granger causality techniques. They found that police hiring occurs approximately six months after a jump in crime and that the increase in police leads to reductions in crime as great as Levitt's (1997) largest estimate. Di Tella and Schargrodsky (2004) examined data decomposed to the level of city blocks. When the city of Buenos Aires reallocated police to temples and mosques in response to terrorist threats against them, Di Tella and Schargodsky observed that auto thefts immediately around those buildings declined abruptly but that the reduction in crime quickly decayed with distance.

Despite the use of different estimation procedures and different data-sets, the second wave of literature

on policing and crime produced quite similar estimates of the crime-reducing effect of police levels. The marginal reduction in crime associated with hiring an additional police officer in large urban environments roughly equals the marginal cost.

Estimates of the causal effect of incarceration rates on crime rates

Empirical analyses of the crime-reducing effect of prisons evolved in a similar manner to studies of policing. Early efforts failed to recognize or address the simultaneity problem and prematurely concluded that imprisonment has neither deterrent nor incapacitating effects (see Zimring and Hawkins, 1991). In the 1990s researchers again applied more sophisticated empirical strategies that attempted to break the simultaneity problem. Marvell and Moody (1994) applied Granger causality techniques to a repeated cross-section of states and found that a ten per cent increase in the prison population produced nearly a two per cent fall in crime rates.

Levitt (1996) disentangled the simultaneity of crime and incarceration by using lawsuits challenging conditions in overcrowded prisons as instrumental variables. He showed that, when the suits produced court orders to reduce overcrowding, states typically complied by releasing prisoners who otherwise would have been incarcerated. His estimates implied that the reduction in crime from incarcerating an additional prisoner was two to three times larger than that predicted by Marvell and Moody (1994).

Although these studies indicate that imprisonment reduces crime, the relevance of their estimated parameters to social policy evaluation of present incarceration levels has already diminished. The prison population has grown so rapidly since the mid-1990s that its margin lies well outside the range in which the parameter estimates were generated. For most reasonable set of assumptions, the current scale of incarceration appears at or above the socially optimal level.

Estimates distinguishing deterrence from incapacitation

Although economists' understanding of the causal relationships among policing, incarceration, and crime has improved, they have made less progress on the question of whether the declines in crime are due to deterrence or incapacitation. Determining the operative effect is crucial for evaluating the economic model of crime and for designing crime-control policy.

A few empirical economic studies attempted to assess the relative importance of deterrence and incapacitation by exploring responses to increased punishments. Kessler and Levitt (1999) studied the effect of a California referendum that provided sentence enhancements for certain serious crimes. The sentence enhancement imposed an additional incapacitating effect only upon completion of the standard prison term, and any decline in crime before that date was arguably attributable to deterrence. Kessler and Levitt found that the rate of crimes covered by the referendum fell relative to other states and that the rate of crimes not covered by the referendum did not change. After the expiration of the standard prison terms, the rate of the affected crimes continued to fall, and this further decline indicated that the full impact of the sentence enhancements included both deterrent and incapacitating effects.

Another effort to distinguish deterrence from incapacitation proceeded from the observation that criminals do not specialize in particular types of offences, but instead are generalists who participate in potentially wide range of offences. Levitt (1998a) noted that, if deterrence is the operative mechanism, a

longer sentence for a particular type of crime implies that generalist criminals should substitute to other kinds of crime. If the primary effect is instead incapacitation, then a longer sentence for a particular crime should lower the rate of alternative offences. Using arrest rate data, Levitt (1998a) found mixed evidence for deterrence.

Levitt (1998b) evaluated the responsiveness of criminal activity to the transition from the juvenile to the adult criminal justice system as another means of distinguishing deterrence from incapacitation. In states where the criminal justice system is substantially more punitive than the juvenile system, deterrence predicts that juveniles should reduce their criminal activity immediately upon reaching the age of majority (before there is time for incapacitation to become a factor). States where the adult system was especially punitive relative to the juvenile system experience sharp declines in crime at the age of majority relative to states where the transition to the adult system is most lenient, consistent with deterrence.

Other empirical analyses of deterrence

Capital punishment seemingly offers a direct test of the deterrence hypothesis, because the alternative sentence for a death-eligible offender is typically life imprisonment, and any crime-reducing effect of capital punishment is therefore arguably attributable to deterrence. Ehrlich (1975; 1977a; 1977b) produced some of the earliest and most contested claims of capital punishment's deterrent effect. Cameron (1994) reviews the large literature on the death penalty, and criticisms of Ehrlich's conclusions focus on the sensitivity of the estimates to the time period, the states, and the control variables included in the analysis. Recently, a number of studies examined the relationship between the death penalty and crime rates using repeated cross-sections of states in the period since the Supreme Court's 1976 reinstatement of capital punishment. These studies use data disaggregated at the monthly (Mocan and Gittings, 2003) or county-level (Dezhbakhsh, Rubin and Shepherd, 2003) and study the impact on different kinds of homicide (Shepherd, 2004). All claim deterrent effects at least as large as Ehrlich's original estimates, despite their continuing sensitivity to minor specification changes. In contrast, Katz, Levitt and Shustorovich (2003) used state-level panel data covering the period 1950–90 and detected no effect of the death penalty on crime rates. Unlike the literature on policing and incarceration, the use of higher frequency data and additional control variables has broadened, rather than narrowed, the range of estimated impacts of the death penalty.

Although most empirical economic analyses of deterrence evaluate the role of public law enforcement, a few studies consider the responsiveness of crime to the precautions taken by potential victims. A victim's precaution may have a general deterrent effect only if the prospective offender cannot observe it before deciding to commit the crime. Otherwise, the observation of a precaution may induce the offender to substitute to a more vulnerable victim but have no effect on the total rate of offending (see Clotfelter, 1978; Shavell, 1991). Ayres and Levitt (1998) analysed a particular kind of anti-theft device for automobiles as an unobservable precaution. The device contained a radio transmitter that allows police with special equipment to track the vehicle, but its lack of outward indications made it unobservable to potential offenders. Ayres and Levitt found that, when the device became available in a city, vehicle thefts fell sharply and that it did not induce car thieves to substitute to other types of crimes or to other geographic areas.

Another purported unobservable precaution that received extensive empirical analysis is surreptitious gun possession. Lott and Mustard (1997) and Lott (1998) claimed that laws relaxing the requirements for concealed weapons permits had a general deterrent effect on crime rates, but numerous researchers challenged the Lott findings. Ayres and Donohue (1999) found that in more recent years the law correlated either positively or not at all with crime rates, and Duggan (2000) showed that crime rates in states that adopted concealed-weapons laws began to decline before the passage of the laws. Other researchers argued that additional tests of the hypothesis failed to confirm it. Ludwig (1998) found that that passage of these laws was associated with large declines in the victimization of juveniles, a group not permitted to carry concealed weapons under these laws. Kovandzic and Marvell (2003) reported no relationship between the number of concealed weapons permits issued and violent crime rates in a single state.

See Also

- Becker, Gary S.
- causality in economics and econometrics
- crime and the city
- deterrence (theory), economics of
- difference-in-difference estimators
- Granger–Sims causality
- law, public enforcement of
- natural experiments and quasi-natural experiments

Bibliography

Ayres, I. and Donohue III, J. 1999. Nondiscretionary concealed weapons laws: a case study of statistics, standards of proof, and public policy. *American Law and Economics Review* 1, 436–70.

Ayres, I. and Levitt, S. 1998. measuring positive externalities from unobservable victim precautions: an empirical analysis of Lojack. *Quarterly Journal of Economics* 113, 43–77.

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Cameron, S. 1994. A review of econometric evidence on the effects of capital punishment. *Journal of Socio-Economics* 23, 197–214.

Clotfelter, C. 1978. Private security and the public safety. *Journal of Urban Economics* 5, 388–402.

Corman, H. and Mocan, H. 2000. A time-series analysis of crime and drug use in New York City.

American Economic Review 90, 584–604.

Dezhbakhsh, H., Rubin, P. and Shepherd, J. 2003. Does capital punishment have a deterrent effect? New evidence from postmoratorium panel data. *American Law and Economics Review* 5, 344–76.

Di Tella, R. and Schargrodsky, E. 2004. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *American Economic Review* 94, 115–33.

Duggan, M. 2000. More guns, more crime. *Journal of Political Economy* 109, 1086–114.

Ehrlich, I. 1975. The deterrent effect of capital punishment: a question of life and death. *American Economic Review* 65, 397–417.

Ehrlich, I. 1977a. The deterrent effect of capital punishment: reply. *American Economic Review* 67, 452–58.

Ehrlich, I. 1977b. Capital punishment: further thoughts and additional evidence. *Journal of Political Economy* 85, 741–88.

Evans, W. and Owens, E. 2005. Flypaper COPS. Mimeo. Department of Economics, University of Maryland. Online. Available at: http://www.bsos.umd.edu/econ/evans/wpapers/Flypaper%20COPS.pdf, accessed 20 October 2005.

Fisher, F. and Nagin, D. 1978. On the feasibility of identifying the crime function in a simultaneous equations model of crime and sanctions. In *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, ed. A. Blumstein, D. Nagin and J. Cohen. Washington, DC: National Academy of Sciences.

Granger, C. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.

Katz, L., Levitt, S. and Shustorovich, E. 2003. Prison conditions, capital punishment, and deterrence. *American Law and Economics Review* 5, 318–43.

Kessler, D. and Levitt, S. 1999. Using sentence enhancements to distinguish between deterrence and incapacitation. *Journal of Law and Economics* 17, 343–63.

Kovandzic, T. and Marvell, T. 2003. Right-to-carry concealed handguns and violent crime: crime control through gun decontrol? *Criminology and Public Policy* 2, 363–96.

Levitt, S. 1996. The effect of prison population size on crime rates: evidence from prison overcrowding

litigation. *Quarterly Journal of Economics* 111, 319–25.

Levitt, S. 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review* 87, 270–90.

Levitt, S. 1998a. Why do increased arrest rates appear to reduce crime: deterrence, incapacitation, or measurement error? *Economic Inquiry* 36, 353–72.

Levitt, S. 1998b. Juvenile crime and punishment. *Journal of Political Economy* 106, 1156–85.

Levitt, S. 2002. Using electoral cycles in police hiring to estimate the effects of police on crime: reply. *American Economic Review* 92, 1244–50.

Lott, J., Jr. 1998. *More Guns, Less Crime*. Chicago: University of Chicago Press.

Lott, J., Jr. and Mustard, D. 1997. Crime, deterrence, and right-to-carry concealed handguns. *Journal of Legal Studies* 26, 1–68.

Ludwig, J. 1998. Concealed-gun-carrying laws and violent crime: evidence from state panel data. *International Review of Law and Economics* 18, 239–54.

Marvell, T. and Moody, C. 1994. Prison population growth and crime reduction. *Journal of Quantitative Criminology* 10, 109–40.

Marvell, T. and Moody, C. 1996. Specification problems, police levels, and crime rates. *Criminology* 34, 609–46.

McCrary, J. 2002. Using electoral cycles in police hiring to estimate the effect of police on crime: comment. *American Economic Review* 92, 1236–43.

Mocan, H. and Gittings, R. 2003. Getting off death row: commuted sentences and the deterrent effect of capital punishment. *Journal of Law and Economics* 46, 453–78.

Shavell, S. 1991. Individual precautions to prevent theft: private versus socially optimal behavior. *International Review of Law and Economics* 11, 123–32.

Shepherd, J. 2004. Murders of passion, execution delays, and the deterrence of capital punishment. *Journal of Legal Studies* 33, 283–322.

Zimring, F. and Hawkins, G. 1991. *The Scale of Imprisonment*. Chicago: University of Chicago Press.

How to cite this article

Levitt, Steven D. and Thomas J. Miles. "deterrence (empirical), economic analyses of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000226> doi:10.1057/9780230226203.0383

The New Palgrave Dictionary of Economics Online

deterrence (theory), economics of

Isaac Ehrlich

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In both its classical and modern versions the economic theory of crime is predicated on ‘the deterrence hypothesis’ – the assumption that potential and actual offenders respond to both positive and negative incentives, and that the volume of offences in the population is influenced by law enforcement and other means of crime prevention. This article traces the evolution of the modern approach to crime from the traditional focus on the interaction between offenders and law enforcers to the development of a more comprehensive ‘market model’ under both partial and general equilibrium settings. Theoretical extensions also emphasize alternative criteria for optimal law enforcement.

Keywords

Beccaria, C.; Becker, G.; Bentham, J.; bribery; deterrence; deterrence hypothesis; Ehrlich, I.; expected utility hypothesis; pecuniary vs. non-pecuniary penalties; public choice; public enforcement of law; punishment; rent seeking

Article

The persistence of criminal activity throughout human history and the challenges it poses for determining optimal law-enforcement activity have already attracted the attention of utilitarian philosophers and early economists like Beccaria, Paley and Bentham. It was not until the late 1960s, however, especially following the seminal work by Becker (1968), that economists reconnected with the subject, using the modern tools of economic theory and econometrics.

In both its utilitarian and modern versions the economic approach to crime is predicated on what the new literature calls ‘the deterrence hypothesis’ – the assumption that potential and actual offenders respond to incentives, and that the volume of offences in the population is therefore influenced by law enforcement and other means of crime prevention. By its common connotation, deterrence generally refers to the threat of a criminal sanction, or any other form of punishment having some moderating

effect on the willingness to engage in criminal activity. To interpret this hypothesis so narrowly misses, however, the basic idea on which it is founded (Ehrlich, 1979). The hypothesis relates to the role of both negative incentives (such as the prospect of apprehension, conviction and punishment) and positive incentives (such as opportunities for gainful employment in legitimate relative to illegitimate occupations) as deterrents to actual or would-be offenders. It follows that not just conventional law enforcement matters in influencing the flow of offences but external market and household conditions as well, to the extent that these affect prospective gains and losses from illegitimate activity. For this approach to provide a useful approximation of the complicated reality of crime, it is not necessary that all those who commit specific crimes respond to incentives, nor is the degree of individual responsiveness prejudged; it is sufficient that a significant number of potential offenders so behave on the margin. By the same token, the theory does not preclude a priori any category of crime, or any class of incentives, as non-conforming. Indeed, economists have applied the deterrence hypothesis to a myriad of illegal activities, from tax evasion, drug abuse and fraud to skyjacking, robbery and murder.

The economic approach

In Becker's analysis the equilibrium volume of crime reflects the interaction between offenders and the law-enforcement authority, and the focus is on optimal probability, severity, and type of criminal sanction – the implicit ‘prices’ society imposes on criminal behaviour – in view of their impact on offenders and the relative social costs associated with their imposition. Subsequent theoretical work has focused on more complete formulations of specific components of the system and their micro foundations – primarily the supply of offences, the production of specific law-enforcement activities, and alternative criteria for achieving socially optimal law enforcement. A later evolution has aimed to expand the analytical setting within which crime is analysed to address the interaction between potential offenders (supply), consumers and potential victims (private actual or indirect ‘demand’), and deterrence and prevention by public authorities (government intervention). This ‘market model of crime’ (Ehrlich, 1981; 1996) has been further explored in recent years to include interactions between criminal activity and the general economy. For the specific articles on which the following discussion is based, see Ehrlich and Liu (2006, vols. 1 and 2).

Supply

The extent of participation in crime is generally modelled as an outcome of the allocation of time among competing legitimate and illegitimate activities by potential offenders. Since illegitimate activity carries the distinct risk of apprehension and punishment for illegitimate behaviour, individuals are assumed to act as expected-utility maximizers. This may generally lead many offenders to be multiple-job holders – being part-time offenders, or going in and out of criminal activity (see Ehrlich, 1973, and the empirical documentation in Reuter, MacCoun and Murphy, 1990). While the mix of pecuniary and non-pecuniary benefits varies across different crime categories, which attract persons of different earning opportunities or attitudes towards risk and moral values (‘preferences’), the basic opportunities affecting choice are identified in all cases as the perceived probabilities of apprehension, conviction and punishment, the marginal penalties imposed, and the expected net return on illegal over legal activity. Net returns from

crime rise with the level of community wealth, which enhances the potential loot from property crime, and fall with the probability of finding employment in the legitimate labour market and the prospective legitimate wages. Entry into criminal activity and the extent of involvement in crime are thus shown to be related inversely to deterrence variables and directly to the differential return it can provide over legitimate activity. Moreover, a one per cent increase in the probability of apprehension is shown to exert a larger deterrent effect than corresponding increases in the conditional probabilities of conviction and of any specific punishment if convicted (Ehrlich, 1975). Essentially due to conflicting income and substitution effects, however, sanction severity can have more ambiguous effects on active offenders: a strong preference for risk may weaken (Becker, 1968) or even reverse (Ehrlich, 1973) the deterrent effect of sanctions, and the results are even less conclusive if one assumes that the length of time spent in crime, not just the moral obstacle to entering it, generates disutility (Block and Heineke, 1975). The results become less ambiguous at the *aggregate* level, however, as one allows for heterogeneity of potential offenders due to differences in legitimate employment opportunities or preferences for risk and crime: a more severe sanction can reduce the crime rate by deterring the entry of potential offenders even if it has little effect on actual ones (Ehrlich, 1973). In addition to heterogeneity across individuals in personal opportunities and preferences, the literature has also addressed the role of heterogeneity in individuals' perceptions about probabilities of apprehension, as affected by learning from past experience (Sah, 1991). As a result, current crime rates may react, in part, to past deterrence measures. A different type of heterogeneity that can affect variability in crime rates across different crime categories and geographical units is identified by Glaeser, Sacerdote and Scheinkman (1996) and Glaeser and Sacerdote (1999) as stemming from the degree of social interaction: that is, the extent to which potential offenders are influenced by the behaviour of their neighbours.

Private ' demand '

The incentives operating on offenders often originate from, and are partially controlled by, consumers and potential victims. Transactions in illicit drugs or stolen goods, for example, are patronized by consumers who generate a direct demand for the underlying offence. But even crimes that inflict pure harm on victims are affected by an indirect (negative) demand, which is derived from a positive demand for safety (Ehrlich, 1981). By their choice of optimal self-protection (lowering the risk of becoming a victim) or self-insurance (reducing the potential loss if victimized) through use of locks, guards, safes, and alarms, or selective avoidance of crime-prone areas (Bartel, 1975; Shavell, 1991; Cullen and Levitt, 1999), potential victims influence the marginal costs to offenders, and thus the implicit return on crime. And since optimal self-protection generally increases with the perceived risk of victimization (the crime rate), private protection and public enforcement will be interdependent. The interaction between the two and its impact on possible fluctuations in the equilibrium volume of offences is explored in Clotfelter (1977), and Phillipson and Posner (1996).

Public intervention

Since crime, by definition, causes a net social loss, and crime control measures are largely a public good,

collective action is needed to augment individual self-protection. Public intervention typically aims to ‘tax’ illegal returns through the threat of punishment, or to ‘regulate’ offenders via incapacitation and rehabilitation programmes. All control measures are costly. Therefore, the ‘optimum’ volume of offences as determined by the law-enforcement authority acting as a social planner cannot be nil, but must be set at a level where the marginal cost of each measure of enforcement or prevention equals its marginal benefit.

To assess the relevant net social loss, however, one must adopt a criterion for public choice. Becker (1968) and Stigler (1970) have chosen variants of aggregate income measures as the relevant social welfare function to be maximized, requiring the minimization of the sum of social damages from offences and the social cost of law-enforcement activities. This approach can lead to powerful propositions regarding the optimal magnitudes of probability and severity of punishments for different crimes and different offenders, or, alternatively, the optimal level and mix of expenditures on police, courts and corrections. The analysis reaffirms the classical utilitarian proposition that the optimal severity of punishment should ‘fit the crime’, and thus be assessed essentially by its deterrent value, as the marginal social cost is higher for offences causing greater marginal social damage. Moreover, it makes a strong case for the desirability of monetary fines as a deterring sanction, since fines involve pure transfer payments between offenders and the rest of society. Different criteria for public choice, however, yield different implications regarding the optimal mix of probability and severity of punishment, as is the case when the social welfare function is expanded to include concerns for the ‘distributional consequences’ of law enforcement on offenders and victims in addition to their aggregate income, in which case even fines can become socially costly. These considerations can be ascribed to aversion to risk (as in Polinsky and Shavell, 1979), or to aversion towards *ex post* inequality under the law as a result of the ‘lottery’ nature of law enforcement, by which only offenders caught and convicted for crime pay for the damage caused by all offenders, including the luckier ones who escape apprehension and conviction (as in Ehrlich, 1982). A positive analysis of enforcement must also address the behaviour of the separate agencies constituting the criminal justice system: police, courts, and prison authorities. For example, Landes's analysis (1971) of the courts, which focuses on the interplay between prosecutors and defence teams, explains why settling cases out of court may be an efficient outcome of many court proceedings.

The optimal enforcement policy arising from the income-maximizing criterion can be questioned from yet another angle: a public-choice perspective. The optimization rule invoked in the aforementioned papers assumes that enforcement is carried out by a social planner. In practice, public law enforcement can facilitate the interests of rent-seeking enforcers who are amenable to malfeasance and bribes. Optimal social policy needs to control malfeasance by properly remunerating public enforcers (Becker and Stigler, 1974) or, where appropriate, setting milder penalties (Friedman, 1999).

Market equilibrium

In Ehrlich's (1981; 1996) ‘market model’, the equilibrium flow of offences results from the interaction between aggregate supply of offences, direct or derived demand for offences (through self-protection), and optimal public enforcement, which operates like a tax on criminal activity. The model derives the equilibrium volume of offences as well as the equilibrium net return, or premium, to offenders from

illegitimate over legitimate activity as a result of the interaction between the relevant aggregate supply, ‘demand’, and government net taxation of crime in a competitive setting. One important application concerns a comparison of deterrence, incapacitation and rehabilitation as instruments of crime control. This is because the efficacy of deterring sanctions cannot be assessed merely by the elasticity of the aggregate supply of offences schedule, as it depends on the elasticity of the private demand schedule as well. Likewise, the efficacy of rehabilitation and incapacitation programmes cannot be inferred solely from knowledge of their impact on individual offenders (see Cook, 1975). It depends crucially on the elasticities of the market supply and demand schedules, as these determine the extent to which successfully rehabilitated offenders will be replaced by others responding to the prospect of higher net returns. This market setting has also been applied by Viscusi (1986), who links observed net returns on specific crimes to underlying parameters of the model, and in works by Schelling (1967), Buchanan (1973), and Garoupa (2000), who analyse various aspects of organized crime by viewing it in a monopolistic rather than a competitive setting.

Crime and the economy

The ‘market model’ has been developed largely in a static, partial-equilibrium setting in which the general economy affects the illegal sector of the economy, but not vice versa. More recently, the model has been extended to deal with the interaction between the two under static and dynamic conditions as well. For example, Ehrlich (1973) argues that income inequality, serving as a proxy for relative earning opportunities in illegal versus legal activities, induces time allocation in favour of illegal activity. A number of subsequent studies interpreted this relation to imply that the volume of offences can be lowered through subsidies to legitimate employment by workers with low legitimate earning capacity. Using a general-equilibrium setting, Imrohoroglu, Merlo and Rupert (2000) show, however, that, to the extent that subsidies must be paid for by raising taxes on legitimate production, such income distribution policies have an ambiguous effect on crime. The subsidy raises the opportunity cost of crime to apprehended offenders, but it also works as a disincentive to legitimate production because of an increased tax rate, which lowers the tax revenue available for crime detection.

The choice between legitimate and illegitimate activity may have not just static effects on the economy's level of output but dynamic growth effects as well if it affects productive human capital formation, which serves as an engine of productivity growth. Bureaucratic corruption is a case in point. As Ehrlich and Lui (1999) argue, this is because, whenever government intervenes in private economic activity, bureaucrats have an opportunity to engage in rent seeking by collecting explicit or implicit bribes, which rise with their bureaucratic status. The return on corruption is thus higher the greater is one's investment in becoming a bureaucrat or attaining higher bureaucratic status, which competes with investment in productive human capital. The analysis explains why corruption is a barrier to growth especially in less developed countries, and why under benevolent autocratic regimes the rate of economic growth can be as high as under democratic regimes.

Investigating and implementing alternative versions of a comprehensive model of crime based on micro foundations remains an intriguing challenge for future research.

See Also

- Bentham, Jeremy
- deterrence (empirical), economic analyses of
- econometrics
- equality
- expected utility hypothesis
- general equilibrium
- labour economics
- law, public enforcement of
- microfoundations
- rent seeking
- risk
- uncertainty
- unemployment

Bibliography

Bartel, A. 1975. An analysis of firm demand for protection against crime. *Journal of Legal Studies* 4, 433–78.

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Becker, G. and Stigler, G. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3, 1–18.

Becker, G. and Landes, W. 1974. *The Economics of Crime and Punishment*. New York: Columbia University Press.

Block, M. and Heineke, J. 1975. A labor theoretic analysis of the criminal choice. *American Economic Review* 65, 314–25.

Buchanan, J. 1973. A defense of organized crime? In *The Economics of Crime and Punishment: A Conference Sponsored by American Enterprise Institute for Public Policy Research*. Washington, DC: American Enterprise Institute for Public Policy Research.

Clotfelter, C. 1977. Public services, private substitutes, and the demand for protection against crime. *American Economic Review* 67, 867–77.

Cook, Philip J. 1975. The correctional carrot: better jobs for parolees. *Policy Analysis* 1, 11–55.

Cullen, J. and Levitt, S. 1999. Crime, urban flight, and the consequences for cities. *Review of Economics and Statistics* 81, 159–69.

Ehrlich, I. 1973. Participation in illegitimate activities: theoretical and empirical investigation. *Journal of Political Economy* 81, 521–65. Reprinted with supplements as ‘Participation in Illegitimate Activities: An Economic Analysis’, in Becker and Landes (1974).

Ehrlich, I. 1975. The deterrent effect of capital punishment: a question of life and death. *American Economic Review* 65, 397–417.

Ehrlich, I. 1979. The economic approach to crime. In *Criminology Review Yearbook*, ed. S. Messingerand and E. Bittner. Beverly Hills: Sage Publications.

Ehrlich, I. 1981. On the usefulness of controlling individuals: an economic analysis of rehabilitation, incapacitation and deterrence. *American Economic Review* 71, 307–22.

Ehrlich, I. 1982. The optimum enforcement of laws and the concept of justice: a positive analysis. *International Review of Law and Economics* 2, 3–27.

Ehrlich, I. 1996. Crime, punishment, and the market for offenses. *Journal of Economic Perspectives* 10 (1), 43–67.

Ehrlich, I. and Brower, G. 1987. On the issue of causality in the economic model of crime and law enforcement: some theoretical considerations and experimental evidence. *American Economic Review, Papers and Proceedings* 77(2), 99–106.

Ehrlich, I. and Liu, Z. 2006. *The Economics of Crime*. International Library of Critical Writings in Economics. Cheltenham, UK and Northampton, MA: Edward Elgar.

Ehrlich, I. and Lui, F. 1999. Bureaucratic corruption and endogenous economic growth. *Journal of Political Economy* 107, S270–S293.

Friedman, D. 1999. Why not hang them all: the virtues of inefficient punishment. *Journal of Political Economy* 107 (6, Part 2), S259–S269.

Garoupa, N. 2000. The economics of organized crime and optimal law enforcement. *Economic Inquiry* 38, 278–88.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111, 507–48.

Glaeser, E. and Sacerdote, B. 1999. Why is there more crime in cities? *Journal of Political Economy* 107 (6, Part 2), S225–S258.

Imrohoroglu, A., Merlo, A. and Rupert, P. 2000. On the political economy of income redistribution and crime. *International Economic Review* 41, 1–25.

Karpoff, J. and Lott, J., Jr. 1993. The reputational penalty firms bear from committing criminal fraud. *Journal of Law and Economics* 36, 757–802.

Landes, W. 1971. An economic analysis of the courts. *Journal of Law and Economics* 14, 61–107.

Philipson, T. and Posner, R. 1996. The economic epidemiology of crime. *Journal of Law and Economics* 39, 405–33.

Polinsky, A. and Shavell, S. 1979. The optimal trade-off between the probability and magnitude of fines. *American Economic Review* 69, 880–91.

Reuter, P., MacCoun, R. and Murphy, P. 1990. *Money from Crime: A Study of the Economics of Drug Dealing in Washington, D.C*. The RAND Corporation, R-3894-RF.

Sah, R. 1991. Social osmosis and patterns of crime. *Journal of Political Economy* 99, 1272–95.

Schelling, T. 1967. Economic analysis of organized crime. Appendix D in *Task Force on Organized Crime: The President's Commission on Law Enforcement and Administration of Justice*. Washington, DC: US Government Printing Office.

Shavell, S. 1991. Individual precautions to prevent theft: private versus socially optimal behavior. *International Review of Law and Economics* 11, 123–32.

Stigler, G. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78, 526–35.

Viscusi, W. 1986. The risks and rewards of criminal activity: a comprehensive test of criminal deterrence. *Journal of Labor Economics* 4, 317–40.

How to cite this article

Ehrlich, Isaac. "deterrence (theory), economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000236> doi:10.1057/9780230226203.0384

The New Palgrave Dictionary of Economics Online

development economics

Debraj Ray
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys the current state of development economics, a subject that studies growth, inequality, poverty and institutions in the developing world. The article is organized around a view that emphasizes the role of history in creating development traps or slow progress. This ‘non-convergence’ viewpoint stands in contrast to a more traditional view, also discussed, based on the notion of economic convergence (across individuals, regions or countries). Some specific research areas in development economics receive closer scrutiny under this overall methodological umbrella, among them political economy, credit markets, legal issues, collective action and conflict.

Keywords

adverse selection; aspirations gap; collective action; complementarity; convergence; convexity; credit; development economics; expectations; human capital; inequality; insurance; land rights; limited liability; measurement error; micro-credit; moral hazard; multiple equilibria; poverty traps; property rights; selection bias; standards of living; underdevelopment

Article

What we know as the ‘developing world’ is approximately the group of countries classified by the World Bank as having ‘low’ and ‘middle’ incomes. An exact description is unnecessary and not too revealing; suffice it to observe that these countries make up over five billion of the world's population, leaving out the approximately one billion who are part of the ‘high’ income ‘developed world’. Together, the low- and middle-income countries generate approximately six trillion (2001) dollars of national income, to be contrasted with the 25 trillion generated by high-income countries. An index of income that controls for purchasing power would place these latter numbers far closer together (approximately 20 trillion and 26 trillion, according to the *World Development Report*, World Bank, 2003), but the per capita disparities are large and obvious, and to those encountering them for the first

time, still extraordinary.

Development economics, a subject that studies the economics of the developing world, has made excellent use of economic theory, econometric methods, sociology, anthropology, political science, biology and demography, and has burgeoned into one of the liveliest areas of research in all the social sciences. My limited approach in this brief article is one of deliberate selection of a few conceptual points that I consider to be central to our thinking about the subject. The reader interested in a more comprehensive overview is advised to look elsewhere (for example, at Dasgupta, 1993; Hoff, Braverman and Stiglitz, 1993; Ray 1998; Bardhan and Udry, 1999; Mookherjee and Ray, 2001; Sen, 1999).

I begin with a traditional framework of development, one defined by conventional growth theory. This approach develops the hypothesis that given certain parameters, say savings or fertility rates, economies inevitably move towards a steady state. If these parameters are the same across economies, then in the long run all economies converge to one another. If in reality we see the utter lack of such convergence – which we do (see, for example, Quah, 1996; Pritchett, 1997) – then such an absence must be traced to a presumption that the parameters in question are *not* the same. To the extent that history plays any role at all in this view, it does so by affecting these parameters – savings, demographics, government interventionism, ‘corruption’ or ‘culture’.

This view is problematic for reasons that I attempt to clarify below. Indeed, the bulk of this article is organized around the opposite presumption: that two societies with the same fundamentals can evolve along very different lines – going forward – depending on past expectations, aspirations or actual history. To some extent, the distinction between evolution and parameter is a semantic one. By throwing enough state variables (‘parameters’) into the mix, one might argue that there is no difference at all between the two approaches. Formally, that would be correct, but then ‘parameters’ would have to be interpreted so broadly as to be of little explanatory value. Ahistorical convergence and historically conditioned divergence express two fundamentally different world views, and there is little that semantic jugglery can do to bring them together.

1 Development from the viewpoint of convergence

Why are some countries poor while others are rich? What explains the success stories of economic development, and how can we learn from the failures? How do we make sense of the enormous inequalities that we see, both within and across questions? These, among others, are the ‘big questions’ of economic development.

It is fair to say that the model of economic growth pioneered by Robert Solow (1956) has had a fundamental impact on ‘big-question’ development economics. An entire literature, including theory, calibration and empirical exercises, emanates from this starting point (see, for example, Lucas, 1990; Mankiw, Romer and Weil, 1992; Barro, 1991; Parente and Prescott, 2000; Banerjee and Duflo, 2005).

Solow's path-breaking work introduced the notion of *convergence*: countries with a low endowment of capital in relation to labour will have a high rate of return to capital (by the ‘law’ of diminishing returns). Consequently, a given addition to the capital stock will have a larger impact on per capita income. It follows that, if we suitably control for parameters such as savings rates and population growth rates, poorer countries will tend to grow faster and hence will catch up or *converge* to the levels of well-

being enjoyed by their richer counterparts. According to this view, development is largely a matter of getting some economic and demographic parameters right and then settling down to wait.

It is true that savings and demography are not the only factors that qualify the argument. Anything that systematically affects the marginal addition to per capita income must be controlled for, including variables such as investment in ‘human capital’ or harder-to-quantify factors such as ‘political climate’ or ‘corruption’. A failure to observe convergence must be traced to one or another of these ‘parameters’.

Convergence relies on diminishing returns to ‘capital’. If this is our assumed starting point, the share of capital in national income gives us rough estimates of the concavity of production in capital. The problem is that the resulting concavity understates observed variation in cross-country income by orders of magnitude. For instance, Parente and Prescott (2000) calibrate a basic Cobb–Douglas production function by using reasonable estimates of the share of capital income (0.25), but then huge variations in the savings rate do not change world income by much. For instance, doubling the savings rate leads to a change in steady-state income by a factor of 1.25, which is inadequate to explain an observed range of around 20:1 (in purchasing-power-parity incomes). Indeed, as Lucas (1990) observes, the discrepancy actually appears in a more primitive way, at the level of the production function. For the same simple production function to fit the data on per capita income differences, a poor country would have to have enormously higher rates of return to capital; say, 60 times higher if it is one-fifteenth as rich. This is implausible. And so begins the hunt for other factors that might explain the difference. What did we not control for, but should have?

This describes the methodological approach. The convergence benchmark must be pitted against the empirical evidence on world income distributions, savings rates, or rates of return to capital. The two will usually fail to agree. Then we look for the parametric differences that will bridge the model to the data.

‘Human capital’ is often used as a first port of call: might differences here account for observed cross-country variation? The easiest way to slip differences in human capital into the Solow equations is to renormalize labour. Usually, this exercise does not take us very far. Depending on whether we conduct the Lucas exercise or the Prescott–Parente variant, we would still be predicting that the rate of return to capital is far higher in India than in the United States, or that per capita income differences are only around half as much (or less) as they truly are. The rest must be attributed to that familiar black box – ‘technological differences’. That slot can be filled in a variety of ways: externalities arising from human capital, incomplete diffusion of technology, excessive government intervention, within-country misallocation of resources, and so on. All these – and more – are interesting candidates, but by now we have wandered far from the original convergence model; and if that model still continues to illuminate, it is by way of occasional return to the recalibration exercise, after choosing plausible specifications for each of these potential explanations.

This model serves as a quick and ready fix on the world, and it organizes a search for possible explanations. Taken with the appropriate quantity of salt, and viewed as a first pass, such an exercise can be immensely useful. Yet playing this game too seriously reveals a particular world view. It suggests a fundamental belief that the world economy is ultimately a great leveller, and that if the levelling is not taking place we must search for that explanation in parameters that are somehow structurally rooted in a society.

While the parameters identified in these calibration exercises go hand in hand with underdevelopment, so do bad nutrition, high mortality rates, or lack of access to sanitation, safe water and housing. Yet

there is no ultimate causal chain: many of these features go hand in hand with low income in self-reinforcing interplay. By the same token, corruption, culture, procreation and politics are all up for serious cross-examination: just because ‘cultural factors’ (for instance) seems more weighty an ‘explanation’, that does not permit us to assign them the status of a truly exogenous variable. In other words, the convergence predicted by technologically diminishing returns to inputs should not blind us to the possibility of non-convergent behaviour when all variables are treated as they should be – as variables that potentially make for underdevelopment, but also as variables that are profoundly affected by the development process.

2 Development from the viewpoint of non-convergence

This leads to a different way of asking the big questions, one that is not grounded in any presumption of convergence. The starting point is that two economies with the same fundamentals can move apart along very different paths. Some of the best-known economists writing on development in the first half of the 20th century were instinctively drawn to this view: Young (1928); Nurkse (1953); Leibenstein (1957); and Myrdal (1957) among them.

Historical legacies need not be limited to a nation's inheritance of capital stock or GDP from its ancestors. Factors as diverse as the distribution of economic or political power, legal structure, traditions, group reputations, colonial heritage and specific institutional settings may serve as initial conditions – with a long reach. Even the accumulated baggage of unfulfilled aspirations or depressed expectations may echo into the future. Factors that have received special attention in the literature include historical inequalities, the nature of colonial settlement, the character of early industry and agriculture, and early political institutions.

2.1 Expectations and development

Consider the role of expectations. Rosenstein-Rodan (1943) and Hirschman (1958) (and several others following them) argued that economic development could be thought of as a massive *coordination failure*, in which several investments do not occur simply because other complementary investments are similarly depressed in the same bootstrapped way. Thus one might conceive of two (or more) equilibria *under the very same fundamental conditions*, ‘ranked’ by different levels of investment.

Such ‘ranked equilibria’ rely on the presence of a *complementarity*, a particular form of externality in which the taking of an action by an agent increases the marginal benefit to other agents from taking a similar action. In the argument above, sector-specific investments lie at the heart of the complementarity: more investment in one sector raises the return to investment in some related sector. Once complementarities – and their implications for equilibrium multiplicity – enter our way of thinking, they seem to pop up everywhere. Complementarities play a role in explaining how technological inefficiencies persist (David, 1985; Arthur, 1994), why financial depth is low (and growth volatile) in developing countries (Acemoglu and Zilibotti, 1997), how investments in physical and human capital may be depressed (Romer, 1986; Lucas, 1988), why corruption may be self-sustaining (Kingston, 2005; Emerson, 2006), the growth of cities (Henderson, 1988; Krugman, 1991), the suddenness of currency crises (Obstfeld, 1994), or the fertility transition (Munshi and Myaux, 2006); I

could easily go on. Even the traditional Rosenstein-Rodan view of demand complementarities has been formally resurrected (Murphy, Shleifer and Vishny, 1989).

An important problem with theories of multiple equilibria is that they carry an unclear burden of history. Suppose, for instance, that an economy has been in a low-level investment trap for decades. Nothing in the theory prevents the very same economy from abruptly shooting into the high-level equilibrium today. There is a literature that studies how the past might weigh on the present when a multiple equilibria model is embedded in real time (see, for example, Adserà and Ray, 1998; Frankel and Pauzner, 2000). When we have a better knowledge of such models we will be able to make more sense of some classical issues, such as the debate on balanced versus unbalanced growth. Rosenstein-Rodan argued that a ‘big push’ – a large, balanced infusion of funds – is ideal for catapulting an economy away from a low-level equilibrium trap. Hirschman argued, in contrast, that certain ‘leading sectors’ should be given all the attention, the resulting imbalance in the economy provoking salubrious cycles of private investment in the complementary sectors. To my knowledge, we still lack good theories to examine such debates in a satisfactory way.

2.2 Aspirations, mindsets and development

The aspirations of a society are conditioned by its circumstances and history, but they also determine its future. There is scope, then, for a self-sustaining failure of aspirations and economic outcomes, just as there is for ever-progressive growth in them (Appadurai, 2004; Ray, 2006).

Typically, the aspirations of an individual are generated and conditioned by the experiences of others in her ‘cognitive neighbourhood’. There may be several reasons for this: the use of role models, the importance of relative income, the transmission of information, or peer-determined setting of internal standards and goals. Such conditioning will affect numerous important socio-economic outcomes: the rate of savings, the decision to migrate, fertility choices, technology adoption, adherence to norms, the choice of ethnic or religious identity, the work ethic, or the strength of mutual insurance motives.

As an illustration, consider the notion of an *aspirations gap*. In a relatively narrow economic context (though there is no need to restrict oneself to this) such a gap is simply the difference between the standard of living that is aspired to and the standard of living that one already has. The former is not exogenous; it will depend on the ambient standards of living among peers or near-peers, or perhaps other communities.

The aspirations gap may be filled, or neglected, by deliberate action. Investments in education, health, or income-generating activities are obvious examples. Does history, via the creation of aspirations gaps, harden existing inequalities and generate poverty traps? Or does the existence of a gap spur individuals on ever harder to narrow the distance? As I have argued in Ray (1998, Sections 3.3.2 and 7.2.4) and Ray (2006), the effect could go either way. A small gap may encourage investments, a large gap stifle it. This leads not only to history-dependence, but also a potential theory of the connections between income inequality and the rate of growth.

These remarks are related to Duflo's (2006) more general (but less structured) hypothesis that ‘being poor almost certainly affects the way people think and decide’. This ‘mindset effect’ can manifest itself in many ways (an aspirations gap being just one of them), and can lead to poverty traps. For instance, Duflo and Udry (2004) find that certain within-family insurance opportunities seem to be inexplicably

forgone. In broadly similar vein, Udry (1996) finds that men and women in the same household farm land in a way that is not Pareto-efficient (gains in efficiency are to be had by simply reallocating inputs to the women's plots). These observations suggest a theory of the poor household in which different sources of income are treated differently by members of the household, perhaps in the fear that this will affect threat points in some intra-household bargaining game. This in itself is perhaps not unusual, but the evidence suggests that poverty itself heightens the salience of such a framework.

2.3 Markets and history dependence

I now move on to other pathways for history dependence, beginning with the central role of inequality. According to this view, historic inequalities persist (or widen) because each individual entity – dynasty, region, country – is swept along in a self-perpetuating path of occupational choice, income, consumption and accumulation. The relatively poor may be limited in their ability to invest productively, both in themselves and in their children. Such investments might include both physical projects, such as starting a business, and ‘human projects’, such as nutrition, health and education. Or the poor may have ideas that they cannot profitably implement, because implementation requires start-up funds that they do not have. Yet, faced with a different level of initial inequality, or jolted by a one-time redistribution, the very same economy may perform very differently. The ability to make productive investments is now distributed more widely throughout the population, and a new outcome emerges with not just lower inequality, but higher aggregate income. These are different steady states, and they could well be driven by distant histories (see, for example, Dasgupta and Ray, 1986; Banerjee and Newman, 1993; Galor and Zeira, 1993; Ljungqvist 1993; Ray and Streufert, 1993; Piketty 1997; Matsuyama, 2000).

The intelligent layperson would be unimpressed by the originality of this argument. That the past systematically preys on the present is hardly rocket science. Yet theories based on convergence *would rule out such obvious arguments*. Under convergence, the very fact that the poor have limited capital in relation to labour allows them to grow faster and (ultimately) to catch up. Economists are so used to the convergence mechanism that they sometimes do not appreciate just how unintuitive it is.

That said, it is time now to cross-examine our intelligent layperson. For instance, if all individuals have access to a well-functioning capital market, they should be able to make an efficient economic choice with no heed to their starting position, and the shadows cast by past inequalities must disappear (or at least dramatically shrink). For past wealth to alter current investments, imperfections in capital or insurance markets must play a central role.

At the same time, such imperfections are not sufficient: the concavity of investment returns would still guarantee convergence. A first response is that ‘production functions’ are simply not concave. A variety of investment activities have substantial fixed costs: business start-ups, nutritional or health investments, educational choices, migration decisions, crop adoptions. Indeed, it is hard to see how the presence of such non-convexities could *not* be salient for the ultra poor. Coupled with missing capital markets, it is easy to see that steady state traps, in which poverty breeds poverty, are a natural outcome (see, for example, Majumdar and Mitra, 1982; Galor and Zeira, 1993). Surveys of the economic conditions of the poor (Fields, 1980; Banerjee and Duflo, 2007) are eminently consistent with this point of view.

A related source of non-convexity arises from limited liability. A highly indebted economic agent may have little incentive to invest. Similarly, poor agents may enter into contracts with explicit or implicit

lower bounds on liability. These bounds can create poverty traps (Mookherjee and Ray, 2002a). Investment activities that go past these minimal thresholds are potentially open to ‘convexification’. There are various stopping points for human capital acquisition, and a household can hold financial assets which are, in the end, scaled-down claims on other businesses. According to this point of view, dynasties that make it past the ultra-poor thresholds will exhibit ergodic behaviour (as in Loury, 1981; Becker and Tomes, 1986) and so the prediction is roughly that of a two-class society: the ultra poor are caught in a poverty trap and the remainder enjoy the benefits of convergence. History would matter in determining the steady-state proportions of the ultra-poor.

But this sort of analysis ignores the endogenous non-convexities brought about by the price system. For instance, even if there are many different education levels, the wage payoff to each such level will generally be determined by the market. There is good reason to argue (see, for example, Ljungqvist, 1993; Freeman, 1996; Mookherjee and Ray, 2002b; 2003) that the price system will sort individuals into different occupational choices, and that there will be persistent inequality across dynasties located at each of these occupational slots. Thus an augmented theory of history dependence might predict a particular proportion of the ultra-poor trapped by physical non-convexities (low nutrition, ill-health, debt, lack of access to primary education), as well as a persistently unequal dispersion of dynasties across different occupational choices, induced by the pecuniary externalities of relative prices.

Note that it is precisely the high-inequality, high-poverty steady states that are correlated with low average incomes for society as a whole, and it is certainly possible to build a view of underdevelopment from this basic premise. The argument can be bolstered by consideration of economy-wide externalities; for instance, in physical and human capital (Romer, 1986; Lucas, 1988; Azariadis and Drazen, 1990).

2.4 History, aggregates and the interactive world

Theories such as these might yield a useful model for the interactive world economy. Take, for instance, the notion of aspirations. Just as domestic aspirations drive the dynamics of accumulation *within* countries, there is a role, too, for national aspirations that are driven by inter-country disparities in consumption and wealth, with implications for the international distribution of income. Even the simplest growth framework that exhibits the usual features of convexity in its technology and budget constraints could give rise in the end to a bipolar world distribution. Countries in the middle of that distribution would tend to accumulate faster, be more dynamic and take more risks as they see the possibility of full catch-up within a generation or less. One might expect the greatest degree of ‘country mobility’ in this range. In contrast, societies that are far away from the economic frontier may see economic growth as too limited and too long-term an instrument, leading to a failure, as it were, of ‘international aspirations’. Groups within these societies may well resort to other methods of potential economic gain, such as rent-seeking or conflict. (The aggregate impact of such activities would reinforce the slide.)

Of course, an entirely mechanical transplantation of the aspirations model to an international context is not a good idea. Countries are not individual units: a more complete theory must take into account the aspirations of various groups in the different countries, and the domestic and international components that drive such aspirations.

Next, consider the role of markets. Once again, tentatively view each country as a single economic agent

in the framework of Section 2.3.The non-convexities to be considered are at the level of the country as a whole – Young's increasing returns on a grand scale, or economy-wide externalities as in Lucas–Azariadis–Drazen. This reinterpretation is fairly standard, but, less obviously, the occupational choice story stands up to reinterpretation as well. To see this, note that the pattern of production and trade in the world economy will be driven by patterns of comparative advantage across countries. But in a dynamic framework, barring non-reproducible resources such as land or mineral endowments, *every* endowment is potentially accumulable, so that comparative advantage becomes endogenous. Thus we may view countries as settling into subsets of occupational slots (broadly conceived), producing an incomplete range of goods and services in relation to the world list, and engaging in trade.

For instance, suppose that country-level infrastructure can be tailored to either high-tech or low-tech production, but not both. If both high-tech and low-tech are important in world production and consumption, then *one* country has to focus on low-tech and *another* on high-tech. Initial history will constrain such choices, if for no other reason than the fact that existing infrastructure (and national wealth) determines the selection of future infrastructure. This is not to say that no country can break free of those shackles. For instance, as the whole world climbs up the income scale, natural non-homotheticities in demand will push commodity compositions increasingly in favour of high-quality goods. As this happens, more countries will be able to make the transition. But on the whole, if national infrastructure is more or less conducive to some (but not the full) range of goods, the non-convergence model that we discussed for the domestic economy must apply to the world economy as well.

This raises an obvious question: what is so specific about ‘national infrastructure’? Why is it not possible for the world to ultimately rearrange itself so that every country produces the same or similar mix of goods, thus guaranteeing convergence? Do current national advantages somehow manifest themselves in future advantages as well, thus ensuring that the world economy settles into a permanent state of global inequality? Might economic underdevelopment across countries, at least in this relative sense, always stay with us?

To properly address such questions we have to drop the tentative assumption that each country can be viewed as an individual unit. In a more general setting, there are individuals within countries, and then there is cross-country interaction. The former are subject to the forces of occupational structure (and possible fixed costs), as discussed in Section 2.3. The latter are subject to the specificities, if any, of ‘national infrastructure’, determining whether countries as a whole have to specialize (at least to some degree). The relative importance of within-country versus cross-country inequalities will rest, in large part, on considerations such as these.

I have not brought in international political economy so far (though see below); yet, as frameworks go, this is not a bad one to start thinking about the effects of globalization. It is certainly preferable to a view of the world as a set of disconnected, autarkic growth models.

2.5 Institutions and history

In many developing countries, the early institutions of colonial rule were directly set up for the purposes of surplus extraction. There would be variation, of course, depending on whether the areas were sparsely or densely populated to begin with, or whether there was widespread availability of mineral deposits. Resource deposits certainly favoured large-scale extractive industry (as in parts of South America), while soil and weather conditions might encourage plantation agriculture, often with the use of slave

labour (as in the Caribbean). On the other hand, a high pre-existing population density would favour extraction of a different hue: the setting up of institutional systems to acquire rents (the British colonial approach in large parts of India).

It has been argued, perhaps most eloquently by Sokoloff and Engerman (2000), that initial institutional modes of production and extraction in distant history had far-reaching effects on subsequent development. In their words, scholars ‘have begun to explore the possibility that initial conditions, or factor endowments broadly conceived, could have had profound and enduring impacts on long-run paths of institutional and economic development’ (2000, p. 220). The inequalities generated by such initial conditions may subsequently be inimical to development in a variety of ways (via the market-based pathways discussed earlier, for instance). In contrast, where initial settlements did not go hand in hand with systems of tribute, land grants, or large-scale extractive industries (as in several regions of North America), one might expect broad-based development to occur.

This is consistent with the market-based processes considered earlier. But a principal strand of the Sokoloff–Engerman argument, as also the lines of reasoning pursued in Robinson (1998), Acemoglu, Johnson and Robinson (2001; 2002) and Acemoglu (2006), emphasizes political economy. In the words of Sokoloff and Engerman,

> [I]nitial conditions had lingering effects … because government policies and other institutions tended to reproduce them. Specifically, in those societies that began with extreme inequality, elites were better able to establish a legal framework that insured them disproportionate shares of political power, and to use that greater influence to establish rules, laws, and other government policies that advantaged members of the elite relative to nonmembers contributing to persistence over time of the high degree of inequality … In societies that began with greater equality or homogeneity among the population, however, efforts by elites to institutionalize an unequal distribution of political power were relatively unsuccessful … (Sokoloff and Engerman, 2000, p. 223–4)

The elite – erstwhile collectors of tribute, land-grant recipients, plantation owners and the like – may survive long after the initial institutions that spawned them are gone. Such survival may nevertheless be compatible with the maximization of aggregate surplus, provided that the elite are the most efficient of the economic citizenry in the generations to come. But there is absolutely no reason why this should be the case. A new generation of entrepreneurs, economic and political, may be waiting to take over in the wings. It is an open question as to what will happen next, but the elite may well engage in policy that has as its goal not economic efficiency but the crippling of political opposition. Some evidence of this reluctance to let go may be seen in literature that argues that more unequal societies redistribute less (see Perotti, 1994; 1996; the survey by Bénabou, 1996).

There are other routes. The elite may be unable to avoid an oppositional showdown. A theory of bad policy may then have to be replaced by a model of social unrest and conflict generated by initial inequality. While this mechanism is clearly different, the end result is the same. The channelling of resources to ongoing conflict will surely inhibit the accumulation of productive resources (Benhabib and Rustichini, 1996; González, 2007). There may also be effects running through legal systems (see, for example, La Porta et al., 1997; 1998) or the varying nature of different colonial systems (see, for

example, Bertocchi and Canova, 2002). There may be effects running through the insecurity of property rights or fear of elite expropriation (see, for example, Binswanger, Deininger and Feder, 1995).

We do not yet have a systematic exploration of these mechanisms, nor an accounting of their relative importance. But there is some reduced-form evidence that historical institutions affect growth in the manner described by Sokoloff and Engerman. The problem in establishing an empirical assertion of this sort is fairly obvious: good institutions and good economic outcomes may simply be correlated via variables we fail to observe or measure, or any observed causality may simply run from outcomes to institutions. Acemoglu, Robinson and Johnson (2001) propose a novel instrument for (bad) institutions: the mortality rate among European settlers (bishops, sailors and soldiers to be exact). This is a clever idea that exploits the following theory: only areas that could be settled by the Europeans developed egalitarian, broad-based institutions. In the other areas, the same Europeans settled for slavery, dictatorship, highly unequal land grants and unbridled extraction instead. (The implied instrument is more convincing when the analysis is combined with controls for the general disease environment, which could have a direct effect on performance.)

The Acemoglu–Johnson–Robinson results, which show that early institutions have an effect on current performance, are provocative and interesting. It bears reiteration, though, that IV estimates are suggestive of an institutional impact on development, but one just cannot be sure of what the mechanism is. By relinquishing more immediate institutional effects on the grounds of, say, endogeneity, it becomes much harder to identify the structural pathways of influence. This appears to be an endemic problem with large, sweeping cross-country studies that attempt to detect an institutional effect. Good instruments are hard to find, and when they exist, their effect could be the echo of one or more of a diversity of underlying mechanisms.

Iyer (2005) and Banerjee and Iyer (2005) consider a somewhat different channel of influence. Both these papers study the differential impact of colonial rule within a single country, India. Iyer studies British annexations of parts of India, and the effect today on public goods provision across annexed and non-annexed parts. There is obvious endogeneity in the areas chosen for annexation (a similar observation applies, in passing, to countries ‘selected’ for colonization). Iyer instruments annexation by exploiting the so-called Doctrine of Lapse, under which the British annexed states in which a native ruler died without a biological heir. Banerjee and Iyer study the effect of variations in the land revenue systems set up by the British, starting from the latter half of the 18th century. In particular, they distinguish between landlord-based institutions, in which large landlords were used to syphon surplus to the British, and other areas based on rent payments, either directly from the cultivator or via village bodies. While these institutions of extraction no longer exist (India has no agricultural income tax), the authors argue that divided, unequal areas in the past cannot come together for collective action. Dispossessed groups are more worried about insecurity of tenure and fear of expropriation than about the absence of public goods, investment (public or private) or development expenditure.

2.6 Institutions and the interactive world

In Section 2.4, we applied market-based theories of occupational choice and persistent inequality to the interactive world economy, (tentatively) treating each country as an economic agent. Recall the main assumption for such an interpretation to be sensible: that countries must face infrastructural constraints

that limit full diversification. With these constraints in place, there will be persistent inequality in the world income distribution, with countries in ‘occupational niches’ that correspond to their infrastructural choices.

Bring to this story the role of institutional origins. Then a particular institutional history may be more suited to particular subsets of occupations, driving the country in question into a determinate slot in the world economy. From that point on, the persistent cross-country inequalities generated by the market-based theory will continue to link past institutions to subsequent growth. In short, initial institutional differences may be correlated with subsequent performance, but the *magnitude* of that under- or over-performance is not to be entirely traced to initial history. Distant history could simply have served as a marker for some countries to supply a particular range of occupations, goods and services. Today's inequality may well be driven, not by that far-away history but simply by the world equilibrium path that follows on those initial conditions. If all goods are needed, there must be banana producers, sugar manufacturers, coffee growers, and high-tech enclaves, but there cannot be too little or too many of any of them.

The ‘inefficient political power’ argument used in Section 2.5 can also be transplanted to international interactions. It may well be that a large part of such interactions – protection of international property rights, restrictions on technology transfer, or barriers to trade – is used to deter the entry of developing countries onto a level playing field in which they can successfully compete with their compatriots in developed countries. It would certainly be naive to disregard this point of view altogether.

Looked at this way, our view of history fits in well with the entire debate on globalization. One might view one side of this debate as emphasizing the convergence attributes of globalization: outsourcing, the establishment of international production standards, technology transfer, political accountability and responsible macroeconomic policies may all be invoked as foot soldiers in the service of convergence. On the other side of the battle lines are equally formidable opponents. A skewed playing field can only keep tipping, so goes the argument. The protection of intellectual property is just a way of maintaining or widening existing gaps in knowledge. Technology transfers are inappropriate because the input mix is not right. Non-convexities and increasing returns are endemic.

My goal here is not to take sides on this debate (though like everyone, I do have an opinion) but to clarify it from a ‘non-convergence perspective’ that has so far received more attention within the closed economy. There is a strong parallel between globalization (and those contented or discontented with it, to borrow a phrase from Joseph Stiglitz, 2002) and the questions of convergence and divergence in closed economies.

3 Digging deeper: the microeconomics of development

There is no getting away from the big questions, even if they cannot be fully answered with the knowledge and tools we have to hand. The issues we have discussed (and our intuitive first-takes on them) determine our world view, the cognitive canvas on which we arrange our overall thoughts. But only the most hard-bitten macroeconomist would feel no trepidation about taking these models literally, and applying them without hesitation across countries, regions and cultures.

The microeconomics of development enables us to dig below the macro questions, unearthing insight and structure with far more confidence than we can hope to have at the world or cross-country level. From the viewpoint of economic theory, the assumptions made can be more carefully motivated and are

open to careful testing. From the viewpoint of empirical analysis, it is far easier to find instruments or natural experiments, or, for that matter, to conduct one's own experiments. There is the philosophical problem of scaling up the results, of using a well-controlled finding to predict outcomes elsewhere. In the end, the choice between the fuzzy, imprecise big picture and the small yet carefully delineated canvas is perhaps a matter of taste.

I need hardly add that my selectivity continues unabated: there is a whole host of issues, and I can but touch on a fraction of them. I focus deliberately on four important topics that are relevant to my overall theme of history dependence, and that have been the subject of much recent attention.

3.1 The credit market

As we have seen, a failure of the credit market to function is at the heart of market-based arguments for divergence.

The fundamental reason for imperfect or missing credit markets is that individuals cannot be counted upon (for reasons of strategy or luck) to fully repay their loans. If borrowers do not have deep pockets, or if a well-defined system for enforcing repayment is missing, then it stands to reason that lenders would be reluctant to advance those loans in the first place. There is little point in asserting that a carefully chosen risk premium will deal with these risks: the premium itself affects the default probability. Therefore some borrowers will be shut out of the market, *no matter what rate of interest they are willing to pay*. Such a market will typically clear by rationing access to credit, and not by an adjustment of the rate of interest.

Three fundamental features characterize different theories of imperfect credit markets. There is classical adverse selection, in which borrower (or project) characteristics may systematically adjust with the terms of the loan contract on offer. Stiglitz and Weiss (1981) initiate this literature for credit markets, arguing that the higher the interest rate, the more likely it is that the borrower pool will be contaminated by riskier types. Then there is the moral hazard problem (see, for example, Aghion and Bolton, 1997), in which the borrower must expend effort *ex post* to increase the chances of project success. Moral hazard also ties into ‘debt overhang’, in which existing indebtedness makes it less credible that a borrower will put in sustained effort in the project. Finally, there is the enforcement problem (see, for example, Eaton and Gersovitz, 1981), in which a borrower may be tempted to engage in strategic default. Ghosh, Mookherjee and Ray (2001) survey some of the literature.

The poor are particularly affected, not because they are intrinsically less trustworthy, but because in the event of a project failure they will not have the deep pockets to pay up. The poor may well possess collateral – a small plot of land or their labour – but such collateral may be hard to adequately monetize: a formal sector bank may be unwilling to accept a small rural plot as collateral, much less bonded labour; but other lenders (a rural landlord, for instance) might. It is therefore not surprising to see interlinkages in credit transactions for the poor: a small farmer is likely to borrow from a trader who trades his crop, while a rural tenant is likely to borrow from his landlord. Even when the entire market looks competitive, these niches may create pockets of exploitative local monopoly (Ray and Sengupta, 1989; Floro and Yotopoulos, 1991; Floro and Ray, 1997; Mansuri, 1997; Genicot, 2002).

In short, the very fact of their limited wealth puts the relatively poor under additional constraints in the credit market. This is why imperfect capital markets serve as a starting point for many of the models that

study market-based history dependence.

The direct empirical evidence on the existence of credit constraints is surprisingly sparse, which is obviously not to say that they do not exist, but only to point out that this is an area for future research. Existing literature in a development context largely uses the existence of (presumably undesirable) consumption fluctuations in households to infer the lack of perfect financial markets (see Morduch, 1994; Townsend, 1995; Deaton, 1997). A direct test for credit constraints yields positive results for Indian firms (Banerjee and Duflo, 2004), though it is unclear how general this finding is (see, for example, Hurst and Lusardi, 2004). There is a sizeable literature dealing with the impact of credit constraints on outcomes such as health (Foster, 1995), education (Jacoby and Skoufias, 1997) or the acquisition of production inputs such as bullocks (Rosenzweig and Wolpin, 1993).

Chiappori and Salanie (2000) and Karlan and Zinman (2006) are two examples of specific tests for different frictions, such as adverse selection and enforcement. Udry's seminal (1994) paper on credit and insurance markets in northern Nigeria may be viewed as singling out enforcement as perhaps the most important binding constraint. The importance of enforcement constraints is, of course, not peculiar to credit or insurance; Fafchamps (2004) develops the point for a variety of markets in sub-Saharan Africa. For more on insurance, see Townsend (1993; 1995); Ligon (1998); Fafchamps (2003); and Fafchamps and Lund (2003). Coate and Ravallion (1993), Ligon, Thomas and Worrall (2002), Kocherlakota (1996) and Genicot and Ray (2003) develop some of the associated theory with limited enforcement.

Finally, there is a literature on *micro-credit*, the lending of relatively small amounts to the very poor; Armendáriz and Morduch (2005) is a good starting point.

3.2 Collective action for public goods

There is a growing literature on the political economy of development. Unlike some mainstream approaches in political science and political economy, this literature appears to largely eschew voting models. In my view this is not a bad thing. Perhaps the most important criticism of voting models is that even in vigorous democracies, most policies are not subject to referenda among the citizenry at large. Certainly, there are periodic elections, and the sum total of enacted policies – and the package of future promises – are then up for voter scrutiny, but, nevertheless, there is a large and significant gap between voting and the enactment of a *particular* policy. Between that policy and the voter falls the shadow of collective action, lobbies, capture and influence, cynical trade-offs across special interests, and covert or open conflict. For countries with a non-democratic history, these considerations are expanded by orders of magnitude.

An important literature concerns the determinants of collective action for the provision of public goods, and how poverty or inequality affects the ability to engage in such action. The relationship here is complex. There are two potential reasons why inequality in a community may enhance collective action. First, the elite in a high-inequality community might largely internalize all the benefits from the resulting public good, and therefore pay for it (Olson, 1965). Good examples involve military alliances (Sandler and Forbes, 1980), technology adoption (Foster and Rosenzweig, 1995) or even ‘top-down interventions’ by local rulers or elites (Banerjee, Iyer and Somanathan, 2007). Second, the elite has a low opportunity cost of money, while the poor have a low opportunity cost of labour; in some situations,

the two resources can be usefully combined for collective action (an alliance for violent conflict, as in Esteban and Ray, 2007a, is a good example). But there are many situations in which inequality can dampen effective collective action: when all agents supply similar inputs – say effort – but their impact or cost of provision is nonlinear (Khwaja, 2004; Ray, Baland and Dagnielie, 2007), when there are unequally distributed private endowments (Baland and Platteau, 1998; Bardhan, Ghatak and Karaivanov, 2006), when different individuals in the same community want different things by virtue of their social differences or inequality (Alesina, Baqir and Easterly, 1999; Banerjee et al., 2001, Miguel and Gugerty, 2005; Alesina and La Ferrara, 2005), or when inequalities in wealth erode the informational basis of collective action (Esteban and Ray, 2006).

The importance of this area of research cannot be overemphasized. Several of the fundamental accompaniments of development require state intervention at a basic level: health, education, social safety nets and infrastructure. This is especially so in poor countries, where privatized health and education are often ruled out by the sheer force of economic necessity. Yet states often are set upon by numerous claims that compete for their attention. How are these claims resolved? The theory and practice of collective action demands more research.

Moreover, while it can be argued (as above) that inequality within a community might go either way in affecting that community's ability to obtain public goods, there is no escaping the fact that at the level of the entire society, high inequality serves to fracture and divide. Simply put, the very rich want state policy that is different from what the very poor desire, and rare is the society that has them in the same camp, and demanding the same things of their government. In the world of the median voter, one might simply resolve these issues by looking at the median voter's ideal policy, but even in this rarefied scenario there are complex issues that deserve our consideration. Political alliances can often redefine the median voter (Levy, 2004) and even without alliances it is unclear just who the median voter is (Bénabou, 2000). When we return to the ‘real world’ of collective action, these issues are magnified considerably. In that world, each citizen does not have an endowment of one vote. The real endowments are labour and money. How these commodities combine (or compete) is fundamental to our understanding of political economy and – via this channel – our views on persistent history-dependence.

3.3 Conflict

A more sinister expression of collective action is conflict. In the second half of the 20th century and well into the first decade of the 21st the loss of human life from conflicts in developing countries was immense; the costs are beyond measurement. Even the narrow economic costs of conflict can be extremely large (Hess, 2003).

That conflict contributes to economic regress is not surprising. But given our focus on history dependence, it is of equal interest to consider the causal chain running from underdevelopment to conflict. That chain has a natural and simple foundation: poverty reduces the opportunity cost of engaging in conflict. The grabbing of resources, often in an organized way, is often a far more lucrative alternative to the steady process of wealth accumulation. It is certainly a quicker alternative. (One might argue that there is less to gain as well, but this effect is attenuated in unequal societies.)

This unfortunate observation has substantial empirical support. For instance, Miguel, Satyanath and Sergenti (2004) use rainfall as an instrument for economic growth in 41 African countries and derive a

striking negative effect of growth on civil conflict: a negative growth shock of five percentage points raises the likelihood of civil conflict by 50 per cent; see also Dube and Vargas (2006) and Hidalgo et al. (2007), both of which also instrument for economic shocks to find significant effects on conflictual outcomes. Collier and Hoeffler (1998), Sambanis (2001), Fearon and Laitin (2003), and Do and Iyer (2006) all establish strong correlations between economic adversity and conflict, the last of these countries establishing this over regions in a single country (Nepal).

Yet conflict is demonstrably wasteful, and if warring parties could sit down at the negotiating table, why would societies engage in it? This is a classical question to which there are a number of possible answers. First, there may be a Prisoner's Dilemma-like quality to conflictual incidents, in the sense that one party can precipitate attacks while the other remains passive (Leventoglu and Slantchev, 2005). Second, while conflict generates waste, there is no reason to believe that *every* group is thereby made worse off by it. It is entirely possible that a group prefers conflict to a peaceful outcome: the former involves a smaller pie, but the group may obtain a larger share of it (Esteban and Ray, 2001). Third, while one should be able to find a system of taxes and transfers that Pareto-dominate the conflict outcome, for various reasons – lack of commitment, a sparse informational base for the levying of taxes, dynamics with rapid power shifts – it may not be possible to implement that system (Fearon, 1995; Powell, 2004; 2006). Fourth, it is certainly possible that conflict is over indivisible resources such as political power or religious hegemony. It may then be absurd to imagine that side *A* compensates side *B* with suitable transfers in exchange for political power: the lack of credibility involved is only too apparent. Finally, conflict may be endemic because both parties to it have incomplete information regarding chances of success, though this view has come under increasing criticism from political scientists (see, for example, Fearon, 1995).

The next question of relevance concerns ethnic and social divisions. Might the presence of potentially divisive markers (caste, religion, geography, ethnicity in general) exacerbate conflictual situations? For instance, Esteban and Ray (2007a) argue that non-economic (‘ethnic’) markers may play a salient role in the outbreak of conflict even when society exhibits high economic inequality and may look prima facie more ripe for a class war.

A standard tool for measuring ethnic and social divisions is that of *fractionalization*, roughly defined as the probability that two individuals drawn at random will come from two distinct groups. While fractionalization seems to have a negative effect on economic outcomes such as per capita GDP (Alesina et al., 2003), growth (Easterly and Levine, 1997), or governance (Mauro, 1995), its effect on civil conflict appears to be insignificant (Collier and Hoeffler, 2004; Fearon and Laitin, 2003). Of course, as Horowitz (2000) and others have observed, it is the presence of large cleavages that is potentially conflictual, whereas fractionalization continues to increase with diversity. The solution is to drop fractionalization altogether. Montalvo and Reynal-Querol (2005) adapt Esteban and Ray's (1994) measure of polarization to show that measures of ethnic and religious polarization *do* indeed have a significant impact on conflict (see also Do and Iyer, 2006). Obviously, more research is called for on questions such as these. For instance, it is unclear how polarization should enter an empirical specification: Esteban and Ray (2007b) argue that highly polarized societies may actually avoid a showdown through deterrence, though *conditional* on the outbreak of conflict, polarization must vary positively with the intensity of conflict.

The continuing study of conflict in development demands our highest priority. Certainly, the social waste of conflict dominates the inefficiency of misallocated resources that so many mainstream economists prefer to emphasize. Indeed, it is entirely possible that the much-maligned (and much-studied) inefficiencies of incomplete information are also of a lower order of magnitude. But, most of all, it is the chain of cumulative causation that must ultimately drive our interest, from underdevelopment to conflict and back again to continuing underdevelopment. Conflict is one channel through which history matters.

3.4 Legal matters

Contract enforcement, property rights, and expropriation risks: these are a few instances of legal matters that are central to development. They bear closely on that much-used catchall phrase, ‘institutional effects on development’. For instance, Acemoglu, Robinson and Johnson (2001) as well as the recent survey by Pande and Udry (2007) clearly have the security of property rights high on the list when discussing ‘institutions’. La Porta et al. (1997; 1998; 2002) and Djankov et al. (2003) begin with the premise that common (English commercial) law and civil (French commercial) law afford different degrees of protection and support to investors, creditors and litigants, and argue that it has had dramatic effects on a variety of indicators across countries: corruption, stock-market participation, corporate valuation, government interventionism, judicial efficiency – and presumably, via these, to economic indicators.

It is little surprise that the security of property rights is generally conducive to investment, and that long-term investment is especially encouraged by such security (see, for example, Demsetz, 1967). Short-term efforts, in contrast, may well be enhanced by insecurity of tenure. Depending on the exact form that property rights assume, there may be further positive effects – for example, via access to credit – that arise from the ability to mortgage or sell property (Feder et al., 1988).

Empirical research into these matters is invariably assailed by questions of endogeneity and omitted variables. For instance, long-gestation investments may provoke – and permit – the establishment of property rights, and high-ability agents might use their ability to both invest and secure their rights. Nevertheless, the evidence on property rights is that by and large they are good for investment and production (Besley, 1995; Banerjee, Gertler and Ghatak, 2002; Do and Iyer, 2003; Goldstein and Udry, 2005), and even more obviously, property values where these are reasonably well-defined (Alston, Libecap and Schneider, 1996; Lanjouw and Levy, 2002). Instances in which property titling creates better access to credit are, intriguingly enough, somewhat harder to come by (Field and Torero, 2006, and Dower and Potamites, 2006, are two of the rarer examples that do document better access, but with some qualifications).

Indeed, economists have little trouble in finding numerous instances of changed (or changing) property rights regimes. This is because there is a plethora of situations in which the absence of well-defined rights is the rule rather than the exception. In rural societies the world over, land rights can be highly ambiguous, and land titles can be missing even when an unambiguous definition of property exists. If one adds to this the sizable proportion of land under tenancy, the effective security for cultivators becomes more tenuous still (and indeed this complicates matters, because their rights may be inversely related to those of the owner!). In non-rural settings, there are substantial uncertainties for those who

operate in the informal sector (such as the periodic ‘cleansing’ of informal retailers from city pavements). If the above studies are to be taken seriously, there are substantial production losses from such states of insecurity.

If imperfections of the law are so inimical to the fortunes of cultivators and producers (and especially for the small and the poor among them), why do we see such institutional ‘failures’ in equilibrium? The Coase–Posner view would presumably have none of this: in their view, legal systems would invariably develop to maximize social surplus. But of course, there could be several reasons for the persistence of ‘inefficient institutions’. When side payments are not feasible or credible, economic agents often prefer a larger share of a reduced pie to a smaller share of a more efficient pie. For instance, domestic businesses that can rely on a trusted network of kin or extended family might prefer an ambiguous legal system, which prevents entry. Or workers might prefer imperfect enforceability of a work norm, so that efficiency wages need to be paid. Borrowers might prefer that loan repayment cannot be fully enforced, so that incentives to repay must be built into the loan contract. And when tenancy is widespread in agriculture, the very design of overall property rights to maximize efficiency can be a highly complex problem.

The last three examples possess another feature that is worth some emphasis: ambiguous property rights often have equity effects that do not go the same way that efficiency-minded economists would like them to go (see Weitzman, 1974; Cohen and Weitzman, 1975; Baland and Platteau, 1996). The ambiguity of property rights can serve as insurance, buffer, or redistributive device. As examples, consider broad access to water resources or grazing land, or the efficiency-wage premia that may need to be paid to workers or borrowers.

Most importantly, the ambiguity of property rights slows down the emergence of an overt assetless class, and that has its own social value (it should not be forgotten that the flip side of unambiguous rights is exclusion). For example, Goldstein and Udry (2005) develop this point of view in the context of rural Ghana, arguing that the ambiguity in property rights prevented the outbreak of extreme poverty (and had an interesting efficiency effect in the bargain, as individuals were reluctant to leave the land fallow – an important investment – in the fear that this would signal a lack of need for land).

The political economy of rights is a messy business, but of central importance in development economics. Poverty in general enhances the social and political need for ambiguity, while to the extent that such ambiguity wears on efficiency, we have an extremely important instance of non-convergence. Sometimes such non-convergence assumes particularly dramatic form. In West Bengal (India) ‘Operation Barga’ provided widespread – and welcome – use rights to registered sharecroppers (see, for example, Banerjee, Gertler and Ghatak, 2001). Those very use rights now lie at the heart of recent difficulties in converting agricultural land in India for use in industry. In the world of the second best, few policies have unambiguously one-directional effects.

4 A concluding note: theory and empirics

While I have tried to provide a conceptual overview in this article, recent research in development economics has been almost entirely empirical. A veritable explosion in computing power, the expansion of institutional data-sets and their increased availability in electronic form, and the growing ease of collecting one's own data have bred a new generation of development economists. Their empirical sensibilities are of a high order; they are extremely sensitive to issues of endogeneity, omitted variables,

measurement error and biases induced by selection. They are constantly on the search for good instruments or natural experiments, and, when these are hard to find, they are adept at creating experiments of their own.

There is little doubt that we know little enough about the world we live in that it is often worth finding out the simple things, rather than continuing to engage in what some would term flights of theoretical fantasy. Are people really credit-rationed? Does rising income automatically make for better nutrition and health? If we had the option to throw in more textbooks, or reduce class size, or add more teachers, or install monitoring devices to track teacher attendance, which policy should we implement? Do women leaders behave differently from men in the policies that they adopt? Do households behave as one frictionless unit? Or, if one is the big-picture sort, have countries indeed converged over the last 200, or 500, years? Are richer countries more democratic? How many excess female deaths have occurred in China or India because of gender bias? Are poorer countries more ‘corrupt’? And so on. The list is practically endless.

The somewhat churlish theoretically minded economist might ask, why are well-trained statisticians unable to answer these questions? Why do we need economists, who are supposed, at the very least, to combine two observations to form a deduction? The answer, at one level, is very simple and not overly supportive of the churlish theorist's complaint. While the questions are straightforward, the answers are often extremely difficult to tease out from the data, and one needs a well-trained *economist*, not a statistician, to understand the difficulty and eliminate it. Because of the aforementioned econometric issues, not a single one of the questions asked above admits a straightforward answer. Development economists spend a lot of time thinking of inventive ways to get around these problems, and it is no small feat of creativity, dedication and extremely hard work to pull off a convincing solution.

It is true that the very desire to obtain a clean, unarguable answer – with its attendant desire to have control over the empirical environment – sometimes narrows the scope of the enquiry. There is often great reluctance to rely on theoretical structure (for such reliance would contaminate the near-lexicographic desire for an unambiguous result). This means that the question to be asked is often akin to that for a simple production function (for example, ‘do students do better in exams if they are given more textbooks?’) or is focused on the direct effect of some policy intervention (‘does the provision of health check-ups improve health outcomes?’). So it is that a boring but well-identified empirical question will often be treated with a great deal more veneration (especially if a clever instrument or randomization device is involved) than a model that relies on intuitive but undocumented assumptions.

That said, it is also a fact that we know very little about the answers to some of the most basic questions, such as the ones we have listed above. The great contribution of empirical development microeconomics is that we are building up this knowledge, piece by piece. Whether the search for that knowledge is informed by theory or not, there will be enough theorists to attempt to put these observations together. There will be enough empirical researchers to keep generating the hard knowledge. Development economics is alive and well.

See Also

- agriculture and economic development
- dual economies

- emerging markets
- endogenous growth theory
- growth and institutions
- poverty traps

Bibliography

Acemoglu, D. and Zilibotti, F. 1997. Was Prometheus unbound by chance? Risk, diversification and growth. *Journal of Political Economy* 105, 709–51.

Acemoglu, D., Johnson, S. and Robinson, J. 2001. The colonial origins of comparative development: an empirical investigation. *American Economic Review* 91, 1369–401.

Acemoglu, D., Johnson, S. and Robinson, J. 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 118, 1231–94.

Adserà, A. and Ray, D. 1998. History and coordination failure. *Journal of Economic Growth* 3, 267–76.

Aghion, P. and Bolton, P. 1997. A theory of trickle-down growth and development. *Review of Economic Studies* 64, 151–72.

Alesina, A., Baqir, R. and Easterly, W. 1999. Public goods and ethnic divisions. *Quarterly Journal of Economics* 114, 1243–84.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. and Wacziarg, R. 2003. Fractionalization. *Journal of Economic Growth* 8, 155–94.

Alesina, A. and La Ferrara, E. 2005. Ethnic diversity and economic performance. *Journal of Economic Literature* 43, 762–800.

Alston, L., Libecap, G. and Schneider, R. 1996. The determinants and impact of property rights: land titles on the Brazilian frontier. *Journal of Law, Economics, and Organization* 12, 25–61.

Appadurai, A. 2004. The capacity to aspire. In *Culture and Public Action*, ed. V. Rao and M. Walton. Stanford, CA: Stanford University Press.

Armendáriz, B. and Morduch, J. 2005. *The Economics of Microfinance*. Cambridge, MA: MIT Press.

Arthur, W. 1994. *Increasing Returns and Path-Dependence in the Economy*. Ann Arbor: University of Michigan Press.

Azariadis, C. and Drazen, A. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105, 501–26.

Baland, J.-M. and Platteau, J.-Ph. 1996. *Halting the Degradation of Natural Resources: Is there a Role for Rural Communities?* Oxford: Clarendon Press.

Baland, J.-M. and Platteau, J.-Ph. 1998. Wealth inequality and efficiency on the commons. Part II: the regulated case. *Oxford Economic Papers* 50, 1–22.

Banerjee, A. and Duflo, E. 2005. Growth theory through the lens of development economics. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Banerjee, A. and Duflo, E. 2007. The economic lives of the poor. *Journal of Economic Perspectives* 21 (1), 141–67.

Banerjee, A., Gertler, P. and Ghatak, M. 2002. Empowerment and efficiency: tenancy reform in West Bengal. *Journal of Political Economy* 110, 239–80.

Banerjee, A. and Iyer, L. 2005. History, institutions and economic performance: the legacy of colonial land tenure systems in India. *American Economic Review* 95, 1190–213.

Banerjee, A., Iyer, L. and Somanathan, R. 2007. Public action for public goods. Working Paper No. 12911. Cambridge, MA: NBER.

Banerjee, A., Mookherjee, D., Munshi, K. and Ray, D. 2001. Inequality, control rights and efficiency: a study of sugar cooperatives in western Maharashtra. *Journal of Political Economy* 109, 138–90.

Banerjee, A. and Newman, A. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101, 274–98.

Bardhan, P. and Udry, C. 1999. *Development Microeconomics*. New York: Oxford University Press.

Bardhan, P., Ghatak, M. and Karaivanov, A. 2006. Wealth inequality and collective action. Mimeo, London School of Economics.

Barro, R. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics* 106, 407–44.

Becker, G. and Tomes, N. 1986. Human capital and the rise and fall of families. *Journal of Labor Economics* 4, S1–39.

Bénabou, R. 1996. Inequality and growth. In *NBER Macroeconomics Annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Bénabou, R. 2000. Unequal societies: income distribution and the social contract. *American Economic Review* 90, 96–129.

Benhabib, J. and Rustichini, A. 1996. Social conflict and growth. *Journal of Economic Growth* 1, 125–42.

Bertocchi, G. and Canova, F. 2002. Did colonization matter for growth? An empirical exploration into the historical causes of Africa's underdevelopment. *European Economic Review* 46, 1851–71.

Besley, T. 1995. Property rights and investment incentives: theory and evidence from Ghana. *Journal of Political Economy* 103, 903–37.

Binswanger, H., Deininger, K. and Feder, G. 1995. Power, distortions, revolt and reform in agricultural land relations. In *Handbook of Development Economics*, vol. 3B, ed. J. Behrman and T. Srinivasan. Amsterdam: North-Holland.

Chiappori, P.-A. and Salanie, B. 2000. Testing for asymmetric information in insurance markets. *Journal of Political Economy* 108, 56–78.

Coate, S. and Ravallion, M. 1993. Reciprocity without commitment: characterization and performance of informal insurance arrangements. *Journal of Development Economics* 40, 1–24.

Cohen, J. and Weitzman, M. 1975. A Marxian view of enclosures. *Journal of Development Economics* 1, 287–336.

Collier, P. and Hoeffler, A. 1998. On economic causes of civil war. *Oxford Economic Papers* 50, 563–73.

Collier, P. and Hoeffler, A. 2004. Greed and grievance in civil war. *Oxford Economic Papers* 56, 563–95.

Dasgupta, P. 1993. *An Inquiry into Well-Being and Destitution*. Oxford: Clarendon Press.

Dasgupta, P. and Ray, D. 1986. Inequality as a determinant of malnutrition and unemployment: theory. *Economic Journal* 96, 1011–34.

David, P. 1985. Clio and the economics of QWERTY. *American Economic Review* 75, 332–7.

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, MD: Johns Hopkins Press, for the World Bank.

Demsetz, H. 1967. Towards a theory of property rights. *American Economic Review* 57, 347–59.

Djankov, S., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. 2003. Courts. *Quarterly Journal of Economics* 118, 453–517.

Do, Q.-T. and Iyer, L. 2006. Poverty, social divisions and conflict in Nepal. Mimeo, Harvard Business School.

Dorfman, R., Samuelson, P. and Solow, R. 1958. *Linear Programming and Economic Analysis*. Tokyo: McGraw-Hill Kogashuka.

Dower, P. and Potamites, E. 2006. Signaling credit-worthiness: land titles, banking practices and access to formal credit in Indonesia. Mimeo, Department of Economics, New York University.

Dube, O. and Vargas, J. 2006. Are all resources cursed? Coffee, oil and armed conflict in Colombia. Documentos de CERAC 002748. Bogota, Colombia.

Duflo, E. 2006. Poor but rational? In *Understanding Poverty*, ed. A. Banerjee, R. Bénabou and D. Mookherjee. New York: Oxford University Press.

Duflo, E. and Udry, C. 2004. Intrahousehold resource allocation in Cote d'Ivoire: social norms, separate accounts and consumption choices. Working Paper No. 10498. Cambridge, MA: NBER.

Easterly, W. and Levine, R. 1997. Africa's growth tragedy: policies and ethnic divisions. *Quarterly Journal of Economics* 111, 1203–50.

Eaton, J. and Gersovitz, M. 1981. Debt with potential repudiation: theoretical and empirical analysis. *Review of Economic Studies* 48, 289–309.

Emerson, P. 2006. Corruption, competition and democracy. *Journal of Development Economics* 81, 193–212.

Esteban, J. and Ray, D. 1994. On the measurement of polarization. *Econometrica* 62, 819–51.

Esteban, J. and Ray, D. 2001. Social rules are not immune to conflict. *Economics of Governance* 2, 59–67.

Esteban, J. and Ray, D. 2006. Inequality, lobbying and resource allocation. *American Economic Review*

96, 257–79.

Esteban, J. and Ray, D. 2007a. On the salience of ethnic conflict. Mimeo, Department of Economics, New York University.

Esteban, J. and Ray, D. 2007b. Polarization, fractionalization and conflict. *Journal of Peace Research*, forthcoming.

Fafchamps, M. 2003. *Rural Poverty, Risk, and Development*. Cheltenham: Edward Elgar.

Fafchamps, M. 2004. *Market Institutions and Sub-Saharan Africa: Theory and Evidence*. Cambridge, MA: MIT Press.

Fafchamps, M. and Lund, S. 2003. Risk sharing networks in rural Philippines. *Journal of Development Economics* 71, 261–87.

Fearon, J. 1995. Rationalist explanations for war. *International Organization* 49, 379–414.

Fearon, J. and Laitin, D. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97, 75–90.

Feder, G., Onchan, T., Chalamwong, Y. and Hongladarom, C. 1988. *Land Policies and Farm Productivity in Thailand*. Baltimore, MD: Johns Hopkins University Press.

Field, E. and Torero, M. 2006. Do property titles increase credit access among the urban poor? Evidence from a Nationwide Titling Program. Mimeo. Department of Economics, Harvard University.

Fields, G. 1980. *Poverty, Inequality and Development*. London: Cambridge University Press.

Floro, M. and Ray, D. 1997. Vertical links between formal and informal financial institutions. *Review of Development Economics* 1, 34–56.

Floro, M. and Yotopoulos, P. 1991. *Informal Credit Markets and the New Institutional Economics: The Case of Philippine Agriculture*. Boulder, CO: Westview.

Foster, A. 1995. Prices, credit constraints, and child growth in low-income rural areas. *Economic Journal* 105, 551–70.

Foster, A. and Rosenzweig, M. 1995. Learning by doing and learning from others: human capital and technical change in agriculture. *Journal of Political Economy* 103, 1176–209.

Frankel, D. and Pauzner, A. 2000. Resolving indeterminacy in dynamic settings: the role of shocks. *Quarterly Journal of Economics* 115, 283–304.

Freeman, S. 1996. Equilibrium income inequality among identical agents. *Journal of Political Economy* 104, 1047–64.

Galor, O. and Zeira, J. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60, 35–52.

Genicot, G. 2002. Bonded labor and serfdom: a paradox of voluntary choice. *Journal of Development Economics* 67, 101–27.

Genicot, G. and Ray, D. 2003. Group formation in risk-sharing arrangements. *Review of Economic Studies* 70, 87–113.

Ghosh, P., Mookherjee, D. and Ray, D. 2001. Credit rationing in developing countries: an overview of the theory. In *Readings in the Theory of Economic Development*, ed. D. Mookherjee and D. Ray. London: Basil Blackwell.

Goldstein, M. and Udry, C. 2005. The profits of power: land rights and agricultural investment in Ghana. Mimeo, Department of Economics, Yale University.

González, F. 2007. Effective property rights, conflict and growth. *Journal of Economic Theory*, forthcoming.

Henderson, J. 1988. *Urban Development: Theory, Fact, and Illusion*. Oxford: Oxford University Press.

Hess, G. 2003. The economic welfare cost of conflict: an empirical assessment. Working Paper Series No. 852, CESifo, Munich.

Hidalgo, F., Naidu, S., Nichter, S. and Richardson, N. 2007. *Occupational choices: economic determinants of land invasions. Mimeo, Department of Political Science*. Berkeley CA: University of California.

Hirschman, A. 1958. *The Strategy of Economic Development*. New Haven, CT: Yale University Press.

Hoff, K., Braverman, A. and Stiglitz, J. 1993. *The Economics of Rural Organization: Theory, Practice and Policy*. London: Oxford University Press.

Hurst, E. and Lusardi, A. 2004. Liquidity constraints, household wealth, and entrepreneurship. *Journal of Political Economy* 112, 319–47.

Iyer, L. 2005. The long-term impact of colonial rule: evidence from India. Working Paper No. 05-041, Harvard Business School.

Jacoby, H. and Skoufias, E. 1997. Risk, financial markets, and human capital in a developing country. *Review of Economic Studies* 64, 311–35.

Karlan, D. and Zinman, J. 2006. Observing unobservables: identifying information asymmetries with a consumer credit field experiment. Mimeo, Department of Economics, Yale University.

Khwaja, A. 2004. Is increasing community participation always a good thing? *Journal of the European Economic Association* 2, 427–36.

Kingston, C. 2005. Social structure and cultures of corruption. Mimeo, Department of Economics, Amherst College.

Kocherlakota, N. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63, 595–609.

Krugman, P. 1991. *Geography and Trade*. Cambridge, MA: MIT Press.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R. 1997. Legal determinants of external finance. *Journal of Finance* 52, 1131–50.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R. 1998. Law and finance. *Journal of Political Economy* 106, 1113–55.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R. 2002. Investor protection and corporate valuation. *Journal of Finance* 57, 1147–70.

Lanjouw, J. and Levy, P. 2002. Untitled: a study of formal and informal property rights in urban Ecuador. *Economic Journal* 112, 986–1019.

Leibenstein, H. 1957. *Economic Backwardness and Economic Growth*. New York: Wiley.

Leventoglu, B. and Slantchev, B. 2005. *The armed peace: a punctuated equilibrium theory of war. Mimeo, Department of Political Science*. San Diego: University of California.

Levy, G. 2004. A model of political parties. *Journal of Economic Theory* 115, 250–77.

Ligon, E. 1998. Risk-sharing and information in village economies. *Review of Economic Studies* 65, 847–64.

Ligon, E., Thomas, J. and Worrall, T. 2002. Mutual insurance and limited commitment: theory and evidence in village economies. *Review of Economic Studies* 69, 209–44.

Ljungqvist, L. 1993. Economic underdevelopment: the case of missing market for human capital. *Journal of Development Economics* 40, 219–39.

Loury, G. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49, 843–67.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Lucas, R. 1990. Why doesn't capital flow from rich to poor countries? *American Economic Review* 80, 92–6.

Majumdar, M. and Mitra, T. 1982. Intertemporal allocation with a non-convex technology: the aggregative framework. *Journal of Economic Theory* 27, 101–36.

Mankiw, N., Romer, D. and Weil, D. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107, 407–38.

Mansuri, G. 1997. Credit layering in rural financial markets: theory and evidence from Pakistan. Ph.D. thesis, Boston University.

Matsuyama, K. 2000. Endogenous inequality. *Review of Economic Studies* 67, 743–59.

Mauro, P. 1995. Corruption and growth. *Quarterly Journal of Economics* 110, 681–712.

Miguel, E. and Gugerty, M. 2005. Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of Public Economics* 89, 2325–68.

Miguel, E., Satyanath, S. and Sergenti, E. 2004. Economic shocks and civil conflict: an instrumental variables approach. *Journal of Political Economy* 112, 725–53.

Montalvo, J. and Reynal-Querol, M. 2005. Ethnic polarization, potential conflict, and civil wars. *American Economic Review* 95, 796–813.

Mookherjee, D. and Ray, D. 2001. *Readings in the Theory of Economic Development*. London: Basil Blackwell.

Mookherjee, D. and Ray, D. 2002a. Contractual structure and wealth accumulation. *American Economic Review* 92, 818–49.

Mookherjee, D. and Ray, D. 2002b. Is equality stable? *American Economic Review* 92, 253–9.

Mookherjee, D. and Ray, D. 2003. Persistent inequality. *Review of Economic Studies* 70, 369–94.

Munshi, K. and Myaux, J. 2006. Social norms and the fertility transition. *Journal of Development Economics* 80, 1–38.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Industrialization and the big push. *Journal of Political Economy* 97, 1003–26.

Myrdal, G. 1957. *Economic Theory and Underdeveloped Regions*. London: Duckworth.

Nurkse, R. 1953. *Problems of Capital Formation in Underdeveloped Countries*. New York: Oxford University Press.

Obstfeld, M. 1994. The logic of currency crises. *Cahiers Economiques et Monétaires (Banque de France)* 43, 189–213.

Olson, M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Pande, R. and Udry, C. 2007. Institutions and development: a view from below. In *Proceedings of the 9th World Congress of the Econometric Society*, ed. R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press.

Parente, S. and Prescott, E. 2000. *Barriers to Riches*. Cambridge, MA: MIT Press.

Piketty, T. 1997. The dynamics of the wealth distribution and the interest rate with credit rationing. *Review of Economic Studies* 64, 173–89.

Perotti, R. 1994. Income distribution and investment. *European Economic Review* 38, 827–35.

Perotti, R. 1996. Growth, income distribution, and democracy: what the data say. *Journal of Economic Growth* 1, 149–87.

Powell, R. 2004. The inefficient use of power: costly conflict with complete information. *American Political Science Review* 98, 231–41.

Powell, R. 2006. War as a commitment problem. *International Organization* 60, 169–203.

Pritchett, L. 1997. Divergence, big time. *Journal of Economic Perspectives* 11(3), 3–17.

Quah, D. 1996. Twin peaks: growth and convergence in models of distribution dynamics. *Economic Journal* 106, 1045–55.

Ray, D., Baland, J.-M. and Dagnielie, O. 2007. Inequality and inefficiency in joint projects. *Economic Journal* 117, 922–35.

Ray, D. and Sengupta, K. 1989. Interlinkages and the pattern of competition. In *The Economic Theory of Agrarian Institutions*, ed. P. Bardhan. Oxford: Clarendon Press.

Ray, D. and Streufert, P. 1993. Dynamic equilibria with unemployment due to undernourishment. *Economic Theory* 3, 61–85.

Ray, D. 1998. *Development Economics*. Princeton, NJ: Princeton University Press.

Ray, D. 2006. Aspirations, poverty and economic change. In *Understanding Poverty*, ed. A. Banerjee, R. Bénabou and D. Mookherjee. New York: Oxford University Press.

Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 92, 1002–37.

Robinson, J. 1998. Theories of ‘bad policy’. *Policy Reform* 1, 1–46.

Rosenstein-Rodan, P. 1943. Problems of industrialization of eastern and southeastern Europe. *Economic Journal* 53, 202–11.

Rosenzweig, M. and Wolpin, K. 1993. Credit market constraints and the accumulation of durable production assets in low-income countries: investments in bullocks. *Journal of Political Economy* 101, 223–44.

Sambanis, N. 2001. Do ethnic and nonethnic civil wars have the same causes? A theoretical and empirical inquiry (part 1). *Journal of Conflict Resolution* 45, 259–82.

Sandler, T. and Forbes, J. 1980. Burden sharing, strategy and the design of NATO. *Economic Inquiry* 18, 425–44.

Sen, A. 1999. *Development as Freedom*. New York: Alfred A. Knopf.

Sokoloff, K. and Engerman, S. 2000. History lessons: institutions, factor endowments, and paths of development in the new world. *Journal of Economic Perspectives* 14(3), 217–32.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70,

65–94.

Stiglitz, J. 2002. *Globalization and its Discontents*. New York: W.W. Norton.

Stiglitz, J. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

Townsend, R. 1993. Risk and insurance in village India. *Econometrica* 62, 539–91.

Townsend, R. 1995. Consumption insurance: an evaluation of risk-bearing systems in low-income economies. *Journal of Economic Perspectives* 99(3), 83–102.

Udry, C. 1994. Risk and insurance in a rural credit market: an empirical investigation in northern Nigeria. *Review of Economic Studies* 61, 495–526.

Udry, C. 1996. Gender, agricultural productivity and the theory of the household. *Journal of Political Economy* 104, 1010–45.

Weitzman, M. 1974. Free access vs. private ownership as alternative systems for managing common property. *Journal of Economic Theory* 8, 225–34.

World Bank. 2003. *World Development Report*. London: Oxford University Press.

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

How to cite this article

Ray, Debraj. "development economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000105> doi:10.1057/9780230226203.0385

The New Palgrave Dictionary of Economics Online

dialectical reasoning

Gareth Stedman Jones
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; civil society; dialectical materialism; dialectical reasoning; Engels, F.; equal exchange; Hegel, G. W. F.; Marx, K. H.; materialist conception of history; political economy; Proudhon, P. J.; Ricardo's theory of value; surplus labour; use value.

Article

This notoriously elusive and multifaceted notion assumed importance in the history of political economy because Marx's ‘critique of political economy’, *Capital*, and particularly its first draft, the *Grundrisse* of 1857–8, was presented in a dialectical form. Part of the difficulty of encapsulating the dialectic within any concise definition derives from the fact that it may be conceived as a method of thought, a set of laws governing the world, the immanent movement of history or any combination of the three. The dialectic originated in ancient Greek philosophy. The original meaning of ‘*dialogos*’ was to reason by splitting in two. In one form of its development, dialectic was associated with reason. Starting with Zeno's paradoxes, dialectical forms of reasoning were found in most of the philosophies of the ancient world and continued into medieval forms of disputation. It was this form of reasoning that Kant attacked in his distinction between the logic of understanding which, applied to the data of sensation, yielded knowledge of the phenomenal world, and dialectic or the logic of reasoning, which proceeded independently of experience and purported to give knowledge of the transcendent order of things in themselves. In another form of dialectic, the focus was primarily upon process: either an ascending dialectic in which the existence of a higher reality is demonstrated, or a descending form in which this higher reality is shown to manifest itself in the phenomenal world. Such conceptions were particularly associated with Christian eschatology, neo-platonism and illuminism, and typically patterned themselves into conceptions of original unity, division or loss, and ultimate reunification.
For practical purposes, however, the form in which the dialectic was inherited and modified by Marx was that in which it had been elaborated by Hegel. ‘Hegel's dialectics is the basic form of all dialectics, but only *after* it has been stripped of its mystified form, and it is precisely this which distinguishes my method’ (Marx, letter to Kugelmann, 6 March 1868).

In Hegel, the dialectic is a self-generating and self-differentiating process of reason (reason being understood both to be the process of cognition and the process of the world). The Hegelian Absolute actualizes itself by alienating itself from itself and then by restoring its self-unity. This corresponds to the three basic divisions of the Hegelian system: the *Logic*, the *Philosophy of Nature* and the *Philosophy of Mind*. It is free because self-determined. Its freedom consists in recognizing that its alienation into its other (nature) is but a free expression of itself. The truth is the whole and it unfolds through a dialectical progression of categories, concepts and forms of consciousness from the most simple and empty to the most complex and concrete. Each category reveals itself to the observer to be incomplete, lacking and contradictory; it thus passes over into a more adequate category capable of resolving the one-sided and contradictory aspects of its predecessor, though throwing up new contradictions in its turn. Against Kant, this process of dialectical reason is not concerned with the transcendent, but is immanent in reality itself. Reflective understanding is not false, but partial. It abstracts from reality and decomposes objects into their elements. Analytic understanding represents a localized standpoint which sets up an unsurpassable barrier between subject and object and thus cannot grasp the systematic interconnection between things or the total process of which it is a part. The absolute subject contains both itself and its other (both being and thought) which is revealed to be identical with itself. Human history, human thought are vehicles through which the absolute achieves self-consciousness, but humanity as such is not the subject of the process. Thus the absolute spirit dwells in human activity without being reducible to it, just as the categories of the *Logic* precede their embodiment in nature and history.

The character of the Marxian dialectic is yet harder to pin down than that of Hegel. In some well-known lines in the Post-Face to the Second Edition of *Capital* in 1873, Marx stated,

> I criticised the mystificatory side of the Hegelian dialectic nearly thirty years ago … [but] the mystification which the dialectic suffers in Hegel's hands by no means prevents him from being the first to present its general form of motion in a comprehensive and conscious manner. With him it is standing on its head. It must be inverted in order to discover the rational kernel within the mystical shell. (Marx, 1873, pp. 102–3)

This statement has satisfied practically no one. How can a dialectic be inverted? How can a rational kernel be extracted from a mystical shell? To critics from empiricist, positivist or structuralist traditions, anxious to free Marx from the clutches of Hegelianism, the dialectic is intrinsically unworkable and must either be dropped or stated in quite other terms (for example, Bernstein, 1899; Della Volpe, 1950; Althusser, 1965; Cohen, 1978; Elster, 1985). To a second group, the dialectical understanding of capitalism is only a particular instance of more general dialectical laws which govern reality as a whole, both natural and social (Engels, dialectical materialism). To a third group, the Hegelian roots of Marx's thought are not sufficiently emphasized in this statement; Marxism is only Hegelianism taken to its logical revolutionary conclusions in the discovery of the proletariat as the subject–object of history and the ‘totality’ as the distinguishing feature of its world-outlook (Lukács, 1923 and much of 20th-century Western Marxism). This *Methodenstreit* cannot be discussed here. All that can be attempted is to give some sense to Marx's statement and in particular to indicate how it informed his critique of political economy.

Marx specifically criticized ‘the mystificatory side of the Hegelian dialectic’ in his 1843 *Critique of*

Hegel's Philosophy of Right and in the concluding section of the *1844 Manuscripts* (both of which were only published in the 20th century). In these texts, Marx followed Feuerbach in considering Hegelian philosophy to be the conceptual equivalent of Christian theology; both were forms of alienation of man's species attributes; Christianity transposed human emotion into a religious Godhead, while Hegel projected human thinking into a fictive subject, the Absolute Idea, which in turn then supposedly generated the empirical world. Employing Feuerbach's ‘transformative method’ (the origin of the inversion metaphor), subject and predicate were reversed and hence the correct starting point of philosophy was the finite, man. Nature similarly was not the alienated expression of Absolute Spirit, it was irreducibly distinct. Thus there could be no speculative identity of being and thought. Man, however, as a natural being, could interact harmoniously with nature, his inorganic body. Once the absolute spirit had been dismantled and the identity of being and thought eliminated, it could be argued that the barrier against the harmonious interpenetration of man and nature and the free expression of human nature, was not ‘objectification’, the division between subject and object constitutive of the finite human condition, but rather the inhuman alienation of man's species life activity in property, religion and the state. True Communism, humanism, meant the re-appropriation of man's essential powers, the generic use of his conscious life activity. In contrast to the predominant Young Hegelian position, therefore, which counterposed Hegel's revolutionary ‘method’ (the dialectic) to his ‘conservative system’, Marx argued that there was no incompatibility between the two. For while Hegel's dialectic ostensibly negated the empirical world, it covertly depended upon it. Not only was the moment of contradiction a prelude to the higher moment of reconciliation and the restoration of identity, but the ideas themselves were tacitly drawn from untheorized experience. The effect of the dialectical chain which embodied the world was not to subvert the existing state of affairs, but to sanctify it.

In the crucial period that followed, that of the *German Ideology* and the *Poverty of Philosophy*, in which the basic architecture of the ‘materialist conception of history’ was elaborated, the attack upon speculative idealism was made more radical. The generic notion of ‘conscious life activity’, ‘praxis’, was replaced by the more specific notion of production. Hegel and the Idealist tradition were given credit for emphasizing the active transformative side of human history, but castigated for recognizing this activity only in the form of thought. Thought itself was now made a wholly derivative activity. The fundamental activity was labour and what developed in history were the productive powers men employed in their interaction with nature, ‘the productive forces’. Stages in the development of these productive forces were accompanied by successive ‘forms of human intercourse’, what became ‘the relations of production’. Finally, ‘man’ as a generic being was dispersed into the struggle between different classes of men, between those who produced and those who owned and controlled the means of production.

In this new theorization of history, explicit references to Hegel were few and the dialectic scarcely mentioned. But Hegel re-entered the story as soon as Marx attempted to write up a systematic theory of the capitalist mode of production in 1857–8. To see why, we must briefly survey his economic writings up to that date.

Marx's 1843 critique of Hegel had led him to the conclusion that civil society was the foundation of the state and that the anatomy of civil society was to be found in political economy. However, if his preoccupation with political economy dated from this point, it was not that of an economist. In the 1844 Manuscripts what is to be found is a humanist critique of both political economy and civil society: not an alternative theory of the economy, but rather a juxtaposition between the ‘economic’ and the

‘human’, the former being judged in terms of the latter. No distinction is made between political economy and the economic reality it purports to address, the one is simply seen as the mirror of the other. The first attempt to define capitalism as an economic phenomenon occurred in the *Poverty of Philosophy* (1847). However, whatever the significance of that work in other respects, it did not outline any specifically Marxian portrayal of the capitalist economy. As in 1844 there was no internal critique of classical political economy. The main difference was that, whereas in 1844 Marx saw that economy through the eyes of Adam Smith, he now saw it through the eyes of Ricardo. In particular, he adopted what he took to be Ricardo's theory of value and belaboured Proudhon for positing as an ideal – the equivalence of value and price – what he considered to be the actual situation under capitalism. The only critique of Ricardo to be found there was a purely external historicist one: that Ricardo was the scientific expression of the epoch of capitalist triumph, but that that epoch had already passed away, that its gravediggers had already appeared and that its collapse was already at hand.

When Marx resumed his economic studies after the 1848 revolutions, Proudhonism was still the main object of attack. It occupied a major part of his unfinished economic manuscripts of 1850–1 and the attack on the Proudhonist banking schemes of Darimon took up the first part of the written-up notebooks of 1857–8, the *Grundrisse*. Proudhonism was the main object of attack because it could be taken for the predominant form of socialist or radical reasoning about the economy. Ricardo could again be utilized to attack such reasoning in order to argue that it represented a nostalgia for petty commodity production under conditions of equal exchange, a situation supposedly preceding modern capitalism rather than representing an emancipation from it. However, if the capitalist mode of production and its historical limits were to be grasped in theory, this would have to involve a critique of Ricardo himself.

The form this critique took, involved problematizing Ricardo's theory of value (or rather Marx's reading of it:). Steedman (1979) has argued strongly that Marx misconstrued Ricardo's theory, though Ricardo's shifting of position between the three editions of the *Principles* and the fact that Marx only used the third edition makes his mistake an understandable one). On the one hand, it raised a question never posed by Ricardo: the source of profit in a system of equal exchange. On the other hand, it involved juxtaposing wealth in the form of productive forces, that is, as a collection of use values against the translation of all wealth into exchange values within capitalism. Ricardo, it was argued, possessed no criterion for distinguishing between the content – or the material elements – and the form of the economy, such as Marx possessed in the distinction between forces and relations of production. Ricardo never problematized the ‘value form’; he linked the object of measurement with the measurement itself. For this reason, Ricardo was considered to possess no conception of the historicity of capitalism. Once the material could be distinguished from the social, the content from the form, the capitalist mode of production could be conceived as a dynamic system whose principle of movement could be located in the contradictory relationship between matter and form.

It is here that Hegel came in. We know that during the writing of the *Grundrisse* at the beginning of 1858, Marx re-read Hegel, in particular the *Science of Logic*. He wrote to Engels, ‘I am getting some nice developments, e.g. I have overthrown the entire doctrine of profit as previously conceived. In the method of working, it was of great service to me that by mere accident I leafed through Hegel's *Logic* again’ (Marx to Engels, 16 January 1858).

What Marx found so useful in his reading of Hegel's *Logic* at this time is not really mysterious. It suggested a way of elaborating the contradictory elements that Marx had discerned in the value form into a theorization of the trajectory of the capitalist mode of production as a whole. The point is

emphasized by Marx in his Post-Face to *Capital*: the dialectic includes in its positive understanding of what exists a simultaneous recognition of its negation, its inevitable destruction; because it regards every historically developed form as being in a fluid state, in motion, and therefore grasps its transient aspect as well (1873, p. 103). The dialectic offered a means of grasping a structure in movement, a process – the subtitle of *Capital*, Volume 1, was ‘the process of capitalist production’. If capitalism could be represented as a process and not just a structure, then concomitantly its building blocks were not factors, but, as in Hegel, ‘moments’. As Marx put it in the *Grundrisse:*

> When we consider bourgeois society in the long view and as a whole, then the final result of the process of social production always appears as the society itself i.e. the human being itself in its social relations. Everything that has a fixed form, such as the product etc., appears as merely a moment, a vanishing moment in this movement. The conditions and objectifications of the process are themselves equally moments of it, and its only subjects are the individuals, but individuals in mutual relationships, which they equally reproduce and produce anew …. in which they renew themselves even as they renew the world of wealth they create. (Marx [1857–8], p. 712)

Marx's attempt to utilize the *Logic* can be seen most clearly in the *Grundrisse*. There one can see the genesis of particular concepts which in *Capital* appear in more polished form. What is clear is that the *Logic* is used as a first means of setting terms in relation to each other. The text is littered with Hegelian expressions and turns of phrase; indeed, sometimes it appears as if lumps of Hegelian ratiocination have simply been transposed, undigested, to sketch the more intractable links in the chain. Here, for instance, is money striving to become capital: ‘… already for that reason, value which insists on itself as value preserves itself through increase; and it preserves itself precisely only by constantly driving beyond its quantitative barrier, which contradicts its character as form, its inner generality’ (p. 270). But at the same time we can see Marx remind himself to correct the ‘idealist manner of presentation, which makes it seem as if it were merely a matter of conceptual determination and of the dialectic of these concepts’ (p. 151).

But the interest of dialectical logic for Marx was not simply that it offered him a way of outlining a structure in movement; more fundamentally it enabled him to depict contradiction as the motor of this movement. This was why the dialectic was ‘in its very essence critical and revolutionary’ (Marx, 1873, p. 103), in that both in Hegel and in ancient Greek usage movement was contradiction. This appears closely in the dramatic relationship that Marx sets up between the circulation system and the production system in *Capital*. The system of exchange of the market is the public face of capitalism. It is ‘in fact a very Eden of the innate rights of Man’ (p. 280). Exchanges are equal. To look for the source of inequality in the exchange system, like the Proudhonists, is to look in the wrong place. Yet, if exchanges are equal, how does capital accumulation take place? Equal exchange implies the principle of identity, of non-contradiction. It is, in Hegel's sense, the sphere of ‘simple immediacy’, the world as it first appears to the senses. It cannot move or develop, because it apparently contains no contradictory relations.

But this surface of things is not self-sufficient. It is ‘the phenomenon of a process taking place behind it’. As a surface it is not nothing, but rather a boundary or limit. Contradiction and therefore movement is located in production. Here there is non-identity, the extraction of surplus labour disguised by the surface value form and its tendency to limitless expansion.

Thus, there are two processes, on the one hand that of the surface, that of immediate identity lacking the motive power of its own regeneration; on the other hand, that beneath the surface, a process of contradiction. Thus in Hegelian terms, the whole could then be defined as ‘the identity of identity and non-identity’. In this whole, contradiction is the overriding moment, but the surface places increasingly formidable obstacles to its development, for instance, so-called ‘realization’ crises. Values can only be realized in an act of exchange and the medium of this exchange is money. But there is no guarantee that these exchanges must take place. The ‘anarchy’ of the market place is such that overproduction or disproportionality between sectors of production can only be seen after the event. Hence trade crises and slumps (see M. Nicolaus, Introduction to Marx [1857–8]).

This is only one example of how Marx employed dialectical principles in his attempt to conceptualize the process or movement of a contradictory whole. Another would be the six books Marx originally planned to write in 1857–8, the original blueprint of *Capital*. Their order would have been: Capital, Wage Labour, Landed Property, State, World Market, Crises. This plan is reminiscent of Hegel's *Encyclopaedia*. It describes a circle in a Hegelian sense. The point of departure is not capital per se, but commercial exchange as appearance, then proceeding through the contradictory world of production and eventually returning to commercial exchange again as the world market, but this time enriched by the whole of the preceding analysis.

There has been much controversy about the proximity or distance between the Hegelian and Marxian dialectics. Those who like Althusser (1965) argue for their radical dissimilarity, are on their strongest ground when arguing that in Marx the terms of the dialectic have been radically transformed. The contradiction between forces and relations of production cannot be reduced to the ultimate simplicity of that between Hegel's master and slave or of that between proletariat and bourgeoisie in the Hegelianized Marxist account of Lukacs. But it is far more difficult to establish as unambiguously the difference in the relationship between the terms in their respective dialectics. On the one hand, the relation between matter and form in Hegel is only one of apparent exteriority. Matter relates to form as other only because form is not yet posited within it. Once the terms are related, they are declared to be identical. Marx, on the other hand, insists upon the irreducible difference between matter and form, between the material and the social (even if he is not wholly successful in keeping them apart). Not only are matter and form different, but the one determines the other: value is determined in relation to the material production of use value; the opposite is not true. Relations of determination would seem to exclude identity, and this is confirmed by Marx's avoidance of the Hegelian notion of ‘sublation’ (*Aufhebung*), the higher moment of synthesis. The dialectical clash between forces and relations of production in the capitalist mode of production does not of itself produce a higher unity (socialism); rather what crises do, is to make manifest the otherwise hidden determination of value by use value, of form by matter. Against this, however, must be set one or two passages, including a famous peroration in *Capital* Volume 1, where Marx does conceive the end of capitalism as a return to a higher but differentiated unity and does employ the notion of the negation of the negation (Marx, 1873, p. 929), and, despite the best efforts of some modern commentators, it is difficult honestly to deny the strongly teleological imagination which underpins the whole enterprise of *Capital*.

Finally, in two important respects, Hegelian dialectic, however surreal, is less vulnerable than that of Marx. Firstly, Hegel's *Science of Logic* takes place outside spatio-temporal constraints. It is a purely logical progression of concepts, even if the principles on which one ontological category is derived from another ‘have resisted analysis to this day’ (Elster, 1985, p. 37). Marx's effort to avoid giving any

impression of the ‘self-determination’ of the concept, took the form of attempting to demonstrate that ‘the ideal is nothing but the material world reflected in the mind of man and translated into forms of thought’ (Marx, 1873, p. 102). In practical terms this implied that there was some systematic relationship between the logical sequence of concepts in the exposition of the argument and the chronological order of their appearance in historical time. But this turned out to impose insurmountable difficulties in terms of presentation (and it is significant that, having begun with the product in the *Grundrisse*, he began with the commodity in *Capital*). Thus Marx both stated his position and violated it, bequeathing insoluble ambiguities surrounding his interpretation of value, of the meaning of ‘reflection’ and of the relationship between history and logic which have plagued even his closest followers ever since. Secondly, when it came to applying his dialectic to history, Hegel was categorical in refusing to project his theory into the future. The philosopher could explain the rationality of what had happened; it was only then that it could be grasped in thought. Marx, despite all his strictures against the voluntarism of other Young Hegelians and some of his fellow revolutionaries, was unable by the very nature of his project, fully to abide by the Hegelian restriction. Thus, while Hegel's owl of Minerva flew at dusk, the Marxian owl, unfortunately, took flight at high noon.

Bibliography

Althusser, L. 1965. *For Marx*. London: Allen Lane, 1969.

Bernstein, E. 1899. *Evolutionary Socialism*. Stuttgart. English trans. E.C. Harvey, London: Independent Labour Party, 1909.

Bhaskar, R. 1983. *Dialectic, Materialism and Human Emancipation*. London: New Left Books.

Cohen, G. 1978. *Karl Marx's Theory of History: A Defence*. London: Oxford University Press.

Della Volpe, G. 1950. *Logica come scienza positiva*. Messina: G. d'Anna.

Elster, J. 1985. *Making Sense of Marx*. Cambridge: Cambridge University Press.

Hegel, G.W.F. 1812–16. *The Science of Logic*. London: Allen & Unwin, 1961.

Kolakowski, L. 1978. *Main Currents of Marxism* Vol. I: *The Founders*, Oxford: Oxford University Press.

Lukács, G. 1923. *History and Class Consciousness*. London: Merlin, 1971.

Marx, K. 1844. *Economic and Philosophical Manuscripts of 1844*. In K. Marx and F. Engels, *Collected Works*, vol. 3. London: Lawrence & Wishart, 1975.

Marx, K. 1847. *The Poverty of Philosophy*. In *Collected Works*, vol. 6, London: Lawrence & Wishart,

1976.

Marx, K. [1857–8]. *Grundrisse*. Harmondsworth: Penguin 1973.

Marx, K. 1873. *Capital*, vol. I. 2nd edn., Harmondsworth: Penguin, 1976.

Rosdolsky, R. 1968. *The Making of Marx's Capital*. Trans. P. Burgeis, London: Pluto Press, 1977.

Steedman, I. 1979. Marx on Ricardo. Discussion Paper No. 10, Department of Economics, University of Manchester.

How to cite this article

Jones, Gareth Stedman. "dialectical reasoning." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000109> doi:10.1057/9780230226203.0386

The New Palgrave Dictionary of Economics Online

D í az-Alejandro, Carlos (1937 – 1985)

José Gabriel Palma
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Carlos Díaz-Alejandro was the most prominent Latin American economist of his generation. In his short professional life he gave us powerful insights into Latin America's trade and development, and its economic and financial history. In true Kindlebergian tradition, he was particularly fascinated by the region's many financial crises. His contributions were characterized by a rare capacity to weave together history and theory, abstract economic theory and complex Latin American socio-political life. In this way, he avoided the sterility of pure formalistic theory that characterized so much of the economics of his own generation and the next.

Keywords

absorption approach to the balance of payments; capital account liberalization; capital controls; current account liberalization; devaluation; Díaz-Alejandro, C.; economic history; elasticity approach to the balance of payments; export-led growth; Hirschman, A. O.; import substitution; import-substituting industrialization; inflation; International Monetary Fund; Kindleberger, C.; Prebisch, R.; redistribution of income

Article

Carlos Díaz-Alejandro was born in Havana and died in New York one day short of his 48th birthday. At 32, he became Yale's youngest ever full professor of economics. In 1983 he moved to Columbia, and at the time of his sudden death he had just accepted a chair at Harvard.
During sabbatical leaves, he visited many Latin American and European universities. Among numerous other activities he was a (dissenting) member of the Kissinger Commission on Central America. He strongly criticized US support for the ‘Contras’ in Nicaragua (para-military groups associated with the Somoza dictatorship, opposed to the Sandinista government), and insisted that if the United States were serious about Central America it should tie economic assistance to human rights and allow Central

American exports free access into its own market. Needless to say, such quixotic attempts to influence US foreign policy were never among his greatest successes!

From a personal point of view I admired his sense of humour and wit, his approachability and ‘bridge-building’ capacity, his aversion to positions of administrative power, his independence of mind and his common sense.

His work gave us powerful insights into Latin America's trade and development, and its economic and financial history. He was particularly fascinated by the region's many financial crises. His contributions were characterized by a rare capacity to weave together history and theory, abstract economic theory and complex Latin American socio-political life.

In his doctoral dissertation at MIT, Díaz-Alejandro revisited the controversy between the ‘elasticity’ and ‘absorption’ approaches to the balance of payments in the context of Argentina's experience of devaluation, concluding that on balance it supported the first approach (1965a). He further argued that one of the main mechanisms through which devaluation influences both the balance of payments and economic growth is through its effects on income distribution. The apparent paradox that many devaluations improve the trade balance but negatively affect the overall growth of output could be explained by the complex redistributive effects of devaluation. In fact, the effectiveness of a devaluation may depend more on the nature of its distributional outcome than on its capacity to change relative prices. Therefore, the exchange rate could be seen as yet another sphere in the struggle between different groups over their shares in national income.

Another peculiar feature of semi-industrialized economies is that ‘[i]n the long run, the success or failure of a stabilisation effort will depend more on the capacity of governments to obtain a national consensus over the objectives and policy instruments than on the approval or help that they could receive from foreign investors or governments and international agencies’ (1985, p. 201).

Díaz-Alejandro also maintained a keen interest in Latin American economic history, writing first on Argentina (1970). Then, in an article on the 1930s crisis, he identified the causes of the dissimilar performances of Latin American economies in the fact that some countries pursued an ‘active’ approach to fighting recession, while others stuck to conventional ‘passive’ adjustment mechanisms (1982). The ‘active’ countries were mainly the large ones, but also included Chile and Uruguay. They performed much better by abandoning the gold standard and by adopting flexible monetary and fiscal policies, real devaluations, moratoria on their foreign debt, and spending massively on public works. This heterodox response of some countries was in part a reaction to the emergence after the 1929 crash of a protectionist, interventionist and nationalistic Centre.

Díaz-Alejandro's articles on trade and development also discussed the high import intensity of import substitution (1965b), and the transition from import-substituting industrialization to export-led growth (1974). Diaz-Alejandro was particularly sceptical about the idea that this transition would help achieve both faster and more equitable growth. He strongly supported export orientation, but did not believe that it could be achieved simply by ‘getting the prices right’; he also feared that it could contribute to ‘stop-go’ macroeconomics. Moreover, he thought that most of the advice given to Third World countries for their trade policies ‘… suggest evangelical fervour rather than scientific analysis’ (1980, p. 332). Díaz-Alejandro re-examined all these issues in his book on Colombia (1975).

He was also a critic of the intervention of the International Monetary Fund (IMF) in markets which were not within its competence:

> It is the business of the IMF to insist on balance of payments targets … It is not the business of the IMF to make loans conditional on … food subsidies, utility rates, or controls over foreign corporations… It was a brilliant administrative stroke for the IMF staff to develop the 'monetary approach to the balance of payments' during the 1950s, allowing the translation of balance of payments targets into those involving domestic credit, but for many LDCs [less developed countries] the assumptions needed to validate such translation, such as a stable demand for money, have become less and less convincing. (1984, p. 169)

He also strongly criticized the IMF intervention in the debt crisis of the 1980s: 'Since August 1982 the world has lived with … a peculiar semi-cartelization shakily managed by central banks and the IMF [which] imposes on countries like Brazil the costs of monopoly (for example, larger spreads and fees) without some of its benefits (the ability to plan ahead)' (1983, p. 32).

The economic reforms of the late 1970s and 1980s provided another major intellectual challenge. Not since the 1930s had Latin America witnessed such dramatic economic and political experiments. The new military regimes of the Southern Cone applied their Chicago-oriented policies with a degree of ferocity that rivalled their treatment of political dissent. As Velasco said, Díaz-Alejandro's wisdom was twice as useful because it was delivered in a timely fashion (1988, p. 5). His papers of the late 1970s contain the basic ideas which later became accepted wisdom regarding the policy mistakes of the pro-Chicago governments in Latin America and the irrational behaviour of borrowers and lenders in (highly liquid) national and international financial markets. He particularly questioned the feasibility of simultaneous current and capital account liberalization, the lack of capital controls on speculative inflows, and the use of exchange rate policy to fight inflation.

Among his many articles from this period, his 'Southern Cone Stabilisation Plans' (1981) stands out. Appearing just before the Mexican moratorium which triggered the debt crisis, his argument ran completely against the tide of dominant opinion. Finally, a detailed analysis of the dynamics of the 1982 crisis was the last – and probably best known – of Díaz-Alejandro's contributions (see Palma, 2003).

Díaz-Alejandro began his studies at MIT at the time when Fidel Castro landed clandestinely in Cuba in 1956, and graduated at the time of the Bay of Pigs invasion (an unsuccessful CIA-planned and funded invasion by Cuban exiles in south-west Cuba in 1961). He felt that the complexity of the situation was such that he opted for the Miltonian hope that 'they also serve who only stand and wait'.

He had a fascination with Latin American economics. His approach was firmly grounded in the real world, and his work on economic history was rooted in the idea that all history is always the history of the present. As Gustav Ranis remarked, he always 'respected history, used data carefully, and theory selectively' (1989, p. xiv). Like his mentors Hirschman, Kindleberger, Lewis and Prebisch he basically belonged to the 'markets are good servants but bad masters' Keynesian school of economic thought, and always studied economic problems in their historical context, thus avoiding the sterility of pure formalistic theory that characterized so much of the economics of his own generation and the next.

See Also

- elasticities approach to the balance of payments

- Furtado, Celso
- Kindleberger, Charles P.
- Lewis, W. Arthur
- Prebisch, Raúl
- structuralism
- terms of trade
- third world debt

Selected works

1965a. *Exchange Rate Devaluation in a Semi-industrialized Country: Argentina 1955–1961*. Cambridge, MA: MIT Press.

1965b. On the import intensity of import substitution. Repr. in Velasco (1988).

1970. *Essays on the Economic History of Argentine*. New Haven, CT: Yale University Press.

1974. Some characteristics of recent export expansion in Latin America. Repr. in Velasco (1988).

1975. *Foreign Trade Regimes and Economic Development: Colombia*. New York: NBER.

1980. Discussions. *American Economic Review* 70, 330–5.

1981. Southern cone stabilization plans. Repr. in Velasco (1988).

1982. Latin America in the 1930's. Repr. in Velasco (1988).

1983. Some aspects of the 1982–83 Brazilian payments crisis. *Brookings Papers in Economic Activity* 1983(2), 515–2.

1984. Some economic lessons of the early 1980s. Repr. in Velasco (1988).

Bibliography

Palma, J.G. 2003. The three routes to financial crises. In *Rethinking Development Economics*, ed. H.-J. Chang. London: Anthem.

Ranis, G. 1989. Carlos Díaz-Alejandro: an appreciation. In *Debt, Stabilization and Development: Essays in Memory of Carlos Díaz-Alejandro*, ed. G. Calvo et al. Oxford: Blackwell.

Velasco, A., ed. 1988. *Trade, Development and the World Economy: Selected Essays of Carlos Díaz-Alejandro*. Oxford: Blackwell.

How to cite this article

Palma, José Gabriel. "Díaz-Alejandro, Carlos (1937–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000267>
doi:10.1057/9780230226203.0387

The New Palgrave Dictionary of Economics Online

Dickinson, Henry Douglas (1899 – 1969)

David Collard

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

CES production function; institutional rents; market socialism.; neoclassical growth theory

Article

Dickinson went from the King's School, Wimbledon, to Emmanuel College, Cambridge, where he took the Part II Tripos in both Economics and History. He carried out research at the London School of Economics under Cannan, then went to teaching posts at Leeds and Bristol, where he held the chair of economics from 1951 to 1964. Although his *Institutional Revenue* (1932) is of interest for generalizing the concept of institutional rents, he is deservedly known for a series of writings which attempted to reconcile choice and individual freedom with socialist planning, in the tradition of market socialism. Together with Taylor, Lange and Lerner he provided a rebuttal (based on actual markets) of von Mises's view that rational allocation under socialism was impossible. He saw ‘the beautiful systems of economic equilibrium’ not as ‘descriptions of society as it is but prophetic visions of a socialist economy of the future’ (1933, p. 247). During the 1930s his writings were well known to intellectuals of the Left, including Cole, Dalton, Durbin and Laski. The best-known of his works is the *Economics of Socialism* (1939). His technical prowess was later exhibited in a *Review of Economic Studies* article of 1954–5 in which he formulated a constant elasticity of substitution production function (CES) for the first time and anticipated some of the neoclassical growth results of Solow and Swan. ‘Dick’, as he was universally known, was a much loved, unworldly, eccentric figure with a keen sense of fun and a most astute mind.

Selected works

1932. *Institutional Revenue: A Study of the Influence of Social Institutions on the Distribution of Wealth*. London: Williams & Norgate.

1933. Price formation in a socialist community. *Economic Journal* 43, 237–50.

1939. *Economics of Socialism*. London: Oxford University Press.

1955. A note on dynamic economics. *Review of Economics Studies* 22(3), 169–79.

How to cite this article

Collard, David. "Dickinson, Henry Douglas (1899–1969)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000111> doi:10.1057/9780230226203.0388

The New Palgrave Dictionary of Economics Online

Dietzel, Heinrich (1857 – 1935)

H.C. Recktenwald

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Dietzel, H.; Eucken, W.; Freiburg School

Article

Born in Leipzig, Dietzel was appointed to a chair at the University of Dorpat in 1885 after studies in economics and law in Heidelberg, Göttingen and Berlin. In 1890 he accepted a chair in the philosophy faculty in Bonn. There he died in 1935.

Dietzel was a respected figure in circles of 19th-century German economists (such as Rau, von Thünen, von Hermann, von Mangoldt and Wagner) who were endeavouring to defend, pursue and modify classical methods and principles. He kept a sceptical distance from both the younger Historical School and the Austrian School, and was sharply opposed to popular Marxism. Nevertheless his excellent biography of Rodbertus and his writings on the early socialists are proof of his academic openness and liberal fairness. Enthusiastically though not successfully engaged in propagating free trade, Dietzel (in contrast to Manchester liberalism) was not dogmatic concerning the functions of the state in a concrete mixed economy.

His most important contribution to theory, the *Theoretische Sozialökonomie* (1895), unfortunately remained a torso. It is a pioneering analysis of the two main orders of an economy, namely, the individualistic system of competitive markets and the collective system of compulsion of the state. This concept of the two (centralized and decentralized) elementary forms replaced the unscientific notions of capitalism and socialism, with their ideological bias. It opened the way to the foundation of an order theory that his disciple in Bonn, Walter Eucken, and the Freiburg School further developed and later on applied in Germany.

Though Dietzel dealt with self-interest, methodological theory (1911) and value theory, he and his followers (as Smithians) did not attempt to unify Smith's three systems of ethics, economics and politics to an integrated order theory via reconstructing and developing his ‘obvious and simple system of natural liberty’. They also failed to produce an analysis of state and collective failures while they originally stressed the state's responsibility for ensuring sufficient market competition.

Nevertheless they made a number of contributions to the field and pointed to the right road to be taken in the future.

Selected works

1882. *Über das Verhältnis der Volkswirtschaftslehre zur Sozialwirtschaftslehre*. Berlin: Puttkammer und Mühlbrecht.

1886–8. *Karl Rodbertus: Darstellung seines Lebens und seiner Lehre*. 2 vols. Jena: G. Fischer.

1895. *Theoretische Sozialökonomie, I*. Leipzig: Winter.

1911. Selbstinteresse und Methodenstreit in der Wirtschaftstheorie. In *Handwörterbuch der Staatswissenschaften*, vol. 7. Jena: G. Fischer.

1921. *Vom Lehrwert der Wertlehre und vom Grundfehler der Marxschen Verteilungslehre*. Leipzig-Erlangen: Scholl.

1922. *Technischer Fortschritt und Freiheit der Wissenschaft*. Bonn–Leipzig: Schroeder.

Bibliography

Recktenwald, H.C. 1985. Über das Selbstverständnis der ökonomischen Wissenschaft. *Jahrbuch der Leibniz-Akademie der Wissenschaften und der Literatur*. Wiesbaden: Steiner.

Recktenwald, H.C. and Samuelson, P.A. 1986. Über Thünen's ‘Der isolierte Staat’. *Wirtschaft und Finanzen*, Darmstadt–Düsseldorf.

How to cite this article

Recktenwald, H. C. "Dietzel, Heinrich (1857–1935)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000114> doi:10.1057/9780230226203.0389

The New Palgrave Dictionary of Economics Online

difference-in-difference estimators

Alberto Abadie
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses difference-in-differences (DID) estimators, which are commonly applied in evaluation research. In particular, the discussion focuses on (*a*) motivation, definition and interpretation of DID estimators, (*b*) conditions under which DID estimators are valid, (*c*) data requirements to compute DID estimators, (*d*) representative applications of DID estimators in the empirical economics literature, (*e*) extensions of DID estimators, and (*f*) a simple indirect test to assess the validity of these estimators.

Keywords

difference-in-differences estimators; evaluation studies; fixed effects; minimum wages

Article

Motivation and definition

Difference-in-differences (DID) estimators are often used in empirical research in economics to evaluate the effects of public interventions and other treatments of interest in the absence of purely experimental data.

The usual goal of evaluation studies is to estimate the average effect of a treatment (for example, participation in a vocational training programme) on some outcome variable of interest (for example, earnings or employment). Often researchers concentrate on estimating the average effect of the treatment on the treated, that is, on those individuals exposed to the treatment or intervention (for example, the trainees). In the typical setting of an evaluation study, we observe an outcome variable, Y_i, for a sample of treated individuals and also for a sample of untreated individuals. The main challenge in evaluation research is to find an appropriate comparison group among the untreated individuals, in the sense that the distribution of the outcome variable for the untreated comparison group can be taken as an approximation to the counterfactual distribution that the outcome variable, Y_i, would have followed for the treated in the absence of the treatment.

Sometimes the sample of untreated individuals may not provide an appropriate comparison group, and therefore differences in the distribution of the outcome variable between treated and untreated reflect not only the effect of the treatment but also intrinsic differences between the two groups. To address this problem, the DID estimator uses the assumption that in the absence of the treatment the average difference in the outcome variable, Y_i, between treated and untreated would have stayed roughly constant. Then, the average difference in the outcome variable between treated and untreated before the treatment can be used to approximate the part of the difference in average outcomes after the treatment that is created by intrinsic differences between the two groups and not by the effect of the treatment.

Let $\bar{Y}_t^T$ and $\bar{Y}_t^C$ be the average outcomes in period t ($t = 1, 2$) in the treated and untreated samples, respectively. Period $t = 1$ takes place before the treatment and period $t = 2$ takes place after the treatment. The difference in average outcomes between treated and untreated after the treatment is $\bar{Y}_2^T - \bar{Y}_2^C$. The same difference for the pre-treatment period is $\bar{Y}_1^T - \bar{Y}_1^C$. Then, the DID estimator is defined as follows:

$$\hat{\alpha} = \left(\bar{Y}_2^T - \bar{Y}_2^C\right) - \left(\bar{Y}_1^T - \bar{Y}_1^C\right). \tag{1}$$

Figure 1 provides a graphical interpretation of the DID estimator. The solid lines represent the evolution in average outcomes for the treated and the untreated comparison group between the pre-treatment period ($t = 1$) and the post-treatment period ($t = 2$). The dashed line approximates the counterfactual evolution that the average outcome would have experienced for the treated in the absence of the treatment. This line is constructed under the DID assumption that, in the absence of the treatment, the difference in average outcomes between treated and untreated would have stayed roughly constant in the two periods. As reflected in Figure 1, an equivalent formulation of the DID assumption is that, in the absence of the treatment, average outcomes for treated and untreated would have followed a common trend. As a result, the untreated comparison group can be used to infer the counterfactual evolution of the average outcome for the treated in the absence of the treatment.

Figure 1

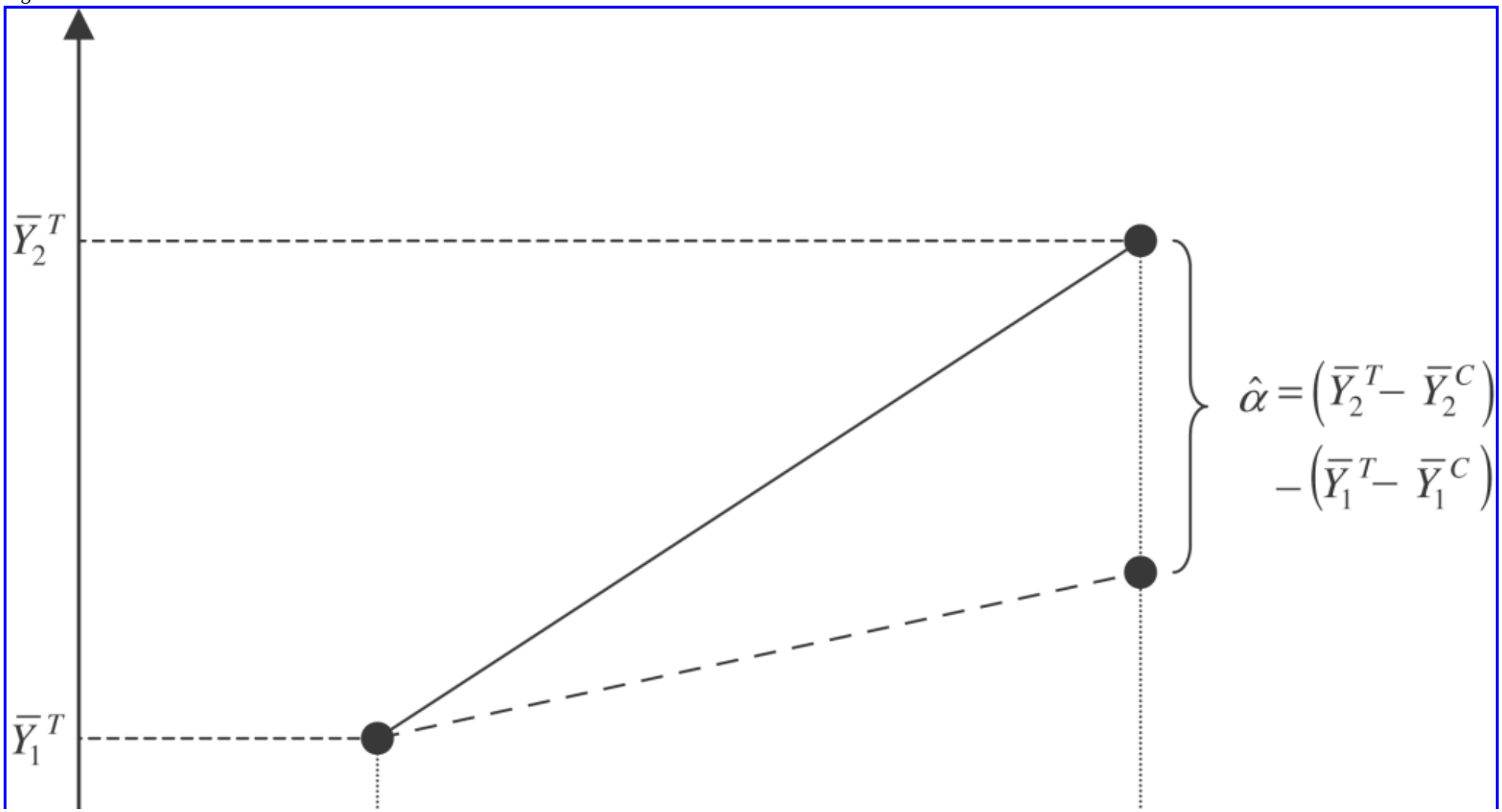

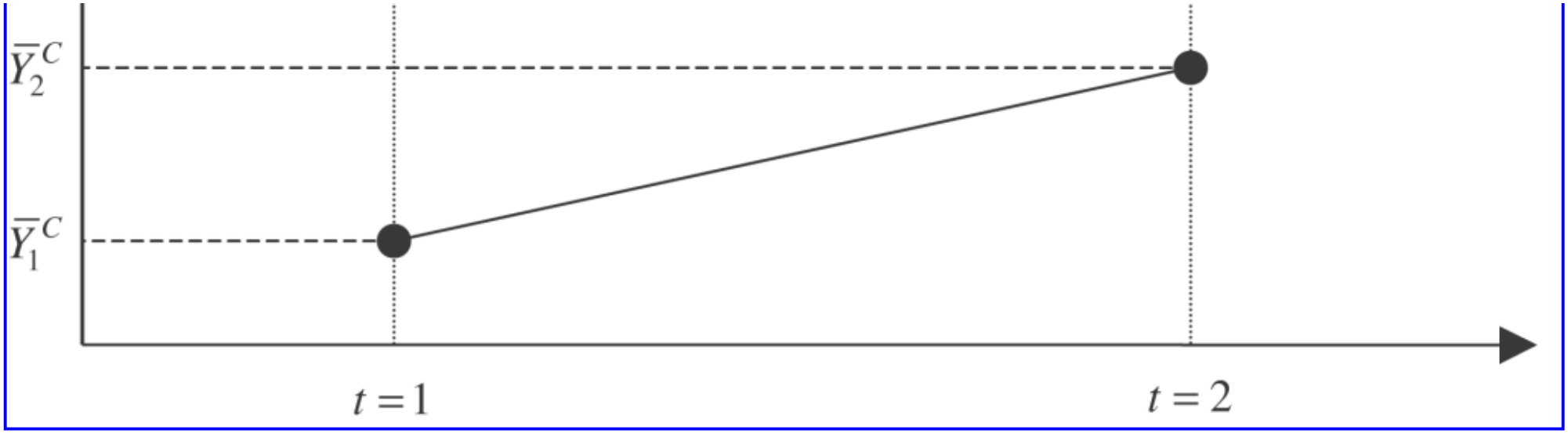

Difference in differences estimators have been applied to the study of a variety of issues in economics. Card and Krueger (1994) evaluate the employment effects of an increase in the minimum wage in New Jersey using a contiguous state (Pennsylvania), which did not increase the minimum wage, to approximate how employment would have evolved in New Jersey in the absence of the raise. Card (1990) applies DID estimators to evaluate the employment effects of the massive flow of Cuban immigrants to Miami during the 1980 Mariel boatlift. To estimate the effects of the boatlift, Card uses a group of four comparison cities to approximate how employment would have evolved in Miami in the absence of the 1980 immigration shock. Other applications of the DID estimator include studies of the effects of disability benefits on time out of work (Meyer, Viscusi and Durbin, 1995), the effect of anti-takeover laws on firms' leverage (Garvey and Hanka, 1999), and the effect of tax subsidies for health insurance on health insurance purchases (Gruber and Poterba, 1994).

The DID estimator has a simple regression representation. Let Y_{it} be the outcome of interest (for example, earnings) for individual i at time t, with $i = 1, \ldots, N$ and $t = 1, 2$. Let D_i be an indicator of membership to the treatment group, so $D_i = 1$ for the treated and $D_i = 0$ for the untreated. Finally, let $\Delta Y_i = Y_{i2} - Y_{i1}$ be the change in the outcome variable between the pre-treatment and the post-treatment period for individual i. The regression representation of the DID estimator is:

$$\Delta Y_i = \mu + \alpha D_i + u_i, \tag{2}$$

where u_i is a regression error, which is mean independent of D_i (that is, $E[u_i|D_i = 1] = E[u_i|D_i = 0]$). It can be easily seen that the ordinary least squares estimator of α in eq. (2) is numerical identical to the DID estimator, $\widehat{\alpha}$, in eq. (1). Regression standard errors along with the point estimate, $\widehat{\alpha}$, can be used to construct confidence intervals for α and perform statistical hypothesis tests. As reflected in eq. (2) and emphasized in Blundell and MaCurdy (1999), the DID estimator is a particular case of fixed effects estimators for panel data, with only two time periods and a fraction of the sample exposed to the treatment in the second time period.

Extensions

In some instances, the common trend assumption adopted for DID is not plausible because treated and untreated differ in the distribution of some variables, X_i, that are thought to affect the trend of the outcome variable. In this situation, treated and untreated may exhibit different trends in the average of the outcome variable between $t = 1$ and $t = 2$, even if the treatment does not have any impact on the outcome of interest. The regression formulation of the DID estimator is useful to compute a conditional version of the DID estimator that corrects for the effect of X_i on the trend of Y_i:

$$\Delta Y_i = \mu + \alpha D_i + X_i'\beta + u_i.$$

Abadie (2005) and Heckman, Ichimura and Todd (1997) develop semiparametric and nonparametric versions of the conditional DID estimator.

Panel data are not always necessary to apply the DID estimator. A simple inspection of eq. (1) indicates that $\hat{\alpha}$ can be estimated from repeated cross sections, using a cross-section at time $t = 2$ to estimate $\bar{Y}_2^T - \bar{Y}_2^C$ and a cross section at time $t = 1$ to estimate $\bar{Y}_1^T - \bar{Y}_1^C$. A regression formulation of the DID estimator is also available for repeated cross sections (see, for example, Meyer, 1995; Abadie, 2005). When the DID estimator is constructed using repeated cross sections, it is important to check whether there exist compositional changes in the sample between the two periods. Compositional changes may constitute a threat to the assumption that the difference in the average outcome between treated and untreated would have stayed constant in the absence of the treatment.

In general, the DID assumption cannot be tested directly with data from $t = 1$ and $t = 2$ only. However, if the common trend assumption extends to more than one pretreatment period for which data are available, pre-existing differences in the trends of the outcome variable between treated and untreated can be detected by applying the DID estimator to pretreatment data. This is done by constructing ΔY_i as the difference in the outcome variable for individual i between two pretreatment periods. Then, a test of the hypothesis $\alpha = 0$ in eq. (2) is a test of the common trend assumption. In addition, the DID assumption can sometimes be rejected when the dependent variable has bounded support (for example, when Y_i is a binary variable). If the dependent variable has bounded support the DID assumption may imply that, in the absence of the treatment, the average outcome for the treated would have lain outside the support of the dependent variable (see Athey and Imbens, 2006).

For a more detailed explanation of the theory behind DID estimators, see Abadie (2005), Angrist and Krueger (1999), Ashenfelter and Card (1985), Blundell and MaCurdy (1999), Heckman, Ichimura and Todd (1997), and Meyer (1995).

See Also

- fixed effects and random effects
- treatment effect

Bibliography

Abadie, A. 2005. Semiparametric difference-in-differences estimators. *Review of Economic Studies* 72, 1–19.

Angrist, J.D. and Krueger, A.B. 1999. Empirical strategies in labor economics. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Ashenfelter, O. and Card, D. 1985. Using the longitudinal structure of earnings to estimate the effects of training programs. *Review of Economics and Statistics* 67, 648–60.

Athey, S.C. and Imbens, G.W. 2006. Identification and inference in nonlinear difference-in-difference models. *Econometrica* 74, 431–98.

Blundell, R. and MaCurdy, T. 1999. Labor supply: a review of alternative approaches. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Card, D. 1990. The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review* 44, 245–57.

Card, D. and Krueger, A.B. 1994. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84, 772–93.

Garvey, G.T. and Hanka, G. 1999. Capital structure and corporate control: the effect of antitakeover statutes on firm leverage. *Journal of Finance* 54, 519–46.

Gruber, J. and Poterba, J. 1994. Tax incentives and the decision to purchase health insurance: evidence from the self-employed. *Quarterly Journal of Economics* 109, 701–33.

Heckman, J.J., Ichimura, H. and Todd, P.E. 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies* 64, 605–54.

Meyer, B.D. 1995. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics* 13, 151–61.

Meyer, B.D., Viscusi, W.K. and Durbin, D.L. 1995. Workers' compensation and injury duration: evidence from a natural experiment. *American Economic Review* 85, 322–40.

How to cite this article

Abadie, Alberto. "difference-in-difference estimators." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000245> doi:10.1057/9780230226203.0390

The New Palgrave Dictionary of Economics Online

diffusion of agricultural technology

David Zilberman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

diffusion of agricultural technology; diffusion of technology; discrete-choice estimation; Green Revolution; imitation model of technology diffusion; technical change; technology treadmill; threshold model of technology diffusion

Article

A high rate of technological change is a major feature of modern agriculture. New technologies are introduced gradually; diffusion is the process through which technologies spread throughout the farm sector over time. While adoption is the decision by an individual producer to use a new technology at a given moment, diffusion is the aggregate measure of adoption decisions. Early studies of diffusion were conducted by sociologists. Rogers (1962) measured technology usage as a fraction of farmers that had adopted a certain technology at a given point in time. Other studies measured diffusion by the fraction of land employed with the new technology. Rogers noticed that diffusion rates of hybrid corn in the United States fit very well as an S-shaped function of time:

$$S_t = \frac{K}{1 + exp^{-(a+bt)}},$$

where S_t is the level of diffusion at time t, K is the diffusion level at the limit and $K \leq 1$, a is a measure of initial diffusion, and b is a measure of the speed of diffusion. Rogers modelled diffusion as a process of imitation. In the early and late stages of diffusion, the level of diffusion is low because either the potential population of adopters or the population of users of the new technology to be imitated is small. During the middle period the diffusion rate takes off as there is a sufficient number of potential adopters, as well as a large population of established users to imitate. Rogers (1962) emphasized the role of

distance from urban centres in explaining diffusion, finding that villages closer to urban centres had higher coefficients of diffusion.

Griliches (1957) argued that diffusion is an economic phenomenon and showed, using the diffusion data for hybrid corn in Iowa, that the three parameters *K*, *a*, and *b* are affected by profit. Other studies also found that the rate of diffusion tends to increase with farm size and the education of the farmer.

However, as the review of Feder, Just and Zilberman (1985) suggests, the imitation model lacks a microeconomic foundation. An alternative model, the threshold model, suggests that the population of potential adopters is heterogeneous, and at every moment there is a critical variable that distinguishes between them. At every moment there is a threshold level of this variable that separates adopters from non-adopters.

Threshold models have three components: microeconomic behaviour, sources of heterogeneity, and a dynamic factor that drives the threshold level up or down. For example, adoption of mechanical innovation reflects the maximization of discounted net benefit. Farms vary in size, and at each moment there is a farm size threshold that distinguishes adopters from non-adopters. Over time, the cost of machinery may go down due to learning by doing, or the gain from adoption may go up because of learning by using, and that will reduce the adoption threshold. Empirical models, based on cross-sections of adopters, use discrete-choice estimation techniques to identify the key sources of heterogeneity. They found in many cases that size increases adoption of mechanical innovation, education explains adoption of more complex crops, and modern irrigation technologies that actually augment land quality are adopted earlier on lower-quality lands.

Much of the research has attempted to explain the diffusion of new ‘Green Revolution’ varieties in developing countries. In those cases, adoption was often partial (meaning farmers switched only a portion of their crops to the new technologies), and adoption rates were sometimes low, even given the significantly higher yields of Green-Revolution varieties. These facts emphasize the importance of risk considerations in explaining diffusion processes. Land allocation choices of risk-averse farmers were modelled as a portfolio, leading farmers to consider partial adoption of modern varieties because of their increased vulnerability to variable weather conditions. In addition to risk, wealth, human capital, and physical conditions, institutional forces have been identified as major determinants of diffusion rates. For example, renters are less likely to adopt new innovations than owners, especially when the rental contract is short. Lack of availability of credit is another deterrent to adoption. On the other hand, government policies, in the forms of output price subsidies and extension services that reduce the fixed costs of adoption, as well as technology and credit subsidies, can enhance the diffusion of modern agricultural technologies. For irrigation technologies, subsidies of water combined with restrictive trading regulations slow the diffusion of improved irrigation practices; water conservation can be enhanced by reducing constraints on water trading.

When demand for agricultural products is inelastic, the main beneficiary of the diffusion of more efficient technology is the consumer, while farmers are stuck on a ‘technology treadmill’. Early adopters also benefit from the introduction of the new technology, but followers, who make up the majority of the farm population, may adopt only to stay competitive, while sometimes the laggards may go out of business. When the demand for agricultural products is elastic, then the gain from adoption of modern technologies contributes to enhanced land values, but the individual farm operators may not gain significantly because of the technology treadmill effect.

See Also

- agricultural research
- diffusion of technology

Bibliography

Feder, G., Just, R.E. and Zilberman, D. 1985. Adoption of agricultural innovations in developing countries: a survey. *Economic Development and Cultural Change* 32, 255–98.

Griliches, Z. 1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25, 501–22.

Rogers, E.M. 1962. *Diffusion of Innovation*. New York: Free Press.

How to cite this article

Zilberman, David. "diffusion of agricultural technology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000192> doi:10.1057/9780230226203.0391

The New Palgrave Dictionary of Economics Online

diffusion of technology

Wolfgang keller

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The diffusion of technology has a major impact on per-capita income. Moreover, international convergence turns on whether technology diffusion is local or global. This article characterizes the creation of technological knowledge and discusses the primary determinants of diffusion. It is shown that even today technology diffusion is to an important degree local, allowing for many technological knowledge levels in the world to coexist. This article focuses on the data and empirical methods employed in the estimation of diffusion patterns.

Keywords

asymmetric information; control functions; diffusion of technology; economic growth; human capital; instrumental variables; intermediate goods; knowledge spillovers; monopoly; neoclassical growth theory; patents; product market competition; research and development; tacit knowledge; technical change; technology spillovers; total factor productivity

Article

The technology of a firm or country determines the efficiency with which inputs are mapped into outputs. Technological change may result in the ability to produce entirely new products, or it may allow an existing product to be produced with fewer inputs. This process has long been viewed as central to economic growth. The question of whether or not there is convergence across firms and countries raises issues related not only to the process of technical change but also to the diffusion of technology. Beginning in the late 1950s, economists have formalized their thinking as to how such technological knowledge diffuses from one economic entity to another. The early efforts were primarily directed to understanding firms' technology adoption decisions that often yield an S-shaped diffusion pattern over time. Since the 1990s, a vibrant literature has emerged in which the issues addressed are considerably broader, and where much more emphasis is placed on seeking high-quality empirical evidence.

A firm's technology and its productivity are closely related, and the two are identical if technology is identified with total factor productivity, an approach frequently adopted since the 1950s. The development of models of endogenous technical change in the early 1990s represented a step forward in that the R&D resources devoted to innovation were separated from the new technological knowledge itself. For example, consider the technology production function

$$\dot{N} = \eta N^{\lambda} H_N, \tag{1}$$

where η and λ are parameters, η, $\lambda > 0$. The term H_N denotes the skilled-labour resources devoted the R&D, which according to eq. (1) lead to a flow of new technological knowledge of $\dot{N}$. A higher level of R&D produces a higher level of technology, *N*, and that in turn can be shown to result in higher productivity.

According to eq. (1), a higher stock of existing technological knowledge facilitates innovation. This stock of technological knowledge will rarely be entirely self-produced, so that (1) typically involves the diffusion of technology – diffusion between different persons, firms or countries. Technology is sometimes purchased or licensed in a market transaction, but, due to asymmetric information and other problems in the market for technology, non-market transactions in the form of technological externalities, called knowledge spillovers, are much more important. What are the nature and the size of these knowledge spillovers? Since technological knowledge is non-rival, such externalities can in principle benefit many economic agents.

A useful benchmark is the complete diffusion of technology, which describes the case where technological knowledge created anywhere in the world is available worldwide immediately. This could underlie the assumptions of common-to-all and free technological knowledge of neoclassical growth theory. Clearly, this is not true in reality, where the diffusion of technology is gradual and uneven. Why? First of all, acquiring technology involves making complementary investments, and the equilibrium choice for such investments often implies that not all technology diffuses. For instance, in Keller's (1996) model, international trade enables domestic producers to raise productivity by importing specialized foreign intermediate goods. Since these goods embody foreign R&D investments, this means the diffusion of technology from one country to another. For this imported technological knowledge to trigger domestic innovation, however, additional investments are necessary. According to Keller (1996), these investments mean additional training of workers so that they have the skills to manufacture products according to new blueprints. In addition, domestic innovators may have to invest resources in reverse engineering the foreign intermediate goods in order to fully comprehend the underlying foreign technological knowledge.

Second, another major determinant of the firm's decision to acquire the existing technology and innovate is the degree of product market competition. For example, in early Schumpeterian endogenous growth models, a higher degree of product market competition leads to lower monopoly profits and thus to a lower rate of innovation. More recent work by Aghion et al. (2001), for example, shows that, if

technological laggards must first catch up with the leading-edge technology before battling for technological leadership in the future, the overall effect of more product market competition may be positive. The reason for this is that, even though more competition means lower monopoly profits, technological leaders now also have an incentive to innovate to avoid competition with technological laggards, and, if the latter effect is strong enough, product competition has a positive effect on technology diffusion and growth.

Third, there is no complete diffusion because it is simply not in the interest of the original creator of the technology, since his market for the technology would shrink if there were additional suppliers. In some cases, innovators obtain a patent that provides government-sanctioned protection of economic interests for a limited period of time in exchange for release of the technological information. Another strategy on the part of the original innovator is to use a varying amount of resources to keep the technological knowledge secret. At the same time, studies show that it often is no more than two years until new technology becomes publicly available.

Another, probably the most important, reason why knowledge spillovers are limited is that only the broad outlines of technology are codified – the remainder is the ‘tacit’ part of the knowledge. A person who is engaged in a problem-solving activity can often not fully define (and hence prescribe) what exactly he or she is doing. Along these lines, technology is only partially codified because it is impossible or at least very costly to fully codify it. For technology diffusion to occur completely, it may be necessary that the person who learns about the new technology can observe another person in the process of applying the technology. Even if this can be dispensed with, person-to-person contacts will generally be beneficial to the diffusion of technology.

Research has now turned to the essential task of assessing the importance of these processes empirically. As an intangible, technology is intrinsically difficult to measure, and economic data is hard to come by. This is even more the case for the non-market effects caused by technological knowledge. The main approach for quantifying technical change has been to study the relationship between R&D investments and productivity (Griliches, 1979). For example, Keller (2002a) estimates

$$tfp_{it} = \beta s_{it} + X'\gamma + \varepsilon_{it},\ i = 1, \ldots, I,\ \text{and}\ t = 1, \ldots, T, \tag{2}$$

where tfp_{it} is log total factor productivity in industry i at time t, s_{it} are industry i's cumulative R&D investments (in logs) in period t, X is a vector of other observed determinants of productivity, and the error ε_{it} picks up unobserved effects. The parameter β, estimated in Keller (2002a) at β=0.15, measures how R&D investments translate into higher productivity, thereby implicitly capturing the rate of technical change.

This approach is attractive since R&D spending is the main cause of technical change, and data on R&D expenditures is relatively easy to collect and compare across units (firms, industries and countries). A drawback is that measuring technical change this way requires an estimate of β. This can be complicated if productivity is badly measured, R&D is endogenous, or unobserved determinants on productivity are important, as in practice is often the case. Applications of instrumental-variable and

control-function approaches have shown much promise in addressing the major estimation concerns (see Gong and Keller, 2003). Patents are an alternative measure of technology, with the advantage that patent data is available for a broader set of countries and a longer time horizon than is data on R&D (Jaffe and Trajtenberg, 2002). While patent counts are an imperfect measure of technology because the distribution of patent values is extremely skewed, recent work using citations-weighted patent data has addressed this point since citations of a particular patent are a plausible indicator of its value. At the same time, patents cannot capture more than the codified part of technological knowledge, apart from the fact that across industries and firms the prevalence of patenting varies strongly for reasons that are difficult to fully ascertain.

Technology spillovers, as the major form of technology diffusion, are mainly analysed by extending eq. (2) above to estimate as well the effects of R&D investments conducted elsewhere. For example, in addition to the effects of own-industry R&D, Keller (2002a) estimates the effects of R&D in other domestic industries (s_{it}^{do}), as well as those of R&D in the same and other foreign industries (s_{it}^{f} and s_{it}^{fo}, respectively):

$$tfp_{it} = \beta_1 s_{it} + \beta_2 s_{it}^{do} + \beta_3 s_{it}^{f} + \beta_4 s_{it}^{fo} + \tilde{X}'\gamma + \varepsilon_{it}. \quad (3)$$

In this framework, the estimates of β_1, β_2, β_3, and β_4 determine the relative strength of intra- and inter-industry, and of domestic and international technology diffusion. For his sample of eight large Organisation for Economic Co-operation and Development (OECD) countries, Keller (2002a) finds that intra-industry effects dominate inter-industry spillovers, and that about 25 per cent of the total effect is due to international technology diffusion.

Other interesting approaches have employed multi-country extensions of recent models of endogenous technical change that include international technology diffusion (Eaton and Kortum, 1999). Because here the economic environment is fully specified, it is straightforward to simulate a model and perform interesting policy experiments. At the same time, typically there is little data on technology diffusion employed in the econometric estimation of these models. Consequently, the model's structure has a great influence on the results, while the implications for the diffusion of technology are not clear.

One major finding has been that the diffusion of technology is geographically localized, both domestically and internationally. For example, Keller (2002b) studies international technology diffusion between the G-5 countries (the United States, Japan, Germany, France, and England) and nine smaller OECD countries by estimating

$$tfp_{it} = \beta\left[s_{it} + \sum_{j \in G5} \exp(-\delta Dist_{ij}) s_{jt}\right] + X'\gamma + \varepsilon_{it}, i = 1, \ldots, I, \text{ and } t = 1, \ldots, T.$$

(4)

Here, $Dist_{ij}$ is the geographic distance between country *i* and G-5 country *j*. The parameter δ determines the extent of geographic localization: the higher is δ, the stronger is the degree of the localization of technological knowledge, while if $\delta=0$, international technology diffusion is complete in the sense that geography has no impact whatsoever. The geographic reach of technology spillovers is a critical determinant of the cross-country income distribution, since global spillovers favour income convergence while local spillovers lead to income divergence. Keller's (2002b) results for the years 1970 to 1995 strongly reject the null hypothesis of complete diffusion. Instead, he estimates that with every additional 1,200 kilometres there is a 50 per cent drop in technology diffusion. The results imply that the benefits of being located next to major technology producers are substantial, highlighting the danger for isolated areas of being left behind.

While distance still shapes technology diffusion in a major way, there is also evidence that between 1970 and 1995 geography's grip on technology diffusion has weakened. Keller (2002b) estimates that the size of the δ parameter in eq. (4) fell substantially from the late 1970s to the 1990s, consistent with the idea that innovations in information and communication technologies have led to a major improvement in technology diffusion.

Such improvements in countries' abilities to draw on international innovations also imply that increasingly the ultimate sources of domestic productivity growth lie abroad. This is especially true for medium-sized and small countries, where the contribution of foreign technology to domestic productivity growth often exceeds 90 per cent. At the same time, because successful technology diffusion requires complementary investments in terms of adaptive R&D and/or human capital, domestic activities have a significant impact on the ease of technology diffusion.

See Also

- transfer of technology

Bibliography

Aghion, P., Harris, J., Howitt, P. and Vickers, J. 2001. Competition, imitation and growth with step-by-step innovation. *Review of Economic Studies* 68, 467–92.

Eaton, J. and Kortum, S. 1999. International technology diffusion: theory and measurement. *International Economic Review* 40, 537–70.

Gong, G. and Keller, W. 2003. Convergence and polarization in global income levels: a review of recent results on the role of international technology diffusion. *Research Policy* 32, 1055–79.

Griliches, Z. 1979. Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics* 10, 92–116.

Jaffe, A. and Trajtenberg, M. 2002. *Patents, Citations, and Innovations: A Window in the Knowledge Economy*. Cambridge, MA: MIT Press.

Keller, W. 1996. Absorptive capacity: on the creation and acquisition of technology in development. *Journal of Development Economics* 49, 199–227.

Keller, W. 2002a. Trade and the transmission of technology. *Journal of Economic Growth* 7, 5–24.

Keller, W. 2002b. Geographic localization of international technology diffusion. *American Economic Review* 92, 120–42.

How to cite this article

keller, Wolfgang. "diffusion of technology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000122> doi:10.1057/9780230226203.0392

The New Palgrave Dictionary of Economics Online

directly unproductive profit-seeking (DUP) activities

Jagdish N. Bhagwati
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chicago School; directly unproductive profit-seeking (DUP) activities; endogeneous tariffs; immiserizing growth; lobbying; optimal tariffs; predation; production subsidies; public choice; regulation; rent seeking; revenue seeking; shadow pricing; smuggling; tariff seeking; tariffs; transfer problem; voluntary export restrictions

Article

Directly unproductive profit-seeking (DUP) activities are defined (Bhagwati, 1982a) as ways of making a profit (that is, income) by undertaking activities which are directly (that is, immediately, in their primary impact) unproductive, in the sense that they produce pecuniary returns but do not produce goods or services that enter a conventional utility function or inputs into such goods and services.

Typical examples of such DUP (pronounced appropriately as ‘dupe’) activities are (i) tariff-seeking lobbying which is aimed at earning pecuniary income by changing the tariff and therefore factor incomes; (ii) revenue-seeking lobbying which seeks to divert government revenues towards oneself as recipient; (iii) monopoly seeking lobbying whose objective is to create an artificial monopoly that generates rents; and (iv) tariff-evasion or smuggling which de facto reduces or eliminates the tariff (or quota) and generates returns by exploiting thereby the price differential between the tariff-inclusive legal and the tariff-free illegal imports.

While these are evidently profitable activities, their *output* is zero. Hence, they are wasteful in their primary impact, recalling Pareto's distinction between production and predation: they use real resources to produce profits but no output.

DUP activities of one kind or another have been analysed by several economic theorists, among them (i) the public-choice school's leading practitioners, their major work having been brought together in Buchanan, Tullock and Tollison (1980), (ii) Lindbeck (1976) who has worked on ‘endogenous politicians’, and (iii) the Chicago ‘regulation’ school, led by Stigler, Peltzman, Posner and also Becker (1983).

However, a central theoretical breakthrough has come from the work of trade theorists who have systematically incorporated the analysis of DUP activities in the main corpus of general equilibrium theory.

The early papers that defined this general-equilibrium-theoretic approach, and which were set in the context of the theory of trade and welfare, were: Bhagwati and Hansen (1973) which analysed the question of illegal trade (that is, tariff-evasion), Krueger (1974) which analysed the question of rent-seeking for rents associated with import quotas specifically and quotas more generally, and (iii) Bhagwati and Srinivasan (1980) who analysed the phenomenon of revenue-seeking, the ‘price’ counterpart of Krueger's rent-seeking, where a tariff resulted in revenues which were then sought by lobbies.

The synthesis and generalization of these and other apparently unrelated contributions, showing that they all related to diversion of resources to zero-output activities, was provided in Bhagwati (1982a) where they were called DUP activities. The following significant aspects of the theoretical analysis of DUP activities are noteworthy.

First, they are generally related to policy interventions (but they need not be: plunder, for instance, pre-dates the organization of governments). In so far as policy interventions induce DUP activities, they are analytically divided into two appropriate categories (Bhagwati and Srinivasan, 1982):

Category I: Policy-triggered DUP activities. One class consists of *lobbying* activities. Examples include: rent-seeking analysis of the cost of protection *via* import licences (Krueger, 1974); revenue-seeking analysis of the cost of tariffs (Bhagwati and Srinivasan, 1980), of shadow prices in cost–benefit analysis (Foster, 1981), of price versus quantity interventions (Bhagwati and Srinivasan, 1982), of non-economic objectives (Anam, 1982), of rank-ordering of alternative distorting policies such as tariffs, production and consumption taxes (Bhagwati, Brecher and Srinivasan, 1984), of the optimal tariff (Dinopoulos, 1984), of the transfer problem (Bhagwati, Brecher and Hatta, 1985), and of voluntary export restrictions relative to import tariffs (Brecher and Bhagwati, 1987).

Another class consists of *policy-evading* activities. Examples include: analysis of smuggling (Bhagwati and Hansen, 1973), its implication for optimal tariffs (Johnson, 1974 and Bhagwati and Srinivasan, 1973), and alternative modelling by Kemp (1976), Sheikh (1974), Pitt (1981) and Martin and Panagariya (1984).

Category II: Policy-influencing DUP activities. The other generic class of DUP activities is not triggered by policies in place but is rather aimed at influencing the formulation of the policy itself. The most prominent DUP-theoretic contributions in this area relate to the analysis of tariff-seeking. Although Brock and Magee (1978; 1980) pioneered here, the general equilibrium analyses of endogeneous tariffs began with Findlay and Wellisz (1982) and Feenstra and Bhagwati (1982), the two sets of authors modelling the government and the lobbying activities in contrasting ways. Notable among the later contributions are Mayer (1984), who extends the analysis formally to include factor income-distribution and therewith voting behaviour, and Wellisz and Wilson (1984). Magee (1984) has an excellent review of many of these contributions. The implication of endogenizing the tariff for conventional measurement of the cost of protection has been analysed in Bhagwati (1980) and Tullock (1981).

The *choice* between alternative policy instruments when modelling the response of lobbies and governments to import competition has also been extensively analysed. The issue was raised by

Bhagwati (1982b) and analysed further by Dinopoulos (1983) and Sapir (1983) in terms of how different agents (for example, ‘capitalists’ and ‘labour’) would profit from different policy responses such as increased immigration of cheap labour and tariffs when import competition intensified. It has subsequently been explored more fully by Rodrik (1986), who compares tariffs with production subsidies.

Second, Bhagwati (1982a) has noted, generalizing a result in Bhagwati and Srinivasan (1980), that DUP activities, while defined to be those that waste resources in their direct impact, cannot be taken as *ultimately* wasteful, that is, immiserizing, since they may be triggered by a suboptimal policy intervention. For, in that event, throwing away or wasting resources may be beneficial. The shadow price of a productive factor in such ‘highly distorted’ economies may be negative. This is the obverse of the possibility of immiserizing growth (Bhagwati, 1980). Thus, Buchanan (1980), who has addressed the issue of DUP activities and *defined* them as activities that (ultimately) cause waste, has been corrected in Bhagwati (1983): the definition of DUP activities cannot properly exclude the possibility that DUP activities are ultimately beneficial rather than wasteful. This central distinction between the direct and the ultimate welfare impacts of DUP activities is now universally accepted. DUP activities are therefore defined now, as in Bhagwati (1982b) and subsequent contributions, as wasteful only in the direct sense.

Third, Bhagwati, Brecher and Srinivasan (1984) have raised yet another fundamental issue concerning DUP activities. Thus, where DUP activities belong to Category II distinguished above, full endogeneity of policy can follow. If so, the conventional rank-ordering of policies is no longer possible. We have the *determinacy paradox:* policy is chosen in the solution to the full ‘political-economy’, DUP-theoretic solution and cannot be varied at will. These authors have therefore suggested that, where full endogeneity obtains, the appropriate way to theorize about policy is to take variations around the observed DUP-theoretic equilibrium. Thus, traditional economic parameters such as factor supply could be varied; similarly now the DUP-activity parameters such as, say, the cost of lobbying could be varied. The impact on actual welfare resulting from such variations can then be a proper focus of analysis, implying a wholly different way of looking at policy questions from that which economists have employed to date.

Finally, DUP activities are related to Krueger's (1974) important category of rent-seeking activities. The latter are a subset of the former, in so far as they relate to lobbying for quota-determined scarcity rents and are therefore part of DUP activities of Category II distinguished above (Bhagwati, 1983).

See Also

- bribery
- rent seeking

Bibliography

Anam, M. 1982. Distortion-triggered lobbying and welfare: a contribution to the theory of directly-unproductive profit-seeking activities. *Journal of International Economics* 13(August), 15–32.

Becker, G.S. 1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 93, 371–400.

Bhagwati, J. 1980. Lobbying and welfare. *Journal of Public Economics* 14, 355–63.

Bhagwati, J. 1982a. Directly-unproductive profit-seeking (DUP) activities. *Journal of Political Economy* 90, 988–1002.

Bhagwati, J. 1982b. Shifting comparative advantage, protectionist demands, and policy response. In *Import Competition and Response*, ed. J. Bhagwati. Chicago: Chicago University Press.

Bhagwati, J. 1983. DUP activities and rent seeking. *Kyklos* 36, 634–7.

Bhagwati, J., Brecher, R. and Hatta, T. 1985. The generalized theory of transfers and welfare: exogenous (policy-imposed) and endogenous (transfer-induced) distortions. *Quarterly Journal of Economics* 100, 697–714.

Bhagwati, J., Brecher, R. and Srinivasan, T.N. 1984. DUP activities and economic theory. In *Neoclassical Political Economy: The Analysis of Rent-seeking and DUP Activities*, ed. D. Colander. Cambridge, MA: Ballinger & Co.

Bhagwati, J. and Hansen, B. 1973. Theoretical analysis of smuggling. *Quarterly Journal of Economics* 87, 172–87.

Bhagwati, J. and Srinivasan, T.N. 1973. Smuggling and trade policy. *Journal of Public Economics* 2, 377–89.

Bhagwati, J. and Srinivasan, T.N. 1980. Revenue-seeking: a generalization of the theory of tariffs. *Journal of Political Economy* 88, 1069–87.

Bhagwati, J. and Srinivasan, T.N. 1982. The welfare consequences of directly-unproductive profit-seeking (DUP) lobbying activities: price *versus* quantity distortions. *Journal of International Economics* 13, 33–44.

Brecher, R. and Bhagwati, J. 1987. Voluntary export restrictions and import restrictions: a welfare-theoretic comparison. In *Essays in Honour of W.M. Corden*, ed. H. Kierzkowski. Oxford: Basil Blackwell.

Brock, W. and Magee, S. 1978. The economics of special interest politics: the case of the tariff. *American Economic Review* 68, 246–50.

Brock, W. and Magee, S. 1980. Tariff formation in a democracy. In *Current Issues in International Commercial Policy and Diplomacy*, ed. J. Black and B. Hindley. New York: Macmillan.

Buchanan, J. 1980. Rent seeking and profit seeking. In *Towards a General Theory of the Rent-seeking Society*, ed. J. Buchanan, G. Tullock and R. Tollison. College Station: Texas A&M University Press.

Buchanan, J., Tullock, G. and Tollison, R., eds. 1980. *Towards a General Theory of the Rent-seeking Society*. College Station: Texas A&M University Press.

Dinopoulos, E. 1983. Import competition, international factor mobility and lobbying responses: the Schumpeterian industry cases. *Journal of International Economics* 14, 395–410.

Dinopoulos, E. 1984. The optimal tariff with revenue-seeking: a contribution to the theory of DUP activities. In *The Neoclassical Political Economy: The Analysis of Rent-seeking and DUP Activities*, ed. D. Colander. Cambridge, MA: Ballinger & Co.

Feenstra, R. and Bhagwati, J. 1982. Tariff seeking and the efficient tariff. In *Import Competition and Response*, ed. J. Bhagwati. Chicago: Chicago University Press.

Findlay, R. and Wellisz, S. 1982. Endogenous tariffs, the political economy of trade restrictions, and welfare. In *Import Competition and Response*, ed. J. Bhagwati. Chicago: Chicago University Press.

Foster, E. 1981. The treatment of rents in cost-benefit analysis. *American Economic Review* 71, 171–8.

Johnson, H.G. 1974. Notes on the economic theory of smuggling. In *Illegal Transactions in International Trade*, ed. J. Bhagwati. Series in International Economics. Amsterdam: North-Holland.

Kemp, M. 1976. Smuggling and optimal commercial policy. *Journal of Public Economics* 5, 381–4.

Krueger, A. 1974. The political economy of the rent-seeking society. *American Economic Review* 64, 291–303.

Lindbeck, A. 1976. Stabilization policies in open economies with endogenous politicians. Richard Ely Lecture. *American Economic Review* 66, 1–19.

Magee, S. 1984. Endogenous tariff theory: a survey. In *Neoclassical Political Economy: The Analysis of Rent-seeking and DUP Activities*, ed. D. Colander. Cambridge, MA: Ballinger & Co.

Martin, L. and Panagariya, A. 1984. Smuggling, trade, and price disparity: a crime-theoretic approach. *Journal of International Economics* 17, 201–18.

Mayer, W. 1984. Endogenous tariff formation. *American Economic Review* 74, 970–85.

Pitt, M. 1981. Smuggling and price disparity. *Journal of International Economics* 11, 447–58.

Rodrik, D. 1986. Tariffs, subsidies and welfare with endogenous policy. *Journal of International Economics* 21, 285–99.

Sapir, A. 1983. Foreign competition, immigration and structural adjustment. *Journal of International Economics* 14, 381–94.

Sheikh, M. 1974. Smuggling, production and welfare. *Journal of International Economics* 4, 355–64.

Tullock, G. 1981. Lobbying and welfare: a comment. *Journal of Public Economics* 16, 391–4.

Wellisz, S. and Wilson, J.D. 1984. Public sector inefficiency, a general equilibrium analysis. Discussion Paper No. 254, International Economics Research Center, Columbia University.

How to cite this article

Bhagwati, Jagdish N. "directly unproductive profit-seeking (DUP) activities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000130> doi:10.1057/9780230226203.0393

The New Palgrave Dictionary of Economics Online

Director, Aaron (1901 – 2004)

Steven G. Medema
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

antitrust; cartels; Director, A.; law and economics; oligopoly; Posner, R.; predatory pricing; resale price maintenance; tie-in sales

Article

Aaron Director's enduring contribution to economics came via his role in the development of the Chicago law and economics tradition. Director was born in Charterisk (in present-day Ukraine) in 1901 and emigrated to the United States with his family in 1913. He received his undergraduate degree from Yale University and his graduate training at the University of Chicago. Although he came to Chicago in 1927 to work with Paul Douglas on labour economics, it was Frank Knight and Jacob Viner who, via their price theory courses, had the greatest influence on him. Director remained at Chicago as a graduate student and part-time instructor until 1934. The 1930s were a heady period at Chicago, where the student body included George Stigler, Paul Samuelson (who credits Director's teaching with stimulating his interest in economics), and Milton Friedman – each of whom helped to reshape economic thinking in the middle third of the 20th century – as well as Rose Director (Aaron's sister and, eventually, Rose Friedman). Aaron Director was very much part of this milieu. He left the University of Chicago for the US Treasury Department in 1934 and, save for an aborted attempt to complete a dissertation on the history of the Bank of England, remained in Washington, DC, until 1946, when he returned to the University of Chicago to take up a position in the Law School, where he remained until his retirement in 1966.

Director's appointment in the Law School was a result of the efforts of Henry Simons, the first economist on the law faculty at Chicago, and Friedrich Hayek, whose *Road to Serfdom* was published in the United States largely because of Director's intervention with the University of Chicago Press. The plan, as laid out by Simons, was for Director to head up the ‘Free Market Study’, a Volker Fund-financed project, housed in the Law School and dedicated to undertaking ‘a study of a suitable legal and institutional framework of an effective competitive system’ (Coase, 1998, p. 246). However, Simons committed suicide in the summer of 1946, and Director was asked to take on Simons's basic Law School

price theory course, ‘Economic Analysis of Public Policy’. This provided Director with an initial forum for bringing the perspective he had learned from Knight and Viner into the Law School's teaching programme.

The transition from having an economist on the Law School faculty to the establishment of a law and economics tradition at Chicago began not long after this, when Edward Levi invited Director to collaborate in the teaching of the antitrust course. Levi would teach a traditional antitrust course for four days each week; Director would then come in on the fifth day and, using the tools of price theory, show that the traditional legal approach could not stand up to the rigours of economic analysis. The basic pattern was very simple: Director would ask whether the practice in question was, in general, consistent with monopolistic profit maximization. The answer was often negative, which meant that there had to be some sort of legitimate rationale for the supposedly anti-competitive practice in question. What Director's price theory showed was that the ‘simple and obvious’ answers were often wrong-headedly simplistic. This process had a profound impact on students and colleagues alike. Director's antitrust students – a group that included Robert H. Bork, Ward Bowman, Kenneth Dam, Edmund Kitch, Wesley J. Liebeler, John S. McGee, Henry Manne, and Bernard H. Siegan – have often spoken of the ‘conversion’ they experienced in this class, and even Levi himself became a partial convert (see Kitch, 1983; Director and Levi, 1951). What was perhaps Director's most significant contribution on the missionary front came after his retirement, when he and Richard Posner spent time together at Stanford in 1968 – Posner's first year on the Stanford Law School faculty. It was Director who taught Posner to think like a Chicago economist, introduced him to Stigler and Ronald Coase, and in this and other ways was instrumental in Posner's move to the Chicago Law School after only one year on the Stanford faculty. The rest, as they say, is history.

Although Director's published output was slight, his influence extended well beyond the classroom. His insights made their way into the antitrust literature – and, eventually, antitrust policy – through the writings of students and colleagues, as Sam Peltzman (2005) has detailed. Director's primary legacies are in the analysis of predatory pricing (via McGee, 1958), resale price maintenance (via Telser, 1960), and tie-in sales (see Director and Levi, 1951; Bowman, 1957; Burnstein, 1960), but his influence was also prominent in Stigler's view of oligopoly and antitrust policy, Posner's (1969) perspective on oligopoly and cartels, and Robert Bork's influential articles on antitrust (for example, Bork and Bowman, 1965; Bork, 1967). These contributions coalesced in a distinctive Chicago approach to antitrust analysis, an approach that Herbert Hovenkamp (1986, p. 1020) says ‘has done more for antitrust policy than any other coherent economic theory since the New Deal’, and whose influence is inescapable.

Director's impact at the Law School went far beyond antitrust: He was also the prime mover in the early professionalization of law and economics. Director formally established the nation's first law and economics programme, which maintained visiting fellowships for law and economics scholars, and, in 1958, founded the *Journal of Law and Economics*. Within a few decades, Director's efforts at Chicago had been replicated in a set of thriving and well-funded law and economics programmes at major law schools around the country. One would be hard pressed to name an individual in our discipline who has had as much influence as Director without a much more extensive bibliography.

See Also

- antitrust enforcement
- Chicago School
- law, economic analysis of

Selected works

1933. *The Economics of Technocracy*. Public Policy Pamphlet No. 2. Chicago: University of Chicago Press.

1940. Does inflation change the economic effects of war? *American Economic Review* 35, 351–61.

1964. Parity and the economic marketplace. *Journal of Law and Economics* 7, 1–10.

1931. (With P. Douglas.) *The Problem of Unemployment*. New York: Macmillan.

1951. (With E. H. Levi.) Law and the future: trade regulation. *Northwestern Law Review* 51, 281–96.

Bibliography

Bork, R.H. 1967. Antitrust and monopoly: the goals of antitrust policy. *American Economic Review* 57, 242–53.

Bork, R.H. and Bowman, W.S., Jr. 1965. The crisis in antitrust. *Columbia Law Review* 65, 363–76.

Bowman, W.S., Jr. 1957. Tying arrangements and the leverage problem. *Yale Law Journal* 67, 19–36.

Burnstein, M.L. 1960. The economics of tie-in sales. *Review of Economics and Statistics* 42, 68–73.

Coase, R.H. 1998. Director, Aaron. In *The New Palgrave Dictionary of Economics and the Law*. London: Macmillan.

Hovenkamp, H. 1986. Chicago and its alternatives. *Duke Law Journal* 1986, 1014–29.

Kitch, E.W. 1983. The fire of truth: a remembrance of Law and Economics at Chicago, 1932–1970. *Journal of Law and Economics* 26, 163–233.

McGee, J.S. 1958. Predatory price cutting: the standard oil (NJ) case. *Journal of Law and Economics* 1, 137–69.

Peltzman, S. 2005. Aaron Director's influence on antitrust policy. *Journal of Law and Economics* 48, 313–30.

Posner, R.A. 1969. Oligopoly and the antitrust laws: a suggested approach. *Stanford Law Review* 21, 1562–1606.

Stigler, S.M. 2005. Aaron Director remembered. *Journal of Law and Economics* 48, 307–11.

Telser, L.G. 1960. Why should manufacturers want fair trade? *Journal of Law and Economics* 3, 86–105.

How to cite this article

Medema, Steven G. "Director, Aaron (1901–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000262> doi:10.1057/9780230226203.0394

The New Palgrave Dictionary of Economics Online

dispute resolution

Amy Farmer and Paul Pecorino

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The high cost of disputes creates an incentive for parties to disputes to settle. In civil litigation and arbitration, settlement failure may arise from asymmetric information or optimism. Devices to induce settlement include voluntary disclosure and mandatory discovery. The effects of these are considered, as are the English rule (whereby the loser at trial pays the reasonable legal costs of the winner), the use of contingency fees, and the operation of conventional arbitration and final offer arbitration. Researchers continue to propose new arbitration mechanisms in the hope of improving the dispute resolution process.

Keywords

asymmetric information; bargaining; signalling; screening; arbitration; final offer arbitration; contingency fees; English rule; fee-shifting; optimism; self-serving bias

Article

Disputes may arise in a variety of settings, including labour negotiations, civil disputes and family conflict. If individuals fail to reach an agreement, there exist a variety of mechanisms for resolving the dispute. These include civil litigation, arbitration and, in labour relations, the strike. Resolving disputes in these ways is costly, thereby creating a contract zone within which both parties strictly prefer to settle. Given the high cost of disputes, considerable research has been devoted to understanding why settlement sometimes fails to occur and how different mechanisms affect the dispute rate. Here we focus on dispute resolution in the context of civil litigation and arbitration.

Why settlement fails

The dominant rational choice explanation for settlement failure is asymmetric information. An alternative explanation, not consistent with the assumption of full rationality, is optimism. If agents have

symmetric information and beliefs about the expected outcome of a dispute, theory suggests a settlement will occur. However, if one party has private information about the expected outcome of the dispute, settlement failure can occur. Similarly, if one or both parties to the dispute are subject to optimism, a contract zone may fail to exist.

There are two basic models in the asymmetric information literature, which make different assumptions about the structure of information. When the uninformed party makes the offer, we have a screening model which was developed by Bebchuk (1984). When the informed party makes the offer we have a signalling model developed by Reinganum and Wilde (1986). Both models' predictions are consistent with the existence of costly disputes in equilibrium.

To explore the intuition behind these models, consider a civil dispute in which the failure of negotiations would result in a trial. Suppose a plaintiff known to be harmed has private information concerning the damages she has incurred and that this information would be revealed at trial. Further, suppose the plaintiff is one of two types: a weak type with a low expected payoff at trial or a strong type with a high expected payoff. A risk-neutral plaintiff will accept a settlement offer if and only if it equals or exceeds her expected net payoff from trial. The defendant knows the probability that he is facing a weak or strong plaintiff but not the plaintiff's exact type. In a screening model, the uninformed defendant makes an offer to the plaintiff. He will choose between a low (screening) offer that only weak plaintiffs would accept and a high (pooling) offer that both types would accept. If he makes the low offer, then a strong plaintiff would proceed to trial. The screening offer is more likely to be optimal for the defendant when there is a high prevalence of weak plaintiffs, when court costs are low, and when the difference in expected trial awards for the two plaintiff types is large.

If the informed plaintiff is allowed to make the offer, this is called the signalling game. While these games generally have multiple equilibria, the D1 refinement (Cho and Kreps, 1987) has been employed to focus on a separating equilibrium in which the weak plaintiff submits a low demand to the defendant, while the strong plaintiff submits a high demand. Under D1, it is assumed that an out of equilibrium offer is made by the plaintiff willing to make that offer for the largest set of acceptance probabilities. In equilibrium, the high demand must be rejected with a sufficiently high probability so as to discourage the weak plaintiff from also making this demand. These rejections lead to a positive probability of trial with the strong plaintiff.

While we used a two-type model to motivate the discussion, the Bebchuk and Reinganum and Wilde models employ a continuum of types whose distribution is known by the uninformed party. These models have been extended in numerous ways by allowing for two-sided information asymmetries (Schweizer, 1989; Daughety and Reinganum, 1994) and multiple offers (Spier, 1992) among other extensions. While the effects of policy variables (such as cost shifting at trial) are often sensitive to the modelling details, the prediction that asymmetric information can result in costly disputes is quite robust. Excellent surveys of the literature are provided by Spier (1998) and Daughety (1999).

The empirical studies by McConnell (1989), Conlin (1999) and Osborne (1999) support the model of asymmetric information.

The optimism or self-serving bias explanation for settlement failure relies on bargainers who have potentially inaccurate beliefs about the expected outcome at trial. For example, the plaintiff's belief about the probability she will prevail at trial may exceed the defendant's belief about this same probability. If these differences in beliefs are not based on differences in information, then we are in the

realm of the optimism model. Versions of this model have been employed by Landes (1971), Posner (1973), Shavell (1982), and Priest and Klein (1984). Optimism violates rationality, but Bar-Gill (2002) finds that cautious optimism can allow the optimistic party to obtain a larger portion of the joint surplus from settlement. As a result, cautious optimism can persist in an evolutionary setting. Babcock and Loewenstein (1997) survey an experimental literature documenting the existence of a self-serving bias which leads players in the role of a plaintiff to expect a greater payoff at trial than the defendant, even though both are exposed to the same set of facts. When players are exposed to the facts of the case before being assigned their role as plaintiff or defendant, the self-serving bias tends to disappear. Waldfogel (1998) and Farmer, Pecorino and Stango (2004) find empirical evidence that is consistent with the optimism model. Note that the optimism and asymmetric information explanations are not mutually exclusive. It is possible that each factor is responsible for some proportion of observed disputes.

Mandatory discovery and voluntary disclosure

If asymmetric information causes disputes, it is logical to ask whether voluntary disclosures and mandatory discovery can eliminate these asymmetries. In a screening model where credible disclosure is costless, Shavell (1989) shows that plaintiffs with strong cases will reveal enough information to ensure that all cases settle. Plaintiffs who do not reveal their information (those with weak cases) receive a pooling offer that all accept. However, the work of Sobel (1989) shows that this result is not robust to the introduction of positive costs of disclosure. He also shows that a costless (to the plaintiff) discovery procedure will lead to greater settlement. Farmer and Pecorino (2005) consider costly discovery and disclosure in both the signalling and the screening games. Costly disclosures may be made in the signalling game but not the screening game, while costly discovery may be invoked in the screening game but not the signalling game. If the cost of these procedures is not too high, the combination of the two will lead to a great deal of information transmission and a large reduction in the dispute rate.
Why then do disputes persist? Perhaps, as Shavell (1989) suggests, private information has strategic value if withheld until trial. Hay (1995) develops a model in which an initial informational asymmetry on the merits of the case is resolved by mandatory discovery, but by the time this occurs a new asymmetry – namely, the extent of attorney preparation – has emerged. This second asymmetry leads to trials in the equilibrium of the model. Hay notes that the extent of attorney preparation is not subject to discovery. This is also true of preferences. Farmer and Pecorino (1994) show that asymmetric information on risk preferences can lead to trial, and that this information is neither subject to discovery nor easy to credibly transmit. As a result, this type of asymmetry may tend to persist in the face of mandatory discovery and opportunities for voluntary disclosure.

Other institutional features

There is a voluminous literature which examines how a variety of institutional features affect settlement in civil litigation. What follows is a much abbreviated discussion of a large and complex literature. One difficulty in addressing this question is that even a single institution is likely to have multiple effects on the litigation process. Thus, a single institution may have conflicting effects on the dispute rate and may

also have important influences on other aspects of the litigation process.

A classic example of this difficulty is reflected in the analysis of the English rule under which the loser at trial pays the reasonable legal costs of the winner. If the probability of a finding for the plaintiff at trial is private information, then fee shifting at trial reduces settlement rates by, in effect, spreading out the distribution of player types (Bebchuk, 1984). If players are optimistic in their assessments of the probability that the plaintiff will prevail, then fee shifting will aggravate this optimism and reduce the probability that a contract zone will exist (Shavell, 1982).

This prediction – that fee shifting will increase the probability of trial – is made with expenditure at trial held constant. It is well established that the fee shifting at trial will increase expenditure (Braeutigam, Owen, and Panzar, 1984). If the expenditure effect is strong enough, it can result in fewer (but more costly) disputes (Hause, 1989). The English rule also affects the mix of cases which are filed. It discourages cases where there are large stakes but a low probability of success, and encourages low stakes cases with a high probability of success (Shavell, 1982).

Many of the theoretical predictions on fee shifting at trial appear to be borne out in the data (see Hughes and Snyder, 1998).

Under a contingency fee, the plaintiff's lawyer receives a percentage of the judgment at trial if the plaintiff wins the case and nothing if she loses. The effects of contingency fees on the litigation process are very complex and wide ranging (see Rubinfeld and Scotchmer, 1998, for a survey). However, one effect of contingency fees on settlement is clear: if the attorney controls the settlement decision, he will have an excessive incentive to settle the case relative to the interests of his client. The reason is that the attorney bears most of the costs of a trial but is paid only a fraction of the award. On the other hand, if the client controls the case, she may have an excessive incentive to reject a settlement offer and bring the case to trial. (This is particularly true if the contingency percentage is not lower for cases which settle early.)

When a single defendant faces multiple plaintiffs in sequence, some interesting issues regarding settlement arise. Spier (2003a; 2003b) and Daughety and Reinganum (2004) have analysed the use of most favoured nation (MFN) clauses in the context of repeat litigation. Suppose a plaintiff settles early under MFN. If another plaintiff later settles for more, the early settlement is adjusted upward. An MFN clause can be a mechanism whereby the defendant commits to not raising his offer to plaintiffs who settle later in the process. While there is some ambiguity of the effects of MFN on settlement rates and the overall dispute costs (see especially Spier, 2003a), the general thrust of these papers suggests that MFN clauses are efficiency enhancing in the sense that they will reduce the expected dispute costs associated with litigation.

Arbitration

Under conventional arbitration (CA), the arbitrator is free to impose her preferred settlement on the bargaining parties. Under final offer arbitration (FOA), each party to the dispute submits an offer to the arbitrator who must pick one of the submitted offers. While there is some evidence that submitted offers affect the outcome in CA (Farber and Bazerman, 1986), for the purpose of the following discussion we assume that they do not. From a modelling standpoint, this makes CA look exactly like a simple version

of civil litigation. Under FOA, the submitted offers clearly affect the outcome, a feature which has important implications for dispute resolution.

Consider the two-type version of the screening model where the plaintiff can have a strong or a weak case. In CA, the defendant will either make an offer that only a weak plaintiff will accept or a pooling offer that both types will accept. If all negotiation takes place prior to the submission of offers to the arbitrator, then under FOA it is possible that the defendant will make an offer that neither type will accept, resulting in a 100 per cent dispute rate (Farmer and Pecorino, 2003). This can occur because the sequentially rational offer submitted to the arbitrator influences the acceptable settlement prior to arbitration. The lack of early settlement allows the defendant to commit to an offer which is optimal against the entire distribution of plaintiff types. Farmer and Pecorino (2003) also show (in contrast to CA) that costless voluntary disclosure never takes place when FOA is the dispute resolution mechanism. The reason is that information has strategic value in this game. Both of the impediments to settlement discussed above disappear if bargaining is permitted after offers are submitted to the arbitrator but prior to the arbitration hearing.

While not totally conclusive on this point, the results of Farmer and Pecorino (1998) also suggest that allowing for bargaining after offers are submitted to the arbitrator can increase settlement for reasons different from those discussed above. Because a submitted offer affects the outcome of arbitration, it will tend to reflect private information. This may in turn promote settlement. Taken together, the results on FOA suggest that the effects of this institution on settlement are sensitive to whether or not bargaining occurs in the face of offers submitted to the arbitrator. In major league baseball, a prominent use of FOA, a good deal of bargaining and settlement occurs after offers have been submitted to the arbitrator.

FOA was proposed by Stephens (1966) and has since become an important alternative to CA. Researchers continue to propose new arbitration mechanisms in the hope of improving the dispute resolution process. Combined arbitration (Brams and Merrill, 1986) is a mixture of FOA and CA. Other proposed mechanisms include tri-offer arbitration (Ashenfelter et al., 1992) and amended final offer arbitration (Zeng, 2003).

See Also

- epistemic game theory: an overview
- epistemic game theory: incomplete information

Bibliography

Ashenfelter, O., Currie, J., Farber, H. and Spiegel, M. 1992. An experimental comparison of dispute rates in alternative arbitration systems. *Econometrica* 60, 1407–33.

Babcock, L. and Loewenstein, G. 1997. Explaining bargaining impasse: the role of self-serving biases. *Journal of Economic Perspectives* 11, 109–26.

Bar-Gill, O. 2002. *The Success and Survival of Cautious Optimism: Legal Rule, and Endogenous Perceptions in Pre-Trial Settlement Negotiations*. Public Law Working Paper No. 35. Cambridge, MA: Harvard Law School.

Bebchuk, L. 1984. Litigation and settlement under imperfect information. *RAND Journal of Economics* 15, 404–15.

Braeutigam, R., Owen, B. and Panzar, J. 1984. An economic analysis of alternative fee shifting systems. *Law and Contemporary Problems* 47, 173–85.

Brams, S. and Merrill, S. 1986. Binding versus final-offer arbitration: a combination is best. *Management Science* 32, 1346–55.

Cho, I. and Kreps, D. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102, 179–222.

Conlin, M. 1999. Empirical test of a separating equilibrium in National Football League contract negotiations. *RAND Journal of Economics* 30, 289–304.

Daughety, A. 1999. Settlement. In *Encyclopedia of Law and Economics*, Vol. 5, ed. B. Bouckaert and G. de Geest. Cheltenham: Edward Elgar.

Daughety, A. and Reinganum, J. 1994. Settlement negotiations with two-sided asymmetric information: model duality, information distribution, and efficiency. *International Review of Law and Economics* 14, 283–98.

Daughety, A. and Reinganum, J. 2004. Exploiting future settlements: a signaling model of most-favored-nation clauses in settlement bargaining. *RAND Journal of Economics* 35, 467–85.

Farber, H. and Bazerman, M. 1986. The general basis of arbitrator behavior: an empirical analysis of conventional and final offer arbitration. *Econometrica* 54, 819–54.

Farmer, A. and Pecorino, P. 1994. Pretrial negotiations with asymmetric information on risk preferences. *International Review of Law and Economics* 14, 273–81.

Farmer, A. and Pecorino, P. 1998. Bargaining with informative offers: an analysis of final offer arbitration. *Journal of Legal Studies* 27, 415–432.

Farmer, A. and Pecorino, P. 2003. Bargaining with voluntary transmission of private information: does the use of final offer arbitration impede settlement? *Journal of Law, Economics and Organization* 19, 64–82.

Farmer, A. and Pecorino, P. 2005. Civil litigation with mandatory discovery and voluntary transmission of private information. *Journal of Legal Studies* 34, 137–59.

Farmer, A., Pecorino, P. and Stango, V. 2004. The causes of bargaining failure: evidence from major league baseball. *Journal of Law and Economics* 47, 543–68.

Hause, J. 1989. Indemnity, settlement, and litigation, or I'll be suing you. *Journal of Legal Studies* 18, 157–79.

Hay, B. 1995. Effort, information, settlement, trial. *Journal of Legal Studies* 24, 29–62.

Hughes, J. W. and Snyder, E. A. 1998. Allocation of litigation costs: American and English rules. In *The New Palgrave Dictionary of Economics and the Law*, Vol. 1, ed. P. Newman. London: Macmillan.

Landes, W. 1971. An economic analysis of the courts. *Journal of Law and Economics* 14, 61–107.

McConnell, S. 1989. Strikes, wages and private information. *American Economic Review* 79, 801–15.

Osborne, E. 1999. Who should be worried about asymmetric information in litigation? *International Review of Law and Economics* 19, 399–409.

Posner, R. 1973. An economic approach to legal procedure and judicial administration. *Journal of Legal Studies* 2, 399–458.

Priest, G. and Klein, B. 1984. The selection of disputes for arbitration. *Journal of Legal Studies* 13, 215–43.

Reinganum, J. and Wilde, L. 1986. Settlement, litigation, and the allocation of litigation costs. *RAND Journal of Economics* 17, 557–66.

Rubinfeld, D. and Scotchmer, S. 1998. Contingent fees. In *The New Palgrave Dictionary of Economics and the Law*, vol. 1, ed. P. Newman. London: Macmillan.

Schweizer, U. 1989. Litigation and settlement under two-sided incomplete information. *Review of Economic Studies* 56, 163–78.

Shavell, S. 1982. Suit, settlement, and trial: a theoretical analysis under alternative methods for the allocation of legal costs. *Journal of Legal Studies* 11, 55–82.

Shavell, S. 1989. Sharing of information prior to settlement or litigation. *RAND Journal of Economics* 20, 183–195.

Sobel, J. 1989. An analysis of discovery rules. *Law and Contemporary Problems* 52, 133–59.

Spier, K. 1992. The dynamics of pretrial negotiation. *Review of Economic Studies* 59, 93–108.

Spier, K. 1998. Settlement of litigation. In *The New Palgrave Dictionary of Economics and the Law*, Vol. 3, ed. P. Newman. London: Macmillan.

Spier, K. 2003a. ‘Tied to the mast’: most-favored-nation clauses in settlement contracts. *Journal of Legal Studies* 32, 91–120.

Spier, K. 2003b. The use of ‘most-favored-nation’ clauses in settlement of litigation. *Rand Journal of Economics* 34, 78–95.

Stephens, C. 1966. Is compulsory arbitration compatible with bargaining? *Industrial Relations*, 5(1), 38–52.

Waldfogel, J. 1998. Reconciling asymmetric information and divergent expectations theories of litigation. *Journal of Law and Economics* 41, 451–76.

Zeng, D. 2003. An amendment to final offer arbitration. *Mathematical Social Sciences*, 46(1), 9–19.

How to cite this article

Farmer, Amy and Paul Pecorino. "dispute resolution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000242> doi:10.1057/9780230226203.0395

The New Palgrave Dictionary of Economics Online

distributed lags

Philip Hans Franses

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews various aspects of distributed lag models. Specific attention is paid to the interpretation of model parameters.

Keywords

Almon lags; ARMA models; cointegration; distributed lags; error correction models; Koyck transformation; maximum likelihood; multicollinearity; nonlinear least squares; ordinary least squares; vector autoregressions

Article

Distributed lag models correlate a single dependent variable with its own lags and with current and lagged values of one or more explanatory variables. Examples concern the current and dynamic correlations between output and investment and between sales and advertising. Distributed lag models typically assume that the explanatory variable is exogenous. (In case of doubt, one usually resorts to vector autoregressive models where two or more variables can be endogenous; see Sims, 1980.)

This article highlights a few aspects of distributed lag models. The two main aspects are representation and interpretation. Useful extended surveys appear in Dhrymes (1971), Griliches (1967) and Hendry, Pagan and Sargan (1984).

Representation

Consider a dependent variable y_t and, for ease of notation, a single explanatory variable x_t. Indicator t runs from 1 to n and it can concern seconds, hours, days or even years. A general (autoregressive) distributed lag model is given by

$$y_t = \mu + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_m x_{t-m} + \varepsilon_t, \quad (1)$$

where p and m can take any positive integer value, and where it is usually assumed that ε_t is an uncorrelated variable with mean zero and variance σ^2. (Part of the literature assumes the label distributed lags model for the case where $p = 0$ and $m = \infty$. Below we will see that such a model is often approximated by a model as in (1).) As the model contains the lagged dependent variables, it is called an autoregressive distributed lag model with orders p and m, in short ADL(p, m). The model allows for delayed effects of x_t, as β_0 can be 0, and it also allows for time gaps in these effects when some β parameters are zero and others are not.

Reducing the number of parameters

Basically, given fixed and finite values of p and m, the parameters in (1) can be consistently estimated with ordinary least squares (OLS). (Typically one uses information criteria as those of Akaike or Schwarz to choose the relevant values of p and m in practice.) In practice, p and m can be large, and in theory even as large as ∞. This can be inconvenient, for two reasons. First, the variables y_t and x_t each can be strongly autocorrelated, and then the regression in (1) suffers from multicollinearity. Second, with many parameters in a model there might be many values to evaluate and interpret.
To reduce the number of parameters and to facilitate interpretation, one can impose restrictions. Early suggestions are the Almon and Shiller lag structures, where the parameters are made functions of i, $i = 0, 1, 2, \ldots, m$ (see Almon, 1965, and Shiller, 1973), and the so-called Koyck transformation (see Koyck, 1954).

Almon and Shiller transformations

Consider the version of (1) with $p = 0$ and $m = m$ and set μ at 0 for convenience, that is, consider

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_m x_{t-m} + \varepsilon_t. \tag{2}$$

Almon (1965) proposes to reduce the number of parameters by assuming the approximation

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \ldots + \alpha_q i^q, \tag{3}$$

with $q > m$. This makes the sequence of β_i parameters a polynomial and hence a smooth function without possibly implausible spikes.
Working out the Almon lags, one can derive that the structure implies that

$$\beta_{i+1} - 2\beta_i + \beta_{i-1} = \gamma_i, \tag{4}$$

where γ_i is a function of α_i values. Shiller (1973) considers this as too restrictive and he proposes to assume that

$$\beta_{i+1} - 2\beta_i + \beta_{i-1} \sim N(0, \zeta^2), \tag{5}$$

for $i = 1, 2, \ldots, m-1$.

Koyck transformation

The Koyck model can be interpreted as a model which includes adaptive expectations. Suppose that

$$y_t = \alpha + \beta x_t^* + \varepsilon_t, \tag{6}$$

where x_t^* denotes the expected value of x_t, an expectation formed at $t-1$. When the adaptive expectations schedule is assumed, like

$$x_t^* = \lambda x_{t-1}^* + (1-\lambda) x_t, \tag{7}$$

with again $|\lambda| < 1$, then substituting (7) into (6) gives

$$y_t = \alpha(1-\lambda) + \lambda y_{t-1} + \beta(1-\lambda) x_t + \varepsilon_t - \lambda\varepsilon_{t-1}. \tag{8}$$

The short-run effect of x_t on y_t is $\beta(1-\lambda)$, while the long-run effect is $\frac{\beta(1-\lambda)}{1-\lambda} = \beta$, as could be expected given

(6).
Consider the case were *m* equals ∞, and where all α parameters are set to zero. When it is further assumed that $\beta_j = \beta_0 \lambda^{j-1}$, with $|\lambda| < 1$ for *j* is 1,2,…, then (1) becomes

$$y_t = \mu + \beta_0 x_t + \beta_0 \lambda x_{t-1} + \beta_0 \lambda^2 x_{t-2} + \ldots + \varepsilon_t. \tag{9}$$

Subtracting λy_{t-1} from this expression gives

$$y_t = (1-\lambda)\mu + \lambda y_{t-1} + \beta_0 x_t + \varepsilon_t - \lambda \varepsilon_{t-1}, \tag{10}$$

which is again (8). This Koyck transformation leads to a rather simple model with a moving average (MA) error term. The appropriate estimation method is maximum likelihood, as it is described in, for example, Hamilton (1994, p. 132) for general ARMA models. Note that the parameter λ appears in the autoregressive part and in the MA part.

Restructuring the model

An alternative way to reduce the number of parameters, also in order to facilitate interpretation, is to restructure the model.
To overcome multicollinearity, one can rewrite model (1) in the so-called error correction format. This format combines levels and differences of levels, which is convenient as these are usually much less correlated than the levels themselves, and hence multicollinearity will be much less of a problem. An additional feature of the error correction format is that it provides an immediate look at key parameters such as the total effect, the current effect, and the speed at which the total effect is accomplished.
With Δ_j denoted as the *j*-th order differencing filter, that is, $\Delta_j y_t = y_t - y_{t-j}$, an error correction representation for (1) reads as

$$\Delta_1 y_t = \mu + \left(\sum_{j=1}^{p} \alpha_j - 1\right)\left[y_{t-1} - \frac{\Sigma_{i=0}^{m} \beta_i}{1 - \Sigma_{j=1}^{p} \alpha_j} x_{t-1}\right] + \beta_0 \Delta_1 x_t - \sum_{i=2}^{m} \beta_i \Delta_{i-1} x_{t-1} - \sum_{j=2}^{p} \alpha_j \Delta_{j-1} y_{t-1} + \varepsilon_t, \tag{11}$$

where lagged levels are suitably combined into differenced variables such that at each lag a higher-order differenced variable appears. This representation even further reduces chances of having multicollinearity. Note

that the model can also be written in terms of lagged levels and first differences only, that is as

$$\Delta_1 y_t = \mu + \left(\sum_{j=1}^{p} \alpha_j - 1\right)\left[y_{t-1} - \frac{\Sigma_{i=0}^{m} \beta_i}{1 - \Sigma_{j=1}^{p} \alpha_j} x_{t-1}\right] + \beta_0 \Delta_1 x_t + \sum_{i=1}^{m} \gamma_i \Delta_1 x_{t-i} + \sum_{j=1}^{p} \theta_j \Delta_1 y_{t-j} + \varepsilon_t \quad (12)$$

With the use of (11), all but two parameters (that is, α_1 and β_1) can be directly estimated by using OLS, while $\hat{\alpha}_1$ and $\hat{\beta}_1$ straightforwardly follow from applying OLS to (1). Note that model (11) can also be written such that the levels (now at $t-1$) enter at $t-2$ or, say, $t-p$.

Interpretation

We now turn to the interpretation of distributed lag models.

Long-run and short-run effects

The error correction model in (11) provides immediate estimates of current and dynamic effects. (Fok et al., 2006, show that when the series y_t and x_t have a unit root and are cointegrated, as defined by Engle and Granger, 1987, one should speak of the long-run effect, while when the series are stationary there is a total or cumulative effect. For the latter, see also Hendry, Pagan and Sargan, 1984.) The current effect is β_0 and the long-run or total effect is

$$\frac{\Sigma_{i=0}^{m} \beta_i}{1 - \Sigma_{j=1}^{p} \alpha_j}. \quad (13)$$

Note that the long-run effect can be larger or smaller than the short-run effect, depending on the values of the parameters. The parameters in the error correction model, when written as

$$\Delta_1 y_t = \mu + \rho[y_{t-1} - \gamma x_{t-1}] + \beta_0 \Delta_1 x_t - \sum_{i=2}^{m} \beta_i \Delta_{i-1} x_{t-1} - \sum_{j=2}^{p} \alpha_j \Delta_{j-1} y_{t-1} + \varepsilon_t, \quad (14)$$

can be estimated using non-linear least squares. This method provides direct estimates of the long-run effect γ

and its associated standard error.

Duration interval

As well as the long-run and short-run effects, one may also be interested in the speed with which the effect of x_t decays over time. To be able to compute this so-called duration interval, one needs explicit expressions for $\frac{\partial y_{t+k}}{\partial x_t}$ for all values of k running from 1 to, potentially, ∞. Given the expression in (1), these expressions are easily derived as

$$\frac{\partial y_t}{\partial x_t} = \beta_0$$

$$\frac{\partial y_{t+1}}{\partial x_t} = \beta_1 + \alpha_1 \frac{\partial y_t}{\partial x_t}$$

$$\frac{\partial y_{t+2}}{\partial x_t} = \beta_2 + \alpha_1 \frac{\partial y_{t+1}}{\partial x_t} + \alpha_2 \frac{\partial y_t}{\partial x_t}$$

$$\vdots$$

$$\frac{\partial y_{t+k}}{\partial x_t} = \beta_k + \sum_{j=1}^{k} \alpha_j \frac{\partial y_{t+(k-j)}}{\partial x_t}$$

where it should be noted that $\alpha_k = 0$ for $k > p$, and that $\beta_k = 0$ for $k > m$. Hence, the final form of a distributed lag model (see Harvey, 1990), is

$$y_t = \sum_{i=0}^{\infty} \delta_i x_{t-i} + error, \tag{15}$$

where

$$\delta_i = \frac{\partial y_{t+i}}{\partial x_t}. \tag{16}$$

With these δ_i, one can derive all kinds of summary effects (like mean and median, or half lives of shocks) of x_t

on y_t.
When δ_i decays monotonically, it is useful to define the decay factor by

$$p_k = \frac{\frac{\partial y_t}{\partial x_t} - \frac{\partial y_{t+k}}{\partial x_t}}{\frac{\partial y_t}{\partial x_t}}. \tag{17}$$

This can be computed only for discrete values of k as there are only discrete time intervals. This decay factor is a function of the model parameters. Through interpolation, one can decide on the time k it takes for the decay factor to be equal to some value of p, which typically is equal to 0.95 or 0.90. This estimated time k is then called the p per cent duration interval. This measure is frequently used in advertising research (see Clarke, 1976; Leeflang et al., 2000; Franses and Vroomen, 2006).

Final issues

Distributed lag models continue to be a standard empirical approach. When the models are applied, there are at least two further issues that one needs to address, that is, next to selecting p and m and a useful transformation. The first concerns the statistical analysis of the model. For example, if y_t and x_t are not stationary, one needs to rely on cointegration techniques that involve non-standard asymptotic theory. The theory that is most relevant here is formulated in Boswijk (1995). Also, in the case of the Koyck model, one faces the so-called Davies (1987) problem. Under the null hypothesis that $\beta_0 = 0$, the model collapses to $y_t = \varepsilon_t$, and hence λ is not identified then.
The second issue concerns aggregation over time. It may be that y_t and x_t are not available at the same sampling frequency. For example, television commercials last for 30 seconds and recur each hour, say, while sales data are available only at the weekly level. Tellis and Franses (2006) have a few recent results, but more work is needed.

Bibliography

Almon, S. 1965. The distributed lag between capital appropriations and expenditures. *Econometrica* 33, 178–96.

Boswijk, H. 1995. Efficient Inference on cointegration parameters in structural error–correction models. *Journal of Econometrics* 69, 133–58.

Clarke, D. 1976. Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research* 8, 345–57.

Davies, R. 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–54.

Dhrymes, P. 1971. *Distributed Lags: Problems of Estimation and Formulation*. San Francisco: Holden-Day.

Engle, R. and Granger, C. 1987. Cointegration and error correction: representation, estimation, and testing. *Econometrica* 55, 251–76.

Fok, D., Horvath, C., Paap, R. and Franses, P. 2006. A hierarchical Bayes error correction model to explain dynamic effects of price changes. *Journal of Marketing Research* 43, 443–61.

Franses, P. and Vroomen, B. 2006. Estimating confidence bounds for advertising effect duration intervals. *Journal of Advertising* 35(Summer), 33–37.

Griliches, Z. 1967. Distributed lags: a survey. *Econometrica* 35, 16–49.

Hamilton, J. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Hansen, B. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64, 413–30.

Harvey, A. 1990. *The Econometric Analysis of Time Series*. London: Philip Allan.

Hendry, D., Pagan, A. and Sargan, J. 1984. Dynamic specification. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Koyck, L. 1954. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.

Leeflang, P., Wittink, D., Wedel, M. and Naert, P. 2000. *Building Models for Marketing Decisions*. Boston: Kluwer.

Shiller, R. 1973. A distributed lag estimator derived from smoothness priors. *Econometrica* 41, 775–88.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Tellis, G. and Franses, P. 2006. The optimal data interval for econometric models of advertising. *Marketing Science* 23, 217–229.

How to cite this article

Franses, Philip Hans. "distributed lags." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000154> doi:10.1057/9780230226203.0396

The New Palgrave Dictionary of Economics Online

distributive politics and targeted public spending

Brian G. Knight

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article analyses common pool problems associated with the provision of local public goods by central legislatures. In response to incentives associated with common pool problems, legislators act to maximize spending for their home jurisdiction but to restrain spending elsewhere due to the associated tax costs. The resolution of this conflict between jurisdictions depends in the United States upon the distribution of political power across Congressional delegations. Incumbents are rewarded for delivering federal spending to their jurisdiction through increased voter support.

Keywords

common pool problems; distributive politics; earmarked projects; lobbying; local public goods; proposal power; targeted public spending

Article

While conventional models of political economy, such as the median voter model, focus on the provision of national public goods, most federal spending programmes, such as the US interstate highway system, are more aptly characterized as local in nature. While in the United States the benefits of federal spending are concentrated in specific geographic units, such as states, counties, and Congressional districts, the associated tax costs are, by contrast, geographically dispersed. This common pool feature of federal spending – concentrated spending but dispersed financing – leads to a geographic tug-of-war in which jurisdictions attempt to increase own-jurisdiction spending but to reduce spending elsewhere due to the associated tax costs. This conflict between jurisdictions is reflected most intensely in the budget process within the US Congress, whose members are locally elected and thus naturally respond to these common pool incentives.

In this article, I first summarize evidence suggesting that Congressional representatives are responsive to the common pool incentives associated with concentrated spending but dispersed costs. Having established the empirical saliency of this common pool problem in Congress, I then summarize the

literature examining how this conflict is resolved. In particular, I analyse the effects of Congressional delegation characteristics, such as size, ideology, seniority, and committee assignments, on the geographic allocation of federal funds. Finally, I review evidence on the effects of the geographic distribution of federal funds on electoral outcomes.

As described in Knight (2006), common pool problems underpin several theoretical models of the legislative process, such as the universalism model of Weingast, Shepsle, and Johnsen (1981) and the legislative bargaining model of Baron and Ferejohn (1989). Whether or not Congressional delegations respond to these incentives in practice, however, is primarily an empirical question. It may be the case, for example, that political parties, or related Congressional organizations, serve as collective mechanisms through which legislators internalize the tax costs in other jurisdictions associated with own-jurisdiction spending. One of the first papers to directly measure the responsiveness of representatives to common pool problems is by DelRossi and Inman (1999), who examine the geographic distribution of water projects authorized by the Water Resources Development Act of 1986. In particular, the authors compare the size of project requests before and after changes in local matching requirements, which significantly increased the fraction of project costs financed by local governments. As hypothesized, districts experiencing larger increases in matching rates requested significantly less funding for water projects. In a similar vein, Knight (2004b) examines Congressional voting in 1998 over whether to finance a set of transportation projects, which were earmarked for specific Congressional districts and were funded primarily via federal gasoline taxes. As predicted, support for funding was concentrated in those districts receiving more in funding and also in those districts with lower gasoline tax burdens.

How is this geographic battle between jurisdictions resolved? Which states and Congressional districts win and why? Regarding the mere size of delegations, an important feature of the US Congress is its bicameral structure in which each state has an equal number of delegates in the Senate but in which seats are apportioned between states according to population in the House of Representatives. This equality of delegation sizes in the US Senate provides small states with power disproportionate to their population; Senators from California, the largest state, currently have over 60 times as many constituents as do senators from Wyoming, the smallest state. In attempting to measure the magnitude of this small-state bias, Atlas et al. (1995) and Lee (1998) find that small states receive significantly more per capita in aggregate federal spending than do large states. While this finding is certainly provocative, it is difficult to distinguish between the role of Senate representation and other factors, such as population density, that make small states inherently different from larger states. In attempting to address this issue of unobserved differences between small and large states, Knight (2004a) demonstrates that small states receive considerably more per-capita funding in projects earmarked in Senate bills; in House bills, by contrast, small and large states receive similar project spending on a per-capita basis. Knight (2004b) also identifies two theoretical channels underlying this small-state bias in the US Senate. Relative to their population, small states are disproportionately represented on key committees (the proposal power channel) but are also cheaper coalition partners (the vote cost channel) given that they pay a smaller share of federal taxes. Interestingly, both channels are shown to be empirically important and, taken together, explain over 90 per cent of the measured small-state bias. In a related study of the size of delegations, Falk (2006) studies discontinuities in the apportionment of seats in the US House arising from both timing (re-apportionment occurs once every ten years) and rounding issues (delegation sizes must be integers). Using this variation in delegation sizes, he finds that increases in seats per capita lead

to statistically significant increases in federal spending per capita.

Delegations of similar sizes, however, may differ significantly in their composition. Key differences between delegations in the degree of political power include majority party affiliation, seniority, and representation on key committees. Regarding majority party affiliation, Levitt and Snyder (1995) find that the Democratic Party used its majority control of Congress to channel federal funds into Congressional districts with a high percentage of Democratic voters during the period 1984–90. However, they find no evidence that, conditional on the percentage of Democratic voters, districts represented by Democrats received higher federal spending. Levitt and Poterba (1999) report that states with very senior Democratic representatives experienced more rapid economic growth than did other states. However, they find no relationship between the partisan affiliation of delegations and the geography of federal spending, a key hypothesized channel of the measured differences in economic growth. Regarding the role of Congressional committees, Knight (2005) finds that Congressional districts represented on key committees received substantially more funding in projects earmarked in transportation bills authorized in 1991 and 1998. He interprets this result as evidence of the importance of proposal power associated with the committee's ability to set the legislative agenda. De Figueiredo and Silverman (2002) examine interactions between committee representation and lobbying in an empirical examination of earmarked projects for universities. In particular, they find a strong correlation between lobbying outlays by universities and the receipt of federal funding; this link between lobbying and spending, however, is found to be much stronger for those universities located in districts that are represented on key appropriations committees.

We have focused throughout this survey on the determinants of the geographic distribution of federal funds. Politicians, however, have an incentive to put forth the effort to secure project funding only if they perceive that the associated political gains are sufficiently high. While clearly important, measurement of the effects of federal spending on incumbent vote shares is plagued with endogeneity problems. For example, incumbents facing the strongest opposition have the strongest incentives to put forth effort in securing funds. Thus, there may be a downward bias in ordinary least squares (OLS) estimates of the effect of federal spending on incumbent vote shares. As an instrument for district-specific federal spending, Levitt and Snyder (1997) use federal spending outside of the district but within the state. The idea is that other actors, such as Senators or governors, also play a role in the geographic distribution of federal funds. Using this exogenous variation in federal spending, they conclude that an additional $100 per capita in spending translates into an additional two percentage points in incumbent vote shares.

We conclude that common pool problems associated with concentrated project benefits but dispersed costs are reflected not only in the behaviour of Congressional delegations but also in the resulting distribution of federal funds. Who wins and who loses in this geographic battle is determined in part by state size and the political power of delegations. Consistent with these results, evidence suggests that incumbent re-election prospects are significantly enhanced by increases in federal spending.

See Also

- campaign finance, economics of
- fiscal federalism

- intergovernmental grants
- local public finance
- political institutions, economic approaches to
- public choice

Bibliography

Atlas, C.M., Gilligan, T.W., Hendershott, R.J. and Zupan, M.A. 1995. Slicing the federal government net spending pie: who wins, who loses, and why. *American Economic Review* 85, 624–29.

Baron, D.P. and Ferejohn, J.A. 1989. Bargaining in legislatures. *American Political Science Review* 83, 1881–207.

de Figueiredo, J.M. and Silverman, B.S. 2002. Academic earmarks and the returns to lobbying. Working Paper No. 9064. Cambridge, MA: NBER.

Falk, J. 2006. The effects of Congressional district size and representative's tenure on the allocation of federal funds. Working paper, University of California, Berkeley.

DelRossi, A.F. and Inman, R.P. 1999. Changing the price of pork: the impact of local cost sharing on legislators’ demands for distributive public goods. *Journal of Public Economics* 71, 247–73.

Knight, B.G. 2004a. Legislative representation, bargaining power, and the distribution of federal funds: evidence from the U.S. Senate. Working Paper No. 10385. Cambridge, MA: NBER.

Knight, B.G. 2004b. Parochial interests and the centralized provision of local public goods: evidence from congressional voting on transportation projects. *Journal of Public Economics* 88, 845–66.

Knight, B.G. 2005. Estimating the value of proposal power. *American Economic Review* 95, 1639–52.

Knight, B.G. 2006. Common tax pool problems in federal systems. In *Democratic Constitutional Design and Public Policy Analysis and Evidence*, ed. R.D. Congleton and B. Swedenborg. Cambridge, MA: MIT Press.

Lee, F.E. 1998. Representation and public policy: the consequences of Senate apportionment for the geographic distribution of federal funds. *Journal of Politics* 60, 34–62.

Levitt, S.D. and Poterba, J.M. 1999. Congressional distributive politics and state economic performance. *Public Choice* 99, 185–216.

Levitt, S.D. and Snyder, J.M. 1995. Political parties and the distribution of federal outlays. *American*

Journal of Political Science 39, 958–80.

Levitt, S.D. and Snyder, J.M. 1997. The impact of federal spending on House election outcomes. *Journal of Political Economy* 105, 30–53.

Weingast, B.R., Shepsle, K.A. and Johnsen, C. 1991. The political economy of benefits and costs: a neoclassical approach to distributive politics. *Journal of Political Economy* 89, 642–64.

How to cite this article

Knight, Brian G. "distributive politics and targeted public spending." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000261>
doi:10.1057/9780230226203.0397

The New Palgrave Dictionary of Economics Online

dividend policy

David J. Denis and John J. McConnell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Dividends represent the primary means by which invested capital is returned to common stockholders. In this article we summarize the development of academic thinking on dividend policy, focusing on three primary perspectives: (*a*) the effect of dividend policy on common stock value and firm performance, (*b*) the determinants of dividend policy, and (*c*) macroeconomic trends in the propensity of firms to pay dividends.

Keywords

agency costs; asymmetric information; capital gains; capital gains taxation; dividend change; dividend policy; dividend taxation; dividends; manager–shareholder conflict; Modigliani–Miller theorem; regular vs special dividends; stock repurchases; stockholders

Article

There are two major ways in which a firm can distribute cash to its common stockholders. The firm can either declare a cash dividend which it pays to all its common stockholders or it can repurchase shares. Stock repurchases may take the form of registered tender offers, open market purchases, or negotiated repurchases from a large shareholder. In a share repurchase, shareholders may choose not to participate. In contrast, dividends are direct cash payments to shareholders and are distributed on a pro rata basis to all shareholders.

Most firms pay cash dividends on a quarterly basis. The dividend is declared by the firm's board of directors on a date known as the ‘announcement date’. The board's announcement states that a cash payment will be made to stockholders who are registered owners on a given ‘record date.’ The dividend checks are mailed to stockholders on the ‘payment date,’ which is usually about two weeks after the record date. Stock exchange rules generally dictate that the stock is bought or sold with the dividend until the ‘ex-dividend date’, which is a few business days before the record date. After the ex-dividend

date, the stock is bought and sold without the dividend.
Dividends may be either labelled or unlabelled. Most dividends are not given labels by management. Unlabelled dividends are commonly referred to as ‘regular dividends’. When managers label a dividend, the most common label is ‘extra’.

A historical perspective

Prior to 1961, academic treatments of dividends were primarily descriptive, as, for example, in Dewing (1953). To the extent that economists considered corporate dividend policy, the commonly held view was that investors preferred high dividend payouts to low payouts (see, for example, Graham and Dodd, 1951). The only question was how much value was attached to dividends relative to capital gains in valuing a security (Gordon, 1959). This view was concisely summarized with the saying that a dividend in the hand is worth two (or some multiple) of those in the bush. The only question was: what is the multiple?
In 1961, scientific inquiry into the motives and consequences of corporate dividend policy shifted dramatically with the publication of a classic paper by Miller and Modigliani. Perhaps the most significant contribution of the Miller and Modigliani paper was to spell out in careful detail the assumptions under which their analysis was to be conducted. The most important of these include the assumptions that the firm's investment policy is fixed and known by investors, that there are no taxes on dividends or capital gains, that individuals can costlessly buy and sell securities, that all investors have the same information, and that investors have the same information as the managers of the firm. With this set of assumptions, Miller and Modigliani demonstrate that a firm's stockholders are indifferent among the set of feasible dividend policies. That is, the value of the firm is independent of the dividend policy adopted by management.
Because investment policy is fixed in the Miller–Modigliani set-up, all feasible dividend policies involve the distribution of the full present value of the firm's free cash flow (that is, cash flow in excess of that required for investment) and are, therefore, equally valuable. If internally generated funds exceed required investment, the excess must be paid out as a dividend so as to hold investment constant. If internally generated funds are insufficient to fund the fixed level of investment, new shares must be sold. It is also possible for managers to finance a higher dividend with the sale of new shares.
The key insight from the Miller–Modigliani analysis is that investors will be indifferent among the feasible dividend choices because they can costlessly create their own dividend stream by buying and selling shares. If investors demand higher dividends than the amount paid by the firm, they can sell shares and consume the proceeds, leaving themselves in the same position as if the firm had paid a dividend. Alternatively, if shareholders prefer to reinvest rather than to consume, they can choose to purchase new shares with any dividends paid. In this instance, shareholders would be in the same position that that they would have been in had no dividends been paid. Thus, regardless of corporate dividend policy, investors can costlessly create their own dividend position. For this reason, stockholders are indifferent to corporate dividend policy, and, as a consequence, the value of the firm is independent of its dividend policy.
After a brief flurry of debate, the Miller–Modigliani irrelevance proposition was essentially universally accepted as correct under their set of assumptions. There nevertheless remained an underlying notion

that dividend policy must ‘matter’ given that managers and security analysts spend time worrying about it. If so, and if the Miller–Modigliani proposition is accepted, it must be due to violation of one or more of the Miller–Modigliani assumptions in the real world.

Since the early 1960s, the dividend debate has been lively and interesting. Economists have analysed theoretically whether the relaxation of the various Miller–Modigliani assumptions alters their irrelevance proposition. In addition, economists have analysed the data from several perspectives. First, they have undertaken an array of analyses to determine the effect, if any, of dividend policy on stock value and firm performance. Second, they have sought to identify the characteristics associated with dividend payments (or the lack thereof) by individual firms. Third, they have attempted to characterize macroeconomic trends in the level and propensity of firms to pay dividends, and in the form of the payout. Our discussion of these issues focuses primarily (though not exclusively) on studies of US firms since these are the studies most accessible to us.

Relaxing the Miller – Modigliani assumptions

Taxes

Perhaps the obvious starting point for an investigation into the effect of relaxing the Miller–Modigliani assumptions is to introduce taxes. In the United States, dividend payments by a corporation do not affect that firm's taxes. However, at least historically, dividends have been taxed at a higher rate than capital gains at the personal level. Thus, superficially, the US tax code appears to favour a low dividend payout policy, with payouts occurring primarily through share repurchases.

Under the assumption that dividends and capital gains are taxed differentially, Brennan (1970) derives a model of stock valuation in which stocks with high payouts have higher required before-tax returns than stocks with low payouts. As a counterpoint to this proposition, Miller and Scholes (1978) argue that under the US tax code there exist sufficient loopholes so that investors can shelter dividend income from taxation, thereby driving the effective tax rate on dividends to zero. Early studies of the association between stock returns and dividend yield (for example, Black and Scholes, 1974; Litzenberger and Ramaswamy, 1979; Miller and Scholes, 1982) yielded mixed results using different definitions of dividend yield. Subsequent studies indicated that the correlation between dividend yield and stock returns (if any) appeared to be due to omitted risk factors that were correlated with dividend yield. For example, Chen, Grundy and Stambaugh (1990) report that dividend yield and risk measures are cross-sectionally correlated. Similarly, Fama and French (1993) show that, when a three-factor model for expected returns is used, there is no significant relation between dividend yields and stock returns.

Other studies have analysed the potential effects of the differential taxation of dividends and capital gains by studying the behaviour of stock prices and trading volume around ex-dividend days. The logic of these studies is that, in order for investors to be indifferent between selling a stock just before it goes ex dividend and just after, stocks should be priced so that the marginal tax liability would be the same for each strategy. Thus, if dividends are taxed more heavily than are capital gains, stock prices should fall by less than the size of the dividend on the ex-dividend day. Evidence consistent with a tax effect in stock price behaviour around ex-dividend days is provided in Elton and Gruber (1970), Eades, Hess and Kim (1984), Green and Rydqvist (1999), Bell and Jenkinson (2002), and Elton, Gruber and Blake

(2005). In addition, evidence of tax-motivated trading around ex-dividend days is provided in Lakonishok and Vermaelen (1986), Michaely and Vila (1995) and Green and Rydqvist (1999). Collectively, the evidence in these studies indicates that the differential taxation of dividends and capital gains affects both ex-dividend day stock returns and trading activity. This conclusion has been reinforced in studies that examine changes in tax laws (for example, Poterba and Summers, 1984; Barclay, 1987; Michaely, 1991). Nonetheless, the fact that individual investors in high tax brackets receive large amounts of taxable dividends each year (Allen and Michaely, 2003) casts doubt on taxes being a first-order determinant of dividend policy.

Agency costs

A second real-world violation of the Miller–Modigliani assumptions is the existence of agency costs associated with stock ownership. In particular, managers of firms maximize their own utility, which is not necessarily the same as maximizing the market value of common stock. The costs associated with this potential conflict of interest include expenditures for structuring monitoring and bonding contracts between shareholders and managers, and residual losses due to imperfectly constructed contracts (Jensen and Meckling, 1976).

Several authors have argued that dividends may be important in helping to resolve manager–shareholder conflicts. If dividend payments reduce agency costs, firms may pay dividends even if these payments are taxed disadvantageously.

Easterbrook (1984) and Rozeff (1982) argue that establishing a policy of paying dividends enables managers to be evaluated periodically by the capital market. By paying dividends, managers are required to tap the capital market more frequently to obtain funds for investment projects. Periodic review by the market is one way in which agency costs are reduced, which in turn raises the value of the firm.

Similarly, Jensen (1986) argues that establishing a policy of paying dividends reduces agency problems of overinvestment by reducing the amount of discretionary cash controlled by managers.

An implication of the agency models is that dividends will be more valuable in mature firms with substantial cash flow and poor investment opportunities. Early tests of this implication focused on the stock price reaction to dividend change announcements and produced mixed results. Lang and Litzenberger (1989) find that firms with less valuable growth opportunities exhibit a larger stock price reaction to dividend increase announcements than firms with more valuable growth opportunities. Although this finding is consistent with the agency cost hypothesis, Denis, Denis and Sarin (1994) find that when they control for other factors, particularly the change in dividend yield, they find no difference in the stock price reaction to dividend changes between firms with good growth opportunities and those with poor growth opportunities. Moreover, they find no evidence that increases in dividends reduce corporate investment.

More recent tests of the agency models have focused on the cross-sectional determinants of dividend policy. Fama and French (2001) find that the propensity to pay dividends is positively related to firm size and profitability, and negatively related to the value of future growth opportunities. DeAngelo, DeAngelo and Stulz (2006) find that the propensity to pay dividends is strongly associated with the proportion of the firm's equity that comes from retained earnings. These findings support the primary

prediction of the agency models that dividends are more valuable for mature firms with high cash flow and poor growth opportunities.

La Porta et al. (2000) and Faccio and Lang (2002) provide further support for the agency models of dividend policy by analysing international evidence. La Porta et al. hypothesize that agency conflicts will differ across countries because of differences in the extent of investor protection. In a sample of 33 different countries, they find that dividend payments are higher in countries with better investor protection. This indicates that when investors are better able to monitor managers, they are able to force higher dividend payouts. Faccio and Lang (2002) show that in western Europe and in Asia dividend payments are higher when controlling shareholders have a higher ratio of voting rights to cash flow rights – that is, those situations in which minority shareholders are otherwise at greatest risk of expropriation by the controlling shareholder.

Asymmetric information

Contrary to the Miller–Modigliani assumption that investors have the same information as managers, a large number of studies assume that managers possess more information about the prospects of the firm than individuals outside the firm, and that dividend changes convey this information to outsiders. This idea was suggested by Miller and Modigliani and has roots in Lintner's (1956) classic study on dividend policy. Lintner interviewed a sample of corporate managers. One of the primary findings of the interviews is that a high proportion of managers attempt to maintain a stable regular dividend. In Lintner's words, managers demonstrate a ‘reluctance (common to all companies) to reduce regular rates once established and a consequent conservatism in raising regular rates’ (1956, p. 84).

If managers change regular dividends only when the earnings potential of the firm has changed, changes in regular dividends are likely to provide some information to the market about the firm's prospects. More formal models in which dividends convey information to outsiders include Bhattacharya (1979; 1980), John and Williams (1985), and Miller and Rock (1985). The common assumption in these models is that managers have information not available to outside investors. Typically, the information has to do with the current or future earnings of the firm.

Empirical evidence on the information content of dividends has taken three forms. First, a large set of studies has analysed whether dividend changes are associated with abnormal stock returns of the same sign. Second, studies have analysed whether dividend changes are associated with subsequent earnings changes. Third, studies have analysed the association between dividend changes and changes in investor expectations regarding future earnings.

Studies have consistently documented that stock returns around the announcement of a dividend change are positively correlated with the change in the dividend (Aharony and Swary, 1980; Asquith and Mullins, 1983; Brickley, 1983; Healy and Palepu, 1988; Grullon, Michaely and Swaminathan, 2002; Michaely, Thaler and Womack, 1995; Pettit, 1972). These studies are robust over time and are robust to controls for contemporaneous earnings announcements. Moreover, in general, the studies indicate that the market reacts more strongly to a dividend decrease than to a dividend increase.

The findings described above indicate that dividend announcements provide information to the market. Subsequent studies have investigated whether this information is correlated with current or future earnings. On this issue, the evidence is more mixed. In a study of dividend initiations and omissions,

Healy and Palepu (1988) find that the initiation of dividends follows a period of abnormal earnings growth and that earnings continue to grow in subsequent years. For omissions, however, earnings decline in the year of omission, then rebound in the following years. Using a comprehensive sample of dividend changes, Benartzi, Michaely and Thaler (1997) find no evidence that dividend changes are associated with subsequent earnings changes of the same sign. Miller's interpretation of the evidence (1987) is that dividends appear to be better described as lagging earnings than as leading earnings.

One difficulty in testing whether dividend changes ‘signal’ unexpected future earnings is that it is difficult to identify what level of earnings would be expected by the market if the dividend change did not take place. To address this issue, Ofer and Siegel (1987) study how analysts alter their estimates of current year earnings when firms announce dividend changes. They find that analysts revise their earnings estimates in the direction of the dividend change and that the size of the earnings revision is positively associated with the stock price reaction to the dividend change. Similarly, Fama and French (1998) report a positive association between dividends and firm value after controlling for past, current and future earnings, as well as investment and debt. They conclude that dividends contain information about value that is not contained in earnings, investment and debt.

The accumulated empirical evidence thus indicates that dividend announcements provide information to the market. Whether they convey information about future earnings is less clear. Moreover, other findings indicate that information signalling is unlikely to be a first-order determinant of dividend policy. For example, as noted earlier, dividends are paid primarily by larger, more mature firms with higher cash flow and poorer growth opportunities. These types of firm would seem to be least in need of signalling their true value to the market.

Firm value and the form of the payout

As with increases in regular cash dividends, specially labelled cash dividends and share repurchases have been shown to be accompanied by permanent increases in stock prices (Brickley, 1983; Dann, 1981; Vermaelen, 1981). However, there is little agreement on the factors that lead managers to choose one method over another.

Given the Miller–Modigliani assumptions, the choice of the payout mechanism, like the choice of dividend policy itself, does not affect the value of the firm. Therefore, if the form of the payout is to matter, it must be due to violation of one or more of the Miller–Modigliani assumptions. To develop a theory to explain the choice of payout mechanism, it must be that there are differential costs or benefits associated with the alternative payout methods. Furthermore, the relative benefits or costs must be especially significant because, in general, dividends have been tax-disadvantaged (at the personal level) relative to share repurchases.

Economists have explored several possible explanations as to why a particular form of payout is chosen, including adverse selection effects (Barclay and Smith, 1988; Miller and McConnell, 1995), the impact on equity ownership structure (Stulz, 1988; Denis, 1990), the signalling power of alternative payout mechanisms (Ofer and Thakor, 1987; Jagannathan, Stephens and Weisbach, 2000), and the impact of executive stock options (Fenn and Liang, 2001). The evidence indicates that share repurchases are more likely when recent earnings increases are temporary, when earnings are riskier, when firms make heavy

use of stock options in executive compensation contracts and when firms seek to protect themselves from a hostile takeover.

As regards the choice between regular cash dividends and specially labelled cash dividends, reasonable explanations have been relatively scarce. Brickley (1983) does provide evidence that specially labelled dividends convey a less positive message about firm value than do increases in regular cash dividends. Nonetheless, it is unclear why this is so. Moreover, there has been little examination of the choice between special dividends and share repurchases.

What managers say

Lintner's (1956) classic empirical study began with a survey of corporate executives. The results of that survey and the accompanying evidence laid the foundation for much of the empirical and theoretical work that has followed over the succeeding half century. Brav et al. (2005) have conducted a new and more extensive survey of chief financial officers (CFOs) regarding their views of corporate payout policy. Their survey yields further insights into what managers think about dividend policy, and complements the existing empirical evidence.

Brav et al. report that CFOs view dividends as inflexible in that, once a dividend level has been established, any dividend cut is likely to have a significantly adverse impact on the company's stock price. Thus, consistent with Lintner's (1956) original observation, managers tend to be conservative when adjusting dividends upward in order to avoid having to cut the dividend at a later date. Rather than establishing a target payout ratio, managers set a per share payment that is downwardly inflexible. According to the survey, managers do not explicitly view dividends as a mechanism for signalling information that would distinguish their companies from competitors, and they consider tax effects only as an afterthought. These observations accord with the conclusions drawn from empirical studies in that both imply that taxes and signalling are not first-order determinants of dividend policy.

In contrast to dividends, repurchases are viewed by managers as a parallel but more flexible way to distribute cash to shareholders in that they can be initiated and discontinued as funds are available. This observation is consistent with the empirical evidence cited earlier that repurchases tend to be associated with temporary increases in earnings, while dividends are associated with earnings changes that are more permanent. Whether the modern survey of Brav et al. leads to the volume of additional empirical work that followed Lintner's study remains to be seen.

Summary and recent trends

Since the mid-1960s, rigorous consideration has added considerably to progress in what is known about dividend policy. We know that firms pay out to stockholders substantial amounts of cash annually in the form of regular cash dividends, share repurchases and specially labelled dividends. We also know that stock prices increase permanently when regular dividends are increased, when special dividends are declared, and when shares are repurchased, and that stock prices decline when regular dividends are reduced. While these findings imply that dividend changes reflect information available to managers that is not otherwise available to outside investors, it is still not clear what information is being conveyed through the dividend payment. Moreover, although we now know a considerable amount about the

empirical determinants of the size of payout and the form of payout, there is little agreement as to whether the level of cash payout affects the value of the firm or and whether the choice of the payout method matters.

We conclude by outlining several recent trends that pose additional challenges to our understanding of dividend policy. First, Fama and French (2001) document that the propensity to pay dividends has declined substantially since the late-1970s. Second, despite this decline in the propensity to pay dividends, aggregate dividends have not declined (DeAngelo, DeAngelo and Skinner, 2004). Rather, dividends and earnings have become increasingly concentrated among larger firms. Third, specially labelled dividends have nearly disappeared (DeAngelo, DeAngelo and Skinner, 2000). Fourth, share repurchases have increased substantially so that aggregate payouts through share repurchases now exceed those through regular dividends (Grullon and Michaely, 2002). These trends are difficult to explain given our current understanding of dividend policy. Undoubtedly, therefore, economists will continue to devote substantial effort to understanding the puzzles of dividend policy.

See Also

- finance (new developments)
- Modigliani–Miller theorem

Bibliography

Aharony, J. and Swary, I. 1980. Quarterly dividend and earnings announcements and stockholders' returns: an empirical analysis. *Journal of Finance* 35, 1–12.

Allen, F. and Michaely, R. 2003. Payout policy. In *Handbook of the Economics of Finance: Volume 1a*, ed. G. Constantinides, M. Harris and R. Stulz. Amsterdam: North-Holland.

Asquith, P. and Mullins, D. 1983. The impact of initiating dividend payments on shareholders' wealth. *Journal of Business* 56, 77–96.

Barclay, M. 1987. Dividends, taxes, and common stock prices: the ex-dividend day behavior of common stock prices before the income tax. *Journal of Financial Economics* 14, 31–44.

Barclay, M. and Smith, C. 1988. Corporate payout policy: cash dividends versus open-market repurchases. *Journal of Financial Economics* 22, 61–82.

Bell, L. and Jenkinson, T. 2002. New evidence of the impact of dividend taxation and on the identity of the marginal investor. *Journal of Finance* 57, 1321–46.

Benartzi, S., Michaely, R. and Thaler, R. 1997. Do changes in dividends signal the future or the past? *Journal of Finance* 52, 1007–43.

Bhattacharya, S. 1979. Imperfect information, dividend policy, and ‘the bird in the hand’ fallacy. *Bell Journal of Economics* 10, 259–70.

Bhattacharya, S. 1980. Nondissipative signaling structures and dividend policy. *Quarterly Journal of Economics* 95, 1–24.

Black, F. and Scholes, M. 1974. The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics* 1, 1–22.

Brav, A., Graham, J., Michaely, R. and Harvey, C. 2005. Payout policy in the 21st century. *Journal of Financial Economics* 77, 483–527.

Brennan, M. 1970. Taxes, market valuation and financial policy. *National Tax Journal* 23, 417–29.

Brickley, J. 1983. Shareholders wealth, information signaling, and the specially designated dividend: an empirical study. *Journal of Financial Economics* 12, 187–209.

Chen, N., Grundy, B. and Stambaugh, R. 1990. Changing risk, changing risk premiums, and dividend yield effects. *Journal of Business* 63, S51–S70.

Dann, L. 1981. Common stock repurchases: an analysis of returns to bondholders and stockholders. *Journal of Financial Economics* 9, 113–38.

DeAngelo, H., DeAngelo, L. and Skinner, D. 2000. Special dividends and the evolution of dividend signaling. *Journal of Financial Economics* 57, 309–354.

DeAngelo, H., DeAngelo, L. and Skinner, D. 2004. Are dividends disappearing? Dividend concentration and the consolidation of earnings. *Journal of Financial Economics* 72, 425–56.

DeAngelo, H., DeAngelo, L. and Stulz, R. 2006. Dividend policy and the earned/contributed capital mix: a test of the life-cycle theory. *Journal of Financial Economics* 81, 227–54.

Denis, D. 1990. Defensive changes in corporate payout policy: share repurchases and special dividends. *Journal of Finance* 45, 1433–56.

Denis, D., Denis, D. and Sarin, A. 1994. The information content of dividend changes: cash flow signaling, overinvestment, and dividend clienteles. *Journal of Financial and Quantitative Analysis* 29, 567–87.

Dewing, A. 1953. *The Financial Policy of Corporations*, vol. 2, 5th edn. New York: Ronald Press Co.

Eades, K., Hess, P. and Kim, E. 1984. On interpreting security returns during the ex-dividend period. *Journal of Financial Economics* 13, 3–34.

Easterbrook, F. 1984. Two agency–cost explanations of dividends. *American Economic Review* 74, 650–9.

Elton, E. and Gruber, M. 1970. Marginal stockholders' tax rates and the clientele effect. *Review of Economics and Statistics* 52, 68–74.

Elton, E., Gruber, M. and Blake, C. 2005. Marginal stockholder tax effects and ex-dividend day behavior: evidence from taxable versus nontaxable closed-end funds *Review of Economics and Statistics* 87, 579–86.

Faccio, M. and Lang, L. 2002. The ultimate ownership of western European corporations. *Journal of Financial Economics* 65, 365–95.

Fama, E.F. and French, K. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 3–56.

Fama, E.F. and French, K. 1998. Taxes, financing decisions, and firm value. *Journal of Finance* 53, 819–43.

Fama, Eugene and French, K. 2001. Disappearing dividends: changing firm characteristics or lower propensity to pay? *Journal of Financial Economics* 60, 3–43.

Fenn, G. and Liang, N. 2001. Corporate payout policy and managerial stock incentives. *Journal of Financial Economics* 60, 45–72.

Gordon, M. 1959. Dividends, earnings and stock prices. *Review of Economics and Statistics* 41, 99–105.

Graham, B. and Dodd, D. 1951. *Security Analysis: Principles and Technique*. New York: McGraw-Hill.

Green, R. and Rydqvist, K. 1999. Ex-day behavior with dividend preference and limitation to short-term arbitrage: the case of Swedish lottery bonds. *Journal of Financial Economics* 53, 145–87.

Grullon, G. and Michaely, R. 2002. Dividends, share repurchases and the substitution hypothesis. *Journal of Finance* 57, 1649–84.

Grullon, G., Michaely, R. and Swaminathan, B. 2002. Are dividend changes a sign of firm maturity? *Journal of Business* 75, 387–424.

Healy, P. and Palepu, K. 1988. Earnings information conveyed by dividend initiations and omissions. *Journal of Financial Economics* 21, 149–76.

Jagannathan, M., Stephens, C. and Weisbach, M. 2000. Financial flexibility and the choice between dividends and stock repurchases. *Journal of Financial Economics* 57, 355–84.

Jensen, M. 1986. Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review* 76, 323–9.

Jensen, M. and Meckling, W. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

John, K. and Williams, J. 1985. Dividends, dilution, and taxes: a signaling equilibrium. *Journal of Finance* 40, 1053–70.

La Porta, R., Lopez-De Silanes, F., Shleifer, A. and Vishny, R. 2000. Agency problems and dividend policy around the world. *Journal of Finance* 55, 1–33.

Lakonishok, J. and Vermaelen, T. 1986. Tax induced trading around ex-dividend dates. *Journal of Financial Economics* 16, 287–319.

Lang, L. and Litzenberger, R. 1989. Dividend announcements: cash flow signaling vs. free cash flow hypothesis. *Journal of Financial Economics* 24, 181–92.

Lintner, J. 1956. Distribution of incomes of corporations among dividends, retained earnings, and taxes. *American Economic Review* 46, 97–113.

Litzenberger, R. and Ramaswamy, K. 1979. The effects of personal taxes and dividends on capital asset prices: theory and empirical evidence. *Journal of Financial Economics* 7, 163–95.

Michaely, R. 1991. Ex-dividend day stock price behavior: the case of the 1986 tax reform act. *Journal of Finance* 46, 845–60.

Michaely, R. and Vila, J. 1995. Investors' heterogeneity, prices and volume around the ex-dividend day. *Journal of Financial and Quantitative Analysis* 30, 171–98.

Michaely, R., Thaler, R. and Womack, K. 1995. Price reactions to dividend initiations and omissions: overreaction or drift? *Journal of Finance* 50, 573–608.

Miller, J. and McConnell, J. 1995. Open-market share repurchase programs and bid-ask spreads on the NYSE: implications for corporate payout policy. *Journal of Financial and Quantitative Analysis* 30,

365–82.

Miller, M. 1987. The information content of dividends. In *Macroeconomics: Essays in Honor of Franco Modigliani*, ed. J. Bossons, R. Dornbush and S. Fischer. Cambridge MA: MIT Press.

Miller, M. and Modigliani, F. 1961. Dividend policy, growth and the valuation of shares. *Journal of Business* 34, 411–33.

Miller, M. and Rock, K. 1985. Dividend policy under asymmetric information. *Journal of Finance* 40, 1031–51.

Miller, M. and Scholes, M. 1978. Dividends and taxes. *Journal of Financial Economics* 6, 333–64.

Miller, M. and Scholes, M. 1982. Dividends and taxes: empirical evidence. *Journal of Political Economy* 90, 1118–41.

Ofer, A. and Siegel, D. 1987. Corporate financial policy, information, and market expectations: an empirical investigation of dividends. *Journal of Finance* 42, 889–911.

Ofer, A. and Thakor, A. 1987. A theory of stock price responses to alternative corporate cash disbursement methods: stock repurchases and dividends. *Journal of Finance* 42, 365–94.

Pettit, R. 1972. Dividend announcements, security performance, and capital market efficiency. *Journal of Finance* 27, 993–1007.

Poterba, J. and Summers, L. 1984. New evidence that taxes affect the valuation of dividends. *Journal of Finance* 39, 1397–415.

Rozeff, M. 1982. Growth, beta and agency costs as determinants of dividend payout ratios. *Journal of Financial Research* 5, 249–59.

Stulz, R. 1988. Managerial control of voting rights: financing policies and the market for corporate control. *Journal of Financial Economics* 20, 25–54.

Vermaelen, T. 1981. Common stock repurchases and market signaling: an empirical study. *Journal of Financial Economics* 9, 139–83.

How to cite this article

Denis, David J. and John J. McConnell. "dividend policy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000171> doi:10.1057/9780230226203.0398

The New Palgrave Dictionary of Economics Online

Divisia index

Charles R. Hulten

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Divisia index, it its modern application, is a continuous-time index related to an underlying economic structure via a potential function. Under certain conditions, the index can retrieve important characteristics of the underlying structure using prices and quantities alone, without full knowledge about the structure itself. The Divisia index is widely used in theoretical discussions of productivity analysis, and has important applications elsewhere. In practice, it is approximated by discrete-time superlative indexes, like the Tornqvist, or by chain indexes. Older applications of the Divisia stressed its discrete-time axiomatic properties.

Keywords

aggregation; chain indexes; continuous-time indexes; discrete-time indexes; Divisia index; Divisia, F.; duality; path dependence; production functions; productivity (measurement problems); Solow, R.; Törnqvist index

Article

The Divisia index is a continuous-time index number formula due to François Divisia (1925–6) that has been widely used in theoretical discussions of data aggregation and the measurement of technical change. It is defined with respect to the time paths of a set of prices $[P_1(t), \ldots, P_N(t)]$ and commodities $[X_1(t), \ldots, X_N(t)]$. Total expenditure on this group of commodities is given by:

$$Y(t) = P_t(t)X_1(t) + \ldots + P_N(t)X_N(t). \quad (1)$$

With dots over variables indicating derivatives with respect to time, total differentiation of (1) yields:

$$\frac{\dot{Y}(t)}{Y(t)} = \sum_{i=1}^{i=N} \frac{P_i(t)X_i(t)}{Y(t)} \frac{\dot{P}_i(t)}{P_i(t)} + \sum_{i=1}^{i=N} \frac{P_i(t)X_i(t)}{Y(t)} \frac{\dot{X}_i(t)}{X_i(t)}. \quad (2)$$

The growth rates of the Divisia price and quantity indexes are the respective weighted averages of the growth rates of the individual $P_i\bullet(t)$ and $X_i\bullet(t)$, where the weights are the components' shares in total expenditure. The levels of these indexes are obtained by line integration over the trajectory followed by the individual prices and quantities over the time interval [0, $T\bullet$]. For the quantity index, the line integral has the following form:

$$I_q(0, T) = \exp\left\{\int\left[\sum_{i=1}^{N} \frac{P_i(t)X_i(t)}{\Sigma P_j(t)X_j(t)} \frac{\dot{X}_i(t)}{X_i(t)}\right]\right\} = \exp\left\{\int_r \varphi(X)\mathrm{d}X\right\}, \quad (3)$$

where ϕ is a vector-valued function whose arguments are $P_i\bullet(t)/Y\bullet(t)$, prices are assumed to be a function of the X_i, and Γ is the curve described by X_i. A similar expression characterizes the Divisia price index (for a more extensive discussion of Divisia line integrals, see Richter, 1966; Hulten, 1973; Samuelson and Swamy, 1974).

The value of the index defined by (3) depends on the solution of the line integral. This can be obtained by identifying a 'potential function' Φ whose partial derivatives are the vector-valued function ϕ, that is, $\varphi = \nabla\Phi$. Writing $\Phi = \log F$ function, the value of the index can be shown to equal $F[X(T)]/F[X(0)]$, implying that the index is unique only up to a scalar multiple.

In economic terms, the solution to (3) is associated with some underlying economic relationship among the variables being indexed. Assume, for example, there is a constant returns to scale production function F(X) and $F_i = \lambda P_i$ (Fi denotes the partial derivative of F with respect to Xi and λ is a factor of proportionality). Then the function log F can serve as the requisite potential function for (3), and in this particular case, the Divisia index of inputs can be interpreted as the ratio of output at time T to output at time zero.

If the form of the potential function is known a priori, the value of the index could be computed directly from the function F. However, the rationale for the Divisia index is that it provides a way of obtaining the ratio $F(X(T))/F(X(0))$ by using data on prices and quantities alone, without direct knowledge of F. Intuitively, this is possible because, under sufficiently restrictive assumptions, information about the

slope of the function F (as estimated by relative prices) over the path followed by the inputs is sufficient to characterize F up to a scalar multiple.

When the objective is to form an index of a subset of inputs – aggregate labour input, for example – the required potential function is a ‘piece’ of a production function. Specifically, if one wants to form a Divisia index of the first M inputs, the production function needs to be weakly separable into a function of these inputs, that is, $F\{G[X_1(t), \ldots, X_M(t)], X_{M+1}(t), \ldots, X_N(t)\}$. The function log *G* serves as the potential function for the line integration (see also Balk, 2005).

These considerations apply to Divisia price indexes as well. The relevant potential function is now the factor price frontier $\Psi[P_1(t), \ldots, P_N(t)]$. A basic result of duality theory shows that the partial derivatives of Ψ are proportional to the corresponding $X_i(t)$.

The discussion suggests that the existence of the Divisia index is closely linked to the conditions for consistent aggregation. Furthermore, the required existence of a potential function implies that aggregation cannot proceed with just any set of prices or quantities. There must be an a priori reason for supposing that the variables to be indexed are theoretically related. This is an important characteristic of the Divisia index, one which it shares with the broader class of economic index numbers (in contrast to the non-structural axiomatic approach associated with Irving Fisher, 1921; see also Balk, 2005). The potential function theorem establishes the conditions under which the Divisia index is an ‘exact’ index number (to use the terminology of Diewert, 1976) for some underlying economic structure.

Divisia indexes have the desirable property that they are invariant when the path of integration lies entirely in the same level set of the potential function. That is, if one input is substituted for another along a given isoquant, the value of the index will not change. However, there is no guarantee of invariance when the path of integration lies across several level sets. This reflects the mathematical property that line integrals are, in general, path dependent.

Path dependence means that the index (3) will generally have a different value for a path $\beta(t) \in \Gamma_1$ than path $\alpha(t) \in \Gamma$, even though the beginning and end points of Γ_1 and Γ are identical. This can lead to the following situation: the economy moves along Γ_1 from X from X $'$ (which is on a different isoquant); the economy then returns along Γ to the original point X; because of path dependence, the vector of quantities represented by the vector X will have a different Divisia index value after the trip around the composite path, and subsequent circuits will produce still different values. The value of the Divisia index at any point X is thus arbitrary under path dependence. The uniqueness of the Divisia index thus involves path independence.

The condition for path independence is the existence of a homothetic potential function, log *F*, such that $\varphi = \nabla \log F$, where ϕ is defined in (3). Given the existence of the potential function, the value of (3) is $F(X(T))/F(X(0))$, implying path independence since (3) depends only on the end points of the path, $X(0)$ and $X(T\bullet)$. Conversely, if (3) is path independent, there exists a potential function log *F* such that $\nabla \log F = \varphi$. In some applications in productivity analysis, the homotheticity condition must be strengthened to linear homogeneity, but this can be weakened depending on data availability (Hulten, 2001, pp. 11–12).

We note, finally, that the Divisia index is defined using time as a continuous variable. Data on prices and quantities typically refer to discrete points in time, and the indexes constructed from them must therefore have a discrete-time form. The continuous-time Divisia index is nevertheless useful, both for informing

the structure of these discrete-time indexes (for example, for the determining which variables are conceptually related), and for interpreting the results. The Divisia framework is also appropriate for the theoretical analysis of many economic problems, such as the use of Divisia indexes by Solow (1957) in growth accounting.

One approach to linking discrete and continuous index numbers is to approximate the continuous variables of (2) with their discrete time counterparts. Under the Törnqvist (1936) approach, the growth rates of prices and quantities are approximated by logarithmic differences, and the continuous weights by two period arithmetic averages. The Törnqvist approximation to the growth rate of the Divisia quantity index can then be written:

$$\sum_{i=1}^{i=T} \frac{1}{2}\left[\frac{P_{i,t}X_{i,t}}{Y_i} + \frac{P_{i,t-1}X_{i,t-1}}{Y_{t-1}}\right][\log X_{i,t} - \log X_{i,t-1}] \tag{4}$$

A similar approximation applies to the growth rate of the Divisia index of prices.

While the Törnqvist index may be regarded as approximate, Diewert (1976) has shown that it is exact when the underlying potential function has the (continuous) translog form. This result is very important in its own right, but can also be regarded as an important conceptual link between the discrete and continuous-time families of index numbers, given the exact properties of the Divisia index in continuous time.

The continuous Divisia index can also be approximated by using chain indexing procedures (the Divisia index is sometimes regarded as a chain whose links are defined over infinitesimal time periods). Other numerical approximation techniques can also be employed.

See Also

- Divisia, François Jean Marie
- index numbers

Bibliography

Balk, B. 2005. Divisia price and quantity indices: 80 years after. *Statistica Neerlandica* 59, 119–58.

Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4(2), May, 115–45.

Divisia, F. 1925–6. L'indice monétaire et la théorie de la monnaie. *Revue d'Economie Politique* 39(4), 842–64; (5), 980–1008; (6), 1121–51; 40(1), 49–81. Also separately: Paris: Société Anonyme du Recueil Sirey, 1926.

Fisher, I. 1921. *The Making of Index Numbers*. Boston, MA: Houghton Mifflin Co.

Hulten, C. 1973. Divisia index numbers. *Econometrica* 41, 1017–25.

Hulten, C. 2001. Total factor productivity: a short biography. In *New Developments in Productivity Analysis*, ed. C. Hulten, E. Dean and M. Harper. Studies in Income and Wealth, vol. 63. Chicago: University of Chicago Press for the NBER.

Richter, M. 1966. Invariance axioms and economic indexes. *Econometrica* 34, 739–55.

Samuelson, P. and Swamy, S. 1974. Invariant economic index numbers and canonical duality: survey and synthesis. *American Economic Review* 64, 566–93.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

Törnqvist, L. 1936. The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin* 10, 1–8.

How to cite this article

Hulten, Charles R. "Divisia index." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000174> doi:10.1057/9780230226203.0400

The New Palgrave Dictionary of Economics Online

Divisia, François Jean Marie (1889 – 1964)

David E.R. Gay
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Statistical Association; Divisia Index; Econometric Society; International Econometric Society.; mathematical economics; statistics and economics

Article

Divisia was born in Tizi-Ouzou, Algeria. He received baccalaureate degrees in mathematics and philosophy at Algiers. After two years in the Ecole Polytechnique he worked for the government as a civil engineer (Ponts et Chaussées). His graduate engineering work at the Ecole Nationale des Ponts et Chaussées was completed in 1919 after the interruption of the First World War. After nearly ten years as a government engineer he joined the ministry of national education to continue research and teaching economics. He became a professor of applied economics at the Ecole Nationale des Ponts et Chaussées (1932–50), the Conservatoire National des Arts et Métiers (1929–59), and the Ecole Polytechnic (1929–59). He was a founding member of the Econometric Society and its president in 1935. Subsequently he was also president of the Paris Statistics Society (1939) and of the International Econometric Society. He was a Fellow of the American Statistical Association and of the American Association for the Advancement of Science.

His major contributions to economics can be found centred in several books on economics and applied statistics. The Divisia Index, a variable-weight price index, was developed in *L'indice monétaire et la théorie de la monnaie* (1926). His *Economique rationnelle* (1928) was widely acclaimed in mathematical economics and was awarded prizes by the Academy of Sciences and by the Academy of Moral Sciences and Politics. Using a microeconomic perspective he cautioned against uncritical acceptance of macroeconomic research in *Traitement économétrique de la monnaie, l'intérêt, l'emploi* (1962).

Selected works

1926. *L'indice monétaire et la théorie de la monnaie*. Paris.

1928. *Economique rationnelle*. Paris.

1931. *L'epargne et la richesse collective*. Paris.

1951–65. *Exposés d'économique*. Paris.

1962. *Traitement économétrique de la monnaie, l'intérêt, l'emploi*. Paris.

How to cite this article

Gay, David E. R. "Divisia, François Jean Marie (1889–1964)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000175>
doi:10.1057/9780230226203.0399

The New Palgrave Dictionary of Economics Online

division of labour

Peter Groenewegen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Division of labour has been a very important topic for economic writings from the earliest times, and was treated in great detail by major economists, including especially Adam Smith and Alfred Marshall. This article surveys the development of ‘division of labour’ from its beginnings in the writings of Greek philosophers through the centuries and up to the 21st century. It therefore also reflects on its offshoots: international division of labour, sexual division of labour and its contemporary revival as an essential adjunct to the theory of economic growth, labour productivity, inter-firm cooperation, and its modern limits in coordination and communication costs.

Keywords

agricultural productivity; Aristotle; Austrian capital theory; Babbage, C.; Bagehot, W.; Beccaria, C. B.; British Association; capital accumulation; capitalism; Carl, E. L.; Carlyle, T.; clustering; communication and coordination costs; Comte, A.; concentration; d'Alembert, C.; Dexterity; Diderot, D.; division of labour; Durkheim, E.; economies of scale; Engels, F.; engineering; equilibrium analysis; Ferguson. A.; firm, theory of; foreign direct investment; Hegel, G. W. F.; Hodgskin, T.; housework; human capital; Hutcheson, F.; Ibn Khaldun; increasing returns; industrial organization; industrial psychology; industrial revolution; inequality; inter-firm cooperation; international division of labour; Kaldor, N.; labour productivity; Mandeville, B.; manufacturing; manufacturing division of labour; Marshall, A.; Marshall, M. P.; McCulloch, J. R.; new classical microeconomics; new growth economics; Nicholson, J. S.; Petty, W.; Plato; productivity growth; putting-out system; Quesnay, F.; Rae, J. (1796–1872); Robbins, L. C.; Ruskin, J.; segmented labour markets; Senior, N.; sexual division of labour; Sidgwick, H.; social division of labour; specialization; Spencer, H.; Taussig, F. W.; technical change; Torrens, R.; Tucker, J.; Turgot, A. R. J.; Ure, A.; Verdoorn's Law; women's work; Xenophon; Young, A. A.

Article

Division of labour, or specialization, may be defined as the division of a process or employment into parts, each of which is carried out by a separate person, or any system of production in which tasks are separated to enable specialization to occur. This includes the separation of employments and professions within society at large or *social division of labour* as well as the division of labour which takes place within the walls of a factory building or within the limits of a of a single industry, the *manufacturing division of labour*. Division of labour as a form of specialization can also be practiced by small firms which all contribute to the production of parts (inputs) for the manufacturing of a complex output, as in the case of aircraft production or sophisticated electronic equipment. This form of business organization requires excellent coordination and communication between its various parts to ensure continuous supply of the necessary parts for the manufacturer of the final output. It is a geographical form of division of labour, developed from the notion of clustering related firms in a particular area or industrial district (for a survey, see Dosi, 1988).

Division of labour and its consequences for productivity were analysed as early as the time of the Greek philosophers, including Plato, Aristotle and Xenophon. Early analysis of the manufacturing division of labour had to await industrial developments of the 17th and 18th centuries and underwent further qualitative change in the 19th, 20th and 21st centuries. Hence manufacturing and more detailed division of labour should not be seen as a simple continuum of the social division of labour. By the end of the Middle Ages, social division of labour was extensively practiced; manufacturing division of labour, generally speaking, came with the Industrial Revolution. Under modern capitalism, social division of labour remains largely a market influenced phenomenon but manufacturing division of labour is enforced by those who plan and control the manufacturing process. Furthermore, the one divides society: the other human activity within the workshop, or within an industry: labour generally enhances ‘the individual and the species, [a manufacturing division] of labour, when carried on without regard to human capabilities and needs, is a crime against the person and against humanity’ (Braverman, 1974, p. 73). Division of labour was first practiced within the household, a *sexual division of labour* between women's activities in or near the house, and those of men further afield. When applied to local specialization of industries both nationally and internationally, it has produced a variety of conceptions of the *territorial* or *international (global) division of labour*.

Adam Smith (1776) placed the division of labour at the forefront of his discussion of economic growth and progress. Neither in its social nor in its manufacturing forms did the idea originate with him. It retained a varying, but often very prominent, place in 19th-century writings (particularly those of Senior, Babbage, John Stuart Mill, Marx and Marshall). ‘About 1890, Schmoller, Semmel, Bücher, Durkheim and Maunier all wrote on religious and sociological aspects of specialization’ (Salz, 1934, p. 284). For much of the 20th century, division of labour and specialization virtually disappeared as a major topic from economic texts. Reasons for this varied. Some economists believed such discussions were more appropriate to technical handbooks of production engineering and factory management. Other writers wished to confine analysis of its effects to sociological studies assessing the general impact of division of labour on society. The return of economic growth as an important part of the economist's research programme from the 1950s onwards, and earlier the work of Young (1928), brought renewed interest in the division of labour in its wake, as did growing dissatisfaction with the narrow view confining economics to studying ‘the disposal of scarce commodities’ (Robbins, 1932, p. 38). Global organization of manufacturing made possible by improvements in transport and communication implies modern

adaptations of the division of labour which economists cannot ignore. An example is the formation of industrialized districts, first observed and analysed by Alfred Marshall (1890), to be rediscovered and adapted to the post-Second World War Italian situation by Becattini (for example, 1990; 2001) and his colleagues (for a survey, see Goodman and Pamford, 1989). The various dimensions of division of labour raised in these introductory paragraphs suggest that a broad-based treatment of the subject is warranted by featuring highlights within its continuous development.

The Greeks

Many of the major Greek philosophers discussed aspects of the division of labour in their writings. In Book 2 of the *Republic*, Plato stated the necessity for a division of labour or specialization in occupations for social well-being and the adequate satisfaction of primary wants linking the phenomenon with exchange, the requirements of ‘a market, and a currency as a medium of exchange’ (Plato, 380 bc, pp. 102–6). Aristotle, though very conscious of the social need for a division of labour, did not depart much from Plato's earlier discussion (see Bonar, 1893, p. 34). More importantly, Xenophon linked division of labour and specialization to great cities, because they provided a substantial demand for individual products while the subdivision of work raised the skill of individual workers. Extracts from the work of these Greek pioneers on the division of labour have been often reprinted (see, for example, Sun, 2005, chs 2–4). Knowledge of these Greek texts among Arabian Islamic scholars during the middle ages enabled them to produce sophisticated treatments of the division of labour. Examples are the writings of Islamic theologian, al-Ghazali (1058–1111) and, more importantly, the writings of fourteenth century Islamic philosopher and historian, Ibn Khaldun, whose *Muqaddima* contains a detailed account of the division of labour (Sun, 2005, pp. 7–8, ch. 5).

Subsequent pre-Smithian developments

Towards the end of the seventeenth century, English economic literature rediscovered the concept of the division of labour and began to analyse the more modern manufacturing forms, linking them to productivity growth, cost reduction, increased international competitiveness and associating its scope with the more extensive markets made possible through urbanization. For example, Petty's *Political Arithmetick* written in 1671 compared the benefits of division of labour in textile production with specialization in ship building:

> For as Cloth must be cheaper made, when one Cards, another Spins, another Weaves, another Draws, another Presses and Packs; than when all the Operations above-mentioned, were clumsily performed by the same hand; so those who command the Trade of Shipping [need] to build…a particular sort of Vessels for each particular Trade. (Petty, 1671, pp. 260–1)

Ten years later, in *Another Essay on Political Arithmetick Concerning the Growth of the City of London* (1683, p. 473), Petty showed that a major gain from a vast city like London came from the improvement

and growth of manufactures it encouraged:

> For in so vast a City *Manufacturers* will beget one another, and each *Manufacture* will be divided into as many parts as possible, whereby the Work of each *Artisan* will be simple and easy; As for example in the making of a *Watch*, if one Man shall make the *Wheels*, another the *Spring*, another shall Engrave the *Dial-plate*, and another shall make the *Cases*, then the *Watch* will be better and cheaper, than if the whole Work be put upon any one man.

In continuing this argument Petty also suggested that specialization benefits could be achieved from concentrating certain manufactures on a particular location, partly because of the savings in transport and communication costs such concentration entailed (Petty, 1683, pp. 471–2). The anonymous author of *Considerations on the East India Trade* (1701, pp. 590–2) illustrated productivity gains from the division of labour by examples drawn from cloth making, watch making and shipbuilding. He clearly indicated that sufficient demand and regular trade were a precondition for such improvements, which lowered manufacturing labour costs without the need to lower wages. During the 18th century, examples of authors aware of the benefits and preconditions for a division of labour become more common. Practical writers like Patrick Lindsay (1733), Richard Campbell (1747) and Joseph Harris (1757) tended to concentrate on manufacturing division of labour using examples from linen and pin production as well as from the familiar watch making. Those writing from the position of moral or political philosophy, like Mandeville (1729), Hutcheson (1755), Ferguson (1767) and Josiah Tucker (1755; 1774) concentrated more on aspects of the social division of labour.

Discussion of the division of labour was of course not confined to English economic literature. A treatise on wealth published in the 1720s by Ernst Ludwig Carl discussed the benefits of the division of labour, applying them also to demonstrate the gains from free trade through an international division of labour based on different climates, resource availability and locational advantages (cited in Hutchison 1988, pp. 161–2). Among the Physiocrats, Quesnay dealt briefly with the social aspects of the division of labour in his article ‘Natural Right’ (1765, p. 51). Turgot developed the subject more thoroughly, making it the starting point of his *Reflections*, subsequently associating it with the introduction of money, the extension of commerce and the accumulation of capital (1766, pp. 44–6, 64, 70). Earlier, Turgot (1751, pp. 242–3) had linked the spread of social division of labour to inequality, arguing that this particular consequence of inequality improved living standards for even the humblest members of society and made possible cultivation of the arts and sciences. Among the general principles with which Beccaria (1771, pp. 387–8) commenced the argument of his *Elementi*, the division of labour and its benefits in terms of increased skills and dexterity are clearly set out. Finally, it may be noted that the *Encyclopédie* of Diderot and d'Alembert in its article ‘Art’ discussed the essentials of the manufacturing division of labour, listing its consequences as improvements in skill, better quality products, saving of time and of materials, and ‘of making the time or the labour go further, whether by the invention of a new machine or the discovery of a more suitable method’. In its article on pins (‘Epingle’) their manufacture is described as being generally subdivided into eighteen separate operations and thereby a prime example of the manufacturing division of labour (see Cannan, 1929, pp. 94–5).

Adam Smith's treatment of the division of labour

Adam Smith's discussion of the division of labour deserves separate treatment not because of its ‘originality’ or ‘completeness of exposition’ (Cannan, 1929, p. 96) but because ‘nobody either before or after [him], ever thought of putting such a burden upon division of labour. With A. Smith, it is practically the only factor in economic progress’ (Schumpeter, 1954, p. 187). The first three chapters of the *Wealth of Nations* were devoted to its analysis because it provided one of the two causes explaining increases in per capita output by which Smith defined the wealth of the nation. Although therefore only one of two causes, the other being ‘the proportion between the number of those who are employed in useful labour, and that of those who are not so employed’ (Smith, 1776, p. 10), it is the dominant one. Smith seems to have believed that scope for substantial increases in the proportion of the labour force to productive activities was limited. Using the equation, $g = (k \cdot p / w) - 1$, developed by Hicks (1965, p. 38) to summarize the Smithian growth progress, if a change in *k*, the proportion of productive labour in the labour force, is more or less ruled out, a substantial growth rate (*g*), given the real wage (*w*), depends exclusively on rising productivity (*p*) through extensions of the division of labour. Smith's emphasis on the division of labour as a factor in growth via its enormous influence on productivity makes his treatment of the subject so novel. Surprisingly, this aspect of his contribution was taken up by few 19th-century writers and had to be largely rediscovered in the work of Young (1928) and Kaldor (1972) who reiterated dynamic aspects of the phenomenon Smith was analysing.

Even though it was the most frequently revised part of his economics (see Meek and Skinner, 1973), Smith's basic account of the division of labour contains a number of weaknesses. First, Smith failed to develop aspects of the manufacturing division of labour with which he ought to have been familiar. Marglin (1974) points out that Smith ignored organizational features from a division of labour taking place within the one building of relevance to some well-established industries like textiles and the manufacture of metal implements. These organizational features which Smith omitted were associated with growing labour discipline problems, wasting time and materials, inherent in the putting-out system, then the dominant form of manufacturing organization. In fact it can be suggested that if this aspect of the division of labour is more fully taken into account, its important role in explaining economic growth so much emphasized by Smith is more easily integrated as a major factor explaining the industrial revolution (see Groenewegen, 1977). Marglin (1974) also questioned the force of ‘the three different circumstances’ by which Smith (1776, p. 17) explained the productivity gains from the division of labour: increased dexterity, saving of time, and invention of machinery. Although increased dexterity is clearly a product of a division of labour in a manufacturing process, its scope there is rather limited when compared to that of the continual practice of surgeons, concert pianists and opera singers, to give some examples. Time saved in eliminating time lost in passing from job to job is trivial and not the ‘very considerable’ benefit Smith (1776, pp. 18–19) had suggested. Savings in materials and time through transforming a putting-out to a factory system, an organizational feature of the division of labour Smith had ignored, was more important, particularly through eliminating losses from pilfering. Rae (1834, pp. 164–5) saw savings in the use of tools as far more significant than time saved, and for him (pp. 352–7) this provided the basic reason for extending the division of labour. Other 19th century writers, particularly Babbage (1832), expanded further on this aspect of the matter. Smith's association of

division of labour with inventions (1776, pp. 19–22) covered both ‘on the job improvements’ and scientific inventions by specialists originating from within a more sophisticated division of professions. It ignores, as Hegel (1821, p. 129) was one of the first to point out (cf. Stewart, 1858–75, vol. 8, pp. 318–19), that as division of labour makes ‘work more and more mechanical,…man is able to step aside and install machines in his place’. This feature of the process was subsequently noted by Babbage (1832, pp. 173–4), Ure (1835, p. 21) and developed by Marx (1867). In short, the three circumstances Smith saw as explaining the productivity consequences from the division of labour derive their basic validity from reasons different to those Smith advanced. Further, Smith's remarks (1776, pp. 16–17) on the smaller benefits from applying the division of labour to agriculture than to manufacturing can be contrasted with his quite different and controversial analysis of the primacy of agricultural investment in terms of its employment of productive labour. Agriculture's more substantial contribution to gross revenue as Smith (1776, Book II, ch. 5) subsequently argued, was used by him to define the ‘natural’ course of economic development (Book II, ch. 1) and recommended as superior practice for newly settled regions like the American colonies. Perelman (1984, p. 185) explained this seeming contradiction in Smith by suggesting Smith was the ‘first theorist of neo-imperialism’ because his strategy of development forces developing regions to specialize in raw material production whose terms of trade with manufactures are invariably poor. More likely, Smith's views on the productivity of agriculture relative to manufacturing are posed in terms of different yardsticks: agricultural activity by the very nature of its processes is less amenable to division of labour, even though its ability to employ productive labour is greater than produced by equal investments in manufacture and trade. However, growing mechanization of agriculture, especially in the 20th century, together with the greater scope for exporting agricultural surplus with modern transport, encouraged specialization in agriculture and very large scale farming (Salz, 1934, p. 283).

A final controversial issue from Smith's treatment of the division of labour concerns its social consequences, an argument he placed in the context of public education. The ‘few simple operations’ which under a division of labour most ordinary labouring people are asked to perform, renders them ‘as stupid and ignorant as it is possible for a human creature to become’ and increased ‘dexterity at his own particular trade’ is purchased with a reduction in ‘intellectual, social and martial virtues…unless government take some pains to prevent it’ through providing general education (Smith, 1776, pp. 781–5). Smith was not alone in presenting this disadvantage in an extensive division of labour: similar views were put by Ferguson (1767, p. 280) and Kames (1774). Ferguson described ‘ignorance as the mother of industry’ and argued that prosperous manufactures arise ‘where the mind is least consulted, and where the workshop may…be considered as an engine, the parts of which are men.’ At the turn of the century, and after, German philosophers (for example, Schiller, 1793, Hegel, 1821 and the young Marx, 1844) developed this into a humanist critique of industrial society, suggesting like Smith that these detrimental consequences were removable by education, especially aesthetic education. Such sentiments were resurrected in mid-19th century England by Carlyle (1843) and Ruskin (1851–3, pp. 197–8). For others, Smith's remarks were an aberration, ‘as unfounded [a statement] as can well be imagined’ (McCulloch, 1850, p. 350) or even a contradiction with the division of labour's ability to inspire inventive faculties in labourers (West, 1964).

Despite its deficiencies, Smith's account of the division of labour proved particularly hardy and was

invariably praised in most general terms by major textbook writers of the 19th century and after, though few followed the emphasis Smith gave it as the key factor explaining growth. Cannan (1929, p. 97) ascribed this success to ‘the popularity of its form’. It can also be attributed to the striking productivity increase inherent in the pin example (cf. Mill, 1821, p. 215) and the unambiguous connection Smith drew between increased division of labour, extending the market and human proclivities ‘to truck and barter’ (McCulloch, 1825, pp. 54–5). The account of the division of labour is undoubtedly one of Smith's best remembered performances in economics.

19th-century developments

With the growth of the factory system and more extensive use of increasingly sophisticated machinery, the manufacturing form of division of labour was considerably expanded. Consequently, some economic writers focused on a number of new aspects of the phenomenon, linking the division of labour with developments in the machine tool industry, large scale production and its advantages, and hence, on a more theoretical level, with increasing returns to scale and explicit recognition of a different pattern of productivity growth in manufacturing from that in agriculture.

Charles Babbage was in many respects the pioneer in presenting the division of labour as ‘the most important principle on which the economy of a manufacture depends’ (1832, p. 169). He therefore carefully revised the advantages of a division of labour as first expounded by Adam Smith. In this discussion, time (and cost) savings were also related to time saved in learning a skill and reduced waste of materials during the learning process (pp. 170–1), as well as economy in tool using (p. 172), while the association between division of labour, dexterity and the introduction of new machines was developed more precisely and rigorously. More significantly, Babbage pointed to a hitherto ignored additional advantage of the division of labour he had derived from observation. This had earlier been discussed by Gioja (1815–17) whose interesting contribution on this subject was analysed by Scazzieri (1981, ch. 3).

> *By dividing the work to be executed into different processes of skill or of force,…the master manufacturer…can purchase exactly that precise quantity of both which is necessary for each process; whereas, if the whole work were executed by one workman, that person must possess sufficient skill to perform the most difficult, and sufficient strength to execute the most laborious, of the operations into which the art is divided.* (Babbage, 1832, pp. 175–6; emphasis in original)

This economy of skill, Babbage demonstrated from a pin example, not only reinforced the cost advantages traditionally associated with division of labour, but was also a major cause of establishing large factories: ‘When the number of processes into which it is most advantageous to divide it, and the number of individuals to be employed in it, are ascertained then all factories which do not employ a direct multiple of this number, will produce the article at a greater cost’ (Babbage, 1832, p. 213). Detailed division of labour, Babbage also argued, as in its manufacturing form, can also be applied to mental labour (p. 191). An illustration of its application to mining highlights these control and information gathering features, two aspects of the division of labour to which Babbage paid particular attention. His analysis of the division of labour is even more important because the process as he

described it is made interdependent with machine production, increased factory size, lower costs and prices from such concentration of industry and hence induces growth in demand and an extended market (see Corsi, 1984).

Ure's (1835) contribution must also be noted. It likewise linked development of the factory system to division of labour, summarizing ‘the principle of the factory system…as substituting mechanical science for hand skills, and the partition of a process into its essential constituents’ (1835, p. 20). Ure commented on two other consequences of the division of labour in modern factories: deskilling of the workforce when workers become ‘mere overlookers of machines’ and the development of mechanical engineering since the ‘machine factory displayed the division of labour in manifold gradations’ and facilitated the substitution of skilled hands by ‘the planning, the key-groove cutting, and the drilling machines’ (pp. 20–1).

Accounts of the division of labour by economists of the middle of the century were generally less innovative than those of Babbage and Ure, though they did occasionally provide some new points of departure. Senior (1836, pp. 74–5, 77), after classifying division of labour as one major advantage from the use of capital, concentrated on listing its benefits additional to those given by Smith. Illustrating from the post office, he argued that the fact that ‘the same exertions which are necessary to produce a single given result are often sufficient to produce many hundreds or many thousands similar results’ was one aspect of the division of labour omitted by Smith. The development of retailing as a separate profession was likewise something Smith had failed to consider adequately. More importantly, for a number of reasons, but particularly the division of labour, Senior suggested ‘additional Labour when employed in Manufactures is MORE, when employed in Agriculture is LESS efficient in proportion’, linking manufacturing activity implicitly to increasing returns to scale (1836, pp. 81–2). Mill (1848) treated division of labour as an important aspect of cooperation, arguing that irrespective of its well known productivity advantages, without this complex cooperation in the modern division of labour ‘few things would be produced at all’ (Mill, 1848, p. 118) In discussing the productivity advantages, Mill cited the modification and additional advantages provided by Babbage (1832) and Rae (1834), adding little to their discussion. However, in Chapter 9 dealing with large scale and small scale production, he highlighted the point, so ‘ably illustrated by Mr Babbage…[that] the larger the enterprise, the farther the division of labour may be carried…as one of the principal causes of large manufactories’ (Mill, 1848, p. 131), thereby bringing the argument firmly into the corpus of economics. Mill's account was largely followed by Fawcett (1863) and in most of its essentials by Nicholson (1893).

Marx's account (1867, chs. 13–15) combines much of this discussion, endowing it in the process with sharper analytical insights derived from his study of both the technical literature and his appreciation of the significance of the qualitative changes underlying the evolution of the division of labour. To Marx is owed the important distinction between manufacturing and social division of labour, as well as the precise assessment of the organizational features of its application to modern manufacture, derived from his careful study of Babbage, Ure and many other sources. No wonder that Nicholson (1893, p. 105) described Marx's treatment as ‘both learned and exhaustive and…well worth reading’. More recently, Rosenberg (1976) expressed regret that Marx's close study of ‘both the history of technology, and its newly emerging forms’ has had so few imitators among contemporary economists.

Marshall is another economist from the second half of the 19th century who fully appreciated the importance of the division of labour and revealed it in its more modern forms. In 1879, the *Economics of*

Industry, written with his wife (Marshall and Marshall 1879), devoted Chapter 8 of Book I to the division of labour, immediately after its Chapter 7 on organization of industry. It distinguished the opportunity to apply a division of labour as inherent in the nature of the work, as dependent on direction and control by an entrepreneur as earlier indicated by Bagehot, and as applied to firms: ‘If there are any producers, large and small, all engaged in the same process, *Subsidiary Industries* will grow up to meet their special wants.’ These include special machine tool makers for the industry, improved transport to enhance communication between related firms, as well as auxiliary enterprises in banking and credit provision (Marshall and Marshall 1879, p. 52). Localization of industry also fosters ‘education of skills and taste’ and ‘diffusion of linked knowledge’, and encourages large firms. Hence division of labour is closely related to economies of scale, where size has enabled specialization to grow more and more. Marshall also devoted no less than three chapters to division of labour in his *Principles* (1890, Book IV, chs 9–11), not only covering points traditionally dealt with under this heading, but often introducing subtle modifications. For example, Marshall (1890, p. 263) discounted detrimental social consequences from monotonous work by pointing to the mental stimulus from the ‘social surroundings of the factory’ and the view that factory work was not inconsistent with ‘considerable intelligence and mental resources’. Likewise, he extended Babbage's principle of ‘economy of skill’ to economy of machinery and materials (1890, p. 265), used it as a major explanatory factor for the localization of specialized industry (p. 271) and made it the chief advantage of large scale production in his famous discussion of economies of scale (p. 278). Later, Marshall applied these aspects of his work to his detailed study of industry and trade to explain such things as America's leadership in standardized production (seen by Marshall, 1919, p. 149, as an ‘unprecedented’ application of Babbage's ‘great principle of economical production’), the successful specialization of plant during the First World War, and new issues concerning the growth of the firm. It is therefore paradoxical that Marshall's work in other respects induced the demise of the division of labour in theoretical literature. This arose from the incompatability of increasing returns to scale with stable demand and supply equilibrium (Marshall, 1890, Appendix H). Apart from this, modern equilibrium analysis found it difficult to come to grips with the dynamic features of the division of labour process, and it is presumably at least partly for this reason that division of labour was dropped as an important subject from the economic textbooks (see Kaldor, 1972). However, the locational aspects of the division of labour were further addressed by Becattini (for example, 1990; 2001) in his development of the notion of industrial districts as a concentration of related firms. Marshall had discovered this aspect of industrial organization through the factory tours in the British midlands and Scotland he engaged in from the late 1860s, on which he first reported in 1879. When division of labour for technical reasons could not take place within the same building, small firms spring up specializing in part of the manufacturing process, thereby generating a division of labour among firms concentrated in a particular geographical area (for a survey, see Goodman and Pamford, 1988).

International division of labour

Torrens (1808) appears to have been the first economist to distinguish the territorial division of labour from the mechanical division, suggesting that the former is inspired by ‘different soils and climates [being] adapted to the growth of different production’ thereby inducing regional specialization in those

products which best suit ‘the varieties of their soil’ and climate. Taking advantage of territorial division of labour through regional and international trade enhances productivity and increases the wealth of nations as much as a manufacturing division of labour. Senior (1836, p. 76) also drew attention to this aspect of the division of labour, attributing its discovery to Torrens. Marshall (1890, pp. 267–77) covered territorial division of labour under localization of industry while Taussig (1911, pp. 41–7) called it ‘the geographical division of labour’ with gains arising from ‘the adaptation of different regions to specific articles’ for climatic and resource endowment reasons as well as from the general increase in proficiency which all specialization brings. During the 1970s a new dimension of the international division of labour was analysed, concentrating on its direct foreign investment aspects. Its novel features were a tendency to ‘undermine the traditional bisection of the world into a few industrialized countries on the one hand, and a great majority of developing countries integrated into the world economy solely as raw material producers on the other, and [secondly, to compel] the increasing subdivision of manufacturing processes into a number of partial operations at different industrial sites throughout the world’ to take advantage of favourable labour market circumstances, relatively cheap transport opportunities, tax breaks and other government inducements for foreign investors (Fröbel, Heinrichs and Kreye, 1980, p. 45). This multinational dimension to application of the division of labour is a direct descendent from the concept as understood by Smith, Babbage, Ure and Marx.

The characteristics of the contemporary global division of labour have been well captured by Hobsbawm (2000, pp. 65–6):

> Thus, while the global division of labour was once confined to the exchange of products between particular regions, today it is possible to produce across the frontiers of states and continents. This is what the process is founded on. The abolition of trade barriers and liberalization of markets is, in my opinion, a secondary phenomenon. This is the real difference between the global economy before 1914 and today. Before the Great War, there was pan global movement of capital, goods and labor. But the emancipation of manufacturing and occasionally agricultural products from the territory in which they were produced was not yet possible. When people talked about Italian, British and American industry, they meant not only industries owned by citizens of these countries, but also something that took place almost entirely in Italy, Britain, or America, and was then traded with other countries. This is no longer the case. How can you say that a Ford is an American car, given that it is made of Japanese and European components, as well as parts manufactured in Detroit?

Sexual division of labour

The first explicit reference to a sexual division of labour in economic literature I could find is Hodgskin (1829, pp. 111–12). He argued that

> There is no state of society, probably, in which division of labour between the sexes does not take place. It is and *must* be practiced the instant a family exists. Among even the most barbarous tribes, *war* is the exclusive business of the males; they are in general the

> principal hunters and fishers…the woman labours in and about the hut…In modern as well as in ancient times,…we find the men as the rule taking the out-door work to themselves, leaving the women most of the domestic occupations.…The aptitude of the sexes for different employments, is only an example of the more general principle, that every human being…is better adapted than another to some particular occupation.

Marx and Engels (1845–6, pp. 42–3) ascribed beginnings of the division of labour ‘originally [to] nothing but the division of labour in the sexual act’ and only later to that ‘spontaneously’ or ‘naturally’ derived from predisposition, needs, accidents, and so on. Engels (1884, esp. p. 311) elaborated further on the matter presenting the sexual division of labour in the family as a barrier to the ‘emancipation of women’. Such an emancipation, he argued, was ‘possibly only as a result of modern large-scale industry [which] actually called for the participation of women in production and moreover, strives to convert private domestic work also into a public industry’. Both aspects of the sexual division of labour to which Engels referred in the context of women's emancipation have been taken up in more recent research. The role of domestic labour has been analysed by contemporary writers (see, for example, Himmelweit and Mohun, 1977; Gershuny, 1983) while attention has also been drawn to the shift in the provision of services from domestic production to production for the market (laundromats, take-away-food) as a result of the gradual break-down of the traditional sexual division of labour within the family (Gouverneur, 1978). Sexual division of labour issues have also been applied in segmented labour market analysis, thereby enriching this particular aspect of labour economics.

Becker (1985) has analysed the sexual division of labour in the context of human capital investment and allocating the work load of parties within the household. Thus both the allocation of effort within a household, and the advantages of investing in specific human capital are designed to enhance the social division of labour and its benefits without necessarily diminishing the exploitative aspects of such arrangements (Becker 1985, p. S41). Social factors are, however, equally important. Increasing returns by itself cannot explain the traditional division of labour within the household; a division of labour itself subject to change. The increased contribution to housework by men during the 1970s is one observed aspect of this social change (Becker 1985, p. S56). Furthermore, as Posner (1992, pp. 54, 129) has noted in particular, women were not fully brought into the work place on a large scale until the two world wars, and this only became a dominant pattern in employment from the 1950s onwards. Cigno (1991) discusses many of these issues as part of his economics of the family.

Decline and rehabilitation of division of labour in the 20th and 21st centuries

The association between division of labour and increasing returns, the consequent possibility of falling supply and cost curves, created problems for equilibrium analysis already noticed as a factor explaining decline in emphasis on the division of labour and induced its virtual elimination from much of the theoretical literature. Attempts to remove division of labour from economics were also based on other grounds. Robbins (1932, pp. 32–8) argued that study of the ‘technical arts of production’ belonged to engineering and not to economics or, in the case of ‘motion study’, to industrial psychology even if this meant removal of traditional topics like division of labour from economics. Robbins's approach followed

Sidgwick's (1883, pp. 104–7) treatment, removing all technical aspects from the topic, leaving only what he called the pure economics side. Others suggested it was better to leave discussion of division of labour to sociologists because Durkheim, and before him Comte and Herbert Spencer, had absorbed it within this emerging discipline. However, some economists in the 20th century objected to removal of the division of labour from economics. In particular, this would reduce understanding of the dynamics of economic progress.

Allyn Young (1928) was one of these economists. He made Adam Smith's theorem that the division of labour is limited by the extent of the market the central theme of his address to section F of the British Association, arguing this was ‘one of the most illuminating and fruitful generalizations which can be found in the whole literature of economics’ (Young, 1928, p. 529). Rather than covering all aspects of the division of labour, Young concentrated on two interdependent matters: ‘growth of indirect and roundabout methods of production and the division of labour [or increased specialization] among industries’ (Young, 1928, p. 529) but the former, as Kaldor (1975, pp. 355–6) pointed out, was not to be confounded with the Austrian capital theoretic notion. From this he deduced division of labour as a cumulative, self-reinforcing process, because every re-organization of production, sometimes described as a new invention, involves fresh application of scientific progress to industry,

> alters the conditions of industrial activity and initiates responses elsewhere in the industrial structure which in turn have further unsettling effects.…The apparatus of supply and demand in their relation to prices does not seem to be particularly helpful for the purpose of an inquiry into these broader aspects of increasing returns. (Young, 1928, p. 533)

However, apart from this damaging conclusion for competitive price theory, the ‘possibility of economic progress’ could not really be grasped by ignoring these factors of greater specialization, better combinations of advantages of location, and a consequent increased number of specialized producers between basic raw materials and final producers (Young, 1928, pp. 538–40).

Kaldor was a major economist who took up Young's challenge in both its critical (Kaldor, 1972; 1975) and more constructive aspects (Kaldor, 1966; 1967). The major thrust of Kaldor's positive argument proclaimed that faster growth is derived from faster growth in the manufacturing sector, partly from the cumulative features linking the growth of manufacturing to growth of labour productivity via static and dynamic economies of scale, or the notion of increasing returns as developed by Young from the division of labour. This strong and powerful interaction of productivity growth and manufacturing growth is also posited in Verdoorn's Law (1949) but its association with aspects of the division of labour is what is relevant here. Faster manufacturing growth draws labour from other sectors of the economy, inducing faster productivity growth, but as the scope of transferring such labour from lower productivity sectors like agriculture dries up, the growth process slows down (see Thirlwall, 1983). A key feature of the process, as Rowthorn (1975, p. 899, n. 1) noticed in one his skirmishes with Kaldor on the subject, is that it is an interdependent, cumulative historical process where ‘higher productivity means more exports which means greater industrial output which via its effects on investment, innovation and *scale of production* reacts back on productivity growth’. The importance of such a process was given detailed

empirical examination in a discussion of the Taiwan machine tools industry in the 1970s as an application of the division of labour, envisaged as increases in output increasing productivity, with ‘technological change, broadly defined, sandwiched in between’ (Amsden, 1985, p. 271). Writers in the new growth economics, who emphasized the impact of increasing returns from specialization on growth performance (Romer, 1987) drew in part for their inspiration on the literature of the division of labour, in Romer's case as represented by Marshall (1890) and possibly Young (1928).

Research from the 1990s has particularly stressed the importance of communication and co-ordination costs of the division of labour. Becker and Murphy (1992) portray these costs as setting limits on the division of labour more important than that exerted by the extent of the market so heavily emphasized by Adam Smith. Subsequent, Camacho (1996) has studied this aspect in more detail, drawing a clear and direct relationship between increases in the division of labour and rises in both communication and co-ordination costs, as an essential extension to the modern theory of the firm and the market. Perlin (1993) has treated inter-firm cooperation and its benefits from a similar angle, assessing the benefits for production from such cooperation as an economy of conventions and inter-firm agreements. This analysis thereby treats division of labour once again as part of the organizational theory of the firm or a production unit in which much emphasis is placed on the potential trade-offs between the economies reaped from specialization and the transaction costs it generates (Yang and Ng 1993). In this way, division of labour has also become an important part of the foundations for a new classical micro-economic analytic framework.

Conclusion

Viewed dynamically within the context of economic growth, as Smith (1776) and others had intended the division of labour to be contemplated, it continues to be a powerful tool for understanding the process of growth and development. On this ground alone it can therefore not be jettisoned from economics as unwanted baggage, as Robbins (1932) mistakenly suggested. When its importance for understanding aspects of the labour process, the labour market, the theory of production and the theory of the firm contemplated at the plant and the industry level are included, this argument is even stronger. As mentioned in the previous paragraph, on these grounds division of labour is making a definite come-back as part of the theory of a new classical micro-economics. Last, but not least, the importance of the division of labour for economics is underlined by the fact that some of the major economic minds from both past and present have invariably included it as an important part of their economic analysis.

Bibliography

Amsden, A.H. 1985. The division of labour is limited by the rate of growth of the market; the Taiwan machine tool industry in the 1970s. *Cambridge Journal of Economics* 9, 271–84.

Anonymous. 1701. Considerations on the East-India trade. In *A Select Collection of Early English Tracts on Commerce*, ed. J. R. McCulloch. London, 1856; Cambridge: Cambridge University Press, 1954.

Babbage, C. 1832. *On the Economy of Machinery and Manufactures*. London. Fourth enlarged edition of 1835; New York: Augustus M. Kelley, 1963.

Becattini, G. 1990. The Marshallian industrial district as a socio-economic notion. In *Industrial Districts and Inter-firm Co-Operation in Italy*, ed. F. Pyke, G. Becattini and W. Sengenberger. Geneva: International Institute of Labour Studies.

Becattini, G. 2001. *The Caterpillar and the Butterfly*. Florence: Felice de Monier.

Beccaria, C. 1771. *Elementi di economia publica*. In *Opere*, ed. S. Romagnoli. Florence: Sansoni, 1958.

Becker, G.S. 1985. Human capital, effort, and the sexual division of labour. *Journal of Labour Economics* 3(1), S33–S58.

Becker, G.S. and Murphy, K.M. 1992. The division of labour, coordination costs, and knowledge. *Quarterly Journal of Economics* 107, 1137–60.

Bonar, J. 1893. *Philosophy and Political Economy*. 3rd edn. London: Allen and Unwin, 1967.

Braverman, H. 1974. *Labour and Monopoly Capital: The Degradation of Work in the Twentieth Century*. New York: Monthly Review Press.

Camacho, A. 1996. *Division of Labour, Variability, Coordination and the Theory of the Firm and Markets*. Dordrecht: Kluwer.

Campbell, R. 1747. *The London Tradesman*. London.

Cannan, E. 1929. *Review of Economic Theory*. London: P.S. King & Son.

Carlyle, T. 1843. *Past and Present*. London: Chapman & Hall. Another edn. London: G. Routledge & Sons, 1893.

Cigno, A. 1991. *Economics of the Family*. Oxford: Oxford University Press.

Corsi, M. 1984. Il sistema di fabbrica e la divisione del lavoro, il pensiero di Charles Babbage. *Quaderni di storia dell'economia politica* 3, 111–23.

Dosi, G. 1988. Sources, procedures, and micro-economic effects of innovation. *Journal of Economic Literature* 26, 1120–71.

Engels, F. 1884. *Origin of the Family, Private Property and the State*. In *Marx-Engels Selected Works*,

vol. 2. Moscow: Progress Publishers.

Fawcett, H. 1863. *Manual of Political Economy*. London and Cambridge: Macmillan & Co.

Ferguson, A. 1767. *An Essay on the History of Civil Society*. Edinburgh.

Fröbel, F., Heinrichs, J. and Kreye, O. 1980. *The New International Division of Labour*, trans. P. Burgess. Cambridge: Cambridge University Press.

Gershuny, J. 1983. *Social Innovation and Division of Labour*. Oxford: Oxford University Press.

Gioja, M. 1815–17. *Nuove prospetto delle scienze economiche*. Milan: G. Pirotta.

Goodman, E. and Pamford, J., eds. 1989. *Small Firms and Industrial Districts in Italy*. London: Routledge.

Gouverneur, J. 1978. *Contemporary Capitalism and Marxist Economics*. Trans. R. le Farnu. Oxford: Robertson, 1983.

Groenewegen, P. 1977. Adam Smith and the division of labour: a bi-centenary estimate. *Australian Economic Papers* 16, 161–74.

Harris, J. 1757. *An Essay Upon Money and Coins*. Part I. London: G. Hawkins.

Hegel, G.W.F. 1821. *Philosophy of Right*. Trans. T.M. Knox, Oxford: Clarendon Press, 1982.

Hicks, J.R. 1965. *Capital and Growth*. Oxford: Clarendon Press.

Himmelweit, S. and Mohun, S. 1977. Domestic Labour and Capital. *Cambridge Journal of Economics* 1, 15–31.

Hobsbawm, E. 2000. *The New Century*. London: Abacus.

Hodgskin, T. 1827. *Popular Political Economy*. London: C. Tait. New York: A.M. Kelley, 1966.

Hutcheson, F. 1755. *A System of Moral Philosophy*. Glasgow.

Hutchison, T.W. 1988. *Before Adam Smith: The Emergence of Political Economy 1662–1776*. Oxford: Basil Blackwell.

Kaldor, N. 1966. *Causes of the Slow Rate of Economic Growth in the United Kingdom*. Cambridge:

Cambridge University Press.

Kaldor, N. 1967. *Strategic Factors in Economic Development*. Ithaca: State School of Industrial and Labour Relations, Cornell University.

Kaldor, N. 1972. The irrelevance of equilibrium economics. *Economic Journal* 82, 1237–55.

Kaldor, N. 1975. What is wrong with economic theory? *Quarterly Journal of Economics* 89, 347–57.

Kames, H.H. 1774. *Sketches of the History of Man*. Edinburgh: W. Creech. London: W. Strahan and T. Cadell.

Lindsay, P. 1733. *The Interest of Scotland Considered*. Edinburgh: R. Fleming & Co.

McCulloch, J.R. 1825. *Principles of Political Economy*. London: Murray, 1870.

McCulloch, J.R. 1850. Introduction and notes to Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*. 4th edn. Edinburgh: A. & C. Black; London: Longman.

Mandeville, B. 1729. *The Fable of the Bees*. London: A. Roberts.

Marglin, S. 1974. What do bosses do? The origins and functions of hierarchy in capitalist production, *Review of Radical Political Economics* 6, 60–112.

Marshall, A. 1890. *Principles of Economics*. 8th edn. London: Macmillan & Co., 1920.

Marshall, A. 1919. *Industry and Trade*. 4th edn. London: Macmillan & Co., 1923.

Marshall, A. and Marshall, M.P. 1879. *The Economics of Industry*. London: Macmillan and Company.

Marx, K. 1844. *Economic and Philosophic Manuscripts*. Moscow: Foreign Languages Publishing House, 1959.

Marx, K. 1867. *Capital*, vol. 1, Moscow: Foreign Languages Publishing House, 1959.

Marx, K. and Engels, F. 1845–6. *The German Ideology*. Moscow: Progress Publishers, 1964.

Meek, R.L. and Skinner, A.S. 1973. The development of Adam Smith's ideas on the division of labour. *Economic Journal* 83, 1094–116.

Mill, J.S. 1821. *Elements of Political Economy*. 3rd edn. In *The Selected Economic Writings of James*

Mill, ed. D.N. Winch. Edinburgh: Oliver & Boyd for the Scottish Economic Society, 1966.

Mill, J.S. 1848. Principles of Political Economy. In *The Collected Works of John Stuart Mill*, vols. 2 and 3, ed. J.M. Robson. Toronto: Toronto University Press, 1965.

Nicholson, J.S. 1893. *Principles of Political Economy*. 2nd edn. London: A. & C. Black.

Perelman, M. 1984. *Classical Political Economy: Primitive Accumulation and the Social Division of Labour*. London: Rowman & Allenheld.

Pernin, J.L. 1993. La cooperation entre firmes: une approche par l'économie des conventions. *Economie appliqué* 46, 105–26.

Petty, W. 1671. *Political Arithmetick*. In *Economic Writings of Sir William Petty*, ed. C.H. Hull. New York: A.M. Kelley, 1963.

Petty, W. 1683. *Another Essay on Political Arithmetick concerning the Growth of the City of London*. In *Economic Writings of Sir William Petty*, ed. C.H. Hull. New York: A.M. Kelley, 1963.

Plato. 380 bc. *The Republic*. Harmondsworth: Penguin Classics, 1955.

Posner, R.A. 1992. *Sex and Reason*. Cambridge, MA: Harvard University Press.

Quesnay, F. 1765. Natural right. Extracts trans. R.L. Meek, *The Economics of Physiocracy*. London: Allen & Unwin, 1962.

Rae, J. 1834. *Statement of Some New Principles on the Subject of Political Economy*. New York: A.M. Kelley, 1964.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan, 1935.

Romer, P.M. 1987. New theories of economic growth: growth based on increasing returns due to specialization. *American Economic Review* 77, 56–62.

Rosenberg, N. 1976. Marx as a student of technology. *Monthly Review* 28, 56–77. In N. Rosenberg, *Inside the Black Box: Technology and Economics*. Cambridge: Cambridge University Press, 1982.

Rowthorn, R.E. 1975. A reply to Lord Kaldor's comment. *Economic Journal* 85, 897–901.

Ruskin, J. 1851–3. *The Stones of Venice*. In *The Complete Works of John Ruskin*, ed. E.T. Cook and A.

Wedderburn. London: George Allen, 1904.

Salz, A. 1934. Specialisation. In *International Encyclopaedia of the Social Sciences*, vol. 13, ed. E.R.A. Seligman. New York: Macmillan Company, 1948.

Scazzieri, R. 1981. *Efficienza, produttività e livelli di attitività*. Bologna: Il Mulino.

Schiller, F. 1793. *On the Aesthetic Education of Man*. Trans. R. Snell, New York: Ungar, 1980.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Oxford University Press.

Senior, N. 1836. *An Outline of the Science of Political Economy*. London: George Allen & Unwin, 1938, 1951.

Sidgwick, H. 1883. *Principles of Political Economy*, 2nd edn. London: Macmillan & Co.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Clarendon Press, 1976.

Stewart, D. 1858–75. *Collected Works of Sir Dugald Stewart*, ed. Sir William Hamilton, Edinburgh.

Sun, G.Z. 2005. *Readings in the Economics of the Division of Labour*. Singapore: World Scientific Publishing Company.

Taussig, F.W. 1911. *Principles of Economics*. 3rd edn. New York: Macmillan, 1936.

Thirlwall, A.P. 1983. A plain man's guide to Kaldor's growth law. *Journal of Post Keynesian Economics* 5, 345–58.

Torrens, R. 1808. *The Economists Refuted*. Sydney: Department of Economics, University of Sydney, Reprints of Economic Classics, 1984.

Tucker, J. 1755. *The Elements of Commerce and the Theory of Taxes*. London.

Tucker, J. 1774. *Four Tracts on Political and Economic Subjects*, 2nd edn. Gloucester: R. Raikes.

Turgot, A.R.J. 1751. Lettre à Madame de Graffigny sur les letters d'un Péruvienne. In *Oeuvres de Turgot et documents le concernant*, vol. 1, ed. G. Schelle. Paris: F. Alcan, 1913.

Turgot, A.R.J. 1766. Reflections on the Production and Distribution of Wealth. In *The Economics of A.R. J. Turgot*, ed. P.D. Groenewegen. The Hague: Nijhoff, 1977.

Ure, A. 1835. *The Philosophy of Manufactures*. London: C. Knight; London: Frank Cass, 1967.

Verdoorn, P.J. 1949. Fattori che regolano lo sviluppo della produttività del lavoro. *L'industria* 1, 45–53.

West, E.G. 1964. Adam Smith's two views on the division of labour. *Economica* 3, 23–32.

Yang, X.-K. and Ng, Y.-K. 1993. *Specialisation and Economic Integration: A New Classical Micro-economic Framework*. New York: Elsevier Science.

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 327–42.

How to cite this article

Groenewegen, Peter. "division of labour." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000176> doi:10.1057/9780230226203.0401

The New Palgrave Dictionary of Economics Online

Dobb, Maurice Herbert (1900 – 1976)

Amartya Sen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Dobb, M.; economic development; equality; exploitation; feudalism; labour theory of value; Langer–Lerner price mechanism; market socialism; Marxism; planning; revealed preference theory; saving and investment; socialist pricing theory; Sraffa, P.; Sweezy. P. M.; utility theory of value

Article

Maurice Dobb was undoubtedly one of the outstanding political economists of this century. He was a Marxist, and was one of the most creative contributors to Marxian economics. As Ronald Meek put it, in his obituary of Dobb for the British Academy, ‘over a period of fifty years [Dobb] established and maintained his position as one of the most eminent Marxist economists in the world’. Dobb's *Political Economy and Capitalism* (1937) and *Studies in the Development of Capitalism* (1946) are his two most outstanding contributions to Marxian economics. The former is primarily concerned with economic theory (including such subjects as value theory, economic crises, imperialism, socialist economies), and the latter with economic history (particularly the emergence of capitalism from feudalism). These two fields – economic theory and economic history – were intimately connected in Dobb's approach to economics. He also wrote an influential book on Soviet economic development. This was first published under the title *Russian Economic Development since the Revolution* (1928), and later in a revised edition as *Soviet Economic Development since 1917* (1948).

Maurice Dobb was born on 24 July 1900 in London. His father Walter Herbert Dobb had a draper's retail business and his mother Elsie Annie Moir came from a Scottish merchant's family. He was educated at Charterhouse, and then at Pembroke College, Cambridge, where he studied economics. This was followed by two postgraduate years at the London School of Economics, where he did his Ph.D. on ‘The Entrepreneur’. The thesis formed the basis of his book *Capitalist Enterprise and Social Progress* (1925). Dobb returned to Cambridge at the end of 1924 on being appointed as a lecturer in economics. He taught in Cambridge until his retirement in 1967. He was a Fellow of Trinity College, and was elected to a University Readership in 1959. He received honorary degrees from the Charles University of Prague, the University of Budapest, and Leicester University, and was elected a Fellow of the British Academy.

After retirement he and his wife, Barbara, stayed on in the neighbouring village of Fulbourn. He died on 17 August 1976.

Dobb was a theorist of great originality and reach. He was also, throughout his life, deeply concerned with economic policy and planning. His foundational critique of ‘market socialism’ as developed by Oscar Lange and Abba Lerner, appeared in the *Economic Journal* of 1933, later reproduced along with a number of related contributions in his *On Economic Theory and Socialism* (1955). His relatively elementary book *Wages* (1928) presented not merely a simple introduction to labour economics, but also an alternative outlook on these questions, including their policy implications, leading to interesting disputations with John Hicks, among others. In later years Dobb was much concerned with planning for economic development. In three lectures delivered at the Delhi School of Economics, later published as *Some Aspects of Economic Development* (1951), Dobb discussed some of the central issues of development planning for an economy with unemployed or underutilized labour, and his ideas were more extensively developed in his later book, *An Essay on Economic Growth and Planning* (1960). Maurice Dobb also published a number of papers on more traditional fields in economic theory, including welfare economics, and some of these papers were collected together in his *Welfare Economics and the Economics of Socialism* (1969). In his *Theories of Value and Distribution since Adam Smith: Ideology and Economic Theory* (1973), he responded *inter alia* to the new developments in Cambridge political economy, including the influential ‘Prelude to a Critique of Economic Theory’ by Piero Sraffa (1960). Maurice Dobb's association with Piero Sraffa extended over a long period, both as a colleague at Trinity College, and also as a collaborator in editing *Works and Correspondence of David Ricardo*, published in 11 volumes between 1951 and 1973 (on the latter, see Pollitt, 1990).

In addition to academic writings, Maurice Dobb also did a good deal of popular writing, both for workers’ education and for general public discussion. He wrote a number of pamphlets, including *The Development of Modern Capitalism* (1922), *Money and Prices* (1924), *An Outline of European History* (1926), *Modern Capitalism* (1927), *On Marxism Today* (1932), *Planning and Capitalism* (1937), *Soviet Planning and Labour in Peace and War* (1942), *Marx as an Economist: An Essay* (1943), *Capitalism Yesterday and Today* (1958), and *Economic Growth and Underdeveloped Countries* (1963), and many others. Dobb was a superb communicator, and the nature of his own research was much influenced by policy debates and public discussions. Dobb the economist was not only close to Dobb the historian, but also in constant company of Dobb the member of the public. It would be difficult to find another economist who could match Dobb in his extraordinary combination of genuinely ‘high-brow’ theory, on the one hand, and popular writing on the other. The author of *Political Economy and Capitalism* (from the appearance of which – as Ronald Meek (1978) rightly notes – ‘that future historians of economic thought will probably date the emergence of Marxist economics as a really serious economic discipline’: was also spending a good deal of effort writing pamphlets and material for labour education, and doing straightforward journalism. It is not possible to appreciate fully Maurice Dobb's contributions to economics without taking note of his views of the role of economics in public discussions and debates.

Another interesting issue in understanding Dobb's approach to economics concerns his adherence to the labour theory of value. The labour theory has been under attack not only from neoclassical economists, but also from such anti-neoclassical political economists as Joan Robinson and, indirectly, even Piero Sraffa. In his last major work, *Theories of Value and Distribution since Adam Smith* (1973), Maurice Dobb speaks much in support of the relevance of Sraffa's (1960) major contribution, which eschews the use of labour values (on this see Steedman, 1977), but without abandoning his insistence on the

importance of the labour theory of value. It is easy to think that there is some inconsistency here, and it is tempting to trace the origin of this alleged inconsistency to Dobb's earlier writings, which made Abram Bergson remark that ‘in Dobb's analysis the labour theory is not so much an analytic tool as excess baggage’ (Bergson, 1949, p. 445).

The key to understanding Dobb's attitude to the labour theory of value is to recognize that he did not see it just as an intermediate product in explaining relative prices and distributions. He took ‘the labour-principle’ as ‘making an important qualitative statement about the nature of the economic problem’ (Dobb, 1937, p. 21). He rejected seeing the labour theory of value as simply a ‘first approximation’ containing ‘nothing essential that cannot be expressed equally well and easily in other terms’ (Dobb, 1973, pp. 148–9). The description of the production process in terms of labour involvement has an interest that extends far beyond the role of the labour value magnitudes in providing a ‘first approximation’ for relative prices. As Dobb (1973, pp. 148–9) put it,

> there is something in the first approximation that is lacking in later approximations or cannot be expressed so easily in those terms (e.g., the first approximation may be a device for emphasising and throwing into relief something of greater generality and less particularity).

Any description of reality involves some selection of facts to emphasize certain features and to underplay others, and the labour theory of value was seen by Dobb as emphasizing the role of those who are involved in ‘personal participation in the process of production *per se*’ in contrast with those who do not have such personal involvement.

> As such ‘exploitation’ is neither something ‘metaphysical’ nor simply an ethical judgement (still less ‘just a noise’) as has sometimes been depicted: it is a factual description of a socio-economic relationship, as much as is Marc Bloch's apt characterisation of Feudalism as a system where feudal Lords ‘lived on the labour of other men’. (Dobb, 1973, p. 145)

The possibility of calculating prices without going through value magnitudes, and the greater efficiency of doing that (on this see Steedman, 1977), does not affect this descriptive relevance of the labour theory of value in any way. Maurice Dobb also outlined the relationship of this primarily descriptive interpretation of labour theory of value with evaluative questions, for example, assessing the ‘right of ownership’ (see especially Dobb, 1937).

The importance for Dobb of descriptive relevance is brought out also by his complex attitude to the utility theory of value. While he rejected the view that the utility picture is the best way of seeing relative values (‘by taking as its foundation a fact of individual consciousness’), he lamented the descriptive impoverishment that is brought about by replacing the subjective utility theory by the ‘revealed preference’ approach.

> If all that is postulated is simply that men *choose*, without anything being stated even as to how they choose or what governs their choice, it would seem impossible for economics to

> provide us with any more than a sort of algebra of human choice. (Dobb, 1937, p. 171)

Indeed, as early as 1929, a long time before the ‘revealed preference theory’ was formally inaugurated by Paul Samuelson, Dobb (1929, p. 32) had warned:

> Actually the whole tendency of modern theory is to abandon such psychological conceptions: to make utility and disutility coincident with observed offers in the market; to abandon a ‘theory of value’ in pursuit of a ‘theory of price’. But this is to surrender, not to solve the problem.

Maurice Dobb's open-minded attitude to non-Marxian traditions in economics added strength and reach to his own Marxist theorizing. He could combine Marxist reasoning and methodology with other traditions, and he was eager to be able to communicate with economists belonging to other schools. Dobb's honesty and lack of dogmatism were important for the development of the Marxist economic tradition in the English-speaking world, because he occupied a unique position in Marxist thinking in Britain. As Eric Hobsbawm (1967, p. 1) has noted,

> for several generations (as these are measured in the brief lives of students) he was not just the only Marxist economist in a British university of whom most people had heard, but virtually the only don known as a communist to the wider world.

The Marxist economic tradition was well served by Maurice Dobb's willingness to engage in spirited but courteous debates with economists of other schools. Dobb achieved this without compromising the integrity of his position. The distinctly Marxist quality of his economic writings was as important as his willingness to listen and dispassionately analyse the claims of other schools of thought with which he engaged in systematic disputation. The gentleness of Dobb's style of disputation arose from strength rather than from weakness.

Dobb's willingness to appreciate positive elements in other economic traditions while retaining the distinctive qualities of his own approach is brought out very clearly also in his truly far-reaching critique of the theory of socialist pricing as presented by Lange, Lerner, Dickinson and others in the 1930s. Dobb noted the efficiency advantages of a price mechanism, especially in a static context. He was, however, one of the first economists to analyse clearly the conflict between the demands of efficiency expressed in the equilibrium conditions of the Langer–Lerner price mechanism (and also of course in a perfectly competitive market equilibrium), and the demands that would be imposed by the requirements of equality, given the initial conditions. In his paper called ‘Economic Theory and the Problems of a Socialist Economy’ published in 1933, Maurice Dobb argued thus:

> If carpenters are scarcer or more costly to train than scavengers, the market will place a higher value upon their services, and carpenters will derive a higher income and have greater ‘voting power’ as consumers. On the side of supply the extra ‘costliness’ of carpenters will receive expression, but only at the expense of giving carpenters a differential ‘pull’ as consumers, and hence vitiating the index of demand. On the other

> hand, if carpenters and scavengers are to be given equal weight as consumers by assuring them equal incomes, then the extra costliness of carpenters will find no expression in costs of production. Here is the central dilemma. Precisely because consumers are also producers, both costs and needs are precluded from receiving simultaneous expression in the same system of market valuations. Precisely to the extent that market valuations are rendered adequate in one direction they lose significance in the other. (1933, p. 37)

The fact that given an initial distribution of resources the demands of efficiency and those of equity may – and typically will – conflict is, of course, one of the major issues in the theory of resource allocation, with implications for market socialism as well as for competitive markets in a private ownership economy. As a matter of fact, Marx had *inter alia* noted this conflict in his *Critique of Gotha Programme*, but in the discussion centring around Langer–Lerner systems, this deep conflict had attracted relatively little attention, except in the arguments presented by Maurice Dobb. The fact that even a socialist economy has to cope with inequalities of initial resource distribution (arising from, among other things, differences in inherited talents and acquired skills) makes it a relevant question for a socialist economy as well as for competitive market economies, and Dobb's was one of the first clear analyses of this central question of resource allocation.

The second respect in which Maurice Dobb found the literature on market socialism inadequate concerns allocation over time. In discussing the achievements and failures of the market mechanism, Maurice Dobb argued that the planning of investment decisions

> may contribute much more to human welfare than could the most perfect micro-economic adjustment, of which the market (if it worked like the textbooks, at least, and there were no income-inequalities) is admittedly more fitted in most cases to take care. (Dobb, 1960, p. 76)

In his book *An Essay in Economic Growth and Planning* (1960), Dobb provided a major investigation of the basis of planned investment decisions, covering overall investment rates, sectoral divisions, choice of techniques, and pricing policies related to allocation (including that over time).

This contribution of Dobb relates closely to his analysis of the problems of economic development. In his earlier book *Some Aspects of Economic Development* (1951), Dobb had already presented a pioneering analysis of the problem of economic development in a surplus-labour economy, with shortage of capital and of many skills. While, on the one hand, he anticipated W.A. Lewis's (1954) more well-known investigation of economic growth with ‘unlimited supplies of labour’, he also went on to demonstrate the far-reaching implications of the over-all savings rates being socially sub-optimal and inadequate. Briefly, he showed that this requires not only policies directly aimed at raising the rates of saving and investment, but it also has implications for the choice of techniques, sectoral balances, and price fixation.

In such a brief note, it is not possible to do justice to the enormous range of Maurice Dobb's contributions to economic theory, applied economics and economic history. Different authors influenced by Maurice Dobb have emphasized different aspects of his many-sided works (see, for example, Feinstein, 1967, and the *Cambridge Journal of Economics*’ Maurice Dobb Memorial Issue (1978)). He

has also had influence even outside professional economics, particularly in history, especially through his analysis of the development of capitalism.

Dobb argued that the decline of feudalism was caused primarily by ‘the inefficiency of Feudalism as a system of production, coupled with the growing needs of the ruling class for revenue’ (1946, p. 42). This view of feudal decline, with its emphasis on *internal* pressures, became the subject of a lively debate in the early 1950s. An alternative position, forcefully presented by Paul Sweezy in particular, emphasized some *external* developments, especially the growth of trade, operating through the relations between the feudal countryside and the towns that developed on its periphery. No matter what view is taken as to ‘who won’ the debates on the transition from feudalism to capitalism, Dobb's creative role in opening up a central question in economic history as well as a major issue in Marxist political economy can scarcely be disputed. Indeed, *Studies in the Development of Capitalism* (1946) has been a prime mover in the emergence of the powerful Marxian tradition of economic history in the English-speaking world, which has produced scholars of the eminence of Christopher Hill, Rodney Hilton, Eric Hobsbawm, Edward Thompson and others.

It is worth emphasizing that aside from the explicit contributions made by Maurice Dobb to economic history, he also did use a historical approach to economic analysis in general. Maurice Dobb's deep involvement in descriptive richness (as exemplified by his analysis of ‘the requirements of a theory of value’), his insistence on not neglecting the long-run features of resource allocation (influencing his work on planning as well as development), his concern with observed phenomena in slumps and depressions in examining theories of ‘crises’, and so on, all relate to the historian's perspective. Dobb's works in the apparently divergent areas of economic theory, applied economics and economic history are, in fact, quite closely related to each other.

Maurice Dobb was not only a major bridge-builder between Marxist and non-Marxist economic traditions (aside from pioneering the development of Marxist economics in Britain and to some extent in the entire English-speaking world): he also built many bridges between the different pursuits of economic theorists, applied economists and economic historians. Dobb's political economy involved the rejection of the narrowly economic as well as the narrowly doctrinaire. He was a great economist in the best of the broad tradition of classical political economy.

Selected works

1925. *Capitalist Enterprise and Social Progress*. London: Routledge.

1928. *Russian Economic Development since the Revolution*. London: Routledge.

1928. *Wages*. London: Nisbet; Cambridge: Cambridge University Press.

1929. A sceptical view of the theory of wages. *Economic Journal* 39, 506–19.

1933. Economic theory and the problems of a socialist economy. *Economic Journal* 43, 588–98.

1937. *Political Economy and Capitalism: Some Essays in Economic Tradition*. London: Routledge.

1946. *Studies in the Development of Capitalism*. London: Routledge.

1948. *Soviet Economic Development since 1917*. London: Routledge.

1950. Reply (to Paul Sweezy's article on the transition from feudalism to capitalism). *Science and Society* 14(2), 157–67.

1951. *Some Aspects of Economic Development: Three Lectures*. Delhi: Ranjit Publishers, for the Delhi School of Economics.

1955. *On Economic Theory and Socialism*. London: Routledge.

1960. *An Essay on Economic Growth and Planning*. London: Routledge.

1969. *Welfare Economics and the Economics of Socialism*. Cambridge: Cambridge University Press.

1973. *Theories of Value and Distribution since Adam Smith: Ideology and Economic Theory*. Cambridge: Cambridge University Press.

Bibliography

Bergson, A. 1949. Socialist economics. In *A Survey of Contemporary Economics*, ed. H.S. Ellis. Philadelphia: Blakiston.

Cambridge Journal of Economics. 1978. Maurice Dobb memorial issue. Vol. 2(2), June.

Hobsbawm, E.J. 1967. Maurice Dobb. In *Socialism, Capitalism and Economic Growth: Essays Presented to Maurice Dobb*, ed. C. Feinstein. Cambridge: Cambridge University Press.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School* 20(2), 139–91.

Meek, R. 1978. Obituary of Maurice Herbert Dobb. *Proceedings of the British Academy 1977* 53, 333–44.

Pollitt, B.H. 1990. Clearing the path for ‘Production of Commodities by Means of Commodities’: notes on the collaboration of Maurice Dobb in Piero Sraffa's edition of ‘The Works and Correspondence of David Ricardo’. In *Essays on Piero Sraffa: Critical Perspectives on the Revival of Classical Theory*, ed. K. Bharadwaj and B. Schefold. London: Unwin Hyman.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Sraffa, P., with the collaboration of M.H. Dobb. 1951–73. *Works and Correspondence of David Ricardo*. 11 vols, Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

How to cite this article

Sen, Amartya. "Dobb, Maurice Herbert (1900–1976)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000178> doi:10.1057/9780230226203.0402

The New Palgrave Dictionary of Economics Online

dollarization

Roberto Chang
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article focuses on dollarization, a situation in which a foreign currency (often the US dollar) replaces a country's currency in performing one or more of the basic functions of money. The distinction between official dollarization and endogenous dollarization is discussed, as are the concepts of currency substitution and liability dollarization. Implications for monetary and exchange rate policy are emphasized.

Keywords

aggregate demand; capital asset pricing model; currency substitution; dollarization; euro; financial dollarization; inflation; lender of last resort; liability dollarization; monetary policy; net worth effects; portfolio balance; search theory; seigniorage; stabilization; transaction costs

Article

Dollarization is a situation in which a foreign currency (often the US dollar) replaces a country's currency in performing one or more of the basic functions of money.
Thus in Ortiz (1983) the term ‘dollarization’ refers to the widespread usage of US dollars for transaction purposes in Mexico. More recently, Ize and Levy-Yeyati (2003) use ‘financial dollarization’ for episodes in which domestic financial contracts are denominated in dollars or another foreign currency.
In some countries, dollarization has been the outcome of official government policy. Examples include Ecuador in 2000 and El Salvador in 2001, where the domestic currency was retired from circulation and the US dollar became the official currency. An immediate implication of such ‘official dollarization’ is that domestic prices of tradable goods are tied to world prices, so domestic inflation is closely related to US inflation. Hence official dollarization has been advocated for countries suffering from chronic, high, and volatile inflation.
On the other side of the ledger, official dollarization implies the surrender of independent monetary

policy, leaving only fiscal policy available as a stabilization tool. In addition, the domestic government gives up seigniorage, or the revenue from money creation, which accrues to the US Federal Reserve. While both effects are widely regarded as costly for the domestic economy, their welfare implications depend on details about the policymaking process and, in particular, on whether the monetary authorities can credibly commit to implement optimal policy (see Chang and Velasco, 2002, for a discussion).

Finally, official dollarization implies that the domestic central bank is no longer available as a lender of last resort, which may be conducive to financial fragility and crises. Calvo (2005) argues, however, that last resort lending can be provided by alternative arrangements.

Impetus for official dollarization as a policy alternative was greatest at the turn of the millennium, as emerging economies had to cope with a sequence of financial and exchange rate crises while several European countries were abandoning their national currencies in favour of the newly created euro. Support for official dollarization appears to have subsided since, however.

More frequently, dollarization has emerged as a spontaneous response of domestic agents to inflation. The special case in which such a process has resulted in the dollar becoming a widespread medium of exchange is known as ‘currency substitution’. Currency substitution has been the subject of a large literature, much of it focused on the determinants of the relative demand for domestic vis-à-vis foreign currencies and on implications for monetary management. Early research followed Girton and Roper (1981) in postulating ad hoc aggregate demand functions for domestic and foreign currency, in the portfolio balance tradition. Somewhat later, Calvo (1985) derived similar demand functions from an optimizing model in which domestic and foreign currencies entered the representative household's utility function. Those approaches emphasized the possibility that increasing substitutability between the domestic and the foreign currencies would lead to monetary and exchange rate instability. However, they did not identify the basic determinants of substitutability, which was buried in the specification of the postulated demand function for foreign currency or the properties of the representative agent's utility function. Hence the early studies were of little use in understanding how to cure the ills associated with dollarization, and, in particular, they failed to trace the consequences of common policies designed to deal directly with currency substitution, such as outright prohibitions on the holdings of foreign currency. Subsequent studies have attempted to address these shortcomings by modelling more explicitly the fundamental frictions underlying currency substitution. Thus Guidotti and Rodriguez (1992) developed a cash-in-advance model of currency substitution on the assumption that using foreign currency entailed fixed transaction costs, while Chang (1994) studied the implications of a similar assumption in an overlapping generations setting. These models still left unexplained where the assumed transaction costs were coming from. Therefore, recent work on this area models currency substitution entirely from first principles, in the search theoretic tradition (see, for instance, Craig and Waller, 2004).

Another focus of recent literature has been the increased use of the dollar as the currency of denomination of the debts of domestic residents in emerging economies, a problem that Calvo (2005) terms ‘liability dollarization’. A substantial degree of liability dollarization places an economy in a vulnerable situation, since presumably many of the agents with dollar debts have assets denominated in domestic currency. Such a currency mismatch situation means that a depreciation of the domestic currency reduces the net worth of domestic agents. If, in turn, aggregate demand depends on net worth (as would be the case in the presence of financial imperfections), a currency depreciation may lead to a reduction in income and employment. In other words, liability dollarization may render depreciations

contractionary, not expansionary as assumed by conventional analysis (Aghion, Bachetta and Banerjee, 2001; Cespedes, Chang and Velasco, 2004). The combination of liability dollarization and net worth effects has been blamed for the severity of the income and output contractions in recent emerging markets crises.

At this point, no consensus exists as to the causes of liability and financial dollarization, although research on this question is rather active. Ize and Levy-Yeyati (2003), in particular, have examined the choice of currency denomination of assets and liabilities from a capital asset pricing model (CAPM) perspective, while Jeanne (2005) models liability dollarization as the private sector response to the lack of credibility in monetary policy. Finally, several studies estimate how measures of financial dollarization depend empirically on other characteristics of an economy. For example, Arteta (2005) has found that the dollarization of bank deposits is empirically more frequent in countries with a higher degree of exchange rate flexibility.

See Also

- currency unions
- money

Bibliography

Aghion, P., Bachetta, P. and Banerjee, A. 2001. Currency crises and monetary policy in an economy with credit constraints. *European Economic Review* 45, 1121–50.

Arteta, C. 2005. Exchange rate regimes and financial dollarization: does flexibility reduce currency mismatches in bank intermediation? *Topics in Macroeconomics* 5(1).

Calvo, G.A. 1985. Currency substitution and the real exchange rate: the utility maximization approach. *Journal of International Money and Finance* 4, 175–88.

Calvo, G.A. 2005. Capital markets and the exchange rate with special reference to the dollarization debate in Latin America. In *Emerging Capital Markets in Turmoil*, ed. G. Calvo. Cambridge, MA: MIT Press.

Cespedes, L., Chang, R. and Velasco, A. 2004. Balance sheets and exchange rate policy. *American Economic Review* 94, 1183–93.

Chang, R. 1994. Endogenous currency substitution, inflationary finance, and welfare. *Journal of Money, Credit, and Banking* 26, 903–16.

Chang, R. and Velasco, A. 2002. Dollarization: analytical issues. In *Dollarization*, ed. E. Levy-Yeyati and F. Sturzenegger. Cambridge, MA: MIT Press.

Craig, B. and Waller, C. 2004. Dollarization and currency exchange. *Journal of Monetary Economics* 51, 671–89.

Girton, L. and Roper, D. 1981. Theory and implications of currency substitution. *Journal of Money, Credit, and Banking* 13, 12–30.

Guidotti, P.E. and Rodriguez, C.A. 1992. Dollarization in Latin America: Gresham's law in reverse? *IMF Staff Papers* 39, 518–44.

Ize, A. and Levy-Yeyati, E. 2003. Financial dollarization. *Journal of International Economics* 59, 323–47.

Jeanne, O. 2005. Why do emerging economies borrow in foreign currency? In *Other People's Money*, ed. B. Eichengreen and R. Hausmann. Chicago: University of Chicago Press.

Ortiz, G. 1983. Currency substitution in Mexico: the dollarization problem. *Journal of Money, Credit and Banking* 15, 174–85.

How to cite this article

Chang, Roberto. "dollarization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000181> doi:10.1057/9780230226203.0403

The New Palgrave Dictionary of Economics Online

Domar, Evsey David (1914 – 1997)

E. Cary Brown

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Domar, E. D.; economic growth; Harrod–Domar growth model; portfolio theory; proportional income tax; slavery; technical change

Article

Domar (Domashevitsky) was born in 1914 in Lodz, Russia (now Poland), spent most of his early life in Harbin, Manchuria, and moved permanently to the United States in 1936. His undergraduate degree in economics (1939) was from the University of California (Los Angeles); his graduate work was at the Universities of Michigan (MA, Mathematical Statistics) and Harvard (Ph.D., 1947), where he studied with Alvin Hansen, the leading American Keynesian and most important single intellectual influence on Domar. Domar is best known for his leadership role, along with Roy Harrod, in the initiation of modern growth theory.

His first position was with the research staff of the Board of Governors of the Federal Reserve System, where he worked on fiscal problems from 1943 to 1946. His subsequent academic career took him briefly to the Carnegie Institute of Technology, the Cowles Foundation and the University of Chicago, the Johns Hopkins University in 1948 for ten years, and the Massachusetts Institute of Technology in 1958, from which he retired in 1984. An avid traveller, he held more than a dozen visiting professorships in universities at home and abroad.

While the claim to the earliest statement of the famous Harrod–Domar growth model was clearly Harrod's (1939), Domar arrived independently at a structurally similar model but from a different point of view (1946; 1947). By incorporating into static Keynesian analysis the capacity changes associated with investment, he found that steady-state capacity growth required investment to grow at a rate equal to the savings rate multiplied by the capital–output ratio. From this simple beginning, growth theory took off to become a major focus, one might almost say obsession, of the profession in the 1950s and 1960s. Domar also made important contributions to some of its conceptual and measurement problems, such as the proper treatment of depreciation (1953) and the measurement of technological change (1961), and he coined the term ‘residual’ for the fraction of expanding output unexplained by the

contribution of factors of production.

In fiscal theory, his early investigation, with Richard Musgrave (1944a), of the effect of a proportional income tax, with and without loss offsets, on portfolio choice was very similar in style and approach to portfolio theory of a decade later. Given individual preferences, the portfolio decision was modelled as a choice between alternative portfolios weighing their expected net returns against their risks (expected losses). The unconventional conclusion was reached that, given risk aversion, the imposition of a proportional income tax with symmetrical treatment of gains and losses would induce individuals to adjust their portfolios towards riskier assets. The reminder that expected risks and yields are both reduced by an income tax was an important correction to a simplistic focus on yields alone.

As an applied theorist, Domar had the knack of getting important results with simple theory. At a time when deficit finance was harshly criticized for increasing the debt burden and tax rate, Domar showed (1944b) that in a growing economy even continuous deficit finance resulted in only limited debt–income ratios and tax rates. Second, he made a fertile historical hypothesis (1970) – that the economic basis for the introduction of serfdom (or slavery) was a low land-to-labour cost. Third, he ingeniously modified the administrative rules that guided the behaviour of collective farms (1966) or that determined the compensation of socialist managers (1974) to induce them towards more efficient price–output decisions.

Domar's work was informed by a rare combination of historical, empirical and theoretical breadth. His profound scholarship, in several languages, periods, and areas, often resurrected important findings of earlier writers previously overlooked.

Selected works

1944a. (With R.A. Musgrave.) Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58, 388–422.

1944b. The burden of the debt and the national income. *American Economic Review* 34, 798–827. Reprinted in Domar (1957).

1946. Capital expansion, rate of growth, and employment. *Econometrica* 14, 137–47. Reprinted in Domar (1957).

1947. Expansion and employment. *American Economic Review* 37, 34–55. Reprinted in Domar (1957).

1948. The problem of capital accumulation. *American Economic Review* 38, 777–94. Reprinted in Domar (1957).

1952. Economic growth: an econometric approach. *American Economic Review, Papers and Proceedings* 42, 479–95. Reprinted in Domar (1957).

1953. Depreciation, replacement and growth. *Economic Journal* 63, 1–32 Reprinted in Domar (1957).

1957. *Essays in the Theory of Economic Growth*. New York: Oxford University Press.

1961. On the measurement of technological change. *Economic Journal* 71, December, 709–29.

1966. The Soviet collective farm as a producer cooperative. *American Economic Review* 56, 734–57.

1970. The causes of slavery or serfdom: a hypothesis. *Journal of Economic History* 30, 18–32.

1974. On the optimal compensation of a socialist manager. *Quarterly Journal of Economics* 88, 1–18.

Bibliography

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49, March, 14–33, Errata, June 1939, 377.

Harrod, R.F. 1948. *Towards a Dynamic Economics. Some Recent Developments of Economic Theory and their Applications to Policy*. London: Macmillan.

How to cite this article

Brown, E. Cary. "Domar, Evsey David (1914–1997)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000182> doi:10.1057/9780230226203.0404

The New Palgrave Dictionary of Economics Online

Dorfman, Joseph (1904 – 1991)

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Dorfman, J.; history of economic thought; Veblen, T.

Article

Historian of American economic thought, Dorfman was born in Russia in 1904 and educated at Reed College and at Columbia University, where he earned a Ph.D. degree in 1935 and taught from 1931 until his retirement 40 years later. Dorfman was a student of Clarence Ayres at Reed, and of Wesley C. Mitchell and John Maurice Clark at Columbia. Mitchell in turn had been a student of Thorstein Veblen. These four economists, all with institutional leanings, stand out among the formative influences that affected Dorfman's early career. He made Veblen the subject of his doctoral dissertation, which was published under the title *Thorstein Veblen and His America* in 1934. This was at the time the only book-length appraisal of a modern economist that gave close attention not only to the subject's writings but also to biographical detail, the contemporary climate of opinion, and the general social and cultural setting of the work.

This type of holistic approach is characteristic also of Dorfman's monumental *The Economic Mind in American Civilization*, a five-volume work that he published from 1946 to 1959. It is dedicated ‘To the pioneering spirit of Thorstein Veblen and the first-born of his intellectual heirs, Wesley C. Mitchell’. The work is a detailed history of American economic thought from colonial times to 1933, the first of its kind and not likely to be replaced for many years. It is based on extensive research and in many instances provides the first comprehensive account of a writer's life and work. Dorfman sees a break of emphasis in the history of American economic thought at the time of the Civil War: it was commerce before, and industry later. He notes with respect the achievements of the past, and is a critical but tactful chronicler of past foibles. He was a pioneer in exploring not only the printed page but also archival material made up of ‘papers’, ‘letters’, and similarly elusive sources of information, the first writer to do so on a large and systematic scale in the history of economic thought.

Selected works

1934. *Thorstein Veblen and His America*. New York: Viking Press.

1935. (With R.G. Tugwell.) *William Beach Lawrence: Apostle of Ricardo*. New York, reprinted from *Columbia University Quarterly*, September, 1935.

1940. *The Economic Philosophy of Thomas Jefferson*. New York: Academy of Political Science.

1946–59. *The Economic Mind in American Civilization*. 5 vols. New York: Viking Press.

1954. Introduction to Adams, H.C., *Relation of the State to Industrial Action, and Economics and Jurisprudence. Two Essays*, ed. J. Dorfman. New York: Columbia University Press.

1960. (With R.G. Tugwell.) *Early American Policy: Six Columbian Contributors*. New York: Columbia University Press.

1967–9. Introduction to W.C. Mitchell, *Types of Economic Theory: From Mercantilism to Institutionalism*, ed. J. Dorfman. New York: A.M. Kelley.

How to cite this article

Spiegel, Henry W. "Dorfman, Joseph (1904–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000188> doi:10.1057/9780230226203.0405

The New Palgrave Dictionary of Economics Online

Dornbusch, Rudiger (1942 – 2002)

Kenneth Rogoff
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Rudiger Dornbusch was one of the leading researchers in international macroeconomics in the late 20th century. He introduced the influential concept of exchange rate ‘overshooting’ to explain the excessive volatility of exchange rates after the break-up of the Bretton Woods system of fixed exchange rates in the early 1970s. Along with Stanley Fischer and Paul Samuelson, he revived the Ricardian theory of international trade whereby trade was driven by differences in technology; their simple tractable framework became similarly influential in the study of international trade.

Keywords

comparative advantage; Dornbusch, R.; exchange rate dynamics; exchange rates (floating vs. fixed); Fischer, S.; Hecksher-Ohlin framework; new open economy macroeconomics; non-traded goods; overshooting; Ricardian trade theory; Samuelson, P.; sticky prices; value at risk (VAR)

Article

Rudiger Dornbusch was born in Germany on 8 June 1942. He received his Licence es Sciences Politiques from the University of Geneva in 1966, and his Ph.D. in Economics from the University of Chicago in 1971. He was an assistant professor at the Department of Economics at the University of Rochester from 1972 to 1974, an associate professor at the Graduate School of Business at Chicago University from 1974 to 1975, and a member of the MIT Department of Economics from 1975 to 1978. He became a Professor of Economics at MIT in 1978. From 1984 until his death from cancer on 25 July 2002, he was Ford International Professor of Economics at MIT.
Dornbusch was, by any measure, one of the giants of late 20th century international macroeconomics. His celebrated *Journal of Political Economy* paper ‘Expectations and exchange rate dynamics’ (1976), which introduced the concept of exchange rate ‘overshooting’, became the workhorse of international macroeconomics over the ensuing two decades. His *American Economic Review* paper (with Stanley

Fischer and Paul Samuelson) ‘Comparative advantage, trade and payments in a Ricardian model with a continuum of goods’ (1977) introduced a simple tractable framework that became similarly influential in the study of international trade.

This entry begins by reviewing Dornbusch's two most important scientific contributions, and goes on to give a brief sketch of his broader influence on the profession through students (he served as an advisor on over 125 doctoral dissertations), through his leading intermediate textbook *Macroeconomics* (written with Stanley Fischer), and through his role as an important voice in the public policy debate.

Exchange rate overshooting

Dornbusch's overshooting model of exchange rates (1976) captured the imagination of policymakers and academics alike during the early years of floating exchange rates. The model attracted enormous attention because, after the break-up of the Bretton Woods system of fixed exchange rates in the early 1970s, exchange rates seemed far too volatile relative to the underlying fundamentals. Although subsequent empirical work has undermined the model's original bold claim to explain floating exchange rates (see Meese and Rogoff, 1983), the model is still viewed as relevant, especially during episodes of major shifts in monetary policy. In fact, an informal survey conducted by Alan Deardorff of eight top economics departments found that, as late as 1990, Dornbusch's overshooting model was the only paper taught in every one of their graduate international finance courses.

The idea of overshooting is so simple and elegant that the small-country version can be illustrated with just a couple of equations (the analysis here draws on Rogoff, 2002). The assumption of ‘uncovered interest parity’ relates the home nominal interest rate to the exogenous foreign nominal interest rate and the expected rate of depreciation of the exchange rate:

$$i_t = i_t^* + E(e_{t+1} - e_t) \tag{1}$$

where i_t is the level home nominal interest rate and e_t is the logarithm of the exchange rate (the home currency price of foreign currency), so that $E(e_{t+1}-e_t)$ is the expected rate of change in the exchange rate. The second key relationship is a money demand equation that relates the real balances to the nominal interest rate.

$$m_t - p_t = -\lambda i_t + \eta y_t \tag{2}$$

where y denotes the log of output, m is the nominal money supply and p is the price level. Higher

interest rates lower the demand for real balances, and an increase in output raises it. Dornbusch posed the question of what would happen if there were a one-time permanent increase in the money supply, m. If prices were fully flexible, it would be possible to maintain equilibrium in the above two equations by having prices and exchange rates all rise permanently in proportion to the increase in the money supply. In this case, money would be neutral and have no real effects.

In reality, however, while asset markets (including the exchange rate) adjust very quickly, goods markets adjust more slowly partly due to temporary price rigidities. Therefore, in this set-up money is neutral only in the long run (in which the price level rises proportionately to the money supply). But with goods markets clearing only slowly, what is the impact of a money shock on exchange rates and interest rates? Assume that output, y, is also fixed. If domestic prices are constant, then a rise in the money supply implies a rise in real balances, $m - p$. But this means that the home nominal interest rate i must fall, so there is a corresponding rise in the demand for real balances. Then, however, the uncovered interest parity equation (eq. (1) above) implies that e_t *must fall, or depreciate, relative to expectations of* e_{t+1}. That is, after any initial movement of the exchange rate in response to an unexpected shock, the currency must subsequently be expected to appreciate. But recall that in the long run, even with sticky prices, money is still neutral, so the exchange rate has to depreciate by the same amount as the rise in the domestic price level, thus producing no real effect.

How is all this possible? The answer, Dornbusch deduced, is that the initial money shock must cause the exchange rate to depreciate by more in the short run than it does in the long run. It ‘overshoots’. Therefore, Dornbusch's model offered a highly plausible explanation of why exchange rates seem to be so volatile relative to fundamentals. At one level of abstraction, of course, ‘overshooting’ is an application of Paul Samuelson's ‘Le Chatelier's principle’ theorem: when prices in some markets are inflexible in the short run, prices in others may overreact in the short run. But Dornbusch's model did much more than innovatively contrast the fast adjustment of asset markets with the slow adjustment of goods markets (an insight that any realistic short-run dynamic macroeconomic model should take into account). It offered a concrete and coherent analysis of an extremely important practical phenomenon. Over the decades since Dornbusch's article appeared, the term ‘overshooting’ has become deeply woven into the popular economic lexicon.

Modern research has advanced considerably beyond the overshooting model, of course, and the Mundell-Fleming-Dornbusch model has largely been supplanted by ‘new open economy macroeconomics’ (see Obstfeld and Rogoff, 1996). And the notion of looking at money shocks via a money demand equation has increasingly been supplanted by frameworks which view the overnight interest rate as the key instrument of monetary policy. Nevertheless, these newer frameworks typically include sticky prices – perhaps the most fundamental, and controversial, element of Dornbusch's model – and hence can all replicate a similar phenomenon to ‘overshooting.’

Although Dornbusch's overshooting paper was his best-known work, with over 900 citations in refereed journals, he published numerous other very well-known articles, including his 1973 *American Economic Review* paper that was among the first to incorporate non-traded goods in a monetary model (see also his elegant 1974 contribution to the collection edited by Robert Aliber), his 1983 *Journal of Political Economy* paper that illustrated how changes in the real interest rate could affect exchange rates and current accounts, and his 1987 *American Economic Review* paper that demonstrated a link between market structure and the adjustment of relative prices to exchange rate movements. Without doubt, however, his other extremely influential paper was not in international finance but in trade.

Ricardian model of trade

Dornbusch's 1977 *American Economic Review* paper with MIT colleagues Stanley Fischer and Paul Samuelson almost single-handedly revived the analysis of Ricardian trade; a ‘Ricardian’ model of trade is one with only one factor of production (usually taken to be labour). Trade is driven by differences in technology. The Ricardian model is contrasted with the Hecksher–Ohlin framework, where countries have identical technologies but different relative endowments of the factors of production (labour and capital, in the simplest canonical case). Prior to Dornbusch–Fischer–Samuelson (DFS), the Ricardian approach had been dormant for years, having been largely supplanted by the Hecksher–Ohlin framework. The Ricardian model had lost out not so much because of poor empirical results but because it had come be viewed as intractable for all but illustrative purposes. By introducing a continuum of goods (rather than a discrete number), DFS were able to analyse elegantly a broad range of comparative static questions that had previously seemed unapproachable. DFS showed, for example, how to mobilize the combination of comparative advantage and trade costs to endogenize the dividing line between ‘traded’ and ‘non-traded’ goods, and how to analyse the classic ‘transfer’ problem where one country owes debt to another. Although at first only a trickle of papers followed DFS, the power of their continuum specification has led to a recent explosion of related research. DFS have become the starting point for a number of applied papers (see, for example, Copeland and Taylor, 1994). In addition, DFS form the basis for a broad range of empirical papers (see, for example, Eaton and Kortum, 2002; Kehoe and Ruhl, 2002; Kraay and Ventura, 2002; Kei-Mu Yi, 2003; Ghironi and Melitz, 2005; see also Feenstra and Hanson, 1996). As the empirical work following DFS deepens, it is fair to say that trade economists have increasing faith in the fundamental underpinnings of the model.

Broader contributions

Aside from his path-breaking research, Dornbusch made important contributions to economics in a number of other dimensions. His intermediate undergraduate textbook with Stanley Fischer, *Macroeconomics*, written in the mid-1970s, became a worldwide best-seller. The book was really the first to integrate modern supply-side economics into the standard demand-driven framework of the day. As such, students were able to gain a far deeper understanding of problems such as the effects of oil price shocks.

Dornbusch was enormously influential as a graduate teacher at MIT. At his regular early-morning international economics ‘breakfasts’, Dornbusch would dissect recent models and serve up provocative questions in a fast-paced freewheeling style; many students remember these unique seminars as their most influential experiences as Ph.D. students. Dornbusch served as thesis advisor to scores of economists (as noted earlier, more than 125 in all), including Jeffrey Frankel, Paul Krugman, Maurice Obstfeld and Kenneth Rogoff. His dynamic, Socratic lecturing style also attracted students from outside MIT to his advanced graduate classes, including the likes of Jeffrey Sachs and Lawrence Summers. Many Dornbusch students went on to become finance ministers and heads of central banks throughout the world.

Through clear and incisive policy analysis embodied in editorials, speeches, and private meetings,

Dornbusch exercised an enormous influence on global macroeconomic policy. He was a frequent guest of leading government officials throughout the world, who greatly valued and respected his advice. Arguably, no other recent economist has had so great an impact on the global macroeconomic policy debate, especially in emerging markets such as Brazil, Korea and Mexico, but also in more advanced countries such as Italy and Germany. Notably, in his later writing he succeeded in drawing ever more concrete insights from contemporary academic research, displaying a magnificent ability to translate complex theoretical models into ideas of immediate practical relevance. For example, his 1994 Brookings paper (with Alejandro Werner) argued that Mexico's pegged exchange rate had become overvalued to an extent that was unsustainable. Dornbusch's comments on markets prior to the currency collapse at the end of 1994 were highly influential. He also advanced a number of innovative ideas for dealing with international debt problems. His policy analysis was notable in that he managed to adopt strong views while continuing to be perceived as an independent and objective thinker. Over the last ten years of his life, Dornbusch became especially well-known for his monthly ‘Economic Perspectives’ newsletter, which covered with panache a broad range of topical global economic problems. One innovative idea, first developed in the newsletter and then formally published in his ‘Primer on Emerging Market Crises’ (2002) was to apply ‘value at risk’ analysis to the balance sheet of a country. In his primer, he wrote:

> …the right answer to crisis avoidance is controlling risk. The appropriate conceptual framework is *value at risk (VAR)* – a model-driven estimate of the maximum risk for a particular balance sheet situation over a specified horizon. There are surely genuine issues with the specifics of VAR surrounding modelling as has been widely discussed with respect to bank risk models used for meeting BIS requirements. But just as surely there is no issue whatsoever in recognizing that this general approach is the right one. If authorities everywhere enforced a culture of risk-oriented evaluation of balance sheets, extreme situations such as those of Asia in 1997 would disappear or, at the least, become a rare species. (2002, pp. 743–54)

In this short space it has not been possible to do full justice to the range and breadth of Dornbusch's contributions. But I hope the reader has gained some perspective on why he will have a lasting influence.

See Also

- comparative advantage
- exchange rate volatility
- extremal quantiles and value-at-risk
- neo-Ricardian economics

Selected works

1973. Devaluation, money and nontraded goods. *American Economic Review* 63, 871–80.

1974. Real and monetary aspects of the effects of exchange rate regime changes. In *National Monetary Policies and the International Financial System*, ed. R. Aliber. Chicago: University of Chicago Press.

1974. Tariffs and nontraded goods. *Journal of International Economics* 4, 177–85.

1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84, 1161–76.

1977. (With S. Fischer and P. Samuelson.) Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67, 823–39.

1980. *Open Economy Macroeconomics*. New York: Basic Books.

1983. Real interest rates, home goods and optimal external borrowing. *Journal of Political Economy* 91, 141–53.

1987. Exchange rates and prices. *American Economic Review* 77, 93–106.

1990. (With S. Fischer.) *Macroeconomics*, 5th edn. New York: McGraw-Hill.

1994. (With A. Werner.) Mexico: stabilization, reform and no growth. *Brookings Papers on Economic Activity*, 1994: 1, 253–315.

2002. A primer on emerging market crises. In *Preventing Currency Crises in Emerging Markets*, ed. S. Edwards and J.A. Frankel. Chicago: The University of Chicago Press.

Bibliography

Copeland, B. and Taylor, M. Scott. 1994. North-South trade and the environment. *Quarterly Journal of Economics* 109, 755–87.

Eaton, J. and Kortum, S. 2002. Technology, geography, and trade. *Econometrica* 70, 1741–79.

Feenstra, R. and Hanson, G. 1996. Globalization, outsourcing, and wage inequality. *American Economic Review* 86, 240–5.

Ghironi, F. and Melitz, M. 2005. International trade and macroeconomic dynamics with heterogeneous firms. Mimeo. Harvard University.

Kehoe, T. J. and Ruhl, K. J. 2002. How important is the new goods margin in international trade? Mimeo. University of Minnesota.

Kei-Mu, Yi. 2003. Can vertical specialization explain the growth of world trade? *Journal of Political Economy* 111, 52–102.

Kraay, A. and Ventura, J. 2002. Trade integration and risk sharing. *European Economic Review* 46, 1023–48.

Meese, R. and Rogoff, K. 1983. Empirical exchange rate models of the seventies: do they fit out of sample? *Journal of International Economics* 14, 3–24.

Obstfeld, M. and Rogoff, K. 1996. *Foundations of International Macroeconomics*. Cambridge, MA: MIT Press.

Rogoff, K. 2002. Dornbusch's overshooting model after 25 years: IMF Mundell-Fleming Lecture. *International Monetary Fund Staff Papers* 49 (Special Issue), 1–35 (including remarks by Rudiger Dornbusch).

How to cite this article

Rogoff, Kenneth. "Dornbusch, Rudiger (1942–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000237> doi:10.1057/9780230226203.0406

The New Palgrave Dictionary of Economics Online

double-entry bookkeeping

Basil S. Yamey
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Double-entry bookkeeping is a system for arranging and organizing accounting information. It requires that each transaction (or other change) recorded in the accounting system must be recorded twice, and for the same money amount, once in debit form and once in credit form. Because it is concerned with the organization of information rather than with the scope and detail of that information, the system of double-entry bookkeeping is highly adaptable. It neither generates nor requires any particular set of valuation rules or profit concepts, and it is compatible with different treatments for changes in the value of money.

Keywords

accounting; assets and liabilities; balance-sheet equation; double-entry bookkeeping; national accounting; Pacioli, L.; Sombart, W.; Spengler, O.; transaction analysis; van Gezel, W.

Article

Firms of all kinds need, in different degrees, to maintain records of their transactions with other firms and persons, of the debts they owe or are owed, and of their assets. The records they keep for this purpose constitute their accounting records. Traditionally they have consisted of account-books of various kinds, but they can take the form also of magnetic tapes and so on. If the records are kept on a systematic basis, one can speak of an accounting system. From the accounting records one can prepare a variety of accounting statements in which the detailed accounting information is rearranged, regrouped and presented in summary form. The balance sheet and the profit-and-loss (or income) account or statement are important examples of such accounting statements.
Double-entry bookkeeping is a system or method for the arrangement and classification of accounting information. It developed in Italy, possibly in the second half of the 13th century. A description of the system was first published in Venice in 1494 as one part of a famous compendium of mathematical and commercial information: Luca Pacioli's *Summa de Arithmetica Geometria Proportioni et*

Proportionalità. Knowledge of the double-entry system spread gradually from Italy to the rest of Europe by way of commercial contacts, schools and published treatises. It is not possible to establish how widely the system was used by merchants and others, say, in the 18th century. But by the late 19th century it had become the standard system for accounting records. Today it is used by virtually all corporate enterprises and many other firms as well as non-profit-making organizations in the West and also elsewhere. It has also proved suitable to serve as a useful scaffolding for the construction of the national income and related accounts for countries or regions.

Double-entry bookkeeping is no more than a system for arranging and organizing accounting information. It does not itself define the scope and detail of that information. Thus, for example, the double-entry system does not require that all transactions with third parties should be recorded, although it is the convention now to record all of them. What is more important, it does not prescribe which occurrences or changes that do not involve external transactions should be recorded in the accounts. Thus it does not prescribe whether changes in the value of the firm's assets should be recorded, how they should be determined, or how they should be recorded. Double entry neither generates nor requires any particular set of valuation rules or profit concepts. Different valuation bases or conventions, and different treatments for changes in the value of money, are all compatible with the use of the double-entry system. The system itself is highly adaptable, since it is concerned with arrangement and organization rather than with scope and content. Its adaptability has made it possible for it to serve as the basis for arranging the records needed by the relatively small-scale merchants in the early modern period of economic expansion as well as for those of the largest corporate enterprises operating today. But this does not mean that asset values were recorded and profits calculated in the same way by 17th-century merchants as they are by today's corporate enterprises. In fact, 17th-century merchants used several alternative bases for recording changes in asset values. And some of these would not be used by companies today.

Moreover, although all the companies within the same jurisdiction are subject to the same laws and the same institutional constraints (for example, those imposed by the stock-market authorities and those reflecting professional accounting standards), there is still scope for considerable variation in the determination and statement of accounting profits and asset values. However, because of developments in legislation and in the other constraining forces operating on corporate enterprises, it is no longer the case that a company chairman in the United Kingdom would be able to say (as Arthur Chamberlain, chairman of Tube Investments said in 1935) that he ‘would almost undertake to draw up two balance-sheets for the same company, both coming within an auditor's statutory certificate, in which practically the only recognizable items would be the name and the capital authorised and issued’.

Double entry requires that each transaction (or other event) recorded in the accounting system must be recorded twice, and for the same money amount, once in debit form and once in credit form. In double entry, as Pacioli expressed it, ‘all the entries placed in the ledger must be double, that is if you make a creditor (entry) you must make a debtor (entry)’. The debit and credit entries are made in the ledger, on the basis of the information entered in preliminary records. The ledger, which may for convenience be subdivided into a series of specialized ledgers, consists of a number of ledger accounts, pertaining, for example, to particular debtors or creditors, particular assets or particular categories of expenditure. It is the convention that the debit entry is made on the left-hand (debit) side of the appropriate ledger account, and the corresponding off-setting credit entry on the right-hand (credit) side of the other appropriate ledger account.

The duality of entries for each transaction (or other recorded event) ties together the ledger accounts into an interlocking system of recorded information. Moreover, as each transaction gives rise to two equal but opposite entries, the system of accounts (if properly kept) is always in balance or equilibrium. The total of debit entries must be equal to the total of credit entries. Similarly, the total of the balances on all ledger accounts that have debit balances must be equal to the total of the balances on all the remaining ledger accounts that have credit balances. (If debit balances are taken as positive amounts and credit balances as negative amounts, the algebraic sum of the balances on all ledger accounts is zero.) The equality of debits and credits is the basis for the trial balance. This is a list of the balances on all open (that is, unbalanced) accounts in the ledger, distinguishing between debit and credit balances. If the trial balance does not balance, there is some error in the ledger. Postlethwayt in his Dictionary (1751) wrote of the ‘agreeable satisfaction’ of getting a trial balance to balance, and said that the trial balance will ‘shew you that this [double entry], of all methods, is the most excellent’. The fact that a trial balance does not balance is proof that the ledger does contain some error. The converse is, of course, not correct.
Roger North, son of the prominent Turkey merchant Sir Dudley North, wrote in 1714 as follows: ‘The making true Drs. (debtors) and Crs. (creditors) is the greatest Difficulty of Accompting, and perpetually exerciseth the Judgment; being an Act of the Mind, intent upon the Nature and Truth of Things.’ Writers of instructional books on bookkeeping and accounts through the centuries have devised various lists, rules or approaches to help the accountant decide which debit and credit entries he should make for the various categories of transaction.
An early rule, widely used, was as follows (taken from a verse, ‘Rules to be Observed’, in a book of 1553 by James Peele):

> To make the thinges Received, or the receiver,
> Debter to the thinges delivered, or to the deliverer.

This rule is obviously readily applicable to many categories of transaction. If cash is received from a debtor, debit the cash account; and credit the debtor's account. If office furniture is bought on credit, debit the furniture account; and credit the supplier's account. If the owner withdraws cash from the business, debit the capital (that is, owner's) account; and credit the cash account. But it is evidently a straining of the language to say, when an amount is written off the book value of, say, a ship, in order to reflect diminution of value due to wear and tear, that the profit-and-loss account, which is to be debited, ‘receives’ something that has been ‘delivered’ to it by the ship account. Teachers and textbook writers not surprisingly looked for a rule that is robust enough to cover comfortably all transactions and events to be recorded, and to indicate unambiguously in each case where the debit and where the credit are to be placed.
The most common rule or approach adopted today in transaction analysis in the double-entry system derives from the so-called balance-sheet equation. The earliest formulation of this approach can be traced to the work of a Dutchman, Willem van Gezel, published in 1681.

> The basic balance-sheet equation is:
> Owner's Equity
> (or the firm's net worth)=Assets – Liabilities=Net Assets; or Owner's Equity
> +Liabilities=Assets.

The ledger contains accounts for the various assets and liabilities; and there are accounts in it for the capital contributed or withdrawn by the owner(s) and for any increases (decreases) in ‘net worth’ resulting from the activities of the firm. In the double-entry system, increases in assets are indicated by debits to an asset account – the extent to which assets are subdivided into separate ledger accounts is for each firm to decide. Conversely, decreases in assets are recorded as credits to asset accounts. The total of a firm's assets is represented by the total of claims on those assets; namely, its liabilities (that is, its debts to third parties) and its owner's equity. The total of these claims must be a credit amount that equals the debit amount representing the assets. An increase (decrease) in a claim is therefore represented by a credit (debit) in a liability account or an equity account. (Again, the extent to which claims are subdivided into various ledger accounts is a matter for each firm to decide. As regards the equity element, it is common for a ledger to contain separate accounts for each major category of business expenditure and income, a trading account, perhaps subdivided by type of activity, for showing the gross profit, and a profit-and-loss account to bring together the results from all the subordinate ledger accounts.)

Transaction analysis follows readily. The payment of salaries reduces the asset ‘cash’ and reduces the owner's equity, since the payment, taken by itself, represents a loss to the firm: hence, debit the salaries (eventually, profit-and-loss) account; and credit the cash account. The depreciation of an asset likewise reduces an asset and reduces the equity: debit the depreciation account (eventually profit-and-loss) account; and credit the ship account.

As has already been emphasized, the double-entry system does not itself dictate whether or in what circumstances increases or decreases in assets are to be recognized in the accounts. Neither does the system dictate the basis on which, or the circumstances in which, assets are to be revalued in the accounts. Decisions of these kinds are accounting decisions; and whenever such decisions are taken, the double-entry system of recording will accommodate them in accordance with its own logical structure. It follows from this that, although the value of the owner's equity in the ledger will always be equal to the value of the firm's net assets (that is, assets minus liabilities to those outside the firm) as stated in the accounts, those two values depend on the bases on which the values of assets are stated in the accounts. Subject to this crucial qualification, it follows from the equilibrium feature of the double-entry system that the change (increase or decrease) in the value of the net assets of a firm over a period will be reflected as entries in the various ledger accounts that represent the owner's equity. Those entries in the various equity accounts that relate to the firm's operations, when they are brought together in the profit-and-loss account, yield a balance that is equal to the change in the value of the net assets over the period. It is the profit (loss) for the period. This profit is equal to the change in the value of the net assets over the period (allowance being made for any contributions or withdrawals of assets by the owner). It may be noted that the same profit figure would be established if one took the difference between the totals of two inventories of the firm's net assets taken, respectively, at the beginning and at the end of the period, provided that the same valuations were used and the same allowance made for the owner's contributions and withdrawals. The method of profit calculation by means of successive inventories of assets and liabilities was widely used in the past. The surviving 16th-century records of the large-scale commercial, financial and mining enterprise of the Fugger family of Augsburg provide examples of this procedure.

The equality – Profit (Loss)=Change in Net Assets – evidently holds only if all the changes recorded in asset and liability accounts (other than the owner's contributions or withdrawals) are also recorded in

equity accounts that, in turn, are closed into the profit-and-loss account. In contemporary corporate financial accounting it is permissible to allow the counter-entries representing certain changes in asset values, depending upon the circumstances, to bypass the profit-and-loss account (for example, by recording these changes as debits or credits to one or other reserve account). This practice breaks the nexus between changes in net asset values and profits. It does, however, allow more ‘realistic’ values to be used in asset accounts where, otherwise, their use might produce ‘distortions’ in the profit figures that could mislead users such as investors and investment advisers. Both ‘realistic’ and ‘distortions’ are words that give rise to much debate in accounting circles. The double-entry recording system can accommodate the practice of bypassing the profit-and-loss account as comfortably as it can the alternative. The system itself imposes no discipline or constraint upon accountant or management – except the constraint that for each transaction or change recorded in the firm's accounting system, equal but offsetting debit and credit entries have to be made in accounts in the ledger.

The German economic historian, Werner Sombart, claimed that ‘capitalism without double-entry bookkeeping is simply inconceivable’, and that double-entry was one of the most significant inventions or creations of the human spirit. In similar vein, Oswald Spengler asserted that the creator of double-entry bookkeeping could take his place worthily beside his contemporaries Columbus and Copernicus. These scholars evidently attributed to the double-entry system a role that goes well beyond what one might think appropriate to ascribe to a system of organizing and arranging accounting data. In a nutshell, Sombart argued that, historically, the double-entry system opened up possibilities and provided stimuli that enabled capitalism to develop fully. It clarified the acquisitive ends of commerce and provided the rational basis on which this acquisition could be carried on. It provided the basis for the continued rational pursuit of profits, and virtually compelled its users to pursue the acquisition of wealth. It also enabled the firm or enterprise to be separated from its owners, thus facilitating the development of corporate enterprises.

These views are in their details either untenable or grossly exaggerated. To note only a few points: the profits of an enterprise and its capital employed can be calculated without double-entry bookkeeping; joint-stock companies, such as the Dutch East India Company, have existed and flourished without double-entry bookkeeping; 16th- and 17th-century merchants, like the Fugger, who did not use the system do not seem to have been any less acquisitive, rational and successful than those who did use the system; and the adoption of the double-entry system could not have changed, or even have reinforced, the temperament, commercial acumen, motivation or goals of those who adopted it for organizing their accounting records.

To reject grandiose claims made for double-entry bookkeeping is not to deny the more workaday usefulness of the system. A method or system for recording and classifying accounting data that has been used increasingly over a period of six centuries must indeed have substantial practical merit. Double entry is a useful and versatile method for organizing accounting data, its value increasing with the volume and complexity of the data to be organized. In turn, the efficient organization of data helps management at various levels in many ways, more notably in large organizations. But its contribution to efficiency does not proceed along the lines emphasized by Sombart.

See Also

- accounting and economics

- assets and liabilities
- Sombart, Werner

Bibliography

Yamey, B.S. 1964. Accounting and the rise of capitalism. *Journal of Accounting Research* 2, 117–36 (for a discussion of Sombart's views on double-entry bookkeeping and capitalism).

How to cite this article

Yamey, Basil S. "double-entry bookkeeping." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000189> doi:10.1057/9780230226203.0407

The New Palgrave Dictionary of Economics Online

Douglas, Paul Howard (1892 – 1976)

Colin G. Clark

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agricultural economics; climacteric (of 1896–1914); Cobb, C. W.; Cobb–Douglas functions; Douglas, P. H.; Phelps Brown, H.; real wage growth

Article

Born in 1892 in Salem, Massachusetts, Paul Douglas attended Bowdoin College in Maine (BA, 1913) and Columbia University (Ph.D., 1921). After holding a number of teaching posts between 1916 and 1920, he joined the faculty of the University of Chicago where he remained (apart from service in the Second World War) until 1948, when he became a United States Senator from Illinois. After his retirement from the Senate in 1966, he taught at the New School for Social Research for two years (1967–9).

Paul Douglas first became well known for his massive theoretical and factual studies (for example, 1930) of all the available information on wages in the United States from 1890. This work required laborious following up of old, obscure records, and repairing gaps in the available knowledge, such as domestic service wages. Douglas also collected information on prices so as to make an estimate of the movement of real wages.

In Britain there was almost complete cessation of the growth of real wages between 1896 and 1914. Understandably, it was a period of growing social tension. Sir Henry Phelps Brown called it the ‘climacteric’. We still do not really understand its cause; there was some sociological evidence about the deterioration of the quality of businessmen. D.H. Robertson found at least a partial explanation in economic causes, namely, that, of the two leading British export industries, cotton was produced under constant returns and coal under diminishing returns.

This problem remains of primary interest to economic historians, and naturally they enquire whether there is any evidence of a similar ‘climacteric’ in other countries. In Germany there was a slowing down of the rate of rise in real wages, but not very marked. Douglas's American data likewise do not show such a ‘climacteric’. Recent research, however, has thrown some doubt not on Douglas's wage data, but on his price data; and perhaps there was some slowing down of the rate of growth of real wages.

Douglas became famous to the whole economic world through the ‘Cobb–Douglas function’ (for example, 1934). Working in conjunction with Charles W. Cobb, a mathematician from Amherst College, and using Massachusetts State annual factory returns, Douglas in 1928 established the following relation: Let product be *P*, labour input *L*, capital input *C*, and *k* a constant. Then $P = kL^aC^b$. (The same formula, with land in place of capital, had already been used by Wicksell – for example, 1900 – but he gave it neither theoretical nor empirical development.)

We may, if we wish, constrain *a* and *b* to add up to 1; but we get much the same results unconstrained. If *a* and *b* add up to more than 1 this is an indication of economies of scale (increasing returns) – a uniform increase in the quantities of inputs giving a more than proportionate increase in product.

Annual data, which many economists have been using, give results mainly dependent on fluctuations in the short-period business cycle – which is not what we want at all. It is only when we have data for such a long period as to make it possible to average out the business cycle that we can draw conclusions about productivity. This has been done by Solow in the United States, Aukrust in Norway, and Niitamo in Finland. In each case it was found, in the long run, that the product was rising much more rapidly than expected from inputs and their exponents. This difference is generally held to be due to technical advance, though some look for economies of scale. Some difficult but promising work by Denison further analyses the labour input by numerous categories, male and female, adult and juvenile, and various levels of education. These methods reduce the unknown factor – but it does not disappear.

Differentiating the Cobb–Douglas formula to obtain marginal productivities, then aggregate earnings of the factors should be proportional to a:b – assuming that each factor is remunerated according to its marginal productivity. When he first made this calculation (so he told me), Douglas fully expected the aggregate income of labour to be below that indicated by its marginal productivity. He was surprised, however, to find that it was almost exactly what was to be expected – about 75 per cent of the product.

The Cobb–Douglas formula has had abundant application in agricultural economics, especially for cross-section studies, where each farm may be considered an independent piece of evidence. Land is introduced as a factor, and also data for other inputs – fertilizers, insecticides, and so on – even (in one study in Sweden) the age of the farmer – a negative factor.

Douglas was very much a political economist. Organized labour in the United States did not attempt to form a political party of its own as in Britain, but instead played the two existing parties off against each other in demanding concessions. But in the 1920s this was not fully agreed. The other element in the population with a grievance against the current state of affairs was the farmers, and an attempt was made to form a Farmer–Labour political party. Douglas took an active part in these negotiations, and was national treasurer of the organization. But with the Roosevelt reforms of the 1930s the prospects of a Farmer–Labour party died away.

Chicago had acquired a worldwide reputation for corruption and crime; and the ruling Democratic Party considered that its ‘image’ would be improved by an upright professor of economics on the city council. Douglas assured me that some improvement had taken place, though less than was hoped for. Later, the despotic Mayor Daley achieved a real reduction in crime. But once I asked Douglas whether, if I wished to set up a milk distribution business in Chicago, he could guarantee my safety. He replied that, ‘regrettably’, he could not.

Douglas was a Quaker, and in the First World War applied for exemption from military service on religious grounds. But in the Second World War he felt very differently. In spite of his age, he obtained a commission in the marines through President Roosevelt's personal intervention, and took part in the

bloody landing on Iwojima. He sustained an injury to his hand which was with him for the rest of his life. From city councillor he advanced to become Senator for Illinois. On the very day that he arrived in Washington he found a vanload of furniture which had been offered to him as a gift. He sent it back. This episode prompted him to write a little book, *Ethics in Government* (1952). He saw no harm in the small presents customarily exchanged among businessmen and politicians – calendars, cigars, and so on – but instructed his staff to return any present valued at over four dollars.

See Also

- Cobb–Douglas functions

Selected works

1928. (With C.W. Cobb.) A theory of production. *American Economic Review*, Supplement, 18, 139–65.

1930. *Real Wages in the United States, 1890–1926*. Boston and New York: Houghton Mifflin Company.

1934. *The Theory of Wages*. New York: Macmillan.

1936. *Social Security in the United States: An Analysis and Appraisal of the Federal Social Security Act*. New York, London: Whittlesey House, McGraw-Hill.

1947. (With E.H. Schoenberg.) Studies in the supply curve of labour; the relation in 1929 between average earnings in American cities and the proportions seeking employment. *Journal of Political Economy* 45, February, 45–79.

1939a. The effect of wage increases upon employment. *American Economic Review* Supplement 29, 138–57.

1939b. (With H.G. Lewis.) Some problems in the measurement of income elasticities. *Econometrica* 7, 208–20.

1939c. (With M. Bronfenbrenner.) Cross-section studies in the Cobb–Douglas function. *Journal of Political Economy* 47, 761–85.

1948. Are there laws of production? *American Economic Review* 38, 1–41.

1952. *Ethics in Government*. The Godkin Lectures at Harvard University, 1951. Cambridge, MA: Harvard University Press.

1972. *In the Fullness of Time: The Memoirs of Paul H. Douglas*. New York: Harcourt Brace Jovanovich.

1976. The Cobb–Douglas production function once again: its history, its testing and some new empirical values. *Journal of Political Economy* 84, 903–15.

Bibliography

Wicksell, K. 1900. Marginal productivity as the basis of distribution in economics. *Ekonomisk Tidskrift*. English trans. in *K. Wicksell: Selected Papers in Economic Theory*, ed. E. Lindahl. London: Allen & Unwin, 1958.

How to cite this article

Clark, Colin G. "Douglas, Paul Howard (1892–1976)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000192> doi:10.1057/9780230226203.0408

The New Palgrave Dictionary of Economics Online

Du Pont de Nemours, Pierre Samuel (1739 – 1817)

Peter Groenewegen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advisers; assignats; Du Pont de Nemours, P. S.; excise taxes; mathematical economics; Mercier de la Rivière, P.-P.; Mirabeau, V. R., Marquis de; Physiocracy; Quesnay, F.; silver; Turgot, A. R. J.

Article

Economic writer and editor. Born in Paris, he trained for various occupations including medicine and watch making. A pamphlet on taxation (1763) brought him in contact with Mirabeau and Quesnay, under whose guidance he wrote a work on the grain trade (1764). He also befriended Turgot, with whom he diligently corresponded until Turgot's death. From 1766 to late 1768 he edited the *Journal de l'Agriculture* in the Physiocratic cause, then the *Ephémérides* until 1772. During this period he also published Quesnay's economics under the title *Physiocratie* (Du Pont, 1767) and summarized Mercier (1767), adding material on the history of the new science (Du Pont, 1768). From the early 1770s he developed a career as economic adviser through correspondence with the King of Sweden and the Margrave of Baden; the correspondence with the latter was subsequently published (Knies, 1892). In 1774 he was appointed tutor to the Polish royal family. On becoming *contrôleur-général*, Turgot required his friend's assistance and Du Pont was back in Paris by early 1775. Financial compensation for loss of his royal tutorship enabled him to purchase landed property near Nemours. Turgot's dismissal from office in 1776 did not end Du Pont's career in giving official economic advice; a highlight of which is his influence on the 1786 Anglo–French Commercial Treaty. Du Pont was politically active in the French Revolution, serving from 1789 as Deputy for Nemours in the National Assembly and becoming its President during 1790; in 1794 to 1797 he was imprisoned for short periods. He migrated to the United States in 1799 but returned to Paris in 1802. From 1803 to 1810 he served in the Paris Chamber of Commerce, and in addition edited Turgot's works (Du Pont, 1808–11). In 1815 he returned to the United States and settled in Delaware, the town where his son Irenée had started the gunpowder factory from which the Du Pont chemical conglomerate developed, and where he died in 1817. Du Pont is now mainly remembered as a major propagator of Physiocracy, an early historian of economics, a pioneer in the use of diagrams in economic argument and, most importantly, as the editor of Quesnay and Turgot,

whose works he helped to preserve. An assessment of his work as economist needs to take all facets of his career into account, as the one full-length attempt at this (McLain, 1977) has in fact done.

Virtually all Du Pont's economic work is characterized by dogmatic adherence to the Physiocracy developed by Quesnay and codified by Mercier de la Rivière. Turgot criticized this ‘servitude to the ideas of the master’ as totally inappropriate in matters of science (Schelle, 1913–23, vol. 2, p. 677). Despite such criticism Du Pont allowed his dogmatism to colour excursions into the history of economics (Du Pont, 1769) and, more importantly, his preparation of Turgot's works for the press (see Groenewegen, 1977), particularly his editions of the *Reflections* (Turgot, 1766). Two examples of his more novel contributions to economics can be given. One is his use of diagrams in explaining economic policy, which Theocharis (1961, p. 60) described as the first use of a diagram by a professional economist for ‘illustrating an economic argument set out in essentially dynamic time’, thereby making Du Pont (1774) ‘the earliest French contribution of importance in mathematical economics’. The problem analysed is the price effects of an excise reduction, the benefits of which are argued to accrue ultimately to the landowning class. The excise reduction's initial income effect on manufacturers and merchants allows them either to reduce their own prices or to pay higher prices for raw materials. By assuming this increased competition for raw materials to raise their price in each period by three-fourths of the increase in the preceding period, Du Pont shows how a new equilibrium price will be reached which transfers the benefits from excise reduction to the rural sector. His proof relies on the properties of diminishing geometrical progressions which also formed the basis for much of the analysis of the *Tableau économique*. Du Pont's analysis of the inflationary consequences from issuing assignats is a second example. Although much of this is similar to Turgot's (1749) analysis, some of it is of interest in explaining Smith's version of the specie mechanism to which Du Pont (1790, p. 28) explicitly refers. Issuing paper money by assignats makes silver superfluous as a circulating medium; this drives the metal out of the country because its only other use is to be sold abroad (Du Pont, 1790, p. 42), a specie mechanism like Smith's (1776, pp. 293–4) that is independent of relative price movements. Both examples of his more original economics relate to matters of economic policy and add force to the claim by McLain (1977, p. 255) that Du Pont represents ‘the first important case of a professional economist turned government policy-maker, a tradition in which he would be followed by [many] others…’.

Selected works

1763. *Réflexions sur l’écrit intitulé: Richesses de l’état*. Paris.

1764. *De l'exportation et de l'importation des grains*. Soissons and Paris.

1767. *Physiocratie, ou Constitution Naturelle du Gouvernement le plus avantageux au genre humain*. Leyden and Paris.

1768. *De l'origine et des progrès d'une science nouvelle*. In *Physiocrates*, ed. E. Daire. Paris, 1846.

1769. *Notice abrégée des différents écrits modernes qui ont concours en France à former la science de*

l'économie politique. In *Oeuvres Oeconomiques et Philosophiques de François Quesnay*, ed. A. Oncken, Frankfurt and Paris, 1888.

1774. *On Economic Curves*, ed. H.W. Spiegel, Baltimore: Johns Hopkins Reprints of Economic Tracts, 1955.

1790. *The Dangers of Inflation*. Trans. E. E. Lincoln, Boston: Kress Library Publications, 1950.

1808–11. *Oeuvres de Turgot*. 9 vols, Paris.

Bibliography

Groenewegen, P.D. 1977. *The Economics of A.R.J. Turgot*. The Hague: Nijhoff.

Knies, K. 1892. *Carl Friedrichs von Baden Brieflicher Verkehr mit Mirabeau und Du Pont*. Heidelberg.

McLain, J.J. 1977. *The Economic Writings of Du Pont de Nemours*. Newark and London: University of Delaware Press.

Mercier de la Rivière, P.P. 1767. *L'Ordre Naturel et Essentiel des Sociétiés politiques*. Paris.

Schelle, G. 1913–23. *Oeuvres de Turgot et documents le concernant*. 5 vols. Paris: F Alcan.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R.H. Campbell and A.S. Skinner, Oxford: Clarendon, 1976.

Theocharis, R.D. 1961. *Early Developments in Mathematical Economics*. London: Macmillan.

Turgot, A.R.J. 1749. *Letter on Paper Money*. In Groenewegen (1977).

Turgot, A.R.J. 1766. *Reflections on the Production and Distribution of Wealth*. In Groenewegen (1977).

How to cite this article

Groenewegen, Peter. "Du Pont de Nemours, Pierre Samuel (1739–1817)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000194>
doi:10.1057/9780230226203.0415

The New Palgrave Dictionary of Economics Online

dual economies

David Vines and Andrew Zeitlin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Dual economies have asymmetric sectors, the interaction between which influences the path of development. These are typically a rural, traditional, or agricultural sector on one hand, and an urban, modern, or industrial sector on the other. The relevant asymmetries are not merely technological but also include institutional, behavioural, and informational aspects. Modern treatments have grown out of the work of W. Arthur Lewis, whose model was based on the existence of surplus labour in agriculture. Subsequent authors have considered the implications of alternative assumptions for the development of a dual economy.

Keywords

agriculture and economic development; asymmetric information; capital accumulation; credit markets in developing countries; development economics; disguised unemployment; dual economies; Engel's Law; Harris–Todaro hypothesis; Hecksher–Ohlin trade theory; immigration; Kuznets curve; labour surplus economies; Lewis, W. A.; Malthus's theory of population; neoclassical growth theory; outsourcing; pure surplus labour; rural–urban migration; scissors problem; subsistence; terms of trade; urban unemployment; wage differentials

Article

Dual economies have asymmetric sectors, the interaction between which influences the path of development. W. Arthur Lewis introduced this idea in his paper, ‘Economic Development with Unlimited Supplies of Labour’ (Lewis, 1954), which earned him the Nobel Prize for Economics in 1979. That paper contains two theoretical models, both designed to explain the intrinsic problems of underdevelopment. When the prize was awarded, Ronald Findlay wrote that ‘a large part of … development economics … can be seen as an extended commentary on the meaning and ramifications [of this article]’ (Findlay, 1980, p. 64). Here we focus primarily on the first of Lewis's two models of

dualism – that of a single underdeveloped economy. We describe that model, trace the evolution of the ideas which grew from it, and discuss the continuing importance of these ideas in the study of economic development.

Long before Lewis wrote his article, there had been much thinking about ‘dual’ economies, conceived of as economies with both an industrial sector and an agricultural sector. Adam Smith and David Ricardo both focused on the interaction between these sectors during the Industrial Revolution; for Ricardo the outlook for industrial growth was ‘dismal’ because of diminishing returns in agriculture (see Hicks, 1965; Pasinetti, 1974). In the early 20th century, there was an extended discussion in the Soviet Union of the ‘scissors problem’, concerning the determination of the terms of trade between these two sectors. Evgeny Preobrazhensky (1924) argued that a decrease in the relative price of agricultural goods could be used to stimulate industrial investment; others replied that sufficient agricultural goods would not be available at lower relative prices and that these goods would need to be seized by force, something which the collectivization of agriculture made possible (see Sah and Stiglitz, 1984). And during the Great Leap Forward in China in the 1950s, Chairman Mao attempted to confiscate an increasing quantity of primary goods from the Chinese countryside in order to facilitate the development of urban manufacturing. These policies led to famine and to the deaths of approximately 30 million people. Thus both theorists and policymakers have long recognized that, in an economy with two very different sectors, growth prospects hinge on how these sectors interact.

In his Nobel Prize autobiography, Lewis (1979) writes that his interest was in the ‘fundamental forces determining the rate of economic growth’. But he was not satisfied with the neoclassical model of growth that was emerging at the time (Solow, 1956; Swan, 1956), out of the work of Roy F. Harrod (1939) and Evsey D. Domar (1945). That neoclassical framework aimed to provide a *general* theory of growth. But to Lewis it seemed inadequate because it did not deal with interactions between the industrial and the agricultural sectors: in Lewis's words, this model contained no discussion ‘of what determines the relative price of steel and coffee [namely, of industrial goods and agricultural goods]. The approach through marginal utility made no sense to me. And the Heckscher–Ohlin framework could not be used, since that assumes that trading partners have the same production functions, whereas coffee cannot be grown in most of the steel-producing countries.’ Furthermore, the neoclassical theory seemed inadequate to him for historical reasons: ‘[a]pparently, during the first fifty years of the industrial revolution, real wages in Britain remained more or less constant while profits and savings soared. This could [also] not be squared with the neoclassical framework, in which a rise in investment should raise wages and depress the rate of return on capital’ (Lewis, 1979).

Then, Lewis continues:

> One day in August, 1952, walking down the road in Bangkok, it came to me suddenly that both problems have the same solution. Throw away the neoclassical assumption that the quantity of labour is fixed. An ‘unlimited supply of labour’ will keep wages down, producing cheap coffee in the first case and high profits in the second case. The result is a dual (national or world) economy, where one part is a reservoir of cheap labour for the other. The unlimited supply of labour derives ultimately from population pressure, so it is a phase in the demographic cycle. (Lewis, 1979, p. 397)

This key insight launched Lewis on the journey that led to his famous article. Spelling out the implications of his insight led him to use the term ‘dualism’ to describe economies in which there are differences between industrial and agricultural sectors that cannot be adequately explained by differences in production technologies or in factor endowments, in the manner normally used by economists.

The Lewis model

Lewis identified three such differences between industry and agriculture, which we term ‘asymmetries’ in this article (following Kanbur and McIntosh, 1987).

First, there are technological differences between the sectors. Labour is used in each sector. In agriculture it is combined with land in production, whereas industrial goods are produced by combining labour with reproducible capital. Moreover, industrial goods can be consumed or invested, whereas agricultural goods can only be consumed.

Second, there are organizational differences between the sectors. The large, rural agricultural sector functions on traditional lines and is primarily based on subsistence; industrial production happens in a modern, market-oriented sector, located in towns and cities. There is ‘an unlimited supply of labour, available at [a] subsistence wage’ (Lewis 1954, p. 139) to both sectors. Lewis interprets the word ‘subsistence’ broadly. The level of the wage is determined in some way by conventions in the underdeveloped agricultural sector. Lewis is non-committal as to whether wages in this sector are set according to actual subsistence needs, or living standards, or workers’ average product. The central idea is that workers are paid above their *marginal* product. Labour can be transferred from agricultural sector to the industrial sector by the migration of workers to towns and cities. The overall stock of labour in the economy is normally fixed in supply (though Lewis, like Ricardo, did sometimes allow for Malthusian features). Workers in the cities are paid not much more than the subsistence wage, although there may be a gap, as discussed below.

Third, and finally, there are differences in the behaviour of the actors in the two sectors. Capitalists in the industrial sector save all their profits, because they are ambitious. Workers save nothing, in either sector, because they are poor (Lewis describes them as not belonging to the ‘the saving class’ – 1954, p. 157). And landlords in agriculture are assumed to consume all their income, which comes to them to the extent that agricultural workers receive a wage below their average product.

The general story is this: the profits in the modern, capitalist, sector create a growing supply of savings. This finances the formation of an increasing stock of capital, which is used to employ more and more labour in the urban workforce.

We can explain the story in detail, using a simplified version of the model. To do this we make four sets of extreme assumptions. (*a*) There is ‘pure’ surplus labour, by which we mean that the marginal product of workers withdrawn from agriculture is *zero*. Wages initially consist only of agricultural goods, the level of wages per worker is exogenous, and workers are indifferent between working in industry and in agriculture at the same wage. (*b*) When one individual worker leaves agriculture and no longer needs to be rewarded there, then all the increase in the agricultural surplus (that is, all the increase in the total of food produced minus the total of wages paid to agricultural workers) accrues to landlords and is spent by them on consumption of industrial goods. (*c*) Industrial capitalists employ labour up to the point at which the marginal physical product of labour is equal to the cost of the wage, measured in industrial

goods. (*d*) All industrial profits are saved and then invested in industrial production.

Given these assumptions, there are two steps to the argument. First, given assumptions (*a*), (*c*) and (*d*), the rate of growth depends negatively on the relative price of agricultural goods in terms of industrial goods. This is because an increase in food prices raises the cost of the wage per worker in terms of steel, causing less labour-intensive methods of production to be adopted, that is, causing production to become more capital intensive. As a result of this, any given amount of savings, and the accumulation of capital that it causes, will ‘go less far’ in employing labour in industry, and, as a result, industrial output will grow less rapidly. Second, assumptions (*a*) and (*b*) determine the relative price of agricultural goods, in the following way: the accumulation of capital in industry increases the demand for industrial workers, who must be transferred from agriculture. The relative price of agricultural goods will need to be high enough to induce the workers’ landlords to offer up those agricultural goods that they would have paid to the transferred workers but now receive as surplus, so as to receive industrial goods for consumption in exchange. Such trade enables workers to be paid in industry, where they now work. As Lewis (1954, p. 188) says, ‘the capitalists need the peasants’ food, and … the demand for food is inelastic’.

Clearly, the relative prices of industrial and agricultural goods, and the growth rate of the economy, are *jointly* determined in this process – as Lewis's intuition had suggested to him. And it will clearly be true that the relative price of agricultural goods will need to be less high – and so the rate of growth will be higher – the lower is the price of agricultural goods required for landlords to release their surplus in exchange for consumable industrial goods.

Note that the share of income that accrues to industrial capitalists will increase during the growth process, as the capitalist sector grows in size. This suggested to Lewis (1954, p. 155) that a growth process of this kind might help to solve what he called the ‘central problem’ of development: the need to raise the savings rate enough to enable rapid growth to take place. In this model it is necessary to transfer labour into industry, in order to increase the overall savings rate of the economy. This is due to the behavioural assumption that agricultural income is not saved; we revisit this assumption below. Interestingly – from today's point of view – Lewis thought that a savings rate of ten to twelve per cent might be sufficient to achieve the ‘rapid capital accumulation’ that he believed integral to the process of development (Lewis, 1954, p. 155). Note also that increasing inequality is a frequent, if not necessary, correlate to this rising savings share, at least in the early stages of development (see, for example, Fei et al., 1979). This story thus also provides an explanation of the ‘Kuznets curve’.

Generalizations

Lewis does sometimes enlist the extreme simplifications made above. They correspond most closely to those made by Gustav Ranis and John C. H. Fei (1961), who used them to explain, more formally than Lewis did, what they call the ‘first phase’ of economic development – a phase in which there is ‘pure’ surplus labour. But Lewis also hints at many ways in which these assumptions could be relaxed. Ranis and Fei, along with Dale W. Jorgenson (and many others), went on to consider the implications of dualism when there are sectoral asymmetries different from those outlined above. In what follows, we consider a number of these extensions.

The first, and most fundamental, generalization of Lewis's model was made by Ranis and Fei (1961), who demonstrated that the dualistic framework continued to give insight into the process of economic

growth even when the condition of pure surplus labour does not hold. They initiated a large body work on this question by examining the microeconomic foundations of surplus labour and exploring what occurs when these conditions come to an end. This occurs when a sufficient number of workers have been removed from agriculture for the marginal productivity of the remaining agricultural workers to become positive. As a result, agricultural output declines as further workers leave. (This may happen even if there is technological progress in agriculture, providing that this progress is not sufficient to fully compensate for lost labour.) Consequently, the marginal agricultural surplus per worker, which accrues to landlords as each worker leaves – and which is traded by landlords for industrial goods – begins to decline, *even if* the wage per worker (measured in terms of agricultural goods) is exogenous. This means that the cost of labour to industry, measured in terms of industrial goods, will begin to rise above the level described in the sketch above – thereby constraining the rate of growth. This is the ‘first turning point’ identified by Ranis and Fei. It corresponds to the onset of Ricardo's ‘dismal’ diminishing returns. Ranis and Fei label what happens beyond this point as the ‘second phase’ of economic development. In that phase the economy is characterized by ‘disguised unemployment’, since labour in agriculture is still paid more than its marginal product.

Lewis himself was accused of not allowing for this possibility, even though he had written that the existence of zero marginal product is ‘not … of fundamental importance to our analysis’ (Lewis, 1954, p. 142). This accusation led to what Lewis later called an ‘irrelevant and intemperate controversy’ about the existence, or not, of ‘pure’ surplus labour (Lewis, 1972, p. 77). Ranis (2003, p. 8) agrees with Lewis's self-defence: in a retrospective assessment, he describes the postulation of a ‘pure’ labour surplus as a red herring. Amartya Sen (1966) helpfully clarifies the debate about this issue.

Growth becomes more difficult in this second stage of development. Recall that Lewis argues that the real wages per worker, and the level of welfare per worker, do not fall as growth proceeds. But growth is driven by the transfer of labour from agriculture to industry, which, in this second phase, causes agricultural output to fall. As a consequence of this the relative price of agricultural goods rises, and real wages can remain constant only if workers are able to substitute towards industrial goods in such a way as to avoid any damage to their welfare.

The agricultural sector as a constraint on growth

To highlight the essential role of such substitution, Mukesh Eswaran and Ashok Kotwal (1993) assume an extreme version of Engel's law. Consumers are assumed to spend *all* their income on food until they reach a particular threshold level of consumption, when they become sated with food. Beyond this point all further increases in consumption are devoted to industrial goods. At the same time they assume that labour always has a positive marginal product in agriculture. Under these assumptions, if workers remain so poor that they are not sated with food, then the transfer of labour across sectors – and therefore accumulation of industrial capital – becomes impossible. The inability of the poor to ‘eat shirts’ – an extreme version of what Ranis (2003) describes as the ‘product’ dimension of dualism – becomes a constraint on whether savings can lead to development. (And this constraint will bind quite independently of how high the marginal *physical* product of labour is in industry.) Any attempt to increase savings rates, in the manner desired by Lewis, so as to draw labour out of agriculture, would fail in these circumstances. The withdrawal of labour would lead to a reduction in the supply of food per

worker – the only thing that matters for workers' real wages – and so to a shortage of food. That shortage would turn the terms of trade against industry, depressing industrial profits and savings until the downward pressure on the supply of food had been removed, or until growth has ceased. As a result, all the gains from any increase in industrial production would accrue, in the form of lower prices, to those who consume industrial goods, rather than enabling growth, as in the Lewis model. It is thus clear that an important influence on whether development can proceed under dualism is the ability to shift workers' demands away from agricultural goods.

Of course, in a small, open, economy, the relative prices of tradables will be tied down, and the economy can respond to any developing shortage of food simply by exporting manufactures and importing food. That was Ricardo's insight, over 100 years earlier, about the gains to Britain from the abolition of the Corn Laws; Lewis's model of dualism in the world economy also incorporates such trade. But Lewis (1972, p. 94) cautions that there may be limits to this if export prices are not really exogenous, and if, instead, the county needs to cheapen its exports to pay for the imports of food – and other goods – that it will need as it grows. Perhaps partly because of this, Lewis (1954, p. 176) argues that a country which exhausts its surplus labour supply might instead export its savings, investing in industrial development in countries where the surplus labour condition continues to hold, and so enabling the output of manufactured goods to grow without driving down the rate of profit. In addition, the country might import labour from these countries. In this way Lewis's early contributions anticipated, and fed into, debates about the roles of outsourcing and immigration in contemporary globalization.

Jorgenson (1961) further develops the study of the dynamics of a dualistic economy in this second phase of development – when there is a positive marginal product of labour in agriculture and disguised unemployment. He incorporates a Malthusian perspective, by supposing that population growth is increasing in the amount of food consumed per capita, up to a biological ceiling that corresponds to the food-consumption threshold of Eswaran and Kotwal. This has the consequence that too rapid a rate of growth of population can cause a Malthusian trap by preventing the emergence of *any* significant agricultural surplus. Growth of manufacturing activity, such as that analysed by Lewis, can then be sustained only if technological progress in agriculture enables food production to outstrip population growth. (Capital accumulation in agriculture could have a similar effect in a model more general than that used by Jorgenson.) Only then can an agricultural surplus emerge, and grow, and so only then can labour progressively move away from agriculture. If this does not happen, then any increases in profits, savings and capital accumulation in industry become self-defeating, since they turn the terms of trade against industry and so bring down profits and savings, and bring growth to an end, in the way described two paragraphs above.

As stressed by Avinash Dixit (1973, p. 346), such a model focuses on 'the constraint on growth imposed by the rate of release of labour from agriculture', whereas in Lewis's model the focus had been on the ability of capital accumulation in industry to soak up the surplus labour force in agriculture. Nevertheless, as Dixit notes, growth paths in the two models will produce similar outcomes. In particular, in both models one would observe an endogenous rise in the savings rate as development proceeds. And in both models, it *may* be the case that any attempt to foster growth in industry, by a 'big push' to save more, is self-defeating. (This can be true in Jorgenson's model, and as we saw above, it can also be true beyond the 'first stage' of growth in the Lewis model, if it is not possible to induce workers to substitute away from agricultural goods.) This is why Jorgenson thought of increases in savings rates as an *outcome* of development, not as a policy tool which can be used to *promote* development

(Jorgenson, 1961, p. 328).

It is worth contrasting this view of potential ‘development traps’ with that which had been put forward in the 1940s by Paul Rosenstein-Rodan (1943), who built on his experience of eastern Europe. Rosenstein-Rodan's viewpoint also came from thinking about the interaction between agriculture and industry; like Lewis, he argued that development could only come to an agricultural economy through a process of industrialization. This, he argued, is because only industrial capitalists could afford to pay for the large fixed costs that are necessary to enable them to produce goods in a modern way, with low marginal costs. But if most people live in an impoverished agricultural sector then this would constrain their incomes, and so would limit their demand for modern industrial goods. That might make it unprofitable to make the required investment, and so might thwart the process of development. Here, just as for Lewis, a shortage of savings can be *the* problem of development. But by contrast with Lewis, a big push might fix it, since, roughly speaking, if all capitalists invested at once and paid their workers higher wages, then the demand for industrial goods would grow, making the investment worthwhile. This insight gave birth to the other great analytical engine of development economics, subsequently formalized by Kevin M. Murphy, Andrei Shleifer and Robert W. Vishny (1989) and Kiminori Matsuyama (1991), and well explained by Paul Krugman (1993). Since the pecuniary externalities that allow an economy to escape from a development trap are accessible only in the ‘modern’ sector, asymmetries between sectors are also central to this view.

Further aspects of labour transfer

The Lewis model was also generalized to explain the gap between the wage paid in the rural sector and that paid in the urban sector and to explore the consequences of such a gap. Lewis himself (1954, p. 150) acknowledged the existence of a wage gap, and suggested that it may result from the psychological costs of lifestyle changes, from the need to reward skills accumulated in the urban sector, or from the ability of workers in cities to bargain for higher wages. (This is particularly relevant when we recognize that the urban sector includes government employment and some services.) Subsequent authors took up this question, arguing, for example, that wage premia may arise because they lead to greater productivity through effects on health or employee motivation (for example, Dasgupta and Ray, 1986; Shapiro and Stiglitz, 1984).

The consequences of such a gap, for the process of labour transfer from agriculture to industry, were set out in the celebrated work of John Harris and Michael P. Todaro (1970). It may be that a wage floor in the urban formal sector prevents the market from clearing there. If a wage floor operates, then workers who choose to leave the rural sector face the prospect of receiving an urban wage which is above that of the rural sector, if they get employed, but also face some probability of becoming unemployed. In the simplest version of this model, equilibrium occurs when labour migration equalizes expected income across sectors – an outcome in which the rural wage equals a weighted average of the incomes received by employed and unemployed urban workers, weighted according to the probability of unemployment in the urban sector. Even without this extreme outcome, there are important policy implications in such a model. The more elastic is labour supply to the urban sector with respect to expected income there, the greater the amount of urban unemployment that will be induced by any policies that increase urban wages. This incorporation of urban unemployment into the model also enables one to begin to discuss

the growth of a third sector: the production of services in cities (see Fields, 1975). Roughly speaking, we can say that services get produced by (some of) those who migrate to cites, but do not get a job in manufacturing.

It is clear that the expansion of the industrial sector will ultimately take the economy beyond the second phase of economic development, in which there is disguised unemployment. This is because withdrawal of labour from agriculture will eventually reach the point at which the marginal product of the remaining labour rises to equality with the subsistence wage. Ranis and Fei call this a ‘second turning point’. At this point the marginal worker, offered a subsistence wage, can now instead offer his or her labour to a higher bidder. From then on the wage (measured in agricultural goods) will begin to rise in both sectors as growth continues. We can say that the ‘dualistic’ structure of the economy then comes to an end, in that the rural economy becomes ‘commercialized’. (Something similar, too, will happen in any services sector.) That leads one back to a labour-scarce economy, the analysis of which is better suited to neoclassical theory. A two-sector neoclassical growth model – something like the model of Hirofumi Uzawa (1961; 1963) – may be a better way to think about growth in these circumstances.

One key strand of the story of dualism that we have been telling is the assumption that capitalists save, but workers (and landlords) do not. Lewis's explanation of this asymmetry is largely behavioural. But such differences in savings rates between the traditional and the modern sectors might also be explained *institutionally*, by means of credit-market imperfections. If a technological asymmetry precludes investment in rural areas, and if limited financial development means that rural residents lack access to investment opportunities in manufacturing, then the agricultural surplus will not be used directly to finance investment. Moreover, typical characterizations of credit-market imperfections highlight the moral hazard problems that persist in rural areas because the poor there are unable to provide the kind of collateral required for formal-sector loans. (Small rural landholdings are of limited use as collateral.) Such lack of collateral stands as a barrier to borrowing, even though loans might be used to facilitate growth by promoting education, or capital accumulation, or technical progress in agriculture. (See Ray, 1998, for a summary of these arguments.) By contrast, Abhijit Banerjee and Andrew Newman (1998) provide an alternative perspective, emphasizing a sectoral asymmetry in the *informational* dimensions of credit-market imperfections, and showing how this can affect the willingness of individual workers to migrate in a dualistic economy. They present a model in which there is access to credit for consumption in rural areas. Given that workers have limited collateral wherever they live, a crucial determinant of their access to credit is the amount of information that lenders have about prospective borrowers. In contrast with the relative anonymity of urban life, small communities of the rural sector may provide superior information about borrowers, and thus foster lending. Banerjee and Newman show that dualism, characterized in terms of this differential severity of information asymmetries, might lead to a suboptimal allocation of labour across sectors. By financing consumption in the rural sector, rural credit might actually provide an incentive for labour to remain there; this incentive could offset the relatively high wages of the modern sector and could thereby impede the development process. Their paper suggests – at the least – that the lens of asymmetric information can shed useful light on the development of such economies.

Defining characteristics of economic dualism

We conclude by noting that we have described a number of reasons for differences between the industrial and agricultural sectors of a developing economy. Just as in Lewis's original article, all these differences go beyond mere asymmetries in production technologies or factor endowments between the sectors. This is why, following Ravi Kanbur and James McIntosh (1987), we would not normally describe the two-sector growth models of Uzawa (1961; 1963) as models of dualism, even though in those models the two sectors have different factor intensities. Nor would we say that that the two-sector Hecksher–Ohlin model of international trade is a model of a dualistic economy – even when its two sectors have different factor intensities, and even when the two sectors are labelled ‘agriculture’ and ‘industry’. Furthermore, although the specificity of factors to sectors appears central to Lewis's set-up (with land specific to agriculture and capital specific to industry), this feature does not seem to be sufficient to merit the label of ‘dualism’. Thus, for example, we would not regard the short-run version of the Heckscher–Ohlin trade model presented by J. Peter Neary (1978), with factors specific in each of the two sectors, as portraying a dualistic economy.

Instead, we would argue that the defining characteristic of modern theories of economic dualism lies – just as it did in Lewis's article – in a focus on sectoral asymmetries that are not simply technological. For Lewis, and for Ranis and Fei, there were *organizational* differences between sectors – in that wages were assumed to be determined by institutional factors in the agricultural sector – and *behavioural* differences between sectors – in that those in the rural sector were assumed to be unwilling to save, while capitalists were assumed to save everything. A focus on these features might imply that ‘pull’ factors drive labour transfer, and hence economic growth, in a dualistic economy. But since Lewis, economists studying economic development have explored alternative asymmetries between sectors and have reached different conclusions. The model of Eswaran and Kotwal, in which the defining asymmetries are *product* asymmetries – an assumption that all income is spent on agricultural goods until some threshold – highlights the need for labour productivity increases in agriculture to avoid stagnation of real wages. This is a need that persists even in the presence of rising productivity in industry. Jorgenson, who coupled such a view with a demonstration that Malthusian pressures can prevent income from ever rising above this threshold, showed clearly that growth can be constrained unless the ‘push’ factor of growth in agricultural technology is strong enough. Banerjee and Newman, by contrast, have emphasized that *informational* asymmetries between traditional and modern sectors can constrain the growth process.

We thus believe that, in the study of any particular economy, it is important to understand which asymmetries impose binding constraints on growth. Different constraints imply the need for different policies. But identifying the relevant asymmetries is even more important if we wish to remove these underlying constraints themselves. Joseph Stiglitz has proposed that we do just this, advocating what he calls ‘growth strategies based on duality's elimination’ (Stiglitz, 1999, p. 56). Much empirical work is necessary if we are to understand what such strategies might require.

See Also

- labour surplus economies
- Lewis, W. Arthur

Bibliography

Banerjee, A.V. and Newman, A.F. 1998. Information, the dual economy, and development. *Review of Economic Studies* 65, 631–5.

Dasgupta, P. and Ray, D. 1986. Inequality as a determinant of malnutrition and unemployment: theory. *Economic Journal* 96, 1011–34.

Dixit, A. 1973. Models of dual economies. In *Models of Economic Growth: Proceedings of a Conference held by the International Economic Association at Jerusalem*, ed. J.A. Mirrlees and N.H. Stern. London: Macmillan.

Domar, E.D. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14, 137–47.

Eckaus, R.S. 1955. The factor proportions problem in underdeveloped areas. *American Economic Review* 45, 539–65.

Eswaran, M. and Kotwal, A. 1993. A theory of real wage growth in LDCs. *Journal of Development Economics* 42, 243–69.

Fei, J.C.H., Ranis, G. and Kuo, S.W.Y. 1979. *Growth with Equity: The Taiwan Case*. Oxford: Oxford University Press.

Fields, G.S. 1975. Rural-urban migration, urban unemployment and underemployment, and job-search activity in LDCs. *Journal of Development Economics* 2, 165–87.

Findlay, R. 1980. On W. Arthur Lewis' contributions to economics. *Scandinavian Journal of Economics* 82, 62–79.

Harris, J.R. and Todaro, M.P. 1970. Migration, unemployment and development: a two-sector analysis. *American Economic Review* 60, 126–42.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49, 13–33.

Hicks, J. 1965. *Capital and Growth*. Oxford: Oxford University Press.

Higgins, B. 1956. The 'dualistic theory' of underdeveloped areas. *Economic Development and Cultural Change* 4, 99–115.

Jorgenson, D.W. 1961. The development of a dual economy. *Economic Journal* 71, 309–34.

Kanbur, R. and McIntosh, J. 1987. Dual economies. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P.K. Newman. London: Macmillan.

Krugman, P. 1993. Toward a counter-counterrevolution in development theory. In *Proceedings of the World Bank Annual Conference on Development Economics, 1992: Supplement to The World Bank Economic Review and The World Bank Research Observer*, ed. L.H. Summers and S. Shah. Washington, DC: World Bank.

Lewis, W.A. 1953. *Report on Industrialisation and the Gold Coast*. Accra, Gold Coast: Government Printing Department.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School* 28, 139–91.

Lewis, W.A. 1955. *The Theory of Economic Growth*. London: Allen & Unwin.

Lewis, W.A. 1972. Reflections on unlimited labour. In *International Economics and Development: Essays in Honour of Raul Prebisch*, ed. L.E. di Marco. London: Academic Press.

Lewis, W.A. 1978. *Growth and Fluctuations, 1870–1913*. London: Allen & Unwin.

Lewis, W.A. 1979. Autobiography. In Lindbeck (1992). Also online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1979/lewis-autobio.html, accessed 9 January 2007.

Lindbeck, A., ed. 1992. *Nobel Lectures: Economic Sciences 1969–1980*. Singapore: World Scientific Publishing.

Matsuyama, K. 1991. Increasing returns, industrialization, and the indeterminacy of equilibrium. *Quarterly Journal of Economics* 106, 617–50.

Murphy, K.M., Shleifer, A. and Vishny, R.W. 1989. Industrialization and the big push. *Journal of Political Economy* 97, 1003–26.

Neary, J.P. 1978. Short-run capacity specificity and the pure theory of international trade. *Economic Journal* 88, 488–510.

Pasinetti, L. 1974. *Growth and Income Distribution—Essays in Economic Theory*. Cambridge: Cambridge University Press.

Preobrazhensky, E. 1924. *The New Economics*. Trans. B. Pearce, Oxford: Clarendon Press, 1965.

Ranis, G. 2003. Is dualism worth revisiting? Discussion Paper No. 870. Economic Growth Center, Yale University.

Ranis, G. and Fei, J.C.H. 1961. A theory of economic development. *American Economic Review* 51, 533–65.

Ray, D. 1998. *Development Economics*. Princeton: Princeton University Press.

Rosenstein-Rodan, P. 1943. Problems of industrialisation of eastern and south-eastern Europe. *Economic Journal* 53, 202–11.

Sah, R.K. and Stiglitz, J.E. 1984. The economics of price scissors. *American Economic Review* 74, 125–38.

Sen, A. 1966. Peasants and dualism with or without surplus labour. *Journal of Political Economy* 74, 425–50.

Shapiro, C. and Stiglitz, J.E. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Stiglitz, J.E. 1999. Duality and development: some reflections on economic policy. In *Development, Duality, and the International Economic Regime: Essays in Honor of Gustav Ranis*, ed. G. Saxonhouse and T.N. Srinivasan. Ann Arbor: University of Michigan Press.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Uzawa, H. 1961. On a two-sector model of economic growth. *Review of Economic Studies* 29, 40–47.

Uzawa, H. 1963. On a two-sector model of economic growth II. *Review of Economic Studies* 30, 105–118.

How to cite this article

Vines, David and Andrew Zeitlin. "dual economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000195> doi:10.1057/9780230226203.0409

The New Palgrave Dictionary of Economics Online

dual track liberalization

Yingyi Qian
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Dual track liberalization is a reform strategy in which a market track is introduced while the plan track is maintained at the same time. Dual track liberation is Pareto improving in the sense that it makes some people better off without making anybody worse off. Because prices are liberalized at the margin, dual track liberalization can also achieve efficiency. China used the dual track reform strategy in liberalizing many markets such as the markets of agricultural goods, industrial goods, consumer goods, foreign exchange, and labour, as well as in creating special economic zones.

Keywords

allocative efficiency; China, economics in; compensatory transfers; corruption; dual track liberalization; foreign exchange control; market liberalization; Pareto efficiency; planning; price control; price liberalization; rationing; rent seeking; special economic zones (China)

Article

Dual track liberalization is a reform strategy of market liberalization in which a market track is introduced while the plan track is maintained at the same time. Under the plan track, economic agents are assigned rights to and obligations for a fixed quantity of goods and services at fixed planned prices as specified in the pre-existing plan. Under the market track, economic agents can participate in the market at free market prices, provided that they fulfil their obligations under the pre-existing plan. The essential feature of the dual track strategy to market liberalization is that prices are liberalized at the margin while inframarginal plan prices and quotas are maintained for some time before being phased out. Although the dual track reform strategy is widely adopted in China during its transition from plan to market, it is also used in other countries. For example, when introducing new legislation, a ‘grandfathering’ clause is often adopted to protect existing interests, which is a form of the dual track approach to reform.

Analysis of dual track liberalization follows two lines of approach. The first focuses on its Pareto-improvement property, that is, dual track liberation makes nobody worse off while it makes somebody better off – and therefore it has a political advantage in implementing reforms. Most efficiency-improving market liberalization reforms potentially create winners and losers, despite the fact that, in theory, efficiency gains should be large enough to allow the potential losers to be compensated. For example, the single track approach to liberalization (that is, where all the prices are freed at once) in general cannot guarantee an outcome without losers. Dual track liberalization means that planned quantity continues to be delivered at plan price but any additional quantity can be sold freely in the market. With the dual track, the surpluses of the rationed users and the planned suppliers remain exactly the same. The purpose of maintaining the plan track is to provide implicit transfers to compensate potential losers from market liberalization by protecting status quo rents under the pre-existing plan. On the one hand, the introduction of the market track provides the opportunity for economic agents who participate in it to be better off. At the same time, the new users and suppliers outside the plan are also better off. Therefore, the intuitive appeal of dual track liberalization for reformers lies precisely in the fact that it represents a mechanism of the implementation of a reform without creating losers (Lau, Qian and Roland, 2000).

The second approach focuses on the efficiency property of dual track liberalization. Pareto-improvement property implies that it always improves efficiency. This is independent of other assumptions, for example, as to whether the market is competitive or not. In contrast, the single track approach to liberalization may improve efficiency under perfect competition, but may not improve efficiency if the market is monopolistic (Li, 1999). The more subtle and deeper point is that the dual track approach to liberalization may achieve

allocative efficiency, despite the fact that it appears inefficient, by maintaining the inefficient planned track. The fundamental reason is that the compensatory transfers, which are implicitly embodied in the planned track, are inframarginal, and thus the distortion can be avoided.

To see this we look at the special case where the pre-reform status quo features efficient rationing and efficient planned supply in the sense that the planned output is allocated to users with the highest willingness to pay and the planned supply is delivered by suppliers with the lowest marginal costs. Nevertheless, the price of the good is fixed at an artificially low level and the production quota is fixed below market equilibrium (Figure 1). When the market track is introduced into this setting, it is clear that the market equilibrium quantity and price would be identical to the case without the planned price and quota to start with. Therefore, dual track liberalization achieves efficiency. Notice that efficiency is achieved without making anyone worse off. Indeed, the rents enjoyed by the buyers under rationing (area A in Figure 1) are preserved under dual track liberalization, but would be lost under single track liberalization.

Figure 1

The case of efficient supply and efficient rationing

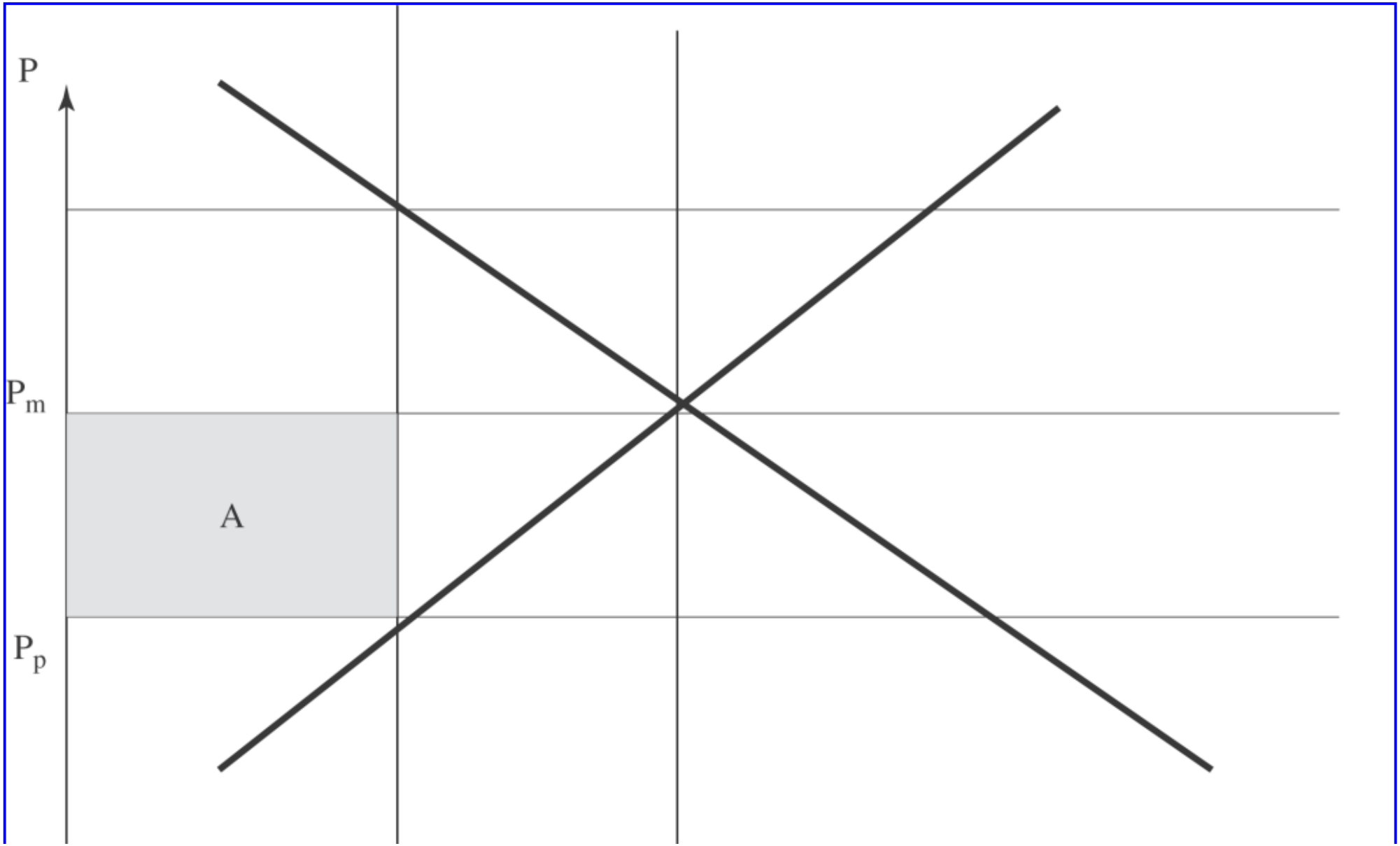

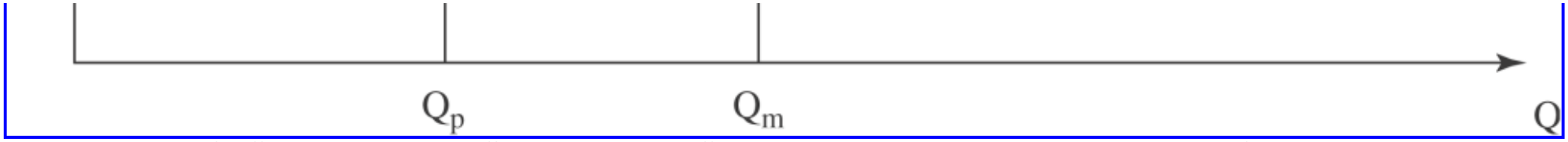

In a more general case of inefficient rationing and/or inefficient planned supply, efficiency can still be achieved provided market liberalization is full, in the sense that market resales of plan-allocated goods and market purchases by planned suppliers for fulfilling planned delivery quotas are permitted after the fulfilment of the obligations of planned suppliers and rationed users under the plan. This removes any inefficiency associated with the original planned prices and quotas and makes imputed rents under planning inframarginal. This type of transaction takes many forms in practice, for example, subcontracting by inefficient planned suppliers to more efficient non-planned suppliers, and labour reallocation when workers in inefficient enterprises keep the housing while taking a new job in more efficient firms. In both examples, after fulfilling the obligations under the plan (planned delivery of supply and welfare support through housing subsidies), the market track functions to undo the inefficiency of the plan track.

This above partial equilibrium analysis can be generalized to a general equilibrium mode (Lau, Qian, and Roland, 1997). Efficiency requires full market liberalization under which market resales, subcontracting, and market purchases for redelivery are all allowed. Indeed, the distinction between limited and full market liberalization is a major difference between Lau, Qian, and Roland (1997) and Byrd (1991), and others who have studied the dual track approach.

If such resales and purchases are not allowed or cannot be achieved, then dual track liberalization is limited and efficiency in general cannot be achieved, although it can be improved. Of course, in the special case discussed above with efficient supply and efficient rationing, dual track with limited market liberalization is the same as dual track with full market liberalization. In general, dual track limited market liberalization need not be the same as dual track full market liberalization.

Sometimes dual track liberalization of the market takes the following sequential form: in a first stage, limited market liberalization is implemented, and then in a second stage full market liberalization is implemented. In the first stage, going from a centrally planned economy to limited market liberalization, Pareto improvement is clearly attained, but efficiency cannot be guaranteed. Specifically, limited market liberalization generally leads to inefficient overproduction due to market entry. In the second stage, when full liberalization is introduced, efficiency is attained but Pareto improvement may not be. This is because the second-stage full market liberalization implies efficiency, and thus there must be a production contraction and some people have to reduce production and are made worse off. Therefore, the sequential dual track liberalization may result in some opposition to further reforms after the first and before the second stage, while the dual track full market liberalization that is implemented in one stroke will not. Nevertheless, it is also clear that, even under the sequential dual track liberalization, there are no losers at the end of the second stage compared with the status quo before the reform.

The dual track approach to market liberalization is an example of reform making the best use of existing information and institutions. First, it utilizes efficiently the existing information embedded in the original plan (that is, existing rents distribution) and its implementation does not require additional information. Second, it also enforces the plan through the existing plan institutions and does not need additional institutions. Enforcement of the plan track is crucial for preserving pre-existing rents. However, contrary to common understanding of the relationship between state power and reform, state enforcement power is needed here not to implement an unpopular reform, but to carry out one that creates only winners, without losers.

Economists sometimes find dual track liberalization puzzling and counter-intuitive, for several reasons. First, economists are used to the law of one price: in a competitive setting, multiple prices entail inefficiency. However, in dual price liberalization, the planned price comes together with planned quantity, when they are fixed, they do not entail inefficiency, at least not additional inefficiency. Second, dual track resembles price control, which is associated with inefficiency and rent seeking. But dual track is not price control; on the contrary, it is a move towards price liberalization. An important difference between the plan track under dual track and price control is that the plan track embodies both fixed prices and fixed quantities; it is a package of price and quantity control, not just price control. Under pure price control, the government fixes only prices, but not quantities. Third, to reformers, dual track seems a partial reform and not a complete reform. This is true under dual track with limited market liberalization, but not true with full market liberalization. Although dual track with limited market liberalization does not achieve efficiency, it improves efficiency and makes nobody worse off.

Dual track liberalization requires enforcement of the rights and obligations under the plan track. In fact, enforcement of the plan track alone would prevent any decline in aggregate output. Can the plan track be enforced? With a collapsing government, it cannot. But enforcing the pre-existing plan is informationally much less demanding for the government than drawing up a new plan. Under central planning, the information requirement for drawing up a plan is huge. Enforcing a pre-existing plan is different. In fact, the dual track approach uses minimal additional information as compared with other possible compensation schemes that may be used with other approaches to reform. Compliance with the plan by economic agents depends on their expectations of the credibility of state enforcement. If state enforcement is not credible, then the economic agents will have no incentive to fulfil their plan obligations. If people think that they are not going to receive the plan-mandated deliveries at plan prices, they will not make the plan-mandated sales at the fixed plan prices. In that case, dual track liberalization degenerates to single track liberalization.

Lack of enforcement of the plan track may result in supply diversion as analysed by Murphy, Shleifer and Vishny (1992). These authors studied a partial reform model with the following two crucial assumptions: (i) suppliers are free to sell to all users, and (ii) buyers who are not covered by the plan can freely purchase inputs at any price, but buyers who are covered by the plan are not allowed to purchase inputs above the plan price. This partial reform model differs from the dual track liberalization model in an important respect: there is

no plan delivery quota enforced on the suppliers.

In their model, partial reform may lead to inefficient supply diversion to such an extent that the outcome can be worse than that without reform. Therefore, the partial reform is not only not Pareto improving, but also total welfare reducing. Consider the case where the initial condition is also characterized by efficient rationing and efficient supply as shown in Figure 1, where the planned price P_p is below the market clearing level P_m. Then, after the partial liberalization as defined above, suppliers can sell the good freely to the highest bidders. While the firms under the plan are forced to buy the good at price P_p, the firms outside the plan are free to buy the good at any price. Then they will bid the good for price $P_p + \varepsilon$ where ε is a positive but small number. Because the firms under the plan are constrained to pay P_p, an amount will be diverted from them to those not covered by the plan. Because the willingness to pay from those not covered by the plan is lower than those covered by the plan (by the assumption of efficient rationing), this kind of partial reform induces a net efficiency loss. While the sector not covered by the plan gains, the sector covered by the plan loses, and the total welfare effect is unambiguously negative. Although the assumption of efficient rationing and efficient supply under central planning is too strong, the result of inefficient supply diversion under partial reform remains valid with weaker conditions about initial rationing and supply.

So which model is more relevant? It depends on the quota enforcement capability of the government. A good enforcement capability makes the dual track liberalization model of Lau, Qian and Roland more relevant, while a poor enforcement would make the partial reform model of Murphy, Shleifer, and Vishny more relevant. The dual track liberalization model is motivated mainly by the practice in China, where enforcement has been reasonably good, while the partial reform model is mainly motivated by the experiences of the last years of the Soviet Union, when the state enforcement power diminished quickly.

In China's context, lack of quota enforcement sometimes takes the following form. The government may be unable to freeze the plan by creating new quotas with (below market equilibrium) planned price and giving windfall rents to some people who are politically connected. This may lead to corruption: firms find it easier to make profits by lobbying the government for allocating more input goods delivery at low planned prices, without the corresponding obligations to deliver low price outputs as under central planning. They then sell the goods at the market price to receive the windfall gains. This type of corruption is often attributed to the dual track approach to liberalization. Indeed, without the coexistence of the planned prices and market prices, the above form of corruption is not possible. By eliminating the two prices, such form of corruption would disappear. However, the essence of the problem is the failure in the enforcement of the original planned track. If the planned track is strictly enforced, no new quotas should be created. (On the other hand, full market liberalization allows for market arbitrage, which may increase the welfare of those who were allocated with goods at below-market prices. This is essential for achieving efficiency. The difference is that the potential rents are inherited from the previous regime in this case, not from a new creation.)

Dual track liberalization in practice

Studies of dual track liberalization focus mostly on China, although other cases, such as that of Mauritius, are also mentioned. The origin of the dual track can be traced to the 1950s when China had two prices for grain, the official price and negotiated price. However, dual track approach to market liberalization as a reform strategy was used only after 1979, first in the agricultural goods markets, and then in other markets (Byrd, 1991; Naughton, 1995; Lau, Qian and Roland, 2000).

Agriculture goods. The agricultural reform in China started with a dual track approach to market liberalization. Under that reform, the commune (and later the household) was assigned the responsibility to sell a fixed quantity of output to the state procurement agency as previously mandated under the plan at predetermined plan prices and to pay a fixed tax (often in kind) to the government. It also had the right (and obligation) to receive a fixed quantity of inputs, principally chemical fertilizers, from state-owned suppliers at predetermined plan prices. Subject to fulfilling these conditions, the commune was free to produce and sell whatever it considered profitable, and retain any profit. Moreover, the commune could purchase from the market grain (or other) output for resale to the state in fulfilment of its responsibility. There was thus a full market liberalization.

Between 1978 and 1988 state procurement of domestically produced grain remained essentially fixed, with 47.8 million tons in 1978 and 50.5 million tons in 1988. During that same period, total grain output increased by almost one-third. But the dual track approach to liberalization applied to agricultural products other than grain: between 1978 and 1990, the share of transactions at plan prices in all agricultural goods fell from 94 per cent to 31 per cent, when the agricultural output in China doubled. There was a huge supply response to the introduction of the market track.

Industrial goods. The most noticeable and often cited application of the dual track approach to liberalization is to industrial goods (Byrd, 1991; McMillan and Naughton, 1992). The Chinese government issued a document in May 1984 stipulating that there would be two forms of production in state-owned enterprises: planned and non-planned. Correspondingly, there were two types of material supplies for enterprises, namely, state allocation and free purchase. Prices of goods in the former were fixed by the state and prices of goods above quota quantity could be sold in the market at price within a range up to 20 per cent higher or lower than of the planned price. In February 1985, the 20 per cent price cap was removed and the dual track for industrial goods was formally in place (Wu and Zhao, 1987). As a result, the share of transactions at plan prices, in terms of output value, fell from 100 per cent before the reform to 45 per cent in 1990.

Coal and steel are the two important industrial commodities most tightly controlled under central planning, and both coal and steel markets were liberalized through the dual track approach. For coal, China's principal energy source, the planned delivery led to some slight increases in absolute terms during the 1980s, but the market track increased dramatically from 293 million tons to 628 million tons over the same period – the supply came mainly from small rural mines run by Township–Village Enterprises. As a result, the share of the

plan allocation declined from 53 per cent in 1981 to 42 per cent in 1990. For steel, the plan track was quite stable in absolute terms during the 1980s, but the share of plan allocation fell from 52 per cent in 1981 to 30 per cent in 1990. In the cases of both coal and steel, because the plan track was essentially frozen, the economy was able to ‘grow out of the plan’ on the basis of the expansion of the market track (Naughton, 1995).

Consumer goods. Prior to the economic reform of 1979, most essential consumer goods and services for urban residents, such as grain, cooking oil, meat, electricity, housing, and the monthly transport pass, were rationed with coupons at values lower than corresponding free market prices. With dual track liberalization, urban residents continued to have the right to purchase grain, meat, electricity and housing at the same pre-reform prices and within the limits of the pre-reform rationed quantities, but, at the same time, they were also free to buy consumer goods from the free market at generally higher prices. The proportion of transactions at plan prices declined from 97 per cent in 1978 to only 30 per cent in 1990.

Foreign exchange. Under central planning, foreign exchange transactions were strictly controlled by the government at the official exchange rate. Exporters were required to surrender to the state all foreign exchange they earned at the official exchange rate, and importers were allocated with planned quotas of foreign exchange, also at the official exchange rate. Foreign visitors to China were required to use ‘foreign exchange certificates’, which were available at the official exchange rate. Starting from May 1988, China allowed trading of foreign exchange at Foreign Exchange Adjustment Centres (more commonly referred to as ‘swap centres’) at the rate determined by market supply and demand, called ‘swap rate’. This was the beginning of the dual track in the foreign exchange market. The swap rate was, not surprisingly, significantly higher than the official rate. The supply of foreign exchange in the swap markets was provided by exporters through the foreign exchange they were allowed to retain from net increases in their export earnings in relation to the base period. By the end of 1993, transactions at official exchange rates accounted only for about 20 per cent of the total; the rest were at the market rate.

Labour. As in many other centrally planned economies, the labour market in China was also distorted: most labour was allocated to unproductive, state-owned enterprises and few to the non-state sector. Dual track liberalization in the labour market takes two forms. In the first, the non-state sector (the liberalized sector) pays market wages and decides on hiring and firing. Between 1978 and 1994, employment in the non-state sector increased by 318.8 per cent, while employment in the state sector (including civil servants in government agencies and non-profit organizations) increased by only 50.5 per cent. Second, even within the state sector there are also two tracks. Beginning in 1980, while pre-existing employees maintained their permanent employment status, most new hires in the state sector were made under the more flexible contract system and often at lower effective wage rates. Employment in the plan track was virtually stationary – it declined from 87.14 million in 1983, on the eve of the introduction of economic reform in industry, to 83.61 million in 1994.

Special economic zones. Dual track liberalization can also have a geographical dimension: special economic zones are such examples. Although similar zones for processing exports can be seen in other Asian economies, special economic zones had a more profound effect in China because the whole country was still under central planning when they were created. Therefore, the purpose of special economic zones was more than for exporting; it was a strategy for market reform.

In 1980, China established four ‘special economic zones’, Shenzhen, Zhuhai and Shantou in Guangdong province and Xiamen in Fujian province. Most transactions relating to activities inside the zones were on the market track, including prices of input and output goods and wages of labour – at a time when the rest of the economy was still operating under central planning. The special economic zones were insulated from the rest of the economy to minimize the impact on and interaction with the rest of the economic system. Initially, firms inside the special economic zones had to import all their inputs and export all their outputs – thus creating no disruption to the domestic aggregate supply and demand. The principal purpose of this approach was to minimize the impact of new economic activities on the old-style domestic state-owned enterprises. Thus, once again, there were two tracks and the reform was Pareto improving.

In order for the special economic zones to work, merely creating them was not enough. One of the crucial conditions was the insulation of the non-liberalized sector from the liberalized sector so that the latter's existing rents could be maintained while the other sector was liberalized. Therefore, creation of special economic zones is a type of limited market liberalization. It is Pareto improving and efficiency enhancing, but cannot be fully efficient.

Phasing out the plan track

With rapid growth, the plan track will become a matter of little consequence to most potential losers, which in turn reduces the cost required for compensating them. In China, the plan track in product markets was largely phased out during the 1990s. By 1996, the plan track was reduced to 16.6 per cent in agricultural goods, 14.7 per cent in industrial producer goods, and only 7.2 per cent in total retail sales of consumer goods. However, this phasing-out of the plan track was generally accompanied by compensation. For example, urban food coupons (grain, meat, oil, and so on) were removed in the early 1990s with lump-sum compensation. But the cost of compensation was much smaller in relative terms as compared to the potential cost of compensation in the early 1980s. The dual track exchange rate ended on 1 January 1994, when the two exchange rates – the official rate and the swap rate – were merged into a single, market rate. In this last step of foreign exchange reform, those organizations that used to receive cheap foreign exchange were provided with annual lump-sum subsidies for a period of three years, which was sufficient for them to purchase the pre-reform allocation of foreign exchange. Because at that time the share of centrally allocated foreign exchange had already fallen to less than 20 per cent of the total, the cost of compensation was not too large.

See Also

- China, economics in

Bibliography

Byrd, W.A. 1991. *The Market Mechanism and Economic Reforms in China*. Armonk, NY: M.E. Sharpe.

Lau, L.J., Qian, Y. and Roland, G. 1997. Pareto-improving economic reforms through dual track liberalization. *Economics Letters* 55, 285–92.

Lau, L.J., Qian, Y. and Roland, G. 2000. Reform without losers: an interpretation of China's dual track approach to transition. *Journal of Political Economy* 108, 120–43.

Li, W. 1999. A tale of two reforms. *RAND Journal of Economics* 30, 120–36.

McMillan, J. and Naughton, B. 1992. How to reform a planned economy: lessons from China. *Oxford Review of Economic Policy* 8, 130–43.

Murphy, K.M., Shleifer, A. and Vishny, R.W. 1992. The transition to a market economy: pitfalls of partial reform. *Quarterly Journal of Economics* 107, 887–906.

Naughton, B. 1995. *Growing Out of the Plan*. Cambridge: Cambridge University Press.

Wu, J. and Zhao, R. 1987. The dual pricing system in China's industry. *Journal of Comparative Economics* 11, 309–18.

How to cite this article

Qian, Yingyi. "dual track liberalization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000256> doi:10.1057/9780230226203.0410

The New Palgrave Dictionary of Economics Online

duality

Lawrence E. Blume
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys duality in producer theory, consumer theory and welfare economics. As opposed to the usual analysis through first-order conditions for optimization, the various dualities are derived here from convex duality theory, using Fenchel transforms and subdifferentials.

Keywords

Antonelli, G.B.; Bergson–Samuelson social welfare function; convex programming; convexity; cost functions; cyclic monotonicity; duality; envelope th; Fenchel transform; firm, theory of the; Hicksian-compensated demand; Hotelling, H.; Hotelling's lemma; hyperplanes; indirect utility function; Lagrange multipliers; marginal revolution; monotonicity; profit functions; quasi-equilibrium; saddlepoints; separation th; Shephard, R.W.; Shephard's lemma

Article

1 Introduction

The word ‘duality’ is often used to invoke a contrast between two related concepts, as when the informal, peasant, or agricultural sector of an economy is labelled as dual to the formal, or profit-maximizing, sector. In microeconomic analysis, however, ‘duality’ refers to connections between quantities and prices which arise as a consequence of the hypotheses of optimization and convexity. Connected to this duality are the relationship between utility and expenditure functions (and profit and production functions), primal and dual linear programs, shadow prices, and a variety of other economic concepts. In most textbooks, the duality between, say, utility and expenditure functions arises from a sleight of hand with the first-order conditions for optimization. These dual relationships, however, are not naturally a product of the calculus; they are rooted in convex analysis and, in particular, in different ways of describing a convex set. This article will lay out some basic duality theory from the point of view of convex analysis, as a remedy for the microeconomic theory textbooks the reader may have suffered.

2 Mathematical background

Duality in microeconomics is properly understood as a consequence of convexity assumptions, such as laws of diminishing marginal returns. In microeconomic models, many sets of interest are closed convex sets. The mathematics here is surveyed in convex programming. The urtext for this material is Rockafellar (1970).

Closed convex sets can be described in two ways: by listing their elements, the ‘primal’ description of the set, and by listing the closed half-spaces that contain it. A closed (upper) half-space in $\mathbf{R^n}$ is a set of the form $h_{pa} = \{x: p \cdot x \geq a\}$, where p is another n-dimensional vector, a is a number and $p \cdot x$ is the inner product. The vector p is the *normal vector* to the half-spaces h_{pa}. Geometrically speaking, this is the set of points lying on or above the line $p \cdot x = a$. The famous separation theorem for convex sets implies that every closed convex set is the intersection of the half-spaces containing it.

Suppose that C is a closed convex set, and that p is a vector in $\mathbf{R^n}$. How do we find all the numbers a such that $C \subset h_{pa}$? If there is

an $x \in C$ such that $p \cdot x < a$, then a is too big. So the natural candidate is $w = \inf_{x \in C} p \cdot x$. If $a > w$ there will be an $x \in C$ such that $p \cdot x < a$ on the other hand, if $a < w$, then $p \cdot x > a$ for all $x \in C$. So the half-spaces h_{pa} for $a \leq w$ are the closed half-spaces containing C.

This construction can be applied to functions. A concave function on $\mathbf{R^n}$ is an $[-\infty, \infty)$ valued function f• such that the *hypograph* of f, the set $\text{hypo } f = \left\{(x, a) \in \mathbf{R^{n+1}} : a \leq f(x)\right\}$, is convex. If hypo f is closed, f is said to be *upper semi-continuous* (usc). The *domain* dom f of concave f is the set of vectors in $\mathbf{R^n}$ for which f is finite-valued. Concave (and convex) functions are very well-behaved on the *relative interiors* of their effective domains. The relative interior ri C of a convex set C is the interior relative to the smallest affine set containing C (see convex programming), and on ri dom f, f (concave or convex) is continuous.

Suppose that f is usc. The minimal level a such that $h_{(p,-1)a}$, the hyperplane in $\mathbf{R^{n+1}}$ with normal vector $(p, -1)$, contains hypo f is $f^*(p) \equiv \inf_x p \cdot x - f(x)$. Why the normal vector $(p, -1)$? Because the graph of the affine function $x \mapsto f^*(p) + px$ is a tangent line to f, the graph of f lies everywhere beneath it, and no other line with the same slope and a smaller intercept has this property. The function $f^*(p)$ is the (concave) *Fenchel transform* or *conjugate* of f, and is traditionally denoted f^*. The construction of the preceding paragraph can be done just this way: the *concave indicator function* of a convex set C is the function $\delta_C(x)$ which is 0 on C and $-\infty$ otherwise, and $\delta_C^*(p) = \inf_{x \in C} p \cdot x$. For any function f, not necessarily usc or concave, the Fenchel transform f^* is usc and concave. If f is in fact both usc and concave, then $f^{**} = f$. This fact is known as the conjugate duality theorem. Convex functions with range $(-\infty, \infty]$ are treated identically. The function f is convex if and only if $-f$ is concave, but the definitions are handled slightly differently in order to preserve the intuition just described. The set $\text{epi } f = \{x, a) : a > f(x)\}$, and the convex Fenchel transform is defined differently: $f^*(p) = \sup_x p \cdot x - f(x)$. The convex indicator function of a convex set C is the function $\delta^C(x)$ which is 0 on C and +otherwise; its (convex) conjugate is $\delta^{C*}(p) = \sup_x p \cdot x$. These facts are discussed in convex programming.

If concave functions have tangent lines, then they must have something like gradients. A vector p is a subgr*adient* of f at x if $f(x) + p \cdot (y - x) \leq f(y)$. If f has a unique subgradient at x, then f is differentiable at x and $p = \nabla f(x)$, and conversely. But the subgradient need not be unique: the set $\partial f(x)$ of subgradients at x is the *subdifferential* of f at x. The *domain* of f, dom f, is the set of x such that $f(x) > -\infty$. The subdifferential is non-empty for all x in its *relative interior*. It follows from the definition of concavity (and is proved in convex programming that the subdifferential correspondence is *monotonic*: if $p \in \partial f(x)$ and $q \in \partial f(y)$, *then* $(p-q) \cdot (x-y) \leq 0$. If f is convex, then the inequality is reversed, and $(p-q) \cdot (x-y) \geq 0$. Finally, suppose f is usc and concave. Then so is its conjugate f^*, and their subdifferentials have an inverse relationship: $p \in \partial f(x)$ if and only if $x \in \partial f^*(p)$.

3 Cost, profit and production

In the theory of the firm, profit functions and cost functions are alternative ways of describing the firms' technology choices. A technology is described by a set of vectors F in $\mathbf{R}^N$. Each vector $Z \in F$ is an input–output vector. We adopt the convention that negative coefficients correspond to input quantities and positive quantities correspond to outputs. Suppose that the first L goods are inputs and the last $M = N - L$ are outputs, so that $F \subset \mathbf{R}_-^L \times \mathbf{R}_+^M$. It is convenient to assume free disposal, so that if $(x, y) \in F$, and both $x' \leq x$ and $y' \leq y$ (more input and less output), then $(x', y') \in F$. Two important dual representations of the technology are the cost and profit functions. The profit function is $\pi(p, w) = \sup_{(x,y) \in F} p \cdot y + w \cdot x$ for $p \in$ and $w \in \mathbf{R}^L$, which is the conjugate of the convex indicator function of F. The cost function too can be obtained through conjugacy. The set $F(y) = \{x : (x, y) \in F\}$ is the set of all input bundles that produce y. Then $C(y, w) = -\sup_{x \in F(y)} w \cdot x$, that is, $C(y, \cdot) = -\delta^{F(y)*}$.

Immediately the properties of the Fenchel transform imply that $\pi(p,w)$ is convex in its arguments and $C(y,w)$ is concave in w, the profit function is lsc and the cost function is usc. (This implies that both functions are continuous on the relative interior of their effective domains.) Cost and profit functions are also linear homogeneous. Doubling all prices doubles both costs and revenues. Cost is also monotonic. If $w_l' < w_l$ for every input l, then $C(y, w') \leq C(y, w)$ and if $w_l' < w_l$ for all l, then $C(y, w') < C(y, w)$. The point of duality is that, if the technology is closed and convex, then cost profit functions each characterize the technology F. The conjugate duality theorem (see convex programming) implies that $\pi^*(x, y) \equiv \delta^{F**}(x, y) = \delta^F(x, y)$, the convex indicator

function of F:

$$\sup_{(p,w)\in\mathbf{R}^{\mathbf{N}}} p\cdot x + w\cdot y - \pi(p,w) = \begin{cases} 0 & \text{if } (x,y)\in F, \\ +\infty & \text{otherwise}. \end{cases}$$

If F *is* closed and convex, then each $F(y)$ is convex. If F is closed then $F(y)$ will also be closed. Then $\delta^{F(y)}$ is concave and usc, so

$$\sup_{w\in\mathbf{R}^{\mathbf{L}}_{+}} w\cdot x + C(y,w) = \sup_{w\in\mathbf{R}^{\mathbf{L}}_{+}} w\cdot x\delta^{F(y)*}(w) = \delta^{F(y)}(x).$$

Hotelling's lemma is a famous result of duality theory. It says that the net supply function of good i is the derivative of the profit function with respect to the price of good i. The usual proof is via the envelope theorem: the marginal change in profits from a change in price p is the quantity of good i times the change in the price plus the price of all goods times the changes in their respective quantities. But the quantity changes are second-order because the quantities solve the profit maximization first-order conditions, that price times the marginal change in quantities in technologically feasible directions is 0. Every advanced microeconomics text proves this. A result like this is true whenever the technology is convex, even if the technology is not smooth.

The convex version of Hotelling's lemma is a consequence of the inversion property of subdifferentials for concave and convex f; that $p\in\partial f(x)$ if and only if $x\in\partial f^{*}(p)$. See convex programming for a brief discussion.

Hotelling's lemma: $(x,y)\in\partial\pi(p,w)$ if and only if (x,y) is profit-maximizing at prices (p,w).

Hotelling's lemma is quickly argued. If $(x,y)\in\partial\pi(p,w) = \partial\delta^{F*}(p,w)$, then $(p,w)\in\partial\delta^{F**}(x,y) = \partial\delta^{F}(x,y)$. Then $\delta^{F}(x,y) + (p,w)\cdot((x',y') - (x,y)) \le \delta^{F}(x',y')$ for all (x',y'). This implies that $x\in F$ and furthermore that $(p,w)\cdot((x',y') - (x,y)) \le 0$ for all $(x,y)\in F$, in other words, that (x,y) is profit-maximizing at prices (p,w). Conversely, suppose that (x,y) is profit maximizing at prices (p,w). Then (p,w) satisfies the subgradient inequality of δ^{F} at (x,y), and so $(p,w)\in\partial\delta^{F}$. Consequently,

$$(x,y)\in\partial\delta^{F*}(p,w) \equiv \partial\pi(p,w).$$

The textbook treatment of duality observes that, if net supply is the first derivative of the profit function, then the own-price derivative of net supply must be the second own-partial derivative of profit with respect to price, and convexity of the profit function implies that this partial derivative should be positive, so net supply is increasing in price. The same fact follows in the convex framework from the monotonicity properties of the subgradients. Suppose that (w,p) and (w',p') are two price vectors, and suppose that (x,y) and (x',y') are two profit-maximizing production plans corresponding to the two price vectors. Then $(w-w',p-p')(x-x',y-y')\ge 0$. If the two price vectors are identical for all prices but, say, $p_k\neq p'_k$, then $(p_k-p'_k)(y_k-y'_k)\ge 0$, and net supply is non-decreasing in price. As with net supplies, some comparative statics of conditional factor demand with respect to input price changes follows from the monotonicity property of subgradients.

Another implication of profit function convexity and (twice continuous) differentiability is symmetry of the derivatives of net supply:

$$\frac{\partial y_k}{\partial p_l} = \frac{\partial^2 \pi}{\partial p_k \partial p_l} = \frac{\partial^2 \pi}{\partial p_l \partial p_k} = \frac{\partial y_l}{\partial p_k}.$$

The convex analysis version of this is that for any finite sequences of goods $i,\ldots,l$,

$$p_i \cdot (y_j - y_i) + p_j \cdot (y_k - y_j) + \cdots + p_l \cdot (y_i - y_l) \leq 0.$$

This requirement, which has a corresponding expression in terms of differences in prices, is called *cyclic monotonicity*. All subdifferential correspondences are cyclicly monotone. The connection with symmetry is not obvious, but it helps to know that Rockafellar (1974) leaves as an exercise (and so do we) that cyclic monotonicity is a property of a linear transformation corresponding to an $n \times n$ matrix M if and only if M is symmetric and positive semi-definite. Monotonicity is cyclic monotonicity for sequences of length 2.

The other famous result in duality theory for production is Shephard's lemma, which does for cost functions what Hotelling's lemma does for profit functions: conditional input demands are the derivatives of the cost functions. This is demonstrated in the same way, since the cost function and the indicator function for the set of inputs from which y is produceable are both convex and have closed hypographs.

4 Utility and expenditure functions

A quasi-concave utility function U defined on the commodity space $\mathbf{R}^{\mathbf{n}}_{+}$ has upper contour sets, the sets R_u of consumptions bundles which have utility at least u, which are convex. If u is usc, these sets are closed as well.

The *expenditure function* gives for each utility level u and price vector p the minimum cost of realizing utility u at prices p: $e(p, u) = \inf\{p \cdot x : u(x) \geq u\}$. If the infimum is actually realized at a consumption bundle x, then x is the *Hicksian* or *compensated real income demand*.

In terms of convex analysis, $e(p, u)$ is the conjugate of the concave indicator function $\phi_u(x)$ of the set $R(u) = \{x : U(x) \geq u\}$, that is, $e(p, u) \equiv \varphi_u^*(p)$. Thus $e(p, u)$ will be usc and concave in p for each u. The expenditure function is also linearly homogeneous in prices. If prices double, then the least cost of achieving u will double as well.

The duality of utility and expenditure functions is that each can be derived from the other; they are alternative characterizations of preference. Since the concave indicator function $\phi_u(x)$ is closed and convex, $e(\cdot, u)^* = \varphi_u(x)$. For fixed u, the Fenchel transform of the expenditure function is the concave indicator function of $R(u)$; $\inf_p p \cdot x - e(p, u)$ is 0 if $U(x) \geq u$ and $-\infty$ otherwise. If $x \in R(u)$, then the cost of x at any price p can be no less than the minimum cost necessary to achieve utility u. The gap between the cost of x and the cost of utility level u is made by taking ever smaller prices, and so its minimum is 0. Suppose that x is not in $R(u)$. The separation theorem for convex sets says there is a price p such that $p \cdot x < \inf_{y \in R(u)} p \cdot y$; there is a price at which x is cheaper than the cost of u. Now, by taking ever larger multiples of p, the magnitude of the gap can be made arbitrarily large, and so the value of the conjugate is $-\infty$. Thus the conjugate is the concave indicator function of $R(u)$.

Among the most useful consequence of the duality between utility and expenditure functions is the relationship between derivatives of the expenditure function and the Hicksian, or compensated, demand. Hicksian demand. The compensated demand at prices p and utility u are those consumption bundles in $R(u)$ which minimize expenditure at prices p. This result is just Shephard's lemma for expenditure functions:

Hicks compensated demand: Consumption bundle x is a Hicks compensated consumption bundle at prices p if and only if $x \in \partial_p e(p, u)$. Furthermore, if x is demanded at prices p and utility u, and y is demanded at prices q and the same utility u, then $(p-q)\cdot(x-y) \leq 0$.

The downward-sloping property just restates the monotonicity property of the subdifferential correspondence. For the special case of changes in a single price, the statement is that demand is non-increasing in its own price.

5 Equilibrium and optimality

The equivalence between Pareto optima and competitive equilibria can also be viewed as an expression of duality. When preferences have concave utility representations, *quasi-equilibrium* emerges from Lagrangean duality. Quasi-equilibrium entails feasibility, profit maximization, and expenditure minimization rather than utility maximization. That is, each trader's consumption allocation is expenditure minimizing for the level of utility it achieves. The now traditional route of Arrow (1952) and Debreu (1951) to the Second Welfare Theorem first demonstrates that a Pareto-optimal allocation can be regarded as a quasi-equilibrium for an appropriate set of prices. Under some additional conditions, the quasi-equilibrium is in fact a competitive equilibrium, wherein utility maximization on an appropriate budget set replaces expenditure minimization. Our concern here is with the first step on this path.

Suppose that each of I individuals has preferences represented by a concave utility function on $\mathbf{R}^{\mathbf{N}}_{+}$, and that production is represented, as in Section 3, by a closed and convex set F of feasible production plans. Suppose that $0 \in F$ (it is possible to produce nothing) and that the aggregate endowment e is strictly positive. Assume, too, that there is free disposal in production. Every Pareto optimum is the maximum of a Bergson–Samuelson social welfare function of the form $\sum_i \lambda_i u_i$ defined on the set of all consumption allocations. An allocation is a vector (x, y) where $x \in \mathbf{R}^{\mathbf{NI}}_{+}$ is a consumption allocation, a consumption bundle for each individual, and y is a production plan. The allocation is *feasible* if $y \in F$ and $y + e - \sum_i x_i \geq 0$. A Lagrangean for this convex program is

$$L(x, y, p) = \begin{cases} \sum_i u_i(x_i) + p \cdot (y + e - \sum_i x_i) & \text{if } x \in \mathbf{R}^{\mathbf{NI}}_{+},\ y \in F \text{ and } p \in \mathbf{R}^{\mathbf{L}}_{+}, \\ +\infty & \text{if } x \in \mathbf{R}^{\mathbf{NI}}_{+},\ y \in F \text{ and } p \notin \mathbf{R}^{\mathbf{L}}_{+}, \\ -\infty & \text{otherwise,} \end{cases}$$

where p is the vector of Lagrange multipliers for the L goods constraints.
The possibility of 0 production and the strict positivity of the aggregate endowment guarantee that the set of feasible solutions satisfies Slater's condition, and so a saddlepoint (x^*, y^*, p^*) exists; that is, $\sup_{x,y} L(x, y, p^*) \leq L(x^*, y^*, p^*) \leq L(x^*, y^*, p)$ for all $x \in \mathbf{R}^{\mathbf{NI}}$, $y \in F$ and $p \in \mathbf{R}^{\mathbf{L}}$. Then (x^*, y^*) is Pareto optimal and p^* solves the dual problem $\min_p \sup_{xy}, L(x, y, p)$. The interpretation of (x^*, y^*, p^*) as a quasi-equilibrium comes from examining the dual problem. The dual problem can be rewritten as

$$\inf_{p \in \mathbf{R}^{\mathbf{L}}_{+}} \sup_{x \in \mathbf{R}^{\mathbf{NI}}_{+},\, y \in F} L(x, y, p) = \inf_{p \in \mathbf{R}^{\mathbf{L}}_{+}} \sup_{x \in \mathbf{R}^{\mathbf{NI}}_{+},\, y \in F} \sum_i u_i(x_i) + p \cdot \left(y + e - \sum_i x_i\right) = \inf_{p \in \mathbf{R}^{\mathbf{L}}_{+}} \sum_i \sup_{x_i \in \mathbf{R}^{\mathbf{L}}_{+}} \left\{\lambda_i u_i(x_i) - p \cdot x_i\right\} + \sup_{y \in F} p \cdot y. \tag{1}$$

In the dual problem, the Lagrange multipliers can be thought of as goods prices. The Second Welfare Theorem interprets the optimal allocation as an equilibrium allocation using the Lagrange multipliers as equilibrium prices. To see this, look at the second line of (1). At prices p, a production plan is chosen from y to maximize profits $p \cdot y$, so the value of this term is $\pi(p)$. Each consumer is asked to solve

$$\max_i \lambda_i u_i(x_0) - p \cdot x = -\min p \cdot x - \lambda_i u_i(x_i) = \lambda_i u_i^* - \min p \cdot x - \lambda_i (u_i(x_i) - u_i^*)$$

where $u_i^* = u_i(x_i^*)$. The term being minimized is the Lagrangean for the problem of expenditure minimization, and so x_i^* is the Hicksian demand for consumer *i* at prices *p* and utility level $u_i^* = u_i(x_i^*)$. Finally, the optimal allocation is feasible, and so (x^*, y^*, p^*) is a quasi-equilibrium.
Given the observation about expenditure minimization, the saddle value of the Lagrangean is

$$\sum_i \lambda_i u_i^* - e_i(p^*, u_i^*) + \pi(p^*)$$

The planner chooses prices to minimize net surplus, which is the sum of profits from production and the excess of total Bergson–Samuelson welfare less the cost of the consumption allocation.

6 Historical notes

Duality ideas appeared very early in the marginal revolution. Antonelli, for instance, introduced the indirect utility function in 1886. The modern literature begins with Hotelling (1932), who provided us with Hotelling's lemma and cyclic monotonicity. Shephard (1953) was the first modern treatment of duality, making use of notions such as the support function and the separating hyperplane theorem.
The results on consumer and producer theory are surveyed more extensively in Diewert (1981), who also provides a guide to the early literature. In its focus on Fenchel duality, this review has not even touched on the duality between direct and indirect aggregators, such as utility and indirect utility, and topics that would naturally accompany this subject such as Roy's identity. Again, this is admirably surveyed in Diewert (1981).

See Also

- convex programming
- convexity
- duality
- Lagrange multipliers
- Pareto efficiency
- quasi-concavity

Bibliography

Arrow, K.J. 1952. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19, 273–92.

Diewert, W.E. 1981. The measurement of deadweight loss revisited. *Econometrica* 49, 1225–44.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply. *Journal of Political Economy* 40, 577–616.

Rockafellar, R.T. 1970. *Convex Analysis*. Princeton, NJ: Princeton University Press.

Rockafellar, R.T. 1974. *Conjugate Duality and Optimization*. Philadelphia: SIAM.

Shephard, R.W. 1953. *Cost and Production Functions*. Princeton, NJ: Princeton University Press.

How to cite this article

Blume, Lawrence E. "duality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000196> doi:10.1057/9780230226203.0411

The New Palgrave Dictionary of Economics Online

D ü hring, Eugen Karl (1833 – 1921)

Tom Bottomore

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bernstein, E.; Dühring, E. K.; Engels, F.; Marx, K. H.; positivism; private property; Schumpeter, F. A.

Article

Dühring was born on 12 January 1833 in Berlin and died on 21 September 1921 at Nowawes bei Potsdam. The son of a Prussian state official, Dühring studied law, philosophy and economics at the University of Berlin and practised law until blindness obliged him to abandon this career. He then became a *Privatdozent* at the University of Berlin, where he taught philosophy and economics from 1863 to 1877, and began to write voluminously on a wide range of subjects, from the natural sciences to philosophy, social theory and socialism, his aim being to construct a system of social reform based upon positive science. His system was expounded in a series of books on capital and labour (1865), the principles of political economy (1866), a critical history of philosophy (1869), a critical history of political economy and socialism (1871), and courses in political economy and philosophy (1873; 1875). Dühring was an adherent of positivism, concerned in his philosophical works to expound a ‘strictly scientific world outlook’, in opposition particularly to the Hegelian dialectic. His economic writings emphasize the role of political factors in the development of capitalism, and he argued that social injustice is not caused primarily by the economic system, but by social and political circumstances, the remedy being to control the misuse of private property and capital (not abolish them) through workers’ organizations and state intervention.

Schumpeter (1954, pp. 509–10), praised Dühring's history of mechanics (1873), which was awarded an academic prize, suggested that he would retain a prominent place in the history of anti-metaphysical and positivist currents of thought, and noted that he made an important criticism of Marxist theory in his argument that political causes had played a major part in constituting the property relations of capitalist society. In other respects, however, Schumpeter considered that Dühring had made no significant contribution to economic theory.

Engels, in his well-known book (originally published as a series of articles), *Herr Eugen Dühring's Revolution in Science* [*Anti-Dühring*] (1877–8), which has done more than anything else to keep

Dühring's name alive, took a much more critical view, deriding his work as a prime example of the ‘higher nonsense’ which infected German academic life. His philosophical views were dismissed by Engels as ‘vulgar materialism’ and compared unfavourably with the ‘revolutionary side’ of Hegel's dialectics; and in the chapter of *Anti-Dühring* devoted to the history of political economy (largely written by Marx, but not published in full until the third edition of the book in 1894), Dühring was castigated for his superficiality and theoretical misconceptions. It was, however, the concern with Dühring's programme of social reform, and its possible baleful effect on the developing labour movement (Eduard Bernstein, for example, was initially impressed by Dühring's *Cursus* of 1873, though soon repelled by his anti-Semitism) that originally provoked Engels's articles, and was countered in the final section of the book (frequently reprinted later as a separate text under the title *Socialism, Utopian and Scientific*) by an exposition of Marxist socialism which became enormously influential.
It seems doubtful that Dühring occupies more than a minor place in the history of economic and social thought, except for this encounter with Marx and Engels, though Schumpeter (1954, p. 509) called him a ‘significant thinker’ and the entry in the *Encyclopedia of the Social Sciences* (1931, vol. 5, p. 273) described his writings as ‘among the important intellectual achievements of the nineteenth century’.

Selected works

1871. *Kritische Geschichte der Nationalökomie und des Sozialismus*. Berlin: T. Grieben.

1873. *Cursus der National- und Sozialökonomie einschliesslich der Hauptpunkte der Finanzpolitik*. Berlin: T. Grieben.

1875. *Cursus der Philosophie als streng wissenschaftlicher Weltanschauung und Lebensgestaltung*. Leipzig: E. Koschny.

Bibliography

Albrecht, G. 1927. *Eugen Dühring: ein Beitrag zur Geschichte der Sozialwissenschaften*. Jena: G. Fischer.

Engels, F. 1877–8. *Anti-Dühring. Herr Eugen Dühring's Revolution in Science*. Moscow: Progress Publishers, 1947.

Schumpeter, J.A. 1954. *A History of Economic Analysis*. London: Allen & Unwin.

How to cite this article

Bottomore, Tom. "Dühring, Eugen Karl (1833–1921)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_D000197> doi:10.1057/9780230226203.0412

The New Palgrave Dictionary of Economics Online

dummy variables

Pietro Balestra

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The dummy-variable method is a useful device for introducing, into a regression analysis, information contained in qualitative or categorical variables, that is, in variables that are not conventionally measured on a numerical scale, such as race, sex, marital status, occupation, or level of education. It is a means for considering a specific scheme of parameter variation, in which the variability of the coefficients is linked to the causal effect of some precisely identified qualitative variable. But when the qualitative effects are generic, as in the cross-section time-series model, an interpretation in terms of random effects may seem more appealing.

Keywords

covariance model; cross-section time-series model; dummy variables; Engel curve; error component model; qualitative variables; random coefficient model

Article

In economics, as well as in other disciplines, qualitative factors often play an important role. For instance, the achievement of a student in school may be determined, among other factors, by his father's profession, which is a qualitative variable having as many attributes (characteristics) as there are professions. In medicine, to take another example, the response of a patient to a drug may be influenced by the patient's sex and the patient's smoking habits, which may be represented by two qualitative variables, each one having two attributes. The dummy-variable method is a simple and useful device for introducing, into a regression analysis, information contained in qualitative or categorical variables; that is, in variables that are not conventionally measured on a numerical scale. Such qualitative variables may include race, sex, marital status, occupation, level of education, region, seasonal effects, and so on. In some applications, the dummy-variable procedure may also be fruitfully applied to a quantitative variable such as age, the influence of which is frequently U-shaped. A system of dummy variables defined by age classes conforms to any curvature and consequently may lead to more significant results.

The working of the dummy-variable method is best illustrated by an example. Suppose we wish to fit an Engel curve for travel expenditure, based on a sample of *n* individuals. For each individual *i*, we have quantitative information on his travel expenditures (y_i) and on his disposable income (x_i), both variables being expressed in logarithms. A natural specification of the Engel curve is:

$$y_i = a + bx_i + u_i$$

where *a* and *b* are unknown regression parameters and u_i is a non-observable random term. Under the usual classical assumptions (which we shall adopt throughout this presentation), ordinary least-squares produce the best estimates for *a* and *b*.

Suppose now that we have additional information concerning the education level of each individual in the sample (presence or absence of college education). If we believe that the education level affects the travel habits of individuals, we should explicitly account for such an effect in the regression equation. Here, the education level is a qualitative variable with two attributes: college education; no college education. To each attribute, we can associate a dummy variable which takes the following form:

$$d_{1i} = \begin{cases} 1 \text{ if college education} \\ 0 \text{ if no college education} \end{cases}$$

$$d_{2i} = \begin{cases} 1 \text{ if no college education} \\ 0 \text{ if college education} \end{cases}$$

Inserting these two dummy variables in the Engel curve, we obtain the following expanded regression:

Specification I:

$$y_i = a_1 d_{1i} + a_2 d_{2i} + bx_i + u_i$$

which may be estimated by ordinary least-squares. Alternatively, noting that $d_{1i} + d_{2i} = 1$ for all *i*, we can write:

Specification II:

$$y_i = a_2 + (a_1 - a_2)d_{1i} + bx_i + u_i$$

which, again, may be estimated by ordinary least-squares.
It is easy to see how the procedure can be extended to take care of a finer classification of education levels. Suppose, for instance, that we actually have *s* education levels (*s* attributes). All we require is that the attributes be exhaustive and mutually exclusive. We then have the two following equivalent specifications:
Specification I:

$$y_i = a_1 d_{1i} + a_2 d_{2i} + \ldots + a_s d_{si} + bx_i + u_i$$

Specification II:

$$y_i = a_s + (a_1 - a_s)d_{1i} + \ldots + (a_{s-1} - a_s)d_{s-1,i} + bx_i + u_i.$$

Obviously, the two specifications produce the same results but give rise to different inpts. Specification I includes all the *s* dummy variables but no constant term. In this case, the coefficient of d_{ji} gives the specific effect of attribute *j*. Specification II includes $s - 1$ dummy variables and an overall constant term. The constant term represents the specific effect of the omitted attribute, and the coefficients of the different d_{ji} represent the contrast (difference) of the effect of the *j*th attribute with respect to the effect of the omitted attribute. (Note that it is not possible to include all dummy variables plus an overall constant term, because of perfect collinearity.)
It is important to stress that by the introduction of additive dummy variables, it is implicitly assumed that the qualitative variable affects only the intercept but not the slope of the regression equation. In our example, the elasticity parameter, *b*, is the same for all individuals; only the intercepts differ from individual to individual depending on their education level. If we are interested in individual variation in slope, we can apply the same technique, as long as at least one explanatory variable has a constant coefficient over all individuals. Take the initial case of only two attributes. If the elasticity parameter varies according to the level of education, we have the following specification:

$$y_i = a_1 d_{1i} + a_2 d_{2i} + b_1 d_{1i} x_i + b_2 d_{2i} x_i + u_i.$$

Simple algebra shows that ordinary least-squares estimation of this model amounts to performing two separate regressions, one for each class of individuals. If, however, the model contained an additional explanatory variable, say z_i, with constant coefficient c, by simply adding the term cz_i to the above equation, we would simultaneously allow for variation in the intercept and variation in the slope (for x). The dummy variable model also provides a conceptual framework for testing the significance of the qualitative variable in an easy way. Suppose we wish to test the hypothesis of no influence of the level of education on travel expenditures. The hypothesis is true if the s coefficients a_i are all equal; that is, if the $s - 1$ differences $a_j - a_s$, $j = 1, \ldots, s - 1$, are all zero. The test therefore boils down to a simple test of significance of the $s - 1$ coefficients of the dummy variables in Specification II. If $s = 2$, the t-test applied to the single coefficient of d_{1i} is appropriate. If $s > 2$, we may conveniently compute the following quantity:

$$\frac{(SS_c - SS) / (s - 1)}{SS / (n - s - 1)}$$

which is distributed as an F-variable with $s - 1$ and $n - s - 1$ degrees of freedom. In the above expression, SS is the sum of squared residuals for the model with the dummy variables (either Specification I or II), and SS_c is the sum of squared residuals for the model with no dummy variables but with an overall constant term.

In some economic applications the main parameter of interest is the slope parameter, the coefficients of the dummy variables being nuisance parameters. When, as in the present context, only one qualitative variable (with s attributes) appears in the regression equation, an easy computational device is available which eliminates the problem of estimating the coefficients of the dummy variables. To this end, it suffices to estimate, by ordinary least-squares, the simple regression equation:

Specification III:

$$y_i^* = bx_i^* + u_i^*.$$

where the quantitative variables (both explained and explanatory) for each individual are expressed as deviations from the mean over all individuals possessing the same attribute. For the dichotomous case presented in the beginning, for an individual with college education, we subtract the mean over all individuals with college education and likewise for an individual with no college education. Note, however, that the true number of degrees of freedom is not $n - 1$ but $n - 1 - s$. The same procedure also applies when the model contains other quantitative explanatory variables. The interested reader may

consult Balestra (1982) for the conditions under which this simple transformation is valid in the context of generalized regression.

The case of multiple qualitative variables (of the explanatory type) can be handled in a similar fashion. However, some precaution must be taken to avoid perfect collinearity of the dummy variables. The easiest and most informative way to do this is to include, in the regression equation, an overall constant term and to add for each qualitative variable as many dummy variables as there are attributes minus one. Take the case of our Engel curve and suppose that, in addition to the education level (only two levels for simplicity), the place of residence also plays a role. Let us distinguish two types of place of residence: urban and rural. Again, we associate to these two attributes two dummy variables, say e_{1i} and e_{2i}. A correct specification of the model which allows for both qualitative effects is:

$$y_i = a_1 + a_2 d_{1i} + a_3 e_{1i} + b x_1 + u_i.$$

Given the individual's characteristics, the measure of the qualitative effects is straightforward, as shown in the following table:

	Urban	Rural
College education	$a_1 + a_2 + a_3$	$a_1 + a_2$
No college education	$a_1 + a_3$	a_1

The specification given above for the multiple qualitative variable model corresponds to Specification II of the single qualitative variable model. Unfortunately, when there are two or more qualitative variables there is no easy transformation analogous to the one incorporated in Specification III, except under certain extraordinary circumstances (Balestra, 1982).

One such circumstance arises in connection with cross-section time-series models. Suppose that we have n individuals observed over t periods of time. If we believe in the presence of both an individual effect and a time effect, we may add to our model two sets of dummy variables, one corresponding to the individual effects and the other corresponding to the time effects. This is the so-called covariance model. The number of parameters to be estimated is possibly quite large when n or t or both are big. To avoid this, we may estimate a transformed model (with no dummies and no constant term) in which each quantitative variable (both explained and explanatory) for individual i and time period j is transformed by subtracting from it both the mean of the ith individual and the mean of the jth time period and by adding to it the overall mean. Note that, by this transformation, we lose $n + t - 1$ degrees of freedom.

To conclude, the purpose of the preceding expository presentation has been to show that the dummy-variable method is a powerful and, at the same time, simple tool for the introduction of qualitative effects in regression analysis. It has found and will undoubtedly find numerous applications in empirical economic research. Broadly speaking, it may be viewed as a means for considering a specific scheme of parameter variation, in which the variability of the coefficients is linked to the causal effect of some precisely identified qualitative variable. But it is not, by any means, the only scheme available. For instance, when the qualitative effects are generic, as in the cross-section time-series model, one may question the validity of representing such effects by fixed parameters. An interpretation in terms of

random effects may seem more appealing. This type of consideration has led to the development of other schemes of parameter variation such as the error component model and the random coefficient model.
A final remark is in order. In the present discussion, qualitative variables of the explanatory type only have been considered. When the qualitative variable is the explained (or dependent) variable, the problem of these *limited* dependent variables is far more complex, both conceptually and computationally.

Bibliography

Balestra, P. 1982. Dummy variables in regression analysis. In *Advances in Economic Theory*, ed. Mauro Baranzini. Oxford: Blackwell.

Goldberger, A.S. 1960. *Econometric Theory*, pp. 218–27. New York: John Wiley.

Maddala, G.S. 1977. *Econometrics*, ch. 9. New York: McGraw-Hill.

Suits, D.B. 1957. Use of dummy variables in regression equations. *Journal of the American Statistical Association* 52, 548–51.

How to cite this article

Balestra, Pietro. "dummy variables." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000198> doi:10.1057/9780230226203.0413

The New Palgrave Dictionary of Economics Online

Dunlop, John Thomas (1914 – 2003)

Richard B. Freeman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; dispute resolution; Dunlop, J.; industrial relations; labour's share of income; marginal productivity theory; mediation; wage contours; wage determination

Article

John Dunlop was an extraordinary labour economist, Professor and Dean of the Faculty at Harvard University, Secretary of Labor of the United States, and mentor to students and practitioners in the world of labour. He was extraordinary because he was more than an economist and because he was driven by a moral vision of what economists and academics should do to make the world better. Labour economists and policymakers paid close attention to Dunlop's thoughts because he combined academic research with unparalleled practical experience in solving problems and building institutions. His academic writings, which include several classic articles as well as major books, reflect Dunlop's participation in events and direct observations of social behaviour.

Dunlop first attracted academic attention with his 1938 *Economic Journal* article on the movement of real and money wages over the business cycle, which forced Keynes to admit that the *General Theory* was wrong on this issue: real wages fall in recessions not in booms, contrary to simple marginal productivity analysis. Quite an achievement for a 24-year-old economist. Dunlop followed this with *Wage Determination Under Trade Unions* (1944), in which he modelled unions as optimizing organizations; with analyses of the cyclic variation of labour's share, with the concept of ‘wage contours’ that captured the notion that product markets influenced wages, and with numerous analysis of wage determination, labour relations, mediation and dispute resolution. Dunlop's book *Industrial Relations Systems* (1958) sought to develop a broader perspective on how labour relations fit into economics.

In the 1980s, concerned that labour economists were limited in their conceptual vision by narrow optimizing models and in their empirical analysis by extant government data-sets, Dunlop carped at them for failing to see what he could see in the labour market. Dunlop saw the labour market as pre-eminently a social institution to resolve labour problems, which should be analysed as such rather than

as a bourse. His mode of analysis was that of a naturalist, who looks at the world with his own eyes and experience, with direct knowledge of the institutions and practitioners, without trying to force observation into a narrow conceptual framework.

Dunlop's career spanned a wide variety of activities. Earning his AB (1935) and Ph.D. (1939) from Berkeley, he rose to become professor of economics at Harvard and Dean of the University (1970–3), when he helped stabilize the university during a period of student disorders, and Lamont University Professor (1970–2003). He worked for the National War Labor Board (1943–54); served as member or chair on various national panels with responsibility for resolving labour disputes; led labour-management committees in areas ranging from missile sites to apparel, the public sector, and health; served as Director of the Cost of Living Council (1973–4), and as Secretary of Labor of the United States (1975–6). From 1993 to 1994 he chaired the Commission on the Future of Worker–Management Relations, popularly known as the Dunlop Commission, which was given the charge ‘to recommend ways to improve labor–management cooperation and productivity’. The politics and economics of the time were not right, however, for bringing management and labour to a consensus on modernizing labour relations, so that much of the Commission's recommendations went unheeded.

Dunlop approached his work –advising presidents and cabinet officials and telling academics about the real world and practitioners about academic theory – with one goal: to help solve problems. The moral principle that guided him – that academics should use their knowledge and skill to help solve problems faced by real people, by workers and firms, and governments – represents social science at its best.

Selected works

1938. The movement of real and money wage rates. *Economic Journal* 48, 413–34.

1944. *Wage Determination under Trade Unions*. New York: Macmillan.

1948. Productivity and the wage structure. In *Income, Employment and Public Policy: Essays in Honor of Alvin H. Hansen*. New York: W.W. Norton.

1948. The development of labor organizations: a theoretical framework. In *Insights into Labor Issues*, ed. R. Lester and J. Shister. New York: Macmillan.

1957. The task of contemporary theory. In *New Concepts in Wage Determination*, ed. G. Taylor and F. Pierson. New York: McGraw-Hill.

1958. *Industrial Relation Systems*. New York: Henry Holt & Co.

1984. *Dispute Resolution*. Dover, MA: Auburn House Publishing Co.

Bibliography

Segal, M. 1986. Post-institutionalism in labor economics: the forties and fifties revisited. *Industrial and Labor Relations Review* 39, 388–403.

How to cite this article

Freeman, Richard B. "Dunlop, John Thomas (1914–2003)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000201>
doi:10.1057/9780230226203.0414

The New Palgrave Dictionary of Economics Online

Dupuit, Arsene-Jules-Emile Juvenal (1804 – 1866)

Robert B. Ekelund, Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

antitrust enforcement; consumer surplus; cost–benefit analysis; deadweight loss; diminishing marginal utility; Dupuit, A.-J.-E. J.; Edgeworth, F.Y.; Hotelling, H.; law of demand; marginal cost pricing; Menger, C.; monopoly profit maximization; price discrimination; public works

Article

French engineer and economic theorist, born at Fossano, Piedmont, Italy on 18 May 1804, when this region was part of the French empire; died 5 September 1866 in Paris. After his parents returned to Paris in 1814, Dupuit continued his education in the secondary schools at Versailles, at Louis-le-Grand and at Saint-Louis, where he finished brilliantly by winning a physics prize in a large group of competitors. Accepted to the Ecole des Ponts et Chaussées in 1824, Dupuit soon distinguished himself as an engineer and, in 1827, was put in charge of an engineering district in the department of Sarthe, where he concentrated on roadway and navigation work. Dupuit's numerous and trenchant engineering studies on such topics as friction and highway deterioration, floods and hydraulics, and municipal water systems made him one of the most creative civil engineers of his day. Decorated for such contributions by the Legion of Honour in 1843, Dupuit ultimately became director-chief engineer in Paris in 1850 and Inspector-General of the Corps of Civil Engineers in 1855.
No less profound were Dupuit's contributions to general economic analysis and to the economic evaluation of public works (cost–benefit analysis). In fact, Dupuit was the most illustrious contributor in the long French tradition of study, teaching and writing on economic topics at the Ecole des Ponts et Chaussées, whose professors and students included Isnard, Henri Navier, Charles Minard, Emile Cheysson and Charles Ellet.
Led by a desire to evaluate the economic or *net* benefits of public provision, Dupuit directed his considerable analytical gifts to the utility foundation of demand and to its relevance to the welfare benefits of public works. In three substantial papers appearing in the *Annales des Ponts et Chaussées* (1844; 1849) and the *Journal des économistes* (1853), Dupuit became the first non-adventitious expositor of the theory of *marginal* utility, of (a variant of) marginal cost pricing, of simple and discriminating monopoly theory, and of pricing principles of the firm where location is a factor in expressing demand.
The font of Dupuit's contribution is the construction of a marginal utility curve and the identification of it with the demand curve or *courbe de consommation* (see Figure 1).
Figure 1

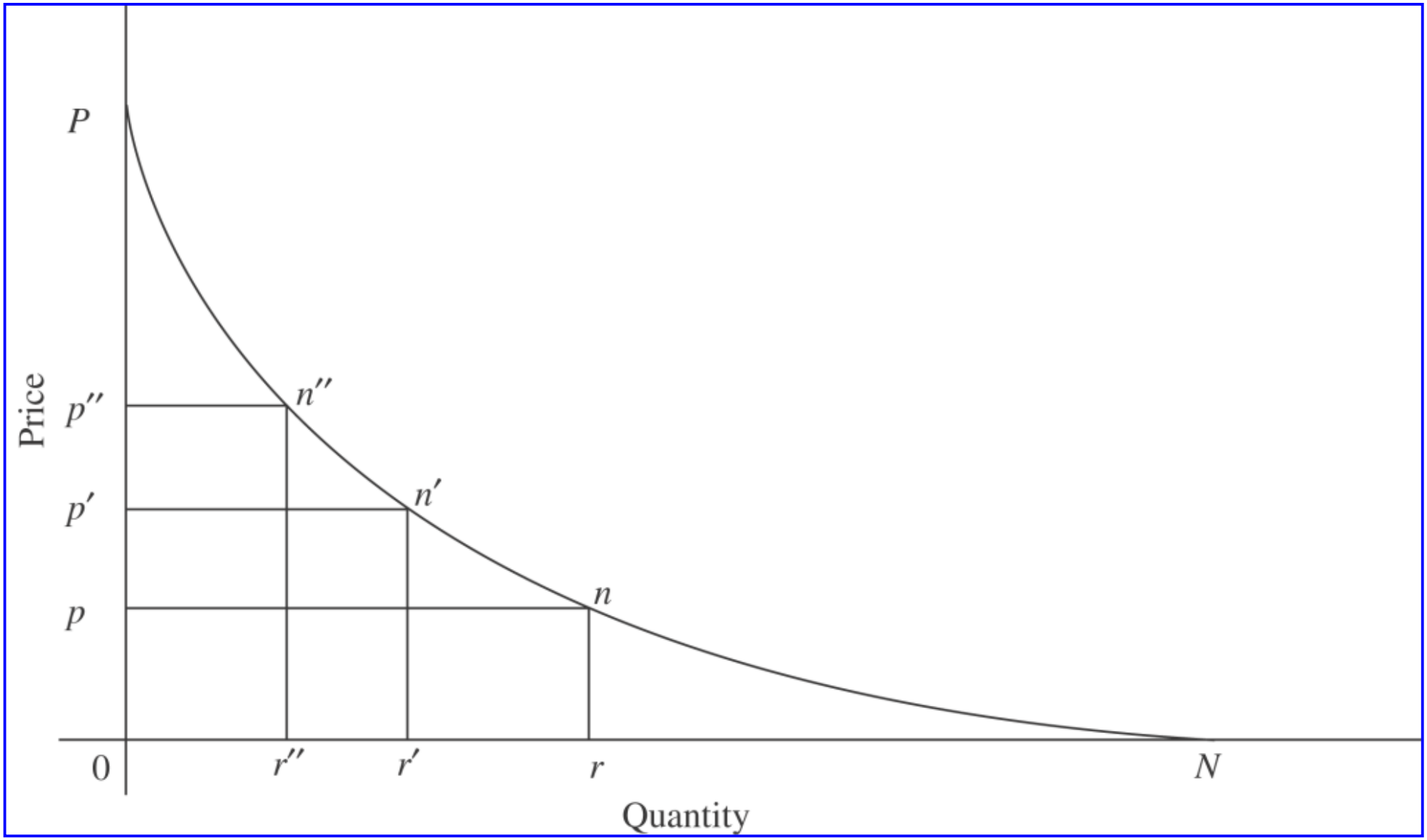

Arguing in the manner of Carl Menger, who later elaborated on the point, Dupuit showed that the marginal utility that an individual obtained from a homogeneous stock of goods is determined by the use to which the last units of the stock are put. In doing so, he clearly pointed out that the marginal utility of a stock or some particular good diminishes with increases in quantity and that each consumer attaches a different marginal utility to the same good according to the quantity consumed. The importance of Dupuit's invention rests in the fact that the psychological concept of diminishing marginal utility, and its ramifications, were carried over to the law of demand. With some, but not all, of the reservations and qualifications of Alfred Marshall, Dupuit *identified* the marginal utility curve with the demand curve, adding up the utility curves of individuals to obtain the market demand curve. Dupuit (1844, p. 106) described his construction (see Figure 1), which applied to all goods, public and private, as follows:

> If … along a line *Op* the lengths *Op*, *Op′*, *Op″* … represent various prices for an article, and that … *pn*, *p′n′*, *p″n″* … represent the number of articles consumed corresponding to these prices, then it is possible to construct a curve *Nn′n″P* which we shall call the curve of consumption. *ON* represents the quantity consumed when the price is zero, and *OP* the price at which consumption falls to zero.

The identification of marginal utility and demand, of course, sets up the demand curve as a welfare tool and Dupuit made specific calculations. A measure of the welfare produced by the good (*utilité absolue*) at quantity *Or* is the definite integral of the demand curve between *O* and *r*. Given that *Op* is the (average) cost of producing quantity *Or*, consumers earn a surplus (*utilité relative*) equal to absolute utility (*OrnP*) less costs of production (*Ornp*). (Relative utility (*pnP*) is none other than Marshall's consumers' surplus without all the reservations that Marshall attached to the concept.) Importantly, Dupuit identified area *rNn* as lost utility (*utilité perdue*). Under competitive conditions this loss was inevitable due to the opportunity cost of resources. Under a monopoly structure, for example, if, in Figure 1, *Op* were a monopoly price with zero production costs assumed, *utilité perdue* would be a loss to society – the 'deadweight' loss associated with excise taxes, tariffs or monopoly. Further, Dupuit advanced the theorem that the loss in utility was proportional to the *square* of the tax of price above marginal cost. This theorem, with attendant analysis, formed the base for large areas of neoclassical welfare economics, including the taxation studies of F.Y. Edgeworth and the marginal cost pricing argument of Harold Hotelling.

From this theoretical base, Dupuit investigated an impressive number of pricing systems and market models (1849). While Dupuit was an ardent and stubborn defender of laissez faire in most markets (1861), he was equally concerned that public works, provided or regulated by government as a *last* resort, should produce the maximum amount of utility possible. Thus tools such as marginal cost pricing find their theoretical foundations in the writings of Dupuit. Although Dupuit did not provide an explicit formulation of the principle, one of his bridge pricing examples and other statements strongly suggest the possibilities of such a technique to maximize welfare, but as a *long-run* proposition.

Dupuit analysed, independently of Cournot, who was apparently unknown to him, the profit-maximizing behaviour of the simple monopolist. He saw monopoly at the apex of a range of problems regarding the production of total welfare, being unconcerned about the 'distribution' of welfare between producers and consumers. His point was that the amount of 'absolute utility' (or what could be called net benefit) was lessened by monopoly profit maximization. This led him to defend the private practice of price discrimination and to produce an economic theory of discrimination. Price discrimination could exist, in Dupuit's view, with differences in 'buyer estimates', with the ability to segment markets either naturally or artificially, and with some degree of monopoly power. The motive was profit maximization, and although Dupuit discussed the effects of discrimination on price and revenue, he was primarily interested in the fact, as was Joan Robinson later, that discrimination could affect the size of the welfare benefit. This view was expanded to include the impact of price discrimination of welfare when buyers were spatially distributed (1849; 1854).

In the matter of policy, Dupuit recommended that tools be carefully fit to specific problems. If industries were to be collectivized or regulated by government, Dupuit proposed the maximization of net benefit under the constraint of covering total costs of production. The recovery of total cost might be achieved through regulated or constrained price discrimination or through a cost-based single price technique. However, Dupuit can hardly be credited with espousing an enlarged role for government or government intervention. A firm adherent of Smith's dictums concerning minimal government, Dupuit believed that free and open competition, along with vigorous antitrust or anticartel enforcement, would ensure optimal provisions in most cases, including transportation. Indeed, in the process of analysing the welfare principles of public works pricing, Dupuit discovered (in an uncommonly complete manner) some of the critical welfare-maximizing properties of a generalized competitive system.

See Also

- consumer surplus
- public utility pricing and finance

Selected works

1844. On the measurement of the utility of public works. Trans. R.H. Barback from the *Annales des Ponts et Chaussées*, in *International Economic Papers*, No. 2, London: Macmillan, 1952.

1849. On tolls and transport charges. Trans. E. Henderson from the *Annales des Ponts et Chaussées*, in *International Economic Papers*, No. 11, London: Macmillan, 1962.

1853. On utility and its measure – on public utility. *Journal des économistes* 36, 1–27.

1854. Péages. In *Dictionnaire de l'Economie Politique*, vol. 2. Paris: Guillaumin.

1861. *La Liberté Commerciale*. Paris: Guillaumin.

1934. *De l'Utilité et sa Mésure: écrits choisis et republiés*, ed. M. de Bernardi. Turin: La Riforma Sociale.

Bibliography

Ekelund, R.B., Jr. 1968. Jules Dupuit and the early theory of marginal cost pricing. *Journal of Political Economy* 76, 462–71.

Ekelund, R.B., Jr. 1970. Price discrimination and product differentiation in economic theory: an early analysis. *Quarterly Journal of Economics* 84, 268–78.

Ekelund, R.B., Jr. and Yeung-Nan Shieh. 1986. Dupuit, spatial economics, and optimal resource allocation: a French tradition. *Economica* 53, 483–96.

How to cite this article

Ekelund, Robert B., Jr. "Dupuit, Arsene-Jules-Emile Juvenal (1804–1866)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000205> doi:10.1057/9780230226203.0416

The New Palgrave Dictionary of Economics Online

durable goods markets and aftermarkets

Michael Waldman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There is an extensive literature on durable-goods markets that starts with the work of Akerlof, Coase, and Swan in the early 1970s. In this entry I survey the literature by starting with the three theoretical building blocks of time inconsistency, adverse selection, and substitutability between new and used units. I then focus on our understanding of three important real-world issues. These are whether firms choose optimal durability levels, whether firms have incentives to eliminate second-hand markets, and reasons for leasing. The article also provides an extensive discussion of aftermarket monopolization.

Keywords

adverse selection; aftermarket monopolization; aftermarkets; Akerlof, G.; asymmetric information; bundling; Coase, R.; commitment; complementary goods; deadweight loss; durable goods markets; durable-goods monopoly problem; hold-up theories; information costs; leasing; market power; optimal durability; price discrimination; product-line pricing problem; second-hand markets; time inconsistency; tying

Article

Durable goods are goods whose useful lifetime spans multiple periods.
This article surveys the extensive literature on durable-goods markets and aftermarkets. I begin with the main theoretical ideas, then turn to specific real-world issues such as durability choice and leasing, discuss aftermarket monopolization and then end with a brief conclusion. (A more in-depth survey appears in Waldman, 2003.)

Three theoretical building blocks

Much of our understanding of durable-goods markets derives from three theoretical contributions. The

first is Coase's (1972) insight concerning time inconsistency. To see the basic logic, consider Bulow's (1982) formalization: a durable-goods monopolist sells its output in each of two periods and cannot commit in the first period to second-period actions. Bulow shows that, because in the second period the firm does not internalize how its actions affect the value of used units, its output is higher than under commitment. First-period purchasers anticipate this, pay less for new units and thus lower overall monopoly profitability.

Coase's insight has spawned a large literature. One branch of this literature focuses on the Coase conjecture, that is, the idea that in an infinite-period setting time inconsistency causes price to drop immediately to marginal cost. A second branch identifies tactics such as leasing that firms can employ to reduce or possibly avoid time inconsistency. Finally, a third branch applies time inconsistency to other issues, including new-product introductions and repurchase prices.

The second major theoretical contribution is Akerlof's (1970) adverse-selection argument. This paper helped start the asymmetric-information revolution, but was not initially thought of as an important contribution to durable-goods theory. However, the paper's main example concerns second-hand markets. In Akerlof's model buyers have higher valuations than sellers, so efficiency requires that all units be traded. Further, each seller is privately informed of his own unit's quality. The result is a single price that reflects average quality, and sellers with high-quality units keep them because prices do not reflect actual quality, that is, trade is below the efficient level. (In Akerlof's analysis there is no trade, but this result is not robust.)

A small empirical literature looks for evidence of adverse selection in durable-goods markets. Most of these papers find some support. For example, Bond (1982) considers the used pickup truck market and finds support for adverse selection for older trucks, while Genesove (1993) finds some supporting evidence in used-car dealer auctions. More recently, Gilligan (2004) finds supporting evidence in business aircraft.

In terms of durable-goods theory, Akerlof's contribution was ignored for almost 30 years. Starting with Hendel and Lizzeri (1999a), however, a number of papers have extended Akerlof's analysis. There are three basic findings. First, Akerlof's main results continue to hold when new units are incorporated into the analysis. Second, because adverse selection in the used-unit market reduces the willingness to pay of new-unit buyers, firms will market new units in a manner that reduces adverse selection. Third, as discussed in detail later, new-unit leasing can be important for reducing adverse selection.

The third major theoretical contribution is that there is a close analogy between the product-line pricing problem and the durable-goods monopoly problem. This analogy is described in Waldman (1996). Consider Mussa and Rosen (1978), which analyses the product-line pricing problem of a non-durable-goods monopolist. The monopolist sells units of varying qualities to consumers who have heterogeneous valuations on quality. Because the substitutability between units links the various prices, the monopolist lowers below efficient levels the quality level sold to all but the highest-valuation group.

Now consider a durable-goods monopolist who controls the quality of a unit at every age. Further, assume heterogeneity in consumers’ valuations for quality and a frictionless second-hand market. Then, if the firm can commit, quality choices are as above. That is, new-unit quality is efficient. But, because of the linkages between the various prices, all used-unit qualities are below efficient levels. As discussed later, a number of recent papers use this result to analyse various real-world issues concerning durable goods.

Three real-world issues

Optimal durability choice

A much debated issue is whether a durable-goods monopolist chooses socially optimal durability. Swan (1970; 1971) considers models that satisfy the once standard assumption that a unit is a bundle of ‘service units’, so some number of used units is a perfect substitute for a new unit. Swan's steady-state analysis shows durability choice to be socially optimal because the firm produces the steady-state flow of service units at minimum cost. (Swan's analysis corrected the conclusions of earlier papers that had concluded that in such settings the monopolist would choose inefficiently low durability levels.)

A large literature investigates the robustness of Swan's conclusions. There are two major findings. The first employs time inconsistency. Bulow (1986) moves away from Swan's assumption of steady-state behaviour by considering a model similar to his earlier one, but now allows endogenous durability choice. He shows that time inconsistency provides a rationale for a durable-goods monopolist to choose less than the socially-optimal durability level. The logic is that durability is what leads to time inconsistency, so reducing durability below the efficient level reduces time inconsistency and thus increases profitability.

The second major finding appears in Waldman (1996) and Hendel and Lizzeri (1999b), which drop the service units assumption and instead assume that new and used units vary in quality and that durability choice controls the speed of quality deterioration. The earlier discussion immediately translates into an incentive for the firm to choose less than the socially optimal durability level. That is, in this setting the incentive for the monopolist to sell output whose used-unit quality is below the efficient level translates into durability below the efficient level. (In Hendel and Lizzeri's analysis durability choice can be above, below, or equal to the first-best level, but it is always below the second-best level defined by actual outputs.)

Eliminating second-hand markets

Do durable-goods producers with market power have incentives to eliminate second-hand markets? For example, do textbook publishers introduce new editions in order to kill off the market for used books? Until recently, the standard argument, found, for example, in Swan (1980), was that, since the new-unit price reflects prices the product will sell for on the second-hand market in subsequent periods, the producer has no such incentive.

Two recent arguments show that this result is, in fact, quite limited. The first, which builds on the discussion above, appears in Waldman (1996; 1997) and Hendel and Lizzeri (1999b). The idea is that, because substitutability between new and used units means the price of a used unit on the second-hand market limits the amount the firm can charge for new units, the firm sometimes eliminates the second-hand market or similarly reduces used-unit availability in order to raise the new-unit price. In particular, this is more likely when consumers of used units have low valuations for the firm's product. This is both because little revenue is lost by not serving such consumers and because serving them means a low used-unit price and thus a lower new-unit price. (A number of earlier papers find similar results starting with

demand functions rather than utility maximization.)

The second argument, found initially in Waldman (1993), employs time inconsistency. As discussed, the early literature on time inconsistency focused on output choice. My 1993 paper shows time inconsistency also applies to actions such as new-product introductions that make used units unavailable because they become obsolete. The difference between this argument and the one above concerns commitment. Above it is assumed the firm can commit, so the firm eliminates the second-hand market only when it is profitable to do so. In contrast, here commitment is not assumed, so the firm may eliminate the second-hand market even though this lowers overall profitability.

A related empirical analysis appears in Iizuka (2004), which shows that the market share of used textbooks is an important determinant of whether or not a publisher introduces a new edition. This is consistent with new editions being used at least partly to eliminate second-hand markets, although Iizuka does not distinguish between the two possibilities described above for why a firm might want to do this. In future research, it might be possible to identify which argument is at work by focusing on how the decision to introduce new editions affects overall profitability.

Reasons for leasing

A number of reasons have been identified for why durable-goods producers frequently lease. (A reason I do not discuss is that there are sometimes tax advantages associated with leasing.) One reason, initially discussed in Coase (1972) and Bulow (1982), is that time inconsistency lowers profitability when a firm sells output because it chooses actions in later periods that inefficiently lower the value of used units. When the firm leases, however, it retains ownership of those units so the incentive to take inefficient actions disappears.

A second reason is also related to a previous discussion. As discussed in Waldman (1997) and Hendel and Lizzeri (1999b), when used-unit prices serve as important constraints on the new-unit price, leasing can be used to eliminate second-hand markets or at least reduce used-unit availability. The logic is that leasing allows a firm to eliminate the second-hand market by allowing the firm to retire returned used units. My 1997 paper shows this formally and argues that it is consistent with classic cases concerning the use of a lease-only policy such as United Shoe in the shoe machinery market, IBM in the computer market, and Xerox in the copier market. (One might argue that leasing is not needed because a firm can sell and then use high repurchase prices to purchase and retire used units. My 1997 paper shows this strategy is inferior to leasing because of time inconsistency.)

Finally, leasing is a response to adverse selection. This argument appears in Hendel and Lizzeri (2002) and Johnson and Waldman (2003). These papers show that, whether the new-unit market is monopolistic or competitive, in a world of asymmetric information leasing in the new-unit market can arise because it means used units are returned to the seller(s), which, in turn, avoids or at least reduces adverse selection in the used-unit market. The two papers develop different variants of the argument and show it is consistent with various empirical findings concerning the automobile market.

Aftermarket monopolization

Aftermarket monopolization is behaviour that stops alternative producers from selling aftermarket

products to the firm's customers. The focus on this subject started after the US Supreme Court's 1992 decision in the case *Eastman Kodak Company v. Image Technical Services*. Aftermarkets are common with durable goods, where aftermarkets refer to markets for complementary products such as maintenance and upgrades. I consider three possibilities: (*a*) hold-up rationales; (*b*) price discrimination and efficiency rationales; and (*c*) other strategic rationales.

Hold-up

There are two distinct hold-up arguments, each of which focuses on aftermarket monopolization by competitive producers. In both, the firm prohibits other firms from selling the aftermarket product – for example, maintenance – and then exploits the locked-in positions of its customers in pricing the product. The result is a standard deadweight loss due to the high aftermarket price, although no transfer between the consumers and the firm since competition in the primary market means firms earn zero profits overall. In the ‘costly-information’ version, consumers ignore the aftermarket price when purchasing the primary product. In the ‘lack-of-commitment’ version, developed in Borenstein, Mackie-Mason and Netz (1995), consumers correctly anticipate the aftermarket price but, because firms cannot commit, time inconsistency causes firms to monopolize the aftermarket and inefficiently raise the aftermarket price after consumers are locked in. (A third hold-up theory is the ‘surprise’ theory. In this argument consumers are surprised by the aftermarket monopolization. Some discussions of this theory describe a transfer between the consumers and the firm, but it is unclear why competition does not result in zero profits, in which case the surprise and costly-information theories are equivalent.)

Price discrimination and efficiency rationales

For various reasons, such as that many buyers in the relevant industries are sophisticated firms for which the costly-information argument is implausible, attention has shifted towards other arguments many of which have either neutral or positive social-welfare implications.

One such argument is the price discrimination argument that appears in Chen and Ross (1993) and Klein (1993). Suppose the primary-good producer has market power. Then the firm may monopolize the aftermarket in order to raise the aftermarket price and in this way price discriminate by charging a high aggregate price to the high-volume/high-valuation consumers. From a social-welfare standpoint, this argument has neutral implications since an improved ability to price discriminate can either raise or lower social welfare. (Klein argues that this argument applies even when firms are competitive, although not perfectly competitive.)

A plausible efficiency rationale follows from Schmalensee's (1974) argument that, given a durable-goods monopolist (which means new units priced above marginal cost) and a competitive maintenance market (which means maintenance is priced at cost), consumers will sometimes inefficiently maintain rather than replace used units. Tirole (1988) shows this can lead to aftermarket monopolization in a durable-goods monopoly setting because having a monopoly in both markets allows the firm to avoid the inefficiency and thus increases its profits.

More recently, Morita and Waldman (2005) and Carlton and Waldman (2006) show that the argument extends to aftermarkets other than maintenance, and to competitive durable-goods markets given

switching costs. In the latter case the inefficient substitution problem arises even with competition because switching costs create market power at the time of the maintenance/replacement decision. Interestingly, because competitive sellers earn zero profits in equilibrium, when aftermarket monopolization eliminates the distortion, both social welfare and consumer welfare increase.

Strategic rationales

There is an extensive literature on strategic rationales for the tying of complementary products. Since the tying of primary and aftermarket products is one potential way to achieve aftermarket monopolization, much of this literature is relevant to aftermarket monopolization.

Whinston (1990) shows that, if the primary good is not essential, tying may force the exit of an alternative producer of the complementary good and in this way increase the firm's profits by monopolizing the segment of the complementary-good market for which the primary good is not required.

In contrast, in Carlton and Waldman (2002) tying is sometimes used to preserve a monopoly in the primary-good market. They consider two-period settings in which a single potential entrant can enter the complementary market in either period but the primary market only in the second. In the presence of fixed costs of entry or network externalities, the primary-good monopolist sometimes ties in order to preserve its primary-good monopoly in the second period. For example, with entry costs tying stops the alternative producer from entering the complementary market in the first period. In turn, because of a possible inability to cover entry costs, the outcome can be no entry in either market in either period.

A third argument appears in Carlton and Waldman (2005). Whinston shows that in one-period settings there is never an incentive to tie if the monopolist's primary product is essential. Carlton and I show that in durable-goods settings, given the presence of complementary-good upgrades and switching costs, tying can be optimal even when the primary product is essential. The basic logic is that some profits are realized in later periods in the sale or lease of the upgraded complementary good, and the only way the monopolist can ensure it captures those profits is by tying and becoming the sole producer of the complementary good.

Conclusion

Starting in the early 1970s with the work of Akerlof, Coase and Swan, significant progress has been made in our understanding of durable-goods markets. In this entry I have surveyed this literature as well as the literature on the related issue of aftermarkets. Although I have referred throughout to various empirical papers, durable-goods markets is a topic for which theory is far ahead of empirical investigation. In the future I expect to see work that extends the theory in various important ways, but also empirical work that tests the validity of the various theoretical approaches that have been explored since the early 1970s.

See Also

- adverse selection

- Akerlof, George Arthur
- bundling and tying
- Coase, Ronald Harry
- resale markets

Bibliography

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Bond, E. 1982. A direct test of the ‘lemons’ model: the market for used pick-up trucks. *American Economic Review* 72, 836–40.

Borenstein, S., Mackie-Mason, J. and Netz, J. 1995. Antitrust policy in aftermarkets. *Antitrust Law Journal* 63, 455–82.

Bulow, J. 1982. Durable goods monopolists. *Journal of Political Economy* 90, 314–32.

Bulow, J. 1986. An economic theory of planned obsolescence. *Quarterly Journal of Economics* 101, 729–49.

Carlton, D. and Waldman, M. 2002. The strategic use of tying to preserve and create market power in evolving industries. *RAND Journal of Economics* 33, 194–220.

Carlton, D. and Waldman, M. 2005. Tying, upgrades, and switching costs in durable goods markets. Mimeo, Cornell University.

Carlton, D. and Waldman, M. 2006. Competition, monopoly, and aftermarkets. Mimeo, Cornell University.

Chen, Z. and Ross, T. 1993. Refusals to deal, price discrimination and independent service organizations. *Journal of Economics and Management Strategy* 2, 593–614.

Coase, R. 1972. Durability and monopoly. *Journal of Law and Economics* 15, 143–9.

Genesove, D. 1993. Adverse selection in the wholesale used car market. *Journal of Political Economy* 101, 644–65.

Gilligan, T. 2004. Lemons and leases in the used business aircraft market. *Journal of Political Economy* 112, 1157–80.

Hendel, I. and Lizzeri, A. 1999a. Adverse selection in durable goods markets. *American Economic Review* 89, 1097–115.

Hendel, I. and Lizzeri, A. 1999b. Interfering with secondary markets. *RAND Journal of Economics* 30, 1–21.

Hendel, I. and Lizzeri, A. 2002. The role of leasing under adverse selection. *Journal of Political Economy* 110, 113–43.

Iizuka, T. 2004. An empirical analysis of planned obsolescence. Mimeo, Vanderbilt University.

Johnson, J. and Waldman, M. 2003. Leasing, lemons, and buy-backs. *RAND Journal of Economics* 34, 247–65.

Klein, B. 1993. Market power in antitrust: economic analysis after *Kodak*. *Supreme Court Economic Review* 3, 43–92.

Morita, H. and Waldman, M. 2005. Competition, monopoly maintenance, and consumer switching costs. Mimeo, Cornell University.

Mussa, M. and Rosen, S. 1978. Monopoly and product quality. *Journal of Economic Theory* 18, 301–17.

Schmalensee, R. 1974. Market structure, durability, and maintenance effort. *Review of Economic Studies* 41, 277–87.

Swan, P. 1970. Durability of consumption goods. *American Economic Review* 60, 884–94.

Swan, P. 1971. The durability of goods and regulation of monopoly. *Bell Journal of Economics* 2, 347–57.

Swan, P. 1980. Alcoa: the influence of recycling on monopoly power. *Journal of Political Economy* 88, 76–99.

Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

Waldman, M. 1993. A new perspective on planned obsolescence. *Quarterly Journal of Economics* 108, 273–83.

Waldman, M. 1996. Durable goods pricing when quality matters. *Journal of Business* 69, 489–510.

Waldman, M. 1997. Eliminating the market for secondhand goods. *Journal of Law and Economics* 40,

61–92.

Waldman, M. 2003. Durable goods theory for real world markets. *Journal of Economic Perspectives* 17 (1), 131–54.

Whinston, M. 1990. Tying, foreclosure, and exclusion. *American Economic Review* 80, 837–59.

How to cite this article

Waldman, Michael. "durable goods markets and aftermarkets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000240>
doi:10.1057/9780230226203.0417

The New Palgrave Dictionary of Economics Online

Durbin-Watson statistic

James G. MacKinnon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Durbin–Watson statistic; linear regression models; Monte Carlo test; ordinary least squares (OLS) estimator; serial correlation; testing; DW Statistic

Article

The well-known Durbin–Watson, or DW, statistic, which was proposed by Durbin and Watson (1950; 1951), is used for testing the null hypothesis that the error terms of a linear regression model are serially independent.
Consider the linear regression model with AR(1) errors,

$$y_t = \boldsymbol{X}_t b + u_t,\ u_t = \rho u_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathrm{IID}(0, \sigma^2). \tag{1}$$

Here the scalar y_t is an observation on a dependent variable, $\boldsymbol{X}_t$ is a $1 \times k$ vector of observations on independent variables that may be treated as fixed, and β is a k-vector of parameters to be estimated. There are n observations, and we wish to test the null hypothesis that $\rho = 0$, under which the model (1) reduces to

$$y_t = \boldsymbol{X}_t b + u_t,\ u_t \sim \mathrm{IID}(0, \sigma^2), \tag{2}$$

for which the ordinary least squares (OLS) estimator is efficient. This estimator is usually written as $\hat{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, where the $n \times k$ matrix $\boldsymbol{X}$ has t^{th} row $\boldsymbol{X}_t$ and the n-vector $\boldsymbol{y}$ has t^{th} element y_t. The Durbin–Watson d statistic for testing (2) against (1) is solely a function of the OLS residuals $\hat{u}_t = y_t - \boldsymbol{X}_t\hat{b}$. It is defined as

$$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2}. \tag{3}$$

It is easy to see that d is approximately equal to $2 - 2\hat{\rho}$, where $\hat{\rho}$ is the OLS estimate of ρ in a regression of $\hat{u}_t$ on $\hat{u}_{t-1}$. Thus d will be approximately equal to 2 if the residuals do not display any serial correlation, and it will be less (greater) than 2 whenever $\hat{\rho}$ is more than a little bit greater (less) than 0.

The exact distribution of the DW statistic

The DW statistic can be written as a ratio of quadratic forms in the n-vector $\hat{\boldsymbol{u}}$ of OLS residuals, the tth element of which is $\hat{u}_t$. Specifically,

$$d = \frac{\hat{\boldsymbol{u}}'\boldsymbol{A}\hat{\boldsymbol{u}}}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}, \tag{4}$$

where $\boldsymbol{A}$ is the $n \times n$ matrix

$$\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

Durbin and Watson (1950) actually considered a number of statistics that can be written in the form of (4) for different choices of the matrix $\boldsymbol{A}$ and chose to focus on d for reasons of computational and theoretical convenience. Because both the numerator and the denominator are proportional to σ^2, d is invariant to σ.

The exact distribution of d depends on $\boldsymbol{X}$ and the distribution of the u_t. When the error terms are i.i.d. normal, Durbin and Watson (1951) tabulated bounds on the critical values for tests based on d against the one-sided alternative that $\rho > 0$. These bounds, denoted d_L and d_U, depend on the sample size and the number of regressors. We can reject the null hypothesis when $d < d_L$, cannot reject it when $d < d_U$, and can draw no firm conclusion when $d_L < d < d_U$. To test against the alternative that $\rho < 0$, we would replace d by $4 - d$ and use the same procedure.

The original Durbin–Watson tables have been extended by various authors, notably Savin and White (1977). However, since $d_U - d_L$ can be quite large, tests based on the bounds often have indeterminate outcomes. It is much better to perform exact tests conditional on $\boldsymbol{X}$, and this is easy to do with modern computing technology. There are two approaches.

The first approach is to calculate an exact P value for d using one of several methods for calculating the distribution of a ratio of quadratic forms in normal random variables. The method of Imhof (1961) is probably the best known of these, but the more recent method of Ansley, Kohn and Shively (1992) is faster. If a suitable computer program is readily available, this approach is the best one.

An alternative approach is to perform a Monte Carlo test. As can be seen from (4), the statistic d depends only on the vector $\boldsymbol{u}$ and the matrix $\boldsymbol{X}$, since $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, where $\boldsymbol{M}_{\boldsymbol{X}} = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Because of its invariance to σ, d does not depend on any unknown parameters. This implies that a Monte Carlo test will be exact.

To perform a Monte Carlo test at level α, we first choose B such that $\alpha(B+1)$ is an integer (999 is often a reasonable choice) and generate B vectors $\boldsymbol{u}_j^*$, each of which is multivariate standard normal. Each of the $\boldsymbol{u}_j^*$ is regressed on $\boldsymbol{X}$ to calculate a vector of residuals $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}_j^*$, which is then used to compute a simulated test statistic d_j^* according to (3). We can then calculate simulated P values for a one-tailed test against either $\rho > 0$ or $\rho < 0$ or for a two-tailed test. For example, the simulated P value for a one-tailed test against $\rho > 0$ is

$$P^*(d) = \frac{1}{B}\sum_{j=1}^{B} I(d_j^* < d),$$

where $I(\cdot)$ is the indicator function that is equal to 1 when its argument is true and equal to 0 otherwise.

We reject the null hypothesis whenever $P^*(d) < \alpha$. For more on the calculation of *P* values for bootstrap and Monte Carlo tests, see Davidson and MacKinnon (2006).

Limitations of the DW statistic

The Durbin–Watson statistic is valid only when all the regressors can be treated as fixed. It is not valid, even asymptotically, when $\boldsymbol{X}_t$ includes a lagged dependent variable or any variable that depends on lagged values of y_t. Because $\hat{\rho}$ is biased towards 0 when $\boldsymbol{X}_t$ includes a lagged dependent variable, *d* is biased towards 2 in this case. Thus, a test based on the DW statistic will tend to under-reject when the null hypothesis is false.

Numerous procedures have been proposed for testing for serial correlation in models that include lagged dependent variables. The simplest is to rerun regression (2), with the addition of the lagged residuals from that regression. The test statistic is then the *t* statistic on the lagged residuals. This procedure, which is due to Durbin (1970) and Godfrey (1978), does not yield an exact test and should be bootstrapped when the sample size is small.

Of course, since the finite-sample distribution of the DW statistic depends on the distribution of the u_t, we cannot expect to obtain an exact test even when the $\boldsymbol{X}_t$ are exogenous if the normality assumption is not a good one. In principle, we could bootstrap *d* by using re-sampled residuals instead of multivariate standard normal vectors for the $\boldsymbol{u}_j^*$. This would probably work very well in most cases, but it would not actually yield an exact test.

See Also

- artificial regressions
- serial correlation and serial dependence

Bibliography

Ansley, C., Kohn, R. and Shively, T. 1992. Computing *p*-values for the generalized Durbin–Watson statistic and other invariant test statistics. *Journal of Econometrics* 54, 277–300.

Davidson, R. and MacKinnon, J. 2006. Bootstrap methods in econometrics. In *Palgrave Handbooks of Econometrics: Volume 1 Econometric Theory*, ed. T. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Durbin, J. 1970. Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica* 38, 410–21.

Durbin, J. and Watson, G. 1950. Testing for serial correlation in least squares regression I. *Biometrika*

37, 409–28.

Durbin, J. and Watson, G. 1951. Testing for serial correlation in least squares regression II. *Biometrika* 38, 159–77.

Godfrey, L. 1978. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica* 46, 1293–301.

Imhof, J. 1961. Computing the distribution of quadratic forms in normal variables. *Biometrika* 48, 419–26.

Savin, N. and White, K. 1977. The Durbin–Watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica* 45, 1989–96.

How to cite this article

MacKinnon, James G. "Durbin-Watson statistic." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000260> doi:10.1057/9780230226203.0418

The New Palgrave Dictionary of Economics Online

Dutch disease

Thorvaldur Gylfason
From The New Palgrave Dictionary of Economics Online, , 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article outlines the ‘Dutch disease’, the fear of de-industrialization first seen in the Netherlands in the wake of the appreciation of the Dutch guilder following the discovery of natural gas deposits in the North Sea around 1960. It considers its symptoms, and asks whether it is indeed a ‘disease’ with negative economic implications. It also briefly reviews some cases of Dutch disease, and the case of Norway, which appears to have successfully avoided it.

Keywords

de-industrialization; Dutch disease; oil

Article

Dutch disease, in the original sense of the term, refers to the fears of de-industrialization that gripped the Netherlands in the wake of the appreciation of the Dutch guilder following the discovery of natural gas deposits in the North Sea around 1960. The appreciation of the guilder following the gas export boom hurt the profitability of manufacturing and service exports. Total exports from the Netherlands decreased markedly relative to Gross Domestic Product (GDP) during the 1960s. The growth of petroleum exports in the 1960s hurt other exports disproportionately. Many feared dire consequences for Dutch manufacturing. The problem proved short-lived, however. From the late 1960s onward, exports of goods and services have increased from less than 40 per cent of GDP to more than 70 per cent, a high export ratio by world standards. The expected de-industrialization did not materialize, but the name stuck. It can be said that, being neither Dutch nor a disease, the Dutch disease is a double misnomer. But when a disease bears the name of the first patient diagnosed with it, it would seem a bit harsh to require the patient to remain sick for the name to stick.
Is it a disease? Some view it as matter of one sector benefiting at the expense of others, without seeing any macroeconomic or social damage done. Others view the Dutch disease as an ailment, pointing to the potentially harmful consequences of the resulting reallocation of resources – from high-tech, high-skill intensive service industries to low-tech, low-skill intensive primary production, for example – for economic growth and diversification.

Symptoms

An overvalued currency was the first symptom associated with the Dutch disease, but later several other symptoms came to light. Figure 1 illustrates how an oil export boom lifts the equilibrium real exchange rate at which total exports of goods, services, and capital match total imports. In the figure, non-oil exports decline from A to C and hence by less than oil exports increase, so that total exports rise from A to B. For total exports to decline the import schedule would have to shift to the left (for instance through capital inflow) by an amount that exceeds the increase in oil exports, measured by the distance between B and C.
Figure 1
How an oil export boom crowds out nonoil exports.

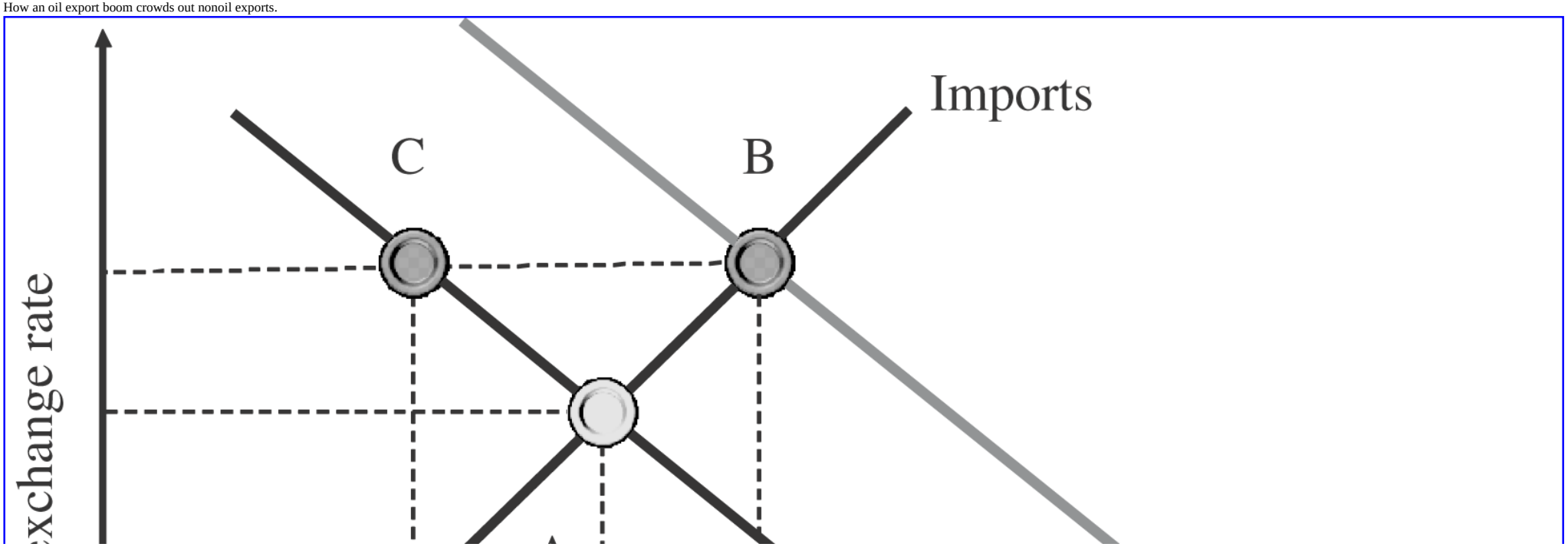

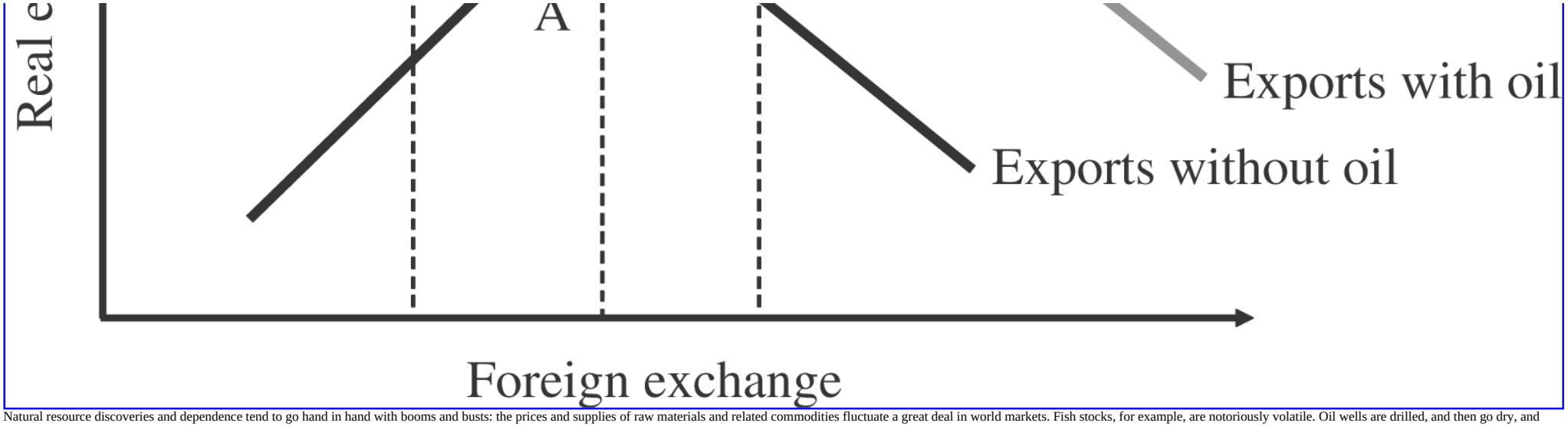

Natural resource discoveries and dependence tend to go hand in hand with booms and busts: the prices and supplies of raw materials and related commodities fluctuate a great deal in world markets. Fish stocks, for example, are notoriously volatile. Oil wells are drilled, and then go dry, and mines are depleted. The resulting fluctuations in export earnings trigger exchange rate volatility, perhaps no less so under fixed exchange rates than under floating rates. Unstable currencies create uncertainty, which tends to hurt exports and imports as well as foreign investment. Further, the Dutch disease can strike even in countries that do not have a national currency of their own (as, for instance, in Greenland, which uses the Danish krone and depends on fish). In this case, the natural-resource-based industry is able to pay higher wages and also higher interest rates than other industries, thus making it difficult for the latter to stay competitive. This problem can become particularly acute in countries with centralized wage bargaining (or with oligopolistic banking systems, for that matter) where the natural-resource-intensive industries set the tone in nation-wide wage negotiations and dictate wage settlements which other industries can ill afford. In one or all of these ways, the Dutch disease tends to reduce the level of total exports or skew the composition of exports away from manufacturing and service exports which could be particularly conducive to economic growth over time. Exports of capital, including inward foreign direct investment, may also suffer.

The Netherlands recovered quickly from the Dutch disease, and has seen a persistent upward trend in its total exports relative to GDP since the mid-1960s. On the other hand, in Norway, the world's third largest oil exporter after Saudi Arabia and Russia, total exports have risen slowly relative to GDP, to a level well below that of the Netherlands (45 per cent in Norway in 2005 compared with 71 per cent in the Netherlands), even if the Dutch economy is almost three times as large as that of Norway. Also, the share of manufactured exports in merchandise exports was 68 per cent in the Netherlands in 2005 compared with 17 per cent in Norway. Exports and manufacturing are good for growth. Openness to trade invigorates imports of goods and services, capital, technology, ideas and know-how. The Dutch disease matters mainly because of its potentially harmful consequences for economic growth.

Channels

Experience seems to suggest six main channels of transmission from heavy natural resource dependence to sluggish economic growth. At the top of the list is the Dutch disease. In second place, huge natural resource rents, especially in conjunction with ill-defined property rights, imperfect or missing markets, and lax legal structures, may lead to rent-seeking behaviour that diverts resources away from more socially fruitful economic activity. The struggle for resource rents may lead to a concentration of economic and political power in the hands of elites which, once in power, use the rent to placate their political supporters and thus secure their hold on power, with stunted or weakened democracy and slow growth as a result. Extensive rent seeking – in other words, seeking to make money from market distortions – can breed corruption, thus reducing both economic efficiency and social equality.

Third, natural resource abundance may imbue people with a false sense of security and lead governments to lose sight of the need for growth-friendly economic management, including free trade, foreign investment, bureaucratic efficiency and good institutions, including democracy. Incentives to create wealth through good policies and institutions may wane because of the relatively effortless ability to extract wealth from the soil or the sea. Fourth, abundant natural resources may likewise weaken incentives to accumulate human capital, even if the rent stream from the resources may enable nations to give a high priority to education. Fifth, natural resource abundance may blunt private and public incentives to save and invest in real capital no less than in human capital, and thereby weaken financial institutions and reduce economic growth. Sixth, natural resource wealth is a fixed factor of production that hampers economic growth potential by causing a growing labour force and a growing stock of capital to run into diminishing returns.

In sum, an abundance of natural capital, if not well managed, may erode or reduce the quality of human, physical, social, financial and foreign capital, and thus stand in the way of rapid economic growth. Manna from heaven can be a mixed blessing. Consider the attitudes of individuals to their own and to other people's money. A person's respect for money tends to vary inversely with his or her distance from the effort expended to make the money. For example, loot tends to be invested with less forethought than honest wages. The same argument applies to unrequited foreign aid. An influx of aid tends to increase the real exchange rate, thereby hurting exports as in Figure 1. Import restrictions exacerbate the appreciation of the currency, hurting exports further. The figure suggests that aid needs to be accompanied by trade liberalization to avoid currency appreciation and its consequences.

Cases

The list of natural-resource-abundant countries beset by economic and political difficulties is a long one. Take Libya. Without its oil export revenues, Libya (population 6 million) would hardly have had the means to purchase 700 military aircraft, submarines and helicopters to pursue the foreign ambitions of Colonel Gaddafi, in power since 1969. In Equatorial Guinea, following oil discoveries, the purchasing power of per capita GDP increased by a factor of six or seven from 1990 to 2005, while life expectancy plunged from 46 years to 42. One child in five dies before reaching its fifth birthday. More than a half of the population of 500.000 lives on less than a dollar a day. President Mbasogo has ruled the country with an iron fist since 1979, usurping the country's oil wealth for himself and his family and cronies. The readiness of the rest of the world to import oil from Equatorial Guinea, and thus to buy stolen goods, is an integral part of the problem because a people's right to its natural resources is a human right proclaimed in primary documents of international law and enshrined in many national constitutions. Article 1 of the International Covenant on Civil and Political Rights states that ‘All people may, for their own ends, freely dispose of their natural wealth and resources.’ Neither Libya nor Equatorial Guinea exports any manufactures to speak of.

The list of countries afflicted by various symptoms of the Dutch disease could be extended to include Iran, Iraq, Mexico, Nigeria, Russia, Saudi Arabia, Sudan and Venezuela, among several others. Some other countries have managed to avoid such afflictions. A prime example is Norway, where, before the first drop of oil emerged, the oil and gas reserves within Norwegian jurisdiction were defined by law as common property resources, thereby clearly establishing the legal rights of the Norwegian people to the resource rents. On this legal basis, the government has absorbed about 80 per cent of the resource rent over the years, having learnt the hard way in the 1970s to use a relatively small portion of the total to meet current fiscal needs. Most oil revenue is set aside in the state petroleum fund, recently renamed the pension fund to reflect its intended use. The government laid down economic as well as ethical principles (commandments) to guide the use and exploitation of the oil and gas for the benefit of current and future generations of Norwegians. The main political parties share an understanding that the national economy needs to be shielded from an excessive influx of oil money to avoid overheating and waste. The Central Bank (Norges Bank), which was granted increased independence from the government in 2001, manages the fund (currently around US$400 billion or $85,000 per Norwegian) on behalf of the Ministry of Finance. This arrangement maintains a distance between politicians and the fund. Almost 40 years after discovering their oil, the Norwegians have a smaller central government than Denmark, Finland and Sweden next door.

Norway's tradition of democracy since long before the advent of oil has probably helped immunize the country from the ailments that afflict most other oil-rich nations. Large-scale rent seeking has been averted in Norway, investment performance has been adequate, and the country's education record is excellent. Even so, some (weak) signs of the Dutch disease can be detected, notably sluggish exports and foreign direct investment and the absence of a large, vibrant high-tech manufacturing sector as in Sweden and Finland. Norway's lack of interest in joining the European Unions can also be viewed in this light.

Then there is Botswana. Having managed its diamonds quite well and used the rents to support rapid growth, Botswana has become the richest country in Africa, measured by the purchasing power of per capita GDP. Its rapid growth since 1965 has been accompanied by political stability and a steady advance of democracy. Unlike Sierra Leone's alluvial diamonds, which are easy to mine by shovel and pan, and easy to loot, Botswana's kimberlite diamonds lie deep in the ground and can only be mined with large hydraulic shovels and other sophisticated equipment. They are therefore not very lootable. This difference probably helped Botswana succeed while Sierra Leone failed, and so, most likely, did South African involvement in Botswana's diamond industry.

See Also

- de-industrialization, ‘premature’ de-industrialization and the Dutch disease
- oil and the macroeconomy

Bibliography

Corden, W.M. and Neary, J.P. 1982. Booming sector and de-industrialisation in a small open economy. *Economic Journal* 92, 825–48.

Corden, W.M. 1984. Booming sector and Dutch disease economics: survey and consolidation. *Oxford Economic Papers* 36, 359–80.

Ross, M. 2001. Does oil hinder democracy? *World Politics* 53, 325–61.

Van Wijnbergen, S. 1984. The ‘Dutch disease’: a disease after all? *Economic Journal* 94, 41–5.

Wenar, L. 2008. Property rights and the resource curse. *Philosophy and Public Affairs* 36, 1–32.

How to cite this article

Gylfason, Thorvaldur. "Dutch disease." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 14 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_D000264> doi:10.1057/9780230226203.1881

Dutch Disease

dutch disease, in the original sense of the term, refers to the fears of de-industrialization that gripped the Netherlands in the wake of the appreciation of the Dutch guilder following the discovery of natural gas deposits in the North Sea around 1960. The appreciation of the guilder following the gas export boom hurt the profitability of manufacturing and service exports. Total exports from the Netherlands decreased markedly relative to Gross Domestic Product (GDP) during the 1960s. The growth of petroleum exports in the 1960s hurt other exports disproportionately. Many feared dire consequences for Dutch manufacturing. The problem proved short-lived, however. From the late 1960s onward, exports of goods and services have increased from less than 40 per cent of GDP to more than 70 per cent, a high export ratio by world standards. The expected de-industrialization did not materialize, but the name stuck. It can be said that, being neither Dutch nor a disease, the Dutch disease is a double misnomer. But when a disease bears the name of the first patient diagnosed with it, it would seem a bit harsh to require the patient to remain sick for the name to stick.

Is it a disease? Some view it as matter of one sector's benefiting at the expense of others, without seeing any macroeconomic or social damage done. Others view the Dutch disease as an ailment, pointing to the potentially harmful consequences of the resulting reallocation of resources – from high-tech, high-skill intensive service industries to low-tech, low-skill intensive primary production, for example – for economic growth and diversification.

Symptoms

An overvalued currency was the first symptom associated with the Dutch disease, but later several other symptoms came to light. Figure 1 illustrates how an oil export boom lifts the equilibrium real exchange rate at which total exports of goods, services, and capital match total imports. In the figure, nonoil exports decline from A to C and hence by less than oil exports increase, so that total exports rise from A to B. For total exports to decline the import schedule would have to shift to the left (e.g., through capital inflow) by an amount that exceeds the increase in oil exports, measured by the distance between B and C.

<Figure 1>

Natural resource discoveries and dependence tend to go hand in hand with booms and busts: the prices and supplies of raw materials and related commodities fluctuate a great deal in world markets. Fish stocks, for example, are notoriously volatile. Oil wells are drilled, and then go dry, and mines are depleted. The resulting fluctuations in export earnings trigger exchange rate volatility, perhaps no less so under fixed exchange rates than under floating rates. Unstable currencies create uncertainty that tends to hurt exports and imports as well as foreign investment. Further, the Dutch disease can strike even in countries that do not have a national currency of their own (as, for instance, in Greenland that uses the Danish krone and depends on fish). In this case, the natural-resource-based industry is able to pay higher wages and also higher interest rates than other industries, thus making it difficult for the latter to stay competitive. This problem can become particularly acute in countries with centralized wage bargaining (or with oligopolistic banking systems, for that matter) where the natural-resource-intensive industries set the tone in nation-wide wage negotiations and dictate wage settlements that other industries can ill afford. In one or all of these ways, the Dutch disease tends to reduce the level of total exports or skew the composition of exports away from manufacturing and service exports that may be particularly conducive to economic growth over time. Exports of capital, including inward foreign direct investment, may also suffer.

The Netherlands recovered quickly from the Dutch disease, and have seen a persistent

upward trend in their total exports relative to GDP since the mid-1960s. On the other hand, in Norway, the world's third largest oil exporter after Saudi Arabia and Russia, total exports have risen slowly relative to GDP to a level well below that of the Netherlands (45 per cent in Norway in 2005 compared with 71 per cent in the Netherlands), even if the Dutch economy is almost three times as large as that of Norway. Also, the share of manufactured exports in merchandise exports was 68 per cent in the Netherlands in 2005 compared with 17 per cent in Norway. Exports and manufacturing are good for growth. Openness to trade invigorates imports of goods and services, capital, technology, ideas, and know-how. The Dutch disease matters mainly because of its potentially harmful consequences for economic growth.

Channels

Experience seems to suggest six main channels of transmission from heavy natural resource dependence to sluggish economic growth. At the top of the list is the Dutch disease. In second place, huge natural resource rents, especially in conjunction with ill-defined property rights, imperfect or missing markets, and lax legal structures, may lead to rent-seeking behavior that diverts resources away from more socially fruitful economic activity. The struggle for resource rents may lead to a concentration of economic and political power in the hands of elites that, once in power, use the rent to placate their political supporters and thus secure their hold on power, with stunted or weakened democracy and slow growth as a result. Extensive rent seeking – i.e., seeking to make money from market distortions – can breed corruption, thus reducing both economic efficiency and social equality. Third, natural resource abundance may imbue people with a false sense of security and lead governments to lose sight of the need for growth-friendly economic management, including free trade, foreign investment, bureaucratic efficiency, and good institutions, including democracy. Incentives to create wealth through good policies and institutions may wane because of the relatively effortless ability to extract wealth from the soil or the sea. Fourth, abundant natural resources may likewise weaken incentives to accumulate human capital, even if the rent stream from the resources may enable nations to give a high priority to education. Fifth, natural resource abundance may blunt private and public incentives to save and invest in real capital no less than in human capital and thereby weaken financial institutions and reduce economic growth. Sixth, natural resource wealth is a fixed factor of production that hampers economic growth potential by causing a growing labor force and a growing stock of capital to run into diminishing returns.

In sum, an abundance of natural capital, if not well managed, may erode or reduce the quality of human, physical, social, financial, and foreign capital, and thus stand in the way of rapid economic growth. Manna from heaven can be a mixed blessing. Consider the attitudes of individuals to their own and to other people's money. A person's respect for money tends to vary inversely with his or her distance from the effort expended to make the money. For example, loot tends to be invested with less forethought than honest wages. The same argument applies to unrequited foreign aid. An influx of aid tends to increase the real exchange rate, thereby hurting exports as in Figure 1. Import restrictions exacerbate the appreciation of the currency, hurting exports further. The figure suggests that aid needs to be accompanied by trade liberalization to avert the currency appreciation and its consequences.

Cases

The list of natural-resource-abundant countries beset by economic and political difficulties is a long one. Take Libya. Without its oil export revenues, Libya (population six million) would hardly have had the means to purchase 700 military aircraft, submarines, and helicopters to

pursue the foreign ambitions of Colonel Gaddafi, in power since 1969. In Equatorial Guinea, following oil discoveries, the purchasing power of per capita GDP increased by a factor of six or seven from 1990 to 2005 while life expectancy plunged from 46 years to 42. One child in five dies before reaching its fifth birthday. More than a half of the population of 500.000 lives on less than a dollar a day. President Mbasogo has ruled the country with an iron fist since 1979, usurping the country's oil wealth for himself and his family and cronies. The readiness of the rest of the world to import oil from Equatorial Guinea, and thus to buy stolen goods, is an integral part of the problem because a people's right to its natural resources is a human right proclaimed in primary documents of international law and enshrined in many national constitutions. Article 1 of the International Covenant on Civil and Political Rights states that "All people may, for their own ends, freely dispose of their natural wealth and resources." Neither Libya nor Equatorial Guinea exports any manufactures to speak of.

The list of countries afflicted by various symptoms of the Dutch disease could be extended to include Iran, Iraq, Mexico, Nigeria, Russia, Saudi Arabia, Sudan, and Venezuela, among several others. Some other countries have managed to avoid such afflictions. A prime example is Norway, where, before the first drop of oil emerged, the oil and gas reserves within Norwegian jurisdiction were defined by law as common property resources, thereby clearly establishing the legal rights of the Norwegian people to the resource rents. On this legal basis, the government has absorbed about 80 per cent of the resource rent over the years, having learnt the hard way in the 1970s to use a relatively small portion of the total to meet current fiscal needs. Most oil revenue is set aside in the state petroleum fund, recently renamed the pension fund to reflect its intended use. The government laid down economic as well as ethical principles (commandments) to guide the use and exploitation of the oil and gas for the benefit of current and future generations of Norwegians. The main political parties share an understanding that the national economy needs to be shielded from an excessive influx of oil money to avoid overheating and waste. The Central Bank (Norges Bank), which was granted increased independence from the government in 2001, manages the fund (currently around $400 billion or $85,000 per Norwegian) on behalf of the Ministry of Finance. This arrangement maintains a distance between politicians and the fund. Almost 40 years after discovering their oil, the Norwegians have a smaller central government than Denmark, Finland, and Sweden next door.

Norway's tradition of democracy since long before the advent of oil has probably helped immunize the country from the ailments that afflict most other oil-rich nations. Large-scale rent seeking has been averted in Norway, investment performance has been adequate, and the country's education record is excellent. Even so, some (weak) signs of the Dutch disease can be detected, notably sluggish exports and foreign direct investment and the absence of a large, vibrant high-tech manufacturing industry as in Sweden and Finland. Norway's lack of interest in joining the European Union can also be viewed in this light.

And then there is Botswana. Having managed its diamonds quite well and used the rents to support rapid growth, Botswana has become the richest country in Africa measured by the purchasing power of per capita GDP. Its rapid growth since 1965 has been accompanied by political stability and a steady advance of democracy. Unlike Sierra Leone's alluvial diamonds that are easy to mine by shovel and pan and easy to loot, Botswana's kimberlite diamonds lie deep in the ground and can only be mined with large hydraulic shovels and other sophisticated equipment and, therefore, are not very lootable. This difference probably helped Botswana succeed while Sierra Leone failed, and so, most likely, did South African involvement in Botswana's diamond industry.

References

Corden, W.M., and J.P. Neary 1982. Booming sector and de-industrialisation in a small open economy. *Economic Journal* 92, 825-848.
Corden, W.M. 1984. Booming sector and Dutch disease economics: Survey and consolidation. *Oxford Economic Papers* 36, 359-380.
Ross, M. 2001. Does oil hinder democracy? *World Politics* 53, 325-361.
Van Wijnbergen, S. 1984. The 'Dutch disease': A disease after all? *Economic Journal* 94, 41-45.
Wenar, L. 2008. Property rights and the resource curse. *Philosophy and Public Affairs* 36, 1-32.

Thorvaldur Gylfason.

Figure 1. How an oil export boom crowds out nonoil exports

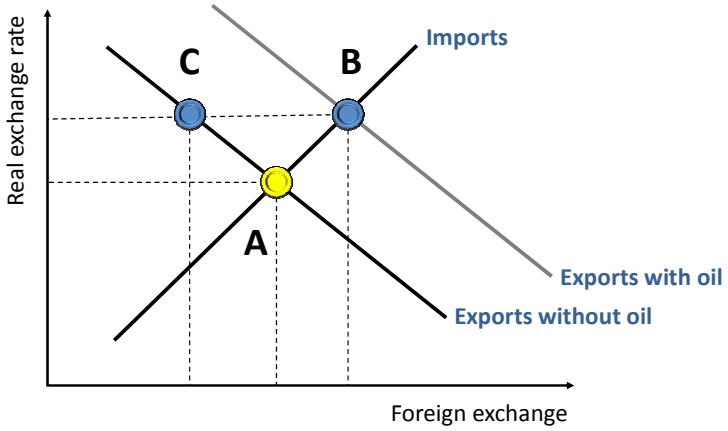

The New Palgrave Dictionary of Economics Online

dynamic models with non-clearing markets

Jean-Pascal Bénassy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article studies a new class of models which synthesize the two traditions of general equilibrium with non-clearing markets and imperfect competition on the one hand, and dynamic stochastic general equilibrium (DSGE) models on the other hand. This line of models has become a central paradigm of modern macroeconomics for at least three reasons: (*a*) it displays solid microeconomic foundations, (*b*) it is a highly synthetic theory, which combines in a unified framework general equilibrium, non-clearing markets, imperfect competition, growth theory and rational expectations, and (*c*) it is also an empirical success, leading to substantial progress towards matching real world statistics.

Keywords

budget constraints; dynamic models with non-clearing markets; dynamic stochastic general equilibrium (DSGE) models; efficiency wages; fixprice macroeconomic model; general equilibrium; growth theory; implicit contracts; IS–LM model; Keynesianism; market clearing; menu costs; microfoundations; monetary shocks; monopolistic competition; nominal rigidities; non-clearing markets in general equilibrium; objective demand curve; Phillips curve; quantity constraint; quantity signals; rational expectations; real business cycles; state dependent price rigidities; sticky prices; technological shocks; time-dependent contracts; Walrasian equilibrium

Article

This article studies a new class of models which synthesize the two traditions of general equilibrium with non-clearing markets and imperfect competition on the one hand, and dynamic stochastic general equilibrium (DSGE) models on the other hand. Although this line of models is still recent, it has clearly become in a short time a central paradigm of modern macroeconomics. The reasons are at least threefold.

The first is that it displays solid microeconomic foundations. This is quite natural since from the two constituent fields above this one inherited a strong general equilibrium framework where all agents (households or firms) maximize their respective objectives subject to well defined constraints.

The second is that it is a highly synthetic theory, which combines in a unified framework general equilibrium, non-clearing markets, imperfect competition, growth theory and rational expectations, so that it can appeal to macroeconomists with very different backgrounds.

The third reason is empirical. A key motivation for DSGE models is to compare the ‘statistics’ generated by these models with the real-world ones. In that respect the addition of non-clearing markets and imperfect competition has led to substantial progress in matching these statistics, and this has certainly been an important factor in the success of these models.

Now such a wide synthesis did not come all at once. So we begin by recalling briefly a little bit of history and some of the antecedents of the field.

We then present a series of models with explicit solutions. These will demonstrate analytically how the introduction of non-clearing markets allows us to substantially improve the ability of DSGE to reproduce a number of macroeconomic facts.

History

Early times

At the time when many of the developments leading to these models were initiated, there was a profound split between microeconomics and macroeconomics. On the one hand microeconomics, in its general equilibrium version, was dominated by Walras's (1874) model, as developed by Arrow and Debreu (1954), Arrow (1963), and Debreu (1959). In these models all adjustments are carried out via fully flexible prices, and agents never experience any quantity constraint. On the other hand in the standard macroeconomic model in the Keynes (1936) and Hicks (1937) tradition, as exemplified by the IS–LM model, there are price and wage rigidities, unemployment is present and most adjustments are carried out through variations in real income, a quantity, not a price.

Confronted with this inconsistency, the strategies of macroeconomists turned out to be quite diverse and they took two different routes.

General equilibrium with non-clearing markets

On the one hand, a first set of authors aimed at achieving a synthesis between the then existing microeconomics and macroeconomics. This was achieved by generalizing the traditional general equilibrium model, by introducing non-clearing markets, introducing quantity signals into demand and supply functions, and endogenizing prices in a framework of imperfect competition.

Patinkin (1956) and Clower (1965) showed that the presence of quantity constraints in non-clearing markets would drastically modify the demands for labour and goods, an insight further emphasized by Leijonhufvud (1968). Barro and Grossman (1971; 1976) combined these insights into a fixprice macromodel. Drèze (1975) and Bénassy (1975; 1982) constructed full general equilibrium concepts with price rigidities, where price movements are partially replaced by endogenous quantity constraints. Bénassy (1976) linked these concepts with general equilibrium under imperfect competition à la Negishi (1961). This link was furthered with the construction of a full general equilibrium concept of objective demand curve based on quantity constraints (Bénassy, 1988; see also Gabszewicz and Vial, 1972, for a Cournotian view). All these developments are reviewed in the dictionary entry ‘non clearing markets in general equilibrium’.

Dynamic market clearing macroeconomics

A second set of authors achieved consistency between microeconomics and macroeconomics by importing into macroeconomics the basic assumption of the then dominant general equilibrium microeconomic models, market clearing. At the same time they paid strong attention to the issues of dynamics and expectations. A central part of these developments was the use of ‘rational expectations’ in the sense of Muth (1961). This was an important addition, as in the Keynesian system it was sometimes difficult to disentangle the results due to price or wage rigidity from those due to incorrect expectations. Rational expectations allowed the suppression of the second type of results. It appeared also that, even with rational expectations and market clearing, it was possible to build rigorous models displaying fluctuations (Lucas, 1972; Kydland and Prescott, 1982; Long and Plosser, 1983).

Non-Walrasian cycles

Starting in the mid-1980s authors began combining elements of the two paradigms described above, achieving the synthesis that is the subject of this article. Svensson (1986) studies a dynamic stochastic general equilibrium monetary economy subject to supply and demand shocks. Prices are preset one period in advance by monopolistically competitive firms, so we have both imperfect competition and sticky prices. Because of price presetting the model has multiple regimes.

Various types of rigidities have been then introduced in dynamic models, leading to different patterns of cycles. Andersen (1994) reviews various causes and consequences of price and wage rigidities.

A first type of rigidities is ‘real’ rigidities, which create an endogenous non-competitive wedge between various prices. As an example, monopolistic competition à la Dixit and Stiglitz (1977) introduces a markup between marginal cost and price. In this class Danthine and Donaldson (1990) introduce efficiency wages, Danthine and Donaldson (1991; 1992) introduce implicit contracts in the vein of Azariadis (1975), Baily (1974) and Gordon (1974). Rotemberg and Woodford (1992; 1995) study imperfect competition.

Models with nominal rigidities study situations where the nominal prices themselves (and not relative prices) are sluggish. Several devices have been used. The first, following the early works on wage and price contracts by Gray (1976), Fischer (1977), Phelps and Taylor (1977), Taylor (1979; 1980) and Calvo (1983), assumes that there is a system of contracts expiring at deterministic or stochastic dates. For that reason they are called ‘time dependent’. Such contracts have been integrated in DSGE models by Cho (1993), Cho and Cooley (1995), Bénassy (1995; 2002; 2003a; 2003b), Yun (1996), Cho, Cooley and Phaneuf (1997), Andersen (1998), Jeanne (1998), Ascari (2000), Chari, Kehoe and McGrattan (2000), Collard and Ertz (2000), Ascari and Rankin (2002), Huang and Liu (2002), Smets and Wouters (2003) and Christiano, Eichenbaum and Evans (2005), to name only a few.

Another type of price rigidity, called ‘state dependent’, is based on costs of changing prices. Two specifications are favourite in the literature: quadratic costs of changing prices

(Rotemberg, 1982a; 1982b), which have been implemented, for example, in Hairault and Portier (1993), and fixed costs of changing prices (Barro, 1972), often renamed 'menu costs'. Clearly these costs should be interpreted as surrogates for other unspecified causes, and identifying these causes is a challenge that faces this line of research.
Now most of the contributions of this field are based on numerical evaluations of various models. So we present next a number of models with explicit solutions which will make clear why this line of models has been successful in solving problems that were difficult to solve in market-clearing models.

An analytical illustration

We shall now show in this section in a series of explicitly solved models how the introduction of nominal rigidities in DSGE models allows to considerably improve the capacities of these models to reproduce the dynamic evolutions of actual economies.
We first present a basic model and compute as a reference its Walrasian equilibrium and dynamics. Then we introduce a first nominal rigidity, one-period wage contracts. This improves some correlations, but cannot create strong persistence as in reality. We next introduce multi-periodic wage contracts, and show that this allows us to obtain a persistent response of output to demand shocks. Finally, simultaneous rigidities of wages and prices are considered, and we show that one can obtain in this way with fairly realistic values of the parameters a persistent and hump-shaped response of both output and inflation.

The basic model

We study a dynamic monetary economy à la Sidrauski (1967) and Brock (1975), where goods are exchanged against money at the (average) price P_t and work against money at the (average) wage W_t. There are two types of agents: households and firms. Firms have a simple technology:

$$Y_t = Z_t N_t^{\alpha} \tag{1}$$

where N_t is the quantity of labour used by firms and Z_t a technological shock common to all firms. Note that we do not introduce capital in this model. Because its rate of depreciation is low, it would not add much to our argument, and would substantially complicate the results and exposition.
The representative household works N_t, consumes C_t, and ends period t with a quantity of money M_t. It maximizes the expectation of its discounted utility:

$$U = E_0 \sum_{t=0}^{\infty} \beta^t \left[\log C_t + \omega \log \frac{M_t}{P_t} - \xi \frac{N_t^{\nu}}{\nu} \right]. \tag{2}$$

At the beginning of period t the household faces a monetary shock à la Lucas (1972), whereby the quantity of money M_{t-1} coming from $t-1$ is multiplied by μ_t, so that its budget constraint for period t is:

$$C_t + \frac{M_t}{P_t} = \frac{W_t}{P_t} N_t + \frac{\mu_t M_{t-1}}{P_t}. \tag{3}$$

There are thus two shocks in this economy, the technology shock Z_t and the monetary shock $\mu_t = M_t / M_{t-1}$. As an illustration we shall use below the following traditional processes (in all that follows lower-case letters represent the logarithm of the variable represented by the corresponding uppercase letter):

$$m_t - m_{t-1} = \frac{\varepsilon_{mt}}{1 - \rho L} z_t = \frac{\varepsilon_{zt}}{1 - \phi L}$$
(4)

where ε_{zt} and ε_{mt}, the innovations in z_t and m_t, are uncorrelated white noises with:

$$\mathrm{var}(\varepsilon_{zt}) = \sigma_z^2 \mathrm{var}(\varepsilon_{mt}) = \sigma_m^2.$$
(5)

Walrasian dynamics

As a benchmark we shall study here the case where both labour and goods markets are in Walrasian equilibrium in each period, as in the first traditional real business cycle (RBC) models, and we shall see how this economy reacts to technological and monetary shocks. Solving the model we find that money holdings are a multiple of consumption:

$$\frac{M_t}{P_t C_t} = \frac{\omega}{1 - \beta}$$
(6)

and that employment N_t is constant:

$$N_t = N = (\alpha / \xi)^{1/\nu}.$$
(7)

Using (1) and (7) we find (we eliminate some irrelevant constant terms):

$$n_t = n y_t = z_t + \alpha n w_t - p_t = y_t - n.$$
(8)

Although we will not do any real calibration in this article, we can note at this stage a few issues that posed a problem to researchers in the RBC domain.
First, real wages are much too pro-cyclical in this Walrasian model. From (8) we see that the real wage–output correlation is equal to 1. Even though this correlation is lower than 1 in calibrated models where N_t varies, it is always quite above what is observed in real economies.
A second problem concerns the inflation–output correlation, a problem related to the literature on the Phillips curve. Whereas it is generally considered that this correlation is positive, the above Walrasian model yields a negative correlation:

$$\mathrm{cov}(\Delta p_t, y_t) = -\frac{\sigma_z^2}{1+\phi} < 0. \tag{9}$$

Finally, an important and recurrent critique of RBC-type models has been that they do not generate any internal propagation mechanism, and that the only persistence in output movements is that already present in the exogenous process of technological shocks z_t (see, for example, Cogley and Nason, 1993; 1995). This appears here in eq. (8), where the dynamics of output y_t is exactly the same as that of the technological shock z_t.
We shall now introduce wage contracts, first lasting one period, and then multi-period overlapping contracts, and we shall see that the above problems find a natural solution in this framework.

Single-period wage contracts

Let us thus assume (Bénassy, 1995, and Bénassy, 2002, for microfoundations), that the wages are predetermined at the beginning of each period at the expected value of the Walrasian wage (in logarithms), and that at this contractual wage the households supply the quantity of work demanded by firms (this type of contract was introduced by Gray, 1976).
Combining (6) and $C_t = Y_t$ we find that the Walrasian wage w_t^* is, up to an unimportant constant, equal to m_t, so that the preset wage w_t is given by:

$$w_t = E_{t-1} w_t^* = E_{t-1} m_t \tag{10}$$

where $E_{t-1} m_t$ is the expectation of m_t formed at the beginning of period t, before shocks are known.
The difference with the Walrasian case is that employment N_t is now variable and demand determined. Equations (8) become:

$$y_t = z_t + \alpha n_t w_t - p_t = y_t - n_t \tag{11}$$

while $n_t = n$ is replaced by (10). So we first obtain the level of employment in period t:

$$n_t = n + m_t - E_{t-1} m_t = n + \varepsilon_{mt}$$

(12)

since $m_t - E_{t-1}m_t = \varepsilon_{mt}$. Contrarily to what happened in the Walrasian version of the model, unanticipated monetary shocks now have an impact on the level of employment, and therefore output. We shall now use the preceding formulas to show that the hypothesis of preset wages allows to substantially improve some correlations relative to the Walrasian model.

Let us start with the real wage which, in the Walrasian model, has a much too high positive correlation with output. Let us combine (11) and (12), to obtain the values of output and real wage:

$$y_t = z_t + \alpha\varepsilon_{mt} w_t - p_t = z_t - (1-\alpha)\varepsilon_{mt}.$$

(13)

We see that supply shocks create a positive correlation between the real wage and output. However, monetary shocks create a negative correlation. Our model thus allows us to combine this last characteristic, typical of traditional Keynesians models, with the usual results of RBC models. If one considers the technological and monetary shocks (4), one obtains the following correlation:

$$\mathrm{corr}(w_t - p_t, y_t) = \frac{\sigma_z^2 - (1-\phi^2)\alpha(1-\alpha)\sigma_m^2}{\left[\left(\sigma_z^2 + (1-\phi^2)\alpha^2\sigma_m^2\right)\right]^{1/2}\left[\sigma_z^2 + (1-\phi^2)(1-\alpha)^2\sigma_m^2\right]^{1/2}}.$$

(14)

We see that the real-wage–output correlation is equal to 1 if there are *only* technological shocks. But this correlation diminishes as soon as there are monetary shocks, and it can even become negative. One can thus reproduce the correlations observed in reality by adequate combinations of technological and monetary shocks.

Let us now consider the relation between inflation and output, which are generally considered to be positively correlated, at least in Keynesian tradition. If we assume again the monetary and technological shocks (4), we find:

$$\mathrm{Covariance}\,(\Delta p_t, y_t) = \alpha(1-\alpha)\,\sigma_m^2 - \frac{\sigma_z^2}{1+\phi}.$$

(15)

Formula (15) shows us that the positive covariance (and thus correlation) between inflation and output is linked to the presence of demand shocks, and that the sign of this correlation may change if there are sufficiently strong technological shocks.

So we just saw that one-period contracts allow us to improve some important correlations. We now naturally ask a question already posed for the standard RBC model: is the response to shocks, and in particular to demand shocks, sufficiently persistent? Let us recall eq. (13):

$$y_t = z_t + \alpha\varepsilon_{mt}.$$

(16)

We see that monetary shocks now have an immediate effect on output (and employment), but that, starting with the second period, the effect of these shocks is completely dampened. One-period contracts allow us to solve the puzzle raised by some correlations, but certainly not the persistence problem. We shall see in the next two sections that multi-periodic contracts allow us to solve that problem.

Multi-periodic wage contracts

The models that we have examined so far share with traditional RBC models the defect of having an extremely weak internal propagation mechanism. In particular, the response of output to monetary demand shocks is almost entirely transitory. But several empirical studies (see, for example, Christiano, Eichenbaum and Evans, 1999; 2005) have pointed out that in reality the response to monetary shocks not only was persistent but also had a hump-shaped response function. We shall now introduce multi-periodic wage contracts in rigorous stochastic dynamic models, and show that they allow us to reproduce these features. Models with such multi-periodic wage or price contracts have been studied notably by Yun (1996), Andersen (1998), Jeanne (1998), Ascari (2000), Chari, Kehoe and McGrattan (2000), Collard and Ertz (2000), and Bénassy (2002; 2003a; 2003b).

In order to make our demonstration analytically, we use a contract, inspired by Calvo (1983) and developed in Bénassy (2002; 2003a), which has three advantages: (*a*) the average duration of contracts can take any value from zero to infinity, (*b*) an analytical solution can be found with both wage and price contracts, and (*c*) it has explicit microfoundations.

In this framework in each period *s* a contract is made for wages at period $t \geq s$. As in the Gray contract, the contract wage is the expectation of the market-clearing wage in period *t*. So if we denote as x_{st} the contract wage made in *s* for period *t*:

$$x_{st} = E_s(w_t^*).$$

(17)

Now, as in Calvo (1983), each wage contract has a probability γ to stay unchanged, and a probability $1-\gamma$ to be broken. If the contract is broken, a new contract is immediately renegotiated on the basis of current period information. So for $\gamma = 0$, wages are totally flexible, for $\gamma = 1$ they are totally rigid.

It is easy to compute the average duration of these contracts. The probability for a contract to be still valid *j* periods after the date it was concluded is equal to $(1-\gamma)\gamma^j$. The expected duration of the contract is thus:

$$\sum_{j=0}^{\infty}(1-\gamma)\,j\gamma^j = \frac{\gamma}{1-\gamma}.$$

(18)

We thus see that varying γ from 0 to 1 the average duration of the contract varies from zero to infinity.

The average wage w_t is the mean of past x_{st}'s weighted by the probability for the corresponding contract to be still in effect. Because of the law of large numbers, and since the probability of survival of wage contracts is γ, the proportion of contracts coming from period $s \leq t$ is $(1-\gamma)\gamma^{t-s}$. Therefore, the average wage in the economy is given by:

$$w_t = (1-\gamma)\sum_{s=-\infty}^{t}\gamma^{t-s}x_{st}.$$
(19)

If we now solve the model with the shocks (4) we find that the dynamics of employment is characterized by (Bénassy, 2002; 2003a):

$$n_t = n + \frac{\gamma\varepsilon_{mt}}{(1-\gamma L)(1-\gamma\rho L)}$$
(20)

where *L* is the lag operator: $L^j x_t = x_{t-j}$. The response of output is deduced from that of employment through:

$$y_t = \alpha n_t + z_t.$$
(21)

Formula (20) shows clearly that, contrarily to the case of one-period contracts, the response to a monetary shock can be quite persistent. We can have an idea of the temporal profile of this response by computing the response function of output and employment to a monetary shock. The value of ρ most often found in the literature is $\rho = 0.5$. As for γ, we saw above (formula 18) that the average duration of wage contracts is equal to $\gamma/(1-\gamma)$. One considers generally that the average duration of wage contracts is about one year (see, for example, Taylor, 1999), which corresponds to $\gamma = 4/5$. Figure 1 shows the response of employment (output is derived via 21) to a monetary shock for $\gamma = 4/5$.

Figure 1

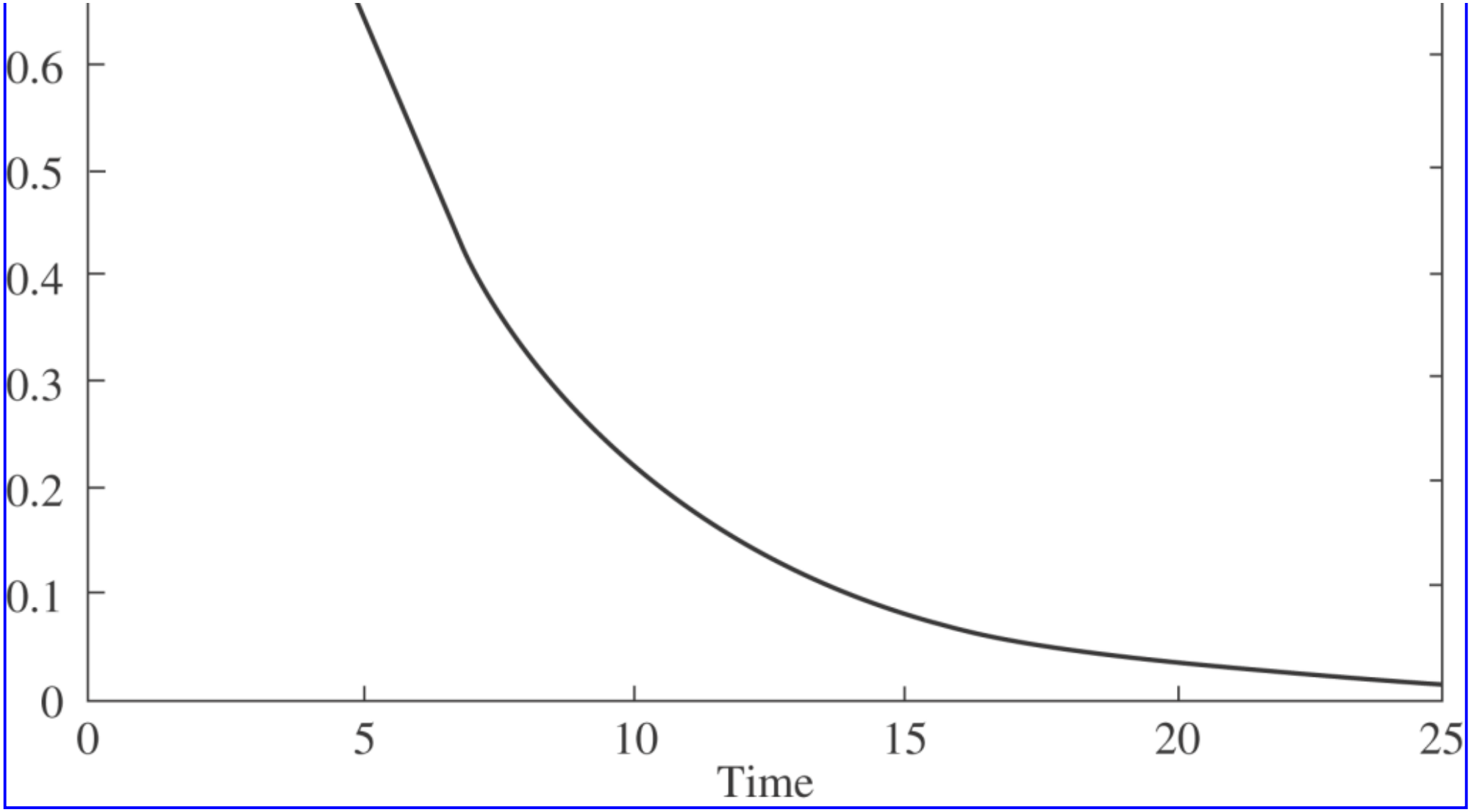

We see that the response function displays persistence in the effects of monetary shocks, and has even a hump-shaped response. If we plot, however, the response function of inflation, we find that it is steadily decreasing after the initial jump, whereas it seems to have a delayed hump-shaped response in reality.

Wage and price multi-periodic contracts

We shall now enlarge our model by considering simultaneously wage and price multi-periodic contracts (see Bénassy, 2003b, for such a model with explicit microfoundations). Numerically solved models with both wage and price multi-periodic contracts are found in Christiano, Eichenbaum and Evans (2005), Huang and Liu (2002), Smets and Wouters (2003).

Wage contracts are exactly the same as in the preceding section: each contract is maintained with probability γ, or renegotiated with probability $1-\gamma$. Symmetrically, price contracts are maintained with probability ϕ, or break down and are renegotiated with probability $1-\varphi$. The average price p_t is given by:

$$p_t = (1 - \varphi) \sum_{s=-\infty}^{t} \varphi^{t-s} q_{st}$$
(22)

where q_{st} is the price contract negotiated in period s for period t. Using again the shock processes (4), and taking $v = 1$, we find the following dynamics for output and inflation:

$$y_t = z_t - \frac{\varphi \varepsilon_{zt}}{1 - \varphi \phi L} + \frac{\alpha \gamma \varepsilon_{rrt}}{(1 - \gamma L)(1 - \gamma \rho L)} + \frac{\varphi \varepsilon_{rrt}}{(1 - \varphi L)(1 - \varphi \rho L)} - \frac{\alpha \gamma \varphi \varepsilon_{rrt}}{(1 - \gamma \varphi L)(1 - \gamma \varphi \rho L)}$$
(23)

$$\pi_t = (1 - L) p_t = (1 - L)(m_t - y_t).$$
(24)

As in the preceding section we take as an illustration $\alpha = 2/3$, $\rho = 1/2$ and $\gamma = 4/5$ (one-year wage contracts). As for prices, we want to take a rather low duration of contracts, so we shall take $\varphi = 1/2$ (one quarter). Simulations show that in that case we obtain a persistent and hump-shaped response for both output and inflation.
So we see that with only reasonable nominal rigidities we obtain some realistic response functions. Clearly the adjunction of ‘real’ rigidities would allow to reproduce even better the actual dynamic macroeconomic patterns.

See Also

- non-clearing markets in general equilibrium
- real business cycles

Bibliography

Andersen, T. 1994. *Price Rigidity: Causes and Macroeconomic Implications*. Oxford: Oxford University Press.

Andersen, T. 1998. Persistency in sticky price models. *European Economic Review* 42, 593–603.

Arrow, K. 1963. The role of securities in the optimal allocation of risk–bearing. *Review of Economic Studies* 31, 91–6.

Arrow, K. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Ascari, G. 2000. Optimising agents, staggered wages and persistence in the real effects of money shocks. *Economic Journal* 110, 664–86.

Ascari, G. and Rankin, N. 2002. Staggered wages and output dynamics under disinflation. *Journal of Economic Dynamics and Control* 26, 653–80.

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83, 1183–202.

Baily, M. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41, 37–50.

Barro, R. 1972. A theory of monopolistic price adjustment. *Review of Economic Studies* 39, 17–26.

Barro, R. and Grossman, H. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

Barro, R. and Grossman, H. 1976. *Money, Employment and Inflation*. Cambridge: Cambridge University Press.

Bénassy, J.-P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42, 503–23.

Bénassy, J.-P. 1976. The disequilibrium approach to monopolistic price setting and general monopolistic equilibrium. *Review of Economic Studies* 43, 69–81.

Bénassy, J.-P. 1977. A neoKeynesian model of price and quantity determination in disequilibrium. In *Equilibrium and Disequilibrium in Economic Theory*, ed. G. Schwödiauer. Boston: Reidel Publishing Company.

Bénassy, J.-P. 1982. *The Economics of Market Disequilibrium*. New York: Academic Press.

Bénassy, J.-P. 1988. The objective demand curve in general equilibrium with price makers. *Economic Journal* 98(supplement), 37–49.

Bénassy, J.-P. 1990. Non-Walrasian equilibria, money and macroeconomics. In *Handbook of Monetary Economics*, ed. B. Friedman and F. Hahn. Amsterdam: North-Holland.

Bénassy, J.-P. 1995. Money and wage contracts in an optimizing model of the business cycle. *Journal of Monetary Economics* 35, 303–15.

Bénassy, J.-P. 2002. *The Macroeconomics of Imperfect Competition and Nonclearing Markets: A Dynamic General Equilibrium Approach*. Cambridge, MA: M.I.T. Press.

Bénassy, J.-P. 2003a. Staggered contracts and persistence: microeconomic foundations and macroeconomic dynamics. *Recherches Economiques de Louvain* 69, 125–44.

Bénassy, J.-P. 2003b. Output and inflation persistence under price and wage staggering: analytical results. *Annales d'Economie et de Statistique*, n° 69, 1–30.

Brock, W. 1975. A simple perfect foresight monetary model. *Journal of Monetary Economics* 1, 133–50.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Chari, V., Kehoe, P. and McGrattan, El. 2000. Sticky price models of the business cycle: can the contract multiplier solve the persistence problem? *Econometrica* 68, 1151–79.

Cho, J.-O. 1993. Money and the business cycle with one-period nominal contracts. *Canadian Journal of Economics* 26, 638–59.

Cho, J.-O. and Cooley, T. 1995. Business cycles with nominal contracts. *Economic Theory* 6, 13–34.

Cho, J.-O., Cooley, T. and Phaneuf, L. 1997. The welfare cost of nominal wage contracting. *Review of Economic Studies* 64, 465–84.

Christiano, L., Eichenbaum, M. and Evans, C. 1999. Monetary policy shocks: what have we learned and to what end? In *Handbook of Macroeconomics*, vol. 1A, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–46.

Clower, R. 1965. The Keynesian counterrevolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cogley, T. and Nason, J. 1993. Impulse dynamics and propagation mechanisms in a real business cycle model. *Economics Letters* 43, 77–81.

Cogley, T. and Nason, J. 1995. Output dynamics in real-business-cycle models. *American Economic Review* 85, 492–511.

Collard, F. and Ertz, G. 2000. Stochastic nominal wage contracts in a cash-in-advance model. *Recherches Economiques de Louvain* 66, 281–301.

Danthine, J.-P. and Donaldson, J. 1990. Efficiency wages and the business cycle puzzle. *European Economic Review* 34, 1275–301.

Danthine, J.-P. and Donaldson, J. 1991. Risk sharing, the minimum wage and the business cycle. In *Equilibrium Theory and Applications: A Conference in Honor of Jacques Drèze*, ed. W. Barnett et al. Cambridge: Cambridge University Press.

Danthine, J.-P. and Donaldson, J. 1992. Risk sharing in the business cycle. *European Economic Review* 36, 468–75.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Drèze, J. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16, 301–20.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85, 191–205.

Gabszewicz, J. and Vial, J.-P. 1972. Oligopoly ‘A la Cournot’ in a general equilibrium analysis. *Journal of Economic Theory* 42, 381–400.

Gordon, D. 1974. A neo-classical theory of Keynesian unemployment. *Economic Inquiry* 12, 431–59.

Gray, J.-A. 1976. Wage indexation: a macroeconomic approach. *Journal of Monetary Economics* 2, 221–35.

Hairault, J.-O. and Portier, F. 1993. Money, new-Keynesian macroeconomics and the business cycle. *European Economic Review* 37, 1533–68.

Hicks, J. 1937. Mr. Keynes and the ‘classics’: a suggested interpretation. *Econometrica* 5, 147–59.

Huang, K. and Liu, Z. 2002. Staggered price-setting, staggered wage-setting, and business cycle persistence. *Journal of Monetary Economics* 49, 405–33.

Jeanne, O. 1998. Generating real persistent effects of monetary shocks: how much nominal rigidity do we really need? *European Economic Review* 42, 1009–32.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. New York: Harcourt Brace.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. Oxford: Oxford University Press.

Long, J. and Plosser, C. 1983. Real business cycles. *Journal of Political Economy* 91, 39–69.

Lucas, R., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Muth, J. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28, 196–201.

Patinkin, D. 1956. *Money, Interest and Prices*, 2nd edn. New York: Harper and Row, 1965.

Phelps, E. and Taylor, J. 1977. Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85, 163–90.

Rotemberg, J. 1982a. Monopolistic price adjustment and aggregate output. *Review of Economic Studies* 44, 517–31.

Rotemberg, J. 1982b. Sticky prices in the United States. *Journal of Political Economy* 90, 1187–211.

Rotemberg, J. and Woodford, M. 1992. Oligopolistic pricing and the effects of aggregate demand on economic activity. *Journal of Political Economy* 100, 1153–207.

Rotemberg, J. and Woodford, M. 1995. Dynamic general equilibrium models with imperfectly competitive product markets. In *Frontiers of Business Cycle Research*, ed. T. Cooley. Princeton: Princeton University Press.

Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review* 57(supplement), 534–44.

Smets, F. and Wouters, R. 2003. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association* 1, 1123–75.

Svensson, L. 1986. Sticky goods prices, flexible asset prices, monopolistic competition and monetary policy. *Review of Economic Studies* 53, 385–405.

Taylor, J. 1979. Staggered wage setting in a macro model. *American Economic Review* 69, 108–13.

Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

Taylor, J. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of Macroeconomics*, vol. 1, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz. Definitive edition trans. W. Jaffe as *Elements of Pure Economics*. London: Allen and Unwin, 1954.

Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37, 345–70.

How to cite this article

Bénassy, Jean-Pascal. "dynamic models with non-clearing markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume.

Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000253> doi:10.1057/9780230226203.0419

The New Palgrave Dictionary of Economics Online

dynamic programming

John Rust

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the history and theory of dynamic programming (DP), a recursive method of solving sequential decision problems under uncertainty. It discusses computational algorithms for the numerical solution of DP problems, and an important limitation in our ability to solve realistic large-scale dynamic programming problems, the ‘curse of dimensionality’. It also summarizes recent research in complexity theory that delineates situations where the curse can be broken (allowing us to solve DPs using fast polynomial time algorithms), and situations where it is insuperable. The literature on econometric estimation and testing of DP models is reviewed, as is another ‘scientific limit to knowledge’, namely, the identification problem.

Keywords

backward induction; Bellman equation; computational complexity; computational experiments; concavity; continuous and discrete time models; curse of dimensionality; decision variables; discount factor; dynamic discrete choice models; dynamic games; dynamic programming; econometric estimation; Euler equations; game tree; identification; independence; indirect inference; infinite horizons; Kalman filtering; Kuhn–Tucker th; Markov chain Monte Carlo methods; Markovian decision problems; maximum likelihood; method of simulated moments; method of simulated scores; minimum residual method; monotonicity; Monte Carlo integration; neural networks; nonlinear regression; nonparametric regression; optimal decision rules; policy iteration; principle of optimality; rational expectations; sequential decision problems; simulated maximum likelihood; simulated method of moments; simulated minimum distance; simulation-based estimation; state variables; stationarity; statistical decision theory; structural estimation; subgame perfection; uncertainty; Wald, A.

Article

1 Introduction

Dynamic programming is a recursive method for solving *sequential decision problems* (hereafter abbreviated as SDP). Also known as *backward induction*, it is used to find *optimal decision rules* in ‘games against nature’ and *subgame perfect equilibria* of dynamic multi-agent games, and competitive equilibria in dynamic economic models. Dynamic programming has enabled economists to formulate and solve a huge variety of problems involving sequential decision-making under uncertainty, and as a result it is now widely regarded as the single most important tool in economics. Section 2 provides a brief history of dynamic programming. Section 3 discusses some of the main theoretical results underlying dynamic programming, and its relation to game theory and optimal control theory. Section 4 provides a brief survey of numerical dynamic programming. Section 5 surveys the experimental and econometric literature that uses dynamic programming to construct empirical models economic behaviour.

2 History

The earliest reference to the use of the method of backward induction to solve decision problems appears to be Arthur Cayley's 1875 solution to the secretary problem (I am grateful to Arthur F. Veinott Jr. for alerting me to this). In the mid-1940s a number of different researchers in economics and statistics appear to have independently discovered backward induction as a way to solve SDPs involving risk or uncertainty. Von Neumann and Morgenstern, in their seminal work on game theory (1944), used backward induction to find what we now call *subgame perfect equilibria of extensive form games*. (‘We proceed to discuss the game Γ by starting with the last move $\mathcal{M}_\nu$ and then going backward from there through the moves $\mathcal{M}_{\nu-1}, \mathcal{M}_{\nu-2} \cdots$’: 1944, p. 126.) Abraham Wald, who is credited with the invention of *statistical decision theory*, extended this theory to sequential decision-making in his 1947 book *Sequential Analysis*. Wald generalized the

problem of gambler's ruin from probability theory and introduced the *sequential probability ratio test* that minimizes the expected number of observations in a sequential generalization of the classical hypothesis test. However, the role of backward induction is less obvious in Wald's work. It was more clearly elucidated in the 1949 paper by Arrow, Blackwell and Girshick. They studied a generalized version of the statistical decision problem and formulated and solved it in a way that is a readily recognizable application of modern dynamic programming. Following Wald, they characterized the optimal rule for making a statistical decision (for example, accept or reject a hypothesis), accounting for the costs of collecting additional observations. In the section 'The Best Truncated Procedure' they show how the optimal rule can be approximated 'Among all sequential procedures not requiring more than N observations …' and solve for the optimal truncated sampling procedure 'by induction backwards' (1949, p. 217).

Other early applications of backward induction include the work of Pierre Massé (1945, p. 196) on statistical hydrology and the management of reservoirs, and Dvoretsky, Kiefer and Wolfowitz's (1952) analysis of optimal inventory policy. Richard Bellman is widely credited with recognizing the common structure underlying SDPs, and showing how backward induction can be applied to solve a huge class of SDPs under uncertainty. Most of Bellman's work in this area was done at the RAND Corporation, starting in 1949. It was there that he invented the term 'dynamic programming' that is now the generally accepted synonym for backward induction. Bellman (1984, p. 159) explained that he invented the name 'dynamic programming' to hide the fact that he was doing mathematical research at RAND under a Secretary of Defense who 'had a pathological fear and hatred of the term, research'. He settled on 'dynamic programming' because it would be difficult give it a 'pejorative meaning' and because 'It was something not even a Congressman could object to'.

3 Theory

Dynamic programming can be used to solve for optimal strategies and equilibria of a wide class of SDPs and multiplayer games. The method can be applied both in discrete time and continuous time settings. The value of dynamic programming is that it is a 'practical' (that is, *constructive*) method for finding solutions to extremely complicated problems. However, continuous time problems involve technicalities that I wish to avoid in this survey. If a continuous time problem does not admit a closed-form solution, the most commonly used numerical approach is to solve an approximate discrete time version of the problem or game, since under very general conditions one can find a sequence of discrete time DP problems whose solutions converge to the continuous time solution the time interval between successive decisions tends to zero (Kushner, 1990). I start by describing how dynamic programming is used to solve single agent 'games against nature'. The approach can be extended to solve multiplayer games, dynamic contracts, principal–agent problems, and competitive equilibria of dynamic economic models. See recursive competitive equilibrium.

3.1 Sequential decision problems

There are two key variables in any dynamic programming problem: a *state variable* s_t, and a *decision variable* d_t (the decision is often called a 'control variable' in the engineering literature). These variables can be vectors in R^n, but in some cases they might be *infinite-dimensional* objects. For example, in Bayesian decision problems, one of the state variables might be a *posterior distribution* for some unknown quantity θ. In general, this posterior distribution lives in an infinite dimensional space of all probability distributions on θ. In heterogeneous agent equilibrium problems state variables can also be distributions. The state variable evolves randomly over time, but the agent's decisions can affect its evolution. The agent has a *utility* or *payoff function* $U(s_1, d_1,\ldots, s_T, d_T)$ that depends on the realized states and decisions from period t=1 to the *horizon* T. In some cases $T=\infty$, and we say the problem is *infinite horizon*. In other cases, such as a life-cycle decision problem, T might be a random variable, representing a consumer's date of death. As we will see, dynamic programming can be adapted to handle either of these possibilities. Most economic applications presume a *discounted*, *time-separable* objective function, that is, U has the form

$$U(s_1, d_1, \ldots, s_T, d_T) = \sum_{t=1}^{T} \beta^t u_t(s_t, d_t) \tag{1}$$

where β is known as a *discount factor* that is typically presumed to be in the (0, 1) interval, and $u_t(s_t, d_t)$ is the agent's *period t utility* (*payoff*) function. Discounted utility and profits are typical examples of time separable payoff functions studied in economics. However, the method of dynamic programming does not require time separability, and so I will describe it without imposing this restriction.

We model the uncertainty underlying the decision problem via a family of history and decision-dependent conditional probabilities $\{p_t(s_t|H_{t-1})\}$ where $H_{t-1} = (s_1, d_1, \ldots, s_{t-1}, d_{t-1})$ denotes the *history*, that is the realized states and decisions from the initial date t=1 to date t=T. Note that this includes all deterministic SDPs as a special case where the transition probabilities p_t are degenerate. In this case we can represent the 'law of motion' for the state variables by deterministic functions $s_{t+1}=f_t(s_t, d_t)$. This implies that in the most general case, $\{s_t, d_t\}$, evolves as a history dependent stochastic process. Continuing the 'game against nature' analogy, it will be helpful to think of $\{p_t(s_t|H_{t-1})\}$ as constituting a 'mixed strategy' played by 'nature' and the agent's optimal

strategy as a 'best response' to nature's strategy.

The final item we need to specify is the *timing of decisions*. Assume that the agent selects d_t *after* observing s_t, which is 'drawn' from the distribution $p_t(s_t|H_{t-1})$. The alternative case where d_t is chosen before s_t is realized can also be handled, but requires a small change in the formulation of the problem. The agent's choice of d_t is restricted to a *state-dependent constraint (choice) set* $D_t(H_{t-1}, s_t)$. We can think of D_t as the generalization of a 'budget set' in standard static consumer theory. The choice set could be a finite set, in which case we refer to the problem as *discrete choice*, or D_t could be a subset of R^k with non-empty interior, then we have a *continuous choice* problem. In many cases, there is a mixture of types of choices, which we refer to as *discrete-continuous choice problems*. An example is commodity price speculation; see for example Hall and Rust (2006), where a speculator has a discrete choice of whether or not to order to replenish his inventory and a continuous decision of how much of the commodity to order. Another example is retirement: a person has a discrete decision of whether to retire and a continuous decision of how much to consume.

Definition: A (single agent) *sequential decision problem* (SDP) consists of (1) a *utility function U*, (2) a sequence of *choice sets* $\{D_t\}$, and (3) a sequence of *transition probabilities* $\{pt(s_t|H_{t-1})\}$ where we assume that the process is initialized at some given initial state s_1.

In order to solve this problem, we have to make assumptions about how the decision-maker evaluates alternative risky strategies. The standard assumption is that the decision-maker maximizes *expected utility*. Backward induction does not necessarily result in optimal strategies for *non-expected utility maximizers*, except for certain classes of *recursive preferences*.

As the name implies, an expected utility maximizer makes decisions that maximize their *ex ante* expected utility. However, since information unfolds over time, it is generally not optimal to *pre-commit* to any fixed sequence of actions $(d_1,\ldots,d_T)$. Instead, the decision-maker can generally obtain higher expected utility by adopting a *history-dependent strategy* or *decision rule* $(\delta_1,\ldots,\delta_T)$. This is a sequence of *functions* such that for each time t the realized decision is a function of all available information. In the engineering literature, a decision rule that does not depend on evolving information is referred to as an *open-loop* strategy, whereas one that does is referred to as a *closed-loop* strategy. In deterministic control problems, the closed-loop and open-loop strategies are the same since both are simple functions of time. However in stochastic control problems, open-loop strategies are a strict subset of closed-loop strategies. Under our timing assumptions the information available at time t is (H_{t-1}, s_t), so we can write $d_t = \delta_t(H_{t-1}, s_t)$. By convention we set $H_0 = \varnothing$ so that the available information for making the initial decision is just s_1. A decision rule is *feasible* if it also satisfies $\delta_t(H_{t-1}, s_t) \in D_t(H_{t-1}, s_t)$ for all (s_t, H_{t-1}). Each feasible decision rule can be regarded as a 'lottery' whose payoffs are utilities, the expected value of which corresponds to expected utility associated with the decision rule. An *optimal decision rule* $\delta^* \equiv (\delta_1^*, \ldots, \delta_T^*)$ is simply a feasible decision rule that maximizes the decision-maker's expected utility

$$\delta^* = \underset{\delta \in \mathcal{F}}{\operatorname{argmax}} E\{U(\{\tilde{s}_t, \tilde{d}_t\}_\delta)\}, \tag{2}$$

where $\mathcal{F}$ denotes the class of feasible history-dependent decision rules, and $\{\tilde{s}_t, \tilde{d}_t\}_\delta$ denotes the stochastic process induced by the decision rule $\delta \equiv (\delta_1, \ldots, \delta_T)$. Problem (2) can be regarded as a static, *ex ante* version of the agent's problem. In game theory, (2) is referred to as the *normal form* or the *strategic form* of a dynamic game, since the dynamics are suppressed and the problem has the superficial appearance of a static optimization problem or game in which an agent's problem is to choose a best response, either to nature (in the case of single agent decision problems) or to other rational opponents (in the case of games). The strategic formulation of the agent's problem is quite difficult to solve since the solution is a *sequence of history-dependent functions* $\delta^* = (\delta_1^*, \ldots, \delta_T^*)$ for which standard finite dimensional constrained optimization techniques (for example, the Kuhn–Tucker th) are inapplicable. (If we consider problems where all states can assume only a finite number of values, it is possible to apply standard finite dimensional Kuhn–Tucker constrained optimization methods, but if the state variables can assume a continuum of possible values, the programming problem becomes an infinite dimensional programming problem for which optimal control and dynamic programming methods are more appropriate. See Luenberger, 1969, for a more thorough discussion of how Lagrange multipliers and Kuhn–Tucker methods can be extended to problems where decisions are infinite-dimensional objects. These methods are usually applied in deterministic context, and there is a specialized literature on optimal control for solving such problems.) See Pontryagin's principle of optimality.

3.2 Solving sequential decision problems by backward induction

To carry out backward induction, we start at the last period, T, and for each possible combination (H_{T-1}, s_T) we calculate the time T *value function* and *decision rule* (we will discuss how backward induction can be extended to cases where T is random or where $T=\infty$ shortly).

$$V_T(H_{T-1}, s_T) = \max_{d_T \in D_T(H_{T-1}, s_T)} U(H_{T-1}, s_T, d_T)$$

$$\delta_T(H_{T-1}, s_T) = \underset{d_T \in D_T(H_{T-1}, s_T)}{\operatorname{argmax}} U(H_{T-1}, s_T, d_T), \tag{3}$$

where we have written $U(H_{T-1}, s_T, d_T)$ instead of $U(s_1, d_1, \ldots, s_T, d_T)$ since $H_{T-1}=(s_1, d_1, \ldots, s_{T-1}, d_{T-1})$. Next we move backward one time period to time T–1 and compute

$$V_{T-1}(H_{T-2}, s_{T-1}) = \max_{d_{T-1} \in D_{T-1}(H_{T-2}, s_{T-1})} E\{V_T(H_{T-2}, s_{T-1}, d_{T-1}, \tilde{s}_T) | H_{T-2}, s_{T-1}, d_{T-1}\} = \max_{d_{T-1} \in D_{T-1}(H_{T-2}, s_{T-1})} \int V_T(H_{T-2}, s_{T-1}, d_{T-1}, s_T)\, p_T(s_T | H_{T-2}, s_{T-1}, d_{T-1})$$

$$\delta_{T-1}(H_{T-2}, s_{T-1}) = \underset{d_{T-1} \in D_{T-1}(H_{T-2}, s_{T-1})}{\operatorname{argmax}} E\{V_T(H_{T-2}, s_{T-1}, d_{T-1}, \tilde{s}_T) | H_{T-2}, s_{T-1}, d_{T-1}\}, \tag{4}$$

where the integral in eq. (4) is the formula for the conditional expectation of V_T, where the expectation is taken with respect to the random variable $\tilde{s}_T$ whose value is not known as of time T–1. We continue the backward induction recursively for time periods T–2, T–3, … until we reach time period t=1. The equation for the value function V_t in an arbitrary period t is defined recursively by an equation that is now commonly called the *Bellman equation*

$$V_t(H_{t-1}, s_t) = \max_{d_t \in D_t(H_{t-1}, s_t)} E\{V_{t+1}(H_{t-1}, s_t, d_t, \tilde{s}_{t+1}) | H_{t-1}, s_t, d_t\} = \max_{d_t \in D_t(H_{t-1}, s_t)} \int V_{t+1}(H_{t-1}, s_t, d_t, s_{t+1})\, p_{t+1}(s_{t+1} | H_{t-1}, s_t, d_t). \tag{5}$$

The decision rule δ_t is defined by the value of d_t that attains the maximum in the Bellman equation for each possible value of (H_{t-1}, s_t)

$$\delta_t(H_{t-1}, s_t) = \underset{d_t \in D_t(H_{t-1}, s_t)}{\operatorname{argmax}} E\{V_{t+1}(H_{t-1}, s_t, d_t, \tilde{s}_{t+1}) | H_{t-1}, s_t, d_t\}. \tag{6}$$

Backward induction ends when we reach the first period, in which case, as we will now show, the function $V_1(s_1)$ provides the expected value of an optimal policy, starting in state s_1 implied by the recursively constructed sequence of decision rules $\delta=(\delta_1,\ldots, \delta_T)$.

3.3 The principle of optimality

The key idea underlying why backward induction produces an optimal decision rule is called

The principle of optimality: : *an optimal decision rule* $\delta^* = (\delta_1^*, \ldots, \delta_T^*)$ *has the property that given any* $t \in \{1,\ldots, T\}$ *and any history* H_t–1 *in the support of the controlled process* $\{s_t, d_t\}_{\delta^*}$, δ^* *remains optimal for the 'subgame' starting at time t and history* H_t*–1. That is,* δ^* *maximizes the "continuation payoff" given by the conditional expectation of utility from period t to T, given history* H_t*–1*:

$$\delta^* = \underset{\delta}{\operatorname{argmax}}\, E\{U(\{s_t, d_t\}_\delta)|H_{t-1}\}. \tag{7}$$

In game theory, the principle of optimality is equivalent to the concept of a *subgame perfect equilibrium* in an *extensive form game*. When all actions and states are discrete, the stochastic decision problem can be diagrammed as a *game tree*. The principle of optimality, which in game theory is equivalent to the concept of a *subgame perfect equilibrium*, guarantees that if δ^* is an optimal strategy (or equilibrium strategy) for the overall game tree, then it must also be an optimal strategy for every subgame, or, more precisely, *all subgames that are reached with positive probability from the initial node*.

It should now be evident why there is a need for the qualification 'for all H_{t-1} in the support of $\{s_t, d_t\}_{\delta^*}$, in the statement of the principle of optimality. There are some subgames that are never reached with positive probability under an optimal strategy. Thus, it is easy to construct alternative optimal decision rules that do not satisfy the principle of optimality because they involve taking suboptimal decisions on 'zero probability subgames'. Since these subgames are never reached, such modifications do not jeopardize *ex ante* optimality. However we cannot be sure *ex ante* which subgames will be irrelevant *ex post* unless we carry out the full backward induction process. Dynamic programming results in strategies that are optimal in *every possible subgame*, even those which will never be reached when the strategy is executed. Since backward induction results in a decision rule δ that is optimal for all possible subgames, it is intuitively clear that δ is optimal for the game as a whole, that is, it is a solution to the *ex ante* strategic form of the optimization problem (2).

For a formal proof of this result for games against nature (with appropriate care taken to ensure measurability and existence of solutions), see Gihman and Skorohod (1979). If in addition to 'nature' we extend the game tree by adding another rational expected utility maximizing player, then backward induction can be applied in the same way to solve this alternating move dynamic game. Assume that player 1 moves first, then player 2, then nature, and so on. Dynamic programming results in a *pair of strategies* for both players. Nature still plays a 'mixed strategy' that could depend on the entire previous history of the game, including all the previous moves of both players. The backward induction process ensures that each player can predict the future choices of their opponent, not only in the succeeding move but in all future stages of the game. The pair of strategies (δ^1, δ^2) produced by dynamic programming are mutual best responses, as well as being best responses to nature's moves. Thus, these strategies constitute a *Nash equilibrium*. They actually satisfy a stronger condition: they are Nash equilibrium strategies in every possible subgame of the original game, and thus are *subgame-perfect* (Selten, 1975). Subgame-perfect equilibria exclude 'implausible equilibria' based on *incredible threats*. A standard example is an incumbent's threat to engage in a price war if a potential entrant enters the market. This threat is incredible if the incumbent would not really find it advantageous to engage in a price war (resulting in losses for both firms) if the entrant called its bluff and entered the market. Thus the set of all Nash equilibria to dynamic multiplayer games is strictly larger than the subset of subgame-perfect equilibria, a generalization of the fact that, in single agent decision problems, the set of optimal decision rules includes ones which take suboptimal decisions on subgames that have zero chance of being reached for a given optimal decision rule. Dynamic programming ensures that the decision-maker would never mistakenly reach any such subgame, similar to the way subgame perfection ensures that a rational player would not be fooled by an incredible threat.

3.4 Dynamic programming for stationary, Markovian, infinite-horizon problems

The *complexity* of dynamic programming arises from the exponential growth in the number of possible histories as the number of possible values for the state variables, decision variables, and/or number of time periods T increases. For example, in a problem with N possible values for s_t and D possible values for d_t in each time period t, there are $[ND]^T$ possible histories, and thus the required number of calculations to solve a general T period, history-dependent dynamic programming problem is $O([ND]^T)$. Bellman and Dreyfus (1962) referred to this exponential growth in the number of calculations as the *curse of dimensionality*. In the next section, I will describe various strategies for dealing with this problem, but an immediate solution is to restrict attention to *time separable Markovian decision problems*. These are problems where the payoff function U is additively separable as in eq. (1), and where both the choice sets $\{D_t\}$ and the transition probabilities $\{p_t\}$ depend only on the contemporaneous state variable s_t and not on the entire previous history H_{t-1}. We say a conditional distribution p_t satisfies the *Markov property* if it depends on the previous history only via the most recent values, that is, if $p_t(s_t|H_{t-1}) = p_t(s_t|s_{t-1}, d_{t-1})$. In this case backward induction becomes substantially easier. For example, in this case the dynamic

programming optimizations have to be performed only at each of the N possible values of the state variable at each time t, so only $O(NDT)$ calculations are required to solve a time T period time separable Markovian problem instead of $O([ND]^T)$ calculations when histories matter. This is part of the reason why, even though time non-separable utilities and non-Markovian forms of uncertainty may be more general, most dynamic programming problems that are solved in practical applications are both time separable and Markovian.

SDPs with random horizons $\tilde{T}$ can be solved by backward induction provided there is some finite time $\bar{T}$ satisfying $Pr\{\tilde{T} \leq \bar{T}\} = 1$. In this case, backward induction proceeds from the maximum possible value $\bar{T}$ and the *survival probability* $\rho_t = Pr\{\tilde{T} > t | \tilde{T} \geq t\}$ is used as to capture the probability that the problem will continue for at least one more period. The Bellman equation for the discounted, time-separable utility with uncertain lifetime is

$$V_t(s_t) = \max_{d \in D_t(s_t)} [u_t(s_t, d) + \rho_t \beta E V_{t+1}(s_t, d)]$$

$$\delta_t(s_t) = \operatorname*{argmax}_{d \in D_t(s_t)} [u_t(s_t, d) + \rho_t \beta E V_{t+1}(s_t, d)], \tag{8}$$

where

$$E V_{t+1}(s, d) = \int_{s'} V_{t+1}(s') p_{t+1}(s' | s, d). \tag{9}$$

In many problems there is no finite upper bound $\bar{T}$ on the horizon. These are called *infinite horizon problems* and they occur frequently in economics. For example, SDPs used to model decisions by firms are typically treated as infinite horizon problems. It is also typical in infinite horizon problems to assume *stationarity*. That is, the utility function $u(s, d)$, the constraint set $D(s)$, the survival probability ρ, and the transition probability $p(s' | s, d)$ do not explicitly depend on time t. In such cases, it is not hard to show the value function and the optimal decision rules are also stationary, and satisfy the following version of Bellman's equation

$$V(s) = \max_{d \in D(s)} [u(s, d) + \rho \beta E V(s, d)]$$

$$\delta(s) = \operatorname*{argmax}_{d \in D(s)} [u(s, d) + \rho \beta E V(s, d)], \tag{10}$$

where

$$EV(s, d) = \int_{S'} V(s') p(s'|s, d). \tag{11}$$

This is a fully recursive definition of *V*, and as such there is an issue of existence and uniqueness of a solution. In addition, it is not obvious how to carry out backward induction, since there is no 'last' period from which to begin the backward induction process. However, under relatively weak assumptions one can show there is a unique *V* satisfying the Bellman equation, and the implied decision rule in eq. (10) is an optimal decision rule for the problem. Further, this decision rule can be approximated by solving an approximate finite horizon version of the problem by backward induction.

For example, suppose that $u(s, d)$ is a continuous function of (s, d), the state space S is compact, the constraint sets $D(s)$ are compact for each $s \in S$, and the transition probability $p(s'|s, d)$ is weakly continuous in (s, d) (that is, $EV(s, d) \equiv \int_{S'} W(s') p(s'|s, d)$ is a continuous function of (s, d) for each continuous function $W{:}S \to R$). Blackwell (1965a; 1965b), Denardo (1967) and others have proved that, under these sorts of assumptions, *V* is the unique fixed point to the *Bellman operator* $\Gamma : B \to B$, where *B* is the Banach space of continuous functions on *S* under the supremum norm, and Γ is given by

$$\Gamma(W)(s) = \max_{d \in D(s)} \left[u(s, d) + \rho\beta \int_{S'} W(s') p(s'|s, d) \right]. \tag{12}$$

The existence and uniqueness of *V* is a consequence of the *contraction mapping th*, since Γ can be shown to satisfy the contraction property,

$$\|\Gamma W - \Gamma V\| \le \alpha \|W - V\|, \tag{13}$$

where $\alpha \in (0,1)$ and $\|W\| = \sup_{s \in S} |W(s)|$. In this case, $\alpha = \rho\beta$, so the Bellman operator will be a contraction mapping if $\rho\beta \in (0,1)$.
The proof of the optimality of the decision rule δ in eq. (10) is somewhat more involved. Using the Bellman equation (10), we will show that (see eq. (34) in section 4),

$$V(s) = u(s, \delta(s)) + \rho\beta \int_{S'} V(s') p(s'|s, \delta(s)) = E\left\{ \sum_{t=0}^{\infty} [\rho\beta]^t u(s_t, \delta(s_t)) \middle| s_0 = s \right\}, \tag{14}$$

that is, *V* is the value function implied by the decision rule δ. Intuitively, the boundedness of the utility function, combined with discounting of future utilities, $\rho\beta \in (0,1)$, implies that if we truncate the infinite horizon problem to a *T* period problem, the error in doing so would be arbitrarily small when *T* is sufficiently large. Indeed, this is the key to understanding how to find approximately optimal decision rules to infinite horizon SDPs: *we approximate the infinite horizon decision rule* δ *by solving an approximate finite horizon version of the problem by dynamic programming*. The validity of this approach can be formalized using a well-known property of contraction mappings, namely, that the *method of successive approximations* starting from any initial guess *W* converges to the fixed point of Γ, that is

$$\lim_{t \to \infty} V_t = \Gamma^t(W) = V \; \forall W \in B, \tag{15}$$

where $\Gamma^t W$ denotes t successive iterations of the Bellman operator Γ,

$$\begin{gathered} V_0 = \Gamma^0(W) = W \\ V_1 = \Gamma^1(W) \\ \cdots \\ V_t = \Gamma^t(W) = \Gamma(\Gamma^{t-1}W) = \Gamma(V_{t-1}) . \end{gathered} \tag{16}$$

If W=0 (that is the zero function in B), then $V_T = \Gamma^T(0)$ is simply the period t=1 value function resulting from the solution of a T period dynamic programming problem. Thus, this result implies that the optimal value function V_T for a T-period approximation to the infinite horizon problem converges to V as $T \to \infty$. Moreover, the difference in the two functions satisfies the bound

$$\|V_T - V\| \le \frac{[\rho\beta]^T \|W\|}{1 - \rho\beta}. \tag{17}$$

Let $\delta_T = \delta_{1,T}, \delta_{2,T}, \ldots, \delta_{T,T}$ be the optimal decision rule to the T period problem. It can be shown that, if we follow this decision rule up to period T and then use $\delta_{1,T}$ in every period after T, the resulting decision rule is approximately optimal in the sense that the value function for this infinite horizon problem also satisfies inequality (17), and thus can be made arbitrarily small as T increases.

In many cases in economics the state space S has no natural upper bound. An example might be where s_t denotes an individual's wealth at time t, or the capital stock of the firm. If the unboundedness of the state space results in unbounded payoffs, the contraction mapping argument must be modified since the Banach space structure under the *supremum* norm no longer applies to unbounded functions. Various alternative approaches have been used to prove existence of optimal decision rules for unbounded problems. One is to use an alternative norm (for example, a *weighted norm*) and demonstrate that the Banach space/contraction mapping argument still applies. However, there are cases where there are no natural weighted norms, and the contraction mapping property cannot hold since the Bellman equation can be shown to have multiple solutions. The most general conditions under which the existence and uniqueness of the solution V to the Bellman equation and the optimality of the implied stationary decision rule δ has been established is in Bhattacharya and Majumdar (1989). However, as I discuss in the next section, considerable care must be taken in solving unbounded problems numerically.

4 Numerical dynamic programming and the curse of dimensionality

The previous section showed that dynamic programming is a powerful tool that has enabled us to formulate and solve a wide range of economic models involving sequential decision-making under uncertainty – at least 'in theory'. Unfortunately, the cases where dynamic programming results in *analytical, closed-form solutions* are rare and often rather fragile in the sense that small changes in the formulation of a problem can destroy the ability to obtain an analytical solution. However even though most problems do not have analytical solutions, the theorems in the previous section guarantee the *existence* of solutions, and these solutions can be calculated (or approximated) by numerical methods. Since the 1980s, faster computers and better numerical methods have made dynamic programming a tool of substantial practical value by significantly expanding the range of problems that can be solved. In particular, it has led to the development of a large and rapidly growing literature on econometric estimation and testing of 'dynamic structural models' that I will discuss in the next section.

However, there are still many difficult challenges that prevent us from formulating and solving models that are as detailed and realistic as we might like, a problem that is especially acute in

empirical applications. The principal challenge is what Bellman and Dreyfus (1962) called *the curse of dimensionality*. We have already illustrated this problem in Section 3.4: for history-dependent SDPs with a finite horizon T and a finite number of states N and actions D, dynamic programming requires $O([ND]^T)$ operations to find a solution. Thus it appears that the time required to compute a solution via dynamic programming increases *exponentially fast* with the number of possible decisions or states in a dynamic programming problem.

Fortunately, computer power (for example, operations per second) has also been growing exponentially fast, a consequence of *Moore's Law* and other developments in information technology, such as improved communications and massive parallel processing. Bellman and Dreyfus (1962) carried out calculations on RAND's 'Johnniac' computer (named in honour of Jon von Neumann, whose work contributed to the development of the first electronic computers) and reported that this machine could do 12,500 additions per second. Nowadays, in 2007, a typical laptop computer can do over a billion operations per second and we now have supercomputers that are approaching a *thousand trillion operations per second* – a level known as a 'petaflop'. In addition to faster 'hardware', research on numerical methods has resulted in significantly better 'software' that has had a huge impact on the spread of numerical dynamic programming and on the range of problems we can solve. In particular, algorithms have been developed that succeed in 'breaking' the curse of dimensionality, enabling us to solve in polynomial time classes of problems that were previously believed to be solvable only in exponential time. The key to breaking the curse of dimensionality is the ability to recognize and exploit *special structure* in an SDP problem. We have already illustrated an example of this in Section 3.4: if the SDP is Markovian and utility is time separable, a finite horizon, finite state SDP can be solved by dynamic programming in only $O(NDT)$ operations, compared to the $O([ND]^T)$ operations that are required in the general history-dependent case. There is only enough space here to discuss several of the most commonly used and most effective numerical methods for solving different types of SDPs by dynamic programming. I refer the reader to Puterman (1994), Rust (1996) and Judd (1998) for more in-depth surveys on the literature on numerical dynamic programming. See computational methods in econometrics.

Naturally, the numerical method that is appropriate or 'best' depends on the type of problem being solved. Different methods are applicable depending on whether the problem has (*a*) finite versus infinite horizon, (*b*) finite versus continuous-valued state and decision variables, and (*c*) single versus multiple players. In finite horizon problems, backward induction is the essentially the only approach, although as we will see there are many different choices about how to most implement it most efficiently – especially in discrete problems where the number of possible values for the state variables is huge (for example, chess) or in problems with continuous state variables. In the latter case, it is clearly not possible to carry out backward induction for every possible history (or value of the state variable at stage t if the problem is Markovian and time separable), since there are infinitely many (indeed a continuum) of them. In these cases, it is necessary to *interpolate* the value function, whose values are only explicitly computed at a finite number of points in the state space. I use the term 'grid' to refer to the finite number of points in the state space where the backward induction calculations are actually performed. Grids might be *lattices* (that is, regularly spaced sets of points formed as Cartesian products of unidimensional grids for each of the continuous state variables), or they may be *quasi-random grids* formed by randomly sampling the state space from some probability distribution, or by generating deterministic sequences of points such as *low discrepancy sequences*. The reason why one might choose a random or low-discrepancy grid instead of regularly spaced lattice is to break the curse of dimensionality, as I discuss shortly. Also, in many cases it is advantageous to refine the grid over the course of the backward induction process, starting out with an initial 'coarse' grid with relatively few points and subsequently increasing the number of points in the grid as the backward induction progresses. I will have more to say about such *multigrid* and *adaptive grid* methods when I discuss solution of infinite horizon problems below.

Once a particular grid is chosen, the backward induction process is carried out in the way it would be normally be done in a finite state problem. On the assumption that the problem is Markovian and the utility is time separable and there are n grid points $\{s_1, \ldots, s_n\}$, this involves the following calculation at each grid point s_i, $i=1,\ldots, n$

$$V_t(s_i) = \max_{d \in D_t(s_i)} [u_t(s_i, d) + \rho\beta\hat{E}V_{t+1}(s_i, d)], \tag{18}$$

where $\hat{E}V_{t+1}(s_i, d)$ is a numerical estimate of the conditional expectation of next period's value function. I will be more specific below about which numerical integration methods are appropriate, but at this point it suffices to note that they are all simple weighted sums of values of the value function at t+1, $V_{t+1}(s)$. We can now see that, even if the actual backward induction calculations are carried out only at the n grid points $\{s_1,\ldots, s_n\}$, we will still have to do numerical integration to compute $\hat{E}V_{t+1}(s_i, d)$ and the latter calculation may require values of $V_{t+1}(s)$ at points s off the grid, that is at points $s \notin \{s_1, \ldots, s_n\}$. This is why some form or interpolation (or in some cases *extrapolation*) is typically required. Almost all methods of interpolation can be represented as weighted sums of the value function at its known values $\{V_{t+1}(s_1), \ldots, V_{t+1}(s_n)\}$ at the n grid points, which were calculated by backward induction at the previous stage. Thus, we have

$$\hat{V}_{t+1}(s) = \sum_{j=1}^{n} w_j(s) V_{t+1}(s_j), \tag{19}$$

where $w_i(s)$ is a weight assigned to the ith grid point that depends on the point s in qst. These weights are typically positive and sum to 1. For example in *multilinear interpolation* or *simplicial interpolation* the $w_i(s)$ weights are those that allow s to be represented as a convex combination of the vertices of the smallest lattice hypercube containing s. Thus, the weights $w_i(s)$ will be zero for all i except the immediate neighbours of the point s. In other cases, such as *kernel density* and *local linear regression*, the weights $w_i(s)$ are generally non-zero for all i, but the weights will be highest for the grid points $\{s_1,\ldots, s_n\}$ which are the *nearest neighbours* of s. An alternative approach can be described as *curve fitting*. Instead of attempting to interpolate the calculated values of the value function at the grid points, this approach treats these values as a *data-set* and estimates parameters θ of a flexible functional form approximation to $V_{t+1}(s)$ by *nonlinear regression*. Using the estimated $\hat{\theta}_{t+1}$ from this nonlinear regression, we can 'predict' the value of $V_{t+1}(s)$ at any $s \in S$

$$\hat{V}_{t+1}(s) = f(s, \hat{\theta}_{t+1}). \tag{20}$$

A frequently used example of this approach is to approximate $V_{t+1}(s)$ as a linear combination of K 'basis functions' $\{b_1(s),\ldots, b_K(s)\}$. This implies that $f(s, \theta)$ takes the form of a *linear regression* function

$$f(s, \theta) = \sum_{k=1}^{K} \theta_k b_k(s), \tag{21}$$

and $\hat{\theta}_{t+1}$ can be estimated by *ordinary least squares*. *Neural networks* are an example where f depends on θ in a nonlinear fashion. Partition θ into subvectors $\theta = (\gamma, \lambda, \alpha)$, where γ and λ are vectors in R^J, and $\alpha = (\alpha_1,\ldots, \alpha_J)$, where each α_j has the same dimension as the state vector s. Then the neural network f is given by

$$f(s, \theta) = f(s, \gamma, \lambda, \alpha) = \sum_{j=1}^{J} \gamma_j \varphi(\lambda_j + (s, \alpha_j)) \tag{22}$$

where (s, α_j) is the inner product between s and the conformable vector α_j, and ϕ is a 'squashing function' such as the logistic function $\varphi(x) = \exp\{x\} / (1 + \exp\{x\})$. Neural networks are known to be 'universal approximators' and require relatively few parameters to provide good approximations to nonlinear functions of many variables. For further details on how neural networks are applied, see the book by Bertsekas and Tsitsiklis (1996) on *Neuro-Dynamic Programming*.

All these methods require extreme care for problems with *unbounded state spaces*. By definition, any finite grid can cover only a small subset of the state space in this case, and thus any of the methods discussed above would require *extrapolation* of the value function to predict its values in regions where there are no grid points, and thus 'data' on what its proper values should be. Not only may mistakes that lead to incorrect extrapolations in these regions lead to errors in the regions where there are no grid points, but the errors can 'unravel' and also lead to considerable errors in approximating the value function in regions where we do have grid points. Attempts to 'compactify' an unbounded problem by arbitrarily truncating the state space may also lead to inaccurate solutions, since the truncation is itself an implicit form of extrapolation (for example, some assumption needs to be made what to do when state variables approach the 'boundary' of the state space: do we assume a 'reflecting boundary', an 'absorbing boundary', and so on?). For example in life-cycle optimization problems, there is no natural upper bound on wealth, even if it is true that there is only a finite amount of wealth in the entire economy. We can always ask the qst, if a person had wealth near the 'upper bound', what would happen to next period wealth if he invested some of it? Here we can see that, if we extrapolate the value function by assuming that the value function is bounded in wealth, this means that by definition there is no incremental return

to saving as we approach the upper bound. This leads to lower saving, and this generally leads to errors in the calculated value function and decision rule far below the assumed upper bound. There is no good general solution to this problem except to solve the problem on a much bigger (bounded) state space than one would expect to encounter in practice, in the hope that extrapolation-induced errors in approximating the value function die out the further one is from the boundary. This property should hold for problems where the probability that the next period state will hit or exceed the 'truncation boundary' gets small the farther the current state is from this boundary.

When a method for interpolating/extrapolating the value function has been determined, a second choice must be made about the appropriate method for *numerical integration* in order to approximate the conditional expectation of the value function $EV_{t+1}(s, d)$ given by

$$EV_{t+1}(s, d) = \int_{s'} V_{t+1}(s') p_{t+1}(s' | s, d). \tag{23}$$

There are two main choices here: (1) deterministic quadrature rules or (2) (quasi-) Monte Carlo methods. Both methods can be written as weighted averages of form

$$\hat{E}V_{t+1}(s, d) = \sum_{i=1}^{N} w_i(s, d) V_{t+1}(a_i), \tag{24}$$

where $\{w_i(s, d)\}$ are *weights*, and $\{a_i\}$ are *quadrature abscissae*. Deterministic quadrature methods are highly accurate (for example, an N-point Gaussian quadrature rule is constructed to exactly integrate all polynomials of degree $2N-1$ or less), but become unwieldy in multivariate integration problems when *product rules* (tensor products of unidimensional quadrature) are used. *Any* sort of deterministic quadrature method can be shown to be subject to the curse of dimensionality in terms of worst-case computational complexity (see Traub and Werschulz, 1998). For example, if $N=O(1/\varepsilon)$ quadrature points are necessary to approximate a univariate integral within ε, then in a d-dimensional integration problem $N^d = O(1/\varepsilon^d)$ quadrature points would be necessary to approximate the integral with an error of ε, which implies that computational effort to find an ε-approximation increases exponentially fast in the problem dimension d. Using the theory of computational complexity, one can prove that *any* deterministic integration procedure is subject to the curse of dimensionality, at least in terms of a 'worst case' measure of complexity. The curse of dimensionality can disappear if one is willing to adopt a Bayesian perspective and place a 'prior distribution' over the space of possible integrands and consider an 'average case' instead of a 'worst case' notion of computational complexity.

Since multivariate integration is a 'sub-problem' that must be solved in order to carry out dynamic programming when there are continuous state variables (indeed, dynamic programming in principle involves infinitely many integrals in order to calculate $EV_{t+1}(s, d)$, one for each possible value of (s, d)), if there is a curse of dimensionality associated with numerical integration of a single multivariate integral, then it should also not be surprising that dynamic programming is also subject to the same curse. There is also a curse of dimensionality associated with global optimization of nonconvex objective functions of continuous variables. Since optimization is also a sub-problem of the overall dynamic programming problem, this constitutes another reason why dynamic programming is subject to a curse of dimensionality. Under the standard worst case definition of computational complexity, Chow and Tsitsiklis (1989) proved that *no* deterministic algorithm can succeed in breaking the curse of dimensionality associated with a sufficiently broad class of dynamic programming problems with continuous state and decision variables. This negative result dashes the hopes of researchers dating back to Bellman and Dreyfus (1962), who conjectured that there might be sufficiently clever deterministic algorithms that can overcome the curse of dimensionality.

However, there are examples of *random algorithms* that can circumvent the curse of dimensionality. Monte Carlo integration is a classic example. Consider approximating the (multidimensional) integral in eq. (23) by using *random* quadrature abscissae $\{\tilde{a}_i\}$ that are N independent and identically distributed (*IID*) draws from the distribution $p_{t+1}(s' | s, d)$ and uniform quadrature weights equal to $w_i(s, d)=1/N$. Then the law of large numbers and the central limit theorem imply that the Monte Carlo integral $\hat{E}V_{t+1}(s, d)$ converges to the true conditional expectation $EV_{t+1}(s, d)$ at rate $1/\sqrt{N}$ *regardless of the dimension of the state space d*. Thus a random algorithm, Monte Carlo integration, succeeds in breaking the curse of dimensionality of multivariate integration. Unfortunately, randomization does *not* succeed in breaking the curse of dimensionality associated with general nonconvex optimization problems with continuous multidimensional decision variables d (see Nemirovsky and Yudin, 1983).

However, naive application of Monte Carlo integration will not necessarily break the curse of dimensionality of the dynamic programming problem. The reason is that a form of *uniform*

convergence (as opposed to pointwise) convergence of the conditional expectations $\hat{E}V_{t+1}(s, d)$ to $EV_{t+1}(s, d)$ is required in order to guarantee that the overall backward induction process converges to the true solution as the number of Monte Carlo draws, N, gets large. To get an intuition why, note that if separate *IID* sets of quadrature abscissae $\{\tilde{s}_i\}$ where drawn for each (s, d) point that we wish to evaluate the Monte Carlo integral $\hat{E}V_{t+1}(s, d)$ at, the resulting function would be an extremely 'choppy' and irregular function of (s, d) as a result of all the random variation in the various sets of quadrature abscissae. Extending an idea introduced by Tauchen and Hussey (1991) to solve rational expectations models, Rust (1997) proved that it is possible to break the curse of dimensionality in a class of SDPs where the choice sets $D_t(s)$ are finite, a class he calls *discrete decision processes*. The restriction to finite choice sets is necessary, since, as noted above, randomization does not succeed in breaking the curse of dimensionality of nonconvex optimization problems with continuous decision variables. The key idea is to choose, as a *random grid*, the same set of random points that are used quadrature abscissae for Monte Carlo integration. That is, suppose $p_{t+1}(s'|s, d)$ is a transition *density* and the state space (perhaps after translation and normalization) is identified with the d-dimensional *hypercube* $S=[0,1]^d$. Apply Monte Carlo integration by drawing N *IID* points $\{\tilde{s}_1, \ldots, \tilde{s}_N\}$ from the this hypercube (this can be accomplished by drawing each component of s_i from the uniform distribution on the [0,1] interval). We have

$$\hat{E}V_{t+1}(s, d) = \frac{1}{N}\sum_{i=1}^{N} V_{t+1}(\tilde{s}_i)\, p_{t+1}(\tilde{s}_i|s, d). \tag{25}$$

Applying results from the theory of *empirical processes* (Pollard, 1989), Rust showed that this form of the Monte Carlo integral does result in uniform convergence (that is, $\mathrm{P}\hat{E}V_{t+1}(s, d) - EV_{t+1}(s, d)\mathrm{P} = O_p(1/\sqrt{N})$), and, using this, he showed that this randomized version of backward induction succeeds in breaking the curse of dimensionality of the dynamic programming problem. The intuition of why this works is, instead of trying to approximate the conditional expectation in (23) by computing *many independent Monte Carlo integrals* (that is, drawing separate sets of random abscissae $\{\tilde{s}_i\}$ from $p_{t+1}(s'|s, d)$ for each possible value of (s, d)), the approach in eq. (25) is to compute a *single Monte Carlo integral* where the random quadrature points $\{\tilde{s}_i\}$ are drawn from the uniform distribution on $[0,1]^d$, and the integrand is treated as the *function* $V_{t+1}(s')\, p_{t+1}(s'|s, d)$ instead of $V_{t+1}(s')$. The second important feature is that eq. (25) has a *self-approximating* property: that is, since the quadrature abscissae are the same as the grid points at which we compute the value function, no auxiliary interpolation or function approximation is necessary in order to evaluate $\hat{E}V_{t+1}(s, d)$. In particular, if $p_{t+1}(s'|s, d)$ is a smooth function of s, then $\hat{E}V_{t+1}(s, d)$ will also be a smooth function of s. Thus, backward induction using this algorithm is extremely simple. Before starting backward induction we choose a value for N and draw N *IID* random vectors $\{\tilde{s}_1, \ldots, \tilde{s}_N\}$ from the uniform distribution on the d-dimensional hypercube. This constitutes a random grid that remains fixed for the duration of the backward induction. Then we begin ordinary backward induction calculations, at each stage t computing $V_t(\tilde{s}_i)$ at each of the N random grid points, and using the self-approximating formula (25) to calculate the conditional expectation of the period t+1 value function using only the N stored values $(V_{t+1}(\tilde{s}_1), \ldots, V_{t+1}(\tilde{s}_N))$ from the previous stage of the backward induction. See Keane and Wolpin (1994) for an alternative approach, which combines Monte Carlo integration with the curve-fitting approaches discussed above. Note that the Keane and Wolpin approach will not generally succeed in breaking the curse of dimensionality since it requires approximation of functions of d variables which is also subject to a curse of dimensionality, as is well known from the literature on *nonparametric regression*.

There are other subclasses of SDPs for which it is possible to break the curse of dimensionality. For example, the family of *linear quadratic/Gaussian* (LQG) can be solved in polynomial time using highly efficient matrix methods, including efficient methods for solving the *matrix Ricatti equation* which is used to compute the *Kalman filter* for Bayesian LQG problems (for example, problems where agents only receive a noisy signal of a state variable of interest, and they update their beliefs about the unknown underlying state variable via Bayes rule).

Now consider stationary, infinite horizon Markovian decision problems. As noted in Section 3.4, there is no 'last' period from which to begin the backward induction process. However, if the utility function is time separable and discounted, then, under fairly general conditions, it will be possible to approximate the solution arbitrarily closely by solving a finite horizon version of the problem, where the horizon T is chosen sufficiently large. As we noted in Section 3.4, this is equivalent to solving for V, the fixed point to the contraction mapping $V=\Gamma(V)$ by the method of *successive approximations*, where Γ is the *Bellman operator* defined in eq. (12) of Section 3.4.

$$V_{t+1} = \Gamma(V_t). \tag{26}$$

Since successive approximations converges at a geometric rate, with errors satisfying the upper bound in eq. (17), this method can converge at an unacceptably slow rate when the discount factor is close to 1. A more effective algorithm in such cases is *Newton's method* whose iterates are given by

$$V_{t+1} = V_t - [I - \Gamma'(V_t)]^{-1}[V_t - \Gamma(V_t)], \tag{27}$$

where Γ' is the *Gateaux* or *directional derivative* of Γ, that is, it is the linear operator given by

$$\Gamma'(V)(W) = \lim_{t \to 0} \frac{\Gamma(V + tW) - \Gamma(V)}{t}. \tag{28}$$

Newton's method converges *quadratically* independent of the value of the discount factor, as long as it is less than 1 (to guarantee the contraction property and the existence of a fixed point). In fact, Newton's method turns out to be equivalent to the method of *policy iteration* introduced by Howard (1960). Let δ be any stationary decision rule, that is, a candidate *policy*. Define the policy-specific conditional expectation operator E_δ by

$$E_\delta V(s) = \int_{s'} V(s')\, p(s' | s, \delta(s)). \tag{29}$$

Given a value function V_t, let δ_{t+1} be the decision rule implied by V_t, that is

$$\delta_{t+1}(s) = \underset{d \in D(s)}{\operatorname{argmax}} \left[u(s, d) + \rho\beta \int_{s'} V_t(s')\, p(s' | s, d) \right]. \tag{30}$$

It is not hard to see that the value of policy δ_{t+1} must be at least as high as V_t, and for this reason, eq. (30) is called the *policy improvement step* of the policy iteration algorithm. It is also not hard to show that

$$\Gamma'(V_t)(W)(s) = \rho\beta E_{\delta_{t+1}} W(s), \tag{31}$$

and this implies that the Newton iteration, eq. (27), is numerically identical to *policy iteration*

$$V_{t+1}(s) = [I - \rho\beta E_{\delta_{t+1}}]^{-1} u(s, \delta_{t+1}(s)), \tag{32}$$

where δ_{t+1} is given in eq. (30). Equation (32) is called the *policy valuation step* of the policy iteration algorithm since it calculates the value function implied by the policy δ_{t+1}. Note that, since E_δ is an expectation operator, it is linear and satisfies $\|E_\delta\| \leq 1$, and this implies that the operator $[I - \rho\beta E_\delta]$ is invertible and has the following geometric series expansion

$$[I - \rho\beta E_\delta]^{-1} = \sum_{j=0}^{\infty} [\rho\beta]^j E_\delta^j, \tag{33}$$

where E_δ^j is the *j* step ahead expectations operator. Thus, we see that

$$[I - \rho\beta E_\delta]^{-1} u(s, \delta(s)) = \sum_{j=0}^{\infty} [\rho\beta]^j E_\delta^j u(s, \delta(s)) = E\left\{\sum_{t=0}^{\infty} [\rho\beta]^t u(s_t, \delta(s_t)) \middle| s_0 = s\right\}, \tag{34}$$

so that value function V_t from the policy iteration (32) corresponds to the expected value implied by policy (decision rule) δ_t.

If there are an infinite number of states, the expectations operator E_δ is an infinite-dimensional linear operator, so it is not feasible to compute an exact solution to the policy iteration eq. (32). However if there are a finite number of states (or an infinite state space is discretized to a finite set of points, as per the discussion above), then E_δ is an $N \times N$ transition probability matrix, and policy iteration is feasible using ordinary matrix algebra, requiring at most $O(N^3)$ operations to solve a system of linear equations for V_t at each policy valuation step. Further, when there are a finite number of possible actions as well as states, there are only a finite number of possible policies $|D|^{|S|}$, where $|D|$ is the number of possible actions and $|S|$ is the number of states, and policy iteration can be shown to converge in a finite number of steps, since the method produces an improving sequences of decision rules, that is $V_t \leq V_{t+1}$. Thus, since there is an upper bound on the number of possible policies and policy iteration cannot cycle, it must converge in a finite number of steps. The number of steps is typically quite small, far fewer than the total number of possible policies. Santos and Rust (2004) show that the number of iterations can be bounded independent of the number of elements in the state space $|S|$. Thus, policy iteration is the method of choice for infinite horizon problems for which the discount factor is sufficiently close to 1. However, if the discount factor is far enough below 1, then successive approximations can be faster since policy iteration requires $O(N^3)$ operations per iteration whereas successive approximations requires $O(N^2)$ operations per iteration. At most $T(\varepsilon, \beta)$ successive approximation iterations are required to compute an ε-approximation to an infinite horizon Markovian decision problem with discount factor β, where $T(\varepsilon, \beta) = \log((1-\beta)\varepsilon) / \log(\beta)$. Roughly speaking, if $T(\varepsilon, \beta) < N$, then successive approximations are faster than policy iteration.

Successive approximations can be accelerated by a number of means discussed in Puterman (1994) and Rust (1996). *Multigrid algorithms* are also effective: these methods begin backward induction with a coarse grid with relatively few grid points *N*, and then as iterations proceed, the number of grid points is successively increased, leading to finer and finer grids as the backward induction starts to converge. Thus, computational time is not wasted early on in the backward induction iterations when the value function is far from the true solution. *Adaptive grid* methods are also highly effective in many problems: these methods can automatically detect regions in the state space where there is higher curvature in the value function, and in these regions more grid points are added in order to ensure that the value function is accurately approximated, whereas in regions where the value function is 'flatter' grid points can be removed, so as to direct computational resources to the regions of the state space where there is the highest payoff in terms of accurately approximating the value function. See Grüne and Semmler (2004) for more details and an interesting application of adaptive grid algorithms.

I conclude this section with a discussion of several other alternative approaches to solving stationary infinite horizon problems that can be extremely effective relative to 'discretization' methods when the number of grid points *N* required to obtain a good approximation becomes very large. Recall the curve-fitting approach discussed above in finite horizon SDPs: we approximate the value

function V by a parametric function as $V_\theta(s) \equiv f(s, \theta)$ for some flexible functional form f, where θ are treated as unknown parameters to be 'estimated'. For infinite horizon SDPs, our goal is to find parameter values $\hat{\theta}$ so that the implied value function satisfies the Bellman equation as well as possible. One approach to doing this, known as the *minimum residual method*, is a direct analogue of nonlinear least squares: if θ is a vector with K components, we select $N \geq K$ points in the state space (potentially at random) and find $\hat{\theta}$ that minimizes the squared deviations or *residuals* in the Bellman equation

$$\hat{\theta} = \underset{\theta \in R^K}{\operatorname{argmin}} \sum_{i=1}^{N} [\hat{\Gamma}(V_\theta)(s_i) - V_\theta(s_i)]^2, \tag{35}$$

where $\hat{\Gamma}$ denotes an approximation to the Bellman operator, where some numerical integration and optimization algorithm are used to approximate the true expectation operator and maximization in the Bellman equation (12). Another approach, called the *collocation method*, finds $\hat{\theta}$ by choosing K grid points in the state space and setting the residuals at those K points to zero:

$$\begin{aligned} V_{\hat{\theta}}(s_1) &= \hat{\Gamma}(V_{\hat{\theta}})(s_1) \\ V_{\hat{\theta}}(s_2) &= \hat{\Gamma}(V_{\hat{\theta}})(s_2) \\ &\cdots \\ V_{\hat{\theta}}(s_K) &= \hat{\Gamma}(V_{\hat{\theta}})(s_K). \end{aligned} \tag{36}$$

Another approach, called *parametric policy iteration*, carries out the policy iteration algorithm in eq. (32) above, but, instead of solving the linear system (32) for the value function V_t at each policy valuation step, they approximately solve this system by finding $\hat{\theta}_t$ that solves the regression problem

$$\theta_t = \underset{\theta \in R^K}{\operatorname{argmin}} \sum_{i=1}^{N} [V_{\hat{\theta}_t}(s_i) - u(s_i, \delta_t(s_i)) - \rho\beta E_{\delta_t} V_{\theta_t}(s_i)]^2. \tag{37}$$

Other than this, policy iteration proceeds exactly as discussed above. Note that, due to the linearity of the expectations operator, the regression problem above reduces to an ordinary linear regression problem when V_θ is approximated as a linear combination of basis functions as in (21) above.

There are variants of the minimum residual and collocation methods that involve parameterizing the *decision rule* rather than the value function. These methods are frequently used in problems where the control variable is continuous, and construct residuals from the *Euler equation* – a functional equation for the decision rule that can in certain classes of problems be derived from the first-order necessary condition for the optimal decision rule. These approaches then try to find $\hat{\theta}$ so that the Euler equation (as opposed to the Bellman equation) is approximately satisfied, in the sense of minimizing the squared residuals (minimum residual approach) or setting the residuals to zero at K specified points in the state space (collocation method). See Judd (1998) for further discussion of these methods and a discussion of strategies for choosing the grid points necessary to implement the collocation or minimum residual method.

There is a variety of other *iterative stochastic algorithms* for approximating solutions to dynamic programming problems that have been developed in the computer science and 'artificial intelligence' literatures on *reinforcement learning*. These methods include *Q-learning*, *temporal difference learning*, and *real time dynamic programming*. The general approach in all these methods is to iteratively update an estimate of the value function, and recursive versions of Monte Carlo integration methods are employed in order to avoid doing numerical integrations to calculate conditional expectations. Using methods adapted from the literature on *stochastic approximation*, it is possible to prove that these methods converge to the true value function in the limit

as the number of iterations tends to infinity. A key assumption underlying the convergence proofs is that there is sufficient stochastic noise to ensure that all possible decisions and decision nodes are visited ‘infinitely often’. The intuition of why such an assumption is necessary follows from the discussion in Section 3: suppose that at some state *s* an initial estimate of the value function for decision that is actually optimal happens to be so low that the action is deemed to be ‘nonoptimal’ relative to the initial estimate. If the agent does not ‘experiment’ sufficiently, and thus fails to choose suboptimal decisions infinitely often, the agent may fail to learn that the initial estimated value was an underestimate of the true value, and therefore the agent might never learn that the corresponding action really is optimal. There is a trade-off between learning and experimentation, of course. The literature on ‘multi-armed bandits’ (Gittins, 1979) shows that a fully rational Bayesian decision-maker will generally not find it optimal to experiment infinitely often. As a result such an agent can fail to discover actions that are optimal in an *ex post* sense. However, this does not contradict the fact that their behaviour is optimal in an *ex ante* sense: rather, it is a reflection that learning and experimentation is a costly activity, and thus it can be optimal to be incompletely informed, a result that has been known as early as Wald (1947). A nice feature of many of these methods, particularly the real time dynamic programming developed in Barto, Bradtke and Singh (1995), is that these methods can be used in ‘real time’, that is, we do not have to ‘precalculate’ the optimal decision rule in ‘offline’ mode. All these algorithms result in steady improvement in performance with experience. Methods similar to these have been used to produce highly effective strategies in extremely complicated problems. An example is IBM's ‘Deep Blue’ computer chess strategy, which has succeeded in beating the world's top human chess player, Garry Kasparov. However, the level of computation and repetition necessary to ‘train’ effective strategies is hugely time consuming, and it is not clear that any of these methods succeed in breaking the curse of dimensionality. For further details on this literature, see Bertsekas and Tsitsiklis (1996). Pakes (2001) applies these methods to approximate Markov perfect equilibria in games with many players. All types of stochastic algorithms have the disadvantage that the approximate solutions can be ‘jagged’ and there is always at least a small probability that the converged solution can be far from the true solution. However, they may be the only feasible option in many complex, high-dimensional problems where deterministic algorithms (for example, the Pakes and McGuire, 1994, algorithm for Markov perfect equilibrium) quickly become intractable due to the curse of dimensionality.

5 Empirical dynamic programming and the identification problem

The developments in numerical dynamic programming described in the previous section paved the way for a new, rapidly growing literature on empirical estimation and testing of SDPs and dynamic games. This literature began to take shape in the late 1970s, with contributions by Sargent (1978) on estimation of dynamic labour demand schedules in a linear quadratic framework, and Hansen and Singleton (1982), who developed a generalized method of moment estimation strategy for a class of continuous choice SDPs using the *Euler equation* as an *orthogonality condition*. About the same time, a number of papers appeared that provided different strategies for estimation and inference in *dynamic discrete choice models* including Gotz and McCall's (1980) model of retirements of air force pilots, Wolpin's (1984) model of a family's decision whether or not to have a child, Pakes's (1986) model of whether or not to renew a patent, and Rust's (1987) model of whether or not to replace a bus engine. Since 1987, hundreds of different empirical applications of dynamic programming models have been published. For surveys of this literature see Eckstein and Wolpin (1989), Rust (1994), and the very readable book by Adda and Cooper (2003) – which also provides accessible introductions to the theory and numerical methods for dynamic programming. The remainder of this section will provide a brief overview of estimation methods and a discussion of the identification problem.

In econometrics, the term *structural estimation* refers to a class of methods that tries to go beyond simply summarizing the behaviour of economic agents by attempting to infer their underlying *preferences* and *beliefs*. This is closely related to the distinction between the *reduced-form* of an economic model and the underlying *structure* that ‘generates’ it. (Structural estimation methods were first developed at the Cowles Commission at Yale University, starting with attempts to structurally estimate the linear simultaneous equations model, and models of investment by firms. Frisch, Haavelmo, Koopmans, Marschak, and Tinbergen were among the earliest contributors to this literature.) The reason why one would want to do structural estimation, which is typically far more difficult (for example, computationally intensive) than reduced-form estimation, is having knowledge of underlying structure enables us to conduct *hypothetical/counterfactual policy experiments*. Reduced-form estimation methods can be quite useful and yield significant insights into behaviour, but they are limited to summarizing behaviour under the status quo. However, they are inherently limited in their ability to forecast how individuals change their behaviour in response to various changes in the environment, or in *policies* (for example, tax rates, government benefits, regulations, laws, and so on) that *change the underlying structure* of agents’ decision problems. As long as it is possible to predict how different policies change the underlying structure, we can use dynamic programming to re-solve agents’ SDPs under the alternative structure, resulting in corresponding decision rules that represent predictions of how their behaviour (and welfare) will change in response to the policy change.

The rationale for structural estimation was recognized as early as Marschak (1953); however, his message appears to have been forgotten until the issue was revived in Lucas's (1976) critique of the limitations of reduced-form methods for policy evaluation. An alternative way to do policy evaluation is via *randomized experiments* in which subjects are randomly assigned to the *treatment group* (where the ‘treatment’ is some alternative policy of interest) and the *control group* (who continue with the policy under the status quo). By comparing the outcomes in the treatment and control groups, we can assess the behavioural and welfare impacts of the policy change. However, human experiments can be very time consuming and expensive to carry out, whereas ‘computational experiments’ using a structural model are very cheap and can be conducted extremely rapidly. The drawback of the structural approach, though, is the issue of *credibility* of the structural model. If the structural model is *misspecified*, it can generate incorrect forecasts of the impact of a policy change. There are numerous examples of how structural models can be used to make policy predictions: see Todd and Wolpin (2005) for an example that compares the prediction of a structural model with the results of a randomized experiment, where the structural model is estimated using subjects from the control group, and *out-of-sample predictions* are made to predict the behavioural response by subjects in the treatment group. They show that the structural model results in accurate predictions of how the treatment group subjects responded to the policy change.

I illustrate the main econometric methods for structural estimation of SDPs in the case of a stationary infinite horizon Markovian decision problem, although all the concepts extend in a straightforward fashion to finite horizon, nonstationary and non-Markovian problems. Estimation requires a specification of the *data generating process*. Assume we observe *N* agents, and we

observe agent i from time period $\underline{T}_i$ to $\bar{T}_i$ (or via appropriate re-indexing, from t=1,…, T_i). Assume observations of each individual are independently distributed realizations from the controlled process $\{s_t,d_t\}$. However, while we assume that we can observe the decisions made by each agent, it is more realistic to assume that we only observe a subset of the agent's state s_t. If we partition $s_t = (x_t, \varepsilon_t)$, assume that the econometrician observes x_t but not ε_t, so this latter component of the state vector constitutes an *unobserved state variable*. Then the reduced-form of the SDP is the decision rule δ

$$d = \delta(x, \varepsilon), \tag{38}$$

since the decision rule embodies all the behavioural content of the SDP model. The *structure* Λ consists of the objects $\Lambda = \{\beta, \rho, u(s, d), p(s'|s, d)\}$. Equation (10) specifies the mapping from the structure Λ into the reduced form, δ. The data-set consists of $\{(x_{i,t}, d_{i,t}), \ t = 1, \ldots, T_i, \ i = 1, \ldots, N\}$. The econometric problem is to infer the underlying structure Λ from our data on the observed states and decisions by a set of individuals. Although the decision rule is potentially a complicated nonlinear function of unobserved state variables in the reduced-form eq. (38), it is often possible to consistently estimate the decision rule under weak assumptions as $N\to\infty$, or as $T_i\to\infty$ if the data consists only of a single agent or a small number of agents i who are observed over long intervals. Thus, the decision rule δ can be treated as a *known function* for purposes of a theoretical analysis of identification. The *identification problem* is the qst, *under what conditions is the mapping from the underlying structure* Λ *to the reduced form* 1 *to* 1 (*that is invertible*)? If this mapping is 1 to 1, we say that the structure is *identified* since in principle it can be inverted to uniquely determine the underlying structure Λ. In practice, we construct an *estimator* $\hat{\Lambda}$ based on the available data and show that $\hat{\Lambda}$ converges to the true underlying structure Λ as $N\to\infty$ and/or $T_i\to\infty$ for each i.

Unfortunately, rather strong a priori assumptions on the form of agents' preferences and beliefs are required in order to guarantee identification of the structural model. Rust (1994) and Magnac and Thesmar (2002) have shown that an important subclass of SDPs, *discrete decision processes* (DDPs), are *nonparametrically unidentified*. That is, if we are unwilling to make any *parametric* functional form assumptions about preferences or beliefs, then in general there are infinitely many different structures Λ consistent with any reduced form δ. In more direct terms, there are many different ways to *rationalize* any observed pattern of behaviour as being 'optimal' for different configurations of preferences and beliefs. It is likely that these results extend to continuous choice problems, since it is possible to approximate a continuous decision process (CDP) by a sequence of DDPs with expanding numbers of elements in their choice sets. Further, for dynamic games, Ledyard (1986) has shown that *any* undominated strategy profile can be a Bayesian equilibrium for some set of preferences and beliefs. Thus, the hypothesis of optimality or equilibrium per se does not have testable empirical content: further a priori assumptions must be imposed in order for SDPs models to be identified and result in empirically testable restrictions on behaviour.

There are two main types of identifying assumptions that have been made in the literature to date: (*a*) parametric functional form assumptions on preferences $u(s,d)$ and components of agents' beliefs $p(s'|s, d)$ that involve unobserved state variables ε and (*b*) *rational expectations*. Rational expectations states that an agent's *subjective beliefs* $p(s'|s, d)$ coincide with *objective probabilities* that can be estimated from data. Of course, this restriction is useful only for those components of s, x, that the econometrician can actually observe. In addition, there are other more general *functional restrictions* that can be imposed to help identify the model. One example is monotonicity and shape restrictions on preferences (for example, concavity and monotonicity of the utility function), and another example is independence or *conditional independence* assumptions about variables entering agents' beliefs. I will provide specific examples below; however, it should be immediately clear why these additional assumptions are necessary.

For example, consider the two parameters ρ (the agent's subjective survival probability) and β (the agent's subjective discount factor). We have seen in Section 3 that *only the product of* ρ *and* β *enter the SDP model, and not* ρ *and* β *separately*. Thus, at most the product $\rho\beta$ can be identified, but without further assumptions it is impossible to separately identify the subjective survival probability ρ from the subjective discount factor β since both affect an agent's behaviour in a symmetrical fashion. However, we can separately identify ρ and β if we assume that an individual has *rational survival expectations*, that is, that their subjective survival probability ρ coincides with the 'objective' survival probability. Then we can estimate ρ 'outside' the SDP model, using data on the lifetime distributions of similar types of agents, and then β can be identified if other restrictions are imposed to guarantee that the product $\rho\beta$ is identified. However, it can be very difficult to make precise inferences about agents' discount factors in many problems, and it is easy to think of models where there is heterogeneity in survival probabilities and discount factors, and unobserved variables affecting one's beliefs about them (for example, family characteristics such as a predisposition for cancer, and so on, that are observed by an agent but not by the econometrician) where identification is problematic.

There are two main approaches for conducting inference in SDPs: (*a*) maximum likelihood and (*b*) 'simulation estimation'. The latter category includes a variety of similar methods such as *indirect inference* (Gourieroux and Monfort, 1997), *simulated method of moments* (McFadden, 1989; Gallant and Tauchen, 1996), *simulated maximum likelihood and method of simulated scores* (see simulation-based estimation), and *simulated minimum distance* (Hall and Rust, 2006). To simplify the discussion I will define these initially for single agent SDPs and at the end discuss how these concepts naturally extend to dynamic games. I will illustrate maximum likelihood and show how a likelihood can be derived for a class of DDPs; however, for CDPs, it is typically much more difficult to derive a likelihood function, especially when there are issues of *censoring*, or problems involving mixed discrete and continuous choice. In such cases simulation estimation is often the only feasible way to do inference.

For discrete decision processes, assume that the utility function has the following parametric, *additively separable* representation

$$u(x, \varepsilon, d) = u(x, d, \theta_1) + \varepsilon(d) \text{ (AS)}. \tag{39}$$

where $\varepsilon = \{\varepsilon(d) | d \in D(x)\}$, and $\varepsilon(d)$ is interpreted as an unobserved component of utility associated with choice of alternative $d \in D(x)$. Further, suppose that the transition density $p(x', \varepsilon' | x, \varepsilon, d)$ satisfies the following *conditional independence assumption*

$$p(x', \varepsilon' | x, \varepsilon, d) = p(x' | x, d, \theta_2) q(\varepsilon', \theta_3) \text{ (CI)}. \tag{40}$$

The CI assumption implies that $\{\varepsilon_t\}$ is an *IID* 'noise' process that is independent of $\{x_t, d_t\}$. Thus all of the serially correlated dynamics in the state variables are captured by the observed component of the state vector x_t. If, in addition, $q(\varepsilon_t, \theta_3)$ is a distribution with unbounded support with finite absolute first moments, one can show that the following *conditional choice probabilities* exist

$$P(d | x, \theta) = \int I_\varepsilon \left\{ d = \delta(x, \varepsilon, \theta) \right\} q(\varepsilon) d\varepsilon, \tag{41}$$

where $\theta = (\rho, \beta, \theta_1, \theta_2, \theta_3)$ constitute the vector of unknown parameters to be estimated. (Identification of fully parametric models is a 'generic' property, that is, if there are two different parameters θ that produce the same conditional choice probability $P(d | x, \theta)$ for all x and $d \in D(x)$ – and thus led to the same limiting expected log-likelihood – small perturbations in the parameterization will 'almost always' result in a nearby model for which θ is uniquely identified.) In general, the parametric functional form assumptions, combined with the assumption of rational expectations and the AS and CI assumptions, are sufficient to identify the unknown parameter vector θ^*. θ^* can be estimated by maximum likelihood, using the *full information* likelihood function L_f given by

$$\mathcal{L}_f(\theta | \{x_{i,t}, d_{i,t}\}, t = 1, \ldots, T_i, \ i = 1, \ldots, N) = \prod_{i=1}^{N} \prod_{t=2}^{T_i} P(d_{i,t} | x_{i,t}, \theta) \times p(x_{i,t} | x_{i,t-1}, d_{i,t-1}, \theta_2). \tag{42}$$

A particularly tractable special case is where $q(\varepsilon, \theta_3)$ has a *multivariate extreme value distribution* where θ_3 is a common scale parameter (linearly related to the standard deviation) for each variable in this distribution (see McFadden, Daniel; logit models of individual choice for the exact formula for this density). This specification leads to a dynamic generalization of the *multinomial logit model*

$$P(d | x, \theta) = \frac{\exp\{v(x, d, \theta) / \theta_3\}}{\sum_{d' \in D(x)} \exp\left\{v(x, d', \theta) / \theta_3\right\}},$$

(43)

where $v(x, d, \theta)$ is the expected, discounted utility from taking action *d* in observed state *x* given by the unique fixed point to the following *smoothed Bellman equation*

$$v(x, d, \theta) = u(x, d, \theta_1) + \rho\beta\int_{x'} \theta_3 \log\left(\sum_{d' \in D(x')} \exp\left\{v(x', d', \theta) / \theta_3\right\}\right) \times p(x'|x, d, \theta_2)\, dx'. \tag{44}$$

Define Γ_θ by

$$\Gamma_\theta(W)(x, d) = u(x, d, \theta_1) + \rho\beta\int_{x'} \theta_3 \log\left(\sum_{d' \in D(x')} \exp\left\{\exp\left\{W(x', d', \theta) / \theta_3\right\}\right\}\right) \times p(x'|x, d, \theta_2)\, dx'. \tag{45}$$

It is not hard to show that under weak assumptions Γ_θ is a contraction mapping, so that $v(x, d, \theta)$ exists and is unique. Maximum likelihood estimation can be carried out using a nested *fixed point* maximum likelihood algorithm consisting of an ‘outer’ optimization algorithm to search for a value of θ that maximizes $\mathcal{L}_f(\theta)$, and an ‘inner’ fixed point algorithm that computes $v_\theta = \Gamma_\theta(v_\theta)$ each time the outer optimization algorithm generates a new trial guess for θ. The implicit function theorem guarantees that v_θ is a smooth function of θ. See Aguirregabiria and Mira (2004) for an ingenious alternative that ‘swaps’ the order of the inner and outer algorithms of the nested fixed-point algorithm resulting in significant computational speedups. See also Rust (1988) for further details on the nested fixed-point algorithm and the properties of the maximum likelihood estimator, and Rust (1994) for a survey of alternative less efficient but computationally simpler estimation strategies.

As noted above, econometric methods for CDPs, that is, problems where the decision variable is continuous (such as firm investment decisions, price settings, or consumption/savings decisions) are harder, since there is no tractable, general specification for the way unobservable state variables to enter the decision rule that result in a *nondegenerate* likelihood function (that is, where the likelihood $\mathcal{L}(\theta)$ is non-zero for any data-set and any value of θ). For this reason, maximum likelihood estimation of CDPs is rare, outside certain special subclasses, such at linear quadratic CDPs (Hansen and Sargent, 1980; Sargent, 1981). However, simulation-based methods of inference can be used in a huge variety of situations where a likelihood is difficult or impossible to derive. These methods have a great deal of flexibility, a high degree of generality, and often permit substantial computational savings. In particular, generalizations of McFadden's (1989) *method of simulated moments* (MSM) have enabled estimation of a wide range of CDPs. The MSM estimator minimizes a quadratic form between a set of moments constructed from the data, h_N and a vector of *simulated moments* $h_{N,S}(\theta)$, that is

$$h_N = \frac{1}{N}\sum_{i=1}^{N} h(\{x_{it}, d_{it}\})$$

$$h_{N,S}(\theta) = \frac{1}{S}\sum_{j=1}^{S}\frac{1}{N}\sum_{i=1}^{N} h\left(\left\{\tilde{x}_{it}^{j}(\theta),\ \tilde{d}_{it}^{j}(\theta)\right\}\right) \tag{46}$$

where h is a vector of $l \geq K$ 'moments' (that is, functionals of the data that the econometrician is trying to 'match'), where K is the dimension of θ, $\{x_{it}, d_{it}\}$ are the data, and $\left\{\tilde{x}_{it}^{j}(\theta),\ \tilde{d}_{it}^{j}(\theta)\right\}$, j=1, …, S are S *IID* realizations of the controlled process. The estimate $\hat{\theta}$ is given by

$$\hat{\theta} = \underset{\theta \in R^{K}}{\operatorname{argmin}}\ [h_N - h_{N,S}(\theta)]'\, W_N\, [h_N - h_{N,S}(\theta)], \tag{47}$$

where W_N is a $J\times J$ positive–definite weighting matrix. The most efficient choice for W_N is $W_N = [\hat{\Omega}_N]^{-1}$ where $\hat{\Omega}_N$ is the variance-covariance matrix formed from the vector of sample moments h_N. Simulation estimators require a nested fixed-point algorithm since each time the outer minimization algorithm tries a new trial value for θ, the inner fixed point problem must be called to solve the CDP problem, using the optimal decision rule $d_{it}^{j}(\theta) = \delta(x_{it}^{j}, \varepsilon_{it}^{j}, \theta)$ to generate the simulated decisions, and the transition density $p(x_{i,t+1}^{j}, \varepsilon_{i,t+1}^{j} | x_{i,t}^{j}, \varepsilon_{i,t}^{j}, d_{i,t}^{j}, \theta_2)$ to generate j=1,…, S *IID* realizations for a simulated panel each potential value of θ. (It is important to simulate using 'common random numbers' that remain fixed as θ varies over the course of the estimation, in order to satisfy the *stochastic equicontinuity conditions* necessary to establish consistency and asymptotic normality of the simulation estimator.)

Simulation methods are extremely flexible for dealing with a number of data issues such as attrition, missing data, censoring and so forth. The idea is that, if we are willing to build a stochastic model of the data 'problem', we can account for it in the process of simulating the behavioural model. For example, Hall and Rust (2006) develop a dynamic model of commodity price speculation in the steel market. An object of interest is to estimate the stochastic process governing wholesale steel prices; however, there is no public commodity market where steel is traded and prices are recorded on a daily basis. Instead, Hall and Rust observe only the actual wholesale prices of a particular steel trader, who records wholesale prices only on the days he actually buys steel in the wholesale market. Since the speculator makes money by 'buying low and selling high', the set of observed wholesale prices are *endogenously sampled*, and failure to account for this can lead to incorrect inferences about wholesale prices – a dynamic analogue of *sample selection bias*. However, in a simulation model it is easy to censor the simulated data in the same way it is censored in the actual data, that is, by discarding simulated wholesale prices on days where no simulated purchases are made. Hall and Rust show that even though moments based on the observed (censored) data are 'biased' estimates, the simulated moments are biased in exactly the same fashion, so minimizing the distance between actual and simulated biased moments nevertheless results in consistent and asymptotically normal estimates of the parameters of the wholesale price process and other parameters entering the speculator's objective function.

Simulation methods have also enabled the use of Bayesian methods, resulting in methods of inference that do not require asymptotic approximations, although they generally use Markov chain Monte Carlo methods to generate simulated draws from a distribution that approximates the exact finite sample posterior distribution for the parameters of interest (see for example, Lancaster, 1997; Imai, Jain and Ching, 2005; Nourets, 2006).

The most recent literature has extended the methods for estimation of single-agent SDPs to multi-agent dynamic games. For example, Rust (1994) described applications of dynamic discrete choice models to multiple-agent *discrete dynamic games*. The unobserved state variables ε_t entering any particular agent's payoff function are assumed to be unobserved both by the econometrician and by the other players in the game. The *Bayesian–Nash equilibria* of this game can be represented as a vector of conditional choice probabilities $(P_1(d_1|x), \ldots, P_n(d_n|x))$, one for each player, where $P_i(d_i|x)$ represents the econometrician's and the other players' beliefs about the probability player i will take action d_i, 'integrating out' over the unobservable states variable $\varepsilon_{i,t}$ affecting player i's decision at time t similar to eq. (41) for single-agent problems. If one adapts the numerical methods for Markov-perfect equilibrium described in Section 4, it is possible to compute Bayesian–Nash equilibria of discrete dynamic games using nested fixed-point algorithms. While it is relatively straightforward to write down the likelihood function for the game, actual estimation via a straightforward application of full information maximum likelihood is extremely computationally demanding since it requires a *doubly nested fixed point algorithm* (that is, an 'outer' algorithm to search over θ to maximize the likelihood, and then an inner algorithm to solve the dynamic game for each value of θ, but this inner algorithm is itself a nested fixed-point algorithm). Alternative, less computationally demanding estimation methods have been proposed by Aguirregabiria and Mira (2007), Bajari and Hong (2006), Bajari, Benkard and Levin (2007), and Pesendorfer and Schmidt-Dengler (2003). This research is at the current frontier of development in numerical and empirical applications of dynamic programming.

Besides econometric methods, which are applied for structural estimation for actual agents in their 'natural' settings, an alternative approach is to try to make inferences about agents' preferences and beliefs (and even their 'mode of reasoning') for artificial SDPs in a laboratory setting. The advantage of a laboratory experiment is *experimental control over preferences and beliefs*. The

ability to control these aspects of decision-making can enable much tighter tests of theories of decision-making. For example, Binmore et al. (2002) structured a laboratory experiment to determine whether individuals do backward induction in one- and two-stage alternating offer games, and 'find systematic violations of backward induction that cannot be explained by payoff-interdependent preferences' (2002, p. 49).

6 Comments

There has been tremendous growth in research related to dynamic programming since the 1940s. The method has evolved into the main tool for solving sequential decision problems, and research related to dynamic programming has led to fundamental advances in theory, numerical methods and econometrics. As we have seen, while dynamic programming embodies the notion of rational decision-making under uncertainty, there is mixed evidence as to whether it provides a good literal description of how human beings actually behave in comparable situations. Although human reasoning and decision-making is undoubtedly both more complex and more 'frail' and subject to foibles and limitations than the idealized notion of 'full rationality' that dynamic programming embodies, the discussion of the identification problem shows that, if we are given sufficient flexibility about how to model individual preferences and beliefs, there exist SDPs whose decision rules provide arbitrarily good approximations to individual behaviour.

Thus, dynamic programming can be seen as a useful 'first approximation' to human decision-making, but it will undoubtedly be superseded by more descriptively accurate psychological models. Indeed, in the future one can imagine behavioural models that are not derived from some a priori axiomatization of preferences, but will result from empirical research that will ultimately deduce human behaviour from yet even deeper 'structure', that is the very underlying neuroanatomy of the human brain.

Even if dynamic programming is unlikely to be a descriptively accurate model of *human* decision-making, it will probably still remain highly relevant for the foreseeable future as the embodiment of *rational* decision-making. There are well-defined problems, for example, profit maximization or cost minimization, where there is agreement on the objective function to be maximized or minimized, and where there will be a demand for dynamic programming methods to find the optimal profit- or cost-minimizing strategies. There are many examples of this in the operations research literature. Practical applications include optimal inventory management (Hall and Rust, 2006) and optimal harvesting of timber (Paarsch and Rust, 2007).

Some observers such as Kurzweil (2005) predict that in the not too distant future (for example, approximately 2050) a *singularity* will occur, 'during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed' (2005, p. 7). The singularity is a complex of accelerating improvements in computer hardware and software, and a merger of machine- and biological-based intelligence that will blur the distinction between 'artificial intelligence' and human intelligence, that will overcome many of current limitations of the human brain and human reasoning: 'By the end of this century, the nonbiological portion of our intelligence will be trillions and trillions of times more powerful than unaided human intelligence' (2005, p. 9). Dynamic programming will undoubtedly continue to be a critical tool in this brave new world.

Whether this prognosis will ever come to pass, or come to pass as soon as Kurzweil forecast, is debatable; but it does suggest that there will be continued interest in and research on dynamic programming. However, the fact that reasonably broad classes of dynamic programming problems are subject to a curse of dimensionality suggests that it may be too optimistic to think that human rationality will soon be superseded by 'artificial rationality'. While there are many complicated problems that we would like to solve by dynamic programming in order to understand what 'fully rational' behaviour actually looks like in specific situations, the curse of dimensionality still limits us to very simple 'toy models' that only very partially and simplistically capture the myriad of details and complexities we face in the real world. Although we now have a number of examples where artificial intelligence based on principles from dynamic programming outstrips human intelligence, for example computerized chess, all these cases are for very specific problems in very narrow domains. I believe that it will be a long time before technological progress in computation and algorithms produce truly general-purpose 'intelligent behaviour' that can compete successfully with human intelligence in widely varying domains and in the immensely complicated situations that we operate in every day. Despite all our psychological frailties and limitations, there is an important unanswered question of 'how do we do it?', and more research is required to determine if human behaviour is simply suboptimal, or whether the human brain uses some powerful implicit 'algorithm' to circumvent the curse of dimensionality that digital computers appear to be subject to for solving problems such as SDPs by dynamic programming. For a provocative theory that deep principles of quantum mechanics can enable human intelligence to transcend computational limitations of digital computers, see Penrose (1989).

See Also

- Bellman equation
- game theory
- logit models of individual choice
- McFadden, Daniel
- recursive competitive equilibrium
- recursive preferences
- sequential analysis
- simulation-based estimation

This article has benefited from helpful feedback from Kenneth Arrow, Daniel Benjamin, Larry Blume, Moshe Buchinsky, Larry Epstein, Chris Phelan and Arthur F. Veinott, Jr.

Bibliography

Adda, J. and Cooper, R. 2003. *Dynamic Economics Quantitative Methods and Applications*. Cambridge, MA: MIT Press.

Aguirregabiria, V. and Mira, P. 2004. Swapping the nested fixed point algorithm: a class of estimators for discrete Markov decision models. *Econometrica* 70, 1519–43.

Aguirregabiria, V. and Mira, P. 2007. Sequential estimation of dynamic discrete games. *Econometrica* 75, 1–53.

Arrow, K.J., Blackwell, D. and Girshik, M.A. 1949. Bayes and minimax solutions of sequential decision problems. *Econometrica* 17, 213–44.

Bajari, P. and Hong, H. 2006. Semiparametric estimation of a dynamic game of incomplete information. Technical Working Paper No. 320. Cambridge, MA: NBER.

Bajari, P., Benkard, L. and Levin, J. 2007. Estimating dynamic models of imperfect competition. *Econometrica* 75, 1331–70.

Barto, A.G., Bradtke, S.J. and Singh, S.P. 1995. Learning to act using real-time dynamic programming. *Artificial Intelligence* 72, 81–138.

Bellman, R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Bellman, R. and Dreyfus, S. 1962. *Applied Dynamic Programming*. Princeton, NJ: Princeton University Press.

Bellman, R. 1984. *Eye of the Hurricane*. Singapore: World Scientific.

Bertsekas, D.P. 1995. *Dynamic Programming and Optimal Control*, vols 1 and 2. Belmont, MA: Athena Scientific.

Bertsekas, D.P. and Tsitsiklis, J. 1996. *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.

Bhattacharya, R.N. and Majumdar, M. 1989. Controlled semi-Markov models – the discounted case. *Journal of Statistical Planning and Inference* 21, 365–81.

Binmore, K., McCarthy, J., Ponti, G., Samuelson, L. and Shaked, A. 2002. A backward induction experiment. *Journal of Economic Theory* 104, 48–88.

Blackwell, D. 1962. Discrete dynamic programming. *Annals of Mathematical Statistics* 33, 719–26.

Blackwell, D. 1965a. Positive dynamic programming. *Proceedings of the 5th Berkeley Symposium* 3, 415–28.

Blackwell, D. 1965b. Discounted dynamic programming. *Annals of Mathematical Statistics* 36, 226–35.

Cayley, A. 1875. Mathematical qsts and their solutions. Problem No. 4528. *Educational Times* 27, 237.

Chow, C.S. and Tsitsiklis, J.N. 1989. The complexity of dynamic programming. *Journal of Complexity* 5, 466–88.

Denardo, E. 1967. Contraction mappings underlying the theory of dynamic programming. *SIAM Review* 9, 165–77.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. 1952. The inventory problem: I. Case of known distributions of demand. *Econometrica* 20, 187–222.

Eckstein, Z. and Wolpin, K.I. 1989. The specification and estimation of dynamic stochastic discrete choice models: a survey. *Journal of Human Resources* 24, 562–98.

Gallant, A.R. and Tauchen, G.E. 1996. Which moments to match? *Econometric Theory* 12, 657–81.

Gihman, I.I. and Skorohod, A.V. 1979. *Controlled Stochastic Processes*. New York: Springer.

Gittins, J.C. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society* B 41, 148–64.

Gotz, G.A. and McCall, J.J. 1980. Estimation in sequential decision-making models: a methodological note. *Economics Letters* 6, 131–6.

Gourieroux, C. and Monfort, A. 1997. *Simulation-Based Methods of Inference*. Oxford: Oxford University Press.

Grüne, L. and Semmler, W. 2004. Using dynamic programming with adaptive grid scheme for optimal control problems in economics. *Journal of Economic Dynamics and Control* 28, 2427–56.

Hall, G. and Rust, J. 2006. Econometric methods for endogenously sampled time series: the case of commodity price speculation in the steel market. Manuscript, Yale University.

Hansen, L.P. and Sargent, T.J. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2, 7–46.

Hansen, L.P. and Singleton, K. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–81.

Howard, R.A. 1960. *Dynamic Programming and Markov Processes*. New York: Wiley.

Imai, S., Jain, N. and Ching, A. 2005. Bayesian estimation of dynamic discrete choice models. Manuscript, University of Illinois.

Judd, K. 1998. *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Keane, M. and Wolpin, K.I. 1994. The solution and estimation of discrete choice dynamic programming models by simulation: Monte Carlo evidence. *Review of Economics and Statistics* 76, 648–72.

Kurzweil, R. 2005. *The Singularity is Near when Humans Transcend Biology*. New York: Viking Press.

Kushner, H.J. 1990. Numerical methods for stochastic control problems in continuous time. *SIAM Journal on Control and Optimization* 28, 999–1048.

Lancaster, A. 1997. Exact structural inference in optimal job search models. *Journal of Business Economics and Statistics* 15, 165–79.

Ledyard, J. 1986. The scope of the hypothesis of Bayesian equilibrium. *Journal of Economic Theory* 39, 59–82.

Lucas, R.E. Jr. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labour Markets*, ed. K. Brunner and A.K. Meltzer. Carnegie-Rochester Conference on Public Policy. Amsterdam: North-Holland.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1426–45.

Luenberger, D.G. 1969. *Optimization by Vector Space Methods*. New York: Wiley.

Magnac, T. and Thesmar, D. 2002. Identifying dynamic discrete decision processes. *Econometrica* 70, 801–16.

Marschak, T. 1953. Economic measurements for policy and prediction. In *Studies in Econometric Method*, ed. W.C. Hood and T.J. Koopmans. New York: Wiley.

Massé, P. 1945. *Application des probabilités en chaîne á l'hydrologie statistique et au jeu des réservoirs*. Report to the Statistical Society of Paris. Paris: Berger-Levrault.

Massé, P. 1946. *Les réserves et la régulation de l'avenir*. Paris: Hermann.

McFadden, D. 1989. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57, 995–1026.

Nemirovsky, A.S. and Yudin, D.B. 1983. *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley.

Nourets, A. 2006. Inference in dynamic discrete choice models with serially correlated unobserved state variables. Manuscript, University of Iowa.

Paarsch, H.J. and Rust, J. 2007. Stochastic dynamic programming in space: an application to British Columbia forestry. Working paper.

Pakes, A. 1986. Patents as options: some estimates of the values of holding European patent stocks. *Econometrica* 54, 755–84.

Pakes, A. 2001. Stochastic algorithms, symmetric Markov perfect equilibria and the ‘curse’ of dimensionality. *Econometrica* 69, 1261–81.

Pakes, A. and McGuire, P. 1994. Computing Markov perfect Nash equilibrium: numerical implications of a dynamic differentiated product model. *RAND Journal of Economics* 25, 555–89.

Penrose, R. 1989. *The Emperor's New Mind*. New York: Penguin.

Pesendorfer, M. and Schmidt-Dengler, P. 2003. Identification and estimation of dynamic games. Manuscript, University College London.

Pollard, D. 1989. Asymptotics via empirical processes. *Statistical Science* 4, 341–86.

Puterman, M.L. 1994. *Markovian Decision Problems*. New York: Wiley.

Rust, J. 1985. Stationary equilibrium in a market for durable goods. *Econometrica* 53, 783–805.

Rust, J. 1987. Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher. *Econometrica* 55, 999–1033.

Rust, J. 1988. Maximum likelihood estimation of discrete control processes. *SIAM Journal on Control and Optimization* 26, 1006–24.

Rust, J. 1994. Structural estimation of Markov decision processes. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Rust, J. 1996. Numerical dynamic programming in economics. In *Handbook of Computational Economics*, ed. H. Amman, D. Kendrick and J. Rust. Amsterdam: North-Holland.

Rust, J. 1997. Using randomization to break the curse of dimensionality. *Econometrica* 65, 487–516.

Rust, J. and Hall, G.J. 2007. The (S, s) rule is an optimal trading strategy in a class of commodity price speculation problems. *Economic Theory* 30, 515–38.

Rust, J. and Phelan, C. 1997. How Social Security and Medicare affect retirement behavior in a world with incomplete markets. *Econometrica* 65, 781–832.

Rust, J., Traub, J.F. and Wo•niakowski, H. 2002. Is there a curse of dimensionality for contraction fixed points in the worst case? *Econometrica* 70, 285–329.

Santos, M. and Rust, J. 2004. Convergence properties of policy iteration. *SIAM Journal on Control and Optimization* 42, 2094–115.

Sargent, T.J. 1978. Estimation of dynamic labor demand schedules under rational expectations. *Journal of Political Economy* 86, 1009–44.

Sargent, T.J. 1981. Interpreting economic time series. *Journal of Political Economy* 89, 213–48.

Selten, R. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4, 25–55.

Tauchen, G. and Hussey, R. 1991. Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models. *Econometrica* 59, 371–96.

Todd, P. and Wolpin, K.I. 2005. Ex ante evaluation of social programs. Manuscript, University of Pennsylvania.

Traub, J.F. and Werschulz, A.G. 1998. *Complexity and Information*. Cambridge: Cambridge University Press.

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press. 3rd edn, 1953.

Wald, A. 1947. Foundations of a general theory of sequential decision functions. *Econometrica* 15, 279–313.

Wald, A. 1947. *Sequential Analysis*. New York: Dover Publications.

Wald, A. and Wolfowitz, J. 1948. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* 19, 326–39.

Wolpin, K. 1984. An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy* 92, 852–74.

How to cite this article

Rust, John. "dynamic programming." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000246> doi:10.1057/9780230226203.0420

The New Palgrave Dictionary of Economics Online

Easterlin hypothesis

Diane J. Macunovich and Richard A. Easterlin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economic and social fortunes of a birth cohort tend to vary as a function of that cohort's relative size, approximated by the crude birth rate surrounding the cohort's birth. Effects have been observed on young men's earnings and unemployment rates, college enrolment rates, marriage and divorce, fertility, crime, and suicide rates. These effects have been found to be asymmetrical about the peak of a baby boom, and the original hypothesis has been extended to suggest a wide range of effects on the economy as a whole, from GDP growth rate, through interest rates and stock market performance, to measures of productivity.

Keywords

aggregate demand; cohort size effects; crowding; demographic transition; Easterlin hypothesis; female labour force participation; fertility; inflation; interest rates; life cycle models; marriage and divorce; saving rates; relative income; relative cohort size; productivity; GDP growth; unemployment rates; college enrolment rates; material aspirations; preferences; crime rates; suicide rates

Article

The Easterlin, or ‘relative cohort size’, hypothesis as originally formulated posits that, other things constant, the economic and social fortunes of a cohort (those born in a given year) tend to vary as a function of its relative size, approximated by the crude birth rate surrounding the cohort's birth (Easterlin, 1987). This hypothesis has since been extended to suggest a wider range of effects on the economy as a whole (Macunovich, 2002).

Although cohort size effects were originally expected to be symmetrical around the peak of the baby boom, which in the United States entered the labour market around 1980, it is now thought that they are tempered by aggregate demand effects and by feedback effects from adjustments made by young adults on the ‘leading edge’ of a baby boom. As a result, cohorts – and the economy generally – on the ‘leading

edge of a baby boom fare much better than those on the ‘trailing edge’, when all else is equal.
The ultimate effects of changing relative cohort size are hypothesized to fall into these three categories:

1. 1. Direct or first-order effects of relative cohort size on male relative income (the earnings of young men relative to their aspirations); male unemployment and hours worked; men's and women's college wage premium (the extra earnings of a college graduate relative to those of a secondary school graduate); and levels of income inequality generally.
2. 2. Second-order effects operating through male relative income, especially the demographic adjustments people make in response to changing relative income, such as changes in women's labour force participation and their occupational choices; men's and women's college enrolment rates; marriage and divorce; fertility; crime, drug use, and suicide rates; out-of-wedlock childbearing and the incidence of female-headed families; and living arrangements.
3. 3. Third-order effects on the economy of changing relative cohort size and the resulting demographic adjustments, such as changes in average wage growth; the overall demand for goods and services in the economy and hence the growth rate of the economy; inflation, interest rates, and savings rates; stock market performance; industrial structure; measures of gross domestic product (GDP); and productivity measures.

The three categories of effect are discussed first in this article, followed by a consideration of feedback effects and a discussion of empirical evidence.

First-order effects

The linkage between higher birth rates and adverse social and economic effects arises from ‘crowding mechanisms’ operating within three major social institutions, the family, school and the labour market. Within the family, a sustained upsurge in the birth rate is likely to entail an increase in the average number of siblings, higher average birth order, and a shorter average birth interval, and there is a substantial literature in psychology, sociology and economics linking child development negatively to one or more of these magnitudes (Ernst and Angst, 1983; Heer, 1985). The negative effects that have been investigated range over a wide variety of phenomena. With regard to mental health, for example, there is evidence that problem behaviours such as fighting, breaking rules, and delinquency are associated with increased family size. Adverse effects on morbidity and mortality of children have been found to be associated with increased family size and shorter birth spacing. A negative association between IQ and number of siblings has been found in a number of studies, and, with IQ controlled for, between educational attainment and family size. The principal mechanism underlying such developments is likely to be the dilution of parental time and energy per child and family economic resources per child, associated with increased family size.
The family mechanisms just discussed imply that, on average, a larger cohort is likely to perform less well in school. But even in the absence of any adverse effects within the family, a large cohort is likely to experience crowding in schools, which reduces average educational performance (Freeman, 1976). At any given time the human and physical capital stock comprising the school system tends to be either fixed in amount or to expand at a fairly constant rate, so that a surge in entrants into the school system

tends to be accompanied by a reduction in physical facilities and teachers per student. In the United States, school planning decisions are divided among numerous local governments and private institutions, and expansion has tended to occur in reaction to, rather than in anticipation of, a large cohort's entry. Moreover, even when expansion occurs it is usually not accompanied by maintenance of curriculum standards, partly because of the diminishing pool of qualified teachers available to supply the needs of educational expansion.

The experience of a large cohort both in the family and in school is likely, in turn, to leave the cohort less well prepared, on reaching adulthood, for success in the labour market. But even if there were no prior effects, the entry of a large proportion of young and relatively inexperienced workers into the labour market creates a new set of crowding phenomena, because the expansion of complementary factor inputs is unlikely to be commensurate with that of the youth labour force. Additions to physical capital stock tend to be dominated by considerations other than the relative supply of younger workers, and the growth in older, experienced, workers is largely governed by prior demographic conditions. Growth in the relative supply of younger workers results, in consequence, in a deterioration of their relative wage rates, unemployment conditions and upward job mobility (Welch, 1979). The adverse effects of labour market crowding tend to reinforce those of crowding within the school and family. For example, the deterioration in relative wage rates of the young translates into lower returns to education and consequent adverse impact on school drop-out rates and college enrolment (Freeman, 1976). Also, problems encountered in finding a good job may reinforce feelings of inadequacy or frustration already stirred up by some prior experiences at home or in school, and lead to lower labour force participation among young men.

Second-order effects

The relative economic standing of successive generations at a given point in time may be altered systematically by fluctuations in relative cohort size. If parents’ living levels play an important role in setting their children's material aspirations, as socialization theory leads one to believe, then an increase in the shortfall of children's wage rates relative to parents will cause the children to feel relatively deprived and under greater pressure to keep up. The importance of relative status influences of this type in affecting attitudes or behaviour has been widely recognized in social science theory (Duesenberry, 1949).

Confronted with the prospect of a deterioration in its living level relative to that of its parents, a large young adult cohort may make a number of adaptations in an attempt to preserve its comparative standing. Foremost among these are changes in behaviour related to family formation and family life (Macunovich and Easterlin, 1990; McNown and Rajbhandary, 2003). To avoid the financial pressures associated with family responsibilities, marriage may be deferred. If marriage occurs, wives are more likely to work and to put off childbearing. If a wife bears children, she is more likely to couple labour force participation with childrearing, and to have a smaller number of children more widely spaced (Macunovich, 2002; Jeon and Shields 2005).

The process of demographic adjustment to changing relative income can best be thought of in terms of *ex ante* and *ex post* income; that is, the disposable per capita income of individuals prior to and then following the adjustments. Analyses of baby boom cohorts in the United States have found that a

cohort's male relative income – individual earning potential of baby boomers relative to that of their parents – was significantly lower than the individual earning potential of *pre-boom* cohorts relative to *their* parents. But after making the type of demographic adjustments indicated above, the boomers managed to bring their per capita disposable income on a par with that of their parents (Easterlin, Macdonald and Manucnovich, 1990).

Other reactions to the psychological stresses induced by large cohort size may be viewed as socially dysfunctional. Feelings of inadequacy and frustration, for example, may lead to disproportionate consumption of alcohol and drugs, to mental depression, and, at the extreme, to a higher rate of suicide (Pampel, 2001; Stockard and O'Brien, 2002). Feelings of bitterness, disappointment and rage may induce a higher incidence of crime (O'Brien, Stockard and Isaacson, 1999). Within marriage, the stresses of conflicting work and motherhood roles for women, and feelings of inadequacy as a breadwinner for men, are likely to result in a higher incidence of divorce (Macunovich, 2002). In the political sphere, the disaffection felt by a large cohort because of its lack of success may make it more responsive to the appeals of those who are politically alienated (O'Brien and Gwartney-Gibbs, 1989).

Third-order effects

The second-order effects described in the previous section will, through reduced marriage rates and increased divorce and female labour force participation rates, reduce the proportion of households with stay-at-home spouses, which increases the tendency to purchase market replacements for the goods and services traditionally produced by women in the home. The result is a ‘commoditization’ of many goods and services that used to be produced in the home. They are now exchanged in the market – and thus counted in official measures of GDP and productivity – whereas previously they were part of the excluded ‘non-market’ economy.

This commoditization of goods and services causes measures of industrial structure to skew strongly toward services and retail, away from agriculture and manufacturing, creating low-wage service jobs. In addition, the influx of inexperienced young workers as members of a large birth cohort – both men and women – into the labour market exacerbates any decline in productivity growth by changing the composition of the workforce to one dominated by inexperienced and therefore lower-productivity workers. This decline in relative wages of younger workers resulting from their oversupply would lead employers to substitute cheaper labour for more expensive capital, thus lowering the young workers’ productivity still further by providing those low-wage workers with less productivity-enhancing machinery and technology.

Although some analysts maintain that the potential age structure effect of the baby boomers on personal savings is not large enough to explain the full drop in US national savings rates since the 1980s, studies of this phenomenon to date have focused only on the behaviour of the baby boomers themselves. However, one might argue that the baby boomers have affected the propensity to save in age groups other than their own. For example, because boomers’ earnings were depressed and they experienced an inflated housing market when they went to buy homes (both the effects of their own large cohort size), many parents of baby boomers drew on their own savings in order to help with down payments.

When the age structure of children is permitted to affect consumption and savings, a very strong age-related pattern of expenditures and saving can be identified. Children induce savings on the part of their

parents between the ages of five and 16, possibly in anticipation of later educational expenses. When the relationships identified in this way are combined with the changing age distribution in the US population during the 20th century, they produce a savings rate that fluctuates by plus or minus 25 per cent around the mean, simply as a result of changing age structure (Macunovich, 2002).

Similarly, a strong effect has been identified of changing age structure (measured simply as the proportion of young to old in the population) on real interest rates and inflation, because of differential patterns of savings and consumption with age (McMillan and Baesel, 1990). A higher proportion of young adults in a population will produce lower aggregate savings levels – and hence higher interest rates. In this model, today's lower interest and inflation rates are the result of the ageing of the baby boomers, as they begin to acquire assets for their retirement years. The converse of this phenomenon – the potential ‘meltdown’ effect of a retiring baby boom on financial markets, asset values and interest rates – has been described as well (Schieber and Shoven, 1994).

Some research has estimated a strong effect of age structure on housing prices in the United States, with the entry of the baby boom into the housing market causing the severe house price inflation of the 1970s and 1980s, and the entry of the baby bust causing house price deflation (Mankiw and Weil, 1989). Although some have disputed the magnitude of the effect estimated there, most researchers have confirmed its existence. A later study, for example, found significant effects of detailed (single year) age structure in the adult population on all forms of consumption, including housing demand, and on money demand (Fair and Dominguez, 1991).

These potential effects on aggregate demand, savings rates, interest rates and inflation suggest that there might have been a connection between changing age structure and macroeconomic fluctuations in the United States and elsewhere during the 20th century. When the population of young adults is expanding, the resultant growth in demand for durable goods creates confidence in investors, while an unexpected slowdown in the growth rate of young adults could cause cutbacks in production and investment in response to inventory buildups, with a snowball effect throughout the economy. There was a close correspondence in the United States in the 20th century between ‘turnaround points’ of growth in the key age group of 15–24, and significant economic dislocations in 1908, 1929, 1938 and 1974. Similarly, there was a correlation between age structure and economic performance in industrialized nations in the 1930s, and in both industrialized and developing nations since the 1980s, with the ‘Asian Tigers’ some of the most recent examples (Macunovich, 2002).

Feedback effects on the relationship between relative cohort size and relative income

Easterlin's original statements recognized the potential effects of outside influences on the relative cohort size mechanism (Easterlin, 1987). However, the dynamic nature of the mechanism – the fact that many of these other factors would, in fact, be secondary and tertiary results of changing relative cohort size, and thus endogenous in any empirical application – has not been fully appreciated in most analyses to date. As a result, it is often concluded that the hypothesis may have been relevant in the post-Second World War period up to about 1980, but that it fails to extend beyond one full cycle to apply to the period since 1980.

The aggregate demand effect of changing relative cohort size, discussed in the previous section, is hypothesized to contribute significantly to the observed asymmetry in relative cohort size effects on

male relative income. Although cohorts on the leading edge of a baby boom experience declining wages relative to those of older workers, they do so in an economy experiencing strong growth in aggregate demand resulting from the increasing relative cohort size among young adults. Cohorts on the lagging edge of a baby boom, however, enter a labour market weakened by the economic slump resulting from a transition from expanding to contracting relative cohort size.

Similarly, as one of the secondary effects of changing relative cohort size discussed earlier, female labour force participation is hypothesized to have increased in response to declining male relative income as the leading edge of the baby boom entered the labour market. If, as hypothesized, these young women also increased their levels of educational attainment in anticipation of future labour market participation, they would have in many cases competed directly with the male members of their cohort and exacerbated the effects of relative cohort size on male relative income. This effect would have been greatest for cohorts on the lagging edge of the boom – those who should have benefited from declining relative cohort size. It is important in empirical analyses to recognize the potential endogeneity of these other factors, rather than treat relative cohort size effects as ‘contingent’ on exogenous changes in female labour force participation, educational attainment and wages. Wage analyses based on relative cohort size which control for a cohort's position in the US baby boom – and thus allow for aggregate demand and female labour force changes – can explain most of the observed change in young men's entry level wages and in their returns to experience and education (Macunovich, 2002).

Empirical analyses

Empirically, the most important application of the hypothesis has been to explain the varying experience of young adults in the United States since the Second World War. There is, however, some evidence of its relevance to the experience of developed countries more generally in this period (Korenman and Neumark, 2000; Pampel, 2001; Stockard and O'Brien, 2002; Jeon and Shields, 2005), and perhaps as a mechanism leading to fertility decline during the demographic transition in developing countries (Macunovich, 2002).

Overall, however, empirical analyses testing various aspects of the Easterlin hypothesis have produced fairly mixed results. By 2007 there have been two comprehensive analyses of the literature on the Easterlin hypothesis, and one meta-analysis of 19 studies completed between 1976 and 2002. The meta-analysis (Waldorf and Byun, 2005) focused on the age structure–fertility link, and concluded that analytical problems contribute to an apparent lack of empirical support for the Easterlin hypothesis. Most significant among these were the failure to recognize the endogeneity of an income variable when combined with a relative cohort size variable, and the use of very broad age groups in defining relative cohort size.

The first of the literature reviews considered a broad range of topics, including labour market experience and education; marriage, fertility and divorce; and crime, suicide and alienation. It concluded:

> [T]he evidence for the Easterlin effect proves mixed at best and plain wrong at worst… Aggregate data support the hypothesis more than individual level data, period-specific or time-series data support the hypothesis more than cohort-specific data, experiences from 1945–1980 support the hypothesis more than the years since 1980, and trends in the

> United States support the hypothesis more than trends in European nations. (Pampel and Peters, 1995, p. 189)

The second literature review evaluated 76 published analyses focused solely on fertility, and concluded:

> With an equal number of micro- and macro-level analyses using North American data (twenty-two), the 'track record' of the hypothesis is the same in both venues, with fifteen providing significant support in each case. The literature suggests unequivocal support for the relativity of the income concept in fertility but is less clear regarding the source(s) of differences in material aspirations, and suggests that the observed relationship between fertility and cohort size has varied across countries and time periods due to the effects of additional factors not included in most models. (Macunovich, 1998, p. 53)

This review suggests that, because of data limitations and idiosyncratic interpretations of the hypothesis by individual researchers, many of the studies with unfavourable findings have been only peripherally related to the Easterlin hypothesis.

Conclusion

Since the early 1980s, demographic concepts have encroached modestly on economic theory, as evidenced by the appearance of life cycle, overlapping generations and vintage models. The cohort size hypothesis might be viewed as another in this sequence. Its roots, however, extend beyond economics, reaching out into sociology, demography and psychology, and it seeks to encompass a wider range of attitudinal and behavioural phenomena than is traditionally considered economic.

See Also

- demographic transition
- economic demography

Bibliography

Duesenberry, J.S. 1949. *Income, Saving, and the Theory of Consumer Behaviour*. Cambridge, MA: Harvard University Press.

Easterlin, R.A. 1980. *Birth and Fortune*, 1st edn. New York: Basic Books.

Easterlin, R.A. 1987. *Birth and Fortune*, 2nd edn. Chicago: University of Chicago Press.

Easterlin, R.A., Macdonald, C. and Macunovich, D.J. 1990. How have the American baby boomers fared? Earnings and well-being of young adults 1964–1987. *Journal of Population Economics* 3, 277–90.

Ernst, C. and Angst, J. 1983. *Birth Order: Its Influence on Personality*. Berlin: Springer.

Fair, R.C. and Dominguez, K. 1991. Effects of the changing U.S. age distribution on macroeconomic equations. *American Economic Review* 81, 1276–94.

Freeman, R.B. 1976. *The Overeducated American*. New York: Academic Press.

Heer, D.M. 1985. Effects of sibling number on child outcome. *Annual Review of Sociology* 11, 27–47.

Jeon, Y. and Shields, M.P. 2005. The Easterlin hypothesis in the recent experience of higher-income OECD countries: a panel-data approach. *Journal of Population Economics* 18, 1–13.

Korenman, S. and Neumark, D. 2000. Cohort crowding and the youth labour market: a cross-national analysis. In *Youth Employment and Joblessness in Advanced Countries*, ed. D.G. Blanchflower and R.B. Freeman. NBER Comparative Labour Market Series. Chicago: University of Chicago Press.

Macunovich, D.J. 1998. Fertility and the Easterlin hypothesis: an assessment of the literature. *Journal of Population Economics* 11, 1–59.

Macunovich, D.J. 2002. *Birth Quake: The Baby Boom and Its After Shocks*. Chicago: University of Chicago Press.

Macunovich, D.J. and Easterlin, R.A. 1990. How parents have coped: the effect of life cycle decisions on the economic status of pre-school age children, 1964–1987. *Population and Development Review* 16, 301–25.

Mankiw, N.G. and Weil, N.D. 1989. The baby boom, the baby bust and the housing market. *Regional Science and Economics* 19, 235–58.

McMillan, H.M. and Baesel, J.B. 1990. The macroeconomic impact of the baby boom generation. *Journal of Macroeconomics* 12, 167–95.

McNown, R. and Rajbhandary, S. 2003. Time series analysis of fertility and female labour market behaviour. *Journal of Population Economics* 16, 501–23.

O'Brien, R.M. and Gwartney-Gibbs, P.A. 1989. Relative cohort size and political alienation: three methodological issues and a replication supporting the Easterlin hypothesis. *American Sociological Review* 54, 476–80.

O'Brien, R.M., Stockard, J. and Isaacson, L. 1999. The enduring effects of cohort characteristics on age-

specific homicide rates 1960–1995. *American Journal of Sociology* 104, 1061–95.

Pampel, F.C. 2001. *The Institutional Context of Population Change: Patterns of Fertility and Mortality across High-Income Nations*. Chicago: University of Chicago Press.

Pampel, F.C. and Peters, H.E. 1995. The Easterlin effect. *Annual Review of Sociology* 21, 163.

Schieber, S.J. and Shoven, J.B. 1994. The consequences of population aging on private pension fund saving and asset markets. Working Paper No. 4665. Cambridge, MA: NBER.

Stockard, J. and O'Brien, R.M. 2002. Cohort effects on suicide rates: international variations. *American Sociological Review* 67, 854–72.

Waldorf, B. and Byun, P. 2005. Meta-analysis of the impact of age structure on fertility. *Journal of Population Economics* 18, 14–40.

Welch, F. 1979. Effects of cohort size on earnings: the baby boom babies’ financial bust. *Journal of Political Economy* 87, 65–97.

How to cite this article

Macunovich, Diane J. and Richard A. Easterlin. "Easterlin hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000002>
doi:10.1057/9780230226203.0421

The New Palgrave Dictionary of Economics Online

Eckstein, Otto (1927 – 1984)

Lester C. Thurow

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Eckstein, O.; forecasting; macroeconometric models

Article

Eckstein was an entrepreneur who moved a whole technology from the research community into the marketplace. Until he founded Data Resources, Inc., macroeconometric models were research vehicles and not vehicles for aiding business decision making. Under his direction Data Resources came to dominate the marketplace for this type of information, but more importantly it changed the nature of the game. To be taken seriously after his innovation, all economic forecasts had to be buttressed with econometric equations and no large firm would attempt to begin its decision-making processes without an understanding of the national and international economic forecasts emanating from such models.
Born in Ulm, Germany, in 1927, Dr Eckstein fled to England in 1938 and came to the United States in 1939. He graduated from Stuyvesant High School in New York City and served in the United States Army Signal Corps from 1946 to 1947. He received an AB degree from Princeton University in 1951 and a Ph.D. from Harvard University in 1955.
In 1968, he and Donald B. Marron founded Data Resources, Inc., which has grown into the largest economic information company in the world. The firm became a subsidiary of McGraw-Hill, Inc. in 1979. He directed the development of the Data Resources Model of the US economy, and was responsible for its forecasting operations.
As an immigrant to the United States from Nazi Germany, Otto Eckstein wanted to contribute something to America's future success. Better economic policies that would lead to a higher American standard of living were not an abstraction to him. They were the centre of his professional life.
His professional career began with the analysis of large scale multi-year water resources projects and how one might better allocate national resources in such projects. In the late 1950s he was the principal intellectual director of a Joint Economic Committee study on how the United States might break out of what was then seen as the stagnation of the mid-1950s. His study on growth, full employment and price stability laid the basis for the successful economic policies that were followed in the first two-thirds of

the 1960s. But he went on to implement those intellectual foundations as a member of the President's Council of Economic Advisers under President Johnson.
No one who knew the enthusiasm of Otto Eckstein for studying, teaching, and practising economics could thereafter think of economics as the dismal science.

Selected works

1958a. *Water Resources Development: The Economics of Project Evaluation*. Cambridge, MA: Harvard University Press.

1958b. (With J.V. Krutilla.) *Multiple Purpose River Development*. Baltimore: Johns Hopkins Press.

1964a. *Public Finance*. New York: Prentice-Hall. 4th edn, 1979.

1964b. (With E.S. Kirschen and others.) *Economic Policy in Our Time*. Amsterdam: North-Holland.

1967. (ed.) *Studies in the Economics of Income Maintenance*. Washington, DC: Brookings.

1970. (ed.) *The Econometrics of Price Determination*. Washington, DC: Board of Governors of the Federal Reserve System and Social Science Research Council.

1976. (ed.) *Parameters and Policies in the U.S. Economy*. Amsterdam: North-Holland.

1978. *The Great Recession*. Amsterdam: North-Holland.

1981. *Core inflation*. New York: Prentice-Hall.

1983. *The DRI Model of the U.S. Economy*. New York: McGraw-Hill.

1984. (With C. Caton, R. Brinner and P. Duprey.) *The DRI Report on U.S. Manufacturing Industries*. New York: McGraw-Hill.

How to cite this article

Thurow, Lester C. "Eckstein, Otto (1927–1984)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000004> doi:10.1057/9780230226203.0422

The New Palgrave Dictionary of Economics Online

ecological economics

Anastasios Xepapadeas

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Ecological economics is the study of the interactions and co-evolution in time and space of human economies and the ecosystems in which human economies are embedded. It uncovers the links and feedbacks between human economies and ecosystems, and so provides a unified picture of ecology and economy. The link between ecology and human economies has been manifested in the development of resource management or bio-economic models, in which the main focus has been on fishery or forestry management where the impact of humans on ecosystems is realized through harvesting. More closed links have been developed, however, as both disciplines evolve.

Keywords

adiabatic parameter; biodiversity; bio-economics; Boulding, K.; Diffusion; ecological economics; ecology; ecosystems; externalities; Fisher–Kolmogorov equation; intrinsic growth rate; Kolmogorov model; logistic function; Lotka, A.; Nash equilibrium; optimality behaviour; predator–prey models; random walk models; Red Queen hypothesis; spatial economics; tragedy of the commons; Turing mechanism; Volterra, V.

Article

Ecology can be regarded as the study of living species such as animals, plants and microorganisms, and the relations among them and their natural environment. In this context, an ecosystem includes these species and their non-living environment, their interactions, and their evolution in time and space (see, for example, Roughgarden, May and Levin, 1989). Economics, meanwhile, is the study of how human societies use scarce resources to produce commodities and to distribute them among their members. The need for an interdisciplinary approach – ‘ecological economics’ – stems from the fact that natural ecosystems and human economies are closely linked. In the process of production and consumption, human beings use ecosystems and their services, influence their evolution, and are the recipients of

feedbacks originating from their actions upon ecosystems. As Kenneth Boulding (1965) notes in his classic paper ‘Earth as a space ship’, which can be regarded as a landmark in the emergence of ecological economics, ‘Man is finally going to have to face the fact that he is a biological system living in an ecological system, and that his survival power is going to depend on his developing symbiotic relationships of a closed-cycle character with all the other elements and populations of the world of ecological systems.’

Thus, ecological economics can be regarded as the study of the interactions and co-evolution in time and space of human economies and the ecosystems in which human economies are embedded. This implies that the task of ecological economics is to bridge the gap between economy and ecology by uncovering the links and the feedbacks between human economies and ecosystems, and by using these links and feedbacks to provide a unified picture of ecology and economy and their interactions and co-evolution. In a sense, ecological economics aims at linking ecological models and economic models in order to provide insights into complex and interrelated phenomena stemming from and affecting both ecosystems and human economies.

The natural link between ecology and human economies has been manifested in the traditional development of resource management or bio-economic models (for example, Clark, 1990), in which the main focus has been on fishery or forestry management where the impact of humans on ecosystems is realized through harvesting. More close links have been developed, however, as both disciplines evolve. Common methodological approaches may also be encountered in ecology and economics. Optimality behaviour, which is fundamental in economics, has also been used to provide insights into the structure of ecological systems, in the context of optimal foraging behaviour, species competition, or net energy maximization by organisms (for example, Tschirhart, 2000; Tilman, Polasky and Lehman, 2005) with the purpose of founding macro-behaviours in ecosystems – such as those emerging from population dynamics – on micro-foundations.

In the same context, the classical phenomenological-descriptive approach to species competition based on Lotka–Volterra systems has recently been complemented by mechanistic resource-based models of species competition for limiting resources (Tilman, 1982; 1988). This approach has obvious links to competition among economic agents for limited resources. Furthermore, by linking the functioning of natural ecosystems with the provision of useful services to humans, or by using concepts such as ecosystems productivity, insurance from the genetic diversity of ecological systems against catastrophic events, or development of new products using genetic resources existing in natural ecosystems (Heal, 2000), new insights into the fundamental issues of the valuation of ecosystems or the valuation of biodiversity have been derived. (Examples of useful services to humans include provisioning services, such as food, water, fuel, genetic material; regulation services, such as climate regulation, disease regulation; and cultural services and supporting services, such as soil formation, nutrient cycling; see Millennium Ecosystem Assessment, 2005.)

Ecological models

The traditional bio-economic models (Clark, 1990), which describe the evolution of the population or the biomass of species when harvesting takes place, have formed the building blocks of ecological-economic modelling. These models can be extended along various lines to provide a more realistic

picture of ecosystems (for a detailed analysis, see Murray, 2003) and help build meaningful ecological-economic models. To start with, let $x(t)$ denote the biomass of a certain species at time t. Then evolution of the biomass is described by an ordinary differential equation

$$\frac{dx(t)}{dt} = \text{birth} - \text{natural death} + \text{migration} - \text{harvesting}. \tag{1}$$

In the analysis of population models it is common, unless it is a specific case, to set the migration rate at zero, and to represent the natural rate of population growth (birth-natural death) by a function $F(x)$. The most common specification of this function is the famous *logistic function*, which is $F(x) = rx(1 - x/K)$. In this function r is a positive constant called *intrinsic growth rate* and K is the *carrying capacity* of the environment which depends on factors such as resource availability or environmental pollution. If we denote by $h(t)$ the rate of harvesting of the species biomass by humans, the population model becomes:

$$\frac{dx(t)}{dt} = F(x) - h(t), \quad x(0) = x_0. \tag{2}$$

If $h(t) \equiv F(x)$, the population remains constant and the harvesting rate corresponds to *sustainable yield*. Harvesting rate is usually modelled as population dependent or $h=qEx$, where q is a positive constant, referred to as a *catchability coefficient* in fishery models, and E is *harvesting effort*. Human activities can affect the species population, in addition to harvesting, by affecting parameters such as the intrinsic growth rates or the carrying capacity. For example, if the stock of environmental pollution of a certain pollutant (such as phosphorus in a lake) in a natural ecosystem is denoted by P, with dynamics described by

$$\frac{dP(t)}{dt} = g(s(t), P(t)), \quad P(0) = P_0, \tag{3}$$

where $s(t)$ is the rate of emissions (such as phosphorus loadings), and the pollutant affects parameters of the population model, then the combined model will be (3) along with

$$\frac{dx(t)}{dt} = r(P)x\left[1 - \frac{x}{K(P)}\right] - qEx,\ r'(P) < 0,\ \ K'(P) < 0.$$
(4)

If the catchability coefficient is affected by technical change, then it can be expressed by a function of time as $q(t)$. In this case (4) is not autonomous. Alternatively q can be a function of technological variables like R&D evolving in the economic module.
The population model (2) can be generalized to age-structured populations and multi-species populations. In multi-species populations the Lotka–Volterra predator–prey models are classic. If we denote the prey population by $x(t)$ and the predator population by $y(t)$ and ignore harvesting for the moment to simplify things, the model can be written as

$$\frac{dx(t)}{dt} = x\left[r\left(1 - \frac{x}{K}\right) - yR(x)\right],\ x(0) = x_0$$
(5)

$$\frac{dy(t)}{dt} = ym\left(1 - \frac{ny}{x}\right),\ y(0) = y_0 \quad m, n > 0$$
(6)

where $R(x)$ is a function called the *predation term*, which can be specified as $\gamma x / (x^2 + \delta^2)$, γ, $\delta > 0$. A more general multi-species model with J prey and J predators can be written, for $i = 1, \ldots, J$, as

$$\frac{dx_i(t)}{dt} = x_i\left[a_i - \sum_{j=1}^{J} \beta_{ij} y_j\right],\ x_i(0) = x_{i0}$$
(7)

$$\frac{dy_i(t)}{dt} = y_i\left[\sum_{j=1}^{J} \gamma_{ij}x_j - \delta_i\right],\ y_i(0) = y_{i0}$$
(8)

where all parameters are positive constants. An even more general model of interacting populations can be obtained by the generalized Kolmogorov model where the evolution of each species biomass is described by:

$$\frac{dx_i(t)}{dt} = x_iF_i(x_1, x_2, x_3, \ldots)i = 1, 2, 3, \ldots$$
(9)

In the mechanistic resource-based models of species competition emerging from the work of Tilman (for example, Tilman, 1982; 1988), species compete for limiting resources. (For the use of this model in ecological-economic modelling, see Brock and Xepapadeas, 2002; Tilman, Polasky and Lehman, 2005.) In these models the growth of a species depends on the limiting resource, and interactions among species take place through the species' effects on the limiting resource. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the vector of species biomasses and *R* the amount of the available limiting resource. Then a mechanistic resource-based model with a single limiting factor in a given area and $i = 1, \ldots, n$ species can be described by the following equations:

$$\frac{\dot{x}_i}{x_i} = g_i(R) - d_i,\ x_i(0) = x_{i0}$$
(10)

$$\dot{R} = S - aR - \sum_{i=1}^{n} w_ix_ig_i(R)$$
(11)

where $g_i(R)$ is resource-related growth, d_i is the species' natural death rate, S is the amount of resource supplied, a is the natural resource removal rate (leaching rate), and w_i is specific resource consumption by species i. The main result in this framework relates to an exclusion principle stating that, in a landscape free of disturbances, the species with the lowest resource requirement in equilibrium will competitively displace all other species, driving the system to a monoculture. Species coexistence and polycultures in equilibrium can be supported in a system with more than one limiting resource, or even in single resource systems if there is temperature-dependent growth and temperature variation in the ecosystem, spatial or temporal variations in resource ratios, differences in local palatabilities and local abundance of herbivores.

In addition to the temporal variation captured by the models described above an important characteristic of ecosystems is that of spatial variation. Biological resources tend to disperse in space under forces promoting 'spreading' or 'concentrating' (Okubo, 2001); these processes, along with intra- and inter-species interactions, induce the formation of spatial patterns for species. A central concept in modelling the dispersal of biological resources is that of *diffusion*. Diffusion is defined as a process whereby the microscopic irregular movement of particles such as cells, bacteria, chemicals, or animals results in some macroscopic regular motion of the group. Biological diffusion is based on random walk models which, when coupled with population growth equations, lead to general reaction-diffusion systems (see, for example, Okubo and Levin, 2001; Murray, 2003). When only one species is examined, the coupling of classical diffusion with a logistic growth function leads to the so-called Fisher–Kolmogorov equation, which can be written as

$$\frac{\partial x(z,t)}{\partial t} = F(x(z,t)) + D_x \frac{\partial^2 x(z,t)}{\partial z^2} \tag{12}$$

where $x(z,t)$ denotes the concentration of the biomass at spatial point z at time t. The biomass grows according to a standard growth function $F(x)$ which determines the resource's kinetics but also disperses in space with a constant diffusion coefficient D_x. (Nonlinear reaction diffusion equations are associated with propagating wave solutions.) In general, a diffusion process in an ecosystem tends to produce a uniform population density, that is, spatial homogeneity. Thus it might be expected that diffusion would 'stabilize' ecosystems where species disperse and humans intervene through harvesting.

There, is however, one exception, known as 'diffusion induced instability' or 'diffusive instability'. It was Alan Turing (1952) who suggested that under certain conditions reaction-diffusion systems can generate spatially heterogeneous patterns. This is the so-called 'Turing mechanism' for generating diffusion instability. With two interacting species evolving according to

$$\frac{\partial x(z,t)}{\partial t} = F(x,y) + D_x \frac{\partial^2 x(z,t)}{\partial z^2}$$

(13)

$$\frac{\partial y(z,t)}{\partial t} = G(x, y) + D_y \frac{\partial^2 y(z,t)}{\partial z^2},$$

(14)

if in the absence of diffusion $(D_x = D_y = 0)$ the system tends to a spatially uniform stable steady state, then under certain conditions, depending on the relationship D_x/D_y, spatially heterogeneous patterns can emerge due to diffusion-induced instability.

Spatial variations in ecological systems can also be analysed in terms of meta-population models. A meta-population is a set of local populations occupying isolated patches which are connected by migrating individuals. Meta-population dynamics can be developed for single or many species (Levin, 1974). For the single species case the dynamics become

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x})\mathbf{x} + \mathbf{D}\mathbf{x}$$

(15)

where $\mathbf{x} = (x_1, \ldots, x_l)$ is a column vector of species densities, **F** has its *i*th row depending on the *i*th row of **x**, and **D**=[d_{ij}] is a connectivity matrix, where d_{ij} is the rate of movement from patch *j* to patch $i\,(j \neq i)$. Thus dynamics are local with the exception of movements from one patch to the other.

A more general model encompassing $i = 1, \ldots, n$ species competing for $j = 1, \ldots, J$ limiting resources, with density-dependent growth and interactions across patches $c = 1, \ldots, C$ in a given landscape, can be written as

$$\frac{\dot{x}_{ci}}{x_{ci}} = F_{ic}(\mathbf{x}_c, \mathbf{x}_{-c})\, g_{ic}(\mathbf{R}_c, d_{ic}), \quad \forall i, c$$

(16)

$$\dot{R}_{jc} = S_{jc}(\mathbf{R}_c, \mathbf{R}_{-c}) - D_{jc}(\mathbf{x}_c, \mathbf{x}_{-c}, \mathbf{R}_c, \mathbf{R}_{-c}), \ \forall j, c$$
(17)

where $\mathbf{R}_{-c}$, $\mathbf{x}_{-c}$ are respectively vectors of resources and species outside patch *c*.
(For a detailed analysis, see Brock and Xepapadeas, 2002.) A more general set-up can be obtained in the context of co-evolutionary models which describe the interactions between population (or biomass) dynamics and mutation (or trait dynamics). Antagonistic co-evolution of species on the one hand and pests or parasites or the other can be described by the so-called Red Queen hypothesis (see, for example, Van Valen, 1973, and Kawecki, 1998). According to this hypothesis, parasites evolve ceaselessly in response to perpetual evolution of species' (or hosts') resistance. The co-evolution of the parasites' ability to attack (virulence) and the hosts' resistance is expected to indicate persistent fluctuations of resistance and virulence. In this context the Red Queen hypothesis generates a continuous need for variation, resulting in a limit cycle or other non-point attractor in trait space dynamics, which are called Red Queen races. Red Queen cycles are observed in a slow time scale, since trait dynamics are assumed to evolve slowly, in contrast to the population, host-parasite, dynamics which are assumed to evolve fast (see Dieckmann and Law, 1996).
A simple co-evolutionary model can be developed in a system with one harvested ('useful') species or host species whose biomass is denoted by *x* and a parasite denoted by *y*, where the abundance of *x* and *y* depends on the evolution of two characteristics or traits denoted by *d* and γ (see, for example, the Red Queen dynamic models developed by Krakauer and Jansen, 2002), where *d* affects the fitness of *x* and γ affects the fitness of *y*.
Let the growth rates of *x* and the pathogen *y* be given by

$$g_x = \frac{\dot{x}}{x} = (s - rx - yQ(d, \gamma))$$
(18)

$$g_y = \frac{\dot{y}}{y} = (xQ(d, \gamma) - \delta).$$
(19)

If we measure fitness by growth rates, then $\frac{\partial Q(d,\gamma)}{\partial d} < 0,$ so that an increase in *d* increases fitness of *x*. In the same way, $\frac{\partial Q(d,\gamma)}{\partial \gamma} > 0$ for an increase in γ to increase fitness of *y*. In equilibrium of the fast population system where $\dot{x} = \dot{y} = 0,$ it holds that

$$\hat{x} = \frac{\delta}{Q(d,\gamma)},\ \hat{y} = \frac{s - r\hat{x}}{Q(d,\gamma)},\ s \geq r\hat{x}.$$
(20)

On the assumption of constant mutation rates μ_d and μ_γ, the evolutionary dynamics for the traits *d* and γ, when population dynamics have reached the asymptotically stable steady state, are given by

$$\dot{d} = -\mu_d \hat{x}\hat{y}\frac{\partial Q(d,\gamma)}{\partial d}$$
(21)

$$\dot{\gamma} = \mu_\gamma \hat{x}\hat{y}\frac{\partial Q(d,\gamma)}{\partial \gamma}.$$
(22)

See Krakauer and Jansen (2002) who, by considering the slow time scale trait dynamics, show that the equilibrium point $(d^*, \gamma^*): \dot{d} = \dot{\gamma} = 0$ is not attracting; the dynamics spiral away from this point. This behaviour is the oscillatory, Red Queen dynamics.

Ecological-economic modelling

The ecological models developed above are the cornerstones of the development of meaningful ecological-economic models. The impact of humans on the population of species can be realized through direct harvesting *h* as described in (1) and (2). This type of impact can be easily incorporated into the more general population dynamic models by selecting the harvested species. Human influence can also be realized in an indirect way by having the environmental carrying capacity affected by environmental

pollution generated in the non-harvesting sector of the economy, as in (3) and (4), or by having technological considerations affecting catchability coefficients. It is also possible that external environmental conditions which are anthropogenic, such as global warming, can make some parameters associated with population dynamics or mutation dynamics change slowly. This can be modelled in (21) and (22) by considering μ_d and μ_γ as slow varying parameters, defined as $\mu_d(\varepsilon t)$ and $\mu_\gamma(\varepsilon t)$, where $0 < \varepsilon \ll 1$ is the *adiabatic parameter*. This slowly varying system could be used to model slow anthropogenic impacts on ecosystem structure.

However, the size and the severity of the impact of human economies in ecosystems depend on the way in which variables, such as harvesting or other variables which can be chosen by humans (such as emissions, investment in harvesting capacity) and which influence the evolution of ecosystems, are actually chosen. These variables can be regarded as *control variables*, and the way in which they are chosen affects the evolution of ecological variables, such as species biomasses or traits, which can be considered as the *state variables* of the problem.

The typical approach in economics is to associate the choice of the control variables with *optimizing behaviour*. Thus, the control variables are chosen so that a criterion function is optimized, and the economic problem of ecosystem management – where management means choice of control variables – is defined as a formal optimal control problem. In this problem the objective is the optimization of the criterion function subject to the constraints imposed by the structure of the ecosystem. These constraints, which provide the transition equations of the optimal control problem, are the dynamic equations of the ecological models described in the previous section.

The solution of the ecological-economic model, provided it exists, will determine the paths of the state and the control variables and the steady state of the system, which will determine the long-run equilibrium values of the ecological populations as well as the approach dynamics to the steady state. In this context, managed ecological systems which are predominantly nonlinear could exhibit dynamic behaviour characterized by multiple, locally stable and unstable steady states, limit cycles, or the emergence of hysteresis, bifurcations or irreversibilities.

The way in which the objective function is set up and the ecological constraints which are taken into account determine the solution of the ecological-economic model. In principle, a *socially optimal solution* can be distinguished from a *privately optimal solution*. The socially optimal solution corresponds to the so-called problem of the *social planner*, where the objective function takes into account not only benefits from harvesting certain resources of the ecological system, which corresponds to harvesting commercially valuable biomass, but in addition a wide spectrum of flows of services generated by the whole ecosystem. These include, as described above, regulation, cultural or supporting services, existence values, or benefits associated with productivity or insurance gains. If $V(\mathbf{h}(t))$ denotes harvesting benefits at time *t* associated with harvesting vector **h**, and $U(\mathbf{x}(t))$ denotes the flow of benefits associated with ecosystem service generated by species biomasses existing in the ecosystem and not removed by harvesting, then the total flow of benefit is $V(\mathbf{h}(t)) + U(\mathbf{x}(t))$. In this formulation, the $V(\cdot)$ and $U(\cdot)$ functions are usually assumed to be monotonically increasing and concave. In a more general setup, the total benefit function can be non-separable, defined as $u(\mathbf{h}(t), \mathbf{x}(t))$.

The objective can then be written as:

$$\max_{\{\mathbf{h}(t)\}} \int_0^{\infty} e^{-\rho t} [V(\mathbf{h}(t)) + U(\mathbf{x}(t))] \, dt$$
(23)

where $\rho \geq 0$ is a discount rate. It should be noted that in principle benefits associated with $V(\mathbf{h}(t))$ can be estimated using market data, while benefits associated with $U(\mathbf{x}(t))$ are hard to estimate because markets for the larger part of the spectrum of ecosystem services are missing. (Valuation of ecosystem services is an open question. For details, see, for example, Bingham et al., 1995.) The social optimum corresponds to the maximization of (23), subject to the constraints imposed by the ecological system. For example, if we use the generalized model of resource competition, the constraints are:

$$\frac{\dot{x}_{ci}}{x_{ci}} = F_{ic}(\mathbf{x}_c, \mathbf{x}_{-c}) g_{ic}(\mathbf{R}_c, d_{ic}) - h_{io} \quad \forall i, c$$
(24)

$$\dot{R}_{jc} = S_{jc}(\mathbf{R}_c, \mathbf{R}_{-c}) - D_{jc}(\mathbf{x}_c, \mathbf{x}_{-c}, \mathbf{R}_c, \mathbf{R}_{-c}), \quad \forall j, c.$$
(25)

A solution $(\mathbf{h}^*(t), \mathbf{x}^*(t))$ is regarded as the socially optimal solution.
The privately optimal solution is distinguished from the socially optimal by the fact that only harvesting benefits enter the objective function. The assumption is that management is carried out by a ‘small’ profit-maximizing private agent that ignores the general flows of ecosystem services. In this case, the private agents do not take into account *externalities* associated with their management practices on ecosystem service flow and $U(\mathbf{x}(t)) \equiv 0$. Market externalities associated with the definition of $V(\mathbf{h})$ could relate to imperfections in the markets for the harvested commodities, or to property rights-related externalities, as the well known ‘tragedy of the commons’ emerging in the harvesting of open access resources.

In general the privately optimal solution $(\mathbf{h}^0(t), \mathbf{x}^0(t))$ will deviate from the socially optimal solution. Another type of externality can be associated with strategic behaviour in resource harvesting if more than one private agent harvests the resource. If $l = 1, \ldots, L$ harvesters are present, then the biomass equation (24) for patch *c* becomes

$$\frac{\dot{x}_{ci}}{x_{ci}} = F_{ic}(\mathbf{x}_c, \mathbf{x}_{-c})\, g_{ic}(\mathbf{R}_c, d_{ic}) - \sum_{l=1}^{L} h_{io} \quad \forall i, c. \tag{26}$$

In this case the privately optimal solution can be obtained as an *open loop or feedback* Nash equilibrium. Privately optimal solutions can also be distinguished from the socially optimal by the extent to which the ecological constraints are taken into account. For example, if resource dynamics or trait dynamics are not taken into account in the optimization problem, the management rule will deviate from the social optimum. Furthermore, since *all* the ecological constraints are operating, there will be discrepancies between the perceived evolution of ecosystems under management that ignores certain constraints, and the actual evolution of the ecosystem. Brock and Xepapadeas (2003), show that, by ignoring genetic constraints associated with the development of resistance to genetically modified organisms, the actual system loses any productivity advantage because of resistance development.

These discrepancies might be a cause for *surprises* in ecosystem management. For example, with reference to the co-evolutionary model (18) – (22), profit-maximizing decisions which ignore evolution might steer the system to a certain steady state on a fast time scale, but then the underlying trait dynamics might move the system in slow time to another attractor.

The deviations between the private solution and the social optimum provide a basis for regulation which is similar to the rationale behind the regulation of environmental externalities. Regulation could take the form, in general spatial models of ecosystem management, of species-specific and site-specific taxes on harvesting, or equivalent quota and zoning systems.

See Also

- approximate solutions to dynamic models (linear methods)
- common property resources
- consumption externalities
- dynamic programming
- environmental economics
- spatial economics
- spatial econometrics

Bibliography

Bingham, G. et al. 1995. Issues inecosystem valuation: improving information for decision making. *Ecological Economics* 14(2), 73–90.

Boulding, K. 1965. Earth as a space ship. Washington State University Committee on Space Sciences.

Kenneth E. Boulding Papers, Archives (Box # 38), University of Colorado at Boulder Libraries. Online. Available at http://www.colorado.edu/econ/Kenneth.Boulding/spaceship-earth.html, accessed 13 July 2005.

Brock, W. and Xepapadeas, A. 2002. Optimal ecosystem management when species compete for limiting resources. *Journal of Environmental Economics and Management* 44, 189–230.

Brock, W. and Xepapadeas, A. 2003. Valuing biodiversity from an economic perspective: a unified economic, ecological and genetic approach. *American Economic Review* 93, 1597–614.

Clark, C. 1990. *Mathematical Bioeconomics: The Optimal Management Of Renewable Resources*. 2nd edn. New York: Wiley.

Dieckmann, U. and Law, R. 1996. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of Mathematical Biology* 34, 579–612.

Kawecki, T. 1998. Red Queen meets Santa Rosalia: arms races and the evolution of host specialization in organisms with parasitic lifestyles. *American Naturalist* 152(4), 635–51.

Krakauer, D. and Jansen, V. 2002. Red Queen dynamics in protein translation. *Journal of Theoretical Biology* 218, 97–109.

Heal, G. 2000. *Nature and the Marketplace: Capturing the Value of Ecosystem Services*. Washington, DC: Island Press.

Levin, S. 1974. Dispersion and population interactions. *American Naturalist* 108, 207–28.

Millennium Ecosystem Assessment. 2005. *Ecosystems and Human Well-being. Volume 3: Policy Responses*. Washington, DC: Island Press.

Murray, J. 2003. *Mathematical Biology*, 3rd edn. Berlin, Springer.

Okubo, A. 2001. Introduction: the mathematics of ecological diffusion. In *Diffusion and Ecological Problems: Modern Perspectives*, 2nd edn., ed. A. Okubo and S. Levin. Berlin: Springer.

Okubo, A. and Levin, S. 2001. The basics of diffusion. In *Diffusion and Ecological Problems: Modern Perspectives*, 2nd edn., ed. A. Okubo and S. Levin. Berlin: Springer.

Roughgarden, J., May, R. and Levin, S. 1989. *Perspectives in Ecological Theory*. Princeton, NJ: Princeton University Press.

Tilman, D. 1982. *Resource Competition and Community Structure*. Princeton, NJ: Princeton University Press.

Tilman, D. 1988. *Plant Strategies and the Dynamics and Structure of Plant Communities*. Princeton, NJ: Princeton University Press.

Tilman, D., Polasky, S. and Lehman, C. 2005. Diversity, productivity and temporal stability in the economies of humans and nature. *Journal of Environmental Economics and Management* 49, 405–26.

Tschirhart, J. 2000. General equilibrium of an ecosystem. *Journal of Theoretical Biology* 203, 13–32.

Turing, A. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 237(641), 37–72.

Van Valen, L. 1973. A new evolutionary law. *Evolutionary Theory* 1, 1–30.

How to cite this article

Xepapadeas, Anastasios. "ecological economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000221> doi:10.1057/9780230226203.0423

The New Palgrave Dictionary of Economics Online

ecological inference

Gary King, Ori Rosen and Martin Tanner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Ecological inference is a general statistical problem where a response variable is not available at the subject level because summary statistics are reported for groups only. It consists of merging information from different databases which are not linked to each other at the record level. We consider an election scenario where in each electoral precinct the fraction of voting-age people who turn out to vote, the fraction of black population and the number of voting-age people are observed. The proportions of blacks and of whites who vote are unobserved because electoral results and census data are not linked.

Keywords

aggregation; ecological inference; likelihood; Markov chain Monte Carlo methods; method of bounds; nonparametric models; statistical approaches

Article

1 The ecological inference problem

For expository purposes, we discuss only an important but simple special case of ecological inference, and adopt the running example and notation from King (1997: ch. 2). The basic problem has two observed variables (T_i and X_i) and two unobserved quantities of interest (β_i^b and β_i^w) for each of p observations. Observations represent aggregate units, such as geographic areas, and each individual-level variable within these units is dichotomous.
To be more specific, in Figure 1 we observe for each electoral precinct $i(i = 1, \ldots, p)$ the fraction of voting age people who turnout to vote (T_i) and who are black (X_i), along with the number of voting age people (N_i). The quantities of interest, which remain unobserved because of the secret ballot, are the proportions of blacks who vote (β_i^b) and whites who vote (β_i^w). The proportions β_i^b and β_i^w are not observed because T_i and X_i are from different data sources (electoral results and census data, respectively) and record linkage is impossible (and illegal), and so the cross-tabulation cannot be computed.
Figure 1
Notation for Precinct i. *Note*: The goal is to estimate the quantities of interest, β_i^b (the fraction of blacks who vote) and β_i^w (the fraction of whites who vote), from the aggregate variables X_i (the fraction of voting age people who are black) and T_i (the fraction of people who vote), along with N_i (the known number of voting age people).

Race of voting age person	Voting decision: Vote	Voting decision: No Vote	
Black	β_i^b	$1-\beta_i^b$	X_i
White	β_i^w	$1-\beta_i^w$	$1-X_i$
	T_i	$1-T_i$	

Also of interest are the district-wide fractions of blacks and whites who vote, which are respectively

$$B^b = \frac{\sum_{i=1}^{p} N_i X_i \beta_i^b}{\sum_{i=1}^{p} N_i X_i}, \text{ and} \quad (1)$$

$$B^w = \frac{\sum_{i=1}^{p} N_i (1-X_i) \beta_i^w}{\sum_{i=1}^{p} N_i (1-X_i)}. \quad (2)$$

These are weighted averages of the corresponding precinct-level quantities. Some methods aim to estimate only B^b and B^w without giving estimates of β_i^b and β_i^w for all i.

2 Deterministic and statistical approaches

The ecological inference literature before King (1997) was bifurcated between supporters of the method of bounds, originally proposed by Duncan and Davis (1953), and supporters of statistical approaches, proposed even before Ogburn and Goltra (1919) but first formalized into a coherent statistical model by Goodman (1953; 1959). (For the historians of science among us: although these two monumental articles were written by two colleagues and friends in the same year and in the same department and university – the Department of Sociology at the University of Chicago – the principal did not discuss their work prior to completion. Even by today's standards, nearly a half century after their publication, the articles are models of clarity and creativity.) Although Goodman and Duncan and Davis moved on to other interests following their seminal contributions, most of the ecological inference literature in the five decades since 1953 was an ongoing war between supporters of these two key approaches, and often without the usual academic decorum.

2.1 Extracting deterministic information: the method of bounds

The purpose of the method of bounds and its generalizations is to extract deterministic information, known with certainty, about the quantities of interest.
The intuition behind these quantities is simple. For example, if a precinct contained 150 African-Americans and 87 people in the precinct voted, then how many of the 150 African-American actually cast their ballot? We do not know exactly, but bounds on the answer are easy to obtain: in this case, the answer must lie between 0 and 87. Indeed, conditional only on the data being correct, [0,87] is a 100 per cent confidence interval. Intervals like this are sometimes narrow enough to draw meaningful inferences, and sometimes they are too wide, but the ability to provide (non-trivial) 100 per cent confidence intervals in even some situations is quite rare in any statistical field.
In general, before any data are seen, the unknown parameters β_i^b and β_i^w are each bounded on the unit interval. Once we observe T_i and X_i they are bounded more narrowly, as:

$$\beta_i^b \in \left[\max\left(0, \frac{T_i - (1 - X_i)}{X_i}\right), \min\left(\frac{T_i}{X_i}, 1\right)\right]$$
$$\beta_i^w \in \left[\max\left(0, \frac{T_i - X_i}{1 - X_i}\right), \min\left(\frac{T_i}{1 - X_i}, 1\right)\right]. \quad (3)$$

Deterministic bounds on the district-level quantities B^b and B^w are weighted averages of these precinct-level bounds.
The bounds then indicate that the parameters in each case fall within these deterministic bounds with certainty, and in practice they are almost always narrower than [0,1]. Whether they are narrow enough in any one application depends on the nature of the data.

2.2 Extracting statistical information: Goodman's regression

Leo Goodman's (1953; 1959) approach is very different from, but just as important as, Duncan and Davis's. He looked at the same data and focused on the statistical information. His approach examines variation in the marginals (X_i and T_i) over the precincts to attempt to reason back to the district-wide fractions of blacks and whites who vote, B^b and B^w. The outlines of this approach and the problems with it have been known at least since Ogburn and Goltra (1919). For example, if in precincts with large proportions of black citizens we observe that many people do not vote, then it may seem reasonable to infer that blacks turn out at lower rates than whites. Indeed, it often is reasonable, but not always. The problem is that it could instead be the case that the whites who happen to live in heavily black precincts are the ones who vote less frequently, yielding the opposite ecological inference to the individual-level truth.

What Goodman accomplished was to formalize the logic of the approach in a simple regression model, and to give the conditions under which estimates from such a model are unbiased. To see this, note first that the accounting identity

$$T_i = X_i\beta_i^b + (1 - X_i)\beta_i^w \quad (4)$$

holds exactly. Then he showed that a regression of T_i on X_i and $(1 - X_i)$ with no constant term could be used to estimate B^b and B^w, respectively. The key assumption necessary for unbiasedness that Goodman identified is that the parameters and X_i be uncorrelated: $\mathrm{Cov}(\beta_i^b, X_i) = \mathrm{Cov}(\beta_i^w, X_i) = 0$. In the example, the assumption is that blacks vote in the same proportions in homogeneously black areas as in more integrated areas. Obviously, this is true sometimes and it is false other times. (King, 1997: ch. 3, showed that Goodman's assumption was necessary but not sufficient. To have unbiasedness, it must also be true that the parameters and N_i are uncorrelated.)

As Goodman recognized, when this key assumption does not hold, estimates from the model will be biased. Indeed, they can be very biased, outside the deterministic bounds, and even outside the unit interval. This technique has been used extensively since the 1950s, and impossible estimates occur with considerable frequency (some estimates range to a majority of real applications; Achen and Shively, 1995).

2 Extracting both deterministic and statistical information: King's EI approach

From 1953 until 1997, the only two approaches used widely in practice were the method of bounds and Goodman's regression. King's (1997) idea was that the insights from these two conflicting literatures in fact do not conflict with each other; the sources of information are largely distinct and can be combined to improve inference overall and synergistically. The idea is to combine the information from the bounds, applied to both quantities of interest for each and every precinct, with a statistical approach for extracting information within the bounds. The amount of information in the bounds depends on the data-set, but for many data-sets it can be considerable. For example, if precincts are spread uniformly over a scatterplot of X_i by T_i, the average bounds on β_i^b and β_i^w are narrowed from [0,1] to less than half of that range – hence eliminating half of the ecological inference problem with certainty. This additional information also helps make the statistical portion of the model far less sensitive to assumptions than previous statistical methods which exclude the information from the bounds.

To illustrate these points, we first present all the information available without making any assumptions, thus extending the bounds approach as far as possible. As a starting point, the left graph in Figure 2 provides a scatterplot of a sample data set as observed, X_i horizontally by T_i vertically. Each point in this figure corresponds to one precinct, for which we would like to estimate the two unknowns. We display the unknowns in the right graph of the same figure; any point in the right graph portrays values of the two unknowns, β_i^b which is plotted horizontally, and β_i^w which is plotted vertically. Ecological inference involves locating, for each precinct, the one point in this unit square corresponding to the true values of β_i^b and β_i^w, since values outside the square are logically impossible.

Figure 2

Two views of the same data. *Note*: The left graph is a scatterplot of the observables, X_i by T_i. The right graph displays this same information as a tomography plot of the quantities of interest, β_i^b by β_i^w. Each precinct *i* that appears as a point in the left graph is a line (rather than a point because of information lost due to aggregation) in the right graph. For example, precinct 52 appears as the dot with a little square around it in the left graph and the dark line in the right graph. *Source*: The data are from King (1997: Figures 5.1 and 5.5).

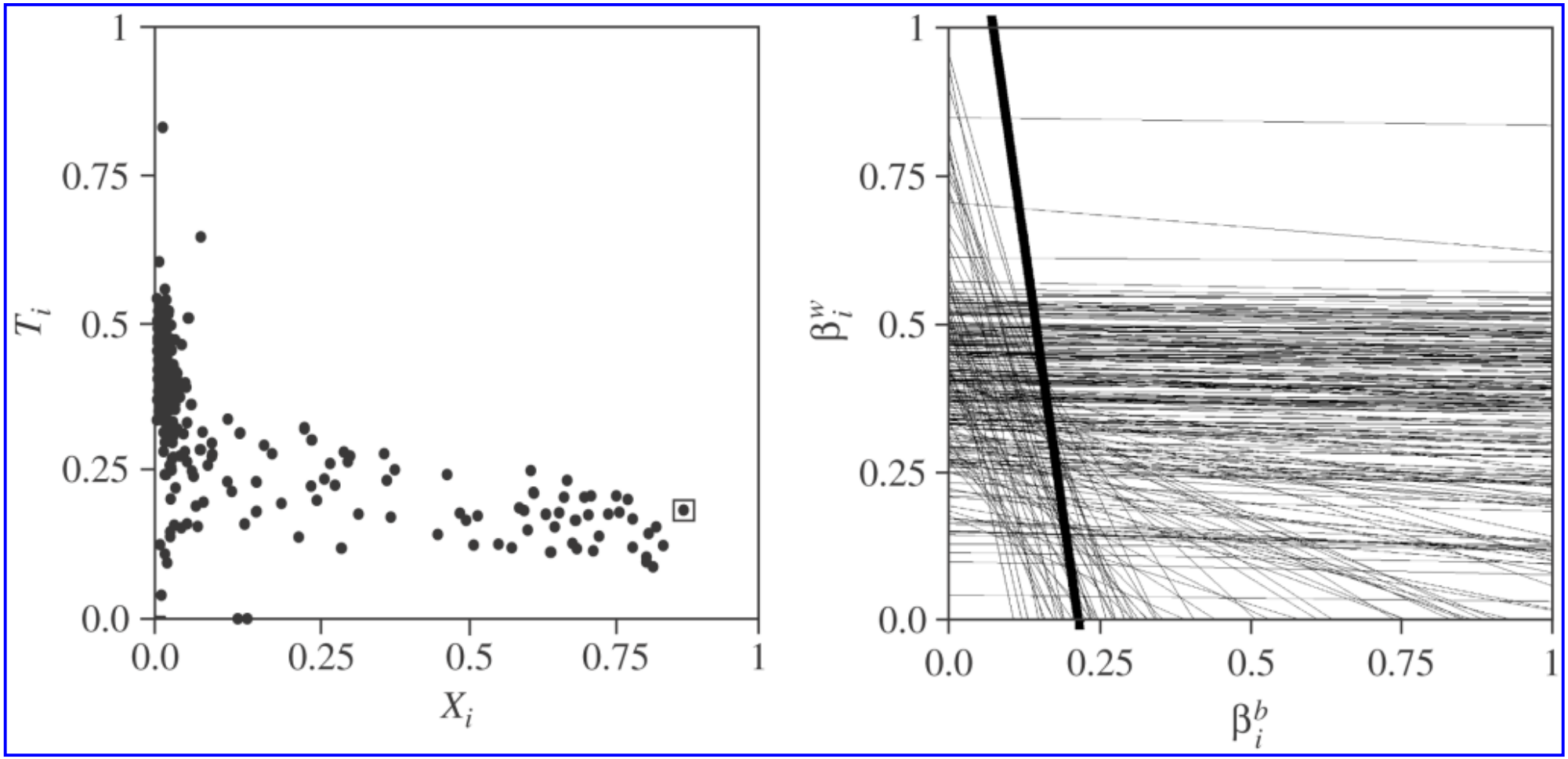

To map the knowns onto the unknowns, King begins Goodman's accounting identity from eq. (4). From this equation, which holds exactly, King solves for one unknown in terms of the other:

$$\beta_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right)\beta_i^b, \tag{5}$$

which shows that β_i^w is a *linear* function of β_i^b, where the intercept and slope are known (since they are functions of the data, X_i and T_i).

King then maps the knowns from the left graph onto the right graph by using the linear relationship in eq. (5). A key point is that each dot on the left graph can be expressed, without assumptions or loss of information, as what King called a ‘tomography’ line within the unit square in the right graph. It is precisely the information lost due to aggregation that causes us to have to plot an entire line (on which the true point must fall) rather than the goal of one point for each precinct on the right graph. In fact, the information lost is equivalent to having a graph of the β_i^b by β_i^w points but having the ink smear, making the points into lines and partly but not entirely obscuring the correct positions of the (β_i^b, β_i^w) points. (King also showed that the ecological inference problem is mathematically equivalent to the ill-posed ‘tomography’ problem of many medical imaging procedures, such as CAT and PET

scans, where one attempts to reconstruct the inside of an object by passing X-rays through it and gathering information only from the outside. Because the line sketched out by an X-ray is closely analogous to eq. (5), King labels the latter a *tomography line* and the corresponding graph a *tomography graph*.)

What does a tomography line tell us? Before we know anything, we know that the true (β_i^b, β_i^w) point must lie somewhere within the unit square. After X_i and T_i are observed for a precinct, we also know that the true point must fall on a specific line represented by eq. (5) and appearing in the tomography plot in Figure 2. In many cases narrowing the region to be searched for the true point from the entire square to the one line in the square can provide a significant amount of information. To see this, consider the point enclosed in a box in the left graph, and the corresponding dark line in the right graph. This precinct, number 52, has observed values of X_{52} =0.88 and T_{52} =0.19. As a result, substituting into eq. (5) gives $\beta_i^w = 1.58 - 7.33\beta_i^b$, which when plotted appears as the dark line on the right graph. This particular line tells us that, in our search for the true $\beta_{52}^b, \beta_{52}^w$ point on the right graph, we can eliminate with certainty all area in the unit square except that on the line, which is clearly an advance over not having the data. Translated into the quantities of interest, this line tells us (by projecting the line downward to the horizontal axis) that, wherever the true point falls on the line, β_{52}^b must fall in the relatively narrow bounds of [0.07,0.21]. Unfortunately, in this case, β_i^w can only be bounded (by projecting to the left) to somewhere within the entire unit interval. More generally, lines that are relatively steep, like this one, tell us a great deal about β_i^b and little about β_i^w. Tomography lines that are relatively flat give narrow bounds on β_i^w and wide bounds on β_i^b. Lines that cut off the bottom left (or top right) of the figure give narrow bounds on both quantities of interest.

If the only information available to learn about the unknowns in precinct *i* is X_i and T_i, a tomography line like that in Figure 2 exhausts all this available information. This line immediately tells us the known bounds on each of the parameters, along with the precise relationship between the two unknowns, but it is not sufficient to narrow in on the right answer any further. Fortunately, additional information exists in the other observations in the same data set (X_j and T_j for all $i \neq j$) which, under the right assumptions, can be used to learn more about β_i^b and β_i^w in our precinct of interest.

In order to borrow statistical strength from all the precincts to learn about β_i^b and β_i^w in precinct *i*, some assumptions are necessary. The simplest version of King's model (that is, the one most useful for expository purposes) requires three assumptions, each of which can be relaxed in different ways.

First, the set of (β_i^b, β_i^w) points must fall in a single cluster within the unit square. The cluster can fall anywhere within the square; it can be widely or narrowly dispersed or highly variable in one unknown and narrow in the other; and the two unknowns can be positively, negatively, or not at all correlated over *i*. An example that would violate this assumption would be two or more distinct clusters of (β_i^b, β_i^w) points, as might result from subsets of observations with fundamentally different data generation processes (such as from markedly different regions). The specific mathematical version of this one-cluster assumption is that β_i^b and β_i^w follow a truncated bivariate normal density

$$\mathrm{TN}(\beta_i^b, \beta_i^w | \check{\mathfrak{B}}, \check{\Sigma}) = \mathrm{N}(\beta_i^b, \beta_i^w | \check{\mathfrak{B}}, \check{\Sigma}) \frac{1(\beta_i^b, \beta_i^w)}{R(\check{\mathfrak{B}}, \check{\Sigma})}, \tag{6}$$

where the kernel is the untruncated bivariate normal,

$$\mathrm{N}(\beta_i^b, \beta_i^w | \check{\mathfrak{B}}, \check{\Sigma}) = (2\pi)^{-1} |\check{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2} (\beta_i - \check{\mathfrak{B}})' \check{\Sigma}^{-1} (\beta_i - \check{\mathfrak{B}}) \right], \tag{7}$$

and $\mathbf{1}(\beta_i^b, \beta_i^w)$ is an indicator function that equals 1 if $\beta_i^b \in [0, 1]$ and $\beta_i^w \in [0, 1]$ and zero otherwise. The normalization factor in the denominator, $R(\check{\mathfrak{B}}, \check{\Sigma})$, is the volume under the untruncated normal distribution above the unit square:

$$R(\check{\mathfrak{B}}, \check{\Sigma}) = \int_0^1 \int_0^1 \mathrm{N}(\beta^b, \beta^w | \check{\mathfrak{B}}, \check{\Sigma}) d\beta^b d\beta^w \tag{8}$$

When divided into the untruncated normal, this factor keeps the volume under the truncated distribution equal to 1. The parameters of the truncated density, which we summarize as

$$\check{\psi} = \left\{ \check{\mathfrak{B}}^b, \check{\mathfrak{B}}^w, \check{\sigma}_b, \check{\sigma}_w, \check{\rho} \right\} = \left\{ \check{\mathfrak{B}}, \check{\Sigma} \right\}, \tag{9}$$

are on the scale of the untruncated normal (and so, for example, $\check{\mathfrak{B}}^b$ and $\check{\mathfrak{B}}^w$ need not be constrained to the unit interval even though β_i^b and β_i^w are constrained by this density). The second assumption, which is necessary to form the likelihood function, is the absence of spatial autocorrelation: conditional on X_i, T_i and T_j are mean independent. Violations of this assumption in empirically reasonable (and even some unreasonable) ways do not seem to induce much if any bias.

The final, and by far the most critical, assumption is that X_i is independent of β_i^b and β_i^w. The three assumptions together produce what has come to be known as King's ‘basic’ EI model. (The use of EI to name this method comes from the name of his software, available at http://GKing.Harvard.edu.) King also generalizes this assumption, in what has come to be known as the ‘extended’ EI model, by allowing the truncated normal parameters to vary as functions of measured covariates, Z_i^b and Z_i^w, giving:

$$\check{\mathfrak{B}}_i^b = \left[\varphi_1 (\check{\sigma}_b^2 + 0.25) + 0.5 \right] + (Z_i^b - \bar{Z}^b) \alpha^b \check{\mathfrak{B}}_i^w = \left[\varphi_2 (\check{\sigma}_w^2 + 0.25) + 0.5 \right] + (Z_i^w - \bar{Z}^w) \alpha^w \tag{10}$$

where α^b and α^w are parameter vectors to be estimated along with the original model parameters and that have as many elements as Z_i^b and Z_i^w have columns. This relaxes the mean independence assumptions to:

$$\mathrm{E}(\beta_i^b | X_i, Z_i) = \mathrm{E}(\beta_i^b | Z_i) \mathrm{E}(\beta_i^w | X_i, Z_i) = \mathrm{E}(\beta_i^w | Z_i).$$

Note that this extended model also relaxes the assumptions of truncated bivariate normality, since there is now a separate density being assumed for each observation. Because the

bounds, which differ in width and information content for each i, generally provide substantial information, even X_i can be used as a covariate in Z_i. (The recommended default setting in EI includes X_i as a covariate with a prior on its coefficient.) In contrast, under Goodman's regression, which does not include information in the bounds, including X_i leads to an unidentified model (King, 1997: sec. 3.2).

These three assumptions – one cluster, no spatial autocorrelation, and mean independence between the regressor and the unknowns conditional on X_i and Z_i – enable one to compute a posterior (or sampling) distribution of the two unknowns in each precinct. A fundamentally important component of EI is that the quantities of interest are not the parameters of the likelihood but instead come from conditioning on T_i and producing a posterior for β_i^b and β_i^w in each precinct. Failing to condition on T_i and examining the parameters of the truncated bivariate normal only makes sense if the model holds exactly and so is much more model-dependent than King's approach. Since the most important problem in ecological inference modelling is precisely model misspecification, failing to condition on T assumes away the problem without justification. This point is widely regarded as a critical step in applying the EI model (Adolph and King, with Herron and Shotts, 2003).

When bounds are narrow, EI model assumptions do not matter much. But, for precincts with wide bounds on a quantity of interest, inferences can become model dependent. This is especially the case with ecological inference problems precisely because of the loss of information due to aggregation. In fact, this loss of information can be expressed by noting that the joint distribution of β_i^b and β_i^w cannot be fully identified from the data without some untestable assumptions. To be precise, distributions with positive mass over *any* curve or combination of curves that connects the bottom left point $(\beta_i^b = 0, \beta_i^w = 0)$ to the top right point $(\beta_i^b = 1, \beta_i^w = 1)$ of a tomography plot cannot be rejected by the data (King, 1997: 191). Other features of the distribution are estimable. This fundamental indeterminacy is, of course, a problem because it prevents pinning down the quantities of interest with certainty, but it can also be something of an opportunity since different distributional assumptions can lead to the same estimates, especially since only those pieces of the distributions above the tomography lines are used in the final analysis.

4 Alternative approaches to ecological inference

In the continuing search for more information to bring to bear on ecological inferences, King, Rosen and Tanner (1999) extend King's (1997) model another step. They incorporate King's main advance of combining deterministic and statistical information but begin modelling a step earlier at the individuals who make up the counts. They also build a hierarchical Bayesian model, using easily generalizable Markov chain Monte Carlo (MCMC) technology (Tanner, 1996).

To define the model formally, let T_i' denote the *number* of voting age people who turn out to vote. At the top level of the hierarchy they assume that T_i' follows a binomial distribution with probability equal to $\theta_i = X_i\beta_i^b + (1 - X_i)\beta_i^w$ and count N_i. Note that at this level it is assumed that the *expectation* of T_i', rather than T_i', is equal to $X_i\beta_i^b + (1 - X_i)\beta_i^w$. In other words, King (1997) models T_i as a continuous proportion, whereas King, Rosen, and Tanner (1999) recognize the inherently discrete nature of the counts of voters that go into computing this proportion. The two models are connected, of course, since T_i/N_i approaches T_i as N_i gets large.

The connection to King's tomography line can be seen in the contribution of the data from precinct i to the likelihood, which is

$$(X_i\beta_i^b + (1 - X_i)\beta_i^w)^{T_i'}(1 - X_i\beta_i^b - (1 - X_i)\beta_i^w)^{(N_i - T_i')}. \tag{11}$$

By taking the logarithm of this contribution to the likelihood and differentiating with respect to β_i^b and β_i^w, King, Rosen and Tanner show that the maximum of (11) is not a unique point, but rather a line whose equation is given by the tomography line in eq. (5). Thus, the log-likelihood for precinct i looks like two playing cards leaning against each other. As long as T_i is fixed and bounded away from 0.5 (and X_i is a fixed known value between 0 and 1), the derivative at this point is seen to increase with N_i, that is, the pitch of the playing cards increases with the sample size. In other words, for large N_i, the log-likelihood for precinct i degenerates from a surface defined over the unit square into a single playing card standing perpendicular to the unit square and oriented along the corresponding tomography line.

At the second level of the hierarchical model, β_i^b is distributed as a beta density with parameters c_b and d_b and β_i^w follows an independent beta with parameters c_w and d_w. While β_i^b

and β_i^w are assumed *a priori* independent, they are *a posteriori* dependent. At the third and final level of the hierarchical model, the unknown parameters c_b, d_b, c_w and d_w follow an exponential distribution with a large mean.

A key advantage of this model is that it generalizes immediately to arbitrarily large $R \times C$ tables. This approach was pursued by Rosen et al. (2001), who also provided a much faster method of moment-based estimator. For an application, see King et al. (2003).

Wakefield (2004) presents an alternative approach based on the Bayesian paradigm using a Markov chain Monte Carlo inference scheme. King, Rosen and Tanner (2004) survey the latest strategies for solving ecological inference problems in various fields, many of which do not fit the textbook case of a 2 × 2 table with known marginals and unknown cell entries. Staniswalis (2005) proposes a nonparametric model for ecological inference with an application to renal failure data.

Bibliography

Achen, C. and Shively, W. 1995. *Cross-Level Inference*. Chicago: University of Chicago Press.

Adolph, C., King, G., Herron, M. and Shotts, K. 2003. A consensus position on second stage ecological inference models. *Political Analysis* 11, 86–94.

Duncan, O. and Davis, B. 1953. An alternative to ecological correlation. *American Sociological Review* 18, 665–6.

Goodman, L. 1953. Ecological regressions and the behavior of individuals. *American Sociological Review* 18, 663–6.

Goodman, L. 1959. Some alternatives to ecological correlation. *American Journal of Sociology* 64, 610–24.

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

King, G., Rosen, O. and Tanner, M. 1999. Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research* 28, 61–90.

King, G., Rosen, O. and Tanner, M., eds. 2004. *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press.

King, G., Rosen, O., Tanner, M. and Wagner, A. 2003. The ordinary election of Adolf Hitler: a modern voting behavior approach. Online. Available at http://gking.harvard.edu/files/naziV.pdf, accessed 16 August 2006.

Ogburn, W. and Goltra, I. 1919. How women vote: a study of an election in Portland, Oregon. *Political Science Quarterly* 34, 413–33.

Rosen, O., Jiang, W., King, G. and Tanner, M. 2001. Bayesian and frequentist inference for ecological inference: the *R*×*C* case. *Statistica Neerlandica* 55(2), 134–56.

Staniswalis, J. 2005. On fitting generalized non-linear models with varying coefficients. *Computational Statistics and Data Analysis* 50, 893–902.

Tanner, M. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd edn. New York: Springer-Verlag.

Wakefield, J. 2004. Prior and likelihood choices in the analysis of ecological data. In *Ecological Inference: New Methodological Strategies*, ed. G. King, O. Rosen and M. Tanner. Cambridge: Cambridge University Press.

How to cite this article

King, Gary, Ori Rosen and Martin Tanner. "ecological inference." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume.

Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000271> doi:10.1057/9780230226203.0424

The New Palgrave Dictionary of Economics Online

econometrics

John Geweke, Joel L. Horowitz and Hashem Pesaran

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

As a unified discipline, econometrics is still relatively young and has been transforming and expanding very rapidly. Major advances have taken place in the analysis of cross-sectional data by means of semiparametric and nonparametric techniques. Heterogeneity of economic relations across individuals, firms and industries is increasingly acknowledged and attempts have been made to take it into account either by integrating out its effects or by modelling the sources of heterogeneity when suitable panel data exist. The counterfactual considerations that underlie policy analysis and treatment valuation have been given a more satisfactory foundation. New time-series econometric techniques have been developed and employed extensively in the areas of macroeconometrics and finance. Nonlinear econometric techniques are used increasingly in the analysis of cross-section and time-series observations. Applications of Bayesian techniques to econometric problems have been promoted largely by advances in computer power and computational techniques. The use of Bayesian techniques has in turn provided the investigators with a unifying framework where the tasks of forecasting, decision making, model evaluation and learning can be considered as parts of the same interactive and iterative process, thus providing a basis for ‘real time econometrics’.

Keywords

acceptance sampling; adaptive expectations hypothesis; ARMA processes; asset pricing models; asset return volatility; auctions; Bachelier, L.; Bayesian computation; Bayesian econometrics; Bayesian inference; Benini, R.; binary logit and probit models; bootstrap; building cycle; bunch maps; causality in economics and econometrics; censored regression models; central limit theorems; cointegration; common factors; conditional hazard functions; conditional mean functions; conditional median functions; confluence analysis; convexity; correlation analysis; Cowles Commission; curse of dimensionality; Davenant, C.; diagnostic tests; discrete choice models; discrete response models; distributed lags; Douglas, P.H.; Duhem–Quine thesis; duration models; dynamic decision models; dynamic specification; dynamic stochastic general equilibrium models; Econometric Society;

econometrics; economic distance; economic laws; Edgeworth expansions; Edgeworth, F. Y.; efficient market hypothesis; Engel curve; error correction models; Euler equations; experimental economics; financial econometrics; Fisher, I.; Fisher, R. A.; fixed effects and random effects; forecast error variances; forecast evaluation; forecasting; Frisch, R. A. K.; full information maximum likelihood; Galton, F.; Gaussian quadrature; generalized method of moments; geometric distributed lag model; Gibbs sampling; Haavelmo, T.; habit persistence; Hastings–Metropolis algorithm; hedonic prices; homogeneity; Hooker, R.H.; identification; impulse response analysis; indirect utility function; inference; instrumental variables; integration; inventory cycle; joint hypotheses; Juglar cycle; Juglar, C.; k-class estimators; kernel estimators; King, G.; Kitchin, J.; Kondratieff, N.; Koopmans, T. C.; Kuznets, S.; labour market search; Lagrange multiplier; latent variables; least absolute deviations estimators; likelihood ratio; limited information maximum likelihood; linear models; local linear estimation; logit models; long waves; longitudinal data; Lucas critique; macroeconometric models; Markov chain Monte Carlo methods; maximum likelihood; measurement; measurement errors; method of simulated moments; microeconometrics; microfoundations; misspecification; Mitchell, W. C.; model evaluation; model selection; model testing; model uncertainty; monotonicity; Monte Carlo simulation; Moore, H.L.; multicollinearity; multinomial probit model; National Bureau of Economic Research; nonlinear simultaneous equation models; non-nested tests; nonparametric models; observed variables; ordinary least squares; parameter uncertainty; Pearson K.; Petty, W.; Phillips curve; policy evaluation; political arithmeticians; probability; probability calculus; probability distribution; purchasing power parity; quantile functions; random assignment; random utility models; random variables; random walk theory; rational expectations; real time econometrics; regional migration; regression analysis; revealed preference theory; saddlepoint expansions; sampling theory; Schultz, H.; semiparametric estimation; sensitivity analysis; series estimators; significance tests; simulated method of moments; simulation methods; simultaneous equations models; simultaneous linear equations; social experimentation in economics; spatial econometrics; specification tests; splines; spurious correlation; state dependence; state space models; statistical inference; statistics and economics; stochastic models; stock return predictability; structural change; structural estimation; structural VAR; survival models; three-stage least squares; time-series analysis; Tinbergen, J.; Tobit models; treatment effect; truncated regression models; uncovered interest parity; unit roots; value distribution; vector autoregressions (VAR); Vining, R.; Waugh, F.; Weibull hazard model; Working, H.

Article

1 What is econometrics?

Broadly speaking, econometrics aims to give empirical content to economic relations for testing economic theories, forecasting, decision making, and for *ex post* decision/policy evaluation. The term ‘econometrics’ appears to have been first used by Pawel Ciompa as early as 1910, although it is Ragnar Frisch who takes the credit for coining the term, and for establishing it as a subject in the sense in which

it is known today (see Frisch, 1936, p. 95, and Bjerkholt, 1995). By emphasizing the quantitative aspects of economic relationships, econometrics calls for a ‘unification’ of measurement and theory in economics. Theory without measurement can have only limited relevance for the analysis of actual economic problems; while measurement without theory, being devoid of a framework necessary for the interpretation of the statistical observations, is unlikely to result in a satisfactory explanation of the way economic forces interact with each other. Neither ‘theory’ nor ‘measurement’ on its own is sufficient to further our understanding of economic phenomena.

As a unified discipline, econometrics is still relatively young and has been transforming and expanding very rapidly since an earlier version of this article was published in the first edition of *The New Palgrave: A Dictionary of Economics* in 1987 (Pesaran, 1987a). Major advances have taken place in the analysis of cross-sectional data by means of semiparametric and nonparametric techniques. Heterogeneity of economic relations across individuals, firms and industries is increasingly acknowledged, and attempts have been made to take them into account either by integrating out their effects or by modelling the sources of heterogeneity when suitable panel data exists. The counterfactual considerations that underlie policy analysis and treatment evaluation have been given a more satisfactory foundation. New time series econometric techniques have been developed and employed extensively in the areas of macroeconometrics and finance. Nonlinear econometric techniques are used increasingly in the analysis of cross-section and time-series observations. Applications of Bayesian techniques to econometric problems have been given new impetus largely thanks to advances in computer power and computational techniques. The use of Bayesian techniques has in turn provided the investigators with a unifying framework where the tasks of forecasting, decision making, model evaluation and learning can be considered as parts of the same interactive and iterative process; thus paving the way for establishing the foundation of ‘real time econometrics’. See Pesaran and Timmermann (2005a).

This article attempts to provide an overview of some of these developments. But to give an idea of the extent to which econometrics has been transformed over the past decades we begin with a brief account of the literature that pre-dates econometrics, and discuss the birth of econometrics and its subsequent developments to the present. Inevitably, our accounts will be brief and non-technical. Readers interested in more details are advised to consultant the specific entries provided in the *New Palgrave* and the excellent general texts by Maddala (2001), Greene (2003), Davidson and MacKinnon (2004), and Wooldridge (2006), as well as texts on specific topics such as Cameron and Trivedi (2005) on microeconometrics, Maddala (1983) on econometric models involving limited-dependent and qualitative variables, Arellano (2003), Baltagi (2005), Hsiao (2003), and Wooldridge (2002) on panel data econometrics, Johansen (1995) on cointegration analysis, Hall (2005) on generalized method of moments, Bauwens, Lubrano and Richard (2001), Koop (2003), Lancaster (2004), and Geweke (2005) on Bayesian econometrics, Bosq (1996), Fan and Gijbels (1996), Horowitz (1998), Härdle (1990), Härdle and Linton (1994) and Pagan and Ullah (1999) on nonparametric and semiparametric econometrics, Campbell, Lo and MacKinlay (1997) and Gourieroux and Jasiak (2001) on financial econometrics, Granger and Newbold (1986), L•tkepohl (1991) and Hamilton (1994) on time series analysis.

2 Quantitative research in economics: historical backgrounds

Empirical analysis in economics has had a long and fertile history, the origins of which can be traced at least as far back as the work of the 16th-century political arithmeticians such as William Petty, Gregory King and Charles Davenant. The political arithmeticians, led by Sir William Petty, were the first group to make systematic use of facts and figures in their studies. They were primarily interested in the practical issues of their time, ranging from problems of taxation and money to those of international trade and finance. The hallmark of their approach was undoubtedly quantitative, and it was this which distinguished them from their contemporaries. Although the political arithmeticians were primarily and understandably preoccupied with statistical measurement of economic phenomena, the work of Petty, and that of King in particular, represented perhaps the first examples of a unified quantitative–theoretical approach to economics. Indeed Schumpeter in his *History of Economic Analysis* (1954, p. 209) goes as far as to say that the works of the political arithmeticians ‘illustrate to perfection, what Econometrics is and what Econometricians are trying to do’.

The first attempt at quantitative economic analysis is attributed to Gregory King, who was the first to fit a linear function of changes in corn prices on deficiencies in the corn harvest, as reported in Charles Davenant (1698). One important consideration in the empirical work of King and others in this early period seems to have been the discovery of ‘laws’ in economics, very much like those in physics and other natural sciences.

This quest for economic laws was, and to a lesser extent still is, rooted in the desire to give economics the status that Newton had achieved for physics. This was in turn reflected in the conscious adoption of the method of the physical sciences as the dominant mode of empirical enquiry in economics. The Newtonian revolution in physics, and the philosophy of ‘physical determinism’ that came to be generally accepted in its aftermath, had far-reaching consequences for the method as well as the objectives of research in economics. The uncertain nature of economic relations began to be fully appreciated only with the birth of modern statistics in the late 19th century and as more statistical observations on economic variables started to become available.

The development of statistical theory in the hands of Galton, Edgeworth and Pearson was taken up in economics with speed and diligence. The earliest applications of simple correlation analysis in economics appear to have been carried out by Yule (1895; 1896) on the relationship between pauperism and the method of providing relief, and by Hooker (1901) on the relationship between the marriage rate and the general level of prosperity in the United Kingdom, measured by a variety of economic indicators such as imports, exports, and the movement in corn prices.

Benini (1907), the Italian statistician was the first to make use of the method of multiple regression in economics. But Henry Moore (1914; 1917) was the first to place the statistical estimation of economic relations at the centre of quantitative analysis in economics. Through his relentless efforts, and those of his disciples and followers Paul Douglas, Henry Schultz, Holbrook Working, Fred Waugh and others, Moore in effect laid the foundations of ‘statistical economics’, the precursor of econometrics. The monumental work of Schultz, *The Theory and the Measurement of Demand* (1938), in the United States and that of Allen and Bowley, *Family Expenditure* (1935), in the United Kingdom, and the pioneering works of Lenoir (1913), Wright (1915; 1928), Working (1927), Tinbergen (1929–30) and Frisch (1933) on the problem of ‘identification’ represented major steps towards this objective. The work of Schultz was exemplary in the way it attempted a unification of theory and measurement in demand analysis; while the work on identification highlighted the importance of ‘structural estimation’ in econometrics

and was a crucial factor in the subsequent developments of econometric methods under the auspices of the Cowles Commission for Research in Economics.

Early empirical research in economics was by no means confined to demand analysis. Louis Bachelier (1900), using time-series data on French equity prices, recognized the random walk character of equity prices, which proved to be the precursor to the vast empirical literature on market efficiency hypothesis that has evolved since the early 1960s. Another important area was research on business cycles, which provided the basis of the later development in time-series analysis and macroeconometric model building and forecasting. Although, through the work of Sir William Petty and other early writers, economists had been aware of the existence of cycles in economic time series, it was not until the early 19th century that the phenomenon of business cycles began to attract the attention that it deserved. Clement Juglar (1819–1905), the French physician turned economist, was the first to make systematic use of time-series data to study business cycles, and is credited with the discovery of an investment cycle of about 7–11 years duration, commonly known as the Juglar cycle. Other economists such as Kitchin, Kuznets and Kondratieff followed Juglar's lead and discovered the inventory cycle (3–5 years duration), the building cycle (15–25 years duration) and the long wave (45–60 years duration), respectively. The emphasis of this early research was on the morphology of cycles and the identification of periodicities. Little attention was paid to the quantification of the relationships that may have underlain the cycles. Indeed, economists working in the National Bureau of Economic Research under the direction of Wesley Mitchell regarded each business cycle as a unique phenomenon and were therefore reluctant to use statistical methods except in a nonparametric manner and for purely descriptive purposes (see, for example, Mitchell, 1928; Burns and Mitchell, 1947). This view of business cycle research stood in sharp contrast to the econometric approach of Frisch and Tinbergen and culminated in the famous methodological interchange between Tjalling Koopmans and Rutledge Vining about the roles of theory and measurement in applied economics in general and business cycle research in particular. (This interchange appeared in the August 1947 and May 1949 issues of the *Review of Economics and Statistics*.)

3 The birth of econometrics

Although, quantitative economic analysis is a good three centuries old, econometrics as a recognized branch of economics began to emerge only in the 1930s and the 1940s with the foundation of the Econometric Society, the Cowles Commission in the United States, and the Department of Applied Economics (DAE) in Cambridge, England. (An account of the founding of the first two organizations can be found in Christ, 1952; 1983, while the history of the DAE is covered in Stone, 1978.) This was largely due to the multidisciplinary nature of econometrics, comprising of economic theory, data, econometric methods and computing techniques. Progress in empirical economic analysis often requires synchronous developments in all these four components.

Initially, the emphasis was on the development of econometric methods. The first major debate over econometric method concerned the applicability of the probability calculus and the newly developed sampling theory of R.A. Fisher to the analysis of economic data. Frisch (1934) was highly sceptical of the value of sampling theory and significance tests in econometrics. His objection was not, however, based on the epistemological reasons that lay behind Robbins's and Keynes's criticisms of econometrics.

He was more concerned with the problems of multicollinearity and measurement errors which he believed were pervasive in economics; and to deal with the measurement error problem he developed his confluence analysis and the method of ‘bunch maps’. Although used by some econometricians, notably Tinbergen (1939) and Stone (1945), the bunch map analysis did not find much favour with the profession at large. Instead, it was the probabilistic rationalizations of regression analysis, advanced by Koopmans (1937) and Haavelmo (1944), that formed the basis of modern econometrics.

Koopmans did not, however, emphasize the wider issue of the use of stochastic models in econometrics. It was Haavelmo who exploited the idea to the full, and argued for an explicit probability approach to the estimation and testing of economic relations. In his classic paper published as a supplement to *Econometrica* in 1944, Haavelmo defended the probability approach on two grounds. First, he argued that the use of statistical measures such as means, standard errors and correlation coefficients for inferential purposes is justified only if the process generating the data can be cast in terms of a probability model. Second, he argued that the probability approach, far from being limited in its application to economic data, because of its generality is in fact particularly suited for the analysis of ‘dependent’ and ‘non-homogeneous’ observations often encountered in economic research.

The probability model is seen by Haavelmo as a convenient abstraction for the purpose of understanding, or explaining or predicting, events in the real world. But it is not claimed that the model represents reality in all its details. To proceed with quantitative research in any subject, economics included, some degree of formalization is inevitable, and the probability model is one such formalization. The attraction of the probability model as a method of abstraction derives from its generality and flexibility, and the fact that no viable alternative seems to be available. Haavelmo's contribution was also important as it constituted the first systematic defence against Keynes's (1939) influential criticisms of Tinbergen's pioneering research on business cycles and macroeconometric modelling. The objective of Tinbergen's research was twofold: first, to show how a macroeconometric model may be constructed and then used for simulation and policy analysis (Tinbergen, 1937); second, ‘to submit to statistical test some of the theories which have been put forward regarding the character and causes of cyclical fluctuations in business activity’ (Tinbergen, 1939, p. 11). Tinbergen assumed a rather limited role for the econometrician in the process of testing economic theories, and argued that it was the responsibility of the ‘economist’ to specify the theories to be tested. He saw the role of the econometrician as a passive one of estimating the parameters of an economic relation already specified on a priori grounds by an economist. As far as statistical methods were concerned, he employed the regression method and Frisch's method of confluence analysis in a complementary fashion. Although Tinbergen discussed the problems of the determination of time lags, trends, structural stability and the choice of functional forms, he did not propose any systematic methodology for dealing with them. In short, Tinbergen approached the problem of testing theories from a rather weak methodological position.

Keynes saw these weaknesses and attacked them with characteristic insight (Keynes, 1939). A large part of Keynes's review was in fact concerned with technical difficulties associated with the application of statistical methods to economic data. Apart from the problems of the ‘dependent’ and ‘non-homogeneous’ observations mentioned above, Keynes also emphasized the problems of misspecification, multicollinearity, functional form, dynamic specification, structural stability, and the difficulties associated with the measurement of theoretical variables. By focusing his attack on Tinbergen's attempt at testing economic theories of business cycles, Keynes almost totally ignored the practical significance of Tinbergen's work for econometric model building and policy analysis (for more

details, see Pesaran and Smith, 1985a).

In his own review of Tinbergen's work, Haavelmo (1943) recognized the main burden of the criticisms of Tinbergen's work by Keynes and others, and argued the need for a general statistical framework to deal with these criticisms. As we have seen, Haavelmo's response, despite the views expressed by Keynes and others, was to rely more, rather than less, on the probability model as the basis of econometric methodology. The technical problems raised by Keynes and others could now be dealt with in a systematic manner by means of formal probabilistic models. Once the probability model was specified, a solution to the problems of estimation and inference could be obtained by means of either classical or of Bayesian methods. There was little that could now stand in the way of a rapid development of econometric methods.

4 Early advances in econometric methods

Haavelmo's contribution marked the beginning of a new era in econometrics, and paved the way for the rapid development of econometrics, with the likelihood method gaining importance as a tool for identification, estimation and inference in econometrics.

4.1 Identification of structural parameters

The first important breakthrough came with a formal solution to the identification problem which had been formulated earlier by Working (1927). By defining the concept of ‘structure’ in terms of the joint probability distribution of observations, Haavelmo (1944) presented a very general concept of identification and derived the necessary and sufficient conditions for identification of the entire system of equations, including the parameters of the probability distribution of the disturbances. His solution, although general, was rather difficult to apply in practice. Koopmans, Rubin and Leipnik (1950) used the term ‘identification’ for the first time in econometrics, and gave the now familiar rank and order conditions for the identification of a single equation in a system of simultaneous *linear* equations. The solution of the identification problem by Koopmans (1949) and Koopmans, Rubin and Leipnik (1950) was obtained in the case where there are a priori linear restrictions on the structural parameters. They derived rank and order conditions for identifiability of a single equation from a complete system of equations without reference to how the variables of the model are classified as endogenous or exogenous. Other solutions to the identification problem, also allowing for restrictions on the elements of the variance–covariance matrix of the structural disturbances, were later offered by Wegge (1965) and Fisher (1966).

Broadly speaking, a model is said to be identified if all its structural parameters can be obtained from the knowledge of its implied joint probability distribution for the observed variables. In the case of simultaneous equations models prevalent in econometrics, the solution to the identification problem depends on whether there exists a sufficient number of a priori restrictions for the derivation of the structural parameters from the reduced-form parameters. Although the purpose of the model and the focus of the analysis on explaining the variations of some variables in terms of the unexplained variations of other variables is an important consideration, in the final analysis the specification of a minimum number of identifying restrictions was seen by researchers at the Cowles Commission to be

the function and the responsibility of ‘economic theory’. This attitude was very much reminiscent of the approach adopted earlier by Tinbergen in his business cycle research: the function of economic theory was to provide the specification of the econometric model, and that of econometrics to furnish statistically optimal methods of estimation and inference. More specifically, at the Cowles Commission the primary task of econometrics was seen to be the development of statistically efficient methods for the estimation of structural parameters of an a priori specified system of simultaneous stochastic equations.

More recent developments in identification of structural parameters in context of semiparametric models is discussed below in Section 12. See also Manski (1995).

4.2 Estimation and inference in simultaneous equation models

Initially, under the influence of Haavelmo's contribution, the maximum likelihood (ML) estimation method was emphasized as it yielded consistent estimates. Anderson and Rubin (1949) developed the limited information maximum likelihood (LIML) method, and Koopmans, Rubin and Leipnik (1950) proposed the full information maximum likelihood (FIML). Both methods are based on the joint probability distribution of the endogenous variables conditional on the exogenous variables and yield consistent estimates, with the former utilizing all the available a priori restrictions and the latter only those which related to the equation being estimated. Soon, other computationally less demanding estimation methods followed, both for a fully efficient estimation of an entire system of equations and for a consistent estimation of a single equation from a system of equations.

The two-stage least squares (2SLS) procedure was independently proposed by Theil (1954; 1958) and Basmann (1957). At about the same time the instrumental variable (IV) method, which had been developed over a decade earlier by Reiersol (1941; 1945), and Geary (1949) for the estimation of errors-in-variables models, was generalized and applied by Sargan (1958) to the estimation of simultaneous equation models. Sargan's generalized IV estimator (GIVE) provided an asymptotically efficient technique for using surplus instruments in the application of the IV method to econometric problems, and formed the basis of subsequent developments of the generalized method of moments (GMM) estimators introduced subsequently by Hansen (1982). A related class of estimators, known as k-class estimators, was also proposed by Theil (1958). Methods of estimating the entire system of equations which were computationally less demanding than the FIML method were also advanced. These methods also had the advantage that, unlike the FIML, they did not require the full specification of the entire system. These included the three-stage least squares method due to Zellner and Theil (1962), the iterated instrumental variables method based on the work of Lyttkens (1970), Brundy and Jorgenson (1971), and Dhrymes (1971) and the system k-class estimators due to Srivastava (1971) and Savin (1973). Important contributions have also been made in the areas of estimation of simultaneous nonlinear equations (Amemiya, 1983), the seemingly unrelated regression equations (SURE) approach proposed by Zellner (1962), and the simultaneous rational expectations models (see Section 7.1 below).

Interest in estimation of simultaneous equation models coincided with the rise of Keynesian economics in early 1960s, and started to wane with the advent of the rational expectations revolution and its emphasis on the GMM estimation of the structural parameters from the Euler equations (first-order

optimization conditions). See Section 7 below. But, with the rise of the dynamic stochastic general equilibrium models in macroeconometrics, a revival of interest in identification and estimation of nonlinear simultaneous equation models seems quite likely. The recent contribution of Fernandez-Villaverde and Rubio-Ramirez (2005) represents a start in this direction.

4.3 Developments in time series econometrics

While the initiative taken at the Cowles Commission led to a rapid expansion of econometric techniques, the application of these techniques to economic problems was rather slow. This was partly due to a lack of adequate computing facilities at the time. A more fundamental reason was the emphasis of the research at the Cowles Commission on the simultaneity problem almost to the exclusion of other econometric problems. Since the early applications of the correlation analysis to economic data by Yule and Hooker, the serial dependence of economic time series and the problem of nonsense or spurious correlation that it could give rise to had been the single most important factor explaining the profession's scepticism concerning the value of regression analysis in economics. A satisfactory solution to the spurious correlation problem was therefore needed before regression analysis of economic time series could be taken seriously. Research on this topic began in the mid-1940s at the Department of Applied Economics (DAE) in Cambridge, England, as a part of a major investigation into the measurement and analysis of consumers' expenditure in the United Kingdom (see Stone et al., 1954). Although the first steps towards the resolution of the spurious correlation problem had been taken by Aitken (1934–5) and Champernowne (1948), the research in the DAE introduced the problem and its possible solution to the attention of applied economists. Orcutt (1948) studied the autocorrelation pattern of economic time series and showed that most economic time series can be represented by simple autoregressive processes with similar autoregressive coefficients. Subsequently, Cochrane and Orcutt (1949) made the important point that the major consideration in the analysis of stationary time series was the autocorrelation of the error term in the regression equation and not the autocorrelation of the economic time series themselves. In this way they shifted the focus of attention to the autocorrelation of disturbances as the main source of concern. Although, as it turns out, this is a valid conclusion in the case of regression equations with strictly exogenous regressors, in more realistic set-ups where the regressors are weakly exogenous the serial correlation of the regressors is also likely to be of concern in practice. See, for example, Stambaugh (1999).

Another important and related development was the work of Durbin and Watson (1950; 1951) on the method of testing for residual autocorrelation in the classical regression model. The inferential breakthrough for testing serial correlation in the case of observed time-series data had already been achieved by von Neumann (1941; 1942), and by Hart and von Neumann (1942). The contribution of Durbin and Watson was, however, important from a practical viewpoint as it led to a bounds test for residual autocorrelation which could be applied irrespective of the actual values of the regressors. The independence of the critical bounds of the Durbin–Watson statistic from the matrix of the regressors allowed the application of the statistic as a general diagnostic test, the first of its type in econometrics. The contributions of Cochrane and Orcutt and of Durbin and Watson marked the beginning of a new era in the analysis of economic time-series data and laid down the basis of what is now known as the 'time-series econometrics' approach.

5 Consolidation and applications

The work at the Cowles Commission on identification and estimation of the simultaneous equation model and the development of time series techniques paved the way for widespread application of econometric methods to economic and financial problems. This was helped significantly by the rapid expansion of computing facilities, advances in financial and macroeconomic modelling, and the increased availability of economic data-sets, cross section as well as time series.

5.1 Macroeconometric modelling

Inspired by the pioneering work of Tinbergen, Klein (1947; 1950) was the first to construct a macroeconometric model in the tradition of the Cowles Commission. Soon others followed Klein's lead. Over a short space of time macroeconometric models were built for almost every industrialized country, and even for some developing and centrally planned economies. Macroeconometric models became an important tool of *ex ante* forecasting and economic policy analysis, and started to grow in both size and sophistication. The relatively stable economic environment of the 1950s and 1960s was an important factor in the initial success enjoyed by macroeconometric models. The construction and use of large-scale models presented a number of important computational problems, the solution of which was of fundamental significance, not only for the development of macroeconometric modelling but also for econometric practice in general. In this respect advances in computer technology were clearly instrumental, and without them it is difficult to imagine how the complicated computational problems involved in the estimation and simulation of large-scale models could have been solved. The increasing availability of better and faster computers was also instrumental as far as the types of problems studied and the types of solutions offered in the literature were concerned. For example, recent developments in the area of microeconometrics (see Section 10 below) could hardly have been possible if it were not for the very important recent advances in computing facilities.

5.2 Dynamic specification

Other areas where econometrics witnessed significant developments included dynamic specification, latent variables, expectations formation, limited dependent variables, discrete choice models, random coefficient models, disequilibrium models, nonlinear estimation, and the analysis of panel data models. Important advances were also made in the area of Bayesian econometrics, largely thanks to the publication of Zellner's textbook (1971), which built on his earlier work including important papers with George Tiao. The Seminar on Bayesian Inference in Econometrics and Statistics (SBIES) was founded shortly after the publication of the book, and was key in the development and diffusion of Bayesian ideas in econometrics. It was, however, the problem of dynamic specification that initially received the greatest attention. In an important paper, T. Brown (1952) modelled the hypothesis of habit persistence in consumer behaviour by introducing lagged values of consumption expenditures into an otherwise static Keynesian consumption function. This was a significant step towards the incorporation of dynamics in applied econometric research, and allowed the important distinction to be made between the

short-run and the long-run impacts of changes in income on consumption. Soon other researchers followed Brown's lead and employed his autoregressive specification in their empirical work.
The next notable development in the area of dynamic specification was the distributed lag model. Although the idea of distributed lags had been familiar to economists through the pioneering work of Irving Fisher (1930) on the relationship between the nominal interest rate and the expected inflation rate, its application in econometrics was not seriously considered until the mid-1950s. The geometric distributed lag model was used for the first time by Koyck (1954) in a study of investment. Koyck arrived at the geometric distributed lag model via the adaptive expectations hypothesis. This same hypothesis was employed later by Cagan (1956) in a study of demand for money in conditions of hyperinflation, by Friedman (1957) in a study of consumption behaviour and by Nerlove (1958a) in a study of the cobweb phenomenon. The geometric distributed lag model was subsequently generalized by Solow (1960), Jorgenson (1966) and others, and was extensively applied in empirical studies of investment and consumption behaviour. At about the same time Almon (1965) provided a polynomial generalization of I. Fisher's (1937) arithmetic lag distribution which was later extended further by Shiller (1973). Other forms of dynamic specification considered in the literature included the partial adjustment model (Nerlove, 1958b; Eisner and Strotz, 1963) and the multivariate flexible accelerator model (Treadway, 1971) and Sargan's (1964) work on econometric time series analysis which formed the basis of error correction and cointegration analysis that followed next. Following the contributions of Champernowne (1960), Granger and Newbold (1974) and Phillips (1986) the spurious regression problem was better understood, and paved the way for the development of the theory of cointegration. For further details see Section 8.3 below.

5.3 Techniques for short-term forecasting

Concurrent with the development of dynamic modelling in econometrics there was also a resurgence of interest in time-series methods, used primarily in short-term business forecasting. The dominant work in this field was that of Box and Jenkins (1970), who, building on the pioneering works of Yule (1921; 1926), Slutsky (1927), Wold (1938), Whittle (1963) and others, proposed computationally manageable and asymptotically efficient methods for the estimation and forecasting of univariate autoregressive-moving average (ARMA) processes. Time-series models provided an important and relatively simple benchmark for the evaluation of the forecasting accuracy of econometric models, and further highlighted the significance of dynamic specification in the construction of time-series econometric models. Initially univariate time-series models were viewed as mechanical ‘black box’ models with little or no basis in economic theory. Their use was seen primarily to be in short-term forecasting. The potential value of modern time-series methods in econometric research was, however, underlined in the work of Cooper (1972) and Nelson (1972) who demonstrated the good forecasting performance of univariate Box–Jenkins models relative to that of large econometric models. These results raised an important question about the adequacy of large econometric models for forecasting as well as for policy analysis. It was argued that a properly specified structural econometric model should, at least in theory, yield more accurate forecasts than a univariate time-series model. Theoretical justification for this view was provided by Zellner and Palm (1974), followed by Trivedi (1975), Prothero and Wallis (1976), Wallis

(1977) and others. These studies showed that Box–Jenkins models could in fact be derived as univariate final form solutions of linear structural econometric models. In theory, the pure time-series model could always be embodied within the structure of an econometric model and in this sense it did not present a ‘rival’ alternative to econometric modelling. This literature further highlighted the importance of dynamic specification in econometric models and in particular showed that econometric models that are outperformed by simple univariate time-series models most probably suffer from specification errors. The papers in Elliott, Granger and Timmermann (2006) provide excellent reviews of recent developments in economic forecasting techniques.

6 A new phase in the development of econometrics

With the significant changes taking place in the world economic environment in the 1970s, arising largely from the breakdown of the Bretton Woods system and the quadrupling of oil prices, econometrics entered a new phase of its development. Mainstream macroeconometric models built during the 1950s and 1960s, in an era of relative economic stability with stable energy prices and fixed exchange rates, were no longer capable of adequately capturing the economic realities of the 1970s. As a result, not surprisingly, macroeconometric models and the Keynesian theory that underlay them came under severe attack from theoretical as well as from practical viewpoints. While criticisms of Tinbergen's pioneering attempt at macroeconometric modelling were received with great optimism and led to the development of new and sophisticated estimation techniques and larger and more complicated models, the disenchantment with macroeconometric models in 1970s prompted a much more fundamental reappraisal of quantitative modelling as a tool of forecasting and policy analysis.

At a theoretical level it was argued that econometric relations invariably lack the necessary ‘microfoundations’, in the sense that they cannot be consistently derived from the optimizing behaviour of economic agents. At a practical level the Cowles Commission approach to the identification and estimation of simultaneous macroeconometric models was questioned by Lucas and Sargent and by Sims, although from different viewpoints (Lucas, 1976; Lucas and Sargent, 1981; Sims, 1980). There was also a move away from macroeconometric models and towards microeconometric research with greater emphasis on matching of econometrics with individual decisions.

It also became increasingly clear that Tinbergen's paradigm where economic relations were taken as given and provided by ‘economic theorist’ was not adequate. It was rarely the case that economic theory could be relied on for a full specification of the econometric model (Leamer, 1978). The emphasis gradually shifted from estimation and inference based on a given tightly parameterized specification to diagnostic testing, specification searches, model uncertainty, model validation, parameter variations, structural breaks, and semiparametric and nonparametric estimation. The choice of approach often governed by the purpose of the investigation, the nature of the economic application, data availability, computing and software technology.

What follows is a brief overview of some of the important developments. Given space limitations there are inevitably significant gaps. These include the important contributions of Granger (1969), Sims (1972) and Engle, Hendry and Richard (1983) on different concepts of ‘causality’ and ‘exogeneity’, the literature on disequilibrium models (Quandt, 1982; Maddala, 1983; 1986), random coefficient models (Swamy, 1970; Hsiao and Pesaran, 2008, unobserved time series models (Harvey, 1989), count

regression models (Cameron and Trivedi, 1986; 1998), the weak instrument problem (Stock, Wright and Yogo, 2002), small sample theory (Phillips, 1983; Rothenberg, 1984), econometric models of auction pricing (Hendricks and Porter, 1988; Laffont, Ossard and Vuong, 1995).

7 Rational expectations and the Lucas critique

Although the rational expectations hypothesis (REH) was advanced by Muth in 1961, it was not until the early 1970s that it started to have a significant impact on time-series econometrics and on dynamic economic theory in general. What brought the REH into prominence was the work of Lucas (1972; 1973), Sargent (1973), Sargent and Wallace (1975) and others on the new classical explanation of the apparent breakdown of the Phillips curve. The message of the REH for econometrics was clear. By postulating that economic agents form their expectations *endogenously* on the basis of the true model of the economy, and a *correct* understanding of the processes generating exogenous variables of the model, including government policy, the REH raised serious doubts about the invariance of the structural parameters of the mainstream macroeconometric models in the face of changes in government policy. This was highlighted in Lucas's critique of macroeconometric policy evaluation. By means of simple examples Lucas (1976) showed that in models with rational expectations the parameters of the decision rules of economic agents, such as consumption or investment functions, are usually a mixture of the parameters of the agents' objective functions and of the stochastic processes they face as historically given. Therefore, Lucas argued, there is no reason to believe that the 'structure' of the decision rules (or economic relations) would remain invariant under a policy intervention. The implication of the Lucas critique for econometric research was not, however, that policy evaluation could not be done, but rather than the traditional econometric models and methods were not suitable for this purpose. What was required was a separation of the parameters of the policy rule from those of the economic model. Only when these parameters could be identified separately given the knowledge of the joint probability distribution of the variables (both policy and non-policy variables) would it be possible to carry out an econometric analysis of alternative policy options.
There have been a number of reactions to the advent of the rational expectations hypothesis and the Lucas critique that accompanied it.

7.1 Model consistent expectations

The least controversial reaction has been the adoption of the REH as one of several possible expectations formation hypotheses in an otherwise conventional macroeconometric model containing expectational variables. In this context the REH, by imposing the appropriate cross-equation parametric restrictions, ensures that 'expectations' and 'forecasts' generated by the model are consistent. In this approach the REH is regarded as a convenient and effective method of imposing cross-equation parametric restrictions on time series econometric models, and is best viewed as the 'model-consistent' expectations hypothesis. There is now a sizeable literature on solution, identification, and estimation of linear RE models. The canonical form of RE models with forward and backward components is given by

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{B}E(\mathbf{y}_{t+1}|F_t) + \mathbf{w}_t,$$

where $\mathbf{y}_t$ is a vector of endogenous variables, $E(\cdot|F_t)$ is the expectations operator, F_t the publicly available information at time t, and $\mathbf{w}_t$ is a vector of forcing variables. For example, log-linearized version of dynamic general equilibrium models (to be discussed) can all be written as a special case of this equation with plenty of restrictions on the coefficient matrices A and B. In the typical case where $\mathbf{w}_t$ are serially uncorrelated and the solution of the RE model can be assumed to be unique, the RE solution reduces to the vector autoregression (VAR)

$$\mathbf{y}_t = \Phi\mathbf{y}_{t-1} + \mathbf{G}\mathbf{w}_t,$$

where Φ and $\mathbf{G}$ are given in terms of the structural parameters:

$$\mathbf{B}\Phi^2 - \Phi + \mathbf{A} = 0, \text{ and } \mathbf{G} = (\mathbf{I} - \mathbf{B}\Phi)^{-1}.$$

The solution of the RE model can, therefore, be viewed as a restricted form of VAR popularized in econometrics by Sims (1980) as a response in macroeconometric modelling to the rational expectations revolution. The nature of restrictions is determined by the particular dependence of A and B on a few ‘deep’ or structural parameters. For general discussion of solution of RE models see, for example, Broze, Gouriéroux and Szafarz (1985) and Binder and Pesaran (1995). For studies of identification and estimation of linear RE models see, for example, Hansen and Sargent (1980), Wallis (1980), Wickens (1982) and Pesaran (1981; 1987b). These studies show how the standard econometric methods can in principle be adapted to the econometric analysis of rational expectations models.

7.2 Detection and modelling of structural breaks

Another reaction to the Lucas critique has been to treat the problem of ‘structural change’ emphasized by Lucas as one more potential econometric ‘problem’. Clements and Hendry (1998; 1999) provide a taxonomy of factors behind structural breaks and forecast failures. Stock and Watson (1996) provide extensive evidence of structural break in macroeconomic time series. It is argued that structural change can result from many factors and need not be associated solely with intended or expected changes in policy. The econometric lesson has been to pay attention to possible breaks in economic relations. There now exists a large body of work on testing for structural change, detection of breaks (single as well as multiple), and modelling of break processes by means of piece-wise linear or non-linear dynamic models

(Chow, 1960; Brown, Durbin and Evans, 1975; Nyblom, 1989; Andrews, 1993; Andrews and Ploberger, 1994; Bai and Perron, 1998; Pesaran and Timmermann, 2005b; 2007. See also the surveys by Stock, 1994; Clements and Hendry, 2006). The implications of breaks for short-term and long-term forecasting have also begun to be addressed (McCulloch, and Tsay, 1993; Koop and Potter, 2004a; 2004b; Pesaran, Pettenuzzo and Timmermann, 2006).

8 VAR macroeconometrics

8.1 Unrestricted VARs

The Lucas critique of mainstream macroeconometric modelling also led some econometricians, notably Sims (1980; 1982), to doubt the validity of the Cowles Commission style of achieving identification in econometric models. Sims focused his critique on macroeconometric models with a vector autoregressive (VAR) specification, which was relatively simple to estimate; and its use soon became prevalent in macroeconometric analysis. The view that economic theory cannot be relied on to yield identification of structural models was not new and had been emphasized in the past, for example, by Liu (1960). Sims took this viewpoint a step further and argued that in presence of rational expectations a priori knowledge of lag lengths is indispensable for identification, even when we have distinct strictly exogenous variables shifting supply and demand schedules (Sims, 1980, p. 7). While it is true that the REH complicates the necessary conditions for the identification of structural models, the basic issue in the debate over identification still centres on the validity of the classical dichotomy between exogenous and endogenous variables (Pesaran, 1981). In the context of closed-economy macroeconometric models where all variables are treated as endogenous, other forms of identification of the structure will be required. Initially, Sims suggested a recursive identification approach where the matrix of contemporaneous effects was assumed to be lower (upper) triangular and the structural shocks orthogonal. Other non-recursive identification schemes soon followed.

8.2 Structural VARs

One prominent example was the identification scheme developed in Blanchard and Quah (1989), who distinguished between permanent and transitory shocks and attempted to identify the structural models through long-run restrictions. For example, Blanchard and Quah argued that the effect of a demand shock on real output should be temporary (that is, it should have a zero long-run impact), while a supply shock should have a permanent effect. This approach is known as ‘structural VAR’ (SVAR) and has been used extensively in the literature. It continues to assume that structural shocks are orthogonal, but uses a mixture of short-run and long-run restrictions to identify the structural model. In their work Blanchard and Quah considered a bivariate VAR model in real output and unemployment. They assumed real output to be integrated of order 1, or $I(1)$, and viewed unemployment as an $I(0)$, or a stationary variable. This allowed them to associate the shock to one of the equations as permanent, and the shock to the other equation as transitory. In more general settings, such as the one analysed by Gali (1992) and Wickens and Motto (2001), where there are m endogenous variables and r long-run or

cointegrating relations, the SVAR approach provides $m(m-r)$ restrictions which are not sufficient to fully identify the model, unless $m = 2$ and $r = 1$ which is the simple bivariate model considered by Blanchard and Quah (Pagan and Pesaran, 2007). In most applications additional short-term restrictions are required. More recently, attempts have also been made to identify structural shocks by means of qualitative restrictions, such as sign restrictions. Notable examples include Canova and de Nicolo (2002), Uhlig (2005) and Peersman (2005).

The focus of the SVAR literature has been on impulse response analysis and forecast error variance decomposition, with the aim of estimating the time profile of the effects of monetary policy, oil price or technology shocks on output and inflation, and deriving the relative importance of these shocks as possible explanations of forecast error variances at different horizons. Typically such analysis is carried out with respect to a single model specification, and at most only parameter uncertainty is taken into account (Kilian, 1998). More recently the problem of model uncertainty and its implications for impulse response analysis and forecasting have been recognized. Bayesian and classical approaches to model and parameter uncertainty have been considered. Initially, Bayesian VAR models were developed for use in forecasting as an effective shrinkage procedure in the case of high-dimensional VAR models (Doan, Litterman and Sims, 1984; Litterman, 1985). The problem of model uncertainty in cointegrating VARs has been addressed in Garratt et al. (2003b; 2006), and Strachan and van Dijk (2006).

8.3 Structural cointegrating VARs

This approach provides the SVAR with the decomposition of shocks into permanent and transitory and gives economic content to the long-run or cointegrating relations that underlie the transitory components. In the simple example of Blanchard and Quah this task is trivially achieved by assuming real output to be I(1) and the unemployment rate to be an I(0) variable. To have shocks with permanent effects some of the variables in the VAR must be non-stationary. This provides a natural link between the SVAR and the unit root and cointegration literature. Identification of the cointegrating relations can be achieved by recourse to economic theory, solvency or arbitrage conditions (Garratt et al., 2003a). Also there are often long-run over-identifying restrictions that can be tested. Once identified and empirically validated, the long-run relations can be embodied within a VAR structure, and the resultant structural vector error correction model identified using theory-based short-run restrictions. The structural shocks can be decomposed into permanent and temporary components using either the multivariate version of the Beveridge and Nelson (1981) decompositions, or the one more recently proposed by Garratt, Robertson and Wright (2006).

Two or more variables are said to be cointegrated if they are individually integrated (or have a random walk component), but there exists a linear combination of them which is stationary. The concept of cointegration was first introduced by Granger (1986) and more formally developed in Engle and Granger (1987). Rigorous statistical treatments followed in the papers by Johansen (1988; 1991) and Phillips (1991). Many further developments and extensions have taken place with reviews provided in Johansen (1995), Juselius (2006) and Garratt et al. (2006). The related unit root literature is reviewed by Stock (1994) and Phillips and Xiao (1998).

8.4 Macroeconometric models with microeconomic foundations

For policy analysis macroeconometric models need to be based on decisions by individual households, firms and governments. This is a daunting undertaking and can be achieved only by gross simplification of the complex economic interconnections that exists across millions of decision-makers worldwide. The dynamic stochastic general equilibrium (DSGE) modelling approach attempts to implement this task by focusing on optimal decisions of a few representative agents operating with rational expectations under complete learning. Initially, DSGE models were small and assumed complete markets with instantaneous price adjustments, and as a result did not fit the macroeconomic time series (Kim and Pagan, 1995). More recently, Smets and Wouters (2003) have shown that DSGE models with sticky prices and wages along the lines developed by Christiano, Eichenbaum and Evans (2005) are sufficiently rich to match most of the statistical features of the main macroeconomic time series. Moreover, by applying Bayesian estimation techniques, these authors have shown that even relatively large models can be estimated as a system. Bayesian DSGE models have also shown to perform reasonably well in forecasting as compared with standard and Bayesian vector autoregressions. It is also possible to incorporate long-run cointegrating relations within Bayesian DSGE models. The problems of parameter and model uncertainty can also be readily accommodated using data-coherent DSGE models. Other extensions of the DSGE models to allow for learning, regime switches, time variations in shock variances, asset prices, and multi-country interactions are likely to enhance their policy relevance (Del Negro and Schorfheide, 2004; Del Negro et al., 2005; An and Schorfheide, 2007; Pesaran and Smith, 2006). Further progress will also be welcome in the area of macroeconomic policy analysis under model uncertainty, and robust policymaking (Brock and Durlauf, 2006; Hansen and Sargent, 2007).

9 Model and forecast evaluation

While in the 1950s and 1960s research in econometrics was primarily concerned with the identification and estimation of econometric models, the dissatisfaction with econometrics during the 1970s caused a shift of focus from problems of estimation to those of model evaluation and testing. This shift has been part of a concerted effort to restore confidence in econometrics, and has received attention from Bayesian as well as classical viewpoints. Both these views reject the ‘axiom of correct specification’ which lies at the basis of most traditional econometric practices, but they differ markedly as how best to proceed.

It is generally agreed, by Bayesians as well as by non-Bayesians, that model evaluation involves considerations other than the examination of the statistical properties of the models, and personal judgements inevitably enter the evaluation process. Models must meet multiple criteria which are often in conflict. They should be relevant in the sense that they ought to be capable of answering the questions for which they are constructed. They should be consistent with the accounting and/or theoretical structure within which they operate. Finally, they should provide adequate representations of the aspects of reality with which they are concerned. These criteria and their interaction are discussed in Pesaran and Smith (1985b). More detailed breakdowns of the criteria of model evaluation can be found in Hendry and Richard (1982) and McAleer, Pagan, and Volker (1985). In econometrics it is, however, the criterion of ‘adequacy’ which is emphasized, often at the expense of relevance and consistency.

The issue of model adequacy in mainstream econometrics is approached either as a model selection problem or as a problem in statistical inference whereby the hypothesis of interest is tested against general or specific alternatives. The use of absolute criteria such as measures of fit/parsimony or formal Bayesian analysis based on posterior odds are notable examples of model selection procedures, while likelihood ratio, Wald and Lagrange multiplier tests of nested hypotheses and Cox's centred log-likelihood ratio tests of non-nested hypotheses are examples of the latter approach. The distinction between these two general approaches basically stems from the way alternative models are treated. In the case of model selection (or model discrimination) all the models under consideration enjoy the same status and the investigator is not committed a priori to any one of the alternatives. The aim is to choose the model which is likely to perform best with respect to a particular loss function. By contrast, in the hypothesis-testing framework the null hypothesis (or the maintained model) is treated differently from the remaining hypotheses (or models). One important feature of the model-selection strategy is that its application always leads to one model being chosen in preference to other models. But, in the case of hypothesis testing, rejection of all the models under consideration is not ruled out when the models are non-nested. A more detailed discussion of this point is given in Pesaran and Deaton (1978).

Broadly speaking, classical approaches to the problem of model adequacy can be classified depending on how specific the alternative hypotheses are. These are the *general specification tests*, *the diagnostic tests*, and the *non-nested tests*. The first of these, pioneered by Durbin (1954) and introduced in econometrics by Ramsey (1969), Wu (1973), Hausman (1978), and subsequently developed further by White (1981; 1982) and Hansen (1982), are designed for circumstances where the nature of the alternative hypothesis is kept (sometimes intentionally) rather vague, the purpose being to test the null against a *broad* class of alternatives. (The pioneering contribution of Durbin, 1954, in this area has been documented by Nakamura and Nakamura, 1981.) Important examples of general specification tests are Ramsey's regression specification error test (RESET) for omitted variables and/or misspecified functional forms, and the Durbin–Hausman–Wu test of misspecification in the context of measurement error models and/or simultaneous equation models. Such general specification tests are particularly useful in the preliminary stages of the modelling exercise.

In the case of diagnostic tests, the model under consideration (viewed as the null hypothesis) is tested against more specific alternatives by embedding it within a general model. Diagnostic tests can then be constructed using the likelihood ratio, Wald or Lagrange multiplier (LM) principles to test for parametric restrictions imposed on the general model. The application of the LM principle to econometric problems is reviewed in the papers by Breusch and Pagan (1980), Godfrey and Wickens (1982), and Engle (1984). An excellent review is provided in Godfrey (1988). Examples of the restrictions that may be of interest as diagnostic checks of model adequacy include zero restrictions, parameter stability, serial correlation, heteroskedasticity, functional forms, and normality of errors. The distinction made here between diagnostic tests and general specification tests is more apparent than real. In practice some diagnostic tests such as tests for serial correlation can also be viewed as a general test of specification. Nevertheless, the distinction helps to focus attention on the purpose behind the tests and the direction along which high power is sought.

The need for non-nested tests arises when the models under consideration belong to separate parametric families in the sense that no single model can be obtained from the others by means of a suitable limiting process. This situation, which is particularly prevalent in econometric research, may arise when models

differ with respect to their theoretical underpinnings and/or their auxiliary assumptions. Unlike the general specification tests and diagnostic tests, the application of non-nested tests is appropriate when specific but rival hypotheses for the explanation of the same economic phenomenon have been advanced. Although non-nested tests can also be used as general specification tests, they are designed primarily to have high power against specific models that are seriously entertained in the literature. Building on the pioneering work of Cox (1961; 1962), a number of such tests for single equation models and systems of simultaneous equations have been proposed (Pesaran and Weeks, 2001).

The use of statistical tests in econometrics, however, is not a straightforward matter and in most applications does not admit of a clear-cut interpretation. This is especially so in circumstances where test statistics are used not only for checking the adequacy of a *given* model but also as guides to model construction. Such a process of model construction involves specification searches of the type emphasized by Leamer (1978) and presents insurmountable pre-test problems which in general tend to produce econometric models whose ‘adequacy’ is more apparent than real. As a result, in evaluating econometric models less reliance should be placed on those indices of model adequacy that are used as guides to model construction, and more emphasis should be given to the performance of models over other data-sets and against rival models.

A closer link between model evaluation and the underlying decision problem is also needed. Granger and Pesaran (2000a; 2000b) discuss this problem in the context of forecast evaluation. A recent survey of forecast evaluation literature can be found in West (2006). Pesaran and Skouras (2002) provide a review from a decision-theoretic perspective.

The subjective Bayesian approach to the treatment of several models begins by assigning a prior probability to each model, with the prior probabilities summing to 1. Since each model is already endowed with a prior probability distribution for its parameters and for the probability distribution of observable data conditional on its parameters, there is then a complete probability distribution over the space of models, parameters, and observable data. (No particular problems arise from non-nesting of models in this framework.) This probability space can then be augmented with the distribution of an object or vector of objects of interest. For example, in a macroeconomic policy setting the models could include VARs, DSGEs and traditional large-scale macroeconomic models, and the vector of interest might include future output growth, interest rates, inflation and unemployment, whose distribution is implied by each of the models considered. Implicit in this formulation is the conditional distribution of the vector of interest conditional on the observed data. Technically, this requires the integration (or marginalization) of parameters in each model as well as the models themselves. As a practical matter this usually proceeds by first computing the probability of each model conditional on the data, and then using these probabilities as weights in averaging the posterior distribution of the vector of interest in each model. It is not necessary to choose one particular model, and indeed to do so would be suboptimal. The ability to actually carry out this simultaneous consideration of multiple models has been enhanced greatly by recent developments in simulation methods, surveyed in Section 16 below; recent texts by Koop (2003), Lancaster (2004) and Geweke (2005) provide technical details. Geweke and Whiteman (2006) specifically outline these methods in the context of economic forecasting.

10 Microeconometrics: an overview

Partly as a response to the dissatisfaction with macroeconometric time-series research and partly in view of the increasing availability of micro data and computing facilities, since the mid-1980s significant advances have been made in the analysis of micro data. Important micro data-sets have become available on households and firms especially in the United States in such areas as housing, transportation, labour markets and energy. These data sets include various longitudinal surveys (for example, University of Michigan Panel Study of Income Dynamics, and Ohio State National Longitudinal Study Surveys), cross-sectional surveys of family expenditures, population and labour force surveys. This increasing availability of micro-data, while opening up new possibilities for analysis, has also raised a number of new and interesting econometric issues primarily originating from the nature of the data. The errors of measurement are likely to be important in the case of some micro data-sets. The problem of the heterogeneity of economic agents at the micro level cannot be assumed away as readily as is usually done in the case of macro data by appealing to the idea of a ‘representative’ firm or a ‘representative’ household.

The nature of micro data, often being qualitative or limited to a particular range of variations, has also called for new econometric models and techniques. Examples include categorical survey responses (‘up’, ‘same’ or ‘down’), and censored or truncated observations. The models and issues considered in the microeconometric literature are wide ranging and include fixed and random effect panel data models (for example, Mundlak, 1961; 1978), logit and probit models and their multinominal extensions, discrete choice or quantal response models (Manski and McFadden, 1981), continuous time duration models (Heckman and Singer, 1984), and microeconometric models of count data (Hausman, Hall and Griliches, 1984; Cameron and Trivedi, 1986).

The fixed or random effect models provide the basic statistical framework and will be discussed in more detailed below. Discrete choice models are based on an explicit characterization of the choice process and arise when individual decision makers are faced with a finite number of alternatives to choose from. Examples of discrete choice models include transportation mode choice (Domenich and McFadden, 1975), labour force participation (Heckman and Willis, 1977), occupation choice (Boskin, 1974), job or firm location (Duncan 1980), and models with neighbourhood effects (Brock and Durlauf, 2002).

Limited dependent variables models are commonly encountered in the analysis of survey data and are usually categorized into truncated regression models and censored regression models. If all observations on the dependent as well as on the exogenous variables are lost when the dependent variable falls outside a specified range, the model is called *truncated*, and, if only observations on the dependent variable are lost, it is called *censored*. The literature on censored and truncated regression models is vast and overlaps with developments in other disciplines, particularly in biometrics and engineering. Maddala (1983, ch. 6) provides a survey.

The censored regression model was first introduced into economics by Tobin (1958) in his pioneering study of household expenditure on durable goods, where he explicitly allowed for the fact that the dependent variable, namely, the expenditure on durables, cannot be negative. The model suggested by Tobin and its various generalizations are known in economics as Tobit models and are surveyed in detail by Amemiya (1984), and more recently in Cameron and Trivedi (2005, ch. 14). Continuous time duration models, also known as survival models, have been used in analysis of unemployment duration, the period of time spent between jobs, durability of marriage, and so on. Application of survival models to analyse economic data raises a number of important issues resulting primarily from the non-controlled

experimental nature of economic observations, limited sample sizes (that is, time periods), and the heterogeneous nature of the economic environment within which agents operate. These issues are clearly not confined to duration models and are also present in the case of other microeconometric investigations that are based on time series or cross-section or panel data.

Partly in response to the uncertainties inherent in econometric results based on non-experimental data, there has also been a significant move towards social experimentation, and experimental economics in general. A social experiment aims at isolating the effects of a policy change (or a treatment effect) by comparing the consequences of an exogenous variation in the economic environment of a set of experimental subjects known as the ‘treatment’ group with those of a ‘control’ group that have not been subject to the change. The basic idea goes back to the early work of R.A. Fisher (1928) on randomized trials, and has been applied extensively in agricultural and biomedical research. The case for social experimentation in economics is discussed in Burtless (1995). Hausman and Wise (1985) and Heckman and Smith (1995) consider a number of actual social experiments carried out in the United States, and discuss their scope and limitations.

Experimental economics tries to avoid some of the limitations of working with observations obtained from natural or social experiments by using data from laboratory experiments to test economic theories by fixing some of the factors and identifying the effects of other factors in a way that allows *ceteris paribus* comparisons. A wide range of topics and issues are covered in this literature, such as individual choice behaviour, bargaining, provision of public goods, theories of learning, auction markets, and behavioural finance. A comprehensive review of major areas of experimental research in economics is provided in Kagel and Roth (1995).

These developments have posed new problems and challenges in the areas of experimental design, statistical methods and policy analysis. Another important aspect of recent developments in microeconometric literature relates to the use of microanalytic simulation models for policy analysis and evaluation to reform packages in areas such as health care, taxation, social security systems, and transportation networks. Cameron and Trivedi (2005) review the recent developments in methods and application of microeconometrics. Some of these topics will be discussed in more detail below.

11 Econometrics of panel data

Panel data models are used in many areas of econometrics, although initially they were developed primarily for the analysis of micro behaviour, and focused on panels formed from cross-section of N individual households or firms surveyed for T successive time periods. These types of panels are often refereed to as ‘micropanels’. In social and behavioural sciences they are also known as longitudinal data or panels. The literature on micro-panels typically takes N to be quite large (in hundreds) and T rather small, often less than ten. But more recently, with the increasing availability of financial and macroeconomic data, analyses of panels where both N and T are relatively large have also been considered. Examples of such data-sets include time series of company data from Datastream, country data from International Financial Statistics or the Penn World Table, and county and state data from national statistical offices. There are also pseudo panels of firms and consumers composed of repeated cross sections that cover cross-section units that are not necessarily identical but are observed over relatively long time periods. Since the available cross-section observations do not (necessarily) relate to

the same individual unit, some form of grouping of the cross-section units is needed. Once the grouping criteria are set, the estimation can proceed using fixed effects estimation applied to group averages if the number of observations per group is sufficiently large; otherwise possible measurement errors of the group averages also need to be taken into account. Deaton (1985) pioneered the econometric analysis of pseudo panels. Verbeek (2008) provides a recent review.

Use of panels can enhance the power of empirical analysis and allows estimation of parameters that might not have been identified using the time or the cross-section dimensions alone. These benefits come at a cost. In the case of linear panel data models with a short time span the increased power is usually achieved under assumptions of parameter homogeneity and error cross-section independence. Short panels with autocorrelated disturbances also pose a new identification problem, namely, how to distinguished between dynamics and state dependence (Arellano, 2003, ch. 5). In panels with fixed effects the homogeneity assumption is relaxed somewhat by allowing the intercepts in the panel regressions to vary freely over the cross-section units, but continues to maintain the error cross-section independence assumption. The random coefficient specification of Swamy (1970) further relaxes the slope homogeneity assumption, and represents an important generalization of the random effects model (Hsiao and Pesaran, 2007). In micro-panels where T is small cross-section dependence can be dealt with if it can be attributed to spatial (economic or geographic) effects. Anselin (1988) and Anselin, Le Gallo and Jayet (2007) provide surveys of the literature on spatial econometrics. A number of studies have also used measures such as trade or capital flows to capture economic distance, as in Conley and Topa (2002), Conley and Dupor (2003), and Pesaran, Schuermann and Weiner (2004).

Allowing for dynamics in panels with fixed effects also presents additional difficulties; for example, the standard within-group estimator will be inconsistent unless $T \to \infty$ (Nickell, 1981). In linear dynamic panels the incidental parameter problem (the unobserved heterogeneity) can be resolved by first differencing the model and then estimating the resultant first-differenced specification by instrumental variables or by the method of transformed likelihood (Anderson and Hsiao, 1981; 1982; Holtz-Eakin, Newey and Rosen, 1988; Arellano and Bond, 1991; Hsiao, Pesaran and Tahmiscioglu, 2002). A similar procedure can also be followed in the case of short T panel VARs (Binder, Hsiao and Pesaran, 2005). But other approaches are needed for nonlinear panel data models. See, for example, Honoré and Kyriazidou (2000) and review of the literature on nonlinear panels in Arellano and Honoré (2001).

Relaxing the assumption of slope homogeneity in dynamic panels is also problematic, and neglecting to take account of slope heterogeneity will lead to inconsistent estimators. In the presence of slope heterogeneity Pesaran and Smith (1995) show that the within-group estimator remains inconsistent even if both N and $T \to \infty$. A Bayesian approach to estimation of micro dynamic panels with random slope coefficients is proposed in Hsiao, Pesaran and Tahmiscioglu (1999).

To deal with general dynamic specifications, possible slope heterogeneity and error cross-section dependence, large T and N panels are required. In the case of such large panels it is possible to allow for richer dynamics and parameter heterogeneity. Cross-section dependence of errors can also be dealt with using residual common factor structures. These extensions are particularly relevant to the analysis of purchasing power parity hypothesis (O'Connell, 1998; Imbs et al., 2005; Pedroni, 2001; Smith et al., 2004), output convergence (Durlauf, Johnson and Temple, 2005; Pesaran, 2007b), the Fisher effect (Westerlund, 2005), house price convergence (Holly, Pesaran and Yamagata, 2006), regional migration

(Fachin, 2006), and uncovered interest parity (Moon and Perron, 2007). The econometric methods developed for large panels has to take into account the relationship between the increasing number of time periods and cross-section units (Phillips and Moon, 1999). The relative expansion rates of N and T could have important consequences for the asymptotic and small sample properties of the panel estimators and tests. This is because fixed T estimation bias tend to magnify with increases in the cross-section dimension, and it is important that any bias in the T dimension is corrected in such a way that its overall impact disappears as both N and $T \to \infty$, jointly.

The first generation panel unit root tests proposed, for example, by Levin, Lin and Chu (2002) and Im, Pesaran and Shin (2003) allowed for parameter heterogeneity but assumed errors were cross-sectionally independent. More recently, panel unit root tests that allow for error cross-section dependence have been proposed by Bai and Ng (2004), Moon and Perron (2004) and Pesaran (2007a). As compared with panel unit root tests, the analysis of cointegration in panels is still at an early stage of its development. So far the focus of the panel cointegration literature has been on residual-based approaches, although there has been a number of attempts at the development of system approaches as well (Pedroni, 2004). But once cointegration is established the long-run parameters can be estimated efficiently using techniques similar to the ones proposed in the case of single time-series models. These estimation techniques can also be modified to allow for error cross-section dependence (Pesaran, 2007a). Surveys of the panel unit root and cointegration literature are provided by Banerjee (1999), Baltagi and Kao (2000), Choi (2006) and Breitung and Pesaran (2008).

The micro and macro panel literature is vast and growing. For the analysis of many economic problems, further progress is needed in the analysis of nonlinear panels, testing and modelling of error cross-section dependence, dynamics, and neglected heterogeneity. For general reviews of panel data econometrics, see Arellano (2003), Baltagi (2005), Hsiao (2003) and Wooldridge (2002).

12 Nonparametric and semiparametric estimation

Much empirical research is concerned with estimating conditional mean, median, or hazard functions. For example, a wage equation gives the mean, median or, possibly, some other quantile of wages of employed individuals conditional on characteristics such as years of work experience and education. A hedonic price function gives the mean price of a good conditional on its characteristics. The function of interest is rarely known a priori and must be estimated from data on the relevant variables. For example, a wage equation is estimated from data on the wages, experience, education and, possibly, other characteristics of individuals. Economic theory rarely gives useful guidance on the form (or shape) of a conditional mean, median, or hazard function. Consequently, the form of the function must either be assumed or inferred through the estimation procedure.

The most frequently used estimation methods assume that the function of interest is known up to a set of constant parameters that can be estimated from data. Models in which the only unknown quantities are a finite set of constant parameters are called ‘parametric’. A linear model that is estimated by ordinary least squares is a familiar and frequently used example of a parametric model. Indeed, linear models and ordinary least squares have been the workhorses of applied econometrics since its inception. It is not difficult to see why. Linear models and ordinary least squares are easy to work with both analytically and computationally, and the estimation results are easy to interpret. Other examples of widely used

parametric models are binary logit and probit models if the dependent variable is binary (for example, an indicator of whether an individual is employed or whether a commuter uses automobile or public transit for a trip to work) and the Weibull hazard model if the dependent variable is a duration (for example, the duration of a spell of employment or unemployment).

Although parametric models are easy to work with, they are rarely justified by theoretical or other a priori considerations and often fit the available data badly. Horowitz (2001), Horowitz and Savin (2001), Horowitz and Lee (2002), and Pagan and Ullah (1999) provide examples. The examples also show that conclusions drawn from a convenient but incorrectly specified model can be very misleading. Of course, applied econometricians are aware of the problem of specification error. Many investigators attempt to deal with it by carrying out a specification search in which several different models are estimated and conclusions are based on the one that appears to fit the data best. Specification searches may be unavoidable in some applications, but they have many undesirable properties. There is no guarantee that a specification search will include the correct model or a good approximation to it. If the search includes the correct model, there is no guarantee that it will be selected by the investigator's model selection criteria. Moreover, the search process invalidates the statistical theory on which inference is based.

Given this situation, it is reasonable to ask whether conditional mean and other functions of interest in applications can be estimated nonparametrically, that is, without making a priori assumptions about their functional forms. The answer is clearly ‘yes’ in a model whose explanatory variables are all discrete. If the explanatory variables are discrete, then each set of values of these variables defines a data cell. One can estimate the conditional mean of the dependent variable by averaging its values within each cell. Similarly, one can estimate the conditional median cell by cell.

If the explanatory variables are continuous, they cannot be grouped into cells. Nonetheless, it is possible to estimate conditional mean and median functions that satisfy mild smoothness conditions without making a priori assumptions about their shapes. Techniques for doing this have been developed mainly in statistics, beginning with Nadaraya's (1964) and Watson's (1964) nonparametric estimator of a conditional mean function. The Nadaraya–Watson estimator, which is also called a kernel estimator, is a weighted average of the observed values of the dependent variable. More specifically, suppose that the dependent variable is Y, the explanatory variable is X, and the data consist of observations $\{Y_i, X_i; i = 1, \ldots, n\}$. Then the Nadaraya–Watson estimator of the mean of Y at $X = x$ is a weighted average of the Y_i's. Y_i's corresponding to X_i's that are close to x get more weight than do Y_i's corresponding to X_i's that are far from x. The statistical properties of the Nadaraya–Watson estimator have been extensively investigated for both cross-sectional and time-series data, and the estimator has been widely used in applications. For example, Blundell, Browning and Crawford (2003) used kernel estimates of Engel curves in an investigation of the consistency of household-level data and revealed preference theory. Hausman and Newey (1995) used kernel estimates of demand functions to estimate the equivalent variation for changes in gasoline prices and the deadweight losses associated with increases in gasoline taxes. Kernel-based methods have also been developed for estimating conditional quantile and hazard functions.

There are other important nonparametric methods for estimating conditional mean functions. Local linear estimation and series or sieve estimation are especially useful in applications. Local linear estimation consists of estimating the mean of Y at $X = x$ by using a form of weighted least squares to fit a linear model to the data. The weights are such that observations (Y_i, X_i) for which X_i is close to x

receive more weight than do observations for which X_i is far from x. In comparison with the Nadaraya–Watson estimator, local linear estimation has important advantages relating to bias and behaviour near the boundaries of the data. These are discussed in the book by Fan and Gijbels (1996), among other places.

A series estimator begins by expressing the true conditional mean (or quantile) function as an infinite series expansion using basis functions such as sines and cosines, orthogonal polynomials, or splines. The coefficients of a truncated version of the series are then estimated by ordinary least squares. The statistical properties of series estimators are described by Newey (1997). Hausman and Newey (1995) give an example of their use in an economic application.

Nonparametric models and estimates essentially eliminate the possibility of misspecification of a conditional mean or quantile function (that is, they consistently estimate the true function), but they have important disadvantages that limit their usefulness in applied econometrics. One important problem is that the precision of a nonparametric estimator decreases rapidly as the dimension of the explanatory variable X increases. This phenomenon is called the ‘curse of dimensionality’. It can be understood most easily by considering the case in which the explanatory variables are all discrete. Suppose the data contain 500 observations of Y and X. Suppose, further, that X is a K-component vector and that each component can take five different values. Then the values of X generate 5^k cells. If $K = 4$, which is not unusual in applied econometrics, then there are 625 cells, or more cells than observations. Thus, estimates of the conditional mean function are likely to be very imprecise for most cells because they will contain few observations. Moreover, there will be at least 125 cells that contain no data and, consequently, for which the conditional mean function cannot be estimated at all. It has been proved that the curse of dimensionality is unavoidable in nonparametric estimation. As a result of it, impracticably large samples are usually needed to obtain acceptable estimation precision if X is multidimensional.

Another problem is that nonparametric estimates can be difficult to display, communicate, and interpret when X is multidimensional. Nonparametric estimates do not have simple analytic forms. If X is one- or two-dimensional, then the estimate of the function of interest can be displayed graphically, but only reduced-dimension projections can be displayed when X has three or more components. Many such displays and much skill in interpreting them can be needed to fully convey and comprehend the shape of an estimate.

A further problem with nonparametric estimation is that it does not permit extrapolation. For example, in the case of a conditional mean function it does not provide predictions of the mean of Y at values of x that are outside of the range of the data on X. This is a serious drawback in policy analysis and forecasting, where it is often important to predict what might happen under conditions that do not exist in the available data. Finally, in nonparametric estimation it can be difficult to impose restrictions suggested by economic or other theory. Matzkin (1994) discusses this issue.

The problems of nonparametric estimation have led to the development of so-called semiparametric methods that offer a compromise between parametric and nonparametric estimation. Semiparametric methods make assumptions about functional form that are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing (though not eliminating) the possibility of specification error. Semiparametric methods permit greater estimation precision than do nonparametric methods when X is multidimensional. Semiparametric estimation results are usually easier to display and interpret than are nonparametric ones, and provide limited

capabilities for extrapolation.

In econometrics, semiparametric estimation began with Manski's (1975; 1985) and Cosslett's (1983) work on estimating discrete-choice random-utility models. McFadden had introduced multinomial logit random utility models. These models assume that the random components of the utility function are independently and identically distributed with the Type I extreme value distribution. (The Type I extreme value distribution and density functions are defined, for example, in eqs (3.1) and (3.2) Maddala, 1983, p. 60.) The resulting choice model is analytically simple but has properties that are undesirable in many applications (for example, the well-known independence-of-irrelevant-alternatives property). Moreover, estimators based on logit models are inconsistent if the distribution of the random components of utility is not Type I extreme value. Manski (1975; 1985) and Cosslett (1983) proposed estimators that do not require a priori knowledge of this distribution. Powell's (1984; 1986) least absolute deviations estimator for censored regression models is another early contribution to econometric research on semiparametric estimation. This estimator was motivated by the observation that estimators of (parametric) Tobit models are inconsistent if the underlying normality assumption is incorrect. Powell's estimator is consistent under very weak distributional assumptions.

Semiparametric estimation has continued to be an active area of econometric research. Semiparametric estimators have been developed for a wide variety of additive, index, partially linear, and hazard models, among others. These estimators all reduce the effective dimension of the estimation problem and overcome the curse of dimensionality by making assumptions that are stronger than those of fully nonparametric estimation but weaker than those of a parametric model. The stronger assumptions also give the models limited extrapolation capabilities. Of course, these benefits come at the price of increased risk of specification error, but the risk is smaller than with simple parametric models. This is because semiparametric models make weaker assumptions than do parametric models, and contain simple parametric models as special cases.

Semiparametric estimation is also an important research field in statistics, and it has led to much interaction between statisticians and econometricians. The early statistics and biostatistics research that is relevant to econometrics was focused on survival (duration) models. Cox's (1972) proportional hazards model and the Buckley and James (1979) estimator for censored regression models are two early examples of this line of research. Somewhat later, C. Stone (1985) showed that a nonparametric additive model can overcome the curse of dimensionality. Since then, statisticians have contributed actively to research on the same classes of semiparametric models that econometricians have worked on.

13 Theory-based empirical models

Many econometric models are connected to economic theory only loosely or through essentially arbitrary parametric assumptions about, say, the shapes of utility functions. For example, a logit model of discrete choice assumes that the random components of utility are independently and identically distributed with the Type I extreme value distribution. In addition, it is frequently assumed that the indirect utility function is linear in prices and other characteristics of the alternatives. Because economic theory rarely, if ever, yields a parametric specification of a probability model, it is worth asking whether theory provides useful restrictions on the specification of econometric models, and whether models that are consistent with economic theory can be estimated without making non-theoretical parametric

assumptions. The answers to these questions depend on the details of the setting being modelled.

In the case of discrete-choice, random-utility models, the inferential problem is to estimate the distribution of (direct or indirect) utility conditional on observed characteristics of individuals and the alternatives among which they choose. More specifically, in applied research one usually is interested in estimating the systematic component of utility (that is, the function that gives the mean of utility conditional on the explanatory variables) and the distribution of the random component of utility. Discrete choice is present in a wide range of applications, so it is important to know whether the systematic component of utility and the distribution of the random component can be estimated nonparametrically, thereby avoiding the non-theoretical distributional and functional form assumptions that are required by parametric models. The systematic component and distribution of the random component cannot be estimated unless they are identified. However, economic theory places only weak restrictions on utility functions (for example, shape restrictions such as monotonicity, convexity, and homogeneity), so the classes of conditional mean and utility functions that satisfy the restrictions are large. Indeed, it is not difficult to show that observations of individuals' choices and the values of the explanatory variables, by themselves, do not identify the systematic component of utility and the distribution of the random component without making assumptions that shrink the class of allowed functions.

This issue has been addressed in a series of papers by Matzkin that are summarized in Matzkin (1994). Matzkin gives conditions under which the systematic component of utility and the distribution of the random component are identified without restricting either to a finite-dimensional parametric family. Matzkin also shows how these functions can be estimated consistently when they are identified. Some of the assumptions required for identification may be undesirable in applications. Moreover, Manski (1988) and Horowitz (1998) have given examples in which infinitely many combinations of the systematic component of utility and distribution of the random component are consistent with a binary logit specification of choice probabilities. Thus, discrete-choice, random-utility models can be estimated under assumptions that are considerably weaker than those of, say, logit and probit models, but the systematic component of utility and the distribution of the random component cannot be identified using the restrictions of economic theory alone. It is necessary to make additional assumptions that are not required by economic theory and, because they are required for identification, cannot be tested empirically.

Models of market-entry decisions by oligopolistic firms present identification issues that are closely related to those in discrete-choice, random utility models. Berry and Tamer (2006) explain the identification problems and approaches to resolving them.

The situation is different when the economic setting provides more information about the relation between observables and preferences than is the case in discrete-choice models. This happens in models of certain kinds of auctions, thereby permitting nonparametric estimation of the distribution of values for the auctioned object. An example is a first-price, sealed bid auction within the independent private values paradigm. Here, the problem is to infer the distribution of bidders' values for the auctioned object from observed bids. A game-theory model of bidders' behaviour provides a characterization of the relation between bids and the distribution of private values. Guerre, Perrigne and Vuong (2000) show that this relation nonparametrically identifies the distribution of values if the analyst observes all bids and certain other mild conditions are satisfied. Guerre, Perrigne and Vuong (2000) also show how to

carry out nonparametric estimation of the value distribution.

Dynamic decision models and equilibrium job-search models are other examples of empirical models that are closely connected to economic theory, though they also rely on non-theoretical parametric assumptions. In a dynamic decision model, an agent makes a certain decision repeatedly over time. For example, an individual may decide each year whether to retire or not. The optimal decision depends on uncertain future events (for example, the state of one's future health) whose probabilities may change over time (for example, the probability of poor health increases as one ages) and depend on the decision. In each period, the decision of an agent who maximizes expected utility is the solution to a stochastic, dynamic programming problem. A large body of research, much of which is reviewed by Rust (1994), shows how to specify and estimate econometric models of the utility function (or, depending on the application, cost function), probabilities of relevant future events, and the decision process.

An equilibrium search model determines the distributions of job durations and wages endogenously. In such a model, a stochastic process generates wage offers. An unemployed worker accepts an offer if it exceeds his reservation wage. An employed worker accepts an offer if it exceeds his current wage. Employers choose offers to maximize expected profits. Among other things, an equilibrium search model provides an explanation for why seemingly identical workers receive different wages. The theory of equilibrium search models is described in Albrecht and Axell (1984), Mortensen (1990), and Burdett and Mortensen (1998). There is a large body of literature on the estimation of these models. Bowlus, Kiefer and Neumann (2001) provide a recent example with many references.

14 The bootstrap

The exact, finite-sample distributions of econometric estimators and test statistics can rarely be calculated in applications. This is because, except in special cases and under restrictive assumptions (for example, the normal linear model), finite sample distributions depend on the unknown distribution of the population from which the data were sampled. This problem is usually dealt with by making use of large-sample (asymptotic) approximations. A wide variety of econometric estimators and test statistics have distributions that are approximately normal or chi-square when the sample size is large, regardless of the population distribution of the data. The approximation error decreases to zero as the sample size increases. Thus, asymptotic approximations can to be used to obtain confidence intervals for parameters and critical values for tests when the sample size is large.

It has long been known, however, that the asymptotic normal and chi-square approximations can be very inaccurate with the sample sizes encountered in applications. Consequently, there can be large differences between the true and nominal coverage probabilities of confidence intervals and between the true and nominal probabilities with which a test rejects a correct null hypothesis. One approach to dealing with this problem is to use higher-order asymptotic approximations such as Edgeworth or saddlepoint expansions. These received much research attention during 1970s and 1980s, but analytic higher-order expansions are rarely used in applications because of their algebraic complexity.

The bootstrap, which is due to Efron (1979), provides a way to obtain sometimes spectacular improvements in the accuracy of asymptotic approximations while avoiding algebraic complexity. The bootstrap amounts to treating the data as if they were the population. In other words, it creates a pseudo-population whose distribution is the empirical distribution of the data. Under sampling from the pseudo-

population, the exact finite sample distribution of any statistic can be estimated with arbitrary accuracy by carrying out a Monte Carlo simulation in which samples are drawn repeatedly from the empirical distribution of the data. That is, the data are repeatedly sampled randomly with replacement. Since the empirical distribution is close to the population distribution when the sample size is large, the bootstrap consistently estimates the asymptotic distribution of a wide range of important statistics. Thus, the bootstrap provides a way to replace analytic calculations with computation. This is useful when the asymptotic distribution is difficult to work with analytically.

More importantly, the bootstrap provides a low-order Edgeworth approximation to the distribution of a wide variety of asymptotically standard normal and chi-square statistics that are used in applied research. Consequently, the bootstrap provides an approximation to the finite-sample distributions of such statistics that is more accurate than the asymptotic normal or chi-square approximation. The theoretical research leading to this conclusion was carried out by statisticians, but the bootstrap's importance has been recognized in econometrics and there is now an important body of econometric research on the topic. In many settings that are important in applications, the bootstrap essentially eliminates errors in the coverage probabilities of confidence intervals and the rejection probabilities of tests. Thus, the bootstrap is a very important tool for applied econometricians.

There are, however, situations in which the bootstrap does not estimate a statistic's asymptotic distribution consistently. Manski's (1975; 1985) maximum score estimator of the parameters of a binary response model is an example. All known cases of bootstrap inconsistency can be overcome through the use of subsampling methods. In subsampling, the distribution of a statistic is estimated by carrying out a Monte Carlo simulation in which the subsamples of the data are drawn repeatedly. The subsamples are smaller than the original data-set, and they can be drawn randomly with or without replacement. Subsampling provides estimates of asymptotic distributions that are consistent under very weak assumptions, though it is usually less accurate than the bootstrap when the bootstrap is consistent.

15 Programme evaluation and treatment effects

Programme evaluation is concerned with estimating the causal effect of a treatment or policy intervention on some population. The problem arises in many disciplines, including biomedical research (for example, the effects of a new medical treatment) and economics (for example, the effects of job training or education on earnings). The most obvious way to learn the effects of treatment on a group of individuals by observing each individual's outcome in both the treated and the untreated states. This is not possible in practice, however, because one virtually always observes any given individual in either the treated state or the untreated state but not both. This does not matter if the individuals who receive treatment are identical to those who do not, but that rarely happens. For example, individuals who choose to take a certain drug or whose physicians prescribe it for them may be sicker than individuals who do not receive the drug. Similarly, people who choose to obtain high levels of education may be different from others in ways that affect future earnings.

This problem has been recognized since at least the time of R.A. Fisher. In principle, it can be overcome by assigning individuals randomly to treatment and control groups. One can then estimate the average effect of treatment by the difference between the average outcomes of treated and untreated individuals. This random assignment procedure has become something of a gold standard in the treatment effects literature. Clinical trials use random assignment, and there have been important economic and social

experiments based on this procedure. But there are also serious practical problems. First, random assignment may not be possible. For example, one cannot assign high-school students randomly to receive a university education or not. Second, even if random assignment is possible, post-randomization events may disrupt the effects of randomization. For example, individuals may drop out of the experiment or take treatments other than the one to which they are assigned. Both of these things may happen for reasons that are related to the outcome of interest. For example, very ill members of a control group may figure out that they are not receiving treatment and find a way to obtain the drug being tested. In addition, real-world programmes may not operate the way that experimental ones do, so real-world outcomes may not mimic those found in an experiment, even if nothing has disrupted the randomization.

Much research in econometrics, statistics, and biostatistics has been aimed at developing methods for inferring treatment effects when randomization is not possible or is disrupted by post-randomization events. In econometrics, this research dates back at least to Gronau (1974) and Heckman (1974). The fundamental problem is to identify the effects of treatment or, in less formal terms, to separate the effects of treatment from those of other sources of differences between the treated and untreated groups. Manski (1995), among many others, discusses this problem. Large literatures in statistics, biostatistics, and econometrics are concerned with developing identifying assumptions that are reasonable in applied settings. However, identifying assumptions are not testable empirically and can be controversial. One widely accepted way of dealing with this problem is to conduct a sensitivity analysis in which the sensitivity of the estimated treatment effect to alternative identifying assumptions is assessed. Another possibility is to forgo controversial identifying assumptions and to find the entire set of outcomes that are consistent with the joint distribution of the observed variables. This approach, which has been pioneered by Manski and several co-investigators, is discussed in Manski (1995; 2003), among other places. Hotz, Mullin and Sanders (1997) provide an interesting application of bounding methods to measuring the effects of teenage pregnancy on the labour market outcomes of young women.

16 Integration and simulation methods in econometrics

The integration problem is endemic in economic modelling, arising whenever economic agents do not observe random variables and the behaviour paradigm is the maximization of expected utility. The econometrician inherits this problem in the expression of the corresponding econometric model, even before taking up inference and estimation. The issue is most familiar in dynamic optimization contexts, where it can be addressed by a variety of methods. Taylor and Uhlig (1990) present a comprehensive review of these methods; for later innovations see Keane and Wolpin (1994), Rust (1997) and Santos and Vigo-Aguiar (1998).

The problem is more pervasive in econometrics than in economic modelling, because it arises, in addition, whenever economic agents observe random variables that the econometrician does not. For example, the economic agent may form expectations conditional on an information set not entirely accessible to the econometrician, such as personal characteristics or confidential information. Another example arises in discrete choice settings, where utilities of alternatives are never observed and the prices of alternatives often are not. In these situations the economic model provides a probability distribution of outcomes conditional on three classes of objects: observed variables, available to the

econometrician; latent variables, unobserved by the econometrician; and parameters or functions describing the preferences and decision-making environment of the economic agent. The econometrician typically seeks to learn about the parameters or functions given the observed variables.

There are several ways of dealing with this task. Two approaches that are closely related and widely used in the econometrics literature generate integration problems. The first is to maintain a distribution of the latent variables conditional on observed variables, the parameters in the model, and additional parameters required for completing this distribution. (This is the approach taken in maximum likelihood and Bayesian inference.) Combined with the model, this leads to the joint distribution of outcomes and latent variables conditional on observed variables and parameters. Since the marginal distribution of outcomes is the one relevant for the econometrician in this conditional distribution, there is an integration problem for the latent variables. The second approach is weaker: it restricts to zero the values of certain population moments involving the latent and observable variables. (This is the approach taken in generalized method of moments, which can be implemented with both parametric and nonparametric methods.) These moments depend upon the parameters (which is why the method works) and the econometrician must therefore be able to evaluate the moments for any given set of parameter values. This again requires integration over the latent variables.

Ideally, this integral would be evaluated analytically. Often – indeed, typically – this is not possible. The alternative is to use numerical methods. Some of these are deterministic, but the rapid growth in the solution of these problems since (roughly) 1990 has been driven more by simulation methods employing pseudo-random numbers generated by computer hardware and software. This section reviews the most important these methods and describes their most significant use in non-Bayesian econometrics, namely, simulated method of moments. In Bayesian econometrics the integration problem is inescapable, the structure of the economic model notwithstanding, because parameters are treated explicitly as unobservable random variables. Consequently simulation methods have been central to Bayesian inference in econometrics.

16.1 Deterministic approximation of integrals

The evaluation of an integral is a problem as old as the calculus itself. In well-catalogued but limited instances analytical solutions are available: Gradshteyn and Ryzhik (1965) is a useful classic reference. For integration in one dimension there are several methods of deterministic approximation, including Newton-Coates (Press et al., 1986, ch. 4; Davis and Rabinowitz, 1984, ch. 2), and Gaussian quadrature (Golub and Welsch, 1969; Judd, 1998, s. 7.2). Gaussian quadrature approximates a smooth function as the product a polynomial of modest order and a smooth basis function, and then uses iterative refinements to compute the approximation. It is incorporated in most mathematical applications software and is used routinely to approximate integrals in one dimension to many significant figures of accuracy. Integration in several dimensions by means of deterministic approximation is more difficult. Practical generic adaptations of Gaussian quadrature are limited to situations in which the integrand is approximately the product of functions of single variables (Davis and Rabinowitz, 1984, pp. 354–9). Even here the logarithm of computation time is approximately linear in the number of variables, a phenomenon sometimes dubbed ‘the curse of dimensionality.’ Successful extensions of quadrature beyond dimensions of four or five are rare, and these extensions typically require substantial analytical

work before they can be applied successfully.

Low discrepancy methods provide an alternative generic approach to deterministic approximation of integrals in higher dimensions. The approximation is the average value of the integrand computed over a well-chosen sequence of points whose configuration amounts to a sophisticated lattice. Different sequences lead to variants on the approach, the best known being the Halton (1960) sequence and the Hammersley (1960) sequence. Niederreiter (1992) reviews these and other variants.

A key property of any method of integral approximation, deterministic or non-deterministic, is that it should provide as a by-product some indicator of the accuracy of the approximation. Deterministic methods typically provide upper bounds on the approximation error, based on worst-case situations. In many situations the actual error is orders of magnitude less than the upper bound, and as a consequence attaining desired error tolerances may appear to be impractical, whereas in fact these tolerances can easily be attained. Geweke (1996, s. 2.3) provides an example.

16.2 Simulation approximation of integrals

The structure of integration problems encountered in econometrics makes them often more amenable to attack by simulation methods than by non-deterministic methods. Two characteristics are key. First, integrals in many dimensions are required. In some situations the number is proportional to the size of the sample, and, while the structure of the problem may lead to decomposition in terms of many integrals of smaller dimension, the resulting structure and dimension are still unsuitable for deterministic methods. The second characteristic is that the integration problem usually arises as the need to compute the expected value of a function of a random vector with a given probability distribution *P*:

$$I = \int_S g(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}, \tag{1}$$

where *p* is the density corresponding to *P*, *g* is the function, **x** is the random vector, and *I* is the number to be approximated. The probability distribution *P* is then the point of departure for the simulation.

For many distributions there are reliable algorithms, implemented in widely available mathematical applications software, for simulation of random vectors *x*. This yields a sample $\left\{g(x^{(m)})\right\}(m = 1, \ldots, M)$ whose arithmetic mean provides an approximation of *I*, and for which a central limit theorem provides an assessment of the accuracy of the approximation in the usual way. (This requires the existence of the first two moments of *g*, which must be shown analytically.) This approach is most useful when *p* is simple (so that direct simulation of **x** is possible) but the structure of *g* precludes analytical evaluation of *I*.

This simple approach does not suffice for the integration problem as it typically arises in econometrics. A leading example is the multinomial probit (MNP) model with *J* discrete choices. For each individual *i* the utility of the last choice u_{iJ} is normalized to be zero, and the utilities of the first $J - 1$ choices are

given by the vector

$$\mathbf{u}_i \sim N(\mathbf{X}_i\beta, \Sigma), \tag{2}$$

where **X** is a matrix of characteristics of individual *i*, including the prices and other properties of the choices presented to that individual, and β and Σ are structural parameters of the model. If the *j*'th element of $\mathbf{u}_i$ is positive and larger than all the other elements of $\mathbf{u}_i$ the individual makes choice *j*, and if all elements of **u** are negative the individual makes choice *J*. The probability that individual *i* makes choice *j* is the integral of the $(n-1)$-variate normal distribution (1) taken over the subspace $\{\mathbf{u}_i : u_{ik} \leq u_{ij} \forall k = 1, \ldots, n\}$. This computation is essential in evaluating the likelihood function, and it has no analytical solution. (For discussion and review, see Sandor and Andras, 2004.)

Several generic simulation methods have been used for the problem (1) in econometrics. One of the oldest is acceptance sampling, a simple variant of which is described in von Neumann (1951) and Hammersley and Handscomb (1964). Suppose it is possible to draw from the distribution *Q* with density *q*, and the ratio $p(\mathbf{x}) / q(\mathbf{x})$ is bounded above by the known constant *a*. If **x** is simulated successively from *Q* but accepted and taken into the sample with probability $p(\mathbf{x}) / [aq(\mathbf{x})]$, then the resulting sample is independently distributed with the identical distribution *P*. Proofs and further discussion are widely available; for example, Press et al. (1992, s. 7.4), Bratley, Fox and Schrage (1987, s. 5.2.5), and Geweke (2005, s. 4.2.1). The unconditional probability of accepting draws from *Q* is 1/*a*. If *a* is too large the method is impractical, but when acceptance sampling is practical it provides draws directly from *P*. This is an important component of many of the algorithms underlying the ‘black box’ generation of random variables in mathematical applications software.

Alternatively, in the same situation all of the draws from *Q* are retained and taken into a stratified sample in which the weight $w(\mathbf{x}^{(m)}) = p(\mathbf{x}^{(m)}) / q(\mathbf{x}^{(m)})$ is associated with the *m*'th draw. The approximation of *I* in (1) is then the weighted average of the terms $g(\mathbf{x}^{(m)})$. This approach dates at least to Hammersley and Handscomb (1964, s. 5.4), and was introduced to econometrics by Kloek and van Dijk (1978). The procedure is more general than acceptance sampling in that a known upper bound of *w* is not required, but if in fact a is large then the weights will display large variation and the approximation will be poor. This is clear in the central limit theorem for the accuracy of approximation provided in Geweke (1989a), which as a practical matter requires that a finite upper bound on *w* be established analytically. This is a key limitation of acceptance sampling and importance sampling.

Markov chain Monte Carlo (MCMC) methods provide an entirely different approach to the solution of the integration problem (1). These procedures construct a Markov process of the form

$$\mathbf{x}^{(m)} \sim p(\mathbf{x}|\mathbf{x}^{(m-1)}) \tag{3}$$

in such a way that

$$M^{-1}\sum_{m=1}^{M} g(x^{(m)})$$

converges (almost surely) to *I*. These methods have a history in mathematical physics dating back to the algorithm of Metropolis et al. (1953). Hastings (1970) focused on statistical problems and extended the method to its present form known as the Hastings–Metropolis (HM) algorithm. HM draws a candidate $\mathbf{x}^{*}$ from a convenient distribution indexed by $\mathbf{x}^{(m-1)}$. It sets $\mathbf{x}^{(m)} = \mathbf{x}$ with probability $\alpha(\mathbf{x}^{(m-1)}, \mathbf{x}^{(m)})$ and sets $\mathbf{x}^{(m)} = \mathbf{x}^{(m)-1}$ otherwise, the function α being chosen so that the process (3) defined in this way has the desired convergence property. Chib and Greenberg (1995) provide a detailed introduction to HM and its application in econometrics. Tierney (1994) provides a succinct summary of the relevant continuous state space Markov chain theory bearing on the convergence of MCMC.

A version of the HM algorithm particularly suited to image reconstruction and problems in spatial statistics, known as the Gibbs sampling (GS) algorithm, was introduced by Geman and Geman (1984). This was subsequently shown to have great potential for Bayesian computation by Gelfand and Smith (1990). In GS the vector **x** is subdivided into component vectors, $\mathbf{x}' = (\mathbf{x}_1', \ldots, \mathbf{x}_B')$, in such a way that simulation from the conditional distribution of each $\mathbf{x}_j$ implied by $p(\mathbf{x})$ in (1) is feasible. This method has proven very advantageous in econometrics generally, and it revolutionized Bayesian approaches in particular beginning about 1990.

By the turn of the century HM and GS algorithms were standard tools for likelihood-based econometrics. Their structure and strategic importance for Bayesian econometrics were conveyed in surveys by Geweke (1999) and Chib (2001), as well as in a number of textbooks, including Koop (2003), Lancaster (2004), Geweke (2005) and Rossi, Allenby and McCulloch (2005). Central limit theorems can be used to assess the quality of approximations as described in Tierney (1994) and Geweke (2005).

16.3 Simulation methods in non-Bayesian econometrics

Generalized method of moments estimation has been a staple of non-Bayesian econometrics since its introduction by Hansen (1982). In an econometric model with $k \times 1$ parameter vector θ economic theory provides the set of sample moment restrictions

$$\mathbf{h}(\boldsymbol{\theta}) = \int_S \mathbf{g}(\mathbf{x})\, p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{x} = 0, \tag{4}$$

where **g**(**x**) is a $p \times 1$ vector and **y** denotes the data including instrumental variables. An example is the MNP model (2). If the observed choices are coded by the variables $d_{ij} = 1$ if individual *i* makes choice *j* and $d_{ij} = 0$ otherwise, then the expected value of d_{ij} is the probability that individual *i* makes choice *j*, leading to restrictions of the form (4).

The generalized method of moments estimator minimizes the criterion function $\mathbf{h}(\boldsymbol{\theta})'\mathbf{W}\mathbf{h}(\boldsymbol{\theta})$ given a suitably chosen weighting matrix **W**. If the requisite integrals can be evaluated analytically, $p \geq k$, and other conditions provided in Hansen (1982) are satisfied, then there is a well-developed asymptotic theory of inference for the parameters that by 1990 was a staple of graduate econometrics textbooks. If for one or more elements of **h** the integral cannot be evaluated analytically, then for alternative values of it is often possible to approximate the integral appearing in (4) by simulation. This is the situation in the MNP model.

The substitution of a simulation approximation

$$M^{-1} \sum_{m=1}^{M} \mathbf{g}(\mathbf{x}^{(m)})$$

for the integral in (4) defines the method of simulated moments (MSM) introduced by McFadden (1989) and Pakes and Pollard (1989), who were concerned with the MNP model (2) in particular and the estimation of discrete response models using cross-section data in general. Later the method was extended to time series models by Lee and Ingram (1991) and Duffie and Singleton (1993). The asymptotic distribution theory established in this literature requires that the number of simulations *M* increase at least as rapidly as the square of the number of observations. The practical import of this apparently severe requirement is that applied econometric work must establish that changes in *M* must have little impact on the results; Geweke, Keane and Runkle (1994; 1997) provide examples for MNP. This literature also shows that in general the impact of using direct simulation, as opposed to analytical evaluation of the integral, is to increase the asymptotic variance of the GMM estimator of θ by the factor M^{-1}, typically trivial in view of the number of simulations required. Substantial surveys of the details of MSM and leading applications of the method can be found in Gourieroux and Monfort (1993; 1996), Stern (1997) and Liesenfeld and Breitung (1999).

The simulation approximation, unlike the (unavailable) analytical evaluation of the integral in (4), can lead to a criterion function that is discontinuous in θ. This happens in the MNP model using the obvious simulation scheme in which the choice probabilities are replaced by their proportions in the *M*

simulations, as proposed by Lerman and Manski (1981). The asymptotic theory developed by McFadden (1989) and Pakes and Pollard (1989) copes with this possibility, and led McFadden (1989) to used kernel weighting to smooth the probabilities. The most widely used method for smoothing probabilities in the MNP model is the Geweke–Hajivassiliou–Keane (GHK) simulator of Geweke (1989b), Hajivassiliou, McFadden and Ruud (1991) and Keane (1990); a full description is provided in Geweke and Keane (2001), and comparisons of alternative methods are given in Hajivassiliou, McFadden and Ruud (1996) and Sandor and Andras (2004).

Maximum likelihood estimation of θ can lead to first-order conditions of the form (4), and thus becomes a special case of MSM. This context highlights some of the complications introduced by simulation. While the simulation approximation of (1) is unbiased, the corresponding expression enters the log likelihood function and its derivatives nonlinearly. Thus for any finite number of simulations *M*, the evaluation of the first-order conditions is biased in general. Increasing *M* at a rate faster than the square of the number of observations eliminates the squared bias relative to the variance of the estimator; Lee (1995) provides further details.

16.4 Simulation methods in Bayesian econometrics

Bayesian econometrics places a common probability distribution on random variables that can be observed (data) and unobservable parameters and latent variables. Inference proceeds using the distribution of these unobservable entities conditional on the data – the posterior distribution. Results are typically expressed in terms of the expectations of parameters or functions of parameters, expectations taken with respect to the posterior distribution. Thus, whereas integration problems are application-specific in non-Bayesian econometrics, they are endemic in Bayesian econometrics.

The development of modern simulation methods had a correspondingly greater impact in Bayesian than in non-Bayesian econometrics. Since 1990 simulation-based Bayesian methods have become practical in the context of most econometric models. The availability of this tool has been influential in the modelling approach taken in addressing applied econometric problems.

The MNP model (2) illustrates the interaction in latent variable models. Given a sample of *n* individuals, the $(J-1)\times 1$ latent utility vectors $\mathbf{u}_1,\ldots,\mathbf{u}_n$ are regarded explicitly as $n(J-1)$ unknowns to be inferred along with the unknown parameters β and Σ. Conditional on these parameters and the data, the vectors $\mathbf{u}_1,\ldots,\mathbf{u}_n$ are independently distributed. The distribution of $\mathbf{u}_i$ is (2) truncated to an orthant that depends on the observed choice *j:* if $j<J$ then $u_{ik}<u_{ij}$ for all $k\neq j$ and $u_{ij}>0$, whereas for choice *J:*, $u_{ik}<0$ for all *k*. The distribution of each u_{ik}, conditional on all of the other elements of $\mathbf{u}_i$, is truncated univariate normal, and it is relatively straightforward to simulate from this distribution. (Geweke, 1991, provides details on sampling from a multivariate normal distribution subject to linear restrictions.) Consequently GS provides a practical algorithm for drawing from the distribution of the latent utility vectors conditional on the parameters.

Conditional on the latent utility vectors – that is, regarding them as observed – the MNP model is a seemingly unrelated regressions model, and the approach taken by Percy (1992) applies. Given conjugate priors the posterior distribution of β, conditional on Σ and utilities, is Gaussian, and the conditional distribution of Σ, conditional on β and utilities, is inverted Wishart. Since GS provides the

joint distribution of parameters and latent utilities, the posterior mean of any function of these can be approximated as the sample mean. This approach and the suitability of GS for latent variable models were first recognized by Chib (1992). Similar approaches in other latent variable models in include McCulloch and Tsay (1994), Chib and Greenberg (1998), McCulloch, Polson and Rossi (2000) and Geweke and Keane (2001).

The Bayesian approach with GS sidesteps the evaluation of the likelihood function and, of any moments in which the approximation is biased given a finite number of simulations, two technical issues that are prominent in MSM. On the other hand, as in all MCMC algorithms, there may be sensitivity to the initial values of parameters and latent variables in the Markov chain, and substantial serial correlation in the chain will reduce the accuracy of the simulation approximation. Geweke (1992; 2005) and Tierney (1994) discuss these issues.

17 Financial econometrics

Attempts at testing of the efficient market hypothesis (EMH) provided the impetus for the application of time series econometric methods in finance. The EMH was built on the pioneering work of Bachelier (1900) and evolved in the 1960s from the random walk theory of asset prices advanced by Samuelson (1965). By the early 1970s a consensus had emerged among financial economists suggesting that stock prices could be well approximated by a random walk model and that changes in stock returns were basically unpredictable. Fama (1970) provides an early, definitive statement of this position. He distinguished between different forms of the EMH: the ‘weak’ form that asserts all price information is fully reflected in asset prices; the ‘semi-strong’ form that requires asset price changes to fully reflect all publicly available information and not only past prices; and the ‘strong’ form that postulates that prices fully reflect information even if some investor or group of investors have monopolistic access to some information. Fama regarded the strong form version of the EMH as a benchmark against which the other forms of market efficiencies are to be judged. With respect to the weak form version he concluded that the test results strongly support the hypothesis, and considered the various departures documented as economically unimportant. He reached a similar conclusion with respect to the semi-strong version of the hypothesis. Evidence on the semi-strong form of the EMH was revisited by Fama (1991). By then it was clear that the distinction between the weak and the semi-strong forms of the EMH was redundant. The random walk model could not be maintained either, in view of more recent studies, in particular that of Lo and MacKinlay (1988).

This observation led to a series of empirical studies of stock return predictability over different horizons. It was shown that stock returns can be predicted to some degree by means of interest rates, dividend yields and a variety of macroeconomic variables exhibiting clear business cycle variations. See, for example, Fama and French (1989), Kandel and Stambaugh (1996), and Pesaran and Timmermann (1995) on predictability of equity returns in the United States; and Clare, Thomas and Wickens (1994), and Pesaran and Timmermann (2000) on equity return predictability in the UK.

Although it is now generally acknowledged that stock returns could be predictable, there are serious difficulties in interpreting the outcomes of market efficiency tests. Predictability could be due to a number of different factors such as incomplete learning, expectations heterogeneity, time variations in

risk premia, transaction costs, or specification searches often carried out in pursuit of predictability. In general, it is not possible to distinguish between the different factors that might lie behind observed predictability of asset returns. As noted by Fama (1991) the test of the EMH involves a joint hypothesis, and can be tested only jointly with an assumed model of market equilibrium. This is not, however, a problem that is unique to financial econometrics; almost all areas of empirical economics are subject to the joint hypotheses problem. The concept of market efficiency is still deemed to be useful as it provides a benchmark and its use in finance has led to significant insights.

Important advances have been made in the development of equilibrium asset pricing models, econometric modelling of asset return volatility (Engle, 1982; Bollerslev, 1986), analysis of high frequency intraday data, and market microstructures. Some of these developments are reviewed in Campbell, Lo and MacKinlay (1997), Cochrane (2005), Shephard (2005), and McAleer and Medeiros (2007). Future advances in financial econometrics are likely to focus on heterogeneity, learning and model uncertainty, real time analysis, and further integration with macroeconometrics. Finance is particularly suited to the application of techniques developed for real time econometrics (Pesaran and Timmermann, 2005a).

18 Appraisals and future prospects

Econometrics has come a long way over a relatively short period. Important advances have been made in the compilation of economic data and in the development of concepts, theories and tools for the construction and evaluation of a wide variety of econometric models. Applications of econometric methods can be found in almost every field of economics. Econometric models have been used extensively by government agencies, international organizations and commercial enterprises. Macroeconometric models of differing complexity and size have been constructed for almost every country in the world. In both theory and practice, econometrics has already gone well beyond what its founders envisaged. Time and experience, however, have brought out a number of difficulties that were not apparent at the start.

Econometrics emerged in the 1930s and 1940s in a climate of optimism, in the belief that economic theory could be relied on to identify most, if not all, of the important factors involved in modelling economic reality, and that methods of classical statistical inference could be adapted readily for the purpose of giving empirical content to the received economic theory. This early view of the interaction of theory and measurement in econometrics, however, proved rather illusory. Economic theory is invariably formulated with *ceteris paribus* clauses, and involves unobservable latent variables and general functional forms; it has little to say about adjustment processes, lag lengths and other factors mediating the relationship between the theoretical specification (even if correct) and observables. Even in the choice of variables to be included in econometric relations, the role of economic theory is far more limited than was at first recognized. In a Walrasian general equilibrium model, for example, where everything depends on everything else, there is very little scope for a priori exclusion of variables from equations in an econometric model. There are also institutional features and accounting conventions that have to be allowed for in econometric models but which are either ignored or are only partially dealt with at the theoretical level. All this means that the specification of econometric models inevitably involves important auxiliary assumptions about functional forms, dynamic specifications, latent

variables, and so on, with respect to which economic theory is silent or gives only an incomplete guide. The recognition that economic theory on its own cannot be expected to provide a complete model specification has important consequences for testing and evaluation of economic theories, for forecasting and real time decision making. The incompleteness of economic theories makes the task of testing them a formidable undertaking. In general it will not be possible to say whether the results of the statistical tests have a bearing on the economic theory or the auxiliary assumptions. This ambiguity in testing theories, known as the Duhem–Quine thesis, is not confined to econometrics and arises whenever theories are conjunctions of hypotheses (on this, see for example Cross, 1982). The problem is, however, especially serious in econometrics because theory is far less developed in economics than it is in the natural sciences. There are, of course, other difficulties that surround the use of econometric methods for the purpose of testing economic theories. As a rule economic statistics are not the results of designed experiments, but are obtained as by-products of business and government activities often with legal rather than economic considerations in mind. The statistical methods available are generally suitable for large samples while the economic data typically have a rather limited coverage. There are also problems of aggregation over time, commodities and individuals that further complicate the testing of economic theories that are micro-based.

Econometric theory and practice seek to provide information required for informed decision-making in public and private economic policy. This process is limited not only by the adequacy of econometrics but also by the development of economic theory and the adequacy of data and other information. Effective progress, in the future as in the past, will come from simultaneous improvements in econometrics, economic theory and data. Research that specifically addresses the effectiveness of the interface between any two of these three in improving policy – to say nothing of all of them – necessarily transcends traditional sub-disciplinary boundaries within economics. But it is precisely these combinations that hold the greatest promise for the social contribution of academic economics.

Bibliography

Aitken, A.C. 1934–5. On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* 55, 42–8.

Albrecht, J.W. and Axell, B. 1984. An equilibrium model of search unemployment. *Journal of Political Economy* 92, 824–40.

Allen, R.G.D. and Bowley, A.L. 1935. *Family Expenditure*. London: P.S. King.

Almon, S. 1965. The distributed lag between capital appropriations and net expenditures. *Econometrica* 33, 178–96.

Amemiya, T. 1983. Nonlinear regression models. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Amemiya, T. 1984. Tobit models: a survey. *Journal of Econometrics* 24, 3–61.

An, S. and Schorfheide, F. 2007. Bayesian analysis of DSGE models. *Econometric Reviews* 26(2–4), 113–72.

Anderson, T.W. and Hsiao, C. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Society* 76, 598–606.

Anderson, T.W. and Hsiao, C. 1982. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* 18, 47–82.

Anderson, T.W. and Rubin, H. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20, 46–63.

Andrews, D.W.K. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821–56.

Andrews, D.W.K. and Ploberger, W. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383–414.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Boston: Kluwer Academic Publishers.

Anselin, L., Le Gallo, J. and Jayet, H. 2007. Spatial panel econometrics. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, 3rd edn. ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer (forthcoming).

Arellano, M. 2003. *Panel Data Econometrics*. Oxford: Oxford University Press.

Arellano, M. and Bond, S.R. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58, 277–97.

Arellano, M. and Honoré, B. 2001. Panel data models: some recent developments. In *Handbook of Econometrics*, vol. 5, ed. J.J. Heckman and E. Leamer. Amsterdam: North-Holland.

Bachelier, L.J.B.A. 1900. *Théorie de la Speculation*. Paris: Gauthier-Villars. Reprinted in *The Random Character of Stock Market Prices*, ed. P. H. Cootner. Cambridge, MA: MIT Press, 1964.

Bai, J. and Ng, S. 2004. A panic attack on unit roots and cointegration. *Econometrica* 72, 1127–77.

Bai, J. and Perron, P. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66, 47–78.

Baltagi, B. 2005. *Econometric Analysis of Panel Data*, 2nd edn. New York: Wiley.

Baltagi, B.H. and Kao, C. 2000. Nonstationary panels, cointegration in panels and dynamic panels: a survey. In *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, ed. B. Baltagi. Advances in Econometrics, vol. 15. Amsterdam: JAI Press.

Banerjee, A. 1999. Panel data unit roots and cointegration: an overview. *Oxford Bulletin of Economics and Statistics* 61, 607–29.

Basmann, R.L. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25, 77–83.

Bauwens, L., Lubrano, M. and Richard, J.F. 2001. *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Benini, R. 1907. Sull'uso delle formole empiriche a nell'economia applicata. *Giornale degli economisti*, 2nd series 35, 1053–63.

Berry, S. and Tamer, E. 2006. Identification in models of oligopoly entry. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. R. Blundell, W.K. Newey and T. Persson. Cambridge: Cambridge University Press.

Beveridge, S. and Nelson, C.R. 1981. A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to measurement of the ‘business cycle’. *Journal of Monetary Economics* 7, 151–74.

Binder, M., Hsiao, C. and Pesaran, M.H. 2005. Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory* 21, 795–837.

Binder, M. and Pesaran, M.H. 1995. Multivariate rational expectations models and macroeconometric modelling: a review and some new results. In *Handbook of Applied Econometrics, Volume 1 – Macroeconomics*, ed. M.H. Pesaran and M.R. Wickens. Oxford: Basil Blackwell.

Bjerkholt, O. 1995. Ragnar Frisch, Editor of *Econometrica*. *Econometrica* 63, 755–65.

Blanchard, O.J. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 1146–64.

Blundell, R.W., Browning, M. and Crawford, I.A. 2003. Nonparametric Engel curves and revealed preference. *Econometrica* 71, 205–40.

Bollerslev, T. 1986. Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics* 51, 307–27.

Boskin, M.J. 1974. A conditional logit model of occupational choice. *Journal of Political Economy* 82, 389–98.

Bosq, D. 1996. *Nonparametric Statistics for Stochastic Processes*. New York: Springer.

Bowlus, A.J., Kiefer, N.M. and Neumann, G.R. 2001. Equilibrium search models and the transition from school to work. *International Economic Review* 42, 317–43.

Box, G.E.P. and Jenkins, G.M. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Bratley, P., Fox, B.L. and Schrage, L.E. 1987. *A Guide to Simulation*. New York: Springer-Verlag.

Breitung, J. and Pesaran, M.H. 2008. Unit roots and cointegration in panels. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, 3rd edn. ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Breusch, T.S. and Pagan, A.R. 1980. The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies* 47, 239–53.

Brock, W. and Durlauf, S. 2002. A multinomial choice model with neighborhood effects. *American Economic Review* 92, 298–303.

Brock, W. and Durlauf, S. 2006. Macroeconomics and model uncertainty. In *Post-Walrasian Macroeconomics: Beyond the Dynamic Stochastic General Equilibrium Model*, ed. D. Colander. New York: Cambridge University Press.

Brown, R.L., Durbin, J. and Evans, J.M. 1975. Techniques for testing the constancy of regression relationships over time (with discussion). *Journal of the Royal Statistical Society*, Series B 37, 149–92.

Brown, T.M. 1952. Habit persistence and lags in consumer behaviour. *Econometrica* 20, 355–71.

Broze, L., Gouriéroux, C. and Szafarz, A. 1985. Solutions of dynamic linear rational expectations models. *Econometric Theory* 1, 341–68.

Brundy, J.M. and Jorgenson, D.N. 1971. Efficient estimation of simultaneous equations by instrumental variables. *Review of Economics and Statistics* 53, 207–24.

Buckley, J. and James, I. 1979. Linear regression with censored data. *Biometrika* 66, 429–36.

Burdett, K. and Mortensen, D.T. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39, 257–73.

Burns, A.F. and Mitchell, W.C. 1947. *Measuring Business Cycles*. New York: Columbia University Press for the NBER.

Burtless, G. 1995. The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives* 9(2), 63–84.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Cameron, A.C. and Trivedi, P.K. 1986. Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1, 29–53.

Cameron, A.C. and Trivedi, P.K. 1998. *Regression Analysis for Count Data*. Econometric Society Monograph No. 30. Cambridge: Cambridge University Press.

Cameron, A.C. and Trivedi, P.K. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Campbell, J.Y., Lo, A.W. and MacKinlay, A.C. 1997. *The Econometrics of Financial Markets*, Princeton: Princeton University Press.

Canova, F. and de Nicolo, G. 2002. Monetary disturbances matter for business fluctuations in the G7. *Journal of Monetary Economics* 49, 1131–59.

Champernowne, D.G. 1948. Sampling theory applied to autoregressive sequences. *Journal of the Royal Statistical Society, Series B* 10, 204–31.

Champernowne, D.G. 1960. An experimental investigation of the robustness of certain procedures for estimating means and regressions coefficients. *Journal of the Royal Statistical Society* 123, 398–412.

Chib, S. 1992. Bayes inference in the tobit censored regression model. *Journal of Econometrics* 51, 79–99.

Chib, S. 2001. Markov chain Monte Carlo methods: computation and inference. In *Handbook of Econometrics*, vol. 5, ed. J. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Chib, S. and Greenberg, E. 1995. Understanding the Metropolis–Hastings algorithm. *The American Statistician* 49, 327–35.

Chib, S. and Greenberg, E. 1998. Analysis of multivariate probit models. *Biometrika* 85, 347–61.

Choi, I. 2006. Nonstationary panels. In *Palgrave Handbooks of Econometrics*, vol. 1, ed. T.C. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Chow, G.C. 1960. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28, 591–605.

Christ, C.F. 1952. *Economic Theory and Measurement: A Twenty-Year Research Report. 1932–52*. Chicago: Cowles Commission for Research in Economics.

Christ, C.F. 1983. The founding of the Econometric Society and Econometrica. *Econometrica* 51, 3–6.

Christiano, L.J., Eichenbaum, M. and Evans, C. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–45.

Clare, A.D., Thomas, S.H. and Wickens, M.R. 1994. Is the gilt–equity yield ratio useful for predicting UK stock return? *Economic Journal* 104, 303–15.

Clements, M.P. and Hendry, D.F. 1998. *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, M.P. and Hendry, D.F. 1999. *Forecasting Non-stationary Economic Time Series*. Cambridge, MA: MIT Press.

Clements, M.P. and Hendry, D.F. 2006. Forecasting with breaks. In *Handbook of Economic Forecasting*, vol. 1, ed. G. Elliott, C.W.J. Granger and A. Timmermann. Amsterdam: North-Holland.

Cochrane, J. 2005. *Asset Pricing*, rev. edn. Princeton: Princeton University Press.

Cochrane, P. and Orcutt, G.H. 1949. Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association* 44, 32–61.

Conley, T.G. and Dupor, B. 2003. A spatial analysis of sectoral complementarity. *Journal of Political Economy* 111, 311–52.

Conley, T.G. and Topa, G. 2002. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics* 17, 303–27.

Cooper, R.L. 1972. The predictive performance of quarterly econometric models of the United States. In *Econometric Models of Cyclical Behavior*, ed. B.G. Hickman. New York: NBER.

Cosslett, S.R. 1983. Distribution free maximum likelihood estimation of the binary choice model. *Econometrica* 51, 765–82.

Cox, D.R. 1961. Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press.

Cox, D.R. 1962. Further results of tests of separate families of hypotheses. *Journal of the Royal Statistical Society*, Series B 24, 406–24.

Cox, D.R. 1972. Regression models and life tables. *Journal of the Royal Statistical Society*, Series B 34, 187–220.

Cross, R. 1982. The Duhem–Quine thesis, Lakatos and the appraisal of theories in macroeconomics. *Economic Journal* 92, 320–40.

Davenant, C. 1698. *Discourses on the Publick Revenues and on the Trade of England*, Vol. 1, London.

Davidson, R. and MacKinnon, J.G. 2004. *Econometric Theory and Methods*. Oxford: Oxford University Press.

Davis, P.J. and Rabinowitz, P. 1984. *Methods of Numerical Integration*. Orlando, FL: Academic Press.

Deaton, A. 1985. Panel data from time series of cross-sections. *Journal of Econometrics* 30, 109–26.

Del Negro, M. and Schorfheide, F. 2004. Priors from equilibrium models for VAR's. *International Economic Review* 45, 643–73.

Del Negro, M., Schorfheide, F., Smets, F. and Wouters, R. 2005. On the fit and forecasting performance of new Keynesian models. Working Paper No. 491. Frankfurt: European Central Bank.

Dhrymes, P. 1971. A simplified estimator for large-scale econometric models. *Australian Journal of Statistics* 13, 168–75.

Doan, T., Litterman, R. and Sims, C.A. 1984. Forecasting and conditional projections using realistic prior distributions. *Econometric Reviews* 3, 1–100.

Domenich, T. and McFadden, D. 1975. *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland.

Duffie, D. and Singleton, K. 1993. Simulated moments estimation of Markov models of asset prices. *Econometrica* 61, 929–52.

Duncan, G. 1980. Formulation and statistical analysis of the mixed continuous/discrete variable model in classical production theory. *Econometrica* 48, 839–52.

Durbin, J. 1954. . Errors in variables. *Review of the International Statistical Institute* 22, 23–32.

Durbin, J. and Watson, G.S. 1950. Testing for serial correlation in least squares regression I. *Biometrika* 37, 409–28.

Durbin, J. and Watson, G.S. 1951. Testing for serial correlation in least squares regression II. *Biometrika* 38, 159–78.

Durlauf, S.N., Johnson, P.A. and Temple, J.R.W. 2005. Growth econometrics. In *Handbook of Economic Growth*, vol. 1A, ed. P. Aghion and S. N. Durlauf. Amsterdam: North-Holland.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1–26.

Eisner, R. and Strotz, R.H. 1963. Determinants of business investment. In *Impacts of Monetary Policy*. Englewood Cliffs, NJ: Prentice-Hall, for the Commission on Money and Credit.

Elliott, G., Granger, C.W.J. and Timmermann, A. 2006. *Handbook of Economic Forecasting*, vol. 1, Amsterdam: North-Holland.

Engle, R.F. 1982. Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

Engle, R.F. 1984. Wald likelihood ratio and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Engle, R.F. and Granger, G. 1987. Cointegration and error-correction: representation, estimation and testing. *Econometrica* 55, 251–76.

Engle, R.F., Hendry, D.F. and Richard, J.-F. 1983. Exogeneity. *Econometrica* 51, 277–304.

Fachin, S. 2006. Long-run trends in internal migrations in Italy: a study in panel cointegration with dependent units. *Journal of Applied Econometrics* (forthcoming).

Fama, E.F. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance*

25, 383–417.

Fama, E.F. 1991. Efficient capital markets: II. *Journal of Finance* 46, 1575–617.

Fama, E.F. and French, K.R. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25, 23–49.

Fan, J. and Gijbels, I. 1996. *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

Fernandez-Villaverde, J. and Rubio-Ramirez, J. 2005. Estimating dynamic equilibrium economies: linear versus nonlinear likelihood. *Journal of Applied Econometrics* 20, 891–910.

Fisher, F.M. 1966. *The Identification Problem in Econometrics*. New York: McGraw-Hill.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan. Reprinted, Philadelphia: Porcupine Press, 1977.

Fisher, I. 1937. Note on a short-cut method for calculating distributed lags. *Bulletin de l'Institut International de Statistique* 29, 323–7.

Fisher, R.A. 1928. *Statistical Methods for Research Workers*, 2nd edn. London: Oliver and Boyd.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Frisch, R. 1933. *Pitfalls in the Statistical Construction of Demand and Supply Curves*. Leipzig: Hans Buske Verlag.

Frisch, R. 1934. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: University Institute of Economics.

Frisch, R. 1936. A note on the term ‘econometrics’. *Econometrica* 4, 95.

Gali, J. 1992. How well does the IS–LM model fit postwar US data? *Quarterly Journal of Economics* 107, 709–38.

Garratt, A., Robertson, D. and Wright, S. 2006. Permanent vs transitory components and economic fundamentals. *Journal of Applied Econometrics* 21, 521–42.

Garratt, A., Lee, K., Pesaran, M.H. and Shin, Y. 2003a. A long run structural macroeconometric model of the UK. *Economic Journal* 113(487), 412–55.

Garratt, A., Lee, K., Pesaran, M.H. and Shin, Y. 2003b. Forecast uncertainty in macroeconometric modelling: an application to the UK economy. *Journal of the American Statistical Association* 98(464), 829–38.

Garratt, A., Lee, K., Pesaran, M.H. and Shin, Y. 2006. *Global and National Macroeconometric Modelling: A Long-Run Structural Approach*. Oxford: Oxford University Press.

Geary, R.C. 1949. Studies in relations between economic time series. *Journal of the Royal Statistical Society*, Series B 10, 140–58.

Gelfand, A.E. and Smith, A.F.M. 1990. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–41.

Geweke, J. 1989a. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57, 1317–0.

Geweke, J. 1989b. Efficient simulation from the multivariate normal distribution subject to linear inequality constraints and the evaluation of constraint probabilities. Discussion paper, Duke University.

Geweke, J. 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, ed. E.M. Keramidas. Fairfax: Interface Foundation of North America, Inc.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, ed. J.M. Bernardo et al. Oxford: Clarendon Press.

Geweke, J. 1996. Monte Carlo simulation and numerical integration. In *Handbook of Computational Economics*, ed. H.M. Amman, D.A. Kendrick and J. Rust. Amsterdam: North-Holland.

Geweke, J. 1999. Using simulation methods for Bayesian econometric models: Inference, development and communication (with discussion and rejoinder). *Econometric Reviews* 18, 1–126.

Geweke, J. 2005. *Contemporary Bayesian Econometrics and Statistics*. New York: Wiley.

Geweke, J. and Keane, M. 2001. Computationally intensive methods for integration in econometrics. In *Handbook of Econometrics*, vol.5, ed. J. Heckman and E.E. Leamer. Amsterdam: North-Holland.

Geweke, J., Keane, M. and Runkle, D. 1994. Alternative computational approaches to statistical

inference in the multinomial probit model. *Review of Economics and Statistics* 76, 609–32.

Geweke, J., Keane, M. and Runkle, D. 1997. Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics* 80, 125–65.

Geweke, J. and Whiteman, C. 2006. Bayesian forecasting. In *Handbook of Economic Forecasting*, ed. G. Elliott, C.W.J. Granger and A. Timmermann. Amsterdam: North-Holland.

Godfrey, L.G. 1988. *Misspecification Tests in Econometrics: The LM principle and Other Approaches*. Cambridge: Cambridge University Press.

Godfrey, L.G. and Wickens, M.R. 1982. Tests of mis-specification using locally equivalent alternative models. In *Evaluation and Reliability of Macro-economic Models*, ed. G.C. Chow and P. Corsi. New York: John Wiley.

Golub, G.H. and Welsch, J.H. 1969. Calculation of Gaussian quadrature rules. *Mathematics of Computation* 23, 221–30.

Gourieroux, C. and Jasiak, J. 2001. *Financial Econometrics: Problems, Models, and Methods*. Oxford: Oxford University Press.

Gourieroux, C. and Monfort, A. 1993. Simulation based inference: a survey with special reference to panel data models. *Journal of Econometrics* 59, 5–33.

Gourieroux, C. and Monfort, A. 1996. *Simulation-Based Econometric Methods*. New York: Oxford University Press.

Gradshteyn, I.S. and Ryzhik, I.M. 1965. *Tables of Integrals, Series and Products*. New York: Academic Press.

Granger, C.W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.

Granger, C.W.J. 1986. Developments in the study of co-integrated economic variables. *Oxford Bulletin of Economics and Statistics* 48, 213–28.

Granger, C.W.J. and Newbold, P. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.

Granger, C.W.J. and Newbold, P. 1986. *Forecasting Economic Time Series*, 2nd edn. San Diego: Academic Press.

Granger, C.W.J. and Pesaran, M.H. 2000a. A decision theoretic approach to forecast evaluation. In *Statistics and Finance: An Interface*, ed. W.S. Chan, W.K. Li and H. Tong. London: Imperial College Press.

Granger, C.W.J. and Pesaran, M.H. 2000b. Economic and statistical measures of forecast accuracy. *Journal of Forecasting* 19, 537–60.

Greene, W.H. 2003. *Econometric Analysis*, 5th edn. New Jersey: Prentice Hall.

Gronau, R. 1974. Wage comparisons – a selectivity bias. *Journal of Political Economy* 82, 1119–43.

Guerre, E., Perrigne, I. and Vuong, Q. 2000. Optimal nonparametric estimation of first-price auctions. *Econometrica* 68, 525–74.

Haavelmo, T. 1943. Statistical testing of business cycle theories. *Review of Economics and Statistics* 25, 13–18.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement), 1–118.

Hajivassiliou, V., McFadden, D. and Ruud, P. 1991. Simulation of multivariate normal rectangle probabilities. Methods and programs mimeo, University of California, Berkeley.

Hajivassiliou, V., McFadden, D. and Ruud, P. 1996. Simulation of multivariate normal rectangle probabilities and their derivatives: theoretical and computational results. *Journal of Econometrics* 72, 85–134.

Hall, A.R. 2005. *Generalized Method of Moments*. Oxford: Oxford University Press.

Halton, J.M. 1960. On the efficiency of evaluating certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik* 2, 84–90.

Hamilton, J.D. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Hammersley, J.M. 1960. Monte Carlo methods for solving multivariate problems. *Annals of the New York Academy of Sciences* 86, 844–74.

Hammersley, J.M. and Handscomb, D.C. 1964. *Monte Carlo Methods*. London: Methuen.

Hansen, L.P. 1982. Large sample properties of generalized method of moments. *Econometrica* 50, 1029–54.

Hansen, L.P. and Sargent, T.J. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2, 7–46.

Hansen, L.P. and Sargent, T.J. 2007. *Robustness*. Princeton: Princeton University Press.

Hart, B.S. and von Neumann, J. 1942. Tabulation of the probabilities for the ratio of mean square successive difference to the variance. *Annals of Mathematical Statistics* 13, 207–14.

Härdle, W. 1990. *Applied Nonparametric Estimation*. Cambridge: Cambridge University Press.

Härdle, W. and Linton, O. 1994. Applied nonparametric methods. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D. McFadden. Amsterdam: North-Holland.

Harvey, A. 1989. *Forecasting, Structural Time Series Models and Kalman Filter*. Cambridge: Cambridge University Press.

Hastings, W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109.

Hausman, J.A. 1978. Specification tests in econometrics. *Econometrica* 46, 1251–72.

Hausman, J.A., Hall, B.H. and Griliches, Z. 1984. Econometric models for count data with application to the patents–R&D relationship. *Econometrica* 52, 909–1038.

Hausman, J.A. and Newey, W.K. 1995. Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica* 63, 1445–76.

Hausman, J.A. and Wise, D.A., eds. 1985. *Social Experimentation*. NBER Conference Report.Chicago: University of Chicago Press.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–94.

Heckman, J.J. and Singer, B. 1984. Econometric duration analysis. *Journal of Econometrics* 24, 63–132.

Heckman, J.J. and Smith, A.J. 1995. Assessing the case for social experimentation. *Journal of Economic Perspectives* 9(2), 85–110.

Heckman, J.J. and Willis, R. 1977. A beta-logistic model for the analysis of sequential labour force participation by married women. *Journal of Political Economy* 85, 27–58.

Hendricks, K. and Porter, R.H. 1988. An empirical study of an auction with asymmetric information.

American Economic Review 78, 865–83.

Hendry, D.F. and Richard, J.-F. 1982. On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics* 20, 3–33.

Holly, S., Pesaran, M.H. and Yamagata, T. 2006. A spatio-temporal model of house prices in the US. Mimeo, University of Cambridge.

Holtz-Eakin, D., Newey, W.K. and Rosen, H.S. 1988. Estimating vector autoregressions with panel data. *Econometrica* 56, 1371–95.

Honoré, B. and Kyriazidou, E. 2000. Panel data discrete choice models with lagged dependent variables. *Econometrica* 68, 839–74.

Hooker, R.H. 1901. Correlation of the marriage rate with trade. *Journal of the Royal Statistical Society* 44, 485–92.

Horowitz, J.L. 1998. *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.

Horowitz, J.L. 2001. Semiparametric models. In *International Encyclopedia of Behavioral and Social Sciences*, ed. N.J. Smelser and P.B. Baltes. Elsevier. Amsterdam: Elsevier.

Horowitz, J.L. and Lee, S. 2002. Semiparametric methods in applied econometrics: do the models fit the data? *Statistical Modelling* 2, 3–22.

Horowitz, J.L. and Savin, N.E. 2001. Binary response models: logits, probits, and semiparametrics. *Journal of Economic Perspectives* 15(4), 43–56.

Hotz, V.J., Mullin, C.H. and Sanders, S.G. 1997. Bounding causal effects using data from a contaminated natural experiment: analyzing the effects of teenage childbearing. *Review of Economic Studies* 64, 575–603.

Hsiao, C. 2003. *Analysis of Panel Data*, 2nd edn. Cambridge: Cambridge University Press.

Hsiao, C. and Pesaran, M.H. 2007. Random coefficient panel data models. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, 3rd edn. ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Hsiao, C., Pesaran, M.H. and Tahmiscioglu, A.K. 1999. Bayes estimation of short-run coefficients in dynamic panel data models. In *Analysis of Panels and Limited Dependent Variables Models*, ed. C. Hsiao et al. Cambridge: Cambridge University press.

Hsiao, C., Pesaran, M.H. and Tahmiscioglu, A.K. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109, 107–50.

Im, K.S., Pesaran, M.H. and Shin, Y. 2003. Testing for unit roots in heterogenous panels. *Journal of Econometrics* 115, 53–74.

Imbs, J., Mumtaz, H., Ravn, M.O. and Rey, H. 2005. PPP strikes back, aggregation and the real exchange rate. *Quarterly Journal of Economics* 120, 1–43.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–54. Reprinted in *Long-run Economic Relationships*, ed. R.F. Engle and C.W.J. Granger. Oxford: Oxford University Press, 1991.

Johansen, S. 1991. Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica* 59, 1551–80.

Johansen, S. 1995. *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Jorgenson, D.W. 1966. Rational distributed lag functions. *Econometrica* 34, 135–49.

Judd, K.L. 1998. *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Juselius, K. 2006. *The Cointegrated VAR Model: Econometric Methodology and Macroeconomic Applications*. Oxford: Oxford University Press.

Kagel, J. and Roth, A.E., eds. 1995. *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

Kandel, S. and Stambaugh, R.F. 1996. On the predictability of stock returns: an asset-allocation perspective. *Journal of Finance* 51, 385–424.

Keane, M.P. 1990. A computationally practical simulation estimator for panel data, with applications to estimating temporal dependence in employment and wages. Discussion paper, University of Minnesota.

Keane, M. and Wolpin, K.I. 1994. The solution and estimation of discrete choice dynamic programming models by simulation: Monte Carlo evidence. *Review of Economics and Statistics* 76, 648–72.

Keynes, J.M. 1939. The statistical testing of business cycle theories. *Economic Journal* 49, 558–68.

Kilian, L. 1998. Small-sample confidence intervals for impulse response functions. *Review of*

Economics and Statistics 80, 218–29.

Kim, K. and Pagan, A.R. 1995. The econometric analysis of calibrated macroeconomic models. In *Handbook of Applied Econometrics: Macroeconomics*, ed. M.H. Pesaran and M. Wickens. Oxford: Basil Blackwell.

Klein, L.R. 1947. The use of econometric models as a guide to economic policy. *Econometrica* 15, 111–51.

Klein, L.R. 1950. *Economic Fluctuations in the United States 1921–1941*. Cowles Commission Monograph No. 11. New York: John Wiley.

Kloek, T. and van Dijk, H.K. 1978. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica* 46, 1–20.

Koop, G. 2003. *Bayesian Econometrics*. Chichester: Wiley.

Koop, G., Pesaran, M.H. and Potter, S.M. 1996. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics* 74, 119–47.

Koop, G. and Potter, S. 2004a. Forecasting and estimating multiple change-point models with an unknown number of change-points. Mimeo, University of Leicester and Federal Reserve Bank of New York.

Koop, G. and Potter, S. 2004b. Prior elicitation in multiple change-point models. Mimeo, University of Leicester and Federal Reserve Bank of New York.

Koopmans, T.C. 1937. *Linear Regression Analysis of Economic Time Series*. Haarlem: De Erven F. Bohn for the Netherlands Economic Institute.

Koopmans, T.C. 1949. Identification problems in economic model construction. *Econometrica* 17, 125–44.

Koopmans, T.C., Rubin, H. and Leipnik, R.B. 1950. Measuring the equation systems of dynamic economics. In *Statistical Inference in Dynamic Economic Models*, ed. T.C. Koopmans. Cowles Commission Monograph No. 10. New York: John Wiley.

Koyck, L.M. 1954. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.

Laffont, J-J., Ossard, H. and Vuong, Q. 1995. Econometrics of first-price auctions. *Econometrica* 63, 953–80.

Lancaster, T. 2004. *An Introduction to Modern Bayesian Econometrics*. Malden MA: Blackwell.

Leamer, E.E. 1978. *Specification Searches: Ad Hoc Inference with Non-experimental Data*. New York: John Wiley.

Lee, B.S. and Ingram, B. 1991. Simulation estimation of time-series models. *Journal of Econometrics* 47, 197–205.

Lee, L.F. 1995. Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. *Economic Theory* 11, 437–83.

Lenoir, M. 1913. *Etudes sur la formation et le mouvement des prix*. Paris: Giard et Brière.

Lerman, S. and Manski, C.S. 1981. On the use of simulated frequencies to approximate choice probabilities. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

Levin, A., Lin, C. and Chu, C.J. 2002. Unit root tests in panel data: asymptotic and finite-sample properties. *Journal of Econometrics* 108, 1–24.

Liesenfeld, R. and Breitung, J. 1999. Simulation based method of moments. In *Generalized Method of Moment Estimation*, ed. L. Tatyas. Cambridge: Cambridge University Press.

Litterman, R.B. 1985. Forecasting with Bayesian vector autoregressions: five years of experience. *Journal of Business and Economic Statistics* 4, 25–38.

Liu, T.C. 1960. Underidentification, structural estimation and forecasting. *Econometrica* 28, 855–65.

Lo, A. and MacKinlay, C. 1988. Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies* 1, 41–66.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R.E. 1973. Some international evidence on output-inflation tradeoffs. *American Economic Review* 63, 326–34.

Lucas, R.E. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A. M. Meltzer. Amsterdam: North-Holland.

Lucas, R.E. and Sargent, T. 1981. Rational expectations and econometric practice. Introduction to *Rational Expectations and Econometric Practice*. Minneapolis: University of Minnesota Press.

L•tkepohl, H. 1991. *Introduction to Multiple Time Series Analysis*. New York: Springer-Verlag.

Lyttkens, E. 1970. Symmetric and asymmetric estimation methods. In *Interdependent Systems*, ed. E. Mosback and H. Wold. Amsterdam: North-Holland.

Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maddala, G.S. 1986. Disequilibrium, self-selection, and switching models. In *Handbook of Econometrics*, vol. 3, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Maddala, G.S. 2001. *Introduction to Econometrics*, 3rd edn. New York: John Wiley and Sons.

Manski, C.F. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C.F. 1985. Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27, 313–34.

Manski, C.F. 1988. Identification of binary response models. *Journal of the American Statistical Association* 83, 729–38.

Manski, C.F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C.F. 2003. *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C.F. and McFadden, D. 1981. *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.

Matzkin, R.L. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of Econometrics*, vol. 4, ed. R. F. Engle and D. L. McFadden. Amsterdam: North-Holland.

McAleer, M., Pagan, A.R. and Volker, P.A. 1985. What will take the con out of econometrics? *American Economic Review* 75, 293–307.

McAleer, M. and Medeiros, M.C. 2007. Realized volatility: a review. *Econometric Reviews*.

McCulloch, R.E. and Rossi, P.E. 1994. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* 64, 207–40.

McCulloch, R.E. and Tsay, R.S. 1993. Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association* 88, 965–78.

McCulloch, R.E. and Tsay, R.S. 1994. Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis* 15, 235–50.

McCulloch, R.E., Polson, N.G. and Rossi, P.E. 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* 99, 173–93.

McFadden, D. 1989. A method of simulated moments for estimation of multinomial probits without numerical integration. *Econometrica* 57, 995–1026.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087–92.

Mitchell, W.C. 1928. *Business Cycles: The Problem in its Setting*. New York: NBER.

Moon, R. and Perron, B. 2004. Testing for unit root in panels with dynamic factors. *Journal of Econometrics* 122, 81–126.

Moon, R. and Perron, B. 2007. An empirical analysis of nonstationarity in a panel of interest rates with factors. *Journal of Applied Econometrics* 22, 383–400.

Moore, H.L. 1914. *Economic Cycles: Their Law and Cause*. New York: Macmillan.

Moore, H.L. 1917. *Forecasting the Yield and the Price of Cotton*. New York: Macmillan Press.

Mortensen, D.T. 1990. Equilibrium wage distributions: a synthesis. In *Panel Data and Labor Market Studies*, ed. J. Hartog, G. Ridder and J. Theeuwes. New York: North Holland.

Mundlak, Y. 1961. Empirical production function free of management bias. *Journal of Farm Economics* 43, 44–56.

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica* 46, 69–85.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Nadaraya, E.A. 1964. On estimating regression. *Theory of Probability and Its Applications* 10, 141–2.

Nakamura, A. and Nakamura, M. 1981. On the relationships among several specification error tests presented by Durbin, Wu, and Hausman. *Econometrica* 49, 1583–88.

Nelson, C.R. 1972. The prediction performance of the FRB-MIT-Penn model of the US economy. *American Economic Review* 62, 902–17.

Nerlove, M. 1958a. Adaptive expectations and the cobweb phenomena. *Quarterly Journal of Economics* 72, 227–40.

Nerlove, M. 1958b. *Distributed Lags and Demand Analysis*. Washington, DC: USDA.

Newey, W.K. 1997. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79, 147–68.

Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49, 1399–416.

Niederreiter, H. 1992. *Random Number Generation and Quasi-Monte Carol Methods*. Philadelphia: SIAM.

Nyblom, J. 1989. Testing for the constancy of parameters over time. *Journal of the American Statistical Association* 84, 223–30.

O'Connell, P. 1998. The overvaluation of purchasing power parity. *Journal of International Economics* 44, 1–19.

Orcutt, G.H. 1948. A study of the autoregressive nature of the time series used for Tinbergen's model of the economic system of the United States, 1919–1932 (with discussion). *Journal of the Royal Statistical Society* Series B 10, 1–53.

Pagan, A.R. and Pesaran, M.H. 2007. On econometric analysis of structural systems with permanent and transitory shocks and exogenous variables. Unpublished manuscript.

Pagan, A. and Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Pakes, A. and Pollard, D. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–58.

Peersman, G. 2005. What caused the early millennium slowdown? Evidence based on autoregressions. *Journal of Applied Econometrics* 20, 185–207.

Pedroni, P. 2001. Purchasing power parity tests in cointegrated panels. *Review of Economics and Statistics* 83, 727–31.

Pedroni, P. 2004. Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econometric Theory* 20, 597–625.

Percy, D.F. 1992. Prediction for seemingly unrelated regressions. *Journal of the Royal Statistical Society*, Series B 54, 243–52.

Pesaran, M.H. 1981. Identification of rational expectations models. *Journal of Econometrics* 16, 375–98.

Pesaran, M.H. 1987a. Econometrics. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Pesaran, M.H. 1987b. *The Limits to Rational Expectations*. Oxford: Basil Blackwell.

Pesaran, M.H. 2006. Estimation and inference in large heterogeneous panels with cross section dependence. *Econometrica* 74, 967–1012.

Pesaran, M.H. 2007a. A simple panel unit root test in the presence of cross section dependence. *Journal of Applied Econometrics* 22, 265–312.

Pesaran, M.H. 2007b. A pair-wise approach to testing for output and growth convergence. *Journal of Econometrics* 138, 312–55.

Pesaran, M.H. and Deaton, A.S. 1978. Testing non-nested nonlinear regression models. *Econometrica* 46, 677–94.

Pesaran, M.H., Pettenuzzo, D. and Timmermann, A. 2006. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies* 73, 1057–84.

Pesaran, M.H., Schuermann, T. and Weiner, S.M. 2004. Modelling regional interdependencies using a global error-correcting macroeconometric model (with discussion). *Journal of Business and Economic Statistics* 22, 129–62, 175–81.

Pesaran, M.H. and Skouras, S. 2002. Decision-based methods for forecast evaluation. In *A Companion to Economic Forecasting*, ed. M.P. Cements and D.F. Hendry. Oxford: Blackwell Publishing.

Pesaran, M.H. and Smith, R.P. 1985a. Keynes on econometrics. In *Keynes' Economics: Methodological Issues*, ed. T. Lawson and M.H. Pesaran. London: Croom Helm.

Pesaran, M.H. and Smith, R.P. 1985b. Evaluation of macroeconometric models. *Economic Modelling* 2, 125–34.

Pesaran, M.H. and Smith, R. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68, 79–113.

Pesaran, M.H. and Smith, R.P. 2006. Macroeconometric modelling with a global perspective. *Manchester School* 74, 24–49.

Pesaran, M.H. and Timmermann, A. 1995. The robustness and economic significance of predictability of stock returns. *Journal of Finance* 50, 1201–28.

Pesaran, M.H. and Timmermann, A. 2000. A recursive modelling approach to predicting UK stock returns. *Economic Journal* 110, 159–91.

Pesaran, M.H. and Timmermann, A. 2005a. Real time econometrics. *Econometric Theory* 21, 212–31.

Pesaran, M.H. and Timmermann, A. 2005b. Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics* 129, 183–217.

Pesaran, M.H. and Timmermann, A. 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137, 134–61.

Pesaran, M.H. and Weeks, M. 2001. Non-nested hypothesis testing: an overview. In *Companion to Theoretical Econometrics*, ed. B.H. Baltagi. Oxford: Basil Blackwell.

Phillips, P.C.B. 1983. Exact small sample theory in the simultaneous equations model. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M. D. Intrilgator. Amsterdam: North-Holland.

Phillips, P.C.B. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–40.

Phillips, P.C.B. 1991. Optimal inference in cointegrated systems. *Econometrica* 59, 283–306.

Phillips, P.C.B. and Moon, H.R. 1999. Linear regression limit theory for nonstationary panel data. *Econometrica* 67, 1057–111.

Phillips, P.C.B. and Xiao, Z. 1998. A primer on unit root testing. *Journal of Economic Surveys* 12, 423–69.

Powell, J.L. 1984. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25, 303–25.

Powell, J.L. 1986. Censored regression quantiles. *Journal of Econometrics* 32, 143–55.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. 1986. *Numerical Recipes: The Art of Scientific Computing*, 1st edn. Cambridge: Cambridge University Press.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. 1992. *Numerical Recipes: The Art of Scientific Computing*, 2nd edn. Cambridge: Cambridge University Press.

Prothero, D.L. and Wallis, K.F. 1976. Modelling macroeconomic time series. *Journal of the Royal Statistical Society*, Series A 139, 468–86.

Quandt, R.E. 1982. Econometric disequilibrium models. *Econometric Reviews* 1, 1–63.

Ramsey, J.B. 1969. Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society*, Series B 31, 350–71.

Reiersol, O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9, 1–24.

Reiersol, O. 1945. Confluence analysis by means of instrumental sets of variables. *Arkiv for Mathematik Astronomi och Fysik* 32, 1–119.

Rossi, P.E., Allenby, G.M. and McCulloch, R. 2005. *Bayesian Statistics and Marketing*. Chichester: Wiley.

Rothenberg, T.J. 1984. Approximating the distributions of econometric estimators and test statistics. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Rust, J. 1994. Structural estimation of Markov decision processes. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Rust, J. 1997. Using randomization to break the curse of dimensionality. *Econometrica* 65, 487–516.

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 41–9.

Sandor, Z. and Andras, P. 2004. Alternative sampling methods for estimating multivariate normal probabilities. *Journal of Econometrics* 120, 207–34.

Santos, M.S. and Vigo-Aguiar, J. 1998. Analysis of a numerical dynamic programming algorithm applied to economic models. *Econometrica* 66, 409–26.

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica*

26, 393–415.

Sargan, J.D. 1964. Wages and prices in the United Kingdom: a study in econometric methodology. In *Econometric Analysis for National Economic Planning*, ed. P.E. Hart, G. Mills and J.K. Whitaker. London: Butterworths.

Sargent, T.J. 1973. Rational expectations, the real rate of interest and the natural rate of unemployment. *Brookings Papers on Economic Activity* 1973(2), 429–72.

Sargent, T.J. and Wallace, N. 1975. Rational expectations and the theory of economic policy. *Journal of Monetary Economics* 2, 169–84.

Savin, N.E. 1973. Systems k-class estimators. *Econometrica* 41, 1125–36.

Schultz, M. 1938. *The Theory and Measurement of Demand*. Chicago: University of Chicago Press.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: George Allen & Unwin.

Shephard, N., ed. 2005. *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press.

Shiller, R.J. 1973. A distributed lag estimator derived from smoothness priors. *Econometrica* 41, 775–88.

Sims, C.A. 1972. Money, income and causality. *American Economic Review* 62, 540–52.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Sims, C.A. 1982. Policy analysis with econometric models. *Brookings Papers on Economic Activity* 1982 (1), 107–64.

Slutsky, E. 1927. The summation of random causes as the source of cyclic processes. In *Problems of Economic Conditions*, vol. 3. Moscow. English trans. in *Econometrica* 5 (1937), 105–46.

Smets, F. and Wouters, R. 2003. An estimated stochastic dynamic general equilibrium model of the euro area. *Journal of the European Economic Association* 1, 1123–75.

Smith, V., Leybourne, S., Kim, T.-H. and Newbold, P. 2004. More powerful panel data unit root tests with an application to mean reversion in real exchange rates. *Journal of Applied Econometrics* 19, 147–70.

Solow, R.M. 1960. On a family of lag distributions. *Econometrica* 28, 393–406.

Srivastava, V.K. 1971. Three-stage least-squares and generalized double k-class estimators: a mathematical relationship. *International Economic Review* 12, 312–16.

Stambaugh, R.F. 1999. Predictive regressions. *Journal of Financial Economics* 54, 375–421.

Stern, S. 1997. Simulation-based estimation. *Journal of Economic Literature* 35, 2006–39.

Stock, J.H. 1994. Unit roots, structural breaks and trends. In *Handbook of Econometrics*, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Stock, J.H. and Watson, M.W. 1996. Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* 14, 11–30.

Stock, J.H., Wright, J.H. and Yogo, M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20, 518–29.

Stone, C.J. 1985. Additive regression and other nonparametric models. *Annals of Statistics* 13, 689–705.

Stone, J.R.N. 1945. The analysis of market demand. *Journal of the Royal Statistical Society*, Series A 108, 286–382.

Stone, J.R.N. 1978. Keynes, political arithmetic and econometrics. Seventh Keynes Lecture in Economics, British Academy.

Stone, J.R.N. et al. 1954. *Measurement of Consumers' Expenditures and Behavior in the United Kingdom. 1920–38*, 2 vols. London: Cambridge University Press.

Strachan, R.W. and van Dijk, H.K. 2006. Model uncertainty and Bayesian model averaging in vector autoregressive processes. Discussion Papers in Economics 06/5, Department of Economics, University of Leicester.

Swamy, P.A.V.B. 1970. Efficient inference in a random coefficient regression model. *Econometrica* 38, 311–23.

Taylor, J.B. and Uhlig, H. 1990. Solving nonlinear stochastic growth models: a comparison of alternative solution methods. *Journal of Business and Economic Statistics* 8, 1–18.

Theil, H. 1954. Estimation of parameters of econometric models. *Bulletin of International Statistics Institute* 34, 122–8.

Theil, H. 1958. *Economic Forecasts and Policy*. Amsterdam: North-Holland; 2nd edn, 1961.

Tierney, L. 1994. Markov chains for exploring posterior distributions with discussion and rejoinder. *Annals of Statistics* 22, 1701–62.

Tinbergen, J. 1929–30. Bestimmung und Deutung von Angebotskurven: ein Beispiel. *Zeitschrift für Nationalökonomie* 1, 669–79.

Tinbergen, J. 1937. *An Econometric Approach to Business Cycle Problems*. Paris: Herman & Cie Editeurs.

Tinbergen, J. 1939. *Statistical Testing of Business Cycle Theories. Vol. 1: A Method and its Application to Investment activity; Vol. 2: Business Cycles in the United States of America. 1919–1932*. Geneva: League of Nations.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26, 24–36.

Treadway, A.B. 1971. On the multivariate flexible accelerator. *Econometrica* 39.

Trivedi, P.K. 1975. Time series analysis versus structural models: a case study of Canadian manufacturing behaviour. *International Economic Review* 16, 587–608.

Uhlig, H. 2005. What are the effects of monetary policy: results from an agnostic identification approach. *Journal of Monetary Economics* 52, 381–419.

Verbeek, M. 2007. Pseudo panels and repeated cross-sections. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, 3rd edn. ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 12, 367–95.

von Neumann, J. 1942. A further remark on the distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 13, 86–8.

von Neumann, J. 1951. Various techniques used in connection with random digits. *Applied Mathematics Series* 12, 36–8. US National Bureau of Standards.

Wallis, K.F. 1977. Multiple time series analysis and the final form of econometric models. *Econometrica* 45, 1481–97.

Wallis, K. 1980. Econometric implications of the Rational Expectations Hypothesis. *Econometrica* 48, 49–73.

Watson, G.M. 1964. Smooth regression analysis. *Sankhyâ*, Series A 26, 359–72.

Watson, M.W. 1994. Vector autoregressions and cointegration. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Wegge, L.L. 1965. Identifiability criteria for a system of equations as a whole. *Australian Journal of Statistics* 7, 67–77.

West, K.D. 2006. Forecast evaluation. In *Handbook of Economic Forecasting*, vol. 1, ed. G. Elliott, C. Granger and A. Timmermann. Amsterdam: North-Holland.

Westerlund, J. 2005. Panel cointegration tests of the Fisher effect. Working Papers 2005:10, Lund University, Department of Economics.

White, H. 1981. Consequences and detection of misspecified nonlinear regression models. J*ournal of the American Statistical Association* 76, 419–33.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–26.

Whittle, P. 1963. *Prediction and Regulation by Linear Least-squares Methods*. London: English Universities Press.

Wickens, M. 1982. The efficient estimation of econometric models with rational expectations. *Review of Economic Studies* 49, 55–68.

Wickens, M.R. and Motto, R. 2001. Estimating shocks and impulse response functions. *Journal of Applied Econometrics* 16, 371–87.

Wold, H. 1938. *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksell.

Wooldridge, J.M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J.M. 2006. *Introductory Econometrics: A Modern Approach*, 3rd edn. Stamford: Thomson-South-Western.

Working, E.J. 1927. What do statistical ‘demand curves’ show? *Quarterly Journal of Economics* 41, 212–35.

Wright, P.G. 1915. Review of economic cycles by Henry Moore. *Quarterly Journal of Economics* 29,

631–41.

Wright, P.G. 1928. *The Tariff on Animal and Vegetable Oils*. London: Macmillan for the Institute of Economics.

Wu, D. 1973. Alternatives tests of independence between stochastic regressor and disturbances. *Econometrica* 41, 733–50.

Yule, G.U. 1895, 1896. On the correlation of total pauperism with proportion of out-relief. *Economic Journal* 5, 603–11; 6, 613–23.

Yule, G.U. 1921. On the time-correlation problem, with special reference to the variate-difference correlation method. *Journal of the Royal Statistical Society* 84, 497–526.

Yule, G.U. 1926. Why do we sometimes get nonsense correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society* 89, 1–64.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57, 348–68.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons.

Zellner, A. 1984. *Basic Issues in Econometrics*. Chicago: University of Chicago Press.

Zellner, A. 1985. Bayesian econometrics. *Econometrica* 53, 253–70.

Zellner, A. and Palm, F. 1974. Time series analysis and simultaneous equation econometric models. *Journal of Econometrics* 2, 17–54.

Zellner, A. and Theil, H. 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica* 30, 54–78.

How to cite this article

Geweke, John, Joel L. Horowitz and Hashem Pesaran. "econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000007>
doi:10.1057/9780230226203.0425

The New Palgrave Dictionary of Economics Online

economic anthropology

Timothy Earle

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic anthropology is an empirical science that describes production, exchange and consumption cross-culturally. All societies have economies, but they are variable. Anthropologists evaluate the operations of individual economies and the applicability of Western theories to these cases. Some economic processes work broadly; for example, strategic decision-making, the law of competitive advantage, and calculations of transaction costs help explain many observed patterns. Human economies, however, are often structured as intertwined sectors with distinctive processes. Differences observed in productivity, specialization, institutional structure and social motivations across history and across modern societies are of theoretical significance when constructing the limits of general theory.

Keywords

behavioural economics; capitalism; cognitive ability; commodity chains; division of labour; domestic mode of production; dual economies; economic anthropology; formalism; gift exchange; household production; hunting and gathering economies; identity; institutional economics; markets; Polanyi, K.; political sector; postmodernism; prestige economy; rationality; reciprocity; social networks; social sector; specialization; staple finance and wealth finance; subsistence sector; substantivism; surplus; technological progress; transactions costs; Weber, M.

Article

Economic anthropology is an empirical science that seeks to describe how production, exchange and consumption operate outside the West (compare Hunt, 1997). The second edition (1952) of Herskovits's (1940) text, titled *Economic Anthropology*, labelled this sub-discipline in anthropology. The broader mission of anthropology has been to make sense of the diversity in the human experience, which became apparent to Europeans during progressive stages of exploration, colonialization and globalization. Underlying anthropological research is the premise that human societies have developed parallel

institutions of aesthetics, religion, kinship, politics, and of course economics. All societies have economies, and the economic patterns observed in non-Western economies both comfort and confront theories developed by Western scholars.

Common economic processes, such as rational decision-making, law of competitive advantage, and institutional economics help explain many patterns across human economies based on variable conditions of cost, demand and availability. Additionally, however, human economies appear often to be structured quite differently from Western models, and these differences in institutional structure and motivation are of theoretical significance. From the beginning, economic anthropology has contained, and more or less successfully resolved, a tension between the desire to find cross-culturally general theories and to recognize the uniqueness of each individual case. In economic anthropology this tension has been represented in the formalist–substantivist debate.

Few anthropologists identify themselves primarily as economic anthropologists, but study economic matters as part of a broadly integrative approach to human societies. Founded in 1980, the Society of Economic Anthropology is the primary organization for anthropologists with such interests. Members include ethnographers, applied development anthropologists, archaeologists and ethnohistorians, suggesting that economic studies bridge the diversity of the discipline. The society sponsors annual meetings on themes that range across topics including key institutions of labour, property, markets and consumption, and special topics from the gift to slow foods. *Research Series in Economic Anthropology* and *Society for Economic Anthropology Monographs* offer edited volumes on the sub-discipline.

History of economic anthropology

From early in the 20th century, anthropologists have questioned whether theories developed to understand Western market economies apply only to those Western societies for which they were generated. To answer this question, anthropologists have described traditional economies, which survived into the 20th century, which existed in the past, and which have been transformed by engagement with the West. Largely empirical, the work is of substantial theoretical significance for understanding economies cross-culturally. Gudeman (1998) has compiled many of the most highly referenced articles.

Economic anthropology's beginning traces to the landmark ethnography *Argonauts of the Western Pacific*, in which Malinowski (1922) described the circulation of shell valuables among the islands of the Kula Ring. Malinowski used the Trobriand Islanders' obsession with certain shell valuables to challenge simplistic notions of 'economic man', and he argued that a non-Western economy could be fundamentally different from modern market economies in values and socialized exchange relationships.

Anthropological studies of traditional economies thrived during the first half of the 20th century. As part of British functionalism, Malinowski and his students developed the approach; in French structuralism, Mauss (1925) focused on the gift as a social phenomenon; and, within American anthropology, Herskovits (1940) defined the sub-field. Much of the work was descriptive, emphasizing how traditional people meet basic needs and how the exchange of primitive valuables fashioned and maintained social relationships.

By mid-century, however, studies of traditional economy were increasingly adopting the terms and concepts of Western economic theory. Both Herskovits (1940) and Firth (1939) revised their original

books on traditional economies so as to clarify underlying similarities across world economies. They each took concepts, like scarcity and specialization, and generalized them to show that they apply well to societies in which market penetration is not great. They were making the essential point that traditional economies were not simply driven by the food quest. Although most anthropologists took pains to emphasize the differences between traditional economies and market-integrated systems, some seemed to homogenize the human experience, and a sharp reaction followed.

In the tradition of Max Weber, economic historian Karl Polanyi (1944) wrote his famous treatise *The Great Transformation* to argue that the integrating structure of modern markets, for which prices are set by supply and demand, are a very recent creation of industrialism and capitalism. Theories based on scarcity, rationality, equilibrating price mechanisms operated, he argued, only in the special case of Western capitalism. Modern market conditions should not be taken as inherent in the human experience, but as a recent social artifact malleable in future societies.

Polanyi's impact on economic anthropology was profound and created the debate between substantivists and formalists that raged in the sub-discipline for a generation. *Trade and Markets in the Early Empires* (Polanyi, Arensberg and Pearson, 1957), the seminal edited book, came out of a discussion group which Polanyi led at Columbia University and which included anthropologists who would be influential in the field. Polanyi's chapter ‘The Economy as Instituted Process’ characterized the substantivist approach. He defined three forms of distribution found in societies with different structured relationships: reciprocity in egalitarian relationships; redistribution in hierarchical relationships; and market exchange in the anonymous relationships of the market. Because economic relations were so deeply embedded in social structure, variation in social organization was thought to explain the differences in the economies. Substantivists recognized that markets were found widely in traditional economies, but argued that those markets were peripheral to most economic activities, which were deeply embedded in social relationships (Bohannan and Dalton, 1962). A compendium collected by Dalton (1967) provided empirical cases that illustrate the embedded nature of traditional economies.

In his critique of those using economic theory in non-Western contexts, Polanyi labelled them as ‘formalist’, meaning that they focused on ‘formal’ (mathematical) maximizing models to predict how individuals choose among alternative possibilities to allocate limited time, money and other resources. The substantivists, in contrast, focused on how economies were embedded within cultural institutions to meet the material desires that particular culture might have. The debate raged between the two factions through the 1960s and 1970s. Much of the argument became focused on how extensive markets were in traditional societies. In a classic cross-cultural study, Pryor (1977) showed that markets were very broadly distributed, sometimes moving primitive valuable, tools and food. They certainly did not originate with modern capitalism. In his famously acerbic article, Cook (1969) criticized substantivists for being romantic and naive; after all, even if they had useful points to make, the penetration of market economies, he argued, was so pervasive that formalist theories were *now* effectively universal. Articles representing the two sides were collected in a reader by LeClair and Schneider (1968) that has been used to teach the debate ever since. Articulating the substantivist position, Sahlins (1972) then argued that many concepts of Western economic theory were inapplicable to traditional economy. He discussed the affluence of hunter-gatherers, underproduction in household economies, and the social determinants of reciprocal exchanges. Schneider (1974) countered with the fully articulated formalist position, summarizing how Western economics can be applied cautiously to a wide range of non-

Western transactions and decisions, including marriage payments, primitive money, the prestige economy and household production. The debate came to focus on definitions of rationality, scarcity and institutional constraints, but those reading the papers increasingly saw that the participants were talking past each other.

The formalist and substantivist factions represented the inherent tension within anthropology: on the one hand, to seek cross-cultural regularities that reflect shared social process; on the other, to recognize the cultural relativity and uniqueness of each culture. The two sides of the debate fought to exhaustion, as both presented compelling approaches that could be seen as more complementary than alternative. In 1980, Schneider helped organize the Society for Economic Anthropology in order to resolve the debate by bringing the full spectrum of economic anthropologists together. The first meeting, published as *Economic Anthropology: Topics and Theories* (Ortiz, 1983) gathered an eclectic group of scholars to bridge the theoretical divides within the sub-discipline, with broad interests in marketing, institutions, Marxism, ecology, and economic development. An edited text, *Economic Anthropology* (Plattner, 1989), provided a new generation of students with the breadth of economies and economic conditions that anthropologists were trying to make sense of.

Important to the new harmony has been respect for the different objectives of economic anthropologists, including ethnographic work on traditional economies, applied work on developing economies, and archaeological and historical studies of economies. The field has recognized diversity in both the theoretical and historical nature of human economies. To maintain a proper balance between substantivists and formalists (relativists and universalists) in economic anthropology, the role of archaeological and historical studies has been especially important. As ethnographers increasingly study variants of a single modern system, historical and archaeological studies continue to study the true variation in how human economies are organized and operate. Earle (2002), for example, looks at the alternative means by which political economies have emerged to finance the evolution of chiefdoms and states, showing that the development of market systems is quite rare and specific in that process. Although no careful comparative study exists, the extent of exchange in prehistory appears to have been highly variable.

During the 1980s and 1990s, as economic anthropology matured as a sub-discipline, it became marginalized within anthropology. As in many of the social sciences and humanities, postmodernism became popular, and its anti-materialist, anti-scientific critiques were antithetical to much of what the sub-discipline advocated. As the excesses of postmodernism have receded, however, economic anthropology has regained some of its former popularity, and its potential significance for anthropology and economics seems promising. Perhaps the greatest challenge now is that economics and economic anthropology have remained far apart because of the strongly formal (theoretical) basis of the former and empirical basis of the later. The two approaches would, however, seem complementary.

Economic anthropology and its perspective on world economies

Economists should consider the empirical value of economic anthropology, and a good place to begin is the compendium *Theory in Economic Anthropology* (Ensminger, 2002a). Economic anthropologists are committed to models of reality. The empirical observations and theoretical inferences of anthropology should help recognize the specific frames of applicability for grand theories. In essence, anthropology

makes clear that all things are never equal. In this section, I summarize a few conclusions derived from economic anthropology that make a difference to studies of economies. These involve human rationality, consumer behaviour, commodity chains, and the multi-sectored quality of human economies. This list is not meant to be exhaustive, but only to illustrate the importance of cross-cultural evaluations for the models that economists develop. As economics begins to look at such concepts as behavioural economics and personalized networks, the relevance of anthropology's research on these topics becomes particularly significant.

Human decision-making is to a degree rational, and empirical anthropological work significantly improves an understanding of decision-making processes from a cross-cultural and evolutionary perspective. Although rationality underpins much economic theorizing, human cognitive abilities and goals have been under theorized. Recent trends to rectify this within behavioural economics emphasize that individuals do not always act rationally with primary economic objectives and it would appear that economic anthropology could provide valuable cross-cultural validation of these new ideas. Humans prove to be fairly poor decision makers; they appear rather to use simplified proximate measures to estimate such considerations as value and cost (Henrich, 2002). Anthropologists have experimented with various economic games given under controlled conditions in non-Western societies, and their results are often counter-intuitive (Ensminger, 2002b). In a sample of societies representing different levels of economic development, for example, as market integration increases cooperation can be shown in such game-playing experiments to become more highly prized.

To understand the evolutionary roots of human rationality, anthropological research has looked at decision-making in small-scale hunting and gathering societies (see for example, Cashdan, 1989). As seen by the rapid expansion in brain size deep in history, humans must have been under strong selective pressure for expanded cognitive abilities, and this selective pressure took place when humans were low-density hunter-gatherers. Such hunter-gatherers make daily a wide range of decisions about what foods to eat, where to camp, what groups to join, and the like, and the relative scarcity and abundance of food and their different nutritional qualities appear to be considered. Human cognitive skill determines the ability of hunter-gatherers to adjust rapidly to changing conditions of food availability, to occupy diverse habitats from the Arctic to the tropical forests, and to intensify food procurement as required by population growth. In short, cognitive abilities in the food quest, in movement through the landscape, and in deciding which groups to join must have provided a strong selective advantage that resulted in the moulding of human rationality.

As illustrated by economic anthropology, human decisions often have little direct relationship to economic factors of cost and financial gain. Although of more interest recently to economists, with the notable exception of Thorstein Veblen, economic theory has not attempted systematically to explain how potential consumer outcomes are ranked. Rather, within the West, consumer behaviour has been studied with a rather eclectic and under-theorized set of assumptions. Anthropologists, however, have tried to understand consumption cross-culturally as a social process involving issues of identity and association (Rutz and Orlove, 1989). From the anthropological literature, we know how valued objects signify social relationships. The giving and receiving of gifts impart form and meaning to social relationships, and materialize the social distance between actors (Sahlins, 1972).

Economic anthropologists frequently study the movements of objects around the globe. These commodity chains describe how goods are produced, distributed and transformed as they move through

a sequence of markets (Hansen, 2002; Obukhova and Guyer, 2002). Commodity chains illustrate how goods, like used clothing, are transformed in value, form and meaning as they pass through a sequence of social worlds and economic sectors. Social considerations of prestige and personal worth are always of great concern in this highly creative process of economic decision-making.

Economic anthropologists have emphasized that economies are multilayered and that the specific character of an economy has historical routes. Although economists often refer to ‘dual economy’, implying a vestigial survival of traditional practices, they have been reluctant to accept that economies are always multilayered mosaics with spheres of exchange that only partially articulate the different sectors. Economic theory thus radically simplifies reality by focusing on decision-making and outcomes under market conditions, and this simplification makes very different economies appear superficial similar. In the emergence and development of capitalism, since wealth was made in the markets, the primary concern of economists became directed there. As anthropologists seek to understand the different motives and dynamics of economies as articulated in specific social contexts, they have, however, realized that human economies are highly variable, combining subsistence, social, political, and market sectors, each with distinct logics and historical traditions.

The subsistence sector is family-based and involves the daily struggles to meet basic needs. It is universal and represents the economic world of survival in which humans evolved as a species. The primary motivation of humans has probably always been the satisfaction of a family's basic needs. The construction of a general theory of human economies should thus start with how households and communities make a living. Until recently, household requirements were handled largely by family production. Although markets have a long history in human societies, they were typically quite marginal to subsistence needs. Theorized as the domestic mode of production (DMP; Sahlins, 1972), households were oriented to meeting their subsistence needs, and distribution involved sharing between family members with different tasks appropriate to an elementary division of labour by age and gender. In the model, the household is economically self-sufficient, and the economy is not inherently growth-oriented. The amazing conclusion of considerable anthropological research is that the DMP is often at least the model of what the economy should be, and the amounts of goods consumed by households that are produced outside the family have often been but a fraction of the households’ overall consumption budget. Prior to the development of full-scale markets, households probably produced 75 per cent or more of everything that they consumed.

The social sector is community-based and involves the lifetime strategies of individuals to define identity and relationships within a broader social group. The social sector is probably universal, finding its roots among early hunter-gatherers and their need to form networks of support, cooperation and exclusion. In cross-cultural perspective, much of the social sector involves reciprocal exchanges within highly social worlds that can be manipulated to emphasize personal prestige. In traditional societies, such competitive exchanges commonly produce social ranking in what has been called a ‘prestige economy’. The social sector was elaborated following the Neolithic revolution, as the creation of local corporate groups must have placed a premium on group identity and status. With deep and enduring roots in human history, the social sector would seem to provide a cross-cultural understanding of consumer behaviour as part of processes much broader than capitalism.

Economics now questions assumptions about anonymous markets organized independently of other social institutions. Goods and services are seen as flowing through personalized networks that create the institutions for expanding economic transactions. Greif (2006), for example, argues that the social

networks of medieval Europe provided the frameworks for an emergent modern economy. Almost self-evident to anthropologists, such conclusions suggest how economic theory can gain from insights from comparative empirical studies of non-industrial political forms.

Political sectors mobilize and allocate goods to finance regional and interregional institutions of domination and stratification (Earle, 2002). Importantly, political economies are not universal. From the fourth millennium bc, the political sector of the economy developed along with chiefdoms and then states. Goods became mobilized as a tax or tribute and then ‘redistributed’ by dominant political organizations as means to finance their activities. Recent archaeology has studied how political sectors were developed and functioned. An inherent contrast is between staple finance and wealth finance. In staple finance, food goods are mobilized and stored centrally as a means to support craftsmen, warriors and labourers working for the state. Many of these systems, especially in chiefdoms but also some states, functioned with few or no markets. Subsistence and social sectors continued largely unchanged, but new patterns of land ownership and domination required the production of a surplus for ruling institutions. Wealth finance worked similarly, but the local surplus was used to support the production of wealth for tribute payments.

And what about the market sector, so fundamental to most economic theorizing? Archaeological evidence documents that exchange and markets were not universal. From case to case, the amount and types of goods exchanged varied greatly according to specific conditions of availability and production costs and to specific objects of value. Based on ethnographic analogies, until quite recently most of the goods traded were probably handled by down-the-line exchanges between social partners. Goods moving any distance were primarily primitive valuables, items of display and tribute. The extent of exchange in Neolithic and later Bronze Age communities, for example, has been discussed for Europe, where the comparative advantage of one region over another would have been based on the availability of special materials (Sherratt, 1997). Subsistence and technological items were rarely exchanged over long distances until the end of the medieval age. Earlier, some market exchange certainly existed, but their extent and elaboration were apparently quite small.

This empirical record from economic anthropology contests economic theories based on asserted long-term trends in the emergence of marketing. A common assumption among economists from Adam Smith onwards has been that the creation of wealth is an outcome of the development of efficiencies associated with specialization and trade. For example, in his analysis of institutional economics, North (1990) argues that states developed to lower transaction costs between locally specialized but political independent regions. To simplify the logic, technological development and specialization should have created increasing productive economies that, with the emergence of integrating political systems to guarantee the peace of the market, would generate the surplus used to support the growth of civilizations. The development of markets, however, was quite late and episodic. Following North, economics might suggest that such failure of markets to develop was an outcome of high transaction costs that made exchange unprofitable. Empirically such a conclusion, however, can be shown to be wrong. As political superstructures were developed and imposed broad regional peace that would have radically lowered transactions costs, markets surprisingly did not emerge. The reason appears to be linked to the nature of finance. When finance was based on staples, markets were only rudimentary and peripheral. The complex Hawaiian chiefdoms, for example, conquered and integrated several islands with local specialties in food, stone and other materials, but trade remained very small-scale and local despite the regional peace. Archaeology has documented only minor trade in basaltic adzes and obsidian in

Hawaiian prehistory, and these exchanges did not increase with the formation of the large-scale chiefdoms. As a dramatic example, the Inka empire conquered a massive territory that extended 3,000• km up the spine of the Andes, imposed an effective regional peace across that territory, and constructed nearly 30,000•km of roads to integrate it. Although these actions would certainly have lowered transaction costs, the regional and distant movements of goods, like metal, ceramics, and foods, remained very limited and completely unchanged from the pre-imperial period (Earle, 2002).

Both markets and currencies seem to have expanded in other circumstances where they were linked with wealth finance of states. In the Aztec empire, tribute to the state was in wealth objects like textiles that could be easily transported long distances, centrally stored, and then used as payment to those working for the state. But the use of wealth objects in payment required that the objects be convertible into the staple goods and other consumables desired by state personnel. The Aztec market system provided the mechanism for conversion and was apparently developed by the state (Brumfiel, 1980). Afterwards, markets appear to have escaped from state sponsorship and control to take on many of the characterizations commonly associated with market systems.

What are the possibilities for a grand theory of economies? The relatively low status of historical and comparative studies within economics is not promising, but economics would do well to test theories claimed for generally applicability by looking closely at the anthropological literature. To the degree that economic models are used to design economic development in non-Western societies, the general relevance of the economic models must demonstrated. Using a uniform method of analysis, the economist Pryor (2005) has compared industrial economies and traditional (hunter-gatherer and agricultural) economies. His primary conclusions are startling, suggesting the advantages of such comparative analyses. All economies appear to consist of a small number of component parts, probably reflecting the processes and constraints involved in the production and movement of material goods. Economies are thus comparable. Furthermore, the factors that affect such variables as gross productivity or volume of exchange appear not to be determined by social structure but by the particular internal characteristics of the economy. Thus, Polanyi would appear to be wrong; economies are rather independent engines of essential processes. As recent work in economics has relaxed simplifying assumptions about information, frictionless trade, and anonymity of markets, the potential links between economics and economic anthropology take on reciprocal value.

See Also

- behavioural economics and game theory
- hunting and gathering economies
- ‘political economy’
- property rights
- stratification

Bibliography

Bohannan, P. and Dalton, G., eds. 1962. *Markets in Africa*. Evanston, IL: Northwestern University Press.

Brumfiel, E. 1980. Specialization, market exchange, and the Aztec state: a view from Huexotla. *Current Anthropology* 21, 459–78.

Cashdan, E. 1989. Hunters and gatherers: economic behavior in bands. In *Economic Anthropology*, ed. S. Plattner. Stanford: Stanford University Press.

Cook, S. 1969. The ‘anti-market’ mentality: a critique of the substantive approach to economic anthropology. *Southwestern Journal of Anthropology* 25, 378–406.

Dalton, G., ed. 1967. *Tribal and Peasant Economies*. Garden City, NY: Natural History Press.

Earle, T. 2002. *Bronze Age Economics*. Boulder, CO: Westview.

Ensminger, J., ed. 2002a. *Theory in Economic Anthropology*. Walnut Creek, CA: AltaMira.

Ensminger, J. 2002b. Experimental economics: a powerful new method for theory testing in anthropology. In Ensminger (2002a).

Firth, R. 1939. *Primitive Polynesian Economy*. London: Routledge & Kegan Paul, 1965.

Greif, A. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge: Cambridge University Press.

Gudeman, S. 1998. *Economic Anthropology*. Cheltenham: Edward Elgar.

Hansen, K.T. 2002. Commodity chains and the international secondhand clothing trade: *Salaula* and the work of consumption in Zambia. In Ensminger (2002a).

Henrich, J. 2002. Decision-making, cultural transmission and adaptation in economic anthropology. In Ensminger (2002a).

Herskovits, M. 1940. *Economic Anthropology*. New York: Knopf, 1952.

Hunt, R. 1997. Economic anthropology. In *The Dictionary of Anthropology*, ed. T. Barfield. Oxford: Blackwell.

Johnson, A. and Earle, T. 2000. *The Evolution of Human Societies*, 2nd edn. Stanford: Stanford University Press.

LeClair, E.E. and Schneider, H.K., eds. 1968. *Economic Anthropology*. New York: Holt, Rinehart.

Malinowski, B. 1922. *Argonauts of the Western Pacific*. London: Routledge.

Mauss, M. 1925. *The Gift: Forms and Functions of Exchange in Archaic Societies*. London: Cohen and West, 1969.

North, D. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.

Obukhova, E. and Guyer, J.I. 2002. Transcending the formal/informal distinction: commercial relations in Africa and Russia in the post-1989 world. In Ensminger (2002a).

Ortiz, S., ed. 1983. *Economic Anthropology: Topics and Theories*. Lanham, MD: University Press of America.

Plattner, S., ed. 1989. *Economic Anthropology*. Stanford: Stanford University Press.

Polanyi, K. 1944. *The Great Transformation*. New York: Rinehart.

Polanyi, K., Arensberg, C. and Pearson, H., eds. 1957. *Trade and Market in the Early Empires*. New York: Free Press.

Pryor, F.L. 1977. *The Origins of the Economy*. New York: Academic Press.

Pryor, F.L. 2005. *Economic Systems of Foraging, Agricultural, and Industrial Societies*. Cambridge: Cambridge University Press.

Rutz, H., Orlove, B., eds. 1989. *The Social Economy of Consumption*. Lanham, MD: University Press of America.

Sahlins, M. 1972. *Stone Age Economics*. Chicago: Aldine.

Schneider, H.K. 1974. *Economic Man: The Anthropology of Economics*. New York: Free Press.

Sherratt, A. 1997. *Economy and Society in Prehistoric Europe*. Edinburgh: Edinburgh University Press.

How to cite this article

Earle, Timothy. "economic anthropology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000009> doi:10.1057/9780230226203.0426

The New Palgrave Dictionary of Economics Online

economic calculation in socialist countries

Michael Ellman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the 1930s, when the classical socialist system emerged, economic decisions were based not on detailed and precise economic methods of calculation but on rough and ready political methods. An important method of economic calculation – particularly in the post-Stalin period – was that of incrementalism. Input norms were a very important method of both inter-industry and consumption planning. Material balances, and later input–output, were also widely used. Project evaluation, linear programming, comparisons with the West, and economic intuition were other methods used. The influence of methods of economic calculation on economic outcomes should not be exaggerated.

Keywords

coefficient of absolute economic effectiveness; coefficient of relative economic effectiveness; consumption norms method; Council for Mutual Economic Assistance (Comecon); economic calculation in social economies; Five Year Plans (USSR); Gosplan; Gossnab; incrementalist method of economic calculation; input norms method of economic calculation; input–output method of economic calculation; Kantorovich, L.; Lemeshev, M.; linear programming; material balances method; planned economy; planning and counter-planning; Stalinism, political economy of; Standard Methodology of economic calculation

Article

Economic calculation and political decisions

An important result of the archival revolution of the 1990s (that is, the access to former Soviet archives made possible by the collapse of the USSR) was the additional knowledge it provided about economic decision-making in the USSR in the Stalin period. This made it clear that in the 1930s, when the socialist economic system emerged, economic decisions were based not on detailed and precise economic methods of calculation but on rough and ready political methods. Interesting light has been thrown on the significance of this for macroeconomic, mesoeconomic and microeconomic decision making.

Macroeconomic policy in the Stalin era aimed to maximize investment subject to the need to provide sufficient consumer goods (mainly food) to maintain labour productivity. The consumer goods were obtained from agriculture by force and allocated by the state in a way which it was hoped would enable investment to be maximized. A schematic representation of short-term macroeconomic calculation under these circumstances is set out in Figure 1.

Figure 1
Maximizing the investible surplus. *Source*: adapted from Gregory and Harrison (2005, p. 732).

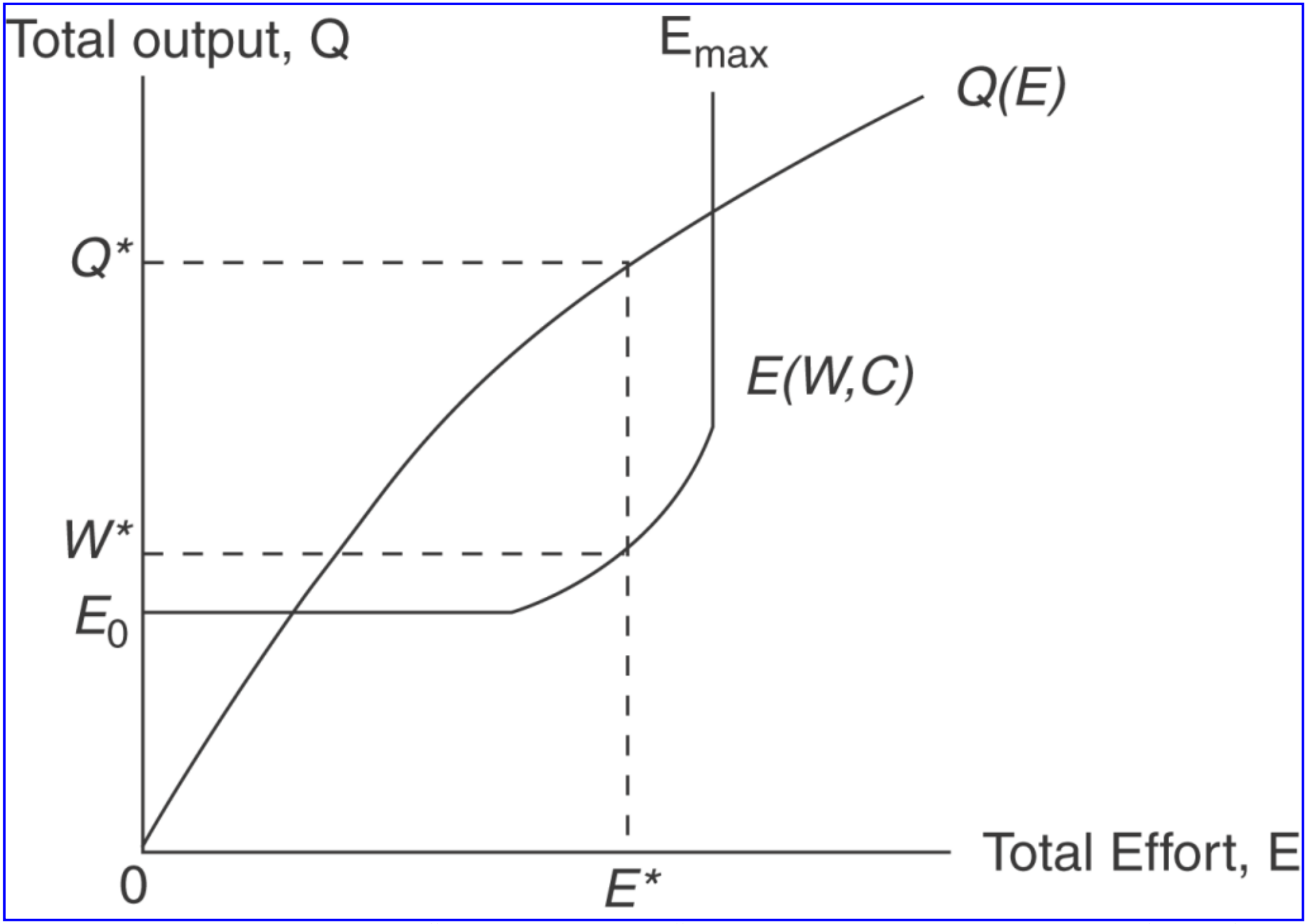

Figure 1 shows an output curve *OQ* which depends on the effort the workers provide, and an effort curve E_0E_{max} which depends on the real wage and the level of coercion. If the state chooses too low a level of wages, output will decline and the intended investment level will be impossible to meet. If wages are set at the fair wage level, output will be maximized but investment less than desired. At the wage level *W**, investment will be maximized. Hence, macroeconomic calculation involved gathering information about worker attitudes (via

the state security organizations), allocating the available food to crucial groups of workers, and using coercive or ideological methods to reduce the food–output ratio.
Mesopolicy aimed at developing heavy industry and the defence sector. An important result was what has been termed the ‘structural militarization’ of the Soviet economy. This resulted from the Soviet view of international relations, the stress on mobilization planning, the lessons of 1941, and the use by the general staff of absurdly inflated estimates of the mobilization capacity of the USA and other countries. An example is the USSR's capacity at the end of the 1980s to produce about four million tons of aluminium annually. This was greatly in excess of the peacetime economy's need for aluminium. However, in the event of mobilization it would have enabled the country to produce huge numbers of military airplanes. This situation arose as a result of using as a method of economic calculation the attainment of Western levels and of these levels in the military sphere being systematically exaggerated.
On the microeconomic level, Lazarev and Gregory (2003) have studied the allocation of motor vehicles (cars/autos and lorries/trucks) from the central reserve fund in 1932 and 1933. This showed that an economic planning model was unable to explain their allocation (in the regressions the economic variables were insignificant and frequently had the wrong signs). But a political model, in which their allocation was explained as part of a gift-exchange process, explained the data quite well.

Incrementalism

A basic method of economic calculation used in the state socialist countries – particularly in the post-Stalin period – was that of incrementalism, or, as it was known in the USSR, ‘planning from the achieved level’. The starting point of all economic plans was the actual or expected outcome of the previous period. The planners adjusted this by reference to anticipated growth rates, current economic policy, shortages and technical progress. For nearly all products, the planned output for next year was the anticipated output for this year plus a few per cent added on. The advantages of incrementalism as a method of economic calculation were its simplicity, realism and compatibility with the functioning of a hierarchical bureaucracy. Its disadvantages were that it provided no method for making technically efficient or consistent decisions, nor did it ensure that the population derived maximum satisfaction from the resources available.

Planning and counter-planning

A widely used method of economic calculation was that of planning and counter-planning. If the plan were simply handed down to the enterprises from above, in accordance with the planners’ view of national economic requirements but in ignorance of the real possibilities of each enterprise, then it would be unfeasible (if it was too high) or wasteful (if it was too low) or both at the same time (that is, unfeasible for some products and wasteful for others). Conversely, if plans were simply drawn up by each enterprise, they might have failed to use resources in accordance with national economic requirements. The process of planning and counter-planning involved a mutual submission and discussion of planning suggestions, designed to lead to the adoption of a plan which was feasible for the enterprise and ensured that the resources of each enterprise were used in accordance with national requirements.
Unfortunately, the bureaucratic complexity of this procedure militated against both efficiency and consistency.

Input norms

The main method of economic calculation used to ensure efficiency was that of input norms. An input norm is simply a number assumed to describe an efficient process of transformation of inputs into outputs. For example, suppose that the norm for the utilization of coal in the production of one ton of steel is x tons. Then the efficient production of z tons of steel is assumed to require zx tons of coal.
The method of norms was widely used in Soviet planning, and considerable effort was devoted to updating them. Very detailed norm fixing took place for expenditures of fuel and energy. Much attention was devoted to the development of norms for the expenditure of metal, cement, and timber in construction. All this work was directed by the department of norms and normatives of Gosplan (the State Planning Commission). Responsibility for elaborating and improving the norms lay with Gosplan's Scientific Research Institute of Planning and Norms.
Nevertheless, the method of norms was incapable of ensuring efficiency. The norms used in planning calculations were simply averages of input requirements, weighted somewhat in favour of efficient producers. Actual technologies showed a wide dispersion in input–output relations. Furthermore, given norms took no account of the possibilities of substitution of inputs for one another in the production process, non-constant returns to scale, and the results of technical progress. Thus in general, the method of norms did not make it possible to calculate efficient input requirements, and plans calculated in this way were always inefficient.
The method of norms was used not only in inter-industry planning but also in consumption planning. In calculating the volume of particular consumer goods and services required, the planners used two main methods. One was forecasts of consumer behaviour, based on extrapolation, expenditure patterns of higher-income groups, income and price elasticities of demand, and consumer behaviour in the more advanced countries. The other method was that of consumption norms. The former method attempted to foresee consumer demand, the

latter to shape it.

An example of the method of norms, and its policy implications, is set out in Table 1.

The Soviet diet

	Norm (kgs/head/year)	Per capita consumption in 1976 as % of norm
Bread and bread products	120	128
Potatoes	97	123
Vegetables and melons	37	59
Vegetable oil and margarine	7	85
Meat and meat products	82	68
Fish and fish products	18	101
Milk and milk products	434	78
Eggs	17	72

Sources: Weitzman (1974); Agababyan and Yakovleva (1979, p. 142).

Table 1 makes clear the logic of the Soviet policy in the Brezhnev era (1964–82) of expanding the livestock sector, and also importing fodder and livestock products. Since the consumption of livestock products was below the norm level, the government sought to make possible an increase in their consumption.

The method of consumption norms was an alternative to the price mechanism for the determination of output. It has also been used, however, in Western countries. It is used there in those cases where distribution on the basis of purchasing power has been replaced by distribution on the basis of need. Examples include the provision of housing, hospitals, schools and parks. Calculations of the desirable number of rooms, hospital beds and school places per person are a familiar tool of planning in welfare states.

There are two main problems with the norm method of consumption planning. The first is that of substitution between products. Although consumers may well have a medically necessary need for x grams of protein per day, they can obtain these proteins from a wide variety of foods. Second, consumers may choose to spend their money ‘irrationally’, for example, to buy spirits instead of children's shoes.

Material balances

A material balance is a balance sheet for a particular commodity showing, on the one hand, the economy's resources and potential output, and, on the other, the economy's need for a particular product. Material (and labour) balances were the main methods used in calculating production and distribution plans for goods, supply plans and labour plans. Soviet planners took great pride in the balance method and considered it one of the greatest achievements of planning theory and practice. Material balances were drawn up for different periods (for example, for annual or five year periods), by different organizations (for example, Gosplan, Gossnab – the body responsible for allocating supplies of inputs – and the ministries) and at different levels (for example, national and republican). The material balances were also drawn up with different degrees of aggregation. Highly aggregated balances were drawn up for the Five Year Plans, and highly disaggregated balances by the chief administrations of Gossnab for annual supply planning. The aim of the material balance method was to ensure the consistency of the plans.

Normally, at the start of the planning work, the anticipated availability of a commodity was not sufficient to meet anticipated requirements. To balance the two, the planners sought possibilities of economizing on scarce products and substituting for scarce materials; they investigated the possibilities of increasing production or importing raw materials or equipment, or in the last resort they determined the priority needs to be fulfilled by the scarce commodity. Even with great efforts, achieving a balance was difficult. The complexity of an economy in which a great variety of goods are produced by different processes, all of which are subject to continuous technological change, was often too great for anything more than a balance that balanced only on paper. Hence it was normal, during the ‘planned’ period, for the plan to be altered, often repeatedly, as imbalances came to light. Particularly important problems with the use of material balances were the highly aggregated nature of the balances and their interrelated nature.

Input – output

A wide variety of input–output tables were regularly constructed in socialist countries. *Ex post* national tables in value terms, planning national tables in value and physical terms, regional tables, and capital stock matrices were widely constructed and used. An interesting and important use concerned variant calculations of the structure of production in medium-term planning.

Because an input–output table can be represented by a simple mathematical model, and because of the assumption of constant coefficients, an input–output table can be utilized for variant calculations.

$$X = (I - A)^{-1} Y$$

On the assumption that *A* is given, *X* can be calculated for varying values of *Y*. Variant calculations of the structure of production were not undertaken with material balances because of their great labour intensity. Variant calculations played a useful role in medium-term planning because they enabled the planners to experiment with a wide range of possibilities. The first major use of variant calculations of the structure of production in Soviet national economic planning was in connection with the 1966–70 Five Year Plan. Gosplan's economic research institute analysed the results of various possible shares of investment in the national income for 1966–70. It became clear that stepping up the share of investment in the national income would increase the rate of growth of the national income, but that this would have very little effect on the rate of growth of consumption (because almost all of the increased output would be producer goods). The results of the calculations are set out in Tables 2 and 3.

Output of steel on various assumptions

	Variants				
	I	II	III	IV	V
Production of steel in 1970 (millions of tonnes)	109	115	121	128	136

Average annual growth rates of selected industries, 1966–1970 (%)

	Variants				
	I	II	III	IV	V
Engineering and metal working	7.1	8.2	9.3	10.4	11.4
Light industry	6.3	6.6	6.8	7.0	7.2
Food industry	7.1	7.3	7.4	7.5	7.6

Source: Ellman (1973, p. 71).

The five variants are for the share of investment in the national income, I being the lowest and V the highest. A sharp increase in the share of investment in the national income in the Five Year Plan 1966–70 would have led to a sharp fall in the share of consumption in the national income, and only a small increase in the rate of growth of consumption (within a Five Year Plan period). What is very sensitive to the share of investment in the national income is the output of the producer goods industries, as Tables 2 and 3 show.

These results are along the lines of what one would expect on the basis of Fel'dman's model, but the input–output technique improves on Fel'dman's model since it enables the effect of different strategies to be seen at the industry level rather than merely in terms of macroeconomic aggregates.

Another example of the use of input–output for economic calculations concerns the statistical data about the relations between industries contained in the national *ex post* tables in value terms. In his controversial 1968 book *Mezhotraslevye svyazi sel'skogo khozyaistva*, M. Lemeshev, then deputy head of the sector for forecasting the development of agriculture of the USSR Gosplan's Economic Research Institute, used the Soviet input–output table for 1959 as the basis for a powerful plea for more industrial inputs to be made available to agriculture.

He began by observing that from the 1959 input–output table it was clear that of the current material inputs into agriculture in that year only 23.4 per cent came from industry, while 54.7 per cent came from agriculture itself (feed, seed and so on). He argued that this was most unsatisfactory. In the section on the relationship between agriculture and engineering Lemeshev argued that the supply to agriculture of agricultural machinery was inadequate, in the section on the relationship between agriculture and the chemical industry he argued that the supply of fertilizers was inadequate, and in the section on agriculture and electricity he argued that the supply of electricity to the villages for both productive and unproductive needs was inadequate. In addition, in the section on the relationship between agriculture and the processing industry he argued that the latter was not helping agriculture as it should do; for example, it was sometimes impossible to accept vegetables (although the consumption of these in the towns was below the norms) because of inadequate processing and distribution facilities. Furthermore, he argued that the supply of concentrated feed was inadequate and the processing of milk wasteful. In view of the inadequate development of the food processing industry, he argued for the development of processing enterprises by the farms themselves.

The chapter on the productive relations between agriculture and the building industry was an extensive critique of the practice of productive, and of housing and communal, building in the villages. Lemeshev argued that the state should take on responsibility for building on the collective farms. The chapter on the relationship between agriculture and transport was critical of the shortage of river freight boats. The chapter on investment argued that investment in agriculture was inadequate, and that in the period 1959–65 there was an unwarranted increase in the proportion of investment in the collective farms which they had to finance themselves. He also argued that a greater proportion of agricultural investment should be financed by bank loans, and that as a criterion of investment efficiency the recoupment period was satisfactory. The concluding chapter was concerned with improving the productive relations between agriculture and the rest of the economy. The author argued for improving central planning by the use of input–output, for replacing procurement plans by free contracts between farms and the procurement organs (if a shortage of a particular product threatened then its price could be raised), and for the elimination of the supply system (that is, the rationing of producer goods) which hindered farms from receiving the goods they wanted and sometimes supplied them with goods that they did not want. Lemeshev also argued for higher pay in agriculture and for the reorganization of the labour process within state and collective farms on the basis of small groups which were paid by results.
This book was a good example of the use of input–output to provide statistical data which could be used, alongside other information, to provide a description of important economic relations and to support a case for important institutional and policy changes.

Project evaluation

In the USSR of the 1930s, it was officially considered that there was no problem of project evaluation to which economists could contribute. The sectoral allocation of investment was a matter for the central political leadership to decide. It was they who decided in which sectors and at which locations production should be expanded. These decisions were based on the experience of the more advanced countries, the traditions of the Russian state (for example, stress on railway building) and of the Bolshevik movement (for example, stress on electrification and on the metal-using industries) and on the needs of defence. As far as decisions within sectors were concerned, here the main idea was to fulfil the plan by using the world's most advanced technology.
The practical study of methods for choosing between variants within sectors was begun by engineers in the electricity and railway industries. The problem analysed was that of comparing the cost of alternative ways of meeting particular plan targets. A classic example of the type of problem considered was the choice between producing electricity by a hydro station and by a thermal station.
During Stalin's lifetime, the elaboration by orthodox economists and the adoption by the planners of economic criteria for project evaluation were impossible because they were outside Stalin's conception of the proper role of economists (apologetics). When economists did make a contribution in this area, as was done by Novozhilov, it was ignored. After Stalin's death, however, it became possible for Soviet economists to contribute to the elaboration of methods of economic calculation for use in the decision-making process. An early and important example was in the field of project evaluation. An official method for project evaluation was adopted in 1960, and revised versions in 1964, 1966, 1969 and 1981. In a very abbreviated and summary form, the 1981 version was as follows.
In evaluating investment projects, a wide variety of factors have to be taken into account, for example, the effect of the investment on labour productivity, capital productivity, consumption of current material inputs (such as metals and fuel), costs of production, environmental effects, technical progress, the location of economic activity and so on. Two indices which give useful synthetic information about economic efficiency (but are not necessarily decisive in choosing between investment projects) are the coefficient of absolute economic effectiveness and the coefficient of relative economic effectiveness.
At the national level, the coefficient of absolute effectiveness is defined as the incremental output–capital ratio.

$$E_p = \frac{\Delta Y}{I}$$

where E_p is the coefficient of absolute effectiveness for a particular project, ΔY is the increase in national income generated by the project, and I is the investment cost. The value of E_p calculated in this way for a particular investment has to be compared with E_a, the normative coefficient of absolute effectiveness, which is fixed for each Five Year Plan and varies between sectors. In the 11th Five Year Plan (1981–85) it was 0.16 in industry, 0.07 in agriculture, 0.05 in transport and communications, 0.22 in construction and 0.25 in trade.

$$\text{If } E_p > E_a$$

then the project is considered efficient.
For calculating the criterion of absolute effectiveness at the level of individual industries, net output is used in the numerator instead of national income. At the level of individual enterprises and associations, in particular when a firm's own money or bank loans are the source of finance, profit is used instead of national income.
The coefficient of relative effectiveness is used in the comparison of alternative ways of producing particular products. In the two products case

$$E = \frac{C_1 - C_2}{K_2 - K_1}$$

where E is the coefficient of relative effectiveness, C_i is the current cost of the ith variant, and K_i is the capital cost of the ith variant.
If E>En, where En is the officially established normative coefficient of relative economic efficiency, then the more capital intensive variant is economically justified. In the 11th Five Year Plan, En was in general 0.12, but exceptions were officially permitted in the range 0.08/0.10–0.20/0.25.
In the more than two variants case, they should be compared according to the formula

$$C_i + E_n K_i \rightarrow \text{minimum}$$

that is, choose that variant which minimizes the sum of current and capital costs.
At one time a rationalist misinterpretation of socialist planning was widespread. According to this view, a planned economy was one in which rational decisions were made after a dispassionate analysis by omniscient and all-powerful planners of all the alternative possibilities. In such a system, the adoption of rational criteria for project evaluation would have been of enormous importance. Socialist planning, however, was just one part of the social relations between individuals and groups in the course of which decisions were taken, all of which were imperfect and many of which produced results quite at variance with the intentions of the top economic and political leadership.
A good example of the factors actually influencing investment decisions under state socialism was the commencement of the construction of the Baoshan steel plant near Shanghai. The site was apparently chosen because of the political influence of a high-ranking Shanghai party official. The location decision ignored the fact that, because of the swampy nature of the site, necessitating large expenditures on the foundations, this was in fact the most expensive of the sites considered. Very expensive, dogged with cost overruns, involving major pollution problems, the whole project was kept alive for some time by a powerful steel lobby. In due course, as a result of a national policy reversal in Beijing, the second phase was deferred and those involved publicly criticized. To judge from its initial costs of production, it produced gold rather than steel.
In general, the choice of projects owed more to inter-organization bargaining in an environment characterized by investment hunger than it did to the detached choice of a cost-minimizing variant. The development of new and better criteria for project evaluation turned out to be no guarantee that project evaluation would improve since the criteria were often not in fact used to evaluate projects. Their main function was to provide an acceptable common language in which various bureaucratic agencies conducted their struggles. Agencies adopted projects on normal bureaucratic grounds and then tried to get them adopted by higher agencies, or defended them against attack, by presenting efficiency calculations using the official methodology but relying on carefully selected data.

Linear programming and extensions

Linear programming was discovered by the Soviet mathematician Kantorovich in the late 1930s. Its relevance for Soviet planning was widely discussed in the USSR in the 1960s and extensive efforts were made actually to use it in Soviet planning in the 1970s. Three examples of its use follow.

Production scheduling in the steel industry

Linear programming was discovered by Kantorovich in the course of solving the problem, presented to him by the Laboratory of the all-Union Plywood Trust, of allocating productive tasks between machines in such a way as to maximize output given the assortment plan. From a mathematical point of view, the problem of optimal production scheduling

for tube mills and rolling mills in the steel industry, which was tackled by Kantorovich in the 1960s, is very similar to the Plywood Trust problem, the difference being its huge dimensions.

The problem arose in the following way. As part of the planning of supply, Soyuzglavmetal (the department of Gossnab concerned with the metal industries), after the quotas had been specified, had to work out production schedules and attachment plans in such a way that all the orders were satisfied and none of the producers received an impossible plan. In the 1960s an extensive research programme was initiated by the department of mathematical economics (which was headed by Academician Kantorovich) of the Institute of Mathematics of the Siberian branch of the Academy of Sciences, to apply optimizing methods to this problem. The chief difficulties were the huge dimensions of the problem and the lack of the necessary data. About 1,000,000 orders, involving 60,000 users, more than 500 producers and tens of thousands of products, were issued each year for rolled metal. Formulated as a linear programming problem it had more than a million unknowns and 30,000 constraints. Collecting the necessary data took about six years. Optimal production scheduling was first applied to the tube mills producing tubes for gas pipelines (these were a scarce commodity in the USSR). In 1970 this made possible an output of tubes 108,000 tons greater than it would otherwise have been, and a substantial reduction in transport costs was also achieved.

The introduction of optimal production scheduling into the work of Soyuzglavmetal was only part of the work initiated in the late 1960s on creating a management information and control system in the steel industry. This was intended to be an integrated computer system which would embrace the determination of requirements, production scheduling, stock control, the distribution of output and accounting. Such systems were widely introduced in Western steel firms in the late 1960s. Work on the introduction of management information and control systems in the Soviet economy was widespread in the 1970s, but by the 1980s there was widespread scepticism in the USSR about their usefulness. This largely resulted from the failure to fulfil the earlier exaggerated hopes about the returns to be obtained from their introduction in the economy.

Industry investment plans

In the state socialist countries investment plans were worked out for the country as a whole, and also for industries, ministries, departments, associations, enterprises, republics, economic regions and cities. An important level of investment planning was the industry. Industry investment planning is concerned with such problems as the choice of products, of plants to be expanded, location of new plants, technology to be used, and sources of raw materials.

The main method used in the 1970s and 1980s in the Council for Mutual Economic Assistance (CMEA, known in the West as Comecon) countries for processing the data relating to possible investment plans into actual investment plans was mathematical programming. After extensive experience in this field, in 1977 a Standard Methodology for doing such calculations was adopted by the Presidium of the USSR Academy of Sciences.

The Soviet Standard Methodology presented models for three standard problems. They were: a static multi-product production problem with discrete variables, a multi-product dynamic production problem with discrete variables, and a multi-product static problem of the production-transport type with discrete variables. The former can be set out as follows: Let $i = 1, \ldots, n$ be the finished goods or resources, $j = 1, \ldots, m$ be the production units, $r = 1, \ldots, R_j$ be the production technique in a unit, a_{ij}^r be the output of good $I = 1, \ldots, n'$ or input of resource $i = n' + 1, \ldots, n$, using technique *r* of production in unit *j*; C_j^r are the costs of production using technique *r* in unit *j*; D_i is the given level of output of good *i*, $i = 1, \ldots, n'$; P_i is the total use of resource *i*, $i = n' + 1, \ldots, n$ allocated to the industry; Z_j^r is the unknown intensity of use of technique *r* at unit *j*.

The problem is to find values of the variables Z_j^r that minimize the objective function

$$\sum_{j=1}^{m} \sum_{j=1}^{R_j} C_j^r Z_j^r \tag{1}$$

that is, minimize costs of production subject to

$$\sum_{j=1}^{m} \sum_{r=1}^{R_j} a_{ij}^r Z_j^r \geq D_i, \; i = 1, \ldots, n'$$

(2)

that is, each output must be produced in at least the required quantities

$$\sum_{j=1}^{m}\sum_{r=1}^{R_j} a_{ij}^r Z_j^r \leq P_i,\ i = n' + 1, \ldots, n$$

(3)

that is, the total use of resources cannot exceed the level allocated to the branch

$$\sum_{r=1}^{R_j} Z_j^r \leq 1,\ j = 1, \ldots, m$$

(4)

$$Z_j^r = 0 \text{ or } 1,\ j = 1, \ldots, m,\ r = 1, \ldots, R_j$$

(5)

that is, either a single technique of production for unit *j* is included in the plan or unit *j* is not included in the plan.
In order to illustrate the method, an example will be given which is taken from the Hungarian experience of the 1950s in working out an investment plan for the cotton weaving industry for the 1961–65 Five Year Plan. The method of working out the plan can be presented schematically by looking at the decision problems, the constraints, the objective function and the results.
The decision problems to be resolved were:

1. (a) How should the output of fabrics be increased, by modernizing the existing weaving mills or by building new ones?
2. (b) For part of the existing machinery, there were three possibilities. It could be operated in its existing form, modernized by way of alterations or supplementary investments, or else scrapped. Which should be chosen?
3. (c) For the other part of the existing machinery, it could be either retained or scrapped. What should be done?
4. (d) If new machines are purchased, a choice has to be made between many types. Which types should be chosen, and how many of a particular type should be purchased?

The constraints consisted of the output plan for cloth, the investment fund, the hard currency quota, the building quota and the material balances for various kinds of yarn. The objective function was to meet the given plan at minimum cost.
The results provided answers to all the decision problems. An important feature of the results was the conclusion that it was cheaper to increase production by modernizing and

expanding existing mills than by building new ones.

It would clearly be unsatisfactory to optimize the investment plan of each industry taken in isolation. If the calculations show that it is possible to reduce the inputs into a particular industry below those originally envisaged, then it is desirable to reduce planned outputs in other industries, or increase the planned output of the industry in question, or adopt some combination of these strategies. Accordingly, the experiments in working out optimal industry investment plans, begun in Hungary in the 1950s, led to the construction of multi-level plans linking the optimal plans of the separate industries to each other and to the macroeconomic plan variables. Multi-level planning of this type was first developed in Hungary, but subsequently spread to the other CMEA countries. Extensive work on the multi-level optimization of investment planning was undertaken in the USSR in connection with the 1976–90 long-term plan. (The 1976–90 plan, like all previous Soviet attempts to compile a long-term plan, was soon overtaken by events. The plan itself seems never to have been finished and was replaced by ten-year guidelines for 1981–90.)

The determination of costs in the resource sector

In view of the wide dispersion of production costs in the resource sector, the use of average costs (and of prices based on average costs) in allocation decisions is likely to lead to serious waste. An important outcome of the work of Kantorovich and his school for practical policy was (after a long lag) official acceptance of this proposition and of linear programming as a way of calculating the relevant marginal costs. For example, in 1979 in the USSR the State Committee for Science and Technology and the State Committee for Prices jointly approved an official method for the economic evaluation of raw material deposits. This was a prescribed method for the economic evaluation of exploration and development of raw material deposits. What was new in principle about this document was that it permitted the output derived from the deposits to be evaluated either in actual (or forecast) wholesale prices or in marginal costs. For the fuel-energy sector, a lot of work was done to calculate actual (and forecast) marginal costs for each fuel at different locations throughout the country and for different periods. These figures were regularly calculated on optimizing models (they were the dual variables to the output maximizing primal) and were widely used in planning practice for many years.

Comparison with the West

An important method of economic calculation in socialist countries was comparison with the West. If a particular product or method of production had already been introduced (or phased out) in the West, this was generally considered a good argument to introduce it (or phase it out) in the socialist countries, subject to national priorities and economic feasibility. Obtaining advanced technology from abroad (by purchase, Lend-Lease, reparations, espionage, direct investment) was an integral part of socialist planning, the importance of the different elements varying over time. Comparisons with the West were particularly important in an economic system which lagged behind the leading countries, lacked institutions which automatically introduced innovations into production (that is, profit-seeking business firms), and found it difficult (because of the ignorance of the planners, stable cost-plus prices and the self-interest of rival bureaucratic agencies) to notice, appraise realistically when noticed, and adopt, innovations.

Economic calculation and economic results

It is important not to exaggerate the influence of methods of economic calculation on the performance of an economy. The performance of an economy is largely determined by external factors (such as the world market), economic policy (for example, the decision to import foreign capital or to declare a moratorium), economic institutions (like collective farms) and the behaviour of the actors within the system (for example, underestimation of investment costs by initiators of investment projects). It is entirely possible for an improvement in the methods of economic calculation to coincide with a worsening of economic performance (as happened in the USSR in the Brezhnev period). Realization of these facts led in the 1970s to a shift from the traditional normative approach (which concentrates on the methods of economic calculation and which regards their improvement as the main key to improved economic performance and the main role of the economist) in the study of planned economies, to the systems and behavioural approaches.

Economic calculation and economic intuition

In view of bounded rationality, and the huge volume, and distorted nature, of the information available to the central leadership, really existing decision-making relied heavily on rules of thumb and the ‘feel’ for reality of the top decision-makers (sometimes known as ‘planning by feel’). This could quickly lead to an equilibrium, but an inefficient one.

See Also

- behavioural public economics
- Kantorovich, Leonid Vitalievich
- Leontief, Wassily
- Soviet Union, economics in
- Stalinism, political economy of

Bibliography

Agababyan, E. and Yakovleva, Ye., eds. 1979. *Problemy raspredeleniya i rost narodnogo blagosostoyaniya*. Moscow: Nauka.

Birman, I. 1978. From the achieved level. *Soviet Studies* 30(2), 153–72.

Birman, I. 1996. Otraslevoe optimal'noe. Ch. 9 of *Ya –ekonomist*. Novosibirsk: EKOR.

Boltho, A. 1971. *Foreign Trade Criteria in Socialist Economies*. Cambridge: Cambridge University Press.

Ellman, M. 1973. *Planning Problems in the USSR: The Contribution of Mathematical Economics to their Solution 1960–1971*. Cambridge: Cambridge University Press.

Ellman, M. 1983. Changing views on central planning: 1958–1983. *ACES Bulletin* [now *Comparative Economic Studies*] 25(1), 11–34.

Gács, J. and Lackó, M. 1973. A study of planning behaviour on the national-economic level. *Economics of Planning* 13, 91–119.

Giffen, J. 1981. The allocation of investment in the Soviet Union. *Soviet Studies* 33(4), 593–609.

Granick, D. 1990. Planning as coordination. Ch. 3 of *Chinese State Enterprises*. Chicago: University of Chicago Press.

Gregory, P. and Harrison, M. 2005. Allocation under dictatorship: research in Stalin's archives. *Journal of Economic Literature* 43, 721–61.

Kornai, J. 1967. *Mathematical Planning of Structural Decisions*. Amsterdam: North-Holland.

Kornai, J. 1980. *Economics of Shortage*, 2 vols. Amsterdam: North-Holland.

Kornai, J. 1992. Planning and direct bureaucratic control. Ch. 7 of *The Socialist System*. Oxford: Oxford University Press.

Kueh, Y. 1985. *Economic Planning and Local Mobilization in Post-Mao China*. London: Contemporary China Institute.

Kushnirsky, F. 1982. *Soviet Economic Planning, 1965–1980*. Boulder, CO: Westview, ch. 4.

Lazarev, V. and Gregory, P. 2003. Commissars and cars: a case study in the political economy of dictatorship. *Journal of Comparative Economics* 31, 1–19.

Lemeshev, M. 1968. *Mezhotraslevye svyazi sel'skogo khozyaistva*. Moscow: Ekonomika.

Levine, H. 1959. The centralized planning of supply in Soviet industry. In *Comparisons of the United States and Soviet Economies*. Washington, DC: Joint Economic Committee, US Congress.

Malinovskii, B. 1995. *Istoriya vychislitel'noi tekhniki v litsakh*. Kyiv: KIT/A.S.K.

Matekon. 1978. Standard methodology for calculations to optimize the development and location of production in the long run. Vol. 15(1), 75–96.

Qian, Y., Roland, G. and Xu, C. 2000. Coordinating activities under alternative organizational forms. In *Planning, Shortage, and Transformation*, ed. E. Maskin and A. Simonovits. Cambridge, MA: MIT Press.

Shlykov, V. 2004. The economics of defense in Russia and the legacy of structural militarization. In *The Russian military*, ed. S. Miller and D. Trenin. Cambridge, MA: MIT Press.

Stalin, J. 1952. Concerning the errors of comrade L.D. Yaroshenko. In *Economic Problems of Socialism in the USSR*. Moscow: Foreign Languages Publishing House.

Tretyakova, A. and Birman, I. 1976. Input–output analysis in the USSR. *Soviet Studies* 28(2), 157–86.

Weitzman, P. 1974. Soviet long term consumption planning: distribution according to rational need. *Soviet Studies* 26(3), 305–21.

World Bank. 1992. *China: Reform and the Role of the Plan in the 1990s*. Washington, DC: World Bank.

How to cite this article

Ellman, Michael. "economic calculation in socialist countries." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000208> doi:10.1057/9780230226203.0427

The New Palgrave Dictionary of Economics Online

economic demography

Allen C. Kelley and Robert M. Schmidt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic demography is an area of study that examines the determinants and consequences of demographic change, including fertility, mortality, marriage, divorce, location (urbanization, migration, density), age, gender, ethnicity, population size and population growth. This article reviews and critically evaluates important macroeconomic dimensions of the ‘population debates’ between the ‘optimists’ and the ‘pessimists’ since 1950. It concludes with an examination of demography in the popular ‘convergence’ growth models of the 1990s.

Keywords

adult equivalency; aging; agricultural growth and population change; capital accumulation; convergence; demographic drag; demographic gift; demographic transition; diffusion of technology; diminishing returns; dismal science; economic demography; economic development; economic growth; education; endogenous growth; fertility; free rider problem; human capital; innovation; Kuznets, S.; labour productivity; learning-by-doing; life expectancy; life-cycle modelling; Malthus., T. R.; mortality; population density; population growth; population size; renewable resources; research and development; rule of law; saving; Simon, J. L.; subsistence; technical change

Article

Economic demography is an area of study that examines the determinants and consequences of demographic change, including fertility, mortality, marriage, divorce, location (urbanization, migration, density), age, gender, ethnicity, population size, and population growth. An applied area of research, economic demography draws upon the theoretical and applied fields of economics. For example, the determinants of fertility or migration primarily draw upon microeconomic theory and labour economics, while the consequences of population growth or ageing primarily draw upon macroeconomic theory and development economics.

The field has had a long tradition of controversy, beginning with the publication in 1798 of *An Essay on the Principle of Population* by the Reverend Thomas Malthus. The basic Malthusian model is founded on two propositions: (*a*) population, *when unchecked*, increases at a geometric rate (for example, 1, 2, 4, 8…) and (*b*) food, in contrast, expands at an arithmetic rate (for example, 1, 2, 3, 4…). The result is a population trapped at a meagre standard of living. Short of 'preventive checks' (birth control), population is constrained to live at subsistence by 'positive checks' (deaths, war, famines and pestilence). In later writings Malthus admitted the possibility of 'moral restraint' that could deter births, primarily through the postponement of marriage. However, he held little hope for a notable attenuation of the 'natural passions' of the working class.

While much of the controversy relating to Malthusianism has focused on the determinants of population growth, a second premise of his model relates to its economic underpinnings: the determinants of agricultural growth. Here Malthus appealed to the historical law of diminishing returns in agriculture. While this proposition engendered relatively little dispute at the time, history has since documented widespread and sometimes notable improvements in agricultural technology. Indeed, food production has represented an engine of growth in many of the areas that Malthus investigated. In some areas today, governments worry about 'excess' food production that depresses prices and farmers' living standards. Unfortunately, the pessimistic food-production predictions, when confronted by rapid population growth, caused economics to be dubbed the 'dismal science'.

The enormous popularity of the Malthusian ideas was the result of several factors: the model's simplicity and its explanation of poverty (the poor failed to exercise moral restraint, ending up with large families); the appeal of the message that subsidizing the poor is of questionable efficacy; and the plausibility of the Malthusian argument given the unexpected 'population explosion' revealed by the 1801 census. These and other elements of the 'Malthusian debate' provide a useful taxonomy for organizing the present article.

Specifically, we highlight the macroeconomic dimensions of the economic consequences of population growth since 1950. As with the early Malthusian debates, an assessment of the macroeconomic impacts of demographic change on economic production has resulted in an outpouring of research, which has spawned further debate. There are periods when vigorous Malthusian-like alarmism has carried the day; there are periods of counter-challenges; and, since the mid-1980s, there has been a productive 'revisionist' movement. In short, the simplistic Malthusian notion of diminishing returns in production has given way to more informed modelling of economic–demographic interactions. An assessment of the historical evolution of this literature will constitute the bulk of this review and appropriately delimits the scope of our essay since a wide range of important microeconomic themes are taken up in other articles in this dictionary (see fertility in developing countries, family decision making, marriage and divorce, retirement, and multiple articles dealing with the topics of gender, ageing and mortality).

We begin by examining population impacts in one-sector growth models. This leads nicely into a more detailed assessment of factor accumulation, and in particular, the impacts of demography on saving, investment and technological change. This is in turn followed by an analytical description of the evolution of economic–demographic thinking since 1950. Such a perspective exposes many of the key analytical and empirical linkages of interest. The article concludes with an examination of 'convergence modelling', a useful paradigm that exposes the roles of changing demographic structures that take place over the demographic transition.

1 Theory: modelling economic – demographic change

1.1 One-sector growth models

The aggregate production function constitutes the primary organizing device for delineating the impacts of demographic change on economic growth. Within this model, labour productivity depends on the availability of complementary factors of production (land, natural resources, human and physical capital) and technology. If we assume, for convenience, that labour is a constant fraction of population, then population size directly affects aggregate output.

In a production function with constant returns to scale, an increase in population growth will lower the average availability of other factors of production – a ‘resource-shallowing’ effect, and, through diminishing returns, reduce the growth of worker productivity. Such an adverse demographic impact can be magnified (or attenuated) if population growth diminishes (raises) the growth rate of complementary factors.

In a standard growth model with factor inputs of labour and capital, and a saving rate and pace of technological change that are exogenous with respect to population growth, demography affects the long-run *level* but not the long-run *growth rate* of output per capita. This is because the capital-shallowing effect of increased population will eventually reduce the capital per worker ratio to a level sufficient to be maintained by a fixed rate of saving. In this case, long-run growth is determined by the pace of technological change. The determinants of the ‘fixed’ saving rate and pace of technology growth, both considered in more detail below, are central to the analysis.

If one relaxes some of the assumptions of this model, the impact of population growth on per capita output growth can be ambiguous. Negative impacts can arise through diminishing returns, diseconomies of scale, and perhaps savings, while positive impacts can arise through induced technological change, economies of scale, and possibly savings. Most economists believe that adverse capital-shallowing impacts will dominate positive feedback effects, although the magnitude of the demographic impacts may not be all that large.

1.2 Saving

Possibly the most investigated linkage of population growth to economic growth has been the impact of demographic change on saving. Two perspectives dominate.

Adult equivalency. Rapid (slow) rates of population growth result in a disproportionate number of children (elderly adults) who consume, but contribute relatively little to, household income. In recognizing that these ‘dependents’ consume less than a working-age adult, the notion of an ‘adult equivalent’ consumer was born. The financing of an additional child's ‘adult-equivalent’ consumption has been hypothesized to be out of saving. Such a view, however, has been challenged by consideration of several offsetting alternatives. Specifically, children may (*a*) substitute for other forms of consumption, (*b*) contribute directly to household market and non-market income, (*c*) encourage parents to work more (or less), (*d*) stimulate the amassing (or reduction) of estates, and (*e*) encourage (or discourage) the accumulation of certain types of assets (for example, education or farm implements). The net impact of changing dependency rates on saving is therefore theoretically ambiguous. This is particularly the case if one views human capital as an investment financed in part by households and

governments. At any rate, empirical evidence showing negative impacts of youth dependency on saving are found in several studies.

The life-cycle. A second population-saving linkage is based on a life-cycle formulation incorporated into a lifetime household utility function. Specifically, households attempt to even out their lifetime consumption by setting aside earnings during working years to finance consumption by their children as well as for their own retirement. This formulation can yield positive or negative impacts on aggregate saving depending on the relative sizes of the dissaving youth and elderly cohorts. While empirical evidence from life-cycle modelling is mixed, those studies do tend to show linkages between age structure and saving. However, the direction and magnitude of that impact depends upon time and place. (See, for example, Mason, 1987; Higgins, 1998; and Lee, Mason and Miller, 2001.)

1.3 Population-sensitive government spending

Government spending on population-sensitive activities such as schooling (youth) and health (elderly) has been alleged both to reduce saving and to crowd out spending on relatively growth-oriented investments. These two hypotheses constitute the core of Ansley J. Coale and Edgar M. Hoover's (1958) path-breaking study of India. While these premises are appealing, they require qualification. Governments have many options to accommodate population pressures. Indeed, limited empirical evidence (for example, Schultz, 1987) has shown that education financing can be met all or in part by (*a*) trade-offs within the public sector, (*b*) reductions in per pupil expenditures, and (*c*) efficiency gains. While the second approach can be expected to reduce the quality of education (and therefore future productivity), the importance of population pressures on government spending or educational quality is uncertain.

1.4 Technological change: density, size and endogenous growth

While development economists have for decades harkened the pace of technological change as a (the?) major source of economic growth, most standard growth theory models take the rate of technological change as exogenous. With technological change independent of demographic change, population growth per se will have no impact on the pace of economic growth in long-run equilibrium. By contrast, if technological change is all or in part *embodied* in new investment, then a vintage specification is appropriate whereby new capital is relatively more productive than old. In this set-up, population growth can be economic-growth enhancing by expanding the rate at which technology is incorporated into production. In yet another specification, population growth can directly affect the rate of technological change and/or its form (factor bias). Kenneth J. Arrow (1962) has hypothesized that learning by doing is quickened in an environment of rapid employment growth.

A fourth linkage between technology and demography is found in ‘endogenous growth’ models that relate the pace of technology directly to population *size*. In particular, the benefits of R&D are assumed to be available to all firms without cost; that is, an R&D industry generates a non-rival stock of knowledge. As a result, if we hold constant the share of resources used for research, an increase in population size advances technological change without limit. This somewhat controversial prediction has been qualified by models that incorporate various firm- or industry-specific constraints on R&D

production. Such models typically reduce, but do not eliminate, the positive impacts of population size which, as in the embodiment models above, are manifested largely during the ‘transition’ to long-run equilibrium.

Evidence on the roles of demographic-technology linkages and growth has been fragmentary and sparse. A pioneering study by Hollis Chenery and Moises Syrquin (1975) draws upon the experience of 101 countries across the income spectrum over the period 1950–70. They find that the structure of development reveals strong and pervasive scale effects (measured by population size) that vary by stage of development. Basically, small countries develop a modern productive industrial structure more slowly and later, while large countries have higher levels of accumulation and (presumably) higher rates of technological change. Although these roles for demography may have been important historically, the impacts plausibly have waned somewhat: (*a*) economies in infrastructure are judged to be substantially exhausted in cities of moderate size; (*b*) specialization through international trade provides a means of garnering some or many of the benefits of size; and (*c*) scale effects are most prevalent in industries with relatively high capital–labour ratios and such industries are inappropriate to the factor proportions of developing countries.

It is in agriculture where the positive benefits of population size have been most discussed. Higher population densities can lower per unit costs and increase the efficiency of transport, irrigation, extension services, markets and communications (Glover and Simon, 1975). Possibly the most cited work is that by Ester Boserup (1965; 1981), who observes that increasingly productive agricultural technologies are made economically attractive in response to higher land densities. While this is probably true, the issue becomes one of identifying the quantitative magnitude of such effects over varying population sizes and in differing institutional settings. One must be cautious in attributing causation. For example, while high population densities may have accounted for a portion of expanded agricultural output in recent decades, in several important Asian countries these densities were sufficiently high decades ago to justify the investments associated with the new technologies. Boserup in more recent writing has been less sanguine about the benefits of population size because densities appropriate to modern technologies in Asia are three to four times the average for Africa and Latin America.

In short, a wide-ranging review of the literature does not provide a strong consensus on the quantitative linkages between the size and growth of population, on the one hand, and the pace of technological change and economic growth, on the other hand.

1.5 The bottom line

An evaluation of population growth on economic growth through the filter of formal economic-growth modelling yields limited results: population growth affects the level but not the growth of per capita income in long-run equilibrium. Moreover, the key determinants of long-run growth are saving and technology. Only if these factors depend on demographic change does population matter. This somewhat constraining limitation of growth theory has caused researchers to branch out and explore a host of economic–demographic interactions using less formal paradigms. This blossoming literature has been extensive, lively and sometimes contentious.

2 Evolution of population-impacts thinking: 1950 – 90

Four major studies, two by the United Nations (1953; 1973) and two by the National Academy of Sciences (1971; 1986), reveal well the evolution of thinking on population matters over the period 1950–90. Three individual scholars, Coale and Hoover and Simon, also played prominent and important roles. (This section draws on Kelley, 2001.)

2.1 United Nations, 1953

The 1953 United Nations report, *Determinants and Consequences of Population Trends*, easily represents the most important contribution to population thinking since the writings of Malthus. Unlike Malthus, however, the UN study was balanced and exhaustive both in detail and in coverage. Some 21 linkages between population and the economy were taken up. For example, the impacts of population on the economy can be: (*a*) positive due to economies of scale and organization; (*b*) negative due to diminishing returns; or (*c*) neutral due to technology and social progress. An evaluation of these and other linkages led to a mildly negative overall assessment that was both cautious and qualified.

The most notable feature of this report was its methodology. More than any major study on population to that time, the UN Report embraced a methodology that would ultimately represent elements of modern-day ‘revisionism’. Specifically, the report (*a*) downgraded the importance of population growth's impact on economic growth by placing it on a par with several other determinants of equal or greater impact; (*b*) assessed the consequences of population over a long period of time; and (*c*) emphasized the importance of feedbacks within and between the economic and political systems.

2.2 Coale and Hoover, 1958

The next major contribution to the population-impacts literature was provided by Ansley J. Coale and Edgar M. Hoover in their 1958 book *Population Growth and Economic Development in Low-Income Countries*. Based on simulations of a mathematical model calibrated with Indian data, they concluded that India's development would be enhanced by lower population growth. This was due to the hypothesized adverse impacts of population on household saving. It was also proffered that ‘unproductive’ investments in human capital (such as health and education) would partially displace investments in ‘relatively productive’ forms (such as machines and factories). Economic growth would diminish in response.

Empirically, the above hypotheses have not been convincingly established. While several studies have exposed negative dependency-rate impacts on saving, there are others that show little or no impact. Overall, the findings are mixed, with a tilt toward supporting the Coale and Hoover formulation. (See Section 1.2 above for a discussion of the trade-offs that households can make to maintain saving in response to expanding family size.)

Similarly, there are alternative ways for governments to organize and finance schooling in response to population pressures. Unfortunately, studies of this are limited, although one by T. Paul Schultz (1987) finds no support for the Coale and Hoover (1958) formulation.

2.3 National Academy of Sciences, 1971

Arguably the most pessimistic assessment of the consequences of population growth was a study compiled by the National Academy of Sciences (NAS). The panel's final submission, *Rapid Population Growth: Consequences and Policy Implications*, issued in 1971, appeared in two volumes: Volume 1, *Summary and Recommendations*, and Volume 2, *Research Papers*. Unfortunately, the *Summary* volume appeared to be more political than academic in goal and orientation, and was not faithful to many of the underlying research reports assembled by the panel. Indeed, the *Summary* volume highlighted some 25 alleged negative consequences of population growth, whereas it downplayed or eliminated impacts that could be considered as ‘neutral’ or ‘favourable’. As a result, the *Summary* represents an upper bound on the negative consequences of population growth. (A detailed documentation exposing the somewhat controversial way in which the *Summary* was compiled is provided by Kelley, 2001.)

What can be learned from the NAS study? First, given its apparent bias and the lack of a systematic vetting of Volume 1 by members of the panel, it is difficult to use that volume, either in full or in part. However, the individual papers are available and they, in total, offer a more balanced treatment. Second, by its own acknowledgment, the study focused on the short run when negative impacts of population change are most likely to prevail. (‘We have limited ourselves to relatively short term issues’; 1971, p. vi.) By contrast, ‘direct’ (short-run) impacts of demographic change are almost always attenuated (and sometimes offset) by ‘indirect feedbacks’ that occur over longer periods of time. Thus the decision by the NAS panel to focus only on the short-run direct impacts resulted in an overly negative assessment of the consequences of population growth.

Third, economists were underrepresented on both the panel and in providing background reports. This is relevant since economists have substantial faith in the capacity of markets, individuals and institutions to adjust in the face of population pressures. Such adjustments, of course, take time and they are not without cost. Finally, this NAS Report provides a striking example of the difficulty of maintaining objectivity when social science research enters the public policy domain.

2.4 United Nations, 1973

In 1973 the United Nations weighed in with an update of its previous seminal work (United Nations, 1953). In contrast to the broadly eclectic stance in the earlier report, the new one ended with a mild to moderate negative overall assessment of rapid population growth. The authors were concerned with the ability of agriculture to feed expanding populations (à la Malthus) and the difficulty of offsetting capital shallowing (à la Coale and Hoover). Still, the 1973 Report, whose conclusions are highly qualified, is not alarmist, nor is it all that pessimistic. The reason for this moderate stance was the exceptionally influential empirical finding of Simon Kuznets (1960, pp. 19–20, 63) that notable negative correlations between population growth and per capita output growth were largely absent in the data. Given the strong priors of some contributors to the UN study, a failure to find a negative association in the aggregate data by a scholar with impeccable credentials had a profound impact. Indeed, this singular finding arguably kept the population debate alive for yet another round of assessments in the 1980s.

2.5 Revisionism, 1980s and beyond

The 1980s represented a decade when many of the underlying assumptions and conclusions of earlier studies of population–development interactions were subjected to critical scrutiny. The result was a revisionist rendering that was both surprising and controversial. Specifically, the revisionists downgraded the prominence of population growth as either a major source of, or a constraint on, economic prosperity in the Third World. The basis of this somewhat startling conclusion was the revisionists' methodology that (*a*) assessed the consequences of demographic change over longer periods of time and (*b*) expanded the analysis to take into account *indirect feedbacks* within economic and political systems. In general, empirical assessments of population growth will be smaller (less negative or less positive) when using the revisionist's methodology than when focusing on the short run and ignoring feedbacks. On net, most revisionists conclude that many, if not most, Third World countries would benefit from slower population growth.

2.6 Julian L. Simon, 1981

No one was more important in stimulating the new round of debates in the 1980s than Julian L. Simon, author of *The Ultimate Resource* (1981). This book attracted enormous attention, substantially because of two factors. First, it concluded that population growth would likely provide a *positive* impact on economic development of many developed, and some less developed, countries. Second, the book was accessible, well written, and organized in a 'debating', confrontational style. This included goading and prodding, the setting up and knocking down of straw men, and an examination of albeit popular, but somewhat extreme, anti-natalist positions. Simon's powerful book helped spawn a group of survey articles in the 1980s.

What accounts for Simon's positive assessments? Simon was an early advocate of evaluating the full effects of population over the intermediate to long run. He argued that the negative 'direct' impacts in the short run will probably be moderated, or sometimes overturned, when households, businesses, and/or governments react to changing prices which signal problems of resource scarcity. Two important examples of responses to population pressures can be cited: those relating to technological change and those relating to natural resource scarcity, both highlighted by Simon.

Technological change. Simon hypothesized and attempted to document that the pace of technological change, and its bias, can be stimulated by population pressures. Technological change, in turn, plays a central role in economic growth theory and has been shown in sources-of-growth studies to be a (the?) key to economic growth. Additionally, with respect to population size impacts in general, Simon observes that major social overhead projects (for example, roads, communications and irrigation) have benefited from expanded populations and scale. (For more detail, see Section 1.4 above.)

Resource depletion. Consider next the impacts of population growth on natural resource depletion. Theoretically an exhaustion of non-renewable resources (for example, coal and minerals) would appear to be inevitable in the long run. However, such a period may be in the indeterminably distant future. By contrast, Simon argued that the most relevant measure of resource scarcity is its price. He prepared many graphs of US non-renewable resource prices (deflated by price indexes in order to focus on 'real' resource trends).

Surprisingly, virtually every resource has experienced a *declining* real price over lengthy periods of time. This means, à la Simon, that resources are becoming *more* abundant over time. It seems that the

more resources are used, the more abundant they become! How can this happen? Simple. A rising resource price, due in part to population pressures, triggers several reactions that reduce or even eliminate the apparent resource scarcity. Specifically, in the short run, rising prices encourage an economizing of the resource at every level of production and consumption. In the longer run, rising prices stimulate exploration, new methods of extraction and process, and the search for substitutes. Nevertheless, Simon recognized that market failures, institutional failures, and political factors can all result in less-than-complete adjustments when population and economic development press against resource availabilities. This is particularly the case with renewable resources (such as rain forests, fisheries, the environment, and so forth) where market or institutional failures are pervasive. Without mechanisms to assign and maintain property rights, internalize externalities, and address free rider problems of public and quasi-public goods, government regulation may be required to safeguard renewable resources over time.

2.7 National Academy of Sciences, 1986

Some 15 years after the 1971 National Academy Report that highlighted 25 negative consequences of population growth, a new National Academy Report was released. In contrast to the previous study, the new report was balanced, eclectic and non-alarmist. A careful examination of its bottom line is instructive.

‘*On balance*, we reach the *qualitative* conclusion that *slower* population growth would be beneficial to economic development of *most* developing countries.’ (1986, p. 90; emphasis added)

This qualified assessment reveals key features found in most population assessments in the 1980s. Specifically: (*a*) there are both positive and negative impacts of demographic change (thus ‘on balance’); (*b*) the magnitude of the net impacts cannot be determined given current evidence (thus ‘qualitative’); (*c*) only the direction of the impact from high to low growth rates can be ascertained (thus ‘slower’ rather than ‘slow’); and (*d*) the net impact varies from country to country. In most cases it will be negative; in some positive; and in others of little impact (thus, ‘most developing countries’).

What accounts for the dramatic turnaround in the two National Academy assessments? Several factors can be advanced. First, the 1986 report extends the short-run time horizon of the 1971 report to examine individual and institutional responses to the initial impacts of population change: conservation in response to scarcity, substitution of abundant for scarce factors of production, innovation and adoption of technologies to exploit profitable opportunities, and the like. These responses are considered to be pervasive and they are judged to be important. According to the report writers: ‘the key [is the] mediating role that human behavior and human institutions play in the relation between population growth and economic processes’ (1986, p. 4).

Second, the 1986 study was assembled almost entirely by economists whose understanding of and faith in markets to induce responses that modify initial direct impacts of population change is far greater than that of other social and biological scientists.

Third, research accumulating over the 15 years between the two reports revealed a need to downgrade: (1) the concern about non-renewable resource exhaustion; (2) the adverse impact of children on the capacity to save, and in turn to undertake productive investments; and (3) the inability to invest in schooling and health facilities.

Finally, the 1986 Report upgrades the concern about population impacts on *renewable* natural resources

(such as fishing areas and rain forests) where property rights are difficult to assign and maintain. Overuse can result. It is recognized that the problems of overuse are not solely due to population growth per se, but rather institutional failure. Cutting population growth by one half, or even to zero, would not solve the problem. Rather it would slow the process and postpone the date of resource exhaustion. Government policies are needed to account for negative externalities and market failure. Slowing population growth provides time for institutional response.

3 New paradigms for modelling demography's role in economic growth: 1990 and beyond

As noted previously, Kuznets's empirical finding of an absence of notable negative correlations between population growth and per capita output growth influenced the population debate throughout the 1970s and 1980s. Simple correlations stimulated research during the 1990s as well. This time, however, statistically significant negative correlations during the 1980s drove the discussion. Interestingly, economic–demographic modelling continued in the ‘revisionist’ vein, incorporating positive and negative as well as short- and long-run influences into an economic growth model. The modelling challenge remains one of accommodating correlations that can be negative, positive or insignificant depending upon time and place.

3.1 Convergence growth models: a framework for assessing demography's impact

Renewed interest in modelling the impacts of demographic change on economic growth coincided with the emergence in the economic growth literature of the ‘technology gap’ or ‘convergence’ model. This model, formulated initially by Barro and Sala-i-Martin (1991), has been used widely to explore many hypothesized influences on economic growth, including openness to trade, form of government, and the rule of law. Since this type of modelling highlights the dynamics of the adjustment process, it is particularly relevant to examining the impacts of major shifts in the population's age distribution associated with birth and death rates that change systematically over the demographic transition. As a result, economic demographers have employed convergence paradigms to explore demographic–economic interactions.

Briefly stated, convergence models focus on the pace at which countries move from their current level of labour productivity to their long-run or steady-state level of labour productivity. The model assumes that all countries converge at the same rate from their current to their long-run levels (which can vary across countries and over time). The greater the productivity gap, the greater are the gaps of physical capital, human capital and technical efficiency from their long-run levels. Large gaps allow for ‘catching up’ through (physical and human) capital accumulation, and technology creation and diffusion across countries and over time. Indeed, many empirical studies indicate that growth rates do slow down as a country approaches its long-run productivity level, especially those studies that provide for country- and period-specific conditions that influence the long-run level of labour productivity.

Since long-run labour productivity is unobservable, empirical implementations of the model substitute a vector of ‘conditioning’ variables thought to influence long-run labour productivity. The actual specification of these conditioning variables varies notably. Consider two of their many representations. The first, by Barro (1997), highlights inflation, government consumption ratios, the rule of law, the form

of the political system, terms of trade, human capital, the total fertility rate, and life expectancy at birth (a proxy for health). The second formulation, by Bloom and Williamson (1998), highlights two categories of growth-rate determinants: economic structure variables (natural resources, schooling, access to ports, location in the tropics, whether landlocked, and extent of coastline); and economic and political policies (openness to trade, quality of institutions, and government savings share of GDP). Clearly there are many defensible perspectives on variable choice, and much is yet to be learned about the appropriate configuration of conditioning variables that influence long-run productivity levels.

3.2 Alternative demographic renderings within a convergence framework

The 1990s witnessed attempts by various researchers to model demography in a manner that accommodates both the insignificant correlations of the 1960s and 1970s as well as the significant negative correlations of the 1980s and 1990s. Three different approaches are described here. All three employ a convergence-type growth model and all employ a broad set of countries spanning the income spectrum.

Modelling through aggregate measures of fertility and mortality. Barro (1997) includes two demographic aggregate measures among his list of conditioning variables, the total fertility rate (TFR) and life expectancy. Barro's formulation thus has demography impacting the long-run equilibrium level of per capita income. The TFR captures, for example, the adverse capital-shallowing impact of more rapid population growth as well as the resource opportunity costs of bringing up children. Furthermore, while Barro treats life expectancy as a human capital proxy for health, demographers consider it to be a demographic variable. Both are statistically significant, with a higher TFR inhibiting, and longer life expectancy enhancing economic growth.

Modelling through population growth components. Kelley and Schmidt (1995) decompose population growth by examining two components (births and deaths) and by modelling their contemporaneous and lagged impacts. This approach allows for disparate impacts of fertility and mortality as well as negative short-run effects (costs of high birth and death rates) and positive long-run effects (favourable impacts of past births on current labour force growth and declining mortality). Consistent with Kuznets's earlier work, they found an absence of a net demographic impact on economic growth in the 1960s and 1970 – the separate impacts of births and deaths are notable but offsetting. Consistent with empirical work of the early 1990s, they found negative impacts throughout the 1980s. These negative correlations were in part the result of (*a*) rising short-run costs of high birth rates, (*b*) declining benefits of mortality reduction, and (*c*) insufficient labour force entry from past births to offset these increased costs.

Modelling through differential age-structure growth. In a series of papers beginning in the late 1990s, several Harvard economists argued for a demographic rendering that incorporates not only population growth but also labour growth (see, for example, Bloom and Williamson, 1998; and Bloom, Canning and Malaney, 2000). They note that, while theorists conceptualize the economic growth process in labour productivity terms, empirical growth models are generally specified in per capita terms. This makes no difference when population and labour grow at the same rate, but does when they grow at different rates.

The authors argue that the post-war period was exactly such a time since during that period demographic transitions took place in different countries at different times and at different paces. At various stages of

the demographic transition, the population and working ages (used within this framework as a proxy for labour) can grow at very different rates. In a predictable pattern, the population initially grows faster, then slower, and then faster than the working-aged population during the transition from a high-fertility, high-mortality to a low-fertility, low-mortality demographic steady-state equilibrium. (For an historical evolution of economic, sociological, and biological factors during the demographic transition, see R.A. Easterlin, 1978.)

Without allowing for differential growth rates of the population and working ages, demographic coefficient estimates (mainly population growth) will be biased. In that case the population–growth coefficient captures net demographic impacts that can be positive, negative, or neutral, depending upon time and place. Bloom and Williamson (1998) demonstrate this point for a broad cross-section of countries over the period 1965–90 in a convergence model that also includes life expectancy as a human capital variable. Consistent with some studies, their simple demographic rendering results in a positive but insignificant coefficient for the population growth rate. When supplemented by the working-age growth rate, however, that coefficient turns negative and the coefficient for the working-age growth rate is positive, both statistically significant.

Effectively, the Harvard economists append an accounting structure to translate labour productivity impacts into per capita terms. The resulting demographic specification is elegant in its simplicity, incorporating only two demographic variables that have unambiguous predicted coefficient values of –1 (for population rate of growth, Ngr) and +1 (for working-age population rate of growth, WAgr) when used to expose demography's impact on income growth per capita relative to income growth per working-age population. In that context, demography exerts its primary impact on the pace at which the long-run equilibrium is reached (Bloom and Williamson, 1998, p. 419) rather than on the long-run equilibrium level of productivity.

This is an intriguing specification. The interpretation is clear: if labour force growth exceeds population growth, then the rate of per capita income growth is boosted by demography. The Harvard economists label this phenomenon the ‘demographic gift’ that may be reaped for several decades after the onset of fertility decline as new labour force entrants from earlier large birth cohorts outpace fertility. The ‘gift’ was large throughout the 1965–90 period for Japan and other Asian Tigers because of the early and rapid pace of their demographic transition. Of course, the converse of the ‘gift’ began to be felt in the 1990s as new labour force entry from smaller birth cohorts was outpaced by labour force exit of the aging population. The model predicts productivity outpacing per capita income growth over several decades into the future in these Asian (and other) countries.

Note that the qualitative predictions are based on theoretically determined coefficients on WAgr and Ngr of +1 and –1, respectively. To the extent that estimated coefficients deviate from +1 and –1, WAgr and Ngr play an additional role in the determination of the long-run productivity level. The Harvard studies provide some guidance in this area. In their earlier study, Bloom and Williamson (1998) estimate coefficients that differ significantly from +1 and –1. However, in a later study that further elucidates the accounting, Bloom, Canning and Malaney (2000) find no significant difference from those values. If that is the case, then the model at once makes an important contribution and is somewhat narrower than many in the literature which admit both short-run and long-run impacts of demographic change as a part of the theoretical structure. Yet modelling demography in growth equations tends to be both imprecise and ad hoc. In contrast, the Bloom and Williamson model is relatively clear in interpretation, and it

targets the shorter-run impacts that are of primary interest to policymakers.

3.3 The bottom line

Bloom and Williamson (1998) estimated that as much as one-third of the average per capita income growth rate in East Asian countries over the period 1965–90 is explained by population dynamics. Kelley and Schmidt (2001) evaluated eight distinct demographic renderings within a convergence model using a consistent set of conditioning variables – those described above for Barro's variant. Among others, these renderings included Barro's TFR; a ‘naive’ variant predating the 1990s work that simply includes Ngr; a ‘components’ model (contemporaneous and lagged birth rates and the death rate: Kelley and Schmidt, 2001); two variants of the Harvard transitions framework; and demographic extensions to several variants.

Kelley and Schmidt (2001) find that on average, across all eight demographic formulations and over their full 86-country sample (covering the full income spectrum), approximately 21 per cent of the combined impacts on change in the per capita income growth rate is accounted for by changes in the demographic variables in the various models. What is striking about this result is that the 21 per cent is fairly stable across all eight demographic renderings, from one that is quite simplistic (Ngr only) to those that incorporate short-, intermediate- and long-term population effects. On the one hand, this should not be terribly surprising because of the interconnectedness of all of the demographic measures. On the other hand, while population matters, it is still important to determine why.

Although there is an emerging consensus that the magnitude of the impacts of population growth have been sizeable (for example, 21 per cent globally and as much as 33 per cent in East Asia), the reasons why this is the case are still both contestable and not well understood. Are the demographic determinants primarily longer-run impacts, or are they mainly shorter-run transitional dynamics that are diminishing? Will the so-called ‘demographic gift’ of these dynamics in the past reveal themselves as a ‘demographic drag’ in the future, deriving from reduced fertility, slow population growth and ageing? Or will a new mechanism reveal itself? For example, (*a*) will future modelling better expose the components of labour force change (for example, utilization rates, age- and/or gender-specific participation rates); and (*b*) will fertility and mortality be endogenously specified to better reveal the dynamics of the demographic transition about which the field of economic demography has much to say? Whatever the outcome, the stage is set for another round of research, pinning down the results of the past with the goal of understanding the future.

See Also

- family decision making
- fertility in developing countries
- marriage and divorce
- retirement

Bibliography

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 29, 155–73.

Barro, R.J. 1997. *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge, MA: MIT Press.

Barro, R.J. and Sala-i-Martin, X. 1991. Convergence. *Journal of Political Economy* 100, 223–51.

Birdsall, N., Kelley, A.C. and Sinding, S., eds. 2001. *Demography Matters: Population Change, Economic Growth and Poverty in the Developing World*. Oxford: Oxford University Press.

Bloom, D.E. and Williamson, J.G. 1998. Demographic transitions and economic miracles in emerging Asia. *World Bank Economic Review* 12, 419–55.

Bloom, D.E., Canning, D. and Malaney, P. 2000. Demographic change and economic growth in Asia. *Population and Development Review* 26, 257–1990.

Boserup, E. 1965. *Conditions of Agricultural Growth*. Chicago: Aldine.

Boserup, E. 1981. *Population and Technological Change*. Chicago: University of Chicago Press.

Chenery, H. and Syrquin, M. 1975. *Patterns of Development: 1950–1970*. Oxford: Oxford University Press.

Coale, A.J. and Hoover, E.M. 1958. *Population Growth and Economic Development in Low-Income Countries*. Princeton: Princeton University Press.

Easterlin, R.A. 1978. The economics and sociology of fertility: a synthesis. In *Historical Studies of Changing Fertility*, ed. C. Tilly. Princeton: Princeton University Press.

Glover, D.R. and Simon, J.L. 1975. The effect of population density on infrastructure: the case of road building. *Economic Development and Cultural Change* 23, 453–68.

Higgins, M. 1998. Demography, national savings, and international capital flows. *International Economic Review* 39, 343–69.

Kelley, A.C. 1988. Economic consequences of population change in the Third World. *Journal of Economic Literature* 26, 1685–728.

Kelley, A.C. 2001. The population debate in historical perspective: revisionism revisited. In Birdsall, Kelley and Sinding (2001).

Kelley, A.C. and Schmidt, R.M. 1995. Aggregate population and economic growth correlations: the role of the components of demographic change. *Demography* 32, 543–55.

Kelley, A.C. and Schmidt, R.M. 2001. Economic and demographic change: a synthesis of models, findings and perspectives. In Birdsall, Kelley and Sinding (2001).

Kuznets, S. 1960. Population change and aggregate output. In *Demographic and Economic Change in Developed Countries*. National Bureau of Economic Research. Princeton: Princeton University Press.

Lee, R.D., Mason, A. and Miller, T. 2001. Saving, wealth and population. In Birdsall, Kelley and Sinding (2001).

Malthus, T.R. 1798. *An Essay on the Principle of Population*. Harmondsworth: Penguin, 1970.

Mason, A. 1987. National saving rates and population growth: a new model and new evidence. In *Population Growth and Economic Development: Issues and Evidence*, ed. D.G. Johnson and R.D. Lee. Madison: University of Wisconsin Press.

National Academy of Sciences. 1971. *Rapid Population Growth: Consequences and Policy Implications*, vol. 1: *Summary and Recommendations;* vol. 2: *Research Papers*. Baltimore: Johns Hopkins University Press.

National Academy of Sciences. 1986. *Population Growth and Economic Development: Policy Questions*. Washington, DC: National Research Council.

Schultz, T.P. 1987. Schooling Expenditures and Enrollments 1960–1980: The Effects on Income, Prices and Population Growth. In *Population Growth and Economic Development Issues and Evidence*, ed. D. Gale Johnson and R.D. Lee. Madison: University of Wisconsin Press.

Simon, J.L. 1981. *The Ultimate Resource*. Princeton: Princeton University Press.

Simon, J.L. 1996. *The Ultimate Resource 2*. Princeton: Princeton University Press.

Srinivasan, T.N. 1988. Modeling growth and economic development. *Journal of Policy Modeling* 10, 7–28.

United Nations. 1953. *The Determinants and Consequences of Population Trends*. New York: United Nations.

United Nations. 1973. *The Determinants and Consequences of Population Trends*. New York: United

Nations.

How to cite this article

Kelley, Allen C. and Robert M. Schmidt. "economic demography." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000075>
doi:10.1057/9780230226203.0428

The New Palgrave Dictionary of Economics Online

economic development and the environment

Ian Coxhead

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic development in low-income economies is initially highly resource-intensive. Resource depletion and pollution damage is often estimated to reduce ‘real’ GDP growth by between one and two per cent per year. Growth and structural change alter the environment–development nexus in nonlinear fashion. Policy reforms, global market integration, and institutional development all alter the propensity for growth to generate environmental damage. The emergence of new trade patterns among developing countries has created new challenges in the measurement and analysis of development–environment interactions. Larger developing economies are now emerging as major sources of emissions that contribute to global climate change.

Keywords

biodiversity; comparative advantage; conservation; economic development and the environment; environmental economics; environmental Kuznets curve; greenhouse gas emissions; growth and international trade; Heckscher–Ohlin trade theory; import substitution; income effects; natural resources; non-use amenities; North–South economic relations; pollution; pollution haven hypothesis; poverty alleviation; Ricardian trade theory; social norms; structural change; sustainable development

Article

Economic development depends on sustained per capita income growth and entails dramatic changes in production structure. In low-income economies, growth typically stimulates markets and promotes the evolution of institutions that constrain behaviour according to social norms. The expansion of trade in relation to GDP is another common accompaniment to growth. Each of these has effects on ‘the environment’, which in a developing-country setting refers not only to phenomena such as water and air quality but also, importantly, to natural resource stocks such as forests, fisheries and soils. Conversely, changes in environmental quality, including resource stock drawdowns, may affect

economic development in a dynamic interaction. This feedback is hard to quantify; however, the World Bank's *World Development Indicators* series now includes ‘adjusted’ national accounts data reporting GDP and savings net of the implied value of resource depletion and environmental damage (Bolt, Matete and Clemens, 2002). These indicate that environmental damage can reduce GDP growth by as much as one to two per cent per year. On a broader scale, growth of large low-income economies like China and India is beginning to have ramifications not only for their own environmental conditions, but also for the global environment through transboundary pollution spillovers and greenhouse gas (GHG) emissions.

The welfare of the poor in low-income countries is intimately linked to their access to environmental assets, and especially to the natural resource base. Despite this, the central concerns of environmental and resource economics – the economic costs of pollution and natural resource depletion – have only recently begun to be linked to models of economic development. Publication of the so-called Brundtland Report (WCED, 1987) was a watershed event; since then, ‘no account of economic development would be regarded as adequate if the environmental-resource base were absent from it’ (Dasgupta and Mäler, 1995, p. 2734).

Growth in low-income economies is inevitably associated with higher resource demands and increased pollution intensity per unit of income generated. Other things equal, more economic activity generates more environmental damage monotonically through a scale effect. The relationship may be nonlinear, however. As income grows, environmental damage per unit of additional income may initially rise, then decline. This conjecture, known as the environmental Kuznets curve (EKC), posits that scale effects dominate all other influences on the growth–environment relationship at low income levels, but that, as incomes rise, changes in the composition of production, technological improvements, and income-elastic preferences for conservation and a cleaner environment become more influential (Grossman and Krueger, 1993). Institutional and legal constraints on pollution and resource depletion, initially so weak as to create a form of open access for polluters and resource depleters, may also evolve or be applied with greater vigour as incomes increase, whether due to income effects or to increased recognition of limits to growth imposed by pollution and resource scarcity (Stokey, 1998). Despite the heuristic value of EKC, however, empirical tests in low-income economies are plagued by data and measurement problems. Most notably, there is no robust evidence of an EKC for resource-depleting activities such as deforestation.

Changes in production structure and factor demands are also inherent to development. The most prominent manifestation of structural change in low-income countries is the relative decline of agricultural and resource sectors as contributors to GDP and employment. This has clear environmental implications when the majority of the population is initially dependent on the natural resource base. In capital-scarce economies, forest and land conversion for agriculture and the exploitation of fisheries and other resource stocks are standard strategies for increasing labour productivity and generating surpluses. Accordingly, early stages of development are characterized by rapid resource depletion – most visibly in the form of tropical deforestation. Such processes are abetted by conditions of open access (Barbier, 2005).

Whether the depletion rate eventually slows – a prerequisite for sustainable development – depends largely on the extent to which surpluses are used to build capacity in secondary and tertiary industries making more intensive use of reproducible resources such as labour, technology and human capital. In

this way, the central story of structural change in low-income economies is intimately linked to the evolution of demands on the environmental and natural resource base. Sustained growth leads to a relative reduction in dependence on natural resources, and thus makes it easier for society to agree to promote conservation, biodiversity retention and non-use amenities. Conversely, macroeconomic failures, often in combination with rapid population growth, high transactions costs and market failures, can lead low-income economies into unsustainable cycles of poverty, resource over-exploitation, and institutional failure.

Trade is another influential source of structural change. Early development policies stressing import substitution and de-emphasizing trade have, in most countries, been supplanted by greater outward orientation. Trade-to-GDP ratios have risen and domestic prices have tended to converge on world market prices, thus altering domestic production and investment incentives. With the exception of resource-poor East Asian countries like Korea and Taiwan, the pursuit of comparative advantage in low-income countries initially means expanded exports of tropical agriculture, forestry and fisheries and of resource-based semi-manufactures such as sawnwood. Both the growth of global demand and the pro-trade effects of policy reforms encourage accelerated resource drawdowns; unless property rights and externalities are adequately dealt with, these are likely to occur at socially excessive rates (Coxhead and Jayasuriya, 2003). A related idea known as the pollution haven hypothesis posits that weak environmental laws and unresolved externalities may lead developing countries to specialize in pollution-intensive industrial activities (Copeland and Taylor, 1994).

Whereas early policy advice to developing countries typically stressed the desirability of exploiting resource wealth to create jobs and earn foreign exchange, contemporary concerns about exhaustibility and the integrity of ecological systems have led to more cautious counsel and an emphasis on sustainable development. Such advice, however, is often difficult to implement as policy in the face of pressures to promote growth and alleviate poverty in the current generation.

New issues in the development–environment relationship continue to emerge as economies grow and become more globalized. Traditionally, trade-environment analyses used Ricardian or Heckscher–Ohlin models of North–South interactions in which welfare growth in resource-abundant South is contingent on trade with industrialized North and on domestic externalities or market failures (for example, Chichilnisky 1994). However, South–South trade – or, in the case of China's emergence as a major market for resource exports from Asia, Africa, and Latin America, ‘East–South’ trade – is now growing much faster than trade of the North–South type. South–South trade is a form of internationally fragmented production in which primary products or semi-manufactures are exported from one low-income country to another to be used in production of final goods. The latter low-income economy thus moves to ‘clean’ growth based on labour-intensive manufactures, while growth in the former becomes more resource-intensive. Countries in the South may have comparative advantage in either clean or dirty goods – or both. Conventional models and measures for evaluating environmental costs of growth must be adapted to such new modalities.

Other new trends reflect the growing global influence of large developing economies. In poor countries, about 50 per cent of carbon dioxide emissions (the primary sources of GHGs) comes from land conversion. But total emissions increase rapidly with energy demands driven by growth, urbanization and industrialization. According to the International Energy Agency, China accounted for 13 per cent of global energy-related CO_2 emissions in 2006, and is expected to overtake the USA as the largest CO_2

source by 2009; India is now following a similar path (IEA, 2006). Under the 1997 Kyoto Protocol, these economies are not required to limit GHG emissions. But, even if they do take major steps to limit pollution intensity, scale effects of their growth will ensure that global pollution externalities will continue to expand for the foreseeable future. In turn, concerns over the global environmental consequences of growth in low-income countries will find increasingly forceful expression in international negotiations not only on the environment but also on trade and other forms of international integration.

See Also

- climate change, economics of
- environmental economics
- environmental Kuznets curve
- poverty alleviation programmes
- sustainability

Bibliography

Barbier, E.B. 2005. *Natural Resources and Economic Development*. Cambridge: Cambridge University Press.

Bolt, K., Matete, M. and Clemens, M. 2002. Manual for calculating adjusted net savings. Mimeo, Environment Department, World Bank.

Chichilnisky, G. 1994. North–South Trade and the global environment. *American Economic Review* 84, 851–74.

Copeland, B.R. and Taylor, M.S. 1994. North–South trade and the environment. *Quarterly Journal of Economics* 109, 755–87.

Coxhead, I. and Jayasuriya, S.K. 2003. *The Open Economy and the Environment: Development, Trade and Resources in Asia*. Cheltenham, UK and Northampton, MA: Edward Elgar.

Dasgupta, P. and Mäler, K.-G. 1995. Poverty, institutions and the natural resource base. In *Handbook of Development Economics*, vol. 3A, ed. J. Behrman and T.N. Srinivasan. Amsterdam: North-Holland.

Grossman, G.M. and Krueger, A.B. 1993. The environmental impacts of a North American Free Trade Agreement. In *The US-Mexico Free Trade Agreement*, ed. P. Garber. Cambridge, MA: MIT Press.

IEA (International Energy Agency). 2006. *World Energy Outlook 2006*. Paris: IEA.

Stokey, N. 1998. Are there limits to growth? *International Economic Review* 39, 1–31.

WCED (World Commission on Environment and Development). 1987. *Our Common Future*. Oxford and New York: Oxford University Press.

How to cite this article

Coxhead, Ian. "economic development and the environment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000308>
doi:10.1057/9780230226203.0429

The New Palgrave Dictionary of Economics Online

economic epidemiology

Tomas J. Philipson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economic analysis of epidemiological issues has different implications for disease and its optimal control from those of traditional analysis of such issues. It views undesirable disease occurrence as the result of self-interested behaviour in the presence of constraints. Unlike with methods used in public health, the effects and desirability of disease-reducing public interventions are then evaluated in terms of how they improve the private behaviour essential to controlling disease. Economic epidemiology has been applied to a wide range of topics, including infectious diseases such as AIDS, and also to non-infectious behaviour such as smoking, obesity, and crime.

Keywords

economic epidemiology; health subsidy; morbidity; mortality; programme evaluation; public health; rational epidemics

Article

The fast-growing literature on the economic analysis of epidemiological issues (see Philipson, 2000, for a review) delivers very different implications about disease occurrence and its optimal control from those of traditional analysis of the same issues in the field of public health. At the risk of vastly oversimplifying the positive component of the public health approach, the traditional analysis comprises empirical methods and analysis aimed at identifying and quantifying the effects of ‘risk factors’ on health outcomes. These factors are typically defined as covariates that negatively affect the measured health outcomes – for example, the effects of smoking on lung cancer or the effects of obesity on heart disease. Thereafter, the normative component of the public health approach is concerned with attempts to reduce the measured risk factors, whether through private or public intervention, and to thereby improve health outcomes.
This approach drastically differs from that of economic epidemiology, which attempts to explain

undesirable disease occurrence as the result of self-interested behaviour in the presence of constraints. The effects and desirability of disease-reducing public interventions are then evaluated in terms of how they improve the private behaviour essential to controlling disease in the first place. In some sense, the public health approach aims to improve health, whereas the economic approach aims to improve economic efficiency, even if that does not necessarily improve health. Just as closing highways would improve health but impair economic efficiency, the two approaches often clash in desired interventions. The public health approach, therefore, more often favours public intervention, and sometimes simply assumes that the existence of a health problem is sufficient cause for intervention, potentially because it lacks a theory about how private incentives affect the observed level of disease across time and populations.

Economic epidemiology and infectious disease

Infectious diseases cause roughly one-third of all deaths worldwide and represent the primary cause of mortality in the world. Historically, the share of worldwide mortality due to infectious diseases has been even greater, although data tend to be less reliable for earlier periods. Morbidity and mortality from infectious diseases such as tuberculosis, malaria and acute respiratory infection have always been at the forefront of public policy in developing countries, where infectious diseases accounted for nearly one-half of mortality in the 1990s.

Worldwide concern about infectious disease has received renewed interest in public policy discussions given the disastrous impacts of HIV/AIDS and the potential threat of bird flu. Like most communicable diseases, especially those that are potentially fatal, HIV has incited an extensive governmental response, consisting of regulatory measures, subsidies for research, education, treatment, testing and counselling. Here we review the main contributions of economic epidemiology in predicting both the short- and the long-run behaviour of infectious disease, as well as the effects and desirability of public health interventions that attempt to reduce such disease.

Philipson and Posner (1993) provide the first systematic analysis of rational infectious disease epidemics in the context of AIDS. Kremer (1996) analyses the effects of a reduction in the number of one's sexual partners on the growth of disease. The predictions of such models rely crucially on the prevalence elasticity of private demand for prevention against disease, that is, the degree to which prevention increases in response to disease occurrence. Prevalence-elastic behaviour has different implications for the susceptibility to infection than standard epidemiological models of disease occurrence as discussed in Philipson (1995). Evidence of the degree of prevalence-elastic demand is discussed in Ahituv, Hotz and Philipson (1996) and Auld (2003; 2006). Oster (2006) attempts to explain the lack of prevalence-elastic demand in Africa by the competing risks that lower the demand for prevention in that part of the world. Lakdawalla, Sood and Goldman (2006) provide evidence that demand is sensitive to overall risk, both in terms of prevalence and the cost of infection as when reduced by new medical technologies.

This type of prevalence-elastic behaviour has two major implications. First, growth of infectious disease is self-limiting because it induces preventive behaviour. Second, since the decline of a disease discourages prevention, initially successful public health efforts actually make it progressively harder to eradicate infectious diseases. Geoffard and Philipson (1996) discuss a very general result concerning the inability of private markets to eradicate disease when demand is prevalence-elastic because a

disappearing disease implies less prevention. Barrett (2003; 2004) also analyses the implications of economic efficiency for optimal eradication. See also Gersovitz and Hammer (2003; 2004; 2005). Regarding the value of public health interventions, Mechoulan (2004) analyses the prevalence and efficiency implications of HIV testing. Geoffard and Philipson (1996) argue that eradication is never Pareto optimal when only the current generation is considered. However, the missing market is dynamic: future generations cannot pay vaccine producers for the benefit they derive from the producers' product. Brito, Sheshinski and Intriligator (1991) analyse the non-standard efficiency implications of mandatory vaccinations.

Moreover, the prevalence elasticity of demand lowers the price elasticity of demand, which implies that Pigouvian-style subsidies to stimulate prevention may have only limited success. This occurs because demand rises among those who are subsidized and falls among those who are not – in the extreme case, total demand is inelastic to subsidies. In addition, prevalence competes with public interventions in inducing protective activity, which makes the timing of the public intervention a crucial factor in determining its economic efficiency. If the subsidy is not prompt enough, the growth in prevalence will have already induced protection.

A growing literature examines the optimal control of infectious diseases in the presence of antibiotic resistance (see, for example, Laxminarayan and Brown, 2001; Laxminarayan and Weitzman, 2002; Laxminarayan, 2002; and Horowitz and Moehring, 2004). The standard, positive external effect of treating more individuals with an infectious disease is partly or fully offset by the negative external effect induced by increased antibiotic resistance. The R&D problem induced by external consumption effects such as antibiotic resistance is discussed in Philipson, Mechoulan and Jena (2006).

Economic epidemiology has also considered the welfare losses induced by disease, the welfare effects of R&D in developing new methods of prevention and treatment (Philipson, 1995), and how these contrast with cost-of-illness studies of disease burden.

Spread of economic epidemiology to other fields

Several other topics have grown out of this more systematic analysis of infectious disease by economists. One strand is the analysis of public health-related issues such as obesity (Philipson and Posner, 2003; Lakdawalla, Philipson and Bhattacharya, 2005). The addictive aspect of obesity is analysed by Cawley (1999). Empirical studies explaining the observed growth in obesity, whether it includes a rise in caloric intake or fall in caloric expenditure, include Cutler, Glaeser and Shapiro (2003). Chou, Grossman and Saffer (2004) and Rashad and Grossman (2004) analyse the co-variation between the growth of obesity and smoking and fast-food establishments in the United States. The important and rich set of issues raised by growth in obesity promises a useful role for economic analysis.

Another area in which economic analysis of epidemiological issues has emerged is the economic analysis of clinical trials (see, for example, Philipson and DeSimone, 1997; Philipson and Hedges, 1998; Malani, 2006). This literature deals with the non-traditional aspects of programme evaluation that are unique to clinical trials – for example, the blinding of subjects. Economic analysis of clinical trials differs from bio-statistical analysis in that subjects are assumed to act in their best interest rather than be passively observed.

The stark difference between economic explanations of disease occurrence on the one hand and the evaluation of public interventions aimed at limiting disease on the other implies that economics may have a very useful role to play in understanding these issues.

See Also

- health economics

Bibliography

Ahituv, A., Hotz, J. and Philipson, T. 1996. Is AIDS self-limiting? Evidence on the prevalence elasticity of the demand for condoms. *Journal of Human Resources* 31, 869–98.

Auld, M.C. 2003. Choices, beliefs, and infectious disease dynamics. *Journal of Health Economics* 22, 361–77.

Auld, M.C. 2006. Estimating behavioral response to the AIDS epidemic. *Contributions to Economic Analysis and Policy* 5(1), Article 1.

Barrett, S. 2003. Global disease eradication. *Journal of the European Economic Association* 1, 591–600.

Barrett, S. and Hoel, M. 2004. Optimal disease eradication. Working Paper No. 50.04, FEEM.

Brito, D., Sheshinski, E. and Intriligator, M. 1991. Externalities and compulsory vaccinations. *Journal of Public Economics* 45, 69–90.

Cawley, J. 1999. Obesity and addiction. Ph.D. dissertation, Department of Economics, University of Chicago.

Chou, S.-Y., Grossman, M. and Saffer, H. 2004. An economic analysis of adult obesity, results from the Behavioral Risk Factor Surveillance System. *Journal of health Economics* 23, 565–87.

Cutler, D.M., Glaeser, E.L. and Shapiro, J.M. 2003. Why have Americans become more obese? *Journal of Economic Perspectives* 17(3), 93–118.

Geoffard, P.-Y. and Philipson, T. 1996. Rational epidemics and their public control. *International Economic Review* 37, 603–24.

Gersovitz, M. and Hammer, J.S. 2003. Infectious diseases, public policy, and the marriage of economics and epidemiology. *The World Bank Research Observer* 18, 129–57.

Gersovitz, M. and Hammer, J.S. 2004. The economical control of infectious diseases. *Economic Journal* 114, 1–27.

Gersovitz, M. and Hammer, J.S. 2005. Tax/subsidy policies toward vector-borne infectious diseases. *Journal of Public Economics* 89, 647–74.

Grossman, M. and Rashad, I. 2004. The economics of obesity. *Public Interest* (156), 104–12.

Horowitz, B.J. and Moehring, B.H. 2004. How property rights and patents affect antibiotic resistance. *Health Economics* 13, 575–83.

Kremer, M. 1996. Integrating behavioral choice into epidemiological models of the AIDS epidemic. *Quarterly Journal of Economics* 111, 549–73.

Lakdawalla, D., Philipson, T. and Bhattacharya, J. 2005. Welfare enhancing technological change and the growth of obesity. *American Economic Review* 95, 253–8.

Lakdawalla, D., Sood, N. and Goldman, D. 2006. HIV breakthroughs and risky sexual behavior. *Quarterly Journal of Economics* 121, 1063–102.

Laxminarayan, R. 2002. How broad should the scope of antibiotics patents be? *American Journal of Agricultural Economics* 84, 1287–92.

Laxminarayan, R. and Brown, G.M. 2001. Economics of antibiotics resistance: a theory of optimal use. *Journal of Environmental Economics and Management* 42, 183–206.

Laxminarayan, R. and Weitzman, M.L. 2002. On the implications of endogenous resistance to medications. *Journal of Health Economics* 21, 709–18.

Malani, A. 2006. Identifying placebo effects with data from clinical trials. *Journal of Political Economy* 114, 236–56.

Mechoulan, S. 2004. HIV testing, a Trojan horse? *Topics in Economic Analysis and Policy* 4(1), article 18.

Oster, E. 2006. HIV and sexual behavior change: why not Africa? Working paper, Graduate School of Business, University of Chicago.

Philipson, T. 1995. The welfare loss of disease and the theory of taxation. *Journal of Health Economics* 14, 387–96.

Philipson, T. 2000. Economic epidemiology and infectious disease. In *Handbook of Health Economics*, ed. T. Culyer and J. Newhouse. Amsterdam: North-Holland.

Philipson, T. and DeSimone, J. 1997. Experiments and subject sampling. *Biometrika* 84, 618–32.

Philipson, T. and Hedges, L. 1998. Subject evaluation in social experiments. *Econometrica* 66, 381–409.

Philipson, T., Mechoulan, S. and Jena, A.B. 2006. IP and external consumption effects: generalizations from health care markets. Working Paper No. 11930. Cambridge, MA: NBER.

Philipson, T. and Posner, R.A. 1993. *Private Choices and Public Health, An Economic Interpretation of the AIDS Epidemic*. Cambridge, MA: Harvard University Press.

Philipson, T. and Posner, R.A. 2003. The long run growth of obesity as a function of technological change. *Perspectives in Biology and Medicine* 46(3), 87–108.

Rashad, I. and Grossman, M. 2004. The economics of obesity. *Public Interest* (156), 104–12.

How to cite this article

Philipson, Tomas J. "economic epidemiology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000261> doi:10.1057/9780230226203.0430

The New Palgrave Dictionary of Economics Online

economic governance

Avinash K. Dixit

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic governance consists of the processes that support economic activity and economic transactions by protecting property rights, enforcing contracts, and taking collective action to provide appropriate physical and organizational infrastructure. These processes are carried out within institutions, formal and informal. The field of economic governance studies and compares the performance of different institutions under different conditions, the evolution of these institutions, and the transitions from one set of institutions to another.

Keywords

accountability; arbitration; authoritarianism and economic development; Better Business Bureaus (USA); Coase, R.; colonization; commitment; common law vs civil law; common-pool resource management; corporate governance; corruption; corruption control; credit rating agencies; democracy and economic development; democracy: parliamentary vs presidential; discriminating alignment hypothesis; economic governance; electoral rules; ethnic trading networks; evolution of institutions; exploitation; formal vs informal (private) contract enforcement; for-profit private institutions; governance; growth and governance; incentive compatibility; inequality; information verifiability; institutions and transition; international contracts; Mafia; market institutions; monitoring mechanisms; opportunism; political stability; Prisoner's Dilemma; property rights enforcement; regulatory burden; relation-based vs rule-based governance; reputation mechanisms; rule of law; slavery; social networks; social norms; subgame perfection; title registration; transaction cost economics; trust; voice; World Bank; World Trade Organization; Yakuza

Article

Formal and informal institutions arise and evolve to underpin economic activity and exchange by

protecting property rights, enforcing contracts, and collectively providing physical and organizational infrastructure. The field of economic governance studies and compares these institutions: state politico-legal institutions, private ordering within the law (credible contracting, arbitration), for-profit governance (credit-rating agencies, organized crime), and social networks and norms. Private institutions can outperform the state's legal system in obtaining and interpreting relevant information, and imposing social sanctions on the violators of norms. But private institutions are often limited in size; as economic activity expands, a transition towards more formal institutions is usually observed.

Concepts and taxonomies

The term ‘governance’ has exploded from obscurity to ubiquity in economics since the 1970s. A search of the EconLit database shows clear evidence of this explosion. In the relevant categories (title, keywords and abstracts), there are just five occurrences of the word from 1970 to 1979. The number jumps to 112 for the 1980s and 3,825 for the 1990s. Since 2000 to the time of this writing (December 2005), there are already 7,948.

The Oxford English Dictionary gives several definitions of the word ‘governance’: (*a*) the action or manner of governing; controlling, directing, or regulating influence; control, sway, mastery; the state of being governed; good order; (*b*) the office, function, or power of governing; authority or permission to govern; that which governs; (*c*) the manner in which something is governed or regulated; method of management, system of regulations; a rule of practice, a discipline; and (*d*) the conduct of life or business; mode of living; behaviour, demeanour; discreet or virtuous behaviour; wise self-command. These diverse meanings allow the word to be used (and sometimes misused) for almost any context of economic decision-making or policy.

Two areas of application merit special mention. One is *corporate governance*. This analyses the internal management of a corporation – organizational structure and the design of incentives for managers and workers – and the rules and procedures by which the corporation deals with its shareholders and other stakeholders.

The second is *economic governance*; Williamson (2005) expresses its theme as the ‘study of good order and workable arrangements’. This includes the institutions and organizations that underpin economic transactions by protecting property rights, enforcing contracts, and organizing collective action to provide the infrastructure of rules, regulations, and information that are needed to lend feasibility or workability to the interactions among different economic actors, individual and corporate. Different economies at different times have used different institutions to perform these functions, with different degrees of success. The field of economic governance studies and compares these different institutions. It includes theoretical models and empirical and case studies of the performance of different institutions under different circumstances, of how they relate to each other, of how they evolve over time, and of whether and how transitions from one to another occur as the nature and scope of economic activity and its institutional requirements change.

Corporate governance and economic governance are connected because the boundary of a corporation is itself endogenous, determined by the same considerations of information and commitment costs that raise problems of internal organization as well as those of property and contract (Coase, 1937). Specifically, the nature of transaction costs may make it more efficient to handle some problems of

governance by merging the two parties, for example by vertical integration (Williamson, 1975; 1995). But it is analytically convenient to separate the two. This article concerns economic governance. To avoid constant repetition, I will simply call it ‘governance’ here unless some explicit reference to corporate governance is relevant.

Governance was neglected by economists for a long time, perhaps because they expected the government to provide it efficiently. However, experience with less developed and reforming economies, and observations from economic history, have led economists to study non-governmental institutions of governance.

Governance is not a field per se; it is an organizing or encompassing concept that bears on issues in many fields, including institutions and organizational behaviour, economic development and growth, industrial organization, law and economics, political economy, comparative economic systems, and various subfields of these.

We can organize the subject by classifying institutions along different dimensions. As is usual with such taxonomies, these are conceptual categories to help organize our thinking and analysis. In reality, there are significant differences within each category and overlaps across categories.

The first dimension concerns the purpose of the institution. The categories are: (*a*) protection of property rights against theft by other individuals and usurpation by the state itself or its agents, (*b*) enforcement of voluntary contracts among individuals, and (*c*) provision of the physical and regulatory infrastructure to facilitate economic activity and the functioning of the first two categories of institutions. We might also consider a fourth category, namely, the deep institutions that are essential to avoid serious cleavages or alienation that threaten the cohesion of the society itself. But this has not been studied in this context so far.

The second dimension concerns the nature of the institution. The categories are: (*a*) the formal state institutions that enact and enforce the laws, including the legislature, police, judiciary and regulatory, agencies, (*b*) institutions of private ordering that function under the umbrella of state law, for example various forums for arbitration, (*c*) private for-profit institutions that provide information and enforcement, and (*d*) self-enforcement within social or ethnic groups and network. My discussion is organized in sections along this dimension.

A third dimension distinguishes institutions that arise and evolve organically from those that are designed purposively; self-enforcing groups are often organic while the first three categories in the second dimension usually require some measure of design. This matters for the evolution of institutions of governance (see Greif, 2006, especially ch. 6; Williamson 2005 p. 1).

Formal institutions of the state

There is broad agreement that the quality of institutions of governance significantly affects economic outcomes. The importance of protecting property rights, both from other individuals and from predation by the state itself, is generally recognized and documented (for example, De Soto, 2000). But serious disputes about the precise measures of quality of institutions, and about many details of the causal mechanisms by which they affect economic outcomes, remain.

At the broadest level, the distinction is between democracy and authoritarianism, each of which comes in many different varieties. Democracy has many normative virtues, but its worth in governance is less

clear. Barro (1999, p. 61) finds an inverse U-shaped relationship between economic growth and a continuous measure of democracy – ‘more democracy raises growth when political freedoms are weak, but depresses growth when a moderate amount of freedom is already established’ – but the fit is relatively poor. Persson (2005), using cross-sectional as well as panel data, finds that the crude distinction between democratic and non-democratic forms of government is not enough. The precise form of democracy matters for policy design and economic outcomes: ‘parliamentary, proportional, and permanent democracies seem to foster the adoption of more growth-promoting structural policies, whereas … presidential, majoritarian, and temporary democracy do not’ (Persson, 2005, p. 22). However, Keefer (2004, p. 10), after surveying a wide-ranging literature on electoral rules and legislative organizations, concludes that they affect policies but are not a crucial determinant of success: ‘electoral rules … almost surely do not explain why some countries grow and others do not’, and ‘the mere fact that developing countries are more likely to have presidential forms of government is unlikely to be a key factor to explain slow development.’

Democracy can be important for governance because its reliance on rules and procedures provides citizens with protection against predation by the state or its agents. Indeed, the elite, which might otherwise prefer to rule unconstrained, may find it in its own interest to make a credible commitment not to steal from the population by creating and fostering democracy (Acemoglu, 2003; Acemoglu and Robinson, 2005). Greif, Milgrom and Weingast (1994) discuss how groups of traders (guilds) in late medieval Europe took collective action to counter rulers’ incentives to violate their members’ property rights.

Even in a democracy, agents of the state may pursue their private interests using corruption, complex regulations to extract rent, and favouritism. In fact, an emerging literature argues that economic growth, at least in its early stages, is better promoted under suitably authoritarian regimes. Glaeser et al. (2004) argue that less developed countries that achieve economic success do so by pursuing good policies, often under dictatorships, and only then do they democratize. While these conclusions are controversial, these authors’ criticisms of the measures of institutions used in the research that argues for the primacy of institutions in general, and of democracy in particular, are telling. Giavazzi and Tabellini (2005) find a positive feedback between economic and political reform, but they also find that the sequence of reforms matters, and countries that implement economic liberalization first and then democratize do much better in most dimensions than those that follow the opposite route. In practice, of course, it is difficult to ensure *ex ante* that an authoritarian ruler will implement good governance.

Many different measures of institutional quality exist. World Bank researchers Kaufman, Kraay and Mastruzzi (2005, which contains citations to their earlier work) have constructed six: (*a*) Voice and Accountability – measuring political, civil and human rights; (*b*) Political Instability and Violence – measuring the likelihood of violent threats to, or changes in, government, including terrorism; (*c*) Government Effectiveness – measuring the competence of the bureaucracy and the quality of public service delivery; (*d*) Regulatory Burden – measuring the incidence of market-unfriendly policies; (*e*) Rule of Law – measuring the quality of contract enforcement, the police, and the courts, as well as the likelihood of crime and violence; and (*f*) Control of Corruption – measuring the exercise of public power for private gain, including both petty and grand corruption and state capture. Of these, (*e*), (*f*) and also (*b*) concern the most basic institutions for protection of property rights and enforcement of contracts, (*a*) relates to governance because voice and accountability can reduce the severity of the agency problem

between the citizens and the agencies of the state, and (*c*) and (*d*) pertain to what I called provision of the infrastructure of governance. Conceptually they are a mixed bag; the quality of some of them can itself depend on the quality of other more basic ones, and some are closer to being measures of effects than of causes. Their method of construction relies on subjective perceptions, and is subject to error. But when used with caution, they have proved significant as explanatory variables in empirical studies of economic growth, and for observing changes in governance quality over time in specific countries. Corruption and regulatory burdens are major themes of the World Bank's research on governance in many countries (see World Bank Institute, website).

Empirical estimations of the level or growth of GDP on various measures of institutional quality confront many conceptual and econometric problems. Researchers have tackled the issue of reverse causation by using various instruments, such as the nationality of colonizers (Hall and Jones, 1999), mortality among colonizers (Acemoglu, Johnson and Robinson, 2001), and whether a colony had rich mineral resources or climatic and soil conditions conducive to plantation agriculture and a large or dense native population, or was sparsely populated and poor in the 1500•s (Engerman and Sokoloff, 2002; Acemoglu, Johnson and Robinson, 2002). The general idea is that in the former circumstances the European colonizers established institutions of slavery and inequality to facilitate the exploitation of labour on a large scale, whereas in the latter conditions, where the colonizers had to exert their own effort, their institutions provided the correct incentives and became conducive to longer-term economic success. The debate on the factual and econometric validity and the economic interpretation of these findings is fierce and continuing; Hoff (2003) surveys and discusses this literature in detail.

La Porta et al. (1998; 1999) contrast different legal traditions for protecting the rights of small shareholders. If such protection is poor, that will inhibit the flows of capital to its most efficient uses. They find that systems based on common law are better in this regard than those based on civil law. But Rajan and Zingales (2003) and Lamoreaux and Rosenthal (2005) argue that in practice there was little difference between the systems during critical periods of industrialization.

These debates are sure to continue, and this section will get out of date very quickly.

At the international level, formal governance works through bodies like the World Trade Organization. Their members are sovereign countries; therefore their procedures must be subject to self-enforcement in repeated interactions, whether through bilateral or multilateral sanctions. These institutions are therefore basically similar to the social networks discussed below. See Maggi (1999) and Bagwell and Staiger (2003) for detailed analyses.

Private institutions

The policing functions for property right protection supplied by the state are often supplemented by private security systems that serve specific clients and purposes – firms employ or hire security personnel, gated communities and neighbourhoods have private (hired or volunteer) patrols. These generally merely supplement the functions of the police for their specific context and work cooperatively with the police, but the two may clash if the private security system goes beyond its permissible functions.

Private institutions of contract enforcement similarly coexist with formal law, and become essential when the latter is weak or nonexistent. Explicit or implicit private contractual arrangements are also

important for assignment of property rights as a part of Coasean contracting for efficient outcomes. Therefore, analyses of private institutions often focus on the governance of contracts.

The basic problem of contract enforcement is control of opportunism. If one or both parties have to make transaction-specific investments, the other can attempt to secure a greater part of the benefit by reneging or demanding renegotiation. The prospect of this can jeopardize the potentially mutually beneficial deal in the first place. Williamson (1975; 1995) pioneered the analysis of this issue under the title of transaction cost economics.

Information constitutes a major source of advantage for private ordering over formal law. Enforcement of a contract in a court requires offering proof of misconduct by the other party in the dispute; the relevant information must be verifiable to outsiders. Therefore, formal contracts can stipulate actions by the parties conditional only on verifiable information. Other or more detailed information may be observable to the parties themselves, or can be inferred by specialist insiders to the industry, but cannot be verified to non-specialist judges or juries of the state's legal system, or can be verified only at excessive cost.

The informational advantage of private ordering may be offset by a disadvantage in enforcement. Informal arrangements must be made to overcome each participant's temptation to behave opportunistically at the others’ expense. Different methods of this kind underlie the various institutions of informal governance, and achieve different degrees of success. Some are able to exert coercion for immediate punishment of misbehaviour. Others create long-run costs, typically in the form of exclusion from future participation or worse future opportunities, to offset the short-run advantages of opportunism. This is the standard theory of self-enforcing cooperation in repeated Prisoner's Dilemmas. The following sections discuss some of these alternatives.

Private ordering with formal law in the background

Perhaps the most remarkable thing about formal legal institutions and mechanisms for the enforcement of commercial contracts is how rarely they are actually used. Business transactions often do have underlying formal contracts, but when disputes arise recourse to the law is often the last resort. Other private alternatives are tried first; these include bilateral negotiation, arbitration by industry experts, and so on. Filing a suit in a formal court of law often signals the end of a business relationship. Most actual practice in business contracting is therefore better characterized as ‘private ordering under the shadow of the law’ (Macaulay, 1963; Williamson, 1995, pp. 95–100, 121–2).

If one of the parties to an ongoing informal relationship behaves opportunistically, the most common alternative is to fall back on a formal contract based on verifiable contingencies alone. Suppose an outcome based on a tacit understanding of what each party should do in any one exchange (including good-faith negotiation to adapt to changing circumstances) yields both of them higher payoffs than does a formal contract. Consider the implicit arrangement where, if one party deviates from the agreed course of action to its own advantage and to the detriment of the other, their future exchanges will be governed by the formal contract. This yields a subgame-perfect (credible) equilibrium of the repeated game if each party's one-time gain from opportunism does not exceed the capitalized value of the future difference of payoffs between the tacit and the formal contracts. Williamson (2005, p. 2) expresses this well: ‘continuity can be put in jeopardy by defecting from the spirit of cooperation and reverting to the letter.’

When such relationship-based implicit contracting prevails, partial improvement in the formal system can worsen the outcome, due to a problem of the second-best. The partial improvement raises the payoffs the two parties could get from the fallback formal contract. This in turn reduces the future cost of a current deviation from the implicit contract or spirit of cooperation. It tightens the incentive-compatibility constraints, and therefore worsens what can be achieved by relational contracting (Baker, Gibbons and Murphy, 1994; Dixit, 2004, ch. 2).

Arbitration comes in two prominent forms. One is industry-specific, based on expert knowledge of insiders. More information is verifiable in such settings; therefore richer contracts specifying actions for more detailed contingencies become feasible. In many industries there is a large common-knowledge basis of custom and practice, which may even make it unnecessary to write down a contingent contract in great detail. Arbitration can also provide an opportunity for the parties to communicate and renegotiate adaptations to new circumstances. Formal legal systems often recognize these advantages of expert arbitration, and courts stand ready to enforce the decisions of arbitrators if the losing party tries to evade the sanctions. However, industry arbitrators often have severe sanctions at their own disposal; they can essentially drive the miscreant out of business, and even ostracize him or her from the social group of that business community. Examples of arbitration institutions include Bernstein's (1992) classic study of the diamond industry. For further discussion and modelling, see Dixit (2004, ch. 2) and Williamson (2005, p. 14).

The other prominent forums of arbitration deal with international contracts (Dezalay and Garth, 1996; Mattli, 2001). There are several of these, specializing in different legal traditions. They lack direct power to enforce their decisions, but are backed by treaties that ensure enforcement by national courts. These forums do not have industry-specific knowledge, their processes can be slow and costly, and their decisions can be somewhat arbitrary. But parties in transnational transactions may prefer them to either country's courts, suspecting that these will be biased in favour of their own nationals.

For-profit private institutions

If the state is unwilling to protect certain kinds of property or enforce certain kinds of contracts (for example in illegal activities), or is unable to do so (for example in weak and failing states), or is itself predatory, then private institutions can emerge to perform these functions for a profit. Organized crime often fills the niches uncovered by the state. Gambetta (1993), Bandiera (2003) and others argue that the Mafia emerged in just such a situation to fill the vacuum of protection in late 19th-century Sicily. Landowners began to hire guards of former feudal lords, and even the toughest among bandits, to protect their property. Gambetta describes how the Mafia's role expanded to providing contract enforcement in illegal or grey markets. Similarly, the Japanese Yakuza was instrumental in organizing markets at the end of the Second World War in August and September 1945 when the Japanese state had collapsed (Dower, 1999, pp. 140–8), and mafias grew in Russia after the collapse of the Soviet regime (Varese, 2001).

Gambetta (1993, p. 19) argues that this ‘business of protection’ is the core business of the Mafia. It may engage in other activities using in-house protection, but that is just downstream vertical integration – the opposite of upstream integration where an ordinary business firm has its in-house security department. A transaction-cost analysis of the internal organization of mafias, and of their vertical integration

decisions, may provide an interesting link between economic governance and corporate governance. Another dimension in which the protection business can expand is extortion; although private protectors may be welcome when state protection has collapsed, ‘protectors, once enlisted, invariably overstay their welcome’ (Gambetta, 1993, p. 198).

The Mafia can provide contract enforcement because, even though two traders may not have sufficiently frequent dealings with each other to achieve good outcomes in an ongoing bilateral relationship, each trader can be a regular customer of the enforcer. This converts multiple one-shot Prisoner's Dilemma games among the whole group of traders into several bilateral repeated games of each trader with the enforcer. The intermediary can provide information (keeping track of previous contract violations and informing a customer of the history of a potential trading partner) and/or actual punishment if a customer's trading partner violates their contract. The information role of the Mafioso is similar to that of credit rating agencies and Better Business Bureaus in the United States. Dixit (2004, ch. 4) constructs a model of such for-profit governance, and establishes the conditions for an equilibrium with for-profit private enforcement. These are lower bounds on the shares of the surplus that the customer and the Mafioso must have, so as to overcome the trader's temptation to cheat and the Mafioso’ temptation to double-cross the customer. Milgrom, North and Weingast (1990) have a related and complementary model of private judges at medieval European trade fairs. They specify the game of each trade, and investigation in the event of cheating, in greater detail, but do not examine the issue of the judges’ honesty.

Group enforcement through social networks and norms

Any institution of contract enforcement must solve three key problems: (*a*) detection of opportunistic deviations from the contractually stipulated behaviour, (*b*) preservation and dissemination of information about the histories of the participants’ behaviour, and (*c*) inflicting appropriate punishments to reduce future payoffs of any deviators. The first is often constrained by the available technology of monitoring, although institutions and regulations such as reporting requirements and auditing can improve the technology. The second and third problems are best resolved in bilateral ongoing relationships: each party has a natural incentive to detect and remember the other's cheating, and can punish the other by breaking off the relationship. However, governance is often needed in groups each of whose members interacts frequently with someone else in the group, but not necessarily bilaterally with the same person every time. Now remembering and transmitting information about your current partner's behaviour to others, and refusing a potentially beneficial deal because the counter-party has cheated someone else in the past, are privately costly activities and therefore require their own governance mechanisms.

Formal state institutions of governance can solve these problems by fiat; the legal system compels the whole group of traders to commit to good behaviour by subjecting themselves to detection and punishment if they cheat. A third-party supplier of information or enforcement serves similar functions. In the case of a Mafia enforcer, anyone who trades with a customer of the Mafioso subjects himself to the grim punishment if he cheats. In the case of a Better Business Bureau, a firm that joins the organization thereby gives hostage to its own good behaviour: if it misbehaves it will get a poor rating or blacklisting. Transactions vary in their characteristics; therefore we should expect the effectiveness of

such reputation mechanisms to vary also, and should not expect universal success from any one.

An institution of social networks and norms can solve the problems of information and punishment in a decentralized manner. Each participant can transmit information about his or her current trading partner's behaviour to others in the group to whom he or she is linked. And each can play his or her assigned part in punishment, typically by refusing to trade, if he or she gets matched with a potential partner who is known to have misbehaved in past dealings with others in the group. Incentives to transmit information or refuse potentially good trades can be established by a norm that regards refusal to do so as itself a punishable offence, as in Abreu's (1986) penal codes for repeated games; see Calvert (1995a; 1995b). Extrinsic incentives may even be unnecessary if people have sufficiently strong natural instincts to punish social cheaters, as found by Fehr and Gächter (2000).

Numerous empirical and case studies of governance based on social relations have been conducted; space constraints allow mention of only a few. Greif's (1993) historical analysis of Maghribi traders' system of communication and collective punishment is well known. So is Ostrom's (1990) synthesis of the evidence on common-pool resource management; she emphasizes the importance of local knowledge and communication, of appropriately designed (generally graduated) punishments, and of incentives for individuals to perform their assigned roles and actions in the system. Fafchamps (2004) studies and compares many different market institutions in Africa; his work highlights the importance of designing systems appropriate to the conditions of each country or group. Ensminger (1992) describes a similarly rich complex of arrangements for trade and employment relationships among the Orma tribe of Kenya, and examines how formal institutions of property right enforcement including title registration can interact dysfunctionally with traditional arrangements based on family and tribal connections. Johnson, McMillan and Woodruff (2002) present and analyse findings from survey research in former socialist economies. Of particular interest are the links between evolving formal and informal governance. Even without a backup of courts, trust in bilateral relationships can build quickly in response to good experiences. New or transient customers are more likely to be offered credit if courts work better, but the effectiveness of courts becomes largely irrelevant for the functioning of established relationships. Casella and Rauch (2002) study the role of ethnic networks in international trade.

Li (2003) points out a key difference between the costs of operating such a system and those of formal governance. A relation-based system of networks and norms has low fixed costs, but high and rising marginal costs. Trading on a small scale naturally starts among the most closely connected people who have sufficiently good communication and common understanding to sustain honesty. No fixed costs need be incurred to establish any formal rules or mechanisms of enforcement. But as trade expands, potential partners added at the margin are almost by definition less well-connected, making it harder to communicate information with them and to ensure their participation in any punishments. By contrast, formal or rule-based governance has high fixed costs of setting up the legal system and the information mechanism, but once these are incurred, marginal costs of dealing with strangers are low. Therefore, relation-based governance is better for small groups and rule-based governance better for large groups. Greif's (1994) comparison between the relation-based system of Maghribi traders and the formal institutions of Genoese traders supports this theory. Dixit (2004, ch. 3) constructs a formal model that compares relation-based and rule-based systems. This characterizes the maximum size of a self-enforcing group, and finds that, when the group exceeds this critical size, the maximum scope of sustainable honesty shrinks absolutely. The intuition is as follows. At the critical size, each trader is

indifferent between honesty and cheating when dealing with the most distant person. When more traders are added, this weakens the communication between the previously marginal person and other almost equally distant ones, tipping the balance toward cheating.

Kranton (1996) models individuals who can either choose bilateral long-lived self-enforcing trading relationships or search for one-time trading partners in an anonymous market with external enforcement. The market thus provides the outside opportunity in the repeated game of bilateral trade. If more people trade in the anonymous market, it becomes thicker and offers better prospects for successful search. Then parties in bilateral relationships have better outside opportunities, which makes it harder to sustain tacit cooperation there, further increasing the relative attraction of the market. Therefore the system can have multiple equilibria – no one uses the market because no one else uses it, or everyone uses the market because everyone else does – and can get locked into a Pareto-inferior equilibrium.

Evolution and transformation of governance institutions

A persistent theme in this survey has been that different governance institutions are optimal for different societies, for different kinds of economic activity, and at different times. Changes in underlying technologies of production, exchange and communication change the relative merits of different methods of governance. As the volume and scope of trade expand, formal institutions generally become superior to informal ones, but informal ones serve useful roles under the shadow of formal ones even in the most advanced economies and sectors. All this raises the question of whether we should expect institutions to adapt and evolve optimally.

Williamson's famous ‘discriminating alignment hypothesis’ says that transactions, with their different attributes, align with institutions, with their different costs and competencies; see his recent exposition (2005, p. 6). This gives ground for optimism for synergistic evolution of the need for governance and the institutions that supply it. Others are less sanguine. North (1990) and others argue that institutional change is subject to long delays due to resistance by organized interests favouring the status quo, problems of coordinating collective action to bring about a discrete change in equilibrium, and so on. Dixit (2004, pp. 79–85) discusses some of these problems for transition from relation-based to rule-based contract enforcement. Eggertson (2005) gives a dramatic example of how institutions restricting fishing and requiring costly mutual insurance persisted in Iceland for centuries after they had become obstacles to good economic performance.

I believe that a balanced approach is needed, recognizing the tendency towards synergistic alignment but also the obstacles to its realization. The net outcome will depend on many specifics of each context. Understanding and predicting the process requires a combination of approaches: case-based and analytical, inductive and deductive. Greif (2006) discusses, develops and applies such methodologies using historical studies of trade in medieval Europe.

See Also

- cooperation
- corporations

- growth and institutions
- hold-up problem
- law, economic analysis of
- law, public enforcement of
- market institutions
- property rights
- social norms
- spontaneous order
- transition and institutions

I thank Tore Ellingsen, Diego Gambetta, Karla Hoff, Eva Meyersson-Milgrom, Dani Rodrik, Oliver Williamson, and the editors for comments on previous drafts, and the National Science Foundation for research support.

Bibliography

Abreu, D. 1986. Extremal equilibria of oligopolistic supergames. *Journal of Economic Theory* 39, 191–225.

Acemoglu, D. 2003. Why not a political Coase Theorem? Social conflict, commitment and politics. *Journal of Comparative Economics* 31, 620–52.

Acemoglu, D., Johnson, S. and Robinson, J. 2001. The colonial origins of comparative development: an empirical investigation. *American Economic Review* 91, 1369–401.

Acemoglu, D., Johnson, S. and Robinson, J. 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 117, 1231–94.

Acemoglu, D. and Robinson, J. 2005. *Economic Origins of Dictatorship and Democracy*. New York: Cambridge University Press.

Bagwell, K. and Staiger, R. 2003. *The Economics of the World Trading System*. Cambridge, MA: MIT Press.

Baker, G., Gibbons, R. and Murphy, K. 1994. Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics* 109, 1125–56.

Bandiera, O. 2003. Land reform, the market for protection and the origins of the Sicilian Mafia: theory and evidence. *Journal of Law, Economics and Organization* 19, 218–44.

Barro, R. 1999. *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge, MA: MIT Press.

Bernstein, L. 1992. Opting out of the legal system: extralegal contractual relations in the diamond industry. *Journal of Legal Studies* 21, 115–57.

Calvert, R. 1995a. The rational choice theory of social institutions: cooperation, communication, and coordination. In *Modern Political Economy: Old Topics, New Directions*, ed. J. Banks and E. Hanushek. Cambridge: Cambridge University Press.

Calvert, R. 1995b. Rational actors, equilibrium, and social institutions. In *Explaining Social Institutions*, ed. J. Knight and I. Sened. Ann Arbor: University of Michigan Press.

Casella, A. and Rauch, J. 2002. Anonymous market and group ties in international trade. *Journal of International Economics* 58, 19–47.

Coase, R. 1937. The nature of the firm. *Economica* 4, 386–406.

De Soto, H. 2000. *Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. New York: Basic Books.

Dezalay, Y. and Garth, B. 1996. *Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Order*. Chicago, IL and London: University of Chicago Press.

Dixit, A. 2004. *Lawlessness and Economics: Alternative Modes of Governance*. Princeton, NJ: Princeton University Press.

Dower, J. 1999. *Embracing Defeat: Japan in the Wake of World War II*. New York: W.W. Norton.

Eggertson, T. 2005. *Imperfect Institutions: Possibilities and Limits of Reform*. Ann Arbor: University of Michigan Press.

Engerman, S. and Sokoloff, K. 2002. Factor endowments, inequality, and paths of development among New World economies. *Economia* 3, 41–109.

Ensminger, J. 1992. *Making a Market: The Institutional Transformation of an African Society*. New York: Cambridge University Press.

Fafchamps, M. 2004. *Market Institutions in Sub-Saharan Africa: Theory and Evidence*. Cambridge, MA: MIT Press.

Fehr, E. and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90, 980–94.

Gambetta, D. 1993. *The Sicilian Mafia: The Business of Private Protection*. Cambridge, MA: Harvard University Press.

Giavazzi, F. and Tabellini, G. 2005. Economic and political liberalizations. *Journal of Monetary Economics* 57, 1297–330.

Glaeser, E., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. 2004. Do institutions cause growth? *Journal of Economic Growth* 9, 271–303.

Greif, A. 1993. Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. *American Economic Review* 83, 525–48.

Greif, A. 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy* 102, 912–50.

Greif, A. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. New York: Cambridge University Press.

Greif, A., Milgrom, P. and Weingast, B. 1994. Coordination, commitment, and enforcement: the case of the merchant guild. *Journal of Political Economy* 102, 745–76.

Hall, R. and Jones, C. 1999. Why do some countries produce so much more output than others? *Quarterly Journal of Economics* 114, 83–116.

Hoff, K. 2003. Paths of institutional development: a view from economic history. *World Bank Research Observer* 18, 205–26.

Johnson, S., McMillan, J. and Woodruff, C. 2002. Courts and relational contracts. *Journal of Law, Economics and Organization* 18, 221–77.

Kaufman, D., Kraay, A. and Mastruzzi, M. 2005. Governance matters IV: Updated governance indicators 1996–2004. Washington, DC: World Bank research paper. Online. Available at http://www.worldbank.org/wbi/governance/pubs/govmatters4.html, accessed 20 April 2006.

Keefer, P. 2004. What does political economy tell us about economic development – and vice versa? *Annual Review of Political Science* 7, 247–72.

Kranton, R. 1996. Reciprocal exchange: a self–sustaining system. *American Economic Review* 86, 830–

51.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R. 1998. Law and finance. *Journal of Political Economy* 106, 1113–55.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R. 1999. The quality of government. *Journal of Law, Economics and Organization* 15, 222–79.

Lamoreaux, N. and Rosenthal, J.-L. 2005. Legal regime and contractual flexibility: a comparison of business's organizational choices in France and the United States during the era of industrialization. *American Law and Economics Review* 7, 28–61.

Li, J.S. 2003. Relation-based versus rule-based governance: an explanation of the East Asian miracle and Asian crisis. *Review of International Economics* 11, 651–73.

Macaulay, S. 1963. Non-contractual relationships in business: a preliminary study. *American Sociological Review* 28, 55–70.

Maggi, G. 1999. The role of multilateral institutions in international trade cooperation. *American Economic Review* 89, 190–214.

Mattli, W. 2001. Private justice in a global economy: from litigation to arbitration. *International Organization* 55, 919–47.

Milgrom, P., North, D. and Weingast, B. 1990. The role of institutions in the revival of trade: the law merchant, private judges, and the Champagne fairs. *Economics and Politics* 2, 1–23.

North, D. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK, and New York: Cambridge University Press.

Persson, T. 2005. Forms of democracy, policy, and economic development. Working Paper No. 11171. Cambridge, MA: NBER.

Rajan, R. and Zingales, L. 2003. The great reversals: the politics of financial development in the twentieth century. *Journal of Financial Economics* 69, 5–50.

Varese, F. 2001. *The Russian Mafia: Private Protection in a New Market Economy*. Oxford: Oxford University Press.

Williamson, O. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

Williamson, O. 1995. *The Mechanisms of Governance*. New York: Oxford University Press.

Williamson, O. 2005. The economics of governance. *American Economic Review* 95, 1–18.

World Bank Institute. Governance and Anti-Corruption. Online. Available at http://www.worldbank.org/wbi/governance, accessed 20 April 2006.

How to cite this article

Dixit, Avinash K. "economic governance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000260> doi:10.1057/9780230226203.0431

The New Palgrave Dictionary of Economics Online

economic growth in the very long run

Oded Galor
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The evolution of economies during the major portion of human history was marked by Malthusian stagnation. The transition from an epoch of stagnation to a state of sustained economic growth has shaped the contemporary world economy and has led to the great divergence in income per capita across the globe in the past two centuries. This article examines the process of development over the course of human history in light of recent advances in unified growth theory.

Keywords

capital accumulation; China; colonialism; convergence clubs; demographic transition; economic growth in the very long run; education; endogenous growth; exogenous growth; fertility; globalization; great divergence; growth take-offs; human capital; income growth; Industrial Revolution; industrialization; inequality; land; land productivity; literacy; Malthusian stagnation; marriage; microfoundations; population growth; post-Malthusian regime; sustained growth regime; technological progress; unified growth theory

Article

The evolution of economies during the major portion of human history was marked by Malthusian stagnation. Technological progress and population growth were minuscule by modern standards, and the average growth rates of income per capita in various regions of the world were even slower due to the offsetting effect of population growth on the expansion of resources per capita.

In the past two centuries the pace of technological progress increased significantly in association with the process of industrialization. Various regions of the world departed from the Malthusian trap and experienced a considerable rise in the growth rates of income per capita and population. Unlike episodes of technological progress in the pre-Industrial Revolution era that failed to generate sustained economic growth, the increasing role of human capital in the production process in the second phase of industrialization ultimately prompted a demographic transition, liberating the gains in productivity from the counterbalancing effects of population growth. The decline in the growth rate of population and the enhancement of human capital formation and technological progress paved the way for the emergence of the modern state of sustained economic growth. Variations in the timing of the transitions from a Malthusian epoch to a state of sustained economic growth across countries lead to a considerable rise in the ratio of GDP per capita between the richest and the poorest regions of the world from 3:1 in 1820 to 18:1 in 2000 (see Figure 1).

Figure 1
The evolution of regional income per capita, 1–2000. *Source*: Maddison (2001).

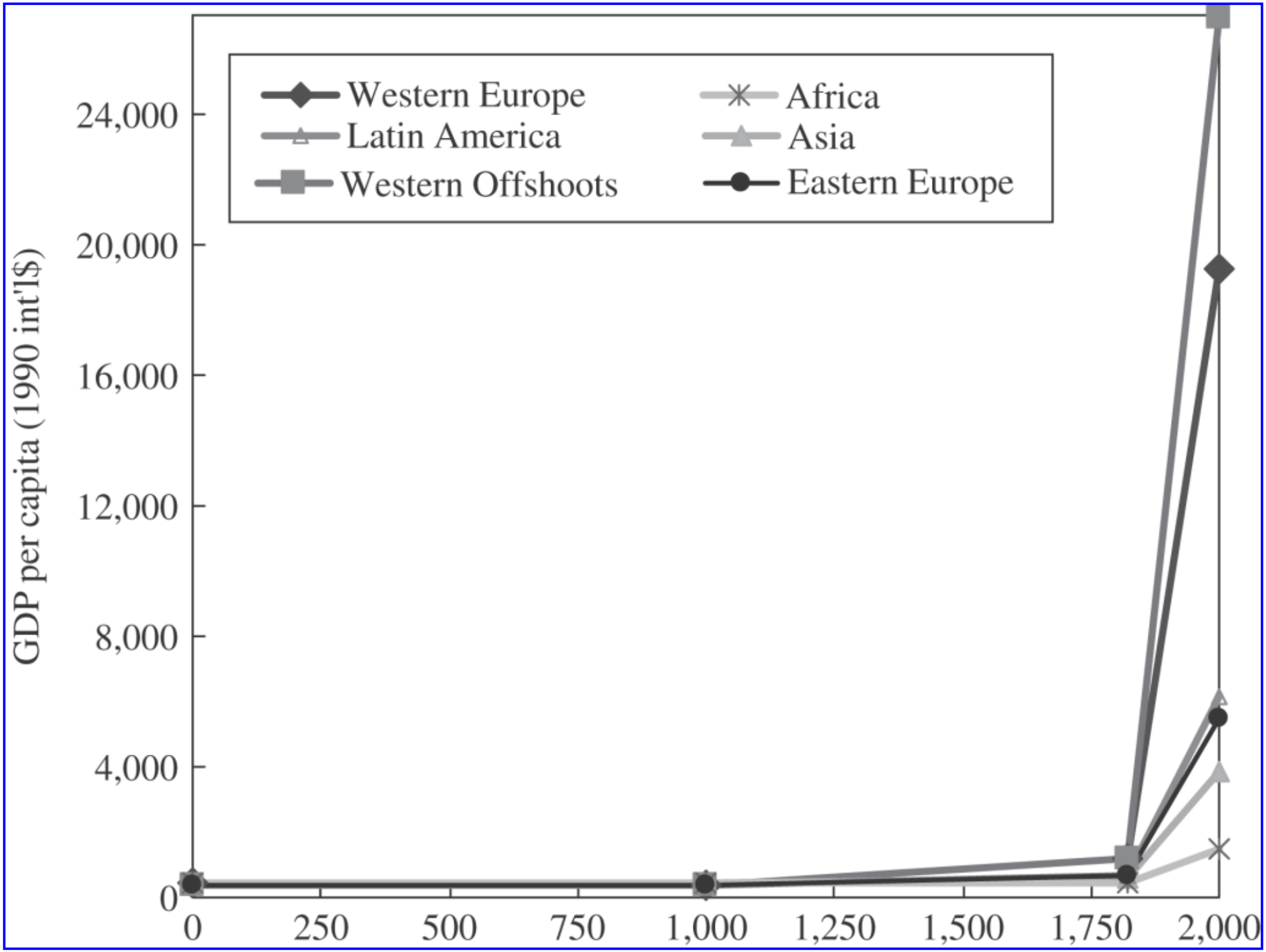

The transition from stagnation to growth and the associated phenomenon of the great divergence have been the subject of intensive research in the growth literature in recent years

(Galor and Weil, 1999; 2000; Galor and Moav, 2002; Lucas, 2002; Hansen and Prescott, 2002; Jones, 2001; Hazan and Berdugo, 2002; Doepke, 2004; Lagerlof, 2003; 2006; Galor and Mountford, 2003; 2006). The inconsistency of exogenous and endogenous growth models with some of the most fundamental features of the process of development has led to a search for a unified theory that would unveil the underlying microfoundations of the growth process in its entirety, and would capture in a single framework the epoch of Malthusian stagnation that characterized most of human history, the contemporary era of modern economic growth, and the driving forces that triggered the recent transition between these regimes.

The advance of unified growth theory was fuelled by the conviction that the understanding of the contemporary growth process would be fragile and incomplete unless growth theory were based on proper microfoundations that reflect the various qualitative aspects of the growth process and their central driving forces. Moreover, it has become apparent that a comprehensive understanding of the hurdles faced by less developed economies in reaching a state of sustained economic growth would remain obscure unless the factors that prompted the transition of the currently developed economies into a state of sustained economic growth could be identified and modified to account for the differences in the growth structure of less developed economies in an interdependent world.

Unified growth theory explores the fundamental factors that generated the remarkable escape from the Malthusian epoch and their significance in understanding the contemporary growth process of developed and less developed economies. Moreover, it sheds light on the perplexing phenomenon of the great divergence in income per capita across regions of the world in the past two centuries. It suggests that the transition from stagnation to growth is an inevitable outcome of the process of development. The inherent Malthusian interaction between the level of technology and the size and the composition of the population accelerated the pace of technological progress and ultimately raised the importance of human capital in the production process. The rise in the demand for human capital in the second phase of industrialization and its impact on the formation of human capital as well as on the onset of the demographic transition brought about significant technological advances along with a reduction in fertility rates and population growth, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita, and paving the way for the emergence of sustained economic growth.

Differences in the timing of the take-off from stagnation to growth across countries (for example, England's earlier industrialization in comparison with China) contributed significantly to the great divergence and to the emergence of convergence clubs. These variations reflect initial differences in geographical factors and historical accidents and their manifestation in diversity in institutional, demographic, and cultural factors, trade patterns, colonial status, and public policy. In particular, once a technologically driven demand for human capital emerged in the second phase of industrialization, the prevalence of human capital-promoting institutions determined the extensiveness of human capital formation, the timing of the demographic transition, and the pace of the transition from stagnation to growth. Thus, unified growth theory provides the natural framework of analysis in which variations in the economic performance across countries and regions could be examined based on the effect of variations in educational, institutional, geographical, and cultural factors on the pace of the transition from stagnation to growth.

The process of development

The process of economic development has been characterized by of three fundamental regimes: the Malthusian epoch, the post-Malthusian regime, and the sustained growth regime.

The Malthusian epoch

During the Malthusian epoch that characterized most of human history, humans were subjected to a persistent struggle for existence. Resources generated by technological progress and land expansion were channelled primarily towards an increase in the size of the population, with a minor long-run effect on income per capita. Improvements in the technological environment or in the availability of land generated temporary gains in income per capita, leading eventually to a larger but not richer population. Technologically superior countries ultimately had denser populations but their standard of living did not reflect the degree of their technological advancement.

During the Malthusian epoch the average growth rate of output per capita was negligible and the standard of living did not differ greatly across countries. The average level of income per capita in the world during the first millennium fluctuated around $450 per year (in 1990 international dollars) and the average growth rate of output per capita was nearly zero (Maddison, 2001). This state of Malthusian stagnation persisted until the end of the 18th century. In the years 1000–1820, the average level of income per capita in the world economy was below $670 per year, and the average growth rate of the world income per capita was minuscule, creeping at a rate of about 0.05 per cent per year. Nevertheless, income per capita fluctuated significantly within regions, deviating from their sluggish long-run trend over decades and sometimes centuries.

Population growth over this era followed the Malthusian pattern as well. The gradual increase in income per capita during the Malthusian epoch was associated with a monotonic increase in the average rate of growth of world population. The slow pace of resource expansion in the first millennium was reflected in a modest increase in the population of the world from 231 million people in 1 ce to 268 million in 1000 ce: a minuscule average growth rate of 0.02 per cent per year. The more rapid (but still very slow) expansion of resources in the period 1000–1500 permitted the world population to increase by 63 per cent, from 268 million in 1000 to 438 million in 1500; a slow 0.1 per cent average growth rate per year. Resource expansion over the period 1500–1820 had a more significant impact on the world population, which grew 138 per cent from 438 million in 1500 to 1,041 million in 1820: an average pace of 0.27 per cent per year.

Variations in population density across countries during the Malthusian epoch reflected primarily cross-country differences in technology and land productivity. Due to the positive adjustment of the population to an increase in income per capita, differences in technology or in land productivity across countries resulted in variations in population density rather than in the standard of living. For instance, China's technological advancement in the period 1500–1820 permitted its share of world population to increase from 23.5 per cent to 36.6 per cent, while its income per capita in the beginning and the end of this time interval remained approximately $600 per year.

The post-Malthusian regime

During the post-Malthusian regime, the pace of technological progress markedly increased in association with the process of industrialization, triggering a take-off from the Malthusian trap. The growth rate of income per capita increased significantly but the positive Malthusian effect of income per capita on population growth was still maintained, generating a sizeable increase in population growth that offset some of the potential gains in income per capita.

The take-off of developed regions from the Malthusian regime occurred at the beginning of the 19th century and was associated with the Industrial Revolution, whereas the take-off of less developed regions occurred towards the beginning of the 20th century and was delayed in some countries well into the 20th century. During the post-Malthusian regime the average growth rate of output per capita increased significantly and the standard of living began to differ considerably across countries. The average growth rate of output per capita in the world soared from 0.05 per cent per year during the period 1500–1820 to 0.53 per cent per year in the years 1820–70, and 1.3 per cent per year during the period 1870–1913. The timing of the take-off and its magnitude differed across regions. The take-off from the Malthusian epoch and the transition to the post-Malthusian regime occurred in western Europe, the Western offshoots (that is, the United States, Canada, Australia and New Zealand), and eastern Europe at the beginning of the 19th century, whereas in Latin America, Asia and Africa it occurred towards the beginning of the 20th century.

The rapid increase in income per capita in the post-Malthusian regime was channelled partly towards an increase in the size of the population. During this period, the Malthusian mechanism linking higher income to higher population growth continued to function. However, the effect of higher population on the dilution of resources per capita was counteracted by accelerated technological progress and capital accumulation, allowing income per capita to rise despite the offsetting effects of population growth.

The western European take-off along with that of the Western offshoots brought about a sharp increase in population growth in these regions and consequently a modest rise in population growth in the world as a whole. The subsequent take-off of less developed regions, and the associated increase in their rates of population growth, brought about a significant rise in population growth in the world. The rate of population growth in the world increased from an average rate of 0.27 per cent per year in the period 1500–1820 to 0.4 per cent per year in the years 1820–70, and to 0.8 per cent per year in the time interval 1870–1913. Despite the decline in population growth in western Europe and the Western offshoots towards the end of the 19th century and the beginning of the 20th century, the delayed take-off of less developed regions, and the significant increase in their income per capita prior to their demographic transitions, generated a further increase in the rate of population growth in the world to 0.93 per cent per year in the period 1913–50, and 1.92 per cent per year in the period 1950–73. Ultimately, the onset of the demographic transition in less developed economies during the second half of the 20th century reduced population growth rates to 1.66 per cent per year in the 1973–98 period (Maddison, 2001).

It appears that the significant rise in income per capita in the post-Malthusian regime increased the desired number of surviving offspring and thus, despite the decline in mortality rates, fertility increased significantly so as to enable households to reach this higher desired level of surviving offspring. Fertility was controlled during this period, despite the absence of modern contraceptive methods, partly via adjustment in marriage rates. Increased fertility was achieved by earlier female age of marriage, and a decline in fertility by a delay in the marriage age.

The take-off in the developed regions was accompanied by a rapid process of industrialization. Per-capita level of industrialization increased significantly in the United Kingdom, rising 50 per cent over the 1750–1800 period, quadrupling in the years 1800–60, and nearly doubling in the time period 1860–1913. Similarly, per capita level of industrialization accelerated in the United States, doubling in the 1750–1800 as well as 1800–60 periods, and increasing sixfold in the years 1860–1913. A similar pattern was experienced in Germany, France, Sweden, Switzerland, Belgium and Canada. The take-off of less developed economies in the 20th century was associated with increased industrialization as well. However, during the 19th century these economies experienced a decline in per capita industrialization, reflecting the adverse effect of the sizeable increase in population on the level of industrial production per capita as well as the forces of globalization and colonialism, which induced less developed economies to specialize in the production of raw materials (Galor and Mountford, 2003; 2006).

The acceleration in technological progress during the post-Malthusian regime and the associated increase in income per capita stimulated the accumulation of human capital in the form of literacy rates, schooling, and health. The increase in the investment in human capital was induced by the rise in income per capita, as well as by qualitative changes in the economic environment that increased the demand for human capital and induced households to invest in the education of their offspring.

In the first phase of the Industrial Revolution, human capital had a limited role in the production process. Education was motivated by a variety of reasons, such as religion, enlightenment, social control, moral conformity, socio-political stability, social and national cohesion, and military efficiency. The extensiveness of public education was therefore not necessarily correlated with industrial development, and it differed across countries due to political, cultural, social, historical and institutional factors. In the second phase of the Industrial Revolution, however, the demand for education increased, reflecting the increasing skill requirements in the process of industrialization. The economic interests of

capitalists were a significant driving force behind the implementation of educational reforms (Galor and Moav, 2006). The process of industrialization has been characterized by a gradual increase in the relative importance of human capital in less developed economies as well and educational attainment increased significantly across all less developed regions in the post-Malthusian regime.

The sustained growth regime

The acceleration in the rate of technological progress in the second phase of industrialization, and its interaction with human capital formation, triggered a demographic transition, paving the way to a transition to an era of sustained economic growth. In the post demographic-transition period, the rise in aggregate income due to technological progress and factors accumulation was no longer counterbalanced by population growth, permitting sustained growth in income per capita in regions that experienced sustained technological progress and accumulation of physical and human capital.

The transition of the developed regions of western Europe and the Western offshoots to the state of sustained economic growth occurred towards the end of the 19th century, and their income per capita in the 20th century has advanced at a stable rate of about two per cent per year. The transition of some less developed countries in Asia and Latin America occurred towards the end of the 20th century. Africa, in contrast, is still struggling to make this transition.

The transition to a state of sustained economic growth was characterized by a gradual increase in the importance of the accumulation of human capital relative to physical capital as well as with a sharp decline in fertility rates. In the first phase of the Industrial Revolution (1760–1830), capital accumulation as a fraction of GDP significantly increased whereas literacy rates remained largely unchanged. Skills and literacy requirements were minimal, the state devoted virtually no resources to raise the level of literacy of the masses, and workers developed skills primarily through on-the-job training (Green, 1990; Mokyr, 1993). Consequently, literacy rates did not increase during the period 1750–1830 (Sanderson, 1995).

In the second phase of the Industrial Revolution, however, the pace of capital accumulation subsided, skills became necessary for production and the education of the labour force markedly increased. The investment ratio in the UK, which increased from six per cent in 1760 to 11.7 per cent in 1831, remained at around 11 per cent on average in the years 1856–1913 (Crafts, 1985). In contrast, the average years of schooling of males in the labour force that did not change significantly until the 1830s tripled by the beginning of the 20th century. The drastic rise in the level of income per capita in England as of 1865 was associated with an increase in school enrolment of ten-year-old children from 40 per cent in 1870 to 100 per cent in 1900. Moreover, total fertility rate in England sharply declined over this period from about five in 1875, to nearly two in 1925.

The demographic transition swept the world in the course of the 20th century. The unprecedented increase in population growth during the post-Malthusian regime was reversed and the demographic transition brought about a significant reduction in fertility rates and population growth in various regions of the world, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita. The demographic transition enhanced the growth process via three channels: (*a*) reductions in the dilution of the stocks of capital and natural resources, (*b*) enhancements in human capital formation, and (*c*) changes in the age distribution of the population, temporarily increasing the size of the labour force relative to the population as a whole.

The timing of the demographic transition differed significantly across regions. The reduction in population growth occurred in Western Europe, the Western offshoots, and eastern Europe towards the end of the 19th century and in the beginning of the 20th century, whereas Latin America and Asia experienced a decline in the rate of population growth only in the last decades of the 20th century. Africa's population growth, in contrast, has been rising steadily.

The process of industrialization was characterized by a gradual increase in the relative importance of human capital in the production process. The acceleration in the rate of technological progress gradually increased the demand for human capital, inducing individuals to invest in education, and stimulating further technological advancement. Moreover, in developed as well as less developed regions, the onset of the process of human capital accumulation preceded the onset of the demographic transition, suggesting that the rise in the demand for human capital in the process of industrialization and the subsequent accumulation of human capital played a significant role in the demographic transition and the shift to a state of sustained economic growth.

Notably, the reversal of the Malthusian relation between income and population growth during the demographic transition corresponded to an increase in the level of resources invested in each child. For example, literacy rate among men in England was stable at around 65 per cent in the first phase of the Industrial Revolution and increased significantly during the second phase, reaching nearly 100 per cent at the end of the 19th century. In addition, the proportion of children aged 5 to 14 in primary schools increased from 11 per cent in 1855 to 74 per cent in 1900. A similar pattern is observed in other European societies (Flora, Kraus and Pfenning, 1983).

The process of industrialization was characterized by a gradual increase in the relative importance of human capital in less developed economies as well. Educational attainment increased significantly across all less developed regions. Moreover, in line with the pattern that emerged among developed economies in the 19th century, the increase in educational attainment preceded or occurred simultaneously with the decline in total fertility rates.

The great divergence

The differential timing of the take-off from stagnation to growth across countries and the corresponding variations in the timing of the demographic transition led to a great divergence in income per capita as well as population growth. Inequality in the world economy was negligible till the 19th century. The ratio of GDP per capita between the richest region and the poorest region in the world was only 1.1:1 in 1000, 2:1 in 1500 and 3:1 in 1820. In the past two centuries, however, the ratio of GDP per capita between the richest group (Western offshoots) and the poorest region (Africa) has widened considerably from a modest 3:1 ratio in 1820, to 5:1 ratio in 1870, 9:1 ratio in 1913, 15:1 ratio in 1950, and 18:1 ratio in 2001.

An equally momentous transformation occurred in the distribution of world population across regions. The earlier take-off of western European countries increased the amount of resources that could be devoted for the increase in family size, permitting a 16 per cent increase in the share of their population in the world from 12.8 per cent in 1820 to 14.8 per cent in 1870. However, the early onset in the western European demographic transition and the long delay in the demographic transition of less developed regions, well into the second half of the 20th century, led to a decline in the share of western European population in the world, from 14.8 per cent in 1870 to 6.6 per cent in 1998. In contrast, the prolongation of the post-Malthusian period among less developed regions, in association with the delay in their demographic transition well into the second half of 20th century, channelled their increased resources towards a significant increase in their population. Africa's share of world population increased from seven per cent in 1913 to 12.9 per cent in 1998, Asia's share of world population increased from 51.7 per cent in 1913 to 57.4 per cent in 1998, and Latin American countries increased their share in world population from two per cent in 1820 to 8.6 per cent in 1998.

Unified growth theory

Galor and Weil (2000) advanced a unified growth theory that captures the three regimes that have characterized the process of development as well as the fundamental driving forces that generated the transition from an epoch of Malthusian stagnation to a state of sustained economic growth. The theory replicates the observed time paths of population, income per capita, and human capital, generating: (*a*) the Malthusian oscillations in population and output per capita during the Malthusian epoch, (*b*) an endogenous take-off from Malthusian stagnation that is associated with an acceleration in technological progress and is accompanied initially by a rapid increase in population growth, and (*c*) a rise in the demand for human capital, followed by a demographic transition and sustained economic growth. These qualitative patterns are confirmed in the calibration of the theory by Lagerlof (2006).

The theory proposes that in early stages of development economies were in the proximity of a stable Malthusian equilibrium. Technology advanced rather slowly, and generated proportional increases in output and population. The inherent positive interaction between population and technology in this epoch, however, gradually increased the pace of technological progress, and due to the delayed adjustment of population, output per capita advanced at a minuscule rate. The slow pace of technological progress in the Malthusian epoch provided a limited scope for human capital in the production process and parents, therefore, had no incentive to reallocate resources towards human capital formation of their offspring.

The Malthusian interaction between technology and population accelerated the pace of technological progress and permitted a take-off to the post-Malthusian regime. The expansion of resources was partially counterbalanced by the enlargement of population, and the economy was characterized by rapid growth rates of income per capita and population. The acceleration in technological progress eventually increased the demand for human capital, generating two opposing effects on population growth. On the one hand, it eased households' budget constraints, allowing the allocation of more resources for raising children. On the other hand, it induced a reallocation of resources towards child quality. In the post-Malthusian regime, due to the modest demand for human capital, the first effect dominated, and the rise in real income permitted households to increase the number as well the quality of their children.

As investment in human capital took place, the Malthusian steady-state equilibrium vanished and the economy started to be attracted by the gravitational forces of the modern growth regime. The interaction between investment in human capital and technological progress generated a virtuous circle: human capital generated faster technological progress, which in turn further raised the demand for human capital, inducing further investment in child quality, and eventually triggering the onset of the demographic transition and the emergence of a state of sustained economic growth.

The theory suggests that the transition from stagnation to growth is an inevitable outcome of the process of development. The inherent Malthusian interaction between the level of technology and the size of the population accelerated the pace of technological progress, and ultimately raised the importance of human capital in the production process. The rise in the demand for human capital in the second phase of the Industrial Revolution and its impact on the formation of human capital as well as on the onset of the demographic transition brought about significant technological advancements along with a reduction in fertility rates and population growth, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita, and paving the way for the emergence of sustained economic growth. Quantitative analysis of unified growth theories (Doepke, 2004); Lagerlof, 2006) indeed suggest that the rise in the demand for human capital was a significant force behind the demographic transition and the emergence of a state of sustained economic growth.

Variations in the timing of the transition from stagnation to growth and thus in economic performance across countries reflect initial differences in geographical factors and historical accidents and their manifestation in diversity in institutional, demographic, and cultural factors, trade patterns, colonial status, and public policy. In particular, once a technologically driven demand for human capital emerged in the second phase of industrialization, the prevalence of human capital-promoting institutions determined the extensiveness of human

capital formation, the timing of the demographic transition, and the pace of the transition from stagnation to growth.

The theory proposes that the growth process is characterized by stages of development and it evolves nonlinearly. Technological leaders experienced a monotonic increase in the growth rates of their income per capita. Their growth was rather slow in early stages of development, increased rapidly during the take-off from the Malthusian epoch, and continued to rise, often stabilizing at higher levels. In contrast, technological followers that made the transition to sustained economic growth experienced a non-monotonic increase in the growth rates of their income per capita. Their growth rate was rather slow in early stages of development, but increased rapidly in the early stages of the take-off from the Malthusian epoch, boosted by the adoption of technologies from the existing technological frontier. However, once these economies reached the technological frontier, their growth rates dropped to the level of the technological leaders. Hence, consistently with contemporary evidence about the existence of multiple growth regimes (Durlauf and Quah, 1999), the differential timing of the take-off from stagnation to growth across economies generated convergence clubs characterized by a group of poor countries in the vicinity of the Malthusian equilibrium, a group of rich countries in the vicinity of the sustained growth equilibrium, and a third group in the transition from one club to another.

See Also

- growth take-offs
- human capital, fertility and growth

Bibliography

Crafts, N.F.R. 1985. *British Economic Growth during the Industrial Revolution*. Oxford: Oxford University Press.

Doepke, M. 2004. Accounting for fertility decline during the transition to growth. *Journal of Economic Growth* 9, 347–83.

Durlauf, S.N. and Quah, D. 1999. The new empirics of economic growth. In *Handbook of Macroeconomics*, ed. J. B. Taylor and M. Woodford. Amsterdam: North-Holland.

Flora, P., Kraus, F. and Pfenning, W. 1983. *State, Economy and Society in Western Europe 1815–1975*. Chicago: St. James Press.

Galor, O. 2005. From stagnation to growth: unified growth theory. In *Handbook of Economic Growth*, ed. P. Aghion and S.N. Duraluf. Amsterdam: North-Holland.

Galor, O. and Moav, O. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117, 1133–92.

Galor, O. and Moav, O. 2006. Das human kapital: a theory of the demise of the class structure. *Review of Economic Studies* 73, 85–117.

Galor, O. and Mountford, A. 2003. Trading population for productivity. Working paper, Brown University.

Galor, O. and Mountford, A. 2006. Trade and the great divergence: the family connection. *American Economic Review* 96, 299–303.

Galor, O. and Weil, D.N. 1999. From Malthusian stagnation to modern growth. *American Economic Review* 89, 150–4.

Galor, O. and Weil, D.N. 2000. Population, technology and growth: from the Malthusian regime to the demographic transition and beyond. *American Economic Review* 110, 806–28.

Green, A. 1990. *Education and State Formation*. New York: St. Martin's Press.

Hansen, G. and Prescott, E. 2002. Malthus to Solow. *American Economic Review* 92, 1205–17.

Hazan, M. and Berdugo, B. 2002. Child labor, fertility and economic growth. *Economic Journal* 112, 810–28.

Jones, C.I. 2001. Was an industrial revolution inevitable? Economic growth over the very long run. *Advances in Macroeconomics* 1, 1–43.

Lagerlof, N. 2003. From Malthus to modern growth: the three regimes revisited. *International Economic Review* 44, 755–77.

Lagerlof, N. 2006. The Galor–Weil model revisited: a quantitative exploration. *Review of Economic Dynamics* 9, 116–42.

Lucas, R.E. 2002. *The Industrial Revolution: Past and Future*. Cambridge, MA: Harvard University Press.

Maddison, A. 2001. *The World Economy: A Millennia Perspective*. Paris: OECD.

Mokyr, J. 1993. The new economic history and the industrial revolution. In *The British Industrial Revolution: An Economic Perspective*, ed. J. Mokyr. Boulder, CO: Westview Press.

Sanderson, M. 1995. *Education, Economic Change and Society in England 1780–1870*. Cambridge: Cambridge University Press.

How to cite this article

Galor, Oded. "economic growth in the very long run." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000283> doi:10.1057/9780230226203.0434

The New Palgrave Dictionary of Economics Online

economic growth nonlinearities

Chih Ming Tan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Nonlinearities in growth have important implications for cross-country income inequality. In particular, they imply that countries may spend long periods of time in a low-growth poverty trap. However, finding evidence of such nonlinearities in the data and accounting for their emergence pose unique challenges to researchers.

Keywords

balanced growth; conditional convergence; convergence; convergence clubs; demographic transition; diffusion of technology; economic development; economic growth nonlinearities; increasing returns; industrialization; market size; multiple-growth regimes; neoclassical growth theory; neoclassical production function; nonconvexity; poverty traps; spillover effects; stages of theory of growth; strategic complementarities; structural change; take-off; technological progress

Article

Nonlinear growth models

Nonlinear growth models are characterized by a country's subsequent performance being critically dependent upon its initial conditions. In particular, these models tend to imply that countries which have unfavourable initial conditions may either experience substantial periods of time in low-growth/low-income poverty traps or be altogether caught in one. In some cases, it has been explicitly suggested that active (exogenous) policy interventions may be necessary in order to kick-start a country into a more favourable equilibrium. Nonlinear growth models can be broadly classified into two classes: structural change (or ‘stages of development’) models, and models that emphasize endogenous technological development and cross-country interactions in terms of technological diffusion.
Structural change models focus on the (internal) transformations of an economy as it transits through

critical phases or ‘stages’ (see Lewis, 1956; Rostow, 1960) leading to industrialization. The aim of this work is to clarify the conditions for such transitions to occur. Early work in the economic development literature (see Rosenstein-Rodan, 1943; Nurkse, 1953; Scitovsky, 1954; Fleming, 1955; formalized by Murphy, Shleifer, and Vishy, 1989) emphasized the importance of increasing returns and the size of the market in industrialization. The key idea behind this view is that countries could be locked in a no-industrialization trap because of the small size of the market for each sector of the economy. No single sector can achieve growth on its own. However, the growth of one sector results in the enlargement of markets for other sectors. The enlargement of markets then encourages investment and growth in the corresponding sectors. These spillover effects and strategic complementarities imply that a ‘big push’ – that is, coordinated investments (or ‘balanced growth’) across sectors – may be sufficient to push the economy out of the trap and into a ‘take-off’ towards industrialization. Other models are explicitly informed by the analysis of historical data (see Maddison, 2004), and emphasize the importance of explaining simultaneously both historical patterns of other state variables associated with growth and growth itself. An important recent work that models the demographic transition in growth take-offs is Galor and Weil (2000). Because these models require that certain conditions be met before countries are able to achieve take-off, those who do not meet these requirements could find themselves trapped in a phase of economic stagnation for extended periods of time.

The second class of models focuses on the role of technological progress in growth. In particular, the emphasis of these models is on the diffusion of technology from countries which are technological leaders to less developed countries. Lucas (2000) is a seminal work in this area (see also Basu and Weil, 1998; Parente and Prescott, 1994; Howitt and Mayer-Foulkes, 2005). Particular attention has been paid to exploring the channels through which less advanced countries imitate or adopt technologies in leader countries. If there are no barriers to technological diffusion across countries, then these models typically predict that rich and poor countries would gradually converge in per capita income. However, if such barriers exist, then countries may differ in their ability to adopt technologies leading to the creation of ‘clusters’ of countries defined by a set of common barriers to technological adoption. Countries within each of these clusters or ‘convergence clubs’ converge to common levels of mean per capita income. Nevertheless, the per capita incomes across convergence clubs need never converge and the polarization of per capita incomes across countries may be permanent.

Growth empirics

In both classes of models, therefore, the primary concern is that countries may become separated – perhaps permanently – into multiple growth regimes corresponding to different levels of long-run per capita income. The fact that nonlinear growth models imply that global inequality may be persistent has sparked major advances in the area of cross-country growth empirics. Driven by such concerns, the central preoccupation of growth empirics has been to evaluate the conditions under which poor countries catch up with rich ones or fail to do so. Initial work along these lines focused on the concept of ‘conditional convergence’. Conditional convergence is said to occur if permanent per capita income differences between countries can be accounted for solely by structural differences (and not initial conditions). Researchers initially argued that because conditional convergence was predicted by the canonical neoclassical growth model (see Ramsey, 1928; Solow, 1956; Swan, 1956; Cass, 1965;

Koopmans, 1965) whereas nonlinear growth models potentially predict dependence on initial conditions, tests for conditional convergence could be used to discriminate between these classes of theories. Following Mankiw, Romer and Weil (1992) and Barro and Sala-i-Martin (1992), the canonical way such tests were conducted was to first construct a linearized version of the neoclassical growth model about the (unique) steady state with average growth rates across a time period as the dependent variable, and measures of physical and human capital, population growth rates, and initial per capita income as covariates. Researchers then applied the linearized neoclassical model to cross-country data with the aim of testing to see whether the data supported a negative coefficient on initial per capita income. A finding of a negative coefficient on initial per capita income was taken to imply that, conditional on countries having similar structural characteristics (as defined by the set of covariates), poorer countries would close the income gap with the rich – that is, conditional convergence.

An important outcome of the, oftentimes heated (see Sala-i-Martin, 1996), convergence debates of the 1990s was precisely to weaken the idea that such tests of convergence could be interpreted as model selection tests. In a highly influential work, Bernard and Durlauf (1996) strongly disputed the interpretation of such ‘conditional convergence’ tests by pointing out that these tests were not able to discriminate against a class of nonlinear growth theories that have dramatically different ergodic implications from the neoclassical model. The class of models they were referring to was developed by Azariadis and Drazen (1990). Azariadis and Drazen extended the spillover models of Lucas (1988) and Romer (1986) and showed that, if (local) nonconvexities in the production function were sufficiently strong, then countries that are similar in all aspects except for initial conditions may nevertheless be organized into multiple growth regimes, each of which corresponds to a different steady state for long-run per capita income.

Bernard and Durlauf showed that the multiple-regimes Azariadis–Drazen model was theoretically consistent with a finding of conditional convergence in the data. Therefore, even in the narrowly restricted sense of countries being structurally similar, the finding of a negative coefficient to initial income in the data was no guarantee that countries would converge to a common steady state. Galor (1997) lent further support to the relevance of the Azariadis–Drazen model by arguing that standard ways of augmenting the traditional Solow model increased the likelihood that the true data-generating process followed a multiple-regimes rather than a single steady-state model. Clearly, evidence of multiple regimes and nonlinearities in growth raises questions about misspecification in empirical studies that assume that all countries follow the same growth process, and casts doubt on inferences and policy recommendations that are drawn from these studies.

The work by Bernard and Durlauf has spurred a large quantity of research searching for the existence of multiple-growth regimes. One direction of this new research has been to argue that the finding of parameter heterogeneity in the neoclassical model may be suggestive of the existence of multiple growth regimes. In a seminal work, Durlauf and Johnson (1995), employing a classification and regression tree methodology, implemented a version of Azariadis and Drazen's model and showed that there was evidence in the data to suggest that countries grouped according to initial per capita income and literacy rates correspond to four different growth regimes. Their work has inspired a long list of confirmatory works using a wide variety of econometric approaches (for example, Bloom, Canning and Sevilla, 2003; Canova, 2004; Durlauf, Kourtellos and Minkin, 2001; Kourtellos, 2005; Liu and Stengos, 1999; and Tan, 2005).

While there now is a strong consensus in the literature that there exists substantial heterogeneity across countries, it should be emphasized that this finding is only suggestive of multiple-growth regimes and is not conclusive evidence of it. These heterogeneities could arise because of small deviations in the specification of the production function (see Masanjala and Papageorgiou, 2004) which need not correspond to multiple-growth regimes. Further, even within the context of Azariadis–Drazen model, if non-convexities in the production function are not strong enough, the finding of parameter heterogeneity would not imply the existence of multiple regimes (see Durlauf and Johnson, 1995, Figure 2).

An alternative approach to investigating the existence of multiple regimes or convergence clubs has focused on the evolution of the world distribution of per capita income. The aim of this research has been to look for evidence of emerging multimodality (typically, bimodality) in the world income distribution. A secondary aim has been to evaluate the degree of churning within the multimodal distribution. If the world income distribution is characterized by emerging multimodality with little evidence of countries moving freely within the distribution (that is, churning), then this finding would suggest, in a manner analogous with the finding of multiple-growth regimes, that global income inequality is real, intensifying and persistent in nature. In fact, these are the precise findings by Quah (1993). By estimating transition probabilities for the cross-country per capita income distribution, Quah finds emerging ‘twin peaks’ in the world income distribution as well as substantial persistence within the distribution. Quah's seminal work has been confirmed by subsequent work (for example, Bianchi, 1997; Fiaschi and Lavezzi, 2003; and Paap and van Dijk, 1998) even though there had been questions about the robustness of his initial methodology (see Kremer, Onatski and Stock, 2001).

While the findings of the ‘twin peaks’ literature have been suggestive of growth nonlinearities and multiple equilibria, it is not definitive. It is quite possible, for instance, that the aggregate production functions across countries actually exhibit decreasing marginal productivity of capital, so that there is only one steady state. However, other growth factors are sufficiently strong to overcome the convergence effect of diminishing marginal returns to produce divergence and bimodality in cross-country incomes nevertheless. Without an explicit theory to explain the observed income divergence, there is also the question of whether the bimodality in the cross-country income distribution is a transitional or permanent feature of growth (see Galor, 1997; Lucas, 2000).

Conclusion

Nonlinearities in growth have been highly influential in shaping the thinking of both growth theorists and empiricists in recent years. The work on multiple-growth regimes and the world income distribution suggests that there may exist growth factors strong enough to overcome the decreasing marginal productivity of the neoclassical production function, thereby producing increasing inequality across countries. Nevertheless, while an increasingly large body of work finds evidence that is suggestive of growth nonlinearities, many questions remain open and are the subject of current research. What are the factors that are responsible for generating multiple growth regimes or convergence clubs? Are the effects of these factors transient or permanent? If the former, what are the applicable timescales? This area of research continues to be promising and fruitful.

See Also

- balanced growth
- diffusion of technology
- economic growth, empirical regularities in
- economic growth in the very long run
- endogenous growth theory
- growth take-offs
- Rosenstein-Rodan, Paul Narcyz
- structural change

Bibliography

Azariadis, C. and Drazen, A. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105, 501–26.

Barro, R.J. and Sala-i-Martin, X. 1992. Convergence. *Journal of Political Economy* 100, 223–51.

Basu, S. and Weil, D.N. 1998. Appropriate technology and growth. *Quarterly Journal of Economics*, 1025–54.

Bernard, A. and Durlauf, S. 1996. Interpreting tests of the convergence hypothesis. *Journal of Econometrics* 71, 161–73.

Bianchi, M. 1997. Testing for convergence: evidence from nonparametric multimodality tests. *Journal of Applied Econometrics* 12, 393–409.

Bloom, D., Canning, D. and Sevilla, J. 2003. Geography and poverty traps. *Journal of Economic Growth* 8, 355–78.

Canova, F. 2004. Testing for convergence clubs in income per capita: a predictive density approach. *International Economic Review* 45, 49–77.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32, 233–40.

Durlauf, S. and Johnson, P. 1995. Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics* 10, 363–84.

Durlauf, S., Kourtellos, A. and Minkin, A. 2001. The local Solow growth model. *European Economic Review* 45, 928–40.

Fiaschi, D. and Lavezzi, M. 2003. Distribution dynamics and nonlinear growth. *Journal of Economic Growth* 8, 355–78.

Fleming, J.M. 1955. External economies and the doctrine of balanced growth. *Economic Journal* 65, 241–56.

Galor, O. 1997. Convergence? Inferences from theoretical models. *Economic Journal* 106, 1056–69.

Galor, O. and Weil, D.N. 2000. Population, technology, and growth: from the Malthusian regime to the demographic transtition and beyond. *American Economic Review* 90, 806–28.

Howitt, P. and Mayer-Foulkes, D. 2005. R&D, implementation, and stagnation: a Schumpeterian theory of convergence clubs. *Journal of Money, Credit and Banking* 37, 147–77.

Koopmans, T.C. 1965. On the concept of optimal growth. In *The Econometric Approach to Development Planning*. Amsterdam: North Holland.

Kremer, M., Onatski, A. and Stock, J. 2001. Searching for prosperity. *Carnegie-Rochester Conference Series on Public Policy* 55, 275–303.

Kourtellos, A. 2005. Modeling parameter heterogeneity in cross-country growth regression models. Working paper, Department of University of Cyprus.

Lewis, A. 1956. *The Theory of Economic Growth*. London: Allen & Unwin.

Liu, Z. and Stengos, T. 1999. Non-linearities in cross-country growth regressions: a semiparametric approach. *Journal of Applied Econometrics* 14, 527–38.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Lucas, R. 2000. Some macroeconomics for the 21st century. *Journal of Economic Perspectives* 11, 159–68.

Maddison, A. 2004. *The World Economy: Historical Statistics*. Paris: Development Studies Centre, OECD.

Mankiw, N.G., Romer, D. and Weil, D. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107, 407–37.

Masanjala, W. and Papageorgiou, C. 2004. The Solow model with CES technology: nonlinearities and parameter heterogeneity. *Journal of Applied Econometrics* 19, 171–202.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Industrialization and the big push. *Journal of Political Economy* 97, 1003–26.

Nurkse, R. 1953. *Problems of Capital Formation in Underdeveloped Countries*. New York: Oxford University Press.

Paap, R. and van Dijk, H. 1998. Distribution and mobility of wealth of nations. *European Economic Review* 42, 1269–93.

Parente, S.L. and Prescott, E.C. 1994. Barriers to technology adoption and development. *Journal of Political Economy* 102, 298–321.

Quah, D.T. 1993. Empirical cross-section dynamics for economic growth. *European Economic Review* 37, 426–34.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Rosenstein-Rodan, P. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53, 202–11.

Rostow, W.W. 1960. *The Stages of Economic Growth*. Oxford: Oxford University Press.

Sala-i-Martin, X. 1996. The classical approach to convergence analysis. *Economic Journal* 106, 1019–36.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62, 143–51.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70 (1), 65–94.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Tan, C.M. 2005. No one true path to development: uncovering the interplay between geography, institutions, and ethnic fractionalization in economic development. Mimeo, Tufts University.

How to cite this article

Tan, Chih Ming. "economic growth nonlinearities." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000270> doi:10.1057/9780230226203.0435

The New Palgrave Dictionary of Economics Online

economic growth, empirical regularities in

Steven N. Durlauf and Paul A. Johnson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The evolution of economic growth theory throughout the post-war period has been deeply influenced by the effort to explain broad patterns in cross-country behaviour. We discuss some of the salient empirical regularities associated with neoclassical and new growth economics and consider the shift in focus that has occurred. We first describe the stylized facts of Kaldor that played an important role in the assessment of neoclassical growth models. Next, we consider how a switch in focus to a different class of regularities is associated with the new growth economics that began in the 1980s and dominates contemporary research.

Keywords

economic growth; elasticity of substitution; endogenous growth; factor shares in national income; Kaldor, N.; labour productivity; labour's share of income; model uncertainty; neoclassical growth theory

Article

The evolution of economic growth theory throughout the post-war period has been deeply influenced by the effort to explain broad patterns in cross-country behaviour. In this entry, we discuss some of the salient empirical regularities associated with neoclassical and new growth economics and consider the shift in focus that has occurred. We first describe the role of empirical regularities in neoclassical growth theory as it emerged in the 1950s. Next, we consider how a switch in focus to a different class of regularities is associated with the new growth economics that developed in the 1980s and continues to dominate contemporary research. Finally, we assess this shift. Durlauf, Johnson and Temple (2005) contains details of the data and methods used to substantiate the claims made here.

Empirical regularities and neoclassical growth

Neoclassical growth theory is commonly associated with Kaldor's (1961) well-known ‘stylized facts’ of long-run economic behaviour, which primarily focused on the invariance of long run behaviour for advanced economies. Four of his six facts – (1) the constancy of the growth rate of output per worker over long time horizons, (2) the constancy of the growth rate of capital which is lower than the growth rate of the labour supply, (3) the absence of any systematic trends in the capital–output ratio and (4) the constancy of the rate of profit (and, by implication with the other facts, factor shares in national income) – emphasize common behaviour across countries. Only the fifth and sixth facts – the presence of substantial differences in output per worker across countries, and the positive relationship between the rate of profit and the investment–output ratio – focus on heterogeneity. Kaldor (1957) cites the prediction of constant factor shares as an important test of alternative growth models. An important empirical study at the time was Klein and Kosobud (1961) who investigated constancy by testing for a trend in labour's share, finding none using US data from 1900 to 1953.

While these facts are generally cited as a motivation of neoclassical growth models, their actual relationship to the theory is in fact more complicated. In Solow (1956), for example, the objective is the explanation of long-run economic growth and the constancy of factor shares is only mentioned in passing as an implication of the Cobb–Douglas technology. Indeed Solow (1958) criticizes the literature studying the constancy of factor shares for lacking a precise notion of constancy given that exact constancy cannot reasonably be expected. Bronfenbrenner (1960) argues that, for a wide range of values of the elasticity of substitution between capital and labour, and for reasonable variation in the capital–labour ratio, the theoretical variation in factor shares is consistent with that observed. He concludes that the constancy or otherwise of factor shares is not useful in the assessment of (distribution) theories. Put differently, the first three of Kaldor's stylized facts seem most important to understanding the motivation of the neoclassical program; Solow (2000, p. 4) (in a discussion originally published in 1970) remarks that growth theory is largely

> devoted to analyzing the properties of steady states and to finding out whether an economy not initially in a steady state will evolve into one …

How do Kaldor's stylized facts appear from the vantage point of modern empirical growth research? Barro and Sala-i-Martin (2004, pp. 12–16) assess the concordance of Kaldor's stylized facts with the data and conclude that, with the exception of the constant rate of profit, each of the first five holds ‘reasonably well’ for developed economies. They cite evidence suggesting some tendency for the real rate of return to decline in some economies. The evidence they present, and that which we discuss below, shows that, at least as far as it concerns the rate of growth of labour productivity, the sixth of Kaldor's facts also fits well with the data.

Kaldor's stylized facts are therefore of contemporary use in understanding long-run output behaviour. That said, the facts are no longer central to the research efforts in growth economics as other regularities (or the lack thereof) have become the primary focus of research. We therefore turn to those regularities that have become the focus of contemporary work.

Empirical regularities and the new growth economics

The renaissance in growth theory associated with the rise of endogenous growth models was influenced by interest in the determinants of heterogeneity in growth experiences. While not usually called stylized facts, there is a set of general propositions about heterogeneity that have been very important in influencing research. The most prominent global features evident in the data are the divergence in living standards over the past three centuries and the large disparities in living standards at the end of the 20th century. By modern standards, all countries were poor in 1700 but since then sustained growth, first in the United Kingdom and parts of Western Europe, and more recently in the United States and parts of the Asia–Pacific region, has resulted in large cross-country differences in living standards. In 2000 average GDP per worker in some countries was about one-fiftieth that in the United States while more than 40 per cent of the world's population lived in countries with average levels of GDP per worker of no more than ten per cent of that in the United States.

Divergence in living standards over the 1960–2000 period is also evident in the large group of countries covered by the Penn World Tables (PWT) (Heston, Summers and Aten, 2002). While a substantial group of countries has exhibited prolonged growth over this period, there remains a large mass of countries at the bottom of the distribution. One result was a hollowing out of the middle of the distribution – a phenomenon labelled ‘twin peaks’ by Quah (1996; 1997). Moreover, there is strong persistence within the cross-country income distribution with a Spearman rank correlation of 0.84 between GDP per worker in 1960 and that in 2000. This degree of correlation is not peculiar to the PWT data. Easterly et al. (1993) report a rank correlation of 0.82 between GDP per capita in 1988 and that in 1870 for the 28 countries in Maddison (1989). This sense of a lack of mobility is reinforced by Bianchi (1997), who found that very few of the possible crossings from one end of the distribution to the other actually occurred between 1970 and 1989.

The persistence in levels of GDP per worker contrasts sharply with the wide cross-country variation in the growth rates of GDP per worker especially for those countries with relatively low levels of GDP per worker in 1960. The data show scant support for the proposition that the countries of the world are converging to a common level of income per person or for the belief that poor countries have always grown slowly. Both growth ‘miracles’ – countries exhibited consistently strong growth over the 1960–2000 period – and growth ‘disasters’ – countries that did poorly, often having negative average growth rates – are present in the data. East and South East Asian countries are well represented among the former group while the later is dominated by countries in sub-Saharan Africa. Taiwan, for example, grew at an average annual rate of over six per cent during this 40-year period and increased GDP per worker by a factor of 11 in the process. Hong Kong, Korea and Singapore were not far behind in either respect. By contrast, Mauritania, Senegal, Chad, Mozambique, Madagascar, Zambia, Mali, Niger, Nigeria, the Central African Republic, Angola and the Democratic Republic of the Congo all had negative average growth over this period.

For most countries, the average growth rate from 1980 to 2000 was lower than that from 1960 to 1980. The notable exceptions to this observation are China and India. Moreover, past growth does not seem to be a good predictor of future growth as, for example, the correlation between growth in 1960–80 and that in 1980–2000 is just 0.40. Easterly et al. (1993) suggest that the lack of persistence in growth rates indicates the importance of good luck in economic development. Nevertheless, the cross-decade correlations in growth rates have tended to increase during the 1960–80 period, indicating a sorting of countries into distinct groups of winners and losers.

There seems to be little relationship between the 1960 level of GDP per worker and subsequent average growth rates. The cross-country dispersion of growth rates tends to fall as initial income rises largely due to the rarity of poor performance among the countries with relatively high levels of GDP per worker in 1960. There is, however, a close relationship between geographical group membership and economic growth between 1960 and 2000. As alluded to above, the countries of sub-Saharan Africa performed poorly over this period, with three-quarters of them growing at an average annual rate of less than just 1.3 per cent. The countries in South and Central America did somewhat better with three-quarters of them having grown at an average of less than 1.5 per cent. Among the East and South East Asian countries, three-quarters grew at an average rate of over 3.8 per cent, and a similar fraction of the South Asian countries grew at over 1.9 per cent.

Many of the poor countries of the world were unable to break out of stagnation between 1960 and 2000. A country growing at two per cent per year for 40 years would enjoy a 120 per cent increase in income per person over that period. Yet, between 1960 and 2000, about a quarter of countries never exceeded their 1960 income level by more than 60 per cent, and about ten per cent of countries never exceeded their 1960 level by more than 30 per cent. One reason for this stagnation is the disposition of some economies to large, abrupt output collapses. About half of countries experienced a three-year output collapse of 15 per cent or more between 1960 and 2000. Over the same period, the largest three-year output collapse in the United States was 5.4 per cent, and in the United Kingdom 3.6 per cent, both in 1979–82.

In sum, there are large cross-countries disparities in GDP per worker and hence in living standards. These disparities have grown wider since 1960 and the middle of cross-country income distribution has thinned since 1960. There is substantial immobility in a country's position in the distribution. Growth rates are much less persistent and have tended to fall since 1980. In general, the countries of sub-Saharan Africa performed poorly over the 1960–2000 period. The countries in South and Central America did somewhat better while the South Asian countries did better still. The East and South East Asian countries did best of all.

The changing empirical focus of growth economics

The two sets of empirical regularities we have described, while appearing to differ greatly in terms of their implications for understanding the determinants of the growth process, may in fact be reconciled. A key difference between neoclassical and modern growth economics is its domain of explanation: whereas neoclassical theory attempted to understand the long-run behaviour of advanced industrialized economies, the new growth economics attempts to understand worldwide growth patterns. As a result, the differences between the advanced industrialized economies and the rest of the world take on primary importance. Lucas (2002, pp. 2–3) describes his motivation as

> to see whether modern growth theory could also be adapted for use as a theory of economic development. Adaptation of some kind was evidently necessary: The balanced path of growth theory, with constant income growth, and the assumed absence of population pressures, obviously did not fit all of economic history or even all the behavior that can be seen in today's world. The theory is, and was designed to be, a model of the recent past of a subset of countries.

Thus, as the domain of inquiry in growth economics has evolved, the stylized facts of interest have shifted to identifying features of international divergence rather than international convergence. Further, the effort to identify patterns that characterize the differences in cross-country growth experiences has led to empirical research that focuses on the identification of particular factors in generating the divergence. Theoretical work in growth economics moved away from the traditional emphasis on factor accumulation and towards the analysis of a wide range of social, historical, geographic, and political factors as sources of cross-country heterogeneity. For example, a major strand of contemporary research focuses on the ways that institutional quality affects growth and development; see Acemoglu, Johnson and Robinson (2005) for a detailed survey. The richness of the modern growth literature has led to the widespread use of regression methods to allow for the simultaneous consideration of multiple growth determinants, with a focus on identifying which determinants in fact matter.

The move towards regression methods as the basis for empirical growth research has altered the nature of the sorts of regularities that link data and theory. It is still the case that theoretical analyses are often motivated by the identification of a bivariate relationship between some factor of interest and growth rates. However, relationships of this type do not represent basic growth regularities in the way that Kaldor's stylized facts did. The reason for this transition is that the different growth factors that have expanded the domain of growth economics are typically mutually consistent (Brock and Durlauf, 2001) and so the empirical significance of one factor can only be assessed when others are considered as well. Put differently, the finding of a bivariate relationship, or lack thereof, can always be rationalized as reflecting a failure to control for other factors.

As a result, the empirical regularities that matter for contemporary research, such as the coefficient relating a measure of institutional quality to growth, are derivative from statistical analyses of the entire growth process. But statistical models of growth are subject to many forms of model uncertainty, ranging from uncertainty about the appropriate theories to employ to uncertainty about the empirical measurement of the qualitative factors identified by a theory to uncertainty about the details of the statistical specification of a model; see Durlauf and Quah (1999) and Durlauf, Johnson and Temple (2005) for a delineation of these issues. Model uncertainty has meant that there is relatively little consensus on the empirically salient determinants of growth and so little consensus on which regularities should be of primary interest. Thus current growth economics has been handicapped as different papers identify different salient empirical regularities, with inadequate attention to the robustness of such claims. The development of sturdy inferences about the growth process thus represents a very active area of current work.

See Also

- economic growth
- endogenous growth theory
- growth accounting
- level accounting

Bibliography

Acemoglu, D., Johnson, S. and Robinson, J. 2005. Institutions as the fundamental cause of long-run growth. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Barro, R. and Sala-i-Martin, X. 2004. *Economic Growth*, 2nd edn. Cambridge, MA, and London: MIT Press.

Bianchi, M. 1997. Testing for convergence: evidence from nonparametric multimodality tests. *Journal of Applied Econometrics* 12, 393–409.

Brock, W. and Durlauf, S. 2001. Growth empirics and reality. *World Bank Economic Review* 15, 229–72.

Bronfenbrenner, M. 1960. A note on relative shares and the elasticity of substitution. *Journal of Political Economy* 68, 284–7.

Durlauf, S., Johnson, P. and Temple, J. 2005. Growth econometrics. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Durlauf, S. and Quah, D. 1999. The new empirics of economic growth. In *Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Easterly, W., Kremer, M., Pritchett, L. and Summers, L. 1993. Good policy or good luck? Country growth performance and temporary shocks. *Journal of Monetary Economics* 32, 459–83.

Heston, A., Summers, R. and Aten, B. 2002. *Penn World Table Version 6.1*. Philadelphia: Center for International Comparisons at the University of Pennsylvania (CICUP).

Kaldor, N. 1957. A model of economic growth. *Economic Journal* 67, 591–624.

Kaldor, N. 1961. Capital accumulation and economic growth. In *The Theory of Capital, Proceedings of a Conference held by the International Economic Association*, ed. F. Lutz and D. Hague. London: Macmillan.

Klein, L. and Kosobud, R. 1961. Some econometrics of growth: great ratios of economics. *Quarterly Journal of Economics* 125, 173–98.

Lucas, R. 2002. *Lectures on Economic Growth*. Cambridge, MA: Harvard University Press.

Maddison, A. 1989. *The World Economy in the 20th Century*. Paris: OECD.

Quah, D. 1996. Twin peaks: growth and convergence in models of distribution dynamics. *Economic Journal* 106, 1045–55.

Quah, D. 1997. Empirics for growth and distribution: stratification, polarization, and convergence clubs. *Journal of Economic Growth* 2(1), 27–59.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Solow, R. 1958. A skeptical note on the constancy of relative shares. *American Economic Review* 48, 618–31.

Solow, R. 2000. *Growth Theory: An Exposition*. Oxford: Oxford University Press.

How to cite this article

Durlauf, Steven N. and Paul A. Johnson. "economic growth, empirical regularities in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000316> doi:10.1057/9780230226203.0433

The New Palgrave Dictionary of Economics Online

economic growth

Peter Howitt and David N. Weil
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic growth is the increase in a country's standard of living over time. Growth economists study how living standards differ across countries as well as across time. This article discusses some of the broad facts of economic growth and some of the main approaches to its study.

Keywords

adjustment costs; capital accumulation; Cobb–Douglas functions; competition policy; creative destruction; cultural determinants of growth; demographic dividend; development accounting; Dutch disease; economic growth; endogenous growth; export promotion; Fertility; general purpose technology; geographical determinants of growth; growth accounting; health; human capital; import substitution; increasing returns; inequality (global); innovation; institutional determinants of growth; learning by doing; Malthus, T. R.; mortality; natural capital; neoclassical growth theory; patents; physical capital; policy determinants of growth; population growth; production functions; productivity growth; research and development; resource curse; savings and investment; schooling; technical change; technical complementarities; technology; technology spillovers; trade and growth

Article

Economic growth is typically measured as the change in per capita gross domestic product (GDP). Sustained long-term economic growth at a positive rate is a fairly recent phenomenon in human history, most of it having occurred in the last 200 years. According to Maddison's (2001) estimates, per capita GDP in the world economy was no higher in the year 1000 than in the year 1, and only 53 per cent higher in 1820 than in 1000, implying an average annual growth rate of only one-nineteenth of one per cent over the latter 820-year period. Some time around 1820, the world growth rate started to rise, averaging just over one-half of one per cent per year from 1820 to 1870, and peaking during what Maddison calls the ‘golden age’, the period from 1950 to 1973, when it averaged 2.93 per cent per year. By 2000, world per capita GDP had risen to more than 8.5 times its 1820 value.
Growth has been uneven not only across time but also across countries. Since 1820, living standards in Western Europe and its offshoots in North America and the Antipodes have raced ahead of the rest of the world, with the exception of Japan, in what is often referred to as the ‘Great Divergence’. As shown in Figure 1 below, the proportional gap in per capita GDP between the richest group of countries and the poorest group (as classified by Maddison) grew from three in 1820 to 19 in 1998. Pritchett (1997) tells a similar story, estimating that the proportional gap between the richest and poorest countries grew more than fivefold from 1870 to 1990.
Figure 1
Per capita GDP, 1650–2000, 1990 international dollars. *Source*: Maddison (2001).

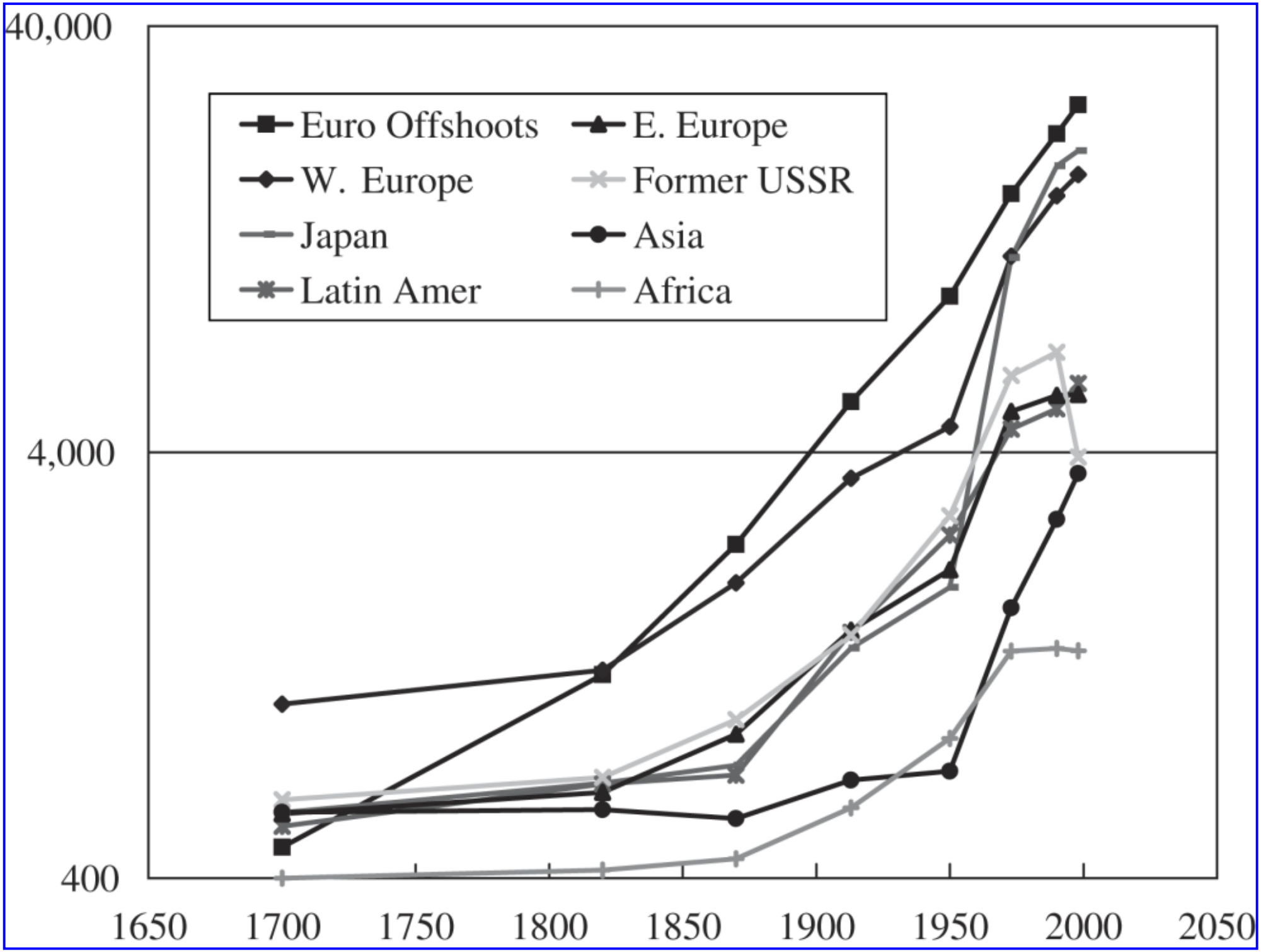

This widening of the cross-country income distribution seems to have slowed during the second half of the 20th century, at least among a large group of nations. Indeed, Figure 1,

which is drawn on a proportional scale, shows that with the acceleration of growth in Asia there has been a narrowing of the spread between the richest and the second poorest group since 1950. Evans (1996) shows a narrowing of the top end of the distribution (that is, among Organisation for Economic Co-operation and Development, OECD, countries) over the period. However, not all countries have taken part in this convergence process, as the gap between the leading countries as a whole and the very poorest countries has continued to widen. In Figure 1 the gap between the Western Offshoots and Africa grew by a factor of 1.75 between 1950 and 1998. Likewise, the proportional income gap between Mayer-Foulkes's (2002) richest and poorest convergence groups grew by a factor of 2.6 between 1960 and 1995.

Jones (1997) argues that continuing divergence of the poorest countries from the rest of the world does not imply rising income inequality among the world's population, mainly because China and India, which contain about 40 per cent of that population, are rising rapidly from near the bottom of the distribution. Indeed, Sala-i-Martin (2006) shows, using data on within-country income distributions, that the cross-individual distribution of world income narrowed considerably between 1970 and 2000, even as the cross-country distribution continued to widen somewhat. But between-country inequality is still extremely important; in 1992 it explained 60 per cent of overall world inequality (Bourguignon and Morrison, 2002). Another reason that growth economists are typically more concerned with the cross-country than the cross-individual distribution is that many of the determinants of economic growth vary across countries but not across individuals within countries.

The production function approach

The main task of growth theory is to explain this variation of living standards across time and countries. One way to organize one's thinking about the sources of growth is in terms of an aggregate production function, which indicates how a country's output per worker *y* depends on the (per worker) stocks of physical, human, and natural capital, represented by the vector *k*, according to

$$y = f(k, A),$$

where *A* is a productivity parameter. Economic growth, as measured by the growth rate of *y*, depends therefore on the rate of capital accumulation and the rate of productivity growth. Similarly, countries can differ in their levels of GDP per capita either because of differences in capital or because of differences in productivity. Much recent work on the economics of growth has focused on trying to identify the relative contributions of these two fundamental factors to differences in growth rates or income levels among countries.

Modern growth theory started with the neoclassical model of Solow (1956) and Swan (1956), who showed that in the long run growth cannot be sustained by capital accumulation alone. In their formulation, the diminishing marginal product of capital (augmented by an Inada condition that makes the marginal product asymptote to zero as capital grows) will always terminate any temporary burst of growth in excess of the growth rate of labour-augmenting productivity. But this perspective has been challenged by more recent endogenous growth theory. In the *AK* theory of Frankel (1962) and Romer (1986), growth in productivity is functionally dependent on growth in capital, through learning by doing and technology spillovers, so that an increase in investment rates in physical capital can also sustain a permanent increase in productivity growth and hence in the rate of economic growth. In the innovation-based theory that followed *AK* theory, the Solow model has been combined with a Schumpeterian theory of productivity growth, in which capital accumulation is one of the factors that can lead to a permanently higher rate of productivity growth (Howitt and Aghion, 1998).

Capital

Having introduced the production function in a general sense, we now examine the accumulation of different types of capital in more detail, and then turn to an assessment of the relative importance of factor accumulation and productivity in explaining income differences among countries and growth over time.

Physical capital

Physical capital is made up of tools, machines, buildings, and infrastructure such as roads and ports. Its key characteristics are, first, that it is produced (via investment), and second that it is in turn used in producing output. Physical capital differs importantly from technology (which, as is discussed below, is also both produced and productive) in that physical capital is rival in its use: only a limited number of workers can use a single piece of physical capital at a time.

Differences in physical capital between rich are poor countries are very large. In the year 2000, for example, physical capital per worker was 148,091 dollars in the United States,

42,991 dollars in Mexico, and 6,270 dollars in India. These large differences in physical capital are clearly contributors to income differences among countries in a proximate sense. That is, if the United States had India's level of capital it would be a poorer country. The magnitude of this proximate effect can be calculated by using the production function. For example, using a value for capital's share of national income of 1/3 (which is consistent with the findings of Gollin (2002) for a cross-section of countries), the ratio of capital per worker in the United States to that in India would by itself explain a ratio of income per capita in the two countries of 7.9 ($=(148{,}091/6{,}270)^{1/3}$).

Differences in physical capital among countries can result from several factors. First, countries may differ in their levels of investment in physical capital relative to output. In an economy closed to external capital flows, the investment rate will equal the national saving rate. Saving rates can differ among countries because of differences in the security of property rights, due to the availability of a financial system to bring together savers and investors, because of government policies like budget deficits or pay-as-you-go old age pensions, differences in cultural attitudes towards present versus future consumption, or simply because deferring consumption to the future is a luxury that very poor people cannot afford.

A second factor that drives differences in investment rates among countries is the relative price of capital. The price of investment goods in relation to consumption goods is two to three times as high in poor countries as in rich countries. If one measures both output and investment at international prices, investment as a fraction of GDP is strongly correlated with GDP per capita (correlation of 0.50), and poor countries have on average between one half and one quarter of the investment rate of rich countries. When investment rates are expressed in domestic prices, the correlation between investment rates and GDP per capita falls to 0.05 (Hsieh and Klenow, 2007).

But levels of capital can also differ among countries for reasons that have nothing to do with the rate of accumulation. Differences in productivity (the *A* term in equation 1) will produce different levels of capital even in countries with the same rates of physical capital investment. Similarly, differences in the accumulation of other factors of production will produce differences in the level of physical capital per worker.

Human capital

Human capital refers to qualities such as education and health that allow a worker to produce more output and which themselves are the result of past investment. Like physical capital, human capital can earn an economic return for its owner. However, the two types of capital differ in several important respects. Most significantly, human capital is ‘installed’ in a person. This makes it very difficult for one person to own human capital that is used by someone else. Human capital investment is a significant expense. In the United States in the year 2000, spending by governments and families on education amounted to 6.2 per cent of GDP; forgone wages by students were of a similar magnitude.

Information on the productivity of human capital can be derived from comparing wages of workers with different levels of education. So called ‘Mincer regressions’ of log wage on years of education, controlling by various means for bias due to the endogeneity of schooling, yield estimated returns to schooling of about ten per cent per year. In the year 2000, the average schooling of workers in advanced countries was 9.8 years and among workers in developing countries 5.1 years. Applying a rate of return of ten per cent implies that the average worker in the advanced countries supplied 56 per cent more labour input because of this education difference. If labour's share in a Cobb–Douglas production function is two-thirds, this would imply that education differences would explain a factor of 1.35 difference in income between the advanced and developing countries, which is very small relative to the observed gap in income. Allowing for differences in school quality increases somewhat the income differences explained by human capital in the form of schooling.

A second form of human capital is health. The importance of health as an input into production can be estimated by looking at microeconomic data on how health affects individual wages. Health differences between rich and poor countries are large, and in wealthy countries worker health has improved significantly over the last 200 years (Fogel, 1997). Weil (2007), using the adult survival rate as a proxy for worker health, estimates that eliminating gaps in worker health among countries would reduce the log variance of GDP per worker by 9.9 per cent.

Natural capital

Natural capital is the value of a country's agricultural and pasture lands, forests and subsoil resources. Like physical and human capital, natural capital is an input into production of goods and services. Unlike other forms of capital, however, it is not itself produced.

Natural capital per worker and GDP per worker are positively correlated, but the link is much weaker than for the other measures of capital discussed above. The poor performance of many resource-rich countries has led many observers to identify a ‘resource curse’ by which the availability of natural capital undermines other forms of capital accumulation or reduces productivity. Among the suggested channels by which this happens are that resource booms lead countries to raise consumption to unsustainable levels, thus depressing saving and investment (Rodriguez and Sachs, 1999); that exploitation of natural resources suppresses the development of a local manufacturing sector, which holds back growth because manufacturing is inherently more technologically dynamic than other parts of the economy (this is the so called Dutch disease); and that economic inefficiencies are associated with political competition or even civil war to appropriate the rents generated by natural resources.

Population and economic growth

Population affects the accumulation of all three forms of capital discussed above, and through them the level of output per worker. Rapid population growth dilutes the quantities of physical and human capital per worker, raising the rates of investment and school expenditure required to maintain output per worker. The interaction of natural capital with population growth is at the centre of the model of Malthus (1798). For a fixed stock of natural capital, higher population lowers output per capita. Combined with a positive feedback from the level of income to population growth, this resource constraint produces a stable steady state level of output per capita and, with technology fixed, a stable level of population as well. This Malthusian feedback is the explanation for the long period of nearly constant living standards that preceded the Industrial Revolution (Galor and Weil, 2000). Because of resource-saving technological progress, as well as expansion of international trade, which allows countries to evade resource constraints, the interaction of population and natural capital is much less important today than in the past, with the exception of very poor countries that are reliant on subsistence agriculture.

In addition to its effect on the level of factors of production per worker, population also matters for economic growth because demographic change produces important changes in the age structure of the population. A reduction in fertility, for example, will produce a long period of reduced dependency, in which the ratio of children and the elderly, on the one hand, to working age adults, on the other, is temporarily below its sustainable steady state level. This is the so-called ‘demographic dividend’ (see population ageing).

In addition to these effects of population on the level of income per capital, there is also causality that runs from the economic to the demographic. Over the course of economic development, countries generally move through a demographic transition in which mortality rates fall first, followed by fertility rates. While the decline in mortality is easily explained as a consequence of higher income and technological progress, the decline in fertility is not fully understood. Among the factors thought to contribute to the decline in fertility are falling mortality, a shift along a quality–quantity trade-off due to rising returns to human capital, the rise of women's relative wages, the reduced importance of children as a means of old age support, and improvements in the availability of contraception.

Growth accounting and development accounting

The discussion above makes clear that stocks of different forms of capital are positively correlated with GDP per capita. Similarly, as countries grow, levels of capital per worker grow as well. It is natural to ask whether these variations in capital are sufficiently large to explain the matching variations in growth. The techniques of growth accounting (Solow, 1957) and development accounting (Klenow and Rodriguez-Clare, 1997; Hall and Jones, 1999) attempt to give quantitative answers to this question. Using a parameterized production function and measures of the quantities of human and physical capital, one can back out relative levels of productivity among countries and rates of productivity growth within a country.

Caselli (2005) presents a review of development accounting along with his own thorough estimates. His finding is that if human and physical capital per worker were equalized across countries, the variance of log GDP per worker would fall by only 39 per cent. In other words, the majority of variation in income is due to differences in productivity, not factor accumulation. Differences in productivity growth, rather than differences in the growth of physical and human capital, are also the dominant determinants of differences in income growth rates among countries (Weil, 2005, ch. 7; Klenow and Rodriguez-Clare, 1997); differences in productivity levels among countries are striking. For example, comparing the countries at the 90th and 10th percentiles of the income distribution (which differ in income by a factor of 21), the former would produce seven times as much output as the latter with equal quantities of human and physical capital.

Productivity, technology and efficiency

Development accounting shows that productivity differences among countries are the dominant explanation for income differences. Similarly, differences in productivity growth are the most important explanation for differences in income growth rates among countries. And as a theoretical matter, the Solow model shows that as long as there are decreasing returns to capital per worker, productivity growth can be the only source of long-term growth. The question is: what explains these changes over time and differences in the level of productivity? Over the long term it is natural to associate productivity growth with technological change. However, especially as an explanation for differences in productivity at a given point in time, a second possibility is that productivity differences reflect differences not in technology, in the sense of inventions, blueprints, and so on, but rather differences in how economies are organized and use available technology and inputs. We label this second contributor to productivity as ‘efficiency’.

Technology

Technology consists of the knowledge of how to transform basic inputs into final utility. This knowledge can be thought of as another form of capital, an intangible intellectual capital. What distinguishes technology from human or physical capital is its non-rival character. For example, the knowledge that a particular kind of corn will be immune to caterpillars, or the knowledge of how to produce a 3•GHz CPU for a portable computer, can be used any number of times by any number of people without diminishing anyone's

ability to use it again. By contrast, if you drive a lorry for an hour, or if you employ the skills of a doctor for an hour, then that lorry or those skills are not available to anyone else during that hour.

Different growth theories have different approaches to modelling the accumulation of technology – that is, technical progress. According to neoclassical theory, for example, the relationship between technology and the economy is a one-way street, with all of the causation running from technology to the economy. It portrays technical progress as emanating from a scientific progress that operates outside the realm of economics, and thus takes the rate of technical progress as being given exogenously.

This neoclassical view has never been accepted universally. Specialists in economic history and the economics of technology have generally believed that technical progress comes in the form of new products, new techniques and new markets, which do not spring directly from the scientific laboratory; instead they come from discoveries made by private business enterprises, operating in competitive markets, and motivated by the search for profits. For example, the transistor, which underlies so much recent technological progress, was discovered by scientists working for the AT&T telephone company on the practical problem of how to improve the performance of switch boxes that were using vacuum tubes. Rosenberg (1981) describes many other examples of scientific and technological breakthroughs that originated in profit-oriented economic activity.

What kept this view of endogenous technology from entering the mainstream of economics until recently was the difficulty of incorporating increasing returns to scale into dynamic general equilibrium theory. Increasing returns arise once one considers technology as a kind of capital that can be accumulated, because of its non-rival nature; that is, the cost of developing a technology for producing a particular product is a fixed set-up cost, which does not have to be repeated when more of the product is produced. Once the technology has been developed then there should be at least constant returns to scale in the factors that use that technology, on the grounds that if you can do something once then you can do it twice. But this means that there are increasing returns in the broad set of factors that includes the technology itself. Increasing returns creates a problem because it generally implies that a competitive equilibrium will not exist, at least not without externalities.

These technical difficulties were overcome by the new ‘endogenous growth theory’ introduced by Romer (1986) and Lucas (1988), which incorporated techniques that had been developed for dealing with increasing returns in the theories of industrial organization and international trade. The first generation of endogenous growth theory to enter the mainstream was the ‘AAK theory’, according to which technological progress takes place as a result of externalities in learning to produce capital goods more efficiently. The second generation was the innovation-based theory of Romer (1990) and Aghion and Howitt (1992), which emphasizes the distinction between technological knowledge and other forms of capital, and analyses technological innovation as a separate activity from saving and schooling.

Historically, technical progress has engendered much social conflict, because it involves what Schumpeter (1942) called ‘creative destruction’; that is, new technologies render old technologies obsolete. As a result, technical progress is a game with losers as well as winners. From the handloom weavers of early 19th century Britain to the former giants of mainframe computing in the late 20th century, many people's skills, capital equipment and technological knowledge have been devalued and their livelihoods imperilled by the same innovations that have created fortunes for others.

The destructive side of technical progress shows up most clearly during periods when a new ‘general purpose technology’ (GPT) is being introduced. A GPT is a basic enabling technology that is used in many sectors of the economy, such as the steam engine, the electric dynamo, the laser or the computer. As Lipsey, Carlaw and Bekar (2005) have emphasized, a GPT typically arrives only partially formed, creates technological complementarities and opens a window on new technological possibilities. Thus it is typically associated with a wave of new innovations. Moreover, the period in which the new GPT is diffusing through the economy is typically a period of rapid obsolescence, costly learning and wrenching adjustment. Greenwood and Yorukoglu (1997) argue that the productivity slowdown of the 1970s is attributable to the arrival of the computer, and Howitt (1998) argues that the rapid obsolescence generated by a new GPT can cause per capita income to fall for many years before eventually paying off in a much higher standard of living.

New technologies are often opposed by those who would lose from their introduction. Some of this opposition takes place within the economic sphere, where workers threaten action against firms that adopt labour-saving technologies and firms try to pre-empt innovations by rivals. But much of it also takes place within the political sphere, where governments protect favoured firms from more technically advanced foreign competitors, and where people sometimes vote for politicians promising to preserve traditional ways of life by blocking the adoption of new technologies.

The leading industrial nations of the world spend large amounts on R&D for generating innovations. In the United States, for example, R&D expenditures constituted between 2.2 and 2.9 per cent of GDP every year from 1957 to 2004. But not much cutting-edge R&D takes place outside a small group of countries. In 1996, for example, 73 per cent of the world's R&D expenditure, as measured by UNESCO, was accounted for by just five countries (in decreasing order of R&D expenditure they are the United States, Japan, Germany, France and United Kingdom). In the majority of countries that undertake very little measured R&D, technology advances not so much by making frontier innovations as by implementing technologies that have already been developed elsewhere. But the process of implementation is not costless, because technologies tend to be context-dependent and technological knowledge tends to be tacit. So implementation requires an up-front investment to adapt the technology to a new environment (see, for example, Evenson and Westphal, 1995). This investment plays the same role analytically in the implementing country as R&D does in the original innovating country.

Implementation is important in accounting for the patterns of cross-country convergence and divergence noted above. This is because a country in which firms are induced to spend on implementation have what Gerschenkron (1952) called an ‘advantage of backwardness’. That is, the further they fall behind the world's technology frontier the faster they will grow with any given level of implementation expenditures, because the bigger is the improvement in productivity when they implement any given foreign technology. In the long run, as Howitt (2000) has shown, this force can cause all countries that engage in R&D or implementation to grow at the same rate, while countries in which firms are not induced to make

such investments will stagnate. But technology transfer through implementation expenditures is no guarantee of convergence, because the technologies that are being developed in the rich R&D-performing countries are not necessarily appropriate for conditions in poor implementing countries (Basu and Weil, 1998; Acemoglu and Zilibotti, 2001) and because financial constraints may prevent poor countries from spending at a level needed to keep pace with the frontier (Aghion, Howitt and Mayer-Foulkes, 2005).

Efficiency

The efficiency with which a technology is used is not likely to play a major role in accounting for long-run growth rates, because there is a finite limit to how high you can raise living standards simply by using the same technologies more efficiently. But there is good reason to believe that differences in efficiency account for much of the cross-country variation in the level of productivity.

Inefficiencies take several different forms. Economic resources are sometimes allocated to unproductive uses, or even unused, as when union featherbedding agreements kick in. Resources can be misallocated as the result of taxes, subsidies and imperfect competition, all of which create discrepancies between marginal rates of substitution. Technologies can be blocked by those who would lose from their implementation and have more market power or political influence than those who would win.

The distinction between differences in technology and differences in efficiency is often unclear. Suppose firms in country A are using the same machinery and the same number of workers per machine as in country B, but output per worker is higher in A than B. This may appear to be an obvious case of inefficiency, since the technology embodied in the machines used by workers in the two countries is the same. But maybe it is just that people in country B lack the knowledge of how best to use the machines, in which case it may actually be a case of differences in technology. As an example, General Motors has had little success in their attempts to emulate the manufacturing methods that Toyota has deployed successfully for many years even in their US operations.

Moreover, identical technologies will have different effects in different countries, because of differences in language, raw materials, consumer preferences, workers expectations and the like. Euro Disney, for example, was plagued initially with labour disputes when it first opened its park in the outskirts of Paris in 1987. It took the American managers several years to realize that the problem was not recalcitrant workers but rather that French workers consider it an intolerable indignity to be forced to wear items such as mouse ears when serving the public. A minor adjustment in amusement park technology was needed to make it as productive in France as it had been in the United States.

Deeper determinants of growth

Even if we knew how much of the cross-country variation in growth rates or income levels to attribute to different kinds of capital or to technology or efficiency, we would still be faced with the deeper question of why these differences in capital and productivity arise. A large number of candidate explanations have been offered in the literature. These candidates can be classified into four broad categories: geography, institutions, policy and culture.

Geographical differences are perhaps the most obvious. As Sachs (2003) has emphasized, countries that are landlocked, that suffer from a hazardous disease environment and that have difficult obstacles in the way of internal transport, will almost certainly produce at a lower level than countries without these problems, even if they use the same technology and the same array of capital. In addition, the lower productivity of these countries will serve to reduce the rate of return to accumulating capital and to generating new technologies.

Institutions matter because of the way they affect private contracts and also because of the way they affect the extent to which the returns to different kinds of investments can be appropriated by the government. The origin of a country's legal system has been shown by La Porta et al. (1998) to have an important effect on private contracts. In particular, these authors show that countries with British legal origins tend of offer greater protection of investor and creditor rights, which in turn is likely to affect both capital accumulation and investment in technology by making outside finance more easily available.

Because long-term productivity growth requires technical progress, it depends on political, institutional and regulatory factors that affect the way the conflict between the winners and losers of technical progress will be resolved, and hence affect the incentives to create and adopt new technologies. For example, the way intellectual property is protected will affect the incentive to innovate, because on the one hand no one will want to spend resources creating new technologies that his or her rivals can easily copy, while on the other hand a firm that is protected from competition by patent laws that make it difficult for rivals to innovate in the same product lines will be under less pressure to innovate. Likewise, a populist political regime may erect barriers to labour-saving innovation, resulting in slower technical progress.

Economic policies matter not only because of the way they affect the return to investing in capital and technology but also because of the inefficiencies that can be created by taxes and subsidies. But how these policies affect economic growth can vary from one country to another. In particular, Aghion and Howitt (2006) have argued that growth-promoting policies in technologically advanced countries are not necessarily growth-promoting in poorer countries, because innovation and implementation are affected differently by the same variables. For example, tighter competition policy in a relatively backward country might retard technology development by local firms that will be discouraged by the threat of foreign entry, whereas in more advanced countries firms will be spurred on to make even greater R&D investments when threatened by competition.

As this example suggests, international trade is one of the policy domains most likely to matter for growth and income differences, because of the huge productivity advantage that is squandered by policies that run counter to comparative advantage, because protected firms tend to become technologically backward firms, and because for many countries

international trade is the only way for firms to gain a market large enough to cover the expense of developing leading-edge technologies. So it is probably no accident that export promotion has been a prominent feature of all the East Asian countries that began escaping from the lower end of the world income distribution towards the end of the 20th century, whereas import substitution was a prominent feature of several Latin American countries that fell from the upper end of the distribution early in the 20th century.

Culture is a difficult factor to measure. In principle, however, it is capable of explaining a great deal of cross-country variation in growth, because a society in which people are socialized to trust each other, to work hard, to value technical expertise and to respect law and order is certainly going to be thriftier and more productive than a society in which these traits do not apply. Recent work has begun to quantify the role of culture using measures of social capital, social capability, ethno-linguistic fractionalization, religious belief, the spread of Anglo-Saxon culture and many other variables.

See Also

- economic growth, empirical regularities in
- economic growth in the very long run
- endogenous growth theory
- growth accounting
- growth and institutions
- growth take-offs
- level accounting
- population ageing

Bibliography

Acemoglu, D. and Zilibotti, F. 2001. Productivity differences. *Quarterly Journal of Economics* 116, 563–606.

Aghion, P. and Howitt, P. 1992. A model of growth through creative destruction. *Econometrica* 60, 323–51.

Aghion, P. and Howitt, P. 2006. Appropriate growth policy: an integrating framework. *Journal of the European Economic Association* 4, 269–314.

Aghion, P., Howitt, P. and Mayer-Foulkes, D. 2005. The effect of financial development on convergence: theory and evidence. *Quarterly Journal of Economics* 120, 173–222.

Basu, S. and Weil, D.N. 1998. Appropriate technology and growth. *Quarterly Journal of Economics* 113, 1025–54.

Bourguignon, F. and Morrison, C. 2002. Inequality among world citizens: 1820–1992. *American Economic Review* 92, 727–44.

Caselli, F. 2005. Accounting for cross-country income differences. In *Handbook of Economic Growth*, vol. 1, ed. P. Aghion and S.N. Durlauf. Amsterdam: North-Holland.

Evans, P. 1996. Using cross-country variances to evaluate growth theories. *Journal of Economic Dynamics and Control* 20, 1027–49.

Evenson, R.E. and Westphal, L.E. 1995. Technological change and technology strategy. In *Handbook of Development Economics*, vol. 3A, ed. T.N. Srinivasan and J. Behrman. Amsterdam: Elsevier.

Fogel, R. 1997. New findings on secular trends in nutrition and mortality: some implications for population theory. In *Handbook of Population and Family Economics*, vol. 1A, ed. M. R. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Frankel, M. 1962. The production function in allocation and growth: a synthesis. *American Economic Review* 52, 995–1022.

Galor, O. and Weil, D.N. 2000. Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. *American Economic Review* 90, 806–28.

Gerschenkron, A. 1952. Economic backwardness in historical perspective. In *The Progress of Underdeveloped Areas*, ed. B.F. Hoselitz. Chicago: University of Chicago Press.

Gollin, D. 2002. Getting income shares right. *Journal of Political Economy* 110, 458–74.

Greenwood, J. and Yorukoglu, M. 1997. *1974 Carnegie-Rochester Conference Series on Public Policy* 46, 49–95.

Hall, R. and Jones, C. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114, 83–116.

Howitt, P. 1998. Measurement, obsolescence, and general purpose technologies. In *General Purpose Technologies and Economic Growth*, ed. E. Helpman. Cambridge, MA: MIT Press.

Howitt, P. 2000. Endogenous growth and cross-country income differences. *American Economic Review* 90, 829–46.

Howitt, P. and Aghion, P. 1998. Capital accumulation and innovation as complementary factors in long-run growth. *Journal of Economic Growth* 3, 111–30.

Hsieh, C.-T. and Klenow, P. 2007. Relative prices and relative prosperity. *American Economic Review* 97.

Jones, C.I. 1997. On the evolution of the world income distribution. *Journal of Economic Perspectives* 11(3), 19–36.

Klenow, P. and Rodriguez-Clare, A. 1997. The neoclassical revival in growth economics: has it gone too far? In *NBER Macro Annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R.W. 1998. Law and finance. *Journal of Political Economy* 106, 1113–55.

Lipsey, R.G., Carlaw, K.I. and Bekar, C.T. 2005. *Economic Transformations: General Purpose Technologies and Long Term Economic Growth*. New York: Oxford University Press.

Lucas, R.E. Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Maddison, A. 2001. *The World Economy: A Millennial Perspective*. Development Centre Studies. Paris: OECD.

Malthus, T.R. 1798. *An Essay on the Principle of Population, as it Affects the Future Improvement of Society with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*. London: printed for J. Johnson in St Paul's Churchyard.

Mayer-Foulkes, D. 2002. Global divergence. Documento de Trabajo del CIDE, SDTE 250, División de Economía. Mexico: CIDE.

Pritchett, L. 1997. Divergence, big-time. *Journal of Economic Perspectives* 11(3), 3–17.

Rodriguez, F. and Sachs, J.D. 1999. Why do resource-abundant economies grow more slowly? *Journal of Economic Growth* 4, 277–303.

Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Romer, P.M. 1990. Endogenous technological change. *Journal of Political Economy* 98, S71–102.

Rosenberg, N. 1981. How exogenous is science? In *Inside the Black Box: Technology and Economics*, ed. N. Rosenberg. New York: Cambridge University Press.

Sachs, J.D. 2003. Institutions don't rule: direct effects of geography on per capita income. Working Paper No. 9490. Cambridge, MA: NBER.

Sala-i-Martin, X. 2006. The world distribution of income: falling poverty and … convergence, period. *Quarterly Journal of Economics* 121, 351–97.

Schumpeter, J.A. 1942. *Capitalism, Socialism and Democracy*. New York: Harper.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Solow, R.M. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Weil, D.N. 2005. *Economic Growth*. Boston: Addison-Wesley.

Weil, D.N. 2007. Accounting for the effect of health on economic growth. *Quarterly Journal of Economics* 122.

How to cite this article

Howitt, Peter and David N. Weil. "economic growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000209> doi:10.1057/9780230226203.0432

The New Palgrave Dictionary of Economics Online

economic history

Alexander J. Field

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic history focuses on the historical study of growth and development. Originating in the German historical school and studies of the Industrial Revolution in England, it became professionally differentiated from economics proper with the establishment of associations in Britain (1926) and the United States (1941). As economics continued on its increasingly mathematical and ahistorical path in the 1960s, the ‘new economic history’ advocated applying theory to history. But its emphasis on data analysis retained a bridge to older traditions. As economists have rediscovered an interest in long-term economic growth, often applying traditional institutional approaches, there is continuing evidence of rapprochement.

Keywords

American Economics Association; anthropometric history; cliometrics; economic development; economic growth; economic history; Economic History Association in the United States; Historical School (German); Great Depression; Industrial Revolution; Kuznets, S.; Marshall, G.; Marx, K.; mathematical models; mercantilism; new economic history; North, D.; Physiocrats; Royal Economic Society; Smith, A.; total factor productivity; Weber, M.

Article

Economic history is a sub-discipline within economics and, to a lesser degree, within history, whose main focus is the study of economic growth and development over time. It is to be distinguished from the history of economic thought, a branch of intellectual history.
Studies in economic growth, whether historical or contemporary, develop and analyse quantitative measures of increases in output and output per capita, emphasizing in particular changes in saving rates and rates of technological innovation and their consequences. Economic development is a larger and more encompassing rubric, also including consideration of the role of cultural changes and changes in

formal institutions.

Economic history has its origins in two main traditions. The first is the German historical school, a group of scholars in the 19th and early 20th centuries, including Gustav Schmoller and Max Weber, who ranged widely over human history with special emphasis on the consequences of institutional variation for economic as well as political performance. The second tradition stems from the efforts of a group of writers who viewed the complex of innovations in steam power, iron manufacture, and textiles in late 18th-century Britain as an epochal event – an industrial revolution – equivalent in its significance for human welfare to the Neolithic revolution which gave birth to agriculture around ten millennia earlier. The study of the causes, dimensions and consequences of the emergence of sustained increases in per capita incomes – what Simon Kuznets (1966) called modern economic growth – along with a focus on the consequences of institutional variation, continues to define much of what economic historians do.

Although historians have practised their craft at least since the time of Herodotus (the fifth century bc), economics emerged as a separate social science with the work of Adam Smith or, perhaps, as some have argued, that of the Mercantilists and the Physiocrats. Classical economists, with the notable exception of Ricardo, were almost all also historical economists. The reader of Smith, Mill, Marx, or even Marshall ploughs through thick volumes in which propositions in economic theory are embedded in often lengthy descriptions of historical events or the course of economic history. Throughout most of the 19th century, the divide between economists and economic historians was weak.

With the professionalization of economics that picked up speed in the 20th century (the American Economics Association was founded in 1885; the Royal Economic Society in 1890), economic history began to emerge as a distinct and to some degree separate sub-discipline. The Economic History Society was founded in Britain in 1926; the Economic History Association in the United States in 1941. The trend towards a separate identity accelerated in the third quarter of the 20th century, with the increasing emphasis within economics on formal mathematical modelling and the weakening within the general profession of ties to historical traditions. In the economic history societies, in contrast, those trained as historians as well as economists remained active; in Britain, distinctiveness was accentuated by the establishment of separate university departments of economic history.

As the intellectual paths taken by economics and economic history seemed increasingly to diverge, a countervailing intellectual movement known variously as ‘cliometrics’ or the ‘new economic history’ emerged. Its pioneers knew their history, but emphasized by argument and example that, if economic history was to remain influential within economics, it had to make more use of formal models as well as place increased emphasis on quantitative (rather than just qualitative) data and more advanced statistical techniques (econometrics) to analyse them.

The use of mathematical models was anathema within historical traditions, but by the 1960s widely accepted in economics. Thus, the new economic history represented something of a gauntlet thrown down to those trained in history or allied with its traditions. The push for quantitative data analysis, in contrast, was more cross-cutting in the challenges it implied. Many traditional economic historians had in fact examined such data, although the statistical techniques they used were often quite rudimentary. Within some economic circles, on the other hand, an emphasis on data was becoming suspect. Here, some scholars were comfortable with the evolution of economic theory as a branch of applied mathematics, constrained and judged by the rules of logic and consistency, but governed in its realism, if at all, by intuition rather than systematic empirical inquiry.

The effort to force formal theory upon traditional economic history often lacked acknowledgement that

the relation between economic history and formal theory might usefully be a two-way street. The emphasis on data analysis, in contrast, offered a bridge between economics and economic history. It helped reaffirm within economics the importance of empirical inquiry, and encouraged those historically trained to become more sophisticated in their statistical analyses.

Nevertheless, the stress on quantitative data could not help but draw attention away from economic history's traditional concern with legal and institutional variation, where the source documents were almost uniformly qualitative. How would this theme, one of the defining features of economic history since its inception, survive the new economic history? The initial ‘solution’ was to try to make institutions endogenous. Blending a mix of influences from technologically deterministic Marxism to the emerging law and economics and public choice literatures, a number of scholars suggested that institutions could be understood as epiphenomenal: reflective of more fundamental givens. The high point of such efforts was probably the short book by North and Thomas (1973).

These efforts, however, gradually disintegrated under the force of the ad hoc twists required to make the framework consistent with known historical evidence (Field, 1981), and even proponents such as North eventually backed away from this agenda. Formal rules often vary where technologies and endowments are similar, and are often similar when more fundamental givens differ, and such variation has consequences for economic performance. Had the endogenization initiative been successful, it would have eliminated from economic history one of the most important perspectives it offers to general economics.

The old economic historians had taken it as obvious that, at critical historical junctures, changes in formal institutions such as laws or constitutions had powerful influences on the course of a country or region's economic development, and that these changes were not always predictable *ex ante*. The breakdown of the former Soviet Empire, and the opportunities afforded to Western scholars actually to influence the design of formal institutions, gave a powerful impetus to returning to thinking about such designs as consequential, and increasingly this perspective came to be reflected in research by scholars who did not necessarily think of themselves as economic historians.

If the main subject of economic history continues to be the history of economic growth and development, the influence of variations in formal and informal institutions in both the private and public arenas will remain an important theme. These institutions and a broader economic culture help structure the environment in which individuals pursue their interests. But the success of an economy in raising output and output per person also depends on available technologies, on the size, composition, and characteristics of the labour force, on natural resources, and on the accumulation of physical capital. The study of the evolution of these inputs suggests some of the other themes around which economic historians organize their work. In particular, there is a rich tradition, particularly in the United States, examining issues in and applying methods from modern labour economics within an historical context.

The basic agenda of economic history has not changed since the first edition of *The New Palgrave*. Interest in the causes and dimensions of the Industrial Revolution, for example, remains strong, particularly in Britain. But the field has evolved in new directions, with several discernable trends. First, scholars have concerned themselves with a broadening range of topics under the umbrella of growth and development. In the 1960s and 1970s, especially in the United States, railroads and slavery dominated much of the discussion. In recent decades, it is not possible to point to one or two issues around which research and discussion has coalesced to the same degree. Instead, there have been a number of new initiatives; one example would be the growing exploitation of anthropometric data to make inferences

about variation in standards of living.

Associated with this has been a broadening of the scope of the discipline, both in terms of the countries in which economic history research is conducted and in the geographical range of topics, which extends, somewhat more so than in earlier decades, beyond Western Europe and North America to Asia, Latin America, Australia, and Africa. One illustration of this has been a range of cross-national studies, exploring such issues as economic convergence.

A third trend has been a growing willingness to think of the 20th century as an historical epoch in its own right. When the new economic history began, the Second World War had barely ended and the Great Depression was recent history. The main focus of research was the 18th and 19th centuries. Treating the 20th century as an historical period promises to reduce the gap between economics and economic history. The Great Depression, of course, continues to attract attention, but interest in the 20th century is beginning to expand beyond this. The data and events of recent decades can now more easily be seen in an historical context. The result can be a smoother continuum between topics understood as economic history and the analysis of contemporary data.

Placing more recent developments within a longer-run perspective has already begun to pay important dividends. Many trends that economists and economic historians expected at mid-century would characterize the 20th century as a whole moderated, became erratic, or in some cases reversed themselves in the last quarter of the century (Field, 2001). In 1950, for example, it looked as if the United States (and other countries) would continue to experience decreases in wealth and income inequality, robust and perhaps rising shares of union membership in the labour force, a growing role for government, and a continuing high contribution of total factor productivity (TFP) growth to growth in output per hour. In fact, inequality has generally increased, union membership has fallen, and TFP growth basically disappeared in much of the developed world between 1973 and the 1990s. The size and role of government, which many predicted would continue to expand, has in fact displayed a more complex dynamic.

A fourth and related trend has been a reinvigoration within mainstream economics of interest in what has always been a primary subject of economic history: economic growth. Much of economic theory in the 1950s and 1960s modelled production and allocation within a static economy. The revived interest in the study of growth, combined with the growing willingness of economists to adopt traditional institutional approaches, reflects the persisting influence of the original concerns and approaches of economic history within the larger profession.

Whatever the labels people apply to themselves and others, if we want better understanding of the processes of growth and development, we will continue to need scholars familiar with how to work with data and interpret the influences on economic outcomes of institutional, political, and cultural variation. Doctoral training with a specialization in economic history is well suited to imparting such knowledge and the skills for acquiring it, capabilities that will remain essential in developing improved theory and policy in the area.

See Also

- cliometrics
- growth and cycles

- growth and institutions
- Historical School, German
- institutionalism, old
- technical change

Bibliography

Cairncross, A. 1989. In praise of economic history. *Economic History Review* 42, 173–85.

Field, A. 1981. The problem with neoclassical institutional economics. *Explorations in Economic History* 18, 174–98.

Field, A. 1987. *The Future of Economic History*. Boston: Kluwer-Nijhoff.

Field, A. 2001. Not what it used to be: the Cambridge Economic History of The United States, vols. II and III. *Journal of Economic History* 61, 806–18.

Kuznets, S. 1966. *Modern Economic Growth: Rate, Structure, and Spread*. New Haven: Yale University Press.

McCloskey, D. 1987. *Econometric History*. Basingstoke: Macmillan Education.

North, D. and Thomas, R. 1973. *The Rise of the Western World*. Cambridge: Cambridge University Press.

Solow, R. 1985. Economic history and economics. *American Economic Review* 75, 328–31.

How to cite this article

Field, Alexander J. "economic history." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000016> doi:10.1057/9780230226203.0436

The New Palgrave Dictionary of Economics Online

economic laws

Stefano Zamagni
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

ceteris paribus; economic laws; law of diminishing returns; law of variable proportions; marginal revolution; natural law; neo-Austrian economics; positive economics; substitution; tendency laws

Article

The social sciences, and economics in particular, separated from moral and political philosophy in the second half of the 18th century when the results of the myriad of intentional actions of people were perceived to produce regularities resembling the laws of a system. Both Physiocratic thought and Smith's *Wealth of Nations* reflect this extraordinary discovery: scientific laws thought to be found only in nature could also be found in society. This extension poses several problems. A serious one refers to the tension of combining individuals' freedom of action with the scientists' desire to discover the systematic aspects of the unintended and quite often unpredictable consequences of human action, that is, the desire to arrive at laws characterized by a certain degree of generality and permanence.
In the history of economic thought this fundamental tension has been solved in different ways. In the 18th century, the mechanistic ideal of the natural sciences, combined with the natural law idea of a harmonious order of nature, determined the way social phenomena were treated. There was a desire to discover the 'natural laws' of economic life and to formulate the natural precepts which rule human conduct. The classical economists upheld the notion that natural laws are embedded in the economic process as beneficial laws, along with the belief in the existence of rules of nature capable of being discovered. Thus the belief that things could follow the beneficial 'natural course' only in a rationally organized society which it was a duty to create according to the precepts of nature. The economic system is the mechanism by which the individual is driven to fostering the prosperity of society while pursuing his private interest. Hence the automatic operation of the economic system may be combined with freedom of individual action. This is the core of the doctrine of economic harmony. Besides being causal laws of a mechanical type, the laws of nature are providentially imposed norms of conduct. In such a setting it would have been pointless to separate means and ends, since the implementation of natural laws is both an end and a means, and even more pointless to think of a tension between

‘explaining’ and ‘understanding’ economic behaviour. Causal and teleological, positive and normative, theoretical and practical started being seen as separate categories only when the economic discourse freed itself from the philosophy of natural law and all its implications.

Post-classical economics set out to be a science of the laws regulating the economic order and of the conditions allowing these laws to operate. It became the basis of a theory that, in Jevons's own terms, proposed to construct a ‘social physics’. The view of a social world ordered according to transcendent ends was abandoned in favour of an ideal of objective knowledge of economic phenomena gained through a ‘positive’ study of the laws that regulate market activities. In so doing, neoclassical ‘positive’ economics solves the aforementioned tension by extrapolating the theoretical model of natural sciences to economics: economics is to produce the laws of motion similar to those of physics, chemistry, astronomy.

But what is a scientific law and which role do laws play within the logical positivist's perspective adopted by neoclassical economics? Laws provide the foundation of a deductive scientific method of inquiry. According to the deductive–nomological conception of explanation, due to C. Hempel, laws are universal statements not requiring reference to any one particular object or spatio–temporal location. To be valid, laws are constrained neither to finite populations nor to particular times and places; they are, in effect, expressions of natural stationarities. This interpretation of the notion of law provides the so-called covering-law model of explanation with an unquestionably firm inferential foundation. Deductive logic is employed to ensure the truth status of propositions and, since the deductions are (by hypothesis) predicated on true universal statements (laws), the empirical validity of these statements may be ascertained. However, what sort of constraints on economic discourse are imposed by this positivistic structure? On the one hand this structure constitutes its object; on the other hand it generates specific economic questions together with their method of solution. Following the model of natural sciences and its success in controlling a natural world made up of objects and unvarying relations among them expressed in the form of laws, the neoclassical approach arrives at a study of regularities conceived of as specifying the nature of its objects.

To capture the different interpretations of the notion of law by classical and neoclassical economists let us refer to one of the most famous of economic laws: the law of diminishing returns, also known as the law of variable proportions. Studying agricultural production, Ricardo had noted that different quantities of labour, assisted by certain quantities of other inputs (farm tools, fertilizers, and so on), could be employed on a given piece of land, that is, it was possible to vary the proportions in which land and complex labour (labour assisted by other inputs) are employed. He accordingly arrived at the law which states that production increases resulting from equal increments in the employment of complex labour, while the quantity of land farmed remains constant, will initially be increasing and then decreasing. (To be sure, the first statement of the law is due to the Physiocratic economist Turgot.)

Three points deserve attention. First, Ricardo and classical authors in general offer no formal demonstration of this law. To them, it is basically an empirical law, on which no functional association between output and variable inputs can be built. Second, the classics’ use of the law refers to their theories of distribution and development: as the supply of land in the whole system is fixed, sooner or later a point will be reached at which economic growth will come to a halt, notwithstanding any countervailing effects due to technical progress. Finally, the law presupposes a comparative statics framework: the pattern of the marginal products of complex labour refers to different observable equilibrium positions and not to hypothetical or virtual variations.

With the advent of the marginalist revolution, two subtle changes in the interpretation of the law took place. (*a*) The *de facto* elimination of the distinction between the extensive case (the case of the simultaneous cultivation of pieces of land of different fertility) and the intensive case (the application of successive doses of capital and labour to the same piece of land) with an over-evaluation of the latter. Classical economists, being interested in the explanation of rent, concentrated on the extensive case; they took also the intensive case into consideration but with many qualifications. Indeed, whereas the various levels of productivity of different qualities of land is a circumstance which may be directly observed in a given situation, the marginal productivity of a given input is related to a virtual increment in output and therefore to a virtual change in the situation. (*b*) The change in the method of analysis – it was preferred to reason in terms of hypothetical rather than observable changes – brought about by the shift of interest towards the intensive margin, supported the thesis of the symmetrical nature of land and other inputs. This in turn favoured the extension of the substitutability between land and complex labour from agricultural production to all kinds of production, including those in which land does not figure as a direct input. It so happened that whereas in classical economics the substitutability between land and complex labour presupposes that simple labour and equipment are strictly complementary, in neoclassical economics this substitutability is applied to all inputs indiscriminately.

However, the neoclassical interpretation of the law poses serious problems. In the first place, there is the problem of justifying, on empirical grounds, the general applicability of the substitution principle. Secondly, and more importantly, in order to allow the substitution of inputs to take place, a certain lapse of time is required during which the required modifications to the productive structure can be made. (It is certainly true that coal can replace oil to provide heating, but before this can happen it will be necessary to change the heating system.) The well-known distinction between the short run and the long run is a partial and indirect way to take the temporal element into consideration. In the short run the plant is fixed by definition. It is therefore the fixed input which, in the neoclassical interpretation of the law, plays the same role as land in the classical interpretation. Now, neoclassical theory correctly states the law of diminishing returns with respect to the short run; however it is in the long run that the substitutability of inputs becomes actually feasible. One is therefore confronted with a dilemma: the neoclassical interpretation of the law seems to be more plausible in a long-run framework when there exists the necessary time to accommodate input adjustments; on the other hand, fixed inputs cannot, by definition, exist in the long run so that the law of variable proportions cannot be stated in such a context. This dilemma is the price neoclassical theory has to pay for its interpretation of the law in accordance with the positivistic statute. Indeed, the power of deductive, truth-preserving rules of scientific inference is not purchased without a cost. A school of economic thought which is not prepared to sustain such a cost is the neo-Austrian. The neo-Austrian economists solve what has been called the fundamental tension by arguing economics cannot and should not provide general laws since, by its very nature, it is an idiographic and not a nomothetical discipline. The general target of economics is ‘understanding’ grounded in *Verstehen* doctrine: by introspection and empathy, the study of the economic process should aim at explaining individual occurrences, not abstract classes of phenomena. It follows that if by a scientific law one should mean a universal conditional statement of type ‘for all *x*, if *x* is *A*, then *x* is *B*’, statements regarding unique events cannot by definition express any regularity for the simple reason that any regularity presupposes the recurrence of what is defined as regular. In the words of L. von Mises, who shares with F. von Hayek the paternity of the neo-Austrian school, what assigns economics its peculiar and unique position in the orbit of pure knowledge ‘… is the fact that its particular theorems are

not open to any verification or falsification on the ground of experience … the ultimate yardstick of an economic theorem's correctness or incorrectness is solely reason unaided by experience' (von Mises, 1949, p. 858).

There is indeed a place for economic 'laws' in the framework of Austrian economics. The familiar 'laws' of economics (diminishing marginal utility, supply and demand, diminishing returns to factors, Say's Law and so on) are seen as 'necessary truths' which explain the essential structure of the economic world but with no predictive worth. In other words, economic laws are not generalizations from experience, as it is the case within the positivistic paradigm, but are theorems which enable us to understand the economic world. It is ironic that Mises' position of radical apriorism joined to Hayek's attack on scientism and methodological monism are completely at variance with the position taken by the father of the Austrian school, Carl Menger (1883), who announced that in economic theories exact laws are defined which are just as rigorous as in fact are the laws of nature.

Between the extreme positions of neoclassical positive economic and neo-Austrian economics are those who, without denying that economics is in search for laws in the same sense in which natural sciences are and that laws perform an explanatory as well as a predictive function, underline that the explicative structure of economics, albeit nomothetical, substantially differs from that of natural sciences. This intermediate position can be traced back to Keynes's (1973) methodology which considers the conditions of truth and universality of the positivistic conception of scientific laws as far too rigid for a discipline such as economics. Two main reasons account for the different epistemological status of laws in natural sciences and in economics. First, the knowledge of economic phenomena is itself an economic variable, that is, it changes, along with the process of its own acquisition, the economic situation to which it refers. The formulation of a new physical law does not change the course of physical processes; it does not influence the truth or falsity of the prognosis. This is not the case in economics where the prognosis, say, that in two years time there will be a boom can cause overproduction and a resulting recession. In turn, this specific aspect is strictly connected to the fact that the object of study of economics possesses an historical dimension. Economics is in time in a way that natural sciences are not. The ensuing mutability of observed regularities is well expressed by Keynes when he writes, 'As against Robbins, economics is essentially a moral science and not a natural science. That is to say it employs introspection and judgements of value' (1973, p. 297) to which he adds, 'It deals with motives, expectations, psychological uncertainties. One has to be constantly on guard against treating the material as constant and homogenous' (p. 300).

Second, the role played by *ceteris paribus* clauses in natural sciences and in economics is substantially different. The modern economists appeal to the 'other things being equal' clause – which according to Marshall is invariably attached to any economic law – in all those cases where the classical economists were talking of 'disturbing causes'. J.S. Mill's (1836) discussion of inexact sciences is suggestive here:

> When the principles of Political Economy are to be applied to a particular case then it is necessary to take into account all the individual circumstances of that case … These circumstances have been called *disturbing causes*. This constitutes the only uncertainty of Political Economy. (1836, p. 300)

Also in natural sciences we find *ceteris paribus* clauses. Indeed, a scientific theory that could dispense

with them would in effect achieve perfect closure, which is a rarity. So where lies the difference? The example of the science of tides used by Mill is revealing. Physicists know the laws of the greater causes (the gravitational pull of the moon) but do not know the laws of the minor causes (the configuration of the sea bottom). The ‘other things’ which scientists hold equal are the lesser causes. So could we conclude that just about all generalizations in both natural sciences and economics express in fact *tendency laws*, in the sense that these ‘laws’ truly capture only the functioning of ‘greater causes’ within some domain? Certainly not, since there is a world of difference between the two cases. Galileo's law of falling bodies certainly presupposes a *ceteris paribus* clause, so much so that he had to employ the idealization of a ‘perfect vacuum’ to get rid of the resistance of air. However, he was able to give estimates of the magnitudes of the amount of distortion that friction and the other ‘accidents’ would determine and which the law ignored. In other words, whereas in natural sciences the ‘disturbing causes’ have their own laws, this is not the case in economics where we find tendency statements with unspecified *ceteris paribus* clauses or, if specified, specified only in qualitative terms. In economics it is generally impossible to list all the conceivable inferences implied in a lawlike statement and to replace the *ceteris paribus* clause with precise conditions. So, for example, the law that ‘less will be bought at a higher price’ is not refuted by panic buying, nor is it confirmed by organized consumer boycotts. No test is decisive unless *ceteris* are really *paribus*.

These remarks help to understand the role acknowledged by Keynes to laws in economic inquiry. Besides general laws, there are also rules and norms which are significant in the explanation of economic behaviour. To Keynes, it makes no sense to reduce all forms of explanation in economics to that of the covering-law model. Indeed, whereas to justify a law one has to show that it is logically derivable from some other more general statements, often called principles or postulates, the justification of rules occurs through the reference to goals and the justification of norms through the reference to values which are not general sentences, but rather intended singular patterns or even ideal entities. Since no scientific law, in the natural scientific sense, has been established in economics, on which economists can base predictions, what are used and have to be used to explain or to predict are tendencies or patterns expressed in empirical or historical generalizations of less than universal validity, restricted by local and temporal limits. Recently, Arrow has amazed orthodox economists when raising doubts about the mechanistically inspired understanding of economic processes: ‘Is economics a subject like physics, true for all time or are its laws historically conditioned?’ (Arrow, 1985, p. 322).

The list of generally accepted economic laws seems to be shrinking. The term itself has come to acquire a somewhat old-fashioned ring and economists now prefer to present their most cherished general statements as theorems or propositions rather than laws. This is no doubt a healthy reaction: for too long economists have been under the nomological prejudice, of positivistic origin, that the only route towards explanation and prediction is the one paved with laws, and laws as forceful as Newton's laws. Images in science are never innocent: wrong images can have disastrous effects.

Bibliography

Arrow, K. 1985. Economic history: a necessary though not sufficient condition for an economist. *American Economic Review, Papers and Proceedings* 75 , 320–3.

Keynes, J.M. 1973. *The General Theory and After*. Part II: *Defence and Development*. In *The Collected Writings of John Maynard Keynes*, vol. 14, London: Macmillan.

Menger, C. 1883. *Unterschungen über die Methode der Sozialwissenschaften*. Leipzig: Duncker & Humblot.

Mill, J.S. 1836. On the definition of political economy and the method of investigation proper to it. Repr. in *Collected Works of John Stuart Mill*, vol. 4, *Essays on Economy and Society*, ed. J.M. Robson, Toronto: University of Toronto Press, 1967.

Mises, L. von. 1949. *Human Action: A Treatise on Economics*. London: William Hodge.

How to cite this article

Zamagni, Stefano. "economic laws." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000019> doi:10.1057/9780230226203.0437

The New Palgrave Dictionary of Economics Online

economic man

Shaun Hargreaves Heap

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic man ‘knows the price of everything and the value of nothing’, so said because he or she calculates and then acts so as to satisfy best his or her preferences. The value of these preferences is immaterial. The hypothesis has nevertheless proved remarkably powerful not only in economics but across the social sciences where it has spawned ‘rational choice’ accounts of many aspects of social life. This ambition has attracted critics both from without and within. The latter have developed, with insights from psychology on how people acquire and use information, a less elegant but arguably more realistic model.

Keywords

behavioural economics; bounded rationality; comparative statics; economic man; expected utility maximization; game theory; *homo economicus* ; Hume, D.; information economics; institutional economics; Kant, I.; law of small numbers; learning; neoclassical economics; preferences; rational behaviour; rational expectations; reference dependence; risk; satisficing; self-serving biases; social norms; subjective probability; welfare economics; white noise

Article

Among the many different portrayals of economic agents, the title of *homo economicus* is usually reserved for those who are rational in an instrumental sense. For example, this is how agency is defined in neoclassical economics. In its ideal type case the agent has complete, fully ordered preferences (defined over the domain of the consequences of his or her feasible actions), perfect information and all the necessary computing power. After deliberation, he or she chooses the action that satisfies their preferences better (or at least no worse) than any other. No questions are raised about the source or worth of preferences, reason focuses on the efficient selection of the means to given ends.
This basic model is then made more sophisticated. The theory of risk allows for the point that an action

may have several possible consequences. When preferences are represented via the device of a utility function, the agent assesses his or her expected utility by discounting the utility of each consequence by how likely it is to be the actual one. That requires the agent to have a probability distribution for the consequences, even if only a subjective one. Other refinements include allowance for costs of acquiring information, of processing it and of action. Then there are complexities, illustrated by game theory, when actions of other agents form part of the environment in which the person acts. The basic vision remains, however, one of agents who are rational in the sense that they maximize an objective function subject to constraints (or act ‘as if’ this were the case).

This vision is not unique to neoclassical economics. For example, Marx's profit-maximizing capitalist fits the same instrumental model of rationality. Institutionalist accounts of, for instance, banks or trade unions often conceive economic bodies as similar unitary rational agents. Nor is the vision confined to any specific motivating desire in agents, like a selfish pleasure-maximizing drive. There is scope for allowing ethical preferences alongside the symptomatic textbook desires for apples and oranges. Agents are, however, regarded as self-interested, in the looser sense that they are moved to satisfy whatever preferences they happen to have. Furthermore, granted that *de gustibus non est disputandum*, this modest base is enough to ground a full-blown social theory on a model of agency which can be exported to other social sciences.

Such a social theory is individualist and contractarian, with a pedigree that includes Hobbes's *Leviathan* and Benthamite utilitarianism. The satisfaction of individual preference, aided by felicific calculation, is what makes the social world go round. Social relations become instrumental, in the sense that they embody exchanges in the service of individual preferences (see Becker, 1976). For instance, marriage has been analysed in this spirit as an arrangement to secure the mutual benefit of exchange between two agents with different endowments. Crime has been claimed to occur because calculation of costs and benefits proves it to be the action that maximizes expected utility. Meanwhile, institutions, which feature in elementary microeconomics as constraints on individual choice, become deposits left by earlier transactions, often deliberately so as devices to prevent preferences being frustrated by situations of the Prisoner's Dilemma type. Government policies are explained on the hypothesis that the political arena is also peopled by individuals maximizing expected utility, who form coalitions in support of policies that will secure re-election (see Downs, 1957). In short, *homo economicus* morphs into a universal *homo sapiens*.

Such a full-blown social theory may be too ambitious because assumptions that are plausible for simple market transactions become suspect when scaled up. For example, the ideal-type case makes agents, so to speak, transparent to themselves, and does not allow for history occurring behind their backs. Freudians would object to transparency of preferences and Marxians would invoke theories of false consciousness. (Although Marx's capitalists are instrumentally rational, their desire to maximize profit is an alienated one, ‘forced’ on them by a competitive capitalist system.) Many other social theorists would object to the treatment of norms and social relations as instrumental, on the grounds that norms are prior to preferences. For instance, cultural forms like the rules of orchestral composition are a source of musical preferences rather than a solution to a priori problems of maximizing musical enjoyment. Or, to put this differently, game theory yields too many instances of indeterminacy for an ambitious programme of reducing all social practices to the exercise of instrumental reason by the individual participating agents.

Such objections, of course, need not affect the more modest enterprise of explaining economic

transactions within the parameters of social institutions like the market. But even here *homo economicus* has critics. Philosophically, it is not plain that preferences can be taken as given in a sense which makes them impervious to the agent's beliefs about the moral quality of his or her actions. In supposing that only desires can motivate agents, the economist is taking sides in a continuing philosophical dispute between Humeans, who regard reason as the slave of the passions, and Kantians, who make place for the rational monitoring of desire. This dispute surfaces plainly in welfare economics, when it is asked whether all preferences should count equally or whether ‘capabilities’ are more appropriate for the evaluation of social states than degrees of preference satisfaction, but bears on the elementary model of action too (see Sen, 1999).

There are also methodological doubts about the empirical standing of the model. What would falsify the claim that economic agents seek the most effective means to satisfy their preferences? Apparent counter-examples can always be dealt with by treating them as evidence that preferences have changed or been dismissed through a careful individuation of outcomes. Indeed, since preferences are unobservable, they can be identified only if the correctness of the model is presupposed. In other words, there is room for deeper dispute about the foundations of orthodox microeconomics than is always realized.

Even within economics there are critics. The most substantial attack comes from those who think that perfect information is not a useful limiting case of imperfect information. Granted that there is often no way of calculating the likely marginal costs and benefits of acquiring extra information (short of actually acquiring it), how shall the agent decide rationally when to stop? Simon (1976) uses the question to argue for ‘satisficing’ models, in place of maximizing ones, and for ‘procedural’ or ‘bounded’ rationality. Rationality, he suggests, is a matter of following a procedure that halts with a good solution, and should not be defined in terms of best solutions. While this is a tempting thought, it is not obvious that searching for a ‘good’ solution is any easier than the best one if ‘good’ is some kind of second-best version of the ‘best’. As a result, ‘behavioural economists’ have been drawn to the large experimental literature in psychology on how people actually behave and have produced economic models of decision-making that incorporate a variety of psychological processes such as ‘self-serving biases’, the ‘law of small numbers’ and ‘reference dependence’ (see Kahneman, 2003). In this way, *homo economicus* has become more psychologically complex and more of an institutional or organizational person than an abstract maximizer.

The rational expectations hypothesis offers a different approach to the information issue. A rational agent who is short of information should not use an information-generating mechanism that gives rise to systematic errors. If errors are systematic, the agent should be able to learn how to eliminate them by amending the mechanism. There is an incentive to do so, because improved estimates of future variables will be profitable. On the face of it this makes rational expectations the natural ally of the pure economic-man models. Economic Man can proceed much as before, in the assurance that inadequate information involves nothing more systematic than ‘white noise’ and with the benefit of fresh analytic results that flow from a rational expectations hypothesis.

But this is to sidestep the informational problem set earlier, unless one sees how rational agents will learn to remove systematic errors. When there are costs to learning then it may not be rational to expend the effort that achieves a rational expectation. If we set such costs aside, in some simple learning situations a Bayesian updating procedure turns a rational expectations-generating process into an approximation of adaptive expectations, which could be construed as a procedural rule of thumb. But no general rapprochement between maximizing and procedural models of rationality follows. In more

general learning situations the rational agent is trying to learn the rational expectations equilibrium relationship between variables – the one which, if used by agents to form their expectations, would reproduce itself in experience (white noise apart). This sounds easy, in that repeated experience of a particular relationship should lead to convergence on accurate parameter estimates. However, ignorance of the rational expectations equilibrium values produces behaviour that departs from those values. So observed values of variables embody a distortion which agents cannot correct without knowing the dimensions of their own ignorance. To know this, however, they would have to know the rational expectations equilibrium values already. To put it as the procedural critics might, learning would be feasible only if there were nothing to learn. The information question has been begged; and the door again opens on to psychology and its rich literature on what people actually do.

Nevertheless, the ideal-type Economic Man remains a powerful model of action not only in neoclassical theories, where insights in comparative statics have been especially notable, but elsewhere too. How powerful it finally is depends, within economics, on what becomes of the informational difficulties and on whether procedural or bounded models can come up with rival results of equal scope and elegance. For the wider social sciences, it offers a tempting analysis of social behaviour at large both for transactions in other social arenas and for the emergence of the institutions that govern those arenas. But the greater its ambitions, the more serious become the unresolved doubts about the origin of preferences and their relation to norms and institutions.

See Also

- altruism, history of the concept
- rational behaviour
- rationality, history of the concept
- utilitarianism and economic theory

Bibliography

Becker, G. 1976. *The Economic Approach to Human Behaviour*. Chicago: Chicago University Press.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper Row.

Kahneman, D. 2003. Maps of bounded rationality: psychology for behavioural economics. *American Economic Review* 93, 1449–75.

Sen, A. 1999. *Commodities and Capabilities*. Oxford: Oxford University Press.

Simon, H.A. 1976. From substantive to procedural rationality. In *Method and Appraisal in Economics*, ed. S. Latsis. Cambridge: Cambridge University Press.

How to cite this article

Heap, Shaun Hargreaves. "economic man." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000020> doi:10.1057/9780230226203.0438

The New Palgrave Dictionary of Economics Online

economic sanctions

Jeffrey J. Schott

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic sanctions are tools of statecraft used to achieve a broad range of foreign policy goals by threat or deployment of coercive measures such as trade embargoes, asset freezes, or withholding of development aid. Throughout the post-war era, the United States and other countries frequently have imposed economic sanctions, even though they have contributed only infrequently to foreign policy successes. Globalization has made the exercise of economic coercion increasingly complex, but has not obviated the utility of sanctions as part of the foreign policy arsenal.

Keywords

economic development; economic sanctions; globalization; terrorism

Article

Economic sanctions are tools of statecraft used to influence the behaviour of foreign countries by the threat or actual withdrawal of trade and sources of finance. Traditional means of coercion include trade embargoes, withholding development assistance, and asset freezes. The objective is to confront a foreign country with a choice: either bear the cost of lost trade and finance, or change policies to comply with the demands of those imposing the sanctions (the sender countries). Projecting power through economic coercion is deemed more forceful than diplomatic reproach yet less drastic than military intervention. In practice, economic measures generally are deployed as part of a broader programme of foreign policy responses encompassing diplomatic entreaties, covert or quasi-military intrusions, and threat of or preparation for military action.

Countries impose sanctions in pursuit of a variety of foreign policy goals. Historically, economic sanctions have preceded and then accompanied military conflict. The oil embargo of Japan was a prelude to the Second World War in the Pacific; so, too, were the United Nations' sanctions against Iraq following its invasion of Kuwait in 1990. Obviously, sanctions are part and parcel of 'hot wars' that sever economic ties between the combatants; but they are also prevalent in 'cold war' episodes, where

the goal is to impair military capabilities through denial of weapons and dual-use technologies (for example, post-war sanctions against the Soviet Union and its satellites under the auspices of the Consultative Group and Coordinating Committee for Multilateral Export Controls, or CoCom, and efforts to blunt the development of nuclear weapons in Iran and North Korea). In addition, sanctions have sought to impede or reverse military incursions across borders (for example, the League of Nations effort to get Italy to withdraw from Abyssinia in 1936) and between warring factions within a country (the sad recent history of several West African states).

Not all sanctions episodes respond to or presage military actions. Many post-war cases have been advanced to counter other types of aberrant behaviour such as state sponsored terrorism, proliferation of weapons of mass destruction, or human rights abuses. In these cases, sender countries impose sanctions in an effort to redress foreign outrages, to deter emulation by others (the rationale in most anti-proliferation cases), and to punish the target regime for its misdeeds (for example, the US grain embargo after the Soviet invasion of Afghanistan in 1989). In a number of cases, sanctions pursue the goal of regime change *sotto voce* – whether the target is Moammar Gaddafi in Libya, Kim Jong-il in North Korea, or the Afrikaaners in South Africa. Sanctions that portend regime change obviously meet stauncher resistance than those that seek narrow changes in governance by the target government.

Do sanctions ‘ work ’ ?

Foreign policy ventures seldom yield unambiguous results. Gauging the effectiveness of sanctions involves a combination of quantitative method and intuition, and often requires subjective evaluation of incomplete results. Sanctions alone seldom are sufficient to change foreign practices, but they can *contribute* to the achievement of policy goals in conjunction with other instruments of statecraft, if properly designed and implemented. That is easier said than done.

Sanctions are blunt policy instruments; they are better at impairing economic performance over time than at inflicting surgical strikes on target countries. Senders that expect immediate gratification often tire of the effort, especially if the sanctions impose significant costs on their own firms and workers. Moreover, when sanctions are hard hitting, it is difficult to avoid innocent victims within the target country and in neighbouring states; in such cases, the debilitating effect of sanctions often results in substantial suffering among the civilian population. Humanitarian exemptions from the sanctions designed to soften the blow to the general public invariably weaken the economic impact of the sanctions and muddy the policy signal to the target regime. To be sure, such loopholes in the sanctions net are important both on moral grounds and to maintain the cohesion of the coalition of sender countries, but the loopholes are prone to abuse (witness the scandalous operation of the United Nations’ oil-for-food programme, which was supposed to channel Iraqi oil export revenues to humanitarian assistance) and reduce the economic pressure to comply with the sender's demands.

Almost all sanctions leak; targeted countries can evade the full thrust of the economic restrictions by redirecting trade and finance to non-sanctioning states or by engaging in clandestine operations. Countries seeking economic or political influence with the target regime often conspire to evade the sanctions; the Cold War period was replete with examples of ‘Black Knight’ countries coming to the rescue of targeted regimes with aid to offset the impact of sanctions imposed by the United States or the Soviet Union. Smugglers still outwit even the most comprehensive embargoes – witness the billions of dollars earned by Saddam Hussein, the former Iraqi president, from illicit oil exports during the period

of ‘comprehensive’ UN sanctions against Iraq. For a price, targeted regimes can still procure goods, services and technologies; the profit motive seems to be an irresistible force regardless of region or culture!

That said, sanctions have contributed to a few notable successes in the post-war era, including the collapse of the apartheid regime in South Africa and the renunciation of terrorism by President Gaddafi in Libya. Hufbauer et al. (2007) found success – measured by the partial fulfillment or better of policy goals – in more than a quarter of the almost 200 sanctions episodes documented in the 20th century. (The third edition of this comprehensive study of economic sanctions contains updated policy analysis and case studies, and an extensive bibliography. See also Baldwin, 1985, for an examination of the tools of economic statecraft, and Martin, 1992, for analysis of the use of multilateral economic sanctions.)

Most of these cases, however, involved relatively modest demands on the target country. When the stakes are high, resistance by the target regime stiffens. Accordingly, most high-profile sanctions cases – like those seeking to oust President Castro in Cuba or to deter support for terrorism and the development of nuclear weapons by the ayatollahs in Iran – have been abject failures.

Can sanctions be effective in an era of rampant globalization?

Economic sanctions traditionally have been the domain of big powers, acting unilaterally or as part of a broader international coalition. Until recently, the big powers controlled the trade lanes and purse strings of international commerce, and held a near monopoly on advanced technologies. Since the mid-1980s, however, the success of post-war economic development, spurred in part by the spread of technological innovation, has eroded the franchise of the big powers and created alternative sources of goods, technology and capital for countries targeted by economic sanctions. Simply put, globalization has made it much harder to design an effective sanctions policy.

In addition, global politics are now more complex than in the period of East–West rivalry. Former allies differ regarding strategies and priorities for using sanctions to deal with regional trouble spots. For example, Europe is more vulnerable than the United States to an interruption of energy supplies from the Middle East, and thus is less willing to constrain oilfield development and to take actions that risk political retaliation. Similarly, China and Japan are highly dependent on imported energy and thus sensitive to sanctions against Iran and other oil-producing states.

Globalization also has contributed to the decentralization of power, allowing smaller countries – especially those rich in energy resources – to provide offsetting assistance to blunt the economic impact of sanctions. But the influence of globalization goes beyond the realm of state-to-state intervention; terrorism, for example, now operates in a stateless domain of sleeper cells and territories outside of governmental control linked through informal financial and telecommunications networks. For that reason, sanctions policies increasingly seek to target individuals and corporations as well as governmental bodies, and to favour financial measures to interdict inter-bank electronic transfers in addition to the more traditional controls on trade, investment and development assistance.

In sum, economic sanctions continue to play a major role in international relations. However, the familiar goals of economic coercion now must be pursued through measures adapted to the changing conditions in global markets. The use of economic sanctions needs to be reconsidered and revamped, but not abandoned.

See Also

- foreign aid
- trade policy, political economy of
- transfer of technology

Bibliography

Baldwin, D.A. 1985. *Economic Statecraft*. Princeton, NJ: Princeton University Press.

Hufbauer, G.C., Elliott, K., Schott, J.J. and Oegg, B. 2007. *Economic Sanctions Reconsidered*, 3rd edn. Washington, DC: Peterson Institute for International Economics.

Martin, L.L. 1992. *Coercive Cooperation: Explaining Multilateral Economic Sanctions*. Princeton, NJ: Princeton University Press.

How to cite this article

Schott, Jeffrey J. "economic sanctions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000306> doi:10.1057/9780230226203.0439

The New Palgrave Dictionary of Economics Online

economic sociology

Richard Swedberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘economic sociology’, used primarily by sociologists, is defined as the application of sociological concepts and methods of analysis to economic phenomena. Founded by Durkheim, Weber, and Simmel, and continued by Schumpeter and Polanyi, it began to flourish in the mid-1980s around the notion that economic actions are embedded in personal networks. The concept of networks and other concepts and perspectives from ‘new economic sociology’ facilitate the analysis of topics like the links between corporations and between firms, job search, production markets, finance markets, insurance markets, industrial markets, consumption, and ethnic entrepreneurship. Its long-term impact on economics remains uncertain.

Keywords

Becker, G.; Capitalism; Weber on; Polanyi on; Coase, R.; division of labour; Smith vs Durkheim on; Durkheim, E.; economic sociology; emotions; entrepreneurship; ethnic; and immigration; forced; Granovetter, M.; Immigration; and ethnic entrepreneurship; Jevons, S.; Marx, K.; networks; strong vs weak ties; and job search; and consumption; and groups of firms; New Institutional Economics; Parsons, T.; Polanyi, K.; on capitalism; Schumpeter, J.; on social economics;; Simmel, G.; on trust; Smith, A.; trust; Simmel on; Weber, M.; on capitalism

Article

The first recorded use of the term ‘economic sociology’ is in a 1879 work by Stanley W. Jevons; and it is clear from the context that Jevons viewed economic sociology as part of the overall enterprise of economics rather than as an area belonging to another social science, such as sociology. Today, in contrast, the term ‘economic sociology’ is used primarily by sociologists, and they define it as *the application of sociological concepts and methods of analysis to economic phenomena*. While it is definitely possible to treat the great concern with institutions in New Institutional Economics, for

example, as a kind of economic sociology, the reader is referred to the entry for this topic for this type of analysis. Similarly, while Gary Becker at times has referred to his extension of the economic model to non-economic topics as ‘economic sociology’, the reader is similarly referred to the entry for his work. Here, the first section, on classical economic sociology, is followed by sections on more recent economic sociology. This way of proceeding not only follows the general development of the field of economic sociology but is often how economic sociology is taught today, since the classics play a somewhat different role in economic sociology (as in sociology itself) to that in economics. In brief, while sociologists are trained through work with the classics as well as modern material, today's economists read the classics primarily when they study the history of their discipline.

Classical economic sociology

The work of Karl Marx (1818–83) can be seen as a type of economic sociology, in the sense just mentioned. More generally, Marx closely linked classical economic categories, such as value, price and capital, to distinctly social categories, such as class, work and relations of production. Nevertheless, Marx has played a marginal role in economic sociology as an academic enterprise – except as a catalyst and inspiration for a number of scholars, including Max Weber and Joseph Schumpeter.

Modern academic sociology is generally regarded as having three founders – Max Weber, Emile Durkheim and Georg Simmel – all of whom were interested in the economy. Georg Simmel (1858–1918), who pioneered sociology in Germany, wrote on the sociological role of money, competition and trust in the economy (Simmel, 1900; 1908). He closely linked different types of money to different types of social authority, and also attempted to show how money is linked to the element of relativism in modern society. Competition, he argued, releases the energy of all participants to the benefit of the public, whereas in a conflict combatants are pitted against each other and block each other's efforts. Trust, finally, is central to the economy as well as society at large; without trust, the economy as well as society would collapse.

Emile Durkheim (1858–1917), unlike Simmel, attempted to institutionalize economic sociology, partly by encouraging some of his students to specialize in this field. Durkheim's own most important contribution to economic sociology can be found in his doctoral study of the division of labour, which contains a sharp critique of the argument in Adam Smith's *The Wealth of Nations* (Durkheim, 1893). According to Durkheim, while Adam Smith had seen the significance of division of labour exclusively from the perspective of the creation of wealth, he had neglected its importance for the cohesion of society. More precisely, Smith had failed to realize that the primary function of the division of labour in modern society is to tie people together: people who do very different things need each other, and this is also what gives cohesion to modern society.

The most sustained effort to lay a solid theoretical foundation for economic sociology and also to carry out empirical studies can be found in the work of Max Weber (1864–1920) (Swedberg, 1998). While Weber is famous for *The Protestant Ethic and the Spirit of Capitalism* (1905), it is less well known that his work is part of a more general attempt to develop a new academic field that would complement economic history and economic theory, namely, economic sociology.

At first Weber carried out empirical and historical studies with this goal in mind, and of these *The Protestant Ethic* is by far the best known (but see also Weber, 1909; 1895). Weber's thesis, which holds

that a certain type of religion (‘ascetic Protestantism’) had helped to create the mentality of modern capitalism in the 16th and 17th centuries (‘rational capitalism’; Weber, 1905), has led to a heated debate. Most commentators have found Weber's thesis unconvincing, but it should be emphasized that the debate is still going on with as much fervour as in the early 20th century (see, for example, Marshall, 1982).

The heart of Weber's economic sociology is to be found in *Economy and Society*, a work that was incomplete when Weber died. It is here, for example, that Weber set out his well-known typology of capitalism: political capitalism, traditional capitalism and rational capitalism. While the former two have existed for thousands of years, rational capitalism has emerged only in modern times and in the West. While traditional capitalism is non-dynamic and centred around small enterprises involving trade and the exchange of money, political capitalism is profit-making that either takes place through the state or under its direct protection, as in imperialism. Rational capitalism, in contrast, gets its name from the strong element of conscious and methodical calculation: the activities of the firm are carried out with the help of accountants and a trained staff; similarly, the activities of the state bureaucracy (including in the legal system) are predictable and rational. All of this makes possible a truly dynamic and revolutionary form of capitalism, according to Weber.

Economy and Society also contains a serious attempt by Weber to develop the central theoretical categories of economic sociology (Weber, 1914, pp. 63–211). The basic unit of analysis is ‘economic social action’, which differs from economic action in economic theory by partly being determined by its social dimension. Economic social action is defined by Weber as behaviour that is (*a*) invested with meaning, (*b*) aimed at utility and (*c*) *oriented to another actor*. Utility is what makes the action ‘economic’; and Weber's definition of ‘social’ is to be found in the formula ‘orientation to another actor’. The emphasis on meaning explains why Weber's sociology is called an interpretive sociology; his economic sociology was to be a form of *interpretive economic sociology*.

Weber then proceeds to economic relationships in which two actors orient their actions to one another. These relationships can be either open or closed; and there is a general tendency for open economic relationships to become closed when there are not enough resources to go round. Economic organizations are defined as closed social relationships of a certain type; there also has to be a staff. Economic systems, finally, can be oriented either to profit-making (as in capitalism) or to the provision for a household (as in socialism or earlier non-market economies). Weber also discusses a host of other topics, including trade, money, division of labour and different ways of appropriation.

After the classics

While the founding fathers of sociology were all interested in economic sociology and promoted it, the topic did not become popular among sociologists until the mid-1980s with the emergence of so-called ‘new economic sociology’. The reason for this is not clear, but may well have been a strong sense among sociologists that the economists were better equipped to deal with economic topics. In any case, very little work on economic sociology was produced between 1920 and the mid-1980s.

There were, however, a few exceptions. For one thing, sociologists did discuss topics relating to the economy, even if they did so under labels other than ‘economic sociology’. One example is industrial sociology, which saw as its main task to analyse situations when people work in groups, in the factory as

well as the office. An important research result is that workers develop norms in a number of areas, including what is seen as the maximum effort. Those who breach these norms are punished (for example, Whyte, 1955).

Three individuals who all made important contributions to economic sociology also appeared during the period after the classics: Joseph Schumpeter, Karl Polanyi and Talcott Parsons. According to Schumpeter (1885–1950), economics should be a broad science (‘social economics’) and encompass four areas: economic theory, economic history, economic statistics and economic sociology (Schumpeter, 1954, pp. 12–24). Schumpeter did work in each of these fields, including economic sociology. According to Schumpeter, economic sociology deals with institutions, while economic theory deals with economic mechanisms. Schumpeter's three most famous essays in economic sociology deal with the issues of social class in economic life, the role of taxation (‘fiscal sociology’) and imperialism (Schumpeter, 1991). Schumpeter thought highly of these essays and they are all considered minor classics today.

But one can also find elements of economic sociology in some of Schumpeter's non-sociological writings. This goes for the famous analysis of entrepreneurship in *Theory of Economic Development*, not least the element of resistance from the environment that the entrepreneur usually confronts (Schumpeter, 1934). Similarly in *Capitalism, Socialism and Democracy*, we find a sociological portrait of contemporary capitalism. The US economy was doing very well, according to Schumpeter, but its institutions were decaying (Schumpeter, 1942).

Like Schumpeter, Karl Polanyi (1886–1964) came from the Austro-Hungarian Empire and ended his life on the American continent. Like Schumpeter, he wrote a famous book on capitalism – *The Great Transformation* – and contributed to the economic sociology of his days (Polanyi, 1957). It is to Polanyi that we owe the term ‘embeddedness’, even if he used it in his own, very political sense: all economies had been embedded in politics and religion before the advent of capitalism, and were disembedded by the traumatic ‘great transformation’. The political task of the day, in other words, was to re-embed the economy into political and human values.

Polanyi covered historical distances with great ease and was as much at home in ancient Babylonia as in 19th-century Britain or 20th-century United States. The scope of his knowledge about the economy is also reflected in one of his most useful sets of categories: the concepts of reciprocity, redistribution and exchange (for example, Polanyi, 1971). In a kinship situation, for example, reciprocity may be used as a way of distributing resources. A political centre, like the state, would in contrast redistribute resources; and a market distributes resources through exchange. Most economic systems draw on each of these three ways of distributing resources, with their corporate sectors (‘exchange’), state sectors (‘redistribution’) and household sectors (‘reciprocity’).

Talcott Parsons (1902–1979) had begun his career as an economist, only to switch to sociology, since he thought that utilitarian thought was unable to properly capture the structure of modern society. Parsons argued for a general systems perspective in social theory, and suggested in *Economy and Society* (together with Neil Smelser) that the economy should be conceptualized as a sub-system of the general system of society (Parsons and Smelser, 1956). Just as each society has to have a distinct goal (‘Polity’) and a value-system (‘Latent-Pattern-Maintenance’), it also has to adapt to nature and reality (‘Economy’). While it is part of society, the economy is also its own society, with a ‘polity’, ‘latent-pattern-maintenance’, and so on.

New economic sociology

Around the mid-1980s American sociologists suddenly started to become interested in economic sociology, and it is this development that is generally known as ‘new economic sociology’. One article in particular operated as a catalyst in this process, and that is Mark Granovetter's ‘Economic action and social structure: the problem of embeddedness’ (1985). Its central argument is that all economic actions are embedded in personal networks, and it is this quality that brings them into the sociologist's domain. While this message was important enough in itself, the article's implicit or subliminal message that sociology had neglected a whole area of social life which lent itself to sociological analysis, namely, the economy, also explains its great impact. Since sociological skills had not been applied to economic problems, sociologists might also be able to solve a number of important puzzles that the economists had failed to do, according to Granovetter.

Since the mid-1980s economic sociology has advanced steadily, and it is now fully institutionalized in the United States. It is routinely taught in sociology departments in all the major universities and also has a strong presence among the major journals of the profession. The American Sociological Association has a special section for economic sociology; a number of readers have been published as well as a huge handbook (Smelser and Swedberg, 1994; 2005).

Economic sociology is becoming increasingly popular and accepted in Europe as well, though in a somewhat different form than in the United States, which is only natural given the various national traditions in sociology. While interesting contributions can be found in many European countries, it is especially in France that one can find highly original contributions that stand up well to international competition (for England, see for example Dodd, 1994; for Scotland, MacKenzie, 2003; for Germany, Beckert, 2004; for Italy, Trigilia, 2002; and for Sweden, Aspers, 2001).

The three key figures in French economic sociology are Pierre Bourdieu, Luc Boltanski and Michel Callon (see also the works of Lebaron, 2000, and Steiner, 2005). Bourdieu (1930–2002) has, among other things, analysed consumption in an innovative manner in his celebrated study *Distinction* (1986); he has also sketched a whole programme for economic sociology, drawing on his three key concepts of habitus, field, and different types of capitals (Bourdieu, 1979; 2005). Luc Boltanski has contributed to the discussion of modern capitalism through an important study of class formation and also co-authored a provocative volume on ‘the new spirit of capitalism’ (Boltanski, 1987; Boltanski and Chiapello, 1999). And Michel Callon (1998) has introduced the so-called theory of performativity or the idea that economic theory may be as successful as an explanatory approach for the simple reason that it analyses phenomena that it has helped to create in the first place.

The number of studies in economic sociology (books and articles) amounts to several thousand by now, which makes it hard to summarize its achievements. One way to convey a sense of this literature, however, would be to discuss the methods that are being used to gather and analyse data as well as some of the most important topics. That economic sociology indeed has a distinctive profile that sets it off from mainstream economics emerges very clearly from a discussion of these two themes.

The data that is being used in economic sociology has often been put together by the analyst, and it is considerably less common than in mainstream economics to draw on official data of the type that is produced by government agencies. One example is historical studies in economic sociology, as

illustrated by Bruce Carruther's *City of Capital* (1996). The focus in this work is the emergence of one of the world's first financial markets, and the author draws heavily on various primary and secondary sources. In particular, Carruthers succeeds in showing that early trade in shares often followed party lines; that is, sellers were reluctant to trade with political opponents.

Comparative studies are long-standing in economic sociology and have also been popular in new economic sociology. In one of these, *Forging Industrial Policy*, Frank Dobbin (1994) compares the ways in which the railroad industry developed in the 19th century in the United States, Britain and France. The author shows that industrial policy has largely mirrored the general political culture in its approach to solving problems in each of these three countries. In the United States, there has been scepticism towards the state and reliance on the corporations; in France, the state has been the central actor; and in Britain there has been an attempt to protect the individual firm from competition as well as from interventions from the state. Dobbin claims to have found that there is no one best way of doing things. Rather, people generalize from how they themselves do things and proclaim this to be the universally rational way to proceed.

Economic sociologists also draw on ethnography and participant observation, two methods that allow the researcher to handle huge amounts of empirical detail and to approach things from the perspective of the actors. Michael Burawoy (1979), for example, worked as a shop steward in order to better understand how workers interact and deal with the demands of their work (especially boredom); and Mitchel Abolafia (1996; 1998) passed an examination as a stockbroker in order to better understand what goes on in various stock and bond exchanges.

By far the most significant single method used by economic sociologists today, however, is that of networks. This is a very flexible tool, which allows for quantification and therefore goes well with a large number of research tasks. It has been used, for example, to analyse the links that exist between corporations by virtue of having the same individual on their boards (so-called interlocks). Through the resultant system of communication, various ways of doing things may be diffused. The so-called poison pill (a measure against hostile takeovers) has, for example, been shown to diffuse quickly among corporations linked by common board members (Davis, 1991). That links between corporations are not to be understood exclusively in terms of instrumental actions may be exemplified by the fact that, when a board member resigns or dies, he or she is only replaced in something like half of the cases (Palmer, 1983).

Using networks is also a popular way in economic sociology to approach collaboration between corporations as well as the relationship between firms and their customers and suppliers (see, for example, Gulati and Gargiulo, 1999). The area where it has been most successful, however, may well be the labour market; and here the classic study is Mark Granovetter's *Getting a Job* (Granovetter, 1974). While one may have thought that the most important source of assistance for a person seeking a job is that's person's closest friends and family (‘strong ties’), in fact it is his or her more casual contacts (‘weak ties’), whose number depends on how many jobs a person has had. The reason for this ‘strength of weak ties’ is simply that, whereas one's ‘strong ties’ all share the same information, ‘weak ties’ can provide access to new and varied information, including information about job opportunities.

In European economic sociology an attempt has also been made to expand the notion of networks to include not only people and organizations in the category of actors but also objects (so-called actor-network-theory; see, for example, Law and Hassard, 1999). That objects can be actors in the

conventional sense of this term is no doubt wrong; the weaker claim that objects can be part of networks is, however, more interesting. One may, for example, see a machine as a link between people, some objects may be used for communication between people, and so on – and all this can affect the structure of the network. More generally, the advocates of actor-network-theory also argue that the traditional approach of economists and sociologists tends totally to ignore the role that objects play in the economy and to focus exclusively on actions, social relations and the like. The perspective that argues for including objects in the analysis is usually referred to as ‘materiality’.

When it comes to the topics that are often analysed, new economic sociologists have first and foremost tried to focus on economic institutions as opposed to phenomena situated at the boundary of, say, religion and the economy or politics and the economy. The reason for this has been a desire to take on truly ‘economic’ topics and go beyond the old division of labour between economics and sociology, when the former dealt with the economy and the latter with society minus the economy. As examples of this is the interest among contemporary economic sociologists in markets and corporations, which have attracted a large number of studies.

One type of study has attempted to develop a general model for markets that differs sharply from the standard economic model of the perfect market. The most prominent example of this is the work of Harrison White (1981; 2002) on so-called production markets, by which he roughly means industrial markets. Production markets, it is argued, differ from so-called exchange markets primarily because their participants have permanent roles as either sellers or buyers and do not switch between these two roles as is common in financial markets.

According to White, the typical production market holds about a dozen actors who closely follow what the other actors are up to. Markets come into being, White argues, precisely because economic actors position themselves in relation to the products of other actors. Prices are not set through demand and supply but by producers relating the revenue of their goods to the volume that is being sold. Individual markets, finally, are connected to each other in giant networks, either ‘upstreams’ (suppliers) or ‘downstreams’ (customers).

A number of studies of financial markets have also been carried out, and here the work of Donald MacKenzie is outstanding (for example, MacKenzie, 2003; MacKenzie and Millo, 2003). MacKenzie has picked up from Callon the theme of performativity, and he uses it, for example, in his analysis of trade in options. The pricing of options was very difficult, the argument goes, until Black, Scholes and Merton suggested a solution for which the latter two would win the Nobel Prize in 1997. While this formula covers most cases with much precision, according to MacKenzie it does not cover all – and this was to have important consequences. Since this fact was not well understood, however, and since economic reality was mistaken for how it was portrayed in finance theory (performativity), there have been cases in which people were unprepared for what was happening (as in the case of Long-Term Capital Management). MacKenzie traces this development and also shows how actors have tried to protect themselves against exceptional cases by keeping a margin against the price predicted according to the Black–Scholes–Merton formula.

Economic sociologists have suggested several new ways to approach consumer markets. Viviana Zelizer (1979), for example, has analysed the growth of the market in life insurance in the United States and shown how the idea of putting a price on a human life initially attracted hostility, for religious reasons. But as people moved into the cities and religion had to adjust to new circumstances, a different view of life insurance emerged. Zelizer has recently also started to look at consumption among children, both

how children are socialized into becoming consumers and the ways in which they themselves relate to objects and goods in their environment (Zelizer, 2005).

DiMaggio and Louch (1998) have attempted to use networks to analyse consumption. While it is well known that people will turn to others in their surroundings to find out where to buy something, and which merchants, traders and so on are reliable (‘search embeddedness’), DiMaggio and Louch examine situations in which people approach someone in their personal network in order to buy something (‘within-networks exchange’). As it happens, this is quite common, especially infrequent purchases of the type that involve legal services, home repair maintenance and the buying of a car or a home.

The number of studies in economic sociology that deal with corporations is very great, but a few studies nonetheless stand out. One of these is Mark Granovetter's pioneering 1994 article on business groups. Against R.H. Coase, Granovetter argues that it is not so much the existence of the individual firm that needs to be explained but the common phenomenon of groups of firms. In many countries, such as India, South Korea and Japan, these business groups control large parts of the economy, but have not received the scholarly attention that they deserve. The impact of business groups in the United States is not clear from Granovetter's work, except that US antitrust legislation has ruled out some common forms of this phenomenon.

The business groups that Granovetter studies lend themselves to a networks approach, and so do the corporations that Ronald Burt (1983) has analysed in his study of US industrial markets. Each firm, according to Burt, can be conceptualized as situated at the centre of a network in which there are a number of competitors, suppliers and customers. The fewer competitors there are, the more suppliers, and the more customers, the more the corporation is characterized by ‘structural independence’. And with more structural independence comes more profit, as Burt shows.

The emphasis on corporations in interaction, as opposed to the single corporation, is also obvious in another landmark study in economic sociology, *Regional Advantage* by AnnaLee Saxenian (1994). Following Alfred Marshall in analysing industrial districts, Saxenian carries out a comparative study of the computer industry during the post-war period in Silicon Valley and the area around Route 128 in Boston. Silicon Valley has clearly overtaken Route 128 during recent decades, and the reason for this, according to Saxenian, has to do with the nature of the interaction in the two regions. While in Route 128 the corporations are loath to cooperate, rely on banks for finance, and prosecute employees who switch to competitors, in Silicon Valley there is plenty of cooperation, finance comes from venture capital firms, and employees are free to switch as they like. A much more decentralized and flexible form of entrepreneurship, in brief, has emerged in Silicon Valley.

Saxenian's fascination with entrepreneurship is shared by many economic sociologists. While she argues that a radical decentralized industrial region represents the best conditions for entrepreneurship, there exist other perspectives as well. Granovetter, for example, argues that entrepreneurs often come from those parts of the social system which are far away from the controlling centre (for example, Granovetter, 2005). While this may be termed a theory of peripheral entrepreneurship, Granovetter suggests several other situations that are favorable to entrepreneurship. An entrepreneur may, for example, be someone who crosses a social boundary in society and thereby becomes the first to unite resources from two otherwise separated regions (for example, Granovetter, 1995). On immigration, Granovetter also points out that some ethnic groups that are not entrepreneurial in their country of origin may be highly entrepreneurial in their new country because they often leave parts of the extended family

behind (Granovetter, 1995). This means that they do not have to provide jobs for their relations or share their wealth with relatives.

Economic sociologists have been very active in studying ethnic entrepreneurship, (for example, Light, 2005). Ethnic entrepreneurs, for example, often have to overcome the fact that their initial market consists of their countrymen ('the ethnic market'), and that they will have to go beyond this market if they are to expand. In many cases they have become entrepreneurs simply because they have no other way of making a living ('forced entrepreneurship').

Economic sociologists have also emphasized the collective nature of entrepreneurship and attempted to explode the myth of the creative Schumpeterian individual. One important example of this can be found in the research by Rosabeth Moss Kanter (1983) on entrepreneurship within the corporation, so-called intrapreneurship. Through a combination of ethnographic studies and survey research, Kanter has attempted to show the conditions under which it is possible to put together creative and entrepreneurial groups in modern corporations. Someone has to suggest the creation of such groups and provide them with resources and legitimacy. The group also has to be defended from outside intervention while it operates, internal conflicts have to be solved, and so on. According to Kanter, this type of group is common among modern corporations.

While economic sociologists have been unable to present a general theory of entrepreneurship, it is nonetheless clear that a number of insights have been accumulated. Economic sociologists are also expanding their work into such topics as social entrepreneurship and the diffusion of courses among business schools (for example, Swedberg, 2000).

Concluding remarks

Economic sociology is currently in a very active phase of its development, and all signs indicate that this trend will continue. Economic sociologists are also gradually expanding their range of topics of study. There has recently, for example, been an attempt to introduce law into the analysis, and some economic sociologists are trying to formulate a position on the relationship between the economy and technology. Some economic sociologists are also in the process of investigating the role of emotions in the economy; and there is a growing number of studies of gender and the economy. What all of this adds up to, again, is a steady growth of studies in economic sociology and a confirmation that economic sociology is established as a distinct and accepted area of sociology. But it remains to be seen whether economic sociology will be able to make inroads into economics itself and gain respect from economists, along the lines of, say, behavioural economics.

See Also

- Akerlof, George Arthur
- cartels
- entrepreneurship

Bibliography

Abolafia, M. 1996. *Making Markets: Opportunism and Restraint on Wall Street*. Cambridge, MA: Harvard University Press.

Abolafia, M. 1998. Markets as culture: an ethnographic approach. In *The Laws of the Markets*, ed. M. Callon. Oxford: Blackwell.

Aspers, P. 2001. *A Market in Vogue: A Study of Fashion Photography in Sweden*. Stockholm: City University Press.

Beckert, J. 2004. *Unverdients Vermögen. Soziologie des Erbrechtes*. Frankfurt: Campus Verlag.

Boltanski, L. 1987. *The Making of a Class: Cadres in French Society*. Cambridge: Cambridge University Press.

Boltanski, L. and Chiapello, E. 1999. *Le Nouvel Esprit du Capitalisme*. Paris: Gallimard.

Bourdieu, P. 1979. Algeria 1960. Cambridge: Cambridge University Press.

Bourdieu, P. 1986. *Distinction: A Social Critique of the Judgment of Taste*. London: Routledge.

Bourdieu, P. 2005. Principles of an economic anthropology. In *The Handbook of Economic Sociology*, 2nd edn., ed. N. Smelser and R. Swedberg. Princeton: Princeton University Press.

Burawoy, M. 1979. *Manufacturing Consent: Changes in the Labor Process under Monopoly Capitalism*. Chicago: University of Chicago Press.

Burt, R. 1983. *Corporate Profits and Cooptation: Networks of Market Constraints and Directorate Ties in the American Economy*. New York: Academic Press.

Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Callon, M., ed. 1998. *The Laws of the Markets*. Oxford: Blackwell.

Carruthers, B. 1996. *City of Capital: Politics and Markets in the English Financial Revolution*. Princeton, NJ: Princeton University Press.

Davis, G. 1991. Agents without principles? The spread of the poison pill throughout the intercorporate network. *Administrative Science Quarterly* 36, 583–613.

DiMaggio, P. and Louch, H. 1998. Socially embedded consumer transactions: for what kind of

purchases do people most often use networks? *American Sociological Review* 63, 619–37.

Dobbin, F. 1994. *Forging Industrial Policy: The United States, Britain, and France in the Railway Age*. New York: Cambridge University Press.

Dodd, N. 1994. *The Sociology of Money: Economics, Reason and Contemporary Society*. Cambridge: Polity Press.

Durkheim, E. 1893. *The Division of Labor in Society*. New York: Free Press, 1984.

Granovetter, M. 1974. *Getting a Job: A Study of Contacts and Careers*. Cambridge. MA: Harvard University Press.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* 91, 481–510.

Granovetter, M. 1994. Business groups. In *The Handbook of Economic Sociology*, ed. N. Smelser and R. Swedberg. Princeton, NJ: Princeton University Press.

Granovetter, M. 1995. The economic sociology of firms and entrepreneurship. In *The Economic Sociology of Immigration*, ed. A. Portes. New York: Russell Sage Foundation.

Granovetter, M. 2005. The impact of social structure on economic outcomes. *Journal of Economic Perspectives* 19(1), 33–50.

Gulati, R. and Gargiulo, M. 1999. Where do interorganizational networks come from? *American Journal of Sociology* 104, 1439–93.

Jevons, S. 1879. *The Theory of Political Economy*, 5th edn. New York: Augustus M. Kelley, 1965.

Kanter, R. 1983. *The Change Masters: Innovation and Entrepreneurship in America*. New York: Simon and Schuster.

Law, J. and Hassard, J., eds. 1999. *Actor Network Theory and After*. Oxford: Blackwell.

Lebaron, F. 2000. *La Croyance Economique: Les Economistes entre Science et Politique*. Paris: Seuil.

Light, I. 2005. The ethnic economy. In *The Handbook of Economic Sociology*, 2nd edn., ed. N. Smelser and R. Swedberg. Princeton, NJ: Princeton University Press.

MacKenzie, D. 2003. Long-term capital management and the sociology of arbitrage. *Economy and*

Society 32, 349–80.

MacKenzie, D. and Millo, Y. 2003. Constructing a market, performing theory: the historical sociology of a financial derivatives exchange. *American Journal of Sociology* 109, 107–45.

Marshall, G. 1982. *In Search of the Spirit of Capitalism: An Essay on Max Weber's Protestant Ethic Thesis*. London: Hutchinson.

Palmer, D. 1983. Broken ties: interlocking directorates and intercorporate coordination. *Administrative Science Quarterly* 28, 40–55.

Parsons, T. and Smelser, N. 1956. *Economy and Society: A Study in the Integration of Economic and Social Theory*. New York: The Free Press.

Polanyi, K. 1957. *The Great Transformation*. Boston: Beacon Hill.

Polanyi, K. 1971. The economy as instituted process. In *Trade and Market in the Early Empires*, ed. K. Polanyi, C. Arensberg and H. Pearson. Chicago, IL: Henry Regnery.

Saxenian, A. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.

Schumpeter, J. 1934. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Schumpeter, J. 1954. *History of Economic Analysis*. London: Allen & Unwin.

Schumpeter, J. 1991. *The Economics and Sociology of Capitalism*, ed. R. Swedberg. Princeton, NJ: Princeton University Press.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. London: Routledge, 1994.

Simmel, G. 1900. *The Philosophy of Money*. London: Routledge, 1978.

Simmel, G. 1908. Competition. In *Conflict and the Web of Group-Affiliation*. New York: The Free Press, 1955.

Smelser, N. and Swedberg, R., eds. 1994. *The Handbook of Economic Sociology*. Princeton, NJ: Princeton University Press.

Smelser, N. and Swedberg, R., eds. 2005. *The Handbook of Economic Sociology*, 2nd edn. Princeton, NJ: Princeton University Press.

Steiner, P. 2005. *L'Ecole Durkheimienne et l'Economie*. Geneva: Droz.

Swedberg, R. 1998. *Max Weber and the Idea of Economic Sociology*. Princeton, NJ: Princeton University Press.

Swedberg, R. 2000. The social science view of entrepreneurship. In *Entrepreneurship: The Social Science View*, ed. R. Swedberg. Oxford: Oxford University Press.

Trigilia, C. 2002. *Economic Sociology: State, Market and Society in Modern Capitalism*. Oxford: Blackwell.

Weber, M. 1895. The national state and economic policy (Freiburg Address). *Economy and Society* 9, 428–49 (1980).

Weber, M. 1905. *The Protestant Ethic and the Spirit of Capitalism*. New York: Charles Scribner's Sons, 1958.

Weber, M. 1909. *The Agrarian Sociology of Ancient Civilizations*. London: New Left Books, 1976.

Weber, M. 1914. *Economy and Society: An Outline of Interpretive Sociology*. Berkeley: University of California Press, 1978.

White, H. 1981. Where do markets come from? *American Journal of Sociology* 87, 517–47.

White, H. 2002. *Markets from Networks: Socioeconomic Models of Production*. Princeton, NJ: Princeton University Press.

Whyte, W. 1955. *Money and Motivation*. New York: Harper.

Zelizer, V. 1979. *Morals and Markets: The Development of Life Insurance in the United States*. New York: Columbia University Press.

Zelizer, V. 2005. Culture and consumption. In *The Handbook of Economic Sociology*, 2nd edn., ed. N. Smelser and R. Swedberg. Princeton, NJ: Princeton University Press.

How to cite this article

Swedberg, Richard. "economic sociology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_E000227> doi:10.1057/9780230226203.0440

The New Palgrave Dictionary of Economics Online

economic surplus and the equimarginal principle

Maurice Allais
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Allais, M.; economic surplus; economy of markets; equimarginal principle; general equilibrium; marginalism; marginal equivalence; marginal rate of substitution; maximum efficiency; preference index; production functions; surplus

Article

Marginal analysis is actually only a particular case of a more general theory, the theory of surpluses and the economy of markets, which, if considered first, facilitates the discussion of the equimarginal principle.

The general theory of surpluses and the economy of markets – fundamental concepts and theorems

To simplify the exposition, it is assumed that one good (*U*), enters all preference and production functions, and that its quantity can vary continuously. Except for the hypothesis of continuity with respect to this good (*U*), the discussion in this first part is free of any restrictive hypothesis of continuity, differentiability or convexity for the goods (*V*),…,(*W*) considered, and the preference indexes and production functions. (For an exposition of the following theory in the case where no one good plays a particular role, see Allais, 1985, Section II, pp. 139–41.)

Structural conditions

The needs of every unit of consumption, individual or collective, can be entirely defined by considering a preference index

$$I_i = f_i(U_i, V_i, \ldots, W_i) \quad (1)$$

increasing as it passes from a given situation to one it finds preferable. Every quantity V_i is counted positively if it refers to a consumption, negatively if it refers to a service supplied. The set of feasible techniques for a unit of production *j* can be represented by a condition of the form

$$f_j(U_j, V_j, \ldots, W_j) \geq 0$$

where every quantity V_j is considered as representing a consumption or an output depending on whether it is positive or negative. The extreme points corresponding to the boundary between possible and impossible situations represent states of maximum efficiency for the production unit considered. They may be represented by the condition

$$f_j(U_j, V_j, \ldots, W_j) = 0. \tag{2}$$

The function f_j may be called the production function. It is defined up to any transformation which leaves its sign unchanged.
From a technical point of view, maximum efficiency implies quite specific conditions. If, for instance, one considers a production technique $A=A(X, Y,\ldots, Z)$ and if n production units are technically preferable to a single one, we should have (Allais, 1943, pp. 187–8; 1981, pp. 319–22)

$$\sum_j A(X_j, Y_j, \ldots, Z_j) > A\left[\sum_j X_j, \sum_j Y_j, \ldots, \sum_j Z_j\right]. \tag{3}$$

In the opposite case we have

$$A\left[\sum_j X_j, \sum_j Y_j, \ldots, \sum_j Z_j\right] > \sum_j A(X_j, Y_j, \ldots, Z_j). \tag{3*}$$

An industry is referred to as differentiated if the use of distinct production units is technically more advantageous than the concentration of all production operations into a single production unit. It is called non-differentiated in the opposite case. Conditions (3) and (3*) are two particular illustrations of differentiation (Allais, 1943, p. 637).
From inequality (3) it is possible to show that the whole production function of a differentiated industry is asymptotically homogeneous. In this case $(n \gg 1)$ there is quasihomogeneity (Allais, 1943, pp. 201–6; 1974b).

Distributable Surplus Corresponding to a Given Modification of the Economy

The distributable surplus σ_u relative to a good (U) and to a realizable modification of the economy which leaves all preference indexes unchanged is defined as the quantity of that good which can be released following this shift (Allais, 1943, pp. 610–16). The surplus considered here differs essentially from the concepts of consumer surplus as normally considered in the literature (for example, Samuelson, 1947, pp. 195–202; Blaug, 1985, pp. 355–70; Allais, 1981, pp. 297–8, and 1985, nn. 12–13).
Let us consider an initial state (ξ_1) characterized by consumption values $U_i, V_I,\ldots,W_i$ and $U_j, V_j,\ldots,W_j$ (positive or negative) of the different units of consumption and production. We have

$$\sum_i U_i + \sum_j U_j = U_0;$$
$$\sum_i V_i + \sum_j V_j = V_0; \ldots; \sum_i W_i + \sum_j W_j = W_0$$
(4)

where U_0, V_0,…, W_0 designate available resources. Let $(\delta\xi_1)$ be a feasible modification of (ξ_1) characterized by finite variations $\delta U_i, \delta V_i, \ldots, \delta W_i, \delta U_j, \delta V_j, \ldots, \delta W_j$, and let

$$(\xi_2) = (\xi_1) + \delta(\xi_1)$$

represent the new state.
According to (4) we naturally have

$$\sum_i \delta V_i + \sum_j \delta V_j = 0$$

for every good (U), (V),…, (W). From (2) we also have for every unit of production j

$$f_j(U_j + \delta U_j, V_j + \delta V_j, \ldots, W_j + \delta W_j) = 0.$$

According to (1) the preference indexes become

$$I_i + \delta I_i = f_i(U_i + \delta U_i, V_i + \delta V_i, \ldots, W_i + \delta W_i).$$

The δI_i can be positive, zero, or negative.
Let us now define a third state (ξ_3) by the condition that by the modification $-\delta\sigma_{ui}$ of just the quantities $U_i + \delta U_i$ all the preference indexes return to their initial values.
We then have the conditions

$$f_i(U_i + \delta U_i - \delta\sigma_{ui}, V_i + \delta V_i, \ldots, W_i + \delta W_i) = f_i(U_i, V_i, \ldots, W_i).$$
(5)

The state (ξ_3) can be termed ‘isohedonous’ with the state (ξ_1). In passing from (ξ_1) to (ξ_3) the quantity

$$\delta\sigma_u = \sum_i \delta\sigma_{ui} \tag{6}$$

of the good (U) is released, as all the units of consumption find themselves again in situations which they consider equivalent, since their preference indexes return to the same values (Allais, 1943, pp. 637–8).

The surplus $\delta\sigma_u$ has been released during the passage from (ξ_1) to (ξ_3). It may then be considered that in the situation (ξ_1) this surplus was both realizable and distributable. It may further be considered that in passing from (ξ_1) to (ξ_2), it has in effect been distributed.

The distributable surplus thus defined covers the whole economy, but this definition can be used for any group of agents. It is necessary only to consider the functions f_i and f_j and the resources relating to this group in the preceding relations.

Any exchange system, with the corresponding production operations it implies, is deemed ‘advantageous’ when a distributable surplus is achieved and distributed, so that the preference index of any consumption unit concerned increases. If an exchange and production system is advantageous, there must be at least one system of prices which allows it, the prices used by each pair of agents being specific to them. The distribution of the realized surplus between agents is determined by the system of prices used in the exchanges between them.

Conditions of equilibrium and maximum efficiency

In essence all economic operations of whatever type may be considered as reducing to the search for, the achievement of, and the distribution of surpluses. Thus stable general economic equilibrium exists if, and only if, in the situation under consideration, there is no realizable surplus, which means

$$\delta\sigma_u \leq 0 \tag{7}$$

for all feasible modifications of the economy (Allais, 1943, pp. 606–12).

In such a situation the distributable surplus is zero or negative for all possible modifications of the economy compatible with its structural relations, and it is impossible to find any set of prices that would permit effective bilateral or multilateral exchanges (accompanied by the implied production operations) which are advantageous to all the agents concerned.

A situation of maximum efficiency can be defined as a situation in which it is impossible to improve the situation of some people without undermining that of others, i.e. to increase certain preference indices without decreasing others. The set of states of maximum efficiency represents the boundary between the possible and the impossible (Figure 1).

Figure 1

Process of dynamic evolution. Illustrative diagram

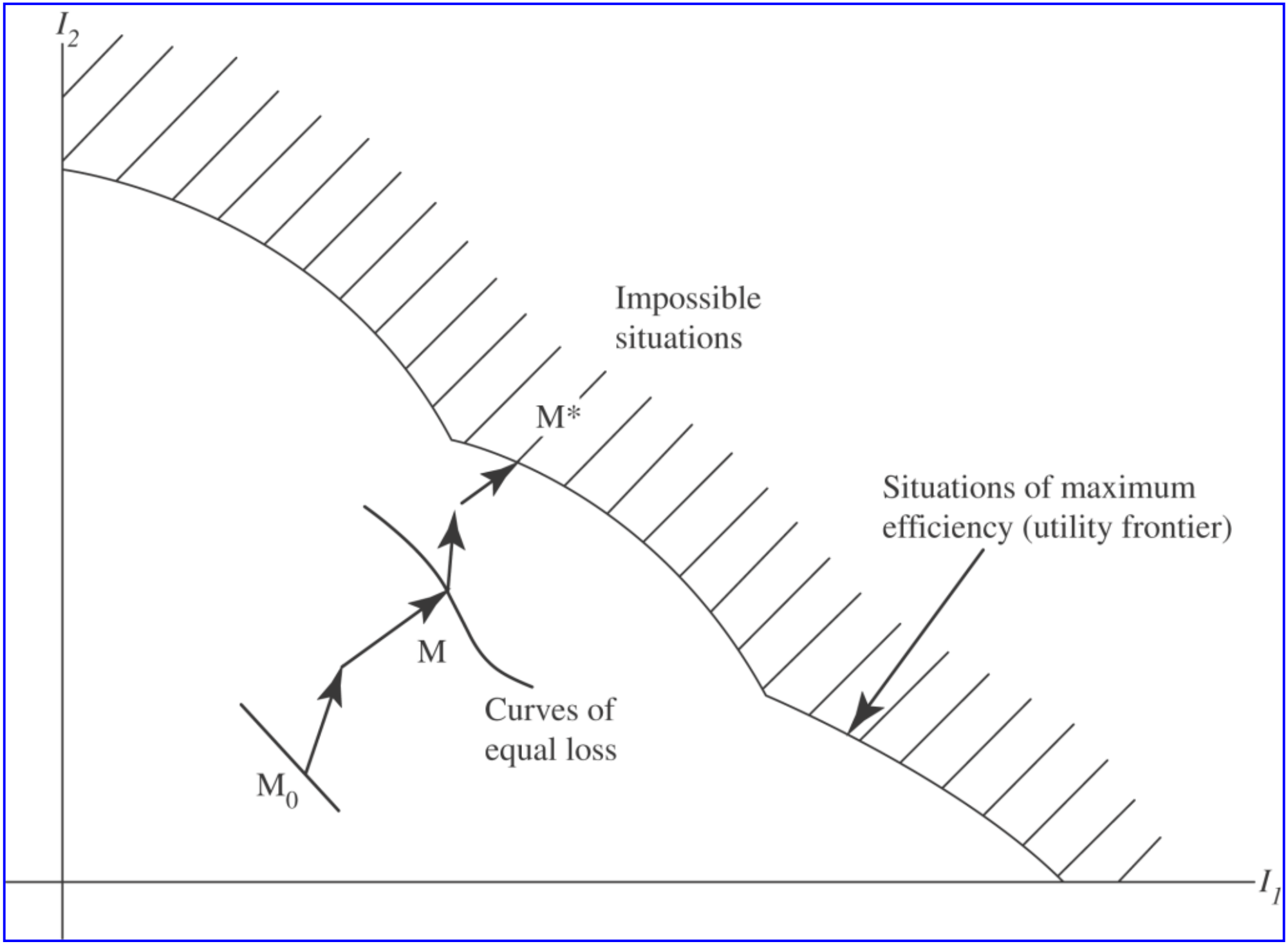

From those definitions of the situations of maximum efficiency and stable general economic equilibrium, it follows, with the greatest generality and without any restrictive hypothesis of continuity, differentiability or convexity, except for the common good (*U*), that:

Any state of stable general economic equilibrium is one of maximum efficiency (*First theorem of equivalence*). Any state of maximum efficiency is one of stable general economic equilibrium (*Second theorem of equivalence*).

Since there can be no stable general economic equilibrium if there is any distributable surplus, every state of stable general economic equilibrium is a state of maximum efficiency. Conversely, if there is maximum efficiency, there is no realizable surplus which could be used to increase at least one preference index without decreasing the others, and consequently, every state of maximum efficiency is a state of stable general economic equilibrium.
Because of the theorems of equivalence, the terms ‘conditions of stable general economic equilibrium’ and ‘conditions of maximum efficiency’ are used interchangeably below.

The dynamic process of the economy: decentralized search for surpluses

In their essence all economic operations, whatever they may be, can be thought of as boiling down to the pursuit, realization and allocation of distributable surpluses. The corresponding model is the Allais model of the economy of markets (1967), defined by the fundamental rule that every agent tries to find one or several other agents ready to accept at specific prices a bilateral or multilateral exchange (accompanied by corresponding production decisions) which will release a positive surplus that can be shared out, and which is realized and distributed once discovered. Thus the evolution of the market's economy is characterized by the condition

$$\delta I_i \geq 0$$

for every consumption unit.
Since in the evolution of an economy of markets surpluses are constantly being realized and allocated, the preference indexes of the consumption units are never decreasing, at the same time as some are increasing. This means that for a given structure, that is to say, for given preferences, resources, and technical know-how, the working of an economy of markets tends to bring it nearer and nearer to a state of stable general economic equilibrium, hence a state of maximum efficiency (Figure 1), which is the third fundamental theorem.
Naturally such evolution takes place only if sufficient information exists about the actual possibilities of realizing surpluses.
To any given initial situation whatsoever, assumed not to be a situation of equilibrium, there corresponds an infinite number of possible equilibrium situations, each corresponding to a particular path and each satisfying the general condition that no index of preference should take on a lower value than in the initial situation (Figure 1).

Economic Loss

The loss σ_u^* which is associated with a given situation is defined as the greatest quantity of the good (U) which can be released in a transformation of the economy for which all the preference indexes remain unchanged (Figure 1) (Allais, 1943, pp. 638–49).
It is a well determined function

$$\sigma_u^* = F[I_1, I_2, \ldots, I_n, U_0, V_0, \ldots, W_0] \quad (8)$$

of the preference indexes I_i and of the resources V_0 which characterize this situation. The loss σ_u^* is an indicator of inefficiency, and $-\sigma_u^*$ an indicator of the efficiency of the economy as a whole.
The loss is minimum and nil in every state of maximum efficiency, and positive in every feasible situation which is not a state of maximum efficiency. It decreases in any modification of the economy, whereby some preference indexes increase, others remaining unchanged, or whereby some surpluses are released with no decline in some preference indexes.

Paths to states of economic equilibrium and maximum efficiency

Since the preference indices I_i are continuous functions of the quantities U_i of the common good (U), the boundary between the possible and the impossible situations in the hyperspace of preference indexes is constituted by a continuous surface. On this surface the loss σ_u^* is nil. This representation allows an immediate demonstration by simple topological considerations of propositions whose proof would otherwise be very difficult. (The paternity of this representation has been unduly attributed to P. Samuelson, 1950, but it was in fact published for the first time in Allais, 1943, and systematically used by Allais in later years especially 1945 and 1947; see Allais, 1971, n.11, p. 385; and 1974a, n.18, pp. 176–7.)
For every feasible situation which is not a state of maximum efficiency, represented by a point such as M_0, there are an infinity of realizable displacements M_0M enabling a situation of maximum efficiency M^* to be approached, such that all the preference indexes have greater values than in the initial situation M_0.

Figure 1 presents an illustration of the process of dynamic evolution by releasing and sharing out of surpluses during which the loss σ_u^* is constantly decreasing (Allais, 1943; 1974b; and 1981, p. 121).

The changing structure of the economy

As psychological patterns vary, as techniques are improved, or as new resources are discovered (or existing resources depleted), the set of situations of maximum efficiency relative to the indexes of preference constantly undergoes change over time. Consequently, situations of equilibrium and maximum efficiency are never reached, and what is really important is to determine the rules of the game which must be applied to come constantly closer to them as rapidly as possible. At a given time t, if information is sufficient and if the adjustments are sufficiently rapid, the point representing the economy will never be very far from the maximum efficiency surface of that time t.

General comment

An economy of markets can be defined as one in which the agents – consumption, production, and arbitrage units – coexist and are free to undertake any exchange transaction or production operation which can result in rendering some distributable surplus available. The principle of the market economy is that any surplus realized is shared among the operators involved. How the surpluses achieved are shared out depends on the specific systems of prices used in the exchanges between the agents concerned. The prices used are always specific to the exchange and production operations considered and there is never a unique system of prices used in common by all the agents.
Diagrammatic representation like that of Figure 1 reveals clearly three basic facts:

1. 1. There is an infinity of situations of maximum efficiency corresponding to a given initial situation characterized by some distribution of property.
2. 2. To each situation of maximum efficiency there corresponds a final distribution of property.
3. 3. This final distribution depends on the initial situation and the distribution of surpluses in the course of the transition.

Thus there is a very strong interdependence between the point of view of efficiency corresponding to the discovery and realization of surpluses and the ethical point of view corresponding to their sharing.
In any event, since only what is produced can be shared, the incentive stemming from the partial or total appropriation of the surpluses by the various agents appears as a fundamental factor for the functioning of the economy of markets.
On the general theory of surpluses and the economy of markets in the general case, and on the fundamental theorems see Allais (1943, pp. 112–77; 181–211; 604–56), (1967, § 8–65), (1968a, vol. 2), (1968b), (1971), (1974a), (1981, pp. 27–48), (1985).

The Equimarginal Principle

Continuity and differentiability

The preceding definitions and theorems are very general and do not make any hypothesis of continuity, derivability or convexity, except the hypothesis of continuity for the common

good (U).

We now assume in addition only that all the quantities and functions considered are continuous and that all functions have first and second order derivatives, the following developments being totally independent of any hypothesis of general convexity.

From the sign conventions adopted earlier it follows that for any i, j and V

$$f'_{iv} = \partial f_i / \partial V_i \geq 0, \; f'_{jv} = \partial f_j / \partial V_j \geq 0.$$

The second partial derivatives are written

$$f''_{ivw} = \partial^2 f_i / \partial V_i \delta W_i, \; f''_{jvw} = \partial^2 f_j / \partial V_j \delta W_j.$$

In the following, the symbol $\overline{d^2 g}$ represents the second differential

$$\overline{d^2 g} = \sum_U^W g''_{v^2} \mathrm{d}V^2 + 2 \sum_{U,V} g''_{vw} \mathrm{d}V \mathrm{d}W$$

of a function $g(U, V, \ldots, W)$ when all parameters in that function are taken as independent, while the symbol $\overline{d^2 g_u}$ represents what this second differential becomes after du has been replaced by its expression derived from

$$\mathrm{d}g = \sum_U^W g'_v \mathrm{d}V = 0$$

(Allais, 1968a, vol. 2, pp. 77–8; 1973b, pp. 151–5; 1981, pp. 688–9).

Convexity and concavity

The local properties of diminishing or increasing marginal returns are related to local conditions of convexity or concavity. Convexity is defined as follows:

Ordinal fields of preference•A field of choice is said to be convex in the whole space (postulate of general convexity) if, at all points of the field, the condition

$$I(M_0) \leq I(M_1)$$

entails

$$I(M_0) \leq I(M)$$

with

$$M = \lambda M_0 + (1 - \lambda) M_1 0 < \lambda < 1 .$$

There is local convexity at M_0 if this condition is satisfied only for

$$|M_0 M_1| < \varepsilon$$

where ε is a given positive number.
When differentiability is assumed local convexity implies

$$\overline{d^2 f_{iu}} \leq 0 \text{for } d f_1 = 0.$$

Fields of production• A field of production is said to be convex over the whole space (postulate of general convexity) if, for any two possible points M_0 and M_1, the centre of gravity defined by the relation

$$M = \lambda M_0 + (1 - \lambda) M_1$$

is likewise a possible point for

$$0 < \lambda < 1 .$$

Local convexity obtains at M_0 if the preceding condition is satisfied only for

$$|M_0 M_1| < \varepsilon$$

where ε is a given positive number.
When differentiability is assumed, local convexity implies

$$\overline{d^2 f_{iu}} \leq 0 \text{for } d f_1 = 0.$$

In fact there is no production operation that does not begin by providing increasing marginal returns, and it is only beyond a certain threshold that diminishing marginal returns are observed. That is a general physical law of nature (Allais, 1943, pp. 193–5; 1968a, vol. 2, pp. 68–96; 1971, pp. 362–4; 1974a, pp. 153–7). Similarly it can be considered as an introspective datum that psychological returns begin by increasing but in the end always decrease beyond certain threshold values. That is a general psychological law (Allais, 1968a, vol. 2, pp. 109–38; 1971, pp. 360–2; 1974a, pp. 153–5). These are two fundamental properties of fields of choice and production. They rule out the postulate of general convexity which is generally accepted in the contemporary literature.

Generation of Distributable Surplus

Consider any economic state (ξ) and a realizable modification $(\delta\xi)$ such that all the preference indexes I_i remain constant (isohedonous modification). Let the conditions of constancy of these indexes and the conditions corresponding to the production functions be written in the same general form

$$g_k(U_k, V_k, \ldots, W_k) = 0 \tag{9}$$

where U_k, V_k,…, W_k represent the consumption of both consumption and production units. By convention, any quantity V_k, if positive, represents consumptions, either by a consumption or a production unit. For any production or consumption unit, any parameter V_k, if negative, represents production of a good or a service.
Let dU_k, dV_k,…, dW_k, be the first order differentials of the variations δU_k, δV_k,…, δW_k of consumptions U_k, V_k,…, W_k in the displacement $(\delta\xi)$. From (9), we have

$$g'_{ku} dU_k + g'_{kv} dV_k + \cdots + g'_{kw} dW_k = 0. \tag{10}$$

Let δV_{kl} be the quantity of (V) received by the consumption or the production unit k from the consumption or production unit l. By definition, we have

$$\delta V_k = \sum_{k \neq 1} \delta V_{kl} \tag{11}$$

$$\delta V_{lk} = -\delta V_{kl}.$$
(12)

Assuming that the displacement $(\delta\xi)$ is such that

$$\sum_k \delta V_k = 0, \ldots, \sum_k \delta W_k = 0.$$
(13)

Let

$$\varepsilon^k_{v,u} = g'_{kv} / g'_{ku}.$$
(14)

The ratio $\varepsilon^k_{v,u}$ is the coefficient of marginal equivalence (or marginal rate of substitution) of goods (V) and (U) for agent k (Allais, 1943, pp. 609–10, and 617–21).
From (10) and (14) we have the relation

$$\mathrm{d}U_k = -\left[E^k_{vu}\mathrm{d}V_k + \cdots + E^k_{wu}\mathrm{d}W_k\right]$$
(15)

between the first order differential $\mathrm{d}U_k$, $\mathrm{d}V_k$,…, $\mathrm{d}W_k$.
If $\mathrm{d}U_k$ is positive, agent k receives a quantity $\mathrm{d}U_k$ to within the second order. If $\mathrm{d}U_k$ is negative, agent k supplies a quantity -$\mathrm{d}U_k$ to within the second order.
From the condition (13), it follows that the displacement considered releases a global distributable surplus

$$\delta\sigma_u = -\sum_k \delta U_k$$

representing the excess of the quantities supplied over the quantities received of good (U) whose first order differential is

$$\delta\sigma_u = -\sum_k dU_k.$$

From (11) and (15)

$$dU_k = -\sum_v^w \left[\varepsilon_{vu}^k \sum_{\substack{k,l \\ k<l}} dV_{kl} \right]$$

and from (12), we have (Allais, 1952c, p. 31; 1968a, vol. 2, p. 174; 1981, p. 88)

$$d\sigma_u = \sum_v^w \sum_{\substack{k,l \\ k<l}} (\varepsilon_{vu}^k - \varepsilon_{kl}^l) dV_{kl}. \tag{16}$$

According to definitions (5) and (6) $d\sigma_u$ is the first differential of the global distributable surplus $\delta\sigma_u$ released in the displacement considered. For all economic agents the unit of value is defined by condition $u_k=u=1$. The marginal values $v_k,\ldots, w_k$ of goods (V),…, (W) for unit k are defined with respect to the u_k by the relations

$$\frac{g'_{ku}}{u_k} = \frac{g'_{kv}}{v_k} = \cdots = \frac{g'_{kw}}{w_k} \tag{17}$$

$$u_k = u = 1. \tag{18}$$

Under the adopted sign convention, all the v_k are positive. We have from (14) and (18)

$$\varepsilon_{vu}^{k} = v_k \tag{19}$$

and relation (16) is written

$$\mathrm{d}\sigma_u = \sum_{v}^{w} \sum_{\substack{k,l \\ k<l}} (v_k - v_l)\mathrm{d}V_{kl} \tag{20}$$

where v_k and v_l are the marginal values of good (V) for units k and l. This summation covers all agents, both consumption and production units. It can thus be seen that all the differences between the marginal values in the situation $\mathcal{E}$ can give rise to the release of potential surpluses which can be released and distributed.

The meaning of relation (20) is immediate. Thus if $v_k > v_l$ the relative value of good (V) is higher for agent k than for agent l. The transfer of a positive quantity $\mathrm{d}V_{kl}$ of good (V) from agent l to agent k therefore creates an additional positive value

$$\mathrm{d}\sigma_{ukl} = (v_k - v_l)\mathrm{d}V_{kl}.$$

If in this ‘isohedone’ transformation surpluses are released, all positive, they can be distributed in such a way as to increase all preference indexes. In such a modification of the economy, the maximum distributable surplus diminishes, and the point representing the economic situation considered moves closer to the surface of maximum efficiency in the hyperspace of preference indexes. Naturally, for this condition to obtain, the corresponding exchanges and the changes of the consumptions and productions they imply in the production system, must effectively occur.

Psychological Values and Marginal Psychological Values

Naturally, the v_k are only marginal values for the agents. The psychological values v_i^* of the consumption V_i of a subject i is defined by the relation

$$f_i(U_i + v_i^* V_i, 0, \ldots, W_i) = f(U_i, V_i, \ldots, W_i)$$

where $v_i^* V_i$ is the sum he would accept to receive to offset the drop in his consumption V_i to zero. The unit value v_i^* is generally much higher than the marginal value v_i corresponding to relations (17), (18) and (19).

In any event, a consumption is only advantageous when its psychological value is higher than its marginal value, because, if this were not so, it would be in the subject's interest to

reduce his consumption V_i.

Conditions of Stable General Economic Equilibrium and Maximum Efficiency of the Economy

From condition (7) it follows that the necessary and sufficient condition for a situation (ξ) to be of stable equilibrium and maximum efficiency is that the distributable surplus $\delta\sigma_u$ defined by (5) and (6) be negative or zero for every feasible modification $(\delta\xi)$, that is every modification that is compatible with the constraint conditions, that is, the structural relations of the economy (2) and (4) above.
Condition (7) implies the two conditions (Allais, 1943, p. 612)

$$d\sigma_u = 0 \text{ (first order condition)} \tag{21}$$

$$d^2\sigma_u \leq 0 \text{ (second order condition)} \tag{22}$$

for any realizable and reversible modification $(\delta\xi)$ in which the expressions of $d\sigma_u$ and $d^2\sigma_u$ represent the first and second differential of $\delta\sigma_u$.
Thus we have according to (21) and (22) using the above notations

$$d\sigma_u = \sum_i d\sigma_{ui} = \sum_i dl_i / l'_{iu} = 0 \tag{23}$$

$$d^2\sigma_u = \sum_i \frac{\overline{d^2 f_{iu}}}{f'_{iu}} + \sum_j \frac{\overline{d^2 f_{ju}}}{f'_{ju}} \leq 0 \text{for } d\sigma_u = 0. \tag{24}$$

Actually, and according to relation (20), the first order condition (23) implies that when the quantities V_k are not nil, all the marginal values v_k are equal to a same value v and a same system of prices $u, v, \ldots, w$ *then exists for all the agents k concerned, such that*

$$\frac{g'_{ku}}{u} = \frac{g'_{kv}}{v} = \cdots = \frac{g'_{kw}}{w}. \tag{25}$$

These equalities condense the general equimarginal principle into a single formulation. They express the fact that in a situation of equilibrium and maximum efficiency, the psychological (or objective) value v_k of the last dollar is the same, for any agent (consumption or production unit), whatever use it is put to.

For the quantities V_k which are nil (terminal equilibria), we necessarily have

$$v_k \leq v$$

since, if this were not true, the operator's interest would be to increase V_k from the value $V_k = 0$; he could indeed do this because of the existence of other operators who are in a situation of tangential equilibrium for good (V).

The second order condition (24) holds whether or not the df_i are equal to zero. It is only subject to the constraint (21). If we consider only the modifications of the economy involving units k and l, condition (24) is written

$$d^2\sigma_u = \frac{\overline{d^2 f_{ku}}}{f'_{ku}} + \frac{\overline{d^2 f_{lu}}}{f'_{lu}} \leq 0 \text{for } d\sigma_{uk} + d\sigma_{ul} = 0$$

shows that when in a situation of maximum efficiency consumption or production units consume (or produce) the same goods, one unit at most is in a situation of local concavity, that is, in a situation of marginal increasing returns (Allais, 1968a, pp. 196–9; 1974a, n.125, p. 184; 1981, p. 65).

Consequently, when maximum efficiency obtains, most operators are in a situation of local convexity and marginal decreasing returns. However, this condition cannot be interpreted as meaning that all fields of choice and production are convex everywhere, this hypothesis being totally contradicted by observed data.

When local convexity obtains for a consumption unit, its index of preference is effectively at a maximum, subject to the budgetary constraint, equilibrium prices being taken as given. Similarly, if local convexity obtains for a production unit, the unit's income is effectively at a maximum, equilibrium prices again being taken as given. However, these two principles, which in any case could be valid only for a situation of maximum efficiency, cannot be considered as corresponding in all cases to optimum behaviour, and they cannot be taken to be of general value. As a matter of fact and for instance, if, in a situation of maximum efficiency, a production unit is in a situation of local concavity, its income is minimum, the equilibrium prices being considered as given.

Conditions (25) and (24) show the total symmetry of the implications of the psychological and technical structures of the economy.

of Maximum Efficiency

The integration of eq. (20) along a path leading to a state of maximum efficiency leads to the following approximate estimate to within third order accuracy of the global loss involved in the initial situation (relation 8)

$$\sigma_u^* \sim \frac{1}{2u} \sum_v \sum_{\substack{k,l \\ k<l}} (v_k - v_l)\, \delta V_{kl}^*. \tag{26}$$

In this relation, the quantities $v_k - v_l$ represent the differences of marginal values in the initial state considered, and the δV_{kl}^* are the quantities of the good (V) received by operator k from operator l in the transition from the initial to the final state. Relation (26) is of the broadest generality, and holds whatever the initial state (Allais, 1952a, pp. 31–2, n. 8; 1968a, vol. 2, p. 207; 1981, p. 110).

Its simplicity is really extraordinary in view of the complexity of the concept it represents, namely the maximum of the distributable surplus for all the modifications which the economy can undergo while leaving the preference indexes unchanged.

In the neighbourhood of a situation of maximum efficiency, the $(v_k - v_l)$ and δV_{kl}^* are of the first order quantities, where as the loss σ_u^* is only of the second order. However, since the δV_{kl}^* are of the first order, the variations δI_i of the preference indexes are also of the first order. As a result, and for instance, in the neighbourhood of a situation of maximum efficiency, taxes have major first order effects on the distribution of income but only second order effects on the efficiency of the economy.

On the theoretical foundations of the equimarginal principle, see Allais (1943, pp. 604–56), (1945), (1952a, pp. 28–32), (1967), (1968a, vol. 2), (1971), (1973a), (1973b), (1974a), (1974b), (1981) and (1986). Illustrative models: Allais (1943, Annexe I, pp. 4–24), (1945, pp. 57–69). On its extension see: cases of perfect and imperfect foresight: Allais (1943, pp. 343–84), (1947, pp. 23–228), (1964), (1967), (1968a, vol. 2). Illustrative models: (1947, pp. 631–771). Capitalistic optimum theory: Allais (1947, pp. 179–228), (1962), (1963). Demographic optimum theory: Allais (1943, pp. 749–85). Case of risk: Allais (1952b). Application of marginal analysis to transport: Allais (1964) and (1987). For a general overview on the meaning, limits, generalizations, and history of the equimarginal analysis see Allais (1987).

General Overview

Theory of Surpluses and Marginal Analysis

As a matter of fact a single relation, the relation (20) (or the equivalent relation (16)) condenses the whole marginal approach as it has developed for over a century. Subject only to the hypotheses of continuity and derivability implied by any marginal theory, it applies in all cases, and its simplicity is really extraordinary.

It also shows that equilibrium and maximum efficiency can obtain only when all marginal values are equal, which is the equimarginal principle.

The equimarginal principle was discovered first by Gossen (1854), and rediscovered, broadened and introduced independently into economics by Jevons (1871), Menger (1871) and Walras (1874–7). In the following years numerous new developments of the principle have been presented by their immediate successors, especially by Edgeworth (1881), Irving Fisher (1892) and Vilfredo Pareto (1896–1911). Particularly striking illustrations of the role of differences in marginal equivalences are Ricardo's theory of comparative costs (1817) and Dupuit's theory of economic losses (1844–53).

This principle corresponds to the outcome of the dynamic process of the economy induced by differences in marginal equivalences. According to Irving Fisher (1892), with whose judgement I agree fully, ‘No idea has been more fruitful in the history of economic science.’ Its applications and generalizations dominate all economic analysis in real terms.

From the foregoing a double conclusion emerges: the classical theory of marginal equivalences is irreplaceable to make understandable the underlying nature of all economic phenomena; the general theory of surplus, of which classical marginal theory is only a special case, allows one to extend the propositions of marginal analysis to the most general case of discrete variations and indivisibilities.

As important as the analysis of the conditions of general equilibrium and maximum efficiency may be, the analysis of the dynamic processes which enable surpluses to be generated from a given situation is much more important. From this point of view the analyses by Dupuit, Jevons, Edgeworth, Pareto, and the marginal school and its predecessors in general, appear much more realistic than the contributions which rest only upon the consideration of Walras's general model of equilibrium.

In fact, what is really important is not so much the knowledge of the properties of a state of maximum efficiency as the rules of the game which have been applied to the economy effectively to move nearer to a state of maximum efficiency.

The decentralized search for surpluses is truly the dynamic principle from which a thorough and yet very simple conception of the operation of the whole economy can be derived. Whereas in the market economy model the search for efficiency is essentially focused on the determination of a certain set of prices, the analysis of the model of the economy of

markets is based on the search for potential surpluses and their realization. Not only is the economy of markets model much more realistic than the market economy model while lending itself to much simpler proofs, but also these proofs are not subordinated to any restrictive assumptions relating to continuity, differentiability of functions, or convexity. All of economic dynamics is reduced to a single principle: the search for and realization of potential surpluses, which leads to the minimization of loss for the economy as a whole.
On all these points see especially: Allais (1971) and (1974a).

The Tendencies of the Contemporary Literature

From Walras on, the literature became progressively – and unduly – concentrated on equilibrium analysis which, however interesting it could be, is less so than the analysis of the processes by which the economy tends at any time towards situations of equilibrium which in fact are never reached.
Today there is a tendency to neglect the dynamic marginal approach based on the consideration of differences in marginal equivalences; and in the name of a so-called rigour it has been replaced by new theories. A fortiori, the general theory of surpluses which generalizes marginal analysis is simply ignored. This development, which in reality, and despite the too-widely held belief to the contrary, represents an immense step backward, basically stems from the unquestioning acceptance of ‘established truths’ taught by the dominant ‘establishments’, whose only real basis is their incessant repetition.
As a matter of fact the guiding principles of the contemporary theories descending from Walras: the adoption of the market economy model; the hypothesis that a common price system applicable to all operators prevails at each instant; the assumption of general convexity; and the exaltation of mathematical formalism of the theory of sets to the detriment of conformity with actual facts, constitute an impediment to any genuine progress in analysis of the economy in real terms.
The essential difference between the market economy model and the model of the economy of markets is that, in the latter, the exchanges leading to equilibrium take place successively at different prices, and that, at any given moment, the price sets used by different operators are not necessarily the same. Whereas in the first model the final situation is determined totally by the initial situation, which correspondingly plays a privileged role without any real justification, in the second the final situation depends both on the initial situation and the path taken from it to the final situation (Figure 1).
Whereas the market economy model postulates perfect competition and a large number, if not an infinity, of operators, the model of the economy of markets applies just as well to the cases of monopoly as to the cases of competition.
Not only is the market economy model unrealistic, but it also gives rise to considerable mathematical difficulties when an attempt is made to demonstrate the above three fundamental theorems. Whether differential calculus or set theory is used, the theorems can only be demonstrated under extremely restrictive conditions, and the difficulties they imply are, from an economic standpoint, completely artificial, for they arise solely from the unrealistic nature of the model used. Paradoxically, whereas these restrictive assumptions are totally unrealistic, most of the theoretical difficulties encountered disappear, as shown above, once they are discarded.
The market economy approach leads to imposing on any economic model, for it to be considered satisfactory, conditions which actually apply to a particular model, which are generally not fulfilled in reality, and for which, at all events, no rigorous justification can be found.
By departing from the great tradition of marginal theory and by adopting an unrealistic model and unrealistic assumptions, the contemporary theories, purely mathematical, have doomed themselves to sterility as regards the understanding of reality.
On the contemporary theories see especially: Samuelson (1947); Arrow (1968); Debreu (1959) and (1985); Blaug (1979 and 1985); Arrow and Hahn (1971); Hutchison (1977, pp. 62–97 and 161–70); Woo (1985); and Allais (1952b; 1968b; 1968e; 1971; 1974a; and 1981).

See Also

- Allais, Maurice
- efficient allocation
- general equilibrium
- optimality and efficiency
- surplus

Bibliography

Allais, M. 1943. *A la recherche d'une discipline économique. Première partie. I’économie pure*. 2 vols. Paris: Ateliers Industria. 2nd edn, *Traité d’économie pure*, 5 vols, Paris: Imprimerie Nationale, Paris, 1952. (The second edition is identical to the first, except for a new Introduction.)

Allais, M. 1945. *Economie pure et rendement social*. Paris: Sirey.

Allais, M. 1947. *Economie et intérêt*. 2 vols. Paris: Imprimerie Nationale and Librairie des Publications Officielles.

Allais, M. 1952a. *Introduction to the 2nd edn of Allais (1943)*. Paris: Imprimerie Nationale.

Allais, M. 1952b. L'éxtension des théories de l'équilibre économique général et du rendement social au cas du risque. *Colloques Internationaux du Centre National de la Recherche Scientifique* 40, *Econométrie*, 81–120. A summarized version was published under the same title in *Econometrica*, (1953), 269–90.

Allais, M. 1962. The influence of the capital output ratio on real national income. *Econometrica* 30, 700–28. Republished in American Economic Association, *Readings in Welfare Economics*, vol. 12, with an additional Note, 1969.

Allais, M. 1963. The role of capital in economic development. In *Study Work on the Econometric Approach to Development Planning*, Pontificiae Academiae Scientiarum Scripta Varia 28, Pontifica Academia Scientiarum, Amsterdam: North-Holland, 1963 and Chicago: Rand McNally, 1965.

Allais, M. 1964. La theorie economique et la tarification optimum de l'usage des infrastructures de transport. *La Jaune et la Rouge* (publication of the Société Amicale des Anciens Eléves de l'Ecole Polytechnique), special issue *Les Transports*, Paris, 1964.

Allais, M. 1967. Les conditions de l'éfficacite dans l'économie. Fourth International Seminar, Centro Studi e Ricerche su Problemi Economico-Sociali, Milan. Italian translation: 'Le condizioni dell' efficienza nell' economia' in *Programmazione E Progresso Economico*, Milan: Franco Angeli, 1969. Original French text in M. Allais, *Les Fondements du Calcul Economique*, vol. 1, Paris: Ecole Nationale Supérieure des Mines de Paris, 1967.

Allais, M. 1968a. *Les fondements du calcul économique*. 3 vols. Paris: Ecole Nationale Supérieure des Mines, Paris, vol. 1, 1967, and vols 2 and 3, 1968.

Allais, M. 1968b. The conditions of efficiency in the economy. *Economia Internazionale* 21(3), 399–420.

Allais, M. 1968c. Pareto, Vilfredo: contributions to economics. In *International Encyclopedia of the Social Sciences*, vol. 2 New York: Macmillan and Free Press.

Allais, M. 1968d. Fisher, Irving. In *International Encyclopedia of the Social Sciences*, vol. 5 New York: Macmillan and Free Press.

Allais, M. 1968e. L'économique en tant que science. *Revue d'Economie Politique*, January-February, 5–30. Trans. as 'Economics as a Science', *Cahiers Vilfredo Pareto*, (1968), 5–24.

Allais, M. 1971. Les théories de l'équilibre economique général et de l'éfficacité maximale – impasse récentes et nouvelles perspectives. Congrés des Economistes de Langue Francçaise, 2–6 June. *Revue d'Economie Politique* 3, 331–409. Spanish translation: 'Las theorias del equilibrio economico general y de la eficacia maxima – recientes callejones sin salida y nuevas perspectivas'. *El Trimestre Economico* (Mexico), 39(1972), 557–633; English translation: see Allais (1974a).

Allais, M. 1973a. La théorie gélnérale des surplus et l'apport fondamental de Vilfredo Pareto. *Revue d'Economie Politique* 6, 1044–97.

Allais, M. 1973b. The general theory of surplus and Pareto's fundamental contribution. *Convegno Internazionale Vilfredo Pareto*. Roma, 25–27 October, Rome: Accademia Nazionale dei Lincei, 1975 (English trans. of Allais, 1973a.)

Allais, M. 1974a. Theories of general economic equilibrium and maximum efficiency. Vienna Institute for Advanced Studies. In *Equilibrium and Disequilibrium in Economic Theory*, ed. G. Schwödiauer. Dordrecht: Reidel, 1977. (English version of Allais, 1971, with some additions.)

Allais, M. 1974b. Les implications de rendements croissants et décroissants sur les conditions de I'équilibre economique général et d'uned éfficacité maximale. In *Hommage à François Perroux*. Grenoble: Presses Universitaires de Grenoble, 1978.

Allais, M. 1981. *La théorie générale des surplus. Economies et Sociétés*, 2 vols. Montrouge institut de sciences mathématique et économies appliquées.

Allais, M. 1985. The concepts of surplus and loss and the reformulation of the theories of stable general economic equilibrium and maximum efficiency. In *Foundations and Dynamics of Economic Knowledge*, ed. M. Baranzini and R. Scazzieri. Oxford: Basil Blackwell.

Allais, M. 1987. *The Equimarginal Principle, Meaning, Limits, and Generalisations*. Centre d'Analyse Economique. *Revista internazionale di scienze economiste e commerciale*.

Arrow, K.J. 1968. Economic equilibrium. In *International Encyclopedia of the Social Sciences*, vol. 4, New York: Macmillan and Free Press.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day; Edinburgh: Oliver.

Blaug, M. 1979. *Economic Theory in Retrospect*. 4th edn, London: Heinemann Educational Books, 1985.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Debreu, G. 1985. *Theoretic Models: Mathematical Form and Economic Content*. Frisch Memorial Lecture, Fifth World Congress of the Econometric Society, MIT, 17–24 August.

Dupuit, J. 1844. De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées*, 2nd series, Mémoires et Documents No. 116, vol. 8.

Dupuit, J. 1849. De l'influence des péages sur l'utilité des voies de communication. *Annales des Ponts et Chaussées*, 2nd series.

Dupuit, J. 1853. De l'utilite et de sa mesure. *Journal des Economistes* 36(147), 1–28.

Edgeworth, F.Y. 1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: Kegan Paul. Reprinted, New York: Kelley, 1953.

Fisher, I. 1892. *Mathematical Investigations in the Theory of Value and Prices*. New Haven: Yale University Press, 1925.

Gossen, H.H. von. 1854. *Entwickelung der Gesetze des menschlichen Verkehrs und der daraus fliessender Regeln für menschliches Handeln*. 3rd edn, introduction by Friedrich Hayek, Berlin: Präger, 1927.

Hutchison, T.W. 1977. *Knowledge and Ignorance in Economics*. Oxford: Blackwell.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan. 5th edn trans. as *La théorie de économie politique*. Paris: Giard, 1909.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Vienna: Braumneller.

Pareto, V. 1896–7. *Cours d'économie politique*. 2 vols. Lausanne: Rougé. Reprinted, Geneva: Droz, 1964.

Pareto, V. 1901. Anwendungen der Mathematik auf Nationalökonomie. *Encyklopädie der Mathematichen Wissenschaften*, vol. 1, Leipzig.

Pareto, V. 1906. *Manuale d'economia politica*. Milan. Trans. as *Manuel d'économie politique*. Paris: Giard et Brière, 1909, and Geneva: Droz, 1966, and as *Manual of Political Economy*. Reprinted, New York: Kelley, 1971.

Pareto, V. 1911. Economie mathématique. *Encyclopedie des Sciences Mathématiques*, Paris: Gauthier–Villars, 1911, also in *Statistique et économie mathématique*, Geneva: Droz, 1966, published in English as 'Mathematical economics', *International Economic Papers* No. 5. New York, 1955.

Ricardo, D. 1817. On the Principles of Political Economy and Taxation. Vol. I of *The Works and Correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951–1955.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. 2nd edn, Cambridge, MA: Harvard University Press, 1948.

Samuelson, P.A. 1950. Evaluation of real national income. *Oxford Economic Papers* NS, 2, 1–29.

Walras, L. 1874–7. *Eléments d'économie politique pure – théorie de la richesse sociale*. 6th edn. Paris: Guillaumin; reprinted, Paris: Pichon et Durand-Auzias, 1952. English trans. of the 6th edn. as *Elements of Pure Economics*, ed. W. Jaffé, London: Allen & Unwin, 1954.

Woo, H.K.H. 1985. *What's Wrong with Formalization in Economics? – An Epistemological Critique*. Hong Kong: Hong Kong Institute of Economic Science.

How to cite this article

Allais, Maurice. "economic surplus and the equimarginal principle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000024> doi:10.1057/9780230226203.0441

The New Palgrave Dictionary of Eco

economics in Belgium

Guido Erreygers
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There is no single Belgian school of economics, but a number of Belgian economists have made significant contributions, including Adolphe Quetelet, Ernest Solvay, Léon Dupriez, Paul Van Zeeland, Gaston Eyskens, Étienne Sadi Kirschen, Robert Triffin and Jacques Drèze.

Keywords

Belgium; De Man Plan; De Molinari, Gustave; Drèze, Jacques; European Economic Association; free trade; Quetelet, Adolphe; Triffin, Robert

Article

Although no genuine Belgian school of economics has ever emerged, except perhaps in the second half of the 20th century, Belgian economists have made original and significant contributions to the discipline and played a major role in the creation of a European community of economists.

Before the First World War

When Belgium gained independence in 1830, economics as a scientific discipline virtually did not exist in the country. By the middle of the 19th century, however, most universities were offering courses on economic subjects, economists began forming associations, and international cooperation was actively pursued. Throughout the 19th century French economic thought was undoubtedly the main source of inspiration for economists working in Belgium, but British, German and Dutch economic

schools were also influential. Another characteristic of that period was the strong separation along ideological lines, with little interaction between Catholic economists, mainly associated to the Catholic University of Louvain, on one side, and liberal and socialist economists, associated to the Free University of Brussels and the State Universities of Ghent and Liège, on the other. The ideological divide is clearly visible in the rather biased account of the history of economic thought in 19th-century Belgium published by Michotte (1904).

Although not in the first place known as an economist, the polymath Adolphe Quetelet (1796–1874) needs to be mentioned for his pathbreaking contributions with regard to the use of statistics in the social sciences. He introduced the notion of the ‘average man’, which in economics influenced the work of both the German Historical School and William Stanley Jevons (Mosselmans, 2005). A fine example of pioneering statistical research is provided by the household surveys of Édouard Ducpétiaux (1804–1868), whose data were used by Ernst Engel to derive relationships between consumption and income.

Not surprisingly, many Belgian economists considered themselves to be part of the liberal family. The first generation of liberal economists is probably best represented by Charles De Brouckere (1796–1860), who combined careers in politics, academics and business. Together with Adolphe Le Hardy de Beaulieu (1814–1894), the Italian émigré Giovanni Arrivabene (1787–1881) and others, he founded a Belgian association of free-traders (Erreygers, 2001). The main accomplishment of the association was the organization of the Congrès des Économistes in September 1847 in Brussels, the very first international conference of economists attended predominantly by ardent free-traders, and also by Karl Marx and Friedrich Engels. The next generation of liberal economists was headed by Gustave De Molinari (1819–1912), who advocated an extreme libertarian form of liberalism, opposing virtually any form of government intervention. Some consider him to have laid the foundations of free-market anarchism, also known as anarcho-capitalism (Hart, 1981–2). Although he spent much of his time in France, where he was for a long time editor of the *Journal des Économistes*, he played a very active role in Belgium. He founded and edited *L'Économiste Belge*, animated the Société Belge d'Économie Politique, and managed to breathe new life into the free-trade movement, both nationally and internationally (Van Dijck, 2008). In many of these initiatives he found a fellow-traveller in Charles Le Hardy de Beaulieu (1816–1871), who published several textbooks on economics. It

must be added, however, that few Belgian economists shared De Molinari's extreme view of liberalism.

In the second half of the 19th century Émile De Laveleye (1822–1892) was the country's most prominent economist. This prolific writer covering a wide area of topics had been strongly influenced by the French philosopher and Christian socialist François Huet, who taught at the University of Ghent. De Laveleye's economic publications included work on the origins and varieties of property rights, on bimetallism and on socialist doctrines. He was professor of political economy at the University of Liège, and authored an often reprinted textbook on economics. He considered his views to be close to those of John Stuart Mill and the German Historical School. Although very much appreciated by his contemporaries – he built up an impressive international network of colleagues and correspondents – his contributions lost most of their influence soon after he died.

At that time the wealthy industrialist Ernest Solvay (1838–1922), founder of the chemical firm Solvay & Cie, started to turn his attention to social and economic issues. He was convinced that a change of the monetary system (replacing the system based on metallic money by a pure-credit system which he called ‘social comptabilism’) combined with a sweeping reform of taxation (replacing all existing taxes by taxes on gifts and bequests, with rates increasing with the number of transfers) would provide the clue to solving society's problems (Erreygers, 1998; Boianovsky and Erreygers, 2005). Solvay, a prominent liberal, worked on these issues in close collaboration with leading socialist economists such as Hector Denis (1845–1913) and Émile Vandervelde (1866–1938). Their monetary propositions led to a debate with Léon Walras, who saw a great similarity with his own views. On a more practical level, Solvay influenced economics in Belgium through his generous funding of various institutions associated to the University of Brussels, the most important of which are the Institut de Sociologie and the École de Commerce Solvay (now Solvay Business School). These institutions allowed economists such as Émile Waxweiler (1867–1916), Maurice Ansiaux (1869–1943) and Boris Chlepner (1890–1964) to do research.

Socialist doctrines found responsive audiences in Belgium, partly as the result of the rapid industrialization, but also because of the presence of exiles such as Karl Marx and Joseph Proudhon. The Saint-Simonians and Fourierists attracted scores of young intellectuals. This created a fertile ground for such figures as Hippolyte Colins de Ham (1783–1859), who proposed to provide all adults with a capital endowment, Joseph Charlier (1816–1896), who launched

the idea of an unconditional basic income, and César De Paepe (1842–1890), who tried to bridge the gap between the Marxists and the anarchists (Cunliffe and Erreygers, 2001).

At the University of Louvain economics had a decidedly Catholic profile in the 19th century. Charles Périn (1815–1905) and Victor Brants (1856–1917) both aimed at developing a social economics in accordance with the doctrine of the Church, along the lines of Le Play.

The interwar period

After the end of the First World War fellowships offered by the Commission for Relief in Belgium and the Belgian American Educational Foundation gave many talented economics students the opportunity to spend at least one year in the United States. This was the case with Léon Dupriez (1901–1986), Paul Van Zeeland (1893–1973) and Gaston Eyskens (1905–1988), who combined their studies at the University of Louvain with stays in respectively Harvard, Princeton and Columbia. As a result Belgian economics gradually obtained a more American character and the University of Louvain became much more prominent in economic research (Maes and Buyst, 2005).

Dupriez introduced to Belgium the statistical business cycle techniques used by Harvard University. He became the driving force of the Institut des Sciences Économiques (later renamed Institut de Recherches Économiques et Sociales, IRES), founded in 1928 at the University of Louvain. Van Zeeland, who had joined the National Bank of Belgium as head of its research department, made a swift career in the bank and was soon considered as one of the country's leading economic experts. In the troubled political and economic climate of the 1930s he was twice prime minister of governments of ‘national unity’. With the scientific backing of IRES, Van Zeeland successfully devalued the Belgian currency in 1936. The Van Zeeland governments also included the socialist Hendrik De Man (1885–1951), who in 1933 had proposed an ambitious ‘Labour Plan’ (also known as the ‘De Man Plan’) as a way out of the economic depression. His project, which was intensively discussed and had broad support in Belgium and in other countries, involved state planning of the economy and a technocratic way of governing.

A few isolated attempts were made to introduce a more mathematical approach in economics. The most ambitious project was that of Bernard Chait (1893–1957), an engineer and businessman who was in close contact with Jan

Tinbergen and François Divisia. In his 1938 Ph.D. thesis he constructed a general mathematical theory capable of explaining business cycle movements, but he failed to convince economists of its usefulness (Erreygers and Jolink, 2007).

After the Second World War

In the first half of the 20th century economics gained increasing recognition as a mature scientific discipline. Universities created special schools and separate faculties for economic sciences, and several economics journals were launched. In the northern part of the country Dutch gradually replaced French as the language of instruction. Both the Universities of Brussels and Louvain were eventually split into Dutch-speaking and French-speaking universities. As early as the 1930s Gaston Eyskens had become the leading figure of the Dutch-speaking section of economists at the University of Louvain. In the years after the war he seemed to be more receptive to the ideas of Keynes than the leading figure on the French-speaking side, Léon Dupriez, who favoured a laissez-faire approach and resisted to the introduction of Keynesian macroeconomic policies. The tension between the Dutch-speaking and French-speaking section at the university led to the foundation, in 1955, of a separate Dutch-speaking research institute at the University of Louvain, the Center for Economic Studies. The Center provided scientific backing for various economic reforms adopted under Eyskens's period as prime minister of Belgium (1958–61, 1968–72). Eyskens also took the initiative to create a government planning bureau (Buyst et al., 2005).

At the University of Brussels Étienne Sadi Kirschen (1913–2000) was the main agent of change in the 1950s. After having worked at the Office of European Economic Cooperation in Paris, he organized a team which estimated the first national accounts for Belgium. In 1957 Kirschen and his collaborators founded the Département d'Économie Appliquée, better known under its acronym DULBEA, the economic research institute of the University of Brussels. It constructed the first input–output table for Belgium. In close cooperation with Tinbergen, who lectured at the University of Brussels in the 1960s, Kirschen emphasized policy-oriented research based on quantitative methods, as exemplified in the ambitious three-volume *Economic Policy in our Time* (1964) by an international team of economists led by Kirschen (Sirjacobs, 1997).

A few Belgian economists emigrated and made a career abroad. The most

striking case is that of Robert Triffin (1911–1993), who initially studied at the University of Louvain. Having obtained his Ph.D. at Harvard, he worked for the Fed and the International Monetary Fund (IMF) – under its first director, the former Belgian Finance Minister Camille Gutt (1884–1971) – and in 1951 became professor of economics at Yale University. He made himself a reputation by pointing out a crucial weakness of the Bretton Woods system (later known as the Triffin dilemma) and arguing for a fundamental reform of the international monetary system. He was an influential economic adviser to key players in the European economic and monetary integration process; he returned to Belgium in the 1970s (Maes and Buyst, 2005). A less well-known émigré is Raymond De Roover (1904–1972), who after his studies at the Antwerp Catholic business school decided to specialize in economic history. He earned his Ph.D. at the University of Chicago, and was later appointed as professor in Boston and New York. His work on early banking in Bruges and Florence made him a leading specialist on late medieval economic history and thought.

Probably the most important development in Belgian economics after the Second World War was initiated by a man who decided not to stay in the United States after completing his Ph.D. thesis. In 1966 the Center for Operations Research and Econometrics (CORE) was founded in Louvain. This was very much the achievement of Jacques Drèze (b. 1929), a student of the University of Liège who thanks to a fellowship of the Commission for Relief in Belgium went to the United States and obtained a Ph.D. at Columbia. After his return to Belgium he was appointed at the University of Louvain, where he rapidly replaced Dupriez as the dominant figure of the economics faculty, but he kept close contacts with his American colleagues, especially at Northwestern and Chicago, where he was visiting professor in the 1960s. The creation of CORE marked the adoption of a decidedly American style of doing research. From the outset CORE was meant as an interdisciplinary research institute, bringing together specialists in econometrics, statistics, operations research, game theory and (mathematical) economics. Thanks to grants from the Ford Foundation and other institutions, it created a stimulating environment for research and offered fellowships to both Ph.D. students and established researchers. CORE was unique not only because it brought together Dutch-speaking and French-speaking economists from the University of Louvain, but also because Drèze managed to get the econometricians of the University of Brussels involved in the project. Moreover, from the very beginning CORE opened itself to the world: it hired the Dutch econometrician

Anton Barten (b. 1930), established strong links with other European institutions focusing on quantitative economics, and welcomed American and other foreign scholars as fellows (Maes and Buyst, 2005).
The University of Brussels and CORE played a major role in the creation of the *European Economic Review* and the European Economic Association. The *European Economic Review* was founded in 1969 by the European Scientific Association for Medium and Long Term Economic Forecasting (ASEPELT), of which Kirschen was the driving force. His younger colleagues Jean Waelbroeck (b. 1927) and Herbert Glejser (b. 1938), both of the University of Brussels, were the founding editors. In 1985 Waelbroeck and three other Belgian economists, Jean Jaskold Gabszewicz (b. 1936), Louis Phlips (b. 1933) and Jacques-François Thisse (b. 1946), all affiliated to CORE, took the initiative to launch the European Economic Association. Drèze was elected as the first president.
Drèze's contributions to economics are extensive, covering uncertainty, general equilibrium theory, macroeconomics, econometrics and much more (Dehez and Licandro, 2005). He is by far the most influential Belgian economist of the second half of the 20th century. Besides those already mentioned, other important Belgian economists include Claude d'Aspremont (b. 1946), Paul De Grauwe (b. 1946), Mathias Dewatripont (b. 1959), Pierre Pestieau (b. 1943), Jean-Philippe Platteau (b. 1947) and Gérard Roland (b. 1954). It is remarkable that the top of the economics profession remains very much dominated by French-speaking economists.

See Also

- Catholic economic thought
- Tinbergen, Jan (1903–1994)
- Triffin, Robert (1911–1993)

Bibliography

Boianovsky, M. and Erreygers, G. 2005. Social comptabilism and pure credit systems: Solvay and Wicksell on monetary reform. In *The Experiment in the History of Economics*, ed. P. Fontaine and R. Leonard. London: Routledge, 98–134.

Buyst, E., Maes, I., Plasmeijer, H.W. and Schoorl, E. 2005. Comparing the development of economics during the twentieth century in Belgium and the

Netherlands. *History of Political Economy* 37, 61–78.

Cunliffe, J. and Erreygers, G. 2001. The enigmatic legacy of Fourier: Joseph Charlier and basic income. *History of Political Economy* 33, 459–84.

Dehez, P. and Licandro, O. 2005. From uncertainty to macroeconomics and back: an interview with Jacques Drèze. *Macroeconomic Dynamics* 9, 429–61.

Erreygers, G. 1998. The economic theories and social reform proposals of Ernest Solvay (1838–1922). In *European Economists of the Early 20th Century. Volume One: Studies of Neglected Thinkers of Belgium, France, The Netherlands and Scandinavia*, ed. W. Samuels. Cheltenham: Edward Elgar, 220–62.

Erreygers, G. 2001. Economic associations in Belgium in the 19th century. In *The Spread of Political Economy and the Professionalisation of Economists: Economic Societies in Europe, America and Japan in the Nineteenth Century*, ed. M. Augello and M. Guidi. London: Routledge, 91–108.

Erreygers, G. and Jolink, A. 2007. Perturbation, networks and business cycles: Bernard Chait's pioneering work in econometrics. *European Journal of the History of Economic Thought* 14, 543–71.

Hart, D.M. 1981–2. Gustave de Molinari and the anti-statist liberal tradition. *Journal of Libertarian Studies* 5, 263–90, 399–434; 6, 83–104.

Maes, I. and Buyst, E. 2005. Migration and Americanization: the special case of Belgian economics. *European Journal of the History of Economic Thought* 12, 73–88.

Michotte, P. 1904. *Études sur les theories économiques qui dominèrent en Belgique de 1830 à 1886*. Louvain: Charles Peeters.

Mosselmans, B. 2005. Adolphe Quetelet, the average man and the development of economic methodology. *European Journal of the History of Economic Thought* 12, 565–82.

Sirjacobs, I. 1997. *L'économiste dans le temps*. Bruxelles: Archives de l'ULB.

Van Dijck, M. 2008. From science to popularization, and back: the science

and journalism of the Belgian economist Gustave de Molinari. *Science in Context* 21, 377–402.

How to cite this article

Erreygers, Guido. "economics in Belgium." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_B000343> doi:10.1057/9780230226203.1899

The New Palgrave Dictionary of Eco

economics of franchises

Roger D. Blair and Jessica S. Haynes
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Franchising, which is common in many advanced economies, is a contractual form of vertical integration. This article examines the economic rationale for choosing franchising over vertical integration. It also examines the influence of the franchisor's ability to maximize its own profit.

Keywords

distribution systems; franchisee; franchising; franchisor; royalties; vertical integration

Article

Franchising is a contractual form of vertical integration. A manufacturer, for example, produces a product that must be distributed to consumers. The manufacturer can perform the distribution function itself through a chain of retail outlets it owns and operates. When a manufacturer both produces and distributes its product, the firm is vertically integrated by ownership. The manufacturer can then control the retail promotion, customer service, pricing, product availability, delivery and other relevant decisions at the distribution stage. It does this through internal managerial decisions designed to maximize the overall profit of the firm. But this is not the only way to organize the production and distribution of the firm's output. Instead of having a network of its own distributors, it can license independent firms to perform the distribution function. These licensees are termed franchisees, and the distribution system is called a franchise system. The manufacturer then

engages in a contractual form of vertical integration. The franchise contract gives the franchisee the right to distribute the manufacturer's product, but also allows the manufacturer to control retail promotion, customer service, resale prices (minimum or maximum), and a host of other things that are important to the manufacturer, who is now a franchisor. But if franchising is just a contractual form of vertical integration, what is the rationale for franchising?

Rationale for franchising

Brands provide information to consumers in a mobile society, and so reduce search costs. When a consumer visits a branded outlet, there are supposed to be no surprises – pleasant or otherwise. Thus, local residents go to Roger's Ribs and Jessica's Java while visitors go to Sonny's Barbecue and Starbucks. The consumer's increased reliance on brand names has resulted in the development of chains of retail outlets. Chains benefit from lower costs as a result of economies of bulk purchasing and economies of scale in production, new product development, and promotion. In principle, all of the retail outlets could be corporately owned and managed, but many of the resulting chains are organized as franchises.

Franchising permits specialization that may lead to higher overall profits. The franchisor is the innovator who specializes in developing the brand, exploiting economies of scale in production and promotion, and negotiating with vendors on behalf of the chain. The franchisees bring their entrepreneurial spirit to their locations. They also contribute their knowledge of the local market. The result is a synergistic effect which increases potential profits (Caves and Murphy, 1976). Franchising succeeds because it allows each party to do what it does best. Further, franchise contracts deliberately organize this relationship to give the franchisor and franchisee incentives to work in tandem to increase revenues.

Types of franchise

Although the lines drawn are somewhat arbitrary, franchising can be divided into *traditional* franchising and *business format* franchising. Traditional franchising is used by a manufacturer to distribute its product through distributors (the franchisees) that are specifically licensed to do so. Traditional franchising is found in several industries: automobiles, beer, gasoline, ice cream, and soft-drink bottling to name a few. Business format franchising involves the use of the franchisor's brand, trademark and trade dress, and

distinctive way of supplying goods and services to the consumer. In this case, the franchisor develops and promotes the concept while the franchisees implement the concept and carry out local production and distribution. The ‘quick serve restaurant’ sector is a good example. with familiar names such as the US-based chains McDonald's, Pizza Hut, Subway and Taco Bell. There are many other sectors in business format franchising including accounting services, automobile servicing, health and fitness, hotels and motels, and real estate.

Franchisor compensation in traditional franchising

Traditional franchisors sell products to their franchisees which are resold to consumers. The franchisor earns its profit on these sales to franchisees. The extent to which franchising is a good substitute for vertical integration depends on two critical factors: the market structure in distribution and the relative efficiency of franchisees versus employee-managers. Assuming that franchisees are neither more nor less efficient than employee-managers, a manufacturer will earn the same profit whether it performs the distribution function itself or franchises the distribution. Its profit will be the same because the cost of performing the distribution function will be in the same in either case. If franchisees are more efficient than employee-managers, then distribution costs will be lower, retail demand will be higher, or both. This will improve the manufacturer's profit, due to increased sales to the franchisees.

If there is monopoly at the distribution stage, the market structure is one of successive monopoly. Assuming equal efficiency, the manufacturer will prefer vertical integration rather than franchising. Compared with the case of competitive distribution, successive monopoly results in double marginalization, which leads to lower output and higher price. The manufacturer's profits are reduced below the level that would result from vertical integration because double marginalization reduces the derived demand for the product.

These ill effects can be offset with other contractual provisions such as maximum resale price constraints, minimum quantity standards and price advertising (Blair and Esquibel, 1996). If these are effective, they should eliminate the exercise of downstream monopoly power and improve the manufacturer's profits.

Things are more complicated when franchisees are more efficient than

employee-managers. The effect of the increased efficiency is to shift the derived demand to the right. This, of course, tends to improve the manufacturer's profits. The exercise of monopoly power by the franchise, however, tends to decrease the manufacturer's profits. The net effect cannot be determined on a priori grounds. Presumably, when a manufacturer expects the net effect to be positive, franchising is selected.

Franchisor compensation in business format franchising

Business format franchisors do not sell products to their franchisees for resale. Instead, they provide a concept, a brand, and a distinctive way of doing business. Business format franchisors have many ways of charging their franchisees for using the licence: initial franchise fees, sales revenue royalties, output royalties, rent and sales of necessary inputs, to name a few.

Nearly all franchisors include franchise fees in their contracts. These fees are structured as either initial or periodic lump-sum payments, with the initial form being far more prevalent. Blair and Lafontaine (2005, p. 61) found that 99.2 percent of all franchisors use initial franchise fees. In principle, both forms allow the franchisor to capture the maximum profit available under vertical integration, provided that the franchisee can obtain all inputs at competitive prices. To realize this level of profit, the franchisor sets the initial franchise fee or the present value of the periodic payments equal to the present value of the stream of future operating profits that will be generated by the franchisee. This, in turn, will allow the franchisee to earn only a competitive return on its investment.

As in traditional franchising, the efficiency of the franchisees relative to employee-managers contributes to the level of profits the franchisor can obtain. If franchisees are more efficient, this will increase the franchisor's profit because franchisees will bid up the franchise fee until it is equal to the stream of future operating profits.

Although theoretically feasible, franchise fees generally are not set at such a high level. This is predominantly a result of the uncertainty surrounding future market conditions, future interactions between the parties, as well as franchisees' wealth constraints. Often franchise fees just cover the start-up costs of opening a new franchise location.

Franchisors can also extract revenue by charging franchisees royalties, based on either a percentage of their sales revenue or a fixed fee on output sold. Sales revenue royalties are the more prominent of the two forms and the

second most utilized charge by franchisors, next to initial franchise fees (Blair and Lafontaine, 2005, p. 66).

As long as the franchisee can acquire all inputs at competitive prices and a competitive market exists among local franchisees, the franchisor can extract the optimum level of profits with royalties. In the case of sales revenue royalties, the franchisor achieves this outcome by setting the royalty rate equal to the ratio of the difference between the profit maximizing price and marginal cost to the profit maximizing price. This rate will result in a post-royalty price equal to the franchisees' marginal cost. The royalties collected will be precisely equal to the maximum profits that vertical integration would yield. Again, franchisees' efficiency relative to employee-managers impacts these profits and the decision to franchise. If the franchisees are more efficient, then the franchisor will earn more profit due to the increase in sales revenue. If instead the employee-manager is more efficient, then the company will choose vertical integration.

If the franchisee has local monopoly power, however, the franchisor will be unable to attain the maximum level of profit through a sales royalty. The ability to exert local market power makes price endogenous for franchisees and allows them to factor sales revenue royalties into output decisions. Therefore, no sales revenue royalty will allow the franchisor to capture the amount of profit available through vertical integration. An equivalent analysis holds for output-based royalties (Blair and Kaserman, 1980). Output-based royalties, however, make little sense in business format franchises because there are often too many products to make tracking outputs feasible.

Another way in which franchisors obtain revenue is through input requirements. In many business format franchise systems, the franchisees are required to buy inputs from the franchisor. For example, a pizza franchisee may be required to buy pizza dough and sauces from the franchisor. If the franchisor charges its franchisees prices above the competitive level, these tying arrangements provide an alternative way of extracting the profit that vertical integration would provide (Blair and Kaserman, 1978). Again, however, this result holds only when there is competition among the franchisees; otherwise, the higher input prices cause distortions that reduce overall profits.

Choosing to franchise

The choice of franchising ultimately depends on whether it is more profitable

than vertical integration. This, in turn, depends on the structure of the particular market and the efficiency of franchisees relative to employee-managers. The less competitive the downstream market structure, the less attractive franchising becomes. But the more efficient franchisees are, the more attractive franchising becomes. The choice of whether and how to franchise will therefore vary from chain to chain.

See Also

- franchising
- vertical integration

Bibliography

Blair, R.D. and Esquibel, A. 1996. Maximum resale price restraints in franchising. *Antitrust Law Journal* 65, 157–80.

Blair, R.D. and Kaserman, D.L. 1978. Vertical integration, tying, and antitrust policy. *American Economic Review* 68, 397–402.

Blair, R.D. and Kaserman, D.L. 1980. Vertical control with variable proportions: ownership integration and contractual equivalents. *Southern Economic Journal* 47, 1118–28.

Blair, R.D. and Kaserman, D.L. 1983. *Law and Economics of Vertical Integration and Control*. Orlando. Fla: Academic Press.

Blair, R.D. and Lafontaine, F. 2005. *The Economics of Franchising*. New York: Cambridge University Press.

Caves, R.E. and Murphy, W.F. 1976. Franchising: firms, markets, and intangible assets. *Southern Economic Journal* 42, 572–86.

Coase, R.H. 1937. The nature of the firm. *Economica* 4, 386–405.

Kaufmann, P.J. and Lafontaine, F. 1994. Cost of control: the source of economic rents for McDonald's franchises. *Journal of Law and Economics* 57, 417–54.

Rubin, P. 1976. The theory of the firm and the structure of the franchise

contract. *Journal of Law and Economics* 21, 223–33.

Spengler, J.J. 1950. Vertical integration and antitrust policy. *Journal of Political Economy* 58, 347–52.

How to cite this article

Blair, Roger D. and Jessica S. Haynes. "economics of franchises." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_F000324> doi:10.1057/9780230226203.1902

The New Palgrave Dictionary of Economics Online

economics, definition of

Roger E. Backhouse and Steven Medema

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economics is difficult to define unambiguously, many definitions having been proposed as the subject has evolved. Definitions are *ex post* constructions, even rationalizations, but they can nonetheless influence what economist do and how they set about doing it. This article considers the main definitions from the late 18th century to the present, pointing out some of the ways in which changing views reflect and have influenced changes in the subject.

Keywords

cameralism; economic science; economics imperialism; economics, definition of; institutional design; Marshall, A.; Mill, J. S.; operations research; political economy; rational choice; Robbins, L. C.; Samuelson, P. A.; scarcity; Steuart, Sir J.; value judgements

Article

The definition of economics has evolved significantly over time, influenced by and influencing the focus of economic study. The definition often attributed to Jacob Viner, ‘economics is what economists do’, reflects the difficulty of providing an unambiguous definition. The problem, of course, is that definitions of the field are proposed *ex post* in an attempt to impose order upon a body of work that has grown up as economists have sought to tackle diverse practical and intellectual problems. Viner's statement suggests that there is no need for a tight, specific definition of the subject, which may explain the tendency of economists blithely to ignore definitions, and hence to not analyse them in detail, except sporadically. However, definitions of the subject do have effects through influencing what economists choose to study and the methods they think legitimate for analysing them.

The root of the word ‘economics’ lies in the Greek οἰκονομία, meaning the management of a household, as in Xenophon's *Οἰκονομικος*, written around 400 bc. In the 18th century, the idea of efficiently providing for the wants of a household was extended to the nation as a whole, under

the heading ‘political economy’, the term first used for the discipline that later became economics. The first systematic English-language book on the subject was James Steuart's *An Inquiry into the Principles of Political Oeconomy* (1767, p. 16). Though Steuart made an analogy between ‘providing for all the wants of a family, with prudence and frugality’ and doing the same for the state, there was a difference, for the ruler of the state could not direct people in the way that the head of a household was able to do. This had the consequence that,

> The great art therefore of political oeconomy is, first to adapt the different operations of it [the state] to the spirit, manners, habits and customs of the people; and afterwards to model these circumstances so, as to be able to introduce a set of new and more useful institutions. (Steuart, 1767, p. 16)

No doubt influenced by German Cameralism, Steuart saw institutional design as lying at the heart of political economy. This usage was followed by Adam Smith, who saw political economy as ‘a branch of the science of a statesman or legislator’ with two objects: providing the people with ‘plentiful revenue or subsistence’ and providing the state with enough revenue to provide public services (Smith, 1776, p. 428).

Many of the classical economists, however, disagreed with the focus on policy, arguing that political economy was concerned with the laws that govern the production, distribution and consumption of wealth, the clearest example of this being Jean Baptiste Say, whose major work (1803) is *Traité d’économie politique, ou simple exposition de la manière dont se forment, se distribuent et se consomment les richesses* (A treatise on political economy, or a simple account of the way in which wealth is formed, distributed and consumed). This definition formed the basis for Nassau Senior's *Outline of the Science of Political Economy* (1836) in which he argued that the science was based on four propositions, the first and most important of which was ‘That every man desires to obtain additional Wealth with as little sacrifice as possible’ (Senior, 1836, p. 26).

Neither of these definitions was acceptable to John Stuart Mill, whose ‘On the Definition of Political Economy; and the Method of Investigation Proper to It’, first published in 1836, was the last of his *Essays on Some Unsettled Questions of Political Economy* (1844). To define political economy as the rules for making a nation rich was to confuse ‘art’ and ‘science’. However, it was not enough to define it as the laws relating to the production and use of wealth, for these included many physical laws that lay outside its remit. He thus favoured a more limited definition: ‘The science which treats of the production and distribution of wealth, so far as they depend upon the laws of human nature’ or ‘The science relating to the moral or psychological laws of the production and distribution of wealth’ (Mill, 1844, p. 318).

Mill went on to argue that even this definition was too broad, for political economy related only to man in society.

The most significant challenge to this definition of political economy as, loosely, the science of wealth, came from Alfred Marshall, who offered the well-known definition:

> Political Economy or Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of wellbeing. (Marshall, 1890, p.

1)

This definition is significant not so much for changing the name of the discipline to economics as for its focus on the study of mankind. For Marshall, as for many of his generation, the evolution of human character was of crucial importance: it was important to study actual human behaviour, but it was important, especially in the longer run, to consider how activities and consumption served to influence character and hence behaviour. Wants could not be taken as given but depended on activities.

In these discussions there was, as Neville Keynes pointed out, an ambiguity in the use of the word ‘economic’. On the one hand it referred to attaining an end ‘with the least possible expenditure of money, time and effort’ (Keynes, 1891, pp. 1–2) whilst on the other hand it was used as an adjective corresponding to the noun, wealth. The economists who laid most emphasis on the first of these were the Austrians – Carl Menger and his successors – who focused on economizing behaviour. It was his familiarity with this literature that led Lionel Robbins to deny originality for his much-quoted definition, ‘Economics is the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses’ (Robbins, 1932, p. 16). Robbins's definition put scarcity and choice at the centre of economic analysis. He emphasized that ‘any kind of human behaviour’ that demonstrates the scarcity aspect falls within the scope of economics, and that there are ‘no limitations on the subject-matter of Economic Science’ beyond involving ‘the relinquishment of other desired alternatives’ (choice) (1932, p. 17). The significance of this definition lies in its analytical nature: instead of defining economics in terms of its subject matter, it defines it as an aspect of behaviour.

In spite of Robbins's claim that he was simply describing professional practice, the initial reaction of the profession to his definition of economics, at least as it surfaced in academic journal articles and introductory textbooks (where the definition of economics was primarily discussed), was negative (for a detailed discussion, see Backhouse and Medema, 2007). Throughout the 1930s and 1940s, textbook writers continued to define economics in terms more reminiscent of Mill and Marshall than Robbins, in that, even where reference was made to scarcity, this was frequently qualified: economics was described as a social science concerned with the study of wealth, of earning a living or a study of the system of free enterprise. Robbins's choice-based definition was seen as too wide, and needed to be restricted so as to rule out matters that did not come within the ‘traditional’ boundaries of economics. The acceptance of the Robbins definition came piecemeal. First, scarcity came to be stressed as important to the subject. The first edition of Paul Samuelson's *Economics* (1948), undoubtedly the leading textbook in the post-war period, captures well the qualified attitude with which the Robbins definition was approached. Samuelson explained that economics was about scarcity, for ‘the American way of life’ required more resources than were available, but he chose to define the subject in terms of ‘what’, ‘how’ and ‘for whom’ – that is, as concerning the production and consumption of goods and services. There is nothing here that is inconsistent with Robbins, but this approach was equally consistent with a more traditional approach. Books such as George Stigler's *Theory of Price* (1946), which adopted the Robbins definition, laid great stress on both scarcity and choice, but others carefully refrained from doing so.

It was only in the late 1950s and 1960s that the use of Robbins's definition became widespread. By the late 1960s, Samuelson's *Economics* was claiming that economists agreed on ‘a general definition something like the following’:

> Economics is the study of how men and society *choose*, with or without the use of money, to employ *scarce* productive resources, which could have alternative uses, to produce various commodities over time and distribute them for consumption, now and in the future, among various people and groups in society. (Samuelson, 1967, p. 5).

However, support for this was still not universal. For Richard Lipsey, whose *Introduction to Positive Economics* was one of the most successful rivals to Samuelson's *Economics*, scarcity was '*one of* the basic problems encountered in *most* aspects of economics', not the entire subject (Lipsey, 1963/71, p. 50). Economics also dealt with questions related to failure to achieve a point on the production possibility frontier, such as explaining unemployment, which could not be reduced to problems of scarcity.

The move by Robbins to define economics as an aspect of behaviour made it just a short step to defining economics in terms of a method – that of rational choice – which could be applied not simply to production and consumption choices, but to all of human behaviour. This move was encouraged by the tendency, in the aftermath of the Second World War, to see economics though the lens of operations research, as social engineering, in which optimization techniques were central and game theory played a significant role. It has also been argued that this move towards emphasising rational choice had ideological attractions during the Cold War. During the 1960s, economics became increasingly conceived as the 'science of choice', without reference to a particular social domain, even, at times, without reference to scarcity: the subject could encompass non-market as well as market activities. The work of Theodore Schulz and Gary Becker on human capital, James Buchanan, Anthony Downs and Gordon Tullock on political processes, and Becker on discrimination and on crime and punishment laid foundation for what came to be called 'economics imperialism', the application of economics to fields including politics, law, history, and sociology. These theoretical moves were reinforced by advances on the empirical side, where the techniques developed by, for example, James Heckman and Daniel McFadden for analysing cross-section data sets on individuals and households were used to investigate phenomena, such as non-marital fertility, that lie outside the traditional domain of economics as concerned with market behaviour.

Robbins's definition of economics in terms of the allocation of scarce resources remains the most widely cited definition of the subject, but it has never commanded universal assent. Though scarcity can be defined in such a way as to make it true, there have always been significant numbers of economists who have considered that it does not encompass all aspects of their discipline and that qualifications or extensions are required. These result in definitions closer to those found in the 19th-century literature, focusing on phenomena such as the production and distribution of wealth. At the other end of the spectrum, there are economists for whom rational choice is more fundamental than scarcity. To this extent, then, there is no universally agreed upon definition of the subject.

The reason this does not present a problem is that economists can proceed with their work irrespective of how their subject is defined. Definitions of fields generally come only after the field is established; as fields change, so definitions change. Despite this, however, definitions can matter. As Mill recognized, questions of method and definition are linked. The clearest example of this is Robbins, who sought to derive all the main propositions of economics from the premise of scarcity. His definition, therefore, was the basis for claiming that economic theory was central to economics – that it was far more important than Marshall had believed it to be. Also significant was his reference to economic science, for the word

science is far from neutral. Robbins had argued that value judgements, including those necessary to make interpersonal welfare comparisons, did not come within the scope of economic science, but belonged instead to the realm of ‘political economy’. In claiming this, he was arguably attempting to clarify the status of economists’ arguments, for, as he later made very clear, offering any advice on economic policy requires such value judgements. Thus if economics includes policy advice it must encompass more than economic science as Robbins defines it. However, such is the prestige of ‘science’ that Robbins's definition caused many economists to try to dispense with value judgements altogether, even in welfare economics. An exercise in clarification (and no doubt a critique of certain views of the subject) thus had the effect of significantly narrowing the subject. Attempting to define economics thus was not and is not simply a descriptive exercise; it has consequences for what economists do, and how they go about doing it.

See Also

- altruism, history of the concept
- Mill, John Stuart
- rationality, history of the concept
- Robbins, Lionel Charles
- Samuelson, Paul Anthony
- United States, economics in (1945 to present)

Bibliography

Backhouse, R.E. and Medema, S.G. 2007. Defining economics: Robbins's definition in theory and practice. SSRN working paper. Abstract online. Available at http://ssrn.com/abstract=969994, accessed 19 May 2007.

Keynes, J.N. 1891. *The Scope and Method of Political Economy*. London: Macmillan.

Lipsey, R.G. 1963. *An Introduction to Positive Economics*, 3rd edn. London: Weidenfeld & Nicolson, 1971.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan, 1949.

Mill, J.S. 1844. *Essays on Some Unsettled Questions of Political Economy*. In *Collected Works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press, 1967.

Robbins, L.C. 1932. *An Essay on the Nature and Significance of Economic Science*, 2nd edn. London: Macmillan, 1935.

Samuelson, P.A. 1948. *Economics*. New York: McGraw Hill.

Samuelson, P.A. 1967. *Economics*, 7th edn. New York: McGraw Hill.

Say, J.-B. 1803. *Traité d'économie politique, ou simple exposition de la manière dont se forment, se distribuent et se consomment les richesses*. Paris: Deterville.

Senior, N. 1836. *An Outline of the Science of Political Economy*. New York: Augustus Kelley, 1965.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vols. Indianapolis: Liberty Press, 1976.

Steuart, J. 1767. *An Inquiry into the Principles of Political Oeconomy*, 2 vols. Edinburgh: Oliver & Boyd, 1966.

Stigler, G.J. 1946. *The Theory of Price*. New York: Macmillan.

How to cite this article

Backhouse, Roger E. and Steven G. Medema. "economics, definition of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000291>
doi:10.1057/9780230226203.0442

The New Palgrave Dictionary of Economics Online

economy as a complex system

A.P. Kirman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Complex systems are composed of particles or agents which interact directly with each other. The rules for this interaction may be very simple and may not reflect the sort of rationality associated with standard economic models. Interaction is not through some exogenously given market, nor does it depend on the complicated reasoning involved in game theory. A complex system exhibits emergent aggregate properties as it organizes itself, and these can explain important phenomena such as bubbles, herding behaviour, and segregation. In each case the aggregate state of the economy or market could not be predicted from the average behaviour of the individuals.

Keywords

agent-based models; bubbles; chaos; clusters; complex systems; complexity; computational complexity; coordination; deterministic chaos; economy as a complex system; econophysics; emergence; equilibrium; ergodicity; evolution; excess volatility; financial market contagion; forecasting; game theory; herding; imperfect competition; law of large numbers; minority game; neighbourhood effects; Prisoner's Dilemma; punctuated equilibria; residential segregation; social interaction; social networks; statistical mechanics; statistical physics; steady state; tit for tat; tournaments

Article

Introduction

The term ‘complex system’ has been widely used in science and many different definitions have been given. Frequently, rather than give a definition of such a system, scientists have fallen back on certain characteristics that these systems exhibit. For example, emergence, self-organization, synergetics, collective behaviour, and non-equilibrium have all been cited in this regard. It is useful at the outset to make the distinction between ‘complexity’ and ‘complex system’. The former involves a number of

ideas which are important in economics and which are inherited from computer science but which will not be dealt with here. In particular, the notion of computational complexity, as it applies to decision-making or to the computation of equilibria or of dynamic programming problems is central to certain aspects of economic theory. However, here the discussion will turn on the idea of the economy as a complex, adaptive, evolving system. For economists the first real incarnation of this approach was with the introduction of deterministic chaos. The idea of complex dynamic behaviour, which would not explode or cycle or converge to a steady state, was fascinating for a science long dominated by the ideas of convergence to a static equilibrium or to a steady state. Jean-Michel Grandmont (1985) developed a simple model of ‘business cycles’ involving the ‘tent map’, which gave rise to such chaotic behaviour. Apart from the idea of the complicated dynamics involved, it was clear that the fact that a small perturbation in the initial conditions governing such a process could produce radically different trajectories was also of great intellectual interest. Two important innovations were involved. Firstly, there was the idea that the economy should be thought of as a truly dynamic system and that the initial conditions of such a system might play a key role. Secondly, there was the idea that there might be no continuity in the dependence on those initial conditions and that small changes might radically influence the trajectory of the system; hence the famous allusion to the influence of the fluttering of a butterfly's wing on the world's weather. These two aspects led economists to focus their attention on deterministic chaos. Yet in making such a close link between complexity and chaos, economists may have lost sight of the broader implications of complexity for the analysis of economic systems.

To see why this is so, consider what sort of systems are referred to as ‘complex’ in other disciplines. Typically they have some, or all, of the following characteristics:

- The agents are heterogeneous and interact directly with each other.
- The interaction and the information of agents are ‘local’.
- The agents’ behaviour is governed by simple ‘rules of thumb’.
- The aggregate behaviour of the system is not that of an ‘average’ or representative agent.
- This aggregate behaviour ‘emerges’ from the complicated interaction between the individuals.

To someone who has not studied theoretical economics, all of these characteristics might seem rather intuitive as features of an economy. Yet they are very different from the traditional view. In that view, the economy is a system in which the only interaction is through the market. By this it is meant that agents react to signals from some central authority such as an auctioneer. In some way the central prices adjust so as to coordinate the activities of the agents. The system adjusts in this way until the activities are coordinated – for example, in a market economy, until aggregate demand for all products is equal to the aggregate supply of those products. Once this is achieved, the signals will not change and no agent has an incentive to modify his behaviour and to deviate from this ‘equilibrium state’.

This description reveals another important feature of the collective model. No agent takes account of any influence that he might have on the outcome of the system. Many economists will react to this description by arguing that models of ‘imperfect competition’ abound, and in these models agents take into account the impact of their actions on the state of the system and know that other agents do the same.

This brings us to a second view of the economy, that based on game theory. Here all agents take account of the reciprocal impacts of their actions and know that all the other agents do the same. This view is

very different from the basic model of the economy. However, it is also very different from that of a complex system, since it attributes unlimited calculating capacity and depth of reasoning to the agents. The vision of the economy as a complex system falls between these two approaches. It requires neither the central coordinating mechanism of the competitive market, nor the analytically sophisticated players of game theory.

A good comparison might be between, on the one hand, an economy organized as a set of markets that are open simultaneously, each with an auctioneer and, on the other, an ants' nest. In the former there are structured central price-giving mechanisms, and the actors gradually reveal their willingness or non-willingness to pay until the goods are allocated efficiently. In the latter, the individuals pursue their own different activities and react to each other and to outside stimuli. The system organizes itself but there is no central mechanism for achieving such organization. No one would think of trying to describe the activity of an ants' nest by examining the behaviour of the 'representative ant' yet many would describe the allocation of effort and resources as 'efficient'.

The sort of system that could be described as complex in the sense outlined above can be physical or biological or social. A typical reaction to the use of physical or biological analogies in economics or other social sciences is that social systems are populated by individuals that have intentions and undertake purposeful activity, while the other systems are composed of purposeless molecules or particles. It is therefore argued that the sort of analysis that can be applied to the other systems is not pertinent to the analysis of economic systems. This reasoning does not stand up to close inspection. If individuals follow well-defined rules and their interaction is well specified, the simple models that are used in physical and biological models can be applied. Precisely why the individuals should follow these rules is a different question.

Why is the complex systems approach of particular interest currently to economists? Economic theory has recently been attacked on two fronts. The first is the problem of aggregation: how is the behaviour of the economic system related to that of the individuals that make it up? The second is the question of why individuals behave as they do. The answer to the first question is simple but undermines much of modern macroeconomics that is based on the idea that the behaviour of the aggregate can be treated as the behaviour of an individual. Yet what is known is that the standard model of a system composed of isolated individuals each solving his own maximizing problem does not allow one to treat the system as an individual. (This is not the place to enter into the details of this assertion but the basic argument is given in Kirman, 1992, and stems from the results of Sonnenschein, Mantel and Debreu.) The second question is that posed by behavioural economics that questions the idea of the isolated maximizing individual. Ideas from Simon (1957) onwards have suggested that individuals reason in a limited and local way. Experiments, observation, and examination of the neural processes utilized in making decisions all suggest that *homo economicus* is not an accurate or adequate description of human decision making. (For a good survey of the relevant literature, see Rabin, 1998.)

All of this suggests that one might want to take a very different view of how the economy functions. In particular, the notion of a complex system as used in many parts of science seems to correspond well to an intuitive vision of the economy. Just as in an ants' nest individuals perform tasks without having any idea of the behaviour of the system, individuals in an economy go about their business and achieve a remarkable degree of coordination. Take a simple example that of bees in a hive. The tasks for house bees are varied but temperature control is one of the important duties. When the temperature is low, bees cluster to generate heat for themselves, but when it is high some of them fan their wings to circulate air

throughout the hive. The general hive temperature required is between 33° and 36°C, while the brood chamber requires a constant heat of 35°. Honey has to be cured in order to ripen, and this also requires the help of circulating air. According to Crane (1999), 12 fanning bees positioned across a hive entrance 25•cm wide can produce an air flow amounting to 50–60 litres per minute. This fanning can go on day and night during the honey-flow season. Honeybees' wings beat 11,400 times per minute, thus making their distinctive buzz.

What is the lesson here for us? The typical economist's response to this phenomenon would be to consider a representative bee and then study how its behaviour responds to the ambient temperature. This would be a smooth function of temperature, wing beats going up or down with the temperature. Yet this is not what happens at all. Bees have different threshold temperatures and they are either on (beating at 11,400 beats per minute) or off. As the temperature rises more bees join in. Thus collectively with very simple 1, 0 rules the bees produce a smooth response. This sort of coordination, with each agent doing something simple, can only be explained by having a distribution of temperature thresholds across bees. Aggregation of individuals with specific local and differentiated behaviour produces smooth and sophisticated aggregate behaviour.

Nobody would argue that, in social systems, all coordination is achieved by simple interaction. Markets make a powerful contribution to economic coordination. Yet the important question is not whether such mechanisms exist, but how they come into being and develop and modify their rules. As already explained, the idea that the existence of such markets facilitates the allocation of resources is clear and generally accepted. What is not so clear is that the abstract idea of a market governed by centralized prices which are adjusted to equilibrate the market has any descriptive value. The idea of markets and networks of communication and transactions as emergent and changing phenomena is much more persuasive.

Considering the economy in this light is far from a new idea. When Adam Smith discusses the 'invisible hand' some of these notions are apparent, Pareto's work contains some of these ideas and Hayek is perhaps he who was closest to this vision. Schelling in his *Micromotives and Macrobehavior* (1978) clearly foresaw the role of self-organization. A recent development of these ideas had an introduction on the formal level by Foellmer (1974), who adopted the basic Ising model. He posited a system in which individuals were situated in space and whose preferences were dependent on those around them. This dependence was stochastic, that is, the probability of having certain preferences depended on the preferences of an individual's neighbours. If all the preferences are independently drawn, then one can determine the expected values of the equilibrium prices. However, if the interdependence of the individuals is too strong, this is no longer true. The 'law of large numbers' no longer applies. There is no easy transition from the micro to the macro level by simple averaging.

Foellmer's contribution was left to one side for a long time. However, the complexity approach to economics took on new life with the work at the Santa Fe Institute of a number of economists, physicists and other scientists such as Arthur, Bak, Blume, Durlauf, Geanakoplos, and Holland. A good picture of this sort of work can be found in *The Economy as an Evolving Complex System* (Anderson, Arrow and Pines, 1988) and the two additional volumes that followed it (Arthur, Durlauf and Lane, 1997; Blume and Durlauf, 2006).

The emphasis on the increasing 'socialization' of economics, which is intrinsic to models of interacting agents, permits one to introduce the influence of neighbours and groups on individual behaviour. Such an approach is standard in sociology and anthropology but has remained a very thinly populated field in

economics. A good survey of this work is to be found in Durlauf and Young (2001).

One important part of the research on complex systems in economics has been that on *agent-based* models. Here the idea is to look at a set of linked individuals whose behaviour is influenced by and which influences their neighbours, and to simulate the dynamics of that interaction. Perhaps the best-known early example of this was Axelrod's work on the Prisoner's Dilemma, which is summarized in Axelrod (1997). He started from a series of tournaments. The strategies used for these were those that individuals proposed for a repeated Prisoner's Dilemma game. These strategies were then played against each other in a series of tournaments and the winning strategy turned out to be ‘tit for tat’, which is basically cooperative.

Axelrod was concerned that those who had entered his tournament had already anticipated the strategies that would be proposed by others. To overcome this he ran simulations in which new strategies were introduced into the pool of existing strategies. To do this he assigned existing strategies randomly to his artificial agents and then modified them using a ‘genetic algorithm’. (For an introduction to the theory and use of genetic algorithms see Mitchell, 1996.) The set of strategies thus evolved in two ways. After the strategies had played against each other a new generation with more of the successful strategies was created. To these were also added new strategies generated by mutations and crossovers from the current population. After a while reciprocating strategies – that is, strategies which respond to cooperation with cooperation but which defect in the face of defection – took over, giving high payoffs. Here we have a selection process working on strategies that evolved rather than were consciously chosen. The behaviour of this basic but complex system – indeed, Axelrod refers to himself as a complexity scientist – led to the evolution of interesting aggregate characteristics. In this context it is also interesting to look at the work of Lindgren (1991), who also allowed the evolution of the strategy pool and generated periods of stability in which one strategy dominated, followed by periods of instability as the population was invaded by another strategy. This corresponds to the idea of ‘punctuated equilibria’ introduced into evolutionary theory by Eldredge and Gould (1972).

The notion of evolution, which can also be interpreted in the human or social context as adaptive learning, is important here. We can think of selection among a population of automata endowed with single strategies or of the idea that individuals learn to use more successful strategies.

Phase transitions

Recalling the characterization of complex systems given above, it is worth considering a few examples. In complex systems governed by local interactions, it may be the case that as a result of some perturbation there is a major change in aggregate behaviour. This is an important idea which is central to *statistical mechanics*. The idea here is that local interaction can generate a rapid transition from one ‘phase’ to another of an economic system and, more importantly, that one cannot simply apply the ‘law of large numbers’ to evaluate the impact of stochastic shocks. An example of this is provided by Bak et al. (1993), who consider a model of ‘self-organized criticality’ to describe an economy composed of a large number of productive units, each supplying a limited number of customers and, in turn, each supplied by a limited number of suppliers; both customers and suppliers are located near the productive unit.

The graph outlining the location of productive units is a cylindrical lattice. In other words, each

production unit is supplied by the firms above it on a vertical line and supplies the customers next to it on a horizontal line. The demand for each final good producer is characterized by stochastic fluctuations, which affects the variability of orders received by the suppliers. Such orders (and shocks) are locally and vertically correlated, as every final producer is supplied by the two upstream firms situated a line up along the network representing the productive system. In such a context, characterized by local interaction, Bak et al. (1993) prove that, if individual costs are non-convex, the aggregation of small independent individual shocks may lead to large aggregate fluctuations in the productive system, breaking therefore the law of large numbers. These small shocks do not cancel each other out but are amplified by their interaction. Thus fluctuations at the aggregate level cannot be explained by reducing the whole model to one of an individual.

Coordination: the Schelling model

Now let us pursue the discussion of the relationship between aggregate and individual behaviour. One of the important features of complex systems is that the system can coordinate on a solution which could not be predicted from a careful analysis of the average or typical individual. In other words, patterns at the aggregate level can emerge as the individuals in an economy or market interact with each other. The emergence of such aggregate patterns cannot be forecast from the specification of the individual characteristics. A good example of this was provided by Tom Schelling at the end of the 1960s (for a summary see Schelling, 1978). He introduced a model of segregation involving local interaction, in the sense that peoples' utility depends on the race of their neighbours. He showed that, even if people have only a very mild preference for living with neighbours of their own colour, as they move to satisfy their preferences complete segregation will occur.

The basic model is very simple. Take a large chess board, and place a certain number of black and white counters on the board, leaving some free places. A counter prefers to be on a square where half or more of the counters in his Moore neighbourhood, (the eight squares around him) are of its own colour (utility 1) to the opposite situation (utility 0). From the counters with utility zero, one is chosen at random and moves to a preferred location. This model, when simulated, yields complete segregation even though people's preferences for being with their own colour are not strong. Indeed, the result holds when individuals are happy even when more than half of their neighbours are of a colour different from their own. This result was greeted with surprise and has generated a large literature.

In fact, this result is not surprising and some simple physical theory (see Vinkovic and Kirman, 2006), can explain the segregation phenomenon. Numerous variants on Schelling's original model have been developed. In particular, the form of the utility function used by Schelling, the size of neighbourhoods, the rules for moving, and the amount of unoccupied space have all been studied (see Pancs and Vriend, 2007, for a survey). The physical model encompasses all of these variants.

An attempt to provide a formal structure has been made by Pollicot and Weiss (2001). They however, examine the limit of a Laplacian process in which individuals' preferences are strictly increasing in the number of like neighbours. In this situation it is intuitively clear that there is a strong tendency to segregation. Yet Schelling's result has become famous because the preferences of individuals for segregation were not particularly strong. The model is of interest because it illustrates the emergence of an aggregate phenomenon which is not directly foreseen from individual behaviour and because it

concerns an important economic problem, that of segregation.

The physical analogue to Schelling's model, developed in Vinkovic and Kirman (2006), exhibits three features of the resultant segregation. The first is the organization of the system into ‘regions’ or clusters, each containing individuals of only one colour. Second, it explains the shape of the frontier between the regions. Lastly, in the case where several clusters of one colour may form it allows one to analyse the size distribution of the clusters.

The basic idea is simple. Think of utility as the negative of energy. Particles with high energy in the physical system correspond to individuals with low utility in the social system. Where are the unhappy or high-energy individuals to be found? Clearly they are individuals on the frontiers of clusters. Those within clusters of their own colour are happy and have no possibility of increasing their utility by moving. Those on the frontier, on the other hand, are in contact with those of the other colour and there may be too many of the latter. In this case these individuals correspond to particles with high energy. A physical system with these characteristics will seek to minimize its energy. The energy is highest on the frontier between clusters. Thus the way for the system to minimize its energy is to reduce the length of these frontiers. It will achieve this by organizing itself into clusters, and the shape and size of these clusters will depend on the precise variant of the model. In the original model the system will organize itself into two giant clusters, each composed of individuals of one colour. If we only allow people to move to currently free places, then the number of these will be important for the outcome. If there are not enough, the system will ‘freeze’ with many small clusters. If, on the other hand, individuals can swap places the system will segregate, but there will be perpetual movement within it. Thus, a simple physical model generates the result obtained by Schelling and, furthermore, shows how the form of the segregation depends on the exact version of the model. (For a discussion of the emergent properties of the Schelling model see emergence.)

The ‘ El Farol Bar ’

Another interesting example of emergent coordination is that provided by Brian Arthur (1994) in his ‘El Farol Bar’ problem. The simple model that he develops and which has been taken up by many physicists under the name of ‘the minority game’ shows how individuals using rules -of thumb can come to coordinate in a way which yields a satisfactory social outcome even though no individual had any such intention. The idea is that the bar can hold 100 people. Being at the bar with fewer than 60 people is, by common consent, better than staying at home. However once attendance goes over 60 the bar becomes too crowded and home is the preferable alternative. The question then is how people will decide whether to go to the bar. Suppose that they all reason strategically. In this case they must decide in function of what their neighbours will decide. Thus, to anticipate whether there will be more than 60 people at the bar they must reflect on the strategies employed by the others. However, they must also take into account that the others are doing the same and know that the others know that they know that they are behaving in this way. This leads to an infinite regress that poses logical problems for the foundations of such game-theoretic reasoning. Rather than attribute such calculating capacities to his agents, Brian Arthur imagined that each was endowed with a set of forecasting rules based on previous attendance at the bar. Given his set of rules the individual chooses that rule which has forecast best up to the present, ‘best’ meaning the forecast that has the smallest sum of squared prediction errors, for example. Now, each agent uses, as information, just the attendance observed at the bar, and updates in consequence.

There is no coordinating mechanism, yet the model quickly settles to the ‘equilibrium’ solution with 60 people at the bar with occasional small deviations. Furthermore, each agent receives a fixed number of forecasting rules, some of which may be rather stupid. Nevertheless, coordination is achieved at the aggregate level.

Some things about this model are worth noting. It is not guaranteed that all agents will learn to forecast correctly; some may persist in erroneous forecasts. The way in which the model is set up means that whenever attendance goes to 61 many people are unhappy, which is not the case when it goes to 59. This asymmetry does not prevent the achievement of collective coordination, however. Thus, the relation between satisfactory performance at the aggregate level and satisfaction at the individual level is tenuous. While many may find this example intriguing, one might enquire as to how it can be directly applied to economic problems. An interesting answer is to be found in a book by some Oxford physicists who specialize in complex systems and who apply the model to financial markets (see Johnson, Jeffries and Hui, 2003, pp. 81–136).

Financial markets

This brings us to another important example, that of financial markets. Models of economies with interacting agents in the spirit of complex systems may, as we have just seen, be able to show how certain aggregate coordination may emerge. They may also help us to analyse some of the observed features of markets which normal economic analysis has difficulty explaining. For example, one of the major problems with the standard model of financial markets is that they do not reproduce certain well-established stylized facts about empirical price series. In standard models, where there is uncertainty about the evolution of prices, the usual way of achieving consistency is to assume that agents have common and ‘rational’ expectations. Yet, if agents have such common expectations, how can there be trade? Indeed there are many ‘no trade’ theorems for such markets. How, then, do we deal with the fact that the volume of trade on financial markets is very important and that agents do, in fact, differ in their opinions and forecasts and that this is one of the main sources of such trade? There is also an old problem of ‘excess volatility,’ that is, prices have a higher variance than the returns on the assets on which they are based. One answer is to allow for direct interaction between agents other than through the market mechanism. Models reminiscent of the Ising model from physics have been used to doing this. For example, one might suggest that individuals may change their opinions or forecasts as a function of those of other agents. In simple models of financial markets such changes may be self-reinforcing. If agents forecast an increase in the price of an asset and others are persuaded by their view, the resultant demand will drive the price up, thereby confirming the prediction. However, the market will not necessarily ‘lock on’ to one view for ever. Indeed, under certain rather reasonable assumptions, if agents make stochastic rather than deterministic choices, then it is certain that the system will swing back to a situation in which another opinion dominates. The stochastic choices are not irrational, however. The better the results obtained when following one opinion, the higher is the probability of continuing to hold that opinion.

Such models will generate swings in opinions, regime changes and ‘long memory’, all of which are hard to explain with standard analysis. An essential feature of these models is that agents are wrong for some of the time, but whenever they are in the majority they are essentially right. Thus they are not

systematically irrational. (For examples of this sort of model see, Lux and Marchesi, 1999; Brock and Hommes, 1997; and Kirman and Teyssiere, 2005, and for a recent survey, De Grauwe and Grimaldi, 2006.) Thus the behaviour of the agents in the market cannot correctly be described as ‘irrational exuberance’, in the well-known words of Alan Greenspan, Chairman of the Board of Governors of the Federal Reserve from 1987 to 2006.

Economists faced with this sort of model are often troubled by the lack of any equilibrium notion. The process is always moving; agents are neither fully rational nor systematically mistaken. Worse, the process never settles down to a particular price even without exogenous shocks. Suppose that we accept this kind of model: can we say anything analytic about the time series that result? If we consider some of these models, for certain configurations of parameters they could become explosive. There are two possible reactions to this. Since we will never observe more than a finite sample, it could well be that the underlying stochastic process is actually explosive, but this will not prevent us from trying to infer something about the data that we observe. Suppose, however, that we are interested in being able, from a theoretical point of view, to characterize the long-run behaviour of the system. In particular, if we treat the process as being stochastic and do not make a deterministic approximation, then we have to decide what, if anything, constitutes an appropriate long-run equilibrium notion. Such a concept provides an answer to those who consider that complex systems, by their nature, are not amenable to formal analysis. Foellmer, Horst and Kirman (2005), examined the sort of price process discussed here and produced some analytical results characterizing the process. Furthermore, they provided a long-run equilibrium notion that is not the convergence to a particular price vector.

If prices change all the time, as they will do in an evolving complex system, how may one speak of ‘equilibrium’? The idea is to look at the evolving distribution of prices and to try to characterize its long-run behaviour. Foellmer, Horst and Kirman (2005) examined the process governing the evolution of asset prices and the profits made by traders, and gave conditions under which it is ergodic, that is, the proportion of time that the price takes on each possible value converges over time and that the *limit distribution* is unique. (For a discussion of the mathematical background, see ergodicity and nonergodicity in economics.) This means that, unlike the ‘anything can happen’ often associated with deterministic chaos, in the long run the price and profits process does have a well-defined structure.

Conclusion

To view the economy as a complex system implies a fundamental rethinking of theoretical economics. The basic idea is that of a decentralized system with no central source of signals, whose aggregate behaviour cannot be reduced to that of an individual. Furthermore, the individuals are endowed with local information and interact directly with each other, and their behaviour can be characterized by simple rules. Such a vision is far from new in economics. Its origins can be traced back at least to Adam Smith and a long chain of economists leads from him to Hayek and Simon, who preceded the developments described here. The most recent contributions borrow heavily from other disciplines such as statistical physics and the appearance of ‘econophysics’ represents a shift from the path that led from classical mechanics to axiomatic mathematical models as the basic paradigm of economic theory. This sort of approach has already allowed economists to analyse problems such as contagion, neighbourhood effects, financial bubbles, and herding behaviour, none of which fits well into the standard economic

framework. In addition, many of the features that are imposed on standard models emerge as a result of the interaction between agents (see emergence).

Perhaps, most importantly, looking at economies in this way provides a very different and more intuitive vision of the economy as a vast interactive system whose aggregate properties reflect the self-organization of the system and its continual adaptation. However, entrenched ideas die hard and it remains to be seen whether Steven Hawking's prediction that the 21st century will be the ‘age of complexity’ will hold true for economics.

See Also

- emergence
- ergodicity and nonergodicity in economics
- interacting agents in finance
- social interactions (empirics)
- social interactions (theory)
- social networks, economic relevance of
- statistical mechanics

Bibliography

Anderson, P.W., Arrow, K.J. and Pines, D. 1988. *The Economy as an Evolving Complex System*. Redwood City, CA: Addison-Wesley.

Arthur, W.B. 1994. Inductive reasoning and bounded rationality. *American Economic Review* 84, 406–11.

Arthur, W.B., Durlauf, S.N. and Lane, D.A., eds. 1997. *The Economy as an Evolving Complex System II*. Reading, MA: Addison-Wesley.

Axelrod, R. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton, NJ: Princeton University Press.

Bak, P., Chen, K., Scheinkman, J. and Woodford, M. 1993. Aggregate fluctuations from independent sectoral shocks: self-organized criticality in a model of production and inventory dynamics. *Ricerche Economiche* 47, 3–30.

Blume, L. and Durlauf, S.N. 2006. *The Economy as an Evolving Complex System III*. Oxford: Oxford University Press.

Brock, W.A. and Hommes, C. 1997. A rational route to randomness. *Econometrica* 65, 1059–95.

Crane, E. 1999. *The World History of Beekeeping and Honey Hunting*. London: Duckworth.

De Grauwe, P. and Grimaldi, M. 2006. *The Exchange Rate in a Behavioral Finance Framework*. Princeton, NJ: Princeton University Press.

Durlauf, S.N. and Young, H.P. 2001. *Social Dynamics: Economic Learning and Social Evolution*. London: MIT Press.

Eldredge, N. and Gould, S.J. 1972. Punctuated equilibria: an alternative to phyletic gradualism. In *Models in Paleobiology*, ed. T.J.M. Schopf. San Francisco: Freeman, Cooper and Co.

Foellmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1, 51–62.

Foellmer, H., Horst, U. and Kirman, A. 2005. Equilibrium in financial markets with heterogeneous agents: a new perspective. *Journal of Mathematical Economics* 41, 123–55.

Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53, 995–1046.

Johnson, N., Jeffries, P. and Hui, P.M. 2003. *Financial Market Complexity*. Oxford: Oxford University Press.

Kirman, A. 1992. What or whom does the representative individual represent? *Journal of Economic Perspectives* 6(2), 117–36.

Kirman, A. and Teyssiere, G. 2005. Testing for bubbles and change points. *Journal of Economic Dynamics and Control* 29, 765–99.

Lindgren, K. 1991. Evolutionary phenomena in simple dynamics. In *Artificial Life II*, ed. C.G. Langton, C. Taylor, J.D. Farmer and S. Rasmussen. Redwood City, CA: AddisonWesley.

Lux, T. and Marchesi, M. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397, 498–50.

Mitchell, M. 1996. *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.

Pancs, R. and Vriend, N.J. 2007. Schelling's spatial proximity model of segregation revisited. *Journal of Public Economics* 91, 1–24.

Pollicott, M. and Weiss, H. 2001. The dynamics of Schelling-type segregation models and a nonlinear graph Laplacian variational problem. *Advances in Applied Mathematics* 27, 17–40.

Rabin, M. 1998. Psychology and economics. *Journal of Economic Literature* 36, 11–46.

Schelling, T. 1978. *Micromotives and Macrobehavior*. New York: W.W. Norton and Co.

Simon, H.A. 1957. *Models of Man: Social and Rational*. New York: Wiley.

Vinkovic, D. and Kirman, A. 2006. A physical analogue of the Schelling model. *Proceedings of the National Academy of Sciences* 103, 19261–5.

How to cite this article

Kirman, A. P. "economy as a complex system." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000246> doi:10.1057/9780230226203.0443

The New Palgrave Dictionary of Economics Online

econophysics

J. Barkley Rosser, Jr.

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Econophysics, a term neologized only in 1995, refers to physicists studying economics problems using conceptual approaches from physics. Certain ideas are emphasized, especially the ubiquity of scaling laws in distributions of financial returns, income and wealth, firm sizes, city sizes, and other economic phenomena. However, economists have been using many of these techniques since much earlier, and the influence of ideas from physics on economics dates as far back as 1801 at least. Arguably, if economics successfully absorbs the most useful of this work, ‘econophysics’ may cease to exist.

Keywords

Bachelier, L.; Black–Scholes formula; bounded rationality; Brownian motion; Canard, N.-F.; econobiology; econochemistry; econophysics; Lévy distribution; lognormal distribution; Pareto distribution; Pareto, V.; random walk; scaling laws; statistical mechanics

Article

According to Bikas Chakrabarti (2005, p. 225), the term ‘econophysics’ was neologized in 1995 at the second Statphys-Kolkata conference in Kolkata (formerly Calcutta), India, by the physicist H. Eugene Stanley, who was also the first to use it in print (Stanley, 1996). Mantegna and Stanley (2000, pp. viii–ix) define ‘the multidisciplinary field of econophysics’ as ‘a neologism that denotes the activities of physicists who are working on economics problems to test a variety of new conceptual approaches deriving from the physical sciences’.
The list of such problems has included distributions of returns in financial markets (Mantegna, 1991; Levy and Solomon, 1997; Bouchaud and Cont, 1998; Gopakrishnan et al., 1999; Sornette and Johansen, 2001; Farmer and Joshi, 2002), the distribution of income and wealth (Dra ˇ gulescu and Yakovenko, 2001; Bouchaud and Mézard, 2000; Chatterjee, Yarlagadda and Charkrabarti, 2005), the distribution of

economic shocks and growth rate variations (Bak et al., 1993; Canning et al., 1998), the distribution of firm sizes and growth rates (Stanley et al., 1996; Takayasu and Okuyama, 1998; Botazzi and Secchi, 2003), the distribution of city sizes (Rosser, 1994; Gabaix, 1999), and the distribution of scientific discoveries (Plerou et al., 1999; Sornette and Zajdenweber, 1999), among other problems, all of which are seen at times not to follow normal or Gaussian patterns that can be described fully by mean and variance. The main sources of conceptual approaches from physics used by the econophysicists have been from models of statistical mechanics (Spitzer, 1971), geophysical models of earthquakes (Sornette, 2003), and 'sandpile' models of avalanches, the latter involving self-organized criticality (Bak, 1996). An early physicist to assert the essential identity of statistical methods used in physics and the social sciences was Majorana (1942).

A common theme among those who identify themselves as econophysicists is that standard economic theory has been inadequate or insufficient to explain the non-Gaussian distributions empirically observed for various of these phenomena, such as 'excessive' skewness and leptokurtotic 'fat tails' (McCauley, 2004). With their sense of creating and developing a new science based on physics that is superior to the older conventional economics, many of the econophysicists have focused their publishing efforts in physics journals, notably *Physica A*, *Physical Review E*, and *European Physical Journal B*, to name some of the most frequently used ones, along with the general science journal *Nature* and some more clearly multidisciplinary journals such as *Quantitative Finance*. However, increasingly some of the econophysicists have begun to publish jointly with economists, with some of these papers appearing in economics journals as well. This should not be surprising in that the emergence of econophysics followed fairly shortly after the influential interactions and discussions that occurred between groups of physicists and economists at the Santa Fe Institute (Anderson, Arrow and Pines, 1988; Arthur, Durlauf and Lane, 1997), with some of the physicists involved in these discussions also becoming involved in the econophysics movement.

Now we come to a great curiosity and irony in this matter: some of the main techniques used by econophysicists were initially developed by economists (with many others developed by mathematicians), and some of the ideas associated with economists were developed by physicists. Thus, in a sense, these efforts by physicists resemble carrying coals to Newcastle, except that it must be admitted that many economists either forgot or never knew of these issues or methods. This is true of the most canonical of such models, the Pareto distribution.

The empirical focus on scaling laws (power laws)

If there is a single issue that unites the econophysicists it is the insistence that many economic phenomena occur according to distributions that obey scaling laws rather than Gaussian normality. Whether symmetric or skewed, the tails are fatter or longer than they would be if Gaussian, and they appear to be linear in figures with the logarithm of a variable plotted against its cumulative probability distribution. They search for physics processes, most frequently from statistical mechanics, that can generate these non-Gaussian distributions that obey scaling laws.

The canonical (and original) version of such a distribution was discovered by the mathematical economist and sociologist, Vilfredo Pareto, in 1897. Let N be the number of observations of a variable that exceed a value x with A and α positive constants. Then

$$N = Ax^{-\alpha}. \quad (1)$$

This exhibits the scaling property in that

$$\ln(N) = \ln A - \alpha \ln(x). \quad (2)$$

This can be generalized to a more clearly stochastic form by replacing *N* with the probability that an observation will exceed *x*. Pareto formulated this to explain the distribution of income and wealth, and believed that there was a universally true value for α that equalled about 1.5. More recent studies (Clementi and Gallegati, 2005) suggest that it is only the upper end of income and wealth distributions that follow such a scaling property, with the lower ends following the lognormal form of the Gaussian distribution that is associated with the random walk, originally argued for the whole of the income distribution by Gibrat (1931).

The random walk and its associated lognormal distribution is the great rival to the Pareto distribution and its relatives in explaining stochastic economic phenomena. It was only a few years after Pareto did his work that the random walk was discovered in a Ph.D. thesis about speculative markets by the mathematician Louis Bachelier (1900), five years prior to Einstein using it to model Brownian motion, its first use in physics (Einstein, 1905). Although the Paretian distribution would have its advocates for explaining stochastic price dynamics (Mandelbrot, 1963), the random walk would become the standard model for explaining asset price dynamics for many decades, although it would be asset returns that would be so modelled rather than asset prices themselves directly as Bachelier did originally. As a further irony, it was a physicist, M. F. M. Osborne (1959), who was among the influential advocates of using the random walk to model asset returns. It was the Gaussian random walk that would be assumed to underlie asset price dynamics when such basic financial economics concepts as the Black–Scholes formula would be developed (Black and Scholes, 1973). If we let *p* be price, *R* be the return due to a price increase, *B* be debt, and σ be the standard deviation of the Gaussian distribution, then Osborne characterized the dynamic price process by

$$\mathrm{d}p = Rp\mathrm{d}t + \sigma p\mathrm{d}B. \quad (3)$$

Meanwhile, a variety of efforts were made over a long time by physicists, mathematicians and economists to model a variety of phenomena using either the Pareto distribution or one its relatives or generalizations, such as the stable Lévy (1925) distribution, prior to the clear emergence of econophysics. Alfred Lotka (1926) saw scientific discoveries as following this pattern. George Zipf (1941) would see city sizes as doing so. Benoit Mandelbrot (1963) saw cotton prices doing so and was inspired to discover fractal geometry from studying the mathematics of the scaling property (Mandelbrot, 1983; 1997). Ijiri and Simon (1977) saw firm sizes also following this pattern, a result more recently confirmed by Axtell (2001).

Economists doing econophysics?

Also, economists would move to use statistical mechanics models to study a broader variety of economic dynamics prior to the emergence of econophysics as such. Those doing so included Hans Föllmer (1974), Lawrence Blume (1993), Steven Durlauf (1993), William Brock (1993), Duncan Foley (1994) and Michael Stutzer (1994), with Durlauf (1997) providing an overview of an even broader set of applications. However, by 1993 the econophysicists were fully active even if they had not yet identified themselves by this term.

While little of this work explicitly focuses on generating outcomes consistent with scaling laws, it is certainly reasonable to expect that many of them could. It is true that the more traditional view of efficient markets with all agents possessing full information rational expectations about a single stable equilibrium is not maintained in these models, and therefore the econophysics critique carries some weight. However, many of these models do make assumptions of at least forms of bounded rationality and learning, with the possibility that some agents may even conform to the more traditional assumptions. Stutzer's (1994) reconciles the maximum entropy formulation of Gibbsian statistical mechanics with a relatively conventional financial economics formulation of the Black–Scholes options formula, based on Arrow–Debreu contingent claims (Arrow, 1974). Brock and Durlauf (2001) formalize heterogeneous agents socially interacting within a utility maximizing, discrete choice framework. Neither of these specifically generates scaling law outcomes, but there is nothing preventing them from doing so potentially.

While some econophysicists seek to integrate their findings with economic theory, as noted above many seek to replace conventional economic theory, seeing it as useless and limited. An irony in this effort is that it has been argued that conventional neoclassical economic theory itself was substantially a result of importing 19th-century physics conceptions into economics, with not all observers approving of this (Mirowski, 1989). The culmination of this effort is seen by many as being Paul Samuelson's *Foundations of Economic Analysis* (1947), whose undergraduate degree was in physics at the University of Chicago. Samuelson himself noted approvingly that Irving Fisher's 1892 dissertation (1926) was partly supervised by the pioneer of statistical mechanics, J. Willard Gibbs (1902), and as far back as 1801 Nicholas-François Canard conceived of supply and demand ontologically being contradicting ‘forces’ in a physics sense. So the interplay between economics and physics has been going on for far longer and is considerably more complicated than is usually conceived.

Related trans-disciplinary movements

Curiously but unsurprisingly given the tremendous attention given to the new econophysics movement, it has spawned imitators since 2000 in the form of *econochemistry* and *econobiology*, although these have not had nearly the same degree of development. The former term is the title of a course of study established at the University of Ulm by Barbara Mez-Starke, and was used to describe the work of Hartmann and Rössler (1998) at a conference in 2002 in Urbino, Italy (see also Padgett, Lee and Collier, 2003, for a more recent effort). The latter term first appeared in Hens (2002), although McCauley (2004, pp. 196–9) dismisses it as not a worthy competitor for econophysics. Nevertheless, there has long been a tradition among economists of advocating drawing more from biology for inspiration than from physics (Hodgson, 1993), going back at least as far as Alfred Marshall's famous declaration that economics is ‘a branch of biology broadly interpreted’ (Marshall, 1920, p. 637), even as Marshall's actual analytical apparatus arguably drew more from physics than from biology.

In any case, one trend we can expect for some time is an increase in coauthoring between economists and physicists within the area of econophysics (Lux and Marchesi, 1999; Li and Rosser, 2004). Very likely we shall eventually see the more useful ideas of econophysics coming to be absorbed into economics proper. As that comes to pass, it may also come to pass that the separate and distinct movement we now know as econophysics will cease to exist and will be forgotten, just as most economists do not think about the physics roots of standard neoclassical economic theory today.

See Also

- economy as a complex system
- evolutionary economics
- Gibrat, Robert Pierre Louis
- Gibrat's Law
- inequality (global)
- inequality (measurement)
- lognormal distribution
- nonlinear time series analysis
- Pareto distribution
- Pareto, Vilfredo
- power laws
- redistribution of income and wealth
- science, economics of
- systems of cities
- transfer of technology
- urban growth

Bibliography

Anderson, P., Arrow, K. and Pines, D., eds. 1988. *The Economy as an Evolving Complex System*. Redwood City, CA: Addison-Wesley.

Arrow, K. 1974. *Essays in the Theory of Risk Bearing*. Amsterdam: North-Holland.

Arthur, W., Durlauf, S. and Lane, D., eds. 1997. *The Economy as an Evolving Complex System II*. Redwood City, CA: Addison-Wesley.

Axtell, R. 2001. Zipf distribution of firm sizes. *Science* 293, 1818–20.

Bachelier, L. 1900. Théorie de la spéculation. *Annales Scientifique de l'École Normale Supérieure* III-17, 21–86. (English translation in P. Cootner, ed. *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press, 1964.)

Bak, Per. 1996. *How Nature Works: The Science of Self-Organized Criticality*. New York: Copernicus Press for Springer-Verlag.

Bak, P., Chen, K., Scheinkman, J. and Woodford, M. 1993. Aggregate fluctuations from independent sectoral shocks: self-organized criticality in a model of production and inventory dynamics. *Ricerche Economiche* 47, 3–30.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Blume, L. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5, 387–424.

Botazzi, G. and Secchi, A. 2003. A stochastic model of firm growth. *Physica A* 324, 213–19.

Bouchaud, J.-P. and Cont, R. 1998. A Langevin approach to stock market fluctuations and crashes. *European Physical Journal B* 6, 543–50.

Bouchaud, J.-P. and Mézard, M. 2000. Wealth condensation in a simple model of economy. *Physica A* 282, 536–45.

Brock, W. 1993. Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance. *Estudios Económicos* 8, 3–55.

Brock, W. and Durlauf, S. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68, 235–60.

Canard, N.-F. 1801. *Principes d'Économie Politique*. Rome: Edizioni Bizzarri, 1969.

Canning, D., Amaral, L., Lee, Y., Meyer, M. and Stanley, H. 1998. A power law for scaling the volatility of GDP growth rates with country size. *Economics Letters* 60, 335–41.

Chakrabarti, B. 2005. Econphys-Kolkata: a short story. In *Econophysics of Wealth Distributions*, ed. A. Chatterjee, S. Yarlagadda and B. Charkrabarti. Milan: Springer.

Chatterjee, A., Yarlagadda, S. and Charkrabarti, B., eds. 2005. *Econophysics of Wealth Distributions*. Milan: Springer.

Clementi, F. and Gallegati, M. 2005. Power law tails in the Italian Personal income distribution. *Physica A* 350, 427–38.

Dra ˇ gulescu, A. and Yakovenko, V. 2001. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A* 299, 213–21.

Durlauf, S. 1993. Nonergodic economic growth. *Review of Economic Studies* 60, 349–66.

Durlauf, S. 1997. Statistical Mechanics approaches to socioeconomic behavior. In *The Economy as a Complex Evolving System II*, ed. W. Arthur, S. Durlauf and D. Lane. Redwood City, CA: Addison-Wesley.

Einstein, A. 1905. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von der ruhenden Flüssigkeiten suspendierten Teichen. *Annalen der Physik* 17, 549–60.

Farmer, J. and Joshi, S. 2002. The price dynamics of common trading strategies. *Journal of Economic Behavior and Organization* 49, 149–71.

Fisher, I. 1926. *Mathematical Investigations into the Theory of Value and Prices*. New Haven: Yale University Press.

Foley, D. 1994. A statistical equilibrium theory of markets. *Journal of Economic Theory* 62, 321–45.

Föllmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1, 51–62.

Gabaix, X. 1999. Zipf's law for cities: an explanation. *Quarterly Journal of Economics* 114, 739–67.

Gibbs, J. 1902. *Elementary Principles in Statistical Mechanics*. New Haven: Yale University Press.

Gibrat, R. 1931. *Les Inégalités Économiques*. Paris: Sirey.

Gopakrishnan, P., Plerou, V., Amaral, L., Meyer, M. and Stanley, H. 1999. Scaling of the distributions of fluctuations of financial market indices. *Physical Review E* 60, 5305–16.

Hartmann, G. and Rössler, O. 1998. Coupled flare attractors – a discrete prototype for economic modelling. *Discrete Dynamics in Nature and Society* 2, 153–9.

Hens, T. 2002. Evolutionary portfolio theory. *Asset Allocation Almanac: Special Report #4*. Merrill Lynch. Online. Available at http://www.evolutionaryfinance.ch/uploads/media/MerrillLynch.pdf, accessed 22 May 2006.

Hodgson, G. 1993. *Economics and Evolution: Bringing Life Back into Economics*. Ann Arbor: University of Michigan Press.

Ijiri, Y. and Simon, H. 1977. *Skew Distributions and the Sizes of Business Firms*. Amsterdam: North-Holland.

Levy, M. and Solomon, S. 1997. New evidence for the power-law distribution of wealth. *Physica A* 242, 90–4.

Lévy, P. 1925. *Calcul des Probabilités*. Paris: Gauthier-Villars.

Li, H. and Rosser, J., Jr. 2004. Market dynamics and stock price volatility. *European Physical Journal B* 39, 409–13.

Lotka, A. 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 12, 317–23.

Lux, T. and Marchesi, M. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397, 498–500.

Majorana, E. 1942. Il valore delle leggi statistiche nelle fisica e nelle scienze sociali. *Scientia* 36, 58–66.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Mandelbrot, B. 1983. *The Fractal Geometry of Nature*. San Francisco: W.H. Freeman.

Mandelbrot, B. 1997. *Fractals and Scaling in Finance*. New York: Springer-Verlag.

Mantegna, R. 1991. Lévy walks and enhanced diffusion in Milan stock exchange. *Physica A* 179, 232–

42.

Mantegna, R. and Stanley, H. 2000. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge: Cambridge University Press.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

McCauley, J. 2004. *Dynamics of Markets: Econophysics and Finance*. Cambridge: Cambridge University Press.

Mirowski, P. 1989. *More Heat than Light: Economics as Social Physics, Physics as Nature's Economics*. Cambridge: Cambridge University Press.

Osborne, M. 1959. Brownian motion in stock markets. *Operations Research* 7, 145–73.

Padgett, J., Lee, D. and Collier, N. 2003. Economic production as chemistry. *Industrial and Corporate Change* 12, 843–77.

Pareto, V. 1897. *Cours d'Économie Politique*. Paris and Lausanne. Trans. A. Schwier, *Manual of Political Economy*. New York: Kelly, 1971.

Plerou, V., Amaral, L., Gopakrishnan, P., Meyer, M. and Stanley, H. 1999. Similarities between the growth dynamics of university research and competitive economic activities. *Nature* 400, 433–37.

Rosser, J., Jr. 1994. Dynamics of emergent urban hierarchy. *Chaos, Solitons & Fractals* 4, 553–62.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sornette, D. 2003. *Why Stock Markets Crash: Critical Events in Complex Financial Systems*. Princeton: Princeton University Press.

Sornette, D. and Johansen, A. 2001. Significance of log-periodic precursors to financial crashes. *Quantitative Finance* 1, 452–71.

Sornette, D. and Zajdenweber, D. 1999. Economic returns of research: the Pareto law and its implications. *European Physical Journal B* 8, 653–64.

Spitzer, F. 1971. *Random Fields and Interacting Particle Systems*. Providence: American Mathematical Society.

Stanley, H., Afanasyev, V., Aamaral, L., Buldyrev, S., Goldberger, A., Havlin, S., Leschhorn, H.,

Maass, P., Mantegna, R., Peng, C.-K., Prince, P., Salinger, M., Stanley, M. and Viswanathan, G. 1996. Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. *Physica A* 224, 302–21.

Stanley, M., Amaral, L., Buldyrev, S., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. and Stanley, H. 1996. Scaling behavior in the growth of companies. *Nature* 379, 804–6.

Stutzer, M. 1994. The statistical mechanics of asset prices. In *Differential Equations, Dynamical Systems, and Control Science: A Festschrift in Honor of Lawrence Markus*, vol. 152, ed. K. Elworthy, W. Everitt and E. Lee. New York: Marcel Dekker.

Takayasu, H. and Okuyama, K. 1998. Country dependence on company size distributions and a numerical model based on competition and cooperation. *Fractals* 6, 67–79.

Zipf, G. 1941. *National Unity and Disunity*. Bloomington, IN: Principia Press.

How to cite this article

Rosser, J. Barkley, Jr. "econophysics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000253> doi:10.1057/9780230226203.0444

The New Palgrave Dictionary of Economics Online

Eden, Frederick Morton (1766 – 1809)

K. Tribe
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Eden, F. M.; poverty

Article

The son of Sir Robert Eden, F.M. Eden was educated at Oxford, gaining a Master's degree in 1789. A co-founder of the Globe Insurance Company, he published in 1797 the three volumes of his investigation into the conditions of the labouring poor, *The State of the Poor*. This work was perhaps the most detailed appraisal of social legislation and its actual workings that had appeared, and the findings provided ample material for ensuing debate on the best form of dealing with poverty and pauperism. In the years that followed Eden wrote a number of pamphlets on related issues.

The greater part of *The State of the Poor* records Eden's findings relating to the actual conditions prevailing in the parishes of England. Stimulated by the high prices prevailing in 1794–5, Eden initially set out to study the condition of the poor, but later extended this to the labouring classes. He encountered at times great resistance from local parish authorities, but despite this he was able to gather a considerable amount of information on wage levels, diet and prices. This was linked to an appraisal of the nutritional value of available foodstuffs, such that it was possible to arrive at some kind of comparative assessment of levels of poverty and want. It emerged from his empirical findings that the actual conditions and treatment of the poor varied greatly from parish to parish, this in part reflecting the patchwork of legislation that had grown up over the years in relation to the pauper and the workless. He argued however that existing legislation implied a policy of support for the indigent, and that in general a civilized society had an obligation to make such provision.

Selected work

1797. F. M. Eden, *The State of the Poor: Or an History of the Labouring Classes in England*, 3 vols. London.

How to cite this article

Tribe, K. "Eden, Frederick Morton (1766–1809)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000039> doi:10.1057/9780230226203.0445

The New Palgrave Dictionary of Economics Online

Edgeworth, Francis Ysidro (1845 – 1926)

John Creedy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Edgeworth was a major figure in the development of neoclassical economics, and one of its most original theorists, making a wide range of lasting contributions. After describing his approach to economics, this article discusses his early work in moral philosophy, which had a strong influence on his economics. His important contribution to the theory of exchange, focusing on indeterminacy and the role of the number of traders, is examined. His later work on monopoly, international trade and taxation are then briefly discussed.

Keywords

arbitration; assumptions; barter; Bentham, J.; Bickerdike, C.; bilateral monopoly; biological analogies; bootstrap; Butler, J.; calculus of variations; coalitions; collusion; combinations; competitive (price-taking) equilibrium; complements: *see* substitutes and complements; conjectural variations; contract curve; correlation coefficient; Cournot, A.; Darwin, C.; deductive method; determinacy and indeterminacy of exchange; distributive justice; Duopoly; Edgeworth box; Edgeworth, F.; egoism; equilibrium in exchange; experimental psychology; first fundamental theorem of welfare economics; Giffen good; Harsanyi, J.; Historical School; Hotelling, H.; idealism; immiserizing growth; indeterminacy of contract; indifference map; inference; international trade; international values, theory of; intuitionism; Jevons, W.; Keynes, J. M.; Lagrange multipliers; Laplace, P.; Launhardt, C.; law of indifference; Marshall, A.; mathematical economics; mechanical analogies; Mill, J. S.; monotonicity; moral philosophy; negative income tax; neoclassical; no-profit entrepreneur; offer curve; optimal distribution; optimal tariffs; Paley A. and M.; *Palgrave's Dictionary of Political Economy* ; partial equilibrium theory; Pearson distributions; physical sciences; Pigou, A.; probability; progressive and regressive taxation; rate of exchange; reciprocal demand curve; recontracting; Royal Statistical Society; rules of conduct; sacrifice theory of tax incidence; saddle point; Schumpeter, J.; Sidgwick, H.; social contract; social welfare function; statistical inference; substitutes and complements; tax incidence; taxation, theory of; transformations; utilitarianism; utility functions; utility maximization; Vickrey, W.; Walras, L.

Article

Biographia

Francis Ysidro Edgeworth (1845–1926) was born in Edgeworthstown in County Longford, Ireland. The background into which he was born was dominated by the ‘larger than life’ figure of his grandfather Richard Lovell Edgeworth (1744–1817), whose life was documented in a two-volume memoir (1820) by his oldest daughter, the famous novelist Maria Edgeworth (1767–1849). Richard Lovell's many scientific and mechanical experiments were helped by his strong association with the Lunar Society of Birmingham, whose members included Watt, Bolton, Wedgwood, Priestley, Darwin and Galton. In addition, Maria's scientific acquaintances included Davy, Humboldt, Herschel, Babbage, Hooker and Faraday. The marriage of F. Y. Edgeworth's cousin Harriet Jessie Edgeworth (daughter of Richard Lovell's seventh and youngest son Michael Pakenham, 1812–81) to Arthur Gray Butler provided links with another large and eminent academic family. These connections extend even further since A. G. Butler's sister, Louisa Butler, married Francis Galton, a cousin of Charles Darwin.

Richard Lovell's sixth son, and 17th surviving child, was Francis Beaufort Edgeworth (1809–46), who met his wife, Rosa Florentina Eroles, the daughter of a Spanish refugee from

Catalonia and then aged 16, while on the way to Germany to study philosophy; they married within three weeks in 1831. F. Y. Edgeworth was their fifth son. With his family background and his knowledge of French, German, Spanish and Italian, Edgeworth had wide international sympathies. On the family background, see Butler and Butler (1927) and for a full-length treatment of Edgeworth's work, see Creedy (1986).

Edgeworth was educated by tutors in Edgeworthstown until the age of 17, when in 1862 he entered Trinity College Dublin to study languages. In 1867 Edgeworth entered Exeter College, Oxford, but after one term transferred to Magdalen Hall. He transferred to Balliol in 1868, where in Michaelmas 1869 he obtained a first in Literae Humaniares. He was called to the bar in 1877, the same year in which his first book, *New and Old Methods of Ethics*, was published. Edgeworth applied unsuccessfully for a professorship of Greek at Bedford College, London, in 1875, but later lectured there on English language and literature for a brief period from late 1877 to mid-1878. He had earlier lectured on logic, mental and moral sciences and metaphysics to prospective Indian civil servants, at a private institution run by a Mr Walter Wren. In 1880 he applied for a chair of philosophy, also unsuccessfully, but began lecturing on logic to evening classes at King's College London. Soon after the publication of his second book, *Mathematical Psychics*, in 1881, he applied for a professorship of logic, mental and moral philosophy and political economy at Liverpool. Testimonials for two of Edgeworth's applications were given by Jevons (see Black, 1977, v, pp. 98, 145) and Marshall.

Edgeworth had to wait until 1890 until he obtained a professorial appointment: this was at King's College London, where he succeeded Thorold Rogers in the Tooke Chair of Economic Science and Statistics. In the next year, 1891, he again succeeded Rogers, this time to become Drummond Professor and Fellow of All Souls' College, Oxford, a position he held until his retirement in 1922. Edgeworth therefore finally settled in Oxford at the age of 46 in what was to become one of the most illustrious British chairs in economics. At the same time he became the first editor of the *Economic Journal*. He was editor or co-editor from its first issue until his death. He was supported by Henry Higgs from 1892 to 1905, when the latter became the Prime Minister's Private Secretary, with further assistance provided at a later stage by Alfred Hoare. Keynes was a co-editor for 15 years. After a tremendously creative period of the late 1870s and 1880s, Edgeworth had become firmly established as the leading economist, after Marshall, in Britain.

In addition to his work in economics, Edgeworth began a series of statistical papers in 1883. He was President of section F of the British Association in 1889, a position he held again in 1922. Edgeworth's work on mathematical statistics played an increasingly important role. Indeed, of about 170 papers which he published, approximately three-quarters were concerned with statistical theory. He became a Guy Medalist (Gold) of the Royal Statistical Society in 1907 and was President of the Society during 1912–14. His main contributions to statistics concern work on inference and the law of error, the correlation coefficient, transformations (what he called 'methods of translation'), and the 'Edgeworth expansion'. The latter, a series expansion which provides an alternative to the Pearson family of distributions, has been widely used (particularly since the work of Sargan, 1976) to improve on the central limit theorem in approximating sampling distributions. It has also been used to provide support for the bootstrap in providing an Edgeworth correction. Edgeworth's work in probability and statistics has been collected by McCann (1996). His third and final book was *Metretike: or the Method of Measuring Probability and Utility* (1887). These contributions are not examined here; see Bowley (1928) and Stigler (1978).

Approach to economics

A dominant characteristic of Edgeworth's approach to economics is that it is mathematical, characterized by an original use of techniques, although he does not appear to have received a formal training in mathematics. However, he came to economics from moral philosophy. The central question of distributive justice, rather than simply the application of mathematics, dominated his attitude towards economics. His main argument was that mathematics provided powerful assistance to 'unaided' reason, and could check the conclusions reached by other methods. Thus:

> He that will not verify his conclusions as far as possible by mathematics, as it were bringing the ingots of common sense to be assayed and coined at the mint of the sovereign science, will hardly realise the full value of what he holds, will want a measure of what it will be worth in however slightly altered circumstances, a means of conveying and making it current. (1881, p. 3)

Edgeworth's approach contrasts sharply with that of Marshall. The contrast between Edgeworth and Marshall was neatly summarized by Pigou as follows:

> During some thirty years until their recent deaths in honoured age, the two outstanding names in English economics were Marshall … and Edgeworth … Edgeworth, the tool-maker, gloried in his tools … Marshall, on the other hand, had what almost amounted to an obsession for hiding his tools away. (Pigou and Robertson, 1931, p. 3)

Although both men turned to economics from mathematics and moral philosophy, Marshall generally used biological analogies, and was concerned with developing maxims. In contrast, Edgeworth generally used mechanical analogies, and was more concerned with developing theorems.

In the 1880s and 1890s the deductive method encountered a great deal of criticism, especially from the 'Historical School' of economists. Edgeworth's defence of the deductive method often involved showing how other economists had advocated its use. His interest in the natural sciences often led him to make comparisons with scientific laws, and especially to show that the physical sciences also relied on abstraction and approximation.

Edgeworth argued carefully that the assumptions used in economics are often untestable, and he therefore took precautions against the accusation of 'plucking assumptions from the air'. He was conscious of the fact that the difficulty is in making the crucial abstractions which make the particular problem under consideration tractable, but which are not question begging. His attitude to many a priori assumptions was directly related to his approach to statistical inference. In *Mathematical Psychics*, for example, he referred to 'the first principle of probabilities, according to which cases about which we are equally undecided … count as equal' (1881, p. 99). This was then transferred to economics. The appropriate assumption was that all feasible values, say, of elasticities, were equally likely, until evidence is obtained. Hence, 'There is required, I think … in order to override the a priori probability, either very definite specific evidence, or the consensus of high authorities' (1925, ii, pp. 390–391). This also illustrates Edgeworth's attitude to authority and his many allusions to the views of other leading economists. Price (1946, p. 38) referred to his frequent 'reference to authority for … support of tentative opinion waveringly advanced'.

Edgeworth was also prone to stress negative results. For example, in discussing taxation, where the criterion of minimum sacrifice does not alone provide a simple tax formula, he stated:

> Yet the premises, however inadequate to the deduction of a definite formula, may suffice for a certain negative conclusion. The ground which will not serve as the foundation of the elaborate edifice designed may yet be solid enough to support a battering-ram capable of being directed against simpler edifices in the neighbourhood. (1925, p. 261)

Edgeworth's position as editor of the *Economic Journal* enabled him to combine both his critical attitude and his appetite for a wide range of reading. He contributed 32 book reviews, and in sending books to other reviewers he would include 'apposite remarks on particular points in the text' (Bowley, 1934, p. 123). These reviews should also be placed beside his 17 reviews in the Academy, and 131 articles in the original *Palgrave's Dictionary of Political Economy*. Furthermore, Edgeworth's later articles in the *Economic Journal*, such as those on international trade and on taxation, took the form of extended commentaries on contemporary work.

Early work in moral philosophy

Before turning to economics, Edgeworth published a brief note in *Mind* in 1876, and his first (privately printed) book on *New and Old Methods of Ethics* in 1877. The description by Keynes of Edgeworth's first book could just as well be applied to his other two books:

> Edgeworth's peculiarities of style, his brilliance of phrasing, his obscurity of connection, his inconclusiveness of aim, his restlessness of direction, his courtesy, his caution, his shrewdness, his wit, his subtlety, his learning, his reserve – all are there full-grown. Quotations from the Greek tread on the heels of the differential calculus. (Keynes, 1972, p. 257)

The main focus of this early work, strongly influenced by the great Cambridge philosopher Henry Sidgwick (1838–1900), was to examine in detail the implications of utilitarianism for the optimal distribution of resources. Edgeworth's special and original contribution was to apply advanced mathematics to this problem. Edgeworth's approach was dominated by his utilitarianism, but the influence of contemporary psychological research and the impact of evolutionary ideas can also be traced. Both aspects led to explicit consideration of differences between individuals and changes which take place over time.

Edgeworth was also influenced by the major fierce debates in the last half of the 19th century between egoism, evolutionism, idealism, intuitionism, and of course utilitarianism. His brand of utilitarianism became extremely eclectic, and embraced the majority of the above principles (except for those of the Hegelian idealists) while regarding utilitarianism as the 'sovereign principle'. His note in *Mind* discussed Matthew Arnold's views of Joseph Butler, who had examined egoism at great length. Arnold had argued that Butler's term 'self love' should be interpreted to mean 'the pursuit of our temporal good'. However, Edgeworth argued that egoism and utilitarianism could be subsumed under the same principle. He believed Butler to be saying, 'duty and interest are perfectly coincident; for the most part in this world, but entirely and in every instance, if we take in the future and the whole' (1876, p. 571).

Edgeworth generally distinguished between 'impure' and 'pure' utilitarianism. In the latter case individuals are assumed to be concerned with the welfare of society as a whole. The former case in fact corresponds more closely with a 'short term' version of egoism. Economic exchange can usefully be analysed in terms of 'jostling egoists', but he believed that ultimately individuals would evolve to become pure utilitarians. A reason for believing that individuals would make such a transition was later to be developed by Edgeworth in the form of his contractarian justification of utilitarianism as the appropriate principle of distributive justice.

Edgeworth's early utilitarianism was influenced by his wide knowledge of work in experimental psychology. In his books of 1877 and 1881 there are many references to the work of

Delboeuf, Fechner, Helmholtz, Weber and Wundt. These references occur in the context of discussing the nature of utility functions and, although Edgeworth at this time was not aware of the earlier work of Jevons, the same range of psychological work was also important to Jevons. Edgeworth in 1877 explicitly suggested, in connection with Fechner, that an additive form would not be appropriate.

A further aspect of Edgeworth's utilitarianism is his attitude towards authority. An important issue for early utilitarians involved the nature of inductive evidence about the consequences of acts. Most people cannot know the full consequences of their acts, so that rules of moral conduct must be followed (in contrast with intuitionism where individuals are assumed to have immediate consciousness of moral rules). In arriving at such rules, the opinions of highly regarded individuals are taken to be credible though it may not be possible to show conclusively that they are ‘correct’. Edgeworth argued, for example, that ‘we ought to defer even to the undemonstrated dicta and opinions of the wise, who have a power of mental vision acquired by experience’ (1925, ii, p. 149).

Edgeworth defined the problem of determining the optimal utilitarian distribution as follows: ‘given a certain quantity of stimulus to be distributed among a given set of sentients … to find the law of distribution productive of the greatest quantity of pleasure’ (1877, p. 43). In treating this problem mathematically Edgeworth used Lagrange multipliers, without any explanation, and concluded that, ‘unto him that hath greater capacity for pleasure shall be added more of the means of pleasure’ (1877, p. 43). In using Lagrange multipliers Edgeworth was also careful to discuss possible complications, referring to the possibility of multiple solutions and explicitly discussing corner solutions and inequality constraints. Further complexities were then examined, where Edgeworth emphasized that utilitarianism implies equality of the ‘means of pleasure’ only under a special set of assumptions, and in the general case the prescribed solution will be some form of inequality. In dealing with the distribution of effort, he argued not surprisingly that most work should be provided by those most capable of providing it. In a yet more general treatment of the problem, Edgeworth used the calculus of variations, but again provided the reader with virtually no help in following his mathematical argument. Edgeworth's analysis of the utilitarian optimal distribution was continued in his paper on ‘The Hedonical Calculus’ (1879), which was later reprinted as the third part of *Mathematical Psychics*.

Early work in economics

The turning point in Edgeworth's work was his introduction to Jevons in 1879 by a mutual friend James Sully, who in 1878 moved to Hampstead, where Edgeworth had lodgings in Mount Vernon and where Jevons also lived; see Sully (1918, pp. 180, 223). His first knowledge of Marshall came from Jevons, who ‘highly praised the then recently published Economics of Industry’ (in Pigou, 1925, p. 66). Edgeworth became interested in the problem of the indeterminacy of the rate of exchange, arising from the existence of only a small number of transactors. This led rapidly to Edgeworth's second and most important book *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences* (1881), which was clearly written in a state of considerable enthusiasm for his new subject. This slim volume of 150 pages was known only to a small group of experts. Marshall's review began, ‘this book shows clear signs of genius, and is a promise of great things to come’ (Whitaker, 1975, p. 265). Jevons began by stating that ‘whatever else readers of this book may think about it, they would probably all agree that it is a very remarkable one’ (1881, p. 581). It was not until the middle of the 20th century that many of its central ideas began to be more fully appreciated.

Part 1 of *Mathematical Psychics* (1881, pp. 1–15) was devoted mainly to a justification of the use of mathematics in economics where precise data are not available. There is probably no other ‘apology’ in the whole of economic literature which compares with Edgeworth's plea for the application of mathematics. For example, when considering individual utility maximization:

> Atoms of pleasure are not easy to distinguish and discern; more continuous than sand, more discrete than liquid; as it were nuclei of the just-perceivable, embedded in circumambient semi-consciousness. We cannot count the golden sands of life; we cannot number the ‘innumerable smile’ of seas of love; but we seem to be capable of observing that there is here a greater, there a less, multitude of pleasure-units; mass of happiness; and that is enough. (1881, pp. 8–9)

Great stress was placed on comparison with Lagrange's ‘principle of least action’ in examining the overall effects produced by the interactions among many particles. The connection with Edgeworth's analysis of competition, involving interaction among a large number of competitors to produce a determinate rate of exchange, is central here. The fact that in the natural sciences so much could be derived from a single principle was important for both Jevons and Edgeworth. But Edgeworth took this to its ultimate limit in arguing that the comparable single principle in social sciences, that of maximum utility, would produce results of comparable value. Referring to Laplace's massive work, *Mécanique Céleste*, he suggested that:

> ‘Mécanique Sociale’ may one day take her place along with ‘Mécanique Celeste’ [sic], throned each upon the double-sided height of one maximum principle, the supreme pinnacle of moral as of physical science … the movements of each soul, whether selfishly isolated or linked sympathetically, may continually be realising the maximum energy of pleasure, the Divine love of the universe. (1881, p. 12)

Jevons's work in the *Theory of Political Economy* involved the application of very basic mathematics and of psychological research to the analysis of exchange in competitive markets. In addition to this direct stimulus, Edgeworth was also influenced by an anonymous review of Jevons's book in the *Saturday Review* (1871).

The crucial development following Edgeworth's contact with Jevons was not simply the realization that mathematics could be used to examine equilibrium in exchange. Rather, it was that in his analysis Jevons explicitly assumed, through his ‘law of indifference’, that all individuals take the equilibrium prices as given, that is, outside their control. In using this law as ‘one of the central pivots of the theory’, Jevons stated that, ‘there can only be one ratio of exchange of one uniform commodity at any moment’ (1871, p. 87). His theory was explicitly limited to the static equilibrium conditions. He deliberately excluded the role of the number of competitors from his analysis via the awkward notion of the ‘trading body’, following correspondence with Fleeming Jenkin (1833–85), who raised the question of indeterminacy with just two traders; see Black (1977, iii, pp. 166–78). Jenkin could not see why two isolated individuals should accept the price-taking equilibrium, whereas Jevons wished to consider the behaviour of two typical individuals in a large market.

In a section on ‘Failure of the Laws of Exchange’, Jevons discussed cases in which some indeterminacy would result. His most notable example was of house sales, where it was suggested that indeterminacy would result from the discrete nature of the good being exchanged. The *Saturday Review* article took exception to this, suggesting that indeterminacy ‘is really owing in our opinion to the assumed absence of competition’ (see Black, 1981, p. 157). The stress on indeterminacy was also influenced by Marshall's discussion of wage bargaining: Edgeworth (1881, p. 48 n.1) referred to Thornton's comparison of the determination of prices in Dutch and English auctions, and cited Alfred and Mary Paley Marshall's joint book on the *Economics of Industry* (1879).

It was this gap in Jevons's analysis that Edgeworth set out to fill. His achievement was to show the conditions under which competition between buyers and sellers, through a barter process, leads to a ‘final settlement’ which is equivalent to one in which all individuals act independently as price takers. As he later stated (1925, p. 453), ‘the existence of a uniform rate of exchange between any two commodities is perhaps not so much axiomatic as deducible from the process of competition in a perfect market’.

Exchange and contract

Having argued that ‘the conception of Man as a pleasure machine may justify and facilitate the employment of mechanical terms and Mathematical reasoning in social science’ (1881, p. 15), Edgeworth moved on to the analysis of the ‘economical calculus’, the starting point of which was the assumption that ‘every agent is actuated only by self-interest’ (1881, p. 16).

In modern economic analysis the analytical tools invented by Edgeworth in 1881, such as the indifference map and the contract curve, are now used in a vast range of contexts. They were introduced by Edgeworth to examine the nature of barter among individuals. He wanted to see if a determinate rate of exchange would be likely to result in barter situations where it is assumed only that individuals wish to maximize their own utility, considered solely as a function of their own consumption. With full knowledge of individuals’ utility functions, and their initial endowments of goods, would it be possible to work out a ‘determinate’ rate of exchange at which trade would take place? Edgeworth's direct statement of the problem is as follows:

> The PROBLEM to which attention is specially directed in this introductory summary is: How far contract is indeterminate – an inquiry of more than theoretical importance, if it show not only that indeterminateness tends to [be present] widely, but also in what direction an escape from its evils is to be sought. (1881, p. 20)

Edgeworth began his analysis of this problem by taking the simplest case of two individuals exchanging fixed quantities of two goods. The basic framework is that described by Jevons, where the first individual holds all of the initial stocks of the first good, and the second individual holds all the stocks of the second good. He wrote the utility functions of each individual in terms of the amounts exchanged rather than consumed, using the general utility function (‘utility is regarded as a function of the two variables, not the sum of two functions of each’, 1881, p. 104). He then immediately defined the contract curve and indifference curves, in that order.

In the sentence which follows Edgeworth's introduction of the general utility function, he raised the question of the equilibrium which may be reached with ‘one or both refusing to move further’. In barter the conditions of exchange must be reached by voluntary agreement, or contract, between the two parties, and of course it is fundamental that no egoist would agree to a contract which would make him worse off than before the exchange. The question thus concerns the nature of the settlement reached by two contracting parties. He immediately answered that contract supplies only part of the answer so that ‘supplementary conditions … supplied by competition or ethical motives’ are required, and then wrote the equation of his famous contract curve (1881, pp. 20–1).

The problem of obtaining the equilibrium values of x and y which, ‘cannot be varied without the consent of the parties to it’ was stated as follows: ‘It is required to find a point (x, y) such that, in whatever direction we take an infinitely small step, $[U_A]$ and $[U_B]$ do not increase together, but that, while one increases, the other decreases’ (1881, p. 21). The locus of such points ‘it is here proposed to call the contract-curve’. Edgeworth's alternative derivations of the contract curve involved the movement, from an arbitrary position, along one person's indifference curve; ‘motion is possible so long as, one party not losing, the other gains’ (1881, p. 23). He thus used the Lagrange multiplier method of maximizing one person's utility subject to the condition that the other person's utility remains constant.

In the diagram drawn by Edgeworth (1881, p. 28) he did not use a box construction. Furthermore the only indifference curves shown fully were those which each individual is able to

reach in isolation, and which therefore specify the limits beyond which each is not prepared to move. Also part of the offer or reciprocal demand curves of each individual were drawn on the same diagram, although they were not defined until ten pages later.

After presenting the results for the two-person two-good case, Edgeworth (1881, p. 26) examined the contract curve in the case where three individuals exchange three goods, stated that it is given by the 'eliminant', and then gave three lines of three sets of partial derivatives. In fact, the contract curve in this context is defined by $\left|\frac{\partial U_i}{\partial x_j}\right| = 0$, where $\frac{\partial U_i}{\partial x_j}$, is the marginal utility of person *i* with respect to good *j*, but Edgeworth did not use the modern notation for determinants and did not set the Jacobian equal to zero. This early use of determinants in economics would probably have confused many of his readers.

The problem of indeterminacy

The concepts of indifference curves and the contract curve therefore help to specify a range of 'efficient exchanges' of goods between individuals. The essential feature of the analysis from Edgeworth's point of view is precisely that there is a range rather than a unique point: 'the settlements are represented by an indefinite number of points' (1881, p. 29). At any particular settlement, the rate of exchange is expressed simply in terms of the amount of one good which is given up in order to obtain a specified amount of the other good. Hence the existence of a range of efficient contracts means that the rate of exchange is 'indeterminate'. The rate of exchange achieved in practice will thus depend to a large extent on bargaining strength. It was this result which led Edgeworth to make his often quoted remark that 'an accessory evil of indeterminate contract is the tendency, greater than in a full market, towards dissimulation and objectionable arts of higgling' (1881, p. 30).

Edgeworth argued that his analysis of indeterminacy in contract between two traders could be applied to a very wide variety of contexts. In particular, the tendency of large groups to form 'combinations', as in the case of trade unions and employers' associations, would serve to increase the extent of indeterminacy. The general applicability of his analysis of contract and indeterminacy was summarized by Edgeworth as follows:

> What it has been sought to bring clearly into view is the essential identity (in the midst of diversity of fields and articles) of contract; a sort of unification likely to be distasteful to those excellent persons who are always dividing the One into the Many, but do not appear very ready to subsume the Many under the One. (1881, p. 146; Plato's expression 'the one in the many' was later used by Marshall as the motto for his 1919 book on *Industry and Trade*.)

Having shown the possibilities of indeterminacy, Edgeworth then went on to show how 'the escape from its evils' requires either competition or arbitration.

Competition and the number of traders

The central question which Edgeworth was trying to resolve in the second part of *Mathematical Psychics* was that of the conditions necessary to remove the indeterminacy which exists in the case of barter between two traders. The question naturally arises as to the extent to which this indeterminacy is the result of the absence of competition in the simple two-person market. Edgeworth thus quickly moved on to the introduction of further traders.

In Edgeworth's earlier problem of two traders exchanging two goods, the definition of a range of efficient exchanges (along the contract curve) is of course analytically separate from the question of whether or not two isolated traders would actually reach a settlement on the contract curve. However, these two aspects were not clearly separated by Edgeworth because at the beginning of his analysis he introduced his stylized description of the process of barter: this is the famous 'recontracting' process. Edgeworth did not wish to assume that individuals initially have perfect knowledge. Instead, he supposed that, 'There is free communication throughout a normal competitive field. You might suppose the constituent individuals collected at a point, or connected by telephones – an ideal supposition, but sufficiently approximate to existence or tendency for the purposes of abstract science' (1881, p. 18). The knowledge of the other traders' dispositions and resources could be obtained by the formation of tentative contracts which are not assumed to involve actual transfers, and can be broken when further information is obtained. Edgeworth introduced this in typical style:

> 'Is it peace or war?' asks the lover of 'Maud', of economic competition, and answers hastily: it is both, pax or pact between contractors during contract, war, when some of the contractors without the consent of others recontract. (1881, p. 17; the allusion here is to Alfred Tennyson's poem *Maud: A Monodrama*, part 1, verse VII.)

An important role of the recontracting process is thus to disseminate information among traders. It allows individuals who initially agree to a contract, which is not on the contract curve, to discover that an opportunity exists for making an improved contract according to which at least one person gains without another suffering.

However, the real importance of the recontracting process lies in the fact that it allows for Edgeworth's analysis of the role of the number of individuals in a market. With numerous individuals, the recontracting process makes it possible to analyse the use of collusion among some of the traders. Individuals are allowed to form coalitions in order to improve

bargaining strength. Recontracting enables the coalitions to be broken up by outsiders who may attract members of a group away with more favourable terms of exchange.

Edgeworth's analysis was extremely terse and the following discussion does not therefore follow his own presentation. The analysis begins by introducing a second person *A* and a second person *B*. The new traders are assumed to be exact replicas of the initial pair, with the same tastes and endowments. This simplification is useful because the dimensions of the Edgeworth box and the utility curves are identical for each pair of traders. Hence, it enables the same diagram to be used as in the case when only two traders are considered in isolation. Two basic points can be stated immediately. First, in the final settlement all individuals will be at a common point in the Edgeworth box. Second, the settlement must be on the contract curve. The first point arises because if two individuals have identical tastes then their total utility is maximized by sharing their resources equally. It is useful to consider other types of contract which will eventually be broken, in order to illustrate the way in which the introduction of additional traders provides a role for some kind of competitive process.

The major question at issue is whether the range of indeterminacy along the contract curve is reduced by the addition of these traders. Consider Figure 1 and suppose that when A_1 and B_1 are trading independently of A_2 and B_2, trader B_1 has all the bargaining power and is able to appropriate all the gains from trade by pushing A_1 to the limit of the contract curve at point *C*. Suppose also that the same applies to A_2 and B_2. If the two pairs of traders are then able to communicate with each other, A_2 can now simply refuse to trade with B_2 at *C*. With no transaction costs, A_2 was previously indifferent between trading at *C* and consuming at the endowment point, *E*. This endowment position is effectively the 'threat point' of the *A*s: it is the position in which they would find themselves if the bargaining process were to break down. But A_2 no longer needs to remain in isolation after refusing to trade with B_2, and instead can trade with A_1, after A_1 has traded with B_1 at *C* and has therefore obtained some of good *Y*. The two *A*s can share their stocks of *X* and *Y* equally, arriving at point *P*; such an equal division maximizes their total utility.

Figure 1

Two pairs of traders

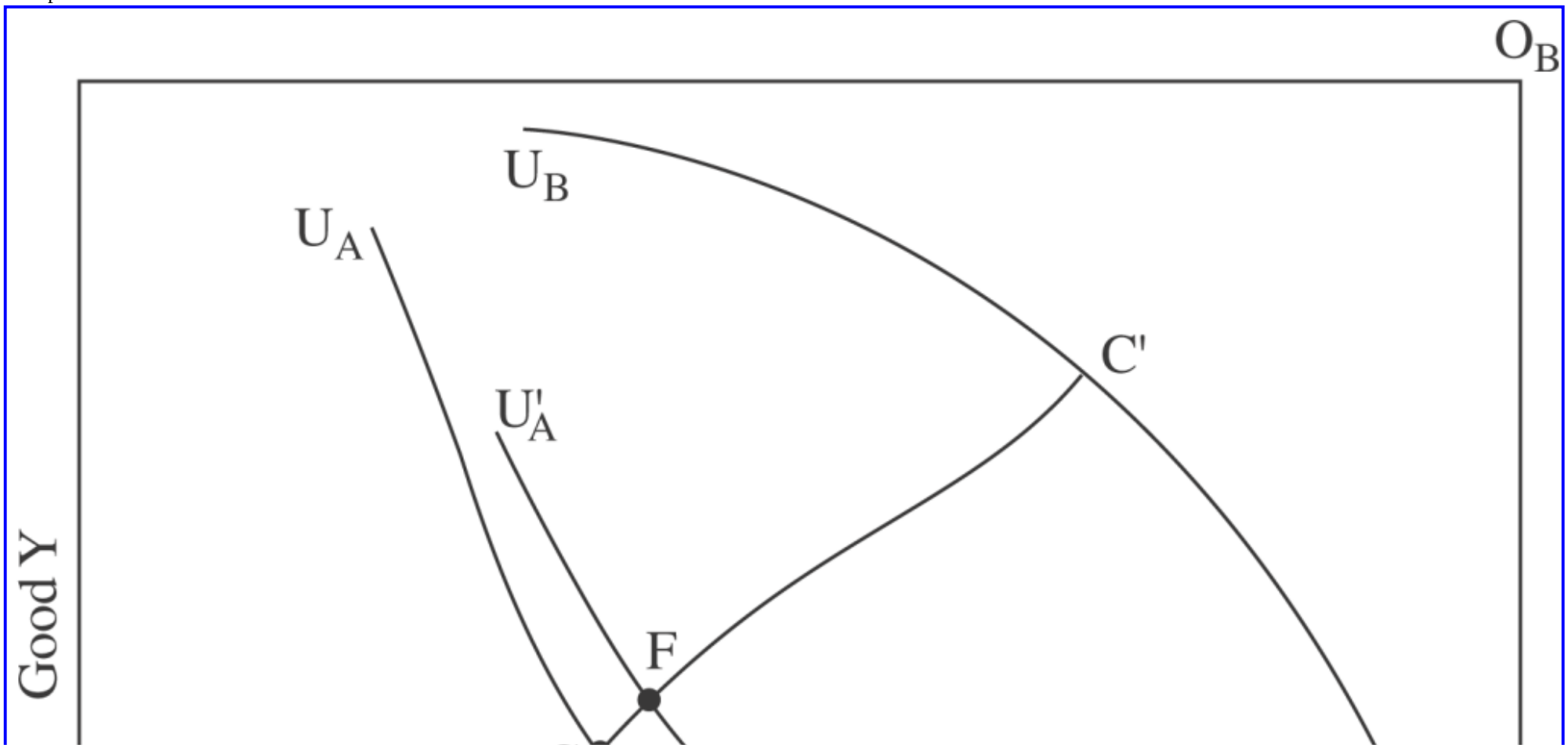

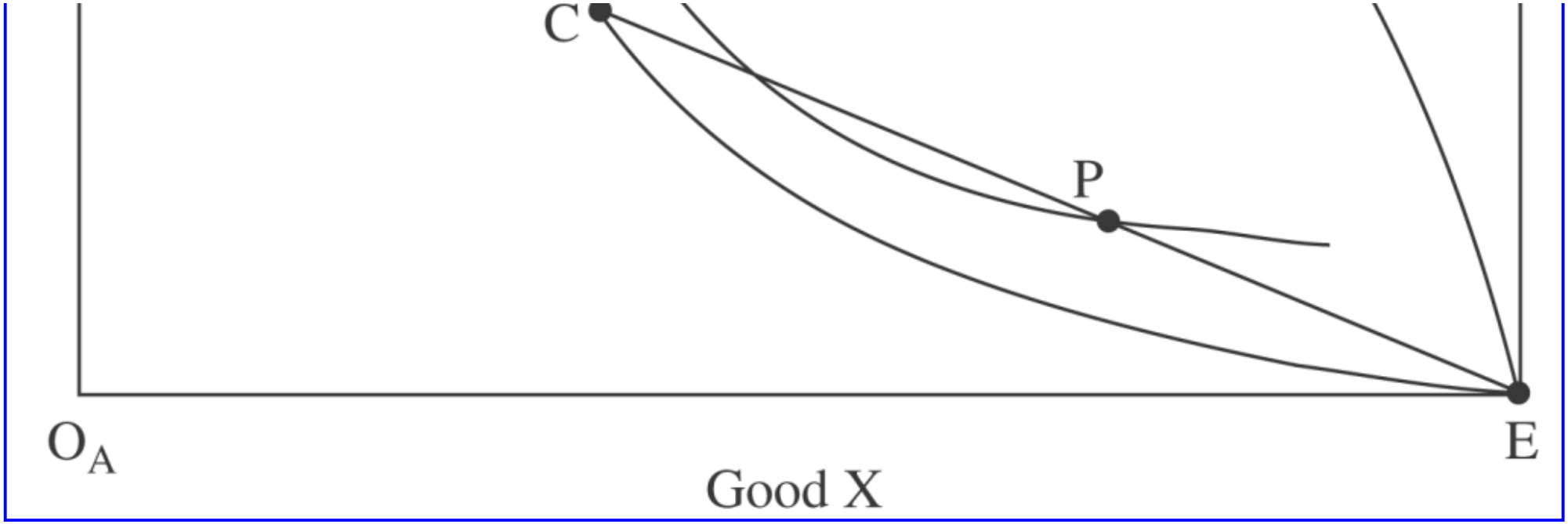

By reaching point *P*, halfway between *C* and *E*, the convexity of the indifference curves implies that they are both better off than anywhere on the no-trade indifference curve. The two *A*s would be on a higher common indifference curve, and thus better-off, if they could consume at a point along the *CPE* which is to the north-west of point *P*. However, they do not have enough resources to move beyond the halfway point *P*.

Trader B_2, who has been isolated, cannot prevent such a bargain. Thus B_1 is at *C*, both *A*s are at *P* and B_2 is at the initial endowment point *E*. In this situation B_1 has no incentive to change, but B_2 has a strong incentive to offer a better deal to one of the *A*s than the one offered by trader B_1. So long as B_2 offers one of the *A*s, say A_2, a trade on the contract curve which allows A_2 to reach a higher indifference curve than U'_A, the initial agreement with B_1 will be broken and recontracting will take place.

The implication is that the ability of the *A*s to turn to someone else, rather than deal with a single trader, means that the *B*s now compete against each other. However, trader B_1, who cannot prevent the recontracting, has an incentive to make yet a better offer. Hence, the recontracting process continues. The stylized process of recontracting with the two *B*s competing against each other will produce a final settlement at the point C^* in Figure 2. This has the property that the indifference curve U''_A passes through C^* and P^*, where P^* is halfway between C^* and *E*. This means that the two *A*s are indifferent between C^* and P^*, and since they cannot both reach any point between C^* and P^* along the line C^*E, they are unable to improve on C^*. Hence there is no need to leave one of the *B*s in isolation and the two *B*s will trade with the two *A*s at point C^*.

Figure 2

The new limit to the contract curve

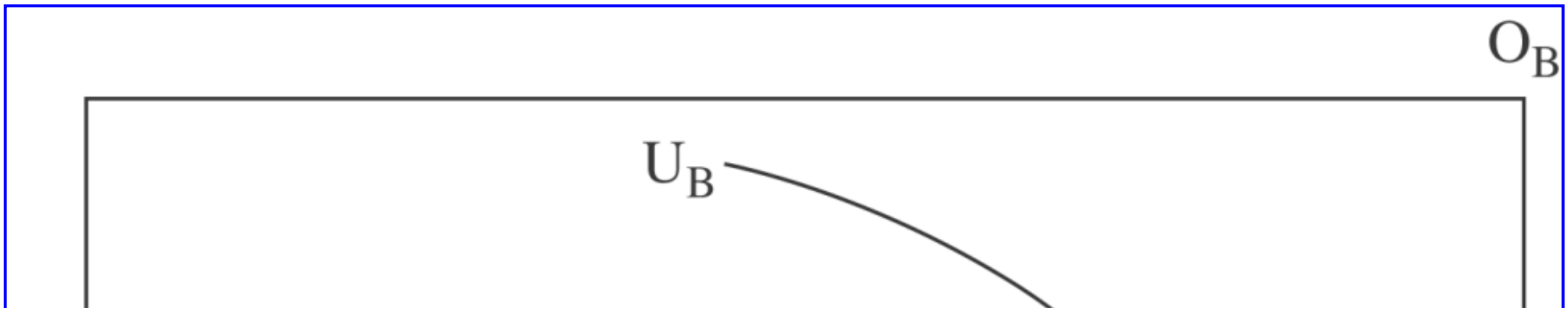

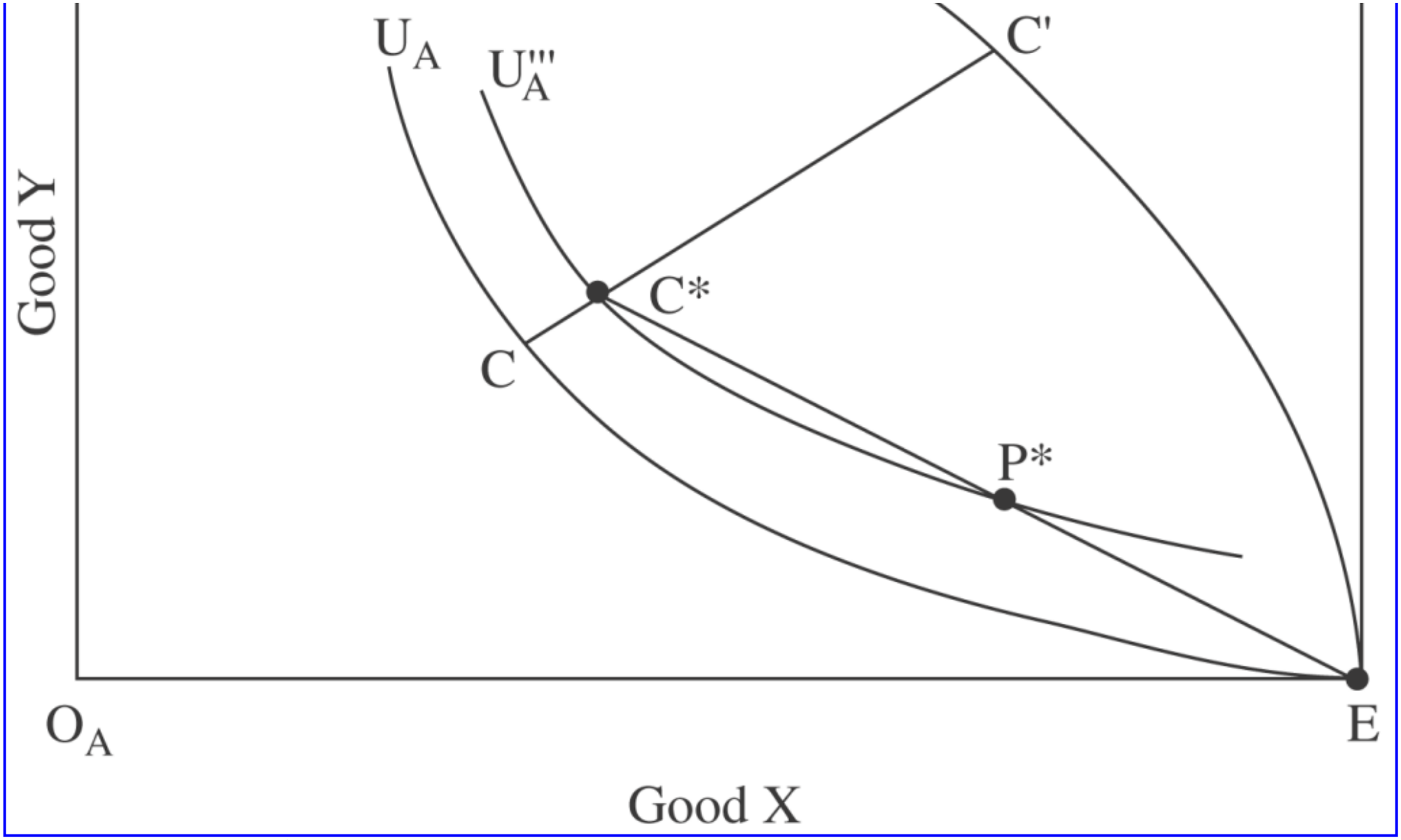

This argument has shown that at the final settlement all traders are at a common point on the contract curve and the limit has moved inwards along the old contract curve. The analysis can be repeated by starting with an alternative situation whereby the *A*s are initially assumed to be able to appropriate all the gains from trade. The point *C* ′ would then no longer qualify as a point on the new contract curve. The introduction of the additional pair of traders means that the contract curve shrinks, and the range of indeterminacy involved in barter is correspondingly reduced.

The extent to which the contract curve shrinks when the additional pair of traders is introduced is influenced by the fact that the *A*s cannot get further than halfway along a ray from a point on the contract curve to the endowment position. However, if there are three pairs of *A*s and three pairs of *B*s, the repetition of the above analysis involves two of the *A*s dealing with two of the *B*s at a point on the contract curve. The two *A*s then share their resources equally with the remaining *A* while the third *B* is isolated. The *A*s are able to consume together at a point which is two-thirds of the way along the ray from the initial endowment position to the point on the contract curve where the trade involving the two *A*s and two *B*s takes place.

With N pairs, the As can reach a proportion $\frac{(N-1)}{N}$ of the way from the endowment point to the contract curve. Thus as N increases, the values of k approaches unity. This means that the As can reach all the way from E to the contract curve, so that the final settlement must be such that the indifference curve is tangential to the ray from the origin. A final settlement with many traders is therefore shown in Figure 3 as point P on the contract curve. The effect of working in from the point C' would lead to an equivalent result for an indifference curve of the Bs, shown as U_B^*.

Figure 3

Final settlement with many traders

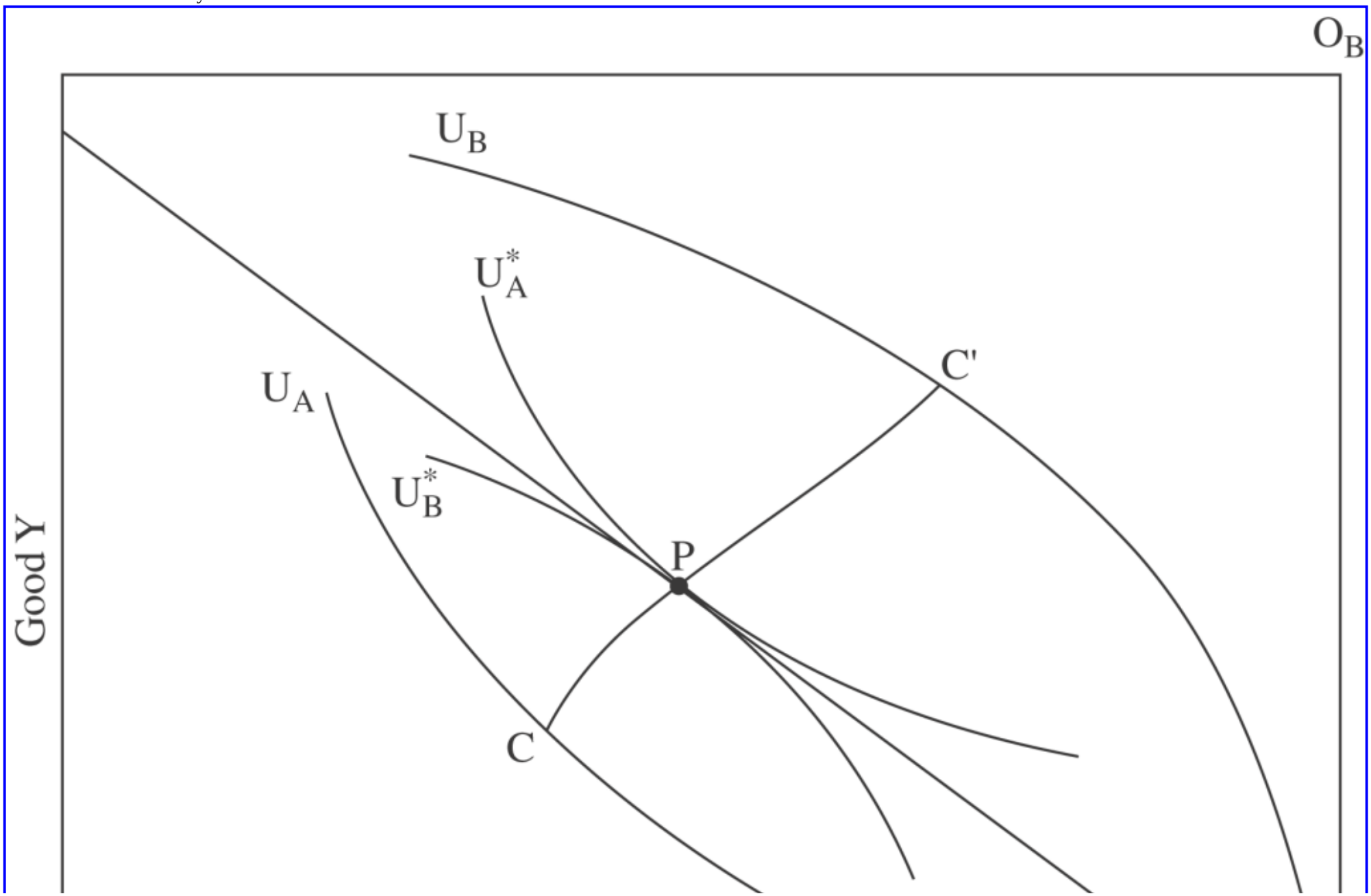

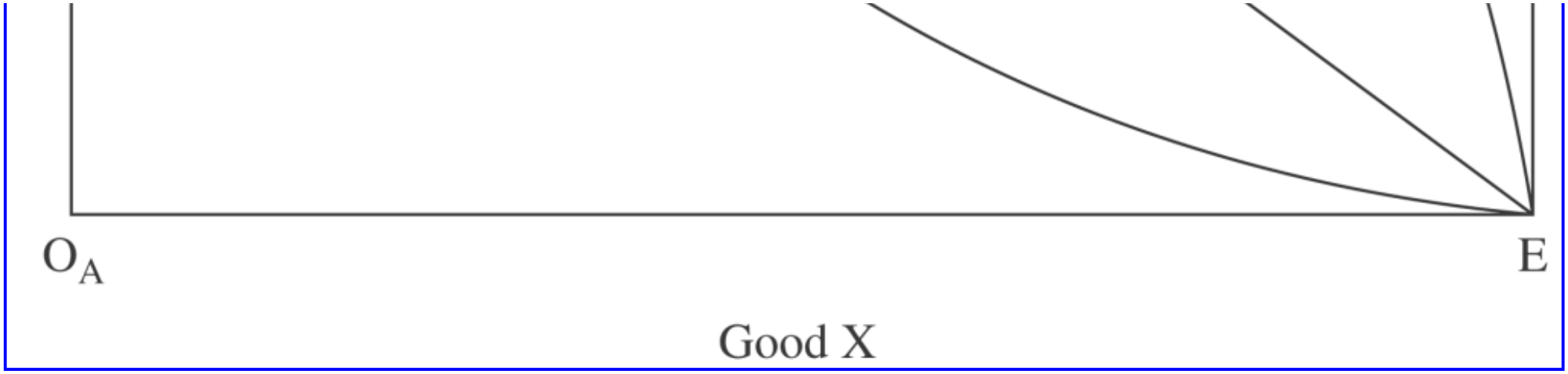

The result is that the final settlement looks just like a price-taking equilibrium. The figure illustrates the case where there is a single price-taking equilibrium. If there are multiple equilibria, the recontracting process causes the number of final settlements, with sufficiently large *N*, to shrink to the number of price-taking equilibria. (For discussion of utility functions involving multiple equilibria, and comparison of bargaining, competitive and utilitarian solutions, see Creedy, 1994a.) This argument relating to the shrinking contract curve, first established by Edgeworth, is often referred to as the *limit theorem*.

After Edgeworth's terse discussion, he stated:

> If this reasoning does not seem satisfactory, it would be possible to give a more formal proof; bringing out the important result that the common tangent to both indifference curves … is the vector from the origin. (1881, p. 38)

The price-taking solution is necessarily on the contract curve. This gives rise to what is now referred to as the ‘first fundamental theorem’ of welfare economics – that a price-taking equilibrium is Pareto efficient. Furthermore, the use of price-taking provides a considerable reduction in the amount of information required by traders when compared with the recontracting process. Given an equilibrium set, individuals need to know only the prices of goods, whereas in the recontracting process they have to learn a considerable amount of information about other individuals’ preferences and endowments. But Edgeworth placed more stress on the equivalence of the competitive price-taking solution with a recontracting barter process involving large numbers.

Given that coalitions among traders are allowed in the recontracting process, a price-taking equilibrium cannot be blocked by a coalition of traders. In this sense the competitive equilibrium is robust. The argument that a complex process of bargaining among a large number of individuals produces a result which replicates a price-taking equilibrium, allowing for the free flow of information using recontracting and enabling coalitions of traders to form and break up, is an important result that is far from intuitively obvious. The recontracting process can be said to represent a competitive process, and the contract curve shrinks essentially because of the competition between suppliers of the same good, although it is carried out in a barter framework in which explicit prices are not used (although rates of exchange are equivalent to price ratios).

The price-taking equilibrium, in contrast, does not actually involve a competitive process. Individuals simply believe that they must take market prices as given and outside their control. They respond to those prices without any reference to other individuals. But the result is that the price-taking equilibrium looks just like a situation in which all activity is perfectly coordinated.

Edgeworth suggested that similar results apply when some of the assumptions are relaxed. Thus, ‘when we suppose plurality of natures as well as persons, we have to suppose a plurality of contract-curves … Then, by considerations analogous to those already employed, it may appear that the quantity of final settlements is diminished as the number of competitors is increased’ (1881, p. 40). He then briefly considered different numbers of *A*s and *B*s, concluding that ‘the theorem admits of being extended to the general case of unequal numbers and natures’ (1881, p. 43). However, some of the results do not hold in the general case; for example, equality within the group of *A*s no longer holds when there are unequal numbers of *A*s and *B*s. A considerable number of articles have been written, since the late 1950s, examining various aspects of the Edgeworth recontract model under different assumptions.

Reciprocal demand curves

It has been mentioned that Edgeworth included in his diagram (1881, p. 28) the reciprocal demand curve, or offer curve, of each individual, although such curves were then called

‘demand-and-supply curves’. Edgeworth mentioned them only briefly in the text (1881, p. 39), but the lack of emphasis is understandable since in imperfect competition they are not relevant. Edgeworth's contribution was to provide the basic ‘analytics’ of the offer curve in terms of indifference curves, whereby it is ‘the locus of the point where lines from the origin touch curves of indifference’ (1881, p. 113).

When there is a lack of competition, giving rise to indeterminacy, there is nothing to ensure that individuals will trade on their offer curves and, as Edgeworth argued, ‘the conceptions of demand and supply at a price are no longer appropriate’ (1881, p. 31). It is this general preference, in favour of the analysis of barter in non-competitive situations, to which Marshall objected and which led to the controversy discussed below.

The utilitarian calculus

Having shown how indeterminacy can be removed by increasing the number of traders, Edgeworth turned to consider the role of arbitration in resolving the conflict between traders, in a ‘world weary of strife’ (1881, p. 51). The principle of arbitration examined was, not surprisingly, the utilitarian principle, which Edgeworth had earlier used to examine the optimal distribution. However, the new context of indeterminacy led him to a deeper justification of utilitarianism as a principle of distributive justice. Having arrived at this new link between ‘impure’ (egoistic) and ‘pure’ utilitarianism, Edgeworth had only to reorientate his earlier analysis of optimal distribution, contained in his paper in *Mind* of 1879.

The need for arbitration with indeterminacy had been stated by Jevons as follows:

> The dispositions and force of character of the parties … will influence the decision. These are motives more or less extraneous to a theory of economics, and yet they appear necessary considerations in this problem. It may be that indeterminate bargains of this kind are best arranged by an arbitrator or third party. (1871, pp. 124–5)

Edgeworth's statement of the same point was as usual rather less prosaic: ‘The whole creation groans and yearns, desiderating a principle of arbitration, and end of strifes’ (1881, p. 51). Edgeworth argument involved two steps. First, he showed that the principle of utility maximization places individuals on the contract curve, because the first-order conditions are equivalent to the tangency of indifference curves.

> It is a circumstance of momentous interest that one of the in general indefinitely numerous settlements between contractors is the utilitarian arrangement … the contract tending to the greatest possible total utility of the contractors. (1881, p. 53)

Edgeworth recognized that this result was not sufficient to justify the use of utilitarianism as a principle of arbitration. It is only a necessary condition of a principle of arbitration that it should place the parties somewhere on the contract curve. Edgeworth's justification for utilitarianism as a principle of justice, comparing points along the contract curve, was as follows:

> Now these positions lie in a reverse order of desirability for each party; and it may seem to each that as he cannot have his own way, in the absence of any definite principle of selection, he has about as good a chance of one of the arrangements as another … both parties may agree to commute their chance of any of the arrangements for … the utilitarian arrangement. (1881, p. 55)

The important point to stress about this statement is that Edgeworth clearly viewed distributive justice in terms of choice under uncertainty. He argued that the contractors, faced with uncertainty about their prospects, would choose to accept an arrangement along utilitarian lines. A crucial component of this argument, also clearly stated by Edgeworth in this quotation, is the use of equal a priori probabilities.

The importance to him of this new justification of utilitarianism cannot be exaggerated. Indeed the whole of *Mathematical Psychics* seems to be imbued with a feeling of excitement generated by his discovery of a justification based on a ‘social contract’. This provided the crucial link between ‘impure’ and ‘pure’ utilitarianism in a more satisfactory way than his earlier appeal to evolutionary forces.

Edgeworth believed that he had provided an answer to an age-old question, stating ‘by what mechanism the force of self-love can be applied so as to support the structure of utilitarian politics, neither Helvetius, nor Bentham, nor any deductive egoist has made clear’ (1881, p. 128). Nevertheless this argument was neglected until restatements along similar lines were made by Harsanyi (1953; 1955) and Vickrey (1960). The maximization of expected utility, with each individual taking the a priori view that any outcome is equally likely, was shown to lead to the use of a social welfare function which maximizes the sum of individual utilities. This approach is now usually described as ‘contractarian neo-utilitarianism’.

In discussing the utilitarian solution as a principle of arbitration in indeterminate contract, Edgeworth did not clearly indicate in 1881 that the utilitarian solution of maximum total utility could specify a position which makes one of the parties worse off than in the no-trade situation. This was nevertheless later made explicit when, after proposing arbitration along utilitarian lines, he added ‘subject to the condition that neither should lose by the contract’ (1925, ii, p. 102). This possibility of course depends largely on the initial

endowments of the individuals.

Later work in economics

After the publication of *Mathematical Psychics*, Edgeworth concentrated increasingly on mathematical statistics, in particular on the problem of statistical inference, but, following his appointment to the Drummond Chair at Oxford, Edgeworth again made important contributions to economics, although this work mainly involved reactions to, and discussions arising from, the later work of other authors.

Demand and exchange

In the *Principles of Economics* (1890, Appendix F) Marshall included a brief discussion of Edgeworth's analysis of barter, and produced a figure showing the contract curve. During the following year, in the course of a review written in Italian (translated in Edgeworth, 1925, ii, pp. 315–19), Edgeworth criticized Marshall for not having dealt sufficiently with the problem of indeterminacy. The basic problem was that Marshall, using a model in which a series of trades are allowed to take place at disequilibrium prices, believed he had shown that prices will eventually settle at the price-taking equilibrium. However, the argument was not transparent. The adjustment process involves moving from the initial endowment point in a series of trades, where trading at ‘false’ prices is allowed at each step. The process must conclude with both individuals at a point on the contract curve. A feature of the process is the assumption that each stage or iteration of the sequence involves Pareto improvements: individuals trade only if it makes them better off. Furthermore, it involves trading at the ‘short end’ of the market, that is, the minimum of supply and demand. This arises from the impossibility of forcing any individual either to buy or sell more than desired at any price.

An example of two disequilibrium trades is shown in Figure 4, where the endowment moves from E to E_1, and then to E_2. With a price line represented by EP, there is an excess supply of good X as person A tries to reach the indifference curve U'_A and person B wishes to reach U'_B. Trade takes place at E_1, the short end of the market. Point E_1 then becomes the new endowment point. At the second trading stage, the price of X must be lowered to induce person B to purchase more. At a price represented by the line E_1P_1 through the new endowment point, the excess supply is lower than formerly and trade takes place at E_2. Comparing U'_A and U'_B with U''_A and U''_B respectively, it can be seen that E_2 is a Pareto improvement relative to E_1. It is also clear that person A is better off the slower the fall is in the price of X relative to Y at each stage.

Figure 4
Disequilibrium trades

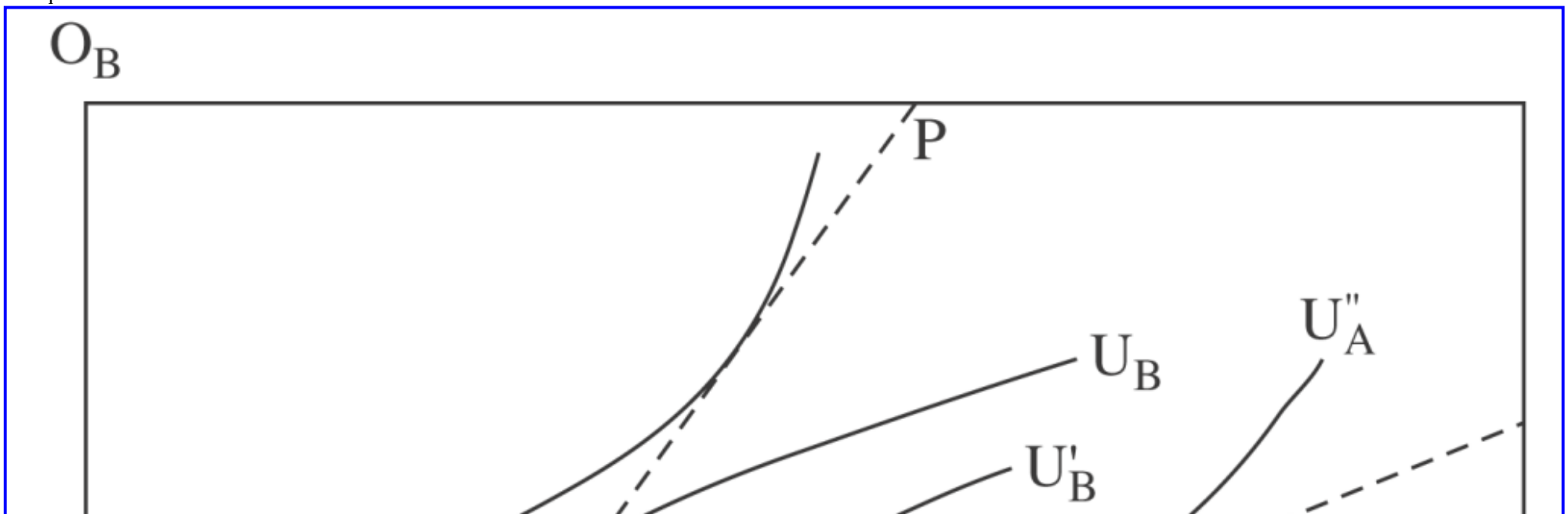

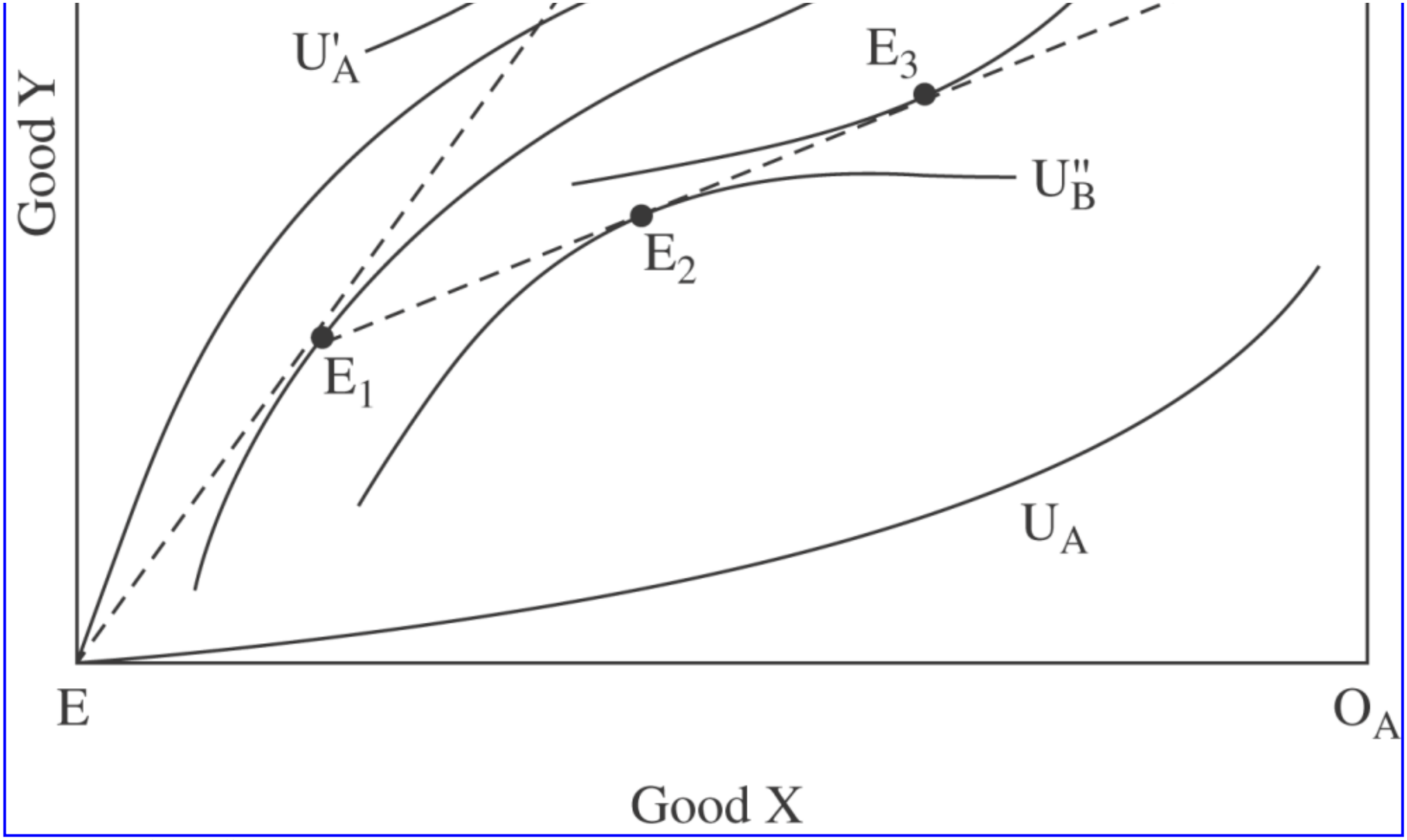

The combination of Pareto-efficient moves at each stage and an adjustment process such that an excess supply leads to a price reduction, and vice versa, produces a stable process that converges to an equilibrium somewhere on the contract curve. (This type of sequence of disequilibrium trades was later used by Launhardt; see Creedy, 1994b.)

The basic problem was that Marshall believed that his assumption of an additive utility function, combined with the assumption that the marginal utility of one good is constant for both individuals, guaranteed a determinate price, if the good having constant marginal utility was money. Indeed, this case was mentioned by Edgeworth (see 1925, ii, p. 317 n.1). The contract curve is a straight line parallel to the *y* axis (where this good is the one with constant marginal utility), along which the rate of exchange is constant. So the equilibrium price does not depend on the sequence of trades. However, Edgeworth's point was that the total amount spent on good *x* remains indeterminate.

There was a later, though much milder, disagreement between Marshall and Edgeworth over the so-called Giffen good. In a book review, Edgeworth argued that, ‘even the milder statement that the elasticity of demand for wheat may be positive, though I know it is countenanced by high authority, appears to me so contrary to a priori probability as to require very strong evidence (1909, p. 104). The ‘authority’ was of course Marshall (1890, p. 132), who replied directly to Edgeworth that, ‘I don't want to argue … But … the matter has not been taken quite at random’ (Pigou, 1925, p. 438). Marshall gave a numerical example involving a journey travelled by two methods, where the distance travelled by the cheaper and

slower method must increase when its price increases. For further details, see Creedy (1990).

It has been mentioned that Edgeworth introduced the generalized utility function. An implication is that it allows for complementarity, although Edgeworth did not explicitly consider this in 1881. The first formal definition of complementarity is attributed to Auspitz and Lieben, and it was used by Edgeworth in his paper on the pure theory of monopoly, and also by Pareto: this amounts to what is now called ‘gross’ complementarity, defined in terms of cross-price elasticities. It is also sometimes referred to, using the initials of the four people mentioned above, as ALEP complementarity.

The first major criticism came from Johnson (1913), who pointed out that the criterion was not invariant with respect to monotonic transformations of the utility function. His treatment was extended by Hicks and Allen (1934), so that the modern definition involves ‘net’ complements in terms of compensated price changes. There is no symmetry between gross substitutes and complements as only the matrix of (compensated) substitution elasticities is assumed to be symmetric.

Monopoly and oligopoly

In a paper first published in Italian in 1897, and not translated until the collected *Papers* (1925), Edgeworth examined several problems relating to monopoly. He began his discussion with Cournot's (1838) example of the ‘source minérale’ in which there are ‘two monopolists’ (that is, duopolists), each owning a spring of mineral water. It would be natural for Edgeworth to expect an indeterminate price in this ‘small numbers’ context. Cournot had arrived at a determinate solution for price and output, but Edgeworth showed that ‘when two or more monopolists are dealing with competitive groups, economic equilibrium is indeterminate’ (1925, p. 116). The daily output from each spring was assumed to be limited to identical fixed amounts, delivery costs were zero and all consumers had the same demand curve (purchasing one unit only of output). Hence demand is $n(1-p)$ where n is the number of customers and p is the price. Cournot's solution was that the price would be $p = 1/4$, but Edgeworth argued that one of the ‘monopolists’ had an incentive to raise the price back to $p = 1/2$, which is the revenue maximizing price, so that there is not a determinate price. He argued that:

> at every stage … it is competent to each monopolist to deliberate whether it will pay him better to lower his price against his rival as already described, or rather to raise it to a higher … for that remainder of customers of which he cannot be deprived by his rival. … Long before the lowest point has been reached, that alternative will have become more advantageous than the course first described’ (1925, p. 120)

Edgeworth went on to say ‘the matter may be put in a clearer light’, and he then defined what are now called the reaction curve and isoprofit lines (in that order) for variations in prices. However, it was not until Bowley's (1924) discussion that these matters began to be presented in a more transparent manner.

Edgeworth then considered the case of complementary demand within the context of ‘bilateral monopoly’, where the two goods are demanded in fixed proportions for use in the production of a further article. An interesting feature is that he wrote the equations of the reaction curves and explicitly dealt with what are now called conjectural variations, reflecting the extent to which one duopolist is expected to change price in response to changes made by the second duopolist. In discussing this problem Edgeworth also introduced the further important concept of the ‘saddle point’, which he called the ‘hog's back’, clearly indicating its importance for stability.

The no-profit entrepreneur

Walras (1874, p. 225) had introduced the concept of the entrepreneur who neither gains nor loses. This result applied only to the competitive equilibrium, where there are no incentives for entrepreneurs to enter any industry. This does not of course mean that there are no profits, in the accounting sense, since the returns to homogeneous units of inputs of organization and management services are subsumed in the costs of the firm.

Edgeworth's criticisms of this concept of the no-profit entrepreneur, reproduced in his *Papers* (1925), recognized that with Walras's assumptions there was nothing illogical about the argument. The theory simply means that nothing remains ‘after the entrepreneur has paid a normal salary to himself•’ (1925, pp. 26, 30). Furthermore, ‘if [the general expenses] are taken into account, the argument becomes a fortiori. For why should not a substantial remuneration for the entrepreneur be included in the general expenses of the business’ (1925, ii, p. 469). Edgeworth's difference with Walras was to some extent ‘only verbal’, but he was also unhappy with the idea that entrepreneurship is homogeneous and divisible.

The theory of taxation

In the 1890s Edgeworth produced two surveys of considerable importance. These surveys, of the pure theory of taxation and of the pure theory of international values, were both published in the *Economic Journal* and subsequently reproduced (with alterations) in his *Papers* (1925, vol. ii). Each survey consisted of three separate parts, and displayed a staggering breadth of knowledge and command of the subject. They represent his most serious attempts to produce any kind of synthesis of a branch of economic literature.

Edgeworth began his survey with the rather strong statement that ‘the science of taxation comprises two subjects to which the character of pure theory may be ascribed; the laws of incidence, and the principle of equal sacrifice’ (1925, p.64). He then considered a variety of special cases and contexts of tax incidence. The basic framework for incidence analysis was the simple partial equilibrium approach, still used in many basic textbooks, in which the incidence depends on the relative values of supply and demand elasticities.

The basic approach to incidence analysis actually stemmed from the important paper by Jenkin (1871). It suggests that in general the price of the taxed good will either remain constant (in the extreme case of inelastic supply) or will increase. However, this result ignores interrelationships among commodities. Edgeworth showed that, when such interrelationships are explicitly allowed, there are some circumstances in which the price of the taxed good will actually fall. When discussing this ‘paradox’, Edgeworth reproduced his argument which had in fact been explored in more detail in his paper on monopoly, published in Italian in the same year (translated in Edgeworth, 1925, i, pp. 111–42). Edgeworth first stated his ‘tax paradox’ in the following terms:

> when the supply of two or more correlated commodities – such as the carriage of passengers by rail first class or third class – is in the hands of a single monopolist, a tax on one of the articles – e.g. a percentage of first class fares – may prove advantageous to the consumers as a whole. … The fares for all the classes might be reduced. (1925, p. 139)

Edgeworth regarded this result as an example of a situation where, ‘the abstract reasoning serves as a corrective to what has been called the “metaphysical incumbus” of dogmatic laisser faire’ (1925, i, p. 139; see also 1925, ii, pp. 93–4). Essentially the two commodities must be substitutes in consumption and production, and the result is partly brought about by the fact that the monopolist has an incentive to increase the supply of the untaxed commodity. Edgeworth also recognized that the result could occur in competitive markets (see 1925, p. 63). As with many of Edgeworth's original results, this tax paradox was not a subject of continuous development. Its main practical importance perhaps arises from the fact that in the early 1930s it attracted the attention of Hotelling (1932). For further discussion of the paradox, see Creedy (1988).

The section of the taxation survey which attracted most immediate attention was Edgeworth's discussion of the various ‘sacrifice’ theories of the distribution of the tax burden, and his qualified support for progressive taxation. Edgeworth's attitude to taxation was similar to that of the major classical economists in that he rejected a benefit approach, on the argument that taxation is not an economic bargain governed by competition. Thus in his view the problem was to determine ‘the distribution of those taxes which are applied to common purposes, the benefits whereof cannot be allocated to particular classes of citizens’ (1925, p. 103). A principle of justice is thus required. His approach can be seen as marking a crucial stage in the transition towards a ‘welfare economics’ view of public finance, rather than using a special set of ‘tax maxims’ such as the famous criteria laid down by Adam Smith.

Not surprisingly, Edgeworth (1925, p. 102) argued along neo-contractarian lines set down in *Mathematical Psychics* that the utilitarian arrangement would be accepted by individuals uncertain of their own prospects and taking an equal a priori view of the probabilities. He suggested that

> each party may reflect that, in the long run of various cases … of all the principles of distribution which would afford him now a greater, now a smaller proportion of the sum-total utility obtainable … the principle that the collective utility should be on each occasion a maximum is most likely to afford the greatest utility in the long run to him individually

Having established the use of utilitarianism as a principle of distributive justice, Edgeworth then succinctly stated the main argument:

> The condition that the total net utility procured by taxation should be a maximum then reduces to the condition that the total disutility should be a minimum … it follows in general that the marginal disutility incurred by each taxpayer should be the same. (1925, p. 103)

The implication is that, if all individuals have the same cardinal utility function, after-tax incomes would be equalized. Edgeworth also clearly recognized that, if there is considerable dispersion of pre-tax incomes relative to the total amount of tax to be raised, where there is ‘not enough tax to go around’ (1925, ii, p. 103), the equi-marginal condition cannot be fully satisfied unless there is a ‘negative income tax’ which raises the incomes of the poorest individuals to a common level. Thus, ‘the acme of socialism is for a moment sighted’ (1925, p. 104). But Edgeworth immediately considered the practical limitations to such high progressive taxation. The following quotation illustrates one of Edgeworth's favourite metaphors, his respect for Sidgwick, his attitude to authority, his views on utilitarianism and the applicability of pure theory, and of course his unmistakable style:

> In this misty and precipitous region let us take Professor Sidgwick as our chief guide. He best has contemplated the crowning height of the utilitarian first principle, from which the steps of a sublime deduction lead to the high tableland of equality; but he also discerns the enormous interposing chasms which deter practical wisdom from moving directly towards that ideal. (1925, p. 104)

Among the various limitations, Edgeworth noted differences in individual utility functions, population effects, the disincentives to work, growth of culture and knowledge, savings, and of course the problem of evasion.

International trade

Edgeworth's survey of the pure theory of international values was in some ways responsible for a change of emphasis in the approach to trade theory, despite the fact that it contained few original analytical contributions. Indeed, he said that, ‘Mill's exposition of the general theory is still unsurpassed’ (1925, p. 20), and acknowledged further that, ‘what is written … after a perusal of [Marshall's] privately circulated chapters … can make no claim to originality’ (1925, p. 46). Edgeworth saw trade theory as an application of the general theory of exchange:

> The fundamental principle of international trade is that general theory … the Theory of Exchange … which … constitutes the ‘kernel’ of most of the chief problems in economics. It is a corollary of the general theory that all the parties to a bargain look to gain by it … This is the generalised statement of the theory of comparative cost. (1925, p. 6)

Thus the gains from trade are analogous to the gains from exchange in simple barter and ‘It is useful … to contemplate the theory of distribution as analogous to that of international trade proper’ (1925, p. 19). Hence trade theory is to Edgeworth simply one more application of the general method of *Mathematical Psychics*. In directly applying the theory of exchange to that of trade, Edgeworth was quite content to use community indifference curves without clearly specifying how aggregation might be carried out. He said only that ‘by combining properly the utility curves for all the individuals, we obtain what may be called a collective utility curve’ (1925, p. 293).

One of Edgeworth's criticisms of Mill (1848) was that the latter took as his measure of the gain from trade the change in the ratio of exchange of exports against imports. Thus Mill in this case ‘confounds “final” with integral utility’ (1925, p. 22). The same point had in fact been made by Jevons (1871, pp. 154–6). However, Edgeworth, while preferring total utility, admitted that Mill was not otherwise led to serious error in using his own measure.

Edgeworth's survey was, as always, extremely wide-ranging, though for later developments the most interesting parts are concerned with his elucidation of Mill's ‘recognition of the case in which an impediment may be beneficial – or an improvement prejudicial – to one of the countries’ (1925, p. 9). These cases would now be discussed under the headings of the ‘optimal tariff’ and ‘immiserizing growth’. In the case of an optimal tariff, a country acts as monopolist and imposes a price which enables that country to attain its highest indifference curve, subject to the other country's offer curve. However, this position is not on the contract curve. The detailed specification of the optimum tariff in terms of elasticities had to wait until Bickerdike (1906), Pigou (1908) and the later revivals of interest in the 1940s. Edgeworth's judgement of Bickerdike was that he had ‘accomplished a wonderful feat. He has said something new about protection’ (1925, ii, p. 344).

Edgeworth could not of course be expected to support the use of such tariffs in practice. He acknowledged the possibility of retaliation, but also:

> For one nation to benefit itself at the expense of … others is contrary to the highest morality … But in an abstract study upon the motion of projectiles in vacuo, I do not think it necessary to enlarge upon the horrors of war. (1925, p. 17 n. 5)

The ‘highest morality’ was, of course, the principal of utilitarianism.

Conclusions

It has been seen that Edgeworth did not begin working and writing in economics until his mid-30s, but in common with the majority of neoclassical economists he soon pursued an academic career as a professor of economics. Indeed, in a period which saw the rapid and widespread professionalization of the subject Edgeworth held an academic position in England that was regarded as second only to that of Alfred Marshall. In spite of his wide range of reading and sympathies, Edgeworth's work was characterized by the fact that it was virtually all addressed to his fellow professional economists. So uncompromising was he in his view that economics is a very difficult subject offering only remote and nearly always negative policy advice that it may fairly be said that his work was addressed to just a small number of ‘fellow travellers’ in the rarefied atmosphere of the ‘higher regions’ of pure theory. However, Edgeworth imposed no geographical limitations, and with his considerable linguistic skills and international sympathies was in contact with the majority of leading economists around the world.

The distinguishing feature of the neoclassical ‘revolution’ was its emphasis on exchange as the central economic problem. The success of this shift of focus from production and distribution to exchange was closely associated with the fact that it had as its foundation a model based on utility maximization. This allowed for a deeper treatment of the gains from

exchange and the wider considerations of economic welfare. Schumpeter summarized the point by stating that utility analysis must be understood in terms of exchange as the central 'pivot' and 'the whole of the organism of pure economics thus finds itself unified in the light of a single principle' (1954, p. 913). This is indeed the context in which Edgeworth's work in economics must be seen. Schumpeter's remark is merely a more prosaic expression of Edgeworth's view quoted above that '"Méchanique Sociale" may one day take her place along with "Méchanique Celeste" [sic], throned each upon the double-sided height of one maximum principle'. The central theme of Edgeworth's work is also clear in his revealing statement, taken from his presidential address to Section F of the Royal Society, that:

> It may be said that in pure economics there is only one fundamental theorem, but that is a very difficult one: the theory of bargain in a wide sense. (1925, ii, p. 288)

This perspective helps the major thread which runs through all Edgeworth's work in economics to be seen. His earlier mathematical analysis of the implications of utilitarianism for the optimal distribution, written before he turned to economics, was not only highly original (and esoteric) but laid the foundation for his work in economics. Thus, the transition from *New and Old Methods of Ethics* to *Mathematical Psychics* was not a shift in major preoccupations but rather a change of emphasis. Distribution was then seen as an important concomitant of exchange, so that the analysis of contract became central for Edgeworth. Edgeworth's emphasis on the indeterminacy (the inability of utility maximization alone to determine the rate of exchange, only a range of efficient exchanges) which results from the existence of a small number of traders led him to his path-breaking analysis of the role of numbers in competition, along with the efficiency properties of competitive equilibria.

The analysis of the utilitarian objective as an arbitration rule led Edgeworth directly to his new 'social contract' argument in explaining the acceptance of utilitarianism as a principle of social justice. It was the realization of this new justification of utilitarianism, using his newly developed analytical tools, which generated the excitement that is clearly evident in his first work in economics. While *Mathematical Psychics* developed the techniques of indifference curves and the contract curve within the 'Edgeworth box' – tools which are now ubiquitous in economic analysis – Edgeworth himself was clearly driven mainly by his ability to link the analysis of private contracts in markets to that of a social contract in which utilitarianism is the 'sovereign principle'. The integration of his analysis of barter, and the effects of the introduction of additional traders into the market, with the demonstration that the utilitarian arrangement prescribes a point on the contract curve of efficient exchanges and is acceptable to risk-averse traders, was to Edgeworth nothing short of 'momentous'. The results are of course highly abstract. In discussing their ultimate value suggested that:

> Considerations so abstract it would of course be ridiculous to fling upon the flood-tide of practical politics … it is at a height of abstraction in the rarefied atmosphere of speculation that the secret springs of action take their rise, and a direction is imparted to the pure foundation of youthful enthusiasm whose influence will ultimately affect the broad current of events. (1881, p. 128)

The intellectual pleasure derived from being able to draw together so many different subjects of analysis, and strands of his enormous range of learning, is clearly evident. However, it is precisely this wide field of vision, combined with the technical level and idiosyncratic style of writing, which made *Mathematical Psychics* so difficult for his contemporaries, and which continue to make the book seem so strange and yet so rewarding to the modern reader.

I am grateful to Denis O'Brien and Steven Durlauf for comments on an earlier draft of this article.

Selected works

1876. Mr. Matthew Arnold on Bishop Butler's doctrine of self love. *Mind* 1, 570–1.

1877. *New and Old Methods of Ethics: or 'Physical Ethics' and 'Methods of Ethics'*. Oxford: Parker.

1879. The hedonical calculus. *Mind* 4, 394–408.

1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: Kegan Paul.

1887. *Metretike, or the Method of Measuring Probability and Utility*. London: Temple.

1909. Review of Free Trade in Being. *Economic Journal* 19, 104–5.

1925. *Papers Relating to Political Economy*, 3 vols. London: Macmillan, for the Royal Economic Society.

Bibliography

Bickerdike, C. 1906. The theory of incipient taxes. *Economic Journal* 16, 529–35.

Black, R., ed. 1977. *Papers and Correspondence of William Stanley Jevons*. London: Macmillan, for the Royal Economic Society.

Black, R., ed. 1981. *Papers and Correspondence of William Stanley Jevons. Vol. VII, Papers on Political Economy*. London: Macmillan, for the Royal Economic Society.

Bowley, A. 1924. *The Mathematical Groundwork of Economics*. Oxford: Clarendon Press.

Bowley, A. 1928. *Edgeworth's Contribution to Mathematical Statistics*. London: Royal Statistical Society.

Bowley, A. 1934. Francis Ysidro Edgeworth. *Econometrica* 1, 113–24.

Butler, J. and Butler, H. 1927. *The Black Book of Edgeworthtown and other Edgeworth Memories 1585–1817*. London: Faber and Gwyer.

Cournot, A. 1838. *Researches into the Mathematical Principles of the Theory of Wealth*, trans. N. Bacon, ed. I. Fisher. London: Stechert-Hafner, 1927.

Creedy, J. 1986. *Edgeworth and the Development of Neoclassical Economics*. Oxford: Basil Blackwell.

Creedy, J. 1988. Wicksell on Edgeworth's tax paradox. *Scandinavian Journal of Economics* 90, 101–12.

Creedy, J. 1990. Marshall and Edgeworth. *Scottish Journal of Political Economy* 37, 18–39.

Creedy, J. 1994a. Exchange equilibria: bargaining, utilitarian and competitive solutions. *Australian Economic Papers* 33, 34–52.

Creedy, J. 1994b. Launhardt's model of exchange. *Journal of the History of Economic Thought* 16, 40–60.

Edgeworth, M. 1820. *Memories of Richard Lovell Edgeworth Esq., Begun by Himself and Concluded by his Daughter, Maria Edgeworth*, 2 vols. London: Hunter.

Harsanyi, J. 1953. Cardinal utility in welfare economics and in the theory of risk taking. *Journal of Political Economy* 61, 434–5.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Hicks, J. and Allen, R. 1934. A reconsideration of the theory of value. *Economica* 14, 52–76, 196–219.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40, 577–616.

Jenkin, F. 1871. On the principles which regulate the incidence of taxes. Reproduced in *Readings in the Economics of Taxation*, ed. R. Musgrave and C. Shoup. London: Allen and Unwin, 1959.

Jevons, W. 1881. Review of *Mathematical Psychics*. *Mind* 6, 581–3.

Jevons, W. 1871. *The Theory of Political Economy*, 5th edn, ed. H. Jevons. New York: Augustus Kelly, 1957.

Johnson, W. 1913. The pure theory of utility curves. *Economic Journal* 23, 483–513.

Keynes, J.M. 1972. *Essays in Biography. Vol. X of the Collected Writings of Keynes*. London: Macmillan, for the Royal Economic Society.

McCann, C., ed. 1996. *F.Y. Edgeworth: Writings in Probability, Statistics and Economics*, 3 vols. Cheltenham: Edward Elgar.

Marshall, A. 1890. *Principles of Economics*, 2 vols, ed. C. Guillebaud (variorum edn). London: Macmillan. 1961.

Marshall, A. and Marshall, M. 1879. *Economics of Industry*. London: Macmillan.

Mill, J.S. 1848. *Principles of Political Economy*. Reprinted with editorial material by W. Ashley. London: Longmans, Green, 1920.

Pigou, A. 1908. *Protective and Preferential Import Duties*. London: Macmillan.

Pigou, A., ed. 1925. *Memorials of Alfred Marshall 1842–1924*. London: Macmillan.

Pigou, A. and Robertson, D. 1931. *Economic Essays and Addresses*. London: King.

Price, L. 1946. Memoirs and notes on British economists 1881–1946. MSS, Brotherton Library, University of Leeds.

Sargan, J. 1976. Econometric estimators and the Edgeworth expansion. *Econometrica* 44, 421–48.

Schumpeter, J. 1954. *History of Economic Analysis*. London: Allen and Unwin.

Stigler, S. 1978. Francis Ysidro Edgeworth, statistician. *Journal of the Royal Statistical Society, A* 141, 287–322.

Sully, J. 1918. *My Life and Friends*. London: Fisher Unwin.

Vickrey, W. 1960. Utility, strategy and social decision rules. *Quarterly Journal of Economics* 74, 507–35.

Walras, L. 1874. *Elements of Pure Economics*, trans. W. Jaffe. London: Allen and Unwin, 1954.

Whitaker, J., ed. 1975. *The Early Economic Writings of Alfred Marshall 1867–1890*. London: Macmillan.

How to cite this article

Creedy, John. "Edgeworth, Francis Ysidro (1845–1926)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000041> doi:10.1057/9780230226203.0446

The New Palgrave Dictionary of Economics Online

education in developing countries

Paul Glewwe
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In many developing countries, children complete few years of schooling and learn little during their time in school. There are many estimation problems that confound attempts to understand the impact of education policies on years of schooling and learning while in school. Recent research has focused on implementing randomized trials to get around estimation problems based on retrospective data. While some useful results have been found, many additional studies are still under way. As these results accumulate it is likely that general conclusions can be drawn, but the evidence to date is too limited to draw general policy recommendations.

Keywords

attenuation bias; credit; education in developing countries; education production functions; returns to schooling; value of time

Article

Most economists who study economic growth agree that an educated citizenry is necessary for sustained economic growth, and virtually all international development organizations concur (UNDP, 1990; World Bank, 2001), and so those organizations provide substantial financial resources and policy advice to promote education in developing countries. Yet in many developing countries, especially the poorest, many children leave school at a young age and learn little during the time they spend in school. These problems have led many economists and other social scientists to turn their attention to education in developing countries.

This article summarizes recent research on the factors that affect the amount of time that children spend in school and the factors that determine how much they learn during their time in school. Thus, it focuses on the factors that shape education outcomes as opposed to the impact of education on income, economic growth and other phenomena (for a recent assessment of the impact of education on other

socio-economic outcomes, see Glewwe, 2002). This article also omits, due to space constraints, a discussion of estimation issues (see Glewwe, 2002, and Glewwe and Kremer, 2006, for thorough discussions of estimation problems and possible solutions).

Factors that determine years of schooling

In developing countries, parents usually decide how many years their children will attend school. Each year, parents consider the costs and expected benefits of an additional year of schooling and then enrol their children for another year if the expected benefits outweigh the estimated costs. The main costs are school fees and other payments required by schools, transportation and (occasionally) meals and housing, and the opportunity cost of the children's time. There may also be an additional, ‘psychic’ cost; some parents may dislike particular values that schools attempt to instil in students. For many parents, the largest of these costs is the value of their children's time; in developing countries, especially in rural areas, children's time is valuable because they can help in household farming activities.

The main benefits of schooling are the skills learned (which usually reap substantial monetary returns in the labour market), increased employment opportunities that come with educational credentials, and the direct satisfaction and social approval that parents receive from having educated children. While the decision rule to continue schooling when the benefits outweigh the costs would seem to hold as a tautology, there are circumstances in which children are not enrolled in school even when the economic benefits outweigh the costs. This could occur because the costs are incurred today while the benefits accrue over many years in the future. In particular, parents who have low incomes and cannot obtain credit may not send their children to school even though the present discounted value at prevailing interest rates is positive.

Given this type of decision making by parents, policies to increase school enrolment must focus on reducing the costs of schooling, increasing the benefits of education, or providing access to credit. Reductions in fees are easy to implement, and in some countries (such as Mexico) parents with low incomes receive monthly payments if their children are enrolled in school. Of course, this entails potentially large budgetary costs, so some governments try to limit fee exemptions and outright subsidies to households or communities that are particularly needy. Evidence from many developing countries indicates that reducing fees or providing payments conditional on school enrolment can lead to large increases in enrolment; studies in Honduras, Kenya, Mexico and Nicaragua document these impacts (see Glewwe and Kremer, 2006, for further details and references).

The main alternative policy for increasing school enrolment is to increase the expected returns. These returns will increase if the relative price of skilled labour increases, and if schools become more effective at providing academic skills. While some economists have shown that increased returns to education does raise school enrollment (Foster and Rosenzweig, 1996), most policy research has focused on what makes schools more efficient at raising students’ skills. This research is discussed in the next section.

Three additional points regarding policies to increase years of schooling deserve attention. First, improvements in the health and nutritional status of both very young and school-age children are another potentially important route to increase the time that children spend in school (see Glewwe and Miguel, 2006, for a review of this literature). Second, many policy discussions presume that the main reason

children are not in school is that no school is available, yet in most countries schools are available but parents opt not to enrol their children because they judge that the costs outweigh the benefits (see Glewwe and Zhao, 2005). Third, the role of credit constraints in determining years in school is an under-researched topic, in terms of both the impact of credit constraints and policies that could loosen those constraints.

Factors that determine student learning

In principle, student learning can be depicted as a production process in which student, household, teacher and school characteristics combine to produce students' academic skills. While the existence of an academic skills production function is true almost by definition, there are serious problems that confound attempts to estimate this process. The main problem is omitted variables bias: students, households, teachers and schools can vary in hundreds of ways, and no data-set contains all variables that are potentially important. Indeed, important factors such as student innate ability, teacher effort and parental encouragement are almost impossible to measure and likely to be correlated with the observed variables. This problem applies to virtually all studies based on retrospective (non-experimental) data; indeed, it is probably the main reason that different studies find very different results (the main alternative explanation is that educational production functions are very different in different countries). A second serious estimation problem is attenuation bias. Much of the data on students, households, teachers and schools has a substantial amount of measurement error. This typically leads to underestimation of the true impacts of variables, which may explain, at least in part, why many variables in estimates of the determinants of student learning are statistically insignificant.

In recent years economists and other social scientists have turned to natural experiments and randomized trials to estimate the impacts of particular school characteristics, policies and programmes on student academic achievement. Natural experiments result from institutions and policies that cause random variation in school or student characteristics, which can be used to analyse the impact of those characteristics on student learning (and on time spent in school). Randomized trials are controlled experiments designed by researchers and school officials that generate random variation in a school characteristic or policy, which again allows one to estimate the impact of the characteristic or policy on learning. Natural experiments are relatively rare, but in recent years randomized trials have been implemented in many countries in Africa, Asia and Latin America.

One of the first randomized trials was conducted in Nicaragua in the late 1970s. The results indicated that workbooks and radio instruction had significant impacts on pupils' math scores. In the Philippines in the early 1980s, provision of textbooks raised students' performance on academic tests, but in Kenya in the late 1990s the only effect of textbooks was among the better students, perhaps because the textbooks provided were too difficult for most students. Other randomized trials conducted in Kenya suggest little impact on test scores from reductions in class size, provision of flip charts, and provision of deworming medicine. On a more positive note, school meals in Kenya raised test scores in schools that had well-trained teachers, but not in schools with poorly trained teachers. In public schools in an urban area of India, a remedial education programme increased test scores at a relatively low cost. Finally, a computer-assisted learning programme in India also appears to have increased test scores. The positive impacts of radio education in Nicaragua and computer instruction in India suggest that using modern technologies may be particularly helpful in schools with weak teachers. (For citations and more detailed

discussion, see Glewwe and Kremer, 2006.)

While natural experiments and especially randomized trials may seem to avoid the estimation problems that plague retrospective studies, more randomized studies are needed before general conclusions can be drawn that can guide policy in countries that have not yet had such studies. Moreover, randomized trials can also suffer from estimation problems. One problem is that parents of students in the control schools (or schools excluded from the evaluation) may try to enrol their children in the treatment schools. This may affect the results by increasing class size (if class size affects learning). This would not occur if the policy were implemented nationwide. In addition, children who transfer into treatment schools may not be a random sample of the general student population. A related problem is that marginal students in the treatment schools are less likely to drop out (if the intervention raises student achievement), which leads to underestimation of the impact of the policy on learning if comparisons are made based on all students currently enrolled in school. A final problem with randomized trials is that the evaluation itself may lead the treatment group to change its behaviour, or the control group to change its behaviour, because both groups know that their results are being used in an evaluation.

In summary, recent research on education in developing countries has provided fairly convincing evidence of the impact on time in school and on learning for particular policies in particular countries. Many additional studies are currently under way, and as these results accumulate it is likely that general conclusions can be drawn. This should lead to better education policies, which will contribute to higher economic growth and, ultimately, a higher quality of life in developing countries.

See Also

- development economics
- education production functions
- human capital
- returns to schooling

Bibliography

Foster, A. and Rosenzweig, M. 1996. Technical change and human capital returns and investments: evidence from the Green Revolution. *American Economic Review* 86, 931–53.

Glewwe, P. 2002. Schools and skills in developing countries: education policies and socioeconomic outcomes. *Journal of Economic Literature* 40, 436–82.

Glewwe, P. and Kremer, M. 2006. Schools, teachers and education outcomes in developing countries. In *Handbook on the Economics of Education*, ed. E. Hanushek and F. Welch. Amsterdam: North-Holland.

Glewwe, P. and Miguel, E. 2006. The impact of child health and nutrition on education in less developed countries. In *Handbook of Agricultural Economics*, vol. 4, ed. R. Evenson and T. Schultz. Amsterdam: North-Holland.

Glewwe, P. and Zhao, M. 2005. Attaining universal primary completion by 2015: how much will it cost? Department of Applied Economics, University of Minnesota.

UNDP (United Nations Development Programme). 1990. *Human Development Report*. New York: UNDP.

World Bank. 2001. *World Development Report 2000/2001: Attacking Poverty*. Washington, DC: World Bank.

How to cite this article

Glewwe, Paul. "education in developing countries." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000248> doi:10.1057/9780230226203.0447

The New Palgrave Dictionary of Economics Online

education production functions

Eric A. Hanushek

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The accumulated economic analysis of education suggests that current provision of schooling is very inefficient. Commonly purchased inputs to schools – class size, teacher experience, and teacher education – bear little systematic relationship to student outcomes, implying that conventional input policies are unlikely to improve achievement. At the same time, differences in teacher quality have been shown to be very important. Unfortunately, teacher quality, defined in terms of effects on student performance, is not closely related to salaries or readily identified attributes of teachers.

Keywords

education production functions; random assignment; school attainment; school resources; student outcomes; teacher quality

Article

A simple production model lies behind much of the analysis in the economics of education. The common inputs are things like school resources, teacher quality, and family attributes; and the outcome is student achievement. Knowledge of the production function for schools can be used to assess policy alternatives and to judge the effectiveness and efficiency of public provided services. This area is, however, distinguished from many because the results of analyses enter quite directly into the policy process.

Historically, the most frequently employed measure of schooling has been attainment, or simply years of schooling completed. The value of school attainment as a rough measure of individual skill has been verified by a wide variety of studies of labour market outcomes (for example, Mincer, 1970; Psacharopoulos and Patrinos, 2004). However, the difficulty with this common measure of outcomes is that it assumes a year of schooling produces the same amount of student achievement, or skills, over time and in every country. This measure simply counts the time spent in schools without judging what

happens in schools – thus, it does not provide a complete or accurate picture of outcomes.
Recent direct investigations of cognitive achievement find significant labour market returns to individual differences in cognitive achievement (for example, Lazear, 2003; Mulligan, 1999; Murnane et al., 2000). Similarly, society appears to gain in terms of productivity; Hanushek and Kimko (2000) demonstrate that quality differences in schools have a dramatic impact on productivity and national growth rates. (A parallel line of research has employed school inputs to measure quality but has not been as successful. Specifically, school input measures have not proved to be good predictors of wages or growth.)
Because outcomes cannot be changed by fiat, much attention has been directed at inputs – particularly those perceived to be relevant for policy such as school resources or aspects of teachers.
Analysis of the role of school resources in determining achievement begins with the Coleman Report, the US government's monumental study on educational opportunity released in 1966 (Coleman et al., 1966). That study's greatest contribution was directing attention to the distribution of student performance – the outputs as opposed to the inputs.
The underlying model that has evolved as a result of this research is very straightforward. The output of the educational process – the achievement of individual students – is directly related to inputs that both are directly controlled by policymakers (for example, the characteristics of schools, teachers, and curricula) and are not so controlled (such as families and friends and the innate endowments or learning capacities of the students). Further, while achievement may be measured at discrete points in time, the educational process is cumulative; inputs applied sometime in the past affect students' current levels of achievement.
Family background is usually characterized by such socio-demographic characteristics as parental education, income, and family size. Peer inputs, when included, are typically aggregates of student socio-demographic characteristics or achievement for a school or classroom. School inputs typically include teacher background (education level, experience, sex, race, and so forth), school organization (class sizes, facilities, administrative expenditures, and so forth), and district or community factors (for example, average expenditure levels). Except for the original Coleman Report, most empirical work has relied on data constructed for other purposes, such as a school's standard administrative records. Based upon this, statistical analysis (typically some form of regression analysis) is employed to infer what specifically determines achievement and what is the importance of the various inputs into student performance.

Measured school inputs

The state of knowledge about the impacts of resources is best summarized by reviewing available empirical studies. Most analyses of education production functions have directed their attention at a relatively small set of resource measures, and this makes it easy to summarize the results (Hanushek, 2003). The 90 individual publications that appeared before 1995 contain 377 separate production function estimates. For classroom resources, only nine per cent of estimates for teacher education and 14 per cent for teacher–pupil ratios yielded a positive and statistically significant relationship between these factors and student performance. Moreover, these studies were offset by another set of studies that found a similarly negative correlation between those inputs and student achievement. Twenty-nine per cent of the studies found a positive correlation between teacher experience and student performance; however,

71 per cent still provided no support for increasing teacher experience (being either negative or statistically insignificant). Studies on the effect of financial resources provide a similar picture. These indicate that there is very weak support for the notion that simply providing higher teacher salaries or greater overall spending will lead to improved student performance. Per pupil expenditure has received the most attention, but only 27 per cent of studies showed a positive and significant effect. In fact, seven per cent even suggested that adding resources would harm student achievement. It is also important to note that studies involving pupil spending have tended to be the lowest-quality studies as defined below, and thus there is substantial reason to believe that even the 27 per cent figure overstates the true effect of added expenditure.

These studies make a clear case that resource usage in schools is subject to considerable inefficiency, because schools systematically pay for inputs that are not consistently related to outputs.

Study quality

The previous discussions do not distinguish among studies on the basis of any quality differences. The available estimates can be categorized by a few objective components of quality. First, while education is cumulative, frequently only current input measures are available, which results in analytical errors. Second, schools operate within a policy environment set almost always at higher levels of government. In the United States, state governments establish curricula, provide sources of funding, govern labour laws, determine rules for the certification and hiring of teachers, and the like. In other parts of the world, similar policy setting, frequently at the national level, affects the operations of schools. If these attributes are important – as much policy debate would suggest – they must be incorporated into any analysis of performance. The adequacy of dealing with these problems is a simple index of study quality.

The details of these quality issues and approaches for dealing with them are discussed in detail elsewhere (Hanushek, 2003) and only summarized here. The first problem is ameliorated if one uses the ‘value added’ versus ‘level’ form in estimation. That is, if the achievement relationship holds at different points in time, it is possible to concentrate on the growth in achievement and on exactly what happens educationally between those points when outcomes are measured. This approach ameliorates problems of omitting prior inputs of schools and families, because they will be incorporated in the initial achievement levels that are measured (Hanushek, 1979). The latter problem of imprecise measurement of the policy environment can frequently be ameliorated by studying performance of schools operating within a consistent set of policies – for example, within individual states in the USA or similar decision-making spheres elsewhere. Because all schools within a state operate within the same basic policy environment, comparisons of their performance are not strongly affected by unmeasured policies (Hanushek, Rivkin and Taylor, 1996).

If the available studies are classified by whether or not they deal with these major quality issues, the prior conclusions about research usage are unchanged (Hanushek, 2003). The best quality studies indicate no consistent relationship between resources and student outcomes.

An additional issue, which is particularly important for policy purposes, concerns whether this analytical approach accurately assesses the causal relationship between resources and performance. If, for example, school decision-makers provide more resources to those they judge as most needy, higher resources could simply signal students known for having lower achievement. Ways of dealing with this

include various regression discontinuity or panel data approaches. When done in the case of class sizes, the evidence has been mixed (Angrist and Lavy, 1999; Rivkin, Hanushek and Kain, 2005).

An alternative involves the use of random assignment experimentation rather than statistical analysis to break the influence of sample selection and other possible omitted factors. With one major exception, this approach nonetheless has not been applied to understand the impact of schools on student performance. The exception is Project STAR, an experimental reduction in class sizes that was conducted in the US state of Tennessee in the mid-1980s (Word et al., 1990). To date, it has not had much impact on research or our state of knowledge. While Project STAR has entered into a number of policy debates, the interpretation of the results remains controversial (Krueger, 1999; Hanushek, 1999).

Magnitude of effects

Throughout most consideration of the impact of school resources, attention has focused almost exclusively on whether a factor has an effect on outcomes that is statistically different from zero. Of course, any policy consideration would also consider the magnitude of the impacts and where policies are most effective. Here, even the most refined estimates of, say, class size impacts does not give very clear guidance. The experimental effects from Project STAR indicate that average achievement from a reduction of eight students in a classroom would increase by about 0.2 standard deviations, but only in the first grade of attendance in smaller classes (kindergarten or first grade) (see Word et al., 1990; Krueger, 1999). Angrist and Lavy (1999), with their regression discontinuity estimation, find slightly smaller effects in grade five and approximately half the effect size in grade four. Rivkin, Hanushek and Kain (2005), with their fixed effects estimation, find effects half of Project STAR in grade four and declining to insignificance by grade seven. Thus, from a policy perspective the alternative estimates are both small in economic terms when contrasted with the costs of such large class size reductions and inconsistent across studies.

Do teachers and schools matter?

Because of the Coleman Report and subsequent studies discussed above, many have argued that schools do not matter and that only families and peers affect performance. Unfortunately, these interpretations have confused measurability with true effects.

Extensive research since the Coleman Report has made it clear that teachers do indeed matter when assessed in terms of student performance instead of the more typical input measures based on characteristics of the teacher and school. When fixed effect estimators that compare student gains across teachers are used, dramatic differences in teacher quality are seen.

These results can also be reconciled with the prior ones. These differences among teachers are simply not closely correlated with commonly measured teacher characteristics (Hanushek, 1992; Rivkin, Hanushek and Kain, 2005). Moreover, teacher credentials and teacher training do not make a consistent difference when assessed against student achievement gains (Boyd et al., 2006; Kane, Rockoff and Staiger, 2006). Finally, teacher quality does not appear to be closely related to salaries or to market decisions. In particular, teachers exiting for other schools or for jobs outside of teaching do not appear to

be of higher quality than those who stay (Hanushek et al., 2005).

Some conclusions and implications

The existing research suggests inefficiency in the provision of schooling. It does not indicate that schools do not matter. Nor does it indicate that money and resources never impact achievement. The accumulated research surrounding estimation of education production functions simply says there currently is no clear, systematic relationship between resources and student outcomes.

See Also

- human capital
- local public finance
- returns to schooling

Bibliography

Angrist, J.D. and Lavy, V. 1999. Using Maimondides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114, 533–75.

Boyd, D., Grossman, P., Lankford, H., Loeb, S. and Wyckoff, J. 2006. How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy* 1, 176–216.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D. and York, R. L. 1966. *Equality of Educational Opportunity*. Washington, DC: US Government Printing Office.

Hanushek, E.A. 1979. Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources* 14, 351–88.

Hanushek, E.A. 1992. The trade-off between child quantity and quality. *Journal of Political Economy* 100, 84–117.

Hanushek, E.A. 1999. Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis* 21, 143–63.

Hanushek, E.A. 2003. The failure of input-based schooling policies. *Economic Journal* 113, F64–F98.

Hanushek, E.A., Kain, J.F., O'Brien, D.M. and Rivkin, S.G. 2005. The market for teacher quality. Working Paper No. 11154. Cambridge, MA: NBER.

Hanushek, E.A. and Kimko, D.D. 2000. Schooling, labor force quality, and the growth of nations. *American Economic Review* 90, 1184–208.

Hanushek, E.A., Rivkin, S.G. and Taylor, L.L. 1996. Aggregation and the estimated effects of school resources. *Review of Economics and Statistics* 78, 611–27.

Kane, T.J., Rockoff, J.E. and Staiger, D.O. 2006. What does certification tell us about teacher effectiveness? Evidence from New York City. Working Paper No. 12155. Cambridge, MA: NBER.

Krueger, A.B. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114, 497–532.

Lazear, E.P. 2003. Teacher incentives. *Swedish Economic Policy Review* 10(3), 179–214.

Mincer, J. 1970. The distribution of labor incomes: a survey with special reference to the human capital approach. *Journal of Economic Literature* 8, 1–26.

Mulligan, C.B. 1999. Galton versus the human capital approach to inheritance. *Journal of Political Economy* 107(pt. 2), S184–S224.

Murnane, R.J., Willett, J.B., Duhaldeborde, Y. and Tyler, J.H. 2000. How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management* 19, 547–68.

Psacharopoulos, G. and Patrinos, H.A. 2004. Returns to investment in education: a further update. *Education Economics* 12, 111–34.

Rivkin, S.G., Hanushek, E.A. and Kain, J.F. 2005. Teachers, schools, and academic achievement. *Econometrica* 73, 417–58.

Word, E., Johnston, J., Bain, H.P., DeWayne Fulton, B., Zaharies, J.B., Lintz, M.N., Achilles, C.M., Folger, J. and Breda, C. 1990. *Student/Teacher Achievement Ratio (STAR), Tennessee's K-3 Class Size Study: Final Summary Report, 1985–1990*. Nashville: Tennessee State Department of Education.

How to cite this article

Hanushek, Eric A. "education production functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000238> doi:10.1057/9780230226203.0448

The New Palgrave Dictionary of Economics Online

educational finance

William A. Fischel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The American system of government-financed education is decentralized among 50 states and more than 15,000 local school districts. Local funds are derived from local property taxes, and this system tends to make local spending unequal. State-government efforts to equalize education spending involve manipulating the local ‘tax price’ with matching grants. School districts with low tax prices are not, however, necessarily populated by rich people, so the distribution of state funds may penalize many low-income districts with large amounts of non-residential property.

Keywords

educational finance; local government; median voter; property taxation; school districts (USA); school vouchers; spatial competition; tax price of school spending; Tiebout hypothesis

Article

This article deals with the government-financed system of education in the United States, which is referred to as ‘public’ education. Educational finance in the United States is different from that of other nations, which typically fund education from national taxes. Within each American state, a substantial portion of education is financed by local governments, although the proportion financed locally has declined from 83.2 per cent in 1920 to 43.2 per cent in 2000.

The state–local system of finance stems from the history and geography of the United States and the federal nature of its government. The 50 states are, in the eyes of the national government, primarily responsible for education. In most states, implementation of this responsibility is delegated to local municipal corporations called ‘school districts’. The school district is more than a local administrative agency of the state. It is a distinct political entity that usually has some correspondence with the geographic area of a municipality. The district, however, has a separate board of directors, which is locally elected. The board then selects a superintendent of schools to manage the district's education. Boards have the authority to levy taxes, which are almost always on property within their district, and spend the revenue they derive from them. The state government may prescribe curricular standards for public schools, but the method of achieving these standards is the responsibility of the local district.

School districts and school boards were once the most common form of local government in the United States, numbering about 200,000 in 1900. The number of school districts declined steadily throughout the 20th century, which can largely be accounted for by the consolidation of rural one-room school districts into larger units. By 1970, one-room schools were essentially extinct, and since 1970 the total number of school districts has declined only slightly, numbering about 16,000 at the beginning of the 21st century.

Despite their numerical decline in rural areas, there are many school districts in most metropolitan areas. Urban households that are already on the move for job-related reasons have the luxury of choosing a home within one of several school districts in most regions of the nation. Choosing among school districts and the resulting competition among districts to obtain residents is consistent with the model proposed by Tiebout (1956). Numerous tests of the Tiebout model indicate that the quality of schooling is important to most home buyers (Oates, 1969; Bradbury, Case and Mayer, 2001). There is also evidence that spatial competition makes school districts more efficient in delivering education services (Hoxby, 2000).

One-room schools of the 19th century were usually ‘ungraded’. Students were instead divided into skill-specific recitation groups, formed without regard for chronological age. In this system, uniformity of education was not critical. New pupils could be placed according to what they knew in particular subjects rather than by age. But when almost all schools were age-graded, it paid for each district to offer an age-specific curriculum that allowed both teachers and pupils to be interchangeable among schools and districts (Fischel, 2006a).

Standardization of age-graded curricula became widespread by about 1940 and was brought about by two forces, one local and the other statewide. Property-owning voters in a given district would find that potential homebuyers would shun them if they did not offer a standard, public-school education. Voters would thus support taxes necessary to fund standardized schools. However, differences in the economic make-up and tax-bases of local districts sometimes made this difficult to do.

Figure 1 illustrates the problem for attempts to fund schools from local sources. It depicts a trade-off between local school spending and other goods for the median voter (the voter with the median income, assumed always to be in the majority in local elections) in two separate communities, a rich district and a poor district. The decisive voter chooses the mix of school spending and private goods that achieves the highest indifference curve that his private–public budget line allows (Bergstrom and Goodman, 1973). Because at the local level education is essentially a private good, the slope of the budget lines is the ‘tax price’ of school spending for the median voter in each community.

Figure 1

School spending in rich and poor districts

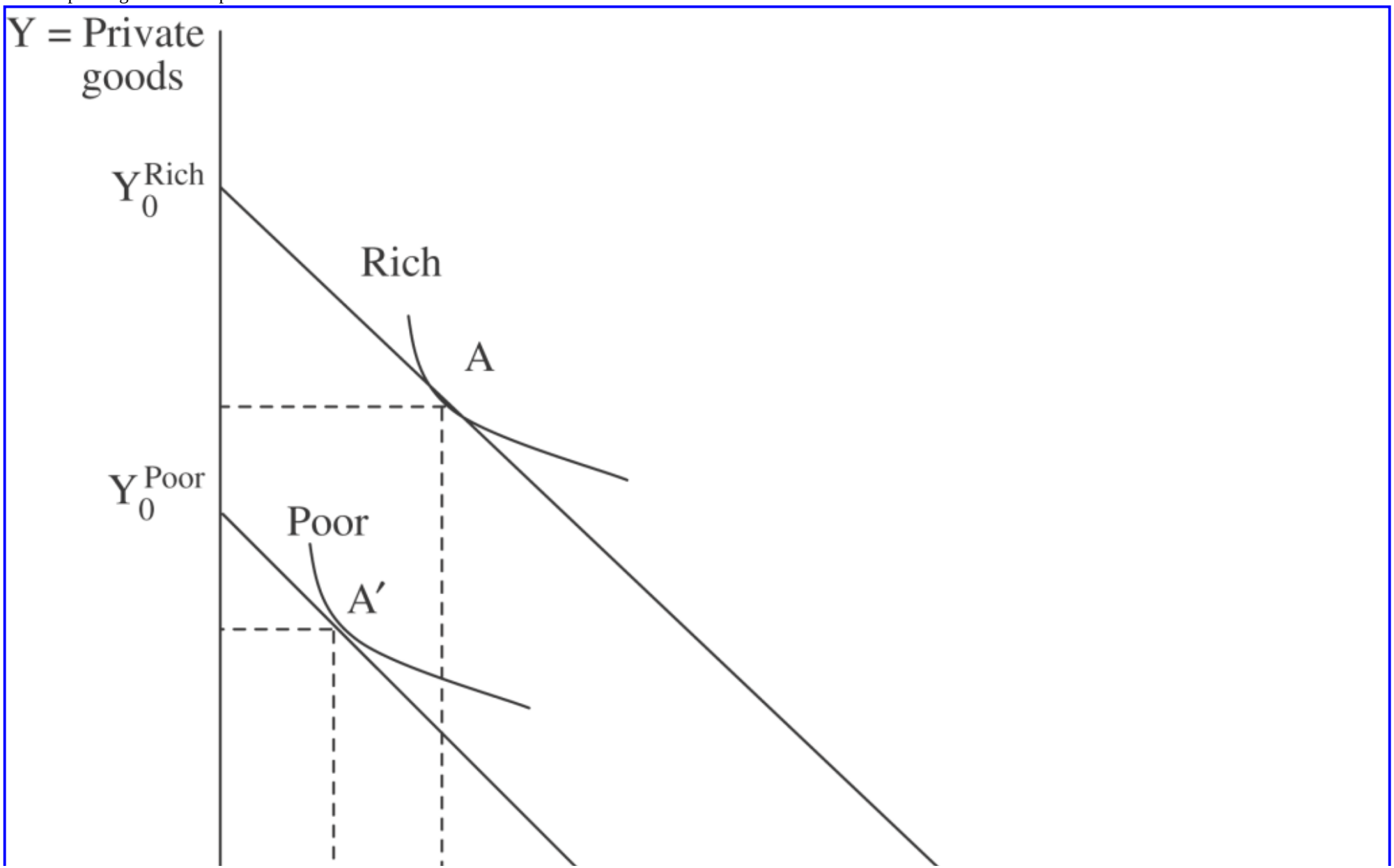

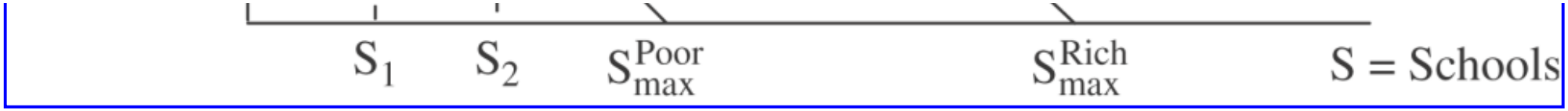

The tax price is not a tax rate. A school district composed exclusively of mansions will have, for a given level of spending, a much lower property tax *rate* than a district composed of modest-sized homes. But if the second moment of the distribution of wealth is the same in both communities, the *tax price* faced by the median voter in each will be the same. A 1,000 dollar increase in per-pupil spending will cost the median voter the same amount of money in both cases, if one assumes that the number of public-school children per household is the same in both.

The other generalization that Figure 1 illustrates is that average income of a district accounts for much of the differences in spending per pupil. Even though the tax prices are the same, the positive income elasticity of demand for education (estimated at somewhere between 0.5 and 1.0) causes the richer community to choose a higher level of school inputs (Bergstrom, Rubinfeld and Shapiro, 1982). While much of the criticism of these differences is based on equity concerns, there are efficiency reasons to promote a relatively uniform system of education (Benabou, 1996).

The way most states have attempted to equalize education opportunities is to reduce the tax price of spending in poorer districts. State funds (from statewide taxes) are offered to the poorer community in proportion to the district's own tax effort. The poorer median voter thus perceives, as indicated by the dotted budget line in Figure 2, that for every dollar raised locally, the state will send it another dollar. The tax price has been cut in half in the graphical example, so that the poorer community will choose to spend an amount closer to that of the richer district.

Figure 2

Subsidies to poor districts

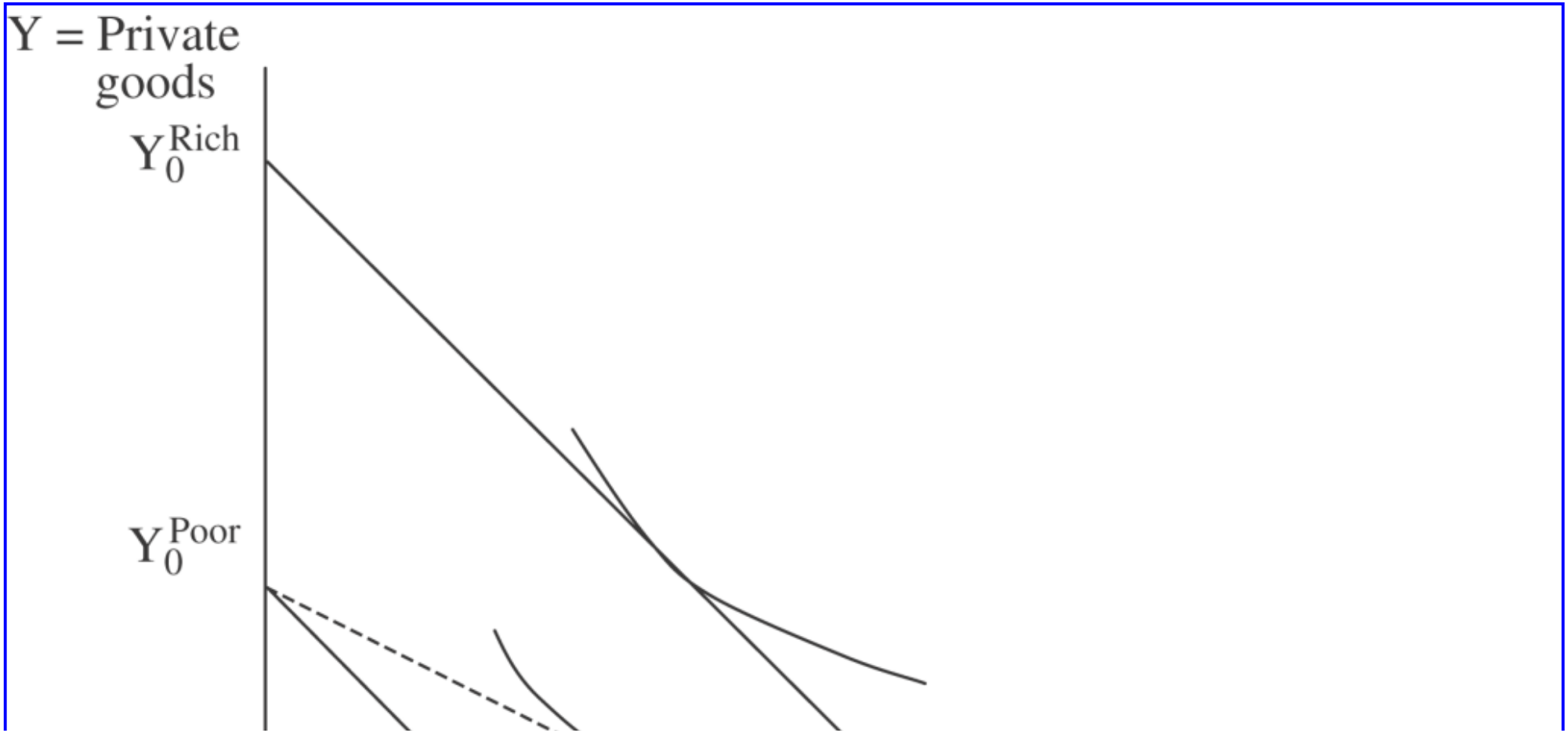

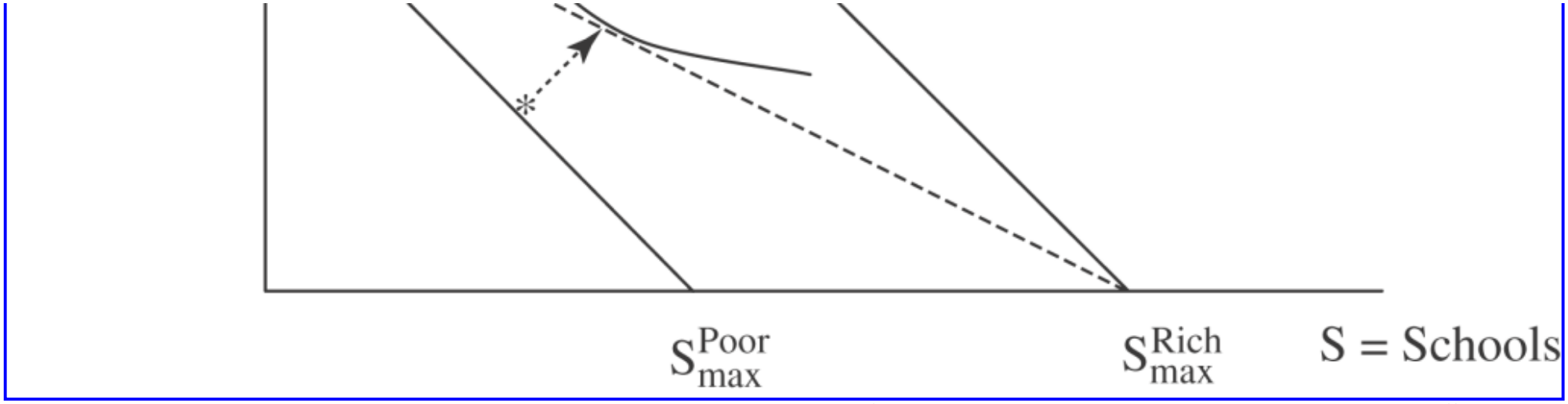

By manipulating the local tax price, state governments can in principle induce a substantial equality of school spending in nominally independent districts, though state officials still seem surprised that there is an income effect as well as a substitution effect from lowering the tax price. They seem to expect that the arrow in Figure 2 should point horizontally to the right. Instead, local voters use the subsidy (the reduced tax price) to both increase local spending on schools, which is the desired substitution effect, and to reduce their own local taxes (nudging the arrow's direction upwards), which is the income effect.

Another factor can also account for differences in local tax prices. The poorer district may have a substantial amount of non-residential property to tax. Commercial and industrial uses do not come with children attached (at least in metropolitan areas, where workers can live in other communities), and so their tax revenues amount to a subsidy to their school district. The effect of this is the same as a matching-grant subsidy by the state. And the effect is not trivial. Nationally, almost one-half of all property taxes are paid by non-residential property owners, which puts them on the same order of magnitude as state funds for public education.

Although both state subsidies and a large non-residential tax base reduce the tax price, they have been treated differently in recent years. The school finance litigation movement began with *Serrano v. Priest* in California in 1971 (Brunner and Sonstelie, 2006). Its objective was to use state constitutional directives (equal protection and school funding clauses) to improve schools in poor districts. For strategic reasons, the movement focused its remedial efforts on differences in tax base per pupil rather than differences in spending per pupil or on educational outcomes. Many state courts thus ruled that unequal tax bases, not unequal spending, were constitutionally suspect and ordered legislatures to transfer funds from the ‘property rich’ to the ‘property poor’.

What this remedy overlooked is that low-income communities are as likely to be ‘property rich’ (on the widely used ‘tax base per-pupil’ standard) as high-income communities. This is because many urban districts have a large non-residential property tax base that offsets the lower valued residential tax base. (The poor may have migrated there for jobs or rezoned land to attract industry, something most affluent suburbs are reluctant to do.) Besides this, poorer cities often have relatively few children in public schools because of an aged population or because low-quality public schools encourage the use of private schools. In any case, many of the court-induced ‘equalization’ remedies have actually caused state funds to be removed from low-income (but ‘property rich’) districts to higher-income districts that are ‘property poor’ because of their modest nonresidential tax base and large school-age population.

An alternative response to the difficulties of distributing state funds to school districts is simply to have the state government run the schools without the intermediation of local school boards and districts. Another is a voucher system, in which the state gives public funds to parents and allows them to select whatever school they want. Both are certainly viable means of school finance, and it is worth asking why they have not been embraced.

Full state funding forgoes the local monitoring of school performance by voters. Capitalization of school quality in local home values creates a feedback mechanism for local governance. The median voter in most jurisdictions is a homeowner, and voters therefore care about the consequences of school governance. School superintendents who waste local taxpayers’ money will find that their tenure is short as voters become dissatisfied. Even if they keep their jobs, the declines in taxable property value due to inefficient policies will leave them with less revenue to spend in the future (Hoxby, 1999). Neither of these desirable feedback effects is likely to occur under a state-managed system.

The drawback of school vouchers appears to be that voters are reluctant to embrace them as a general practice. American voters appear to perceive benefits from local public schools that go beyond educational qualities. One benefit I have advanced is that public schools create location-specific social capital among adults (Fischel, 2006b). Adults with children are more likely to know the parents of their children's schoolmates. This creates a network of adult social capital that lowers the transaction costs of public participation in municipal affairs. A voucher system disperses children to various schools and thus does not create the same location-specific social capital that public schools do. In any case, America's continuing embrace of locally run and locally financed public education reflects the school's central role in facilitating local self-governance.

See Also

- exit and voice
- fiscal federalism
- local public finance
- property taxation
- public choice
- school choice and competition
- Tiebout hypothesis

Bibliography

Benabou, R. 1996. Heterogeneity, stratification, and growth: macroeconomic implications of community structure and school finance. *American Economic Review* 86, 584–609.

Bergstrom, T.C. and Goodman, R.P. 1973. Private demand for public goods. *American Economic Review* 63, 280–96.

Bergstrom, T.C., Rubinfeld, D.L. and Shapiro, P. 1982. Micro-based estimates of demand functions for local school expenditures. *Econometrica* 50, 1183–205.

Bradbury, K.L., Case, K.E. and Mayer, C. 2001. Property tax limits, local fiscal behavior, and property values: evidence from Massachusetts under proposition 2 1/2. *Journal of Public Economics* 80, 287–311.

Brunner, E.J. and Sonstelie, J. 2006. California's school finance reform: an experiment in fiscal federalism. In *The Tiebout Model at Fifty*, ed. W.A. Fischel. Cambridge, MA: Lincoln Institute of Land Policy.

Fischel, W.A. 2006a. Will I see you in September? An economic explanation for the standard school calendar. *Journal of Urban Economics* 59, 236–51.

Fischel, W.A. 2006b. Why voters veto vouchers: public schools and community-specific social capital. *Economics of Governance* 7, 109–32.

Hoxby, C.M. 1999. The productivity of schools and other local public goods producers. *Journal of Public Economics* 74, 1–30.

Hoxby, C.M. 2000. Does competition among public schools benefit students and taxpayers? *American Economic Review* 90, 1209–38.

Oates, W.E. 1969. The effects of property taxes and local public spending on property values: an empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy* 77, 957–71.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

How to cite this article

Fischel, William A. "educational finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000237> doi:10.1057/9780230226203.0449

The New Palgrave Dictionary of Economics Online

effective demand

J.A. Kregel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

By ‘effective demand’ Keynes meant the forces determining changes in the scale of output and employment as a whole. It was intended to replace Say's Law. For Keynes, since entrepreneurs maximized monetary returns, not employment or physical output, there was no reason why their investment decisions should lead to an equilibrium at full employment. Since this account permitted any level of employment to emerge as a stable equilibrium, including full employment, it is more general than the classical Say's Law position, in which the only stable equilibrium was the limit set by full employment as given in the labour market.

Keywords

effective demand; excess supply; full employment equilibrium; gold; Hume, D.; liquidity preference; Malthus, T. R.; marginal efficiency of capital; mercantilism; Mill, J. S.; multiplier; natural price; natural rate and market rate of interest; propensity to consume; quantity theory of money; rate of interest; saving and investment; Say's Law; Smith, A.; subjective theory of value; supply and demand; velocity of circulation

Article

‘Effective demand’ is the term used by Keynes in his *General Theory* (1936a) to represent the forces determining changes in the scale of output and employment as a whole. Keynes attributed the first discussions of the determinants of the supply and demand for output as a whole to the classical economists, in particular the debate between Ricardo and Malthus concerning the possibility of ‘general gluts’ of commodities, or what has come to be known as Say's Law of Markets. Indeed, Keynes's theory was intended to replace Say's Law, although the emergence of effective demand from his *Treatise on Money* (1930) critique of the quantity theory of money, and his insistence on its application in what he originally called a ‘monetary production economy’, suggests that it should also be seen in antithesis to

classical monetary theory. For Adam Smith (1776, p. 285), ‘A man must be perfectly crazy who … does not employ all the stock which he commands, whether it be his own or other peoples’ on consumption or investment. As long as there was what Smith called ‘tolerable security’, economic rationality implied that it was impossible for demand for output as a whole to diverge from aggregate supply. Although Smith (1776, p. 73) did call the demand ‘sufficient to effectuate the bringing of the commodity to the market’, the ‘effectual demand’ ‘of those who are willing to pay the natural price’ of the commodity, the idea referred to divergence of market from natural price of particular commodities and the process of gravitation of prices to their natural values. J.B. Say's discussion of the problem of the ‘disposal of commodities’ adopted Smith's position. Against those who held that ‘products would always be abundant, if there were but a ready demand, or market for them,’ Say's ‘law of markets’ argued ‘that it is production which opens a demand for products’ (1855, pp. 132–3); if production determined ability to buy, then demand could not be deficient. While excesses in particular markets were admitted, they would always be offset by deficiencies in others. Ricardo used similar arguments against Malthus, who responded by suggesting that:

> from the want of a proper distribution of the actual produce, adequate motives are not furnished to continued production,•…•the grand question is whether it [actual produce] is distributed in such a manner between the different parties concerned as to occasion the most effective demand for future produce … (Malthus, 1821)

Malthus argues that the composition of output affects its quantity by producing doubts in the minds of Smith's rational entrepreneurs concerning the ‘security’ of their future profit.

The final word in the classical debate was J.S. Mill's ‘On the Influence of Consumption on Production’, which sought exceptions to the proposition that ‘All of which is produced is already consumed, either for the purpose of reproduction or enjoyment’ so that ‘There will never, therefore, be a greater quantity produced, of commodities in general, than there are customers for’ (1874, pp. 48–9). Mill accused those who argued that demand limits output of a fallacy of composition, for the individual shopkeeper's failure to sell is due to a disproportion of demand which cancels out for the nation as a whole. Mill also notes that the argument that every purchaser must be a seller presumes barter, for money enables exchange ‘to be divided into two separate acts’ so one ‘need not buy at the same moment when he sells’ (p. 70). To avoid this problem ‘money must itself be considered as a commodity’, for ‘there cannot be an excess of all other commodities, and an excess of money at the same time’ (p. 71). Mill admits that if money were ‘collected in masses’, there might be an excess of all commodities, but this would mean only a temporary fall in the value of all commodities relative to money. Similarly to Smith's ‘tolerable security’, Mill explains an excess of commodities in general by ‘a want of commercial confidence’, which he denies may be caused by an overproduction of commodities (p. 74).

Mill's defence of Say's Law highlights the importance of the classical quantity theory, which was originally formulated to oppose the undue emphasis given to precious metals as components of national wealth by the mercantilists. Hume noted that labour, not gold, produced the commodities which composed national wealth; that gold was only as good as the labour it commanded to produce output. Thus the classical position that the velocity of circulation of money was independent of its quantity was built on the view that money would only be held to be spent. Money could at best cause temporary general gluts; in the long term, ‘rational’ men would not choose to hold money rather than spend it.

On the eve of the marginal revolution, classical theory thus admitted the temporary occurrence of general gluts explained by cyclical disproportions in demand for money and commodities due to crises of confidence. It is paradoxical that, while the marginal revolution was motivated by the failure of classical theory to give sufficient attention to the role of demand in value theory, it failed to extend its analysis of demand to output as a whole in either the long or the short period. Indeed, the emphasis on individual equilibrium produced by the subjective theory of value which replaced the classical theory, made separate discussion of aggregate supply and demand redundant. Thus Keynes's reference to ‘the disappearance of the theory of demand and supply for output as a whole, that is the theory of employment *after* it has been for a quarter of a century the most discussed thing in economics’ (Keynes, 1936c).

But it was discussion, not Say's Law, which disappeared from neoclassical economics. Thus Keynes classed economists from Smith and Ricardo to Marshall and Pigou as ‘Classical’, for, despite antagonistic theories of value and distribution, they all held a similar theory of supply and demand for output as a whole.

Keynes suggests that this was due more to the failure of neoclassical economists to heed Mill's warning concerning the extension of the conditions faced by the individual to the economy as a whole, than to positive analysis. If consumers (producers) maximize utility (profit) subject to an income (cost) constraint, reaching the maximum by substituting in consumption (production) goods (inputs) which were cheaper per unit of utility (output), then excess supply of any good (resource) is due to its price exceeding its marginal utility (productivity). Market competition would lead to relative price adjustments which eliminate excess supply. Since it was impossible for any single good (resource) to be unsold (unemployed), it was natural to extend this analysis to the aggregate level to deny the possibility of general gluts without further analysis.

Any divergence from this position was explained, not by reference to hoarding money due to crises of confidence, but by temporary impediments to the automatic adjustment of relative prices in competitive markets. Thus, despite their new marginal theory of value, Keynes's contemporaries reached a similar result that divergence of employment from its full employment level would be determined by temporary non-persistent causes eliminated in the long run.

From 1921 to 1939 the unemployment rate in the United Kingdom never fell below ten per cent, peaking in 1932 at 22.5 per cent (over 2.7 million). This exceeded the limits that most economists attributed to short-period frictions. The self-adjusting nature of the neoclassical version of Say's Law that Keynes chose to criticize was thus contradicted by reference to economic events as well as by Keynes's conception of effective demand.

Keynes was not concerned with impediments to the equality of the supply and demand, but with the

> problem of the equilibrium of supply and demand for output as a whole, in short, of effective demand … When one is trying to discover the volume of output and employment, it must be this point of equilibrium for which one is searching.

While the Classics solved the problem by assuming the identity of savings and expenditure on investment goods, neoclassical theory presumed Say's Law ‘without giving the matter the slightest discussion’ (1936b, p. 215).

Keynes's theory of effective demand thus had to replace Say's Law. To do this Keynes departed from the Classical position on two points. The first was to assume that wages exceed subsistence so that expenditure on consumption goods does not exhaust factor incomes. As expressed in Keynes's psychological law of consumption, this implied that as output increased, the gap between aggregate expenditure and factor costs increased, so that unless investment expenditure expanded to fill the gap, entrepreneurs would experience losses.

The second departure was from the assumption that rationality dictated that entrepreneurs' savings represented productive investment expenditure. If investment could produce losses, or changes in interest rates change capital values, then greater future enjoyment might be assured by not investing; holding money might be 'rational' in such conditions. Further, in a monetary economy, nothing guarantees that maximization of returns in money will maximize either productive capacity or the demand for labour.

In Keynes's theory the propensity to consume and the multiplier produce the proposition that it is the level of output which adjusts saving to investment, rather than the rate of interest, while the explanation of the decisions over the level of investment in a monetary economy requires an explanation of rates of interest in money terms. The two factors are closely related.

In a 1934 letter to Kahn, Keynes gives a 'precise definition of what is meant by effective demand' (1934a, p. 422). If O is the level of output, W the marginal prime cost of production for that output, and P the expected selling price, 'Then OP is effective demand'. The classical theory that 'supply creates its own demand' assumes that OP equals OW, irrespective of the value of O, 'so that effective demand is incapable of setting a limit to employment which consequently depends on the relation between marginal product in wage-goods industries and marginal disutility of employment'. Thus, what Keynes later called (1936a, ch. 2) the two 'classical' postulates limit O at full employment. In contrast,

> On my theory $OW \neq OP$ for *all* values of O, and entrepreneurs have to choose a value of O for which it is equal – otherwise the equality of price and marginal prime cost is infringed. This is the real starting point of everything.

The key point was thus the impact of different levels of O on the difference between costs and prices, that is on entrepreneurs' profits. Keynes took up this question, in an undated exchange with Sraffa of about the same time (1934b, pp. 157ff). Keynes notes that a non-unitary marginal propensity to consume implies $OP \neq OW$ for any O, and generates

> the general principle that *any* expansion of output gluts the market unless there is a *pari passu* increase of investment appropriate to the community's marginal propensity to consume; and any contraction leads to windfall profits to producers unless there is an appropriate *pari passu* contraction of investment.

The level of O at which $OP = OW$ will be determined by the level of investment and the propensity to consume. Changes in the rate of investment, based on entrepreneurs' expectations of their future profits, will determine O.

In an early draft of the *General Theory* Keynes (1973a, p. 439) put it this way:

> Effective demand is made up of the sum of two factors based respectively on the expectation of what is going to be consumed and on the expectation of what is going to be invested.

Thus the theory of effective demand required, in addition to explanation of consumption based on the propensity to consume, an explanation of variations in the level of investment. Since neoclassical theory resolved this problem by presuming that investment was brought into balance with full employment saving by means of the rate of interest, Keynes located the 'flaw being largely due to the failure of the Classical doctrine to develop a satisfactory theory of the rate of interest' (1934c, p. 489).

Keynes concentrated his efforts to produce a theory of interest compatible within this theory of effective demand within what he called a monetary production economy. The *Treatise on Money* (1930) had explained changes in prices in terms of households' consumption decisions relative to entrepreneurs' production decisions. If these decisions were incompatible, investment diverged from saving and prices of consumption goods adjusted producing windfall profits or losses. The prices of investment goods were determined separately from this process, by means of the interaction of the bearishness of the public reflecting their decisions to hold bank deposits or securities on the one hand, and the monetary policy of the banking system on the other.

Investment goods are held because their present costs or supply prices are lower than the present value of their anticipated future earnings or demand prices; the larger this difference, the higher the expected rate of return. Since any change in the price of a durable capital asset will influence its rate of return, a theory that explains the price of capital assets also explains rates of return (which Keynes called marginal efficiency). With the demand price of an asset based on the value of expected future earnings discounted by the rate of interest, it is clear why a satisfactory theory of interest is crucial to the explanation of effective demand.

But money was a durable asset like any other, and as such it has a spot or demand price and a supply price or forward price, which determine the money rate of interest. Keynes thus transformed his concept of bearishness into liquidity preference which, together with banking policy, would determine the rate of interest. For Keynes, 'the money rate of interest … is nothing more than the percentage excess of a sum of money contracted for forward delivery … over what we may call the "spot" or cash price of the sum thus contracted for forward delivery' (1936a, p. 222), it is:

> the premium obtainable on current cash over deferred cash … No one would pay this premium unless the possession of cash served some purpose, that is had some efficiency. Thus we may conveniently say that interest on money measures the marginal efficiency of money measured in terms of itself as a unit. (1937a, p. 101)

Since both money and capital assets had marginal efficiencies representing their rates of return, profit-maximizing individuals in a monetary economy would demand money and capital assets in proportions which equated their respective returns. The equilibrium level of output chosen by entrepreneurs would then be represented by equality of the marginal efficiency of capital and the rate of interest (the marginal efficiency of money). The question of the effect of an increase in output on profit raised by a propensity to consume less than unity can now be seen as the effect of an increase in investment on the marginal

efficiency of money relative to the marginal efficiencies of capital assets. Since these marginal efficiencies reflect pairs of spot and forward asset prices, the question can also be put as the effect of an increase in investment on relative money prices. Thus Keynes's independent variables, the propensity to consume, the efficiency of capital and liquidity preference, given expectations and monetary policy, interact to determine effective demand.

Since this equilibrium could be described by $S = I$, or equality between the rate of interest and the marginal efficiency of capital, the level of output which equates aggregate demand and supply also equates marginal efficiency with the rate of interest. To complete his theory of effective demand, Keynes faced the question first raised by Wicksell of the causal relation between the natural and the money rate of interest. Just as Keynes rejected the determination of the level of O at which $OP = OW$ by the equality of the marginal productivity and disutility of labour, he rejected marginal productivity as the determinant of marginal efficiency and the real rate of interest determining the money rate because it was based on ‘circular reasoning’ (1937b, p. 212).

Keynes argues instead that it is the marginal efficiency of capital assets which adapts to the money rate of interest rather than vice versa. These two points of departure are discussed in Chapters 16 and 17 of the *General Theory*, where Keynes points out that the money rate of return to be expected from a capital asset depends on the relation of anticipated money receipts relative to expected money costs, and that there is no reason to believe that these will be related in any predictable way to the asset's physical productivity. Wicksell's natural rate, derived from physical relations of production and exchange, has no application in a monetary economy; Keynes thus substitutes the concept of marginal efficiency.

Keynes also notes that increased investment in particular capital assets increases supply prices and reduces demand prices, causing a decline in marginal efficiencies; an increase in output thus leads to investment in assets with lower rates of return. At some point the marginal efficiency of money will make investment in money as profitable as the purchase of capital assets. At this point the rate of interest equals the marginal efficiency of capital, and any further increase in output would confirm Keynes's ‘general principle’ that any further expansion in output gluts the market, for increased income is not spent but held in the form of money which becomes a ‘generalised sink for purchasing power’.

The question that distinguishes Keynes's theory is thus why money's liquidity premium does not fall as output expands, for this is what prevents investment from rising by just the amount to fill the gap created by the propensity to consume being less than one. To describe these ‘essential properties of interest and money’, Keynes departs from Mill's position that money is just another commodity. When money is the debt of the banking system its price and quantity behaviour will differ from physical commodities, for it has no real costs of production nor real substitutes. Thus an asset which has a negligible elasticity of production and substitution with respect to a change in effective demand, will have a rate of return which responds less rapidly to an expansion in demand. As long as the rate of interest falls less rapidly than the marginal efficiencies of capital assets, its rate will be the one which sets the point at which further expansion creates losses.

Thus the propensity to consume shows that investment will have to increase by the amount of the gap between incomes and expenditures as incomes rise if entrepreneurs are not to make losses, while the marginal efficiency of capital and liquidity preference in a monetary production economy explain why the behaviour of the rate of interest relative to the marginal efficiency of capital makes it unlikely that the rate of investment should adjust by just that amount. Since entrepreneurs maximize monetary returns, not employment or physical output, there is no reason why their investment decisions should

lead to an equilibrium at full employment. Keynes's explanation of the limit to the level of employment permits any level as a stable equilibrium, including full employment; it is thus more general than the classical Say's Law position, in which the only stable equilibrium was the limit set by full employment as given in the labour market.

See Also

- Say's Law

Bibliography

Keynes, J.M. 1930. *A Treatise on Money*. Reprinted in Keynes (1971).

Keynes, J.M. 1934a. Letter to R.F. Kahn, 13 April. Reprinted in Keynes (1973b).

Keynes, J.M. 1934b. Letter to P. Sraffa, undated. Reprinted in Keynes (1979).

Keynes, J.M. 1934c. Poverty in plenty: is the economic system self-adjusting? Reprinted in Keynes (1973b).

Keynes, J.M. 1936a. *The General Theory of Employment, Interest and Money*. Reprinted in Keynes (1973a).

Keynes, J.M. 1936b. Letter to A. Lerner, 16 June. Reprinted in Keynes (1979).

Keynes, J.M. 1936c. Letter to R.F. Harrod, 30 August. Reprinted in Keynes (1973c).

Keynes, J.M. 1937a. The theory of the rate of interest. Reprinted in Keynes (1973c).

Keynes, J.M. 1937b. Alternative theories of the rate of interest. Reprinted in Keynes (1973c).

Keynes, J.M. 1971–83. *The Collected Writings of John Maynard Keynes*, ed. D. Moggridge. London: Macmillan for the Royal Economic Society: 1971. Vols. 5 and 6. *A Treatise on Money* (1930). 1973a. Vol. 7. *The General Theory of Employment, Interest and Money* (1936). 1973b. Vol. 13. *The General Theory and After: Part I – Preparation*. 1973c. Vol. 14 *The General Theory and After: Part II – Defence and Development*. 1979. Vol. 29. *The General Theory and After – A Supplement*.

Malthus, T.M. 1821. Letter from Malthus to Ricardo, 7 July. Reprinted in Ricardo (1952), 9–10.

Mill, J.S. 1874. On the influence of consumption on production. In J.S. Mill, *Essays on Some Unsettled Questions of Political Economy*, 2nd edn, reprinted Clifton, NJ: A.M. Kelley, 1974.

Ricardo, D. 1952. *Works and Correspondence of David Ricardo*, vol. 9, ed. P. Sraffa with the collaboration of M. Dobb. Cambridge: Cambridge University Press.

Say, J.B. 1855. *A Treatise on Political Economy*. 6th American edn, Philadelphia: J.B. Lippincott.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Oxford University Press, 1976.

How to cite this article

Kregel, J. A. "effective demand." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000044> doi:10.1057/9780230226203.0450

The New Palgrave Dictionary of Economics Online

efficiency bounds

Han Hong
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In large sample analysis, the performances of estimators can be approximated by the asymptotic variances. In parametric models, maximum likelihood estimators often achieve the efficient Cramer–Rao lower bound, while efficient GMM estimation can be achieved by choosing the weighting matrix and the instruments optimally. Semiparametric efficiency bound is defined by the supremum of the Cramer–Rao bounds for all parametric models that satisfy the semiparametric restrictions. The efficiency bounds for asymptotically linear semiparametric estimators are given by the variances of the efficient influence functions, which are the projections of the linear influence functions onto the tangent spaces of the semiparametric models.

Keywords

asymptotic efficiency; asymptotic variance; Asymptotically linear estimators; average risk optimality; Cramer–Rao lower bound; delta method; efficiency bounds; Euler equations; generalized method of moment; Hodges' estimator; invariance principle; maximum likelihood; mean square errors; measurement error models; measures of closeness; minimax optimality; parametric models; partial linear model; semiparametric efficiency bound; semiparametric models; unbiased estimators

Article

Oftentimes we want to compare estimators. For a given parameter in which we are interested, there are typically many estimators that can estimate it consistently. We need to choose the best estimator, or the estimator that is the closest to the true parameter value. The mean square error (MSE), $E(\hat{\theta} - \theta)^2$, is frequently used as a measure of closeness. However, there can be many other various measures of closeness, and often they do not agree with each other. See, for example, Amemiya (1994, pp. 116–24).

Even with a given measurement of closeness, such as the MSE, it is typically not possible to rank two estimators. For two estimators X and Y of θ, X is better than Y only if $E(X - \theta)^2 \leq E(Y - \theta)^2$ for all $\theta \in \Theta$. An estimator that is not dominated by another estimator in the above sense is called admissible. A uniformly 'most' efficient estimator does not exist. To find an efficient estimator, one needs to confine the analysis to a limited class of estimators, such

as unbiased estimators or equivariant estimators. Alternatively, one can rely on a subjective strategy such as average risk optimality which requires a prior distribution over the parameter space, or use a pessimistic and risk-averse approach such as minimax optimality.

In large sample analysis, the performance measures of estimators can often be approximated by their asymptotic distribution. Under suitable regularity conditions, many estimators are consistent and converge to the true parameter values at $\sqrt{n}$ rate. These estimators can be compared based on their asymptotic variance. The notation of efficiency bound usually refers to the largest lower bound for the variances that can be achieved by $\sqrt{n}$ consistent and asymptotically normal estimators under suitable regularity conditions.

Asymptotic efficiency in parametric models

In parametric models, the variance of an unbiased estimator has to be larger than the Cramer–Rao lower bound, which is defined as the inverse of the information matrix:

$$V(\hat{\theta}) \geq -\left(E\frac{\partial^2 \log L}{\partial \theta^2}\right)^{-1},$$

where L is the likelihood function. Proofs of this result can be found, for example, in Amemiya (1994, pp. 138–39; 1985, pp. 14–17). A consistent estimator is said to be asymptotically efficient if its asymptotic variance achieves the Cramer–Lao lower bound. Under suitable regularity assumptions such as those given in Theorem 4.1.3 in Amemiya (1985), the maximum likelihood estimator is asymptotically efficient.

There exist super-efficient estimators whose asymptotic variances are smaller than the Cramer–Rao lower bound on a set of parameter θ with Lebesgue measure zero, such as Hodges's estimator defined as

$$w_T = \begin{cases} 0 & \text{if} \quad |\hat{\theta}| < T^{1/4} \\ \hat{\theta} & \text{if} \quad |\hat{\theta}| \geq T^{-1/4} \end{cases}.$$

where $\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, v(\theta))$. One can show that $\sqrt{T}(w_T - \theta) \xrightarrow{d} N(0, v(\theta))$ if $\theta \neq 0$ and $\sqrt{T}(w_T) \xrightarrow{d} 0$ if $\theta = 0$. However, the better behaviour of w_T at $\theta = 0$ comes at the expense of erratic behaviour when θ is close to 0. See for example, van der Vaart (1999, p. 110).

A common alternative to maximum likelihood is generalized method of moment estimators (GMM). Its asymptotic efficiency is extensively discussed in Newey and McFadden (1994). While GMM estimators are less efficient than maximum likelihood (see, for example, the proof in Newey and McFadden (1994, p. 2163), oftentimes they are easy to compute, especially when maximum likelihood is computationally infeasible. For a given set of unconditional moment conditions, a proper choice of the weighting matrix or the linear combination matrix minimizes the asymptotic variance. For a given set of

conditional moment conditions, a proper choice of instruments can also minimize the asymptotic variance.

A GMM estimator can be formed from the over-identified moment conditions $Em(z;\theta)\equiv 0$ by minimizing a quadratic form based on a weighting matrix W:

$$\frac{1}{T}\sum_{t=1}^{T} m(z_t;\hat{\theta})\, W \frac{1}{T}\sum_{t=1}^{T} m(z_t;\hat{\theta}) .$$

The resulting estimator has asymptotic variance $(G'WG)^{-1}(G'W\Omega WG)(G'WG)^{-1}$, where $G = E\frac{\partial}{\partial\theta}m(z;\theta)$ and $\Omega = Var(m(z;\theta))$. Hansen (1982) showed that the optimal choice of $W=\Omega^{-1}$, which equates $G'WG=G'W\Omega WG$. In this case the asymptotic variance is reduced to $(G'\Omega^{-1}G)^{-1}$.

Alternatively, a set of over-identified moment conditions $Em(z;\theta)\equiv 0$ can be translated into a set of exactly identified moment conditions by a linear combination matrix $AEm(z;\theta)\equiv 0$. Given A, the resulting method of moment estimator that equates $A\Sigma_{t=1}^{T} m(z_t;\hat{\theta})$ to zero has asymptotic variance $(AG)^{-1}(A\Omega A')(G'A')^{-1}$. As a rule of thumb, the optimal choice of A should simplify this asymptotic variance, by equating $AG = A\Omega A' = G'A'$. The resulting optimal $A = G'\Omega^{-1}$ gives rise to the same asymptotic distribution as the above optimally weighted GMM estimator of Hansen (1982), which minimizes

$$\frac{1}{T}\sum_{t=1}^{T} m(z_t;\hat{\theta})\, \Omega^{-1} \frac{1}{T}\sum_{t=1}^{T} m(z_t;\hat{\theta}).$$

Many economic models, such as those based on Euler equations, are stated in terms of conditional moment conditions of the form $E(m(z;\beta)|x) = 0$ for almost all x. These conditional moment conditions can be translated into exactly identified unconditional moment conditions using an instrument matrix $A(x)$: $EA(x)m(z;\beta) = 0$. The question arises as to what is the optimal instrument matrix $A(x)$. For a given choice of $A(x)$, the resulting method of moment estimator that equates $\frac{1}{T}\Sigma_{t=1}^{T}(x_t)m(z_t;\beta) = 0$ has asymptotic variance $(EA(x)G(x))^{-1}EA(x)\Omega(x)A(x)'(EG(x)'A(x)')^{-1}$, where $G(x) = E(\frac{\partial}{\partial\theta}m(z;\theta)|x)$ and $\Omega(x)=Var(m(z;\beta)|x)$. We can then equate

$$EA(x)G(x) = EA(x)\Omega(x)A(x)'$$

to obtain the optimal instrument matrix $A(x) = G(x)'\Omega(x)^{-1}$. The resulting efficient asymptotic variance is therefore $(EG(x)'\Omega(x)^{-1}G(x))^{-1}$.

Formal proofs of these derivations can be found in, for example, Newey and McFadden (1994). Estimators that achieve these efficiency bounds typically involve two-step or multi-step procedures and possibly nonparametric methods, such as Newey and Powell (1990).

Asymptotic efficiency in semiparametric models

Semiparametric models are extensions of parametric models where some components are specified nonparametrically with unknown functional forms. Generalized method of moment models are semiparametric models if the data-generating process is not fully specified. A partial linear model is another example. Other popular semiparametric models are surveyed in Powell (1994).

Intuitively, the variance of an estimator for a semiparametric model should be larger than the Cramer–Rao lower bound for any parametric sub-model that satisfies the semiparametric restrictions. The semiparametric efficiency bound is therefore defined to be the supremum of the Cramer–Rao bounds for all parametric models that satisfy the semiparametric restrictions. Extensive results for semiparametric efficiency bounds are developed in, among others, Bickel et al. (1993) and Newey (1990). In this section we give a brief summary of some of the results presented in Newey (1990). The next section will apply these results to a particular estimation problem.

Because of pathological cases such as the super-efficient estimator, the semiparametric efficiency bound is used to provide a lower bound only for *regular* estimators. Consider a parameter of interest that is a smooth function of the underlying parametric path: $\beta(\theta)$. A regular estimator $\hat{\beta}$ is one where for each θ_0 the limiting distribution of $\sqrt{T}(\hat{\beta} - \beta(\theta_T))$ does not depend on θ_T as long as $\sqrt{T}(\theta_T - \theta_0)$ is bounded. The super-efficient estimator is not regular.

Most estimators in econometrics are asymptotically linear, in the sense that they have an influence function representation as

$$\sqrt{T}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi(z_t) + o_p(1).$$

In particular, almost all econometric estimators asymptotically solve some moment conditions $\frac{1}{\sqrt{T}} \Sigma_{t=1}^{T} m(z_t; \hat{\beta}) = o_p(1)$, in which case the linear influence function is given by $\psi(z_t) = -G^{-1} m(z_t; \beta)$ for $G = E\frac{\partial}{\partial \beta} m(z_t; \beta)$.

Asymptotically linear estimators are regular if and only if for all parametric sub-models $\frac{\partial}{\partial \theta}\beta(\theta) = E\psi S_\theta'$. When $\psi(z_t) = -G^{-1} m(z_t; \beta)$, this follows from differentiating $E_\theta m(z; \beta(\theta)) = 0$ with respect to θ. The asymptotic variance of an asymptotically linear estimator is $E\psi\psi'$, which is apparently larger than that of the maximum likelihood estimator $\beta(\hat{\theta})$ of any parametric sub-model, which is given through information matrix and the delta method as

$$\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)(E(S_\theta S_\theta'))^{-1}\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)' = E[\psi S_\theta'](ES_\theta S_\theta')^{-1}E[S_\theta \psi'].$$

A starting point for calculating the semiparametric efficiency bound is to restrict attention to differentiable parameters $\beta(\theta)$ which satisfies $\frac{\partial \beta(\theta)}{\partial \theta} = E(dS_\theta')$ for some *d* and all parametric sub-models. Such *d* are not unique. Adding a random vector that is orthogonal to S_θ preserves the validity of *d*. In fact, any linear influence function ψ can serve as a *d*. For differentiable parameters, if we use the invariance principle and the delta method, the Cramer–Rao lower bound for estimating $\beta(\theta)$ is

$$\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)(E(S_\theta S_\theta'))^{-1}\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)' = E[dS_\theta'](ES_\theta S_\theta')^{-1}E[S_\theta d'].$$

Obviously, this is the variance of $d_\theta = E[dS_\theta](E[S_\theta S_\theta])^{-1}S_\theta$, which is the projection of *d* onto the linear space spanned by the score functions S_θ. As the class of parametric sub-models expands, the linear space it spans also increases and the variance of d_θ also increases. The semiparametric efficiency bound should be the limit of this progress of increments. Formally, the tangent space is defined to be the mean square closure of all linear combinations of scores S_θ for smooth parametric sub-models, and the efficiency bound is given by the variance of the projection of *d* onto the tangent space *T*. In other words, the efficiency bound is given by $V = E[\delta\delta']$ where $\delta \in T$ and $E[(d-\delta)'\iota] = 0$ for all $\iota \in T$.

Application

In this section we illustrate the computation of semiparametric efficiency bound using a model of non-classical measurement errors, studied in Chen, Hong and Tamer (2005) and Chen, Hong and Tarozzi (2004), where information from a primary data-set and from an auxiliary data-set need to be efficiently combined. Their models extend the results in the treatment effect literature on the mean parameter (see Hahn, 1998, Hirano, Imbens and Ridder, 2003 and Imbens, Newey and Ridder, 2005), to measurement error models where parameters are generically defined through nonlinear moment conditions. Consider the following model. The researcher is interested in a parameter β defined by the moment condition $Em(Y;\beta) = 0$ if and only if $\beta = \beta_0$. The researcher has access to a primary data-set which is a random sample from the population of interest. However, the true variable *Y* is not always observed in the primary data-set. Instead, a proxy variable *X* is observed throughout the primary data. For a subset of the primary data-set, which we will call the auxiliary data-set, *X* is validated so that both *Y* and *X* are observed. We will use the random variable *D*=0 to denote observations in the auxiliary data-set where both *X* and *Y* are observed, and will use *D*=1 to denote the rest of the primary data-set where only *X* is observed. Chen, Hong and Tarozzi (2004) call this the 'verify-in-sample' case. They make the following conditional independence assumption:

Assumption 4.1: $Y \perp D|X$.

Under this assumption, we follow the framework of Newey (1990) to show that the efficiency bound for estimating β is given by $\left(J_\beta \Omega_\beta^{-1} J_\beta\right)^{-1}$, where for $p(X) = p(D=1|X)$:

$$J_\beta = \frac{\partial}{\partial \beta}E[m(Y;\beta)] \text{ and } \Omega_\beta = E\left[\frac{1}{1-p(X)}V[m(Y;\beta)|X] + \xi(X;\beta)\xi(X;\beta)'\right].$$

To demonstrate this result, we follow the steps in the efficiency framework of Newey (1990). First we characterize the properties of the tangent space under assumption 4.1. Next we write the parameter of interest in its differential form and therefore find a linear influence function *d*. Finally, we conjecture and verify the projection of *d* onto the tangent space and the variance of this projection gives rise to the efficiency bound. We first go through these three steps under the assumption that the moment conditions exactly identify β. Finally, the results are extended to over-identified moment conditions by considering their optimal linear combinations.

First we assume that the moment conditions exactly identify β.

Step 1. Consider a parametric path θ of the joint distribution of *Y*, *X* and *D*. Define $p_\theta(x) = P_\theta(D = 1|x)$. Under assumption 1, the joint density function for *Y*, *D* and *X* can be factorized into

$$f_\theta(y, x, d) = f_\theta(x)\, p_\theta(x)^d [1 - p_\theta(x)]^{1-d} f_\theta(y|x)^{1-d}. \tag{1}$$

The resulting score function is then given by

$$S_\theta(d, y, x) = (1 - d) s_\theta(y|x) + \frac{d - p_\theta(x)}{p_\theta(x)(1 - p_\theta(x))} \dot{p}_\theta(x) + t_\theta(x),$$

where

$$s_\theta(y|x) = \frac{\partial}{\partial \theta} \log f_\theta(y|x),\ \dot{p}_\theta(x) = \frac{\partial}{\partial \theta} p_\theta(x),\ t_\theta(x) = \frac{\partial}{\partial \theta} \log f_\theta(x).$$

The tangent space of this model is therefore given by:

$$\mathcal{T} = \{(1 - d) s_\theta(y|x) + a(x)(d - p_\theta(x)) + t_\theta(x)\} \tag{2}$$

where $\int s_\theta(y|x) f_\theta(y|x)\,dy = 0$, $\int t_\theta(x) f_\theta(x)\,dx = 0$, and $a(x)$ is any square integrable function.

Step 2. As in the method of moment model in Newey (1990), the differential form of the parameter β can be written as

$$\frac{\partial \beta(\theta)}{\partial \theta} = -(\mathcal{J}_\beta)^{-1} E\left[m(Y;\beta)\frac{\partial \log f_\theta(Y,X)}{\partial \theta'}\right] = -(\mathcal{J}_\beta)^{-1}\left\{E\left[m(Y;\beta)(s_\theta(Y|X)' + t_\theta(X)')\right]\right\} = -(\mathcal{J}_\beta)^{-1}\left\{E\left[m(Y;\beta)s_\theta(Y|X)'\right] + E[\xi(X)t_\theta(X)']\right\}. \tag{3}$$

Therefore $d = -\mathcal{J}_\beta^{-1} m(Y;\beta)$. Since $\mathcal{J}_\beta$ is only a constant matrix of nonsingular transformation. The projection of d onto the tangent space will be $-\mathcal{J}_\beta$ multiplied by the projection of $m(Y;\beta)$ onto the tangent space. Therefore we only need to consider the projection of $m(Y;\beta)$ onto the tangent space.

Step 3. We conjecture that this projection takes the form of

$$\tau(Y,X,D) = \frac{1-D}{1-p(X)}[m(Y;\beta) - \xi(X)] + \xi(X)$$

To verify that this is the efficient influence function we need to check that $\tau(Y,X,D)$ lies in the tangent space and that

$$E[(m(Y;\beta) - \tau(Y,X,D))s_\theta(Y,X)] = 0.$$

or that

$$E[m(Y;\beta)s_\theta(Y,X)] = E[\tau(Y,X,D)s_\theta(Y,X)]. \tag{4}$$

To see that $\tau(Y,X,D)$ lies in the tangent space, note that the first term in $\tau(Y,X,D)$ has mean zero conditional on X, and corresponds to the first term of $(1-d)s_\theta(y|x)$ in the tangent space. The second term in $\tau(Y,X,D)$, $\xi(x)$, has unconditional mean zero and obviously corresponds to the $t_\theta(x)$ in the tangent space.

To verify (4), one can make use of the representation of $E[m(Y;\beta)s_\theta(Y,X)]$ in (3), by verifying the two terms in $\tau(Y,X,D)$ separately. The second term is obvious and tautological. The first part,

$$E\left[\frac{1-D}{1-p(X)}[m(Y;\beta)-E(X)]s_\theta(Y,X)\right]=E[m(Y;\beta)s_\theta(Y,X)],$$

follows from the conditional independence assumption 4.1 and the score function property $E[s\theta(Y,X)|X]=0$. Therefore we have verified that $\tau(Y,X,D)$ is the efficient projection and that the efficiency bound is given by

$$V=(\mathcal{J}_\beta)^{-1}E\left[\tau(Y,X,D)\tau(Y,X,D)'\right](J_\beta)'^{-1}=(\mathcal{J}_\beta)^{-1}E\left[\frac{1}{1-p(X)}Var(m(Y;\beta)|X)+\xi(X)\xi(X)'\right](\mathcal{J}_\beta)'^{-1}.$$

Finally, consider the extensions of these results to the over-identified case. When $d_m>d_\beta$, the moment condition is equivalent to the requirement that for any matrix A of dimension $d_\beta\times d_m$ the following exactly identified system of moment conditions holds

$$\mathcal{A}E[m(Y;\beta)]=0.$$

Differentiating under the integral again, we have

$$\frac{\partial\beta(\theta)}{\partial\theta}=-\left(\mathcal{A}E\left[\frac{\partial m(Y;\beta)}{\partial\beta}\right]\right)^{-1}E\left[\mathcal{A}m(Y;\beta)\frac{\partial\log f_\theta(Y,X|D=1)}{\partial\theta'}\right].$$

Therefore, any regular estimator for β will be asymptotically linear with influence function of the form

$$-\left(\mathcal{A}E\left[\frac{\partial m(Y;\beta)}{\partial\beta}\right]\right)^{-1}\mathcal{A}m(Y;\beta).$$

For a given matrix $\mathcal{A}$, the projection of the above influence function onto the tangent set follows from the previous calculations, and is given by

$$-[\mathcal{A}J_\beta]^{-1}\mathcal{A}\tau(Y, X, d).$$

The asymptotic variance corresponding to this efficient influence function for fixed $\mathcal{A}$ is therefore

$$[\mathcal{A}J_\beta]^{-1}\mathcal{A}\Omega\mathcal{A}'[J_\beta\mathcal{A}']^{-1} \tag{5}$$

where

$$\Omega = E[\tau(Y, X, D)\tau(Y, X, D)']$$

as calculated above. Therefore, the efficient influence function is obtained when $\mathcal{A}$ is chosen to minimize this efficient variance. It is easy to show that the optimal choice of $\mathcal{A}$ is equal to $J_\beta'\Omega^{-1}$, so that the asymptotic variance becomes

$$V = \left(\mathcal{J}_\beta'\Omega^{-1}\mathcal{J}_\beta\right)^{-1}.$$

Different estimation methods can be used to achieve this semiparametric efficiency bound. In particular, Chen, Hong and Tarozzi (2004) showed that both a semiparametric conditional expectation projection estimator and a semiparametric propensity score estimator based on a sieve nonparametric first-stage regression achieve this efficiency bound.

Conclusion

As discussed in Newey (1990), while the calculation of the tangent space and the efficient projection is easy in several important examples, including the one above, it can be difficult in general. A variety of techniques are available to characterize the tangent space and the efficient projection. Some of these are discussed in details in Newey (1990) and Bickel et al. (1993).

Even in parametric models, the notion of asymptotic efficiency is more complex when one compares estimators that do not converge at $\sqrt{n}$ rate or are not asymptotically distributed. Comparing these estimators requires the choice of a loss function, and different loss functions can lead to different efficiency rankings (see Ibragimov and Has'minskii, 1981). In econometrics, these estimators sometimes arise in structural models in labour economics and in industrial organization. The efficiency properties of these estimators are analysed in Hirano and Porter (2003) and Chernozhukov and Hong (2004).

See Also

- generalized method of moments estimation
- maximum likelihood
- measurement error models
- nonparametric structural models
- semiparametric estimation
- stratification

Bibliography

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Amemiya, T. 1994. *Introduction to Statistics and Econometrics*. Cambridge, MA: Harvard University Press.

Bickel, P., Klaassen, C.A., Ritov, Y. and Wellner, J. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.

Chen, X., Hong, H. and Tamer, E. 2005. Measurement error models with auxiliary data. *Review of Economic Studies* 72, 343–66.

Chen, X., Hong, H. and Tarozzi, A. 2004. Semiparametric efficiency in GMM models of nonclassical measurement errors. Working paper, Duke University and New York University.

Chernozhukov, V. and Hong, H. 2004. Likelihood inference for a class of nonregular econometric models. *Econometrica* 72, 1445–80.

Hahn, J. 1998. On the Role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66, 315–32.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Hirano, K., Imbens, G. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–89.

Hirano, K. and Porter, J. 2003. Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* 71, 1307–38.

Ibragimov, I. and Has'minskii, R. 1981. *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag.

Imbens, G., Newey, W. and Ridder, G. 2005. Mean-squared-error calculations for average treatment effects. Working paper.

Newey, W. 1990. Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5(2), 99–135.

Newey, W. and McFadden, D. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

Newey, W. and Powell, J. 1990. Efficient estimation of linear and type in censored regression models under conditional quantile restrictions. *Econometric Theory* 6, 295–317.

Powell, J. 1994. Estimation of semiparametric models. In *Handbook of Econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

van der Vaart, A. 1999. *Asymptotic Statistics*. Cambridge: Cambridge University Press.

How to cite this article

Hong, Han. "efficiency bounds." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000251> doi:10.1057/9780230226203.0451

The New Palgrave Dictionary of Economics Online

efficiency wages

Andrew Weiss

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Efficiency wages capture the effect of compensation on the behaviour of workers, as well as on the quality of workers attracted and retained by the firm. This effect has greater significance in some areas than others, and can be used to explain wage differentials among firms and industries, as well as to explain why firms respond to demand shocks by reducing their labour force rather than cutting wages, and may ration jobs even in normal times. At the macroeconomic level efficiency wages can explain persistent long-term unemployment as an equilibrium outcome in a competitive labour market.

Keywords

capital cost; capital market imperfections; efficiency wages; firm size; firm-specific human capital; gift exchange; involuntary unemployment; labour supply; layoffs; low-wage probation period; monitoring; nutrition models; productivity; retirement; sorting effect of wages; wage differentials

Article

‘Efficiency wages’ is a term used to express the idea that labour costs can be described in terms of efficiency units of labour rather than in terms of hours worked, and that wages affect the performance of workers. In this respect, labour differs from most other inputs (with the notable exception of credit), in which inputs are well defined independently of prices. Models of efficiency wages explore the implications of the interconnections between compensation and productivity. On the macroeconomic level, efficiency wages can explain persistent unemployment without relying on either structural imperfections such as search costs or fixed-length contracts or irrational behaviour such as money illusion, which would cause real wages to fail to adjust to market conditions. (For some of the earliest such models, see Futia, 1977; Salop, 1979; Solow, 1979; Shapiro and Stiglitz, 1984; Weiss, 1981.) At the level of the firm, efficiency wages can result in job queues (excess supply of labour) and can explain why seemingly identical workers may receive different wages at different firms, and why these observed

wage differentials are positively correlated with firm characteristics such as profitability, high capital–labour ratios, and establishment size (Brown and Medoff, 1989). These market imperfections arise because employers cannot costlessly observe the ability and productivity of workers or because of capital market imperfections that prevent workers from 'buying' the high-wage jobs.
Efficiency wage models have one or more of the following characteristics:

1. 1. Compensation levels and rules affect the types of workers who are attracted to, and retained by, the firm – this is normally referred to as the sorting effect of wages.
2. 2. Compensation rules create incentives for workers to behave in ways that increase firm profits.
3. 3. Wages affect the nutrition and health of workers and thus higher wages directly increase productivity (these 'nutrition' models are most applicable in poor countries).

Consequences of the use of efficiency wages are:

1. 1. Compensation levels within a firm may not be proportionate to relative productivity.
2. 2. Compensation could be a function of characteristics of the establishment employing the worker.
3. 3. Wages could rise more steeply with tenure than does productivity.
4. 4. Some firms could have an excess supply of workers.
5. 5. A frictionless economy could be in a long-run equilibrium with unemployment.

The sorting effects of wages enable a firm to benefit from private information that the employee knows about himself and that is either not available to the firm or would be costly for the firm to acquire. High compensation enables the firm to draw from a larger and better pool of workers. Firms that test job applicants will also find that, by offering a higher wage, the expected quality of the worker hired, conditional on the applicants test score, will also be higher.
The test could be in the form of a low-wage probation period for new hires. Using a low-wage probation period, followed by a significant wage increase, followed by high wages for workers who perform well during the probation period, the firm can attract job applicants with positive private information about their ability. If the test is imperfect, the use of a low-wage probation period will also discourage applications from risk-averse applicants as well as applicants with a higher cost of capital. Wages that increase steeply with tenure will attract workers who have low quit propensities (aside from their incentive effect of deterring quits). Groshen and Loh (1993) have found that much of the return to tenure takes place at the end of low-wage probationary periods.
Sorting effects of efficiency wages may also explain why firms do not cut wages in response to a fall in demand. If a firm were to cut the wages, it may find that its better workers are most likely to quit. Thus, a profit maximizing firm could find that its best response to a fall in demand for its product would be to fire workers rather than to cut wages.
Most of the efficiency wage models have focused on the ways in which compensation affects the behaviour of workers.
The incentive effects of wages stem from the effect of the level of compensation on the cost to the worker of being fired. Thus, wages above the market clearing level will increase effort, decrease

employee theft, decrease absenteeism, and decrease quits. See, for example, Salop and Salop (1976), Klein, Spady and Weiss (1991), and Weiss (1984) on quits; Shapiro and Stiglitz (1984) on effort; Lazear and Rosen (1981), Weiss (1985) on absenteeism.

Levels of compensation also affect the attitude of the employee towards the firm. Thus, paying wages above the market clearing level may have multiple beneficial effects for the firm including: reducing employee theft, increasing unobserved effort, and inducing higher levels of care, which will decrease costs incurred from damage to the firm's property. Greater loyalty to the firm will also encourage workers to acquire firm-specific human capital, to report theft of firm property, and to allocate the worker's effort in ways that benefit the firm. See Akerlof (1984) on gift exchange.

Higher levels of compensation will also reduce the time needed to fill vacancies (Lang, 1991). In this case the behaviour being affected is the application process.

Wages directly affect the productivity of workers through their effect on the nutrition of workers as well as their access to clean water and medical care and other goods and services that directly improve their productivity. These ‘nutrition’ effects are strongest in poor countries and could also possibly explain poverty traps for particularly poor workers who do not have access to firms that are offering efficiency wages.

The importance of these effects will vary across firms. For instance, we would expect that capital-intensive firms will derive the greatest benefit from reductions in absenteeism and quits, and from increased productivity of their employees. Capital-intensive firms will also tend to be most vulnerable to careless behaviour by workers that would damage the valuable property. Larger firms have more difficulty monitoring individual effort and directing the effort in ways that fit the needs of the firm. Consequently, the efficiency wage models would predict that compensation would be correlated with firm size. The direct effects of wages through better nutrition and health take some time to affect productivity, so we would expect that firms with lower costs of capital will offer higher wages – in poor countries these tend to be foreign firms. (In poor countries, in which the nutrition effects are strongest, we might see that wages would be correlated with a firm's cost of capital as well as with the ability of the firm to retain workers after their productivity has been enhanced by the higher wages. The nutrition effects of wages may take some time to affect productivity.) Finally, if high wages are used to attract better workers, then we would expect that when workers are laid off from firms in high-wage industries they will tend to get jobs in other high-wage industries (see Gibbons and Katz, 1992).

All of these implications of the efficiency wage model have been confirmed by empirical studies of the relationship between firm characteristics and wages. (In cases in which wages directly affects productivity we would expect that firms that are likely to be able to retain their workers will also pay higher wages. However, since wages directly affects turnover, and prices vary according to the presence of competitive firms, this implication of the nutrition version of the efficiency wage model is more difficult to verify.) Of course, many if not all of these empirical findings can be explained by other models. For example, the relationship between prior and posterior industry wages for laid-off workers can be explained by competitive models in which workers are being selected based on attributes, such as pulchritude, that are directly observed by the firm but not by the researchers.

Thus, efficiency wages can explain why empirical studies of the relationship between wage and characteristics of establishments find that large, capital-intensive establishments are most likely to pay wages that are above market clearing levels – and in the case of poor countries why foreign firms tend to

pay higher wages. The efficiency wage models also can explain why firms fire workers rather than cutting wages, offer wages that attract an excess supply of workers, and pay some of their workers to take early retirement or seek to impose mandatory retirement. See, for instance, Brown and Medoff (1989). Finally, efficiency wage theory can explain the persistence of involuntary unemployment in a free market economy.

Bibliography

Akerlof, G.A. 1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97, 543–69.

Akerlof, G.A. 1984. Gift exchange and efficiency-wage theory: four views. *American Economic Review* 74(2), 79–83.

Brown, C. and Medoff, J. 1989. The employer size-wage effect. *Journal of Political Economy* 97, 1027–59.

Futia, C. 1977. Excess supply equilibria. *Journal of Economic Theory* 14, 200–20.

Gibbons, R. and Katz, L. 1992. Does unmeasured ability explain inter-industry wage differentials? *Review of Economic Studies* 59, 515–35.

Groshen, E. and Loh, E.S. 1993. What do we know about probationary periods? Proceedings of the 45th Annual Meeting of the Industrial Relations Research Association, Madison, WI.

Klein, R., Spady, R. and Weiss, A. 1991. Factors affecting the output and quit propensities of production workers. *Review of Economic Studies* 58, 929–54.

Landau, H. and Weiss, A. 1984. Wages, hiring standards and firm size. *Journal of Labor Economics* 2, 477–99.

Lang, K. 1991. Persistent wage dispersion and involuntary unemployment. *Quarterly Journal of Economics* 106, 181–202.

Lazear, E. and Rosen, S. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89, 841–64.

Salop, J. and Salop, S. 1976. Self-selection and turnover in the labor market. *Quarterly Journal of Economics* 90, 619–27.

Salop, S. 1979. A model of the natural rate of unemployment. *American Economic Review* 69, 117–25.

Shapiro, C. and Stiglitz, J. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Solow, R. 1979. Another possible source of wage stickiness. *Journal of Macroeconomics* 1, 79–82.

Weiss, A. 1981. Job queues and layoffs in labor markets with flexible wages. *Journal of Political Economy* 88, 526–38.

Weiss, A. 1984. Determinants of quit behavior. *Journal of Labor Economics* 2, 371–87.

Weiss, A. 1985. Absenteeism and wages. *Economics Letters* 19, 277–9.

Weiss, A. and Wang, R. 1998. Probation, layoffs, and wage-tenure profiles: a sorting explanation. *Labour Economics* 5, 359–83.

How to cite this article

Weiss, Andrew. "efficiency wages." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000245> doi:10.1057/9780230226203.0452

The New Palgrave Dictionary of Economics Online

efficient allocation

Stanley Reiter
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Analysis of efficiency in the context of resource allocation has been a central concern of economic theory from ancient times, and is an essential element of modern microeconomic theory. The ends of economic action are seen to be the satisfaction of human wants through the provision of goods and services. These are supplied by production and exchange and limited by scarcity of resources and technology. In this context efficiency means going as far as possible in the satisfaction of wants within resource and technological constraints. This is expressed by the concept of Pareto optimality, which can be stated informally as follows: a state of affairs is Pareto optimal if it is within the given constraints and it is not the case that everyone can be made better off in his own view by changing to another state of affairs that satisfies the applicable constraints.

Because knowledge about wants, resources and technology is dispersed, efficient outcomes can be achieved only by coordination of economic activity. Hayek (1945) pointed out the role of knowledge or information, particularly in the context of prices and markets, in coordinating economic activity. Acquiring, processing and transmitting information are costly activities themselves subject to constraints imposed by technological and resource limitations. Hayek pointed out that the institutions of markets and prices function to communicate information dispersed among economic agents so as to bring about coordinated economic action. He also drew attention to motivational properties of those institutions, or incentives. In this context, the concept of efficiency takes account of the organizational constraints on information processing and transmission in addition to those on production of ordinary goods and services. The magnitude of resources devoted to business or governmental bureaucracies, and to some of the functions performed by industrial salesmen, attests to the importance of these constraints. Economic analysis of efficient allocation has formally imposed only the constraints on production and exchange, and until recently recognized organizational constraints only in an informal way. But it is these constraints that motivate the pervasive and enduring interest in decentralized modes of economic organization, particularly the competitive mechanism.

It is necessary to limit the scope of this essay so that it is not coextensive with microeconomic theory. The main limitation imposed here is to confine attention to models in which either the role of information is ignored, or in which agents do not behave strategically on the basis of private information. In so doing, a large and important class of models involving problems of efficient allocation in the presence of incentive constraints is excluded.

The main ideas of efficient resource allocation are present in their simplest form in the linear activity analysis model of production. We begin with that model.

Efficiency of Production: Linear Activity Analysis

The analysis of production can to some extent be separated from that of other economic activity. The concept of efficiency appropriate to this analysis descends from that of Pareto optimality, which refers to both productive and allocative efficiency in the full economy in which production is embedded. It is useful to begin with a model in which technological possibilities afford constant returns to scale, that is, with the (linear) activity analysis model of production pioneered by Koopmans (1951a, 1951b, 1957), and closely related to the development of linear programming associated with Dantzig (1951a, 1951b) and independently with the Russian mathematician Kantorovitch (1939, 1942) and Kantorovitch and Gavurin (1949).

The two primitive concepts of the model are *commodity* and *activity*. A list of n commodities is postulated; a commodity *bundle* is given by specifying a sequence of n numbers a_1, a_2, …, a_n. Technological possibilities are thought of as knowledge of how to transform commodities. Such knowledge may be described in terms of collections of activities called *processes*, much as knowledge of how to prepare food is described by recipes. A recipe commonly has two parts, a list of ingredients or inputs and of the output(s) of the recipe, and a description of how the ingredients are to be combined to produce the output(s). In the activity analysis model the description of productive activity is suppressed. Only the

specification of inputs and outputs is retained; this defines the production process.

Commodities are classified into ‘desired’, ‘primary’ and ‘intermediate’ commodities. Desired commodities are those whose consumption or availability is the recognized goal of production; they satisfy wants. Primary commodities are those available from nature. (A primary commodity that is also desired is listed separately among the desired commodities and must be transformed by an act of production into its desired form.) Intermediate commodities are those that merely pass from one stage of production to another. Each commodity can exist in any non-negative amount *(divisibility)*. Addition and subtraction of the numbers measuring the amount of a commodity represent joining and separating corresponding amounts of the commodity.

An activity is characterized by a *net output number* for each commodity, which is positive if the commodity is a net output, negative if it is a net input and zero if it is neither. The term *input-output vector* is also used for this ordered array of numbers. Activity analysis postulates a finite number of basic activities from which all technologically possible activities can be generated by suitable combination. Allowable combinations are as follows. If two activities are known to be possible, then the activity given by their algebraic sum is also possible, i.e. if $a=(a_1, a_2,\ldots,a_n)$ and $b=(b_1, b_2,\ldots,b_n)$, then $a+b=(a_1+b_1, a_2+b_2,\ldots,a_n+b_n)$ is also possible. Thus, additivity embodies an assumption of non-interaction between productive activities, at least at the level of knowledge. Furthermore, if an activity is possible, then so is every non-negative multiple of it (*proportionality*), i.e. if $a=(a_1, a_2, \ldots,a_n)$ is possible, then so is $\mu a=(\mu a_1, \mu a_2,\ldots,\mu a_n)$ for any non-negative real number μ. This expresses the assumption of constant returns to scale. The family of activities consisting of all non-negative multiples of a given one forms a process. Since there is a finite number of basic activities, there is also a finite number of basic processes, each intended to describe a basic method of production capable of being carried out at different levels, or intensities.

The assumptions of additivity and proportionality determine a linear model of technology that can be given the following form. Let A be an n by k matrix whose jth column is the input-output vector representing the basic activity that defines the jth basic process, and let $x=(x_1, x_2,\ldots,x_n)$ be the vector whose jth component x_j is the scale (level or intensity) of the jth basic process. Let $y=(y_1, y_2,\ldots,y_n)$ be the vector of commodities. Technology is represented by a linear transformation mapping the space of activity levels into the commodity space, i.e.

$$y = Ax x \geq 0.$$

With the properties assumed, a process can be represented geometrically in the commodity space by a halfline from the origin including all non-negative multiples of some activity in that process. The finite number of halflines representing basic processes generate a convex polyhedral cone consisting of all activities that can be expressed as sums of activities in the basic processes, or equivalently, as non-negative linear combinations of the basic activities, sometimes called a *bundle of basic activities*. This cone is called the *production set*, or set of *possible productions*.

Two other assumptions are made about the production set itself, rather than just the individual activities. First, there is no activity, whether basic or derived, in the production set with a positive net output of some commodity and non-negative net outputs of all commodities. This excludes the possibility of producing something from nothing, whether directly or indirectly. Second, it is assumed that the production set contains at least one activity with a positive net output of some commodity.

If the availability of primary commodities is subject to a bound, the technologically possible productions described by the production set are subject to another restriction; only those possible productions that do not require primary inputs in amounts exceeding the given bounds can be produced. Furthermore, because intermediate commodities are not desired in themselves, their net output is required to be zero. (Strictly speaking, the technological constraint on intermediate commodities is that their net output be non-negative. The requirement that they be zero can be viewed as one of elementary efficiency, excluding accumulation or necessity to dispose of unwanted goods.) With these restrictions the model can be written

$$y = Ax, \quad x \geq 0, \quad y_i = 0$$

if i is an intermediate commodity, and

$$y_i \geq r_i$$

if i is a primary commodity, where r_i is the (non-positive) limit on the availability of primary commodity i. This leads to the concept of an *attainable* activity.

A bundle of basic activities is *attainable* if the resulting net outputs are non-negative for all desired commodities, zero for intermediate commodities and non-positive for primary commodities, and if the total inputs of primary commodities do not exceed (in absolute amount) the prescribed bounds of availability of those commodities. The set of activities satisfying these conditions is a truncated convex polyhedral cone in the commodity space called the *set of attainable productions*.

The concept of productive efficiency in this model is as follows. An activity (a bundle of basic activities) is *efficient* if it is attainable and if every activity that provides more of some desired commodity and no less of any other is not attainable.

This concept can be seen to be a specialization of Pareto optimality. If for each desired commodity there is at least one consumer who is not satiated in that commodity, at least in the range of production attainable within the given resource limitations, then increasing the amount of any desired commodity without decreasing any other can improve the state of some non-satiated consumer without worsening that of any other.

Characterizing efficient production in terms of prices

Efficient production can be characterized in terms of *implicit prices*, also called *shadow prices*, or in the context of linear programming, *dual variables*. Efficient activities are precisely those that maximize profit for suitably chosen prices. The profit returned by a process carried out at the level x is

$$x\sum_i p_i a_i,$$

where the prices are $p=(p_1,\ldots,p_n)$, and $a=(a_1,\ldots,a_n)$ is the basic activity defining the process; the profit on the bundle of activities Ax at prices p is given by the inner product $py=pAx$.

This characterization is the economic expression of an important mathematical fact about convex sets in n–1 dimensional Euclidean space, namely that through every point of the space not interior to the convex set in question there passes a hyperplane that contains the set in one of its two halfspaces (Fenchel, 1950; Nikaido, 1969, 1970). (A hyperplane in n dimensional space is a level set of a linear function of n variables, and thus is a translate of an n–1 dimensional linear subspace. A hyperplane is given by an equation of the form $c_1x_1+c_2x_2+\cdots+c_nx_n=k$, where the x's are variables, the c's are coefficients defining the linear function and k is a constant identifying the level set. A hyperplane divides the space into two halfspaces corresponding to the two inequalities $c_1x_1+c_2x_2+\cdots+c_nx_n \gtreqless k$ respectively.) It can also be seen that a point of a convex set is a boundary point if and only if it maximizes a linear function on the (closure of the) set. These facts can be used to characterize efficient production because the attainable production set is convex and efficient activities are boundary points of it. Because the efficient points are those, roughly speaking, on the ‘north-east’ frontier of the set, the linear functions associated with them have non-negative coefficients, interpreted as prices. On the other hand, if a point of the attainable set maximizes a linear function with strictly positive coefficients (prices), then it is on the ‘north-east’ frontier of the set.

In Figure 1 the set enclosed by the broken line and the axes is the projection of the attainable set on the output coordinates; inputs are not shown. The point y' in the figure is efficient; the point y' is not; both y' and y'' maximize a linear function with non-negative coefficients (the level set containing y' is labelled a and also contains y''). However, y' maximizes a linear function with positive coefficients (one such, whose level set through y' is labelled b, is shown), while y'' does not.

Figure 1

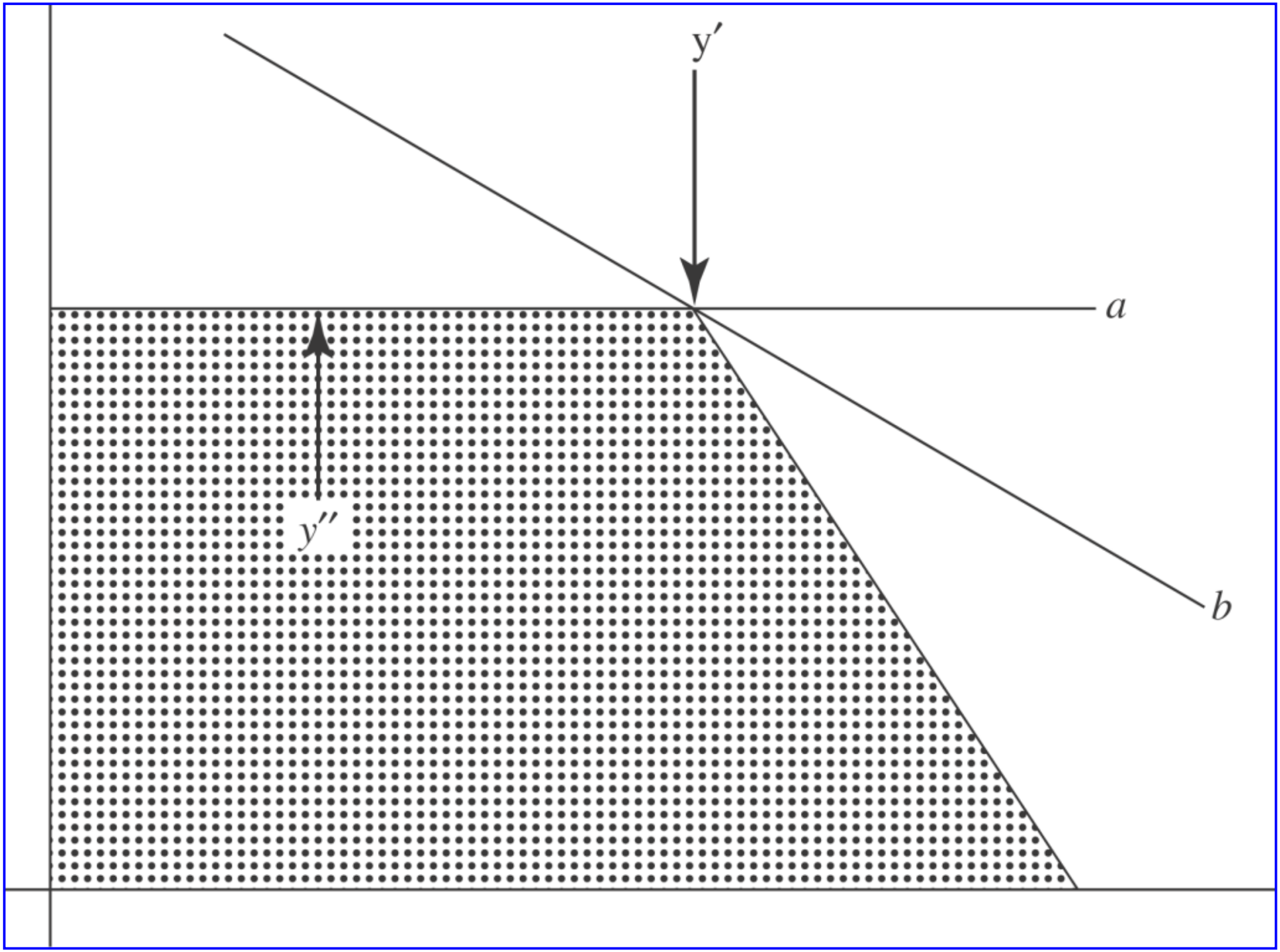

These implicit, or efficiency prices arise from the logic of efficiency or maximization when the relevant sets are convex, not from any institutions such as markets or exchange. An

important reason for interest in them is the possibility of achieving efficient performance by decentralized methods. As described above, under the assumptions of additivity and constant returns to scale the production set can be seen to be generated by a finite number of basic processes, each of which consists of the activities that are non-negative multiples of a basic activity, the multiple being the scale (level, or intensity) at which the process is operated. Following the presentation of Koopmans (1957), each basic process is controlled by a manager, who decides on its level. The manager of a process is assumed to know only the input-output coefficients of his process. Each primary resource is in the charge of a resource holder, who knows the limit of its availability. Efficiency prices are used to guide the choices of managers and resource holders. (Under constant returns to scale, if an activity yields positive profit at a given system of prices, then increasing the scale of the process containing that activity increases the profit. Since the scale can be increased without bound, if the profitability of a process is not zero or negative, then, in the eyes of its manager, who does not know the aggregate resource constraints, it can be made infinite. Therefore, the systems of prices that can be considered for the role of efficiency prices must be restricted to those *compatible with the given technology*, namely prices such that no process is profitable and at least one process breaks even.) Two propositions characterize efficient production by prices and provide the basis for an interpretation in terms of decentralized control of production.

> In a given linear activity analysis model, if there is a given system of prices compatible with the technology, in which the prices of all desired commodities are positive, then any attainable bundle of basic activities selected only from processes that break even and which utilizes all positively priced primary commodities to the limit of their availability and does not use negatively priced primary commodities at all, is an efficient bundle of activities.
>
> In a given linear activity analysis model, each efficient bundle of activities has associated with it at least one system of prices compatible with the technology such that every activity in that bundle breaks even and such that prices of desired commodities are positive, and the price of a primary commodity is non-negative, zero or non-positive, according as its available supply is full, partly, or not used at all (Koopmans, 1957).

These propositions are stated in a static form. There is no reference to managers raising or lowering the levels of the processes they control, or to resource holders adjusting prices. A dynamic counterpart of these propositions would be of interest, but because of the linearity of the model such dynamic adjustments are unstable (Samuelson, 1949).

It should also be noted that the concept of decentralization is not explicitly defined in this literature; the interpretation is by analogy with the competitive mechanism. Nevertheless, the interest in characterizing efficiency by prices and their interpretation in terms of decentralization is an important theme in the study of efficient resource allocation.

The linear activity analysis model has been generalized in several directions. These include dropping the assumption of proportionality, dropping the restriction to a finite number of basic activities, dropping the restriction to a finite number of commodities and dropping the restriction to a finite number of agents. Perhaps the most directly related generalization is to the nonlinear activity analysis, or nonlinear programming, model.

Efficiency of production: nonlinear programming

In the nonlinear programming model there is, as in the linear model, a finite number of basic processes. Their levels are represented by a vector $x=(x_1,\bullet x_2,\ldots,\bullet x_k)$, where k is the number of basic processes. Technology is represented by a nonlinear transformation from the space of process levels to the commodity space (still assumed to be finite dimensional), written

$$y = F(x),\ x \geq 0.$$

The production set in this model is the image in the commodity space of the non-negative orthant of the space of process levels. Under the assumptions usually made about F, the production set is convex, though, of course, not a polyhedral cone.

In this model as in the linear activity analysis model a central result is the characterization of efficient production in terms of prices. The simplest case to begin with is that of one desired commodity, say, one output, with perhaps several inputs. In this case the (vector-valued) function F can be written

$$F(x) = [f(x), g_1(x), g_2(x), \ldots, g_m(x)],$$

where the value of f is the output, and $g_1,\ldots,g_m$ correspond to the various inputs. Resource constraints are expressed by the conditions

$$g_j(x) \geq 0, \quad for\ j = 1, 2, \ldots, m,$$

and non-negativity of process levels by the condition, $x \geq 0$. (Here the resource constraints $r_j \leq h_j(x) \leq 0$ are written more compactly as $h_j(x) - r_j = g_j(x) \geq 0$.)
In this model the definition of efficient production given in the linear model amounts to maximizing the value of f subject to the resource and non-negativity constraints just mentioned.
Problems of constrained maximization are intimately related to saddle-point problems. Let L be a real valued function defined on the set $X \times Y$ in R^n. A point (x^*, y^*) in $X \times Y$ is a *saddle point* of L if

$$L(x, y^*) \leq L(x^*, y^*) \leq L(x^*, y),$$

for all x in X and all y in Y. The concept of a concave function is also needed. A real valued function f defined on a convex set X in R^n is a *concave function* if for all x and y in X and all real numbers $0 \leq a \leq 1$

$$f(ax + (1 - a)y) \geq af(x) + (1 - a)f(y).$$

The following mathematical theorem is fundamental.
Theorem: (Kuhn and Tucker, 1951; Uzawa, 1958): Let f and $g_1,\cdot g_2,\ldots,\cdot g_m$ be real valued concave functions defined on a convex set X in R^n. If f achieves a maximum on X subject to $g_j(x) \geq 0$, $j=1,\cdot 2,\ldots,\cdot m$ at the point x^* in X, then there exist non-negative numbers $p_0^*, p_0^*, \ldots, p_m^*$, not all zero, such that $p_0^* f(x) + p^* g(x) \leq p_0^* f(x^*)$ for all x in X, and furthermore, $p^* g(x^*) = 0$. (Here the vectors $p^* = (p_1^*, p_2^*, \ldots, p_m^*)$, and $g(x)=[g_1(x),\cdot g_2(x)\cdot,\ldots,\cdot g_m(x)]$) The vector p^* may be chosen so that

$$\sum_0^m p_j^* = 1.$$

An additional condition (Slater, 1950) is important. (It ensures that the coefficient p_0 of f is not zero.)
Slater's Condition: : There is a point x' in X at which $g_j(x')>0$ for all $j=1,\cdot 2,\ldots,\cdot m$.
If attention is restricted to concave functions, as in the Kuhn-Tucker-Uzawa Theorem, the relation between constrained maxima and saddle points can be summarized in the following theorem.
Theorem: : If f and g_j, $j=1,\cdot 2,\ldots,\cdot m$ are concave functions defined on a convex subset X in R^n, and if Slater's Condition is satisfied, then x^* in X maximizes f subject to $g_j(x) \geq 0$, $j=1,\cdot$

2•,…,•*m*, if and only if there exists $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*)$, $\lambda_j^* \geq 0$ for *j*=1,•2,…,•*m*, such that (x^*, λ^*) is a saddle point of $L(x,\bullet\lambda)=f(x)+\lambda g(x)$ on $X \times R_+^n$.
This theorem is easily seen to cover the case where some constraints are equalities, as in the case of intermediate commodities. The sufficiency half of this theorem holds for functions that are not concave.
The auxiliary variables $\lambda_1,\bullet\lambda_2\bullet,\ldots,\bullet\lambda_m$, called *Lagrange multipliers*, play the role of efficiency prices, or shadow prices; they evaluate the resources constrained by the condition $g(x)\geq 0$. The maximum characterized by the theorem is a global one, as in the case of linear activity analysis.
If the functions involved are differentiable, a saddle point of the Lagrangean can be studied in terms of first-order conditions. The first-order conditions are necessary conditions for a saddle point of *L*. If the functions *f* and the *g*'s are concave on a convex set *X*, then the first-order conditions at a point (x^*, λ^*) are also sufficient; that is, they imply that (x^*, λ^*) is a saddle point of *L*. Thus,
Theorem: : If *f*, $g_1,\bullet g_2,\ldots,\bullet g_m$ are concave and differentiable on an open convex set *X* in R^n, and if Slater's Condition is satisfied, then x^* maximizes *f* subject to $g_j(x)\geq 0$ for *j*=1,•2,…,•*m* if and only if there exists numbers $\lambda_1^*, \lambda_2^*, \ldots \lambda_m^*$ such that the first-order conditions for a saddle point of $L(x,\bullet\lambda)=f(x)+\lambda g(x)$ are satisfied at (x^*, λ^*).
If there are non-negativity conditions on the *x*'s,

$$g_j(x) \geq 0,\ x \geq 0,\ \ x \text{ in } R^n$$

and the first-order conditions can be written

$$\begin{aligned} f_x^* + \lambda^* g_x^* \leq 0, \quad & (f_x^* + \lambda^* g_x^*)x^* = 0, \\ \lambda^* g(x^*) = 0, \quad & g(x^*) \geq 0,\ g(x^*) \geq 0, \\ \lambda^* \geq 0 \quad & \text{and} \lambda^* g(x^*) = 0, \end{aligned}$$

where f_x^* denotes the derivative of *f* evaluated at x^*. In more explicit notation, the conditions $f_x^* + \lambda^* g_x^* = 0$ can be written as

$$\partial f / \partial x_i + \sum_{j=1}^{m} \lambda_j^* \partial g_j / \partial x_i = 0,\ i = 1, 2, \ldots, n$$

When the assumption of concavity is dropped, it is no longer possible to ensure that the local maximum is also a global one. However, it is still possible to analyse local constrained maxima in terms of local saddle-point conditions. In this case a condition is needed to ensure that the first-order conditions for a saddle point are indeed necessary conditions. The Kuhn-Tucker Constraint Qualification is such a condition. Arrow, Hurwicz and Uzawa (1961) have found a number of conditions, more useful in application to economic models, that imply the Constraint Qualification.
The case of more than one desired commodity leads to what is called the *vector maximum problem*, Kuhn and Tucker (1951). This may be defined as follows. Let $f_1,\bullet\bullet f_2,\ldots,\bullet\bullet f_k$ and $g_1,\bullet g_2\bullet,\ldots,\bullet g_m$ be real valued functions defined on a set *X* in R^n. We say x^* in *X* achieves a (global) *vector maximum* of $f=(f_1,\bullet\bullet f_2,\ldots,\bullet\bullet f_k)$ subject to $g_j(x)\geq 0$, *j*=1, 2•,…,•*m* if,

1. (I) $g_j(x^*)\geq 0$, *j*=1,•2,…,•*m*,

2. (II) there does not exist x' in X satisfying $f_i(x') \geq f_i(x^*)$ for i=1, 2,…, k with $f_i(x')>f_i(x^*)$ for some value of i, and $g_j(x')\geq 0$ for j=1,•2,…,•m.

This is just the concept of an efficient point expressed in the present notation.
A vector maximum has a saddle-point characterization similar to that for a scalar valued function.
Theorem: : Let f_1, f_2•,…,•f_k and g_1, g_2•,…,•g_m be real valued concave functions defined on a convex X set in R^n. Suppose there is x^0 in X such that $g_j(x^0)>0$, j=1, 2,…, m (Slater's Condition). If x^* achieves a vector maximum of f subject to $g(x)\geq 0$ then there exist $a=(a_1,•a_2,…,•a_k)$ and $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*)$ with $a_j\geq 0$ for all j, $a\neq 0$ and $\lambda \geq 0$ such that (x^*, λ^*) is a saddle point of the Lagrangean $L(x,• \lambda)=af(x)+ \lambda g(x)$.
Several different 'converses', to this theorem are known. One states that if x^* maximizes $L(x, \lambda^*)$ for some strictly positive vector a and non-negative λ^*, and if $\lambda^* g(x^*)=0$ and $g(x^*) \geq 0$, then x^* gives a vector maximum of f subject to $g(x)\geq 0$, and x in X. Another, parallel to the result for the case of one desired commodity, is the following.
Theorem: : Let f and g be functions as in the theorem above. If there are positive real numbers a_1,•a_2,…,•a_k and if (z^*, λ^*) is a saddle point of the Lagrangean L (defined as above) then (I) x^* achieves a maximum of f subject to $g(x)\geq 0$ on X, and (II) $\lambda^* g(x^*)=0$.

The positive numbers a_1•,…,•a_k are interpreted as prices of desired commodities, and the non-negative numbers λ_j^* are prices of the remaining commodities. The condition $\lambda^* g(x^*) =0$ which arises in these theorems states that the value of unused resources at the efficiency prices λ^* is zero; that is, resources not fully utilized at a vector maximum have a zero price.
The connection between vector maxima and Pareto optima is as follows. Because a vector maximum is an efficient point (for the vectorial ordering of the commodity space), it is a Pareto optimum for appropriately specified (non-satiated) utility functions, as was already pointed out in the case of the linear activity analysis model. Furthermore, if the functions f_1, …,•f_k are themselves utility functions, and the variable x denotes allocations, with the constraints g defining feasibility, then a vector maximum of f subject to the constraints $g(x)\geq 0$ and x in X is a Pareto optimum, and vice versa. Hence the saddle-point theorems give a characterization of Pareto optima by prices. The interpretation of prices in terms of decentralized resource allocation described in the linear activity analysis model also applies in this nonlinear model. The proofs of these theorems reveal an important logical role played by the principle of marginal cost pricing.
The basic theorems of nonlinear programming, especially the Kuhn-Tucker-Uzawa Theorem in the setting of the vector maximum problem, have been extended to the case of infinitely many commodities. (Hurwicz, 1958, first obtained the basic results in this field.) Technicalities aside, the theorems carry over to certain infinite dimensional spaces, namely linear topological spaces, or in the case of first-order conditions, Banach spaces.
Dropping the restriction to a finite number of basic processes leads to classical production or transformation function models of production, whose properties depend on the detailed specifications made.
Samuelson (1947) used Lagrangean methods to analyse interior maxima subject to equality constraints in the context of production function models, as well as that of optimization by consumers. He also gave the interpretation of Lagrange multipliers as shadow prices.

Efficient allocation in an economy with consumers and producers

In an economy with both consumption and production decisions, efficiency is concerned with distribution as well as production. Data about restrictions on consumption and the wants of consumers must be specified in addition to the data about production. The elements of the models are as follows.
The commodity space is denoted X; it might be l-dimensional Euclidean space, or a more abstract space such as an additive group in which, for example, some coordinates are restricted to have integer values. There is a (finite) list of consumers, 1,•2,…,•n, and a similar list of producers, 1,•2,…,•m. A *state* of the economy is an array consisting of a commodity bundle for each agent in the economy, consumer or producer. This may be written ($\langle x^i\rangle$, $\langle y^j\rangle$), where $\langle x^i\rangle=(x^1,•x^2,…,•x^n)$ and $\langle y^j\rangle=(y^1,•y^2,…,•y^m)$ and x^i and y^j are commodity bundles. Absolute constraints on consumption are expressed by requiring that the allocation $\langle x^i\rangle$ belong to a specified subset X of the space X^n of allocations. Examples of such constraints are:

1. 1. The requirement that the quantity of a certain commodity be non-negative.
2. 2. The requirement that a consumer requires certain minimum quantities of commodities in order to survive.

Each consumer i has a preference relation, denoted $\succsim_i$, defined on X. This formulation admits externalities in consumption, including physical externalities and externalities in

preferences; for example, preferences that depend on the consumption of other agents, termed non-selfish preferences. The consumption set of the *i*th consumer is the projection X^i of X onto the space of commodity bundles whose coordinates refer to the holdings of the *i*th consumer.

Technology is specified by a production set Y, a subset of X^m, consisting of those arrays $\langle y^j \rangle$ of input-output vectors that are jointly feasible for all producers. The production set of the *j*th producer, denoted Y^j, is the projection of Y onto the subspace of X^m whose coordinates refer to the *j*th producer.

The (aggregate) initial endowment of the economy is denoted by w, a commodity bundle in X.

These specifications define an *environment*, a term introduced by Hurwicz (1960) in this usage and according to him suggested by Jacob Marschak. This term refers to the primitive or given data from which analysis begins. Each environment determines a set of *feasible* states. These are the states $(\langle x^i \rangle, \langle y^j \rangle)$ such that $\langle x^i \rangle$ is in X, $\langle y^j \rangle$ is in Y and

$$\sum x^i - \sum y^j \leq w.$$

An environment determines the set of states that are Pareto optimal for that environment. Explicitly, they are the states $((x^{*i}), (y^{*j}))$ that are feasible in the given environment, and such that if any other state $((x^i), (y^j))$ has the property that $(x^i) \succsim_i (x^{*i})$ for all i with $(x^i) \succ_{i'} (x^{*i})$ for some i', then $((x^i), (y^j))$ is not feasible in the given environment.

It is important to note that the set of feasible states and the set of Pareto optimal states are completely determined by the environment; specification of economic organization is not involved.

At this level of generality, where externalities in consumption and production are admitted as possibilities, and where commodities may be indivisible, no general characterization of Pareto optima in terms of prices is possible. (Indeed, Pareto optima may not exist. Conditions that make the set of feasible allocations non-empty and compact and preferences continuous suffice to ensure the existence of Pareto optima.) In environments with externalities, or other non-neoclassical features, Pareto optima are generally not attainable by decentralized processes.

If the class of environments under consideration is restricted to the neoclassical environments, the fundamental theorems of welfare economics provide a characterization of Pareto optimal states via efficiency prices. That characterization has a natural interpretation in terms of a decentralized mechanism for allocation of resources.

The framework for these results is obtained by restricting the class of environments specified above as follows. The commodity space is to be Euclidean space of l dimensions, i.e. $X=R^l$. The consumption set for the economy is to be the product of its projections, i.e. $X = X^1 \times X^2 \times \cdots \times X^n$. This expresses the fact that if each agent's consumption is feasible for him, the total array is jointly feasible. Furthermore, each agent is restricted to having selfish preferences; that is, agent i's preference relation depends only on the coordinates of the allocation that refer to his holdings. In that case the preference relation $\succsim_i$ may be defined only on X^i, for each i. Similarly, externalities are ruled out in production, i.e. $Y = Y^1 \times Y^2 \times \cdots \times Y^m$.

The concept of an *equilibrium relative to a price system* (Debreu, 1959) serves to characterize Pareto optima by prices. A price system, denoted p, is an element of R^l; the environment $e = [(X^i), (\succsim_i), (Y^j), w]$ is of the restricted type specified above (free of externalities and indivisibilities).

A state $[(x^{*i}), (y^{*j})]$ of e is an *equilibrium relative to price system p* if:

1. 1. For every consumer i, x^{*i} maximizes preference $\succsim_i$ on the set of consumption bundles whose value at the prices p does not exceed the value of x^{*i} at those prices, i.e., if x^i is in $\{x^i \text{ in } x^i : px^i \leq px^{*i}\}$ then $x^i \precsim_i x^{*i}$.
2. 2. For every producer j, y^{*j} maximizes profit py^j on Y^j.
3. 3. Aggregate supply and demand balance, i.e.

$$\sum_i x^{*i} - \sum_i y^{*j} = w.$$

An equilibrium relative to a price system differs from a competitive equilibrium (see below) in that the former does not involve the budget constraints applying to consumers in the

latter concept. In an equilibrium relative to a price system the distribution of initial endowment and of the profits of firms among consumers need not be specified.

The first theorem of neoclassical welfare economics states, subject only to the exclusion of externalities and a mild condition that excludes preferences with thick indifference sets, that a state of an environment *e* that is an equilibrium relative to a price system *p* is a Pareto optimum of *e* (Koopmans, 1957).

The second welfare theorem is deeper and holds only on a smaller class of environments, sometimes referred to in the literature as the *classical environments* (called neoclassical above). One version of this theorem is as follows. Let $e = [(X^i), (\succsim_i), (Y^j), w]$ be an environment such that for each *i*

1. 1. X^i is convex.
2. 2. The preference relation $\succsim_i$ is continuous.
3. 3. The preference relation $\succsim_i$ is convex.
4. 4. The set $\Sigma_j Y_j$ is convex.

Let $[(x^{*i}), (y^{*j})]$ be a Pareto optimum of *e* such that there is at least one consumer who is not satiated at x^{*i}. Then there is a price system *p*, with not all components equal to 0, such that–except for Arrow's (1951) ‘exceptional case’, where *p* is such that for some *i* the expenditure px^{*i} is a minimum on the consumption set X^i–the state $[(x^{*i}), (y^{*j})]$ is an equilibrium relative to *p*.

(The condition that preferences are convex and not satiated is sufficient to exclude ‘thick’ indifference sets. A preference relation on X^i is convex if whenever x' and x'' are points of X^i with x' strictly preferred to x'' then the line segment connecting them (not including the point x'') is strictly preferred to x'. The consumption set X^i must be convex for this property to make sense. A preference relation is not satiated if there is no consumption preferred to all others.)

Hurwicz (1960) has given an alternative formalization of the competitive mechanism in which Arrow's exceptional case presents no difficulties.

If the exceptional case is not excluded, then it can still be said that:

1. 1. x^{*i} minimizes expenditure at prices *p* on the upper contour set of x^{*i}, for every *i*, and
2. 2. y^{*j} maximizes ‘profit’ py^j on the production set Y^j, for every *j*.

The state (x^*, y^*) together with the prices *p*, constitute a *valuation equilibrium* (Debreu, 1954).

As in the case of efficiency prices in pure production models, these prices have in themselves no institutional significance. They are, however, in the same way as other efficiency prices, suggestive of an interpretation in terms of decentralization.

If, in addition to the restriction to classical environments, the economic organization is specified to be that of a system of markets in a private ownership economy, and if agents are assumed to take prices as given, then the welfare theorems can translate into the assertion that the set of Pareto optima of an environment *e* and the set of competitive equilibria for *e* (subject to the possible redistribution of initial endowment and ownership shares) are identical. More precisely, the specification of the environment given above is augmented by giving each consumer a bundle of commodities, his initial endowment, denoted w^i. The total endowment is $w = \Sigma_i w^i$. Furthermore, each consumer has a claim to a share of the profits of each firm; the claims for the profit of each firm are assumed to add up to the entire profit. When prices and the production decisions of the firms are given, the profits of the firms are determined and so is the value of each consumer's initial endowment. Therefore, the income of each consumer is determined. Hence, the set of commodity bundles a consumer can afford to buy at the given prices, called his *budget set*, is determined; this consists of all bundles in his consumption set whose value at the given prices does not exceed his income at the given prices. Competitive behaviour of consumers means that each consumer treats the prices as given constants and chooses a bundle in his budget set that maximizes his preference: that is, a bundle x^i that is in X^i and such that if any other bundle x'^i is preferred to it, then x'^i is not in his budget set.

Competitive behaviour of firms is to maximize profits computed at the given prices *p*, regarded by the firms as constants; that is, a firm chooses a production vector y^j in its production set with the property that any other vector affording higher profits than py^j is not in the production set of firm *j*.

A *competitive equilibrium* is a specification of a commodity bundle for each consumer, a production vector for each firm, and a price system, together denoted $[(x^{*i}), (y^{*j}), p^*]$, where p^* has no negative components, satisfying the following conditions:

1. 1. For each consumer *i* the bundle x^{*i} maximizes preference on the budget set of *i*.
2. 2. For each firm *j* the production vector y^{*j} maximizes profit p^*y^j on the production set Y^j.

3. 3. For each commodity, the total consumption does not exceed the net total output of all firms plus the total initial endowment, i.e. $\Sigma_i x^{*i} - \Sigma_j y^{*j} \leq w = \Sigma_i w^i$;
4. 4. For those commodities *k* for which the inequality in 3 is strict; that is, the total consumption is less than initial endowment plus net output, the price p_k^* is zero.

The welfare theorems stated in terms of equilibrium relative to a price system translate directly into theorems stated in terms of competitive equilibrium. Briefly, every competitive equilibrium allocation in a given classical environment is Pareto optimal in that environment, and every Pareto optimal allocation in a given classical environment can be made a competitive equilibrium allocation of an environment that differs from the given one only in the distribution of the initial endowment. (Arrow (1951), Koopmans (1957), Debreu (1959) and Arrow and Hahn (1971) give modern and definitive treatment of the classical welfare theorems.)

It should be noted that the equilibria involved must exist for these theorems to have content. Sufficient conditions for existence of competitive equilibrium, which, since a competitive equilibrium is automatically an equilibrium relative to a price system, are also sufficient for existence of an equilibrium relative to a price system, include convexity and continuity of consumption sets and preferences and of production sets, as well as some assumptions which apply to the environment as a whole, restricting the ways in which individual agents may fit together to form an environment (Arrow and Debreu, 1954; Debreu, 1959; McKenzie, 1959).

The second welfare theorem involves redistribution of initial endowment. This is essential because the set of competitive equilibria from a given initial endowment is small (essentially finite) (Debreu, 1970), while the set of Pareto optima is generally a continuum. The set of Pareto optima cannot in general be generated as competitive allocations without varying the initial point. If redistribution is done by an economic mechanism, then it should be a decentralized one to support the interpretation given of the second welfare theorem. No such mechanism has been put forward as yet. Redistribution of initial endowment by lump-sum taxes and transfers has been discussed. A customary interpretation views these as brought about by a process outside economics, perhaps by a political process; no claim is made that such processes are decentralized. Some economists consider dependence on redistribution unsatisfactory because information about initial endowment is private; only the individual agent knows his own endowment. Consequently the expression of that information through political or other action can be expected to be strategic. The theory of second-best allocations has been proposed in this context. Redistribution of endowment is excluded, and the mechanism is restricted to be a price mechanism, but the price system faced by consumers is allowed to be different from that faced by producers; all agents behave according to the rules of the (static) competitive mechanism. The allocations that satisfy these conditions, when the price systems are variable, are maximal allocations in the sense that they are Pareto optimal within the restricted class just defined. These are so-called *second-best* allocations. This analysis was pioneered by Lipsey and Lancaster (1956) and Diamond and Mirrlees (1971).

Efficient allocation in non(neo)classical environments

The term *nonclassical* refers to those environments that fail to have the properties of classical ones; there may be indivisible commodities, nonconvexities in consumption sets, preferences or production sets, or externalities in production or consumption. An example of nonconvex preference would arise if a consumer preferred living in either Los Angeles or New York to living half the time in each city, or living halfway between them, depending on the way the commodity involved is specified. A production set representing a process that affords increasing returns to scale is an example of nonconvexity in production. A large investment project such as a road system is an example of a significant indivisibility. Phenomena of air or water pollution provide many examples of externalities in consumption and production.

The characterization of optimal allocation in terms of prices provided by the classical welfare theorems does not extend to nonclassical environments. If there are indivisibilities, equilibrium prices may fail to exist. Lerner (1934, 1947) has proposed a way of optimally allocating resources in the presence of indivisibilities. It would typically require adding up consumers' and producers' surplus.

Increasing returns to scale in production generally results in non-existence of competitive equilibrium, because of unbounded profit when prices are treated as given. Nash equilibrium, a concept from the theory of games, can exist even in cases of increasing returns. The difficulty is that such equilibria need not be optimal. Similar difficulties occur in cases of externalities.

Failure of the competitive price mechanism to extend the properties summarized in the classical welfare theorems to nonclassical environments has led economists to look for alternative ways of achieving optimal allocation in such cases. Such attempts have for the most part sought institutional arrangements that can be shown to result in optimal allocation. Ledyard (1968, 1971) analysed a mechanism for achieving Pareto optimal performance in environments with externalities. The use of taxes and subsidies advocated by Pigou (1932) to achieve Pareto optimal outcomes in cases of externalities is such an example. In a similar spirit Davis and Whinston (1962) distinguish externalities in production that leave marginal costs unaffected from those that do change marginal costs. In the former case they propose a pricing scheme, but one that involves lump-sum transfers. Marginal cost pricing, including lump-sum transfers to compensate for losses, which was extensively discussed as a device to achieve optimal allocation in the presence of increasing returns (Lerner, 1944; Hotelling, 1938; and many others) is another example of a scheme to realize optimal outcomes in nonclassical environments in a way that seeks to capture the benefits associated with decentralized resource allocation. In the case of production under conditions of increasing returns, the use of nonlinear prices has been suggested in an effort to

achieve optimality with at least some of the benefits of decentralization. (See Arrow and Hurwicz, 1960; Heal, 1971; Brown and Heal, 1982; Brown et al., 1986; Jennergren, 1971; Guesnerie, 1975.)

In the case of indivisibilities, and in the context of productive efficiency, integer programming algorithms exist for finding optima in specific problems, but a general characterization in terms of prices such as exists for the classical environments is not available. A decentralized process, involving the use of randomization, whose equilibria coincide with the set of Pareto optima has been put forward by Hurwicz, Radner and Reiter (1975). This process has the property that the counterparts of the classical welfare theorems hold for environments in which all commodities are indivisible, and the set of feasible allocations is finite, or in which there are no indivisible commodities, or externalities, but there may be nonconvexities in production or consumption sets, or in preferences. This, of course, includes the possibility of increasing returns to scale in production.

The schemes and processes that have been proposed, including many not described here, are quite different from one another. If attention is confined to pricing schemes without additional elements, such as lump-sum transfers, it may be satisfactory to proceed on the basis of an informal intuitive notion of decentralization. This amounts in effect to identifying decentralization with the competitive mechanism, or more generally with price or market mechanisms. If a broader class of processes is to be considered, including some already mentioned in this discussion, then a formal concept of decentralized resource allocation process is needed.

Efficient Allocation through Informationally Decentralized Processes

A formal definition of a concept of *allocation process* was first given by Hurwicz (1960). He also gave a definition of *informational decentralization* applying to a broad class of allocation mechanisms, based in part on a discussion by Hayek (1945) of the advantages of the competitive market mechanism for communicating knowledge initially dispersed among economic agents so that it can be brought to bear on the decisions that determine the allocation of resources. Hurwicz's formulation is as follows.

There is an initial dispersion of information about the environment; each agent is assumed to observe directly his own characteristic, e^i, but to know nothing directly about the characteristics of any other agent. In the absence of externalities, specifying the array of individual characteristics specifies the environment, i.e. $e=(e^1,\ldots,\bullet e^n)$. When there are externalities, an array of individual characteristics, each component of which corresponds to a possible environment, may not together constitute a possible environment. In more technical language, when there are externalities the set of environments is not the Cartesian product of its projections onto the sets of individual characteristics.

The goal of economic activity, whether efficiency, Pareto optimality or some other desideratum such as fairness, can be represented by a relation between the set of environments and the set of allocations, or outcomes. This relation assigns to each environment the set of allocations that meet the criterion of desirability. In the case of the Pareto criterion, the set of allocations that are Pareto optimal in a given environment is assigned to that environment. Formally, this relation is a correspondence (a set-valued function) from the set of environments to the set of allocations.

An allocation process, or mechanism, is modelled as an explicitly dynamic process of communication, leading to the determination of an outcome. In formal organizations standardized forms are frequently used for communication; in organized markets like the Stock Exchange, these include such things as order forms; in a business, forms on which weekly sales are reported; in the case of the Internal Revenue Service, income tax forms. A form consists of entries or blanks to be filled in a specified way. Thus, a form can be regarded as an ordered array of variables whose values come from specified sets. In the Hurwicz model, each agent is assumed to have a *language*, denoted M^i for the ith agent, from which his (possibly multi-dimensional) *message*, m^i, is chosen. The *joint message* of all the agents, $m=(m^1,\ldots,\bullet m^n)$ is in the *message space* $M = M^1 \times \cdots \times M^n$. Communication takes place in time, which is discrete; the message $m_t = (m_t^1, \ldots, m_t^n)$ denotes the message at time t. The message an agent emits at time t can depend on anything he knows at that time. This consists of what the agent knows about the environment by direct observation, by assumption, (*privacy*) his own characteristics, e^i for agent i, and what he has learned from others via the messages received from them. The agents' behaviour is represented by *response functions*, which show how the current message depends on the information at hand. Agent i's message at time t is

$$m_t^i = f^i(m_{t-1}, m_{t-2}, \ldots, e^i),\ i = 1, \ldots, n,\ t = 0, 1, 2, \ldots$$

If it is assumed that memory is finite, and bounded, it is possible without loss of generality to take the number of past periods remembered to be one. (If memory is unbounded, taking the number of periods remembered to be one excludes the possibility of a finite dimensional message space.) In that case the response equations become a system of first order temporally homogeneous difference equations in the messages. Thus:

$$m_t^i = f_i(m_{t-1}; e^i) i = 1, \ldots, n, t = 0, \ldots,$$

which can be written more compactly as

$$(^*) m_t = f(m_{t-1}; e).$$

(This formulation can accommodate the case of directed communication, in which some agents do not receive some messages; if agent *i* is not to receive the message of *j*, then f^i is independent of m^j, although m^j appears formally as an argument.) Analysis of informational properties of mechanisms is to begin with separated from that of incentives. When the focus is on communication and complexity qsts, the response functions are not regarded as chosen by the agent, but rather by the designer of the mechanism.

The iterative interchange of messages modelled by the difference equation system (*) eventually comes to an end, by converging to a stationary message. (It is also possible to have some stopping rule, such as to stop after a specified number of iterations.) The stationary message, which will be referred to as an *equilibrium message*, is then translated into an outcome, by means of the *outcome function*:

$$h: M \to Z,$$

where *Z* is the space of outcomes, usually allocations or trades. An allocation mechanism so modelled is called an *adjustment process*; it consists of the triple (*M*, *f*, *h*). Since no production or consumption takes place until all communication is completed, these processes are *tâtonnement* processes.

A more compact and general formulation was given by Mount and Reiter (1974) by looking only at message equilibria when attention is restricted to static properties. A correspondence is defined, called the *equilibrium message correspondence*. It associates to each environment the set of equilibrium messages for that environment. In order to satisfy the requirement of privacy, namely that each agent's message depend on the environment only through the agent's characteristic, the equilibrium message correspondence must be the intersection of individual message correspondences, each associating a set of message acceptable to the individual agent as equilibria in the light of his own characteristic. Thus the equilibrium message correspondence

$$\mu: E \to M,$$

is given by

$$\mu(e) = \bigcap_i \mu^i(e^i),$$

where $\mu^i: E^i \to M$, is the individual message correspondence of agent *i*. Note that here the message space *M* need not be the Cartesian product of individual languages. In the case of an adjustment process, the equilibrium message correspondence is defined by the conditions

$$\mu^i(e^i) = \{\min M \mid f^i(m; e^i) = m^i\},\ i = 1, \ldots, n$$

together with the condition that μ is the intersection of the μ^i. Specification of the outcome function $h: M \to Z$ completes the model, (M, μ, h).

The performance of a mechanism of this kind can be characterized by the mapping defined by the composition of the equilibrium message correspondence μ and the outcome function h. The mapping $h\mu: E \to Z$, possibly a correspondence, specifies the outcomes that the mechanism (M, μ, h) generates in each environment in E. A mechanism, whether in the form of an adjustment process, or in the equilibrium form, is called *Pareto-satisfactory* (Hurwicz, 1960) if for each environment in the class under consideration, the set of outcomes generated by the mechanism coincides with the set of Pareto optimal outcomes for that environment. Allowance must be made for redistribution of initial endowment, as in the case of the second welfare theorem. (A formulation in the framework of mechanisms is given in Mount and Reiter, 1977.)

The competitive mechanism formalized as a static mechanism is as follows. (Hurwicz, 1960, has given a different formulation, and Sonnenschein, 1974, has given an axiomatic characterization of the competitive mechanism from a somewhat different point of view.) The message space M is the space of prices and quantities of commodities going to each agent (it has dimension $n(l-1)$ when there are n agents and l commodities, taking account of budget constraints and Walras' Law), the individual message correspondence μ^i maps agent i's characteristic e^i to the graph of his excess demand function. The equilibrium message is the intersection of the individual ones, and is therefore the price-quantity combinations that solve the system of excess demand equations. The outcome function h is the projection of the equilibrium message onto the quantity components of M. Thus $h\mu(e)$ is a competitive equilibrium allocation (or trade) when the environment is e. The classical welfare theorems state that for each e in E_c, $h[\mu(e)] = p(e)$, where E_c denotes the set of classical environments and P is the Pareto correspondence. (Allowance must be made for redistribution of initial endowment in connection with the second welfare theorem. Explicit treatment of this is omitted to avoid notational complexity. The decentralized redistribution of initial endowment is, as in the case of the second welfare theorem, not addressed.) The welfare theorems can be summarized in the Mount-Reiter diagram (Figure 2) (Reiter, 1977).

Figure 2

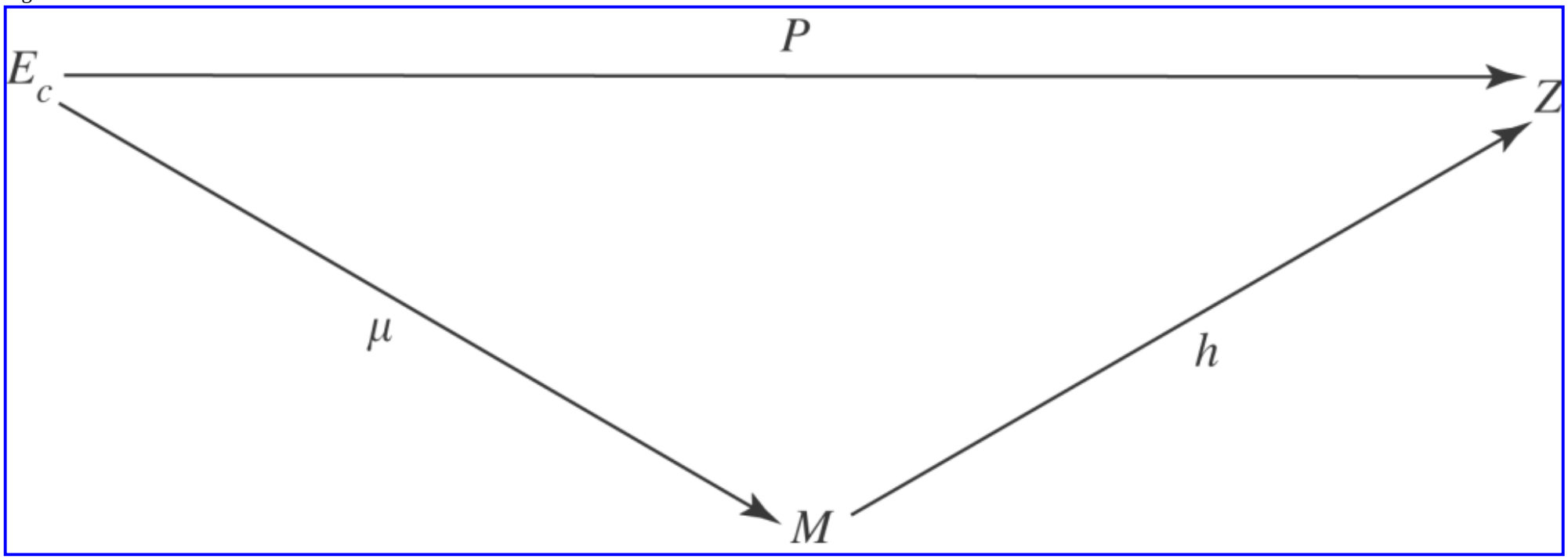

The welfare theorems state that this diagram *commutes* in the sense that starting from any environment e in E_c one reaches the same allocations via the mechanism, that is, via $h\mu$, as via the Pareto correspondence P.

With welfare theorems as a guide, the class of environments E_c can be replaced by some other class E, and the Pareto correspondence can be replaced by a correspondence, P,

embodying another criterion of optimality, and one can ask whether there is a mechanism, (M, μ, h) that makes the diagram commute, or, in other words, *realizes* P•? Without further restrictions on the mechanism, this is a triviality, because one agent can act as a central agent to whom all others communicate their environmental characteristics; the central agent then has the information required to evaluate P.

The concept of an *informationally decentralized mechanism* defined by Hurwicz (1960) makes explicit intuitive notions underlying the view that the price mechanism is decentralized. Informationally decentralized processes are a subclass of so-called *concrete processes*, introduced by Hurwicz (1960). These are processes that use a language and response rules that allow production and distribution plans to be specified explicitly. The informationally decentralized processes are those whose response rules permit agents to transmit information only about their own actions, and which in effect require each agent to treat the rest of the economy either as one aggregate, or in a symmetrical way that, like the aggregate, gives anonymity to the other agents.

In the case of static mechanisms, the requirements for informational decentralization boil down to the condition that the message space have no more than a certain finite dimension, and in some cases only that it be of finite dimension. In the case of classical environments this can be seen to include the competitive mechanism, and to exclude the obviously centralized one mentioned above.

Without going deeply into the matter, an objective of this line of research is to analyse explicitly the consequences of constraints on economic organization that come from limitations on the capacity of economic agents to observe, communicate and process information. One important result in this field is that there is no mechanism (M, μ, h) where μ preserves privacy, that uses messages smaller (in dimension) than those of the competitive mechanism (Hurwicz, 1972b; Mount and Reiter, 1974; Walker, 1977; Osana, 1978). Similar results have been obtained for environments with public goods, showing that the Lindahl mechanism uses the minimal message space (Sato, 1981). Another objective is to analyse effects on incentives arising from private motivations in the presence of private information; that is, information held by one agent that is not observable by others, except perhaps at a cost. (There is a large literature on this subject under the rubric ‘incentive compatibility’, or ‘strategic implementation’ (Dasgupta, Hammond and Maskin, 1979; Hurwicz, 1971, 1972a). The informational requirements of achieving a specified performance taking some aspects of incentive compatibility into account have been studied by Hurwicz (1976), Reichelstein (1984a, 1984b) and by Reichelstein and Reiter (1985).

Some important results for non-neoclassical environments can be mentioned. Hurwicz (1960, 1972a) has shown that there can be no informationally decentralized mechanism that realizes Pareto optimal performance on a class of environments that includes those with externalities. Calsamiglia (1977, 1982) has shown in a model of production that if the set of environments includes a sufficiently rich class of those with increasing returns to scale in production, then the dimension of the message space of any mechanism that realizes efficient production cannot be bounded.

Efficient allocation with infinitely many commodities

An infinite dimensional commodity space is needed when it is necessary to make infinitely many distinctions among goods and services. This is the case when commodities are distinguished according to time of availability and the time horizon in the model is not bounded or when time is continuous, or according to location when there is more than a finite number of possible locations; differentiated commodities provide other examples, and so does the case of uncertainty with infinitely many states. The bulk of the literature deals with the infinite horizon model of allocation over time, though recently more attention is given to models of product differentiation. Ramsey (1928) studied the problem of saving in a continuous time infinite horizon model with one consumption good and an infinitely lived consumer. He used as the criterion of optimality the infinite sum (integral) of undiscounted utility. Ramsey's contribution was largely ignored, and rediscovered when attention returned to problems of economic growth. A model of maximal sustainable growth based on a linear technology with no unproduced inputs was formulated by von Neumann (1937 in German; English translation, 1945–6). This contribution was unknown among English-speaking economists until after World War II. Study of intertemporal allocation by Anglo-American economists effectively began with the contributions of Harrod (1939) and Domar (1946). These models were concerned with stationary growth at a constant sustainable rate (stationary growth paths) rather than full intertemporal efficiency. Malinvaud (1953) first addressed this problem in a pioneering model of intertemporal allocation with an infinite horizon.

Efficient allocation over (discrete) time would be covered by the finite dimensional models described above if the time horizon were finite. It might be thought that a model with a sufficiently large but still finite horizon would for all practical purposes be equivalent to one with an infinite horizon, while avoiding the difficulties of infinity, but this is not the case, because of the dependence of efficient or optimal allocations on the value given to final stocks, a value that must depend on their uses beyond the horizon.

Malinvaud (1953) formulated an important infinite horizon model, which is the infinite dimensional counterpart of the linear activity analysis model of Koopmans. In Malinvaud's model time is discrete. The time horizon consists of an infinite sequence of time periods. At each date there are finitely many commodities. All commodities are desired in each time period, and no distinction is made between desired, intermediate and primary commodities. As in the activity analysis model, there is no explicit reference to preferences of consumers. Productive efficiency over time is analysed in terms of the output available for consumption, rather than the resulting utility levels.

Technology is represented by a production set X^t for each time period t=1, 2,…, an element of X^t being an ordered pair (a^t, b^{t+1}) of commodity bundles where a^t represents inputs to a production process in period t, and b^{t+1} represents the outputs of that process available at the beginning of period t+1. Here both a^t and b^{t+1} are non-negative. The set X^t is the

aggregate production set for the economy during period *t*. The net outputs available for consumption are given by

$$y^t = b^t - a^t, \text{ for } t \geq 1,$$

where b^1 is the initial endowment of resources available at the beginning of period 1. A *programme* is an infinite sequence $((a^t, b^{t+1}))$; it is a *feasible programme* if (a^t, b^{t+1}); is in X^t, and $b^t - a^t \geq 0$ for each $t \geq 1$, given b^1. The sequence $y=\langle y^t \rangle$ is called the *net output programme* associated with the given programme; it is a *feasible net output programme* if it is the net output programme of a feasible programme. A programme is *efficient* if it is (1) feasible and (2) there is no other programme that is feasible, from the same initial resources b^1, and provides at least as much net output in every period and a larger net output in some period. This is the concept of efficient production, already seen in the linear activity analysis model, now extended to an infinite horizon model. The main aim of this research is to extend to the infinite horizon model the characterization of efficient production by prices seen in the finite model. This goal is not quite reached, as is seen in what follows.

The main difficulties presented by the infinite horizon are already present in a special case of the Malinvaud model with one good and no consumers. Let *Y* be the set of all non-negative sequences $y=(y_t)$ that satisfy $0 \leq y_t = f(a_{t-1}) - a_t$ for $t \geq 1$, and $0 \leq y^0 = b^1 - a^0$, $b^1 > 0$, where *f* is a real-valued continuous concave function on the non-negative real numbers (the production function), $f(0)=0$, and b^1 is the given initial stock. The set *Y* is the set of all feasible programmes. A programme $y' - y > 0$. A price system is an infinite sequence $p=(p^t)$ of non-negative numbers. Denote by *P* the set of all price systems.

Malinvalud recognized the possibility that an efficient net output programme (y^t) need not have an associated system of non-zero prices (p^t) relative to which the production programme generating *y* satisfies the condition of intertemporal profit maximization, namely that

$$p^{t+1} f(a^t) - p^t a^t \geq p^{t+1} f(a) - p^t a$$

for all *t* and every $a \geq 0$. (Here (a^t) is the sequence of inputs producing *y*.) A condition introduced by Malinvaud, called *nontightness*, is sufficient for the existence of such non-zero prices. Alternative proofs of Malinvaud's existence theorem were given by Radner (1967) and Peleg and Yaari (1970). (An example showing the possibility of non-existence given by Peleg and Yaari (1970) is as follows. Suppose *f* is as shown in Figure 3.

Figure 3

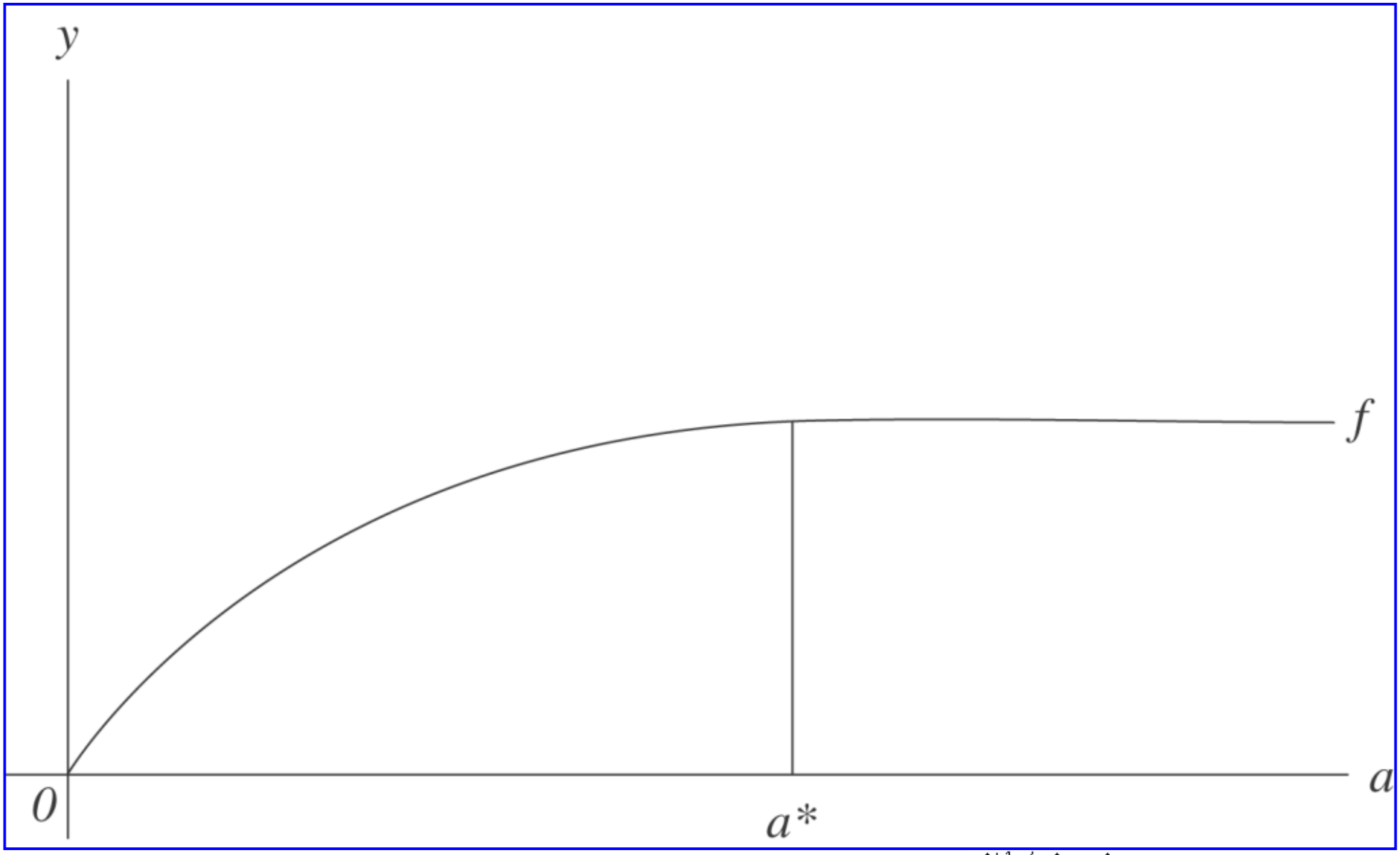

At an interior efficient, and therefore value maximizing, programme the first-order necessary conditions for a maximum imply $p^{t+1}f'(a^t) = p^t$. If there is a time at which $a^t=a^*$, in an efficient programme, then, since $f'(a^*) = 0$, it follows that prices at all prior and future times are 0. (Nontightness rules out such examples.)

On the side of sufficiency, Malinvaud showed that intertemporal profit maximization relative to a strictly positive price system p is not enough to ensure that a feasible programme is efficient. An additional (transversality) condition is needed. In the present model the following to such a condition;

$$\lim_{t \to \infty} p^t y^t = 0.$$

Cass (1972) has given a criterion that completely characterizes the set of efficient programmes in a one-good model with strictly concave and smooth production technology that satisfies endpoint conditions $0 \leq f'(\infty) < 1 < f'(x) < \infty$ for some $x>0$. Cass's criterion, states that a programme is *inefficient* if and only if the associated competitive prices – that is, satisfying $p^{t+1} f'(a^t) = p^t$ – also satisfy $\Sigma_{t=1}^{\infty}(1/p^t) < \infty$. This criterion may be interpreted as requiring the terms of trade between present and future to deteriorate sufficiently fast. Other similar conditions have been presented (Benveniste and Gale, 1975; Benveniste, 1976; Majumdar, 1974; Mitra, 1979). It is hard to see how any transversality condition can be interpreted in terms of decentralized resource allocation.

An alternate approach to characterizing efficient programmes was taken by Radner (1967), based on value functions as introduced in connection with valuation equilibrium by Debreu (1954). (Valuation equilibrium was discussed in connection with Arrow's exceptional case, above.) The value function approach was followed up by Majumdar (1970, 1972) and by Peleg and Yaari (1970). A price system defines a continuous linear functional, (a real-valued linear function) on the commodity space. This function assigns to a programme its present value. The present value may not be well-defined, because the infinite sequence that gives it diverges. This creates certain technical problems passed over here. A more important difficulty is that linear functionals exist that are not defined by price systems. Radner's approach was to characterize efficient programmes in terms of maximization of present value relative to a linear functional on the commodity space. Radner showed, technical matters aside, that:

1. 1. If a feasible programme maximizes the value of net output (consumption) relative to a strictly positive continuous linear functional, then it is efficient.
2. 2. If a given programme is efficient, then there is a nonzero non-negative continuous linear functional such that the given programme maximizes the value of net output relative to that functional on the set of feasible programmes.

These propositions seem to be the precise counterparts of the ones characterizing efficiency in the finite horizon model. Unfortunately, a linear functional may not have a representation in the form of the inner product of a price sequence with a net output sequence. (The production function $f(a) = a^{\beta}$, with $0 < \beta < 1$ provides an example. It is known that the programme with constant input sequence $x_t = (1/\beta)^{\beta/\beta-1}$ and output sequence $y_t = (1/\beta)^{\beta/\beta-1} - (1/\beta)^{1/\beta-1} t = 1, 2, \ldots,$ is efficient, and therefore there is a continuous linear functional relative to which it is value maximizing. But there is no price sequence (p^t) that represents that linear functional.) This presents a serious problem, because in the absence of such a representation it is unclear whether this characterization has an interpretation in terms of decentralized allocation processes; profit in any one period can depend on ‘prices at infinity’.

This approach has the advantage that it is applicable not only to infinite horizon models, but to a broader class in which the commodity space is infinite dimensional. Bewley (1972), Mas-Colell (1977) and Jones (1984) among others discuss Pareto optimality and competitive equilibrium in economies with infinitely many commodities. Hurwicz (1958) and others analysed optimal allocation in terms of nonlinear programming in infinite dimensional spaces. Theorems of programming in infinite dimensional spaces are also used in some of the models mentioned in this discussion.

The basic difficulties encountered in the one-good model, apart from the numerous technical problems that tend to make the literature large and diverse as different technical structures are investigated, are on the one hand the fact that transversality conditions are indispensable, and on the other the possibility that linear functionals, even when they exist, may not be representable in terms of price sequences. These problems raise strong doubt about the possibility of achieving efficient intertemporal resource allocation by decentralized means, though they leave open the possibility that some other decentralized mechanism, not using prices, might work. Analysis of this possibility has just begun, and is discussed below.

The difficulties seen in the one-good production model persist in more elaborate ones, including multisectoral models with efficiency as the criterion, and models with consumers in which Pareto optimality is the criterion. McFadden, Mitra and Majumdar (1980) studied a model in which there are firms, and overlapping generations of consumers, as in the model first investigated by Samuelson (1958). Each consumer lives for a finite time and has a consumption set and preferences like the consumers in a finite horizon model. A model with overlapping generations of consumers presents the fundamental difficulty that consumers cannot trade with future consumers as yet unborn. This difficulty can appear even in a finite horizon model if there are too few markets. The economy is closed in the sense that there are no nonproduced resources; the von Neumann growth model is an example of such a model. Building on the results of an earlier investigation (Majumdar, Mitra and McFadden, 1976), these authors introduced several notions of price systems, of competitive equilibrium, efficiency and optimality, and sought to establish counterparts of the classical welfare theorems. To summarize, in the 1976 paper they strengthen an earlier result of Bose (1974) to the effect that the problem of proper distribution of goods in essentially a short-run problem, and that the only long-run problem, one created by the infinite horizon, is that of inefficiency through overaccumulation of capital. In the 1980 paper the focus is on the relationships among various notions of equilibrium and Pareto optimality. The force of their results is, as might be expected, that the difficulties already seen in one-good model without consumers persist in this model. A transversality condition is made part of the definition of competitive equilibrium in order to obtain the result that an equilibrium is optimal. A partial converse requires some additional assumptions on the technology

(reachability) and on the way the economy fits together (nondecomposability). These results certainly illuminate the infinite horizon model with overlapping generations of consumers and producers, but the possibility of efficient or optimal resource allocation by decentralized means is not different from that in the one-good Malinvaud model. Hurwicz and Majumdar in an unpublished manuscript dated 1983, and later Hurwicz and Weinberger (1984), have addressed this issue directly, building on the approach of mechanism theory.

Hurwicz and Majumdar have studied the problem of efficiency in a model with an infinite number of periods. In each period there are finitely many commodities, one producer who is alive for just one period, and no consumers' choices. The criterion is the maximization of the discounted value of the programme (well-defined in this model). The producer alive in any period knows only the technology in that period. The question is whether there is a (static) privacy preserving mechanism using a finite dimensional message space whose equilibria coincide with the set of efficient programmes. The question can be put as follows. In each period a message is posted. The producer alive in that period responds 'Yes' or 'No'. If every producer over the entire infinite horizon answers 'Yes', the programme is an outcome corresponding to the equilibrium consisting of the infinite succession of posted messages. Since each producer knows only the technology prevailing in the period when he is alive, the process preserves privacy. If in addition the message posted in each period is finite dimensional, the process is informationally decentralized. Period-by-period profit maximization using period-by-period prices is a mechanism of this type; the message posted in each period consists of the vector of prices for that period, and the production plan for that period, both finite dimensional. The object is to characterize all efficient programmes as equilibria of such a mechanism. This would be an analogue of the classical welfare theorems, but without the restriction to mechanisms that use prices in their messages.

The main result is in the nature of an impossibility theorem. If the technology is constant over time, and that fact is common knowledge at the beginning, the problem is trivial since knowledge of the technology in the first period automatically means knowledge of it in every period. On the other hand, if there is some period whose technology is not known in the first period, then there is no finite dimensional message that can characterize efficient programmes, and in that sense, production cannot be satisfactorily decentralized over time.

Hurwicz and Weinberger (1984) have studied a model with both producers and consumers. As with producers, there is a consumer in each period, who lives for one period. The consumer in each period has a one-period utility function, which is not known by the producer; similarly the consumer does not know the production function. The criterion of optimality is the maximization of the sum of discounted utilities over the infinite horizon. Hurwicz and Weinberger show that there is no privacy preserving mechanism of the type just described whose equilibria correspond to the set of optimal programmes. It should be noted that their mechanism requires that the first-period actions (production, consumption and investment decisions) be made in the first period, and not be subject to revision after the infinite process of verification is completed. (On the other hand, under tâtonnement assumptions it may be possible to decentralize. In this model tâtonnement entails reconsideration 'at infinity'.)

If attention is widened to efficient programmes, and if technology is constant over time, there is an efficient programme with a fixed ratio of consumption to investment. This programme can be obtained as the equilibrium outcome of a mechanism of the specified type. However, this corresponds to only one side of the classical welfare theorems. It says that the outcome of such a mechanism is efficient; but it does not ensure that every efficient programme can be realized as the outcome of such a mechanism. The latter property fails in this model.

See Also

- incentive compatibility
- linear programming
- welfare economics

Bibliography

Arrow, K. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22(July), 265–90.

Arrow, K. and Hahn, F. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Arrow, K. and Hurwicz, L. 1960. Decentralization and computation in resource allocation. In *Essays in Economics and Econometrics*, ed. R.W. Pfouts. Chapel Hill: University of North Carolina Press, 34–104.

Arrow, K., Hurwicz, L. and Uzawa, H. 1961. Constraint qualifications in maximization problems. *Naval Research Logistics Quarterly* 8(2), June, 175–91.

Benveniste, L. 1976. Two notes on the Malinvaud condition for efficiency of infinite horizon programs. *Journal of Economic Theory* 12, 338–46.

Benveniste, L. and Gale, D. 1975. An extension of Cass' characterization of infinite efficient production programs. *Journal of Economic Theory* 10, 229–38.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4, 514–40.

Bose, A. 1974. Pareto optimality and efficient capital accumulation. Discussion Paper No. 74–4. Department of Economics, University of Rochester.

Brown, D. and Heal, G. 1982. Existence, local-uniqueness and optimality of a marginal cost pricing equilibrium in an economy with increasing returns. Cal. Tech. Social Science Working Paper No. 415.

Brown, D., Heal, G., Ali Khan, M. and Vohra, R. 1986. On a general existence theorem for marginal cost pricing equilibria. *Journal of Economic Theory* 38, 371–9.

Calsamiglia, X. 1977. Decentralized resource allocation and increasing returns. *Journal of Economic Theory* 14, 263–83.

Calsamiglia, X. 1982. On the size of the message space under non-convexities. *Journal of Mathematical Economics* 10, 197–203.

Cass, D. 1972. On capital over-accumulation in the aggregative neoclassical model of economic growth: a complete characterization. *Journal of Economic Theory* 4(2), April, 200–23.

Dantzig, G.B. 1951a. The programming of interdependent activities. In *Activity Analysis of Production and Allocation*, ed. T. Koopmans, Cowles Commission Monograph No. 13, New York: Wiley, ch. 2, 19–32.

Dantzig, G.B. 1951b. Maximization of a linear function of variables subject to linear inequalities. In *Activity Analysis of Production and Allocation*, ed. T. Koopmans, Cowles Commission Monograph No. 13, New York: Wiley, ch. 21, 339–47.

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules: some general results on incentive compatibility. *Review of Economic Studies* 46, 185–216.

Davis, O.A. and Whinston, A.B. 1962. Externalities welfare and the theory of games. *Journal of Political Economy* 70, 214–62.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. In *Proceedings of the National Academy of Sciences of the USA* 40(7), 588–92.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38(3), May, 387–92.

Diamond, P. and Mirrlees, J. 1971. Optimal taxation and public production. I: Production efficiency; II: Tax rules. *American Economic Review* 61, 8–27; 261–78.

Domar, E. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14(April), 137–47.

Fenchel, W. 1950. Convex cones, sets, and functions. Princeton University (hectographed).

Guesnerie, R. 1975. Pareto optimality in non-convex economies. *Econometrica* 43, 1–29.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49, 14–33.

Hayek, F. von. 1945. The use of knowledge in society. *American Economic Review* 35, 519–53. Reprinted in F. von Hayek, *Individualism and Economic Order*, Chicago: University of Chicago Press, 1949, 77–92.

Heal, G. 1971. Planning, prices and increasing returns. *Review of Economic Studies* 38, 281–94.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6, 242–69.

Hurwicz, L. 1958. Programming in linear spaces. In *Studies in Linear and Non-Linear Programming*, ed. K. Arrow, L. Hurwicz and H. Uzawa. Stanford: Stanford University Press.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical Methods in the Social Sciences, 1959*, ed. K.J. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

Hurwicz, L. 1971. Centralization and decentralization in economic processes. In *Comparison of Economic Systems: Theoretical and Methodological Approaches*, ed. A. Eckstein. Berkeley: University of California Press, ch. 3.

Hurwicz, L. 1972a. On informationally decentralized systems. In *Decision and Organization*, ed. C. McGuire and R. Radner. Amsterdam, London: North-Holland, ch. 14, 297–336.

Hurwicz, L. 1972b. On the dimensional requirements of in formationally decentralized Pareto-satisfactory processes. Presented at the Conference Seminar in Decentralization Northwestern University. In *Studies in Resource Allocation Processes*, ed. K.J. Arrow and L. Hurwicz. Cambridge: Cambridge University Press, 1977.

Hurwicz, L. 1976. On informational requirements for nonwasteful resource allocation systems. In *Mathematical Models in Economics: Papers and Proceedings of a US-USSR Seminar, Moscow*, ed. S. Shulman. New York: National Bureau of Economic Research.

Hurwicz, L., Radner, R. and Reiter, S. 1975. A stochastic decentralized resource allocation process. *Econometrica* 43: Part I, 187–221; Part II, 363–93.

Hurwicz, L. and Weinberger, H. 1984. Paper presented at IMA seminar in Minneapolis.

Jennergren, L. 1971. Studies in the mathematical theory of decentralized resource-allocation. PhD dissertation, Stanford University.

Jones, L. 1984. A competitive model of commodity differentiation. *Econometrica* 52, 507–30.

Kantorovitch, L. 1939. *Matematicheskie metody organizatii i planirovania proizvodstva* (Mathematical methods in the organization and planning of production). Izdanie Leningradskogo Gosudarstvennogo Universiteta, Leningrad. Trans. in *Management Science* 6(4), July 1960, 363–422.

Kantorovitch, L. 1942. On the translocation of masses. (In English.) *Comptes Rendus (Doklady) de l'Academie des Sciences d l'URSS* 37(7–8).

Kantorovitch, L. and Gavurin, M. 1949. Primenenie matematicheskikh metodov v voprosakh analyza grusopotokov (The application of mathematical methods to problems of freight flow analysis). In *Problemy Povysheniia Effektivnosty Raboty Transporta* (Problems of raising the efficiency of transportation), ed. V. Zvonkov. Moscow and Leningrad: Izdatel'stvo Akademii Nauk SSSR.

Koopmans, T.C. 1951a. Analysis of production as an efficient combination of activities. In *Activity Analysis of Production and Allocation*, ed. T. Koopmans, Cowles Commission Monograph No. 13, New York: Wiley, ch. 3, 33–97.

Koopmans, T.C. 1951b. Efficient allocation of resources. *Econometrica* 19, 455–65.

Koopmans, T.C. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill, 66–104.

Kuhn, H. and Tucker, A. 1951. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press, 481–92.

Ledyard, J. 1968. Resource allocation in unselfish environments. *American Economic Review* 58, 227–37.

Ledyard, J. 1971. A convergent Pareto-satisfactory non-tâtonnement adjustment process for a class of unselfish exchange environments. *Econometrica* 39, 467–99.

Lerner, A. 1934. The concept of monopoly and measurement of monopoly power. *Review of Economic Studies* 1(3), June, 157–75.

Lerner, A. 1944. *The Economics of Control*. New York: Macmillan.

Lipsey, R. and Lancaster, K. 1956. The general theory of second best. *Review of Economic Studies* 24, 11–32.

McFadden, D., Mitra, T. and Majumdar, M. 1980. Pareto optimality and competitive equilibrium in infinite horizon economies. *Journal of Mathematical Economics* 7, 1–26.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27(1), January, 54–71.

Majumdar, M. 1970. Some approximation theorems on efficiency prices for infinite programs. *Journal of Economic Theory* 2, 399–410.

Majumdar, M. 1972. Some general theorems of efficiency prices with an infinite dimensional commodity space. *Journal of Economic Theory* 5, 1–13.

Majumdar, M. 1974. Efficient programs in infinite dimensional spaces: a complete characterization. *Journal of Economic Theory* 7, 355–69.

Majumdar, M., Mitra, T. and McFadden, D. 1976. On efficiency and Pareto optimality of competitive programs in closed multisector models. *Journal of Economic Theory* 13, 26–46.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21, 233–68.

Mas-Colell, A. 1977. Regular nonconvex economies. *Econometrica* 45, 1387–407.

Mitra, T. 1979. On optimal economic growth with variable discount rates: existence and stability results. *International Economic Review* 20, 133–45.

Mount, K. and Reiter, S. 1974. The informational size of message spaces. *Journal of Economic Theory* 8, 161–92.

Mount, K. and Reiter, S. 1977. Economic environments for which there are Pareto satisfactory mechanisms. *Econometrica* 45, 821–42.

Neumann, J. von. 1937. A model of general economic equilibrium. *Ergebnisse eines mathematischen Kolloquiums*, No. 8. Trans. from German, *Review of Economic Studies* 13(1), (1945–6), 1–9.

Nikaido, H. 1969. *Convex Structures and Economic Theory*. New York: Academic Press.

Nikaido, H. 1970. *Introduction to Sets and Mappings in Modern Economics*. Trans. K. Sato, Amsterdam: North-Holland (Japanese original, Tokyo, 1960).

Osana, H. 1978. On the informational size of message spaces for resource allocation processes. *Journal of Economic Theory* 17, 66–78.

Peleg, B. and Yaari, M. 1970. Efficiency prices in an infinite dimensional commodity space. *Journal of Economic Theory* 2, 41–85.

Pigou, A. 1932. *The Economics of Welfare*. 4th edn, London: Macmillan.

Radner, R. 1967. Efficiency prices for infinite horizon production programs. *Review of Economic Studies* 34, 51–66.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Reichelstein, S. 1984a. Dominant strategy implementation, incentive compatibility and informational requirements. *Journal of Economic Theory* 34(1), October, 32–51.

Reichelstein, S. 1984b. Information and incentives in economic organizations. PhD dissertation, Northwestern University.

Reichelstein, S. and Reiter, S. 1985. Game forms with minimal strategy spaces. Discussion Paper No. 663, The Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston, Ill.

Reiter, S. 1977. Information and performance in the (new)^2 welfare economics. *American Economic Review* 67, 226–34.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Samuelson, P. 1949. Market mechanisms and maximization, I, II, III. Hectographed memoranda, The RAND Corporation, Santa Monica.

Samuelson, P. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66(December), 467–82.

Sato, F. 1981. On the informational size of message spaces for resource allocation processes in economies with public goods. *Journal of Economic Theory* 24, 48–69.

Slater, M. 1950. Lagrange multipliers revisited: a contribution to non-linear programming. *Cowles Commission Discussion Paper*, Math. 403, also RM–676, 1951.

Sonnenschein, H. 1974. An axiomatic characterization of the price mechanism. *Econometrica* 42, 425–34.

Uzawa, H. 1958. The Kuhn-Tucker Theorem in concave programming. In *Studies in Linear and Non-Linear Programming*, ed. K. Arrow, L. Hurwicz and H. Uzawa. Stanford: Stanford University Press.

Walker, M. 1977. On the informational size of message spaces. *Journal of Economic Theory* 15, 366–75.

How to cite this article

Reiter, Stanley. "efficient allocation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The

New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000049>
doi:10.1057/9780230226203.0453

The New Palgrave Dictionary of Economics Online

efficient markets hypothesis

Andrew W. Lo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The efficient markets hypothesis (EMH) maintains that market prices fully reflect all available information. Developed independently by Paul A. Samuelson and Eugene F. Fama in the 1960s, this idea has been applied extensively to theoretical models and empirical studies of financial securities prices, generating considerable controversy as well as fundamental insights into the price-discovery process. The most enduring critique comes from psychologists and behavioural economists who argue that the EMH is based on counterfactual assumptions regarding human behaviour, that is, rationality. Recent advances in evolutionary psychology and the cognitive neurosciences may be able to reconcile the EMH with behavioural anomalies.

Keywords

adaptive markets hypothesis; agent-based models; altruism; Arbitrage; asset price anomalies; behavioural biases; behavioural economics; behavioural finance; bid–ask bounce; biology and economics; bounded rationality; capital asset pricing model; consumer choice; creative destruction; deductive inference; dividend smoothing; dividend-discount model; Dutch books; economic complexity; efficient markets hypothesis; emotions; equilibrium; equity risk premium; evolutionary economics; evolutionary game theory; evolutionary psychology; Fama, E. F.; financial economics; herding; hyperbolic discounting; inductive inference; information aggregation; informational efficiency; January effect; joint hypotheses; learning; long-term memory; loss aversion; market efficiency; martingales; miscalibration of probabilities; natural selection; neuroeconomics; noise traders; optimization; overconfidence; overreaction; post-earnings announcement drift; preferences; present value; price reversals; psychological accounting; punctuated equilibrium; random walk hypothesis; rational expectations; regret; relative efficiency; risk aversion; risk preferences; risk–reward relation; Samuelson, P. A.; satisficing; serial correlation; Simon, H.; size effect; social norms; sociobiology; statistical inference; stock price volatility; survival of the fittest; uncertainty; utility maximization; variance

bounds; variance decomposition

Article

There is an old joke, widely told among economists, about an economist strolling down the street with a companion. They come upon a $100 bill lying on the ground, and as the companion reaches down to pick it up, the economist says, ‘Don't bother – if it were a genuine $100 bill, someone would have already picked it up’. This humorous example of economic logic gone awry is a fairly accurate rendition of the efficient markets hypothesis (EMH), one of the most hotly contested propositions in all the social sciences. It is disarmingly simple to state, has far-reaching consequences for academic theories and business practice, and yet is surprisingly resilient to empirical proof or refutation. Even after several decades of research and literally thousands of published studies, economists have not yet reached a consensus about whether markets – particularly financial markets – are, in fact, efficient.

The origins of the EMH can be traced back to the work of two individuals in the 1960s: Eugene F. Fama and Paul A. Samuelson. Remarkably, they independently developed the same basic notion of market efficiency from two rather different research agendas. These differences would propel the them along two distinct trajectories leading to several other breakthroughs and milestones, all originating from their point of intersection, the EMH.

Like so many ideas of modern economics, the EMH was first given form by Paul Samuelson (1965), whose contribution is neatly summarized by the title of his article: ‘Proof that Properly Anticipated Prices Fluctuate Randomly’. In an informationally efficient market, price changes must be unforecastable if they are properly anticipated, that is, if they fully incorporate the information and expectations of all market participants. Having developed a series of linear-programming solutions to spatial pricing models with no uncertainty, Samuelson came upon the idea of efficient markets through his interest in temporal pricing models of storable commodities that are harvested and subject to decay. Samuelson's abiding interest in the mechanics and kinematics of prices, with and without uncertainty, led him and his students to several fruitful research agendas including solutions for the dynamic asset-allocation and consumption-savings problem, the fallacy of time diversification and log-optimal investment policies, warrant and option-pricing analysis and, ultimately, the Black and Scholes (1973) and Merton (1973) option-pricing models.

In contrast to Samuelson's path to the EMH, Fama's (1963; 1965a; 1965b, 1970) seminal papers were based on his interest in measuring the statistical properties of stock prices, and in resolving the debate between technical analysis (the use of geometric patterns in price and volume charts to forecast future price movements of a security) and fundamental analysis (the use of accounting and economic data to determine a security's fair value). Among the first to employ modern digital computers to conduct empirical research in finance, and the first to use the term ‘efficient markets’ (Fama, 1965b), Fama operationalized the EMH hypothesis – summarized compactly in the epigram ‘prices fully reflect all available information’ – by placing structure on various information sets available to market participants. Fama's fascination with empirical analysis led him and his students down a very different path from Samuelson's, yielding significant methodological and empirical contributions such as the event study, numerous econometric tests of single- and multi-factor linear asset-pricing models, and a host of empirical regularities and anomalies in stock, bond, currency and commodity markets.

The EMH's concept of informational efficiency has a Zen-like, counter-intuitive flavour to it: the more efficient the market, the more random the sequence of price changes generated by such a market, and the most efficient market of all is one in which price changes are completely random and unpredictable. This is not an accident of nature, but is in fact the direct result of many active market participants attempting to profit from their information. Driven by profit opportunities, an army of investors pounce on even the smallest informational advantages at their disposal, and in doing so they incorporate their information into market prices and quickly eliminate the profit opportunities that first motivated their trades. If this occurs instantaneously, which it must in an idealized world of ‘frictionless’ markets and costless trading, then prices must always fully reflect all available information. Therefore, no profits can be garnered from information-based trading because such profits must have already been captured (recall the $100 bill on the ground). In mathematical terms, prices follow martingales.

Such compelling motivation for randomness is unique among the social sciences and is reminiscent of the role that uncertainty plays in quantum mechanics. Just as Heisenberg's uncertainty principle places a limit on what we can know about an electron's position and momentum if quantum mechanics holds, this version of the EMH places a limit on what we can know about future price changes if the forces of economic self-interest hold.

A decade after Samuelson's (1965) and Fama's (1965a; 1965b; 1970) landmark papers, many others extended their framework to allow for risk-averse investors, yielding a ‘neoclassical’ version of the EMH where price changes, properly weighted by aggregate marginal utilities, must be unforecastable (see, for example, LeRoy, 1973; M. Rubinstein, 1976; and Lucas, 1978). In markets where, according to Lucas (1978), all investors have ‘rational expectations’, prices do fully reflect all available information and marginal-utility-weighted prices follow martingales. The EMH has been extended in many other directions, including the incorporation of non-traded assets such as human capital, state-dependent preferences, heterogeneous investors, asymmetric information, and transactions costs. But the general thrust is the same: individual investors form expectations rationally, markets aggregate information efficiently, and equilibrium prices incorporate all available information instantaneously.

The random walk hypothesis

The importance of the EMH stems primarily from its sharp empirical implications many of which have been tested over the years. Much of the EMH literature before LeRoy (1973) and Lucas (1978) revolved around the random walk hypothesis (RWH) and the martingale model, two statistical descriptions of unforecastable price changes that were initially taken to be implications of the EMH. One of the first tests of the RWH was developed by Cowles and Jones (1937), who compared the frequency of *sequences* and *reversals* in historical stock returns, where the former are pairs of consecutive returns with the same sign, and the latter are pairs of consecutive returns with opposite signs. Cootner (1962; 1964), Fama (1963; 1965a), Fama and Blume (1966), and Osborne (1959) perform related tests of the RWH and, with the exception of Cowles and Jones (who subsequently acknowledged an error in their analysis – Cowles, 1960), all of these articles indicate support for the RWH using historical stock price data.

More recently, Lo and MacKinlay (1988) exploit the fact that return variances scale linearly under the RWH – the variance of a two-week return is twice the variance of a one-week return if the RWH holds –

and construct a variance ratio test which rejects the RWH for weekly US stock returns indexes from 1962 to 1985. In particular, they find that variances grow faster than linearly as the holding period increases, implying positive serial correlation in weekly returns. Oddly enough, Lo and MacKinlay also show that individual stocks generally do satisfy the RWH, a fact that we shall return to below.

French and Roll (1986) document a related phenomenon: stock return variances over weekends and exchange holidays are considerably lower than return variances over the same number of days when markets are open. This difference suggests that the very act of trading creates volatility, which may well be a symptom of Black's (1986) noise traders.

For holding periods much longer than one week – for example, three to five years – Fama and French (1988) and Poterba and Summers (1988) find negative serial correlation in US stock returns indexes using data from 1926 to 1986. Although their estimates of serial correlation coefficients seem large in magnitude, there is insufficient data to reject the RWH at the usual levels of significance. Moreover, a number of statistical artifacts documented by Kim, Nelson and Startz (1991) and Richardson (1993) cast serious doubt on the reliability of these longer-horizon inferences.

Finally, Lo (1991) considers another aspect of stock market prices long thought to have been a departure from the RWH: long-term memory. Time series with long-term memory exhibit an unusually high degree of persistence, so that observations in the remote past are non-trivially correlated with observations in the distant future, even as the time span between the two observations increases. Nature's predilection towards long-term memory has been well-documented in the natural sciences such as hydrology, meteorology, and geophysics, and some have argued that economic time series must therefore also have this property.

However, using recently developed statistical techniques, Lo (1991) constructs a test for long-term memory that is robust to short-term correlations of the sort uncovered by Lo and MacKinlay (1988), and concludes that, despite earlier evidence to the contrary, there is little support for long-term memory in stock market prices. Departures from the RWH can be fully explained by conventional models of short-term dependence.

Variance bounds tests

Another set of empirical tests of the EMH starts with the observation that in a world without uncertainty the market price of a share of common stock must equal the present value of all future dividends, discounted at the appropriate cost of capital. In an uncertain world, one can generalize this *dividend-discount model* or *present-value relation* in the natural way: the market price equals the conditional expectation of the present value of all future dividends, discounted at the appropriate risk-adjusted cost of capital, and conditional on all available information. This generalization is explicitly developed by Grossman and Shiller (1981).

LeRoy and Porter (1981) and Shiller (1981) take this as their starting point in comparing the variance of stock market prices to the variance of *ex post* present values of future dividends. If the market price is the conditional expectation of present values, then the difference between the two, that is, the forecast error, must be uncorrelated with the conditional expectation by construction. But this implies that the variance of the *ex post* present value is the sum of the variance of the market price (the conditional expectation) and the variance of the forecast error. Since volatilities are always non-negative, this

variance decomposition implies that the variance of stock prices cannot exceed the variance of *ex post* present values. Using annual US stock market data from various sample periods, LeRoy and Porter (1981) and Shiller (1981) find that the variance bound is violated dramatically. Although LeRoy and Porter are more circumspect about the implications of such violations, Shiller concludes that stock market prices are too volatile and the EMH must be false.

These two papers ignited a flurry of responses which challenged Shiller's controversial conclusion on a number of fronts. For example, Flavin (1983), Kleidon (1986), and Marsh and Merton (1986) show that statistical inference is rather delicate for these variance bounds, and that, even if they hold in theory, for the kind of sample sizes Shiller uses and under plausible data-generating processes the sample variance bound is often violated purely due to sampling variation. These issues are well summarized in Gilles and LeRoy (1991) and Merton (1987).

More importantly, on purely theoretical grounds Marsh and Merton (1986) and Michener (1982) provide two explanations for violations of variance bounds that are perfectly consistent with the EMH. Marsh and Merton (1986) show that if managers smooth dividends – a well-known empirical phenomenon documented in several studies of dividend policy – and if earnings follow a geometric random walk, then the variance bound is violated in theory, in which case the empirical violations may be interpreted as *support* for this version of the EMH.

Alternatively, Michener constructs a simple dynamic equilibrium model along the lines of Lucas (1978) in which prices do fully reflect all available information at all times but where individuals are risk averse, and this risk aversion is enough to cause the variance bound to be violated in theory as well. These findings highlight an important aspect of the EMH that had not been emphasized in earlier studies: tests of the EMH are always tests of joint hypotheses. In particular, the phrase ‘prices fully reflect all available information’ is a statement about two distinct aspects of prices: the information content and the price formation mechanism. Therefore, any test of this proposition must concern the *kind* of information reflected in prices, and *how* this information comes to be reflected in prices.

Apart from issues regarding statistical inference, the empirical violation of variance bounds may be interpreted in many ways. It may be a violation of EMH, or a sign that investors are risk averse, or a symptom of dividend smoothing. To choose among these alternatives, more evidence is required.

Overreaction and underreaction

A common explanation for departures from the EMH is that investors do not always react in proper proportion to new information. For example, in some cases investors may overreact to performance, selling stocks that have experienced recent losses or buying stocks that have enjoyed recent gains. Such overreaction tends to push prices beyond their ‘fair’ or ‘rational’ market value, only to have rational investors take the other side of the trades and bring prices back in line eventually. An implication of this phenomenon is price reversals: what goes up must come down, and vice versa. Another implication is that *contrarian* investment strategies – strategies in which ‘losers’ are purchased and ‘winners’ are sold – will earn superior returns.

Both of these implications were tested and confirmed using recent US stock market data. For example, using monthly returns of New York Stock Exchange (NYSE) stocks from 1926 to 1982, DeBondt and Thaler (1985) document the fact that the winners and losers in one 36-month period tend to reverse their

performance over the next 36-month period. Curiously, many of these reversals occur in January (see the discussion below on the ‘January effect’). Chopra, Lakonishok and Ritter (1992) reconfirm these findings after correcting for market risk and the size effect. And Lehmann (1990) shows that a zero-net-investment strategy in which long positions in losers are financed by short positions in winners almost always yields positive returns for monthly NYSE/AMEX stock returns data from 1962 to 1985.
However, Chan (1988) argues that the profitability of contrarian investment strategies cannot be taken as conclusive evidence against the EMH because there is typically no accounting for risk in these profitability calculations (although Chopra, Lakonishok and Ritter, 1992 do provide risk adjustments, their focus was not on specific trading strategies). By risk-adjusting the returns of a contrarian trading strategy according to the capital asset pricing model, Chan (1988) shows that the expected returns are consistent with the EMH.
Moreover, Lo and MacKinlay (1990c) show that at least half of the profits reported by Lehmann (1990) are not due to overreaction but rather the result of positive cross-autocorrelations between stocks. For example, suppose the returns of two stocks A and B are both serially uncorrelated but are positively cross-autocorrelated. The lack of serial correlation implies no overreaction (which is characterized by negative serial correlation), but positive cross-autocorrelations yields positive expected returns to contrarian trading strategies. The existence of several economic rationales for positive cross-autocorrelation that are consistent with EMH suggests that the profitability of contrarian trading strategies is not sufficient evidence to conclude that investors overreact.
The reaction of market participants to information contained in earnings announcements also has implications for the EMH. In one of the earliest studies of the information content of earnings, Ball and Brown (1968) show that up to 80 per cent of the information contained in the earnings ‘surprises’ is anticipated by market prices.
However, the more recent article by Bernard and Thomas (1990) argues that investors sometimes underreact to information about future earnings contained in current earnings. This is related to the ‘post-earnings announcement drift’ puzzle first documented by Ball and Brown (1968), in which the information contained in earnings announcement takes several days to become fully impounded into market prices. Although such effects are indeed troubling for the EMH, their economic significance is often questionable – while they may violate the EMH in frictionless markets, very often even the smallest frictions – for example, positive trading costs, taxes – can eliminate the profits from trading strategies designed to exploit them.

Anomalies

Perhaps the most common challenge to the EMH is the anomaly, a regular pattern in an asset's returns which is reliable, widely known, and inexplicable. The fact that the pattern is regular and reliable implies a degree of predictability, and the fact that the regularity is widely known implies that many investors can take advantage of it.
For example, one of the most enduring anomalies is the ‘size effect’, the apparent excess expected returns that accrue to stocks of small-capitalization companies – in excess of their risks – which was first discovered by Banz (1981). Keim (1983), Roll (1983), and Rozeff and Kinney (1976) document a

related anomaly: small capitalization stocks tend to outperform large capitalization stocks by a wide margin over the turn of the calendar year. This so-called ‘January effect’ seems robust to sample period, and is difficult to reconcile with the EMH because of its regularity and publicity. Other well-known anomalies include the Value Line enigma (Copeland and Mayers, 1982), the profitability of short-term return-reversal strategies in US equities (Rosenberg, Reid and Lanstein,1985; Chan, 1988; Lehmann, 1990; and Lo and MacKinlay, 1990c), the profitability of medium-term momentum strategies in US equities (Jegadeesh, 1990; Chan, Jegadeesh and Lakonishok, 1996; and Jegadeesh and Titman, 2001), the relation between price/earnings ratios and expected returns (Basu, 1977), the volatility of orange juice futures prices (Roll, 1984), and calendar effects such as holiday, weekend, and turn-of-the-month seasonalities (Lakonishok and Smidt, 1988).

What are we to make of these anomalies? On the one hand, their persistence in the face of public scrutiny seems to be a clear violation of the EMH. After all, most of these anomalies can be exploited by relatively simple trading strategies, and, while the resulting profits may not be riskless, they seem unusually profitable relative to their risks (see, especially, Lehmann, 1990).

On the other hand, EMH supporters might argue that such persistence is in fact evidence in favour of EMH or, more to the point, that these anomalies cannot be exploited to any significant degree because of factors such as risk or transactions costs. Moreover, although some anomalies are currently inexplicable, this may be due to a lack of imagination on the part of academics, not necessarily a violation of the EMH. For example, recent evidence suggests that the January effect is largely due to ‘bid–ask bounce’, that is, closing prices for the last trading day of December tend to be at the bid price and closing prices for the first trading day of January tend to be at the ask price. Since small-capitalization stocks are also often low-price stocks, the effects of bid–ask bounce in percentage terms are much more pronounced for these stocks – a movement from bid to ask for a $5.00 stock on the NYSE (where the minimum bid-ask spread was $0.125 prior to decimalization in 2000) represents a 2.5 per cent return.

Whether or not one can profit from anomalies is a question unlikely to be settled in an academic setting. While calculations of ‘paper’ profits of various trading strategies come easily to academics, it is virtually impossible to incorporate in a realistic manner important features of the trading process such as transactions costs (including price impact), liquidity, rare events, institutional rigidities and non-stationarities. The economic value of anomalies must be decided in the laboratory of actual markets by investment professionals, over long periods of time, and even in these cases superior performance and simple luck are easily confused.

In fact, luck can play another role in the interpretation of anomalies: it can account for anomalies that are not anomalous. Regular patterns in historical data can be found even if no regularities exist, purely by chance. Although the likelihood of finding such spurious regularities is usually small (especially if the regularity is a very complex pattern), it increases dramatically with the number of ‘searches’ conducted on the same set of data. Such *data-snooping* biases are illustrated in Brown et al. (1992) and Lo and MacKinlay (1990b) – even the smallest biases can translate into substantial anomalies such as superior investment returns or the size effect.

Behavioural critiques

The most enduring critiques of the EMH revolve around the preferences and behaviour of market

participants. The standard approach to modelling preferences is to assert that investors optimize additive time-separable expected utility functions from certain parametric families – for example, constant relative risk aversion. However, psychologists and experimental economists have documented a number of departures from this paradigm, in the form of specific behavioural biases that are ubiquitous to human decision-making under uncertainty, several of which lead to undesirable outcomes for an individual's economic welfare – for example, overconfidence (Fischoff and Slovic, 1980; Barber and Odean, 2001; Gervais and Odean, 2001), overreaction (DeBondt and Thaler, 1985), loss aversion (Kahneman and Tversky, 1979; Shefrin and Statman, 1985; Odean, 1998), herding (Huberman and Regev, 2001), psychological accounting (Tversky and Kahneman, 1981), miscalibration of probabilities (Lichtenstein, Fischoff and Phillips, 1982), hyperbolic discounting (Laibson, 1997), and regret (Bell, 1982). These critics of the EMH argue that investors are often – if not always – irrational, exhibiting predictable and financially ruinous behaviour.

To see just how pervasive such behavioural biases can be, consider the following example which is a slightly modified version of an experiment conducted by two psychologists, Kahneman and Tversky (1979). Suppose you are offered two investment opportunities, A and B: A yields a sure profit of $240,000, and B is a lottery ticket yielding $1 million with a 25 per cent probability and $0 with 75 per cent probability. If you had to choose between A and B, which would you prefer? Investment B has an expected value of $250,000, which is higher than A's payoff, but this may not be all that meaningful to you because you will receive either $1 million or zero. Clearly, there is no right or wrong choice here; it is simply a matter of personal preferences. Faced with this choice, most subjects prefer A, the sure profit, to B, despite the fact that B offers a significant probability of winning considerably more. This behaviour is often characterized as ‘risk aversion’ for obvious reasons. Now suppose you are faced with another two choices, C and D: C yields a sure loss of $750,000, and D is a lottery ticket yielding $0 with 25 per cent probability and a loss of $1 million with 75 per cent probability. Which would you prefer? This situation is not as absurd as it might seem at first glance; many financial decisions involve choosing between the lesser of two evils. In this case, most subjects choose D, despite the fact that D is more risky than C. When faced with two choices that both involve losses, individuals seem to be ‘risk seeking’, not risk averse as in the case of A versus B.

The fact that individuals tend to be risk averse in the face of gains and risk seeking in the face of losses can lead to some very poor financial decisions. To see why, observe that the combination of choices A and D is equivalent to a single lottery ticket yielding $240,000 with 25 per cent probability and –$760,000 with 75 per cent probability, whereas the combination of choices B and C is equivalent to a single lottery ticket yielding $250,000 with 25 per cent probability and –$750,000 with 75 per cent probability. The B and C combination has the same probabilities of gains and losses, but the gain is $10,000 higher and the loss is $10,000 lower. In other words, B and C is formally equivalent to A and D plus a sure profit of $10,000. In light of this analysis, would you still prefer A and D?

A common response to this example is that it is contrived because the two pairs of investment opportunities were presented sequentially, not simultaneously. However, in a typical global financial institution the London office may be faced with choices A and B and the Tokyo office may be faced with choices C and D. Locally, it may seem as if there is no right or wrong answer – the choice between A and B or C and D seems to be simply a matter of personal risk preferences – but the globally consolidated financial statement for the entire institution will tell a very different story. From that

perspective, there *is* a right and wrong answer, and the empirical and experimental evidence suggests that most individuals tend to select the wrong answer. Therefore, according to the behaviouralists, quantitative models of efficient markets – all of which are predicated on rational choice – are likely to be wrong as well.

Impossibility of efficient markets

Grossman and Stiglitz (1980) go even farther – they argue that perfectly informationally efficient markets are an *impossibility* for, if markets are perfectly efficient, there is no profit to gathering information, in which case there would be little reason to trade and markets would eventually collapse. Alternatively, the degree of market *inefficiency* determines the effort investors are willing to expend to gather and trade on information, hence a non-degenerate market equilibrium will arise only when there are sufficient profit opportunities, that is, inefficiencies, to compensate investors for the costs of trading and information gathering. The profits earned by these attentive investors may be viewed as 'economic rents' that accrue to those willing to engage in such activities. Who are the providers of these rents? Black (1986) gave us a provocative answer: 'noise traders', individuals who trade on what they consider to be information but which is, in fact, merely noise.

The supporters of the EMH have responded to these challenges by arguing that, while behavioural biases and corresponding inefficiencies do exist from time to time, there is a limit to their prevalence and impact because of opposing forces dedicated to exploiting such opportunities. A simple example of such a limit is the so-called 'Dutch book', in which irrational probability beliefs give rise to guaranteed profits for the savvy investor. Consider, for example, an event E, defined as 'the S&P 500 index drops by five per cent or more next Monday', and suppose an individual has the following irrational beliefs: there is a 50 per cent probability that E will occur, and a 75 per cent probability that E will *not* occur. This is clearly a violation of one of the basic axioms of probability theory – the probabilities of two mutually exclusive and exhaustive events must sum to 1 – but many experimental studies have documented such violations among an overwhelming majority of human subjects.

These inconsistent subjective probability beliefs imply that the individual would be willing to take both of the following bets B_1 and B_2:

$$B_1 = \begin{cases} \$1 \text{ if } E \\ -\$1 \text{ otherwise} \end{cases}, \quad B_2 = \begin{cases} \$1 \text{ if } E^c \\ -\$1 \text{ otherwise} \end{cases}$$

where E^c denotes the event 'not E'. Now suppose we take the opposite side of both bets, placing \$50 on B_1 and \$25 on B_2. If E occurs, we lose \$50 on B_1 but gain \$75 on B_2, yielding a profit of \$25. If E^c occurs, we gain \$50 on B_1 and lose \$25 on B_2, also yielding a profit of \$25. Regardless of the outcome, we have secured a profit of \$25, an 'arbitrage' that comes at the expense of the individual with inconsistent probability beliefs. Such beliefs are not sustainable, and market forces – namely, arbitrageurs such as hedge funds and proprietary trading groups – will take advantage of these

opportunities until they no longer exist, that is, until the odds are in line with the axioms of probability theory. (Only when these axioms are satisfied is arbitrage ruled out. This was conjectured by Ramsey, 1926, and proved rigorously by de Finetti, 1937, and Savage, 1954.) Therefore, proponents of the classical EMH argue that there are limits to the degree and persistence of behavioural biases such as inconsistent probability beliefs, and substantial incentives for those who can identify and exploit such occurrences. While all of us are subject to certain behavioural biases from time to time, according to EMH supporters market forces will always act to bring prices back to rational levels, implying that the impact of irrational behaviour on financial markets is generally negligible and, therefore, irrelevant.
But this last conclusion relies on the assumption that market forces are sufficiently powerful to overcome any type of behavioural bias, or equivalently that irrational beliefs are not so pervasive as to overwhelm the capacity of arbitrage capital dedicated to taking advantage of such irrationalities. This is an empirical issue that cannot be settled theoretically, but must be tested through careful measurement and statistical analysis. The classic reference by Kindleberger (1989) – where a number of speculative bubbles, financial panics, manias, and market crashes are described in detail – suggests that the forces of irrationality can overwhelm the forces of arbitrage capital for months and, in several well-known cases, years.
So what does this imply for the EMH?

The current state of the EMH

Given all of the theoretical and empirical evidence for and against the EMH, what can we conclude? Amazingly, there is still no consensus among economists. Despite the many advances in the statistical analysis, databases, and theoretical models surrounding the EMH, the main result of all of these studies is to harden the resolve of the proponents of each side of the debate.
One of the reasons for this state of affairs is the fact that the EMH, by itself, is not a well-defined and empirically refutable hypothesis. To make it operational, one must specify additional structure, for example, investors' preferences or information structure. But then a test of the EMH becomes a test of several auxiliary hypotheses as well, and a rejection of such a joint hypothesis tells us little about which aspect of the joint hypothesis is inconsistent with the data. Are stock prices too volatile because markets are inefficient, or due to risk aversion, or dividend smoothing? All three inferences are consistent with the data. Moreover, new statistical tests designed to distinguish among them will no doubt require auxiliary hypotheses of their own which, in turn, may be questioned.
More importantly, tests of the EMH may not be the most informative means of gauging the efficiency of a given market. What is often of more consequence is the efficiency of a particular market *relative* to other markets – for example, futures vs. spot markets, auction vs. dealer markets. The advantages of the concept of relative efficiency, as opposed to the all-or-nothing notion of absolute efficiency, are easy to spot by way of an analogy. Physical systems are often given an efficiency rating based on the relative proportion of energy or fuel converted to useful work. Therefore, a piston engine may be rated at 60 per cent efficiency, meaning that on average 60 per cent of the energy contained in the engine's fuel is used to turn the crankshaft, with the remaining 40 per cent lost to other forms of work, such as heat, light or noise.
Few engineers would ever consider performing a statistical test to determine whether or not a given

engine is perfectly efficient – such an engine exists only in the idealized frictionless world of the imagination. But measuring relative efficiency – relative, that is, to the frictionless ideal – is commonplace. Indeed, we have come to expect such measurements for many household products: air conditioners, hot water heaters, refrigerators, and so on. Therefore, from a practical point of view, and in light of Grossman and Stiglitz (1980), the EMH is an idealization that is economically unrealizable, but which serves as a useful benchmark for measuring relative efficiency.

The desire to build financial theories based on more realistic assumptions has led to several new strands of literature, including psychological approaches to risk-taking behaviour (Kahneman and Tversky, 1979; Thaler, 1993; Lo, 1999), evolutionary game theory (Friedman, 1991), agent-based modelling of financial markets (Arthur et al., 1997; Chan et al., 1998), and direct applications of the principles of evolutionary psychology to economics and finance (Lo, 1999; 2002; 2004; 2005; Lo and Repin, 2002). Although substantially different in methods and style, these emerging sub-fields are all directed at new interpretations of the EMH. In particular, psychological models of financial markets focus on the the manner in which human psychology influences the economic decision-making process as an explanation of apparent departures from rationality. Evolutionary game theory studies the evolution and steady-state equilibria of populations of competing strategies in highly idealized settings. Agent-based models are meant to capture complex learning behaviour and dynamics in financial markets using more realistic markets, strategies, and information structures. And applications of evolutionary psychology provide a reconciliation of rational expectations with the behavioural findings that often seem inconsistent with rationality.

For example, in one agent-based model of financial markets (Farmer, 2002), the market is modelled using a non-equilibrium market mechanism, whose simplicity makes it possible to obtain analytic results while maintaining a plausible degree of realism. Market participants are treated as computational entities that employ strategies based on limited information. Through their (sometimes suboptimal) actions they make profits or losses. Profitable strategies accumulate capital with the passage of time, and unprofitable strategies lose money and may eventually disappear. A financial market can thus be viewed as a co-evolving ecology of trading strategies. The strategy is analogous to a biological species, and the total capital deployed by agents following a given strategy is analogous to the population of that species. The creation of new strategies may alter the profitability of pre-existing strategies, in some cases replacing them or driving them extinct.

Although agent-based models are still in their infancy, the simulations and related theory have already demonstrated an ability to understand many aspects of financial markets. Several studies indicate that, as the population of strategies evolves, the market tends to become more efficient, but this is far from the perfect efficiency of the classical EMH. Prices fluctuate in time with internal dynamics caused by the interaction of diverse trading strategies. Prices do not necessarily reflect ‘true values’; if we view the market as a machine whose job is to set prices properly, the inefficiency of this machine can be substantial. Patterns in the price tend to disappear as agents evolve profitable strategies to exploit them, but this occurs only over an extended period of time, during which substantial profits may be accumulated and new patterns may appear.

The adaptive markets hypothesis

The methodological differences between mainstream and behavioural economics suggest that an alternative to the traditional deductive approach of neoclassical economics may be necessary to reconcile the EMH with its behavioural critics. One particularly promising direction is to view financial markets from a biological perspective and, specifically, within an evolutionary framework in which markets, instruments, institutions and investors interact and evolve dynamically according to the ‘law’ of economic selection. Under this view, financial agents compete and adapt, but they do not necessarily do so in an optimal fashion (see Farmer and Lo, 1999; Farmer, 2002; Lo, 2002; 2004; 2005).

This evolutionary approach is heavily influenced by recent advances in the emerging discipline of ‘evolutionary psychology’, which builds on the seminal research of E.O. Wilson (1975) in applying the principles of competition, reproduction, and natural selection to social interactions, yielding surprisingly compelling explanations for certain kinds of human behaviour, such as altruism, fairness, kin selection, language, mate selection, religion, morality, ethics and abstract thought (see, for example, Barkow, Cosmides and Tooby, 1992; Gigerenzer, 2000). ‘Sociobiology’ is the rubric that Wilson (1975) gave to these powerful ideas, which generated a considerable degree of controversy in their own right, and the same principles can be applied to economic and financial contexts. In doing so, we can fully reconcile the EMH with all of its behavioural alternatives, leading to a new synthesis: the adaptive markets hypothesis (AMH).

Students of the history of economic thought will no doubt recall that Thomas Malthus used biological arguments – the fact that populations increase at geometric rates whereas natural resources increase at only arithmetic rates – to arrive at rather dire economic consequences, and that both Darwin and Wallace were influenced by these arguments (see Hirshleifer, 1977, for further details). Also, Joseph Schumpeter's view of business cycles, entrepreneurs and capitalism have an unmistakeable evolutionary flavour to them; in fact, his notions of ‘creative destruction’ and ‘bursts’ of entrepreneurial activity are similar in spirit to natural selection and Eldredge and Gould's (1972) notion of ‘punctuated equilibrium’.

More recently, economists and biologists have begun to explore these connections in several veins: direct extensions of sociobiology to economics (Becker, 1976; Hirshleifer, 1977); evolutionary game theory (Maynard Smith, 1982); evolutionary economics (Nelson and Winter, 1982); and economics as a complex system (Anderson, Arrow and Pines, 1988). And publications like the *Journal of Evolutionary Economics* and the *Electronic Journal of Evolutionary Modeling and Economic Dynamics* now provide a home for research at the intersection of economics and biology.

Evolutionary concepts have also appeared in a number of financial contexts. For example, Luo (1995) explores the implications of natural selection for futures markets, and Hirshleifer and Luo (2001) consider the long-run prospects of overconfident traders in a competitive securities market. The literature on agent-based modelling pioneered by Arthur et al. (1997), in which interactions among software agents programmed with simple heuristics are simulated, relies heavily on evolutionary dynamics. And at least two prominent practitioners have proposed Darwinian alternatives to the EMH. In a chapter titled ‘The Ecology of Markets’, Niederhoffer (1997, ch. 15) likens financial markets to an ecosystem with dealers as ‘herbivores’, speculators as ‘carnivores’, and floor traders and distressed investors as ‘decomposers’. And Bernstein (1998) makes a compelling case for active management by pointing out that the notion of equilibrium, which is central to the EMH, is rarely realized in practice and that market dynamics are better explained by evolutionary processes.

Clearly the time is now ripe for an evolutionary alternative to market efficiency.

To that end, in the current context of the EMH we begin, as Samuelson (1947) did, with the theory of the individual consumer. Contrary to the neoclassical postulate that individuals maximize expected utility and have rational expectations, an evolutionary perspective makes considerably more modest claims, viewing individuals as organisms that have been honed, through generations of natural selection, to maximize the survival of their genetic material (see, for example, Dawkins, 1976). While such a reductionist approach can quickly degenerate into useless generalities – for example, the molecular biology of economic behaviour – nevertheless, there are valuable insights to be gained from the broader biological perspective. Specifically, this perspective implies that behaviour is not necessarily intrinsic and exogenous, but evolves by natural selection and depends on the particular environment through which selection occurs. That is, natural selection operates not only upon genetic material but also upon social and cultural norms in *homo sapiens*; hence Wilson's term ‘sociobiology’.

To operationalize this perspective within an economic context, consider the idea of ‘bounded rationality’ first espoused by Nobel-prize-winning economist Herbert Simon. Simon (1955) suggested that individuals are hardly capable of the kind of optimization that neoclassical economics calls for in the standard theory of consumer choice. Instead, he argued that, because optimization is costly and humans are naturally limited in their computational abilities, they engage in something he called ‘satisficing’, an alternative to optimization in which individuals make choices that are merely satisfactory, not necessarily optimal. In other words, individuals are bounded in their degree of rationality, which is in sharp contrast to the current orthodoxy – rational expectations – where individuals have unbounded rationality (the term ‘hyper-rational expectations’ might be more descriptive). Unfortunately, although this idea garnered a Nobel Prize for Simon, it had relatively little impact on the economics profession. (However, his work is now receiving greater attention, thanks in part to the growing behavioural literature in economics and finance. See, for example, Simon, 1982; Sargent, 1993; A. Rubinstein, 1998; Gigerenzer and Selten, 2001.) Apart from the sociological factors discussed above, Simon's framework was commonly dismissed because of one specific criticism: what determines the point at which an individual stops optimizing and reaches a satisfactory solution? If such a point is determined by the usual cost–benefit calculation underlying much of microeconomics (that is, optimize until the marginal benefits of the optimum equals the marginal cost of getting there), this assumes the optimal solution is known, which would eliminate the need for satisficing. As a result, the idea of bounded rationality fell by the wayside, and rational expectations has become the de facto standard for modelling economic behaviour under uncertainty.

An evolutionary perspective provides the missing ingredient in Simon's framework. The proper response to the question of how individuals determine the point at which their optimizing behaviour is satisfactory is this: such points are determined not analytically but through trial and error and, of course, natural selection. Individuals make choices based on past experience and their ‘best guess’ as to what might be optimal, and they learn by receiving positive or negative reinforcement from the outcomes. If they receive no such reinforcement, they do not learn. In this fashion, individuals develop heuristics to solve various economic challenges, and, as long as those challenges remain stable, the heuristics will eventually adapt to yield approximately optimal solutions to them.

If, on the other hand, the environment changes, then it should come as no surprise that the heuristics of the old environment are not necessarily suited to the new. In such cases, we observe ‘behavioural biases’ – actions that are apparently ill-advised in the context in which we observe them. But rather than

labelling such behaviour ‘irrational’, it should be recognized that suboptimal behaviour is not unlikely when we take heuristics out of their evolutionary context. A more accurate term for such behaviour might be ‘maladaptive’. The flopping of a fish on dry land may seem strange and unproductive, but under water the same motions are capable of propelling the fish away from its predators.

By coupling Simon's notion of bounded rationality and satisficing with evolutionary dynamics, many other aspects of economic behaviour can also be derived. Competition, cooperation, market-making behaviour, general equilibrium, and disequilibrium dynamics are all adaptations designed to address certain environmental challenges for the human species, and by viewing them through the lens of evolutionary biology we can better understand the apparent contradictions between the EMH and the presence and persistence of behavioural biases.

Specifically, the adaptive markets hypothesis can be viewed as a new version of the EMH, derived from evolutionary principles. Prices reflect as much information as dictated by the combination of environmental conditions and the number and nature of ‘species’ in the economy or, to use the appropriate biological term, the *ecology*. By ‘species’ I mean distinct groups of market participants, each behaving in a common manner. For example, pension funds may be considered one species; retail investors, another; market-makers, a third; and hedge-fund managers, a fourth. If multiple species (or the members of a single highly populous species) are competing for rather scarce resources within a single market, that market is likely to be highly efficient – for example, the market for 10-Year US Treasury Notes reflects most relevant information very quickly indeed. If, on the other hand, a small number of species are competing for rather abundant resources in a given market, that market will be less efficient – for example, the market for oil paintings from the Italian Renaissance. Market efficiency cannot be evaluated in a vacuum, but is highly context-dependent and dynamic, just as insect populations advance and decline as a function of the seasons, the number of predators and prey they face, and their abilities to adapt to an ever-changing environment.

The profit opportunities in any given market are akin to the amount of food and water in a particular local ecology – the more resources present, the less fierce the competition. As competition increases, either because of dwindling food supplies or an increase in the animal population, resources are depleted which, in turn, causes a population decline eventually, decreasing the level of competition and starting the cycle again. In some cases cycles converge to corner solutions, that is, certain species become extinct, food sources are permanently exhausted, or environmental conditions shift dramatically. By viewing economic profits as the ultimate food source on which market participants depend for their survival, the dynamics of market interactions and financial innovation can be readily derived.

Under the AMH, behavioural biases abound. The origins of such biases are heuristics that are adapted to non-financial contexts, and their impact is determined by the size of the population with such biases versus the size of competing populations with more effective heuristics. During the autumn of 1998, the desire for liquidity and safety by a certain population of investors overwhelmed the population of hedge funds attempting to arbitrage such preferences, causing those arbitrage relations to break down. However, in the years prior to August 1998 fixed-income relative-value traders profited handsomely from these activities, presumably at the expense of individuals with seemingly ‘irrational’ preferences (in fact, such preferences were shaped by a certain set of evolutionary forces, and might be quite rational in other contexts). Therefore, under the AMH, investment strategies undergo cycles of profitability and loss in response to changing business conditions, the number of competitors entering and exiting the industry, and the type and magnitude of profit opportunities available. As opportunities shift, so too will

the affected populations. For example, after 1998 the number of fixed-income relative-value hedge funds declined dramatically – because of outright failures, investor redemptions, and fewer start-ups in this sector – but many have reappeared in recent years as performance for this type of investment strategy has improved.

Even fear and greed – the two most common culprits in the downfall of rational thinking according to most behaviouralists – are the product of evolutionary forces, adaptive traits that enhance the probability of survival. Recent research in the cognitive neurosciences and economics, now coalescing into the discipline known as ‘neuroeconomics’, suggests an important link between rationality in decision-making and emotion (Grossberg and Gutowski, 1987; Damasio, 1994; Elster, 1998; Lo and Repin, 2002; and Loewenstein, 2000), implying that the two are not antithetical but in fact complementary. For example, contrary to the common belief that emotions have no place in rational financial decision-making processes, Lo and Repin (2002) present preliminary evidence that physiological variables associated with the autonomic nervous system are highly correlated with market events even for highly experienced professional securities traders. They argue that emotional responses are a significant factor in the real-time processing of financial risks, and that an important component of a professional trader's skills lies in his or her ability to channel emotion, consciously or unconsciously, in specific ways during certain market conditions.

This argument often surprises economists because of the link between emotion and behavioural biases, but a more sophisticated view of the role of emotions in human cognition shows that they are central to rationality (see, for example, Damasio, 1994; Rolls, 1999). In particular, emotions are the basis for a reward-and-punishment system that facilitates the selection of advantageous behaviour, providing a numeraire for animals to engage in a ‘cost–benefit analysis’ of the various actions open to them (Rolls, 1999, ch. 10.3). From an evolutionary perspective, emotion is a powerful adaptation that dramatically improves the efficiency with which animals learn from their environment and their past (see Damasio, 1994). These evolutionary underpinnings are more than simple speculation in the context of financial market participants. The extraordinary degree of competitiveness of global financial markets and the outsize rewards that accrue to the ‘fittest’ traders suggest that Darwinian selection – ‘survival of the richest’, to be precise – is at work in determining the typical profile of the successful trader. After all, unsuccessful traders are eventually eliminated from the population after suffering a certain level of losses.

The new paradigm of the AMH is still under development, and certainly requires a great deal more research to render it ‘operationally meaningful’ in Samuelson's sense. However, even at this early stage it is clear that an evolutionary framework is able to reconcile many of the apparent contradictions between efficient markets and behavioural exceptions. The former may be viewed as the steady-state limit of a population with constant environmental conditions, and the latter involves specific adaptations of certain groups that may or may not persist, depending on the particular evolutionary paths that the economy experiences. More specific implications may be derived through a combination of deductive and inductive inference – for example, theoretical analysis of evolutionary dynamics, empirical analysis of evolutionary forces in financial markets, and experimental analysis of decision-making at the individual and group level.

For example, one implication is that, to the extent that a relation between risk and reward exists, it is unlikely to be stable over time. Such a relation is determined by the relative sizes and preferences of

various populations in the market ecology, as well as institutional aspects such as the regulatory environment and tax laws. As these factors shift over time, any risk–reward relation is likely to be affected. A corollary of this implication is that the equity risk premium is also time-varying and path-dependent. This is not so revolutionary an idea as it might first appear – even in the context of a rational expectations equilibrium model, if risk preferences change over time, then the equity risk premium must vary too. The incremental insight of the AMH is that aggregate risk preferences are not immutable constants, but are shaped by the forces of natural selection. For example, until recently US markets were populated by a significant group of investors who had never experienced a genuine bear market – this fact has undoubtedly shaped the aggregate risk preferences of the US economy, just as the experience since the bursting of the technology bubble in the early 2000s has affected the risk preferences of the current population of investors. In this context, natural selection determines who participates in market interactions; those investors who experienced substantial losses in the technology bubble are more likely to have exited the market, leaving a markedly different population of investors. Through the forces of natural selection, history matters. Irrespective of whether prices fully reflect all available information, the particular path that market prices have taken over the past few years influences current aggregate risk preferences. Among the three fundamental components of any market equilibrium – prices, probabilities, and preferences – preferences is clearly the most fundamental and least understood. Several large bodies of research have developed around these issues – in economics and finance, psychology, operations research (also called ‘decision sciences’) and, more recently, brain and cognitive sciences – and many new insights are likely to flow from synthesizing these different strands of research into a more complete understanding of how individuals make decisions (see Starmer, 2000, for an excellent review of this literature). Simon's (1982) seminal contributions to this literature are still remarkably timely and their implications have yet to be fully explored.

Conclusions

Many other practical insights and potential breakthroughs can be derived from shifting our mode of thinking in financial economics from the physical to the biological sciences. Although evolutionary ideas are not yet part of the financial mainstream, the hope is that they will become more commonplace as they demonstrate their worth – ideas are also subject to ‘survival of the fittest’. No one has illustrated this principal so well as Harry Markowitz, the father of modern portfolio theory and a Nobel laureate in economics in 1990. In describing his experience as a Ph.D. student on the eve of his graduation, he wrote in his Nobel address (Markowitz, 1991, p. 476):

> … [W]hen I defended my dissertation as a student in the Economics Department of the University of Chicago, Professor Milton Friedman argued that portfolio theory was not Economics, and that they could not award me a Ph.D. degree in Economics for a dissertation which was not Economics. I assume that he was only half serious, since they did award me the degree without long debate. As to the merits of his arguments, at this point I am quite willing to concede: at the time I defended my dissertation, portfolio theory was not part of Economics. But now it is.

In light of the sociology of the EMH controversy (see, for example, Lo, 2004), the debate is likely to continue. However, despite the lack of consensus in academia and industry, the ongoing dialogue has given us many new insights into the economic structure of financial markets. If, as Paul Samuelson has suggested, financial economics is the crown jewel of the social sciences, then the EMH must account for half the facets.

See Also

- financial market anomalies
- rational expectations
- rationality, bounded

I thank John Cox, Gene Fama, Bob Merton, and Paul Samuelson for helpful discussions.

Bibliography

Anderson, P., Arrow, K. and Pines, D., eds. 1988. *The Economy as an Evolving Complex System*. Reading, MA: Addison-Wesley Publishing Company.

Arthur, B., Holland, J., LeBaron, B., Palmer, R. and Tayler, P. 1997. Asset pricing under endogenous expectations in an artificial stock market. In *The Economy as an Evolving Complex System II*, ed. B. Arthur, S. Durlauf, and D. Lane. Reading, MA: Addison Wesley.

Ball, R. and Brown, P. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6, 159–78.

Banz, R. 1981. The relationship between return and market value of common stock. *Journal of Financial Economics* 9, 3–18.

Barber, B. and Odean, T. 2001. Boys will be boys: gender, overconfidence, and common stock investment. *Quarterly Journal of Economics* 116, 261–29.

Barkow, J., Cosmides, L. and Tooby, J. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Basu, S. 1977. The investment performance of common stocks in relation to their price–earnings ratios: a test of the efficient market hypothesis. *Journal of Finance* 32, 663–82.

Becker, G. 1976. Altruism, egoism, and genetic fitness: economics and sociobiology. *Journal of Economic Literature* 14, 817–26.

Bell, D. 1982. Risk premiums for decision regret. *Management Science* 29, 1156–66.

Bernard, V. and Thomas, J. 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics* 13, 305–40.

Bernstein, P. 1998. Why the efficient market offers hope to active management. In *Economics and Portfolio Strategy*, 1 October. New York: Peter Bernstein, Inc.

Black, F. 1986. Noise. *Journal of Finance* 41, 529–44.

Black, F. and Scholes, M. 1973. Pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Brown, S., Goetzmann, W., Ibbotson, R. and Ross, S. 1992. Survivorship bias in performance studies. *Review of Financial Studies* 5, 553–80.

Campbell, J. and Shiller, R. 1988. The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195–228.

Chan, K. 1988. On the contrarian investment strategy. *Journal of Business* 61, 147–64.

Chan, L., Jegadeesh, N. and Lakonishok, J. 1996. Momentum strategies. *Journal of Finance* 51, 1681–713.

Chan, N., LeBaron, B., Lo, A. and Poggio, T. 1998. Information Dissemination and Aggregation in Asset Markets with Simple Intelligent Traders. Laboratory Technical Memorandum No. 1646. Cambridge, MA: MIT Artificial Intelligence.

Chopra, N., Lakonishok, J. and Ritter, J. 1992. Measuring Abnormal Performance: Do Stocks Overreact? *Journal of Financial Economics* 31, 235–86.

Cootner, P. 1962. Stock prices: random vs. systematic changes. *Industrial Management Review* 3, 24–45.

Cootner, P. 1964. *The Random Character of Stock Market Prices*. London: Risk Publications.

Copeland, T. and Mayers, D. 1982. The Value Line enigma (1965–1978): a case study of performance evaluation issues. *Journal of Financial Economics* 10, 289–322.

Cowles, A. 1960. A revision of previous conclusions regarding stock price behavior. *Econometrica* 28, 909–15.

Cowles, A. and Jones, H. 1937. Some a posteriori probabilities in stock market action. *Econometrica* 5, 280–294.

Damasio, A. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

de Finetti, B. 1937. La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré* 7, 1–68. English translation in *Studies in Subjective Probability*, ed. H. Kyburg and H. Smokler. New York: John Wiley & Sons, 1964.

DeBondt, W. and Thaler, R. 1985. Does the stock market overreact? *Journal of Finance* 40, 793–807.

Eldredge, N. and Gould, S. 1972. Punctuated equilibria: an alternative to phyletic gradualism. In *Models in Paleobiology*, ed. T. Schopf. San Francisco: Freeman, Cooper.

Elster, J. 1998. Emotions and economic theory. *Journal of Economic Literature* 36, 47–74.

Fama, E. 1963. Mandelbrot and the stable Paretian hypothesis. *Journal of Business* 36, 420–29.

Fama, E. 1965a. The behavior of stock market prices. *Journal of Business* 38, 34–105.

Fama, E. 1965b. Random walks in stock market prices. *Financial Analysts Journal* 21, 55–9.

Fama, E. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25, 383–417.

Fama, E. and Blume, M. 1966. Filter rules and stock market trading profits. *Journal of Business* 39, 226–41.

Fama, E. and French, K. 1988. Permanent and temporary components of stock prices. *Journal of Political Economy* 96, 246–73.

Farmer, D. 2002. Market force, ecology and evolution. *Industrial and Corporate Change* 11, 895–953.

Farmer, D. and Lo, A. 1999. Frontiers of finance: evolution and efficient markets. *Proceedings of the National Academy of Sciences* 96, 9991–2.

Fischoff, B. and Slovic, P. 1980. A little learning…: confidence in multicue judgment tasks. In *Attention and Performance, VIII*, ed. R. Nickerson. Hillsdale, NJ: Erlbaum.

Flavin, M. 1983. Excess volatility in the financial markets: a reassessment of the empirical evidence. *Journal of Political Economy* 91, 929–56.

French, K. and Roll, R. 1986. Stock return variances: the arrival of information and the reaction of traders. *Journal of Financial Economics* 17, 5–26.

Friedman, D. 1991. Evolutionary games in economics. *Econometrica* 59, 637–66.

Gervais, S. and Odean, T. 2001. Learning to be overconfident. *Review of Financial Studies* 14, 1–27.

Gigerenzer, G. 2000. *Adaptive Thinking: Rationality in the Real World*. Oxford: Oxford University Press.

Gigerenzer, G. and Selten, R. 2001. *Bounded Rational: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Gilles, C. and LeRoy, S. 1991. Econometric aspects of the variance-bounds tests: a survey. *Review of Financial Studies* 4, 753–92.

Grossberg, S. and Gutowski, W. 1987. Neural dynamics of decision making under risk: affective balance and cognitive-emotional interactions. *Psychological Review* 94, 300–18.

Grossman, S. and Shiller, R. 1981. The determinants of the variability of stock market prices. *American Economic Review* 71, 222–7.

Grossman, S. and Stiglitz, J. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70, 393–408.

Hirshleifer, J. 1977. Economics from a biological viewpoint. *Journal of Law and Economics* 20, 1–52.

Hirshleifer, D. and Luo, G. 2001. On the survival of overconfident traders in a competitive securities market. *Journal of Financial Markets* 4, 73–84.

Huberman, G. and Regev, T. 2001. Contagious speculation and a cure for cancer: a nonevent that made stock prices soar. *Journal of Finance* 56, 387–96.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *Journal of Finance* 45, 881–98.

Jegadeesh, N. and Titman, S. 2001. Profitability of momentum strategies: an evaluation of alternative explanations. *Journal of Finance* 56, 699–720.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica*

47, 263–91.

Keim, D. 1983. Size-related anomalies and stock return seasonality: further empirical evidence. *Journal of Financial Economics* 12, 13–32.

Kim, M., Nelson, C. and Startz, R. 1991. Mean reversion in stock prices? a reappraisal of the empirical evidence. *Review of Economic Studies* 58, 515–28.

Kindleberger, C. 1989. *Manias, Panics, and Crashes: A History of Financial Crises*. New York: Basic Books.

Kleidon, A. 1986. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94, 953–1001.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 62, 443–77.

Lakonishok, J. and Smidt, S. 1988. Are seasonal anomalies real? A ninety-year perspective. *Review of Financial Studies* 1, 403–25.

Lehmann, B. 1990. Fads, martingales, and market efficiency. *Quarterly Journal of Economics* 105, 1–28.

Leroy, S. 1973. Risk aversion and the martingale property of stock returns. *International Economic Review* 14, 436–46.

LeRoy, S. and Porter, R. 1981. The present value relation: tests based on variance bounds. *Econometrica* 49, 555–74.

Lichtenstein, S., Fischoff, B. and Phillips, L. 1982. Calibration of probabilities: the state of the art to 1980. In *Judgment Under Uncertainty: Heuristics and Biases*, ed. D. Kahneman, P. Slovic and A. Tversky. Cambridge: Cambridge University Press.

Lo, A. 1991. Long-term memory in stock market prices. *Econometrica* 59, 1279–313.

Lo, A., ed. 1997. *Market Efficiency: Stock Market Behavior in Theory and Practice*, 2 vols. Cheltenham: Edward Elgar Publishing Company.

Lo, A. 1999. The three P's of total risk management. *Financial Analysts Journal* 55, 87–129.

Lo, A. 2001. Risk management for hedge funds: introduction and overview. to appear in *Financial Analysts Journal* 57, 16–33.

Lo, A. 2002. Bubble, rubble, finance in trouble? *Journal of Psychology and Financial Markets* 3, 76–86.

Lo, A. 2004. The adaptive markets hypothesis: market efficiency from an evolutionary perspective. *Journal of Portfolio Management* 30, 15–29.

Lo, A. 2005. Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis. *Journal of Investment Consulting* 7, 21–44.

Lo, A. and MacKinlay, C. 1988. Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies* 1, 41–66.

Lo, A. and MacKinlay, C. 1990a. An econometric analysis of nonsynchronous trading. *Journal of Econometrics* 45, 181–212.

Lo, A. and MacKinlay, C. 1990b. Data snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3, 431–68.

Lo, A. and MacKinlay, C. 1990c. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* 3, 175–206.

Lo, A. and MacKinlay, C. 1999. *A Non-Random Walk Down Wall Street*. Princeton, NJ: Princeton University Press.

Lo, A. and Repin, D. 2002. The psychophysiology of real-time financial risk processing. *Journal of Cognitive Neuroscience* 14, 323–39.

Loewenstein, G. 2000. Emotions in economic theory and economic behavior. *American Economic Review* 90, 426–32.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–46.

Luo, G. 1995. Evolution and market competition. *Journal of Economic Theory* 67, 223–50.

Markowitz, H. 1991. Foundations of portfolio theory. *Journal of Finance* 46, 469–77.

Marsh, T. and Merton, R. 1986. Dividend variability and variance bounds tests for the rationality of stock market prices. *American Economic Review* 76, 483–98.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science*

4, 141–83.

Merton, R. 1987. On the current state of the stock market rationality hypothesis. In *Macroeconomics and Finance: Essays in Honor of Franco Modigliani*, ed. R. Dornbusch, S. Fischer and J. Bossons. Cambridge, MA: MIT Press.

Michener, R. 1982. Variance bounds in a simple model of asset pricing. *Journal of Political Economy* 90, 166–75.

Nelson, R. and Winter, S. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap Press of Harvard University Press.

Niederhoffer, V. 1997. *Education of a Speculator*. New York: John Wiley & Sons.

Odean, T. 1998. Are investors reluctant to realize their losses? *Journal of Finance* 53, 1775–98.

Osborne, M. 1959. Brownian motion in the stock market. *Operations Research* 7, 145–73.

Poterba, J. and Summers, L. 1988. Mean reversion in stock returns: evidence and implications. *Journal of Financial Economics* 22, 27–60.

Ramsey, F. 1926. Truth and probability. In *Foundations of Mathematics and Other Logical Essays*, ed. R. Braithwaite. New York: Harcourt Brace & Co.

Richardson, M. 1993. Temporary components of stock prices: a skeptic's view. *Journal of Business and Economics Statistics* 11, 199–207.

Roberts, H. 1959. Stock-market ‘patterns’ and financial analysis: methodological suggestions. *Journal of Finance* 14, 1–10.

Roberts, H. 1967. Statistical versus clinical prediction of the stock market. Unpublished manuscript, Center for Research in Security Prices, University of Chicago.

Roll, R. 1983. Vas is das? The turn-of-the-year effect and the return premia of small firms. *Journal of Portfolio Management* 9, 18–28.

Roll, R. 1984. Orange juice and weather. *American Economic Review* 74, 861–80.

Rolls, E. 1999. *The Brain and Emotion*. Oxford: Oxford University Press.

Rosenberg, B., Reid, K. and Lanstein, R. 1985. Persuasive evidence of market inefficiency. *Journal of*

Portfolio Management 11, 9–17.

Rozeff, M. and Kinney, W., Jr. 1976. Capital market seasonality: the case of stock returns. *Journal of Financial Economics* 3, 379–402.

Rubinstein, A. 1998. *Modeling Bounded Rationality*. Cambridge, MA: MIT Press.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7, 407–25.

Samuelson, P. 1947. *Foundations of Economics Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 41–9.

Sargent, T. 1993. *Bounded Rationality in Macroeconomics*. Oxford: Clarendon Press.

Savage, L. 1954. *Foundations of Statistics*. New York: John Wiley & Sons.

Schumpeter, J. 1939. *Business Cycles: A Theoretical, Historical, And Statistical Analysis of the Capitalist Process*. New York: McGraw-Hill.

Shefrin, M. and Statman, M. 1985. The disposition to sell winners too early and ride losers too long: theory and evidence. *Journal of Finance* 40, 777–90.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71, 421–36.

Simon, H. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69, 99–118.

Simon, H. 1982. *Models of Bounded Rationality*, 2 vols. Cambridge, MA: MIT Press.

Starmer, C. 2000. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38, 332–82.

Thaler, R., ed. 1993. *Advances in Behavioral Finance*. New York: Russell Sage Foundation.

Tversky, A. and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science* 211, 453–8.

Wilson, E. 1975. *Sociobiology: The New Synthesis*. Cambridge, MA: Belknap Press of Harvard

University Press.

How to cite this article

Lo, Andrew W. "efficient markets hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000050> doi:10.1057/9780230226203.0454

The New Palgrave Dictionary of Economics Online

egalitarianism

Harry Brighouse and Adam Swift

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys a variety of egalitarian theories. We look at a series of different answers to the question of what the metric of justice should be. Then we survey different interpretations of the egalitarian distributive rule, including ‘equality’, ‘prioritizing benefit to the least advantaged’ and ‘sufficiency’. Theories also differ by whether they see equality as properly holding within social institutions or being a principle that applies more cosmically. Finally, we observe that egalitarian theories differ as to the weight they grant to egalitarian values relative to other values.

Keywords

compensation; distributive justice; Dworkin, R.; egalitarianism; envy; equality; equality of capabilities; equality of opportunity; equality of resources; happiness; interpersonal well-being comparisons; justice; luck egalitarianism; opportunity cost; positional goods; prioritarianism; property rights; Rawls, J.; relational egalitarianism; Sen, A.; sufficientarianism; value pluralism; welfare; well-being

Article

All modern political theories assume that persons are in some relevant sense moral equals, entitled to equal concern, respect or treatment, and that a theory of justice must interpret and reflect that moral equality. This commitment is sometimes dubbed the ‘egalitarian plateau’, and it has been a common foundational moral assumption since Locke. Contemporary theories differ in how they *interpret* the egalitarian plateau. Two kinds of theory of justice are usually counted as egalitarian. Theories of distributive equality concern themselves with the relative standing of individuals in the distribution of benefits and burdens; theories of relational equality concern themselves with the relative standing of individuals when they face each other in the public sphere.

The metric

One key question concerns the metric of equality: what, precisely, is it that egalitarians should seek to equalize? The literature falls into three main camps. Resourcists argue that people should be equal in the space of resources, meaning that they should have equal opportunity for achieving holdings of alienable goods. How are holdings priced? Ronald Dworkin imagines a hypothetical auction in which persons with equal holdings of some currency bid for available goods until markets clear (Dworkin, 2000). The distribution after the auction is equal if no one prefers anyone else's bundle of goods to her own; the distribution is then said to pass the ‘envy test’. The intuitive idea is that the price of some good is set by the opportunity cost to others of that good. We have to tailor our preferences to our resources; equality is achieved when all face the same budget constraint, not when all achieve equal satisfaction.

Equality of resources has difficulty with the intuition that those with less socially valued talent, and in particular those with serious impairments, should receive compensation. Two strategies are available. One is to adopt a view that talent is socially constructed, so that much of the disadvantage faced by the less talented and the impaired is a consequence not of their lack of talent but of the fact that social institutions are maladapted to their natural endowments (Pogge, 2003). This view allows resourcists to call for the reform of social institutions in the name of equality, without demanding compensation for impairments. The problem with this strategy is that some mental and physical impairments *intrinsically* cause disadvantage; there is no feasible set of social arrangements that would not make it more difficult for people with the impairments to derive satisfaction from resources. So an alternative strategy is to make the cut between persons and resources in a different place, regarding talents as resources and disabilities as resource-deficits. Dworkin's own version of this strategy proposes compensating the less talented with additional income, the amount calculated by looking at the insurance that talented individuals would have bought against a lack of talents if they had no knowledge of their probability of having the talents.

An alternative metric is welfare; egalitarians of welfare would seek to equalize levels of welfare (understood sometimes as idealized preference satisfaction, sometimes in terms of internal states such as happiness). This view handles talent-inequality in a straightforward manner; the less talented and the disabled should be compensated up to the level where they enjoy as much welfare as anyone else. But it faces the problem that there is no reason for people to moderate their preferences; since welfare is a direct target, those with expensive tastes receive more resources than those with inexpensive tastes, which is widely regarded as intuitively unfair. An alternative view – equality of opportunity for welfare – deals with this problem by seeking equality of welfare except when inequalities are the result of voluntary well-informed choices rather than bad luck or circumstances outside the agent's control (Arneson, 1989). Again, the less talented are straightforwardly compensated for the way in which they find it harder than others to derive satisfaction, but those who cultivate expensive tastes are not. However, those with non-cultivated expensive preferences are also compensated, even if they could easily be overcome; this view does not see lack of talent, and disability in particular, as morally more urgent than expensive preferences. (See Roemer, 1986, for an argument that equality of resources implies equality of welfare.)

All of the views deploying an ‘opportunity’ metric, including Dworkin's resourcist view, presume the desirability of holding people accountable for their voluntary choices, but compensating them for deficits that are beyond their control. Views of this kind are sometimes referred to as varieties of ‘luck egalitarianism’. Inequalities resulting from voluntary choice are acceptable because they reflect a deeper

sense in which we are equal as moral agents; choice legitimizes inequality, brute luck does not. (For an elegant attempt both to conceptualize and operationalize equality of opportunity *tout court*, see Roemer, 1998.)

The main rival account – namely, the capabilities approach developed by Amartya Sen and Martha Nussbaum – focuses on the *preconditions* of agency (Sen, 1999; Nussbaum, 2000). Equality of capabilities demands that people be equal in the space of the functionings or livings that they are substantively able to achieve. Walking is a functioning, so are eating, reading, mountain climbing, and chatting. 'The concept of functionings … reflects the various things a person may value doing or being – varying from the basic (being adequately nourished) to the very complex (being able to take part in the life of the community)' (Sen, 1999, p. 75). But when we make interpersonal comparisons of well-being we should find a measure that incorporates references to functionings but also reflects the intuition that what matters is not merely achieving the functioning but being free to achieve it. So we should look at 'the freedom to achieve actual livings that one can have a reason to value' (Sen, 1999, p. 73) or, to put it another way, 'substantive freedoms – the capabilities – to choose a life one has reason to value'. The idea is that people should be equal in this space.

The capabilities approach avoids the problems of the standard welfarist approaches by focusing on *choice* (thus treating inequalities arising from voluntary choices differently from those arising from circumstances). It avoids the difficulty resourcist accounts have with unequal talent by focusing on *functionings*; talent deficits are compensated for by looking not at what others would pay to avoid them but at the valuable activities the deficits deprive people of access to. Some theorists place the capabilities account in the welfarist camp (Williams, 2002) but it is not implausible to think of it as a variant of resourcism, distinguished by its approach to the valuation of talents.

A major recent development in the debates about egalitarianism has involved criticisms of luck egalitarianism. Each of the luck egalitarian principles, taken alone, imposes heavy costs on those who endure misfortunes for which they can be held responsible, even if those costs place the agent below the threshold for full participation in social affairs. An alternative has developed which is best described as 'relational egalitarianism'. Relational egalitarianism is not directly concerned with equality in terms of the distribution of any particular currency, but endorses the idea that individuals should have equal standing in the public sphere. This vague idea has several instantiations. Elizabeth Anderson (1999, p. 304) talks of seeking 'a social order in which persons stand in relations of equality'; Nancy Fraser (1998, p. 30) says that 'Justice requires social arrangements that permit all (adult) members of society to interact with one another as peers'. Both fill out their theories with more details. According to Fraser (1998, p. 24), 'It is unjust that some individuals and groups are denied the status of full partners in social interaction, simply as a consequence of institutionalized patterns of interpretation and evaluation in whose construction they have not equally participated and that disparage their distinctive characteristics or the distinctive characteristics assigned to them'. A third variant of relational egalitarianism spells it out specifically in terms of political equality, the idea being that it is particularly important that people enjoy equal availability of or opportunity for political power or influence (Christiano, 1995). This variant is typically less hostile than other variants to luck egalitarianism.

Each of the views reviewed in this section allows inequality along some dimensions. Relational egalitarianisms allow such inequalities of income, wealth, welfare or capabilities as are compatible with equal political influence, or interaction as peers, or 'equal opportunity for participation as a peer'. These

permitted inequalities may be great or very small, and how great or small may vary by social context. Principles demanding equality of opportunity are consistent with great inequalities in outcome, and consistent also with some being very badly off in absolute terms. While equality of opportunity conceptions place no limit on how badly off someone may be as a result of her own imprudent choices, equality of social standing demands that no one fall below the threshold needed for equal participation, even if she makes numerous imprudent choices.

The distributive rules

Do egalitarians even care about *equality*? Principles demanding equality of X seem vulnerable to an obvious objection. In some dynamic situations it is possible to produce more of X by distributing X unequally, and to ensure that even those with least have more than under an equal distribution. For example, we can sometimes produce more wealth by judiciously attaching higher income to more productive positions in the economy, and to longer work hours; the higher income acts both as a signal and as an incentive to produce more. That greater production can be turned to the benefit of those with least. But, the objection goes, it would be perverse to prefer an equal situation in which everyone has less to one in which everyone has more, even if we have to sacrifice equality for the sake of that additional product.

This is known as the ‘levelling down’ objection to equality. Egalitarians make two distinct responses. The first is to concede the argument, abandoning ‘equality’ and replacing it with ‘giving priority to the interests of the least advantaged’. John Rawls's difference principle, which states that ‘social and economic inequalities are to be arranged to the maximum benefit of the least advantaged’, embodies one variant of this response, a variant that gives *absolute* priority to the prospects of the least advantaged (Rawls, 1971; 2001). A weaker variant in this family of views, usually known as ‘prioritarianism’, simply says that it is more urgent to provide benefits to those with less advantage than to those with more (Parfit, 2000).

An alternative response is to assert value pluralism. This response acknowledges that priority to the least advantaged is an important value and perhaps more important than equality, so that when it comes to policy or action prioritarian principles should govern. But it says that equality nevertheless matters some; there is one way in which an unequal distribution is worse than an equal distribution, even if, all things considered, it is better; the way in which it is worse is that it is unequal and for that reason unfair (Temkin, 2002). This response is bolstered by the observation that there is nothing eccentric about endorsing a principle that values distributions that benefit nobody; the retributive principle of proportionality between punishment and crime, for example, calls for harming the criminal even when there is no gain to anyone else in harming him.

Some reject principles of equality and priority on the grounds that all that matters for the purposes of justice is that all have enough. Sufficientarian theories are not usually counted as within the egalitarian family, because they eschew any fundamental concern with relativities. Relativities may matter in determining what is enough for people to live a decent life in any given social environment, but ultimately what matters is not where someone ranks in the distribution of resources (or anything else) but whether she has enough. However, as suggested above, sufficientarian principles also have a place in some variants of egalitarianism. While relational egalitarianism places no principled limits on the level

of material or welfare inequality, and gives no general priority to the least advantaged, it does set a floor – all must have sufficient resources to be full participants in social interaction. Equality of political influence demands that all have sufficient resources, personal and financial, to play an equal role in political life, but, as long as it is possible to insulate politics from residual inequalities of wealth, it is not concerned with equalizing or prioritizing benefit to the least advantaged.

Many theories of justice that do not fit the above characterizations of egalitarianism nevertheless incorporate some elements of egalitarian thinking. John Rawls's theory of justice, for example, prioritizes the principle that certain basic liberties (not including strong property rights) be equally distributed, then demands that within that constraint fair equality of opportunity should be implemented, and then that social and economic inequalities be arranged to the greatest benefit of the least advantaged in so far as that is possible without jeopardizing the equal liberty and fair equality of opportunity principle (Rawls, 1971; 2001). Michael Walzer's (1983) theory of ‘complex equality’ takes seriously widely shared intuitions that different goods are subject to different distributive rules. For example, while income should be distributed according to productive contribution, as will tend to result from market interactions, the inequalities this norm generates should be prevented from translating into unequal access to certain key goods like health care and educational opportunities, the distribution of which should be governed by need and the requirements of equal opportunity respectively. It is unclear in what sense Walzer's ‘complex equality’ is genuinely an egalitarian position, since it is in principle consistent with unequal and coinciding distributions of all goods that are not themselves governed by egalitarian norms.

Priority and equality coincide in practice for one class of goods: positional goods. These have the property that the contribution an individual's share of the good makes to her absolute position is determined by how much of the good she has relative to others. The credentialing aspect of education is a paradigm case; how useful a degree is in landing a job (as opposed to the learning one achieved in the process of getting the degree) depends entirely on the credentials of one's competitors for that job. (Other cases are detailed in Hirsch, 1976.) Those who give priority to the worst-off will countenance inequalities in positional goods only in so far as they are required by or result in the least advantaged benefiting overall (Brighouse and Swift, 2006).

The scope of equality

Whatever the right distribuendum, and whatever the appropriate distributive principle, it is a further question who should be equal to whom. Some limit the application of their egalitarianism to members of the same society or system of cooperation, or to those subject to the same coercive structure (Nagel, 2005), or hold that it is states that owe their citizens a particular duty to treat them with equal concern and respect (Dworkin, 2000). Others believe that egalitarian principles should apply to all human beings, irrespective of the relations that obtain between them. If we restrict the application of egalitarian principles to schemes of cooperation, that does not exclude the possibility of a global egalitarianism, since most now accept that in the modern world social cooperation extends well beyond national boundaries (Julius, 2006). But consider this version of Derek Parfit's divided world case. All the people in A are half as well off as all the people in B, but A and B have no knowledge of or contact with each other (Parfit, 2000). Is there anything regrettable from the perspective of injustice about this inequality?

If so, then the scope of justice is cosmic, not simply social. In the stated version of the divided world case this difference is motivationally inert, since the people in B do not have the relevant knowledge. But, if they did, cosmic egalitarianism would give them a reason to try to find a way to contact and interact with the people on A, while intra-societal egalitarianism would provide them with no such reason.

The divided world case brings out another difference in orientation. Where members of A and B have no interaction, or even knowledge of each other, equality can be valued only intrinsically rather than because of its effects on members of A or B. Often, however, inequality with respect to some goods is devalued, and equality valued, instrumentally, because of its absolute effects on those subject to the unequal distribution – usually its effects on the relatively disadvantaged. Thus, for example, economic inequalities are thought to undermine the fairness of legal or political processes, or occupational or other status hierarchies are claimed to harm the health of those on the lower rungs. Those who value equality intrinsically would hold that there is a reason to level down for the sake of equality or fairness, whereas instrumental egalitarians might seek the more equal distribution of some goods, not for egalitarian reasons *stricto sensu*, but to eliminate the bad effects of certain kinds of inequality.

The subject of justice

A further dividing line between egalitarians concerns the subject of justice. Rawls stipulates that the subject is the ‘basic structure of society’, which consists of some of the central, interaction-shaping institutions of a society: for example, the constitution, the legally recognized forms of property, the structure of the economy, the design of the legislature, and the judiciary. The idea is that these institutions govern the division of the advantages that accrue from social cooperation, and they assign the basic rights and responsibilities to citizens. So a society is just when those institutions are arranged according to the correct principles.

Rawls officially exempts individual actions and motives from evaluation from the perspective of egalitarian justice, as long as individuals obey the rules set by a just basic structure. But this has the consequence that a society in which talented individuals take advantage of the prerogatives not to serve the least advantaged that are built into the principles that he thinks justice requires of coercive institutions is no less just than one in which they are much more strongly motivated by the desire to benefit the least advantaged through their choices regarding work. A society with an egalitarian governing ethos, on this view, is no more just than one without, even when the least advantaged are much better off. But the motivations and actions of talented individuals affect the prospects and status of others in ways that have ‘profound and pervasive influence on persons’ (Rawls, 2001, p. 55), which is Rawls's central reason for focusing on the basic structure. So some egalitarians regard justice as commenting not only on the broad coercive outline of society, but also on less officially coercive institutions such as a society's ethos (Cohen, 1997). For a powerful defence of an account intermediate between Cohen's and Rawls's, see Julius, 2003).

Other values

Most egalitarian theorists are value pluralists; they believe that equality (or priority) of their preferred

metric matters, but so do other principles. Observing that equality or priority is sometimes in conflict with liberty or privacy or efficiency does not require us to reject one of the conflicting values. It requires us, instead, to evaluate reasons for considering one of the values more morally important than the others, and, in the light of that evaluation, to establish which should give way in different conflicts. Unless the relationship between values is one of lexical priority (in which case the prior value always trumps subordinate values, which can be pursued only when there is no conflict), different trade-offs between values will be mandated in different conflicts. But lexical priority is unlikely to hold between genuine values. If a value matters *at all*, it is hard to believe it could never be the case that a very large amount of it was greater than a very small amount of a conflicting value *however great that conflicting value is*.

See Also

- equality of opportunity
- ethics and economics
- liberalism and economics
- libertarianism
- Pareto efficiency
- satisficing
- Sen, Amartya

Bibliography

Anderson, E. 1999. What is the point of equality? *Ethics* 109, 287–337.

Arneson, R. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56, 77–93.

Arneson, R. 2002. Egalitarianism. *Stanford Encyclopedia of Philosophy*. Online. Available at http://plato.stanford.edu/entries/egalitarianism, accessed 14 October 2006.

Brighouse, H. and Swift, A. 2006. Equality, priority and positional goods. *Ethics* 116, 471–97.

Christiano, T. 1995. *The Rule of the Many*. Boulder, CO: Westview Press.

Cohen, G.A. 1997. Where the action is: on the site of distributive justice. *Philosophy and Public Affairs* 26, 3–30.

Dworkin, R. 1981. What is equality? Part 2: equality of resources. *Philosophy and Public Affairs* 10, 283–345.

Dworkin, R. 2000. *Sovereign Virtue*. Cambridge, MA: Harvard University Press.

Fraser, N. 1998. Social justice in the age of identity politics. *The Tanner Lectures on Human Values*. Stanford University, 29 April–2 May. Online. Available at http://www.tannerlectures.utah.edu/lectures/Fraser98.pdf, accessed 14 October 2006.

Hirsch, F. 1976. *Social Limits to Growth*. London: Routledge & Kegan Paul.

Julius, A.J. 2003. Basic structure and the value of equality. *Philosophy and Public Affairs* 31, 321–55.

Julius, A.J. 2006. Nagel's atlas. *Philosophy and Public Affairs* 34, 176–92.

Nagel, T. 2005. The problem of global justice. *Philosophy and Public Affairs* 33, 113–47.

Nussbaum, M. 2000. *Women and Human Development*. Cambridge: Cambridge University Press.

Parfit, D. 2000. Equality or priority? In *The Ideal of Equality*, ed. M. Clayton and A. Williams. London and New York: Palgrave Macmillan and St Martin's Press.

Pogge, T. 2003. Can the capability approach be justified? *Philosophical Topics* 30(2), 167–228.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 2001. *Justice as Fairness*. Cambridge, MA: Harvard University Press.

Roemer, J.E. 1986. Equality of resources implies equality of welfare. *Quarterly Journal of Economics* 101, 751–84.

Roemer, J.E. 1998. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Sen, A. 1999. *Development as Freedom*. New York: Knopf.

Temkin, L. 2002. Equality, priority, and the levelling down objection. In *The Ideal of Equality*, ed. M. Clayton and A. Williams. London and New York: Palgrave Macmillan and St Martin's Press.

Walzer, M. 1983. *Spheres of Justice*. New York: Basic Books.

Williams, A. 2002. Dworkin on capability. *Ethics* 113, 23–39.

How to cite this article

Brighouse, Harry and Adam Swift. "egalitarianism." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000273> doi:10.1057/9780230226203.0455

The New Palgrave Dictionary of Economics Online

Einaudi, Luigi (1874 – 1961)

F. Caffè

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic liberalism; Einaudi, L.; Mill, J. S.; public finance; tax incentives for saving; taxable income

Article

An outstanding Italian economist and influential figure on the broader political and cultural scene, Einaudi was born in Carru (Piedmont) on 24 March 1874 and died in Rome on 30 October 1961. He graduated in law from Turin in 1895 and then, while continuing with this studies, embarked on a career in journalism. The success he achieved in both fields underlined his rare talent and his endless capacity for work. In fact, his academic progress was so rapid that in 1907 he was appointed as professor of public finance at the University of Turin. Meanwhile, he wrote articles for the most influential Italian daily newspaper of the period, the *Corriere delle Serra*, which not only brought him national recognition but also earned him the reputation of ‘educator’ of the entire country. He became a member of the Senate in 1919, but retired from all political and public activity with the advent of fascism. Towards the end of the First World War he went into exile in Switzerland. On his return, he was appointed Governor of the Bank of Italy (1945), Vice-President of the Cabinet and Minister in charge of the Budget (1947), and was finally elected President of the Republic of Italy (1948–1955). At the end of his seven-year presidential term of office, he was made a life member of the Senate.
The most important aspect of Einaudi's achievements is the use he made of his academic and journalistic ability, as foundations for his activity as a statesman and politician. In addition, close study of his strictly scientific works reveals the extent to which he drew on the wealth of knowledge and experience which he had gained also in other fields. The 3,800 recorded items of Einaudi's works cover such a wide range of interests that it is necessary here to concentrate on his contributions to the study of public finance and his ideas on economic policy. Einaudi's main contributions to the study of public finance were investigations, based on the classical ideas of John Stuart Mill, which gave a solid logical basis to the principle of the exclusion of savings from taxable income; his research into the theory of capitalization of taxation; his critical and constructive contributions on the effects of certainty and stability of fiscal principles; his important analysis of the concept of taxable income which he identified with normal

income, or, in other words, with the average income potentiality of the person subject to taxation. Einaudi's position vis-à-vis public intervention in the economy was not hostile in principle, though he undoubtedly took a limited view of state interference in economic life. Since, for Einaudi, ‘All liberties were jointly liable’, autonomous sources of income were a necessity to prevent people from being subjected to a single centralizing order of the state. He asserted this during the 20 years of fascism, when he continued to teach with the same independence of mind and without compromising his fidelity to economic liberalism. Even though Einaudi had been stressing the usefulness of productive public expenditure since 1919, he showed a singular lack of comprehension of the Keynesian contribution, in the belief that it would be an inevitable cause of inflation.

Selected works

On Luigi Einaudi himself there is a *Bibliografia degli scritti* edited by Luigi Firpo under the auspices of the Bank of Italy, Turin, 1971. It is useful to divide his work into the three main areas which he outlined: theory, politics and history. Representative works of the three sections are as follows:

1912. *Intorno al concetto di reddito imponibile e di un sistema di imposte sul reddito consumato*. Turin: V. Bona.

1919. *Osservazioni critiche intorno alla teoria dell'ammortamento dell'imposta e teoria delle variazioni nei redditi e nei valori capitali sussequenti all'imposta*. Turin: Fratelli Bocca.

1929. *Contributo all ricerca della ‘ottima imposta’*. Milan: Bocconi.

1938. *Miti e paradossi delli giustizia tributaria*. Turin: Luigi Einaudi.

The following handbooks are available:

1914. *Corso di scienza delle finanze*. Turin: Tip. e Bono.

1932–66. *Principi di scienza delle finanze*. Turin: La Riforma Sociale.

1932. *Il sistema tributario italiano*. Turin: La Riforma Sociale.

With reference to the history of finance and the history of ideas see:

1908. *La finanza sabauda all'aprirsi del secolo XVIII e durante la guerra di successione spagnola*. Turin: Società Tip. Editrice Nazionale.

1927. *La guerra e il sistema tributario italiano*. Bari: Laterza.

1953. *Saggi bibliografici e storici intorno alle dottrine economiche*. Rome: Ediz. Storia e Litteratura.

Einaudi's journalistic work has been largely collected in eight volumes comprising the *Cronache economiche e politiche di un trentennio* (1893–1925), Turin: Ed. Einaudi, 1959–65, and in *Lo scrittoio del Presidente 1948–1955*, Turin: Ed. Einaudi, 1956. For many years Einaudi was Italian correspondent for the *Economist*.

How to cite this article

Caffè, F. "Einaudi, Luigi (1874–1961)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000054> doi:10.1057/9780230226203.0456

The New Palgrave Dictionary of Economics Online

Eisner, Robert (1922 – 1998)

James K. Galbraith

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robert Eisner, a leading American macroeconomist and theorist of the investment function, was an architect of the Keynesian ascendancy in post-war America. He developed the accounting foundations of Keynesian macroeconomics, finally producing a Total Income System of Accounts. His ideas found application in his later, policy-oriented writings on the budget deficit, the current account, and the Social Security system. His embrace of capital budgeting underpinned a strong advocacy of liberal expenditure on infrastructure, education, and research and development. He was throughout motivated by a commitment to larger social goals, especially full employment, peace, and justice.

Keywords

American Economic Association; budget deficits; capital budgeting; China; one-child policy; collective bargaining; depreciation allowance; Economists Allied for Arms Reduction; Eisner, R.; full employment; invariant multiplier; investment function; Keynesian Revolution; liquidity preference; liquidity trap; natural rate of unemployment (NAIRU); Office of Price Administration; peace economics; permanent income hypothesis; replacement costs; Social Security; Total Income System of Accounts; unit relative price elasticity

Article

Robert Eisner, a leading American macroeconomist and theorist of the investment function, graduated in history from College of the City of New York in 1940, took an MA in sociology from Columbia University in 1942 and, following service in the army and the Office of Price Administration, a Ph.D. in economics under Fritz Machlup at Johns Hopkins University in 1951. He joined the faculty of Northwestern University in 1952, rising to hold the William R. Kenan Professorship of Economics from 1974 until his retirement in 1994. He served as President of the American Economic Association in 1988. Eisner was an architect of the Keynesian ascendancy in post-war America. Much of his work was

devoted to technical developments in that tradition; his singular distinction lay in taking the accounting foundations of Keynesian macroeconomics seriously and in developing their implications with utmost rigour. This thread runs through his writing from his earliest papers on the ‘Invariant Multiplier’, the permanent income hypothesis, liquidity preference and the liquidity trap. It reaches its apogee in his work on a Total Income System of Accounts (TISA). It suffuses his later, policy-oriented writings on the meaning and implications of deficits in the budget, current account, and Social Security system.

No shrinking violet, Eisner liked to call his shots. Thus, H. S . Houthaker ‘has not performed [a] test correctly’; ‘Bronfenbrenner and Mayer... confound... issues of elasticity with those of slope’; ‘Re-estimation with Pifer's data and application of appropriate statistical tests contradict Pifer's conclusions’ (1998a, pp. 8, 27, 48). The tone is ever tactful, the intent always the pursuit of truth, the subtext a certain delight in finding the exact, fatal weakness of an opposing view. Late in his life, this author heard Eisner speak to a room of senior officials in China on the error and futility of the one-child policy, a delicate issue which he raised in the same spirit and with deeply impressive effect.

Underpinning his technical precision lay an unflagging commitment to larger social goals, especially full employment, peace, and justice. Eisner actively advocated all three throughout his career, but especially in the later years when he appeared frequently on the opinion pages of the *Wall Street Journal*, as a leading director of Economists Allied for Arms Reduction, and in causes devoted to the advancement of women in the economics profession.

For instance, in a 1952 paper in the *American Economic Review* (1998a, 106–17) Eisner analysed the relationship of replacement costs to depreciation allowances in a growing economy. In doing so he called attention to the fact that growth in the latter usually exceeded that in the former, resulting in reported profits that were understated for purposes of both taxation and collective bargaining. Pointedly, he suggested the work ought to interest both revenue officers and trade unionists.

Yet Eisner's views were often unfashionable and politically inconvenient. In important papers in the 1980s, at a time when Democrats had taken the veil of fiscal virtue, he undertook with Paul Pieper to show that (among numerous other difficulties with budget accounting) inflation had rendered the deficit meaningless, introducing vast inconsistencies between the nominal budget deficit and the change in the real public debt. Thus, the Reagan deficits were far smaller than normally supposed, while those of Carter were surpluses in real terms – likely to produce fiscal drag and so to bear partial responsibility for the stagnation of those years. Correctly accounting for inflation, Eisner argued, might have forestalled the new classical critique that led many in those years to abandon Keynesian principles.

A closely related cause was the misunderstanding of ‘national saving’ and the fallacious popular argument that to reduce deficits would lead to increased capital formation. In 1995, Eisner argued that to take the accounting relation between public and private saving

> as evidence that reducing the federal deficit must raise national saving should be recognized, on even the slightest reflection, as patently absurd. It is startlingly akin to the assumption, more than half a century ago, that saving and investment would be increased if we all undertook to save more by consuming less. Perhaps! But that is exactly the proposition to be proved, or supported by empirical evidence, not assumed. (1998a, p. 322)

Second only to correct reasoning, evidence mattered. In the 1990s Eisner took up arms against the ‘governing myth’ of economic policy, the natural rate of unemployment introduced by Friedman and

Phelps in 1968. From this strangely self-damaging justification for perpetually high unemployment, Eisner hoped for a ‘NAIRU escape’. His method was largely econometric, and in what may have been his final paper, published in 1998 (1998a, pp. 454–87), he argued that a separate analysis of low-unemployment cases showed no relationship between full employment and rising inflation. This position was to be vindicated dramatically in the two years following his death.

Eisner embraced capital budgeting, so that the liabilities acquired by the government might be properly offset against corresponding assets. This position helped underpin a strong advocacy of liberal expenditure on infrastructure, education, and research and development. It also provides one bridge between the Keynesian Eisner and his counterpart, the theorist of investment, public finance, and peace economics and stalwart defender of Social Security, all of which he was.

Eisner's investigations of investment involved pioneering use of corporate records. They permitted cross-section analysis of firm decisions, showing that the concepts of macro models, such as the accelerator, operated differently on firms from different industries or with differing recent growth histories. In numerous studies, Eisner criticized neoclassical investment theories. Rejecting the notions of a desired capital stock and unit relative price elasticity, he adhered to a Keynesian relation of investment to expected profitability and of expected profits to the rate of growth. An important theme in this work concerns the appropriate level of aggregation at which to take measurements. Eisner found that firms appropriately assess the growth of their own industry to be the most relevant to profit prospects, not the inherently variable growth of individual firms or the potentially irrelevant growth of generalized aggregate demand.

Eisner's Total Income System of Accounts marked the peak of his campaign to rationalize economic measurement and theory. The importance of changing household relations appears vividly in his initial motivation for this work: ‘What happens to income, output, and productivity when clotheswashing moves from the washtub and the professional laundry to the laundromat and to the automatic washer and dryer...?’ (1998b, p. 188). Particularly noteworthy is capital accumulation by households in a country where transportation is provided mainly and increasingly by private car. The challenge of TISA remains to be taken up by most economists and national income statisticians.

Finally, midway through the Vietnam War Eisner deflated the view that President Johnson might have forestalled inflation by raising taxes; the only sure way to that end, he showed, would have been to avoid the war. This insight led to papers on the ‘staggering cost’ of the Vietnam war, much in the spirit of total accounts, and on post-cold war disarmament. Equally, to the end of his life Robert Eisner defended Social Security from all those who would cut it. Spurious and persistent allegations of financial ‘crisis’ notwithstanding, he believed that a rich and civilized society can, and should, provide decent incomes and care for its old.

See Also

- government budget constraint
- labour supply
- national accounting, history of
- Social Security in the United States
- war and economics

Selected works

1986. *How Real is the Federal Deficit?* New York: Free Press.

1994. *The Misunderstood Economy: What Counts and How to Count It*. Cambridge, MA: Harvard Business School Press.

1997. *The Great Deficit Scares: The Federal Budget, Trade, and Social Security*. New York: The Century Foundation.

1998a. *The Keynesian Revolution, Then and Now: The Selected Essays of Robert Eisner*, vol. 1. Cheltenham: Edward Elgar.

1998b. *Investment, National Income and Economic Policy: The Selected Essays of Robert Eisner*, vol. 2. Cheltenham: Edward Elgar.

How to cite this article

Galbraith, James K. "Eisner, Robert (1922–1998)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000232> doi:10.1057/9780230226203.0457

The New Palgrave Dictionary of Economics Online

elasticities approach to the balance of payments

Murray C. Kemp

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

elasticities approach to the balance of payments; excess demand

Article

The substance of a theory is independent of the manner in which it is dressed. In particular, it is a matter of style only whether or not formulae are expressed in terms of elasticities of demand and supply, or in terms of ordinary derivatives. To speak of an ‘elasticities approach’ to the balance of payments is therefore to speak no sense at all.

However, behind the nonsensical label there hides a coherent and distinctive theory of what determines the response of a country's balance of payments to parametric changes in its rate of exchange, that is, to changes in the terms on which its currency exchanges for other currencies. The theory goes back to a paper published by Charles Bickerdike (1920).

Consider a simplified world containing just two countries (the ‘home’ country and the ‘foreign’) and producing and trading just two commodities. Let R be the price of foreign currency in terms of home currency, let p_i be the home price of the ith commodity in terms of home currency (so that, in arbitrage equilibrium, $p_i^* \equiv p_i / R$ is the foreign price of the commodity in terms of foreign currency), and let B be the home balance of trade in terms of foreign currency. Then, writing $z_i(p_i)$ and $z_i^*(p_i^*)$ as the home and foreign excess demands for the ith commodity, Bickerdike's model of the balance of payments reduces to the system of three equations

$$z_i(p_i) + z_i^*(p_i / R) = 0 (i = 1, 2) B = -(1 / R)[p_1 z_1(p_1) + p_2 z_2(p_2)] \quad (1)$$

In this system the rate of exchange R is treated as a parameter and p_1, p_2 and B as variables to be determined. Differentiating (1) with respect to R, solving for dB and the dp_i, and converting to elasticities, we obtain

$$\mathrm{d}B = \left\{ -p_2^* z_2^* \left[\frac{\eta_1^*(1+\eta_1)}{\eta_1^* - \eta_1} - \frac{\eta_2^*(1+\eta_2)}{\eta_2^* - \eta_2} \right] - B \right\} \frac{\mathrm{d}R}{R} \tag{2}$$

and

$$\frac{\mathrm{d}p_i}{p_i} = \frac{\eta_i^*}{\eta_i^* - \eta_i} \frac{\mathrm{d}R}{R}, \; i = 1, 2 \tag{3}$$

where $\eta_i \equiv (\mathrm{d}z_i / \mathrm{d}p_i)(p_i / z_i)$ and $\eta_i^* \equiv (\mathrm{d}z_i^* / \mathrm{d}p_i^*)(p_i^* / z_i^*)$. In the special case in which B is initially zero, (2) takes the simpler form

$$\mathrm{d}B = -p_2^* z_2^* \left[\frac{\eta_1^*(1+\eta_1)}{\eta_1^* - \eta_1} - \frac{\eta_2^*(1+\eta_2)}{\eta_2^* - \eta_2} \right] \frac{\mathrm{d}R}{R}. \tag{2'}$$

Equation (2) is often referred to as the Bickerdike–Robinson–Metzler formula: however, the role of Robinson (1947) and of Metzler (1949) was that of expositor only.

Suppose for concreteness that the home country exports the first commodity and imports the second, so that η_1 and η_2^* are export-supply elasticities and η_2 and η_1^* import-demand elasticities. Suppose further that all marginal propensities to buy are positive, so that η_1 and η_2^* are positive, η_2 and η_1^* negative. Then for the balance of payments to improve in response to devaluation it suffices that the sum of the two import demand elasticities exceed 1 in magnitude, that is, that the Marshall–Lerner condition be

satisfied. Thus eq. (2′) can be rewritten as

$$dB = -p_2^* z_2^* \left[\frac{\eta_1 \eta_2^* (1 + \eta_1^* + \eta_2) - \eta_1^* \eta_2 (1 + \eta_1 + \eta_2^*)}{(\eta_1 - \eta_1^*)(\eta_2 - \eta_2^*)} \right] \frac{dR}{R}$$

with all terms of known sign except $(1 + \eta_1^* + \eta_2)$. For a positive response of the balance of payments to devaluation it suffices also that the terms of trade improve, or at least that they not worsen. For changes in the terms of trade are indicated by changes in p_1/p_2 and, from eq. (3),

$$\frac{d(p_1/p_2)}{p_1/p_2} = \frac{dp_1}{p_1} - \frac{dp_2}{p_2} = \left(\frac{\eta_1^*}{\eta_1^* - \eta_1} - \frac{\eta_2^*}{\eta_2^* - \eta_2} \right) \frac{dR}{R}$$

If this expression is non-negative then, from (2′), d*B* must be positive.

Bickerdike's theory is very special in that the excess demand for each commodity depends on the money price of that commodity only. Implicitly, all ‘cross’ price elasticities are set equal to zero. For more general theories and, in particular, more general versions of (2′), the reader is referred to Negishi (1968), Kemp (1970), Dornbusch (1975) and Kyle (1978).

Bibliography

Bickerdike, C.F. 1920. The instability of foreign exchange. *Economic Journal* 30(March), 118–22.

Dornbusch, R. 1975. Exchange rates and fiscal policy in a popular model of international trade. *American Economic Review* 65, 859–71.

Kemp, M.C. 1970. The balance of payments and the terms of trade in relation to financial controls. *Review of Economic Studies* 37, 25–31.

Kyle, J.F. 1978. Financial assets, non-traded goods and devaluation. *Review of Economic Studies* 45, 155–63.

Metzler, L.A. 1949. The theory of international trade. In *A Survey of Contemporary Economics*, ed. H.S. Ellis. Philadelphia: Blakiston.

Negishi, T. 1968. Approaches to the analysis of devaluation. *International Economic Review* 9, 218–27.

Robinson, J. 1947. *Essays in the Theory of Employment*. 2nd edn. Oxford: Basil Blackwell.

How to cite this article

Kemp, Murray C. "elasticities approach to the balance of payments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000056>
doi:10.1057/9780230226203.0458

The New Palgrave Dictionary of Economics Online

elasticity of intertemporal substitution

Christopher Bliss

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The elasticity of intertemporal substitution (EIS) measures the willingness on the part of the consumer to substitute future consumption for present consumption. It plays a key role in the theory of consumption and saving, in particular in the life-cycle version of that theory.

Keywords

consumption smoothing; consumption theory; elasticity of intertemporal substitution; elasticity of substitution; multiple equilibria; overlapping generations models; risk aversion; savings

Article

The EIS and consumption theory

The elasticity of intertemporal substitution (EIS) is an important number in macroeconomic theory. It measures the willingness on the part of the consumer to substitute future consumption for present consumption. This parameter plays a key role in the theory of consumption and saving, in particular in the life-cycle version of that theory. For a start we examine the role of the EIS in a basic life-cycle model. In that model there is complete certainty concerning prices, future income, and preferences present and future. The consumer can lend and borrow at will at a single invariant rate of interest, subject only to a lifetime budget constraint. Preferences are additively separable. The consumer chooses present and future consumption to maximize:

$$\sum_{t=1}^{T} \delta^{t-1} U[c_t] \tag{1}$$

where c_t is consumption in period t, and $0< \delta <1$, and where δ is the rate at which utility is discounted. Lifetime utility (1) is maximized subject to the lifetime budget constraint:

$$\sum_{t=1}^{T} c_t\left(\frac{1}{1+r}\right)^{t-1} \leq \sum_{t=1}^{T} y_t\left(\frac{1}{1+r}\right)^{t-1} \tag{2}$$

where the y values are incomes in the various periods, and r is the real rate of interest. Assuming positive consumptions in all periods, the maximization of (1) requires:

$$\frac{dU[c_t]}{dc_t}\delta^{t-1} - \lambda\left(\frac{1}{1+r}\right)^{t-1} = 0 \tag{3}$$

where λ is the Lagrange multiplier. From (3), taking logs:

$$\ln\frac{dU[c_{t+1}]}{dc_{t+1}} - \ln\frac{dU[c_t]}{dc_t} = \ln\left(\frac{1}{1+r}\right) - \ln\ \delta \tag{4}$$

Differentiating (4) with respect to r and holding c_t constant gives:

$$\frac{\dfrac{d^2U[c_{t+1}]}{dc_{t+1}^2}}{\dfrac{dU[c_{t+1}]}{dc_{t+1}}}\frac{dc_{t+1}}{dr} = -\frac{1}{(1+r)} \tag{5}$$

A useful way of writing (5) is:

$$\frac{1}{c_{t+1}}\frac{dc_{t+1}}{dr} = -\frac{1}{(1+r)}\frac{\dfrac{dU[c_{t+1}]}{dc_{t+1}}}{c_{t+1}\dfrac{d^2U[c_{t+1}]}{dc_{t+1}}} \tag{6}$$

Or equivalently:

$$\frac{1}{c_{t+1}}\frac{dc_{t+1}}{dr} = \frac{\sigma(c_{t+1})}{1+r} \tag{7}$$

where σ is the EIS, defined as:

$$\sigma(c) = -\frac{\dfrac{dU[c]}{dc}}{c\dfrac{d^2U[c]}{dc^2}} \tag{8}$$

Equation (7) indicates that the size of the EIS will be a crucial determinant of how far consumption levels will respond to changes in the interest rate.

The effect of a small change in r analysed above is a standard partial equilibrium result, in which enough is held constant to obtain a definite result. The calculation shows how two solution paths compare with regard to c_{t+1}, as r is varied slightly, when for each of these paths c_t takes the same optimal value. For that special case, (7) says that c_{t+1} increases with r, which is to say that c_{t+1} increases relative to c_t. In that particular sense a small increase in r encourages saving. Even for the two-period model popular for classroom exposition, it cannot be shown that a rise in r encourages saving. However in the two-period model it is true for any separable lifetime utility function, as (1), that c_1 declines as r increases, provided that $c_2 > y_2$, the usual case. When r increases the substitution effect always favours lower early

consumption. When $y_2 > c_2$, however, the income effect opposes the substitution effect, and the outcome is uncertain.

Equation (8) shows that the second derivative of the utility function, how curvy it is if one likes, is crucial in giving a specific value to the EIS. If:

$$U[c] = c^{\frac{\sigma-1}{\sigma}} \tag{9}$$

then the EIS is constant, independent of c, and equal to σ.

Consumption smoothing and risk aversion

The EIS as defined in (8) is the same as the Arrow–Pratt measure of relative risk aversion. It is no accident that consumption substitution through time, with no uncertainty whatsoever, and risk aversion, where uncertainty is necessarily involved, should involve the same parameter. Absolute risk aversion is related to the willingness of a consumer to accept a lottery ticket in preference to a sum of money available for certain, the certain sum being lower than the expected value of the lottery. One can think of the extra expected value in the better-than-fair lottery as a premium needed to entice the agent to accept the risk. The higher is relative risk aversion, the larger must be the expected-value premium in the lottery. Arrow (1971, ch. 3) provides a detailed discussion, and references the parallel and independent work of Pratt.

Now consider the life-cycle maximization of (1) subject to (2). To make the explanation as simple as possible let δ and r both be zero. The consumer maximizes:

$$\sum_{t=1}^{T} U[c_t] \tag{10}$$

subject to the lifetime budget constraint:

$$\sum_{t=1}^{T} c_t \leq \sum_{t=1}^{T} y_t \tag{11}$$

With U [•] a concave function, it is evident that the consumer will consume at the same level in each period:

$$c_t = \frac{\sum_{t=1}^{T} y_t}{T} \tag{12}$$

In the particular sense defined by this special case, the consumer is averse to consumption variability over time. It is the same as the risk-averse consumer disliking variations in wealth when different states of the world are realized. That each period of time will certainly arrive, whereas only one state of the world will be realized, is irrelevant in the *ex ante* view of the consumer facing uncertainty. A risk-averse agent can be induced to accept a gamble if the odds are sufficiently favourable, that is, if the expected-value premium is sufficiently large. Similarly, a life-cycle planner will opt for a non-constant consumption plan if it provides a larger total consumption sufficient to compensate for the unattractive variability. A positive rate of interest plays the same role as an expected-value premium. It is the sweetener that persuades the consumer to accept variability. For this reason it is no surprise to find that the extent to which the consumer will respond to the sweetener, in either case, is governed by precisely how much the consumer dislikes variability. And the EIS, or the coefficient of relative risk aversion, as the case may be, measures that dislike of variability.
The argument just completed ignores the part played by δ, the utility discount rate. The presence of a positive δ means that, were *r* zero, the consumer would choose a plan with consumption falling through time. Then a positive *r*, and especially an *r* greater than δ, persuades the consumer to select a consumption plan with consumption falling less rapidly or rising through time. How far an optimal plan responds to a given change in *r* is governed again by the EIS.

A constant or a variable coefficient?

The EIS has been compared above to the coefficient of relative risk aversion. In the theory of risk aversion the emphasis is on the variability of the coefficient. On this turns the issue of whether the wealthy will be more or less willing to undertake risk than the poor. With the EIS the most common assumption is that it is a constant. A popular special case of (1) is:

$$U[c_1, c_2, \ldots, c_n] = c_1^{\frac{\sigma-1}{\sigma}} + \delta c_2^{\frac{\sigma-1}{\sigma}} + \ldots + \delta^{n-1} c_n^{\frac{\sigma-1}{\sigma}} \tag{13}$$

This is the love-of-variety utility function of Dixit and Stiglitz (1977), with discounting added. The EIS measured at any of the consumptions above is σ.

The elegance and convenience of forms such as (13) has made them appealing. Thus Barro and Sala-i-Martin (1995), in their influential study of economic growth, assume that different countries or regions solve independent Ramsey optimal model problems. This leads to the condition:

$$\frac{1}{c}\frac{dc}{dt} = \sigma[AF_1\{k, 1\} - \delta] \tag{14}$$

where F_1 is the marginal product of capital, c is consumption, k is capital, δ is the utility discount rate, A measures total factor productivity as it is affected by policy, culture, corruption, and so on, and σ is the EIS. The lower is k the larger is F_1. If this effect is not offset by poor countries having lower total factor productivities, and if all countries share the same values of δ and σ, then conditional β-convergence follows from (14), meaning that poor countries grow faster.

The poor will be reluctant to save if their value of σ is low. And this is a most plausible specification. When all the meals that one eats are small, it is rationally more difficult to postpone eating now for a larger meal later. This point has been recognized in the literature. For example, King and Rebelo (1993) allow for a utility function of the Stone–Geary form, where the consumer gives priority to a fixed basket of essentials until that basket has reaches a critical scale. With those preferences, the poorest consumers will not save at all, and there is the possibility of a poverty trap. The Stone–Geary utility function implies a zero value for the EIS at low consumptions, and positive values for higher consumptions.

The EIS in consumption studies

Many applied economists used to take the view that the value of σ is close to zero (see Hall, 1988; Mankiw, Rotenberg and Summers, 1985). This reflects the failure of consumption studies to find a significant effect of the rate of interest on saving. Such estimates are seriously biased if the consumer is constrained from borrowing freely (a feature ignored in the computations above) or if, as in Deaton (1992), most consumers save only to replenish precautionary balances following negative shocks. Then the optimizing substitution-based theory does not apply. Blundell, Browning and Meghir (1994) and Attanasio and Browning (1995) show that representative consumer models give seriously misleading results when applied to aggregate consumption data. They use UK household expenditure data to model consumption at the individual level and obtain a greatly improved fit when they allow the rich to have a higher EIS than the poor. Does that mean that as economies grow richer over time, the average EIS will increase? This remains an unanswered question.

VEIS functions

Let the utility function be chosen from a class of which the simplest case is:

$$U[c] = \int_0^c \exp\left\{\frac{1}{\beta x}\right\} dx \tag{15}$$

where β is a positive constant and *c* is the level of consumption. This is a VEIS utility function, where VEIS stands for *variable elasticity of intertemporal substitution*. Then:

$$\frac{dU[c]}{dc} = \exp\left\{\frac{1}{\beta x}\right\} > 0 \tag{16}$$

and:

$$\frac{d^2U[c]}{dc^2} = -\exp\left\{\frac{1}{\beta c}\right\}\frac{1}{\beta c^2} < 0 \tag{17}$$

U [•] is an increasing concave function. Now the EIS may be computed as:

$$-\frac{\frac{dU[c]}{dc}}{c\frac{d^2U[c]}{dc^2}} = \beta c \tag{18}$$

This increases linearly with consumption at rate β. The poor have a lower EIS and β-convergence will not necessarily prevail.

A variable EIS in the Diamond capital model

In their deep study of the Diamond overlapping generations model with capital, De La Croix and Michel (2002) more or less dismiss the importance of multiple stable equilibria. To summarize, it is possible to obtain multiple stable steady-state solutions with simple functional forms, but these cases are unsatisfactory at best. If the production function is Cobb–Douglas and with a simple separable utility function, there are no cases of multiple stable steady states. With a logarithmic utility function and the constant elasticity of substitution in production $\rho > 0$, there can be two positive steady-states, but it may be that only the corner degenerate outcome is stable.

Rather than using given simple functional forms and looking for a few steady-state solutions, try for a continuum of solutions as follows. Assume:

$$-\frac{\frac{dU[c]}{dc}}{c\frac{d^2U[c]}{dc^2}} = \sigma(c) \tag{19}$$

where $\sigma(c)$ is an arbitrary positive increasing function of c. Then:

$$\frac{\frac{d^2U[c]}{dc^2}}{\frac{dU[c]}{dc}} = -\frac{1}{\sigma(c)c} \tag{20}$$

Integrating (20) gives:

$$\ln\frac{dU[c]}{dc} = -\int_a^c \frac{1}{\sigma(x)x}dx + \ln D \tag{21}$$

where a is a positive constant, and D is a constant of integration.

In a steady state solution to the Diamond model we must have:

$$\ln \frac{dU[c_1]}{dc} - \ln \frac{dU[c_2]}{dc} = \ln \delta - \ln R$$
(22)

where c_1 and c_2 are consumption in respectively the first and second period of a life, R is the gross rate of return to saving, and δ is the discount factor. From (21) and (22):

$$\int_{c_1}^{c_2} \frac{1}{\sigma(x)\,x} dx = \ln \delta - \ln R$$
(23)

Now in steady state c_1, c_2 and R all depend upon capital per head k. If over some range of values of k every value gives a steady state, then (23) will be an identity in k. Let the per capita production function be Cobb–Douglas with coefficient α. Then (23) takes the form:

$$\int_{(1-\alpha)k^{\alpha}-k}^{k+\alpha k^{\alpha}} \frac{1}{\sigma(x)\,x} dx = \ln \delta - \ln(1 + \alpha k^{\alpha-1})$$
(24)

When (24) is an identity in k, over an interval at least, then differentiating both sides of (24) gives:

$$\frac{1}{\sigma(k+\alpha k^{\alpha})} \frac{1}{k+\alpha k^{\alpha}} - \frac{1}{\sigma\left((1-\alpha)k^{\alpha}-k\right)} \frac{1}{(1-\alpha)k^{\alpha}-k} = \frac{\alpha(1-\alpha)k^{\alpha-2}}{1+\alpha k^{\alpha-1}}$$
(25)

Take a given a value of k, and let σ (c_1) values be known for the c_1 value implied by that k all the way up to the c_2 defined by the same k. Then σ (c_2) values are determined by (25), which rolls out a solution for σ such that all values of k on a connected interval are steady-state equilibrium levels. The contrast to the case advanced by De La Croix and Michel is striking.

Concluding remarks

The EIS is an important value, just as is its cousin, the coefficient of relative risk aversion. The use of a simple functional form has too often frozen the EIS as a constant. When it is allowed to vary, the β-convergence of growth theory is no longer secure; cross-section consumption studies perform better; and multiple equilibrium in the Diamond capital model is seen to be far more probable than previous studies indicate.

See Also

- consumer expenditure
- consumer expenditure (new developments and the state of research)

Bibliography

Arrow, K.J. 1971. *Essays in the Theory of Risk Bearing*. Amsterdam: North-Holland.

Attanasio, O.P. and Browning, M. 1995. Consumption over the life-cycle and over the business cycle. *American Economic Review* 85, 1118–37.

Barro, R.J. and Sala-i-Martin, X. 1995. *Economic Growth*. New York: McGraw-Hill.

Blundell, R., Browning, M. and Meghir, C. 1994. Consumer demand and the life-cycle allocation for household expenditures. *Review of Economic Studies* 61, 57–80.

Deaton, A. 1992. *Understanding Consumption*. Clarendon Lectures in Economics, Oxford: Clarendon Press.

De La Croix, D. and Michel, P. 2002. *A Theory of Economic Growth: Dynamics and Policy in Overlapping Generations*. Cambridge: Cambridge University Press.

Dixit, A.K. and Stiglitz, J.E. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Hall, R.E. 1988. Intertemporal substitution in consumption. *Journal of Political Economy* 96, 339–57.

King, R.T. and Rebelo, S.T. 1993. Transitional dynamics and economic growth in the neoclassical model. *American Economic Review* 83, 908–31.

Mankiw, N.G., Rotenberg, J.J. and Summers, L. 1985. Intertemporal substitution in macroeconomics. *Quarterly Journal of Economics* 100, 225–81.

How to cite this article

Bliss, Christopher. "elasticity of intertemporal substitution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000311>
doi:10.1057/9780230226203.0460

The New Palgrave Dictionary of Economics Online

elasticity of substitution

D.R. Helm

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

CES production functions; Cobb–Douglas functions; elasticity; elasticity of substitution; Hicks, J. R.; income effects; indifference curves; production functions; Robinson, J. V.; substitution; substitution effects; value theory

Article

The concept of the elasticity of substitution, developed by Joan Robinson and John Hicks separately in the 1930s, represented an important addition to the marginal theory of the 1870s, in the tradition of Marshall, Edgeworth and Pareto. It brought together two concepts which were already well established in the literature – the ideas of elasticities (which derive from Mill) and those of substitution (which go back to Smith). The relationship defined by the concept is a mathematical one relating to utility and production functions, with considerable economic implications. It has two applications: to the theory of production, and in particular the isoquant relationship between factor inputs, and to consumer behaviour and the indifference curve. Let us look at each in turn.

The two inventors of the concept – Joan Robinson, in her *Economics of Imperfect Competition* (1933), and John Hicks in his *Theory of Wages* (1932) – each developed Marshall's formula for the elasticity of derived demand. Each defined the concept somewhat differently. For Hicks, the definition was the percentage change in the relative amount of the factors employed resulting from a given percentage change in the relative marginal products or relative prices, that is (following Samuelson, 1968):

$$\sigma = \sigma_{12} = (F_1 F_2 / F F_{12}) = \sigma_{21},$$

where $F(V_1, V_2)$ is a standard neoclassical production function, and the subscripts are the partial derivatives. This is sometimes called the direct elasticity of substitution. For Joan Robinson, on the other

hand, concerned with relative shares and hence distributional issues, the elasticity of substitution was defined as ‘the proportionate change in the ratio of the amounts of the factors employed divided by the proportionate change in the ratio of their prices’ (1933, p. 256):

$$\sigma = -\frac{\partial (V_1 / V_2) / (V_1 / V_2)}{\partial (W_1 / W_2) / (W_1 / W_2)}$$

where W_1 is the price of the V_1 factor.

These two definitions of the concept gave rise to a considerable debate in the early issues of the *Review of Economic Studies*, with in particular a notable contribution from Kahn (1933) concerned to identify how these concepts related to each other. It turns out that these two original definitions are identical when the production function is confined to two factors of production, where the partial derivatives of the production function are the marginal productivities of the factor inputs and yield the relevant factor prices. In addition, the contributors to the debate attempted to identify the implications of these somewhat abstract concepts. Amongst these were the joint determination by the elasticity of substitution and the factor supplies of the relative shares of the factor reward (wages and profits), and implications for the definition of imperfect competition with increasing returns to scale.

It is not surprising that it is with the cases where the restrictive neoclassical assumptions for the production function are not met that most interest arises. Two important developments are where production function involves three or more factors and in extending from Cobb–Douglas to constant elasticity of substitution (CES) production functions. But although considerable emphasis has been placed on the elasticity of substitution in production, it remains a technical concept concerning factor substitutability. It has no direct allocation consequence. Diminishing elasticity of substitution does not imply diminishing returns to scale, since for returns we must have prices. Thus it is restricted to describing the technical conditions of production. But, being a technical concept, it can be generalized to all forms of transformation. Thus, as we noted above, along with a number of other concepts, these tools developed for production were taken over to consumer theory. Because of the implications the concept had for the development of consumer behaviour, and because of the insight which the resulting difficulties threw up concerning the concept more generally, this application is of special interest.

It was Hicks (and Allen) who made that step. While Joan Robinson's development of the concept was closely related to her extension of Marshall's theory of the industry, Hicks was familiar with a very different approach to value theory, that of Edgeworth, Pareto and Walras. While Joan Robinson had focused on production substitutions, and hence isoquants, Hicks took the idea developed in that domain, and translated it across to consumer theory, and to the indifference curves which he had got from Edgeworth. In the two goods case, price elasticity could be represented in terms of his fundamental formula, according to which:

$$\text{Price elasticity} = k(\text{income elasticity}) + (1 - k)(\text{e.s.})$$

where *k* is the total expenditure that is spent on the commodity. Thus, with income elasticity, consumer theory led into a representation of the effect of a price change in terms of the income and substitution effects, with elasticity being thus of prime importance in classifying goods by their demand characteristics.

But whereas the elasticity concept in production theory naturally led on to the possibility of measurement, that step in consumer theory was more contentious. For although this technical concept represented one important step in the development of the marginalist approach to the theory of value, the theory of demand behaviour requires a behavioural theory of choice. The elasticity of substitution with respect to the indifference curve is one technical component. But, as with production theory, prices, and in this case the budget line, are also required.

Technical concepts thus aided the formulation of modern consumer theory as outlined in Hicks and Allen's ‘A Reconsideration of the Theory of Value’ (1934) and the opening chapters of *Value and Capital* (1939), a path from which it has scarcely deviated. But, despite the mathematical elegance of this construction, it may be argued that it disguised many of the important underlying questions. The increased power of the indifference curve analysis begged the question of whether consumer preferences could in reality be represented in this abstract way. Ultimately, whether consumer behaviour is well described by concepts like the elasticity of substitution, depends upon whether preferences can be represented by complete, transitive, utility functions. Much recent evidence from psychologists and decision theorists suggests otherwise. Likewise for production theory, the concepts of capital and labour may be themselves ambiguous.

See Also

- CES production function
- Cobb–Douglas functions
- production functions

Bibliography

Hicks, J.R. 1932. *The Theory of Wages*. London: Macmillan.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hicks, J.R. 1970. Elasticity of substitution again: substitutes and complements. *Oxford Economic Papers* 22, 289–96.

Hicks, J.R. and Allen, R.G.D. 1934. A reconsideration of the theory of value I–II. *Economica* 1, Pt 1, February, 52–76; Pt II, May, 196–219.

Kahn, R.F. 1933. The elasticity of substitution and the relative share of a factor. *Review of Economic*

Studies 1(October), 72–8.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Samuelson, P.A. 1968. Two generalizations of the elasticity of substitution. In *Value, Capital and Growth*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

How to cite this article

Helm, D. R. "elasticity of substitution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000059> doi:10.1057/9780230226203.0461

The New Palgrave Dictionary of Economics Online

elasticity

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Formally invented by Marshall, the concept of elasticity of demand goes beyond the notion, which can be found in classical economics, that demand varies less or more than price. The crucial property that alone makes elasticity so important in pure and applied economics is that the elasticity measure is *invariant* to changes in units of measurement of quantities and prices.

Keywords

arc elasticity; Cournot, A. A.; dimensional analysis; duality; elasticity; elasticity of demand; elasticity of substitution; elasticity of supply; invariance to transformation; Mill, J. S.

Article

One day in the winter of 1881–2 Alfred Marshall came down from the sunny rooftop of his hotel in Palermo ‘highly delighted’, for he had just invented elasticity of demand (Keynes, 1925, pp. 39 n. 3, 45 n. 2). So delighted was he that within a mere four years he had introduced the word *elasticity* into the technical literature of economics (Marshall, 1885), which by his own standards was rushing pell-mell into print. But if the speed of its introduction was uncharacteristic the manner of it was not, tucked away as it was at the end of a lecture dull even for its time, and giving no hint that elasticity was new and exciting (1885, p. 187).

The notion that demand varies less or more than price can of course be found rather often in classical economics, especially in John Stuart Mill (Edgeworth, 1894, p. 691). But to turn that trite idea into something useful requires a firm grip on the prior idea of quantity demanded *at a price*. So it is not surprising that the only ancient who came close to Marshall's idea was Cournot himself, the inventor of (among much else) the demand function.

In fact Cournot came so close that it is hard to understand, first, why he did not go all the way, and second, why Marshall gave him no credit for showing that way. Such lack of generosity is the more

puzzling since we know that between the time when (according to Mrs Marshall) he invented elasticity, and the late spring of 1882 when he first drafted the chapter on Elasticity for the *Principles*, Marshall reread Cournot (Whitaker, 1975, vol. 1, p. 85).

Starting with the demand function $D = F(p)$, Cournot pointed out that $pF(p)$ is total revenue, so that for maximum revenue the price p must be such that $F(p) + pF'(p) = 0$ (1838, p. 56). Thus total revenue will increase or decrease with increase in price according as $\Delta D/\Delta p$ is larger or smaller than D/p, where ΔD is the absolute value of the change in quantity demanded.

> Commercial statistics should therefore be required to separate articles of high economic importance into two categories, according as their current prices are above or below the value which makes a maximum of $pF(p)$. We shall see that many economic problems have different solutions, according as the article in question belongs to one or other of these two categories. (Bacon's translation, 1897, p. 54)

Let f be a real-valued nonzero differentiable function whose domain is some open interval I of the real line. In conformity with Marshall's Mathematical Appendix (1890, Note IV, pp. 738–40), the *elasticity of f at the point x*, denoted by $\eta_f(x)$, is defined here to be the *number* $xf'(x)/f(x)$. *The function* η_f defined by this formula is called the *elasticity of f*. To define the elasticity of *demand*, some authors prefer to follow the convention $f(x) = -xf'(x)/f(x)$, which is not used here. Unfortunately there is no standard notation for elasticity, since the obvious candidates are already taken, e for e and E for the expectations operator.

Cournot's critical value of p, his criterion for sorting out commodities, is simply that p^* for which $\eta_f(p^*) = -1$; he was close indeed. However, unlike Marshall (who is crystal clear on the point) there is no trace in Cournot of the crucial property that the elasticity measure is *invariant* to changes in units of measurement of quantities and prices, and it is this property alone that makes it so important in pure and applied economics.

A little calculus will prove such invariance, but is more enlightening to apply the dimensional analysis of Jevons and Wicksteed. Let the dimension of x be X and that of $f(x) = y$ be Y, so that $f'(x)$ has dimension YX^{-1}. The dimension of $\eta_f(x)$ is then $X \cdot YX^{-1} \cdot Y^{-1}$ and everything cancels. The elasticity of f at x is a pure number, unaffected by change in the units of either x or y. (This application is so obvious that the most plausible explanation of why it was not included in Wicksteed, 1894, is that his entry was actually written before Marshall's *Principles* appeared.) Although invariance to transformation of units is the key property of elasticities, partly as a consequence the measure has a number of other agreeable properties. For example, it is easily seen that $\eta_f(x) = d\log f(x)/d\log x$, which paves the way for a whole calculus of elasticities in terms of logarithmic derivatives (Champernowne, 1935; Allen, 1938, pp. 251–4). One simple application of this calculus is the formula $\eta_{fg}(x) = \eta_f(x) + \eta_g(x)$, where fg is the product of f and g (with a corresponding formula for the quotient function f/g), while another is the characterization of constant elasticity functions as those which are linear in logarithms, that is, of Wicksell–Cobb–Douglas type. Incidentally, Douglas's paper of 1927 was apparently intended to

introduce elasticity of supply, which is odd since it had already appeared 20 years before (and rather late at that) in the fifth edition of the *Principles* (see Marshall, 1961, vol. 2, p. 521).

The extension of elasticity to functions of more than one variable is easy – one simply uses the partial derivatives f_i rather than the derivative f' – and is staple fare in textbooks (see for example Allen, 1938, pp. 310–12). However, many of those textbooks underplay another useful property of elasticities of strictly monotonic functions (such as the usual demand and supply curves) which follows from the inverse function theorem. Considering just functions of one variable, if we write $\Phi = f^{-1}$ then from that theorem $\Phi' = f^{-1}$, so from this and the definition of elasticity,

$$\eta_\Phi(y) = y\Phi'(y) / \Phi(y) = f(x) / xf'(x) = (\eta_f(x)^{-1}),$$

that is, the elasticity of the inverse function is the inverse of the elasticity. Two obvious applications of this to the elementary theory of the firm are:

(i) Since the revenue function is $R(q) = pq = q\Phi(q)$,

$$\text{marginal revenue } (mr) = \Phi(q) + q\Phi'(q) = \Phi(q)[1 + (q\Phi'(q) / \Phi(q)] = \Phi(q)(1 + \eta_\Phi(q)),$$

from which one can derive the more usual but less intuitive formula $mr = p[1 + (1 / \eta_f(p))]$; and (ii) since at the firm's profit maximizing output marginal cost $mc = mr$, the Lerner (1934) measure of monopoly power $(p–mc)/p$ may be written

$$[\Phi(q) - mr] / \Phi(q) = 1 - [\Phi(q)(1 + \eta_\Phi(q)] / \Phi(q) = -\eta_\Phi(q).$$

Arc elasticity, which is really ordinary elasticity with the index number problem thrown in, was introduced quite early by Dalton (1920, pp. 192–7). But the heyday of elasticities of all kinds came later, in the 1930s, so much so that it is small wonder that in the immediate post-war period Samuelson (1947, pp. 4–5) used elasticity statements to exemplify what he meant both by ‘meaningful theorems’ and by non-meaningful theorems in economics. A peculiar aspect of some of the elasticity measures introduced then was their definition not in terms of the properties of a given *function f* (as here), but rather as the ratio of proportionate change in one variable to proportionate change in another, allegedly causative, variable, without any explicit functional relationship intervening. Thus with Hicks's ‘elasticity of expectations’ (1939, p. 205) there *is* no ‘expectation function’ of which it is an elasticity, as that term is defined above. Similarly, although the elasticity of substitution (σ) invented by Hicks (1932) and Robinson (1933) immediately provoked many articles in response (for example, Lerner, 1933), at no time was a ‘substitution function’ introduced whose elasticity it was. The lack of a generating function for σ might help to explain why its use often occasions technical difficulty.

It is of some interest to apply duality theory to the problem of deriving simple formulas for entities like

σ (cf. Woodland, 1982, p. 31). Consider the elasticity of substitution σ between two consumer's goods *x* and *y*, with no restriction being placed on preferences apart from the smoothness conditions implicit at this level of analysis. First, take advantage of homogeneity in both the ordinary and compensated demand functions to write the former function as $f(p, m)$ and the latter as $h(p, t)$, where *p* is the price of *x* in terms of *y*, *m* is the consumer's income in terms of *y*, and *t* is the *maximized* level of utility for the price-income situation (p, m). Put $x^* = f(p, m)$. Finally, observe that σ is wholly determined by the price slope corresponding to *p* together with the indifference curve corresponding to *t*, so that we may write $\sigma = \sigma(p, t)$.

From a modern version of the fundamental equation of value theory (Hicks, 1939, p. 309),

$$f_p(p, m) = h_p(p, t) - x^* f_m(p, m) \tag{1}$$

where f_p, h_p and f_m are, in sequence, the partial derivatives of *f* and *h* with respect to *p*, and of *f* with respect to *m*. Multiplying (1) by $p/f(p, m)$ and writing η_{fp}, η_{fm} for the two partial elasticities of *f*, we obtain

$$\eta_{fp}(p, m) = ph_p(p, t) / x^* - px^* m f_m(p, m) / (mf(p, m)) = ph_p(p, t) / x^* - k\eta_{fm}(p, m) \tag{2}$$

where $k = px^* m$, that is, the fraction of *m* spent on *x*. Now since *t* is the maximized level of utility, given local non-satiation $x^* = h(p, t)$. Hence, the first term on the right-hand side of (2) is $\eta_{hp}(p, t)$, the partial elasticity of *h* with respect to *p*, and (2) becomes

$$\eta_{fp}(p, m) = \eta_{hp}(p, t) - k\eta_{fm}(p, m). \tag{3}$$

A standard result of Hicks and Allen (1934; see Hicks, 1981, p. 20) for the two-good case can be written in the present notation as

$$-\eta_{fp}(p, m) = k\eta_{fm}(p, m) + (1 - k)\sigma(p, t)$$
(4)

so from (3) and (4),

$$(k - 1)\sigma(p, t) = \eta_{hp}.$$
(5)

Let the cost (expenditure) function for this problem be $c(p, t)$, and denote its partial derivative with respect to p by c_p. Then, writing η_{cpp} for the partial elasticity of c_p with respect to p, since Shephard's Lemma implies $c_p = h$ we have

$$\eta_{hp}(p, t) = \eta_{cpp}(p, t).$$
(6)

Now $k = px^* / m = ph(p, m) / m = pc_p(p, t) / m$. Because t is the maximized level of utility $m = c(p, t)$, so $k = pc_p(p, t) / c(p, t) = \eta_{cp}(p, t)$, where η_{cp} is the partial elasticity of c with respect to p. Substituting from this and (6) into (5),

$$\sigma(p, t) = \eta_{cpp}(p, t) / (\eta_{cp}(p, t) - 1).$$
(7)

Thus the elasticity of substitution in this two-good case can be expressed entirely in terms of the cost function.

See Also

- Marshall, Alfred

Bibliography

Allen, R.G.D. 1938. *Mathematical Analysis for Economists*. London: Macmillan.

Champernowne, D.G. 1935. A mathematical note on substitution. *Economic Journal* 15, 246–58.

Cournot, A.A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. New edn, ed. G. Lutfalla, Paris: Riviére, 1938. English trans. by N.T. Bacon, 1897. Reprinted, New York: A.M. Kelley, 1960.

Dalton, H. 1920. *Some Aspects of the Inequality of Incomes in Modern Communities*. London: Routledge.

Douglas, P.H. 1927. Elasticity of supply as a determinant of distribution. In *Economic Essays Contributed in Honor of John Bates Clark*, ed. J.H. Hollander. New York: Macmillan.

Edgeworth, F.Y. 1894. Elasticity. In *Dictionary of Political Economy*, vol. 1, ed. R.H.I. Palgrave. London: Macmillan.

Hicks, J.R. 1932. *The Theory of Wages*. London: Macmillan.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hicks, J.R. 1981. *Collected Essays on Economic Theory*. Vol. 1: Wealth and Welfare. Cambridge, MA: Harvard University Press.

Hicks, J.R. and Allen, R.G.D. 1934. A reconsideration of the theory of value. *Economica* 1, 52–76, 196–219. Reprinted in Hicks (1981, vol. 1).

Keynes, J.M. 1925. Alfred Marshall, 1842–1924. In Pigou (1925).

Lerner, A.P. 1933. The diagrammatical representation of elasticity of substitution. *Review of Economic Studies*. Reprinted in Lerner (1953).

Lerner, A.P. 1934. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1(June), 157–75. Reprinted in Lerner (1953).

Lerner, A.P. 1953. *Essays in Economic Analysis*. London: Macmillan.

Marshall, A. 1885. The graphic method of statistics. *Journal of the Royal Statistical Society*. Reprinted in Pigou (1925).

Marshall, A. 1890. *Principles of Economics*, vol. 1, London: Macmillan.

Marshall, A. 1961. *Principles of Economics*. 9th (variorum) edn, with annotations by C.W. Guillebaud. 2 vols. London: Macmillan.

Pigou, A.C., ed. 1925. *Memorials of Alfred Marshall*. London: Macmillan. Reprinted, New York: A.M. Kelley, 1966.

Robinson, J.V. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of Economics Analysis*. Cambridge, MA: Harvard University Press.

Whitaker, J.K., ed. 1975. *The Early Economic Writings of Alfred Marshall, 1867–1890*. 2 vols. New York: Free Press.

Wicksteed, P.H. 1894. Dimensions of economic quantities. In *Dictionary of Political Economy*, vol. I, ed. R.H.I. Palgrave. London: Macmillan.

Woodland, A.D. 1982. *International Trade and Resource Allocation*. Amsterdam: North-Holland.

How to cite this article

Newman, Peter. "elasticity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000058> doi:10.1057/9780230226203.0459

The New Palgrave Dictionary of Economics Online

electricity markets

James Bushnell and Catherine Wolfram
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost-based regulation; electricity industry; electricity markets; experimental economics; game theory; incentive regulation; investment; liberalization; market power; natural monopoly; oligopoly simulation analysis; privatization; real-time pricing; state-owned utilities

Article

In many parts of the world buyers and sellers now trade electrical energy in liberalized markets. These markets have partially replaced cost-based regulation and government ownership.
Since the 1980s, governments in many countries have privatized and restructured their electricity industries. Liberalized electricity markets now operate in much of Europe, North and South America, New Zealand and Australia. These changes were primarily motivated by the perception that the previous regimes of either state ownership or cost-of-service regulation yielded inefficient operations and poor investment decisions. Liberalization of the electricity industry also reflected the progression of a deregulation movement that had already transformed infrastructure industries, including water, communications and transportation, in many countries. Although electricity shares many characteristics with other deregulated industries, the differences have proven to be more important than the similarities. Electricity has been one of the most challenging industries to liberalize and in most places new layers of regulations have replaced the old.
Historically, electricity was viewed as a natural monopoly. Typically, a single utility company generated, transmitted and distributed all electricity in its service territory. In much of the world, the monopoly was a state-owned utility. Within the United States, private investor-owned companies supplied the majority of customers, although federally and municipally owned companies played an important minority role. These companies operated under multiple layers of local, state and federal regulation.
Restructured electricity markets share a common basic organization. The three segments – generation, transmission, and distribution – have been unbundled. Wholesale generation, no longer viewed as a natural monopoly, is priced through a market process. Transmission and distribution remain regulated,

although in many cases some form of incentive regulation has replaced cost-of-service regulation or state ownership.

Most wholesale electricity is traded through long-term (a week or longer) forward contracts. Many markets also feature day-ahead auction-based exchanges. Because supply and demand must be continually balanced to preserve transmission stability, transmission system operators run real-time balancing markets. Prices in these high-frequency markets can be highly volatile since electricity is non-storable and real-time demand fluctuates dramatically. To meet unforeseen contingencies, transmission system operators also contract for and occasionally use standby or reserve generation services. Many markets reflect price differences across geographical locations when parts of the transmission grid are congested (Schweppe et al., 1988; Chao and Peck, 1992). Game theorists and experimental economists are involved in the ongoing process of designing electricity markets (Wilson, 2002), while empirical researchers have used detailed auction data to estimate how well predictions from theoretical models describe firm behaviour (Wolak, 2000; Hortascu and Puller, 2004).

At the retail level, the vision of liberalization was to provide customers a choice among competing retailers who would operate as either resellers or integrated providers with access to customers through a regulated common-carriage distribution network. In most restructured US markets, retail competition for residential customers is very weak (Joskow, 2005). Retail competition is more advanced in the United Kingdom, although evidence suggests that customers have been slow to take advantage of the ability to switch to a lower-priced retailer (Waddams, 2004). Several authors have noted the economic benefits of allowing retail prices to vary to reflect real-time changes in the wholesale prices, although this sort of real-time retail pricing has been slow to take hold in practice (Borenstein and Holland, 2005; Joskow and Tirole, 2004).

Oligopoly simulation analysis indicates the potential for serious market power problems because suppliers face extremely inelastic demand and entry requires long lead times (Green and Newbery, 1992). Empirical work has indicated that market power has indeed been present, although to varying degrees in different markets. Wolfram (1999) found that prices in England and Wales were lower than static oligopoly models would suggest. By contrast, extreme levels of market power in California contributed to record high prices in 2000–1 (Borenstein, Bushnell and Wolak, 2002). The explanations for these differences have focused on variations in the threat of future regulation and in the extent of long-term fixed price contracts (Bushnell, Mansur and Saravia, 2005).

Although the main motivation for market liberalization was to improve economic efficiency, there have been few attempts to measure efficiency changes. Newbery and Pollitt (1997) and Fabrizio, Rose and Wolfram (2004) find modest positive effects of market liberalization on, respectively, industry efficiency in the United Kingdom and plant-level efficiency in the United States.

As electricity industry restructuring moves forward, the major unresolved question is the degree to which public policy will influence investment decisions. Electric generating plants are long-lived, so while operating efficiency gains appear to be real, the potential gains from improved investment stand to be larger. Also, policies to limit the environmental impact of electricity generation could affect the types of technologies in which we invest.

See Also

- competition
- energy economics
- privatization impacts in transition economies

Bibliography

Borenstein, S., Bushnell, J. and Wolak, F. 2002. Measuring market inefficiencies in California's deregulated wholesale electricity market. *American Economic Review* 92, 1376–405.

Borenstein, S. and Holland, S. 2005. On the efficiency of competitive electricity markets with time-invariant retail prices. *RAND Journal of Economics* 36, 469–93.

Bushnell, J., Mansur, E. and Saravia, C. 2005. Vertical arrangements, market structure, and competition: an analysis of restructured US electricity markets. CSEM Working Paper WP-126. Berkeley: University of California Energy Institute.

Chao, H. P. and Peck, S. 1992. A market mechanism for electric power transmission. *Journal of Regulatory Economics* 10, 25–60.

Fabrizio, K., Rose, N. and Wolfram, C. 2004. Does competition reduce costs? Assessing the impact of regulatory restructuring on US electric generation efficiency. CSEM Working Paper WP-135. Berkeley: University of California Energy Institute.

Green, R. and Newbery, D. 1992. Competition in the British electricity spot market. *Journal of Political Economy* 100, 929–53.

Hortascu, A. and Puller, S. 2004. Understanding strategic bidding in restructured electricity markets: a case study of ERCOT. Working paper. Chicago: University of Chicago. Online. Available at http://econweb.tamu.edu/puller/AcadDocs/Hortacsu_Puller.pdf, accessed 20 July 2005.

Joskow, P. 2005. The difficult transition to competitive electricity markets in the US. In *Electricity Restructuring: Choices and Challenges*, ed. J. Griffen and S. Puller. Chicago: University of Chicago Press.

Joskow, P. and Tirole, J. 2004. Retail electricity competition. CMI Working Paper 44. Cambridge, MA: MIT.

Newbery, D. and Pollitt, M. 1997. The restructuring and privatization of Britain's CEGB – was it worth it? *Journal of Industrial Economics* 45, 269–303.

Schweppe, F., Caramanis, M., Tabors, R. and Bohn, R. 1988. *Spot Pricing of Electricity*. Norwell, MA: Kluwer Academic Publishers.

Waddams, C. 2004. Spoilt for choice? The costs and benefits of opening UK residential energy markets. Working Paper 04–1. Norwich: Centre for Competition and Regulation, University of East Anglia.

Wilson, R. 2002. Architecture of power markets. *Econometrica* 70, 1299–340.

Wolak, F. 2000. An empirical analysis of the impact of hedge contracts on bidding behavior in a competitive electricity market. *International Economic Journal* 14, 1–39.

Wolfram, C. 1999. Measuring duopoly power in the British electricity spot market. *American Economic Review* 89, 805–26.

How to cite this article

Bushnell, James and Catherine Wolfram. "electricity markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000242>
doi:10.1057/9780230226203.0462

The New Palgrave Dictionary of Economics Online

electronic commerce

Avi Goldfarb

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Electronic commerce is the exchange, distribution, or marketing of goods or services over the Internet. This article first reviews electronic commerce adoption across US industries. While the Internet is used in most industries, it has had a profound impact only on a small number. Businesses that rely heavily on electronic commerce can be divided into four groups: retail, media, business-to-business and other intermediaries. Each of these is discussed. The article concludes with a discussion of some features of electronic commerce that are of special interest to economists: lower economic frictions, lower communication costs, lower marginal costs and rich data.

Keywords

advertising; bundling; communication costs; computer industry; economic frictions; electronic commerce; Internet, economics of the; media; menu costs; price dispersion; retail; search costs; switching costs

Article

In this article, electronic commerce is defined as the exchange, distribution, or marketing of goods or services over the Internet.

There is, unfortunately, no standard definition used in the academic literature or the popular press. A broader definition would include all business facilitated by telephones, fax machines, televisions, and other technologies that are ‘electronic’. This broad definition, however, becomes so large that it encompasses a substantial fraction of all economic activity since the 1950s. A narrower definition would focus only on items sold over the World Wide Web, the browser-enabled portion of the Internet. This definition omits much of the important business-to-business segment of electronic commerce and the numerous advertising-supported websites.

The definition used here encompasses a variety of ways in which businesses have used the Internet. The

Internet is a worldwide network of computers that connect to each other using the communication protocols defined by TCP/IP. Electronic commerce includes businesses that have used the Internet to reach other businesses and to reach consumers directly. It includes businesses that sell products directly to their customers and businesses that function as intermediaries. This definition also includes businesses that operate only online, the online business of those that operate online and offline, and businesses that use the Internet but not as their primary business function.

Adoption of electronic commerce by industry

While most attention has focused on those few businesses where the Internet is a fundamental part of their strategy, electronic commerce is just one aspect of business processes for most businesses. As of 2000, nearly 90 per cent of large US establishments used the Internet (Forman, Goldfarb and Greenstein, 2002). Nearly all industries and cities had adoption rates well over 70 per cent. For the vast majority of these establishments, the Internet was used to send and receive email, to help automate some basic processes like inventory management, and/or for web browsing. This basic level of use was particularly important to establishments in rural areas (Forman, Goldfarb and Greenstein, 2005). Overall, the impact on most industries, from nursing homes to construction to furniture manufacturing to petrol stations, has been limited. The Internet is used in day-to-day business activities, but it is a small piece in a much larger puzzle. Even in retail, the US Census reported that Internet sales (totalling $26.3 billion) were just 2.7 per cent of total US retail sales in the second quarter of 2006 (U.S. Census Bureau, 2006b).

Still, a small portion of businesses have used the Internet to enhance business processes at a deep level. While little research has examined why some industries adopted quickly and others did not, it is the businesses that adopted quickly that get the majority of the attention. The Internet has had a profound effect on publishing, securities trading, some wholesaling, and some retailing (for example, books and computers). In particular, businesses that rely heavily on electronic commerce can be divided into four (not necessarily mutually exclusive) groups: retail, media, business-to-business (B2B), and other intermediaries.

Retail

Electronic commerce represents the introduction of a new sales channel. While the size of the online channel is still small relative to the entire retail sector, electronic commerce has had a large effect on some retail markets. According to the U.S. Census, Internet sales made up over ten per cent of 2004 retail sales in two broad categories if online-only stores are included: electronics and appliance stores (that is, NAICS 443) and sporting goods, hobby, book, and music stores (that is, NAICS 451) (U.S. Census Bureau, 2006a). Much of the literature on electronic commerce has focused on these categories, as well as motor vehicles and travel.

A new channel has the potential to create channel conflict. There is considerable evidence that consumers compare prices and options across channels (Prince, 2006; Ellison and Ellison, 2006). Forman, Ghose and Goldfarb (2006) show that use of the online channel depends on local offline retail options. Also, Hendershott and Jie Zhang (2006) argue that manufacturers may face resistance from their retailers to setting up a direct online channel. They show that the benefits of selling directly to

consumers (rather than though a retailer) depend on the relative online–offline search costs. The benefits of the online channel are largest for goods that are not widely available in retail stores (that is, high offline search costs) and for goods that do not need to be touched to assess quality (that is, low online search costs).

Media websites

In addition to a new retail channel, the Internet has provided a new media outlet. This outlet has developed a market structure similar to the magazine industry (Goldfarb, 2004). Media websites provide information to visitors and earn money (mostly) through advertising. In particular, entry is easy but distribution is difficult to achieve; concentration is largely determined by market size and distribution costs; large media conglomerates coexist with small niche players; and there is a high mortality rate. Online media appear to be particularly important to overcome local isolation (Sinai and Waldfogel, 2004). The two-sided nature of the media market and the digital nature of the product mean that competition between media websites is different in nature from competition between online retailers.

Intermediaries

According to Alexa.com, six of the top seven most popular websites in October 2006 had roles as intermediaries: Yahoo, MSN, Google, MySpace, YouTube, and eBay. While these intermediaries may share features of media websites (Google) or retailers (eBay), their primary business is to facilitate online interactions. Without physical storefronts or displays, intermediaries help individuals (and firms) find each other online. Intermediaries allow people with heterogeneous tastes to find better matches in terms of media, products, and people (Scott Morton, 2006).

Business to business

Business-to-business (B2B) electronic commerce is a relatively under-researched area, perhaps because of the difficulties in obtaining data. Still, B2B transactions are many times the size of business-to-consumer transactions. Lucking-Reiley and Spulber (2001) summarize many of the key questions and opportunities in B2B electronic commerce including B2B exchanges, automatic ordering, and outsourcing. Some aspects of the Internet, such as asynchronous communication, may be particularly important for international B2B interactions. Many B2B applications can also be done on electronic data interchange (EDI) rather than the Internet.

Key features of electronic commerce for general economic research

In addition to its widespread usage across industries and its profound impact on a small set of them, electronic commerce has a number of features that make it a particularly interesting area of study for economists.

Fewer economic frictions

The Internet reduces a number of economic frictions that are often cited as key contributors to observed imperfections in markets. To the consumer, search and switching costs are reduced substantially. To the firm, menu and distribution costs may fall.

For consumers, the Internet makes it relatively easy to search through several retail options. Instead of having to walk from store to store, consumers can simply click from one company to another without leaving their desks. Furthermore, a number of intermediaries exist that reduce search costs even further. These ‘shopbots’ allows consumers to compare prices and features from several websites during a single keyword search. In addition to lower search costs, switching costs are also lower online than offline. It is not difficult to switch from one competitor to another. Much of the earliest research examining electronic commerce focused on why price dispersion persisted in this environment. Broadly speaking, this literature concluded that, all else equal, search and switching costs are lower online; however, firms created search and switching costs to overcome this challenge (Ellison and Ellison, 2004).

Consequently, there is still substantial price dispersion online. Still, low search costs do not mean zero search costs. Visibility matters to the long-term prospects of any business-to-consumer company. Many early Internet companies struggled because they misinterpreted low search costs as zero search costs, mistakenly assuming customers would arrive once they set up the website.

Firms also benefit from fewer frictions online. In particular, the menu costs of changing prices and updating product offerings are much lower online than offline. In addition to the reduction in menu costs, some firms benefit from lower distribution costs: for digital goods (namely, music, news and images) online distribution costs are near zero. Low menu costs combined with the digital nature of many online products allow for mass customization of products (Murthi and Sarkar, 2003) and creative bundling, licensing, versioning and pricing strategies. Shapiro and Varian (1999) and Bakos and Brynjolfsson (1999) provide examples of a number of situations in which online firms are better able to match customers needs and therefore are better able to price discriminate.

Lower communication costs

The Internet reduces communication costs considerably. It provides an additional means of communication that creates new potential to interact with customers, suppliers and with other branches of the same firm. Internet communication differs from telephone communication in two primary ways. First, the marginal cost of communication is effectively zero, even over long distances. While establishing a connection is costly, each additional e-mail, web page viewed, and instant messaging interaction has no monetary cost to the communicator. Second, Internet communication is often asynchronous. Unlike telephone communications, the people communicating do not necessarily have to be available at the same time. This has many important applications. For example, it facilitates communication across time zones. Together, these features of Internet communication mean that geography may be less important online. Given access, people can communicate with any other person who has access, irrespective of location. Still, despite the substantial fall in long-distance communications costs, most online communication is local because social networks are local (Wellman, 2001).

Lower marginal costs

Many goods sold over the Internet are digital in nature (for example, newspaper content, music, information). The marginal cost of replication for digital goods is near zero. Depending on the particular good, fixed costs may be high (software) or low (blogs). Shapiro and Varian (1999) discuss in detail the economics of goods with high fixed and low marginal costs. If fixed costs are high enough, this cost structure allows monopolists with broad flexibility in pricing, versioning and bundling policies. It also leads to substantial economies of scale and incentives to sell a broad scope of products. In markets with more than one player, this cost structure can lead to fierce competition and little profit. If fixed costs are low and entry is easy then prices should approach zero.

One misunderstood aspect of electronic commerce is that many Internet business models have not benefited from low marginal costs, and therefore have no cost advantage over offline competition. Low marginal costs apply only to digital goods and services. In the late 1990s, many companies failed because their business models shipped heavy items to consumers. For example, taking orders for pet food and shipping it to customers involves very high marginal costs per item sold.

Rich data

By definition, all online activity is digital. This means that it is relatively easy to record and store information on the behaviour of consumers and firms online. In contrast, it is extremely expensive to track all a shopper's activity in a typical offline store. Online, however, every item browsed and the time spent looking is easily recorded. This presents an opportunity for both firms and researchers. Firms can use this data to better understand their customers, which leads to more effective customization. Researchers can use this data to answer many questions that previously could not be answered due to data constraints. Online data has greatly enhanced of our understanding of a number of economic concepts including auctions (for example, Bajari and Hortacsu, 2003), the economics of information (for example, Jin and Kato, 2005), and social interactions (for example, Mayzlin and Chevalier, 2006).

In summary, this article has identified some important features of electronic commerce and the some of the main areas of related economic research. Useful surveys of electronic commerce and related subjects include Scott Morton (2006), Hendershott (2007), and Ellison and Ellison (2005).

See Also

- computer industry
- information technology and the world economy
- Internet, economics of the
- price dispersion

Bibliography

Bajari, P. and Hortacsu, A. 2003. The winner's curse, reserve prices, and endogenous entry: empirical

insights from eBay auctions. *RAND Journal of Economics* 34, 329–55.

Bakos, Y. and Brynjolfsson, E. 1999. Bundling information goods: price, profits, and efficiency. *Management Science* 45, 1613–30.

Ellison, G. and Ellison, S.F. 2004. Search, obfuscation, and price elasticities on the Internet. Working Paper, No. 10570. Cambridge, MA: NBER.

Ellison, G. and Ellison, S.F. 2005. Lessons about markets from the Internet. *Journal of Economic Perspectives* 19(2), 139–58.

Ellison, G. and Ellison, S.F. 2006. Internet retail demand: taxes, geography, and online-offline competition. Working paper No. 12242. Cambridge, MA: NBER.

Forman, C., Ghose, A. and Goldfarb, A. 2006. Geography and electronic commerce: measuring convenience, selection, and price. Working Paper No. 06–15. New York: NET Institute.

Forman, C., Goldfarb, A. and Greenstein, S. 2002. Digital dispersion: an industrial and geographic census of commercial Internet use. Working Paper No. 9287. Cambridge, MA: NBER.

Forman, C., Goldfarb, A. and Greenstein, S. 2005. How did location affect adoption of the commercial Internet? Global village vs. urban leadership. *Journal of Urban Economics* 58, 389–420.

Goldfarb, A. 2004. Concentration in advertising-supported online markets: an empirical approach. *Economics of Innovation and New Technology* 13, 581–94.

Hendershott, T., ed. 2007. *Handbook of Economics and Information Systems*. Amsterdam: North-Holland.

Hendershott, T. and Jie Zhang. 2006. A model of direct and intermediated sales. *Journal of Economics & Management Strategy* 15, 279–316.

Jin, G.Z. and Kato, A. 2005. Price, quality, and reputation: evidence from an online field experiment. *RAND Journal of Economics* (forthcoming).

Lucking-Reiley, D. and Spulber, D.F. 2001. Business-to-business electronic commerce. *Journal of Economic Perspectives* 15(1), 55–68.

Mayzlin, D. and Chevalier, J.A. 2006. The effect of word-of-mouth on sales: online book reviews. *Journal of Marketing Research* 43, 345–54.

Murthi, B.P.S. and Sarkar, S. 2003. The role of the management sciences in research on personalization. *Management Science* 49, 1344–62.

Prince, J. 2006. The beginning of online/retail competition and its origins: an application to personal computers. in the *International Journal of Industrial Organization* (forthcoming).

Scott Morton, F. 2006. Consumer benefit from use of the Internet. In *Innovation Policy and the Economy*, vol. 6, ed. A.B. Jaffe, L. Lerner and S. Stern. Cambridge, MA: MIT Press.

Shapiro, C. and Varian, H.R. 1999. *Information Rules: A Strategic Guide to the Network Economy*. Boston: Harvard Business School Press.

Sinai, T. and Waldfogel, J. 2004. Geography and the Internet: is the Internet a substitute or a complement for cities? *Journal of Urban Economics* 56, 1–24.

U.S. Census Bureau. 2006a. E-Stats, 25 May. Online. Available at http://www.census.gov/eos/www/papers/2004/2004reportfinal.pdf, accessed 13 January 2007.

U.S. Census Bureau. 2006b. Quarterly retail e-commerce sales, 2nd quarter 2006. U.S. Census Bureau News, 17 August. Online. Available at http://www.census.gov/mrts/www/data/html/06Q2.html, accessed 13 January 2007.

Wellman, B. 2001. Computer networks as social networks. *Science* 29, 2031–4.

How to cite this article

Goldfarb, Avi. "electronic commerce." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000310> doi:10.1057/9780230226203.0463

The New Palgrave Dictionary of Economics Online

elites and economic outcomes

Elise S. Brezis and Peter Temin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Elites are a necessary part of economic activity. It therefore matters how elites are recruited and how they act. History is full of examples of elites that have acted well and also badly. Modern research has examined the training of elites, recruitment schemes and incentives for elites to discover how they can be used to promote, rather than impede, economic growth. The literature has also emphasized the effect of elite interconnection and elite recruitment on social mobility; it has shown that the standardization of elite education over the years may lead to uniformity and the creation of a transnational oligarchy.

Keywords

circulation of elites; class; corruption; cultural capital; economic growth; education; elites and economic outcomes; globalization; human capital; inequality; iron law of oligarchy; meritocracy; power elite; social capital; social mobility; symbolic capital; technical change

Article

A ruling elite (from the Latin *eligere*, ‘to elect’) is a small, dominant group that enjoys the power of decision in the various sectors of the economic and social organization of a state. It includes the bureaucrats and civil servants who rule the macro-environment; the political elite that governs and operates the executive, legislative and judicial structures; and the business elite. Non-ruling elites include the members of the media, academia and the intelligentsia.

Even in a democratic regime in which the power is meant to reside in the *demos* (‘the people’), power is really concentrated in the hands of a few. All political organizations, even democracies, tend towards domination by an oligarchy, which Mills (1956) called the *power elite*. This is the *iron law of oligarchy* as stated by Michels (1915). This stratification of society based on the accumulation of decision-making power therefore differs from the familiar stratification based on income and economic means, or on ownership of the factors of production as emphasized by Marx.

The effects of elite actions on the economy operate through several channels: economic growth and development; social mobility; inequality; and the political system, which in turn affects the economy. The characteristics that affect these economic realms are (*a*) the extent of the intertwining and inter-connections of elites; and (*b*) the stability and recruitment of the elite.

Elites ’ interconnections

The ruling elite can display unity and collusion, acting as a monolithic group, or it can be fragmented and characterized by dissociation and diversification of power, a ‘polyarchy’ that permits competition among its members.

The elite in non-democratic polities displays unity, has unlimited political and economic power, and typically acts on behalf of its own interests. But democracy should a priori impose some control on the power of the ruling elite. Indeed, Schumpeter (1954) claimed that the democratic process permits ‘free competition among would-be leaders for the vote of the electorate’ and that the masses can choose between various elites. In contrast, classical elite theorists such as Mosca (1939), Pareto (1935), Michels and Mills emphasized that there can be collusion even in democracies. Numerous elites may not be mutually competitive and may not control and balance each other; instead, they may be intertwined as a unanimous, cohesive power elite.

Economic consequences of the extent of interconnection

Inequality

The elite's plurality and competition ensures its responsiveness to the demands of the public, while a consensual elite might use its power for its own interests. Etzioni-Halevi (1997) claims that a unified elite does not use its power to reduce inequality and promote the development of a more egalitarian society, due to common recruitment and common interests. It is the plurality and differentiation of the members of the elite that enables them to countervail each others’ power and to increase their responsiveness to the will of public. In consequence, elite homogeneity might actually increase the gap between the elite and the masses.

When the political elite controls wealth and the main factors of production, then elite and class stratifications coincide, and consequently power and wealth are in the hands of the same happy few. Engerman and Sokoloff (1997) showed that members of the elite who have power and wealth establish institutions that serve their own interests and exclude the masses from benefits. In consequence, inequality persists through institutional development in the elite's own favour. Justman and Gradstein (1999) added that elite unity leads to greater inequality through regressive redistribution policy. A power elite that controls wealth may refrain from investment in human capital of the majority because education would increase the latter's political voice and weaken the elite's hold on power (Easterly, 2001); yet in some cases, the elite deliberately decides to forfeit power by investing in human capital as a consequence of a cost–benefit analysis (Bourguignon and Verdier, 2000).

The extent of elite unity can be endogenously determined (Sokoloff and Engerman, 2000), and elite

unity can also be affected by revolutions, wars and economic growth. Justman and Gradstein (1999) argue that economic growth dilutes the power of the elite by broadening political participation and reducing inequality.

Economic growth

A strong interconnection among elites has the consequence that all sectors of the economy are ruled by a group that thinks in a monolithic way. Two lines of thoughts have related a monolithic group to economic growth. The first one underlines that a monolithic group leads to the stagnation of ideas and attitudes, which in turn may prevent the adoption of major technological breakthroughs (Bourdieu, 1977). The lack of competition in a monolithic powerful group also generates corruption, with harmful consequences for growth.

The second line of thought argues that wealthy elites with enough political power to block changes will not accept adopting institutions that would enhance growth, since they might hurt them. Acemoglu, Johnson and Robinson (2001) developed this line of thought in relation to colonial impacts, showing that, wherever colonial governments were composed of few elite members, economic progress was reduced.

Following the same line of reasoning, Acemoglu and Robinson (2000) and Gradstein (2007) stressed that elite plurality, in which the political and economic elites are separate, explains the adoption of political franchise and industrialization in western Europe; while 19th-century eastern Europe, where elite unity was strong, did not adopt growth-enhancing institutions, since its elites held on to their wealth and power.

Paradoxically, in countries in which the elite was united and consensual, with common aims, the transition to capitalist production in the 1990s took place without violence, as in Poland and the Czech Republic. In contrast, wherever the elite was divided and fragmented, there were conflicts, especially on the ethno-nationalist level, as in Yugoslavia and Romania (Pakulski, 1999).

Recruitment and training of elites

Plato claimed that government should be in the hands of the most able members of society, that is, the *aristocracy* (Greek for ‘rule by the best’), a term that became pejorative and was later changed to *meritocracy* (coined by Young, 1958). Pareto argued that a stable economic system needs a *circulation of elites*, so that the most capable and talented are in the governing class. He stressed that the quality of the ruling class can be maintained only if social mobility is allowed, so that the non-elite has the possibility of entering the elite: ‘History is a cemetery of aristocracies’ (Pareto, 1935). His theory may be viewed as a sort of social Darwinism in which mobility is needed, just as evolution relied upon competition and selection.

For millennia, recruitment of the Western elite was based on social inheritance and was carried out via heredity, nepotism and violence. Hereditary monarchy was considered the most legitimate means of recruitment for rulers, and the upper elite was made up of wealthy large landowners, an *état de fait* considered normal in agrarian societies. Nevertheless, there were some channels of entrance into the

elite, such as military prowess and exploits or involvement in government finance (Brezis and Crouzet, 2004).

In democracies, the political elite came to be recruited mainly by election. Yet for a long time, the franchise was not for all. Big landowners and members of the upper middle class were the overwhelming majority in parliaments and cabinets, even though some prominent business people entered the political elite. Only in the late 19th century did members of the lower middle class and working class enter the political elite.

From the 19th century onwards, the circulation of the business elite took two differing yet concurrent paths. The first was that economic growth led to spurts of new firms and the decline of others, allowing a new business elite to emerge (Schumpeter, 1961). The second path was the rise of the professions, with competitive and meritocratic exams that led to circulation of elites (Perkin, 1978). After the Second World War, the elite was mainly recruited through education into elite universities to which admission started to be conferred following success at meritocratic exams.

Economic consequences of the recruitment of elites

Social mobility in the economy

Prior to recruitment through meritocracy, social mobility, and in particular the potential for non-elite members to enter the elite, was low. Temin (1999a; 1999b) showed that today, as in the 1900s, and despite meritocracy, the American economic elite is composed almost entirely of white Protestant males who have been educated for the most part in Ivy League colleges. Although in 1900 the political elite was quite similar to the business elite, today the former is more diversified; the political elite has changed in its recruitment, while the economic elite has not. In other words, minorities have not penetrated the economic elite in the United States (see also Friedman and Tedlow, 2003, which summarizes studies on US elite mobility, and Foreman-Peck and Smith, 2004 on British elites).

Recruitment to a university through meritocratic entrance exams, does not, indeed, lead to enrolment from all classes of society according to distribution or ability, nor does it necessarily lead to the admission of the most talented. Recruitment by entrance exam still encompasses a bias in favour of elite candidates because this type of exam requires a pattern of aptitude and thinking that favours candidates from an elite background. All elite positions may be open to all applicants with the right qualifications, but they are more accessible to those with specific social, cultural and symbolic capital (Arrow, Bowles and Durlauf, 2000). Thus the power elite maintains its status and power by a *strategy of distinction*, or a cultural bias that is necessary for accessing it (Bourdieu, 1977). A small difference in culture and education leads to narrow recruitment, and in turn to class-based stratification in the recruitment of the elite, despite meritocratic selection for universities (Brezis and Crouzet, 2006).

The relationship between mobility and the political system, as emphasized by Pareto, has been analysed by sociologists. For instance, Lengyel (1999) showed that circulation in the elite occurs at times of political upheavals and revolutions: the existing elite is eliminated and replaced by a new one. The first-generation members of the elite following a political change have neither specific training and education nor specific origin; they are the trailblazers, the entrepreneurs who seized power on the strength of their

competence. In the next generation, the elite becomes narrowly recruited from the best educated, and members are selected mostly by training and education. The elite returns to an occupational specialization, similar to the meritocratic profession criterion of earlier industrialization (Perkin, 1978).

Economic growth

A crucial element of economic growth is that the recruited elite be of the highest quality. Countries in which elites are recruited in a non-meritocratic way face the problem of the quality of their elites. However, the prevalence of meritocratic recruitment does not necessarily lead to the selection of the best ruling elites. Brezis and Crouzet (2006) argue that, when a country faces only mild technological and structural changes, the narrow recruitment, due to meritocracy, optimally fulfils its purpose, since the cultural bias of the elites is an advantage in the given type of technology. However, at times of major changes in technology, elites recruited this way are not the best for adopting new technologies. Moreover, the homogeneity of the recruitment of elites through similar curricula leads to convergence of views; this, in turn, leads to a monolithic elite, which, as we have claimed above, may have negative consequences for economic growth.

Conclusion

In this short article, we have summarized the modern research that has examined recruitment schemes and incentives for elites to discover how they can be used to promote, rather than impede, economic growth. There is also an entire economic history literature that has enriched us with a wealth of knowledge on the business elite. The main works in this literature are by Cassis (1997), Crouzet (1999) and Lachmann (2000).

The literature cited herein seems to show that the structure of this small group called the elite has numerous effects on the world economy. In the opposite direction, globalization will also affect the elite, as we are now facing a globalization of education of the elite.

In its first wave, globalization of education will probably create a new collection of elites and elicit some changes, yet the unity and uniformity of the elite will be even greater, not only at the national level but also at the global level. National elites will be replaced by a worldwide elite, along with uniformity in culture and education. We will face an international technocratic elite with its own norms, ethos, and identity, as well as its private clubs like the Davos World Economic Forum – a transnational oligarchy.

See Also

- economic growth, empirical regularities in
- income mobility
- Pareto, Vilfredo
- social status, economics and

Bibliography

Acemoglu, D., Johnson, S. and Robinson, J.A. 2001. The colonial origins of comparative development: an empirical investigation. *American Economic Review* 91, 1369–401.

Acemoglu, D. and Robinson, J. 2000. Political losers as a barrier to economic development. *American Economic Review* 90, 126–44.

Arrow, K., Bowles, S. and Durlauf, S. 2000. *Meritocracy and Economic Inequality*. Princeton: Princeton University Press.

Bourdieu, P. 1977. *Reproduction in Education, Society and Culture*. London: Sage.

Bourguignon, F. and Verdier, T. 2000. Oligarchy, democracy, inequality and growth. *Journal of Development Economics* 62, 285–313.

Brezis, E.S. and Crouzet, F. 2004. Changes in the training of the power elites in Western Europe. *Journal of European Economic History* 33, 33–58.

Brezis, E.S. and Crouzet, F. 2006. The role of higher education institutions: recruitment of elites and economic growth. In *Institutions and Economic Growth*, ed. T. Eicher and Garcia-Penalosa. Cambridge, MA: MIT Press.

Brezis, E.S. and Temin, P. eds. 1999. *Elites, Minorities, and Economic Growth*. Amsterdam: Elsevier.

Cassis, Y. 1997. *Big Business. The European Experience in the Twentieth Century*. Oxford: Oxford University Press.

Crouzet, F. 1999. Business dynasties in Britain and France. In Brezis and Temin (1999).

Easterly, W. 2001. The middle class consensus and economic development. *Journal of Economic Growth* 6, 317–35.

Engerman, S.L. and Sokoloff, K.L. 1997. Factor endowments, institutions, and differential paths of growth among New World economies: a view from economic historians of the United States. In *How Latin America Fell Behind: Essays on the Economic History of Brazil and Mexico, 1800–1914*, ed. S. Haber. Stanford: Stanford University Press.

Etzioni-Halevy, E. 1997. *Classes and Elites in Democracy and Democratization*. New York: Garland.

Foreman-Peck, J. and Smith, J. 2004. Business and social mobility into the British elite 1870–1914. *Journal of European Economic History* 55, 485–518.

Friedman, W.A. and Tedlow, R.S. 2003. Statistical portraits of American business elites: a review essay. *Business History* 45, 89–113.

Gradstein, M. 2007. Inequality, democracy and the protection of property rights. *Economic Journal* 117, 252–69.

Higley, J. and Burton, M. 1989. The elite variable in democratic transitions and breakdowns. *American Sociological Review* 54, 17–32.

Justman, M. and Gradstein, M. 1999. Industrial Revolution, political transition and the subsequent decline in inequality in nineteenth-century Britain. *Explorations in Economic History* 36, 109–127.

Lachmann, R. 2000. *Capitalist in Spite of Themselves: Elite Conflict and Economic Transition*. Oxford: Oxford University Press.

Lengyel, G. 1999. Two waves of professionalization of the Hungarian economic elite. In Brezis and Temin (1999).

Michels, R. 1915. *Political Parties*. New York: Dover, 1959.

Mills, C.W. 1956. *The Power Elite*. New York: Oxford University Press.

Mosca, G. 1939. *The Ruling Class*. New York: McGraw.

Pakulski, J. 1999. Elites, ethnic mobilization and democracy in postcommunist Europe. In Brezis and Temin (1999).

Pareto, V. 1935. *The Mind and Society*. New York: Harcourt Brace.

Perkin, H. 1978. The recruitment of elites in British society since 1800. *Journal of Social History* 12, 222–34.

Sokoloff, K. and Engerman, S. 2000. Institutions, factor endowment and paths of development in the New World. *Journal of Economic Perspectives* 14(3), 217–32.

Schumpeter, J.A. 1954. *Capitalism, Socialism and Democracy*. London: Routledge.

Schumpeter, J.A. 1961. *Theory of Economic Development*. New York: Oxford University Press.

Temin, P. 1999a. The American business elite in historical perspective. In Brezis and Temin (1999).

Temin, P. 1999b. The stability of the American business elite. *Industrial and Corporate Change* 8, 189–210.

Young, M. 1958. *The Rise of Meritocracy, 1870–2033*. London: Thames & Hudson.

How to cite this article

Brezis, Elise S. and Peter Temin. "elites and economic outcomes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000263>
doi:10.1057/9780230226203.0464

The New Palgrave Dictionary of Economics Online

Ellet, Charles, Jr. (1810 – 1862)

Robert B. Ekelund, Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

differential calculus; duopoly; Ellet, C., Jr; firm, theory of; market area analysis; monopoly; price discrimination

Article

American engineer and economic theorist, Ellet was born on 1 January 1810 at Penn's Manor, Pennsylvania, and died on 21 June 1862, a victim of the Civil War. Ellet grew up on a family farm but showed little inclination for agriculture: at age 17 he joined a surveying crew. With no formal education or training, he soon became an assistant engineer to Benjamin Wright, chief engineer of the Chesapeake and Ohio Canal. With ability and hard work Ellet taught himself mathematics and French, earning the respect of influential engineers. Letters of introduction to Lafayette and the American ambassador helped secure Ellet a place at the Ecole des Ponts et Chaussées, Dupuit's alma mater, in 1830. On his return to America in 1832 Ellet became the premier suspension bridge designer in America, building in 1849 the (then) longest suspension bridge in the world across the Ohio River at Wheeling. Colonel Ellet designed, constructed and commanded the ram fleet of the Union forces at the naval battle at Memphis, Tennessee. He died as a result of a wound received in the heat of that battle.

Ellet spent most of his professional life as an engineer, but, in one major work and in a number of contributions to the *Journal of the Franklin Institute* between 1840 and 1844, he significantly advanced the economic theory of monopoly, input selection, spatial economics, benefit–cost theory and econometric estimation. All Ellet's contributions were facilitated by the use of the differential calculus, which permitted him to express the simple theory of the firm, and some of its extensions, in mathematical terms. In his *Essay on the Laws of Trade* (1839) Ellet established the demand curve for a monopoly railroad with distance as a variable. Utilizing first-order conditions and solving for the gross toll on passenger traffic, Ellet demonstrated that the profit-maximizing toll would be equal to one-half the costs of transportation added to a constant quantity, a well-known result.

Ellet considered not one monopoly model but a multiplicity of them, including those dealing with freight transport, duopoly conditions and the principles of monopoly price discrimination. Further, Ellet's

particular insights into simple and discriminatory pricing systems led him to provide, with distance as a variable, an amazingly complete mathematical and graphical analysis of the impact of changes in the pricing system upon the market area served by a profit-maximizing railroad (1840a). In this important contribution to market area analysis Ellet argued that a set of (constrained) discriminatory tolls inverse to distance, in contrast to tolls proportional to distance, could be devised whereby all interested parties (management, shippers, the state) could be made better off. In a series of papers (1842–4) Ellet extended his theoretical analysis of inputs and input selection (1839) to one of the earliest attempts to develop, empirically specify and test a theoretical cost function. Utilizing a ‘law’ of costs which included his selected determinants of annual total railway costs, Ellet estimated the empirical dimensions from data collected from the mid-1830s. He then reaffirmed the power of his initial equation with new and supplementary data.

In all, the calibre and completeness of Ellet's theoretical and empirical inventions would not compare unfavourably with those of von Thünen, Cournot, Dupuit or Lardner. Ellet, who was primarily an engineer, was America's best representative among the pioneer contributors to scientifically oriented economics in the 19th century.

Selected works

1839. *An Essay on the Laws of Trade in Reference to the Works of Internal Improvement in the United States*. Richmond. Reprint from the 1st edn, New York: Augustus Kelley, 1966.

1840a. The laws of trade applied to the determination of the most advantageous fare for passengers on railroads. *Journal of the Franklin Institute* 30, 369–79.

1840b. A popular exposition of the incorrectness of the tariffs of tolls in use on the public improvements of the United States. *Journal of the Franklin Institute* 29, 225–32.

1842–4. Cost of transportation on railways. *Journal of the Franklin Institute*, various issues.

Bibliography

Baumol, W. and Goldfeld, S.M. eds. 1968. *Precursors in Mathematical Economics: An Anthology*. London: London School of Economics and Political Science.

Calsoyas, C.D. 1950. The mathematical theory of monopoly in 1839: Charles Ellet, Jr. *Journal of Political Economy* 58(April), 162–70.

Ekelund, R.B., Jr. and Hooks, D. 1972. Joint demand, discriminating two-part tariffs and location theory: an early American contribution. *Western Economic Journal* 10, 84–94.

Viner, J. 1958. *The Long View and the Short: Studies in Economic Theory and Policy*. New York:

Glencoe.

How to cite this article

Ekelund, Robert B., Jr. "Ellet, Charles, Jr. (1810–1862)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000062> doi:10.1057/9780230226203.0465

The New Palgrave Dictionary of Economics Online

Ely, Richard Theodore (1854 – 1943)

A.W. Coats

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Commons, J.R.; Ely, R.T.; German Historical School; institutionalism; Knies, K.G.A.

Article

Ely was born in Ripley, New York, on 13 April 1854 and died at Old Lyme, Connecticut, on 4 October 1943.

Ely's long and vigorous career epitomizes the general proposition that an economist can exert a major constructive influence on his subject and profession even though his original contribution to economic theory is negligible. A highly effective teacher and maker of careers for his former students; prolific author of popular articles, scholarly volumes, and publications series; organizer and fund-raiser for major research projects; founder of various academic institutes and associations; leader or participant in numerous reform societies; and centre of innumerable controversies, Ely was the most widely known, even notorious, economist in the USA around the turn of the 20th century.

After a brief spell as a country schoolteacher and a preliminary year at Dartmouth College, Ely graduated from Columbia College in 1876 and was awarded a three-year fellowship to study philosophy in Germany. He soon switched to political economy, came under the influence of Karl Knies at Heidelberg, where he obtained a Ph.D., summa cum laude, in 1878, and later attended Adolph Wagner's lectures in Berlin. Returning to the USA he was unemployed for more than a year before his appointment, initially on a half-time basis, at Johns Hopkins, where he taught from 1881 to 1892. He then moved to Wisconsin, founding an outstanding school of Economics, Political Science and History including such luminaries as F.J. Turner, E.A. Ross, and J.R. Commons. A unique collaboration developed between the social scientists and the state legislators, especially under the La Follette governorship, which pioneered major social and economic reform legislation. In 1925 Ely took his Institute for Research in Land Economics and Public Utilities, founded in 1920, from Madison to Northwestern University, and remained there until 1932, when he launched a new, but impoverished Institute for Economic Research in New York City. Eventually hit by the depression, Ely was forced to

depend on the support of friends and former students as he completed his autobiography and failed to complete a massive history of American economic thought initiated 50 years earlier.

An ardent Christian Socialist and outspoken critic of laissez-faire individualism and ‘old school’ English classical economics, Ely delighted social reformers and outraged conservatives by his writings on such controversial current topics as socialism and the American labour movement. Prone to emotional overstatement and careless in exposition, his public pronouncements and reputation frequently embarrassed the aspiring young professional economists with whom he founded the American Economic Association, in 1885, and for a time discouraged some moderate and conservative economists from joining. Although Ely's original draft prospectus had been rejected, and the association's original constitution was toned down, and then dropped, the organization hovered uneasily between missionary evangelism and scholarly objectivity until he was obliged to relinquish his secretaryship in 1892.

Two years later, at Wisconsin, Ely's fellow professionals rallied around him when he was denounced for preaching socialism and encouraging strikes, and, although he was completely exonerated in a ‘trial’ that attracted national attention, Ely gradually became more conservative. Ironically, in the 1920s his institute was attacked, no doubt unfairly, as a tool of the public utilities, and was referred to disparagingly in a report on professional ethics by a committee of the American Association of University Professors, in 1930.

During his long lifetime Ely wrote extensively on an extraordinarily wide variety of topics, often in a popular and journalistic fashion. Nevertheless, he repeatedly opened up new research topics that were developed by his colleagues and former students – for example, in labour history, state taxation, land economics, and natural resources – and his various textbooks, especially the multi-edition *Outlines of Economics* which sold 350,000 copies, were both widely used and highly regarded.

At Wisconsin he helped to launch the American Association for Labor Legislation, of which he became President, and raised private resources to finance John R. Commons's massive *Documentary History of American Society* (11 vols, 1910–11). He served as President of the American Economic Association in 1900–1901.

Ely was a stimulating teacher whose ideas formed a direct link between the doctrines of the German Historical School and American institutionalism, a link most clearly evident in his neglected two-volume study of *Property and Contract in their Relations to the Distribution of Wealth* (1914). Many of his students went on to distinguished careers in academic and/or public life. He was undoubtedly an outstanding academic entrepreneur, and his contribution to the American Economic Association is recognized in its annual invited Richard T. Ely lecture, which was inaugurated in 1963.

Selected works

1883. *French and German Socialism in Modern Times*. New York: Harper & Brothers. Reprinted, 1911.

1884a. *The Past and the Present of Political Economy*. Baltimore: N. Murray for Johns Hopkins.

1884b. *Recent American Socialism*. Baltimore: N. Murray for Johns Hopkins. Reprinted, 1885.

1886. *The Labour Movement in America*. New York: T.Y. Crowell Co. New edn, revised and enlarged,

New York: Macmillan Co, 1905.

1888a. *Taxation in American States and Cities*. New York: T.Y. Crowell Co.

1888b. *Social Aspects of Christianity*. Boston: W.L. Greene and Co.

1889a. *An Introduction to Political Economy*. New York: Chautauqua Press. New and revised edn, New York: Eaton and Mains; Cincinnati: Jennings and Pye, 1901.

1889b. *Social Aspects of Christianity and other Essays*. New York: T.Y. Crowell Co. Reprinted, 1895.

1893. *Outlines of Economics*. Meadville, Pennsylvania and New York: Flood and Vincent. 6th edn with Ralph Hess, New York: Macmillan Co., 1938.

1894. *Socialism: An Examination of its Nature, its Strength, its Weakness. With Suggestions for Social Reform*. London: S. Sonnenschein Co.; New York and Boston: T.Y. Crowell Co.

1900. *Monopolies and Trusts*. New York: Macmillan Co. Reprinted, 1912.

1903. *Studies in the Evolution of Industrial Society*. New York and London: Macmillan Co. Reprinted, 1918.

1914. *Property and Contract in their Relations to the Distribution of Wealth*. 2 vols. New York: Macmillan Co. Reprinted, 1922.

1924. (With E.W. Morehouse.) *Elements of Land Economics*. New York: Macmillan Co. Reprinted, 1932.

1928. (With G.S. Wehrwein.) *Land Economics*. Ann Arbor, Michigan: Edwards Bros. Revised edn, Madison: University of Wisconsin Press, 1964.

1938. *Ground Under Our Feet: An Autobiography*. New York: Macmillan Co.

Bibliography

Rader, B.G. 1966. *The Academic Mind and Reform: The Influence of Richard T. Ely in American Life*. Lexington: University of Kentucky Press.

How to cite this article

Coats, A. W. "Ely, Richard Theodore (1854–1943)." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000063> doi:10.1057/9780230226203.0466

The New Palgrave Dictionary of Economics Online

emergence

Yannis M. Ioannides
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

With its philosophical pedigree and its use especially among life scientists and science writers since the early 1990s, the term ‘emergence’ in economics is more evocative than precise, reflects influence from physics and biology, and is now associated with phenomena where economic structures evolve into qualitatively different forms. These exhibit properties that are *emergent* in that they apply at an aggregate level but lack individual analogues and therefore are not describable at the individual level. This article emphasizes applications that possess firm economic foundations, from the evolution of patterns in international trade to the establishment of a common currency.

Keywords

aggregation; autarky; class; Cobb–Douglas functions; complementarities; congestion; division of labour; economic geography; elasticity of substitution; emergence; fiat money; financial market globalization; first fundamental theorem of welfare economics; foreign aid; herding; innovation; international currencies; international currency; international finance; international trade; invisible hand; Mill, J. S.; multiple equilibria; neighbours and neighbourhoods; poverty traps; power laws; residential segregation; specialization; spontaneous order; stochastic stability theory; tipping points; trade costs; urban agglomeration; urbanization

Article

Having acquired widespread use among life scientists and science writers since the early 1990s, the term ‘emergence’ in economics is more evocative than precise, reflects influence from physics and biology, and has come to be associated with phenomena involving evolution of economic structures into qualitatively different forms. These phenomena exhibit properties that are emergent in the sense that they are novel and apply at an aggregate more ‘complex’ level but lack individual analogues and therefore are not describable at, or reducible to, the individual level. A good case in point is the statement that consciousness is an emergent property of the brain. The notion of emergence originates in the philosophy of science, with John Stuart Mill being an important precursor (see Stanford Encyclopedia of Philosophy, 2002).

This article reviews, albeit selectively, the recent usage of the term by emphasizing applications with predominantly economic phenomena where emergence of macroscopic properties may be elucidated by means of economic arguments. These range from neighbourhood tipping and evolution of patterns in international trade to emergence of urban structure and the establishment of norms and institutions and of a common currency, among many others.

More generally, emergent properties or behaviours have been studied in a variety of circumstances in nature, such as emergence of differentiated behaviour in colonies of animals, of herding behaviour in organizations and markets, of specialization of individuals into occupations and of cities and of regions and countries in specific products, of groups of biological cells in multicellular biological organisms and even of groups of processors in computer simulations involving cellular automata (see Holland, 1998). The World Wide Web is an example of a decentralized engineering system that is continuously being modified by human initiatives in the form of actions by individuals and firms. The web has not been deliberately designed and no central organization administers how different sites are linked to others. Some of the properties of the graph topology of the web may be termed as *emergent*, such as that the number of links pointing to each page follows approximately a power law, with a few pages being pointed to by many others and most others seldom, and the fact that any pair of pages can be connected to each other through a relatively short chain of links in the average.

The presence of ‘emergence’ within the vocabulary of economists does suggest some interplay with multidisciplinary research by scientists who have been associated with the Santa

Fe Institute (http://www.santafe.edu). To quote from Kauffman (1995, p. 24), an alternative definition of emergence is that '[t]he whole is greater than the sum of its parts'. And 'life itself is an emergent phenomenon … arising as the molecular diversity of a prebiotic chemical system increases beyond a threshold of complexity. If true, then life is not located in the property of any single molecule – in the details – but is a collective property of systems of interacting molecules.' The entirety of complex molecules together is able to reproduce and evolve, a 'stunning property'.

Blume and Durlauf (2001) argue that emergence plays an important role even within the body of neoclassical economics proper. For example, the extent to which macroeconomics is a distinct discipline from microeconomics would be explained by emergent properties as alluded to by the statement 'aggregation is not summation' (see Kirman, 1992). Consider, within microeconomics and general equilibrium theory, the metaphor of the invisible hand of the market (which goes back to Adam Smith), whereby individuals' pursuit of their own selfish aims leads to social outcomes that obey important social properties. Under certain conditions, after markets have brought about an equilibrium, it is impossible to make anyone better off without making someone worse off. Thus, the first fundamental theorem of welfare economics is an emergent property of social outcomes. However, the more modern work on emergence in economics has emphasized emergence of patterns. Similarly, Hayek's concept of spontaneous order may be considered an instance of emergence.

There are numerous other contexts where emergence has been alleged to occur. This article explores a number of examples of emergence that are limited to social and economic settings. They underscore the scope of the concept of emergence in such settings. As discussed earlier, there are many other contexts in socioeconomic settings and beyond, ranging from computation to the life sciences.

Emergent social interconnections

Suppose that a society consists of *I* individuals, where *I* is large, where any two individuals may be linked in a way that allows for communication, social relations, or social interactions. Let p_k denote the probability that each individual is connected with exactly *k* other individuals. A literature going back to Erdös and Renyi (1960) and continuing at the time of writing up to Newman, Strogatz and Watts (2001) has studied the topological properties of the (random) graph formed by the agents as nodes and connections between agents as edges when each agent's connections with other follows a given distribution p_k and the number of agents is large. According to Newman, Strogatz, and Watts (2001), depending upon whether the quantity $E[k^2] - 2E[k]$ is greater than or equal to 0, or falls below 0, there emerges, as *I* tends to infinity, a *proportion* of all individuals being interconnected, or, alternatively, the economy consists of different groups of finite sizes. In other words, the social structure undergoes a *phase transition* when this quantity exceeds 0: a giant interconnected component emerges. Intuitively, starting from a connected component of the graph, consider adding a new edge that connects with a previously isolated node of degree *k*. Doing so will change the number of nodes on the boundary of the connected component by $-1 + (k-1) = k-2$. The likelihood that a node is on the boundary of the connected component is proportional to *k*. The expected change in the number of nodes on the boundary when an additional node is connected is given by $\Sigma_i k_i(k_i - 2) / \Sigma_i k_i$. If this quantity is negative, then the number of nodes on the boundary decreases and therefore the connected component will stop growing. If it is positive, on the other hand, then the number of boundary nodes will grow and the connected component will grow, limited only by the size of the network.

In the simple case of the Erdös and Renyi random graph, where the number of connections is proportional to the number of individuals, the phase transition occurs when the factor of proportionality is equal to $\frac{1}{2}$ and the corresponding average number of connections per person is equal to 1. Below this value, there are too few edges and the components of the random graph are small; above that value, a proportion of the entire graph belongs to a single, *giant* component. In this case, emergence of a qualitatively different social structure depends on the value of a single parameter (Kirman, 1983; Ioannides, 1990; Durlauf, 1997). Individual behaviour that leads to a law for the number of individuals' connections does not necessarily imply the same macroscopic outcome in all circumstances. Similarly, social outcomes are not described by means of mere summation of individual actions; aggregation is not summation (Kirman, 1992). Kauffman (1995, p. 57) invokes this in the context of autocatalytic reactions and goes as far as seeing this 'as a toy version of phase transition that I believe led to the origin of life'.

Patterns of residential segregation

Now we turn to a description of neighbourhood tipping, which is originally due to Thomas C. Schelling (1978) and has been adapted here from recent works. Suppose that individual *i* is white and would live in a neighbourhood provided that the percentage of whites among her neighbours, $\omega \in [0, 1]$, is at least w_i, $\omega \geq w_i$. She moves out otherwise. Individuals differ in terms of preference characteristic w_i, which is assumed to be distributed in a typical neighbourhood according to $F(w)$, when the analysis starts. For any neighbourhood with a share of white residents equal to ω, the percentage of white individuals who would find living there acceptable are those with $w < \omega$. Their share is given by the value of the cumulative distribution function at ω, $F = F(\omega)$.

In Figure 1, let the horizontal axis e_1 denote ω and w_i, the vertical axis e_2 the cumulative distribution *F*, and $(0, \bar{0})$ the 45-degree line. As long as $\omega > F(\omega)$, whites have an incentive to exit the neighbourhood, causing a reduction of ω, and this process continues until there are no whites left; $\omega = 0$. If, on the other hand, $\omega \leq F(\omega)$, additional whites have

an incentive to enter, and this process continues until $\omega = 1$. Thus, the process has three equilibria, $(0, 0^*, \bar{0})$, of which the two extreme ones, either only blacks or no blacks in the neighbourhood, are stable, and the mixed one, with ω^* whites in the neighbourhood, where $\omega^* = F(\omega^*)$, unstable. The mixed equilibrium defines the *tipping point*. Individuals' preferences differ widely, but only extreme outcomes *emerge* at the social equilibrium. Schelling (1978) underscores how outcomes that persist may not be what individuals had intended.

Figure 1

Neighbourhood tipping, poverty traps

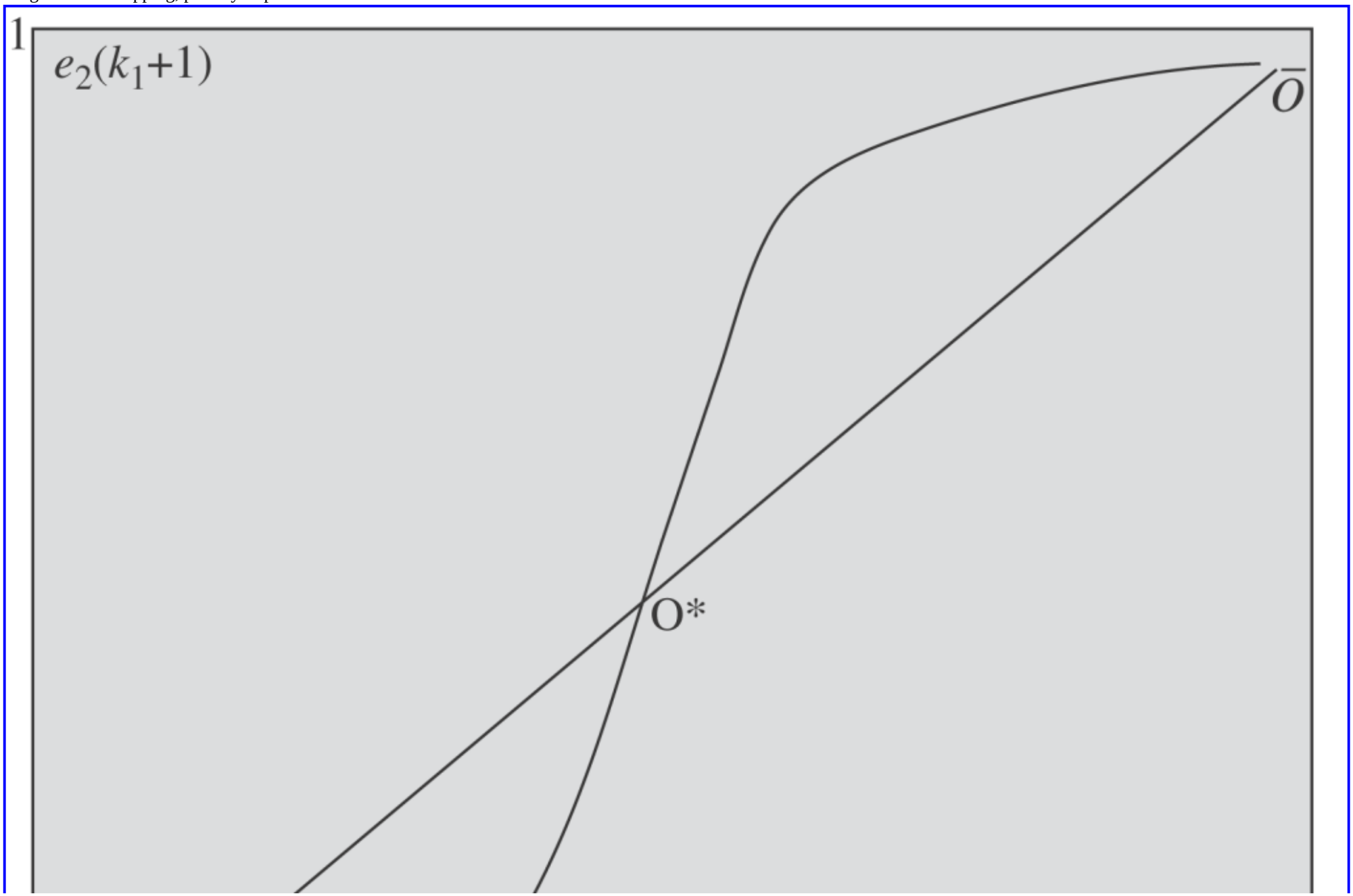

Could such a stark outcome be due to the fact that the respective populations of individuals are not being replenished? It turns out that, if one goes deeper and allows for turnover and stochastic shocks, persistence of stable states may be rigorously characterized by means of the tools of stochastic stability theory (Blume and Durlauf, 2003; Young, 1998). Multiplicity of equilibria allows, of course, for accidents of history to become reinforced over time.

Emergence of urbanization

The concentrated economic activity that we associate with the emergence of cities punctuates the physical and economic landscape throughout the world. How did it emerge? While small-scale agriculture and home production could be reasonably accurately referred to as spatially uniform distribution of economic activity, the world population is increasingly concentrated in cities. Also, urbanization has been closely associated with economic development.

Let us consider a simple setting where utility *U* depends on individual productivity, itself an increasing function $f(n_\ell)$, of the total number of others in the same location, n_ℓ, and on the share of a fixed resource, $\frac{R}{n_\ell}$. Even when utility is assumed to be increasing and concave in both arguments, it is initially increasing, as a function of n_ℓ, may reach a peak at n^*, and then may start decreasing. In other words, a larger population initially means more innovation and mutually beneficial interaction until congestion offsets them. Consider then two alternative locations, $\ell = 1, 2$, that do not interact spatially, and a total of *N* individuals who wish to locate so as to maximize utility. At a locational equilibrium, individuals must be indifferent as to where they locate. If $N < 2n^*$, the symmetric equilibrium, where $n_1 = n_2 = \frac{1}{2}N$, is unstable and agglomeration – that is, either site occupied by the entire population – is stable. Therefore, the trade-off between the value of agglomeration and the cost of congestion moves the economy away from the symmetric outcome (Anas, 1992).

Consider next a setting where interactions do explicitly depend on distance to others, as with accessibility to others being valued and congestion disliked. If individuals are allowed to relocate, with probabilities that depend on expected utilities in each site relative to all other sites, then a dynamic model may be formulated that describes locational outcomes for an entire population. The economy may attain steady states that are either uniform (populations are equal across all sites) or uneven (with some sites having large and others small populations). Such a stylized reduced-form model of spatial patterns of human settlements (see Papageorgiou and Smith, 1983) yields spatially uniform outcomes that are either stable or unstable. Agglomeration is determined by the interplay between the value of agglomeration and the cost of congestion. If the former dominates, spatially uniform steady states are unstable. Fujita, Krugman and Venables (1999, chs 6 and 17) develop a model with ingredients from economic geography that incorporates trading costs and also allows for uniform distributions of economic activity to exhibit different stability properties. Again, conditions under which agglomerations prevail possess intuitive economic appeal.

Emergence of poverty traps

In a standard neoclassical growth model that extends over discrete time, with a demographic structure consisting of two overlapping generations and individuals living for two periods, working only in the first and retiring in the second, individual savings would be proportional to the wage rate under Cobb–Douglas preferences. Let the aggregate production function expressing output Y_t as a function of capital, labour and total factor productivity, K_t, L_t, A, respectively, be of the constant elasticity of substitution form,

$$Y_t = A\left(\delta K_t^{1-\frac{1}{\sigma}} + (1-\delta)L_t^{1-\frac{1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}.$$

If the elasticity of substitution is sufficiently small – that is, complementarity between capital and labour is high – and total factor productivity sufficiently large, the time map of the economy – that is, the amount of capital per person next period (axis e_2) as a function of the amount of capital per person in the present period (axis e_1) – may be loosely graphed, as in Figure 1. Therefore, depending upon the economy's starting point, it may end up at a steady state either with high or with low capital per person at a steady state. The mid-range (‘symmetric’) steady state is unstable. Therefore, conditions of productive complementarities, (even small) initial differences in capital per person, and possibly historical accidents as well across countries in terms of characteristics and endowments when growth starts, mitigate in favour of an explanation for inequalities in incomes per person across different countries. The same mechanism worldwide produces sharply different outcomes (see Azariadis and Stachurski, 2006, for an in-depth treatment).

Similar arguments may be developed in order to understand persistence in the inequality of the distribution of wealth within an economy. Matsuyama (2006) presents a model of emergent class structure, in which a society inhabited by inherently identical households may, depending upon parameter values, be endogenously split into the rich bourgeoisie and the poor proletariat. For some parameter values, the model has no steady state where all households remain equally wealthy. The model predicts emergent class structure or the rise of class societies. Even if every household starts with the same amount of wealth, the society will experience ‘symmetry breaking’ and will be polarized into two classes in steady state, where the rich maintain a high level of wealth partly due to the presence of the poor, who have no choice but to work for the rich at a wage rate strictly lower than the ‘fair’ value of labour.

It is worth noting that similar modelling tools may be used to express Adam Smith's famous dictum that ‘the division of labour is limited by the extent of the market’ and thus endogenize specialization (Weitzman, 1994). The division of labour emerges as individuals in an economy acquire specialized roles.

Emergent structures in international economics: autarky, specialization, and international currencies

Krugman (1995) and Matsuyama (1995) discuss how a world economy where all countries are initially identical and live in autarky (a ‘symmetric’ outcome) leads to a world that is separated into rich and poor regions, once countries engage in international trade. International trade *causes* specialization and agglomeration of different economic activities in different regions of the world to emerge, with some countries being rich and others poor. In several similarly motivated papers, Matsuyama (in particular, 2004; 2006) shows the effects of financial market globalization on the cross-country pattern of development in the world economy. In the absence of the international financial market, the world economy converges to the symmetric steady state, and the cross-country difference disappears in the long run. Financial market globalization causes the instability of the symmetric steady state and generates stable asymmetric steady states, in which the world economy is polarized into the rich and the poor. The world output is smaller, the rich are richer and the poor are poorer in these asymmetric steady states than in the (unstable) symmetric steady state. The model thus demonstrates the possibility that financial market globalization may cause, or at least magnify, inequality among nations, and that the international financial market is a mechanism through which some countries become rich at the expense of others. Furthermore, the poor countries cannot jointly escape from the poverty trap by merely cutting their links to the rich. Nor would foreign aid from the rich to the poor eliminate inequality; as in a game of musical chairs, some countries must be excluded from being rich.

Especially at times of political and economic upheavals, many different national currencies may circulate simultaneously within and across countries. From a modelling viewpoint, such circumstances fit neatly multiplicity of equilibria. Emergence of a particular currency as an international currency, which in turn depends on the degree of economic and financial integration, may be more of a decentralized phenomenon then the emergence and establishment of a national currency (Matsuyama, Kiyotaki and Matsui, 1993). To start with, a national currency is typically fiat money, whose use is decreed although not necessarily ensured. World monetary history suggests that a bewildering variety of commodities have served as medium of exchange, unit of account and store of value, and may have coexisted at times of financial uncertainties. It has been known at least since Menger (1892) that fiat money comes to dominate other options, thus leading to establishment of monetary equilibria, because individuals accept fiat money in trade when it is convenient and they trust that others will do the same. Such an outcome may be fragile, when trust in the currency is weakened, especially in time of war and other upheavals. Howitt and Clower (2000) employ ‘rules’ concerning transactor behaviour (instead of relying on a priori principles of equilibrium and rationality) to show computationally commodity ‘money’ as a possible emergent property of interactions between gain-seeking transactors who are unaware of any system-wide consequences of their own actions. Similar is the emergence of standards in new industries described by many writers.

Concluding remarks

The scientific literature, along with popular science literature, on emergence has sought to explain the emergence of persistent patterns as outcomes of dynamic interactions between individuals, groups of individuals and other entities. Such emergence is typically intrinsic to specific nonlinear dynamic processes and represents international currency. Not all possible outcomes may be sustained at equilibrium, and economic and political structures emerge as a result of self-organization. Future research needs to go beyond evolutionary thinking and also deal with emergence in the context of purposeful action by forward-looking agents, as opposed to social outcomes of decentralized interactions of many agents.

See Also

- poverty traps
- spontaneous order

Bibliography

Anas, A. 1992. On the birth and growth of cities: laissez-faire and planning compared. *Regional Science and Urban Economics* 22, 243–58.

Azariadis, C. and Stachurski, J. 2006. Poverty traps. In *Handbook of Economic Growth*, ed. P. Aghion and S.N. Durlauf. Amsterdam: North-Holland.

Blume, L.E. and Durlauf, S.N. 2001. The interactions-based approach to socioeconomic behavior. In *Social Dynamics*, ed. S.N. Durlauf and H. Peyton Young. Princeton, NJ: Princeton University Press.

Blume, L.E. and Durlauf, S.N. 2003. Equilibrium concepts for social interaction models. *International Game Theory Review* 5, 193–209.

Durlauf, S.N. 1997. Statistical mechanics approaches to socioeconomic behavior. In *The Economy as an Evolving complex System II*, ed. W.B. Arthur, S.N. Durlauf and D. Lane. Redwood City, CA: Addison-Wesley.

Erdös, P. and Renyi, A. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 17–61.

Fujita, M., Krugman, P.R. and Venables, A.J. 1999. *The Spatial Economy*. Cambridge, MA: MIT Press.

Holland, J.H. 1998. *Emergence: From Chaos to Order*. Reading, MA: Addison-Wesley.

Howitt, P. and Clower, R. 2000. The emergence of economic organization. *Journal of Economic Behavior and Organization* 41, 55–84.

Ioannides, Y.M. 1990. Trading uncertainty and market form. *International Economic Review* 31, 619–38.

Kauffman, S.A. 1995. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford: Oxford University Press.

Kirman, A.P. 1983. Communication in markets: a suggested approach. *Economic Letters* 12, 1–5.

Kirman, A.P. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6(2), 117–36.

Krugman, P.R. 1995. Complexity and emergent structure in the international economy. In *New Directions in Trade Theory*, ed. A.V. Deardorff, J. Levinsohn and R.M. Stern. Ann Arbor: University of Michigan Press.

Matsuyama, K. 1995. Comment on P. Krugman, ‘Complexity and Emergent Structure in the International Economy’. In *New Directions in Trade Theory*, ed. A.V. Deardorff, J. Levinsohn and R.M. Stern. Ann Arbor: University of Michigan Press.

Matsuyama, K. 2004. Financial market globalization, symmetry breaking and endogenous and endogenous inequality of nations. *Econometrica* 72, 853–84.

Matsuyama, K. 2006. The 2005 Lawrence R. Klein lecture: emergent class structure. *International Economic Review* 47, 327–60.

Matsuyama, K., Kiyotaki, N. and Matsui, A. 1993. Toward a theory of international currency. *Review of Economic Studies* 60, 283–307.

Menger, C. 1892. On the origins of money. *Economic Journal* 2, 239–255.

Newman, M.E.J., Strogatz, S.H. and Watts, D.J. 2001. Random graphs with arbitrary degree distribution and their applications. *Physical Review E* 64, 026118.

Papageorgiou, Y.Y. and Smith, R.S. 1983. Agglomeration as a local instability of spatially uniform steady-states. *Econometrica* 51, 1109–19.

Schelling, T.C. 1978. *Micromotives and Macrobehavior*. New York: Norton.

Stanford Encyclopedia of Philosophy. 2002. Emergent properties. Online. Available at http://plato.stanford.edu/entries/properties-emergent, accessed 17 November 2006.

Weitzman, M.L. 1994. Monopolistic competition with endogenous specialization. *Review of Economic Studies* 61, 45–56.

Young, H.P. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.

How to cite this article

Ioannides, Yannis M. "emergence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000313> doi:10.1057/9780230226203.0467

The New Palgrave Dictionary of Economics Online

emerging markets

Joshua Aizenman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The club of high-performing emerging markets is fairly concentrated in East Asia. Their TFP growth may not be extraordinary, though their growth rate is unprecedented. Factors argued to promote growth include trade, investment, external financing, and good governance. The importance of external financing is overrated – higher growth induces higher saving rate, allowing investment to be self-financed. Institutional changes as the key for take-off remains debatable – India and China took off without any prior major institutional overhaul. Allowing newcomers to challenge incumbents and the capacity to adjust policies to shocks may be the keys for sustainable growth.

Keywords

agency problems; Asian miracle; emerging markets; external financing; financial liberalization; financial risk; growth and governance; growth and institutions; growth and international trade; high-performing Asian economies; moral hazard; savings; shocks; Solow, R.; take-off; total factor productivity

Article

‘Emerging markets’ are countries or markets that are not well established economically and financially, but are making progress in that direction.
The growing focus on emerging markets follows exciting developments during the second half of the 20th century – the emergence of a growing class of (formerly) poor countries that took off, and managed to close half of their income gap with the OECD countries within a generation or two. Remarkably, from 1960 to 1989 seven high-performing Asian economies (HPAEs) experienced unprecedented growth rates of the real GDP per capita in the range of four to seven per cent. This phenomenon has been the focus of a notable research report by the World Bank (1992), whose title *The East Asian Miracle* suggests a possible, though controversial, interpretation. The big story of recent years has been that the two most populous countries, China and India, joined the HPAE club. With few exceptions (such as

Chile and Botswana), the club of high-performing emerging markets is fairly concentrated in East Asia. The HPAEs' remarkable growth rates during recent decades imply a sizable drop in global poverty rates, also entailing greater concentration of the incidence of extreme poverty, mostly in Africa (see Fischer, 2003). Yet the emerging markets phenomenon goes well beyond Asia, encompassing a growing share of developing countries that are closing, though at a lower rate than the HPAEs, their income gap with the OECD countries.

These developments were in sharp contrast to the pessimistic predictions made in the 1950–60s by several influential economic growth models (for a review, see Easterly, 1999). The HPAE experience dispelled most of these fears. The superior performance of the HPAEs illustrated that the fast growth option is viable, raising pertinent questions, and stirring a lively debate. While the World Bank (1992) dubbed the experience of the HPAEs a 'miracle', Young (1995) questioned this 'miraculous' interpretation, arguing that it is in line with Solow's growth model. Specifically, he reasoned that most of the growth has been the outcome of very high rates of investment in tangible and human capital, and a sizable increase in labour market participation. Controlling for these factors, Young found that the HPAEs' total factor productivity growth is in line with the historical experience of other countries. The debate about the role of accumulation in accounting for the HPAE experience is not over, yet the large drop of the growth rate of Japan in the 1990s, and the East Asian financial crisis of 1997, somehow deflated the 'East Asian miracle' hypothesis, suggesting the onset of Solow's growth convergence. Even if Young's thesis is correct, the speed and relative smoothness of the convergence of the HPAEs to the OECD's development level are without precedent. It raises questions about the obstacles preventing other countries from accomplishing this task, and about the ways to facilitate the take-off process in other regions.

The HPAE take-offs have been associated with fast growth of exports climbing, over time, the technology ladder of trade. This led to a lively debate about the importance of exports as the engine of growth: is the dominant causal association from exports to growth or vice versa? Earlier studies inferred that trade liberalization enhances growth (Ben-David, 1993; Edwards, 1998), a point disputed by Rodríguez and Rodrik (2001). Several authors revisited this issue, applying better controls, inferring strong growth effects of trade openness. Frankel and Romer (1999) applied measures of the geographic component of countries' trade to obtain instrumental variables estimates of the effect of trade on income. They inferred that ordinary least square (OLS) estimates understate the effects of trade, and that trade has a significant large positive effect on income. The contrast between the economic performance of the Soviet Union and that of China in the second part of the 20th century suggests another advantage of export orientation: it imposes a powerful market test on domestic output. Since exports must meet the quality and pricing tests of the global market, export-led growth limits potential distortions induced by 'growth promoting' domestic policies. Specifically, it prevents Soviet Union-type superficial economic growth induced by forced investment, growth that may result in inferior products that would be wiped out in the absence of protection. Export-oriented growth also forces countries to move faster towards the technological frontier in order to survive competitive global pressures.

Some of the obstacles preventing countries from taking off arise from political economy factors. Specifically, as growth is frequently associated with the emergence of new sectors and new elites, incumbent policymakers opt to block development in an attempt to preserve their rents and their grip on power. This phenomenon was vividly illustrated at the micro level by De Soto (1989), and was shown to

be a major impediment to growth (see Parente and Prescott, 2005). As the burden of the low growth would mostly affect future generations, the low growth equilibrium may persist with limited opposition. Proponents of this view point out that free commerce, both internal (between provinces or states in a union) and international, provides a powerful constraint on an incumbent's ability to block development. The importance of external financing and financial integration in the development process remains a hotly debated topic. Advocates of financial liberalization in the early 1990s argued that external financing would alleviate the scarcity of saving in developing countries, inducing higher investments and thus higher growth rates. In contrast, Rodrik (1998) and Stiglitz (2002) questioned the gains from financial liberalization. Indeed, the 1990s experience with financial liberalization suggests that the gains from external financing are overrated – the bottleneck inhibiting economic growth is less the scarcity of saving and more the scarcity of good governance. This can be illustrated by tracing the patterns of self-financing ratios, measuring the share of tangible capital financed by past national saving (see Aizenman, Pinto and Radziwill, 2004). Higher self-financing rates of the nation's stock of capital are associated with a significant *increase* in growth rates. Remarkably, the wave of financial reforms in the 1990s led to deeper diversification, where greater inflows from the OECD financed comparable outflows from developing countries, with little effect on the availability of resources to finance tangible investment. These findings are consistent with several interpretations. The first deals with risk: agents in various countries may react to exposure to financial risk differently. The desire to diversify these risks may lead to two-way capital flows, with little change in net positions (see Dooley, 1988). The ultimate obstacles limiting external financing may be related to acute moral hazard and agency problems – sovereign states, decision makers and corporate insiders pursue their own interests at the expense of outside investors (see Gertler and Rogoff, 1990; Stulz, 2005). An alternative interpretation follows Caroll and Weil (1994), who found that statistical causality runs from higher growth rates to higher saving rates. They conjectured that the growth-saving causality may be explained by habit formation, where consumers' utility depends on both present and past consumption. 'Habit formation', however, may be observationally equivalent to adaptive learning in the presence of uncertainty – in countries where private savings are taxed in arbitrary and unpredictable ways, credibility must be acquired as an outcome of a time-consuming learning process. In these circumstances, a higher growth rate provides a positive signal about the competence and the intentions of the administration, increasing saving and investment over time. Consequently, agents in countries characterized by greater political instability and polarization would be more cautious in increasing their saving and investment rates following a reform. Hence, accomplishing take-offs in Latin America may be much harder than in Asia, explaining Latin America's relatively low growth rate. (Various studies pointed out that policy uncertainty and political instability reduce private investment and growth; see Ramey and Ramey, 1995; Aizenman and Marion, 1999).

I close this review with an outline of open issues. The positive association between the equality of institutions and growth is well documented, yet the precise role of institutions in the development process remains debatable. Acemoglu et al. (2003) inquired how the colonial history of a developing country affects the quality of institutions, concluding that distortionary macroeconomic policies are more likely to be symptoms of underlying institutional problems rather than the main causes of economic volatility. Yet this interpretation does not satisfactorily explain the role of institutions in the growth process. The remarkable take-offs of China and India in recent decades, episodes directly

affecting about a third of the global population, cannot obviously be explained by reference to institutional changes. This suggests that there is no simple correspondence or causality between growth and institutions. A tentative answer is provided by Rodrik (1999), who identifies a nonlinear interaction between shocks, polarization of a society and the quality of institutions. This argument suggests the key importance of the capacity of societies to adjust policies to shocks. A deeper understanding of the interaction between history, geography, polarization and institutions remains a challenge awaiting future research.

The exciting developments associated with the emergence of a growing class of (formerly) poor countries that took off implies that the rewards for adopting the proper growth incentives are high. A remaining challenge is how to facilitate the widening of the emerging market club, and how to minimize the prospects of new conflicts associated with the emergence of new economic powers like China and India.

See Also

- development economics
- growth and institutions
- growth and international trade
- Solow, Robert

Bibliography

Acemoglu, D., Johnson, S., Robinson, J. and Thaicharoen, Y. 2003. Institutional causes, macroeconomic symptoms: volatility, crises and growth. *Journal of Monetary Economics* 50, 49–123.

Aizenman, J. and Marion, N. 1999. Volatility and investment: interpreting evidence from developing countries. *Economica* 66, 157–79.

Aizenman, J., Pinto, B. and Radziwill, A. 2004. Sources for financing domestic capital – is foreign saving a viable option for developing countries? Working Paper No. 1007. Department of Economics, University of California, Santa Cruz.

Ben-David, D. 1993. Equalizing exchange: trade liberalization and income convergence. *Quarterly Journal of Economics* 108, 653–79.

Caroll, C. and Weil, D. 1994. Saving and growth: a reinterpretation. *Carnegie Rochester Conference Series* 40, 133–92.

De Soto, H. 1989. *The Other Path*. New York: Harper and Row.

Dooley, M. 1988. Capital flight: a response to differences in financial risks. *IMF Staff Papers* 35, 422–

36.

Easterly, W. 1999. The ghost of financing gap: testing the growth model used in the international financial institutions. *Journal of Development Economics* 60, 423–38.

Edwards, S. 1998. Openness, productivity and growth: what do we really know? *Economic Journal* 108, 383–98.

Fischer, S. 2003. Globalization and its challenges. *American Economic Review* 93, 1–30.

Frankel, J. and Romer, D. 1999. Does trade cause growth? *American Economic Review* 89, 379–99.

Gertler, M. and Rogoff, K. 1990. North-South lending and endogenous domestic capital market inefficiencies. *Journal of Monetary Economics* 26, 245–66.

Parente, S. and Prescott, E. 2005. A unified theory of the evolution of international income levels. In *Handbook of Economic Growth*, vol. 1, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Ramey, G. and Ramey, V. 1995. Cross-country evidence on the link between volatility and growth. *American Economic Review* 85, 1138–51.

Rodríguez, F. and Rodrik, D. 2001. Trade policy and economic growth: a skeptic's guide to the cross-national evidence, *Macroeconomics Annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press for NBER.

Rodrik, D. 1998. Who needs capital-account convertibility? In *Should the IMF Pursue Capital Account Convertibility? Essays in International Finance*, No. 207, ed. P. Kenen. Princeton, NJ: Princeton University Press.

Rodrik, D. 1999. Where did all the growth go? External shocks, social conflict, and growth collapses. *Journal of Economic Growth* 4, 358–412.

Stiglitz, J. 2002. *Globalization and Its Discontents*. New York: W. W. Norton.

Stulz, R. 2005. The limits of financial globalization. *Journal of Finance* 60, 1595–638.

World Bank. 1992. *The East Asian Miracle*. Washington, DC: World Bank.

Young, A. 1995. The tyranny of numbers: confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 110, 641–80.

How to cite this article

Aizenman, Joshua. "emerging markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000065> doi:10.1057/9780230226203.0468

The New Palgrave Dictionary of Economics Online

empirical likelihood

Yuichi Kitamura

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Empirical likelihood (EL) is a method for estimation and inference without making distributional assumptions. Viewed as a nonparametric maximum likelihood estimation procedure (NPMLE), it approximates the unknown distribution function with a discrete distribution, then applies the ML estimation method. Alternatively, EL can be regarded as a minimum divergence estimation procedure. EL works well for estimating moment condition models, though it applies to other models as well. The large deviation principle (LDP) and other techniques show that EL has many optimality properties.

Keywords

blockwise empirical likelihood; empirical likelihood; empirical likelihood ratio; generalized empirical likelihood; generalized method of moments; kernel regression technique; Lagrange multiplier; large deviation principle; maximum likelihood; nonparametric maximum likelihood estimation; semiparametric estimation; vector autoregressions

Article

1 Introduction

Empirical likelihood (EL) is a method for estimation and inference without making distributional assumptions. The main feature of EL is the use of a discrete distribution to approximate the unknown distribution function nonparametrically, where the approximating discrete distribution is typically supported by empirical observations. Owen (1988) and subsequent papers considered applications of this approach to moment condition models. Their important discovery is that EL, which can be interpreted as a nonparametric maximum likelihood estimation (NPMLE) method, possesses many desirable asymptotic properties that are analogous to those of parametric likelihood procedures. To describe more details of empirical likelihood, consider i.i.d. data $\{z_i\}_{i=1}^n$, where each z_i is distributed according to an

unknown probability distribution F_0. Suppose the expectation of an $\mathbb{R}^q$-valued function $g(z, \theta_0)$, which is known up to the finite-dimensional parameter θ_0 in $\Theta \subset \mathbb{R}^k$, is restricted to be zero:

$$E[g(z, \theta_0)] = \int g(z, \theta_0) dF_0(z) = 0. \tag{1.1}$$

Let Δ denote the simplex $\{(p_1, \ldots, p_n) : \Sigma_{i=1}^n p_i = 1, \quad 0 \le p_i, i = 1, \ldots, n\}$. Each vector $(p_1, \ldots, p_n) \in \Delta$ 'parametrizes' the unknown distribution F_0 by $\hat{F}_n(z) = \Sigma_{i=1}^n p_i 1\{z_i \le z\}$, $z \in \mathbb{R}$ (1{·} signifies the usual indicator function). This is the approximating discrete distribution mentioned above. The nonparametric log-likelihood function to be maximized is

$$\ell_{\mathrm{NP}} = \sum_{i=1}^n \log p_i, \sum_{i=1}^n g(z_i, \theta) p_i = 0, (p_1, \ldots, p_n) \in \Delta, \theta \in \Theta.$$

Let $(\hat{\theta}_{\mathrm{EL}}, \hat{p}_{\mathrm{EL}1}, \ldots, \hat{p}_{\mathrm{EL}n})$ denote the value of $(\theta, p_1, \ldots, p_n) \in \Theta \times \Delta$ that maximizes $\bullet_{\mathrm{NP}}$. This is called the (maximum) empirical likelihood estimator. The NPMLE for θ and F are $\hat{\theta}_{\mathrm{EL}}$ and $\hat{F}_{\mathrm{EL}} = \Sigma_{i=1}^n \hat{p}_{\mathrm{EL}i} 1\{z_i \le z\}$. One might expect that the high dimensionality of the parameter space $\Theta \times \Delta$ makes the above maximization problem intractable for any practical application. Fortunately, that is not the case, if one uses the following nested procedure. First, fix θ at a value in Θ and consider the log-likelihood with the parameters $(p_1,\ldots,p_n)$ 'profiled out':

$$\ell(\theta) = \max \ell_{\mathrm{NP}}(p_1, \ldots, p_n) \text{ subject to} \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g(z_i, \theta) = 0. \tag{1.2}$$

A straightforward application of the Lagrange multiplier method shows that $\bullet(\theta)$ is represented by

$$\ell(\theta) = \min_{\gamma \in \mathbb{R}^q} - \sum_{i=1}^n \log(1 + \gamma' g(z_i, \theta))) - n \log n$$

(1.3)

(see, for example, Kitamura, 2006). The numerical evaluation of the function •(·) is easy, because (1.3) is a low-dimensional convex maximization problem, for which a simple Newton algorithm works. Second, obtain the empirical likelihood estimator $\hat{\theta}_{EL}$ as the maximizer of (1.3). The maximization of •(θ) with respect to θ is typically carried our using a nonlinear optimization algorithm.

Basic properties of the empirical likelihood procedure are now well-understood. The EL estimator $\hat{\theta}_{EL}$ is $n^{1/2}$-consistent and asymptotically normal. Let D and S denote $E[\nabla_\theta g(z, \theta_0)]$ and $E[g(z, \theta_0)g(z, \theta_0)']$, then its asymptotic distribution is given by $N(0,(D'SD)^{-1})$. Also, suppose R is a known $\mathbb{R}^s$-valued function of θ, and the econometrician poses a hypothesis that θ_0 is restricted as $R(\theta_0)=0$, where the s restrictions are independent. This can be tested by forming a nonparametric analogue of the parametric likelihood ratio statistic. Let $r = -2(\sup_{\theta:R(\theta)=0}\ell(\theta) - \sup_{\theta\in\Theta}\ell)$, then this obeys the chi-square distribution with s degrees of freedom asymptotically under the null. The factor r is called the empirical likelihood ratio (ELR) statistic. ELR also applies to testing overidentifying restrictions: see Section 2. These properties and other basics of EL and related methods have been studied extensively in the literature (see Qin and Lawless, 1994; Imbens, 1997; Kitamura, 1997; Kitamura and Stutzer, 1997; Smith, 1997; Imbens, Spady and Johnson, 1998; Newey and Smith, 2004).

An alternative way to motivate EL is to use a minimum divergence estimation framework. Let f and g denote the density functions or the probability functions of distribution functions F and G. Define a 'divergence measure' between F and G to be

$$D(F, G) = \int \varphi\left(\frac{f(z)}{g(z)}\right) g(z)\mathrm{d}z, \tag{1.4}$$

for a convex function ϕ. It is easy to see that $D(\cdot,G)$ is minimized at G. Let

$$\mathcal{F}(\theta) = \left\{F: \int g(z, \theta)\mathrm{d}F = 0,\ F \text{ is a CDF}\right\}.$$

Then $\mathcal{F} = \cup_{\theta\in\Theta}\mathcal{F}(\theta)$ is the set of all probability distributions that are compatible with the moment restriction (1.1). Now consider the problem of minimizing the divergence $D(F,F_0)$ with respect to $F\in\mathcal{F}$. In other words, a distribution that is 'closest' to the true distribution F_0 in the class of distributions $\mathcal{F}$ is

sought. Pick a value $\theta \in \Theta$ and define

$$v(\theta) = \inf_F D(F, F_0) \text{subject to} \int g(z, \theta) f \, dz = 0, \int f \, dz = 1.$$
(P)

The value $v(\theta)$ is regarded as the minimum divergence between F_0 and the set of distributions that satisfy the moment restriction with respect to $g(z, \theta)$. The non-negativity of f is maintained if ϕ is modified so that $\varphi(z) = \infty$ for $z<0$ (see Borwein and Lewis, 1991). The primal problem (P) has a dual problem

$$v^*(\theta) = \max_{\lambda \in \mathbb{R}, \gamma \in \mathbb{R}^q} \left[\lambda - \int \varphi^*(\lambda + \gamma' g(z, \theta)) dF_0(z) \right],$$
(DP)

where φ^* is the convex conjugate (or the Legendre transformation) of ϕ, that is $\varphi^*(y) = \sup_x [xy - \varphi(x)]$. (DP) is a finite-dimensional unconstrained convex maximization problem. The Fenchel duality theorem implies that $v(\theta) = v^*(\theta)$. Since the true value θ_0 minimizes $v(\theta)$ over Θ, it follows that

$$\theta_0 = \operatorname{argmin}_{\theta \in \Theta} v^*(\theta).$$
(1.5)

Note that the integral in the definition of v^* is the expected value of $\varphi^*(\lambda + \gamma' g(z, \theta))$ with respect the true distribution F_0, which is unknown in practice. A feasible procedure is obtained by replacing the expectation with the sample average, that is

$$\hat{v}^*(\theta) = \max_{\lambda \in \mathbb{R}, \gamma \in \mathbb{R}^q} \left[\lambda - \frac{1}{n} \sum_{i=1}^{n} \varphi^*(\lambda + \gamma' g(z_i, \theta)) \right].$$
(1.6)

Corresponding to (1.5), an appropriate minimum distance estimator takes the form

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \hat{V}^{*}(\theta).$$

This minimum divergence framework yields empirical likelihood as a special case with $\varphi(x) = -\log(x)$ (or equivalently, $\varphi^{*}(x) = -1 - \log(-y)$). Other choices for ϕ are, of course, possible. For example, $\varphi(x) = x\log(x)$ yields the 'exponential tilt' estimator (Kitamura and Stutzer, 1997), while $\varphi(z) = \frac{1}{2}(x^2 - 1)$ corresponds to the continuous updating GMM estimator (CUE) (Hansen, Heaton and Yaron, 1996). A convenient parametric family of convex functions known as the Cressie–Read family (Read and Cressie, 1988) subsumes these three important cases. If ϕ belongs to the Cressie–Read family, one can show that the minimum divergence estimator can be written as

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \max_{\gamma \in \mathbb{R}^q} \left[\frac{1}{n} \sum_{i=1}^{n} \kappa(\gamma' g(z_i, \theta)) \right] \tag{1.7}$$

where $\kappa(y) = -\varphi^{*}(y + 1)$. This is essentially equivalent to the generalized empirical likelihood (GEL) estimator by Smith (1997). Smith (2004) provides a detailed account for GEL.

2 EL and the large deviation principle

Like the conventional asymptotic method, the large deviation principle (LDP) offers first order approximations for various estimators and tests. Unlike the conventional theory, which produces local linear approximations, the LDP provides global nonlinear approximations. It is the latter feature that enables the LDP to yield results not obtained by the conventional linear approximations. For example, the LDP shows that EL enjoys many optimality properties that are not shared by, for example, the conventional GMM estimator.

To introduce the concept of the LDP in the context of moment condition models, suppose the econometrician observes i.i.d. data ($z_1, \ldots, z_n$), where z_i satisfies the restriction (1.1). Let A_n be an event as a result of estimation or testing: for example, if one uses an estimator θ_n to estimate θ_0, one may consider $A_n = 1\{\|\theta_n - \theta_0\| > c\}$ for a constant c. Then $\Pr\{A_n\}$ is the probability of the estimator missing the true value by a margin larger than c. Or, in testing a null hypothesis H_0, A_n can represent the

event that H_0 is accepted. If the null is incorrect, $\Pr\{A_n\}$ is the probability of type II errors. In either way, $\lim_{n\to\infty}\Pr\{A_n\} = 0$ if the estimator or the test is consistent. The LDP also deals with asymptotic properties, but it is concerned with the limit of the form $\lim_{n\to\infty}\frac{1}{n}\log\Pr\{A_n\}$. (If the limit does not exist, one needs to consider lim inf or lim sup, depending on the purpose of analysis.) Let $-d\leq 0$ denote the above limit so that $\Pr\{A_n\}\approx e^{-nd}$, which characterizes how fast $\Pr\{A_n\}$ decays. The goal is to obtain a procedure that maximizes the speed of decay d.

Kitamura and Otsu (2005) study the estimation of models of the form (1.1) using the LDP. One complication in the application of the LDP to an estimation problem in general is that an estimator that maximizes the limiting decay rate d with $A_n = 1\{\|\theta_n - \theta_0\| > c\}$ uniformly in unknown parameters does not exist in general, unless the model belongs to the exponential family. A possible way around this issue is to pursue minimax optimality, rather uniform optimality. See Puhalskii and Spokoiny (1998) for a general discussion on such a minimax framework. Note that the probability of the event $A_n = 1\{\|\hat{\theta} - \theta_0\| > c\}$ depends on θ_0 and F_0, therefore the worst case scenario is given by the pair (allowed in the model (1.1)) that maximizes $\Pr\{A_n\}$. Suppose an estimator θ_n minimizes this worst-case probability, thereby achieving minimaxity. The limit inferior of the minimax probability provides an asymptotic minimax criterion. Kitamura and Otsu (2005) show that an estimator that attains the lower bound of the asymptotic minimax criterion can be obtained from the EL objective function $\cdot(\theta)$ in (1.2) as follows:

$$\hat{\theta}_{ld} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, Q_n(\theta),\ Q_n(\theta) = \sup_{\theta^*\in\Theta:\|\theta^*-\theta\|>c} \mathcal{E}(\theta^*).$$

Calculating $\hat{\theta}_{ld}$ in practice is straightforward. If the dimension of θ is high, it is also possible to focus on a low-dimensional sub-vector of θ and obtain a large deviation minimax estimator for it, treating the rest as nuisance parameters.

Kitamura (2001) shows that empirical likelihood dominates other methods in terms of the LDP when applied to overidentifying restrictions testing. Researchers routinely test overidentifying restrictions of the form

$$\int g(z,\theta)\,dF = 0 \text{ for some } \theta\in\Theta \text{ and for some distribution function } F, \tag{O}$$

with $\dim(\Theta) = k$ and $g \in \mathbb{R}^q$, $q>k$. The log empirical likelihood under the restriction (O) is $\sup_{\theta \in \Theta} \ell(\theta)$; without the restriction, it is $-n \cdot \log \cdot n$. The ELR test statistic for (O) is the difference of the two multiplied by –2. It is asymptotically distributed according to the χ^2 distribution with $q-k$ degrees of freedom under (O) (Qin and Lawless, 1994). Using the notation in the previous section, rewrite the above null in an equivalent form: $(O)' : F_0 \in \mathcal{F}$. It turns out that ELR for $(O)'$ has a property of being uniformly most powerful in an LDP criterion. To state this optimality property of ELR formally, let $\mathcal{F}$ denote the set of all probability distribution functions. Practically all reasonable tests for (O) (or $(O)'$) can be represented by a partition $\Omega = (\Omega_1, \Omega_2)$ of $\mathcal{F}$, such that if the empirical distribution function F_n falls into Ω_1 (Ω_2) one rejects (accepts) (O). It is a straightforward exercise to show that the ELR test rejects the null if the Kullback–Leibler divergence $K(F_n,G)$ between F_n and G, minimized over $G \in \mathcal{F}$, is too large. Therefore ELR is represented by the following partition of $\mathcal{F}$: $\Lambda = (\Lambda_1, \Lambda_2)$, $\Lambda_1 = \{F : \inf_{G \in \mathcal{F}} K(F, G) < \eta\}$, $\Lambda_2 = \Lambda_1^c$ for a positive number η. Following Owen (2001), for an event A_n that involves observations $z_1,\ldots,z_n$ that are randomly sampled from F, let $\Pr\{A_n;F\}$ denote the probability of the event. By applying a mathematical result called Sanov's theorem, it can be shown that

$$\sup_{F^* \in \mathcal{F}} \limsup_{n \to \infty} \frac{1}{n} \log \Pr\left\{F_n \in \Lambda_2;\ F^*\right\} \le -\eta.$$

Kitamura (2001) also shows that if the following inequality holds for a test $\Omega = (\Omega_1, \Omega_2)$ that satisfies some regularity conditions (see Kitamura, 2001, for the regularity conditions):

$$\sup_{F^* \in \mathcal{F}} \limsup_{n \to \infty} \frac{1}{n} \log \Pr\left\{F_n \in \Omega_2;\ F^*\right\} \le -\eta,$$

then it must be that

$$\limsup_{n \to \infty} \frac{1}{n} \log \Pr\left\{F_n \in \Omega_1;\ F^{**}\right\} \ge \limsup_{n \to \infty} \frac{1}{n} \log \Pr\left\{F_n \in \Lambda_1;\ F^{**}\right\}$$

for every $F^{**} \notin \mathcal{F}$. The first two of the above three inequalities mean that the ELR test Λ and the

arbitrary regular test Ω are comparable in terms of its LDP property of type I error probabilities. But the third inequality implies that the ELR test is no less powerful than the arbitrary test if the LDP of type II error probabilities are used to measure the asymptotic powers of the tests. Note that the third inequality holds for every $F^{**} \notin \mathcal{F}$: that is, it holds uniformly over alternatives. Since the test (Ω_1, Ω_2) is arbitrary, this shows that ELR is uniformly most powerful in an LDP sense. Such a property is sometimes referred to as the Generalized Neyman–Pearson (GNP) optimality.

3 Higher-order asymptotics

An alternative way to see why EL works well is to analyse it using higher-order asymptotics. Newey and Smith (2004) investigate higher-order properties of the GEL family of estimators. To illustrate their findings, it is instructive to look at the first-order condition that the EL estimator satisfies, that is $\nabla_\theta \ell(\hat{\theta}_{\mathrm{EL}}) = 0$. A straightforward calculation shows that this condition, using the notation $\hat{D}(\theta) = \sum_{i=1}^{n} \hat{p}_{\mathrm{EL},i} \nabla_\theta g(z_i, \theta)$ and $\hat{S}(\theta) = \sum_{i=1}^{n} \hat{p}_{\mathrm{EL},i} g(z_i, \theta) g(z_i, \theta)'$, can be written as

$$\hat{D}(\hat{\theta}_{\mathrm{EL}})' \hat{S}^{-1}(\hat{\theta}_{\mathrm{EL}}) \bar{g}(\hat{\theta}_{\mathrm{EL}}) = 0; \tag{3.1}$$

see Theorem 2.3 of Newey and Smith (2004). The factor $\hat{D}(\hat{\theta}_{\mathrm{EL}})' \hat{S}^{-1}(\hat{\theta}_{\mathrm{EL}})$ can be interpreted as a feasible version of the optimal weight for the sample moment $\bar{g}(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(z_i, \theta)$. Equation (3.1) is similar to the first-order condition for GMM, though there are important differences. Notice that the Jacobian term *D* and the variance term *S* are estimated by $\hat{D}(\hat{\theta}_{\mathrm{EL}})$ and $\hat{S}(\hat{\theta}_{\mathrm{EL}})$ in (3.1). It can be shown that these are semiparametrically efficient estimators of *D* and *S* under the moment restriction (1.1). This means that they are asymptotically uncorrelated with $\bar{g}(\theta_0)$, removing the important source of the second-order bias of GMM. Moreover, the EL estimator does not involve a preliminary estimator, thereby eliminating another source of the second-order bias in GMM. Newey and Smith (2004) formalize this intuition and obtain an important conclusion that the second-order bias of the EL estimator is equal to that of the infeasible method-of-moments estimator that optimally weights $\bar{g}$ by the unknown factor $D' S^{-1}$. In contrast, the first-order condition of GMM takes a similar form, but the terms that correspond to *D* and *S* are inefficiently estimated, causing bias. Newey and Smith (2004) note that the first-order conditions of GEL estimators have a form where *D* is efficiently estimated but *S* is not, leaving a source of bias that is not present for EL.

Higher-order properties of ELR tests have been studied in the literature as well. One of the significant findings in the early literature of empirical likelihood is the Bartlett correctability of the empirical likelihood ratio test, discovered by DiCiccio, Hall and Romano (1991). Consider the ELR test statistic

for H_0: $\theta = \theta_0$ in the model (1.1) with $q=k$. DiCiccio, Hall and Romano (1991) show that the accuracy of the χ^2 asymptotic approximation for the distribution of the ELR statistic can be improved from the rate n^{-1} to the much faster rate n^{-2} by multiplying it by a factor called the Bartlett coefficient.

4 Some variations of EL

EL is applicable to many problems other than (1.1), but they sometimes require extending and modifying the standard EL method described so far. For example, suppose economic theory implies that the conditional mean of $g(z, \theta_0)$ given a vector of covariates x is zero:

$$E[g(z, \theta_0)|x] = 0. \tag{4.1}$$

This restriction is stronger than (1.1). Though one can choose an arbitrary function $a(x)$ of x as an instrument, this can be problematic since (*a*) choosing an instrument that delivers strong identification may be a difficult task, and (*b*) an arbitrary instrument does not achieve efficiency in general. Kitamura, Tripathi and Ahn (2004) use the kernel regression technique to incorporate the information in the conditional moment restriction into empirical likelihood. Their estimator achieves the semiparametric efficiency bound of the model (4.1) under weak regularity conditions. While there exist estimators that achieve efficiency in the model, the EL-based estimator has an advantage that finding a preliminary estimator that is consistent is not necessary. A simulation study in Kitamura, Tripathi and Ahn (2004) indicates that the conditional EL estimator and tests based on it work remarkably well in finite samples. Donald, Imbens and Newey (2003) propose an alternative estimator for (4.1). Their idea is to use a sequence functions of x as a vector of instruments, then apply EL to the resulting unconditional moment restriction model. By letting the dimension of the instrument vector grow with the sample size in such a way that it spans the ‘optimal instrument’ asymptotically, their procedure also achieves the semiparametric efficiency bound.

A topic that is closely related to the above is nonparametric specification testing. Suppose, for example, one is interested in testing the specification of a parametric regression model $E[y|x] = m(x, \theta_0)$, where m is parametrized by a vector $\theta_0 \in \Theta$. The null hypothesis of correct specification can be written in terms of a conditional moment restriction for the function $g(z, \theta) = y - m(x, \theta)$, $z = (x', y)'$:

$$E[g(z, \theta)|x] = 0 \text{ for some } \theta \in \Theta. \tag{C}$$

Tripathi and Kitamura (2003) shows that a conditional version of the ELR test applies to the above problem. They propose a simple procedure: reject (C) if the maximized value of the conditional empirical likelihood function, which is essentially the one used in Kitamura, Tripathi and Ahn (2004), is too small. They also calculate the asymptotic power of their test. Their analysis shows that the EL-based testing procedure has an asymptotic optimality property in terms of an average power criterion.
Another example in which EL needs an appropriate modification is a time series model. Suppose the researcher observes a strictly stationary and weakly dependent time series $\{z_1,\ldots,z_t\}$, and each z_t satisfies the moment condition $E[g(z_t, \theta_0)] = 0, \theta_0 \in \Theta$. Applying EL to this model ignoring dependence is inappropriate; it leads to efficiency loss, and the chi-square asymptotics of the ELR test break down. There are at least three alternative ways to deal with the problem caused by dependence. The first approach is to parametrize the dynamics using a reduced form time series model such as a vector autoregression (VAR) model (Kitamura, 2006). While straightforward, this approach involves the risk of mis-specifying the dynamics, and reduces the appeal of EL as nonparametric likelihood. The second approach is the blocking method proposed by Kitamura and Stutzer (1997) and Kitamura (1997). The idea is to form data blocks by taking consecutive observations, and apply EL to them. This is termed blockwise empirical likelihood (BEL). BEL preserves the dependence information in the data, in a fully nonparametric manner. The third approach is a hybrid of the first and the second approaches (Kitamura, 2006). That is, one applies a low order parametric filter to lessen the degree of dependence in the data, then apply BEL to the filtered data. While this does not change the desirable asymptotic property of BEL, it appears to have advantages in finite samples when applied to a time series that is highly persistent.

See Also

- generalized method of moments estimation
- semiparametric estimation
- vector autoregressions

Bibliography

Borwein, J.M. and Lewis, A.S. 1991. Duality relationships for entropy-type minimization problems. *SIAM Journal of Control and Optimization* 29, 325–38.

DiCiccio, T., Hall, P. and Romano, J. 1991. Empirical likelihood is Bartlett-correctable. *Annals of Statistics* 19, 1053–61.

Donald, S.G., Imbens, G.W. and Newey, W.K. 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics* 117, 55–93.

Hansen, L.P., Heaton, J. and Yaron, A. 1996. Finite-sample properties of some alternative GMM

estimators. *Journal of Business and Economic Statistics* 14, 262–80.

Imbens, G.W. 1997. One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies* 64, 359–83.

Imbens, G.W., Spady, R.H. and Johnson, P. 1998. Information theoretic approaches to inference in moment condition models. *Econometrica* 66, 333–57.

Kitamura, Y. 1997. Empirical likelihood methods with weakly dependent processes. *Annals of Statistics* 25, 2084–102.

Kitamura, Y. 2001. Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* 69, 1661–72.

Kitamura, Y. 2006. Empirical likelihood methods in econometrics: theory and practice. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. R. Blundell, W.K. Newey and T. Persson. Cambridge: Cambridge University Press.

Kitamura, Y. and Otsu, T. 2005. Minimax estimation and testing for moment condition models via large deviations. Manuscript, Department of Economics, Yale University.

Kitamura, Y. and Stutzer, M. 1997. An information theoretic alternative to generalized method of moments estimation. *Econometrica* 65, 861–74.

Kitamura, Y., Tripathi, G. and Ahn, H. 2004. Empirical likelihood based inference in conditional moment restriction models. *Econometrica* 72, 1667–714.

Newey, W.K. and Smith, R.J. 2004. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72, 219–55.

Owen, A. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–49.

Owen, A. 2001. *Empirical Likelihood*. New York: Chapman and Hall/CRC.

Puhalskii, A. and Spokoiny, V. 1998. On large-deviation efficiency in statistical inference. *Bernoulli* 4, 203–72.

Qin, J. and Lawless, J. 1994. Empirical likelihood and general estimating equations. *Annals of Statistics* 22, 300–25.

Read, T.R.C. and Cressie, N.A.C. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Berlin: Springer.

Smith, R.J. 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal* 107, 503–19.

Smith, R.J. 2004. GEL criteria for moment condition models. Working paper, University of Warwick.

Tripathi, G. and Kitamura, Y. 2003. Testing conditional moment restrictions. *Annals of Statistics* 31, 2059–95.

How to cite this article

Kitamura, Yuichi. "empirical likelihood." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000249> doi:10.1057/9780230226203.0469

The New Palgrave Dictionary of Economics Online

encompassing

Grayham E. Mizon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The concept of encompassing is defined and the role that it and congruence have in econometric modelling is discussed. Empirically, more than one model can appear to be congruent, but that which encompasses its rivals is dominant and will encompass all models nested within it and accurately predict the mis-specifications of non-congruent models. These results are consistent with a general-to-specific modelling strategy being successful in practice. Alternative forms and applications of encompassing tests are discussed.

Keywords

congruence; encompassing; Gaussian linear regression models; indirect inference; models; non-nested hypotheses; simulation; testing

Article

Introduction and motivation

Imaginative and productive disciplines like economics generate many new theories, partly to extend the range of phenomena that they embrace but also to improve on existing theories. New theories require rigorous evaluation to establish their worth if they are to be relevant, reliable, and robust. In addition to checking their logical consistency and relevance it is important to assess their coherence with observation. The latter usually involves the development of a model that embodies the essential characteristics of the theory and has observable implications.
The analysis presented here concentrates on the evaluation of empirical models. Numerous criteria have been proposed for assessing the coherence of an empirical model with observation. Measures of goodness of fit and selection criteria based on likelihood functions (usually degrees of freedom adjusted) are common (Schwarz, 1978), and are often used both to assess coherence with observation and to select

the preferred model. Probably the most comprehensive and demanding criterion for data coherence is that of congruence (Hendry, 1995; Bontemps and Mizon, 2003), which requires a model to be a valid reduction of whatever process actually generates the observed data – the data generation process (DGP). When $\mathbf{x}_t$ contains the full set of variables involved in an investigation, let the DGP be denoted by the joint density $D_X(\mathbf{x}_t|\mathbf{X}_{t-1}, \phi)$ for $\mathbf{x}_t$ conditional on its history $\mathbf{X}_{t-1}$ with parameters ϕ. Knowledge of the DGP endows one with omniscience and in particular the ability to derive the properties of all models involving the same variables such as $f_X(\mathbf{x}_t|\mathbf{X}_{t-1}, \chi)$, but, alas, for practical purposes it is unattainable. In empirical modelling, therefore, congruence means that, given the available information, the model is indistinguishable from the DGP for the chosen variables, that is, no evidence has been evinced that the model is not the DGP. Testing the latter requires that extensive, not limited, searching is done for evidence of non-congruence. This leads to the adoption of statistical tests of model mis-specification (for example, wrong functional form, heteroskedastic or serially correlated residuals) as indirect but practical tests of congruence (Hendry, 1995; Mizon, 1995). Since in practice a congruent model will not be the DGP, it will not necessarily be able to explain the properties of other models, and in particular those that constitute the current best knowledge and practice. Thus, a valuable part of the evaluation of a model is an assessment of whether it represents an advance on existing knowledge. ‘The encompassing principle is concerned with the ability of a model to account for the behaviour of others, or less ambitiously, to explain the behaviour of relevant characteristics of other models’ (Mizon, 1984, p. 136). A well-known illustration in physics, discussed by Okasha (2002), for example, is provided by Newton's laws of motion and gravitation that encompassed Kepler's laws of motion and gravitation as well as Galileo's law of free-fall, and as a result the same laws explained the motion of bodies in both the terrestrial and the celestial domains. This added credence to Newton's laws, as it does for all models that encompass their rivals. It was widely believed for a long time that Newton's theory revealed the workings of nature and had the ability to explain everything in principle. However, Newton's laws have been superseded or encompassed by Einstein's relativity theory and quantum mechanics. This illustrates the fact that modelling, like discovery, is not a once-for-all event, but a continuous process of development. Progress in science, however, is achieved in many ways, with confidence and persistence playing a role in some instances as a consequence of rejection not being accepted as final or corroboration of models that are subsequently superseded not being taken as definitive.

Background

The idea underlying the encompassing principle has a long pedigree; for example, the comparison of competing theories has been long recognized as a basic ingredient of a scientific research strategy (Nagel, 1961). The implementation via a statistical contrast equally has a long history; Cox (1961; 1962) are the most significant early examples. These papers introduced statistical tests for separate families of hypotheses, and discussed several examples to illustrate their practical relevance. The tests were later developed in the literature on non-nested hypothesis testing (Pesaran, 1974; Davidson and MacKinnon, 1981), and encompassing (Mizon, 1984). The latter paper contains a general presentation of the concept of encompassing and discussion of numerous applications, and Mizon and Richard (1986) provides a theoretical framework for encompassing, on which other theoretical papers have built extensions.

Davidson et al. (1978) is one of the first attempts to develop a framework for a scientific comparison of alternative economic theories and econometric models implementing them. Different econometric models for the series of UK consumption, which rely on different economic hypotheses about consumption behaviour, were embedded in a general model and shown to imply different testable restrictions on its coefficients.

Distinguished natural scientists have expressed surprise that social scientists are able to learn anything from empirical observation when they rarely have experimental evidence. However, the encompassing principle provides precisely the analogue of the physical experiment. Experiments enable physicists and chemists to sift through alternative theories by evaluating the veracity of their implications or predictions in controlled conditions, and thus to eliminate those theories whose predictions perform badly. Congruence is the analogue of setting up controlled experimental conditions. The need to distinguish between alternative theories that each appear to be coherent with outcomes, experimental or non-experimental, leads to the search for dominant theories. For disciplines that are largely non-experimental, having a principle such as encompassing is essential for discriminating between alternative models. Typically, alternative empirical models use different information sets and possibly different functional forms, and are thus separate or non-nested. This non-nested feature enables more than one model to be congruent with respect to sample information – each can be congruent with respect to its own information set – and so it is important to assess their relative merits. Using the encompassing principle, Ericsson and Hendry (1999) analyse this issue and show that the corroboration of more than one model can imply the inadequacy of each, and Mizon (1989) provides an illustration by comparing a Keynesian and a monetarist model of inflation. Hence, congruence and encompassing are inextricably linked; in particular, encompassing comparisons of non-congruent models can be misleading. For example, general models will not always encompass simplifications of themselves even though that might seem to be an obvious characteristic of a general model, but a congruent general model will always encompass simpler models (Hendry, 1995; Gouriéroux and Monfort, 1995; Bontemps and Mizon, 2003).

Principle

Underlying all empirical econometric analyses is an information set (collection of variables or their sigma field), and a corresponding probability space. This information set has to be sufficiently general to include all the variables thought to be relevant to the empirical implementation of theoretical models in the form of statistical models. It is also important that this information set include the variables needed for all competing models that are to be compared. When these variables are $\mathbf{x}_t$ the DGP for the observed sample is the joint density $D_X(\mathbf{x}_t|\mathbf{X}_{t-1}, f)$ at the particular parameter value $\boldsymbol{\varphi} = \boldsymbol{\varphi}_0$. Let a parametric statistical model of the joint distribution be $M_f = \{f_X(\mathbf{x}_t|\mathbf{X}_{t-1}, x) x \in \Xi \subset \mathrm{R}^k\}$. Let $\hat{x}$ be the maximum likelihood estimator of ξ so that $\hat{x} \overset{p}{\underset{M_f}{\square}} x$ and $\hat{x} \overset{p}{\underset{DGP}{\square}} x(f_0) = x_0$ which is the pseudo-true value of $\hat{x}$. Note that the parameters of a model are not arbitrary in that M_f and its parameterization ξ are chosen to correspond to phenomena of interest such as elasticities and partial responses within the chosen

probability space. For the two alternative models $M_1 = \{f_1(\mathbf{x}_t|\mathbf{X}_{t-1}, q_1), q_1 \in \Theta_1 \subset R^{p_1}\}$ and $M_2 = \{f_2(\mathbf{x}_t|\mathbf{X}_{t-1}, q_2), q_2 \in \Theta_2 \subset R^{p_2}\}$ the concept of parametric encompassing, in accordance with the approach in Mizon (1984), Mizon and Richard (1986), and Hendry and Richard (1989), can be defined as follows. M_1 encompasses M_2 (denoted $M_1 \xi M_2$) if and only if $q_{20} = \mathbf{h}_{21}(q_{10})$ when θ_{i0} is the pseudo-true value of the maximum likelihood estimator $\hat{q}_i$ of $q_i, i = 1, 2$, and $\mathbf{h}_{21}(q_{10})$ is the binding function given by $\hat{q}_2 \overset{p}{\underset{M_1}{\square}} \mathbf{h}_{21}(q_{10})$ (Mizon and Richard, 1986; Hendry and Richard, 1989; Gouriéroux and Monfort, 1995). Note that this definition of encompassing applies when M_1 and M_2 are non-nested as well as nested. However, Hendry and Richard (1989) showed that when M_1 and M_2 are non-nested $M_1 \xi M_2$ is equivalent to M_1 being a valid reduction of the minimum completing model $M_c = M_1 \cup M_2^{\perp}$ (so that $M_1, M_2 \subset M_c$) when $M_2^{\perp}$ is the model which represents all aspects of M_2 that are not contained in M_1. When this condition is satisfied, M_1 is said to parsimoniously encompass M_c (denoted $M_1 \xi_p M_c$). Parsimonious encompassing is the property that a model is a valid reduction of a more general model. When a general-to-simple modelling strategy is adopted, the general unrestricted model (GUM) will have been chosen to embed the different econometric models implementing rival economic theories for the phenomenon of interest. Hence searching for the model that parsimoniously encompasses the congruent GUM is an efficient way to find congruent and encompassing models in practice. Hendry and Krolzig (2003) describe and illustrate the performance of a computer program that implements a general-to-specific modelling strategy.

The comparison of Gaussian linear regression models provides a simple and convenient framework to illustrate the main ideas. Consider the two models M_1 and M_2 defined in:

$$\begin{aligned} M_1 &\sim \mathbf{y} = \mathbf{Z}_1 b + \mathbf{u}_1, & \mathbf{u}_1 &\sim N\left(0, s_1^2 \mathbf{I}_n\right) \\ M_2 &\sim \mathbf{y} = \mathbf{Z}_2 g + \mathbf{u}_2, & \mathbf{u}_2 &\sim N\left(0, s_2^2 \mathbf{I}_n\right) \\ M_c &\sim \mathbf{y} = \mathbf{Z}_1 \mathbf{b} + \mathbf{Z}_2 \mathbf{c} + e & e &\sim N\left(0, s_c^2 \mathbf{I}_n\right) \end{aligned} \tag{1}$$

when $\mathbf{y}$ is $n \times 1$, and $\mathbf{Z}_i$ is $n \times k_i (i = 1, 2)$ containing n observations on the independent and two sets of explanatory variables respectively with no variables in common. The explanatory variables are distributed independently of the error vectors $\mathbf{u}$, $\mathbf{v}$, and ε. When M_1, M_2 and M_c are each hypotheses about the distribution of $M_{\mathbf{Z}}$, the models M_1 and M_2 are non-nested in that neither is a special case of the other, whereas both M_1 and M_2 are nested within M_c. A test of the hypothesis that M_1 encompasses M_2 (denoted $M_1 \xi M_2$) is possible using the contrast $\hat{\boldsymbol{\Psi}}_Y = \hat{\boldsymbol{\Upsilon}} - \hat{\boldsymbol{\Upsilon}}_1 = (\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{Q}_1\mathbf{y}$ with

$\mathbf{Q}_1 = (\mathbf{I}_n - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1')$ between the maximum likelihood estimator of γ, $\hat{g} = (\mathbf{Z}_1'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{y}$, and an estimate $\hat{g}_1 = (\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{y}$ of the pseudo-true value of $\hat{g}$ under M_1 given by $g_1 = p\lim_{n\to\infty}(\hat{g})_{M_1}$. The sample complete parametric encompassing test statistic is given by $\eta_c = \hat{\boldsymbol{\psi}}_\gamma'(\mathbf{Z}_2'\mathbf{Z}_2)(\mathbf{Z}_2'\mathbf{Q}_1\mathbf{Z}_2)^{-1}(\mathbf{Z}_2'\mathbf{Z}_2)\hat{\boldsymbol{\psi}}_\gamma / k_2\hat{\sigma}_c^2$ when $\hat{\sigma}_c^2 = \mathbf{y}'(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{y} / (n - k_1 - k_2)$ is the unbiased estimator of σ_c^2 with $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. Under the complete parametric encompassing hypothesis $H_c: \boldsymbol{\psi}_\gamma = \boldsymbol{\gamma} - \boldsymbol{\gamma}_1 = 0$ the statistic η_c is distributed as $F(k_2, n - k_1 - k_2)$. Mizon and Richard (1986) showed that this is precisely the same statistic as that for testing the hypothesis $\mathbf{c} = 0$ in (1), that is, the test statistic for $M_1 \xi M_2$ is exactly the same as that for $M_1 \xi p M_c$ in this case. Variance encompassing is based on the contrast $\hat{\psi}_{\sigma_2^2} = \hat{\sigma}_2^2 - \hat{\sigma}_{21}^2$ between $\hat{\sigma}_2^2$ and an estimator of $\sigma_{21}^2 = \sigma_2^2 + (\sigma_1^2 / n)b_1'(\mathbf{Z}_1'\mathbf{Q}_2\mathbf{Z}_1)b_1$ the pseudo-true value $\hat{\sigma}_2^2$ under M_1 when $\mathbf{Q}_2 = (\mathbf{I}_n - \mathbf{Z}_2(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2')$. Mizon and Richard (1986) showed that the resulting variance encompassing test statistic is asymptotically equivalent to each of the one degree of freedom non-nested test statistics developed by Cox (1961; 1962), Perasan (1974), and Davidson and MacKinnon (1981), among others. The fact that variance dominance is a necessary but not a sufficient condition for variance encompassing highlights a serious limitation of choosing models on the basis of goodness-of-fit selection criteria rather than comparing the alternative models using encompassing test statistics.

Further developments

This analysis illustrates the fact that the choice of statistic for the encompassing contrast is very important, and may depend very much on the purpose of the analysis or the nature of the models being investigated. For example, when the GUM is not easily available or the calculation of pseudo-true values for other encompassing test statistics is difficult, comparison of the forecasting abilities provides an alternative basis for an encompassing test. Although selecting models on the basis of forecast performance can be very misleading for some purposes in a non-stationary environment with regime shifts (Hendry and Mizon, 2005), the concept of forecast encompassing is a valuable method of model comparison. Forecast encompassing statistics were presented by Chong and Hendry (1990), and Ericsson (1993) and Lu and Mizon (1991) extend this analysis in several directions, including multi-step ahead forecasts from nonlinear dynamic models with estimated coefficients. Similarly, when the analytic calculation of pseudo-true values is intractable simulation methods may be used to estimate the pseudo-true values and hence compute the non-nested test statistics (Hendry and Richard, 1989; Pesaran and Pesaran, 1993). Gouriéroux, Monfort and Renault (1993) developed a comprehensive framework for such simulation known as indirect inference, which allows choice of auxiliary functions as the basis for parameter estimation. A consistent estimator of the parameters involved in the encompassing contrast can be obtained when a correction based on the simulated pseudo-true values of the testing statistics is applied. This approach has the potential to extend the application of the encompassing principle

enormously. The relationship between encompassing and conditional moment or m-tests (Newey, 1985) is discussed in White (1994) and Lu and Mizon (1996). The possibility that the encompassing principle be used as a generator of test statistics is discussed in Mizon and Richard (1986). Govaerts, Hendry and Richard (1994) consider the application of encompassing in dynamic models, and Hendry and Mizon (1993) apply it to the comparison of alternative dynamic simultaneous equations models containing integrated and cointegrated variables. A Bayesian approach to encompassing is presented in Florens, Hendry and Richard (1996) and, as a result of using statistical procedures rather than pseudo-true values as in Mizon and Richard (1986), argues that encompassing can be interpreted as a property of model specificity analogous to that of sufficiency for statistics. The encompassing relationship between nonparametric models is considered in Bontemps, Florens and Richard (2006). Finally, Marcellino and Mizon (2006) contains a comprehensive statement and analysis of encompassing as well as many applications of the principle.

See Also

- artificial regressions
- forecasting
- model selection
- models
- testing

Bibliography

Bontemps, C., Florens, J. and Richard, J. 2006. Encompassing in regression models: parametric and non-parametric procedures. In *Progressive Modelling: Non-nested Testing and Encompassing*, ed. M. Marcellino and G. Mizon. Oxford: Oxford University Press.

Bontemps, C. and Mizon, G. 2003. Congruence and encompassing. In *Econometrics and the Philosophy of Economics*, ed. B. Stigum. Princeton: Princeton University Press.

Chong, Y. and Hendry, D. 1990. Econometric evaluation of linear macro-economic models. In *Modelling Economic Series*, ed. C. Granger. Oxford: Clarendon Press.

Cox, D. 1961. Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1, 105–23. Berkeley: University of California Press.

Cox, D. 1962. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society B* 24, 406–24.

Davidson, J., Hendry, D., Srba, F. and Yeo, J. 1978. Econometric modelling of the aggregate time-series

relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88, 661–92.

Davidson, R. and MacKinnon, J. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49, 781–93.

Ericsson, N. 1993. Comment on 'On the limitations of comparing mean squared forecast errors', by M. P. Clements and D. F. Hendry. *Journal of Forecasting* 12, 644–51.

Ericsson, N. and Hendry, D. 1999. Encompassing and rational expectations: how sequential corroboration can imply refutation. *Empirical Economics* 24, 1–21.

Florens, J.-P., Hendry, D. and Richard, J.-F. 1996. Encompassing and specificity. *Econometric Theory* 12, 620–56.

Gouriéroux, C. and Monfort, A. 1995. Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* 11, 195–228.

Gouriéroux, C., Monfort, A. and Renault, E. 1993. Indirect inference. *Journal of Applied Econometrics* 8, 85–118.

Govaerts, B., Hendry, D. and Richard, J.-F. 1994. Encompassing in stationary linear dynamic models. *Journal of Econometrics* 63, 245–70.

Hendry, D. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D. and Krolzig, H.-M. 2003. New developments in automatic general-to-specific modelling. In *Econometrics and the Philosophy of Economics*, ed. B. Stigum. Princeton: Princeton University Press.

Hendry, D. and Mizon, G. 1993. Evaluating dynamic econometric models by encompassing the VAR. In *Models, Methods and Applications of Econometrics*, ed. P. Phillips. Oxford: Basil Blackwell.

Hendry, D. and Mizon, G. 2005. Forecasting in the presence of structural breaks and policy regime shifts. In *Identification and Inference for Econometric Models: Festschrift in Honor of Tom Rothenberg*, ed. D. Andrews and J. Stock. Cambridge: Cambridge University Press.

Hendry, D. and Richard, J.-F. 1989. Recent developments in the theory of encompassing. In *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, ed. B. Cornet and H. Tulkens. Cambridge, MA: MIT Press.

Lu, M. and Mizon, G. 1991. Forecast encompassing and model evaluation. In *Economic Structural*

Change, Analysis and Forecasting, ed. P. Hackl and A. Westlund. Berlin: Springer-Verlag.

Lu, M. and Mizon, G. 1996. The encompassing principle and hypothesis testing. *Econometric Theory* 12, 845–58.

Marcellino, M. and Mizon, G., eds. 2006. *Progressive Modelling: Non-nested Testing and Encompassing*. Oxford: Oxford University Press.

Mizon, G. 1984. The encompassing approach in econometrics. In *Econometrics and Quantitative Economics*, ed. D. Hendry and K. Wallis. Oxford: Blackwell.

Mizon, G. 1989. The role of econometric modelling in economic analysis. *Revista Espanola de Economia* 6, 167–91.

Mizon, G. 1995. Progressive modelling of macroeconomic time series: the LSE methodology. In *Macroeconometrics: Developments, Tensions and Prospects*, ed. K. Hoover. Dordrecht: Kluwer Academic Press.

Mizon, G. and Richard, J.-F. 1986. The encompassing principle and its application to non-nested hypothesis tests. *Econometrica* 54, 657–78.

Nagel, E. 1961. *The Structure of Science*. New York: Harcourt Brace.

Newey, W. 1985. Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047–70.

Okasha, S. 2002. *Philosophy of Science: A Very Short Introduction*. Oxford: Oxford University Press.

Pesaran, M. 1974. On the general problem of model selection. *Review of Economic Studies* 41, 153–71.

Pesaran, M. and Pesaran, B. 1993. A simulation approach to the problem of computing Cox's statistic for testing non-nested models. *Journal of Econometrics* 57, 377–92.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–64.

White, H. 1994. *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.

How to cite this article

Mizon, Grayham E. "encompassing." The New Palgrave Dictionary of Economics. Second Edition. Eds.

Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000243> doi:10.1057/9780230226203.0470

The New Palgrave Dictionary of Economics Online

endogeneity and exogeneity

John Geweke
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Endogeneity and exogeneity are properties of variables in economic or econometric models. The specification of these properties in variables is an essential component of the process of model specification. This article considers their application in the specification of, in turn, deterministic and stochastic models.

Keywords

Cowles Commission; endogeneity and exogeneity; model specification; simultaneous equations models; statistical inference

Article

Endogeneity and exogeneity are properties of variables in economic or econometric models. The specification of these properties for respective variables is an essential component of the entire process of model specification. The words have an ambiguous meaning, for they have been applied in closely related but conceptually distinct ways, particularly in the specification of stochastic models. We consider in turn the case of deterministic and stochastic models, concentrating mainly on the latter.
A deterministic economic model typically specifies restrictions to be satisfied by a vector of variables **y**. These restrictions often incorporate a second vector of variables **x**, and the restrictions themselves may hold only if **x** itself satisfies certain restrictions. The model asserts

$$\forall \mathbf{x} \in R, \quad G(\mathbf{x}, \mathbf{y}) = 0 .$$

The variables **x** are exogenous and the variables **y** are endogenous. The defining distinction between **x** and **y** is that **y** may be (and generally is) restricted by **x**, but not conversely. This distinction is an essential part of the specification of the functioning of the model, as may be seen from the trivial model,

$$\forall \mathbf{x} \in \mathbf{R}^1, \quad x + y = 0 .$$

The condition $x + y = 0$ is symmetric in *x* and *y*; the further stipulation that *x* is exogenous and *y* is endogenous specifies that in the model *x* restricts *y* and not conversely, a property that cannot be derived from $x + y = 0$. In many instances the restrictions on **y** may *determine* **y**, at least for $\mathbf{x} \in R^* \subset R$, but the existence of a *unique* solution has no bearing on the endogeneity and exogeneity of the variables.
The formal distinction between endogeneity and exogeneity in econometric models was emphasized by the Cowles Commission

in its path-breaking work on the estimation of simultaneous economic relationships. The class of models it considered is contained in the specification

$$\mathbf{B}(L)\mathbf{y}(t) + \mathbf{\Gamma}(L)\mathbf{x}(t) = \mathbf{u}(t);\ \mathbf{A}(L)\mathbf{u}(t) = \boldsymbol{\epsilon}(t);\ \mathrm{cov}[\boldsymbol{\epsilon}(t), \mathbf{y}(t-s)] = \mathbf{O},\ \ s > 0;\ \mathrm{cov}[\boldsymbol{\epsilon}(t), \mathbf{x}(t-s)] = \mathbf{O},\ \ \text{all } s;\ \varepsilon(t) \sim \mathbf{IIDN}(\mathbf{O}, \Sigma)\ .$$

The vectors $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are observed, whereas $\mathbf{u}(t)$ and $\varepsilon(t)$ are underlying disturbances not observed but affecting $\mathbf{y}(t)$. The lag operator L is defined by $L\mathbf{x}(t)=\mathbf{x}(t-1)$; the roots of $|\mathbf{B}(L)|$ and $|\mathbf{A}(L)|$ are assumed to have modulus greater than 1, a stability condition guaranteeing the non-explosive behaviour of **y** given any stable path for **x**. The Cowles Commission definition of exogeneity in this model (Koopmans and Hood, 1953, pp. 117–20) as set forth in Christ (1966, p. 156) is as follows:

> An exogenous variable in a stochastic model is a variable whose value in each period is statistically independent of the values of all the random disturbances in the model in all periods.

All other variables are endogenous. In the prototypical model set forth above **x** is exogenous and **y** is endogenous.
The Cowles Commission distinction between endogeneity and exogeneity applied to a specific class of models, with linear relationships and normally distributed disturbances. The exogenous variables **x** in the prototypical model have two important but quite distinct properties. First, the model may be solved to yield an expression for $\mathbf{y}(t)$ in terms of current and past values of **x** and ε,

$$y(t) = \mathbf{B}(L)^{-1}\mathbf{\Gamma}(L)\mathbf{x}(t) + \mathbf{B}(L)^{-1}\mathbf{A}(L)\varepsilon(t)\ .$$

Given suitably restricted $\mathbf{x}(t)$ (for example, all **x** uniformly bounded, or being realizations of a stationary stochastic process with finite variance) it is natural to complete the model by specifying that it is valid for all **x** meeting the restrictions, and this is often done. The variables **x** are therefore exogenous here as **x** is exogenous in a deterministic economic model. A second, distinct property of these variables is that in estimation $\mathbf{x}(t)$ $(-\infty < t < \infty)$ may be regarded as fixed, thus extending to the environment of simultaneous equation models methods of statistical inference initially designed for experimental settings. It was generally recognized that exogeneity in the prototypical model was a sufficient but not a necessary condition to justify treating variables as fixed for purposes of inference. If $\mathbf{u}(t)$ in the model is serially independent (that is, $\mathbf{A}(L)=\mathbf{I}$) then lagged values of **y** may also be treated as fixed for purposes of the model; this leads to the definition of ‘predetermined variables’ (Christ, 1966, p. 227) following Koopmans and Hood (1953, pp. 117–21):

> A variable is predetermined at time *t* if all its current and past values are independent of the vector of current disturbances in the model, and these disturbances are serially independent.

These two properties were not explicitly distinguished in the prototypical model (Koopmans, 1950; Koopmans and Hood, 1953) and tended to remain merged in the literature over the next quarter-century (for example, Christ, 1966; Theil, 1971; Geweke, 1978). By the late 1970s there had developed a tension between the two, due to the increasing sophistication of estimation procedures in nonlinear models, treatment of rational expectations, and the explicit consideration of the respective dynamic properties of endogenous and exogenous variables (Sims, 1972; 1977; Geweke, 1982). Engle, Hendry and Richard (1983), drawing on this literature and discussions at the 1979 Warwick Summer Workshop, formalized the distinction of the two properties we have discussed. Drawing on their definitions 2.3 and 2.5 and the discussions in Sims (1977) and Geweke (1982), **x** is *model exogenous* if given $\{\mathbf{x}(t), t \leq T\} \in \boldsymbol{R}(T)$ the model may restrict $\{\mathbf{y}(t), t \leq T\}$, but given

$$\{\mathbf{x}(t), t \leq T + J\} \in R(T + J)$$

there are no further restrictions on $\{\mathbf{y}(t), t \leq T\}$, for any $J>0$. If the model in fact does restrict $\{\mathbf{y}(t), t \leq T\}$, then **y** is model endogenous. As examples consider

$$\text{Model 1}: y(t) = ay(t-1) + bx(t) + u(t), x(t) = cx(t-1) + v(t);$$

$$\text{Model 2}: \quad y(t) = ay(t-1) + bx(t) + u(t), \quad x(t) = cx(t-1) + dy(t) + v(t);$$

$$\text{Model 3}: y(t) = ay(t-1) + b\{x(t) + E[x(t)|x(t-s), s>0]\} + u(t), x(t) = cx(t-1) + v(t).$$

In each case $u(t)$ and $v(t)$ are mutually and serially independent, and normally distributed. The parameters are assumed to satisfy the usual stability restrictions guaranteeing that x and y have normal distributions with finite variances. In all three models y is model endogenous, and x is model exogenous in Models 1 and 3 but not 2. For estimation the situation is different. In Model 1, treating $x(t)$, $x(t-1)$ and $y(t-1)$ as fixed simplifies inference at no cost; $y(t-1)$ is a classic predetermined variable in the sense of Koopmans and Hood (1953) and Christ (1966). Similarly in Model 2, $x(t-1)$ and $y(t-1)$ may be regarded as fixed for purposes of inference despite the fact that x and y are both model endogenous. When Model 3 is re-expressed

$$y(t) = ay(t-1) + bx(t) + bcx(t-1) + u(t), x(t) = cx(t-1) + v(t),$$

it is clear that $x(t)$ cannot be treated as fixed if the parameters are to be estimated efficiently since there are cross-equation restrictions involving the parameter c. Model exogeneity of a variable is thus neither a necessary nor a sufficient condition for treating that variable as fixed for purposes of inference.

The condition that a set of variables can be regarded as fixed for inference can be formalized, following Engle, Hendry and Richard (1983) along the lines given in Geweke (1984). Let

$$\mathbf{X} \equiv [\mathbf{x}(1), \ldots, \mathbf{x}(n)] \text{ and } \mathbf{Y} \equiv [\mathbf{y}(1), \ldots, \mathbf{y}(n)]$$

be matrices of n observations on the variables **x** and **y** respectively. Suppose the likelihood function $L(\mathbf{X}, \mathbf{Y}|\Theta)$ can be reparameterized by $\boldsymbol{\lambda} = F(\Theta)$ where F is a one-to-one transformation; $\lambda' = (\lambda_1, \lambda_2)', \quad (\lambda_1, \lambda_2) \in \Lambda_1 X \Lambda_2;$ and the investigator's loss function depends on parameters of interest λ_1 but not nuisance parameters λ_2. Then **x** is *weakly exogenous* if

$$L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = L_1(\mathbf{Y}|\mathbf{X}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_1) \cdot L_2(\mathbf{X}|\boldsymbol{\lambda}_2),$$

and in this case **y** is *weakly endogenous*. When this condition is met the expected loss function may be expressed using only $L_1(\mathbf{Y}|\mathbf{X}, \boldsymbol{\lambda}_1)$, that is, **x** may be regarded as fixed for purposes of inference.

The concepts of model exogeneity and weak exogeneity play important but distinct roles in the construction, estimation, and evaluation of econometric models. The dichotomy between variables that are model exogenous and model endogenous is a global property of a model, drawing in effect a logical distinction between the inputs of the model $\{x(t), t \leq T\} \in R(T)$ and the set of variables restricted by the model $\{\mathbf{y}(t), t \leq T\}$. Since model exogeneity stipulates that $\{\mathbf{x}(t), t \leq T + J\}$ places no more restrictions on $\{\mathbf{y}(t), t \leq T\}$ than does $\{\mathbf{x}(t), t \leq T\}$, the global property of model exogeneity is in principle testable, either in the presence or absence of other restrictions imposed by the model. When conducted in the absence of most other restrictions this test is often termed a ‘causality test’, and its use as a test of specification was introduced by Sims (1972). The distinction between weakly exogenous and weakly endogenous variables permits a simplification of the likelihood function that depends on the subset of the model's parameters that are of interest to the investigator. It is a logical property of the model: the same results would be obtained using $L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ as using $L(\mathbf{Y}|\mathbf{X}, \boldsymbol{\lambda}_1)$. The stipulation of weak exogeneity is therefore not, by itself, testable.

See Also

- causality in economics and econometrics
- identification
- simultaneous equations models

Bibliography

Christ, C.F. 1966. *Econometric Models and Methods*. New York: Wiley.

Engle, R.F., Hendry, D.F. and Richard, J.-F. 1983. Exogeneity. *Econometrica* 51, 277–304.

Geweke, J. 1978. Testing the exogeneity specification in the complete dynamic simultaneous equation model. *Journal of Econometrics* 7, 163–85.

Geweke, J. 1982. Causality, exogeneity, and inference. In *Advances in Econometrics*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.

Geweke, J. 1984. Inference and causality. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Koopmans, T.C. 1950. When is an equation system complete for statistical purposes? In *Statistical Inference in Dynamic Economic Models*, ed. T.C. Koopmans. New York: Wiley.

Koopmans, T.C. and Hood, W.C. 1953. The estimation of simultaneous economic relationships. In *Studies in Econometric Method*, ed. W.C. Hood and T.C. Koopmans. New York: Wiley.

Sims, C.A. 1972. Money, income, and causality. *American Economic Review* 62, 540–52.

Sims, C.A. 1977. Exogeneity and causal ordering in macroeconomic models. In *New Methods in Business Cycle Research*, ed. C. A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Theil, H. 1971. *Principles of Econometrics*. New York: Wiley.

How to cite this article

Geweke, John. "endogeneity and exogeneity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000077>
doi:10.1057/9780230226203.0471

The New Palgrave Dictionary of Economics Online

endogenous growth theory

Peter Howitt
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Endogenous growth theory explains long-run growth as emanating from economic activities that create new technological knowledge. This article sketches the outlines of the theory, especially the ‘Schumpeterian’ variety, and briefly describes how the theory has evolved in response to empirical discoveries.

Keywords

aggregate production function; capital accumulation; competition; creative destruction; economic growth; endogenous growth; human capital; innovations; intellectual capital; intermediate products; intertemporal utility maximization; law of large numbers; marginal product of capital; neoclassical growth theory; physical capital; product variety; productivity growth; research and development; saving rate; Schumpeterian growth; steady state; technological progress; technology; technology frontier; total factor productivity; transfer of technology

Article

Endogenous growth is long-run economic growth at a rate determined by forces that are internal to the economic system, particularly those forces governing the opportunities and incentives to create technological knowledge.

In the long run the rate of economic growth, as measured by the growth rate of output per person, depends on the growth rate of total factor productivity (TFP), which is determined in turn by the rate of technological progress. The neoclassical growth theory of Solow (1956) and Swan (1956) assumes the rate of technological progress to be determined by a scientific process that is separate from, and independent of, economic forces. Neoclassical theory thus implies that economists can take the long-run growth rate as given exogenously from outside the economic system.

Endogenous growth theory challenges this neoclassical view by proposing channels through which the

rate of technological progress, and hence the long-run rate of economic growth, can be influenced by economic factors. It starts from the observation that technological progress takes place through innovations, in the form of new products, processes and markets, many of which are the result of economic activities. For example, because firms learn from experience how to produce more efficiently, a higher pace of economic activity can raise the pace of process innovation by giving firms more production experience. Also, because many innovations result from R&D expenditures undertaken by profit-seeking firms, economic policies with respect to trade, competition, education, taxes and intellectual property can influence the rate of innovation by affecting the private costs and benefits of doing R&D.

AK theory

The first version of endogenous growth theory was AK theory, which did not make an explicit distinction between capital accumulation and technological progress. In effect it lumped together the physical and human capital whose accumulation is studied by neoclassical theory with the intellectual capital that is accumulated when innovations occur. An early version of AK theory was produced by Frankel (1962), who argued that the aggregate production function can exhibit a constant or even increasing marginal product of capital. This is because, when firms accumulate more capital, some of that increased capital will be the intellectual capital that creates technological progress, and this technological progress will offset the tendency for the marginal product of capital to diminish.
In the special case where the marginal product of capital is exactly constant, aggregate output *Y* is proportional to the aggregate stock of capital *K*:

$$Y = AK \tag{1}$$

where *A* is a positive constant. Hence the term ‘AK theory’.
According to AK theory, an economy's long-run growth rate depends on its saving rate. For example, if a fixed fraction *s* of output is saved and there is a fixed rate of depreciation δ, the rate of aggregate net investment is:

$$\frac{dK}{dt} = sY - \delta K$$

which along with (1) implies that the growth rate is given by:

$$g \equiv \frac{1}{Y}\frac{dY}{dt} = \frac{1}{K}\frac{dK}{dt} = sA - \delta.$$

Hence an increase in the saving rate *s* will lead to a permanently higher growth rate.
Romer (1986) produced a similar analysis with a more general production structure, under the assumption that saving is generated by intertemporal utility maximization instead of the fixed saving rate of Frankel. Lucas (1988) also produced a similar analysis focusing on human capital rather than physical capital; following Uzawa (1965) he explicitly assumed that human capital and technological knowledge were one and the same.

Innovation-based theory

AK theory was followed by a second wave of endogenous growth theory, generally known as ‘innovation-based’ growth theory, which recognizes that intellectual capital, the source of technological progress, is distinct from physical and human capital. Physical and human capital are accumulated through saving and schooling, but intellectual capital grows through innovation.
One version of innovation-based theory was initiated by Romer (1990), who assumed that aggregate productivity is an increasing function of the degree of product variety. In this theory, innovation causes productivity growth by creating new, but not necessarily improved, varieties of products. It makes use of the Dixit–Stiglitz–Ethier production function, in which final output is produced by labour and a continuum of intermediate products:

$$Y = L^{1-\alpha}\int_0^A x(i)^{\alpha}\, di,\ 0 < \alpha < 1 \qquad (2)$$

where *L* is the aggregate supply of labour (assumed to be constant), *x*(*i*) is the flow input of intermediate product *i*, and *A* is the measure of different intermediate products that are available for use. Intuitively, an increase in product variety, as measured by *A*, raises productivity by allowing society to spread its intermediate production more thinly across a larger number of activities, each of which is subject to diminishing returns and hence exhibits a higher average product when operated at a lower intensity.
The other version of innovation-based growth theory is the ‘Schumpeterian’ theory developed by Aghion and Howitt (1992) and Grossman and Helpman (1991). (Early models were produced by Segerstrom, Anant and Dinopoulos, 1990, and Corriveau, 1991). Schumpeterian theory focuses on quality-improving innovations that render old products obsolete, through the process that Schumpeter (1942) called ‘creative destruction.’
In Schumpeterian theory aggregate output is again produced by a continuum of intermediate products, this time according to:

$$Y = L^{1-\alpha}\int_0^1 A(i)^{1-\alpha}\, x(i)^{\alpha}\, di, \tag{3}$$

where now there is a fixed measure of product variety, normalized to unity, and each intermediate product *i* has a separate productivity parameter *A*(*i*). Each sector is monopolized and produces its intermediate product with a constant marginal cost of unity. The monopolist in sector *i* faces a demand curve given by the marginal product: $\alpha \cdot (A(i)L / x(i))^{1-\alpha}$ of that intermediate input in the final sector. Equating marginal revenue (α time this marginal product) to the marginal cost of unity yields the monopolist's profit-maximizing intermediate output:

$$x(i) = \xi L A(i)$$

where $\xi = \alpha^{2/(1-\alpha)}$. Using this to substitute for each *x*(*i*) in the production function (3) yields the aggregate production function:

$$Y = \theta A L \tag{4}$$

where $\theta = \xi^{\alpha}$, and where *A* is the average productivity parameter:

$$A \equiv \int_0^1 A(i)\ di.$$

Innovations in Schumpeterian theory create improved versions of old products. An innovation in sector *i* consists of a new version whose productivity parameter *A*(*i*) exceeds that of the previous version by the fixed factor $\gamma > 1$. Suppose that the probability of an innovation arriving in sector *i* over any short interval of length *dt* is $\mu \cdot dt$. Then the growth rate of *A*(*i*) is

$$\frac{dA(i)}{A(i)} \cdot \frac{1}{dt} = \begin{Bmatrix} (\gamma - 1) \cdot \frac{1}{dt} & \text{with probability } \mu \cdot dt \\ 0 & \text{with probability } 1 - \mu \cdot dt \end{Bmatrix}.$$

Therefore the expected growth rate of $A(i)$ is:

$$E(g) = \mu(\gamma - 1). \tag{5}$$

The flow probability μ of an innovation in any sector is proportional to the current flow of productivity-adjusted R&D expenditures:

$$\mu = \lambda R / A \tag{6}$$

where R is the amount of final output spent on R&D, and where the division by A takes into account the force of increasing complexity. That is, as technology advances it becomes more complex, and hence society must make an ever-increasing expenditure on research and development just to keep innovating at the same rate as before.

It follows from (4) that the growth rate g of aggregate output is the growth rate of the average productivity parameter A. The law of large numbers guarantees that g equals the expected growth rate (5) of each individual productivity parameter. From this and (6) we have:

$$g = (\gamma - 1)\lambda R / A.$$

From this and (4) it follows that the growth rate depends on the fraction of GDP spent on research and development, $n = R / Y$, according to:

$$g = (\gamma - 1)\lambda\theta L n. \tag{7}$$

Thus, innovation-based theory implies that the way to grow rapidly is not to save a large fraction of output but to devote a large fraction of output to research and development. The theory is explicit about how R&D activities are influenced by various policies, who gains from technological progress, who loses, how the gains and losses depend on social arrangements, and how such arrangements affect society's willingness and ability to create and cope with technological change, the ultimate source of economic growth.

Empirical challenges

Endogenous growth theory has been challenged on empirical grounds, but its proponents have replied with modifications of the theory that make it consistent with the critics' evidence. For example, Mankiw, Romer and Weil (1992), Barro and Sala-i-Martin (1992) and Evans (1996) showed, using data from the second half of the 20th century, that most countries seem to be converging to roughly similar long-run growth rates, whereas endogenous growth theory seems to imply that, because many countries have different policies and institutions, they should have different long-run growth rates. But the Schumpeterian model of Howitt (2000), which incorporates the force of technology transfer, whereby the productivity of R&D in one country is enhanced by innovations in other countries, implies that all countries that perform R&D at a positive level should converge to parallel long-run growth paths.
The key to this convergence result is what Gerschenkron (1952) called the 'advantage of backwardness'; that is, the further a country falls behind the technology frontier, the larger is the average size of innovations, because the larger is the gap between the frontier ideas incorporated in the country's innovations and the ideas incorporated in the old technologies being replaced by innovations. This increase in the size of innovations keeps raising the laggard country's growth rate until the gap separating it from the frontier finally stabilizes.
Likewise, Jones (1995) has argued that the evidence of the United States and other OECD countries since 1950 refutes the 'scale effect' of Schumpeterian endogenous growth theory. That is, according to the growth equation (7) an increase in the size of population should raise long-run growth by increasing the size of the workforce L, thus providing a larger market for a successful innovator and inducing a higher rate of innovation. But in fact productivity growth has remained stationary during a period when population, and in particular the number of people engaged in R&D, has risen dramatically. The models of Dinopoulos and Thompson (1998), Peretto (1998) and Howitt (1999) counter this criticism by incorporating Young's (1998) insight that, as an economy grows, proliferation of product varieties reduces the effectiveness of R&D aimed at quality improvement by causing it to be spread more thinly over a larger number of different sectors. When modified this way the theory is consistent with the observed coexistence of stationary TFP growth and rising population, because in a steady state the growth-enhancing scale effect is just offset by the growth-reducing effect of product proliferation.
As a final example, early versions of innovation-based growth theory implied, counter to much evidence, that growth would be adversely affected by stronger competition laws, which by reducing the profits that imperfectly competitive firms can earn ought to reduce the incentive to innovate. However, Aghion and Howitt (1998, ch. 7) describe a variety of channels through which competition might in fact spur economic growth. One such channel is provided by the work of Aghion et al. (2001), who show

that, although an increase in the intensity of competition will tend to reduce the absolute level of profits realized by a successful innovator, it will nevertheless tend to reduce the profits of an unsuccessful innovator by even more. In this variant of Schumpeterian theory, more intense competition can have a positive effect on the rate of innovation because firms will want to escape the competition that they would face if they lost whatever technological advantage they have over their rivals.

Much more work needs to be done before we can claim to have a reliable explanation for why economic growth is faster in some countries and in some time periods than in others. But the fact that much of the cross-country variation in growth rates is attributable to differences in productivity growth rather than differences in rates of capital accumulation suggests that endogenous growth theory, which aims to provide an economic explanation of these differences in productivity growth, will continue to attract economists' attention for years to come.

See Also

- Schumpeterian growth and growth policy design

Bibliography

Aghion, P. and Howitt, P. 1992. A model of growth through creative destruction. *Econometrica* 60, 323–51.

Aghion, P. and Howitt, P. 1998. *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Aghion, P., Harris, C., Howitt, P. and Vickers, J. 2001. Competition, imitation and growth with step-by-step innovation. *Review of Economic Studies* 68, 467–92.

Barro, R.J. and Sala-i-Martin, X. 1992. Convergence. *Journal of Political Economy* 100, 223–51.

Corriveau, L. 1991. Entrepreneurs, growth, and cycles. Doctoral dissertation, University of Western Ontario.

Dinopoulos, E. and Thompson, P. 1998. Schumpeterian growth without scale effects. *Journal of Economic Growth* 3, 313–35.

Evans, P. 1996. Using cross-country variances to evaluate growth theories. *Journal of Economic Dynamics and Control* 20, 1027–49.

Frankel, M. 1962. The production function in allocation and growth: a synthesis. *American Economic Review* 52, 995–1022.

Gerschenkron, A. 1952. Economic backwardness in historical perspective. In *The Progress of*

Underdeveloped Areas, ed. B.F. Hoselitz. Chicago: University of Chicago Press.

Grossman, G.M. and Helpman, E. 1991. *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.

Howitt, P. 1999. Steady endogenous growth with population and R&D inputs growing. *Journal of Political Economy* 107, 715–30.

Howitt, P. 2000. Endogenous growth and cross-country income differences. *American Economic Review* 90, 829–46.

Jones, C.I. 1995. R&D-based models of economic growth. *Journal of Political Economy* 103, 759–84.

Lucas, R.E., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Mankiw, N.G., Romer, D. and Weil, D.N. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107, 407–37.

Peretto, P.F. 1998. Technological change and population growth. *Journal of Economic Growth* 3, 283–311.

Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Romer, P.M. 1990. Endogenous technological change. *Journal of Political Economy* 98, S71–S102.

Schumpeter, J.A. 1942. *Capitalism, Socialism and Democracy*. New York: Harper.

Segerstrom, P.S., Anant, T.C.A. and Dinopoulos, E. 1990. A Schumpeterian model of the product life cycle. *American Economic Review* 80, 1077–91.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Uzawa, H. 1965. Optimal technical change in an aggregative model of economic growth. *International Economic Review* 6, 18–31.

Young, A. 1998. Growth without scale effects. *Journal of Political Economy* 106, 41–63.

How to cite this article

Howitt, Peter. "endogenous growth theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000079> doi:10.1057/9780230226203.0472

The New Palgrave Dictionary of Economics Online

endogenous market incompleteness

Christopher Sleet
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Endogenously incomplete models derive restrictions on asset trading from primitive constraints on the enforcement and monitoring technologies available to societies. They have been applied to a wide variety of macroeconomic problems. This article reviews some of these applications and the models that underpin them.

Keywords

asset pricing models; autarky; bilateral insurance games; commitment; default; dynamic moral hazard models; endogenously incomplete markets; equity premium; Euler equations; incomplete markets; information revelation; limited enforcement models; risk sharing; taxation of assets

Article

An asset trading arrangement is incomplete if it is too restrictive to ensure a fully Pareto-optimal allocation of risk. Endogenously incomplete models derive such trading arrangements from primitive frictions. They are to be contrasted with models that *assume* a particular incomplete asset markets structure.

Recent contributions to the endogenous incompleteness literature have emphasized imperfections in the enforcement and monitoring technologies available to societies. They derive endogenous market structures, sometimes supplemented with a tax system, as decentralizations of planning problems in which the planner faces one or both of these imperfections. These market-tax structures ensure that agents are provided with incentives to honour promises that cannot be costlessly enforced or that are contingent on states that cannot be costlessly observed. By construction they admit equilibria that are constrained efficient.

Models with endogenous incompleteness have received a variety of applications in macroeconomics. They have been used to enhance understanding of risk sharing, asset pricing and business cycles; on the

normative side they have been applied to analyses of optimal fiscal policy. Here I review some of these applications and the models that underpin them.

Limited enforcement

The canonical example of a limited enforcement model is the bilateral insurance game of Kocherlakota (1996). In this game, two risk-averse agents are endowed with random and imperfectly correlated income processes. Neither agent can be compelled to deliver resources to the other, even if they have promised to do so in the past.

Equilibrium allocations in this setting can be implemented with strategies that revert to autarky following an agent defection. Agents with high-income shocks can be induced to share some of their resources by the threat of such reversion and, when this is insufficient, by promises of extra resources in the future. Such promises introduce additional dynamics into optimal equilibrium allocations; shocks that cannot be smoothed over states are smoothed over time instead, ensuring that individual consumption is persistent even when aggregate consumption is not.

Constrained-efficient allocations in limited enforcement economies can be decentralized using a complete set of Arrow security markets coupled with endogenous debt limits (see Alvarez and Jermann, 2000). Intuitively, agents can borrow only up to the amount that they are willing to pay back in the future given that the penalty for default is consignment to autarky. Thus, the limited enforcement friction provides a micro-foundation for the often-made assumption of a debt limit tighter than that implied by an agent's intertemporal budget constraint. In the limited enforcement case, however, the debt limit is state-contingent; it depends upon the value of autarky to the agent. Since this value is a function of individual and aggregate shocks, the parameters of the shock process and, in richer models, the agent's opportunities for self- or public insurance after exclusion from markets, so too is the debt limit.

When agents’ endogenous debt limits periodically bind, risk sharing is disrupted; individual consumption, conditional on the aggregate state, is positively correlated with current and past individual income. Qualitatively, such departures from full risk-sharing cohere well with evidence on individual consumption. In Alvarez and Jermann's (2001) quantitative analysis of a calibrated limited enforcement model, the endogenous debt limits bind fairly often and permit relatively little risk sharing. This is consistent with evidence on the sharing of low-frequency risks. Alvarez and Jermann's analysis also has implications for asset pricing. They obtain a volatile asset pricing kernel and risk premia that are large and time varying. These implications are consistent with asset pricing data, but contrast with those of the benchmark representative agent asset pricing model.

Cross-country consumption data also exhibit apparent departures from full risk sharing. Standard models (with complete markets) imply co-movements in consumption that exceed those in output, yet the data suggests the reverse. Kehoe and Perri (2002) show that a limited enforcement model augmented with production and physical capital accumulation can go some way to explaining this anomaly.

Recent papers have considered alternative penalties for default including the confiscation of an endogenously valued collateral asset (see, Lustig, 2005) or the payment of a fixed default cost (Cooley, Marimon and Quadrini, 2004). These contributions illustrate the scope of limited enforcement models: Lustig explores the implications of endogenously valued collateral for asset pricing and obtains a large and time-varying price of risk; Cooley, Marimon and Quadrini examine the role of limited enforcement

frictions in propagating business cycle shocks. Cordoba (2005) and Arpad and Cárceles-Poveda (2005), however, sound cautionary notes. They provide calibrated models in which the introduction of collateral relaxes endogenous debt limits so much that agents can fully diversify risk.

Private information

An alternative line of research has analysed environments in which risk-averse agents privately observe shocks to their endowments, tastes or productivity (see, for example, Atkeson and Lucas, 1992). In this setting, agents must be provided with incentives to reveal information. The socially efficient provision of incentives requires the conditioning of current consumption on an agent's history of shock reports. Intuitively, agents are rewarded for reporting a low current need for resources with the promise of more consumption in the future. Thus, intertemporal consumption smoothing is enhanced and interstate smoothing disrupted.

Albanesi and Sleet (2006) and Kocherlakota (2005) show that optimal information-constrained allocations can be implemented with a mixture of non-contingent debt markets and taxes. Thus, these authors derive joint restrictions on the market structure *and* the tax system from primitive informational frictions. Central to their analyses is an ‘inverted Euler equation’. If $\{c_t^*\}_{t=0}^{\infty}$ denotes the optimal consumption allocation, this equation is given by:

$$\frac{1}{u'(c_t^*(z^t, \theta^t))} = \beta\lambda_{t+1}(z^{t+1})E_t\left[\frac{1}{u'(c_{t+1}^*)}|z^{t+1}, \theta^t\right]. \tag{1}$$

Here θ^t denotes an agent's period t history of privately observed shocks, z^t and z^{t+1} denote t and t+1 histories of observable aggregate shocks, β is the agent's discount factor and u' her marginal utility of consumption. λ_{t+1} is a social stochastic discount factor (SSDF) that ‘prices’ resources delivered after each history z^{t+1}. Golosov, Kocherlakota and Tsyvinski (2003) show that such equations hold in a large class of dynamic moral hazard models. They imply a wedge between an agent's conditional expected intertemporal marginal rate of substitution (IMRS) and the SSDF. This wedge provides a rationale for asset taxation; intuitively, agents must be discouraged from saving at date t since greater wealth at t+1 undermines incentives at that date. However, the implications for asset taxation are subtle. The optimal allocation cannot be implemented with an asset tax that merely ‘matches the wedge’ and equates the conditional expectation of an agent's IMRS to the SSDF. Instead, marginal asset taxes at t+1 are used to generate a positive covariance between the after-tax asset return and the agent's consumption that deters savings. In some cases, the expected asset tax is zero and the wedge is entirely generated by this covariance effect.

Positive analyses of dynamic moral hazard are relatively scarce. Green and Oh (1991) contrast the

empirical implications of various incomplete market models, including those with moral hazard. Kocherlakota and Pistaferri (2005) identify λ_{t+1} with the market discount factor, assume that utility has the constant relative risk aversion property and use (1) to derive expressions for λ_{t+1} in terms of cross-sectional moments of the consumption distribution. They then investigate the implications of this dynamic moral hazard model for asset pricing and, in particular, the equity premium and risk-free rate. They find that plausible values of the coefficient of relative risk aversion set the equity premium pricing error to zero.

In all of the dynamic moral hazard models described so far, the consumption of agents is observable. An alternative assumption is that agents can undertake asset trades that are hidden from society. Agents must now be given incentives to reveal information *and* save an appropriate amount. This places additional constraints on risk sharing. When agents can control their publicly observable histories and can save at the prices implied by an exogenously given sequence of SSDFs, these constraints are severe. In this case, the optimal allocation is identical to that in an economy with riskless debt (see Cole and Kocherlakota, 2001). This result is important as it provides a micro-foundation for models that exogenously restrict agents to the trading of such debt.

Government incentive problems

Governments or mechanism designers may also have difficulty keeping their promises. There is a long tradition of considering commitment problems in Ramsey models. In these, a socially benevolent government typically has access to a restricted set of linear tax mechanisms and an asset market in which it can trade claims to resources. *Ex ante* optimal policy entails implicit promises over future allocations and, in particular, the expected value of the government's future stream of primary surpluses that it is rarely in the government's interests to keep. For example, if the government can default on its debt it will, since in this way it can avoid the distortionary taxes necessary for debt repayment. As in the limited-enforcement models described above, reversion to autarky after a default can sustain some equilibrium borrowing by the government, though typically it implies a tight endogenous debt limit (Chari and Kehoe, 1993). Sleet (2004) and Sleet and Yeltekin (2006a) consider models in which the government's true spending needs are not publicly observable. Although the government has access to a complete set of contingent claims markets, in equilibrium it is required to adopt a debt-trading policy consistent with truthful revelation of its spending needs. This limits its ability to buy claims against high spending-needs states and sell them against low spending-needs ones. The outcome is enhanced intertemporal, as opposed to inter-state, smoothing of taxes.

The optimal allocations and market-tax implementations implied by dynamic moral hazard models also involve promises from a planner (or government) to an agent. These allocations often entail the absorption of almost all agents by a minimal utility immiserating state; they thus place strong demands on the planner's ability to commit. Sleet and Yeltekin (2006b) remove this ability. They show that optimal allocations without planner commitment solve the problems of *committed* planners who discount the future *less heavily* than agents. Coupling this result with the work of Farhi and Werning (2005), who directly assume a planner discount factor in excess of the agents, suggests that constrained optimal allocations can be implemented with non-contingent debt, an income tax and a progressive estate tax. Analysis of dynamic moral hazard models without societal commitment is, however, still in

its infancy and much remains to be done.

See Also

- default and enforcement constraints
- optimal fiscal and monetary policy (without commitment)
- social insurance

Bibliography

Arpad, A. and Cárceles-Poveda, E. 2005. Endogenous trading constraints with incomplete markets. Working paper, University of Rochester.

Albanesi, S. and Sleet, C. 2006. Dynamic optimal taxation with private information. *Review of Economic Studies* 73, 1–30.

Alvarez, F. and Jermann, U. 2000. Efficiency, equilibrium and asset pricing with risk of default. *Econometrica* 68, 775–97.

Alvarez, F. and Jermann, U. 2001. Quantitative asset pricing implications of endogenous solvency constraints. *Review of Financial Studies* 14, 1117–51.

Atkeson, A. and Lucas, R. 1992. On efficient distribution with private information. *Review of Economic Studies* 59, 427–53.

Chari, V. and Kehoe, P. 1993. Sustainable plans and debt. *Journal of Economic Theory* 60, 175–95.

Cole, H. and Kocherlakota, N. 2001. Efficient allocations with hidden income and storage. *Review of Economic Studies* 68, 523–42.

Cooley, T., Marimon, R. and Quadrini, V. 2004. Aggregate consequences of limited contract enforceability. *Journal of Political Economy* 112, 817–47.

Cordoba, J.-C. 2005. US inequality: debt constraints or incomplete markets? Working paper, Rice University.

Farhi, E. and Werning, I. 2005. Inequality, social discounting and estate taxation. Working Paper No. 11408. Cambridge, MA: NBER.

Golosov, M., Kocherlakota, N. and Tsyvinski, A. 2003. Optimal indirect and capital taxation. *Review of Economic Studies* 70, 569–87.

Green, E. and Oh, S. 1991. Contracts, constraints and consumption. *Review of Economic Studies* 58, 883–99.

Kehoe, P. and Perri, F. 2002. International business cycles with endogenously incomplete markets. *Econometrica* 70, 907–28.

Kocherlakota, N. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63, 595–609.

Kocherlakota, N. 2005. Zero expected wealth taxes: a Mirrleesian approach to dynamic optimal taxation. *Econometrica* 73, 1587–622.

Kocherlakota, N. and Pistaferri, L. 2005. Asset pricing implications of Pareto optimality with private information. Discussion Paper No. 4930. London: CEPR.

Lustig, H. 2005. The market price of aggregate risk and the wealth distribution. Working Paper No. 11132. Cambridge, MA: NBER.

Sleet, C. 2004. Optimal taxation with private government information. *Review of Economic Studies* 71, 1217–39.

Sleet, C. and Yeltekin, S. 2006a. Optimal taxation with endogenously incomplete markets. *Journal of Economic Theory* 127, 36–73.

Sleet, C. and Yeltekin, S. 2006b. Credibility and endogenous societal discounting. *Review of Economic Dynamics* 9, 410–37.

How to cite this article

Sleet, Christopher. "endogenous market incompleteness." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000297> doi:10.1057/9780230226203.0473

The New Palgrave Dictionary of Economics Online

energy economics

Ying Fang and Robin Sickles

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Energy economics studies energy resources and energy commodities. It includes forces motivating firms and consumers to supply, convert, transport, use energy resource; market and regulatory structures; distributional and environmental consequences; economically efficient use. The fact that energy use is dominantly depletable resources, particularly fossil fuels, makes this study unique. The energy industry has moved into the 21st century with promises of both profits and a short-term future. With added pressure from government, cleaner fuels are being introduced on a continual basis. Additionally, the expanding energy demand from developing countries is changing the energy market.

Keywords

conservation; depletable resources; derived demand; dynamic models; ecological economics; energy economics; energy policies; environmental economics; essential goods; Framework Convention on Climate Change; intertemporal choices; Kyoto Protocol; oil; Organization of Petroleum Exporting Countries; renewable resources; Strategic Petroleum Reserve (USA)

Article

Energy is crucial to the economic progress and social development of nations. Energy can be neither created nor destroyed but its form can be changed. Energy comes from the physical environment and ultimately returns there. The demand for energy is a derived demand. The value of energy is assessed by its ability to provide a set of desired services in both industry and in the household.
Energy commodities are economic substitutes. Energy resources are depletable or renewable and storable or non-storable. On a global scale the 20th century was dominated by the use of fossil fuels. According to the US Department of Energy, in the year 2000 global commercial energy consumption consisted of petroleum (39 per cent), coal (24 per cent), natural gas (23 per cent), hydro (6 per cent), nuclear (7 per cent) and others (1 per cent). In 1999, of the total sources of energy consumed in the

United States, 92 per cent were from depletable resources and only 8 per cent from renewable resources (EIA, 2001). No one doubts that fossil fuels are subject to depletion, and that depletion leads to scarcity, which in turn leads to higher prices. Resources are defined as ‘non-conventional’ when they cannot be produced economically at today's prices and with today's technology. With higher prices, however, the gap between conventional and non-conventional oil resources narrows. Ultimately, a combination of escalating prices and technological enhancements can transform the non-conventional into the conventional. Much of the pessimism about oil resources has been focused entirely on conventional resources.

Demand for energy

Bohi and Toman (1996) suggest a link between energy and economy. An abundance of empirical research suggests a strong correlation between increases in oil prices and decreases in macroeconomic performance for oil-importing industrialized countries. Higher import costs may lead to higher price levels and inflation. Industrial energy demand increases most rapidly at the initial stages of development, but growth slows steadily throughout the industrialization process (Medlock and Soligo, 2001). Energy demand for transportation rises steadily, and takes the major share of total energy use at the latter stages of developments.

Elasticity of energy demand

Is energy an essential good? In economics, an essential good is one for which the demand remains positive no matter how high its price. Energy is often described as an essential good because human activity would be impossible absent use of energy. Although energy is essential to humans, neither particular energy commodities nor any purchased energy commodities are essential goods because consumers can convert one form of energy into another.

The income elasticity of energy demand is defined as the percentage change in energy demand given a one per cent change in income holding all else constant, or

$$\varepsilon_y = \frac{\%\Delta e}{\%\Delta y} = \frac{de}{dy} \cdot \frac{y}{e}$$

where *e* denotes energy demand and *y* denotes income. ‘The household sector's share of aggregate energy consumption tends to fall with income, the share of transportation tends to rise, and the share of industry follows an inverse-U pattern’ (Judson, Schmalensee and Stoker, 1999).

The price elasticity of energy demand is defined as the percentage change in energy demand given a one per cent change in price, with all else held constant, or

$$\varepsilon_p = \frac{\%\Delta e}{\%\Delta p} = \frac{de}{dp} \cdot \frac{p}{e}$$

where p denotes the price of energy.

Cooper (2003) uses a multiple regression model derived from an adaptation of Nerlove's (1958) partial adjustment model to estimate both the short-run and the long-run elasticity of demand for crude oil in 23 countries over a 30-year period from 1971 to 2000. The estimates so obtained confirm that the demand for crude oil internationally is highly insensitive to changes in price.

Demand substitution between energy commodities and others

Denny, Fuss, and Waverman (1981) used time-series data for 18 US manufacturing two-digit industries (1948–71) and 18 Canadian manufacturing industry groups (1962–75). Their results were also mixed: for both the United States and Canada, energy and capital were substitutes in the food industry, but they were complements in the tobacco industry.

Energy consumption can be modelled either as providing utility to households or as an input in the production process for firms. To express the former problem mathematically, a representative consumer maximizes utility, $U(z,e)$, which is function of energy consumption, e, and all other consumption, z, subject to the constraint that expenditures cannot exceed income, y. Let the energy variable be a vector of n energy products, $e = (e_1, e_2, \ldots, e_n)$; we could examine the substitution possibilities across energy products. Allowing the price of good j to be represented as p_j, the consumer is assumed to

$$\max_{z, e_1, \ldots e_n} U(z, e_1, \ldots e_n)$$
$$\text{subject to: } y \geq p_z z + p_{e_1} e_1 + \ldots + p_{e_n} e_n$$

The first order necessary conditions for a maximum for this problem can be solved to yield demand equations for each of the energy products and for all other consumption. With some adjustments, the above method can be applied to a representative firm.

Recent research focuses mainly on dynamic models. Dynamic models allow for a more complete analysis of the energy demand because they are capable of capturing factors that generate the asymmetries. In addition, dynamic models incorporate the intertemporal choices that a consumer/firm must make when maximizing utilities or profits over some time horizon. Medlock and Soligo (2002) developed a useful framework. Let z_t be multiple types of capital and e_t be multiple types of energy consumption. Denoting time using the subscript t, the consumer will maximize the discounted sum of lifetime utility, $\sum_{t=0}^{T} \beta^t U(z_t, e_t)$, subject to the constraint whereby capital goods purchases (i_t), purchases of other goods (z_t), purchases of energy (e_t), and savings (s_t) in each period cannot exceed this

period's income (y_t), plus the return of last period's saving ($(1+r)s_{t-1}$). It is assumed that capital goods depreciate at a rate δ, savings earn a rate return r, the discount rate is $0 < \beta < 1$, and all initial conditions are given.
Consumers will

$$\max_{z,e,s} \sum_{t=0}^{T} \beta^t U(z_t, e_e)$$
$$\text{subject to } p_{zt}z_t + p_{et}e_t + p_{kt}i_t + s_t \leq y_t + (1+r)s_{t-1}$$
$$i_t = k_t - (1-\delta)k_{t-1}$$
$$z_t, u_t, k_t \geq 0 \text{ for } t = 1, \ldots, T$$

Medlock and Soligo (2002) indicate that the income elasticity of passenger vehicle demand is decreasing as the real GDP per capita increases, no matter in the long run or in the short run. For example, with 1988 purchasing power parity dollar, if the real GDP per capita is $500, the short-run elasticity is 0.74 and the long-run elasticity is 3.61; if the real GDP per capita is $20,000, the short-run and the long-run elasticity are 0.02 and 0.09, respectively.

Energy supply

OPEC

The Organization of the Petroleum Exporting Countries (OPEC) comprises countries that have organized for the purpose of negotiating with oil companies on matters of petroleum production, prices, and future concession rights. Founded on 14 September 1960 at a Baghdad conference, OPEC originally consisted of only five countries – Iran, Iraq, Kuwait, Saudi Arabia and Venezuela – but has since expanded to include several others: Algeria, Indonesia, Libya, Nigeria, Qatar and United Arab Emirates. The members of OPEC, which constitute a cartel, agree on the quantity and the prices of the oil exported. OPEC seeks to regulate oil production, and thereby manage oil prices, primarily by setting quotas for its members. Member countries hold about 75 per cent of the world's oil reserves, and supply 40 per cent of the world's oil. Loury (1990) is an excellent clarification; it studies a dynamic, quantity-setting duopoly game. The author considers a model of competition between two independent firms, *A* and *B*, facing indivisibility in production, with given limitations on their cumulative capacities to produce. At date *t* the flow rates of production of firms *A* and *B* are denoted by q_t^a and q_t^b respectively. The demand side of the market is passively modelled; buyers do not behave strategically. There is an inverse demand function, $P(\cdot)$, which is time invariant and dependent only on the total rate of flow of output of the two firms. Define the discount factor δ^t, a dollar received on date *t* is worth dollars at date zero. Then their respective payoffs are V_A and V_B where:

$$V_A = \beta \sum_t \delta^t \left[q_t^a P(Q_t) \right]; \text{ and } V_B = \beta \sum_t \delta^t \left[q_t^b P(Q_t) \right]$$

for β, the lump sum equivalent of the flow of one dollar. It is shown that the ability to precommit can be disadvantageous. Loury (1990) also formalizes the intuition that, when indivisibilities are important, tacit coordination of plans so as to avoid destructive competition is facilitated by establishing a convention of ‘taking turns’, that is, a self-enforcing norm of mutual, alternate forbearance. Since worldwide oil sales are denominated in US dollars, changes in the value of the dollar against other world currencies affect OPEC's decisions on how much oil to produce. After the introduction of the euro, Iraq unilaterally decided it wanted to be paid for its oil in euros instead of US dollars.

OPEC decisions have a strong influence on international oil prices. A good example is the 1973 energy crisis, in which OPEC refused to ship oil to Western countries that had supported Israel in its conflict with Egypt, the Yom Kippur War. This refusal caused a fourfold increase in oil prices, which lasted five months, starting on 17 October 1973 and ending on 18 March 1974. OPEC nations then agreed, on 7 January 1975, to raise crude oil prices by ten per cent. The high and rising price of oil burdens industrial oil-importing countries in two ways. First, it renders the standard of living lower than otherwise. Second, it affects the economy in ways that are difficult for policymakers to manage: on the one hand, the rising oil price spurs general inflation; on the other hand, it depresses domestic demand and employment.

Unlike many other cartels, OPEC has been successful at increasing the price of oil for extended periods. Much of OPEC's success can be attributed to Saudi Arabia's flexibility. It has tolerated cheating on the part of other cartel members, and cut its own production to compensate for other members exceeding their production quotas. This actually gives them good leverage because, with most members at full production, Saudi Arabia is the only member with spare capacity and the ability to increase supply, if needed. The policy has been successful. However, OPEC's ability to raise prices does have some limits. An increase in oil price decreases consumption, and could cause a net decrease in revenue. Furthermore, an extended rise in price could encourage systematic behaviour change, such as alternative energy utilization, or increased conservation. As of August 2004, OPEC has been communicating that its members have little excess pumping capacity, indicating that the cartel is losing influence over crude oil prices.

The six major non-OPEC oil-producing nations are Norway, Russia, Canada, Mexico, the United States and Oman. Russian production increases dominated non-OPEC production growth from 2000 onward and was responsible for most of the non-OPEC increases since the turn of the century. In 2001, a weakening US economy and increases in non-OPEC production put downward pressure on prices. In response OPEC once again entered into a series of reductions in member quotas, cutting production by 3.5 million barrels per day by 1 September 2001. In the absence of the September 11, 2001 terrorist attack this would have been sufficient to moderate or even reverse the trend.

In the wake of that attack the crude oil price plummeted. Under normal circumstances a drop in price of this magnitude would have resulted in another round of quota reductions, but, given the political climate, OPEC delayed additional cuts until January 2002, when it reduced its quota by 1.5 million barrels per day and was joined by several non-OPEC producers, including Russia, which promised combined daily

production cuts of an additional 462,500 barrels. This had the desired effect, with oil prices moving into the $25 per barrel range by March 2002. By mid-year the non-OPEC members were restoring their production cuts, but prices continued to rise and US inventories reached a 20-year low later in the year. By year's end oversupply was not a problem. Problems in Venezuela led to a strike at Petroleos de Venezuela (PDVSA) causing Venezuelan production to plummet. In the wake of the strike Venezuela was never able to restore capacity to its previous levels. On 19 March 2003, just as some Venezuelan production was beginning to return, military action began in Iraq. Meanwhile, inventories remained low in the United States and other OECD countries. With an improving economy US demand was increasing, and Asian demand for crude oil was growing at a rapid pace. The loss of production capacity in Iraq and Venezuela, combined with increased production to meet growing international demand, led to the erosion of excess oil production capacity. During much of 2004 and 2005 the spare capacity to produce oil has been less than one million barrels per day. A million barrels per day is not enough spare capacity to cover an interruption of supply from almost any OPEC producer. In a world that consumes over 80 million barrels of petroleum products per day, that adds a significant risk premium to crude oil price and is largely responsible for prices in excess of $40 per barrel. For further information, see Energy Information Administration (EIA).

Future energy supply

Undoubtedly, depletable resource use cannot dominate forever. Therefore, a future transition from depletable resources, particularly from fossil fuels, is inevitable. However, which renewable energy sources will dominate future consumption is unclear. And there is great uncertainty about the timing of a shift to renewable energy resources. Although this is a formidable question, Wiser et al. (2004) introduce green pricing programmes, which represent one way whereby consumers can voluntarily support renewable energy. Their analysis yields several interesting results. Programme duration affects customer response. The longer a programme has been operating, the more likely it is that its message has spread and the higher the probability of strong programme success. Initial customer participants in green pricing programmes may not be highly sensitive to cost, and may be willing to purchase higher quantities of renewable energy, which makes the case for utilities focusing on maximizing renewable energy sales, not customer participation rates. Price premiums and minimum monthly costs are not the primary determinants of programme success. Price may become a more important determinant as green pricing programmes expand beyond the early innovator customers. And smaller utilities appear to have a greater likelihood of achieving success.

The prospect of producing clean, sustainable power in substantial quantities from renewable energy sources is arousing renewed interest worldwide. Hydroelectricity is the only renewable energy source today that makes a large contribution to world energy production. Its long-term technical potential is believed to be 9 to 12 times current production, but increasingly environmental concerns block new dams. The large areas affected may have a negative environmental impact. Hydroelectricity dams, like the Aswan Dam, have adverse consequences both upstream and downstream.

Wind power is one of the most cost-competitive renewable sources today. Its long-term technical potential is believed to be five times current global energy consumption. But this requires 12.7 per cent of all land area and the facilities have to be built at certain height. Geothermal power and tidal power are the only renewable sources not dependent on the sun, but are today limited to special locations. Most

renewable sources are diffuse and require large land areas and great quantities of construction material for significant energy production. There is some doubt that they can be built rapidly enough to replace fossil fuels. The large and sometimes remote areas may also increase energy loss and cost from distribution. On the other hand, some forms allow small-scale production and may be placed very close to or directly at consumer households, businesses, and industries. We may forecast the future coexistence of multi-renewable energy sources. Boyle (1996) provides a comprehensive overview of the principal renewable energy sources: solar thermal, biomass, tidal, wave, photovoltaic, hydro, wind and geothermal.

Forecasts of the energy markets

According to Energy Information Administration (EIA, 2005b), based on its expectations for world energy prices, world energy consumption is projected to increase by 57 per cent from 2002 to 2025. World oil use is expected to grow from 78 million barrels per day in 2002 to 103 million barrels per day in 2015 and 119 million barrels per day in 2025. The projected increment in worldwide oil use would require an increment in world oil production capacity of 42 million barrels per day above 2002 levels. Members of OPEC are expected to be the major suppliers of the increased production that will be required to meet demand, accounting for 60 per cent of the projected increase in world capacity. In addition, non-OPEC suppliers are expected to add nearly 17 million barrels per day of oil production capacity between 2002 and 2025. Substantial increments in new non-OPEC supply are expected to come from the Caspian Basin, Western Africa, and Central and South America.

Natural gas is projected to be the fastest-growing component of world primary energy consumption. Consumption of natural gas worldwide increases in the forecast by an average of 2.3 per cent annually from 2002 to 2025, compared with projected annual growth rates of 1.9 per cent for oil consumption and 2.0 per cent for coal consumption. From 2002 to 2025, consumption of natural gas is projected to increase by 69 per cent, and its share of total energy consumption is projected to grow from 23 to 25 per cent.

Natural gas is seen as a desirable alternative to electricity generation in many parts of the world, given its relatively efficiency in comparison with other energy sources, as well as the fact that it burns more cleanly than either coal or oil and thus is an attractive alternative for countries pursuing reductions in greenhouse gas emission.

World coal consumption is projected to increase at an average rate of 2.5 per cent per year. From 2015 to 2025, the projected rate of increase in world coal consumption slows to 1.3 per cent annually. Coal is expected to maintain its importance as an energy source in both the electric power and industrial sectors. Hydroelectricity and other renewable energy sources are expected to maintain their 8 per cent share of total energy use worldwide throughout the projection period. Much of the projected growth in renewable electricity generation is expected to result from the completion of large hydroelectric facilities in emerging economies, particularly in Asia.

Energy policies

The study of depletable resource economics began with articles by H. Hotelling (1931), which examined

economically intertemporal optimal extraction from a perfectly known stock of the resource, with perfectly predictable future prices of the extracted commodity. Sweeney (1977) and Stiglitz (1976) both clarified the Hotelling rule in the presence of monopoly, and Gilbert and Richard (1978) and Salant (1976) extended this to the case of a dominant producer with a competitive fringe and several dominant producers, analogous to the case of OPEC. Pindyck (1982) and Kolstad (1994) extended the model to several imperfectly substitutable exhaustible resources.

Energy security refers to loss of economic welfare that may occur as a result of a change in the price of availability of energy. In the years following the 1973 oil price rise, US energy policy could be characterized as generally suspicious of the market. Supply augmentation was a major strategy pursued by the US government in addressing the ‘energy crisis’. The security dimensions of energy supply have always been viewed as appropriate concerns of the government. One could argue that the Gulf War in the early 1990s was simply a form of energy policy, protecting Western oil supplies originating in the Middle East. Countries other than the United States (such as Japan and China) have tried to diversify their sources of energy to reduce the risk of disruption. Security was also viewed as threatened by sudden fluctuations in the price of oil, hence the establishment in the United States of the Strategic Petroleum Reserve (SPR): petroleum.

Stocks are maintained by the federal government for use during periods of major supply interruption. The idea is that, if the price of oil were to rise rapidly due to disruption in supply, then the SPR could be called upon to provide supplies, thus reducing the price shock.

Nuclear power was declared dead in the United States because it is too expensive and unacceptably risky. Around the world, nuclear plant ended up achieving less than ten per cent of the new capacity and one per cent of the new orders (all from countries with centrally planned energy systems) forecast in the early 1980. The industry has suffered the greatest collapse of any enterprise in industrial history. Scientists still have not developed reliable ways to handle nuclear wastes and decommissioned plants, which remain dangerously radioactive for far longer than societies last or geological foresight extends.

Strong economic growths across the globe and new global demands for more energy have meant the end of sustained surplus capacity in hydrocarbon fuels and the beginning of capacity limitations. In fact, the world is currently precariously close to utilizing all of its available oil-production capacity, raising the chances of an oil-supply crisis with more substantial consequences than seen since the early 1970. These limits mean that the United States can no longer assume that oil-producing states will provide more oil. Nor is it strategically and politically desirable for the United States to remedy its present tenuous situation by simply increasing its dependence on a few foreign sources. As a result, expanding demand for energy will change US policy towards the Middle East, Russia and China. A recent example is that, in 2005, the state-owned Chinese company CNOOC eventually abandoned its bid for Unocal due to strong political opposition in the United States.

Effects of energy demand

Energy and macroeconomics

In fact, almost every recession since the Second World War in the United States, as well as many other energy-importing nations, has been preceded by a spike in the price of energy (Hamilton, 1983;

Ferderer, 1996; Mork, Mysen and Olsen, 1994). The oil price movement affects certain sectors: oil-dependent manufacturing such as paper and packaging, consumer-related sectors such as autos, refiners' margins, the energy-intensive utility sector, and of course exploration companies and the big oil majors themselves.

Energy, economy and environment

Many important environment damages stem from the production, conversion, and consumption of energy. The costs of these environmental damages generally are not incorporated into prices for energy commodities and resources; this omission leads to overuse of energy. It has been shown that estimates of damage costs resulting from combustion of fossil fuels, if internalized into the price of the resulting output of electricity, could clearly lead to a number of renewable technologies being financially competitive with generation from coal plants. Environmental impacts currently receiving most attention are associated with the release of greenhouse gases in the atmosphere, primarily carbon dioxide, from the combustion of fossil fuels. During combustion, carbon combines with oxygen to produce carbon dioxide, the primary greenhouse gas. Carbon dioxide accumulates in the atmosphere and is expected to result in significant detrimental impacts on the world's climate, including global warming, rises in the ocean levels, increased intensity of tropical storms, and losses in biodiversity. Concern about this issue is common to energy economics, environmental economics, and ecological economics. Cropper and Oates (1992) suggest measuring benefits and costs with a review of cases where benefit–cost analyses have actually been used in the setting of environmental standards. Owen (2004) suggests that penalizing high pollutant-emitting technologies not only creates incentives for 'new' technologies but also encourages the adoption of energy-efficiency measures with existing technologies and consequently lower pollutants per unit of output.

World carbon dioxide emissions are expected to increase by 1.9 per cent annually between 2001 and 2025. Much of this increase is expected to occur in developing countries. The United States produces about 25 per cent of global carbon dioxide emissions from burning fossil fuels, primarily because of it has the largest economy in the world and meets 85 per cent of its energy needs through burning fossil fuels. The United States is projected to lower its carbon intensity by 25 per cent from 2001 to 2025. There are numerous proposals aimed at reducing the carbon dioxide emissions, of which the Kyoto Protocol is a well-known and influential one. During 1–11 December 1997, more than 160 nations met in Kyoto, Japan, to negotiate binding limitations on greenhouse gases for the developed nations, pursuant to the objectives of the Framework Convention on Climate Change of 1992. The outcome of the meeting was the Kyoto Protocol, in which the developed nations agreed to limit their greenhouse gas emissions relative to the levels emitted in 1990. The United States agreed to reduce emissions from 1990 levels by seven per cent during the period 2008 to 2012.

Sickles and Jeon (2004) evaluate the role that undesirable outputs of the economy, such as carbon dioxide and other greenhouse gases, play on the frontier production process. This paper also explores implications for growth of total factor productivity in the OECD and Asian economies.

Natural disasters shock the energy market, too. According to the Minerals Management Service (MMS, 2005), Gulf of Mexico daily oil production was reduced by 89 per cent as a result of Hurricane Katrina in 2005. The MMS also reports that 72 per cent of daily Gulf of Mexico natural gas production was shut

in. In 2004, Hurricane Ivan caused lasting damage to the energy infrastructure in the Gulf of Mexico and interrupted oil supplies to the United States. US Secretary of Energy Spencer Abraham agreed to release 1.7 million barrels of oil in the form of a loan from the Strategic Petroleum Reserve.

A concluding comment

The world runs on energy, primarily energy generated from coal and petroleum. The current war against terrorism and the tensions in the Middle East have raised new questions about the reliability of America's oil supply from that region. Concerns about global climate change have also focused increased attention on the search for cleaner fuels and energy-generating methods. Russia's determination to become a major petroleum supplier, OPEC's periodic moves to restrict oil production and the rising energy needs in China and other developing countries are all important issues forming the future world energy market.

I would like to thank Robert Thomure, Rice University, for his research assistance.

See Also

- environmental economics
- oil and the macroeconomy
- Organization of the Petroleum Exporting Countries (OPEC)

Bibliography

Bohi, D. and Toman, M. 1996. *The Economics of Energy Security*. Boston: Kluwer Academic Publishers.

Boyle, G. 1996. *Renewable Energy: Power for a Sustainable Future*. Oxford: Oxford University Press.

Cooper, J. 2003. Price elasticity of demand for crude oil: estimates for 23 countries. *OPEC Review: Energy Economics & Related Issues* 27(1), 1–8.

Cropper, M. and Oates, W. 1992. Environmental economics: a survey. *Journal of Economic Literature* 30, 675–740.

Denny, M., Fuss, M. and Waverman, L. 1981. Substitution possibilities for energy: evidence from U.S. and Canadian manufacturing. In *Modeling and Measuring Natural Resource Substitution*, ed. E. Berndt and B. Field. Cambridge, MA: MIT Press.

EIA (Energy Information Administration). 2001. *Annual Energy Review 2000*. Washington, DC: EIA.

EIA. 2004. *Annual Energy Review 2004*. Washington, DC: EIA.

EIA. 2005a. *Annual Energy Review 2005*. Washington, DC: EIA.

EIA. 2005b. *Annual Energy Outlook 2005 with Projections to 2025*. Washington, DC: EIA.

Fang, F. and Sickles, R. 2004. The role of environmental factors in growth accounting. *Journal of Applied Econometrics* 19, 567–91.

Ferderer, P. 1996. Oil price volatility and macroeconomy. *Journal of Macroeconomics* 18, 1–26.

Gilbert, R. and Richard, J. 1978. Dominant firm pricing policy in a market for an exhaustible resource. *Bell Journal of Economics* 9, 385–95.

Hamilton, J. 1983. Oil and the macroeconomy since World War II. *Journal of Political Economy* 91, 228–48.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Judson, R., Schmalensee, R. and Stoker, T. 1999. Economic development and the structure of the demand for commercial energy. *Energy Journal* 20(2), 29–7.

Kolstad, C. 1994. Hotelling rents in hotelling space: product differentiation in exhaustible resource markets. *Journal of Environmental Economics and Management* 26, 163–80.

Loury, G. 1990. Tacit collusion in a dynamic duopoly with indivisible production and cumulative capacity constraints. Working Paper No. 557. Department of Economics, MIT.

Medlock, K. and Soligo, R. 2001. Economic development and end-use energy demand. *Energy Journal* 22(2), 77–105.

Medlock, K. and Soligo, R. 2002. Car ownership and economic development with forecasts to 2015. *Journal of Transport Economics and Policy* 36, 163–88.

MMS (Minerals Management Service). 2005. *Hurricane Information*. Washington, DC: MMS.

Mork, K., Mysen, H. and Olsen, O. 1994. Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy Journal* 15(4), 19–35.

Nerlove, M. 1958. Distributed lags and demand analysis for agricultural and other commodities. In *Agriculture Handbook* No. 141. Washington, DC: US Department of Agriculture.

Owen, A. 2004. Environmental externalities, market distortions and the economics of renewable energy

technologies. *Energy Journal* 25(3), 127–56.

Pindyck, R. 1982. Jointly produced exhaustible resources. *Journal of Environmental Economics and Management* 9, 291–303.

Salant, S. 1976. Exhaustible resources and industrial structure – Nash–Cournot approach to the world oil market. *Journal of Political Economy* 84, 1079–93.

Sickles, R.B. and Jeon, B.M. 2004. The role of environmental factors in growth accounting. *Journal of Applied Econometrics* 19, 567–91.

Stiglitz, J. 1976. Monopoly and rate of extraction of exhaustible resources. *American Economic Review* 66, 655–61.

Sweeney, J. 1977. Economics of depletable resources – market forces and intertemporal bias. *Review of Economic Studies* 44, 125–41.

Wiser, R., Olson, S., Bird, L. and Swezey, B. 2004. *Utility Green Pricing Programs: A Statistical Analysis of Program Effectiveness*. Berkeley, CA and Golden, CO: Ernest Orlando Lawrence Berkeley National Laboratory and National Renewable Energy Laboratory. Online. Available at http://www.eere.energy.gov/greenpower/resources/pdfs/lbnl_54437.pdf, accessed 1 August 2006.

How to cite this article

Fang, Ying and Robin Sickles. "energy economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000082> doi:10.1057/9780230226203.0474

The New Palgrave Dictionary of Economics Online

Engel curve

Arthur Lewbel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An Engel curve describes how a consumer's purchases of a good like food varies as the consumer's total resources such as income or total expenditures vary. Engel curves may also depend on demographic variables and other consumer characteristics. A good's Engel curve determines its income elasticity, and hence whether the good is an inferior, normal, or luxury good. Empirical Engel curves are close to linear for some goods, and highly nonlinear for others. Engel curves are used for equivalence scale calculations and related welfare comparisons, and determine properties of demand systems such as aggregability and rank.

Keywords

aggregation; consumers’ expenditure; consumer demand; demand equations; Engel curves; Engel equivalence scales; Engel's Law; law of one price; nonparametric methods; Rothbarth scales; separability; utility theory; Working, H.; Working–Leser model

Article

An Engel curve is the function describing how a consumer's expenditures on some good or service relate to the consumer's total resources, with prices fixed, so $q_i = g_i(y, z)$, where q_i is the quantity consumed of good i, y is income, wealth, or total expenditures on goods and services, and z is a vector of other characteristics of the consumer, such as age and household composition. Usually y is taken to be total expenditures, to separate the problem of allocating total consumption to various goods from the decision of how much to save or dissave out of current income. Engel curves are frequently expressed in the budget share form $w_i = h_i[\log(y), z]$ where w_i is the fraction of y that is spent buying good i. The goods are typically aggregate commodities such as total food, clothing or transportation, consumed over some weeks or months, rather than discrete purchases. Engel curves can be defined as Marshallian demand functions, with the prices of all goods fixed.

The term ‘Engel curve’ is also used to describe the empirical dependence of q_i on y,z in a population of consumers sampled in one time and place. This empirical or statistical Engel curve coincides with the above theoretical Engel curve definition if the law of one price holds (all sampled consumers paying the same prices for all goods), and if all consumers have the same preferences after conditioning on z and possibly on some well-behaved error terms. Since these conditions rarely hold, it is important in practice to distinguish between these two definitions.

Using data from Belgian surveys of working class families, Ernst Engel (1857; 1895) studied how households’ expenditures on food vary with income. He found that food expenditures are an increasing function of income and of family size, but that food budget shares decrease with income. This relationship of food consumption to income, known as Engel's law, has since been found to hold in most economies and time periods, often with the function h_i for food i close to linear in $\log(y)$.

Engel curves can be used to calculate a good's income elasticity, which is roughly the percentage change in q_i that results from a one per cent change in y, or formally $\partial \log g_i(y,z)/\partial \log(y)$. Goods with income elasticities below zero, between zero and 1, and above 1 are called inferior goods, necessities and luxuries respectively, so by these definitions what Engel found is that food is a necessity. Elasticities can themselves vary with income, so a good that is a necessity for the rich can be a luxury for the poor.

Some empirical studies followed Engel (1895), such as Ogburn (1919), but Allen and Bowley (1935) firmly connected their work to utility theory. They estimated linear Engel curves $q_i = a_i + b_i y$ on data-sets from a range of countries, and found that the resulting errors in these models were sometimes quite large, which they interpreted as indicating considerable heterogeneity in tastes across consumers.

Working (1943) proposed the linear budget share specification $w_i = a_i + b_i \log(y)$, which is known as the Working–Leser model, since Leser (1963) found this functional form to fit better than some alternatives. However, Leser obtained still better fits with what would now be called a rank-three model, namely, $w_i = a_i + b_i \log(y) + c_i y^{-1}$, and in a similar, earlier, comparative statistical analysis Prais and Houthakker (1955) found $q_i = a_i + b_i \log(y)$ to fit best. More recent work documents sometimes considerable nonlinearity in Engel curves. Motivated by this nonlinearity, one of the earlier empirical applications of nonparametric regression methods in econometrics was kernel estimation of Engel curves. Examples include Bierens and Pott-Buter (1990), Lewbel (1991), and Härdle and Jerison (1991).

More recent studies that control for complications like measurement error and other covariates z, including Hausman, Newey and Powell (1995) and Banks, Blundell and Lewbel (1997), find Engel curves for some goods are close to Working–Leser, while others display considerable curvature, including quadratics or S shapes. Even Allen and Bowley (1935, p. 123) noted ‘there is a good fit, allowance being made for observation and sampling errors,…, to a linear expenditure relation and occasionally to a parabolic relation’.

Other variables z also help explain cross-section variation in demand. Commonly used covariates include the number, ages and gender of family members, location measures, race and ethnicity, seasonal effects, and labour market status. Variables indicating ownership of a home, a car or other large durables can also have considerable explanatory power, though these are themselves consumption decisions.

Engel's original work showed the relevance of family size, and later studies confirm that larger families typically have larger budget shares of necessities than smaller families at the same income level. Adult equivalence scales model the dependence of utility functions on family size, and use this dependence to

compare welfare across households, assuming that a large family with a high income is as well off as a smaller family with a lower income if both families have demands that are similar in some way, such as equal food budget shares or equal expenditures on adult goods such as alcohol. The ratio of total expenditures needed to equate food budget shares across households are known as Engel equivalence scales, while the ratio that equates expenditures on adult goods are called Rothbarth scales (Rothbarth, 1943).

Shape invariance assumes that budget share Engel curves for one type of consumer, such as a household with children, is a linear transformation of the budget-share Engel curves for other types of consumers, such as households without children. Shape invariance is necessary for constructing what are known as exact or independent of base equivalence scales, and has been found to at least approximately hold in some data-sets. See Lewbel (1989), Blackorby and Donaldson (1991), Gozalo (1997), Pendakur (1999), and Blundell, Browning and Crawford (2003).

The level of aggregation across goods affects Engel curve estimates. Demand for a narrowly defined good like apples varies erratically across consumers and over time, while Engel curves based on broad aggregates like food are affected by variation in the mix of goods purchased. The aggregate necessity food could include inferior goods like cabbage and luxuries like caviar, which may have very different Engel curve shapes.

Other empirical Engel curve complications include unobserved variations in the quality of goods purchased, and violations of the law of one price. When price or quality variation is unobserved, their effects may correlate with, and so be erroneously attributed to, y or z. Examples of such correlations could include the wealthy systematically favouring higher quality goods, and the poor facing higher prices than other consumers because they cannot afford to travel to discount stores.

Assume a consumer (household) h determines demands q_{hi} facing prices p_i for each good i by maximizing a well-behaved utility function over goods (which could depend on z_h), subject to a budget constraint $\sum_i p_i q_{hi} \leq y_h$. This yields Marshallian demand functions $q_{hi} = G_{hi}(p, y_h, z_h)$, with Engel curves given by these functions with the price vector p fixed. Utility functions that yield Engel curves of the form $q_{hi} = b_i(z)\, y_h$ are called homothetic, and $q_{hi} = a_i(z) + b_i(z)\, y_h$ are quasihomothetic. Many theoretical results regarding two-stage budgeting and aggregation across goods require homotheticity or quasihomotheticity, most notably Gorman (1953).

The shape of Engel curves plays an important role in the determination of macroeconomic demand relationships. For example, if we ignore z for now, suppose individual consumers h each have Engel curves of the quasihomothetic form $q_{hi} = a_{hi} + b_i y_h$. Then, letting Q_i and Y be aggregate per capita quantities and total expenditures in the population, we get $Q_i = A_i + b_i Y$ by averaging q_{hi} across consumers h. This is a representative consumer model, in the sense that the distribution of y affects aggregate demand Q_i only through its mean $E(y) = Y$. Gorman (1953) showed that only linear Engel curves have this property, though linear Engel curve aggregation dates back at least to Antonelli (1886). Gorman's linearity requirement, which does not usually hold empirically, can be relaxed given restrictions on the distribution of y; for example, Lewbel (1991) shows that $E(y \log y) / Y - \log(Y)$ is very close to constant in US data, and if it is constant then Working–Leser household Engel curves yield Working–Leser aggregate, representative consumer demands.

Exactly aggregable demands are defined by $q_i = \Sigma_{j=1}^{J} A_{ji}(p) c_j(y, z)$, and so have Engel curves $q_i = \Sigma_{j=1}^{J} a_{ji} c_j(y, z)$ that are linear in the functions $c_j(y,z)$. These models have the property that aggregate demands Q_i depend only on the means of $c_j(y,z)$. Utility theory imposes constraints on the functional forms of $c_j(y,z)$. Properties of exactly aggregable demands and associated Engel curves are derived in Muellbauer (1975), Jorgenson, Lau and Stoker (1982), and Lewbel (1990), but primarily by Gorman (1981), who proved the surprising result that utility maximization forces the matrix of Engel curve coefficients a_{ji} to have rank three or less.

Lewbel (1991) extends Gorman's rank idea to arbitrary demands, not just those in the exactly aggregable class, by defining the rank of a demand system as the dimension of the space spanned by its Engel curves. Engel curve rank can be nonparametrically tested, and has implications for utility function separability, welfare comparisons, and for aggregation across goods and across consumers. Many empirical studies find demands have rank three.

One area of current research concerns the observable implications of collective models, that is, households that determine expenditures based on bargaining among members. For example, the Engel curves of such households could violate Gorman's rank theorem, even if each member had exactly aggregable preferences. Another topic attracting current attention is the role of errors in demand models, particularly their interpretation as unobserved preference heterogeneity, random utility model parameters. This matters in part because another of Allen and Bowley's (1935) findings remains true today, namely, Engel curve and demand function models still fail to explain most of the observed variation in individual consumption behaviour.

See Also

- aggregation (theory)
- consumer expenditure
- demand theory
- Engel, Ernst
- Engel's Law
- equivalence scales
- Gorman, W.M. (Terence)
- utility

Bibliography

Allen, R. and Bowley, A. 1935. *Family Expenditure: A Study of its Variation*. London: P.S. King and Son.

Antonelli, G. 1886. Sulla teoria metematica della economia politica. Pisa: Nella Tipografia del Fochetto. Translated as ‘On the mathematical theory of political economy’, in *Preferences, Utility and Demand*,

ed. J. Chipman et al. New York: Harcourt Brace Jovanovich, 1971.

Banks, J., Blundell, R. and Lewbel, A. 1997. Quadratic Engel curves and consumer demand. *Review of Economics and Statistics* 79, 527–39.

Bierens, H. and Pott-Buter, H. 1990. Specification of household expenditure functions and equivalence scales by nonparametric regression. *Econometric Reviews* 9, 123–210.

Blackorby, C. and Donaldson, D. 1991. Adult-equivalence scales, interpersonal comparisons of well-being, and applied welfare economics. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. Roemer. Cambridge: Cambridge University Press.

Blundell, R., Browning, M. and Crawford, I. 2003. Nonparametric Engel curves and revealed preference. *Econometrica* 71, 205–40.

Engel, E. 1857. Die Productions- und Consumptionsverhaeltnisse des Koenigsreichs Sachsen. *Zeitschrift des Statistischen Bureaus des Koniglich Sachsischen Ministeriums des Inneren, No. 8 und 9*. Reprinted in the Appendix of Engel (1895).

Engel, E. 1895. Die Lebenskosten Belgischer Arbeiter-Familien Fruher and jetzt. *International Statistical Institute Bulletin* 9, 1–74.

Gorman, W. 1953. Community preference fields. *Econometrica* 21, 63–80.

Gorman, W. 1981. Some Engel curves. In *Essays in the Theory and Measurement of Consumer Behaviour in Honor of Sir Richard Stone*, ed. A. Deaton. Cambridge: Cambridge University Press.

Gozalo, P. 1997. Nonparametric bootstrap analysis with applications to demographic effects in demand functions. *Journal of Econometrics* 81, 357–93.

Härdle, W. and Jerison, M. 1991. Cross section Engel curves over time. *Recherches Economiques de Louvain* 57, 391–431.

Hausman, J., Newey, W. and Powell, J. 1995. Nonlinear errors in variables: estimation of some Engel curves. *Journal of Econometrics* 65, 205–53.

Jorgenson, D., Lau, L. and Stoker, T. 1982. The transcendental logarithmic model of aggregate consumer behavior. In *Advances in Econometrics*, ed. R. Basman and G. Rhodes. Greenwich: JAI Press.

Leser, C. 1963. Forms of Engel functions. *Econometrica* 31, 694–703.

Lewbel, A. 1989. Household equivalence scales and welfare comparisons. *Journal of Public Economics* 39, 377–91.

Lewbel, A. 1990. Full rank demand systems. *International Economic Review* 31, 289–300.

Lewbel, A. 1991. The rank of demand systems: theory and nonparametric estimation. *Econometrica* 59, 711–30.

Muellbauer, J. 1975. Aggregation, income distribution, and consumer demand. *Review of Economic Studies* 62, 269–83.

Ogburn, W. 1919. Analysis of the standard of living in the District of Columbia in 1916. *Journal of the American Statistical Association* 16, 374–89.

Pendakur, K. 1999. Estimates and tests of base-independent equivalence scales. *Journal of Econometrics* 88, 1–40.

Prais, S. and Houthakker, H. 1955. *The Analysis of Family Budgets*. Cambridge: Cambridge University Press (2nd edn. 1971).

Rothbarth, E. 1943. Note on a method of determining equivalent income for families of different composition. Appendix 4 in *War-time Pattern of Saving and Spending*, ed. C. Madge. Cambridge: Cambridge University Press.

Working, H. 1943. Statistical laws of family expenditures. *Journal of the American Statistical Association* 38, 43–56.

How to cite this article

Lewbel, Arthur. "Engel curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000085> doi:10.1057/9780230226203.0476

The New Palgrave Dictionary of Economics Online

Engel, Ernst (1821 – 1896)

H.S. Houthakker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agricultural economics; Engel curve; Engel, E.; Engel's Law; household surveys; International Statistical Institute

Article

Born in Dresden, Engel was a German statistician best known for the discovery of the Engel curve and of Engel's Law. In his early years he was associated with the French sociologist Frédéric Le Play, whose interest in the family led him to conduct household surveys. The expenditure data collected in these surveys convinced Engel that there was a relation between a household's income and the allocation of its expenditures between food and other items. This was one of the first functional relations ever established quantitatively in economics. Furthermore, he observed that households with higher incomes tended to spend more on food than poorer households, but that the share of food expenditures in the total budget tended to vary inversely with income. From this empirical regularity he went on to infer that in the course of economic development agriculture would decline relative to other sectors of the economy (Engel, 1857). From 1860 to 1882 Engel was director of the Prussian statistical bureau in Berlin, in which capacity he did much to expand and strengthen official statistics. His resignation resulted from his opposition to Bismarck's protectionist policies. In his own research he dealt particularly with the value of human life (Engel, 1877), which he approached from the cost side. He also investigated the influence of price on demand. His influence on official statistics extended well beyond Germany, and in 1885 he was among the founders of the International Statistical Institute. He died in Radebeul in 1896.

Selected works

1857. Die Productions- und Consumptionsverhaeltnisse des Koenigsreichs Sachsen. Reprinted with Engel (1895), *Anlage* I, 1–54.

1877. *Der Kostenwerth des Menschen*. Berlin.

1895. Die Lebenskosten Belgisher Arbeiter-Familien fruether und jetzt. Reprinted in *International Statistical Institute Bulletin* 9, 1–124.

How to cite this article

Houthakker, H. S. "Engel, Ernst (1821–1896)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000086> doi:10.1057/9780230226203.0475

The New Palgrave Dictionary of Economics Online

Engel's Law

Martin Browning

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agriculture and economic development; equivalence scales; Engel's Law

Article

Engel's law states that food is not a luxury. This is one of the earliest empirical regularities in economics and also one of the most robust. The widespread finding is that regressions of food expenditures, quantities or budget shares on income or total expenditure and other variables such as prices, demographics and regional dummies uniformly imply that the income elasticity of food is less than 1 (and greater than zero). For example, time series from individual countries, cross-sections within countries and cross-country analyses all find the same qualitative empirical finding.

This correlation seems to have been highlighted for a number of reasons. First, food is an important component of household budgets everywhere so that it is intrinsically of interest. Second, the finding suggests that over the long run countries experiencing significant growth will find that agriculture provides an increasingly unimportant part of national income. This argues against balanced growth in long-run development. Third, we do not observe such a consistent pattern for any other wide commodity grouping such as clothing or durables. Finally, the fact that the food budget share is a decreasing function of the material standard of living (if other factors are held constant) suggested at one time that it can be used as an indicator of the latter. In particular, iso-prop (‘same proportion’) methods have been used to compute adult equivalence scales by finding the level of income that would equate the food budget share across different demographic groups. The conditions under which the iso-prop method is valid are very strong – essentially, extra people in the household have to make the household behave as though it is poorer and should not cause any change in the structure of demands above this – and such methods have fallen out of favour (see Deaton and Muellbauer, 1986, for discussion and references).

Despite the venerability of the literature on Engel's law, the inferences that can be drawn from it are limited. For example, the cross-section finding is consistent with all households having a decreasing relationship so that increasing the income of a household will lead to a decrease in the food budget share. On the other hand, the correlation might be completely spurious if it is due to poorer households

having a higher ‘taste’ for food. In this case the apparent dependence is simply due to heterogeneity in tastes, which is correlated with income. The fact that studies using aggregate time series-data find different elasticities from those found in cross-section data from the same country and time period suggests that the empirical finding is a combination of both causes. The paucity of panel data with full expenditure information makes any inference hazardous. Thus Engel's law remains what it has always been: a very robust but unsurprising partial correlation with many alternative interpretations.

See Also

- Engel, Ernst
- Engel curve
- equivalence scales

Bibliography

Deaton, A. and Muellbauer, J. 1986. On measuring child costs: with applications to poor countries. *Journal of Political Economy* 94, 720–44.

How to cite this article

Browning, Martin. "Engel's Law." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000087> doi:10.1057/9780230226203.0477

The New Palgrave Dictionary of Economics Online

Engels, Friedrich (1820 – 1895)

Gareth Stedman Jones
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost of production theory of value; Engels, F.; Kautsky, K.; labour theory of value; Marx, K. H.; Marxian transformation problem; peasants; rent; subjective theory of value

Article

Born in Barmen, the eldest son of a textile manufacturer in Westphalia, Engels was trained for a merchant's profession. From school onwards, however, he developed radical literary ambitions which eventually brought him into contact with the Young Hegelian circle in Berlin in 1841. In 1842, Engels left for England to work in his father's Manchester firm. Already converted by Moses Hess to a belief in ‘communism’ and the imminence of an English social revolution, he used his two-year stay to study the conditions which would bring it about. From this visit came two works which were to make an important contribution to the formation of Marxian socialism: *Outlines of a Critique of Political Economy* (generally called the *Umrisse*) published in 1844, and *The Condition of the Working Class in England*, published in Leipzig in 1845.

Returning home via Paris in 1844, Engels had his first serious meeting with Marx. Their lifelong collaboration dated from this point with an agreement to produce a joint work (*The Holy Family*), setting out their positions against other tendencies within Young Hegelianism. This was followed by a second unfinished joint enterprise (*The German Ideology*, 1845–6), where their materialist conception of history was expounded systematically for the first time.

Between 1845 and 1848, Engels was engaged in political work among German communist groups in Paris and Brussels. In the 1848 revolution itself, he took a full part, first as a collaborator of Marx on the *Neue Rheinische Zeitung* and subsequently in the last phase of armed resistance to counter-revolution in the summer of 1849.

In 1850, Engels returned once more to Manchester to work for his father's firm and remained there until he retired in 1870. During this period, in addition to numerous journalistic contributions, including attempts to publicize Marx's *Critique of Political Economy* (1859) and *Capital*, Volume 1 (1867, second edition 1873), he first developed his interest in the relationship between historical materialism and the

natural sciences. These writings were posthumously published as *The Dialectics of Nature* (1925). In 1870 Engels moved to London.

As Marx's health declined, Engels took over most of his political work in the last years of the First International (1864–72) and took increasing responsibility for corresponding with the newly founded German Social Democratic Party and other infant socialist parties. Engels's most important work during this period was his polemic against the positivist German socialist, Eugen Dühring. The *Anti-Dühring* (1877) was the first comprehensive exposition of a Marxian socialism in the realms of philosophy, history and political economy. The success of this work, and in particular of extracts from it like *Socialism, Utopian and Scientific*, represented the decisive turning point in the international diffusion of Marxism and shaped its understanding as a theory in the period before 1914.

In his last years after Marx's death in 1883, Engels devoted most of his time to the editing and publishing of the remaining volumes of *Capital* from Marx's manuscripts. Volume 2 appeared in 1885, Volume 3 in 1894, a year before his death. Engels had also hoped to prepare the final volume dealing with the history of political economy. But the difficulty of deciphering Marx's handwriting, his own failing eyesight and the formidable editorial problems encountered in constructing Volumes 2 and 3, induced him to hand over this task to Karl Kautsky, who subsequently published it under the title *Theories of Surplus Value*.

Engels's work was of importance, both in the construction and interpretation of Marxian economic theory and in the laying down of important guidelines in the subsequent development of Marxist economic policy.

In the realm of theory, his contribution is of particular significance in three respects.

First, and of real importance in the formation of a distinctively Marxian stance towards political economy was Engels's *Outlines of a Critique of Political Economy* (the *Umrisse*), published in 1844. In 1859 in his own *Critique of Political Economy*, Marx acknowledged this sketch as ‘brilliant’, and its impact is discernible in Marx's 1844 writings. The *Umrisse* represented the first systematic confrontation between the ‘communist’ strand of Young Hegelianism and political economy. The communist aspiration was expressed in Feuerbachian language, while the mode of analysis was Hegelian. But, as has recently been demonstrated (Claeys, 1984), the content of Engels's critique was first and foremost a product of his early stay in Manchester. For, apart from some indebtedness to Proudhon's *What is Property?* (1841), the main source of Engels's essay was John Watts, *The Facts and Fictions of Political Economy* (1842), a resumé of the Owenite case against the propositions of political economy. At this stage, Engels's own acquaintance with the work of political economists seems to have been mainly at second-hand.

The *Umrisse* was an attempt to demonstrate that all the categories of political economy presupposed competition which in turn presupposed private property. He began with an analysis of value, which juxtaposed a ‘subjective’ conception of value as utility ascribed to Say with an ‘objective’ conception as cost of production attributed to Ricardo and McCulloch. Reconciling these two definitions in Hegelian fashion, Engels defined value as the relation of production costs to utility. This was the equitable basis of exchange, but one impossible to implement on the basis of competition which was responsive to market demand rather than social need. (Engels still adhered to this definition of value 30 years later in the *Anti-Dühring*. Discussing the disappearance of the ‘law of value’ with the end of commodity production, he wrote:

> As long ago as 1844, I stated that the above mentioned balancing of useful effects and expenditure of labour would be all that would be left, in a communist society, of the concept of value as it appears in political economy•…•The scientific justification for this statement, however,… was only made possible by Marx's *Capital*. (Engels, 1877, pp. 367–8)

This shows how much greater continuity of thought there was between the young and the old Engels than is normally imagined.)

He next analysed rent, counterposing a Ricardian notion of differential productivity to one attributed to Smith and T.P. Thompson based upon competition. Interestingly, in this analysis Engels differed both from Watts and Proudhon, in denying the radical form of the labour theory – the right to the whole product of labour – both by citing the case of the need to support children and in querying the possibility of calculating the share of labour in the product.

Finally, after an attack on the Malthusian population theory, which closely followed Alison and Watts, Engels attacked competition itself, both because it provided no mechanism of reconciling general and individual interest, and because it was argued to be self-contradictory. Competition based on self-interest bred monopoly. Competition as an immanent law of private property led to polarization and the centralization of property. Thus private property under competition is self-consuming.

What particularly impressed Marx was the argument that all the categories of political economy were tied to the assumption of competition based on private property. This, for him, represented an important advance over Proudhon whose notion of equal wage would lead to a society conceived as ‘abstract capitalist’ and whose conception of labour right presupposed private property. Proudhon had not seen that labour was the essence of private property. His critique was of ‘political economy from the standpoint of political economy’. He had not ‘considered the further creations of private property, e.g. wages, trade, value, price, money etc. as forms of private property in themselves’ (Engels and Marx, 1844b). The *Umrisse* suggested a new means of underpinning the Marxian ambition to transcend the categorical world of political economy and private property altogether. Moreover, by representing competition as a law which would produce its opposite, monopoly, the elimination of private property and revolution, Engels preceded Marx in positing the ‘free trade system’ as a process moving towards self-destruction through the operation of laws immanent within it.

These conclusions were amplified in Engels's other major work of this period, *The Condition of the Working Class in England*. Here, the law of competition by engendering ‘the industrial revolution’ had created a revolutionary new force, the working class. The single thread underlying the development of the working class movement had been the attempt to overcome competition. Such an analysis prefigured the famous statement in the *Communist Manifesto* that the capitalists were begetting their own gravediggers (Stedman Jones, 1977).

Between the mid-1840s and the mid-1870s, Engels played no discernible part in the elaboration of *Capital* beyond supplying Marx with practical business information. His vital contributions to the prehistory of the theory were forgotten and it was only in his better-known role as interpreter and publicist of Marx's work that his writings received widespread attention. During the Second International period, these writings attained almost canonical status, but in the 20th century they generally provided a polemical target for all those attempting to re-theorize Marx in the light of the publication of his early writings.

In the realm of political economy more narrowly conceived, Engels helped to set up the ‘transformation’ debate by his dramatization of Marx's switch from value to production price in his introductions to Volumes 2 and 3 of *Capital*. Engels's own contribution to this debate in his last published article in *Neue Zeit* in 1895 (now published as ‘Supplement and Addendum’ to Volume 3 of *Capital*) was to argue that the shift from value to production price was not merely a logical development entailed by the enlargement of the scope of investigation to include circulation and the ‘process of capitalist production as a whole’, but also reflected a real historical transition from the stage of simple commodity production to that of capitalism proper. ‘The Marxian law of value has a universal economic validity for an era lasting from the beginning of the exchange that transforms products into commodities down to the fifteenth century of our epoch’ (Marx, 1894, p. 1037).

Leaving aside the empirical question whether during the pre-capitalist era commodities were exchanged in accordance with the amount of labour embodied in them, commentators as diverse as Bernstein and Rubin have objected that this makes no sense in terms of Marx's theory, since during this epoch there existed ‘no mechanism of the general equalisation of different individual labour expenditures in separate economic units on the market’ and that consequently it was not appropriate to speak of ‘abstract and socially necessary labour which is the basis of the theory of value’ (Rubin, 1928, p. 254). They have further objected, appealing to Marx's 1857 ‘Introduction to the Critique of Political Economy’, that there is no necessary connection between the logical and historical sequence of concepts, and that the order of appearance of concepts in *Capital* is determined simply by the logical place they occupy in an exposition of the theory of the capitalist mode of production.

Engels could certainly claim explicit textual support from Volume 3 for his historical interpretation of value (‘It is also quite apposite to view the value of commodities not only as theoretically prior to the prices of production, but also as historically prior to them. This applies to those conditions in which the means of production belong to the worker…’; Marx, 1894, p. 277). It should also be stressed that there was nothing new in Engels's representation of the character of Marx's theory. Back in 1859, in a review of Marx's *Critique of Political Economy*, Engels stated, ‘Marx was, and is, the only one who could undertake the work of extracting from the Hegelian Logic the kernel which comprised Hegel's real discoveries… and to construct the dialectical method divested of its idealistic trappings’; and in characterizing that method as a form of identity between logical and historical progression, he continued, ‘the chain of thought must begin with the same thing that this history begins with, and its further course will be nothing but the mirror image of the historical course in abstract and theoretically consistent form…’ (Engels, 1859). It is implausible to suppose that Marx at this time should have sanctioned a fundamental distortion of his method and it is suggestive that he himself, describing his relationship to Hegel, should have endorsed the metaphor of discovering ‘the rational kernel in the mystical shell’ in his 1873 Postface to the second edition of *Capital* , Volume 1 (Marx, 1873, p. 103).

Perhaps the real difficulty lies not in Engels but in Marx himself. It may be, as Louis Althusser has claimed, that Marx did not find a suitable language in which to characterize the distinctiveness of his approach, or it may be more simply that Marx remained ambivalent about how to characterize the theory. In any event, it is not difficult to establish disjunctions between the way he proceeds and the descriptions he gives of his procedures. Engels stuck fairly closely to Marx's descriptions of his procedures and can hardly be reproached for taking Marx at his word.

The problem of Engels's role as an interpreter of Marx's theory debouches onto a third and potentially yet more contentious aspect of Engels's legacy, his role as editor of *Capital*, Volumes 2 and 3. Engels's

work was not confined to the transcription of Marx's illegible handwriting. He had to make active editorial choices. The published versions of these volumes contain over 1,300 pages, but the original manuscripts amount to almost twice as many. For Volume 2, for instance, Marx had composed eight versions of his treatment of the process of circulation, from which Engels made a collation. In the absence of an independent transcription and publication of the manuscripts, from which Engels worked, it is impossible to assess whether the emphasis and meaning of the published volumes differ in any significant way from the original. What seems clear, is that in his cautious desire to reproduce as much of the original material as possible, Engels produced a much bulkier and more repetitive version than Marx originally intended. Marx, it seems, always hoped that *Capital* should consist of two volumes and a further volume on the history of political economy (Rubel, 1968; Levine, 1984). From a detailed comparison of Volume 2, Part 1, with the original manuscripts, it appears that Engels also occasionally committed inaccuracies in the citation of the manuscripts he had used (Levine, 1984). Much more doubtful, given all we know of Engels's caution as an editor, is the further suggestion that Engels's editing procedures may have shifted the meaning of the text in ways that lent support to a ‘collapse theory’ of capitalism (*Zusammenbruchstheorie*) (Levine, 1984). Apart from the smallness of the sample and Engels's own reservations about such a theory, the fact is that proponents of such a position already had sufficient ammunition from *Capital*, Volume 1. Moreover, it simply begs the question whether Marx's attitude to the collapse of capitalism was any more or less apocalyptic than that of Engels.

This discussion by no means exhausts Engels's importance in the history of economic theory or policy. A fuller treatment would have to discuss his analysis of the ‘peasant question’ which included the important prescription that collectivization must be by example rather than force, his definition of political economy in the *Anti-Dühring*, his interpolations in *Capital*, Volume 3, on banks, the stock exchange and cartels which set the agenda for the early 20th-century discussion of finance capital, his various writings on the relationship between the state and economic forces and his later surveys of English developments since 1844 which prepared the way for later Marxist theories of labour aristocracy. These are only some of the more salient examples.

Finally, at a time when it seems that the technical debate on value seems to have reached a moment of exhaustion, it is perhaps worth going back to Engels if only to remind us of the anti-economic purpose underlying Marx's attempt to construct a theory of value in the first place.

Selected works

1844a. *Outlines of a Critique of Political Economy*. In Karl Marx and Frederick Engels, *Collected Works* [MECW], vol. 3. London: Lawrence & Wishart, 1975.

1844b. (With K. Marx.) *The Holy Family*. In MECW, vol. 4.

1845. *The Condition of the Working Class in England*. In MECW, vol. 4.

1845–6. (With K. Marx.) *The German Ideology*. London: Lawrence & Wishart, 1987.

1859. Karl Marx, ‘A Contribution to the Critique of Political Economy’. *Das Volk*, Nos. 14 and 16, 6

and 20 August.

1877. *Anti-Dühring*. Moscow: Foreign Languages Publishing House, 1954.

1894. *The Peasant Question in France and Germany*. In Karl Marx and Frederick Engels, *Selected Works*, vol. 3, Moscow: Progress Publishers, 1970.

1925. *The Dialectics of Nature*. Moscow: Foreign Languages Publishing House, 1954.

1938. *Engels on Capital*. London: Lawrence & Wishart.

Bibliography

Claeys, G. 1984. Engels' *Outlines of a Critique of Political Economy* (1843) and the origins of the Marxist critique of capitalism. *History of Political Economy* 16, 207–32.

Levine, N. 1984. *Dialogue within Dialectics*. London: Allen & Unwin.

Marx, K. 1859. *A Contribution to the Critique of Political Economy*. In MECW, vol. 16.

Marx, K. 1873. *Capital*, vol. 1, 2nd edn. Harmondsworth: Penguin, 1976.

Marx, K. 1894. *Capital*, vol. 3. Harmondsworth: Penguin, 1981.

Rubel, M. ed. 1968. *Karl Marx, Oeuvres*, vol. 2. Paris: Gallimard.

Rubin, I. 1928. *Essays on Marx's Theory of Value*. Detroit: Black & Red, 1972.

Stedman Jones, G. 1977. Engels and the history of Marxism. In *The History of Marxism*, ed. E.J. Hobsbawm. Hassocks: Harvester, 1983.

Bibliographic addendum

T. Carver, *Friedrich Engels: His Life and Thought*, London, Palgrave Macmillan, 1990 and J.D. Hunley, *The Life and Thought of Friedrich Engels: A Reinterpretation of His Life and Thought*, New Haven: Yale University Press, 1991, are useful in understanding Engels as an original thinker in his own right.

How to cite this article

Jones, Gareth Stedman. "Engels, Friedrich (1820–1895)." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000088> doi:10.1057/9780230226203.0478

The New Palgrave Dictionary of Economics Online

Engle, Robert F. (born 1942)

Tim Bollerslev

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robert Engle has published widely on topics ranging from urban economics to band spectrum regression, electricity demand, state-space modelling, testing, exogeneity, seasonality, option pricing, and market microstructure finance. Most notable, however, are his seminal contributions on cointegration and AutoRegressive Conditional Heteroskedasticity (ARCH), which have revolutionized the field of time series econometrics and the practice of empirical macroeconomics and asset pricing finance, respectively. The research field of financial econometrics and corresponding developments in practical risk management and measurement also derive largely from the insights afforded by the ARCH class of models and Engle's many other research contributions since the 1980s.

Keywords

ARCH models; asset pricing finance; cointegration; Engle, R.; financial econometrics; Fischer, F.; Friedman, M.; Granger Representation Theorem; Phillips curve; risk management; risk measurement; Rothenberg, J.; Solow, R.; time series econometrics

Article

Robert F. Engle was born in Syracuse, upstate New York, on 10 November 1942. Shortly thereafter his family moved to Philadelphia, and Engle graduated from high school there in 1960. He majored in physics as an undergraduate at Williams College, and went on to enrol as a Ph.D. student in physics at Cornell University. However, after one year he decided to switch to the Ph.D. programme in economics, where he wrote his thesis on temporal aggregation and the relationship between macroeconomic models estimated at different frequencies, under the direction of T.C. Liu. After graduating from Cornell in 1969, Engle was hired as an assistant professor at MIT. He moved on to University College at San Diego (UCSD) in 1975, where he was promoted to full professor in 1977 and a Chancellors' Associates Chair in 1993. He also chaired the UCSD Economics Department from 1990 to 1994. In 2000 his

growing interest in financial markets prompted him to accept the Michael Armellino Professorship in Finance at the Stern School of Business at New York University, and he now lives on Manhattan with his wife of many years, Marianne, for most of the year. Together they have two grown children.

Engle has written and published extensively on a wide array of topics, ranging from urban economics to band spectrum regression, electricity demand, state-space modelling, testing, exogeneity, seasonality, option pricing, and market microstructure finance. However, he is particularly well-known for his contributions to time series econometrics and his path-breaking work on cointegration and AutoRegressive Conditional Heteroskedasticity (ARCH). The 2003 Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel was explicitly awarded to Engle for ‘methods of analyzing economic time series with time-varying volatility (ARCH)’, a prize he shared with Clive W. J. Granger for his seminal contributions to the theory of cointegration. It is hardly an exaggeration to say that since the 1980s the concepts of cointegration and ARCH have completely revolutionized the field of time series econometrics and the practice of empirical macroeconomics and asset pricing finance, respectively. The blossoming new research field of financial econometrics and corresponding developments in practical risk management and measurement may also in large part be attributed to the insights afforded by the ARCH class of models and some of Engle's many other pioneering research contributions.

Encouraged by his senior colleagues Franklin M. Fisher, Robert Solow and Jerome Rothenberg, much of Engle's work as an assistant professor at MIT was in the area of urban economics. In fact, Engle was hired by UCSD as an urban economist, and he continued to teach, and occasionally publish in, urban economics almost up until he left San Diego in 2000. It was Clive Granger, whom Engle had first met at the 1970 World Congress of the Econometric Society in Cambridge, who persuaded Engle to move to the West Coast. Granger had himself just accepted a permanent position at UCSD in 1974 and, only a few years after Engle's arrival in 1975, Halbert White also joined the department. The ensuing two decades may rightfully be referred to as the golden age of modern time series econometrics, and UCSD, along with Yale, home of the group led by Peter Phillips, was *the* place to be. The list of visitors to the UCSD Economics Department over this period reads like a who's who in time series econometrics. Engle's hospitality and generosity with his time, as well as the many successful conferences he organized in San Diego, played a crucial role in fostering this nexus. The group was further strengthened by the arrival of James Hamilton, Graham Elliott and Allan Timmermann as additional faculty members in the early 1990s, and the Engle–Granger UCSD econometrics tradition continues to this day. Many of Engle's former Ph.D. students from that period have also gone on to successful academic careers, continuing the UCSD legacy.

Albert Einstein's famous maxim ‘Everything should be made as simple as possible, but not simpler’ succinctly characterizes Engle's approach to econometric modelling. Consider his early research on band spectrum regression. The static OLS regression approach routinely employed throughout economics implicitly assumes that the identical linear relationship holds across all frequencies. Yet in many situations this is obviously a gross oversimplification. For instance, the relation between interest rates and housing starts arguably differs between the short run and the long run. Similarly, the Phillips-curve trade-off between unemployment and inflation may be primarily a business cycle phenomenon. Rather than building a fully fledged complicated dynamic model for analysing these types of temporal dependencies, the band spectrum regression approach offers a simple way of estimating separate regression coefficients, and therefore different relationships, for different frequencies. The idea of

estimating different short-run and long-run regressions may also be seen as a precursor to Engle's later work on cointegration and error correction models.

The original idea of cointegration came from Granger. Nonetheless, it was the seminal joint paper by Engle and Granger (1987a) that devised the first empirical test for cointegration and formally established the link between cointegration and the error-correction type models popularized by Denis Sargan and David Hendry at the LSE during the 1960s and 1970s. More specifically, suppose that the two univariate time series y_t and x_t are both non-stationary, or I(1), so that their first differences, $\Delta y_t \equiv y_t - y_{t-1}$ and $\Delta x_t \equiv x_t - x_{t-1}$, are stationary, or I(0). Most nominal macroeconomic and financial time series may be characterized in this way. Any linear combination of the two series, say $z_t = y_t - \beta x_t$, will then generally also be non-stationary. However, it is possible that z_t may actually be stationary, or I(0), in which case y_t and x_t are said to be cointegrated, with cointegrating vector $(1, -\beta)$. Indeed, many of the 'classical ratios' in macroeconomics and finance (such as consumption/income and dividends/prices) are naturally thought of as cointegrating relationships when expressed in logs. Engle and Granger showed that in this situation a satisfactory vector autoregression for the stationary bivariate process of first differences, $\{\Delta y_t, \Delta x_t\}$, must necessarily include the z_t 'error-correction' term in at least one of the two equations, the so-called Granger Representation Theorem. Intuitively, while both y_t and x_t are stochastically trending, they trend together, so that in the long run they do not stray too far apart. The inclusion of the stationary z_t term as an additional explanatory variable ensures this condition. On the other hand, if the two variables are not cointegrated z_t will be non-stationary, resulting in an unbalanced regression. Hence, empirically the null hypothesis of no cointegration may be assessed on the basis of the popular Engle–Granger cointegration test for a unit root in z_t or, if β is not known, a least-squares estimate thereof. The cointegration concept has had a profound impact on practical macroeconomic time series modelling in government and private institutions around the world. The academic literature also abounds with hundreds, if not thousands, of papers expanding upon the basic testing and modelling approach first developed by Engle and Granger. Engle's subsequent work on common features may also be seen as a natural extension of the cointegration concept.

Another more technical theme brought to the fore by Engle's research entails the powerful use of one-step-ahead prediction error decompositions and conditional Gaussian likelihoods. For instance, the beauty of his influential work on testing, including the simple-to-implement Lagrange Multiplier (LM) chi-square type test statistics constructed by multiplying the number of time series observations with the R^2 from an auxiliary regression of either unity on the vector of scores evaluated under the null hypothesis, or, alternatively, a regression of the squared residuals on the derivatives of the conditional mean, hinges directly on recursively expressing the likelihood function in terms of conditional one-step-ahead densities. Engle's pioneering contributions on dynamic factor models and Kalman filtering are similarly based on the powerful idea of representing the likelihood function in terms of successive conditional densities. Most important, however, the seminal ARCH class of models is also formulated directly in terms of one-step-ahead conditional expectations and densities.

The ARCH model (aptly named so by David Hendry) was conceived during Engle's sabbatical visit to the LSE in 1979. Engle's interest in modelling variance dynamics was spurred by the assertion in Milton Friedman's 1976 Nobel Lecture on a trade-off between unemployment and inflationary uncertainty rather than a trade-off between unemployment and the level of inflation as stipulated by the conventional

Phillips curve. The actual formulation of the first ARCH model was also influenced by Granger's ongoing work on bilinear models. At the time Granger had noted that in a non-Gaussian setting white noise series need not necessarily be unpredictable, and, in particular, when the squared residuals from otherwise well-specified linear models were regressed on their own lagged squared values, the regression coefficients often turned out to be highly significant. Engle realized that this was not actually a test for bilinearity but rather the optimal LM test for some other nonlinear model. Putting this together, Engle brought forth the ARCH model.

The particular ARCH(p) model first analysed and estimated by Engle (1982) may be succinctly expressed as

$$y_t = m_t + \varepsilon_t,\ \varepsilon_t | I_{t-1} \sim N(0, h_t), \text{and} \qquad h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_p \varepsilon_{t-p}^2,$$

where I_{t-1} refers to the set of information available at time t–1, m_t denotes the conditional mean of the y_t time series, and all of the $\alpha_0, \ldots, \alpha_p$ parameters are restricted to be non-negative. The first equation for the conditional mean is, of course, completely standard (in his original application to UK consumer prices Engle used an error correction model for the mean). However, the key difference – Engle's brilliant new insight – comes from recognizing that even though the residuals, ε_t, must be serially uncorrelated, their conditional variance, and therefore the conditional variance of y_t, need not be constant but may in fact be predictable. Moreover, by explicit parameterizing h_t as a function of the past squared residuals and by assuming conditional normality, the joint density for all of the observations, say y_t, t=1, 2, …, T, may easily be evaluated through a prediction error decomposition type argument, and the log likelihood function maximized with respect to all of the model parameters, in turn resulting in a time series of positively serially correlated conditional variance estimates, $\hat{h}_t$, t=1, 2, …, T (that is, estimates of inflationary uncertainty in Engle's original application).

While Engle's initial work and empirical applications of the ARCH model were rooted in macroeconomics, the model has shone most brightly in the area of finance. Since Mandelbrot's work in the early 1960s on the behaviour of speculative prices, it had been recognized that, even though most returns are approximately serially uncorrelated (at least over shorter daily or weekly horizons), ‘large changes tend to be followed by large changes – of either sign – and small changes tend to be followed by small changes’ (Mandelbrot, 1963). However, the empirical finance literature up until the mid-1980s had largely ignored this fact, focusing instead on best characterizing the unconditional return distributions. Meanwhile, Engle soon realized that the ARCH model was ideally suited to this type of data: little, or no, serial correlation in the mean, but strong serial correlation in the second moments. Moreover, the ability to directly quantify the risk through a parametric model for the conditional variance, or more generally the conditional covariance matrix, for the returns strikes directly at the heart of the risk-return trade-off central to asset pricing finance. Consequently, Engle quickly shifted the focus of his research agenda to finance. Over the next 20 years, along with his many students and other collaborators, he developed numerous refinements to the basic ARCH model described above designed

to account better for specific features of the data and/or questions of economic import: richer ARMA-type representations for the variance, including unit-root and long-memory type dependencies, models in which the variance directly influences the conditional mean, asymmetries or leverage effects in the variance, alternative parametric and non-parametric conditional distributions in place of the normal, multivariate factor models and cointegration in variance, to mention but a few. The corresponding long list of new acronyms is also legendary: ARCH-M, GARCH, IGARCH, EGARCH, TARCH, GJR-GARCH, NARCH, QARCH, STARCH, VGARCH, SWARCH, FIGARCH – the list goes on. Empirical applications of these models have in turn resulted in many important new insights into the pricing and hedging of financial instruments and functioning of financial markets, and it is no exaggeration to say that the day-to-day risk management and monitoring in financial institutions have been completely altered by the advent of the ARCH class of models.

Not one to rest on his laurels, Engle continues to push forward the research frontier in financial econometrics. Most recently he has worked extensively on new methods for analysing ultra high-frequency, or tick-by-tick, financial data. In particular, whereas most procedures in time series econometrics, including most of Engle's own earlier work, are explicitly designed for modelling discretely sampled equidistant observations, high-frequency financial data are typically not observed at fixed time intervals. Engle's recent Autoregressive Conditional Duration (ACD) model, which derives many of its statistical properties from the ARCH class of models, provides a particularly convenient way of accommodating this feature by explicitly modelling the times between observations as a serially correlated process. His Dynamic Conditional Correlation (DCC) model, which allows for the estimation of large-scale dynamic covariance matrices, represents another recent noteworthy advance. In keeping with his trademark, this latest research represents the perfect blend between sophisticated yet simple-to-implement econometric techniques explicitly designed for answering genuinely interesting economic questions. Like most of his research since the 1970s, his latest work has already found widespread use both inside and outside academia, and spurred a number of ongoing new developments by other researchers in the field.

In addition to the much-deserved recognition bestowed on him by the Nobel Prize Committee, Engle is a long-standing fellow of the Econometric Society, of the American Statistical Association, and of the American Academy of Arts and Sciences. He is also an excellent speaker, and he has a long list of invited talks and keynote addresses to his name, including the prestigious A. W. Philips and Fisher-Schultz lectures sponsored by the Econometric Society. (For a more in-depth discussion of Engle's work along with some personal reflections, see Diebold, 2004; 2003.)

In conclusion, it is simply impossible to imagine what the field of time series econometrics, let alone the new field of financial econometrics, would have looked like today had it not been for Engle's seminal contributions, both direct and indirect, through the substantial subsequent research programmes his work has helped stimulate. But Engle isn't merely one of the greatest econometricians of his time. He has a wide range of other interests and talents. For example, he is an outstanding ice skater, having competed at the US national level, finishing second in the 1996 and 1999 ice dancing championship competition.

See Also

- ARCH models

- cointegration
- econometrics
- extremal quantiles and value-at-risk
- forecasting
- Granger, Clive W. J.
- measurement error models
- risk

Selected works

1972. (With F. Fisher, J. Harris and J. Rothenberg.) An econometric simulation model of intra-metropolitan housing location: housing, business, transportation and local government. *American Economic Review* 62, 87–97.

1973. Band spectrum regression. *International Economic Review* 15, 1–11.

1976. Interpreting spectral analysis in terms of time domain models. *Annals of Economic and Social Measurement* 5, 89–109.

1979. (With C. Granger, R. Ramanathan and A. Andersen.) Residential load curves and time-of-day pricing: an econometric analysis. *Journal of Econometrics* 9, 13–32.

1981. (With M. Watson.) A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association* 76, 774–81.

1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* 50, 987–1008.

1983a. (With D. Hendry and J. Richard.) Exogeneity. *Econometrica* 51, 277–304.

1983b. (With M. Watson). Alternative algorithms for the estimation of dynamic factor, MIMIC, and varying coefficient regression models. *Journal of Econometrics* 23, 385–400.

1984. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, vol. 3, ed. Z. Griliches and M. Intrilligator. Amsterdam: North-Holland.

1986a. (With C. Granger, J. Rice and A. Weiss.) Semi-parametric estimation of the relation between weather and electricity demand. *Journal of the American Statistical Association* 81, 310–20.

1986b. (With T. Bollerslev.) Modeling the persistence in conditional variances. *Econometric Reviews* 5, 1–50.

1987a. (With C. Granger.) Co-integration and error correction: representation estimation and testing. *Econometrica* 55, 251–76.

1987b. (With D. Lilien and R. Robins.) Estimation of time varying risk premia in the term structure: the ARCH-M model. *Econometrica* 55, 391–407.

1987c. (With S. Yoo.) Forecasting and testing in co-integrated systems. *Journal of Econometrics* 35, 143–59.

1988. (With T. Bollerslev and J. Wooldridge.) A capital asset pricing model with time varying covariances. *Journal of Political Economy* 96, 116–31.

1990a. (With T. Ito and W. Lin.) Meteor showers or heat waves? Heteroskedastic intra-daily volatility in the foreign exchange market. *Econometrica* 58, 525–42.

1990b. Asset pricing with a factor ARCH covariance structure: empirical estimates for Treasury bills (with V. Ng and M. Rothschild). *Journal of Econometrics* 45, 213–37.

1990c. Seasonal integration and cointegration (with S. Hylleberg, C.W.J. Granger and B.S. Yoo). *Journal of Econometrics* 40, 45–62.

1991a. (With C. Granger.) *Long Run Economic Relations: Readings in Cointegration*. Oxford: Oxford University Press.

1991b. (With G. Gonzales.) Semi-parametric ARCH models. *Journal of Business and Economic Statistics* 9, 345–59.

1992. (With C. Mustafa.) Implied ARCH models from options prices. *Journal of Econometrics* 52, 289–311.

1993a. (With T. Bollerslev.) Common persistence in conditional variances. *Econometrica* 61, 167–86.

1993b. (With S. Kozicki). Testing for common features. *Journal of Business and Economic Statistics* 11, 369–80.

1994a. *Handbook of Econometrics*, vol. 4, ed. with D. McFadden. Amsterdam: North-Holland.

1994b. (With T. Bollerslev and D. Nelson.) ARCH models. In *Handbook of Econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

1995a. *ARCH: Selected Readings*. Oxford: Oxford University Press.

1995b. (With K. Kroner.) Multivariate simultaneous ARCH. *Econometric Theory* 11, 122–50.

1998. (With J. Russell). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* 66, 1127–62.

1999. (With G. Lee). A permanent and transitory component model of stock return volatility. In *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger*, ed. R. Engle and H. White. Oxford: Oxford University Press, 475–97.

2000. The econometrics of ultra high frequency data. *Econometrica* 68, 1–22.

2002a. Dynamic conditional correlation: a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics* 20, 339–50.

2002b. (With J. Rosenberg.) Empirical pricing kernels. *Journal of Financial Economics* 64, 341–72.

2004. Risk and volatility: econometric models and financial practice. *American Economic Review* 94, 405–20.

Bibliography

Diebold, F. 2003. The ET interview: Professor Robert F. Engle. *Econometric Theory* 19, 1159–93.

Diebold, F. 2004. The Nobel memorial for Robert F. Engle. *Scandinavian Journal of Economics* 106, 165–85.

Mandelbrot, B. 1963. The variation in certain speculative prices. *Journal of Business* 36, 394–419.

How to cite this article

Bollerslev, Tim. "Engle, Robert F. (born 1942)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000210> doi:10.1057/9780230226203.0479

The New Palgrave Dictionary of Economics Online

English School of political economy

A.M.C. Waterman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The ‘English School’ of political economy describes the tradition of economic thought that began with Malthus, and included Henry Thornton, Chalmers, James Mill, Torrens, West, Ricardo, and Thomas Tooke in the first generation; Whately, Senior, McCulloch and J.S. Mill in the second; and the Fawcetts, Cairnes, Jevons, Bagehot, Foxwell, Sidgwick, J.N. Keynes and Nicholson in the third. J.-B. Say was an honorary member. Karl Marx identified his own work with that school. Its most important production was J.S. Mill's *Principles of Political Economy*, which continued to be used as a textbook until the mid-20th century.

Keywords

American Economic Association; Anderson, J.; Bagehot, W.; Bentham, J.; Böhm-Bawerk, E. von; Boisguilbert, P. de; Brougham, H.; Bullion Committee; Cairnes, J. E.; Cassel, G.; Chadwick, E.; Chalmers, T.; Christian Political Economy; Clark, J. B.; classical economics; Cliffe Leslie, T. E.; comparative advantage; Condillac, E. B. de; diminishing returns; Edgeworth, F. Y.; English School of political economy; factor substitution; Fawcett, H.; Fawcett, M. G.; Fisher, I.; Foxwell, H. S.; free trade; Harrod, R. F.; Horner, F.; Hume, D.; increasing returns to scale; Jeffrey, F.; Jevons, W. S.; Jones, R.; Keynes, J. N.; labour theory of value; Lauderdale, Eighth Earl of; Lloyd, W. F.; Longfield, M.; Malthus, T. R.; Mandeville, B.; marginal revolution; marginal utility; Marshall, A.; Marx, K. H.; mathematics and economics; Mayerne-Turquet, L. de; McCulloch, J. R.; Menger, C.; mercantilism; Mill, J.; Mill, J. S.; Montchrétien, A. de; natural price; neutrality of money; Nicholson, J. S.; Norman, G. W.; Pareto, V.; Pigou, A. C.; political economy; Political Economy Club; Poor Law; productive and unproductive labour; Quesnay, F.; Ricardo, D.; Robbins, L. C.; Say, J.-B.; Say's identity; Say's Law; Scrope, G. P.; Senior, N. W.; Sidgwick, H.; Smith, A.; Smith, S.; Sraffa, P.; stationary state; Steuart, J.; Stewart, D.; Sumner, J. B.; surplus; Thornton, H.; Thornton, W. T.; Thünen, J. H. von; Tooke, T.; Torrens, R.; Turgot, A. R. J.; Utilitarianism; Walker, F. A.; Walras, L.; Wayland, F.; Weiser,; West, E.; Whately, R.;

Whewell, W.; Wicksell, J. G. K.; Wicksteed, P. H.; Young, A. A.

Article

The ‘English School’ of political economy comprises all major British economists of the 19th century, together with J.-B. Say and perhaps Karl Marx.

‘Important changes have taken place in the meaning of the term “political economy,” as used by leading writers, since it was first employed’, wrote Henry Sidgwick in Palgrave's original *Dictionary of Political Economy* (Palgrave, 1899, pp. 128–9). As first used by Mayerne-Turquet and Montchrétien, ‘*œconomie politique*’ signified an attempt to extend the art of estate management to the entire kingdom of Louis XIII and his successors (Waterman, 2004, p. 225). This usage, generalized to mean a ‘system’ of policy designed to ‘increase the riches and power’ of a country (Smith, 1776, I.xi.n.1; II.5.31; IV.1.3) remained current until the end of the 18th century and was so employed by Steuart.

Adam Smith disliked the usage because of its implicit mercantilism. He recognized it, but proposed a better definition. ‘What is properly called Political Œconomy’ is ‘a branch of the science of a statesman or legislator’: namely ‘*an inquiry*’, which is in principle disinterested and open-ended, into ‘*the nature and causes of the wealth of nations*’ (Smith, 1776, IV. intro; IV.ix.38; emphasis added). The prestige that *Wealth of Nations* quickly acquired, amplified by Dugald Stewart's widely influential Edinburgh lectures in the new science, redefined ‘political economy’ as a ‘part of the science of human society’ (Palgrave,1899, p. 129; cf. Winch, 1983, who appears to disagree with this interpretation) and created a circle of younger thinkers committed both to criticizing and refining Smith's ideas and to propagating them among the governing classes. Though the *Edinburgh Review*, founded in 1802, was at first the principal means of propagation, most of the prime movers soon migrated to London, which from the second or third decade of the 19th century became the home of what was soon called the ‘English School’.

It is important to recognize that to describe the small community of anglophone political economists in the 1820s and after as a ‘school’ is to imply neither a quasi-apostolic succession of doctrine in some leading university nor a closed shop of experts defined by their adherence to any orthodoxy. It is rather the fact, as T.S. Eliot observed of all such intellectual circles in general, that ‘they are driven to each other's company by their common dissimilarity from everybody else, and by the fact that they find each other the most profitable people to disagree with’ (Kojecky, 1971, p. 244). Members of the English School, like all subsequent economists, were notorious for their disagreements, both with Adam Smith and with each other. But they did not find it profitable to disagree with hostile critics of their enterprise, such as the Lake Poets (from whom they were all indeed markedly ‘dissimilar’), because the latter chose not to acquire the viewpoint and vocabulary of the new, political-economy conversation, but resorted rather to the idioms of a very different conversation: that of Romantic aesthetics and non-utilitarian ethics.

In attending to the conversation of the English School it is necessary first to establish its identity, secondly to consider its members and its literature, and thirdly to distinguish its chief analytical features, especially as these differ both from the economic thought that preceded it and from economics of the present day. Finally, since the boundary between the political economy of the English School and what is generally thought of as ‘modern’ economics is vague and permeable, some attention should be paid to

continuity and ‘revolution’, if any.

Identity of the English School of political economy

Writing of ‘the English School of Political Economy’ in the original Palgrave dictionary, James Bonar (Palgrave, 1894, p. 730) observed that ‘The English writers on political economy before Adam Smith do not at any time present the marks of a “school” properly so called.’ What Bonar called ‘Modern Economics’ – meaning ‘political economy’ in the new, Smithian sense – he then divided into four periods headed respectively: Adam Smith, Malthus and Ricardo, John Stuart Mill, and W.S. Jevons; with all other authors subsumed under these canonical names.

Adam Smith was not an Englishman and he died before Malthus and Ricardo had begun to write. Though the *Edinburgh Review* (Anon., 1837, p. 73) referred to the English School as ‘the school of which Adam Smith was the founder’, this is Caledonian hyperbole. Smith founded no ‘school’. His most influential disciple, Dugald Stewart, was the intermediary between Smith and those the *Edinburgh Review* more accurately described later in the article as the ‘followers of Dr Smith’ practising ‘Political Economy, using the word in the sense of Ricardo and Malthus’ (Anon., 1837, pp. 77, 79). Subject to this important qualification, Bonar's chronology is helpful. Roughly speaking the English School lasted for about three generations. The first generation, from 1798 to the 1830s is that of which Malthus, Henry Thornton, Chalmers, James Mill, Torrens, West, Ricardo, and Thomas Tooke are now the best remembered. A second generation, whose members were active in some cases before 1830, but who flourished for the most part until the 1860s or even later, included Whately, Senior, McCulloch and J.S. Mill. Political economy of the English School never really died out. It changed, very gradually and almost imperceptibly, into the international, professionalized ‘economics’ of the mid-20th century. Yet a third and last generation can be detected – and was in fact detected in the 1890s – which included W.T. Thornton, the Fawcetts, Cairnes, Jevons, Bagehot, Foxwell, Sidgwick, J.N. Keynes and Nicholson. The positions of Marshall, Edgeworth and Wicksteed are problematic and will be considered below.

The English School was recognized by its difference from ‘the foreign school’ (Anon., 1837, p. 77) which included Sismondi, Cherbuliez and Villeneuve, but not J.-B. Say who from the first was deemed an honorary Englishman. The English writers distinguished ‘the art of government’ from the ‘science’ of political economy. With respect to the former, the latter is ‘only one of many subservient sciences; which involves the consideration only of motives, of which the desire for wealth is only one among many, and aims at objects to which the possession of wealth is only a subordinate means’ (Senior, 1836, pp. 129–30). The foreign writers rejected this minimalist construal, labelled it ‘chrematistics’ or ‘chrysology’, and continued to maintain that political economy embraces both the art and the science of government.

The incipient distinction between ‘art’ and ‘science’ seemed to imply that any practitioner of the latter must abstain – qua political economist – from political judgements. His analytical conclusions, being strictly positive and abstracted from ethical considerations, ‘do not authorize him in adding a single syllable of advice’. His business is ‘neither to recommend nor to dissuade, but to state general principles which it is fatal to neglect’ (Senior, 1836). McCulloch for one strongly disagreed with Senior on this point: the general principles, he thought, had already been completely enounced by Ricardo. What remained was ‘to exhibit some of their more important applications’ (McCulloch, 1843, p. vi). Though

Senior's view of the scope of political economy was tidied up and assimilated by the end of the 19th century (J.N. Keynes, 1891), all members of the ‘school’ were agreed at the outset on at least one most important ‘application’. The ‘great principles of free exchange and natural distribution’ that Smith had developed from ‘the philosophers of the Continent’ (that is, Quesnay, Turgot and so on) showed it to be economically unprofitable ‘for the legislature to intermeddle’ with trade and income distribution (Anon., 1837, pp. 80, 78). Though Cairnes later averred that ‘political economy has nothing to do with laisser faire’, Sidgwick thought this ‘too daring a paradox’.

> There can be no doubt that the interest of Adam Smith's book for ordinary readers is largely due to the decisiveness with which he offers to statesmen the kind of practical counsels which, according to Senior and Cairnes, he ought carefully to have abstained from giving. (Palgrave, 1899, pp. 130–1)

Rightly or wrongly, the political economy of the English School was associated in the popular mind with free trade and attacks on corporate privilege, and was denounced for these disturbing ideas by a wide variety of hostile critics.

Both the methodological tendencies of the new science and its ‘more important applications’ owed much to Dugald Stewart: the former to his influential *Philosophy of the Human Mind* (1792, 1814, 1827) the latter to his annual public lectures at the University of Edinburgh, beginning in the winter of 1800/1. Though preferring a broader definition of ‘political economy’ than that of either Smith or his English followers, and emphasizing the historical character of economic knowledge, Stewart argued in *Human Mind* that the hypothetical and a priori reasoning so characteristic of what he called the ‘new science’ – and which became one of the hallmarks of the English School – was perfectly legitimate, and compatible with the testing of theories against experience (Fontana, 1985, pp. 99–102; Waterman, 2004, ch. 8).

Stewart's Edinburgh lectures were crucial in what a recent author has aptly called ‘the process of Anglicisation of Scottish thought after 1790’ (Fontana, 1985, p. 9). Not only were they attended by Jeffrey, Horner, Brougham, Chalmers and the newly arrived Englishman, Sydney Smith, all of whom were influential in propagating political economy; Pryme (1823, p. vii) records that they ‘attracted so much attention that several members of our own university [namely, Cambridge] went from the South of England to pass the Winter at Edinburgh, for the purpose of attending them’: one of these seems to have been John Bird Sumner (Waterman, 1991a, pp. 159–60). According to a later account, ‘a wave of young Englishmen … went North in lieu of the grand tour made impossible by the renewal of war’ (Checkland, 1951, p. 43). Though the lectures were diffuse and circumspect, their underlying message was that contained in an early paper that Adam Smith had entrusted to Stewart before his death:

> Little else is required to carry a state to the highest degree of opulence from the lowest barbarism, but peace, easy taxes and a tolerable administration of justice; all the rest being brought about by the natural course of things (Smith, 1755). (Winch, 1996, p. 90)

Leading members of Stewart's circle – Jeffrey, Horner, Brougham, and Sydney Smith – founded the *Edinburgh Review* to urge this message upon the Holland House Whigs from whom they hoped to

receive patronage. First Smith, then Jeffrey, served as editor until 1829, when replaced by McVey Napier. By that date its contributors on political economy had included all the leading members of the English School save Ricardo (who declined out of modesty, and who died in 1823): Malthus, James Mill, Chalmers, Torrens and McCulloch (Fontana, 1985, p. 8).

Of these authors, all save Chalmers were members of the Political Economy Club, a London dining club founded in 1821 which, in addition to Malthus, Mill and Torrens, included from the outset Ricardo, George Warde Norman and Thomas Tooke. J.-B. Say was elected as an Honorary Foreign Member in 1822, the only such member until 1919. McCulloch was elected in 1829, shortly after his migration from Scotland; Senior (in 1823), Pryme (1828) and Whately (1831) were elected as Honorary Members by virtue of their professorships in political economy. Cairnes (1862), Cliffe Leslie (1862), Fawcett (1862), Jevons (1873), Foxwell (1882), Marshall (1886), Nicholson (1888) and Edgeworth (1891) were all subsequently elected under this rule. Among those political economists now remembered as influential authors of the English School, only Henry Thornton, Sir Edward West, Archbishop J. B. Sumner, Thomas Chalmers, Poulett Scrope, and Richard Jones were never members of the Club: Thornton because he died in 1815, West because he went to India, Chalmers because he stayed in Scotland, and Sumner because he announced in 1818 – to Ricardo's regret – that he intended to give up political economy for the study of theology (Waterman, 1991a, p. 157). Scrope and Jones were on the outer edge of the ‘School’.

It has been suggested that the English School was a ‘scientific community’ of which the Political Economy Club was a ‘vital hub’ (O’Brien, 2004, pp. 12–13). There is merit in this suggestion, but it should be recognized that the original purpose of the club, though including the ‘mutual instruction’ of members, was chiefly propagandist: ‘the diffusion amongst others of the just principles of Political Economy’ and

> to watch carefully the proceedings of the Press, and to ascertain if any doctrines hostile to sound views on Political Economy have been propagated … to refute such erroneous doctrines, and counteract their influence … and to limit the influence of hurtful publications. (Political Economy Club, 1921, p. 375)

Many members were Whig or liberal statesmen who knew a ‘hurtful publication’ when they saw one: 52 of the 115 elected between 1821 and 1870 sat in either the upper or lower House of Parliament; and included Lord Althorp, the Marquis of Landsdown (a descendent of Sir William Petty), Earl Grey and W. E. Gladstone. Fetter (1980) has documented the activities of ‘the economists in Parliament’.

Almost from the first there was a desire by the Club to recognize and foster the academic study of political economy. Though there had been high-level economic analysis at British universities before the end of the 18th century, it was but a small ingredient of ‘moral and political philosophy’ (for example, see Waterman, 1995) and never known as ‘political economy’. But in the decade of the 1820s chairs in political economy were established in Oxford, London and Cambridge and their incumbents immediately co-opted (Checkland, 1951).

We may therefore identify the English School roughly speaking as that subset of Political Economy Club members in the 19th century who published and disputed with each other on the subject, together with half a dozen or so other major authors who at some time or other were part of their conversation.

Despite its name, several leading members were Scotch immigrants, and it included one Frenchman. Though Karl Marx lived in London from 1848 and thoroughly digested the literature of anglophone political economy over the next two decades, he was not known or recognized by the Club. But in the ‘Afterword’ to the second German edition of *Capital* (Marx, 1873, vol. 1, p. 26) he explicitly identified his own work, in method at least, with that of the English School.

Literature of the English School

Literature of the English School begins with Malthus's first *Essay on Population* (1798). For as an unintended consequence of his Whiggish polemic against Godwin's (1793) romantic anarchism, Malthus analysed the effect of population growth under land scarcity to show what was later called ‘diminishing returns’ (Stigler, 1952). Though diminishing returns in agriculture had been identified by Steuart (1767) and Turgot (1768), and had actually been used by Anderson (1777) to adumbrate the ‘Ricardian’ theory of rent, the concept was not integrated into 18th-century economic thought. Notwithstanding Samuelson's influential interpretation, land scarcity plays little or no analytical part in *Wealth of Nations* (Samuelson, 1978; cf. Hollander, 1998; Waterman, 1999). When Malthus (1815a), West (1815), Torrens (1815) and Ricardo (1815) worked out the implications of Malthus (1798) they believed that they were correcting Smith and saying something new and important (McCulloch, 1845, p. 68). Diminishing returns immediately became part of the hard core of the so-called classical political economy of the English School.

Ricardo made diminishing returns in agriculture the cornerstone of his *Principles* (1817), combined it with ‘Malthusian’ population theory, Smith's account of accumulation and growth, and an ad hoc ‘93% Labor Theory of Value’ (Stigler, 1958) to produce a complete account of value, distribution and growth in a two-sector market economy. The labour theory of value (LTV) was also the key concept in Ricardo's rigorous and elegant analysis of comparative advantage in international trade. Looking back 30 years later, McCulloch (1845, p. 16) called the LTV ‘the fundamental theorem of the science of value’. An authoritative and exhaustive account of Ricardo's contribution – which it treats, à la McCulloch, as virtually identical with ‘classical economics’ – appeared in the first edition of *The New Palgrave Dictionary of Economics* (Blaug, 1987).

In addition to the above works, the ‘English’ literature that already existed by the time the Political Economy Club was founded in 1821 included Malthus's (1800) *High Price of Provisions*, which formally specified a demand function of price and inaugurated the supply-and-demand value theory that eventually ‘won out’ over the Ricardo–Marx LTV (Smith, 1956; Schumpeter, 1954, p. 48) which it generalizes, Thornton's (1802) *Paper Currency*, which analysed the macroeconomic relations between monetary and real variables in a manner reinvented by Wicksell a century later, and numerous pamphlets by many authors on monetary questions provoked by the Parliamentary Bullion Committee of 1810. It was this controversy that brought Malthus and Ricardo together, and which seems to have been a catalyst for the nascent ‘scientific community’. The pre-1821 literature also includes J.-B. Say's (1803) *Traité d’économie politique*; Lauderdale's (1804) *Inquiry*, dismissed by McCulloch (1845, p. 15) as without value; Chalmers's (1808) strikingly original but completely neglected *Nature and Stability of National Resources* (see Waterman, 1991b); Malthus's (1815b) heretical pamphlet, ‘Restricting the

Importation of Foreign Corn', which led to his excommunication by the *Edinburgh Review* (Fontana, 1985, p. 75); J. B. Sumner's (1816) *Records of the Creation* that Ricardo (1951–73, vol. 7, pp. 247–8) deemed a 'clever book' and which McCulloch (1845, p. 261 described as 'an excellent work'; the fifth edition of Malthus's Essay on Population (1817) substantially modified as a result of Sumner's arguments; Mrs Marcet's (1817) influential work of popularization, *Conversations on Political Economy*; and Copleston's (1819a; 1819b) two brilliant and penetrating *Letters to Peel* that grasped more clearly than Malthus himself the connection between population and poverty, and between the latter and inflation of the currency – and which Ricardo so admired that he made a detailed paragraph-by-paragraph summary (Waterman, 1991a, pp. 186–95; Hollander, 1932, p. 135–45). Finally, shortly before or just after the first meeting of the Club there appeared important monographs by three of the founding fathers: Malthus's (1820) *Principles*, which quarrelled with Ricardo over value theory and put forward a heterodox macroeconomics of 'general gluts' that Keynes was later to find so appealing, James Mill's (1821) *Elements of Political Economy*, and Torrens's (1821) long undervalued *Essay on the Production of Wealth*.

It is apparent that during the first two decades of the 19th century, and for a further ten years or more, Malthus was at the centre of the political-economy conversation of the English School. This fact has been obscured by the excessive attention paid to Ricardo by those eager to praise or blame him for present-day economics, and by textbook authors wanting a handle on which to hang a student-friendly chapter on 'classical economics'. A long process of reappraisal, beginning with J. M. Keynes's (1972, vol. 10) biographical essay of 1933, has gradually restored the true picture (Waterman, 1998). Donald Winch's (1996) *Riches and Poverty* is the latest and most authoritative intellectual history of political economy, covering the period 1750–1834. Nearly half his book is concerned with Malthus. Ricardo, 'treated largely as a foil to Malthus' (Winch, 1996, p. 15) gets a few scattered references. Samuel Hollander's (1997) magisterial *Economics of Thomas Robert Malthus* shows that the analytical differences between Malthus and Ricardo have been exaggerated, and that the former was a theoretician of the same order, and of at least as much historical importance as the latter.

Malthus was central because the first *Essay* began a century-long transformation of 'political economy' (the science of wealth) into 'economics' (the science of scarcity). The theological implications of this, totally ignored by most historians, are a vital part of the intellectual context of the English School. Economic thought of the 18th century was believed by all to be wholly compatible with Christianity. But the seeming inevitability of 'misery' or 'vice' produced by human fecundity and resource scarcity challenges the goodness of God; and the political economy of Malthus and Ricardo was therefore condemned as 'hostile to religion'. For most of the 19th century, England was both officially and actually a Christian society. In such a society it is part of the duty of a scientist – essential if his work is to receive serious attention – to reconcile his findings with Christian theology. Malthus attempted this in 1798 and failed. His failure stimulated an important branch of the literature of the English School now known as 'Christian Political Economy' (Waterman, 1991a). Works by William Paley (1802), by Malthus himself (1803; 1817), and by J. B. Sumner (1816) who eventually became Archbishop of Canterbury, demonstrated that the new science could be co-opted as theodicy; and even better, be used to demonstrate the benevolent 'design' of the Creator. The approval that Ricardo and McCulloch evinced for Sumner's 'clever book' had less to do with their own religious convictions than

with their relief that political economy had been convincingly defended against the damaging charge of irreligion.

Quite different circumstances in the 1820s revived the need to defend political economy against religion, and created a new need: to defend religion against political economy. Jeremy Bentham, James Mill and other Benthamites, who were later called the ‘Philosophic Radicals’, founded the *Westminster Review* in 1824 to propagate a ‘radical’ reformism as against the Whiggish reformism of the *Edinburgh*. Anti-clerical and at times anti-religious, the radicals hijacked political economy to mount a strictly utilitarian attack on the Establishment in Church and State. Animated by James Mill's puritanical hatred of the Arts, the *Westminster* compounded the injury by gratuitous attacks on the Lake Poets and other romantic authors. Influential Tories at the two universities (then exclusively Anglican) were alarmed, and opposition was made to the teaching of, and the establishment of chairs in, political economy. In this crisis, both political economy and Christian theology were authenticated and insulated against mutual encroachment by two Oxford men, Richard Whately, a former pupil and friend of Copleston, and Nassau Senior, Whately's former pupil and friend (Waterman, 1991a, pp. 196–215).

Whately engineered the election of Senior as first Drummond Professor of Political Economy in 1826, and accepted the chair himself when it fell vacant in 1830. His seminal Introductory Lectures (1831) argued for an epistemological demarcation between ‘religious and ‘scientific’ knowledge; and explained how, like all scientific knowledge, political economy depends upon both a priori deduction and the possibility of falsification. Whately thus established the methodological tradition of the English School that runs through Senior, J.S. Mill, J.N. Keynes and Lionel Robbins. Pietro Corsi (1987) has shown that Whately's philosophical apparatus was based on Dugald Stewart's *Philosophy of the Human Mind*, transmitted to Oxford through the friendship between Stewart and Copleston created by the migration from Edinburgh to Oxford in 1799 of J.W. Ward, 1st Earl of Dudley.

Whately's decisive intervention healed a potentially disastrous schism in the young ‘scientific community’ between Benthamite radicals and Malthusian Whigs. Elections to the Political Economy Club in the 1820s and 1830s included both Whigs and radicals and even the liberal Tory, Lord Althorp. When McVey Napier edited the 1824 *Supplement to the Encyclopaedia Britannica* he commissioned articles on political economy from Malthus and Sumner on the one hand, and from Mill and McCulloch on the other. (Ricardo's contribution, on the Funding System, was posthumous.) The Royal Commission on the Poor Laws (1832) which included Sumner, then Bishop of Chester, united all in the common cause once again. Malthus was the most important witness. The report, which led to the Poor Law Amendment Act (1834), was jointly written by the Benthamite Chadwick and the Whatelian Senior, and was based on Copleston's (1819b, p. 28) crucial distinction between ‘propagation’ and ‘preservation’ of human life.

One of the most interesting, certainly the most revealing, contributions to literature of the English School is McCulloch's compendious *Literature of Political Economy* (1845) which appeared about halfway through the life of the ‘school’. The usual English and Scotch authors from Mun and Petty are listed, and many of their works praised or censured in light of McCulloch's doctrinal preconceptions. Malthus is predictably belittled. All the leading French authors of the 18th and early 19th centuries appear save Boisguilbert and Cournot. Condillac's path-breaking *Le Commerce et le gouvernement* (1776) is dismissed with a patronizing comment of J.-B. Say (McCulloch, 1845, p. 63; cf. Eltis and Eltis, 1997, pp. 30–4). Considerable respect is paid to Italian authors (McCulloch, 1845, pp. 28–31, 86), but the Spanish are written off as intellectually impotent until Napoleon's invasion (1845, pp. 31–2, 326).

McCulloch seems never to have heard of Thünen, and no other German author is mentioned. Omissions of anglophone authors are equally telling. Whewell's pioneering mathematical economics is ignored, presumably for the same reason as the omission of Cournot and Thünen. Dugald Stewart is cited merely as a biographer of Adam Smith and Robertson (1845, pp. 8, 104, 162). McCulloch seems not have read or understood either Chalmers (1808) or Copleston (1818), nor to have grasped the analytical significance of Malthus (1800). Everything is viewed through the powerful but slightly distorting lenses of Adam Smith and Ricardo.

Three years later there appeared the single most important production of the School: J.S. Mill's *Principles of Political Economy* (1848), perceptively reviewed by Bagehot (1848) among many others. Mill's *Principles* is the definitive statement of the English School of political economy. It went through seven editions in the author's lifetime; the 1909 scholarly edition by Ashley was based on the seventh (1871), and may be taken as the terminus ad quem of the English School. For though Mill continued to be the principal textbook in political economy until the 1930s at many universities throughout the English-speaking world, Anglophone economic literature of the 20th century gradually became less insular (Palgrave, 1894, p. 735) and was formed in the cautiously new idiom of Marshall and Pigou, with at least some peripheral awareness of Jevons and Edgeworth, Walras and Pareto, Weiser and Böhm-Bawerk, Cassel and Wicksell, J.B. Clark and Fisher.

Though Mill dominated, there were many other significant contributions to the literature in the last third of the 19th century. Henry Fawcett's *Manual of Political Economy* (1863) encapsulated Mill's *Principles* for faint-hearted undergraduates; his wife's even more elementary *Political Economy for Beginners* (1870) went through ten editions over the next 41 years. W.T. Thornton's *On Labour* (1869) introduced the concept of multiple equilibria, as Mill (1869, p. 637) admitted. Cairnes's *Leading Principles* first appeared in 1874, Cliffe Leslie's *Essays* in 1879, Bagehot's posthumous *Economic Studies* in 1880, and Henry Sidgwick's *Political Economy* in 1883. Sidgwick's importance in the incipient ‘Cambridge’ mutation of the English School has lately been documented (Backhouse, 2006). J.N Keynes's classic *Scope and Method* first appeared in 1891. Perhaps the last major production of the English School was J. Shield Nicholson's three-volume *Principles of Political Economy* (1893–1901), a basically Millian exposition with the occasional bow to Marshall, used as a textbook in many parts of the British Empire in the early 20th century. Nicholson's appears to be the last widely read work of political economy to consider explicitly the relation between that science and Christian theology (1893–1901, vol. 3, ch. 20).

Stanley Jevons (1871, p. 275) went out of his way to challenge ‘the noxious influence of authority’ in the English School, above that of Mill. Though elected to the Political Economy Club as a professor in 1873 and as an Ordinary Member in 1882 (the year of his death), he was therefore handled with caution by his fellow-economists – including the powerfully influential Marshall. Whilst crediting him with the intellectual defeat of Ricardian and Marxian value theory, Bonar (Palgrave, 1894, p. 735) thought that ‘the ideas of Jevons have had greater power since his death than during his life’. Jevons and his two most creative English followers, Edgeworth and Wicksteed, were ‘often spoken of as a school by itself, the mathematical school’ (Palgrave, 1894). The original Palgrave article on ‘Recent Developments of Political Economy’ (Palgrave, 1894, p. 148) alludes to Jevons's *State in Relation to Labour* (1882) but ignores his *Theory of Political Economy* (1871).

Literature of the English School was augmented and popularized by *The Economist* newspaper, founded

in 1843 and edited by Walter Bagehot from 1860 to 1877, which, like the Political Economy Club, sought to relate economic analysis to public policy in the spirit of Adam Smith. That literature may be said to have culminated in the three-volume *Dictionary of Political Economy* (1894–1899) edited by R. H. Inglis Palgrave.

Some analytical features of the English School

Political economists of the English School inherited much of their economic analysis from their 18th-century predecessors, especially Cantillon, Hume, Quesnay, Smith and Turgot. However, some features of their analysis were as ‘novel’ as any idea ever is in the social sciences. And despite loose talk about a ‘marginal revolution’, much of their analysis, both what they inherited and what they originated, has become part of the stock-in-trade of present-day economics. The standard account by D.P. O’Brien (2004) should be supplemented by S.J. Peart's and D. Levy's (2003) review of the period 1830–1870, which considers catallactics, methodological egalitarianism and the new ideological alliance – a mutation of the old Whig-Liberal orthodoxy – between political economists and reformist Evangelicals in the Church of England.

The central conception of 18th-century economic thought was that of a surplus of production in one period over and above what is necessary (as inputs into production) to sustain that level of production in the next. The agricultural sector is an obvious source of the surplus since land normally produces more than the (food) cost of necessary labour and capital inputs. But Smith generalized the concept to include all produced goods capable of use as inputs. Masters incur production costs in advance, hence control the entire output at the end of the process. Some of this they consume either directly, or in the employment of unproductive labour. The remainder is used to feed and equip productive labour. This unconsumed portion of output is the (circulating) capital stock of a master, firm or community, the growth, stationarity or decay of which depends on a psychological propensity of masters: the extent of their ‘frugality’ or parsimony (Eltis, 2000, pp. 75–100). These ideas, and the necessarily dynamic analytical framework they imply, were taken for granted by most the English School despite its seeming incompatibility with such other conceptions as comparative advantage in trade (Blaug, 1987, vol. 1, pp. 439–42). Other characteristically 18th-century ideas accepted by ‘the followers of Dr Smith’ included that of a labour supply perfectly elastic in the (Malthusian) long period at a socially determined zero-population-growth real wage; enough factor mobility to produce uniform rates of wages and profit throughout the economy; a negative relation between the real wage and the rate of profit; a positive relation between the general price level and the stock of money, and the Cantillon–Hume price-specie-flow mechanism of international monetary adjustment which follows from that relation. Most accepted Smith's account of natural prices that correspond, more or less, to Marshall's long-period equilibrium prices, but O’Brien (2004, ch. 4) has shown in detail how much variation there was in this matter.

Perhaps the most important 18th-century idea, certainly that which gave the English School its ideological momentum, was Boisguilbert's vision – derived from the Jansenist theology of Pierre Nicole and Jean Domat – of a self-regulating market economy driven by ‘self-love’ and producing some kind of social optimum at competitive equilibrium (Faccarello, 1999). This powerful conception was transmitted by Mandeville, Cantillon and Quesnay and canonized by Smith in *Wealth of Nations*.

As we have seen, the English School made at least one sharp analytical break with 18th-century thought.

The explicit incorporation of diminishing returns (though as yet in agricultural production only) created a fundamentally different view of the economic universe. Though all recognized increasing returns to scale (IRS) resulting from the division of labour, IRS plays a small or negligible part in the implicit growth models of Malthus and his successors (Eltis, 2000). The salient feature of the new growth theory was rather a tendency for the rate of profit to fall: either because of rising costs in agriculture as in Malthus and Ricardo, or because of increasing capital intensity in manufactures as in Marx. In the former case, falling real factor payments retarded the growth of capital and labour, leading to a stationary state in the absence of technical progress. Samuelson (1978) has shown that the variable factor in agriculture was conceived as a single ‘labor-cum-capital’ unit, and though all ‘classical’ economists recognized the possibility of factor substitution especially in manufacturing, the capital–labour ratio was generally taken as a parameter. The same was true of technique. Improvements were seen to occur from time to time, and their effect upon wages, profits and employment analysed. Malthus, and perhaps some others, recognized that technical progress could become endogenous (Eltis, 2000, pp. 150 ff.) and few if any of the English School regarded it, as some do today, as ‘manna from Heaven’.

Two other new, or somewhat new, analytical features of the English School deserve note. The first is the LTV theory of comparative advantage, later improved by Mill's analysis of reciprocal demand. The second is Say's Law of Markets, which in its strong form (Say's identity) implies the neutrality of money (Blaug, 1996, pp. 143–60). Whether Samuel Hollander (for example, 1987, pp. 6–7) is correct in maintaining that Ricardo and his contemporaries and successors, including Marx, recognized ‘a fundamentally important core of general-equilibrium economics accounting for resource allocation in terms of the rationing function of relative prices’ is still a matter of debate (Blaug, 1987, vol. 1, pp. 442–3).

It is evident that most of these analytical characteristics, both those inherited from the 18th century and those that were new, have been transmitted to present-day economic thought. The obvious exception is the concept of a surplus with its concomitant distinction between ‘productive’ and ‘unproductive’ labour; though in the spirit of Feyerabend's (1988) methodological anarchism this venerable doctrine has lately been brought back to useful life (Bacon and Eltis, 1976). For the most part however, present-day economists prefer to rely on a putatively constant-returns-to-scale (CRS) general equilibrium model that abstracts from time, and in which each factor-owner is paid the value of his factor's marginal product. The surplus is therefore regarded as a museum piece and left to heterodox Marxists and Sraffians (Walsh and Gram, 1980; cf. Blaug, 1987, vol. 1, pp. 440–2). It is important to recognize, however, that the eventual disappearance of the surplus in a neoclassical theory of distribution was brought about by an ever wider application of the marginal analysis originally applied by Steuart, Turgot, and Anderson, and then by Malthus, Ricardo and their contemporaries to agricultural production costs alone (Blaug, 1987, vol. 1, p. 441). Authors of the next generation such as Longfield and Lloyd began the analysis of marginal utility (O’Brien, 2004, pp. 119–22). Replacement of the dynamic surplus macroeconomics by a static general-equilibrium microeconomics dependent on universal CRS created perhaps the most significant analytical difference between political economists of the English School and the new professionalized economists of the early 20th century: an almost complete lack of interest among the latter in macroeconomics and growth theory. Not until Keynes's rediscovery of Malthus (Kates, 1994) and Harrod's (1939) critique of Keynesian ‘equilibrium’ did these return to the theoretical agenda. As for

Adam Smith's IRS, quietly forgotten by most of the English School – save Marshall – for most of the time and ignored by their successors, its reintroduction by Sraffa (1926) and Young (1928) has remained a thorn in the flesh for general equilibrium theorists.

Revolution and continuity

Present-day ‘economics’ looks quite different from ‘political economy’ of the English School. Yet despite Samuelson's remarks about Marshall in *Foundations* (1947, pp. 6, 142, 311–12) and despite his focus on Walrasian general equilibrium in that work, the microeconomic part of his immensely influential *Economics* (1948) is unmistakeably Marshallian, at any rate as mediated by Chamberlin (1933) and Joan Robinson (1933). And though Marshall had digested Thünen and Cournot, knew the work of Menger and the Austrian School, and admitted that ‘there are few writers of modern times who have approached as near to the brilliant originality of Ricardo as Jevons has done’ (Marshall, 1920, p. 673), yet he ‘consistently discounted the “Jevonian revolution”’ (Schumpeter, 1954, p. 826) and used all his influence, which was great, to insist that in science, as in the world it contemplates, *Natura non facit saltum*. There are few references to Jevons in his famous *Principles*, and in the most extended of these (Appendix I) Marshall went out of his way to counter the former's ‘antagonism to Ricardo and Mill’ and to defend their value theory against his intemperate exaggerations’ (Marshall, 1920, pp. 673–6; see also O’Brien, 1994, vol. 2, pp. 325–61).

Upon the evidence of Palgrave's original dictionary it appears that by the last decade of the 19th century the effect of Marshall's efforts had been to co-opt Jevons and his ‘marginalist’ followers into the mainstream of English political economy with a minimum of fuss, and with a minimum of attention to the continental marginalists. Jevons's ‘final utility’ became ‘marginal utility’ in Marshall's *Principles* (1920, pp. 78–85), and there was used with deceptive innocence (see Blaug, 1996, p. 322–37) to generate a market demand function of price. Though Edgeworth himself contributed 17 articles to the dictionary, including ‘Cournot’, ‘Curves’ and ‘Demand Curve’ in volume 1, ‘Mathematical Methods’ in volume 2 and ‘Pareto’, ‘Pareto's Law’, ‘Supply Curve’ and ‘Utility’ in volume 3, his own work was ignored in the general surveys of ‘Political Economy’ and ‘The English School’ and his name omitted from the index of volume 1, along with those of Menger and J.B. Clark. Walras received three short references in that volume. Not until volume 3 (1899, pp. 652–5) was his work recognized, and then only for its use of marginal utility. There is no awareness of general equilibrium in that article, and the term appears nowhere else in the original *Dictionary*.

It would appear from the foregoing that if there really was any such thing as a marginal revolution in Anglophone political economy, it began as early as 1767 with Steuart's *Political Œconomy* and still had some way to go by the time volume 3 of the Palgrave dictionary appeared in 1899. Thünen's (1826) generalization of diminishing returns to all factors of production remained unnoticed by any save Marshall. Though Wicksteed (1894) and Flux (1894) reinvented this wheel, Wicksteed's (Palgrave, 1899, pp. 140–2) own contribution to the Palgrave article on ‘Political Economy’ only hints at what later became known as the neoclassical theory of distribution. In 1895 Edgeworth rejected Barone's submission to the *Economic Journal* showing that product exhaustion is implied by Walras's (1894) cost-minimization equations. A companion article to Wicksteed's baldly states that ‘the law of diminishing

returns points to an increase in the cost of agricultural produce accompanying increase of population' (Palgrave, 1899, p. 140). For that author at any rate, nothing had changed since Malthus.
In summary, it would appear that the English School was alive and well in the first decade of the 20th century. Elections to the Political Economy Club included Pigou (in 1906) and J.M. Keynes (1912), along with the Bishop of Stepney (1904), the Rt Hon. Herbert Samuel MP, the Viscount Ridley (1907) and John Buchan (1909). Mill's *Principles* was still perhaps the most widely used textbook. Questions on Adam Smith still appeared in university examinations in political economy (for example, at Edinburgh, 21 November 1898, 17 March 1899). Mathematics was still an unwelcome eccentricity. Jevons (1871, p. vii) had asserted that economics 'must be a mathematical science in matter if not in language'. Marshall (for example, 1890, p. ix) threw all his influence against this doctrine and locked up his own sophisticated mathematics in well-guarded appendices (Keynes, 1972, pp. 182–8). Despite his dependence upon mathematical reasoning and his prominence in the emerging profession of economics, Edgeworth's deference for Marshall deterred him from challenging a Cambridge, anti-mathematical orthodoxy that persisted until the 1950s.
Edgeworth was unusual, too, in his ability and willingness to read foreign authors and to recognize their contributions (Keynes, 1972, pp. 263–5). In general, the insularity of the English School persisted until well into the 20th century. When Harrod was about to begin his studies in economics, Keynes advised him not to waste his time on the Continent 'where they knew nothing at all of economics' (Harrod, 1952, pp. 317–19). The 'market socialists' of the 1930s, none of whom was English, were the first to specify the complete set of marginal conditions required for a welfare optimum in general competitive equilibrium. J.R. Hicks (1939, p. 6) believed himself to be the first English author to 'free the Lausanne School from the reproach of sterility brought against it by the Marshallians'.
It might have been expected that political economists in the United States, at any rate, would have identified with the English School. In the early 19th century authors such as Wayland (1837) had assimilated Malthus and Ricardo, and as late as 1888 Amasa Walker regarded Jevons and Marshall as 'an extension of the English School' (Goodwin, 1972, p. 562). But throughout much of the century protectionist sentiment in the USA was at variance with the ideology of free trade promoted by the English School. And towards the end of that century there was 'an estrangement from British scholarly life' created by a 'growing attachment to German thought' (Goodwin, 1972, p. 563). The American Economic Association was originally formed to promote the Liberal-Protestant 'social gospel', very different in spirit and substance from the aristocratic Whiggery of the Political Economy Club.

Bibliography

Anderson, J. 1777. *Observations on the Means off Exciting a Spirit of National Industry; Chiefly Intended to Promote the Agriculture, Commerce, Manufactures, and Fisheries, of Scotland*. Dublin: Price.

Anon. 1837. 1. *An Outline of the Science of Political Economy*. By Nassau W. Senior. London: 1836. 2. *Principes Fondamentaux de l'Économie Politiques, tirés de leçons édites et inédites, de M. N. W. Senior*. Par le Compte Jean Arrivabe. Paris: 1836. *Edinburgh Review* (October), 73–102.

Backhouse, R.E. 2006. Sidgwick, Marshall and the Cambridge school of economics. *History of Political Economy* 38, 15–44.

Bacon, R. and Eltis, W. 1976. *Britain's Economic Problem: Too Few Producers*. London: Macmillan.

Bagehot, W. 1848. Review of J.S. Mill's Principles of Political Economy. *Prospective Review* 4, 460–502.

Bagehot, W. 1880. *Economic Studies*, ed. R.H. Hutton. London: Longmans, Green.

Blaug, M. 1987. Classical economics. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Blaug, M. 1996. *Economic Theory in Retrospect*, 5th edn. Cambridge: Cambridge University Press.

Cairnes, J.E. 1874. *Some Leading Principles of Political Economy Newly Expounded*. London: Macmillan.

Chalmers, T. 1808. *An Enquiry into the Nature and Stability of National Resources*. Edinburgh: Moir.

Chamberlin, E.H. 1933. *The Theory of Monopolistic Competition: a Reorientation of the Theory of Value*. Cambridge, MA: Harvard University Press.

Checkland, S.G. 1951. The advent of academic economics in England. *The Manchester School* 19, 43–70.

Cliffe Leslie, T.E. 1879. *Essays in Political and Moral Philosophy*. London: Longmans, Green.

Copleston, E. 1819a. *A Letter to the Right Hon. Robert Peel, MP for the University of Oxford, on the Pernicious Effect of a Variable Standard of Value, especially as it regards the Condition of the Lower Orders and the Poor Laws* … Oxford: Murray.

Copleston, E. 1819b. *A Second Letter to the Right Hon. Robert Peel, MP for the University of Oxford, on the Causes of the Increase in Pauperism, and on the Poor Laws* … Oxford: Murray.

Corsi, P. 1987. The heritage of Dugald Stewart: Oxford philosophy and the method of political economy. *Nuncius* 2, 89–143.

Eltis, S. and Eltis, W. 1997. The life and contribution to economics of the Abbé de Condillac. In E.B. Abbé de Condillac, *Commerce and Government*, trans. S. Eltis. Cheltenham: Elgar.

Eltis, W. 2000. *The Classical Theory of Economic Growth*, 2nd edn. Basingstoke: Palgrave.

Faccarello, G. 1999. *The Foundations of Laissez-Faire: The Economics of Pierre de Boisguilbert*. London: Routledge.

Fawcett, H. 1863. *Manual of Political Economy*. London: Macmillan.

Fawcett, M.G. 1870. *Political Economy for Beginners*. London: Macmillan.

Fetter, F. 1980. *The Economist in Parliament, 1780–1868*. Durham, NC: Duke University Press.

Feyerabend, P. 1988. *Against Method*. London: Verso.

Flux, A.W. 1894. Review of Wicksteed. Repr. in *Precursors in Mathematical Economics*, ed. W. Baumol and S.M. Goldfield. London School of Economics, 1968.

Fontana, B. 1985. *Rethinking the Politics of a Commercial Society: The Edinburgh Review 1802–1832*. Cambridge: Cambridge University Press.

Godwin, W. 1793. *Enquiry Concerning Political Justice and its Influence on Morals and Happiness*. London: Robinson.

Goodwin, C.G.W. 1972. Marginalism moves to the New World. *History of Political Economy* 4, 551–70.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49, 14–33.

Harrod, R.F. 1952. *The Life of John Maynard Keynes*. London: Macmillan.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hollander, J.H. 1932. *Minor Papers on the Currency Question, 1809–1823 by David Ricardo*. Baltimore: Johns Hopkins Press.

Hollander, S. 1987. *Classical Economics*. Oxford: Blackwell.

Hollander, S. 1997. *The Economics of Thomas Robert Malthus*. 2 vols. Toronto: University of Toronto Press.

Hollander, S. 1998. The canonical classical growth model: content, adherence and priority. *Journal of the History of Economic Thought* 20, 253–77.

Jevons, W.S. 1871. *Theory of Political Economy*. London: Macmillan.

Jevons, W.S. 1882. *The State in Relation to Labour*. London: Macmillan.

Kates, S. 1994. The Malthusian origins of the General Theory or how Keynes came to write a book about Say's Law and effective demand. *History of Economics Review* 21, 10–20.

Keynes, J.M. 1972. Essays in biography. In *The Collected Writings of John Maynard Keynes*, vol. 10, ed. E. Johnson and D. Moggridge. London: Macmillan.

Keynes, J.N. 1891. *The Scope and Method of Political Economy*. London: Macmillan.

Kojecky, R. 1971. *Eliot's Social Criticism*. London: Faber.

Lauderdale, J. and Maitland, E. 1804. *An Inquiry into the Nature and Origin of Public Wealth: and into the Means and Causes of its Increase*. Edinburgh: Constable.

Malthus, T.R. 1798. *An Essay on the Principle of Population as It Affects the Future Improvement of Society, with Remarks upon the Speculations of Mr Godwin, M. Condorcet, and Other Writers*. London: Johnson.

Malthus, T.R. 1800. *An Investigation of the Cause of the Present High Price of Provisions. By the Author of the Essay on the Principle of Population*. London: Johnson.

Malthus, T.R. 1803. *An Essay on the Principle of Population, or, A View of its Past and Present Effects on Human Happiness, with an Inquiry into our Prospects Respecting the Future Removal or Mitigation of the Evils which it Occasions*. London: Johnson.

Malthus, T.R. 1815a. *An Inquiry into the Nature and Progress of Rent, and the Principles by which is Regulated*. London: Murray.

Malthus, T.R. 1815b. *The Grounds of an Opinion on the Policy of Restricting the Importation of Foreign Corn* … London: Murray.

Malthus, T.R. 1817. *An Essay on the Principle of Population … 4th edn of Malthus (1803), described as 5th edn*. London: Hunter.

Malthus, T.R. 1820. *Principles of Political Economy, Considered with a View to their Practical Application*. London: Pickering.

[Mrs Marcet (Jane Haldimand)] 1817. *Conversations on Political Economy: in which the Elements of*

that Science are Familiarly Explained. London: Longman et al.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan1952.

Marx, K. 1873. *Capital: A Critique of Political Economy*, ed. F. Engels, trans. S. Moore and E. Aveling. 3 vols. Moscow: Progress Publishers, 1954.

McCulloch, J.R. 1843. *Principles of Political Economy*. Edinburgh: Tait.

McCulloch, J.R. 1845. *Literature of Political Economy: A Classified Catalogue*. London: Longmans.

Mill, J. 1821. *Elements of Political Economy*. London: Baldwin et al.

Mill, J.S. 1869. Thornton on labour and its claims. In *Collected Works of John Stuart Mill*. vol. 5, ed. J. M. Robson. Toronto: University of Toronto Press, 1967.

Mill, J.S. 1871. *Principles of Political Economy*, 7th edn. ed. W.J. Ashley. London: Longmans, Green, 1909.

Nicholson, J.S. 1893–1901. *Principles of Political Economy*. 3 vols. London: Macmillan.

O'Brien, D.P. 1994. *Methodology, Money and the Firm*. 2 vols. Aldershot: Elgar.

O'Brien, D.P. 2004. *The Classical Economists Revisited*. Princeton, NJ: Princeton University Press.

Paley, W. 1802. *Natural Theology*. London: Wilkes and Taylor.

Palgrave, R.H. 1894–1899. *Dictionary of Political Economy*. 1899. Vol. 1, 1894; vol. 2, 1896; vol. 3, London: Macmillan.

Peart, S.J. and Levy, D.M. 2003. 1830–1870: Post-Ricardian British economics. In *A Companion to the History of Economic Thought*, ed. W.J. Samuels, J.E. Biddle and J.B. Davis. Oxford: Blackwell.

Political Economy Club. 1921. *Minutes of Proceedings, 1899–1920, Roll of Members and Questions Discussed, 1821–1920, with Documents bearing on the History of the Club*. London: Macmillan.

Pryme, G. 1823. *Introductory Lecture and Syllabus*. Cambridge: Cambridge University Press.

Ricardo, D. 1815. *An Essay on the Influence of a Low Price of Corn on the Profits of Stock*. London:

Murray.

Ricardo, D. 1817. *On the Principles of Political Economy and Taxation*. London: Murray.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*, 11 vols, ed. P. Sraffa. Cambridge: Cambridge University Press.

Robinson, J.V. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge MA: Harvard University Press.

Samuelson, P.A. 1948. *Economics: an Introductory Analysis*. New York: McGraw Hill.

Samuelson, P.A. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16, 1415–34.

Say, J.-B. 1803. *Traitè d'economie politique: ou, Simple exposition de la manière dont se forment, se distribuent et se consomment les richesses*. Paris: Dèterville.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen & Unwin.

Senior, N.W. 1836. *An Outline of the Science of Political Economy*. London: W. Clowes.

Sidgwick, H. 1883. *The Principles of Political Economy*. London: Macmillan.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vols. ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Oxford University Press, 1976.

Smith, V.E. 1956. Malthus's theory of demand and its influence on value theory. *Scottish Journal of Political Economy* 3, 205–20.

Sraffa, P. 1926. The laws of return under competitive conditions. *Economic Journal* 35, 535–50.

Steuart, J. 1767. *An Inquiry into the Principles of Political Œconomy: being an Essay on the Science of domestic Policy in free Nations. In which are particularly considered Population, Agriculture, Trade, Industry, Money, Coin, Interest, Circulation, Banks, Exchange, Public Credit, and Taxes*. 2 vols. London: Millar and Cadell.

Stewart, D. 1792, 1814, 1827. *Elements of the Philosophy of the Human Mind*, 3 vols. Reprinted in *The Collected Works of Dugald Stewart*, 11 vols. ed. W. Hamilton. Edinburgh: Constable, 1854–60.

Stigler, G. 1952. The Ricardian theory of value and distribution. *Journal of Political Economy* 60, 187–207.

Stigler, G. 1958. Ricardo and the 93% labor theory of value. *American Economic Review* 48, 357–67.

Sumner, J.B. 1816. *A Treatise on the Records of the Creation: with Particular Reference to Jewish History, and the Consistency of the Principle of Population with the Wisdom and Goodness of the Deity*. 2 vols. London: Hatchard.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. London: Hatchard.

Thornton, R.W.T. 1869. *On Labour: its Wrongful Claims and Rightful Dues, its Actual Present and Possible Future*. London: Macmillan.

Thünen, J.H. von. 1826. *Der isolirte Staat in Beziehung auf Landwirthschaft und Nationalökonomie*, part 1. Hamburg: Pethes.

Torrens, R. 1815. *An Essay on the External Corn Trade*. London: Longman et al.

Torrens, R. 1821. *An Essay on the production of Wealth: with an Appendix, in which the Principles of Political Economy are applied to the Actual Circumstances of this Country*. London: Longman et al.

Turgot, A.R.J. 1768. Observations sur le mémoire de M. de Saint-Péravy en faveur de l'impôt indirect. In *Écrits Économiques*, ed. B. Cazes. Paris: Calman-Lévey, 1970.

Walras, L. 1894. *Élèments d'èconomie politique pure: ou, Thèorie de la richesse sociale*. Lausanne: Rouge.

Walsh, V. and Gram, H. 1980. *Classical and Neoclassical Theories of General Equilibrium*. New York: Oxford University Press.

Waterman, A.M.C. 1991a. *Revolution, Economics and Religion: Christian Political Economy, 1798–1833*. Cambridge: Cambridge University Press.

Waterman, A.M.C. 1991b. The 'canonical classical model' in 1808 as viewed from 1825: Thomas Chalmers on the national resources. *History of Political Economy* 23, 221–41.

Waterman, A.M.C. 1995. Why William Paley was 'The first of the Cambridge Economists'. *Cambridge Journal of Economics* 20, 673–86.

Waterman, A.M.C. 1998. Reappraisal of ‘Malthus the economist’, 1933–97. *History of Political Economy* 30, 293–334.

Waterman, A.M.C. 1999. Hollander on the ‘canonical classical growth model’: a comment. *Journal of the History of Economic Thought* 21, 311–13.

Waterman, A.M.C. 2004. *Political Economy and Christian Theology since the Enlightenment*. Basingstoke: Palgrave Macmillan.

Wayland, Francis. 1837. *The Elements of Political Economy*. Boston: Gould and Lincoln.

West, E. 1815. *Essay on the Application of Capital to Land; with Observations Shewing the Impolicy of Any Great Restriction on the Importation of Corn … by a Fellow of University College, Oxford*. London: Underwood.

Whately, R. 1831. *Introductory Lectures in Political Economy*. London: Fellowes.

Wicksteed, P.H. 1894. *Essay on the Coordination of the Laws of Distribution*. London: Macmillan.

Winch, D.N. 1983. Science and the legislator: Adam Smith and after. *Economic Journal* 93, 501–20.

Winch, D.N. 1996. *Riches and Poverty. An Intellectual History of Political Economy in Britain, 1750–1834*. Cambridge: Cambridge University Press.

Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

How to cite this article

Waterman, A. M. C. "English School of political economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000355>
doi:10.1057/9780230226203.0480

The New Palgrave Dictionary of Economics Online

Enlightenment, Scottish

John Robertson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Scottish contribution to the Europe-wide intellectual movement of Enlightenment in the 18th century was unusually rich, covering moral philosophy, history, and political economy. It was not the simple product of the Union with England in 1707; more important were the gradual opening up of intellectual life and reform of the country's intellectual institutions, notably the universities, and economic growth, rapid by the last quarter of the century. The Scots set the investigation of economic phenomena in a broad framework; led by David Hume and Adam Smith, they were particularly interested in the comparative development prospects of rich and poor nations.

Keywords

Chalmers, T.; commercial society; division of labour; Enlightenment; Ferguson, A.; Galiani, F.; Genovesi, A.; Hume, D.; Hutcheson, F.; individual liberty; innovation; invisible hand; justice; Kames, Lord; Law, J.; luxury; Mandeville, B.; Melon, J.-F.; Millar, J.; natural jurisprudence; physiocracy; political economy; private property; protectionism; Pufendorf, S.; Quesnay, F.; Reid, T.; Robertson, W.; rule of law; Scottish Enlightenment; Shaftesbury, Lord; Smith, A.; specialization; stages theory of development; Steuart, Sir J.; Stewart, D.; subsistence; *Tableau économique*

Article

Between 1740 and 1790 Scotland provided one of the most distinguished branches of the European Enlightenment. David Hume and Adam Smith were the pre-eminent figures in this burst of intellectual activity; and around them clustered a galaxy of major thinkers, including Francis Hutcheson, Lord Kames, Adam Ferguson, William Robertson, Thomas Reid, Sir James Steuart and John Millar. The interests of individual thinkers ranged from metaphysics to the natural sciences; but the distinctive achievements of the Scottish Enlightenment as a whole lay in those fields associated with the enquiry into ‘the progress of society’ – history, moral and political philosophy and, not least, political economy.

‘Enlightenment’ and ‘Scottish Enlightenment’ were usages unknown in the 18th century: the term ‘Scottish Enlightenment’ was first coined in the early 20th century, and began to be generally used by historians in the 1960s. (*Lumières* and *Aufklärung* were in 18th-century use, but not to denote a European Enlightenment as a whole.) As a historian's construction, however, the term ‘Scottish Enlightenment’ is supported by the consciousness of those named above that they shared common intellectual interests (which did not preclude disagreement between them) and a common standing as men of letters in 18th-century Scottish society. This awareness of belonging to a broad intellectual movement extended to the continent of Europe: led by Hume, the Scottish thinkers cultivated connections with Paris, the Enlightenment's acknowledged metropolitan centre. But the Scottish Enlightenment is perhaps best understood when it is compared with the Enlightenment in Italy or in Germany. The concern with economic improvement and its moral and political conditions and consequences was as urgent, for instance, in the distant Kingdom of Naples as in Scotland; and political economy was equally absorbing to the Neapolitan philosophers Antonio Genovesi and Ferdinando Galiani.

At the same time, the experience of Scotland in the 18th century was distinctive in a number of respects, which offered a particular stimulus to Scottish thinkers. First of all, there was the actual achievement of economic growth. The late 17th-century Scottish economy supported an uneasy balance between population and food supply; bad harvests, which occurred in a sequence in the 1690s, could cause severe shortages and even localized famine. Overseas trade was likewise vulnerable. Nevertheless the elites, both landed and urban, were committed to economic development, and showed a marked propensity to invest. Agriculture gradually became commercialized, and landowners joined merchants to invest in manufactures, and, most spectacularly, in the ‘Darien venture’, intended to establish a Scottish trading colony in Panama. The failure of the latter persuaded many of the elite that economic development could only come through closer union with England. In the event, the economic fruits of the Union were disappointingly slow in coming; but by the third quarter of the 18th century it was clear to contemporaries that agriculture, trade and manufactures were all on an upward curve. The thinkers of the Scottish Enlightenment thus enjoyed an unusually direct acquaintance with the phenomena of economic development.

Scotland's political position was also unusual. Many of Europe's monarchies sought to bring their constituent kingdoms into closer union over the 18th century, for economic as well as administrative reasons. But none did so as successfully as the British monarchy. The Union of 1707 with England was in no simple or direct sense the cause of Scotland's economic growth (or of its Enlightenment). But it secured a common framework of law and a common market, and it also established that the Scottish Presbyterian and the English Anglican Churches should coexist in peace. These gains were important to the great majority of the Scottish elites, and it was never in their interest to back the Jacobite challenge to the Hanoverian monarchy.

Culturally and intellectually, the position of Scotland looked unpropitious before 1700. There were pockets of interest in the new science, Newton having a group of Scottish adherents; but the latest developments in French philosophy were shunned for their Epicurean, materialist and sceptical tendencies. After the Revolution of 1688, however, change gradually got under way in the institutions most important for intellectual life, making possible the infiltration of new ideas. The fierce, covenanting Presbyterianism of the 17th century was dissipated, as the ‘Moderate’ group of clergy rose to power in the Kirk. The universities of Edinburgh, Glasgow, Aberdeen and St Andrews were reformed,

allowing professorial specialization; and around the universities there developed a vigorous informal culture of voluntary clubs, most famous of which was the Select Society of Edinburgh, founded by David Hume and his friends in 1754. Together these changes secured for Scottish thinkers unprecedented intellectual freedom and social support; and they provided an object lesson in the importance of the moral and cultural as well as the material dimensions of progress.

The intellectual interests which distinguished the Scottish Enlightenment had two more specific sources. One was the explicit preoccupation with the conditions and means of economic development which was fostered by the debate which preceded the Union of 1707. The preoccupation was by no means unique to the Scots, but the contributions of John Law (the future author of the French Mississippi Scheme) and others ensured a high quality of discussion. The other, two decades later, was the initiative taken by two very different philosophers, Francis Hutcheson and David Hume, to transform the agenda by which philosophy was taught and discussed in Scotland. Drawing on the moral philosophy of Shaftesbury and the natural jurisprudence of Pufendorf, Hutcheson taught his Glasgow students, who included Adam Smith, a moderate, benevolent, providential Stoicism. More disturbingly, Hume drew on the scepticism of Pierre Bayle and the Epicurean morals of Bernard Mandeville to offer in his *Treatise of Human Nature* (1739–40) and his two later *Enquiries* (1748; 1751) an account of justice and morals which had no need of divine support. Most of those now associated with the Scottish Enlightenment found Hutcheson's philosophy more congenial; but it was Hume's challenge which galvanized them. It was Hume, moreover, who turned their attention back to economic matters. Recognizing that philosophy alone would never make the Scots into virtuous atheists, Hume decided instead to educate them in political economy, the subject of the leading essays in his *Political Discourses* of 1752.

For Hume as for all the Scottish thinkers, political economy was not a science apart. It belonged within a wider enquiry into the ‘progress of society’. There were three principal dimensions to this enquiry: the historical, the moral and the political.

The historical theory of the Scottish Enlightenment developed a line of argument from later 17th-century natural jurisprudence, a tradition made familiar to the Scots by its incorporation in the moral philosophy curriculum of the reformed universities. Discarding the older jurisprudential thesis of the contractual foundations of society and government, the Scots focused on the new insights of Pufendorf and Locke into the origin and development of property. According to Pufendorf, there had never been an original state of common ownership of land and goods; from the first, property was the result of individual appropriation. As increasing numbers made goods scarce, individual property became the norm, and systems of justice and government were established to secure it. What the Scots added to this argument was a scheme of specific stages of social development, the hunting, the pastoral, the agricultural and the commercial. At each of the four stages the extent of property ownership was related to the society's means of subsistence, and these shaped the nature and sophistication of the society's government. Different versions of the theory were offered by Adam Ferguson in his *Essay on the History of Civil Society* (1767) and by John Millar in his *Origin of the Distinction of Ranks* (1770), and it underlay both Lord Kames's investigations into legal history and William Robertson's historical narratives. The locus classicus of the theory, however, was Adam Smith's Lectures on Jurisprudence, delivered to his students in Glasgow in the early 1760s.

As Smith's exposition makes particularly clear, the stages theory of social development provided the historical premises for political economy. An explicitly conjectural theory – a model of society's ‘natural’ progress – it provided a framework for a comparably theoretical treatment of economic

development as ‘the natural progress of opulence’. By positing the systematic interrelation of economic activity, property and government, with consequences which could be neither foreseen nor controlled by individuals, the theory also underlined the limits of effective government action. ‘Reason of state’, the standby of rulers and their advisers for over two centuries, still had the capacity to distort and obstruct the economic activity of subjects and those with whom they would trade; but the Scots’ historical perspective showed it to be a doctrine inadequate to the complexity of a modern commercial economy.

The moral thought of the Scottish Enlightenment was closely related to the historical, sharing a common origin in 17th-century natural jurisprudence. Here the inspiration was the jurisprudential thinkers’ increasingly sophisticated treatment of needs. These, it was recognized, could no longer be thought of primarily in relation to subsistence; with the progress of society, needs must be understood to cover a much wider range of scarce goods, luxuries as well as necessities. The potential of this insight was seen by every Scottish moral philosopher, but again it was Smith who exploited it to the full, in the *Theory of Moral Sentiments* (1759). Beyond the most basic necessities, Smith acknowledged, men's needs were always relative, a matter of status and emulation, of bettering one's individual condition. But it was precisely the vain desires of the rich and the envy of others which served, by ‘an invisible hand’, to stimulate men's industry and hence to increase the stock of goods available for all ranks.

Such an argument, however, had to overcome two of the most deeply entrenched convictions of European moral thought: the Aristotelian view that the distribution of goods was a matter for justice, and the classical or civic humanist view that luxury led to corruption and the loss of moral virtue. The Scots answered the first more confidently (but perhaps less satisfactorily) than the second. Following Grotius, Hobbes and Pufendorf, they defined justice in exclusively corrective terms, setting aside questions of distribution. On the issue of corruption, they were divided. Hume, who ridiculed fears of luxury, was the most confident; Ferguson, who defiantly reasserted the ancient ideal of virtue, was the most pessimistic. Smith was closer to Hume in preferring propriety to virtue, at least for the great majority; but he showed that he shared Ferguson's doubts when he added, at the end of his life, that the disposition to admire the rich and the great did tend to corrupt moral sentiments. At a fundamental level, however, there was general agreement. As a consequence of the progress of society, the multiplication of needs was not only irreversible; it was the essential characteristic of a ‘cultivated’ or ‘civilized’ as distinct from a ‘barbarian’ society. And civilization, however morally ambiguous, was preferable to barbarism. With consensus on this, the moral premises of political economy were secure.

The definition of justice in simple corrective terms provided the starting-point for the political dimension of the Scottish enquiry. The priority of any government, the Scots believed, must be the security of life and property, ensuring every individual liberty under the law. This, as Smith put it, was freedom ‘in our present sense of the word’; and there was a general confidence that it was tolerably secure under the governments of modern Europe, including the absolute monarchies. In principle, individual liberty was a condition of a fully commercial society: its provision, therefore, was the institutional premise of political economy.

Few of the Scots took their analysis beyond this relatively simple, if vital, point; the theory of the modern commercial state was not a Scottish achievement. Both Hume and Smith were more concerned to limit the opportunities for enlarging government at the expense of ‘productive’ society, by confining the former to the minimum necessary provision of justice, defence and public works. But they also recognized that the proliferation of interests in a commercial society would require more sophisticated institutional mechanisms to ensure their adequate representation within the political system. Smith's

analysis in Book IV of the *Wealth of Nations* of the growing alienation of the colonial elites in North America from parliamentary authority was an object lesson in the need for such representation – and a strong hint that it was incompatible with maintaining an extended empire.

A large part of the originality of the Scottish Enlightenment's conception of political economy lay in this exploration of the historical, moral and institutional framework of economic activity. But of course the Scots also engaged directly in economic analysis; and one such work of analysis, Adam Smith's *Wealth of Nations* (1776), would so outshine all others that it came to be regarded as having established political economy as a science in its own right.

The Scots' attention focused on growth in a context of international rivalry. In contemporary terms, Hont has shown, the issue was the means by which poor countries (of which Scotland might be regarded as one) could best hope to catch up on rich countries (such as England certainly was). What is striking is the hard-headedness with which Hume and Smith tackled the issue. Responding to French economists – Hume to Jean-François Melon, Smith to the Physiocrats – who argued that agriculturally endowed countries should follow a different path from purely commercial nations, the Scots insisted that one analysis applied to all. Protection for agricultural economies and their manufactures, a policy supported by the former Jacobite exile Sir James Steuart in his *Principles of Political Economy* (1767), was futile and damaging. But theirs was no naive optimism in the equalizing powers of commerce. The ideal of *doux commerce*, by which trade would be the agent of global peace and prosperity, was as much of a panacea as the belief that commercial success would be self-cancelling, because the advantage of low labour costs would always pass on to others. Instead, Hume and Smith suggested that rich countries could expect to maintain their advantage over poorer ones, whether by flexible specialization and product innovation (Hume) or by constantly increasing industrial productivity through the division of labour (Smith). What distinguished commercial superiority from military conquest was that it was achieved 'without malice'; poor countries would also develop if they followed the same route, even if they might never catch up on the rich.

Brilliant as Hume's economic essays were, it was Adam Smith's *Wealth of Nations* (1776) which set the standard of Enlightenment political economy. To be systematic and comprehensive had earlier been the ambition, at least, of Quesnay's *Tableau Economique* (1758–9), Genovesi's *Lezioni di Commercio* (1765) and Steuart's *Principles;* but the *Wealth of Nations* eclipsed them all. Its success, moreover, was such as to suggest that political economy had an identity all of its own. Smith himself did not admit such an implication, continuing to insist that political economy was but 'a branch of the science of a statesman or legislator': his own engagement with both jurisprudence and moral philosophy left him disinclined to drop the wider intellectual framework in which political economy had been conceived. But a work at once as extensive and as self-contained as the *Wealth of Nations* made it at least plausible to suppose that what it presented was a distinct, autonomous science of political economy.

Smith's death in 1790 coincided with the end of the Scottish Enlightenment. In Scotland as throughout Europe, the French Revolution transformed the conditions and assumptions of intellectual life, while political economy had to come to terms with the increasingly obvious impact of machinery. Within Scotland Dugald Stewart set himself to adapt the Enlightenment conception of political economy to these new circumstances; but while he had French admirers, his expansive, didactic approach had few followers in Britain. Another Scot, Thomas Chalmers, took the lead alongside Malthus in attaching political economy to newly urgent theological concerns, while Ricardo and his followers simply took a narrower view of the subject. Even so, it would be a mistake to see 19th-century classical political

economy as a new departure. As the philosophical analysis of Hegel (who learnt much from Steuart) and the radical critiques of Marx and the early socialists pointed out, the historical, moral and institutional premises on which political economy rested were still those elucidated by the Scots.

See Also

- Hume, David
- Hutcheson, Francis
- Mandeville, Bernard
- Pufendorf, Samuel von
- Smith, Adam
- Steuart, Sir James
- Stewart, Dugald

Bibliography

Berry, C.J. 1997. *Social Theory and the Scottish Enlightenment*. Edinburgh: Edinburgh University Press.

Broadie, A. ed. 2003. *The Cambridge Companion to the Scottish Enlightenment*. Cambridge: Cambridge University Press.

Devine, T.M. 1994. *The Transformation of Rural Scotland: Social change and the Agrarian Economy 1660–1815*. Edinburgh: Edinburgh University Press.

Hont, I. and Ignatieff, M. eds. 1983. *Wealth and Virtue: The Shaping of Political Economy in the Scottish Enlightenment*. Cambridge: Cambridge University Press.

Hont, I. 2005. *Jealousy of Trade: International Competition and the Nation-State in Historical Perspective*. Cambridge, MA: Harvard University Press.

Phillipson, N.T. 1981. The Scottish Enlightenment. In *The Enlightenment in National Context*, ed. R. Porter and M. Teich. Cambridge: Cambridge University Press.

Robertson, J.C. 2005. *The Case for the Enlightenment. Scotland and Naples 1680–1760*. Cambridge: Cambridge University Press.

Sakamoto, T. and Tanaka, H. eds. 2003. *The Rise of Political Economy in the Scottish Enlightenment*. London: Routledge.

Sher, R.B. 1985. *Church and University in the Scottish Enlightenment*. Princeton and Edinburgh:

Princeton: Princeton University Press; Edinburgh: Edinburgh University Press.

Sher, R.B. 2006. *The Enlightenment and the Book. Scottish Authors and their Publishers in Eighteenth-Century Britain, Ireland and America*. Chicago and London: Chicago University Press.

Winch, D. 1996. *Riches and Poverty: An Intellectual History of Political Economy in Britain 1750–1834*. Cambridge: Cambridge University Press.

How to cite this article

Robertson, John. "Enlightenment, Scottish." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000049> doi:10.1057/9780230226203.0481

The New Palgrave Dictionary of Economics Online

enterprise zones

Leslie E. Papke

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Enterprise zones are geographically targeted economic development incentives used in the United States by individual states since the early 1980s and the federal government since 1993. Research on state zone programmes that accounts for the endogeneity of zone designation finds little improvement in the employment and incomes of zone residents, but some evidence that firms respond to tax incentives for capital. In contrast, the federal empowerment zone programme combines tax incentives with local initiatives and access to large federal grants. Recent research on round one of the federal programme finds mixed evidence on zone resident employment.

Keywords

enterprise zones; investment subsidies; labour subsidies; tax competition; tax credits; tax incentives

Article

Enterprise zone programmes are geographically targeted tax, expenditure, and regulatory inducements used by US state and local governments since the early 1980s and by the federal government since 1993. While they differ in their specifics, all the programmes provide development incentives, including tax preferences to capital and/or labour, in an attempt to induce private investment location or expansion to depressed areas and to enhance employment opportunities for zone residents. Most enterprise zones are designated in urban areas, but there are some rural zones. Typically, state and local zone programmes provide larger tax credits for business investment than for employment incentives. investment incentives include the exemption of business-related purchases from state sales and use taxes, investment tax credits and corporate income or unemployment tax rebates. Labour subsidies include employer tax credits for all new hires or zone-resident new hires, employee income tax credits and job-training tax credits. Some programmes assist firms financially with investment funds or industrial development bonds.

Enterprise zones have been criticized as ineffective and inefficient in stimulating new economic activity.

This criticism is part of a long-standing debate on the effects of intersite tax differentials on the location of capital investment. It is argued that if tax-induced investment represents only relocation from another state, then tax competition is a zero-sum game for the country as a whole. In addition, the preferential treatment of certain types of investment or employment within enterprise zones may induce decisions that would not be economically sound in the absence of the tax incentives. Often, however, redistribution of economic activity within a state may be a desirable goal. If investment is relocated from local labour markets with low unemployment to local labour markets with higher unemployment, the incentives may generate efficiency gains for the economy as underutilized resources are tapped (Bartik, 1991). Efficiency gains may also result if reductions in unemployment produce positive externalities, such as reductions in social unrest.

A partial equilibrium model predicts that a labour subsidy or an equal-cost subsidy to both zone capital and zone resident labour will raise zone wages. A capital subsidy alone may actually reduce zone wages – yet many of the subsidies are for capital investment in the zone (Gravelle, 1992; Papke, 1994).

Empirical evaluations of zone programmes typically measure the amount of investment undertaken after the designation, for example, or the increase in the number of firms in the zone, and the change in zone employment. Two key methodological issues in empirical evaluations are (*a*) to separate the effects of zone designation from jobs and investments arising from other factors – for example, general upswings in the economy; (*b*) to account for the depressed economic characteristics that led to the initial zone designation. If zone sites are better randomly selected, the effect of the programme can be measured by comparing the performance of the experimental and control groups. But zone designation in the 43 state and local programmes in the United States depends on comparative unemployment rates, population levels and trends, poverty status, median incomes, and percentage of welfare recipients, so the data are non-experimental. This sample selection problem can be addressed with a variety of econometric techniques.

Econometric analysis of a zone's success faces a practical difficulty in that conventional economic data are not available by zone. In most states, zones do not coincide with census tracts or taxing jurisdictions. As a result, zone areas cannot be pinpointed in standard data collections. Zip code level data is available from the Census, but outcome measures are ten years apart.

Econometric evaluations of the Indiana and New Jersey programmes find mixed effects on investment and employment. Indiana zones are estimated to have greater inventory growth and fewer unemployment claims than they would have in the absence of the zone designation (from 1983 to 2006, an inventory tax credit was the most lucrative incentive). However, in the 1980s, inventory investment came at the cost of a drop in the value of depreciable property (Papke, 1994). Moreover, despite the reduction in unemployment rates in the zones, a comparison of incomes from the 1980 and 1990 Censuses suggests that zone residents are not appreciably better off after the first decade of the Indiana zone programme (Papke, 1993) and there is no discernable increase in capital investment or land values (Papke, 2001). Similar econometric analysis of the New Jersey enterprise zone programme finds no positive effects on either business investment or employment (Boarnet and Bogart, 1996). Multi-state econometric analyses that combine data from many states – thereby assuming zone programmes have similar effects in every state – typically find no positive zone effects on business activity or employment (Bondonio, 2003; Bondonio and Engberg, 2000). Peters and Fisher (2002) survey state evaluations.

Cost-per-job estimates from zone programmes are rare. The literature also lacks a discussion of the

distribution of the cost of the zone programme between state and local governments. For example, local governments may bear the brunt of the cost of a state enterprise zone programme if tax incentives are provided against local taxes without state reimbursement.

Congress established the Empowerment Zone and Enterprise Community (EZ/EC) programme in 1993 and the Renewal Community (RC) programme in 2000 to provide assistance to the nation's distressed communities. By 2007, there had been three rounds of EZs, two rounds of ECs, and one round of RCs leading to a total of 40 empowerment zones (30 urban and 10 rural), 95 enterprise communities (65 urban, 30 rural) and 40 renewal communities.

Empowerment zone incentives include a 20 per cent employer wage credit for the first 15,000 dollars of wages for zone residents who work in the zone, additional expensing of equipment investments of qualified zone businesses, and expanded tax exempt financing for certain zone facilities. Each zone is eligible for 100 million dollars in Social Services Block Grant funds. Selected areas needed to demonstrate pervasive poverty, unemployment and general distress, and applicants had to outline a plan of action that included local business and community interests. The residence-based approach of the income tax credit differs significantly from another federal programme designed to increase employment of the disadvantaged. The Targeted Jobs Tax Credit provides firms with a similar-sized subsidy for wages paid to targeted individuals – primarily welfare recipients and poor youth. Providing a subsidy based on individual characteristics may create a stigma that actually reduces the probability of being hired. Residence-based eligibility may eliminate this problem and encourage individuals who become employed to continue to live in the zone.

Features of the programmes have changed over time. Round I and II EZs and ECs received different combinations of grant funding and tax benefits. By round III, EZs and the RCs received mainly tax benefits. The GAO (1991; 2004; 2006) reports that Round I and II EZs and ECs are continuing to access their grant funds and Internal Revenue Service (IRS) data show that businesses are claiming some tax benefits (Brashares, 2000). However, the IRS does not collect data on other tax benefits and cannot always identify the communities in which they were used. The lack of tax benefit data limits evaluation of the programmes.

Evaluation of the federal programme is also confounded by its hybrid structure. The federal EZ/EC programme is based on the idea that effective community revitalization results when the strategy is tailored to the local site. The diverse nature of the Round I EZ/ECs – each may differ in terms of objective, size of targeted area, type of designation, governance structure, projects used, grant money, and strategies for implementation – has made it difficult to generate general conclusions about even the early stages of Round I implementation (GAO, 2004; 2006). Further, the tax incentives changed over the three rounds of the federal programme. Third, no easy method of data collection was included in the tax forms so even usage is hard to measure.

Using Census data, Hanson (2007) finds no effect of the first round zone programme on local employment or poverty rates in the targeted areas, but instead finds capitalization into property values. Busso and Kline (2006) find modest improvements in labour market conditions, but sizable increases in owner-occupied housing values and rents along with small changes in the demographic composition of neighbourhoods. Taken together, these two papers suggest that improvements for residents have been limited at best, but that property owners have benefited from the federal programme.

See Also

- economic development and the environment
- fiscal federalism
- local public finance
- public finance
- regional development, geography of
- taxation and poverty

Bibliography

Bartik, T.J. 1991. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Boarnet, M.G. and Bogart, W.T. 1996. Enterprise zones and employment: evidence from New Jersey. *Journal of Urban Economics* 40, 198–215.

Bondonio, D. and Engberg, J. 2000. Enterprise zones and local employment: evidence from the states' programs. *Regional Science and Urban Economics* 30, 519–49.

Brashares, E. 2000. Empowerment zone tax incentive use: what the 1996 data indicate. *Statistics of Income Bulletin*.

Busso, M. and Kline, P. 2006. Do local economic development programs work? Evidence from the federal empowerment zone program. Mimeo, University of Michigan.

Engberg, J. and Greenbaum, R. 1999. State enterprise zones and local housing markets. *Journal of Housing Research* 10, 163–87.

GAO (General Accounting Office). 1991. Businesses' use of empowerment zone incentives. RCED-99–253. US Government Accounting Office, Washington, DC.

GAO. 2004. Community development: federal revitalization programs are being implemented, but data on the use of tax programs are limited. RCED 04–306, US Government Accounting Office, Washington, DC.

GAO. 2006. Empowerment zone and enterprise community program: improvements occurred in communities but the effect of the program is unclear. RCED-06–727, US Government Accounting Office, Washington, DC.

Gravelle, J.G. 1992. Enterprise zones: the design of tax incentives. CRS Report for Congress 92–476 S.

Congressional Research Service, Library of Congress, Washington, DC.

Hanson, A. 2007. Poverty reduction and local employment effects of geographically targeted tax incentives: an instrumental variables approach. Mimeo, Syracuse University.

HUD (U.S. Department of Housing and Urban Development). 1992. *State Enterprise Zone Update: Summaries of the State Enterprise Zone Programs*. Washington, DC: U.S. Department of Housing and Urban Development.

Papke, L.E. 1993. What do we know about enterprise zones? In *Tax Policy and the Economy*, vol. 7, ed. J.M. Poterba. Cambridge, MA: MIT Press.

Papke, L.E. 1994. Tax policy and urban development: evidence from the Indiana enterprise zone program. *Journal of Public Economics* 54, 37–49.

Papke, L.E. 2001. The Indiana enterprise zone revisited: effects on capital investment and land values. National Tax Association Proceedings of the Ninety-Third Annual Conference, National Tax Association, Washington DC.

Peters, A.H. and Fisher, P.S. 2002. *State Enterprise Zone Programs: Have They Worked?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

How to cite this article

Papke, Leslie E. "enterprise zones." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000281> doi:10.1057/9780230226203.0482

The New Palgrave Dictionary of Economics Online

entitlements in laboratory experiments

Sheryl Ball
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

altruism; behavioural economics; desert; entitlement programmes; entitlements; entitlements in laboratory experiments; fair allocation; justice; other-regarding behaviour; procedural fairness; psychological games; Rawls, J.; reciprocity; self-interest

Article

Entitlements are rights granted by contract, law or practice. Under the assumption of pure self-interest, modelling games with entitlements is fairly straightforward; however, work in behavioural economics has consistently demonstrated the existence of other-regarding preferences, with strong effects of perceptions of what is fair. In the laboratory, behaviour is affected not only by the entitlement per se but also by the procedure by which entitlements come about. One form of laboratory entitlement is a more advantageous position in an economic game, where the advantage arises from a larger endowment, favourable exchange rules or greater decision-making authority. A second type of entitlement is a guaranteed payoff or a payoff floor. Experimental results show that the means by which entitlements are acquired is one cue that influences the nature of other-regarding behaviour. This is important both for understanding behaviour and the design of experiments.

In early experimental work on entitlements, Hoffman and Spitzer (1985) demonstrate that both the existence of an entitlement and its source determine economic outcomes. They study bilateral bargaining problems where one of the two subjects, called the ‘controller’, has unilateral authority to decide the outcome of a negotiation game in the event of disagreement. Authority is assigned based on either the outcome of a coin flip or the result of a simple test of a skill that is irrelevant to the experimental task. They find that controllers are most willing to exploit their power when they are assigned their role based on the skill test and are told that they ‘earned’ the right to be the controller – that is, that they have moral authority. These results are consistent with Burrows and Loomes (1994).

The subjects’ behaviour illustrates Rawls's (1971) notion of ‘desert’, which requires that people deserve the conditions underlying their actions as well as the fruits of their actions. Thus subjects divided an

endowment equally when the controller was chosen according to the flip of a coin and had low moral authority. On the other hand, both earning the right to be controller and higher moral authority triggered changes in observed allocations, so that outcomes favoured the controller. Entitlements that were earned or that involved ‘morally unequal’ agents were sufficient to trigger unequal outcomes. Equity theory developed by social psychologists is similar in spirit to this theory of justice.

Ideas of procedural fairness also affect perceptions of government entitlements. Fong (2001) looks at poll data on perceptions of poverty and opportunity, and finds that beliefs about others’ effort, luck and opportunity play the largest role in determining support for government entitlement programmes. In particular these beliefs outweigh concerns about tax costs in supporting these programmes. These results are consistent with the experimental results discussed above, where low payoffs are acceptable if one displays low effort. If one's situation is determined by poor luck, however, one will give up some of one's earnings to increase the earnings of others.

A number of experimental studies on income redistribution examine Rawls's claim that individuals prefer an income redistribution rule that maximizes the position of the poorest member of society (Frohlich and Oppenheimer, 1990). Studies where subjects must choose a principle of distributive justice and a tax system in addition to participating in a production task find that people choose rules that maximize the productivity of society while maintaining a minimum floor for the worst off members. Subjects generate greater output in experiments where they are able to determine the entitlements for the worst off individual in their group, again demonstrating that the source of entitlements matters.

These results show that researchers need to pay attention to how entitlements are determined. This is a complication for theories of behavioural economics or psychological games. People do not have a pure taste for fair allocations; they are more self-interested, altruistic or fair according to circumstances that depend on how advantage arises. This behaviour is closely related to reciprocity, but that is often modelled as ‘if you are nice to me I'll be nice to you’ (Bowles and Gintis, 2001). In contrast, this collection of results can be interpreted as, ‘I will respect your entitlement if you deserve it’.

A preference for procedural factors also complicates experimental design, since subjects behave in a more self-interested manner when entitlements are earned than when they are randomly assigned. Researchers must be careful to consider how subjects will interpret the rules by which advantages are assigned or they may risk introducing nuisance variables. Future work might deliberately award entitlements in a manner that subjects view as unjust to see whether that produces yet another pattern of behaviour.

See Also

- behavioural game theory
- Coase theorem
- experimental economics
- fair allocation
- justice
- psychological games

Bibliography

Burrows, P. and Loomes, G.. 1994. The impact of fairness on bargaining behavior. *Empirical Economics* 19, 201–21.

Bowles, S. and Gintis, H. 2001. The inheritance of economic status: education, class and genetics. In *Genetics, Behavior and Society*, ed. M. Feldman. In *International Encyclopedia of the Social and Behavioral Sciences*, ed. N. Smelser and P. Baltes. Oxford: Elsevier.

Fong, C. 2001. Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics* 82, 225–46.

Frohlich, N. and Oppenheimer, J.A. 1990. Choosing justice in experimental democracies with production. *American Political Science Review* 84, 461–77.

Hoffman, E. and Spitzer, M.L. 1985. Entitlements, rights and fairness: an experimental examination of subjects' concepts of distributive justice. *Journal of Legal Studies* 14, 259–97.

Lissowski, G., Tyszka, T. and Okrasa, W. 1991. Principles of distributive justice: experiments in Poland and America. *Journal of Conflict Resolution* 35, 98–119.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap, Harvard University Press.

How to cite this article

Ball, Sheryl. "entitlements in laboratory experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000314> doi:10.1057/9780230226203.0483

The New Palgrave Dictionary of Economics Online

entrepreneurship

William J. Baumol and Melissa A. Schilling

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes the recent expansion of research on entrepreneurship, innovation and growth. Although the entrepreneur is widely credited with critical contributions to innovation and growth, the subject of entrepreneurship has virtually disappeared from mainstream theory and standard textbooks. Reasons explaining this gap are indicated. In addition to some brief materials on earlier writings, the rich body of recent work on the subject, both theoretical and empirical, is surveyed, illustrating the wide variety of subjects explored and the insights offered by the new literature.

Keywords

Baumol, W.; Cantillon, R.; economic growth; entrepreneurship; equilibrium; innovation; Kirzner, I.; Knight, F.; Marshall, A.; Mill, J.S.; neoclassical theory; opportunity costs; patents; product development; property rights; risk; Say, J.-B.; Schumpeter, J.; self-employment; social status; transfer of technology; uncertainty premium; venture capital

Article

An entrepreneur is an individual who organizes, operates, and assumes the risk of creating new businesses. There are two types. A replicative entrepreneur organizes a new business firm that is like other firms already in existence. An innovating entrepreneur provides something new – a new product or process, or a new type of business structure, a new approach to marketing, and so on. These innovations need not be productive or beneficial. For example, Richard Cantillon (one of the first great economic theorists) spoke of thieves who are entrepreneurs (Cantillon, 1730, pp. 54–5). And Joseph A. Schumpeter, arguably the contributor of the most important analysis of entrepreneurship, included as an entrepreneurial act '…the creation of a monopoly position (for example, through trustification) ….' (Schumpeter, 1911, p. 66). Entrepreneurs (interpreted as the self-employed) are estimated to constitute about seven percent of the labour force in the United States (U.S. Bureau of the Census,

2004). Most of them are probably replicative, not innovative, entrepreneurs.

It is widely agreed that the entrepreneur plays an important role in economic growth. But the evidence shows little correlation between an economy's number of replicative entrepreneurs and its growth rate. Innovative entrepreneurs do make a substantial difference to a nation's growth rate, having introduced many breakthrough innovations like the telephone and the airplane. The primary social contribution of replicative entrepreneurship is as a means for individuals to escape poverty, because such undertakings require little capital, education or experience. Still, the data show that entrepreneurs, on average, earn less than employees with similar education and experience (Freeman, 1978; Astebro, 2003; Benz and Frey, 2004).

Although economists have recently exhibited a resurgence of interest in entrepreneurship, the entrepreneur nevertheless rarely shows up in contemporary mainstream economic theory.

Early writings and the origin of the term

Until the 20th century, writings in English referred to entrepreneurs as ‘adventurers’ or ‘undertakers’ (see, for example, Marshall, 1923, p. 172). Apparently, the term ‘entrepreneur’ was introduced by Cantillon in the French translation of his great work, *Essai Sur la Nature de Commerce en Général* (1730, p. 54), but what is apparently *his* English text uses the word ‘undertaker’. The early writings on entrepreneurship were descriptive rather than theoretical. Cantillon's discussion (1730, ch. 11) is brief, focusing on replicative entrepreneurs: ‘…wholesalers in Wool and Corn, Bakers, Butchers, Manufacturers and Merchants of all kinds….’ (1730, p. 51). Cantillon's main point, like that of Frank H. Knight (1921), was the task's riskiness: ‘These Undertakers can never know how great will be the demand in their City, nor how long their customers will buy of them since their rivals will try all sorts of means to attract customers from them. All this causes so much uncertainty among these Undertakers that every day one sees some of them become bankrupt’ (1730, p. 51).

Nearly a century later, Jean-Baptiste Say's (1819) discussion is still brief, but richer. Say seems interested primarily in innovating entrepreneurs, dealing with three types of ‘producers’: scientists, entrepreneurs and labourers. Using mechanical locks as an example, the scientist investigates ‘…the properties of iron, the method of extracting from the mine and refining the ore…’ The entrepreneurs deal with ‘…application of this knowledge to a useful purpose…,’ while the third group – the workers – actually make the product (1819, p. 80). And any successful economy needs all three: ‘Nor can [industry] approximate to perfection in any nation, till that nation excel in all three branches’ (1819, p. 80).

Thus, Say blames poverty in Africa on the absence of scientists and entrepreneurs. Lack of entrepreneurs alone can undercut prosperity, even with scientific knowledge abundant, for without the entrepreneur, ‘…that knowledge might possibly have lain dormant in the memory of one or two persons, or in the pages of literature’ (1819, p. 81). This is precisely the explanation that one of the present authors proposed for the failure of medieval China and the Soviet Union to translate an abundance of non-military inventions into viable consumer products (Baumol, 2002, chs. 5, 14). Say also foreshadows some of Schumpeter's analysis (see below): ‘In manufacture…if success [in innovation] ensue, the adventurer is rewarded by a longer period of exclusive advantage, because his process is less open to observation’ (1807, p. 84).

Finally, Say mentions the spillovers of innovation and their justification for governmental financing: ‘The charges of experiment, when defrayed by the government… [are] hardly felt at all, because the burthen is divided among innumerable contributors; and the advantages resulting from success being a common benefit to all, it is by no means inequitable that the sacrifices, by which they are obtained, should fall on the community at large’ (1819, p. 85).

Before Schumpeter's breakthrough (see below), the subject was touched upon by economists like J.S. Mill, Alfred Marshall and (a bit later) Knight. Generally, their focus was not on innovative entrepreneurship, and they emphasized management's directing of going concerns rather than establishment of new firms. (But Marshall, 1923, p. 172, does digress briefly to mention Matthew Boulton's significant role as an entrepreneur dealing with James Watt's inventions.) Today, however, these discussions would hardly be considered theory. Rather, they are usually narratives containing illuminating observations. They assert that the entrepreneur's payment is a residual after other inputs are compensated, and that compensation is determined by the entrepreneur's ability and the supply of entrepreneurship in the market. They note that entrepreneurs employ themselves, so that unlike other inputs there is no demand function, as for other inputs.

Disappearance of the entrepreneur from modern mainstream economics

Given the acknowledged importance of the entrepreneur's role, it could be hoped that modern theoretical economics, with its powerful analytic tools, would have produced an extensive entrepreneurship analysis. Instead, the opposite happened – the entrepreneur became the ‘invisible man’ in mainstream theory. There are at least two reasons for this. First, the most advanced and powerful microeconomic models predominately study timeless static equilibria. But, for the entrepreneur, the transition process is the heart of the story. Schumpeter (1911) shows the entrepreneur as a destroyer of equilibria by constant innovation, while Israel Kirzner (1979) tells how the alert entrepreneur seeks out the arbitrage opportunities presented by disequilibria, thereby moving the economy back toward equilibrium. Such a relentless attack upon both equilibria and disequilibria does not fit a stationary model from which firm creation and invention are excluded.

The second reason for the entrepreneur's disappearance from mainstream theory is that, by definition, an invention is something never available before. So invention is the ultimate heterogeneous product. This impedes the optimality analysis underlying most microeconomic theory. Explicitly or implicitly, an optimality calculation entails a comparison among possible substitute choices, while the innovating entrepreneur normally deals with no well-defined substitutes with quantifiable attributes. In contrast, the standard theory of the firm analyses repetitious decisions of management in fully operational enterprises where the entrepreneur has already completed his job and left to create other firms.

Thus, neoclassical theory is justified in excluding the entrepreneur, because it deals with subjects for which the entrepreneur is irrelevant. That does not mean that no theory of entrepreneurship is needed, or that such a theory is lacking, but it means that a theory of entrepreneurship must be sought elsewhere, and that is what Schumpeter succeeded in doing.

Brief summary: Schumpeter's model – the supply and earnings of entrepreneurial activity

The basic Schumpeterian model (1911) notes that the successful innovative entrepreneur's reward is profit temporarily exceeding that of perfect competition. This attracts rivals who seek to share those profits by imitating the innovation, and thereby erode its super-competitive earnings. To prevent termination of these rewards, the entrepreneur can never desist from further innovation and cannot rest on his laurels.

Perhaps most important, the Schumpeterian analysis shows how the entrepreneur is driven to work without let-up for economic growth. Thus, it clearly reveals the tight association between innovative entrepreneurship and growth.

Allocation between productive and unproductive entrepreneurship

Some work of one of the present authors (Baumol, 2002, ch. 14) tells much of the rest of the story about the supply and allocation of productive entrepreneurship and the key role of evolving institutions. In the economic growth literature, it has often been asserted that an expanded supply of entrepreneurs effectively stimulates growth, while shrinkage in the supply undermines growth. But the standard explanation of the entrepreneurs' appearance and disappearance is shrouded in mystery, with hints about cultural developments and vague psychological and sociological changes. The historical evidence suggests a more mundane explanation: that entrepreneurs are always present but, as the structure of rewards in the economy changes, entrepreneurs switch their activities, moving to where payoffs become more attractive. In doing so, they move in and out of the activities usually recognized as entrepreneurial, exchanging them for other activities that also require enterprising talent but are often distant from production of goods and services. The generals of ancient Rome, the Mandarins of the Tang, Sung, and Ming Chinese empires, the captains of late medieval private and mercenary armies, the rent-seeking contemporary lawyers, and the Mafia Dons – all are clearly enterprising and often successful. And when institutions have changed so as to modify profoundly the relative payoffs offered by the different enterprising activities, the supply of entrepreneurs has shifted accordingly. Here, it is helpful to distinguish two categories of entrepreneurs, the productive and the unproductive entrepreneurs, with the latter, in turn, divided into subgroups such as rent-seeking entrepreneurs and destructive entrepreneurs, including the organizers of private armies or criminal groups. Once there is a pertinent change in the institutions that govern the relative rewards, the entrepreneurs will shift their activities between productive and unproductive occupations, so the set of productive entrepreneurs will appear to expand or contract autonomously. For example, when institutions change to prohibit private armies, entrepreneurs are led to look elsewhere to realize their financial ambitions. If, simultaneously, rules against confiscation of private property and for patent protection of inventions are adopted, entrepreneurial talent will shift into productive, innovative directions.

Recent studies: other disciplines and empirical approaches

Outside mainstream economic analysis, research on entrepreneurship has expanded rapidly since the 1980s, particularly that by specialists in management, psychology, and sociology. We focus here on three streams of work that have attracted the most scholarly attention: (*a*) how differences among individuals influence entry into (and success in) entrepreneurship, (*b*) how environment influences

entrepreneurship, and (*c*) the strategies and forms of organization used by entrepreneurs.

Differences among individuals

There are numerous studies investigating how differences among individuals (in attributes such as education, age, experience, social position and psychology) are associated with a propensity to become self-employed and the likelihood of success at entrepreneurship. A wide variety of studies have indicated that individuals with higher education than the general population are more likely to become entrepreneurs (Shane, 2003). Robinson and Sexton (1994) and others have found that number of years of education is significantly related to likelihood of becoming self-employed, and Bates (1995) found that individuals with a graduate education were significantly more likely to become self-employed. Age appears to have an inverted U-shaped relationship with likelihood of forming a new venture. Entrepreneurship first increases with age because of experience, and then decreases with age because of opportunity costs and uncertainty premiums (Bates, 1995; Shane, 2003).

A number of studies that look at how experience influences likelihood of starting a business and the success of the new venture have found that general business experience (Evans and Leighton, 1989; Robinson and Sexton, 1994), experience specific to the industry in which the entrepreneur later founds a business (Aldrich, 1999), and prior self-employment (Carroll and Mosakowski, 1987) all increase the likelihood that an individual will found a new business. Furthermore, such experience tends to improve new venture performance and survival rates (Gimeno et al., 1997).

Studies have revealed that, in general, social status increases the likelihood of forming a new venture (for example, Stuart, Huang and Hybels, 1999). The number and diversity of an individual's social ties also increase the likelihood of founding a company (Aldrich, Rosen and Woodward, 1987), as well as the success of the venture (Hansen, 1995). Psychological factors also influence an individual's likelihood of becoming an entrepreneur (Shane, 2003). In particular, extraversion (Babb and Babb, 1992), need for achievement (Hornaday and Aboud, 1973), risk-taking propensity (Astebro, 2003), self-efficacy (Zietsma, 1999), overconfidence (Arabsheibani et al., 2000), and creativity (Ames and Runco, 2005) have all been shown to be significantly related to an individual's likelihood of becoming an entrepreneur.

Environmental factors

A number of industry characteristics influence new venture formation. Market size (Pennings, 1982) and growth (Dean and Meyer, 1992) increase the likelihood of new firm formation, while uncertainty from technological change decreases the rate of business start-ups (Audretsch and Acs, 1994). Capital intensity also reduces new firm formation by raising entry costs (Dean and Meyer, 1992). The density of firms has an inverted U-shaped relationship with new firm formation (Carroll and Wade, 1991). Too few firms in an industry may signal that there is no opportunity worth pursuing, or scarcity of market information. Thus, initial increases in the density of firms in the industry encourage business start-ups (Shane, 2003), although high density can increase competition for resources and create an entry barrier. Not surprisingly, the institutional environment of an industry or region also affects new firm formation.

Capital availability (for example, low-cost debt or venture capital) enhances firm formation (McMillan and Woodruff, 2002). Higher marginal federal income tax rates decrease self-employment (Gentry and Hubbard, 2000) and business tax concessions increase business start-ups (Dana, 1987). Stronger property rights encourage entrepreneurship, presumably because they assure entrepreneurs that they can appropriate the fruits of their efforts (McMillan and Woodruff, 2002). Researchers have also investigated the role of university technology-transfer offices on entrepreneurship, with most research indicating that such offices increase rates of new venture formation, particularly when technology-transfer offices are structured to profit from the transfers (Markman et al., 2005). Finally, socio-cultural norms about the desirability of self-employment or the risks of failure are significantly related to rates of business start-ups in a nation or ethnic group (Butler and Herring, 1991).

Strategy and organization

The area of entrepreneurial strategy that has received most research attention is method of financing. Consistent with Knight's (1921) argument that self-financing is needed to overcome moral hazard problems, most entrepreneurs finance their ventures primarily with their own capital (Aldrich, 1999; Shane, 2003). However, funds provided by ‘angel’ investors (wealthy individuals who invest in entrepreneurial companies, usually at an early stage) and venture capitalists are also important. The research on angel investment is sparse, but there is more research on venture capitalst investment. A number of researchers have investigated how venture capitalists choose their investments, mitigate risk, and influence new venture survival and growth (Bygrave and Timmons, 1992). Some studies have also examined how entrepreneurs identify opportunities (Shane, 2003), their degree of reliance on patent protection (Shane, 2001), the effect of entrepreneurs’ new product development strategies (Zahra and Bogner, 2000), and their breadth of market focus (Bhide, 2000; Gimeno et al., 1997).
Finally, there also has been some research on the organization of new ventures – how they are formed as legal entities, the performance implication of this choice (Delmar and Shane, 2004), and the effect of venture team size and background (Eisenhardt and Schoonhoven, 1990). In general, formation as a legal entity and a large, diverse venture team appear to improve new venture performance.

On the state of the theory of entrepreneurship

Our discussion demonstrates that the beginnings of a significant theory of entrepreneurship already exist. The analysis uses little mathematics to derive any formal theorems, and its results are primarily qualitative. But this nascent theory of entrepreneurship does tell us about its supply and earnings, its role in the pricing of its products and the role of the price mechanism in its allocation among alternative activities. The Schumpeterian model tells us about the determination of entrepreneurs’ profits and the prices of their products, as well as their influence on the supply of their activity. The model of productive and unproductive entrepreneurship tells us more about supply, as well as about the allocation of this resource. The empirical research adds further insight into the factors that increase the likelihood of individuals engaging in, and being successful at, entrepreneurship.
Beyond the stationary analysis of standard microeconomic theory, we see that the entrepreneurship

models enable us to deal with such important questions as what features of the structure of the free market economy have caused it to outperform by an order of magnitude the innovation and growth of any alternative economic system. The institutional changes that reallocated much of entrepreneurship from redistributive to productive activities are, according to the model, the key to the answer. And this has profound policy implications both for developing countries seeking desperately to escape their poverty and for developed economies seeking to keep up the pace of their growth.

See Also

- Cantillon, Richard
- growth and institutions
- intellectual property
- Knight, Frank Hyneman
- Schumpeterian growth and growth policy design

Bibliography

Aldrich, H. 1999. *Organizations Evolving*. London: Sage.

Aldrich, H., Rosen, B. and Woodward, W. 1987. The impact of social networks on business foundings and profit: a longitudinal study. In *Frontiers of Entrepreneurship Research*, ed. N. Churchill et al. Babson Park, MA: Babson College.

Ames, M. and Runco, M. 2005. Predicting entrepreneurship from ideation and divergent thinking. *Creativity and Innovation Management* 14, 311–15.

Arabsheibani, G., De Meza, D., Maloney, J. and Pearson, B. 2000. And a vision appeared unto them of a great profit: evidence of self-deception among the self-employed. *Economics Letters* 67, 35–41.

Astebro, T. 2003. The return to independent invention: evidence of unrealistic optimism, risk seeking or skewness loving. *Economic Journal* 113, 226–38.

Audretsch, D. and Acs, Z. 1994. New firm startups, technology, and macroeconomic fluctuations. *Small Business Economics* 6, 439–49.

Babb, E. and Babb, S. 1992. Psychological traits of rural entrepreneurs. *Journal of Socio-Economics* 21, 353–62.

Bates, T. 1995. Self-employment entry across industry groups. *Journal of Business Venturing* 10, 143–56.

Baumol, W. 2002. *The Free-Market Innovation Machine: Analyzing the Growth Miracle of Capitalism*. Princeton, NJ: Princeton University Press.

Benz, M. and Frey, B. 2004. Being independent raises happiness at work. *Swedish Economic Policy Review* 11, 95–134.

Bhide, A. 2000. *The Origin and Evolution of New Businesses*. New York: Oxford University Press.

Butler, J. and Herring, C. 1991. Ethnicity and entrepreneurship in America: toward an explanation of racial and ethnic group variations in self-employment. *Sociological Perspectives* 34, 79–95.

Bygrave, W. and Timmons, J. 1992. *Venture Capital at the Crossroads*. Boston, MA: Harvard Business School Press.

Cantillon, R. 1730. *Essai Sur la Nature de Commerce en Général*, trans. H. Higgs. London: Macmillan, 1931.

Carroll, G. and Mosakowski, E. 1987. The career dynamics of self employment. *Administrative Science Quarterly* 32, 570–89.

Carroll, G. and Wade, J. 1991. Density dependence in organizational evolution of the American brewing industry across different levels of analysis. *Social Science Research* 20, 271–302.

Dana, L. 1987. Entrepreneurship and value creation – an international comparison of five commonwealth nations. In *Frontiers of Entrepreneurship Research*, ed. N. Churchill et al. Babson Park, MA: Babson College.

Dean, T. and Meyer, G. 1992. New venture formation in manufacturing industries: a conceptual and empirical analysis. In *Frontiers of Entrepreneurship Research*, ed. N. Churchill et al. Babson Park, MA: Babson College.

Delmar, F. and Shane, S. 2004. Legitimating first: organizing activities and the survival of new ventures. *Journal of Business Venturing* 19, 385–410.

Eisenhardt, K. and Schoonhoven, K. 1990. Organizational growth: linking founding team, strategy, environment, and growth among U.S. semiconductor ventures, 1978–1988. *Administrative Science Quarterly* 35, 504–29.

Evans, D. and Leighton, L. 1989. Some empirical aspects of entrepreneurship. *American Economic Review* 79, 519–35.

Freeman, R. 1978. Job satisfaction as an economic variable. *American Economic Review* 68, 135–41.

Gentry, W. and Hubbard, R. 2000. Tax policy and entrepreneurial entry. *American Economic Review Papers and Proceedings* 90, 283–92.

Gimeno, J., Folta, T., Cooper, A. and Woo, C. 1997. Survival of the fittest? Entrepreneurial human capital and the persistence of underperforming firms. *Administrative Science Quarterly* 42, 750–83.

Hansen, E. 1995. Entrepreneurial networks and new organization growth. *Entrepreneurship Theory and Practice* 19(4), 7–19.

Hornaday, J. and Aboud, J. 1973. Characteristics of successful entrepreneurs. *Personnel Psychology* 24, 141–53.

Kirzner, I. 1979. *Perception, Opportunity and Profit*. Chicago: University of Chicago Press.

Knight, F. 1921. *Risk, Uncertainty and Profit*. Boston and New York: Houghton Mifflin Company.

Markman, G., Phan, P., Balkin, D. and Gianiodis, P. 2005. Entrepreneurship and university-based technology transfer. *Journal of Business Venturing* 20, 241–63.

Marshall, A. 1923. *Industry and Trade*. London: Macmillan.

McMillan, J. and Woodruff, C. 2002. The central role of entrepreneurs in transition economies. *Journal of Economic Perspectives* 16(3), 153–70.

Pennings, J. 1982. Organizational birth frequencies: an empirical investigation. *Administrative Science Quarterly* 27, 120–44.

Robinson, T. and Sexton, E. 1994. The effect of education and experience on self-employment success. *Journal of Business Venturing* 9, 141–56.

Say, J.-B. 1819. *Traite d'économie politique*. 4th edn, trans. C. Prinsep. Boston: Wells and Lilly, 1821.

Schumpeter, J. 1911. *The Theory of Economic Development*, trans. R. Opie. Cambridge, MA: Harvard University Press, 1934.

Shane, S. 2001. Technology opportunities and new firm creation. *Management Science* 47, 205–20.

Shane, S. 2003. *A General Theory of Entrepreneurship: The Individual-Opportunity Nexus*. Northampton, MA: Edward Elgar.

Stuart, T., Huang, H. and Hybels, R. 1999. Interorganizational endorsements and the performance of entrepreneurial ventures. *Administrative Science Quarterly* 44, 315–49.

U.S. Bureau of the Census. 2004. *Statistical Abstract of the United States: 2004–2005*. Washington, DC: Bureau of the Census.

Zahra, S. and Bogner, W. 2000. Technology strategy and software new ventures' performance: exploring the moderating effect of the competitive environment. *Journal of Business Venturing* 15, 135–73.

Zietsma, C. 1999. Opportunity knocks – or does it hide? An examination of the role of opportunity recognition in entrepreneurship. In *Frontiers of Entrepreneurship Research*, ed. P. Reynolds et al. Babson Park, MA: Babson College.

How to cite this article

Baumol, William J. and Melissa A. Schilling. "entrepreneurship." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000274>
doi:10.1057/9780230226203.0484

The New Palgrave Dictionary of Economics Online

envelope theorem

Eugene Silberberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The envelope theorem appeared in economics following the 1931 Viner–Wong diagram (incorrectly drawn in the original paper). This famous paper indicated that, starting at some minimum cost input combination, the change of average cost when output changed was the same whether or not other inputs were allowed to vary or were held fixed. This puzzling result remained mostly a curiosity until the 1970s when, with the use of a generalization of this diagram, the modern theory of duality was developed. This new approach to comparative statics provided a clearer explanation for the appearance of refutable implications in maximization models.

Keywords

comparative statics; constrained optimization models; cost functions; envelope theorem; Lagrange multipliers; primal–dual model; shadow pricing

Article

The origin of this famous theorem is the discussion between Jacob Viner (1931) and his draughtsman Y.K. Wong concerning the relationship between short- and long-run average cost curves. Viner had apparently reasoned that since in the long run average costs should be at a minimum, the long-run average cost (LRAC) curve should not only always be below the short-run average cost (SRAC) curves, but should also pass through the minimum points of each short-run curve. Wong pointed out the impossibility of this joint occurrence, and Viner opted to draw the long-run curve through the minimum points, thereby necessarily passing above sections of the short run curves. It was also puzzling (in the now corrected diagram) that at the point of tangency between the LRAC and a SRAC, the rate of change of average cost with respect to output was the same when capital was fixed as when it was allowed to vary. The puzzle was solved by Samuelson (1947), who showed in a general way why the long-run curve would be the ‘envelope’ curve to the set of short-run curves. Perhaps the most surprising result of all was that this seeming mathematical curiosity turned out to be the fundamental basis for the development of refutable comparative statics implications in economics.

Unconstrained maximization models

The most general comparative statics model with explicit maximizing behaviour is *maximize* $y=f(x, \alpha)$ subject to $g(x, \alpha)=0$, where $x=(x_1,\ldots,x_n)$ is a vector of decision variables, $\alpha = (\alpha_1,\ldots, \alpha_m)$ is a vector of parameters (though for simplicity, we treat α as a scalar in the discussion below), and $g(\cdot)$ represents one or more constraints. Models at this level of generality, however, imply no refutable implications and are hence largely uninteresting. In particular, there are never refutable implications for parameters that enter the constraint (see, for example, Silberberg and Suen, 2000). We therefore initially restrict the analysis to models of unconstrained maximization:

$$\text{maximize } y = f(x, \alpha) \tag{1}$$

The necessary first-order conditions (NFOC) are

$$f_i(x, \alpha) = 0 \quad i = 1, \ldots, n \tag{2}$$

The sufficient second-order conditions (SSOC) are that the *Hessian* matrix $\mathbf{H}=(f_{ij})$ is negative definite. Alternatively, the principal minors of order (size) k of the Hessian determinant $H = |f_{ij}|$ have sign $(-1)^k$. Assuming the sufficient second-order conditions hold, we can in principle 'solve' for the n explicit choice functions $x=x^*(\alpha)$. Of course, since these choice functions are the result of solving the NFOC simultaneously, each individual x_i is a function of *all* the parameters, not just ones which might appears in some f_i.

Substituting the x_i^*'s into the objective function yields the *indirect objective function* $\varphi(\alpha)=f(x^*(\alpha), \alpha)$, the maximum value of f for given α. Since $\varphi(\alpha)$ is by definition a maximum value, $\Phi(\alpha) \geq f(x, \alpha)$, but $\varphi(\alpha)=f(x, \alpha)$ when $x=x^*$. In Figure 1, a typical $\varphi(\alpha)$ is plotted. For an arbitrary α^0, an $x^0 = x^*(\alpha^0)$ is implied. Consider the behaviour of $f(x, \alpha)$ when the x_i's are held fixed at x^0 as opposed to when they are variable. When $\alpha = \alpha^0$, the 'correct' x_i's are chosen, and therefore $\varphi(\alpha)=f(x^0, \alpha)$ at that one point. However, both to the left and to the right of α^0, the 'wrong' (that is non-maximizing) x_i's are chosen, and, since $\varphi(\alpha)$ is the *maximum* value of f for given α, $f(x^0, \alpha) \leq \Phi(\alpha)$ in any neighbourhood around α^0. This implies that φ and f must be tangent at α^0 (assuming differentiability), and, moreover, f must be either more concave or less convex than φ there. Since this must happen for arbitrary α, similar tangencies occur at other values of α. It is apparent from Figure 1 that $\varphi(\alpha)$ is the *envelope* of the $f(x_1,x_2, \alpha)$'s for each α. What surprised most researchers was the discovery that all comparative statics theorems in maximization models are in fact consequences of the relative curvatures of φ and f.

Figure 1

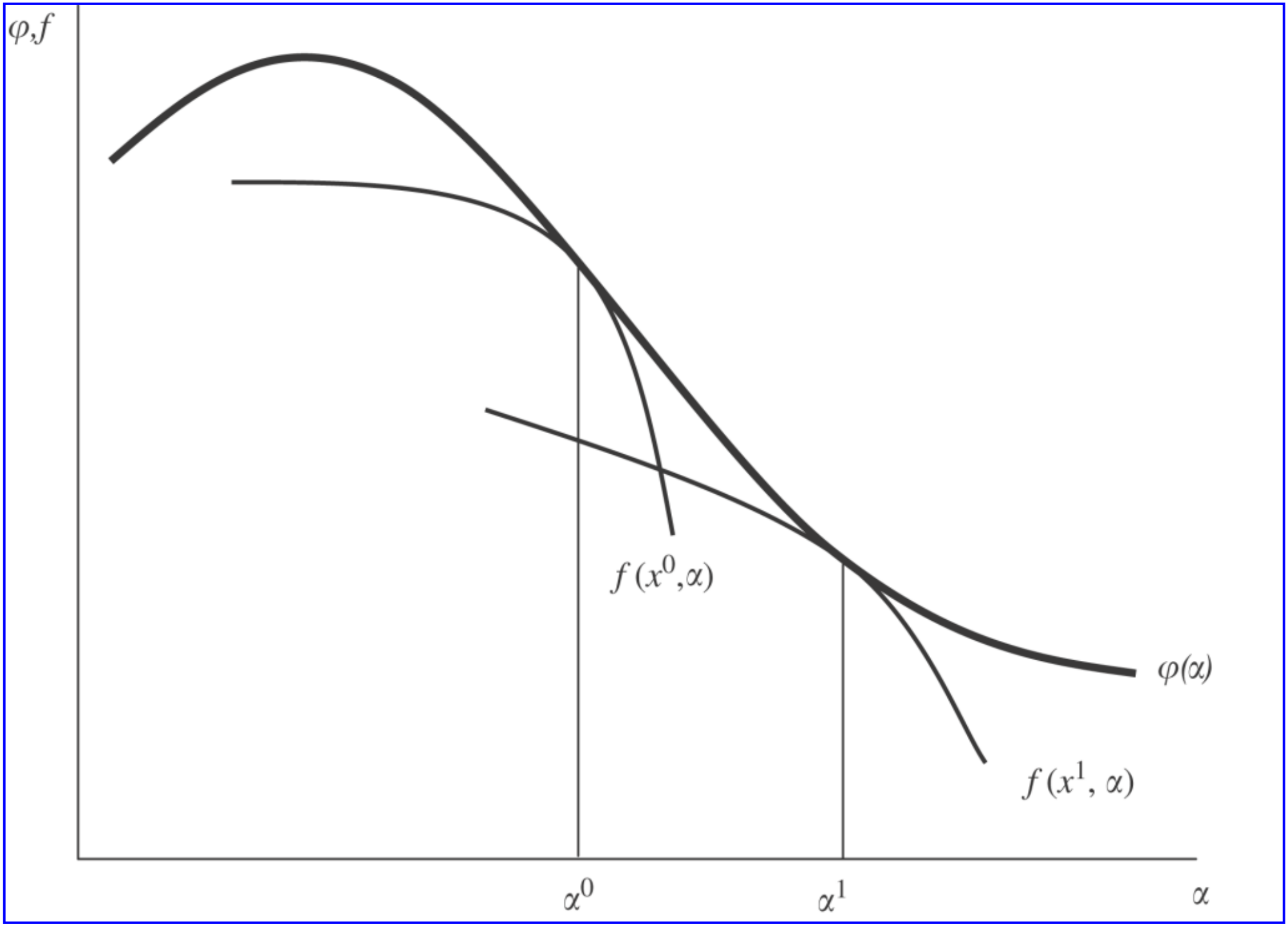

From the above discussion, the function $F(x, \alpha)=f(x, \alpha)-\varphi(\alpha)$ has a maximum of zero, with respect to both x and α. Thus we consider the *primal–dual* model

$$\text{maximize } F(x, \alpha) = f(x, \alpha) - \phi(\alpha) \tag{3}$$

where the maximization runs over x and also α. (In the latter instance, we ask, for given x_i's, what values of the parameters would make these x_i's the maximizing values?) The NFOC with respect to x are the same as in the original model. With respect to α, the NFOC yield the famous 'envelope theorem' which is the tangency of f and φ in Figure 1:

$$F_\alpha = f_\alpha - \phi_\alpha = 0 \tag{4}$$

In the α dimensions, the second-order conditions are simply

$$F_{\alpha\alpha} = f_{\alpha\alpha} - \phi_{\alpha\alpha} \leq 0. \tag{5}$$

This inequality says that in the α dimensions, f is relatively more concave than φ. (When α is a vector, this second-order condition is that the Hessian matrix ($F_{\alpha\alpha}$) is negative semi-definite.)
This is the fundamental geometrical property that underlies all comparative statics relationships. The NFOC (4) are identities when $x=x^*$. That is,

$$\phi_\alpha(\alpha) \equiv f_\alpha(x^*(\alpha), \alpha) \tag{6}$$

Differentiating with respect to α,

$$\phi_{\alpha\alpha} \equiv \sum_1^n f_{\alpha i} \frac{\partial x_i^*}{\partial \alpha} + f_{\alpha\alpha} \tag{7}$$

Rearranging terms, using (5) and invariance to the order of differentiation,

$$\phi_{\alpha\alpha} - f_{\alpha\alpha} \equiv \sum_{1}^{n} f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0$$
(8)

This is the fundamental relation of comparative statics. From it, we can derive Samuelson's famous ‘conjugate pairs’ theorem that refutable implications occur in maximization models when and only when a parameter enters one and only one first-order condition. For in that case, where say α enters only f_i=0, $f_{j\alpha} \equiv 0$, $j \neq i$, and so (8) reduces to one term:

$$f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0$$
(9)

In this case we can say that the response of x_i is in the same direction as the disturbance to the equilibrium (or, in the case of minimization models, in the opposite direction). For example, consider the profit-maximization model

$$maximize\ \pi = f(x, w, p) = p\theta(x_1 \ldots, x_n) - \sum w_i x_i$$

Each parameter w_i enters only the ith NFOC, and $f_{x_i w_i} = -1$, so that (9) yields the slope property $\partial x_i^* / \partial w_i \leq 0$; the factor demand functions are downward sloping in their own price.
The envelope theorem also yields the non-intuitive ‘reciprocity’ conditions. Suppose there are two parameters α and β. Then from invariance of second partial derivatives to the order of differentiation (Young's theorem), $\varphi_{\alpha\beta} = \varphi_{\beta\alpha}$. Using equation (6) above,

$$\sum f_{i\alpha} \frac{\partial x_i^*}{\partial \beta} = \sum f_{i\beta} \frac{\partial x_i^*}{\partial \alpha}$$
(10)

When the objective function contains a linear expression such as in the profit maximization model, that is, $w_1x_1+\cdots+w_nx_n$, we have $f_{x_i w_i} = -1$ and $f_{x_i w_j} = 0, \quad i \neq j$. In that case, (11) reduces to the simple expression $\frac{\partial x_i^*}{\partial w_j} = \frac{\partial x_j^*}{\partial w_i}$. This result also occurs in consumer theory for the Hicksian demands.

Constrained maximization models

Consider now the general comparative statics model with constraints, *maximize* $y=f(x, \alpha)$ subject to $g(x, \alpha)=0$, where $g(\cdot)$ represents one or more constraints. Assuming just one constraint for the moment, the Lagrangian for this model is $L=f(x, \alpha)+\lambda g(x)$, producing the NFOC

$$L_i = f_i(x, \alpha) + \lambda g_i(x, \alpha) = 0 \quad i = 1, \ldots, n$$
(11)

$$L_\lambda = g(x, \alpha) = 0$$
(12)

Assuming the SSOC, we can in principle 'solve' for the n+1 explicit choice functions $x=x^*(\alpha)$ and $\lambda^*(\alpha)$. We derive the indirect objective function as before by substituting the x_i^*'s into the objective function producing $\varphi(\alpha)=f(x^*(\alpha), \alpha)$, the maximum value of f for given α, now also subject to the constraint. Proceeding as above, since $\varphi(\alpha)$ is by definition a maximum value, $\phi(\alpha) \geq f(x, \alpha)$, but $\varphi(\alpha)=f(x, \alpha)$ when $x=x^*$. Thus the function $F(x, \alpha)=f(x, \alpha)-\varphi(\alpha)$ has a (constrained) maximum of zero, with respect to both x and α. Thus we consider the *primal–dual* model

$$\textit{maximize } F(x, \alpha) = f(x, \alpha) - \phi(\alpha)$$
(13)

$$\textit{subject to } g(x, \alpha) = 0$$
(14)

where the maximization runs over x and also α. The Lagrangian for this model is

$$L = f(x, \alpha) - \phi(\alpha) + \lambda g(x, \alpha)$$
(15)

The first-order conditions with respect to x are the same as in the original model. With respect to α, we get the envelope theorem in its most general form,

$$\phi_\alpha = L_\alpha = f_\alpha + \lambda g_\alpha$$
(16)

At this level of generality, it is not possible to generate any useful curvature properties of $\varphi(\alpha)$. However, consider the case where α does not enter any constraint. In that case, $g_\alpha \equiv 0$ and the NFOC reduce to (4) above, that is, $F_\alpha = f_\alpha - \varphi_\alpha = 0$. Moreover, when α does not enter the constraint, the primal–dual model is *an unconstrained maximization in* α. Hence in the α dimensions, the second-order conditions are as before:

$$F_{\alpha\alpha} = f_{\alpha\alpha} - \phi_{\alpha\alpha} \leq 0. \tag{17}$$

Thus in this important class of models, the comparative statics are identical to the models with no constraints. We obtain the inequalities (8) and (9) in the same manner as above. Consider now an important class of models having the structure maximize $f(x)$ subject to $g(x)=k$, where we suppress all parameters except k, which is the focus of this analysis. The Lagrangian for this model is $L=f(x)+\lambda(k-g(x))$; assuming the NFOC and SSOC are valid, we solve for the explicit choice functions $x=x^*(k)$ and $\lambda^*(k)$. The indirect objective function is $\varphi(k)=f(x^*(k))$, the maximum value of f for given k. The envelope theorem (16) yields

$$\phi_k = \lambda^*(k) \tag{18}$$

Suppose the function f represents the value of output, and the constraint describes a limitation on that value due to the scarcity of some resource, measured by the value of k. Then the Lagrange multiplier imputes a ‘shadow price’, a marginal evaluation of that resource, since $\lambda^*(k)$ is the rate of change of the maximum value of output with respect to a change in the availability of that resource. This is a very widespread use of Lagrangian analysis in economics. For example, the fundamental model from which we derive the cost curves for a firm is, *minimize* $C=\sum w_i x_i$ subject to $f(x)=y$, where y is a parameter. Using (17), the Lagrange multiplier in this model is the marginal cost function $\partial C^*/\partial y=\lambda^*(w,y)$.
To further show the powerful nature of this analysis, consider the two-factor, two-goods model that plays an important part of international trade theory:

$$\begin{aligned}
&\textit{maximize } NNP = p_1 y_1 + p_2 y_2 \\
&\textit{subject to:} \\
&y_1 = f^1(L_1, K_1) \quad y_2 = f^2(L_2, K_2) L_1 + L_2 = L K_1 + K_2 = K
\end{aligned}$$

where f^1 and f^2 are production functions using labour (L) and capital (K) in each of two industries with outputs y_1 and y_2; output prices p_1 and p_2 and labour and capital endowments L and K are parametric. We can enumerate the salient properties of this model just by inspection, using the above results. The Lagrangian for this model is

$$L = p_1 y_1 + p_2 y_2 + \lambda_1(f^1(L_1, K_1) - y_1) + \lambda_2(f^2(L_2, K_2) - y_2) + \lambda_L(L - L_1 - L_2) + \lambda_K(K - K_1 - K_2).$$

Assuming the NFOC and SSOC hold, we solve the NFOC for the output supply functions $y_1^*(p_1, p_2, L, K)$ and $y_2^*(p_1, p_2, L, K)$, and the Lagrange multipliers, particularly $\lambda_L^*(p_1, p_2, L, K)$ and $\lambda_K^*(p_1, p_2, L, K)$. Substituting $y_1^*(\cdot)$ and $y_2^*(\cdot)$ into the objective function, we get the maximum value of *NNP* for given prices and resource constraints, φ (p_1,p_2,L,K). Since prices enter the objective function only, and in the classic linear form, (9) immediately yields the envelope relations $\phi_{p_i} = y_i^*(\cdot)$. We also note $\phi_L = \lambda_L^*(\cdot)$ *and* $\phi_K = \lambda_K^*(\cdot)$. The primal–dual model is, *maximize* $F = p_1 y_1 + p_2 y_2 - \phi(p_1, p_2, L, K)$ subject to the same constraints above. Since p_1 and p_2 do not enter the constraints, F is concave in p_1 and p_2. Since the first two terms are linear and φ enters negatively, φ is convex in p_1 and p_2, and thus $\phi_{p_i p_i} = \partial y_i^* / p_i > 0$; the supply curves are upward sloping. Furthermore, from (17), the Lagrange multipliers λ_L^* and λ_K^* are the imputed values of labour and capital. If an additional increment of labour, say, became available, λ_L^* would represent its marginal value product, and hence its implied wage in a competitive economy. Without further assumptions (for example, concavity of the production functions), we cannot determine a sign for how these imputed values change when the resource endowment changes: $\partial \lambda_L^* / \partial L / 0$. The reciprocity relationships are straightforward: $\phi_{p_1 p_2} = \partial y_1^* / \partial p_2 = \partial y_2^* / \partial p_1 = \phi_{p_2 p_1}$, and similarly, $\phi_{LK} = \partial \lambda_L^* / \partial K = \partial \lambda_K^* / \partial L = \phi_{KL}$. We also find $\phi_{p_1 L} = \partial y_1^* / \partial L = \partial \lambda_L / \partial p_1 = \phi_{L p_1}$, and so on. It seems unlikely that Jacob Viner could have imagined what the corrected version of his diagram would eventually lead to!

See Also

- cost functions
- duality
- Hicksian and Marshallian demands
- Le Chatelier principle

Bibliography

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Silberberg, E. and Suen, W. 2000. *The Structure of Economics*, 3rd edn. New York: McGraw-Hill.

Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3, 23–46. Repr. in American Economic Association, *Readings in Price Theory*, Homewood, IL: Irwin, 1952.

How to cite this article

Silberberg, Eugene. "envelope theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000095> doi:10.1057/9780230226203.0485

The New Palgrave Dictionary of Economics Online

environmental economics

Robert N. Stavins

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An overview is provided of the economics of environmental policy, including the setting of goals and targets, notably the Kaldor–Hicks criterion and the related method of assessment known as benefit–cost analysis. Also reviewed are the means of environmental policy, that is, the choice of specific policy instruments, featuring an examination of potential criteria for assessing alternative instruments, with focus on cost-effectiveness. The theoretical foundations and experiential highlights of individual instruments are reviewed, including conventional command-and-control mechanisms and market-based instruments.

Keywords

asymmetric information; averting behaviour; benefit–cost analysis; bequest value; Coase, R.; command-and-control instruments; contingent valuation (CV); cost-effectiveness; efficiency; environmental economics; existence value; general equilibrium analysis; hedonic pricing methods; hedonic wage method; insurance premium taxes; Kaldor–Hicks criterion; market-based instruments; net present value (NPV) analysis; non-use value; Pareto, V.; partial equilibrium analysis; Pigou, A.; Pigouvian tax; pollution charges; random utility models; recreation; reservation price; revealed preference; revenue cycling; risk reduction; social discount rate; tax differentiation; tradable permits; travel cost method; use value; value of a statistical life (VSL); willingness to accept; willingness to pay

Article

The fundamental theoretical argument for government activity in the environmental realm is that pollution is an externality – an unintended consequence of market decisions which affect individuals other than the decision maker. Providing incentives for private actors to internalize the full costs of their actions was long thought to be the theoretical solution to the externality problem. The primary advocate of this view was Arthur Pigou, who in *The Economics of Welfare* (1920) proposed that the government

should impose a tax on emissions equal to the cost of the related damages at the efficient level of control. A response to the Pigouvian perspective was provided by Ronald Coase in ‘The problem of social cost’ (1960). Coase demonstrated that, in a bilateral bargaining environment with no transaction costs, wealth or income effects, or third-party impacts, two negotiating parties will reach socially desirable agreements, and the overall amount of pollution will be independent of the assignment of property rights. At least some of the specified conditions are unlikely to hold for most environmental problems. Hence, private negotiation will not – in general – fully internalize environmental externalities.

Criteria for environmental policy evaluation

More than 100 years ago Vilfredo Pareto (1896) enunciated the well-known normative criterion for judging whether a social change makes the world better off: a change is *Pareto efficient* if at least one person is made better off and no one is made worse off. This criterion has considerable normative appeal, but virtually no public policies meet the test. Nearly 50 years later Nicholas Kaldor (1939) and John Hicks (1939) postulated a more pragmatic criterion that seeks to identify ‘potential Pareto improvements’: a change is welfare-improving if those who gain from the change could – in principle – fully compensate the losers, with (at least) one gainer still being better off.
The Kaldor–Hicks criterion – a test of whether total social benefits exceed total social costs – is the theoretical foundation for the use of the analytical device known as benefit–cost (or net present value) analysis. If the objective is to maximize the difference between benefits and costs (net benefits), then the related level of environmental protection (pollution abatement) is defined as the efficient level of protection:

$$\max_{\{q_i\}} \sum_{i=1}^{N} [B_i(q_i) - C_i(q_i)] \rightarrow q_i^* \qquad (1)$$

where q_i is abatement by source i ($i = 1$ to N), $B_i(\cdot)$ is the benefit function for source i, $C_i(\cdot)$ is the cost function for the source, and q_i^* is the efficient level of protection (pollution abatement). The key necessary condition that emerges from the maximization problem of equation (1) is that marginal benefits be equated with marginal costs (on the assumption of convexity of the respective functions).
The Kaldor–Hicks criterion is clearly more practical than the strict Pareto criterion, but its normative standing is less solid. Some have argued that other factors should be considered in a measure of social well-being, and that criteria such as distributional equity should trump efficiency considerations in some collective decisions (Sagoff, 1993). Many economists would agree with this assertion, and some have noted that the Kaldor–Hicks criterion should be considered neither a necessary nor a sufficient condition for public policy (Arrow et al., 1996).

Benefit – cost analysis of environmental regulations

The soundness of empirical benefit–cost analysis rests upon the availability of reliable estimates of social benefits and costs, including estimates of the social discount rate. The present value of net benefits (PVNB) is defined as:

$$PVNB = \sum_{t=0}^{T} \{(B_t - C_t) \cdot (1 + r)^{-t}\} \quad (2)$$

where B_t are benefits at time t, C_t are costs at time t, r is the discount rate, and T is the terminal year of the analysis. A positive PVNB means that the policy or project has the potential to yield a Pareto improvement (meets the Kaldor–Hicks criterion). Thus, carrying out benefit–cost or ‘net present value’ (NPV) analysis requires discounting to translate future impacts into equivalent values that can be compared. In essence, the Kaldor–Hicks criterion provides the rationale both for benefit–cost analysis and for discounting (Goulder and Stavins, 2002).

Choosing the discount rate to be employed in an analysis can be difficult, particularly where impacts are spread across a large number of years involving more than a single generation. In theory, the social discount rate could be derived by aggregating the individual time preference rates of all parties affected by a policy. Evidence from market behaviour and from experimental economics indicates that individuals may employ lower discount rates for impacts of larger magnitude, higher discount rates for gains than for losses, and rates that decline with the time span being considered (Cropper, Aydede and Portney, 1994; Cropper and Laibson, 1999). In particular, there has been support for the use of hyperbolic discounting and similar approaches with declining discount rates over time (Ainslie, 1991; Weitzman, 1994; 1998), but most of these approaches are subject to time inconsistency.

The costs of environmental regulations

In the environment context, the economist's notion of cost (or, more precisely, opportunity cost) is a measure of the value of whatever must be sacrificed to prevent or reduce the risk of an environmental impact. A full taxonomy of environmental costs ranges from the most obvious to the least direct (Jaffe et al., 1995).

Methods of direct compliance cost estimation, which measure the costs to firms of purchasing and maintaining pollution-abatement equipment plus costs to government of administering a policy, are acceptable when behavioural responses, transitional costs, and indirect costs are small. Partial and general equilibrium analysis allows for the incorporation of behavioural responses to changes in public policy. Partial equilibrium analysis of compliance costs incorporates behavioural responses by modelling supply and/or demand in major affected markets, but assumes that the effects of a regulation are confined to one or a few markets. This may be satisfactory if the markets affected by the policy are

small in relation to the overall economy; but, if an environmental policy is expected to have large consequences for the economy, general equilibrium analysis is required, such as through the use of computable general equilibrium models (Hazilla and Kopp, 1990; Conrad, 2002). The potential interaction of abatement costs with pre-existing taxes indicates the importance of employing general equilibrium models for comprehensive cost analysis. Revenue recycling (using emission tax or auctioned permit revenues to reduce distortionary taxes) can make the costs of pollution control significantly less than they would otherwise be (Goulder, 1995).

In a retrospective examination of 28 environmental and occupational safety regulations, Harrington, Morgenstern, and Nelson (2000) found that 14 cost estimation analyses had produced *ex ante* cost estimates that exceeded actual *ex post* costs, apparently due to technological innovation stimulated by market-based instruments (see below).

The benefits of environmental regulations

Protecting the environment usually involves active employment of capital, labour, and other scarce resources. The benefits of an environmental policy are defined as the sum of individuals' aggregate willingness to pay (WTP) for the reduction or prevention of environmental damages or individuals' willingness to accept (WTA) compensation to tolerate such environmental damages. In theory, which measure of value is appropriate for assessing a particular policy depends upon the related assignment of property rights, the nature of the status quo, and whether the change being measured is a gain or a loss; but under a variety of conditions the difference between the two measures may be expected to be relatively small (Willig, 1976). Empirical evidence suggests larger than expected differences between willingness to pay and willingness to accept (Fisher, McClelland and Schulze, 1988). Theoretical explanations include psychological aversion to loss and poor substitutes for environmental amenities (Hanemann, 1991).

The benefits people derive from environmental protection can be categorized as (*a*) related to human health (mortality and morbidity), (*b*) ecological (both market and non-market), or (*c*) materials damage. The distinction between use value and non-use value is critical. In addition to the direct benefits (use value) people receive through protection of their health or through use of a natural resource, they derive passive or non-use value from environmental quality, particularly in the ecological domain. For example, an individual may value a change in an environmental good because she wants to preserve the good for her heirs (bequest value). Still other people may envision no current or future use by themselves or their heirs, but still wish to protect the good because they believe it should be protected or because they derive satisfaction from simply knowing it exists (existence value).

How much would individuals sacrifice to achieve a small reduction in the probability of death during a given period of time? How much compensation would individuals require to accept a small increase in that probability? These are reasonable economic questions because most environmental regulations result in very small changes in individuals' mortality risks. Hedonic wage studies, averted behaviour, and contingent valuation (all discussed below) can provide estimates of marginal willingness to pay or willingness to accept related to small changes in mortality risk, and such estimates can be normalized as the 'value of a statistical life' (VSL).

The VSL is *not* the value of an individual life, whether in ethical or technical, economic terms. Rather it

is simply a convention:

$$VSL = \frac{MWTP \text{ or } MWTA \text{ (from hedonic wage or CV)}}{\text{Small risk change}} \quad (3)$$

where *MWTP* and *MWTA*, respectively, refer to marginal willingness to pay and marginal willingness to accept. For example, if people are willing, on average, to pay $12 for a risk reduction from 5 in 500,000 to 4 in 500,000, equation (3) would yield:

$$VSL = \frac{\$12}{0.000002} = \$6,000,000 \quad (4)$$

Thus, VSL quantifies the aggregate amount that a group of individuals are willing to pay for small reductions in risk, standardized (extrapolated) for a risk change of 1.0. It is not the economic value of an individual life because the VSL calculation does not signify that an individual would pay $6 million to avoid (certain) death this year, or accept (certain) death this year in exchange for $6 million.

Revealed preference methods of environmental benefit estimation

The *averting behaviour method*, in which values of willingness to pay are inferred from observations of people's behavioural responses to changes in environmental quality, is grounded in the household production function framework (Bockstael and McConnell, 1983). People sometimes take actions to reduce the risk (averting behaviour) or lessen the impacts (mitigating behaviour) of environmental damages, for example by purchasing water filters or bottled water. In theory, people's perceptions of the cost of averting behaviour and its effectiveness should be measured (Cropper and Freeman, 1991), but in practice actual expenditures on averting and mitigating behaviours are typically employed. An additional challenge is posed by the necessity of disentangling attributes of the market good or service. Recreational activities represent a potentially large class of benefits that are important in assessing policies affecting the use of public lands. The models used to estimate recreation demand fall within the class of household production models. *Travel cost models* (or Hotelling–Clawson–Knetsch models) use information about time and money spent visiting a site to infer the value of that recreational resource (Bockstael, 1996). The simplest version of the method involves one site and uses data from surveys of users from various geographic origins, together with estimates of the cost of travel and opportunity cost of time, to infer a demand function relating the number of trips to the site to a function of people's willingness to pay for the experience. *Random utility models* explicitly model the consumer's decision to

choose a particular site from among recreation locations, assessing the probability of visiting each location. Such models can be used to value changes in environmental quality by comparing decisions to visit alternative sites (Phaneuf and Smith, 2004).

All recreation demand models share limitations. First, the valuation of costs depends on estimates of the opportunity cost of (leisure) time, which is notoriously difficult to estimate. Also, most trips to a recreation site are part of a multi-purpose experience. In addition, random utility models rely on people's perceptions of environmental quality changes. Finally, like all revealed-preference approaches, recreation demand models can be used to estimate use value only; non-use value cannot be examined.

An alternative approach to assessing people's willingness to pay for recreational experiences is to draw on evidence from *private options to use public goods*. This approach also fits within the household production framework, and is based upon the notion of estimating the derived demand for a privately traded option to utilize a freely available public good. In particular, the demand for state fishing licences has been used to infer the benefits of recreational fishing. Using panel data on fishing license sales and prices, combined with data on substitute prices and demographic variables, Bennear, Stavins and Wagner (2005) estimated a licence demand function from which the expected benefits of a recreational fishing day were derived.

Hedonic pricing methods are founded on the proposition that people value goods in terms of the bundles of attributes that constitute those goods. *Hedonic property value methods* employ data on residential property values and home characteristics, including structural, neighbourhood, and environmental quality attributes (Palmquist, 2003). By regressing the property value on key attributes, the hedonic price function is estimated:

$$P = f(\underset{\sim}{x}, \underset{\sim}{z}, e) \tag{5}$$

where P=housing price (includes land); $\underset{\sim}{x}$=vector of structural attributes; $\underset{\sim}{z}$=vector of neighbourhood attributes; and e=environmental attribute of concern.

From the estimated hedonic price function of equation (5), the marginal implicit price of any attribute, including environmental quality, can be calculated as the partial derivative of the housing price with respect to the given attribute:

$$\frac{\partial P}{\partial e} = \frac{\partial f(\cdot)}{\partial e} = P_e \tag{6}$$

This marginal implicit price, P_e, measures the aggregate marginal willingness to pay for the attribute in

question. For purposes of benefit estimation, the demand function for the attribute is required, and so it is necessary to examine how the marginal implicit price of the environmental attribute varies with changes in the quantity of the attribute and other relevant variables. If the hedonic price equation (5) is nonlinear, then fitted values of P_e can be calculated as e is varied, and a second-stage equation can be estimated:

$$\hat{P}_e = g(e, \underset{\sim}{y}) \tag{7}$$

where $\hat{P}_e$=the fitted value of the marginal implicit price of e from the first-stage equation; and $\underset{\sim}{y}$=a vector of factors that affect marginal willingness to pay for e, including buyer characteristics.
Equation (7), above, has been interpreted as the demand function for the environmental attribute, from which benefits (consumers surplus) can be estimated in the usual way; but there are problems. Most important among these is the question of whether a demand function has actually been estimated, since environmental quality may affect both the demand for housing and its supply, raising the classic identification problem. In addition, informational asymmetries may distort the analysis. Also, because the hedonic property method is based on analysis of marginal changes, it should not be applied to analysis of policies with large anticipated effects.
A related benefit-estimation technique is the *hedonic wage method*, based on the reality that individuals in well-functioning labour markets make trade-offs between wages and risk of on-the-job injuries (or death). A job is a bundle of characteristics, including its wage, responsibilities and risk, among others factors. Two jobs that require the same skill level but have different risks of on-the-job mortality will pay different wages. On the labour supply side, employees tend to require extra compensation to accept jobs with greater risks; and on the labour demand side, employers are willing to offer higher wages to attract workers to riskier jobs. Hence, labour market data on wages and job characteristics can be used to estimate people's marginal implicit price of risk, that is, their valuation of risk. By regressing the wage on key attributes, the hedonic price function is estimated:

$$W = h(\underset{\sim}{x}, r) \tag{8}$$

where W=wage (in annual terms); $\underset{\sim}{x}$=vector of worker and job characteristics; and r=mortality risk of job.
The marginal implicit price of risk is calculated as the partial derivative of the annual wage with respect to the measured mortality risk:

$$\frac{\partial W}{\partial r} = \frac{\partial h(\cdot)}{\partial r} = W_r \tag{9}$$

This marginal implicit price of risk is the average annual income necessary to compensate a worker for a marginal change in risk throughout the year, and it varies with the level of risk.

Many of the issues that arise with the hedonic property value method have parallels here. First, there is the possibility of simultaneity: causality between risk and wages can run in both directions. Also, if individuals' perceptions of risk do not correspond with actual risks, then the marginal implicit price of risk calculated from a hedonic wage study will be biased, and imperfections in labour markets (less than perfect mobility) can cause further problems.

Direct application of the method in the environmental realm is limited to occupational (as opposed to environmental) exposures and risks. Yet hedonic wage methods are of considerable importance in the environmental policy realm, because the results from hedonic wage studies have frequently been used through 'benefit transfer' to infer the VSL. In such applications, the hedonic wage method brings with it possible bias, because studies typically focus on risky occupations, which may attract workers who are systematically less risk-averse.

Standard economic theory would suggest that younger people would have higher values for risk reduction because they have a longer expected life remaining before them and thus a higher expected lifetime utility (Moore and Viscusi, 1988; Cropper and Sussman, 1990). In contrast, some models and empirical evidence suggest that older people may in fact have a higher demand for reducing mortality risks than younger people, and that the value of a life may follow an 'inverted-U' shape over the life cycle, with its peak during mid-life (Shepard and Zeckhauser, 1982; Mrozek and Taylor, 2002; Viscusi and Aldy, 2003; Alberini et al., 2004).

Stated preference methods of environmental benefit estimation

In the best known stated preference method, *contingent valuation* (CV), survey respondents are presented with scenarios that require them to trade off, hypothetically, something for a change in an environmental good or service (Mitchell and Carson, 1989; Boyle, 2003). The simplest approach is to ask people for their maximum willingness to pay, but as there are few real markets in which individuals are actually asked to generate their reservation prices, this method is considered unreliable. In a bidding game, the researcher begins by stating a willingness-to-pay number, asks for a yes–no response, and then increases or decreases the amount until indifference is achieved. The problem with this approach is starting-point bias. A related approach is the use of a payment card shown to the respondent, but the range of WTP on the card may introduce bias, and the approach cannot be used with telephone surveys. Finally, the referendum (discrete choice) approach is favoured by researchers. Each respondent is offered a different WTP number, to which a simple yes–no response is solicited.

The primary advantage of contingent valuation is that it can be applied to a wide range of situations,

including use as well as non-use value; but potential problems remain. Respondents may not understand what they are being asked to value. This may introduce greater variance, if not bias, in responses. Likewise, respondents may not take the hypothetical market seriously because no budget constraint is imposed. This can increase variance and bias. Yet if the scenario is ‘too realistic,’ strategic bias may be expected to show up in responses. Finally, the ‘warm glow effect’ may plague some stated preference surveys: people may purchase moral satisfaction with large but unreal statements of their willingness-to-pay (Andreoni, 1995).

The 1989 Exxon Valdez oil spill off the coast of Alaska led to massive litigation, and resulted in the most prominent use ever of the concept of non-use value and the method of contingent valuation for its estimation. The result was a symposium sponsored by the Exxon Corporation attacking the CV method (Hausman, 1993), and the subsequent creation of a government panel – established by the National Oceanic and Atmospheric Administration (NOAA) and chaired by two Nobel laureates in economics – to assess the scientific validity of the CV method. The NOAA panel concluded that ‘CV studies can produce estimates reliable enough to be the starting point of a judicial process of damage assessment, including lost passive (non-use) values’ (Arrow et al., 1993, p. 4610). The panel offered its approval of CV methods subject to a set of best-practice guidelines.

It is important to distinguish between legitimate methods of benefit estimation and approaches sometimes encountered in the policy process that do not measure willingness-to-pay or willingness-to-accept. Frequently misused techniques include: (*a*) employing, as proxies for the benefits of a policy, estimates of the ‘cost avoided’ by not using the next most costly means of achieving the policy's goals; (*b*) ‘societal revealed preference’ models, which seek to infer the benefits of a proposed policy from the costs of previous regulatory actions; and (*c*) cost-of-illness or human-capital measures which estimate explicit market costs resulting from changes in morbidity or mortality. Because none of these approaches provides estimates of WTP or WTA, these techniques do not provide valid measures of economic benefits.

Choosing instruments: the means of environmental policy

Even if the goals of environmental policies are given, economic analysis can bring insights to the assessment and design of environmental policies. One important criterion is *cost-effectiveness*, defined as the allocation of control among sources that results in the aggregate target being achieved at the lowest possible cost, that is, the allocation which satisfies the following cost-minimization problem:

$$\min_{\{r_i\}} C = \sum_{i=1}^{N} c_i(r_i) \tag{10}$$

$$s.t. \quad \sum_{i=1}^{N} [u_i - r_i] \leq \bar{E}$$
(11)

$$and \quad 0 \leq r_i \leq u_i$$
(12)

where r_i=reductions in emissions (abatement or control) by source i (i=1 to N); $c_i(r_i)$=cost function for source i; C=aggregate cost of control; u_i=uncontrolled emissions by source i; and $\bar{E}$=the aggregate emissions target imposed by the regulatory authority.
If the cost functions are convex, then necessary and sufficient conditions for satisfaction of the constrained optimization problem posed by equations (10) to (12) are the following (among others) (Kuhn and Tucker, 1951):

$$\frac{\partial c_i(r_i)}{\partial r_i} - \lambda \geq 0$$
(13)

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - \lambda \right] = 0$$
(14)

Equations (13) and (14) together imply the crucial condition for cost-effectiveness that all sources (that exercise some degree of control) experience the same marginal abatement costs (Baumol and Oates, 1988). Thus, when one examines environmental policy instruments, a key question is whether marginal abatement costs are likely to be being equated across sources.

Command-and-control versus market-based instruments

Conventional approaches to regulating the environment – frequently characterized as command-and-control – allow relatively little flexibility in the means of achieving goals. Such policy instruments tend to force firms to take on equal shares of the pollution-control burden, regardless of the cost. The most prevalent form of uniform command-and-control standards is technology standards that specify the adoption of specific pollution-control technologies, and performance standards that specify uniform limits on the amount of pollution a facility can generate. In theory, non-uniform performance standards could be made to be cost-effective, but the government typically lacks the requisite information (on marginal costs of individual sources).

Market-based instruments encourage behaviour through market signals rather than through explicit directives regarding pollution-control levels or methods. Market-based instruments fall within four categories: pollution charges, tradable permits, market-friction reductions, and government subsidy reductions. Liability rules may also be thought of as a market-based instrument, because they provide incentives for firms to take into account the potential environmental damages of their decisions.

Where there is significant heterogeneity of abatement costs, command-and-control methods will not be cost-effective. In reality, costs can vary enormously due to production design, physical configuration, age of assets, and other factors. For example, the marginal costs of controlling lead emissions have been estimated to range from $13 to $56,000 per ton (Hartman, Wheeler and Singh, 1994; Morgenstern, 2000). But where costs are similar among sources, command-and-control instruments may perform as well as (or better than) market-based instruments, depending on transactions costs, administrative costs, possibilities for strategic behaviour, political costs, and the nature of the pollutants (Newell and Stavins, 2003).

In theory, market-based instruments allow any desired level of pollution clean-up to be realized at the lowest overall cost by providing incentives for the greatest reductions in pollution by those firms that can achieve the reductions most cheaply. Rather than equalizing pollution levels among firms, market-based instruments equalize their marginal abatement costs (Montgomery, 1972). In addition, market-based instruments have the potential to bring down abatement costs over time by providing incentives for companies to adopt cheaper and better pollution-control technologies. This is because, with market-based instruments, most clearly with emission taxes, it pays firms to clean up a bit more if a sufficiently low-cost method (technology or process) of doing so can be identified and adopted (Downing and White, 1986; Maleug, 1989; Milliman and Prince, 1989; Jaffe and Stavins, 1995). However, the ranking among policy instruments in terms of their respective impacts on technology innovation and diffusion is ambiguous (Jaffe, Newell and Stavins, 2003).

Closely related to the effects of instrument choice on technological change are the effects of vintage-differentiated regulation on the rate of capital turnover, and thereby on pollution abatement costs and environmental performance. Vintage-differentiated regulation is a common feature of many environmental policies, whereby the standard for regulated units is fixed in terms of their date of entry, with later vintages facing more stringent regulation. Such vintage-differentiated regulations can be expected to retard turnover in the capital stock, and thereby to reduce the cost-effectiveness of regulation. Under some conditions the result can be higher levels of pollutant emissions than would occur in the absence of regulation. Such economic and environmental consequences are not only

predictions from theory (Maloney and Brady, 1988); both types of consequences have been validated empirically (Gruenspecht, 1982; Nelson, Tietenberg and Donihue, 1993).

Pollution charges

Pollution charge systems assess a fee or tax on the amount of pollution that firms or sources generate (Pigou, 1920). By definition, actual emissions are equal to unconstrained emissions minus emissions reductions, that is, $e_i = u_i - r_i$. A source's cost minimization problem in the presence of an emissions tax, t, is given by:

$$\min_{\{r_i\}} \left[c_i(r_i) + t \cdot (u_i - r_i) \right] \tag{15}$$

$$s.t. \quad r_i \geq 0 \tag{16}$$

The result for each source is:

$$\frac{\partial c_i(r_i)}{\partial r_i} - t \geq 0 \tag{17}$$

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - t \right] = 0 \tag{18}$$

Equations (17) and (18) imply that each source (that exercises a positive level of control) will carry out

abatement up to the point where its marginal control costs are equal to the tax rate. Hence, marginal abatement costs are equated across sources, satisfying the condition for cost-effectiveness specified by equations (13) and (14), at least in the simplest case of a uniformly mixed pollutant. In the non-uniformly mixed pollutant case, where ‘hot spots’ can be an issue, the respective cost-effective instrument is an ‘ambient charge’.

A challenge with charge systems is identifying the appropriate tax rate. For social efficiency, it should be set equal to the marginal benefits of clean-up at the efficient level of clean-up (Pigou, 1920); but policymakers are more likely to think in terms of a desired level of clean-up, and they do not know beforehand how firms will respond to a given level of taxation. An additional problem is that, although such systems minimize aggregate social costs, these systems may be *more* costly than comparable command-and-control instruments *for regulated firms*, because firms pay both their abatement costs *and* taxes on their residual emissions.

If charges are broadly defined, many applications can be identified (Stavins, 2003). Coming closest to true Pigouvian taxes are the increasingly common *unit-charge* systems for financing municipal solid waste collection, where households and businesses are charged the incremental costs of collection and disposal. Another important set of charge systems has been *deposit refund systems*, whereby consumers pay a surcharge when purchasing potentially polluting products, and receive a refund when returning the product to an approved centre for recycling or disposal. A number of countries and states have implemented this approach to control litter from beverage containers and to reduce the flow of solid waste to landfills (Bohm, 1981; Menell, 1990), and the concept has also been applied to lead-acid batteries. There has also been considerable use of *environmental user charges*, through which specific environmentally related services are funded. Examples include *insurance premium taxes* (Barthold, 1994). Another set of environmental charges are *sales taxes* on motor fuels, ozone-depleting chemicals, agricultural inputs, and low-mileage motor vehicles. Finally, *tax differentiation* has been used to encourage the use of renewable energy sources.

Tradable permit systems

Tradable permits can achieve the same cost-minimizing allocation as a charge system, while avoiding the problems of uncertain firm responses and the distributional consequences of taxes. Under a tradable permit system, an allowed overall level of pollution, E^{-}, is established, and allocated among sources in the form of permits. Firms that keep emission levels below allotted levels may sell surplus permits to other firms or use them to offset excess emissions in other parts of their operations. Let q_{0i} be the initial allocation of emission permits to source i, such that:

$$\sum_{i=1}^{N} q_{0i} = \bar{E} \tag{19}$$

Then, if *p* is the market-determined price of tradable permits, a single firm's cost minimization problem is given by:

$$\min_{\{r_i\}} \left[c_i(r_i) + p \cdot (u_i - r_i - q_{0i}) \right] \tag{20}$$

$$s.t. \quad r_i \geq 0 \tag{21}$$

The result for each source is:

$$\frac{\partial c_i(r_i)}{\partial r_i} - p \geq 0 \tag{22}$$

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - p \right] = 0 \tag{23}$$

Equations (22) and (23) together imply that each source (that exercises a positive level of control) will carry out abatement up to the point where its marginal control costs are equal to the market-determined permit price. Hence, the environmental constraint, $\bar{E}$, is satisfied, and marginal abatement costs are equated across sources, satisfying the condition of cost-effectiveness. The unique cost-effective equilibrium is achieved independently of the initial allocation of permits (Montgomery, 1972), which is of great political significance.

The performance of a tradable permit system can be adversely affected by: concentration in the permit

market (Hahn, 1984; Misolek and Elder, 1989); concentration in the product market (Maleug, 1990); transaction costs (Stavins, 1995); non-profit maximizing behaviour, such as sales or staff maximization (Tschirhart, 1984); the pre-existing regulatory environment (Bohi and Burtraw, 1992); and the degree of monitoring and enforcement (Montero, 2003).

Tradable permits have been the most frequently used market-based system (US Environmental Protection Agency, 2000). Significant applications include: the emissions trading programme (Tietenberg, 1985; Hahn, 1989); the leaded gasoline phase-down; water quality permit trading (Hahn, 1989; Stephenson, Norris and Shabman, 1998); CFC trading (Hahn and McGartland, 1989); the sulphur dioxide (SO_2) allowance trading system for acid rain control (Schmalensee et al., 1998; Stavins, 1998; Carlson et al., 2000; Ellerman et al., 2000); the RECLAIM programme in the Los Angeles metropolitan region (Harrison, 1999); tradable development rights for land use; and the European Union's greenhouse gas emission trading scheme.

Market friction reduction

Market friction reduction can serve as a policy instrument for environmental protection. *Market creation* establishes markets for inputs or outputs associated with environmental quality. Examples of market creation include measures that facilitate the voluntary exchange of water rights and thus promote more efficient allocation and use of scarce water supplies (Howe, 1997), and policies that facilitate the restructuring of electricity generation and transmission. Since well-functioning markets depend, in part, on the existence of well-informed producers and consumers, *information programmes* can help foster market-oriented solutions to environmental problems. These programmes have been of two types. *Product labelling requirements* have been implemented to improve information sets available to consumers, while other programmes have involved *reporting requirements* (Hamilton, 1995; Konar and Cohen, 1997; Khanna, Quimio and Bojilova, 1998).

Government subsidy reduction

Government subsidy reduction constitutes another category of market-based instruments. Subsidies are the mirror image of taxes and, in theory, can provide incentives to address environmental problems. Although subsidies can advance environmental quality (see, for example, Jaffe and Stavins, 1995), it is also true that subsidies, in general, have important disadvantages relatives to taxes (Dewees and Sims, 1976; Baumol and Oates, 1988). Because subsidies increase profits in an industry, they encourage entry, and can thereby increase industry size and pollution output (Mestelman, 1982; Kohn, 1985). In practice, rather than internalizing externalities, many subsidies promote economically inefficient and environmentally unsound practices. In such cases, reducing subsidies can increase efficiency and improve environmental quality. For example, because of concerns about global climate change, increased attention has been given to cutting inefficient subsidies that promote the use of fossil fuels.

Implications of uncertainty for instrument choice

The dual task facing policymakers of choosing environmental goals and selecting policy instruments to achieve those goals must be carried out in the presence of the significant uncertainty that affects the benefits and the costs of environmental protection. Since Weitzman's (1974) classic paper on ‘Prices vs. quantities’, it has been widely acknowledged that benefit uncertainty on its own has no effect on the identity of the efficient control instrument, but that cost uncertainty can have significant effects, depending upon the relative slopes of the marginal benefit (damage) and marginal cost functions. In particular, if uncertainty about marginal abatement costs is significant, and if marginal abatement costs are flat relative to marginal benefits, then a quantity instrument is more efficient than a price instrument. In the environmental realm, benefit uncertainty and cost uncertainty are usually both present, with benefit uncertainty of greater magnitude. When marginal benefits are positively correlated with marginal costs (which, it turns out, is not uncommon), then there is an additional argument in favour of the relative efficiency of quantity instruments (Stavins, 1996). Nevertheless, the regulation of stock pollutants will often favour price instruments, because the marginal benefit function – linked with the stock of pollution – will tend to be flatter than the marginal cost function – linked with the flow of pollution (Newell and Pizer, 2003). In theory, there would be considerable efficiency advantages in the presence of uncertainty of hybrid systems – for example, quotas combined with taxes – or nonlinear taxes (Roberts and Spence, 1976; Weitzman, 1978; Kaplow and Shavell, 2002; Pizer, 2002), but such systems have not been adopted.

Conclusion

The growing use of economic analysis to inform environmental decision-making marks greater acceptance of the usefulness of these tools in improving regulation. But debates about the normative standing of the Kaldor–Hicks criterion and the challenges inherent in making benefit–cost analysis operational will continue. Nevertheless, economic analysis has assumed a significant position in the regulatory state. At the same time, despite the arguments made for decades by economists, there is only limited political support for broader use of benefit–cost analysis to assess proposed or existing environmental regulations. These analytical methods remain on the periphery of policy formulation. In a growing literature (not reviewed here), economists have examined the processes through which political decisions regarding environmental regulation are made (Stavins, 2004).

The significant changes that have taken place over the past 20 years with regard to the means of environmental policy – that is, acceptance of market-based environmental instruments – may provide a model for progress with analysis of the ends – the targets and goals – of public policies in this domain. The change in the former realm has been dramatic. Market-based instruments have moved centre stage, and policy debates today look very different from those of 20 years ago, when these ideas were routinely characterized as ‘licences to pollute’ or dismissed as completely impractical. Market-based instruments are now considered seriously for nearly every environmental problem that is tackled, ranging from endangered species preservation to regional smog and global climate change. Of course, no individual policy instrument – whether market-based or conventional – is appropriate for all environmental problems. Which instrument is best in any given situation depends upon a variety of characteristics of the environmental problem, and the social, political, and economic context in which it is regulated.

See Also

- climate change, economics of
- common property resources
- contingent valuation
- ecological economics
- energy economics
- environmental Kuznets curve
- hedonic prices
- household production and public goods
- pollution haven hypothesis
- pollution permits
- social discount rate
- value of life

Bibliography

Ainslie, G. 1991. Derivation of rational economic behavior from hyperbolic discount curves. *American Economic Review* 81, 334–40.

Alberini, A., Cropper, M., Krupnick, A. and Simon, N. 2004. Does the value of a statistical life vary with age and health status? Evidence from the US and Canada. *Journal of Environmental Economics and Management* 48, 769–92.

Andreoni, J. 1995. Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110, 1–21.

Arrow, K., Cropper, M., Eads, G., Hahn, R., Lave, L., Noll, R., Portney, P., Russell, M., Schmalensee, R., Smith, K. and Stavins, R. 1996. Is there a role for benefit–cost analysis in environmental, health, and safety regulation? *Science* 272, 221–2.

Arrow, K., Solow, R., Portney, P., Leamer, E., Radner, R. and Schuman, H. 1993. Report of the NOAA Panel on Contingent Valuation. *Federal Register* 58, 4601–14.

Barthold, T. 1994. Issues in the design of environmental excise taxes. *Journal of Economic Perspectives* 8(1), 133–51.

Baumol, W. and Oates, W. 1988. *The Theory of Environmental Policy*. Cambridge: Cambridge University Press.

Bennear, L., Stavins, R. and Wagner, A. 2005. Using revealed preferences to infer environmental benefits: evidence from recreational fishing. *Journal of Regulatory Economics* 28, 157–79.

Bockstael, N. 1996. Travel cost methods. In *The Handbook of Environmental Economics*, ed. D. Bromley. Oxford: Blackwell Publishers.

Bockstael, N. and McConnell, K. 1983. Welfare measurement in the household production framework. *American Economic Review* 73, 806–14.

Bohi, D. and Burtraw, D. 1992. Utility investment behavior and the emission trading market. *Resources and Energy* 14, 129–53.

Bohm, P. 1981. *Deposit-Refund Systems: Theory and Applications to Environmental, Conservation, and Consumer Policy*. Baltimore, MD: Resources for the Future/Johns Hopkins University Press.

Boyle, K. 2003. Contingent valuation in practice. In *A Primer on Nonmarket Valuation*, ed. P. Champ, K. Boyle and T. Brown. Dordrecht: Kluwer Academic Publishers.

Carlson, C., Burtraw, D., Cropper, M. and Palmer, K. 2000. Sulfur dioxide control by electric utilities: what are the gains from trade? *Journal of Political Economy* 108, 1292–326.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Conrad, K. 2002. Computable general equilibrium models in environmental and resource economics. In *The International Yearbook of Environmental and Resource Economics 2002/2003: A Survey of Current Issues*, ed. T. Tietenberg and H. Folmer. Northampton: Edward Elgar.

Cropper, M., Aydede, S. and Portney, P. 1994. Preferences for life saving programs: how the public discounts time and age. *Journal of Risk and Uncertainty* 8, 243–65.

Cropper, M. and Freeman, A. 1991. Environmental health effects. In *Measuring the Demand for Environmental Quality*, ed. B. Braden and C. Kolstad. Amsterdam: Elsevier Science Publications.

Cropper, M. and Laibson, D. 1999. The implications of hyperbolic discounting for project evaluation. In *Discounting and Intergenerational Equity*, ed. P. Portney and J. Weyant. Washington, DC: Resources for the Future.

Cropper, M. and Sussman, F. 1990. Valuing future risks to life. *Journal of Environmental Economics and Management* 19, 160–74.

Dewees, D. and Sims, W. 1976. The symmetry of effluent charges and subsidies for pollution control.

Canadian Journal of Economics 9, 323–31.

Downing, P. and White, L. 1986. Innovation in pollution control. *Journal of Environmental Economics and Management* 13, 18–27.

Ellerman, D., Joskow, P., Schmalensee, R.,Montero, J. and Bailey, E. 2000. *Markets for Clean Air: The US Acid Rain Program*. New York: Cambridge University Press.

Fisher, A., McClelland, G. and Schulze, W. 1988. Measures of willingness to pay versus willingness to accept: evidence, explanations, and potential reconciliation. In *Amenity Resource Valuation: Integrating Economics with Other Disciplines*, ed. G. Peterson, B. Driver and R. Gregory. State College, PA: Venture.

Goulder, L. 1995. Environmental taxation and the double dividend: a reader's guide. *International Tax and Public Finance* 2, 157–83.

Goulder, L. and Stavins, R. 2002. An eye on the future: how economists’ controversial practice of discounting really affects the evaluation of environmental policies. *Nature* 419, 673–4.

Gruenspecht, H. 1982. Differentiated regulation: the case of auto emissions standards. *American Economic Review Papers and Proceedings* 72, 328–31.

Hahn, R. 1984. Market power and transferable property rights. *Quarterly Journal of Economics* 99, 753–65.

Hahn, R. 1989. Economic prescriptions for environmental problems: how the patient followed the doctor's orders. *Journal of Economic Perspectives* 3, 95–114.

Hahn, R. and McGartland, A. 1989. Political economy of instrument choice: an examination of the U.S. role in implementing the Montreal Protocol. *Northwestern University Law Review* 83, 592–611.

Hamilton, J.T. 1995. Pollution as news: media and stock market reactions to the toxic release inventory data. *Journal of Environmental Economics and Management* 28, 98–113.

Hanemann, M. 1991. Willingness to pay and willingness to accept: how much can they differ? *American Economic Review* 81, 635–47.

Harrington, W., Morgenstern, R. and Nelson, P. 2000. On the accuracy of regulatory cost estimates. *Journal of Policy Analysis and Management* 19, 297–322.

Harrison, D., Jr. 1999. Turning theory into practice for emissions trading in the Los Angeles air basin. In

Pollution for Sale: Emissions Trading and Joint Implementation, ed. S. Sorrell and J. Skea. Edward Elgar.

Hartman, R., Wheeler, D. and Singh, M. 1994. The cost of air pollution abatement. Policy Research Working Paper No. 1398. Washington, DC: World Bank.

Hausman, J., ed. 1993. *Contingent Valuation: A Critical Assessment*. Amsterdam: North-Holland.

Hazilla, M. and Kopp, R. 1990. Social cost of environmental quality regulations: a general equilibrium analysis. *Journal of Political Economy* 98, 853–73.

Hicks, J. 1939. The foundations of welfare economics. *Economic Journal* 49, 696–712.

Howe, C. 1997. Increasing efficiency in water markets: examples from the western United States. In *Water Marketing – The Next Generation*, ed. T. Anderson and P. Hill. Lanham, MD: Rowman and Littlefield Publishers.

Jaffe, A., Newell, R. and Stavins, R. 2003. Technological change and the environment. In *The Handbook of Environmental Economics*, ed. K. Mäler and J. Vincent. Amsterdam: North-Holland/ Elsevier Science.

Jaffe, A. and Stavins, R. 1995. Dynamic incentives of environmental regulation: the effects of alternative policy instruments on technological diffusion. *Journal of Environmental Economics and Management* 29, S43–S63.

Jaffe, A., Peterson, S., Portney, P. and Stavins, R. 1995. Environmental regulation and the competitiveness of U.S. manufacturing: what does the evidence tell us? *Journal of Economic Literature* 33, 132–65.

Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49, 549–52.

Kaplow, L. and Shavell, S. 2002. On the superiority of corrective taxes to quantity regulation. *American Law and Economics Review* 4, 1–17.

Khanna, M., Quimio, W. and Bojilova, D. 1998. Toxic release information: a policy tool for environmental protection. *Journal of Environmental Economics and Management* 36, 243–66.

Kohn, R. 1985. A general equilibrium analysis of the optimal number of firms in a polluting industry. *Canadian Journal of Economics* 18, 347–54.

Konar, S. and Cohen, M. 1997. Information as regulation: the effect of community right to know laws on toxic emissions. *Journal of Environmental Economics and Management* 32, 109–24.

Kuhn, H. and Tucker, A. 1951. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Maleug, D. 1989. Emission credit trading and the incentive to adopt new pollution abatement technology. *Journal of Environmental Economics and Management* 16, 52–7.

Maleug, D. 1990. Welfare consequences of emission credit trading programs. *Journal of Environmental Economics and Management* 18, 66–77.

Maloney, M. and Brady, G. 1988. Capital turnover and marketable property rights. *Journal of Law and Economics* 31, 203–26.

Menell, P. 1990. Beyond the throwaway society: an incentive approach to regulating municipal solid waste. *Ecology Law Quarterly* 17, 655–739.

Mestelman, S. 1982. Production externalities and corrective subsidies: a general equilibrium analysis. *Journal of Environmental Economics and Management* 9, 186–93.

Milliman, S. and Prince, R. 1989. Firm incentives to promote technological change in pollution control. *Journal of Environmental Economics and Management* 17, 247–65.

Misolek, W. and Elder, H. 1989. Exclusionary manipulation of markets for pollution rights. *Journal of Environmental Economics and Management* 16, 156–66.

Mitchell, R. and Carson, R. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington, DC: Resources for the Future.

Montero, J. 2003. Tradeable permits with imperfect monitoring: theory and evidence. Working paper. Cambridge, MA: Center for Energy and Environmental Policy, MIT.

Montgomery, D. 1972. Markets in licenses and efficient pollution control programs. *Journal of Economic Theory* 5, 395–418.

Moore, M. and Viscusi, W. 1988. The quantity-adjusted value of life. *Economic Inquiry* 26, 369–88.

Morgenstern, R. 2000. Decision making at EPA: economics, incentives and efficiency. Draft paper for a conference: EPA at thirty: evaluating and improving the environmental protection agency, Duke

University, 7–8 December.

Mrozek, J. and Taylor, L. 2002. What determines the value of life? A meta analysis. *Journal of Policy Analysis and Management* 21, 253–70.

Nelson, R., Tietenberg, T. and Donihue, M. 1993. Differential environmental regulation: effects on electric utility capital turnover and emissions. *Review of Economics and Statistics* 75, 368–73.

Newell, R. and Pizer, W. 2003. Regulating stock externalities under uncertainty. *Journal of Environmental Economics and Management* 45, 416–32.

Newell, R. and Stavins, R. 2003. Cost heterogeneity and the potential savings from market-based policies. *Journal of Regulatory Economics* 23, 43–59.

Palmquist, R. 2003. Property value models. In *Handbook of Environmental Economics*, vol. 2, ed. K. Mäler and J. Vincent. Amsterdam: North-Holland/Elsevier Science.

Pareto, V. 1896. *Cours d'Economie Politique*, vol. 2. Lausanne: F. Rouge.

Phaneuf, D. and Smith, V. 2004. Recreation demand models. In *Handbook of Environmental Economics*, ed. K. Mäler and J. Vincent. Amsterdam: North-Holland/Elsevier Science.

Pigou, A. 1920. *The Economics of Welfare*. London: Macmillan.

Pizer, W. 2002. Combining price and quantity controls to mitigate global climate change. *Journal of Public Economics* 85, 409–34.

Porter, M. and van der Linde, C. 1995. Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives* 9(4), 97–118.

Revesz, R. and Stavins, R. 2005. Environmental law and policy. In *The Handbook of Law and Economics*, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland/Elsevier Science.

Roberts, M. and Spence, M. 1976. Effluent charges and licenses under uncertainty. *Journal of Public Economics* 5, 193–97.

Sagoff, M. 1993. Environmental economics: an epitaph. *Resources* (111), 2–7.

Schmalensee, R., Joskow, P., Ellerman, A., Montero, J. and Bailey, E. 1998. An interim evaluation of sulfur dioxide emissions trading. *Journal of Economic Perspectives* 12(3), 53–68.

Shepard, D. and Zeckhauser, R. 1982. Life-cycle consumption and willingness to pay for increased survival. In *The Value of Life and Safety*, ed. M. Jones-Lee. Amsterdam: North-Holland.

Stavins, R. 1995. Transaction costs and tradeable permits. *Journal of Environmental Economics and Management* 29, 133–46.

Stavins, R. 1996. Correlated uncertainty and policy instrument choice. *Journal of Environmental Economics and Management* 30, 218–25.

Stavins, R. 1998. What have we learned from the grand policy experiment: lessons from SO_2 allowance trading. *Journal of Economic Perspectives* 12(3), 69–88.

Stavins, R. 2003. Experience with market-based environmental policy instruments. In *The Handbook of Environmental Economics*, ed. K. Mäler and J. Vincent. Amsterdam: North-Holland/Elsevier Science.

Stavins, R. 2004. *The Political Economy of Environmental Regulation*. Northampton, MA: Edward Elgar.

Stephenson, K., Norris, P. and Shabman, L. 1998. Watershed-based effluent trading: the nonpoint source challenge. *Contemporary Economic Policy* 16, 412–21.

Tietenberg, T. 1985. *Emissions Trading: An Exercise in Reforming Pollution Policy*. Washington, DC: Resources for the Future.

Tschirhart, J. 1984. Transferable discharge permits and the control of stationary source air pollution: a survey and synthesis. In *Economic Perspectives on Acid Deposition Control*, ed. T. Crocker. Boston: Butterworth.

US Environmental Protection Agency. 2000. *Guidelines for Preparing Economic Analyses*. Washington, DC: US EPA.

Viscusi, W. and Aldy, J. 2003. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27(1), 5–76.

Weitzman, M. 1974. Prices vs. quantities. *Review of Economic Studies* 41, 477–91.

Weitzman, M. 1978. Optimal rewards for economic regulation. *American Economic Review* 68, 683–91.

Weitzman, M. 1994. On the environmental discount rate. *Journal of Environmental Economics and Management* 26, 200–9.

Weitzman, M. 1998. Why the far-distant future should be discounted at its lowest possible rate. *Journal of Environmental Economics and Management* 36, 201–8.

Willig, R. 1976. Consumer's surplus without apology. *American Economic Review* 66, 589–97.

How to cite this article

Stavins, Robert N. "environmental economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000096> doi:10.1057/9780230226203.0486

The New Palgrave Dictionary of Economics Online

environmental Kuznets curve

Arik Levinson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Pollution often appears first to worsen and later to improve as countries' incomes grow. Because of its resemblance to the pattern of inequality and income described by Simon Kuznets, this pattern of pollution and income has been labelled an 'environmental Kuznets curve'. While many pollutants exhibit this pattern, peak pollution levels occur at different income levels for different pollutants, countries and time periods. This link between income and pollution cannot be interpreted causally, and is consistent with either efficient or inefficient growth paths. The evidence does, however, refute the claim that environmental degradation is an inevitable consequence of economic growth.

Keywords

carbon emissions; economic growth: and pollution; environmental Kuznets curve; environmental regulation; local externalities; pollution; and income

Article

Some forms of pollution appear first to worsen and later to improve as countries' incomes grow. The world's poorest and richest countries have relatively clean environments, while middle-income countries are the most polluted. Because of its resemblance to the pattern of inequality and income described by Simon Kuznets (1955), this pattern of pollution and income has been labelled an 'environmental Kuznets curve' (EKC).

Grossman and Krueger (1995) and the World Bank (1992) first popularized this idea, using a simple empirical approach. They regress data on ambient air and water quality in cities worldwide on a polynomial in GDP per capita and other city and country characteristics. They then plot the fitted values of pollution levels as a function of GDP per capita, and demonstrate that many of the plots appear inverse-U-shaped, first rising and then falling. The peaks of these predicted pollution-income paths vary across pollutants, but 'in most cases they come before a country reaches a per capita income of $8000'

in 1985 dollars (Grossman and Krueger, 1995, p. 353).

In the years since these original observations were made, researchers have examined a wide variety of pollutants for evidence of the EKC pattern, including automotive lead emissions, deforestation, greenhouse gas emissions, toxic waste and indoor air pollution. Some investigators have experimented with different econometric approaches, including higher-order polynomials, fixed and random effects, splines, semi- and nonparametric techniques, and different patterns of interactions and exponents. Others have studied different groups of jurisdictions and different time periods, and have added control variables, including measures of corruption, democratic freedoms, international trade openness, and even income inequality (bringing the subject full circle back to Kuznets's original idea).

Some generalizations across these approaches emerge. Roughly speaking, pollution involving local externalities begins improving at the lowest income levels. Fecal coliform in water and indoor household air pollution are examples. For some of these local externalities, pollution appears to decrease steadily with economic growth, and we observe no turning point at all. This is not a rejection of the EKC; pollution must have increased at some point in order to decline with income eventually, and there simply are no data from the earlier period. By contrast, pollutants involving very dispersed externalities tend to have their turning points at the highest incomes, or even no turning points at all, as pollution appears to increase steadily with income. Carbon emissions provide one such example. This, too, is not necessarily a rejection of the EKC; the turning points for these pollutants may come at levels of income per capita higher than in today's wealthiest economies.

Another general empirical result is that the turning points for individual pollutants differ across countries. This difference shows up as instability in empirical approaches that estimate one fixed turning point for any given pollutant. Countries that are the first to deal with a pollutant do so at higher income levels than following countries, perhaps because the following countries benefit from the science and engineering lessons of the early movers.

Most researchers have been careful to avoid interpreting these reduced-form empirical correlations structurally, and to recognize that economic growth does not automatically cause environmental improvements. All of the studies omit country characteristics correlated with both income and pollution levels, the most important being environmental regulatory stringency. The EKC pattern does not provide evidence of market failures or efficient policies in rich or poor countries. Rather, there are multiple underlying mechanisms, some of which have begun to be modelled theoretically.

In theory, the EKC relationship can be divided into three parts: scale, composition, and technique (see Brock and Taylor, 2005). If as an economy grows the *scale* of all activities increases proportionally, pollution will increase with economic growth. If growth is not proportional but is accompanied by a change in the *composition* of goods produced, then pollution may decline or increase with income. If richer economies produce proportionally fewer pollution-intensive products, because of changing tastes or patterns of trade, this composition effect can lead to a decline in pollution associated with economic growth. Finally, if richer countries use less pollution-intensive production *techniques*, perhaps because environmental quality is a normal good, growth can lead to falling pollution. The EKC summarizes the interaction of these three processes.

Beyond this aggregate decomposition of the EKC, some attempts have been made to formalize structural models that lead to inverse-U-shaped pollution-income patterns. Many describe economies at some type of corner solution initially, where residents of poor countries are willing to trade environmental quality for income at a faster rate than possible using available technologies or resources. As the model

economies become wealthier and their environments dirtier, eventually the marginal utility of income falls and the marginal disutility from pollution rises, to the point where people choose costly abatement mechanisms. After that point, the economies are at interior solutions, marginal abatement costs equal marginal rates of substitution between environmental quality and income, and pollution declines with income (see Stokey, 1998). In frameworks of this type, there is typically zero pollution abatement until some threshold income level is crossed, after which abatement begins and pollution starts declining with income.

To date, the practical lessons from this theoretical literature are limited. Most of the models are designed to yield inverse-U-shaped pollution-income paths, and succeed using a variety of assumptions and mechanisms. Hence, any number of forces may be behind the empirical observation that pollution increases and then decreases with income. Moreover, that pattern cannot be interpreted causally, and is consistent with either efficient or inefficient growth paths. Perhaps the most important insight is in Grossman and Krueger's original paper: ‘We find no evidence that economic growth does unavoidable harm to the natural habitat’ (1995, p. 370). Economists have long argued that environmental degradation is not an inevitable consequence of economic growth. The EKC literature provides empirical support for that claim.

See Also

- environmental economics
- growth and international trade
- pollution haven hypothesis

Bibliography

Brock, W. and Taylor, M. 2005. Economic growth and the environment: a review of theory and empirics. In *The Handbook of Economic Growth*, vol. 1, ed. S. Durlauf and P. Aghion. Amsterdam: North-Holland.

Grossman, G. and Krueger, A. 1995. Economic growth and the environment. *Quarterly Journal of Economics* 110, 353–77.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45, 1–28.

Stokey, N. 1998. Are there limits to growth? *International Economic Review* 39, 1–31.

World Bank. 1992. *World Development Report 1992*. New York: Oxford University Press.

How to cite this article

Levinson, Arik. "environmental Kuznets curve." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000211> doi:10.1057/9780230226203.0487

The New Palgrave Dictionary of Economics Online

Eph é m é rides du citoyen ou chronique de l'esprit national

Peter Groenewegen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Baudeau, N.; Du Pont de Nemours, P.S.; *Ephémérides du Citoyen* ; Mirabeau, V.R. Marquis de; Physiocracy

Article

French economic periodical issued in three series under different names from 1766 to 1772, 1774 to 1776 and in 1788. Published first as a bimonthly by its founder and first editor, l'Abbé Baudeau, it became a monthly as from January 1767 after Baudeau's conversion to Physiocracy by Mirabeau and Le Trosne. Its contents included contributed articles on economic and political subjects, book reviews, comments and letters to the editor, together with a chronicle of public events of interest to its readership. This provided its format from January 1769, when Du Pont de Nemours took over the editorship. Although censorship problems troubled the journal persistently (as disclosed in the Turgot–Du Pont correspondence, for this reason many issues appeared well after the ostensible month of publication) the first series was terminated by l'Abbé Terray in November 1772, presumably because it contained much vigorous criticism of his abolition of domestic free trade in grain. The first series produced therefore six issues in 1766 as a bi-monthly and 63 monthly issues from January 1767 to March 1772 inclusive. Under the title *Nouvelles Ephémérides ou Bibliothèque raisonnée de l'histoire, de la morale et de la politique*, it was revived by Baudeau after Turgot became Contrôleur-général in 1774, publishing 18 issues in all from January 1775 to June 1776, that is, the month after Turgot's dismissal from the ministry. A third series, *Nouvelles Ephémérides économiques* published three issues from January to March 1788, again under Baudeau's editorship, but his failing mental powers were presumably the reason why this final series ended so quickly.

Although initially set up by Baudeau in imitation of the English *Spectator*, within a year of its inception economics began to dominate its contents and many of the leading Physiocrats, in particular Mirabeau, Baudeau and Du Pont de Nemours, contributed most of the articles. A detailed discussion of its contents is given in Bauer (1894) and in Coquelin and Guillaumin (1854, pp. 710–12). Perhaps the most important piece it contained is Turgot's *Réflexions sur la formation et distribution des richesses* in serial

form (*Ephémérides*, 1769, No. 11, pp. 12–56; No. 12, pp. 31–98; and 1770, No. I, pp. 113–73), although with considerable unauthorized alterations and notes by Du Pont (see Groenewegen, 1977, pp. xix–xxi). It also published foreign contributions in French translation, including Beccaria's inaugural lecture with copious notes and comments by Du Pont (*Ephémérides*, 1769, No. 6, pp. 57–152) and a contribution by Franklin on the increasing troubles between England and her American colonies (*Ephémérides*, 1768, No. 8, pp. 159–92). As an early, if not the first, economic journal, the *Ephémérides* remains an important part of economic literature and an indispensable source for those interested in the study of Physiocracy.

See Also

- physiocracy

Bibliography

Bauer, S. 1894. Ephémérides. In *Dictionary of Political Economy*, vol. 1, ed. R.H.I. Palgrave. London: Macmillan.

Coquelin, C. and Guillaumin, H., ed. 1854. Ephémérides. In *Dictionnaire de l'économie politique*, vol. 1. Paris: Guillaumin, Hachette.

Groenewegen, P.D. 1977. *The Economics of A.R.J. Turgot*. The Hague: Martinus Nijhoff.

How to cite this article

Groenewegen, Peter. "Ephémérides du citoyen ou chronique de l'esprit national." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000098> doi:10.1057/9780230226203.0488

The New Palgrave Dictionary of Economics Online

epistemic game theory: an overview

Adam Brandenburger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

common knowledge; complete information; epistemic game theory; epistemic game theory: an overview; game theory; Harsanyi, J. C.; incomplete information; Morgenstern, O.; non-cooperative game theory; rational behaviour; uncertainty; von Neumann, J.

Article

The following three articles survey some aspects of the foundations of non-cooperative game theory. The goal of work in foundations is to examine in detail the basic ingredients of game analysis.
The starting point for most of game theory is a ‘solution concept’ – such as Nash equilibrium or one of its many variants, backward induction, or iterated dominance of various kinds. These are usually thought of as the embodiment of ‘rational behaviour’ in some way and used to analyse game situations.
One could say that the starting point for most game theory is more of an endpoint of work in foundations. Here, the primitives are more basic. The very idea of rational – or irrational – behaviour needs to be formalized. So does what each player might know or believe about the game – including about the rationality or irrationality of other players. Foundational work shows that even what each player knows or believes about what other players know or believe, and so on, can matter.
Investigating the basis of existing solution concepts is one part of work in foundations. Other work in foundations has uncovered new solution concepts with useful properties. Still other work considers changes even to the basic model of decision making by players – such as departures from the expected utility model or reasoning in various formal logics.
The first article, epistemic game theory: beliefs and types, by Marciano Siniscalchi, describes the formalism used in most work on foundations. This is the ‘types’ formalism going back to Harsanyi (1967–8). Originally proposed to describe the players’ beliefs about the structure of the game (such as the payoff functions), the types approach is equally suited to describing beliefs about the play of the game or beliefs about both what the game is and how it will be played. Indeed, in its most general form, the formalism is simply a way to describe any multi-person uncertainty. Harsanyi's conception of a

‘type’ was a crucial breakthrough in game theory. Still, his work left many fundamental questions about multi-person uncertainty unanswered. Siniscalchi's article surveys these later developments.

The second and third articles apply these tools to the two kinds of uncertainty mentioned. The second article, epistemic game theory: complete information, concerns the case where the matrix or tree itself is ‘transparent’ to the players, and what is uncertain are the actual strategies chosen by the players. The third article, epistemic game theory: incomplete information, by Aviad Heifetz, has the opposite focus: it covers the case of uncertainty about the game itself. (Following Harsanyi, the third article focuses on uncertainty about the payoffs, in particular.)

Both cases are important to the foundations programme. Because Nash equilibrium is ‘as if’ each player is certain (and correct) about the strategies chosen by the other players (Aumann and Brandenburger, 1995, Section 7h), uncertainty of the first kind has played a small role in game theory to date. Uncertainty of the second kind is the topic of the large literatures on information asymmetries, incentives, and so on.

Interestingly, though, von Neumann and Morgenstern (1944) already appreciated the significance of both complete and incomplete information environments. Indeed, they asserted that phenomena often thought to be characteristic of incomplete-information settings could, in fact, arise in complete-information settings (1944, p. 31):

> Actually, we think that our investigations – although they assume ‘complete information’ without any further discussion – do make a contribution to the study of this subject. It will be seen that many economic and social phenomena which are usually ascribed to the individual's state of ‘incomplete information’ make their appearance in our theory and can be satisfactorily interpreted with its help.

This is indeed true, as work in the modern foundations programme shows. (Some instances are mentioned in what follows.) Overall, the foundations programme aims at a ‘neutral’ and comprehensive treatment of all ingredients of a game.

See Also

- epistemic game theory: beliefs and types
- epistemic game theory: complete information
- epistemic game theory: incomplete information
- game theory
- Nash equilibrium, refinements of

My thanks to Rena Henderson and Michael James. The Stern School of Business provided financial support.

Bibliography

Aumann, R. and Brandenburger, A. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63, 1161–80.

Harsanyi, J. 1967–8. Games with incomplete information played by ‘Bayesian’ players, I–III. *Management Science* 14, 159–82, 320–34, 486–502.

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press. 60th anniversary edn, 2004.

How to cite this article

Brandenburger, Adam. "epistemic game theory: an overview." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000257>
doi:10.1057/9780230226203.0489

The New Palgrave Dictionary of Economics Online

epistemic game theory: beliefs and types

Marciano Siniscalchi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Modelling what each agent believes about her opponents, what she believes her opponents believe about her, and so on, plays a prominent role in game theory and its applications. This article describes Harsanyi's formalism of type spaces, which provides a simple, elegant representation of probabilistic belief hierarchies. A special emphasis is placed on the construction of rich type spaces, which can generate all ‘reasonable’ belief hierarchies in a given game. Recent developments, employing richer representation of beliefs, are also considered.

Keywords

belief hierarchies; common knowledge; epistemic game theory: beliefs and types; Harsanyi, J.C.; Kolmogorov's extension theorem; Monotonicity; Polish spaces; preferences; recursive preferences; type spaces; universal type space

Article

John Harsanyi (1967–8) introduced the formalism of type spaces to provide a simple and parsimonious representation of belief hierarchies. He explicitly noted that his formalism was not limited to modelling a player's beliefs about payoff-relevant variables: rather, its strength was precisely the ease with which Ann's beliefs about Bob's beliefs about payoff variables, Ann's beliefs about Bob's beliefs about Ann's beliefs about payoff variables, and so on, could be represented.

This feature plays a prominent role in the epistemic analysis of solution concepts (see epistemic game theory: complete information), as well as in the literature on global games (Morris and Shin, 2003) and on robust mechanism design (Bergemann and Morris, 2005). All these applications place particular emphasis on the expressiveness of the type-space formalism. Thus, a natural question arises: just how expressive is Harsanyi's approach?

For instance, solution concepts such as Nash equilibrium or rationalizability can be characterized by means of restrictions on the players’ mutual beliefs. In principle, these assumptions could be formulated directly as

restrictions on players' hierarchies of beliefs; but in practice the analysis is mostly carried out in the context of a type space à la Harsanyi. This is without loss of generality only if Harsanyi type spaces do not themselves impose restrictions on the belief hierarchies that can be represented. Similar considerations apply in the context of robust mechanism design.

A rich literature addresses this issue from different angles, and for a variety of basic representations of beliefs. This article focuses on hierarchies of probabilistic beliefs; however, some extensions are also mentioned. For simplicity, attention is restricted to two players, denoted '1' and '2' or 'i' and '$-i$.'

Probabilistic type spaces and belief hierarchies

Begin with some mathematical preliminaries. A topology on a space X is deemed Polish if it is separable and completely metrizable; in this case, X is itself deemed a Polish space. Examples include finite sets, Euclidean space $\mathbb{R}^n$ and closed subsets thereof. A countable product of Polish spaces, endowed with the product topology, is itself Polish. For any topological space X, the notation $\Delta(X)$ indicates the set of Borel probability measures on X. If the topology on X is Polish, then the weak* topology on $\Delta(X)$ is also Polish (for example, Aliprantis and Border, 1999, Theorem 14.15). A sequence $\{\mu^k\}_{k\geq 1}$ in $\Delta(X)$ converges in the weak* sense to a measure $\mu \in \Delta(X)$, written $\mu^k \overset{w^*}{\square} \mu$, if and only if, for every bounded, continuous function $\psi: X \to \mathbb{R}$, $\int_X \psi d\mu^k \to \int_X \psi d\mu$. The weak* topology on $\Delta(X)$ is especially meaningful and convenient when X is a Polish space: see Aliprantis and Border (1999, ch. 14) for an overview of its properties. Finally, if μ is a measure on some product space $X \times Y$, the marginal of μ on X is denoted $\text{marg}_X \mu$.

The basic ingredient of the players' hierarchical beliefs is a description of payoff-relevant or fundamental uncertainty. Fix two sets S_1 and S_2, hereinafter called the *uncertainty domains*; the intended interpretation is that S_{-i} describes aspects of the strategic situation that Player i is uncertain about. For example, in an independent private-values auction, each set S_i could represent bidder i's possible valuations of the object being sold, which is not known to bidder $-i$. In the context of interactive epistemology, S_i is usually taken to be Player i's strategy space. It is sometimes convenient to let $S_1 = S_2 \equiv S$; in this case, the formalism introduced below enables one to formalize the assumption that each player observes different aspects of the common uncertainty domain S (for instance, different signals correlated with the common, unknown value of an object offered for sale).

An (S_1, S_2)-*based type space* is a tuple $\mathcal{T} = (T_i, g_i)_{i=1,2}$ such that, for each i=1, 2, T_i is a Polish space and $g_i \bullet : \bullet T_i \to \Delta(S_{-I} \times T_{-i})$ is continuous. As noted above, type spaces can represent hierarchies of beliefs; it is useful to begin with an example. Let $S_1 = S_2 = \{a,b\}$ and consider the type space defined in Table 1. To interpret, for every i=1,2, the entry in the row corresponding to t_i and (s_{-i}, t_{-i}) is $g_i(t_i)(\{(s_{-i}, t_{-i})\})$. Thus, for instance, $g_1(t_1)(\{(a, t'_2)\}) = 0$; $g_2(t_2)(\{b\} \times T_1) = 0.5$.

A type space

T_1	$\bullet a, t_2$	a, t'_2	b, t_2	b, t'_2
t_1	•1	0	0	0
t'_1	•0	0.3	0	0.7

T_2	•a, t_1	a, t'_1	b, t_1	b, t'_1
t_2	•0	0.5	0.5	0
t'_2	•0	0	0	1

Consider type t_1 of Player 1. She is certain that $s_2=a$; furthermore, she is certain that Player 2 believes that $s_1=a$ and $s_1=b$ are equally likely. Taking this one step further, type t_1 is certain that Player 2 assigns probability 0.5 to the event that Player 1 believes that $s_2=b$ with probability 0.7.

These intuitive calculations can be formalized as follows. Fix an (S_1, S_2)-based type space $\mathcal{T} = (T_i, g_i)_{i=1,2}$; for every i=1,•2 define the set X^0_{-i} and the function $h^1_i : T_i \to \Delta(X^{-i}_0)$ by

$$X^0_{-i} = S_{-i} \text{ and } \forall t_i \in T_i,\ h^1_i(t_i) = \text{marg}_{S_{-i}} g_i(t_i). \tag{1}$$

Thus, $h^1_i(t_i)$ represents the *first-order beliefs* of type t_i in type space T – her beliefs about the uncertainty domain S_{-i}. Note that each $X^0_{-i} = S_{-i}$ is Polish. Proceeding inductively, assuming that $X^0_{-i}, \ldots, X^{k-1}_{-i}$ and $h^1_i, \ldots, h^k_i$ have been defined up to some k>0 for i=1,2, and that all sets X^{ℓ}_{-i}, $\ell = 0, \ldots, k-1$ are Polish, define the set X^k_{-i} and the functions $h^{k+1}_i : T_i \to \Delta(X^k_{-i})$ for i=1,•2 by

$$X^k_{-i} = X^{k-1}_{-i} \times \Delta(X^{k-1}_i) \text{ and } \forall t_i \in T_i,\ h^{k+1}_i(t_i)(E) = g_i(t_i)\left(\left\{(s_{-i}, t_{-i}) \in S_{-i} \times T_{-i} : (s_{-i}, h^k_{-i}(t_{-i})) \in E\right\}\right) \tag{2}$$

for every Borel subset E of X^k_{-i}. Thus, $h^2_1(t_1)$ represents the *second-order beliefs* of type t_1 – her beliefs about *both* the uncertainty domain $S_2 = X^0_2$ and Player 2's beliefs about S_1, which by definition belong to the set $\Delta(X^0_1) = \Delta(S_1)$. Similarly, $h^{k+1}_i(t_i)$ represents type t_i's (k+1)-th order beliefs.

Observe that type t_i's second-order beliefs are defined over $X^0_2 \times \Delta(X^0_1) = S_2 \times \Delta(S_1)$, rather than just over $\Delta(X^0_1) = \Delta(S_1)$; a similar statement holds for her (k+1)-th order beliefs. This is crucial in many applications. For instance, a typical assumption in the literature on epistemic foundations of solution concepts is that Player 1 believes that Player 2 is rational. Letting S_i be the set of actions or strategies of Player i in the game under consideration, this can be modelled by assuming that the support of $h^2_1(t_1)$

consists of pairs $(s_2, \mu_1) \in S_2 \times \Delta(S_1)$ wherein s_2 is a best response to μ_1. Clearly, such an assumption could not be formalized if $h_1^2(t_1)$ only conveyed information about type t_1's beliefs on Player 2's first-order beliefs: even though type t_1's beliefs about the action played by Player 2 could be retrieved from $h_1^1(t_1)$, it would be impossible to tell whether each action that type t_1 expects to be played is matched with a belief that rationalizes it.

Note that, since X_i^{k-1} and X_{-i}^{k-1} are assumed Polish, so are $\Delta(X_i^{k-1})$ and X_{-i}^k. Also, each function h_i^k is continuous.

Finally, it is convenient to define a function that associates to each type $t_i \in T_i$ an entire *belief hierarchy*: to do so, define the set H_i and, for i=1,•2, the function h_i:$T_i \rightarrow H_i$ by

$$H_i = \prod_{k \geq 0} \Delta(X_{-i}^k) \text{and } \forall t_i \in T_i,\ h_i(t_i) = (h_i^1(t_i), \ldots, h_i^{k+1}(t_i), \ldots). \tag{3}$$

Thus, H_i is the set of all hierarchies of beliefs; notice that, since each X_{-i}^k is Polish, so is H_i.

Rich type spaces

The preceding construction suggests a rather direct way to ask how expressive Harsanyi's notion of a type space is: can one construct a type space that generates *all* hierarchies in H_i?

A moment's reflection shows that this question must be refined. Fix a type space $(T_i, g_i)_{i=1,2}$ and a type $t_i \in T_i$; recall that, for reasons described above, the first- and second-order beliefs of type t_i satisfy $h_i^1(t_i) \in \Delta(S_{-i})$ and $h_i^2(t_i) \in \Delta(X_{-i}^0 \times \Delta(X_i^0)) = \Delta(S_{-i} \times \Delta(S_i))$ respectively. This, however, creates the potential for redundancy or even contradiction, because both $h_i^1(t_i)$ and $\text{marg}_{S_{-i}} h_i^2(t_i)$ can be viewed as 'type t_i's beliefs about S_{-i}'. A similar observation applies to higher-order beliefs. Fortunately, it is easy to verify that, for every type space $(T_i, g_i)_{i=1,2}$ and type $t_i \in T_i$, the following *coherency* condition holds:

$$\forall k > 1, \text{marg}_{X_{-i}^{k-2}} h_i^k(t_i) = h_i^{k-1}(t_i); \tag{4}$$

To interpret, recall that $h_i^k(t_i) \in \Delta(X_{-i}^{k-1}) = \Delta(X_{-i}^{k-2} \times \Delta(X_{-i}^{k-2}))$. Thus, in particular, $\text{marg}_{S_{-i}} h_i^2(t_i) = h_i^1(t_i)$.

Since H_i is defined as the set of *all* hierarchies of beliefs for Player *i*, some (in fact, 'most') of its elements are not coherent. As noted above, no type space can generate incoherent hierarchies; more importantly, coherency can be viewed as an integral part of the interpretation of interactive beliefs. How could an individual simultaneously hold (infinitely) many distinct first-order beliefs? Which of these should be used, say, to verify whether she is rational? This motivates restricting attention to coherent hierarchies, defined as follows:

$$H_i^c = \left\{(\mu_i^1, \mu_i^2, \ldots) \in H_i:\ \forall k > 1, \operatorname{marg}_{X_{-i}^{k-2}} \mu_i^k = \mu_i^{k-1}\right\}. \tag{5}$$

Since $\operatorname{marg}_{X_{-i}^{k-2}} : \Delta(X_{-i}^{k-1}) \to \Delta(X_{-i}^{k-2})$ is continuous, H_i^c is a closed, hence Polish subspace of H_i. Brandenburger and Dekel (1993, Proposition 1) show that there exist homeomorphisms $g_i^c : H_i^c \to \Delta(S_{-i} \times H_{-i})$: that is, every coherent hierarchy corresponds to a distinct belief over the uncertainty domain and the hierarchies of the opponent, and conversely. Furthermore, this homeomorphism is canonical, in the following sense. Note that $S_{-i} \times H_{-i} = S_{-i} \times \prod_{k \geq 0} \Delta(X_i^k) = X_{-i}^k \times \prod_{\ell > k} \Delta(X_i^\ell)$. Then it can be shown that, if $\mu_i = (\mu_i^1, \mu_i^2, \ldots) \in H_i^c$, then $\operatorname{marg}_{X_{-i}^k} g_i^c(\mu_i) = \mu_i^{k+1}$. Intuitively, the marginal belief associated with μ_i over the first k orders of the opponent's beliefs is precisely what it should be, namely μ_i^{k+1}. The proof of these results builds upon Kolmogorov's extension theorem, as may be suggested by the similarity of the coherency condition in eq. (5) with the notion of Kolmogorov consistency: cf. for example Aliprantis and Border (1999, theorem 14.26).

This result does not quite imply that all coherent hierarchies can be generated in a suitable type space; however, it suggests a way to obtain this result. Notice that the belief on $S_{-i} \times H_{-i}$ associated by the homeomorphism g_i^c to a coherent hierarchy μ_i may include *in*coherent hierarchies $\nu_{-i} \in H_{-i} \setminus H_{-i}^c$ in its support. This can be interpreted in the following terms: if Player *i*'s hierarchical beliefs are given by μ_i, then she is coherent, but she is not certain that her opponent is. On the other hand, consider a type space $(T_i, g_i)_{i=1,2}$; as noted above, for every player *i*, each type $t_i \in T_i$ generates a coherent hierarchy $h_i(t_i) \in H_i^c$. So, for instance, if (s_1, t_1) is in the support of $g_2(t_2)$ then t_1 also generates a coherent hierarchy. Thus, not only is type t_2 of Player 2 coherent: he is also certain (believes with probability one) that Player 1 is coherent. Iterating this argument suggests that *hierarchies of beliefs generated by type spaces display common certainty of coherency*.

Motivated by these considerations, let

$$H_i^0 = H_i^c \text{ and } \forall k > 0,\ H_i^k = \left\{\mu_i \in H_i^{k-1} : g_i^c(\mu_i)(S_{-i} \times H_{-i}^{k-1}) = 1\right\}. \tag{6}$$

Thus, H_i^0 is the set of coherent hierarchies for Player *i*; H_i^1 is the set of hierarchies that are coherent and correspond to beliefs that display certainty of the opponent's coherency; and so on. Finally, let $H_i^* = \cap_{k \geq 0} H_i^k$. Each element of H_i^* is intuitively consistent with coherency and common certainty of coherency.

Brandenburger and Dekel (1993, Proposition 2) show that the restriction g_i^* of g_i^c to H_i^* is a homeomorphism between H_i^* and $\Delta(S_{-i} \times H_{-i}^*)$; furthermore, it is canonical in the sense described above. This implies that the tuple $(H_i^*, g_i^*)_{i=1,2}$ is a type space in its own right – the (S_1, S_2)-based *universal type space*.

The existence of a universal type space fully addresses the issue of richness. Since the homeomorphism g_i^* is canonical, it is easy to see that the hierarchy generated as per eqs (1) and (2) by any 'type' $t_i = (\mu^1, \mu^2, s) \in H_i^*$ in the universal type space $(H_i^*, g_i^*)_{i=1,2}$ is t_i itself; thus, since H_i^* *consists* of all hierarchies that are coherent and display common certainty of consistency, the universal type space also *generates* all such hierarchies.

The type space $(H_i^*, g_i^*)_{i=1,2}$ is rich in two additional, related senses. First, as may be expected, every belief hierarchy for Player *i* generated by an arbitrary type space is an element of H_i^*; this implies that every type space $(T_i, g_i)_{i=1,2}$ can be uniquely embedded in $(H_i^*, g_i^*)_{i=1,2}$ as a 'belief-closed' subset: see Battigalli and Siniscalchi (1999, Proposition 8.8). Call a type space *terminal* if, like $(H_i^*, g_i^*)_{i=1,2}$, it embeds all other type spaces as belief-closed subsets.

Second, since each function g_i^* is a homeomorphism, in particular it is a surjection (that is, onto). Call a type space $(T_i, g_i)_{i=1,2}$ *complete* if every map g_i is onto. (This should not be confused with the topological notion of completeness.) Thus, the universal type space $(H_i^*, g_i^*)_{i=1,2}$ is complete. It is often the case that, when a universal type space is employed in the epistemic analysis of solution concepts, the objective is precisely to exploit its completeness. Furthermore, for certain representations of beliefs, it is not known whether universal type spaces can be constructed; however, the existence of complete type spaces can be established, and is sufficient for the purposes of epistemic analysis. The next section provides examples.

Alternative constructions and extensions

The preceding discussion adopts the approach proposed by Brandenburger and Dekel (1993), which has the virtue of relying on familiar ideas from the theory of stochastic processes. However, the first constructions of universal type spaces consisting of hierarchies of beliefs are due to Armbruster and Böge (1979), Böge and Eisele (1979) and Mertens and Zamir (1985).

From a technical point of view, Mertens and Zamir (1985) assume that the state space *S* is compact

Hausdorff and beliefs are regular probability measures. Heifetz and Samet (1998b) instead drop topological assumptions altogether: in their approach, both the underlying set of states and the sets of types of each player are modelled as measurable spaces. They show that a terminal type space can be explicitly constructed in this environment.

In all the contributions mentioned so far, beliefs are modelled as countably additive probabilities. The literature has also examined other representations of beliefs, broadly defined.

A *partitional structure* (Aumann, 1976) is a tuple $(\Omega,(\sigma_i, P_i)_{i=1,2})$, where Ω is a (typically finite) space of 'possible worlds', every $\sigma_i : \Omega \to S_i$ indicates the realization of the basic uncertainty corresponding to each element of Ω, and every P_i is a partition of Ω. The interpretation is that, at any world $\omega \in \Omega$, Player *i* is only informed that the true world lies in the cell of the partition P_i containing ω, denoted $P_i(\omega)$. The *knowledge operator* for Player *i* can then be defined as

$$\forall E \subset \Omega,\ K_i(E) = \{\omega \in \Omega : P_i(\omega) \subseteq E\}.$$

Notice that no probabilistic information is provided in this environment (although it can be easily added). Heifetz and Samet (1998a) show that a terminal partitional structure does not exist. This result was extended to more general 'possibility' structures by Meier (2005). Brandenburger and Keisler (2006) establish related non-existence results for complete structures. However, recent contributions show that topological assumptions, which play a key role in the constructions of Mertens and Zamir (1985) and Brandenburger and Dekel (1993), can also deliver existence results in non-probabilistic settings. For instance, Mariotti, Meier and Piccione (2005) construct a structure that is universal, complete and terminal for possibility structures.

Other authors investigate richer probabilistic representations of beliefs. Battigalli and Siniscalchi (1999) construct a universal, terminal, and complete type space for *conditional probability system*, or collections of probability measures indexed by relevant conditioning events (such as histories in an extensive game) and related by a version of Bayes's rule. This type space is used in (2002) to provide an epistemic analysis of forward induction. Brandenburger, Friedenberg and Keisler (2006) construct a complete type space for *lexicographic sequences*, which may be thought of as an extension of lexicographic probability systems (Blume, Brandenburger and Dekel, 1991) for infinite domains. They then use it to provide an epistemic characterization of iterated admissibility.

Non-probabilistic representations of beliefs that reflect a concern for ambiguity (Ellsberg, 1961) have also been considered. Heifetz and Samet (1998b) observe that their measure-theoretic construction extends to beliefs represented by continuous *capacities*, that is non-additive set functions that preserve monotonicity with respect to set inclusion. Motivated by the multiple-priors model of Gilboa and Schmeidler (1989), Ahn (2006) constructs a universal type space for sets of probabilities.

Epstein and Wang (1996) approach the richness issue taking *preferences*, rather than beliefs, as primitive objects. In their setting, an *S*-based type space is a tuple $(T_i, g_i)_{i=1,2}$, where, for every type t_i, $g_i(t_i)$ is a suitably regular preference over *acts* defined on the set $S \times T_{-i}$. The analysis in the preceding section can be viewed as a special case of Epstein and Wang (1996), where preferences conform to expected-utility theory.

Epstein and Wang construct a universal type space in this framework (see also Di Tillio, 2006). Finally, constructions analogous to that of a universal type space appear in other, unrelated contexts. For instance, Epstein and Zin (1989) develop a class of recursive preferences over infinite-horizon temporal lotteries; to construct the domain of such preferences, they employ arguments related to Mertens and Zamir's. Gul and Pesendorfer (2004) employ analogous techniques to analyse self-control preferences over infinite-horizon consumption problems.

See Also

- epistemic game theory: an overview
- epistemic game theory: complete information
- epistemic game theory: incomplete information

Bibliography

Ahn, D. 2007. Hierarchies of ambiguous beliefs. *Journal of Economic Theory* 136, 286–301.

Aliprantis, C. and Border, K. 1999. *Infinite Dimensional Analysis*, 2nd edn. Berlin: Springer.

Armbruster, W. and Böge, W. 1979. Bayesian game theory. In *Game Theory and Related Topics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.

Aumann, R. 1976. Agreeing to disagree. *Annals of Statistics* 4, 1236–9.

Battigalli, P. and Siniscalchi, M. 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory* 88, 188–230.

Battigalli, P. and Siniscalchi, M. 2002. Strong belief and forward induction reasoning. *Journal of Economic Theory* 106, 356–91.

Bergemann, D. and Morris, S. 2005. Robust mechanism design. *Econometrica* 73, 1521–34.

Blume, L., Brandenburger, A. and Dekel, E. 1991. Lexicographic probabilities and choice under uncertainty. *Econometrica* 59, 61–79.

Böge, W. and Eisele, T. 1979. On solutions of Bayesian games. *International Journal of Game Theory* 8, 193–215.

Brandenburger, A. and Dekel, E. 1993. Hierarchies of beliefs and common knowledge. *Journal of Economic Theory* 59, 189–98.

Brandenburger, A., Friedenberg, A. and Keisler, H.J. 2006. Admissibility in games. Unpublished, Stern

School of Business, New York University.

Brandenburger, A. and Keisler, J. 2006. An impossibility theorem on beliefs in games. *Studia Logica* 84, 211–40.

Di Tillio, A. 2006. Subjective expected utility in games. Working Paper No. 311, IGIER, Università Bocconi.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Epstein, L. and Wang, T. 1996. Beliefs about beliefs without probabilities. *Econometrica* 64, 1343–73.

Epstein, L. and Zin, S. 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. *Econometrica* 57, 937–69.

Gilboa, I. and Schmeidler, D. 1989. Maxmin-expected utility with a non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Gul, F. and Pesendorfer, W. 2004. Self-control and the theory of consumption. *Econometrica* 72, 119–58.

Harsanyi, J. 1967–8. Games of incomplete information played by Bayesian players. Parts I, II, III. *Management Science* 14, 159–82, 320–34, 486–502.

Heifetz, A. and Samet, D. 1998a. Knowledge spaces with arbitrarily high rank. *Games and Economic Behavior* 22, 260–73.

Heifetz, A. and Samet, D. 1998b. Topology-free typology of beliefs. *Journal of Economic Theory* 82, 324–81.

Mariotti, T., Meier, M. and Piccione, M. 2005. Hierarchies of beliefs for compact possibility models. *Journal of Mathematical Economics* 41, 303–24.

Meier, M. 2005. On the nonexistence of universal information structures. *Journal of Economic Theory* 122, 132–9.

Mertens, J.F. and Zamir, S. 1985. Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory* 14, 1–29.

Morris, S. and Shin, H. 2003. Global games: theory and applications. In *Advances in Economics and Econometrics (Proceedings of the Eight World Congress of the Econometric Society)*, ed. M. Dewatripont, L. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

How to cite this article

Siniscalchi, Marciano. "epistemic game theory: beliefs and types." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000341> doi:10.1057/9780230226203.0490

The New Palgrave Dictionary of Economics Online

epistemic game theory: complete information

Adam Brandenburger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The epistemic programme can be viewed as a methodical construction of game theory from its most basic elements – rationality and irrationality, belief and knowledge about such matters, beliefs about beliefs, knowledge about knowledge, and so on. To date, the epistemic field has been mainly focused on game matrices and trees – that is, on the non-cooperative branch of game theory. It has been used to provide foundations for existing non-cooperative solution concepts, and also to uncover new solution concepts. The broader goal of the programme is to provide a method of analysing different sets of assumptions about games in a precise and uniform manner.

Keywords

admissibility; backward induction; common knowledge; conditional probability systems; correlation; epistemic game theory; epistemic game theory: complete information; finite games; invariance; iterated dominance; lexicographic probability systems; rational behaviour; rationalizability; strong dominance; type structures; uncertainty; weak dominance

Article

1 Epistemic analysis

Under the epistemic approach, the traditional description of a game is augmented by a mathematical framework for talking about the rationality or irrationality of the players, their beliefs and knowledge, and related ideas.

The first step is to add sets of *types* for each of the players. The apparatus of types goes back to Harsanyi (1967–8), who introduced it as a way to talk formally about the players' beliefs about the payoffs in a game, their beliefs about other players' beliefs about the payoffs, and so on. (See epistemic game theory: incomplete information.) But the technique is equally useful for talking about uncertainty about the actual play of the game – that is, about the players' beliefs about the strategies chosen in the game, their beliefs about other players' beliefs about the strategies, and so on. This survey focuses on this second source of uncertainty. It is also possible to treat both kinds of uncertainty together, using the same technique.

We give a definition of a type structure as commonly used in the epistemic literature, and an example of its use.

Fix an n-player finite strategic-form game $(S^1, \ldots, S^n, \pi^1, \ldots, \pi^n)$. Some notation: given sets $X^1, \ldots, X^n$, let $X = \times_{i=1}^{n} X^i$ and $X^{-i} = \times_{j \neq i} X^j$. Also, given a finite set Ω, write $\mathcal{M}(\Omega)$ for set of all probability measures on Ω.

Definition 1.1: An $(S^1, \ldots, S^n)$*-based* ***(finite) type structure*** *is a structure*

$$\left(S^1, \ldots, S^n; T^1, \ldots, T^n; \lambda^1, \ldots, \lambda^n\right),$$

where each T^i *is a finite set, and each* $\lambda^i : T^i \to \mathcal{M}(S^{-i} \times T^{-i})$. *Members of* T^i *are called* ***types*** *for player i. Members of S × T are called* ***states (of the world)****.*

For some purposes – see, for example, Sections 4 and 6 – it is important to consider infinite type structures. Topological assumptions are then made on the type spaces T_i.

A particular state (s^1, t^1, …, s^n, t^n) describes the strategy chosen by each player, and also each player's type. Moreover, a type t^i for player i induces, via a natural induction, an entire hierarchy of beliefs – about the strategies chosen by the players $j \neq i$, about the beliefs of the players $j \neq i$, and so on. (See epistemic game theory: beliefs and types.)

The following example is similar to one in Aumann and Brandenburger (1995, pp. 1166–7).

Example 1.1: (**A coordination game**). *Consider the coordination game in* Figure 1.1 *(where Ann chooses the row and Bob the column), and the associated type structure in* Figure 1.2.

Figure 1.1

	L	R
U	2, 2	0, 0
D	0, 0	1, 1

Figure 1.2

$\lambda^b(t^b)$

T^b	L	R
u^b	0	½
t^b	0	½

S^b

T^a	U	D
u^a	0	½
t^a	0	½

$\lambda^b(u^b)$

T^b	L	R
u^b	½	0
t^b	0	½

S^b

T^a	U	D
u^a	½	0
t^a	0	½

S^a S^a

There are two types t^a, u^a for Ann, and two types t^b, u^b for Bob. The measure associated with each type is as shown. Fix the state (D, t^a, R, t^b). At this state, Ann plays D and Bob plays R. Ann is 'correct' about Bob's strategy. (Her type t^a assigns probability 1 to Bob's playing R.) Likewise, Bob is correct about Ann's strategy. Ann, though, thinks it possible Bob is wrong about her strategy. (Her type assigns probability 1/2 to type u^b for Bob, which assigns probability 1/2 to Ann's playing U, not D.) Again, likewise with Bob.

What about the rationality or irrationality of the players? At state (D, t^a, R, t^b), Ann is rational. Her strategy maximizes her expected payoff, given her first-order belief (which assigns probability 1 to R). Likewise, Bob is rational. Ann, though, thinks it possible Bob is irrational. (She assigns probability 1/2 to (R, u^b). With type u^b, Bob gets a higher expected payoff from L than R.) The situation with Bob is again symmetric.

Summing up, the example is just a description of a game situation, not a prediction. A type structure is a descriptive tool. Note, too, that the example includes both rationality and irrationality, and also allows for incorrect as well as correct beliefs (for example, Ann thinks it possible Bob is irrational, though in fact he isn't). These are typical features of the epistemic approach.

Two comments on type structures. First, we can ask whether Definition 1.1 above is to be taken as primitive or derived. Arguably, hierarchies of beliefs are the primitive, and types are simply a convenient tool for the analyst. See epistemic game theory: beliefs and types for further discussion.

Second, note that Definition 1.1 applies to finite games. These will be the focus of this survey. There is nothing yet approaching a developed literature on epistemic analysis of infinite games.

2 Early results

A major use of type structures is to identify conditions on the players' rationality, beliefs, and so on, that yield various solution concepts.

A very basic solution concept is iterated dominance. This involves deleting from the matrix all strongly dominated strategies, then deleting all strategies that become strongly dominated in the resulting submatrix, and so on until no further deletion is possible. (It is easy to check that in finite games – as considered in this survey – the residual set will always be non-empty.) Call the remaining strategies the *iteratively undominated* (*IU*) strategies. There is a basic equivalence: a strategy is not strongly dominated if and only if there is a probability measure on the product of the other players' strategy sets under which it is optimal. Using this, IU can also be defined as follows: delete from the matrix any strategy that isn't optimal under some measure on the product of the other players' strategy sets. Consider the resulting sub-matrix and delete strategies that don't pass this test on the sub-matrix, and so on.

The second definition suggests what a formal epistemic treatment of IU should look like. A rational player will choose a strategy which is optimal under some measure. This is the first round of deletion. A player who is rational and believes the other players are rational will choose a strategy which is optimal under a measure that assigns probability 1 to the strategies remaining after the first round of deletion. This gives the second round of deletion. And so on.

Type structures allow a formal treatment of this idea. First the formal definition of rationality. This is a property of strategy-type pairs. Say (s^i, t^i) is **rational** if s^i maximizes player i's expected payoff under the marginal on S^{-i} of the measure $\lambda^i(t^i)$.

Say type t^i of player i **believes** an event $E \subseteq S^{-i} \times T^{-i}$ if $\lambda^i(t^i)(E) = 1$, and write

$$B^i(E) = \left\{ t^i \in T^i : t^i \text{ believes } E \right\}.$$

Now, for each player i, let R_1^i be the set of all rational pairs (s^i, t^i), and for m>0 define R_m^i inductively by

$$R_{m+1}^i = R_m^i \cap [S^i \times B^i(R_m^{-i})].$$

Definition 2.1: If $(s^1, t^1, \ldots, s^n, t^n) \in R_{m+1}$, *say there is* ***rationality and****m* ***th-order belief of rationality (****R****m****BR****) at this state. If* $(s^1, t^1, \ldots, s^n, t^n) \in \cap_{m=1}^{\infty} R_m$, *say there is* ***rationality and common belief of rationality (RCBR****) at this state*.

These definitions yield an epistemic characterization of IU: *Fix a type structure and a state* ($s^1, t^1, \ldots, s^n, t^n$) *at which there is RCBR. Then the strategy profile* ($s^1, \ldots, s^n$) *is IU. Conversely, fix an IU profile* ($s^1, \ldots, s^n$). *There is a type structure and a state* ($s^1, t^1, \ldots, s^n, t^n$) *at which there is RCBR*. Results like this can be found in the early literature – see, among others, Brandenburger and Dekel (1987) and Tan and Werlang (1988).

An important stimulus to the early literature was the pair of papers by Bernheim (1984) and Pearce (1984), which introduced the solution concept of *rationalizability*. This differs from IU by requiring on each round that a player's probability measure on the product of the other players' (remaining) strategy sets be a product measure – that is, be independent. Thus the set of rationalizable strategy profiles is contained in the IU set. It is well known that there are games (with three or more players) in which inclusion is strict.

The argument for the independence assumption is that in non-cooperative game theory it is supposed that players do not coordinate their strategy choices. Interestingly though, correlation is consistent with the non-cooperative approach. This view is put forward in Aumann (1987). (Aumann, 1974, introduced the study of correlation into non-cooperative theory.) Consider an analogy to coin tossing. A correlated assessment over coin tosses is possible, if there is uncertainty over the coin's parameter or 'bias'. (The assessment is usually required to be conditionally i.i.d., given the parameter.) Likewise, in a game, Charlie might have a correlated assessment over Ann's and Bob's strategy choices, because, say, he thinks Ann and Bob have observed similar signals before the game (but is uncertain what the signal was).

The same epistemic tools used to understand IU can be used to characterize other solution concepts on the matrix. Aumann and Brandenburger (1995, Preliminary Observation) point out that pure-strategy Nash equilibrium is characterized by the simple condition that each player is rational and assigns probability 1 to the actual strategies chosen by the other players. (Thus, in Example 1.1 above, these conditions hold at the state (D, t^a, R, t^b), and (D, R) is indeed a Nash equilibrium.) As far as mixed strategies are concerned, in the epistemic approach to games these don't play the central role that they do under equilibrium analysis. Built into the set-up of Section 1 is that each player makes a definite choice of (pure) strategy. (If a player does have the option of making a randomized choice, this can be added to the – pure – strategy set. Indeed, in a finite game, a finite number of such choices can be added.) It is the other players who are uncertain about this choice. Harsanyi (1973) originally proposed this shift in thinking about randomization. Aumann and Brandenburger (1995) give an epistemic treatment of mixed-strategy Nash equilibrium along these lines.

Aumann (1987) asks a question about an outside observer of a game. He provides conditions under which the observer's assessment of the strategies chosen will be the distribution of a correlated equilibrium (as defined in his 1974 paper). The distinctive condition in (1987) is the so-called Common Prior Assumption, which says that the probability assessment associated with each player's type is the same as the observer's assessment, except for being conditioned on what the type in question knows. A number of papers have investigated foundations for this assumption – see, among others, Morris (1994), Samet (1998), Bonanno and Nehring (1999), Feinberg (2000), Halpern (2002), and also the exchange between Gul (1998) and Aumann (1998).

2 Next steps: the tree

An important next step in the epistemic programme was extending the analysis to game trees. A big motivation for this was to understand the logical foundation of *backward induction* (*BI*). At first sight, BI is one of the easiest ideas in game theory. If Ann, the last player to move, is rational, she will make the BI choice. If Bob, the second-to-last player to move, is rational and thinks Ann is rational, he will make the choice that is maximal given that Ann makes the BI choice – that is, he too will make the BI choice. And so on back in the tree, until the BI path is a identified (Aumann, 1995).

For example, Figure 3.1 is three-legged centipede (Rosenthal, 1981). (The top payoffs are Ann's, and the bottom payoffs are Bob's.) BI says Ann plays *Out* at her first node. But what if she doesn't? How will Bob react? Perhaps Bob will conclude that Ann is an irrational player, who plays *Across*. That is, Bob might play *In*, hoping to get a payoff of 6 (better than 4 from *Out*). Perhaps, anticipating this, Ann will in fact play *Down*, hoping to get 4 (better than 2 from playing *Out*).

Figure 3.1

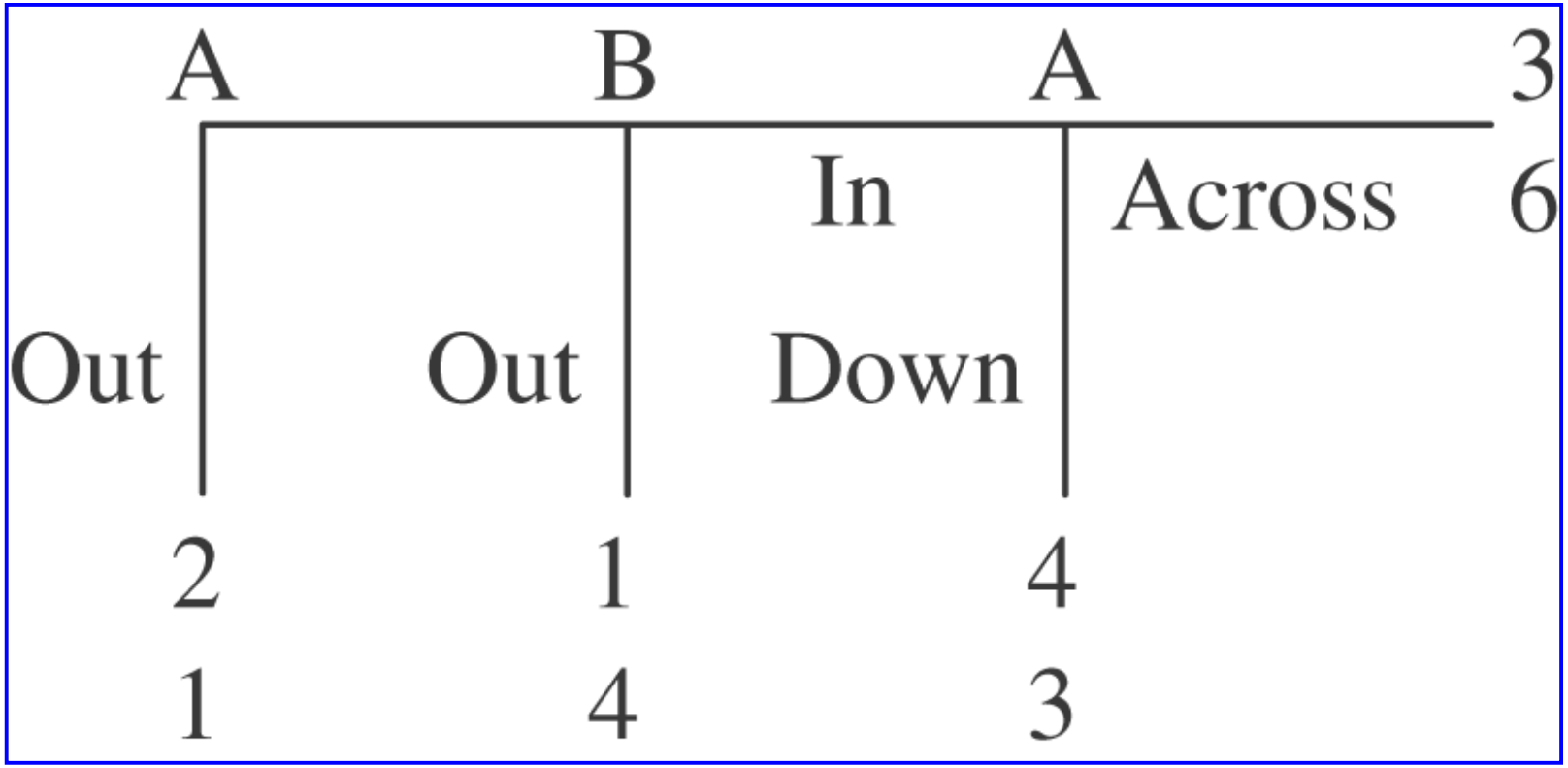

Many papers have examined this conceptual puzzle with BI – see, among others, Binmore (1987), Bicchieri (1988, 1989), Basu (1990), Bonanno (1991), and Reny (1992).

A key step in resolving the puzzle is extending the epistemic tools of Section 1, to be able to talk formally about rationality, beliefs and so on in the tree.

Example 3.1: (**three-Legged centipede**). Figure 3.2 *is a type structure for three-legged Centipede*.

Figure 3.2

$\lambda^a(t^a)$

T^b	Out	In
u^b	0	0
t^b	0	1

S^b

$\lambda^a(u^a)$

T^b	Out	In
u^b	1 [0]	0 [0]
t^b	0 [0]	0 [1]

S^b

$\lambda^b(t^b)$

T^a	Out	Down	Across
u^a	1 [0]	0 [0]	0 [0]
t^a	0 [0]	0 [0]	0 [1]

S^a

$\lambda^b(u^b)$

T^a	Out	Down	Across
u^a	1 [0]	0 [0]	0 [0]
t^a	0 [0]	0 [1]	0 [0]

S^a

There are two types t^a, u^a for Ann. Type t^a for Ann has the measure shown in the top-left matrix. It assigns probability 1 to (In, t^b) for Bob. Type u^a has two associated measures – shown in the top-right matrix. The first measure (the numbers without parentheses) assigns probability 1 to (Out, u^b) for Bob. In this case, we also specify a second measure for Ann, because we want to specify what Ann thinks at her second node, too. Reaching this node is assigned positive probability (in fact, probability 1) under Ann's type t^a, but probability 0 under her type u^a. So, for type u^a, there isn't a well-defined conditional probability measure at Ann's second node. This is why we (separately) specify a second measure for Ann's type u^a: it is the measure in square brackets. If type u^a, Ann assigns probability 1 to (In, t^b) at her second node.

There are also two types t^b, u^b for Bob. Both types initially assign probability 1 to Ann's playing Out. For both of Bob's types, there isn't a well-defined conditional probability measure at his node. At his node, Bob's type t^b assigns probability 1 to {(Across, t^a)}, while his type u^b assigns probability 1 to {(Down, t^a)}.

This is a simple illustration of the concept of a *conditional probability system* (*CPS*), due to Rényi (1955). A CPS specifies a family of conditioning events E and a measure p_E for each such event, together with certain restrictions on these measures. The interpretation is that p_E is what the player believes, after observing E. Even if $p_\Omega(E)=0$ (where Ω is the entire space), the measure p_E is still specified. That is, even if E is 'unexpected', the player has a measure if E nevertheless happens. This is why CPS's are well-suited to epistemic analysis of game trees – where we need to be able to describe how players react to the unexpected.

Myerson (1991, ch. 1) provided a preference-based axiomatization of a class of CPS's. Battigalli and Siniscalchi (1999; 2002) further developed both the pure theory and the game-theoretic application of CPS's (see below).

Suppose the true state in Figure 3.2 is (*Down*, t^a, *In*, t^b). In particular, Ann plays *Down*, expecting Bob to play *In*. Bob plays *In*, expecting (at his node) Ann to play *Across*. Ann expects a payoff of 4 (and gets this). Bob expects a payoff of 6 (but gets only 3). In everyday language, we can say that Ann successfully bluffs Bob. (At the state (*Down*, t^a, *In*, t^b), the bluff works. By contrast, at the state (*Down*, t^a, *Out*, u^b), Ann attempts the bluff and it fails.)

But what about epistemic conditions? Are the players rational in this situation? Does each think the other is rational? And so on.

To answer, we need a definition of rationality with CPS's. Fix a strategy-type pair (s^i, t^i), where t^i is associated with a CPS. Call this pair **rational** (**in the tree**) if the following holds: fix any information set H for i allowed by s^i, and look at the measure on the other players' strategies, given H. (This means given the event that the other players' strategies allow H.) Require that s^i maximizes i's expected payoff under this measure, among all strategies r^i of i that allow H.

With this definition, the rational strategy-type pairs in Figure 3.2 are (*Down*, t^a), (*Out*, u^a), (*In*, t^b), and (*Out*, u^b).

Next, what does Ann think about Bob's rationality? To answer, we need a CPS-analogue to belief (as defined in Section 2). Ben Porath (1997) proposed the following (we have taken the liberty of changing terminology, for consistency with 'strong belief' below): Say player i **initially believes** event E if, under i's CPS, E gets probability 1 at the root of the tree. (Formally, the conditioning event consists of all strategy profiles of the other players.) Battigalli and Siniscalchi (2002) strengthened this definition to: Say player i **strongly believes** event E if, under i's CPS, E gets probability 1 at every information set at which E is possible. Under initial belief, E also gets probability 1 at any information set H that gets positive probability under i's initial measure (that is, i's measure given the root). This is just standard conditioning on non-null events. But under strong belief, this conclusion holds for any information set H which has a non-empty intersection with E – even if H is null under i's initial measure. This is why strong belief is stronger than initial belief.

Let us apply these definitions to Figure 3.2. Does Ann initially believe that Bob is rational? Yes. Both of Ann's types initially believe Bob is rational. Type t^a initially assigns probability 1 to the rational pair (*In*, t^b). Type u^a initially assigns probability 1 to the rational pair (*Out*, u^b). In fact, both types strongly believe Bob is rational. Since, under type t^a, Ann's second node gets positive probability (in fact, probability 1) under her initial measure, we need only check this for type u^a. But at Ann's second node, type u^a assigns probability 1 to the rational pair (*In*, t^b).

Turning to Bob, both of his types initially believe that Ann is rational. Type u^b even strongly believes Ann is rational; but type t^b doesn't. This is because, at Bob's node, type t^b assigns positive probability (in fact, probability 1) to the irrational pair (*Across*, t^a).

Staying with initial belief (we come back to strong belief below), we can parallel Definition 2.1 and define inductively **rationality and *m*th-order initial belief of rationality** (**R*m*IBR**) at a state of a type structure, and **rationality and common initial belief of rationality (RCIBR)** (see Ben Porath, 1997). In Figure 3.2, since all four types initially believe the other player is rational, a simple induction gives that at the state (*Down*, t^a, *In*, t^b) for instance, RCIBR holds.

In words, Ann plays across at her first node, believing (initially) that Bob will play *In*, so she can get a payoff of 4. Why would Bob play In? Because he initially believes that Ann plays *Out*. But in the probability-0 event that Ann plays across at her first node, Bob then assigns probability 1 to Ann's playing across at her second node – that is, to Ann's being irrational. He therefore (rationally) plays *In*. All this is consistent with RCIBR.

4 Conditions for backward induction

Interestingly, this is exactly the line of reasoning which, as we said, was the original stimulus for investigating the foundations of BI. So, there is no difficulty with it – we've just seen a formal set-up in which it holds. The resolution of the BI puzzle is simply to accept that the BI path may not result.

But one can also argue that RCIBR is not the right condition: it is too weak. In the above example, Bob realizes that he might be 'surprised' in the play of the game – that's why he has a CPS, not just an ordinary probability measure. If he realizes he might be surprised, should he abandon his (initial) belief that Ann is rational when he is surprised? Bob's type t^b does so. This is the step taken by Battigalli and Siniscalchi (2002) with their concept of strong belief. The argument says that we want t^b to strongly believe, not just initially believe, that Ann is rational. Type t^b will strongly believe Ann is rational if we move the probability-1 weight (in square brackets) on (*Across*, t^a) to (*Down*, t^a). But now (*In*, t^b) isn't rational for Bob, so Ann doesn't (even initially) believe Bob is rational. It looks as if the example unravels.

We can again parallel Definition 2.1 and define inductively **rationality and *m*th-order strong belief of rationality** (**R*m*SBR**), and **rationality and common strong belief of rationality (RCSBR)** (see Battigalli and Siniscalchi, 2002). The question is then: does RCSBR yield BI?

The answer is yes. *Fix a CPS-based type structure for n-legged Centipede* (Figure 4.1), *and a state at which there is RCSBR. Then Ann plays Out*. The result follows from Friedenberg (2002), who shows that in a PI game (satisfying certain payoff restrictions), RCSBR yields a Nash-equilibrium outcome. In Centipede, there is a unique Nash path and it coincides with the BI path. Of course, this isn't true in general.

Example 4.1: **(A second coordination game)** *Consider the coordination game in* Figure 4.2 *and the associated CPS-based type structure in* Figure 4.3.

Figure 4.1

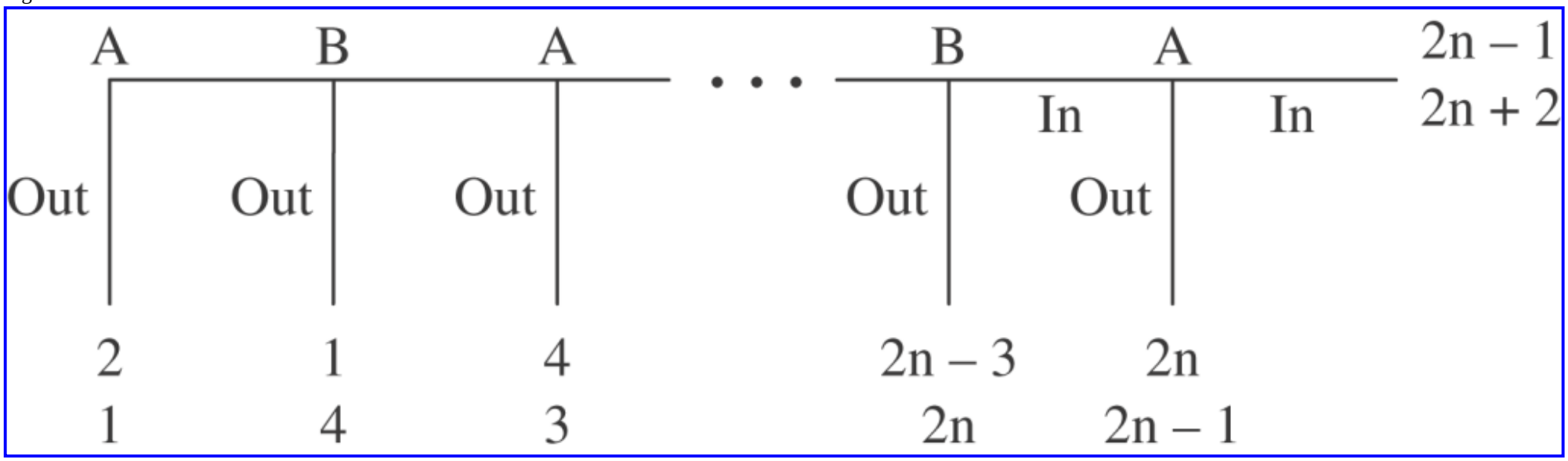

Figure 4.2

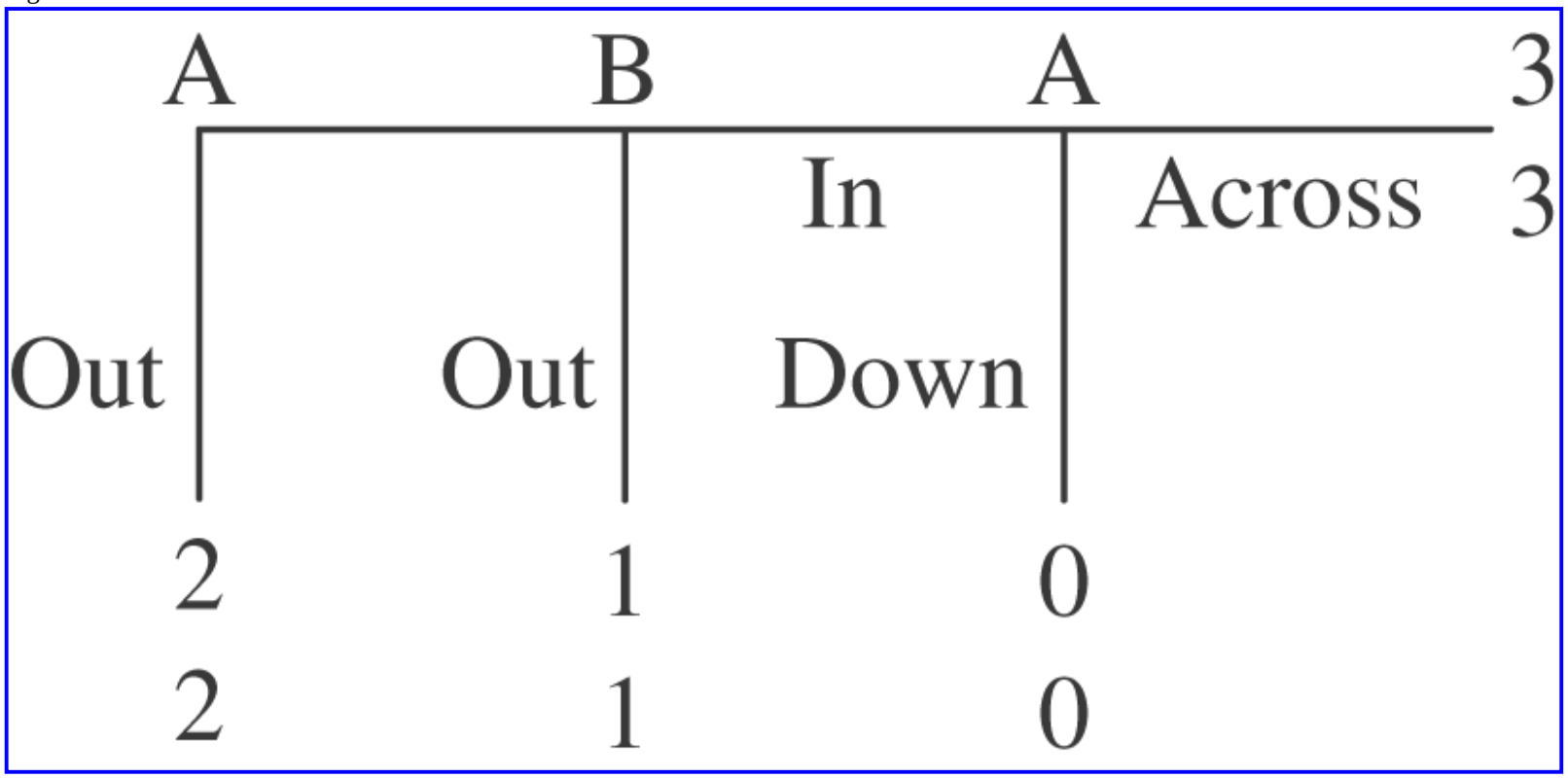

Figure 4.3

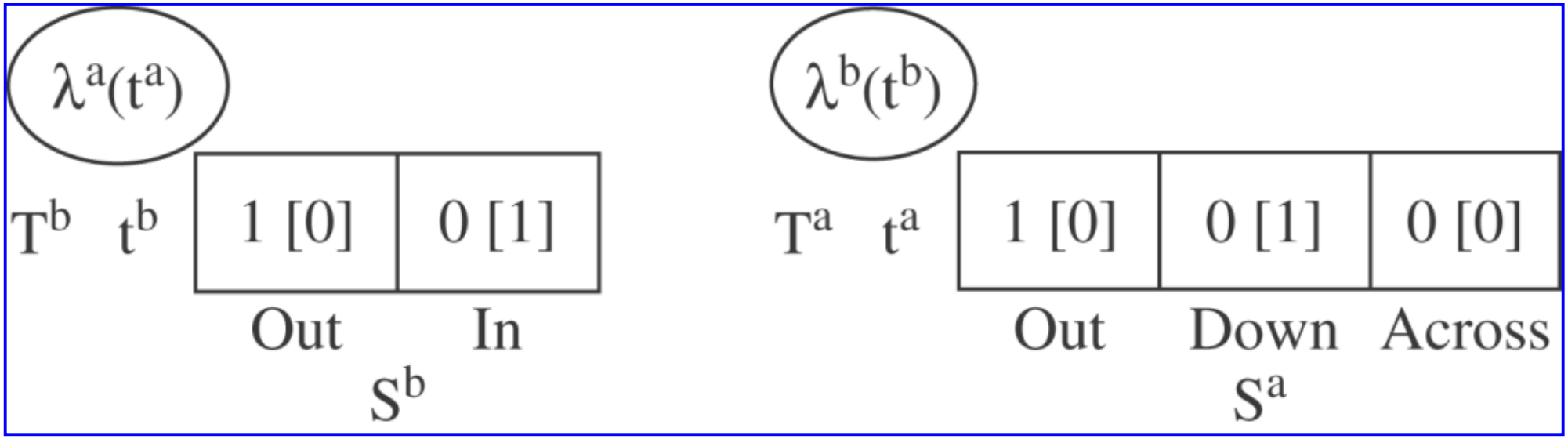

The rational strategy-type pairs are (*Out*, t^a) *and* (*Out*, t^b) *for Ann and Bob respectively. Ann's type* t^a *strongly believes* {(*Out*, t^b)}, *and Bob's type* t^b *strongly believes* {(*Out*, t^a)}. *By induction, RCSBR holds at the state* (*Out*, t^a, *Out*, t^b).

Here, the BI path need not be played under RCSBR. The key is to see that both (*Down*, t^a) and (*Across*, t^a) are irrational for Ann, since she (strongly) believes Bob plays *Out*. So at his node, Bob can't believe Ann is rational. If he considers it sufficiently more likely Ann will play *Down* rather than *Across*, he will rationally play *Out* (as happens). In short, if Ann doesn't play *Out*, she is irrational and so 'all bets are off' as to what she will do. She could play *Down*.

This situation may be surprising, at least at first blush, but there does not appear to be anything conceptually wrong with it. Indeed, it points to an interesting way in which the players in a game can literally be trapped by their beliefs – which here prevent them from getting their mutually preferred (3, 3) outcome.

But one can also argue differently. If Ann forgoes the payoff of 2 she can get by playing *Out* at the first node, then surely she must be playing *Across* to get 3. Playing *Down* to get 0 makes little sense since this is lower than the payoff she gave up at the first node. (This is forward-induction reasoning à la Kohlberg and Mertens, 1986, Section 2.3, introduced in the context of non-PI games. Interestingly, epistemic analysis makes clear that the issue already arises in PI games, such as Figure 4.2.) But if Bob considers *Across* (sufficiently) more likely than *Down*, he will play *In*. Presumably then, Ann will indeed play *Across*, and the BI path results.

There is no contradiction with the previous analysis because in Figure 4.3 Ann is irrational once she doesn't play *Out*, so we can't say Ann should then rationally play *Across* not *Down*. To make *Across* rational for Ann, we have to add more types to the structure – specifically, we would want to add a second type for Ann that assigns (initial) probability 1 to Bob's playing *In* not *Out*. This key insight is due to Stalnaker (1998) and Battigalli and Siniscalchi (2002).

Battigalli and Siniscalchi formulate a general result of this kind. They consider a **complete** CPS-based type structure, which contains, in a certain sense, every possible type for each player (a complete type structure will be uncountably infinite), and prove: *Fix a complete CPS-based type structure. If there is RCSBR at the state* (s^1, t^1, …, s^n, t^n), *then the strategy profile* (s^1, …, s^n) *is extensive-form rationalizable. Conversely, if the profile* (s^1, …, s^n) *is extensive-form rationalizable, then there is a state* (s^1, t^1, …, s^n, t^n) *at which there is RCBR.* The extensive-form rationalizability strategies (Pearce, 1984) yield the BI outcome in a PI game (under an assumption ruling out certain payoff ties; Battigalli, 1997), so the Battigalli and Siniscalchi analysis gives epistemic conditions for BI.

There are other routes to getting BI in PI games. Asheim (2001) develops an epistemic analysis using the properness concept (Myerson, 1978). Go back to Example 4.1. The properness idea says that Bob's type t^b should view (*Across*, t^a) as infinitely more likely than (*Down*, t^a) since *Across* is the less costly 'mistake' for Ann, given her type t^a. Unlike the completeness route taken above, the irrationality of both *Down* and *Across* (given Ann's type t^a) is accepted. But the relative ranking of these 'mistakes' must be in the right order. With this ranking, Bob is irrational to play *Out* rather than *In*. Ann presumably will play *Across*, and we get BI again. Asheim (2001) formulates a general such result.

Another strand of the literature on BI employs knowledge models rather than belief models. As pointed out in Example 1.1, players' beliefs don't have to be correct in any sense. For example, a type might even assign probability 1 to a strategy-type pair for another player different from the actual one. Knowledge as usually formalized is different, in that if a player knows an event *E*, then *E* indeed happens.

Aumann (1995) formulates a knowledge-based epistemic model for PI trees. In his set-up, the condition of common knowledge of rationality implies that the players choose their BI strategies. Stalnaker (1996) finds that non-BI outcomes are possible, under a different formulation of the same condition. The explanation lies in differences in how counterfactuals

are treated. These play an important role in a knowledge-based analysis, when we talk about what a player thinks at an information set that cannot be reached given what he knows. Halpern (2001) provides a synthesis in which these differences can be understood. See also the exchange between Binmore (1996) and Aumann (1996), and the analyses by Samet (1996), Balkenborg and Winter (1997), and Halpern (1999).

Aumann (1998) provides knowledge-based epistemic conditions under which Ann plays *Out* in Centipede. The conditions are weaker than in his (1995) paper, and the conclusion weaker (about outcomes not strategies). There is an obvious parallel between this result and the belief-based result on Centipede we stated above (also about outcomes). More generally, there may be an analogy between counterfactuals in knowledge models and extended probabilities in belief models. But, for one thing, completeness is crucial to the belief-based approach, as we have seen, and an analogous concept does not appear to be present in the knowledge-based approach. As yet, there does not appear to be any formal treatment of the relationship between the two approaches.

5 Next steps: weak dominance

Extending the epistemic analysis of games from the matrix to the tree has been the focus of much recent work in the literature. Another area has been extending the analysis on the matrix from strong dominance (described in Section 2) to weak dominance.

Weak dominance (admissibility) says that a player considers as possible (even if unlikely) any of the strategies for the other players. In the game context, we are naturally led to consider *iterated admissibility* (*IA*) – the weak-dominance analogue to IU. This is an old concept in game theory, going back at least to Gale (1953). Like BI, it is a powerful solution concept, delivering sharp answers in many games – Bertrand, auctions, voting games, and others. (Mertens, 1989, p. 582, and Marx and Swinkels, 1997, pp. 224–5, list various games involving weak dominance.)

But, also like BI, there is a conceptual puzzle. Suppose Ann conforms to the admissibility requirement, so that she considers possible any of Bob's strategies. Suppose Bob also conforms to the requirement, and this leads him not to play a strategy, say *L*. If Ann thinks Bob adheres to the requirement (as he does), then she can rule out Bob's playing *L*. But this conflicts with the requirement that she not rule anything out (see Samuelson, 1992).

Can a sound argument be made for IA? To investigate this, the epistemic tools of Section 1 have to be extended again.

Example 5.1: (**Bertrand)** Figure 5.1 *is a Bertrand pricing game, where each firm chooses a price in* {*0,1,•2,•3*}. (*Ken Corts kindly provided this example*.) *The left payoff is to A, the right payoff to B. Each firm has capacity of two units and zero cost. Two units are demanded. If the firms charge the same price, they each sell one unit*. Figure 5.2 *is an associated type structure* (*with one type for each player*).

Figure 5.1

	B			
	3	2	1	0
3	3, 3	0, 4	0, 2	0, 0

A 2	4, 0	2, 2	0, 2	0, 0
1	2, 0	2, 0	1, 1	0, 0
0	0, 0	0, 0	0, 0	0, 0

Figure 5.2

$\lambda^a(t^a)$

T^b t^b	1	0	0	0
	0	1	2	3
	S^b			

$\lambda^b(t^b)$

T^a t^a	1	0	0	0
	0	1	2	3
	S^a			

The rational strategy-type pairs are $R_1^a = \{0, 1, 2, 3\} \times \{t^a\}$ *and* $R_1^b = \{0, 1, 2, 3\} \times \{t^b\}$. *Since both types assign positive probability only to a rational strategy-type pair for the other player, we get* $R_m^a = R_1^a$ *and* $R_m^b = R_1^b$ *for all m. In particular, there is RCBR at the state* $(3, t^a, 3, t^b)$.

But a price of 3 is inadmissible (as is a price of 0). The IA set is just {(1,1)}, where each firm charges the lowest price above cost. (This is a plausible scenario: while pricing at cost is inadmissible, competition forces price down to the first price above cost.)

A tool to incorporate admissibility is *lexicographic probability systems* (*LPS*'s), introduced and axiomatized by Blume, Brandenburger and Dekel (1991a; 1991b). An LPS specifies a sequence of probability measures. The interpretation is that the first measure is the player's primary hypothesis about the true state. But the player recognizes that his primary hypothesis might be mistaken, and so also forms a secondary hypothesis. This is his second measure. Then his tertiary hypothesis, and so on. The primary states can be thought of as infinitely more likely than the secondary states, which are infinitely more likely than the tertiary states, and so on. Stahl (1995), Stalnaker (1998), Asheim (2001), Brandenburger, Friedenberg and Keisler (2006), and Asheim and Perea (2005), among other papers, use LPS's.

Example 5.2: (**Bertrand contd.)** Figure 5.3 *is a type structure for Bertrand* (Figure 5.1) *that now specifies LPS's.*

Figure 5.3

$\lambda^a(t^a)$

T^b t^b	1	(1/3)	(1/3)	(1/3)
	0	1	2	3
	S^b			

$\lambda^b(t^b)$

T^a t^a	1	(1/3)	(1/3)	(1/3)
	0	1	2	3
	S^a			

Each player has a primary hypothesis which assigns probability 1 to the other player's charging a price of 0. But each player also has a secondary hypothesis that assigns equal probability to each of the three remaining choices for the other player. This measure is shown in parentheses. Note that every state (that is, strategy-type pair) gets positive probability under some measure. But states can also be ruled out, in the sense that they can be give infinitely less weight than other states.

What about epistemic conditions? Are the players rational in this situation? Does each think the other is rational? And so on.

To answer, we need a definition of rationality with LPS's. Fix strategy-type pairs (s^i, t^i) and (r^i, t^i) for player i, where t^i is now associated with an LPS. Calculate the tuple of expected payoffs to i from s^i, using first the primary measure associated with t^i, then the secondary measure associated with t^i, and so on. Calculate the corresponding tuple for r^i. If the first tuple lexicographically exceeds the second, then s^i is preferred to r^i. (If $x=(x_1, \ldots, x_n)$ and $y=(y_1, \ldots, y_n)$, then x lexicographically exceeds y if $y_j>x_j$ implies $x_k>y_k$ for some $k<j$.) A strategy-type pair (s^i, t^i) is **rational** (**in the lexicographic sense**) if s^i is maximal under this ranking.

So (3, t^a) and (3, t^b) are irrational. All choices give each player an expected payoff of 0 under the primary measure. But a price of 2 gives each player an expected payoff of 2 under the secondary measure, as opposed to an expected payoff of 1 from a price of 3. Conceptually, we want (3, t^a) and (3, t^b) to be irrational (because a price of 3 is inadmissible).

What does each player think about the other's rationality? For this, we again need an LPS-based definition. An early candidate in the literature was: Say player *i* **believes** event *E* **at the** 1**st level** if *E* gets primary probability 1 under *i*'s LPS (Börgers, 1994; Brandenburger, 1992). A stronger concept is: Say *i* **assumes** *E* if all states not in *E* are infinitely less likely than all states in *E*, under *i*'s LPS (Brandenburger, Friedenberg and Keisler, 2006). In other words, a player who assumes *E* recognizes *E* may not happen, but is prepared to 'count on' *E* versus not-*E*.

In Figure 5.3, type t^a doesn't 1st-level believe (so certainly doesn't assume) the other player is rational. Likewise with t^b. Again, this is right conceptually.

6 Conditions for iterated admissibility

Once again we can parallel Definition 2.1 and define inductively **rationality and** *m***th-order** 1**st-level belief of rationality** (**R***m*1**BR**) at a state of a type structure, and **rationality and common** 1**st-level belief of rationality** (**RC**1**BR**). Likewise, one can define **rationality and** *m***th-order assumption of rationality (R***m***AR), and rationality and common assumption of rationality (RCAR)**. What do these conditions yield?

In fact, just as we saw in Sections 3 and 4 that neither RCIBR not RCSBR yields BI, so neither RC1BR nor RCAR yields IA. RC1BR is characterized by the $S^{\infty}W$ concept (Dekel and Fudenberg, 1990), that is , the set of strategies that remain after one round of deletion of inadmissible strategies followed by iterated deletion of strongly dominated strategies. RCAR is characterized by the self-admissible set concept (Brandenburger, Friedenberg and Keisler, 2006). Self-admissible sets may be viewed as the weak-dominance analogue to Pearce (1984) best-response sets.

But while the IA set is one self-admissible set in a game, there may well be others. To select the IA set, a completeness assumption is needed, similar to Section 4: *Fix a complete LPS-based type structure. If there is RmAR at the state* (s^1, t^1, …, s^n, t^n), *then the strategy profile* (s^1, …, s^n) *survives* (m+1) *rounds of iterated admissibility. Conversely, if the profile* (s^1, …, s^n) *survives* (m+1) *rounds of iterated admissibility, then there is a state* (s^1, t^1, …, s^n, t^n) *at which there is RmAR* (Brandenburger, Friedenberg and Keisler, 2006).

This result is stated for R*m*AR and not RCAR. See the next section for the reason. Of course, for a given game, there is an *m* such that IA stabilizes after *m* rounds.

IA yields the BI outcome in a PI game (again ruling out certain payoff ties; Marx and Swinkels, 1997), so, understanding IA gives, in particular, another analysis of BI.

Related analyses of IA include Stahl (1995) and Ewerhart (2002). Stahl uses LPS's and directly assumes that Ann considers one of Bob's strategies infinitely less likely than another if the first is eliminated on an earlier round of IA than the second. Ewerhart gives an analysis of IA couched in terms of provability (from mathematical logic).

7 Strategic versus extensive analysis

Kohlberg and Mertens (1986, Section 2.4) argued that a 'fully rational' analysis of games should be invariant – that is , should depend only on the fully reduced strategic form of a game. (This is the strategic form after elimination of any – pure – strategies that are duplicates or convex combinations of other strategies.) In this, they appealed to early results in game theory (Dalkey, 1953; Thompson, 1952) which established that two trees sharing the same reduced strategic form differ from each other by a (finite) sequence of elementary transformations of the tree, each of which can be argued to be 'strategically inessential'. Kohlberg and Mertens added a fourth transformation involving convex combinations, to get to the fully reduced strategic form.

In decision theory, invariance is implied by (and implies) admissibility. (Kohlberg and Mertens, 1986, Section 2.7, gave the essential idea. See Brandenburger, 2007, for the decision-theory argument.) If we build up our game analysis using a decision theory that satisfies admissibility, we can hope to get invariance at this level too. LPS-based decision theory satisfies admissibility. Indeed, IA, and also the $S^{\infty}W$ and self-admissible set concepts, are invariant in the Kohlberg–Mertens sense. The extensive-form rationalizability concept (Section 4) is not.

There does appear to be a price paid for invariance, however. The extensive-form conditions of RCSBR and (CPS-based) completeness are consistent (in any tree). That is, for any tree, we can build a complete type structure and find a state at which RCSBR holds. But Brandenburger, Friedenberg and Keisler (2006) show the strategic-form conditions of RCAR and (LPS-based) completeness are inconsistent (in any matrix satisfying a non-triviality condition).

A possible interpretation is that rationality, even as a theoretical concept, appears to be inherently limited. There are purely theoretical limits to the Kohlberg-Mertens notion of a 'fully rational' analysis of games.

The epistemic programme has uncovered a number of impossibility results (see epistemic game theory: beliefs and types for some others). We don't see this as a deficiency of the programme, but rather as a sign it has reached a certain depth and maturity. Also, central to the programme is the analysis of scenarios (we have seen several in this survey) that are 'a long way from' these theoretical limits. Under the epistemic approach to game theory there is not one right set of assumptions to make about a game.

See Also

- epistemic game theory: an overview
- epistemic game theory: beliefs and types
- epistemic game theory: incomplete information
- game theory
- Nash equilibrium, refinements of

This survey is based on Brandenburger (2007). I am grateful to Springer for permission to use this material. I owe a great deal to joint work and many conversations with Robert Aumann, Eddie Dekel, Amanda Friedenberg, Jerry Keisler and Harborne Stuart. My thanks to Konrad Grabiszewski for important input, John Nachbar for very important editorial advice, and Michael James for valuable assistance. The Stern School of Business provided financial support.

Bibliography

Asheim, G. 2001. Proper rationalizability in lexicographic beliefs. *International Journal of Game Theory* 30, 453–78.

Asheim, G. and Perea, A. 2005. Sequential and quasi-perfect rationalizability in extensive games. *Games and Economic Behavior* 53, 15–42.

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96.

Aumann, R. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55, 1–18.

Aumann, R. 1995. Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8, 6–19.

Aumann, R. 1996. Reply to Binmore. *Games and Economic Behavior* 17, 138–46.

Aumann, R. 1998. On the centipede game. *Games and Economic Behavior* 23, 97–105.

Aumann, R. 1998. Common priors: a reply to Gul. *Econometrica* 66, 929–38.

Aumann, R. and Brandenburger, A. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63, 1161–80.

Balkenborg, D. and Winter, E. 1997. A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics* 27, 325–45.

Basu, K. 1990. On the existence of a rationality definition for extensive games. *International Journal of Game Theory* 19, 33–44.

Battigalli, P. 1997. On rationalizability in extensive games. *Journal of Economic Theory* 74, 40–61.

Battigalli, P. and Siniscalchi, M. 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory* 88, 188–230.

Battigalli, P. and Siniscalchi, M. 2002. Strong belief and forward-induction reasoning. *Journal of Economic Theory* 106, 356–91.

Ben Porath, E. 1997. Rationality, Nash equilibrium, and backward induction in perfect information games. *Review of Economic Studies* 64, 23–46.

Bernheim, D. 1984. Rationalizable strategic behavior. *Econometrica* 52, 1007–28.

Bicchieri, C. 1988. Strategic behavior and counterfactuals. *Synthese* 76, 135–69.

Bicchieri, C. 1989. Self-refuting theories of strategic interaction: a paradox of common knowledge. *Erkenntnis* 30, 69–85.

Binmore, K. 1987. Modelling rational players I. *Economics and Philosophy* 3, 179–214.

Binmore, K. 1996. A note on backward induction. *Games and Economic Behavior* 17, 135–7.

Blume, L., Brandenburger, A. and Dekel, E. 1991a. Lexicographic probabilities and choice under uncertainty. *Econometrica* 59, 61–79.

Blume, L., Brandenburger, A. and Dekel, E. 1991b. Lexicographic probabilities and equilibrium refinements. *Econometrica* 59, 81–98.

Bonanno, G. 1991. The logic of rational play in games of perfect information. *Economics and Philosophy* 7, 37–65.

Bonanno, G. and Nehring, K. 1999. How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory* 28, 409–34.

Börgers, T. 1994. Weak dominance and approximate common knowledge. *Journal of Economic Theory* 64, 265–76.

Brandenburger, A. 1992. Lexicographic probabilities and iterated admissibility. In *Economic Analysis of Markets and Games*, ed. P. Dasgupta, D. Gale, O. Hart and E. Maskin. Cambridge, MA: MIT Press.

Brandenburger, A. 2007. The power of paradox: some recent results in interactive epistemology. *International Journal of Game Theory* 35, 465–92.

Brandenburger, A. and Dekel, E. 1987. Rationalizability and correlated equilibria. *Econometrica* 55, 1391–402.

Brandenburger, A., Friedenberg, A. and Keisler, H.J. 2006. Admissibility in games. Unpublished, Stern School of Business, New York University.

Dalkey, N. 1953. Equivalence of information patterns and essentially determinate games. In *Contributions to the Theory of Games*, vol. 2, ed. H. Kuhn and A. Tucker. Princeton: Princeton University Press.

Dekel, E. and Fudenberg, D. 1990. Rational behavior with payoff uncertainty. *Journal of Economic Theory* 52, 243–67.

Ewerhart, C. 2002. Ex-ante justifiable behavior, common knowledge, and iterated admissibility. Unpublished, Department of Economics, University of Bonn.

Feinberg, Y. 2000. Characterizing common priors in terms of posteriors. *Journal of Economic Theory* 91, 127–79.

Friedenberg, A. 2002. When common belief is correct belief. Unpublished, Olin School of Business, Washington University.

Gale, D. 1953. A theory of *n*-person games with perfect information. *Proceedings of the National Academy of Sciences* 39, 496–501.

Gul, F. 1998. A comment on Aumann's Bayesian view. *Econometrica* 66, 923–7.

Halpern, J. 1999. Hypothetical knowledge and counterfactual reasoning. *International Journal of Game Theory* 28, 315–30.

Halpern, J. 2001. Substantive rationality and backward induction. *Games and Economic Behavior* 37, 425–35.

Halpern, J. 2002. Characterizing the common prior assumption. *Journal of Economic Theory* 106, 316–55.

Harsanyi, J. 1967–8. Games with incomplete information played by ‘Bayesian’ players, I–III. *Management Science* 14, 159–82, 320–34, 486–502.

Harsanyi, J. 1973. Games with randomly disturbed payoffs: a new rationale for mixed strategy equilibrium points. *International Journal of Game Theory* 2, 1–23.

Kohlberg, E. and Mertens, J.-F. 1986. On the strategic stability of equilibria. *Econometrica* 54, 1003–37.

Marx, L. and Swinkels, J. 1997. Order independence for iterated weak dominance. *Games and Economic Behavior* 18, 219–45.

Mertens, J.-F. 1989. Stable equilibria – a reformulation. *Mathematics of Operations Research* 14, 575–625.

Morris, S. 1994. Trade with heterogeneous prior beliefs and asymmetric information. *Econometrica* 62, 1327–47.

Myerson, R. 1978. Refinements of the Nash equilibrium concept. *International Journal of Game Theory* 1, 73–80.

Myerson, R. 1991. *Game Theory*. Cambridge, MA: Harvard University Press.

Pearce, D. 1984. Rational strategic behavior and the problem of perfection. *Econometrica* 52, 1029–50.

Reny, P. 1992. Rationality in extensive form games. *Journal of Economic Perspectives* 6(4), 103–18.

Rényi, A. 1955. On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungaricae* 6, 285–335.

Rosenthal, R. 1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* 25, 92–100.

Samet, D. 1996. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior* 17, 230–51.

Samet, D. 1998. Common priors and the separation of convex sets. *Games and Economic Behavior* 24, 172–4.

Samet, D. 1998. Iterated expectations and common priors. *Games and Economic Behavior* 24, 131–41.

Samuelson, L. 1992. Dominated strategies and common knowledge. *Games and Economic Behavior* 4, 284–313.

Stahl, D. 1995. Lexicographic rationalizability and iterated admissibility. *Economic Letters* 47, 155–9.

Stalnaker, R. 1996. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy* 12, 133–63.

Stalnaker, R. 1998. Belief revision in games: forward and backward induction. *Mathematical Social Sciences* 36, 31–56.

Tan, T. and Werlang, S. 1988. The Bayesian foundations of solution concepts of games. *Journal of Economic Theory* 45, 370–91.

Thompson, F. 1952. Equivalence of games in extensive form. Research Memorandum RM-759. The RAND Corporation.

How to cite this article

Brandenburger, Adam. "epistemic game theory: complete information." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000258> doi:10.1057/9780230226203.0491

The New Palgrave Dictionary of Economics Online

epistemic game theory: incomplete information

Aviad Heifetz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In a game of incomplete information some of the players possess private information which may be relevant to the strategic interaction. Private information is modelled by a *type space*, in which every type of each player is associated with a belief about the basic issues of uncertainty (like payoffs) and about the other players' types. At a *Bayesian equilibrium* each type chooses a strategy which maximizes its expected payoff given the choice of strategies by the other players' types. Bayesian equilibrium payoffs are often inefficient relative to the equilibrium payoffs that would result had the players been fully informed.

Keywords

Bayesian equilibrium; Bayesian strategies; common knowledge; epistemic game theory: incomplete information; games with incomplete information; private information

Article

A game of incomplete information is a game in which at least some of the players possess private information which may be relevant to the strategic interaction. The private information of a player may be about the payoff functions in the game, as well as about some exogenous, payoff-irrelevant events. The player may also form beliefs about other players' beliefs about payoffs and exogenous events, about their beliefs about the beliefs of others, and so forth.

Harsanyi (1967–8) introduced the idea that such a state of affairs can be succinctly described by a *type space*. With this formulation, T_i denotes the set of player i's *types*. Each type $t_i \in T_i$ is associated with a belief $\lambda_i(t_i) \in \Delta(K \times T_{-i})$ about some basic space of uncertainty, K, and the combination T_{-i} of the other players' types. The basic space of uncertainty K is called the space of *states of nature*, and $\Omega = K \times \Pi_{i \in I} T_i$, where I is the set of players, is called the space of *states of the world*.

A type space models a game of incomplete information once each state of nature $k \in K$ is associated with

a payoff matrix of the game, or, more generally, with a payoff function u_i^k for each player $i \in I$. This payoff function specifies the player's payoff $u_i^k(s)$ for each combination of strategies $s=(s_i)_{i \in I} \in S = \prod_{i \in I} S_i$ of the players. (In the particular case in which k is associated with a payoff matrix, that is, the game is such that each player has finitely many strategies, the payoffs $u_i^k(s)$ to the players $i \in I$ appear in the entry of the matrix corresponding to the combination of strategies $s=(s_i)_{i \in I}$.) As usual, the set of strategies S_i of player $i \in I$ may be a complex object by itself. For instance, it may be the set of mixed strategies over some set of pure strategies S_i^0. The payoff function of player i in the state of nature k is $u_i^k : S \to \mathbb{R}$.

Obviously, different types of a player may want to choose different strategies. Thus, a *Bayesian strategy* of player i in a game of incomplete information specifies the strategy $\sigma_i(t_i) \in S_i$ that the player chooses given each one of her types $t_i \in T_i$.

Given a profile of Bayesian strategies $\sigma = (\sigma_j : T_j \to S_j)_{j \in I}$ of the players, the expected payoff of player i of type t_i is

$$U_i(\sigma, t_i) = \sum_{(k, t_{-i}) \in K \times T_{-i}} u_i^k(\sigma_i(t_i), \sigma_{-i}(t_{-i})) \times \lambda_i(t_i)(k, t_{-i})$$

where $\sigma_{-i}(t_{-i}) = (\sigma_j(t_j))_{j \neq i}$. If there is a continuum of states of nature and types, the sum becomes an integral:

$$U_i(\sigma, t_i) = \int_{K \times T_{-i}} u_i^k(\sigma_i(t_i), \sigma_{-i}(t_{-i})) \, d\lambda_i(t_i)(k, t_{-i})$$

(In this case, the expected payoff function $U_i(\sigma, t_i)$ is well defined if the Bayesian strategies $\sigma_j : T_j \to S_j$ are measurable functions and if the payoff function $u_i^k : K \times S \to R$ is measurable as well; we omit the details of this technical requirement).

We assume that the players are expected payoff maximizers. Thus, player i prefers the Bayesian strategy σ over σ' if and only if $U_i(\sigma, t_i) \geq U_i(\sigma', t_i)$ for each of her types $t_i \in T_i$. It follows that given a Bayesian strategy profile σ_{-i} of the other players, the Bayesian strategy σ_i is a *best reply* of player i if for any other strategy σ_i' of hers, $U_i((\sigma_i, \sigma_{-i}), t_i) \geq U_i((\sigma', \sigma_{-i}), t_i)$ for each of her types $t_i \in T_i$. A *Bayes–Nash equilibrium* or a *Bayesian equilibrium* is a profile of Bayesian strategies $\sigma^* = (\sigma_i^*)_{i \in I}$ such

that σ_i^* is a best reply against σ_{-i}^* for every player $i \in I$.
A simple, discrete variant of an example by Gale (1996) may clarify these abstract definitions. There are two investors i=1,•2 and three possible states of nature $k \in K = \{-1, 0, 1\}$. Each investor i only knows her own type

$$t_i \in T_i = \{-10, -6, -2, 2, 6, 10\}.$$

Every type t_i of investor i believes that all of the other investor's types $t_j \in T_j$, $j \neq i$, are equally likely, so that each of them has probability $\frac{1}{6}$. Moreover, every type t_i believes that the state of nature is k=1 when $t_i+t_j>0$; that the state of nature is k=0 when $t_i+t_j=0$; and that the state of nature is k=–1 when $t_i+t_j<0$. Formally, the belief $\lambda_i(t_i)$ of type $t_i \in T_i$ is defined by

$$\lambda_i(t_i)(k, t_j) = \begin{cases} \frac{1}{6} & k \text{ has the same sign as } t_i + t_j \\ 0 & \text{otherwise}. \end{cases}$$

The investors cannot communicate their types to one another. They can invest in at most one of two available investment periods. Each investor has three relevant strategies: invest *immediately*, in the first period; *wait* to the second period and invest only if the other investor has invested in the first period; or *never* invest. The payoff of each of the investors depends on the state of nature $k \in K=\{-1,0,1\}$ and on her own investment strategy, but not on the investment strategy of the other investor. The payoffs are as follows:

- Investing *immediately* when the state of nature is k yields investor i a payoff of k

$$u_i^k(\text{'immediately'}, \cdot) = k$$

  (The · stands for the investment decision of the other investor $j \neq i$, which, as we said, does not effect the payoff of investor i.)
- If investor i chooses to *wait* to the second period and invest only if the other investor has invested in the first period, investor i's payoff in the state of nature k is

$$u_i^k(\text{'wait'}, \cdot) = \frac{3}{4}k.$$

- If the investor *never* invests, her payoff is 0 irrespective of the state of nature:

$$u_i^k(\text{'never'}, \cdot) = 0.$$

How will the different types behave at a Bayesian equilibrium?
The type t_i= 10 assesses that by investing immediately her expected payoff is

$$U_i(\text{'immediately'}, 10) = \frac{1}{6} \times 0 + \frac{5}{6} \times 1 = \frac{5}{6}$$

(immediate investment yields 0 in case t_j=–10, and yields 1 in case t_j=–6,–2,•2,•6,•10). This is higher than $\frac{3}{4}$, the maximum payoff she could possibly get by waiting for the second period, and higher than the payoff 0 of never investing. So at a Bayesian equilibrium

$$\sigma_i^*(10) = \text{'immediately'}, \quad i = 1, 2.$$

Next, the expected payoff to the type t_i=6 from immediate investment is

$$U_i(\text{'immediately'}, 6) = \frac{1}{6} \times (-1) + \frac{1}{6} \times 0 + \frac{4}{6} \times 1 = \frac{1}{2}$$

(immediate investment yields 1 unless t_j=–10, in which case the payoff is –1, or t_j=–6, in which case the payoff is 0). So investing immediately is preferred for her over never investing. But how about waiting until the second period? That's an inferior option as well, since the types t_j=–10,–6,–2 will never invest in the first period (this would yield them a negative expected payoff). So only the positive types

t_j=2,6,10 could *conceivably* invest immediately, with overall probability reaching at most $\frac{3}{6}$. So waiting to see if they invest yields to the type t_i=6 an expected payoff not higher than $\frac{3}{6} \times \frac{3}{4} = \frac{3}{8}$, which is smaller than $\frac{1}{2}$. We conclude that the preferable strategy of t_i=6 at equilibrium is

$$\sigma_i^*(6) = \text{'immediately'}, \quad i = 1, 2.$$

What about t_i=2? Immediate investment yields her

$$U_i(\text{'immediately'}, 2) = \frac{2}{6} \times (-1) + \frac{1}{6} \times 0 + \frac{3}{6} \times 1 = \frac{1}{6}$$

(–1 is the payoff when t_j=–10,–6; 0 is the payoff when t_j=–2; the payoff is 1 otherwise). However, given that the types t_j=6,10 invest immediately at equilibrium, and that the negative types t_j=–10,–6,–2 do not invest immediately, the type t_i=2 figures out that by waiting and investing only if the other investor has invested first would yield her an expected payoff

$$U_i(\text{'wait'}, 2) \geq \frac{2}{6} \times \frac{3}{4} = \frac{1}{4} > \frac{1}{6}$$

($\frac{2}{6}$ is the probability assigned by t_i=2 to the event that $t_j \in \{6,10\}$ and hence j invests immediately, and $\frac{3}{4}$ is the payoff from the second period investment). The preferred strategy of t_i=2 at equilibrium is therefore

$$\sigma_i^*(2) = \text{'wait'}, \quad i = 1, 2.$$

We can now compute inductively, in a similar way, that also

$$\sigma_i^*(-2) = \text{'wait'}, \quad i = 1, 2 \quad \sigma_i^*(-6) = \text{'wait'}, \quad i = 1, 2$$

and that

$$\sigma_i^*(-10) = \text{'never'}, \quad i = 1, 2.$$

Notice that the equilibrium in the example is inefficient. For instance, when the pair of types is (t_1,t_2)= (2,•2) the investment is profitable, but both investors wait to see if the other one invests, and thus end up not investing at all. In this case, behaviour would become efficient if the investors could communicate their types to each other. Indeed, they would have been happy to do so, because their interests are aligned.

Obviously, there are other strategic situations with incomplete information in which the interests of the players are not completely aligned. For example, a potential seller of an object would like to strike a deal with a potential buyer at a price which is as high as possible, while the potential buyer would like the price to be as low as possible. That's why the traders might not volunteer to communicate honestly their private valuations of the object, even if they are technically able to do so. Still, in case the buyer values the object more than the seller, they would both prefer to trade at some price in-between their valuations rather than forgoing trade altogether. Therefore, the traders would nevertheless like to avoid a complete lack of communication. Myerson and Satterthwaite (1983) phrase general conditions under which no Bayesian equilibrium of any trade mechanism is ever fully efficient due to this tension between interests alignment and interests mismatch. Under these conditions, even if the traders are able to communicate their private information, at no Bayesian equilibrium does trade take place in all instances in which there exist gains from trade.

In the above variant of Gale's example we were able to find the unique Bayesian equilibrium using iterative dominance arguments. We have iteratively crossed out strategies that are inferior for some types, which enabled us to eliminate inferior strategies for other types, and so forth. As in games of complete information, this technique is not applicable in general, and there are games with incomplete information in which a Bayesian equilibrium is not the outcome of any process of iterative elimination of dominated strategies (Battigalli and Siniscalchi, 2003; Dekel, Fudenberg and Morris, 2007).

Games with incomplete information are discussed in many game theory textbooks (for example, Dutta, 1999; Gibbons, 1992; Myerson, 1991; Osborne, 2003; Rasmusen, 1989; Watson, 2002). Aumann and Heifetz (2002), Battigalli and Bonanno (1999) and Dekel and Gul (1997) are advanced surveys.

See Also

- epistemic game theory: an overview
- epistemic game theory: beliefs and types
- epistemic game theory: complete information
- game theory

Bibliography

Aumann, R.J. and Heifetz, A. 2002. Incomplete information. In *Handbook of Game Theory*, vol. 3, ed. R. J. Aumann and S. Hart. Amsterdam: North-Holland.

Battigalli, P. and Bonanno, G. 1999. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics* 53, 149–225.

Battigalli, P. and Siniscalchi, M. 2003. Rationalization with incomplete information. *Advances in Theoretical Economics* 3(1), article 3. Online. Available at http://www.bepress.com/bejte/advances/vol3/iss1/art3, accessed 25 April 2007.

Dekel, E., Fudenberg, D. and Morris, S. 2007. Interim correlated rationalizability. *Theoretical Economics* 2, 15–40.

Dekel, E. and Gul, F. 1997. Rationality and knowledge in game theory. In *Advances in Economics and Econometrics*, ed. D. Kreps and K. Wallis. Cambridge, UK: Cambridge University Press.

Dutta, P.K. 1999. *Strategies and Games: Theory and Practice*. Cambridge, MA: MIT Press.

Gale, D. 1996. What have we learned from social learning. *European Economic Review* 40, 617–28.

Gibbons, R. 1992. *Game Theory for Applied Economists*. Princeton: Princeton University Press.

Harsanyi, J.C. 1967–8. Games with incomplete information played by Bayesian players, parts I–III. *Management Science* 14, 159–82, 320–34, 486–502.

Myerson, R. 1991. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Myerson, R. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29, 265–81.

Osborne, M. 2003. *Introduction to Game Theory*. Oxford: Oxford University Press.

Rasmusen, E. 1989. *Games and Information: An Introduction to Game Theory*. Oxford: Basil Blackwell.

Watson, J. 2002. *Strategy: An Introduction to Game Theory*. New York: W.W. Norton.

How to cite this article

Heifetz, Aviad. "epistemic game theory: incomplete information." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000191>
doi:10.1057/9780230226203.0492

The New Palgrave Dictionary of Economics Online

equality of opportunity

J.E. Roemer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Whereas the ethic of equality of outcome does not hold individuals responsible for actions that may create inequality of outcomes, equality of opportunity ‘levels the playing field’ so that all have potential to achieve equal outcomes; inequalities of outcome that then transpire are not compensable at the bar of justice. The influences on the outcome a person experiences comprise *circumstances* (for which he should not be held responsible) and *effort* (for which he should be). Equal-opportunity policy compensates persons for their disadvantaged circumstances, ensuring that, finally, only effort counts in achieving outcomes.

Keywords

affirmative action; and parameters of disadvantage;; compensation; distribution; and effort vs circumstances; Dworkin, R.; on equality; on insurance market; educational finance; efficiency; and equity; equality of opportunity; vs equality of outcome; equality of outcome; vs equality of opportunity; vs equality of resources; and compensation; and efficiency; income-tax regime; insurance market; Dworkin on; meritocracy; property rights; and welfarism; welfarism; Sen on; and property rights; Sen, A.; and welfarism

Article

Equality of opportunity is to be contrasted with equality of outcome. While advocacy of equality of outcome has been traditionally associated with left-wing political philosophy, equality of opportunity has been championed by conservative political philosophy. Equality of outcome does not hold individuals responsible for imprudent actions that may, absent redress, reduce the values of the outcomes they enjoy, or for wise actions that would raise the value of the outcomes above the levels of others’. Equality of opportunity, in contrast, ‘levels the playing field’ so that all have the potential to achieve the same outcomes; whether in the event they do depends upon individual choice.

This traditional political alignment was upset by Ronald Dworkin (1981a; 1981b) who posed the question: if one is egalitarian, then what should one seek to equalize, welfare or resources? He argued, first, that equalizing welfare (outcome) was undesirable, even if interpersonal comparisons of welfare could be made, because doing so would fail to hold individuals accountable for their preferences. The issue of ‘expensive tastes’ was important for Dworkin; he argued that, if a person were glad he possessed an expensive taste, or identified with it, as opposed to viewing it as an addiction – a taste he would prefer not to have – then society owed him no extra resources to satisfy it. Dworkin argued that egalitarians should advocate the equalization of resources, as opposed to outcomes, but his conception of what comprised resources was broad. Resources consisted in not only transferable goods and wealth but internal talents as well. The question became: what allocation of *transferable* resources would count as equalizing the *entire bundle* of resources across persons, that is, would count as appropriately compensating individuals for their endowment of non-transferable resources? Dworkin's answer was to construct a kind of market for contingent claims behind a thin veil of ignorance in which traders knew their preferences (importantly, over risk) but not what resources they would come to have in the (actual) world. The desirable tax scheme, in the world, would mimic the allocation of transferable resources that would be implemented at the equilibrium in this market for contingent claims, after the birth lottery occurred (see Roemer, 1996, ch. 7, for a formal model).

Dworkin's contribution, importantly, attempted to integrate the issue of responsibility into egalitarian theory – which amounted to taking the most important tool of the political right and harnessing it for use by the political left. In Dworkin's theory, individuals are held responsible for their preferences, and this is implemented through the insurance market behind the veil of ignorance, where traders representing persons use their persons’ preferences to enter into insurance contracts. But persons are not held responsible for their resources, including internal talents, and the families into which they are born, and this is implemented through allowing the traders behind the veil to insure against bad luck in the birth lottery in so far as the distribution of these resources is concerned.

Several years later, G. A. Cohen (1989) and Richard Arneson (1989) criticized Dworkin's theory. Cohen argued that ‘Dworkin's cut’ between preferences and resources, was, for the purpose of ethics, the wrong way to separate characteristics. Suppose a person developed champagne tastes because she grew up in an aristocratic family in which she was never exposed to beer. Was it correct to later deny her the resources to buy champagne to achieve the level of welfare that beer drinkers could achieve more cheaply? Or suppose that a person, who grew up in a disadvantaged home that lacked resources, developed no ambition to develop his talents; indeed, he was satisfied with his unambitious tastes. Should he likewise be held responsible, even though his tastes were the consequence, at least in part, of an indigent childhood?

Arneson argued that Dworkin was right to argue against taking the ‘equalisandum’ as welfare, but said that replacing it with ‘resources’ was wrong – rather, it should be replaced with *opportunity for welfare*. What did it mean, then, to equalize opportunities for welfare? In what sense did this differ from ‘equalizing resources’ à la Dworkin? Arneson struggled to formulate an alternative, but did not succeed in proposing one that was clearly feasible.

Following Arneson and Cohen, Roemer (1993; 1998) proposed a model that would attempt to capture the insights of this philosophical discussion and permit one to compute, for a given situation, the policy that constituted the ‘equal opportunity’ policy. He separated the influences on the outcome a person experiences into *circumstances* and *effort*: circumstances are attributes of the person's environment for

which he should not be held responsible, and effort is the choice variable for which he should be held responsible. An equal-opportunity (EOp) policy is an intervention (such as the provision of resources by a state agency) that makes it the case that all those who expend the same degree of effort end up with the same outcome, regardless of their circumstances. Thus, EOp ‘levels the playing field’ in the sense of compensating persons for their deficits in circumstances, ensuring that, finally, only effort counts with regard to outcome achievement.

A more precise formulation follows. Suppose there is an *objective* for whose acquisition a planner wishes to equalize opportunities; this might be a wage-earning capacity, a life expectancy, or an income level. Denote the achievement of the objective as a function $u(\alpha, x; \beta)$ where α is the (scalar) level of effort expended by the person, *x* is the policy of the planner, and β is the vector of circumstances of the person. *u* is monotone increasing in the argument α – thus, effort enhances the acquisition of the objective. Nevertheless, effort may be subjectively costly for the individual: thus, *u* is not to be thought of as the usual economist's utility function, in which effort is costly. For example, *u* might be the wage-earning capacity a person comes to have, where α is the number of years of schooling and β measures family background, natural talent, and so on. The policy *x* can be chosen from some feasible set of policies *X*: it might, for example, be the distribution of a resource possessed by the planner. The set of individuals with a given value of β is called a ‘type’.

Suppose that, for each ordered pair (x, β) there ensues a distribution of effort in type β denoted by its cumulative distribution function $F(\alpha; x, \beta)$. The distribution of effort, classically, would result from the maximization of a preference order by the individuals of the type, one in which effort is differentially costly for those individuals. Typically, these distributions *F* differ across types (β). By hypothesis, individuals are not to be held responsible for their type. We now ask how one should interpret the stricture to choose a policy that equalizes the values of the objective *at constant effort levels* across types. The problem is that the *distribution* of effort in a type is a characteristic of the type and, if individuals are to be compensated for their types, they should likewise be compensated for the characteristics of those distributions. For example, if a disadvantaged type has a distribution of effort with a low mean, that itself should be taken into account in the compensation scheme. Roemer's solution was to propose that the *degree* of a person's effort should be measured by her *rank* in the effort distribution of her type. Thus, define the rank π by

$$\pi = F(\alpha; x, \beta)$$

and define the ‘indirect’ objective function

$$v(\pi; x, \beta) = u(F^{-1}(\pi; x, \beta), x; \beta).$$

Then *x* is an equal-opportunity policy just in case it equalizes the value of objective across types at every

degree of effort, that is:

$$\forall \pi \in [0, 1]\ \forall \beta, \beta' v(\pi; x, \beta) = v(\pi; x, \beta'). \tag{1}$$

Here, the process by which the effort distributions *F* emerge is black-boxed; of course, in actual applications, the black box would be unpacked with the specification of utility functions that individuals maximize to derive their efforts.

In general there will be no policy which equalizes opportunities in the sense of (1). For example, let *u* be wage-earning capacity, α be years of schooling, β be the educational level of the individual's parents, and *x* be investment in the education of the individual by the state. Suppose policies can be targeted to types, and there is a per capita social endowment of $\bar{x}$ for education. Suppose we partition the population into a finite set of types, $\{\beta_i | i = 1, \ldots, n\}$ where the population frequency of type *i* is p_i. A feasible policy is a vector $(x_1, \ldots, x_n)$ such that $\sum p_i x_i = \bar{x}$. We have as data, as well, the distribution functions $F(\cdot; x, \beta)$. For this general specification, there will generally not exist a feasible policy satisfying (1). Some alternative is therefore required. One may proceed as follows. We desire to equalize the values of *v* across different β's, at each π. As a second-best, we desire to maximize the minimum value of *v* across different β's, at each π. Thus, define

$$\Phi(\pi; x) = \min_{\beta} v(\pi; x, \beta).$$

We define a policy *x* to be *efficient* if

$$\text{there is no } x' \in X \quad s.t. (\forall \pi)(\Phi(\pi; x') \geq \Phi(\pi; x)), \tag{2}$$

where the inequality sign in (2) is understood to mean that, for some value(s) of π, there is strict inequality. We are interested only in efficient policies. There may, however, be many, even a continuum, of these; and the theory, thus far, gives us no way of choosing among them.

To see this, let us consider a special case in which effort responses within types are insensitive to the policy: thus, we may write those distributions as $F(\alpha; \beta)$. Suppose that there are just two types, $\beta = 1$ and $\beta = 2$, indicating the level of education of parents; each type comprises one-half the population. The Department of Education has one unit per capita of an educational resource to be invested in children. A

policy is an ordered pair (y,•2–y), indicating the per capita investment in children of the two types. Suppose that $u(\alpha, y, \beta) = \alpha^a y^c \beta^b$ is the value of the objective (perhaps, the child's future wage) where y is the amount of educational resource invested in the child. We will denote a policy by the value of its first component, y. Then

$$\Phi(\pi, y) = \min_y [(F^{-1}(\pi, 1))^a y^c, (F^{-1}(\pi, 2))^a 2^b (2 - y)^c]. \tag{3}$$

We may compute that the two arguments of the min function in (3) are equalized exactly when

$$y = \frac{2}{1 + \left(\frac{1}{2}\right)^{b/c} \left(\frac{F^{-1}(\pi, 1)}{F^{-1}(\pi, 2)}\right)^{a/c}}. \tag{4}$$

Now define

$$m = \min_\pi \left(\frac{F^{-1}(\pi, 1)}{F^{-1}(\pi, 2)}\right), \; M = \max_\pi \left(\frac{F^{-1}(\pi, 1)}{F^{-1}(\pi, 2)}\right). \tag{5}$$

Then any policy

$$y \in \left[\frac{2}{1 + \left(\frac{1}{2}\right)^{b/c} M^{a/c}}, \frac{2}{1 + \left(\frac{1}{2}\right)^{b/c} m^{a/c}}\right] \tag{6}$$

is efficient, and this interval comprises exactly the efficient policies. Thus, there is a continuum of

efficient policies.

There has been no general agreement concerning how to narrow the set of efficient policies to a single choice – in other words, how to rank efficient policies from the equal-opportunity viewpoint. Roemer (1998) proposed to choose a single policy by solving the problem:

$$\max_{x} \int_0^1 \Phi(\pi;\, x)\, d\pi; \tag{7}$$

Van de Gaer (1993) proposed to solve

$$\max_{x} \min_{\beta} \int V(\pi;\, x,\, \beta)\, d\pi. \tag{8}$$

Each of these proposals is somewhat arbitrary. Fleurbaey and Maniquet (2004) summarize the axiomatic approach to the problem, to which they and others have made substantial contributions. I believe that appeal to the equal-opportunity principle as such cannot resolve the issue; we must bring additional ethical considerations to bear.

How does the theory of equal opportunity fit into social choice theory? There are a number of ways one may answer this question; I believe the most salient point is that the equal-opportunity approach is distinguished from classical social-choice theory in being *non-welfarist*. Welfarism is the view that only the set of vectors of outcome (welfare) possibilities matters for the social decision. To be precise, if we represent individual preferences over social alternatives by utility functions, then the choice of a social alternative should depend only upon the information that is recoverable from the utility possibilities sets of the possible societies. In this sense, welfarism is a consequentialist view. Sen (1979) criticized the welfarist postulate for ignoring the issue of civil rights (the right not to be beaten by another, for instance); Roemer (1996) criticized it, with regard to the theory of distributive justice, for ruling out any theories which mention property rights. The equal-opportunity approach says that one cannot judge the goodness of a social outcome by knowing only the distribution of outcomes; one must also know *how hard people tried* in order to evaluate that goodness – in other words, one must know the correlation of effort with achievement to pass judgement on the fairness of a distribution scheme. Put this way, it is clear that the equal-opportunity approach formalizes a view that is held quite generally by citizens in many countries, to judge from opinion surveys. In judging how just schemes of distribution are, the proverbial man on the street usually wants to know if reward is ‘proportionate’ to effort expended. Knowing only the distribution of outcomes does not suffice.

Several empirical studies have applied these ideas. In Roemer et al. (2003), the authors asked: in a set of

11 OECD countries, what income-tax regime would equalize opportunities for income acquisition among workers? All workers in a country were assumed to have a quasi-linear utility function over income and labour, with a constant labour-supply elasticity with respect to the marginal tax rate (or the wage). The sole circumstance was taken to be the level of education of the mother of the worker. Young male workers were partitioned into three types, according to whether their mothers had low, medium, or high levels of education. The set of policies, X, was taken to be the set of feasible affine income tax regimes, that is, ones postulating constant marginal tax rates and a lump-sum payment to all. The objective was the post-fisc income (not utility) of the individual. Using the EOp objective of (7) turns out to be equivalent to choosing that income-tax regime which maximizes the average post-fisc income of the least advantaged type, those whose mothers did not complete secondary school. Table 1 summarizes the observed marginal tax rates in the countries of the sample and the equal-opportunity tax rates, so computed, under the assumption that the (male) labour-supply elasticity with respect to taxation is –.06.

EOp marginal income tax rates for 11 countries

Country	Observed marginal income-tax rate	EOp marginal income tax rate
Belgium	.53	.54
Denmark	.44	0
France	.31	.58
Great Britain	.36	.71
Italy	.23	.82
Netherlands	.53	.47
Norway	.39	0
Spain	.38	.61
Sweden	.52	0
United States	.24	.65
West Germany	.36	0

Source: Roemer et al. (2003).

Countries can be partitioned into three groups: those for which observed tax rates are much greater than the EOp tax rate (West Germany, Denmark, Sweden and Norway), those for which the observed and EOp tax rates are approximately the same (Belgium and the Netherlands), and those for which observed tax rates are much lower than the EOp tax rates (Italy, Spain, France, the United States and Great Britain). The pattern is not particularly surprising given common perceptions of the depth of income-transfer programmes in these countries.

A comment upon the countries in the first category is in order. To say that the EOp tax rate is zero in the northern European countries means that, with the postulated labour-supply effects of taxation, the average post-fisc income of the least advantaged type would be maximized with a lump-sum tax to finance public goods, and no other transfer payments. This comes about precisely because the *pre-fisc* distributions of income across the three types of worker are already very close in these countries. In the other countries in the sample, these pre-fisc distributions are sufficiently far apart that positive marginal

tax rates will, despite their incentive effects, increase the average post-fisc income of the least advantaged type.

Should one conclude from Table 1 that, from the equal-opportunity viewpoint, marginal income taxation should be abandoned in northern Europe? Hardly, for the division of workers into only three types is quite coarse. There are many other circumstances besides the education of the mother for which society might wish to compensate citizens. Indeed, the article under discussion studies as well a typology for four of the countries (where data exist) into six types, where workers are typed not only by three maternal education levels but also by two levels of native ability, as measured by performance in IQ tests. It turns out that a positive marginal EOp tax rate is then recommended for Sweden, although Denmark retains its zero tax rate! (With a sufficiently low labour-supply elasticity, this result, too, would be changed.)

Income taxation may not be the instrument of choice to equalize opportunities for income; one naturally thinks of using educational finance policy as a method for compensating children from disadvantaged families. Betts and Roemer (2003) partitioned American male workers who were attending secondary school in the late 1960s into four types, defined by four levels of maternal education. They took *wage-earning capacity* as the objective and state educational investment in the child as the policy instrument, and asked: What distribution of educational finance would have equalized opportunities for wage-earning capacity among these four types of worker? Wage elasticities with respect to educational investment were computed for the four types using data from the US Panel Studies on Income Dynamics (PSID). Based on the assumption of a per capita educational budget of $2,500, the recommended allocation is presented in Table 2.

EOp allocation of investment with per capita budget of $2,500 per student per annum

Parental ed'n	<8 years	8<ed<12 yrs	12 yrs	>12 yrs
EOp investment	$5,360	$3,620	$1,880	$1,100

Source: Betts and Roemer (2003).

In other words, equal-opportunity investment would allocate almost five times as much to the most disadvantaged type of student as to the most advantaged type. Interestingly, we computed that the average wage of workers under this allocation would have risen by 2.6 per cent over the observed average wage. In other words, there is no observed trade-off between equity and ‘efficiency’.

The authors computed that if the allocation of Table 2 had been implemented there would have been very little change in the fraction of black workers who would have risen above the bottom quintile of the wage distribution. They proceeded to compute the EOp policy for a different typology of workers into four types, defined as:

LB: low maternal education, black
HB: high maternal education, black
LW: low maternal education, white
HW: high maternal education, white

The results are presented in Table 3.

EOp allocation of educational investment, four types, race x maternal education

Type of worker	LB	HB	LW	HW
EOp investment	$8,840	$16,260	$2,610	$679

Source: Betts and Roemer (2003).

For this typology, the investment ratios are huge. Moreover, the total wage bill would fall by 2 per cent under the allocation of Table 3, showing that an equity-‘efficiency’ trade-off does exist with respect to this typology.

At the least, the calculations of Betts and Roemer demonstrate that there is a large difference between an *equal-resource policy*, which invests the same amount in all children, and an *equal-opportunity policy*, which invests in children in such a way as to attempt compensation for differential social circumstances. The United States, with its system of locally financed public education, is in most places less equitable even than the equal-resource policy would be: that is, usually more is invested in the public education of advantaged children than in that of disadvantaged children.

I have earlier distinguished between the equal-opportunity approach and the more classical welfarist approach in welfare economics. A second important distinction is between equal opportunity, as a concept of equity, and meritocracy. Consider the problem of admissions to university or professional school. The equal-opportunity approach would suggest admitting the highest-effort candidates from each of a set of types, distinguished by their levels of advantage in background. The meritocratic approach would suggest admitting those who are most likely to be high achievers. EOp focuses upon fair treatment *among the pool of candidates*, while meritocracy has a double focus: treating the candidates fairly but also considering the quality of services those candidates will, in the future, provide to society at large. (On the other hand, meritocracy is *not* concerned with candidates’ effort in its measurement of fair treatment, but only with their ability to perform.) Thus the two approaches are in conflict.

Clearly, the quality of services provided to society at large must count – the unadorned EOp approach cannot in general be the right one. Generally speaking, society should follow a mixture of equal-opportunity and meritocratic policies. To calibrate the right mixture would require, as well as data to calculate the relevant elasticities, a general theory of justice for society at large, in which account is taken not only of fairness to those competing to occupy social positions but of the welfare of those who eventually consume the products those individuals will produce. In the US debate around affirmative action, one can hear different emphases. With respect to school admissions, most citizens seem concerned with fairness to the candidates, although there is a dispute as to what traits should or should not count in judging fairness; but, with respect to employment, many believe meritocratic principles are primary. Thus, race-based affirmative action policies in universities are under challenge for focusing on the wrong parameters of disadvantage (which, many argue, should be ones of social class, not race), while affirmative-action employment policies are challenged for paying insufficient attention to competence in employing workers.

In the applications discussed above, the policymakers – whether fictitious ones in the minds of scholars or actual ones in social institutions – have generally contemplated only the effects of policies in a single sector, whether it be in education or employment. Calsamiglia (2005) has posed the following problem. Suppose individuals are competing for positions in several sectors simultaneously (in her example, for admission to a university and to an athletic team), and the admissions officer in each sector is attempting

to design an equal-opportunity policy for the candidates in his sector alone. Thus, the university admissions officer knows the abilities and circumstances and efforts of candidates for university, and the athletic coach knows the same information as it applies to performance in her sector. Each designs a *local* equal-opportunity admissions policy for his own sector. When will the combination of policies equalize opportunities *globally*? The tension here is that policies in each sector will, if not properly designed, distort the efforts of candidates in other sectors. Calsamiglia demonstrates that, under suitable conditions, locally designed EOp policies aggregate into a global EOp policy if and only if they *equalize rewards to effort* across types in each sector. For example, assigning disadvantaged students who are applying to law school ‘extra points’ to compensate them does *not* equalize rewards to effort as between them and more advantaged students: rather, one requires a policy which, for each unit of effort expended, increases the probability of admission by the *same amount* across all types of student. One can say, that is, that equalizing rewards to effort is the necessary and sufficient condition for decentralizing the social problem of equalizing opportunities across the board into policy formation at the sectoral level. Whether or not Calsamiglia's insight will be important in policy design depends upon the degree to which individuals are involved in inter-sectoral effort allocation decisions.

See Also

- justice
- redistribution of income and wealth

Bibliography

Arneson, R. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56, 77–93.

Betts, J. and Roemer, J. 2003. Equalizing opportunity through educational finance reform. In *Schools and the Equal Opportunity Problem*, ed. P. Peterson and L. Woessmann. Cambridge, MA: MIT Press.

Calsamiglia, C. 2005. Decentralizing equality of opportunity and issues concerning the equality of educational opportunity. Ph.D. dissertation. New Haven, CT: Department of Economics, Yale University.

Cohen, G. 1989. On the currency of egalitarian justice. *Ethics* 99, 906–44.

Dworkin, R. 1981a. What is equality? Part one: equality of welfare. *Philosophy & Public Affairs* 10, 185–246.

Dworkin, R. 1981b. What is equality? Part two: equality of resources. *Philosophy & Public Affairs* 10, 283–345.

Fleurbaey, M. and Maniquet, F. 2004. Compensation and Responsibility. In *Handbook of Social Choice*

and Welfare, vol. 2, ed. K. Arrow, A. Sen and K. Suzumura. Amsterdam: North-Holland.

Roemer, J. 1993. A pragmatic theory of responsibility for the egalitarian planner. *Philosophy & Public Affairs* 22, 146–66.

Roemer, J. 1996. *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.

Roemer, J. 1998. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Roemer, J., Aaberge, R., Colombino, U., Fritzell, J., Jenkins, S., Marx, I., Page, M., Pommer, E., Ruiz-Castillo, J., San Segundo, M.J., Tranaes, T., Wagner, G. and Zubiri, I. 2003. To what extent do fiscal regimes equalize opportunities for income acquisition among citizens? *Journal of Public Economics* 87, 539–65.

Sen, A. 1979. Utilitarianism and welfarism. *Journal of Philosophy* 76, 463–89.

Van de Gaer, D. 1993. Equality of opportunity and investment in human capital. Ph.D. dissertation. Leuven: Catholic University of Leuven.

How to cite this article

Roemer, J. E. "equality of opportunity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000214> doi:10.1057/9780230226203.0494

The New Palgrave Dictionary of Economics Online

equality

James S. Coleman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Equality’ is used to mean equality before the law, equality of opportunity, and equality of result, among other things. These types of equality are not necessarily mutually compatible. Equal distribution of benefits is often taken to be ‘natural’ (by Rawls, for example), partly because envy is ubiquitous. In welfare economics the presumed diminishing marginal utility of money implies that equality of incomes maximizes welfare, but if interpersonal utility comparisons are impossible no such presumption can be made. As well, the interdependencies between individuals in terms of welfare are such that enforced equalization is likely to reduce overall welfare.

Keywords

Berlin, I.; Coleman, J. S.; Edgeworth, F. Y.; efficiency vs. equality; envy; equality; equality before the law; equality of opportunity; equality of result; inequality; input–output analysis; interdependencies of welfare; interpersonal utility comparisons; liberty; marginal utility of money; Nozick, R.; optimal taxation; Pigou, A. C.; progressive taxation; property rights; Rawls, J.; Robbins, L. C.; welfare economics

Article

The very use of the term ‘equality’ is often clouded by imprecise and inconsistent meanings. For example, ‘equality’ is used to mean equality before the law (equality of treatment by authorities), equality of opportunity (equality of chances in the economic system), and equality of result (equal distribution of goods), among other things. These different meanings often conflict, and are almost never wholly consistent. See Hayek (1960, p. 85; 1976, pp. 62–4) for a discussion of equality before the law and equality of result, and Rawls (1971) for a discussion of equality of opportunity within a theory of distributive justice. Elsewhere I have discussed the difference between equality of opportunity and equality of result in education (Coleman, 1975). See also Pole (1978) for a detailed examination of the

changing conceptions of equality in American history.

Some order can be brought into the confusion among the different uses of the term ‘equality’ by first conceiving of a system that constitutes an abstraction from reality. The system consists of:

1. (a) a set of positions which have two properties:
    1. (i) when occupied by persons, they generate activities which produce valued goods and services;
    2. (ii) the persons in them are rewarded for these activities, both materially and symbolically;
2. (b) a set of adult persons who are occupants of positions;
3. (c) children of these adults;
4. (d) a set of normative or legal constraints on certain actions.

What is ordinarily meant by equality under the law has to do with (b), (c), and (d): that the normative or legal constraints on actions depend only on the nature of the action, and not on the identity of the actor. That is, the law treats persons in similar positions similarly, and does not discriminate among them according to characteristics irrelevant to the action.

What is ordinarily meant by equality of opportunity has to do with (a), (b), and (c): that the processes through which persons come to occupy positions give an equal chance to all. More particularly, this ordinarily means that a child's opportunities to occupy one of the positions (a) do not depend on which particular adults from set (b) are that child's parents. What is ordinarily meant by equality of result has to do with (a.ii): that the rewards given to the position occupied by each person are the same, independent of the activity.

These three conceptions, equality under the law, equality of opportunity, and equality of result can also be seen as involving different relations of the State to the inequalities that exist or spontaneously arise in ongoing social activities. Equality before the law implies that the laws of the State do not recognize distinctions among persons that are irrelevant to the activities of the positions they occupy, but otherwise make no attempt to eliminate inequalities that arise. Equality of opportunity implies that the State intervenes to insure that inequalities in one generation do not cross generations, that children have opportunities unaffected by inequalities among their parents. Equality of result implies a continuous or periodic intervention and redistribution by the State to insure that the inequalities which arise through day-to-day activities are not accumulated, but are continuously or periodically eliminated.

The relations between the first two kinds of equality differ according to how close a society is to a legally minimalist society or a legally maximalist society. In a society that is legally minimalist, equality before the law is compatible with a high degree of inequality of opportunity – depending on the distribution of opportunity provided by other institutions in society, such as the family. In a legally maximalist society, in which many functions of traditional institutions have been taken over by institutions that are creatures of the State (e.g., functions of the family taken over by the public school), equality before the law implies a high degree of equality of opportunity. Only in a society in which the law was far more intrusive than found anywhere, and children were taken from their families to be raised ‘with equal opportunity’ by the State, could it be said that equality before the law would coincide with equality of opportunity.

The relation between equality of opportunity and equality of result is somewhat different, for it implies two different kinds of interventions of the State. Equality of opportunity implies intervention to provide

each person with resources that give equal *chances* to obtain the material and symbolic rewards that arise from productive activity, while equality of result implies intervention in the distribution of these rewards, to provide each person with equal *amounts*. The two concepts become indistinguishable only when the State intervenes to insure that each position (in (a) above) provides the same set of material and symbolic rewards; and in such a circumstance, ‘opportunity’ loses meaning altogether.

Is equality ‘natural’?

There are certain philosophical positions that take equality of result as a ‘natural’ point, from which all others are deviations. Isaiah Berlin probably states this as well as any other

> No reason need be given for … an equal distribution of benefits for that is ‘natural’, self evidently right and just, and needs no justification, since it is in some sense conceived as being self justified … The assumption is that equality needs no reasons, only inequality does so; that uniformity, regularity, similarity, symmetry,… need not be specially accounted for, whereas differences, unsystematic behavior, changes in conduct, need explanation and, as a rule, justification. If I have a cake and there are ten persons among whom I wish to divide it, then if I give exactly one tenth to each, this will not, at any rate automatically, call for justification; whereas if I depart from this principle of equal division I am expected to produce a special reason. It is some sense of this, however latent, that makes equality an idea which has never seemed intrinsically eccentric … (1961, p. 131).

This quotation describes a view with which Berlin does not necessarily identify himself. In the same paper, he states that ‘equality is one value among many … it is neither more nor less rational than any other ultimate principle … rational or non-rational’ It is, however, the position implicitly taken by John Rawls in his *Theory of Justice*, for the book is addressed to the question, ‘When can inequalities (of result) be regarded as just?’ Rawls's answer can be paraphrased as ‘Only those inequalities are just which make the least well off person better off than that person would be (other things being equal) in the absence of the inequalities.’

Whether equality of result is ‘natural’ or not, and whether the position of Berlin and Rawls is correct or incorrect, would appear to depend on how the distribution of goods occurs: If goods are initially the property of a single central source (e.g., ‘the State’), then Berlin's position and that of Rawls appear correct. If all rights and resources originate with the State (or with the king, as in early political theory), than an equal distribution has some claim to be seen as natural. (If, for example, the revenue from oil discovered on public lands is a major component of GNP, as in some Middle Eastern states, equal distribution constitutes a natural point.) But if goods are seen to arise from the activities of a set of independent actors each with certain initial property rights, and each with a certain amount of zeal and skill, ‘equality’ (meaning equality of result) is hardly natural, and is inconsistent with the distribution of property rights including rights to the fruits of one's own activity.

Equality, envy and resentment

The idea of equality as ‘natural’ appears also to derive in part from the ubiquity of envy and resentment in society, with the demand for ‘equality’ as an expression of these feelings which carries legitimacy. A number of sociologists have pointed to this connection. For example, Simmel writes (1922, translated in Schoeck, 1969, pp. 236–7):

> Characteristically, no one is satisfied with his position in relation to his fellow beings, but everyone wishes to achieve a position that is in some way an improvement. When the needy majority experiences the desire for a higher standard of living, the most immediate expression of this will be a demand for equality in wealth and status with the upper ten thousand.

Simmel follows with an anecdote: at the time of the 1848 revolution, a woman coal-carrier remarked to a richly dressed lady, ‘Yes, madam, everything's going to be equal now; I shall go in silks and you'll carry coal.’
Helmut Schoeck, in an extensive examination of the role of envy in society, argues that

> social philosophers have largely failed to see how little the individual is concerned with being *equal* to someone else. For very often his sense of justice is outraged by the very fact that he is denied the measure of inequality which he considers to be right and proper (1969, p. 234).

Feelings of envy and resentment constitute a challenge to the existing distribution of *rights* in society, between those held collectively and those held individually. In particular, it is a challenge to the existence of individual property rights. The centrality of property rights for conceptions of equality is seen most clearly in neoclassical economic theory, which assumes a distribution of property rights among a set of independent actors, accompanied by a free market. (See Meade, 1964, for a discussion of property rights and the market in relation to equality.) It is to economic theory that I now turn.

The role of ‘equality’ in economic theory

The concept of ‘equality’ has no place in positive economic theory. In this it is unlike the concept of ‘liberty’, for economic theory is predicated on the assumption of liberty, that is, free choice (subject only to resource constraints) among alternative actions. There is, in the concept of free choice, however, something closer to the idea of equality before the law than to equality of opportunity, and closer to the latter than to equality of result. Equality of result implies a distribution process that is the antithesis of the market.
But normative economics, that is, welfare economics, makes up for the absence of ‘equality’ from positive economic theory, for the idea of equality of result is a part of the very atmosphere surrounding welfare economics. The question of what policies will maximize social welfare is not often answered directly in terms of equality in the distribution of valued goods, but the idea seems always to hover nearby. The most direct expression of the central importance of equality in welfare economics was probably that of Pigou (1938); see also (Bergson (1966, ch. 9) who reasoned that because money, like

everything else, had declining marginal utility, and thus a dollar was worth much less to a person when he had a million others than when it was the only one he had, then the maximum of social welfare could only be achieved when incomes were made equal. (Neither Pigou nor any other welfare economist followed this implication with actual policy recommendations for equality of income, thus raising the question: if the criterion is correct, then why not recommend implementing it?)

The rock on which Pigou's argument is often regarded as foundering is that of interpersonal comparison of utility. To move from the relative importance for one person of a dollar when he is rich and when he is poor to its relative importance to different persons is a move which, as has been often reiterated, cannot be justified on positive grounds. Perhaps the most widely quoted statement to this effect is that of Lionel Robbins (1938):

> But, as time went on, things occurred which began to shake my belief in the existence between so complete a continuity between politics and economic analysis … I am not clear how these doubts first suggested themselves; but I well remember how they were brought to a head by my reading somewhere – I think in the work of Sir Henry Maine – the story of how an Indian official had attempted to explain to a high-caste Brahmin the sanctions of the Benthamite system. ‘But that,’ said the Brahmin, ‘cannot possibly be right – I am ten times as capable of happiness as that untouchable over there.’ I had no sympathy with the Brahmin. But I could not escape the conviction that, if I chose to regard men as equally capable of satisfaction and he to regard them as differing according to a hierarchial schedule, the difference between us was not one which could be resolved by the same methods of demonstration as were available in other fields of social judgement … ‘I see no means,’ Jevons had said, ‘whereby such comparison can be accomplished.’

Edgeworth expressed the same point, ‘The Benthamite argument that equality of means tends to maximum happiness, presupposes a certain equality of natures; but if the capacity for happiness of different classes is different, the argument leads not to equal, but to unequal distribution’ (1897, p. 114). Such arguments are ordinarily taken as conclusive within the domain of economics, and with their acceptance, the very programme of welfare economics – not to speak of the foundations for a policy designed to bring equality – is emasculated.

A philosopher might argue, of course, that there is no logical difference between the comparison of utilities of two persons and the comparison of utilities of one person at two different times. Neither, by this argument, is warranted. See, for example, Parfit (1984).

However, Pigou's conclusion has, quite apart from problems of interpersonal comparison of utility, another deficiency. It assumes that each person is an island, and contributes nothing to the welfare of others, nor has his welfare contributed to by others. Yet is is the essence of social and economic systems that there is interdependence, that one person's activities do affect the welfare of others, whether intended or not. One person spends money on loud radios that cause disturbance, while another plants flowers that others enjoy. Or one uses income for training which is productive, benefiting general welfare, while another uses income on drink and becomes alcoholic, requiring public-expense hospitalization.

But if this is so, then maximization of welfare one time period into the future would require that these

interdependencies be taken into account. Maximization would occur only if resources were distributed among persons in accordance with the positive impact of their activities on those events which bring welfare to others. But in general persons do not capture the full benefits of their welfare-generating activities, nor do persons pay the full costs of their welfare–diminishing activities.

The matter can also be seen as a problem in input–output economics: What current allocation of resources among productive activities (i.e., among positions in the system as described earlier) will achieve some desired distribution of final consumption? If the aim is to maximize the sum of final consumption (‘maximizing welfare’?), it is quite unlikely that either the current allocation necessary to achieve that, or the distribution of final consumption itself, will approach equality. Even if the desired final distribution is equality, and even if that is achievable within the system of activities, it is highly unlikely that the allocation at time *0* necessary to achieve that at time *t* will be equal. And it may well be that the only distribution at time *0* that would achieve equality at time *t* would do so at a low level of welfare, with each having less than if there were inequality at time *t* resulting from a different distribution at time *0*. If Pareto optimality is taken as a self-evident necessary condition for optimal policies, then because of the processes described above, a criterion of equal distribution (either initially or subsequently) would violate the condition. This suggests that Rawls's question was misdirected, and should have been ‘when (assuming non-violation of constitutional rights) is *equality* of distribution justified?’ and should have been answered, ‘Only when there is no unequal distribution that would subsequently make each better off.’

Thus even if Pigou's point that maximizing welfare requires equalizing marginal utilities is accepted, and noncomparability of utilities is ignored, the policy implication of equalizing incomes appears shortsighted in the extreme. Another way of seeing so is by use of Robert Nozick's Wilt Chamberlain example, an example designed to argue against theories of distributive justice which, like that of Rawls, use the resulting distribution of goods (‘end state theories’, to use Nozick's term) as a criterion.

> Now suppose that Wilt Chamberlain is greatly in demand by basketball teams, being a great gate attraction. (Also suppose contracts run only for a year, with players being free agents.) He signs the following sort of contract with a team: In each home game, twenty five cents from the price of each ticket of admission goes to him. (We ignore the question of whether he is ‘gouging’ the owners, letting them look out for themselves.) The season starts, and people cheerfully attend his team's games; they buy their tickets, each dropping a separate twenty five cents for their admission price into a special box with Chamberlain's name on it. They are excited about seeing him play; it is worth the total admission price to them. Let us suppose that in one season one million persons attend his home games, and Wilt Chamberlain winds up with $250,000, larger even than anyone else has. Is he entitled to this income? (Nozick, 1974, p. 161)

Thus as Nozick points out, an equal distribution at one point will lead to an unequal distribution at a later point, due to the very system of activities through which persons satisfy their interests.

There are only three ways to prevent this, all of which, carried to their limit, can be shown to reduce welfare. One is to prevent the economic exchange through which persons spend their quarters as they see fit, for such exchanges may lead to a large accumulation in the hands of the Wilt Chamberlains. A second is to attack the system of activities itself, the system which generates that matrix of

coefficients that transform equality into inequality – that is, shutting down professional basketball, which redistributes income from those with low incomes to those with high incomes. The third way is to allow the exchange, but then to tax the high incomes back down to equality. This effectively eliminates the activity, because if income is an incentive to carry out the activity that is paid for, the Wilt Chamberlains lose all incentive to carry the activity.

Indeed, unless there is a perfect positive correspondence of those activities which are intrinsically pleasurable with those which produce benefits for others, and a perfect negative correspondence with those that produce harm for others, the absence of any extrinsic incentives will lower the welfare for all. The more interrelated the activities of individuals, the greater the reduction in social welfare when extrinsic incentives are absent.

It is true that taxation which is not carried to the limit, but is merely ‘progressive’, does not eliminate the incentive for activities that bring high income, for these activities continue in societies that have progressive taxation. But this taxation may lead to underprovision of welfare-generating activities. That is, efficiency may be sacrificed to achieve some distributional goals. The potential conflicts between efficiency and equality are discussed in the literature on optimal taxation (e.g., Atkinson and Stiglitz, 1980, part II). (A device which is informally used in social systems to reduce the disincentive effect of regimes of taxation and redistribution that shift incomes in the direction of equality is the attachment of social stigma to the receiving of income thus redistributed, for example, stigma associated with being ‘on welfare’. The existence of this stigma constitutes a means of informally reconstituting the differential incentives that are reduced by redistribution.)

All three approaches to preventing inequalities from arising out of equality give, at their extreme, the same result: elimination of the very system of activities that generates welfare in the first place; for it is these activities which not only generate welfare, but also transform equality at one time into inequality at a later time.

Thus it becomes clear that the source of inequalities is embedded in the very matrix of social and economic activities through which individuals increase the welfare of themselves and one another. If, through technology for example, this matrix changes in such a way that individuals’ satisfaction of wants is more concentrated in a few hands (e.g., by the invention and development of television), then inequalities will necessarily increase.

More generally, the degree of inequality seems related to the degree of interdependence in this matrix of social and economic activities. In a social system that has very low interdependence (e.g., a social system composed largely of subsistence farmers, a condition that was once the case for nearly all societies), the welfare of each in future periods depends largely on his own initial distribution of resources (including zeal and skill). If that distribution is near equality, then near equality is perpetuated into the future, modified only by random events. More important, even if the initial distribution is unequal, the low interdependence of the system of activities means that these inequalities (also modified by random events) are merely carried forward into the future. In a system with a high degree of interdependence, however, there are a great many configurations which constitute ‘inequality-generating’ activity structures. In such activity structures, initial distribution of equality will lead to highly unequal distributions. This inequality in turn will lead in the next generation to inequality of opportunity, constrained only by random processes or explicit policies towards non-inheritance of position, i.e., toward equality of opportunity. (In a system in which attention to basketball was directed not to televised professional teams, but to games of the local high school, both the material and

nonmaterial rewards among basketball player would be more equally distributed. There would be greater equality of results, which would arise not through a change in the set of persons (b), the distribution of children (c), or the normative and legal constraints (d), but only through a change in the distribution of positions.)

Does this mean that there tends to be a negative relation between the interdependence of activities in a social system (and thus the total social product) and the equality with which the activities of the system distribute the product? If so, this is a discouraging result for those who would prefer a social system in which incomes are not increasingly unequal, for it specifies an opposition between two goals both regarded as desirable.

This question has two parts, a within-generation part and a between-generation part. Within generations, it appears likely that there is a negative relation, that increased interdependence does, except in unlikely activity structure, increase inequality. It is possible that this negative relation is responsible for the rise in redistributive actions of governments as interdependence of economic activities increases.

Between generations, the answer would appear to hinge largely upon the relative rates of increase of interdependence of activities and of equality of opportunity (i.e., non-inheritance of position). The latter can occur through regression to the mean as well as through explicit policy intervention (see Becker and Tomes, 1986, for a discussion). If equality of opportunity increases more slowly than interdependence of activities, then (except for unlikely configurations of the activity matrix) there will be a decrease in equality of result among lineages of persons. If equality of opportunity increases more rapidly than the increase in interdependence of activities, there will be an increase in equality of result among lineages, even with a decrease in equality of result within generations.

Altogether, there has been little investigation of the matters discussed above, that is, just how the structure of social and economic activities itself affects inequalities. Such investigations would lead toward taking work on equality partly out of the realm of normative theory, bringing it partly into the realm of positive theory.

See Also

- poverty

Bibliography

Atkinson, A.B. and Stiglitz, J.E. 1980. *Lectures on Public Economics*. New York: McGraw-Hill.

Becker, G. and Tomes, N. 1986. Inequality, human capital, and the rise and fall of families. In *Approaches to Social Theory*, ed. S. Lindenberg, J. Coleman, and S. Nowak. New York: Russell Sage.

Bergson, A. 1966. *Essays in Normative Economics*. Cambridge, MA: Harvard University Press.

Berlin, I. 1961. Equality. In *Justice and Social Policy*, ed. F.A. Olafson. Englewood Cliffs: Prentice Hall.

Coleman, J. 1975. What is meant by ‘an equal educational opportunity’? *Oxford Review of Education* 1 (1), 27–9.

Edgeworth, F.Y. 1897. *Papers Relating to Political Economy*. London: Macmillan for the Royal Economic Society, 1925.

Hayek, F.A. 1960. *The Constitution of Liberty*. Chicago: University of Chicago Press.

Hayek, F.A. 1976. *Law, Legislation and Liberty*. vol. 2. Cambridge: Cambridge University Press.

Meade, J.E. 1964. *Efficiency, Equality and the Ownership of Property*. London: Allen & Unwin.

Nozick, R. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pigou, A.C. 1938. *Economics of Welfare*. 4th edn , New York: Macmillan.

Pole, J.R. 1978. *The Pursuit of Equality in American History*. Cambridge: Cambridge University Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Robbins, L. 1938. Interpersonal comparisons of utility. *Economic Journal* 48, 635–41.

Schoeck, H. 1969. *Envy: A Theory of Social Behavior*. New York: Harcourt, Brace and World.

Simmel, G. 1962. *Soziologie*. 2nd edn, Munich and Leipzig: Duncker & Humblot.

How to cite this article

Coleman, James S. "equality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000101> doi:10.1057/9780230226203.0493

The New Palgrave Dictionary of Economics Online

equation of exchange

Michael D. Bordo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

One of the oldest formal relationships in economics, the equation of exchange, as a basic accounting identity of a money economy, demonstrates that the sum of expenditures must equal the sum of receipts. It is useful both as a classification scheme for analysing the underlying forces at work in a money economy and as a building block or engine of analysis for monetary theory and in particular for the quantity theory of money. It can also be regarded as a building block for a macro theory of aggregate demand and supply, and used to construct a theory of nominal income.

Keywords

Briscoe, J.; Cairnes, J.E.; cash balance (Cambridge) approach; circular flow of income; De Foville, A.; equation of exchange; equation of societary circulation; Fisher, I.; Friedman, M.; Hadley, A. T.; Kemmerer, E. W.; Keynes, J. M.; Lang, J.; Law, J.; Levasseur, E.; Lloyd, H.; Lubbock, J.; Mill, J. S.; naive quantity theory; Newcomb, S.; nominal income; Norton, J. P.; Pantaleoni, M.; Pigou, A. C.; quantity equation; quantity theory of money; Rau, K. H.; Ricardo, D.; Senior, N. W.; stocks and flows; Thornton, H.; transactions velocity of circulation; Turner, S.; velocity of circulation; Walras, L.

Article

The equation of exchange (often referred to as the quantity equation) is one of the oldest formal relationships in economics, early versions of both verbal and algebraic forms appearing at least in the 17th century. Perhaps the best known variant of the equation of exchange is that expressed by Irving Fisher (1922):

$$MV = PT. \tag{1}$$

Equation (1) represents a simple accounting identity for a money economy. It relates the circular flow of money in a given economy over a specified period of time to the circular flow of goods. The left-hand side of eq. (1) stands for money exchanged, the right-hand side represents the goods, services and securities exchanged for money during a specified period of time. *M* is defined as the total quantity of money in the economy, *T* as the total physical volume of transactions, where a transaction is defined as any exchange of goods, including physical capital, services and securities for money, *P* is an appropriate price index representing a weighted average of the prices of all transactions in the economy. Finally, to make the stock of money comparable with the flow of the value of transactions (*PT*), and to make the two sides of the equation balance, it is multiplied by *V*, the transactions velocity of circulation, defined as the average number of times a unit of currency turns over (or changes hands) in the course of effecting a given year's transactions.

An alternative variant of the equation of exchange is the income version by Pigou (1927). Empirical difficulties in measuring an index of transactions, and the special price index related to it, led, with the development of national income accounting, to the formulation of eq. (2):

$$MV = PY \tag{2}$$

where *Y* represents national income expressed in constant dollars, *P* the implicit price deflator and *V* the income velocity of circulation defined as the average number of times a unit of currency turns over in the course of financing the year's final activity.

Equations (1) and (2) differ from each other because the volume of transactions in the economy includes intermediate goods and the exchange of existing assets, in addition to final goods and services. Thus vertical integration and other factors which affect the ratio of transactions to income would also alter the ratio of transactions velocity to income velocity.

A third version of the equation of exchange, the Cambridge cash balance approach (Pigou, 1917; Marshall, 1923; Keynes, 1923), converts the flow of spending into units comparable to the stock of money

$$M = kPY \tag{3}$$

where $k=1/V$ is defined as the time duration of the flows of goods and services money could purchase, for example, the average number of weeks income held in the form of money balances.

Equations (2) and (3) are arithmetically equivalent to each other but they rest on fundamentally different

notions of the role of money in the economy. Both eqs. (2) and (1) view money primarily as a medium of exchange and the quantity of money is represented as continually ‘in motion’ – constantly changing hands from buyer to seller in the course of a time period. Equation (3) views money as a temporary abode of purchasing power (an asset) forming part of a cash balance ‘at rest’. Consequently, the items included in the definition of money in the transactions and income versions of the equation of exchange are assets used primarily to effect exchange – currency and checkable deposits, whereas the cash balance approach includes, in addition to these items, non-checkable deposits and possibly other liquid assets.

The equation of exchange is useful both as a classification scheme for analysing the underlying forces at work in a money economy and as a building block or engine of analysis for monetary theory and in particular for the quantity theory of money.

As a classification scheme, the equation as a basic accounting identity of a money economy demonstrates the two-sided nature of the circular flow of income – that the sum of expenditures must equal the sum of receipts. The left-hand side of the equation shows the market value of goods and services purchased (dollar value of goods exchanged) and the money received. The equation also relates the stock of money to the circular flow of income by multiplying M by its velocity. Finally, the equation is useful in creating definitional categories – M, V, P, T – amenable both to empirical measurement and to theoretical analysis.

The equation of exchange is best known as a building block for the quantity theory of money. The traditional approach has been to make behavioural assumptions about each of the variables in the equation, converting it from an identity to a theory. The simplest application, dubbed the ‘naive quantity theory’ (Locke, 1691) treated V and T in eq. (1) as constants, with P varying in direct proportion to M. A more sophisticated version (Fisher, 1911) treats each of M, V and T as being normally determined by independent sets of forces, with V as determined by slowly changing factors such as those affecting the payments process and the community's money holding habits.

The Cambridge cash balance approach, based on eq. (3), views the quantity theory as encompassing both a theory of money demand and money supply. In this approach the nominal money supply is determined by the monetary standard and the banking system while the nominal quantity of money demanded is proportional to nominal income, with k the factor of proportionality, representing the community's desired holding of real cash balances. k in turn is determined by economic variables such as the rate of interest in addition to the factors stressed by the Fisher approach. The price level (value of money) is then determined by the equality of money supply and demand.

The equation of exchange can also be regarded as a building block for a macro theory of aggregate demand and supply (Schumpeter, 1954). If we view MV as aggregate demand and T or Y as aggregate supply, then P would be determined in the familiar Marshallian way.

Finally, the equation can be used to construct a theory of nominal income. According to this approach (Friedman and Schwartz, 1982), nominal income is determined by the interaction of the money supply and a stable demand for real cash balances. The decomposition of a given change in nominal income into a change in the price level and in real output is determined in the short run by inflation (deflation) forecast errors and in the long run by the natural rate of output.

The equation of exchange both as a classification scheme and as a building block for the quantity theory of money can be traced back to the earliest development of economic science.

The pre-classical writers of the 17th and 18th centuries viewed the equation in both senses. Locke

(1691), Hume (1752) and Cantillon (1735) each organized his approach to monetary issues using the equation. Locke had a clear statement of the naive quantity theory assuming both *V* and *T* to be immutable constants. Hume followed Locke but made a clear distinction between long-run statics and short-run dynamics. In the long run the price level would be proportional to *M* but, in the short run or transition period, changes in *M* would produce changes in *T*. Cantillon had a clear understanding of the relationship between the stock of money and the circular flow of income. Indeed, he was the first to define explicitly the concept of velocity of circulation, viewing *V* not as a constant but as a variable influenced in a stable way by both technological and economic variables. Furthermore, like Hume, Cantillon distinguished between the long-run equilibrium nature of the quantity theory and short-run disequilibrium. Both Locke and Hume viewed the equation from the perspective of money 'at rest' forming a cash balance whereas Cantillon viewed money as continuously in 'motion'.

John Law (1705) understood the equation of exchange but used it to derive a link between changes in the quantity of *M* and changes in *T*.

The classical economists Thornton, Ricardo, Mill, Senior and Cairnes followed the Locke/Hume/Cantillon tradition of the quantity theory of money using a verbal version of the equation of exchange in their monetary analysis.

Algebraic versions of the equation first appeared in the 17th and 18th centuries (see Marget, 1942; Humphrey, 1984). The British writers Briscoe (1694) and Lloyd (1771) both expressed a rudimentary version of eq. (1), unfortunately omitting a term for velocity. Turner (1819) formulated the equation without breaking *PT* into separate components. The most complete early statement of the equation was by Sir John Lubbock (1840), who not only included all the items of the equation but (preceding Fisher) distinguished between the quantities and velocities of hard currency, bank notes and bills of exchange. Similar complete algebraic statements of the equation were made by the German writers Lang (1811) and Rau (1841); the Italian Pantaleoni (1889); the Frenchmen Levasseur (1858), Walras (1874) and de Foville (1907); and the Americans Newcomb (1885), Hadley (1896), Norton (1902) and Kemmerer (1907). Of this group Newcomb presented the clearest statement. Newcomb started with the concept of exchange as involving the transfer of money for wealth. Summing up all exchanges in the economy he arrived at his equation of societary circulation:

$$VR = KP \tag{4}$$

where *V* represents the total value of currency, *R* the rapidity (velocity) of circulation, *K* the volume of real transactions, *P* a price index.

The clearest and best known algebraic expressions of the equation were by the neoclassical economists Irving Fisher (1922) and A.C. Pigou (1917). Fisher (1911, pp. 15–17), directly following Newcomb, defined the equation of exchange as

> a statement, in mathematical form, of the total transaction: effected in a certain period in a

> given community. … [I]n the grand total of all exchanges for a year, the total money paid is equal to the total value of goods bought. The equation thus has a money side and a goods side. The money side is the total money paid, and may be considered as the product of the quantity of money multiplied by its rapidity of circulation. The goods side is made up of the products of quantities of goods exchanged multiplied by their respective prices.

This statement expressed as in eq. (1) or in an expanded version distinguishing between currency and deposits payable by check,

$$MV + M'V' = PT \tag{5}$$

where M' is defined as checkable deposits and V' their velocity, Fisher then used to analyse the forces determining the price level.
Fisher's approach followed the ‘motion’ theory tradition of Cantillon with velocity determined primarily by technological and institutional factors. In contrast, Pigou (1917) and other writers in the Cambridge tradition, Marshall (1923) and Keynes (1923), followed the ‘rest’ approach of Locke and Hume, expressing the equation as

$$1/P = kR/M \tag{6}$$

where R represents total resources enjoyed by the community, k the proportion of resources the community chooses to keep in the form of titles to legal tender, M the number of units of legal tender and P a price index. For Pigou the fundamental difference between his approach and that of Fisher was that by focusing

> attention on the proportion of their resources that people *choose* to keep in the form of titles to legal tender instead of focusing on the ‘velocity of circulation’ … it brings us … into relation with *volition* – an ultimate *cause of demand* – instead of with something that seems at first sight *accidental and arbitrary*. (1917, p. 174, emphasis added)

The Cambridge cash balance version of the equation of exchange, by focusing on the demand for money and volition rather than emphasizing mechanical aspects of the circular flow of money, can be viewed as the starting point for the Keynesian approach to the demand for money (Keynes, 1936), for modern choice theoretic approaches to money demand (Hicks, 1935) and for the modern quantity theory of

money (Friedman, 1956).

See Also

- Newcomb, Simon
- quantity theory of money

Bibliography

Bordo, M.D. 1983. Some aspects of the monetary economics of Richard Cantillon. *Journal of Monetary Economics* 12, 234–58.

Briscoe, J. 1694. *Discourse on the Late Funds*•…•. London.

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. ed. H. Higgs. London: Macmillan, 1931; reprinted New York: Augustus M. Kelley, 1964.

Fisher, I. 1911. *The Purchasing Power of Money*. New York: Macmillan.

Fisher, I. 1922. *The Purchasing Power of Money*, 2nd edn. Reprinted New York: Augustus M. Kelley, 1963.

Foville, A. de. 1907. *La monnaie*. Paris.

Friedman, M. 1956. The quantity theory of money – a restatement. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M. and Schwartz, A.J. 1982. *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices and Interest Rates, 1867–1975*. Chicago: University of Chicago Press for the NBER.

Hadley, A.T. 1896. *Economics*. New York.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* 2, 1–19.

Holtrop, M.W. 1929. Theories of the velocity of circulation of money in earlier economic literature. *Economic Journal* 39(January), 503–24.

Hume, D. 1752. Of money. In *Essays, Moral, Political and Literary*, vol. 1 of *Essays and Treatises*, a new edition, Edinburgh: Bell and Bradfute; Cadell and Davies, 1804.

Humphrey, T.M. 1984. Algebraic quantity equations before Fisher and Pigou. *Federal Reserve Bank of Richmond Economic Review* 70(5), 13–22.

Kemmerer, E.W. 1907. *Money and Credit Instruments in Their Relation to General Prices*. New York: H. Holt & Co.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. Reprinted, London: Macmillan for the Royal Economic Society, 1971.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. Reprinted, London: Macmillan for the Royal Economic Society, 1973.

Lang, J. 1811. *Grundlinien der politischen Arithmetik*. Kharkov.

Law, J. 1705. *Money and Trade Considered with a Proposal for Supplying the Nation with Money*. New York: Augustus Kelley, 1966.

Levasseur, E. 1858. *La question de l'or: les mines de Californie et d'Australie*. Paris.

Lloyd, H. 1771. *An Essay on the Theory of Money*. London.

Locke, J. 1691. *The Works of John Locke*, vol. 5. London, 1823.

Lubbock, J. 1840. *On Currency*. London.

Marget, A.W. 1942. *The Theory of Prices*. New York: Prentice-Hall.

Marshall, A. 1923. *Money, Credit and Commerce*. London: Macmillan. Reprinted, New York: Augustus M. Kelley, 1965.

Newcomb, S. 1885. *Principles of Political Economy*. New York: Harper & Brothers.

Norton, J.P. 1902. *Statistical Studies in the New York Money Market*. New York: Macmillan.

Pantaleoni, M. 1889. *Pure Economics*. Trans. T.B. Bruce, London: Macmillan, 1898.

Pigou, A.C. 1917. The value of money. *In Quarterly Journal of Economics* 32(November), 38–65. Reprinted in *Readings in Monetary Theory*, ed. F.A. Lutz and L.W. Mints for the American Economic Association, Homewood, IL: Irwin, 1951.

Pigou, A.C. 1927. *Industrial Fluctuations*, 2nd edn. London: Macmillan, 1929.

Rau, K.H. 1841. *Grundsatze der Volkswirtsaftslehre*, 4th edn. Leipzig and Heidelberg.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Turner, S. 1819. *A Letter Addressed to the Right Hon. Robert Peel with Reference to the Expediency of the Resumption of Cash Payments at the Period Fixed by Law*. London.

Walras, L. 1874–7. *Eléments d'économie politique pure*. Lausanne: Corbaz.

How to cite this article

Bordo, Michael D. "equation of exchange." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000102> doi:10.1057/9780230226203.0495

The New Palgrave Dictionary of Economics Online

equilibrium (development of the concept)

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

ceteris paribus; equilibrium; expectations; game theory; general equilibrium; intertemporal equilibrium; long-run equilibrium; natural and normal conditions; partial equilibrium; short-run equilibrium; stationary state; Steuart, J.; temporary equilibrium

Article

From what appears to have been the first use of the term in economics by James Steuart in 1769, down to the present day, equilibrium analysis (together with its derivative, disequilibrium analysis) has been the foundation upon which economic theory has been able to build up its not inconsiderable claims to ‘scientific’ status. Yet despite the persistent use of the concept by economists for over 200 years, its meaning and role have undergone some quite profound modifications over that period.

At the most elementary level, ‘equilibrium’ is spoken about in a number of ways. It may be regarded as a ‘balance of forces’, as when, for example, it is used to describe the familiar idea of a balance between the forces of demand and supply. Or it can be taken to signify a point from which there is no endogenous ‘tendency to change’: stationary or steady states exhibit this kind of property. However, it may also be thought of as that outcome which any given economic process might be said to be ‘tending towards’, as in the idea that competitive processes tend to produce determinate outcomes. It is in this last guise that the concept seems first to have been applied in economic theory. Equilibrium is, as Adam Smith might have put it (though he did not use the term), the centre of gravitation of the economic system – it is that configuration of values towards which all economic magnitudes are continually tending to conform.

There are two properties embodied in this original concept which when taken into account begin to impart to it a rather more precise meaning and a well-defined methodological status. Into this category enters the formal definition of ‘equilibrium conditions’ and the argument for taking these to be a useful object of analysis.

There are few better or more appropriate places to isolate the first two properties of ‘equilibrium’ in this original sense than in the seventh chapter of the first book of Adam Smith's *Wealth of Nations*. The argument there consists of two steps. The first is to define ‘natural conditions’:

> There is in every society … an ordinary or average rate of both wages and profits… . When the price of any commodity is neither more nor less than what is sufficient to pay … the wages of the labour and the profits of the stock employed … according to their natural rates, the commodity is then sold for what may be called its natural price. (Smith, 1776, I.vii, p. 62)

The key point here is that ‘natural conditions’ are associated with a general rate of profit – that is, uniformity in the returns to capital invested in different lines of production under existing best-practice technique. In the language of the day, this property was thought to be the characteristic of the outcome of the operation of the process of ‘free competition’.
The second step in the argument captures the analytical status to be assigned to ‘natural conditions’:

> The natural price … is, as it were, the central price, to which the prices of all commodities are continually gravitating. Different accidents may sometimes keep them suspended a good deal above it, and sometimes force them down even somewhat below it. But whatever may be the obstacles which hinder them from settling in this center of repose and continuance, they are constantly tending towards it. (I.vii, p. 65)

This particular ‘tendency towards equilibrium’ was held to be operative in the *actual* economic system at any given time. It is not to be confused with the familiar question concerning the stability of competitive equilibrium in modern analysis. There the question about convergence to equilibrium is posed in some *hypothetical* state of the world where none but the most purely competitive environment is held to prevail. It is also essential to observe that in defining ‘natural conditions’ in this fashion, nothing has yet been said (nor need it be said) about the forces which act to determine the natural rates of wages and profits, or the natural prices of commodities. It will therefore be possible to refrain from discussing the *theories* offered by various economists for the determination of these variables in most of what follows. Treatment of these matters may be found elsewhere in this dictionary. Similarly, there will be no discussion here of existence or uniqueness of equilibrium (see existence of general equilibrium).
‘Natural conditions’ so defined and conceived are the formal expression of the idea that certain systematic or persistent forces, regular in their operation, are at work in the economic system. Smith's earlier idea, that ‘the co-existent parts of the universe … contribute to compose one immense and connected system’ (1759, VII. ii, 1.37), is translated in this later formulation into an analytical device capable of generating conclusions with a claim to general (as opposed to a particular, or special) validity. These general conclusions were customarily referred to as ‘statements of tendency’, or ‘laws’, or ‘principles’ in the economic literature of the 18th and 19th centuries. It is worth emphasizing that there was no implication that these general tendencies were either swift in their operation or that they were not subject at any time to interference from other obstacles. Like sea level, ‘natural conditions’ had an unambiguous meaning, even if subject to innumerable cross-currents.
To put it another way, the distinction between ‘general’ and ‘special’ cases (like its counterpart, the distinction between ‘equilibrium’ and ‘disequilibrium’), refers neither to the immediate practical relevance of these kinds of cases to actual existing market conditions, nor to the prevalence, frequency,

or probability of their occurrence. In fact, as far as simple observation is concerned, it might well be that 'special' cases would be the order of the day. John Stuart Mill expressed this idea especially clearly when he held that the conclusions of economic theory are only applicable 'in the *abstract*', that is, 'they are only true under certain suppositions, in which none but general causes – causes common to the *whole class* of cases under consideration – are taken into account' (Mill, 1844, pp. 144–5). Marshall, of course, understood their application as being subject not only to this qualification (which he spoke about in terms of 'time'), but also to the condition that 'other things are equal' (1890, I.iii, p. 36). There will be cause to return to this matter below.

To unearth these regularities, one had to inquire behind the scene, so to speak, to reveal what otherwise might remain hidden. Adam Smith had set out the basis of this procedure in an early essay on 'The Principles which Lead and Direct Philosophical Enquiries':

> Nature, after the largest experience that common observation can acquire, seems to abound with events which appear solitary and incoherent … by representing the invisible chains which bind together all these disjointed objects, [philosophy] endeavours to introduce order into this chaos of jarring and discordent appearances. (Smith, 1795, p. 45)

In short, 'equilibrium', if we may revert to the modern terminology for a moment, became the central organizing category around which economic theory was to be constructed. It is no accident that the formal introduction of the concept into economics is associated with those very writers whose names are closely connected with the foundation of 'economic science'. It could even be argued that its introduction marks the foundation of the discipline itself, since its appearance divides quite neatly the subsequent literature from the many analyses of individual problems which dominated prior to Smith and the Physiocrats.

Cementing this tradition, Ricardo spoke of fixing his 'whole attention on the permanent state of things' which follows from given changes, excluding for the purposes of general analysis 'accidental and temporary deviations' (1817, p. 88). Marshall, though substituting the terminology 'long-run normal conditions' for the older 'natural conditions', excluded from this category results upon which 'accidents of the moment exert a preponderating influence' (1890, p. vii). J.B. Clark followed suit and held that 'natural or normal' values are those to which 'in the long run, market values tend to conform' (1899, p. 16). Jevons (1871, p. 86), Walras (1874–7, p. 380), Böhm-Bawerk (1899, vol. 2, p. 380) and Wicksell (1901, vol. 1, p. 97) all followed the same procedure.

Not only was the status of 'equilibrium' as the centre of gravitation of the system (the benchmark case, so to speak) preserved, but it was defined in the manner of Smith. The primary theoretical object of all these writers was to explain that situation characterized by a uniform rate of profit on the supply price of capital invested in different lines of production. Walras, whose argument is quite typical, stated the nature of the connection forcefully:

> uniformity of … the price of net income [rate of profit] on the capital goods market … [is one] condition by which the universe of economic interests is governed. (1874–7, p. 305)

From an historical point of view, the novelty of these arguments which were worked out in the 18th

century by Smith and the Physiocrats is not that they recognized that there might be situations which could be described as ‘natural’, but that they associated these conditions with the outcome of a specific process common to market economies (free competition) and utilized them in the construction of a general economic analysis of market society. Earlier applications of ‘natural order’ arguments were little more than normative pronouncements about some existing or possible state of society. They certainly made no ‘scientific’ use of the idea of systematic tendencies, even if these might have been involved. This is particularly apparent in the case of the ‘natural law’ philosophers, but is also true of the early liberals like Locke and Hobbes. Even Hume, who to all intents and purposes had in his possession all of the building blocks of Smith's position, drew back from the one crucial step that would have led him to Smith's ‘method’ – he was just not prepared to admit that thinking in terms of regularities, however useful it might prove to be in dispelling theological and other obfuscations (and thus in advancing ‘human understanding’), was anything more than a convenient and satisfying way of thinking. The question as to whether the social and economic world was actually governed by such regularities, so central to Smith and the Physiocrats, just did not concern Hume.

Yet the earlier normative connotations of ideas like ‘natural conditions’, ‘natural order’, and the like, quite rapidly disappeared when the terminology was appropriated by economic theory. Nothing was ‘good’ simply by virtue of its being ‘natural’. This, of course, is not to say that once the theoretical analysis of the natural tendencies operating in market economies had been completed, and the outcomes of the competitive process had been isolated in abstract, an individual theorist might not at that stage wish to draw some conclusions about the ‘desirability’ of its results (a normative statement, so to speak). But such statements are not implied by the concept of equilibrium – they are value judgements about the characteristics of its outcomes.

Indeed, contrary to the view sometimes expressed, even Smith's use of deistic analogies and metaphors in the *Theory of Moral Sentiments*, where we read about God as the creator of the ‘great machine of the universe’, and where we encounter for the first time the famous ‘invisible hand’, is no more than the extraneous window-dressing which surrounds a well-defined *theoretical* argument based upon the operation of the so-called ‘sympathy’ mechanism. Thus, as W.E. Johnson noted when writing for the original edition of Palgrave's *Dictionary*, ‘the confusion between scientific law and ethical law no longer prevails’, and he observes that ‘the term normal has replaced the older word natural’ – to be understood by this terminology as ‘something which presents a certain empirical uniformity or regularity’ (Palgrave, 1899, p. 139).

While ‘natural conditions’ or ‘long-run normal conditions’ represent the original concept of ‘equilibrium’ utilized in economic theory, John Stuart Mill's *Political Economy* seems to have been the source from which the actual term equilibrium gained widespread currency (though, like so much else, it is also to be found in Cournot's *Recherches*). More significant, however, is the fact that in Mill's hands the meaning and status of the concept undergoes a modification. While maintaining the idea of equilibrium as a long-period position, Mill introduces the idea that the equilibrium theory is essentially ‘static’. The relevant remarks appear at the beginning of the fourth book:

> We have to consider the economical condition of mankind as liable to change … thereby adding a theory of motion to our theory of equilibrium – the Dynamics of political economy to the Statics. (Mill, 1848, IV.i, p. 421)

Since he retained the basic category of ‘natural and normal conditions’, Mill's claim had the effect of adding a property to the list of those associated with the concept of equilibrium. However, over the question of whether this additional property was necessary to the concept of equilibrium, there was to be less uniformity of opinion. Indeed, this matter gave rise to a debate in which at one time or another (until at least the 1930s) almost all theorists of any repute became contributors. The problem was a simple one – are natural or long-period normal conditions the same thing as the ‘famous fiction’ of the stationary or steady state. Much hinged upon the answer; a ‘yes’ would have limited the application of equilibrium to an imaginary stationary society in which no one conducts the daily business of life.

On this question, as might be expected, Marshall vacillated. The thrust of his argument (as well as those of his major contemporaries, with the important exception of Pareto) seems to imply that such a property was not essential to his purpose, but as was his habit on so many occasions, in a footnote he qualified that position (1890, p. 379, n.1). In the final analysis, the answer seems to have depended rather more on the explanation given for the determination of equilibrium values, than upon the concept of equilibrium proper. It was not until the 1930s that the issue seems to have been resolved to the general satisfaction of the profession. But then its ‘resolution’ required the introduction of a new definition of equilibrium (the concept of intertemporal equilibrium) due in the main to Hicks.

However, some further embellishments and modifications were worked upon the concept of equilibrium before the 1930s. Here, two developments stand out. The first concerns the distinction between partial equilibrium analysis and general equilibrium analysis. The second concerns a trend that seems to have developed consequent upon Marshall's treatment of the element of time, which led him to his threefold typology of periods (‘market’, ‘short’, and ‘long’ – we shall leave to one side the further category of ‘secular movement’). The upshot of this trend which is decisive, is that it became common to speak of the possibility of ‘equilibrium’ in each of these Marshallian periods.

The analytical basis for partial equilibrium analysis was laid down in 1838 by Cournot in his *Recherches*. Mathematical convenience, more than methodological principle, seems to have been responsible for his adopting it (see, for example, 1838, p. 127). Though this small volume failed to exercise any widespread influence on the discipline much before the 20th century, it was known and read by Marshall (who spoke of Cournot as his ‘gymnastics master’), from whose *Principles* the popularity of partial equilibrium analysis is largely derived (though it would be remiss to overlook Auspitz, Lieben and von Mangoldt). Unlike the case of Cournot, however, it would be difficult to argue that Marshall came across the method in anything other than a roundabout way (though some have argued that its principal attraction for him lay in its facility in allowing him to express his theory in a manner which required little recourse to mathematics).

When Marshall first introduced the idea of assuming ‘other things equal’ in the *Principles*, the *ceteris paribus* condition which is taken as the hallmark of the partial equilibrium approach, he seems to have done so not in order to justify the procedure of analysing ‘one bit at a time’, but in order to make a quite different point – that a long-run normal equilibrium would only *actually* emerge if none but the most general causes were allowed to operate without interference (see, for example, 1890, pp. 36, 366, and 369–70). In other words, the ‘other things’ that were being held ‘equal’ were the given data of the theory and the external environment – if the data remained the same and the external environment was freely competitive, then a long-run normal equilibrium would result. Indeed, Walrasian general equilibrium holds ‘other things equal’ in this sense. To put it another way, in Marshall's initial argument nothing was said about the possibility of assuming the interdependencies between long-run variables themselves to

be of secondary importance, as is customary in partial equilibrium analysis.

This latter requirement of Marshallian analysis, the idea of the negligibility of indirect effects when one looks at individual markets (1919, p. 677 ff.), seems to have sprung from his habit of presenting equilibrium *theory* in terms of *particular* market demand and supply curves (with their attendant notions of representative consumers and firms). It is here, in fact, that Marshall's presentation of demand and supply theory differs so markedly from its presentation by Walras. To the extent that this is so, it would seem to be better to recognize that the idea of ‘partial’ versus ‘general’ equilibrium has more to do with the presentation of a particular theory, and Marshall's propensity to consider markets one at a time, than it has to do with the abstract category of equilibrium with which this discussion is concerned. This view would accord, incidentally, with the fact that the great disputes over the relative merits of these two modes of analysis (for example, that between Walras on the one hand, and Auspitz and Lieben on the other) were fought over the specification of demand and cost functions.

Another modification to the concept of equilibrium that has become more significant in recent literature also makes an appearance in Marshall; though it is not carried as far as it has been in recent literature. The second, third and fifth chapters of the fifth book of Marshall's *Principles* set out the conditions for the determination of what he calls the ‘temporary equilibrium’, the ‘short-run equilibrium’ and the ‘long-run equilibrium’ of demand and supply. The last of these categories, as Marshall makes perfectly clear in the text, corresponds to Adam Smith's ‘natural conditions’ (1890, p. 347). The first two are to a greater or lesser degree ‘more influenced by passing events, and by causes whose action is fitful and short lived’ (p. 349). What is striking about Marshall's terminology is the fact that situations which from an analytical point of view would traditionally have been regarded as ‘deviations’ from long-period normal equilibrium (that is, disequilibria) are explicitly referred to as different cases of ‘equilibrium’.

This trend has taken on an entirely new significance in recent literature, and has had dramatic consequences for the meaning and status of the concept of equilibrium in economic theory. But just as important in comprehending this development is the introduction of the notion of intertemporal equilibrium into theoretical discourse.

The notion of intertemporal equilibrium (introduced by Hayek, Lindahl and Hicks in the inter-war years and developed in the 1950s by Malinvaud, Arrow and Debreu) warrants special consideration since ‘equilibrium conditions’ under this notion are defined quite differently from ‘natural’ or ‘long-run normal’ conditions. Intertemporal equilibrium defines as its object the determination of *nt* market-clearing prices (for *n* commodities over *t* elementary time periods commencing from an arbitrary short-period starting point). The chief implication of this definition of equilibrium conditions, and that which sets it apart from long-run normal conditions, is that not only will the price of the same commodity be different at different times but also that the stock of capital need not yield a uniform return on its supply price.

This fundamental change in the concept of equilibrium did not mean that intertemporal equilibrium positions were immediately divested of the status that had been given to ‘equilibrium’ ever since Adam Smith. In certain circles they continued to be regarded as positions towards which the economic system could actually be said to be ‘tending’ (or as benchmark cases).

However, once the *sequential* character of this equilibrium concept came to be better understood, it became apparent that there could be no ‘tendency’ towards it – at least not in the former meaning of that idea. One was either in it, in which case the sequence was ‘inessential’, or one was not, in which case the sequence was ‘essential’ (see Hahn, 1973, p. 16). And the probabilities overwhelming suggested the

latter. Attention was thus turned to the individual points in the sequence; the temporary equilibria, as Hicks had dubbed them (applying the terminology of Marshall in a new context). A whole new class of cases, disequilibrium cases from the point of view of full intertemporal equilibrium, began to be examined. The discipline has now accumulated so many varieties that it is impossible to document them all here. Instead, two broad features of this development may be noted here, the first concerning the role that expectations were thereby enabled to play, the second the common designation now uniformly applied to all such cases: ‘equilibrium’.

When equilibrium is interpreted as a solution concept in the sense that *all* solutions to *all* models (for which solutions exist) enjoy equal analytical status and differ only in that they become ‘significant’, as von Neumann and Morgenstern put it, when they are ‘similar to reality in those respects which are essential in the investigation at hand’ (1944, p. 32), it is sometimes said that economics has availed itself of a very powerful notion of equilibrium. On this line of argument, Walrasian equilibrium and, say, conjectural equilibrium compete with one another not for the title ‘general’ (since, in the traditional sense at least, there is no such category), but for the title ‘significant’. Furthermore, at any given time they are competing for this title with as many other models as are available to the profession.

It seems to be the case that the status of equilibrium in economic analysis has come full circle since its introduction in the late 18th century. From being derived from the idea that market societies were governed by certain systematic forces, more or less regular in their operation in different places and at different times, it now seems to be based on an opinion that nothing essential is ‘hidden’ behind the many and varied situations in which market economies might actually find themselves. In fact, it seems that these many cases are to be thought of as being more or less singular from the point of view of modern theory. From being the central organizing category around which the whole of economic theory was constructed, and therefore the ultimate basis upon which its practical application was premissed, equilibrium has become a category with no meaning independent of the exact specification of the initial conditions for *any* model. Instead of being thought of as furnishing a theory applicable, as Mill would have said, to the whole class of cases under consideration, it is increasingly being regarded by theorists as the solution concept relevant to a particular model, applicable to a limited number of cases. The present fashion for replacing economic theory proper by game theory, an approach which could be regarded by no less a theorist than Professor Arrow as contributing only ‘mathematical tools’ to economic analysis not many years ago (1968, p. 113), seems to exemplify the trend of modern economics.

See Also

- Arrow–Debreu model of general equilibrium
- competition, classical
- conjectural equilibria
- general equilibrium
- temporary equilibrium

Bibliography

Arrow, K.J. 1968. Economic equilibrium. In *International Encyclopedia of the Social Sciences*, as reprinted in *The Collected Papers of Kenneth J. Arrow*, vol. 2. Cambridge, MA: Harvard University Press.

Böhm-Bawerk, E. von. 1899. *Capital and Interest*. 3 vols. Reprinted, South Holland, IL: Libertarian Press, 1959.

Clark, J.B. 1899. *The Distribution of Wealth*. London: Macmillan.

Cournot, A.A. 1838. *Researches into the Mathematical Principles of the Theory of Wealth*. Trans. N.T. Bacon with an introduction by I. Fisher, 1897; 2nd edn, London and New York: Macmillan, 1927.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value. In *Modern Capital Theory*, ed. M. Brown et al. Amsterdam: North-Holland.

Hahn, F.H. 1973. *On the Notion of Equilibrium in Economics*. Cambridge: Cambridge University Press.

Hicks, J.R. 1939. *Value and Capital*. 2nd edn, Oxford: Clarendon Press, 1946.

Jevons, W.S. 1871. *Theory of Political Economy*. Edited from the 2nd edition (1879) by R.D.C. Black, Harmondsworth: Penguin, 1970.

Marshall, A. 1890. *Principles of Economics*. 9th (variorum) edition, taken from the text of the 8th edition, 1920. London: Macmillan.

Marshall, A. 1919. *Industry and Trade*. 2nd edn. London: Macmillan.

Mill, J.S. 1844. *Essays on Some Unsettled Questions of Political Economy*. 2nd edn 1874; reprinted, New York: Augustus M. Kelley.

Mill, J.S. 1848. *Principles of Political Economy*. 1871 (reprinted 1909), 6th edn, London: Longmans, Green & Company.

Palgrave, R.H.I., ed. 1899. *Dictionary of Political Economy*, vol. 3, London: Macmillan.

Pareto, V. 1909. *Manual of Political Economy*. Trans. from the French edition of 1927 and ed. A.S. Schwier and A.N. Page, New York: Augustus M. Kelley, 1971.

Ricardo, D. 1817. *The Principles of Political Economy and Taxation*. Edited from the 3rd edition of 1821 by P. Sraffa with the collaboration of M. Dobb, vol. 1 of *The Works and Correspondence of David Ricardo*, 11 vols, Cambridge: Cambridge University Press, 1951–73.

Smith, A. 1759. *The Theory of Moral Sentiments*. Ed. D.D. Raphael and A.L. Macfie from the 6th edn of 1790, Oxford: Oxford University Press, 1976.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 2 vols, ed. E. Cannan, London: Methuen, 1961.

Smith, A. 1795. *Essays on Philosophical Subjects*. Ed. W.P.D. Wrightman and J.C. Bryce, Oxford: Oxford University Press, 1980.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. 3rd edn, Princeton: Princeton University Press, 1953.

Walras, L. 1874–7. *Elements of Pure Economics*. Trans. and ed. W. Jaffé from the definitive edition of 1926, London: Allen & Unwin, 1954.

Wicksell, K. 1901. *Lectures on Political Economy*. 2 vols, ed. L. Robbins London: Routledge and Kegan Paul, 1934.

How to cite this article

Milgate, Murray. "equilibrium (development of the concept)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000104>
doi:10.1057/9780230226203.0497

The New Palgrave Dictionary of Economics Online

equilibrium-correction models

David F. Hendry

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The equilibrium-correction class of econometric models is surprisingly large, and includes regression equations, autoregressive-error models, autoregressive distributed-lags, simultaneous equations, autoregressive conditional heteroskedastic processes and generalized ARCH, vector autoregressions and dynamic stochastic general equilibrium systems, among others. Moreover, its properties are relatively generic for all members. Following an historical overview of its origins in error corrections and control mechanisms on the one hand and cointegration on the other, its properties are described, leading to an explanation as to why the ubiquitous class of equilibrium-correction models is prone to forecast failure in processes that are non-stationary from location shifts.

Keywords

adjustment costs; autoregressive distributed-lag models; autoregressive-error models; cointegration; common factors; control mechanisms; differencing; dynamic stochastic general equilibrium (DGSE) models; equilibrium-correction models; error-correction models; forecast failure; GARCH processes; linear-quadratic models; partial equilibrium; stationarity; unit roots; vector autoregressions

Article

1 Introduction

An equilibrium is a state from which there is no inherent tendency to change. Since we deal with stochastic processes, the equilibrium is the expected value of the variable in an appropriate representation, since that is the state to which the process would revert in the absence of further shocks. Then, we define an equilibrium-correction model (EqCM) as one (*a*) which has a well-defined equilibrium, and (*b*) in which adjustment takes place towards that equilibrium. A key aspect of an EqCM is that deviations from its expected value are attenuated, and eventually eliminated if no additional outside influences impinge. As such, equilibrium-correction models are a very broad class, comprising all regressions, autoregressions,

autoregressive-distributed lag (ADL) models, linear simultaneous equations, vector autoregressions (VARs), vector equilibrium-correction systems based on cointegration (VEqCMs), dynamic stochastic general equilibrium systems (DSGEs), autoregressive conditional heteroscedastic processes as in Engle (1982) (ARCH), and generalized ARCH (GARCH, see Bollerslev, 1986) processes among others. Their formulation (in levels or differences) determines the equilibrium to which they converge (level or steady state). For example, a random walk without drift is a non-stationary process in levels, but is stationary in differences (its non-integrated representation), and has an expectation of zero, so the differences equilibrium corrects to zero.

We first address the broad nature of the equilibrium-correction class in Section 2, then review the history of equilibrium-correction model formulation in Section 3, and consider its links to cointegration in Section 4. The roles of cointegration and equilibrium correction in economic forecasting are examined in Section 5, in particular the non-robustness of EqCMs to location shifts in the underlying equilibria, and consequently their proneness to forecast failure. Section 6 concludes.

2 The equilibrium-correction class

Often it is not realized that the model being used is a member of the equilibrium-correction class, so this section establishes that the models listed above are indeed in the EqCM class. The properties of the class are partly specific to the precise model, but primarily generic, as Section 5 emphasizes. We consider six cases.

2.1 Regression as an equilibrium-correction model

Consider a conditional linear equation of the form in (1) for $t = 1, \ldots, T$:

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i z_{i,t} + \varepsilon_t = \beta_0 + \beta' \mathbf{z}_t + \varepsilon_t \tag{1}$$

with $\varepsilon_t \sim \mathrm{IN}[0, \sigma_\varepsilon^2]$ (normally and independently distributed, mean zero, variance σ_ε^2) independently of the past and present of the k regressors $\{\mathbf{z}_t\}$. Then:

$$\mathrm{E}\left[\left(y_t - \beta_0 - \beta' \mathbf{z}_t\right) | \mathbf{z}_t\right] = 0 \tag{2}$$

defines the conditional equilibrium, where adjustment to that equilibrium is instantaneous as entailed by (1). Re-expressing (1) in differences ($\Delta x_t = x_t - x_{t-1}$ for any x) and lagged deviations from (2) delivers the (isomorphic) EqCM formulation:

$$\Delta y_t = \beta' \Delta \mathbf{z}_t - \left(y_{t-1} - \beta_0 - \beta' \mathbf{z}_{t-1}\right) + \varepsilon_t \tag{3}$$

where the feedback coefficient is –1. Then (3) is an EqCM where the equilibrium-correction term is $(y_{t-1} - \beta_0 - \beta' \mathbf{z}_{t-1})$. Notice that differencing is a linear transformation, not an operator, in any setting beyond a scalar time series.

The existence of (2) does not require that y_t and $\mathbf{z}_t$ are stationary, provided the linear combination is; and could hold, for example, for growth rates rather than the original levels if y_t and $\mathbf{z}_t$ were differences of those original variables.

2.2 Autoregressive-error models as equilibrium-corrections

Even extending a static regression like (1) by (say) a first-order autoregressive error as in:

$$y_t = \beta_0 + \beta' \mathbf{z}_t + u_t \text{where} u_t = \rho u_{t-1} + \varepsilon_t \tag{4}$$

leads to:

$$y_t = \beta_0 + \beta' \mathbf{z}_t + \rho\left(y_{t-1} - \beta_0 - \beta' \mathbf{z}_{t-1}\right) + \varepsilon_t$$

or:

$$\Delta y_t = \beta' \Delta \mathbf{z}_t + (\rho - 1)\left(y_{t-1} - \beta_0 - \beta' \mathbf{z}_{t-1}\right) + \varepsilon_t \tag{5}$$

showing that the common-factor model class (see Sargan, 1980; Hendry and Mizon, 1978) is also a restricted equilibrium-correction mechanism, constrained by the impact effects (from $\Delta \mathbf{z}_t$) being the same as the long-run effects (from $\mathbf{z}_{t-1}$).

2.3 ADLs as equilibrium-correction models

A first-order autoregressive distributed-lag (ADL) model is:

$$y_t = \beta_0 + \beta_1' \mathbf{z}_t + \beta_2 y_{t-1} + \beta_3' \mathbf{z}_{t-1} + \varepsilon_t \text{where} \varepsilon_t \sim \mathrm{IN}\left[0, \sigma_\varepsilon^2\right]. \tag{6}$$

The error $\{\varepsilon_t\}$ on (6) is an innovation against the available information, and its serial independence is part of the definition of the model, whereas normality and homoscedasticity are just for convenience. The condition $|\beta_2| < 1$ is needed to ensure a levels' equilibrium solution: Ericsson (2007) provides an extensive discussion. We consider (6) for both stationary and integrated $\{\mathbf{z}_t\}$, the latter denoting that some of the $\mathbf{z}_t$ have unit roots in their levels representations, but are stationary in differences.

First, under stationarity, taking expectations in (6) where $\mathrm{E}[y_t] = y^*$ and $\mathrm{E}[\mathbf{z}_t] = \mathbf{z}^* \ \forall t$:

$$\mathrm{E}\left[(1-\beta_2) y_t - \beta_0 - (\beta_1 + \beta_3)' \mathbf{z}_t\right] = 0 \tag{7}$$

so:

$$y^* = \frac{\beta_0}{1-\beta_2} + \frac{1}{1-\beta_2}(\beta_1 + \beta_3)' \mathbf{z}^* = \kappa_0 + \kappa_1' \mathbf{z}^*. \tag{8}$$

Since many economic theories have long-run partial equilibria like (8), they could be modelled by this class. Transforming (6) to differences and the equilibrium-correction term $(y - \kappa_0 - \kappa_1' \mathbf{z})_{t-1}$ delivers:

$$\Delta y_t = \beta_1' \Delta \mathbf{z}_t + (\beta_2 - 1)\left(y_{t-1} - \kappa_0 - \kappa_1' \mathbf{z}_{t-1}\right) + \varepsilon_t. \tag{9}$$

The immediate impact of a change in $\mathbf{z}_t$ on y_t is β_1, and the rapidity with which Δy_t converges to zero, which is its equilibrium outcome under stationarity, depends on the magnitude of $(\beta_2 - 1) < 0$; when both changes and ε_t are zero (their expectations), (7) results.

When y_t and $\mathbf{z}_t$ are integrated of order 1 (denoted I(1)), so are stationary in differences, the reformulation in (9) remains valid provided $|\beta_2| < 1$ in which case $(y_t - \kappa_0 - \kappa_1' \mathbf{z}_t)$ is a cointegration relation, as discussed in Section 4. Let $\mathrm{E}[\Delta \mathbf{z}_t] = \delta$ (say) so $\mathrm{E}[\Delta y_t] = \kappa_1' \delta = g_y$ where $\mathrm{E}[y_t - \kappa_1' \mathbf{z}_t] = \mu$, then taking expectations in (9) using (7):

$$g_y = \beta_1' \delta + (\beta_2 - 1)(\mu - \kappa_0) \tag{10}$$

and subtracting (10) from (9) delivers:

$$\Delta y_t = g_y + \beta_1' (\Delta \mathbf{z}_t - \delta) + (\beta_2 - 1)\left(y_{t-1} - \kappa_1' \mathbf{z}_{t-1} - \mu\right) + \varepsilon_t. \tag{11}$$

Re-specifying deterministic terms as in (11) plays an important role in EqCMs, both by helping to orthogonalize the regressors, and because of the pernicious effects of shifts in μ, a topic addressed in Section 5. It is so well known that the standard error of the mean of an IID random variable is the standard deviation of the data divided by the square root of the sample size that it hardly bears reiterating: except that it is somehow almost always ignored in this context. The standard error of the intercept in an EqCM equation like (11) should, therefore, be $\hat{\sigma}_\varepsilon / \sqrt{T}$ but is often a hundred times larger in reported empirical models, revealing a highly collinear specification (a similar comment applies to VARs). Moreover, a check on the model formulation follows from using sample means to estimate δ and μ, then checking that g_y has a sensible value, which may be given by theory (for example, no autonomous inflation, so $g_y = 0$).

Finally, if $\beta_2 = 1$, (9) equilibrium corrects in differences. An autoregression is the special case where $\beta_1 = \beta_3 = 0$, so is also an EqCM; and partial adjustment is another special case where now $\beta_3 = 0$.

2.4 GARCH as an equilibrium-correction model

As a fourth example, consider a non-integrated GARCH(1,1) process for ε_t, where $E[\varepsilon_t^2 | I_{t-1}] = \sigma_t^2$ when I_{t-1} denotes past information, and:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \theta\sigma_{t-1}^2 \tag{12}$$

with $0 < \alpha < 1,\ 0 < \theta < 1$ and $0 < \alpha + \theta < 1$. Let $\sigma_t^2 = \varepsilon_t^2 - v_t$, where $E[v_t] = 0$, then:

$$\varepsilon_t^2 = \omega + (\alpha + \theta)\varepsilon_{t-1}^2 + v_t - \theta v_{t-1} \tag{13}$$

where the equilibrium is:

$$\sigma_\varepsilon^2 \equiv E\left[\varepsilon_t^2\right] = \frac{\omega}{1 - (\alpha + \theta)}. \tag{14}$$

Substituting $\omega = (1 - (\alpha + \theta))\sigma_\varepsilon^2$ from (14) into the equation for σ_t^2:

$$\Delta\sigma_t^2 = (\theta - 1)\left(\sigma_{t-1}^2 - \sigma_\varepsilon^2\right) + \alpha\left(\varepsilon_{t-1}^2 - \sigma_\varepsilon^2\right). \tag{15}$$

Thus, the change in the conditional variance σ_t^2 responds less than proportionally ($\theta < 1$) to the previous disequilibrium between the conditional variance and the long-run variance, perturbed by the zero-mean discrepancy between the previous squared disturbance ε_{t-1}^2 and the long-run variance σ_ε^2, so the model equilibrium corrects to σ_ε^2, consistent with (14). ARCH is simply a special case.

2.5 VARs as equilibrium-correction models

The fifth example is an *n*-dimensional VAR with *m* lags and an innovation error $\varepsilon_t \sim \mathrm{IN}_n[0, \Omega_\varepsilon]$:

$$\mathbf{x}_t = \pi + \sum_{i=1}^{m} \mathbf{\Pi}_i x_{t-i} + \varepsilon_t \tag{16}$$

where the *nm* eigenvalues of the polynomial $|\mathbf{I}_n - \Sigma_{i=1}^{m} \mathbf{\Pi}_i L^i|$ in *L* determine the characteristics of the time series. If all the eigenvalues are inside the unit circle, (16) is stationary (when all the parameters are constant and the initial conditions also satisfy the process). In that case, $\mathbf{\Gamma} = (I_n - \Sigma_{i=1}^{m} \mathbf{\Pi}_i)$ is invertible and has all its eigenvalues inside the unit circle, so the process equilibrium corrects to $\psi = \mathbf{\Gamma}^{-1}\pi$. To illustrate for $m = 2$, (16) can be expressed as:

$$\Delta \mathbf{x}_t = (\mathbf{\Pi}_1 - \mathbf{I}_n)\Delta \mathbf{x}_{t-1} - \mathbf{\Gamma}(\mathbf{x}_{t-2} - \psi) + \varepsilon_t \tag{17}$$

where $\mathrm{E}[\Delta x_t] = 0$ by stationarity, so $\mathrm{E}[x_t - \psi] = 0$ is indeed the equilibrium to which x_t converges in the absence of further shocks. Conversely, if all the eigenvalues are unity, x_t is I(1) with $\mathbf{\Gamma} = 0$ in (17), so does not equilibrium correct in levels, but does so in the differences (unless their polynomial has further unit roots, making the process doubly integrated, I(2)). Finally, for a combination of eigenvalues inside and on the unit circle, Γ has reduced rank $0 < r < n$ equal to the number of non-unit eigenvalues, so can be expressed as $\mathbf{\Gamma} = \alpha\beta'$ where α and β also have rank *r*. Then π in (16) can be decomposed into the unconditional growth rate of x_t, denoted γ, and $\alpha\mu$ such that in place of (17), we have:

$$\Delta \mathbf{x}_t = \gamma + (\mathbf{\Pi}_1 - \mathbf{I}_n)(\Delta \mathbf{x}_{t-1} - \gamma) - \alpha\left(\beta' \mathbf{x}_{t-2} - \mu\right) + \varepsilon_t \tag{18}$$

so that $\mathrm{E}[\beta' x_t - \mu] = 0$ and the system converges to that equilibrium when the original variables are I(1), hence $\beta' x_{t-2}$ is an I(0) process which equilibrium corrects to μ. At the same time, Δx_t is an I(0) process which equilibrium corrects to γ, noting that $\beta'\gamma = 0$, whereas x_t drifts.

Linear simultaneous equations systems of time series are a restriction on a VAR, so are also EqCMs.

2.6 DSGEs as equilibrium-correction models

As a final brief example, well-defined general equilibrium systems have equilibria. Using Taylor-series expansions around the steady-state values of the discretized representation of a system of differential equations, Bårdsen, Hurn and Lindsay (2004) show that any dynamic system with a steady-state solution has a linear EqCM representation. Thus, they argue that linearizations of DSGEs imply linear EqCM representations. In principle, these could be in terms of changes only, corresponding to a steady-state path. More usually, level solutions result.

3 Historical overview

Equilibrium-correction models are a special case of the general class of proportional, derivative and integral control mechanisms, so have a long pedigree in that arena: for economics examples, see Phillips (1954); Phillips and Quenouille (1960); and Whittle (1963), with the links summarized in Salmon (1988). Explicit examples of EqCMs are presented in Sargan (1964) and were popularized by Davidson et al. (1978), although they were called ‘error-correction mechanisms’ (ECMs) by those authors. The major developments underlying cointegration in Engle and Granger (1987) established its isomorphism with equilibrium correction for integrated processes, leading to an explosion in the application of EqCMs and the development of a formal analysis of vector EqCM systems in Johansen (1988; 1995). We now review the two stages linking control mechanisms with error correction, then that with equilibrium correction.

3.1 Error correction and control mechanisms

Phillips (1954; 1957), in particular, pioneered the application of control methods for macroeconomic stabilization, specifically techniques for derivative, proportional and integral control servomechanisms. In this form of control, a target (say an unemployment rate of five per cent) is to be achieved by adjusting an instrument (say government expenditure), and changes to the instrument, its level, and cumulative past errors may need to be included in the rule to stabilize the target.
That approach is a precursor to the well-known linear-quadratic model in which one optimizes a quadratic function of departures from target trajectories for a linear dynamic system over a finite future horizon (see, for example, Holt et al., 1960; Preston and Pagan, 1982). For example, consider the quadratic cost function C_H which penalizes the deviations of a variable x_{t+j} from a pre-specified target trajectory $\left\{x^*_{t+j}\right\}$ subject to costs of adjustment from changes $\Delta x_t = x_t - x_{t-1}$ over an H-period horizon commencing at time t:

$$C_H = \sum_{j=0}^{H} c_{t+j} = \sum_{j=0}^{H} \frac{1}{2}\left[\left(x_{t+j} - x^*_{t+j}\right)^2 + \alpha\left(\Delta x_{t+j}\right)^2\right]. \tag{19}$$

To minimize c_{t+j} at time $t+j$, differentiate with respect to x_{t+j}, noting the intertemporal link that $\Delta x_{t+j+1} = x_{t+j+1} - x_{t+j}$ also depends on x_{t+j}, which yields (ignoring the end point for simplicity):

$$\frac{\partial C_H}{\partial x_{t+j}} = \frac{\partial c_{t+j}}{\partial x_{t+j}} + \frac{\partial c_{t+j+1}}{\partial x_{t+j}} = x_{t+j} - x^*_{t+j} + \alpha(\Delta x_{t+j}) - \alpha(\Delta x_{t+j+1}), \tag{20}$$

so equating to zero for a minimum for any *j*, and hence for $j = 0$:

$$x_t - x^*_t + \alpha\Delta x_t - \alpha\Delta x_{t+1} = 0.$$

Expressed as a polynomial in leads and lags in the operator *L* (for $\alpha \neq 0$):

$$\left(L^{-1} - \left(2 + \alpha^{-1}\right) + L\right)x_t = \left(L^{-1} - \lambda_2\right)(1 - \lambda_1 L)x_t = -\frac{x^*_t}{\alpha}. \tag{21}$$

The polynomial in (21) has roots λ_1 and λ_2 with a product of unity (so they are inverses, with λ_1 inside and λ_2 outside the unit circle) and a sum of $(2 + \alpha^{-1})$. Inverting the first factor $(L^{-1} - \lambda_2)$, using $(1/\lambda_2) = \lambda_1 < 1$ and expanding the last term as a power series in L^{-1}expresses x_t as a function of lagged x_s and current and future values of x^*_{t+k}:

$$(1 - \lambda_1 L)x_t = \frac{\lambda_1}{\alpha}\left(1 + \lambda_1 L^{-1} + \lambda_1^2 L^{-2} + \cdots\right)x^*_t = \frac{\lambda_1}{\alpha}\sum_{k=0}^{\infty}\lambda_1^k x^*_{t+k}. \tag{22}$$

Since $(1-\lambda_1) = \lambda_1 / \alpha(1-\lambda_1)$, let:

$$x_t^{**} = (1-\lambda_1)\sum_{k=0}^{\infty}\lambda_1^k x_{t+k}^{*} \tag{23}$$

denote the 'ultimate' target (scaled so that the weights sum to unity as in, for example, Nickell, 1985) then from (22) using (23), for $t < H$:

$$\Delta x_t = -(1-\lambda_1)\left(x_{t-1} - x_t^{**}\right) = (1-\lambda_1)\Delta x_t^{**} - (1-\lambda_1)\left(x_{t-1} - x_{t-1}^{**}\right). \tag{24}$$

Thus, x_t adjusts to changes in the ultimate target, and to the previous error from that target, and is an EqCM when $-1 < \lambda_1 < 1$. Mistakes in plans, errors in expectations, and relations between the ultimate target and its determinants all need to be modelled for an operational rule. To hit a moving target requires a feedforward rule, and the role of $\alpha(\Delta x_{t+j})^2$ in (19) is to penalize the controller from making huge changes to x_t when doing so. However, it is difficult to imagine real world adjustment costs being proportional to changes, which in any case then depend on the specification of x_t as logs, levels, proportions or even changes (see, for example, Nickell, 1985). Moreover, the entire class is partial adjustment, as (24) shows.

For 1-period optimization (so $H = 0$: see, for example, Hendry and Anderson, 1977), only the end point is relevant, so (20) delivers the planned value x_t^p as a function of $x_t^{**} = x_t^{*}$:

$$x_t^p - x_{t-1} = \frac{1}{1-\alpha}\left(x_t^{*} - x_{t-1}\right) = \rho\left(x_t^{*} - x_{t-1}\right). \tag{25}$$

When the error on the plan is $\varepsilon_t = x_t - x_t^p$, where $E[x_t^p \varepsilon_t] = 0$ under rationality, and $x_t^{*} = \beta' \mathbf{z}_t$ (say), (25) becomes:

$$\Delta x_t = \rho\left(\beta' \mathbf{z}_t - x_{t-1}\right) + \varepsilon_t = \rho\beta'\Delta\mathbf{z}_t - \rho\left(x_{t-1} - \beta'\mathbf{z}_{t-1}\right) + \varepsilon_t.$$

This is a partial adjustment again. The static regression in Section 2.1 has a more restrictive dynamic structure, but otherwise the properties of the ADL in Section 2.3 can vary over a wide range (see Hendry, 1995, ch. 6).

3.2 From error correction to equilibrium correction

The model in Sargan (1964) was explicitly an ECM for wages and prices (w_t and p_t denote their respective logs), building on previous models of wage and price inflation written as:

$$\Delta w_t = \beta_0 + \beta_1 \Delta p_t + \beta_2 \Delta w_{t-1} + \varepsilon_t. \tag{26}$$

When $\mathrm{E}[\varepsilon_t] = 0$ and the differenced variables are stationary with means $\mathrm{E}[\Delta w_t] = \dot{w}$ and $\mathrm{E}[\Delta p_t] = \dot{p}$, then the long-run steady-state solution to (26) is:

$$\dot{w} = \frac{\beta_0 + \beta_1 \dot{p}}{1 - \beta_2}.$$

As formulated, (26) does not establish any relationship between the levels w_t and p_t, hence these could drift apart. Since economic agents are concerned about the level of real wages, $w_t - p_t$, Sargan postulated the equilibrium:

$$(w - p)_{e,t} = \delta_0 + \delta_1 \Delta p_t + \delta_2' \mathbf{z}_t, \tag{27}$$

where $\mathbf{z}_t$ denotes a vector of additional variables, such as unemployment (u), productivity (q) and political factors. The disequilibrium is:

$$v_t = w_t - p_t - \delta_0 - \delta_1 \Delta p_t - \delta_2' \mathbf{z}_t \tag{28}$$

and, to re-establish equilibrium whenever the levels drift apart, he used the explicit adjustment equation:

$$\Delta w_t = \alpha\left(w_{t-1} - p_{t-1} - (w - p)_{e,t-1}\right) = \alpha v_{t-1}. \tag{29}$$

If a relation like (28) is well defined with v_t being I(0) when the levels are I(1), so the differences are I(0), then w_t forms a non-integrated combination with p_t and $\mathbf{z}_t$ so these variables are cointegrated (see, among many others, Engle and Granger, 1987; Phillips and Loretan, 1991; Banerjee, et al., 1993.).
A less restricted specification than (26) entails including the levels terms $(w - p)_{t-1}$ and $\mathbf{z}_{t-1}$ (and their differences), so if contemporaneous variables are excluded:

$$\Delta w_t = \pi_0 + \pi_1 \Delta p_{t-1} + \pi_2 \Delta w_{t-1} - \pi_3 (w - p)_{t-1} + \pi_4' \mathbf{z}_{t-1} + \pi_5' \Delta \mathbf{z}_{t-1} + u_t. \tag{30}$$

When $\pi_3 \neq 0$, the long-run levels equilibrium solution to (30) matching (27) is ($\varphi_4 = \pi_4 / \pi_3$):

$$\mathrm{E}\left[w - p - \varphi_4' \mathbf{z}\right] = f(\dot{w}, \dot{p}, \dot{\mathbf{z}}).$$

The model in (30) has both derivative and proportional control (e.g., Δp_{t-1} and $(w - p)_{t-1}$) following up Phillips (1954; 1957) (see Salmon, 1982). The proportional mechanism ensures the disequilibrium adjustment, based on the (possibly detrended) log-ratio of two nominal levels (see, for example, Bergstrom, 1962). The equivalent of g_y in Section 2.3 should be $\dot{p} + \dot{q}$ in (30) to avoid having ‘autonomous wage inflation’ independent of all economic forces.
The long-run stability of the ‘great ratios’ in Klein (1953) was often implicitly assumed to justify such transformations, but had come under question (see, for example, Granger and Newbold, 1977, and the discussion in Hendry, 1977), although Hendry and Mizon (1978) had argued that what mattered was that the errors in (30) were stationary, not that all the variables were stationary. Granger (1981) related the type

of model in (30) to cointegration, and Granger (1986) showed the important new result that one of Δw_t or Δp_t must depend on the equilibrium correction if w_t and p_t were cointegrated: the assumption in (29) is that w_t adjusts to the disequilibrium. If both variables w_t and p_t adjust to the disequilibrium, then p_t is not weakly exogenous for the $\{\pi_i\}$ (see Phillips and Loretan, 1991; Hendry, 1995). It is primarily because of cointegration that equilibrium-correction models like (30) have proved a popular specification. Engle and Granger (1987) showed that cointegration and proportional EqCM were equivalent, linking time-series approaches more closely with econometric modelling. Davidson and Hall (1991) also linked VARs as in Section 2.5 to target relations as discussed in Section 3.1 using cointegration analysis, so we now turn to the topic of cointegration in more detail.

4 Equilibrium-correction and cointegration

4.1 From the ADL to a VAR

To complete (6), a process is needed for $\{\mathbf{z}_t\}$. Let:

$$\mathbf{z}_t | y_{t-1}, \mathbf{z}_{t-1} \sim \mathrm{N}_k[\pi_{20} + \pi_{21} y_{t-1} + \pi_{22}\mathbf{z}_{t-1}, \Omega_{zz}]. \tag{31}$$

Given (31), the joint distribution is the first-order VAR:

$$\begin{pmatrix} y_t \\ \mathbf{z}_t \end{pmatrix} | y_{t-1}, \mathbf{z}_{t-1} \sim \mathrm{N}_{k+1}\left[\begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} + \begin{pmatrix} \pi_{11} & \pi_{12}' \\ \pi_{21} & \pi_{22} \end{pmatrix}\begin{pmatrix} y_{t-1} \\ \mathbf{z}_{t-1} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12}' \\ \sigma_{12} & \Omega_{zz} \end{pmatrix}\right]. \tag{32}$$

Consequently, to match (6):

$$\mathrm{E}[y_t | \mathbf{z}_t, y_{t-1}, \mathbf{z}_{t-1}] = \pi_{10} + \pi_{11} y_{t-1} + \pi_{12}\mathbf{z}_{t-1} + \sigma_{12}'\Omega_{zz}^{-1}(\mathbf{z}_t - \pi_{20} - \pi_{21} y_{t-1} - \pi_{22}\mathbf{z}_{t-1}), \tag{33}$$

so $\beta_0 = (\pi_{10} - \varphi'\pi_{20})$, $\beta_1 = \varphi$, $\beta_2 = \pi_{11} - \varphi'\pi_{21}$ and $\beta_3 = (\pi_{12} - \varphi'\pi_{22})$ when $\varphi = \Omega_{zz}^{-1}\sigma_{12}$, and

$\sigma_\varepsilon^2 = \sigma_{11} - \sigma_{12}'\Omega_{zz}^{-1}\sigma_{12}$. When $\mathbf{z}_t$ is weakly exogenous for $(\beta_0, \ldots, \beta_3')$, the model in (31) can be ignored when analysing (6); also $\pi_{21} = 0$ then ensures the strong exogeneity of $\mathbf{z}_t$ for $(\beta_0, \ldots, \beta_3')$.
Sufficient conditions for stationarity of (32) are that all the eigenvalues λ_i of the matrix of the $\{\pi_{ij}\}$ are inside the unit circle, but a more realistic setting allows for unit roots in π_{22}. On that basis, we now investigate the properties of the VAR in (32) letting $\mathbf{x}_t' = (y_t : \mathbf{z}_t')$ as in (16).

4.2 Cointegration

Linear combinations of I(1) processes are usually I(1) as well: differencing is still needed to remove the unit root. Sometimes integration cancels between series to yield an I(0) outcome and thereby deliver cointegration. Cointegrated processes in turn define a ‘long-run equilibrium trajectory’ for the economy, departures from which induce ‘equilibrium correction’ to move the economy back towards its path. A rationale for integrated–cointegrated data is that economic agents use fewer equilibrium corrections than there are variables they need to control. We can see that effect as follows.
Consider the bivariate VAR:

$$\begin{aligned} x_{1,t} &= \pi_{10} + \pi_{11}x_{1,t-1} + \pi_{12}x_{2,t-1} + \varepsilon_{1,t} \\ x_{2,t} &= \pi_{20} + \pi_{21}x_{1,t-1} + \pi_{22}x_{2,t-1} + \varepsilon_{2,t}, \end{aligned} \tag{34}$$

where $(\varepsilon_{1,t}, \varepsilon_{2,t})$ are bivariate independent normal. To determine when the system is I(1) and if so, whether or not some linear combinations of variables are cointegrated, rewrite (34) as:

$$\begin{pmatrix} \Delta x_{1,t} \\ \Delta x_{2,t} \end{pmatrix} = \begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} + \begin{pmatrix} (\pi_{11} - 1) & \pi_{12} \\ \pi_{21} & (\pi_{22} - 1) \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \tag{35}$$

or as (a special case of (18)):

$$\Delta \mathbf{x}_t = \pi + \Pi \mathbf{x}_{t-1} + \varepsilon_t. \tag{36}$$

Three cases are of interest. First $\Pi = 0$, so (36) is a vector random walk without any levels relationships, and so $\mathbf{x}_t$ is I(1) with $\Delta \mathbf{x}_t$ being I(0) and equilibrium correcting to π. Secondly, if Π has full rank, then $\mathbf{x}_t$ is I(0) and equilibrium corrects to $\Pi^{-1}\pi$. The most interesting case is when Π is reduced rank so can be expressed as:

$$\Pi = \alpha\beta' = \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \end{pmatrix} (\beta_{11} \quad \beta_{12}),$$

where we will normalize $\beta_{11} = 1$. Then in (35):

$$\begin{pmatrix} \Delta x_{1,t} \\ \Delta x_{2,t} \end{pmatrix} = \begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} + \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \end{pmatrix} (1 \quad \beta_{12}) \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} = \begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} + \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \end{pmatrix} (x_{1,t-1} + \beta_{12} x_{2,t-1}) + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \tag{37}$$

which is an EqCM with $(x_{1,t-1} + \beta_{12} x_{2,t-1})$ stationary. Thus, cointegration entails EqCM and vice versa when the feedback relation is I(0). However, prior to Granger (1981) the EqCM literature did not visualize a single cointegration relation affecting several variables, and thereby making them integrated, but instead just took the non-stationarity of the observed data as due to the behaviour of the non-modelled variables. Consequently, system cointegration ‘endogenizes’ data integrability in a consistent way, and so represents a significant step forward. The extensive literature on cointegration analysis also addresses most of the estimation and formulation issues that arise when seeking to conduct inference in integrated-cointegrated processes: much of this is summarized in Hendry and Juselius (2001), to which the interested reader is referred for bibliographic perspective.

5 Equilibrium correction and forecast failure

Recent research on the impact of structural breaks, particularly location shifts, on cointegrated processes has emphasized the need to distinguish equilibrium correction, which operates successfully only within regimes, from error correction, which stabilizes in the face of other non-stationarities (see, for example, Clements and Hendry, 1995). The assumptions concerning the stationarity, or otherwise, of the entity to be controlled in Section 3.1 were rarely explicitly stated, but suggest an implicitly stationary system (or perhaps steady-state growth). In such a setting, equilibrium-correction or cointegration relationships prevent the levels of the variables from ‘drifting apart’, and so improve the properties of forecasts. Practical work, however, must allow the data generation process to be non-stationary both from unit roots

(that is, I(1) or possibly I(2)) and from a lack of time invariance. When data processes are non-stationary even after differencing and cointegration, equilibrium-correction mechanisms tend to suffer from forecast failure, defined as a significant deterioration in forecast performance relative to in-sample behaviour. Since most empirical model forms are members of the EqCM class, this is a serious practical problem. To illustrate, reconsider the special case of (18) with just one lag, written as:

$$\Delta \mathbf{x}_t = \gamma + \alpha\left(\beta' \mathbf{x}_{t-1} - \mu\right) + \varepsilon_t. \tag{38}$$

The shift of interest here is $\nabla\mu^* = \mu^* - \mu$, where μ^* denotes the post-break equilibrium mean (reasonable magnitude shifts in γ, α and Ω_ε rarely entail forecast failure). Denote the forecast origin as time *T*, then following a change to μ^* immediately after forecasting, the next outcome is:

$$\Delta \mathbf{x}_{T+1} = \gamma + \alpha\left(\beta' \mathbf{x}_T - \mu^*\right) + \varepsilon_{T+1} = \gamma + \alpha\left(\beta' \mathbf{x}_T - \mu\right) + \varepsilon_{T+1} - \alpha\nabla\mu^* \tag{39}$$

where $-\alpha\nabla\mu^*$ is the unanticipated break, and becomes the mean forecast error for known parameters. Importantly, the 1-step ahead forecast at $T+1$ using an unchanged model suffers the same mistake:

$$\mathrm{E}\left[\Delta \mathbf{x}_{T+2} - \left(\gamma + \alpha\left(\beta' \mathbf{x}_{T+1} - \mu\right)\right)\right] = -\alpha\nabla\mu^* \tag{40}$$

so the shift in the equilibrium mean induces systematic mis-forecasting. The impact on multi-step forecasts of the levels is even more dramatic, as the mean forecast error increases at every horizon, eventually converging to $\alpha(\beta'\alpha)^{-1}\nabla\mu^*$, which can be very large (see Clements and Hendry, 1999). Thus, EqCMs are a non-robust forecasting device in the face of equilibrium-mean shifts, a comment which therefore applies to all members of this huge class of model, including GARCH (as noted earlier), where the pernicious shift is in the unconditional variance σ_ε^2 in (15).

To avoid forecast failure, more adaptive methods merit consideration. One generic approach to improving robustness to location shifts is to difference the forecasting device (although that may well worsen the impact of large measurement errors at the forecast origin). Differencing can be before estimation, as in a double-differenced VAR, or after, as in differencing the estimated EqCM to eliminate the equilibrium

mean and growth intercept. Such devices perform as badly as the EqCM in terms of forecast biases when a break occurs after forecasts are announced (see Clements and Hendry, 1999), and have a larger error variance. The key difference is their performance when forecasting after a break has already occurred, in which case the EqCM continues to perform badly (as shown above in (48)), but a DEqCM becomes relatively immune to the earlier break. Taking (47) as an example, an additional difference yields:

$$\Delta^2 \mathbf{x}_{T+1} = \Delta\left(\gamma - \alpha\mu^*\right) + \alpha\beta'\Delta\mathbf{x}_T + \Delta\varepsilon_{T+1} = -\alpha\nabla\mu^* + \alpha\beta'\Delta\mathbf{x}_T + \Delta\varepsilon_{T+1}$$

so there is no benefit when forecasting immediately after the break (as $\Delta\mu^* = \nabla\mu^*$), whereas (48) becomes:

$$\Delta^2 \mathbf{x}_{T+2} = \alpha\beta'\Delta\mathbf{x}_{T+1} + \Delta\varepsilon_{T+2}$$

since $\Delta\mu^* = 0$. Thus, there is no longer any systematic failure. The same comment applies to double-differenced devices, although Hendry (2006) shows how to improve these while retaining robustness.

A further consequence is that, when a location shift is not modelled, since most econometric estimators minimize mis-fitting, the coefficients of dynamic models will be driven towards unity, which induces differencing to convert a location shift into a ‘blip’. Thus, estimates that apparently manifest ‘slow adjustment’ may just reflect unmodelled breaks.

An alternative approach to avoiding forecast failure would be to construct a genuine error-correction model, adjusting more or less rapidly to wherever the target variable moves: for example, exponentially weighted moving averages do so for some processes. In essence, either the dynamics must ensure correction or the target implicit in the econometric model must move when the regime alters. This last result also explains why models in differences are not as susceptible to certain forms of structural break as equilibrium-correction systems (again see Clements and Hendry, 1999), and in turn helps to account for many of the findings reported in the forecasting competitions literature. When the shift in question is a change in a policy regime, Hendry and Mizon (2005) suggest approaches to merging robust forecasts with policy models.

6 Conclusion

Equilibrium-correction models have a long pedigree as an ‘independent’ class, related to optimal control theory. However, their isomorphism with cointegrated relationships has really been the feature that has ensured their considerable popularity in empirical applications. In both cases, part of the benefit from the EqCM specification came from expressing variables in the more orthogonalized forms of differences and equilibrium-correction terms, partly from the resulting insights into both short-run and long-run

adjustments, partly from discriminating between the different components of the deterministic terms, and partly from ‘balancing’ regressors of the same order of integration, namely I(0).

Unfortunately, science is often two steps forward followed by one back, and that backwards step came from an analysis of EqCMs when forecasting in the face of structural breaks. Unmodelled shifts in the equilibrium mean (and less so in the growth rate) induce forecast failure, making EqCMs a non-robust device with which to forecast when data processes are prone to breaks, as many empirical studies suggest they are (see, for example, Stock and Watson, 1996). Since cointegration hopefully captures long-run causal relations, and ties together the levels of I(1) variables, eliminating its contribution should not be undertaken lightly, hence the suggestion in Section 5 of using the differenced version of the estimated EqCM for forecasting.

See Also

- cointegration

Financial support from the ESRC under Professorial Research Fellowship RES051270035 is gratefully acknowledged, as are helpful comments from Gunnar Bårdsen, Julia Campos, Jennifer Castle, Mike Clements, Søren Johansen and Graham Mizon.

Bibliography

Bårdsen, G., Hurn, S. and Lindsay, K.A. 2004. Linearizations and equilibrium correction models. *Studies in Nonlinear Dynamics and Econometrics* 8(4), Article 5.

Banerjee, A., Dolado, J.J., Galbraith, J.W. and Hendry, D.F. 1993. *Co-integration, Error Correction and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Bergstrom, A.R. 1962. A model of technical progress, the production function and cyclical growth. *Economica* 29, 357–70.

Bollerslev, T. 1986. Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics* 51, 307–27.

Clements, M.P. and Hendry, D.F. 1995. Macro-economic forecasting and modeling. *Economic Journal* 105, 1001–13.

Clements, M.P. and Hendry, D.F. 1999. *Forecasting Non-stationary Economic Time Series*. Cambridge, MA: MIT Press.

Davidson, J.E.H. and Hall, S. 1991. Cointegration in recursive systems. *Economic Journal* 101, 239–51.

Davidson, J.E.H., Hendry, D.F., Srba, F. and Yeo, J.S. 1978. Econometric modelling of the aggregate time-

series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88, 661–92.

Engle, R.F. 1982. Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

Engle, R.F. and Granger, C.W.J. 1987. Cointegration and error correction: representation, estimation and testing. *Econometrica* 55, 251–76.

Ericsson, N.R. 2007. *Econometric Modeling*. Oxford: Oxford University Press.

Granger, C.W.J. 1981. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–30.

Granger, C.W.J. 1986. Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics* 48, 213–28.

Granger, C.W.J. and Newbold, P. 1977. The time series approach to econometric model building. In *New Methods in Business Cycle Research*, ed. C.A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Hendry, D.F. 1977. On the time series approach to econometric model building. In *New Methods in Business Cycle Research*, ed. C.A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Hendry, D.F. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D.F. 2006. Robustifying forecasts from equilibrium-correction models. *Journal of Econometrics* 135, 399–426. Special issue in honor of Clive Granger.

Hendry, D.F. and Anderson, G.J. 1977. Testing dynamic specification in small simultaneous systems: an application to a model of building society behaviour in the United Kingdom. In *Frontiers in Quantitative Economics*, vol. 3, ed. M.D. Intriligator. Amsterdam: North-Holland.

Hendry, D.F. and Juselius, K. 2001. Explaining cointegration analysis: Parts I and II. *Energy Journal* 21, 1–42; 22, 75–120.

Hendry, D.F. and Mizon, G.E. 1978. Serial correlation as a convenient simplification, not a nuisance: a comment on a study of the demand for money by the Bank of England. *Economic Journal* 88, 549–63.

Hendry, D.F. and Mizon, G.E. 2005. Forecasting in the presence of structural breaks and policy regime shifts. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D.W.K. Andrews and J.H. Stock. Cambridge: Cambridge University Press.

Holt, C., Modigliani, F., Muth, J.F. and Simon, H. 1960. *Planning Production, Inventories and Work Force*. Englewood Cliffs, NJ: Prentice-Hall.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–54.

Johansen, S. 1995. *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Klein, L.R. 1953. *A Textbook of Econometrics*. Evanston, IL: Row, Peterson and Company.

Nickell, S.J. 1985. Error correction, partial adjustment and all that: an expository note. *Oxford Bulletin of Economics and Statistics* 47, 119–30.

Phillips, A.W.H. 1954. Stabilization policy in a closed economy. *Economic Journal* 64, 290–333.

Phillips, A.W.H. 1957. Stabilization policy and the time form of lagged response. *Economic Journal* 67, 265–77.

Phillips, P.C.B. and Loretan, M. 1991. Estimating long-run economic equilibria. *Review of Economic Studies* 58, 407–36.

Phillips, A.W.H. and Quenouille, M.H. 1960. Estimation, regulation and prediction in interdependent dynamic systems. *Bulletin de l'Institut de Statistique* 37, 335–43.

Preston, A.J. and Pagan, A.R. 1982. *The Theory of Economic Policy*. Cambridge: Cambridge University Press.

Salmon, M. 1982. Error correction mechanisms. *Economic Journal* 92, 615–29.

Salmon, M. 1988. Error correction models, cointegration and the internal model principle. *Journal of Economic Dynamics and Control* 12, 523–49.

Sargan, J.D. 1964. Wages and prices in the United Kingdom: a study in econometric methodology (with discussion). In *Econometric Analysis for National Economic Planning*, vol. 16 of *Colston Papers*, ed. P.E. Hart, G. Mills and J.K. Whitaker. London: Butterworth.

Sargan, J.D. 1980. Some tests of dynamic specification for a single equation. *Econometrica* 48, 879–97.

Sims, C.A., ed. 1977. *New Methods in Business Cycle Research*. Minneapolis: Federal Reserve Bank of Minneapolis.

Stock, J.H. and Watson, M.W. 1996. Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* 14, 11–30.

Whittle, P. 1963. *Prediction and Regulation by Linear Least-Square Methods*. Princeton: D. Van Nostrand.

How to cite this article

Hendry, David F. "equilibrium-correction models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000284> doi:10.1057/9780230226203.0496

The New Palgrave Dictionary of Economics Online

equivalence scales

Arthur Lewbel and Krishna Pendakur

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An equivalence scale is a measure of the cost of living of a household of a given size and demographic composition, relative to the cost of living of a reference household (usually a single adult), when both households attain the same level of utility or standard of living. Equivalence scales are difficult to construct because household utility cannot be directly measured, which results in economic identification problems. Applications of equivalence scales include measurement of social welfare, economic inequality, poverty, and costs of children; indexing payments for social benefits, life insurance, alimony, and legal compensation for wrongful death.

Keywords

consumer expenditure; Engel scales; equivalence scales; happiness, economics of; interpersonal utility comparisons; Marshallian demand functions; neuroeconomics; poverty lines; revealed preference theory; Rothbard scales; Shephard's Lemma; well-being

Article

History

Providing two different households with the same standard of living, making them equally well off, requires some definition of well-being. In the early literature on equivalence scales, a household's well-being was defined in terms of needs, such as having a nutritionally adequate diet.
Engel (1895) observed that a household's food expenditures are an increasing function of income and of family size, but that richer households tend to spend a smaller share of their total budget on food than poorer households. He therefore proposed that this food budget share could be a measure of a household's welfare or standard of living. The resulting Engel equivalence scale is defined as the ratio of incomes of two different sized households that have the same food budget share. This is essentially the

method used by the US Census Bureau to measure poverty. The bureau first defines the poverty line for a typical household as three times the cost of a nutritionally adequate diet, then uses food shares (Engel scales) to derive comparable poverty lines for households of different sizes and compositions, and finally adjusts the results annually by the consumer price index to account for inflation (see Fisher, 1997).

Given two households that differ only in their number or age distribution of children, Rothbarth (1943) equivalence scales are similar to Engel scales. They can be defined as the ratio of incomes of the two households when each household purchases the same quantity of some good that is only consumed by adults, such as alcohol, tobacco, or adult clothing.

Modern equivalence scales measure well-being in terms of utility, using cost (expenditure) functions estimated from consumer demand data via revealed preference theory. Engel or Rothbarth scales are equivalent to valid cost function based equivalence scales only under strong restrictions regarding the dependence of demand functions on characteristics such as age and family size, and on the links between demand functions and utility for these different household types.

One strand of the equivalence scale literature focuses on the former issue, and so deals primarily with the empirical question of how best to model the dependence of household Marshallian demand functions on demographic characteristics. Examples are Sydenstricker and King (1921), Prais and Houthakker (1955), and Barten (1964) scales, in which a different Engel type scale is constructed for every good people purchase, roughly corresponding to a different economies of scale measure for each good. Other examples are Gorman's (1976) general linear technologies, Lewbel's (1985) modifying functions, and Pendakur's (1999) shape invariance.

The second, closely related literature, focuses on the joint restrictions on both preferences and interpersonal comparability of utility required for measuring the relative costs of providing one household with the same utility level as another. Examples include Jorgenson and Slesnick (1987), Lewbel (1989), Blackorby and Donaldson (1993), and Donaldson and Pendakur (2004; 2006).

Definition

Consider a consumer (an individual or a household) with a vector of demographic characteristics **z** and nominal total expenditures x that faces the M vector **p** of prices of M different goods. The consumer chooses a bundle of goods to maximize utility given a linear budget constraint. Define the cost (expenditure) function $x=C(\mathbf{p}, u, \mathbf{z})$ which equals the minimum expenditure required for a consumer with characteristics **z** to attain utility level u when facing prices **p**. $C(\mathbf{p}, u, \mathbf{z})$ is a conditional cost function in the sense of Pollak (1989) because it gives the expenditure necessary to attain a utility level u, conditional on the consumer having characteristics **z**.

Equivalence scales relate the expenditures of a consumer with characteristics **z** to a consumer with a reference vector of characteristics $\bar{\mathbf{z}}$. The reference vector of characteristics may describe, for example, a single, medically healthy, middle-aged childless man. The equivalence scale is defined by $D(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, u, \mathbf{z}) / C(\mathbf{p}, u, \bar{\mathbf{z}})$. Equivalent-expenditure $X(\mathbf{p}, x, \mathbf{z})$ is defined as the expenditure level needed to bring the well-being of a reference household to the level of well-being of a household with characteristics **z**, so $X(\mathbf{p}, x, \mathbf{z}) = x / D(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, u, \bar{\mathbf{z}})$ where u is replaced by the indirect utility function, that is, $x=C(\mathbf{p}, u, \mathbf{z})$ solved for u.

Identification

In economics, a parameter is said to be ‘identified’ if its numerical value can be determined given enough observable data. Here we show why identification of equivalence scales requires either strong untestable assumptions regarding preferences or unusual types of data. Equivalence scales depend on utility, which cannot be directly observed and so must be inferred from consumer demand data, that is, from the quantities that consumers buy of different goods in varying price regimes and at various income levels. The observable (Marshallian) demand functions for goods derived from a conditional cost function $C(\mathbf{p}, u, \mathbf{z})$ are the same as those obtained from $C(\mathbf{p}, \phi(u, \mathbf{z}), \mathbf{z})$ for any function $\phi(u, \mathbf{z})$ that is strictly monotonically increasing in u. By revealed preference theory, demand data identifies the shape and ranking of a consumer's indifference curves over bundles of goods, but not the actual utility level associated with each indifference curve. Changing $\phi(u, \mathbf{z})$ just changes the utility level associated with each indifference curve.

Therefore, given any $C(\mathbf{p}, u, \mathbf{z})$ derived from demand data, the consumer's true cost of attaining a utility level u is $C(\mathbf{p}, \phi(u, \mathbf{z}), \mathbf{z})$ for some unknown function ϕ, so true equivalence scales are $D(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, \varphi(u, \mathbf{z}), \mathbf{z}) / C(\mathbf{p}, \varphi(u, \bar{\mathbf{z}}), \bar{\mathbf{z}})$. This is the source of equivalence scale non-identification. We cannot identify $D(\mathbf{p}, u, \mathbf{z})$ because the change from $\bar{\mathbf{z}}$ to $\mathbf{z}$ has an unobservable affect on D through ϕ. The problem is that revealed preferences over goods identify one set of indifference curves for households of type $\mathbf{z}$ and another set for households of type $\bar{\mathbf{z}}$, but we have no way of observing which indifference curve of type $\mathbf{z}$ yields the same level of utility as any given indifference curve of type $\bar{\mathbf{z}}$.

Given only goods demand data, Blundell and Lewbel (1991) show that changes in equivalence scales that result from price changes can be identified, but the levels of equivalence scales are completely unidentified, because for any cost function C and any positive number d, there exists a $\phi(u, \mathbf{z})$ function that makes $D(\mathbf{p}, u, \mathbf{z})=d$. Changes in D resulting from price changes can be identified because the ratio $D(\mathbf{p}_1, u, \mathbf{z})/D(\mathbf{p}_0, u, \mathbf{z})$ equals a ratio of ordinary identifiable cost of living (inflation) indices.

Identification of equivalence scales therefore requires either additional information or untestable assumptions regarding preferences over characteristics $\mathbf{z}$ and hence regarding ϕ. There are also other identification issues associated with equivalence scales. For example, different members of a household may have different standards of living, so a single level of utility that applies to the entire household to be compared or equated to anything may simply not exist. Lewbel (1997) lists additional equivalence scale identification issues.

Identification from demand data

Let w^j be the fraction of total expenditures a household spends on the jth good (its budget share) and let $\mathbf{w}$ be the vector of budget shares of all purchased goods. Shephard's Lemma states that $\mathbf{w} = \omega(\mathbf{p}, u, \mathbf{z}) = \nabla_{\ln \mathbf{p}} \ln C(\mathbf{p}, u, \mathbf{z})$, the price elasticity of cost. Let $w_f = \omega_f(\mathbf{p}, u, \mathbf{z})$ indicate the food equation. Engel's method notes that since ω_f is monotonically declining in utility u, w_f may be taken as an indicator of well-being. If, in addition, w_f indicates the same level of well-being for all household

types **z**, then the expenditure levels which equate the food share w_f, across household types are the equivalent-expenditure function, whose ratios give the equivalence scale. Monotonicity of ω_f in u is observable, but the second restriction concerning utility levels for different types of households refers to ϕ and so is not testable.

The Rothbarth approach is similar. Let $q_a = h_a(\mathbf{p}, u, \mathbf{z})$ indicate the quantity demanded for a good consumed only by adults, such as alcohol. If h_a is increasing in utility (a testable restriction), q_a may be taken as an indicator of the well-being of adult household members. If, in addition, q_a indicates the same level of adult well-being for adults living in all types of households (untestable), then the expenditure levels which equate q_a across household types are the equivalent-expenditure function, whose ratios again give the (Rothbarth) equivalence scale.

Lewbel (1989) and Blackorby and Donaldson (1993) consider the case where the equivalence scale function is independent of utility, which they call ‘independence of base’ (IB) and ‘equivalence-scale exactness’ (ESE), respectively. In this case there is a function Δ such that $D(\mathbf{p}, u, \mathbf{z}) = \Delta(\mathbf{p}, \mathbf{z})$ and $C(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, u, \bar{\mathbf{z}})\Delta(\mathbf{p}, \mathbf{z})$. The special case where $D(\mathbf{p}, u, \mathbf{z})$ is also independent of **p** yields Engel scales.

Given IB/ESE, Shephard's Lemma implies that $\omega(\mathbf{p}, u, \mathbf{z}) = \omega(\mathbf{p}, u, \bar{\mathbf{z}}) + \mathbf{n}(\mathbf{p}, \mathbf{z})$, where $\mathbf{n}(\mathbf{p}, \mathbf{z}) = \nabla_{\ln \mathbf{p}} \ln \Delta(\mathbf{p}, \mathbf{z})$. Since households with the same equivalent expenditure have the same utility, and since in this case, equivalent expenditure is given by $x/\Delta(\mathbf{p}, \mathbf{z})$, we may write the relation as $\mathbf{w}(\mathbf{p}, x, \mathbf{z}) = \mathbf{w}(\mathbf{p}, x/\Delta(\mathbf{p}, \mathbf{z}), \bar{\mathbf{z}}) + \mathbf{n}(\mathbf{p}, \mathbf{z})$, where $\mathbf{w}(\cdot)$ is the Marshallian budget share vector. Here, $\Delta(\mathbf{p}, \mathbf{z})$ ‘shrinks’ the budget share functions in the expenditure direction, and the amount of ‘shrinkage’ identifies the equivalence scale. Pendakur (1999) shows that this ‘shape invariance’ expression equals the testable implications required for IB/ESE. The untestable restriction, which uniquely defines $\phi(u, \mathbf{z})$ (up to transformations of u that do not depend on **z**) is that all households with the same value of $x/\Delta(\mathbf{p}, \mathbf{z})$ have the same level of utility. Blackorby and Donaldson (1993) show when cost functional forms uniquely identify IB/ESE. Donaldson and Pendakur (2004; 2006) consider identification for equivalence scales with more general functional forms.

Other sources of identification

Equivalence scale identification depends on how we define utility or well-being. Identification is not a problem if what we mean by making households equally well off refers to some observable characteristic such as nutritional adequacy of diet. As an alternative to revealed preference, identification may be based on surveys that ask respondents to either report their happiness (and hence utility) on some ordinal scale, or ask, based on introspection, how their utility or costs would change in response to changes in household characteristics. An early example is Kapteyn and Van Praag (1976), who estimate equivalence scales based on surveys where households rank income levels as ‘excellent’, ‘sufficient’, and so on. Identification requires comparability of these ordinal utility measures across consumers. Happiness studies by psychologists and experimental economists may prove useful for validating these types of subjective responses regarding utility, especially with recent neuroeconomic results measuring brain activity associated with pleasure, regret, and economic decision-making (see, for example, McFadden, 2005).

Another possible source of identification is when consumers can choose **z**, and we can collect information relevant to these choices. Assuming **z** is chosen to maximize utility can provide information about how utility varies with **z**, and hence may restrict the set of possible ϕ transformations. With enough information regarding how **z** is chosen one could identify ‘unconditional’ cost or utility functions over both goods and **z** and thereby identify the dependence of ϕ on **z**. Pollak (1989) refers to the use of unconditional versus conditional data to calculate the cost of demographic changes as ‘situation comparisons’ versus ‘welfare comparisons’.

Traditional equivalence scales assign a single level of utility to a household, implicitly assuming that all household members have the same utility level and hence ignoring the effects of the within-household distribution of resources. Features of this intra-household allocation of resources can be identified and estimated with demand data. Given the indifference curves and resource shares of each household member, instead of trying to calculate the cost of making an individual as well off as a household, one may instead calculate the cost of putting the individual on the same indifference curve when living alone that he attained as a member of a household. Whereas the former calculation requires a welfare comparison, the latter calculation only involves comparing the same individual in two different price and income environments. Browning, Chiappori and Lewbel (2006) call this type of comparison an ‘indifference scale’, and provide one set of conditions under which such scales can be non-parametrically identified.

Applications of equivalence scales

Equivalent expenditures and equivalence scales may be used for social evaluation, for example, inequality and poverty analysis. Given an equivalence scale, *di*, and household expenditure, x_i, for each person *i* in a population, one constructs equivalent expenditure for each person: $x_i^e = x_i / d_i$.

Expenditure data are observed at the level of the household, but x_i^e is constructed for each individual. By construction, the population distribution of equivalent expenditures is equivalent in welfare terms to the actual distribution of expenditures across households. Therefore, one can use this ‘as if’ distribution for constructing population measures of poverty or inequality, or for calculating the welfare implications of tax and transfer programmes.

Equivalence scales can also be used to calibrate social benefits payments and poverty lines. For example, if the social benefit rate (or poverty line) $\bar{x}$ is agreed upon for a single household type, for example, a single childless adult, then one could use equivalence scales to set rates for other household types **z** as $D(\mathbf{p}, u, \mathbf{z})\bar{x}$ where *u* is the utility level of the reference type with expenditures $\bar{x}$. Some statistical agencies flow information in the other direction: poverty lines are constructed for each household type, which can then be use to construct an implicit ‘poverty relative’ equivalence scale. If scales are IB/ESE, this provides enough information to identify equivalence scales for all households. Other applications of equivalence scales are for life insurance, alimony, and wrongful death calculations (see Lewbel, 2003), and for indirectly measuring the cost of children based on equivalence scales for households of different sizes.

See Also

- consumer expenditure
- cost functions
- demand theory
- Engel curve
- Engel's law
- Hicksian and Marshallian demands
- identification
- inequality (measurement)
- interpersonal utility comparisons
- poverty lines
- welfare economics

Bibliography

Barten, A.P. 1964. Family composition, prices, and expenditure patterns. In *Econometric Analysis for National Economic Planning: 16th Symposium of the Colston Society*, ed. P. Hart, L. Mills and J.K. Whitaker. London: Butterworth.

Blackorby, C. and Donaldson, D. 1993. Adult-equivalence scales and the economic implementation of interpersonal comparisons of well-being. *Social Choice and Welfare* 10, 335–61.

Blundell, R.W. and Lewbel, A. 1991. The information content of equivalence scales. *Journal of Econometrics* 50, 49–68.

Browning, M., Chiappori, P.-A. and Lewbel, A. 2006. Estimating consumption economies of scale, adult equivalence scales, and household bargaining power. Working Paper No. 588. Boston College, MA.

Donaldson, D. and Pendakur, K. 2004. Equivalent-expenditure functions and expenditure-dependent equivalence scales. *Journal of Public Economics* 88, 175–208.

Donaldson, D. and Pendakur, K. 2006. The identification of fixed costs from consumer behaviour. *Journal of Business and Economic Statistics* 24, 255–65.

Engel, E. 1895. Die Lebenskosten Belgischer Arbeiter-Familien Fruher and jetzt. *International Statistical Institute Bulletin* 9, 1–74.

Fisher, G.M. 1997. The development and history of the U.S. poverty thresholds – a brief overview. Newsletter of the Government Statistics Section and the Social Statistics Section of the American Statistical Association. Washington, DC.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in Economic Analysis: Proceedings of the 1975 AUTE Conference*, ed. M. Sheffield, J. Artis and A.R. Nobay. Cambridge: Cambridge University Press.

Jackson, C.A. 1968. Revised equivalence scale for estimating equivalent incomes or budget cost by family type. Bulletin 1570–2. Washington, DC: U.S. Bureau of Labor Statistics.

Jorgenson, D.W. and Slesnick, D.T. 1987. Aggregate consumer behavior and household equivalence scales. *Journal of Business and Economic Statistics* 5, 219–32.

Kapteyn, A. and Van Praag, B. 1976. A new approach to the construction of family equivalence scales. *European Economic Review* 7, 313–35.

Lewbel, A. 1985. A unified approach to incorporating demographic or other effects into demand systems. *Review of Economic Studies* 52, 1–18.

Lewbel, A. 1989. Household equivalence scales and welfare comparisons. *Journal of Public Economics* 39, 377–91.

Lewbel, A. 1997. Consumer demand systems and household equivalence scales. In *Handbook of Applied Econometrics*, vol. 2. *Microeconomics*, ed. M.H. Pesaran and P. Schmidt. Oxford: Blackwell.

Lewbel, A. 2003. Calculating compensation in cases of wrongful death. *Journal of Econometrics* 113, 115–28.

McFadden, D. 2005. The new science of pleasure, consumer behavior and the measurement of well-being. Frisch Lecture. London: Econometric Society World Congress.

Pendakur, K. 1999. Estimates and tests of base-independent equivalence scales. *Journal of Econometrics* 88, 1–40.

Pollak, R.A. 1989. *The Theory of the Cost of Living Index*. New York: Oxford University Press.

Prais, S.J. and Houthakker, H.S. 1955. *The Analysis of Family Budgets*. Cambridge: Cambridge University Press.

Rothbarth, E. 1943. Note on a method of determining equivalent income for families of different composition. In *War-time Pattern of Saving and Spending*, ed. C. Madge. Cambridge: Cambridge University Press.

Sydenstricker, E. and King, W.I. 1921. The measurement of the relative economic status of families.

Quarterly Publication of the American Statistical Association 17, 842–57.

How to cite this article

Lewbel, Arthur and Krishna Pendakur. "equivalence scales." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000240>
doi:10.1057/9780230226203.0498

The New Palgrave Dictionary of Economics Online

ergodicity and nonergodicity in economics

Ulrich Horst

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A random economic system is called ergodic if it tends in probability to a limiting form that is independent of the initial conditions. Breakdown of ergodicity gives rise to path dependence. We illustrate the importance of ergodicity and breakdown thereof in economics by reviewing some work of non-market interactions. This includes microeconomic models of endogenous preference formation, macroeconomics models of economic growth, and models of social interaction.

Keywords

ergodicity and non-ergodicity in economics; path dependence; endogenous preference formation; Ising economy; Gibbs distribution theory; Markov processes; social interaction

Article

A stochastic system is called ergodic if it tends in probability to a limiting form that is independent of the initial conditions. Breakdown of ergodicity gives rise to path dependence. Path-dependent features of economics range from small-scale technical standards to large-scale institutions. Prominent examples include technical standards, such as the ‘QWERTY’ standard typewriter keyboard and the ‘standard gauge’ of railway track. Ergodicity and breakdown thereof is of particular relevance to models of social interaction. We illustrate this importance, summarizing some work on endogenous preference formation, dynamic population games, and models of non-market interaction.

Endogenous preference formation

In his pioneering paper on endogenous preference formation, Föllmer (1974) developed an equilibrium analysis of large exchange economies where the conditional excess demand $z(x^a, p)$ of the agent $a \in A$ given a price system p is subject to a random shock x^a and where the probabilities $(\pi_a)_{a \in A}$ governing

this randomness have an interactive structure.

In a benchmark model where the *states* x^a are independent across agents, the distribution μ of the vector of states $x = (x^a)_{a \in A}$ takes the product form $\mu = \prod_{a \in A} \pi_a$, and the law of large numbers yields

$$\lim_{n \to \infty} \frac{1}{|A_n|} \sum_{a \in A_n} z(x^a, p) = \int z(x^a, p) \mu(dx) \mu - \text{almost surely} \tag{1}$$

for an increasing sequence of finite populations $\{A_n\}_{n \in \mathbb{N}}$. Under standard conditions on $z(x^a, \cdot)$ there exists a unique price system p^* for which per capita excess demand is small in economies with many agents, that is, for which

$$\int z(x^a, p^*) \mu(dx^a) = 0. \tag{2}$$

The assumption of independence of states can be dropped as long as μ is *ergodic*, that is, as long as (1) holds. However, when preferences are interactive, the probabilities π_a specifying the dependence of the individual states on the states of others do not necessarily determine the *joint* distribution μ of all the states. This effect can best be illustrated by means of an 'Ising economy' where the agents are indexed by the two dimensional integer lattice $(A = \mathbb{Z}^2)$, the set of possible states is {–1, 1}, and where the conditional distribution of agent *a*'s state depends on all the other states x^{-a} only through the states x^b of his four nearest neighbours $b \in N(a) := \{\hat{a} \in A : |a - \hat{a}| = 1\}$. The distribution also depends on some constant $h \in \mathbb{R}$ which assigns an intrinsic value to private states and on a non-negative quantity *J* that measures the strength of social interactions. Specifically,

$$\pi_a(x^a; x^{-a}) = \frac{\exp\left\{x^a h + x^a \Sigma_{b \in N(a)} J x^b\right\}}{\exp\left\{x^a h + x^a \Sigma_{b \in N(a)} J x^b\right\} + \exp\left\{-x^a h - x^a \Sigma_{b \in N(a) = 1} J x^b\right\}}. \tag{3}$$

A probability measure μ on $S = \{x = (x^a)_{a \in A} : x^a \in \{-1, +1\}\}$ is called a *global phase* if its one-dimensional marginal distributions are consistent with the microscopic data given by the individual

characteristics $(\pi_a)_{a \in A}$, that is, if

$$\mu(x^a = \pm 1 | x^{-a}) = \pi_a(\pm 1; x^{-a}). \tag{4}$$

An ergodic phase μ can be *equilibrated* if there exist prices p^* for which (2) holds. For independent preferences ($J = 0$) global phases and, hence, equilibrium prices are always determined uniquely by the agents' characteristics. However, if the distribution of states depends only on the states of others ($h = 0$) and the interaction is sufficiently strong, that is, if J exceeds some critical value, two ergodic global phases μ_+ and μ_- exist. In this case aggregate behaviour cannot be inferred from looking at microscopic characteristics alone. Moreover, there is typically no price system that equilibrates both phases *simultaneously*. Thus, randomness in preferences becomes a source of uncertainty about market clearing prices.

Stochastic strategy revision in population games

The pioneering work by Blume (1993) puts Föllmer's model into a dynamic framework of interactive choice and exploits the link of discrete choice models with Gibbs distribution theory. It is mainly concerned with the aggregate behaviour in population games of bounded rational play, looking for 'Nash-like play in the aggregate rather than at the level of an individual player'. In Blume's model choice opportunities arise randomly according to individual players' Poisson 'alarm clocks'. When a choice opportunity arises for player $a \in A$ at time t, his choice x_t^a results in an instantaneous payoff $G(x_t^a, x_t^b)$ from each neighbor $b \in N(a)$ and in a total payoff

$$\sum_{b \in N(a)} G(x_t^a, x_t^b). \tag{5}$$

The conditional probability $\pi_a(x_t^a, x_t^{-a})$ with which player $a \in A$ selects an action x_t^a at time t, given the current states x_t^{-a} of all the other agents takes the form (3) with $h = \beta\hat{h}$ and $J = \beta\hat{J}$. Here $\beta \geq 0$ specifies the strength of interaction. For $\beta = 0$ the agents choose the actions with equal probability while a best response dynamics corresponds to the limiting case when β tends to infinity. The constants $\hat{h}$ and $\hat{J}$ are determined *endogenously* by the payoff matrix G through (5). Specifically,

$$\pi_a(x_t^a, x_t^{-a}) = \frac{\exp\left\{\beta\left[x_t^a \hat{h} + x_t^a \Sigma_{b \in N(a)} \hat{J} x_t^b\right]\right\}}{\exp\left\{\beta\left[x_t^a \hat{h} + x_t^a \Sigma_{b \in N(a)} \hat{J} x_t^b\right]\right\} + \exp\left\{-\beta\left[x_t^a \hat{h} + x_t^a \Sigma_{b \in N(a)} \hat{J} x_t^b\right]\right\}}.$$

These *flip rates* generate a continuous time Markov process *X* on *S* which describes the evolution of the agents' choices through time. A probability measure μ is called an *ergodic measure* for *X* if the distribution of choices does not change over time and empirical averages converge to a deterministic limit if the initial state is chosen according to μ. The process *X* is called *ergodic* if it has a unique ergodic measure. It is well known from the theory of interacting particle systems that the set of all ergodic probability measures for *X* is given by the ergodic global phases corresponding to the local specification (3). As a result, Blume's stochastic strategy revision process is ergodic if $\hat{h} \neq 0$ and $\hat{J} \geq 0$. This is the case if *G* describes a two person coordination game. Ergodicity breaks down for games with symmetric payoff matrices ($\hat{h} = 0$) when the interaction gets too strong. In this case, the long-run average choice depends on the starting point. The long-run macroscopic behaviour is as unpredictable as equilibrium prices in Föllmer's model by looking at microscopic characteristics only.

Non-ergodic economic growth

The evolution of individual choices in Blume (1993) is described by a *continuous* time Markov process with *asynchronous* updating. In local interaction models with *synchronous* updating, the dynamics of individual behaviour is typically described by a Markov chain whose transition operator takes the product form

$$\Pi(x_t; \cdot) = \prod_{a \in A} \pi_a\left(\cdot; \left\{x_t^b\right\}_{b \in N(a)}\right). \quad (6)$$

Thus, the distribution of the state x_{t+1}^a in period *t*+1 depends on the neighbours' states $\left\{x_t^b\right\}_{b \in N(a)}$ in period *t*. The long-run dynamics of such Markov chains plays an important role in macroeconomic models of economic growth.

The substantial differences in output levels and growth rates across countries have long been a major focus of macroeconomic research. A hallmark of the stochastic growth model pioneered by Brock and Mirman is the convergence of economies with identical preferences and production functions to a common level of aggregate output. Yet many analyses of long-run output movements have concluded that per capita production is not equalizing across countries. To explain this divergence, Durlauf (1993) studies a dynamic model of capital accumulation of an economy with an infinite set A of interacting

companies where local technological externalities affect the process of production. Each company $a \in A$ chooses a capital stock sequence $\{K_t^a\}_{t \in N}$ that maximizes the present value of future profits, and the technique-specific production functions generate output

$$Y_t^a = f(K_{t-1}^a, x_t^a, F(x_t^a)) \tag{7}$$

where $x_t^a \in \{0, 1\}$. Technique $x_t^a = 1$ is more productive, but comes at a higher fixed cost: $F(1) > F(0)$. Local technological complementarities affect the production as the distribution of x_t^a depends on the techniques implemented by the nearest neighbours $b \in N(a)$ in the previous period. The dynamics of production technologies is then described by an interactive Markov chain of the form (6). Assuming that past choices of technique 1 improve the current relative productivity of the technique and that the high-productivity state $x_t^a = 1$ for all $a \in A$ is an equilibrium, Durlauf (1993) shows that the high-productivity state is the only long-run outcome if the complementarities are weak enough: there exists $0 < \underline{\theta} < 1$ such that

$$\lim_{t \to \infty} P\left[x_t^a = 1 | x_0^a = 0\right] = 1 \text{ if } \pi_a\left(1; \left\{x_{t-1}^b\right\}_{b \in N(a)}\right) \geq \underline{\theta}.$$

Even when one starts with all low-production industries, an economy eventually coordinates on the high-production technology when negative feedbacks from low-production technologies are sufficiently weak. Powerful negative complementarities, on the other hand, can generate a non-ergodic growth path. In fact, there exists $0 < \overline{\theta} < \underline{\theta} < 1$, such that

$$\lim_{t \to \infty} P\left[x_t^a = 1 | x_0^a = 0\right] < 1 \text{ if } \pi_a\left(1; \left\{x_{t-1}^b\right\}_{b \in N(a)}\right) \leq \overline{\theta}.$$

If the complementarities are too strong, industries fail to coordinate on high-productivity equilibria, and economies may get trapped in low-productivity equilibria.

Models of social interaction – mean-field interaction

Much of the literature on social interactions assumes very special interaction structures such as nearest

neighbour interactions as in Blume (1993), or Durlauf (1993) or mean-field interaction. If agents care about the average behaviour throughout the whole population, the analysis is most naturally done in the context of an infinity of agents, as in Brock and Durlauf (2001). These authors analyse aggregate behavioural outcomes when individual utility exhibits social interaction effects. In the simplest setting agents take actions x^a from the binary action set {−1, +1} and their utilities consists of three components:

$$U^a(x^a, m^a, \varepsilon(x^a)) = u(x^a) + Jx^a m^a + \varepsilon(x^a), \tag{8}$$

Here m^a denotes agent a's expectation about the average choice of all the other agents. The second term in the utility function may thus be viewed as a social utility expressing an agent's desire for conformity ($J > 0$). The quantity $u(x^a)$, on the other hand, represents the private utility associated with a choice while $\varepsilon(x^a)$ is a random utility term independent of other agents' utilities and extreme-value distributed with parameter $\beta > 0$. The extreme-value distribution assumption for the random utility term yields conditional choice probabilities π_a of the form (3) if we replace the dependence of *actual* actions by a dependence on *expected* actions. When agents have homogeneous expectations about the behaviour of others ($m^a \equiv m$), then

$$\pi_a(x^a; m) = \frac{\exp\{\beta(u(x^a) + Jx^a m)\}}{\exp\{\beta(u(1) + Jm)\} + \exp\{\beta(u(-1) - Jm)\}}. \tag{9}$$

In the limit of an infinite economy all uncertainty about the average action vanishes because the agents' choices are conditionally independent given their expectations about aggregate behaviour. The average action is $\tanh(\beta h + \beta J m)$ where $h = \frac{1}{2}(u(1) - u(-1))$. If the agents have rational expectations the average satisfies the fixed point condition

$$m = \tanh(\beta h + \beta J m). \tag{10}$$

This equation has a unique solution if $h \neq 0$ and β is large enough. For large enough β the uniqueness property breaks down if $h = 0$, in which case (10) has three roots.

Models of social interaction – local and global interaction

When agents care about both the average action and the choices of neighbours, the equilibrium analysis becomes more involved. Horst and Scheinkman (2006) provide a general framework for analysing systems of social interactions with an infinite set of locally and globally interacting agents located on an integer lattice (for example, $A = Z^2$), continuous action spaces and random preferences. Specifically, they consider utility functions of the form

$$u^a(x, \theta^a) = U(x^a, \{x^b\}_{b \in N(a)}, \rho(x), \theta^a)$$

where $\rho(x)$ denotes the average choice associated with the action profile *x*, and the random variables θ^a specify the distribution of taste shocks. While the distinction between local and global interactions is unnecessary for models with finitely many agents, it is important for the analysis of infinite economies. The continuity of the utility functions $u^a(\cdot, \theta^a)$ in the product topology on the configuration space requires, implicitly, that the dependence of an agent's utility function on another agent's action decays sufficiently fast as the distance from that other agent grows. Thus, if preferences depend on average actions, utility functions are typically discontinuous. To overcome this problem, Horst and Scheinkman (2006) separated the local and global impact of an action profile $x = (x^a)_{a \in A}$ on individual preferences by viewing the average action as an additional parameter, ρ, of a *continuous* utility function on an *extended* state space. The parameter ρ can be seen as the agents' common expectation about the average behaviour. Under standard curvature conditions on *U* an equilibrium x^ρ exists for any such expectation ρ. If some form of spatial homogeneity prevails and under a weak interaction condition that restricts the influence of an agent's choice on the optimal decisions of others, x^ρ is unique. Furthermore, there exists a unique ρ that coincides with the average action $\rho(x^\rho)$ associated with x^ρ. In this case the agents correctly anticipate the average behaviour, and x^ρ turns out to be the unique equilibrium. The weak interaction condition also guarantees spatial ergodicity: the equilibrium of the infinite system is the limit of equilibria of finite systems when the number of agents grows to infinity; see Horst and Scheinkman (2005) for details.

Dynamic models of social interaction

When dynamic models of social interaction are studied the analysis is often confined to the case of backward-looking myopic dynamics, either as a simple explicit dynamic process with random sequential choice or as an equilibrium selection procedure. Rational expectations equilibria of economies with local interactions are studied in Bisin, Horst and Özgür (2006). While agents interact locally in these models, they are forward-looking. Their choices are optimally based on the past actions in their neighbourhood as well as on their anticipations of the future actions of their neighbours. The resulting population

dynamics can be described by an interactive Markov chain of the form (6) but the transition probabilities π_a are *endogenously* specified in terms of the agents' policy functions. Bisin, Horst and Özgür (2006) also allow for local and global interactions and combine spatial and temporal ergodicity results. The dynamics on the level of aggregate behaviour is deterministic (spatial ergodicity) and the distribution of individual choices settles down in the long run (temporal ergodicity) when the interaction is weak enough. The analysis, however, is confined to one-sided interactions. It is an open problem to fully embed the theory of social interactions into a dynamics analysis of equilibrium.

See Also

- agent-based models
- social interactions (theory)
- social interactions (empirics)
- social multipliers

Bibliography

Bisin, A., Horst, U. and Özgür, O. 2006. Rational expectations equilibria of economies with local interactions. *Journal of Economic Theory* 127, 74–116.

Blume, L. 1993. The statistical mechanics of strategic interactions. *Games and Economic Behavior* 5, 387–424.

Brock, W. and Durlauf, S. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68, 235–260.

Brock, W. and Mirman, L. 1972. Optimal growth under uncertainty: the discounted case. *Journal of Economic Theory* 4, 479–513.

Durlauf, S. 1993. Nonergodic economic growth. *Review of Economic Studies* 60, 349–66.

Föllmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1, 51–62.

Horst, U. and Scheinkman, J. 2005. A limit theorem for systems of social interactions. Working paper.

Horst, U. and Scheinkman, J. 2006. Equilibria in systems of social interactions. *Journal of Economic Theory* (forthcoming).

How to cite this article

Horst, Ulrich. "ergodicity and nonergodicity in economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000096>
doi:10.1057/9780230226203.0499

The New Palgrave Dictionary of Economics Online

Erlich, Alexander (1913 – 1985)

Diane Flaherty

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bukharin, N.I.; Erlich, A.; industrialization; Preobrazhensky, E.A.; socialism; Soviet Union, economics in

Article

Alexander Erlich was born in St Petersburg on 6 December 1913 and died on 7 January 1985. He moved to Poland with his family in 1918. In 1914, his father Henryk Erlich, a leader in the Socialist movement in Poland, was executed. In the same year, after university studies in Berlin and Warsaw, Erlich emigrated to the United States, where he earned a Ph.D. at the New School for Social Research and joined the faculty of Columbia University in 1955. From 1966 until his retirement in 1981 Erlich was professor of economics at Columbia, teaching in the economics department, the Russian Institute and the Institute for East Central Europe. Professor Erlich was revered by his students for his unstinting help and encouragement and respected by his colleagues for his breadth of knowledge and understanding of socialist economics.

Alexander Erlich's main contribution to the economics of socialism is his work on the critical issue of industrialization policy in the USSR in the 1920s. To this issue, Erlich brought an unusual blend of sophisticated economic reasoning and penetrating political analysis. His major thesis concerning Soviet policy in this period is that the structural disproportions in the Soviet economy were so deep that virtually any policy would have had negative side effects on reconstruction. Specifically, Erlich argued throughout his career that the economic policies of both the left *and* the right opposition were equally problematic. While the left analysis was correct in pointing out that future growth was limited after 1925 by the existing high-capacity utilization and scarce investment funds, Preobrazhenskii and others were wrong in underestimating the reaction of the peasantry to an industrialization policy that would squeeze peasant incomes. On the other hand, the right opposition did not appreciate the implications of high-capacity utilization for continued growth through small profit margins and high turnover of consumer goods and light manufacturers. The right, and Bukharin in particular, were seen by Erlich to be naive on the intensity of the conflict between consumption and investment once existing capacity was fully

utilized. This, his major work, exhibits a detailed knowledge of the Soviet experience and a dispassionate and rigorous analysis of policy choices that set the standard for such work in the field.

Selected works

1950. Preobrazhenskii and the economics of Soviet industrialization. *Quarterly Journal of Economics* 64 (February), 57–88.

1959. The Polish economy after October 1956: background and outlook. *American Economic Review, Papers and Proceedings* 49(May), 94–112.

1960. *Soviet Industrialization Debate, 1924–1928*. Cambridge, MA: Harvard University Press.

1967a. Development strategy and planning: the Soviet experience. In *National Economic Planning*, ed. M.F. Millikan. New York: Columbia University Press.

1967b. Notes on a Marxian model of capital accumulation. *American Economic Review, Papers and Proceedings* 57(May), 599–616.

1973. A Hamlet without the Prince of Denmark. *Politics and Society* 4(1), 35–53.

1977. Stalinism and Marxian growth models. In *Stalinism: Essays in Historical Interpretation*, ed. R.C. Tucker. New York: Norton.

1978. Dobb and the Marx–Fel'dman model: a problem in Soviet economic strategy. *Cambridge Journal of Economics* 2, 203–14.

Bibliography

Desai, P., ed. 1983. *Marxism, Central Planning and the Soviet Economy: Economic Essays in Honour of Alexander Erlich*. Cambridge, MA and London: MIT Press.

How to cite this article

Flaherty, Diane. "Erlich, Alexander (1913–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000114> doi:10.1057/9780230226203.0500

The New Palgrave Dictionary of Economics Online

estate and inheritance taxes

Wojciech Kopczuk
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article briefly describes features of real-life estate and inheritance taxes, economic arguments for and against these types of taxation and empirical evidence on economic distortions associated with such instruments.

Keywords

bequests; capital gains and losses; charitable contributions; estate taxation; gift taxation; inheritance taxation; progressive and regressive taxation; redistribution of income and wealth; tax avoidance; taxation of income; wealth accumulation

Article

Taxes imposed on intergenerational transfers are among the oldest types of taxation, apparently dating back at least to the Roman Empire (Pechman, 1987). There is substantial variation in their design in actual tax systems. The tax may be imposed on the donor or the donee side: it can apply either to the total estate (the total value of assets left by the decedent) or it can apply separately to transfers received by each beneficiary. This distinction matters when there are multiple beneficiaries and the tax is not simply proportional, or if the inheritance tax interacts with other forms of taxation (such as income tax). Further sources of variation in how these types of tax appear around the world include differences in how family members are treated, deductions allowed, treatment of certain categories of assets, treatment of capital gains and interaction with other types of tax. Two additional types of tax are closely associated with estate and inheritance taxation. First, some countries impose additional tax on transfers that skip generations. Such transfers would otherwise avoid taxation at death of an intermediate generation and hence would provide tax savings. Second, taxes on *inter vivo* gifts are imposed to protect the base of estate taxation (this is not their sole purpose, however: they also reduce the incentive for income shifting across individuals subject to different individual income tax brackets).

Most developed countries impose some form of taxation of intergenerational transfers; the exceptions are Canada, Australia and New Zealand. European countries usually impose inheritance taxes. See Gale and Slemrod (2001) for more details.

Estate taxation in the United States

In the United States, estate and gift taxes are ‘integrated’, that is, gifts over the lifetime influence computation of the estate tax burden at death. On top of federal taxation, many states impose their own taxes (in some cases inheritance rather than estate), although since the 1970s most states have changed their taxes to only ‘soak up’ federal credit for state taxation without imposing any incremental tax liability for those who are subject to the federal tax. The modern federal estate tax was introduced in 1916, although many states imposed their own taxes before that and the federal government made two earlier attempts to tax estates (during the Civil War and the Spanish–American War). The structure of estate taxation changed often before the Second World War, when the top marginal tax rates hit 77 per cent. Marginal tax rates were not reduced until the early 1980s, when the top rate was cut to 55 per cent. Further reductions are taking place as a part of the phase-out of estate tax initiated in 2001 that is supposed to culminate in a repeal in 2010. The repeal is a part of the set of provision that sunset in 2011, and hence the future of this tax is uncertain at the time of this writing. The US estate tax has always been characterized by a large tax exemption. At the peak in 1976, slightly over seven per cent of adult deaths corresponded to taxable estates, but, other than during the period of the growth in the reach of the tax in the 1960s and 1970s induced by ‘bracket creep’ (brackets not indexed for inflation), only two per cent or less of all estates were subject to the tax. Revenue collected by this tax has always been relatively small, constituting one to two per cent of total federal revenue after the Second World War.

Arguments for and against estate and inheritance taxation

A number of arguments are often given in favour of this type of taxation:

- Administrative convenience – taxation occurs at the time when assets have to be valued anyway, thereby reducing the burden of compliance relative to other forms of wealth taxation.
- Presumed lack of distortions if bequests are mostly ‘accidental’, that is, when taxpayers save for their own lifetime consumption rather than for bequests (see Kopczuk, 2003, for a critique of this argument).
- Redistribution (although, Kaplow, 2001 suggests that income taxation may be sufficient for redistribution).
- Backstop to avoidance of income taxes.
- Providing equality of opportunities and breaking down concentration of wealth.
- ‘Carnegie effect’ – inherited wealth is a ‘bad’ because it makes children unproductive members of society (see Holtz-Eakin, Joulfaian and Rosen, 1993, for supporting empirical evidence).
- Providing incentives for charity.

These are countered by the following:

- Distortions introduced by estate taxation.
- Theoretical arguments for zero capital taxation in the long run (recently challenged in the context of the estate tax by Farhi and Werning, 2007).
- Horizontal inequity due to unequal treatment of ‘savers’ and ‘spenders’ (see for example McCaffery, 1994).
- Easy tax avoidance.
- Gift externality (providing an argument for subsidizing transfers).

A broader overview of the normative issues can be found in Gale and Slemrod (2001) and Kaplow (2001).

Economic distortions

Among the types of economic distortions often discussed in this context are effects on saving, investment and labour supply, tax avoidance and damage that is potentially done to small (family) firms when the owner dies (see Brunetti, 2006, for weakly supporting evidence that survival of small businesses is affected by the presence of this tax; note, though, that small firms already enjoy significant preferences in the US tax code). A related argument involves forcing taxpayers to pursue ‘deathbed’ planning and implications of the tax for ‘widows and orphans’ (see Kopczuk, 2007). Because of the presence of a deduction for charitable contributions in the United States, an important topic is the effect of the estate tax on charitable contributions. Some of the more important empirical findings regarding US estate tax are discussed in what follows.

Estate tax avoidance is thought to be very easy. Cooper (1979) suggested that a motivated tax planner could easily reduce tax liability very significantly, if not altogether. Others have challenged this view: for example, Schmalbeck (2001) argues that most avoidance strategies involve losing control over assets. There is some evidence in support of both views. Anecdotal evidence of widespread estate tax avoidance is easy to obtain, and the existence of a large estate tax planning industry is a prima facie evidence that a lot of effort goes into such planning. At the same time, it has been established that some simple tax avoidance strategies are not pursued enough from the tax minimization point of view. McGarry (1999) and Poterba (2001) show that taxpayers do not take full advantage of annual gift tax exemption (annual gifts of less than $11,000 per donee are exempt from taxation). Kopczuk (2007) shows that significant adjustments take place following the onset of a terminal illness, thereby revealing that not enough planning took place earlier in life. Kopczuk and Slemrod (2003) argue that the widespread reliance by married decedents on the unlimited marital deduction implies that taxpayers do not take full advantage of tax savings from splitting an estate. This is so despite the existence of trust instruments that allow for separating tax planning from other considerations such as taking care of the surviving spouse. On the other hand, while gifts do not seem to be fully utilized as a tax-planning device, they are nevertheless responsive to tax considerations, as demonstrated by Bernheim, Lemke and Scholz (2004) and Joulfaian (2004).

A number of papers have focused on estimating the responsiveness of estates to tax rates. Kopczuk and

Slemrod (2001), Holtz-Eakin and Marples (2001) and Joulfaian (2006) all found small but positive elasticities implying that higher marginal tax rates lead to lower estate values. Kopczuk and Slemrod (2001) and Joulfaian (2006) rely on estate tax data and therefore cannot distinguish between tax avoidance and the effect on wealth accumulation. Holtz-Eakin and Marples (2001) use actual wealth, but their results are based on a relatively low-wealth sample and hence are hard to generalize from. Due to the nature of estate taxation, these studies are based on cross-section, repeated cross-section or time series, and hence the econometric assumptions that underlie them are strong.

The estate tax is a part of the tax code, and considering it in isolation is not appropriate. Poterba and Weisbenner (2001) find that over 50 per cent of the value of estates over $10 million are unrealized capital gains that would escape taxation at death due to step-up provisions (capital gains unrealized at the time of death are not subject to the capital gains tax, and the base for the recipient is stepped up to the current value of the asset). Auten and Joulfaian (2001) find that lower estate tax rates reduce capital gains realizations, and thus exacerbate the lock-in effect. The estate tax constitutes a backstop to this type of avoidance and a repeal of the tax would require a modification of the step-up rule. Bernheim (1987) questioned whether the estate tax raises any net revenue once its interaction with other taxes is taken into account.

The effect of estate taxes on charitable contributions is theoretically ambiguous due to offsetting income and substitution effects. Most studies find that higher marginal estate tax rates stimulate charitable giving, but the magnitude of the overall effect, accounting for both price and wealth effects, remains controversial: Joulfaian (2005) provides a recent overview of the empirical literature.

Other than dealing with tax avoidance (the issue that may be better handled by fixing the income tax), the strongest arguments in favour of the tax are based on its role in redistribution, breaking up concentration of wealth and providing equality of opportunities. Kopczuk and Saez (2004) use historical estate tax return data to provide estimates of wealth concentration over the course of the 20th century, and discuss the role that the estate tax might have played in shaping trends in concentration. Piketty and Saez (2007) document the contribution of the estate tax to overall progressivity. Understanding how estate taxation influences the distribution of wealth should be the top priority for anyone interested in an honest assessment of its value as a policy instrument.

See Also

- bequests and the life cycle model
- capital gains taxation
- excess burden of taxation
- inheritance and bequests
- optimal taxation
- redistribution of income and wealth
- tax compliance and tax evasion
- taxation of income
- taxation of wealth

Bibliography

Auten, G. and Joulfaian, D. 2001. Bequest taxes and capital gains realizations. *Journal of Public Economics* 81, 213–29.

Bernheim, B.D. 1987. Does the estate tax raise revenue? In *Tax Policy and the Economy*, vol. 1, ed. L.H. Summers. Chicago: NBER; Cambridge, MA: MIT Press.

Bernheim, B.D., Lemke, R. and Scholz, J.K. 2004. Do estate and gift taxes affect the timing of private transfers? *Journal of Public Economics* 88, 2617–34.

Brunetti, M. 2006. The estate tax and the demise of the family business. *Journal of Public Economics* 90, 1975–93.

Cooper, G. 1979. *A Voluntary Tax? New Perspectives on Sophisticated Tax Avoidance*. Washington, DC: Brookings Institution Press.

Farhi, E. and Werning, I. 2007. Progressive Estate Taxation. MIT Department of Economics. Mimeo. http://econ-www.mit.edu/faculty/download_pdf.php?id=1403.

Gale, W.G., Hines, J.R., Jr. and Slemrod, J., eds. 2001. *Rethinking Estate and Gift Taxation*. Washington, DC: Brookings Institution Press.

Gale, W.G. and Slemrod, J. 2001. Rethinking the estate and gift taxes: overview. In Gale, Hines and Slemrod (2001).

Holtz-Eakin, D. and Marples, D. 2001. Distortion costs of taxing wealth accumulation: income versus estate taxes. Working Paper No. 8261. Cambridge, MA: NBER.

Holtz-Eakin, D., Joulfaian, D. and Rosen, H.S. 1993. The Carnegie conjecture: some empirical evidence. *Quarterly Journal of Economics* 108, 413–35.

Joulfaian, D. 2004. Gift taxes and lifetime transfers: time series evidence. *Journal of Public Economics* 88, 1917–29.

Joulfaian, D. 2005. Estate taxes and charitable bequests: evidence from two tax regimes. Working Paper No. 92. Office of Tax Policy Analysis, US Department of the Treasury.

Joulfaian, D. 2006. The behavioral response of wealth accumulation to estate taxation: time series evidence. *National Tax Journal* 59, 253–68.

Kaplow, L. 2001. A framework for assessing estate and gift taxation. In Gale, Hines and Slemrod (2001).

Kopczuk, W. 2003. The trick is to live: is the estate tax social security for the rich? *Journal of Political Economy* 111, 1318–41.

Kopczuk, W. 2007. Bequest and tax planning: evidence from estate tax returns. *Quarterly Journal of Economics* 122(4).

Kopczuk, W. and Saez, E. 2004. Top wealth shares in the United States: evidence from estate tax returns. *National Tax Journal* 57, 445–88.

Kopczuk, W. and Slemrod, J. 2001. The impact of the estate tax on the wealth accumulation and avoidance behavior of donors. In Gale, Hines and Slemrod (2001).

Kopczuk, W. and Slemrod, J. 2003. Tax consequences on wealth accumulation and transfers of the rich. In *Death and Dollars: The Role of Gifts and Bequests in America*, ed. A.H. Munnell and A. Sunden. Washingon, DC: Brookings Institution Press.

McCaffery, E. 1994. The uneasy case for wealth transfer taxation. *Yale Law Journal* 104, 283–365.

McGarry, K. 1999. Inter vivos transfers and intended bequests. *Journal of Public Economics* 73, 321–51.

Pechman, J. 1987. Inheritance taxes. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Piketty, T. and Saez, E. 2007. How progressive is the U.S. federal tax system? An international and historical perspective. *Journal of Economic Perspectives* 21(1), 3–24.

Poterba, J.M. 2001. Estate and gift taxes and incentives for inter vivos giving in the US. *Journal of Public Economics* 79, 237–64.

Poterba, J.M. and Weisbenner, S. 2001. The distributional burden of taxing estates and unrealized capital gains at death. In Gale, Hines and Slemrod (2001).

Schmalbeck, R. 2001. Avoiding federal wealth transfer taxes. In Gale, Hines and Slemrod (2001).

How to cite this article

Kopczuk, Wojciech. "estate and inheritance taxes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.

dictionaryofeconomics.com/article?id=pde2008_E000296> doi:10.1057/9780230226203.0501

The New Palgrave Dictionary of Economics Online

ethics and economics

Marc Fleurbaey
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In recent decades, an important corpus of theories in normative economics (social choice theory, the theory of fair allocation, and inequality and poverty measurement in particular) has developed in which formal analytical tools of economic theory are mobilized in order to relate basic principles of social ethics to precise criteria for the evaluation of social states of affairs. The efficacy of arguments based on veil-of-ignorance devices has been questioned and the scope of impossibility theorems has been circumscribed, leaving the stage to a variety of constructive proposals in several fields of application (voting, resource allocation, public policy, social indicators).

Keywords

Arrow, K.; axiomatic bargaining; axiomatic theory of indices; bargaining; Bergson–Samuelson social welfare function; Borda rule; business ethics; capability sets; coalition formation; competitive equilibrium; consequentialism; consumer sovereignty; cooperative games; corporate social responsibility; cost–benefit analysis; cost–surplus sharing; dominance criteria; Dworkin, R.; egalitarian equivalence; egalitarianism; equality of capabilities; equality of opportunity; equality of resources; ethics and economics; fair allocation; Gini coefficient; Harsanyi, J.; hypothetical insurance (Dworkin); identity; impartial observer (J. Harsanyi); impartiality; impossibility theorem; income distribution; indexing problem; inequality (measurement); inequality aversion; interest aggregation; interpersonal comparisons of utility; intertemporal preferences; judgements vs interests; liberal paradoxes; libertarianism; Lorenz curve; marginal rate of substitution; maximin principle; multidimensional inequality; Nash, J.; New Welfare Economics; normative economics; objective advantage; ordinal non-comparable (ONC) preferences; original position (Rawls); other-regarding preferences; Pareto principle; positive economics; potential Pareto improvements; poverty measurement; preferences; Rawls, J.; responsibility; rights; risk aversion; Samuelson, P.; Sen, A.; Shapley, L.; social choice; social contract; social justice; social welfare functions; subjective utility; tastes; utility functions; utility possibility sets; value judgements; veil of ignorance (Rawls); well-being

Article

Economics, as a discipline connected to policy decisions, has always been involved in the analysis of social objectives and of their underlying ethical values.
More recently, an important corpus of theories has been developed in which the formal analytical tools of economic theory are mobilized in order to relate basic principles of social ethics to precise criteria for the evaluation of social states of affairs. This corpus is not yet fully unified, and is still replete with debates and open questions, but the outlook of the field is rapidly evolving.
Economics is also connected to individual ethics, as economic decisions by households and firms sometimes involve issues of morality and responsibility toward partners and co-traders. Experimental economics has revealed that altruism, fairness and reciprocity considerations are a key component of strategic interactions between individuals. Ethical issues for firms relate to ‘business ethics’ and ‘corporate social responsibility’.

Normative versus positive economics

‘Positive economics’ seeks to understand and explain economic mechanisms, whereas ‘normative economics’ deals with the assessment of policies or states of affairs. While this distinction is useful, one must not be misled into believing that a clear-cut separation is obtained in practice. Positive economics is naturally inclined to focus on ethically important social and economic issues. Moreover, forecasting the consequences of various policies belongs to its realm and, even though this is in principle a purely positive task, it is practically relevant when it is associated with normative conclusions. Meanwhile, normative economics contains many results which merely consist in a logical analysis of the content of given concepts or the relations between concepts. In such work the economist can largely put his own value judgements aside. As Samuelson (1947, p. 220) aptly wrote, ‘it is a legitimate exercise of economic analysis to examine the consequences of various value judgments, whether or not they are shared by the theorist’. The role of ethical judgements in economics has received valuable scrutiny in Sen (1987), Hausman and McPherson (2006) and Mongin (2006).

Wariness about value judgements has often led economists to shun issues of distribution and to focus on efficiency as encapsulated in the Pareto principle. In particular it is tempting to think that a Pareto improvement (that is, a new situation that everyone in the population prefers to the status quo) cannot be questioned since, by definition, everyone would approve it. But focusing on Pareto improvements does not protect from controversy. The most pressing reason is that there are important ethical values, especially regarding the distribution, which are not captured by the Pareto principle. As a consequence, there may exist a non-Pareto improving reform that is much better for social welfare than any Pareto improvement. In particular, in the presence of large inequalities, focusing on Pareto improvements implicitly amounts to condoning the status quo, as noted in Arrow (1951). A second reason is that people's immediate preferences, which are relied upon in applications of the Pareto principle, may not correctly represent people's interests. The difficult issue of defining individual interests is examined below.

The Pareto criterion, on the other hand, is very restrictive because it applies only when unanimity of preference is obtained. New Welfare Economics (surveyed in Chipman and Moore, 1978) and cost–benefit analysis have developed less restrictive criteria dealing with potential Pareto improvements (that is, reforms that would be Pareto improving if certain transfers from the gainers to the losers were performed in addition, even though they are not actually implemented). Such criteria have been condemned by many authors, from Arrow (1951) to Blackorby and Donaldson (1990), for being inconsistent (they can produce cyclic social preferences), and also unethical because they may approve reforms that seriously hurt the worst-off sub-populations (because the transfers that could compensate the losers may not be actually made). More sophisticated versions of cost–benefit analysis (Drèze and Stern, 1987) rely on consistent and distribution-sensitive social welfare functions.

The branches of normative economics

Four main branches of normative economics can be distinguished. The first – the theory of social choice, sparked off by Arrow (1951) with a provocative impossibility theorem – examines the properties of functions that define an ordering of a set of alternatives (policies, social states, candidates) on the basis of the ordinal non-comparable (ONC) preferences of the population over these alternatives. In economics individual ONC preferences, which can be retrieved from observable choices, are usually considered the natural informational basis. However, the theory of social choice has been extended (following Sen, 1970a) to cover the possibility of incorporating more information about individual utilities. This theory has achieved a thorough understanding of the properties of voting rules (surveyed in Brams and Fishburn, 2002; Pattanaik, 2002) and of the informational requirements about individual utilities of various social welfare functions (surveyed in d'Aspremont and Gevers, 2002; Bossert and Weymark, 2004). The second – the theory of fair allocation, initiated by Kolm (1972) – studies the allocation of resources among individuals with heterogeneous tastes and abilities, in terms of fairness criteria such as no-envy (no agent should prefer another's bundle), lower bounds (for example, no agent should prefer the equal-split allocation), solidarity (for example, no agent should be hurt by an increase in available resources), among others (see survey in Moulin and Thomson, 1997). The third – the theory of inequality and poverty measurement, originally anchored to the Gini coefficient and the Lorenz curve – focuses on income distributions and has developed after Kolm (1969), Atkinson (1970) and Sen (1973) into an axiomatic theory of indices, dominance criteria (that is, criteria ascertaining that a distribution dominates another for a family of indices), and, more recently, statistical tests to be performed on samples. Surveys of this field can be found in Atkinson and Bourguignon (2000) and Silber (1999). The fourth – the theory of axiomatic bargaining and cooperative games, initiated by Nash (1950) and Shapley (1953) – analyses how to find a fair compromise in utility possibility sets, under different assumptions about coalition formation. A synthesis on axiomatic bargaining is available in Thomson (1999).

Three other branches must be mentioned, which can be viewed perhaps as sub-branches of the main ones. Connected to social choice theory, Harsanyi's impartial observer argument (1953) and aggregation theorem (1955), offered in defence of utilitarianism, have generated an important literature and some debates (see Broome, 1991; Weymark, 1991; Roemer, 2002). Sen's (1970b) and Gibbard's (1974) liberal paradoxes have also triggered debates about how to formalize rights and incorporate them in the theory of social choice (see in particular Gaertner, Pattanaik and Suzumura, 1992; Arrow, Sen and Suzumura, 1997, ii, part IV). The theory of axiomatic cost and surplus sharing, which lies somewhere between cooperative games and fair allocation, studies the allocation of cost or surplus shares across individuals not as a function of their preferences but as a function of their actions (demands or contributions). It is surveyed in Moulin (2002).

The various branches have developed more or less independently but, if one puts aside the theory of cooperative games and the theory of cost-surplus sharing, they can all be formally described as seeking to rank alternatives of various sets X on the basis of the population's utility functions $U_1,\bullet\ldots,\bullet U_n$, and possibly other personal characteristics such as abilities and

needs. The theory of social choice usually considers only one given set *X* and in Arrow's initial version, retains information only about individual ONC preferences; the theory of fair allocation takes the *X*s to be sets of feasible allocations and usually seeks only to identify a subset of optimal allocations in each *X* on the basis of ONC preferences; the theory of inequality and poverty indices usually takes *X* to be the set of income distributions and focuses on a special aspect of distributions instead of a general notion of social welfare, although it sometimes establishes a link between special indices and social welfare functions; the theory of axiomatic bargaining usually ignores the structure of the sets *X* and directly examines the utility possibility sets $\{(U_1(x), \ldots, U_n(x)) | x \in X\}$. This formal similarity between the various branches is favourable to cross-fertilizing. One observes that the frontiers between these fields, which are largely due to the contingent circumstances of their creation, are progressively vanishing, to be replaced by more substantial differences in principles, such as whether ONC preferences provide morally sufficient information about individual well-being.

Cross-fertilizing with political philosophy is also an essential part of the history of the field. Rawls's (1971) theory of justice, borrowing many features from economics, has had a profound influence in return in at least three ways. It has rekindled interest for equality and the maximin principle among economists; it has popularized the idea that putting individuals behind a ‘veil of ignorance’ concerning their own circumstances, as in Harsanyi's impartial observer argument (with differences in assumptions and conclusions which have aroused controversies between these two authors and commentators), is a way to define justice; it has also provided an implicit justification of the theory of fair allocation by defining justice in terms of equality of resources (even if by resources he meant ‘primary goods’, that is, all-purpose goods rather than ordinary commodities), firmly rejecting interpersonal comparisons of utility across individuals with incommensurable preferences. Dworkin's (2000) theory of equality of resources makes a clear reference to the no-envy criterion and combines it with the veil of ignorance in a hypothetical insurance market in which individuals can insure against bad personal characteristics. Social policy should then, according to this theory, mimic the hypothetical insurance premiums and indemnities by suitable taxes and transfers. Sen's (1992) theory of capabilities proposes to shift the focus from resources to functionings, a general notion of individual achievement, and to seek equality of capability sets, that is, the sets of functionings that are accessible to individuals. Arneson (1989) and Cohen (1989) have also proposed to focus on opportunities rather than achievements on the ground that individuals should be viewed as responsible for seizing the opportunities that are offered to them. These recent theories of justice have generated an increased interest in normative economics for the notions of freedom and responsibility (see, for example, Roemer, 1998, and, for surveys, Fleurbaey and Maniquet, 2006; Peragine, 1999; Barberà, Bossert and Pattanaik, 2004). Among many other philosophical contributions that have been influential in normative economics, one must also mention Parfit's (1984) thought-provoking essay on utilitarianism, identity and population issues.

The measurement of individual well-being

With reference to the traditional social welfare function $W(U_1,\bullet\ldots,\bullet U_n)$, it is convenient to decompose the problem of defining a criterion for the evaluation of social states into two sub-problems, namely, the problem of assessing each individual situation and the problem of constructing a synthetic measure for the whole population. This decomposition, however, does not imply that the former is any less normative than the latter. The measurement of individual well-being is not just an empirical exercise and raises many ethical issues.

First, in such measurement one must consider whether one should take account of individuals’ political and social preferences or only of their tastes about their personal situation. It appears that these two kinds of preferences belong to different levels of social evaluation. Political and social preferences are relevant in the democratic debate about general principles, while personal tastes belong to the concrete evaluation of social situations. Mixing the two levels may yield absurd consequences. In particular, making the allocation of resources directly depend on the satisfaction of individual political preferences may produce grossly unfair distributions, for instance when a simple summation of utilities representing heterogeneous political preferences induces the altruist to transfer their resources to the malevolent. As Sen (1977) nicely put it, one must not confuse the aggregation of ‘judgments’ with the aggregation of ‘interests’. It is worth emphasizing that, concerning judgement aggregation, the contribution of normative economics is not limited to studying voting procedures with the goal of neutrally aggregating judgements, because it may play an active role in shaping those judgements. Indeed, by scrutinizing issues in *interest* aggregation it clarifies the substance of the debates and may help in the formation of personal *judgements* on these matters. It may thereby make a useful contribution in the deliberation process and the construction of a consensus.

When dealing with the aggregation of ‘interests’, personal tastes themselves are not necessarily appropriate as a basis for ranking social states, since they may be influenced by unjust social pressures and conditioning, or based on mistaken beliefs. It may then appear preferable to try to guess what individual tastes would be if formed in correct conditions. Practical procedures eliciting such authentic preferences have yet to be invented, but one may observe that ‘safety belt’ policies are commonplace and are usually justified by reference to people's well-considered interests.

A more radical questioning of the reference to personal tastes comes with the observation that, even in absence of conditioning or bad information, those with demanding tastes do not necessarily deserve more resources than those with more modest wishes. Should individuals not assume responsibility, in some cases, for their ‘expensive tastes’ (Dworkin, 2000)? The emergence of principles of freedom and responsibility in recent theories of justice has in fact revealed how important such considerations have always been in the selection of relevant dimensions of individual well-being. Even the standard principle of consumer sovereignty according to which every individual is the best judge of his own interests implies that the social criterion will let the allocation of personal resources be managed under the individual's sole responsibility and will at most cater to a synthetic measure of his satisfaction.

More importantly, if individuals are considered responsible for how they transform ordinal satisfaction into numerical utility, then social evaluation can disregard their utility functions and take care of their ONC preferences only. The focus of Arrovian social choice and of the theory of fair allocation on ONC preferences instead of utilities can hereby find a justification in terms of individual responsibility, in addition to more traditional arguments about the difficulty to compare subjective utility across individuals. Similarly, in Sen's theory of capabilities, the social criterion deals with capability sets and disregards what combinations of functionings individuals choose in those sets. The normative appreciation of what individuals should be held responsible for (or be left free to handle by themselves) and what they should not is a difficult domain of philosophical debating to which economists are not necessarily well equipped to contribute. But economic models are very convenient to examine the consequences of various choices in this matter and it is instructive to relate various policy choices to underlying attributions of responsibility to the target population (see, for example, Roemer, 1998).

The important divide between welfarist and non-welfarist approaches is largely connected to this issue. A welfarist approach retains subjective utility (interpreted in terms of happiness or in terms of satisfaction) as the ultimate metric of well-being. A non-welfarist approach will typically discount subjective utility and take account of objective achievements, resources, opportunities or rights, although Sen's theory, for instance, does retain utility as one relevant functioning among others. Critiques of welfarism invoke not only the (at least partial) responsibility of individuals for their subjective utility, but also the idea that there are some objective dimensions of achievement which matter independently of their effect on satisfaction. For instance, a physical disability may justify help even if the concerned individual has perfectly adapted to his situation in terms of subjective utility. Or granting basic rights and freedoms may be viewed as so essential to the constitution of a community of morally autonomous agents that they should be granted uniformly, independently of their potentially unequal effect on individuals' various utility functions. More generally, fairness in the allocation of resources typically involves non-welfarist concerns. For instance, the axiomatic theory of bargaining, as recalled above, disregards the economic allocations and focuses exclusively on utility possibility sets. This has counter-intuitive consequences. Consider problem 1 which consists in deciding with which probability Ann or Bob will win a ten-dollar prize, and problem 2 which also consists in deciding winning probabilities, except that if Ann wins she gets ten dollars whereas if Bob wins he only gets one dollar. Nash's bargaining solution (which maximizes the product of utility gains) selects the fifty–fifty solution for both problems, whereas it would seem more reasonable to give Bob a greater probability of winning in the second case. See Roemer (1996) for a detailed criticism of welfarism in axiomatic bargaining.

The welfarist–non-welfarist distinction, however, is mainly philosophical and the economist can always reinterpret the utility index U_i as an index of capability, opportunity or objective advantage instead of subjective utility, without changing much to the formal analysis of normative criteria. What is more important for economic analysis is whether the relevant data about individual situations consist in a numerical index permitting comparisons across individuals or in ONC preferences only (a third possibility, considered in the theory of multidimensional inequality, is when individual situations are described by vectors of numerical indices, with no synthetic index or ordering). In the first case, one has a kind of 'formal welfarism' and the standard framework of social welfare functions is readily available. In the second case (as well as the third), one may eventually be able to construct a comparable index, but such an index is not given a priori and its construction must be justified.

The indexing problem is considered a vexing issue for non-welfarist theories of justice such as Rawls's (involving an index of primary goods) or Sen's (involving an index of capabilities). Indeed, it appears that if this index is personalized so as to espouse each individual's preferences, it is then a utility representation of preferences and one is back into the welfarist framework. The alternative is to impose a uniform index to all individuals, but this is tantamount to adopting a special view of how to weight the various goods (or capabilities), that is, a dogmatic or perfectionist definition of the good life. At this point economic analysis is helpful because it shows that a non-welfarist approach can nonetheless respect individual preferences. Consider the simple case of an exchange market with identical prices at the various allocations under consideration. Then, on the assumption that individuals are free to choose their consumption in their budget set, indexing their well-being by the market value of their consumption is congruent with their preferences over consumption bundles, although it is certainly non-welfarist since across individuals there is no relation between utility and wealth. More generally, with each preference ordering one can always associate an index function which is ordinally equivalent to the welfarist measure of utility without coinciding with it, as noted in Roemer (1996). In conclusion, between the pure welfarist approach and the 'perfectionist' non-welfarist approach, there is room for 'Paretian' non-welfarist approaches which respect individual preferences. This distinction can be formally described as follows. In the problem of ranking alternatives on the basis of the profile $(U_1,\bullet\ldots,\bullet U_n)$, the welfarist approach relies on the utility values $(U_1(x),\bullet\ldots,\bullet U_n(x))$ at each alternative x; the perfectionist approach ignores individual preferences, imposes an index U^* and evaluates x in terms of $(U^*(x),\bullet\ldots,\bullet U^*(x))$; the 'Paretian' non-welfarist approach retains the ONC preferences represented by $U_1,\bullet\ldots,\bullet U_n$ and constructs an ordering of alternatives obeying the Pareto principle (additional examples are provided below).

Another issue for economic analysis is whether the social state must be described in terms of consequences or in terms of procedures. Most of normative economics is still largely consequentialist, but the growing focus on opportunities, rights and freedom of choice definitely enlarges the scope of analysis beyond narrow consequentialism. So far, the studies of rights and freedom have remained rather abstract, dealing with the general definition of rights, the foundations of a measure of individual freedom and the analysis of distributions of opportunity sets, but there is some interest for more concrete economic settings (on these various approaches, see, for example, Laslier et al., 1998; Pattanaik, Salles and Suzumura, 2004). Contractarian theories of justice, which analyse justice norms as being shaped by individuals' interests in mutual cooperation, also appeal to game theorists. For instance, Binmore (1994–8) relates various degrees of egalitarianism and libertarianism to different time horizons of social interaction, arguing that the latter prevails in the long run.

The definition of social criteria

When a suitable index of well-being U_1,•…,•U_n is given, as in the welfarist or the perfectionist approaches, the only problem that remains is to choose a social welfare function W from the menu offered by the theory of social choice. For instance, the sum-utilitarian function

$$W(U_1(x), \ldots, U_n(x)) = U_1(x) + \ldots + U_n(x)$$

displays no aversion to inequality in utilities, the maximin function

$$W(U_1(x), \ldots, U_n(x)) = \min\{U_1(x), \ldots, U_n(x)\}$$

has an infinite aversion, while the product (or Nash) function

$$W(U_1(x), \ldots, U_n(x)) = U_1(x) \ldots U_n(x)$$

is an example of an intermediate function. With a small or zero inequality aversion over utilities, as with the utilitarian function, priority is given to individuals with a high marginal utility, independently of their utility level, whereas with a high-inequality aversion, as with the maximin, priority is given to the worst-off (in terms of utility level), even if they have a low marginal utility.

In the choice of an appropriate degree of inequality aversion, it is often thought that the veil of ignorance provides a helpful guide. In the simple version of this device (Harsanyi, Rawls), the observer simply imagines that she could become any individual of the considered society. One may introduce variations about whether this implies taking on all the personal characteristics of each possible individual or simply some of them (his ability, his preferences). In more complex versions of the scheme (Dworkin's hypothetical insurance, or Rawls's original position under some interpretations), individuals may all be put behind the veil and be left free to bargain or make transactions. The attraction of the veil of ignorance comes from the obvious fact that it guarantees the impartiality of decisions. But impartiality is a very weak requirement and any symmetric social welfare function such as those listed above is impartial. The important issue in the choice of a social welfare function is not that it must be impartial, but how averse to inequality it should be. In this respect, the veil-of-ignorance device appears actually ill-suited. It links inequality aversion to the degree of risk aversion of the observer. A very risk-averse observer will come close to some maximin criterion, whereas a risk-neutral observer will adopt a utilitarian kind of rule. If she maximizes her expected utility, her decision will at any rate be structurally utilitarian in some metric. How this translates into a choice of social welfare function W depends on the specific form of the ignorance veil (that is, what personal characteristics are inherited by the observer when she becomes a particular individual). In Dworkin's hypothetical insurance, expected-utility maximizers will adopt insurance contracts that will produce utilitarian kinds of allocations. It is hard to find a reason why risk aversion about the possibility to becoming different persons should determine distributive judgements about actually existing individuals. One could as well imagine other devices, such as living all the lives of the population, one after the other, in a reincarnation process. Again, one does not see why intertemporal preferences over sequences of reincarnations should be especially relevant for distributive judgements over coexisting individuals. The attraction of the veil of ignorance may perhaps come from the mistaken belief that impartiality is all there is to social justice. But the theory of social choice clearly shows that impartiality is just a minimal requirement. The multiplicity of impartial observer devices (lottery, reincarnation…) proves that each of them is just one way among many others to achieve impartiality, and presumably none of them reaches a comprehensive view of justice (on these issues, see, for example, Roemer, 2002).

If one forgets the veil and genuinely thinks about inequality aversion in terms of fairness between existing individuals, one still faces a moral dilemma, since typically the social welfare functions that do not give full priority to the worst-off have the repugnant feature that a very small gain to many well-off individuals can always, if these individuals are sufficiently numerous, outweigh a large loss to a badly-off person. On the other hand, the maximin function always prefers giving a very small gain to the worst-off, no matter how costly this may be to all the other individuals. The way out of this dilemma has yet to be invented.

Aggregating preferences

Let us now turn to the case in which ONC preferences are the only relevant data about individual utilities. Arrow's impossibility theorem of social choice suggests that there is a conflict between the Pareto principle and impartiality in this context, in contrast to the context of the previous paragraph in which many social welfare functions were simultaneously increasing and symmetric in utilities. But this alleged conflict occurs only when one requires the social ranking of two alternatives to depend only on how individuals rank them, to the exclusion of any other alternative. This ‘independence’ requirement is very restrictive and precludes using information about individual preferences such as their marginal rate of substitution, how they compare their current bundle to natural benchmarks, and so on. The theory of fair allocation actually features many fairness criteria (such as no-envy or lower bounds) which violate this requirement by examining extended portions of indifference curves in order to evaluate an allocation.

Consider the following example. A certain quantity of divisible goods has to be distributed. At any allocation in which all individuals receive a personal bundle, evaluate every individual's bundle by the fraction of the social endowment that is equivalent according to this individual's preferences (see Figure 1).

Figure 1

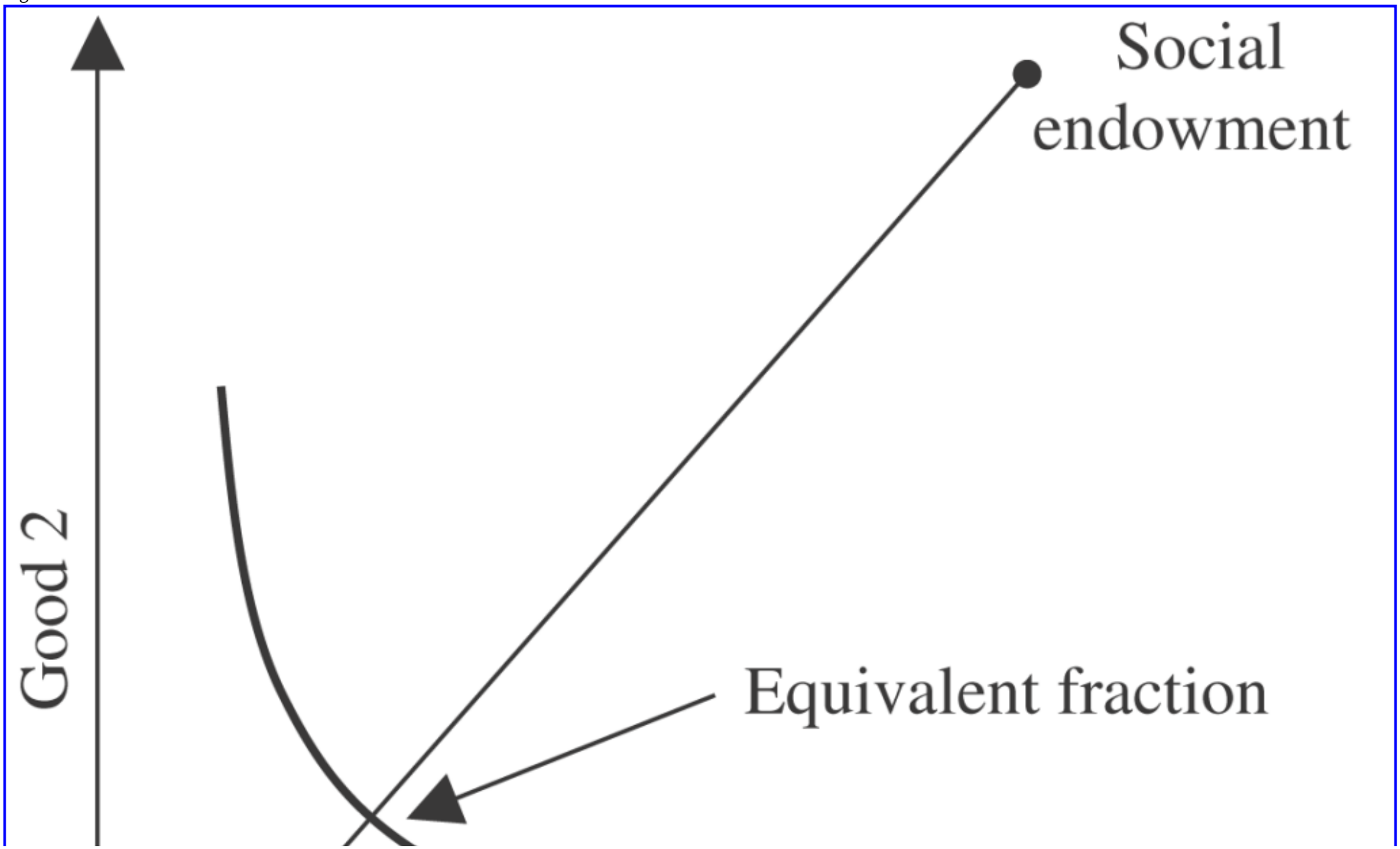

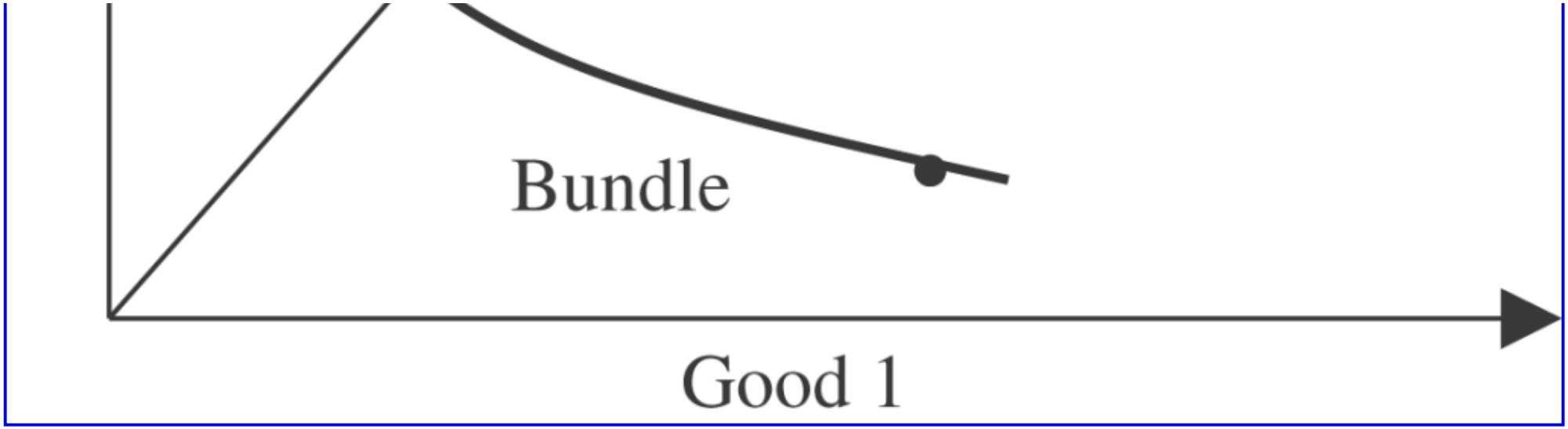

This is a standard way of representing individual preferences by an index function, and such indexes can then serve as arguments of a social welfare function *W*. This exemplifies a Paretian and impartial way to aggregate individual ONC preferences. This example has an interesting life of its own in the literature. It is briefly examined in Arrow (1951, p. 31), and rejected on the ground that it violates the above independence requirement. This observation, however, could as well suggest abandoning the requirement. A variant of the example is mentioned in Kolm (1969). It is invoked by Samuelson (1977) in order to show that it is possible to construct a Bergson–Samuelson social welfare function on the sole basis of ONC preferences. As explained in Samuelson (1947), such a function is a mapping

$$E(x) = W(U_1(x), \ldots, U_n(x))$$

where U_1,•…,•U_n are suitable indices representing individual preferences. (That the construction depends only on ONC preferences is verified by the fact that the same function *E* can be written with other, ordinally equivalent, indices V_1,•…,•V_n provided *W* is correspondingly adapted – some commentators have identified *W* instead of *E* as the fixed Bergson–Samuelson function, thereby concluding that Samuelson must have been wrong.) Eventually, it is used by Pazner and Schmeidler (1978) in the definition of the concept of egalitarian equivalence, which has become quite important in the theory of fair allocation (an allocation is egalitarian-equivalent if it is Pareto-equivalent to a resource egalitarian allocation). This example shows a simple way to aggregate preferences: first construct an index and then apply a standard social welfare function. The Borda rule, in the voting context, is another example of the same vein. The selection of the index need not be arbitrary and the above example, for instance, refers to fractions of the social endowment and thereby makes sure that individuals who prefer their bundle to the equal-split are always considered better-off than those in the opposite situation. There are less simple aggregation methods. Consider for instance the index

$$U_i^p(x_i) = e_i(U_i(x_i), p) - p\omega_i,$$

where x_i and ω_i denote *i*'s personal bundle and endowment, e_i his expenditure function, *p* a price vector. At every feasible allocation *x* (that is, such that $x_1 + \ldots + x_n \square \omega_1 + \ldots + \omega_n$) and for every price vector *p*, one has

$$U_1^p(x_1) + \ldots + U_n^p(x_n) \leq \sum_i (px_i - p\omega_i) \leq 0,$$

while a feasible allocation x^* is a competitive equilibrium associated to price vector p^* if and only if one has $U_i^{p^*}(x_i^*) = 0$ for all i. Therefore, for any inequality-averse social welfare function W, the function

$$E(x) = \max_{p \in S} W(U_1^p(x_1), \ldots, U_n^p(x_n)),$$

(where S is the simplex of appropriate dimension) exactly selects the competitive equilibria as the best allocations in the set of feasible allocations. This function, which bears some similarity to cost–benefit criteria without sharing their drawbacks, is slightly more complex than the previous examples as it makes the evaluation of individual situations depend on a price vector that itself depends on the whole allocation. Observe how even the maximin criterion, which is just a particular case of inequality-averse social welfare function, can rationalize a competitive equilibrium, no matter how unequal the endowments are. This is due to the deduction of the value of endowments in the index U_i^p, which is justified if individuals are held responsible for their endowments, so that one is not interested in individual total consumptions but only in the value difference between their consumption and their endowment.

Even though such constructions are based on ONC preferences, they always involve some kind of interpersonal comparison (of the relevant indexes) in order to determine who should be given priority in the allocation of resources (for a synthesis on interpersonal comparisons in general, see Fleurbaey and Hammond, 2004).

Hard issues

A few hard ethical issues have already been described. There are many others. Consider, for instance, Figure 2. It describes two individuals' utility under three different policies, depending on a random state. Policy B is better than A, since for the same *ex post* distribution of utilities it provides a less unequal distribution of prospects *ex ante*. And Policy C is better than B, since for the same distribution of *ex ante* prospects it guarantees an equal distribution of utilities *ex post*.

Figure 2

Policy A

	State (or effort) 1	State (or effort) 2
Individual 1	1	1
Individual 2	0	0

Policy B

	State (or effort) 1	State (or effort) 2
Individual 1	1	0
Individual 2	0	1

Policy C

	State (or effort) 1	State (or effort) 2
Individual 1	0	1
Individual 2	0	1

However, a social criterion that computes the expected value of social welfare will be indifferent between A and B, while a criterion satisfying the Pareto criterion with respect to expected utilities will be indifferent between B and C. Harsanyi's utilitarian criterion, which satisfies both properties, is indifferent between the three policies. The search for a better criterion in this context is still going on (see, for example, Ben Porath, Gilboa and Schmeidler, 1997).

A similar difficulty plagues the theory of equal opportunities, since, in spite of essential differences, there is an obvious similarity between random prospects and opportunities. In the same figure, replace the random states by effort levels exerted by the individuals. Then one can read the rows as depicting opportunity sets for the individuals. Under this new interpretation, Policy B is still better than Policy A because opportunity sets are less unequal, and Policy C is even better because it perfectly equalizes the opportunity sets. None of the social criteria in the literature displays this pattern of preference, because each of them focuses either on the distribution of *ex ante* opportunities or on the *ex post* neutralization of the effect of variables for which the agents are not responsible (this issue is discussed in Fleurbaey and Maniquet, 2006).

Another issue is the comparison of social welfare across different populations. The theory of social choice is curiously restricted, in its standard formulation, to the ranking of options

for a given population (with a specific ranking for each possible profile of preference of this population). But economic analysis is recurrently asked to compare standards of living across time (measurement of growth) or space (international comparisons). The framework of social choice should then be extended in order to rank not just allocations but pairs (allocation, population) involving populations with different preferences and different sizes. Interestingly, there are contexts in which size should be a neutral matter (for example, comparison of living standards between big and small countries) and other contexts (demographic policy) for which a theory of the optimal size is needed. Optimal demography is a famously hard domain. Classical utilitarianism, which is based on the total sum of utilities, has the unappealing feature that a population with arbitrarily low average welfare is always better, if it is sufficiently large, than any given population (Parfit, 1984). The criteria proposed by Blackorby, Bossert and Donaldson (2005) (see also Broome, 2004) avoid this 'repugnant conclusion' by computing individual welfare as the surplus $U_i - U^*$ over some positive threshold of utility U^* and then apply a social welfare function to these surpluses. The U^* threshold corresponds to the minimal welfare level that an individual must reach in order for his addition to society to be an improvement. Such criteria may thus induce judgements that a given population would be better off without its members whose utility is below the threshold, even when these members have positive utility. There is again a dilemma here.

This is just a sample of those hard ethical issues about social evaluation that economic analysis may never be able to render easy, but is able to clarify and to which it does, sometimes, give inventive solutions.

See Also

- Arrow, Kenneth Joseph
- bargaining
- Bergson, Abram
- cost-benefit analysis
- egalitarianism
- equality of opportunity
- fair allocation
- Harsanyi, John C.
- inequality (measurement)
- libertarianism
- Pareto principle
- philosophy and economics
- Samuelson, Paul Anthony
- Sen, Amartya
- social choice (new developments)
- social contract
- social welfare function
- voting paradoxes

Bibliography

Arneson, R. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56, 77–93.

Arrow, K. 1951. *Social Choice and Individual Values*, 2nd edn. New Haven: Yale University Press, 1963.

Arrow, K., Sen, A. and Suzumura, K., eds. 1997. *Social Choice Re–examined*, 2 vols. International Economic Association New York: St Martin's Press; London: Macmillan.

Arrow, K., Sen, A. and Suzumura, K., eds. 2002–2006. *Handbook of Social Choice and Welfare*, vol. 1 (2002) and vol. 2 (2006). Amsterdam: North-Holland.

d'Aspremont, C. and Gevers, L. 2002. Social welfare functionals and interpersonal comparability. In Arrow, Sen and Suzumura (2002).

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 244–63.

Atkinson, A. and Bourguignon, F., eds. 2000. *Handbook of Income Distribution*, vol. 1, Amsterdam: North-Holland.

Barberà, S., Bossert, W. and Pattanaik, P. 2004. Ranking sets of objects. In Barberà, Hammond and Seidl (2004).

Barberà, S., Hammond, P. and Seidl, C., eds. 2004. *Handbook of Utility Theory*, vol. 2, Dordrecht: Kluwer.

Ben Porath, E., Gilboa, I. and Schmeidler, D. 1997. On the measurement of inequality under uncertainty. *Journal of Economic Theory* 75, 194–204.

Binmore, K. 1994–8. *Game Theory and the Social Contract*. Vol. 1: *Playing Fair*. Vol. 2: *Just Playing*. Cambridge, MA: MIT Press.

Blackorby, C., Bossert, W. and Donaldson, D. 2005. *Population Issues in Social-Choice Theory, Welfare Economics and Ethics*. Cambridge: Cambridge University Press.

Blackorby, C. and Donaldson, D. 1990. Review article: The case against the use of the sum of compensating variations in cost–benefit analysis. *Canadian Journal of Economics* 23, 471–94.

Bossert, W. and Weymark, J. 2004. Utility in social choice. In Barberà, Hammond and Seidl (2004).

Brams, S. and Fishburn, P. 2002. Voting procedures. In Arrow, Sen and Suzumura (2002).

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty, and Time*. Oxford: Basil Blackwell.

Broome, J. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Chipman, J. and Moore, J. 1978. The New Welfare Economics, 1939–1974. *International Economic Review* 19, 547–84.

Cohen, G. 1989. On the currency of egalitarian justice. *Ethics* 99, 906–44.

Drèze, J. and Stern, N. 1987. The theory of cost–benefit analysis. In *Handbook of Public Economics*, vol. 2, ed. A. Auerbach and Feldstein. Amsterdam: North-Holland.

Dworkin, R. 2000. *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, MA: Harvard University Press.

Fleurbaey, M. and Hammond, P. 2004. Interpersonally comparable utility. In Barberà, Hammond and Seidl (2004).

Fleurbaey, M. and Maniquet, F. 2006. Compensation and responsibility. In Arrow, Sen and Suzumura (2006).

Gaertner, W., Pattanaik, P. and Suzumura, K. 1992. Individual rights revisited. *Economica* 59, 161–77.

Gibbard, A. 1974. A Pareto-consistent libertarian claim. *Journal of Economic Theory* 7, 388–410.

Harsanyi, J. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61, 434–35.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Hausman, D. and McPherson, M. 2006. *Economic Analysis, Moral Philosophy, and Public Policy* (2nd edn of *Economic Analysis and Moral Philosophy*). Cambridge: Cambridge University Press.

Kolm, S. 1969. The optimal production of social justice. In *Public Economics*, ed. J. Margolis and H. Guitton. London: Macmillan.

Kolm, S. 1972. *Justice et équité*. Paris: Ed. du CNRS. Translated as *Justice and Equity*, Cambridge, MA: MIT Press, 1997.

Laslier, J., Fleurbaey, M., Gravel, N. and Trannoy, A., eds. 1998. *Freedom in Economics: New Perspectives in Normative Analysis*. London: Routledge.

Mongin, P. 2006. Value judgments and value neutrality in economics. *Economica* 73, 257–86.

Moulin, H. 2002. Axiomatic cost and surplus sharing. In Arrow, Sen and Suzumura (2002).

Moulin, H. and Thomson, W. 1997. Axiomatic analysis of resource allocation problems. In Arrow, Sen and Suzumura (1997), vol. 1.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Pattanaik, P. 2002. Positional rules of collective decision-making. In Arrow, Sen and Suzumura (2002).

Pattanaik, P., Salles, M. and Suzumura, M., eds. 2004. Non welfaristic issues in normative economics. Special issue of *Social Choice and Welfare* 22(1).

Pazner, E. and Schmeidler, D. 1978. Egalitarian equivalent allocations: a new concept of economic equity. *Quarterly Journal of Economics* 92, 671–87.

Peragine, V. 1999. The distribution and redistribution of opportunity. J*ournal of Economic Surveys* 13, 37–69.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Roemer, J. 1996. *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.

Roemer, J. 1998. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Roemer, J. 2002. Egalitarianism against the veil of ignorance. *Journal of Philosophy* 99, 167–84.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P. 1977. Reaffirming the existence of ‘reasonable’ Bergson–Samuelson social welfare functions. *Economica* 44, 81–8.

Sen, A. 1970a. *Collective Choice and Social Welfare*. San Francisco: Holden Day.

Sen, A. 1970b. The impossibility of a Paretian liberal. *Journal of Political Economy* 78, 152–7.

Sen, A. 1973. *On Economic Inequality*. Oxford: Clarendon Press (expanded edn, Sen, A. and Foster, J., 1997).

Sen, A. 1977. Social choice theory: a re-examination. *Econometrica* 45, 53–90.

Sen, A. 1987. *On Ethics and Economics*. Oxford: Basil Blackwell.

Sen, A. 1992. *Inequality Re-examined*. Oxford: Clarendon Press.

Shapley, L. 1953. A value for N-person games. In *Contributions to the Theory of Games*, vol. 2, ed. H. Kuhn and A. Tucker. Princeton: Princeton University Press.

Silber, J. (ed.) 1999. *Handbook of Income Inequality Measurement*. Dordrecht: Kluwer.

Thomson, W. 1999. *Bargaining Theory: The Axiomatic Approach*. New York: Academic Press.

Weymark, J. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. Roemer. Cambridge: Cambridge University Press.

How to cite this article

Fleurbaey, Marc. "ethics and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000272> doi:10.1057/9780230226203.0502

The New Palgrave Dictionary of Economics Online

Eucken, Walter (1891 – 1950)

Josef Molsberger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic order vs. economic process; Eucken, W.; Freiburg School; German Historical School; laissez-faire; social market economy

Article

Head of the Freiburg School of German neo-liberalism and founder of the yearbook *Ordo*, Eucken was born at Jena on 17 January 1891. Eucken earned his doctoral degree at Bonn (1913). After the *Habilitation* in Berlin (1921), he was professor of economics at Tübingen (1925) and Freiburg (1927–50). He died on 20 March 1950 in London, during a lecture series at the London School of Economics. Eucken's works mark the return to (neo)classical theory in German economics after the dominance of the Historical School. He stressed, however, the theorist's task to explain reality and rejected model-building if it was purely an intellectual game. Eucken's outstanding analytical contributions include a masterly explanation of the German inflation and currency depreciation on quantity-theoretical grounds (1923), a capital theory (1934) building on Böhm-Bawerk and Wicksell and, in particular, his theory of economic systems (1940) and of economic policy (1952).

Eucken's theory of economic policy starts from the distinction between the *economic order*, the legal and institutional framework of economic activity, and the *economic process*, the daily transactions of economic agents. Under laissez-faire the state neither shapes the economic order nor intervenes in the economic process; in a centrally planned economy the state dominates both. Eucken conceived a *Wettbewerbsordnung* (competitive system) different from both systems: Government should abstain from directly intervening into market processes, but it has to shape the economic order by guaranteeing, through *Ordnungspolitik*, the ‘constituent principles’ of the market economy (monetary stabilization, free entry, private property, freedom of contract, liability, consistency in economic policy and, primarily, maintaining competition). Subsidiary are the ‘regulatory principles’: monopoly regulation, social policy, process stabilization policy. Eucken's theory laid the ground for West Germany's ‘social market economy’.

Selected works

1923. *Kritische Betrachtungen zum deutschen Geldproblem*. Jena: Gustav Fischer.

1932. Staatliche Strukturwandlungen und die Krise des Kapitalismus. *Weltwirtschaftliches Archiv* 36, 297–323.

1934. *Kapitaltheoretische Untersuchungen*. Jena: Gustav Fischer. 2nd enlarged edn, Tübingen: Mohr; Zürich, Polygraphischer Verlag, 1954.

1940. *Die Grundlagen der Nationalökonomie*. Jena: Gustav Fischer. 8th edn, Berlin, Heidelberg and New York: Springer, 1965. Trans. as *The Foundations of Economics*, London, 1950.

1948a. On the theory of the centrally administered economy: an analysis of the German experiment. *Economica* 15; Pt I, May, 79–100; Pt II, August, 173–93.

1948b. Das ordnungspolitische Problem. *Ordo* 1, 56–90.

1949. Die Wettbewerbsordnung und ihre Verwirklichung. *Ordo* 2, 1–99.

1951. *Unser Zeitalter der Misserfolge: 5 Vorträge zur Wirtschaftspolitick*. Tübingen: Mohr.

1952. *Grundsätze der Wirtschaftspolitik*, ed. E. Eucken and K.P. Hensel. Bern: Francke; Tübingen: Mohr. 5th edn, Tübingen: Mohr, 1975.

Bibliography

Böhm, F. 1950. Die Idee des Ordo im Denken Walter Euckens. *Ordo* 3, xv–1xiv.

Jöhr, W.A. 1950. Walter Euckens Lebenswerk. *Kyklos* 4(4), 257–78.

Lenel, H.O. 1975. Walter Euckens ordnungspolitische Konzeption, die wirtschaftspolitische Lehre in der Bundesrepublik und die Wettbewerbstheorie von Heute. *Ordo* 26, 22–78.

Lutz, F.A. 1961. Eucken, Walter. In *Handwörterbuch der Sozialwissenschaften*, vol. 3. Stuttgart: Fischer; Tübingen: Mohr; Göttingen: Vandenhoeck & Ruprecht.

Schmidtchen, D. 1984. German ‘Ordnungspolitik’ as institutional choice. *Zeitschrift für die gesamte Staatswissenschaft* 140(1), 54–70.

Stackelberg, H. von. 1940. Die Grundlagen der Nationalökonomie. *Weltwirtschaftliches Archiv* 51(2),

245–86.

Welter, E. 1965. Walter Eucken. In *Lebensbilder grosser Nationalökonomen: Einführung in die Geschichte der Politischen Öknonomie*, ed. H.C. Recktenwald. Cologne and Berlin: Kiepenheuer & Witsch.

How to cite this article

Molsberger, Josef. "Eucken, Walter (1891–1950)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000121> doi:10.1057/9780230226203.0503

The New Palgrave Dictionary of Economics Online

Euler equations

Jonathan A. Parker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An Euler equation is a difference or differential equation that is an intertemporal first-order condition for a dynamic choice problem. It describes the evolution of economic variables along an optimal path. It is a necessary but not sufficient condition for a candidate optimal path, and so is useful for partially characterizing the theoretical implications of a range of models for dynamic behaviour. In models with uncertainty, expectational Euler equations are conditions on moments, and thus directly provide a basis for testing models and estimating model parameters using observed dynamic behaviour.

Keywords

calculus of variations; continuous-time models; differential equations; discrete-time models; dynamic programming; Euler equations; expectations; generalized method of moments; Lagrange multipliers; liquidity constraints; optimal control; precautionary saving; Ramsey model; shadow pricing; uncertainty

Article

An Euler equation is an intertemporal version of a first-order condition characterizing an optimal choice as equating (expected) marginal costs and marginal benefits.
Many economic problems are dynamic optimization problems in which choices are linked over time, as for example a firm choosing investment over time subject to a convex cost of adjusting its capital stock, or a government deciding tax rates over time subject to an intertemporal budget constraint. Whatever solution approach one employs – the calculus of variations, optimal control theory or dynamic programming – part of the solution is typically an Euler equation stating that the optimal plan has the property that any marginal, temporary and feasible change in behaviour has marginal benefits equal to marginal costs in the present and future. On the assumption that the original problem satisfies certain regularity conditions, the Euler equation is a necessary but not sufficient condition for an optimum. This differential or difference equation is a law of motion for the economic variables of the model, and as

such is useful for (partially) characterizing the theoretical implications of the model for optimal dynamic behaviour. Further, in a model with uncertainty, the expectational Euler equation directly provides moment conditions that can be used both to test these theoretical implications using observed dynamic behaviour and to estimate the parameters of the model by choosing them so that these implications quantitatively match observed behaviour as closely as possible.

The term 'Euler equation' first appears in text-searchable JSTOR in Tintner (1937), but the equation to which the term refers is used earlier in economics, as for example (not by name) in the famous Ramsey (1928). The mathematics was developed by Bernoulli, Euler, Lagrange and others centuries ago jointly with the study of classical dynamics of physical objects; Euler wrote in the 1700s 'nothing at all takes place in the universe in which some rule of the maximum … does not appear' (Weitzman, 2003, p. 18).

The application of this mathematics in dynamic economics, with its central focus on optimization and equilibrium, is almost as universal. As in physics, Euler equations in economics are derived from optimization and describe dynamics, but in economics variables of interest are controlled by forward-looking agents, so that future contingencies typically have a central role in the equations and thus in the dynamics of these variables.

For general, formal derivations of Euler equations, see calculus of variations or dynamic programming.

This article illustrates by means of example the derivation of a discrete-time Euler equation and its interpretation. The article proceeds to discuss issues of existence, necessity, sufficiency, dynamics systems, binding constraints and continuous-time. Finally, the article discusses uncertainty and the natural estimation framework provided by the expectational Euler equation.

The Euler equation

Consider an infinitely-lived agent choosing a control variable (c) in each period (t) to maximize an intertemporal objective: $\sum_{t=1}^{\infty} \beta^{t-1} u(c_t)$ where $u(c_t)$ represents the flow payoff in t, $u' > 0$, $u'' > 0$, and β is the discount factor, $0 < \beta < 1$. The agent faces a present-value budget constraint:

$$\sum_{t=1}^{\infty} R^{1-t} c_t \leq W_1 \tag{1}$$

where R is the gross interest rate (R=1+r where r is the interest rate) and W_1 is given.

By the theory of the optimum, if a time-path of the control is optimal, a marginal increase in the control at any t, dc_t, must have benefits equal to the cost of the decrease in t+1 of the same present value amount, $-Rdc_t$:

$$\beta^{t-1}u'(c_t)dc_t - \beta^t u'(c_{t+1})Rdc_{t+1} = 0.$$

Reorganization gives the Euler equations

$$u'(c_t) = \beta R u'(c_{t+1}) \text{ for } t = 1, 2, 3\ldots$$
(2)

This set of Euler equations consists of nonlinear difference equations that characterize the evolution of the control along any optimal path. We considered a one-period deviation; several period deviations can be considered, but they follow from sequences of one-period deviations and so doing so does not provide additional information $(u'(c_t) = \beta^2 R^2 u'(c_{t+2}))$. These equations imply that the optimizing agent equalizes the present-value marginal flow benefit from the control across periods.
The canonical application of this problem is to a household or representative agent: call *c* consumption, *u* utility, and let $W_1 = \Sigma_{t=1}^{\infty} R^{1-t} y_t$, the present value of (exogenous) income, *y*. In this case, equations (2) imply the theoretical result that variations in income do not cause consumption to rise or fall over time. Instead, marginal utility grows or declines over time as $\beta R1$; for $\beta R=1$, consumption is constant.

Existence, necessity and sufficiency

In general, to ensure that the Euler equation characterizes the optimal path, one typically requires that the objective is finite (in this example, $u' > 0$) and that some feasible path exists.
Further, since Euler equations are first-order conditions, they are necessary but not sufficient conditions for an optimal dynamic path. Thus, theoretical results based only on Euler equations are applicable to a range of models. On the other hand, the equations provide an incomplete characterization of equilibria. In the example, only by using the budget constraint also can one solve for the time path of consumption; its level is determined by the present value of income.

Dynamic analysis

More generally, complete characterization of optimal behaviour uses the Euler equation as one equation in a system of equations. For example, replacing the budget constraint (eq. (1)) with the capital-accumulation equation

$$k_{t+1} = f(k_t) - c_t + (1-\delta)k_t$$
(3)

where k is capital, $f(k)$ is output, $f' > 0$, $f'' < 0$, $f(0)=0$, $\lim_{k \to 0} f_1 > \beta^{-1} - (1-\delta)$, and $\lim_{k \to \infty} f_1 < \beta^{-1} - (1-\delta)$, and adding the constraints k_1 given, $k_t \geq 0$, and $c_t \geq 0$, gives the basic Ramsey growth model. The constant real interest rate of eq. (2) is replaced by the marginal product of capital in the resulting Euler equation

$$u'(c_t) = \beta\left(1 - \delta + f'(k_{t+1})\right)u'(c_{t+1}). \tag{4}$$

Equations (3) and (4) form a system of two differential equations with two steady states that has been widely studied as a model of economic growth. Linearization shows that the interesting ($k > 0$) steady state is locally saddlepoint stable, and there is a unique feasible convergence path that pins down the dynamic path of consumption and capital.

Binding constraints

The above Euler equations are interior first-order conditions. When the economic problem includes additional constraints on choice, the resulting Euler equations have Lagrange multipliers. Consider adding a ‘liquidity constraint’ to our example: that the household maintain positive assets in every period s: $\Sigma_{t=1}^{s} R^{1-t} y_t - \Sigma_{t=1}^{s} R^{1-t} c_t \geq 0$ for all s. In this case, the program is more easily solved in a recursive formulation. Equation (2) holds with a single Lagrange multiplier, $\lambda_{t+1} \geq 0$, on the constraint that assets are positive in t+1 since prior to t+1 assets levels are unaffected by the choice of c_t and in period t+1 the present value of future consumption is unchanged by the one-period deviation considered:

$$u'(c_t) = \beta R u'(c_{t+1}) + \lambda_{t+1}.$$

The multiplier λ_{t+1} has the interpretation of a shadow price. When the constraint does not bind, $\lambda_{t+1} = 0$, the interior version of the Euler equation holds, and the marginal-benefit-marginal-cost interpretation is straightforward. When the constraint binds, the interpretation still holds, but almost tautologically: the change in utility of an extra marginal unit of consumption in t is equal to the change in utility from the marginal decreases in consumption in t+1 plus the shadow price (in terms of marginal

utility) of marginally relaxing the constraint on c_t. For example, if $\beta R=1$ and $y_t = \bar{y} \ \forall t \neq 2$ and $y_2 = y < \bar{y}$, then $\lambda_{t+1} = 0 \ \forall t \neq 2$, $\lambda_3 = u'\left(\frac{y+R\bar{y}}{1+R}\right) - u'(\bar{y}) > 0$, and $c_1 = c_2 = \frac{y+R\bar{y}}{1+R}$, $c_t = \bar{y} \ \forall t \geq 3$. This example illustrates that, relative to the unconstrained equilibrium $(c_t = \bar{y} - r(\bar{y} - y))$, the constraint can postpone consumption (t=1,2 relative to $t \geq 3$), create a causal link from an increase in income to consumption (t=2 to 3), and can lower consumption in unconstrained periods (t=1).

Continuous time

In general, continuous-time models have differential Euler equations that are equivalent to the difference-equation versions of their discrete-time counterparts. In the example, replacing t+1 with $t+\Delta t$, $c_{t+\Delta t} = c_t + \Delta t$, expanding $u'(c_t + \Delta c_t)$ around c_t, and letting $\Delta t \to 0$ gives:

$$\frac{\dot{c}_t}{c_t} = \sigma_t(r + 1 - \beta)$$

where $\sigma_t = -\frac{u'(c_t)}{c_t u''(c_t)}$. While the marginal-costs-marginal-benefit interpretation of the equation is less obvious in continuous time, it is still clear that consumption rises or falls with the difference between the interest rate (r) and the discount rate ($\beta - 1$), and more obvious that the strength of this response is governed by σ_t, which for this reason is called the elasticity of intertemporal substitution.

Generalized Euler equations

Dynamic games can also lead to ‘generalized’ Euler equations. For example, Harris and Laibson (2001) considers a modification of the example as a game among agents at different times who disagree because their preferences are not time consistent due to hyperbolic discounting. At any s, an agent has objective: $u(c_t) + \beta \sum_{r=1}^{\infty} \delta^t u(c_{s+r})$, where $0 < \delta < 1$. Defining recursively $W_{t+1} = R(W_t - c_t)$, the generalized Euler equation is

$$u'(c_t) = R \text{“}\textit{Effective}\text{ discount factor”} \underline{\left[\beta\delta\left(\frac{\partial c_{t+1}(W_{t+1})}{\partial W_{t+1}}\right) + \delta\left(1 - \frac{\partial c_{t+1}(W_{t+1})}{\partial W_{t+1}}\right)\right]} u'(c_{t+1}).$$

where $c_{t+1}(W_{t+1})$ is the optimal consumption choice made in t+1 as a function of W_{t+1}. The effective discount rate is a function of the (endogenous) marginal propensity to consume wealth in t+1.

Uncertainty

Models that contain uncertainty lead to expectational Euler equations. Add to the discrete-time example that the agent believes income y_s for $s>t$ to be stochastic from the perspective of period t. The Euler equation becomes

$$u'(c_t) = \beta R \hat{E}\left[u'(c_{t+1}) | I_t\right] \tag{5}$$

where $\hat{E}[\,.\,|I_t]$ represents the agent's expectation given information set I_t. The stochastic version of the consumption Euler equation has an analogous interpretation to that under certainty: the household equates *expected* (discounted) marginal utility over time.
Taking a second-order approximation to marginal utility in t+1 around c_t and re-organizing gives

$$\hat{E}\left[\frac{c_{t+1} - c_t}{c_t} | I_t\right] = \sigma_t\left(1 - (\beta R)^{-1}\right) + \frac{1}{2}\varphi_t E\left[(c_{t+1} - c_t)^2 | I_t\right]$$

where $\varphi_t = -\frac{c_t u'''(c_t)}{u''(c_t)}$ is the coefficient of relative prudence (see for example Dynan, 1993). It is now *expected* consumption growth that rises with the real interest rate and falls with impatience. Additionally, for $\varphi_t > 0$, risk leads to precautionary saving: higher expected consumption growth (much like liquidity constraints). Finally, actual consumption growth is also driven by the realization of uncertainty about current and future income.

Testing and estimation

An expectational Euler equation is a powerful tool for testing and estimating economic models in large samples, because, along with a model of expectations, it provides orthogonality conditions on which estimation can be based. Only randomization, as under experimental settings, delivers such a clean basis for estimation without near-complete specification of an economic model, including the sources of

uncertainty.

Considering our main example, define $\varepsilon_{t+1} = u'(c_{t+1}) - (\beta R)^{-1} u'(c_t)$. Hall (1978) pointed out that eq. (5) implies that $\hat{E}[\varepsilon_{t+1} z_t | I_t] = \hat{E}[\varepsilon_{t+1} | I_t] = 0$ for *any* z_t in the agent's information set, I_t. Under the assumption of rational expectations, mathematical expectations can be used in place of the agent's expectations. Thus, this equation predicts that observed changes in discounted marginal utility are unpredictable using I_t, or that marginal utility is a martingale, a strong theoretical prediction that Hall (1978) tests. Hansen and Singleton (1983) use a version of the stochastic Euler equation with a portfolio choice as the basis for estimation (and testing) of the parameters of the representative agent's parameterized utility function.

Following these papers (and others), large-sample testing and estimation of Euler equations under the assumption of rational expectations has played a central role in the evaluation of dynamic economic models. Most research applies the generalized method of moments (GMM) of Hansen (1982) using the restrictions on the moments of time series implied by the expectational Euler equation. Considering a $J \times 1$ vector of z_t's, $\mathbf{z}_t$, and, based on our example, define the column vector $\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t) = (\beta R u'(c_{t+1}) - u'(c_t))\mathbf{z}_t$, so that we have the J moment restrictions $E[\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t)] = 0_{J \times 1}$. For example, letting $u'(c_t) = c_t^{-(1/\sigma)}$ and assuming that second moments exist and the model is covariance stationary, the time-series average of $\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t)$ should converge to $E[\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t)]$ for the true σ, β and R. The GMM estimates of σ, β and R are those that minimize the difference (according to a given metric) between the observed empirical moments and their theoretical counterparts, $0_{J \times 1}$.

This general approach has the advantage that complete specification of the model is not necessary. In our example, the stochastic process for income need not be specified nor the stochastic process for consumption determined (which can be quite demanding in terms of computer programming and run-time). That said, more complete specification can give more theoretical restrictions and thus more power in asymptotic estimation. Gourinchas and Parker (2002), for example, uses numerical methods to bring more theoretical structure to bear in estimation. Further, more complete specification can allow one to use small-sample distribution theory and thus avoid the approximations inherent in using asymptotic distribution theory for inference in finite samples. A recent cautionary example is provided by the literature showing that standard asymptotic inference can be highly misleading in large samples with ‘weak instruments’.

See Also

- Bayesian methods in macroeconometrics
- calculus of variations
- dynamic programming
- generalized method of moments estimation
- instrumental variables

- liquidity constraints
- permanent-income hypothesis
- precautionary saving and precautionary wealth
- Ramsey model
- rational expectations models, estimation of

Bibliography

Dynan, K. 1993. How prudent are consumers? *Journal of Political Economy* 101, 1104–13.

Gourinchas, P.O. and Parker, J.A. 2002. Consumption over the life cycle. *Econometrica* 70, 47–89.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86, 971–87.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Hansen, L.P. and Singleton, J.K. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249–65.

Harris, C. and Laibson, D. 2001. Synamic choices of hyperbolic consumers. *Econometrica* 69(4), 935–58.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Tintner, G. 1937. Monopoly over time. *Econometrica* 5, 160–70.

Weitzman, M.L. 2003. *Income, Wealth and the Maximum Principle*. Cambridge, MA: Harvard University Press.

How to cite this article

Parker, Jonathan A. "Euler equations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000287> doi:10.1057/9780230226203.0504

The New Palgrave Dictionary of Economics Online

Euler's theorem

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adding-up problem; calculus of variations; Euler equations; Euler's Theorem; homogeneous functions; Lagrange multipliers

Article

Euler's Theorem on homogeneous functions is one of those useful pieces of multivariable calculus that has tended not to receive the attention in mathematical textbooks that its importance in economic theory warrants. An analogous case is Lagrange multipliers, though there the analysis in most textbooks falls far short of the rigour and depth that are needed for fruitful economic applications, as it often does of Euler's other discovery of direct importance in economics, the so-called Euler equations in the calculus of variations (for a critical discussion, see Young, 1969). With Euler's Theorem there are no such worries, however, and the discussion in a work like that of Courant (1936, vol. 2, pp. 108–10) is quite adequate.

Once the necessary notation and terminology is established, the statement of the theorem follows easily. Given any real number k, let F be a real-valued function defined on some non-empty subset S of vectors $x \in R^n$. Then F is said to be *homogeneous of degree k* (*h.d.k.*) if the equation

$$F(tx) = t^k F(x) \tag{1}$$

holds for every $x \in S$ and every real number t.

Let f be a differentiable function defined on a non-empty open subset $G \subset R^n$, and denote the *gradient* of f at x, that is, the n-dimensional vector of its partial derivatives f_i evaluated at x, by $\nabla f(x)$. The inner

product of any two vectors a and b is written $\langle a,\bullet b\rangle$.

Euler's Theorem: The differentiable function f is homogeneous of degree k if and only if the following Euler relation holds for every $x \in G$,

$$\langle x, \nabla f(x)\rangle = kf(x) \tag{2}$$

For a proof, see Courant (1936). Notice that this theorem *characterizes* homogeneous functions, that is, any function satisfying (2) for all x must satisfy (1), hence be *h.d.k.* A simple but often useful corollary of the theorem is that, if f is r-times differentiable and $m \leq r$, then each of its partial derivatives of order m is homogeneous of degree $k–m$, so that each f_i is *h.d.*(k–1), each f_{ij} is *h.d.*(k–2), and so on. Since homogeneous functions crop up almost everywhere in economics, Euler's Theorem is a standard tool with innumerable applications. So it is slightly odd that what was apparently its first use occurred so late, and that it was not by an established mathematical economist. In his review of Wicksteed (1894), A. W. Flux (1894) pointed out that Wicksteed could have saved himself a great deal of trouble if he had simply cited Euler's Theorem instead of, in essence, proving it all over again. It was indeed in the controversy over the so-called adding-up problem in the theory of distribution that Euler's Theorem first gained notoriety. For details of the adding-up problem, see Steedman (1987); here only a few of the main points will be lightly sketched in.

Assume that the firm wishes to minimize the cost of producing a scalar output η by the use of factors $x_1, x_2, \ldots, x_n = x$ bought at competitive prices $p_1, p_2, \ldots, p_n = p$. Under standard assumptions the first order conditions for this minimization yield

$$p_i = \lambda f_i(x) \tag{3}$$

where $i = 1, 2, \ldots, n$ and λ is the associated Lagrange multiplier. This multiplier is of course the marginal cost of output, a fact which can be guessed at from (3) on purely dimensional grounds alone. Assume now that the production function f, where $\eta = f(x)$, is homogeneous of some unknown degree k. Then, substituting from (3) into (2) and remembering the meaning of $\nabla f(x)$,

$$\lambda^{-1}\langle x, p\rangle = kf(x). \tag{4}$$

If there is competition in the product market as well, that is, free entry, then in long-run equilibrium marginal cost will equal the price q of the product, so that (4) becomes

$$[x, p] = kqf(x) . \quad (5)$$

The left-hand side of (5) is total factor payments. If constant returns prevail f is *h.d.*l and so k=1. All is well, since the right-hand side is then total revenue, equal to the sum of factor payments. If on the other hand k<1, so that returns to scale are decreasing, (5) shows that there will be something left over after all the purchased inputs have been paid.

What this residual really means is not clear. Some writers have interpreted it to be the returns (rent) to some non-marketed factor internal to the firm. But in that case why isn't the factor sold by its owner to the firm (after all, we are in long-run equilibrium, so that quasi-rents do not apply)? Or is there no external market for the factor?

This is not the place to go into such qsts, but it may be suggested that the incompleteness resides more in the theory than in either the markets or the factor payments. If k>1 there are increasing returns to scale, and (5) then suggests that there will not be enough revenue to meet total factor payments. But with increasing returns the hypothesis of perfect competition in the product market has to be abandoned, so the passage from (4) to (5) is illegitimate and (5) does not hold.

Bibliography

Courant, R. 1936. *Differential and Integral Calculus*, 2 vols. London: Blackie & Son.

Flux, A.W. 1894. Review of Wicksteed (1894). *Economic Journal* 4, 305–8.

Steedman, I. 1987. Adding-up problem. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Wicksteed, P.H. 1894. *An Essay on the Co-ordination of the Laws of Distribution*. London: Macmillan.

Young, L.C. 1969. *Lectures on the Calculus of Variations and Optimal Control Theory*. Philadelphia: W. B. Saunders Co.

How to cite this article

Newman, Peter. "Euler's theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of

Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000122> doi:10.1057/9780230226203.0505

The New Palgrave Dictionary of Economics Online

euro

Adam S. Posen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The euro has been a limited success at home, but it has not challenged the dollar as a global reserve currency. This reflects the limited impact of the euro on member economies. While European financial markets and trade are far more integrated than before adoption of the euro, that is the result of broader international trends. Factors inhibiting growth in the eurozone's real economy prevent truly deep financial integration as well, despite the removal of currency risk. The euro has delivered price stability and credible monetary policy, but not induced the convergence and reform necessary to improve European economic performance.

Keywords

currency unions; euro; European Central Bank; European Monetary Union; eurozone; Exchange Rate Mechanism (EU); exchange rate uncertainty; financial market integration; inflation; inflationary expectations; international capital flows; money market integration; portfolio investment; Stability and Growth Pact (EU)

Article

The launch of the euro, the European Union's currency (at least for 12 of the 27 current members), on 1 January 1999, was a birth long foretold. From at least the 1992 Maastricht Treaty onwards, its creation was at the forefront of the European overall integration agenda, and the meeting of criteria for eurozone entry dominated macroeconomic policymaking in Western Europe. The academic and policy discussion of European Monetary Union's (EMU) potential advantages and disadvantages began even earlier (see Canzoneri, Grilli and Masson, 1992; De Cecco and Giovannini, 1989; De Grauwe, 2000; Cecchini, 1988; as well as the seminal European Commission, 1990. Most of these studies concerned how best to make EMU work, taking the goal as a given, or assessing the optimality of the EU as a currency area.) New international reserve currencies, as the euro has begun to be, do not come along every day, or even every century. New currencies in general are launched usually out of need, due to replacement of a currency of hyperinflation-eroded value or to political fragmentation or secession; when currency unions are formed, they are usually done as pegs to a previously existing anchor currency of the largest and/or most stable member economy. The voluntary adoption of the euro by sovereign but not politically unified nations, and its replacement of already stable currencies (notably the Deutschmark), is thus an extraordinary monetary experiment and policy undertaking.

While the euro certainly has had no shortage of champions among economists – including beyond Euroland's borders the economists Bergsten (1997), Eichengreen (1999), Mundell (1998), and Portes and Alogoskoufis (1991) – many monetary economists observing the euro have tended to be sceptical: first of the virtues of the goal of monetary integration in Europe itself, then of the project's political viability, and then of its economic sustainability, in turn asserting that the euro was a solely political project. (Notable examples of this scepticism include, on the political side, Currie, Levine and Pearlman, 1992; Walters, 1990; and, famously, Feldstein, 1997; and on the economic side Arestis and Sawyer, 2001; De Grauwe, 1996; Dornbusch, 1989; Giavazzi and Spaventa, 1990; and Weber, 1991. See also the essays by eurosceptics in the face of mounting contrary evidence collected in *Cato Journal*, 2004.) Only as the euro passed its eighth birthday in wide usage, remained well past parity with the US dollar (see Figure 1) and experienced a strong cyclical recovery in the eurozone has sentiment changed. Increasingly, the question is being raised whether the euro might appreciate against the dollar for an extended period, be the beneficiary of substantial international portfolio adjustments, or even begin to supplant the dollar as the dominant global reserve currency. (Recent examples include Chinn and Frankel, (2004,

2007); Obstfeld and Rogoff, 2004; and Summers, 2004.) The euro's viability in its own large economic area may not be sufficient to set it on a path to monetary leadership, but its existence now presents an alternative for capital markets to turn to should the dollar's own appeal diminish.

Figure 1

Dollar-euro exchange rate and real GDP growth differential, 1999–2006. *Source*: IMF, IFS Statistics.

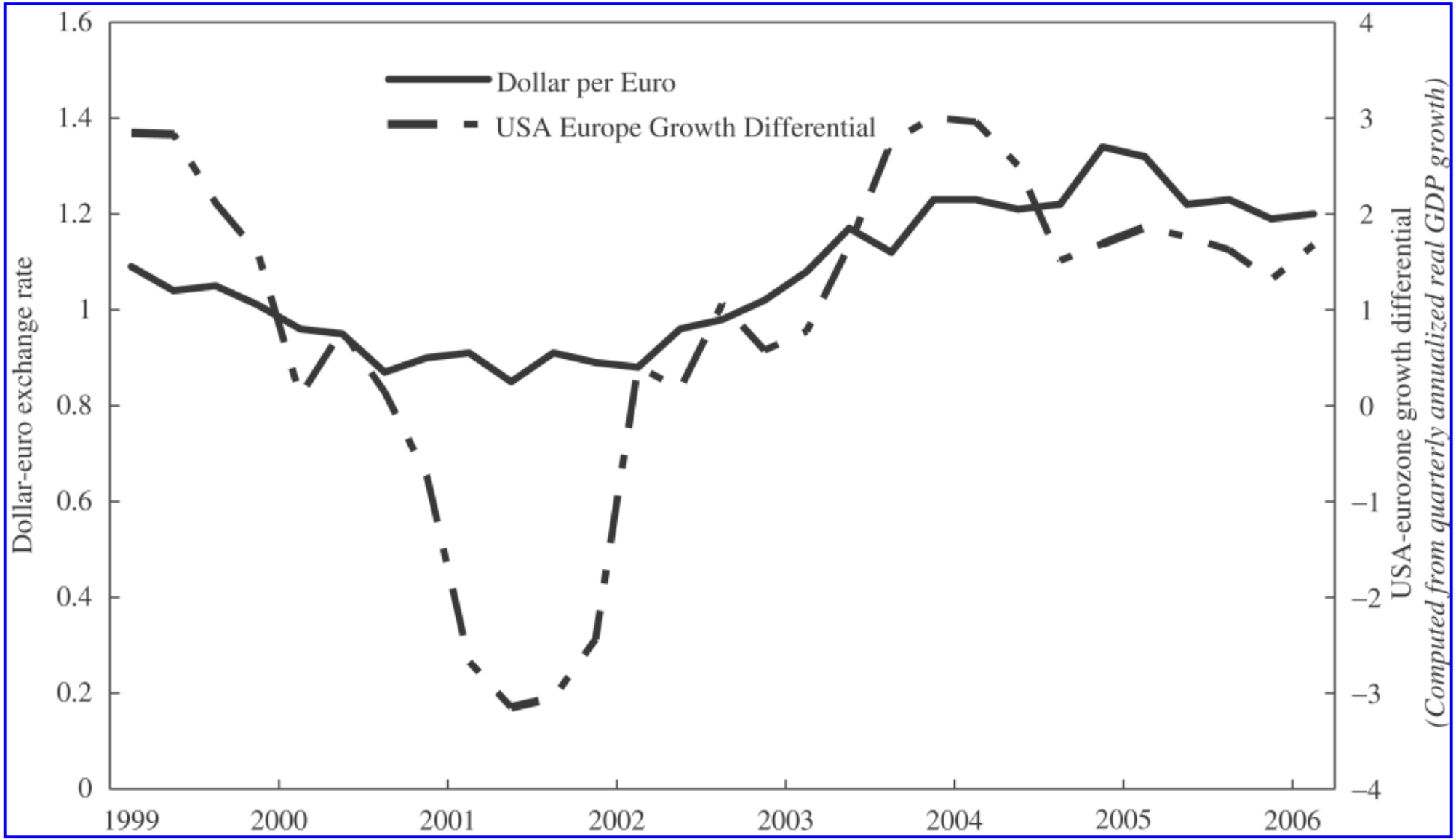

The waiting for US missteps for the euro to rise in importance, however, is a critical commentary on the limitations of the euro's importance to the eurozone member economies' performance in and of itself. When the euro was first proposed, a number of studies claimed that monetary integration would bring significant direct benefits to the economic performance of member states. Emerson, Gros and Italianer (1992) estimated that the elimination of transaction costs from moving to a single European currency would yield direct benefits of up to 0.4 per cent of EU GDP; the European Commission (1996) estimated cost savings of 1.0 per cent of GDP simply from eliminating transaction costs. The European Commission (1990) made the case that the reduction of nominal and real exchange rate uncertainty would lead to significant growth in intra-EU trade and investment. Financial markets in particular were expected to benefit from the introduction of the euro – McCauley and White (1997) and the European Commission (1997) forecast a rapid deepening and liquidity increase in European bond and lending markets, and perhaps even a 'decoupling' of European interest rates from those of the United States.

While empirical investigations to date of these effects remain mixed in interpretation, there is no question that the real economic effects of the euro's launch on the eurozone member countries have been something of a disappointment. In particular, European financial markets and trade integration are far deeper today than they were before the adoption of the euro, yet how much this represents the effect of the euro on EU integration as opposed to the broader international trends towards global integraton that benefited non-euro members as well is in doubt (see Forbes, 2005; Lane and Wälti, 2006; Mann and Meade, 2002; and Rey, 2005). The eurozone's interest rates remain asymmetrically affected by US interest rates, at least through the early 2000s, as established by Chinn and Frankel (2004). The effect of the euro on price convergence and on macroeconomic discipline cannot be all that substantial if on net there has been limited visible improvement in either of these areas (see the assessments of price convergence in Bradford and Lawrence (2003; 2004); and Rogers, 2003 and of macroeconomic discipline in Posen (2005a; 2005b). It seems that the euro has proven on net ‘irrelevant’ to real growth performance of large Continental European economies, neither a harm nor a boon to them, as Posen (1998) forecast it would be.

The external opportunities and shortfalls for the euro

The degree to which the euro comes into wider usage beyond intra-eurozone transactions, for example as an invoicing currency in world trade, is a major issue because of the eurozone's already large share of world output and trade (roughly comparable to that of the United States) and of the established ‘domestic’ monetary stability of the eurozone. Size does matter for international currency purposes. Yet insufficient integration and depth of European financial markets as well as lagging economic performance remain constraints on the euro's wider adoption and usage. Also important is the lack of coherent institutional representation for the eurozone in international monetary forums. Compared with the EU's one voice in global trade negotiations, the inability of the eurozone to speak as a single entity is striking, especially given the unconsolidated overrepresentation of the eurozone in the Bretton Woods institutions.

History also plays a role, however, in the global demand for currencies and their strength. Inertia and incumbency clearly contributed to the lingering of the British pound in a significant share of international reserves well after the Second World War. Yet the combination of macroeconomic mismanagement and growth underperformance in the United Kingdom from the 1920s to the 1980s eroded that role, and it is worth remembering that the passing of international monetary leadership from the pound to the dollar in the mid-20th century was in large part driven by these factors undermining the pound's reserve status. The steady accumulation of international debt by the United States since 1991 could contribute to a similar switch now that the euro is available. An extended dollar depreciation, the natural reaction to a multi-year series of widening US current account deficits, could induce a persistent portfolio diversification into euros by private and official holders of dollars.

In Washington, Frankfurt and Brussels, however, the widespread governmental opinion remains that the euro will not close the gap in usage with the dollar until the eurozone closes the gap with the US economy in per capita GDP growth and employment on a sustained basis (Figure 1 shows the growth differential of the USA over the eurozone). In a typical official expression of this sentiment, Quarles (2005, p. 40) finds that ‘too much attention is being focused on exchange rate[s] … and too little on what seems … of far greater importance: namely, the more effective functioning of economies with regards to growth in output and employment’. Successive US governments have viewed both the short-term international adjustment process and the longer-term role of the euro vis-à-vis the dollar as driven by the gap in growth rates between the USA and Europe – with the burden on European economies to catch up by raising their growth rates. EU officials’ disappointment with the degree of structural reforms catalysed by the introduction of the euro echoes this view, as does the promotion of the Lisbon Agenda announced in March 2000 for promotion of growth in the EU.

Such an external relative focus overlooks one achievement of the launching of the euro – ending the succession of devaluations, competitive depreciations and currency crises that had beset the members of the Exchange Rate Mechanism (ERM) prior to 1999. Certainly, the experiences of intra-European depreciations upon countries leaving the ERM, especially those of 1992–3, and their impact on economic performance and political outcomes in member states were in the forefront of European policymakers’ minds when the run-up to the euro was under way in the late 1990s. And, despite the divergence in histories of some eurozone members, inflation and inflation expectations have remained stable and low in the eurozone. That could have been expected to assist in trade promotion among the already interdependent eurozone economies (see currency unions).

Still there has been little or no expansion in trade as a result of the adoption of the euro – among other evidence, the share of total eurozone exports destined for other members of the eurozone did not increase with the introduction of the currency, as would have been likely if the common currency had promoted trade (Baldwin, 2005 provides an excellent analytical summary of the evidence on this score). As shown in Rogers (2003), the bulk of convergence in traded-goods prices within the eurozone occurred between 1990 and 1994, in response to the creation of the single market, and not after 1999 and the introduction of the euro. As for the global dimension, there has been little change in the share of foreign exchange transactions denominated in euros globally from that previously denominated in Deutschmarks. Similarly, the use of the euro as an invoicing currency is somewhat higher than that for the eurozone home currencies prior to EMU, but remains far from universal within Europe or even comparable to the dollar's usage (with the regional exception of some of the newest members of the European Union).

Even the spreading use of the euro in the EU's new members in the east has been far less than many might have expected. A critical part of this outcome has been the insistence on the part of the European Central Bank (ECB) that all prospective eurozone members go through the full Maastricht Treaty-specified process for qualification, including not just fiscal discipline and nominal convergence but also a two-year period in the ‘waiting room’ of a new ERM-II mechanism. Early, expedited or unilateral adoption of the euro in EU member

countries has in fact been discouraged by the ECB (with the exception of Estonia's pre-existing currency board with the euro). Arguably, this has as much to do with the ECB's desire for perceived control over monetary developments, given the ECB's Bundesbank-esque ‘two pillar’ strategy (of looking at both monetary growth and inflation goals when setting policy), and for keeping decision-making in the ESCB manageable, as with maintaining necessary discipline on eurozone members (see European Central Bank). The ECB has also been explicitly opposed to ‘euroization’ (dollarization with euros) by non-EU member countries, again partly for monetary control reasons, albeit acknowledging its contribution to stability in the post-conflict Balkan economies.

The limited impact of the euro on the eurozone financial integration and performance

The euro has delivered monetary stability in the face of a long list of economic shocks and a large initial decline against the dollar, only to rebound strongly since Autumn 2001 (see Figure 1). Europe has failed to follow the creation of the euro with the complementary policy reforms that were widely expected, however. This leaves an underlying tension between the constraints on national economic policy measures such as those in the Stability and Growth Pact on fiscal policy (see stability and growth pact) and the national frustrations with poor economic performance – a tension that raises recurrent doubts in eurosceptic financial markets about the sustainability of the euro itself, despite its lack of obvious vulnerabilities or viable exit options for any member country.

The euro was widely expected to transform two aspects of the eurozone economies: the integration and depth of their financial markets, and the conduct of their macroeconomic policies. Particularly with regard to the former, there has been beneficial change at least partly attributable to the euro's introduction and acceptance. Money market integration, which is critical to the implementation of a single monetary policy for the eurozone, given the need to transmit monetary policy in a decentralized fashion across the member economies, has succeeded. It took European money markets less than a month in 1999 to ‘learn’ how the new operational framework functioned, and to eliminate most of the volatility and cross-border dispersion in overnight interest rates. The evidence of integration in the unsecured lending rates in the European money market is similarly clear. Rey (2005) finds that government bond markets have seen intra-eurozone interest rate spreads virtually disappear, and benchmark securities of different countries have begun to emerge. Corporate bond markets went from ‘almost non-existent’ prior to EMU to 150 billion euro of issuance in 2003, and the euro swap market has become the largest financial market in the world. Eurozone financial markets, however, still have a long way to go to become a global competitor with those based in London or New York. Factors in the non-financial economy, such as legal differences, obstacles to more rapid real growth, transaction costs, and institutional gaps in financial supervision combine to keep the eurozone from achieving truly deep, integrated financial markets, despite the removal of currency risk. Thus, there remains a striking contrast between the repo (repurchase of safe assets at central banks) and unsecured market in the degree of cross-national differences in interest rates due to the ongoing lack of harmonization in legal and procedural treatment of financial instruments in the eurozone countries. The costs of making cross-border securities transfers within the eurozone can still be ten times more than the cost of securities transfers within a given eurozone country.

Given the surge in capital flows across borders worldwide, following the recovery from the 1997–98 Asian financial crisis, almost half of which were in the form of portfolio investment, one would expect greater influence of market opinion about assets in a given currency or region upon the actual allocation of capital between regions. It seems that prospects for economic growth drive the relative demand for a region's assets, mostly by determining where trade and investment expands, which then in turn sets the pace of stock market integration of that region with the rest of the world. Given the medium-term outlook for European growth, this appears to militate against an increase in investment and therefore in integration (and influence) of European capital markets, which might be partially offset by some diversification incentives. In the long run, though, a slow growth rate in Europe would also translate into a smaller share of global GDP, and less incentive for central banks to hold euro-denominated reserves. In this context, Forbes (2005) and Lane and Wälti (2006) independently investigate whether the euro's launch prompted greater co-movement of stock prices within the eurozone across national borders, indicating greater financial integration as a result of EMU. Both investigations find that stock market correlations of eurozone member markets with the United States increased after the introduction of the euro more than those between the eurozone countries.

Prospects for the euro

The euro therefore occupies something of a halfway house. In terms of its purely technical functions it has been a resounding success, with no problems in acceptance at home or abroad, or in the payments system, and there has been convergence in key eurozone money market interest rates. There has also been evidence of stable low-inflation expectations for the varied eurozone membership as a whole, which remains an outstanding achievement of European central banking. None of the broader forecasts of economic doom or internal political conflict predicted by (mostly American) Chicken Littles came to pass, and those predictions look less credible than they ever did. European financial markets have significantly deepened and added liquidity since the advent of the euro, particularly for fixed-income securities. The sheer size of the eurozone economy as well as the ongoing adjustment of the world economy to US current account deficits propel the euro towards a prominent global role.

At the same time, however, European relative economic performance and growth potential will continue to fall short of that of many other advanced economies and large emerging markets for the foreseeable future. The adoption of the euro and the associated convergence process have failed to induce, let alone produce, the needed transformation in European economic structures, policies and performance. In most scenarios, a collapse of the dollar in coming years, or even an ongoing orderly adjustment involving higher US long-term

interest rates and lower net imports, will have at least as great a contractionary effect on the eurozone as it will on the US economy – even if the Asian currencies take on their share of the adjustment burden. And if the Asian currencies, notably the Chinese yuan and Japanese yen, play their part, reserve switches accruing to euro-denominated securities, and their political benefits, will diminish along with the euro's share in the adjustment process. And as yet there has been little evidence of a change in global invoicing patterns from dollars to euros for traded good transactions.

In short, the euro has been a success within limits at home, but the eurozone economy is not yet strong enough – and is unlikely to be so for some time – to challenge the dollar as a global reserve currency or even to be widely utilized outside its borders. The euro, however, is not judged solely on its own merits, either by markets or by the international community, but rather is judged also in relative terms against developments in the dollar zone and elsewhere.

See Also

- currency unions
- European Central Bank
- European Monetary Union
- Stability and Growth Pact

Bibliography

Arestis, P. and Sawyer, M. 2001. Will the euro bring economic crisis to Europe? Economics Working Paper Archive: Washington University in St Louis.

Baldwin, R. 2005. The euro's trade effects. Presented at the ECB workshop: What effects is EMU having on the euro area and its member countries? Presented at the European Central Bank workshop. Frankfurt, 16 June.

Bergsten, F.C. 1997. The impact of the euro on exchange rates and international policy cooperation. In *EMU and the International Monetary System*, ed. M. Paul, T.H. Krueger and B. Turtelboom. Washington, DC: International Monetary Fund.

Bradford, S. and Lawrence, R.Z. 2004. *Has Globalization Gone Far Enough? The Costs of Fragmented International Markets*. Washington, DC: Institute for International Economics.

Canzoneri, M.B., Grilli, V. and Masson, P. 1992. *Establishing a Central Bank for Europe*. Cambridge: Cambridge University Press.

Cato Journal. 2004. *The Future of the Euro*, vol. 24 (Special Issue). Washington, DC: Cato Institute.

Cecchini, P. 1988. *Research on the 'Cost of Non-Europe': Basic Findings* (Cecchini Report). Luxembourg: EUR-OP.

Chinn, M.D. and Frankel, J.A. 2004. The euro area and world interest rates. Working Paper Series 1016. Center for International Economics, UC Santa Cruz.

Chinn, M.D. and Frankel, J.A. 2007. Will the euro eventually surpass the dollar as leading international reserve currency? In *G7 Current Account Imbalances: Sustainability and Adjustment*, ed. R. Clarida. Chicago: University of Chicago Press.

Currie, D., Levine, P. and Pearlman, J. 1992. European monetary union or hard EMS? *European Economic Review* 36, 1185–204.

De Cecco, M. and Giovannini, A. 1989. *A European Central Bank? Perspectives on Monetary Unification after Ten Years of the EMS*. Cambridge: Cambridge University Press.

De Grauwe, P. 1996. Monetary union and convergence economics. *European Economic Review* 40, 1091–101.

De Grauwe, P. 2000. *Economics of Monetary Union*. Oxford: Oxford University Press.

Dornbusch, R. 1989. The dollar in the 1990s: competitiveness and the challenges of new economic blocs. In *Monetary Policy in the 1990s: A Symposium*, sponsored by the Federal Reserve Bank of Kansas City, >Jackson Hole, WY.

Eichengreen, B. 1999. Will EMU work? In *Euroen og den norske kronens skjebne*, ed. A.J. Isachsen and O.B. Roste. Bergen: Fagbokforlaget. English version online. Available at http://emlab.berkeley.edu/users/eichengr/policy/merrill.pdf, accessed 25 February 2007.

Emerson, M., Gros, D. and Italianer, A. 1992. *One Market, One Money: An Evaluation of the Potential Benefits and Costs of Forming an Economic and Monetary Union*. Oxford: Oxford University Press.

European Commission. 1990. One market, one money. *European Economy* 44.

European Commission. 1996. *Economic Evaluation of the Internal Market*. European Economy: Reports and Studies No. 4. Brussels.

European Commission. 1997. External aspects of economic and monetary union. Euro Papers No. 1. Brussels.

Feldstein, M. 1997. The political economy of the European economic and monetary union: political sources of an economic liability. *Journal of Economic Perspectives* 11(4), 23–42.

Forbes, K. 2005. The euro and financial markets. In Posen (2005a).

Giavazzi, F. and Spaventa, L. 1990. The ‘new’ EMS. Discussion Paper No. 369, CEPR.

Lane, P. and Wälti, S. 2006. The euro and financial integration. Discussion paper. Dublin: Institute for International Integration Studies.

Mann, C.L. and Meade, E.E. 2002. Home bias, transaction costs and prospects for the euro: a more detailed analysis. Working paper. Washington: Institute for International Economics.

McCauley, R.N. and White, W.R. 1997. The euro and European financial markets. Working Paper No. 41. Basle: BIS.

Mundell, R. 1998. What the euro means for the dollar and the international monetary system. *Atlantic Economic Journal* 26, 227–37.

Obstfeld, M. and Rogoff, K. 2004. The unsustainable US current account position revisited. Working Paper No. 10869. Cambridge, MA: NBER.

Portes, R. and Alogoskoufis, G. 1991. International costs and benefits from EMU. *In The Economics of EMU. European Economy* 1 (Special Issue), 231–45.

Posen, A.S. 1998. Why EMU is irrelevant for the German economy. Working Paper No. 1998/11, Center for Financial Studies, University of Frankfurt.

Posen, A.S., ed. 2005a. *The Euro at Five: Ready for a Global Role?* Washington, DC: Institute for International Economics.

Posen, A.S. 2005b. Can Rubinomics work in the eurozone? In Posen (2005a).

Quarles, R. 2005. Discussion of the euro and the dollar. In Posen (2005a).

Rey, H. 2005. The euro and financial markets. In Posen (2005a).

Rogers, J.H. 2003. Monetary union, price level convergence, and inflation: how close is Europe to the United States? International Finance Discussion Papers No. 740. Washington, DC: Board of Governors of the Federal Reserve System.

Summers, L.H. 2004. The United States and the global adjustment process. Third Annual Stavros S. Niarchos Lecture. Washington, DC: Institute for International Economics. Online. Available at http://www.iie.com/publications/papers/paper.cfm?ResearchID=200, accessed 25 February 2007.

Walters, A. 1990. Monetary constitutions for Europe. Speech at the 28th meeting of the Mont Pèlerin Society, Munich.

Weber, A.A. 1991. EMU and asymmetries and adjustment problems in the EMS: some empirical evidence. *European Economy* 44, 187–207.

How to cite this article

Posen, Adam S. "euro." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000123> doi:10.1057/9780230226203.0506

The New Palgrave Dictionary of Economics Online

European Central Bank

Michael Binder and Volker Wieland

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The establishment of the European Central Bank (ECB) and with it the launch of the euro has arguably been a unique endeavour in economic history, representing an experiment of hitherto unknown magnitude in central banking. This article aims to describe the main aspects of the set-up and the responsibilities, strategy and operations of the ECB. It also aims to summarize some of the main lessons learned from the establishment of the ECB for monetary economics, and to sketch some of the prospects for the ECB and the euro.

Keywords

budget deficits; business cycles; Economic and Monetary Union (EMU); euro; European Central Bank (ECB); European System of Central Banks (ESCB); euro area; Exchange Rate Mechanism (ERM); globalization; Harmonized Index of Consumer Prices (HIPC); inflationary expectations; Maastricht criteria; monetary transmission mechanism; open market operations; price stability; reserve requirements; standing facilities

Article

The European Central Bank (ECB) was established on 1 June 1998 and since 1 January 1999 has been responsible for the conduct of a single monetary policy for its member countries, namely, Austria, Belgium, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Portugal and Spain (with Greece subsequently becoming a member country on 1 January 2001 and Slovenia on 1 January 2007). Among European Union (EU) member countries Bulgaria, Czech Republic, Denmark, Estonia, Cyprus, Latvia, Lithuania, Hungary, Malta, Poland, Romania, Slovakia, Sweden and the United Kingdom are, as of 2007, not member countries of the ECB. These countries for the time being either have opted out of becoming member countries of the ECB (Denmark, Sweden and the United Kingdom) or have – according to the judgement of the EU Council – not yet achieved the necessary degree of

economic convergence.

The launch of the ECB was the culmination of a process of monetary and economic integration that dates back at least to the efforts of the French government official Jean Monnet and others in the 1950s and gained decisive momentum with the April 1989 report of a committee headed by the then President of the European Commission, Jacques Delors, which drew up a blueprint for the progressive realization of the European Economic and Monetary Union (EMU). The establishment of the ECB and with it the launch of the euro (the currency of the ECB member countries for which banknotes and coins first went into circulation on 1 January 2002) has arguably been a unique endeavour in economic history, representing an experiment of hitherto unknown magnitude in central banking. In what follows, we shall describe the main aspects of the set-up and the responsibilities, strategy and operations of the ECB, discuss what appear to be the lessons learned from this experiment for monetary economics, and sketch some of the prospects for the ECB and the euro.

Lesson one: How to converge?

There can be little doubt that the European Council's June 1989 decision to pursue the Delors Committee's blueprint of a feasible path towards monetary union for its member countries was primarily driven by political considerations, viewing monetary union as a building block towards tighter political and economic integration of the member countries of the EU. However, given the broad consensus among economists and policymakers that, ideally, economic similarity rather than political boundaries should define the geographic area spanned by a common currency, the Delors report put considerable emphasis on realizing economic convergence before the establishment of a single European central bank. Key elements of the three stages to realization of the EMU as envisioned by the Delors Report were

- Stage 1 (1 July 1990): improvement of economic convergence; abolition of restrictions on cross-country flows of capital; increased cooperation between national central banks.
- Stage 2 (1 January 1994): strengthening of economic convergence; establishment of the European Monetary Institute (EMI) as predecessor of the ECB to strengthen cooperation between national central banks and increase coordination of monetary policy.
- Stage 3 (1 January 1999): completion of the necessary economic convergence; irrevocable fixing of currency conversion rates; single monetary policy to be conducted by the European System of Central Banks (ESCB).

It was envisioned in the Delors plan (and enacted in the Maastricht Treaty, which established the EU, as signed in February 1992) that only those countries should become member countries of the EMU that were successful in accomplishing economic convergence. The convergence criteria (Maastricht criteria) were meant to specify a sufficient degree of economic similarity of member countries with respect to price stability, sustainability of fiscal policy, exchange rate stability and the level of long-term interest rates. In particular, with respect to price stability member countries' average rate of inflation in the year preceding completion of the EMU was to fall within a one and a half per cent interval of average inflation in the three member countries displaying the highest degree of price stability. With respect to sustainability of fiscal policy, member countries were supposed not to carry an 'excessive deficit' –

which would occur if the actual or planned government deficit to GDP ratio exceeded three per cent or if the ratio of government debt to GDP exceeded 60 per cent. Concerning exchange rate stability, member countries would in the two years preceding completion of the EMU have to keep the fluctuations of the value of their currency within the bands provided for by the European Exchange Rate Mechanism (ERM) and in particular not initiate any devaluation of their currency against that of any other member countries. Finally, with respect to the level of long-term interest rates, member countries' average long-term interest rates (on government bonds or comparable securities) in the year preceding completion of the EMU were to fall within a two per cent interval of average long-term interest rates in the three member countries displaying the highest degree of price stability.

Of course, economic similarities desirable for an optimal currency area do not end with these four criteria, but inter alia also include similarities in the monetary transmission mechanism, the coherence of the shocks and of the propagation mechanisms driving national business cycles as well as similarities in the prospects for trend output growth. These latter criteria were not part of the Maastricht criteria, though it was widely hoped that the economic convergence process prior to or immediately after the formation of the ECB would result in these latter criteria being approximately met as well.

Despite the relatively modest requirements for economic convergence in the Maastricht Treaty, the goal of EMU was jeopardized during the 1992–3 crisis of the ERM when foreign exchange market participants widely viewed the ERM's margins of fluctuation of two and a quarter per cent as not sustainable in the light of at best limited coordination of monetary policy, especially in Germany, with that in several other countries in the EU, specifically that in Italy and in the United Kingdom. The fact that despite the widening of the ERM's margins of fluctuation to 15 per cent in August 1993 the goal of EMU was maintained appears to have been due to the commitment of some of the then political leaders of the EU – perhaps most notably the then German Chancellor Helmut Kohl – who saw their vision of building a united Europe jeopardized. Owing to this political commitment as well as the fact that markets increasingly gave weight to complying with the Maastricht criteria as a signal for sound monetary and fiscal policy, convergence as outlined by the Maastricht criteria was sufficiently advanced in May 1998 for the heads of state and government of the EU to decide to proceed with Stage 3 of EMU as planned, if only for the 11 initial member countries of the ECB.

While it is a valuable lesson to have observed in the context of the establishment of the ECB that the prospect of a monetary union may itself help to induce partial economic convergence, it appears key to keep in mind that the process of formation of the ECB would probably not have been successful without the strong desire of the member countries' political leadership to see commonalities in cultural heritage also reflected in increasingly cohesive institutional entities, trusting that a common European currency would help the emergence of a single European identity.

Structural economic diversities between euro area member countries continue today (in 2007). Among these diversities perhaps most notable are persistent differences in trend output growth rates. The widely voiced hope expressed at the time of the signing of the Maastricht Treaty – that formation of the ECB would significantly spur convergence of trend output growth rates for euro area member countries through alignment of structural reforms of labour and product markets – has so far proven to be wishful thinking. While some critics of the ECB have argued that this is due to the mandate of the ECB being too narrowly focused on price stability, it may have been exactly this focus that allowed the ECB to successfully establish itself as a credible safeguard of price stability, an issue which we will discuss further below.

Lesson two: How to design and implement a monetary policy strategy

The starting point for any discussion of the ECB's monetary policy strategy has to be the mandate that the ECB was given by the Maastricht Treaty. Article 105 of that treaty specifies: ‘The primary objective of the ESCB is to maintain price stability. Without prejudice to the objective of price stability the ECB shall support the general economic policies in the Community with a view to contributing to the achievement of the objectives of the Community as laid down in Article 2.’ Article 2 specifies these objectives to be a high level of employment as well as sustainable and non-inflationary growth. (The Maastricht Treaty refers to the ESCB rather than the ECB since it envisioned that all member countries of the European Union would eventually adopt the euro and that even before this was to happen all national central banks of member countries not part of the euro area would be bound by the same objectives.)

While the Maastricht Treaty does not specify a precise quantitative definition of price stability, the ECB, particularly on the basis of the argument that such quantification would strengthen its commitment to its primary objective as well as strengthen its accountability, in October 1998 defined price stability as a year-on-year increase in the Harmonized Index of Consumer Prices (HICP) for the euro area of below two per cent over the medium run. While this definition of price stability does exclude deflation as being consistent with price stability and leaves the ECB with no degree of freedom to potentially remove more volatile and/or temporary components of overall consumer prices in order to declare price stability, the definition does leave the ECB some flexibility in that a time horizon as to what would constitute the medium run was not established.

In its pursuit of price stability, the ECB decided to base its monetary policy framework on two pillars: ‘monetary analysis’ and ‘economic analysis’. In declaring monetary aggregates as providing information valuable to the objective of price stability that should be separated from other economic and financial variables, the ECB has so far maintained that monetary aggregates do not just offer incremental information relative to such other variables for purposes of projecting inflation, but that at longer horizons (stretching beyond those typically adopted by central banks for the computation of their inflation projections but still essential for medium-run price stability) monetary aggregates provide information qualitatively different from that which other economic variables can provide. The ECB in this context has so far also maintained that money demand (as measured by the monetary aggregate M3) for the euro area has been stable at least over longer horizons, with some short-run instabilities being due to an exceptionally prolonged (but still temporary) period of high asset price volatility. Finally, the ECB has so far maintained that conventional macroeconomic analysis is not sufficiently advanced to combine the analysis of real economic phenomena with monetary trends within a single pillar framework. Driven by these considerations, the ECB therefore initially decided to announce annual reference values for the growth rate of M3 as a benchmark for keeping monetary growth in line with the objective of price stability.

The ‘economic analysis’ pillar of the ECB's monetary policy framework aims at identifying and quantifying short- to medium-term non-monetary risks to price stability. Variables entering this analysis include (*a*) gap measures of the discrepancy between actual output as well as its factors of production on the one hand and their medium- to long-run equilibrium values on the other hand; (*b*) labour cost measures; (*c*) exchange rates for the euro and international prices; and (*d*) asset prices other than

exchange rates, particularly yield curve measures. Reflecting the sizeable degree of persistence of consumer price inflation in the euro area, considerable weight in the economic analysis is also given to recent consumer price dynamics.

The ECB's two-pillar strategy has been heavily criticized and remains controversial. Critics argue that monetary aggregates such as M3 – specifically due to the lack of sufficient stability of money demand – lack the degree of reliability needed to separate information in such monetary aggregates from other economic and financial variables. These critics inter alia also argue that, if the transparency and accountability of the ECB's decisions were to be improved, this would be helped most by the publication of inflation forecasts by the ECB as well as the publication of the minutes of the meetings of the ECB's Governing Council (for more on the latter, see below). The two-pillar strategy was reaffirmed in a broad internal assessment by the ECB in 2003, but two clarifications were provided. First, the Governing Council noted that it aims to maintain inflation rates below, but close to, two per cent over the medium run. A number of arguments in favour of tolerating a low rate of inflation – and not aiming at zero inflation – were acknowledged, among which the most important are the need for a safety margin against potential risks of deflation and the ‘zero bound’ on nominal interest rates. While this ‘zero bound’ renders central bank interest-rate management less effective at low rates of inflation, ECB studies argued that inflation rates below, but close to, two per cent would provide a sufficient safeguard against these risks. Second, the Governing Council emphasized that the ‘monetary analysis’ pillar was meant to serve mainly as a means of cross-checking, from a medium- to long-term perspective, the short- to medium-term indications provided by the ‘economic analysis’ pillar. To underscore the longer-term nature of the reference value for monetary growth, the practice of an annual review of the latter was discontinued.

It will be interesting to observe whether eventually the monetary pillar comes to be viewed as having been of importance only in the early years of operation of the ECB when the ECB had to establish its credibility by being as committed to price stability as the Deutsche Bundesbank (the German central bank) had been prior to 1999 and when the ECB was confronted with sizeable problems regarding the measurement of harmonized euro area-wide real economic aggregates, or whether ECB-style cross-checking by means of monetary analysis will become a common practice of central banks around the globe.

The operational framework used by the ECB to implement its monetary policy strategy is less controversial than the strategy itself and includes three main instruments: open market operations, standing facilities and reserve requirements. Among the open market operations of primary importance are the ‘main refinancing options’ that provide the bulk of refinancing to the financial sector and, through signalling the ECB's monetary policy stance, are supposed to steer market interest rates. The ‘main refinancing options’ are executed by the national central banks of the euro area member countries on a weekly basis through a tender procedure spanning three working days. ‘Standing facilities’ aim at providing and absorbing overnight liquidity, and ‘minimum reserve requirements’ (the ECB imposes minimum reserves on all credit institutions in the proportion of two per cent of the reserve base) aim at stabilizing market interest rates.

By way of evaluating the overall success of the ECB in terms of it being able to adhere to its price stability objective, we may observe that inflation rates in the euro area since 1999 have on an annual basis on average been slightly above two per cent (in the range of up to 30 basis points above two per cent). Also, given that surveys of average long-term inflation expectations in the euro area have

consistently measured such expectations as below, but close to, two per cent, its track record has quite firmly established the ECB's credibility with regard to safeguarding price stability.

Lesson three: One central bank for many countries: how to organize decision-making

The most important decision-making body of the ECB is its Governing Council, which is made up of the Executive Board of the ECB (which in turn is made up of its president, vice-president and four other members) as well as the governors of all the national central banks of euro area member countries. It is the responsibility of the Governing Council to formulate monetary policy for the euro area, including decisions about intermediate objectives and key interest rates. The Executive Board is in charge of implementing the monetary policy decisions taken by the Governing Council, and to this purpose cooperates with the national central banks through open market activities. Each member of the Governing Council has one vote. Given that at present slightly more than two-thirds of the votes in the Governing Council, therefore, belong to national central banks, the latter have a strong influence on the ECB's monetary policy decisions.
This organizational structure implies an asymmetry between the economic size of euro area member countries and their influence on decisions arrived at by the Governing Council. Indeed, more than half the euro area member countries at present have an economic weight (as measured by the ratio of their national GDP to euro area GDP) that is smaller than their voting weight within the Governing Council. This is quite different from the structure of, say, the US. Federal Reserve, which is significantly more centralized. While decentralization of the implementation of the ECB's monetary policy arguably is useful, particularly as long as there are important differences among national financial markets and institutions in the euro area, the decentralized institutional set-up of the ECB has risks, particularly during episodes of real divergence. It will be interesting to see whether the ‘one person, one vote’ principle for the Governing Council will be maintained after possible enlargement of the euro area to incorporate (some of) the EU member countries not presently member countries of the ECB. Even if the ‘one person, one vote’ principle is to be maintained, there appears to be considerable scope for future revision of the organizational system of the ECB, such as requiring approval of nominations of new central bank presidents by the Executive Board of the ECB.

Lesson four: Common currency and monetary policy: gains and losses

In general, the principal advantages of a common currency are widely held to include the reduction of transaction and information costs implied by the use of a common medium of exchange as well as the stimulus the common currency provides for the convergence of organizational principles used in business, in turn stimulating trade in goods and services and of cross-country flows of capital. The principal disadvantages of a common currency for multiple countries are widely held to include the loss of shock-absorber properties of flexible exchange rates and of independent national monetary policies. Furthermore, if a single monetary policy is accompanied by a diverse set of national fiscal policies, inappropriate fiscal policy in one country will – through its effect on interest rates – directly spread to other countries in the monetary union. Thus macroeconomic stability could be affected for the worse. How has the euro area so far fared on these counts? Trade within the euro area increased from approximately 26.5 per cent of (euro area) GDP in 1998 to approximately 31 per cent of GDP in 2005;

one and a half per cent of this increase was due to trade in services. Taking into account the limited time span, it is difficult to assess, however, to what extent this increase in trade was indeed driven by the creation of a single currency and to what extent it may instead have been driven by the process of economic globalization. We do know, in fact, that trade with trading partners outside the euro area over this same time period rose by a slightly larger margin than intra-euro area trade, from approximately 24 per cent of GDP in 1998 to approximately 30 per cent of GDP in 2005.

Regarding financial markets, for which the volume of transactions is probably still more sensitive to even small costs and risks associated with the use of multiple currencies, by a variety of measures deeper, broader and more liquid markets have emerged for the euro area member countries since establishment of the ECB. On the money market, issues of their interpretation aside, cross-country standard deviations for average overnight lending rates fell from 130 basis points in January 1998 to three basis points one year later, and since then have decreased to approximately one basis point. Cross-country standard deviations for rates at longer maturities (one and 12 months) for unsecured money market instruments have fallen to less than one basis point also, with the spreads still somewhat larger in the collateralized repurchase agreement (repo) market (due to continued differences in legal structures across euro area countries). In the interest rate derivatives market, the euro interest rate swap market at a daily volume of 250 billion euro was in 2006 one and a half times as large as the corresponding US dollar market. In the government bond market also, spreads have fallen to low levels, suggesting – in the likely absence of major changes in default risks – a significant fall of liquidity risk. The holdings of euro-denominated debt securities overall since 1999 have increased by well over ten per cent to approximately one-third of the global market (through holdings tend to be concentrated in countries neighbouring the euro area).

In the equity and retail banking markets integration has progressed more slowly. For example, despite a decrease in the number of credit institutions in the euro area member countries by almost 50 per cent between 1997 and 2006, less than one-third of the mergers and acquisitions driving this consolidation process have been cross-border. Also, the cross-country standard deviation of interest rates on consumer credit from 2004 to 2006 has still been close to one per cent.

While, just as for trade, it is difficult to disentangle the euro's contribution to the process of financial integration in euro area member countries from the global trend towards financial integration, the euro surely has greatly facilitated the task of bringing the European financial system closer to US standards in terms of market depth and liquidity. Further improvements in this direction, including the creation of a single payment system for the euro area member countries, are likely to intensify the debate about the potential role of the euro as a complement or competitor to the US dollar as an international reserve currency.

Finally, to turn to macroeconomic stability and the potential cost of losing flexible exchange rates and independent national monetary policies as shock absorbers, some such costs clearly have been observed since 1999. While the cross-country standard deviation of consumer price changes has fallen from approximately six per cent in the late 1990s to one per cent with the launch of the euro, and has been rather stable at this level in the following eight years, there have been persistent deviations from euro area average inflation rates for some countries, implying sizeable (and potentially destabilizing) differences in real interest rates. For example, for a sizeable part of the time period since 1999, real interest rates have been significantly lower in a booming Irish economy than in a German economy experiencing weak growth. When it comes to assessing the implications of the establishment of the ECB

for macroeconomic stability, these costs have to be subtracted from benefits owed to factors such as the elimination of intra-euro area exchange rate crises and the fact that inflation rates for some euro area member countries have been falling sizeably in the eight years since 1999. However, a stronger degree of real convergence through aligned policies aimed at removing structural deficiencies in European product and labour markets would have helped to render the benefits yet larger.

Conclusion

While this article has suggested that on various counts (such as the monetary policy strategy and the organizational set-up) there is as of 2007 no consensus as to whether the ECB adheres to best international practice in central banking, it would appear rather questionable to label the establishment of the ECB and with it the introduction of the euro as anything but an enormous success. The ECB has successfully mastered the technical challenges of establishing a new common currency across a set of countries comprising one of the largest economic regions in the world, has in a short period of time established a strong track record of success in preserving price stability, and has on many counts, particularly in the area of financial markets, helped lead the way to a stronger integration of European markets. While it is undisputable that this integration of markets along with structural reforms needs to proceed much further, the key decisions that could facilitate such integration and structural reforms fall outside the core domain of responsibility of the ECB and, for that matter, should probably remain so for any central bank primarily entrusted with maintaining price stability.

See Also

- euro
- European Monetary Union
- federal reserve system
- inflation targeting
- Stability and Growth Pact

Bibliography

European Central Bank. 2004. *The Monetary Policy of the ECB*. Frankfurt am Main: ECB.

European Central Bank. 2006. *ECB Statistical Data Warehouse*. Online. Available at http://sdw.ecb.int, accessed 14 March 2007.

Issing, O. 2003. *Background Studies for the ECB's Evaluation of its Monetary Policy Strategy*. Frankfurt am Main: ECB.

Issing, O., Gaspar, V., Angeloni, I. and Tristani, O. 2001. *Monetary Policy in the Euro Area*. Cambridge: Cambridge University Press.

Padoa-Schioppa, T. 2004. *The Euro and Its Central Bank: Getting United after the Union*. Cambridge, MA: MIT Press.

Posen, A.S. 2005. *The Euro at Five: Ready for a Global Role*? Washington: Institute for International Economics.

How to cite this article

Binder, Michael and Volker Wieland. "European Central Bank." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000128>
doi:10.1057/9780230226203.0507

The New Palgrave Dictionary of Economics Online

European labour markets

Giuseppe Bertola

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Labour taxes and subsidies, collective wage bargaining, and employment protection legislation affect labour market outcomes in European countries more strongly than in other advanced countries. This article outlines theoretical approaches to their motivation and consequences and reviews empirical insights from comparative cross-country studies of how employment, unemployment, and wage dynamics are shaped by the interaction between institutions, macroeconomic developments, and structural features.

Keywords

active labour market policies; centralized wage determination; efficiency wages; employment protection legislation; European labour markets; health insurance; inflation; job creation; labour market search; labour supply; labour taxes; minimum wages; pensions; structural change; tax wedges; training; unemployment insurance; unemployment subsidies; wage floors; wage rigidities; women's work and wages

Article

European labour markets, especially those of Continental countries, are characterized by more unionized wage setting and more stringent regulation of employment relationships than those of other OECD countries. Within that group of advanced countries, their unemployment rates used to be relatively low, and became very high. Around 1970, the unemployment rate was approximately 3.1 per cent in the OECD aggregate and five per cent in the United States, but the unemployment rate hardly exceeded four per cent in any European country. In the aggregate of 11 core European Union countries that later adopted the euro at its inception, unemployment stood at only 2.2 per cent in 1970. It then rose rapidly, exceeding ten per cent in 1984 and hovering around 12 per cent in the second half of the 1990s, while both the United States and the OECD aggregate unemployment rates fluctuated between four per cent

and nine per cent.

The wide variety of labour market developments over the last quarter of the 20th century has motivated extensive modelling efforts and comparative empirical studies of institutional features' motivation and effects. This article reviews the roles of institutions, shocks, and structural change in shaping aggregate and disaggregate labour market outcomes.

Labour market policies

To illustrate the spirit of more general approaches to the relevant issues, it is useful to focus initially on the simplest models and the best understood labour market institutions (Prescott, 2004). Consider inverse demand and supply functions

$$w^d = a(l), \quad w^s = s(l), \tag{1}$$

where l denotes log employment and w^s and w^d denote log wage rates. The wage w^* and employment l^* that equate supply and demand satisfy the condition

$$s(l^*) = a(l^*) = w^* \tag{2}$$

in static competitive equilibrium. As the simplest example of how institutions can change this outcome, consider a labour income tax that, inserting a wedge τ between employers' labour costs and workers' take home pay, changes the equilibrium condition to

$$s(\underline{l}) = a(\underline{l}) - \tau \tag{3}$$

and lowers employment by about

$$\underline{l} - l^* \approx -\tau / (\eta + \varepsilon) \tag{4}$$

where $s'(l) = \varepsilon \geq 0$ and $a'(l) = -\eta$, $0 < \eta < 1$. It is also simple to characterize formally the effects of binding legal or contractual minimum wage levels. If the wage is $\underline{w} > w^*$, the employment levels corresponding to $\underline{w}$ on the supply and demand curves are defined by $s(L) = \underline{w}$ and $a(\underline{l}) = \underline{w}$, and differ by the number

$$L - \underline{l} \approx (\underline{w} - w^*)(\varepsilon + \eta) / (\varepsilon\eta) \tag{5}$$

of unemployed workers, who would be willing to work at the going wage but cannot obtain employment. From this simple perspective it is obvious that differences in taxation and wage floors may explain cross-country differences in employment and unemployment. Qualitatively similar insights can be derived in the context of more complex and realistic models of unemployment, and can be applied to other institutions. When unemployment is due to matching frictions, efficiency wages and other imperfect allocation mechanisms, taxes and wage rigidities can affect search efforts and equilibrium employment and unemployment, which are affected in turn by the market's structure (such as the extent of mismatch between workers' qualifications and vacancies) and by other institutional features (such as the scope and efficiency of employment agencies). In both competitive and frictional models of the labour market, benefits paid to out-of-work individuals can affect labour supply and search effort, and there can be similar effects from less visible policy aspects, such as the availability of public-sector employment opportunities at favourable wage–effort ratios (Algan, Cahuc and Zylberberg, 2002).

At the same time as it offers obvious explanations for labour market outcomes, institutional variation raises the less obvious issues of why institutions should be as different across countries as they are observed to be, and of how their configuration and impact may depend on structural labour market features.

The relevance of distributional issues and of market imperfections can explain some of the labour market institutions' heterogeneity. The equilibrium condition (1) efficiently equates employed labour's marginal productivity with its non-employment opportunity cost, and distorting this outcome reduces the welfare of a perfectly competitive economy's representative individual. If workers disregard non-labour income, however, their total surplus can be increased by trading lower employment against higher pay along downward-sloping labour demand curves such as (2). It is maximized when the wage exceeds the marginal opportunity cost of employment by a monopolistic markup factor, and employment is set at a level $\underline{l}$ such that

$$a(\underline{l}) - s(\underline{l}) = \log[1 / (1 - \eta)] \approx \eta. \tag{6}$$

All workers' welfare can be increased if the higher wages earned by those who are employed more than compensate for the labour income lost by those who would be employed at the competitive wage. Such compensation may take place within families, or over individual lifetimes, and can also be explicit if the revenue raised by employment taxes is spent subsidizing non-employed individuals.
Institutions that decrease employment and increase labour costs can be rationalized recognizing that they affect not only the amount of production but also its distribution across heterogeneous individuals, and that markets (especially financial markets) are not perfect in real-life economies. Higher wages and lower employment can benefit workers who have negligible non-labour income, and households' limited access to formal financial markets can rationalize collectively administered risk-sharing schemes (Agell, 2002). In European countries, legislation meant to endow workers with some bargaining power and to insure them against health, unemployment and old-age hazards was introduced at times of actual or feared social unrest, in Bismarck's industrializing Germany or in Lord Beveridge's post-war United Kingdom. In principle, it can be efficient to try to provide insurance through mandatory government schemes when information and legal enforcement problems make it difficult for private markets to do so. But public schemes are not immune from such problems, and tend to reduce employment as, for example, recipients of unemployment subsidies reduce work effort. Such efficiency losses are more easily affordable by richer societies, and Europe's fast and stable post-war growth was unsurprisingly accompanied by development of increasingly extensive legislation and co-decision powers by unions.
By the early 1970s, the institutional structure of labour markets was distinctively different not only across the United States and Europe as a whole, but also across countries within Europe, where labour market policies play different roles in different welfare state models (Bertola et al., 2001). In Nordic countries, a tradition of full employment and universal welfare is based on generous unemployment benefits and a very important role for active labour market policies (including job creation in the public sector). The Bismarckian model of Continental countries such as France and Germany features centralized wage determination and stringent employment protection legislation, and contributory pension, health, and unemployment insurance programmes. The Beveridgian model of the United Kingdom and other Anglo-Saxon countries features social assistance safety financed by general taxation and comparatively light regulation of wage determination and employment relationships.

The dynamics of European labour market outcomes

Even though relief from the need to work should in general reduce employment, until the 1970s, and even in the aftermath of the late 1960s period of worker unrest, increasingly generous pro-worker institutions coexisted in Europe with low unemployment rates; much lower, in fact, than in the comparatively unregulated United States. The first oil shock and the following decades of slower growth saw the inception and persistence of high unemployment in most European countries, and increasing attention to the effect of institutions on labour market performance. If wages are preset, shocks can cause employment and unemployment fluctuations, the size and persistence of which depends on the extent of *ex post* wage flexibility and on the character of wage bargaining. Nominal shocks are a more relevant source of real wage misalignments and unemployment in labour markets with more pervasive and longer-term collective wage contracts. Conversely, real wages react more promptly to productivity

shocks or growth slowdowns if bargaining parties are in a better position to take into account their employment implications. Reactions to country-wide shocks are quicker, and the unemployment consequences of such shocks less severe, when wage bargaining is more centralized and better coordinated across industries (Calmfors and Driffill, 1988).

This can explain why unemployment began to increase, more or less sharply, when in the 1970s European countries were hit by oil shocks and other macroeconomic developments that reduced the amount of labour demanded at any given wage. Inflation and output dynamics subsequently appear to drive European unemployment fluctuations around a natural level that, after having raised sharply until the early 1980s, has remained essentially flat since the mid-1980s (Blanchard, 2006). The prolonged upward trend and the resilience of high unemployment levels naturally draw attention to non-cyclical, structural aspects of labour market dynamics. Wage floors can prevent underbidding by the unemployed of eq. (5), but it is difficult for that static relationship to explain why, in the absence of institutional changes that would further increase unions' wage-setting power, unemployment remained high in the aftermath of the 1970s crises.

A more suitable dynamic perspective is offered by models where labour demand shocks can permanently affect the link between wages and outside options, for example because job losers no longer have a say in wage determination, or because replacement of employed workers would entail large turnover costs (Lindbeck and Snower, 1988). The persistence of employment and unemployment dynamics, however, is in fact influenced not only by limited wage-setting flexibility but also by regulatory constraints on hiring and firing. In European countries, employment protection legislation (EPL) typically requires that the reasons for individual dismissals be stated by employers and subject to court appeal, and that collective dismissals be conditional on administrative procedures involving formal negotiations with workers' organizations and with local or national authorities.

Such provisions do have the intended effect of 'protecting' jobs at times of declining labour demand, when firing costs smooth out job losses and reduce downward wage pressure. Just because such a situation is costly for employers, however, it is optimal for them to refrain from hiring in upturns, so as to reduce the desirability of labour shedding in downturns. In terms of simple demand-and-supply relationships such as those introduced above, the marginal productivity of labour should be lower than the wage when employment is declining and firing a marginal worker entails firing costs as well as wage-cost savings, but it should symmetrically be higher than the wage when employment is increasing, and the marginal worker's costs include expected future firing costs as well as the current wage. Thus, the implications of EPL are similar to those of labour taxes for expanding firms, and to those of employment subsidies for downsizing firms. If employment fluctuations are efficient in laissez-faire, EPL obviously reduces production and profits. Unlike labour taxes, however, it does do not do so by reducing employment on average (Bentolila and Bertola, 1990), because its contrasting effects on employers' propensity to hire and fire reduce employment volatility but affect its average level ambiguously. Empirically, in fact, there is no convincing evidence of any relationship between EPL and the employment or unemployment level. As discussed in some detail below, correlations have to be treated with caution in this context, but more stringent EPL is associated with more stable aggregate employment paths and with longer unemployment durations within the pool of unemployed workers (Bertola, 1999). There is also some evidence that EPL affects the demographic composition of employment and unemployment – as it should in theory, since it reduces job finding rates for young job

market entrants and female workers with intermittent labour force participation at the same time as it reduces job-loss rates for mature workers.

Another important related difference across labour markets pertains to the extent and character of wage inequality. Earnings are typically less dispersed in Europe than in other advanced countries. The extent of underlying heterogeneity in workers' characteristics is an important determinant of earnings dispersion, but institutional wage-setting constraints also appear very relevant, both theoretically and empirically. While centralized bargaining may be better able to coordinate reactions to aggregate shocks, it tends to result in less detailed, more homogenous wage structures across firms, sectors, regions and individuals. Similar wages for heterogeneous workers imply divergence of employment outcomes, for example across demographic groups (Kahn, 2000) and across regions in Italy, Germany and Spain, where the uniformity of centrally bargained wages (and of other national institutions) tends to lower employment where labour is less productive. Empirically, relative wage variation appears to be heavily constrained in the same countries where EPL is most stringent (Bertola and Rogerson, 1997). This is unsurprising, because quantitative firing restrictions could hardly be binding if, in the face of negative labour demand shocks, wages could fall so as to make stable employment profitable, or to induce voluntary quits. Across countries, the combination of wage and quantity rigidities indeed appears to protect employed workers from labour income volatility, as individuals enjoy more stable wages and longer tenures.

At the aggregate level, the role of institutions in shaping heterogeneous dynamics across labour markets is not as immediately apparent. Institutions vary widely across countries but, within each country, they are much more stable than unemployment, wage inequality and other labour market outcome variables. As discussed above, however, wage-setting institutions can shape an economy's reaction to aggregate shocks. More generally, the same dynamic developments can produce very different employment and wage outcomes in countries with different (albeit stable) institutions. This can explain why, in the 1970s and 1980s, countries with more extensively regulated labour markets experienced more pronounced unemployment increases in the aftermath of similar productivity, inflation and wage shocks (Blanchard and Wolfers, 2000). Empirically, in fact, the forces that interact with labour market institutions in driving dynamic trajectories can be almost equally well represented by period-specific dummy variables as by observable macroeconomic variables, which tend to behave rather similarly over time across industrialized countries. Thus, the evidence can be consistent with a role for common structural trends rather than for country-specific shocks.

For example, the relationship between country-specific labour market institutions and unemployment and wage dispersion dynamics can be interpreted in the light of skill-biased technological progress trends, or of increasing opportunities for advanced countries to import unskilled labour-intensive goods and export skill-intensive ones. Over the last three decades of the 20th century unemployment displayed a trend increase in Continental European countries but remained trendless in the United States and other Anglo-Saxon countries, while earnings inequality remained stable (or even declined) in the former group of countries but trended upward in the latter. If technological progress or international trade increase laissez-faire wage inequality, they also increase the relevance of wage floors: if in European countries low wages cannot decline, employment of unskilled workers must decline (Krugman, 1994). Similar insights into the changing implications of unchanging institutions can be gained by considering other structural aspects. More intense product market competition, as implied by Europe's economic integration process and by more general globalization trends, increases the elasticity of labour demand.

In the context of the simple example above, a smaller η implies larger employment losses from any given tax wedge in eq. (4), and higher unemployment from any given wage floor in eq. (5). In more complex dynamic models, if reallocation towards higher-paying jobs is costly, then institutions that tend to prevent wage inequality and restrict mobility have sharper implications for employment and unemployment when more volatile shocks affect labour demand (Ljungqvist and Sargent, 1998). Structural change can magnify the unemployment and employment effects of institutions meant to redistribute income and remedy financial market imperfection, or it can make them redundant (for example, because financial market development makes labour income fluctuations less problematic). Then, institutions should be reformed. In the simple formal framework above, the same smaller η that amplifies the negative employment implications of given institutions also calls for a smaller markup in eq. (6). And, in reality, policy frameworks introduced in the 1990s, such as those recommended by the OECD Jobs Study (OECD, 1994) and by the European Union's Lisbon Strategy (Council of the European Union, 2000), de-emphasize income support for job seekers and job losers in favour of job creation spurred by wage and employment flexibility, and the role of training and other active labour market policies aimed at bringing workers' productivity in line with wage aspirations.

Reforms are at least partly motivated by better theoretical and empirical understanding of the effects of labour market institutions. But while it is in principle obvious that institutional interference can be responsible for high unemployment and low employment, just because such effects depend on potentially heterogeneous structural parameters, that it is hard to assess their impact in data where many relevant confounding factors cannot be controlled. Simple correlation can be very misleading. For example, a negative cross-country correlation between EPL and employment rates is fully accounted for by low female employment–population ratios in southern Europe (Nickell, 1997), while effects on prime-age male employment rates tend to be positive. Both policies and outcomes can jointly respond to underlying cultural differences in this and other cases, and it is difficult to obtain reliable estimates from cross-sectional relationships between institutions and outcomes (Baker et al., 2005). More articulate and robust insights may be obtained from specifications where time-series variation and interactions play important roles (Bassanini and Duval, 2006). As the time dimension of available data increases, however, it will be increasingly important when interpreting time-series evidence to focus on the economics and politics of reform processes rather than on institutions at each point in time (Saint-Paul, 2000), and to be aware of plausible channels of institutional endogeneity. If shocks or structural changes make job loss more or less likely or trigger painful changes in the generosity of unemployment insurance or in the stringency of employment protection legislation, for example, the correlation between such institutions and employment performances may be largely spurious. The wide and changing variety of labour market policies across countries offers opportunities to try to disentangle their effects in increasingly available disaggregated data, at the same time as it makes it necessary to take into account the many important and related respects, besides labour market structure, in which countries differ.

See Also

- European Monetary Union

- globalization and the welfare state
- Phillips curve
- Phillips curve (new views)
- skill-biased technical change
- unemployment

Bibliography

Agell, J. 2002. On the determinants of labour market institutions: rent seeking vs. social insurance. *German Economic Review* 3, 107–35.

Algan, Y., Cahuc, P. and Zylberberg, A. 2002. Public employment and labour market performance. *Economic Policy* 34, 9–65.

Baker, D., Glyn, A., Howell, D. and Schmitt, J. 2005. Labour market institutions and unemployment: a critical assessment of the cross-country evidence. In *Fighting Unemployment: The Limits of Free Market Orthodoxy*, ed. D. Howell. Oxford: Oxford University Press.

Bassanini, A. and Duval, R. 2006. Employment patterns in OECD countries: reassessing the role of policies and institutions. Economics Department Working Paper No. 486; Social, Employment and Migration Working Paper No. 35. Paris: OECD.

Bentolila, S. and Bertola, G. 1990. Firing costs and labour demand: how bad is Eurosclerosis? *Review of Economic Studies* 57, 381–402.

Bertola, G. 1999. Microeconomic perspectives on aggregate labour markets. In *Handbook of Labour Economics*, vol. 3C, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Bertola, G., Jimeno, J.F., Marimon, R. and Pissarides, C. 2001. Welfare systems and labour markets in Europe: what convergence before and after EMU? In *Welfare and Employment in a United Europe*, ed. G. Bertola, T. Boeri and G. Nicoletti. Cambridge, MA: MIT Press.

Bertola, G. and Rogerson, R. 1997. Institutions and labour reallocation. *European Economic Review* 41, 1147–71.

Blanchard, O.J. 2006. European unemployment: the evolution of facts and ideas. *Economic Policy* 45, 7–59.

Blanchard, O.J. and Wolfers, J. 2000. The role of shocks and institutions in the rise of European unemployment: the aggregate evidence. *Economic Journal* 110, C1–C33.

Calmfors, L. and Driffill, J. 1988. Centralization of wage bargaining. *Economic Policy* 3, 14–61.

Council of the European Union. 2000. *Conclusions of the Lisbon European Council*, Council of the European Union SN 100/00, 23–24 March.

Kahn, L.M. 2000. Wage inequality, collective bargaining and relative employment 1985–94: evidence from 15 OECD countries. *Review of Economics and Statistics* 82, 564–79.

Krugman, P. 1994. Past and prospective causes of high unemployment. *Federal Reserve Bank of Kansas City Economic Review* 1994(4), 23–43.

Lindbeck, A. and Snower, D.J. 1988. *The Insider–Outsider Theory of Employment and Unemployment*. Cambridge, MA: MIT Press.

Ljungqvist, L. and Sargent, T.J. 1998. The European unemployment dilemma. *Journal of Political Economy* 106, 514–50.

Nickell, S. 1997. Unemployment and labour market rigidities: Europe versus North America. *Journal of Economic Perspectives* 11(3), 55–74.

OECD (Organisation for Economic Co-operation and Development). 1994. *The OECD Jobs Study: Evidence and Explanations*. Paris: OECD.

Prescott, E.C. 2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28(1), 2–13.

Saint-Paul, G. 2000. *The Political Economy of Labour Market Institutions*. Oxford and New York: Oxford University Press.

How to cite this article

Bertola, Giuseppe. "European labour markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000309> doi:10.1057/9780230226203.0508

The New Palgrave Dictionary of Economics Online

European Monetary Union

Charles Wyplosz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The European Monetary Union is an original, complex undertaking in the making. The single currency has so far operated smoothly, yet many of its details are controversial. Inflation has been low and the European Central Bank has achieved some credibility, though its strategy is often criticized. The euro has not displaced the dollar as the world currency, but has allowed for the emergence of Europe-wide bond market. The downside has been low growth in many countries, presumably a consequence of market rigidities. Similarly, the Stability and Growth Pact, designed to eliminate the budget deficit bias, has had a checkered history.

Keywords

budget deficits; deflation; euro; European Central Bank; European Monetary Union; European Overnight Interest Average; European Union; Eurosystem; excessive deficit procedure (EU); Exchange Rate Mechanism (EU); externalities; Federal Reserve System; foreign exchange risk; Harmonized Index of Consumer Prices (HICP); inflation targeting; Maastricht Treaty (EU); optimum currency area; price stability; price transparency; Stability and Growth Pact; sticky wages; Taylor rule; time consistency of monetary and fiscal policy; trade diversion

Article

In 2007, 13 independent countries shared the same currency, the euro. The adoption of the common currency is a major step in the deep integration process followed in Europe since the Second World War.

Facts

The European Monetary Union (EMU) started to operate on 1 January 1999, when the authority to carry out monetary policy was transferred from national central banks to the Eurosystem, which is described

below. Banknotes and coins were introduced three years later; until then, national currencies continued to circulate, but their exchange rates were irrevocably fixed. The union started with 11 member countries (Austria, Belgium, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Portugal and Spain); Greece joined in 2001 and Slovenia followed in 2007. All European Union (EU) countries are expected to join, with the exception of Denmark and the UK, which have been given, and have exercised, an opting-out clause. Sweden has decided not to join. Potentially, therefore, the euro could be the currency of 25 to 30 countries.

Within the Eurozone, there is a single short-term interest rate (European Overnight Interest Average, EONIA), which is steered by the monetary authorities though their biweekly auctions and deposit and lending facilities. Longer bond rates have converged; the remaining difference largely reflects different risk factors.

The euro is slowly becoming an international currency. The euro was involved in 37.3 per cent on one side of all exchange market transactions recorded in the BIS (2005) survey, far behind the US dollar (88.7 per cent) and ahead of the Japanese yen (20.3 per cent) and the pound sterling (16.9 per cent). Similarly, in 2004 the market share of euro-denominated bonds stood at about 31 per cent, not too far below the dollar's share of 43 per cent (since every exchange market transaction involves two currencies, the shares add up to 200 per cent; ECB, 2005).

Admission and OCA principles

In order to join the Eurozone a country must satisfy five criteria – the Maastricht criteria, in reference to the Maastricht Treaty, which established the monetary union in 1992. These criteria are: (*a*) at least two years of membership in the Exchange Rate Mechanism, without devaluation; (*b*) an inflation rate that does not exceed by more than 1.5 percentage points the average inflation rate observed in the three European Union countries where it is lowest; (*c*) a short-term interest rate that does not exceed by more than two percentage points the average interest rate observed in the same three EU countries; (*d*) a budget deficit that does not exceed three per cent of GDP; and (*e*) a public debt that does not exceed 60 per cent of GDP, or that is declining towards this level.

Surprisingly, perhaps, the admission criteria are totally unrelated to the optimum currency area (OCA) theory. OCA theory, which evaluates the costs and benefits from joining a monetary union, focuses on the likely occurrence of asymmetric shocks and on the ability to face such shocks without having recourse to the exchange rate. The admission criteria were designed instead to ensure that countries would be allowed into the monetary union only after having demonstrated their commitment to price stability. OCA principles were not seen as helpful, if only because the decision to adopt a single currency was largely political. Policymakers were also eager to stabilize intra-European exchange rates once they had decided to lift all restrictions to capital movements. In addition, Germany would not abandon the Deutschmark for a weaker currency. In the end, among the interested countries, only Greece could not meet the criteria and postponed its application for two years. Among the countries that joined the EU in 2004, Slovenia joined in 2007, Lithuania failed in 2006 and Estonia was invited not to apply. By end 2006, the others had not announced any plan, although four more had joined the Exchange Rate Mechanism (Cyprus, Latvia, Malta, Slovakia).

Structure

Monetary policy decisions lie with the Eurosystem, which consists of the newly created European Central Bank and the national central banks (NCBs) of all member countries. The arrangement very much resembles the US Fed, with its Federal Reserve Board and regional Reserve Banks. All decisions are made by the Governing Council, which is made up of the ECB Board's six members and the (currently 13) governors of the NCBs. The president of the ECB acts as the president of the Eurosystem. The meetings of the Governing Council are highly secretive. Although decisions are made by simple majority on a one-person-one-vote principle, the Council reports that it decides by consensus. The Council does not publish minutes but each of its monthly monetary policy meetings is immediately followed by a communiqué and a press conference by the president and vice-president of the ECB (the transcript of which is promptly posted on the ECB's website).

As more countries are expected to join, the size of Council will expand. As a large Council is unlikely to operate efficiently, it has been decided to cap its size at 25 and to rotate the NCB governors, giving more representation to large countries, again in a way that resembles the Federal Open Market Committee. Outside observers have criticized the current arrangement as involving too many participants, and have expressed the view that 25 is far too many for effective discussion. It is widely believed that the decisions are in fact prepared by the ECB's Executive Board and that the full meeting of the Governing Council adds little substance. The whole question of the role of the NCBs – some of which still employ several thousand people – is highly sensitive and rarely addressed in public by central bank and government officials.

Performance: inflation and growth

The Maastricht Treaty, which created the monetary union, sets price stability as its primary objective. It adds that ‘without prejudice to the objective of price stability, [the Eurosystem] shall support the general economic policies in the Community’. In other words, there is a lexicographic ordering of priorities: growth and employment are a concern only if inflation is not. The Eurosystem has provided its own definition of price stability: ‘an inflation rate close to but below 2% over the medium run.’ It has indicated that it would monitor the area-wide price level (called Harmonized Index of Consumer Prices, HICP).

During its first 90 months of existence, the HICP inflation rate has been below two per cent only 26 times, most which occurred early on – the starting annualized inflation rate in January 1999 stood at 0.8 per cent. On the other hand, it has been between 1.5 and 2.5 per cent for 75 months, most of the exceptions occurring again during the early period. The rate exceeded 2.5 per cent only six times. In that sense, the Eurosystem has not delivered on its commitment; yet, with inflation almost always below 2.5 per cent, price stability can be considered as achieved. The surprise has been the high degree of persistence of inflation. Given the short period of time since the euro was launched, little is known about the reasons for inflation persistence. Early findings indicate that persistence is particularly high in the services sector, which seems to suggest that wage stickiness plays an important role.

Since 1999, the Eurozone growth performance has been disappointing. Although there are large differences between countries, average annual growth over 1999–2006 stands at two per cent, significantly lower than over the previous decade. It has been widely noted that the three EU members

which have decided not to join (Denmark, Sweden and the UK) have achieved a better growth performance than the Eurozone. Early critics have argued that the Eurosystem has been too restrictive; some consider that the policy framework (presented below) is partly responsible as well. As time goes by, any real effect of monetary policy fades away, so this criticism is increasingly less convincing. Furthermore, various Taylor rule estimates reject the view that the Eurosystem has been unduly restrictive. The likely implication is that potential growth has slowed down throughout many Eurozone countries – especially in the larger ones (France, Germany and Italy), which jointly account for two-thirds of the area GDP – and therefore that monetary union cannot be blamed for this.

Controversies

Strategy, transparency and communication

The Eurosystem has adopted a two-pillar strategy inspired by the Bundesbank practice. The monetary pillar calls for a money and credit growth that is consistent with long-term price stability. The economic pillar looks at all factors that affect inflation in the shorter run. Initially, the Eurosystem put the monetary pillar first and even issued a numerical ‘reference value’ for M3 growth rate. Criticism has been widespread. Critics have argued that monetary targeting – which the Eurosystem denied pursuing – was outdated. They observed that actual money growth far exceeded the reference value, which also happened in Germany during the last decade of existence of the Deutschmark and contributed to the demise of monetary aggregate targeting.

In 2003, the Eurosystem promoted ‘economic analysis’ to the status of first pillar, still retaining the monetary pillar. This has not fully quieted criticism as a number of ‘ECB watchers’ have been recommending the flexible inflation-targeting strategy. The Eurosystem has repeatedly rejected inflation targeting. Initially, it even refused to publish its inflation forecast but now does so twice a year. In practice, however, the difference between the Eurosystem's strategy and flexible inflation targeting is slim. Taylor rule estimates indicate that the central bank cares about driving inflation to its desired level in the medium term; it does so by taking into account various constraints, growth obviously being one of them.

The main departure from inflation targeting concerns communication. In its monthly press conferences, the president does not refer to inflation forecasts; rather, he offers a presentation on how the Eurosystem interprets the signals from the two pillars. Critics argue that this reduces transparency and predictability, but there is little evidence that the Eurosystem is less predictable than other major central banks. Somehow, it seems, central bank watchers can always decipher the signals they receive. Yet the president's highly standardized presentation and the refusal to publish minutes of the Governing Council's deliberations provide arguments supporting the claim that the Eurosystem lacks transparency (Buiter, 1999; Svensson, 1999), a charge strenuously rejected by the Bank (Issing, 1999).

One size fits all

A key message from OCA theory is that the main costs of monetary union membership come from asymmetric shocks or from asymmetric effects of common shocks, including monetary policy. Early

studies concluded that Europe is not an optimum currency area, even if some criteria are reasonably well met. An implication was that there would be costs to Eurozone membership, as well as benefits. The benefits include the elimination of exchange risk and costs, enhanced price transparency and deeper financial and trade integration.

Discussions about the costs, largely ignored during the project preparation, have gradually risen to the surface under the question: does one size fit all? The issue is whether a single monetary policy can effectively deal with the needs of all member countries. It is too early to tell. Casual observation suggests that some countries, like Ireland and Spain, have undergone fast growth and relatively high inflation, while others have languished, Germany being a prime example. With a common nominal interest rate, the real rate has been low in the former countries and high in the latter, implying that monetary policy is pro-cyclical. Unsurprisingly, politicians have openly complained that monetary policy is ill-adapted to local needs. Yet, so far at least, the divergences have remained manageable, if not uncontroversial. Serious concern is emerging about Italy and Portugal, two countries with low growth and rapidly rising labour costs. Indications are that both countries have lost ground in price competitiveness. Without the exchange rate tool, they need to claw competitiveness back. With a Eurozone-wide inflation rate at two per cent, this may require lasting deflation. An alternative would be reductions in labour or other production taxes, but these countries are already facing excessive budget deficits.

The Stability and Growth Pact

The Maastricht Treaty introduces the concept of ‘excessive deficits’. While the three monetary entry criteria – low inflation, low interest rates and exchange rate stability – become moot once a country has joined the monetary union, the two budget entry criteria – public deficit and debt – remain a live issue. To that effect, the treaty calls for the adoption of an excessive deficit procedure (EDP) that makes fiscal discipline a permanent requirement. The Stability and Growth Pact (SGP) defines and implements this procedure.

The pact stipulates that the budget deficit should not exceed three per cent of GDP. If it does, an EDP is triggered. The delinquent government is first given a warning. If the deficit remains excessive, it is given specific policy recommendations and a deadline to bring the deficit back to below three per cent. If the deficit remains excessive, a fine is imposed. The fine takes the form of a deposit. It is reimbursed if the budget is not in deficit within one year; if not, the deposit is lost and redistributed to the other member countries. The pact applies to all EU countries but only Eurozone members are subject to the fine.

The SGP raises a number of important questions. First, do countries that share the same currency need a collective fiscal discipline instrument? Many federations have adopted compulsory restraints, but not all, and some restraints are voluntary. As for any coordination arrangement, the pact must be justified by the presence of externalities, in this case specific to governments that share a common currency. The most frequently quoted externality is that default by one government might force a bail-out by the others and the common central bank. The Maastricht Treaty already includes a no-bail-out clause, which explicitly prohibits official lending to a member government, but many believe that the clause could be violated in an emergency situation. Another possible externality is the possibility that one country's indiscipline affects the common exchange and interest rates. The existence and quantitative relevance of this externality remains to be documented.

On the assumption that some collective arrangement for fiscal restraint is justified, the next question is how it should be set up. The SGP relies on two principles: a deficit rule and collective surveillance, delegated to the European Commission, backed by a sanction mechanism. Both features have been controversial since the inception of the pact. The three per cent deficit limit is largely arbitrary, the focus on year-to-year deficits is a poor guide to fiscal discipline, and its implementation is time-inconsistent since restrictive policies are suboptimal during a recession. Externally imposed discipline is seen as an infringement of domestic sovereignty and of parliament's prerogative in voting budgets. A number of proposals have suggested that the pact should instead promote institutions, if possible national institutions, that deliver better policy outcomes (Von Hagen and Harden, 1995; Wyplosz, 2005).

In the early 2000s, several countries were found to be in excessive deficit, including France and Germany, the two largest ones. Their budgetary positions had deteriorated partly due to sluggish growth and partly because policy has been undisciplined in better years. They were not willing to tighten fiscal policy while their economies were stagnating. In November 2003, just as they were about to be sanctioned by the Council of Ministers, these two countries managed to obtain a majority that declared the pact in abeyance. The European Commission, which was duty bound to ask for sanctions, appealed to the European Court of Justice. The Court issued a balanced judgment: the Council of Ministers was allowed not to impose sanctions but not to put the pact in abeyance. A new Council resolution quickly followed the Court's guidelines.

With the smaller countries bitterly complaining, following a cooling period a reform was agreed in June 2005. The reformed pact calls for taking into account economic circumstances, such as a prolonged period of slow growth. It recognizes that the size of the public debt and the evolution of the cyclically adjusted budget, which were previously ignored, are important in reaching a view on fiscal policy. Importantly, governments can argue that they need to temporarily increase spending on particularly important items such as health reform, R&D or infrastructure. All of these amendments provide the Commission – which acts as the pact's watchdog – and the Council of Ministers with considerable flexibility. Officials argue that the SGP will now work better because it is less rigid; critics say that the pact is so flexible that it is useless.

Real effects of the monetary union

The European Commission and national governments have portrayed the monetary union as a source of important and large benefits. At the macroeconomic level, exchange rate stability, low inflation, lower interest rates and better fiscal policies were expected to deliver a better environment. Most of the benefits, however, were expected to be microeconomic. Reduced transaction costs, more trade and therefore more competition were seen as the source of static and, especially, dynamic efficiency gains. Although the Commission did not attempt to quantify these gains, subsequent studies suggested that they could be very substantial. It will take years to find out. Early studies have deflated the initial estimates by Rose (2000). Baldwin (2006) estimates that trade has increased by between five per cent and 15 per cent, not just within the Eurozone but also between the Eurozone and the rest of the world, thus without trade diversion – quite the contrary. What is surprising is that these effects do not seem to be triggered by reduced transaction costs but by a reduction in the fixed cost of introducing new goods into Eurozone markets.

See Also

- currency unions
- euro
- European Central Bank
- Stability and Growth Pact
- Taylor rules

Bibliography

Baldwin, R. 2006. *The Euro's Trade Effect*. London: CEPR.

Baldwin, R. and Wyplosz, C. 2006. *The Economics of European Integration*. Maidenhead: McGraw Hill.

BIS (Bank for International Settlements). 2005. *Triennial Central Bank Survey*. Basel: BIS. Online. Available at http://www.bis.org/publ/rpfx05t.pdf, accessed 26 October 2006.

Buiter, W. 1999. Alice in Euroland. *Journal of Common Market Studies* 37, 181–209.

Brunila, A., Buti, M. and Franco, M., eds. 2001. *The Stability and Growth Pact*. Basingstoke: Palgrave.

ECB (European Central Bank). 2005. *Review of the International Role of the Euro*. Frankfurt: ECB.

European Commission. 1990. One money, one market. *European Economy* 44. Luxembourg: Office for Official Publications of the European Communities.

Issing, O. 1999. The Eurosystem: transparent and accountable or ‘Willem in Euroland’. *Journal of Common Market Studies* 37, 503–19.

Rose, A.K. 2000. One money, one market: the effect of currency unions on trade. *Economic Policy* 30, 7–46.

Svensson, L.E.O. 1999. Monetary policy issues for the Eurosystem. *Carnegie-Rochester Conferences Series on Public Policy* 51, 79–136.

Von Hagen, J. and Harden, I. 1995. Budget processes and commitment to fiscal discipline. *European Economic Review* 39, 771–9.

Wyplosz, C. 2005. Fiscal policy: institutions versus rules. *National Institute Economic Review* 191, 70–84.

How to cite this article

Wyplosz, Charles. "European Monetary Union." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000300> doi:10.1057/9780230226203.0509

The New Palgrave Dictionary of Economics Online

Evans, Griffith Conrad (1887 – 1973)

Herbert A. Simon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

calculus of variations; comparative statics; Evans, G. C.; mathematical economics; non-equilibrium dynamics; two-sector models

Article

A distinguished American mathematician and pioneer mathematical economist, Evans was born on 11 May 1887 in Boston, Massachusetts. Educated in mathematics at Harvard University (AB, 1907; MA, 1908; Ph.D., 1910), he spent two years as a postdoctoral fellow studying with Vito Volterra at the University of Rome, then joined the faculty of the Rice Institute in Houston, Texas, where he taught from 1912 to 1934. In 1934, he became chairman of the mathematics department at the University of California, Berkeley, retaining that position until his retirement in 1954. He died on 8 December 1973, at the age of 86.

Evans's important contributions to mathematics, especially in functional analysis and potential theory, earned him membership in the National Academy of Sciences in 1933, as well as numerous other professional honours. His interest in mathematical economics became evident about 1920, when he gave his first series of lectures on that subject at the Rice Institute, and it continued up to the time of his retirement, his last publication on the subject appearing in 1954. It is likely that his initial contact with mathematical economics took place in Italy and France, for he shows great familiarity with the work of such writers as Pareto, Amoroso and Divisia, who were flourishing during and after his early Continental sojourn. Among earlier writers in mathematical economics, he mainly cities Cournot and Jevons; among his contemporaries, Irving Fisher, Henry Schultz, and Henry Moore.

Evans's most important work in economics is his *Mathematical Introduction to Economics* (1930), which also contains materials from his earlier papers. In the book and his other publications, he applied the calculus and the calculus of variations to problems of monopoly, duopoly and competition, and to a whole range of problems of comparative statics, including the incidence of taxes and the effects of tariffs. His approach was quite different from that of Walras (to whom he does not refer in his book), in that most of his models dealt with one or a few actors in a single market, or a small number of markets.

In his ‘Maximum Production Studied in a Simplified Economic System’ (1934, p. 37), he gave clear expression to his attitude toward general equilibrium models: ‘Large numbers of simultaneous equations in a large number of variables convey little information … about an economic system.’

In Evans's models, supply was generally described in terms of a cost function rather than a production function. To deal with macroeconomic problems, he constructed aggregate models, and, in order to provide a rationale for such models, he and his students made a deep study of the problem of index number construction. Their starting point was the work of Irving Fisher and François Divisia.

Evans's books and his articles were an important resource for early American students with an appetite for mathematical economics, who prior to their publication found an extremely sparse literature on which to graze. Samuelson, for example, mentions Evans as one whose works he ‘pored over’ when working on the *Foundations of Economic Analysis*. Moreover, Evans's methods of modelling economic situations gave new impetus to the approach of comparative statics, especially in application to macroeconomic problems. He constructed an early (perhaps the first) two-sector aggregative model containing a consumption good and a capital good; and he saw the power of second-order conditions of stability in reasoning about comparative statics, thereby anticipating by more than a decade Samuelson's important contributions to that topic.

Evans did only a little work in non-equilibrium dynamics, although he saw clearly the need for further development of that subject. In his ‘Simple Theory of Economic Crises’ (1931, p. 61), he complained that ‘the fact of lack of equilibrium in economic systems continually, and practically, stares us in the face; yet the principal discussion from a theoretical point of view has been of equilibrium, and thus at one stroke has eliminated a major issue.’

Evans was a fellow of the Econometric Society, and one of its founders. His principal influence upon the progress of economics came through the methodologies employed in his book, and through the work of his students, among whom were Francis W. Dresch, Kenneth May, C.F. Roos and Ronald W. Shephard, and one step removed, Lawrence W. Klein and Herbert A. Simon, who were colleagues or pupils of these students.

Selected works

1922. A simple theory of competition. *American Mathematical Monthly* 29, 371–80.

1924. The dynamics of monopoly. *American Mathematical Monthly* 31, 77–83.

1925a. Economics and the calculus of variations. *Proceedings of the National Academy of Sciences of the USA* 11, 90–95.

1925b. The mathematical theory of economics. *American Mathematical Monthly* 32, 104–10.

1929. Cournot on mathematical economics. *Bulletin of the American Mathematical Society* 35, 269–71.

1930. *Mathematical Introduction to Economics*. New York: McGraw-Hill.

1931. A simple theory of economic crises. *American Statistical Association Journal* 26(supplement), March, 61–8.

1932a. Stabilité et dynamique de la production dans l'économie politique.*Mémorial des Sciences Mathématiques 61*. Paris: Gauthier-Villars.

1932b. The role of hypothesis in economic theory. *Science* 75, 321–4.

1934. Maximum production studied in a simplified economic system. *Econometrica* 2, 37–50.

1937. Indices and the simplified system. In Cowles Commission for Research in Economics, *Report of Third Annual Research Conference on Economics and Statistics*. Chicago: University of Chicago Press.

1939. (With K. May.) Stability of limited competition and cooperation. *Reports of a Mathematical Colloquium*, 2nd series. Notre Dame University.

1950. Mathematics for theoretical economists. *Econometrica* 18, 203–4.

1952. Note on the velocity of circulation of money. *Econometrica* 20, 1.

1954. Subjective values and value symbols in economics. In *From Symbols and Values, an Initial Study: Thirteenth Symposium of the Conference on Science, Philosophy, and Religion*. New York: Harper & Row.

How to cite this article

Simon, Herbert A. "Evans, Griffith Conrad (1887–1973)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000139> doi:10.1057/9780230226203.0510

The New Palgrave Dictionary of Economics Online

evolutionary economics

Ulrich Witt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the way of thinking about economic problems and the research agenda associated with the evolutionary approach to economics. This approach generally focuses on the processes that transform the economy from within and on their consequences for firms and industries, production, trade, employment and growth. The article highlights the major contributions to evolutionary economics and explains its key concepts together with some of their implications.

Keywords

behavioural theory of the firm; bounded rationality; capital accumulation; competition and selection; Darwin, C.; diffusion of technology; endogeneity and exogeneity; entrepreneurship; evolutionary economics; evolutionary game theory; Georgescu-Roegen, N.; growth; industrial clusters; innovation; institutionalism; institutions; knowledge; learning; life cycle; national innovation systems; novelty; natural selection; path dependence; preferences; production theory; productivity growth; replicator equation; research and development; routines; rules of conduct; Schumpeter, J.; selection; structural change; technical progress; technology; Veblen, T.

Article

Evolutionary economics focuses on the processes that transform the economy from within and investigates their implications for firms and industries, production, trade, employment and growth. These processes emerge from the activities of agents with bounded rationality who learn from their own experience and that of others and who are capable of innovating. The diversity of individual capabilities, learning efforts, and innovative activities results in growing, distributed knowledge in the economy that supports the variety of coexisting technologies, institutions, and commercial enterprises. The variety drives competition and facilitates the discovery of better ways of doing things. The question in evolutionary economics is therefore not how, under varying conditions, economic resources are

optimally allocated in equilibrium given the state of individual preferences, technology and institutional conditions. The questions are instead why and how knowledge, preferences, technology, and institutions change in the historical process, and what impact these changes have on the state of the economy at any point in time.

Posing the questions this way has consequences for the way theorizing is done in evolutionary economics. First, preferences, technology and institutions become objects of analysis rather than being treated as exogenously given. Second, following from the very notion that evolution is a process of self-transformation, the causes of economic change are in part considered to be endogenous, and not exclusively exogenous shocks. More specifically, these causes are identified with the motivation and capacity of economic agents to learn and to innovate. Third, the evolutionary process in the economy is assumed to follow regular patterns on which explanatory hypotheses can be based, rather than forming an erratic sequence of singular historical events.

These three meta-premises are widely shared in evolutionary economics. However, the details of the argument, methods, and even the specification of the attribute ‘evolutionary’ vary, corresponding to the different theoretical traditions in which evolutionary economics is rooted. The concept of evolution has a long history in economics and social philosophy. This antedates – and, to a certain extent, has influenced – Darwin's theory of the origin of species by means of natural selection. Where the concept of evolution originally stood for a process of betterment (of human society), the Darwinian revolution in the sciences purged these progressive, teleological connotations. Today, evolutionary thought usually defines itself in relation to the Darwinian theory of evolution, the contributions to evolutionary economics not excepted. Some authors consider Darwinian theory to be the master theory. Others borrow from it at a heuristic level for their analogy-driven theorizing in economics. Yet others explicitly dissociate themselves from Darwinian thought.

Schumpeter and the neo-Schumpeterian synthesis

Schumpeter avoided the term ‘evolution’. He considered it a Darwinian concept and denied such concepts any economic relevance. However, in his theory of capitalist development, Schumpeter (1934) clearly subscribes to the three meta-premises above. The restructuring of the economy is explained as emerging endogenously from ever new waves of major innovations implemented by pioneering entrepreneurs with unique capabilities and motivation. Technology and the institutions of capitalism are endogenized. The transformation process of the economy is assumed to be governed by regular patterns, that is, cycles of investment and growth – booms and depressions – triggered by the innovations that occur ‘in waves’ and diffuse throughout the economy in competitive imitation processes.

In Schumpeter (1942, p. 83) innovations that ‘incessantly revolutionize the economic structure from within’ remain central, but the innovating agents change. Previously viewed as achievements of unique promoter-entrepreneurs, innovations now appear as the routine output of trained specialists in large corporations. Correspondingly, the driving force of capitalist development is identified in the risky R&D investments of the large trusts – undertaken only if they expect proper returns to be earned. To protect these returns from being competed away immediately, the large, innovative corporations tend to engage in monopolistic practices. Such practices are incompatible with the ideal of perfect competition, but without them there would be significantly fewer R&D investments and innovations. Moreover, Schumpeter (1942, ch. 8) claims that monopolistic practices work for only a limited time before

innovations are eventually imitated or invalidated by rival innovations. Despite temporary monopolistic practices, competition by innovation thus boosts economic growth and raises prosperity more than fiercer price competition could ever do. This notion of ‘Schumpeterian competition’ induced a long debate about the relationships between firm size, market structure and innovativeness in which, however, the broader concept of endogenous economic change was lost from sight.

Endogenous change returns to centre stage in Nelson and Winter's (1982) neo-Schumpeterian restatement of evolutionary economics that blends Schumpeter's ideas with Darwinian concepts on the one hand and elements of the behavioural theory of the firm on the other. Schumpeter (1942) had not been specific about the innovative operations of the large corporations. To fill the gap, Nelson and Winter assume that, because of bounded rationality, firms operate on the basis of organizational routines. Different firms develop different routines for producing, investing, price setting, using profits, searching for innovations, and so forth, resulting in a diversity of competitive behaviours in the industry. By analogy with the principle of natural selection, Nelson and Winter argue that this diversity tends to be eroded whenever competing routines lead to differences in the firms’ market performance and profitability. The better the firms perform, the more likely they are to grow, and the less reason they have to change their routines. The opposite holds for poorly performing firms. Much as differential reproductive success raises the share of better adapted genes in the gene pool of a population, differential firm growth thus raises the relative frequency of the better adapted routines in the ‘routine pool’ of the entire industry.

Instead of being a matter of optimal, deliberate substitution between given alternatives, in this view, the firms’ competitive adaptations to changing market conditions are forced on them by selection processes operating on their routines. However, in a Schumpeterian spirit, Nelson and Winter also account for innovative moves – a breaking away from old routines – in an industry's response to changing market conditions. New ways of doing things, for example in responding to rising input prices, are established by search processes which are themselves guided by higher-level routines. Modelled as random draws from a distribution of productivity increments, innovations raise the average performance of the industry and regenerate the diversity of firm behaviours for selection to operate on. Some of the firms are driven out of the market, while the surviving ones tend to grow. Under innovation competition, technology and industry structure thus co-evolve and feed a non-equilibrating economic growth process. Regarding the debate on Schumpeterian competition, Nelson and Winter's analysis suggests a reversal of cause and effect: a high degree of concentration within an industry (an indicator of monopolistic power) may evolve as a consequence of, rather than being a prerequisite for, a high rate of innovativeness in the industry.

Selection principles and processes

Analogies between natural selection and market competition are not new. Better-adapted variants of firm behaviour have often been argued to prevail in an industry just as better-adapted variants tend to prevail under natural selection pressure in the population of a species (an argument that has sometimes been misunderstood as vindicating profit-maximizing behaviour). The logic of the argument can be rendered more precise (Metcalfe, 1994). Consider an industry with firms $i = 1, \ldots, n$ producing a homogeneous output with unit cost $c_i = \text{const}$. Assume that the firms use different organizational routines which result

in a non-degenerate unit cost distribution. Let $s_i(t)$ denote the market share of firm i at time t measured by output. In a competitive market in which trade takes place at a uniform price $p(t)$,

$$p(t) = c(t) = \sum_i s_i(t) \cdot c_i, \tag{1}$$

with $c(t)$ as the average level of unit cost in the industry. By eq. (1), the average profit in the industry is zero. For at least one firm i, however, individual profit $\pi_i = p(t) - c_i > 0$ unless the entire market is served by the firm with the lowest level of unit cost.

Let the firm's growth be expressed in terms of the rate of change of its market share $(ds_i(t)/dt)/s_i(t)$ that is assumed to be a monotonic function φ of the firm's profit. With (1) inserted into the individual profit equation, the rate of change of the firm's market share can therefore be written as

$$\frac{ds_i(t)\,/\,dt}{s_i(t)} = \phi(c(t) - c_i) = \phi(\pi_i(t) - \pi(t)). \tag{2}$$

Hence, performance differences across firms and their routines translate into corresponding differential growth rates of the firms.

The ‘replicator’ eq. (2) corresponds to what is called ‘Fisher's principle’ in population genetics (Hofbauer and Sigmund, 1988, ch. 3). Let the fitness of an organism carrying a certain genetic trait be a constant. If it exceeds the average fitness in a population, the relative share of that trait in the population increases and vice versa. Consequently, natural selection raises average fitness over time to the level of the highest individual fitness. The change of the mean population fitness is proportional to the variance of the individual fitness. Analogously, with $c(t)$ as the measure for ‘population fitness’ in eq. (2), $dc(t)\,/\,dt = f(\mathrm{Var}(c_i)) \leq 0$.

If individual fitness is not constant, Fisher's principle no longer applies. Suppose individual unit costs decrease with the firms’ output, for example because of scale economies. The replicator equation can then have several fixed points representing multiple selection equilibria associated with a different average cost level (Metcalfe, 1994). Which of the multiple equilibria the process converges to – and, consequently, whether the *ex ante* most profitable cost practice is eventually selected – depends on the initial conditions. Selection does not necessarily drive fitness or, for that matter, profits to the largest maximum. (Replicator equations with multiple equilibria can also result if the individual fitness terms depend on the population shares of their carriers. Such a frequency dependency is characteristic of models in evolutionary game theory; see Hofbauer and Sigmund, 1988, ch. 16.)

To influence the underlying distribution of traits or behaviours, selection requires sufficiently inert conditions. In economic transformation processes this condition is often systematically violated. For example, firms facing a declining market share and/or profitability have strong incentives to modify their operations, that is, to replace inferior routines and/or to search for innovations. In general, with innovations playing a central role – as in Schumpeterian capitalist development – the volatility of the firms' environment increases and makes inertia rather unlikely. Industry dynamics are then more likely to be shaped by the generation and diffusion of innovations following their own time patterns rather than by selection processes. While in the case of selection processes theorizing focuses at the population level ('population thinking'), the explanation of the generation and diffusion of innovations can benefit from reconstructing motives and capabilities at the individual level.

Emergence and diffusion of innovations

Important as innovations are for economic transformation processes, the possibilities for analysing how they emerge are limited because the underlying cognitive processes are basically unknown. What can nonetheless be analysed is why and when agents are motivated to search for innovations, provided their motivation is not made contingent on the – as yet unknown – outcome of the search (as in models of optimal choices between known alternatives that are therefore not applicable here). Often search motivation is triggered by a state of dissatisfaction or deprivation that the agents want to overcome by actions still to be found. Among the causes may be unsatisfied curiosity, a motivation to achieve something (Schumpeter, 1934), or an agent's aspiration level that is temporarily not satisfied (Nelson and Winter, 1982, ch. 9). Where individual motivations like these occur in an uncorrelated way, they induce a base rate of innovative activity in the economy. If, in contrast, search motivation arises in a correlated way, for example in an economic crisis or when an industry is exposed to major innovations, the rate of innovative activities can rise far above the base rate. This is the case, for example, when firms need to innovate or be fast imitators with sufficient absorptive capacity in order to survive and therefore routinely engage in R&D.

Once an innovation is created or discovered by an agent, its implications can be grasped. Suppose, after assessing its benefits and costs, an agent implements an innovation. The implementation can usually be observed by competitors and/or other potential users. Since, in the absence of independent, own experience, people often draw conclusions from observing what others do, some observers may thus infer that the innovation is profitable and may start imitating it. Other observers may draw this conclusion only after a number of competitors and/or potential users have also signalled that they expect to benefit from adopting the innovation. Observational learning of this kind implies a dependency of the individual imitation or adoption behaviour – and, hence, the diffusion of the innovation – on the relative frequency of adopters.

The logic of this dependency can be captured by a function $q(t) = g(F(t))$, depicting the probability $q(t)$ that an agent who decides in t will adopt the innovation against the relative frequency of adopters $F(t)$ at time t. For $q(t) > F(t)$ the expected relative share of adopters grows with each additional decision and vice versa for $q(t) < F(t)$. The diffusion dynamics

$$\frac{dF(t)}{dt} = q(t) - F(t) \tag{3}$$

therefore hinge on the shape of the function g. For the quadratic function $q(t) = aF(t) - aF(t)^2$, $a > 1$, for instance, $F(t)$ converges to a fixed point F^a, $0 < F^a \square 1$, that depends on the size of a. (By integration of eq. (3) the diffusion path can in this case be shown to follow the well-known S-shaped logistic trend.) For the cubic function $q(t) = 3F(t)^2 - 2F(t)^3$, to take that example, the condition $q(t) = F(t)$ is satisfied if F equals 0, $\frac{1}{2}$, or 1. Inserting the cubic function into eq. (3), $F = 0$ and $F = 1$ can be shown to represent stable fixed points of eq. (3) while $F^* = \frac{1}{2}$ represents an unstable fixed point. This implies that for $F(t) < F^*$ the probability of adopting the innovation is too small to induce a spontaneous diffusion process. If $F(t)$ were for some reason to exceed F^* – representing a ‘critical mass’ of adopters – the innovation would however spread. The reason could be fluctuations of $F(t)$ that randomly cumulate, but are not represented in this simple deterministic model. (This explanation also plays a role in evolutionary game theory where the question is, for example, whether a new convention can emerge in a coordination game; see Young, 1993.) Another reason could be that somebody organizes a collective action by which the critical mass of agents is made to believe that more than the share F^* of agents will adopt the innovation.

With major technological innovations, competing variants or designs that serve the same user needs are often spawned simultaneously. The diffusion processes of the competing variants are interdependent if, for each of the variants, the users’ utility varies with the number of adopters. Such ‘economies to adoption’ of alternative variants have been diagnosed, for example, for electric current transmission, video recorder systems, or the layout of typewriter keyboards. The underlying pattern is again a frequency-dependency effect that can be analysed as before, if only two rival variants are assumed and the decision of agents who adopt neither of these is neglected.

Let $q(t)$ denote the probability of adopting the first variant and $F(t)$ its share of adopters at time t. Suppose both variants become available simultaneously and offer the same inherent benefits. For the first variant the development is captured by the cubic function above, interpreted as the mean process of a stochastic adoption process. With an identical number of initial adopters, $F(0) = F^* = \frac{1}{2}$ and $q(t) = \frac{1}{2}$. Once $F(t) \neq \frac{1}{2}$ for $t > 0$, economies to adoption raise the individual adoption probability of one of the variants over that of the other. As a consequence, the realization of the stochastic diffusion process initially fluctuates around F^*. Over time, however, small historical events and cumulative random fluctuations drive the process in the direction of either $F = 0$ (first variant disappearing) or $F = 1$ (second variant disappearing). In competitive diffusion processes of this kind, the prevailing state of the technology is thus ‘path-dependent’, and the process can be ‘locked in’ to the one variant if it is assumed, in addition, that over time the number of adopters grows beyond all bounds (Arthur, 1994, ch. 3). This means that, for $t \to \infty$, the likelihood of passing F^* by cumulative random fluctuations goes to

zero.

The evolution of industries and the institutions backing innovativeness

The substitution processes that the diffusion of new products and techniques induces shake up the established production structures. Factor owners and producers are forced to make adjustments – often painful ones that depreciate earlier investments and acquired competencies. While such ‘pecuniary externalities’ are inevitable concomitants of innovations, the longer-run consequence of innovativeness is – as Schumpeter (1942) had postulated – a rising standard of living of the masses. As a result of innovativeness, labour productivity and per capita income increase. New products and services absorb the growing consumption expenditures where established markets tend to be satiated. New employment opportunities emerge in new industries. To understand the working of the innovative transformation process and its policy implications, it is often useful to reconstruct the historical record of the evolution of entire industries (Malerba et al., 1999). Many of them, like the auto industry or the computer industry, grow out of a few major innovations for which new markets can be established or existing ones can substantially be expanded. Industries continue to grow over time under the pressure of imitative competition, often following a path of technical improvements that evolves within a ‘technological paradigm’ (cf. Dosi, 1988).

Such regular patterns of change at the industry level can for many, though not all, industries be characterized in a stylized way by a life-cycle metaphor (Klepper, 1997). Soon after their markets have been established by early innovators, the industries experience heavy entry and exit activities by competitors who partly imitate and partly add new varieties. While the market is expanding, a drastic shake-out in the number of firms occurs so that eventually a few large firms dominate the industry, and diversity in products and processes is reduced. In the beginning, product innovations are a main source of competitive advantages. Over time, however, the importance of process innovations increases. They raise productivity, drive down unit costs, and tend to intensify price competition. One cause of these patterns of industry evolution seems to be increasing returns to process innovations. These favour first movers that have been able to attain a sufficient size to spread development costs over larger output bases. With fiercer price competition, the firms with higher unit costs tend to be driven out of the industry, as in the selection model discussed above. Market concentration rises. With fewer innovations at that stage in the industry, its growth slows down, if the industry is not stagnating or declining.

Industry evolution is often connected with spatial effects. Innovative production techniques and new products often grow out of initiatives, competencies, endowments, and institutional settings in particular locations (Antonelli, 2001). If such complementary and interdependent local innovative activities gain momentum and trigger a self-augmenting process of firm growth and firm founding activities in close spatial proximity, an ‘industrial cluster’ can emerge. During early phases of the industry life cycle, a substantial share of the corresponding national or international industrial innovative activity may even be concentrated in such locations, Silicon Valley being the paradigmatic case. In such regions, income and employment are boosted. For policymaking the question therefore arises under what conditions innovative industrial clusters emerge and how and when their emergence can be fostered (Brenner, 2004).

The early growth of innovative industries creates new employment opportunities. At later stages of the

industry life cycle, when price competition and substitution pressure from innovative industries force the industry to raise labour productivity to reduce costs, employment is usually gradually lost. (For this reason, an industrial cluster that dominates a region can, in later stages of the industry life cycle, become a drag on local employment and prosperity.) At the macroeconomic level, the stages reached in the life cycles of the industries interact in a complex way with productivity and income growth rates, and with the overall changes in employment (Metcalfe, Foster and Ramlogan, 2006). Although these interactions have not yet been fully explored, it seems clear that at least two conditions must be met to maintain a high level of aggregate employment. First, innovative industries with new employment opportunities must emerge at the right times to compensate for the labour-saving technical progress. Second, the workforce must be able to adjust to the qualification requirements of the innovative industries and technologies. Since there is no self-regulating mechanism fulfilling the first condition, and because of delays and frictions in satisfying the second condition, the evolution of the industries is not necessarily a smooth transformation process. Aggregate employment and domestic income can vary substantially with the pace at which innovative industries emerge and expand.

However, high levels of education and training are likely to raise innovativeness and the qualifications of the workforce. Ensuring this with an adequate institutional infrastructure – a productive national system of innovations – is an important policy option in supporting and smoothing the transformation process. This is even more true from a global perspective. A country's growth potential and its competitive advantage in trade hinge on when the country gains access to newly emerging technological opportunities and where in the innovative industries' life cycle it enters the market. History shows that differences between countries in this respect correspond to differences in their national innovation systems (Fagerberg, 2002).

Darwinian perspectives on economic evolution

The neo-Schumpeterian approach considers the concept of selection as constitutive for evolutionary economics. Economic selection processes, operating on the diversity of individual behaviours, force adaptations on populations of agents who are prevented by their bounded rationality from deliberately adapting optimally. The import of the selection concept is not meant to extend Darwinism to the economic domain. Such an extension was, however, advocated by Veblen (1898) under the influence of the Darwinian revolution of his time. He coined the term 'evolutionary' economics for such an approach (Hodgson, 2004). A Darwinian perspective on the economic domain can indeed help to clarify how evolutionary economics fits with the Darwinian world view now prevailing in the sciences and in this way offer new insights (Witt, 2003).

In the economic domain, the bulk of change to be explained occurs within single generations. In contrast, the Darwinian theory of natural selection focuses on inter-generational change and is therefore relevant only for explaining the basis on which economic evolution rests. These are, first, the long-term constraints man-made economic evolution is subjected to and, second, the innate dispositions and adaptation mechanisms in humans (shaped earlier in human history by natural selection) that define the basic behavioural repertoire. Veblen (1898) focused on habits, including habits of thought, which he assumed to emerge from hereditary traits and past experiences, given the traditions, conventions and material circumstances of the time. (Habits play the crucial role in Veblen's explanation of the

‘cumulative causation’ of institutions which, in turn, he regarded as the key to understanding the different forms of economic life and their genesis.)

In a similar vein one may focus on human preferences that emerge from the interplay of inherited dispositions and innate conditioning learning mechanisms – both of these shared by all humans with the usual genetic variance. A prominent example of innate dispositions is the altruistic attitudes that play a prominent role in evolutionary game theory (Hofbauer and Sigmund, 1988, ch. 14). Other examples of innate dispositions can be found in certain forms of consumption. The genetically fixed learning mechanism accumulates the influence of a lifelong history of reinforcement and conditioning. It is responsible for the emerging variety of individual preferences and keeps them changing over time.

Following Hayek (1988, ch. 1), innate behaviour can be conjectured to play a key role in the evolution of human institutions. They emerge, he argues, through social learning of ‘rules of conduct’ that starts from primitive, genetically fixed, forms of social behaviour and add on new elements by trial and error. Over their history, different groups or whole societies thus build up a diversity of rules that regulate their interactions. The group members’ innovativeness is channelled into economic activities provided institutional regulations do not discourage this or fail to protect the capital accumulation that is necessary to realize innovations. Those groups that succeed in developing and passing on rules able to better meet these conditions can therefore be expected to grow and prosper in terms of population size and per capita income. Their differential success may enable such groups to conquer and/or absorb less well-equipped, competing groups and thus propagate better adapted institutions.

Economic evolution is, of course, also shaped in an essential way by human intelligence. By cognitive learning, problem solving and inventiveness, knowledge about institutions, opportunities and technologies is created (Mokyr, 2002). In the longer run, the enabling effects of cumulative knowledge generation emerging over time matter more than the effects of economizing on scarce resources at each point in time. From a Darwinian perspective the most significant tendency in the use of cumulative knowledge is the manipulation of natural constraints to better accord them with human preferences. This has enlarged the niche for the human species and has improved living conditions for an ever-increasing number of its members. At the same time, however, knowledge accumulation has contributed to dramatically increasing the human share in the use of natural resources. According to Georgescu-Roegen's (1971) evolutionary approach to production theory, this way of solving problems implies a risky long-term impact on nature, the ultimate basis of the human economy. To account for these risks further innovative efforts that transform the economy from within seem indispensable.

See Also

- analogy and metaphor
- competition and selection
- deterministic evolutionary dynamics
- diffusion of technology
- learning and evolution in games: an overview
- path dependence
- Schumpeter, Joseph Alois

- Schumpeterian growth and growth policy design
- structural change
- Veblen, Thorstein Bunde

Bibliography

Antonelli, C. 2001. *The Microeconomics of Technological Systems*. Oxford: Oxford University Press.

Arthur, W. 1994. *Increasing Returns and Path Dependence in the Economy*. Ann Arbor: Michigan University Press.

Brenner, T. 2004. *Local Industrial Clusters – Existence, Emergence and Evolution*. London: Routledge.

Dosi, G. 1988. Sources, procedures, and microeconomic effects of innovation. *Journal of Economic Literature* 26, 1120–171.

Fagerberg, J. 2002. *Technology, Growth and Competitiveness*. Cheltenham: Edward Elgar.

Georgescu-Roegen, N. 1971. *The Entropy Law and the Economic Process*. Cambridge, MA: Harvard University Press.

Hayek, F. 1988. *The Fatal Conceit*. London: Routledge.

Hodgson, G. 2004. *The Evolution of Institutional Economics*. London: Routledge.

Hofbauer, J. and Sigmund, K. 1988. *The Theory of Evolution and Dynamical Systems*. Cambridge: Cambridge University Press.

Klepper, S. 1997. Industry life cycles. *Industrial and Corporate Change* 6, 145–81.

Malerba, F. Nelson, R., Orsenigo, L. and Winter, S. 1999. ‘History-friendly’ models of industry evolution: the computer industry. *Industrial and Corporate Change* 8, 3–40.

Metcalfe, J. 1994. Competition, Fisher's principle and increasing returns in the selection process. *Journal of Evolutionary Economics* 4, 327–46.

Metcalfe, S., Foster, J. and Ramlogan, R. 2006. Adaptive economic growth. *Cambridge Journal of Economics* 30, 7–32.

Mokyr, J. 2002. *The Gifts of Athena*. Princeton: Princeton University Press.

Nelson, R. and Winter, S. 1982. *An Evolutionary Theory of Economic Change*. Cambridge: Harvard University Press.

Schumpeter, J. 1934. *The Theory of Economic Development*. Cambridge: Harvard University Press.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. New York: Harper.

Veblen, T. 1898. Why is economics not an evolutionary science? *Quarterly Journal of Economics* 12, 373–97.

Witt, U. 2003. *The Evolving Economy*. Cheltenham: Edward Elgar.

Young, P. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

How to cite this article

Witt, Ulrich. "evolutionary economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000295> doi:10.1057/9780230226203.0511

The New Palgrave Dictionary of Economics Online

ex ante and ex post

Otto Steiger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand and supply; ex ante and ex post; expectations; Lindahl, E. R.; Myrdal, G.; Ohlin, B. G.; Stockholm School; temporary equilibrium

Article

The concepts of *ex ante* and *ex post* are the most popular terminological innovations developed by the famous so-called Stockholm School in the 1930s. The terminology was introduced into macroeconomic theory, especially with regard to the savings-investment relation by Gunnar Myrdal (1933; 1939) and clarified and incorporated into sequence or period analysis by Erik Lindahl (1934; 1939b), whose conceptual system of ‘prospective’ and ‘retrospective’ values achieved ‘world-citizenship’ as a method for drawing up national budgets (Hansen, 1951, p. 27). The popularization of the method of *ex ante* and *ex post* is due to Ohlin's seminal articles on the Stockholm School (1937) which made it ‘generally accepted over the whole world with a rapidity unusual to economics’ (Palander, 1941, p. 34).
The significance of the distinction between *ex ante* and *ex post* ‘as one of the most transforming insights that theoretical economics has had’ (Shackle, 1972, p. 440) does not follow so much, as often stressed in the literature, from the simple fact that there exist always two alternative definitions of flow-related economic magnitudes like income, production, and so on, depending on whether they are looked at ‘from before’ or ‘from after’. The central idea of the necessity to distinguish between *ex ante* and *ex post* stems rather from the recognition of the fundamental difference, originally expressed by Frank Knight (1921, pp. 35 f.) and definitely formulated by Myrdal (1939, pp. 59 f.), between ‘foreseen’ and ‘unforeseen’ changes where only the latter result in ‘gains and losses’ which, as shown by Lindahl (1939b, pp. 103 f.), have to be ‘windfalls’. Therefore, in the analysis of expectations under uncertainty time has to be included in an essential way by two alternative methods of calculation of economic variables: (i) an *ex ante* computation or business calculation which refers to a point of time at the beginning of a period and (ii) an *ex post* computation or bookkeeping referring to the development in time at the end of the period (Myrdal, 1939, pp. 45–7). As a consequence, economic analysis can be

divided into (i) an *ex ante* analysis explaining how expectations determine an economic magnitude and (ii) an *ex ante/ex post* analysis explaining the possible divergence between the expected and realized value of this variable.

The emergence of the concepts of *ex ante* and *ex post* can be dated to Lindahl's and Myrdal's early writings in the 1920s. In his first treatise in macroeconomic theory (1924, ch. 3) Lindahl stressed the time factor for economic analysis and used the notion of ‘subjective calculations of the future’ as well as the term *ex post* when he discussed ‘a negative investment recognized only ex post’ (p. 33). A more coherent analysis of these concepts was given in Myrdal's dissertation on expectations and price changes (1927, pp. 67 f.) where he showed, emphasizing Knight's idea of the difference between certain and uncertain changes, how divergences between incomes and costs of an investment calculated ‘*before*’ will be balanced by gains and losses calculated ‘*after*’.

The first application of these ideas to macroeconomic problems was made by Lindahl in *Penningpolitikens medel* (1930; cf. 1939a). However, the dynamic method of temporary equilibrium used in this treatise, that is, ‘an analysis dividing time into a number of short equilibrium periods during which no changes occur’, led to a ‘theoretically inadmissible mixture of the *ex ante* and *ex post* analysis’ (Myrdal, 1939, p. 122). In case of the same but wrong expectations, for example, the equality between savings and investment *ex ante* is not a guarantee for temporary equilibrium (Palander, 1941, p. 44; see also Siven, 2006, pp. 684–5). As shown by Myrdal in the original Swedish version of *Monetary Equilibrium* (1932, pp. 228–30), this mixture was especially obvious in Lindahl's discussion of the relation between investment and saving, where he could not demonstrate in a satisfactory way how an initial discrepancy due to a shift in the rate of interest will always be balanced by changes in the distribution of income between borrowers and lenders. If, however, Lindahl would give up his method of temporary equilibrium, that is, allow for disequilibrium during a period, his analysis could be interpreted in an *ex ante/ex post* framework.

It was exactly this disequilibrium analysis which enabled Myrdal to clarify the relation between investment and saving in his three different versions of *Monetary Equilibrium*, where he introduced the notions of *ex ante* and *ex post* first in the German edition (1933, § 29). In his discussion of these concepts Myrdal (1939, pp. 59–62, 116–25; cf. 1933, §§ 32, 55–6) allowed for a discrepancy between investment and saving *ex ante* based on ‘anticipatory calculations’ at a point of time demarcating the beginning of a period, while at its end their values were constructed by ‘a subsequent “bookkeeping”’in such a way that there is *always* an *ex post* balance ‘regardless of how short the period’. Therefore, it is not ‘this meaningless balance’ which is of interest to economic analysis but ‘*the very changes during the period which are required to bring about this ex post balance*’. Myrdal assumed that these balancing factors arise out of ‘unanticipated changes’ in ‘revenues and costs’, that is, in incomes, during the period for which they can be calculated only *ex post:* gains and losses. The reason why *ex ante* and *ex post* values may differ is that expectations formulated in the beginning of a period are *ex ante* values of prices and quantities that may not be realized because expectations may be disappointed during the period (Siven, 2006, p. 681). As later shown by Lindahl (1939b, pp. 103 f.; cf. 1958), these values have to be windfalls and must not be confused, as sometimes in Myrdal's analysis (1939, p. 65), with entrepreneurial gains or losses, which are already included in the *ex ante* values and which, therefore, cannot serve as balancing factors.

Although Myrdal always spoke of ‘income changes’ as the balancing factor *ex post* of discrepancies *ex*

ante between investment and saving, his examples implied almost exclusively '*price changes*' (1939, p. 60; cf. Palander, 1941, pp. 42f.; Hansson, 1982, p. 149). It was left to Lindahl (1934, 1939b) to demonstrate how the *ex post* equality was achieved in a disequilibrium process via a change in quantities. Lindahl presented his solution in an aggregate demand and supply framework, where demand was identified with the purchase plans of consumers and producers and supply with the expectations of the sellers of production and consumption goods. In his analysis he made the fundamental assumption that the purchase plans as well as the supply prices of the sellers at the beginning of a period 'have been actually realized during the period' (1934, p. 207, cf. 1939b, p. 92). Under this assumption a possible deviation between expected and realized sales, which Lindahl took as given if the future is not foreseen with certainty, must be considered as a result of a difference between investment and saving *ex ante* which in turn will cause a divergence between expected and realized total real income. These changes in income represent gains or losses to the producers which in a form of 'unintentional', not 'forced', saving or dissaving equalize investment and saving *ex post*.

The purpose of Lindahl's rigid assumption that purchase plans are always fulfilled, which made it impossible to apply his analysis to the conditions of full employment (Hansen, 1951, pp. 29–32), was to demonstrate that, once prices are given, the *actions* of the economic subjects during the period 'can be directly deduced from the plans at the beginning of the period' (1939b, p. 92). With this demonstration Lindahl had taken the first step to a sequence analysis, that is, a single-period analysis where *ex ante* plans determine *ex post* results. The second step consisted of a continuation analysis where the *ex post* events of the current period lead to revisions of the *ex ante* plans for the consecutive period at the transition point between these periods, 'especially as regards the supply prices and the producers' and consumers' demand' (1934, p. 211). However, as Lindahl 'never succeeded in formulating 'laws of motion' for revisions of plans … this promising branch of dynamic theory became abortive' (Hansen, 1966, p. 3).

Of greater influence for economic analysis was Lindahl's second contribution to the development of the *ex ante/ex post* method, his discussion of the relations between 'prospective' and 'retrospective' values of micro- and macroeconomic variables (1939b) which contained 'the germ of many lines in later works' on the methodology of national accounting (Ohlsson, 1953, p. 266). Although Lindahl's accounting structure was criticized as 'deficient' (Ohlsson) in the treatment of government accounting, it has been emphasized recently by Hicks (1985, p. 80) that Lindahl's system 'does have some continuing merits … for the accounting of the public sector': 'In this field at least, it may still be contended, *ex ante ex post* remains respectable.'

For a long time it was argued that the exposition of the Keynesian system 'requires the language of *ex ante* and *ex post*' (Shackle, 1972, p. 172; see also Patinkin, 1976, pp. 139–40; Siven, 2006, p. 700), with the emphasis placed on the possible divergences, due to uncertainty, between disappointed *ex ante* expectations and *ex post* results as one of the relevant factors in determining the level of employment. However, the posthumous publication of Keynes's 1937 lecture notes have shown that Keynes (1937a, p. 183; see also 1937b) emphasized that even under the assumption of an 'identity of *ex post* and *ex ante*', that is, with expectations always fulfilled but without having to assume for this case, as did Myrdal and Lindahl, the absence of uncertainty, 'the theory of effective demand is substantially the same' (p. 181). Moreover, as Keynes regarded the '*time* relationship' between the concepts of *ex ante* and *ex post* as 'incapable of being made precise' (p. 179), he rejected this method as an inadequate tool in handling the

problems of uncertainty and time.

See Also

- Lindahl, Erik Robert
- Myrdal, Gunnar
- Stockholm School

Bibliography

Hansen, B. 1951. *A Study in the Theory of Inflation*. London: Allen & Unwin.

Hansen, B. 1966. *Lectures in Economic Theory. Part I: General Equilibrium Theory*. Lund: Studentlitteratur.

Hansson, B. 1982. *The Stockholm School and the Development of Dynamic Method*. London: Croom Helm.

Hicks, J. 1985. *Methods of Dynamic Economics*. Oxford: Clarendon Press.

Keynes, J.M. 1937a. Ex post and ex ante. Notes from the 1937 lectures. In *The General Theory and After*, Part II: *Defence and Development*, vol. 14 of *The Collected Writings of John Maynard Keynes*, ed. D. Moggridge. London: Macmillan, 1973.

Keynes, J.M. 1937b. The ‘ex-ante’ theory of the rate of interest. *Economic Journal* 47, 663–69.

Knight, F.H. 1921. *Risk, Uncertainty, and Profit*. Chicago: University of Chicago Press, 1971.

Lindahl, E. 1924. *Penningpolitikens mål och medel. Del I* [The aims of means of monetary police. Part I]. Lund: Gleerup; Malmö: Försäkringsaktiebolaget.

Lindahl, E. 1930. *Penningpolitikens medel* [The means of monetary policy]. Lund: Gleerup; Malmö: Försäkringsaktiebolaget; enlarged version of 1st edn, 1929; revised version trans. as Lindahl (1939a).

Lindahl, E. 1934. A note on the dynamic pricing problem. Mimeo, Gothenburg, 13; quoted from the corrected version published in Steiger (1971).

Lindahl, E. 1939a. The rate of interest and the price level. In Lindahl (1939c); revised version of Lindahl (1930).

Lindahl, E. 1939b. Algebraic discussion of the relations between some fundamental concepts. In Lindahl (1939c).

Lindahl, E. 1939c. *Studies in the Theory of Money and Capital*. London: Allen & Unwin.

Lindahl, E. 1958. The concept of gains and losses. In *Festskrift til Frederik Zeuthen*, Copenhagen: Nationalkonomisk Forening.

Myrdal, G. 1927. *Prisbildningsproblemet och färänderligheten* [The problem of price formation and changeability]. Uppsala and Stockholm: Almqvist & Wiksell.

Myrdal, G. 1932. Om penningteoretisk jämvikt. En studie över den ‘normala räntan’ i Wicksells penninglära [On monetary equilibrium. A study on the ‘normal rate of interest’ in Wicksell's monetary thought]. *Ekonomisk Tidskrift* 33(5–6) (1931; printed 1932), 191–302; revised version trans. as Myrdal (1933).

Myrdal, G. 1933. Der Gleichgewichtsbegriff als Instrument der geldtheoretischen Analyse. In *Beiträge zur Geldtheorie*, ed. F.A. Hayek. Vienna: J. Springer; 1st revised version of Myrdal (1932); 2nd revised version trans. as Myrdal (1939).

Myrdal, G. 1939. *Monetary Equilibrium*. London: Hodge; revised version of Myrdal (1933).

Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment I–II. *Economic Journal* 47, 53–69, 221–40.

Ohlsson, I. 1953. *On National Accounting*. Stockholm: Konjunkturinstitutet.

Palander, T. 1941. Om ‘Stockholmsskolans’ begrepp och metoder. Metodologiska reflexioner kring Myrdals ‘Monetary Equilibrium’. *Ekonomisk Tidskrift* 43(1), 88–143; quoted from and trans. as ‘On the concepts and methods of the “Stockholm School”: some methodological reflections on Myrdal's “Monetary Equilibrium”. *International Economic Papers* No. 3 (1953), 5–57.

Patinkin, D. 1976. *Keynes's Monetary Thought: A Study of Its Development*. Durham, NC: Durham University Press.

Shackle, G.L.S. 1972. *Epistemics & Economics. A Critique of Economic Doctrines*. Cambridge: Cambridge University Press.

Siven, C.-H. 2006. Monetary equilibrium. *History of Political Economy* 38, 668–709.

Steiger, O. 1971. *Studien zur Entstehung der Neuen Wirtschaftslehre in Schweden. Eine Anti-Kritik*.

Berlin: Duncker & Humblot.

How to cite this article

Steiger, Otto. "ex ante and ex post." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000142> doi:10.1057/9780230226203.0512

The New Palgrave Dictionary of Economics Online

excess burden of taxation

James R. Hines Jr.

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The excess burden of taxation is the efficiency cost, or deadweight loss, associated with taxation. Excess burden is commonly measured by the area of the associated Harberger triangle, though accurate measurement requires the use of compensated demand and supply schedules. The generation of empirical excess burden studies that followed Arnold Harberger's pioneering work in the 1960s measured the costs of tax distortions to labour supply, saving, capital allocation, and other economic decisions. More recent work estimates excess burdens based on the effects of taxation on more comprehensive measures of taxable income, reporting sizable excess burdens of existing taxes.

Keywords

Dupuit, A.-J.-E. J.; excess burden of taxation; Harberger triangle; Jenkin, H. C. F.; path dependence; tax evasion; taxable income

Article

The excess burden of taxation is the efficiency cost, or deadweight loss, associated with taxation.
The total economic burden of a tax includes both payments that taxpayers make to the government and any lost economic value from inefficient activities undertaken in reaction to taxes. Since direct tax burdens take the form of revenue that taxpayers remit to governments, the excess burden of taxation is the magnitude of the economic costs of accompanying economic distortions. For example, a tax on labour income typically discourages work by encouraging inefficient substitution of untaxed leisure for taxed paid work. At low tax rates this substitution entails only modest excess burdens, since, in the absence of other distortions, the welfare cost of substituting an untaxed for a taxed activity simply equals the tax rate, the difference between pre-tax and after-tax returns to the taxed activity. At high tax rates this difference is quite large, and as a result residents of economies with high tax rates may face substantial excess burdens of taxation. Indeed, it is entirely possible for the excess burden of a tax to exceed the revenue collected; a tax imposed at so high a rate that it eliminates the taxed activity clearly

has this feature.

The excess burden of taxation is commonly measured by the area of the associated 'Harberger triangle' (Hines, 1999). The base of the Harberger triangle is the amount by which economic behaviour changes as a result of price distortions introduced by the tax, and the height of the Harberger triangle is the magnitude of the tax burden per unit of economic activity.

The many excess burdens

One of the difficulties that arise in evaluating the excess burden of taxation is that there is more than one possible measure of excess burden. This multiplicity does not imply that all measures are equally desirable or useful. For example, the use of uncompensated (Marshallian) demand and supply curves to construct Harberger triangles produces measures of the excess burden of taxation with a number of known problems. In the (realistic) case in which a government uses multiple taxes, a measure of total excess burden based on uncompensated demand and supply curves is path dependent, meaning that its value depends on the order in which the taxes are imagined to be imposed. As the order of the taxes is perfectly arbitrary, path dependence is troubling – most importantly because it reflects the imprecision of excess burden measures constructed in this way.

Path dependence is one consequence of this imprecision; another is that a tax system that produces a higher level of economic welfare might have a greater measured excess burden than an alternative that raises the same revenue. If excess burden is to be useful in the evaluation and formation of tax policies, it is necessary that the measure should correspond, at least approximately, to the economic cost of taxation – and assign greater excess burden to tax systems that are in fact more burdensome.

Path dependence and inaccurate welfare orderings need not arise if excess burden is measured by Hicksian consumers' surplus, based on schedules that hold utility, rather than income, constant as prices vary. Because actual tax policy changes typically do not hold utility constant, it is necessary to construct a measure based on a conceptual experiment that does. One intuitive experiment is to imagine that, as a tax is imposed, utility is held constant at its pre-tax level. Excess burden is then defined as the amount, in excess of tax revenue, that the government must compensate consumers to maintain initial utility in the face of a tax-induced price change. The amount of compensation, which corresponds to the Hicksian measure of the *compensating variation* of the price change, may be calculated in roughly the same way that Harberger triangles are commonly measured.

An alternative conceptual experiment is to begin with the tax already in place and then remove it, extracting from consumers in lump-sum fashion an amount that prevents them from changing their utility levels while the tax is removed. Because the initial tax is distortionary, it is necessary to extract more from consumers than the tax revenue, the difference representing the excess burden of the initial tax. This differs from the previous measure in corresponding to a Hicksian *equivalent variation* measure of excess burden. One virtue of an equivalent variation measure of excess burden, compared to the compensating variation measure, lies in the fact that, in comparing tax systems that raise equal revenue, the tax system with the lowest excess burden as measured by equivalent variation also produces the highest level of consumer welfare (Kay, 1980).

Although these compensating variation and equivalent variation measures are the most intuitive, they are actually just examples drawn from a class of measures based on arbitrary levels of utility and arbitrary

reference price vectors. As King (1983) and others note, the use of compensated supply and demand schedules together with fixed reference price vectors guarantees that resulting excess burden measures have desirable properties, though the interpretation of the resulting magnitudes depends on the choice of utility levels and price vectors. These measures then can be naturally generalized to include marginal excess burden, the change in excess burden arising from a given tax change, and to treat excess burden in settings in which costs of production vary with output levels (Auerbach and Hines, 2002).

Empirical measurement of excess burden

While the theory of excess burden measurement has a long and colourful history that dates back to the 19th century contributions of Jules Dupuit (1844) and Fleeming Jenkin (1871–2), economists seldom measured actual excess burdens prior to the pioneering work of Arnold Harberger in the 1960s. In two influential papers published in 1964, Harberger (1964a, 1964b) derived an approximation used to measure excess burden and (1964b) applied the method to estimate excess burdens of income taxes in the United States. Harberger shortly thereafter (1966) produced estimates of the excess burden of US capital taxes. A generation of empirical studies by other scholars followed the publication of Harberger's subsequent survey article (1971).

The empirical work that followed Harberger's efforts focused on the use of simple excess burden formulas to estimate the welfare impact of a wide array of tax-induced distortions, including those to labour supply (Browning, 1975), saving (Feldstein, 1978), corporate taxation (Shoven, 1976), and the consumption of goods, such as housing and non-housing consumption items, that are taxed to differing degrees (King, 1983). In addition, some attention was devoted to refining the approximations used in applying estimated behavioural parameters to calculate excess burdens. A variant of the excess burden formula used by Harberger, in which a form of uncompensated demand is used in place of compensated demand, approximates a compensated measure of welfare change. One question of interest to subsequent investigators is the practical difference between results obtained using Harberger-style approximations and those available from more exact measures. As Mohring (1971) and subsequent authors note, it is often the case that the same demand information necessary to calculate approximations can, if properly modified, be used to calculate Hicksian excess burden measures. The extent to which these two methods generate different answers is, of course, an empirical question. Rosen (1978) finds that measures of excess burden based on compensated and uncompensated demand and supply schedules track each other rather closely, but Hausman (1981) offers some examples in which they differ considerably.

A major practical difficulty in measuring the excess burden of a single tax, or of a system of taxes, is that excess burden is a function of interactions that are potentially very difficult to measure. For example, a tax on labour income is expected to affect hours worked, but may also affect the accumulation of human capital, the intensity with which people work, the timing of retirement, and the extent to which compensation takes tax-favoured (for example, pensions, health insurance, and workplace amenities) in place of tax-disfavoured (for example, wage) form. In order to estimate the excess burden of a labour income tax, it is in principle necessary to estimate the effect of the tax on these and other decision margins. Analogous complications are associated with estimating the excess burdens of most other taxes. In practice, it can be very difficult to obtain reliable estimates of the impact of

taxation on just one of these variables.

It is in reaction to the complicated nature of the problem of separately estimating the effect of taxation on all of a taxpayer's decision margins that a number of recent studies estimate excess burdens based on the effects of taxation on reported taxable income. Taxable income incorporates not only any effects of taxation on work effort, but also tax avoidance of various forms, including deliberate hiding of income and legal avoidance such as making tax-deductible charitable contributions. Properly measured, excess burden, as calculated by the effect of taxation on taxable income, should accurately capture all the necessary interactions to evaluate the welfare consequences of taxation (Feldstein, 1999).

Several empirical studies, including Feldstein (1995), Auten and Carroll (1999), and Goolsbee (2000), consider the responsiveness of taxable income to tax rates, relying on major US tax changes to provide variation in tax rates. The evidence indicates that taxable income is generally quite responsive to tax changes, particularly among the high-income population, thereby implying an excess burden of US taxes considerably greater than that produced by studies using estimated effects of taxation on work hours and saving. The estimates suggest excess burdens of taxation that might be as high as 75 per cent of tax revenue collected (Feldstein, 2006), though there is still considerable uncertainty over its true magnitude.

See Also

- neutral taxation
- optimal taxation
- Pigouvian taxes

Bibliography

Auerbach, A.J. and Hines, J.R., Jr. 2002. Taxation and economic efficiency. In *Handbook of Public Economics*, vol. 3, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Auten, G. and Carroll, R. 1999. The effect of income taxes on household income. *Review of Economics and Statistics* 81, 681–93.

Browning, E.K. 1975. Labor supply distortions of social security. *Southern Economic Journal* 42, 243–52.

Dupuit, A.J.É.J. 1844. De la mesure de l'utilité des travaux publics. *Annales des ponts et chaussées*, 2nd ser., 8. Tr. R.H. Barback as ‘On the Measurement of the Utility of Public Works’, *International Economic Papers* 2 (1952), 83–110. Repr. in *Readings in Welfare Economics*, ed. K.J. Arrow and T. Scitovsky. Homewood, IL: Richard D. Irwin, 1969.

Feldstein, M. 1978. The welfare cost of capital income taxation. *Journal of Political Economy* 86, S29–S51.

Feldstein, M. 1995. The effect of marginal tax rates on taxable income: a panel study of the 1986 tax reform act. *Journal of Political Economy* 103, 551–72.

Feldstein, M. 1999. Tax avoidance and the deadweight loss of the income tax. *Review of Economics and Statistics* 81, 674–80.

Feldstein, M. 2006. The effects of taxes on efficiency and growth. Working Paper No. 12201. Cambridge, MA: NBER.

Goolsbee, A. 2000. What happens when you tax the rich? Evidence from executive compensation. *Journal of Political Economy* 108, 352–78.

Harberger, A.C. 1964a. The measurement of waste. *American Economic Review* 54, 58–76.

Harberger, A.C. 1964b. Taxation, resource allocation, and welfare. In *The Role of Direct and Indirect Taxes in the Federal Revenue System*, ed. J.F. Due. Princeton, NJ: Princeton University Press.

Harberger, A.C. 1966. Efficiency effects of taxes on income from capital. In *Effects of Corporation Income Tax*, ed. M. Krzyzaniak. Detroit: Wayne State University Press.

Harberger, A.C. 1971. Three basic postulates for applied welfare economics: an interpretive essay. *Journal of Economic Literature* 9, 785–97.

Hausman, J.A. 1981. Exact consumer's surplus and deadweight loss. *American Economic Review* 71, 662–76.

Hines, J.R., Jr. 1999. Three sides of Harberger triangles. *Journal of Economic Perspectives* 13, 167–88.

Jenkin, H.C.F. 1871–2. On the principles which regulate the incidence of taxes. *Proceedings of the Royal Society of Edinburgh* 7, 618–31. Repr. in *Papers, Literary, Scientific, &c. by the Late Fleeming Jenkin*, vol. 2, ed. S. Colvin and J. A. Ewing. London: Longmans Green, 1887.

Kay, J.A. 1980. The deadweight loss from a tax system. *Journal of Public Economics* 13, 111–20.

King, M.A. 1983. Welfare analysis of tax reforms using household data. *Journal of Public Economics* 21, 183–214.

Mohring, H. 1971. Alternative welfare gain and loss measures. *Western Economic Journal* 9, 349–68.

Rosen, H.S. 1978. The measurement of excess burden with explicit utility functions. *Journal of Political Economy* 86, S121–35.

Shoven, J.B. 1976. The incidence and efficiency effects of taxes on income from capital. *Journal of Political Economy* 84, 1261–83.

How to cite this article

Hines, James R., Jr. "excess burden of taxation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000315> doi:10.1057/9780230226203.0513

The New Palgrave Dictionary of Economics Online

excess volatility tests

Stephen F. LeRoy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Stock prices frequently undergo big changes that do not coincide with commensurate changes in fundamentals (earnings, dividends, interest rates). Shiller and LeRoy and Porter formalized the idea that price volatility is excessive relative to fundamentals by deriving the implications for price volatility of the hypothesis that stock prices equal the present value of discounted dividends. Subsequent discussion focused on the extent to which their results were subject to econometric problems. Also, analysts observed that the adopted version of the present-value model presumed constant discount rates, as would be the case under risk neutrality. This possibly biases the results.

Keywords

asset pricing; equity premium puzzle; excess volatility tests; payoff volatility; present value; price volatility; price–consumption ratio; risk aversion; risk neutrality; Samuelson, P.; vector autoregressions

Article

It has been known for many years that stock prices frequently undergo big changes that do not coincide with commensurate changes in prospective corporate earnings or dividends or in variables that can readily be connected to discount factors, such as interest rates (see for example Cutler, Poterba and Summers, 1989). The best-known episode occurred on 19 October 1987, when stock prices dropped around the world – by 22 per cent in the United States – in the complete absence of news about fundamentals. Such events appear to conflict with finance theory: if stock prices equal the present value of future expected dividend streams, changes in prices should be attributable to news about dividends or discount factors. However, it is difficult to draw reliable conclusions about price volatility from individual episodes, if only because there is no obvious way to evaluate statistical significance.

Is price volatility systematically excessive?

The question is not whether stock price volatility appears to exceed that justified by fundamentals in individual episodes, but whether it does so *systematically*. The latter question was first addressed by Shiller (1981) and LeRoy and Porter (1981). These papers, written independently and approximately contemporaneously, derived the existence of bounds on the volatility of prices and returns that are implied by the present-value relation. They found excess volatility. However, the papers made this point using different analytical methods. Shiller observed that, if stock prices equal the expectation of summed discounted dividends, then stock price volatility should be bounded above by the volatility of what he called ex-post rational stock prices, defined as actual summed discounted dividends. Ex-post rational prices p_t^*, he pointed out, obey the relation

$$p_t^* = \beta(p_{t+1}^* + d_{t+1}), \tag{1}$$

where β is a discount factor (assumed constant). From (1), a time series for p_t^* can be constructed by a backward recursion, at least given an initial condition. Shiller constructed graphs of p_t^* and p_t and argued from visual inspection of these graphs that the former was much smoother than the latter, proving that volatility is excessive. Since Shiller did not specify the model assumed to generate the data, he had no way to evaluate statistical significance.

LeRoy and Porter, in contrast, adopted a model-based analytical procedure. They specified that dividends, stock prices and any auxiliary variables that serve as predictors for future dividends are generated by a linear vector autoregression (to use a term that was not yet in vogue). They proved that a certain function of the variance of stock price and the variance of stock payoffs can be derived from the parameters describing the bivariate autoregression for dividends and prices. The reason both price volatility and dividend volatility enter the expression is that, if the auxiliary variables are accurate predictors of future dividend innovations, then price volatility will be high but payoff volatility will be low. The opposite will be the case if the auxiliary variables do not give accurate predictions of future dividends.

Using this result, LeRoy and Porter constructed a joint test of price volatility and payoff volatility from a bivariate model for dividends and prices. It was unnecessary to estimate the forecasting power of the auxiliary variables, or even to specify them. LeRoy and Porter conducted this test and reported a confidence interval based on the asymptotic distribution of the coefficients of the bivariate process for dividends and prices. They found excess volatility, but it appeared to be of borderline significance. See LeRoy (1989) for a fuller, but still brief, summary of the variance-bounds tests in the context of the efficient capital markets literature.

Econometric problems

Both Shiller's and LeRoy and Porter's procedures had econometric problems (for a survey of these problems, see Gilles and LeRoy, 1991). These problems were serious enough to invalidate the results, in the opinion of some. Discussion focused on Shiller's paper. Kleidon (1986a; 1986b) and Flavin (1983) pointed out that, while the present-value model implied that the unconditional variance of p_t^* exceeded that of p_t, one would expect the variance of p_t^* conditional on its neighbours p_{t+j}^* to be lower than the corresponding conditional variance of p_t. This is so because p_t^* is much more highly autocorrelated than p_t, as is evident from the absence of an error term in (1). In visually evaluating the volatility of p_t^* and p_t it is not easy to distinguish unconditional from conditional volatility. This is a major drawback of Shiller's procedure. Kleidon computed simulations of p_t^* and p_t in which the present-value model was true by construction, and argued that they looked much the same as the actual data.
LeRoy and Porter's procedure had the drawback that the assumed linear process for dividends and prices implies that these are stationary in level. That being so, some sort of trend correction to remove the upward trend in both variables must be imposed, and LeRoy and Porter did so. Trend-correction algorithms can easily distort the time-series properties of variables, and that may have happened in this case. It is possible that LeRoy and Porter's finding that excess price and return volatility is only marginally significant statistically reflects difficulties with the trend correction.

Interpreting excess volatility

Analysts questioned the interpretation of the variance-bounds tests along other lines. They pointed out that the volatility implications of the present-value relation were just repackaged versions of the fundamental implication of the present-value relation (plus the assumption of rational expectations) that stock returns should be serially uncorrelated. If so, why do direct tests of the return orthogonality implication of the present-value relation tend to accept the null, whereas the variance bounds tests appear to reject it? Part of the resolution of this apparent contradiction is that the evidence on return uncorrelatedness began to look less favourable to the null hypothesis in the late 1980s (see for example Campbell and Shiller, 1988). Another possibility, discussed by LeRoy and Steigerwald (1995), is that the volatility tests are more powerful than the return non-autocorrelatedness tests under whatever alternative hypothesis generated the data.

Risk aversion

Discussion of these issues ended fairly suddenly in the mid-1990s. This happened mostly because of a growing realization that the hypothesis being tested required an assumption of risk neutrality: in general, stock prices equal the discounted value of expected dividends (when the expectation is taken under the natural probabilities) with a non-stochastic discount factor only if agents are risk neutral). If agents are risk averse there is no reason to presume that either the orthogonality implications or the volatility implications of the present-value relation will be satisfied. This dependence on risk neutrality had not been brought out clearly in the major papers developing the orthogonality implications of market

efficiency. For example, Samuelson's otherwise superb paper (1965) developing the connection between martingale models and the present-value relation glossed over this point. Similarly, Fama's classic 1970 survey of the efficient capital markets literature observed that ‘market efficiency’ (meaning, presumably, rational expectations) could be tested only if the analyst committed himself to a particular model specifying how returns are generated, so that the joint hypothesis of efficiency and the assumed model is tested. Despite this, Fama did not go on to observe that the returns model that underlies conventional market efficiency tests took no account of risk aversion. Shortly it became clear that the non-autocorrelatedness of returns would not occur in general if agents are risk averse (LeRoy, 1973; 1976; Lucas, 1978).

Initially the dependence of the variance-bounds tests on risk neutrality appeared to be a somewhat abstract point. However, arguments were shortly presented that this might not be so. LaCivita and LeRoy (1981), using a two-state version of Lucas's (1978) tree model, showed that allowing for risk aversion could be expected to increase the predicted volatility of stock prices. Risk-averse agents will try to smooth consumption across time by transferring consumption from low-marginal-utility states to high-marginal-utility states. However, in an exchange economy they cannot do so in the aggregate. The representative agent must consume the aggregate endowment in equilibrium, so prices must counteract preferences. If stock prices are very high when the marginal utility of consumption is low, and vice versa, agents must buy financial assets when they are expensive and sell them when they are cheap if they are to transfer claims on consumption as desired. This price pattern decreases their desire to do so. If price volatility is high enough, this effect will induce agents to consume the endowment. A related argument was presented by Grossman and Shiller (1981).

Relation to the equity premium puzzle

The excess volatility debate shifted with the arrival of Mehra and Prescott's well-known paper (1985) on the equity premium puzzle. This paper was relevant to the excess volatility question for several reasons. The Mehra and Prescott paper followed LaCivita and LeRoy in specializing Lucas's tree model to two states, but modified the states so that they described the growth rate of the aggregate endowment rather than its level, as in Lucas and LaCivita and LeRoy. This leads to a tractable model when the representative agent has homothetic utility, as with power utility. A major advantage of Mehra and Prescott's specification is that when consumption growth rates rather than levels are stationary there is no need for trend correction.

Mehra and Prescott imposed drastic simplifications so as to obtain a tractable model. For example, their model did not distinguish among corporate earnings, dividends and aggregate consumption, despite the fact that these variables behave differently. Some analysts expected that Mehra and Prescott's finding that the equity premium is excessive would be reversed when these simplifications were reversed, but that has turned out not to be the case (see, for example, Kocherlakota, 1996).

LeRoy and Parke (1992) observed that Mehra and Prescott's framework can be adapted to the investigation of volatility by imposing the assumption that consumption growth rates are independently distributed. This is a special case of the Markov distribution that Mehra and Prescott specified. In that case the ratio of equity value to consumption follows a stationary process. The volatility of that variable depends on how much information agents have about future consumption beyond that contained in

present consumption. In the simplest case, when agents have no such information, price is a constant markup of consumption, implying that stock returns have the same volatility as the consumption growth rate. This is true regardless of the degree of risk aversion, implying that the implications of risk aversion for volatility are very different in stationary-growth-rate models than in the stationary-levels models discussed above. This prediction is rejected by the US data: the standard deviation of annual consumption growth is about two per cent, whereas that of annual stock returns is on the order of 20 per cent. In contrast to the equity premium puzzle, which can in principle be resolved with sufficiently high risk aversion, the prediction that equity returns should have standard deviations around two per cent holds for any level of risk aversion.

Even if one accepts that consumption is a geometric random walk, assuming that agents have no information variables for future consumption other than current consumption is unacceptable. If agents do have such information, equity prices will not be a constant markup of consumption. However, LeRoy and Parke showed that in that case the variances of the price–consumption ratio and the return on stock obey a relation similar to that obtained by LeRoy and Porter. They found that the resulting joint test on the volatility of the price–consumption ratio and the volatility of stock returns results in excess volatility for either variable or both.

Most analysts believe that no single convincing explanation has been provided for the volatility of equity prices. The conclusion that appears to follow from the equity premium and price volatility puzzles is that, for whatever reason, prices of financial assets do not behave as the theory of consumption-based asset pricing predicts.

See Also

- risk aversion

Bibliography

Campbell, J. and Shiller, R. 1988. The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195–228.

Cutler, D., Poterba, J. and Summers, L. 1989. What moves stock prices? *Journal of Portfolio Management* 15, 4–12.

Fama, E. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25, 283–417.

Flavin, M. 1983. Excess volatility in the financial markets: a reassessment of the empirical evidence. *Journal of Political Economy* 91, 929–56.

Gilles, C. and LeRoy, S. 1991. Econometric aspects of the variance bounds tests: a survey. *Review of Financial Studies* 4, 753–91.

Grossman, S. and Shiller, R. 1981. The determinants of the variability of stock prices. *American Economic Review Papers and Proceedings* 71, 222–7.

Kleidon, A. 1986a. Bias in small-sample tests of stock price rationality. *Journal of Business* 59, 237–61.

Kleidon, A. 1986b. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94, 953–1001.

Kocherlakota, N. 1996. The equity premium: it's still a puzzle. *Journal of Economic Literature* 34, 42–71.

LaCivita, C. and LeRoy, S. 1981. Risk aversion and the dispersion of asset prices. *Journal of Business* 54, 535–47.

LeRoy, S. 1973. Risk aversion and the martingale model of stock prices. *International Economic Review* 14, 436–46.

LeRoy, S. 1976. Efficient capital markets: comment. *Journal of Finance* 31, 139–41.

LeRoy, S. 1989. Efficient capital markets and martingales. *Journal of Economic Literature* 27, 1583–621.

LeRoy, S. and Parke, W. 1992. Stock price volatility: tests based on the geometric random walk. *American Economic Review* 82, 981–92.

LeRoy, S. and Porter, R. 1981. The present value relation: tests based on implied variance bounds. *Econometrica* 49, 555–74.

LeRoy, S. and Steigerwald, D. 1995. Volatility. In *Handbook of Finance*, ed. R. Jarrow, V. Maksimovic and W. Ziemba. Amsterdam: North-Holland.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

Mehra, R. and Prescott, E. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 41–9.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71, 421–36.

How to cite this article

LeRoy, Stephen F. "excess volatility tests." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000294> doi:10.1057/9780230226203.0514

The New Palgrave Dictionary of Economics Online

exchange rate dynamics

Nelson C. Mark

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Exchange-rate dynamics refers to the response path of the exchange rate following the revelation of some economic shock when the country in question operates under a pure flexible exchange-rate system. The issue attracts research attention because of the volatile nature of the exchange rate and the belief that the exchange rate may affect the allocation of resources across countries and over time. If observed exchange-rate dynamics cannot be shown to have a rational basis, efficiency will suffer from decisions made conditioned on disequilibrium values of the exchange rate.

Keywords

Bretton Woods system; Dornbusch, R.; exchange rate overshooting; exchange rate volatility; exchange-rate dynamics; fixed exchange rates; flexible exchange rates; incomplete markets; local-currency pricing; monetary shocks; Mundell–Fleming model; new open-economy macroeconomics; producer-currency pricing; real exchange rates; Redux model; risk aversion; sticky prices; uncertainty; uncovered interest parity; vector autoregressions

Article

Exchange-rate dynamics refers to the response path of the exchange rate following the revelation of some economic shock (news) when the country in question operates under a pure flexible exchange-rate system.
The topic has attracted research attention since 1973 when the industrialized world abandoned the Bretton Woods system of fixed exchange rates. The initial experience in the 1970s with this new exchange-rate system surprised economists along two dimensions. The first surprise was that exchange-rate returns turned out to be much more volatile than expected. This volatility, which is similar in magnitude to the volatility of stock returns, is much higher than the volatility of macroeconomic fundamentals such as the growth rate of money or income. The second surprise was the very high

persistence of the exchange rate. The logarithm of the exchange rate evolves approximately as a random walk so that all shocks appear to have a permanent effect on the exchange rate level. As a result, percentage changes of the exchange rate (returns) over short horizons are nearly unpredictable. These features of quarterly data from 1973Q1 to 2002Q1 are illustrated in Table 1.

Volatility (sample standard deviation) and autocorrelations of selected currencies, and other variables, 1973–2002

	DM	¥	£	S&P	M1	US GDP	T-Bill
Volatility	23.384	25.040	20.660	31.719	12.200	3.384	3.118
ρ_1	0.156	0.145	0.167	0.113	–0.273	–0.447	0.539
ρ_2	–0.072	–0.034	–0.128	–0.026	0.301	0.061	0.473
ρ_2	0.126	0.149	0.082	0.068	–0.309	–0.084	0.534
ρ_4	0.118	0.057	0.030	0.071	0.741	0.004	0.595

Notes: Data are annualized quarterly growth rates. They show volatility (sample standard deviation) and autocorrelations of real annualized returns on the Deutschmark, yen, and pound sterling relative to the US dollar, returns on the Standard and Poor's index, real M1 growth, and real US gross domestic product growth, and the three-month Treasury-bill rate. All variables are in real terms.

Sources: Standard and Poor's from Robert Shiller's website http://www.econ.yale.edu/~shiller/data.htm. All other data from Federal Reserve Economic Data (FRED).

Academic and policy interest in understanding exchange-rate dynamics stems from the belief that the exchange rate affects the current account, a country's international indebtedness, and the rate of capital formation, and is therefore an important macroeconomic variable that has real allocative implications for the open economy. An important question is whether observed exchange-rate dynamics have a rational basis or if it reflects an irrational overreaction to shocks. If it is the latter, economic efficiency may be compromised when allocative decisions are made conditional on disequilibrium or non-fundamental values of the exchange rate.

Due to limited experience with flexible exchange rates combined with high degrees of capital mobility during the Bretton Woods era, the volatile nature of exchange rates was difficult to anticipate. Moreover, the generally accepted framework at the time, for modeling the open economy, was the static Mundell–Fleming model which was poorly equipped for understanding these issues. The immediate post Bretton Woods experience stimulated a large body of theoretical work aimed at improved modelling the exchange rate. This work culminated with Dornbusch's (1976) celebrated exchange-rate overshooting model which provided a rational explanation for high exchange-rate volatility in the presence of relatively stable macroeconomic fundamentals. The overshooting model is a deterministic perfect-foresight dynamic generalization of Mundell–Fleming. The critical features are the differential speeds of adjustment between the goods (gradual) and asset (immediate) markets, uncovered interest parity, and the central importance of monetary shocks as the underlying source of uncertainty. Subsequent econometric work and quantitative analyses of dynamic general equilibrium models known as the ‘new open-economy macroeconomics’ suggest that an exact understanding of the mechanism that generates overshooting and excess exchange-rate volatility remains elusive.

Exchange-rate overshooting

Let *e* denote the natural logarithm of the home-currency price of the foreign currency. Under this definition of the nominal exchange rate, an increase in *e* means the home currency has weakened. Let the forward-looking steady state log exchange rate be $\bar{e}$. Assume that commodity prices are sticky, output is fixed, real money demand is inversely related to the interest rate *i*, financial capital is perfectly internationally mobile and domestic and foreign currency assets are perfect substitutes. In a deterministic setting, the perfect foresight instantaneous change in the exchange rate *e* can be shown to be proportional to the current exchange-rate gap $(\bar{e} - e)$. Let the exogenous foreign interest rate be i^*. Then uncovered interest parity, which says that an excess yield on domestic nominal assets is offset by an expected capital loss on the domestic currency through movements in the exchange rate, can be expressed as

$$i - i^* = \theta(\bar{e} - e).$$

This is the asset market equilibrium condition.
Now consider a one-time permanent surprise increase of one per cent in the home country's money supply. At the instant the shock is revealed, $\bar{e}$ increases by 0.01 in accordance with long-run monetary neutrality. Noting that the price-level is instantaneously fixed, the monetary shock creates a liquidity effect that lowers the interest rate *i*. To maintain uncovered interest parity, the instantaneous exchange rate *e* must increase by more than the 0.01 increase in $\bar{e}$, thereby ‘overshooting’ its long-run steady state value. Thus, the short-run variability of the exchange rate is seen to exceed the variability of the macroeconomic fundamentals in response to monetary shocks. In order to provide a general explanation for exchange rate volatility, however, it must be the case that nominal shocks are relatively important drivers of aggregate uncertainty because the model does not generate overshooting in response to real shocks such as shifts in the aggregate expenditure function. The overshooting model represented a significant contribution by giving an explanation for high exchange-rate variability in the context of a rational equilibrium model.
In stochastic generalizations of this model, as considered by Mussa (1982) and Obstfeld (1985), one would say that the exchange rate exhibits excess volatility in the sense that the volatility of the exchange rate exceeds that of the macroeconomic fundamentals (money supply and income). Moreover, in the stochastic environment the equilibrium exchange rate has a present-value representation that discounts the expected future flow of the macroeconomic fundamentals. The analogy to the present-value model of stock and bond pricing rationalizes why the exchange rate and other asset prices have many common properties.

Emerging doubts about the overshooting mechanism

Doubts about the precise mechanism that generates overshooting have been raised in vector

autoregression (VAR) studies. Eichenbaum and Evans (1995) modelled monetary shocks both as innovations to non-borrowed reserves and as innovations in the federal funds rate, and used a recursive ordering strategy to identify the VAR. Following a monetary shock, their impulse–response analysis did now show the immediate overshooting and the subsequent exponential decay of the exchange rate predicted by the overshooting model. Instead, they found a ‘hump-shaped’ response in the exchange rate: while it did overshoot the long-run equilibrium value, it did so gradually. This so-called ‘delayed overshooting’ result was confirmed in subsequent studies with structural VARs, such as Clarida and Gali (1994).

While such VAR studies create a dilemma for the overshooting mechanism, there is the possibility that the appearance of delayed overshooting was created by imposing false restrictions in the identification of VARs. Faust and Rogers (2003) argue that restrictions implied by recursive orderings employed in VAR analyses are implausible because they rule out contemporaneous responses of many important variables, such as the foreign interest rate, to domestic monetary policy shocks. Experimenting with alternative sets of more plausible restrictions on contemporaneous interactions, they find that immediate overshooting sometimes occurs and sometimes does not. In those situations where immediate overshooting does occur, they find that it is driven by deviations in uncovered interest parity. The implication then is that, if immediate overshooting is actually present in the data, it does not appear to be generated by the mechanism articulated by Dornbusch.

The preceding development of overshooting was discussed in terms of the nominal exchange rate, but the qualitative predictions would not be affected if it were recast in terms of the real exchange rate. At this point, we see that three pillars of the overshooting explanation of exchange-rate volatility are sticky nominal goods prices, uncovered interest parity, and the principal importance of monetary shocks. The question of whether prices are sticky is a tough problem that macroeconomists struggle to answer. Some international macroeconomic evidence has emerged to suggest that sticky goods prices are an important feature of the data. It has been pointed out by Mussa (1986) that the real exchange rate eP^*/P is much more volatile under flexible exchange rates than under fixed exchange rates. When the exchange rate is flexible, the correlation between nominal and real exchange rate movements is exceedingly high over relatively short horizons, say, from a month to a year (except in very high-inflation countries). It has also been observed that international violations of the law of one price are more severe than within-country violations. Engel and Rogers (1996) find that the volatility of the difference between the log price of a particular good sampled in two locations is much higher if the sample points are in different countries than if they are within the same country.

Exchange-rate dynamics in ‘new open-economy macroeconomic’ models

The exchange-rate models of the 1970s and 1980s were typically built from sets of ad hoc macroeconomic relations that lack rigorous microeconomic foundations. As exchange-rate models have evolved towards dynamic general equilibrium analysis, the research continues to try to understand the necessary conditions for an equilibrium rational-expectations monetary model to explain observed exchange-rate volatility and persistence. A major vein of this work is done within the ‘new open-economy macroeconomics’ (NOEM) class of theories that features rational optimizing agents, imperfect competition and sticky goods prices in a dynamic general equilibrium open-economy setting.

The NOEM began with the Obstfeld and Rogoff (1995) Redux model. Money is introduced through the utility function, and agents maximize expected lifetime utility defined over consumption, leisure and real money balances. Goods prices are set one period in advance by monopolistically competitive firms that produce output with labour but no physical capital. The precise way in which the exporting firm sets the price of output for foreign and domestic buyers has significant implications for exchange-rate dynamics. In the Redux model, firms engage in ‘producer-currency pricing’. That is, an exporter will set the domestic price of the good at P and foreigners pay P/e units of the foreign currency. While P is fixed for the period, e is not, so a within-period change in the nominal exchange rate will alter the relative price between home and foreign goods. This is the basis of the ‘expenditure-switching’ effect of the exchange rate stressed in the Mundell–Fleming model. Even though goods prices are sticky, the Redux model generates the surprising result that a monetary shock causes the nominal exchange rate to jump immediately to its long-run value. In other words, there is no overshooting or excess volatility of the exchange rate in the Redux model.

An alternative to producer-currency pricing is ‘pricing-to-market’ (also known as ‘local-currency pricing’). In this scenario, segmentation between the domestic and foreign goods markets allows firms to set a foreign currency export price that is different from the domestic currency price of the good. Here, foreigners pay the pre-set price P^* and the exporting firm receives eP^* units of home currency. Within-period changes in the exchange rate affect firm revenues but not the relative price between imported and domestic goods. Betts and Devereux (2000) apply this idea in modifying the Redux model and find that increasing the fraction of firms that price-to-market reduces the strength of the expenditure-switching effect, which attenuates the allocative role of the exchange rate. As a result, a larger change in the exchange rate is required to restore equilibrium following a monetary shock. Accordingly, if the degree of pricing-to-market is sufficiently high, the overshooting result can be restored.

Some NOEM modellers ask their models to quantitatively match the data. A quantitative analysis of the exchange-rate dynamics implied by an NOEM model might suppose that agents operate in a complete markets setting where they can trade a full set of state-contingent claims, which allows complete international risk-sharing in a model. Production requires labour and durable physical capital. There is a final good that is not traded internationally and produced in a competitive industry. Intermediate goods are traded internationally and produced by monopolistically competitive price-setting firms that change prices according to a staggered price-setting rule. The period utility is defined over consumption C_t, real money balances M_t/P_t, and leisure L_t. With asterisks denoting foreign country variables, the exchange rate is determined by the equilibrium risk-sharing condition

$$\frac{E_t P_t^*}{P_t} = \Gamma \frac{U_{c^*}(C_t^*, M_t^* / P_t^*, L_t^*)}{U_c(C_t, M_t / P_t, L_t)}.$$

where E_t is the nominal exchange rate level and U_c is the marginal utility of consumption. The real exchange rate is proportional to the ratio of foreign to domestic marginal utility. If utility is separable in its arguments and has the constant relative risk-aversion form, the real exchange rate will be

proportional to relative consumption levels in the two countries and the factor of proportionality Γ will be increasing in the coefficient of relative risk aversion. This exchange-rate pricing condition is invariant to whether commodity prices are sticky or flexible. The question is under what conditions will the model generate consumption responses to monetary shocks to create exchange-rate dynamics like those found in the data.

Chari, Kehoe and McGrattan (2002) calibrate this two-country model to the United States and an aggregate 'European' country which they employ to simulate model-implied observations of the endogenous variables. Using the staggered-price setting rule of Taylor (1999), they find that firms cannot be allowed to change prices more frequently than once a year and that the coefficient of relative risk aversion must be at least 5 in order for the model-generated real exchange rate to match the persistence and volatility found in the data. While these are not unreasonable conditions, the degrees of price stickiness and of risk aversion required both seem to be on the high side. Kollman (2001), on the other hand, is able to obtain exchange-rate overshooting with a relative risk-aversion coefficient of 2, but uses the Calvo (1983) rule for price setting.

While this line of work shows that the interaction of monetary shocks and sticky prices is able to explain volatile and persistent exchange-rate behaviour, the models often generate counterfactual predictions for other dimensions of the data. One of these counterfactuals, obtained from the exchange-rate determination equation, is that the real exchange rate should be nearly perfectly correlated with relative foreign-to-home consumption levels. In the data, however, the real exchange rate and relative consumption levels are uncorrelated – a problem known as the Backus–Smith (1993) puzzle.

The restrictive feature of models with complete markets is the international risk-sharing condition which constrains the extent to which the exchange rate can move. One might think that limiting the menu of tradable assets and working in an incomplete markets environment might loosen up this restriction. A common specification of an incomplete markets environment is to allow only a non-state-contingent bond to be internationally traded. However, as demonstrated by Baxter and Crucini (1995), the quantitative properties of dynamic general equilibrium models are little affected when full risk sharing is replaced by this version of incomplete markets, especially when shocks to the environment are not permanent. The reason is that uncovered interest parity replaces the risk-sharing condition in determining the exchange rate, so relatively smooth interest rates replace smooth consumption levels in limiting the range of exchange-rate movements.

Accordingly, studies by Devereux and Engel (2002), Kollman (2001) and Duarte and Stockman (2005) of exchange-rate dynamics in dynamic general equilibrium models under incomplete markets find it necessary to allow exogenous deviations from uncovered interest parity in order to explain exchange-rate volatility in the presence of smooth fundamentals. There is ample empirical work, surveyed by Hodrick (1987), Engel (1996), and Lewis (1995), to justify these violations, and theoretical work by Mark and Wu (1998) and Jeanne and Rose (2002) gives an explanation for the deviations through participation of noise traders in the foreign-exchange market.

High exchange-rate volatility, persistence, and overshooting can be generated from fully specified dynamic stochastic general equilibrium models. However, a fully convincing story about exchange-rate dynamics remains elusive because an accepted model that provides reasonably good ability to account for all of the salient features of the data is not yet available.

See Also

- exchange rate volatility
- foreign exchange market microstructure
- real exchange rates

Bibliography

Backus, D. and Smith, G. 1993. Consumption and real exchange rates in dynamic economies with non-traded goods. *Journal of International Economics* 35, 297–316.

Baxter, M. and Crucini, M. 1995. Business cycles and the asset structure of foreign trade. *International Economic Review* 36, 821–54.

Betts, C. and Devereux, M. 2000. Exchange rate dynamics in a model of pricing-to-market. *Journal of International Economics* 50, 215–44.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Chari, V., Kehoe, P. and McGrattan, E. 2002. Can sticky price models generate volatile and persistent real exchange rates? *Review of Economic Studies* 69, 533–64.

Clarida, R. and Gali, J. 1994. Sources of real exchange-rate fluctuations: how important are nominal shocks? *Carnegie-Rochester Conference Series on Public Policy* 41, 1–56.

Devereux, M. and Engel, C. 2002. Exchange rate pass-through, exchange rate volatility, and exchange rate disconnect. *Journal of Monetary Economics* 49, 913–40.

Dornbusch, R. 1976. Expectations and exchange-rate dynamics. *Journal of Political Economy* 84, 1161–76.

Duarte, M. and Stockman, A. 2005. Rational speculation and exchange rates. *Journal of Monetary Economics* 52, 3–29.

Eichenbaum, M. and Evans, C. 1995. Some empirical evidence on the effects of shocks to monetary policy on exchange rates. *Quarterly Journal of Economics* 110, 975–1009.

Engel, C. 1996. The forward discount anomaly and the risk premium: a survey of recent evidence. *Journal of Empirical Finance* 3, 123–92.

Engel, C. and Rogers, J. 1996. How wide is the border? *American Economic Review* 86, 1112–25.

Faust, J. and Rogers, J. 2003. Monetary policy's role in exchange rate behavior. *Journal of Monetary Economics* 50, 1403–24.

Hodrick, R. 1987. *The Empirical Evidence on the Efficiency of Forward and Futures Foreign Exchange Markets*. Chur: Switzerland Harwood Academic Publishers.

Jeanne, O. and Rose, A. 2002. Noise trading and exchange rate regimes. *Quarterly Journal of Economics* 117, 537–69.

Kollman, R. 2001. The exchange rate in a dynamic-optimizing business cycle model with nominal rigidities: a quantitative investigation. *Journal of International Economics* 55, 243–62.

Kollman, R. 2005. Macroeconomic effects of nominal exchange rate regimes: new insights into the role of price dynamics. *Journal of International Money and Finance* 24, 275–92.

Lewis, K. 1995. Puzzles in international financial markets. In *Handbook of International Economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: North-Holland.

Mark, N. and Wu, Y. 1998. Rethinking deviations from uncovered interest parity: the role of covariance risk and noise. *Economic Journal* 108, 1686–706.

Mussa, M. 1982. A model of exchange rate dynamics. *Journal of Political Economy* 90, 74–104.

Mussa, M. 1986. Nominal exchange rate regimes and the behavior of real exchange rates: evidence and implications. *Carnegie Rochester Conference Series on Public Policy* 25, 117–213.

Obstfeld, M. 1985. Floating exchange rates: experience and prospects. *Brookings Papers on Economic Activity* 1985(2), 369–450.

Obstfeld, M. and Rogoff, K. 1995. Exchange rate dynamics Redux. *Journal of Political Economy* 103, 624–60.

Taylor, J. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of Macroeconomics*, vol. 1B, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

How to cite this article

Mark, Nelson C. "exchange rate dynamics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000267> doi:10.1057/9780230226203.0516

The New Palgrave Dictionary of Economics Online

exchange rate exposure

Kathryn M.E. Dominguez and Linda L. Tesar

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Exchange rate exposure describes the influence of exchange rate movements on the value of a firm or sector of the economy. Exposure is typically measured as the correlation of firm or industry stock returns and exchange rate changes in the context of a market model. Exposure appears to be most prevalent in firms that are small (these are less likely to engage in hedging activities) or involved in international activities. While studies have linked *ex ante* exchange rate risk with firm investment strategies, it has proven difficult to identify the *ex post* consequences of exposure on firm or industry behaviour.

Keywords

capital asset pricing model; exchange rate exposure; exchange rate risk; hedging; operating exposure; trade-weighted basket of currencies; transaction exposure; translations exposure; value-weighted market returns

Article

Exchange rate exposure measures the extent to which firm value is influenced by exchange rate movements. Narrow definitions of exchange rate exposure, such as transaction exposure, translations exposure, or operating exposure, focus on the effect of the exchange rate on specific components of a firm's balance sheet or on different types of transactions. More broadly, economists use the term ‘economic exposure’ to describe the impact of exchange rate movements on a firm or a sector of the economy. A firm's ‘exposure’ could be measured as the responsiveness of profits, cash flow, operating costs, total assets or liabilities, markups or even wage-setting behaviour to fluctuations in the exchange rate.

If one has detailed information about a firm's operations, it may be possible to trace the effect of exchange rate shifts on a specific line item in the firm's accounting data, with other factors controlled

for. More generally, however, measures of exchange rate exposure involve an indirect test of the link between the exchange rate and a firm's stock return, where the return is taken as a measure of the firm's profitability. Under the capital asset pricing model (CAPM), the expected risk premium on a company's share price is proportional to its covariance with the market portfolio. In theory, investors will require a return only on the non-diversifiable portion of firm risk and, if the market return captures all systematic risk, no variable other than the market return should play a role in determining asset returns. Therefore, a test for exchange rate exposure involves including the change in the exchange rate on the right-hand side of a standard CAPM regression and testing whether its coefficient is significantly different than zero (Adler and Dumas, 1984):

$$R_{i,t} = \beta_{0,i} + \beta_{1,i} R_{m,t} + \beta_{2,i} \Delta s_t + \varepsilon_{i,t} \quad (1)$$

where $R_{i,t}$ is the return on firm i at time t, $R_{m,t}$ is the return on the market portfolio, $\beta_{1,i}$ is the firm's beta, Δs_t is the change in the relevant exchange rate and $\beta_{2,i}$ measures a firm's exposure to exchange rate movements after the overall market's exposure to currency fluctuations is taken into account. If $\beta_{2,i}$ is zero, this implies that firm i has the same exchange rate exposure as the market portfolio (not necessarily that the firm has no exposure). Alternatively, if one rejects the hypothesis that $\beta_{2,i}$ is, on average, zero, this indicates that the firm is more exposed to exchange rate movements than the market. Note that the interpretation of $\beta_{2,i}$ as a measure of exposure is conditional on eq. (1) being the ‘true’ model of asset returns. Evidence that $\beta_{2,i}$ is non-zero can be interpreted as evidence against the joint hypothesis that the CAPM holds (that is, the market efficiently prices systematic risk) and that exchange rate risk is unimportant for stock returns.

An alternative approach is to measure total exposure, or the unconditional correlation of exchange rates and returns that would involve omitting the market return as an explanatory variable in eq. (1). The advantages to measuring total exposure are that it allows one to measure the exposure of all firms as a group rather than individual firms relative to the group average, and it requires no assumption about the validity of the CAPM. The disadvantage of total exposure is that it does not allow one to distinguish between the direct effect of exchange rate changes and the effects of macroeconomic shocks that simultaneously affect firm value and exchange rates (which we assume are captured in eq. (1) by the market return).

Predicting which firms are likely to be affected by changes in the exchange rate and the direction that exposure might take can be quite complicated. There are a number of channels through which the exchange rate might affect the profitability of a firm. Firms that export to foreign markets may benefit from a depreciation of the local currency if their products become more affordable to foreign consumers. On the other hand, firms that rely on imported intermediate products may see their profits shrink as a consequence of increasing costs of production. Firms with foreign subsidiaries may shift activities across national boundaries, taking on exchange rate risk to take advantage of tax differentials. Firms that do no international business may be influenced indirectly by foreign competition. Furthermore, firms in

the non-traded as well as the traded sectors of the economy compete for factors of production, whose returns may be affected by changes in the exchange rate.

A number of specific issues arise when implementing eq. (1) as a test for exchange rate exposure. First, one must identify the relevant exchange rate. Many studies use a trade-weighted exchange rate to measure exposure. The problem with using a trade-weighted basket of currencies in exposure tests is that the results lack power if the nature of firm exposure does not correspond to the exchange rates (and the relative weights) included in the basket. More generally, one should expect variation in individual firm and industry exposure to various exchange rates. Dominguez and Tesar (2001a) find that any test that restricts the measurement of exposure to one exchange rate (whether it be a trade-weighted rate or a bilateral rate) is likely to be biased downward. Exposure may also be a function of horizon. It may be that firms can use financial derivatives to hedge exchange rate risk in the short run, but remain exposed to low-frequency movements in exchange rates over long horizons. Dominguez and Tesar (2006) find that exposure generally increases with return horizon.

A second issue is sample selection – which firms should be included in empirical tests for exposure. Much of the literature has focused on exposure in multinational firms (Jorion, 1990; He and Ng, 1998), or in firms that actively engage in international trade (Bodnar, Dumas and Marston, 2002). However, there are good reasons to think that exposure will not be limited to these firms. Dominguez and Tesar (2001b) find that firms that engage in greater trade exhibit *lower* degrees of exposure, reflecting the fact that they are also the most aware of exchange-rate risk and, therefore, the most likely to hedge their exposure. Exposure may also be affected by the competitive structure at the industry level. Less competitive industries with higher markups can adjust prices in response to exchange rates, and the impact on profitability and returns will thus be smaller. Allyanis and Ihrig (2001) show that the extent of exposure varies systematically with the extent of industry-level markups.

A third issue that arises in tests for exchange rate exposure is the specification of the market index. Empirical tests of the standard CAPM model generally include a country-specific value-weighted market return to proxy for ‘the market’. Value-weighted market returns are likely to be dominated by large firms that are more likely to be multinational or export-oriented, and are more likely than other firms to experience negative cash flow reactions to home currency appreciations. In a world of perfectly integrated capital markets the ‘market return’ should be proxied by a global portfolio. Dominguez and Tesar (2006) sort out the impact of the choice of market index on exposure for a sample of firms in eight (non-US) markets and find little difference in estimated exposure level using value-weighted and equal-weighted indices, but find a significant increase in measured exposure using a global index. Part of the explanation for this is that the global index generally does a poor job of explaining returns, so that more firms appear to be exposed simply because the exchange rate is picking up more of the variability in returns, and the (global) market index is picking up substantially less.

A final issue is whether exposure is predictable. The standard way to test this is to run a second-stage regression that takes the estimated exposure betas from eq. (1) and regresses these on a variety of potential explanatory variables. Using this approach Dominguez and Tesar (2006) find that exposure is more prevalent in small (rather than large or medium-sized) firms, and in firms engaged in international activities (measured by multinational status, holdings of international assets, and foreign sales).

Once one has identified a set of firms that are exposed to exchange rate risk, the question remains whether such exposure affects firm behaviour in some way, such as its level of investment or market

entry. In general, it has proven difficult to identify a link between *ex post* exposure (as measured by estimates of exposure in eq. (1)) and such economic outcomes. Numerous studies, however, have linked firm strategies for handling exchange rate risk *ex ante* with their investment decisions in domestic and foreign markets.

See Also

- capital asset pricing model
- hedging
- risk

Bibliography

Adler, M. and Dumas, B. 1984. Exposure to currency risk: definition and measurement. *Financial Management* 13, 41–50.

Allyanis, G. and Ihrig, J. 2001. Exposure and markups. *Review of Financial Studies* 14, 805–35.

Bodnar, G., Dumas, B. and Marston, R. 2002. Pass-through and exposure. *Journal of Finance* 57, 199–231.

Dominguez, K. and Tesar, L. 2001a. A re-examination of exchange rate exposure. *American Economic Review Papers and Proceedings* 91, 396–9.

Dominguez, K. and Tesar, L. 2001b. Trade and exposure. *American Economic Review Papers and Proceedings* 91, 367–70.

Dominguez, K. and Tesar, L. 2006. Exchange rate exposure. *Journal of International Economics* 68, 188–218.

He, J. and Ng, L. 1998. The foreign exchange exposure of Japanese multinational corporations. *Journal of Finance* 53, 733–53.

Jorion, P. 1990. The exchange rate exposure of U.S. multinationals. *Journal of Business* 63, 33–45.

How to cite this article

Dominguez, Kathryn M. E. and Linda L. Tesar. "exchange rate exposure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008

<http://www.dictionaryofeconomics.com/article?id=pde2008_E000150>
doi:10.1057/9780230226203.0517

The New Palgrave Dictionary of Economics Online

exchange rate target zones

John Driffill
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A target zone attempts to limit the movement of an exchange rate, avoiding the pitfalls of both a pegged rate and a freely floating rate. The European Monetary System was the prime example. An elegant model of Paul Krugman demonstrates that in theory a target zone does indeed stabilize an exchange rate. But in practice it has been substantially rejected empirically. Williamson's ‘crawling bands’ around a ‘fundamental equilibrium exchange rate’ develop the concept. Target zones survive among candidates for membership of the Eurozone who take part in the Exchange Rate Mechanism mark II.

Keywords

Bretton Woods System; Brownian motion; capital controls; Economic and Monetary Union (EMU); euro; European Monetary System; Exchange Rate Mechanism (EU); exchange rate target zone; fixed exchange rates; floating exchange rates; fundamental equilibrium exchange rate; inflation differentials; monetary theory of the exchange rate; nominal exchange rates; purchasing power parity; real exchange rates; uncovered interest parity; Wiener process

Article

An exchange rate target zone is a scheme intended to limit the flexibility of an exchange rate without going as far as fixing or pegging the value of one currency against another. It is a band, or zone, of values for the exchange rate, around a central or target rate. Within the zone, the exchange rate is allowed to fluctuate freely without any intervention from the authorities or, at least, with less intervention than there is elsewhere. At the edge of the band, and outside, if the rate strays there, there is more vigorous intervention to keep the rate within, or return it to, the band. There are many varieties of target zone. The edges may be hard or soft. It may be defined in terms of nominal or real exchange rates. The central rate – the target – may be either constant over time, possibly with provision for occasional discrete changes; or it may be adjusted continuously. The bands may be narrow or wide.

The most celebrated and ambitious target zones were those introduced in 1979 by the European Monetary System (EMS). They operated until the end of 1998. Under the EMS, member states were initially required to keep their bilateral exchange rates within a band of ±2.25 per cent around a grid of central parities (Giavazzi and Giovannini, 1989). They were required to use unlimited intervention in the foreign exchange markets to defend the bands if an exchange rate strayed to the edge. Member countries could adjust central parities occasionally by mutual agreement when perceived misalignments had built up. The system evolved over time. As capital controls were progressively removed, orderly realignments became more difficult to manage. The system became less flexible, notably after 1987. The gradual movement towards complete fixity of exchange rates, intended to prevail under Economic and Monetary Union, was thrown off course by massive speculative attacks in September 1992 and August 1993. The system was unable to withstand them and the bands were widened to 15 per cent. But they were subsequently narrowed again and the EMS gave way to the euro on the 1 January 1999.

The use of target zones sprang from a desire to avoid the pitfalls of fixed rates and free floating. Under the fixed exchange rates of the Bretton Woods System (1944–73), exchange rate misalignments had become progressively worse as inflation rates diverged, and weak currency countries put off devaluation, deterred by costly speculation. Under floating exchange rates during the 1970s, exchange rates fluctuated excessively, unrelated to fundamentals like relative price levels and current accounts. The ‘disconnect’ between exchange rates and economic fundamentals has been confirmed by widespread experience and has become a central tenet of international macroeconomics.

The EMS was intended to allow exchange rates to offset inflation differentials among members. Realignments were to be sufficiently timely to avoid giving the markets a one-way

bet. The bands were intended to enable markets to determine exchange rate movements without official intervention for most of the time, at the same as discouraging destabilizing speculation.

The questions of how target zones might work in theory, whether they worked in practice as the theory predicted, and whether they did indeed cut exchange fluctuations, have generated enormous amounts of research.

The key theoretical contribution is that of Krugman (1991). He showed that a fully credible target zone would reduce the volatility of an exchange rate and reduce its sensitivity to fundamentals. His theoretical model assumes a monetary theory of the exchange rate for a small open economy in a world of perfectly flexible prices and perfect capital mobility, in which purchasing power parity and uncovered interest parity hold good. Then the log of the exchange rate (*e*) can be expressed as a function its own anticipated rate of change over time ($E_t(de) / dt$) and a driving fundamental (f)

$$e = f + \alpha E_t(de) / dt$$

The parameter α denotes the semi-elasticity of money demand with respect to the interest rate. The fundamental reflects money supply and demand. He considers a stochastic model in continuous time, in which the fundamental (f) follows Brownian motion, the continuous-time analogue of a random walk. That is

$$df = \sigma dz$$

where *dz* is the innovation in a standard Wiener process, and σ is the variance of the innovation in the fundamental per unit time. *dz* has mean zero and variance $E(dz^2) = dt$. When the exchange rate is allowed to float freely, the exchange rate will be a linear function of the fundamental

$$e = f.$$

But, when a target zone limits movements of the exchange rate, this solution does not apply. Assume for simplicity that under the target zone the central parity for the logarithm of the exchange rate is equal to zero and the limits of the zone are $\bar{e}$ and $\underline{e}$. Further assume that the zone is symmetrical and $\bar{e} = -\underline{e}$. Using stochastic calculus and methods widely used in the theory of options pricing, Krugman shows that the exchange rate is related to the driving fundamental by the relationship

$$e = f + A(\exp(\xi f) - \exp(-\xi f))$$

where

$$\xi = \sqrt{\frac{2}{\sigma^2 \alpha}}.$$

The constant A is such that at the top and bottom of the band the value-matching conditions

$$\bar{e} = \bar{f} + A(\exp(\xi\bar{f}) - \exp(-\xi\bar{f}))$$

and

$$\underline{e} = \underline{f} + A(\exp(\xi\underline{f}) - \exp(-\xi\underline{f}))$$

hold, and the smooth pasting conditions also hold. These require that the derivative of the exchange rate with respect to the fundamental at the edges of the band is equal to zero. Viz.

$$1 + A(\xi\exp(\xi\bar{f}) + \xi\exp(-\xi\bar{f})) = 0$$

and

$$1 + A(\xi\exp(\xi\underline{f}) + \xi\exp(-\xi\underline{f})) = 0.$$

From these conditions the value of the constant *A*, and the value of the fundamental at the limits of the band $(\bar{f} = -\underline{f})$, can be determined. The value of the parameter A that emerges from this analysis is negative. Thus the value of the exchange rate, corresponding to any particular value of the fundamental, is closer to the parity rate under a target zone than it would be under a free float.

Krugman's analysis establishes that the exchange rate within the target zone enjoys the ‘bias in the band’ or the ‘honeymoon effect’. The relationship between the fundamental and the exchange rate in the target zone is an S-shaped curve. It is flatter everywhere than the relationship under a free float, which is a 45 degree line. The perfectly credible commitment of the authorities to intervene, should the exchange rate ever reach the edge of the band, so as to prevent any movement beyond it, discourages deviations from the central parity even without any actual intervention. It is illustrated in Figure 1.

Figure 1

Source: Krugman (1991).

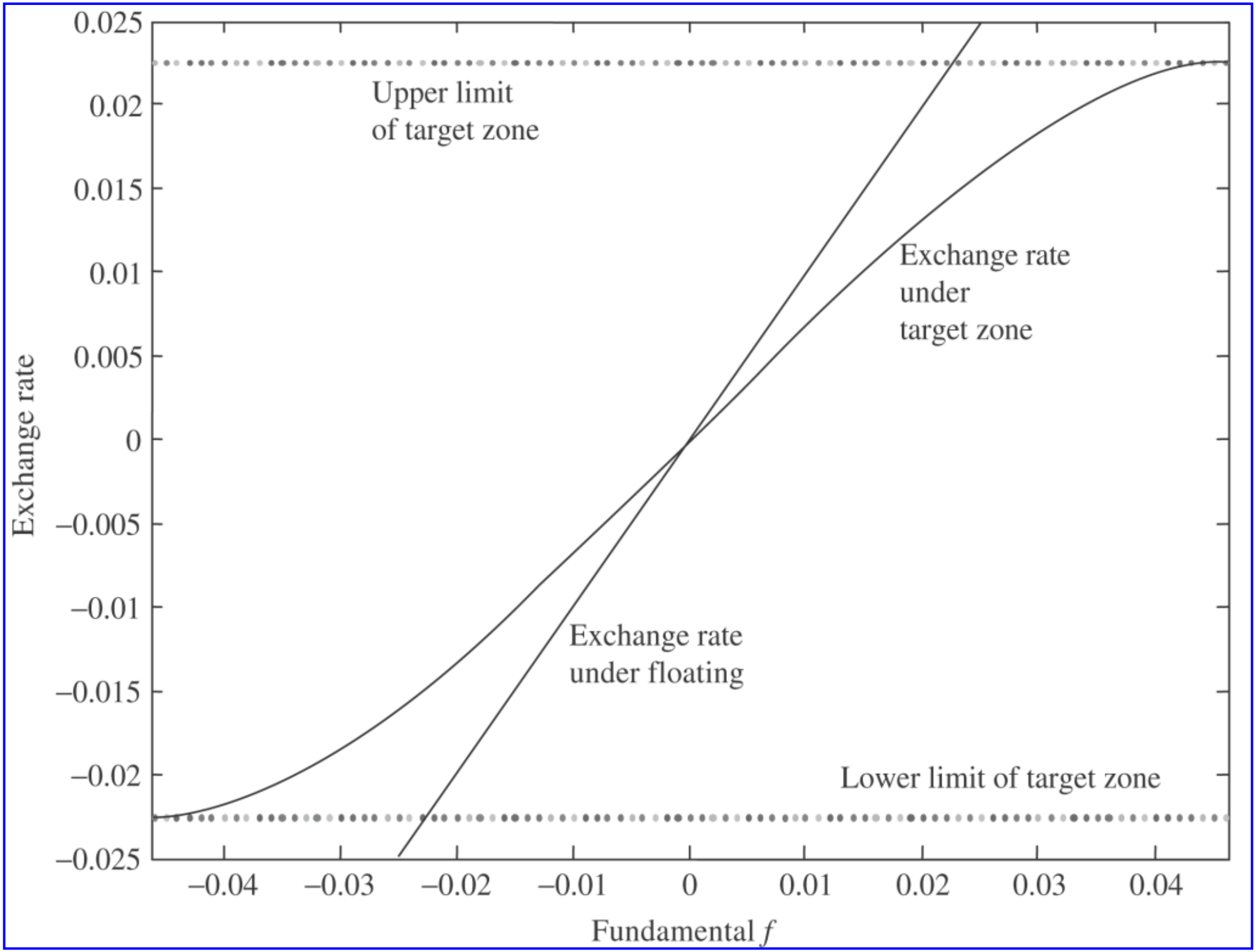
0.025
0.02
0.015
0.01
0.005
0
–0.005
–0.01
–0.015
–0.02
–0.025
Exchange rate
Upper limit
of target zone
Exchange rate
under
target zone
Exchange rate
under floating
Lower limit of target zone
–0.04
–0.03
–0.02
–0.01
0
0.01
0.02
0.03
0.04
Fundamental f

This elegant theory has very strong empirical predictions. Three predictions that do not rely on any assumptions about the value of the unobserved fundamentals are as follows. First, the exchange rate should spend a lot of time near the edges of the band, and relatively little near the centre. The unconditional distribution of the fundamental within the band is uniform, and the distribution of the exchange rate is U-shaped. Second, the uncovered interest parity condition implies that, when the exchange rate is strong (e is close to $\bar{e}$) and thus likely to weaken, the domestic interest rate should exceed the foreign rate, while when the exchange rate is weak, the domestic interest rate should be relatively low. And third, the converse of the second prediction, at any point in time the expected future exchange rate implied by the uncovered interest parity condition

$$E_t(e_{t+dt}) = e_t + (r_t - r_t^*)dt$$

(r_t is the domestic instantaneous nominal interest rate and r_t^* the foreign one) should lie within the band.

Unfortunately, all three of these predictions are comprehensively rejected by empirical evidence, of which a great deal has been accumulated. The work of Flood, Rose and Mathieson (1991) led the way in this. They and many subsequent studies have found that exchange rates had tended to be concentrated near the middle of the band, not at the edges, as predicted. They also found that, when the exchange rate was weak, there was no tendency for the interest rate to be relatively low. The expected future exchange rate implied by uncovered interest parity was found to spend a great deal of time outside the band, suggesting a lack of credibility of the target zone.

Direct tests of the relationship between the exchange rate and the fundamental driving variable have generally found very little evidence of non-linearity or S-shapedness. There appears to be little evidence of any ‘honeymoon effect’. Svensson (1992) remarks that the comprehensive rejection of this theory looks like what T.H. Huxley called ‘the great tragedy of science – the slaying of a beautiful hypothesis by an ugly fact’. But in fact, while descriptively unrealistic, Krugman's model of target zones maintains its conceptual grip. A number of minor amendments to the theory have gone some way to reconciling it with the evidence while leaving its central ideas intact. The theory makes many clearly unrealistic assumptions. Two changes that have been particularly important are allowing for imperfect credibility of the target zone and allowing for intra-marginal intervention. Intra-marginal intervention in particular alters the process driving the fundamentals and can cause the theoretically predicted distribution of the exchange rate within the zone to have the empirical humped shape. Expectations that a zone is not fully credible and that the authorities might adjust the central parity rather than defend the zone also have the effect of reducing the curvature of the S-shaped curve and bringing it closer to the 45-degree line that would prevail under free floating.

With the empirical failure of the theory and the disappearance of the most prominent practical example when the euro replaced the EMS in 1999, interest in target zones has subsided. Nevertheless, John Williamson (1998) emphasizes the empirical observation that within a target zone the forward rate responds less than one-for-one with a change in the spot rate, whereas the same is not true of a floating exchange rate, as an indication that even an imperfectly credible target zone exerts a stabilizing influence on exchange rates. He has proposed looser arrangements for exchange rates, such as ‘crawling bands’ and ‘monitoring bands’, in which the band is defined around an equilibrium real exchange rate (his concept of the fundamental equilibrium exchange rate), which naturally implies a central nominal exchange rate that crawls over time. While a crawling band involves a commitment to keep the exchange rate within a wide announced band, a monitoring band involves a weaker commitment. A number of countries used crawling bands during the 1990s, including Chile, Colombia, Israel, Indonesia, Ecuador, and Russia. The IMF (2006) reports that by the end of 2005 no countries were using crawling bands. Several countries were using ‘pegged exchange rates within horizontal bands’, mostly countries in ERM II, the revised form of the Exchange Rate Mechanism of the former EMS: Cyprus, Denmark, the Slovak Republic, Slovenia, and Hungary; with Tonga alone outside ERM II.

See Also

- Bretton Woods system
- bubbles
- currency crises
- European Monetary Union
- exchange rate volatility
- irreversible investment

Bibliography

Flood, R.P., Rose, A.K. and Mathieson, D.J. 1991. An empirical examination of exchange rate target zones. *Carnegie-Rochester Series on Public Policy* 35, 7–65.

Giavazzi, F. and Giovannini, A. 1989. *Limiting Exchange Rate Flexibility: The European Monetary System*. Cambridge, MA: MIT Press.

IMF (International Monetary Fund). 2006. De facto classification of exchange rate regimes and monetary policy frameworks. 31 December. Online. Available at http://www.imf.org/external/np/mfd/er/2005/eng/1205.htm, accessed 23 October 2006.

Krugman, P.R. 1991. Target zones and exchange rate dynamics. *Quarterly Journal of Economics* 106, 669–82.

Svensson, L.E.O. 1992. An interpretation of recent research on exchange rate target zones. *Journal of Economic Perspectives* 6(4), 119–44.

Williamson, J. 1998. Crawling bands or monitoring bands: how to manage exchange rates in a world of capital mobility. *International Finance* 1, 59–79.

How to cite this article

Driffill, John. "exchange rate target zones." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000301> doi:10.1057/9780230226203.0518

The New Palgrave Dictionary of Economics Online

exchange rate volatility

Roberto Rigobon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Exchange rate volatility is at the forefront of academic, policy, and market participant discussions in developed and developing economies. This article reviews the benefits and cost of exchange rate variations, and its implications for the economy. It also summarizes the general understanding that exists regarding the relationship between real and nominal exchange rates.

Keywords

aggregation bias; Bretton Woods system; consumer price index (CPI); contagion; European Monetary Union; exchange rate dynamics; Exchange Rate Mechanism (EU); exchange rate regimes; exchange rate volatility; financial market contagion; fixed exchange rates; flexible exchange rates; inflation targeting; monetary base; nominal exchange rates; price revelation; protection; purchasing power parity; real exchange rates; speculation; sticky prices; uncovered interest parity

Article

The volatility of the exchange rate is perhaps one of the most studied topics in international economics; from the real exchange rate to the nominal exchange rate, from the theories of exchange rate volatility to the predictability of exchange rate movements, from the explanations of the volatility to its implications, and from the independence of exchange rates around the world to contagion, it can be argued that all dimensions have been heavily scrutinized in the literature. (As it is virtually impossible to cite all the relevant sources, I will mostly refer to the classics and to relevant surveys in the literature.)
Why is so much attention paid to exchange rate volatility? For a start, exchange rate volatility has almost unanimously been seen as a *bad thing* by policymakers, practitioners, and academics – whatever ‘bad’ or ‘thing’ means. Indeed, an important policy objective in the recent past, for several emerging and developed nations, has been exchange rate stability. The road to this stability, however, has been long and with many bumps and pitfalls. For instance, one of the goals of the Bretton Woods system was to

control the excessive exchange rate variability that surfaced during the interwar period – it was unsuccessful, and in the early 1970s the world abandoned fixed exchange rates. After teasing the world with flexible rates for less than a decade, the European nations moved again to controlling their exchange rates with the Exchange Rate Mechanism (ERM) in early 1980s. It was a more comprehensive set of rules devoted to enhance cooperation and coordination among the members with the purpose of reducing exchange rate variability. It was a good idea, but unfortunately the system did not last, and it collapsed in the early 1990s. Europe stirred itself in 1999 to adopt an even stronger set of rules to foster even more cooperation and coordination – the European Monetary Union (EMU). Will it last? Too close to call, yet.

Emerging markets show even stronger variability of exchange rates, and more frequent failures of fixed regimes. Countries usually embark on periods of controlled nominal exchange rates in the hope that they can achieve stability, both internal and external. Those efforts rarely last, and most such countries are forced to devalue. In the end, the search for the Holy Grail of exchange rate stability looks more like an electrocardiogram than a smooth path.

Interestingly, this path has been frustrating enough that some countries have given up their exchange rate stabilization objective. The recent shift towards inflation targeting by many central banks is an indication that today they are assigning a smaller weight to exchange rate stabilization in the design of monetary policy than in the past. However, as the present global imbalances debate on the sustainability of the US current account deficit and the Asian and European current account surpluses shows, the exchange rate is still at the forefront of the concerns of policymakers, practitioners and academics.

Several questions come to mind. First, from the economics point of view, which volatility is the relevant one: nominal or real? Second, what are the costs and benefits of exchange rate stability? Third, what causes exchange rate instability, and how can it be controlled? The following sections answer these questions.

Nominal vs real exchange rate volatility

If we were to analyse the impact of higher exchange rate volatility on growth or trade, the first question would be whether the nominal or the real exchange rate should be considered. In practice, nominal and real exchange rates are closely intertwined. In his seminal paper, Mussa (1986) reports one of the most robust facts in international economics: the nominal and real exchange rates move almost one to one; periods of nominal exchange rate stability are always associated with periods of stable real exchange rates, while periods of nominal exchange rate instability are accompanied by excessive real exchange rate variability. Furthermore, a country that shifts from a fixed nominal exchange rate to a flexible nominal exchange rate experiences an increase in the variance of both nominal and real exchange rates. This is particularly true at shorter frequencies (quarterly or monthly). This evidence points to two important lessons. First, prices are sticky in the short run, so that nominal and real exchange rates are driven by the same factor, namely, shocks to the nominal exchange rate. Second, demand shocks – or nominal shocks – govern the short- run dynamics of the nominal exchange rate.

An active recent area of research has been the collection of empirical evidence regarding the degree of stickiness of prices using micro data. The seminal paper in this area is by Bils and Klenow (2004), who study price stickiness in the items used to construct the consumer price index (CPI) for the United States. They find that the median stickiness is four to five months. By contrast, Alvarez et al. (2005) find

a degree of stickiness of about a year in European items constituting the CPI. Furthermore, for the US prices used to construct the import and export indexes, Gopinath and Rigobon (2006) find that the degree of stickiness is greater than a year. The significance of sticky prices and the degree of stickiness continue to be an important area of research. The recent evidence suggests a substantial degree of stickiness, meaning that short-run movements of the nominal rate will be transmitted one-to-one to the real exchange rate.

Finally, this short-run volatility pans out in the long run, at least in terms of real exchange rate deviations. As summarized by Rogoff (1996), the consensus view in the profession is that purchasing power parity holds in the long run. The average half-life of real exchange rate deviations from trend is around three to four years. More recent evidence has challenged this view. Imbs et al. (2005) have found that at the sectoral level the mean reversion is even stronger (half-lives of around a year). They argue that the very strong persistence of the aggregate real exchange rate measures is due to aggregation bias. The jury is still out on this matter. Nevertheless, it is clear that fluctuations of the real exchange rate are short-lived – where short is to be determined.

In summary, short-run fluctuations of the nominal and real exchange rates come in tandem, and therefore, from the policy point of view, stabilizing one is equivalent to stabilizing the other.

Implications of exchange rate volatility

What are the advantages or disadvantages of exchange rate volatility? Because the variance is regime dependent – meaning fixed exchange rate regimes will have lower variance and flexible regimes will have higher variance – answering this question is closely linked to the advantages and disadvantages of the different exchange rate regimes. To simplify the discussion, let us concentrate on the two extremes: fixed and flexible.

As highlighted by Friedman (1953), one of the main advantages of flexible exchange rates is that they allow an independent monetary policy. Under fixed exchange rates, shifts in the demand for currency imply portfolio recompositions that are met instantaneously by changes in the international reserves. In this sense, when exchange rates are fixed an expansion of the monetary base implies an immediate loss of reserves. Therefore, if the fixed exchange rate regime is credible, the interest rate differential between the domestic rate and the international rate has to be zero, or small, limiting the scope of monetary policy management.

Under a flexible regime, the appreciation or depreciation of the exchange rate, and the risk premium implied by those movements, entails the possibility that domestic and international interest rates differ. Hence, in the presence of a supply shock and sticky prices, flexible nominal rates allow for an easier adjustment of the external account and unemployment. Conversely, in a fixed exchange rate regime, a negative shock implies that the economy would need large fluctuations in real activity to produce the desired price and wage changes – that is, under fixed exchange rates a negative shock implies a large increase in unemployment to achieve a drop in the real wage.

Finally, as has been argued by Tornell and Velasco (1995), flexible regimes react immediately to any shock – hence, irresponsible fiscal policy (for example) is felt immediately through a nominal depreciation, instead of a prolonged decline in international reserves, and an ultimate collapse.

In all these dimensions, a flexible exchange rate – and therefore a volatile exchange rate – seems to be

superior to a fixed exchange rate. Can the exchange rate be excessively volatile, and generate costs to the economy?

Indeed, it is possible that exchange rates are excessively volatile. The first explanation was provided by Dornbusch, and the second was advanced by Mussa.

As Rudiger Dornbusch used to say in most seminars at Massachusetts Institute of Technology, ‘exchange rates are determined in stock markets’, emphasizing the asset price view of the exchange rate he shared, and developed. Indeed, one of the most influential theories in international economics was Dornbusch's (1976) *overshooting* theory of the exchange rate. (In my view, his paper is the most influential idea in international open macroeconomics since the mid-1970s.) The simple intuition of the overshooting model is that the nominal exchange rate has to satisfy two equations or restrictions: first, an asset equation (in this case the uncovered interest rate parity); and second, a ‘real’ equation (long-run purchasing power parity). In a world of fixed prices in the short run, changes in the stance of monetary policy create excessive volatility of the nominal exchange rate – where excess is here defined as overshooting. For instance, after a loosening of monetary policy, we know that in the long run the exchange rate has to depreciate. However, the current increase in the money supply implies a reduction of the nominal interest rate. In order to satisfy the uncovered interest rate parity while prices adjust, the exchange rate should be appreciating (instead of depreciating). Therefore, the exchange rate has to depreciate excessively today, so that it can appreciate on the path towards the new equilibrium. This theory implies that exchange rates in the short run are more volatile than fundamentals.

The second theory explaining excessive exchange rate volatility is based on the fact that speculation in the market can be destabilizing. In standard models with flexible prices, speculation in the market is usually good because price movements will reveal the fundamentals that private agents are using to value the asset. However, if there is no full price revelation, speculation might destabilize the exchange rate market. This has been called in the literature the *magnification effect* (Mussa, 1976). The reason is that expectations might not imply that speculators purchase assets when prices are going down, and sell them when prices are going up. Their actions depend on the properties behind the exchange rate stochastic process, and how agents form their expectations. Therefore, it is possible that excessive volatility of the nominal exchange rate is due to shocks to expectations, and not to fundamentals. This *noise* in the nominal exchange rate plays no role in the adjustment process.

It is possible to argue that excess exchange-rate volatility has real costs. The change in the real exchange rate has implications for production allocation – tradable versus non-tradable sectors, on hedging, on investment, and so on. It has been documented that a more volatile exchange rate reduces the amount of trade (Frankel and Wei, 1993), increases the pressures toward protectionism (see Dornbusch and Frankel, 1987, for a survey of the protectionist forces that appeared during the 1980s in the United States), increases the degree of persistence of inflation or deflation, so slowing the adjustment of the real exchange rate (Obstfeld, 1995), and reduces the development of the financial sector (Aghion et al., 2005; Eichengreen, Hausmann and Panizza, 2003). It is important to highlight that this empirical evidence is mostly suggestive. There are problems of simultaneous equations and omitted variables in answering any of these questions. Furthermore, there are no available instruments for the exchange rate volatility – leaving the literature mainly reporting correlations as opposed to causal relationships. Indeed, this is perhaps an area of research that will need to be revived in the future. However, the ‘consensus’ points to the costs emphasized before. Interestingly, even if the academic literature does not have a clear view on

whether the volatility of the exchange rate is good or bad for the economy, it seems that policymakers and practitioners agree that exchange rate volatility is costly.

Dealing with exchange rate volatility

Is exchange rate volatility a policy choice? Can exchange rate volatility be controlled? In other words, if we were to accept that exchange rate volatility is detrimental to the economy because it is excessive, then what is the policy advice? How can it be reduced?

These questions are related to the theories behind exchange rate determination. For example, if fundamentals determine the exchange rate, is should be the case that the exchange rate, and its volatility, can be controlled, or at least ameliorated, by changing the fundamentals. On the other hand, if exchange rates and their volatility are unrelated to fundamentals, then very little can be done from the policy point of view.

A substantial literature reports a ‘disconnect’ between exchange rates and fundamentals in the short run. The seminal paper by Messe and Rogoff (1983) argues that the best predictor of tomorrow's exchange rate is today's exchange rate. They show that in the short run the random-walk assumption beats most models of the exchange rate. Recently, however evidence has emerged that models of the exchange rate have performed better than the random walk for medium- and long-run horizons. (See Chin and Messe, 1995, for the first paper in this area, and see the subsequent papers by Menzie Chin on the subject.) This disconnect implies that very little can be done in the short run to control for the exchange rate volatility – other than fixing it. In fact, Evans and Lyons (2002) show that 70 per cent of the variation of the nominal exchange rate in the short run is explained by order flows in the market – meaning that market micro-structure factors dominate the fluctuations of the exchange rate.

More interesting is the fact that in small open economies the exchange rate is governed by the fundamentals of other countries. The very well-known anomaly called ‘contagion’ implies that excess exchange rate volatility is affected by crises experienced by trading partners and in countries that share financial linkages. (See Forbes and Rigobon, 2001; 2002, and Kaminsky, Reinhart and Vegh, 2003, for detailed surveys of the empirical literature, and see Pavlova and Rigobon, 2006, 2007, for the theories of contagion.) These are rarely subject to the influence of policymakers, and very little can be done in this respect.

In summary, the short-run volatility of a flexible exchange rate cannot be controlled by policymakers. The short-run fluctuations depend either on market participants’ views or on other countries’ fundamentals. In both cases, the only response open to the monetary authorities is to move toward a fixed regime – which is indeed what most countries do. This is a very imperfect policy measure, but unfortunately the only one available. Hence, there are two types of country. In the first, market participants’ views are very volatile, or are subject to massive external shocks; these countries end up adopting fixed exchange rate regimes. For those countries the short-run fluctuation of the exchange rate is so costly that giving up monetary policy seems a minor issue. The second type of country consists of those that can bear the cost of the short-run exchange rate volatility, and have been moving towards a flexible rate – and towards inflation targeting. In fact, in the life of a single country it is easy to think of times in which nominal shocks dominate the economy (Argentina in the 1990s), which they responded to by fixing the exchange rate. However, after they introduced a fixed exchange rate the economy was

mostly governed by supply shocks (high unemployment) and needed a flexible regime, which Argentina also implemented in 2002.

Final remarks

Exchange rate volatility is one of the most important policy matters in developed economies, and possibly *the* topic in developing ones. The rhetoric of public policy is that exchange rate volatility is costly. In this article we have tried to understand the economic forces behind this claim.

First, we observed that in the short run the nominal exchange rate and the real exchange rate move in tandem. This is mainly due to the presence of sticky prices and real rigidities. This means that the discussion about volatility must embrace both the real and the nominal exchange rate.

Second, we have reviewed some of the evidence pointing out that exchange rate volatility is indeed costly for the economy. Investment is lower, growth is lower, real interest rates are higher, there are costly resource allocations between tradable and non-tradable sectors, and so on. This evidence, however, is not conclusive. There are econometric challenges, such as simultaneity and omitted variable biases, that have not yet been overcome. Nevertheless, it seems fair to say that the consensus points out to exchange rate volatility being costly to economies.

Third, we have observed that, even if we were to accept that exchange rate volatility is costly, there is very little that policymakers can do. There is a tremendous disconnect between exchange rates and fundamentals in the short run, and therefore policymakers are left with only one instrument to deal with exchange rate volatility: to fix it. That implies that only countries that face large demand shocks end up fixing their exchange rates, while most countries have been moving towards more flexible regimes.

See Also

- purchasing power parity

Bibliography

Aghion, P., Angeletos, G., Banerjee, A. and Manova, K. 2005. Volatility and growth: credit constraints and productivity-enhancing investment. Working Paper No. 11349. Cambridge, MA: NBER.

Alvarez, L.J, Dhyne, E., Hoeberichts, M., Kwapil, C., Le Bihan, H., Lunnemann, P., Martins, F., Sabbatini, R., Stahl, H., Vermeulen, P. and Vilumen, J. 2005. Sticky price in the Euro area: a summary of new micro evidence. Working Paper No. 563. Frankfurt am Main: European Central Bank.

Bils, M. and Klenow, P. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112, 947–85.

Chin, M. and Messe, R. 1995. Banking on currency forecasts: how predictable is change in money? *Journal of International Economics* 38, 161–78.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84, 1161–76.

Dornbusch, R. and Frankel, J. 1987. Macroeconomics and protection. In *US Trade Policies in a Changing World Economy*, ed. R. Stern. Cambridge, MA: MIT Press.

Eichengreen, B., Hausmann, R. and Panizza, U. 2003. Currency mismatches, debt intolerance and original sin: why are not the same and why it matters. Working Paper No. 10036. Cambridge, MA: NBER.

Evans, M. and Lyons, R. 2002. Order flow and exchange rate dynamics. *Journal of Political Economy* February 110, 170–80.

Forbes, K. and Rigobon, R. 2001. Contagion in Latin America: definitions, measurement, and policy implications. *Economia* 1(2), 1–46.

Forbes, K. and Rigobon, R. 2002. No contagion, only interdependence: measuring stock market co-movements. *Journal of Finance* 57, 2223–61.

Frankel, J. and Wei, S.-J. 1993. Trade blocs and currency blocs. In *The Monetary Future of Europe*, ed. G. de la Dehesa. London: CEPR.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Gopinath, G. and Rigobon, R. 2006. Sticky borders. Working Paper No. 12095. Cambridge, MA: NBER.

Imbs, J., Mumtaz, H., Ravn, M. and Rey, H. 2005. PPP strikes back: aggregation and the real exchange rate. *Quarterly Journal of Economics* 120, 1–43.

Kaminsky, G.L., Reinhart, C.M. and Vegh, C.A. 2003. The unholy trinity of financial contagion. *Journal of Economic Perspectives* 17(4), 51–74.

Kydland, F. and Prescott, E. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

Messe, R. and Rogoff, K. 1983. Empirical exchange rate models of the seventies: do they fit out of sample? *Journal of International economics* 14, 3–24.

Mussa, M. 1976. The exchange rate, the balance of payments, and monetary and fiscal policy under a regime of controlled floating. *Scandinavian Journal of Economics* 78, 229–48.

Mussa, M. 1986. Nominal exchange rate regimes and the behavior of real exchange rates: evidence and implications. In *Real Business Cycles, Real Exchange Rates, and Actual Policies*, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

Obstfeld, M. 1995. International currency experience: new lessons and lessons relearned. *Brookings Papers on Economic Activity* 1995(1), 119–220.

Obstfeld, M. and Rogoff, K. 2005. Global current account imbalances and exchange rate adjustments. *Brookings Papers on Economic Activity* 2005(1), 67–123.

Pavlova, A. and Rigobon, R. 2006. The role of portfolio constraints in the international propagation of shocks. Mimeo, Massachusetts Institute of Technology.

Pavlova, A. and Rigobon, R. 2007. Asset prices and exchange rates. *Review of Financial Studies* 20, 1139–81.

Rogoff, K. 1996. The purchasing power parity puzzle. *Journal of Economic Literature* 34, 647–68.

Tornell, A. and Velasco, A. 1995. Fiscal discipline and the choice of exchange rate regime. *European Economic Review* 39, 759–70.

How to cite this article

Rigobon, Roberto. "exchange rate volatility." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000312> doi:10.1057/9780230226203.0519

The New Palgrave Dictionary of Economics Online

exchange

Robert B. Wilson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic studies of exchange examine processes in which agents obtain gains from trade, e.g. bilateral bargaining and contracting, auctions, and multilateral markets. Prominent descriptive theories include idealized versions that assume agents simply respond to prices that clear markets. Realistic versions recognize effects of procedural rules and strategic behavior, and various impediments such as incomplete or unenforceable contracts, insufficient markets, imperfect information, and incomplete observability. Normative versions design market procedures, contract forms, and settlement rules that strengthen incentives and promote efficient outcomes.

Keywords

adverse selection; arbitrage; asymmetric information; auctions; bargaining; barter; bundling; coalitions; collusion; commitment; competition; complementary goods; complete information; cores; decentralization; double auctions; Edgeworth, F. Y.; efficiency prices; efficient allocation; entry; excess demand and supply; exchange; existence of equilibria; expected utility; fixed-point problems; game theory; impatience; implicit contracts; incentive compatibility; incomplete contracts; incomplete information; incomplete observability; incomplete markets; information aggregation and prices; inside information; labour market contracts; marginal rates of substitution; market design; mechanism design; monopoly rents; moral hazard; multiple equilibria; nonlinear pricing; optimal contracts; preference falsification; price discrimination; principal and agent; prisoner's dilemma; private information; product differentiation; quality differentiation; rational behaviour; rational expectations; reputation; resale markets; reservation prices; revelation principle; risk sharing; sealed-bid auctions; sequential rationality; Shapley value; signalling; specialization; strategic behaviour; temporary equilibrium; terms of trade; uniqueness of equilibrium; Vickrey, W. S.; Walras, L.; Walras's Law; Walrasian exchange model

Article

The scope of economics includes allocation of scarce resources. Allocation comprises production and exchange, according to a division between processes that transform commodities and those that transfer control. For production or consumption, exchange is essential to efficient use of resources. In production it allows decentralization and specialization; and, for consumption, agents with diverse endowments or preferences require exchange to obtain maximal benefits. Voluntary exchange involves trading bundles of commodities or obligations to the mutual advantage of all parties to the transaction. If two agents have differing marginal rates of substitution, then there exists a trade benefiting both. The advantages of barter extend widely, for example, to trade among nations and among legislators (‘vote trading’), but here it suffices to emphasize markets with enforceable contracts for trading private property unaffected by externalities, and with money as a medium of exchange.

In a market economy using money or credit, terms of trade are usually specified by prices denominated in money. Besides purchases at prices posted by producers and distributors, exchange occurs in bargaining, auctions, and other contexts with repeated or competitive offers. In institutionalized ‘exchanges’ for trading commodities, brokers offer bid and ask prices; and, for trading financial instruments, specialists cross buy and sell orders and maintain markets continually by quoting bid and ask prices and trading for their own accounts.

Theories of exchange

Records of transaction prices and quantities are the raw data of many empirical studies of economic activity, and explanation of these data is a main purpose of economic theory. Theories of exchange attempt to predict the terms of trade and the resulting transactions from the market structure and the agents’ attributes, such as endowments, productive opportunities, preferences, and information. Also relevant are the markets accessible, the trading rules used, and the contracts available. These may depend on property rights, search or transaction costs, and on events observable or verifiable to enforce contracts. If a particular trading rule is used, such as an auction, then it specifies the actions allowed each agent in each contingency, and the trades resulting from each combination of the agents’ actions. These features are the ingredients of experimental designs to test theories, and they motivate models used for empirical estimates of market behaviour. Normative considerations are also relevant, so welfare analyses study the distributional consequences of alternative trading procedures and contracts.

Most theories hypothesize that each agent acts purposefully to maximize its (expected utility of) gain from trade. Some behaviour may be erratic, customary, or reflect dependency on a status quo, but experimental and empirical evidence substantially affirms the hypothesis of ‘rational’ behaviour, at least in the aggregate. Although more general theories are available, the main features are explained by preferences that are quite regular, as assumed here: monotone, convex, and possibly allowing risk aversion and impatience.

Typically there are many efficient allocations of a fixed endowment, since any allocation that equates agents’ marginal rates of substitution is efficient. In the case of risk sharing, for example, an allocation is efficient if all agents achieve the same marginal rates of substitution between incomes in every two states. The distribution of endowments among agents evidently matters, however, and a major accomplishment has been the identification of a small set of salient efficient allocations. Named for Léon Walras, this set is a focus of nearly all theories, in the sense that other allocations are explained by

departures from the Walrasian model.

An allocation is Walrasian if it is obtainable by trading at prices such that it would cost each agent more to obtain a preferable allocation. That is, items are bought at uniform prices available to all, and each agent chooses a preferred trade within a budget constraint imposed by the values of goods bought and sold. If markets are complete, then a Walrasian allocation is necessarily efficient, because another allocation preferred by every agent would cost each agent more at the current prices and therefore more in total, which cannot be true if the preferred allocation is a redistribution of the present one. Conversely, each efficient allocation is Walrasian without further trade, since the agents' common marginal rates of substitution serve as the price ratios. The basic formulation considers trade for delivery in all future contingencies, but refined formulations elaborate the realistic case that markets reopen continually and trade is confined to a limited variety of contracts for immediate and future delivery, possibly contingent on events.

Sufficient conditions for Walrasian allocations to exist have been established. Mainly these require that agents' preferences are convex and insatiable, and that each agent has an endowment sufficient to obtain a positive income. For 'most' economies the number of Walrasian allocations is finite, but uniqueness requires strong assumptions on substitution and income effects. Walrasian allocations and prices for a specific model can be computed by solving a fixed-point problem, for which general methods have been devised. The task is complex (for example, linear models with integer data can yield irrational prices) but an important simplifying feature is that Walrasian prices depend only on the distribution of agents' attributes, and in particular only on the aggregate excess demand function. Essentially, any continuous function satisfying Walras's law (at any prices the value of excess demand is zero) and homogeneity in prices is the excess demand for some economy.

The key requirement for a Walrasian allocation is that each agent's benefit is maximized within its budget imposed by the assigned prices, and that markets clear at those prices. However, complete exploitation of all gains from trade may be precluded by incomplete markets, pecuniary externalities (such as absence of necessary complementary goods), insufficient contracts, or strategic behaviour. If producers with monopoly power restrain output to elevate prices, or practise any of the myriad forms of price discrimination, then the resulting allocation need not be Walrasian. Much discrimination segments markets via quality differentiation or bundling, but equally common is discriminatory pricing of the various conditions of delivery (for example, spatial, temporal, service priority) or if purchases can be monitored and resale markets are absent, by nonlinear pricing of quantities (for example, quantity discounts, loyalty programmes, two-part and block-declining tariffs).

The Walrasian model of exchange is substantially defined by the absence of such practices affecting prices. It also relies on a fixed specification of markets, agents, products, and contracts. The theory of economies with large firms having power to influence prices and to choose product designs is significantly incomplete. The deficiencies derive partly from inadequate formulations, and partly from technical considerations – characterizations and even the existence of equilibria depend on special structural features. For example, the simplest models positing simultaneous choices of qualities and prices by several firms lack equilibria; models with sequential choices encounter similar obstacles but to lesser degrees. In addition, if lump-sum assessments imposed on customers are precluded, then recovering firms' large fixed costs may require nonlinear pricing or price discrimination.

Market clearing is also essential to the Walrasian model since prices are determined entirely by the required equality of demand and supply, including inventories in dynamic contexts. In contrast,

successive markets with overlapping generations of traders need not clear ‘at infinity’. Such markets can exhibit complicated dynamics even if the underlying data of the economy are stationary. Similarly, continually repeated markets where buyers and sellers arrive at the same rate that others depart after completing transactions (for example, real estate) admit non-Walrasian prices or may have persistent excess on one side of the market if search or dispersed bargaining prevents immediate clearing.

Factors affecting exchange

When it is that among the feasible allocations the best prediction might be one of the Walrasian allocations, has been answered in several ways.

Competition is the first answer. On the supply side, for instance, with many sellers each one's incentive to defect from collusive pricing arrangements is increased. Absent collusion, if prices reflect total supplies offered on the market and each seller chooses its optimal supply in response to anticipations of other sellers’ supplies, then each seller's optimal percentage profit margin declines inversely with the number of sellers offering substitutes. Price discrimination, such as nonlinear pricing, is inhibited if there are many sellers, resale markets are available, or customers’ purchases are difficult to monitor. Absent capacity limitations, direct price competition among close or perfect substitutes erodes profits since undercutting is attractive. Although these conclusions are weakened to the extent that buyers incur search or switching costs, easy entry incurring low sunk costs remains important to ensure that markets are contestable and monopoly rents are eliminated. Monopoly rents are often substantially dissipated in entry deterrence, price wars, and other competitive battles to retain or capture monopoly or oligopoly positions. This is true both when entrants bring perfect substitutes and also more generally, since entrants tend to fill in the spectra of quality attributes and conditions of delivery.

Arbitrage is important in markets for commodities with standardized qualities, especially financial assets and derivatives such as options. To the extent that the contingent returns from one asset replicate those from a bundle of other assets, or from some trading strategy, its price is linked to the latter. Importantly, repeated opportunities to trade contingent on events enable a few securities to substitute for a much wider variety of contingent contracts.

One form of the competitive hypothesis emphasizes that each subset of the traders can redistribute their endowments among themselves. For example, a seller and those buyers who purchase from him are a coalition redistributing their resources among themselves. A core allocation is such that no coalition can redistribute its endowments to advantage every member. The core allocations include the Walrasian allocations. A basic result first explored by F.Y. Edgeworth establishes that as the economy is enlarged by adding replicates of the original traders, the set of core allocations shrinks to the Walrasian allocations. Deeper analyses of core allocations take account of agents’ private information, but in this context the relation to Walrasian allocations is tenuous.

Another view emphasizes that in an economy so large that each agent's behaviour has an insignificant effect on the terms of trade, every trader's best option is to maximize its gain from trade at the prevailing prices. For example, any one trader's potential gain from behaviour that influences prices becomes insignificant as the set of traders expands, provided the limit distribution is ‘atomless’. Similar results obtain for various models of markets with explicit price formation via auctions or bilateral bargaining. Generally, an efficient allocation is necessarily Walrasian if each agent is unnecessary to attainment of others’ gains from trade. An idealized formulation considers an atomless measure space of agents in

which only measurable sets of agents matter and thus the behaviour of each single agent is inconsequential. In this case the Walrasian allocations are the only core allocations. Similarly, a Walrasian allocation results from the Shapley value in which each agent shares in proportion to his average marginal contributions to randomly formed coalitions.

Structural features of trading processes suggest alternative hypotheses. Matching problems (for example, workers seeking jobs) admit procedural rules that with optimal play yield core allocations, and for a general exchange economy an appropriately designed auction yields a core allocation. Other games have been devised for which optimal strategies of the agents result in a Walrasian allocation. Continual bilateral bargaining among dispersed agents with diverse preferences, in which agents are repeatedly matched randomly and one designated to offer some trade to the other, also results in a Walrasian allocation from optimal strategies. In a related vein, several methods of selecting allocations create incentives for agents to falsify reports of their preferences, but if they do this optimally then a Walrasian allocation results. Quite generally, any process that is fair in the sense that all agents enjoy the same opportunities for net trades yields a Walrasian allocation. In one axiomization, some signal is announced publicly and then based on his preferences each agent responds with a message that affects the resulting trades: if a core allocation is required, and each signal could be the right signal for some larger economy, then the signal must be essentially equivalent to announcement of a Walrasian price to which each agent responds with his preferred trade within his budget specified by the price.

Traders' impatience can also affect the terms of trade. In the simplest form of impatience, agents discount delayed gains from trade. Dynamic play is assumed to be sequentially rational in the sense that a strategy must specify an optimal continuation from each contingency – this strong requirement severely restricts the admissible equilibria. For example, if a seller and a buyer alternate proposing prices for trading an item, then in the unique equilibrium trade occurs immediately at a price dependent on their discount rates. As the interval between offers shrinks, the seller's share of the gains from trade becomes proportional to the relative magnitude of the buyer's discount rate; for example, equal rates yield equal division. Extensions to multilateral contexts produce analogous results. A monopolist with an unlimited supply selling to a continuum of buyers might plausibly extract favourable terms, but actually, in any equilibrium in which the buyers' strategies are stationary, as the interval between offers shrinks the seller's profit disappears and all trade occurs quickly at a Walrasian price. Similarly, a durable-good manufacturer lacking control of resale or rental markets has an incentive to increase the output rate as the production period shrinks or to pre-commit to limited capacity. This emphasizes that monopoly power depends substantially on powers of commitment stemming from increasing marginal costs, capacity limitations, or other sources. However, impatience and sequential rationality can produce inefficiencies in product design, since then a manufacturer may prefer inferior durability, or in market structure, since a seller may prefer to rent rather than sell durable goods.

Complete information is a major factor justifying predictions of Walrasian prices. Many theories predict Walrasian outcomes when there is complete information and agents have symmetric trading opportunities, but incomplete information often produces departures from the Walrasian norm. Although information may be productively useful, in an exchange economy the arrival of information may be disadvantageous to the extent that risk-averse agents forgo insurance against its consequences. A basic result considers an exchange economy that has reached an efficient allocation before some agents receive further private information, and this fact is common knowledge: the predicted response is no further trade, though prices may change.

Each efficient allocation has ‘efficiency prices’ that reflect the marginal rates of substitution prevailing – in the Walrasian case all trades are made at these prices. They summarize a wealth of information about technology, endowments and preferences. Prices and other endogenous observables are therefore not only sufficient instruments for decentralization but also carriers of information. If information is dispersed among agents then Walrasian prices are signals, possibly noisy, that can inform agents’ trading. Models of ‘temporary equilibrium’ envision a succession of markets, in each of which prices convey information about future trading opportunities. ‘Rational expectations’ models assume that each agent maximizes an expected utility conditioned on both his private information and the informational content of prices. In simple cases prices are sufficient statistics that swamp an agent's private information. In complex real economies the informational content of prices may be elusive; nevertheless, markets are affected by inferences from prices (e.g. indices of stock and wholesale prices) and various models attempt to include these features realistically. Conversely, responses of prices to events and disclosures by firms are studied empirically.

The privacy of each agent's information about his preferences and endowment affects the realized gains and terms of trade. In some cases the relative prices of ‘qualities’ provide incentives for self-selection. An example is a product line comprised of imperfect substitutes, in which price increments for successive quality increments induce customers to select according to their preferences. Several forms of discrimination in which prices depend on the quality (for example, the time, location, priority or other circumstances of delivery) or, if resale is prevented, the quantity purchased, operate similarly.

Absence of the relevant contingent contracts is implicitly a prime source of inefficiencies and distributional effects. Trading may fail if adverse selection precludes effective signalling about product quality: without quality assurances or warranties, each price at which some quality can be supplied attracts sellers offering lesser qualities. Investments in signals, possibly unproductive ones, that are more costly for sellers supplying inferior qualities induce signal-dependent schedules in which the price paid depends on the signal offered. For example, to signal his ability a worker may over-invest in education or work in a job for which he would be underqualified on efficiency grounds. If buyers make repeat purchases based on the quality experienced from trying a product then the initial price itself, or even dissipative expenditures such as uninformative advertising, can be signals used by the seller to induce initial purchases.

Principal–agent relationships in which a risk-averse agent has superior information or his actions cannot be monitored completely by the principal require complex contracts. For example, in a repeated context with perfect capital markets and imperfect insurance, the optimal contract provides the agent with a different reward for each measurable output, and the total remuneration is the accumulated sum of these rewards. Contracting is generally affected severely by limited observation of contingencies (either events or actions relevant to incentives) and in asymmetric relationships nonlinear pricing is often optimal. For example, insurance premia may vary with coverage to counter the effects of adverse selection or moral hazard.

Labour markets are replete with complex incentives and forms of contracting, partly because workers cannot contract to sell labour forward and partly because labour contracts substitute for imperfect loan markets and missing insurance markets (for example, against the risk of declining productivity). Workers may have superior information about their abilities, technical data, or effort and actions taken; and firms may have superior information about conditions affecting the marginal product of labour. Incentives for immediate productivity may be affected by conditioning estimates of ability on current

output, or by procedures selecting workers for promotion to jobs where the impact of ability is multiplied by greater responsibilities. The complexity of the resulting incentives and contracts reflects the multiple effects of incomplete markets and imperfect monitoring.

In the context of trading rules that specify price determination explicitly, analyses of agents' strategic behaviour emphasize the role of private information. The trading rule and typically the probability distribution of agents' privately known attributes are assumed to be common knowledge; consequently, formulations pose games of incomplete information. An example is a sealed-bid auction in which the seller awards an item to the bidder submitting the highest price. Suppose each bidder observes an imperfect estimate that is independently and identically distributed (i.i.d.) conditional on the unknown value of the item. With equilibrium bidding strategies, as the number of bidders increases the maximal bid converges in probability to the expectation of the value conditional on knowing the maximum of all the samples; for the common distributions this implies convergence to the underlying value. Alternative auction rules are preferred by the seller according to the extent that the procedures dilute the informational advantages of bidders (for example, progressive oral bidding has this effect) and exploit impatience and risk aversion. Rules can be constructed that maximize the seller's expected revenue: if bidders' valuations are i.i.d. then for the common distributions awarding the item to the highest bidder at the first or second-highest bid is optimal, subject to an optimal reservation price set by the seller. In such a second-price or oral progressive auction with no reservation price, bidders offer their valuations, so the price is Walrasian.

Another example is a double auction, used in various commodity and financial markets, in which multiple buyers and sellers submit bid and ask prices and then a clearing price is selected from the interval obtained by intersecting the resulting demand and supply schedules. For a restricted class of models, requiring sufficiently many buyers and sellers with i.i.d. valuations, a double auction is incentive efficient, in the sense that there is no other trading rule that is sure to be preferred by every agent; also, as the numbers increase the clearing price converges to a Walrasian price.

The effects of privileged 'inside' information held by some traders have been studied in the context of markets mediated by brokers and specialists, as in most stock and option markets. The results show that specialists' strategies impose all expected losses from adverse selection on uninformed traders. Specialists may further profit from knowledge of the order book and immediate access to trading opportunities.

Private information severely affects bargaining. With alternating offers even the simplest examples have many equilibria, plausible criteria can select different equilibria, and a variety of allocations are possible. In most equilibria, delay in making a serious offer is a signal that a seller's valuation is not low or a buyer's is not high; or the offers made limit the inferences the other party can make about one's valuation. When both valuations are privately known, signalling must occur in some form to establish that gains from trade exist. Typically all gains from trade are realized eventually, but with significant costs of delay. Applications extend beyond purely economic contexts, such as to negotiations to settle a law suit.

In a special case, a seller with a commonly known valuation repeatedly offers prices to a buyer with a privately known valuation: assume that the buyer's strategy is a stationary one that accepts the first offer less than a reservation price depending on his valuation. As mentioned previously for the monopoly context, as the period between offers shrinks the seller's offers decline to a price no more than the least possible buyer's valuation and trade occurs quickly – thus the buyer captures most of the gains. Even

with alternating offers, the buyer avoids serious offers if his valuation is high and the periods are short. Thus, impatience, frequent offers, and asymmetric information combine to skew the terms of trade in favour of the informed party.

The premier instance of exchange is the commodity trading pit in which traders around a ring call out bid and ask prices or accept others' offers. These markets operate essentially as multilateral versions of bargaining but with endogenous matching of buyers and sellers: delay in making or accepting a serious offer can again be a signal about a trader's valuation, but with the added feature that 'competitive pressure' is a source of impatience. That is, a trader who delays incurs a risk that a favourable opportunity is usurped by a competing trader. These markets have been studied experimentally with striking results: typically most gains from trade are realized, at prices eventually approximating a Walrasian clearing price, especially if the subjects bring experience from prior replications. However, if complicated 'rational expectations' features are added, then subjects may fail to infer all the information revealed by offers and transactions.

Trading rules can be designed to maximize the expected realized gains from trade, using the 'revelation principle'. Each trading rule and associated equilibrium strategies induce a 'direct revelation game' whose trading rule is a composition of the original trading rule and its strategies; thus in equilibrium each agent has an incentive to report accurately his privately known valuation. In the case that a buyer and a seller have valuations drawn independently according to a uniform distribution, the optimal revelation rule is equivalent to a double auction in which trade occurs if the buyer's bid exceeds the seller's offer, and the price used is halfway between these. Basic theorems establish that private information among buyers and sellers precludes realization of all the gains from trade – but if traders are symmetric, as when a trader might buy or sell depending on the price, then sometimes full efficiency can be attained. Generally, with many buyers and sellers and an optimal trading rule, the expected unrealized gain from trade declines quickly as the numbers of buyers and sellers increase. Such static models depend, however, on the presumption that subsequent trading opportunities are excluded.

Enforceable contracts facilitate exchange, and most theories depend on them, but they are not entirely essential. Important in practice are 'implicit contracts' that are not enforceable except via threats of discontinuing the relationship after the first betrayal. Similarly, in an infinitely repeated situation, if a seller chooses a product's quality (say, high or low) and price before sale, and a buyer observes the quality only after purchasing, then the buyer's strategy of being willing to pay currently only the price associated with the previously supplied quality suffices to induce continual high quality.

Studies of exchange without enforceable contracts focus on the Prisoner's Dilemma game in which both parties can gain from exchange, but each has an incentive to renege on his half of an agreement. In any finite repetition of this game with complete information the equilibrium strategies predict no agreements, since each expects the other to renege. Infinite repetitions can sustain agreements enforced by threats of refusal to cooperate later. With incomplete information, reputational effects can sustain agreements until near the end. For example, if one party thinks the other might surely reciprocate cooperation then he has an incentive to cooperate until first betrayed, and the other has an incentive to reciprocate until defection becomes attractive near the end. Reputations are important also in competitive battles among firms with private cost information: wars of attrition select the efficient survivors.

In sum, the Walrasian model remains a paradigm for efficient exchange under 'perfect' competition in which equality of demand and supply is the primary determinant of the terms of trade. Further analysis

of agents' strategic behaviour with private information and market power elaborates the causes of incomplete or imperfectly competitive markets that impede efficiency, and it delineates the fine details of endogenous product differentiation, contracting, and price formation essential to the application of the Walrasian model.

Game-theoretic analysis of exchange

Studies of exchange rely increasingly on game-theoretic methods. These are useful to study strategic behaviour in dynamic contexts; to elaborate the roles of private information, impatience, risk aversion, and other features of agents' preferences and endowments; to describe the consequences of incomplete markets and contracting limited by monitoring and enforcement costs; and to establish the efficiency properties of the common trading rules. They also integrate theories of exchange with theories of oligopolistic collusion, product differentiation, discriminatory pricing, and other strategic behaviour by producers. Technically, the game-theoretic approach enables a transition from theories of a large competitive economy with a specified distribution of agents' attributes, to theories of an economy with few agents, each having private information and acting strategically to exploit opportunities.
Game-theoretic methods are especially useful in market design, that is, devising trading rules and procedures that yield outcomes that are efficient subject to the limitations imposed by participants' private information and strategic behaviour. Innovative designs are used for auctions of government procurements and privatization of assets (for example, spectrum licences, Treasury securities), for wholesale commodity markets such as electricity and gas among many others, and some markets for transport. Some of these are 'smart markets' in that the allocation is derived from an elaborate optimization that takes account of bids and offers for several commodities and various technical constraints – such as transmission limits and reserve requirements in the case of electricity. With the advent of electronic commerce, innovative designs are also used in some retail markets conducted as auctions.
A salient feature of the developing theory of market design is the role of alternative settlement rules. Unlike the Walrasian rule of settling all transactions at the market clearing prices, these settlements provide incentives for accurate reporting of benefits and costs. Most of these rules are derivatives of one proposed by W. Vickrey (1961) in which each buyer in an auction pays the highest rejected bid; and more generally (even for public goods), for the allocation he receives each trader is charged for the benefit thereby denied to others. Novel settlement rules that promote 'incentive compatibility' and thus discourage strategic behaviours that impair efficiency are hallmarks of the general theory of 'mechanism design'. Essentially, this approach to exchange replaces the idealized descriptive approach in the Walrasian paradigm with a normative 'engineering' construction of optimal trading and settlement rules, that is, a comprehensive contract that governs participation in the market. Participants' perceptions of the effectiveness of this contract can affect competition for trading volume among alternative market venues.

See Also

- adverse selection

- agency problems
- auctions (theory)
- bargaining
- contract theory
- cores
- efficient allocation
- epistemic game theory: incomplete information
- experimental economics
- fair allocation
- game theory
- general equilibrium
- incomplete contracts
- incomplete markets
- matching and market design
- mechanism design
- perfect competition
- price discrimination (theory)
- principal and agent (i)
- Walras's Law

Bibliography

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 89, 488–500.

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Ausubel, L.M., Cramton, P.C. and Deneckere, R.J. 2002. Bargaining with incomplete information. *Handbook of Game Theory*, vol. 3, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Debreu, G. 1959. *Theory of Value*. New York: John Wiley & Sons.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic*

Review 4, 235–46.

Fudenberg, D. and Tirole, J. 1993. *Game Theory*. Cambridge, MA: MIT Press.

Gül, F. and Postlewaite, A. 1992. Asymptotic efficiency in large exchange economies with asymmetric information. *Econometrica* 60, 1273–92.

Gül, F. and Sonnenschein, H. 1988. On delay in bargaining with one-sided uncertainty. *Econometrica* 56, 601–11.

Gül, F., Sonnenschein, H. and Wilson, R.B. 1986. Foundations of dynamic monopoly and the Coase conjecture. *Journal of Economic Theory* 39, 155–90.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Hölmstrom, B.R. and Milgrom, P.R. 1986. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55, 303–28.

Hölmstrom, B.R. and Myerson, R.B. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51, 1799–820.

Kreps, D.M., Milgrom, P.R., Roberts, D.J. and Wilson, R.B. 1982. Rational cooperation in the finitely repeated Prisoners' Dilemma. *Journal of Economic Theory* 27, 245–52.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27, 54–71.

Milgrom, P.R. 1979. A convergence theorem for competitive bidding with differential information. *Econometrica* 47, 679–88.

Milgrom, P.R. 2004. *Putting Auction Theory to Work*. Cambridge: Cambridge University Press.

Milgrom, P.R. and Stokey, N. 1982. Information, trade, and common knowledge. *Journal of Economic Theory* 26, 17–27.

Myerson, R.B. and Satterthwaite, M.A. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29, 265–81.

Radner, R. 1972. Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica* 40, 289–303.

Roberts, D.J. and Postlewaite, A. 1976. The incentives for price-taking behavior in large exchange economies. *Econometrica* 44, 115–28.

Roberts, D.J. and Sonnenschein, H. 1977. On the foundations of the theory of monopolistic competition. *Econometrica* 45, 101–13.

Roth, A.E., ed. 1995. *Game-Theoretic Models of Bargaining*. Cambridge: Cambridge University Press.

Roth, A.E. and Sotomayor, M. 1990. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge: Cambridge University Press.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50, 97–109.

Satterthwaite, M.A. and Williams, S.R. 1989. The rate of convergence to efficiency in the buyers' bid double auction as the market becomes large. *Review of Economic Studies* 56, 477–98.

Scarf, H. (With T. Hansen.) 1973. *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Schmeidler, D. 1980. Walrasian analysis via strategic outcome functions. *Econometrica* 48, 1585–93.

Schmeidler, D. and Vind, K. 1972. Fair net trades. *Econometrica* 40, 637–42.

Smith, V. 1982. Microeconomic systems as experimental science. *American Economic Review* 72, 923–55.

Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40, 549–63.

Sonnenschein, H. 1974. An axiomatic characterization of the price mechanism. *Econometrica* 42, 425–34.

Spence, A.M. 1973. *Market Signalling: Information Transfer in Hiring and Related Process*. Cambridge, MA: Harvard University Press.

Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders, *Journal of Finance* 16, 8–37.

Wilson, R.B. 1985. Incentive efficiency of double auctions. *Econometrica* 53, 1101–16.

Wilson, R.B. 1993. Design of efficient trading procedures. In *The Double Auction Market: Institutions, Theories, and Evidence*, ed. D. Friedman and J. Rust. Santa Fe Institute Studies in the Sciences of

Complexity. Reading, MA: Addison-Wesley.

How to cite this article

Wilson, Robert B. "exchange." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000146> doi:10.1057/9780230226203.0515

The New Palgrave Dictionary of Economics Online

exchangeability

David Draper

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Exchangeability is an invariance property of probability distributions that is central to the process of specifying Bayesian statistical models.
Exchangeability judgements play a role in Bayesian modelling analogous to judgments in frequentist modelling that observable quantities may be regarded as realizations of independent identically distributed (IID) random variables. Judgements of conditional exchangeability (given the values of relevant covariates), when combined with Bayesian nonparametric modelling, provide a principled and rather general approach to Bayesian model specification that can lead to well-calibrated inferences and predictions; other approaches to achieving this goal include cross-validation and Bayesian model averaging.

Keywords

Bayes' th; Bayesian nonparametric methods; Bayesian statistics; Bernoulli distributions; cumulative distribution functions; Dirichlet process priors; exchangeability; frequentist statistics; Markov chain Monte Carlo methods; model averaging; model specification; model uncertainty; Pólya trees; prior distributions; random variables; uncertainty

Article

Definition: A sequence $Y = (Y_1, \ldots, Y_n)$ of random variables (for $n \geq 1$) is *(finitely) exchangeable* if the joint probability distribution $p(y_1,\ldots,y_n)$ of the elements of y is invariant under permutation of the indices $(1,\ldots,n)$, and a countably infinite sequence $(y_1, y_2,\ldots)$ is *(infinitely) exchangeable* if every finite subsequence is finitely exchangeable.
The idea of exchangeability seems (Good, 1965; Bernardo and Smith, 1994) to be traceable back to Johnson (1924), who used the term *permutable*, and independently to Haag (1924). Other writers who made early use of the concept include Khintchine (1932); Fréchet (1943); Savage (1954), who called an

exchangeable sequence *symmetric*; and Hewitt and Savage (1955). But the deepest implications of the idea are due to de Finetti (1930), who in (1938) still referred to exchangeable sequences as *equivalent*; by (1970) the word had been translated from the Italian as *exchangeable*, and since then usage has stabilized around this terminology under the influence of de Finetti and his translators.

The concept is important because it plays a fundamental role in the specification of statistical models from a Bayesian point of view. Following the example of Good (1950) by referring to You as a generic rational person making uncertainty assessments, suppose that You will in the future get to see a finite sequence $Y = (y_1, \ldots, y_n)$ of binary observables; to illustrate the interplay between context and model, consider as an example the mortality outcomes (within 30 days of admission, say: 1=died, 0=lived) for a sequence of *n* patients with the same admission diagnosis (heart attack, say) at one particular hospital *H*, starting on the first day of next month. You acknowledge Your uncertainty about which elements in the sequence will be 0s and which 1s; suppose further that You find it natural (as in the Bayesian approach to statistics) to use random variables to quantify Your uncertainty. As de Finetti (1970) noted, in this situation Your fundamental imperative is to construct a *predictive* distribution $p(y_1,\ldots,y_n)$ that expresses Your uncertainty about the future observables, rather than – as is perhaps more common – to reach immediately for a standard family of *parametric* models for the y_i (that is, to posit the existence of a vector $\theta = (\theta_1, \ldots, \theta_k)$ of parameters and to model the observables by appeal to a family $p(y_i|\theta)$ of probability distributions indexed by θ).

Even though the y_i are binary, with all but the smallest values of *n* it still seems a formidable task to *elicit* from Yourself an *n*-dimensional predictive distribution $p(y_1,\ldots,y_n)$; it was while facing this challenge that de Finetti developed his version of the idea of exchangeability and its implications. As de Finetti observed, in the absence of any further information about the patients, You notice that Your uncertainty about them is exchangeable: if someone (without telling You) were to rearrange the order in which their mortality outcomes become known to You, Your predictive distribution would not change. This still seems to leave $p(y_1,\ldots,y_n)$ substantially unspecified, but de Finetti (1930) proved a remarkable theorem which shows (in effect) that all exchangeable predictive distributions for a vector of binary observables are representable as mixtures of Bernoulli sampling distributions. More formally,

Theorem 1: **(representation of exchangeable predictive distributions for binary observables [de Finetti, 1930])**: Suppose that You're willing to regard $(y_1,\ldots, y_n)$ as the first *n* terms in an infinitely exchangeable binary sequence $(y_1, y_2, \ldots)$; then, with $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$,

- $\theta = \lim_{n\to\infty} \bar{y}_n$ must exist, and the marginal distribution (given θ) for each of the y_i must be $p(y_i|\theta) = \text{Bernoulli}(\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$;
- $H(t) = \lim_{n\to\infty} P(\bar{y}_n \le t)$, the limiting cumulative distribution function (CDF) of the $\bar{y}_n$ values, must also exist for all *t* and must be a valid CDF, where *P* is Your joint probability distribution on $(y_1, y_2, \ldots)$; and
- Your predictive distribution for the first *n* observations can be expressed as

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} dH(\theta). \quad (1)$$

When (as will essentially always be the case in realistic applications) Your joint distribution *P* is sufficiently regular that *H* possesses a density (with respect to Lebesgue measure), $dH(\theta) = p(\theta)\, d\theta$, (1) can be written in a more accessible way as

$$p(y_1, \ldots, y_n) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n} p(\theta)\, d\theta, \quad (2)$$

where $s_n = \Sigma_{i=1}^{n} y_i = n\bar{y}_n$.

The interpretation of (2) provides a link with non-Bayesian statistical modelling, as follows. In the frequentist (repeated-sampling) approach to statistics, to bring probability into the picture it's necessary to tell a story in which the observable y_i are either literally a random sample from some population $\mathcal{P}$ or *like* what You would get if You took a random sample from $\mathcal{P}$. This is a somewhat awkward story to tell in the medical example above, because the patients whose mortality outcomes are $(y_1,\ldots,y_n)$ are not a random sample of anything; they're simply the exhaustive list of all patients arriving at hospital *H* (itself not randomly chosen), with heart attack as their admission diagnosis, in a particular (not randomly chosen) window of time. In spite of this difficulty, the standard frequentist model (with the same information base as that assumed above) would define θ as the mortality rate in $\mathcal{P}$ (whatever $\mathcal{P}$ might be) and would treat the y_i as measurements on a random sample from $\mathcal{P}$ by regarding random variables Y_i (whose observed values are y_i) as independent and identically distributed (IID) draws from the Bernoulli (θ) distribution $p(Y_i = y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$, which leads to the joint sampling distribution $p(Y_1 = y_1, \ldots, Y_n = y_n) = \Pi_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{s_n}(1-\theta)^{n-s_n}$. Thus, in interpreting Theorem 1 (with reference to equation (2) above), in Your predictive modelling of the binary y_i You may as well proceed *as if*

- there is a quantity called θ, interpretable both as the marginal death probability $p(y_i = 1|\theta)$ for each patient and as the long-run mortality rate in the infinite sequence $(y_1, y_2, \ldots)$;
- conditional on θ, the y_i are IID Bernoulli (θ); and
- θ can be viewed as a realization of a random variable (this is of course how all unknown quantities are treated in the Bayesian paradigm) with density $p(\theta)$.

In other words, exchangeability of Your uncertainty about a binary process is functionally equivalent to assuming the simple Bayesian *hierarchical* model (see, for example, Draper, 2007)

$$\begin{aligned} \theta &\sim p(\theta) \\ (y_i|\theta) &\overset{\text{IID}}{\sim} \text{Bernoulli}(\theta) . \end{aligned} \tag{3}$$

A number of points are worth noting.

First, exchangeability is not a property of the world; it's a judgment by You concerning Your uncertainty about the world. Two reasonable people, with different knowledge bases or different views on how that knowledge should be brought to bear on the issue at hand, may consider the same set of observables, and one may judge her uncertainty about those observables exchangeable while the other may not make the same judgement about his uncertainty (for example, if I know the gender of the patients in the medical example and You do not, and if there's evidence that the mortality rate for male and female heart attack patients differs by an amount that's large in clinical terms, then You may well judge Your uncertainty about the mortality outcomes exchangeable but I would be ill-advised to adopt an exchangeable model; see partial/conditional exchangeability below). This distinction between {the world} and {Your uncertainty about the world} is sometimes blurred by terminology – I might casually say ‘These patients are exchangeable’ when what I mean is ‘my uncertainty about these patients is exchangeable, as far as mortality is concerned’ – but failing to observe the distinction can lead to what Jaynes (2003) terms the *mind projection fallacy*, with undesirable consequences for clarity of thought.

Second, in both the frequentist and Bayesian modelling approaches it's helpful to employ a fiction involving random variables, but for different purposes: in the standard frequentist approach, You regard the y_i as realizations of random variables (as a way to build a useful probability model), even though (in observational settings like the medical example above) no random sampling was performed to arrive at the observables; and in the Bayesian approach, You regard θ as random (as a way to make good predictions of observables), even though in both the frequentist and Bayesian approaches θ has the same logical status, as a fixed unknown constant.

Third, since, for any random variables X and Y for which the following symbols have meaning, the density of Y can be expressed as $p(y) = \int p(y|x)\, p(x)\, dx$ – in other words, Y can be modelled either directly or as a mixture of the conditional distribution $p(y|x)$ with $p(x)$ serving as a mixing distribution – the predictive distribution in eq. (2) can be regarded as a mixture of Bernoulli sampling distributions with $p(\theta)$ as the mixing weights.

Fourth, mathematically $p(\theta)$ is just a mixing distribution, but (of course) statistically it has a more useful inpt. The second line of eq. (3) defines the *likelihood function* $l(\theta|y) = c\, p(y|\theta)$ (an arbitrary positive constant c times the joint sampling distribution of the data vector y, reinterpreted as a function of θ for fixed y); this is where all the information about θ internal to the data-set y is stored, and – under the logic of *Bayes's theorem* – the first line of eq. (3) defines $p(\theta)$ as the place where all the

information about θ *external* to the data-set *y* is stored. It has become traditional to call this $p(\theta)$ a *prior distribution* for θ; this terminology is unfortunate (it sounds as though only information gathered before the data-set *y* arrives can go into $p(\theta)$, and this is not true), but it has been used for so long that it's unlikely it can be changed now. Equation (3) implies that (*a*) learning about θ on the basis of *y* can occur via Bayes's theorem: the *posterior distribution* $p(\theta|y)$, which combines the information about θ contained in the prior and the likelihood, is just a renormalized version of their product: $p(\theta|y) = c\, p(\theta)\, l(\theta|y)$, with *c* chosen so that $p(\theta|y)$ integrates to (1), and (*b*) predictive distributions for future data given past data may also readily be calculated (for example, for $1 < m < n$ the predictive for $(y_{m+1}, \ldots, y_n)$ based on $(y_1,\ldots, y_m)$ is

$$p(y_{m+1}, \ldots, y_n|y_1, \ldots, y_m) = \int_0^1 \theta^{s_n - s_m}(1-\theta)^{n-m-(s_n-s_m)}\, p(\theta|y_1, \ldots, y_m)\, d\theta, \tag{4}$$

in which $s_m = \sum_{i=1}^{m} y_i$ and $p(\theta|y_1, \ldots, y_m)$ is the posterior distribution for θ based only on the first *m* observations.

Fifth, exchangeability evidently plays a role in Bayesian modelling that's somewhat analogous to the role of IID sampling in the frequentist approach, but exchangeability and IID are not the same: IID random variables are exchangeable, and exchangeable random variables are identically distributed, but they're not independent (for example, if You're about to observe a binary process whose tendency to yield a 1 is not known to You, and You judge Your uncertainty about future outcomes to be exchangeable, the information in the first n outcomes would definitely help You to predict outcome $(n+1)$; it's only when (somehow) the knowledge of the ‘underlying’ θ becomes available to You that there's no information in any of the outcomes to help predict any other outcomes – this situation might be summarized by saying that the past and the future become conditionally independent given the truth). Exchangeable observables are thus not IID, but they may often be usefully regarded as *conditionally* IID given a parameter vector θ, as in eq. (3) above.

Sixth, some awkwardness arose above in the frequentist approach to modelling the medical data, because it was not clear what population $\mathcal{P}$ the data could be regarded as *like* a random sample from. This awkwardness also arises in Bayesian modelling: even though in practice You are only going to observe $(y_1,\ldots,y_n)$, de Finetti's representation theorem requires You to extend Your judgement of finite exchangeability to the countably-infinite collective $(y_1, y_2,\ldots)$, *and this is precisely like viewing* $(y_1,\ldots,y_n)$ *as a random sample from* $\mathcal{P} = (y_1, y_2, \ldots)$. (Finite versions of de Finetti's representation theorem are available – for example, Diaconis and Freedman, 1980 – which informally say that, if You're willing to extend Your judgement of exchangeability from $(y_1,\ldots,y_n)$ to $(y_1,\ldots,y_N)$ for $N > n$, the larger *N* is the harder it becomes for Your predictive distribution $p(y_1,\ldots,y_n)$ to differ by a large amount from something representable by eq. (2) without violating the basic rules of probability.) The key point is that the difficulty arising from lack of clarity about the scope of valid generalizability from a given set of

observational data is a fundamental scientific problem that emerges whenever purely observational data are viewed through an inferential or predictive lens, whether the statistical methods You use are frequentist or Bayesian.

The entire discussion so far has been in the context of binary outcomes y with no covariates; in practice, predictor variables x are also generally available, and extensions to non-binary data are evidently needed as well. An example involving a covariate arose in the discussion of the mind projection fallacy above: if, in the medical setting considered here, the gender x of the patients is available, and if this has a clinically meaningful bearing on mortality from heart attack, then it would be more scientifically appropriate to assert exchangeability separately and in parallel within the two gender groups {male, female}. With this in mind de Finetti (1938) defined the concept of *partial exchangeability*, which is also known as *conditional exchangeability* (Lindley and Novick, 1981; Draper et al., 1993); with this newer terminology You would say that Your uncertainty about the mortality observables for these patients is conditionally exchangeable given gender. Conditional exchangeability is related to the notion, introduced by Fisher (1956), of *recognizable subpopulations*.

Suppose now that the observable You will measure on the patients in the medical example is a severity of illness score y_i, scaled as a continuous quantity from $-\infty$ to ∞, and return temporarily to the situation with no covariate information. As before Your uncertainty about the future y_i values is (unconditionally) exchangeable, but now a representation theorem is needed for continuous real-valued outcomes; de Finetti (1937) supplied this as well.

Theorem 2: **(representation of exchangeable predictive distributions for continuous observables [de Finetti, 1937])**: If You're willing to regard $(y_1,\ldots,y_n)$ as the first n terms in an infinitely exchangeable sequence $(y_1, y_2,\ldots)$ of continuous values on $\mathbb{R}$, then

- $F(t) = \lim_{n\to\infty} F_n(t)$ must exist for all t and must be a valid CDF, where F_n is the *empirical CDF* based on $(y_1,\ldots, y_n)$ (i.e., $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le t)$, in which $I(A)$ is the indicator function (1 if A is true, otherwise 0)), and the marginal distribution (given F) for each of the y_i must be $(y_i|F) \sim F$;
- $G(F) = \lim_{n\to\infty} P(F_n)$ must also exist, where P is Your joint probability distribution on $(y_1, y_2,\ldots)$; and
- Your predictive distribution for the first n observations can be expressed as

$$p(y_1, \ldots, y_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(y_i)\, dG(F), \tag{5}$$

where $\mathcal{F}$ is the space of all possible CDFs on $\mathbb{R}$.

Equation (5) says informally that exchangeability of Your uncertainty about an observable processing

unfolding on the real line is functionally equivalent to assuming the Bayesian hierarchical model

$$\begin{aligned} F &\sim p(F) \\ (y_i|F) &\overset{\text{IID}}{\sim} F, \end{aligned} \tag{6}$$

where $p(F)$ is a prior distribution on $\mathcal{F}$.

With binary observables, Theorem 1 (which is evidently a special case of Theorem 2) focuses attention on θ, the underlying rate of 1s in the population $\mathcal{P} = (y_1, y_2, \ldots)$ from which You're in effect regarding $(y_1,\ldots,y_n)$ as like a random sample; in the continuous case the analogous theorem focuses attention on F, the underlying CDF defined by $\mathcal{P}$. This makes Theorem 2 harder to implement in practice, because it's one thing to specify a prior distribution on a quantity $\theta \in (0, 1)$ and quite another to put a scientifically relevant prior distribution on the space $\mathcal{F}$ of all possible CDFs on the real line. Placing probability distributions on functions is the topic addressed by the field of *Bayesian nonparametric methods* (see, for example, Dey, Müller and Sinha, 1998), an area of statistics that has recently moved completely into the realm of day-to-day implementation and relevance through advances (since the early 1990s) in *Markov chain Monte Carlo* (MCMC) simulation-based methods of computation (see, for example, Gilks, Richardson and Spiegelhalter, 1995). Two rich families of prior distributions on CDFs about which a wealth of practical experience has recently accumulated include (mixtures of) *Dirichlet process priors* (see, for example, Ferguson, 1973) and *Pólya trees* (see, for example, Lavine, 1992).

As an example of the use of de Finetti's representation theorem for continuous outcomes, consider a randomized controlled trial or observational study with a treatment (T) and a control (C) group in which the outcome of interest y is modelled continuously on $\mathbb{R}$. A judgement of unconditional exchangeability of Your uncertainty about the y values for all subjects in the study would be equivalent to assuming that the T and C conditions had the same effect on the subjects, which (since the point of the study is presumably to see if this is true) would not be a good starting point; instead, in the absence of any other covariate information, it would be reasonable for You to model Your uncertainty about the y values as conditionally exchangeable given the indicator variable x that identifies which group each subject is in. With F_C and F_T as the underlying control and treatment CDFs and y_i^T as the observable for subject i in the treatment group (and similarly for y_j^C), a straightforward extension of Theorem 2 then leads to the following Bayesian nonparametric model for the observables (for $i = 1, \ldots, n_T$ and $j = 1, \ldots, n_C$):

$$\begin{aligned} (F_C, F_T) &\sim p(F_C, F_T) \\ (y_i^T|F_C, F_T) &\overset{\text{IID}}{\sim} F_T \quad \text{and} \quad (y_j^C|F_C, F_T) \overset{\text{IID}}{\sim} F_C. \end{aligned} \tag{7}$$

A nonparametric joint prior can then be placed on (F_C, F_T) using either of the Dirichlet process prior or Pólya tree methodologies mentioned above, and an appropriate functional of (F_C, F_T) (such as the difference or ratio of the underlying treatment and control means) can be monitored in the MCMC simulation. Note that this model arose solely from exchangeability considerations and (a simple extension of) Theorem 2.

Model specification has been a vexing topic in both frequentist and Bayesian statistics throughout much of the last century. Referring both to the conditional exchangeability judgements and to choices made in specifying the prior on F_C and F_T in the example above as *structural assumptions*, a popular approach to model specification (practised with equal vigour by both frequentists and Bayesians since the work of Tukey, 1962, and others on exploratory data analysis) involves (*a*) enlisting the aid of the data to conduct a search among possible structural assumptions, (*b*) choosing a single favourite structural specification S^*, and (*c*) pretending You knew all along that S^* was ‘correct’, even though it was arrived at via a data-driven search. From a Bayesian perspective this approach is clearly unsound, since it amounts to using the data to specify the prior distribution on the space $\mathcal{S}$ of all possible structural assumptions and then using the same data again to update the prior on $\mathcal{S}$; the result will often be inferences and predictions that are not well *calibrated*, with interval estimates that are not as wide as they need to be to fully acknowledge *model uncertainty*. *Bayesian model averaging* (Leamer, 1978; Draper, 1995), in which predictive distributions $p(y_f|y)$ for future observables y_f given past data y are computed by averaging over the model uncertainty uncovered by the search through $\mathcal{S}$ (rather than ignoring it), through calculations of the form

$$p(y_f|y) = \int_{\mathcal{S}} p(y_f|y, S)\, p(S|y)\, dS,$$

can provide one principled, satisfying and rather general method for solving the problem of Bayesian model specification in a well-calibrated manner; methods based on *cross-validation* (see, for example, Stone, 1974), in which (in effect) part of the data is used to specify the prior on $\mathcal{S}$ and the rest of the data is employed to update that prior, can provide another; and the approach illustrated above in the study with treatment and control groups – which combines conditional exchangeability judgments (driven by the context of the problem) with Bayesian nonparametric methods – can provide yet another.

See Also

- calibration
- model averaging
- model uncertainty

Bibliography

Bernardo, J.M. and Smith, A.F.M. 1994. *Bayesian Theory*. New York: Wiley.

de Finetti, B. 1930. Funzione caratteristica di un fenomeno aleatorio. *Memorie della Reale Accademia dei Lincei* 4, 86–133.

de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut H. Poincaré* 7, 1–68. Reprinted in translation as: Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability*, ed. H.E. Kyburg and H.E. Smokler. New York: Dover, 1980.

de Finetti, B. 1938. Sur la condition d'équivalence partielle. *Actualités Scientifiques et Industrielles* 739. Reprinted in translation as: On the condition of partial exchangeability. In *Studies in Inductive Logic and Probability*, ed. R.C. Jeffrey. Berkeley: University of California Press, 1980.

de Finetti, B. 1970. *Teoria delle Probabilità*, 2 vols. Turin: Eunaudi. Reprinted in translation as de Finetti B. 1974–75. *Theory of Probability*, 2 vols. Chichester: Wiley, 1974–75.

Dey, D.D., Müller, P. and Sinha, D., eds. 1998. *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer.

Diaconis, P. and Freedman, D. 1980. Finite exchangeable sequences. *Annals of Probability* 8, 745–64.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, Series B 57, 45–97.

Draper, D. 2007. Bayesian multilevel analysis and MCMC. In *Handbook of Multilevel Analysis*, ed. J. de Leeuw and E. Meijer. New York: Springer.

Draper, D., Hodges, J., Mallows, C. and Pregibon, D. 1993. Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society*, Series A 156, 9–37.

Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1, 209–30.

Fisher, R.A. 1956. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fréchet, M. 1943. *Les Probabilités Associées à un Système d'Événements Compatibles et Dépendents*, vol. 2. Paris: Hermann et Cie.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., eds. 1995. *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall/CRC.

Good, I.J. 1950. *Probability and the Weighing of Evidence*. London: Charles Griffin.

Good, I.J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Haag, J. 1924. Sur un problème général de probabilités et ses diverses applications. *Proceedings of the International Congress of Mathematics (Toronto)* 2, 659–74.

Hewitt, E. and Savage, L.J. 1955. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society* 80, 470–501.

Jaynes, E.T. 2003. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

Johnson, W.E. 1924. *Logic. III. The Logical Foundations of Science*. Cambridge: Cambridge University Press.

Khintchine, A.I. 1932. Sur les classes d'événements équivalents. *Mathematics of the USSR – Sbornik* 39, 40–3.

Lavine, M. 1992. Some aspects of Pólya tree distributions for statistical modeling. *Annals of Statistics* 20, 1222–35.

Leamer, E.E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Lindley, D.V. and Novick, M.R. 1981. The role of exchangeability in inference. *Annals of Statistics* 9, 45–58.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Stone, M. 1974. Cross-validation choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, Series B 36, 111–47.

Tukey, J.W. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33, 1–67.

How to cite this article

Draper, David. "exchangeability." The New Palgrave Dictionary of Economics. Second Edition. Eds.

Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000278> doi:10.1057/9780230226203.0520

The New Palgrave Dictionary of Economics Online

excise taxes

James R. Hines Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Excise taxes are selective taxes on the sale or use of specific goods and services, such as alcohol and petrol. Over time, governments have relied less on excise taxes, though, as of 2007, excise taxes still contribute 12 per cent of total government revenues in OECD countries. In addition to generating needed revenue, excise taxes can control externalities and impose tax burdens on those who benefit from government spending. Rather more controversially, they also can be used to discourage consumption of potentially harmful substances (such as tobacco and alcohol) that individuals might over-consume in the absence of taxation.

Keywords

excise tax incidence; excise taxes; externalities; leisure; optimal taxation; Pigouvian taxes; Ramsey model

Article

Excise taxes are selective taxes on the sale or use of specific goods and services, such as alcohol and petrol.
Excise taxes have existed for centuries and are widely used by governments today, at the beginning of the 21st century. The 20th-century spread of income taxation and value-added taxation reduced the significance of excise taxation as a source of government revenue, but most governments still collect sizeable taxes on petroleum products, tobacco products and alcohol. For example, in 2004 the US federal government collected 72 billion dollars in excise taxes, representing four per cent of its total tax revenues, of which petroleum taxes accounted for 33 billion dollars, or 45 per cent of total excises. The United States relies the least on excise taxes of the 30 wealthy nations that are members of the Organisation for Economic Co-operation and Development (OECD), whose excise tax collections in 2000 averaged 12 per cent of total government revenues (Hines, 2007). As recently as 1969–71, excise

taxes contributed 23 per cent of total tax collections of high-income countries, and 27 per cent of tax collections of developing countries (Cnossen, 1977). Among OECD countries in 2000, those with the lowest incomes, the most centralized governments and the greatest openness to international trade had the highest ratios of excise tax collections to total government revenues (Hines, 2007).

Excise taxes take the form either of specific taxes or of ad valorem taxes. A specific tax (or unit tax) is defined per unit of the taxed good or service, whereas an ad valorem tax is defined per sales value. Thus, the current US federal taxes of 0.184 dollar per gallon of petrol and 0.39 dollar per pack of cigarettes are specific taxes, while the US tax of 11 per cent of the value of bows and arrows and three per cent of the value of fish-finding sonar devices are ad valorem taxes. In competitive markets specific and ad valorem taxes have identical consequences, other than any differences stemming from compliance and enforcement. In imperfectly competitive markets the difference is more consequential, since ad valorem taxes automatically impose higher per-unit tax rates as firms restrict quantities to drive up prices. Hence, in imperfectly competitive markets, ad valorem taxes produce lower consumer prices and lower deadweight loss than do specific taxes raising the same revenue (Suits and Musgrave, 1953; Delipalla and Keen, 1992).

Four motivations underlie the use of most excise taxes. The first is revenue generation: excise taxes can produce significant government revenues, and may do so at lower political or economic cost than alternatives such as income taxation. The second motivation is the application of the benefit principle of taxation: excise taxes can be tailored to impose tax burdens on those who benefit from government services financed by excise taxes. Petrol fuel taxes are often justified as user fees for government-provided roads, and the tax on sonar devices is justified by government expenditures to maintain lakes and fisheries. The third motivation is control of externalities, which is the goal of a number of excise taxes on polluting substances, such as taxes on ozone-depleting chemicals. And the fourth motivation is that excise taxes may discourage consumption of potentially harmful substances (such as alcohol and tobacco) that individuals might over-consume in the absence of taxation.

Excise tax incidence

It is customary to think of the burden of excise taxes as being borne by consumers of taxed goods and services in the form of higher after-tax prices, but there is considerable scope for the shifting of tax burdens. In a simple competitive partial equilibrium setting, the burden of an excise tax depends on elasticities of demand and supply: if demand for a taxed good or service is elastic, and supply relatively inelastic, then the burden of an excise tax is borne by sellers, whereas buyers bear the burden of a tax on a good or service with inelastic demand and elastic supply. Similar demand and supply elasticities imply equal sharing of excise tax burdens between buyers and sellers.

There are plausible circumstances in which consumers can bear more than 100 per cent of the burden of an excise tax, or alternatively, might actually benefit from the introduction of a tax. If the market for a good or service is imperfectly competitive, then, depending on the nature of competition and cost conditions, consumers may well bear more than 100 per cent of the burden of an excise tax (Delipalla and Keen, 1992). Even with perfect competition, the nature of demand and supply spillovers among multiple markets in general equilibrium can produce anomalous outcomes, such as consumer prices that rise by more than the amount of excise taxes, or (after-tax) prices that even fall with the introduction of

excise taxes (Hotelling, 1932).

There is mixed evidence of excise tax incidence in practice. Poterba (1996) offers evidence that consumers bear the full burden of US excise and sales taxes in the form of higher prices, but Besley and Rosen (1999) find that many consumer prices rise by more than 100 per cent of excise and sales taxes imposed by US states and localities.

One concern frequently expressed about excise taxation is the potential regressivity of the resulting tax burdens. Excise tax rates do not rise with consumption in the way that income tax rates rise with income; furthermore, since the poor spend higher fractions of their incomes than do the wealthy, taxes based on expenditure rather than income may put greater relative burdens on low-income individuals. This second consideration is considerably diminished by adopting a lifetime perspective, however, since individuals ultimately either spend or give away all of their incomes. Hence the distributional impact of excise taxes depends critically on the income elasticities of demand for goods and services subject to high rates of excise taxation. Poterba (1991) analyses US petrol taxes from the standpoint of lifetime incidence, finding that petrol consumption rises more than proportionately with affluence over much of the range of total spending, suggesting that petrol taxes are progressive, albeit less so than income taxes.

Optimal excise taxation

Ramsey (1927) initiated the modern theory of optimal taxation with his analysis of excise taxation in a model with identical consumers, finding that, far from being uniform, optimal excise tax rates vary inversely with elasticities of demand for taxed goods. Ramsey's set-up restricts the government to raising a given amount of revenue exclusively with excise taxes, and the resulting optimal tax pattern reflects that the excess burden of a tax increases with its behavioural impact. Diamond (1975) generalized the Ramsey rule to settings with heterogeneous individuals, showing that the resulting modified optimal excise taxes reflect both efficiency (lower tax rates on elastically demanded goods) and distributional (higher tax rates on goods purchased by wealthy individuals) considerations. As noted by Corlett and Hague (1953–4), the government's inability to tax leisure is what prevents uniform excise taxes from being optimal in the Ramsey model; as a second-best correction, the optimal Ramsey tax configuration entails imposing heavier excise taxes on goods and services that are complementary with untaxed leisure.

Under what circumstances would a government with access to a full range of income tax instruments want to impose excise taxes at differentiated rates? Atkinson and Stiglitz (1976) showed that, if consumers have identical utility functions that are weakly separable in consumption and leisure, then there is nothing to be gained by supplementing an optimal nonlinear income tax with differentiated excise taxes. The reason is that, in such a setting, patterns of commodity consumption fail to convey information to the government that is not already captured by income levels.

Excise taxes can nevertheless serve the function of controlling externalities, a consideration omitted from the Atkinson–Stiglitz framework. Pigou (1920) famously proposed the imposition of corrective excise taxes at rates equal to marginal external damages, noting that doing so restores economic efficiency, and Sandmo (1975) illustrates the optimal application of Pigouvian excise taxes when the government relies on excise taxes to raise revenue. In practice, governments impose heavy taxes on energy products, motor vehicles and other transport, waste management, ozone-depleting substances,

and other products and activities that arguably create externalities in degrading the environment. In 2000, OECD countries raised an average of 5.5 per cent of their total tax revenues from these environmental taxes, with European Union members averaging 6.8 per cent, and the United States the lowest in the OECD at 3.4 per cent (Hines, 2007).

Excise taxes can also play a role in discouraging consumption of goods that may not have external effects, but are nonetheless harmful to the individuals who consume them. Examples of such goods include tobacco products, alcohol, and food with poor nutritional content. Irrational consumers may begin consuming these items without fully appreciating the regret they will experience years later, in which case there could be a role for optimal excise taxation to help consumers by making consumption more expensive, and therefore reducing the likelihood of consumers starting early on the path of over-consumption (O'Donoghue and Rabin, 2006).

Finally, excise taxes raise enforcement concerns, as do all taxes. In the United Kingdom, which boasts the highest cigarette taxes in Europe, one cigarette out of every five is purchased on the black market (Cnossen and Smart, 2005). Hence the choice of excise tax policy needs to be sensitive to smuggling and other evasion opportunities.

See Also

- consumption taxation
- excess burden of taxation
- optimal taxation
- Pigouvian taxes
- progressive and regressive taxation
- tax compliance and tax evasion
- tax incidence
- value-added tax

Bibliography

Atkinson, A.B. and Stiglitz, J.E. 1976. The design of tax structure: direct versus indirect taxation. *Journal of Public Economics* 6, 55–75.

Besley, T.J. and Rosen, H.S. 1999. Sales taxes and prices: an empirical analysis. *National Tax Journal* 52, 157–78.

Cnossen, S. 1977. *Excise Systems*. Baltimore: Johns Hopkins University Press.

Cnossen, S. and Smart, M. 2005. Taxation of tobacco. In *Theory and Practice of Excise Taxation*, ed. S. Cnossen. Oxford: Oxford University Press.

Corlett, W.J. and Hague, D.C. 1953–4. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21, 21–30.

Delipalla, S. and Keen, M. 1992. The comparison between ad valorem and specific taxation under imperfect competition. *Journal of Public Economics* 49, 351–68.

Diamond, P.A. 1975. A many-person Ramsey rule. *Journal of Public Economics* 4, 335–42.

Hines, J.R., Jr. 2007. Taxing consumption and other sins. *Journal of Economic Perspectives* 21(1), 49–68.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of supply and demand functions. *Journal of Political Economy* 40, 577–616.

O'Donoghue, T. and Rabin, M. 2006. Optimal sin taxes. *Journal of Public Economics* 90, 1825–49.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Poterba, J.M. 1991. Is the gasoline tax regressive? In *Tax Policy and the Economy*, vol. 5, ed. D.F. Bradford. Cambridge, MA: MIT Press.

Poterba, J.M. 1996. Retail price reactions to changes in state and local sales taxes. *National Tax Journal* 49, 165–76.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Sandmo, A. 1975. Optimal taxation in the presence of externalities. *Swedish Journal of Economics* 77, 86–98.

Suits, D.B. and Musgrave, R.A. 1953. Ad valorem and unit taxes compared. *Quarterly Journal of Economics* 67, 598–604.

How to cite this article

Hines, James R., Jr. "excise taxes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000162> doi:10.1057/9780230226203.0521

The New Palgrave Dictionary of Economics Online

exhaustible resources

Geoffrey Heal
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides a brief overview of the ideas that have emerged in economics in connection with exhaustible resources. A resource is exhaustible if, the more we consume today, the less will be available for consumption at later dates. The dynamics of resource allocation, and the attainment of efficiency in intertemporal models, are thus key aspects of any economic theory of exhaustibility. The exhaustibility paradigm is widely applicable, including to climate change, biodiversity loss and even such non-environmental phenomena as antibiotic resistance.

Keywords

biodiversity; calculus of variations; capital–resource ratio; certainty-equivalence theorem; CES production function; climate change; Cobb–Douglas function; control theory; depletion; discount rate; exhaustible resources; Hamiltonians; Hartwick rule; Hotelling model; Hotelling rule; imperfect competition; intertemporal substitutability; neoclassical growth model; resource rent; shadow prices; sustainability

Article

Exhaustible resources are among the most important inputs to economic activities. Conventional crude oil and natural gas are good topical examples, with their pricing and availability currently a major source of concern. Of course, all minerals and extractive resources are exhaustible, as the volume of the earth is finite, but for many resources exhaustibility is not a matter of everyday concern as it is for oil and gas. There is a concern that the exhaustibility of low-cost deposits of oil and gas could restrict the long-run growth potential of the industrial world, and understanding the economic implications of exhaustibility is essential to grappling with this issue.

The availability of natural resources is a substantial topic in geology, with an extensive discussion there of the type of information available about resource supplies (see Barnett and Morse, 1963; Brobst, 1979;

Smith, 1980; Goeller, 1979). Geologists classify resource stocks as proven, probable or possible reserves, which differ in the costs of extraction and the certainty with which their scale is known. An important lesson from the geological perspective is that the size of the total reserves of most resources is unknown. Even today, after many years of intensive prospecting, that remains true for a resource as important as oil.

The paradigm of exhaustibility is not limited in its applications to mineral resources: underground aquifers are exhaustible, and the capacity of the atmosphere to absorb greenhouse gases without radical change to the climate system has also been modelled as an exhaustible resource (Heal, 1984). Another related and interesting exhaustible resource is the capacity to store carbon dioxide in underground rock formations: according to some perspectives on climate change, this could be a vitally important – and exhaustible – resource up to about 2050 (Butt et al., 1999). A very different recent application of this paradigm is in the field of drug resistance: Laxminarayan and Brown (2001) have modelled as an exhaustible resource the extent to which a drug can be used before resistance develops among the pathogens it is intended to kill. The world's stock of biodiversity can be seen as an exhaustible resource, too. Every time a species is driven to extinction, this stock falls in an irreversible way. We are depleting that stock, and do not fully understand the consequences.

The first analytical discussion of exhaustibility can be found in the work of L.C. Gray in 1914, but the work that is regarded as the classic work in this area is that of Harold Hotelling in 1931. In an extraordinarily prescient article he provided the foundations of the modern theory of exhaustible resources, while simultaneously giving one of the earliest applications of calculus of variations in economics and developing an arbitrage-free model of equilibrium pricing. His work was so much ahead of its time that it was 30 years or more before it was fully appreciated. His name is now widely attached to the basic model of resource depletion, the ‘Hotelling model’, and to the movement of resource prices in a competitive market, which follow the ‘Hotelling rule’.

A central feature of exhaustible resources is that they force us to think about resource allocation over time. For an exhaustible resource, the more we consume today, the less will be available for consumption at later dates. The dynamics of resource allocation, and the attainment of efficiency in intertemporal models, are key aspects of any economic theory of exhaustibility. This is probably the main reason why Hotelling's contribution was neglected for so long: in the 1930s economists were not ready to consider these issues analytically. It took the development of theories of growth, descriptive and optimal, to bring these to the fore.

The Hotelling model

The simplest economic model of the use of an exhaustible resource is as follows. There is an initial stock S_0 of the resource, and we denote by $S_t \leq S_0$ the stock remaining at time t. Consumption of the resource at date t is c_t and the benefits from consumption are represented by a utility function $u(c_t)$, generally assumed to be strictly concave, twice differentiable and to satisfy a boundary condition such as $du/dc \to \infty$ as $c \to 0$ or $\lim_{c \to 0} u(c) = -\infty$. Either of these conditions penalizes zero consumption very sharply and keeps consumption away from zero. Within this model we seek the time path of resource consumption that maximizes the present value of welfare:

$$Max \int_0^\infty u(c_t)\, e^{-\delta t} dt \text{ subject to} \int_0^\infty c_t dt \leq S_0$$
(1)

Here $\delta \geq 0$ is a discount rate, and the second integral reflects the constraint imposed on cumulative use of the resource by its exhaustibility. It is this intertemporal conflict between present and future resource use that lies at the centre of theories of exhaustibility. Problem 1 is a classical problem in the calculus of variations, an isoperimetric problem, so called because it was first solved in the context of finding the closed plane curve having a given length and enclosing the greatest area. Economists have reformulated this problem so that it can be solved by methods from control theory rather than calculus of variations, and the reformulation usually used is

$$Max \int_0^\infty u(c_t)\, e^{-\delta t} dt \text{ subject to} \frac{dS_t}{dt} = -c_t,\ S_t \geq 0$$
(2)

This replaces an integral constraint with a differential equation and a non-negativity constraint on a state variable. Solving this problem requires use of the Hamiltonian

$$H = u(c_t)\, e^{-\delta t} + \lambda_t e^{-\delta t}[-c_t]$$

where λ is a current value co-state variable. The first order conditions that a solution must satisfy are

$$u(c_t)' = \lambda_t \text{ and } \frac{1}{\lambda_t}\frac{d\lambda_t}{dt} = \delta$$
(3)

where a prime denotes the derivative of a function with respect to its argument, and two primes denote the second derivative. The first condition is intuitively obvious: the marginal utility of consumption must equal the shadow price of the resource. The second condition, which we call the Hotelling rule, states that the shadow price must rise at the discount rate. Another way of looking at this is that in present value terms the shadow prices must be constant, so the present value of the marginal utility of

consumption must be constant. Again, this is an intuitively satisfactory condition. It implies, of course, that consumption must fall over time.

To give an illustration of this result, consider the logarithmic utility function $u(c_t) = \ln c_t$. From 3 $\lambda_t = u(c_t)' = \frac{1}{c_t} = \lambda_0 e^{\delta t}$ so $c_t = c_0 e^{-\delta t}$ and $c_0 = \delta S_0$.

The Hotelling problem has also been known as the ‘cake-eating problem’ as it can be seen as choosing how to divide a finite cake between different periods. As long as the discount rate is positive there is a well-defined answer, as we have seen above. Matters are different if we try to treat all generations equally and set $\delta = 0$. Then, as we can see from the logarithmic example above, consumption is zero: indeed, it clearly goes to zero at all dates as the discount rate goes to zero. On an optimal path, as the discount rate falls we are trying to spread consumption ever more thinly over time and keep more for the future. This makes sense, except in the limiting case with a zero discount rate, in which it tells us never to consume anything and always to keep everything for the future. In this case problems 1 and 2 have no solution: although they look reasonable they are in fact ill-posed. There is an extensive literature on what exactly goes wrong in this case and what alternative formulations are available (Heal, 1985; Gale, 1967).

Extensions of the basic model

A simple change in the basic model is to allow for extraction costs, something that Hotelling did back in 1931. Consider a corporation extracting resources for a profit, rather than a national planning problem. The rate of extraction is again c_t and there is a cost of extraction of x per unit. The sale price of the resource is given by the demand curve $p(c_t)$, so the company is not a price-taker. The present value of profits is

$$\int_0^\infty [p(c_t)c_t - xc_t]e^{-\delta t}dt$$

and the corporation seeks to maximize this subject to the same constraints as above. The result is that

$$[p(c_t)'c_t + p - x] = \lambda_t \text{ and } \frac{1}{\lambda_t}\frac{d\lambda_t}{dt} = \delta$$

The first term is just the marginal revenue minus marginal cost: this difference again has to grow at the discount rate. We summarize this by saying that the rent to the resource must grow at the discount rate. Another simple extension is to place a value on the remaining stock of the resource, $u(c_t, S_t)$. Think here of biodiversity in general or of a slow-growing forest. We value the flow of services that comes from depleting these, but we also value having some of the stock remaining (Krautkramer, 1985; Heal,

1998). Looking at the problem

$$Max\int_0^{\infty} u(c_t, S_t)\, e^{-\delta t} dt \text{ subject to} \frac{dS_t}{dt} = -c_t,\ S_t \geq 0 \tag{4}$$

leads to very different results. The first order conditions for optimality are that

$$u_c = \lambda \text{ and } \frac{d\lambda}{dt} - \delta\lambda = -u_S$$

where u_c, u_S are respectively the derivatives of u with respect to c_t and S_t. This solution is analysed at length in Heal (1998): what is interesting is that, provided we drop the boundary condition on $u(0)$ or $u'(0)$, there may be a stationary solution to this system of equations, at which $\frac{d\lambda}{dt} = 0$, $c_t = 0$ and $\delta = u_c / u_S$. This means that the discount rate equals the slope of an indifference curve in the $c - S$ plane, that is, it equals the marginal rate of substitution between the contributions to welfare of the stock and the flow.

Resources in production

The natural next step in considering how exhaustible resources affect an economic system is to model how they enter into the production process and whether their exhaustibility acts as a drag on the economy, limiting its long-run growth. This takes into the discussion of ‘sustainability’ and exhaustibility. The simplest way to do this is to take the basic neoclassical growth model and replace labour as an input by an exhaustible resource. This gives us as a production function $F(K_t, R_t)$ where K_t and R_t are respectively the capital stock and rate of resource use at date t. Typically F is assumed to be twice differentiable and to show constant or diminishing returns to the two inputs. With this formulation we can consider the problem

$$Max\int_0^{\infty} (c_t)\, e^{-\delta t} dt,\ c = F(K_t, R_t) - I_t, \frac{dK_t}{dt} = I_t,\ \frac{dS_t}{dt} = -R_t,\ S_t \geq 0 \tag{5}$$

Here, obviously, I_t is investment at t, the rate of change of the capital stock. This problem is considerably more complex than any of the previous ones, and is analysed at length by Dasgupta and Heal (1974). Let $y_t = \frac{K_t}{R_t}$, $f(x_t) = F\left(\frac{K_t}{R_t}, 1\right)$, $\eta(c) = \frac{-cu''(c)}{u'(c)}$ and let σ be the elasticity of substitution between capital and the resource,

$$\sigma = \frac{-f'(x)f(x) - xf'(x)}{xf(x)f''(x)}$$

Then we can state the conditions characterizing a solution to eq. (5) as follows (θ is the price of the produced good):

$$\frac{1}{\theta}\frac{d\theta}{dt} = F_K - \delta, \quad \frac{1}{c}\frac{dc}{dt} = [F_K - \delta] / \eta(c) \tag{6}$$

$$\frac{1}{x}\frac{dx}{dt} = \sigma\frac{f(x)}{x} \text{ and } \lambda_t F_R e^{-\delta t} \text{ is constant} \tag{7}$$

Clearly in an economy with capital that can be reproduced and accumulated, and a resource that is exhaustible, growth must take place through the substitution of capital for resources. Equation (7) captures this process. It tells us that the capital–resource ratio changes at a rate that is the product of the elasticity of substitution and the average product per unit of fixed capital. The former indicates the ease with which substitution can be carried out, and the latter can be thought of as a measure of the importance of capital in production. So the easier it is to substitute, and the more important capital is, the more we substitute capital for resources.

What impact does the resource have on long-run growth possibilities? Clearly if $F(K, 0) > 0$ for $K > 0$ then the exhaustibility of the resource does not matter: we can continue producing when it runs out. The interesting case is $F(K, 0) = 0$ for any K. The fact that no production is possible without the resource does not mean that production must go to zero: consumption of the resource, as we saw in the Hotelling model, can be spread thinly over the indefinite future. If we can produce enough to support a good living

standard with only a very small amount of the resource and a lot of capital, again exhaustibility may not matter. And, of course, technical progress may come to the rescue: it could increase the productivity of the scarce resource, or release the constraint that it imposes on production. In the light of these considerations Dasgupta and Heal (1979, p. 198) advance the following definitions: ‘We shall regard an exhaustible resource as being *inessential* if there is a feasible program along which consumption is bounded away from zero: or in other words, if a positive sustainable level of consumption is feasible. Likewise, regard a resource as *essential* if feasible consumption must necessarily decline to zero in the long run.’ This discussion anticipates many more recent discussions of sustainability. We can explore these issues further in the context of a CES production function, and in this case Dasgupta and Heal show that, if the elasticity of substitution between the resource and capital exceeds one, then the exhaustibility of the resource does not pose a fundamental problem. In the opposite case, an elasticity less than one, output must eventually fall to zero in the absence of technical progress. And in the borderline case, a unit elasticity, we have the Cobb–Douglas production function. In this case, if the elasticity of output with respect to capital exceeds that with respect to the resource, there is a feasible policy on which consumption is bounded away from zero. But in the remaining cases there is not: absent technical progress, output must fall to zero. In fact, if we have a Cobb–Douglas production function with the elasticity of output with respect to capital greater than that with respect to the resource, not only are there paths on which output is bounded away from zero, but there are paths on which output is constant and on which output grows continuously. The paths on which output is constant may be characterized by a constant level of investment that maintains the total value of all stock constant, with the investment equal to the rent on the exhaustible resource. This is the Hartwick rule (Hartwick, 1977; Asheim, Buchholtz and Withagen, 2003).

Backstop technologies

Clearly the issue of substitutability is central to an understanding of the economic consequences of exhaustibility of important inputs. There are several dimensions to substitutability. One dimension, the one that we have discussed so far, is substitutability between capital and the resource. We can reduce oil use by insulating our buildings and wearing warmer clothes: this is substituting capital for oil. So is buying more expensive but more efficient furnaces. Both reduce oil consumption but require more capital. But there is another aspect of substitutability. Conventional crude oil is exhaustible and will be fully depleted at some point. But then there will be alternatives. For example, we can extract oil from coal – this is what Germany did during the Second World War and what South Africa did while it was subject to a trade embargo because of its apartheid policies. It is expensive by the standards of historical oil prices – perhaps $40 per barrel – but completely feasible. Similarly, oil can be extracted from tar sands – indeed, it is currently so extracted – but again at a cost that is high by historical oil price standards. And there are vast reserves of tar sands – they can probably provide more oil than all the conventional crude oil deposits in the Middle East. So other resources can replace oil when it runs out. This is a form of substitutability. Dasgupta and Heal (1974) modelled this by assuming that at a date T, which was unknown, the constraint imposed by the exhaustible resources would be lifted and an abundant substitute would become available. This is not unlike the situation described above with respect to oil and coal or tar sands. Another interpretation is that T is when nuclear fusion finally

becomes a reality and, in the much-quoted phrase of the 1960s, energy finally becomes ‘too cheap to meter’. (See also Nordhaus, 1973, for simulation models of the effect of a backstop technology.)

To formalize this, assume that prior to *T* the technology is as in the previous section, but that after *T* there is a new technology that does not depend on the resource as an input. We can think of this as a dramatic technical change (such as fusion) or the appearance of an abundant substitute for the resource (such as tar sands for oil). Production from *T* onwards depends only on the capital stock available at *T* and not on the resource stock at that date. So we can write a state valuation function $V(K_T)$ giving the present value of welfare along an optimal policy from *T* onwards, discounted back to *T*. Then the overall problem is to

$$Max\int_0^T u(c_t)e^{-\delta t}dt + V(K_T), \text{with for } t \in [0, T],\ c = F(K_t, R_t) - I_t, \frac{dK_t}{dt} = I_t, \frac{dS_t}{dt} = -R_t, S_t \geq 0$$

Dasgupta and Heal assumed that *T* was unknown with density function ω_t. Then the maximand is the expected value of $\int_0^T u(c_t)e^{-\delta t}dt + V(K_T)$, which is

$$\int_0^\infty w_t \int_0^T [u(c_t)e^{-\delta t}dt + V(K_T)]$$

which on integration by parts and letting $\Omega_t = \int_t^\infty \omega_T dT$ can be written as

$$\int_0^\infty e^{-\delta t}[u(c_t)\Omega_t + \omega_t V(K_t)]\,dt$$

Dasgupta and Heal (1974) explore possible solutions to this problem in some special cases, and characterize paths that are optimal. In the special case in which the valuation function *V* is independent not only of the resource stock S_T (which we have assumed) but also of the capital stock K_T (so that the new source makes all existing capital obsolete) the optimum path satisfies

$$\frac{1}{x}\frac{dx}{dt} = \sigma\frac{f(x)}{x} \tag{8}$$

$$\frac{1}{c}\frac{dc}{dt} = \frac{F_K - \delta - \Psi}{\eta(c)} \tag{9}$$

where $\Psi_t = \frac{\omega_t}{\Omega_t}$. These are very similar to eqs. (6) and (7) from the deterministic case with no backstop technology, except for the term Ψ that is added to the discount rate. Dasgupta and Heal also establish a certainty-equivalence theorem showing when the possibility of a backstop arriving can be subsumed entirely into a modification of the discount rate, by the addition of a term Ψ_t to the discount rate that reflects the conditional probability of the backstop arriving now, given that it has not yet arrived.

Of particular interest is the behaviour of the resource price when there is a possibility of a substitute arriving. A more direct way of understanding this is to look at a different model (Heal, 1976), again with a backstop technology available as a replacement for the resource. Assume that there is a cost to extraction of the resource, and that this depends on and increases with the amount extracted to date. This is in many ways a natural assumption, in keeping what we know about the grade-tonnage distribution for most minerals. There are small amounts available at low extraction costs, more at larger costs and almost unlimited amounts if we are prepared to pay a sufficiently high cost. There is also a backstop available at a fixed unit cost, and in effectively unlimited supply.

One way of formalizing this is to let $z_t = \int_0^t c_\tau d\tau$ be the amount extracted to date, and have the current unit extraction cost depend on this:

$$\text{extraction cost} = g(z),\ g' > 0 \text{ for } 0 \le z \le \bar{z} \text{ and } g(z) = \beta = g(\bar{z}) > 0 \text{ for } z \ge \bar{z}$$

So the unit extraction cost is an increasing function of cumulative extraction up to a certain total extraction and a corresponding extraction cost, at which point a backstop becomes available in more or less unlimited amounts at a constant cost of β. In the case of a fixed extraction cost, which is the classic Hotelling model considered above, the price rises away from the marginal extraction cost at the discount rate – the rent on the resource satisfied Hotelling's rule and increases exponentially. This is natural: the rent on the resource reflects its scarcity, and this scarcity is rising over time, so it is natural for the rent to rise too. In the present case we should expect a different outcome. Overall the resource is not scarce: high-cost sources, to which we move over time, are abundant. It is only low-cost sources that are scarce. During the depletion of these, scarcity reigns, but once they have gone there is no longer scarcity in the

sense of a limited availability. Any amount is available at the right price. So the dynamics of scarcity are in effect reversed. The solution reflects this, and is found by piecing together the solutions to two distinct problems:

$$Max \int_0^\infty u(c_t)\, e^{-\delta t} dt \text{ subject to } c_t + \frac{dK_t}{dt} = F(K_t, R_t) - g(z_t) R_t \tag{10}$$

and

$$Max \int_0^\infty u(c_t)\, e^{-\delta t} dt \text{ subject to } c_t + \frac{dK_t}{dt} = F(K_t, R_t) - \beta R_t \tag{11}$$

Problem 10 reflects the constraints on the economy up to the time when the backstop comes into use, and problem 11 represents the economy after all lower-cost reserves are depleted and once it is dependent on the backstop. We expect that a path that is optimal overall will first follow a solution to 10 and then one to 11, and this intuition can be verified formally (Heal, 1976). Clearly, once we are in the second regime, we expect that the price of the resource will equal its marginal extraction cost β. What is less clear is how price and extraction cost are related during the phase corresponding to the solution to 10. It is possible to show the following: if λ is the price of the resource and θ that of the produced good, then

$$\frac{1}{\lambda}\frac{d\lambda}{dt} = \delta\left(1 - \frac{\theta g(z)}{\lambda}\right) + \frac{\theta g(z)}{\lambda}\left(\frac{1}{\theta}\frac{d\theta}{dt}\right) \tag{12}$$

Although it looks complicated, eq. (12) in fact bears a simple and intuitive interpretation. It expresses the rate of change of the price of the resource as the weighted average of two terms, where the weights are $\left(1 - \frac{\theta g(z)}{\lambda}\right)$ and $\frac{\theta g(z)}{\lambda}$, which are respectively the fraction of the price that is pure rent and the fraction that is extraction cost. The two terms whose weighted average equals the rate of change of the resource price are the discount rate and the rate of change of the price of the produced good. So, if most of the price of the resource is rent, then its price rises at close to the discount rate, whereas if most of the

price is extraction costs, then it rises at the rate at which extraction costs rise. This suggests a path of price movements rather different from the classic Hotelling model: in the present case prices will contain a large rent element early in the life cycle of the resource and then no rent at all towards the end of the life cycle.

The relationship between price and extraction cost, and so the movement of the scarcity rent over time, has been the subject of many additional papers, including Solow and Wan (1976), Hanson (1980), Farzin (1992) and Oren and Powell (1985). Oren and Powell extend the basic model presented above to consider a class of related issues, and Farzin focuses on the issue of whether the movement of scarcity rent must be monotonic. He concludes that there are reasonable cases in which the scarcity rent moves non-monotonically over time.

Imperfect competition and resource use

So far we have considered patterns of resource use that are socially optimal, which are also the patterns of resource use that would emerge from a set of complete competitive markets with no elements of market failure present. While the answers to these questions are interesting and informative, it is clearly necessary to understand the impact of market imperfections on this picture. There is, not surprisingly, an extensive and sophisticated literature on this. I have space for only one or two basic insights.

We expect that moving from competition to monopoly will raise the prices of a good. To the extent that this is true for exhaustible resources, then monopoly will reduce the rate of extraction. Hence Robert Solow's comment that ‘The monopolist is the conservationist's best friend’. But, with a fixed stock to sell, as in the basic Hotelling model, if the monopolist sells less now because of a higher price, then he must sell more in the future, when the benefits are discounted. This is unattractive to him: the intertemporal substitutability inherent in any choices about a time pattern of resource use brings a new dimension to the impact of monopoly or imperfect competition on price and output.

An interesting illustration of this is seen clearly in the simplest possible case, that of a monopolistic supplier of a resource with a zero marginal extraction cost. In this case, whether the monopolist charges a higher or a lower price initially than the competitive price depends on the behaviour of the price elasticity of demand for the resource along the demand curve. Consider the family of constant-elasticity demand curves, $p(c) = Ac^{-B}$ where $1 > B > 0$. The monopolist's problem is to

$$Max \int_0^{\infty} p(c)\, c e^{-\delta t} dt$$

subject to the usual constraint on the total availability of the resource. This requires that the marginal revenue from sale of the resource grows over time at the discount rate. Now, letting $\varepsilon(c)$ be the demand elasticity when consumption is c we can write marginal revenue as

$$MR_t = p_t(1 + 1/\varepsilon(c_t)) = p_t\gamma_t \text{ where } \gamma = (1 + 1/\varepsilon(c_t))$$

Hence

$$\frac{d\ln MR}{dt} = \delta = \frac{d\ln p}{dt} + \frac{d\ln\gamma}{dt}$$

and so

$$\frac{d\ln p}{dt} = \delta - \frac{d\ln\gamma}{dt}$$

For a constant elasticity demand curve it is easy to check that $\frac{d\ln\gamma}{dt} = 0$ so that in this case a monopolist will want the price to rise at the discount rate – just as in the competitive case. More generally, we can show that

$$\text{If } \frac{d\eta}{dc} > (<) 0 \text{ then } \frac{d\ln p}{dt} < (>) \delta$$

so that the nature of the bias from first best introduced by monopoly depends on the way the demand elasticity changes along the demand curve (Dasgupta and Heal, 1979; Stiglitz, 1976).

The case of a monopoly supplier is the simplest entry point to imperfect competition, but does not do justice to the sophistication of the results that are available in this area. One of the most interesting developments was motivated by the role of the OPEC cartel in the oil market, and a desire to understand its real long-run impact. This is the development of models of a market with a cartel and a ‘competitive fringe’, which seems to describe accurately the relationship between OPEC and non-OPEC members. Closely associated with this is the idea of limit pricing to keep a backstop technology out of the market. Models incorporating these ideas are summarized in Dasgupta and Heal (1979), and some of the key original articles are by Sweeney (1977), Gilbert (1978) and Pindyck (1978).

Conclusions

Exhaustible resources are economically important. In addition, exhaustibility is an analytically interesting property: it forces us to think even in the very simplest case about intertemporal issues, about the present versus the future. Without this conflict there is no exhaustible resource. So dynamics are integral in even the most basic thinking here, which is one reason why serious discussion of exhaustibility was so slow to emerge. It is not surprising that exhaustibility has featured centrally in discussions of sustainability, and earlier in the neo-Malthusian diatribes of the Club of Rome. Many of the issues that have emerged in the debates about sustainability have in fact been analysed by economists in the discussions of exhaustibility in the 1970s (see Heal, 2003), usually with an emphasis on substitutability and technical progress as long-run solutions to the constraints imposed by exhaustibility, solutions that are typically more apparent to economists than to most others, though no less realistic for that. As I emphasized in the introduction, an interesting aspect of exhaustibility is the rather widespread applicability of the paradigm – to climate change, biodiversity loss and even such totally non-environmental phenomena as antibiotic resistance.

In this article I have been able to review only a fraction of a large and original literature on exhaustibility. I have certainly short-changed the literature on imperfect competition in markets for exhaustible resources, and have not touched at all on work on the empirical testing of the models of price movements discussed here (Heal and Barrow, 1980; Miller and Upton, 1985; Slade, 1982; Agbeyegbe, 1989). Another big gap is the theory of non-optimal growth with exhaustible resources, which asks how a market economy with imperfect futures markets will evolve. Because of the intrinsically intertemporal nature of the allocation problem with exhaustibility, the absence of a complete set of futures markets has particularly serious consequences. Again, there is an interesting and original literature on this (for a review, see Dasgupta and Heal, 1979).

See Also

- arbitrage pricing theory
- Hotelling, Harold
- Miller, Merton
- Pontryagin's principle of optimality
- Solow, Robert
- Stiglitz, Joseph E.

Bibliography

Agbeyegbe, T. 1989. Interest rates and metal price movements. *Journal of Environmental Economics and Management* 16, 184–92.

Asheim, G., Buchholz, W. and Withagen, C. 2003. The Hartwick rule: myths and facts. *Environment and Resource Economics* 25, 129–50.

Barnett, H. and Morse, C. 1963. *Scarcity and Growth*. Baltimore, MD: Johns Hopkins University Press.

Brobst, D. 1979. Fundamental concepts for the analysis of resource availability. In *Scarcity and Growth Reconsidered*, ed. V. Kerry Smith. Baltimore, MD: Johns Hopkins Press for Resources for the Future.

Butt, D., Lackner, K., Wendt, C., Nomura, K. and Yanagisawa, Y. 1999. The importance of and a method for disposing of carbon dioxide in a thermodynamically stable form. *World Resource Review* 11, 196–219.

Dasgupta, P. and Heal, G. 1974. The optimal depletion of exhaustible resources. *Review of Economic Studies: Symposium on the Economics of Exhaustible Resources*, 3–28.

Dasgupta, P. and Heal, G. 1979. *Economic Theory and Exhaustible Resources*. Cambridge: Cambridge University Press.

Farzin, H. 1992. The time path of scarcity rent in the theory of exhaustible resources. *Economic Journal* 102, 813–30.

Gale, D. 1967. On optimal development in a multi sector economy. *Review of Economic Studies* 34, 1–18.

Gilbert, R. 1978. Dominant firm pricing policy in a market for an exhaustible resource. *Bell Journal of Economics* 9, 385–95.

Goeller, H. 1979. The age of substitutability: a scientific appraisal of natural resource adequacy. In *Scarcity and Growth Reconsidered*, ed. V. Kerry Smith. Baltimore, MD: Johns Hopkins University Press for Resources for the Future.

Gray, L. 1914. Rent under the assumption of exhaustibility. *Quarterly Journal of Economics* 28, 466–89.

Hanson, D. 1980. Increasing extraction costs and resource prices: some further results. *Bell Journal of Economics* 11, 335–42.

Hartwick, J. 1977. Intergenerational equity and investing the rents from exhaustible resources. *American Economic Review* 66, 9072–74.

Heal, G. 1976. The relationship between price and extraction cost for a resource with a backstop technology. *Bell Journal of Economics* 7, 371–8.

Heal, G. 1984. Interactions between economy and climate: a framework for policy design under uncertainty. In *Advances in Applied Microeconomics*, vol. 3, ed. V. Kerry Smith and A. White. Greenwich, CT and London: JAI Press.

Heal, G. 1985. Depletion and discounting: a classical issue in the economics of exhaustible resources. In *Environmental and Natural Resource Mathematics*, ed. R. McKelvey. *Proceedings of Symposia in Applied Mathematics*, vol. 32. Providence, RI: American Mathematical Society.

Heal, G. 1998. *Valuing the Future: Economic Theory and Sustainability*. New York: Columbia University Press.

Heal, G. 2003. Optimality or sustainability? In *Economics for an Imperfect World: Essays in Honor of Joseph Stiglitz*, ed. R. Arnott et al. Cambridge, MA: MIT Press.

Heal, G. and Barrow, M. 1980. The relationship between interest rates and metal price movements. *Review of Economic Studies* 47, 161–81.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39(2), 137–75.

Krautkramer, J. 1985. Optimal growth, resource amenities and the preservation of natural environments. *Review of Economic Studies* 52, 153–70.

Laxminarayan, R. and Brown, G. 2001. Economics of antibiotics resistance: a theory of optimal use. *Journal of Environmental Economics and Management* 42, 183–206.

Miller, M. and Upton, C. 1985. A test of the Hotelling valuation principle *Journal of Political Economy* 93, 1–25.

Nordhaus, W. 1973. The allocation of energy resources. *Brookings Papers on Economic Activity* 3, 529–76.

Oren, S. and Powell, S. 1985. Optimal supply of a depletable resource with a backstop technology: Heal's theorem revisited. *Operations Research* 33, 277–92.

Pindyck, R. 1978. Gains to producers from the cartelization of exhaustible resources. *Review of Economics and Statistics* 60, 238–51.

Slade, M. 1982. Trends in natural resource commodity prices: an analysis of the time domain. *Journal of Environmental Economics and Management* 9, 122–37.

Smith, V. Kerry. 1980. The evaluation of natural resource adequacy: elusive quest or frontier of economic analysis? *Land Economics* 56, 257–98.

Stiglitz, J. 1976. Monopoly and the rate of extraction of exhaustible resources. *American Economic*

Review 66, 655–61.

Solow, R. and Wan, F. 1976. Extraction costs in the theory of exhaustible resources. *Bell Journal of Economics* 7, 359–70.

Sweeney, J. 1977. Economics of depletable resources: market forces and intertemporal bias. *Review of Economic Studies* 44, 125–41.

How to cite this article

Heal, Geoffrey. "exhaustible resources." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000165> doi:10.1057/9780230226203.0522

The New Palgrave Dictionary of Economics Online

existence of general equilibrium

Gerard Debreu
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes the history of the attempts to prove the existence of general equilibrium from those of Wald and others in Vienna in the 1930s to those of von Neumann and Nash, and of the solutions provided by Arrow, Debreu and McKenzie in the 1950s and their subsequent development.

Keywords

Arrow, K.; convexity; existence of general equilibrium; general equilibrium; Nash, J.; Neumann, J. von; Wald, A.; Walras, L.; Walras's Law

Article

Léon Walras provided in his *Eléments d'économie politique pure* (1874–7) an answer to an outstanding scientific question raised by several of his predecessors. Notably, Adam Smith had asked in *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776) why a large number of agents motivated by self-interest and making independent decisions do not create social chaos in a private ownership economy. Smith himself had gained a deep insight into the impersonal coordination of those decisions by markets for commodities. Only a mathematical model, however, could take into full account the interdependence of the variables involved. In constructing such a model Walras founded the theory of general economic equilibrium.

Walras and his successors were aware that his theory would be vacuous in the absence of an argument supporting the existence of its central concept. But for more than half a century that argument went no further than counting equations and unknowns and finding them to be equal in number. Yet for a non-linear system this equality does not prove that there is a solution. Nor would it provide a proof even for a linear system, especially when some of the unknowns are not allowed to take arbitrary real values.

A successful attack on the problem of existence of a general equilibrium was made possible by an exceptional conjunction of circumstances in Vienna in the early 1930s. It started from the formulation of

the Walrasian model in terms of demand functions which had been given by Gustav Cassel in 1918. As Hans Neisser (1932) noted, certain values of commodity quantities and prices appearing in the solutions of Cassel's system of equations might be negative in such a way as to render those solutions meaningless. Heinrich von Stackelberg (1933) also made a cogent remark. Let x_i be the quantity of the ith final good demanded by consumers, a_{ij} the fixed technical coefficient specifying the input of the jth primary resource required for a unit output of the ith final good, and r_j the available quantity of the jth primary resource. The equality of demand and supply for every resource is expressed by

$$\sum_i a_{ij} x_i = r_j \text{ for all } j.$$

Von Stackelberg observed that if there are fewer final goods than primary resources, the preceding linear system of equations in $(x_1, \ldots, x_m)$ has, in general, no solution. Karl Schlesinger (1933–4) then remarked that equalities should be replaced by inequalities

$$\sum_i a_{ij} x_i \leqq r_j \text{ for all } j,$$

with the condition that a resource for which the strict inequality holds has a zero price. This suggestion, which had already been hinted at by Frederik Zeuthen (1932) in a different context, was essential to the proper formulation of the existence problem.

The problem thus posed received its first solution from Abraham Wald (1933–4), whose work on the existence of a general equilibrium gave rise to three published articles. The first two appeared in *Ergebnisse eines mathematischen Kolloquiums* in 1933–4 and in 1934–5. The third appeared in *Zeitschrift für Nationalökonomie* (1936) and was translated into English in *Econometrica* (1951). In that body of work Wald separately studied a model of production and a model of exchange and proved the existence of an equilibrium for each one.

By the standards prevailing in economic theory at that time, his mathematical arguments were of great complexity, and the major contribution that he had made did not attract the attention of the economics profession. A two-decade pause followed, and when research on the existence problem started again after 1950 it was under the dominant influence of work done, also in the early 1930s, by John von Neumann. His article on the theory of growth, published in *Ergebnisse eines mathematischen Kolloquiums* (1935–6) and translated into English in the *Review of Economic Studies* (1945), contained in particular a lemma of critical importance. That lemma was reformulated in the following far more convenient form, and was also given a significantly simpler proof, by Shizuo Kakutani (1941). Let K be a non-empty, compact, convex set of finite dimension. Associate with every point x in K a non-empty,

convex subset $\phi(x)$ of K, and assume that the graph $G = \{(x, y) \in K \times K | y \in \varphi(x)\}$ of the transformation ϕ is closed. Then ϕ has a fixed point x^*, i.e., a point x^* that belongs to its image $\phi(x^*)$. Kakutani's theorem was applied by John Nash, in a one-page note of 1950, to establish the existence of an equilibrium for a finite game. It can be used as well (Debreu, 1952) to prove the existence of an equilibrium for a more general system composed of n agents. The ith agent chooses an action a_i in a set A_i of *a priori* possible actions. A state of the social system is therefore described by the list $a = (a_1, \ldots, a_n)$ of the actions chosen by the n agents. The preferences of the ith agent are represented by a real-valued utility function u_i defined for every a in the set of states $A = \times_{i=1}^{n} A_i$. Moreover the ith agent is restricted in the choice of his action in A_i, by the actions chosen by the other agents. Formally let N denote the set $\{1,\ldots, n\}$ of all the agents and $N\backslash i$ denote the set of the agents other than the ith. Let also $a_{N\backslash i}$ denote the list of the actions $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ chosen by the agents in $N\backslash i$. The ith agent is constrained to choose his own action in a subset $\phi_i(a_{N\backslash i})$ of A_i depending on $a_{N\backslash i}$. In these conditions the ith agent, considering $a_{N\backslash i}$ as given, chooses his action in $\mu_i(a_{N\backslash i})$, the set of the elements of $\phi_i(a_{N\backslash i})$ at which the maximum of the utility function $u_i(\cdot, a_{N\backslash i})$ in $\phi_i(a_{N\backslash i})$ is attained. Consider now the transformation $a \mapsto \mu(a) = \times_{i=1}^{n} \mu_i(a_{N\backslash i})$ associating with any element a of A, the subset $\mu(a)$ of A. A state a^* is an equilibrium if and only if for every $i \in N$, the action a_i^* of the ith agent is best according to his preferences given the actions $a_{N\backslash i}^*$ of the others, that is, if and only if for every $i \in N$, $a_i^* \in \mu_i(a_{N\backslash i}^*)$, that is, if and only if $a^* \in \mu(a^*)$. Thus the concept of an equilibrium for the social system is equivalent to the concept of a fixed point for the transformation $a \mapsto \mu(a)$ of elements of A into subsets of A. Ensuring that the assumptions of Kakutani's theorem are satisfied for the transformation μ yields a proof of existence of an equilibrium for the social system.

In the revival of interest in the problem of existence of a general economic equilibrium after 1950, the first solutions were published in 1954 by Kenneth Arrow and Gerard Debreu, and by Lionel McKenzie. The article by McKenzie emphasized international trade aspects, and the article by Arrow and Debreu dealt with an integrated model of production and consumption. Both rested their proofs on Kakutani's theorem. They were followed over the next three decades by a large number of publications (a bibliography is given in Debreu, 1982) which confirmed the concept of a Kakutani fixed point as the most powerful mathematical tool for proofs of existence of a general equilibrium.

A simple prototype of the various economies that were the subject of those numerous existence results is (following Arrow–Debreu) composed of m consumers and n producers, producing, exchanging and consuming l commodities. The consumption of the ith consumer $(i = 1, \ldots, m)$ is a vector x_i in R^l whose positive (or negative) components are his inputs (or outputs) of the l commodities. Similarly the production of the jth producer $(j = 1, \ldots, n)$ is a vector y_j in R^l whose negative (or positive) components are his inputs (or outputs) of the l commodities. The ith consumer has three characteristics. (1) His consumption set X_i, a non-empty subset of R^l, is the set of his possible consumptions. (2) A binary relation $\precsim_i$ on X_i defines his preferences, and '$x_i \precsim_i x_i'$' is read as 'x_i' is at least as desired as x_i by the ith consumer'. Formally the preference relation of the ith consumer is the set

$\{(x, x') \in X_i \times X_i | x \precsim_i x'\}$. (3) A vector e_i in R^l describes his initial endowment of commodities. The jth producer has one characteristic, his production set Y_j, a non-empty subset of R^l defining his possible productions. Finally the number $\theta_{ij} \geq 0$ specifies the fraction of the profit of the jth producer distributed to the ith consumer. These numbers satisfy the equality $\sum_{i=1}^{m} \theta_{ij} = 1$ for every j. In summary, the economy $\mathcal{E}$ is characterized by the list of mathematical objects

$$\left[(X_i, \precsim_i, e_i)_{i=1,\ldots,m}, (Y_j)_{j=1,\ldots,n}, (\theta_{ij})_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}} \right].$$

Given a price-vector p in R^l different from 0, the jth producer $(j = 1, \ldots, n)$ chooses a production y_j in Y_j that maximizes his profit, that is, such that the value $p \cdot y_j$ of y_j relative to p satisfies the inequality $p \cdot y_j \geq p \cdot y$ for every y in Y_j. Thus the ith consumer receives in addition to the value $p \cdot e_i$ of his endowment, $\sum_{j=1}^{n} \theta_{ij} p \cdot y_j$ as the sum of his shares of the profits of the n producers. The value $p \cdot x$ of his consumption x is therefore constrained by the budget inequality $p \cdot x \leqq p \cdot e_i + \sum_{j=1}^{n} \theta_{ij} p \cdot y_j$. Under that constraint he chooses a consumption x_i in X_i that is best according to his preferences. The list $[p, (x_i)_{i=1,\ldots,m}, (y_j)_{j=1,\ldots,n}]$ of a non-zero price-vector, m consumptions and n productions forms a general equilibrium of the economy $\mathcal{E}$ if for every commodity, the excess of demand over supply vanishes,

$$\sum_{i=1}^{m} x_i - \sum_{j=1}^{n} y_j - \sum_{i=1}^{m} e_i = 0.$$

The existence of a general equilibrium can be proved (following Arrow–Debreu) by casting the economy $\mathcal{E}$ in the form of a social system of the type defined above. For this it suffices to introduce, in addition to the m consumers and to the n producers, a fictitious price-setting agent whose set of actions and whose utility function are now specified. Note first that the definition of a general equilibrium is invariant under multiplication of the price-vector p by a strictly positive real number. In the simple case where all prices are non-negative, one can therefore restrict p to be an element of the simplex $P = \{p \in R^l_+ | \sum_{h=1}^{l} p^h = 1\}$, the set of the vectors in R^l whose components are non-negative and add up to one. The set of actions of the price-setter is specified to be P. Given the consumptions $(x_i)_{i=1,\ldots,m}$ chosen by the m consumers, and the productions $(y_j)_{j=1,\ldots,n}$ chosen by the n producers,

there results an excess demand

$$z = \sum_{i=1}^{m} x_i - \sum_{j=1}^{n} y_j - \sum_{i=1}^{m} e_i.$$

The utility function of the price-setter is specified to be $p \cdot z$. Maximizing the function $p \mapsto p \cdot z$ over *P* carries to one extreme the idea that the price-setter should choose high prices for the commodities that are in excess demand, and low prices for the commodities that are in excess supply.

Some of the assumptions on which the theorems of Arrow–Debreu (1954) are based are weak technical conditions: closedness of the consumption-sets, of the production-sets and of the preference relations, existence of a lower bound in every coordinate for each consumption-set, possibility of a null production for each producer. Other assumptions were later shown to be superfluous for economies with a finite set of agents: irreversibility of production (if both *y* and –*y* are possible aggregate productions, then $y = 0$), free disposal (any aggregate production $y \leq 0$ is possible), and completeness and transitivity of preferences. Convexity of preferences can be dispensed with, and convexity of consumption-sets can be weakened, in economies with a large number of small consumers. Insatiability of consumers is an acceptable behavioural postulate. There remain, however, two overly strong assumptions. They are the hypothesis that for every *i*, the endowment e_i yields a possible consumption for the *i*th consumer (after disposal of a suitable commodity-vector if need be), and the assumption of convexity on the total production-set $Y = \Sigma_{j=1}^{n} Y_j$ which implies non-increasing returns to scale in the aggregate.

An alternative approach to the problem of existence of a general equilibrium, closer to traditional economic theory, is centred on the concept of excess demand function, or of excess demand correspondence. Given an economy $\mathcal{E}$ defined as before, consider a price-vector *p* in R_+^l different from 0. The productions $(y_1,\ldots,y_n)$ chosen by the producers, and the consumptions $(x_1,\ldots,x_m)$ chosen by the consumers in reaction to the price-vector *p* result in an excess demand *z* in the commodity-space R^l. If *z* is uniquely determined, the excess demand function *f* from $R_+^l \setminus 0$ to R^l is thereby defined. If *z* is not uniquely determined, the set of excess demands in R^l associated with *p* is denoted by $\phi(p)$, and the excess demand correspondence ϕ is thereby defined on $R_+^l \setminus 0$. Both *f* and ϕ are homogeneous of degree zero since $f(p)$ and $\phi(p)$ are invariant under multiplication of *p* by a strictly positive real number. This permits various normalizations of *p*. For instance, *p* may be restricted to the simplex *P*. Moreover, for every $i = 1, \ldots, m$, one has $p \cdot x_i \leq p \cdot e_i + \Sigma_{j=1}^{n} \theta_{ij} p \cdot y_j$. By summation over *i*, one obtains

$$p \cdot \sum_{i=1}^{m} x_i \leq p \cdot \sum_{i=1}^{m} e_i + p \cdot \sum_{j=1}^{n} p \cdot y_j,$$

or equivalently $p \cdot z \leq 0$. Therefore for every p in $R^l_+ \backslash 0$, one has either $p \cdot f(p) \leq 0$ or $p \cdot \varphi(p) \leq 0$. This observation leads to the following proof of existence of a general equilibrium (Gale, 1955; Nikaidô, 1956; Debreu, 1956). Let ϕ be a correspondence transforming points of the simplex P into non-empty convex subsets of R^l. If ϕ is bounded, has a closed graph and satisfies $p \cdot \varphi(p) \leq 0$ for every p in P, then, by Kakutani's theorem, there are a point p^* in P and a point z^* in R^l such that $z^* \in \varphi(p^*)$ and $z^* \leq 0$. In economic terms, there is a price-vector p^* in P yielding an associated excess demand z^* in $\phi(p^*)$, all of whose components are negative or zero.

If all the consumers in the economy $\mathcal{E}$ are insatiable, every individual budget constraint is binding, and one has for every i, $p \cdot x_i = p \cdot e_i + \Sigma_{j=1}^{n} \theta_{ij} p \cdot y_j$. By summation over i, $p \cdot z = 0$. Thus in the case where the vector z associated with p is uniquely determined, the excess demand function satisfies

$$\text{Walras's Law: for every } p \text{ in } R^l_+ \backslash 0,\ p \cdot f(p) = 0.$$

In geometric terms, in the commodity-price space R^l the vectors p and $f(p)$ are orthogonal. This prompts one to normalize the price-vector p so that it belongs to the positive part of the unit sphere $\bar{S} = \{p \in R^l_+ \mid \|p\| = 1\}$, for then $f(p)$ can be represented as a vector tangent to $\bar{S}$ at p. The excess demand function is now seen as a vector field on $\bar{S}$. This in turn suggests another proof of existence of a general equilibrium (Dierker, 1974) for the particular case of an exchange economy $\mathcal{E}$ whose consumers have continuous demand functions, monotone preferences and strictly positive endowments of all commodities. In that case for every $i = 1, \ldots, m$, the consumption-set X_i of the ith consumer is R^l_+ and $x < x'$ implies $x \prec_i x'$ (if x' is at least equal to x in every component and $x' \neq x$, then x' is preferred to x). Since the demand of a consumer with monotone preferences is not defined when some prices vanish, one must restrict the price-vector p to be strictly positive in every component, that is, to belong to $S = \{p \in \text{Interior } R^l_+ \mid \|p\| = 1\}$. Moreover let p_q be a sequence of price-vectors in S converging to p_0 in the boundary $\bar{S} \backslash S$ of S. Thus for every q, the vector p_q is strictly positive in each component, while, in the limit, p_0 has some zero components. Then the associated sequence of excess demands $f(p_q)$ is unbounded. As a consequence, the vector field f points inward towards S near the boundary of S. In these conditions Brouwer's fixed point theorem yields the existence of an equilibrium price-vector p^* in S for which excess demand vanishes, $f(p^*) = 0$.

The preceding solutions of the problem of existence of a general equilibrium all rest directly on fixed point theorems. Three different lines of approach are provided by (1) combinatorial algorithms for the computation of approximate general equilibria, (2) differential processes converging to general

equilibria, (3) the theory of the fixed point index of a mapping.

1. (1) The past two decades have witnessed the development of algorithms of a combinatorial nature for the computation of an approximate general equilibrium (*see* computation of general equilibria). Given any number $\varepsilon > 0$, a constructive procedure thereby yields a price-vector p such that the norm $|f(p)|$ of the associated excess demand is smaller than ε. A compactness argument then gives a sequence of price-vectors p_q in S converging to p_0 for which $|f(p_q)|$ tends to 0. In the limit, $f(p_0) = 0$.
2. (2) Global analysis was introduced into economic theory at the beginning of the 1970s to study the set of general equilibria of an economy and the manner in which it depends on the economy. In that framework Stephen Smale proposed in (1976) a differential process which starts from a point in the boundary of the set of normalized price-vectors, and which converges to the set of equilibria provided that the initial point does not lie in a negligible exceptional set (*see* global analysis in economic literature). Another constructive procedure thus gives, from a differentiable viewpoint, conditions under which the set of general equilibria is not empty.
3. (3) In the same differentiable framework Egbert Dierker (1972) used the theory of the fixed point index of a mapping to prove that a regular economy (as defined by him in regular economies below) whose excess demand points inward near the boundary of S has an odd (hence non-zero) number of general equilibria. The significance of this theorem rests on the fact that under its assumptions almost every economy is regular.

The previous existence results have been extended in many directions. The study of the core of an economy led to the consideration of a set of agents, all of whom are negligible relative to their totality. This concept was formalized first as an atomless measure space of agents, and later by means of non-standard analysis. In both cases the existence of a general equilibrium had to be proved for economies with infinitely many agents.

In order to specify a commodity one lists its physical characteristics, the date, the location, and the event at which it is available. As soon as one of those four variables can take infinitely many values, the analysis of general equilibrium must be set in the framework of infinite-dimensional commodity spaces. Several existence results were obtained in that context.

In yet another direction, external effects called for extensions. When the characteristics of each agent (e. g. his preferences, his production set, …) depend on the actions chosen by the other agents, formulating the economy as a social system of the type described earlier immediately yields an existence theorem. Still other extensions have covered economies with public goods, with indivisible commodities, and with non-convex production sets.

See Also

- Arrow–Debreu model of general equilibrium
- fixed point theorems
- general equilibrium

Bibliography

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Arrow, K.J. and Intriligator, M.D., eds. 1981–6. *Handbook of Mathematical Economics* Vols I–III, Amsterdam: North-Holland.

Cassel, K.G. 1918. *Theoretische Sozialökonomie*. Leipzig: C.F. Winter.

Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences* 38, 886–93.

Debreu, G. 1956. Market equilibrium. *Proceedings of the National Academy of Sciences* 42, 876–8.

Debreu, G. 1982. Existence of competitive equilibrium. Ch. 15 in Arrow and Intriligator (1981–6).

Dierker, E. 1972. Two remarks on the number of equilibria of an economy. *Econometrica* 40, 951–3.

Dierker, E. 1974. *Topological Methods in Walrasian Economics*. Berlin, New York: Springer-Verlag.

Gale, D. 1955. The law of supply and demand, *Mathematica Scandinavica* 3, 155–69.

Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8, 457–9.

McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22, 147–61.

Nash, J.F. 1950. Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences of the USA* 36, 48–9.

Neisser, H. 1932. Lohnhöhe und Beschäftigungsgrad im Marktgleichgewicht. *Weltwirtschaftliches Archiv* 36, 415–55.

Neumann, J. von. 1935–36. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, 73–83. Trans. by G. Morgenstern as ‘A model of general economic equilibrium’, *Review of Economic Studies* 13(1), 1945, 1–9.

Nikaidô, H. 1956. On the classical multilateral exchange problem. *Metroeconomica* 8, 135–45.

Schlesinger, K. 1933–4. Über die Produktionsgleichungen der ökonomischen Wertlehre. *Ergebnisse eines mathematischen Kolloquiums* 6, 10–20.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3, 107–20.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 2 vols, ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Clarendon Press, 1976.

von Stackelberg, H. 1933. Zwei kritische Bemerkungen zur Preistheorie Gustav Cassels. *Zeitschrift für Nationalökonomie* 4, 456–72.

Wald, A. 1933–4. Über die eindeutige positive Lösbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines mathematischen Kolloquiums* 6, 12–20.

Wald, A. 1934–5. Über die Produktionsgleichungen der ökonomischen Wertlehre. *Ergebnisse eines mathematischen Kolloquiums* 7, 1–6.

Wald, A. 1936. Über einige Gleichungssysteme der mathematischen ökonomie. *Zeitschrift für Nationalökonomie* 7, 637–70. Trans. by Otto Eckstein as ‘On some systems of equations of mathematical economics’, *Econometrica* 19(4), October 1951, 368–403.

Walras, L. 1874–7. *Éléments d’économie politique pure*. Lausanne: L. Corbaz. Trans. by William Jaffé as *Elements of Pure Economics*, Homewood, Ill.: Richard D. Irwin, 1954.

Zeuthen, F. 1932. Das Prinzip der Knappheit, technische Kombination und ökonomische Qualität. *Zeitschrift für Nationalökonomie* 4, 1–24.

How to cite this article

Debreu, Gerard. "existence of general equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000166> doi:10.1057/9780230226203.0523

The New Palgrave Dictionary of Economics Online

exit and voice

Albert O. Hirschman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Exit and voice are alternative responses to an unsatisfactory relationship: exit is the withdrawal from it, voice is the attempt to improve it through communication. They are not mutually exclusive responses: thus, the market is the archetypal exit mechanism, yet it usually involves voice. When available jointly, exit and voice may reinforce or undercut each other: the exit option enhances the influence of customers' voice on an unsatisfactory supplier but also reduces its volume. Exit–voice analysis has been applied to trade unions, hierarchies, public services, migration and political action, political party systems, marriage and divorce, and adolescent development.

Keywords

adolescent development; asymmetric information; city–suburb migration; collective voice; democratization; education finance; exit and voice; free rider problem; health finance; hierarchy; Hirschman, A. O.; horizontal vs vertical voice; identity; incomplete information; international capital flows; international migration; loyalty; marriage and divorce; Montesquieu, C. de; multiparty systems; political parties; public services; rural–urban migration; Smith, A.; Smith, A.; trade unions; two-party systems; vouchers; welfare state

Article

A central place is held in economics and social science in general by principles and forces making for order or equilibrium in economic and social systems. Disorder and disequilibrium are then understood as resulting from some malfunction of these principles or forces. Explanations of order–disorder or equilibrium–disequilibrium have typically been discipline-bound, dealing with either the political or the economic world. Since the two are interrelated it would be useful to have a construct that bridges them. Such is the claim of the exit–voice perspective. It addresses the changing balance of order and disorder in the social world by pointing out that social actors who experience developing disorder have available

to them two activist reactions and perhaps remedies: *exit*, or withdrawal from a relationship that one has built up as a buyer of merchandise or as a member of an organization such as a firm, a family, a political party or a state; and *voice*, or the attempt at repairing and perhaps improving the relationship through an effort at communicating one's complaints, grievances and proposals for improvement. The voice reaction belongs in good part to the political domain since it has to do with the articulation and channelling of opinion, criticism and protest. Much of the exit reaction, on the contrary, involves the economic realm as it is precisely the function of the markets for goods, services, and jobs to offer alternatives to consumers, buyers and employees who are for various reasons dissatisfied with their current transaction partners.

The exit–voice alternative was proposed and explored in *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States* (Hirschman, 1970, henceforth *EVL*). Attempts to apply the book's perspective were made over many areas of social life. In the following, the basic concepts will be recapitulated and, where necessary, reformulated. Subsequently some major applications of the exit–voice polarity will be reviewed.

Basic concepts

Exit

Exit means withdrawal from a relationship with a person or organization. If this relationship fulfils some vital function, then the withdrawal is possible only if the same relationship can be re-established with another person or organization. Exit is therefore often predicated on the availability of choice, *competition*, and well-functioning *markets*.

Exit of customers (or employees) serves as a signal to the management of firms and organizations that something is amiss. A search for causes and remedies will then be undertaken and some plan of action designed to restore performance will be adopted. This is one way in which markets and competition work to prevent decay and to maintain and perhaps improve quality.

Exit is a powerful but indirect and somewhat blunt way of alerting management to its failings. Most of the time, those customers and members of organizations who exit have no interest in improving them by their withdrawal, so that exit does not provide management with much information on what is wrong.

Voice

The direct and more informative way of alerting management is to alert it: this is *voice*. Its role is, or should be, paramount in situations where exit is either not available at all or is difficult, costly, and traumatic. This is so for certain primordial groupings one is born into – the family, the ethnic or religious community, the nation – or for those organizations one joins with the intention of staying for a prolonged period – school, marriage, political party, firm. With regard to buying and selling, voice should take over from exit when competition is weak or nonexistent as in the case of goods and services being produced under oligopolistic or monopolistic conditions, or when exit is expensive for both parties as in certain interfirm relations.

Unlike exit in the case of well-functioning markets, voice is never easy. It can even be dangerous. Many

organizations and their agents are not at all keen on being told about their shortcomings by members and the latter often expose themselves to reprisals if they utter any criticism (Birch, 1975). Even in the absence of reprisals, the cost of voice to an individual member will often exceed, in terms of time and effort, any conceivable benefit from voicing. Frequently, moreover, any effective channelling of individual voices requires a number of members to join together so that voice formation depends on the potential for collective action.

In spite of these problems, voice exists or, rather, it has come into being. Its history is to a considerable extent the history of the right to dissent, of due process, of safeguards against reprisal, and of the advance of trade unions and of consumer and many other organizations articulating the demands of individuals and groups who once were silent. Similarly, the history of exit is the history of the broadening of the market, of the right to move freely, to emigrate, to be a conscientious objector, to divorce, etc. Being two basic, complementary ingredients of democratic freedom, the right to exit and the right to voice have on the whole been enlarged or restricted jointly. Yet, there are important instances of unilateral advances or retreats of either the one or the other response mechanism (Rokkan, 1975; Finer, 1974).

Interaction of exit and voice

As noted, exit is paramount as a reaction to discontent in some circumstances and voice holds a similarly privileged position in others, but frequently both mechanisms are available jointly. In such situations they may either reinforce or undercut each other. The availability and threat of exit on the part of an important customer or group of members may powerfully reinforce their voice. On the other hand, the actual recourse to exit will often diminish the volume of voice that would otherwise be forthcoming and, should the organization be more sensitive to voice than to exit, the stage could be set for cumulative deterioration. For example, after an incipient deterioration of public schools or inner cities, the availability of private schools or suburban housing would lead, via exit, to further deterioration – a turn of events that might have been prevented if the parents sending their children to private school or the inner city residents who move to the suburbs had instead used their voice to press for reform. In their aggregate effects, the individual exit decisions are harmful – an instance of the ‘tyranny of small decisions’ – also because they are likely to be taken on the basis of a short-run private-interest calculus only and do not take into account the ‘public bad’ that will be inflicted, even on those who exit, by decaying inner cities and segregated education (Levin, 1983; Breneman, 1983).

These kinds of situations are sufficiently numerous and important to be of interest not only as curious paradoxes showing that under some circumstances the availability of exit (that is, of competition) could have undesirable effects. In this connection, *EVL* stressed the value of *loyalty* as a factor that might delay over-rapid exit. Loyalty would make a member reluctant to leave an organization upon the slightest manifestation of decline even though rival organizations were available. Provided it is not ‘blind’, loyalty would also activate voice as loyal members are strongly motivated to save ‘their’ organization once deterioration has passed some threshold.

The difficulties of combining exit and voice in an optimal manner are in a sense ‘problems of the rich’: they relate to situations and societies where exit and voice are both forthcoming more or less abundantly, but where, for best results, one would wish for a different mix. Historically more frequent are cases

where exit and voice are both in short supply, in spite of many reasons for discontent and unhappiness. There is no doubt, as many commentators have pointed out, that passivity, acquiescence, inaction, withdrawal, and resignation have held sway much of the time over wide areas of the social world. This is largely the result of repression of both exit and voice – a repression that has flourished in spite of the fact that all human organizations could put to good use the feedback provided by the two reaction modes.

Problems in voice formation

The development of voice among customers of firms or members of organizations poses a number of problems that were not fully explored in *EVL*. Critics have asserted that, in its endeavour to present voice as a ready alternative to exit, the book understated the difficulties of voice formation. In examining this issue it is useful to start with the extreme no-voice case: the authoritarian state which is dedicated to repressing and suppressing voice. This situation has given rise to a useful distinction between *horizontal* and *vertical* voice (O'Donnell, 1986). The latter is the actual communication, complaint, petition, or protest addressed to the authorities by a citizen and, more frequently, by an organization representing a group of citizens. Horizontal voice is the utterance and exchange of opinion, concern and criticism *among* citizens: in the more open societies it is today regularly ascertained through opinion polls revealing the approval rating of presidents, prime ministers, mayors, etc. Horizontal voice is a necessary precondition for the mobilization of vertical voice. It is the earmark of the more frightful authoritarian regimes that they suppress not only vertical voice – any ordinary tyranny does that – but horizontal voice as well. The suppression of horizontal voice is generally the side-effect of the terrorist methods used by such regimes in dealing with their enemies.

The distinction between vertical and horizontal voice is relevant to the ‘free ride’ argument in relation to voice formation (Barry, 1974). For *vertical* voice to come about, that is, for members of the organization to engage management in meaningful dialogue, it is frequently necessary for members to forge a tie among themselves, to create an organization which will agitate for their demands, etc. But the hoped-for result of collective voice is a freely available public good; hence, so goes the critical argument, self-interested, ‘rational’ individuals may well withhold their contribution to the voice enterprise in the expectation that others will take on the entire burden. Important as it is, this argument has its limitations. First of all, it is addressed only to vertical voice which it mistakenly equates (as *EVL* did) with voice in general. Horizontal voice is not subject to the strictures of the free-rider argument: it is free, spontaneous activity of men and women in society, akin to breathing. As just noted, extraordinary violence has to be deployed if it is to be suppressed. Under ordinary circumstances, horizontal voice is continuously generated and has an impact even without becoming vertical: in many environments managers of organizations cannot help noticing and reacting to critical opinions and hostile moods of the members, whether or not organized protest movements break out. That the planned economies of Eastern Europe function to the extent they do has been explained on precisely this ground (Bender, 1981, p. 30).

Another limitation of the free-rider argument lies in its assumption that individuals will always act instrumentally. Just because the desired result of collective voice is typically a *public* good – or, better, some aspect of the *public* happiness – participation in voice provides an alternative to self-centred, instrumental action. It therefore has the powerful attractions of those activities that are characterized by the fusion of striving and attaining and can be understood as investments in individual or group identity

(Hirschman, 1985).

Some areas of application

Trade unions

In economics, the major application of the exit–voice theme has been the analysis of trade unions as collective voice by Freeman and Medoff in their book *What Do Unions Do?* (1984). Instead of looking at unions as a monopolistic device raising wages for unionized workers beyond the 'market-clearing' equilibrium level or – much the same zero-sum interpretation in different language – as a tool in the class struggle serving to reduce the degree of exploitation, the book finds that a major function of unions is that of channelling information to management about workers' aspirations and complaints. Collective voice, in the form of union bargaining, is more efficient in conveying information about workers' discontent – and in doing something about it – than individual decisions to quit, as voice carries more information than exit. The presence of union voice is shown to reduce costly labour turnover. Moreover, the fringe benefits, workplace practices, and seniority rules which unions negotiate often result in offsetting labour productivity increases.

Markets and hierarchies vs. exit and voice

Renewed attention has been given in recent years to the question why some kinds of economic activities are carried on through many independent firms while others, to the contrary, are tied together through bureaucratic and hierarchical relations. In accounting for hierarchy, one approach has directed attention to such matters as uncertainty about the evolution of the market and the technology and in particular to asymmetric availability of information to buyer and seller, creating opportunities for deceitful behaviour (Williamson, 1975). Hierarchy is then seen as superior to markets whenever there is need for a sustained and frank dialogue between the contracting parties. Critics of this position have argued: (1) relations between independent firms, such as contractors and subcontractors, are often quite effective in discouraging malfeasance; (2) correlatively, hierarchy frequently leads to characteristic patterns of concealment and control evasion (Eccles, 1981; Granovetter, 1985); and (3) industry structure varies substantially from one country to another as well as within the same country over time: in Japan, for example, subcontracting is much more widely practised than in the West and in Italy subcontracting has become more widespread in the last 10–20 years.

A formulation in terms of exit–voice is helpful here. The characteristics which are said to justify hierarchy – incomplete information, considerable apprenticing of one firm by the other, openings for 'opportunistic' (i.e., dishonest) behaviour, etc. – all make for situations in which there is need for voice: the firms contracting together must intensively consult with, and watch over, each other. *But the need for voice does not necessarily imply that hierarchy is in order*. Whether voicing is done best within the same organization or from one independent firm to another is by no means a foregone conclusion. Moreover, when the two parties are independent and resort a great deal to voice, the possibility of exit from the relationship often looms in the background. The implicit threat of exit could carry as much clout as that of sanctions in hierarchical relationships.

The argument for hierarchy in cases where voice has an important role to play may arise from thinking of market relationship only in terms of the ideal, anonymous market where voice is wholly absent. But most markets involve voice: commerce *is* communication, and is premised on frequent and close contact of the contracting parties who deliver promises, trust them, and engage in mutual adjustment of claims and complaints – all of this was implicit in the eighteenth-century notion of *doux commerce* (Hirschman, 1977, 1982). Adam Smith even conjectured that it was man's ability to communicate through speech that lies at the source of his ‘propensity to truck and barter’. How odd, then, that the need for frequent and intensive communication should be adduced as a conclusive argument for hierarchy.

Public services: education, health, others

The organization of public services represents a privileged area for the application of exit–voice reasoning – significantly the exit–voice idea had its origin in the analysis of a public service in trouble, the Nigerian railroads (*EVL*, Preface). Public services are typically sold or delivered by a single public or publicly regulated supplier, for various well-known reasons.

With the production of most public services being thus deprived of the ‘discipline of the market’, problems of productive efficiency and quality maintenance arise necessarily. An obvious way of mitigating these problems is to attempt to reintroduce market pressures in some fashion. For example, when certain categories of goods and services are to be made available either to all citizens regardless of their income or to some deprived social groups, the state and its agencies can sometimes refrain from producing or distributing these goods directly, and instead issue special purpose money or *vouchers* enabling the beneficiaries to acquire the goods or services through ordinary market channels. In this manner the voucher system reintroduces the market and the possibility of exit. A particularly successful example of the voucher system is the distribution of Food Stamps to low-income persons in the United States. Instead of creating and administering its own food distribution network the state hands out vouchers (food stamps) which the beneficiaries can then use at existing, competitive commercial outlets. In part because of the success of this programme and in part because of the belief in ‘market solutions’ as the remedy for all that ails government programmes, voucher schemes have been proposed for a large number of other public services, from education to low-cost housing to the supply of certain health services. Voucher systems are appropriate primarily under the following conditions (Bridge, 1977): (1) there are widespread differences in tastes and these differences are recognized as legitimate; (2) individuals are well informed about quality and different qualities are easily compared and evaluated; (3) purchases are recurrent and relatively small in relation to income so that buyers can learn from experience and easily switch from one brand and supplier to another.

These conditions are ideally present in the case of foodstuffs, but much less so in the case of, say, health and educational services. Hence the development of voice constitutes here an important alternative strategy for assuring and maintaining product quality. In other words, the beneficiaries of certain public services should be induced to become active on their own behalf, individually or collectively As always, development of voice is arduous because of apathy and passivity of the members, but also because it will often be resisted by the organizations that have been set up to deliver the services. A number of proposals and attempts have been made to introduce more voice into the administration of both health and educational services (Stevens, 1974; Klein, 1980).

EVL had insisted on the see-saw character of exit and voice interventions in these fields. Education and health systems seemed particularly exposed to the danger that premature exit – of the potentially most influential members – would undermine voice. The opposite relation may also occur, however, for the opening up of the exit perspective could serve to strengthen voice: parents who have been wholly passive because of feelings of powerlessness and fear of reprisals may feel empowered for the first time once they are given vouchers that could be used 'against' the schools currently attended by their children, and will be more ready than before to speak out with regard to desirable changes in those very schools.

Spatial mobility (migration) and political action

Another substantial area of exit–voice applications opens up when exit is taken in the literal, spatial sense. Here exit–voice boils down to the familiar flight or fight alternative. While often institutionalized among nomadic groups (Hirschman, 1981, ch. 11), this alternative is not necessarily available in sedentary societies. Here the traditionally available choice is fight or submit in silence. The option of removing oneself from an oppressive environment has become available on a massive scale only in modern times, with the advances in transportation and the *uneven* opening up of economic opportunity, religious tolerance, and political freedom. Where the option has existed, the interaction of exit and voice has been on display in three principal types of migration: (1) that from the countryside to the city, the oldest and no doubt largest of the modern migrations; (2) the migration from the city to the suburbs, which was most intense in the United States during the fifties and sixties, owing to the spread of the automobile and also to the large-scale migration of blacks and Hispanics *into* the cities; (3) finally, of course, international migration with its numerous economic and political determinants and constraints. Under this rubric, the international movement of capital also deserves attention.
Looking at the varieties of exit-voice interplay in these diverse settings, it is possible, on the basis of the numerous studies now available, to distinguish the following patterns:

1. (1) In accordance with the basic hypothesis of *EVL*, exit-migration deprives the geographical unit which is left behind (countryside, city, nation) of many of the more activist residents, including potential leaders, reformers, or revolutionaries. Exit weakens voice and reduces the prospects for advance, reform, or revolution in the area that is being left.
   Something of this pattern can be observed in all three types of migration. Massive rural-urban migration could obviously reduce the potential as well as the need for land reforms which the voice of the countryside might otherwise have precipitated (Huntington and Nelson, 1976, pp. 103ff.). The large outward migration from Europe to the United States in the 19th century up to World War I probably functioned as a political safety-valve for the rapidly industrializing European societies of that period, as has been shown for Italy (MacDonald, 1963–4). In a similar vein, the possibility of westward migration within the United States has been invoked as an explanation for the lack of a militant working-class movement in that country. Finally, the city-to-suburbs migration in the United States has led, at least initially, to cumulative deterioration in the urban areas affected by out-migration in spite of, and in some cases because of, reduced density.
   At times, the voice-weakening effect of exit is consciously utilized by the authorities: permitting,

favouring, or even ordering the exit of enemies or dissidents has long been one – comparatively civilized – means for autocratic rulers to rid themselves of their critics, a practice revived on a large scale by Castro's Cuba and, on a more selective basis, by the Soviet Union.

2. (2) But the basic see-saw pattern – the more exit the less voice – does not exhaust the rich historical material. The mechanism through which voice is strengthened rather than weakened as a result of exit is distinctive in the case of migration. In some societies the accumulated social pressures could be so high that authoritarian political controls will only be relaxed if a certain amount of out-migration takes place concurrently. This is what happened in the fifty years prior to World War I when the franchise was extended in many European states from which large contingents of people were departing. In other words, the state accommodated some of the pressures toward democratization because it could be reasonably surmised, in part as a result of out-migration, that opening the door slightly to voice would not blow away the whole structure. A similar positive relation between exit and voice may exist today with regard to such southern European countries as Spain, Portugal, and Greece: here the large-scale emigration to northern Europe may also have eased the transition to a more democratic (more vociferous) order.
3. (3) Exit–voice theory posits remedial or preventive responses to any large-scale out-migration on the part of the entity that is being left. A firm losing customers or a party losing members will normally undertake a search for the reasons of such declines in fortune and then determine upon a strategy for recovery. For out-migration such reactions are not easy to identify. In the case of massive rural–urban migration, for example, there is usually no organized entity such as the ‘countryside’ that registers the flight from it and can undertake corrective action. With regard to migration from the city to the suburbs, the situation is not too different. Here entities exist – city administrations – but they have generally been ineffective in modifying the individual decisions of millions of people to move into their own homes in the suburbs.

The analogy to the firm is – or should be – most applicable when the geographic entity losing residents is the State, which is after all a highly organized, self-reflective body with considerable means of action. There is, of course, the already noted possibility that out-migration relieves economic or political stress in a country, is therefore *welcome*, and may even be encouraged by the state. But massive emigration is at some point bound to be viewed as dangerous. Just like a business firm, the state may then take measures to make itself more attractive to its citizens. One example of this reaction is the national plan for economic recovery and industrialization adopted by Ireland in 1958, in the midst of very high levels of emigration, mostly to England (Burnett, 1976). It has also been shown that the pioneering welfare state measures of the late 19th and early 20th century, starting in Bismarck's Germany in the eighties and then spreading to the Scandinavian countries and Great Britain, were all taken in countries with high rates of overseas migration. These measures can be seen as attempts of states to make themselves more attractive to their citizens (Kuhnle, 1981).

The international movement of capital was first commented upon from the exit perspective in the 18th century. Montesquieu and Adam Smith both thought that the threat of exit on the part of movable capital could play a useful role in preventing arbitrary and confiscatory measures against the legitimate interests of commerce and industry. The threat of exit or exit itself was expected to function, like the customer's exit, as a curb on misconduct, this time on the part of the state. While this relationship is still pertinent, exit of capital often plays a less constructive role today. In the more peripheral capitalist countries the

owners of capital have become fully alive to the possibility of removing part of their holdings to the United States or other reliable places in case they become unhappy about the ‘investment climate’. In this manner, capital exit (or flight) will often be practised on a large scale as soon as the state undertakes some, perhaps long overdue, reforms with respect to such matters as land tenure or fiscal equity. Instead of preventing arbitrary and ill-considered policies, exit can thus complicate and render more hazardous certain *needed* reforms. Moreover, exit undercuts voice: as long as the capitalists are able to remove their patrimony to a safe place, they will have that much less incentive to raise their voice for the purpose of making a responsible contribution to national problem-solving. Capital mobility and propensity to exit may thus be a major reason for the instability of states in the capitalist periphery (Hirschman, 1981, ch. 11).

Political parties

Two principal propositions were put forward by *EVL* with regard to the dynamics of political parties in a democracy:

1. (1) In a two-party system, the tendency of the parties to move toward the non-ideological centre in order to capture the (allegedly) voluminous middle-of-the-road vote is countered by those party members and militants who are on the parties’ ideological fringes, have ‘nowhere else to go’, but just because of that are maximally motivated to exert influence inside the party, by forceful uses of voice.
2. (2) In a multi-party system, with the ideological distance from one part to the next being presumably shorter than in two-party systems, dissatisfaction with party performance is more likely to lead to exit than in two-party systems; in the latter, voice will play the more important role as switching to the other party requires too big an ideological jump. One inference is that parties in two-party systems may be expected to exhibit more internal divisions, but also more internal democracy and less bureaucratic centralism than parties in multiple-party systems.

The first of these propositions has been strongly supported by events subsequent to the publication of the book. At that time, only the nomination of Barry Goldwater to be the standard bearer of the Republican Party in 1964 could be cited in support. Since then, additional evidence has accumulated: from the nomination by the Democrats of George McGovern to contend the Presidential elections in 1972 to the increasing power of the more radical wing of the Labour party and the ascendancy of Margaret Thatcher within the Conservative party and of Ronald Reagan among the Republicans. The theory that in a two-party system the two parties would increasingly converge toward some middle ground has been amply disconfirmed.

The second proposition on political parties which was deduced from the *EVL* framework has undergone several qualifications. For example, in democracies with old cleavages along ethnic, linguistic, and religious lines, the distances between the several parties rooted in ethnic, etc., identities could actually be wider than that between the parties of two-party systems. Under these conditions, the exitvoice logic would in fact predict that member participation (voice) in parties of multi-party systems would also be vigorous and exit infrequent (Lorwin, 1971; Hirschman, 1981, ch. 9).

A more serious complication is being stressed in a work by S. Kernell still in progress. In two-party systems, exit is a particularly powerful move for dissatisfied members as by casting their vote for the other party they are doubling its impact, something they cannot be sure of in multi-party systems. Hence, in case of disappointment with the performance of one's own party, there could arise a special temptation in two-party systems to switch to the other party so as to *punish* one's own. Such a preference for exit is likely to come to the fore primarily when a party in power is perceived as having seriously mishandled its mandate. Under the circumstances, the prospect of being able to punish that party retrospectively could overcome party loyalty and past ideological commitment. This constellation was an important factor in the sharp defeat of the Democratic ticket in the 1980 Presidential elections in the United States.

The family: marriage and divorce

Modern marriage is one of the simplest illustrations of the exit–voice alternative. When a marriage is in difficulty, the partners can either make an attempt, usually through a great deal of voicing, to reconstruct their relationship or they can divorce. The complexities of the interplay between exit and voice are well in evidence here. Just as the threat of strike in labour-management relations, so is the threat of divorce important in inducing the parties to ‘bargain seriously’; but as exit becomes ever easier and less costly (and perhaps even profitable to one of the parties – see Weitzman, 1985), its availability will undermine voice: rather than being an action of last resort, divorce could become the automatic response to marital difficulty with less and less effort made at communication and reconciliation.

This is exactly what appears to have happened in the United States during the last fifteen years, i.e. since *EVL* stated that ‘the expenditure of time, money and nerves’ necessitated by complicated divorce procedures serves the useful, if unintended purpose of ‘stimulating voice in deteriorating, yet recuperable organizations which would be prematurely destroyed through free exit’ (p. 79). In 1970 California adopted a new ‘no-fault’ law on divorce which spread, though often in attenuated form, to most other states (Weitzman, 1985). The California law drastically altered divorce procedures: instead of requiring proof that one of the parties was guilty of some specific type of behaviour constituting grounds for divorce, the new law permitted divorce when both *or just one* of the two parties asserted that the marriage had irretrievably broken down. The possibility of a unilateral decision, of just ‘walking out’, is symbolic of the way in which the California law undercuts the recourse to voice.

With the new regime, the pendulum has swung quite far in the direction of facilitating exit and of thereby weakening voice. It was of course a reaction to the many abuses of the older fault-based system which required costly and degrading adversarial proceedings, and in effect discriminated against the poor. But the framers of the new legislation probably did not realize the extent to which the earlier obstacles to divorce indirectly encouraged attempts at mending the so easily frayed conjugal relationship and how much the new freedom to exit would torpedo such attempts, with the results that one of every two new marriages now ends in divorce.

The family: adolescent development

This is another family situation for whose analysis a formulation in terms of exit and voice has been found useful (Gilligan, 1986). Adolescent development has often been portrayed as a process through

which the ‘dependent’ child becomes an ‘independent’ adult through progressive ‘detachment’ from the parents. Freud saw this as ‘one of the most significant, but also one of the most painful psychic accomplishments of the pubertal period … a process that alone makes possible the opposition, which is so important for the progress of civilization, between the new generation and the old’ (1905, p. 227). Here is a celebration of exit; Freud's statement neglects a complementary aspect and task of adolescent development which is to maintain and enrich the bond with the older generation through continued, if conflict-ridden, communication. In other words, voice has an important role to play in transforming the adolescent's relationship to the parents. The peculiar poignancy of the adolescent–parents conflict resides in fact in the impossibility of relying *wholly* on voice in resolving it: given the closeness of the relationship, a full accord that would be the outcome of successful voicing risks ending up in incest, as the ‘meeting of minds would suggest a meeting of bodies’ (Gilligan, 1986). It is because of the incest taboo that exit must be part of the solution, but different generations of adolescents are likely to achieve emancipation by practising very different characteristic mixes of exit and voice. Moreover, as Gilligan stresses, the balance of exit and voice differs according to gender. Girls place a greater value than boys on continued attachment to the family, and are therefore less attracted to the masculine ideal of independence-isolation. Hence they experience a greater tension between exit and voice.
With this imaginative use of the exit–voice concept, the outer limits of its sphere of influence may have been reached.

See Also

- Tiebout hypothesis

Bibliography

Barry, B. 1974. Review article: ‘Exit, Voice, and Loyalty’. *British Journal of Political Science* 4, 79–107.

Bender, P. 1981. *Das Ende des ideologischen Zeitalters*. Berlin: Severin und Siedler.

Birch, A.H. 1975. Economic models in political science: the issue of ‘Exit, Voice, and Loyalty’. *British Journal of Political Science* 5, 69–82.

Breneman, D.W. 1983. Where would tuition tax credit take us? Should we agree to go? In *Public Dollars for Private Schools*, ed. T. James and H.M. Levin. Philadelphia: Temple University Press.

Bridge, G. 1977. Citizen choice in public services: voucher systems. In *Alternatives for Delivering Public Services*, ed. E.S. Savas. Boulder, Colorado: Westview.

Burnett, N.R. 1976. Emigration and modern Ireland. Unpublished PhD dissertation, School of Advanced International Studies, Johns Hopkins University.

Eccles, R.G. 1981. The quasifirm in the construction industry. *Journal of Economic Behavior and Organization* 2(4), December, 335–57.

Fainstein, N.I and Fainstein, S.S. 1980. Mobility, community, and participation: the American way out. In *Residential Mobility and Public Policy*, ed. E.G. Moore and W.A.V. Clark. Beverly Hills: Sage.

Finer, S.E. 1974. State-building, state boundaries and border control in the light of the Rokkan-Hirschman model. *Social Science Information* 13(4–5), 79–126.

Freeman, R.B. and Medoff, J.L. 1984. *What Do Unions Do?* New York: Basic Books.

Freud, S. 1905. Three essays on the theory of sexuality. In *Complete Psychological Works*, vol. 7. London: Hogarth, 1953.

Gilligan, C. 1986. Exit–voice dilemmas in adolescent development. In *Development, Democracy and the Art of Trespassing: Essays in Honor of A.O. Hirschman*, ed. A. Foxley et al., Notre Dame, IN: University of Notre Dame Press.

Granovetter, M. 1985. Economic action and social structure: a theory of embeddedness. *American Journal of Sociology* 91, 481–510.

Hirschman, A.O. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press.

Hirschman, A.O. 1977. *The Passions and the Interests: Political Arguments for Capitalism before its Triumph*. Princeton: Princeton University Press.

Hirschman, A.O. 1981. *Essays in Trespassing: Economics to Politics and Beyond*. Cambridge: Cambridge University Press.

Hirschman, A.O. 1982. *Shifting Involvements: Private Interest and Public Action*. Princeton: Princeton University Press.

Hirschman, A.O. 1985. Against parsimony: three easy ways of complicating some categories of economic discourse. *Economics and Philosophy* 1, 7–21.

Huntington, S.P. and Nelson, J.M. 1976. *No Easy Choice: Political Participation in Developing Countries*. Cambridge, MA: Harvard University Press.

Kernell, S. 1987. *Retrospective Voting and Contemporary Macrodemocracy*. Washington, DC: Brookings Institution.

Klein, R. 1980. Models of man and models of policy: reflections on *Exit, Voice and Loyalty* ten years later. *Milbank Memorial Fund Quarterly* 58, 413–29.

Kuhnle, S. 1981. Emigration, democratization, and the rise of the European welfare states. In *Mobilization, Center-Periphery Structures, and Nation-Building*, (a volume in commemoration of Stein Rokkan), ed. P. Torsvik. Bergen: Universitetsforlaget.

Levin, H.M. 1983. Educational choice and the pains of democracy. In *Public Dollars for Private Schools: The Case for Tuition Tax Credits*, ed. T. James and H.M. Levin. Philadelphia: Temple University Press.

Lorwin, V. 1971. Segmented pluralism: ideological cleavages and political cohesion in the smaller European democracies. *Comparative Politics* 3(2), 141–75.

MacDonald, J.S. 1963–4. Agricultural organization, migration and labour militancy in rural Italy. *Economic History Review* 16, 61–75.

O'Donnell, G. 1986. On the convergences of Hirschman's *Exit, Voice and Loyalty* and *Shifting Involvements*. In *Development, Democracy and the Art of Trespassing: Essays in Honor of A.O. Hirschman*, ed. A. Foxley et al., Notre Dame, IN: University of Notre Dame Press.

Rokkan, S. 1975. Dimensions of state formation and nation-building: a possible paradigm for research on variations in Europe. In *The Formation of National States in Western Europe*, ed. C. Tilly. Princeton: Princeton University Press.

Stevens, C.M. 1974. Voice in medical care markets: ‘consumer participation’. *Social Science Information* 13(3), 33–48.

Weitzman, L.J. 1985. *The Divorce Revolution: The Unexpected Social and Economic Consequences for Women and Children in America*. New York: Free Press.

Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

How to cite this article

Hirschman, Albert O. "exit and voice." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000167> doi:10.1057/9780230226203.0524

The New Palgrave Dictionary of Economics Online

expectations

Robert J. Shiller

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The modelling of economic expectations is central to economics. Expectations of future economic conditions can be represented in econometric models by survey data, expectations proxies such as adaptive expectations, expert forecasts, or market expectations. The theory that expectations are rational, that is, optimal forecasts given the model, can be a useful modelling device, but evidence from behavioural economics shows that it has important limitations.

Keywords

adaptive expectations; aspirations; behavioural economics; capital asset pricing model; certainty equivalent; distributed lag; error-learning hypothesis; Euler equations; expectations; expectations proxies; expected utility; extrapolative expectations; forecasting; Iowa Electronic Market; law of iterated projections; market expectations; mathematical expectations; Muth, John F.; Phillips curve; Phillips, A. W.; prediction markets; quadratic expected utility function; rational expectations; rational expectations equilibrium; regressive expectations; stochastic optimal control; subjective probability; survey expectations; transformations of variables; unexpected inflation; variance

Article

Most decisions that economic agents must make involve uncertainty about the future. Thus, any economic model that is intended to be descriptive of human behaviour is likely to involve human expectations about uncertain future economic variables. Areas in economics that involve expectations in fundamental ways include the theories of intertemporal consumption or labour supply decisions, theories of firms' pricing, sales, investment, or inventory decisions, theories of financial markets and money, theories of insurance, and of search behaviour, signalling, agency and bidding. If our purpose is to describe human behaviour, then the study of human expectations is inseparable from the study of the behavioural models in which these expectations are embedded. Only a few general observations can be

discussed here.

Economic expectations, surveys and proxies

Applied econometric research often relies on simple models involving *expectations variables* that represent the expectations of economic agents for some specified economic variables. For example, the total savings of individuals may be related to a variable purporting to measure their expectations for their pension benefits on the date of their retirement, years in the future. Or an individual's decision whether to purchase a long-term or short-term bond may be related to a variable representing expectations as to the course of future short-term interest rates over the life of the long-term bond. Expectations variables included in such models are often referred to as measuring, perhaps imperfectly, some idealized *economic expectations*. What economic expectations actually represent is usually not spelled out. Different people have different perceptions about the outlook for future variables, and so there is an index number problem in reducing their divergent opinions into a single measure. Moreover, when asked for their expectation for some economic variable people may answer that they have no expectation. If pressed, they may hazard a guess. Certainly, most individuals make some economic decisions without making an effort to learn about relevant economic variables. From time to time, circumstances require making difficult or important decisions, and then people may trouble themselves more to find out about economic variables. Economic models that speak of ‘the’ expectation of an economic variable presumably are talking about some average of the expectations of some people and guesses other people would make if pressed, or about averages of the better or worse forecasts of the same people at different times.

The expectations variables used in econometric work to measure economic expectations may be *survey expectations*, representing the average expectation respondents reported on a public opinion survey, or they may be *expectations proxies*, consisting of transformations of other variables that appear to the econometrician to be plausible guesses as to the public expectations. For example, a moving average of lagged inflation rates may serve as an expectations proxy for future inflation.

Expectations surveys commonly take two forms: those that survey individuals representative of the general population and those that sample experts. The former provide measures of expectations that are relevant to decisions, like individual decisions as to how much to save in a given month or whether to put money in a savings account as against corporate bonds, on which decision-makers do not attach great importance. The latter provide measures of expectations that are relevant to decisions, such as firm decisions on whether to market a new product or invest in a new plant, on which decision-makers are likely to spend the resources to obtain informed forecasts.

Day-to-day expectations of the general population

According to Katona (1975), survey research finds that most people can be induced to make a guess as to the direction of change in the near future of major macroeconomic variables, but are reluctant to give quantitative estimates of the extent of the change. The information on which most people base their expectations is fragmentary. Based on decades of survey research on the general public in the United States, Katona concluded that the majority knew whether unemployment had increased or decreased in

the preceding months, whether profits or retail sales had gone up or down, and also whether interest rates had risen or fallen, but did not know how much larger or smaller any of these magnitudes were. The extent of knowledge about macroeconomic variables is generally greater the more important or dramatic the recent changes in these variables, and of course, the more the variable has been emphasized in the mass media.

Since we generally want to incorporate expectations variables in an economic model that describes human behaviour, we are likely to want any variable measuring economic expectations to represent the actual thoughts of individuals *before* they were forced to sit down at a questionnaire and carefully think about how to forecast an economic variable. In modelling, say, income expectations for the purpose of studying the saving decision, we want to get into the individual's frame of mind at the times when saving decisions are made.

When we try to characterize a person's frame of mind at these times, we should recognize that the expectations are likely to differ through time qualitatively as well as quantitatively. For example, an expectation of a future rise in income may become more vividly impressed on individuals' consciousness by some public event that reminds that person of the reasons to expect income to rise. At the same time, the expectation as measured on a survey may be unchanged. Psychologists who study the saving decision have emphasized the importance of changing *aspirations* as distinct from changing expectations.

Individuals who are not thinking at all about economic theories and who are merely confronted with economic variables whose stochastic properties are difficult to comprehend may fall back on simple expectations mechanisms such as that proposed by Fisher (1930). In his, expected inflation is a *distributed lag* or weighted average, with weights that decline linearly with time, of actual inflation. A variation on Fisher's expectation mechanism is *adaptive expectations* (Cagan, 1956) in which expectations are formed as a weighted average, with weights that decline exponentially with time into the past, and that sum to 1, of actual past inflation. The rate at which weights decline might be determined by the rate at which human memory decays. It may be natural to form expectations of a future variable (for example, inflation) by thinking back over the recent past of experience of the variable, and hence such memory decay may result in a distributed lag pattern like that hypothesized by Cagan.

With adaptive expectations, the change in the expectations variable is proportional to the difference between its previous value and the latest value of the variable to be forecasted. This construction resembles that of the *error-learning hypothesis* (Meiselman, 1962). However, in the error learning hypothesis the change in the expectation for a variable at a specific future date is proportional to the error just discovered in the forecast for the variable for today's date.

Alternatives to adaptive expectations are *regressive expectations*, in which variables are expected to return gradually to a fixed level independent of their recent past behaviour (this term has also been used as a synonym for adaptive expectations), and *extrapolative expectations* in which the recent direction of change in the variables is expected to continue (see for example Modigliani and Sutch, 1966.) Any of these expectations mechanisms may be consistent with optimal forecasts of the future variables under certain special circumstances (see Sargent and Wallace, 1973).

We may not want to use such simple models of expectations in periods when individuals may think a great deal about economic theories. During a period of hyperinflation, for example, it is perhaps unlikely that people will form expectations adaptively, since the inflation affects them so noticeably. They may

seek out the opinions of experts at such times.

Expectations of experts

It is often the case that it is much easier for surveyors to find the expectations of randomly sampled individuals than the expert opinions. Expert opinions may be generated only at the time a crucial decision is made, and not when an expert is asked to fill out a questionnaire. Moreover, experts may feel that their time is too valuable to merit attending carefully to a questionnaire.

It is now the case that economic forecasting has become a profession in which practitioners regularly publish their forecasts of macroeconomic variables, and thereby open themselves up to systematic evaluation by outsiders. Usually these forecasts have some basis in econometric models subject to judgmentally introduced ‘add factors’.

Professional forecasters now make available regularly tabulated forecasts of macroeconomic variables for the succeeding few years. The accuracy of these forecasts is now regularly computed by independent evaluators, and this provides a genuine incentive to forecast well. The marketplace will tend to reduce the numbers of those who do not forecast well. Professional forecasts made in organizations, when not made in anticipation of the kind of ‘forecasting race’ judged by outside evaluators, may not be serious individual attempts to predict. Instead, they may be ‘conventional’ forecasts using methods and information that are perceived as having sanction in the organization. Organizations may stipulate what information a forecaster is to use and how the information is to be translated into a forecast. The aim of such sanctions may be to produce uniformity in the organization as to factual premises on which decisions are made, but they may also lead to forecasts that are not accurate. The costs to individuals of violating the assumptions of the organization may be very large relative to the possible benefits of forecasting well.

The distinction between day-to-day expectations of individuals and the expectations of experts may in practice not be an important one. The advantages that experts have, of access to data, understanding of economic theory and use of statistical methods, may confer little advantage in circumstances when the structure of the economy is changing. Then the data may be viewed as of little help, as it is generated by a different model, and statistical analysis also may be of little help. Experts may then fall back on adaptive expectations or other methods of producing guesses like those of ordinary individuals.

Mathematical expectations

When we use the term *mathematical expectations*, we are referring to a probabilistic model, from which we can compute the expectations as first moment conditioned on the information set available to agents. There are of course other candidates to represent economic expectations, for example, other measures of central tendency such as the median or mode, or measures of central tendency applied to transformations of the random variable.

Ultimately, many of economic models that involve mathematical expectations as economic expectations derive from the assumption of maximization of the mathematical expectation of a utility function. The mathematical expectations operator is initially brought into the assumptions of the model because such expected utility maximization is viewed as a good way to represent human behaviour. Expected utility maximization has been shown to follow from some plausible axioms representing an idealization of

‘rational’ human behaviour. But it is only in certain special cases that maximization of expected utility produces simple behavioural relations involving mathematical expectations as ‘economic expectations’ of the kind that many applied econometricians have been using.

Linear utility functions representing risk-neutral agents may give rise to models in which agents care only about the mathematical expectations of variables, as in the models in finance in which the mathematical expectations of returns on various assets are equalized. A quadratic expected utility function may also produce models that depend on mathematical expectations. It is a result of Simon (1956) that if there are no terms of degree higher than two in control variables and exogenous stochastic processes then optimal behaviour depends linearly on a ‘certainty equivalent’ equal to the conditional expected value of future values of the stochastic process, and not on any other characteristics of their conditional distribution. Simon set up a problem in which there was nothing that could be done by the maximizing agent about the variance of the outcome. In contrast, in the capital asset pricing model in finance, a utility function quadratic in wealth (but where there are terms of degree higher than two in control variables and wealth) yields a behavioural relation that involves both a mathematical expectation and a variance matrix of the underlying stochastic variables.

More generally, expected utility function models that are not linear or quadratic will produce Euler-equation type first-order conditions involving the mathematical expectation operator and economic variables. These models can then give rise to relations involving mathematical expectations that may be interpreted as economic expectations. Suppose economic agents at time t maximize subject to a budget constraint an intertemporal utility function $\sum_{k=0}^{\infty}\beta^{k}u(C_{t+k})$ where β is the subjective discount factor, u represents instantaneous utility and C_{t+k} represents consumption k periods after time t. Then the Euler equation implied by agents’ optimization states that, for any liquid asset whose price is P_t at time t that pays no dividend between t and t+1, $P_t = E_t(\beta\frac{u'(C_{t+1})}{u'(C_t)}P_{t+1})$ where E_t denotes mathematical expectation. If we can find some reason to assume that the term $\beta\frac{u'(C_{t+1})}{u'(C_t)}$ can be disregarded, such as by assuming either that the time interval is small and that there is little variation in consumption over this time interval, then the price P_t itself can be interpreted as the mathematical expectation of price P_{t+1} next period.

Many models start from behavioural relations involving mathematical expectations and do not derive these from the hypothesis of expected utility maximization. In these cases, the popularity of mathematical expectations as representations of economic expectations may derive from some intuitively desirable and convenient properties of mathematical expectations, properties that are not shared by other measures of central tendency. The mathematical expectation of the sum of two random variables is equal to the sum of their mathematical expectations whether or not the two variables are independent, a property not shared by the median or mode, even if the variables are independent. If we have a joint distribution of two random variables, x and y, and we define the conditional distribution of x given y, then the mathematical expectation E(x|y) of x in the conditional distribution is a function of y. The *law of iterated projections* states that the mathematical expectation of the mathematical expectation of x, E(E(x|y)) equals the mathematical expectation of x, E(x). In simple terms, this law might be

described as saying that people do not expect to change their expectations. Again, this law does hold in general for the mode or median of *x*. On the other hand, the median has the desirable property that the median of any monotonic transformation of a random variable is the transformation of the median, a property not shared by the mathematical expectation.

Market expectations

In 1907, the statistician Francis Galton did a statistical analysis of a contest in which participants paid, for the possibility of a monetary prize, to play a game in which they guessed the weight of an ox. His conclusion that the average guess was very close to the actual weight of the ox led over the years to a general public appreciation of the idea that markets may predict very well, and hence that market expectations represent optimal forecasts.

Economic theory may, under some conditions, support the idea that financial market prices may represent mathematical expectations conditioned on public information. If we have a security whose value in the near future P_{t+1} unambiguously represents some economic value that is highly variable, then we might conclude that its price today is its mathematical expectation. (If we assume it is variable relative to the potential variability in $\beta\frac{u'(C_{t+1})}{u'(C_t)}$ in the Euler equation, then the price today P_t might be regarded as approximately the mathematical expectation of that economic value.) On the other hand, if the security represents a claim on the distant future, then these assumptions seem problematic, and even the validity of the Euler equation itself becomes questionable (Shiller, 2005).

The Iowa Political Stock Market was created in 1988 by Robert Forsythe, Forrest Nelson and George Neumann at the University of Iowa to allow participants to buy securities that pay one dollar plus the candidate's final vote margin in an election in the near future. The price of such a security may be interpreted under our assumptions as 1 plus the mathematical expectation of the candidate's vote margin. Their market, now renamed the Iowa Electronic Market, also trades securities that pay one dollar if an event occurs, otherwise nothing. Then the price of that security has a possible interpretation as the probability that the event will occur.

In fact, of course, not everyone has the same subjective probability of the event. The differences of opinion may be necessary for a market to function, for it may be only the differences of opinion that make the market interesting to traders. Manski (2006) shows that there is no theoretical support for the idea that the market prices should represent the average (over all market participants) subjective probability of the event, and, according to his assumptions, theory allows us to define only a (rather broad) interval within which this average subjective probability lies. Still, the market prices might turn out to be useful in helping us to judge probabilities of future events (Wolfers and Zitzewitz, 2006).

The Iowa people have claimed that the prices of their securities representing final vote margins have produced remarkably accurate forecasts, better than that generated by public opinion polls of voter intentions (Berg et al., 2008). However, the economics profession has not yet generated many studies that test the interpretation of these new markets as generators of optimal expectations.

Interest in market expectations remains very high, and there has been an explosion of prediction markets. Some are related to universities: the Austrian Electronic Market run by Vienna University; the University of British Columbia Election Stock Market. Others are private companies such as intrade.

com; cityindex.co.uk; igindex.co.uk; tradesports.com; hedgestreet.com; newsfutures.com; and ideosphere.com.

The Economic Derivatives Market was created in 2002 by Deutsche Bank and Goldman Sachs to trade claims that pay out a fixed sum if an economic variable falls in a specified range, using a trading platform created by Longitude, Inc. that allows people to express complex demands for this security (Lange and Economides, 2005). The market is now managed in a partnership between Goldman Sachs and the Chicago Mercantile Exchange. Today US non-farm payrolls, the Institute of Supply Management's Purchasing Managers Index (PMI), weekly initial jobless claims, retail sales, the European harmonized index of consumer prices (HICP), the international trade balance and gross domestic product (GDP) are currently traded. Preliminary studies, with only limited data available as yet, support the notion that these markets yield useful forecasts (Gurkaynak and Wolfers, 2006).

In 2006 the Chicago Mercantile Exchange with MacroMarkets LLC (a company I helped found) created futures and options markets that are cash-settled based on the Standard & Poor's/Case-Shiller Home Price Indices, and has plans to create such markets based on a commercial property price index. These markets are beginning to offer market expectations for real estate prices.

Other examples of potential new markets for economic variables are described in Shiller (2003). The value of these markets for generating market expectations or optimal forecasts will not be in until many more years’ data are at hand.

Modelling rational expectations

Even if we can use markets or surveys to measure expectations, if we are to understand their movements through time we need to model their determination. The idea that expectations are rational has been an important modelling device.

Why does the idea sound plausible that economic expectations of future inflation may be proxied fairly well by adaptive expectations or other distributed lag on actual inflation? Is it just because of the theory of psychologists that human memory decays gradually through time and the notion that casual guesses of future inflation would correspond to recent memories of inflation? Perhaps it is instead that a distributed lag on inflation is not a bad way to forecast inflation.

Suppose people were asked on a monthly basis to forecast the rate of increase of the price of some seasonal commodity, let us say, fresh tomatoes. Certainly, many of them would be aware that fresh tomatoes are more expensive in the winter, when they must be grown in hothouses or brought in from greater distances. Not all people would know this, and many who did know about the seasonality in price would not know its magnitude. But certainly a distributed lag with smoothly declining coefficients on actual tomato price changes is not what we would think of first to model their expectations. Such a distributed lag would imply some seasonality in expectations but would also generally imply that people misforecast the month of highest price.

If there is any doubt as to the value of simple expectations proxies for modelling the expectations of tomato consumers, there is certainly no doubt that it would be inappropriate to use such proxies to model the expectations of tomato producers. Some producers specialize in producing hothouse tomatoes, and time their production for the winter months. Surely *they* know in which month prices are higher, and by how much they tend to be higher.

How then should we build a model that describes the supply of tomatoes over time? Since tomatoes

must be planted months in advance of the anticipated demand for them, the supply function for tomatoes must depend on expectations formed at this time by producers, as well as on seasonal factors affecting the cost of production. We might then model the supply of tomatoes by finding a good way to predict the price of tomatoes (using, say, seasonal dummies and other information) and substituting the prediction in place of the expectation in the supply function. The result would be a *rational expectations model*.

One could use such a model to predict the supply response to some variable that has been found to predict price. If, let us say, we found that bad weather in Mexico, which might later reduce supply of winter tomatoes to the United States, tended to cause the seasonal peak in tomato prices in the United States to be higher than usual, then we might in these circumstances forecast the supply of domestic tomatoes in the United States to be higher than usual. A rational expectations model would produce such a forecast if the model was based on an empirical forecasting relation for price that used the weather variable as an explanatory variable.

Of course, for the purpose of forecasting supply we might also have used the ‘naive’ approach of estimating a forecasting equation directly for supply (without the intermediate step of developing a forecasting equation for price) depending on such variables as earlier weather and on seasonal dummies. Such a method may also satisfactorily predict supply, but it might not do as well since it would not make use of the information in economic theory that weather affects supply only through its effect on rationally expected price. For example, suppose we had a long time series of data on various weather variables and prices but only a few observations on quantities supplied. We could not include all the weather variables directly in a ‘naive’ forecasting equation for supply, since we would thereby exhaust degrees of freedom. But we could first find how these weather variables predict price and then use a single price expectations variable to predict supply.

Rational expectations in equilibrium models

The above example of the use of a rational expectations model was very special in that the model consisted only of a single equation relating supply to an earlier expectation of price. Moreover, the equation was used only to forecast supply in a situation where we expect the correlations observed in the past with explanatory variables to continue. Very often we wish instead to predict the effect on supply of some change in government policy or other structural change that is expected to change the correlations with other variables.

Suppose for example we wish to know the effect on the seasonal pattern of tomato supply in the United States of a government policy of blocking the further international trade of tomatoes. Here, the naive forecasting model that related tomato supply to weather and seasonal dummy variables would be of no value. An estimated rational expectations model relating supply to expected price might still be of value. We need to model only the determination of expected price.

Suppose we then also estimate (using a sample period in which some tomatoes were imported) a domestic demand function for tomatoes, relating, say, total quantities demanded in the United States to contemporaneous price. Consider, then, a two-equation model consisting of this demand equation and the rational expectations domestic supply equation for tomatoes described above. In the sample period domestic demand did not equal domestic supply because of imports. After the policy change the domestic supply and demand will be equal. Can we now predict how the seasonal pattern of quantities

supplied may be changed by the government policy?

To answer this question, we cannot just solve the two-equation model with the two endogenous variables, quantity and price, because both price and expectations appear separately in the model. However, the expectation of price, if it is a rational expectation, ought to be determined by the very model in which it appears. How can we find the rational expectation of price?

One approach is first to guess a function relating expected price to the exogenous variables in the model, in our example, the seasonal dummies and weather variable. If one substitutes this guess into the model in place of the expected price, one then has an ordinary simultaneous equation model in price and quantity in terms of exogenous variables. However, unless one made a lucky guess, one would then find that the model that resulted from the guess was inconsistent with the guess, in that the model implies that a different way of forecasting price is optimal, given the expectations function.

What we need to find is an equation defining the expectation of price which, on substituting into the model, produces a model in which that equation gives the optimal forecast of price. Muth (1961) showed how this can be done if the simultaneous equations model is linear and if rational expectations are defined as mathematical expectations conditioned on variables in the model that are in the public's information set.

Using such a solution method, we might find how the seasonality of both quantities and price will be changed under the new government policy. In this simple example, doing this would seem to be preferable to using a model with an expectations proxy for price that did not take into account how the changing seasonal pattern of price would change the way expectations are formed.

Rational expectations models, stochastic processes and optimal control

The advent of rational expectations in econometric models has marked a revolution in economic thinking that is comparable in the magnitude of its impact on the economics profession to the Keynesian revolution in the mid-20th century.

Muth (1961) and those who carried on the rational expectations literature have borrowed heavily from another literature that was once outside economics, namely the theory of stochastic processes and optimal control. What is substantially new about the rational expectations models derives ultimately from these theories, which were developed for the most part since 1950. The implications of these theories were so profound that it was inevitable that they should make themselves felt in economics, just as they have in many fields in science and engineering.

The rational expectations revolution is not primarily the result of any failure of conventional econometric models to forecast well, as some (for example, Lucas and Sargent, 1981) have argued. It is true that initial optimism for the forecasting ability of such models has been tempered by experience, but it has not been established that shortcomings of the expectations modelling methods has been the major fault. It has certainly not been established empirically that rational expectations models can predict better.

Interest by economists in optimal control and the theory of stochastic processes was initially expressed in their efforts to apply control methods to existing econometric models, to achieve their stabilization. However, the optimal control of conventional ‘Keynesian’ econometric models involving expectations proxies like adaptive expectations has never became as influential in the profession as its developers had hoped. Perhaps the general profession thought that the methods of control were too refined for the crude

models that they were applied to. More concern was felt for improving the models themselves.
The idea that optimal control might be applied to conventional Keynesian econometric models did have the effect of generating hopes that the macroeconomy might be controlled very well, ‘fine tuned’ so to speak, and thus great importance was placed on the structural stability of these models. Much of the polemics against ‘Keynesian’ economics waged by those who promoted rational expectations models as alternatives was really directed against these efforts to apply optimal control systematically to the models (see for example Sargent and Wallace, 1981). A central criticism of these models was their heavy reliance on crude expectations proxies such as adaptive expectations.
The rational expectations models applied stochastic optimal control theory by assuming in effect that human behaviour could be modelled as if everyone all along had been applying the principles of optimal control to their own economic decisions. Given the natural interest of economists in rational behaviour, the optimal filtering and extrapolation that was developed as part of the theory of stochastic processes would naturally be used in modelling how individuals forecast.
Of course, there are strict limits to the extent to which people's actual behaviour can be described in such terms. Rational expectations models thus often sacrifice descriptive accuracy with the hope that the models would exhibit stability in the presence of interventions of the kind envisioned by makers of government macroeconomic policy. The models may not be generally well suited to forecasting when the policy regime is unchanged. They are most appropriately considered as policy analysis tools.

Criticisms of rational expectations models

The simple supply and demand model for tomatoes described above was chosen as an ideal example of the application of rational expectations models. In this example there is substantial seasonal variation in price, which ought to be forecastable. Moreover, as the model was set up, only producers’ expectations entered the model, and producers are far more likely than others to have rational expectations about price. But few of the applications of the theory of rational expectations have been to such ideal examples.
The best-known application of rational expectations models has been to an interpretation of the observed relation between unemployment and inflation. A.W. Phillips (1958) noted a negative relation between the unemployment rate and the wage inflation rate in the United Kingdom between 1861 and 1957. A similar relation was found also in the United States for much of the same sample period. Since then, the negative relation has broken down. Lucas (1976) and Sargent and Wallace (1973) offered interpretations of the Phillips relation and its subsequent breakdown. In its simplest terms, this interpretation asserts that there may be a stable relation between unemployment and *unexpected* inflation. Unexpected inflation may cause job seekers to misperceive the real value of wage offers they have received, and thus to accept offers that they would not have accepted if they had known the true real wage they were getting. By accepting these jobs, they lower the unemployment rate. In the period Phillips studied the price level might have been well-enough approximated by a random walk that actual inflation may have approximately equalled unexpected inflation. Since then, when inflation has become much more serially correlated, actual and expected inflation may have diverged widely.
In its general idea, the Lucas–Sargent–Wallace theory of the Phillips curve sounds like an appealing possibility. The question for econometric testing of the theory is whether we want to assume that expectations of unemployed workers are fully rational.

The tests Sargent (1976) made of the model are illustrative of the manner in which rational expectations models are often tested. Sargent tested whether the model holds under the assumption that unemployed workers are making optimal use in their forecasts of inflation of current and lagged values of the real government surplus, real and money government expenditures, the price level, the money supply, and a wage index. It is commonplace today in the rational expectations literature to see similar extravagant assumptions about the information sets of ordinary individuals.

The most basic criticism of many rational expectations models is that they make implausible claims for individual economic agents' ability and willingness to compute. But the criticism of these models goes beyond that: see for example Friedman (1979), Tobin (1980).

The rational expectations models assume that economic agents behave as if they know the structure of the economy so that they can compute the optimal forecasts that represent their expectations. But the structure of the economy is always changing, as technology, tastes and government interventions change. These changes themselves vary qualitatively from time to time, and so it may not be possible for economic agents to group instances in such a way as to allow dealing with the changes in statistical terms. If these changes occur frequently relative to the speed at which people can figure out the economy, it may never be appropriate to assume that their forecasts are optimal forecasts.

In most rational expectations models, the behaviour of the economic variables that individual economic agents must forecast is itself affected by the way the economic agents form expectations. This fact was noted above in connection with our efforts to solve the supply and demand rational expectations model for tomatoes. Thus, if economic agents learn something about how to forecast an economic variable, the random properties of the economic variable may change in consequence. A rational expectations equilibrium is achieved only when people have adopted a way of forecasting that is consistent with the implications for the economy of their own way of forecasting. How do they find such a way of forecasting? Achievement of a rational expectations equilibrium might take place as a consequence of a long iterative process, each step representing the learning by economic agents of how to forecast in the preceding step, and thereby necessitating the next step of learning anew how to forecast. In models that are more complicated than the simple supply and demand models, for example, models of the entire macroeconomy, the time required for each step may need to be enormous. The problem of convergence of forecasting methods to a rational expectations equilibrium recalls the problem in mathematical economics of the convergence of a price vector to Walrasian equilibrium. However, the former problem has received much less attention. Moreover, convergence may well be orders of magnitude slower in the former. It would appear likely, given the complexity of the macroeconomy, that economic agents learn very slowly about how to forecast given the present structure of the economy. Each step in the iteration requires sifting through large amounts of data and learning how these are related statistically.

The behavioural economics revolution, still very much under way since its beginnings around 1980, can be described as groping for alternatives to rational expectations models, alternatives that have some structure from research in psychology and that do not impose unrealistic complexity on individual decision-making. George Akerlof and Janet Yellen (1985) have argued for 'near-rational expectations' models. Richard Thaler (1991) has argued that we must turn to something he calls 'quasi-rational economics'. Roman Frydman and Michael Goldberg (2007) have argued that we must work towards something they call 'imperfect knowledge economics'. These are important beginnings, though today none of the alternatives has yet won widespread acceptance among economists.

Despite the criticisms, rational expectations models may well be useful for some applications when

compared with alternative models based on expectations proxies. As regards the assumptions in the models for the ability and willingness of economic agents to store and process information, there is no alternative for model builders to that of judging for plausibility on a case-by-case basis.

See Also

- adaptive expectations
- behavioural economics and game theory
- certainty equivalence
- prediction markets
- rational expectations

Bibliography

Akerlof, G.A. and Yellen, J.L. 1985. A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100, 823–38.

Berg, J., Forsythe, R. Nelson, F. and Reitz, T. 2008. Results from a dozen years of election futures markets research. In *Handbook of Experimental Economic Results*, ed. C. Plott and V. Smith. Amsterdam: North-Holland.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Friedman, B.M. 1979. Optimal expectations and the extreme information assumptions of ‘rational’ expectations models. *Journal of Monetary Economics* 5, 23–41.

Frydman, R. and Goldberg, M.D. 2007. *Imperfect Knowledge Economics: Exchange Rates and Risk*. Princeton: Princeton University Press.

Galton, F. 1907. Vox populi. *Nature* 75, 450–1.

Gurkaynak, R. and Wolfers, J. 2006. Macroeconomic derivatives: an initial analysis of market-based macro forecasts, uncertainty, and risk. Working Paper No. 11929. Cambridge, MA: NBER.

Katona, G. 1975. *Psychological Economics*. New York: Elsevier.

Lange, J. and Economides, N. 2005. A parimutuel market microstructure for contingent claims.

European Financial Management 11, 25–49.

Lucas, R.E. 1976. Econometric policy evaluation: a critique. *Carnegie-Rochester Conference Series on Public Policy* 1, 19–46.

Lucas, R.E., Jr. and Sargent, T.J. 1981. After Keynesian macroeconomics. In *Rational Expectations and Econometric Practice*, ed. R.E. Lucas, Jr. and T.J. Sargent. Minneapolis: University of Minnesota Press.

Manski, C. 2006. Interpreting the predictions of prediction markets. *Economics Letters* 91, 425–9.

Meiselman, D. 1962. *The Term Structure of Interest Rates*. Englewood Cliffs, NJ: Prentice Hall.

Modigliani, F. and Sutch, R. 1966. Innovations in interest rate policy. *American Economic Review, Papers and Proceedings* 56, 178–97.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25, 283–99.

Sargent, T.J. 1976. A classical macroeconomic model for the United States. *Journal of Political Economy* 84, 207–37.

Sargent, T.J. and Wallace, N. 1973. Rational expectations and the dynamics of hyperinflation. *International Economic Review* 14, 328–50.

Sargent, T.J. and Wallace, N. 1981. ‘Rational’ expectations, the optimal monetary instrument, and the optimal money supply rule. In *Rational Expectations and Econometric Practice*, ed. R.E. Lucas and T.J. Sargent. Minneapolis: University of Minnesota Press.

Shiller, R.J. 2003. *The New Financial Order: Risk in the 21st Century*. Princeton: Princeton University Press.

Shiller, R.J. 2005. *Irrational Exuberance*, 2nd edn. Princeton: Princeton University Press.

Simon, H.A. 1956. Dynamic programming under uncertainty with a quadratic objective function. *Econometrica* 24, 74–81.

Thaler, R.H. 1991. *Quasi-Rational Economics*. New York: Russell Sage Foundation.

Tobin, J. 1980. *Asset Accumulation and Economic Activity*. Yrjo Jahnsson Lecture. Oxford: Basil

Blackwell.

Wolfers, J. and Zitzewitz, E. 2004. Prediction markets. *Journal of Economic Perspectives* 18(2), 107–26.

Wolfers, J. and Zitzewitz, E. 2006. Interpreting prediction markets prices as probabilities. Working paper, Rodney White Center for Financial Research, Wharton School, University of Pennsylvania.

How to cite this article

Shiller, Robert J. "expectations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000173> doi:10.1057/9780230226203.0525

The New Palgrave Dictionary of Economics Online

expected utility hypothesis

Mark J. Machina
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The expected utility hypothesis – that is, the hypothesis that individuals evaluate uncertain prospects according to their expected level of ‘satisfaction’ or ‘utility’ – is the predominant descriptive and normative model of choice under uncertainty in economics. It provides the analytical underpinnings for the economic theory of risk-bearing, including its applications to insurance and financial decisions, and has been formally axiomatized under conditions of both objective (probabilistic) and subjective (event-based) uncertainty. In spite of evidence that individuals may systematically depart from its predictions, and the development of alternative models, expected utility remains the leading model of economic choice under uncertainty.

Keywords

Arrow–Pratt index of absolute risk aversion; Bernoulli, D.; Bernoulli, N.; Cobb–Douglas functions; comparative likelihood; consumer theory; Cramer, G.; environmental economics; expected utility hypothesis; first-order stochastic dominance preference; increasing risk; Independence Axiom; inequality (measurement); international trade; lotteries; Malinvaud, E.; Marschak, J.; Menger, K.; objective vs. subjective uncertainty; ordinal revolution; preference functions; preference orderings; probability; risk; risk aversion; St Petersburg Paradox; stochastic dominance; subjective probability; sure-thing principle; transitivity; uncertainty; von Neumann–Morgenstern utility function

Article

The expected utility hypothesis is the predominant descriptive and prescriptive theory of individual choice under conditions of risk or uncertainty.

The expected utility hypothesis of behaviour towards risk is the hypothesis that the individual possesses (or acts as if possessing) a ‘von Neumann–Morgenstern utility function’ $U(\cdot)$ or ‘von Neumann–Morgenstern utility index’ $\{U_i\}$ defined over some set $\mathcal{X}$ of alternative possible outcomes, and when faced with alternative risky prospects or ‘lotteries’ over these outcomes, will choose the prospect that maximizes the expected value of $U(\cdot)$ or $\{U_i\}$. Since the outcomes could be alternative wealth levels, multidimensional commodity bundles, time streams of consumption, or even non-numerical consequences (such as a trip to Paris), this approach can be applied to a tremendous variety of situations, and most theoretical research in the economics of uncertainty, as well as virtually all applied work in the field (for example, insurance or investment decisions) is undertaken in the expected utility framework.

As a branch of modern consumer theory (for example, Debreu, 1959, ch. 4), the expected utility model proceeds by specifying a set of objects of choice and assuming that the individual possesses a preference ordering over these objects which may be represented by a real-valued maximand or ‘preference function’ $V(\cdot)$, in the sense that one object is preferred to another if and only if it is assigned a higher value by this preference function. However, the expected utility model differs from the theory of choice over non-stochastic commodity bundles in two important respects. The first is that, since it is a theory of choice under uncertainty, the objects of choice are not deterministic outcomes but rather uncertain prospects. The second difference is that, unlike in the non-stochastic case, the expected utility model imposes a very specific restriction on the functional form of the preference function $V(\cdot)$.

The formal representation of the objects of choice, and hence of the expected utility preference function, depends upon the set of possible outcomes. When the outcome set

$\mathcal{X} = \{x_1, \ldots, x_n\}$ is finite, we can represent any probability distribution over this set by its vector of probabilities $\boldsymbol{P} = (p_1, \ldots, p_n)$ (where $p_i = \text{prob}(x_i)$) and the expected utility preference function takes the form

$$V(\boldsymbol{P}) = V(p_1, \ldots, p_n) \equiv \sum U_i p_i.$$

When the outcome set consists of the real line or some interval subset of it, probability distributions can be represented by their cumulative distribution functions $F(\cdot)$ (where $F(x) = \text{prob}(\tilde{x} \leq x)$), and the expected utility preference function takes the form $V(F) \equiv \int U(x)\,dF(x)$ (or $\int U(x) f(x)\,dx$ when $F(\cdot)$ possesses a density function $f(\cdot)$). When the outcomes are commodity bundles of the form $(z_1,\ldots,z_m)$, cumulative distribution functions are multivariate, and the preference function takes the form $\int\ldots\int U(z_1, \ldots, z_m)\,dF(z_1, \ldots, z_m)$. The expected utility model derives its name from the fact that in each case the preference function consists of the mathematical expectation of the von Neumann–Morgenstern utility function $U(\cdot)$, $U(\cdot,\ldots,\cdot)$ or utility index $\{U_i\}$ with respect to the probability distribution $F(\cdot)$, $F(\cdot,\ldots,\cdot)$ or $\boldsymbol{P}$.

Mathematically, the hypothesis that the preference function $V(\cdot)$ takes the form of a statistical expectation is equivalent to the condition that it be 'linear in the probabilities', that is, either a weighted sum of the components of $\boldsymbol{P}$ (i.e. $\sum U_i p_i$) or else a weighted integral of the functions $F(\cdot)$ or $f(\cdot)$ ($\int U(x)\,dF(x)$ or $\int U(x) f(x)\,dx$). Although this still allows for a wide variety of attitudes towards risk depending upon the shape of the utility function $U(\cdot)$ or utility index $\{U_i\}$, the restriction that $V(\cdot)$ be linear in the probabilities is the primary empirical feature of the expected utility model, and provides the basis for many of its observable implications and predictions.

It is important to distinguish between the preference function $V(\cdot)$ and the von Neumann–Morgenstern utility function $U(\cdot)$ (or index $\{U_i\}$) of an expected utility maximizer, in particular with regard to the prevalent though mistaken belief that expected utility preferences are somehow 'cardinal' in a sense not exhibited by preferences over non-stochastic commodity bundles. As with any real-valued representation of a preference ordering, an expected utility *preference function* $V(\cdot)$ is 'ordinal' in that it may be subject to any increasing transformation without affecting the validity of the representation – thus, the preference functions $\int U(x)\,dF(x)$ and $[\int U(x)\,dF(x)]^3$ represent identical risk preferences. On the other hand, the von Neumann–Morgenstern utility function $U(\cdot)$ is 'cardinal' in the sense that a different utility function $U^*(\cdot)$ will generate an ordinally equivalent preference function $V^*(F) \equiv \int U^*(x)\,dF(x)$ if and only if it satisfies the cardinal relationship $U^*(x) \equiv a \cdot U(x) + b$ for some $a>0$ (in which case $V^*(F) \equiv a \cdot V(F) + b$. However, the same distinction holds in the theory of preferences over non-stochastic commodity bundles: the Cobb–Douglas preference function $\alpha \cdot \ln(z_1) + \beta \cdot \ln(z_2) + \gamma \cdot \ln(z_3)$ (written here in its additive form) can be subject to any increasing transformation and is clearly ordinal, even though a vector of parameters $(\alpha^*, \beta^*, \gamma^*)$ will generate an ordinally equivalent additive form $\alpha^* \cdot \ln(z_1) + \beta^* \cdot \ln(z_2) + \gamma^* \cdot \ln(z_3)$ if and only if it satisfies the cardinal relationship $(\alpha^*, \beta^*, \gamma^*) = \lambda \cdot (\alpha, \beta, \gamma)$ for some $\lambda > 0$.

In the case of a simple outcome set of the form $\{x_1,x_2,x_3\}$, it is possible to graphically illustrate the 'linearity in the probabilities' property of expected utility preferences. Since every probability distribution (p_1,p_2,p_3) over these outcomes must satisfy $p_1 + p_2 + p_3 = 1$, we may represent such distributions by points in the unit triangle in the (p_1,p_3) plane, with p_2 given by $p_2 = 1 - p_1 - p_3$ (Figures 1 and 2). Since they represent the loci of solutions to the equations

$$U_1 p_1 + U_2 p_2 + U_3 p_3 = U_2 - [U_2 - U_1] \cdot p_1 + [U_3 - U_2] \cdot p_3 = \text{constant}$$

Figure 1
Expected utility indifference curves

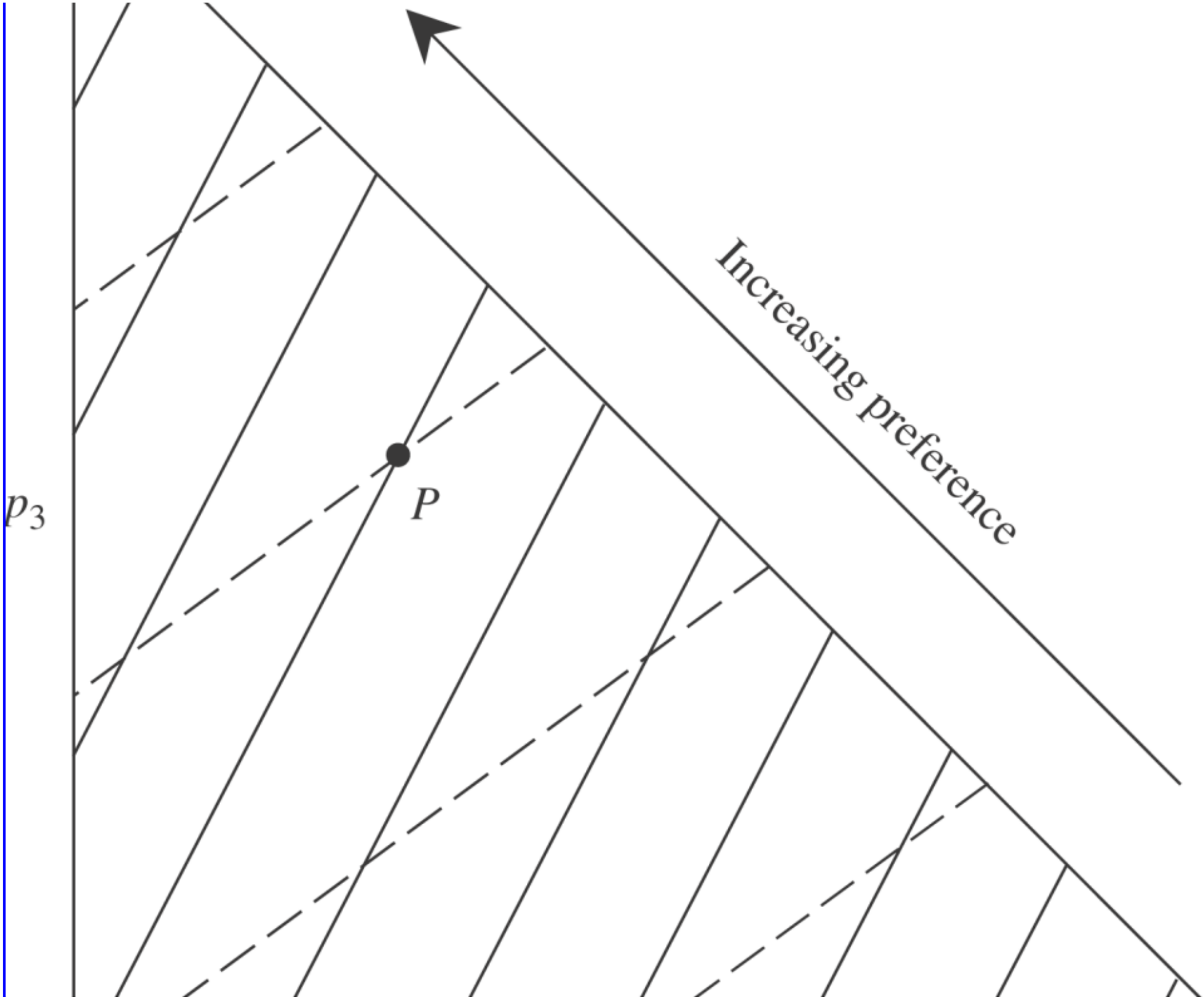
Increasing preference
P
p_3

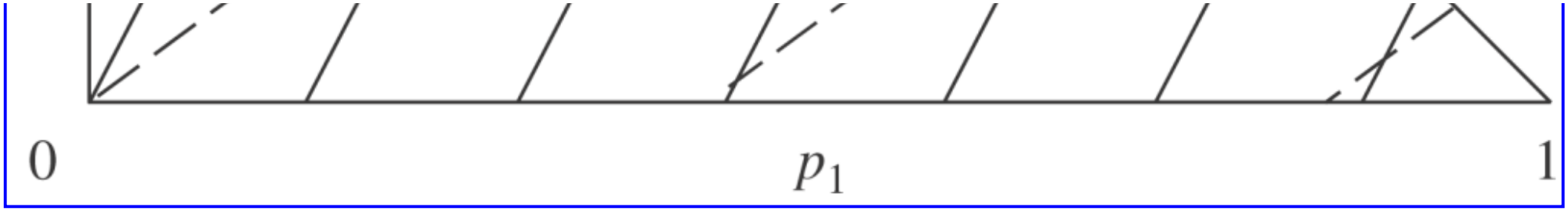

Figure 2
Non-expected utility indifference curves

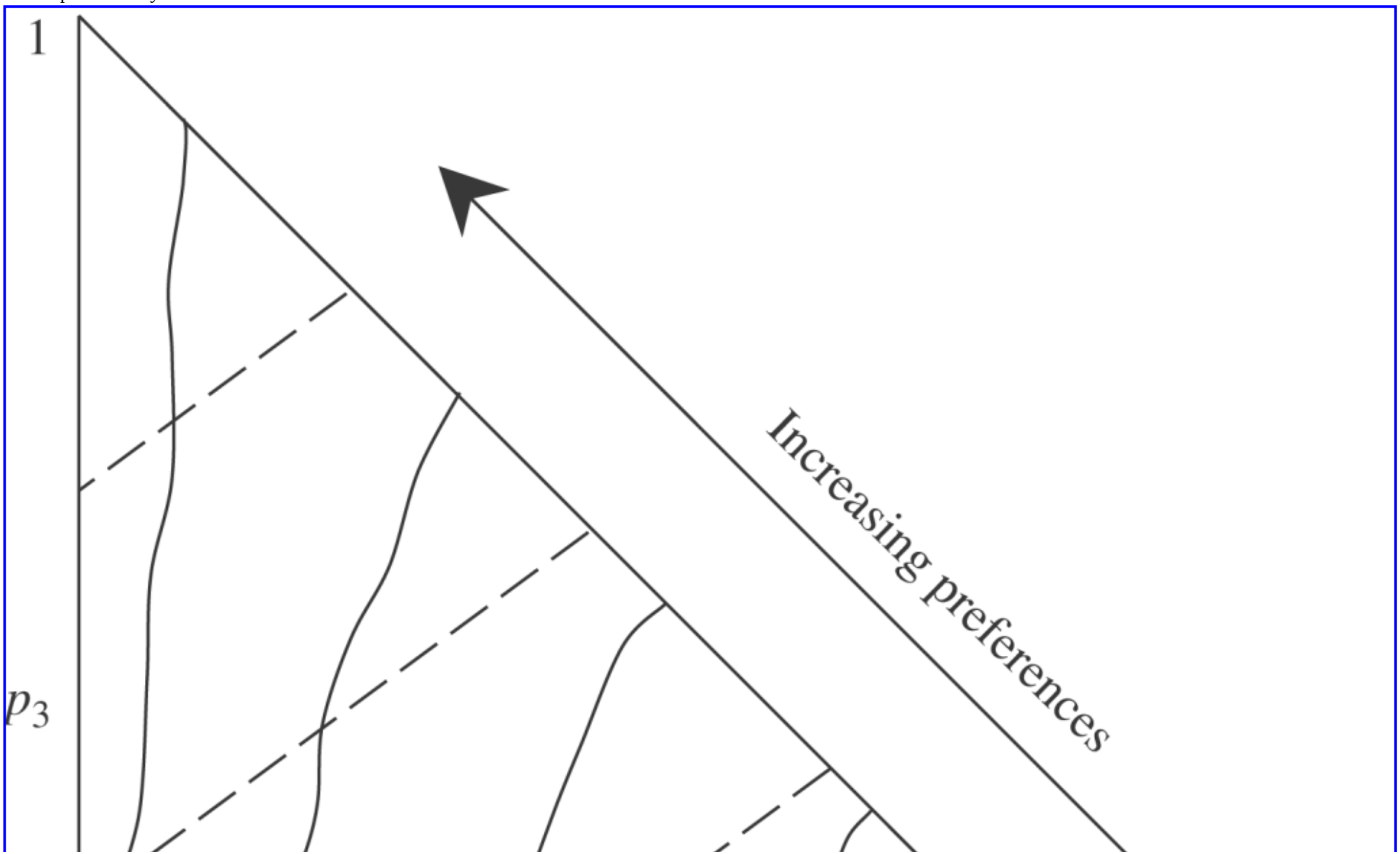

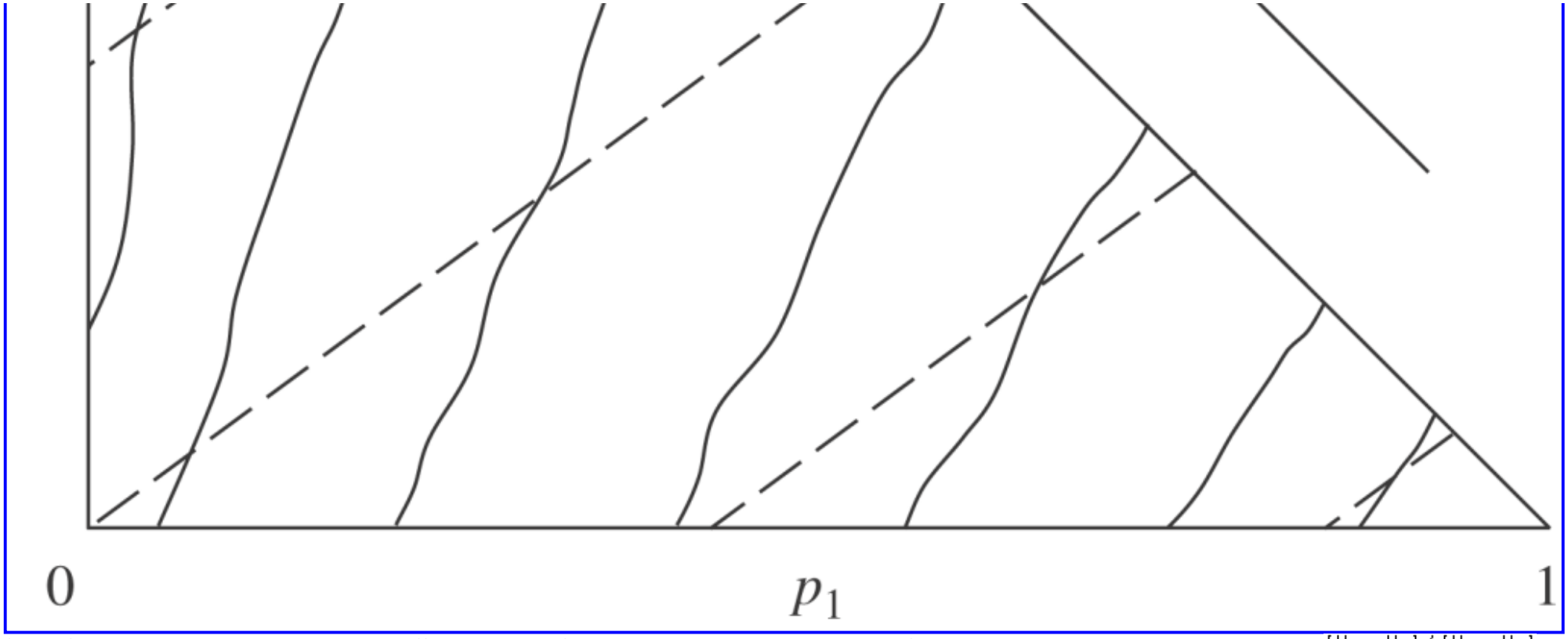

for the fixed utility indices $\{U_1,U_2,U_3\}$, the indifference curves of an expected utility maximizer consist of parallel straight lines in the triangle, with slope $[U_2 - U_1] / [U_3 - U_2]$, as illustrated by the solid lines in Figure 1. Indifference curves which do *not* satisfy the expected utility hypothesis (that is, are not linear in the probabilities) are illustrated by the solid curves in Figure 2.

When the outcomes consist of different wealth levels $x_1<x_2<x_3$, this diagram can be used to illustrate other possible features of an expected utility maximizer's attitudes towards risk. On the principle that more wealth is better, it is typically postulated that any change in a distribution (p_1, p_2, p_3) which increases p_3 at the expense of p_2, increases p_2 at the expense of p_1, or both, will be preferred: this property is known as ‘first-order stochastic dominance preference’. Since such shifts of probability mass are represented by north, west, or north-west movements in the diagram, first-order stochastic dominance preference is equivalent to the condition that indifference curves are upward sloping, with more preferred indifference curves lying to the north-west. Algebraically, this is equivalent to the condition $U_1<U_2<U_3$.

Another widely (though not universally) hypothesized aspect of attitudes towards risk is that of ‘risk aversion’ (for example, Arrow, 1974, ch. 3; Pratt, 1964). To illustrate this property, consider the dashed lines in Figure 1, which represent loci of solutions to the equations

$$x_1 p_1 + x_2 p_2 + x_3 p_3 = x_2 - [x_2 - x_1] \cdot p_1 + [x_3 - x_2] \cdot p_3 = \text{constant}$$

and hence may be termed ‘iso-expected *value* loci’. Since north-east movements along any of these loci consist of increasing the tail probabilities p_1 and p_3 at the expense of the middle probability p_2 in a manner which preserves the mean of the distribution, they correspond to what are termed ‘mean-preserving increases in risk’ (Rothschild and Stiglitz, 1970; 1971). An individual is said to be ‘risk averse’ if such increases in risk always lead to less preferred indifference curves, which is equivalent to the graphical condition that the indifference curves be steeper than the iso-expected value loci. Since the slope of the latter is given by $[x_2 - x_1] / [x_3 - x_2]$, this is equivalent to the algebraic condition that

$[U_2 - U_1] / [x_2 - x_1] > [U_3 - U_2] / [x_3 - x_2]$. Conversely, individuals who *prefer* mean-preserving increases in risk are termed 'risk loving': such individuals' indifference curves will be flatter than the iso-expected value loci, and their utility indices will satisfy $[U_2 - U_1] / [x_2 - x_1] < [U_3 - U_2] / [x_3 - x_2]$.

Note finally that the indifference map in Figure 1 indicates that the lottery ***P*** is indifferent to the origin, which represents the degenerate lottery yielding x_2 with certainty. In such a case the amount x_2 is said to be the 'certainty equivalent' of the lottery ***P***. The fact that the origin lies on a lower iso-expected value locus than ***P*** reflects a general property of risk-averse preferences, namely, that the certainty equivalent of any lottery will always be less than its mean. (For risk lovers, the opposite is the case.)

When the outcomes are elements of the real line, it is possible to represent the above (as well as other) aspects of preferences in terms of the shape of the von Neumann–Morgenstern utility function $U(\cdot)$, as seen in Figures 3 and 4. In each figure, consider the lottery which assigns the probabilities 2/3:1/3 to the outcome levels $x' : x''$. The expected value of this lottery, $\bar{x} = 2/3 \cdot x' + 1/3 \cdot x''$, lies between these two values, two-thirds of the way towards x'. The expected *utility* of this lottery, $\bar{u} = 2/3 \cdot U(x') + 1/3 \cdot U(x'')$ lies between $U(x')$ and $U(x'')$ on the vertical axis, two-thirds of the way towards $U(x')$. The point $(\bar{x}, \bar{u})$ thus lies on the line segment connecting the points $(x', U(x'))$ and $(x'', U(x''))$, two-thirds of the way towards the former. In each figure, the certainty equivalent of this lottery is given by the sure outcome c that also yields a utility level of $\bar{u}$.

Figure 3

Von Neumann-Morgenstern utility function of a risk averse individual

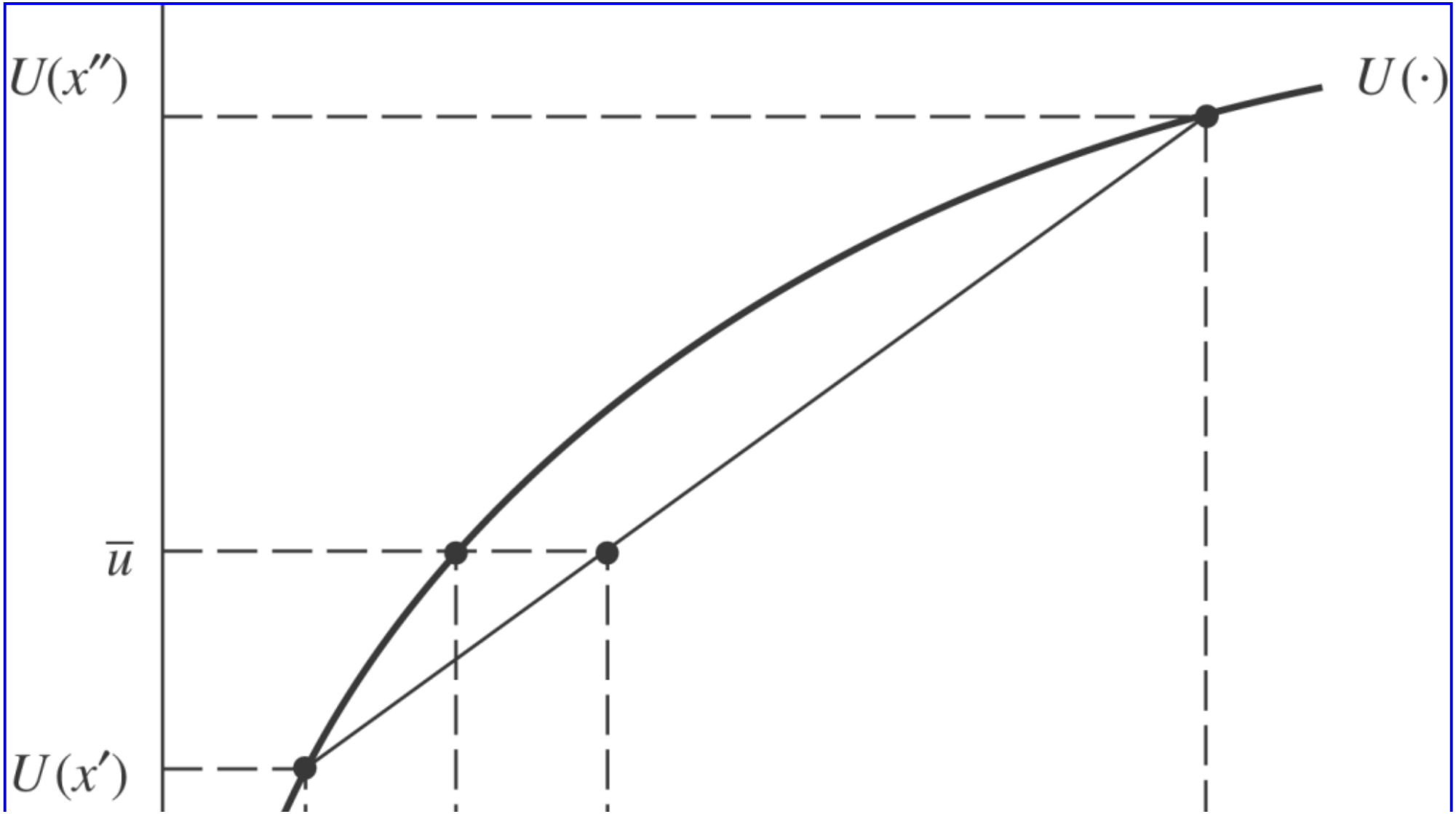

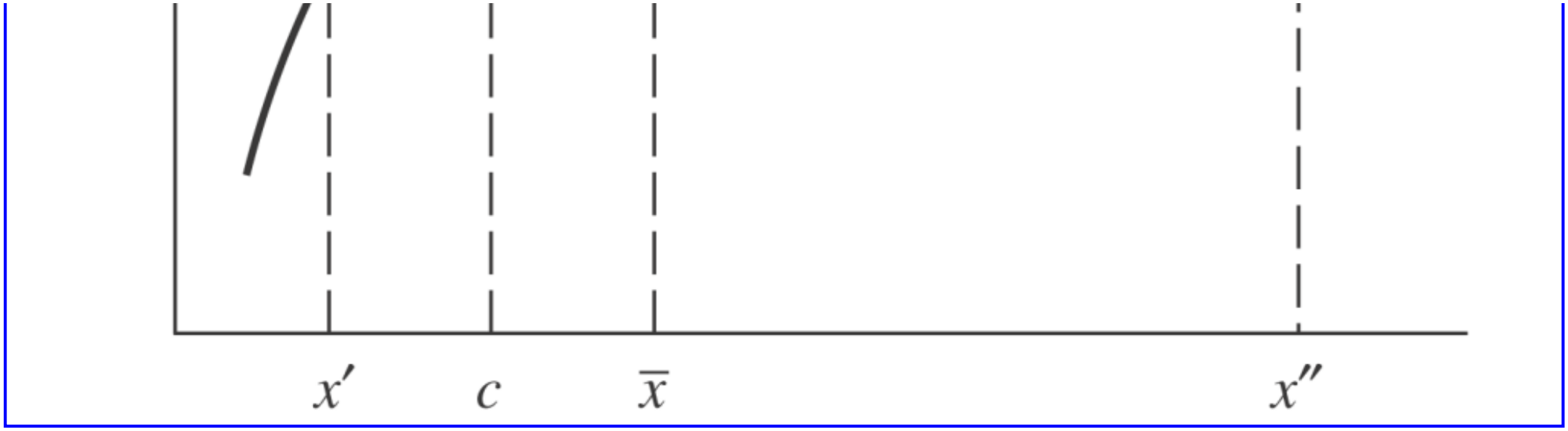

Figure 4
von Neumann–Morgenstern utility function of a risk loving individual

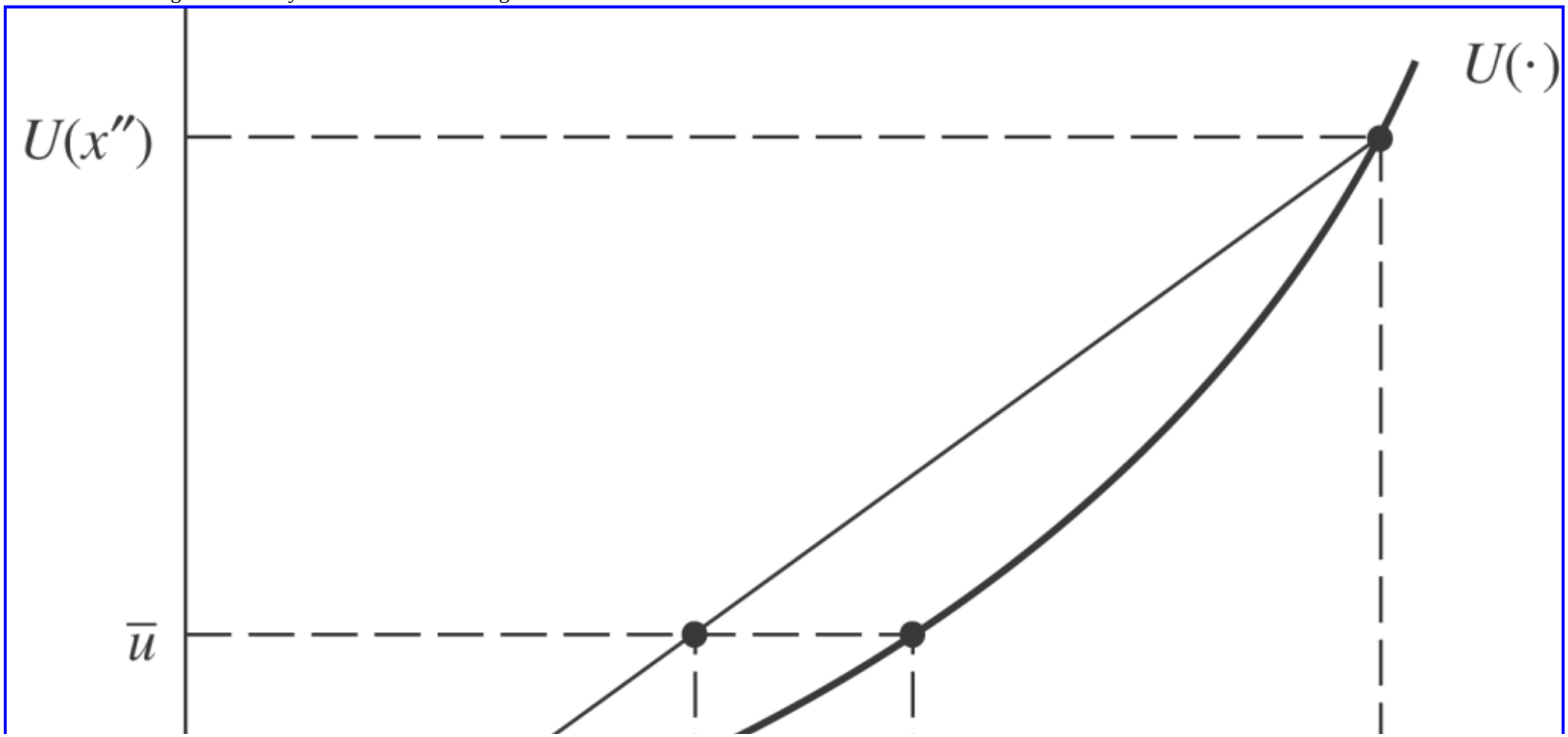

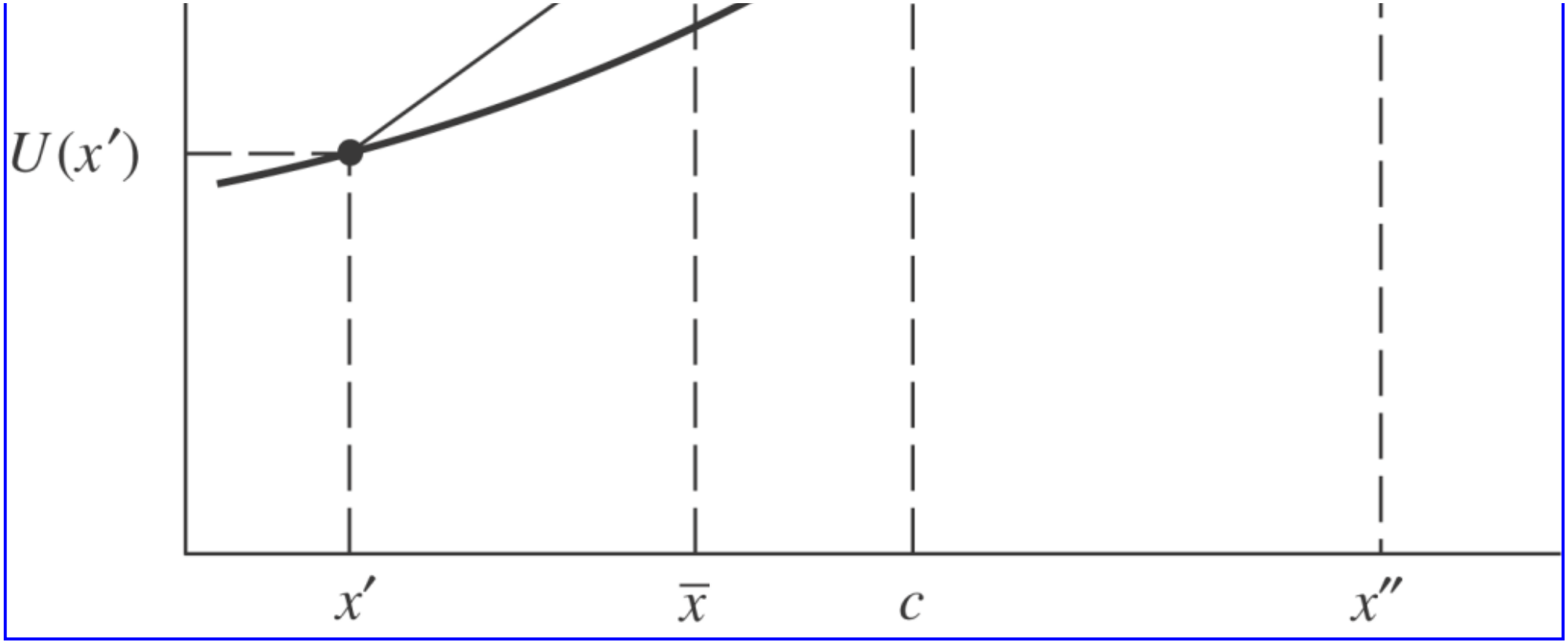

The property of first-order stochastic dominance preference can be extended to the case of distributions over the real line (Quirk and Saposnick, 1962), and it is equivalent to the condition that $U(x)$ be an increasing function of x, as in Figures 3 and 4. It is also possible to generalize the notion of a mean-preserving increase in risk to density functions or cumulative distribution functions (Rothschild and Stiglitz, 1970; 1971), and the earlier algebraic condition for risk aversion generalizes to the condition that $U''(x) < 0$ for all x, that is, that the von Neumann–Morgenstern utility function $U(\cdot)$ be concave, as in Figure 3. As before, risk aversion implies that the certainty equivalent of any lottery will lie below its mean, as seen in Figure 3; the opposite is true for the convex utility function of a risk lover, as in Figure 4. Two of the earliest and most important analyses of risk attitudes in terms of the shape of the von Neumann–Morgenstern utility function are those of Friedman and Savage (1948) and Markowitz (1952).

Analytics

The tremendous analytic capabilities of the expected utility model derive largely from the work of Arrow (1974) and Pratt (1964), who showed that the ‘degree’ of concavity of the utility function provides a measure of an expected utility maximizer's ‘degree’ of risk aversion. Formally, the Arrow–Pratt characterization of comparative risk aversion is the result that the following conditions on a pair of (increasing, twice differentiable) von Neumann–Morgenstern utility functions $U_a(\cdot)$ and $U_b(\cdot)$ are equivalent:

- $U_a(\cdot)$ is a concave transformation of $U_b(\cdot)$ (that is, $U_a(x) \equiv \rho(U_b(x))$ for some increasing concave function $\rho(\cdot)$),
- $-U_a''(x) / U_a'(x) \geq -U_b''(x) / U_b'(x)$ for each x,
- if c_a and c_b solve $U_a(c_a) = \int U_a(x)dF(x)$ and $U_b(c_b) = \int U_b(x)dF(x)$ for some distribution $F(\cdot)$, then $c_a \leq c_b$,

and if $U_a(\cdot)$ and $U_b(\cdot)$ are both concave, these conditions are in turn equivalent to:

- if $r>0$, $E[\tilde{z}] > r$, $\text{prob}(\tilde{z} < r) > 0$, and α_a and α_b maximize $\int U_a((I-\alpha)\cdot r+\alpha\cdot z)dF(z)$ and $\int U_b((I-\alpha)\cdot r+\alpha\cdot z)dF(z)$ respectively, then $\alpha_a \leq \alpha_b$.

The first two of these conditions provide equivalent formulations of the notion that $U_a(\cdot)$ is a more concave function than $U_b(\cdot)$. The curvature measure $R(x) \equiv -U''(x)/U'(x)$ is known as the 'Arrow–Pratt index of (absolute) risk aversion', and plays a key role in the analytics of the expected utility model. The third condition states that the more risk averse utility function $U_a(\cdot)$ will never assign a higher certainty equivalent to any lottery $F(\cdot)$ than will $U_b(\cdot)$. The final condition pertains to the individuals' respective demands for risky assets. Specifically, assume that each must allocate $\$I$ between two assets, one yielding a riskless (gross) return of r per dollar, and the other yielding a risky return $\tilde{z}$ with a higher expected value but with some chance of doing worse than r. This condition says that the less risk-averse utility function $U_b(\cdot)$ will generate at least as great a demand for the risky asset as the more risk-averse utility function $U_a(\cdot)$. It is important to note that it is the *equivalence* of the above concavity, certainty equivalent and asset demand conditions which makes the Arrow-Pratt characterization such an important result in expected utility theory. (Ross, 1981, provides an alternative, stronger, characterization of comparative risk aversion.)

Although the applications of the expected utility model extend to virtually all branches of economic theory (for example, Hey, 1979), much of the flavour of these analyses can be gleaned from Arrow's (1974, ch. 3) analysis of the portfolio problem of the previous paragraph. If we rewrite $(I-\alpha)\cdot r+\alpha\cdot z$ as $I\cdot r+\alpha\cdot(z-r)$, the first-order condition for this problem can be expressed as:

$$\int z\cdot U'(I\cdot r+\alpha\cdot(z-r))dF(z) - \int r\cdot U'(I\cdot r+\alpha\cdot(z-r))dF(z) = 0,$$

that is, the marginal *expected* utility of the last dollar allocated to each asset is the same. The second-order condition can be written as:

$$\int (z-r)^2\cdot U'''(I\cdot r+\alpha\cdot(z-r))dF(z) < 0$$

and is ensured by the property of risk aversion (i.e. $U''(\cdot)<0$).

As usual, we may differentiate the first-order condition to obtain the effect of a change in some parameter, say initial wealth I, on the optimal level of investment in the risky asset (the optimal value of α). Differentiating the first-order condition (including α) with respect to I, solving for $d\alpha/dI$, and invoking the second-order condition and the positivity of r yields that this derivative possesses the same sign as:

$$\int (z-r)\cdot U''(I\cdot r+\alpha\cdot(z-r))dF(z).$$

Substituting $U''(x) \equiv -R(x)\cdot U'(x)$ and subtracting $R(I\cdot r)$ times the first-order condition yields that this expression is equal to:

$$-\int (z-r)\cdot[R(I\cdot r+\alpha\cdot(z-r)) - R(I\cdot r)]\cdot U'(I\cdot r+\alpha\cdot(z-r))dF(z).$$

On the assumption that α is positive and $R(\cdot)$ is monotonic, the expression $(z-r)\cdot[R(I\cdot r+\alpha\cdot(z-r))-R(I\cdot r)]$ will possess the same sign as $R'(\cdot)$. This implies that the derivative $d\alpha/dI$ will be positive (negative) whenever the Arrow–Pratt index $R(x)$ is a decreasing (increasing) function of the individual's wealth level x. In other words, an increase in initial wealth will always increase (decrease) demand for the risky asset if and only if $U(\cdot)$ exhibits decreasing (increasing) absolute risk aversion in wealth. Further examples of the analytics of the expected utility model may be found in the above references, as well as the surveys of Hirshleifer and Riley (1979), Lippman and McCall (1981), Machina (1983) and Karni and Schmeidler (1991).

Axiomatic development

Although there exist dozens of formal axiomatizations of the expected utility model, most proceed by specifying an outcome space and postulating that the individual's preferences over probability distributions on this outcome space satisfy the following four axioms: completeness, transitivity, continuity and the Independence Axiom. Although it is beyond the scope of this entry to provide a rigorous derivation of the expected utility model in its most general setting, it is possible to illustrate the meaning of the axioms and sketch a proof of the expected utility representation theorem in the simple case of a finite outcome set $\{x_1,\ldots,\bullet x_n\}$.

Recall that in such a case the objects of choice consist of probability distributions $\boldsymbol{P}=(p_1,\ldots,p_n)$ over $\{x_1,\ldots,x_n\}$, so that the following axioms refer to the individuals' weak preference relation $\succcurlyeq$ over these prospects, where $\boldsymbol{P}^*\succcurlyeq\boldsymbol{P}$ is read '$\boldsymbol{P}^*$ is weakly preferred (that is, preferred or indifferent) to $\boldsymbol{P}$' (the associated strict preference relation $\succ$ and indifference relation ~ are defined in the usual manner):

- *Completeness*: For any two distributions $\boldsymbol{P}$ and $\boldsymbol{P}^*$, either $\boldsymbol{P}^*\succcurlyeq\boldsymbol{P}$, $\boldsymbol{P}\succcurlyeq\boldsymbol{P}^*$, or both.
- *Transitivity*: If $\boldsymbol{P}^{**}\succcurlyeq\boldsymbol{P}^*$ and $\boldsymbol{P}^*\succcurlyeq\boldsymbol{P}$, then $\boldsymbol{P}^{**}\succcurlyeq\boldsymbol{P}$.
- *Mixture continuity*: If $\boldsymbol{P}^{**}\succcurlyeq\boldsymbol{P}^*\succcurlyeq\boldsymbol{P}$, then there exists some $\lambda\in[0,1]$ such that $\boldsymbol{P}^*\sim\lambda\cdot\boldsymbol{P}^{**}+(1-\lambda)\cdot\boldsymbol{P}$.
- *Independence*: For any two distributions $\boldsymbol{P}$ and $\boldsymbol{P}^*$, $\boldsymbol{P}^*\succcurlyeq\boldsymbol{P}$ if and only if $\lambda\cdot\boldsymbol{P}^*+(1-\lambda)\cdot\boldsymbol{P}^{**}\succcurlyeq\lambda\cdot\boldsymbol{P}+(1-\lambda)\cdot\boldsymbol{P}^{**}$ for all $\lambda\in(0,1]$ and all $\boldsymbol{P}^{**}$

where $\lambda\cdot\boldsymbol{P}+(1-\lambda)\cdot\boldsymbol{P}^{**}$ denotes the $\lambda:(1-\lambda)$ 'probability mixture' of $\boldsymbol{P}$ and $\boldsymbol{P}^{**}$, that is, the lottery with probabilities $(\lambda\cdot p_1+(1-\lambda)\cdot p_1^{**},\ldots,\lambda\cdot p_n+(1-\lambda)\cdot p_n^{**})$.

The completeness and transitivity axioms are analogous to their counterparts in standard consumer theory. Mixture continuity states that if the lottery P** is weakly preferred to P* and P* is weakly preferred to P, then exists some probability mixture of the most and least preferred lotteries which is indifferent to the intermediate one.

As in standard consumer theory, completeness, transitivity and continuity serve to establish the existence of a real-valued preference function $V(p_1,\ldots,p_n)$ which represents the relation $\succcurlyeq$, in the sense that $\boldsymbol{P}^*\succcurlyeq\boldsymbol{P}$ if and only if $V(p_1^*,\ldots,p_n^*)\geq V(p_1,\ldots,p_n)$. It is the Independence Axiom which gives the theory its primary empirical content by implying that $\succcurlyeq$ can be represented by a linear preference function of the form $V(p_1,\ldots,p_n)\equiv\Sigma U_i p_i$. To see the meaning of this axiom, assume that individuals are always indifferent between a two-stage compound lottery and its probabilistically equivalent single-stage lottery, and that $\boldsymbol{P}^*$ happens to be weakly preferred to $\boldsymbol{P}$. In that case, the choice between the mixtures $\lambda\cdot\boldsymbol{P}^*+(1-\lambda)\cdot\boldsymbol{P}^{**}$ and $\lambda\cdot\boldsymbol{P}+(1-\lambda)\cdot\boldsymbol{P}^{**}$ is equivalent to being presented with a coin that has a $(1-\lambda)$ chance of landing tails (in which case the prize will be $\boldsymbol{P}^{**}$) and being asked *before the flip* whether one would rather win $\boldsymbol{P}^*$ or $\boldsymbol{P}$ in the event of a head. The normative argument for the Independence Axiom is that either the coin will land tails, in which case the choice won't have mattered, or it will land heads, in which case one is 'in effect' facing a choice between $\boldsymbol{P}^*$ and $\boldsymbol{P}$ and one 'ought' to have the same preferences as before. Note finally that the above statement of the axiom in terms of the weak preference relation $\succcurlyeq$ also implies its counterparts in terms of strict preference and indifference.

In the following sketch of the expected utility representation theorem, expressions such as '$x_i\succcurlyeq x_j$' should be read as saying that the individual weakly prefers the degenerate lottery yielding x_i with certainty to that yielding x_j with certainty, and '$\lambda\cdot x_i+(1-\lambda)\cdot x_j$' will be used to denote the $\lambda:(1-\lambda)$ probability mixture of these two degenerate lotteries.

The first step in the proof is to define the von Neumann–Morgenstern utility index $\{U_i\}$ and the expected utility preference function $V(\cdot)$. Without loss of generality, we may order the outcomes so that $x_n\succcurlyeq x_{n-1}\succcurlyeq\ldots\succcurlyeq x_2\succcurlyeq x_1$. Since $x_n\succcurlyeq x_i\succcurlyeq x_1$ for each outcome x_i, mixture continuity implies that there exist scalars $\{U_i\}\subset[0,1]$ such that $x_i\sim U_i\cdot x_n+(1-U_i)\cdot x_1$ for each i (which implies $U_1=0$ and $U_n=1$). Given this, define $V(\boldsymbol{P})=\Sigma U_i p_i$ for each $\boldsymbol{P}$.

The second step is to show that each lottery $\boldsymbol{P}=(p_1,\ldots,p_n)$ is indifferent to the mixture $\lambda\cdot x_n+(1-\lambda)\cdot x_1$ where $\lambda=\Sigma U_i p_i$. Since $(p_1,\ldots,p_n)$ can be written as the n-component probability mixture $p_1\cdot x_1+p_2\cdot x_2+\ldots+p_n\cdot x_n$, and each outcome x_i is indifferent to the mixture $U_i\cdot x_n+(1-U_i)\cdot x_1$, an n-fold application of the Independence Axiom

yields that $\boldsymbol{P} = (p_1, \ldots, p_n)$ is indifferent to the mixture

$$p_1 \cdot [U_1 \cdot x_n + (1 - U_1) \cdot x_1] + p_2 \cdot [U_2 \cdot x_n + (1 - U_2) \cdot x_1] + \ldots + p_n \cdot [U_n \cdot x_n + (1 - U_n) \cdot x_1],$$

which is equivalent to $(\Sigma_{i=1}^{n} U_i p_i) \cdot x_n + (1 - \Sigma_{i=1}^{n} U_i p_i) \cdot x_1$.

The third step is to demonstrate that a mixture $\lambda^* \cdot x_n + (1 - \lambda^*) \cdot x_1$ is weakly preferred to a mixture $\lambda \cdot x_n + (1 - \lambda) \cdot x_1$ if and only if $\lambda^* \geq \lambda$. This follows immediately from the Independence Axiom and the fact that $\lambda^* \geq \lambda$ implies that these two lotteries may be expressed as the respective mixtures $(\lambda^* - \lambda) \cdot x_n + (1 - \lambda^* + \lambda) \cdot \boldsymbol{Q}$ and $(\lambda^* - \lambda) \cdot x_1 + (1 - \lambda^* + \lambda) \cdot \boldsymbol{Q}$, where $\boldsymbol{Q}$ is defined as the lottery $(\lambda / (1 - \lambda^* + \lambda)) \cdot x_n + ((1 - \lambda^*) / (1 - \lambda^* + \lambda)) \cdot x_1$.

The completion of the proof is now simple. For any two distributions $\boldsymbol{P}^* = (p_1^*, \ldots, p_n^*)$ and $\boldsymbol{P} = (p_1, \ldots, p_n)$, transitivity and the second step imply that $\boldsymbol{P}^* \succsim \boldsymbol{P}$ if and only if

$$\left(\sum_{i=1}^{n} U_i p_i^*\right) \cdot x_n + \left(1 - \sum_{i=1}^{n} U_i p_i^*\right) \cdot x_1 \succsim \left(\sum_{i=1}^{n} U_i p_i\right) \cdot x_n + \left(1 - \sum_{i=1}^{n} U_i p_i\right) \cdot x_1,$$

which by the third step is equivalent to the condition $\Sigma U_i p_i^* \geq \Sigma U_i p_i$, or in other words, that $V(\boldsymbol{P}^*) \geq V(\boldsymbol{P})$.

As mentioned, the expected utility model has been axiomatized many times and in many contexts. The most comprehensive accounts of the axiomatics of the model are undoubtedly Fishburn (1982) and Kreps (1988).

Subjective expected utility

In addition to the above setting of ‘objective’ (that is, probabilistic) uncertainty, it is possible to define expected utility preferences under conditions of ‘subjective’ uncertainty. In this case, uncertainty is represented by a set $\mathcal{S}$ of mutually exclusive and exhaustive ‘states of nature,’ which can be a finite set $\{s_1,\ldots,s_n\}$ (as with a horse race), a real interval $[\underline{s}, \overline{s}] \subseteq R^1$ (as with tomorrow's temperature), or a more abstract space. The objects of choice are then ‘acts’ $a(\cdot): \mathcal{S} \rightarrow \mathcal{X}$ which map states to outcomes. In the case of a finite state space, acts are usually expressed in the form $\{x_1 \text{ if } s_1;\ldots; x_n \text{ if } s_n\}$. When the state space is infinite, finite-outcome acts can be expressed in the form $a(\cdot) = [x_1 \text{ on } E_1; \ldots; x_m \text{ on } E_m]$ for some partition of $\mathcal{S}$ into a family of mutually exclusive and exhaustive ‘events’ $\{E_1,\ldots,E_m\}$. Except for casino games and state lotteries, virtually all real-world uncertain decisions (including all investment or insurance decisions) are made under conditions of subjective uncertainty.

In such a setting, the ‘subjective expected utility hypothesis’ consists of the *joint* hypothesis that the individual possesses *probabilistic beliefs*, as represented by a ‘personal’ or ‘subjective’ probability measure $\mu(\cdot)$ over the state space, and *expected utility risk preferences*, as represented by a von Neumann–Morgenstern utility function $U(\cdot)$ over outcomes, and evaluates acts according a preference function of the form $W(x_1 \text{ if } s_1; \ldots; x_n \text{ if } s_n) \equiv \Sigma_{i=1}^{n} U(x_i) \cdot \mu(s_i)$, $W(x_1 \text{ on } E_1; \ldots; x_m \text{ on } E_m) \equiv \Sigma_{i=1}^{m} U(x_i) \cdot \mu(E_i)$, or more generally, $W(a(\cdot)) \equiv \int U(a(s)) d\mu(s)$. Whereas all individuals facing a given *objective* prospect $\boldsymbol{P} = (x_1, p_1; \ldots; x_n, p_n)$ are assumed to ‘see’ the same probabilities $(p_1,\ldots,p_n)$ (though they may have different utility functions), individuals facing a given *subjective* prospect $\{x_1 \text{ if } s_1;\ldots; x_n \text{ if } s_n\}$ or $[x_1 \text{ on } E_1;\ldots; x_m \text{ on } E_m]$ will generally possess differing subjective probabilities over these states or events, reflecting their different beliefs, past experiences, and so on.

Researchers such as Arrow (1974), Debreu (1959, ch. 7) and Hirshleifer (1965; 1966) have shown how the analytics of the objective expected utility model can be extended to both the positive and normative analysis of decisions under subjective uncertainty. As a simple example, consider an individual deciding whether to purchase earthquake insurance, and if so, how much. A simple specification of this decision involves the state space $\mathcal{S} = \{s_1, s_2\} = \{\text{earthquake}, \text{ no earthquake}\}$, the individual's von Neumann–Morgenstern utility of wealth function $U(\cdot)$, their subjective probabilities $\{\mu(s_1), \mu(s_2)\}$ (which sum to unity), and the price γ of each dollar of insurance coverage. An individual with initial wealth w would then purchase q dollars’ worth of coverage, where q was the solution to

$$\max_{q} \; [U(w-\gamma q+q)\cdot\mu(s_1)+U(w-\gamma q)\cdot\mu(s_2)]$$

Note that this formulation does not require that the individual and the insurance company agree on the likelihood of an earthquake.
As in the objective case, subjective expected utility can be derived from axiomatic foundations. Completeness and transitivity carry over in a straightforward way, and continuity with respect to mixture probabilities is replaced by continuity with respect to small changes in the events. The existence of additive personal probabilities is obtained by the following axiom:

- *Comparative likelihood:* For all events A, B and outcomes $x^* \succ x$ and $y^* \succ y$, [x^* on A; x on $\sim A$] $\succsim$ [x^* on B; x on $\sim B$] implies [y^* on A; y on $\sim A$] $\succsim$ [y^* on B; y on $\sim B$].

This axiom states that if the individual ‘reveals’ event A to be at least as likely as event B by their preference for staking the preferred outcome x^* on A rather than on B, then this likelihood ranking will hold for all other pairs of ranked outcomes $y^* \succ y$. Finally, under subjective uncertainty the Independence Axiom is replaced by its subjective analogue, first proposed by Savage (1954):

- *Sure-Thing Principle*: For all events E and acts $a(\cdot)$, $a^*(\cdot)$, $b(\cdot)$ and $c(\cdot)$, [$a^*(\cdot)$ on E; $b(\cdot)$ on $\sim E$] $\succsim$ [$a(\cdot)$ on E; $b(\cdot)$ on $\sim E$] implies [$a^*(\cdot)$ on E; $c(\cdot)$ on $\sim E$] $\succsim$ [$a(\cdot)$ on E; $c(\cdot)$ on $\sim E$].

where [$a(\cdot)$ on E; $b(\cdot)$ on $\sim E$] denotes the act yielding outcome $a(s)$ for all $s \in E$ and $b(s)$ for all $s \in \sim E$.
Under subjective uncertainty, an individual's utility of outcomes might sometimes depend upon the particular state of nature. Given a health insurance decision with a state space of $S = \{s_1, s_2\} = \{\text{cancer, no cancer}\}$, an individual may feel a greater need for $100,000 in state s_1 than in state s_2. This can be modelled by means of a ‘state-dependent’ utility function $\{U(\cdot|s)|s \in S\}$ and a ‘state-dependent expected utility’ preference function $\widehat{W}(x_1 \text{ if } s_1; \ldots; x_n \text{ if } s_n) = \Sigma_{i=1}^{n} U(x_i|s_i)\cdot\mu(s_i) \text{ or } \widehat{W}(a(\cdot)) = \int U(a(s)|s)\,d\mu(s)$. The analytics of state-dependent expected utility preferences have been extensively developed by Karni (1985).

History

The hypothesis that individuals might maximize the expectation of ‘utility’ rather than of monetary value was proposed independently by mathematicians Gabriel Cramer and Daniel Bernoulli, in each case as the solution to a problem posed by Daniel's cousin Nicholas Bernoulli (see Bernoulli, 1738). This problem, now known as the ‘St Petersburg Paradox’, considers the gamble which offers a 1/2 chance of $1, a 1/4 chance of $2, a 1/8 chance of $4, and so on. Although the expected value of this prospect is

$$(1/2)\cdot\$1+(1/4)\cdot\$2+(1/8)\cdot\$4+\cdots = \$0.50+\$0.50+\$0.50+\cdots = \$\infty,$$

common sense suggests that no one would be willing to forgo a very substantial certain payment in order to play it. Cramer and Bernoulli proposed that, instead of using expected value, individuals might evaluate this and other lotteries by means of their expected ‘utility’, with utility given by a function such as the natural logarithm or the square root of wealth, in which case the certainty equivalent of the St Petersburg gamble becomes a moderate (and plausible) amount.
Two hundred years later, the St Petersburg paradox was generalized by Karl Menger (1934), who noted that, whenever the utility of wealth function was unbounded (as with the natural logarithm or square root functions), it would be possible to construct similar examples with infinite expected utility and hence infinite certainty equivalents (replace the payoffs $1, $2, $4 … in the above example by x_1, x_2, x_3 …, where $U(x_i)=2^i$ for each i). In light of this, von Neumann–Morgenstern utility functions are typically (though not universally) postulated to be bounded functions of wealth.
The earliest formal axiomatic treatment of the expected utility hypothesis was developed by Frank Ramsey (1926) as part of his theory of subjective probability, or individuals’

‘degrees of belief’ in the truth of alternative propositions. Starting from the premise that there exists an ‘ethically neutral’ proposition whose degree of belief is 1/2, and whose validity or invalidity is of no independent value, Ramsey proposed a set of axioms on how the individual would be willing to stake prizes on its truth or falsity, in a manner which allowed for the derivation of the ‘utilities’ of these prizes. He then used these utility values and betting preferences to determine the individual's degrees of belief in other propositions. Perhaps because it was intended as a contribution to the philosophy of belief rather than to the theory of risk bearing, Ramsey's analysis did not have the impact among economists that it deserved.

The first axiomatization of the expected utility model to receive widespread attention was that of John von Neumann and Oskar Morgenstern, presented in connection with their formulation of the theory of games (von Neumann and Morgenstern, 1944; 1947; 1953). Although this development was recognized as a breakthrough, the mistaken belief that von Neumann and Morgenstern had somehow mathematically overthrown the Hicks–Allen ‘ordinal revolution’ led to some confusion until the difference between ‘utility’ in the von Neumann-Morgenstern and the ordinal (that is, non-stochastic) senses was illuminated by writers such as Ellsberg (1954) and Baumol (1958).

Another factor which delayed the acceptance of the theory was the lack of recognition of the role played by the Independence Axiom, which did not explicitly appear in the von Neumann–Morgenstern formulation. In fact, the initial reaction of researchers such as Baumol (1951) and Samuelson (1950) was that there was no reason why preferences over probability distributions must *necessarily* be linear in the probabilities. However, the independent discovery of the Independence Axiom by Marschak (1950), Samuelson (1952) and others, and Malinvaud's (1952) observation that it had been implicitly invoked by von Neumann and Morgenstern, led to an almost universal acceptance of the expected utility hypothesis as both a normative and positive theory of behaviour towards risk. This period also saw the development of the elegant axiomatization of Herstein and Milnor (1953) as well as Savage's (1954) joint axiomatization of utility and subjective probability, which formed the basis of the state-preference approach described above.

While the 1950s essentially saw the completion of foundational work on the expected utility model, subsequent decades saw the flowering of its analytic capabilities and its application to fields such as portfolio selection (Merton, 1969), optimal savings (Levhari and Srinivasan, 1969; Fleming and Sheu, 1999), international trade (Batra, 1975; Lusztig and James, 2006), environmental economics (Wolfson, Kadane and Small, 1996), medical decision-making (Meltzer, 2001) and even the measurement of inequality (Atkinson, 1970). This movement was spearheaded by the development of the Arrow–Pratt characterization of risk aversion (see above) and the characterization, by Rothschild–Stiglitz (1970; 1971) and others, of the notion of ‘increasing risk’. This latter work in turn led to the development of a general theory of ‘stochastic dominance’ (for example, Whitmore and Findlay, 1978; Levy, 1992), which has further expanded the analytical powers of the model.

Although the expected utility model received a small amount of experimental testing by economists in the early 1950s (for example, Mosteller and Nogee, 1951; Allais, 1953) and continued to be examined by psychologists, economists’ interest in the empirical validity of the model waned from the mid-1950s through the mid-1970s, no doubt due to both the normative appeal of the Independence Axiom and model's analytical successes. However, since the late 1970s there has been a revival of interest in the testing of the expected utility model; a growing body of evidence that individuals’ preferences *systematically* depart from linearity in the probabilities; and the development, analysis and application of alternative models of choice under objective and subjective uncertainty. It is fair to say that today the debate over the descriptive (and even normative) validity of the expected utility hypothesis is more extensive than it has been in over half a century, and the outcome of this debate will have important implications for the direction of research in the economics of uncertainty.

See Also

- Bernoulli, Daniel
- non-expected utility theory
- Ramsey, Frank Plumpton
- risk
- risk aversion
- Savage's subjective expected utility model
- uncertainty
- utility

Bibliography

Allais, M. 1953. Fondements d'une théorie positive des choix comportant un risque et critique des postulats et axiomes de l’école Américaine. *Colloques Internationaux du Centre National de la Recherche Scientifique* 40, 257–332. Trans. as: The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School, in *Expected Utility Hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen. Dordrecht: D. Reidel, 1979.

Arrow, K. 1974. *Essays in the Theory of Risk-Bearing*. Amsterdam: North-Holland.

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 244–63.

Batra, R. 1975. *The Pure Theory of International Trade under Uncertainty*. London: Macmillan.

Baumol, W. 1951. The Neumann–Morgenstern utility index: an ordinalist view. *Journal of Political Economy* 59, 61–6.

Baumol, W. 1958. The cardinal utility which is ordinal. *Economic Journal* 68, 665–72.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*. Trans. as Exposition of a new theory on the measurement of risk. *Econometrica* 22 (1954), 23–36.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New Haven: Yale University Press.

Ellsberg, D. 1954. Classical and current notions of ‘measurable utility’. *Economic Journal* 64, 528–56.

Fishburn, P. 1982. *The Foundations of Expected Utility*. Dordrecht: D. Reidel.

Fleming, W. and Sheu, S.-J. 1999. Optimal long term growth rate of expected utility of wealth. *Annals of Applied Probability* 9, 871–903.

Friedman, M. and Savage, L. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56, 279–304.

Herstein, I. and Milnor, J. 1953. An axiomatic approach to measurable utility. *Econometrica* 21, 291–7.

Hey, J. 1979. *Uncertainty in Microeconomics*. Oxford: Martin Robinson; New York: New York University Press.

Hirshleifer, J. 1965. Investment decision under uncertainty: choice theoretic approaches. *Quarterly Journal of Economics* 79, 509–36.

Hirshleifer, J. 1966. Investment decision under uncertainty: applications of the state-preference approach. *Quarterly Journal of Economics* 80, 252–77.

Hirshleifer, J. and Riley, J. 1979. The analytics of uncertainty and information – an expository survey. *Journal of Economic Literature* 17, 1375–421.

Karni, E. 1985. *Decision Making under Uncertainty: The Case of State-Dependent Preferences*. Cambridge, MA: Harvard University Press.

Karni, E. and Schmeidler, D. 1991. Utility theory with uncertainty. In *Handbook of Mathematical Economics*, vol. 4, ed. W. Hildenbrand and H. Sonnenschein. Amsterdam: North-Holland.

Kreps, D. 1988. *Notes on the Theory of Choice*. Boulder, CO: Westview Press.

Levhari, D. and Srinivasan, T.N. 1969. Optimal savings under uncertainty. *Review of Economic Studies* 36, 153–64.

Levy, H. 1992. Stochastic dominance and expected utility: survey and analysis. *Management Science* 38, 555–93.

Lippman, S. and McCall, J. 1981. The economics of uncertainty: selected topics and probabilistic methods. In *Handbook of Mathematical Economics*, vol. 1, ed. K. Arrow and M. Intriligator. Amsterdam: North-Holland.

Lusztig, M. and James, P. 2006. How does free trade become institutionalised? An expected utility model of the Chrétien era. *World Economy* 29, 491–505.

Machina, M. 1983. *The Economic Theory of Individual Behavior Toward Risk: Theory, Evidence and New Directions*. Technical Report No. 433. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Malinvaud, E. 1952. Note on von Neumann–Morgenstern's strong Independence Axiom. *Econometrica* 20, 679–80.

Markowitz, H. 1952. The utility of wealth. *Journal of Political Economy* 60, 151–8.

Marschak, J. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18, 111–41.

Meltzer, D. 2001. Addressing uncertainty in medical cost-effectiveness analysis: implications of expected utility maximization for methods to perform sensitivity analysis and the use of cost-effectiveness analysis to set priorities for medical research. *Journal of Health Economics* 20, 109–29.

Menger, K. 1934. Das Unsicherheitsmoment in der Wertlehre. *Zeitschrift für Nationalökonomie*. Trans. as: The role of uncertainty in economics, in *Essays in Mathematical Economics in Honor of Oskar Morgenstern*, ed. M. Shubik. Princeton: Princeton University Press, 1967.

Merton, R. 1969. Lifetime portfolio selection under uncertainty: the continuous time case. *Review of Economics and Statistics* 51, 247–57.

Mosteller, F. and Nogee, P. 1951. An experimental measurement of utility. *Journal of Political Economy* 59, 371–404.

Pratt, J. 1964. Risk aversion in the small and in the large. *Econometrica* 32, 122–36.

Quirk, J. and Saposnick, R. 1962. Admissibility and measurable utility functions. *Review of Economic Studies* 29, 140–6.

Ramsey, F. 1926. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*, ed. R. Braithwaite. New York: Harcourt, Brace and Co, 1931. Reprinted in *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, ed. D. Mellor. New Jersey: Humanities Press, 1978.

Ross, S. 1981. Some stronger measures of risk aversion in the small and in the large, with applications. *Econometrica* 49, 621–38.

Rothschild, M. and Stiglitz, J. 1970. Increasing risk I: a definition. *Journal of Economic Theory* 2, 225–43.

Rothschild, M. and Stiglitz, J. 1971. Increasing risk II: its economic consequences. *Journal of Economic Theory* 3, 66–84.

Samuelson, P. 1950. Probability and attempts to measure utility. *Economic Review* 1, 167–73.

Samuelson, P. 1952. Probability, utility, and the Independence Axiom. *Econometrica* 20, 670–8.

Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley & Sons. Revised edn, New York: Dover, 1972.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press. 2nd edn, 1947; 3rd edn, 1953.

Whitmore, G. and Findlay, M., eds. 1978. *Stochastic Dominance: An Approach to Decision Making Under Risk*. Lexington, MA: D.C. Heath.

Wolfson, L., Kadane, J. and Small, M. 1996. Expected utility as a policy making tool: an environmental health example. In *Bayesian Biostatistics*, ed. D. Berry and D. Stangl. New York: Marcel Dekker.

How to cite this article

Machina, Mark J. "expected utility hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000178> doi:10.1057/9780230226203.0526

The New Palgrave Dictionary of Economics Online

experimental economics, history of

Francesco Guala
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Contemporary experimental economics was born in the 1950s from the combination of the experimental method used in psychology and new developments in economic theory. Early experimental studies of bargaining behaviour, social dilemmas, individual decision making and market institutions were followed by a long period of underground growth, until the booming of the field in the 1980s and 1990s.

Keywords

Allais paradox; Allais, M.; auctions (experiments); bargaining; behavioural economics; cardinal utility; Convergence; decision theory; expected utility; experimental economics; experimental economics, history of; Friedman, M.; game theory; gaming; Kahneman, D.; learning; mathematics and economics; mechanism design; Mill, J. S.; models; other-regarding preferences; Plott, C.; positive economics; preference reversals; prisoner's dilemma; public goods; Ramsey, F.; Savage, L.; Selten, R.; Siegel, S.; Simon, H.; Smith, V.; Stanford Value Project; statistics and economics; subgame perfection; subjective probability; Tversky, A.; ultimatum game; utility measurement; von Neumann and Morgenstern

Article

Experimental economics has experienced one of the most stunning methodological revolutions in the history of science. In just a few decades, economics has been transformed from a discipline where the experimental method was considered impractical, ineffective and largely irrelevant to one where some of the most exciting advancements are driven by laboratory data.
Like many other new developments in the social sciences during the second half of the 20th century, experimental economics is largely a by-product of the combination of massive investments in science, a fertile intellectual culture and socio-political conditions in the 1940s and 1950s in the United States. Although it is possible in principle to identify earlier experimental or proto-experimental work being done in economics and psychology (see Roth, 1995), there is hardly any direct intellectual, personal or

institutional continuity between these isolated episodes and today's fully institutionalized experimental programme.

A proper history of experimental economics is yet to be written, and one challenge faced by historians of the discipline is its strikingly interdisciplinary character. The rise of experimental economics takes the form of several, partly independent and partly intertwined threads that can be brought under a single coherent narrative only with difficulty. It is partly for this reason that most of the existing historical literature consists of personal recollections or reconstructions of individual trajectories rather than of a collective enterprise. It is possible, however, to identify some key moments and achievements that have helped to establish experimentation as a legitimate method of investigation in economics.

Historical background and early years

The traditional view of economics as a primarily non-experimental science was outlined in the methodological writings of 19th-century economists. John Stuart Mill (1836, p. 124), for example, identifies several practical obstacles to the use of the experimental method, in particular the impossibility of controlling key economic variables and of keeping background conditions fixed so that the effect of manipulating each cause in isolation can be checked. This was Mill's main justification for adopting the so-called ‘a priori deductive’ method, a mix of introspection and theoretical reasoning, to determine what an idealized *homo oeconomicus* would do in given circumstances. Despite various changes in economists’ methodological rhetoric and practice, it took a century and a half for philosophical scepticism towards experimentation to fade away.

Like many methodological revolutions in science, the experimental turn in economics was primarily made possible not by a change in philosophical perspective but by a number of innovations at the level of scientific practice and theoretical commitment. At a very general level, in the middle of the 20th century economics was in the process of becoming a ‘tool-based’ science (Morgan, 2003): from the old, discursive ‘moral science’ of political economy, it was changing into a discipline where models, statistics and mathematics played the role both of instruments and, crucially, of *objects* of investigation. During this conceptual revolution economists came to accept that the path towards the understanding of a real-world economy might have to go through the detailed analysis of several tools that had apparently only a vague resemblance to the final target of investigation. Theoretical models and computer simulations entered the economists’ toolkit first, with laboratory experiments following shortly after.

The birth of experimental economics owes much to the publication of von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (1944) and to the subsequent developments of game and decision theory. Although game theory is often seen primarily as a contribution to the theoretical corpus of economics, this was not how it was perceived at the time. Von Neumann and Morgenstern's work initially found fertile ground in a community of scientists devoted to the simultaneous development of a great variety of approaches and research methods and interested in their application to solve scientific, policy, and management problems across the disciplinary boundaries – from conflict resolution in international relations to group psychology, cybernetics, and the organization of the firm, to name just a few.

‘Gaming’ – playing game-theoretic problems for real – was common practice in the mathematical community at Princeton in the 1940s and 1950s, and quickly spread elsewhere as game theory increased

in popularity. This practice did not involve sophisticated experimental design, but was conceived mainly as a useful way of illustrating game theoretic puzzles, as well as a check on abstract speculation and a guide to the theoretician's intuitions. Traces of this attitude can be found in the writings of some pioneers in game theory in the 1950s, who explicitly advocated a combination of formal theorizing and empirical evidence of various kinds, and engaged in (mostly casual) forms of experimenting to back up their theoretical claims (see, for example, Schelling, 1960; Shubik, 1960).

The first event devoted specifically to ‘The Design of Experiments in Decision Processes’ was a 1952 two-month seminar sponsored by the Ford Foundation, organized in Santa Monica by a group of researchers at the University of Michigan. The seminar's location was intended to facilitate the participation of members of the RAND Corporation, a think tank sponsored by the US Air Force, where among others Merrill Flood was conducting game-theoretic experiments (including famously the first Prisoner's Dilemma experiments). It is difficult to assess at all precisely the role of the Santa Monica seminar in the birth of experimental economics because, apart from an important minority, most of the published papers (in Thrall, Coombs and Davis, 1954) are theoretical rather than experimental in character. Several later protagonists, however, first became familiar with the idea of experiments in economics through the Santa Monica seminar, which therefore functioned as a catalyst in various indirect ways (see Smith, 1992).

The most extensive experimental projects of the 1950s were pursued at Penn State, Michigan, and Stanford. In collaboration with Lawrence Fouraker, the psychologist Sidney Siegel conducted a systematic investigation of bargaining behaviour at Pennsylvania State University, trying to combine what he took to be the most advanced aspects of economics (the theory) and psychology (the experimental method). The project came to an abrupt end with Siegel's death in 1961, but the resulting book (Siegel and Fouraker, 1960) won the American Academy of Arts and Sciences best monograph prize. Siegel and Fouraker's experiments focused on several aspects of bargaining behaviour, but are particularly significant for the systematic study of variations in the monetary payoffs and in the information made available to the subjects. Interestingly, this research project was rather disconnected from current developments in axiomatic bargaining theory, focusing instead on testing various hypotheses from the psychological literature. ‘Level of aspiration theory’ emerged eventually as the best predictor of bargaining behaviour.

From the point of view of experimental design, Siegel is often credited with being the first experimenter to highlight the importance of using real incentives to motivate subjects but, with hindsight, his experiments with Fouraker are also remarkable for the implementation of strict between-subjects anonymity. The latter practice would become very common in later experimental economics, usually as an attempt to implement economic theory's standard atomistic assumptions (especially the ban on other-regarding preferences). Contrary to the standard economic theory, Fouraker and Siegel recognized that interpersonal reactions do matter, but left a systematic investigation of their effects for later research.

More or less simultaneously, Ward Edwards at Michigan pioneered the experimental study of expected utility theory, as axiomatized in the second edition of von Neumann and Morgenstern's *Theory of Games* (1947). Amos Tversky, a student of Edwards and Coombs, would play a major role in the institutionalization of behavioural economics two decades later, as we shall see. In the mid-1950s an interdisciplinary group was also at work on the new theory of individual decision making, under the heading of the Stanford Value Project. Donald Davidson and Pat Suppes (both to become famous later for their contributions to philosophy) published with Siegel one of the first monographs of experimental

decision theory (Davidson, Suppes and Siegel, 1957). At the centre of their research were measurement issues, in particular the implementation of learning theory and Frank Ramsey's method for measuring utilities and subjective probabilities.

Another major centre of interdisciplinary research in those years was the Carnegie group working on the psychology of organizations. Herbert Simon – working at Carnegie and the RAND Corporation, himself a participant in the Santa Monica seminar – is usually credited with being a pivotal player in this connection, although his influence on experimental economics is mostly indirect. The Carnegie group made use of a variety of methodologies, among which experimental ‘role playing’, ‘business games’, and simulations were central. In their larger projects, like the Carnegie Tech Management Game, human decision makers took managerial decisions in an environment simulated by a computer. Although primarily devised for pedagogic and illustrative purposes, such games were also used to shed light on the ‘boundedly rational’ processes of decision making that guide behaviour in big organizations. There is little continuity, however, between this body of work and contemporary experimental economics, with Simon playing a role more as a source of moral support and intellectual inspiration than as a direct contributor to experimental research.

The most famous experimental discovery of this period is due to a scholar who was to have little to do with later developments in experimental economics. Maurice Allais had been developing in France his own version of utility theory as a cardinal measurable quantity well before the publication of the *Theory of Games*. At a conference he organized in Paris in 1952, during a lunch break Allais presented Leonard Savage with a ‘questionnaire’ that was to become famous as the ‘Allais paradox’ experiment. When Savage gave answers that were inconsistent with the expected utility model he himself supported, Allais was encouraged to extend his questionnaire and to circulate it more widely.

The results were partially published in French in *Econometrica* (Allais, 1953) but received little attention in the short term. The main immediate result of the Allais experiment was Savage's switch to a purely normative defence of expected utility (Jallais and Pradier, 2005). Milton Friedman at the time was developing his methodology of positive economics which accorded no importance to the accuracy of the models of individual decision used to predict aggregate phenomena; and Allais's chauvinistic polemic against the ‘American School’ probably did little to attract sympathy. For about two decades Allais did not pursue research in this area any further.

The only large-scale experimental research project in Europe during this period was led by Reinhard Selten in Frankfurt, under the auspices of Heinz Sauermann. Like other early game theorists, Selten was convinced that the theory could contribute to the solution of important social science problems only if used in conjunction with empirical evidence. Indeed, even his most celebrated theoretical achievement (the concept of subgame perfection) was conceived in the context of a larger experimental project (see Selten, 1995).

The last piece of the puzzle of experimental economics in the 1950s is at the same time the most important and the most idiosyncratic. Vernon Smith had been experimenting at Purdue since 1956, focusing on the properties of different market institutions and their effects on the convergence towards equilibrium (see Smith, 1981). Smith had an engineering background and, unlike most experimenters at the time, did not approach experiments from a game-theoretic perspective. In the 1940s and 1950s Edward Chamberlin at Harvard had been performing little classroom experiments for illustrative purposes, to show his graduate students the falsity of the competitive theory of markets. Although the

results of such experiments had been published in the *Journal of Political Economy* (1948), nobody at the time, including Chamberlin, attributed particular scientific value to them. Smith was the exception: a few years after leaving graduate school he came to question the design used by Chamberlin and to test the robustness of the ‘no convergence’ results to variations in the exchange institution and repetition of the task.

Overcoming several obstacles, Smith managed to publish his counter-experiments to Chamberlin (Smith, 1962). For many years Smith led the only experimental project carried on fully within the boundaries of the economics discipline. In the early 1960s his work received funding from the National Science Foundation, but, apart from a brief attempt to collaborate with the Carnegie group (see Lee, 2004), his work in this phase was mostly carried out in isolation. One important exception is Smith's brief but important encounter with Sidney Siegel at Stanford in 1961. Smith perceived Siegel as much more advanced in methodological matters, and took from him several insights in experimental design that were to become the hallmark of economic experimentation (Smith, 1981; 1992).

From the underground to the big bang

Like other innovations of the previous two decades, experimental economics went through a period of slow, quiet growth in the 1960s. Some early contributors, like Allais, disappeared from the scene; others, like Smith, quit experimenting for some time (1967–74) and generally struggled to find an audience. Some areas, like social dilemmas and bargaining experiments, were booming in psychology but had little impact on the economics literature (see Leonard, 1994). In the 1970s, however, the landscape of experimental economics changed considerably, partly thanks to the formation of a few key partnerships. During 1968–9 Amos Tversky began collaborating with Daniel Kahneman at the Hebrew University, initially on judgement and then on decision making. In Europe, by 1972, Selten had moved to Bielefeld and started a collaboration with Werner Güth, later author of the first experiments on the ultimatum game. Allais in the meantime returned to expected utility in 1974, and was persuaded to publish a full report in English of his 1952 results (in Allais and Hagen, 1979). Allais's legacy would also begin to bear some fruits on the theoretical front. The late 1970s and early 1980s were characterized by a proliferation of alternative models to expected utility, mostly inspired by the experimental evidence that had been accumulated up until then.

After the happy anarchy of the earlier period, the 1970s were marked by the beginning of some controversies and the partial separation of the experimental community into sub-disciplines. In 1974 an article by Tversky and Kahneman in *Science* was widely read as a challenge to the view that human beings were rational agents, and, although it made experiments on judgement and decision making enter the intellectual debate at large, it also fed some deep cross-disciplinary misconceptions. A few years later Lichtenstein and Slovic's seminal experiments on preference reversals were introduced into the economics literature by Grether and Plott (1979), kicking off a series of theoretical and experimental papers that would fill the pages of the *American Economic Review* for years.

Charles Plott had been in close contact with Vernon Smith since the early 1960s, and started to run experiments a decade later, after his move to Caltech. Their collaboration led not only to important experimental projects but also to the creation of the Caltech laboratory and the training of the second and third generations of experimental economists. An important outcome of this period was also the attempt

to systematize the methodology of experimental economics around a set of rules or ‘precepts’ of experimental design (Smith, 1976; 1982). Smith in these papers highlighted the importance of monetary incentives to control subjects’ preferences, a practice that he had borrowed from Siegel – a psychologist – but that ironically was to become the main distinguishing feature of the ‘economic’ way of experimenting, as opposed to the more liberal ‘psychological’ way. With hindsight these methodological papers are also striking for their effective use of the language and conceptual framework of mechanism design theory. In this sense they reflected Smith's (and Plott's) attitude towards the use of experiments to tackle real-world problems of institutional design and policymaking (see Guala, 2005).

With the slow exhaustion of general equilibrium theory, the turmoil in macroeconomics, and an increasing disillusionment about econometrics, the 1970s created the conditions for the seeds of the 1940s and 1950s to finally blossom. Experimental economists were in a position to take advantage of this situation. By the early 1980s most of the ‘paradigmatic’ experiments that would inform subsequent research had already been published (Smith and Plott's experiments on auctions and markets, Lichtenstein and Slovic (1971) on preference reversals, Plott and others on public goods (Isaac, McCue and Plott, 1985), Güth on the ultimatum game (Güth, Schmittberger and Schwartz, 1982), Alvin Roth and others on bargaining (Roth and Malouf, 1979)). Consolidation meant also differentiation. A persistent low-intensity conflict at the methodological and theoretical level led to the creation of so-called ‘behavioural economics’. Whereas experimental economics refers primarily to a method of investigation, the work of behavioural economists is unified by a substantial project of revision of economic theory (especially the replacement of *homo oeconomicus* with a more realistic psychological model), with experimentation constituting a major but by no means exclusive source of evidence.

The history of experimental economics in the 1980s and 1990s is the story of a booming research programme, increasingly influential within the discipline and the social sciences at large, expanding in new directions – neuroscience, for example – and attracting some of the most talented graduate students. Together with game theorists, experimenters have also been increasingly involved in policymaking, notably by contributing to the design of new market institutions for the allocation of sensitive goods – from telecommunication licences to space stations, airport slots, and physicians and surgeons (see Roth, 2002). In 2002 the Nobel Memorial Prize in economics awarded to Vernon Smith and Daniel Kahneman provided official acknowledgement of this remarkable revolution.

See Also

- Allais Paradox
- behavioural game theory
- experimental economics
- field experiments
- Kahneman, Daniel
- neuroeconomics
- Smith, Vernon
- Tversky, Amos

Bibliography

Allais, M. 1953. La psychologie de l'homme rationnel devant le risque: critique des postulats et axiomes de l’école Américaine. *Econometrica* 21, 503–46.

Allais, M. and Hagen, O., eds. 1979. *Expected Utility Hypothesis and the Allais Paradox*. Dordrecht: Reidel.

Chamberlin, E.H. 1948. An experimental imperfect market. *Journal of Political Economy* 56, 95–108.

Davidson, D., Suppes, P. and Siegel, S. 1957. *Decision Making: An Experimental Approach*. Stanford: Stanford University Press.

Grether, D. and Plott, C. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69, 623–38.

Guala, F. 2005. *The Methodology of Experimental Economics*. New York: Cambridge University Press.

Güth, W., Schmittberger, R. and Schwartz, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3, 363–88.

Isaac, R., McCue, K. and Plott, C. 1985. Public goods provision in an experimental environment. *Journal of Public Economics* 26, 51–74.

Jallais, S. and Pradier, P. 2005. The Allais paradox and its immediate consequences for expected utility theory. In *The Experiment in the History of Economics*, ed. P. Fontaine and R. Leonard. London: Routledge.

Lee, K. 2004. Rationality, minds, and machines in the laboratory: a thematic history of Vernon Smith's experimental economics. Ph.D. thesis, University of Notre Dame.

Leonard, R. 1994. Laboratory strife: higgling as experimental science in economics and social psychology. In *Higgling*, ed. N. De Marchi and M. Morgan. Durham. NC: Duke University Press.

Lichtenstein, S. and Slovic, P. 1971. Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89, 46–55.

Mill, J.S. 1836. On the definition of political economy and the method of investigation proper to it. In *Collected Works of John Stuart Mill*, Vol. 4. Toronto: University of Toronto Press, 1967.

Morgan, M. 2003. Economics. In *The Cambridge History of Science, Volume 7: The Modern Social*

Sciences, ed. T. Porter and D. Ross. Cambridge: Cambridge University Press.

Neumann, J. Von and Morgenstern, O. 1944. *The Theory of Games and Economic Behavior*, 2nd edn. Princeton: Princeton University Press, 1947.

Roth, A. 1995. Introduction to experimental economics. In *The Handbook of Experimental Economics*, ed. J. Kagel and A. Roth. Princeton: Princeton University Press.

Roth, A. 2002. The economist as engineer: game theory, experimentation, and computation as tools for design economics. *Econometrica* 70, 1341–78.

Roth, A. and Malouf, M. 1979. Game-theoretic models and the role of information in bargaining. *Psychological Review* 86, 574–94.

Schelling, T. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Selten, R. 1995. Autobiography. In *Les Prix Nobel/The Nobel Prizes 1994*. Stockholm: Nobel Foundation.

Shubik, M. 1960. Bibliography on simulation, gaming, artificial intelligence and allied topics. *Journal of the American Statistical Association* 55, 736–51.

Siegel, S. and Fouraker, L. 1960. *Bargaining and Group Decision Making*. New York: McGraw-Hill.

Smith, V. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70, 111–37.

Smith, V. 1976. Experimental economics: induced value theory. *American Economic Review* 66, 274–7.

Smith, V. 1981. Experimental economics at Purdue. In *Papers in Experimental Economics*. Cambridge: Cambridge University Press.

Smith, V. 1982. Microeconomic systems as an experimental science. *American Economic Review* 72, 923–55.

Smith, V. 1992. Game theory and experimental economics: beginnings and early influences. In *Towards a History of Game Theory*, ed. E. Weintraub. Durham, NC: Duke University Press.

Thrall, R., Coombs, C. and Davis, R., eds. 1954. *Decision Processes*. New York: Wiley.

Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185,

1124–30.

How to cite this article

Guala, Francesco. "experimental economics, history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000289> doi:10.1057/9780230226203.0528

The New Palgrave Dictionary of Economics Online

experimental economics

Vernon L. Smith
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Experimental methods have features in common across all the sciences. All tend to use the framework of falsification, but there is inherent ambiguity in knowing which of the many hypotheses necessary to construct a test are negated by observations contrary to predictions. This ambiguity tends to engender much discussion, contestability and the design of new experiments that attempt to resolve the open qsts. This social process is not part of the logic of scientific testing, but it explains what scientists do and how new results become established.

Keywords

bounded rationality; Brownian motion; constructivist rationality; dictator games; double auction; Duhem–Quine problem; ecological rationality; equilibrium; excess demand; experimental economics; experimental knowledge; extensive form games; falsificationism; first-price auctions; Friedman, M.; Hayek, F. A.; Internet, economics of; market power; mechanism design; neuroscience; perfect competition; perfect information; positive economics; prediction; private information; repeated games; risk neutrality; strategic games; subjective probability; subjective utility; testing; ultimatum games

Article

> But I believe that there is no philosophical highroad in science, with epistemological signposts…we are in a jungle and find our way out by trial and error, building our road *behind* us as we proceed. We do not *find* signposts at crossroads, but our scouts *erect them*, to help the rest.
> —Max Born, *Experiment and Theory in Physics* (1943)

> … they were criticized [those studying observational learning in a social context] for being unscientific and performing *uncontrolled* experiments. In science, there's nothing ‘worse’ than an experiment that's uncontrolled.
> —Temple Grandin, *Animals in Translation* (2005, bracketed comments added).

The subject matter of this article is rationality in science particularly as it applies to experimental methods. In this context ‘rationality’ is commonly used to refer to a particular conception that Hayek (1967, p 85) has called:

> *Constructivist Rationality*, which, applied to individuals, associations or organizations, involves the conscious deliberate use of reason to analyze and prescribe actions judged to be better than alternative feasible actions that might be chosen; applied to institutions it involves the deliberate design of rule systems to achieve desirable performance. The latter include ‘optimal design’ where the intention is to provide incentives for agents to choose better actions than would result from alternative arrangements.

Rationality in socioeconomic systems, including scientific communities, cannot be adequately understood by restricting one's perspective to this traditional Cartesian framework. In the discussion that follows I want to draw upon a second conception of rationality:

> *Ecological rationality* refers to emergent order in the form of practices, norms and rules governing action by individuals, groups and institutions that are part of our cultural and biological heritage, created by human interactions, but not by conscious human design.

I have argued (Smith, 2003) that rationality in the economy depends on individuals whose behaviour is conditioned by cultural norms and emergent institutions that evolve from human experience, neither of which is ultimately derived from constructivist reason; although, clearly, constructivist ideas are important sources of variation for the gristmill of ecological selection. Parallel considerations apply to rationality in scientific method, the extension to be treated here.
Stated briefly, here is the argument that I will present: scientific methodology reveals a predominately constructivist theme largely guided by the following:

- falsification criteria for hypotheses derived from theories;
- experimental designs for testing hypotheses;
- statistical tests; and
- liturgies of reporting style that have become standard in scientific papers.

But all tests of theory are necessarily joint tests of hypotheses derived from theory and the set of auxiliary hypotheses necessary to implement, construct and execute the tests: this is the well-known Duhem–Quine (D–Q) problem. Thus, whatever might be the testing rhetoric of scientists, they do not reject hypotheses, and their antecedent theories, on the basis of falsifying outcomes. But this is not cause for despair, let alone retreat into a narrow postmodern sea of denial in which science borders on unintended fraud. D–Q is a property of inquiry, a truth, and as such a source for deepening our understanding of what *is*, not a clever *touché* for exposing the rhetorical pretensions of science. But the failure of all philosophy of science programs to articulate a rational constructivist methodology of science that serves to guide scientists, or explain what they do, as well as what they say about what they do, does not mean that science is devoid of rationality or that scientific communities fail to generate rational programmes of scientific inquiry. Thus, scientists engage in commentary, reply, rebuttal and vigorous discussions over whether the design is appropriate, the test adequate, whether the procedures and measurements might be flawed, and the conclusions and inpts correct. One must look to this conversation in the scientific community in asking whether and how science sorts out competing primary and auxiliary hypotheses after each new set of tests results are made available. If this conversation does not read like a theorem and its proof, and fails to reduce methodology to a consciously rigorous science of inquiry, this is because we can never reduce the testing enterprise to a simple up or down test of an isolated non-trivial hypothesis; so be it.
If emergent method is rational in science it must be a form of ecological rationality; this means that it rightly and inevitably grows out of the norms, practices and conversation that characterize meaningful interactions in the scientific community. Listen not only to what scientists say about what they do, ignoring the arrogant tone in their little knowledge, but also examine what they do. The power behind the throne of accomplishment in the human career is our *sociality*, and the unintended mansions that are built by that sociality. The long view of that career is in sharp focus: our accumulation of knowledge and its expression in technology enabled us to survive the Pleistocene, people the Earth, penetrate the heavens, and explore the ultimate particles and forces of matter, energy and life. That achievement hardly deserves to be described as either irrational or non-rational.
What does it mean to test a theory or a hypothesis derived from a theory? Scientists everywhere say and believe that the unique feature of science is that theories, if they are to be acceptable, require rigorous support from facts based on replicable observations. But the deeper one examines this belief the more elusive becomes the task of giving it precise meaning and content in the context of conventional rational programs of inquiry.

Can we derive theory directly from observations?

Prominent scientists through history have believed that the answer to this question is ‘yes’, and that this was their modus operandi. Thus, quoting from two of the most influential scientists in history:

> I frame no hypotheses; for whatever is not deduced from the phenomena … have no place in experimental philosophy … (in which) … particular propositions are inferred from the phenomena … (Isaac Newton, *Principia*, 1687; quoted in Segrè, 1984, p. 66)

> Nobody who has really gone into the matter will deny that in practice the world of phenomena uniquely determines the theoretical system, in spite of the fact that there is no theoretical bridge between phenomena and their theoretical principles. (Einstein, 1934, pp. 22–3)

But these statements are without validity. ‘One can today easily demonstrate that there can be no valid derivation of a law of nature from any finite number of facts’ (Lakatos, 1978, vol. 1, p. 2). Yet in economics critics of standard theory call for more empirical work on which to base development of *the* theory, and generally the idea persists that the essence of science is its rigorous observational foundation.

But how are the facts and theories of science to be connected so that each constructively informs and enriches the other?

Newton passionately believed not just that he was proffering lowly hypotheses, but that his laws were derived directly, by logic, from Kepler's discovery that the planets moved in ellipses. But Newton only showed that the path was an ellipse if there are n=2 planets. Kepler was wrong in thinking that the planets followed elliptical paths, and to this day there is no solution for the n(>2)-body problem, and in fact the paths can be chaotic. Thus, when he published the *Principia*, Newton's model could not account for the motion of our nearest and most accurately observable neighbour, the moon, whose orbit is strongly influenced by both the sun and the earth.

Newton's sense of his scientific procedure is commonplace: one studies an empirical regularity (for example, the ‘trade-off’ between the rate of inflation and the unemployment rate), and proceeds to articulate a model from which a functional form can be derived that yields the regularity. In the above confusing quotation, Einstein seems to agree with Newton. At other times he appears to articulate the more qualified view that theories make predictions, which are then to be tested by observations (see his insightful comment below on Kaufmann's test of special relativity theory), while on other occasions his view is that reported facts are irrelevant compared to theories based on logically complete meta theoretical principles, coherent across a broad spectrum of fundamentals (see Northrup, 1969, pp. 387–408). Thus, upon receiving the telegraphed news that Eddington's 1919 eclipse experiments had ‘confirmed’ the general theory, Einstein showed it to a doctoral student who was jubilant, but he commented unmoved: ‘I knew all the time that the theory was correct.’ But what if it had been refuted? ‘In that case I'd have to feel sorry for God, because the theory is correct’ (quoted in Fölsing, 1997, p. 439).

The main theme I want to develop in this and subsequent sections is captured by the following quotation from a lowbrow source, the mythical character Phaedrus in *Zen and the Art of Motorcycle Maintenance*, ‘… the number of rational hypotheses that can explain any given phenomena is infinite’ (Pirsig, 1981, p. 100).

Proposition 1: Particular hypotheses derived from any testable theory imply certain observational outcomes; the converse is false (Lakatos, 1978, vol. 1, pp. 2, 16, *passim*).

Theories produce mathematical theorems. Each theorem is a mapping from postulated statements (assumptions) into derived or concluding statements (the theoretical results). Conventionally, the concluding statements are what the experimentalist uses to formulate specific hypotheses (models) that motivate the experimental design that is implemented. The conditions that underpin the hypotheses are the objects of control in an economics experiment, insofar as they can be controlled. Since not every assumption can always be reproduced in the experimental design the problem of the ‘controlled experiment’ is one of trying to minimize the risk that the results will fail to be interpretable as a test of the theory because one or more assumptions were violated. An uncontrolled assumption that is postulated to hold in interpreting test results is one of many possible contingent auxiliary hypotheses to be discussed below.

The wellspring of testable hypotheses in economics and game theory is to be found in the marginal conditions defining equilibrium points or strategy decision functions that constitute a theoretical equilibrium. In games against nature the subject-agent is assumed to choose among alternatives in the feasible set that which maximizes his or her outcome (reward, utility or payoff), subject to the technological and other constraints on choice. Strategic games are solved by the device of reducing them to games against nature, as in a non-cooperative (Cournot–Nash) equilibrium (pure or mixed) where each agent is assumed to maximize his or her own outcome, given (subject to the constraints of) the maximizing behaviour of all other agents. The equilibrium strategy when used by all but agent i reduces i's problem to an own maximizing choice of that strategy. Hence, in economics, all testable hypotheses come from the marginal conditions (or their discrete opportunity cost equivalent) for maximization that define equilibrium for an individual or across individuals interacting through an institution. These conditions are implied by the theory from which they are derived, but given experimental observations consistent with (that is, supporting) these conditions there is no way to reverse the steps used to derive the conditions, and deduce the theory from a set of observations on subject choice. Behavioural point observations conforming to an equilibrium theory cannot be used to deduce or infer either the equations defining the equilibrium or the logic and assumptions of the theory used to derive the equilibrium conditions.

Suppose, however, that the theory is used to derive non-cooperative best reply functions for each agent that maps one or more characteristics of each individual into that person's equilibrium decision. Suppose next that we perform many repetitions of an experiment varying some controllable characteristic of the individuals, such as their assigned values for an auctioned item, and obtain an observed response for each value of the characteristic. This repetition of course must be assumed always to support equilibrium outcomes. Finally, suppose we use this data to estimate response functions obtained from the original maximization theory. First order conditions defining an optimum can always be treated formally as differential equations in the original criterion function. Can we solve these equations and ‘deduce’ the original theory?

An example is discussed in Smith, McCabe and Rassenti (1991) from first-price auction theory. Briefly the idea is this. Each of N individuals in a repeated auction game is assigned value $v_i(t)$ $(i = 1, \ldots, N; \quad t = 1, 2, \ldots, T)$ from a distribution, and on each trial, t, i bids $b_i(t)$. On each trial each i is assumed to choose a bid that maximizes expected utility.

$$\max_{0 \le b_i \le v_i} (v_i - b_i)^{r_i} G_i(b_i)$$

(1)

where r_i $(0 < r_i \leq 1)$ is i's measure of constant relative risk aversion, and $G_i(b_i)$ is i's probability belief that a bid of b_i will win. This leads to the first order condition,

$$(v_i - b_i)G_i'(b_i) - r_iG_i(b_i) = 0.$$
(2)

If all i have identical common rational probability expectations

$$G_i(b_i) = G(b_i).$$
(3)

This leads to a closed form equilibrium bid function (see Cox, Smith and Walker, 1988).

$$b_i = (N-1)v_i / (N-1+r_i), \quad b_i \leqq \bar{b} = 1 - G(\bar{b}) / G'(b), \quad \text{for all } i.$$
(4)

The data from experimental auctions strongly support linear subject bid rules of the form

$$b_i = \alpha_i v_i,$$
(5)

obtained by linear regression of b_i on v_i using the T observations on (b_i, v_i), for given N, with $\alpha_i = (N-1)v_i / (N-1+r_i)$. Can we reverse the above steps, then integrate (2) to get (1)? The answer is ‘no’. Equation (2) can be derived from maximizing either (1) or the criterion

$$(v_i - b_i)G(b_i)^{1/r_i},$$
(1 ′)

in which subjective probabilities, rather than profit, are ‘discounted’ in computing the expectation. That is, without all the assumptions used to get (4), we cannot uniquely conclude

(1). In (1′) we have $G_i(b_i) = G(b_i)^{1/r_i}$ instead of (3), and this is not ruled out by the data.
In fact, instead of (1) or (1′) we could have maximized

$$(v_i - b_i)^{\beta_i} G(b_i)^{\beta_i / r_i}, \quad \text{with} \quad r_i \le \beta_i \le 1 \qquad (1')$$

giving an infinite mixture of subjective utility and subjective probability models of bidding.
There is a special case of the above model that is reversible: all bidders are risk neutral with, r_i=1. While that model is often defended in the abstract, and is the workhorse assumption for deriving theorems under uncertainty, it fails all the lab and field empirical tests known to me. Risk neutrality trivializes decisions by requiring all humans to have identical preferences. It fails because people are inherently heterogeneous in making decisions under uncertainty, and this empirical diversity is captured by an appearance (or a mirage) in the data of non-neutral ‘risk’. Equation (1′) above provides a hint of the manner in which individual diversity can appear in data inpts that confound measures of risk with other sources of heterogeneity.
Thus, in general, we cannot backward infer from empirical equilibrium conditions, even when we have a large number of experimental observations, to arrive at the original parameterized model within the general theory. *The purpose of theory is precisely one of imposing much more structure on the problem than can be inferred from the data*. This is because the assumptions used to deduce the theoretical model contain more information, such as (3), than the data – the theory is underdetermined.

Economics: is it an experimental science?

All editions of Paul Samuelson's *Principles of Economics* refer to the inability of economists to perform experiments. This continued for a short time after William Nordhaus joined Samuelson as a coauthor. Thus, ‘Economists … cannot perform the controlled experiments of chemists and biologists because they cannot easily control other important factors’ (Samuelson and Nordhaus, 1985, p. 8).
My favourite quotation, however, is supplied by one of the 20th century's foremost Marxian economists, Joan Robinson. To wit, ‘Economists cannot make use of controlled experiments to settle their differences’ (Robinson, 1979, p. 1319). Like Samuelson, she was not accurate – economists do indeed perform controlled experiments – but how often have they, or their counterparts in any science, used them to ‘settle their differences?’ Here she was expressing the popular image of science, which is indeed one in which ‘objective’ facts are the arbiters of truth that in turn ‘settle’ differences. The caricatured image is that of two scientists, who, disagreeing on a fundamental principle, go to the lab, do a ‘crucial experiment’, and learn which view is assuredly right. The hypothesis they are testing is not underdetermined by the test data. They do not argue about the result; their question is answered and they move on to a new topic that is not yet ‘settled.’
Although these quotations provide telling commentaries on the state of the profession's knowledge of the development of experimental methods in economics since the 1950s, there is a deeper question of whether there are more than a very small number of non-experimentalists in economics that understand key features of our methodology. These are twofold: (*a*) employ a reward scheme to motivate individual behaviour in the laboratory within an economic environment defining gains from trade that are controlled by the experimenter – for example, the supply of and demand for an abstract item in an isolated market or an auction; and (*b*) use the observations to test predictive hypotheses derived from one or more models (formal or informal) of behaviour in these environments using the rules of a particular trading institution – for example, the equilibrium clearing price and corresponding exchange volume when subjects trade under some version of an oral or electronic double auction, posted pricing, sealed bidding, and so on. This differs from the way that economics is commonly researched, taught and practised, which implies that it is largely an a priori science in which economic problems come to be understood by thinking about them. This generates logically correct, internally consistent theories and models. The data of econometrics are then used for ‘testing’ between alternative model specifications within basic equilibrium theories that are not subject to challenge, or to estimate the supply and/or demand parameters assumed to generate data representing equilibrium outcomes by an unspecified process. (Leamer, 1978, and others have challenged the interpretation of this standard econometric methodology as a scientific ‘testing’ programme as distinct from a programme for specification searches of data.) Theories are not so much subject to doubt as used to impose restrictions on the data that allow parameters to be estimated. *Its constructivism all the way down*.
I want to report two examples indicating how counter-intuitive it has been for prominent economists to see the function of laboratory experiments in economics. The first example is contained in a quotation from Hayek whose Nobel citation was for his theoretical conception of the price system as an information system for coordinating agents with dispersed information in a world where no single mind or control centre possesses, or can ever have knowledge of, this information. His critique and rejection of mainstream quantitative

methods, ‘scientism’, in economics are well known (see, for example, Hayek, 1942; 1945). But in his brilliant paper interpreting competition as a discovery process, rather than a model of equilibrium price determination, he argues:

> … wherever the use of competition can be rationally justified, it is on the ground that we do not know in advance the facts that determine the actions of competitors … competition is valuable only because, and so far as, its results are unpredictable and on the whole different from those which anyone has, or could have, deliberately aimed at.… The necessary consequence of the reason why we use competition is that, in those cases in which it is interesting, the validity of the theory can never be tested empirically. We can test it on conceptual models, and *we might conceivably test it in artificially created real situations, where the facts that competition is intended to discover are already known to the observer. But in such cases it is of no practical value, so that to carry out the experiment would hardly be worth the expen*se. (F.A. Hayek, 1978, p. 255; emphasis added)

Hayek describes with clarity an important use (unknown to him) that has been made of experiments – testing competitive theory ‘in artificially created real situations, where the facts which competition is intended to discover are already known to the observer’ – then proceeds to completely fail to see how such an experiment could be used to test his own proposition that competition is a discovery procedure, under the condition that neither agents as a whole nor any one mind needs to know what each agent knows. Rather, his concern for dramatizing what is arguably the most important socio-economic idea of the 20th century seems to have caused him to interpret his suggested hypothetical experiment as ‘of no practical value’ since it would (if successful) merely reveal what the observer already knew!

I find it astounding that one of the most profound thinkers in the 20th century did not see the demonstration potential and testing power of the experiment he suggests for testing the proposition: with competition *no one in the market* need know in advance the actions of competitors, and that competition is valuable only because, and so far as, its results are unpredictable *by anyone in the market* and on the whole different from those which anyone *in the market* has, or could have, deliberately aimed at. Yet, unknown to me at the time, this is precisely what my first experiment conducted in January 1956, published later as ‘Test 1’, was all about (Smith, 1962).

I assembled a considerable number of experiments for a paper ‘Markets as Economizers of Information: Experimental Examination of the “Hayek Hypothesis”’, presented at the 50th Jubilee Congress of the Australian and New Zealand Association for the Advancement of Science, in Adelaide, Australia, 12–16 May 1980. A version of this paper was reprinted in Smith (1991, pp. 221–35). Here is what I called the Hayek Hypothesis. Strict privacy together with the trading rules of a market institution (the oral double auction in this case) is sufficient to produce efficient competitive market outcomes. The alternative was called the Complete Knowledge Hypothesis: competitive outcomes require perfectly foreseen conditions of supply and demand, a statement attributable to many economists, including Paul Samuelson who refers to ‘foreseen changes in supply and demand’ (Samuelson, 1966, p. 947 and *passim*), that can be traced back to W.S. Jevons in 1871. (Stigler, 1957, provides a historical treatment of the concept of perfect competition.) In this empirical comparison the Hayek Hypothesis was strongly supported. This theme had been visited earlier (before I had become aware or at least fully appreciative of Hayek's 1945 contribution that equilibrium theory was a tautology) in Smith (1976), wherein eight experiments comparing private information with complete information showed that complete information was neither necessary nor sufficient for convergence to a competitive equilibrium: complete information interfered with, and slowed, convergence compared with private information. Shubik (1959, pp. 169–71) had noted earlier, and correctly, the confusion inherent in ad hoc claims that perfect knowledge is a requirement of pure (or sometimes perfect) competition. The experimental proposition that private information increases support for non-cooperative, including competitive, outcomes applies not only to markets but also to the two-person extensive-form repeated games reported by McCabe, Rassenti and Smith (1998). Hence it is clear that without knowledge of the other's payoff it is not possible for players to identify and consciously coordinate on a cooperative outcome. Thus, as we have learned, payoff information is essential to conscious coordination in two-person interactions, but irrelevant, if not pernicious, in impersonal market exchange. We note in passing that the large number of experiments demonstrating the Hayek Hypothesis in no sense implies that there may not be exceptions (Holt, 1989). It's the other way around: this large set of experiments demonstrates clearly that there are exceptions almost everywhere to the Complete Knowledge Hypothesis, and these exceptions were not part of a prior design created to provide a counter example.

Holt and his coauthors have asked, ‘Are there any conditions under which double-auction markets do not generate competitive outcomes? The only exception seems to be an experiment with a “market power” design reported by Holt, Langan and Villamil (1986) and replicated by Davis and Williams (1991)’ (Davis and Holt, 1993, p. 154; also see Holt, 1989). The example reported in this exception was a market in which there was a constant excess supply of only one unit – a market with inherently weak equilibrating properties. Actually, there were two earlier reported exceptions, neither of which required market power: (*a*) one in which information about private circumstances is known by all traders (the alternative to the Hayek Hypothesis as stated above), and (*b*) an example in which the excess supply in the market was only two units. Exception (*a*) was reported in Smith (1980; 1991, pp. 104–5) and (*b*) in Smith (1991, p. 67). The above cited exceptions, attributed to market power, would need to be supplemented with comparisons in which there was just one unit of excess demand, but no market power, in order to show whether or not each exception was driven by market power, and not the fact that there is only one unit of excess supply which may be enabling of above equilibrium prices even is there is no market power. Sometimes missing in the standard toolkit of experimentalists are routines for challenging our own interpretation of data where there are confounding elements in the explanations of the results. But in this respect our methodology keeps getting better.

My second example involves the same principle as the first. It derives from a personal conversation in the early 1980s with one of my favorite Nobel Laureates in economics, a

prominent theorist. In response to a qst, I described the experimental public goods research I had been doing in the late 1970s and early 1980s comparing the efficacy of various public good mechanisms: Lindahl, Groves–Ledyard, the so-called auction election or mechanism. (See the public goods papers reprinted in Smith, 1991.) He wondered how I had achieved control over the efficient allocation as the benchmark used in these comparisons. So I explained what I had naively thought was commonly understood by then: I give each subject a payoff function (table) in monetary payoffs defined jointly over variable units of a public (common outcome) good, and variable units of a private good. This allows the experimenter to solve for the social optimum and then use the experimental data to judge the comparative performance of alternative public good incentive mechanisms. Incredibly, he objected that if, as the experimenter, I have sufficient information to know what constitutes the socially optimal allocation then I did not need a mechanism! I can just impose the optimal allocation! Baldly stated, economics is about deducing best actions from theory, not finding ways to test its propositions. So there I was, essentially an anthropologist on Mars, unable to convey to one of the best and brightest in the traditional ways of thinking that the whole idea of laboratory experiments was to evaluate mechanisms in an environment where *the Pareto optimal outcome was known by the experimental designer but not by the agents* so that performance comparisons could be made; that in the field such knowledge was never possible, and we had no criteria, other than internal theoretical properties such as incentive compatibility to judge the efficacy of the mechanism. He didn't get it; psychologically this testing procedure is not comprehensible if somehow your thinking has accustomed you to believe that allocation mechanisms require agents to have complete information, but not mechanism designers who presumably slipped through by assuming their agents were fully informed. In fact, with that worldview what is there to test in mechanism theory?

The issue of whether economics is an experimental science is moot among experimental economists who are, and should be, too busy having fun doing their work to reflect on the methodological implications of what they do. But when we do speak of methodology, as in comprehensive introductions to the field, what do we say? Quotations from impeccable sources will serve to introduce the concepts to be developed next. The first emphasizes that an important category of experimental work '… includes experiments designed to test the predictions of well articulated formal theories and to observe unpredicted regularities, in a controlled environment that allows these observations to be unambiguously interpreted in relation to the theory' (Kagel and Roth, 1995, p. 22). Experimental economists strongly believe, I think, that this is our most powerful scientific defence of experimental methods: we ground our experimental inquiry in the firm bedrock of economic or game theory. A second crucial advantage, recognizing that field tests involve hazardous joint tests of multiple hypotheses, is the sentiment that 'Laboratory methods allow a dramatic reduction in the number of auxiliary hypotheses involved in examining a primary hypothesis' (Davis and Holt, 1993, p. 16).

Hence the strongly held belief that, in the laboratory, we can test well-articulated theories, interpret the results unambiguously in terms of the theory, and do so with minimal, trivial or at least greatly reduced dependence on auxiliary hypothesis. This view and the idea that theories can be derived directly from observations are not unique to any science, but they are *illusions*. Fortunately, such illusions do not constitute a barrier to great scientific achievement because they appear to affect the rhetoric of science far more than its substance. Perhaps this is because the beliefs of scientists are important in reinforcing their commitment to discovery whether or not they are defensible. I cannot imagine that Newton would have been more accomplished if he had been methodologically more sophisticated.

What is the scientist's (qua experimentalist) image of what he does?

The standard experimental paper within and without economics uses the following format in outline form: (1) state the theory; (2) implement it in a particular context (with 'suitable' motivation in economics); (3) summarize the implications in one or more testable hypotheses; (4) describe the experimental design; (5) present the data and results of the hypothesis tests; (6) conclude that the experiments either reject or fail to reject the theoretical hypotheses. This format is shown in Figure 1. In the case in which we have two or more competing theories and corresponding hypotheses, the researcher offers a conclusion as to which one is supported by the data using some measure of statistical distance between the data and each of the predictive hypotheses, reporting which distance is statistically the shortest.

Figure 1

The scientist's image of scientific procedure

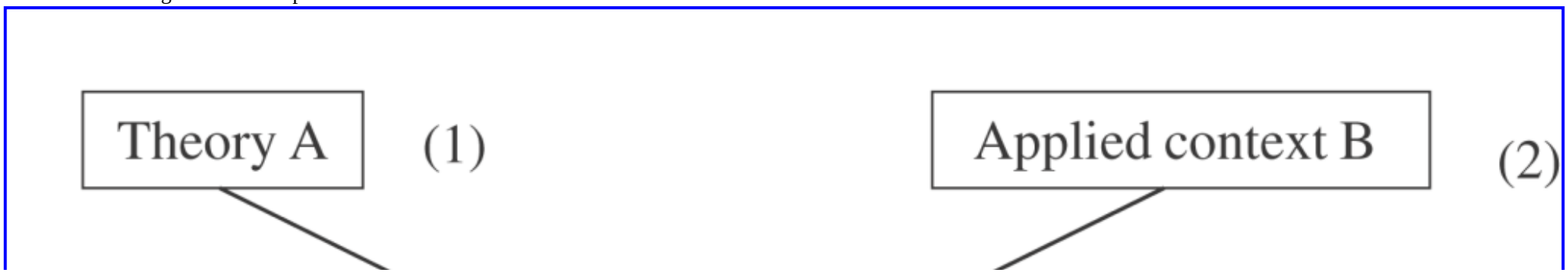

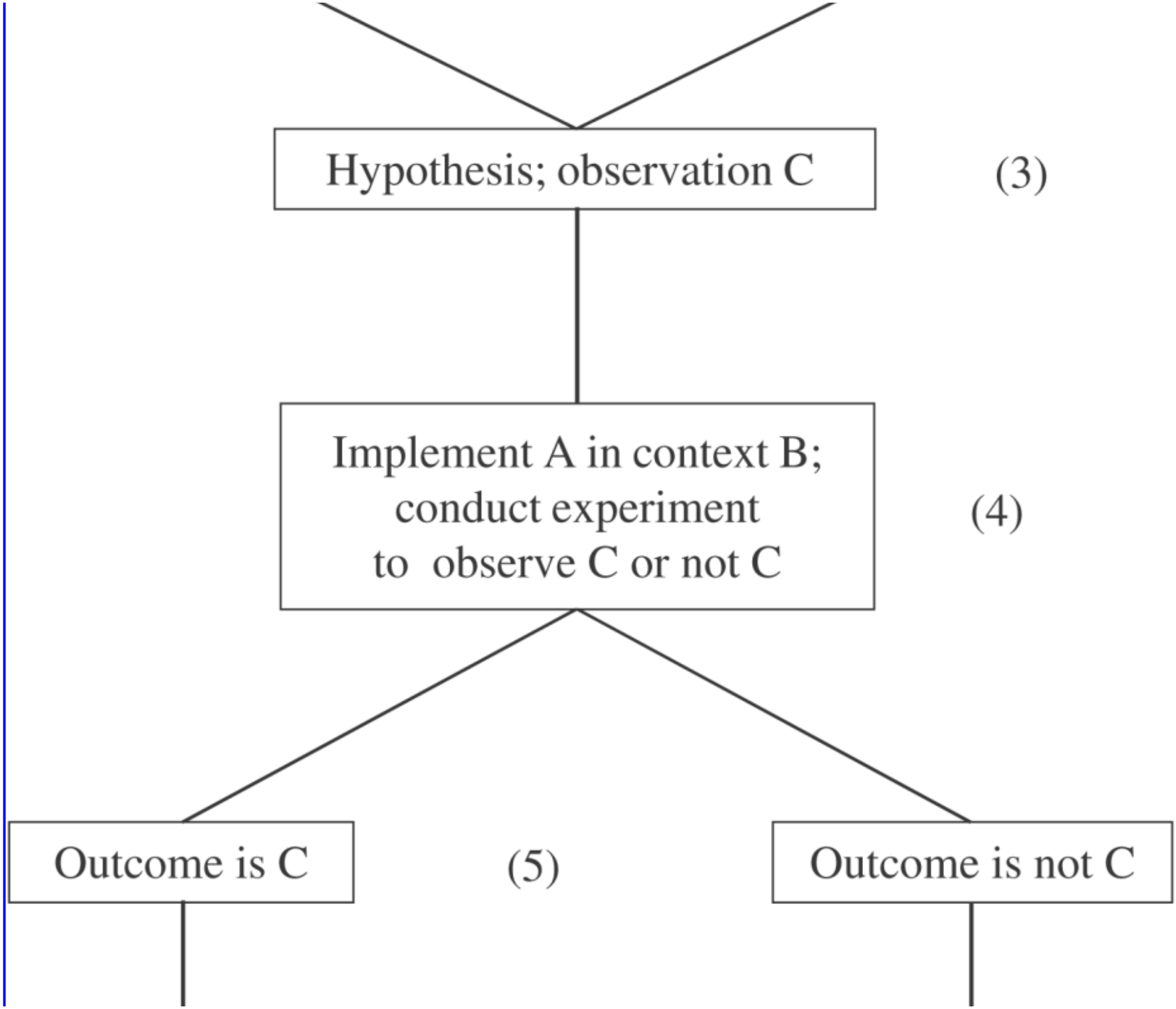
Hypothesis; observation C
(3)
Implement A in context B;
conduct experiment
to observe C or not C
(4)
Outcome is C
(5)
Outcome is not C

Suppes (1969; also see Mayo, 1996, ch. 5) has observed that there exists a hierarchy of models behind the process in Figure 1. The primary model or theory is contained in steps (1) and (2), which generate particular topical hypotheses that address primary qsts. Experimental models are contained in (3) and (4). These serve to link the primary theory with data. Finally, we have data models, steps (5) and (6), that link operations on raw data (not the raw data itself) to the testing of experimental hypotheses.

This process describes much of the rhetoric of science, and reflects the self-image of scientists, but it does not adequately articulate what scientists actually do. Furthermore, the rhetoric does not constitute a viable, defensible and coherent methodology. But what we actually do, I believe, is highly defensible and on the whole positively affects what we think we know from experiment. Implicitly, as experimentalists, we understand that every step, (1)–(6), in the above process is subject to judgments, learning from past experiments, our knowledge of protocols and technique, and to error. This is reflected in what we do as a professional community, if not in what we say about what we do in the standard scientific paper, or when we try to describe the science that we do.

As I have noted, the problem with the above image is known as the 'D–Q problem': experimental results always present a joint test of the theory (however well articulated, formally) that motivated the test, *and all the things you had to do to implement the test*. (For good discussions of the D–Q thesis and its relevance for experimental economics, see Soberg, 2005, and Guala, 2005, pp. 54–61 and *passim*. Soberg provides interesting theoretical results showing how the process of replication can be used, in the limit, to inductively eliminate clusters of alternative hypotheses and lend increasing weight to the conclusion that the theory itself is in doubt.) Thus, if theoretical hypothesis H is implemented with context specific auxiliary hypotheses required to make the test operational, $A_1, A_2, \ldots, A_n$; then it is $(H|A_1, A_2, \ldots, A_n)$ that implies observation C. If you observe not-C, this can be because any of the antecedents $(H; A_1, \ldots, A_n)$ can represent what is falsified. Thus, the interpretation of observations in relation to a theoretical hypothesis is *inherently and inescapably ambiguous*, contrary to our accustomed rhetoric.

The reality of what we do, and indeed must do, is implied by the truth that 'No theory is or can be killed by an observation. Theories can always be rescued by auxiliary hypotheses' (Lakatos, 1978, vol. 1, p. 34).

A D – Q example from physics

Here is a historical example from physics: in 1905 Kaufmann (cited in Fölsing, 1997, p. 205), a very accomplished experimentalist (in 1902 he showed that the mass of an electron is increased by its velocity!), published a paper 'falsifying' Einstein's special theory of relativity the same year in which the latter was published (Einstein, 1905). Subsequently, Einstein (1907) in a review paper reproduced Kaufmann's Figure 2, commenting that

> The little crosses above the [Kaufmann's] curve indicate the curve calculated according to the theory of relativity. In view of the difficulties involved in the experiment one would be inclined to consider the agreement as satisfactory. However, the deviations are systematic and considerably beyond the limits of error of Kaufmann's experiment. That the calculations of Mr. Kaufmann are error-free is shown by the fact that, using another method of calculation, Mr. Planck arrived at results that are in full agreement with those of Mr. Kaufmann. Only after a more diverse body of observations becomes available will it be possible to decide with confidence whether the systematic deviations are due to a not yet recognized source of errors or to the circumstance that the foundations of the theory of relativity do not correspond to the facts. (Einstein, 1907, p. 283)

Kaufmann was testing the hypothesis that (H|A) implies C, where H was an implication of special relativity theory, A was the auxiliary hypothesis that his context-specific test and enabling experimental apparatus was without 'a not yet recognized source of errors', C was the curve indicated by the 'little crosses', and not-C was Kaufmann's empirical curve (Einstein, 1907, Figure 2, p. 283).

Einstein's comment in effect says that either of the antecedents (H|A) represents what is falsified by Kaufmann's results. Others, such as Planck and Kaufmann himself, however,

acknowledged that the observation might conceivably be in error (Fölsing, 1997, p. 205). Such acknowledgements are not unusual in the scientific community, which means that scientists informally recognize the D–Q problem as it arises in particular contexts, that it is part of the scientific conversation and that they seek solutions even if this modus operandi is not part of their rhetoric. Thus, in less than a year, Bucherer (see Fölsing, 1997, p. 207) showed that there indeed had been a ‘problem’ with Kaufmann's experiments and proceeded to obtain new results supporting Einstein's theory.

There is an important lesson in this example if we develop it a little more fully. Suppose Bucherer's experiments *had not changed Kaufmann's results enough to change the conclusion*. (There is never a shortage of claims that a given experimental result may be in doubt: see Mayo, 1996, for the imaginative arguments proffered by the Newtonians in response to Eddington's eclipse observations.) Then Einstein could still have argued that there may be ‘a not yet recognized source of errors’. If so, the implication is that H is not falsifiable, for the same argument can be made after each new test in which the new results are outside the range of error for the apparatus! Recall that the deviations were alleged to be ‘considerably beyond the limits of error of Kaufmann's experiment’. But here ‘error’ is used in the sense of internal variations arising from the apparatus and procedure, not in the sense that there is a problem with the apparatus or procedure itself. We can go still further in explicating the problem of testing H conditional on any A. The key is to note in this example the strong dependence of any test outcome on the *state of experimental knowledge*: Bucherer found a way to ‘improve’ Kaufmann's experimental technique so as to rescue Einstein's ‘prediction’. But the predictive content of H (and therefore of the special theory) was inextricably bound up with A. Einstein's theory did not embrace any of the elements of Kaufmann's (or Bucherer's) *apparatus:* A is based on experimental knowledge of testing procedures and operations in the physics laboratory, and has nothing to do with the theory of relativity, a separate and distinct body of theoretically coherent knowledge.

A proposition and an economics example

Here is the most common casual empiricist objection to economics experiments: the payoffs are too small. This objection is one of several principal issues in a target article by Hertwig and Ortmann (2001), with comments by 34 experimental psychologists and economists. This objection sometimes is packaged with an elaboration to the effect that economic or game theory is about situations that involve large stakes, and you ‘can't’ study these in the laboratory. (Actually, of course, you can, but funding is not readily available.) Suppose, therefore that we have the following:

- H (from theory): subjects will choose the equilibrium outcome (for example, Nash or subgame perfect).
- A (auxiliary hypothesis): payoffs are adequate to motivate subjects.

Proposition 2: Suppose a specific rigorous test rejects (H|A), and someone (say, T), protests that what must be rejected is A not H. Let E replicate the original experiment with an *n*-fold increase in payoffs. There are only two outcomes and corresponding inpts, neither of which is comforting to the rhetorical image of science as conducting falsification tests of predictive hypotheses:

1. 1. The test outcome is negative. Then T can imagine a still larger payoff multiple $N > n$, and still argue for rejecting A not H. But this implies that H *cannot be falsified*.
2. 2. After repeated increases in payoffs, the test outcome is positive for some $N \geqq n^*$. Then H has *no predictive content*. E, with no guidelines from the theory, has searched for and discovered an empirical payoff multiple, n^*, that ‘confirms’ the theory, but n^* is an *extra theoretical property of considerations outside H and the theory* that was being tested. Finding this multiple is not something for T or E to crow about, but rather an event that should send T or E back to his desk. The theory is inadequately specified to embrace the observations from all the experiments.

Proposition 2 holds independent of any of the following considerations:

- how well articulated, rigorous or formal the theory is; game theory in no sense constitutes an exception;
- how effective the experimental design is in reducing the number of auxiliary hypotheses – it only takes one to create the problem; and
- the character or nature of the auxiliary hypothesis – *A* can be anything not contained in the theory.

In experimental economics, reward adequacy is just one of a standard litany of objections to experiments in general and to many experiments in particular. Here are three additional categories in this litany:

1. *Subject sophistication*. The standard claim is that undergraduates are not sophisticated enough. They are not out there participating in the ‘real world’. In the ‘real world’ where the stakes are large, such as in the FCC spectrum rights auctions, bidders use game theorists as consultants (Banks et al., 2003). (For an investigation of the hypothesis that

undergraduates are insufficiently sophisticated, see McCabe and Smith, 2000, who report a comparison of undergraduate and graduate students, and these with economics faculty, in a two-person trust game. The first two groups were indistinguishable, and both earned more than the faculty because with greater frequency they risked defection in offering to cooperate, as against opting for the subgame perfect outcome.)

2. *Subjects need an opportunity to learn*. This is a common response from both experimentalists and theorists when you report the results of single play games in which 'too many' people cooperate. The usual proposed 'fix' is to do repeat single protocols in which distinct pairs play on each trial, and apply a model of learning to play non-cooperatively. (See McCabe, Rassenti and Smith, 1998, for a trust game with the option of punishing defection in which support for the cooperative outcome does not decrease in repeat single relative to single play across trials, and therefore subjects *do not* 'learn' to play non-cooperatively.) But there are many unanswered qsts implicit in this auxiliary hypothesis: since repeat single protocols require large groups of subjects (20 subjects to produce a 10-trial sequence), have any of these games been run long enough to measure adequately the extent of learning? In single play two-person anonymous trust games data have been reported showing that group size matters; that is, it makes a difference whether you are running 12 subjects (6 pairs) or 24 subjects (12 pairs) simultaneously in a session (Burnham, McCabe and Smith, 2000). Also, in the larger groups pairs were found to be less trusting than in the small groups – perhaps not too surprising. But in repeat single games, in which a game is repeated with distinct pairs of subjects on each repetition, larger groups are needed for longer trial sequences. Hence, learning and group size as auxiliary hypotheses loses independence, and we have knotty new problems of complex joint hypothesis testing. The techniques, procedures and protocol tests we fashion for solving such problems are the sources of our experimental knowledge. All testing depends on, and is limited by, the state of that experimental knowledge at any given time. Over time it expands incrementally in the design problem-solving context of particular new testing challenges. This is a community development enterprise that is largely outside individual conscious awareness, but an integral part of the sociality of scientific change.

3. *Instructions are adequate* (or decisions are robust with respect to instructions, and so on). What does it mean for instruction to be adequate? Clear? If so, clear about what? What does it mean to say that subjects understand the instructions? Usually this is interpreted to mean that they can perform the task, which is judged by how they answer qsts about what they are to do. In two-person interactions, instructions often matter so much that they must be considered a (powerful) treatment. (Thus, Hertwig and Ortmann, 2001, section 2, argue that scripts – instructions – are important for replication, and that 'ad-libbing' should be avoided.) Instructions can be important because they define context, and context matters. Ultimatum and dictator game experiments yield statistically and economically significant differences in results due to differing instructions and protocols (Hoffman et al., 1994; Hoffman, McCabe and Smith, 1996).

Positive economics: judge theories by their predictions not their assumptions

There is a methodological perspective associated with Milton Friedman (1953), which fails to provide an adequate foundation for experimental (field or laboratory) science, but which influenced economists for decades and still has some currency. Friedman's proposition is that the truth value of a theory should be judged by its testable and tested predictions not by its assumptions. This proposition is deficient for at least three reasons:

1. 1. If a theory fails a test, we should ask why, not always just move on to seek funding for a different new project; obviously, one or more of its assumptions may be wrong, and it behoves us to design experiments that will probe conjectures about which assumptions failed. Thus, if first price auction theory fails a test is it a consequence of the failure of one of the axioms of expected utility theory, for example, the compound lottery axiom? If a subgame perfect equilibrium prediction fails, does the theory fail because the subjects do not satisfy the assumption that the agents choose dominant strategies? Or did the subjects fail to use backward induction? Or was it none of the above because the theory was irrelevant to how some motivated agents solve such problems in a world of bilateral (or multilateral) reciprocity in social exchange? When a theory fails there is no more important question to ask than what it is about the theory that has failed.
2. 2. Theories may have the if-and-only if property that one (or more) of their 'assumptions' can be derived as implication(s) from one (or more) of their 'results'. These cases if trivial lead to the reversible property of testing illustrated above for risk-neutral agents bidding in first price auctions with linear density functions on value.
3. 3. If a theory passes one or more tests, this need not be because its assumptions are correct. A subject may choose the subgame perfect equilibrium because she believes she is paired with a person that is not trustworthy, and not because she always chooses dominant strategies, or assumes that others always so choose or that this is common knowledge. This is why you are not done when a theory is corroborated by a test. You have examined only one point in the parameter space of payoffs, subjects, tree structure, and so on. Your results may be a freak accident of nature due to a complex of suitabilities or in any case may have other explanations.

In view of Proposition 2, what are experimentalists and theorists to do?

Consider first the example in which we have a possible failure of A: rewards are adequate to motivate subjects. Experimentalists should do what comes naturally, namely, do new experiments that increase rewards and/or lower decision costs by simplifying experiment procedures. The literature is full of examples (for surveys and inpts see Smith and Walker,

1993; Camerer and Hogarth, 1999; Holt and Laury, 2001; Harrison, McInnes and Rutstroem, 2005).

Theorists should ask whether the theory is extendable to include *A*, so that the effect of payoffs is explicitly modelled. It is something of a minor scandal when economists – whose models predict the optimal outcome independently of payoff levels, and however gently rounded the payoff in the neighbourhood of the optimum is – object to an experiment because the payoffs are inadequate. What is adequate? Why are payoff inadequacies a complaint rather than a spur to better and more relevant modelling of payoffs? A step toward modelling both H and A (as payoffs) is provided in Smith and Szidarovszky (2004). Economic intuition tells us that payoffs should matter. But if they do, it must mean that some cost, which is impeding equilibrium, is being overcome at the margin by higher payoffs. The natural psychological assumption is that there are cognitive costs of getting the best outcome, and more reward enables higher cognitive cost to be incurred to increase net return.

Generally, both groups must be aware that for any H, any A and any experiment, one can say that if the outcome of the experiment rejects (H|A), then both should assume that either H or A may be false, which is an obvious corollary to Proposition 2. This was Einstein's position concerning the Kaufmann results, and was the correct position, whatever later tests might show. After every test, either the theory (leading to H) is in error or the circumstances of the test, A, is in error.

The experimentalist has much to do, but primarily more experiments, which is precisely what experimentalists do in response to the many conjectures about what is wrong with the experiment – re-examine the instructions, payoffs, subjects, anything and everything the experimentalist did to formulate the test.

The theorist should also ask, especially if further tests continue to reject (H|A), whether the auxiliary hypothesis can be incorporated into the theory.

If the outcome fails to reject (H|A), the experimentalist should escalate the severity of the test. At some point does H fail? This identifies the limits of the theory's validity, and gives the theorist clues for modifying the theory.

Experimental knowledge drives our methods

Philosophers have written eloquently and argued intently over the implications of D–Q and related issues for the interpretation of science. Popper tried to demarcate science from pseudoscience with the basic rule requiring the scientist to specify in advance the conditions under which he will give up his theory. This is a variation on the idea of a so-called ‘crucial experiment’, a concept which cannot withstand examination (Lakatos, 1978, vol. 1, pp. 3–4, 146–8), as is evident from our Proposition 2.

The failure of all programmes attempting to articulate a defensible rational science of scientific method has bred postmodern negative reactions to science and scientists. These exercises and controversies make fascinating reading, and provide a window on the social context of science, but I believe they miss the essence of what is most important in the working lives of all practitioners. Popper was wrong in thinking he could demarcate science from pseudoscience by an exercise in logic, but that does not imply that the Popperian falsification rule failed as a milestone contribution to the scientific conversation; nor does it mean that ‘anything goes’ (Feyeraband, 1975). Rather, what one can say is much less open ended: anything goes only in so far as what can be concluded about *constructive rationality in science*. But the scientific enterprise is also about *ecological rationality in science*, which is about discovery, about probes into Max Born's ‘jungle’, about thinking outside the box and, as I shall argue below, about the technology of observation in science that renders obsolete long-standing D–Q problems while introducing new ones for a time.

You do not have to know anything about D–Q and statements like Proposition 2 to appreciate that the results of an experiment nearly always suggest new qsts precisely because the interpretation of results in terms of the theory are commonly *ambiguous*. This ambiguity is reflected in the discussion whenever new results are presented at seminars and conferences. Without ambiguity there would be nothing to discuss. What is the response to this ambiguity? Invariably, if it is a matter of consequence, experimentalists design new experiments with the intention of confronting the issues in the controversy, and in the conflicting views that have arisen in interpreting the previous results. This leads to new experimental knowledge of how results are influenced, or not, by changes in procedures, context, instructions and control protocols. The new knowledge may include new techniques that have application to areas other than the initiating circumstance. This ecological process is driven by the D–Q problem, but practitioners need have no knowledge of the philosophy of science literature to take the right next steps, subject to error, in the laboratory.

This is because the theory or primary model that motivates the qsts tells you nothing definitive or even very useful about precisely how to construct tests. Tests are based on extra theoretical intuition, conjectures, and experiential knowledge of procedures. The context, subjects, instructions, parameterization, and so on are determined outside the theory, and their evolution constitutes the experimental knowledge that defines our methodology. The forms taken by the totality of individual research testing programmes cannot be accurately described in terms of the rhetoric of falsification, no matter how much we speak of the need for theories to be falsifiable, stating in advance the conditions under which the theory will be rejected, the tests discriminating or ‘crucial’ and the results robust.

Whenever negative experimental results threaten perceived important new theoretical tests, the resulting controversies galvanize experimentalists into a search for different or better tests – tests that examine the robustness of the original results. Hence, Kaufmann's experimental apparatus was greatly improved by Bucherer, although there was no question about Kaufmann's skill and competence in laboratory technique. The point is that with escalated personal incentives, and a fresh perspective, scientists found improved techniques. This scenario is common as new challenges bring forth renewed effort. This process generates the constantly changing body of experimental and observational knowledge whose insights in solving one problem often carry over to applications in many others.

Just as often experimental knowledge is generated from curiosity about the properties of phenomena that we observe long before a body of theory exists that deals specifically with

the phenomenon at issue. An example in experimental economics is the continuous double auction trading mechanism (Smith, 1962; 2003).

An example from physics is Brownian motion, discovered by the botanist Robert Brown in 1827, who first observed the unpredictable motion of small particles suspended in a fluid. This motion is what keeps them from sinking under gravity. This was 78 years before Einstein's famous paper (one of three) in 1905 developed the molecular kinetic theory that was able to account for it, although he did not know that the applicable observations of Brownian motion were already long familiar (see Mayo, 1996, ch. 7, for references and details).

The long ‘inquiry into the cause of Brownian motion has been a story of hundreds of experiments … [testing hypotheses attributing the motion]…either to the nature of the particle studied or to the various factors external to the liquid medium … ‘(Mayo, 1996, pp. 217–18). The essential point is ‘that these early experiments on the possible cause of Brownian motion were not testing any full-fledged theories. Indeed it was not yet known whether Brownian motion would turn out to be a problem in chemistry, biology, physics, or something else. Nevertheless, a lot of information was turned up and put to good use by those later researchers who studied their Brownian motion experimental kits’ (Mayo, 1996, p. 240). The problem was finally solved by drawing on the extensive bag of experimental tricks, tools and past mistakes that constitute ‘a log of the extant experimental knowledge of the phenomena in qst’ (1996, p. 240).

Again, ‘the infinitely many alternatives really fall into a few categories. Experimental methods (for answering new qsts) coupled with experimental knowledge (for using techniques and information already learned) enable local qsts to be split off and answered’ (Mayo, 1996, p. 242).

The bottom line is that good-enough solutions emerge to the baffling infinity of possibilities, as new measuring systems emerge, experimental toolkits are updated, and understanding is sharpened. This bottom line also goes far towards writing the history of experimental economics and its many detailed encounters with data, and the inevitable ambiguity of subsequent inpt. And in most cases the jury remains in session on whether we are dealing with a problem in psychology (perception), economics (opportunity cost and strategy), social psychology (equality, equity or reciprocity), neuroscience (functional imaging and brain modeling) or all of the above. So be it.

The machine builders

Mayo's (1996) discussion and examples of experimental knowledge leave unexamined the question of how technology affects the experimentalist's toolkit. The heroes of science are neither the theorists nor the experimentalists but the unsung tinkers, mechanics, inventors and engineers who create the new generation of machines that make obsolete yesterday's observations and heated arguments over whether it is T or A that has been falsified. Scientists, of course, are sometimes a part of this creative destruction, but what is remembered in academic recognition is the new scientific knowledge they created, not the instruments they invented that made possible the new knowledge. Michael Faraday, ‘one of the greatest physicists of all time’ (Segrè, 1984, p. 134), had no formal scientific education. He was a bookbinder, who had the habit of reading the books that he bound. He was pre-eminently a tinker for whom ‘some pieces of wood, some wire and some pieces of iron seemed to suffice him for making the greatest discoveries’ (quoted from a letter by Helmholz in Segrè, 1984, p. 140). Yet he revolutionized how physicists thought about electromagnetic phenomena, invented the concept of lines of force (fields), and inspired Maxwell's theoretical contributions. ‘He writes many times that he must experience new phenomena by repeating the experiments, and that reading is totally insufficient for him’ (Segrè, 1984, p. 141).

This is what I mean, herein, when I use the term ‘experimental knowledge’. It is ‘can do’ knowledge acquired by trial, error and discovery. And it is what Mayo (1996) is talking about. It is also why doing experiments changed the way I thought about economics.

Technology and science

With the first moon landing, theories of the origin and composition of our lunar satellite, contingent on the state of existing indirect evidence, were upstaged by direct observation; the first Saturn probe sent theorists back to their desks and computers to re-evaluate her mysterious rings, whose layered richness had not been anticipated. Similar experiences have followed the results of ice core sampling in Greenland, and instrumentation for mapping the genome of any species. Galileo's primitive telescope opened a startling window on the solar system, as do Roger Angel's multiple mirror light-gathering machines (created under the Arizona football stadium) that open the mind to a view of the structure of the universe. (For a brief summary of the impact of past, current, and likely future effects of rapid change in optical and infrared (terrestrial and space) telescopes on astronomy see Angel, 2001.)

The technique of tree ring dating, invented by an Arizona astronomer, has revolutionized the interpretation of archeological data from the last 5,000 years.

Yesterday's reductionisms, shunned by mainstream ‘practical’ scientists, create the demand for new deeper observations, and hence for the machines that can deliver answers to entirely new qsts. Each new machine – microscope, telescope, Teletype, computer, the Internet, fMRI imaging – changes the way teams of users think about their science. The host of auxiliary hypotheses needed to give meaning to theory in the context of remote and indirect observations (inferring the structure of Saturn's ring from earth-based telescopes) are suddenly made irrelevant by deep and more direct observations of the underlying phenomena (fly by computer-enhanced photos). It's the machines that drive the new theory, hypotheses, and testing programmes that take you from atoms, to protons, to quarks. Yet with each new advance comes a blizzard of auxiliary hypotheses, all handcuffed to new theory, giving expression to new controversies seeking to rescue T and reject A, or to accept A and reject T.

Experimental economics and computer/communication technology

In 1976 when Arlington Williams (1980) created the first electronic double-auction software program for market experiments, all of us thought we were simply making it easier to run experiments, collect more accurate data, observe longer time series, facilitate data analysis, and so on. What were computerized were the procedures, recording and accounting that heretofore had been done manually. No one was anticipating how this tool might impact on and change the way we thought about the nature of doing experiments. But with each new electronic experiment we were ‘learning’ (affected by, but without conscious awareness of) the fact that traders could be matched at essentially zero cost, that the set of feasible rules that could be considered was no longer restricted by costly forms of implementation and monitoring, that vastly larger message spaces could be accommodated, and that optimization algorithms could now be applied to the messages to define new electronic market forms for trading energy, air emission permits, water and other network industries. In short, the transaction cost of running experimental markets became minuscule in comparison with the pre-electronic days, and this opened up new directions that previously had been unthinkable.

This quickly led to the concept of smart computer-assisted markets, which appeared in the early 1980s (Rassenti, 1981; Rassenti, Smith and Bulfin, 1982), extended conceptually to electric power and gas pipelines in the late 1980s (Rassenti and Smith, 1986; McCabe, Rassenti and Smith, 1989), with practical applications to electric power networks and the trading of emission permits across time and regions in the 1990s (Rassenti, Smith and Wilson, 2002). These developments continue as major new efforts in which the laboratory is used as a test bed for measuring, modifying, and further testing the performance characteristics of new institutional forms.

What is called e-commerce has spawned a rush to reproduce on the Internet the auction, retailing and other trading systems people know about from past experience. But the new experience of being able to match traders at practically zero cost is sure to change how people think about trade and commerce, and ultimately this will change the very nature of trading institutions. In the short run, of course, efforts to protect existing institutions will spawn efforts to shield them from entry by deliberately introducing cost barriers, but in the long run these efforts will be increasingly uneconomical.

Neuroscience carries the vision of changing the experimental study of individual, two-person interactive and market decision making. The neural correlates of decision making, how it is affected by rewards, cognitive constraints, working memory, institutions, repeat experience and a host of factors that in the past we could neither control or observe can in the future be expected to become an integral part of the way we think about and model decision making. Models of decision, now driven by game and utility theory, and based on trivial, patently false, models of mind, must take account of new models of cognitive, calculation and memory properties of mental function that are accessible to more direct observational inpt. Game-theoretic models assume consciously calculating, rational mental processes, but models of mind include non-self-aware processes just as accessible to neural brain imaging as the conscious. For the first time we may be able to give some observational content to the vague and slippery idea of ‘bounded rationality’ (see Camerer, Loewenstein and Prelec, 2005).

Conclusion

In principle the D–Q problem is a barrier to any defensible notion of a rational science that selects theories by a logical process of confrontation with scientific evidence. This is cause for joy not despair. Think how dull a life of science would be if, once we were trained, all we had to do was to turn on the threshing machine of science, feed it the facts, send its output to the printer, and run it through the formulas for writing a scientific paper.

As I see it, there is no rationally constructed science of scientific method. The attempt to do it has led to important insights and understanding, and has been a valuable exercise. But all construction must ultimately pass ecological or ‘fitness’ tests based on the totality of our experience. Control is of course important; it is why we do laboratory and field experiments. But control is always limited in scope, and above all the rhetoric of control should not restrict the examination and re-examination of our own assumptions, both in the theory and in its testing, or limit our capacity to think outside the professional box. We do this in the reality underneath our rhetoric because we cannot help it, so much is it part of our deep human sociality and the workings of our social brains.

See Also

- experimental economics, history of

Bibliography

Angel, R. 2001. Future optical and infrared telescopes. *Nature Insights* 409, 427–30.

Banks, J., Olson, M., Porter, D., Rassenti, S. and Smith, V. 2003. Theory, experiment and the Federal Communications Commission spectrum auctions. *Journal of Economic Behavior & Organization* 51, 303–50.

Born, M. 1943. *Experiment and Theory in Physics*. Cambridge: Cambridge University Press.

Burnham, T., McCabe, K. and Smith, V. 2000. Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization* 43, 57–73.

Camerer, C.F. and Hogarth, R.M. 1999. The effects of financial incentives in experiments: a review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19, 7–42.

Camerer, C., Loewenstein, G. and Prelec, D. 2005. Neuroeconomics: how neuroeconomics can inform economics. *Journal of Economic Literature* 43, 9–64.

Cox, J., Smith, V. and Walker, J. 1988. Theory and individual behavior of first price auctions. *Journal of Risk and Uncertainty* 1, 61–99. Reprinted in Smith (1991).

Davis, D. and Holt, C. 1993. *Experimental Economics*. Princeton: Princeton University Press.

Davis, D. and Williams, A. 1991. The Hayek hypothesis in experimental auctions: institutional effects and market power. *Economic Inquiry* 29, 261–74.

Einstein, A. 1905. On the electrodynamics of moving bodies. In *The Collected Papers of Albert Einstein*, vol. 2, trans. A. Beck, Princeton: Princeton University Press, 1989.

Einstein, A. 1907. On the relativity principle and the conclusions drawn from it. In *The Collected Papers of Albert Einstein*, vol. 2, trans. A. Beck, Princeton: Princeton University Press, 1989.

Einstein, A. 1934. *The World as I see it*. New York: Covici Friede Publishers.

Feyeraband, P. 1975. *Against Method*. London: Versa.

Fölsing, A. 1997. *Albert Einstein*. New York: Viking.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Grandin, T. and Johnson, C. 2005. *Animals in Translation*. New York: Scriber.

Guala, F. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Harrison, G., McInnes, J. and Rutstroem, L. 2005. Risk aversion and incentive effects: comment. *American Economic Review* 95, 897–901.

Hayek, F.A. 1942. Scientism and the study of society. *Economica*; reprinted in *The Counter-Revolution in Science*. Indianapolis: Liberty Press, 1979.

Hayek, F.A. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

Hayek, F.A. 1967. *Studies in Philosophy, Politics and Economics*. London: Routledge & Kegan Paul.

Hayek, F.A. 1978. Competition as a discovery procedure. In *New Studies in Philosophy, Politics, Economics, and the History*. Chicago: Chicago University Press. Reprinted in *The*

Essence of Hayek. Stanford: Hoover Institution Press, 1984.

Hertwig, R. and Ortmann, A. 2001. Experimental practices in economics. *Behavioral and Brain Sciences* 24, 383–451.

Hoffman, E., McCabe, K. and Smith, V. 1996. On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory* 25, 289–301.

Hoffman, E., McCabe, K., Shachat, K. and Smith, V. 1994. Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior* 7, 346–80.

Holt, C. 1989. The exercise of market power in experiments. *Journal of Law and Economics* 32S, 107–30.

Holt, C., Langan, L. and Villamil, A. 1986. Market power in oral double auction experiments. *Economic Inquiry* 24, 107–23.

Holt, C. and Laury, S. 2001. Varying the scale of financial incentives under real and hypothetical conditions. *Behavioral and Brain Sciences* 24, 417–18.

Kagel, J. and Roth, A. 1995. *Handbook of Experimental Economics*. Princeton: Princeton University Press.

Kahneman, D., Knetsch, J. and Thaler, R. 1986. Fairness as a constraint on profit seeking: entitlements in the market. *American Economic Review* 76, 728–41.

Lakatos, I. 1978. *The Methodology of Scientific Research Programmers*, 2 vols. Cambridge: Cambridge University Press.

Leamer, E. 1978. *Specification Searches*. New York: Wiley.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

McCabe, K., Rassenti, S. and Smith, V.L. 1989. Designing smart computer assisted markets for gas networks. *European Journal of Political Economy* 5, 259–83.

McCabe, K., Rassenti, S. and Smith, V.L. 1998. Reciprocity, trust and payoff privacy in extensive form bargaining. *Games and Economic Behavior* 24, 10–24. Reprinted in Smith (2000).

McCabe, K. and Smith, V. 2000. A comparison of naïve and sophisticated subject behavior with game theoretic prediction. *Proceedings National Academy of Sciences* 97, 3777–81.

Northrup, F.S.C. 1969. Einstein's conception of science. In *Albert Einstein: Philosopher–Scientist*, ed. P.A. Schilpp. LaSalle, IL: Open Court.

Pirsig, R.M. 1981. *Zen and the Art of Motorcycle Maintenance*. New York: Banton Books.

Rassenti, S. 1981. O-1 decision problems with multiple resource constrains: algorithms and applications. Ph.D. thesis, University of Arizona.

Rassenti, S. and Smith, V. 1986. Electric utility deregulation. In *Pricing Electric Gas and Telecommunication Services*. Washington, DC: Institute for the Study of Regulation.

Rassenti, S., Smith, V. and Bulfin, R. 1982. A combinatorial auction mechanism for airport time slot allocation. *Bell Journal of Economics* 13, 402–17.

Rassenti, S., Smith, V. and Wilson, B. 2002. Using experiments to inform the privatization/deregulation movement in electricity. *Cato Journal* 21, 515–44.

Robinson, J. 1979. What are the qsts? *Journal of Economic Literature* 15, 1318–39.

Samuelson, P. 1966. Intertemporal price equilibrium: a prologue to the theory of speculation. In *The Collected Papers of Paul A. Samuelson*, vol. 2, ed. J. Stiglitz. Cambridge, MA: MIT Press.

Samuelson, P. and Nordhaus, W. 1985. *Economics*. New York: McGraw-Hill.

Segrè, E. 1984. *From Falling Bodies to Radio Waves*. New York: Freeman.

Shubik, M. 1959. *Strategy and Market Structure*. New York: Wiley.

Smith, V.L. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70, 111–37.

Smith, V.L. 1965. Experimental auction markets and the Walrasian hypothesis. *Journal of Political Economy* 73, 387–93.

Smith, V.L. 1976. Experimental economics: induced value theory. *American Economic Review Proceedings* 66, 274–79. Reprinted in Smith (1980; 1991).

Smith, V.L. 1980. *Evaluation of Econometric Models*. New York: Academic Press.

Smith, V.L. 1991. *Papers in Experimental Economics*. New York: Cambridge University Press.

Smith, V.L. 2000. *Bargaining and Market Behavior. Essays in Experimental Economics*. New York: Cambridge University Press.

Smith, V. 2003. Constructivist and ecological rationality in economics. *American Economic Review* 93, 465–508.

Smith, V., McCabe, K. and Rassenti, S. 1991. Lakatos and experimental economics. In *Appraising Economic Theories*, ed. N. De Marchi and M. Blaug. London: Edward Elgar.

Smith, V.L. and Szidarovszky, F. 2004. Monetary rewards and decision cost in strategic interactions. In *Models of a Man: Essays in Memory of Herbert A. Simon*, ed. M. Augier and J. March. Cambridge MA: MIT Press.

Smith, V. and Walker, J. 1993. Monetary rewards and decision cost in experimental economics. *Economic Inquiry* 31, 245–61. Best Article Award, Western Economic Association, 1993.

Soberg, M. 2005. The Duhem–Quine thesis and experimental economics: a reinpt. *Journal of Economic Methodology* 12, 581–97.

Stigler, G. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65, 1–17.

Suppes, P. 1969. Models of data. In *Studies in the Methodology and Foundations of Science*. Dordrecht: Reidel.

Williams, A. 1980. Computerized double-auction markets: some initial experimental results. *Journal of Business* 53, 235–57.

How to cite this article

Smith, Vernon L. "experimental economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000277>

doi:10.1057/9780230226203.0527

The New Palgrave Dictionary of Economics Online

experimental labour economics

Armin Falk and Simon Gächter

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Experimental labour economics uses experimental techniques to improve our understanding of labour economics issues. We start by putting experimental data into perspective with the data-sets typically used by empirical labour economists. We then discuss several examples of how experiments can inform labour economics.

Keywords

double auction; efficiency wages; employment relation; experimental labour economics; gift exchange; implicit incentives; incomplete contracts; involuntary unemployment; laboratory experiments; labour economics; labour market institutions; minimum wages; moral hazard; opportunistic behaviour; performance incentives; piece rates; repeated games; reservation wage; tournaments; wage rigidity

Article

Scientific progress relies on testing theories. In labour economics different data sources are available for performing such tests. An important distinction is between circumstantial data and experimental or questionnaire data. Circumstantial data is the by-product of uncontrolled, naturally occurring economic activity. In contrast, experimental data is created explicitly for scientific purposes under controlled conditions. In labour economics, the data most commonly and traditionally used is circumstantial data such as unemployment rates or data on wages, education, or income, complemented by survey data. Labour economists have only recently started to use laboratory experiments.

Laboratory experiments have several important advantages in comparison with data sets typically used in labour economics. A key advantage is the unparalleled opportunity to control crucial aspects of the economic environment. This includes control over information conditions, technology, market structure, and trends in economic fundamentals. Control over the decision environment makes it possible to identify the theoretical equilibrium in an experimental labour market, which is basically impossible with

field data. Knowing the equilibrium allows the study of convergence properties, stability and efficiency. Experiments are particularly useful for investigating the economic consequences of important labour market institutions, such as minimum wages or employment protection legislations. The reason is that experiments allow the exogenous changes of institutions, holding everything else constant. In the field, by contrast, institutions are always adopted endogenously. Econometric strategies such as instrumenting for policy changes with political variables can help ameliorate this problem, but do not achieve the unequivocal exogenous variation provided by a laboratory experiment. Laboratory experiments also make it possible to observe behaviour at the level of individual economic agents. This is important given that theoretical predictions typically involve such micro behaviours. For example, it is possible to directly observe individual reservation wages or individual wage bargaining behaviour. Yet another advantage is that with laboratory experiments one can study, at relatively low cost, institutions that do not yet exist. Analogous to experimental tests of new medicines, where the medication is administered to a small subset of the population initially, laboratory experiments can be used as a first step, before experimenting with institutions in the field. Finally, experimental evidence is replicable, which is a prerequisite in establishing solid empirical knowledge.

Data-sets and the comparative advantage of laboratory experiments

Although we believe that laboratory experiments offer important advantages for studying institutions, and should thus be exploited more often, it is important to recognize that there are also drawbacks to this method, which calls for a complementary use of different methods. A potential disadvantage is limited generalizability. Note, however, that this critique holds with respect to any data-set, given that any empirical observation is time and space contingent. Another concern is that experiments may be overly simple, missing potentially relevant aspects of the labour market. This is in fact both a problem and an advantage of experiments. Just as economic models are simpler than reality, so experiments are designed to simplify as much as possible, without losing the essentials. Thus, simplicity need not be a defect of an experiment. The key challenge, just as in the case of building economic models, is to include those features that are essential to the question at hand.

Examples

In this section we discuss a selected set of examples of experiments that were designed to shed light on important issues in labour economics. The examples concern the nature of the employment relationship and its contractual regulation, wage rigidity, performance incentives and their potentially detrimental effects, and labour market institutions.

The employment relation

The employment relation is an incomplete contract, which typically leaves many important aspects unspecified. This holds in particular for the content of work effort, which is unregulated and thereby non-enforceable by third parties. Contractual incompleteness gives opportunistic agents an incentive to shirk and therefore leads to an inefficiently low surplus. Thus, voluntary cooperation is necessary to ensure

efficiency. Akerlof (1982) argued that many employment relationships are therefore governed by a gift exchange: the firm pays a higher wage than necessary to keep the employee, and the employee returns the gift by providing above minimum effort. Akerlof supported his arguments by a case study and casual observations.

The gift-exchange game by Fehr, Kirchsteiger and Riedl (1993) provided the first experimental test of the existence of gift exchanges in the framework of a formal game-theoretic model designed to mimic an incomplete employment contract. In their experiment, participants assumed the roles of ‘workers’ and ‘firms’. A firm made a wage offer that a worker could accept. If the worker accepted, he or she had then to choose a costly effort level. Parameters were such that a self-interested worker would always choose the lowest possible effort, since effort was costly. In turn, the firm had no incentive to pay an above-minimal wage, because a self-interested worker would shirk anyway. The results of numerous experiments in this framework showed, however, that wages and effort levels are positively correlated. Higher wages were reciprocated by higher effort levels, a finding which is consistent with the gift-exchange argument by Akerlof. This observation is also consistent with field evidence regarding the link between personnel policy and work morale (Bewley, 1999).

In these experiments the employment relationship was modelled as a one-shot game, because this allows an unambiguous prediction under the joint assumptions of rationality and self-interest. Yet in reality, employment relationships are long-term relationships. To test the impact of repeated interaction, Gächter and Falk (2002) conducted the gift-exchange experiment in the form of repeated games in which the same firm–worker pair interacted for ten periods. These repeated games were compared with one-shot games in which each firm was matched with ten different workers. The results showed a significantly higher effort in the repeated game than in the one-shot games. Gächter and Falk showed that the reason for this result is that in the repeated games the selfish types imitate the reciprocal types. This result provides support for theoretical arguments (for example, MacLeod and Malcomson, 1998) that incomplete employment relations allow for implicit incentives for non-opportunistic behaviour.

In the experiments by Gächter and Falk (2002) the experimenter determined the duration of the employment relationship exogenously. In reality, however, the duration of employment relationships arises endogenously. Contract theory suggests that the duration might be linked to contractual incompleteness. Specifically, when contracts are incomplete, a long-term relationship provides implicit incentives that constrain opportunistic behaviour – an argument supported by the cited experimental evidence. If contracts are complete then implicit incentives are not necessary to constrain opportunism. Thus, employment relationships will tend to be short term under contractual completeness. Brown, Falk and Fehr (2004) tested these arguments experimentally and found strong support for them.

Efficiency wages, wage rigidity, and involuntary unemployment

Efficiency wage theories explain why even in the absence of market interventions wages might be downwardly rigid, causing involuntary unemployment. Akerlof's (1982) gift-exchange theory is one efficiency wage theory that can explain involuntary unemployment. The main idea is simple. If gift exchanges exist, then firms have no incentive to lower wages because this would lead to low performance. Thus, paying high wages is profitable to the firm – wages are downwardly rigid and can cause involuntary unemployment. Fehr et al. (1998) demonstrated the behavioural validity of this

argument experimentally.

Fehr and Falk (1999) provide the most stringent confirmation that gift exchanges can lead to downward wage rigidity. In their experiment an employment relationship was embedded in a ‘double auction’ market institution in which there were more workers than firms. This institution is known for its competitive properties; under complete contracts experimental double auction markets tend to clear very quickly. In the Fehr–Falk experiments both workers and firms could make wage offers. This enables us to observe whether workers underbid each other and firms therefore have the possibility of employing a worker at a low wage. There was indeed fierce competition among workers who underbid each other down to the theoretically predicted wage. Underbidding occurred in both treatments, the ‘complete contract treatment’ and the ‘incomplete contract treatment’. In the latter, the striking finding was that firms did not take advantage of the possibility of paying low wages; instead they deliberately paid very high wages. The workers’ reciprocal effort choice explains why firms had an incentive to pay high wages. In the control experiments with complete contracts gift exchanges were precluded by design and actual wages were very close to market clearing wages. Thus, incomplete contracts and gift exchange can explain wage rigidity and involuntary unemployment.

Performance incentives (and their detrimental effects)

Compensation and performance incentives have always been central topics in labour economics. Compensation may take different forms. The simplest form is a piece rate where a worker receives a certain wage for each unit she produces. Compensation may also depend on relative performance and be coupled with the possibility of moving up the career ladder. Tournament theory (Lazear and Rosen, 1981) is an important theoretical framework for understanding career incentives and relative performance incentives.

Bull, Schotter and Weigelt (1987) provide the first experimental analysis of piece rate and tournament incentives. They designed their experiments so that the incentive schemes were directly comparable, that is, the predicted effort level was the same both under piece rates and under tournament incentives. The results confirmed the theoretical predictions in both treatments. As it turned out, however, the support for tournament theory is weaker than for piece rate theory. In various treatment conditions these authors find that average effort choices converged close to the equilibrium prediction, but the variance was up to 30 times higher under tournament incentives than under the piece rate system.

The results by Bull, Schotter and Weigelt (1987) provide clear evidence that incentives influence behaviour very strongly. However, numerous experiments as well as field evidence (Bewley, 1999) suggest that employment relationships are also governed by ‘good will’ and voluntary cooperation. This raises the question how explicit performance incentives affect voluntary cooperation – a fertile area of current research in experimental labour economics. A nice illustration of the potentially dysfunctional effects of introducing explicit incentives is the field experiment by Gneezy and Rustichini (2000). These authors studied the parents’ response to the introduction of a fixed fine for picking up their children too late from kindergarten. The experiment lasted for 20 weeks and there were two conditions. In the baseline condition no fine existed. In the treatment condition the experimenters implemented a fixed fine after week four for picking up a child too late. The fine was removed after week 16. From week seven onwards, there was a steep *increase* in the number of latecomers until their number was roughly twice as

high as in the baseline condition. Moreover, when the fine was removed at the end of week 16 the number of tardy parents remained roughly twice as high as in the baseline condition. This result clearly contradicts standard incentive theory, which predicts that the introduction of the fine should lower the incidence of late coming. A likely explanation of this finding is that the implicit contract that governed the employment relationship was changed from a good-will-based one to a market-like transaction, in which ‘a fine is a price’ and parents bought the commodity of being late.

Labour market institutions

A particularly important advantage of laboratory experiments concerns the possibility to test the economic effects of (labour market) institutions in a controlled way. An example of such an institutional test is the paper on minimum wages by Falk, Fehr and Zehnder (2006). In their experiment firms make wage offers to workers in labour markets either with or without minimum wages. The key insight of their study is that minimum wages may affect the reservation wages of workers in a non-trivial way: first, when minimum wages are introduced, workers stipulate reservation wages above the level of the minimum wage level, because being paid at just the level of the minimum wage is considered unfair. Second, while the introduction of a minimum wage increases reservation wages, the removal of a minimum wage legislation changes reservation wages only marginally. These findings help explain several empirical minimum wage puzzles. First, there exists an anomalously low utilization of sub-minimum wages in situations where employers actually could pay workers less than the minimum; second there exist so-called spillover effects, that is, wages are often increased by an amount in excess of that necessary for compliance with the minimum wage; and third, minimum wages do not always cause a decrease in employment, in particular if the minimum wage increase is modest (see also Card and Krueger, 1995).

The finding that minimum wages affect workers’ fairness perceptions of wages is also supported by Brandts and Charness (2004) who introduced a minimum wage in the context of an experimental labour market with worker moral hazard where workers’ fairness concerns drive effort. They show that workers provide less effort for the same wage level in the presence of the minimum wage. This supports the view that the impact of minimum wages on workers’ attributions of fairness intentions to firms partially shapes their effort responses.

Concluding remarks

Experimental economics is a method of empirical investigation, not a separate subfield of economics. Experimental methods can therefore in principle be utilized in all areas of economics. In this article we have illustrated some selected applications of experimental methods to important issues in labour economics. Further discussions of the issues raised here can be found in Fehr and Gächter (2000), Gächter and Fehr (2002), Fehr and Falk (2002), Falk and Fehr (2003) and Falk and Huffman (2007).

See Also

- behavioural economics and game theory
- incomplete contracts
- institutional economics
- labour economics
- personnel economics
- reciprocity and collective action

Bibliography

Akerlof, G. 1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97, 543–69.

Bewley, T. 1999. *Why Wages Don't Fall in a Recession*. Cambridge, MA: Harvard University Press.

Brandts, J. and Charness, G. 2004. Do labour market conditions affect gift exchange? Some experimental evidence. *Economic Journal* 114, 684–708.

Brown, M., Falk, A. and Fehr, E. 2004. Relational contracts and the nature of market interactions. *Econometrica* 72, 747–80.

Bull, C., Schotter, A. and Weigelt, K. 1987. Tournaments and piece rates: an experimental study. *Journal of Political Economy* 95, 1–33.

Card, D. and Krueger, A. 1995. *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton, NJ: Princeton University Press.

Falk, A. and Fehr, E. 2003. Why labor market experiments? *Labor Economics* 10, 399–406.

Falk, A., Fehr, E. and Zehnder, C. 2006. Fairness perceptions and reservation wages: the behavioral effects of minimum wage laws. *Quarterly Journal of Economics* 121, 1347–81.

Falk, A. and Huffman, D. 2007. Studying labor market institutions in the lab: minimum wages, employment protection and workfare. *Journal of Institutional and Theoretical Economics* 163, 30–45.

Fehr, E. and Falk, A. 1999. Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy* 107, 106–34.

Fehr, E. and Falk, A. 2002. The psychological foundations of incentives. *European Economic Review* 46, 687–724.

Fehr, E. and Gächter, S. 2000. Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives* 14(3), 159–81.

Fehr, E., Kirchler, E., Weichbold, A. and Gächter, S. 1998. When social norms overpower competition – gift exchange in experimental labor markets. *Journal of Labor Economics* 16, 324–51.

Fehr, E., Kirchsteiger, G. and Riedl, A. 1993. Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108, 437–60.

Gächter, S. and Falk, A. 2002. Reputation and reciprocity: consequences for the labour relation. *Scandinavian Journal of Economics* 104, 1–27.

Gächter, S. and Fehr, E. 2002. Fairness in the labour market: a survey of experimental results. In *Surveys in Experimental Economics. Bargaining, Cooperation and Election Stock Markets*, ed. B. Friedel and M. Lehmann-Waffenschmidt. Heidelberg and New York: Physica.

Gneezy, U. and Rustichini, A. 2000. A fine is a price. *Journal of Legal Studies* 29, 1–17.

Lazear, P. and Rosen, S. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89, 841–64.

MacLeod, B. and Malcomson, J. 1998. Motivation and markets. *American Economic Review* 88, 388–411.

How to cite this article

Falk, Armin and Simon Gächter. "experimental labour economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000276>
doi:10.1057/9780230226203.0529

The New Palgrave Dictionary of Economics Online

experimental macroeconomics

John Duffy

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Experimental macroeconomics is a sub-field of experimental economics that makes use of controlled laboratory methods to understand aggregate economic phenomena and to test the specific assumptions and predictions of macroeconomic models. This article reviews important contributions of experimental macroeconomics research, which include an understanding of when equilibration works, when it fails, and the means by which macro-coordination problems are resolved. It also discusses important methodological issues including the choice of market institution, the implementation of representative agent and overlapping generations models, discounting and infinite horizons, and the external validity of experimental macroeconomic findings.

Keywords

anchoring effects; asset pricing; bubbles; business cycles; contagion; coordination failure; double auction; equilibration; equilibrium selection; experimental macroeconomics; forecasting; general equilibrium; infinite horizons; Laffer curve; learning; microfoundations; money-search models; multiple equilibria; optimal growth; overlapping generations models; partial equilibrium; representative agent; search models of money; speculative attacks; strategic uncertainty; sunspot variables; time consistency

Article

Experimental macroeconomics is a subfield of experimental economics that makes use of controlled laboratory methods to understand aggregate economic phenomena and to test the specific assumptions and predictions of macroeconomic models. Surveys of experimental macroeconomics are found in Ochs (1995), Duffy (1998) and Ricciuti (2004). Macroeconomic topics that have been studied in the laboratory include convergence to Walrasian competitive equilibrium (Lian and Plott, 1998), growth and development (Lei and Noussair, 2002; Capra et al., 2005), specialization and trade (Noussair, Plot and Riezman, 1995), Keynesian coordination failures (Cooper, 1999; Van Huyck, Battalio and Beil, 1990),

the use of money as a medium of exchange (Brown, 1996; Duffy and Ochs, 1999; 2002) and as a store of value (McCabe, 1989; Lim, Prescott and Sunder, 1994; Marimon and Sunder, 1993; 1994), exchange rate determination (Arifovic, 1996; Noussair, Plot and Riezman, 1997), money illusion (Fehr and Tyran, 2001), asset price bubbles and crashes (Smith, Suchanek and Williams, 1988; Lei, Noussair and Plott, 2001; Hommes et al., 2005) sunspots (Marimon, Spear and Sunder, 1993; Duffy and Fisher, 2005), bank runs (Schotter and Yorulmazer, 2003; Garratt and Keister, 2005), contagions (Corbae and Duffy, 2006), speculative currency attacks (Heinemann, Nagel and Ockenfels, 2004), and the economic impact of various fiscal and monetary policies (Riedl and Van Winden, 2001; Arifovic and Sargent, 2003; Marimon and Sunder, 1994; Bernasconi and Kirchkamp, 2000).

The use of laboratory experiments, involving small groups of subjects interacting with one another for short periods of time, to analyse aggregate, economy-wide phenomena or to test macroeconomic model predictions or assumptions might be met with some scepticism. However, there are many insights to be gained from controlled laboratory experimentation that cannot be obtained using standard macroeconometric approaches, namely, econometric analyses of the macroeconomic data reported by government agencies. Often the data most relevant to testing a macroeconomic model are simply unavailable. There may also be identification, endogeneity and equilibrium selection issues that cannot be satisfactorily addressed using econometric methods. Indeed, Robert Lucas (1986) was the first macroeconomist to make such observations, and he invited laboratory tests of rational expectations macroeconomic models; much of the subsequent experimental macroeconomics literature may be viewed as a response to Lucas's (1986) invitation. It is also worth noting that experimental methodologies have been improbably applied to the study of many other aggregate phenomena including astronomy, epidemiology, evolution, meteorology, political science and sociology.

Insights from macroeconomic experiments

To date, experimental macroeconomics research has yielded some important insights, including an understanding of when equilibration works, when it fails, and the means by which equilibrium selection or coordination problems are resolved. Equilibration, the process by which competitive equilibrium is achieved, is often ignored by modern macroeconomic modellers, who typically assume that market clearing is friction-free and instantaneous. Experimentalists, following the lead of Smith (1962), have explored mechanisms such as the double auction, the availability of information, futures markets and other means by which this equilibration might be achieved or enhanced (see, for example, Forsythe, Palfrey and Plott, 1982; Plott and Sunder, 1982; Sunder, 1995 for partial equilibrium approaches; and Lian and Plott, 1998 for a general equilibrium approach). A general finding is that, with enough trading experience and information feedback about transaction prices, bids, and asks, even small populations of five to ten subjects can learn to trade at prices and achieve efficiency consistent with competitive equilibrium in a large class of market environments. Indeed, the institutional rules, for instance of the double auction, may be all that is necessary to assure equilibration, as shown in the zero-intelligence trader approach of Gode and Sunder (1993).

Experimental insights regarding equilibration have enabled experimentalists to design market environments where equilibration may fail to obtain; in its place are observed price bubbles and crashes

(Smith, Suchanek and Williams, 1988; Lei, Noussair and Plott, 2001; Hommes et al., 2005). Explaining these laboratory asset price bubbles has proved challenging. Lei, Noussair and Plott (2001) show that speculative motives alone cannot explain bubble formation and suggest that it may have more to do with subject boredom. Duffy and Ünver (2006) suggest that anchoring effects may factor in subjects' bidding up of prices until binding budget constraints force a crash. A further puzzle is that experienced subjects in laboratory asset markets learn to avoid price bubbles and crashes, and generally price assets in line with fundamental values. An explanation for why bubbles and crashes occur among inexperienced but not experienced subjects has yet to be provided. Experiments with mixtures of experienced and inexperienced subjects show no tendency for bubbles to arise (Dufwenberg, Lindqviist and Moore, 2005).

In environments with multiple equilibria, theory is typically silent as to which equilibrium agents will select or whether there will be transitions between equilibria. Understanding how agents coordinate on an equilibrium is of great interest to macroeconomists, as coordination problems are thought to play an important role in the persistence of business cycle fluctuations. Experimental evidence can and has been used to address the issue of which, among multiple equilibria, is most likely to be achieved, and why. For instance, Van Huyck, Battalio and Beil (1990) have shown how minimum effort, team production payoff functions can lead to Keynes-type coordination failures – that is, coordination by groups of subjects on Pareto inferior equilibria. Such inefficiencies do not arise from conflicting objectives or from asymmetries of information; rather, they arise from individuals' *strategic uncertainty* with regard to the actions of other market participants. Similarly, Duffy and Ochs (1999; 2002) report that subjects have no difficulty coordinating on efficient monetary exchange equilibria in Kiyotaki–Wright-type money-search models when theory calls for the use of fundamental, cost-minimizing strategies, but subjects have much greater difficulty coordinating on efficient monetary equilibria that require them to employ more costly and forward-looking, speculative strategies, due perhaps to the unwillingness of other subjects to adopt those same speculative strategies.

Not all the experimental evidence points to inefficiencies in macro-coordination problems. Marimon and Sunder (1993) show that when subjects are presented with a Laffer-curve-type trade-off between two inflation rates, the efficient, low-inflation equilibria is more likely to be selected than is the inefficient, high-inflation equilibrium. They show that the low-inflation equilibrium is stable under the adaptive learning dynamics that subjects use whereas the high-inflation equilibrium is not. Similarly, Arifovic and Sargent (2003) study behaviour in a Kydland–Prescott model of expected inflation output trade-offs and find that a majority of subjects acting in the role of central bank are able to choose policies so as to induce subjects, in the role of the public, to coordinate their expectations on the efficient but time-inconsistent Ramsey equilibrium. Still, they report occasional instances of 'backsliding' to the less efficient, time-consistent Nash equilibrium.

Finally, Duffy and Fisher (2005) explore subjects' use of non-fundamental 'sunspot' variables as coordination devices in an environment with multiple equilibria. They show that, when information is highly centralized, as in a call market, subjects use realizations of a sunspot variable as a device for coordinating on low- or high-price equilibria, but that this coordination mechanism may break down when information is more decentralized, as in a double auction, or when the mapping from realizations of the sunspot variable to the action space is unclear.

Methodological issues

Methodologically, macroeconomic experiments typically involve some kind of centralized market-clearing mechanism through which subjects interact with one another, for instance as buyers or sellers, or both. The double auction market mechanism (Friedman and Rust, 1991) is the most commonly used market-clearing mechanism, as it allows for continuous information on bids, asks, transaction prices and volume – information which is thought to be critical to rapid equilibration and high levels of allocative efficiency (Lian and Plott, 1998; Noussair, Plott and Riezman, 1995; 1997). The simultaneous, sealed-bid ‘call’ market version of this mechanism has also been used by some researchers (Cason and Friedman, 1997; Duffy and Fisher, 2005; Capra et al., 2005).

Some less centralized market mechanisms have also been used. For instance, Brown (1996), Duffy and Ochs (1999; 2002) study a money-search model in which subjects are randomly paired and may trade goods with one another at a fixed exchange rate. In addition, game-theoretic models are also commonly employed, especially in studies of coordination failure, contagion and speculative attacks (Van Huyck, Battalio and Beil, 1990; Corbae and Duffy, 2006; Heinemann, Nagel and Ockenfels, 2004).

A hallmark of modern macroeconomic modelling is the characterization of the economy using recursive dynamical systems where expectations of future endogenous variables determine current outcomes. Several experimental researchers testing such models have found it useful to separate subjects’ forecast decisions from market-trading decisions. For instance, Marimon and Sunder (1993; 1994; 1995) and Hommes et al. (2005) elicit subjects’ forecasts of the next period's price level. Using these individual forecasts, they determine subjects’ individual demands for the consumption good in the current period and, as supply is fixed, they simultaneously determine the current period price. Similarly, Adam (2007) elicits forecasts of inflation one and two periods ahead, consistent with the monetary sticky price model that he investigates; these expectations are then used to determine output and inflation in the current period. Marimon and Sunder (1994) refer to this type of experimental design as a ‘learning to forecast’ framework, which they contrast with a ‘learning to optimize’ framework. Of course, in macroeconomic models, it is assumed that agents are able to both forecast and optimize at the same time.

Many macroeconomic models have representative agents and infinite horizons or an infinity of agents and goods which pose some challenges for laboratory implementation and testing of theoretical predictions. The representative agent assumption has been examined by Noussair and Matheny (2000) and Lei and Noussair (2002). They compare consumption and investment decisions made by individual subjects operating as ‘representative agent-social planners’ in the standard Cass–Koopmans optimal growth framework with the decisions made by groups of subjects who first trade shares of capital via a double-auction market clearing mechanism and then allocate their income between consumption and investment. They find that the double-auction market mechanism results in allocations that are far closer to the theoretical predictions than are the decisions made by subjects in the representative agent role attempting to solve the optimization problem on their own.

To implement infinite horizons, researchers have adopted two designs. One design, used for example by Marimon and Sunder (1993), is to recruit subjects for a fixed period of time but terminate the session early, without advance notice, following the end of some period of play. As Marimon and Sunder use a forward-looking dynamic model, they use the one-step-ahead forecasts made by a subset of subjects who

are paid for their forecast accuracy to determine final period allocations. A second design is to introduce a constant small probability, $1-\delta$, that each period will be the last one played in a sequence, and allow enough time for several indefinite sequences to be played in an experimental session (Duffy and Ochs, 1999; 2002; Lei and Noussair, 2002; Capra et al., 2005). This design has the advantage of inducing both the stationarity associated with an infinite horizon and discounting of future payoffs at rate $(1-\delta)/\delta$ per period (equivalently a discount *factor* of δ). Related to the infinite horizon problem, overlapping generations models, as studied by Marimon and Sunder (1993; 1994; 1995) and Marimon, Spear and Sunder (1993) have an infinity of agents (and goods). Marimon and Sunder cope with this difficulty by recycling subjects – allowing each subject to live several two-period lives over the course of an indefinite sequence of periods. Marimon and Sunder (1993) argue that this repeated entry and exit of subjects does not induce any strategic opportunities that are not already present in the overlapping generations model without 'rebirth'. Indeed, the need for a large number of agents to study macroeconomic behaviour is a common issue confronted by researchers. However, results from many double auction experiments suggest that competitive equilibrium can be quickly achieved with as few as three to five subjects operating on each side of the market. Similarly, while search models of money assume a continuum of agents, Duffy and Ochs (2002) argue that the strategic incentives generated by having finite subject populations do not alter the equilibrium predictions of those models under the assumption of a continuum of agents.

Perhaps the most difficult methodological issue is the external validity of macroeconomic experimental findings. While external validity is generally a problem for all experimental economists, it might be regarded as a greater problem for macro-experimentalists seeking to explain economy-wide aggregate macroeconomic phenomena using necessarily small-scale laboratory evidence. Experimental macroeconomists have several responses to this issue. First, as noted earlier, modern macroeconomic models have explicit microfoundations as to how individual agents make decisions (for example, agents recognize the relevant trade-offs, form rational expectations) which can be directly tested in the laboratory. Indeed, in the laboratory one can be more certain about micro-level *causal relationships*, that is, that an experimenter induced change in a variable is the source of any observed change in subject behaviour as opposed to some other, unaccounted-for factors. Macroeconometric analyses of field data cannot claim the same degree of *internal validity*. A second response is to make use of highly experienced subjects – those who have participated in the same experiment many times – as a means of better proxying real-world behaviour. As noted earlier, asset price bubbles and crashes seem to disappear with experienced subjects. A third response has been to use parametric forms or calibrations of macroeconomic models that are of interest to macroeconomists, or to present subjects with real macroeconomic data as part of the experimental design (for example, Bernasconi, Kirchkamp and Parulo, 2004). Finally, many experimentalists would argue that all experimental work, including macroeconomic experiments, should be judged by the findings obtained and not by biases concerning the suitability of laboratory versus other empirical methods, all of which have their strengths and weaknesses.

See Also

- behavioural finance
- coordination problems and communication
- experimental economics

Bibliography

Adam, K. 2007. Experimental evidence on the persistence of output and inflation. *Economic Journal* 117, 603–36.

Arifovic, J. 1996. The behavior of the exchange rate in the genetic algorithm and experimental economies. *Journal of Political Economy* 104, 510–41.

Arifovic, J. and Sargent, T.J. 2003. Laboratory experiments with an expectational Phillips curve. In *Evolution and Procedures in Central Banking*, ed. D.E. Altig and B.D. Smith. Cambridge: Cambridge University Press.

Bernasconi, M. and Kirchkamp, O. 2000. Why do monetary policies matter? An experimental study of saving and inflation in an overlapping generations model. *Journal of Monetary Economics* 46, 315–43.

Bernasconi, M., Kirchkamp, O. and Paruolo, P. 2004. Do fiscal variables affect fiscal expectations? Experiments with real world and lab data. SPF 504 Discussion Paper No. 04–26: University of Mannheim.

Brown, P.M. 1996. Experimental evidence on money as a medium of exchange. *Journal of Economic Dynamics and Control* 20, 583–600.

Capra, C.M., Tanaka, T., Camerer, C.F., Munyan, L., Sovero, V., Wang, L. and Noussair, C. 2005. The impact of simple institutions in experimental economies with poverty traps. Working paper.

Cason, T.N. and Friedman, D. 1997. Price formation in single call markets. *Econometrica* 65, 311–45.

Cooper, R.W. 1999. *Coordination Games*. Cambridge: Cambridge University Press.

Corbae, D. and Duffy, J. 2006. Experiments with network formation. Working paper.

Duffy, J. 1998. Monetary theory in the laboratory. *Federal Reserve Bank of St. Louis Economic Review* 80, 9–26.

Duffy, J. and Fisher, E.O.N. 2005. Sunspots in the laboratory. *American Economic Review* 95, 510–29.

Duffy, J. and Ochs, J. 1999. Emergence of money as a medium of exchange: an experimental study.

American Economic Review 89, 847–77.

Duffy, J. and Ochs, J. 2002. Intrinsically worthless objects as media of exchange: experimental evidence. *International Economic Review* 43, 637–73.

Duffy, J. and Ünver, U. 2006. Asset price bubbles and crashes with near-zero-intelligence traders. *Economic Theory* 27, 537–63.

Dufwenberg, M., Lindqviist, T. and Moore, E. 2005. Bubbles and experience: an experiment. *American Economic Review* 95, 1731–7.

Fehr, E. and Tyran, J.-F. 2001. Does money illusion matter? *American Economic Review* 91, 1239–62.

Forsythe, R., Palfrey, T.R. and Plott, C.R. 1982. Asset valuation in an experimental market. *Econometrica* 58, 537–68.

Friedman, D. and Rust, J. 1991. *The Double Auction Market: Institutions, Theories and Evidence*. Cambridge, MA: Perseus Publishing.

Garratt, R. and Keister, T. 2005. Bank runs: an experimental study. Working paper, University of California, Santa Barbara.

Gode, D.K. and Sunder, S. 1993. Allocative efficiency of markets with zero intelligence traders: market as a partial substitute for individual rationality. *Journal of Political Economy* 101, 119–37.

Heinemann, F., Nagel, R. and Ockenfels, P. 2004. The theory of global games on test: experimental analysis of coordination games with public and private information. *Econometrica* 72, 1583–99.

Hommes, C.H., Sonnemans, J., Tuinstra, J. and van de Velden, H. 2005. Coordination of expectations in asset pricing experiments. *Review of Financial Studies* 18, 955–80.

Lei, V. and Noussair, C.N. 2002. An experimental test of an optimal growth model. *American Economic Review* 92, 549–70.

Lei, V., Noussair, C.N. and Plott, C.R. 2001. Nonspeculative bubbles in experimental asset markets: lack of common knowledge of rationality vs. actual irrationality. *Econometrica* 69, 831–59.

Lian, P. and Plott, C.R. 1998. General equilibrium, markets, macroeconomics and money in a laboratory experimental environment. *Economic Theory* 12, 21–75.

Lim, S.S., Prescott, E.C. and Sunder, S. 1994. Stationary solution to the overlapping generations model

of fiat money: experimental evidence. *Empirical Economics* 19, 255–77.

Lucas, R.E. 1986. Adaptive behavior and economic theory. *Journal of Business* 59, S401–S426.

Marimon, R., Spear, S.E. and Sunder, S. 1993. Expectationally driven market volatility: an experimental study. *Journal of Economic Theory* 61, 74–103.

Marimon, R. and Sunder, S. 1993. Indeterminacy of equilibria in a hyperinflationary world: experimental evidence. *Econometrica* 61, 1073–107.

Marimon, R. and Sunder, S. 1994. Expectations and learning under alternative monetary regimes: an experimental approach. *Economic Theory* 4, 131–62.

Marimon, R. and Sunder, S. 1995. Does a constant money growth rule help stabilize inflation? Experimental evidence. *Carnegie-Rochester Conference Series on Public Policy* 43, 111–56.

McCabe, K.A. 1989. Fiat money as a store of value in an experimental market. *Journal of Economic Behavior and Organization* 12, 215–31.

Noussair, C.N. and Matheny, K.J. 2000. An experimental study of decisions in dynamic optimization problems. *Economic Theory* 15, 389–419.

Noussair, C.N., Plott, C.R. and Riezman, R.G. 1995. An experimental investigation of the patterns of international trade. *American Economic Review* 85, 462–91.

Noussair, C.N., Plott, C.R. and Riezman, R.G. 1997. The principles of exchange rate determination in an international financial experiment. *Journal of Political Economy* 105, 822–61.

Ochs, J. 1995. Coordination problems. In *The Handbook of Experimental Economics*, ed. J.H. Kagel and A.E. Roth. Princeton: Princeton University Press.

Plott, C.R. and Sunder, S. 1982. Efficiency of experimental security markets with insider information: an application of rational-expectations models. *Journal of Political Economy* 90, 663–98.

Ricciuti, R. 2004. Bringing macroeconomics into the lab. Working paper No. 26, International Centre for Economic Research.

Riedl, A. and van Winden, F. 2001. Does the wage tax system cause budget deficits? A macro-economic experiment. *Public Choice* 109, 371–94.

Schotter, A. and Yorulmazer, T. 2003. On the severity of bank runs: an experimental study. Working

paper, Center for Experimental Social Science, New York University.

Smith, V.L. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70, 111–37.

Smith, V.L., Suchanek, G.L. and Williams, A.W. 1988. Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica* 56, 1119–51.

Sunder. 1995. Experimental asset markets: a survey. In *The Handbook of Experimental Economics*, ed. J. H. Kagel and A.E. Roth. Princeton: Princeton University Press.

Van Huyck, J.B., Battalio, R.C. and Beil, R.O. 1990. Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review* 80, 234–48.

How to cite this article

Duffy, John. "experimental macroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000275> doi:10.1057/9780230226203.0530

The New Palgrave Dictionary of Economics Online

experimental methods in economics

Vernon L. Smith
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the mid-20th century economists became involved in the design and conduct of laboratory experiments to examine propositions implied by economic theory. This development brought new standards of rigour to the data gathering process. This article gives an account of the author's experiment in 1956 to test the hypothesis that the competitive market process yields welfare improving (and, under certain limiting ideal conditions, welfare maximizing) outcomes, provides an interpretive history of the development of experimental economics, discusses the functions of market experiments in microeconomic analysis, and classifies the application of experimental methods.

Keywords

agenda processes; antitrust economics; auctions (theory); buyer's surplus; collusion; compensated unanimity processes; competitive equilibrium; contestable markets hypothesis; decentralization; demand functions; dominance; double oral auction; expected utility theory; experimental methods in economics; hypotheses; incentive compatibility; industrial organization; institutions; labour supply; market as positive sum game; monopoly; monopsony; natural monopoly; non-satiation; parallelism; preference theory; privacy; public enforcement of law; public good allocation; research programmes; saliency; scale economics; seller's surplus; Smith, A.; Smith, V. L.; speculative gains; supply functions; testing; Vickrey, W.; vouchers

Article

Historically, the method and subject matter of economics have presupposed that it was a non-experimental (or ‘field observational’) science more like astronomy or meteorology than physics or chemistry. Based on general, introspectively ‘plausible’, assumptions about human preferences, and about the cost and technology based supply response of producers, economists have sought to understand the functioning of economies, using observations generated by economic outcomes realized over time. The data of the astronomer is of this same type, but it would be wrong to conclude that astronomy and economics are methodologically equivalent. There are two important differences between astronomy and economics which help to illuminate some of the methodological problems of economics. First, based upon parallelism (the maintained hypothesis that the same physical laws hold everywhere), astronomy draws on all the relevant theory from classical mechanics and particle physics – theory which has evolved under rigorous laboratory tests. Traditionally, economists have not had an analogous body of tested behavioural principles that have survived controlled experimental tests, and which can be assumed to apply with insignificant error to the microeconomic behaviour that underpins the observable operations of the economy. Analogously, one might have supposed that there would have arisen an important area of common interest between economics and, say, experimental psychology, similar to that between astronomy and physics, but this has only started to develop in recent years. Second, the data of astronomy are painstakingly gathered by professional observational astronomers for scientific purposes, and these data are taken seriously (if not always non-controversially) by astrophysicists and cosmologists. Most of the data of economics has been collected by government or private agencies for non-scientific purposes. Hence astronomers are directly responsible for the scientific credibility of their data in a way that economists have not been. In economics, when things appear not to turn out as expected the quality of the data is more likely to be questioned than the relevance and quality of the abstract reasoning. Old theories fade away, not from the weight of falsifying evidence that catalyses theoretical creativity into developing better theory, but from lack of interest, as intellectual energy is attracted to the development of new techniques and to the solution of new puzzles that remain untested.

At approximately the mid-20th century, professional economics began to change with the introduction of the laboratory experiment into economic method. In this embryonic research programme economists (and a psychologist, Sidney Siegel) became directly involved in the design and conduct of experiments to examine propositions implied by economic theories of markets. For the first time this made it possible to introduce *demonstrable* knowledge into the economist's attempt to understand markets.

This laboratory approach to economics also brought to the economist direct responsibility for an important source of scientific data generated by controlled processes that can be replicated by other experimentalists. This development invited economic theorists to submit to a new discipline, but also brought an important new discipline and new standards of rigour to the data gathering process itself.

An untested theory is simply a hypothesis. As such it is part of our *self*-knowledge. Science seeks to expand our knowledge of *things* by a process of testing this type of self-knowledge. Much of economic theory can be called, appropriately, ‘ecclesiastical theory’; it is accepted (or rejected) on the basis of authority, tradition, or opinion about assumptions, rather than on the basis of having survived a rigorous falsification process that can be replicated.

Interest in the replicability of scientific research stems from a desire to answer the question ‘Do you see what I see?’ Replication and control are the two primary means by which we attempt to reduce the error in our common knowledge of economic processes. However, the question ‘Do you see what I see?’ contains three component questions, recognition of which helps to identify three different senses in which a research study may fail to be replicable:

1. (1) *Do you observe what I observe?* Since economics has traditionally been confined to the analysis of non-experimental data, the answer to this question has been trivially, ‘yes’. We observe the same thing because we use the same data. This non-replicability of our traditional data sources has helped to motivate some to turn increasingly to experimental methods. We can say that you have replicated my experiments if you are unable to reject the hypothesis that your experimental data came from the same population as mine. This means that the experimenter, his/her subjects, and/or procedures are not significant treatment variables.
2. (2) *Do you interpret what we observe as I interpret it?* Given that we both observe the same, or replicable data, do we put the same interpretation on these data? The interpretation of observations requires theory (either formal or informal), or at least an empirical interpretation of the theory in the context that generated the data. Theory usually requires empirical interpretation either because (i) the theory is not developed directly in terms of what can be observed (e.g. the theory may assume risk aversion which is not directly observable), or (ii) the data were not collected for the purpose of testing, or estimating the parameters of a theory. Consequently, failure to replicate may be due to differences in interpretation which result from different meanings being ascribed to the theory. Thus two researchers may apply different transformations to raw field data (e.g. different adjustments for the effect of taxes), so that the results are not replicable because their theory interpretations differ.
3. (3) *Do you conclude what I conclude from our interpretation?* The conclusions reached in two different research studies may be different even though the data and their interpretation are the same. In economics this is most often due to different model specifications. This problem is inherent in non-experimental methodologies in which, at best, one usually can estimate only the parameters of a prespecified model and cannot credibly test one model or theory against another. An example is the question of whether the Phillips’ curve constitutes a behavioural trade-off between the rates of inflation and unemployment, or represents an equilibrium association without causal significance.

I Markets and market experiments

Markets and how they function constitute the core of any economic system, whether it is highly decentralized – popularly, a ‘capitalistic’ system, or highly centralized – popularly, a ‘planned’ system. This is true for the decentralized economy because markets are the spontaneous institutions of exchange that use prices to guide resource allocation and human economic action. It is true for the centralized economy because in such economies markets always exist or arise in legal form (private agriculture in Russia) and clandestine or illegal form (barter, bribery, the trading of favours, and underground exchange in Russia, Poland and elsewhere). Markets arise spontaneously in all cultures in response to the human desire for betterment (to ‘profit’) through exchange. Where the commodity or service is illegal (prostitution, gambling, the sale of liquor under Prohibition or of marijuana, cocaine, etc.) the result is not to prevent exchange, but to raise the risk and therefore the costs of exchange. This is because enforcement is itself costly, and it is never economical for the authorities (whether Soviet or American) even to approximate perfect enforcement. The spontaneity with which markets arise is perhaps no better illustrated than when (1979–80) US airlines for promotional purposes issued travel vouchers to their passengers. One of these vouchers could be redeemed by the bearer as a cash substitute in the purchase of new airline tickets. Consequently vouchers were of value to future passengers. Furthermore, since (as Hayek would say) the ‘circumstances of time and place’ for the potential redemption of vouchers were different for different individuals, there existed the preconditions for the active voucher market that was soon observed in all busy airports. Current passengers with vouchers who were unlikely to be travelling again soon held an asset worth less to themselves than to others who were more certain of their future or impending travel plans. The resulting market established prices that were discounts from the redemption or ‘face’ value of vouchers. Sellers who were unlikely to be able to redeem their vouchers preferred to sell them at a discount for cash. Buyers who were reasonably sure of their travel plans could save money by purchasing vouchers at a discount. Thus the welfare of every active buyer and seller increased via this market. Without a market, many – perhaps most – vouchers would not have been exercised and would thus have been ‘wasted’.

The previous paragraph illustrates a fundamental hypothesis (theorem) of economics: the (‘competitive’) market process yields welfare improving (and, under certain limiting ideal conditions, welfare maximizing) outcomes. But is the hypothesis ‘true’, or at least very probably true? (Lakatos (1978) would correctly ask ‘Has it led to an empirically progressive

research programme?') I think it is 'true', but how do I know this? Do you see what I see? A Marxist does not see what I see in the above interpretation of a market. The young student studying economics does not see what I see, although if they continue to study economics eventually they (predictably) come to see what I see (or, at least, they say they do). Is this because we have inadvertently brainwashed them? The gasoline consumer does not see what I see. They see themselves in a *zero* sum game with an oil company: any increase in price merely redistributes wealth from the consumer to the company, which is not 'fair' since the company is richer. What I see in a market is a *positive* sum game yielding gains from exchange, which constitutes the fundamental mechanism for creating, not merely redistributing wealth. The traditional method by which the economist gets others to see this 'true' function of markets is by logical arguments (suppose it were not true, then …), examples, and 'observations', such as are contained in my description of the voucher market, in which what is 'observed' is hortatively described and interpreted in terms of the hypothesis itself. But if this knowledge of the function of markets is 'true', can it be demonstrated? Experimentalists claim that laboratory experiments can provide a uniquely important technique of demonstration for supplementing the theoretical interpretation of field observations. I conducted my first experiment in the spring of 1956. Since then hundreds of similar, as well as environmentally richer experiments have been conducted by myself and by others. In 1956, my introductory economics class consisted of 22 science and engineering students, and although this might not have been the 'large number' traditionally thought to have been necessary to yield a competitive market, I though it was large enough for a practice run to initiate a research programme capable of falsifying the standard theory. I conducted the experiment before lecturing on the theory and 'behaviour' of markets in class so as not to 'contaminate' the sample. The 22 subjects were each assigned one card from a well-shuffled deck of 11 white and 11 yellow cards. The white cards identified the sellers, and the yellow cards identified the buyers. Each white card carried a price, known only to that seller, which represented that seller's minimum selling price for one unit, and each yellow card identified a price, known only to that buyer, representing that buyer's maximum buying price for one unit. On the left of Figure 1 is listed these so-called 'limit' prices, identified by buyer, B1, B2 etc. (in descending order, D) and by seller, S1, S2 etc. (in ascending order, S).

To keep things simple and well controlled each buyer (seller) was informed that he/she was a buyer (seller) of at most one unit of the item in each of several trading periods. Thus demand, D (supply, S) was 'renewed' in each trading period as a steady state flow, with no carry-over in unsatisfied demand (or unsold stock), from one period to the next. In the airline voucher example, imagine the vouchers being issued, followed by trading; the vouchers then expire, new vouchers are issued, traded and so on. In the experiment, suppose real motivation is provided by promising to pay (in cash) to each buyer the difference between that buyer's assigned limit buying price and the price actually paid in each period that a unit is purchased in the market. Thus suppose seller 5 sells their unit to buyer 2 at the price 2.25. Then buyer 2 earns a 'profit' of $0.75 from this exchange. In this way we induce on each buyer a value (or hypothesized willingness-to-pay) equal to the assigned limit buy price. Similarly, suppose each seller is paid the difference between that seller's actual sales price and assigned limit price ('cost', or willingness-to-sell) in each trading period that a unit is sold. Thus in the previous exchange example, seller 5 earns $0.50 from the transaction.

Figure 1

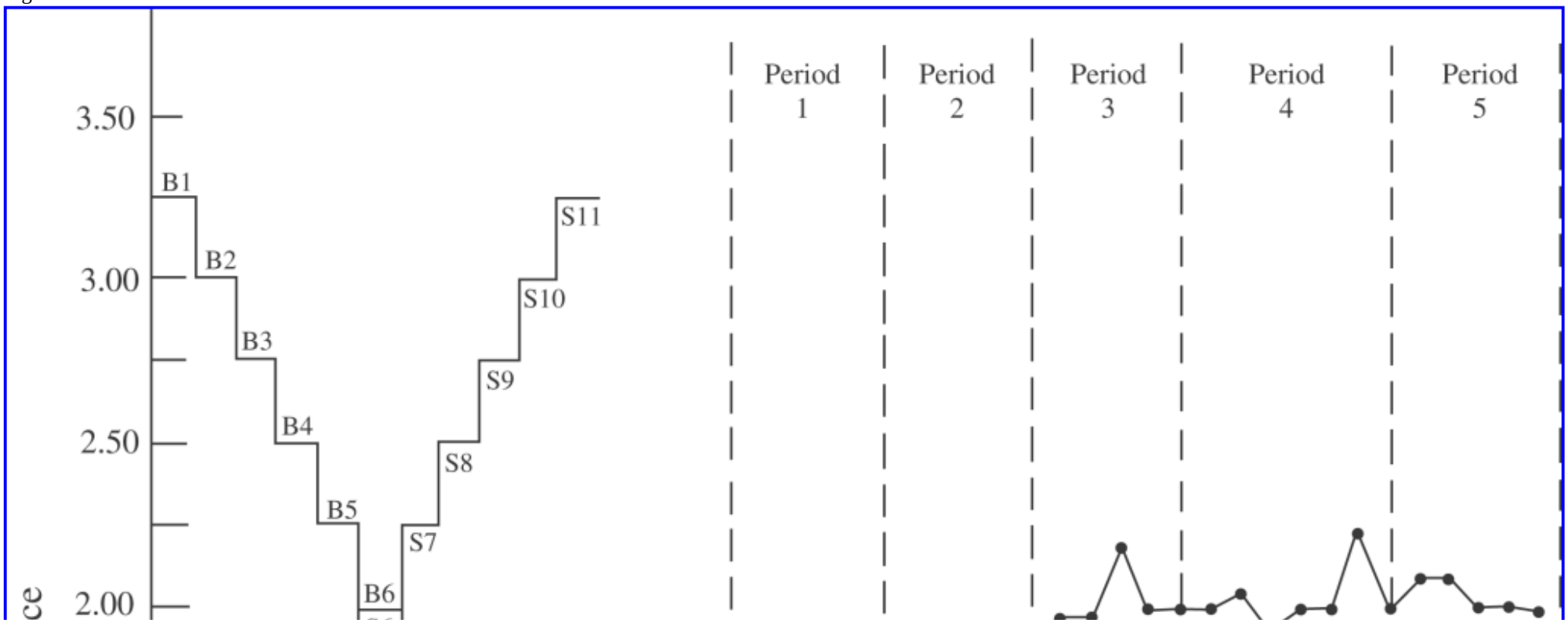

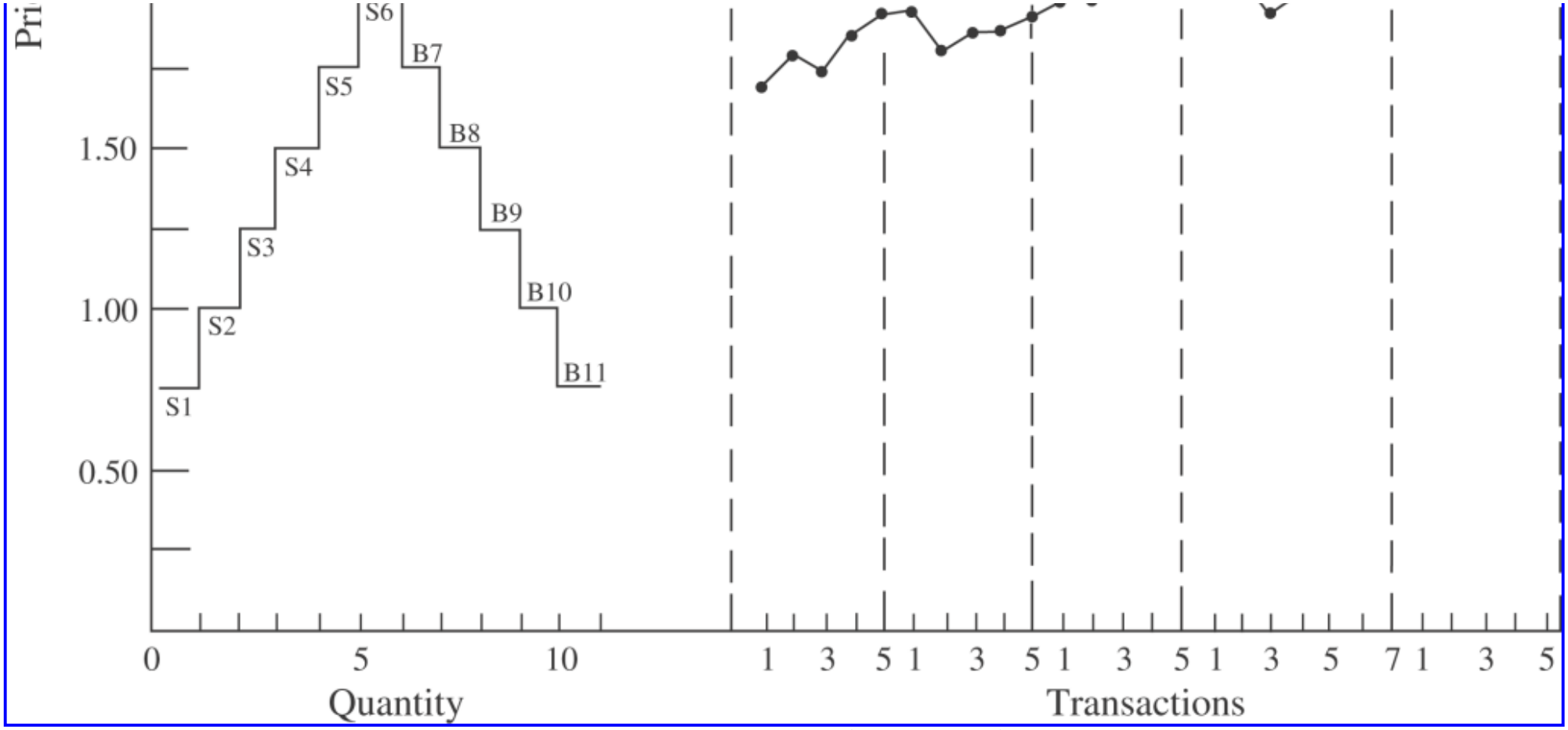

This experimental procedure operationalizes the market preconditions that (1) ‘the circumstances of time and place’ for each economic agent are dispersed and known only to that agent (as in the above voucher market) and (2) agents have a secure property right in the objects of trade and the private gains (‘profits’) from trade (an airline travel voucher was transferable and redeemable by any bearer). The reader should note that ‘profit’ is identified as much with the act of buying as with that of selling. This is because ‘profit’ is the surplus earned by a buyer who buys for less than his willingness-to-pay, just as a seller's ‘profit’ is the surplus earned when an item is sold for more than the amount for which they are willing to sell. Willingness-to-sell need not have, and usually does not have anything to do with accounting ‘cost’, or production ‘cost’, from which one computes accounting profit. Willingness-to-sell, like willingness-to-buy, is determined by the immediate circumstances of each agent. Hence, a passenger might be prepared to pay the regular full fare premium on a first-class ticket for an emergency trip to visit a sick relative. The accountant's concept of profit cannot be applied to the passenger's decision any more than it can be applied to that of a passenger willing to sell a voucher at a deep discount. In what follows I will use the term ‘buyer's surplus’ or ‘seller's surplus’ instead of ‘profit’ to refer to the gains from exchange enjoyed by buyers or sellers because the term ‘profit’ is so strongly, exclusively and misleadingly associated with selling activities.

Now let us interpret the previously cited fundamental theorem of economics in the context of the experimental design contained in Figure 1. We note first that the ordered set of seller (buyer) limit prices defines a supply (demand) function (Figure 1). A supply (demand) function provides a list of the total quantities that sellers (buyers) would be willing to sell (buy) at corresponding hypothetical fixed prices. Neither of these functions is capable of being observed, scientifically, in the field. This is because the postulated limit prices are inherently private and not publicly observable. We could poll every potential seller (buyer) of vouchers in Chicago's O'Hare airport on 20 December 1979 to get each person's reported limit price, but we would have no way of validating the ‘observations’ thus obtained. Referring to Figure 1, we see that in my 1956 experiment, sellers (hypothetically) were just willing to sell three units at price 1.25, nine units at 2.75 and so on. Similarly buyers (hypothetically) were just willing to buy four units at 2.50, seven units at 1.75 and so on. If seller 3 is indifferent between selling and not selling at 1.25, and if every seller (buyer) is likewise indifferent at his/her limit price, then any particular unit may not be sold (purchased) at this limit price. One means of dealing with this problem in laboratory markets is to promise to pay a small ‘commission’, say 5 cents, to each buyer and seller for each unit bought or sold. Thus seller 3 has a small inducement to sell at 1.25 if he can do no better, and buyer 6 has a small inducement to buy at 2.00 if she can do no better.

Economic theory defines the competitive equilibrium as the price and corresponding quantity that clears the market; that is, it sets the quantity that sellers are willing to sell equal to the quantity that buyers are willing to buy. This assumes that the subjective cost of transacting is zero; otherwise any units with limit prices equal to the competitive equilibrium price will not exchange. In Figure 1 this competitive equilibrium price is 2.00. If the 5 cent ‘commission’ paid to each trading buyer and seller is sufficient to compensate for any subjective cost of transacting, then buyer 6 and seller 6 will each trade and the competitive equilibrium quantity exchanged will be 6 units. At the competitive equilibrium price, buyer 1 earns a surplus of 3.25–2.00=1.25 (plus commission) per period and so on. Total surplus, which measures the maximum possible gains from exchange, or maximum wealth *created* by the existence of the market institution, is 7.50 per period, at the competitive equilibrium.

If by some miracle the competitive equilibrium price and exchange quantity were to prevail in this market, sellers 1–6 would sell, buyers 1–6 would buy, while sellers 7–11 would make no sales and buyers 7–11 would make no purchases. It might be thought that this is unfair – the market should permit some or all of the ‘submarginal’ buyers (sellers) 7–11 to trade – or that more wealth would be created if there were more than six exchanges. But these interpretations are wrong. By definition, buyer 10 is not willing to pay more than 1.00. Consequently, it is a peculiar notion of fairness to argue that buyer 10 should have as much priority as buyer 1 in obtaining a unit. In the airline voucher example, this would mean that a buyer who is unlikely to redeem a voucher should have the same priority as a buyer who is likely to redeem a voucher. One can imagine a market in which, say, buyer 1 is paired with seller 9 at price 3.00, buyer 2 with seller 8 at price 2.75, and so on with nine units traded. If this were to occur it would mean buyers 7–9, who are less likely to use vouchers, have purchased them, and sellers 7–9, who initially held vouchers, and were more likely to use them than buyers 7–9, have sold their vouchers. Furthermore, this allocation yields additional possible gains from exchange, and is thus *not sustainable*, even if it were thought to be desirable. That is, buyer 9, who bought from seller 1 at price 1.00, could resell the unit to seller 9 (who sold her unit to buyer 2), at price (say) 2.00. Why? Because, by definition a voucher is worth 2.75 to seller 9 and only 1.25 to buyer 9. Similar additional trades can be made by buyers (sellers) 7 and 8. The end result would be that buyers 1–6 and sellers 7–11 would be the terminal holders of vouchers, just as if the competitive equilibrium had been reached initially.

Hence, either the competitive equilibrium prevails, or if inefficient trades occur at dispersed prices, then further ‘speculative’ gains can be made by some buyers and sellers. If these gains are fully captured the end result is the same allocation as would occur at the competitive equilibrium price and quantity.

Having specified the environment (individual private values) of our experimental market, what remains is to specify an exchange institution. In my 1956 experiment I elected to use trading rules similar to those that characterize trading on the organized stock and commodity exchanges. These markets use the ‘double oral auction’ procedure. In this institution as soon as the market ‘opens’ any buyer is free to announce a bid to buy and any seller is free to announce an offer to sell. In the experimental version each bid (offer) is for a single unit. Thus a buyer might say ‘buy, 1.00’, while a seller might say ‘sell, 5.00’, and it is understood that the buyer bids 1.00 for a unit and the seller offers to sell one unit for 5.00. Bids and offers are freely announced and can be modified. A contract occurs if any seller accepts the bid of any buyer, or any buyer accepts the offer of any seller. In the simple experimental market, since each participant is a buyer or seller of at most one unit per trading period, the contracting buyer and seller drop out of the market for the remainder of the trading period, but return to the market when a new trading ‘day’ begins. The experimenter announces the close of each trading period and the opening of the subsequent period, with each trading period timed to extend, say, five minutes. Each contract price is plotted on the right of Figure 1 for the five trading periods of the experiment. This result was not as expected. The conventional view among economists was that a competitive equilibrium was like a frictionless ideal state which could not be conceived as actually occurring, even approximately. It could be conceived of occurring only in the presence of an abstract ‘institution’ such as a Walrasian *tâtonnement* or an Edgeworth recontracting procedure. It was for teaching, not believing.

From Figure 1 it is evident that in the strict sense the competitive equilibrium was not attained in any period, but the accuracy of the competitive equilibrium theory is easily comparable to that of countless physical processes. Certainly, the data clearly do not support the monopoly, or seller collusion model. The total return to sellers is maximized when four units are sold at price 2.50. Similarly, the monopsony, or buyer collusion model requires four units to exchange at price 1.50.

Since 1956, several hundred experiments using different supply and demand conditions, experienced as well as inexperienced subjects, buyers and sellers with multiple unit trading capacity, a great variation in the numbers of buyers and sellers, and different trading institutions, have established the replicability and robustness of these results. For many years at the University of Arizona and Indiana University we have been using various computerized (the PLATO system) versions of the double ‘oral’ auction, developed by Arlington Williams, in which participating subjects trade with each other through computer terminals. These experiments establish that the 1956 results are robust with respect to substantial reductions in the number of buyers and sellers. Most such experiments use only four buyers and four sellers, each capable of trading several units. Some have used only two sellers, yet the competitive equilibrium model performs very well under double auction rules. Figure 2 shows the supply and demand design and the market results for a typical experiment in which subjects trade through PLATO computer terminals under computer-monitored double auction rules.

Figure 2

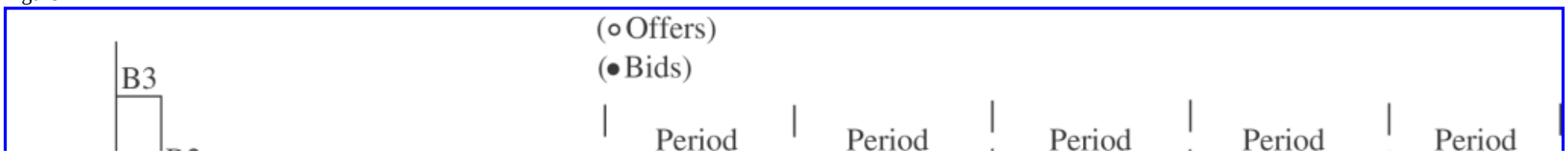

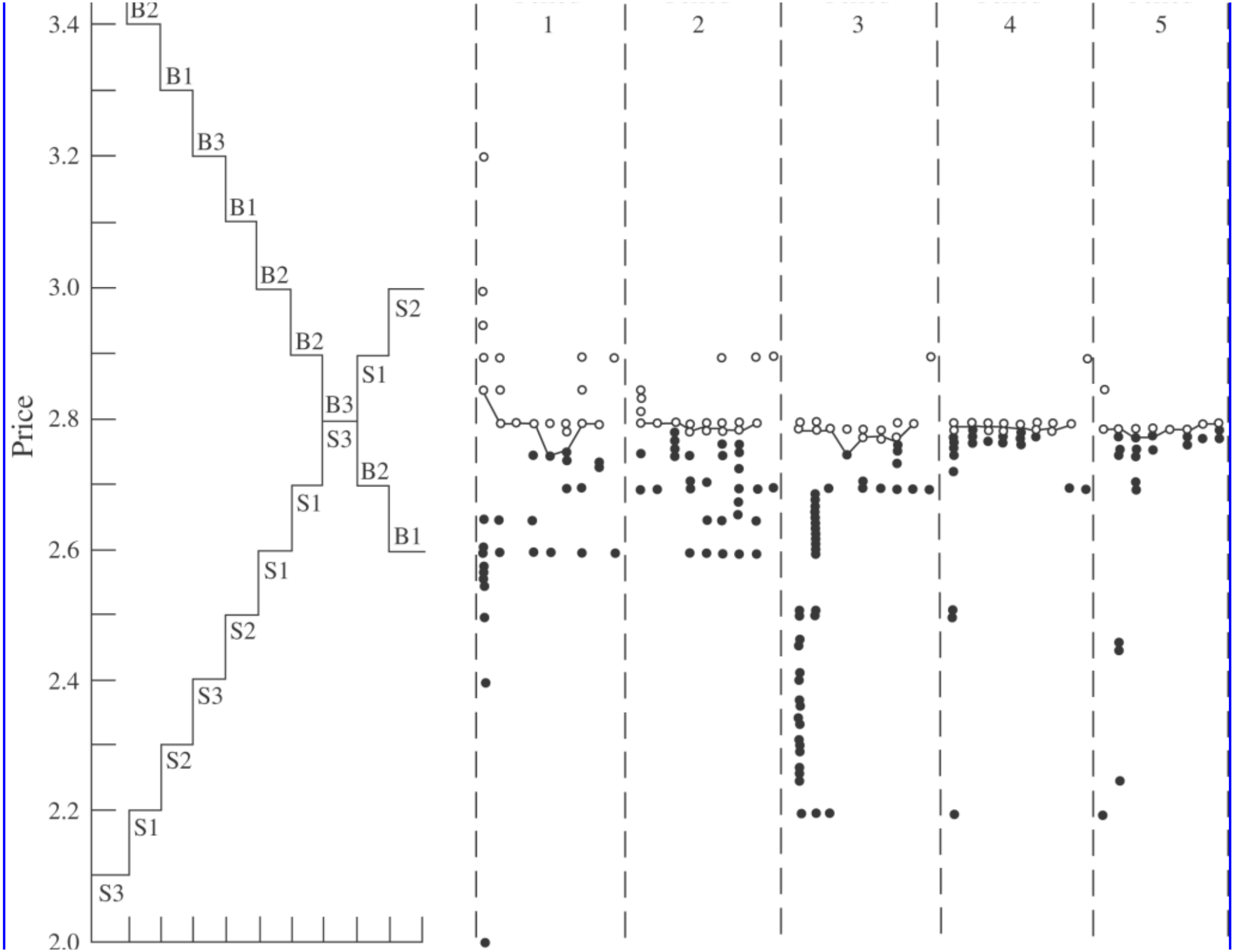
Price
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
B2
B1
B3
B1
B2
B2
B3
S3
S1
S2
S3
S2
S1
S1
B2
B1
S2
S1
S3
1
2
3
4
5

In addition to its antiquarian value, Figure 1 illustrates the problem of monitoring the rules of a ‘manual’ experiment. Observe that in period 4 there were seven contracts which are recorded as occurring in the price range between $1.90 and $2.25. This is not possible since there are only six buyers with limit buy prices above $1.90. Either a buyer violated his budget constraint, or the experimenter erred in recording a price in his first experiment. In Figure 2 there is plotted each contract (an accepted bid if the contract line passes through a ‘dot’; an accepted offer if the line passes through a ‘circle’) and the bids (‘dots’) and offers (‘circles’) that preceded each accepted bid or offer. One of the several advantages of computerized experimental markets is that the complete data of the market (all bids, offers, and contracts at their time of execution) are recorded accurately and non-invasively, and all experimental rules are enforced perfectly. In particular the violation of a budget constraint revealed in Figure 1, which is a perpetual problem with manually executed experiments, is not a problem when trading is perfectly computer monitored.

The rapid convergence shown in Figures 1 and 2 has not always extended to trading institutions other than the double auction. For example, the ‘posted offer’ pricing mechanism (associated with most retail markets), in which sellers post take it or leave it non-negotiable prices at the beginning of each period, yields higher prices and less efficient allocations than the double auction. This difference in performance becomes smaller with experienced subjects and with longer trading sequences in a given experiment (Ketcham et al., 1984). Similarly, a comparison of double auction with a sealed bid-offer auction finds the latter to be less efficient and to deviate more from the competitive equilibrium predictions (Smith et al., 1982). Thus, institutions have been demonstrated to make a difference in what we observe. The data and analysis strongly suggest that institutions make a difference because the rules (legal environment) make a difference, and the rules make a difference because they affect individual incentives.

II Brief interpretive history of the development of experimental economics

The two most influential early experimental studies represent the two most primary poles of experimental economics: the study of individual preference (choice) under uncertainty (Mosteller and Nogee, 1951) and of market behaviour (Chamberlin, 1948). The investigation of uncertainty and preference has focused on the testing of von Neumann–Morgenstern–Savage subjective expected utility theory. Battalio, Kagel and others have pioneered in the testing of the Slutsky–Hicks commodity demand and labour supply preferences using humans (1973) and animals (1975). A series of large-scale field experiments in the 1970s extended the experimental study of individual preference to the measurement of the effect of the negative income tax and other factors on labour supply and to the measurement of the demand for electricity, housing and medical services.

Since the human species has been observed to participate in market exchange for thousands of years, the experimental study of market behaviour is central to economics. Preferences are not directly observable, but preference theory, as an abstract construct, has been *postulated* by economists to be fundamental to the explanation and understanding of market behaviour. In this sense the experimental study of group market behaviour depends upon the study of individual preference behaviour. But this intellectual history should not obscure the fact that the study of markets and the study of preferences need not be construed as inseparable. Adam Smith clearly viewed the human ‘propensity to truck, barter and exchange’ (and *not* the existence of human preferences) as axiomatic to the scientific study of economic behaviour. Obversely, the work of Battalio and Kagel showing that animals behave as if they had Slutsky–Hicks preferences makes it plain that substitution behaviour is an important cross species characteristic, but that such phenomena need not be associated with market exchange.

A significant feature of Chamberlin's (1948) original work is that it concerned the study of behaviourally complete markets; that is all trades, including purchases as well as sales, were executed by active subject agents. This feature has continued in the subsequent bilateral bargaining experiments of Siegel and Fouraker (1960) and in market experiments (Smith, 1962, 1982; Williams and Smith, 1984) such as those discussed in section I. This feature was not present in the early and subsequent experimental oligopoly literature (Hoggatt, 1959; Sauermann and Selten, 1959; Shubik, 1962; Friedman, 1963), in which the demand behaviour of buyers was simulated, that is, programmed from a specified demand function conditional on the prices selected in each ‘trading’ period by the sellers. This simulation of demand behaviour is justified as an intermediate step in testing models of seller price behaviour that assume passive, simple maximizing, demand-revelation behaviour by buyers. But the conclusions of such experimental studies should not be assumed to be applicable, even provisionally, to any observed complete market without first showing that the experimental results are robust with respect to the substitution of subject buyers for simulated buyers.

III The functions of market experiments in microeconomic analysis

A conceptual framework for clarifying some uses and functions of experiments in microeconomics can be articulated by suitable modification and adaptation (Smith, 1982) of the concepts underlying the adjustment process, as in the welfare economics literature (see references to Hurwicz and Reiter in Smith, 1982). In this literature a microeconomic environment consists of a list of agents $\{1, \ldots, N\}$, a list of commodities and resources $\{1, \ldots, K\}$, and certain characteristics of each agent i, such as the agent's preferences (utility)

u_i, technological (knowledge) endowment T^i, and commodity endowment w_i. Thus agent i is defined by the triplet of characteristics $E^i = (u^i, T^i, w^i)$ defined on the K-dimensional commodity space. A microeconomic *environment* is defined by the collection $E = (E^1, \ldots, E^N)$ of these characteristics. This collection represents a set of primitive circumstances that condition agents' interaction through institutions. The superscript i, besides identifying a particular agent, also means that these primitive circumstances are in their *nature* private: it is the individual who likes, works, knows and makes.

There can be no such thing as a credible institution-free economics. Institutions define the property right rules by which agents communicate and exchange or transform commodities within the limits and opportunities inherent in the environment, E. Since markets require communication to effect exchange, property rights in messages are as important as property rights in goods and ideas. An institution specifies a language, $M = (M^1, \ldots, M^N)$, consisting of message elements $m = (m^1, \ldots, m^N)$, where M^i is the set of messages that can be sent by agent i (for example, the range of bids that can be sent by a buyer). An institution also defines a set of allocation rules $h = (h^1(m), \ldots, h^N(m))$ and a set of cost imputation rules $c = (c^1(m), \ldots, c^N(m))$, where h^i(m) is the commodity allocation to agent i and $c^i(m)$ is the payment to be made by i, each as a function of the messages sent by all agents. Finally, the institution defines a set of adjustment process rules (assumed to be common to all agents), $g(t_0, t, T)$, consisting of a starting rule, $g(t_0, \cdot, \cdot)$, a transition rule, $g(\cdot, t, \cdot)$, governing the sequencing of messages, and a stopping rule, $g(\cdot, \cdot, T)$, which terminates the exchange of messages and triggers the allocation and cost imputation rules. Each agent's property rights in communication and exchange is thus defined by $I^i = (M^i, h^i(m), c^i(m), g(t_0, t, T))$. A microeconomic *institution* is defined by the collection of these individual property right characteristics, $I = (I^1, \ldots, I^N)$.

A microeconomic *system* is defined by the conjunction of an environment and an institution, $S = (E, I)$. To illustrate a microeconomic system, consider an auction for a single indivisible object such as a painting or an antique vase. Let each of N agents place an independent, certain, monetary value on the item $v_1, \ldots, v_N$, with agent i knowing his own value, v_i, but having only uncertain (probability distribution) information on the values of others. Thus $E^i = (v^i; P(v), N)$. If the exchange institution is the 'first price' sealed-bid auction, the rules are that all N bidders each submit a single bid any time between the announcement of the auction offering at t_0, and the closing of bids, at T. The item is then awarded to the maker of the highest bid at a price equal to the amount bid. Thus, if the agents are numbered in descending order of the bids, the first price auction institution $I_1 = I_1^1 = [h^1(m) = 1, c^1(m) = b_1]$ and $I_1^i = [h^i(m) = 0, c^1(m) = 0], i > 1$, where $m = (b_1, \ldots, b_N)$ consists of all bids tendered. That is, the item is awarded to the high bidder, $i = 1$, who pays b_1, and all others receive and pay nothing. This contrasts with the 'second price' sealed-bid auction $I_2 = (I_2^1, \ldots, I_2^N)$ in which $I_2^1 = [h^1(m) = 1, c^1(m) = b_2]$ and $I_2^i = [h^i(m) = 0, c^i(m) = 0], i > 1$; that is, the highest bidder receives the allocation but pays a price equal to the second highest bid submitted.

Another example is the English or progressive oral auction, whose rules are discussed under the entry auctions (experiments). It should be noted that the 'double oral' auction, used extensively in stock and commodity trading and in the two experimental markets discussed in section I, is a two-sided generalization of the English auction.

A microeconomic system is activated by the behavioural choices of agents in the set M. In the static, or final outcome, description of an economy, agent *behaviour* can be defined as a function (or correspondence) $m^i = \beta^i(E^i|I)$ carrying the characteristics E^i of agent i into a message m^i, conditional upon the property right specifications of the operant institution I. If all exchange-relevant agent characteristics are included in E^i, then $\beta \equiv \beta^i$ for all i. Given the message-sending behaviour of each agent, $\beta(E|I)$, the institution determines the outcomes

$$h^i(m) = h^i[\beta(E^1|I), \ldots, \beta(E^N|I)]$$

and

$$c^i(m) = c^i[\beta(E^1|I), \ldots, \beta(E^N|I)].$$

Within this framework we see that agents do not choose allocations directly; agents choose messages with institutions determining allocations under the rules that carry messages into

allocations. (You cannot choose to 'buy' an auctioned item; you can only choose to raise the standing bid at an English auction or submit a particular bid in a sealed bid auction.) However, the allocation and cost imputation rules may have important incentive effects on behaviour, and therefore messages will in general depend on these rules. Hence, market outcomes will result from the conjunction of institutions' and agents' behaviour.

A proper theory of agents' behaviour allows one to deduce a particular β function based on assumptions about the agent's environment and the institution, and his motivation to act. Auction theory is perhaps the only part of economic theory that is fully institution specific. For example, in the second price sealed bid auction it is a dominant strategy for each agent simply to bid his or her value; that is

$$b^i = \beta(E^i|I_2) = \beta(v_i|I_2) = v_i, \quad i = 1, \ldots, N.$$

The resulting outcome is that $b_1 = v_1$ is the winning bid and agent 1 pays the price v_2. Similarly, in the English auction, agent 1 will eventually exclude agent 2 by raising the standing bid to v_2 (or somewhat above), and obtain the item at this price. In the first price auction Vickrey proved that if each agent maximizes expected surplus ($v_i - b_i$) in an environment with $P(v) = v$ (the v_i are drawn from a constant density on [0, 1]), then we can deduce the noncooperative equilibrium bid function, $b_i = \beta(E^i|I_1) = \beta[v_i, P(v), N|I_1] = (N-1)v_i/N$ (see the entry on auctions (experiments) for a more complete discussion).

With the above framework it is possible to explicate the roles of theory and experiment, and their relationship, in a progressive research programme (Lakatos, 1978) of economic analysis. But to do this we must first ask two questions:

(1) 'Which of the elements of a microeconomic system are not observable?' The nonobservable elements are (i) preferences, (ii) knowledge endowments, and (iii) agent message behaviour, $\beta(E^i|I)$. Even if messages are available and recorded, we still cannot observe message behaviour functions because we cannot observe, or vary, preferences. The best we can do with field observations of outcomes is to interpret them in terms of models based on assumptions about preferences (Cobb-Douglas, constant elasticity of substitution, homothetic), knowledge (complete, incomplete, common), and behaviour (cooperative, noncooperative). Any 'tests' of such models must necessarily be joint tests of all of these unobservable elements. More often the econometric exercise is parameter estimation, which is conditional upon these same elements.

(2) 'What would we like to know?' We would like to know enough about how agents' behaviour is affected by alternative environments and institutions so that we can classify them according to the mapping they provide into outcomes. Do some institutions yield Pareto optimal outcomes, and/or stable prices, and, if so, are the results robust with respect to alternative environments?

These two questions together tell us that what we want to know is inaccessible in natural experiments (field data) because key elements of the equation are unobservable and/or cannot be controlled. If laboratory experiments are to help us learn what we want to know, certain precepts that constitute proposed sufficient conditions for a valid controlled microeconomic experiment must be satisfied:

1. (1) *Non-satiation* (or monotonicity of reward). Subject agents strictly prefer any increase in the reward medium, π; that is $U_i \bullet (\pi_i)$ is monotone increasing for all i.
2. (2) *Saliency*. Agents have the unqualified right to claim rewards that increase (decrease) in the good (bad) outcomes, x_i, in an experiment; the institution of an experiment renders these rewards salient by defining outcomes in terms of the message choices of agents.
   In both the field and the laboratory it is the institution that induces value on messages, given each agent's (subjective) value of commodity outcomes. In the laboratory we use a monetary reward function to induce utility value on the abstract accounting outcomes ('commodities') of an experiment. Thus, agent i is given a concave schedule, $V_i(x_i)$, defining the 'redemption value' in dollars for x_i units purchased in an experimental market, and is assured of receiving a net payment equal to $V_i(x_i)$ less the purchase prices of the x_i units in the market. If the x_i units are all purchased at price p (which is the assumption used to derive a hypothetical demand schedule) the agent is paid $\pi_i = V_i(x_i) - px_i$, with utility $u^i(x_i) = U_i(\pi_i(x_i))$. In defining demand it is assumed that the agent directly chooses x_i (that is $x_i = m_i$). Therefore, if i maximizes $u^i(x_i) = U_i[V_i(x_i) - px_i]$, then at a maximum we have $U_i'[V_i'(x_i) - p] = 0$, giving the demand function $x_i = V_i'^{(-1)}(p)$ if $U_i' > 0$, where $V_i'^{(-1)}$ is the inverse of i's marginal redemption value of x_i units. (The same procedure for a seller using a cost function $C_j(x_j)$ and paying $px_j - C_j(x_j)$ allows one to induce a marginal cost supply of j.)
   This illustration generalizes easily: if the joint redemption value is $V_i(x_i, y_i)$ for two abstract commodities (x_i, y_i), $u^i = U_i[V^i(x_i, y_i)]$ induces an indifference map given by the level curves of $V^i(x_i, y_i)$, on (x_i, y_i), with marginal rate of substitution $U_i'V_x^i / U_i'V_y^i = V_x^i / V_y^i$, if $U_i' > 0$. If $V^i(x_i, \bullet X)$ the reward function, with x_i a private and X a common

(public) outcome good, we are able to control preferences in the study of public good allocation mechanisms, or if

$$X = \sum_{i=1}^{N} x_i$$

we are poised to study allocation with an ‘atmospheric’ externality (Coursey and Smith, 1985).

The first two precepts are sufficient to allow us to assert that we have created a microeconomic sysem $S = (E, I)$ in the laboratory. But to assure that we have created a controlled microeconomy, we need two additional precepts:

3. (3) *Dominance*. Own rewards dominate any subjective costs of transacting (or other motivation) in the experimental market.
   As with any person, subject agents may have variables other than money in their utility functions. In particular, if there is cognitive and kinesthetic (observe the traders on a Stock Exchange floor) disutility associated with the message-transaction process of the institution, then utility might be better written $U_i(\pi_i, m^i)$. To the extent that this is so we induce a smaller demand on i with the payoff $V_i(x_i)$ than was computed above, and we lose control over preferences. As a practical matter experimentalists think the problem can usually be finessed by using rewards that are large relative to the complexity of the task, and by adopting experimental procedures that reduce complexity (e.g. using the computer to record decisions, perform needed calculations, provide perfect recall, etc.). Another approach, as noted in Section I, is to pay a small commission for each trade to compensate for the subjective transaction costs.
4. (4) *Privacy*. The subjects in an experiment each receive information only on his/her own reward schedule.
   This precept is used to provide control over interpersonal utilities (payoff externalities). Real people may experience negative or positive utilities from the rewards of others, and to the extent that this occurs we lose control over induced demand, supply and preference functions. Remember that the reward functions have the same role in an experiment that preference functions have in the economy, and the latter preferences are private and non-observable.
   If our interest is confined to testing hypotheses from theory, we are done. Precepts (1)–(4) are sufficient to provide rigorous tests of the theorist's ability to model individual and market behaviour. But one naturally asks if replicable results from the laboratory are transferable to field environments. This requires:
5. (5) *Parallelism*. Propositions about behaviour and/or the performance of institutions that have been tested in one microeconomy (laboratory or field) apply also to other microeconomies (laboratory or field) where similar *ceteris paribus* conditions hold.

Astronomy, meteorology, biology and other sciences use the maintained hypothesis that the same physical laws hold everywhere. Economics postulates that when the environment and institution are the same, behaviour will be the same; that is, behaviour is determined by a relatively austere subset of life's parameters. Whether this is ‘true’ is an empirical question. Hence, when one experimentalist studies variations on the treatment variables of another it is customary to replicate the earlier work to check parallelism. Similarly, one must design field experiments, or devise econometric models using non-experimental field data, that provide tests of the transferability of experimental results to any particular market in the field. Only in this way can questions of parallelism be answered. They are not answered with speculations about alleged differences between the experimental subject's behaviour and (undefined) ‘real world’ behaviour. The experimental laboratory *is* a real world, with real people, real institutions, real payoffs and commodities just as real as stock certificates and airline travel vouchers, both of which have utility because of the claim rights they legally bestow on the bearer.

IV Classifying the application of experimental methods

There are many types of experiments and many fields of economic study to which experimental methods have been applied.

The experimental study of auctions makes the most extensive use of models of individual behaviour based explicitly on the message requirements of the different institutions. This literature provides test comparisons of predicted behaviour, $m^i = \beta^i(E_i|I)$, with observations on individual choice, $\hat{m}^i = \beta^i(\hat{E}^i|I)$ for given realizations, $\hat{E}^i$ (such as values, $\hat{v}_i$, where they are assigned at random). The large literature on experimental double auctions makes no such individual comparisons, because the theoretical literature had not yielded tractable models of individual bid-offer behaviour (but recent contributions by Friedman (1984), and Wilson (1984) are providing such models). Here as in most other areas of experimental research the comparisons are between the predicted price–quantity outcomes of static theory (such as competitive, monopoly, and Cournot models), and observed outcomes. But double auctions have been studied (see references in Smith, 1982) in a variety of environments; for example, the effect of price floors and ceilings have been examined (see references in Plott, 1982). In all cases these studies are making comparisons. In *nomotheoretical* experiments one compares theory and observation, whereas in *nomoempirical* experiments one

compares the effect of different institutions and/or environments as a means of documenting replicable empirical ‘laws’ that may stimulate modelling energy in new directions. The idea that formal theory must precede meaningful observation does not account for most of the historical development of science. *Heuristic* or exploratory experiments that provide empirical probes of new topics and new experimental methods should not be discouraged.

In industrial organization, and antitrust economics, experimental methods have been applied to examine the effects of monopoly, conspiracy, and alleged anticompetitive practices, and to study the concept of natural monopoly and its relation to scale economics, entry cost and the contestable markets hypothesis (see references in Plott, 1982; Smith, 1982; Coursey et al., 1984).

An important development in the experimental study of allocation processes has been the extension of experimental market methods to majority rule (and other) committee processes, and to market-like group processes for the provision of goods which have public or common outcome characteristics (loosely, public goods). These studies have examined public good allocation under majority (and Roberts’) rules for committee including the effect of the agenda (see the references to Fiorina and Plott, and Levine and Plott in Smith, 1982), and under compensated unanimity processes suggested by theorists (see the references in Coursey and Smith, 1985). Generally, this literature reports substantial experimental support for the theory of majority rule outcomes, the theory of agenda processes (the sequencing of issues for voting decisions), and for incentive compatible models of the provision of public goods.

See Also

- Allais Paradox
- efficient allocation
- preference reversals

Bibliography

Battalio, R., Kagel, J., Winkler, R., Fisher, E., Basmann, R. and Krasner, L. 1973. A test of consumer demand theory using observations of individual consumer purchases. *Western Economic Journal* 11, 411–28.

Chamberlin, E. 1948. An experimental imperfect market. *Journal of Political Economy* 56, 95–108.

Coursey, D., Isaac, M., Luke, M. and Smith, V. 1984. Market contestability in the presence of sunk (entry) costs. *RAND Journal of Economics* 15(1), 69–84.

Coursey, D. and Smith, V. 1985. Experimental tests of an allocation mechanism for private, public or externality goods. *Scandinavian Journal of Economics* 86, 468–84.

Friedman, D. 1984. On the efficiency of experimental double auction markets. *American Economic Review* 74, 60–72.

Friedman, J. 1963. Individual behavior in oligopolistic markets: an experimental study. *Yale Economic Essays* 3, 359–417.

Hoggatt, A. 1959. An experimental business game. *Behavioral Science* 4(3), 192–203.

Kagel, J., Battalio, R., Rachlin, H., Green, L., Basmann, R. and Klemm, W. 1975. Experimental studies of consumer behavior using laboratory animals. *Economic Inquiry* 13(1), 22–38.

Ketcham, J., Smith, V. and Williams, A. 1984. A comparison of posted-offer and double-auction pricing institutions. *Review of Economic Studies* 51, 595–614.

Lakatos, I. 1978. *The Methodology of Scientific Research Programmes, Philosophical Papers*, vol. 1, ed. J. Worrall and G. Currie. Cambridge: Cambridge University Press.

Mosteller, F. and Nogee, P. 1951. An experimental measurement of utility. *Journal of Political Economy* 59, 371–404.

Plott, C. 1982. Industrial organization theory and experimental economics. *Journal of Economic Literature* 20, 1485–527.

Sauermann, H. and Selten, R. 1959. Ein Oligopolexperiment. *Zeitschrift für die Gesamte Staatswissenschaft* 115, 427–71.

Shubik, M. 1962. Some experimental non zero sum games with lack of information about the rules. *Management Science* 81, 215–34.

Siegel, S. and Fouraker, L. 1960. *Bargaining and Group Decision Making*. New York: McGraw-Hill.

Smith, V. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70, 111–37.

Smith, V. 1982. Microeconomic systems as experimental science. *American Economic Review* 72, 923–55.

Smith, V., Williams A., Bratton, K. and Vannoni, M. 1982. Competitive market institutions: double auctions versus sealed bid-offer auctions. *American Economic Review* 72, 58–77.

Williams, A. and Smith, V. 1984. Cyclical double-auction markets with and without speculators. *Journal of Business* 57(1) Pt 1, 1–33.

Wilson, R. 1984. Multilateral exchange. Working Paper No. 7, Stanford University, August.

How to cite this article

Smith, Vernon L. "experimental methods in economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000186> doi:10.1057/9780230226203.0531

The New Palgrave Dictionary of Economics Online

experimental methods in environmental economics

Jason F. Shogren

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Experimental methods have long played a role in environmental economics. The strong link emerged due to the need to make decisions within the complex confluences of markets, missing markets, and no markets. Two broad areas of experimental work are discussed, institutional and valuation. Institutional experiments help reveal how good ideas for environmental protection can go badly with poorly understood rules and incentives; valuation experiments help illustrate how values for environmental protection depend on the socialization created, directly or indirectly, by the exchange institutions in operation.

Keywords

asymmetric information; bargaining; Coase Theorem; common property resources; endowment effect; environmental economics; experimental methods in environmental economics; hypothetical bias; institutional experiments; land conservation; non-market valuation; Pigouvian taxes; pollution permits; provision point mechanism; public goods; social costs; transaction costs; valuation experiments; value elicitation; willingness to accept compensation; willingness to pay.

Article

Environmental policy is designed within the confluence of markets, missing markets, and no markets. Within this mixture, economists offer working rules to help make outcomes more efficient, usually based on ideas formed by rational choice theory. The rules ask decision-makers to compare benefits in relation to costs, to account for the risks and gains across time and space for winners and losers, to facilitate the movement of resources from low-value uses to high-value uses, and to equate incremental gains per cost across policy actions. The environmental economic challenge is to find effective decision rules that will help move an economy towards efficient resource allocation in the face of market failure, for example, externalities, non-rival consumption, non-excludable net benefits, nonconvexities and

asymmetric information (see Hanley, Shogren and White, 2007).

Experimental methods have proven to be a useful tool in addressing this challenge. Environmental economists used experimental methods relatively early on, following the lead of Vernon Smith, Charles Plott and other pioneers. Experimental methods began to take hold in the 1980s, primarily in the area of non-market valuation (see Bohm, 1972; Bennett, 1983; Knetsch and Sinden, 1984; Coursey, Hovis and Schulze, 1987). Today, experimental economic research is commonplace in environmental economic discussions and research programmes, with data being generated both in the laboratory and field (see for example the research in Cherry, Kroll and Shogren, 2007). Experiments in this area can be grouped broadly into two categories, *institutional* and *valuation*. Institutional experiments test-bed new institutions such as marketable pollution permits and ambient non-point pollution taxes prior to implementation; valuation experiments use the laboratory or field to study how people value goods and services that are not otherwise bought and sold in markets.

Institutional experiments build on traditional designs to test the efficiency of alternative exchange mechanisms under different economic circumstances. Usually the institutions under examination are those theoretically argued to correct for some market failure. Benefits and costs in these institutional experiments are *induced* by the experimenter – buyers have pre-assigned resale values; sellers have designated induced costs; and the goal is to measure the efficiency of a set of alternative incentive schemes.

In contrast, valuation experiments flip the institutional experiment on its head, using experimental methods to elicit preferences for some particular private or public good given alternative market and non-market circumstances. Here eliciting *homegrown* preferences or values – those residing within the minds of people – is of ultimate interest. Research in value elicitation has been environmental economics' most unique contribution to experimental economics. The work has produced insight into how the framing of a question affects values, how different demand-revealing incentives elicit different values, and how unintentional cues affect a person's value for a good. Consider now a few examples of institutional and valuation experiments used in environmental economics.

Institutional experiments

Institutional experiments focus on evaluating market and non-market solutions to environmental problems. The key to these institutional experiments rests in the dialogue between the laboratory and potential or actual applications to environmental policy. For decades, environmental policy around the globe has been proposed and implemented in the real world with minimal input from insight gathered using experimental economics methods. Today, however, this is changing. Researchers are now using experiments to help understand and affect policy development, and this link between the laboratory and policy is probably more rigorously explored in environmental economics than any other area (Bohm, 2003).

Institutional environmental economic experiments can be categorized as three broad areas – institutions to provide incentives to control externality problems that arise from pollution or land use; institutions to increase the voluntary provision of public goods, such as climate change, or to manage effectively common property, such as fishing zones; and institutions designed to manage resources through negotiation and cooperation, that is, the Coase theorem. We now briefly consider each in turn, starting

with early work, moving to current applications, and general principles.

First, experiments examining economic solutions to externality problems began in earnest with Plott's (1983) work on Pigovian taxation. Plott designed a competitive market of buyers and sellers who trade a valuable good. After first establishing that traders ignored negative social costs in a competitive market, he explored whether Pigovian taxes or tradable permits could equate private incentives with social costs. Both increased efficiency with repeated trading periods and quickly hit 100 per cent efficiency. Since then there has been an explosion of work examining incentive systems in a variety of settings, producing a growing and positive dialogue between policy proposals and insight from experimental studies.

Probably the most active area today remains the experimental work that tests the efficiency of tradable permit systems. Experimental methods have evaluated the efficacy of different trading rules in a variety of settings (for example, Bohm and Carlén, 1999). An important early example is the US Environmental Protection Agency's Acid Rain emission trading. This work revealed a basic flaw in the original design of the permit auction run by the Environmental Protection Agency (EPA) (see Cason, 1995; Cason and Plott, 1996). The laboratory results revealed how the EPA could increase the efficiency of the auction by changing how permits were allocated. Originally, buyers and sellers submitted bids and offers for emission permits, and the EPA set the market price discriminatively off the demand curve by first matching the seller with the lowest offer to the buyer with the highest bid. The matching then continued with the second lowest offer to the second highest bid, and so on, until the equilibrium quantity is reached. Rational sellers should see through this auction, and begin capturing rents by understating their true offer so they would be matched with a high bidder. Cason's laboratory results confirmed this intuition – sellers undercut each other to get into the high end of the market. The end result was an inefficient auction. Such lessons can be profitable, but insight like this should be made available before the regulatory tool is already in place, thus avoiding wasting resources due to inefficient design features. (For another important example comparing alternative trading institutions, see the tests of the RECLAIM market for the Los Angeles Basin by Ishikida et al., 2000).

Land conservation is a second area in which experimentally informed market designs have improved policy implementation. The Bush Tender auctions were designed to conserve land in Australia by creating a market where landowners bid to set aside specific units of land. Cason and Gangadharan (2004) examine how information about environmental benefits and a market clearing auction mechanism affect efficiency. Their results reveal an interesting pattern: people who did not know the environmental benefits provided by their private land were less likely to bid strategically in a conservation auction. Private ignorance reduces public expenditures. Based on this they suggest a provocative policy – a regulator might restrict the biological information publicly provided to landowners prior to running the auction. Another example of test-bedding is Parkhurst et al.'s (2002) agglomeration-bonus and smart-subsidy coordination game experiments, which illustrate an incentive scheme that can induce private landowners to create contiguous protected areas voluntarily. They compare a *smart subsidy* proposal, which creates an explicit link between neighbouring landowners with adjacent parcels, in relation to two standard policy options, compulsion and a standard fixed-fee subsidy. Their results show that a no-bonus mechanism always created fragmented habitat, whereas with the bonus, players found the first-best habitat reserve.

Second, environmental policy has long confronted the inherent efficiency issues associated with public goods and common property resources. These experimental games capture the elemental economic

problem that drives many environmental goods: non-rival and non-exclusive consumption lead to free riding and inefficient production levels. Experimental evidence reveals neither complete free riding nor full cooperation (Ledyard, 1995). As noted by Ostrom (2000), three types of people commonly inhabit public-good and common-property experiments: the standard rational egoists, the conditional cooperators, and the willing punishers. Conditional cooperators cooperate when they expect others to reciprocate; otherwise they do not. Within the standard game (see public goods experiments) rules can be manipulated to induce more or less cooperation depending on the mix of subject types, marginal payoffs, group size, communication, and voting with third-party enforcement.

For global environmental goods like climate change, a key policy issue is the impact on efficiency when a collective agreement has costly third-party enforcement. Punishment of free riders is a second-order public good; cooperation means bearing some private cost to sanction others. One relevant policy question is whether an institution based on a voting rule with a punishment mechanism can work to increase contributions to a public good. Kroll, Cherry and Shogren (2007) examined this in the laboratory and observed that voting alone does not increase cooperation; rather, if voters can pay to punish violators, contributions increase significantly. Overall efficiency for a voting-with-punishment rule exceeds the level observed for a voting-without-punishment rule. This result has implications for how policymakers think about institutions such as International Environmental Agreements (IEA), which are more likely to be successful if one nation is willing to act like the ‘global police’, and pay the costs of punishing violators (Barrett, 2003).

Another real-world policy issue is whether policymakers can use economic incentive devices in real-world applications to reveal public good demand (Bohm, 1972). Any proposed system has to ‘work’ – to provide the good when benefits exceed the provision costs – and has to be straightforward enough to be implemented in the field, characteristics that can be tested in the laboratory. One such mechanism is the *provision point* mechanism: if contributions meet or exceed a targeted provision cost, the public good is supplied to the group; otherwise, it is not. In a design that mimics field conditions, Rondeau, Schulze and Poe (1999) explored a mechanism in which contributions are returned if costs were not met. Their results suggest the provision point mechanism was ‘demand revealing in aggregate’ for a large group with heterogeneous preferences, suggesting that a relatively simple mechanism could be used in the field to elicit preferences, leading to the efficient provision of a public good.

Third, many observers and policymakers see place-based collaboration and bargaining as the future of environmental policy – arguing for more local control through negotiation and accountability (for example, Sabel, Fung and Karkkainen, 2000). Collaborative decision-making groups have begun to flourish in rural settings such as the western United States, and now number in the hundreds, ranging from informal grass-roots gatherings to government-mandated advisory councils. To an economist, this is the direct application of the Coase theorem – parties in dispute negotiating on a jointly acceptable agreement over resource use. Starting with Hoffman and Spitzer (1982), researchers have used experimental methods to test the robustness of collaborative decision-making underlying the Coase theorem. Hoffman and Spitzer's initial results supported the Coase theorem in that bargains were highly efficient. Harrison and McKee (1985) confirmed that Coasean bargaining under unilateral and joint property rights regimes can be efficient. Both experiments assumed the transaction costs of bargaining were zero. Recent experimental work has explored how bargaining efficiency is affected by positive transaction costs and addition friction due to large numbers of bargainers, property right insecurity,

delay costs, imperfect contract enforcement, asymmetric information, and uncertain final authority. The lesson from over two decades of Coase bargaining research is that people have to address the nature of transaction costs and friction.

We illustrate using Rhoads and Shogren's (2003) policy-driven Coasean bargaining experiment. Experts see two elements of consensus-based environmental protection as crucial for effective regulatory outcomes: *final authority*, so that a collaborative agreement is binding; and *information symmetry*, in which bargainers create a common information pool about player payoffs. Their results are consistent with the findings of the experts: final authority and information symmetry were necessary conditions for efficient Coasean bargaining. Without final authority, efficiency falls by two-thirds, and falls further with asymmetric information. If the policy objective is to make a negotiated agreement efficient, the policy challenge is to understand the trade-offs associated with granting or denying final authority to the local bargainers.

In summary, the general principle in institutional environmental economics experiments that has emerged over the years is that the germ of a good idea can be codified into a bad one if the rules of implementation trigger unintended incentives that undercut the efficiency of the system. Experimental methods can be used to reveal which good ideas are actually beneficial to control externalities, provide public goods, and facilitate collaboration – and which ideas are ultimately counterproductive.

Valuation experiments

Economists also use experimental methods to understand better the behavioural underpinnings of environmental valuation. Experiments can be used to address incentive and contextual questions that arise in assessing values through direct statements of preferences. Three general areas have emerged: rational valuation, direct elicitation of values, and exploring the effectiveness of hypothetical non-market valuation surveys (see Shogren, 2006).

First, economists assume people can provide rational statements of their preferences and values towards the environment. Rather than assume that people make rational choices and reveal consistent values for environmental protection, environmental economists use experiments to examine whether people's choices and stated values meet these criteria. Enough evidence of behavioural anomalies now exists to undercut this presumption (Kahneman and Tversky, 2000). Without an exchange institution to arbitrage his or her irrational choices, the unsocialized person can engage in behaviours inconsistent with rational choice theory (see Akerlof, 1997).

The key behavioural regularity that potentially undercuts all valuation work is the WTP–WTA gap. Rational choice theory suggests that with small income effects and many available substitutes, the willingness to pay (WTP) for a commodity and the willingness to accept (WTA) compensation to sell the same commodity should be about equal. But evidence suggests that WTA exceeds WTP by up to tenfold. The experimental WTP–WTA work can be divided into two camps: research that suggests the gap is based on a psychological *endowment effect* and that which points to weak market institutions. A person who assigns greater value to a good he or she already owns exhibits the endowment effect, which leads to higher WTA to sell the good than WTP to buy the identical good (Kahneman, Knetsch and Thaler, 1990). The market experience explanation says that people have naive expectations about what they can sell the good for outside an active market place. This experimental work showed that market-

like experience can remove the gap (see Shogren et al., 1994; 2001).

Second, valuation experiments are used to measure actual values for public and private goods (Lusk and Shogren, 2007). Direct valuation experiments are designed so that people buy and sell actual goods to elicit real values, in which researchers test how alternative exchange institutions affect these values. They entail real payments and binding budget constraints, and use auctions to sell goods for money, albeit within a stylized setting. Experimental designs are used to understand the balance between laboratory control and natural context, enabling researchers to learn things about behaviour that would have been impossible to discover from alternative tools. Subtle changes in experimental procedure affect behaviour, such as paying people before as opposed to after bidding, reporting the market-clearing price, and the novelty of the good.

Third, in the 1980s, Coursey and Schulze (1986) hoped the laboratory would be used more to test-bed field surveys. Today, in 2007, experiments are commonly employed to address problems in stated preference surveys such as hypothetical bias, calibration, surrogate bidding, and incentive compatibility. For instance, experiments have revealed time and again that *hypothetical bias* is real – people frequently promise more than they actually deliver. Experimental work has focused on trying to measure the degree of bias and what methods can be used to eliminate or reduce it in survey work. A good example is Cummings and Taylor (1999) who find that they can remove the hypothetical bias by telling a respondent about it.

In summary, choices and economic values emerge in the social context of an active exchange institution, and thus the measurement of value should not be separated from the interactive experience provided by an exchange institution. Institutions and the institutional context matter because experience can make rational choice more transparent to a person. Institutions also dictate the rules under which exchange occurs, and these rules can differ across settings. People can interpret differently the information conveyed by such settings. The reality is that most people make allocation decisions in several institutional settings each day – markets, missing markets, and unidentified markets. How does this institutional mix affect how people make their choices and form or state their preferences for environmental protection? This question is fundamental because it gives a reason for the purposeful actions underlying all valuation work.

Experimental work like the *rationality spillover* treatments in Cherry, Crocker and Shogren (2003) reveal that exposure to competition and discipline is needed to achieve rationality. In becoming rational, people refine their statements of value to better match their preferences. The contact with others who are making similar decisions in an exchange institution puts in context the economic maxim that choices have consequences and stated values have meaning for environmental valuation. Relying on rational theory to guide environmental valuation and policy makes more sense if people make, or act as if they make, consistent and systematic choices about certain and risky events. Valuation work in the laboratory needs to continue to address the economic conditions under which the presumption of rationality is supported and when it is not, which in turn has implications for the values we directly elicit.

Concluding remarks

Through the use of experimental methods, environmental economists now understand better how people learn about and react to incentives, institutions and information. They can compare how decisions are

made with and without real economic commitments, within and without active exchange institutions, and with and without signals of value. They can then delve into what the results suggest for *ex ante* questionnaire design, *ex post* statistical evaluation, and, more importantly perhaps, economic theory itself. The environmental economics literature continues to follow the classic experimental strategy: start simply and add complexity slowly so as to understand which factors matter, and why.

In addition, all experiments in environmental economics reveal the perpetual tension between *control* and *context*. At the core, the experimental method is about *control*. One controls the experimental circumstances by trying to change only one variable at a time, which will reduce problems of confounding. Without control, it is unclear whether unpredicted behaviour is due to a poor theory or experimental design, or both. In contrast, others argue that *context* is desirable to avoid a setting that is too sterile and too removed from reality for something so real as environmental policy. Context affects participants' motivation.

Finally, as evidence continues to accumulate, a clearer and more definitive picture will emerge of how our institutions affect the efficiency and perceived value of environmental policies. The future of experimental work will be to design institutions that address the combination of market failure and behavioural anomalies. Otherwise we could find environmental economics falling into a new second-best problem: if we correct market failure without addressing behavioural biases, we might actually reduce overall social welfare.

See Also

- Coase theorem
- environmental economics
- experimental economics
- market failure
- pollution permits
- public goods experiments
- value elicitation

Bibliography

Akerlof, G. 1997. Social distance and social decisions. *Econometrica* 65, 1005–27.

Barrett, S. 2003. *Environment and Statecraft: The Strategy of Environmental Treaty-Making*. New York: Oxford University Press.

Bennett, J. 1983. Validating revealed preferences. *Economic Analysis and Policy* 13, 2–17.

Bohm, P. 1972. Estimating demand for public goods: an experiment. *European Economic Review* 3, 111–30.

Bohm, P. 2003. Experimental evaluations of policy instruments. In *Handbook of Environmental Economics*, vol. 1, ed. K.G. Mäler and J. Vincent. Amsterdam: North-Holland, pp. 438–60.

Bohm, P. and Carlén, B. 1999. Emission quota trade among the few: laboratory evidence of joint implementation among committed countries. *Resource and Energy Economics* 21, 43–66.

Cason, T. 1995. An experimental investigation of the seller incentives in the EPA's emission trading auction. *American Economic Review* 85, 905–22.

Cason, T. and Plott, C. 1996. EPA's new emissions trading mechanism: a laboratory evaluation. *Journal of Environmental Economics and Management* 30, 133–60.

Cason, T. and Gangadharan, L. 2004. Auction design for voluntary conservation programs. *American Journal of Agricultural Economics* 86, 1211–7.

Cherry, T., Crocker, T. and Shogren, J. 2003. Rationality spillovers. *Journal of Environmental Economics and Management* 45, 63–84.

Cherry, T., Kroll, S. and Shogren, J. 2007. *Experimental Methods, Environmental Economics*. London: Routledge.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Coursey, D. and Schulze, W. 1986. The application of laboratory experimental economics to the contingent valuation of public goods. *Public Choice* 49, 47–68.

Coursey, D., Hovis, J. and Schulze, W. 1987. The disparity between willingness to accept and willingness to pay measures of value. *Quarterly Journal of Economics* 102, 679–90.

Cummings, R. and Taylor, L. 1999. Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. *American Economic Review* 83, 649–65.

Hanley, N., Shogren, J. and White, B. 2007. *Environmental Economics in Theory and Practice*, 2nd edn. London and New York: Palgrave.

Harrison, G. and McKee, M. 1985. Experimental evaluation of the Coase theorem. *Journal of Law and Economics* 28, 653–70.

Hoffman, E. and Spitzer, M. 1982. The Coase theorem: some experimental tests. *Journal of Law and Economics* 25, 73–98.

Ishikida, T., Ledyard, J., Olson, M. and Porter, D. 2000. Experimental testbedding of a pollution trading system: southern California's RECLAIM emissions market. In *Research in Experimental Economics*, vol. 8, ed. R.M. Isaac. Greenwich, CN: JAI Press, Elsevier Science.

Kahneman, D., Knetsch, J. and Thaler, R. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98, 1325–48.

Kahneman, D. and Tversky, A. 2000. *Choices, Values and Frames*. Cambridge: Cambridge University Press.

Knetsch, J. and Sinden, J.A. 1984. Willingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of values. *Quarterly Journal of Economics* 99, 507–21.

Kroll, S., Cherry, T. and Shogren, J. 2007. Voting, punishment and public goods. *Economic Inquiry* 45, 557–70.

Ledyard, J. 1995. Public goods: a survey of experimental research. In *Handbook of Experimental Economics*, ed. J. Kagel and A. Roth. Princeton, NJ: Princeton University Press.

Lusk, J. and Shogren, J. 2007. *Experimental Auctions: Methods and Applications in Economic and Marketing Research*. New York: Cambridge University Press.

Ostrom, E. 2000. Collective action and the evolution of social norms. *Journal of Economic Perspectives* 14(3), 137–48.

Parkhurst, G., Shogren, J., Bastian, C., Kivi, P., Donner, J. and Smith, R. 2002. Agglomeration bonus: an incentive mechanism to reunite fragmented habitat for biodiversity conservation. *Ecological Economics* 41, 305–28.

Plott, C. 1983. Externalities and corrective policies in experimental markets. *Economic Journal* 93, 106–27.

Rhoads, T. and Shogren, J. 2003. Regulation through collaboration: final authority and information symmetry in environmental Coasean bargaining. *Journal of Regulatory Economics* 24, 63–89.

Rondeau, D., Schulze, W. and Poe, G. 1999. Voluntary revelation of the demand for public goods using a provision point mechanism. *Journal of Public Economics* 72, 455–70.

Sabel, C., Fung, A. and Karkkainen, B. 2000. *Beyond Backyard Environmentalism*. Boston, MA: Beacon Press.

Shogren, J. 2006. Experimental methods and valuation. In *Handbook of Environmental Economics*, vol. 2. ed. K.G. Mäler and J. Vincent. Amsterdam: North-Holland.

Shogren, J., Shin, S., Hayes, D. and Kliebenstein, J. 1994. Resolving differences in willingness to pay and willingness to accept. *American Economic Review* 84, 255–70.

Shogren, J., Cho, S., Koo, C., List, J. Park, C., Polo, P. and Wilhelmi, R. 2001. Auction mechanisms and the measurement of WTP and WTA. *Resource and Energy Economics* 23, 97–109.

How to cite this article

Shogren, Jason F. "experimental methods in environmental economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000299>
doi:10.1057/9780230226203.0532

The New Palgrave Dictionary of Economics Online

experiments and econometrics

Daniel E. Houser

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Experimetrics’ refers to formal procedures used in designed investigations of economic hypotheses. Fundamental experimetric contributions by Ronald A. Fisher provided the foundation for a rich literature informing the design and analysis of economics experiments. Key components of this foundation include the concepts of randomization, independence and blocking. Experimetric analysis plays a central role in advancing economic models, and will gain further importance as scholars adopt increasingly sophisticated designed research programmes to illuminate positive economic theory.

Keywords

blocking; causal inference; Exact Test (R. A. Fisher); experimental economics; experiments and econometrics; experimetrics; Fisher, R. A.; individual learning in games; neuroeconomics; quantal response equilibrium; randomization

Article

1 Introduction

‘Experimetrics’ refers to formal procedures used in designed investigations of economic hypotheses. A series of pathbreaking experimetric contributions by Ronald A. Fisher, written largely during the 1920s and early 1930s, elucidated fundamental concepts in the design and analysis of experiments (see, for example, Box, 1980, for a survey). He was first to obtain rigorous experimetric results on the importance of randomization, independence and blocking, and he created many powerful analysis tools that remain widely used, including Fisher's nonparametric Exact Test (Fisher, 1926; see also Fisher, 1935). Controlled experiments allow compelling scientific inferences with respect to hypotheses of interest. Many economic experiments inform hypotheses regarding primitives assumed to be constant within an experiment (for example, preferences or decision strategies), or the effects on economic outcomes of

changes in institutions (for example, comparing different auction rules or unemployment regulations; see experimental economics). One conducts controlled experiments to inform economic hypotheses because relevant naturally occurring data typically include noise of unknown form and magnitude outside the investigator's control. Econometric procedures can go some distance towards solving this problem, but even sophisticated approaches often allow only limited conclusions.

For example, suppose one wanted to investigate the (causal) effect of caffeine on heart rhythms. One approach is to obtain a random sample of ‘heavy’ coffee drinkers and compare them with a random sample of people who do not use caffeine. Because it is not possible with naturally occurring data to control the reason a person falls into a category, discovering that people with greater caffeine consumption have more cardiac episodes need not imply a causal caffeine effect. The reason is that a preference for coffee may stem from a biological characteristic that is itself causally tied to irregular cardiac events.

An advantage of designed investigations is that they allow cogent inference regarding causal effects through the appropriate use of randomization, independence and blocking.

1.1 Randomization

Experiments with randomized designs allow compelling causal inference. The reason is that randomly assigning participants to treatments, and randomly assigning treatments to dates and times, minimizes the possibility of systematic error. In the caffeine example, intentionally assigning heavy caffeine drinkers exclusively to a caffeine treatment generates a systematic error and invalidates causal inference. However, an experiment where subjects are randomly assigned to caffeine and no-caffeine treatments independent of their typical caffeine use allows one to draw appropriate inferences regarding causal relationships.

1.2 Independence

Randomization also helps to ensure independence both within and between treatments’ observations. Loosely speaking, observations are independent if information about one observation does not provide information about another. Independence is critical for many experimetric analyses, and its failure can lead to misleading conclusions. An objective randomization procedure for treatment assignments insures against the possibility that participants in one treatment might unintentionally systematically vary from other treatments’ participants.

1.3 Blocking

Causal relationships can be assessed with greater precision through ‘blocking’. Blocking is a design procedure with which an experimenter can separate treatment effects from nuisance sources of data variation. In the above, heart rhythms might be affected by both caffeine and anxiety over the process of measuring heart rhythms. Especially because it is expected to differ between participants, anxiety is a source of nuisance variation that clouds inferences regarding caffeine effects. To address this one could ‘block’ by participant. This involves measuring each subject both with and without caffeine (in separate,

randomly ordered trials). Caffeine effects are measured as the difference between trials, thus mitigating noise due to individual anxiety effects.

2 Experimetrics toolbox

Although many specialized experimetric tools have been developed, the experimetrics toolbox also includes a large number of general purpose procedures that have become standard in the experimental economics literature. A regular concern is that independence is not satisfied. The failure of independence can occur because of ‘session’ effects, meaning that there is less behavioural variation within than between sessions. Violations of independence can also occur if repeated measurements are taken on the same individual due to individual effects. Standard procedures can address this. Sessions can be treated as fixed effects, and random effects can be used to control for individual differences. The resulting ‘mixed effect’ model can be analysed using standard parametric, panel-data procedures (see, for example, Frechette, 2005).

Also in the toolbox is the McKelvey and Palfrey (1995) ‘quantal response equilibrium’ (QRE) framework (see quantal response equilibria). QRE is a parametric procedure for analysing data from finite games. The key idea is to incorporate errors into players’ best response functions, thus creating ‘quantal response’ functions. This results in an extremely flexible model that can rationalize a wide variety of behaviours. Haile, Hortacsu and Kosenok (2006) point out that this flexibility comes at a cost: in general QRE can rationalize any distribution of behaviour in any normal form game, and imposes no falsifiable restrictions without additional assumptions on the stochastic components of the model. Thus, those who wish to implement QRE analyses face the experimetric challenge of creating designs within which such assumptions are defensible.

For reasons including sample size and robustness, the experimetrics toolbox includes many nonparametric procedures (see Siegel and Castellan, 1988, for a user-friendly textbook treatment of popular nonparametric approaches). For example, Mann–Whitney tests, and their k-sample generalization due to Jonckheere (1954), are frequently used to compare medians among treatments’ data. Also common is Fisher's Exact Test, which uses all the information in the data and is the most powerful nonparametric approach to inference with respect to differences among treatments. Its use is limited by the fact that it can be computationally cumbersome to implement when the numbers of treatments or observations are large.

3 External validity

An experiment's conclusions are ‘externally valid’ if they can be extrapolated to other environments. To rigorously address external validity requires that the source of treatment effects can be identified, which in turn implies a fundamental rule of experiment design: within any good experiment, any treatment can be matched with another that differs from it in exactly one way.

External validity is both important and subtle. For example, consider the well-known ‘dictator game’ where one participant is assigned the role of ‘dictator’, and the other ‘receiver’. The dictator is given $20, and the receiver nothing. The dictator is told to split the $20 between herself and her receiver in any way she likes, after which the experiment ends. A widely replicated result is that a large fraction of

dictators send half ($10) to an anonymous stranger, and one might question whether this finding is externally valid. In particular, there is no evidence that this behaviour is prevalent among winners of naturally occurring lotteries.

There are clear similarities between the situations of lottery winners and dictators. Still, the fact that actions of dictators in laboratory games do not match actions of lottery winners does not necessarily mean that dictator games lack external validity. The reason is that identical decision strategies can imply different decisions in different environments. For example, recent research provides compelling evidence that dictators' decisions are tightly connected to their beliefs regarding the decisions of others who have faced this same situation: dictators give because they believe other dictators give (Bicchieri and Xiao, 2007). This mechanism plausibly guides decisions in naturally occurring environments. In particular, lottery winners do not give because they believe other lottery winners do not give large fractions of their winnings to anonymous strangers. Thus, external validity does not require that one be able to match actions in an experiment to actions in another environment. Rather, an experiment is externally valid if one can extrapolate to novel contexts its conclusions with respect to individual or strategic decision processes.

4 Applied experimetrics research

An important application of experimetrics is to discriminate between many competing theories of learning that have emerged (see individual learning in games). Doing this includes significant experimetric challenges, as it requires one to account for heterogeneity in the way subjects learn. The reason is that not doing so will tend to bias fit statistics in favour of reinforcement (and hybrid) models. Wilcox (2006) shows the reason is that reinforcement models condition behaviour on informative functions of past choices, and in the presence of learning heterogeneity these choices will carry idiosyncratic parameter information not otherwise incorporated into the specification. Having said this, it is also the case that many data-sets from typical learning experiments can be roughly equally well described by many different learning models (Salmon, 2001). Consequently, the 'best' model can be highly sensitive to the particular criterion one uses for model selection, as well as the particular experiment under consideration (Feltovich, 2000). As a result, in-sample fit is often good, but this does not necessarily imply that much has been learned about the way in which people actually learn and make choices (Salmon, 2001).

Knowing how people make choices is critical to advance both economic theory and institution design. Consequently, a significant experimetric literature explores how people make decisions in complex environments, with a focus on characterizing the nature and number of different 'decision rules' at use in a population. Most approaches to accomplishing this require pre-specifying the decision rules the researcher believes people could follow, and then using choice data to assign one of those rules to each member of the population (see, for example, El-Gamal and Grether, 1995). However, in some cases one might be unwilling or unable to pre-specify the decision rules, and it turns out that doing so is not necessary. In particular, Houser, Keane and McCabe (2004) detail a Bayesian experimetric procedure that uses individual choice data to determine endogenously the nature and number of decision rules in a population. The approach requires only that one specify the information relevant to individuals' decisions.

Substantive experimetric advances have been obtained in far too many areas to detail here. Although no general survey is available, Houser, Keane and McCabe (2004), Ashley, Ball and Eckel (2005), and Loomes (2005), include excellent summaries of experimetric contributions to a variety of widely-studied games and decision problems.

5 Conclusion

Experimetrics continues to evolve as scholars adopt highly sophisticated design and analysis procedures to inform new questions. A ready example is the rapidly expanding research in neuroeconomics (see neuroeconomics). The massive spatial-panel data structure that characterizes brain images poses unique inferential problems. Progress on these problems requires significant complementary innovations to both design and analysis strategies. The resulting experimetric advances are sure to have significant impact on economic theory and policy analysis.

See Also

- experimental economics
- experimental economics, history of
- experimental methods in economics
- individual learning in games
- neuroeconomics
- quantal response equilibria
- Smith, Vernon

Bibliography

Ashley, R., Ball, S. and Eckel, C. 2005. Motives for giving. A reanalysis of two classic public goods experiments, Manuscript, Virginia Institute of Technology.

Bicchieri, C. and Xiao, E. 2007. Do the right thing: But only if others do. Manuscript, University of Pennsylvania.

Box, J.F. 1980. R.A. Fisher and the design of experiments, 1922–1926. *American Statistician* 34, 1–7.

Feltovich, N. 2000. Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games. *Econometrica* 68, 605–41.

Fisher, R.A. 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* 33, 503–13.

Fisher, R.A. 1935. *Design of Experiments*. Edinburgh: Oliver and Boyd.

Frechette, G. 2005. Session effects in the laboratory. Manuscript, New York University.

El-Gamal, M.A. and Grether, D.M. 1995. Are people Bayesian? Uncovering behavioral strategies. *Journal of the American Statistical Association* 90, 1137–45.

Haile, P.A., Hortacsu, A. and Kosenok, G. 2006. On the empirical content of quantal response equilibrium. Mimeo, Yale University.

Houser, D., Keane, M. and McCabe, K. 2004. Behavior in a dynamic decision problem: an analysis of experimental evidence using a Bayesian type classification algorithm. *Econometrica* 72, 781–822.

Jonckheere, A.R. 1954. A distribution-free k-sample test against ordered alternatives. *Biometrika* 41, 133–45.

Loomes, G. 2005. Modelling the stochastic component of behavior in experiments: some issues for the interpretation of data. *Experimental Economics* 8, 301–23.

McKelvey, R.D. and Palfrey, T.R. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10, 6–38.

Salmon, T.C. 2001. An evaluation of econometric models of adaptive learning. *Econometrica* 69, 1597–628.

Siegel, S. and Castellan, N., Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. Boston: McGraw Hill.

Wilcox, N. 2006. Theories of learning in games and heterogeneity bias. *Econometrica* 74, 1271–92.

How to cite this article

Houser, Daniel E. "experiments and econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000298> doi:10.1057/9780230226203.0533

The New Palgrave Dictionary of Economics Online

explanation

Julian Reiss

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Explaining socio-economic phenomena is one important aim of economics. There is very little agreement, however, on what precisely constitutes an adequate economic explanation. Starting from the very influential but defective ‘deductive-nomological model’ of explanation, this article describes and criticizes the major contemporary competitors for such an account (the probabilistic–causal, the mechanistic–causal and the unificationist models) and argues that none of them can by itself capture all aspects of a good explanation. When seeking to explain a socio-economic phenomenon it should therefore be borne in mind that different types of explanation serve different purposes.

Keywords

causality in economics and econometrics; economic laws; equilibrium; explanation; laws of nature; methodology of economics; positive economics; probability; statistical inference

Article

In the early 1950s Milton Friedman famously declared that the ‘ultimate goal of a positive science is the development of a ‘theory’ or ‘hypothesis’ that yields valid and meaningful (that is, not truistic) predictions about phenomena not yet observed’ (Friedman, 1953, p. 7). Today, after the demise of logical positivism in philosophy and positivistic trends in economics, economists tend to regard the explanation of phenomena as one legitimate aim of economics besides the more directly policy-oriented aims of prediction and control. Perhaps, following Friedman, explaining a phenomenon is primarily of instrumental value for the preparation and guidance of policy. But perhaps economists seek to explain in order to increase our understanding of the economic world, for purely cognitive reasons. Whether derivative or fundamental, explanation is a major goal that economists pursue and understanding what exactly is sought is an important task for economic methodology.
An adequate account of explanation in economics should satisfy at least three desiderata:

1. (a) it should be *descriptively adequate*; that is, it should be consistent with economic practice;
2. (b) it should be *epistemically adequate*; that is, it should give reason to believe that that which it identifies as an explanation is indeed explanatory; and
3. (c) it should be *empirically adequate*; that is, it should not identify something as an explanation unless it is based on sufficient evidence.

The so-called deductive–nomological or DN model of explanation (Hempel and Oppenheim, 1948) can rightly be regarded as the received view of scientific explanation. Although the theory is now generally regarded as untenable, it is it is useful to consider its guiding ideas as a starting point because its flaws motivate the alternative, more satisfactory accounts.

The deductive – nomological model

According to the DN model, an explanation is an argument whose premises constitute the so-called explanans (or ‘that which explains’) and whose conclusion constitutes the so-called explanandum (or ‘that which is to be explained’). The explanandum will usually be a description of a noteworthy singular event (such as ‘Black Monday’, ‘the rise of the dot.com industry’ or ‘the collapse of the Tiger economies’) or a repeated pattern of events, which may be called a ‘phenomenon’ (such as ‘hyperinflations’, ‘the J-curve effect’ or ‘the price drop of cars that have just left the showroom’).
The adjectives ‘deductive’ and ‘nomological’ indicate that the argument must meet at least two criteria in order to count as an explanation. First, the argument must be deductively valid, that is, the explanandum must follow logically from the explanans. Second, among the premises of the explanans there must be at least one law of nature (the Greek word *nómos* means habit or law). Typically, it is also demanded that the premises of the explanans be true or at least verified. However, none of these criteria is individually necessary nor are the criteria jointly sufficient.
In many cases explanations are probabilistic rather than deterministic and thus the explanandum does not always logically follow from the explanans. John Doe's exposure to asbestos explains his contraction of lung cancer but the statement ‘John contracted lung cancer’ is not entailed by the statement ‘John was exposed to asbestos’. Second, and related, laws of nature in the sense of universal regularities are few and far between, especially in non-fundamental sciences such as economics. All so-called ‘laws’ in economics, such as the law of supply and demand, the iron law of wages, Okun's law, Say's Law and so forth are, at best, true *ceteris paribus*, that is, if nothing intervenes and relative to a specific institutional structure. For example, we can use the law of supply and demand to predict that demand for a good will decrease when a tax is imposed. However, depending on what else happens in the economy actual demand may or may not decrease. If disposable incomes rise sufficiently or if preferences change in the right way, demand may in fact increase.
Third, it is not clear whether laws in the sense used by proponents of the DN model are explanatory at all. Suppose that it is a law – a universal regularity – that economic expansions follow monetary expansions. Economists no doubt regard knowledge of this kind as very valuable, but unless more is told about the relationship it would hardly count as explanatory. The DN model is therefore neither descriptively nor epistemically adequate.
In response to these and other difficulties of the DN model (for a valuable discussion of many of the

criticisms, see van Fraassen, 1980, pp. 103–29) philosophers have developed alternative accounts of explanation. One tradition holds that to explain a phenomenon means to cite the causes of the phenomenon. It therefore roughly agrees with the 'nomological' part of the DN model but replaces the notion of law with that of cause. Another tradition holds that to explain a phenomenon means to show how it fits into a systematization of our beliefs about the world. It agrees with the 'deductive' part of the DN model and insists that good explanations are those that unify diverse sets of beliefs. Both traditions can be found in economics, and both come in two variants.

The probabilistic – causal model

The chief difficulty for the causalist, who maintains that to explain a phenomenon is to provide information about its causes, is to elucidate the notion of cause. We believe that a tightening of the money stock explains the subsequent increase in interest rates; a change in minimum wages explains changes in the employment rate; veteran status explains earnings. In none of these cases is there a universal regularity between event-types; rather, earlier events appear to be probabilistic causes in the sense that they are *statistically relevant*.

One view thus held that event X explains event Y if the probability of Y in some population described by Z is different when X is present from when it is absent: $P(Y|X, Z) \neq P(Y|Z)$ (cf. Salmon, 1971). In econometrics this idea is akin to the notion of a multiple regression:

$$Y = \alpha X + \boldsymbol{\beta} \mathbf{Z} + \varepsilon,$$

where Y is the explained variable, X is the explanatory variable and $\mathbf{Z}$ is a vector of background variables. X is statistically relevant to Y if and only α is different from zero, and can thus be used to explain Y.

Not all statistically relevant events appear to be explanatorily relevant, however. A drop in the barometer reading raises the probability of a storm but the barometer reading does not explain the storm. It is a common cause – the change in atmospheric pressure – that explains its joint effects, the barometer reading and the storm.

This suggests that X plus the set of background factors must constitute the full set of causes of the explained variable (cf. Cartwright, 1983, Essay 1): in any population that is homogenous with respect to atmospheric pressure, the barometer reading is statistically irrelevant to the occurrence of the storm. An obvious drawback of this account is that it asks for immense amounts of background knowledge for identifying explanatory factors from statistics. It requires that all other causes of a phenomenon (that is, all confounding factors) to be known or known to be distributed equally between a treatment and a control group, as in a randomized trial. Given the complexities of the social world, one can expect this requirement to be met only exceptionally.

There are also more principled difficulties. One problem arises because some factors may act differently depending on what other causes are present. An increase in the money stock may have different effects on the economy depending on the interest rate and investor behaviour. In extreme situations increasing

money may have no effects on the economy at all, and so the government's ability to conduct monetary policy is thus incapacitated. It therefore seems exaggerated to demand that explanatory factors raise the probability of the explained variable in *all* causally homogeneous populations (which is presupposed by the linear models favoured by econometricians). But is it enough for a factor to raise the probability of the effect in one single population or should it raise its probability on average (cf. Dupré, 1984)?

Moreover, when factors act genuinely probabilistically, factors can be statistically relevant despite the fact that they are not explanatorily relevant, even when all other causes have been included. Suppose that in some causally homogeneous background Z, money M is a probabilistic cause of both nominal income Y as well as the level of prices L. Let $P(Y|M, Z)=P(L|M, Z)=0.8$. Now, let money cause income on precisely those occasions that it causes prices and vice versa. Then, $1=P(Y|L, M, Z)>P(Y|M, Z)=0.8$ even though the change in prices does not explain the change in nominal income – it is a mere correlate (cf. Cartwright, 1999, ch. 5).

But it is important to keep practical and epistemic issues apart. *If* one knows that C is a probabilistic cause of a phenomenon of interest E, then there is no reason to deny that one can use C in an explanation of E. The epistemic adequacy of the probabilistic–causal model derives from the general acceptance of causes as explanatory factors. However, finding out if C is a probabilistic cause of E will often face insurmountable practical difficulties.

The mechanistic – causal model

In philosophy of science, the mechanistic–causal model has been mostly associated with the name Wesley Salmon (see for instance Salmon, 1984). It attempts to improve upon the probabilistic model on two counts. On the one hand, for practical purposes it may be easier to find out whether C causes E by investigating whether or not there is a mechanism running from C to E than by statistical inference. For instance, Milton Friedman and Anna Schwartz (1963, p. 59) write:

> However consistent may be the [statistical] relation between monetary change and economic change, and however strong the evidence for the autonomy of the monetary changes, we shall not be persuaded that the monetary changes are the source [that is, cause] of the economic changes unless we can specify in some detail the mechanism that connects the one with the other.

Indeed, if C causes E we expect there to be a mechanism running from C to E. Evidence about a mechanism from C to E can thus provide evidence for a causal connection. In turn, according to this view, the mechanism can be used to explain E.

On the other hand, causal explanations that are based on statistical inferences often cite relationships among aggregate factors such as the money stock, the unemployment rate, inflation and so on, and can arguably be said to be somewhat shallow. Perhaps a monetary expansion can be used to explain a subsequent economic expansion because there is statistical evidence that the former is the cause of the latter. In this way one learns at best *that* the monetary change causes the economic change. Describing the transmission mechanism one further learns *how* the monetary change causes the economic change. The explanation is thus arguably more detailed, deeper.

It cannot be said, however, that the mechanistic account wins unequivocally over the probabilistic account on both fronts. In order to meet the empirical adequacy desideratum, a mechanistic explanation must be based on evidence no less than a probabilistic-causal explanation. A mere ‘sketch’ of a mechanism (such as the sketch of the transmission mechanism that follows above quotation by Friedman and Schwartz) does not explain anything. Usually, mechanistic explanations cite relationships among individuals, their preferences and external constraints. The argument that such hypotheses are more readily verifiable (for instance, because they may be verifiable by introspection) goes back at least to the writings of John Stuart Mill (1830). But it is not clear whether it is always easier to provide evidence for causal mechanisms that run at the micro level than for aggregate causal relationships. For instance, the problem of confounding factors is in no way confined to statistical inferences among aggregate variables, and, at the micro level, can only seemingly be alleviated by assuming away the operation of confounders a priori.

Moreover, although with some justification it can be said that mechanistic explanations are deeper than aggregate explanations, there are situations in which information about exactly how some variable influences another is entirely irrelevant. A policymaker, for instance, may be more interested in what is common among expansion episodes rather than in the exact processes that made them happen – which may be different on each occasion.

How models explain: unificationism

Let us now move from the applied side to the more theoretical side of economics. Consider the following quotation (Akerlof, 1970):

> From time to time one hears either mention of or surprise at the large price difference between new cars and those which have just left the showroom. The usual lunch table justification for this phenomenon is the pure joy of owning a ‘new’ car. We offer a different explanation.

Akerlof then describes an asymmetric-information model in which low-quality second-hand cars drive higher-quality cars out of the market, which leads to a decrease in average quality and prices. One way to interpret what Akerlof does is to regard the explanatory power of models such as his as consisting in an ability to suggest *schemas* that allow the description of a wide variety of different and seemingly unconnected phenomena. In Akerlof's original article, for instance, the model of the second-hand car market is regarded as a mere ‘finger exercise’ for further application in markets as diverse as insurance, labour, other goods and credit markets. Other economists invoke transaction costs to explain the existence of firms, intergovernmental collaboration, why crime rates are higher among the poor and ‘fair use’ doctrines about the use of copyrighted material among many other phenomena. Many other salient theoretical concepts in economics play a similar unifying role.

Philip Kitcher developed the idea that to explain a phenomenon means to derive a description of the phenomenon from an instance of an argument pattern, instances of which can be used for deriving descriptions of many different kinds of phenomena into a formal account (Kitcher, 1989). Despite its intuitive appeal, however, it is quite clear that unification cannot be all there is to explanation. How

could we tell whether those factors that are salient in a model are the ones that drive the results in the real world? This is a particular problem in economics as many of the concepts that do the alleged explanatory work in models such as Akerlof's are not very discriminatory. There is hardly any market transaction that is not characterized by asymmetric information because it is virtually always the case that one party knows more or something different about a contractually relevant property. Similar observations can be made about other concepts such as human capital or transaction costs or imperfect information. In some situations such factors will be the ones that drive the result, in others they will merely provide a background against which other factors operate. But this is a dominantly qualitative question that should be decided by empirical means, not by means of models alone. Moreover, it seems unlikely that unification is necessary for explanation. Many economic events will be explained with reference to very local and idiosyncratic processes such as wars, innovations and individuals' decisions that lack the power to unify whole classes of events (for further criticism of the unification model, see Woodward, 2003, ch. 8).

Nevertheless, unification plays at least two important roles in economic explanation. First, models such as Akerlof's suggest factors that may be causes of real phenomena. Unifying model schemas thus have an important heuristic role. Second, unifying explanations are in some sense desirable explanations. Even though the causal role a factor plays in bringing about a phenomenon is that which makes a model that describes the operation of this factor explanatory, it is its ability to systematize our beliefs and to reduce the number of 'brute facts' we have to accept as given that makes the explanation attractive to economists.

Equilibrium explanation

Economics is full of equilibrium notions such as the Nash equilibrium, evolutionarily stable equilibrium, sunspot equilibrium, partial and general equilibrium theories. For a variety of reasons, economists tend to downplay explanatory accounts if these accounts do not have a bearing on theory (Heckman, 2000, p. 85):

> Applications of this approach [that aims at the statistically analysis of 'natural experiments'] often run the risk of producing estimates of causal parameters that are difficult to interpret. Like the evidence produced in VAR [vector autoregression] accounting exercises, the evidence produced by this school is difficult to relate to the body of evidence about the basic behavioural elasticities of economics. The lack of a theoretical framework makes it difficult to cumulate findings across studies, or to compare the findings of one study with another. Many applications of this approach produce estimates very similar to biostatistical 'treatment effects' without any clear economic interpretation.

Equilibrium explanations, of course, have exactly this virtue: they show how some phenomenon can be systematized in a theoretical framework. Equilibrium explanations obtain at a level in between aggregate probabilistic explanations and causal-mechanical explanations. Unlike aggregate explanations, they are always formulated in terms of micro entities such as preferences, production possibilities and so forth. But, unlike causal–mechanical explanations, they rarely specify the exact details of how an equilibrium

is reached or how an economy moves from one equilibrium to another. Equilibrium explanations abstract from the causal dynamics and focus on static end-points. Elliott Sober therefore argues (Sober, 1983, p. 204, emphasis in original):

> Equilibrium explanations present *disjunctions* of possible causal scenarios; the actual cause is given by one of the disjuncts, but the explanation doesn't say which.

Because of this, equilibrium explanations are more unifying than explanations that describe the actual causal mechanism that lead to the equilibrium. Nevertheless, equilibrium explanations (at least in economics) tend to cite a lot of information about causes such as preferences, productivity growth, technology and so on. Although abstracting from *some* causal detail, equilibrium explanations can thus safely be regarded as a species of causal explanation.

The greatest challenge for equilibrium explanations is, however, to meet the desideratum of empirical adequacy. In order to derive any results in an equilibrium model, usually a large number of highly distorting idealizations have to be made: consumers maximize utilities and producers their profits; they operate under perfect information; markets clear instantaneously; goods are infinitely divisible and so on and so forth. Furthermore, results derived from a model making such idealizations tend to be very sensitive to specification changes. There is therefore little reason to believe that those forces that drive the equilibrium results obtain also outside the model. Hence, unless it can be shown that this is the case, equilibrium models should be regarded as mere potential explanations.

Conclusion: the variety of causal explanations

The different types of explanation perform different epistemological roles. Very detailed causal–mechanistic explanations can be contrastive: they can provide information about what is special about the way in which a phenomenon came about, the way in which its causal history differs from the causal histories of other, similar phenomena. Aggregate and unifying explanations, by contrast, are comparative: they provide information about what similar or different phenomena have in common (cf. Pettit, 1993, pp. 253–7). For those who are interested in explanation mostly for the practical goals of prediction and policy, aggregate explanations will often be the relevant type. However, mechanistic knowledge can be used to improve predictions, for instance, because it may provide information about the ways in which aggregate relationships sometimes fail to hold. Those with more purely cognitive goals will often prefer explanations that unify.

See Also

- causality in economics and econometrics
- models

Bibliography

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Cartwright, N. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Cartwright, N. 1999. *The Dappled World*. Cambridge: Cambridge University Press.

Dupré, J. 1984. Probabilistic causality emancipated. In *Midwest Studies in Philosophy* 9, ed. P. French, T. Uehling, Jr. and H. Wettstein. Minneapolis: University of Minnesota Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Friedman, M. and Schwartz, A. 1963. Money and business cycles. *Review of Economics and Statistics* 45, 32–64.

Heckman, J. 2000. Causal parameters and policy analysis in economics: a twentieth century perspective. *Quarterly Journal of Economics* 115, 45–97.

Hempel, C. and Oppenheim, P. 1948. Studies in the logic of explanation. *Philosophy of Science* 15, 135–75.

Kitcher, P. 1989. Explanatory unification and the causal structure of the world. In *Scientific Explanation*, ed. P. Kitcher and W. Salmon. Minneapolis: University of Minnesota Press.

Mill, J.S. 1830. *Essays on Some Unsettled Questions of Political Economy*. London: Parker.

Pettit, P. 1993. *The Common Mind*. Oxford: Oxford University Press.

Salmon, W. 1971. Statistical explanation. In *Statistical Explanation and Statistical Relevance*. Pittsburgh: Pittsburgh University Press.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Sober, E. 1983. Equilibrium explanation. *Philosophical Studies* 43, 201–10.

van Fraassen, B. 1980. *The Scientific Image*. Oxford: Clarendon Press.

Woodward, J. 2003. *Making Things Happen*. Oxford: Oxford University Press.

How to cite this article

Reiss, Julian. "explanation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000290> doi:10.1057/9780230226203.0534

The New Palgrave Dictionary of Economics Online

exploitation

Gilbert L. Skillman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This term ‘exploitation’ is used to characterize social relationships in which one party takes advantage of the attributes or position of another. Used normatively, it conveys the stronger sense that the exploiter takes *inappropriate* or *unfair* advantage of another's condition. While the concept has been invoked in a number of economic contexts, it has been treated most extensively in Marxist analysis of class relations in market economies, featuring in particular *orthodox*, *neo-Ricardian*, and *rational-choice Marxist* approaches to the phenomenon. There is ongoing debate concerning both the systemic basis and the normative significance of exploitation.

Keywords

capitalism; class; Engels, F.; equality of opportunity; exploitation; feudalism; fundamental Marxian theorem; labour power vs. labour; labour theory of value; Leontief production processes; Marx's analysis of capitalist production; means of production; neo-Ricardian economics; orthodox Marxist analysis; profit; rational-choice Marxist analysis; Roemer, J.; slavery; Sraffa, P.; surplus; surplus labour; surplus value; transformation problem

Article

As applied to social settings, ‘exploitation’ characterizes relationships in which one party takes advantage of the attributes or situation of another. A complete positive account of a given instance of exploitation would include, in addition to the identities of the parties to the relationship, three elements: the process by which one party exploits the other, the nature of advantage taken, and the conditions that make exploitation possible.
Used normatively, the term conveys the stronger connotation that the exploiter takes *inappropriate* or *unfair* advantage of another's condition. Understood in this latter sense, exploitation theory thus provides an alternative to the Pareto–utilitarian principle as a basis for assessing social outcomes, inasmuch as a

given relationship may be deemed exploitative even if it improves the welfare of the exploited party. The concept of exploitation has been invoked across a range of distinct economic contexts and paradigms. It has been applied, for example, in mainstream economic analysis of monopsonistic wage practices, and in feminist economics to characterize gender relations within households. The term has been treated most extensively, however, in Marxist political economy, which treatment is therefore the focus of the remaining discussion.

Marxian economic theorizing with respect to this phenomenon can, at some risk of oversimplification, be traced through three stages of development corresponding to distinct systems of analysis brought successively to bear on the question: *orthodox Marxist* analysis grounded in Karl Marx's reformulation of the classical labour theory of value; *Sraffian* or *neo-Ricardian* analysis based on the mathematical properties of linear production systems; and *rational-choice Marxist* analysis utilizing formal methods of optimization and equilibrium analysis. The ensuing clash in perspectives both within the Marxian framework and between the latter and the liberal tradition undergirding the mainstream paradigm has generated a lively debate on the positive and normative significance of exploitation.

Marx's account of capitalist exploitation

Karl Marx analysed societies in terms of relations between *classes*, defined in terms of ownership and control of alienable productive assets or *means of production*. In class systems such as feudalism or slavery, Marx held, those who controlled the means of production directly exploited workers by compelling them to expend more labour than that necessary to meet their own consumption needs. Capitalism, in Marx's account, is a specific historical form of class society in which antagonistic class relations are mediated by market transactions. The central analytical problem, in Marx's view, is thus to explain how exploitation of labour might arise in individually voluntary exchanges between traders with formally equal property rights.

Marx's solution to this problem, developed in the first volume of *Capital* (1867), is grounded in the postulate, previously advanced by David Ricardo, that a commodity's *value* is determined by the labour time necessary to produce it under average production conditions. On the basis of this value framework, Marx advanced three propositions: (*a*) capitalist profit is based on the extraction of surplus labour, and thus the exploitation of workers; (*b*) in industrial capitalism, the locus of exploitation is the capitalist-directed production process rather than the marketplace, although exchange relations are the necessary pretext for exploitation to occur; and (*c*) the systemic basis for exploitation under capitalism is that workers are ‘free in the double sense’, that is, both legally able to offer their services in the labour market and ‘free’ of owning any substantial means of production themselves – a condition Marx saw as the outcome of historical processes of expropriation such as the enclosure movement in pre-industrial Great Britain.

Marx argued the first proposition, which in later formulations has been termed the ‘fundamental Marxian theorem’ (FMT), on the heuristic premise that commodity prices are proportional to their respective labour values. The key to Marx's demonstration is his distinction between *labour power*, that is, the capacity for productive effort, sold as a commodity in capitalist labour markets, and *labour*, the exercise of that capacity. If commodities exchange at their respective values, then positive profit is possible if and only if the labour expended by workers in capitalist production exceeds the value of their labour power – that is, workers perform surplus labour for capitalist firm owners.

Marx acknowledged that commodities typically do not exchange at their values, but maintained that prices were nonetheless ‘regulated’ by values, and attempted in the third volume of *Capital* (edited and published in 1894 after Marx's death by his collaborator Frederick Engels) to demonstrate the aggregate correspondence of profit and surplus labour given competitive market valuation of commodities resulting in economy-wide equalization of profit rates. His treatment of this *transformation problem* is suspect, however, due to an evidently inconsistent ‘transformation’ of value magnitudes in his expressions determining market prices.

Marx's second proposition emerged as a corollary of his distinction between labour power and labour. Marx argued that what capitalists purchase in the market is not a specified amount of labour time or bundle of labour services, but rather just individuals’ capacity to work for a given period of time. The use value of this capacity must therefore be realized by extracting surplus labour in the process of capitalist-controlled production. It should be noted, however, that Marx did not insist as a matter of *definition* that capitalist exploitation occurs in the arena of production. In earlier drafts of *Capital* (including that which provided the basis for Engels's edition of the third volume), he frequently alluded to instances of exploitation arising from transactions involving the finance of commodity production without direct capitalist supervision.

With respect to the third proposition, Marx argued in his theory of economic colonization that workers would be unwilling to provide capitalists with surplus value on a regular basis if they were able to produce independently to meet their own needs. Thus, capitalist profit could not in his assessment be systematic unless workers were on the whole divested of substantial ownership in the means of production, such that their livelihoods would be significantly compromised if they chose not to work for wages.

Neo-Ricardian refinements of the FMT

The calculation of commodity labour values, necessary for the determination of exploitation status in an exchange economy, is most straightforward in the case of *Leontief* production processes, characterized by fixed input coefficients and the absence of jointly produced goods, a scenario consistent with Marx's account in the first volume of *Capital* and his treatment of the transformation problem in the third volume. However, even with this basic representation of production conditions, competitively determined commodity prices are typically disproportionate to their values. In light of his problematic ‘transformation’ procedure, Marx's account raises a question concerning the generality of the FMT. Beginning in the 1960s, this question was taken up by a number of economists who, following a mode of inquiry introduced by Sraffa (1960), applied mathematical results characterizing the formal properties of linear systems understood to represent unit input requirements in a capitalist economy based on Leontief production. This literature is informed by an important theorem due independently to Frobenius and Perron (see Kurz and Salvadori, 1995) which states in effect that, if it is possible to produce a surplus net of physical input requirements and real wages, and if all commodities require the direct or indirect input of other commodities, then there exists a competitive price vector supporting a strictly positive rate of profit. On the basis of this result, one can establish the FMT, framed as the formal equivalence of positive profit and the exploitation of labour given any vector of positive competitively determined commodity prices. This equivalence was subsequently extended to scenarios involving joint production and the presence of multiple alternative Leontief techniques for producing each commodity,

subject to a reformulation of labour values, to be discussed below.

The Sraffian approach prompted a number of debates in the Marxian literature. An overriding concern is methodological in nature, as orthodox Marxists have questioned the relevance of ahistorical mathematical models to Marx's analytical methods and insights. There are, moreover, substantive correlates of their methodological concerns. On the one hand, the neo-Ricardian approach is silent with respect to Marx's second proposition, as the formal representation of production conditions central to the new framework simply takes as given the quantitative translation of labour capacity into productive labour expenditures. Similarly, the question of the systemic conditions enabling the extraction of surplus labour is necessarily begged in this analysis. Consequently, the equivalence between profit and capitalist exploitation established in the new versions of the FMT does not lend itself readily to an assessment of the causal connection between surplus labour and profit; the asserted equivalence might, for example, simply be a reflection of some unspecified underlying condition.

On the other hand, the Sraffian approach challenges the explanatory primacy of labour values, in so far as competitive prices are seen to be calculable directly from production and distribution conditions without any intermediate derivation of labour values. Furthermore, generalization to the case of multiple techniques and joint production introduces an element of ambiguity in the standard calculation of individual commodity values because multiple imputations of labour expenditure are then possible for each individual commodity. Morishima (Morishima and Catephores, 1978) addressed the latter problem by redefining labour values in terms of the minimum possible direct labour required to produce a given bundle of net outputs. This procedure preserves the FMT, but at a potential cost of empirical relevance, as it is at best not obvious that actual capitalists would pursue this objective in selecting among alternative techniques.

While thus casting doubt on Marx's hypothesis that individual commodity prices are somehow regulated by their respective labour values, the Sraffian approach does not directly challenge Marx's labour-based definition of exploitation. Although it provides a basis for calculating economic surplus without reference to labour values, and indicates that the operation of a viable competitive economy is consistent with a range of class distributions of that surplus, this framework establishes no independent basis for judging any such distributive outcome to be the consequence of exploitation.

Rational-choice Marxism and the systemic basis of exploitation

Although the orthodox Marxian account of capitalist exploitation is clearly not methodologically individualist in nature, presumably its exponents understand it to be consistent with the interactions of self-seeking individuals responding intelligently to their available options. Granting the latter possibility raises plausible questions as to why, on the one hand, workers allow themselves to be persistently exploited by capitalists, or on the other, why capitalists choose to exercise direct control over production rather than using simple contractual means for extracting surplus labour. The corresponding normative concern asks how exploitation might be considered morally objectionable if understood to occur in voluntary transactions between rational individuals.

These and similar questions motivate a body of inquiry that has come to be termed ‘rational-choice Marxism’, which utilizes mainstream tools of optimization and equilibrium analysis in investigating the systemic basis and features of exploitation in market economies. A central point of reference for this stage of the Marxian literature is Roemer's *General Theory of Exploitation and Class* (1982), which

poses a cogent and fundamental challenge to the canonical Marxian account. Roemer proposes a re-conceptualization of economic exploitation that dispenses entirely with labour value calculations and makes explicit the systemic basis for moral objections to its existence.

Roemer investigates the traditional labour value-based approach to exploitation in the context of both subsistence and accumulating exchange economies, initially maintaining the standard Leontief specification of production conditions and calculation of commodity values in terms of direct and indirect unit labour requirements. In this sense, his analysis commences where the Sraffian approach leaves off, embedding its representation of production conditions in a general equilibrium framework with optimizing agents.

Roemer expands this framework in two ways. First, in the scenario of subsistence exchange economies, markets for produced commodities are alternatively combined with labour and credit markets facilitating the organization of production activity. Second, in the context of accumulating market economies, he allows for unequal endowments of labour capacity and a more general representation of production possibilities.

Roemer's argument is chiefly organized around three analytical results roughly corresponding to Marx's original propositions. First, the *class-exploitation correspondence principle* (CECP) reflects Marx's fundamental assessment of the class basis of exploitation: ‘capitalists’, that is, those who hire or extend credit to workers, exploit, while those who work for wages or borrow to finance production activities are exploited. In the strong version of this theorem, class status is furthermore consistently linked to the market value of alienable endowments, such that the sufficiently wealthy become exploiting capitalists and the relatively indigent become exploited workers in market equilibrium.

The CECP deepens and extends the sense of the FMT by linking exploitation status to class position. Roemer demonstrates that the CECP obtains for both subsistence and accumulating economies given Leontief technology and identical preferences and labour endowments, subject only to modifications in the labour-based index of exploitation status to accommodate the distinction between subsistence and surplus economies.

Second, Roemer argues that neither the degree nor the structure of equilibrium exploitation depend in general on whether production activities are supported by labour or credit markets. At first glance, this *isomorphism theorem* appears to rebut Marx's identification of production as the primary locus of exploitation. However, as in the Sraffian approach, Roemer's framework abstracts from the problem of translating labour capacity into labour performed. Thus the implied rebuttal of Marx's second proposition is contingent: on the assumption that capitalists face no significant obstacles in contracting for given labour services or loan repayments with interest, exchange rather than production relations provide the essential framework for capitalist exploitation. This qualified refutation nonetheless carries theoretical force, in so far as Marx's value-theoretic account of capitalist exploitation in the first volume of *Capital* does not explicitly feature conditions that would in contemporary terms be described as contracting failures.

Third, Roemer's general equilibrium framework affords a characterization of the systemic conditions for exploitation to occur in either subsistence or accumulating exchange economies. Roemer identifies *differential ownership of scarce productive assets* (DOSPA) as the necessary and (absent significant contracting failures) sufficient basis for the existence of exploitation. Differential ownership refers to inequality in the market value of individuals’ endowments of productive assets, specifically (given the assumption of identical labour capacities) their endowments of *alienable* productive assets, that is,

‘capital’. Significantly, the sense of *capital scarcity* in Roemer's framework is that of *absolute* scarcity, such that the market demand for capital assets (either direct or imputed) is positive at the level of total available capital supply. Otherwise, since Roemer's actors don't discount future payoffs or demand risk premia, capital goods would not command a positive rate of return.

It is readily seen why wealth inequality and capital scarcity are jointly necessary for exploitation to arise. If productive assets are equally distributed, then even if capital is scarce no agent wishes to supply it because of the resulting sacrifice in increased labour expenditure. If capital is not scarce in the absolute sense, then no agent demands it even if others enjoy greater wealth. This result presumes that individuals are identical except in their endowments of alienable assets.

Realistic generalizations of production and endowment conditions, Roemer argues, undermine the sense and scope of Marx's labour-based theory of exploitation. First, representing production possibilities as a convex cone, which maintains the assumption of constant returns to scale but allows for such practical features as factor substitution, fixed capital and joint production, introduces a tension between the labour-based conceptions of commodity value and capitalist exploitation: even if Morishima's revised valuation procedure were adopted to accommodate these more inclusive production conditions, the CECP no longer obtains in general.

To address this problem, Roemer suggests a plausible reformulation of labour value in which capitalists, rather than adopting the labour-minimizing technique for given net outputs, are instead understood to choose techniques which minimize costs at going market prices. As he demonstrates, this modification preserves the CECP, but at the necessary cost of rendering commodity values strictly dependent on commodity prices, thus reversing the causal connection asserted by Marx.

The CECP is more fundamentally compromised by quantitative and qualitative variations in individuals’ inalienable endowments of labour capacity. Merely quantitative variations in labour endowments, corresponding for example to the case that some workers are able to work harder or faster than others, weakly preserve the CECP, but eliminate the monotonic relationship between class status and alienable wealth. Thus, from the standpoint of Marx's canonical account, perversities can arise such as rich but exploited wage workers or poor but exploiting capital suppliers. (A similar difficulty arises, it might be noted, given additional non-produced factors such as ‘land,’ which further disrupt the connection between embodied labour times and individual wealth.)

Heterogeneous labour endowments, reflecting differential abilities across distinct production tasks, raise an even more fundamental problem with respect to defining embodied labour magnitudes, as qualitatively distinct labour expenditures must somehow be aggregated before individuals’ exploitation status can be assessed. But, even if a consistent procedure for aggregating heterogeneous labour inputs were identified, such as weighting them with their respective wage rates, no meaningful and consistent relationship among wealth, class position and exploitation status can thereby be established in general. Thus, the labour value-based conception of exploitation appears to break down entirely in this realistic scenario.

The normative significance of exploitation

These anomalies prompt Roemer to advance a new conception of exploitation that assesses economic outcomes through comparison with some alternative property rights regime reflecting a given distributional norm. Besides sidestepping the inconsistencies arising in the labour-based approach to

exploitation, this formulation offers the advantage of making explicit the grounds for judging given relationships as exploitative. As such it has potential applications beyond the scope of Marxian analysis.
Roemer's approach can be represented in terms of a cooperative game with transferable payoffs. Let N denote a set of economic actors with representative subset C, called a *coalition*. Let (C_i, C_j) designate disjoint subsets of N representing distinct coalitions in the larger society. (In Roemer's account, these coalitions are defined more specifically as complements, so that $C_j = N - C_i$.) Suppose that coalitional payoffs in this game are determined by the function $u(C)$, and define a withdrawal rule $\pi(C)$ for the game, interpreted as the payoff to coalition C should it choose to reject the allocation $u(C)$ and freely withdraw with the resources permitted it by the game's rules.
Then, given the payoff allocation rule $u(C)$, coalition C_i is said to be exploited by coalition C_j relative to withdrawal rule $\pi(C)$ if (*i*) $u(C_i) < \pi(C_i)$ (C_i would be made better off by withdrawing), (*ii*) $u(C_j) > \pi(C_j)$ (C_j would be made worse off by withdrawing), and (*iii*) C_j *dominates* C_i. Roemer doesn't specify the meaning of ‘domination’ in the *General Theory*, and offers alternative formulations in subsequent work without settling on a definitive statement of the condition. A plausible interpretation is that the payoffs $(u(C_i), u(C_j))$ are somehow imposed by C_j in lieu of a feasible alternative payoff rule $\tilde{u}(C)$ such that $\tilde{u}(C_i) \geq \pi(C_i)$ (coalition C_i isn't made better off by withdrawing, given $\tilde{u}$) and $u(C_j) > \tilde{u}(C_j)$ (coalition C_j benefits by choosing u over $\tilde{u}$).
An allocation $u(C)$ is then called ‘non-exploitative’ (NE) relative to the game characterized by $\pi(C)$ if no coalition $C \subset N$ is exploited. NE allocations are thus related to the well-known concept of the *core* of a game characterized by $\pi(C)$. All allocations in the core are NE, but not necessarily vice versa, since an allocation may be NE yet Pareto-inefficient.
The normative significance of this formulation clearly depends in part on the specification and justification of the withdrawal rule $\pi(C)$. Roemer posits three forms or levels of exploitation corresponding to different types of economic organization. *Feudal* exploitation is defined relative to a rule allowing individuals to withdraw with the free use of their existing endowments, the interpretation being that this form of exploitation involves coercive or strategic strictures on others’ utilization of their property. *Capitalist* exploitation is said to exist in the context of a rule allowing actors to withdraw with the per capita share of the economy's alienable productive assets, reflecting the assessment that wealth inequality is a prerequisite for Marxian exploitation in private ownership economies. Finally, *socialist* exploitation is defined relative to a rule allowing individuals to withdraw with the per capita share of both *alienable* and *inalienable* productive assets, including skills and innate abilities.
In subsequent work (for example, Roemer, 1985), Roemer has argued that moral objections to exploitation are most appropriately understood as normative demands for the provision of equal economic opportunities to all. This is a plausible inference, and one that may in any case be compelling to those who characterize capitalist economic relations as exploitative. However, in light of the dominance condition in Roemer's three-part definition, one might with equal justification condemn the *relational* aspect of exploitation, specifically the acts by which exploiters take advantage of the vulnerable position of others, whatever the material conditions that created this asymmetry.
This point can be illustrated with reference to two strands in the rational-choice Marxist literature subsequent to Roemer's *General Theory*. One involves the criticism that Roemer's use of the competitive general equilibrium framework abstracts from contracting failures and thus ignores crucial aspects of the

process by which workers are exploited in capitalist economies. In their influential treatment of this issue, Bowles and Gintis (1990) argue that the strategic response of capital suppliers to contracting failures in labour markets creates a situation of *contested exchange* in which a form of unilateral economic power is exercised over workers in capital-owned firms. They argue for the institution of democratically organized firms as a safeguard against the exercise of such power, without qualifications as to the distributional basis for this form of exchange.

A second line of criticism concerns Roemer's use of a static analytical framework in treating the essentially dynamic phenomenon of capital accumulation. This point is developed by Veneziani (2006), who argues that the condition of absolute capital scarcity, necessary for the existence of exploitation in Roemer's models, becomes extremely fragile once embedded in a genuinely intertemporal equilibrium context. On this basis, Veneziani suggests that the competitive model premised on perfect contracting arrangements is not a suitable vehicle for illuminating the Marxian theory of exploitation.

The validity of this claim, and more generally the systemic basis for capital scarcity and profit, remain open theoretical and empirical questions that stand at the boundary distinguishing Marxian and mainstream economic perspectives. In any case, there are distinct senses in which enduring capital scarcity might be deemed consistent with the manifestation of exploitative economic relationships. One possibility is that capital scarcity is somehow preserved by unequal material conditions. If, for example, time or risk preferences were income elastic, then persistent capital scarcity might arise from inequalities of the sort targeted in Roemer's account. However, even if such preferences were innate and generally uncorrelated with wealth, consistent with the standard neoclassical account, exploitation might still be said to arise if capital suppliers used inappropriate means (such as those criticized by Bowles and Gintis) to maximize the advantage derived from their unique position.

See Also

- equality of opportunity
- Marx's analysis of capitalist production
- neo-Ricardian economics

Bibliography

Bowles, S. and Gintis, H. 1990. Contested exchange: new microfoundations for the political economy of capitalism. *Politics and Society* 18, 165–222.

Goodin, R. 1987. Exploiting a situation and exploiting a person. In *Modern Theories of Exploitation*, ed. A. Reeve. London: Sage Publications.

Kurz, H. and Salvadori, N. 1995. *Theory of Production: A Long-Period Analysis*. Cambridge: Cambridge University Press.

Marx, K. 1867. *Capital: A Critique of Political Economy*, vol. 1. London: Penguin Books, 1990.

Marx, K. 1894. *Capital: A Critique of Political Economy*, vol. 3, ed. F. Engels. London: Penguin Books, 1991.

Morishima, M. and Catephores, G. 1978. *Value, Exploitation and Growth*. Maidenhead: McGraw-Hill.

Nielsen, K. and Ware, R. 1997. *Exploitation*. Atlantic Highlands, NJ: Humanities Press.

Roemer, J. 1982. *A General Theory of Exploitation and Class*. Cambridge, MA: Harvard University Press.

Roemer, J. 1985. Should Marxists be interested in exploitation? *Philosophy and Public Affairs* 14, 30–65.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Veneziani, R. 2006. Exploitation and time. *Journal of Economic Theory* (forthcoming).

How to cite this article

Skillman, Gilbert L. "exploitation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000190> doi:10.1057/9780230226203.0535

The New Palgrave Dictionary of Economics Online

external economies

Peter Bohm
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

External economies and diseconomies nowadays generally mean unpaid side effects of one producer's output or inputs on other producers. External economies in this sense imply as a rule that market prices in a competitive market economy will not reflect marginal social costs of production, giving rise to a ‘market failure’. But, with technological external economies and diseconomies now most often replaced by the well-defined concept of externality, and with pecuniary external economies and diseconomies being synonymous with general market interdependence, external economies no longer have much of a role to play in economic analysis.

Keywords

balanced growth; division of labour; external economies; externalities; investment criteria; Knight, F. H.; linkages; market failure; Marshall, A.; pecuniary external economies; Pigou, A. C.; Robertson, D. H; Scitovsky, T.; technological external economies; Young, A. A.

Article

‘The concept of external economies is one of the most elusive in economic literature.’ This is how Tibor Scitovsky began his article ‘Two Concepts of External Economies’ (1954). His statement is still true, and it may be added that there are at least two such concepts.

The meaning of external economies and its counterpart, external diseconomies, has changed over time. Nowadays, it is essentially synonymous with externality or external effects in the sphere of production. That is, external economies (diseconomies) or positive (negative) external effects in production are unpaid side effects of one producer's output or inputs on other producers. (As an illustrative example, we can take the case where a dam constructed by a hydroelectric power plant eliminates flooding of farmers’ crop fields – external economy – or reduces the catches of fishermen downstream – external diseconomies; a producer's pollution which increases the costs of, *inter alia*, other producers is perhaps the most important case of externalities.) Sometimes, external economies also refer to unpaid side effects of or on consumption activities, but this meaning is disregarded here.

External economies in this modern sense imply as a rule that market prices in a competitive market economy will not reflect marginal social costs of production. Hence, a ‘market failure’ arises, meaning that the market economy cannot attain a state of efficiency on its own. Specifically, in an otherwise ‘perfect’ market economy, a producer who has external economies (positive external effects) on other producers would not extend his externality generating activity, say, his output, to the point where marginal cost of production equals marginal social benefits of production, which amounts to the market value of his marginal output plus the market value of the side effect on the output of other producers.

At an earlier stage, external economies in the meaning now given were called *technological external economies*, reflecting the fact that the effects were transmitted outside the market mechanism and altered the technological relationship between the recipient firm's output and the inputs under its control. Formally, we have that the output q_i, of the ith producer is affected not only by changes in his control variables, a vector x_i, but also by e_j, a variable controlled by some other producer j. This gives us the following production function:

$$q_i = f_i(x_i; e_j).$$

The reason for specifying such effects as technological is that the concept of ‘external economies’ has been given a broad meaning ever since its introduction. During the early part of the 20th century, external economies (diseconomies) were defined so as to include beneficial (detrimental) *price effects* of producer activities. Thus, in principle, the concept included cases where increases in factor inputs by one firm lowered or raised input prices for other firms. However, much of the discussion centred on the case where increases in *industry* output lowered or raised input prices for the individual member firm. The case of reduced input prices presupposes that the supply side of the market for inputs is characterized either by imperfect competition (say, a profit-maximizing monopoly producing at decreasing marginal costs, decreasing at a rate sufficient for an increase in demand to lower price) or by a competitive industry having a downward-sloping ‘supply’ curve, which in turn reflects external economies in this industry. Supply conditions in the original industry, as well as in the industry producing inputs for the first industry, are shown in Figure 1. Here, $\sum s_i(Q_0)$ is the aggregate supply of the firms in the industry when actual industry output is Q_0. When industry output increases to Q_1, input prices drop (or technological external economies arise), causing a downward shift in individual cost and supply curves and hence in the aggregate supply ($\sum s_i(Q_1)$). The curve *M*, the downward-sloping ‘supply’ curve, is actually a market equilibrium curve showing the equilibrium price/output combinations at different levels of demand (see Bohm, 1967).

Figure 1

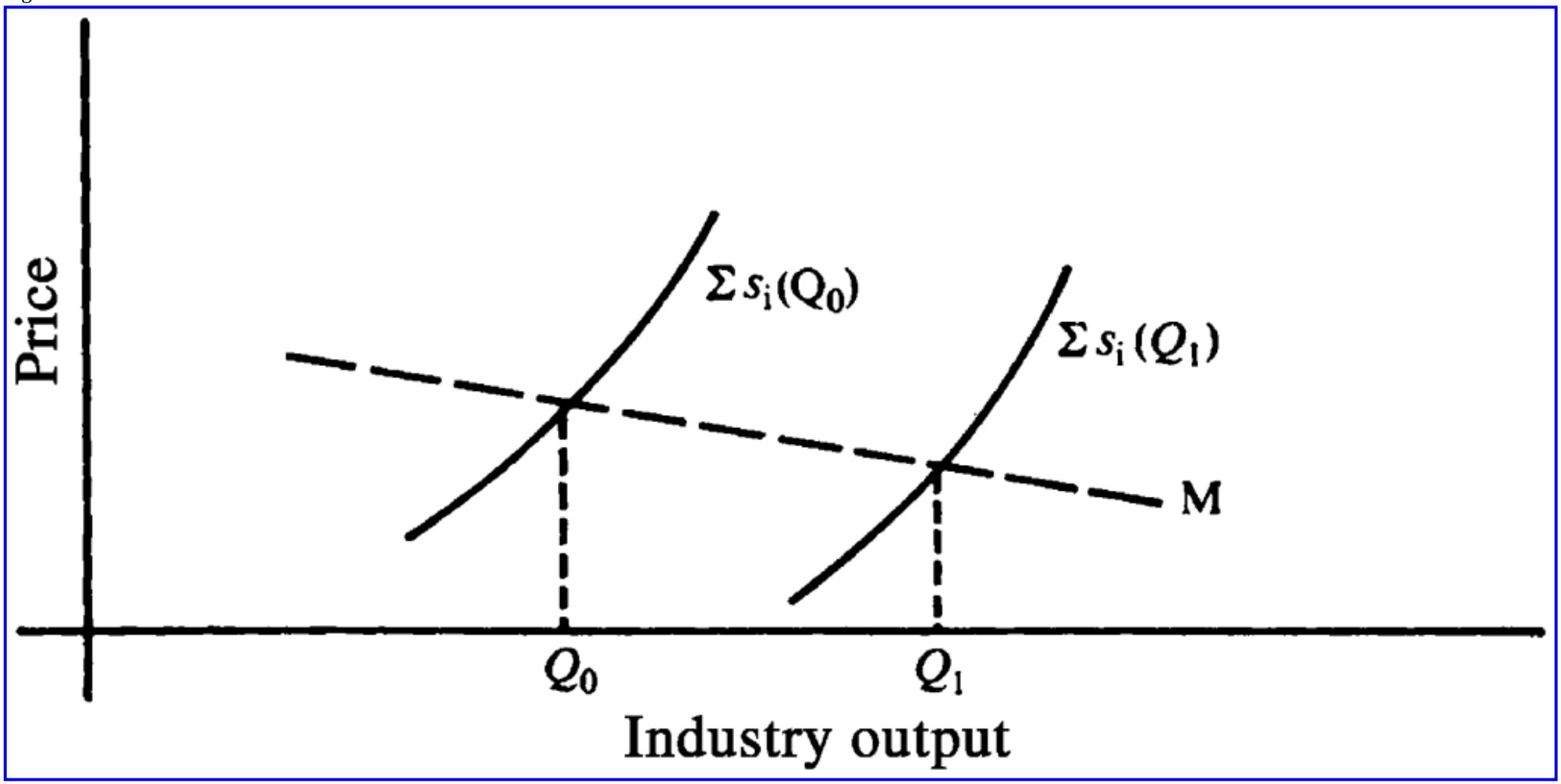

The price effects between firms or between industry and its firms were termed *pecuniary external economies (diseconomies)* by Viner (1931). Before him, A.C. Pigou (1920) had argued (although he phrased it differently) that external economies and diseconomies, both technological and pecuniary, would call for government intervention in order for the industry to attain a socially efficient level of output. Specifically, Pigou argued that, if expansion of a competitive industry would increase prices of inputs sold to the industry, thus

creating pecuniary external diseconomies for the individual firms in the industry, the aggregate ‘supply’ (or market equilibrium) curve would not reflect social marginal costs. Pigou's argument can be illustrated in Figure 2, where *M* is the upward-sloping long-run ‘supply’ curve for the industry due to rising input prices as a consequence of increasing industry output. *SMC* is the curve showing the total marginal outlay on inputs, where the difference to *M* is the increased outlay for *infra*-marginal inputs. Pigou contended (but later rescinded this position) that the *SMC* curve indicated the true social marginal costs and hence, that the price/output combination attained by the market, as shown by the intersection of market ‘supply’ *M* and market demand *D*, was suboptimal. He argued that the optimal level could not be attained unless a tax were levied on this industry so that, in equilibrium, price and output would be those shown by the intersection of *SMC* and *D*. (Similarly, a bounty would be required in the case of pecuniary external economies (see Figure 1), where *SMC* would be downward-sloping and steeper than the *M* curve.)

Figure 2

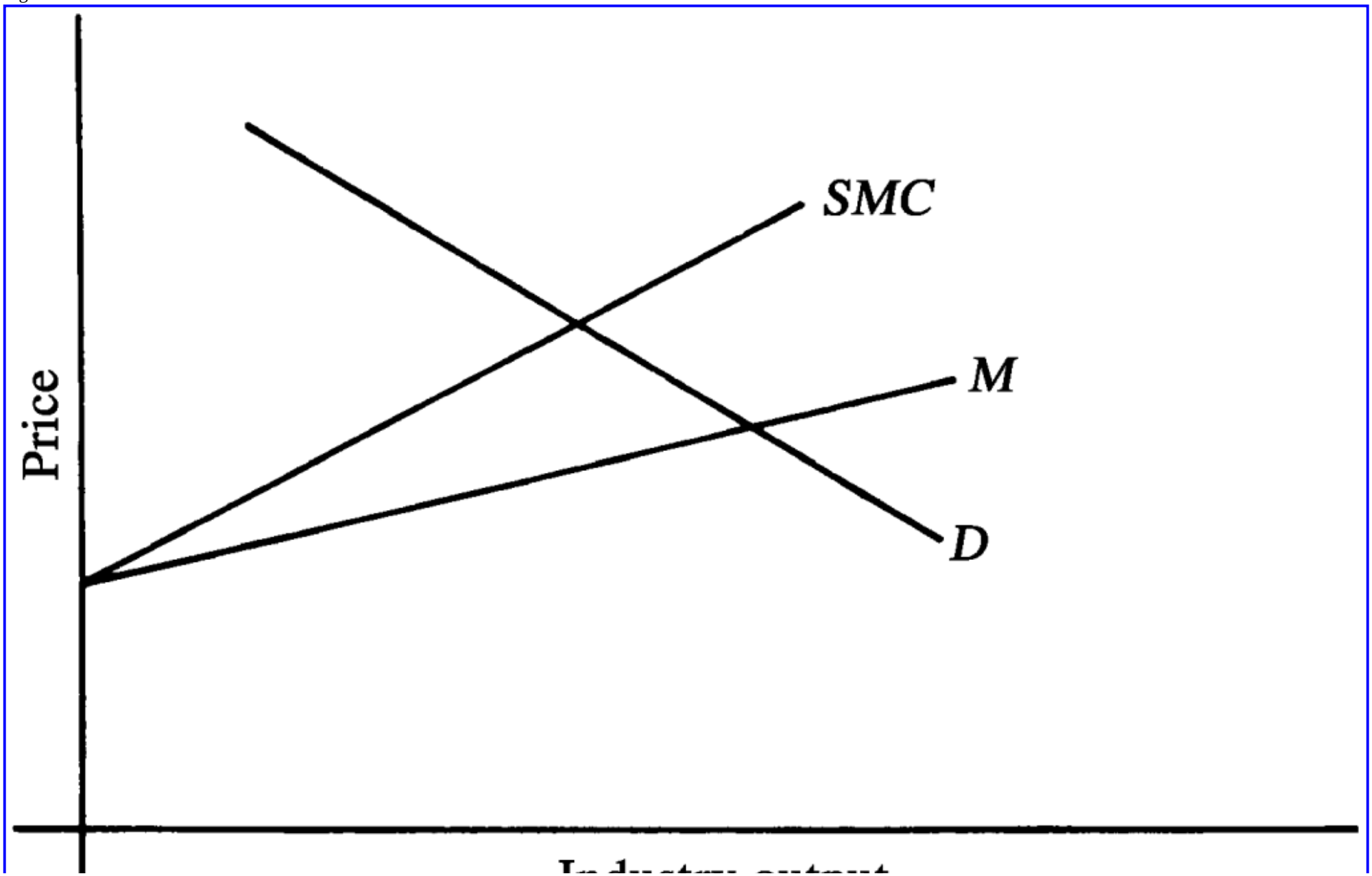

It was demonstrated by F.H. Knight (1924) and D.H. Robertson (1924) and further elaborated by Ellis and Fellner (1943) that the total marginal outlay is irrelevant as an indicator of marginal social costs. The effect on outlay for intramarginal units of inputs represents an increment to rent on these units and hence, it is not part of the real costs of increased output; the real marginal costs are only those which have to be paid for the required marginal inputs. In other words, the rising costs for marginal inputs are those shown by the industry ‘supply’ curve *M*. Hence, pecuniary external economies or diseconomies would not call for government intervention. The case is different, however, for technological economies and diseconomies. Assume, for example, that output of a geographically concentrated industry pollutes the area where it is located and hence reduces the productivity of labour inputs of the individual firms in the industry. Here, all effects on input productivity, for marginal as well as intramarginal units, constitute real costs. So if the rising *M* and *SMC* reflect such effects only, Pigou's argument holds.

When the debate had arrived at this point, pecuniary external economies could be dropped as a cause of market failure and, hence, the concept lost its specific economic interest. But, now, what do the *technological* external economies and diseconomies of industry output on the individual firms in the industry actually represent? Alfred Marshall introduced the term external economies when analysing industry production costs as a function of output:

> We may divide the economies arising from an increase in the scale of production of any kind of goods, into two classes – firstly, those dependent on the general development of the industry; and secondly, those dependent on the resources of individual houses of business engaged in it, on their organization and the efficiency of their management. We may call the former *external economies*, and the latter *internal economies*. (Marshall, 1920, p. 266)

The latter concept is now recognized as economies of scale in the individual firm. Marshall elaborated the meaning of the former concept using scattered examples, the most explicit of which is perhaps the increased knowledge accompanying the expansion of industry output materialized in the publication of trade journals and other forms of improved information about markets and technology in the industry. But, in addition, he argued that industry growth, especially when concentrated to a particular region, might create a market for skilled labour, advance subsidiary industries or give rise to specialized service industries as well as improve railway communication and other infrastructure.

Thus, external economies emerged here essentially as cost reductions for individual firms as a consequence of industry growth, that is, as economies external to the firm but internal to the industry. To remain firmly within the framework of static analysis, these economies should be thought of as being reversible. But some of Marshall's examples alluded to irreversible phenomena and to dynamic effects of industry growth. This was particularly obvious when he at times referred to external economies as being dependent on the ‘general progress of industrial environment’.

Marshall's claim that external economies were important in the long run – in fact, more important than internal economies – seemed to have little immediate impact on the thinking of his fellow economists. For example, the discussion of ‘empty economic boxes’ in the early 1920s (Clapham, 1922; Pigou, 1922; Robertson, 1924) centred on questioning the relevance of external economies and diseconomies, the latter concept deriving from issues raised by Pigou (1920). Actually, the significance of technological external economies and diseconomies in the *static* analysis of an industry or among producers in general escaped most economists all the way up to the post-war years when traffic congestion (although already mentioned by Pigou) and the common pool problem of interdependent producers in an oil field or a fishing area became the leading examples of the industry case and environmental pollution the leading example of the general case. Eventually, diseconomies emerged as the important case and economies as the exceptional case, whereas earlier hardly any importance was attached to technological *dis*economies (see, for example, Robertson, 1924).

If static external effects turned out to be the most important legacy of Marshall's original contribution, external economies as a *dynamic* concept, which seems closer to Marshall's own ideas, also came to play a role in economic analysis. Dynamic external economies refer to increased division of labour resulting from industry growth, the emergence of firms specializing in new activities, some of which aim at developing capital equipment for, or servicing, other firms. An early elaboration of these ideas was made by Young (1928). Later, the concept of external economies came to play a prominent role in development planning, primarily for underdeveloped countries or regions. The general idea here was hardly at all related to the non-market interdependence of technological external economies, but rather to the market interdependence of which pecuniary external economies were part, now on an economy-wide basis. It was argued, in particular by Rosenstein-Rodan (1943) and – with a somewhat different focus – by Scitovsky (1954), that development must be planned so that supply and demand relationships among different sectors of the economy are taken into account. Specifically, it was pointed out that major investments in industrial capacity in a poor economy risk being a failure unless a required increase in the supply of inputs to meet the need of the expanding industry, as well as a sufficient stimulus of demand for the output of the expanding industry, occur at the same time. This doctrine of ‘balanced growth’ called for a simultaneous expansion of investment in several sectors of the economy, a ‘Big Push’ (Rosenstein-Rodan, 1943) determined by input–output relations or general market interdependence. Scitovsky argued specifically that indivisibilities in capital formation may call for investment criteria, which – in contrast to those of individual firms – take into account the indirect supply and demand effects on profitability elsewhere in the economy.

Here, external economies came to be synonymous with what was later called linkage effects (Hirschman, 1958), where backward linkages refer to the supply to the investing sector and forward linkages to the demand for the output of the investing sector. Hirschman's own analysis led him to recommend a strategy for economic development in poor countries that was opposite to that of a ‘balanced growth’. Focusing on the shortage of entrepreneurial capacity and the small impact of subtle market signals in backward areas, he advocated a

strategy of ‘unbalanced growth’, according to which investments should be undertaken so as to reinforce market signals and create pressure for investments in sectors related to the original investments via strong linkage effects.

That linkage effects offered a more precise terminology than dynamic external economies is one reason why the latter concept now has lost ground. Another is, of course, that this meaning of external economies referred only or primarily to market interdependence, which is a general economic phenomenon. Thus, with technological external economies and diseconomies now most often replaced by the well-defined concept of external effects, and with pecuniary external economies and diseconomies being synonymous with general market interdependence, external economies no longer have much of a role to play in economic analysis. Aside from occasional use as a synonym for external effects, the concept now stands for interdependence that does not clearly fall into any of the categories mentioned here. That is, when firms affect one another in a way not covered by static equilibrium analysis or by interdependence among existing markets in the context of the dynamic analysis of economic development, external economies are still used by economists as a convenient catchall.

Again, one may ask – as did economists in the interwar period – what do these external economies actually stand for? At least some examples can be given. Growth of an industry may create a supply of new skills which turn out to provide a starting point for an altogether new line of business. Or, growth of a technologically advanced industry in a particular region leading to the location of a school of higher learning to this region may in turn stimulate – or reduce – growth of other activities. These cases are awkward to handle in traditional, well-structured economic analysis. So the main characteristic of these external economies, very much like most of those suggested by Marshall, is that we cannot yet say in any systematic way exactly what they represent.

See Also

- externalities
- linkages
- Young, Allyn Abbott

Bibliography

American Economic Association (AEA). 1952. *Readings in Price Theory*, vol. 6. Homewood, IL: Irwin.

Arndt, H.W. 1955. External economies in economic growth. *Economic Record* 31(November), 192–214.

Arrow, K. and Scitovsky, T., eds. 1969. *Readings in Welfare Economics*. Homewood, IL: Irwin.

Bohm, P. 1967. *External Economies in Production*. Stockholm: Almqvist & Wiksell.

Clapham, J.H. 1922. On empty economic boxes. *Economic Journal* 32, 305–14. Reprinted in AEA (1952).

Ellis, H. and Fellner, W. 1943. External economies and diseconomies. *American Economic Review* 33, 493–511. Reprinted in AEA (1952).

Heller, W.P. and Starrett, D.A. 1976. On the nature of externalities. In *Theory and Measurement of Economic Externalities*, ed. S. Lin. New York: Academic Press.

Hirschman, A.O. 1958. *The Strategy of Economic Development*. New Haven: Yale University Press.

Knight, F.H. 1924. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38, 582–606. Reprinted in Arrow and Scitovsky (1969).

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Meade, J. 1952. External economies and diseconomies in a competitive situation. *Economic Journal* 62, 54–67. Reprinted in Arrow and Scitovsky (1969).

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Pigou, A.C. 1922. Empty economic boxes. *Economic Journal* 32, 458–65. Reprinted in AEA (1952).

Robertson, D.H. 1924. Those empty boxes. *Economic Journal* 34, 16–30. Reprinted in AEA (1952).

Rosenstein-Rodan, P.N. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53, 202–11.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62, 70–82. Reprinted in Arrow and Scitovsky (1969).

Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3, 23–46. Reprinted in AEA (1952).

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42. Reprinted in Arrow and Scitovsky (1969).

How to cite this article

Bohm, Peter. "external economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000199> doi:10.1057/9780230226203.0536

The New Palgrave Dictionary of Economics Online

externalities

J.J. Laffont

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Externalities are *indirect* effects of consumption or production activity, that is, effects on agents other than the originator of such activity which do not work through the price system. In a private competitive economy, equilibria will not be in general Pareto optimal since they will reflect only *private* (direct) effects and not *social* (direct plus indirect) effects of economic activity. This article explains how this outcome arises and considers the policy responses that have been advanced to remedy the market failures stemming from externalities.

Keywords

asymmetric information; coalitions; competitive equilibrium; cooperative game theory; environmental economics; externalities; imperfect information; lump-sum transfers; non-convexity; pecuniary externalities; pollution rights; strategic behaviour; taxation of externalities; technological externalities

Article

Competitive equilibria are Pareto optimal when they exist if preferences are locally non-satiated and if externalities are not present in the economy. Why externalities upset the first fundamental theorem of welfare economics and which economic policies can remedy this failure are the major questions addressed below.

Technological externalities

Let us call technological externality the indirect effect of a consumption activity or a production activity on the consumption set of a consumer, the utility function of a consumer or the production function of a producer. By ‘indirect’ we mean that the effect concerns an agent other than the one exerting this economic activity and that this effect does not work through the price system.

Externalities may be positive or negative and are quite diverse. Major examples include pollution activities (air pollution, water pollution, noise pollution …), malevolence and benevolence, positive interaction of production activities. From a practical point of view the most significant are negative pollution activities, so that we can say that the theory of technological externalities is essentially the foundation of environmental economics.

The formalization of technological externalities is achieved in microeconomics by making production sets, utility functions and production sets (or functions) affected by externalities functionally dependent on the activities of the other agents creating these indirect effects.

For example, the utility function of a consumer is made dependent on the level of production of a firm polluting the air breathed by the consumer. This modelling option that we will implicitly adopt here is right as long as the link between production and air pollution is not alterable.

If de-polluting activities are possible the link between the level of pollution and the economic activities generating them must be made explicit. An important difficulty in analysing these activities is due to the non-convexities which they usually introduce.

Pecuniary externalities

During the 1930s, a confused debate occurred between economists on the relevance of pecuniary externalities, that is, on externalities which work through the price system. A quite general consensus was that pecuniary externalities are irrelevant for welfare economics: the fact that by increasing my consumption of whisky I affect your welfare through the consequent increase in price does not jeopardize the Pareto optimality of competitive equilibria.

This is true when all the assumptions required for the competitive equilibria to be Pareto optimal are satisfied. In such a framework prices only equate supply and demand and pecuniary externalities do not matter. As soon as we move away from this set of assumptions prices generally play additional roles. For example, in economies with incomplete contingent markets, prices span the subspace in which consumption plans can be chosen. In economies with asymmetric information, prices transmit information. When agents affect prices, they affect the welfare of the other agents by altering their feasible consumption sets or their information structures. Pecuniary externalities matter for welfare economics.

In what follows we focus only on technological externalities.

Competitive equilibrium with externalities

How is the characterization of Pareto optima in convex economies affected by externalities? Very simply, as Pigou early understood. The classical equality of marginal rates of substitution and marginal rates of transformation must now be expressed using *social* marginal rates and not only *private* marginal rates as in an economy without externalities. Social marginal rates must be computed taking into account direct *and* indirect effects of economic activities. For example, the marginal cost of a polluting activity must include not only the direct marginal cost of production, but also the marginal cost imposed on the environment.

Note that Pareto optima do not exclude polluting activities, but set them at levels such that their social marginal benefit equates their social marginal cost well computed.

It is now easy to understand that in a private competitive economy, equilibria will not be in general Pareto optimal since the private decentralized optimizations of economic agents lead them to the equalization of *private and not social* marginal rates through the price system.

Markets for pollution rights

Consider for concreteness a firm polluting a consumer. One potential solution is to create a market for this externality. Before producing, the firm must buy from the consumer the right to pollute. If both actors were behaving competitively with respect to the price of this right, the competitive equilibrium in the economy with an extended price system would be Pareto optimal, since there is no externality left.
A number of difficulties exist with this approach. In general we cannot expect agents to behave competitively unless we are in the special case of impersonal externalities. Then, there is a fundamental non-convexity in the case of negative externalities since as a negative externality increases the production set shrinks, but there is a limit to this effect which is the zero production level. Competitive equilibria cannot then exist unless bounds are set on supplies of pollution rights. (For a positive price, a firm would like to offer an infinite amount of pollution rights and close down.)
In the above set-up, the implicit status quo was the absence of externalities. The initial rights are a clean environment: ‘Polluters must pay.’ We can instead give to the polluting firm the right to pollute and then ask the consumer to buy from the firm a decrease of his pollution. This different allocation of initial rights does not upset the Pareto optimality of the competitive equilibrium, but of course has distributional effects.

Taxation of externalities

The likely strategic behaviour by agents on markets of pollution rights makes taxation of externalities the most common policy tool. The polluter must then pay for each unit of a polluting activity a tax which equals the marginal cost imposed by this activity on the other agents. The polluter then internalizes the externality and Pareto optimality is restored. If the externality is positive he must be similarly subsidized.
Note that nothing is said about the amount of taxes so obtained by the government. There is no presumption that it is given to the polluters. In fact, the implicit assumption is that it is redistributed through lump-sum transfers which do not affect agents’ behaviours (in the sense of their first-order conditions). From the point of view of Pareto optimality, the important goal is to modify polluters’ behaviours.
If lump-sum transfers are not available, the budget of the government must be balanced and then goods different from the polluting activities must be taxed or subsidized to solve the ensuing second-best problem.
The major difficulty with this solution is informational.

Imperfect information

The traditional theory of externalities has proceeded as if the regulators had complete knowledge of the

economy and were therefore able to compute optimal taxes, or as if agents were not behaving strategically with respect to their private information. Very often this is not the case and the problem is to elicit this private information and use it to compute taxes, a more difficult problem.
Intuitively, the solution of what is now a second-best problem is to have taxes which depend nonlinearly on polluting activities. This nonlinearity may sometimes take the extreme form of a zero tax up to a given amount and a very large tax above, a mechanism which is equivalent to a quota.

Planning and externalities

Externalities are not only a problem of market economies with an insufficient number of markets. One way to suppress an externality between two agents is to have them integrate into a single agent. All externalities would be internalized if the whole economy was integrated.
If we leave aside imperfect information and the associated strategic behaviours, the planning problem of these integrated agents is more complicated than if externalities were not present. Planning procedures appropriated to externalities have been provided.

Externalities and cooperative game theory

Suppose we attempt to represent the outcome of cooperation in an economy with externalities by the core. The core is the set of allocations which are not blocked by any coalition. A coalition blocks an allocation if it can do better for all its members than this allocation.
Externalities introduce a difficulty in the definition of a blocking coalition. When a group of agents envision forming a coalition they must conjecture what will be the behaviour of the complementary coalition since it is affected by the externalities of this complementary coalition.
Two extreme notions have been proposed. In the α-core a coalition is said to block an allocation if it can do better, whatever the actions of the complementary coalition. This is extremely prudent. In the β-core a coalition is said to block an allocation if, for any action of the complementary coalition, it can do better. The β-core is of course included in the α-core.
Results depends a lot on these conjectures about the actions of the complementary coalition, an unsatisfactory feature. One lesson, however, is that the core may be empty, that is, that externalities introduce an element of instability in economic games.

Historical note

Following the pioneering work by Sidgwick (1887) and Marshall (1890), Pigou (1920) has provided the basic theory of static technological externalities. Coase (1960) has explained how initial rights could be assigned in various ways. Arrow (1969) has explained how externalities could be internalized by the creation of additional markets. Starrett (1972) has pointed out the associated problem of non-convexity. The first theorem of existence of an equilibrium with externalities has been provided by McKenzie (1955). Shapley and Shubik (1969) studied the core with externalities. A large number of authors have studied various second-best problems associated with externalities (Buchanan, 1969; Plott, 1966;

Diamond, 1973; Sandmo, 1975).

See Also

- clubs
- Coase theorem
- external economies
- incentive compatibility

Bibliography

Arrow, K. 1969. The organization of economic activity: issues pertinent to the choice of market versus non-market allocation. In Joint Economic Committee, *The Analysis and Evaluation of Public Expenditures: the PPB System*, Washington, DC: Government Printing Office.

Buchanan, J.M. 1969. External diseconomies, corrective taxes and market structure. *American Economic Review* 59, 174–6.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Diamond, P. 1973. Consumption of externalities and imperfect corrective pricing. *Bell Journal of Economics and Management Science* 4, 526–38.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

McKenzie, L. 1955. Competitive equilibrium with dependent consumer preferences. In *Proceedings of the Second Symposium in Linear Programming*, ed. H.A. Antosiewicz. Washington, DC: National Bureau of Standards.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Plott, C.R. 1966. Externalities and corrective taxes. *Economica* 33, 84–7.

Sandmo, A. 1975. Optimal taxation in the presence of externalities. *Swedish Journal of Economics* 77, 96–8.

Shapley, L. and Shubik, M. 1969. On the core of an economic system with externalities. *American Economic Review* 59, 687–9.

Sidgwick, H. 1887. *Principles of Political Economy*, 2nd edn. London: Macmillan.

Starrett, D. 1972. Fundamental non-convexities in the theory of externalities. *Journal of Economic Theory* 4, 180–99.

How to cite this article

Laffont, J. J. "externalities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000200> doi:10.1057/9780230226203.0537

The New Palgrave Dictionary of Economics Online

extremal quantiles and value-at-risk

Victor Chernozhukov and Songzi Du
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article examines the theory and empirics of extremal quantiles in economics, in particular value-at-risk. The theory of extremes has gone through remarkable developments and produced valuable empirical findings since the late 1980s. We emphasize conditional extremal quantile models and methods, which have applications in many areas of economic analysis. Examples of applications include the analysis of factors of high risk in finance and risk management, the analysis of socio-economic factors that contribute to extremely low infant birthweights, efficiency analysis in industrial organization, the analysis of reservation rules in economic decisions, and inference in structural auction models.

Keywords

bootstrap; extremal conditional quantiles; extremal quantiles; gamma variables; inference; maximum likelihood; production frontiers; quantile regression function; value-at-risk

Article

1 Introduction

1.1 Some basics

Let a real random variable *Y* have a continuous distribution function $F_Y(y) = \text{Prob}\,[Y \le y]$. A τ-quantile of *Y* is the number $F_Y^{-1}(\tau)$ such that $\text{Prob}\,[Y \le F_Y^{-1}(\tau)] = \tau$ for some $\tau \in (0,1)$. (More generally, let $F_Y^{-1}(\tau) = \inf\left\{y: F_Y(y) > \tau\right\}$.) The quantile function $F_Y^{-1}(\tau)$, viewed as a function of probability index τ, is the inverse of the distribution function $F_Y(y)$. The quantile function is therefore a complete description of the distribution.
Let *X* be a vector of regressor variables. Let $F_Y(y|x) = \text{Prob}\,[Y \le y|x]$ denote the conditional distribution function of *Y* given *X*=*x*. The conditional τ-quantile of a random variable *Y* with a continuous conditional distribution function is the number $F_Y^{-1}(\tau|x)$ such that $\text{Prob}\,[Y \le F_Y^{-1}(\tau|x)|X = x] = \tau$. The conditional quantile function $F_Y^{-1}(\tau|x)$ viewed as a function of *x* is called the τ-quantile regression function. The main use of the quantile regression function $F_Y^{-1}(\tau|x)$ is to measure the effect of covariates on outcomes, both in the centre and in the upper and lower tails of an outcome distribution. To this effect, a quantile or a conditional τ-quantile will be referred to as *extremal* whenever the probability index τ is either low, $\tau \le 0.15$, or high, $\tau \ge 0.85$. Without loss of generality, we focus the discussion on the low quantiles.

1.2 Examples as a motivation

There are many applications of extremal quantiles in economics, particularly of extremal conditional quantiles. Here we give a sample of these applications as a motivation for what follows.

Example 1: Conditional value-at-risk. Value-at-risk analysis seeks to forecast or explain very low conditional quantiles $F_Y^{-1}(\tau|X)$ of an institution's portfolio return, *Y*, tomorrow, using today's available information, *X* (Chernozhukov and Umantsev, 2001). Typically, extremal quantiles $F_Y^{-1}(\tau|X)$ with τ =0.01 and τ =0.05 are of interest. Value-at-risk analysis is a daily activity for banking and other financial institutions, as required by the Securities and Exchange Commission and the Basle Committee on Banking Supervision. As a risk measure, value-at-risk is motivated by the safety-first decision principle formalized by Roy (1952), in which one makes optimal decisions subject to the constraint that the probability of the risk of a large loss is kept small. This and similar measures are commonly used in real-life financial management, insurance, and actuarial science (Embrechts, Klüppelberg and Mikosch, 1997).

Example 2: Determinants of very low birthweights. In the analysis of infant birthweights, we may be interested in how smoking, absence of prenatal care, and other types of maternal behaviour affect various birthweights (Abrevaya, 2001). Of special interest, however, are the very low quantiles, since low birthweights have been linked to subsequent health problems. Chernozhukov (2006) provides an empirical study of extreme birthweights.

Example 3: Probabilistic production frontiers. An important form of efficiency analysis in the economics industrial organization and regulation is the determination of efficiency or production frontiers (Timmer, 1971). Given cost of production and possibly other factors, *X*, we are interested in the highest production levels that only a small fraction of firms, the most efficient firms, can attain. These (nearly) efficient production levels can be formally described by extremal quantile regression function, $F_Y^{-1}(\tau|X)$ for $\tau \in [1-\varepsilon, 1]$ and $\varepsilon > 0$; so only an ε -fraction of firms produce $F_Y^{-1}(\tau|X)$ or more. The models and methods discussed in this article are highly pertinent for inference on the probabilistic frontiers.

Example 4: (S,s)-rules and other approximate reservation rules in economic decisions. A related example is that of (S,s)-adjustment models, which arise as optimal policies in many economic models (Arrow, Harris and Marschak, 1951). For example, the capital stock *Z* is adjusted up to the level *S* once it has depreciated to some low level *s*. In terms of an econometric specification, we may think that the observed capital stock satisfies the equation $Z_i=s(X_i)+v_i$, where X_i are covariates, and v_i is a disturbance that is positive most of the time, that is, $\text{Prob}(v_i \geq 0)$ is close to 1. Once the capital stock Z_i reaches the critical level, that is $Z_i \leq s(X_i)$, it is adjusted in the next period. We assume, as in Caballero and Engel (1999), that when the disturbance $v_i=Z_i-s(X_i)$ is negative, it captures unobserved heterogeneity and small decision mistakes that are independent of observed covariates X_i. (In this example, Z_i could be any monotone transformation of the stock variable. For instance, the log transformation gives an accelerated failure time model for the capital stock. Caballero and Engel, 1999, explore such specifications for empirical (S,s) models in detail.) In a given cross-section or time series, adjustment will occur infrequently, so in fact data at or below the lower adjustment band $s(X_i)$ will be observed with a small probability $\text{Prob}(v_i \leq 0)$; hence $F_Z^{-1}(\tau|X) = s(X) + F_V^{-1}(\tau)$ for $\tau \in (0, \text{Prob}(v \leq 0))$. The lower-band function $s(X)$ therefore coincides with the lower conditional quantile function up to an additive constant. A similar argument works for the upper-band function $S(X)$.

Example 5: Structural auction models. In the standard specification of the first-price procurement auction where bidders hold independent valuations, the winning bid, B_i, satisfies the equation $B_i = c(X_i)\beta(n_i) + \varepsilon_i,\ \varepsilon_i \geq 0$, where $c(X_i)$ is the efficient cost function and $\beta(n_i) \geq 1$ is a mark-up that approaches 1 as the number of bidders, n_i, approaches infinity (Donald and Paarsch, 2002). By construction, $c(X_i)\,\beta(n_i)$ is the extreme conditional quantile function. In empirical analysis, it is realistic to let the disturbance ε_i take some small negative value so that, when negative, these disturbances capture small decision mistakes that are independent of included explanatory variables. In this case the quantile function satisfies $F_B^{-1}(\tau|X, n) = C(X)\beta(n) + F_\varepsilon^{-1}(\tau),\ \text{for } \tau \in (0, P[\varepsilon \leq 0])$. Quantile regression methods can be employed to make inference on $c(X)$ and $\beta(n)$.

1.3 Organization of the article

The rest of the article is organized as follows. Section 2 describes the basic model of extremal quantiles and extremal conditional quantiles. Section 3 describes basic estimation theory and inference theory. Section 4 reviews the key empirical applications and provides an illustrative example. Section 5 concludes.

2 Basic models of extremal quantiles

2.1 A basic model of extremal quantiles

Towards discussing inference methods, assume that the distribution function of the response variable *Y* has Pareto-type tails, which means that tails behave approximately like power functions. Such tails are prevalent in economic data, as discovered by the prominent Italian econonometrician Vilfredo Pareto in 1895 (see Pareto, 1964). Pareto-type tails encompass or approximate a rich variety of tail behaviour, including that of thick-tailed and thin-tailed distributions, having either bounded or unbounded support, and their mathematical theory

in connection to extreme value theory has been developed by Gnedenko (1943) and de Haan (1970).

Consider a random variable Y and define a random variable U as $U \equiv Y$, if the lower end-point of the support of Y is $-\infty$, and $U \equiv Y - F_Y^{-1}(0)$, if the lower end-point of the support of Y is $F_Y^{-1}(0) > -\infty$. The distribution function of U, denoted by F_U then has the lower end-point $F_U^{-1}(0) = -\infty$ or $F_U^{-1}(0) = 0$. The assumption that the distribution function F_U and its quantile function F_U^{-1} exhibit Pareto-type behaviour in the tails can be formally stated as the following two equivalent conditions (the notation $a \sim b$ means that $a/b \to 1$ as appropriate limits are taken):

$$F_U(u) \sim \bar{L}(u) \cdot u^{-1/\xi} \text{ as } u \searrow F_U^{-1}(0), \tag{2.1}$$

$$F_U^{-1}(\tau) \sim L(\tau) \cdot \tau^{-\xi} \text{ as } \tau \searrow 0, \tag{2.2}$$

for some real number $\xi \neq 0$, where $\bar{L}(u)$ is a nonparametric, slowly varying function at $F^{-1}(0)$, and $L(\tau)$ is a nonparametric slowly varying function at 0. (A function $u \mapsto L(u)$ is said to be slowly varying at 0 if $\lim_{l \searrow s}[L(l) / L(ml)] = 1$ for any $m>0$.) The prime examples of slowly varying functions are the constant function $L(y)=L$ and the logarithmic function. The number ξ defined in (2.1) and (2.2) is called the extreme value (EV) index.

The absolute value $|\xi|$ of the EV index ξ measures heavy tailedness of distributions. A distribution F_Y with Pareto-type tails necessarily has a finite lower support point if $\xi<0$ and an infinite lower support point if $\xi>0$. Distributions with $\xi>0$ include stable distributions, Pareto distributions, t distributions, and many others. For example, the t distribution with v degrees of freedom has the EV index $\xi=1/\nu$ and exhibits a wide range of tail behaviour. In particular, setting $\nu=1$ yields the Cauchy distribution which has heavy tails (with $\xi=1$), while setting $\nu=30$ yields approximately the normal distribution which has light tails (with $\xi=1/30$). On the other hand, distributions with $\xi<0$ include the uniform, exponential, Weibull distributions, and others.

The assumption of Pareto-type tails can be equivalently cast in terms of the regular variation assumption, as is commonly done in EV theory. Distribution function F_U is said to be *regularly varying* at $F_U^{-1}(0)$ with index of regular variation $-1/\xi$ if $\lim_{y \searrow F_U^{-1}(0)}(F_U(ym) / F_U(y)) = m^{-1/\xi}$, for any $m>0$. This condition is equivalent to the regular variation of quantile function F_U^{-1} at 0 with index $-\xi$:

$$\lim_{\tau \searrow 0}(F_U^{-1}(\tau m) / F_U^{-1}(\tau)) = m^{-\xi}, \text{ for any } m > 0.$$

It should be mentioned that the case of $\xi=0$ corresponds to the class of rapidly varying distribution functions. These distribution functions have exponentially light tails, with the normal and exponential distributions being the chief examples. To simplify exposition, we do not discuss this case explicitly. However, since the limit distribution of main statistics are continuous in ξ, including at $\xi=0$, inference theory for the case of $\xi=0$ can be adequately approximated by the case of $\xi \approx 0$.

2.2 A basic model of extremal conditional quantiles

Consider the classical linear functional form for the conditional quantile function of Y given $X=x$:

$$F_Y^{-1}(\tau|x) = x'\beta(\tau), \text{ for all } \tau \in \mathcal{T} = (0, \eta], \text{ some } \eta \in (0, 1], \tag{2.3}$$

and for every x in the support of X. This linear functional form is flexible in the sense that it has good approximation properties. Given the original regressor X^*, the final set of regressors X to be used in estimation can be formed as a vector of approximating functions. For example, X may include power functions, splines, and other transformations of X^*. The linear functional form also provides computational convenience.

The following model for the tails and its generalizations were developed in Chernozhukov (2005). The main assumption is that the response variable Y, transformed by some auxiliary regression line, has regularly varying tails with EV index ξ. Indeed, in addition to (2.3), suppose there exists an auxiliary parameter β_e such that the disturbance $U \equiv Y - X'\beta_e$ has conditional end-point 0 or $-\infty$ a.s. and its conditional quantile function $F_U^{-1}(\tau|x)$ satisfies the following tail-equivalence relationship as $\tau \searrow 0$, uniformly for x in the support of X:

$$F_U^{-1}(\tau|x) = F_Y^{-1}(\tau|x) - x'\beta_e \sim F_U^{-1}(\tau), \tag{2.4}$$

where F_U^{-1} is a quantile function such that

$$F_U^{-1}(\tau) \sim L(\tau)\tau^{-\xi}, \tag{2.5}$$

where $L(\tau)$ is a non-parametric slowly varying function at 0. Equation (2.5) imposes Pareto-type behaviour on the conditional law, while equation (2.4) requires this behaviour to hold uniformly across conditioning values. Since this assumption only affects the tails, it allows covariates to impact the extremal quantiles and the central quantiles very differently; the impact of covariates on extremal quantiles is approximated by β_e, which could differ sharply from, for example, the impact on the median given by $\beta(1/2)$. Chernozhukov (2005) provides further generalizations of this model.

3 Basic estimation methods

3.1 Estimates based on sample quantiles

Given T observations $\{Y_t, \bullet t=1, \ldots, T\}$, the τ-sample quantile can be obtained by solving the following optimization problem:

$$\hat{F}_Y^{-1}(\tau) \in \arg\min_{\beta \in R^d} \sum_{t=1}^{T} \rho_\tau(Y_t - \beta),$$

(3.6)

where $\rho_\tau(u) = (\tau - 1(u < 0))u$ is the asymmetric absolute deviation function of Fox and Rubin (1964).

Sample quantiles are also order statistics, and we will refer to τT as the *order* of τ-quantile. The sequence of quantile index-sample size pairs (τ, T) will be said to be an *extreme order* sequence if $\tau \searrow 0$ and $\tau T \to k > 0$, an *intermediate order* sequence if $\tau \searrow 0$ and $\tau T \to \infty$, and a *central order* sequence if τ is fixed and $T \to \infty$. Each different type of the sequence leads to different asymptotic approximations to the finite-sample distributions of sample quantiles. Extreme order sequences lead to non-normal (extreme-value) distributions (EV) that approximate the finite sample distributions of extremal (high and low) quantiles much better than the normal distributions do. In particular, EV distributions work much better than the normal distribution if $\tau T \leq 30$.

3.1.1 Extreme order quantiles

Consider an extreme-order sequence. The following is the classical result on the limit distribution of order statistics: for any integer $k \geq 1$ and $\tau = k/T$, as $T \to \infty$,

$$A_T(\hat{F}_Y^{-1}(\tau) - F_Y^{-1}(\tau)) \to_d \Gamma_k^{-\xi} - k^{-\xi} \tag{3.7}$$

where

$$A_T = 1 / F_U^{-1}(1/T), \Gamma_k = \xi_1 + \cdots + \xi_k \tag{3.8}$$

and $(\xi_1, \xi_2, \ldots)$ is an independent and identically distributed sequence of standard exponential variables.

Result (3.7) was obtained by Gnedenko (1943) under the assumption that $Y_1, Y_2, \bullet\ldots$ is a sequence of independently and identically distributed (i.i.d.) random variables. Result (3.7) continues to hold for stationary weakly dependent series, provided the probability of extreme events occurring in clusters is negligible relative to the probability of a single extreme event (Meyer, 1973). The results have been generalized to more general time series processes (Leadbetter, Lindgren and Rootzén, 1983).

Result (3.7) gives an EV distribution as an approximation to the finite-sample distribution of $\hat{F}_Y^{-1}(\tau)$. The EV distribution is characterized by the EV index ξ, which can be estimated by one of the methods described below. Variables Γ_k, entering the definition of the EV distribution, are known as *gamma* random variables. The limit distribution of the *k*th-order statistic is therefore a transformation of a gamma variable. The EV distribution is not symmetric and may have significant (median) bias. The EV distribution has finite moments if $\xi < 0$ and has finite moments of up to order $1/\xi$ if $\xi > 0$.

The classical result is not feasible for purposes of inference on $F_Y^{-1}(\tau)$, since the scaling constant A_T is not easily estimable consistently. One way to overcome this problem is to make additional strong assumptions in order to estimate A_T consistently. For instance, suppose that $F_U^{-1}(\tau) \sim L\tau^{-\xi}$, then one can estimate ξ using methods described below and L by $\hat{L} = (\hat{F}_Y^{-1}(2\tau) - \hat{F}_Y^{-1}(\tau)) / (2^{-\hat{\xi}} - 1)\tau^{-\hat{\xi}}$.

Another way to overcome the aforementioned infeasibility is to consider the asymptotics of self-normalized extreme order quantiles, as in Chernozhukov (2006):

$$Z_T(\tau) = \mathcal{A}_T(\hat{F}_Y^{-1}(\tau) - F_Y^{-1}(\tau)) \to_d \frac{\sqrt{k}(\Gamma_k^{-\xi} - k^{-\xi})}{\Gamma_{mk}^{-\xi} - \Gamma_k^{-\xi}}$$

(3.9)

where for $m>1$ such that mk is an integer,

$$\mathcal{A}_T = \frac{\sqrt{\tau T}}{\hat{F}_Y^{-1}(m\tau) - \hat{F}_Y^{-1}(\tau)}.$$

(3.10)

Here, the scaling factor $\mathcal{A}_T$ is feasible in that it is completely a function of data. The limit distribution only depends on the EV index ξ, and its quantiles can be easily calculated analytically or by simulation.

3.1.2 Intermediate order quantiles

Consider next an intermediate order sequence. As $\tau \searrow 0$ and $\tau T \to \infty$, under further regularity conditions,

$$Z_T(\tau) = \mathcal{A}_T(\hat{F}_Y^{-1}(\tau) - F_Y^{-1}(\tau)) \to_d \mathcal{N}\left(0, \frac{\xi^2}{(m^{\xi} - 1)^2}\right),$$

(3.11)

where $\mathcal{A}_T$ is defined as in (3.10). This result, obtained by Dekkers and de Haan (1989), gives a normal asymptotic approximation to the finite-sample distribution of sample quantile $\hat{F}_Y^{-1}(\tau)$. The main condition for application of this distribution is that $\tau T \to \infty$. In finite samples, we may interpret this as requiring that $\tau T \geq 30$ at the minimum.
The normal approximation (3.11) is convenient, but extreme approximation (3.9) is always better, because it does not fail when $\tau T \to k < \infty$ and it coincides with the normal approximation (3.11) once k is large.

3.1.3 Extremal bootstrap

In many cases it is convenient to implement inference using the following approach. Consider the sample of i.i.d. variables:

$$(Y_1, \ldots, Y_T) = \left(\frac{\xi_1^{-\xi} - 1}{-\xi}, \ldots, \frac{\xi_T^{-\xi} - 1}{-\xi}\right),$$

(3.12)

where $(\xi_1, \ldots, \xi_T)$ is an i.i.d. sequence of standard exponential variables. (Y_t, defined in this way, follows generalized extreme value distribution, which nests the Frechet, Weibull, and Gumbell distributions. There are other possibilities, for example, $(\xi_1, \ldots, \xi_T)$ in (3.13) can be replaced by uniform variables $(U_1, \ldots, U_T)$, in which case Y_t follows the generalized Pareto distribution.) Variables generated in this way have the quantile function:

$$F_Y^{-1}(\tau) = \frac{[-\ln(1-\tau)]^{-\xi} - 1}{-\xi}. \tag{3.13}$$

Observe that $F_Y^{-1}(\tau) - 1/\xi \sim \tau^{-\xi}/\xi$, so condition (2.1) is satisfied. We propose to estimate the finite-sample distributions of $Z_T(\tau) = A_T(\hat{F}_Y^{-1}(\tau) - F_Y^{-1}(\tau))$ by the finite-sample distribution of $Z_T(\tau)$ for the case when the data follow (3.13). In this way, we reproduce both the EV limit (3.9) and the normal limit (3.11) under extreme and intermediate sequences, and also guarantee good finite-sample performance for the case when (3.13) holds exactly. The simulation can be done using the following algorithm:

1. 1. For each $i \leq B$, draw $(Y_1, \ldots, Y_T)$ as i.i.d. according to (3.13), replacing ξ with a suitable estimate $\hat{\xi}$. Compute the statistic $Z_{T,i}(\tau) = A_T(\hat{F}_Y^{-1}(\tau) - F_Y^{-1}(\tau))$.
2. 2. Use quantiles of the simulated sample $(Z_{T,i}(\tau), i \leq B)$ for inference purposes.

This scheme could be used to estimate distributions of other statistics, including estimators of the EV index and extrapolation estimators.
Another method, developed in Chernozhukov (2006), is based on subsampling the self-normalized quantile statistic. This method is less accurate than the extremal bootstrap. However, it applies under more general conditions. It should be noted that the canonical (nonparametric) bootstrap does not work in these settings (Bickel and Freedman, 1981).

3.1.4 Confidence intervals for $F_Y^{-1}(\tau)$ and bias correction for $F_Y^{-1}(\tau)$

Let the α-quantile of $Z_T(\tau)$ be denoted by $c(\alpha)$. The estimates of $c(\alpha)$ can be obtained using either EV approximation, normal approximation, or the extremal bootstrap, also having replaced ξ with a suitable estimate. Denote the resulting estimates by $\hat{c}(\alpha)$. Then, the median bias-corrected estimate and α%-confidence region for $F_Y^{-1}(\tau)$ can be constructed as

$$F_Y^{-1}(\tau) - \frac{\hat{c}(1/2)}{A_T} \text{ and } \left[\hat{F}_Y^{-1}(\tau) - \frac{\hat{c}(1-\alpha/2)}{A_T}, \hat{F}_Y^{-1}(\tau) - \frac{\hat{c}(\alpha/2)}{A_T}\right]. \tag{3.14}$$

3.1.5 Estimators of the EV index ξ

There are two principal estimators. The first estimator, due to Pickands (1975), relies on the ratio of sample quantile spacings:

$$\hat{\xi} = -\ln\left[\frac{\hat{F}_Y^{-1}(4\tau) - \hat{F}_Y^{-1}(2\tau)}{\hat{F}_Y^{-1}(2\tau) - \hat{F}_Y^{-1}(\tau)}\right] / \ln 2, \tag{3.15}$$

such that $\tau \to 0$ and $\tau T \to \infty$ as $T \to \infty$. Under further regularity conditions,

$$\sqrt{\tau T}(\hat{\xi} - \xi) \to_d \mathcal{N}\left(0, \frac{\xi^2(2^{2\xi+1} + 1)}{(2(2^{\xi} - 1)\ln 2)^2}\right). \tag{3.16}$$

Another estimator, developed by Hill (1975), is a moments estimator (notation $(x)_-$ means $(x)_- = -x$ if $x<0$ and $(x)_- = 0$ if $x \geq 0$):

$$\hat{\xi} = \frac{\Sigma_{t=1}^{T} \ln(Y_t / \hat{F}_Y^{-1}(\tau) -}{T\tau} \tag{3.17}$$

such that $\tau \to 0$ and $\tau T \to \infty$ as $T \to \infty$. This estimator is applicable only for the case of $\xi > 0$. The estimator can be motivated by a maximum likelihood method that fits an exact power law to the tail data. Under further regularity conditions,

$$\sqrt{\tau T}(\hat{\xi} - \xi) \to_d \mathcal{N}(0, \xi^2). \tag{3.18}$$

The methods for choosing τ are described in Embrechts, Klüppelberg and Mikosch (1997). The variance of estimators decreases as τ increases, but the bias (relative to the true ξ) goes up. Another view on the choice of τ is the following: statistical models are approximations, not literal descriptions of the data. In practice, dependence of $\hat{\xi}$ on the threshold τ reflects that power laws with different values of ξ fit better different tail regions. Therefore, if the interest lies in making inference on $F_Y^{-1}(\tau)$ for a particular τ, it seems reasonable to use $\hat{\xi}$ constructed using the same τ or most similar τ' subject to the condition that $\tau' T \geq 30$. (The latter condition requires that a sufficient sample be available to estimate ξ.) The limit results above can be used for the construction of confidence regions. As an alternative, we can apply extremal bootstrap to statistic $Z_T = \sqrt{\tau T}(\hat{\xi} - \xi)$ to estimate the quantiles of Z_T. Given the estimated α-quantiles $\hat{c}(\alpha)$, we can construct the median bias-corrected estimate and α%-confidence regions for ξ.

3.1.6 Extrapolation estimators

When very extreme quantiles cannot be estimated precisely, the following strategy is sensible: estimate less extreme quantiles reliably, and then extrapolate these estimates using the

assumptions on tail behaviour stated earlier. Dekkers and de Haan (1989) developed the following extrapolation estimator:

$$\hat{F}_Y^{-1}(\tau_e) = \frac{(\tau_e/\tau)^{-\xi} - 1}{2^{-\xi} - 1}[\hat{F}_Y^{-1}(2\tau) - \hat{F}_Y^{-1}(\tau)] + \hat{F}_Y^{-1}(\tau), \tag{3.19}$$

where $\tau_e \ll \tau$. Another useful estimator, which is valid only for the case of $\xi > 0$, is the following:

$$\hat{F}_Y^{-1}(\tau_e) = (\tau_e/\tau)^{-\xi} \cdot \hat{F}_Y^{-1}(\tau), \tag{3.20}$$

where $\tau_e \ll \tau$. The above estimators have good properties provided the quantities on the right-hand side of (3.19) and (3.20) are well estimated, which requires that τT be large, and that the tail model be a good approximation of the underlying true tail.

3.2 Estimates based on sample regression quantiles

Given T observations $\{Y_t, X_t, t=1, \ldots, T\}$, the quantile regression estimate of $F_Y^{-1}(\tau|x)$ is given by:

$$\hat{F}_Y^{-1}(\tau|x) = x'\hat{\beta}(\tau),\ \hat{\beta}(\tau) = \arg\min_{\beta \in \mathbb{R}^d} \sum_{t=1}^{T} \rho_\tau(Y_t - X_t'\beta), \tag{3.21}$$

where $\rho_\tau(u) = (\tau - 1(u < 0))u$. Quantile regression was introduced by Laplace (1818) for the median case. Koenker and Bassett (1978) extended this formulation to other quantiles.

3.2.1 Extreme order asymptotics

Chernozhukov (2005) derives asymptotic distributions of regression quantiles under extreme-order sequences. Consider the canonically normalized QR statistic

$$\hat{Z}_T(k) = A_T(\hat{\beta}(\tau) - \beta(\tau)),\ \text{where } A_T = 1/F_U^{-1}(1/T), \tag{3.22}$$

and the self-normalized QR statistic

$$Z_T(k) = \mathcal{A}_T(\hat{\beta}(\tau) - \beta(\tau)), \text{ where } \mathcal{A}_T = \frac{\sqrt{\tau T}}{\bar{X}'(\hat{\beta}(m\tau) - \hat{\beta}(\tau))},$$
(3.23)

where $\tau T(m-1)>d$. The first statistic uses an infeasible canonical normalization, while the second statistic uses a feasible normalization. Then as $\tau T \to k > 0$ and $T \to \infty$

$$\hat{Z}_T(\tau) \to_d Z_\infty(k) - k^{-\xi},$$
$$Z_\infty(k) = \arg\min_{z \in \mathbb{R}^d} \left[-kE[X]'z + \sum_{i=1}^{\infty} \left[X_i' z - \Gamma_i^{-\xi} \right]_+ \right] (\xi < 0)$$
$$Z_\infty(k) = \arg\min_{z \in \mathbb{R}^d} \left[-kE[X]'z + \sum_{i=1}^{\infty} \left[X_i' z + \Gamma_i^{-\xi} \right]_+ \right] (\xi > 0)$$
(3.24)

where $\{\Gamma_1, \Gamma_2, \ldots\} := \{\xi_1, \xi_1 + \xi_2, \ldots\}$ and $\{\xi_1, \xi_2, \ldots\}$ is an i.i.d. sequence of exponential variables that is independent of $\{X_1, X_2, \ldots\}$. Further, for any m such that $k(m-1)>d$,

$$Z_T(\tau) \to_d \frac{\sqrt{k} Z_\infty(k)}{E[X]'(Z_\infty(mk) - Z_\infty(k))}.$$
(3.25)

The results hold under the assumption that the data come from either an i.i.d. sequence or a stationary weakly dependent sequence with extreme events satisfying a non-clustering condition.
Related results for canonically normalized statistics for the case where $\tau T \to 0$ as $T \to \infty$ have been obtained by Knight (2001) and Portnoy and Jure•ková (1999).

3.2.2 Intermediate order asymptotics

Chernozhukov (2005) shows that under intermediate order sequences, as $\tau \searrow 0$ and $\tau T \to \infty$,

$$Z_T(\tau) = \mathcal{A}_T(\hat{\beta}(\tau) - \beta(\tau)) \to_d \mathcal{N}\left(0, [E(XX')]^{-1} \frac{\xi^2}{(m^{-\xi} - 1)^2}\right),$$
(3.26)

where A_T is defined as in (3.23). Like the result under extreme order sequences, this result holds under the assumption that the data come either from an i.i.d. sequence or from a stationary weakly dependent sequence with extreme events satisfying a non-clustering condition.

3.2.3 Extremal bootstrap

In practice, it is convenient to implement inference by constructing a bootstrap model that approximates the tail features of the true conditional quantile model under the assumptions of Section 2.2; then using this model to simulate the distributions of estimators of extreme quantiles and tail parameters.
Consider the sample

$$((Y_1, X_1), \ldots, (Y_T, X_T)) = \left(\left(\frac{\xi_1^{-\xi} - 1}{-\xi}, X_1\right), \ldots, \left(\frac{\xi_T^{-\xi} - 1}{-\xi}, X_T\right)\right), \tag{3.27}$$

where $(\xi_1, \ldots, \xi_T)$ is an i.i.d. sequence of standard exponential variables and $(X_1, \ldots, X_T)$ is a fixed set of observations on regressors that we have. Variable Y_t generated in this way has the conditional quantile function

$$F_{Y_t}^{-1}(\tau|X_t) = X_t'\beta(\tau) = \frac{[-\ln(1-\tau)]^{-\xi} - 1}{-\xi}, \text{where } \beta(\tau) = \left(\frac{[-\ln(1-\tau)]^{-\xi} - 1}{-\xi}, 0, \ldots, 0\right)'. \tag{3.28}$$

Observe that $F_Y^{-1}(\tau|X_i) - 1/\xi \sim \tau^{-\xi}/\xi$, so the model satisfies conditions (2.4) and (2.5), as does the true conditional quantile model under our assumptions. Hence we can estimate the finite-sample distributions of $Z_T(\tau) = A_T(\hat{\beta}(\tau) - \beta(\tau))$ by the finite-sample distribution of $Z_T(\tau)$ in the case when data follows (3.27). In this simple way, we can replicate both the EV approximation (3.26) and the normal approximation (3.24) and also guarantee good finite-sample performance for the case when the model (3.27) holds. The simulation can be done using the following algorithm:

1. 1. For each $i \le B$, draw data according to (3.27), replacing ξ with a suitable estimate $\hat{\xi}$. Compute the statistic $Z_{T,i}(\tau) = A_T(\hat{\beta}(\tau) - \beta(\tau))$.
2. 2. Use the empirical distribution of the simulated sample $(Z_{T,i}(\tau), i \le B)$ for inference.

This method can also be used to estimate distributions of estimators of the EV index and extrapolation estimators described below.
As mentioned before, there is another inference method proposed by Chernozhukov (2006) which uses subsampling to estimate the distribution of the self-normalized statistic $Z_T(\tau)$. This method is less accurate than the extremal bootstrap, but it applies under more general conditions.

3.2.4 Confidence intervals and bias corrected estimates

Suppose we are interested in the parameter $\psi'\beta(\tau)$ for some non-zero vector ψ. Let the α-quantile of $\psi' Z_T(\tau)$ be denoted by $c(\alpha)$. Having replaced ξ with a suitable estimate, the estimates of $c(\alpha)$ can be obtained using either the EV approximation, normal approximation or the extremal bootstrap. Denote the resulting estimates by $\hat{c}(\alpha)$. The

median-bias corrected estimator and the α%-confidence interval for $\psi'\beta(\tau)$ can be constructed as

$$\psi'\hat{\beta}(\tau) - \frac{\hat{\tau}(1/2)}{\mathcal{A}_T} \text{ and } \left[\psi'\hat{\beta}(\tau) - \frac{\hat{\tau}(\alpha/2)}{\mathcal{A}_T}, \psi'\hat{\beta}(\tau) - \frac{\hat{\tau}(1-\alpha/2)}{\mathcal{A}_T}\right]. \tag{3.29}$$

3.2.5 Estimators of the EV index ξ

The following estimators are regression analogs of the Pickands and Hill estimators. The first estimator takes the form

$$\hat{\xi} = -\ln\left[\frac{\hat{F}_Y^{-1}(4\tau|\bar{X}) - \hat{F}_Y^{-1}(2\tau)|\bar{X})}{\hat{F}_Y^{-1}(2\tau|\bar{X}) - \hat{F}_Y^{-1}(\tau|\bar{X})}\right] / \ln 2, \tag{3.30}$$

where $\bar{X}$ is the average value of X_t. Under additional regularity conditions, as $\tau \searrow 0$ and $\tau T \to \infty$

$$\sqrt{\tau T}(\hat{\xi} - \xi) \to_d \mathcal{N}\left(0, \frac{\xi^2(2^{2\xi+1}+1)}{(2(2^{\xi}-1)\ln 2)^2}\right). \tag{3.31}$$

The second estimator, which is applicable when $\xi > 0$, takes the form:

$$\hat{\xi} = -\frac{\Sigma_{t=1}^{T}\ln(Y_t/\hat{F}_Y^{-1}(\tau|X_t)) -}{T\tau}. \tag{3.32}$$

Under additional regularity conditions, as $\tau \searrow 0$ and $\tau T \to \infty$

$$\sqrt{\tau T}(\hat{\xi} - \xi) \to_d \mathcal{N}(0, \xi^2). \tag{3.33}$$

The limit results above can be used for the construction of confidence regions. An alternative approach is to apply the extremal bootstrap to statistic $Z_T = \sqrt{\tau T}(\hat{\xi} - \xi)$ to estimate the quantiles of this statistic. Then we can use estimated α-quantiles $\hat{t}(\alpha)$ for constructing the median bias-corrected estimate and α%-confidence regions for ξ:

$$\hat{\xi} - \frac{\hat{t}(1/2)}{\sqrt{\tau T}} \text{ and } \left[\hat{\xi} - \frac{\hat{t}(1-\alpha/2)}{\sqrt{\tau T}}, \hat{\xi} - \frac{\hat{t}(\alpha/2)}{\sqrt{\tau T}}\right]. \tag{3.34}$$

3.2.6 Extrapolation estimators

By analogy with the unconditional case, the extrapolation estimators for $\hat{F}_Y^{-1}(\tau_e|x)$, where τ_e is a very low value, can be constructed as

$$\hat{F}_Y^{-1}(\tau_e|x) = \frac{(\tau_e/\tau)^{-\hat{\xi}} - 1}{m^{-\hat{\xi}} - 1}[\hat{F}_Y^{-1}(m\tau|x) - \hat{F}_Y^{-1}(\tau|x))] + \hat{F}_Y^{-1}(\tau|x), \tag{3.35}$$

$$\hat{F}_Y^{-1}(\tau_e|x) = (\tau_e/\tau)^{-\hat{\xi}}\hat{F}_Y^{-1}(\tau|x)(\xi > 0), \tag{3.36}$$

where $\tau_e \ll \tau$. Note that the estimator (3.36) is valid only in the case $\xi > 0$. The comments given for the unconditional case apply here as well. Also, we can construct confidence regions for $F_Y^{-1}(\tau_e|x)$ based on extrapolation estimators. This can be done by applying the extremal bootstrap to statistic $Z_T = A_T(\hat{F}_Y^{-1}(\tau_e|x) - \hat{F}_Y^{-1}(\tau_e|x))$ where $A_T = \sqrt{\tau T}/\bar{X}'(\hat{\beta}(2\tau) - \hat{\beta}(\tau))$.

4 Empirical applications: an overview and an illustration

4.1 A simple overview

The following review is not exhaustive by any means; it aims to provide only a few quintessential references.

4.1.1 Extremal unconditional quantiles

As mentioned in Section 2, Pareto analysed income and wealth data in 1895 and suggested that power laws accurately describe the tail data. Pareto's discovery, although rmkably simple, had a profound effect on both empirics and the theory of extremes. Zipf (1949), Mandelbrot (1963), Fama (1965), Praetz (1972), Sen (1973), Jansen and de Vries (1991), and Longin (1996), among others, gave further empirical evidence on the nature and prevalence of Pareto-type laws in economic data, including city sizes, incomes, and financial returns. It should be mentioned that many of the early studies were highly informal in nature. The theoretical work in extreme value theory has opened paths for better analysis. From this aspect, the study of Jansen and de Vries (1991) can be singled out as it gave, to our knowledge, the first highly rigorous analysis of the tail properties of financial returns. Jansen and de Vries (1991) estimate the EV indices for various primary US stocks to be between ξ =1/5 and ξ =1/3. Using quantile extrapolation estimators to estimate value-at-risk, they also conclude that the 1987 market crash was not an outlier. Rather it was a rare event, the magnitude of which could have been predicted using prior data. This study stimulated numerous other studies that rigorously document the tail properties of economic data (Embrechts, Klüppelberg and Mikosch, 1997).

4.1.2 Extremal conditional quantiles

There has been considerably less work on conditional methods. However, following recent theoretical advances we expect that this area will see active development in the near future. In what follows, we merely highlight some of the topics and directions.

In what might be the earliest example of conditional quantile analysis, Quetelet (1871) fitted various conditional quantile curves to age–height data. Remarkably, Quetelet's work included tabulations of very high and very low quantiles of heights as a function of age. There is a great potential for the applications of extremal quantile regression methods in similar problems. In a recent study, Chernozhukov (2006) estimates the impact of smoking and maternal behaviour on extremely low birthweights in the United States, focusing on black mothers. He finds that the impact of these variables on birthweights in the ranges between 250 and 1,500 grams sharply differs from their impact on the central birthweights. For instance, smoking is not correlated with extremal birthweights, while quality of prenatal medical care is strongly linked to extremal birthweights.

Aigner and Chu (1968), Timmer (1971), and Aigner, Amemiya and Poirier (1976) pioneered a large empirical literature on production frontiers. A major problem of the subsequent empirical literature has been the lack of statistical methods for construction of reliable estimates and confidence regions. The new methods discussed in Section 3 solve the problem, and should improve the rigour of the empirical work in this area.

There is a considerable appeal for the use of extremal conditional quantile methods in auction models. An important study that illustrates the potential is by Donald and Paarsch (2002), who analyse an empirical structural auction model. They estimate the conditional support function of bids using extreme order statistics for each covariate cell, then project the estimated function onto a lower dimensional structural function implied by the model via a minimum-distance method. A generalization of this approach is to employ extremal quantile regression for estimation of the (approximate) support function in the first stage. (The use of near-extreme quantile regression for estimation of approximate support functions allows the researcher to discard some outliers that do not conform the model, in the spirit of the discussion given in Section 1.)

Value-at-risk is another potentially important area of applications of the extremal quantile regression. Chernozhukov and Umantsev (2001) apply these methods to the problem of forecasting value-at-risk of a major US oil company. They estimate extremal conditional quantiles, using both ordinary and extrapolation methods, and implement confidence regions, using subsampling methods described in Chernozhukov (2006). The section below briefly revisits some of the main qsts of this study.

4.2 An illustrative example

Here we consider a problem of forecasting conditional value-at-risk. We revisit some of the qsts asked in Chernozhukov and Umantsev (2001) with an improved methodology. To implement the analysis, we use algorithms written in R language that rely on Koenker's (2006) quantreg package as the basic platform. The algorithms as well as the data-set can be downloaded from http://www.mit.edu/~vchern/EQR (accessed 23 April 2007). A detailed description of the data-set is given in Chernozhukov and Umantsev (2001).

We estimate the following conditional quantile function for various low values of τ *:*

$$F_{Y_t}^{-1}(\tau|X_t) = X_t'\beta(\tau) = \beta_0(\tau) + \beta_1(\tau)X_{t,1} + \beta_2(\tau)X_{t,2} + \beta_3(\tau)X_{t,3}, \tag{4.37}$$

where Y_t is the daily (log) return on the stock of Occidental Petroleum, $X_{t,1}$ is the lagged return on spot oil price, $X_{t,2}=Y_{t-1}$ is the lagged own return, and $X_{t,3}$ is the lagged return on the

Dow Jones Industrial Index. There are 2527 observations in the sample.

We first estimate and plot the function $(\tau, t) \mapsto X_t'\hat{\beta}(\tau)$ in (τ ,*t*) space, with $\tau \leq 0.25$. The graph of this function, shown in Figure 1, gives a good picture of the evolution of risk over time, indicating dates where the predicted risk is especially high. Let us next determine what causes these risk fluctuations.

Figure 1

The fit $X_t'\hat{\beta}(\tau)$ as a function of time *t* and τ . *Source*: Chernozhukov and Umantsev (2001).

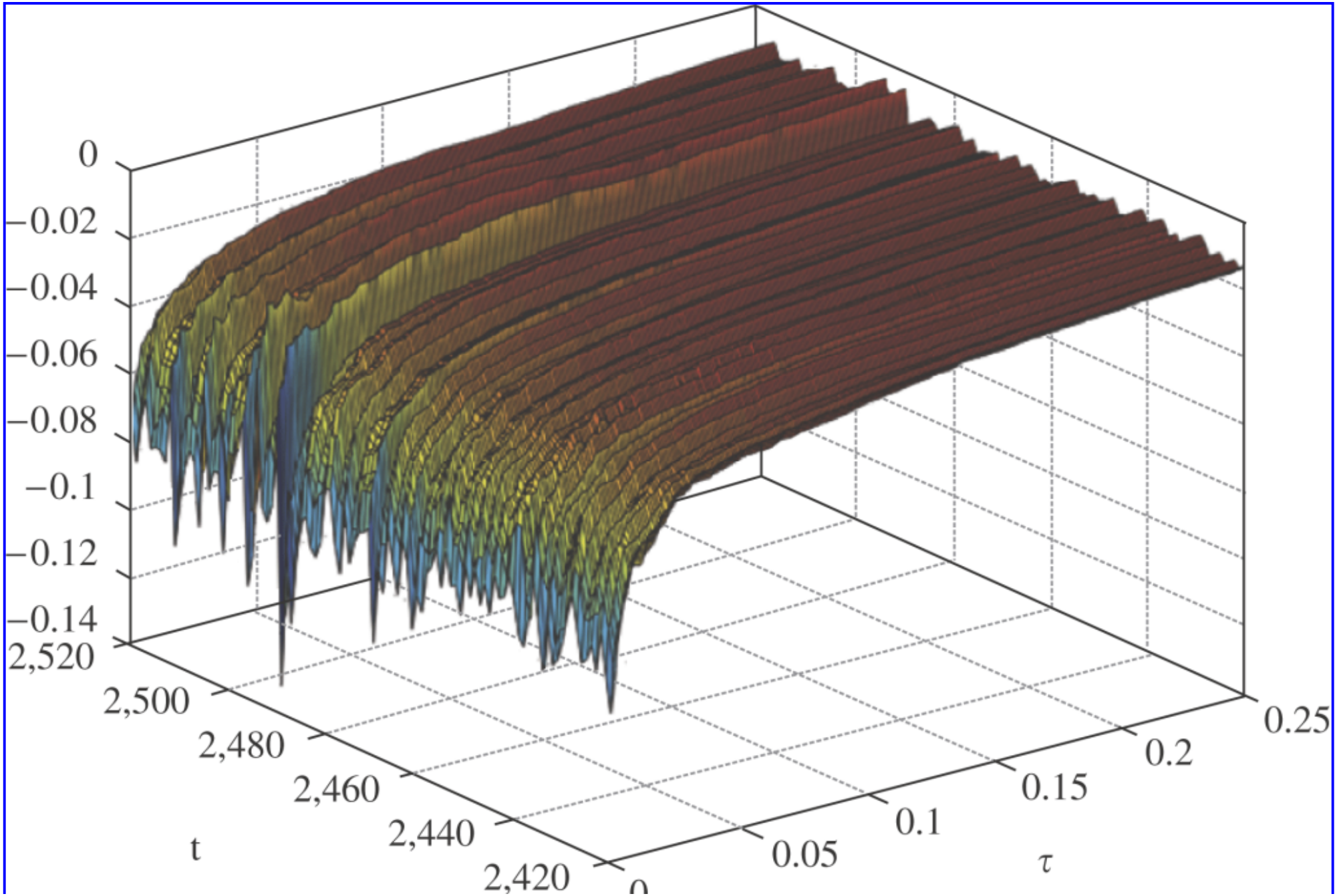

Table 1 reports the estimates of the coefficients of the model (4.37). The table also reports median bias-corrected estimates and 90 per cent confidence regions, which were obtained using the extremal bootstrap approach described in Section 3. The results show that the primary determinant of the high-risk levels is the market. The further in the tail we go, the larger is the magnitude of the point estimate of the coefficient on the DJI return. Moreover, the confidence region for this coefficient excludes zero even at $\tau = 0.001$.

Estimation results

Coefficients	Estimate	Bias-corrected	90% conf. region
$\tau = 0.001$			
Intercept	–0.08	–0.08	[–0.11,–0.06]
Lag return	0.12	0.01	[–0.19, 0.63]
Oil price	–0.05	–0.05	[–0.42, 0.11]
DJI return	0.71	0.73	[0.05, 1.08]
$\tau = 0.01$			
Intercept	–0.05	–0.05	[–0.05,–0.04]
Lag return	–0.04	–0.06	[–0.16, 0.12]
Oil price	–0.05	–0.06	[–0.17, 0.02]
DJI return	0.49	0.50	[0.24, 0.65]
$\tau = 0.05$			
Intercept	–0.03	–0.03	[–0.03,–0.02]
Lag return	–0.03	–0.04	[–0.10, 0.04]
Oil price	0.01	0.01	[–0.04, 0.06]
DJI return	0.29	0.30	[0.17, 0.38]

We next characterize the tail properties of the conditional quantile model. Table 2 reports the estimates of the EV index ξ obtained using estimator (3.32). The table also reports median bias-corrected estimates, and 90 per cent confidence regions, which were obtained using the nested approach described in Sections 3.2.3 and 3.2.5. The bias-corrected estimates tend to be stable with respect to the start of the tail determined by probability index τ. On the basis of Table 2, we take $\hat{\xi} \approx 1/4$ to be the estimate of the EV index.

Estimation results for the EV index ξ

	Estimate	Bias-corrected.	Estimate 90% conf. region
$\tau = 0.005$	0.24	0.22	[0.08, 0.34]
$\tau = 0.01$	0.23	0.17	[0.05, 0.25]
$\tau = 0.025$	0.32	0.24	[0.14, 0.30]
$\tau = 0.05$	0.35	0.23	[0.16, 0.27]

Having characterized the EV index, we can now estimate very extreme quantiles using extrapolation methods. We set the risk level at $\tau = 0.0001$, so that the return falls below $F_{Y_t}^{-1}(0.0001|X_t)$ only once per about 30 years, a very rare, extreme event. The extrapolated estimates of 0.0001-quantile are obtained using equation (3.36) with $\tau = 0.05$ and $\hat{\xi} = 1/4$. The resulting extrapolation fit $\hat{F}_{Y_t}^{-1}(0.0001|X_t)$ sharply differs from the ordinary fit $X_t\hat{\beta}(\tau)$ obtained by quantile regression. The reason for this is simple: the ordinary fit uses sample data that likely contains no observations on the extreme events defined above. In sharp contrast, the extrapolated fit uses the tail model and a reliably estimated conditional 0.05-quantile to predict the magnitude of such events. The quality of this prediction clearly depends on whether the tail model is accurate. Figure 2

Figure 2

Extrapolated and ordinary estimates of the conditional 0.0001-quantile

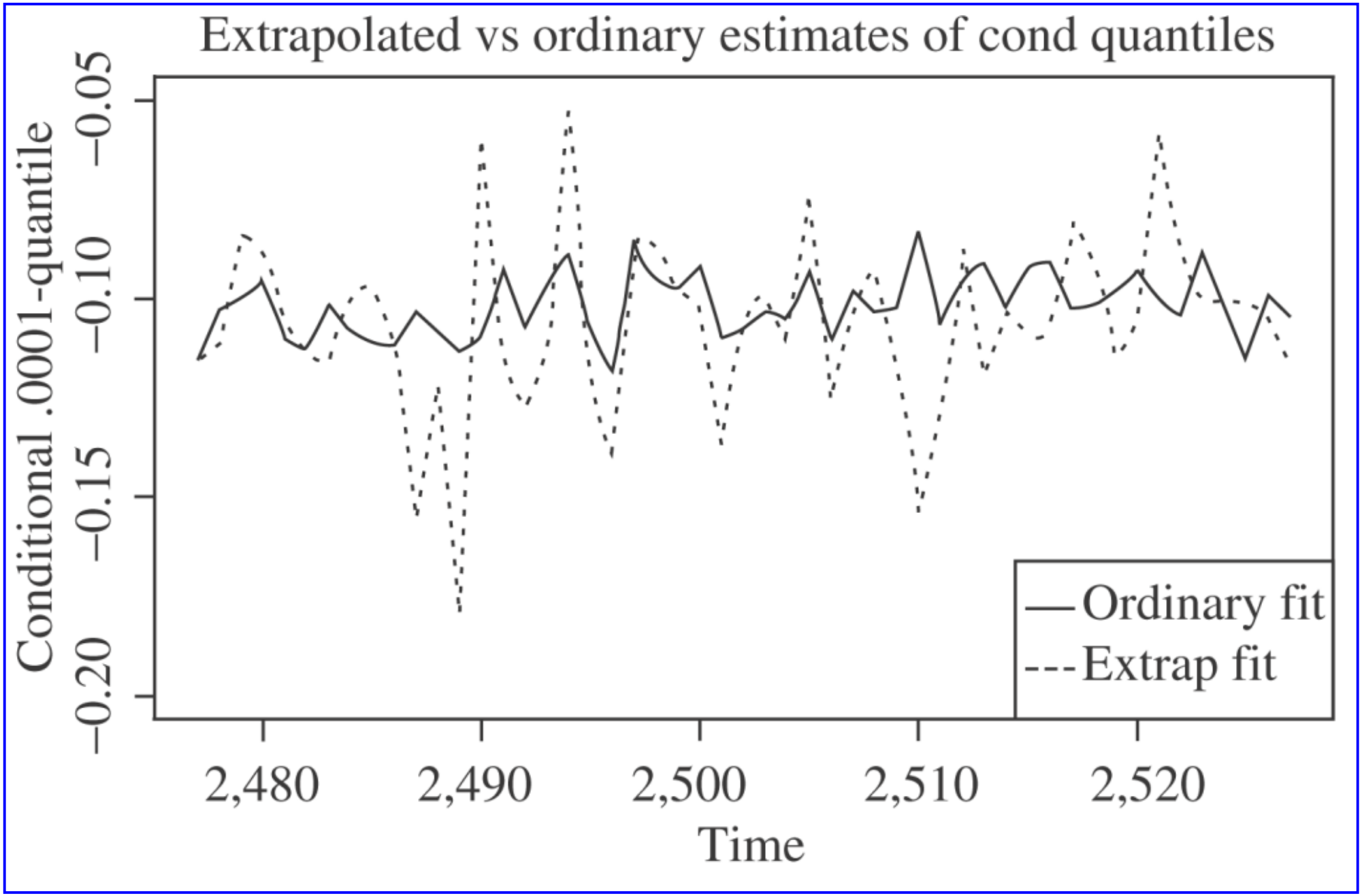

5 Conclusion

This article examines the theory and empirics of extremal quantiles in economics. The theory of extremes provides a set of applicable methods that have generated numerous valuable empirical findings. There is equally promising scope for the use of the extremal conditional quantile methods. The latter methods are new – there are great opportunities for further empirical and theoretical developments.

We thank Emily Gallagher, Greg Fischer and Raymond Guiteras for their help and valuable comments.

Bibliography

Abrevaya, J. 2001. The effect of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics* 26, 247–59.

Aigner, D.J., Amemiya, T. and Poirier, D.J. 1976. On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review* 17, 377–96.

Aigner, D.J. and Chu, S.F. 1968. On estimating the industry production function. *American Economic Review* 58, 826–39.

Arrow, K., Harris, T. and Marschak, J. 1951. Optimal inventory policy. *Econometrica* 19, 205–72.

Bickel, P. and Freedman, D. 1981. Some asymptotic theory for the bootstrap. *Annals of Statistics* 9, 1196–217.

Caballero, R. and Engel, E. 1999. Explaining investment dynamics in U.S. manufacturing: a generalized (S,s) approach. *Econometrica* 67, 783–826.

Chernozhukov, V. 2005. Extremal quantile regression. *Annals of Statistics* 33, 806–39.

Chernozhukov, V. 2006. Inference for extremal conditional quantile models, with an application to birthweights. Working paper, Department of Economics, Massachusetts Institute of Technology.

Chernozhukov, V. and Umantsev, L. 2001. Conditional value-at-risk: aspects of modeling and estimation. *Empirical Economics* 26, 271–93.

de Haan, L. 1970. On regular variation and its applications to the weak convergence. Tract No. 2, Mathematical Centre, Amsterdam.

Dekkers, A. and de Haan, L. 1989. On the estimation of the extreme-value index and large quantile estimation. *Annals of Statistics* 17, 1795–832.

Donald, S.G. and Paarsch, H.J. 2002. Superconsistent estimation and inference in structural econometric models using extreme order statistics. *Journal of Econometrics* 109, 305–40.

Embrechts, P., Klüppelberg, C. and Mikosch, T. 1997. *Modelling Extremal Events*, vol. 33 of *Applications of Mathematics*. Berlin: Springer.

Fama, E.F. 1965. The behavior of stock market prices. *Journal of Business* 38, 34–105.

Fox, M. and Rubin, H. 1964. Admissibility of quantile estimates of a single location parameter. *Annals of Mathematics and Statistics* 35, 1019–30.

Gnedenko, B. 1943. Sur la distribution limité du terme d’ une série alétoire. *Annals of Mathematics* 44, 423–53.

Hill, B.M. 1975. A simple general approach to inference about the tail of a distribution. *Annals of Statistics* 3, 1163–74.

Jansen, D.W. and de Vries, C.G. 1991. On the frequency of large stock returns: putting booms and busts into perspective. *Review of Economics and Statistics* 73, 18–24.

Knight, K. 2001. Limiting distributions of linear programming estimators. *Extremes* 4(2), 87–103.

Koenker, R. 2006. *Quantreg: Quantile Regression*. R package version 3.90. Online. Available at http://www.r-project.org, accessed 23 April 2007.

Koenker, R. and Bassett, G.S. 1978. Regression quantiles. *Econometrica* 46, 33–50.

Laplace, P.-S. 1818. *Théorie analytique des probabilités*. Paris: Éditions Jacques Gabay, 1995.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. 1983. *Extremes and Related Properties of Random Sequences and Processes*. New York and Berlin: Springer.

Longin, F.M. 1996. The asymptotic distribution of extreme stock market returns. *Journal of Business* 69, 383–408.

Mandelbrot, M. 1963. The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Meyer, R.M. 1973. A poisson-type limit theorem for mixing sequences of dependent ‘rare’ events. *Annals of Probability* 1, 480–3.

Pareto, V. 1964. *Cours d’économie politique*. Genéve: Droz.

Pickands, III, J. 1975. Statistical inference using extreme order statistics. *Annals of Statistics* 3, 119–31.

Portnoy, S. and Jure•ková, J. 1999. On extreme regression quantiles. *Extremes* 2, 227–43 (2000).

Praetz, V. 1972. The distribution of share price changes. *Journal of Business* 45, 49–55.

Quetelet, A. 1871. *Anthropométrie*. Brussels: Muquardt.

Roy, A.D. 1952. Safety first and the holding of assets. *Econometrica* 20, 431–49.

Sen, A. 1973. *On Economic Inequality*. New York: Oxford University Press.

Timmer, C.P. 1971. Using a probabilistic frontier production function to measure technical efficiency. *Journal of Political Economy* 79, 776–94.

Zipf, G. 1949. *Human Behavior and the Principle of Last Effort*. Cambridge, MA: Addison-Wesley.

How to cite this article

Chernozhukov, Victor and Songzi Du. "extremal quantiles and value-at-risk." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_V000065> doi:10.1057/9780230226203.0538

The New Palgrave Dictionary of Economics Online

extreme bounds analysis

Edward E. Leamer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Extreme bounds analysis is a global sensitivity analysis that applies to the choice of variables in a linear regression. Rather than a discrete search over models that include or exclude subsets of the variables, this sensitivity analysis answers the question: how extreme can the estimates be if any linear homogenous restrictions on a selected subset of the coefficients are allowed? When these bounds are too wide to be useful, narrower bounds can be found by restricting the set of prior distributions that underlie the sensitivity analysis.

Keywords

Bayesian econometrics; extreme bounds analysis; heteroscedasticity; nonparametric models and methods; probability; White-corrected standard errors

Article

The analysis of economic data necessarily depends on assumptions that our weak data-sets do not allow us to test. We are forced to choose a limited number of variables in a multivariate analysis, to restrict the functional form, to limit the considered interdependence among observations to special forms, and to make special distributional assumptions. We make these assumptions, not because we believe them, but because we have to. Absent assumptions, our data-sets are utterly useless.

We sometimes put aside the discomfort that our choice of assumptions entails by doing what is conventional, like using a normal distribution, a linear functional form and the same limited set of variables studied by almost everyone else.

We sometimes pretend to treat the problem of choice of assumptions by using ‘nonparametric’ methods that masquerade as assumption-free. These methods are assumption-free only in Asymptopia, the land where all data-sets are unlimited. To get to the happy land of Asymptopia, we need only let the number of tested assumptions grow in the future more slowly than the number of observations. Then we can be sure to test all assumptions during our journey into the future. But here on Earth we have limited data-sets, and inevitably the way we analyse these data works well for some sets of assumptions and not so well for others. We cannot really know what we are doing unless we can draw some kind of line between the assumptions for which our inferences are valid and the assumptions for which our inferences are not valid. Thus, for example, ‘heteroscedasticity-consistent’ standard errors are now commonly deployed as if they were corrections for any form of heteroscedasticity. But these corrections of the standard errors, which leave the point estimates unchanged, are an appropriate treatment given the actual limited data only for some forms of heteroscedasticity, not for all. Unfortunately, with these ‘nonparametric’ methods, the border between the dealt-with assumptions for which the method works and the not-dealt-with assumptions is impossible to draw. Incidentally, these ‘heteroscedasticity-consistent’ standard errors are often called White-corrected standard errors, to which I respond rhetorically by calling the method ‘White-washing’.

If conventions and nonparametric methods are not enough to soften the discomfort with the assumptions we make, we can always fantasize that the choice of assumptions doesn't really matter. We ‘know’ the methods we use work well under conditions that are ‘close’ to the assumptions that underlie them but not well if the departures are great. We hope that the neighbourhood in which the assumptions work is wide enough to encompass the problem at hand. For example, we don't really think the distribution is normal, but how much could that matter for linear regression? Isn't it enough to have symmetric unimodal distributions, shaped ‘sort of like’ a normal distribution?

Neither conventions, nor nonparametric sleight-of-hand, nor hope that it doesn't really matter form an adequate scientific response to doubt about the assumptions that underlie a data analysis. The correct way to deal with ambiguity in the choice of assumptions that are beyond the range of statistical tests is a sensitivity analysis that demonstrates that our assumptions do not in fact matter much. The most common sensitivity analysis involves the choice of variables in linear regression. Rather than reporting just one regression, many researchers offer a table of results, all based on different subsets of the variables. Typically, all the reported regressions have a common set of ‘core’ variables but differ depending on whether or not the regressions include selected ‘doubtful’ variables.

Although it can be comforting to discover that the coefficients of the core variables do not change much when doubtful variables are excluded, this kind of sensitivity analysis leaves open the possibility that there is some combination of doubtful variables that would radically change the result. Has the analyst worked hard enough to find the oddball estimates that these data allow? ‘Extreme bounds analysis’ answers this question. In an extreme bounds analysis, the computer chooses the linear combinations of doubtful variables that, when included in the regressions along with the core variables, produce the most extreme (minimum and maximum) estimates for the coefficient on a selected core variable. There is no way of fiddling with the doubtful variables that can produce an estimate outside the extreme bounds.

If the extreme bounds interval is small enough to be useful, that is the end of the story, and the result is reported to be ‘sturdy’. This would occur, for example, when the core variables and the doubtful variables are ‘independent’, in which case the coefficients of the core variables don't change at all when doubtful variables are excluded. But quite often with highly correlated economics data these extreme bounds can be uncomfortably wide, and we are forced either to retreat in dismay or to seek some way to make the bounds narrower.

One way of restricting the range of alternative models is to allow only inclusion/exclusion options, not the all linear combinations embodied in the extreme bounds. These inclusion/exclusion restrictions are the basis for the tables of alternative results that are commonly offered as evidence of inferential sturdiness, and the set of alternative estimates thus presented is smaller than the extreme bounds set.

But why? Why restrict to inclusion/exclusion options? Classical inference is not well suited to respond to this ‘why?’ question since the answer depends on the state of mind of the analyst and since classical inference presumes a researcher and an audience with a ‘blank slate’. Moreover, the effect on the inferences of setting a regression coefficient to zero depends on the coordinate system for defining the parameters. (If you use x and z as explanatory variables, and I use $x+z$ and $x-z$, we get different answers.) Indeed, the extreme bounds are formed by setting coefficients to zero in an appropriately defined coordinate system, and therefore restriction to inclusion/exclusion restrictions is a meaningless restriction absent advice on how to define the coordinate system.

A Bayesian analysis can help to choose a coordinate system and can be a basis for a sensitivity analysis with a set of models that is sensibly smaller than the set of models underlying the extreme bounds. Bayesians allow the state of mind to influence the analysis by letting a researcher act as if the vector of regression coefficients on the doubtful variables, θ, comes from a normal distribution with a mean vector $\mathbf{0}$ and covariance matrix $\mathbf{V}_0$, selected by the researcher. The smaller is the covariance matrix $\mathbf{V}_0$, the more likely are the coefficients θ to hug close to zero and the more doubtful are the doubtful variables.

To parallel the decision always to include the core variables in the equation, it is natural to deploy a prior probability distribution for the core coefficients with an infinite variance – the blank-slate initial-ignorance option. Then, corresponding to each choice of covariance matrix for the coefficients of the doubtful variables are estimates of the coefficients of the core variables, $\hat{\beta}(\mathbf{V}_0)$. If, for example, the covariance matrix $\mathbf{V}_0$ is set to zero, this is equivalent to assuming the coefficients of the doubtful variables are all zero, and we should be running the regression with all these doubtful variables omitted. Conversely, if the covariance matrix $\mathbf{V}_0$ is set to an ‘infinitely large’ matrix (unlimited variances), then this is the ignorance option for the doubtful

variables, and the way to do the estimation is simply to include all the doubtful variables along with the core variables in the regression.

A 'global' sensitivity analysis in this setting is carried out by building a correspondence between sets of prior covariance matrices for the doubtful variables, $\mathbf{V}_0$, and the corresponding sets of estimates of the coefficients on the core variables, $\hat{\beta}(\mathbf{V}_0)$. Three possibilities are discussed in Leamer (1978) and Leamer and Chamberlain (1976): (*a*) $\mathbf{V}_0$ unrestricted, (*b*) $\mathbf{V}_0$ diagonal, (*c*) $\mathbf{V}_0$ diagonal with all diagonal elements the same.

1. 1. The extreme bounds apply when the covariance $\mathbf{V}_0$ is any positive semi-definite matrix.
2. 2. The 2^p regressions, found by including different subsets of the p doubtful variables, define the bound when $\mathbf{V}_0$ is a (non-negative) *diagonal* matrix. (Thus the 'right' coordinate system is the one in which the regression coefficients are a priori independent of each other in the sense that knowledge about one doesn't affect your thinking about the others.)
3. 3. A still narrower set of bounds is found by estimating the p+1 'principal component' regressions with the principal component restrictions ordered by their eigenvalues. This applies when $\mathbf{V}_0$ is proportional to the identity matrix.

Each of these three sets of prior covariance matrices includes the dogmatic priors that set certain linear combinations of the coefficients exactly to zero and also complete ignorance priors that allow certain linear combinations to be completely free. Neither of these two extremes is sensible in practice, since in the first case the data evidence is completely ignored (the restriction is imposed without testing) and in the second case the prior information is completely ignored. These can be excluded to restricting the prior covariance matrix from above and below: $\mathbf{V}_L<\mathbf{V}_0<\mathbf{V}_U$ where $\mathbf{A}<\mathbf{B}$ means that the matrix $\mathbf{B}-\mathbf{A}$ is positive definite. The theorem that then applies comes from Leamer (1981, 1982) as is reported below. First, a statement about the posterior mean of the regression coefficient vector.

Theorem (Bayes estimate): : If, conditional on the observable matrix $\mathbf{X}$, and the unobservable parameters β, and σ^2, an observable vector $\mathbf{y}$ is normally distributed with mean $\mathbf{X}\beta$ and covariance matrix $\sigma^2\mathbf{I}$, and if the coefficient vector β comes from a normal distribution with mean $\mathbf{b}_0$ and covariance matrix $\mathbf{V}_0$, then the conditional mean of β given $\mathbf{y}$ is approximately

$$\mathbf{b}_2 = (\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_0^{-1})(\mathbf{X}'\mathbf{X}\mathbf{b}/s^2 + \mathbf{V}_0^{-1}\mathbf{b}_0)$$

where s^2 is the sample estimate of σ^2: $s^2 = \mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}/(n-k)$, where n is the number of observations and k is the number or regression coefficients and where $\mathbf{b}$ is the ordinary least squares estimator (a solution to the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$): $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Theorem (posterior bounds): : Given $\mathbf{V}_L<\mathbf{V}_0<\mathbf{V}_U$ with $\mathbf{V}_L$ and $\mathbf{V}_U$ positive definite and with $\mathbf{V}_L<\mathbf{V}_0$ signifying that $\mathbf{V}_0 - \mathbf{V}_L$ is positive definite, then the posterior mean $\mathbf{b}_2$ lies in the ellipsoid

$$(\mathbf{b}_2 - \mathbf{f})'\mathbf{H}(\mathbf{b}_2 - \mathbf{f}) < c$$

where

$$\mathbf{H} = (\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1}) + (\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1})(\mathbf{V}_L^{-1} - \mathbf{V}_U^{-1})^{-1}(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1})\mathbf{f} = \left[(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_L^{-1})\right]^{-1}\left[\mathbf{X}'\mathbf{X}\mathbf{b}/s^2 + (\mathbf{V}_L^{-1} - \mathbf{V}_U^{-1})(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}/2s^2\right]c = \left(\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/s^2\right)(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1})^{-1}(\mathbf{V}_L^{-1} - \mathbf{V}_U^{-1})(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_L^{-1})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}/4s^2$$

See Also

- Bayesian econometrics
- data mining

Bibliography

Leamer, E. 1978. *Specification Searches: Ad Hoc Inference with Non Experimental Data*. New York: John Wiley and Sons.

Leamer, E. 1981. Sets of estimates of location. *Econometrica* 49, 193–204.

Leamer, E. 1982. Sets of posterior means with bounded variance priors. *Econometrica* 50, 725–36.

Leamer, E. and Chamberlain, G. 1976. Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society*, Series B 38, 73–84.

How to cite this article

Leamer, Edward E. "extreme bounds analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000250> doi:10.1057/9780230226203.0539

The New Palgrave Dictionary of Economics Online

extreme poverty

Jeffrey D. Sachs and Gordon C. McCord

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Households living in extreme poverty face deprivations that cost millions of lives annually. Ending extreme poverty requires an understanding of poverty traps, including the effects of adverse biophysical and geographical factors, a lack of resources required for the investments needed to escape poverty, and poor governance. Policies must focus both on promoting market-oriented economic growth and on directly addressing the needs of the poor. Foreign aid will be required to finance interventions that poor countries cannot finance themselves, and aid to well-governed poor countries should be increased, consistent with the rich-country promise of 0.7 per cent of GNP as official development assistance.

Keywords

business capital; capabilities; corruption; democracy; development assistance; diffusion of technology; economic growth; education; extreme poverty; family planning; fertility; food security; foreign aid; geography and economic development; governance; health care; human capital; Human Development Index; human rights approach to poverty alleviation; infant mortality; innovation; knowledge capital; Millennium Development Goals; moderate poverty; natural capital; nutrition; patents; poverty; poverty alleviation; poverty lines; poverty traps; public goods; public infrastructure; public institutional capital; relative poverty; research and development; Sen, A.; state failure; sub-Saharan Africa; subsistence; Washington Consensus; water stress

Article

There are many definitions of poverty, as well as intense debates about the exact numbers of the poor, where they live and how their numbers are changing over time. As a matter of definition, it is useful to distinguish between three degrees of poverty. Extreme (or absolute) poverty, moderate poverty and relative poverty. extreme poverty can be thought of as ‘poverty that kills’, meaning that households cannot reliably meet basic needs for survival. Households living in extreme poverty are chronically

undernourished, unable to access health care, lacking the amenities of safe drinking water and sanitation, unable to afford education for some or all of the children, and perhaps lacking rudimentary shelter – a roof to keep the rain out of the hut, a chimney to remove the smoke from the cooking stove – and basic articles of clothing such as shoes. Such deprivations cost lives, by the millions, every year. Life expectancy is considerably lower and mortality rates are considerably higher in countries in which large proportions of the population live in extreme poverty.

Unlike moderate and relative poverty, extreme poverty currently occurs only in developing countries. Moderate poverty generally refers to the conditions of life in which basic needs are met, but only barely. Relative poverty is generally construed as a household income level below a given proportion of average national income. The relatively poor, in high-income countries, lack access to cultural goods, entertainment, recreation, and to quality health care, education, and other perquisites of social mobility. They may also live outside of the 'mainstream' of social life, and thus without dignity and social respect.

In order to estimate the number of extreme poor, most analysts use a poverty line – a level of income below which the person is 'extremely poor' by some definition. Most countries set their own poverty lines, based on the per capita cost of a consumption basket that attempts to measure basic needs. Since the poorest people in poor countries spend most of their money on food, most of the basket used for national poverty lines consists of food, usually in terms of meeting a minimum intake of 2,000 calories (Deaton, 2004). These poverty lines are surely imperfect: they suffer from the measurement error inherent in household surveys; they are rarely updated with regards to spending on nutrition; they do not account for differences in rural versus urban calorie consumption; and they do not capture all dimensions of extreme poverty (for example, access to health care, safe water, sanitation, education or political voice).

Moreover, they can lead to undesirable policy results (a person just below the poverty lines could be treated very differently from someone just above the line, despite having almost equal incomes). Governments judged solely according to the number of people below the poverty line could choose to focus only on those closest to the line and ignore the poorest of the poor. Finally, as Nobel Laureate Amartya Sen has emphasized, poverty should be defined more broadly than having a low income; rather, it is the absence of basic capabilities to function in society. This could include not only income poverty (involving a lack of food, clothing, or shelter), but also lack of access to public goods, social standing, and political participation. Despite these shortcomings, most dimensions of extreme poverty that people would like to improve are correlated with household income, thus making a poverty line a helpful, though rough, first approximation of poverty rates. Measures that combined household income with provisions of public goods (disease control, public health, primary education) would surely be preferable.

In the late 1980s, and especially with the 1990 *World Development Report*, the World Bank introduced a single measure of extreme poverty – an income of one dollar per day or less (in 1985 purchasing power parity, PPP, dollars) – in order to compare rates of extreme poverty across countries and to track extreme poverty over time. The one dollar per day number was chosen since it corresponds roughly to the highest national poverty rate among low-income countries (around 360 dollars per year). In 2000, the World Bank used improved PPP estimates to adjust its global poverty line to 1.08 dollars per person per day (in 1993 PPP dollars). This global extreme poverty line has been criticized by some for not being high enough and thus undervaluing the needs of the poor (Pritchett, 2003) and by others for being too arbitrary and detached from the country-specific needs of the poor (Srinivasan, 2004). Nevertheless, it provides a useful, albeit highly imperfect, measuring tool to look at extreme poverty around the world.

Another important indicator for poverty is the Human Development Index (HDI), published by the United Nations Development Programme (UNDP) since 1990. The UNDP sought to incorporate the multidimensional aspects of poverty into a new indicator, and to emphasize that development should expand human capabilities, particularly those that are universally valued and basic to life: the capability to lead a long and healthy life, to be knowledgeable, and to have access to the resources needed for a decent standard of living (UNDP, 2004). The result was the HDI, which averages normalized 0–1 indexes for income per capita, life expectancy, and education school enrolment and literacy. Countries classified as ‘low human development’ have a very strong overlap with those countries that have a high proportion of the population living under one dollar per day according to the World Bank.

Where are the poor?

The most recent estimates of extreme poverty around the world (using the one dollar per day estimate) were made by Shaohua Chen and Martin Ravallion at the World Bank (see Table 1). They estimated that roughly 1.1 billion people were living in extreme poverty in 2001, down from 1.5 billion in 1981 (Chen and Ravallion, 2004). The overwhelming share of the world's extreme poor, 93 per cent in 2001, live in three regions, East Asia, South Asia and Sub-Saharan Africa. Since 1981, the absolute numbers of extreme poor have risen in Sub-Saharan Africa, but have fallen in East Asia and South Asia. In terms of proportions, nearly half Africa's population is judged to live in extreme poverty, and that proportion has risen slightly over the period. The proportion of the extreme poor in East Asia has plummeted, from 58 per cent in 1981 to 15 per cent in 2001; in South Asia the progress has also been marked, although slightly less dramatically, from 52 per cent to 31 per cent. Latin America's extreme poverty rate is around ten per cent, and relatively unchanged; Eastern Europe's rose from a negligible level in 1981 to around four per cent in 2001, the results of the upheavals of Communist collapse and economic transition to a market economy. It is worth noting that these numbers are debated heatedly; other researchers have relied on national income accounts, which tend to show somewhat faster progress in the reduction of Asian poverty, and sometimes very different estimates for the total amount of people living in extreme poverty (Sala-i-Martin, 2002; Bhalla, 2002). The general picture, however, remains true in all these studies: extreme poverty is concentrated in East Asia, South Asia and Sub-Saharan Africa. It is rising in Africa in absolute numbers and as a share of the population, while it is falling both in absolute numbers and as a proportion of the population in the Asian regions.

Number of poor people by region, 1981–2001

	$ 1.08 per day (million)							
	1981	1984	1987	1990	1993	1996	1999	2001
East Asia	795.6	562.2	425.6	472.2	415.4	286.7	281.7	271.3
•Of which China	633.7	425.0	308.4	374.8	334.2	211.6	222.8	211.6
Eastern Europe and Central Asia	3.1	2.4	1.7	2.3	17.4	19.8	29.8	17.6
Latin America and Caribbean	35.6	46.0	45.1	49.3	52.0	52.2	53.6	49.8
Middle East and North Africa	9.1	7.6	6.9	5.5	4.0	5.5	7.7	7.1
South Asia	474.8	460.3	473.3	462.3	476.2	461.3	428.5	431.1

•Of which India	382.4	373.5	369.8	357.4	380.0	399.5	352.4	358.6
Sub-Saharan Africa	163.6	198.3	218.6	226.8	242.3	271.4	294.0	315.8
Total	**1481.8**	**1276.8**	**1171.2**	**1218.5**	**1207.5**	**1096.9**	**1095.1**	**1092.7**

Source: Chen and Ravallion (2004, p. 153).

There are some defining circumstances specific to the poorest of the poor. They are found mainly in rural areas (though with a growing proportion in the cities); the rural poor tend to have fewer opportunities to earn income, have less access to education and health care, and are often more vulnerable to the forces of nature. The extreme poor face challenges almost unknown in the rich world today – malaria, famines, lack of roads and motor vehicles, great distances to regional and world markets, lack of electricity and modern cooking fuels. Women tend to be at a disadvantage compared with men, since they often have less access to property rights (land ownership, inheritance), and since they bear the physical burden of lack of infrastructure (collecting water and fuel wood at great distances). Girls have historically received less primary and secondary education than boys. Labour markets often discriminate against women, and women tend to work longer when one counts unpaid labour at home. Domestic violence continues to burden the lives of millions of women around the world (World Bank, 2001). Finally, large pockets of poverty exist within many countries due to racial and ethnic discrimination, or low social (for example, caste) status.

Consequences of extreme poverty

When individuals suffer from extreme poverty and lack the meagre income needed to cover even basic needs, a single episode of disease, a drought, or a pest that destroys a harvest can be the difference between life and death. In households suffering from extreme poverty, life expectancy is often around half that in the high-income world, 40 years instead of 80 years. It is common that, in the poorest countries of Sub-Saharan Africa, of every 1,000 children born more than 100 die before their fifth birthday, compared with fewer than ten in the high-income world. An infant born in Sub-Saharan Africa today has only a one-in-three chance of surviving to age 65.

At the most basic level, the poorest of the poor lack the minimum amount of capital necessary to get a foothold on the first rung of the ladder of economic development. The extreme poor tend to lack six major kinds of capital:

- Human capital: health, nutrition, and skills – education – needed for each person to be economically productive.
- Business capital: the machinery, facilities, motorized transport used in agriculture, industry and services.
- Infrastructure: roads, power, water and sanitation, airport and seaports, and telecommunications systems, which are critical inputs into business productivity.
- Natural capital: arable land, healthy soils, biodiversity, and well-functioning ecosystems that provide the environmental services needed by human society.
- Public institutional capital: commercial law, judicial systems, government services and policing that underpin the peaceful and prosperous division of labour.

- Knowledge capital: the scientific and technological know-how that raises productivity in business output and the promotion of physical and natural capital.

Importantly, the poorest of the poor tend to have higher fertility rates, for several reasons. Infant mortality rates are high when there are inadequate health services, so high fertility provides ‘insurance’ to parents that they will succeed in raising a child who will survive to adulthood. In rural areas, children are often perceived as economic assets who provide supplementary labour for the farm household. Poor and illiterate women have few job opportunities away from the farm, and so may place a low value on the opportunity (time) costs of bringing up children. In addition, women are frequently unaware of their reproductive rights (including the right to plan their families) and lack access to reproductive health information, services, and facilities, leading to high unmet demands for contraception in low-income countries and among poorer members of all developing countries. Finally, poor households lack the income to purchase contraceptives and family planning, even when they are available. For these reasons, high fertility rates are prevalent among families living in extreme poverty, resulting in very low investments in the health and education of each child (what is known as the quantity–quality trade-off). Poor and hungry societies are much more likely than high-income societies to fall into violent conflicts over scarce vital resources, such as watering holes and arable land – and over scarce natural resources, such as oil, diamonds and timber (United Nations, 2004). This relationship between violence and high rates of extreme poverty holds with a high degree of statistical significance. A country with a civil war within its borders typically has only one-third of the per capita income of a country with similar characteristics but at peace. Moreover, poor countries – even those not in conflict – risk conflict in the future. A country with a per capita income of 500 dollars is about twice as likely to have a major conflict within five years as a country with an income of about 4,000 dollars per capita (UN Millennium Project, 2005). In addition, low economic growth rates are associated with higher risks of new conflict; one study finds that a negative growth shock of five per cent increases the risk of civil war by 50 per cent in the following year, and that economic conditions are probably the most important determinants of civil conflict in Sub-Saharan Africa (Miguel, Satyanath and Sergenti, 2004). The most comprehensive study of state failure, carried out by the State Failure Task Force established by the Central Intelligence Agency in 1994, confirms the importance of the economic roots of state failure (defined as revolutionary war, ethnic war, genocide, politicide, or adverse or disruptive regime change). The Task Force studied all 113 cases of state failure between 1957 and 1994 in countries of half a million people or more, and found that the most significant variables explaining these conflicts were the infant mortality rate (suggesting that overall low levels of material well-being are a significant contributor to state failure), openness of the economy (more economic linkages with the rest of the world diminish the chances of state failure), and democracy (democratic countries show less propensity to state failure than authoritarian regimes). The linkage to democracy also has a strong economic dimension, however, because research has shown repeatedly that the probability of a country's being democratic rises significantly with its per capita income level. In refinements of the basic study, the Task Force found that in Sub-Saharan Africa, where many societies live on the edge of subsistence, temporary economic setbacks (measured as a decline in gross domestic product per capita) were significant predictors of state failure (State Failure Task Force, 1999). Similar conclusions have been reached in studies on African conflict, which find that poverty and slow economic growth raise the probability of conflict.

Theories of extreme poverty

For decades, observers have tried to explain why extreme poverty persists. Many theories have looked for single-factor explanations for a lack of economic growth, often grounded in racist beliefs (poor countries do not grow because their cultures, races, or religions fail to promote economic growth). The increasing number of success stories of growth proved all these theories to be wrong. However, despite the complexity of an economy and the number of things that can go wrong, single-factor explanations persist. The most common is that poverty is a result of corrupt leadership, which impedes modern development.

Governance is indeed important: economic development stalls when governments do not uphold the rule of law, pursue sound economic policy, make appropriate public investments, manage a public administration, protect basic human rights, and support civil society organizations – including those representing poor people – in national decision-making. Importantly, long-term poverty reduction in developing countries will not happen without sustained economic growth, which requires a vibrant private sector. Government, therefore, needs to provide the economic policy framework and the support that the private sector needs to grow.

However, many well-governed poor countries may be too poor to help themselves out of extreme poverty. Many well-intentioned governments lack the fiscal resources to invest in infrastructure, social services, environmental management, and even the public administration necessary to improve governance. Further, dozens of heavily indebted poor and middle-income countries have been forced by creditor governments to spend large proportions of their limited tax receipts on debt service, undermining their ability to finance vital investments in human capital and infrastructure. The reason these poor countries cannot grow is not poor governance, but a poverty trap. They lack the basic infrastructure, human capital, and public administration – the foundations for economic development and private sector-led growth. Without roads, soil nutrients, electricity, safe cooking fuels, clinics, schools, and adequate and affordable shelter, people are chronically hungry, burdened by disease and unable to save. As mentioned above, fertility rates tend to be high, preventing families from investing enough in each child. Without adequate public sector salaries and information technologies, public management is chronically weak. For all of these interlocking reasons, these countries are then unable to attract private investment flows or retain their skilled workers, and can therefore find themselves with low or negative growth. In short, they are stuck in a poverty trap.

The concept of a low-level poverty trap is a long-standing hypothesis in the theories of economic growth and development. The earliest mathematical formalization was by Nelson (1956), who put emphasis on demography. The theoretical possibility of poverty traps in the neoclassical growth model is covered briefly in the economic growth textbook by Barro and Sala-i-Martin (1998), which also discusses briefly the possible case for large-scale development assistance to overcome such traps. The connection of a low-level trap to subsistence consumption needs is spelled out in Ben-David (1998), and connections to agriculture and education are described in the *World Economic and Social Survey* (UN, 2000). Two recent empirical studies claiming that such poverty traps exist in poor countries are UNCTAD (2002) and Bloom, Canning and Sevilla (2003). A close look at a poverty trap in Sub-Saharan Africa is in Sachs et al. (2004).

An often overlooked characteristic of poverty is that some countries and regions are clearly more

vulnerable than others to falling into a poverty trap. While a history of violence of colonial rule or poor governance can leave any country bereft of basic infrastructure and human capital, physical geography plays special havoc with certain regions. Some regions need more basic infrastructure than others simply to compensate for a difficult physical environment. Some of the barriers that must be offset by investments include adverse transport conditions (landlocked economies, small island economies far from major markets, inland populations far from coasts and navigable rivers, populations living in mountains, long distances from major world markets, very low population densities); adverse agro-climatic conditions (low and highly variable rainfall, lack of suitable conditions for irrigation, nutrient-poor and nutrient-depleted soils, vulnerability to pests and other post-harvest losses, susceptibility to the effects of climate change); adverse health conditions (high ecological vulnerability to malaria and other tropical diseases, high AIDS prevalence); and other adverse conditions (lack of domestic energy sources, small internal market and lack of regional integration, vulnerability to natural hazards, artificial borders that cut across cultural and ethnic groups, proximity to countries in conflict). Adam Smith was acutely aware of the role of geography in hindering economic development. He stressed, in particular, the advantages of proximity to low-cost, sea-based trade as critical, noting that remote economies would be the last regions to achieve economic development. More recent studies have found statistical significance of these relationships between geography and economic outcome (Gallup, Sachs and Mellinger, 1999; Mellinger, Sachs and Gallup, 2000; Sachs and Gallup, 2001).

In the rich countries of North America, Western Europe and East Asia, the process of massive investment in research and development, leading to sales of patent-protected products to a large market, stands at the core of economic growth. Advanced countries are typically investing two per cent or more of their gross national product directly into the research and development process, and sometimes more than three per cent. That investment is very sizeable, with hundreds of billions of dollars invested each year in research and development activities. Moreover, these investments are not simply left to the market. Governments invest heavily, especially in the early stages of R&D. In most poor countries, especially smaller ones, the innovation process usually never gets started. Potential inventors do not invent because they know that they will not be able to recoup the large, fixed costs of developing a new product. Impoverished governments cannot afford to back the basic sciences in government laboratories and in universities. The result is an inequality of innovative activity that magnifies the inequality of global incomes. While the innovation gap is reduced in the case of some poor countries through technological diffusion, even diffusion is limited in the poorest countries, because they face distinctive ecological problems not addressed by ‘rich-world science’ (for example, tropical diseases and tropical farming systems), because they cannot afford high-tech capital goods and because they fail to attract foreign businesses that would bring the technology with them.

Policy responses

Theories on how to tackle extreme poverty are varied and controversial. For the most part, they can be divided into two camps: strategies that focus on promoting market-oriented economic growth, and strategies that focus on directly addressing the needs of the poor. Of course the two approaches can be combined. The Washington Consensus, a set of policy recommendations especially prevalent from 1980 to the late 1990s, embodies the first type, with its focus on macroeconomic stability, greater economic openness to trade and investment, and improved environment for private business. The idea was that

these policies would lead growth of the private sector, thus increasing demand for labour and thereby improving the welfare of the poor.

A second set of strategies focuses instead on providing what the poor need in order to increase their productivity. These investments in ‘human development’ argued for directing health and education investments towards the poor, and providing social safety nets. Many of these strategies became popular in the 1990s as a reaction to the Washington Consensus. There were three kinds of critiques. One held that growth would not be achieved with market reforms alone, because of the poverty trap. A second held that growth must in any event be combined with increased public investments, for example for health and education. A third, and more extreme position, held that growth per se would have adverse effects on the poorest of the poor. For example, the 1996 *Human Development Report* warned that in some cases growth can fail to create jobs and provide benefits, and can even increase empowerment of the rich, wreck cultural identities and destroy the environment (UNDP, 1996).

Numerous studies have shown that growth *does*, in fact, tend to be good for the poor (Dollar and Kraay, 2002; Roemer and Gugerty, 1997; Gallup, Radelet and Warner, 1999). Yet growth may not be achievable for countries trapped in poverty, and growth may not be sufficient to enable the poorest of the poor to meet their basic needs. The emerging consensus is that *both* economic growth and direct investments for the poor are necessary, in order to break the poverty trap and to provide vital public goods. International institutions are paying more attention than before to the possibility of poverty traps, and to the non-income dimensions of extreme poverty (for example, health and education). Five of the eight Millennium Development Goals (the world's time-bound and quantified targets for addressing extreme poverty, discussed below) are about promoting health and education, and individual countries are giving more priority to these broader measures than ever before.

Another dimension of the fight against extreme poverty is referred to as the rights-based approach. The guarantee that all people can live in dignity and meet their basic needs is also a basic human right – the right of each person on the planet to health, education, shelter and security as pledged in the Universal Declaration of Human Rights and various UN covenants, treaties and inter-governmental documents (such as the UN Millennium Declaration). The human rights approach seeks to use national and international human rights accountability mechanisms to monitor action on behalf of a human right rather than a development target. Economic evaluations often measure whether a given policy action contributes to reaching a target. Conceived in terms of rights, the same evaluation would measure not only those reached by a given action, but several other considerations as well: (*a*) the numbers not being reached; (*b*) the empowerment of the poor to achieve their rights; (*c*) the protection of these rights in legislation; and so forth. To date, there has been insufficient effort to integrate development planning with a human rights framework, even though such integration has tremendous potential and relevance.

Since the creation of the United Nations in 1945, the international system has been working to reduce poverty around the world, but often with results that fall short of laudable rhetoric. In January 1961, the United Nations resolved that the decade of the 1960s would be the Decade of Development. US President Kennedy launched the decade at the UN in New York. Earlier, in his inaugural address as President, he had signalled a new sense of purpose in international affairs. He declared: ‘To those peoples in the huts and villages of half the globe struggling to break the bonds of mass misery, we pledge our best efforts to help them help themselves’ (History Place, 2007). The second Development Decade resolved to emphasize measures deliberately targeted at the poor – to help them meet their basic needs for food, water, housing, health and education. The UN held a series of international conferences:

on environment (Stockholm, 1972); population (Bucharest, 1974); food (Rome, 1974); women (Mexico City, 1975); human settlements (Vancouver, 1976); employment (Geneva, 1976); water (Mar del Plata, 1977); and desertification (Nairobi, 1977). In 1978, the governments of the world came together to sign the Alma Ata Declaration that promised 'Health for All by 2000', a promise the world failed miserably in delivering. The 1980s – the third Development Decade – were very difficult for developing countries as they suffered from a worldwide recession that hit the developing world and debtor countries with special force. Nevertheless, important improvements were made in some areas, such as nutrition, access to safe drinking water, and reductions in child mortality. One result was the international conference held in 1990 under the auspices of UNDP, UNESCO, UNICEF and the World Bank in Jomtien (Thailand), which set the target of 'Education for All by the Year 2000', another goal not met.

The 1990s also became a decade in which the response of the UN system to the flagging development movement was to embark on a series of global conferences. The UN Conference on Environment and Development (Rio de Janeiro, 1992) was followed by conferences on nutrition (Rome, 1992); human rights (Vienna, 1993); population and development (Cairo, 1994); social development (Copenhagen, 1995); women (Beijing, 1995), human settlements (Istanbul, 1996). The decade ended with the landmark Millennium Summit in 2000, which resulted in the Millennium Development Goals (MDGs), and the Financing for Development Conference in Monterrey in 2002, where rich countries renewed their pledge to provide 0.7 per cent of their GDP in foreign aid. Also relevant was the Brussels Programme of Action for the Least Developed Countries, which suggests that they require greatly increased official development assistance, since private capital flows will not finance needed public investments. The programme outlines several priority areas for cooperation including human and institutional resource development, removing supply side constraints and enhancing productive capacity, protecting the environment, and attaining food security and reducing malnutrition.

As the UN Millennium Project has pointed out, the Millennium Development Goals are the most broadly supported, comprehensive, and specific poverty reduction targets the world has ever established, so their importance is manifold. For the international political system, they are the fulcrum on which development policy is currently based. For the billion-plus people living in extreme poverty, they represent the means to a productive life (UN Millennium Project, 2005). Besides aiming to reduce the 1990 proportion of people in extreme poverty by half by 2015, the MDGs tackle poverty in its many dimensions – income poverty, hunger, disease, lack of adequate shelter, and exclusion – while promoting gender equality, education and environmental sustainability. Thus, while supporting the need for economic growth, the MDGs emphasize that the growth needs to be pro-poor. In 2005, the UN Millennium Project presented the Secretary General with 'A Practical Plan to Achieve the Millennium Development Goals', which outlined specific interventions to address the multiple causes of poverty traps in poor countries around the world (UN Millennium Project, 2005). Moreover, it emphasized that foreign aid will be needed to finance the interventions that the poor countries cannot finance themselves. In the case of well-governed poor countries, the report recommended that foreign assistance should be scaled up immediately, significantly, and on a sustained basis, consistent with the promise of 0.7 per cent of GNP as official development assistance.

Prospects

There are reasons to be optimistic about the elimination of extreme poverty on the planet. Economic

development has lifted more than 100 million people out of extreme poverty since the mid-1990s, and the pace is probably accelerating in Asia. While the population of developing countries rose from about four billion people to about five billion, average per capita incomes rose by more than 21 per cent. With 130 million fewer people in extreme poverty in 2001 than a decade before, the proportion of people living on less than one dollar a day declined by seven percentage points, from 28 to 21 per cent.

Despite the good news, however, Africa remains mired in seemingly intractable extreme poverty. Africa faces difficult structural challenges (very high transport costs and small markets, low-productivity agriculture, very high disease burden, a history of adverse geopolitics, and slow diffusion of technology from abroad), but, in countries where governments are committed, these challenges can be overcome if addressed through an intensive programme that directly confronts them (Sachs et al., 2004). Ending the poverty trap in Africa and meeting the MDGs will require a comprehensive strategy for public investment in conjunction with improved governance. The good news is that the amount of investment required, although out of reach of African governments alone, is within the amount already promised in foreign aid by the rich countries (UN Millennium Project, 2005).

One final point is that a sustained reduction in extreme poverty requires tackling long-term challenges that the human family faces, in particular environmental challenges. Raising the incomes of billions of people around the world is surely desirable. Nevertheless, the increased income will come with increased demand for food, energy, and consumer goods, which may push our planet's already stressed ecosystems beyond what they can support. As the world works towards eliminating extreme poverty, it must do so with a conscious plan to limit the environmental burden that humanity places on the planet. Moreover, in many cases, the environmental challenges (such as water stress) may prove to be the biggest barriers to poverty reduction even in the short term.

See Also

- foreign aid
- poverty
- poverty alleviation programmes
- poverty lines
- poverty traps
- Sen, Amartya

Bibliography

Barro, R. and Sala-i-Martin, X. 1998. *Economic Growth*. Cambridge, MA: MIT Press.

Ben-David, D. 1998. Convergence clubs and subsistence economies. *Journal of Development Economics* 55, 155–71.

Bhalla, S. 2002. *Imagine There's No Country – Poverty, Inequality and Growth in the Era of Globalization*. Washington, DC: Institute for International Economics.

Bloom, D., Canning, D. and Sevilla, J. 2003. Geography and poverty traps. *Journal of Economic Growth* 8, 355–78.

Chen, S. and Ravallion, M. 2004. How have the world's poorest fared since the early 1980s? *World Bank Research Observer* 19, 141–70.

Deaton, A. 2004. Measuring poverty. Working paper, Research Program in Development Studies, Princeton University.

Dollar, D. and Kraay, A. 2002. Growth is good for the poor. *Journal of Economic Growth* 7, 195–225.

Gallup, J., Radelet, S. and Warner, A. 1999. Economic growth and the income of the poor. CAER II Discussion Paper No. 36, Harvard Institute for International Development.

Gallup, J.L., Sachs, J.D. and Mellinger, A.D. 1999. Geography and economic development. *International Regional Science Review* 22, 179–232.

History Place. 2007. Great Speeches Collection: John F. Kennedy Inaugural Address. Online. Available at http://www.historyplace.com/speeches/jfk-inaug.htm. Accessed 25 April 2007.

Mellinger, A., Sachs, J.D. and Gallup, J. 2000. Climate, coastal proximity, and development. In *Oxford Handbook of Economic Geography*, ed. G.L. Clark, M.P. Feldman and M.S. Gertler. New York: Oxford University Press.

Miguel, E., Satyanath, S. and Sergenti, E. 2004. Economic shocks and civil conflict: an instrumental variables approach. *Journal of Political Economy* 112, 725–53.

Nelson, R. 1956. A theory of the low-level equilibrium trap in underdeveloped economies. *American Economic Review* 46, 894–908.

Perkins, D.H., Radelet, S. and Lindauer, D.L. 2006. *Economics of Development*, 6th edn. New York: W. W. Norton.

Pritchett, L. 2003. Who is *not* poor? Proposing a higher international standard for poverty. Working Paper No. 33, Center for Global Development.

Roemer, M. and Gugerty, M.K. 1997. Does economic growth reduce poverty? CAER II Discussion Paper No. 5, Harvard Institute for International Development.

Sachs, J.D. 2005. *The End of Poverty*. New York: Penguin.

Sachs, J.D. and Gallup, J.L. 2001. The economic burden of malaria. *American Journal of Tropical Medicine and Hygiene* 64(supplement), 85–96.

Sachs, J.D., McArthur, J.W., Schmidt-Traub, G., Kruk, M., Bahadur, C., Faye, M. and McCord, G. 2004. Ending Africa's poverty trap. *Brookings Papers on Economic Activity* 2004(1), 117–240.

Sala-i-Martin, X. 2002. The world distribution of income. Working Paper No. 8933. Cambridge, MA: NBER.

Srinivasan, T.N. 2004. The unsatisfactory state of global poverty estimation. In *In Focus: Dollar a Day: How Much Does it Say?* Brasilia: International Poverty Centre, UNDP.

State Failure Task Force. 1999. *Environmental Change and Security Project Report*, vol. 5. Washington, DC: Woodrow Wilson Center.

UN (United Nations). 2000. *World Economic and Social Survey 2000*. New York: United Nations.

UN. 2004. *A More Secure World: Our Shared Responsibility. Report of the Secretary-General's High-Level Panel on Threats, Challenges and Change*. New York: United Nations.

UNCTAD (United Nations Conference on Trade and Development). 2002. *The Least Developed Countries Report 2002: Escaping the Poverty Trap*. Geneva: UNCTAD.

UNDP (United Nations Development Programme). 1996. *Human Development Report 1996*. New York: UNDP/Oxford University Press.

UNDP. 2004. *Human Development Report 2004: Cultural Liberty in Today's Diverse World*. New York: UNDP/Oxford University Press.

UN Millennium Project. 2005. *Investing in Development: A Practical Plan to Achieve the Millennium Development Goals*. New York: Earthscan.

World Bank. 1990. *World Development Report 1990: Poverty*. Washington, DC: World Bank.

World Bank. 2001. *Engendering Development*. New York: Oxford University Press.

How to cite this article

Sachs, Jeffrey D. and Gordon C. McCord. "extreme poverty." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008

<http://www.dictionaryofeconomics.com/article?id=pde2008_E000293>
doi:10.1057/9780230226203.0540

The New Palgrave Dictionary of Economics Online

Fabian economics

Elizabeth Durbin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N.Durlauf and LawrenceE.Blume

Keywords

Beveridge, W. H.; capitalism; Clark, C. G.; Cole, G. D. H.; Dalton, H.; Durbin, E.F. M.; equality; exchange value; Fabian economics; Gaitskell, H. T. N.; imperfect competition; individual freedom; inequality; inheritance; Keynesian Revolution; Lewis, W. A.; marginal productivity theory; marginal revolution; market socialism; Meade, J. E.; mixed economy; monopoly capitalism; nationalization; planning; property; public ownership; redistribution of income and wealth; Shaw, G. B.; social democracy; socialism; Tawney, R. H.; trade cycle; Webb, B. and S.

Article

Despite the Webbs' disdain for abstract economics ('sheer waste of time'), economic arguments have always held a central place in the Fabian case for socialism. As in most matters the Fabian Society has approached the dismal science eclectically. Some members have accepted market economics, others have rejected it; some embraced the Keynesian Revolution, others remained sceptical; some have believed in market pricing, others have been convinced that controls are essential for centralized planning. There is no consistent body of thought which could properly be described as Fabian economics. There is nonetheless a distinctive Fabian approach to economics, which this essay identifies while tracing the significant shifts in its key elements.

The Fabians and the marginal revolution

When the small group including Sydney Olivier, Bernard Shaw and Sidney Webb first started to meet at Mrs Charlotte Wilson's house in Hampstead, they set themselves the task of reading Marx's *Das Kapital* chapter by chapter. Graham Wallas, who joined the group in February 1885, later recalled how they were astonished to find 'that we did not believe in Karl Marx at all' (Wallas, 1923). Webb, Wallas and Shaw were also members of the Economic Circle, an offshoot of the Bedford Chapel Debating Society, where Professor Edgeworth helped to expound the principles of the new marginal economics with

another economist, Philip Wicksteed. Thus, according to Wallas, under Webb's leadership the group thrashed out ‘the Jevonian anti-Marx value theory as the basis of our socialism’. Shaw apparently needed more convincing than others. In the *Fabian Essays* he later described how he had been converted from his earlier Marxist faith that the working class revolution would take place in Britain by 1889 ‘at latest’ (Shaw, 1908, pp. 218–19). Instead of manning the barricades that year, Shaw was busily explicating the new Fabian economic basis for socialism.

In his preface to the essays, Shaw explained that the writers were all social democrats, ‘with a common conviction of the necessity of vesting the organisation of industry and the material of production in a State identified with the whole people by complete Democracy’. In his contributions he propounded the theory of marginal productivity, demonstrating that Ricardian economic rent, or ‘surplus value’, can accrue to all the factors of production, to land and to labour, and not just to capital as in the Marxist version. Similarly, he rejected the labour theory of value, and advanced the neoclassical version, which he called ‘exchange value’; in other words value was determined by the interaction of supply and demand in the marketplace. Shaw concluded:

> What the achievement of Socialism involves economically, is the transfer of rent from the class which now appropriates it to the whole people. Rent being that part of the produce which is individually unearned, this is the only equitable method of disposing of it. (1908, p. 220)

The method proposed to accomplish the transition was the common ownership of property, or as Webb put it: ‘the gradual substitution of organized operation for the anarchy of competitive struggle’ (p. 62). The original essayists all shared Marx's moral outrage at the evils of capitalism, particularly as a cause of hopeless poverty, inhuman working conditions and excessive inequality, and they also identified the institution of private property as its prime motivating force. However, they did not share the Marxist belief that capitalism must inevitably collapse. Although they recognized that periodic slumps were endemic to the system, they were more struck by its spectacular long-run growth and saw no reason to suppose that it would not continue to reap the benefits of technological change. Thus, as Schumpeter later explained, they were the kind of socialists who believed in the productive success of capitalism while they deplored its distributive results (Schumpeter, 1942, pp. 61–2). They thought that through the gradual extension of public property socialism would evolve from democratic efforts to mitigate the effects of industrialization. Indeed, Webb provided an extraordinary two-page catalogue of socialism's accomplishments to date, which ranged from the army and navy to public baths and cow meadows (Shaw, 1908, pp. 66–7). William Clarke described the growth of joint stock companies, and more recently of ‘rings’ and ‘trusts’, through which ownership became ever more divorced from entrepreneurial function and ‘capitalism ever more inconsistent with democracy and the public interest’. These changes provided the other main Fabian justifications for the public ownership of industry.

Their views on the actual operations of a socialist system were hazy. Shaw and Webb both imply that socialism will have arrived when the entire market operation is administered through nationalization, municipalization and government regulation. Shaw described the aim of social democracy:

> to gather the whole people into the state, so that the state may be trusted with the rent of

> the country, and finally with the land, the capital, and the organisation of the national industry—with all sources of production, in short, which are now abandoned to the cupidity of irresponsible private individuals. (1908, p. 224)

Yet, in other Fabian tracts, Shaw extolled the virtues of competition and of individual freedom, asserting that the latter was ‘as highly valued by the Fabian Society as Freedom of Speech, Freedom of Press, or any other article in the charter of popular liberties’ (Shaw, 1896, p. 327).

Later, of course, the Webbs provided a far more detailed view of their ideas for the organization of a Social Parliament to decide economic policy and to administer public enterprises. Beatrice herself remained ambivalent as to whether unemployment was caused by personal failings or ‘the disease of industry’; their apparently countercyclical unemployment scheme only shifted existing projects without requiring fundamental changes in government policy (Harris, 1972, pp. 42–3). The Webbs’ ideas about state planning were based on administrative principles, not economic science.

In the next Fabian generation, Hugh Dalton, a student of Pigou, used Pigou's revised version of neoclassical theory to demonstrate the critical differences between factor incomes and personal incomes. He introduced and defined the nature of inheritance and its role in maintaining wealth differentials; he broadened its concept to include educational opportunity, access to public services and institutional customs (Dalton, 1920). According to Gaitskell's later assessment of the British tradition, Dalton's work was a decisive influence in shifting socialist thought from the ‘sterile, out-of-date, somewhat academic arguments of earlier writers’ to the practical issues of progressive taxation and educational reform (Gaitskell, 1955, pp. 936–7). Although still grounded in neoclassical criteria of allocative efficiency, Dalton's analysis dealt directly with income equality, opening up ways to achieve socialism other than through Webbian public ownership. Thus, Gaitskell believed that the case for socialist equality could be stated on ‘straightforward ethical principles’, rather than on ‘complicated arguments about economic abstractions’.

The Fabians, the Keynesian revolution and economic planning in the 1930s

The Great Depression threatened both the political and economic stability of capitalist systems. Inspired by the Russian Revolution and its apparent success in replacing capitalism and avoiding mass unemployment, many leftist sympathisers turned to Marxism. They struggled through *Das Kapital*, they visited the Soviet Union, and they recommended the Soviet political philosophy and economic system. The Webbs fell in love with Russia; in their last major work, *Soviet Communism:A New Civilization?*, they advocated a totally controlled economy, visualising Soviet planning as the ultimate Fabian collective. In *New Fabian Essays* Crossman argued that they had simply superimposed Marxism on their basic utilitarianism; he believed that only John Strachey successfully re-thought the entire system ‘in Anglo-Saxon terms’ (Crossman, 1970, p. 5).

It fell to the younger generation to restate the traditional Fabian case against Marxist economic thought and revolutionary methods and to redefine the democratic socialist alternative. Hugh Gaitskell and Evan Durbin organized the Economic Section of the New Fabian Research Bureau, which had been founded by G.D.H. Cole in March 1931 and merged with the parent Fabians in 1938; their purpose was to explore the implications of the theoretical economic controversies for socialism and to make policy

recommendations to the Labour Party (Durbin, 1985). At the same time the obvious failures of the market system were challenging economists to rethink the role of government intervention and to redesign their toolkit. Keynesian macroeconomics, the economics of imperfect competition and the principles of economic planning embodied in the new ‘market socialism’ were first developed during the 1930s. After the war they were incorporated into the orthodox case for the mixed economy.

In pointed contrast to official policy, Keynes had begun pressing British governments to expand, not to contract, public expenditure to cope with unemployment. In the early 1930s his position was largely intuitive; *The General Theory* published in 1936 was the first systematic exposition of his theoretical case. Until then the most fundamental cleavage on the unemployment issue was between those who advocated government intervention in the market and those who did not. Socialists were naturally allied with the interventionists on social and political grounds, as well as economic, and thus were sympathetic to Keynes's policy efforts: but they were suspicious of his political ties to the Liberal Party, and some of the professional; economists were sceptical about his expansionist policies. James Meade and Colin Clark, who were working alongside Keynes, were convinced expansionists by August 1931. Together they were responsible for converting the New Fabians well before 1936. Amongst the sceptics were Gaitskell and Durbin, who were strongly influenced by Hayek's trade cycle theories and who were deeply concerned to demolish ‘treasured dogma’ within the Labour party, namely the myths that capitalism was collapsing and that socialism could easily replace it and automatically solve the unemployment problem. As early as 1932 Gaitskell explained why, although ‘prosperity’ was an important socialist goal, it was not ‘the distinguishing characteristic of the Socialist ideal’ (Gaitskell, 1932).

Meade also played an important role in converting Douglas Jay, whose influential book, *The Socialist Case*, published in 1937, was the first to propose that Keynesian fiscal and monetary measures to control output and employment be explicitly incorporated as part of socialist planning methods. Cole, who thought that the *General Theory* was the most important economics book published since Marx's *Das Kapital* and Ricardo's *Principles*, was quick to point out that because Jay gave such a low priority to nationalization his book contained very little of ‘what most people habitually think of as socialism’ (Cole, 1935). Thus, the introduction of Keynesian methods also served to weaken the case for public ownership as the basis of the socialist economic alternative.

By the late 1930s most democratic socialists in Britain had recognized the importance of the Keynesian message for socialism, and by the end of the war the Labour Party had officially adopted a Keynesian full employment policy. The new macroeconomic analysis provided an obvious answer to the problem of dealing with capitalist collapse. It also reinforced distributive goals, since lower-income families had a higher propensity to consume, and it underscored the importance of central planning to control the economy, since only the government had the power to offset insufficient private spending. So compelling were these arguments that they also converted at least one influential Marxist, John Strachey, to the Fabian cause.

Yet Fabian acceptance of Keynes's economics and of Keynes's basic individualism is often overstated, particularly in the pre-war context. Anthony Wright (in Pimlott, 1984) has suggested that the Tawney approach to equality is fundamentally different from the liberal philosophy behind Beveridge's welfare state. A similar contrast can be made between Fabian conceptions about economic planning in the 1930s and Keynesian macroeconomic management. Fabians were explicit about their opposition to the capitalist system, which Keynes wanted to repair, but which they wanted to replace. They were emphatic

about the need for major reform of Britain's financial institutions and for substantial growth of the public sector; indeed, they believed that both were essential to implement a successful full employment policy. At least one Fabian economist, Evan Durbin, never accepted *The General Theory* model as the solution to all macroeconomic problems; he believed that it failed to explain the trade cycle, and was therefore unsuitable for the long-term growth problems which the socialist state must solve in order to improve upon capitalism's record.

The principles of market socialism grew out of work initiated by Durbin and Gaitskell, who undertook a systematic reconsideration of the Marshallian microeconomic grounds for intervention and the implications for socialist planning. Together with H.D. Dickinson they demonstrated that the market system by definition could neither price collective goods nor reflect the true social value of externalities, and, therefore, that it could not determine the appropriate allocation of resources for their production. They also incorporated the new economics of imperfect competition associated with Joan Robinson to restate the objections to the existing system, which they termed ‘monopoly capitalism’. A planning authority would be able to correct these deficiencies and use the principles of optimal allocation to guide its decisions; in other words, neoclassical criteria should serve as the handmaiden to collective decision-making. In the 1930s and 1940s, many Fabians contributed to the further elaboration of these ideas into a socialist economic system based on free choices in the labour market, consumers’ sovereignty through market pricing and marginal cost pricing in nationalized industries. The importance of this analysis was that it added strong theoretical arguments for a mixed economy as an explicit complement to the macroeconomic Keynesian ones.

There were, however, other Fabians who found such arguments hard to take and/or to follow. Barbara Wootton, whose planning schemes were an updated version of the Webbian administrative structure, was clear that prices would have to be controlled in the public interest. Even Dalton, who recognized that planning was not necessarily socialist, still maintained the early Fabian belief that ‘Socialism is primarily a question of ownership’ (Dalton, 1935, p. 247). With more appreciation for the problems of allocative efficiency under socialism, Cole attempted to fashion a different socialist economics, one which was neither Marxist nor neoclassical (Cole, 1935). Although his own system remained a rather sketchy attempt to incorporate socialist distributional goals into decisions about production, he had some telling arguments against his neoclassical comrades, pointing out that market prices reflected the existing income distribution, and thus could not provide the proper signals for socialist allocation. His efforts are particularly interesting for the light they throw on the need to mesh social policy with economic planning, and on the problem of applying neoclassical analysis to meet essentially political goals.

By the end of the 1930s, most Fabians had come to accept the necessity for a mixed economy, if only on practical grounds, because the legislation necessary to secure socialism by parliamentary methods could not be accomplished by one Labour government. Government planning was necessary to ensure aggregative and allocative efficiency and to redistribute income and wealth. Control of what were later known as ‘the commanding heights’ of the economy was essential to implement the planning alternative, and a central authority was required to make sure that sectional interests, such as bankers, business and trade unionists, did not subvert the public good. However, in an important change of emphasis, Durbin and Gaitskell were explicit that their objections to capitalism and to the Marxist alternative were social and political, not economic (Gaitskell, 1935; Durbin, 1940). The essence of their socialism was social justice as Tawney defined it. In short, the mixed economy was not simply politically expedient, it was

central to the economic operation of the democratic socialist state.

The Fabians and the mixed economy in practice

As authors in the *New Fabian Essays* later pointed out, the war substantially altered the balance of power between the government and the private sector. And in comprehensive plans for recovery, the wartime coalition laid the foundations for bipartisan support of full employment, a unified system of social services and educational reform. Thus, when the Labour government took over in 1945, there was not much resistance to its programme or to its Fabian philosophy.

In 1948 the Fabian Society commissioned W. Arthur Lewis to write a pamphlet on ‘the economic perplexities of the moment’. These turned out to be so numerous that Lewis ended up writing a short book, *The Principles of Economic Planning* (1949), an influential statement of the revised conception of market socialism. Like Meade in *Planning and the Price Mechanism*, published in the same year, he argued the case for planning on general interventionist grounds, implicitly rejecting the Durbin–Gaitskell notion that only a socialist government could run the economy efficiently, although one might still believe only a socialist government would. To paraphrase Lewis, socialism was not about the state, any more than it was about property; ‘socialism is about equality’. There could be many ways to handle property and to plan the economy, which were not inconsistent with socialism (Lewis, 1949, pp. 10–11).

Lewis argued that the crucial issue was whether the state should operate ‘through the price mechanism or in supersession of it’; the real choice was ‘between planning by inducement, and planning by direction’. Lewis himself was neutral on the issue, believing that Britain needed some of both. Although insistent that there must be free consumer goods and labour markets, he argued that demand was not sacred and that it should be manipulated in specific markets and in the aggregate to achieve policy goals. Similarly, he did not believe that nationalization should be taken on its merits. Lewis wanted ‘more than we have already got’ (steel, banking and chemicals were his candidates), but in no circumstances the whole economy; ‘a country whose people love freedom will not wish the state to become the sole employer’ (p. 104).

Shortly after this book was published in 1949, Cole as chairman of the Society organized a conference to begin to rethink the way forward now that the main components of the first Fabian stage to socialism were in place. *New Fabian Essays* published in 1952 was the end result of this effort to take account of important societal changes and the Keynesian Revolution. The essayists were all agreed that the British version of the mixed economy was a permanent Fabian accomplishment, and that the Tories would not dismantle the welfare state nor renege on full employment. Yet, despite the enormous gains, substantial inequities remained and new problems emerged: in particular, the great concentrations of bureaucratic power in the public and private sectors which threatened individual freedom. In general terms the way forward was to continue to pursue equality, to improve labour–management relations and to disperse power as much as possible.

However, the Fabians were still united in their dissatisfaction with that system. Although they were clear that the postwar version of welfare capitalism did not meet their conception of socialism, many of the essayists were vague about what they did want. Writing about equality in *New Fabian Essays*, Roy Jenkins explained that a classless society was one ‘in which men will be separated from each other less sharply by variations in wealth and origin than by differences in character’, but it was impossible to describe ‘the exact shape of the goal’. Of contributors to *New Fabian Essays*, only Crosland was willing

to be explicit in the negative sense that he specified four policies which would *not* achieve equality; the continued extension of free social services, more nationalization, the proliferation of controls and further redistribution of income by direct taxation. In an important shift, many Fabians had come not only to believe in the mixed economy, but also to accept its current structural form.

Crosland outlined the main features of what he called ‘post-capitalist society’: he concluded that it was more equal and more planned than before, but that it was still based on unacceptable class divisions. While individual property rights were no longer the essential basis of economic and social power, they still affected the distribution of wealth. He felt that the power of the state had been expanded sufficiently to exert control over the economy: if anything, physical controls should be reduced as they were unpopular and inefficient. Similarly, nationalization had secured government power in the central sectors of the economy, social legislation had ensured a national minimum welfare level, and full employment policies had removed insecurity and demonstrated that central planning could be directed to meet social ends. Keynesian policies were crucial to maintaining this system, but as these were now well understood, Crosland argued that ‘the new society may prove to be a very enduring one’. In *The Future of Socialism* (1956), Crosland spelled out his ideas on planning in more detail; he believed its ‘essential role’ was Keynesian economic management, that the techniques were no longer controversial nor the preserve of any one party, and that political will, not planning theory, were required to plan effectively; ‘if socialists want bolder planning, they must choose bolder ministers.’

One lone dissenter from the general Fabian romance with Keynes was G.D.H. Cole. Although enthusiastic about the *General Theory* when it was published, he had become increasingly concerned about these new directions after the war. Indeed, this was precisely why he had initiated the process of rethinking, and why, as the discussions progressed, he resigned his position as chairman of the Fabian Society. In 1950 he published a short book, *Socialist Economics*, which spelled out his disagreements with the new Fabian approach. First, he thought that Keynesian economics was too involved with aggregates and not sufficiently concerned with the structural problems necessary for a socialist economy to replace the capitalist system. As far as he was concerned the new direction provided a diluted form of socialism, which was ‘little more than Keynesian Liberalism with frills’. Second, although Cole had advocated using a wide range of industry controls as early as 1929 and was opposed to total public ownership, he was also explicit in rejecting the current version of the mixed economy ‘as a permanent resting place’.

See Also

- social democracy

Bibliography

Clarke, P. 1978. *Liberals and Social Democrats*. Cambridge: Cambridge University Press.

Cole, G.D.H. 1935. *Principles of Economic Planning*. London: Macmillan.

Crosland, A. 1956. *The Future of Socialism*. London: Cape.

Crossman, R.H., ed. 1952. *New Fabian Essays*. London: Turnstile Press; 3rd impression, London: Dent, 1970.

Dalton, H. 1920. *Some Aspects of the Inequality of Incomes in Modern Communities*. London: Routledge & Kegan Paul.

Dalton, H. 1935. *Practical Socialism for Britain*. London: George Routledge.

Durbin, E.F.M. 1940. *The Politics of Democratic Socialism*. London: Routledge & Kegan Paul.

Durbin, E. 1985. *New Jerusalems: The Labour Party and the Economics of Democratic Socialism*. London: Routledge & Kegan Paul.

Gaitskell, H.T.N. 1932. Socialism and wage policy. Fabian Society Papers, Box J24/2 in Nuffield College, Oxford.

Gaitskell, H.T.N. 1935. Financial policy in the transition period. In *New Trends in Socialism*, ed. G.E.G. Catlin. London: Lovat Dickson & Thompson.

Gaitskell, H.T.N. 1955. The ideological development of democratic socialism in Britain 1955. *Socialist International Information* 5, 52–3.

Harris, J. 1972. *Unemployment and Politics*. Oxford: Oxford University Press.

Lewis, W.A. 1949. *The Principles of Economic Planning*. London: Dobson.

Pimlott, B., ed. 1984. *Fabian Essays in Economic Thought*. London: Gower Publishing.

Schumpeter, J.A. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper & Row; Torchbook edition, 1962.

Shaw, G.B. 1884, 1896. *Fabian Tract No. 2*. (1884) *Tract No. 70* (1896). Quoted in Crosland (1956).

Shaw, G.B., ed. 1908. *Fabian Essays in Socialism*. New York: Doubleday edn, 1967.

Wallas, G. 1923. Article in *Morning Post*, 1 January. See Clarke (1978) for further details.

Wright, A.W. 1979. *G.D.H. Cole and Socialist Democracy*. Oxford: Clarendon Press.

How to cite this article

Durbin, Elizabeth. "Fabian economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000001> doi:10.1057/9780230226203.0541

The New Palgrave Dictionary of Economics Online

factor content of trade

Donald R. Davis

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Trade in goods is also implicitly trade in the services of the factors used to produce those goods. This insight underlies the Heckscher–Ohlin–Vanek model of factor service trade, and provides a laboratory to test our theories concerning world general equilibrium. In recent years this theory has undergone close empirical scrutiny. Early tests strongly rejected the simplest variants of the theory. More recent tests have imposed a modest number of additional restrictions suggested by the data. These involve cross-country heterogeneity in productivity, factor prices, consumption patterns, and the incorporation of non-traded goods. With these restrictions, the model fares well.

Keywords

factor content of trade; factor price equalization; Heckscher–Ohlin trade theory; Heckscher–Ohlin–Vanek factor content of trade theory; integrated equilibrium; total factor productivity

Article

International trade is the cross-border exchange of goods (and services), both final and intermediate. These goods are produced with factors of production located in specific countries, hence the trade in these goods is implicitly also trade in the services of the factors used to produce them. This converts the standard Heckscher–Ohlin model of trade in goods into the Heckscher–Ohlin–Vanek (HOV) model of the factor content of trade.

The most important reason to study the factor content of trade is that it provides a laboratory to test our understanding of world general equilibrium. Countries have specific endowments, technologies, tastes, locations, and distributions of incomes (among other characteristics). The simplest statement of general equilibrium is that these elements are supposed to ‘hang together’ in a coherent way. Tests of the factor content of trade thus become a first test of the adequacy of our understanding of this world general equilibrium. If we should fail to correctly predict the factor content of trade, then we know that our theory seriously misunderstands at least one element of the underlying reality. If our theory does a good

job of making sense of the factor content of trade, then this is a suggestion that the main thrust of our theory is working well. This would give us more confidence, then, in using the theory in policy applications.

The canonical Heckscher–Ohlin–Vanek model of factor service trade can be described simply (see Vanek, 1968). Assume that there are *G* goods, each produced under perfect competition with constant returns to scale. Assume as well that there are *F* primary factors of production with factor markets competitive. Let $\mathbf{A}$ be an input–output matrix that links net output $\mathbf{Y}$ to gross output $\mathbf{X}$ via $\mathbf{Y} = (\mathbf{I} - \mathbf{A})\mathbf{X}$. Let $\mathbf{B}$ be a matrix of direct factor inputs, with dimension $F \times G$, where columns denote factor inputs required to produce a unit output of a single good and rows show factor inputs for a single factor across all goods. Let $\overline{\mathbf{B}} \equiv \mathbf{B}(\mathbf{I} - \mathbf{A})^{-1}$ be the corresponding matrix of direct plus indirect factor inputs, where both primary and intermediate usage represent cost-minimizing choices. Let c be an index for countries, and W represent the aggregate for the world as a whole.

Let technologies for all goods and quality of all factors be common for all countries of the world, and let there be at least as many goods as factors, so that $G \geq F$. Assume that trade between countries is free, so that goods prices are equalized. Assume that the distribution of world endowments among countries satisfies the requirements to replicate what has been termed the ‘integrated equilibrium’ (see Helpman and Krugman, 1985). In such a case, the division of world endowments between the countries is of no economic consequence, since outputs adjust across countries so that the countries jointly produce exactly the same output and use the same input ratios as they would if the factors were perfectly mobile across countries. Then factor prices will be equalized (FPE), and for all countries $c \in C$, there are common technology matrices: $\mathbf{B} = \mathbf{B}^c$, and $\overline{\mathbf{B}} = \overline{\mathbf{B}}^c$. For country *c* with gross output vector $\mathbf{X}^c$ and primary input vector $\mathbf{V}^c$, $\mathbf{B}\mathbf{X}^c = \mathbf{V}^c$. We further assume that demand is identical across countries and homothetic. Let $\mathbf{D}^c$ be country *c*'s vector of final goods demand, $\mathbf{Y}^W$ be the world net output vector, and s^c be country *c*'s share of world spending. Then, with free trade equalizing goods prices, $\mathbf{D}^c = s^c\mathbf{Y}^W$. This identifies the demand for goods, and, by pre-multiplying by the common technology matrix $\overline{\mathbf{B}}$, we can convert this to a statement about the factor content of consumption. $\overline{\mathbf{B}}\mathbf{D}^c = s^c\mathbf{V}^W$. Net trade is $\mathbf{T}^c = \mathbf{Y}^c - \mathbf{D}^c$. Hence the prediction of the net factor content of trade is:

$$\text{(HOV)} \quad \overline{\mathbf{B}}\mathbf{T}^c = \mathbf{V}^c - s^c\mathbf{V}^W.$$

Early empirical work, such as Bowen, Leamer and Leo (1987) and Trefler (1995), examined this under the assumption that all countries use the technology matrices of the United States. Without reservation, the conclusion of these papers was that the simplest version of the model is an utter failure. Trefler characterized the central failing as the ‘mystery of the missing trade’. If we term $\overline{\mathbf{B}}\mathbf{T}^c$ the *measured* factor content of trade and $\mathbf{V}^c - s^c\mathbf{V}^W$ the *predicted* factor content of trade, then the mystery is that the measured factor content of trade is much smaller than that predicted. Much of the subsequent literature has focused on identifying reasons for the mystery of the missing trade and finding solutions for it. Virtually every assumption underlying the Heckscher–Ohlin–Vanek model is in principle open to

question. The strategic issue has been to bring more data to bear on the question in order to identify which of the assumptions is violated most seriously and what amendments to the theory and data work are needed to fit the pieces of the puzzle together.

Various approaches have been considered. Trefler (1993) develops a model that assumes net factor trade is correctly measured, and calibrates factor quality differences across countries that would rationalize the measured trade. This can be thought of as a model of adjusted factor price equalization. While the theoretical model of quality-adjusted factor service trade is an important addition to the toolkit of researchers, this proposed resolution has not fared well empirically (Davis and Weinstein, 2003).

Increasingly, researchers moved to a wider set of departures from the standard Heckscher–Ohlin–Vanek model. These include differences across countries in total factor productivity (TFP); a breakdown in factor price equalization, even adjusted for the TFP differences; specialization in different traded goods within industries; differences in factor input ratios in both traded and non-traded sectors; and costly trade. Davis et al. (1997) examined the adequacy of assuming a common technology matrix (in this case, that of Japan) for a set of OECD countries. Instead of looking directly at the factor content of trade, $\mathbf{B}^{Japan}\mathbf{T}^{c} = \mathbf{V}^{c} - s^{c}\mathbf{V}^{W}$, they looked at the factor content of production for these countries, that is, $\mathbf{B}^{Japan}\mathbf{X}^{c} = \mathbf{V}^{c}$. This is such a poor fit in the data that they conclude that much of the problem lies in cross-country differences in technology matrices. They went on to develop a theory to predict the factor content of Japanese regions under the assumption that these share FPE, even though Japan does not share FPE with the world as a whole. In this sample, this largely eliminated the mystery of the missing trade.

This left open the larger question of why technologies differed, how they differed, and whether a parsimonious set of departures from the HOV theory could get the model of factor service trade to work well. Davis and Weinstein (2001a) brought a great deal more data to bear on the problem, developing technology matrices for ten rich OECD countries and a composite rest of the world. Technologies differed systematically, even among these rich OECD countries, so that more capital-abundant countries use more capital-intensive methods industry by industry. As it turned out, this happened in both traded and non-traded goods sectors, the latter being important in identifying a breakdown in relative FPE (because there is less likelihood of aggregation issues impinging). Moreover, recognizing that non-traded sectors in different countries use systematically different input coefficients has a large impact on predicted factor contents. For example, a capital-abundant country uses more capital per worker than would be suggested in an FPE model. For this reason, and because non-traded sectors are large, the capital-abundant country has less ‘excess’ capital to export through factor services. All told, the adjustments made allow the measured factor content of trade to be approximately 60–80 per cent of that predicted.

The subsequent literature has focused on a number of elaborations and challenges to this work. Feenstra and Hanson (2000) explore in more detail issues of aggregation bias in measurements of net factor service trade. In related work, Davis and Weinstein (2001b) have developed a more elaborate model of gross trade in factor services that helps to understand even North–North trade. In effect, they argue that much of the mystery of the missing trade arose because the focus on net goods trade ignored the fact that when factor intensities are not identical even intra-industry goods trade conveys net factor content. Choi and Krishna (2004) implement alternative tests, based on Helpman (1984), of the net factor content of trade which has the advantage of being robust to breakdowns in FPE, but the disadvantage of needing to

have confidence that we can adequately measure the differences in factor prices, including returns to capital. For the sample of bilateral predictions they consider, the model performs well. Reimer (2006) has aimed to incorporate a more elaborate model of trade in intermediates and argues that this diminishes when measured against predicted factor contents.

Research on the factor content of trade is important because it represents the greatest effort on the part of trade economists to assemble all of the pieces of general equilibrium into a single coherent framework relating underlying endowments, production, technology, consumption and trade. The early theoretical and empirical work provided a starting place and a number of anomalies, such as the mystery of the missing trade, that motivated ongoing research. Subsequent literature has gone a long way towards resolving the mystery of the missing trade. But new questions continue to arise, particularly related to trade in intermediates and issues of aggregation. No doubt these will invite further investigation.

See Also

- comparative advantage
- factor price equalization (historical trends)
- factor prices in general equilibrium
- Heckscher–Ohlin trade theory
- international economics, history of
- tradable and non-tradable commodities

Bibliography

Bowen, H.P., Leamer, E.E. and Leo, S. 1987. Multicountry, multifactor tests of the factor abundance theory. *American Economic Review* 77, 791–809.

Choi, Y.S. and Krishna, P. 2004. The factor content of bilateral trade: an empirical test. *Journal of Political Economy* 112, 887–914.

Davis, D.R. and Weinstein, D.E. 2001a. An account of global factor trade. *American Economic Review* 91, 1423–54.

Davis, D.R. and Weinstein, D.E. 2001b. Do factor endowments matter for North–North trade? Working Paper No. 8516. Cambridge, MA: NBER.

Davis, D.R. and Weinstein, D.E. 2003. The factor content of trade. In *Handbook of International Trade*, ed. J.C. Harrigan. London: Blackwell.

Davis, D.R., Weinstein, D.E., Bradford, S. and Kazushige, S. 1997. Using international and Japanese regional data to determine when the factor abundance theory of trade works. *American Economic Review* 87, 421–46.

Feenstra, R.C. and Hanson, G.H. 2000. Aggregation bias in the factor content of trade: evidence from U. S. manufacturing. *American Economic Review* 90, 155–60.

Helpman, E. 1984. The factor content of foreign trade. *Economic Journal* 94, 84–94.

Helpman, E. and Krugman, P. 1985. *Market Structure and Foreign Trade*. Cambridge, MA: MIT Press.

Reimer, J.J. 2006. Global production sharing and trade in the services of factors. *Journal of International Economics* 68, 384–408.

Trefler, D. 1993. International factor price differences: Leontief was right! *Journal of Political Economy* 101, 961–87.

Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85, 1029–46.

Vanek, J. 1968. The factor proportions theory: the N-factor case. *Kyklos* 21, 749–56.

How to cite this article

Davis, Donald R. "factor content of trade." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 30 December 2008 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000321> doi:10.1057/9780230226203.0542

The New Palgrave Dictionary of Economics Online

factor models

Jushan Bai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Factor models explain correlations among a set of variables. By postulating that the variables are linked with a small number of latent components, factor models imply a particular structure for the correlation matrix. This article discusses the model's identification and estimation as well as their applications in economics.

Keywords

approximate factor model; arbitrage pricing theory; cointegration; common factors; covariance matrix; diffusion index forecasting; dynamic factors; factor analysis; factor models; forecasting; principal components analysis; sample correlation matrix; unit roots

Article

The primary objective of factor analysis is to explain, in a parsimonious way, the correlation among a set of variables. For example, cross-sectional correlation of asset returns may be explained by a single factor, according to capital asset pricing theory (Sharpe, 1964; Lintner, 1965). The correlation among a large number of macroeconomic variables could be explained by some common shocks (Sargent and Sims, 1977; Bernanke and Boivin, 2003). Historically, factor models were used by psychologists to examine correlations among a set of test scores. Students' performance across different subjects (maths, philosophy, history, and so on) may potentially be accounted for by a single factor (for example, overall intelligence) (see Lawley and Maxwell, 1971). In these examples, a common theme is that a large number of variables are linked with a small number of unobservable variables which give rise to the cross correlations.

While sample correlation matrix may serve the same purpose in describing the linkage of variables, it is neither parsimonious nor reliable when the number of variables is large relative to the number of observations. Suppose there are N variables, each with T observations. The sample correlation matrix

estimates $N(N-1)/2$ parameters without any restriction. When the number of variables exceeds the number of observations $(N > T)$, the sample correlation matrix is not of full rank, even though the underlying true correlation matrix is positive definite. A factor model with, say, a single factor attempts to explain the correlation with far fewer parameters, and the resulting correlation matrix will be positive definite. If a factor structure truly (or approximately) characterizes the data generating process, the estimated correlation matrix implied by the factor model constitutes a better estimate than the sample correlation matrix. Even if the data generating process does not follow a factor model, under large *N*, shrinking the sample correlation matrix towards a correlation matrix with a factor structure may be desirable, in light of Ledoit and Wolf (2003). Most importantly, sample correlation is purely statistical, but factor models have structural interpretations.

In this article, we first present the mathematical form of the factor model, then we state the assumptions employed by classical factor analysis, in which the statistical theory is developed under a fixed *N*. We then go on to discuss modern factor analysis in which both *N* and *T* are large, and in particular, the number of variables (*N*) can be much larger than the number of observations (*T*). In each case, we discuss issues related to identification and estimation, and the determination of number of factors. More attention is paid to modern factor analysis. We also present a few applications of factor models in economics, including diffusion index forecasting, panel unit root and cointegration analysis. Finally, we briefly highlight the difference between principal component analysis and factor analysis.

The model

A factor model takes the form

$$X_{it} = \mu_i + \lambda_i' f_t + e_{it}; \; i = 1, 2, \ldots, N; \;\; t = 1, 2, \ldots, T$$

where X_{it} is the observation on variable *i* at period *t*; μ_i is the mean of X_{it}, $\lambda_i\,(r \times 1)$ is vector of factor loadings, $f_t(r \times 1)$ is vector of factor processes, and e_{it} is the idiosyncratic error term. For example, X_{it} may represent the output growth rate for country *i* in quarter *t*, μ_i is the mean growth rate, f_t is a vector of common shocks (technology shocks, financial crises, oil price shocks, and so on) that influence output, λ_i represents the impact of shocks on country *i*, and e_{it} is the country-specific growth rate. As a further example, X_{it} is the return of asset *i* in period *t*, μ_i is the mean return, f_t is a vector of factor returns with zero mean (risk premia adjusted), λ_i is a vector of factor loadings, e_{it} is the idiosyncratic return. The arbitrage pricing theory of Ross (1976) implies restrictions between μ_i and λ_i.

Introducing the following notation

$$X_t = \begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{Nt} \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}, F = \begin{bmatrix} f_1' \\ f_2' \\ \vdots \\ f_T' \end{bmatrix}, \Lambda = \begin{bmatrix} \lambda_1' \\ \lambda_2' \\ \vdots \\ \lambda_N' \end{bmatrix}, e_t = \begin{bmatrix} e_{1t} \\ e_{2t} \\ \vdots \\ e_{Nt} \end{bmatrix}$$

the factor model can be rewritten as

$$X_t = \mu + \Lambda f_t + e_t; \; t = 1, 2, \ldots, T. \quad (1)$$

Below we separately discuss classical factor analysis and modern factor analysis; they are based on different assumptions and inferential theory also differs.

Classical factor analysis

The main assumptions under classical factor analysis are: (i) f_t and e_t are i.i.d. random variables with zero means; (ii) for normalization purposes, $Ef_t f_t' = I_r$, an identity matrix; (iii) e_t and f_t are uncorrelated; (iv) Λ is a matrix of fixed constants. Let $\Sigma = E(X_t - \mu)(X_t - \mu)'$, the covariance of X_t and let $\Phi = Ee_t e_t'$, the covariance of e_t. It follows that

$$\Sigma = \Lambda\Lambda' + \Phi. \quad (2)$$

Another key assumption under classical factor analysis is that N is fixed. This assumption appears to be at odds with the essence of factor analysis because this analysis was motivated by large N problems. Nevertheless, traditional applications had been on problems of relatively small N. Furthermore, fixed N assumption makes the statistical inference more tractable. For example, Σ is consistently estimable under fixed N (for example, by the sample covariance matrix) as T goes to infinity. Thus for identification purposes, Σ is assumed to be known.
Without further restrictions, the parameter matrices Λ and Φ are not identifiable since Φ alone would have the same number of parameters as the number of equations in (2). Classical factor analysis thus

assumes that Φ is a diagonal matrix. This assumption is not too restrictive since correlation among the variables is supposedly explained by the common factors f_t. In addition, a rotational indeterminacy exists for Λ since $\Lambda G(\Lambda G)' = \Lambda\Lambda'$, where G is such that $GG' = I_r$. To remove this rotational indeterminacy, it is often assumed that $\Lambda' \Phi^{-1} \Lambda$ is a diagonal matrix. A diagonal matrix imposes $r(r-1)/2$ number of restrictions. Thus the number of parameters on the right hand side (2) is $N + Nr - r(r-1)/2$. The number of equations in (2) is $N(N+1)/2$. Thus in order to identify the parameters, we must have

$$s = N(N+1)/2 - N - Nr + r(r-1)/2 = [(N-r)^2 - (N+r)]/2 \geq 0.$$

This is known as the order restriction, meaning that the number of equations must be no smaller than the number of parameters. This implies that for a factor model to be identifiable, N cannot be smaller than three. When N is exactly three, there can only be one factor $(r = 1)$. In this case, $s = 0$ and the number of parameters in the factor model coincides with the number of elements in Σ. When this occurs, no simplification is achieved via factor analysis. Nevertheless, structural interpretation of the model is still of interest since it indicates that the three variables are related to a single common component.

Even for $s = 0$, there may not exist solutions for Λ and Φ to satisfy (2) because factor models further restrict non-negativity for the diagonal elements of Φ; see examples in Lawley and Maxwell (1971, pp. 10–11).

For a larger N and small r, we usually have $s > 0$. In this case, overidentification occurs. Model estimation entails finding Λ and Φ to make the distance between S and $\Lambda\Lambda' + \Phi$ small, where S is the sample covariance matrix. The model is usually estimated by the principal-factor analysis or the maximum-likelihood method (see Mardia, Kent and Bibby, 1979; Anderson, 1984).

A special case is that $\Phi = \sigma^2 I_N$, a scalar multiple of an identity matrix. In this case, the smallest N that permits identification is $N = 2$, with $r = 1$. To see this, consider

$$\Sigma = \lambda\lambda' + \sigma^2 I_2$$

where $\lambda = (\lambda_1, \lambda_2)'$ is a vector. Reparameterize $\lambda\lambda' = \tau^2\delta\delta'$ with $\|\delta\|^2 = \delta'\delta = 1$, where $\tau^2 = \|\lambda\|^2 = \lambda'\lambda$. The two eigenvalues of the matrix on the right hand side are σ^2 and $\sigma^2 + \tau^2$. The eigenvector associated with the larger eigenvalue is simply δ. Thus we can identify σ^2 as the smaller eigenvalue, and τ^2 as the difference between the two eigenvalues. Moreover, δ is the eigenvector associated with the larger eigenvalue of Σ.

On the assumption that the model is identifiable, the estimated factor loadings will be consistent, the limiting distribution can be found in Anderson (1984) under the assumption of fixed N and large T.

Given Λ and Φ, the factor scores f_t can also be estimated by either the generalized least squares (GLS) or the Bayesian method. For example, the GLS estimator of f_t is $\hat{f}_t = (\Lambda'\Phi^{-1}\Lambda)^{-1}\Lambda'\Phi^{-1}X_t \ (t = 1, 2, \ldots, T)$. While f_t is unbiased for f_t, it is not consistent since N is fixed, even if Λ and Φ are known. Finally, the number of factors r is determined via hypothesis testing.

Modern factor analysis

Modern factor analysis takes model (1) as the starting point, but then proceeds under different assumptions. First, the number of variables N is assumed to be large, and the limit theory is developed under the assumption that both N and T go to infinity. In particular, N can be much larger than the number of observations T. Second, both f_t and e_t can be serially correlated. Third, Φ needs not to be a diagonal matrix, and in fact, none of the off-diagonal elements needs to be zero. Thus the number of parameters in Φ can be as many as the equations. This is called 'approximate factor model' by Chamberlain and Rothschild (1983). The main interest of this large dimensional factor analysis is to estimate r, Λ, and F. One key assumption of the approximate factor model is that the largest eigenvalue of Φ is bounded uniformly in N. This implies that cross-correlations in the idiosyncratic errors must be weak.

Identification and estimation

Let $X = (X_{it})$ be the $N \times T$ data matrix and $e = (e_{it})$ be the error matrix of the same dimension. Then

$$X = \Lambda F' + e.$$

Here we assume the constant vector μ to be zero, but without assuming F to have zero mean. If $\mu \neq 0$, the demeaned data matrix should be used in the following discussion, and zero mean for F should also be imposed. Now both Λ and F are to be estimated. Since $\Lambda F' = \Lambda A A^{-1} F'$ for an arbitrary invertible matrix A $(r \times r)$. As an arbitrary $r \times r$ invertible matrix has r^2 free parameters, we need to impose r^2 restrictions. We may impose $\frac{1}{T}F'F = \frac{1}{T}\sum_{t=1}^{T} f_t f_t' = I_r$ together with $\Lambda'\Lambda$ being diagonal. Alternatively, we may impose $\frac{1}{N}\Lambda'\Lambda = I_r$ together with $F'F$ being diagonal. Either way, it will uniquely fix Λ and F (up to a column sign change) given the product $\Lambda F'$.
Under the least squares objective function $S(\Lambda, F) = tr[(X - \Lambda F')(X - \Lambda F')']$, the optimal solution $(\tilde{F}, \tilde{\Lambda})$ is simply the principal-components estimator. More specifically, under the first set of normalization restrictions, $\tilde{F}$ is the $T \times r$ matrix consisting of the first r eigenvectors (multiplied by $\sqrt{T}$) associated with the first r largest eigenvalues of the $T \times T$ matrix $X'X$, and $\tilde{\Lambda} = \frac{1}{T}X\hat{F}$. Under the second

set of identification restrictions, the optimal solution $(\bar{F}, \bar{\Lambda})$ is also an eigenvalue problem associated with the matrix of XX', which is $N \times N$. That is, $\bar{\Lambda}$ is the matrix of the first *r* eigenvectors (multiplied by $\sqrt{N}$) of the matrix XX' and $\bar{F} = X'\bar{\Lambda} / N$ (see Connor and Korajzcyk, 1986; Stock and Watson, 2002a). The relationships between the two sets of solutions are given by $\tilde{F} = \bar{F}V^{-1/2}$ and $\tilde{\Lambda} = \bar{\Lambda}V^{1/2}$, where *V* is an $r \times r$ diagonal matrix consisting of the eigenvalues of the matrix $\frac{1}{NT}XX'$. The statistical properties of $\tilde{F}$ and $\tilde{\Lambda}$ are analysed by Bai (2003).

The number of factors

The number of factors *r* can be consistently estimated using the information criterion approach of Bai and Ng (2002). Let $\hat{\sigma}^2(k)$ denote the sum of squares residuals (divided by *NT*) when *k* factors are allowed, that is, $\hat{\sigma}(k)^2 = S(\tilde{\Lambda}^k, \tilde{F}^k) / (TN)$. Consider the following criterion

$$IC(k) = \log\hat{\sigma}(k)^2 + kg(N, T).$$

If $g(N, T)$ is such that $g(N, T) \to 0$ and $\min[N, T]g(N, T) \to \infty$, then $P(\hat{k} = r) \to 1$, where $\hat{k}$ minimizes the information criterion. For example, $g(N, T) = (N + T)\log(NT) / (NT)$ satisfies the above condition.

Nonstationary factor analysis

When the factor process f_t is a vector of I(1) or integrated processes such that $f_t = f_{t-1} + \eta_t$, X_{it} is nonstationary. Examples include nominal exchange rates series (see Banerjee, Marcellino and Osbat, 2005). When the idiosyncratic process e_{it} is I(0) both Λ and *F*, as well as *r* can be consistently estimated, as shown by Bai (2004). Since X_{it} (for all *i*) share the same common stochastic trends f_t, X_{it} are cointegrated among themselves.

When the idiosyncratic process e_{it} is I(1) for all *i* such that $e_{it} = e_{i,t-1} + \varepsilon e_{it}$, there is no cointegration among the observable X_{it}. But still, the common stochastic trends are well defined and can be estimated consistently up to a rotation, a striking contrast with a fixed *N* spurious system. In a small *N* system, common stochastic trends and cointegration are synonymous. A spurious system has no common trends or at least cannot be discerned. To see how large *N* makes a difference, consider the system in differenced form

$$\Delta X_{it} = \lambda_i' \eta_t + \varepsilon_{it}$$

where $\eta_t = \Delta f_t$ and $\varepsilon_{it} = \Delta e_{it}$. This is a standard factor model, and η_t can be estimated under large N and large T. Recumulating η_t will obtain f_t up to a location (unless $f_1 = 0$) and scale shift. When the initial observation $X_{i1} = \lambda_i' f_1 + \varepsilon_{i1}$ is included in estimation, there is only a scale shift in the estimated f_t.

The above idea is implemented in Bai and Ng (2004) for testing panel unit roots. The process X_{it} will have a unit root if either f_t or e_{it} has a unit root. The key is to consistently estimate f_t and e_{it} without knowing a priori their integration orders. Bai and Ng propose to test separately the nonstationarity property for the common component and the idiosyncratic components. This permits us to trace the source of a nonstationary property arising from a common or idiosyncratic component.

Moon and Perron (2004) and Phillips and Sul (2003) propose methods for testing unit roots in the idiosyncratic errors. Related studies can be found in the surveys by Breitung and Pesaran (2008) and Choi (2006).

Diffusion-index forecasting

Large-dimensional factor models have proven useful in forecasting macroeconomic variables. Let y_t be the variable to be forecasted, say inflation. Consider the h-period-ahead forecasting equation,

$$y_{t+h} = \alpha' w_t + \beta' f_t + \varepsilon_{t+h}$$

where w_t is a set of observable predictors, such as the lags of y_t and the unemployment rate under the Philips curve model. Here, f_t is not observable, but it captures the co-movement among a large number of macroeconomic variables X_{it}, which links to f_t according to (1). Stock and Watson (2002a; 2002b) suggest that f_t be extracted from X_{it} to obtain $\hat{f}_t$, and then use $\hat{f}_t$ in place of f_t in the forecasting equation. This method is referred to as diffusion index forecasting, which outperforms many competing methods. Bai and Ng (2006a) analyse the statistical properties of this method. A modified diffusion index approach is proposed in Bai and Ng (2006b). The modified approach consists of two steps. The first step selects a subset of X_{it} that is relevant to y_t based on certain criteria. The second step proceeds as the usual diffusion approach using the selected subset of X_{it} only.

Large dimensional covariance matrix

A large dimensional covariance matrix is useful in financial risk management and portfolio construction. For $N>T$, the sample covariance matrix $S(N\times N)$ as an estimator for Σ is not full rank. Thus we consider factor-model based estimator. For this purpose, we use a demeaned data matrix, denoted by X_c (remove the time series mean for each series). Note that the sample covariance matrix is $S = \frac{1}{T-1} X_c X_c'$. Estimate the factor $\tilde{F}$ in the same manner as above using $X_c' X_c$, then $\tilde{\Lambda} = X_c \tilde{F} / T$, and $\tilde{e} = X_c - \tilde{\Lambda}\tilde{F}'$. Given these estimates, a factor-model based covariance matrix is then defined as

$$\hat{\Sigma} = \alpha \tilde{\Lambda}\tilde{\Lambda}' + D$$

where D is a diagonal matrix with typical element $\hat{d}_i = \frac{1}{T-1}\Sigma_{t=1}^{T} \hat{e}_{it}^2 \ (i = 1, 2, \ldots, N)$, $\alpha = T / (T-1)$. The diagonal elements of $\hat{\Sigma}$ coincide with the corresponding elements of the sample covariance matrix S. In essence, $\hat{\Sigma}$ is an estimator that shrinks the off-diagonal elements of S towards zero. Also the inverse of this matrix is quite easy to compute, it is given by

$$\tilde{\Sigma}^{-1} = D^{-1} - \alpha D^{-1}\tilde{\Lambda}(I_r + \alpha\tilde{\Lambda}' D^{-1}\tilde{\Lambda})^{-1}\tilde{\Lambda}' D^{-1},$$

which only requires the inverse of a diagonal matrix and that of an $r \times r$ matrix. Other covariance estimators are discussed by Ledoit and Wolf (2003; 2004), and Fan, Fan, and Lv (2006).

Dynamic factor models

In model (1), the factor process f_t is allowed to be a general dynamic process. However, the relationship between X_{it} and f_t is static. A general dynamic factor model is defined as

$$X_{it} = \mu_i + \gamma_i(L)' u_t + e_{it}$$

where u_t are i.i.d. random vectors, and $\gamma_i(L) = \Sigma_{i,k=0}^{\infty} \gamma_{ik} L^k$ with L being the lag operator. Sargent and Sims (1977), Quah and Sargent (1993) and Geweke and Singleton (1981) are among the early researchers who have studied the dynamic factor models in economics. Identification and estimation of the general dynamic factor model is studied by Forni et al. (2000), who extend the dynamic principal

components analysis of Brillinger (1981) to large N. If $\gamma_i(L)$ is a finite order polynomial such that $\gamma_i(L)'u_t = \gamma_{i0}'u_t + \cdots + \gamma_{ip}'u_{t-p}$ (a finite order moving average) then the dynamic factor model can be written as a static factor model by defining $f_t = (u_t', \ldots, u_{t-p}')'$, and $\lambda_i = (\gamma_{i0}', \ldots, \gamma_{ip}')'$ so that $\gamma_i(L)'u_t = \lambda_i' f_t$. The usual principal components method is still applicable. In general, when the coefficients in $\gamma_i(L)$ decays to zero quickly, $\gamma_i(L)u_t$ can be approximated by a finite order moving average.

Relationship with principal components analysis

Principal components analysis (PCA) seeks linear combinations of the observable variables that give rise to maximum variations. The aim is to summarize the data with as few components as possible without losing too much information. In doing so, it imposes no restrictions on the covariance matrix, as does factor analysis. As such, PCA is a pure dimension-reduction technique.

Factor analysis aims to explain the correlations or co-movements among the observable variables. It assumes that observable variables are linked with a small number of unobservable variables (factors), which are responsible for the correlation. Thus factor analysis is conducted based on a model. In contrast, PCA can be considered as a model-absence method.

Factor models can be estimated by the principal components method. The so-called principal-factor analysis is an iterated principal components method (see Mardia, Kent and Bibby, 1979). There are three situations in which the principal components method (without iteration) will give either identical or similar results as other factor estimation methods: (i) the idiosyncratic covariance is a scalar multiple of an identity matrix, that is, $\Phi = \sigma^2 I_N$; (ii) the idiosyncratic error variance is small, that is, Φ is close to zero; (iii) the number of variables of N is large.

Summary

Factor analysis is a model for correlations, postulating that correlations be induced by a few unobservable common components. The model implies a structure on the covariance matrix, which has far fewer free parameters than unrestricted covariance matrix. Therefore, factor models are employed in problems where a reduction in the number of parameters is desired. Applications in economics include modelling cross-sectional correlation, capturing co-movements, forecasting, panel unit root and cointegration analysis, as well as financial risk management and optimal portfolio construction.

See Also

- arbitrage pricing theory
- cointegration
- forecasting
- longitudinal data analysis

- time series analysis
- unit roots

Bibliography

Anderson, T.W. 1984. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Angelini, E., Henry, J. and Mestre, R. 2001. Diffusion index-based inflation forecasts for the Euro area. Working Paper No. 6, European Central Bank.

Bai, J. 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71, 135–73.

Bai, J. 2004. Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics* 122, 137–83.

Bai, J. and Ng, S. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.

Bai, J. and Ng, S. 2004. A panic attack on unit roots and cointegration. *Econometrica* 72, 1127–77.

Bai, J. and Ng, S. 2006a. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74, 1133–50.

Bai, J. and Ng, S. 2006b. Forecasting using targeted predictors. Unpublished manuscript, Department of Economics, New York University.

Banerjee, A., Marcellino, M. and Osbat, C. 2005. Testing for PPP: should we use panel methods? *Empirical Economics* 30, 77–91.

Bernanke, B. and Boivin, J. 2003. Monetary policy in a data rich environment. *Journal of Monetary Economics* 50, 525–46.

Breitung, J. and Pesaran, M.H. 2008. Unit roots and cointegration in panels. In *The Econometrics of Panel Data*, 3rd edn, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Brillinger, D.R. 1981. *Time Series: Data Analysis and Theory*. San Francisco: Holden-Day.

Chamberlain, G. and Rothschild, M. 1983. Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1305–24.

Choi, I. 2006. Nonstationary panels. In The Palgrave *Handbook of Econometrics*, vol. 1, ed. T. Mills and

K. Patterson. Basingstoke: Palgrave Macmillan.

Connor, G. and Korajzcyk, R. 1986. Performance measurement with the arbitrage pricing theory: a new framework for analysis. *Journal of Financial Economics* 15, 373–94.

Fan, J., Fan, Y. and Lv, J. 2006. High dimensional covariance matrix estimation using a factor model. Unpublished manuscript, Princeton University.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. 2000. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* 82, 540–54.

Geweke, J. and Singleton, K.J. 1981. Maximum likelihood confirmatory factor analysis of economic time series. *International Economic Review* 22, 37–54.

Lawley, D.N. and Maxwell, A.E. 1971. *Factor Analysis as a Statistical Method*. London: Butterworth.

Ledoit, O. and Wolf, M. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10, 603–21.

Ledoit, O. and Wolf, M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88, 365–411.

Lintner, J. 1965. Security prices, risk, and maximal gains from diversification. *Journal of Finance* 20, 587–615.

Mardia, K.V., Kent, J.T. and Bibby, J.M. 1979. *Multivariate Analysis*. New York: Academic Press.

Moon, H.R. and Perron, B. 2004. Testing for unit root in panels with dynamic factors. *Journal of Econometrics* 122, 81–126.

Phillips, P.C.B. and Sul, D. 2003. Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal* 6, 217–59.

Quah, D. and Sargent, T. 1993. A dynamic index model for large cross sections. In *Business Cycles, Indicators and Forecasting*, ed. J. Stock and M. Watson. Chicago: University of Chicago Press.

Ross, S. 1976. The arbitrage theory of capital asset pricing. *Journal of Finance* 13, 341–60.

Sargent, T. and Sims, C. 1977. Business cycle modelling without pretending to have too much a priori economic theory. In *New Methods in Business Cycle Research*, ed. C. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Stock, J.H. and Watson, M.W. 2002a. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–79.

Stock, J.H. and Watson, M.W. 2002b. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20(2), 147–62.

How to cite this article

Bai, Jushan. "factor models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000309> doi:10.1057/9780230226203.0543

The New Palgrave Dictionary of Economics Online

factor price equalization (historical trends)

Kevin H. O'Rourke
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Heckscher–Ohlin prediction that international trade should lead to relative factor prices converging internationally is one that receives abundant empirical support for the period that was the focus of interest for these two economists, namely the 'long 19th century'. In labour-abundant regions, wage–rental ratios increased, whereas they declined in land-abundant countries.

Keywords

factor price equalization; Heckscher, E. F.; Heckscher–Ohlin trade theory; Ohlin, B. G.; skill-biased technical change; transportation costs; wage–rental ratio

Article

When 21st-century undergraduate economists are taught trade theory, they inevitably encounter the standard Heckscher–Ohlin trade theory, which was for many years, and perhaps still is, the workhorse of international trade theory. The '2×2×2' version of the theory which they first study surely strikes many of them as unrealistic in the extreme, with its strong predictions of factor price equalization, which logically follows when both countries produce both goods using the same technology. However, the origins of the theory lie in the attempts of two Swedish economists, Eli Heckscher (who was an economic historian) and Bertil Ohlin, to understand the world around them, and in particular to make sense of the global economy of the late 19th century. Not surprisingly, perhaps, their theoretical predictions find ample empirical support in the historical records of that time.
Bertil Ohlin presented the theory as follows:

> Australia has a small population and an abundant supply of land, much of it not very fertile. Land is consequently cheap and wages high, in relation to most other countries. It would therefore seem profitable to produce goods requiring large areas of less fertile land but relatively little labour. Such is the case, for example, in wool production … Similarly, regions well endowed with technically trained labor and capital will specialize in industrial production … Exports from one region to the other will on the whole consist of goods that are intensive in those factors with which this region is abundantly endowed and the prices of which are therefore low … In short, commodities that embody large quantities of particularly scarce factors are imported, and commodities intensive in relatively abundant factors are exported. … Australia exchanges wool and wheat for industrial products since the former embody much land and little labour while the opposite is true of industrial products. Australian land is thus exchanged for European labor. (Flam and Flanders, 1991, p. 90)

He then argued that the level of trade integration helped determine factor prices in both regions:

> If, for example, Australia produced its own industrial products rather than importing them from Europe and America in exchange for agricultural products, then, on the one hand, the demand for labor would be greater and wages consequently higher, and on the other the demand for land, and therefore rent, lower than at present. At the same time, in Europe the scarcity of land would be greater and that of labor less than at present if the countries of Europe were constrained to produce for themselves all

> their agricultural products instead of importing some of them from abroad. Thus trade increases the price of land in Australia and lowers it in Europe, while tending to keep wages down in Australia and up in Europe. The tendency, in other words, is to approach an equalization of the prices of productive factors. (Flam and Flanders, 1991, pp. 91–2)

Three points should be noted about these quotations. First, Ohlin presented the theory using an example that seems to lend itself more easily to formalization via a three-factor, two-good model (in which land, labour and capital produce agricultural and industrial products) than to formalization via the 2×2 framework that is often associated with Heckscher–Ohlin theory today. Second, he speaks of a ‘tendency’ to ‘approach an equalization’ of factor prices, but not of factor price equality per se. That is, his prediction is that there would be factor price *convergence*. Third, the metaphor that motivated him was one that reflected the international economy of the late 19th century, in which intercontinental trade flows for the most part reflected an exchange of resource-intensive products coming from the New World, but also from resource-abundant regions in Asia and Africa, for labour-intensive (and also capital-intensive) manufactured goods produced in western Europe and parts of North America (Findlay and O’Rourke, 2007, ch. 7).

The 19th century, and particularly the period from roughly 1840 onwards, offers the perfect context in which to study the empirical relevance of such a theory, for it was a period that saw a dramatic, worldwide decline in transport costs (O’Rourke and Williamson, 1999, ch. 3). For example, Knick Harley's (1988) index of British ocean freight rates declines by about 70 per cent between 1840 and 1910, after having remained roughly constant for a century or so. Table 1 presents freight factors (that is to say, transport costs as a percentage of either the import or the export price of a commodity) for several commodities and routes, and the picture which emerges is one of sharply falling transport costs on many routes. The implication is that the relative prices of imported goods should have been steadily declining across continents, as commodity market integration lowered intercontinental price gaps. Furthermore, these declining transport costs were linking continents with very different factor endowments, implying that there should have been scope for trade to have had the sort of impact on factor prices that Heckscher and Ohlin said it should.

Freight factors, 1820–1910 (per cent)

Commodity	From	To	Basis	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910
Wheat	Baltic	UK	Import	8.0	7.1	7.2		6.8	9.6	4.5	3.5	5.9	3.4
Wheat	Black Sea	UK	Import	15.5	16.3			15.0	17.3	9.2	9.7	10.8	6.8
Wheat	East coast, USA	UK	Import		10.3		7.5	10.9	8.1	8.6	5.0	8.2	3.2
Wheat	New York	UK	Export				10.5					6.9	
Wheat	New York	UK	Import				9.4					6.2	
Wheat	Chicago	UK	Export						33.0	21.7	13.3	15.9	7.4
Wheat	South America	UK	Import								15.6	18.5	7.4
Wheat	Rio de la Plata	UK	Import								15.4		6.9
Wheat	Australia	UK	Import								22.3	26.7	15.4
Coal	Britain	Genoa	Export			213.1	224.5	246.1	194.0	163.1	69.7	64.5	53.8
Coal	Nagasaki	Shanghai	Export							84.0	57.0	35.0	20.0
Copper ore	West coast, S. America	UK	Import					21.3			7.8		
Guano	West coast, S. America	UK or European Continent	Import					24.9			18.5		
Nitrate	West coast, S. America	UK or European Continent	Import					34.1			23.0		9.7
Coffee	Brazil	UK or European Continent	Import					5.2			2.0		1.5
Salted hides	Rio de la Plata	UK	Import					3.1			3.8		
Wool	Rio de la Plata	UK	Import					1.3			1.3		
Rice	Rangoon	Europe	Export							73.8			18.1

Source: Findlay and O’Rourke (2007, Table 7.2).

In this historical context, one key prediction of the theory is as follows. In labour-abundant regions, such as the crowded countries of Europe, declining transport costs should have led to the relative price of agricultural commodities falling, as they were imported from land-abundant regions, and thus to the ratio of wages to land rents rising. In land-abundant countries, such as the frontier societies of the New World, declining transport costs should have led to the rise of relative price of agricultural commodities, and thus to a fall in wage–

rental ratios. Economic historians have examined this prediction at great length. O'Rourke, Taylor and Williamson (1996) presented evidence for seven affluent 'Atlantic economy' economies, while more recently Jeffrey Williamson, in a series of papers summarized in Williamson (2002), has expanded the work to include data for several developing economies. By and large, the Heckscher–Ohlin predictions hold good for the late 19th century, both for western Europe and the New World, and for those Third World countries that participated in the late 19th-century global economy (O'Rourke and Williamson, 1999, ch. 4; Williamson 2002, Table 3, p. 73). In land-scarce economies such as those of Japan, Korea, Taiwan or the United Kingdom, the wage–rental ratio increased substantially; while it fell sharply in land-abundant food exporting nations and regions such as Argentina, Uruguay, Burma, Siam, Egypt, the United States, Canada, Australia and the Punjab (see Table 2).

Wage–rental ratio trends, 1870–1914 (1911=100)

Period	1870–1874	1875–1879	1880–1884	1885–1889	1890–1894	1895–1899	1900–1904	1905–1909	1910–1914
Land-abundant countries or regions									
Argentina	•		580.4	337.1	364.7	311.1	289.8	135.2	84
Australia	•416.2	253	239.1	216.3	136.2	147.7	130	97.9	100.6
Burma	•				190.9	189.9	186.8	139.4	106.9
Egypt	•196.7	174.3	276.6	541.9	407.5	160.1	166.7	64.4	79.8
Punjab	•	198.5	147.2	150.8	108.7	92	99.8	92.4	80.1
Siam	•4699.1	3908.7	3108.1	2331.6	1350.8	301.3	173	57.2	109.8
USA	•233.6	195	188.3	182.1	173.5	175	172.4	132.7	101.1
Uruguay	•1112.5	891.3	728.3	400.2	377.2	303.6	233	167.8	117.9
Land-scarce countries									
Britain	•56.6	61.4	64.9	73.1	79.1	87.3	91.4	98.1	102.7
Denmark	•44.8	43.5	44.8	56.6	66.7	87.9	103.8	99.7	100
France	•63.5	62.9	67.3	73.8	80.4	91.8	103.2	106.4	99.8
Germany	•84.4	80	82.3	86	98	108.2	107.6	104.6	100.2
Ireland	•51.3	62.2	72.7	86.4	102.7	122.1	111.2	101.7	94.1
Japan	•			79.9	68.6	91.3	96.1	110.4	107.5
Korea	•							102.8	121.9
Spain	•42.7	55.8	58.6	73	81.8	85.5	74.9	85.7	86.4
Sweden	•	43.7	50.7	57.8	65.3	78.6	87.9	92.5	99.1
Taiwan	•						68.1	85.2	96.6

Source: Williamson (2002, Tables 3 and 4, pp. 73–4).

Of course, the fact that wage–rental ratios were systematically trending upwards in labour-abundant economies, and downwards in land-abundant economies, cannot be taken as proof that the Heckscher–Ohlin theory was at work, any more than rising skill premia in many Organisation for Economic Co-operation and Development (OECD) economies can be taken as evidence of factor price equalization today, in 2007. As today's debate about 'trade and wages' suggests, other forces might be at work driving up the ratio of skilled to unskilled wages in skill-abundant economies, most notably perhaps technological change biased in favour of skilled workers. (Recent contributions to the literature on this controversy include Collins, 1998; Feenstra, 2000; and Feenstra and Hanson, 2004.) For the late 19th century, both econometrics and simulation exercises indicate that the wage–rental ratio trends documented in Table 2 were indeed in part due to Heckscher–Ohlin forces. That is to say, the price of agricultural products relative to manufactured products was negatively related to the wage–rental ratio during this period (O'Rourke and Williamson, 1994; 1999; O'Rourke, Taylor and Williamson, 1996). It is noticeable that wage–rental ratios increased by less in protectionist economies such as those of France and Germany than in the free-trading United Kingdom. Further evidence in favour of this Heckscher–Ohlin interpretation of 19th-century distributional trends comes from a comparison of the pre-1800 and 19th-century periods (O'Rourke and Williamson, 2005). Before 1800, British land–labour ratios were trending downwards, as population expanded but land supplies remained relatively constant. In a closed economy setting, this would be expected to lead to an increase in the relative price of agricultural commodities, and to a decline in wage–rental ratios; and this is indeed what happened (Figure 1, Panel A). From 1840 onwards, by contrast, relative agricultural

prices stopped rising, and eventually started falling, while the wage–rental ratio stopped falling and started rising (Panel B). This switch occurred despite an acceleration in British population growth, and is consistent with a British economy opening up to trade and becoming more exposed to the factor price convergence forces identified by Heckscher and Ohlin.

Figure 1

Endowments and relative prices, Britain 1500–1936 (1900=100). *Source*: O'Rourke and Williamson (2005).

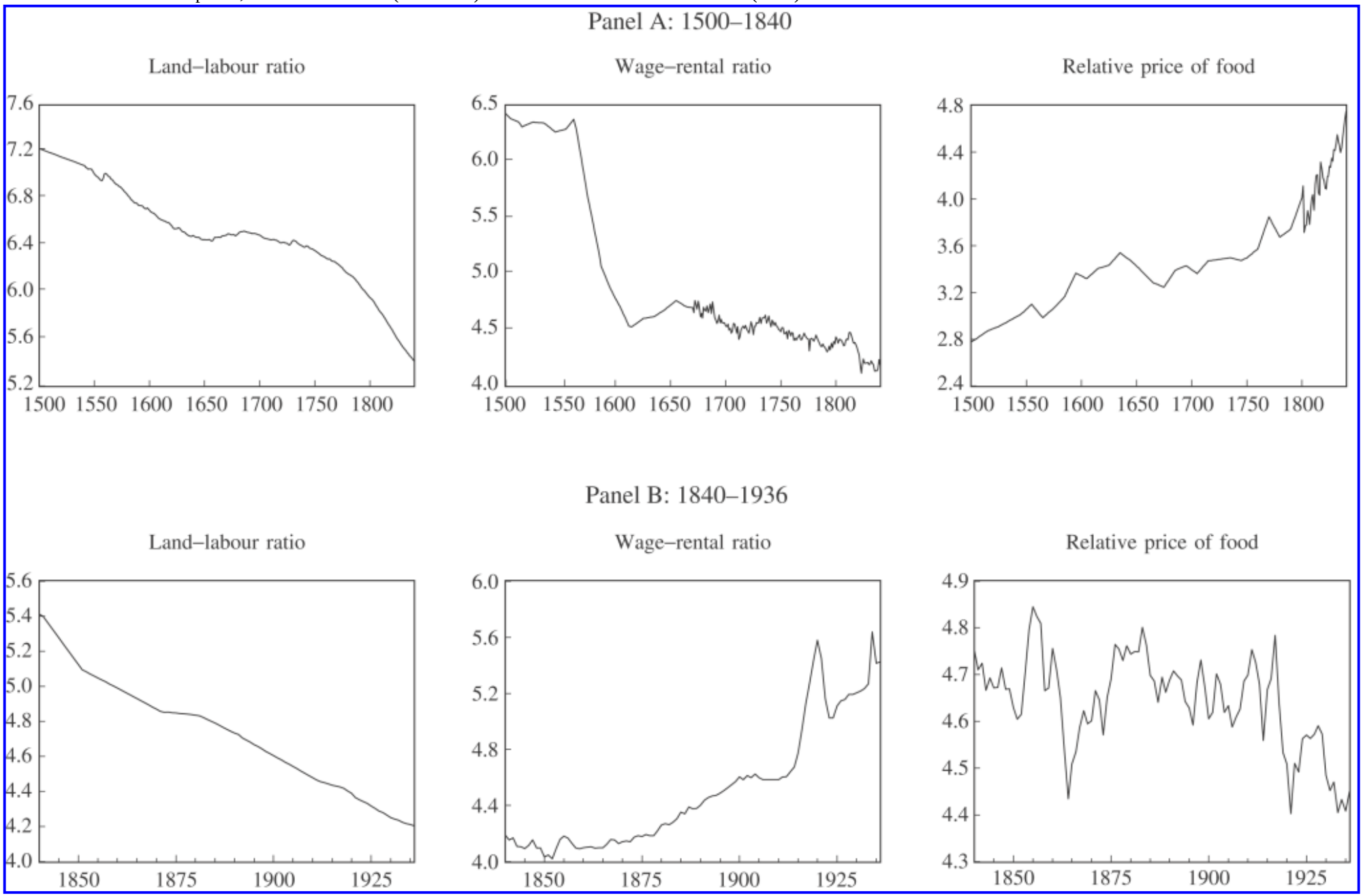

Factor prices were certainly not equalized during the late 19th century, any more than they have been equalized today. But they converged between continents, in precisely the manner envisaged by Heckscher and Ohlin.

See Also

- cliometrics
- Heckscher–Ohlin trade theory
- international trade theory

Bibliography

Collins, S.M. 1998. *Imports, Exports, and the American Worker*. Washington DC: Brookings Institution Press.

Feenstra, R.C. 2000. *The Impact of International Trade on Wages*. Chicago: University of Chicago Press and NBER.

Feenstra, R.C. and Hanson, G.H. 2004. Global production sharing and rising inequality: a survey of trade and wages. In *Handbook of International Trade*, vol. 1, ed. E. Kwan Choi and J. Harrigan. Oxford: Blackwell.

Findlay, R. and O'Rourke, K.H. 2007. *Power and Plenty: Trade, War and the World Economy in the Second Millennium*. Princeton: Princeton University Press.

Flam, H. and Flanders, M.J. 1991. *Heckscher–Ohlin Trade Theory*. Cambridge, MA: MIT Press.

Harley, C.K. 1988. Ocean freight rates and productivity, 1740–1913: the primacy of mechanical invention reaffirmed. *Journal of Economic History* 48, 851–76.

O'Rourke, K.H., Taylor, A.M. and Williamson, J.G. 1996. Factor price convergence in the late nineteenth century. *International Economic Review* 37, 499–530.

O'Rourke, K.H. and Williamson, J.G. 1994. Late 19th century Anglo-American factor price convergence: were Heckscher and Ohlin right? *Journal of Economic History* 54, 892–916.

O'Rourke, K.H. and Williamson, J.G. 1999. *Globalization and History: The Evolution of a Nineteenth-Century Atlantic Economy*. Cambridge, MA: MIT Press.

O'Rourke, K.H. and Williamson, J.G. 2005. From Malthus to Ohlin: trade, industrialisation and distribution since 1500. *Journal of Economic Growth* 10, 5–34.

Williamson, J.G. 2002. Land, labor and globalization in the third world, 1870–1940. *Journal of Economic History* 62, 55–85.

How to cite this article

O'Rourke, Kevin H. "factor price equalization (historical trends)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000322> doi:10.1057/9780230226203.0544

The New Palgrave Dictionary of Economics Online

factor price frontier

Heinz D. Kurz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The constraint binding changes in the distributive variables, in particular the real wage rate (w) and the rate of profit (r), was discovered (though not consistently demonstrated) by Ricardo: ‘The greater the portion of the result of labour that is given to the labourer, the smaller must be the rate of profits, and vice versa’ (Ricardo, 1971, p. 194). He was thus able to dispel the idea, generated by Adam Smith's notion of price as a sum of wages and profits, that the wage and the rate of profit are determined *independently* of each other. Ever since the inverse relationship between the distributive variables played an important role in long-period analysis of both classical and neoclassical descent. In more recent times it was referred to by Samuelson (1957), who later dubbed it ‘factor price frontier’ (cf. Samuelson, 1962). Hicks (1965, p. 140, n.1) objected that this term is unfortunate, since it is the earnings (quasi-rents) of the (proprietors of) capital goods rather than the rate of profit which is to be considered the ‘factor price’ of capital (services). A comprehensive treatment of the problem under consideration within a classical framework of the analysis, including joint production proper, fixed capital and scarce natural resources, such as land, was provided by Sraffa (1960). The relationship is also known as the ‘wage frontier’ (Hicks, 1965), the ‘optimal transformation frontier’ (Bruno, 1969) and the ‘efficiency curve’ (Hicks, 1973).
The duality of the w–r relationship and the c–g relationship, that is, the relationship between the level of consumption output per worker (c) and the rate of growth (g) in steady-state capital theory has been demonstrated by the latter two authors and in more general terms by Burmeister and Kuga (1970); for a detailed account, see Craven (1979).
To begin with, suppose for simplicity that there are only single-product industries with labour as the only primary input and that only one (indecomposable) system of production is known (cf. Sraffa, 1960, Part I). Then, with gross outputs of the different products all measured in physical terms and made equal to unity by choice of units and with wages paid at the end of the uniform production period, we have the price system.

$$p = (1 + r)ap + wa_0, \tag{1}$$

where p is the column vector of normal prices, a is the square matrix of material inputs, which is assumed to be productive, and a_0 is the column vector of direct labour inputs. Using the consumption basket d as standard of value or *numéraire*,

$$dp = 1, \tag{2}$$

we can derive from (1) and (2) the w–r relationship for system (a, a_0)

$$w = \{d[I - (1 + r)a]^{-1}a_0\}^{-1} \quad (3)$$

The relationship is illustrated in Figure 1. At r=0 the real wage in terms of d is at its maximum value W; it falls monotonically with increases in r, approaching zero as r approaches its maximum value R. (The w–r relationship can be shown to be a straight line if Sraffa's Standard commodity s is used as *numéraire*, where s is a row vector such that s=(1+R) sa; cf. Sraffa, 1960, chap. IV.)

Figure 1

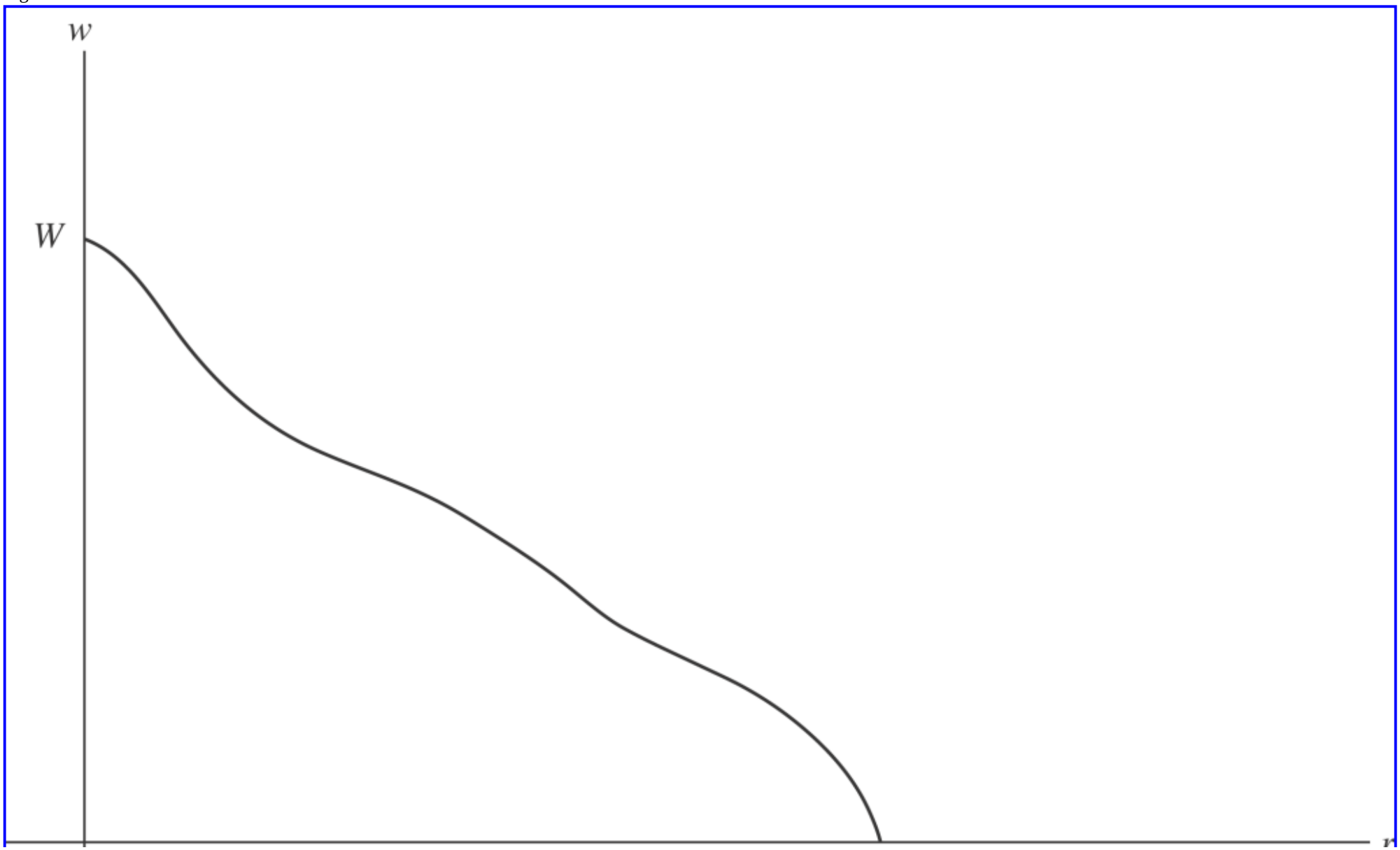

Let us now assume that several systems are available for the production of the different commodities and that all the production processes exhibit constant returns to scale. We call the set of all the alternative methods (or processes) of production known the *technology* of the economic system. From this set a series of alternative *techniques* can be formed by grouping together these methods of production, one for each commodity. Hence there is the question of the *choice of technique*. Under competitive conditions this choice will be exclusively grounded on cheapness, that is, the criterion of choice is that of *cost-minimization*. In the case depicted, it can be shown that the competitive tendency of entrepreneurs to adopt whichever technique is cheapest in the existing price situation, will for a given w (or, alternatively, r) lead to the technique yielding the highest $r(w)$, whereas techniques yielding the same $r(w)$ for the same $w(r)$ are equiprofitable and can co-exist (cf. Garegnani, 1970, p. 411).

What has just been said is illustrated in Figure 2. It is assumed that only three alternative techniques, α β and γ, are available, each of which is represented by the associated w–r relationship; since w is always measured in terms of the consumption basket d, all three relationships can be drawn in the same diagram. Obviously, technique γ is inferior and will not be adopted. Technique α will be chosen for $0<w<w_1$ and $w_2<w\leq w_\alpha$, while technique dominates at $w_1<w<w_2$; there are two *switch points* (at $w=w_1$ and $w=w_2$, respectively) at which both techniques are equiprofitable. The heavy line represents the economy's *w–r frontier* (or ‘factor price frontier’) and is the outer envelope of the w–r relationships. At a level of the wage rate w^*, for example, technique β will be adopted giving a rate of profit r^*. (For a discussion of more general cases of single production, see Pasinetti, 1977, ch. VI; for a reformulation of some results in capital theory in terms of the so-called ‘dual’ cost and profit functions, see Salvadori and Steedman, 1985; on the maximum number of switch points between two production systems, see Bharadwaj, 1970.)

Figure 2

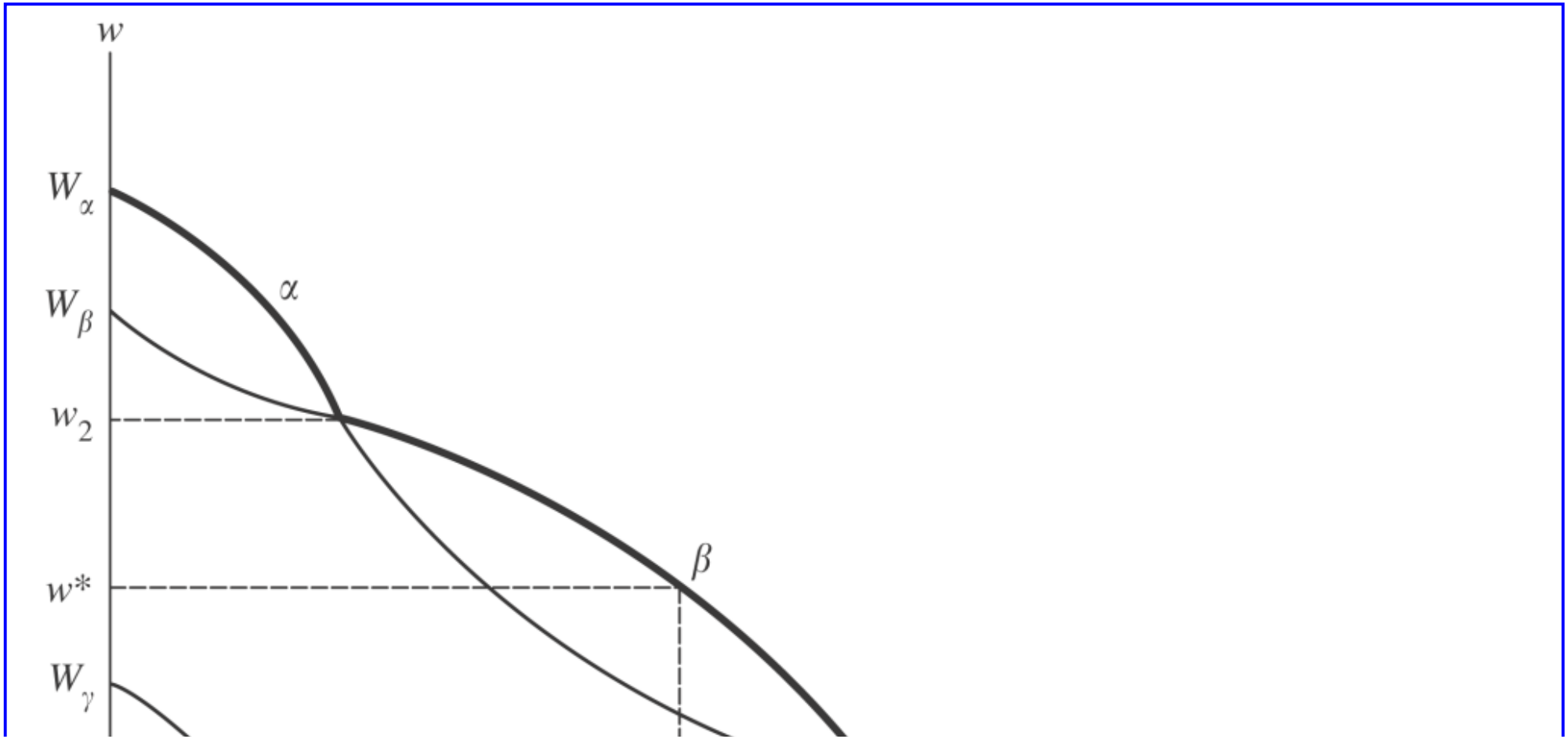

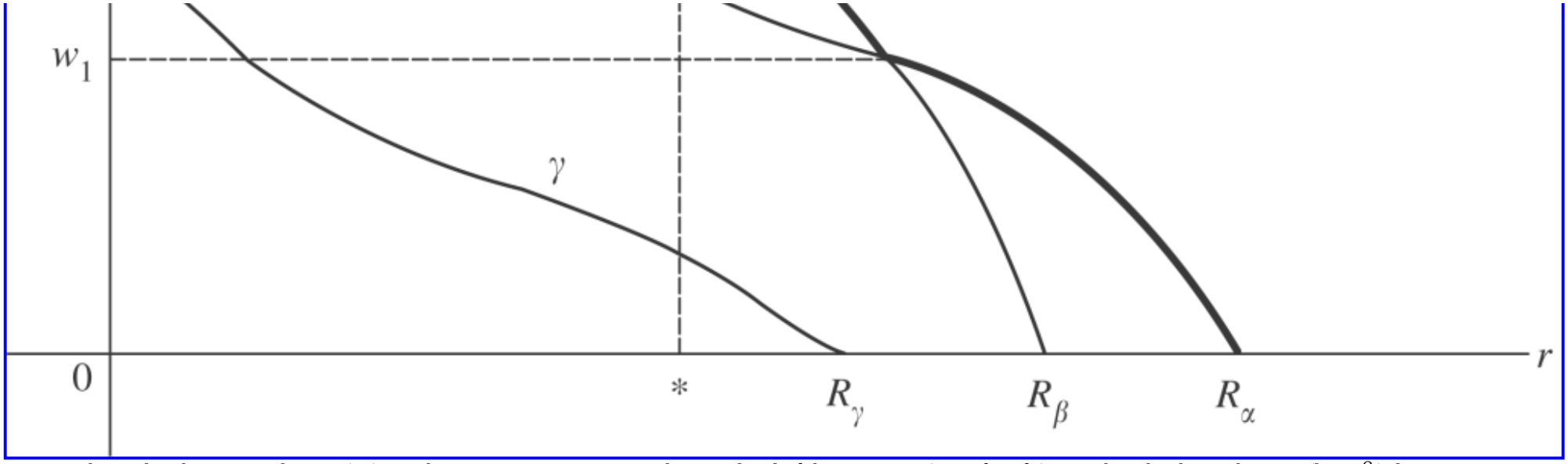

Figure 2 shows that the same technique (α) may be costminimizing at more than one level of the wage rate (rate of profit) even though other techniques (here β) dominate at wage rates in between. The implication of this possibility of the *reswitching* of techniques (and of the related possibility of *reverse capital deepening*) is that the direction of change of input proportions cannot be related unambiguously to changes in the distributive variables. This can be demonstrated by making use of the duality between the *w–r* and the *c–g* frontier. Denoting the value of net output per labour unit by *y* and the value of capital per labour unit by *k*, we have in steady-state equilibrium

$$y = w + rk = c + gk. \tag{4}$$

Solving for *k* we get

$$k = (c - w) / (r - g) \tag{5}$$

except in golden rule equilibrium (*g*=*r*), where *k* can be shown to be (minus) the slope of the golden rule *w–r* relationship at the going level of *r*. In Figure 3(a) the frontier built of two techniques, α and β , is depicted. The rate of growth is fixed at the level $\bar{g}$, to which correspond c_α , and c_β . For values of $r \geq \bar{g}$, that is, on the right side of the golden rule, Fig. 3(b) gives the corresponding value of $k\,(r,\bar{g})$. For example, at $\hat{w}$ technique β will be chosen, yielding a rate of profit $\hat{r}$ the associated capital intensity is given by

$$\tan e = (c_\beta - \hat{w}) / (\hat{r} - \hat{g}) = \hat{k}.$$

Figure 3

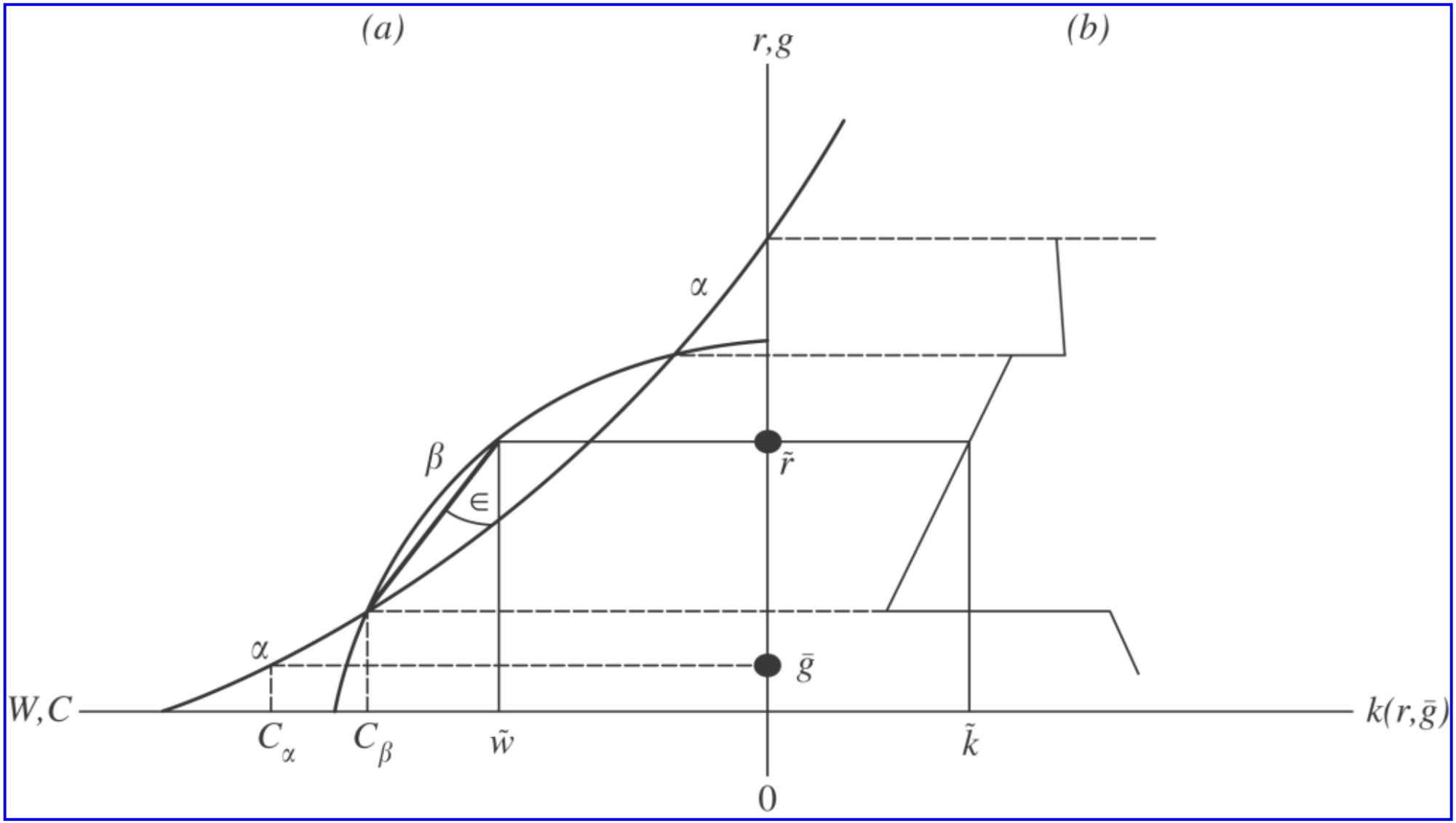

Figure 3(b) shows that the capital–labour ratio need not be inversely related to the rate of profit as neoclassical long-period theory maintained. In more general terms, it cannot be presumed that input uses, per unit of output, are related to the corresponding ‘factor prices’ in the conventional way (see Metcalfe and Steedman, 1972, and Steedman, 1985). This result calls in question the validity of the traditional demand and supply approach to the determination of quantities, prices and income distribution.

The results stated above essentially carry over to the more general case with fixed capital, pure joint production and several primary inputs, such as land and labour of different qualities, provided the formalization of the problem is appropriately adapted to the specific case under consideration. Here it suffices to point out a few additional aspects of the choice of technique problem.

With fixed capital there is always such a problem to be solved. This concerns both the choice of the system of operation of plant and equipment, that is, for example, whether a single or a double-shift system is to be adopted; and the choice of the economic lifetimes of fixed capital goods. During the capital theory debates of the 1960s and early 1970s attention focussed on the latter aspect of the use of capital. It was shown that with decreasing or changing efficiency of the durable capital good, cost minimization implies that for a given level of the rate of profit, premature truncation is advantageous as soon as the price (book value) of the partly worn out item becomes negative. While the *w–r* relationship for a given truncation may slope upwards over some range of *r*, the *w–r* frontier consists only of those parts of the *w–r* relationships that are downward-sloping. Moreover, it was demonstrated that the frontier can display the *return of the same truncation* (cf., for example, Hagemann and Kurz, 1976). As to the other aspect of capital utilization, a similar possibility can be

shown to exist: the *return of the same system of operation of plant and equipment* (cf. Kurz, 1986). Both phenomena are of course variants of the reswitching of techniques.

In systems with pure joint production a choice of technique is inherent, even where the number of processes available does not exceed the number of products. Sraffa's approach to joint production is in terms of ‘square’ systems of production, that is, systems where the number of processes operated is equal to the number of commodities (i.e. positively-priced products). However, as Salvadori (1982) has shown, in such a framework a cost-minimizing system does not need to exist. A way out of this impasse may be seen in a formalization of joint production that is similar to von Neumann's. In such a formalization the free disposal assumption plays a crucial role. It can be shown that the *w–r* frontier is downward-sloping, even though individual *w–r* relationships may have positive ranges.

See Also

- reswitching of technique
- Sraffian economics
- two-sector models

Bibliography

Bharadwaj, K. 1970. On the maximum number of switches between two production systems. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 106(December), 409–29.

Bruno, M. 1969. Fundamental duality relations in the pure theory of capital and growth. *Review of Economic Studies* 36(January), 39–53.

Burmeister, E. and Kuga, K. 1970. The factor price frontier, duality and joint production. *Review of Economic Studies* 37(January), 11–19.

Craven, J. 1979. Efficiency curves in the theory of capital: a synthesis. In *The Measurement of Capital, Theory and Practice*, ed. K.D. Patterson and K. Schott. London: Macmillan.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37(July), 407–36.

Hagemann, H. and Kurz, H.D. 1976. The return of the same truncation period and reswitching of techniques in neo-Austrian and more general models. *Kyklos* 29(December), 678–708.

Hicks, J.R. 1965. *Capital and Growth*. Oxford: Oxford University Press.

Hicks, J.R. 1973. *Capital and Time, A Neo-Austrian Theory*. Oxford: Oxford University Press.

Kurz, H.D. 1986. ‘Normal’ positions and capital utilisation. *Political Economy* 2(May), 37–54.

Metcalfe, J.S. and Steedman, I. 1972. Reswitching and primary input use. *Economic Journal* 82(March), 140–57.

Pasinetti, L.L. 1977. *Lectures on the Theory of Production*. London: Macmillan.

Ricardo, D. 1971. *The Works and Correspondence of David Ricardo*, Edited by P. Sraffa in collaboration with M.H. Dobb, Vol. VIII. Cambridge: Cambridge University Press.

Salvadori, N. 1982. Existence of cost-minimizing systems within the Sraffa framework. *Zeitschrift für Nationalökonomie* 42(September), 281–98.

Salvadori, N. and Steedman, I. 1985. Cost functions and produced means of production: duality and capital theory. *Contributions to Political Economy* 4(March), 79–90.

Samuelson, P.A. 1957. Wages and interest: a modern dissection of Marxian economic models. *American Economic Review* 47(December), 884–912.

Samuelson, P.A. 1962. Parable and realism in capital theory: the surrogate production function. *Review of Economic Studies* 29(June), 193–206.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1985. On input ‘demand curves’. *Cambridge Journal of Economics* 9(June), 165–72.

How to cite this article

Kurz, Heinz D. "factor price frontier." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000006>
doi:10.1057/9780230226203.0545

The New Palgrave Dictionary of Economics Online

factor prices in general equilibrium

Michael Mandler

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In general equilibrium models with linear or nonlinear activities, factor prices can be indeterminate and agents will have an incentive to non-competitively manipulate prices even if they are small relative to the market. The indeterminacy cannot occur at generic endowments, but the non-generic endowments where it does occur will arise endogenously as an equilibrium outcome when some factors, such as capital goods, are produced. This endogenous indeterminacy creates a hold-up problem since investors need not earn the rate of return that obtains in an intertemporal competitive equilibrium. Unlike the classical hold-up problem, factor-price indeterminacy is not attributable to there being few agents or bilateral monopoly.

Keywords

Arrow–Debreu model of general-equilibrium; differentiable production function; excess demand and supply; factor prices in general equilibrium; factor-price indeterminacy; hold-up; imperfect competition; intertemporal efficiency; intertemporal equilibrium; Leontiev production function; linear activities model; non-market institutions; regular economies; sequential-trading equilibrium; uniqueness of equilibrium; Walras's Law

Article

1 Introduction

At first glance, the Walrasian general equilibrium model does not offer a theory of factor prices. Factors are goods supplied by agents to firms which then use them to produce outputs. In the general equilibrium model, there is no such class of goods: one and the same good can simultaneously be used as an input by some firms, produced as an output by other firms, sold by some consumers, and purchased and consumed by other consumers. Indeed, the general equilibrium model's abstraction from

the minutiae of how particular goods are used is one of the theory's great advantages. For many of the classical concerns of the Walrasian tradition – the existence of equilibrium, optimality – these details are irrelevant.

Even if there is a category of factors that consumers sell and firms buy, it is hard to see any distinctive properties of these goods. While factor supply functions can exhibit perverse responses to price changes, so can output demand functions. The responses of firms to price changes are better behaved, and firm factor demands may seem to be governed by a distinctive principle: a firm's demand for a factor diminishes in its own price while a firm's supply of an output increases in its own price. While correct, these two fundamental rules of producer comparative statics are really reflections of a single law, as Samuelson (1947) showed long ago. Suppose in an ℓ-good economy that a profit-maximizing firm with production set $Y \subset R^{\ell}$ chooses $y = (y_1, \ldots, y_\ell) \in Y$ when facing prices $p = (p_1, \ldots, p_\ell)$ and $\hat{y} \in Y$ when facing $\hat{p}$. Since each decision is profit-maximizing, $p \cdot y \geq p \cdot \hat{y}$ and $\hat{p} \cdot \hat{y} \geq \hat{p} \cdot y$ and hence $(\hat{p} - p) \cdot (\hat{y} - y) \geq 0$. If only one price differs at *p* compared to $\hat{p}$, say the first, then $(\hat{p}_1 - p_1)(\hat{y}_1 - y_1) \geq 0$. So if $\hat{p}_1 > p_1$ then $\hat{y}_1 \geq y_1$. Both of the comparative statics rules now follow from the appropriate sign restrictions on y_1 and $\hat{y}_1$: when both are positive we conclude that the output of good 1 supplied by the firm must be weakly increasing in its price, while if both are negative we conclude that the factor demand for good 1 must be weakly decreasing in its price (since $\hat{y}_1 \geq y_1$ and $(\hat{y}_1, y_1) \leq 0$ imply $|\hat{y}_1| \leq |y_1|$). It is tempting to conclude that there is no special general equilibrium principle of factor demands, just a specific application that follows when the sign convention for factors is inserted.

2 Factor-price indeterminacy

The demand for and supply of factors can nevertheless exhibit distinctive properties, although they are consistent with the generalities pointed out in the previous section. These properties do not matter for the most of the classical results of general equilibrium theory, but they can undermine one result, the generic determinacy (local uniqueness) of equilibria.

The first distinguishing trait of factors is that sometimes they do not provide any direct utility and are useful only as inputs in production. Consumers will supply to the market their entire endowment of such ‘pure’ factors and hence supply will be inelastic with respect to price changes. As we will see, what matters is local unresponsiveness to prices. Perhaps when a factor such as iron ore is sufficiently cheap in terms of consumption goods consumers will find some direct use for it and hence have an excess demand that locally varies as a function of prices. But above some minimum price, consumers will not consume any iron ore and in this range consumers’ excess demand will be inelastic. Second, technology can restrict the number of ways in which factors can be productively combined. The extreme case occurs with fixed coefficients – the Leontiev production function – where to produce one unit of a good just one combination of factors will do. More flexible is the linear activities model where finitely many constant-return-to-scale techniques are available to produce one or more goods. Factors then may be combined in various configurations but some factor proportions cannot be used productively (that is, without disposing of some of the factors). Nonlinear activities are qualitatively similar but do not require constant returns to scale. In all these cases, production sets have a kinked rather than smooth

(differentiable) surface. Consequently factor prices can be adjusted at least slightly from one equilibrium configuration without changing the quantity of factors that profit-maximizing producers will demand when producing a given quantity of output (or vector of outputs). In the Leontiev case, picture the multiple price lines that can support the model's L-shaped isoquants. Of course, production sets do not have to exhibit kinks; for example, they will be smooth when each output is a differentiable function of factor inputs. Any change in relative factor prices will then lead to a change in factor demand.

Factors of production thus are distinctive in that both demand and supply can be unresponsive to certain types of price changes. Factor demand and supply do not have to display this unresponsiveness, but under plausible circumstances permitted by the general equilibrium model they will. Inelastic factor demand and supply in turn can lead to an indeterminacy of factor prices. For a simple example, suppose an economy has one consumption good, produced by a single linear activity that requires a_1 units of one factor and a_2 units of a second factor to yield one unit of output. Set the price of consumption equal to 1, let w_1 and w_2 be the two factor prices, let the endowments of the two inelastically supplied factors be $e_1 \geq 0$ and $e_2 \geq 0$, and let y be the sole activity usage level. An equilibrium $(w_1, w_2, y) \geq 0$ where the consumption good is produced and has a positive price must satisfy three conditions: (i) $a_1w_1+a_2w_2=1$ (the activity breaks even), (ii) $a_iy \leq e_i$ for i=1, 2 (market-clearing for factors), and (iii) $a_iy < e_i \Rightarrow w_i = 0$ for i=1, 2 (factors in excess supply have a 0 price). On the assumption that the demand for output equals factor income, which is a form of Walras's law, (i)–(iii) imply that the market for output clears. Evidently equilibrium must satisfy $y = \min[e_1/a_1, e_2/a_2]$. By (iii), the two factors will both have a strictly positive price only if

$$\frac{e_1}{a_1} = \frac{e_2}{a_2}, \tag{1}$$

in which case any $w = (w_1, w_2) \geq 0$ that satisfies (i) will be an equilibrium w: indeterminacy therefore obtains when (1) holds. We defer for a little while the question of whether this knife-edge condition is likely to be satisfied.

Fixed coefficients and inelastic factor supply do not always lead to indeterminate factor prices. Prior to the invention of the differentiable production function and for a while thereafter, the standard cure for factor-price indeterminacy was to argue that, even if each industry uses factors in fixed proportions, those proportions will differ across industries; variations in factor prices will then lead to changes in relative output prices, and thus to changes in output demand that feedback to changes in factor demand (Cassel, 1924; Wieser, 1927). Substitution in consumption can thereby play the same equilibrating role as the technological substitution of inputs in production. For the simplest example, suppose we supplement the above single-sector economy with a new sector that uses b_1 units of the first factor and b_2 units of the second factor to produce one unit of a second consumption good. If we keep the price of the first consumption good equal to 1, and let p_b be the price of the second consumption good, then when both activities break even the equalities

$$a_1 w_1 + a_2 w_2 = 1, \quad b_1 w_1 + b_2 w_2 = p_b$$

must be satisfied. As long as $a_1 / a_2 \neq b_1 / b_2$, it will not be possible to adjust *w* without also changing the relative price of the consumption goods p_b. When w_1 / w_2 increases, the consumption good that uses factor 1 more intensively will rise in price, presumably diminishing demand for that good and thus diminishing the demand for factor 1. Even if demand for consumption is a perverse function of prices, this two output–two factor model will still typically have determinate prices as long as both activities break even.

A general linear activity analysis model will clarify when the determinate and indeterminate cases arise. The linearity of the activities serves only to simplify the model's equilibrium conditions. There will be two types of goods: factors, which give no utility and are inelastically supplied, and consumption goods, which do give utility. Despite their name, consumption goods can be used as inputs and nonproducible but they must provide utility to some agents. We now adopt the standard sign convention and define an activity to be a vector, with as many coordinates as there are commodities, whose positive coordinates give the quantities of goods produced and negative coordinates give the quantities of goods used when the activity is operated at the unit level. In equilibrium the excess demand for each good must be non-positive, each good in excess supply must have a 0 price, each activity must earn non-positive profits, and each activity in use must earn 0 profits. Since determinacy and indeterminacy are purely local events, a search for equilibrium prices and activity near a reference equilibrium can ignore the ‘slack’ equilibrium conditions, the market-clearing condition for any good in excess supply and the no-positive-profits condition for any activity that either makes strictly negative profits or utilizes and produces only goods in excess supply: for small adjustments of prices and activity levels, the excluded goods will remain in excess supply and the excluded activities will continue to make negative profits or continue to use and produce only goods in excess supply (and hence continue to break even). Call any good not in excess supply and any activity that breaks even and that uses or produces at least one good not in excess supply ‘operative’. Given some reference equilibrium with ℓ operative consumption goods, *m* operative factors, and *n* operative activities, let *A* be the $(\ell + m) \times n$ activity analysis matrix whose rows and columns correspond to the operative goods and activities, let *y* be the *n*-vector of operative activity levels, let *p* be the ℓ-vector of prices for the operative consumption goods, let *w* be the *m*-vector of prices for the operative factors, let *z*(*p*, *w*) be the excess demand function for the operative consumption goods, which we assume is homogeneous of degree 0 in (*p*, *w*), and finally let *e* be the *m*-vector of inelastic supplies of the operative factors. Walras's law then states that $p \cdot z(p, w) = w \cdot e$. Equilibria $(p, w, y) \geq 0$ are locally characterized by the equalities

$$(z(p, w), \; -e) = Ay, \quad (p, w)' A = 0. \tag{2}$$

(All vectors are column vectors and $'$ denotes transposition.) Bear in mind that the market-clearing and no-positive-profit inequalities excluded from (2) vary by equilibrium; the activities and goods operative in one equilibrium need not be operative in another. We assume henceforth that, at any equilibirum, each of the operative activities is used at a strictly positive level and that each operative good has a strictly positive price, $(p, w, y) \gg 0$. As usual, the homogeneity of demand allows us to set one of the positively priced goods to be the numéraire and Walras's law implies that one of market-clearing conditions is redundant. So we set the price of the first consumption good not in excess supply to equal 1 and put aside the market-clearing condition for this good. Letting $\bar{z}(p, w)$ denote *z* (*p*, *w*) without the first coordinate, $\bar{A}$ denote *A* without the first row, and $\bar{p}$ denote *p* with the first coordinate set equal to 1, (2) can be written

$$(\bar{z}(\bar{p}, w), -e) = \bar{A}y \tag{3}$$

$$(\bar{p}, w)' A = 0. \tag{4}$$

Any small change in $(\bar{p}, w, y)$ that satisfies (3)–(4) will then continue be an equilibrium: the variables $(\bar{p}, w, y)$ will remain positive, all excluded goods will remain in excess supply, and all excluded activities will continue to make negative profits or continue to use and produce only goods in excess supply.

The most conspicuous case of factor-price indeterminacy occurs when *m*>*n*, that is, when there are more operative factors than operative activities. If, beginning at some reference equilibrium, we fix *y* at its equilibrium value, then as $(\bar{p}, w)$ varies the market-clearing conditions for factors in (3) will continue to be satisfied. But the remaining equilibrium conditions – (4) and the market-clearing conditions for consumption goods in (3) – comprise $n + \ell - 1$ equations in the $\ell - 1 + m$ variables $(\bar{p}, w)$. Hence, if *m*>*n* and as long these remaining equilibrium conditions satisfy a rank condition, which allows the implicit function theorem to be applied, indeterminacy will occur. The economy considered earlier where two factors are used by one activity qualifies as an example of the *m*>*n* type of indeterminacy, while the economy where two factors are used to produce two goods does not.

A slight variation of this argument applies to a subset of factors. Suppose that $\hat{m}$ of the *m* operative factors are used by only $\hat{n}$ of the *n* operative activities, and that $\hat{m} > \hat{n}$. Thus the remaining $n - \hat{n}$

operative activities have 0 entries in the rows of *A* that correspond to these $\hat{m}$ factors. If we fix the $\hat{n}$ coordinates of *y* for the activities that do use these $\hat{m}$ factors, then, as the remaining endogenous variables (the other $n - \hat{n}$ activity levels, $\bar{p}$, and *w*) change, the market-clearing conditions for the $\hat{m}$ factors will continue to be satisfied. Moreover, the number of remaining endogenous variables is $n - \hat{n} + \ell - 1 + m$ while the number of remaining equilibrium conditions is $\ell - 1 + m - \hat{m} + n$. The difference between the number of remaining variables and remaining equilibrium conditions is therefore $\hat{m} - \hat{n}$ and so there are more variables than equilibrium conditions. Indeterminacy therefore obtains (again, given a rank condition).

Factor-price indeterminacy, whether for an economy as a whole or for a subset of an economy's factors, depends critically on production sets that exhibit kinks. By fixing a set of activity levels, the above indeterminacy argument fixes a vector of factor demands and finds a multiplicity of prices at which firms will demand exactly those quantities. If the aggregate production set were smooth, a fixed vector of firm factor demands would be supported by only one vector of relative factor prices.

Factor-price indeterminacy brings dramatic behavioural consequences: agents have a strong incentive to manipulate factor prices and hence markets cannot function competitively. In the two factor–one activity example, where the endowments satisfy (1), the tiniest withdrawal of either factor *i*=1 or *i*=2 from the market will lead the other factor to be in excess supply and have price 0 and hence cause factor *i*'s price to jump to $1 / a_i$. No matter how small an owner of factor *i* is as a proportion of the market, it will be in his or her interest to remove a small amount of *i* from the market. Agents therefore will not behave like price-takers. When more activities are present, the jump in factor prices need not be as large, but a jump will still occur for an arbitrarily small withdrawal of a factor, and hence the incentive to manipulate will remain. The distinctive mathematical feature of factor-price indeterminacy that drives this conclusion is that the equilibrium correspondence fails to be lower hemicontinuous. (The equilibrium correspondence is the correspondence from the parameters of the model, such as the endowments *e*, to the endogenous variables $(\bar{p}, w, y)$.) When the endowments of factors lead to an indeterminate equilibrium, it will usually be impossible at nearby endowment levels to find equilibrium prices near to the prices of the indeterminate equilibrium. Other varieties of indeterminacy in the general equilibrium model, such as the indeterminacy of the overlapping generations model, do not suffer from such a failure of lower hemicontinuity and therefore do not invite market manipulation (see Mandler, 2002).

3 The emergence of factor-price indeterminacy through time

We saw in the two factor–one activity example that indeterminacy occurs only if a knife-edge condition on endowments is satisfied. This observation applies to the broader species of factor-price indeterminacy as well. Suppose again that at some reference equilibrium $\hat{m}$ operative factors are used by $\hat{n} < \hat{m}$ operative activities, let $\hat{e}$ be the endowments of these $\hat{m}$ factors, let $\hat{y}$ be the activity levels for the $\hat{n}$ activities, and let $\hat{A}$ be the $\hat{m} \times \hat{n}$ submatrix of *A* formed by the rows for the $\hat{m}$ factors and the columns for the $\hat{n}$ activities. Then $\hat{A}\hat{y} = \hat{e}$. But since $\hat{A}$ has more rows than columns, for almost every value of $\hat{e}$, $\hat{A}\hat{y} = \hat{e}$ will have no solution. Hence, for most levels of an economy's endowments, there will be no equilibrium at which $\hat{m}$ operative factors are used by fewer than $\hat{m}$ operative activities. While the failure

in these so-called generic cases of the indeterminacy arguments we have given does not show that equilibria are generically locally unique, the literature on *regular economies* (see in particular Mas-Colell, 1975; 1985; and Kehoe 1980; 1982) has shown that, for generic endowments and preferences, general equilibrium models with linear or nonlinear production activities do have locally unique equilibria.

The determinacy qst, however, does not end here. An economy's endowments of produced inputs – capital goods – are in any long-term view endogenous variables not parameters. Consequently, even though factor-price indeterminacy does not arise for generic endowments, it is conceivable that those special endowments that lead to indeterminacy will systematically arise as the equilibrium activity of an economy unfolds through time. To see that this can indeed happen, we partition an intertemporal economy's dates into two periods, a first period where goods are either consumed or invested in the production of factors, and a second period where the factors produced by first-period activities and natural endowments are used to create consumption goods (possibly also with the aid of intermediate inputs produced within the second period). To test whether the nongeneric factor endowments that lead to indeterminacy are likely to appear, we consider intertemporal economies where the endogenous equilibrium production of second-period factors leads the total stock of these factors to assume the nongeneric values where indeterminacy arises. If this endogenous second-period indeterminacy obtains for a robust family of equilibria (the equilibria of a nonempty open set of economies), then *sequential indeterminacy* occurs (Mandler, 1995).

In the Arrow–Debreu view of an intertemporal economy, agents trade just once at the beginning of economic time; after these initial contracts are signed, no further trade occurs, goods are just delivered. To allow for trade at multiple dates, and thus give indeterminacy in later time periods a chance to appear, we assume instead that agents transfer wealth between periods by borrowing or lending assets. Agents then will typically trade every period, and the economies that appear in later periods will have endowments that are endogenously determined by trade in the initial periods. Moreover any indeterminacy of prices in later periods will change the quantities of goods exchanged and hence change agents' utilities. In our setting, with just two periods, we can let the activities that produce second-period factors serve as assets: agents in the first period will buy or sell rights to the outputs of the activities that produce the second-period factors and then in the second period receive or deliver the second-period factors they contracted for in the first period and use their income to trade for consumption. The allocation achieved by a two-period Arrow–Debreu intertemporal equilibrium will occur in an equilibrium with two sequential periods of trade if (*a*) agents in the first period unanimously anticipate a second-period price vector, (*b*) given those expectations, goods and asset markets in the first period clear, and (*c*) given asset deliveries, second-period markets clear at the anticipated prices. We omit the routine details of how to decompose an intertemporal equilibrium into a sequential-trading equilibrium (see Radner, 1972) and will just write one equilibrium condition explicitly, the market-clearing equality for second-period factors.

As usual, we consider some reference equilibrium and ignore those goods in excess supply and those activities that make strictly negative profits or that use and produce only goods in excess supply. If there are k operative goods in period 1, and ℓ operative consumption goods and m operative factors in period 2, the activity analysis matrix for the operative goods and activities takes the form

$$A = \begin{array}{c} \begin{pmatrix} A_1 & 0 \\ 0 & A_{c2} \\ A_{f1} & A_{f2} \end{pmatrix} \end{array} \begin{array}{c} k \\ \ell \\ m \end{array}$$

where the subscript *c* or *f* indicates whether the rows are for consumption goods or factors and the subscript 1 or 2 indicates the time activities begin operation. Since presumably the second-period factors are the outputs of time 1 activities and the inputs of time 2 activities, it makes sense to suppose $A_{f1} \geq 0$ and $A_{f2} \leq 0$. If we let y_i denote the activity levels for operative activities that begin in period *i* and *e* the endowment of operative second-period factors, the market-clearing equality for operative second-period factors is

$$A_{f1} y_1 + A_{f2} y_2 + e = 0. \tag{5}$$

In the background lie the remaining equilibrium conditions: market-clearing conditions for excess-supply factors and for all consumption goods, and nonpositive profit conditions for activities. Consider the restrictions that (5) places on the number of operative factors. If the number of operative activities in the two periods that produce or use the *m* operative second-period factors is less than *m*, then, for almost every *e*, (5) will have no solution $y = (y_1, y_2) \geq 0$. Similarly if there is a subset of $\hat{m}$ operative second-period factors where the number of operative activities in the two periods that produce or use these factors is less than $\hat{m}$, then again (5) will usually have no solution. We may therefore dismiss these cases as unlikely, in line with the literature on regular economies. In the remaining cases, where for each subset of $\hat{m}$ operative second-period factors the number of operative activities in the two periods that produce or use these factors is greater than or equal to $\hat{m}$, then (5) can have a solution $y \geq 0$ for a robust (open) choice of endowment levels *e*. But in these latter cases it could well be that some subset of operative second-period factors – say the entire set of all *m* of these factors – is used by fewer than *m* operative *second-period* activities. For an example, let *m*=2, suppose that the first factor has no endowment but is produced by an activity with factor output coefficient c_1 while the second factor has a positive endowment in the second period and is not produced. In the second period, both factors are used by one activity with factor usage coefficients a_1 and a_2. Then (5) consists of the two equalities

$$\begin{aligned} c_1 y_1 + a_1 y_2 &= 0, \\ a_2 y_2 + e_2 &= 0. \end{aligned} \tag{6}$$

Evidently if $a_1<0$, $a_2<0$, $c_1>0$, and $e_2>0$, then a solution $y \gg 0$ to (6) exists and is robust: for a small variation in the production coefficients or the endowment, a solution $y \gg 0$ will continue to exist. In this equilibrium, factor 2 is produced in just the quantity necessary to ensure that neither factor 1 nor factor 2 is in excess supply. For a second example, suppose that factor 2 is produced as well and also has no endowment, and let y_{1i} denote the usage level of the activity that produces factor *i*. Then (6) is replaced by $c_1y_{11}+a_1y_2=0$ and $c_2y_{12}+a_2y_2=0$. Now efficiency and hence equilibrium will usually *require* that the two factors are produced in quantities that leave neither in excess supply in period 2; if, say, factor 1 were in excess supply and if y_{11} could be lowered, thereby increasing the output of some first-period consumption good, an inefficiency would exist, which is impossible in equilibrium.

Once agents arrive at period 2, they trade again but now the factor outputs produced by the activities that began in period 1 are exogenously given. So in the example given by (6) the endowment of factor 1 in period 2 equals c_1y_1 and one may readily check that this quantity along with e_2 of factor 2 satisfies the knife-edge condition (1). Thus, despite seeming to be unlikely at a given point in time, the endowments that lead to indeterminacy can endogenously arise.

Intertemporal general equilibrium economies therefore can be sequentially indeterminate. Moreover, factor-price indeterminacy is typically the only source of endogenous indeterminacy. Let us call the equilibria that occur in the later periods of operation of a sequential-trading equilibrium and that confirm the expectations formed in the initial period ‘continuation equilibria’. A continuation equilibrium is *indeterminate* if it sits amid a continuum of other (usually non-continuation) equilibria.

Sequential indeterminacy th: (Mandler, 1995). For a generic set of intertemporal economies with linear activities, a continuation equilibrium is indeterminate at some date *t* if and only if there is a set of m operative factors appearing at *t* or later that are used or produced by fewer than m operative activities that begin at *t* or later.

In contrast, when production sets are smooth, endogenous endowments do not lead to indeterminacy; typically continuation equilibria are locally unique (Mandler, 1997).

4 Factor price indeterminacy and the hold-up problem

The endogenous factor-price indeterminacy of the previous section is *not* an indeterminacy of the equilibria of the entire intertemporal economy or of the corresponding sequential-trading equilibria. As long as the non-produced endowments of every period of an intertemporal economy avoid certain nongeneric values, and barring flukes in preference or technology coefficients, only a finite number of intertemporal equilibria will exist. It follows that in a two-period model that displays sequential indeterminacy, almost all of the infinite multiplicity of equilibria of the second-period economy could not form part of a two-period sequential-trading equilibrium: if the prices of almost any of the second-period equilibria were anticipated in period 1, they would be inconsistent with market clearing. Specifically, if anticipated second-period prices were to vary slightly from the values that hold in a sequential-trading equilibrium, then either assets would no longer share the same rate of return or the common rate of return on assets would change, and hence typically markets would not clear. But

bygones are bygones: once period 1 is past, even the second-period equilibria that violate the requirements of an intertemporal equilibrium are equilibria nonetheless when the economy arrives at its second period.

Moreover second-period indeterminacy will prevent sequential-trading equilibria from proceeding smoothly through time: they will be virtually certain to unravel. Since factor prices are indeterminate in the second period, rational agents will predict that an investment in an activity producing a second-period factor will not except by chance earn the rate of return anticipated in the first period of a sequential-trading equilibrium. Investments will therefore differ from their Walrasian levels. The predictions of the general equilibrium model thus become untenable when agents trade repeatedly through time and factor-price indeterminacy is present, even though all the classical presuppositions of the model – price-taking agents, no distortions and so on – obtain.

The inability of second-period markets to ensure that assets earn the rate of return necessary for efficiency amounts to a hold-up problem, but the cause of the problem differs from the conventional diagnosis. In the classical hold-up problem, the owners of two complementary factors Nash bargain over the revenue they jointly earn; hence, if the owner of one of the factors invests to improve the quality of his factor, the owner recoups only a fraction of the increment to revenue, and consequently investment is inefficiently low (Hart, 1995). The problem, it would seem, is that the factor owners form a bilateral monopoly and cannot purchase each other's services on a competitive market. What we have seen, however, is that a hold-up problem can arise with perfectly competitive markets. Even if factor owners can purchase all complementary factors on competitive markets, factor-price indeterminacy can prevent investments in factors from earning the rate of return required in intertemporal equilibrium (and hence the rate necessary for efficiency): an unguided market has no means to select from the continuum of equilibrium factor prices the specific prices that deliver intertemporal efficiency. Factor markets moreover will not operate competitively in the presence of factor-price indeterminacy, which is another cause for the rate of return to deviate from its competitive equilibrium value. For both reasons, the efficient Walrasian levels of investment need not occur.

Just as in the classical hold-up problem, long-term contracts can mitigate the troubles that factor-price indeterminacy brings. If labour is among the factors in an economy displaying factor-price indeterminacy, then a labour contract may be able to force trading at prices that allow intertemporal efficiency and prevent labourers or capital goods owners from manipulating factor prices by withdrawing their services from the market. Of course, as in the classical hold-up problem, the incompleteness of contracts may hamper the ability of this solution to deliver first-best efficiency.

Alternatively, when a set of complementary factors displays factor-price indeterminacy and consists solely of produced goods, then a bundling of the complementary factors in an asset portfolio – that is, in a ‘firm’ – can eliminate the incentive to manipulate prices. From the vantage point of factor-price indeterminacy, unions and labour contracts and the firm as an institution emerge as devices to enforce competitive equilibria, not as consequences of imperfect competition in factor markets.

5 Conclusion: factor-price indeterminacy past and present

Prior to the Arrow–Debreu transformation of general-equilibrium theory, economists were well aware that linear activities could lead to an indeterminacy of factor prices. The problem was considered from a long-run perspective: a change in a factor price was presumed to persist for many periods, and, although

such a change might not lead to an instantaneous change in either the supply or demand for the factor, arguments were deployed for why demand and supply responses would eventually kick in. For example, in response to a wage increase, although existing capital equipment might have fixed labour requirements, newly constructed capital equipment could be built to use labour less intensively. In addition, a wage increase would eventually lead the price of labour-intensive consumption goods to rise, diminishing the demand for these goods and therefore ultimately for labour as well. This effect does not operate immediately since a wage increase will lead to an offsetting fall in the prices of existing stocks of complementary capital inputs. But the prices of newly produced capital inputs are constrained by break-even requirements; hence, given enough time, the prices of labour-intensive consumption goods will increase. (Robertson, 1931, and Hicks, 1932, offered the most detailed long-run theories. See Mandler, 1999, ch. 2.) Although pre-modern explanations of factor prices faced the indeterminacy problem explicitly, and marshalled a rich array of counter-arguments for why the problem normally will not be severe, the long-run perspective had its drawbacks: the attention to persistent changes in factor prices masked an inability to explain why factor prices cannot temporarily change. The older long-run theories simply assumed that, in the absence of demand or supply shocks, factor prices will be maintained at their long-run equilibrium values. This presumption amounts to a rudimentary version of the rule that in an intertemporal equilibrium prices should fulfill the expectations that agents formed in earlier periods. As we have seen, the market mechanism will not enforce this rule; a supplementary theory of contracts and institutions is necessary. The Arrow–Debreu treatment of factors (and other goods) at different dates as fully distinct goods naturally raises the question of whether prices can deviate from previously anticipated values even in the absence of shocks, and curiously, therefore, the Arrow–Debreu account of markets points to the need for a theory of non-market institutions. Unfortunately, the Arrow–Debreu tradition also took the model of trading at a single point in time as its benchmark. It is only with the combination of goods rigorously distinguished by date, sequential trading, and production sets with kinks that factor-price indeterminacy will appear.

See Also

- determinacy and indeterminacy of equilibria
- general equilibrium
- hold-up problem

Bibliography

Cassel, G. 1924. *The Theory of Social Economy*. New York: Harcourt.

Hart, O. 1995. *Firms, Contracts, and Financial Structure*. New York: Oxford University Press.

Hicks, J. 1932. Marginal productivity and the principle of variation. *Economica* 25, 79–88.

Kehoe, T. 1980. An index theorem for general equilibrium models with production. *Econometrica* 48,

1211–32.

Kehoe, T. 1982. Regular production economies. *Journal of Mathematical Economics* 10, 147–7.

Mandler, M. 1995. Sequential indeterminacy in production economies. *Journal of Economic Theory* 66, 406–36.

Mandler, M. 1997. Sequential regularity in smooth production economies. *Journal of Mathematical Economics* 27, 487–504.

Mandler, M. 1999. *Dilemmas in Economic Theory*. New York: Oxford University Press.

Mandler, M. 2002. Classical and neoclassical indeterminacy in one-shot vs. ongoing equilibria. *Metroeconomica* 53, 203–22.

Mas-Colell, A. 1975. On the continuity of equilibrium prices in constant returns production economies. *Journal of Mathematical Economics* 2, 21–33.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: A Differentiable Approach*. Cambridge: Cambridge University Press.

Radner, R. 1972. Existence of equilibrium of plans, prices, and price expectations in a sequence of markets. *Econometrica* 40, 289–303.

Robertson, D. 1931. Wage grumbles. In *Economic Fragments*. London: P.H. King.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Wieser, F. 1927. *Social Economics*. New York: Greenberg.

How to cite this article

Mandler, Michael. "factor prices in general equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000298> doi:10.1057/9780230226203.0546

The New Palgrave Dictionary of Economics Online

fair allocation

William Thomson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We survey the theory of equity in a variety of concretely specified resource allocation models: classical economies with private goods, economies with production, economies with indivisible goods, when monetary compensations are feasible and when they are not, economies with single-peaked preferences, and economies in which the dividend is a non-homogeneous continuum. We present the central fairness punctual notions, no-envy, egalitarian-equivalence, concepts of equal or equivalent opportunities and the relational principles of monotonicity and consistency.

Keywords

consistency; convexity; efficient allocation; egalitarian-equivalent allocation; envy-free allocation; equality of opportunity; fair allocation; indivisible goods; monotonicity; Shapley value; single-peaked preferences; solidarity

Article

We survey the theory of equity in concretely specified economic environments. The literature concerns the existence of allocation rules satisfying various requirements of fairness expressed in terms of resources and opportunities understood in their physical sense (and not in terms of abstract entities such as utilities or functionings). For lack of space, we often give only representative references. Detailed treatments of the subject are Young (1994), Brams and Taylor (1996), Moulin (1995; 2003), and Thomson (1995b; 2006c).

1 Concepts

We introduce concepts central to the classical problem of fair division. These have much broader applicability, but for other models they sometimes have to be reformulated. Also, as models vary in their

mathematical structures, the implications of a given concept may differ significantly from one to the other.

In an *economy*, there is a social endowment of resources to be distributed among a group of agents who are collectively entitled to them. For what we call a classical problem of fair division, the resources are infinitely divisible private goods, and preferences are continuous, usually monotonic (sometimes strictly so), and convex. In an *economy with individual endowments*, each agent starts out with a share of society's resources; the issue in this case is to redistribute endowments. In a *generalized economy*, some resources are initially owned collectively and others are individual endowments (Thomson, 1992; Dagan, 1995). A *solution* associates with each economy a non-empty subset of its set of feasible allocations. A *rule* is a single-valued solution.

An axiomatic study begins with the formulation of requirements on solutions (or rules). Their logical relations are clarified and their implications, when imposed in various combinations, are explored. For each combination of the requirements, do solutions exist that satisfy all of them? If the answer is ‘yes’, can one characterize the class of admissible solutions?

A *punctual requirement* applies to each economy separately. The main question then is the existence, for each economy in the domain under consideration, of allocations satisfying the requirement. First are bounds on welfares defined agent-by-agent, in an intra-personal way. Some are lower bounds, offering agents welfare guarantees. Others are upper bounds, specifying ceilings on their welfares. An allocation satisfies *no-domination of, or by, equal division*, if no agent receives a bundle that contains at least as much as an equal share of the social endowment of each good, and more than an equal share of the social endowment of at least one good, or a bundle that contains at most as much as an equal share of the social endowment of each good, and less than an equal share of the social endowment of at least one good (Thomson, 1995b). It satisfies the *equal-division lower bound* if each agent finds his bundle at least as desirable as equal division (Kolm, 1972; Pazner, 1977; and many others).

Second are requirements based on interpersonal comparisons of bundles, or more generally, ‘opportunities’, involving exchanges of, or other operations performed on, these objects. An allocation satisfies *no domination across agents* if no agent receives at least as much of all goods as, and more of at least one good than, some other agent (Thomson, 1983a). It satisfies *no-envy* if each agent finds his bundle at least as desirable as that of each other agent (Foley, 1967; Kolm, 1973, proposes a definition that encompasses many variants of the concept). The final definition is quite different in spirit: an allocation is *egalitarian-equivalent* if there is a reference bundle that each agent finds indifferent to his own bundle (Pazner and Schmeidler, 1978). Given a direction r in commodity space, it is *r-egalitarian-equivalent* if it is egalitarian-equivalent with a reference bundle proportional to r. Of particular interest is when r is the social endowment.

A *relational requirement* prescribes how a rule should respond to changes in some parameter(s) of the economy. The idea of solidarity is central: if the environment changes, and whether or not the change is desirable, but no one in particular is responsible for the change, that is, no one deserves any credit or blame for it (or no one in a particular group of agents is responsible for the change), the welfares of all agents (or all agents in this particular group), should be affected in the same direction: all ‘relevant’ agents should end up at least as well off as they were initially, or they should all end up at most as well off. In implementing this idea, the focus is usually on a particular parameter. When the parameter belongs to a space that has an order structure, as is frequent, one can speak of the parameter being given

a ‘greater’ or ‘smaller’ value in that order. Then, together with efficiency, the solidarity idea often implies a specific direction in which welfares should be affected: when a Pareto improvement is possible, all relevant agents should end up at least as well off as they were initially; otherwise, all should end up at most as well off. Thus, solidarity takes the form of a ‘monotonicity’ requirement. Examples are *resource monotonicity*: if the social endowment increases, all agents should end up at least as well off as they were initially (Thomson, 1978; Roemer, 1986a; 1986b; Chun and Thomson, 1988); *technology monotonicity*, a similar requirement when technology expands (Roemer, 1986a; Moulin and Roemer, 1989); *population monotonicity*: if population expands, all agents initially present should end up at most as well off as they were initially (Thomson, 1983b; Chichilnisky and Thomson, 1987; for a survey, see Thomson, 2006a).

When the parameter that varies does not belong to a space equipped with an order structure, solidarity retains its general form. For example, *welfare domination under preference replacement* says that if the preferences of some agents change, all agents whose preferences have not changed should end up at least as well off as they were initially, or that all should all end up at most as well off (Moulin, 1987b; for a survey, see Thomson, 1999). Whether or not the parameter belongs to a space with an order structure, one can imagine simply replacing the initial value taken by the parameter with another value (to which, if there is an order, it may or may not be comparable in the order), and still require that the welfares of all relevant agents should be affected in the same direction.

Another application of the idea is to situations where some agents leave with the resources assigned to them. The requirement that the welfares of all remaining agents should be affected in the same direction, when imposed on efficient rules, often means that these agents should be assigned the same bundles as initially. It can be expressed more simply as *consistency*: given an allocation chosen by a solution for some economy, let us imagine the departure of some agents with their components of it. In the resulting ‘reduced economy’, the remaining agents should receive the same bundles as initially (for a survey, see Thomson, 2006b).

Requirements relative to private endowments, when such exist, may be imposed on rules. For instance, the *individual-endowments lower bound* is the punctual requirement that each agent should end up with a bundle that he finds at least as desirable as his endowment; *individual-endowment monotonicity* is the relational requirement that if an agent's endowment increases, he should end up with a bundle that he finds at least as desirable as the one he got initially (Aumann and Peleg, 1974).

Logical relations*; existence*. Under standard assumptions, efficient allocations meeting the equal-division lower bound exist, and so do envy-free and efficient allocations. If preferences are strictly monotonic, no envy implies no-domination. An allocation meeting the equal-division lower bound is not necessarily envy-free. *Equal-division Walrasian allocations* are both envy-free and efficient, and under standard assumptions, they exist. In an economy with an infinite population of agents modelled as a continuum, and if preferences are sufficiently diverse, a partial converse holds: if an allocation is envy-free and efficient, it is an equal-income Walrasian allocation (Varian, 1974; Kleinberg, 1980; Champsaur and Laroque, 1981; Mas-Colell, 1987; Zhou, 1992). If preferences are not convex, the existence of envy-free and efficient allocations can be derived from certain assumptions about the structure of the efficient set itself (Varian, 1974; Svensson, 1983a; 1994b; Diamantaras, 1992). In the absence of such assumptions, efficient allocations satisfying no envy, even no domination, may not exist (Maniquet, 1999).

An egalitarian-equivalent and efficient allocation may violate no domination. When $|N| > 2$, and if the reference bundle is proportional to the social endowment, then obviously the equal-division lower bound is met, although not no domination. In fact, for some economies, all egalitarian-equivalent and efficient allocations violate no domination (Daniel, 1978). An equal-division Walrasian allocation may not be egalitarian-equivalent.

The existence of *r*-egalitarian-equivalent and efficient allocations holds under weak assumptions (Pazner and Schmeidler, 1978; Sprumont and Zhou, 1999, offer a proof for economies with a continuum of agents).

Variants: extensions. Some solutions are based on comparing across agents the number of agents whom each agent envies and the number of agents who envy him. Envy is *balanced* if, for each agent, these two numbers are equal (Daniel, 1975). The existence of allocations with balanced envy holds more generally than is common for other concepts. Other natural ideas are to require of an allocation that all agents should envy the same number of agents, or that all agents should be envied by the same number of agents. But neither definition will do, as soon as efficiency is imposed, because in any economy whose set of feasible allocations is closed under permutations, at an efficient allocation, at least one agent envies no one, and at least one agent is envied by no one (Varian, 1974; Feldman and Kirman, 1974).

Selections. When envy-free allocations exist, there may be a large number of them and the question of selection arises. A variety of proposals have been made. Some are based on quantifying the extent to which the no envy constraints are exceeded. Conversely, when envy-free allocations do not exist, the extent to which they are violated can also be measured. Measures based on counts of envy relations, or on the adjustments in commodity bundles required to eliminate envy have been proposed (Feldman and Kirman, 1974; Varian, 1976; Chaudhuri, 1985; 1986; Diamantaras and Thomson, 1990; Kolpin, 1991a). These operations can be adapted so as to extend, or select from, other equity notions, and in a second step, rankings of allocations can be derived (Chaudhuri, 1986; Thomson, 1995c).

Group fairness. Most of the concepts of the previous pages can be applied to compare the welfares of groups of agents. Central among them are the *equal-division core*, whose definition is straightforward, and *group no envy*: no group should be able to improve the welfares of all of its members if given access to the resources assigned to some other group of the same size. The definition can be adapted to handle groups of different sizes (Kolm, 1972; Feldman and Kirman, 1974; Green, 1972; Khan and Polemarchakis, 1978). Under replication, there is a sense in which the set of efficient allocations that are group envy-free converges to the set of equal-division Walrasian allocations (Varian, 1974; Kolpin, 1991b).

Fairness of trades. The concepts formulated above for allocations can be adapted in various ways to assess the fairness of individual trades when agents are individually endowed (Kolm, 1972; Schmeidler and Vind, 1972), and to assess the fairness of the trades of groups (Jaskold-Gabszewicz, 1975; Yannelis, 1983). Walrasian trades satisfy most of the definitions that have been proposed and under weak assumptions on preferences, for several of the definitions, a converse inclusion holds (Schmeidler and Vind, 1972; Shitovitz, 1992).

Interesting conceptual issues arise in relating the fairness of allocations and the fairness of trades (Goldman and Sussangkarn, 1980; Thomson, 1983a).

2 Economies with production

A fundamental issue is fair allocation when agents have contributed differently to production, because they have supplied unequal amounts of their time or because they are unequally productive.

A first way to extend the notion of an envy-free allocation to such situations is by having each agent $i \in N$ compare his *complete* consumption bundle (including his consumption of leisure) to those of the other agents. Unfortunately, and even if preferences are quite well-behaved, envy-free and efficient allocations may not exist (Pazner and Schmeidler, 1974). Limited exceptions are when all abilities or all preferences are the same (Varian, 1974). Another exception is in the two-good case, under a 'single-crossing' assumption on preferences and when the technology is linear (Piketty, 1994).

Egalitarian-equivalent and efficient allocations exist quite generally (Pazner and Schmeidler, 1978). Also, under appropriate convexity assumptions, existence still holds if the reference bundle in the definition of egalitarian-equivalence is required to be proportional to the average consumption bundle.

An alternative proposal is to recognize the envy of agent $j \in N$ by agent $i \in N$ only after agent *i*'s consumption of leisure is adjusted for him to produce what agent *j* produces (Varian, 1976; Otsuki, 1980). The concept is well defined only if the production set is additive. A proof of the existence of such *productivity-adjusted envy-free* and efficient allocations can be given along the lines of the 'Walrasian' proof of existence of envy-free and efficient allocations in exchange economies and under similar assumptions (Varian, 1974). Some have objected to the definition because it lets agents with high productivity appropriate the benefits of their greater skills. Alternative concepts have been defined that attempt to distribute across agents these benefits (Pazner and Schmeidler, 1978; Varian, 1974; Pazner, 1977). The main proposal here has been to take advantage of the instrumental value of the Walrasian solution in delivering envy-free allocations when there is no production and in providing equal opportunities: here, one operates the Walrasian solution from equal division of all goods, including time endowments. Svensson (1994b) states an existence result for allocations at which implicit incomes are equal.

Non-convexities in technologies present another difficulty for the existence of envy-free and efficient allocations (Vohra, 1992). Vohra proposes to weaken no envy by imposing a certain symmetry among all agents with respect to possible occurrences of envy (see also Varian, 1974). Existence holds without any convexity assumption on either preferences or technologies. A critical one, however, is that there be no agent-specific input (Vohra, 1992).

Next, we turn to criteria that, by contrast with the previous ones, can be evaluated agent-by-agent, just like the equal-division lower bound. First, for each agent, we imagine an economy composed of agents having the same preferences as his, and we identify their common welfare under efficiency and *equal treatment of equals*. We take this welfare as a bound, thereby defining the *identical-preferences lower bound*. For nowhere-increasing returns-to-scale, it can be met (Gevers, 1986; Moulin, 1990d).

Alternatively, we could imagine each agent in turn controlling an equal share of the social endowment and the technology, obtaining the *equal-division free-access upper bound* (Moulin, 1990d; Yoshihara, 1998). This definition can be generalized by imagining each group of agents in turn controlling a proportion of the social endowment equal to its relative size in the economy and the technology (Foley,

1967). This yields the *equal-division free-access core*. There are economies with a concave production function in which no allocation is envy-free, efficient, and meets the equal-division free-access upper bound (Moulin, 1990c). However, the bound is met on that domain by selections from the Pareto solution, in particular by the constant-returns-to-scale-equivalent solution defined later (Mas-Colell, 1980; Moulin, 1987b). For nowhere-decreasing returns-to-scale, the equal-division free-access bound becomes a lower bound: here, no sub-solution of the Pareto solution satisfies no envy for trades and meets the bound (Moulin, 1987b). Systematic studies of lower and upper bounds are Moulin (1990a; 1990e; 1991; 1992b).

For one-input one-output production economies, *an allocation is proportional* if there are prices such that each agent, facing these prices, maximizes his preferences at his component of the allocation. These allocations can be used to define another lower bound on welfares (Maniquet, 1996b; 2002; see also Roemer and Silvestre, 1993).

The *constant-returns-to-scale lower bound* is defined, for each agent, by reference to the best bundle he could achieve if given access to a constant-returns-to-scale technology, the same for all agents; the *work-alone lower bound* is defined for each agent, by reference to the best bundle he could obtain if given access to the actual technology but under the obligation to provide bundles to the other agents to which he would not prefer his own (Fleurbaey and Maniquet, 1996a; 1999).

Another study relating bounds in a class of two-good economies with convex production sets, the identical preferences lower bound and the free-access upper bound is due to Watts (1999).

3 Equal opportunities as equal, or equivalent, choice sets

The notion of ‘equal opportunities’ is of course central in the theory of economic justice (for a general discussion, see Fleurbaey, 1995c). The expression has been given a variety of meanings. In economies affected by uncertainty, it may mean ‘equal treatment *ex ante*’. Uncertainty may also be endogenously generated by an allocation rule. Consider the problem of allocating an indivisible good. A lottery giving all agents equal chances might be deemed equitable *ex ante* although the final allocation may well appear inequitable. Alternatively, if agents’ opportunities today are determined by decisions they made yesterday, equal opportunities may mean that they all had access to the same set of decisions. It is often argued that, because of incentive considerations, we should not attempt to equalize end results but instead should limit ourselves to giving people equal chances to develop their potential. If we do so, equal opportunities are provided by the mechanism that converts the choices agents make into a final outcome.

Another way to give substance to the idea of equal opportunities is to let each agent choose his consumption bundle from a common choice set (for example, see Kolm, 1973). For the list of choices they make to constitute a feasible allocation, one should have access to a ‘rich enough’ family of choice sets. In addition, one would prefer efficiency to hold whenever feasibility does. Let $\mathcal{B}$ be a family of choice sets. An allocation is an *equal-opportunity allocation relative to* $\mathcal{B}$ (Thomson, 1994a) if there is a member of $\mathcal{B}$ on which each agent maximizes his preferences at his component of the allocation. Such an allocation is of course envy-free. *The family* $\mathcal{B}$ *is satisfactory* on a domain if the resulting equal-opportunity allocations are always efficient. Under standard assumptions on preferences, the *equal-*

income Walrasian family is satisfactory.

Another concept, *equal-opportunity–equivalence relative to a family $\mathcal{B}$ of choice sets*, generalizes the reasoning underlying egalitarian-equivalence. Check, whether, for some member of $\mathcal{B}$, each agent is indifferent between what he receives and the bundle he prefers in that set (Thomson, 1994a). For the family of linear choice sets, and adding efficiency, we obtain any efficient allocation such that each agent finds his component of it indifferent to the best bundle he could achieve if endowed with an equal share of the social endowment and given access to a constant-returns-to-scale technology, the same for all agents (Mas-Colell, 1980). Such an allocation is a *constant returns-to-scale equivalent allocation*.

Other solutions are obtained by having all agents face a hypothetical technology obtained from the actual one by imagining the productivity of one specific factor of production (alternatively, of some subset of the factors of production) to be multiplied by some number, or by introducing a fixed cost of some factor of production (alternatively, introducing a fixed cost proportional to some fixed vector). Radial expansions and contractions of the production set can also be considered. An application of the concept is by Nicolò and Perea (2005).

The next definition generalizes proposals by Archibald and Donaldson (1979) and Varian (1976). An allocation *exhibits no envy of opportunities relative to a family $\mathcal{B}$ of choice sets* if for each agent, there is a member of $\mathcal{B}$ that contains the agent's maximizer on the union of everyone's sets (Thomson, 1994a). For the family of linear choice sets, the resulting solution coincides with the equal-income Walrasian solution. If $\mathcal{B}$ is the family of $|N|$-lists of bundles, we obtain a concept that generalizes both no envy and egalitarian-equivalence. An allocation is *envy-free–equivalent* if there is a list of reference bundles, one for each agent, such that each agent is indifferent between his component of the allocation and his reference bundle and he finds his reference bundle at least as desirable as anyone else's reference bundle (Pazner, 1977).

4 Monotonicity

Monotonicity properties are quite strong when imposed in conjunction with no envy and even no domination. Indeed, (*a*) no selection from the no-domination and Pareto solution is *resource-monotonic* (Moulin and Thomson, 1988); (*b*) no selection from the no envy and Pareto solution is *population-monotonic* (Kim, 2004); (*c*) no selection from the no domination and Pareto solution satisfies *welfare-domination under preference-replacement* (Thomson, 1996). Other versions of these results are available, some of which involving significantly weaker distributional requirements (Geanakoplos and Nalebuff, 1988; Moulin and Thomson, 1988; Maniquet and Sprumont, 2000; Kim, 2001). However, if preferences satisfy gross substitutability and all goods are normal, the equal-division Walrasian solution is an example of a selection from the no envy and Pareto solution that is *resource-monotonic* and *population-monotonic* (Moulin and Thomson, 1988; Fleurbaey, 1995c).

On the other hand, no special assumptions are required for the existence of selections from the egalitarian-equivalence and Pareto solution that are *resource-monotonic*, or *population monotonic*, or satisfy *welfare-domination under preference-replacement*. Other rules based on the notion of equal-opportunity equivalence have these properties as well (Thomson, 1987).

For economies with quasi-linear preferences satisfying certain additional assumptions, the Shapley value

can provide the basis for a solution that is *resource-monotonic* (Moulin, 1992a). The Shapley value has in fact proved useful on other domains to obtain this and other desirable properties of rules, although at the price of no envy, egalitarian-equivalence, and their variants.

The solidarity requirement can be applied to the joint replacement of resources and preferences (Sprumont, 1996).

Technology-monotonicity is satisfied by certain selections from the egalitarian-equivalence and Pareto solution. For two goods, a characterization of a particular one is obtained by imposing it together with a few other minimal requirements. Suppose first that good 1 is used to produce good 2 according to a nowhere-decreasing-returns-to-scale technology. Given a group *N* of agents with preferences defined on $\mathbb{R}^2_+$, given some social endowment of good 1, which can be consumed as such or used as input in the production of good 2, the *equal-division free-access lower bound solution* selects the set of allocations such that each agent finds his bundle at least as desirable as the best bundle he could achieve if endowed with an equal share of the social endowment and given access to the technology.

Under alternative assumptions on technologies, (*a*) the only selection from the equal-division free-access lower bound and Pareto solution satisfying *Pareto-indifference* and *technology-monotonicity* is the constant-returns-to-scale–equivalence solution; (*b*) parallel characterizations hold for selections from the equal-division free-access upper bound (Moulin, 1987b; 1990d).

Although in (*a*), the bounds on welfares are individual bounds, the solution that emerges happens to satisfy the requirement that no group of agents should be able to make each of its members at least as well off, and at least one of them better off, if each of its members is endowed with an equal share of the social endowment and the group is given access to the technology. A similar strengthening holds for (*b*).

Suppose now that resources and technologies both change. Dutta and Vohra (1993) require of a solution that if the set of feasible profiles of welfare levels enlarges, each allocation chosen initially should be welfare-dominated by some allocation chosen after the change, and that each allocation chosen after the change should welfare dominate some allocation chosen initially. Let us refer to this requirement as *opportunity-monotonicity*. The requirement of *r-equity* is that in an exchange economy in which there is only some amount of good *r* to divide, equal division should be chosen. Dutta and Vohra consider an invariance requirement that also depends on the choice of a good, say *r*, so we call it *r-invariance*. It is not motivated by normative considerations, so we only note that it is a weak version of an invariance requirement that has been important in the theory of implementation. The results are: up to Pareto-indifference, (*a*) the *r*-egalitarian equivalence and Pareto solution is the only selection from the Pareto solution satisfying *r-equity* and *opportunity-monotonicity*; (*b*) on the sub-domain of exchange economies, it is the only selection from the Pareto solution satisfying *r-equity*, *r-invariance* and *opportunity-monotonicity*.

Economies with production. In situations where agents are differentiated by their input contributions, a first monotonicity requirement is that if the contribution of an agent increases, he should end up at least as well off as he was initially. In situations in which agents differ in their productivities, a corresponding requirement is that if an agent's productivity increases, then again, he should end up at least as well off as he was initially.

The solidarity requirement, applied to the *joint* replacement of preferences and population in conjunction with the self-explanatory *replication-invariance*, leads to the selection from the egalitarian-equivalence solution for which the reference bundle is proportional to the social endowment (Sprumont and Zhou,

1999; these authors also prove a version of this result for a model with infinitely many agents modelled as a continuum).

Economies with individual endowments. If the issue is that of allocating gains from trade, an appealing requirement is that when an agent's endowment increases, he should end up at least as well off as he was initially, *endowment monotonicity*. Another is that under the same hypotheses, nobody else should be made worse off than he was initially, *no negative effects on others*.

It is easy to define selections from the individual-endowments lower-bound and Pareto solution that are *own-endowment monotonic*. However, there are impossibilities too: (*a*) no selection from the no envy in trades and Pareto solution satisfies either *endowment monotonicity* or *no negative effect on others* (Thomson, 1987); (*b*) no selection from the egalitarian-equivalence and Pareto solution satisfies *no negative effect on others* (Thomson, 1987).

The appropriate expression of *population-monotonicity* here is that the welfares of all agents who are present before and after the change should be affected in the same direction. The Walrasian solution violates the property. However, the selections from the egalitarian-equivalence in trades and Pareto solution obtained by requiring the reference trade to lie on a monotone path satisfy the requirement. They also meet the individual-endowments lower bound (Thomson, 1995a).

5 Consistency and related properties

Here, we also consider situations in which both the population of agents and the resources may vary, but this time, our focus is on a variety of invariance properties. These properties can be interpreted as formalizing trade-offs between equity and efficiency objectives with objectives of informational simplicity.

A converse of *replication-invariance*, *division-invariance*, says that if an allocation that is chosen for a replica economy happens to be a replica allocation (of the same order), then the model allocation should be chosen for the model economy.

The central notion, *consistency*, was defined in Section 1. Conversely, given some allocation that is feasible for some economy, check whether the restriction of the allocation to each subgroup of two agents is chosen for the problem of allocating between them what they have received in total. If the answer is always yes, then one can say that each agent is treated fairly in relation to each other agent; then, *converse consistency* requires that the allocation itself should be chosen for the initial economy.

The Pareto solution is *consistent*. If preferences are smooth and corners excluded, it is also *conversely consistent* (Goldman and Starr, 1982). The no-envy solution is both *consistent* and *conversely consistent*. The egalitarian-equivalence solution is *consistent* but not *conversely consistent*. This is also true for the equal-division Walrasian solution although, if preferences are smooth and corners excluded, this solution is *conversely consistent*.

We have the following characterizations: (*a*) if a sub-solution of the equal-division core is *replication-invariant*, then it is a sub-solution of the equal-division Walrasian solution (this is because under replication, the core ‘shrinks’ to the set of Walrasian allocations; Debreu and Scarf, 1963; Thomson, 1988; Nagahisa, 1994, gives full characterizations of the Walrasian solution); (*b*) if a sub-solution of the group no-envy solution is *replication-invariant*, then it is a sub-solution of the equal-division Walrasian solution (Varian, 1974); (*c*) under smoothness, if a sub-solution of the equal-division lower bound and

Pareto solution is *replication-invariant* and *consistent*, then it is a sub-solution of the equal-division Walrasian solution (Thomson, 1988); (*d*) under smoothness, if a sub-solution of the equal-division lower bound and Pareto solution is *anonymous* and *conversely consistent*, then on the sub-domain of two-agent economies, it is a sub-solution of the equal-division Walrasian solution; if in fact coincidence occurs on that sub-domain, then it contains the equal-division Walrasian solution for all other cardinalities (Thomson, 1995b).

Consistency has been studied in economies with a large number of agents modelled as a continuum (Zhou, 1992). For economies with possibly satiated preferences, a characterization of the ‘equal-slack Walrasian solution’ (Mas-Colell, 1992) is available (Thomson and Zhou, 1993). This solution differs from the standard Walrasian notion in that each agent's income is the sum of the value of his endowment at the prices announced by the auctioneer (they may have negative or 0 components) and a supplement, the same for all agents, which, like prices, is determined endogenously. Economies with both atoms and an atomless sector have also been studied (Zhou, 1992; Shitovitz, 1992).

Juxtaposition-invariance says that if an efficient allocation happens to be obtained by juxtaposing two allocations that are chosen for two sub-economies with equal per-capita social endowments, then it should be chosen (Thomson, 1988). Under smoothness of preferences, the equal-division Walrasian solution is the only sub-solution of the Pareto solution satisfying a weak symmetry property, *juxtaposition-invariance*, and *consistency* (Maniquet, 1996a).

In formulating *consistency* for a production economy, the issue arises of how to define the opportunities open to a group of agents after the members of the complementary group leave with their bundles. The simplest idea is to translate the production set by the sum of the bundles the departing agents took with them. Standard classes of technologies are not closed under this operation however, and adjustments have to be made to ensure that the ‘reduced’ production set is admissible. For economies with one-input one-output and inelastic demands, characterizations of *proportional cost sharing* and *serial cost sharing* (which can be understood as an extension of the Shapley value) are available (Moulin and Shenker, 1994).

The *equal-wage-equivalent and Pareto solution* selects the efficient allocations for which there is a reference wage such that each agent finds his bundle indifferent to the best bundle he could achieve by maximizing his preferences on a budget set defined by this wage. The *output-egalitarian-equivalence and Pareto solution* selects the efficient allocations that each agent finds indifferent to a common consumption consisting of only some amount of the output.

Under appropriate assumptions on technologies, (*a*) the former is the only *essentially single-valued* selection from the constant-returns-to-scale lower bound solution satisfying *Pareto indifference*, *equal welfares for equal preferences* (self-explanatory), *contraction independence* (as in bargaining theory), and *consistency*; (*b*) the latter is the only *essentially single-valued* selection from the work-alone lower bound solution satisfying *Pareto indifference*, *equal welfares for equal preferences*, and *consistency* (Fleurbaey and Maniquet, 1999).

Roemer (1986a; 1986b; 1988) formulates consistency requirements with respect to changes in the number of goods.

When a solution is not *consistent*, it has a *minimal consistent enlargement* (Thomson, 1994d). For instance, the minimal *consistent* enlargement of the equal-division lower bound and Pareto solution is ‘essentially’ the Pareto solution. That of the Ω-egalitarian-equivalence and Pareto solution is

‘essentially’ the egalitarian-equivalence and Pareto solution. The *maximal consistent sub-solution* of a solution can be defined in a symmetric way provided the solution contains at least one *consistent* solution.

Notions of consistency have been proposed for economies with individual endowments (Thomson, 1992; van den Nouweland, Peleg and Tijs, 1996; Serrano and Volij, 1998; Korthues, 2000).

6 Indivisible goods

Estate or divorce settlements often involve items that cannot be divided (houses, family heirlooms), or can only be divided at a cost that would make the division undesirable (silverware). Other examples are positions in schools or organs for transplant patients. We call such goods ‘objects’. We consider situations in which the social endowment also contains some amount of an infinitely divisible good, ‘money’. We focus on situations in which each agent can consume at most one object. An illustration is the problem of allocating rooms to students in the house they share, and specifying how much each of them should contribute to the rent.

Let *A* be a set of objects. Each agent has preferences defined over $\mathbb{R} \times A$ (or over $\mathbb{R}_+ \times A$). They are continuous and strictly monotonic with respect to money, and such that the switch from any object to any other object can be compensated by an appropriate adjustment in the consumption of money. The simplest case is when there are as many objects as agents. Situations where there are fewer objects than agents are accommodated by introducing a ‘null object’; there are always enough copies of the null object for each agent to end up with one (real or null) object. If there are fewer agents than objects, some objects are unassigned. In some applications, it is natural to require that the null object should not be assigned until all real objects are, even if these objects are undesirable, or undesirable for some agents. They could be tasks to be assigned to housemates that none of them enjoys performing; alternatively, some of them may find a given task enjoyable and the others not (cooking).

Punctual requirements. It is clear that if consumptions of money have to be non-negative, envy-free allocations may not exist. Otherwise, or if the social endowment of money is large enough, existence holds (Svensson, 1983a; Maskin, 1987; Alkan, Demange and Gale, 1991; Tadenuma and Thomson, 1993; Ichiishi and Idzik, 1999; Su, 1999). For quasi-linear preferences, several algorithms leading to envy-free allocations are available (Aragones, 1995; Klijn, 2000; Ünver, 2003; Abdulkadiro•lu, Sönmez and Ünver, 2004). Remarkably, envy-free allocations are always efficient (Svensson, 1983b). A variety of selections from the no-envy solution have been proposed (Tadenuma, 1989; 1994; Alkan, Demange and Gale, 1991; Aragones, 1995; Tadenuma and Thomson, 1995).

Egalitarian-equivalent and efficient allocations exist very generally, when preferences are defined over $\mathbb{R} \times A$ and the compensation assumption holds. When preferences are defined over $\mathbb{R}_+ \times A$, existence holds under similar assumptions as the ones guaranteeing that of envy-free allocations. Just as in the classical case, there are economies in which all egalitarian-equivalent and efficient allocations violate no-envy.

The case of one object is special, and the solution that selects the envy-free allocation at which the winner receives the least amount of money has a number of interesting properties and has been characterized on the basis of these properties. This allocation is egalitarian-equivalent, with the losers’

bundle serving as reference bundle (Tadenuma and Thomson, 1993; Thomson, 1998).

The *Walrasian solution* can easily be adapted to the present model but here, an allocation is an equal-income Walrasian allocations if and only if it is envy-free and efficient, and if and only if it is group envy-free (Svensson, 1983b).

Another requirement is that each agent should be made at least as well off as he would be at the (essentially) unique envy-free allocation of the hypothetical economy in which everyone had his preferences. If $|N| = 2$, meeting this *identical-preferences lower bound* is equivalent to no-envy, but if $|N| > 2$, the identical-preferences lower bound is weaker (Bevia, 1996a). Thus, this concept gives us another chance of obtaining positive results when no-envy is too demanding. Unfortunately, there are quasi-linear economies with equal numbers of objects and agents in which all egalitarian-equivalent and efficient allocations violate not only no-envy, as already noted, but in fact the identical-preferences lower bound. When there are more objects than agents, an allocation may be envy-free and efficient without meeting the identical-preferences lower bound, but it does meet the variant of the lower bound obtained by using only the objects that are assigned. No-envy remains incompatible with this bound however (Thomson, 2003).

Relational requirements. Selections from the no-envy solution exist that satisfy a form of *money-monotonicity* (Alkan, Demange and Gale, 1991). Any selection from the egalitarian-equivalence and Pareto solution obtained by fixing the reference object is *money-monotonic*.

Object-monotonicity, the requirement that when additional objects become available, all agents should end up at least as well off as they were initially, makes sense if the objects are desirable or when there are undesirable objects, they do not have to be assigned. To study it, in specifying an economy, we now have to allow the numbers of objects and agents to differ. Then, an envy-free allocation is not necessarily efficient and we explicitly impose efficiency. Unfortunately, no selection from the no-envy and Pareto solution is *object-monotonic*, even if preferences are quasi-linear (Alkan, 1994).

Suppose now that all real objects have to be assigned before any null object is, independently of whether they are desirable. For instance, objects may be activities that some agents enjoy and others do not, but these activities have to be carried out if there are enough agents for that, an example mentioned earlier. Even if preferences are quasi-linear, no selection from a natural weakening of the identical-preferences lower bound and Pareto solution is *weakly object monotonic*, that is, such that the welfares of all agents should be affected in the same direction by an enlargement of the set of objects (Thomson, 2003).

Even if preferences are quasi-linear, no selection from the no-envy solution satisfies *welfare-domination under preference-replacement* (Thomson, 1998).

A first requirement in the context of a variable population is that if the social endowment of money is non-negative and the objects are all desirable, none of the agents initially present should benefit from the arrival of additional agents. Even if preferences are quasi-linear, *population-monotonicity* is incompatible with no-envy (Alkan, 1994; Moulin, 1990b). In fact, an agent could be better off at any envy-free allocation than if he were alone, so that the *free-access upper bound* is incompatible with no-envy.

If there is a single object, which is desirable, and the social endowment of money is zero, a *population-monotonic* selection from the identical-preferences lower bound and Pareto solution can be defined, based on the Shapley value (Moulin, 1990b; Bevia, 1996c). Other positive results can be obtained for that case.

The selection from the egalitarian-equivalence and Pareto solution obtained by requiring the reference bundle to contain a fixed object is *weakly population-monotonic* (the arrival of new agents affects the welfares of all existing agents in the same direction), but it is not guaranteed to be a selection from the no-envy solution any more. In fact, no selection from the no-envy solution is *weakly population-monotonic* (Tadenuma and Thomson, 1995). Weaker requirements pertaining to changes in resources or population are defined and investigated by Alkan (1994).

Turning to *consistency*, we have the following result: if a sub-solution of the no-envy solution is *neutral* (that is, invariant under exchanges of bundles that leave all agents indifferent) and *consistent*, then in fact, it coincides with the no-envy solution (Tadenuma and Thomson, 1991). As always, the no-envy solution is *conversely consistent*, but many proper sub-solutions of it are too (as well as *neutral*). On the other hand, the Pareto solution is not (unless the objects are identical). However, if a sub-solution of the no-envy solution is *neutral*, *bilaterally consistent*, and *conversely consistent*, then in fact it coincides with the no-envy solution (Tadenuma and Thomson, 1991).

The identical-preferences lower bound solution is *conversely consistent* but not *consistent*. The *minimal consistent enlargement* of its intersection with the Pareto solution is the Pareto solution itself. This is true when there is at most one object, when there are multiple identical objects, and when there are multiple and possibly different objects. The maximal *consistent* sub-solution of the identical-preferences lower bound and Pareto solution is the no-envy solution (Bevia, 1996a).

Related models. When each agent can consume several objects (in addition to the infinitely divisible good), the situation is quite different from what it is in the one-object-per-agent case, unless severe additional restrictions are imposed on preferences. In fact, many of the special relations that exist in the one-object-per-agent case disappear, and the situation resembles the classical situation (Tadenuma, 1996; Haake, Raith and Su, 2002).

For preferences that have additive representations, a rule proposed by Knaster (1946) is generalized by Steinhaus (1949) and advocated by Samuelson (1980). An alternative is the selection from the egalitarian-equivalence and Pareto solution obtained by choosing the null object as reference object. Interestingly, it is a selection from the no-envy solution (Willson, 2003). Each is *money-monotonic* and satisfies a form of *object-monotonicity*.

Even if preferences are quasi-linear and no other fairness requirement is imposed, no selection from the Pareto solution is *population-monotonic* (Bevia, 1996b). In contrast to the one-object-per-person case, there are *consistent* sub-solutions of the no-envy and Pareto solution, and *converse consistency* becomes a much stronger requirement. Characterizations have been obtained under an additional invariance requirement on solutions (Bevia, 1998). The *population-monotonicity* of rules that select lotteries is examined by Ehlers and Klaus (2003b).

When monetary compensations are not possible. This situation has recently been much studied, mainly in the one-object-per-agent case when preferences are strict. It is clear that punctual requirements of fairness such as no-envy and egalitarian-equivalence are not generally achievable here (think of situations where all agents have the same preferences). However, most of our relational requirements remain meaningful. The main lesson of the literature is that they can be satisfied, but in a rather limited way, by sequential priority rules and variants. If the objective is to respect an exogenously given priority order of agents, then of course more positive results can be obtained (Svensson, 1994a; Balinski and

Sönmez, 1999; Ergin, 2000; 2002; Ehlers and Klaus, 2006; 2007; Kesten, 2006).

Now, imagine that agents can consume several objects. Herreiner and Puppe (2002) propose a maximin-type criterion, and define an iterative procedure that produces, among the efficient allocations, the one that is best according to this criterion (see also Ramaekers, 2006). In that situation, no selection from the Pareto solution satisfies *welfare-domination under preference replacement* (Klaus and Miyagawa, 2001).

Brams and Fishburn (2000) for $|N|=2$ and Edelman and Fishburn (2001) for $|N|>2$ examine the special case when agents have the same preferences over individual objects but possibly different preferences over sets of objects. Brams, Edelman and Fishburn (2003) drop the assumption that preferences over individual objects are the same, and propose, in addition to requirements related to no-envy, some that are based on comparing the numbers of objects received by the various agents.

The possibility that agents are endowed with objects is considered by Shapley and Scarf (1974), and situations when some objects are initially individually owned and others are commonly owned (residential housing on a university campus being an illustrative example; kidney exchange is another application) are discussed by Roth, Sönmez and Ünver (2004) and Sönmez and Ünver (2005).

Various notions of efficiency for rules that select lotteries are examined by Hylland and Zeckhauser (1979), Demko and Hill (1988), Abdulkadiro•lu and Sönmez (1998; 1999), and Bogomolnaia and Moulin (2001; 2002).

When objects cannot be transferred. Consider the problem of allocating a single infinitely divisible good, ‘money’, among agents characterized by variables that cannot be transferred (talent or handicaps for examples), and thus can be thought of ‘objects’. How should money be divided to compensate agents for possible differences in these variables? This question, formulated by Fleurbaey (1994; 1995a), has given rise to a large literature. For a detailed survey, see Fleurbaey and Maniquet (2008).

7 Single-peaked preferences

Consider the problem of allocating a social endowment of an infinitely divisible and non-disposable commodity among a group of agents whose preferences over $\mathbb{R}_+$ are single-peaked: up to some critical level, his *peak amount*, an increase in an agent's consumption increases his welfare but beyond that level, the opposite holds. Since there is no possibility of disposal, the social endowment has to be fully distributed. If the sum of the peak amounts is greater than the social endowment, ‘there is not enough’, and for efficiency, no agent should consume more than his peak amount. If the inequality goes the other way, ‘there is too much’; here, for efficiency, no agent should consume less than his peak amount (Sprumont, 1991).

Punctual requirements. Efficient allocations meeting the equal-division lower bound, or no-envy, in fact both, always exist. The equal-division core and the group-no-envy solution may be empty, but natural variants of these solutions are not.

A number of interesting rules can be defined: the commodity can be divided proportionally to the peak amounts, or so that all agents’ consumptions are at the same distance from their peak amounts subject to non-negativity, or so that the sizes of their upper contour sets at their assigned consumptions are equal, or as equal as possible. The following rule, called the *uniform rule*, will be central: if there is not enough, and given $\lambda \geq 0$, assign to each agent the amount he prefers in $[0, \lambda]$; choose λ so that the sum

of these assignments is equal to the social endowment; if there is too much, given $\lambda \geq 0$, assign to each agent the amount he prefers in $[\lambda, \infty]$; here too, choose λ so that the sum of these assignments is equal to the social endowment.

The uniform rule depends only on the profile of peak amounts – it satisfies the *peak-only* requirement – and it is the only subsolution of the no-envy and Pareto solution to do so (Thomson, 1994c). Also, it is the only selection from the Pareto solution minimizing (*a*) the difference between the smallest amount anyone receives and the greatest amount anyone receives; (*b*) alternatively, the variance of the amounts they all receive (Schummer and Thomson, 1997).

Relational requirements. Here, the natural expression of the idea of solidarity when the social endowment varies is that all agents should be made at least as well off as they were initially *or* that they should all be made at most as well off. This requirement is incompatible with no-envy (or with the equal-division lower bound). This is because a change in the social endowment can be so disruptive that it turns an economy in which there is not enough to one in which there is too much, or converse. This suggests limiting its application to situations in which no such switches occur, yielding *one-sided resource-monotonicity*. This property is much less demanding. Solidarity requirements with respect to changes in population or preferences can similarly be modified by limiting their application to situations in which the direction of the inequality between the sum of the peak amounts and the social endowment is not reversed by the change under consideration. We add the suffix 'one-sided' to indicate the weaker versions so defined. We also consider *separability*, which says that given two economies having a group of agents in common, if the agents in this group receive the same aggregate amount in both, then each of them should receive the same amount in both.

We have the following characterizations, some of which require that each preference relation be such that if its peak amount is positive, there is an amount greater than the peak amount that is indifferent to 0. The uniform rule is (*a*) the only selection from the no-envy and Pareto solution to be *one-sided resource-monotonic* (Thomson, 1994b); (*b*) the only selection from the no-envy and Pareto solution to satisfy *replication-invariance* and *one-sided welfare-domination under preference-replacement* (Thomson, 1997); (*c*) the only selection from the no-envy and Pareto solution to be *replication-invariant* and *one-sided population-monotonic* (Thomson, 1995a); (*d*) the only selection from the no-envy and Pareto solution to be *resource-continuous* and *separable* (Chun, 2003; 2006). (*d*) the smallest (in terms of inclusion) subsolution of the no-envy and Pareto solution to be *resource upper hemi-continuous* and *consistent* (Thomson, 1994c); (*e*) the only *single-valued* selection from the equal-division lower bound and Pareto solution to be *replication-invariant* and *consistent*, or to be *anonymous* and *conversely consistent* (Thomson, 1995c).

Many refinements and variants of these results are available (Sönmez, 1994; Klaus, 1997; 1999; 2006; Dagan, 1996; Moulin, 1999; Herrero and Villar, 1998; 2000; Ehlers, 2002a; 2002b; Kesten, 2004b). An application to a pollution problem is by K•br•s (2003).

Related models. Fairness issues have been analysed for the variant of the model obtained by introducing individual endowments (Thomson, 1995c; Klaus, 1997; 2001; Klaus, Peters and Storcken, 1997; Moreno, 2002).

For economies with both individual endowments and a social endowment, different ways of adapting the punctual fairness requirements have been proposed, and issues of *monotonicity*, with respect to the individual endowments and the social endowment, in addition to *consistency* and *population-*

monotonicity, have been addressed (Thomson, 1995c; Klaus, 1997; Herrero, 2002). In these studies, a rule that is the natural extension of the uniform rule most frequently emerges.

A multi-commodity version of the single-peaked assumption is easily defined. For such a model, a generalization of the equal-slacks Walrasian solution (Mas-Colell, 1992) is axiomatized along the lines of Schummer and Thomson's (1997) axiomatization of the uniform rule (Amoros, 1999). A probabilistic version of the uniform rule is characterized by Sasaki (1997).

8 Non-homogeneous continuum

Here, we consider the problem of dividing a heterogeneous commodity modelled as measure space, each agent having preferences defined over the measurable subsets, and the question being how to select partitions consisting of measurable subsets, one for each agent. Think of a cake on which frosting and decorations are distributed unevenly. Often, this commodity is embedded in a finite-dimensional Euclidean space: an example is land.

Punctual requirements. In such situations, equal division has no economic meaning, even when it can be defined in physical terms (surface area, say, or weight). However, our central criteria (no-envy; egalitarian-equivalence) remain applicable. A large literature concerns preferences that can be represented by atomless measures, a somewhat restrictive assumption that precludes complementarities between different parts of the dividend. Additional topological and geometric criteria are sometimes meaningful (Hill, 1983). The construction of iterative procedures leading to partitions satisfying some fairness requirement, exactly or in some approximate sense, has been important in the literature, but until recently, efficiency had often been ignored.

If no restrictions are imposed on preferences apart from continuity and monotonicity with respect to set inclusion, envy-free and efficient partitions may not exist (Berliant, Dunz and Thomson, 1992). However, when preferences are representable by atomless measures, they do (Weller, 1985). An existence result for group envy-free partitions is also available (Berliant, Dunz and Thomson, 1992).

An interesting special case is the one-dimensional case when the dividend is an interval that has to be partitioned into subintervals, one for each agent. It has many applications: division of an interval of time, a length of road, and so on. When preferences are represented by atomless measures, envy-free partitions exist (Woodall, 1980), but in fact existence then holds under much weaker assumptions (Stromquist, 1980; Su, 1999). In the case of a closed curve, the situation is much less satisfactory (Barbanel and Brams, 2005; Thomson, 2007). Under monotonicity of preferences, no-envy implies efficiency (Berliant, Dunz and Thomson, 1992). Under continuity and strict monotonicity of preferences, egalitarian-equivalent and efficient allocations exist (Berliant, Dunz and Thomson, 1992).

When preferences are represented by atomless measures, the $\frac{1}{n}$ – *lower-bound* is that for each agent, the value to him of his assignment should be at least $\frac{1}{n}$ times his value of the dividend. Some of the early literature searched for partitions such that for each agent, this bound is met as an equality. Given a list $\alpha \in \Delta^N$ of ‘shares’, the α – *lower-bound* is that for each $i \in N$, the value to agent *i* of his assignment should be at least α_i times his value of the dividend. Partitions satisfying these notions and generalizations exist (Berliant, Dunz and Thomson, 1992; Barbanel and Zwicker, 2001; Reijnierse and

Potters, 1998). An existence result is available when preferences are representable by atomless concave capacities (Maccheroni and Marinacci, 2003). The existence of envy-free partitions is also known for a more general notion of a partition, where agents receive ‘fractional’ consumptions of each point of the dividend (Akin, 1995).

A succession of attempts at generalizing to more than two agents the classical two-person divide-and-choose scheme (one agent divides and the other chooses one of the two pieces; the divider receives the other), have been made over the years that generate partitions that are either envy-free or meet the $\frac{1}{n}$ – *lower-bound*. It took many years until an algorithm that produces an envy-free partition in the *n*-person case, for arbitrary *n*, was discovered (Brams and Taylor, 1995). None of the solutions proposed necessarily attains efficiency.

Brams and Taylor (1996) survey the literature. Robertson and Webb (1998) focus on algorithms and pay little attention to efficiency. On the other hand, Barbanel (2005) provides an in-depth analysis of the shape of the image of the set of feasible partitions in a Euclidean space of dimension equal to the number of agents, using their measures as representations of their preferences. It offers characterizations of its subset of efficient points. It also gives existence results for efficient and envy-free partitions.

9 Other domains

We conclude this survey by tying it to literatures concerning other models but also addressing fairness issues. They concern (*a*) the Arrovian model of extended sympathy; (*b*) rights assignments; (*c*) quasi-linear social choice; (*d*) intertemporal allocation; (*e*) public choice from an interval or a closed curve when agents have single-peaked preferences; (*f*) public good production; (*g*) cost sharing; (*h*) queuing, scheduling, and sequencing; (*i*) matching.

See Also

- efficient allocation
- equality of opportunity
- justice
- justice (new perspectives)
- Shapley value

The author would like to thank the National Science Foundation for its support under grant SES 0214691.

Bibliography

Abdulkadiro•lu, A. and Sönmez, T. 1998. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66, 689–701.

Abdulkadiro•lu, A. and Sönmez, T. 1999. House allocation with existing tenants. *Journal of Economic Theory* 88, 233–60.

Abdulkadiro•lu, A., Sönmez, T. and Ünver, U. 2004. Room assignment-rent division: a market approach. *Social Choice and Welfare* 22, 515–38.

Akin, E. 1995. Vilfredo Pareto cuts the cake. *Journal of Mathematical Economics* 24, 23–44.

Alkan, A. 1994. Monotonicity and envyfree assignments. *Economic Theory* 4, 605–16.

Alkan, A., Demange, G. and Gale, D. 1991. Fair allocation of indivisible goods and criteria of justice. *Econometrica* 59, 1023–39.

Amoros, P. 1999. Efficiency and income redistribution in the single-peaked preference model with several commodities. *Economics Letters* 63, 341–49.

Aragones, E. 1995. A derivation of the money Rawlsian solution. *Social Choice and Welfare* 12, 267–76.

Archibald, P. and Donaldson, D. 1979. Notes on economic inequality. *Journal of Public Economics* 12, 205–14.

Aumann, R. and Peleg, B. 1974. A note on Gale's example. *Journal of Mathematical Economics* 1, 209–11.

Balinski, M. and Sönmez, T. 1999. A tale of two mechanisms: student placement. *Journal of Economic Theory* 84, 73–94.

Barbanel, J. 2005. *The Geometry of Efficient Fair Division*. Cambridge: Cambridge University Press.

Barbanel, J. and Brams, S. 2005. Cutting a pie is not a piece of cake. Mimeo, New York University.

Barbanel, J. and Zwicker, W. 2001. Two applications of a theorem of Dvoretsky, Wald, and Wolfovitz to cake division. *Theory and Decision* 43, 639–50.

Berliant, M., Dunz, K. and Thomson, W. 1992. On the fair division of a heterogeneous commodity. *Journal of Mathematical Economics* 21, 201–16.

Bevia, C. 1996a. Identical preferences lower bound and consistency in economies with indivisible goods. *Social Choice and Welfare* 13, 113–26.

Bevia, C. 1996b. Population monotonicity in a general model with indivisible goods. *Economics Letters*

50, 91–7.

Bevia, C. 1996c. Population monotonicity in economies with one indivisible good. *Mathematical Social Sciences* 32, 125–37.

Bevia, C. 1998. Fair allocation in a general model with indivisible goods. *Review of Economic Design* 3, 195–213.

Bogomolnaia, A. and Moulin, H. 2001. A new solution to the random assignment problem. *Journal of Economic Theory* 100, 295–328.

Bogomolnaia, A. and Moulin, H. 2002. A simple random assignment problem with a unique solution. *Economic Theory* 19, 298–317.

Brams, S., Edelman, P. and Fishburn, P. 2003. Fair division of indivisible items. *Theory and Decision* 55, 147–80.

Brams, S. and Fishburn, P. 2000. Fair division of indivisible items between two people with identical preferences: envy-freeness, Pareto-optimality, and equity. *Social Choice and Welfare* 17, 247–67.

Brams, S. and Taylor, A.D. 1995. An envy-free cake division protocol. *American Mathematical Monthly* 102, 9–18.

Brams, S. and Taylor, A.D. 1996. *Fair Division: from Cake-Cutting to Dispute Resolution*. Cambridge: Cambridge University Press.

Champsaur, P. and Laroque, G. 1981. Fair allocations in large economies. *Journal of Economic Theory* 25, 269–82.

Chaudhuri, A. 1985. Formal properties of interpersonal envy. *Theory and Decision* 18, 301–12.

Chaudhuri, A. 1986. Some implications of an intensity measure of envy. *Social Choice and Welfare* 3, 255–70.

Chichilnisky, G. and Thomson, W. 1987. The Walrasian mechanism from equal division is not monotonic with respect to variations in the number of consumers. *Journal of Public Economics* 32, 119–24.

Chun, Y. 2003. One-sided population-monotonicity, separability, and the uniform rule. *Economics Letters* 78, 343–49.

Chun, Y. 2006. The separability principle in economies with single-peaked preferences. *Social Choice and Welfare* 26, 239–53.

Chun, Y. and Thomson, W. 1988. Monotonicity properties of bargaining solutions when applied to economics. *Mathematical Social Sciences* 15, 11–27.

Dagan, N. 1995. Consistent solutions in exchange economies: a characterization of the price mechanism. Mimeo, Pompeu Fabra University.

Dagan, N. 1996. A note on Thomson's characterization of the uniform rule. *Journal of Economic Theory* 69, 255–61.

Daniel, T. 1975. A revised concept of distributional equity. *Journal of Economic Theory* 11, 94–109.

Daniel, T. 1978. Pitfalls in the theory of fairness: comment. *Journal of Economic Theory* 19, 561–4.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

Demko, S. and Hill, T. 1988. Equitable distribution of indivisible objects. *Mathematical Social Sciences* 16, 145–58.

Diamantaras, D. 1992. On equity with public goods. *Social Choice and Welfare* 9, 141–57.

Diamantaras, D. and Thomson, W. 1990. An extension and refinement of the no-envy concept. *Economics Letters* 33, 217–22.

Dutta, B. and Vohra, R. 1993. A characterization of egalitarian equivalence. *Economic Theory* 4, 465–79.

Edelman, P. and Fishburn, P. 2001. Fair division of indivisible items among people with similar preferences. *Mathematical Social Sciences* 41, 327–47.

Ehlers, L. 2002a. Resource-monotonic allocation when preferences are single-peaked. *Economic Theory* 20, 113–31.

Ehlers, L. 2002b. A characterization of the uniform rule without Pareto-optimality. Mimeo, University of Montreal.

Ehlers, L. and Klaus, B. 2003a. Coalitional strategy-proofness and resource-monotonicity for multiple assignment problems. *Social Choice and Welfare* 21, 265–80.

Ehlers, L. and Klaus, B. 2003b. Probabilistic assignments of identical indivisible objects and the probabilistic uniform correspondence. *Review of Economic Design* 8, 249–68.

Ehlers, L. and Klaus, B. 2003c. Resource-monotonicity for house allocation problems. *International Journal of Game Theory* 32, 545–60.

Ehlers, L. and Klaus, B. 2006. Efficient priority rules. *Games and Economic Behavior* 55, 372–84.

Ehlers, L. and Klaus, B. 2007. Consistent house allocation. *Economic Theory* 30, 561–74.

Ehlers, L., Klaus, B. and Pápai, S. 2002. Strategy-proofness and population-monotonicity for house allocation problems. *Journal of Mathematical Economics* 38, 329–39.

Ergin, H. 2000. Consistency in house allocation problems. *Journal of Mathematical Economics* 34, 77–97.

Ergin, H. 2002. Efficient resource allocation on the basis of priorities. *Econometrica* 70, 2489–97.

Feldman, A. and Kirman, A. 1974. Fairness and envy. *American Economic Review* 64, 995–1005.

Fleurbaey, M. 1994. On fair compensation. *Theory and Decision* 36, 277–307.

Fleurbaey, M. 1995a. The requisites of equal opportunity. In *Social Choice, Welfare, and Ethics*, ed. W. A. Barnett, H. Moulin, M. Salles and N. Schofield. Cambridge: Cambridge University Press.

Fleurbaey, M. 1995b. Three solutions for the compensation problem. *Journal of Economic Theory* 65, 505–21.

Fleurbaey, M. 1995c. Equal opportunity or equal social outcomes. *Economics and Philosophy* 11, 25–55.

Fleurbaey, M. and Maniquet, F. 1996a. Cooperative production: a comparison of welfare bounds. *Games and Economic Behavior* 17, 200–8.

Fleurbaey, M. and Maniquet, F. 1996b. Fair allocation with unequal production skills: the no-envy approach to compensation. *Mathematical Social Sciences* 32, 71–93.

Fleurbaey, M. and Maniquet, F. 1999. Cooperative production with unequal skills: the solidarity approach to compensation. *Social Choice and Welfare* 16, 569–83.

Fleurbaey, M. and Maniquet, F. 2008. Compensation and responsibility. In *Handbook of Social Choice and Welfare*, vol. 2, ed. K. Arrow, A. Sen and K. Suzumura. Amsterdam: North-Holland.

Foley, D. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7, 45–98.

Geanakoplos, J. and Nalebuff, B. 1988. On a fundamental conflict between equity and efficiency. Mimeo, Yale University.

Gevers, L. 1986. Walrasian social choice: some simple axiomatic approaches. In *Social Choice and Public Decision Making, Essays in Honor of K.J. Arrow*, ed. W.P. Heller, R.M. Starr and D.A. Starrett. Cambridge: Cambridge University Press.

Ginés, M. and Marhuenda, F. 1996. Cost monotonic mechanisms. *Investigaciones Economicas* 20, 89–103.

Goldman, S. and Starr, R. 1982. Pairwise, *t*-wise and Pareto-optimality. *Econometrica* 50, 593–606.

Goldman, S. and Sussangkarn, C. 1980. On equity and efficiency. *Economics Letters* 5, 29–31.

Green, J. 1972. On the inequitable nature of core allocations. *Journal of Economic Theory* 4, 132–43.

Haake, C.-J., Raith, M. and Su, F. 2002. Bidding for envy-freeness: a procedural approach to *n*-player fair-division problems. *Social Choice and Welfare* 19, 723–49.

Herreiner, D. and Puppe, C. 2002. A simple procedure for finding equitable allocations of indivisible goods. *Social Choice and Welfare* 19, 415–30.

Herrero, C. 2002. General allocation problems with single-peaked preferences: path-independence and related topics. *Spanish Economic Review* 4, 19–40.

Herrero, C. and Villar, A. 1998. The equal-distance rule in allocation problems with single-peaked preferences. In *Current Trends in Economic Theory and Applications*, ed. R. Aliprantis and N. Yannelis. Berlin: Springer.

Herrero, C. and Villar, A. 2000. An alternative characterization of the equal-distance rule for allocation problems with single-peaked preferences. *Economics Letters* 66, 311–7.

Hill, T.P. 1983. Determining a fair border. *American Mathematical Monthly* 90, 438–42.

Hylland, A. and Zeckhauser, R. 1979. The efficient allocations of individual to positions. *Journal of Political Economy* 87, 293–314.

Ichiishi, T. and Idzik, A. 1999. Equitable allocation of divisible goods. *Journal of Mathematical Economics* 32, 389–400.

Jaskold-Gabszewicz, J.-J. 1975. Coalitional fairness of allocations in pure exchange. *Econometrica* 43, 661–8.

Kesten, O. 2004a. Coalition strategy-proofness and resource monotonicity for house allocation problems. Mimeo, Carnegie-Melon University.

Kesten, O. 2004b. More on the uniform rule: characterizations without Pareto-optimality. Mimeo, Carnegie-Melon University.

Kesten, O. 2006. On two competing mechanisms for priority-based allocation problems. *Journal of Economic Theory* 27, 155–71.

Khan, A. and Polemarchakis, H. 1978. Unequal treatment in the core. *Econometrica* 46, 1475–81.

K•br•s, O. 2003. Permit allocation problems. *Social Choice and Welfare* 20, 353–62.

Kim, H. 2001. Essays on fair allocations, Ph.D. thesis.

Kim, H. 2004. Population monotonicity for fair allocation problems. *Social Choice and Welfare* 23, 59–70.

Klaus, B. 1997. The characterization of the uniform reallocation rule without side-payments. In *Game Theoretic Applications to Economics and Operations Research*, ed. T. Parthasarathy, B. Dutta, J.A.M. Potters, T.E.S. Raghavan, D. Ray and A. Sen. Dordrecht: Kluwer-Academic.

Klaus, B. 1999. The role of replication-invariance: two answers concerning the problem of fair division when preferences are single-peaked. Mimeo, University of Maastricht.

Klaus, B. 2001. Uniform allocation and reallocation revisited. *Review of Economic Design* 6, 85–98.

Klaus, B. 2006. A note on the separability principle in economies with single-peaked preferences. *Social Choice and Welfare* 26, 255–61.

Klaus, B. and Miyagawa, E. 2001. Strategy-proofness, solidarity, and consistency for multiple assignment problems. *International Journal of Game Theory* 30, 421–35.

Klaus, B., Peters, H. and Storcken, T. 1997. Reallocation of an infinitely divisible good. *Economic Theory* 10, 305–33.

Kleinberg, N. 1980. Fair allocations and equal incomes. *Journal of Economic Theory* 23, 189–200.

Klijn, F. 2000. An algorithm for envy-free allocations in an economy with indivisible objects and money. *Social Choice and Welfare* 17, 201–15.

Knaster, B. 1946. Sur le problème du partage pragmatique de H. Steinhaus. *Annales de la Société Polonaise de Mathématique* 19, 228–30.

Kolm, S. 1972. *Justice et Equité*. Paris: Editions du Centre National de la Recherche Scientifique, 1972. English edn. Cambridge, MA: MIT Press, 1988.

Kolm, S. 1973. Super-equité. *Kyklos* 26, 841–3.

Kolpin, V. 1991a. Resolving open questions on the λ^*-envy-free criterion. *Economics Letters* 36, 17–20.

Kolpin, V. 1991b. Equity and the core. *Mathematical Social Sciences* 22, 137–50.

Korthues, B. 2000. Characterization of an extended Walrasian concept for open economies. *Journal of Mathematical Economics* 33, 449–61.

Maccheroni, F. and Marinacci, M. 2003. How to cut a pizza fairly: fair division with decreasing marginal evaluations. *Social Choice and Welfare* 20, 457–65.

Maniquet, F. 1996a. Horizontal equity and stability when the number of agents is variable in the fair division problem. *Economics Letters* 50, 85–90.

Maniquet, F. 1996b. Allocation rules for a commonly owned technology: the average cost lower bound. *Journal of Economic Theory* 69, 490–508.

Maniquet, F. 1999. A strong incompatibility between efficiency and equity in non-convex economies. *Journal of Mathematical Economics* 32, 467–74.

Maniquet, F. 2002. A study of proportionality and simplicity in the cooperative production problem. *Review of Economic Design* 7, 1–15.

Maniquet, F. and Sprumont, Y. 2000. On resource-monotonicity in the fair division problem. *Economics Letters* 68, 299–302.

Mas-Colell, A. 1980. Remarks on the game theoretic analysis of a simple distribution of surplus problems. *International Journal of Game Theory* 9, 125–40.

Mas-Colell, A. 1987. On the second welfare theorem for anonymous net trades ain exchange economies with many agents. In *Information, Incentives and Economic Mechanisms*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Mas-Colell, A. 1992. Equilibrium theory with possibly satiated preferences. In *Equilibrium and Dynamics; Essays in honor of D. Gale*, ed. M. Majumdar. London: Macmillan.

Maskin, E. 1987. On the fair allocation of indivisible goods. In *Arrow and the Foundations of the Theory of Economic Policy*, ed. G. Feiwel. London: Macmillan.

Moreno, B. 2002. Single-peaked preferences, population-monotonicity and endowments. *Economics Letters* 75, 87–95.

Moulin, H. 1987a. A core selection for pricing a single-output monopoly. *Rand Journal of Economics* 18, 397–407.

Moulin, H. 1987b. The pure compensation problem: egalitarianism versus laissez-fairism. *Quarterly Journal of Economics* 101, 769–83.

Moulin, H. 1990a. Uniform preference externalities: two axioms for fair allocation. *Journal of Public Economics* 43, 305–26.

Moulin, H. 1990b. Monotonic surplus-sharing and the utilization of common property resources. In *Game Theory and Applications*, ed. T. Ichiishi, A. Neyman and Y. Tauman. New York: Academic Press.

Moulin, H. 1990c. Fair division under joint ownership: recent results and open problems. *Social Choice and Welfare* 7, 149–70.

Moulin, H. 1990d. Joint ownership of a convex technology: comparisons of three solutions. *Review of Economic Studies* 57, 439–52.

Moulin, H. 1990e. Interpreting common ownership. *Recherches Economiques de Louvain* 56, 303–26.

Moulin, H. 1991. Welfare bounds in the fair division problem. *Journal of Economic Theory* 54, 321–37.

Moulin, H. 1992a. An application of the Shapley value to fair division with money. *Econometrica* 60, 1331–49.

Moulin, H. 1992b. All sorry to disagree: a general principle for the provision of non-rival goods. *Scandinavian Journal of Economics* 94, 37–51.

Moulin, H. 1993. On the fair and coalitionproof allocation of private goods. In *Frontiers of Game Theory*, ed. K. Binmore, A. Kirman and P. Tani. Cambridge, MA: MIT Press.

Moulin, H. 1995. *Cooperative Microeconomics: a Game-Theoretic Introduction*. Princeton: Princeton University Press.

Moulin, H. 1996. Cost sharing under increasing returns: a comparison of simple mechanisms. *Journal of Economic Theory* 13, 225–51.

Moulin, H. 1999. Rationing a commodity along fixed paths. *Journal of Economic Theory* 84, 41–72.

Moulin, H. 2003. *Fair Division and Collective Welfare*. Cambridge, MA: MIT Press.

Moulin, H. and Roemer, J. 1989. Public ownership of the external world and private ownership of self. *Journal of Political Economy* 97, 347–67.

Moulin, H. and Shenker, S. 1994. Average cost pricing versus serial cost sharing: an axiomatic comparison. *Journal of Economic Theory* 64, 178–207.

Moulin, H. and Thomson, W. 1988. Can everyone benefit from growth? Two difficulties. *Journal of Mathematical Economics* 17, 339–45.

Nagahisa, R. 1994. A necessary and sufficient condition for Walrasian social choice. *Journal of Economic Theory* 62, 186–208.

Nicolò, A. and Perea, A. 2005. Monotonicity and equal-opportunity equivalence in bargaining. *Mathematical Social Sciences* 49, 221–43.

Otsuki, M. 1980. On distribution according to labor: a concept of fairness in production. *Review of Economic Studies* 47, 945–58.

Pazner, E. 1977. Pitfalls in the theory of fairness. *Journal of Economic Theory* 14, 458–66.

Pazner, E. and Schmeidler, D. 1974. A difficulty in the concept of fairness. *Review of Economic Studies* 41, 441–3.

Pazner, E. and Schmeidler, D. 1978. Egalitarian equivalent allocations: a new concept of economic equity. *Quarterly Journal of Economics* 92, 671–87.

Piketty, T. 1994. Existence of fair allocations in economies with production. *Journal of Public Economics* 55, 391–405.

Ramaekers, E. 2006. Fair allocation of indivisible goods without monetary compensations. Mimeo, University of Namur.

Reijnierse, J.H. and Potters, J.A.M. 1998. On finding an envy-free Pareto-optimal division. *Mathematical Programming* 83, 291–311.

Robertson, J. and Webb, W. 1998. *Cake-Cutting Algorithms*. Natick: A.K. Peters.

Roemer, J. 1986a. The mismarriage of bargaining theory and distributive justice. *Ethics* 97, 88–110.

Roemer, J. 1986b. Equality of resources implies equality of welfare. *Quarterly Journal of Economics* 101, 751–84.

Roemer, J. 1988. Axiomatic bargaining on economic environments. *Journal of Economic Theory* 45, 1–35.

Roemer, J. 1996. *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.

Roemer, J. and Silvestre, J. 1993. The proportional solution for economies with both private and public ownership. *Journal of Economic Theory* 59, 426–44.

Roth, A., Sönmez, T. and Ünver, U. 2004. Kidney exchange. *Quarterly Journal of Economics* 119, 457–88.

Samuelson, W. 1980. The object distribution problem revisited. *Quarterly Journal of Economics* 94, 85–98.

Sasaki, H. 1997. Randomized uniform allocation mechanism and single-peaked preferences of indivisible good. Working paper, Waseda University.

Schmeidler, D. and Vind, K. 1972. Fair net trades. *Econometrica* 40, 637–42.

Schummer, J. and Thomson, W. 1997. Two derivations of the uniform rule. *Economics Letters* 55, 333–7.

Serrano, R. and Volij, O. 1998. Axiomatizations of neoclassical concepts for economies. *Journal of Mathematical Economics* 30, 87–108.

Shapley, L. and Scarf, H. 1974. On cores and indivisibility. *Journal of Mathematical Economics* 1, 23–37.

Shitovitz, B. 1992. Coalitional fair allocations in smooth mixed markets with an atomless sector. *Mathematical Social Sciences* 25, 27–40.

Sönmez, T. 1994. Consistency, monotonicity and the uniform rule. *Economics Letters* 46, 229–35.

Sönmez, T. and Ünver, U. 2005. House allocation with existing tenants: an equivalence. *Games and Economic Behavior* 52, 153–85.

Sprumont, Y. 1991. The division problem with single-peaked preferences: a characterization of the uniform allocation rule. *Econometrica* 59, 509–19.

Sprumont, Y. 1996. Axiomatizing ordinal welfare egalitarianism when preferences vary. *Journal of Economic Theory* 68, 77–110.

Sprumont, Y. and Zhou, L. 1999. Pazner-Schmeidler rules in large society. *Journal of Mathematical Economics* 31, 321–39.

Steinhaus, H. 1949. Sur la division pragmatique. *Econometrica* 17, 315–9.

Stromquist, N. 1980. How to cut a cake fairly. *American Mathematics Monthly* 87, 640–4.

Su, F. 1999. Rental harmony: Sperner's lemma in fair division. *American Mathematical Monthly* 106, 930–42.

Svensson, L.-G. 1983a. On the existence of fair allocations. *Zeitschrift für National Oekonomie* 43, 301–8.

Svensson, L.-G. 1983b. Large indivisibilities: an analysis with respect to price equilibrium and fairness. *Econometrica* 51, 939–54.

Svensson, L.-G. 1994a. Queue allocation of indivisible goods. *Social Choice and Welfare* 11, 323–30.

Svensson, L.-G. 1994b. σ-optimality and fairness. *International Economic Review* 35, 527–31.

Tadenuma, K. 1989. On the single-valuedness of a solution for the problem of fair allocation in economies with indivisibilities. Mimeo, University of Rochester.

Tadenuma, K. 1994. On strongly envy-free allocations in economies with indivisible goods. Mimeo, University of Rochester.

Tadenuma, K. 1996. Trade-off between equity and efficiency in a general model with indivisible goods.

Social Choice and Welfare 13, 445–50.

Tadenuma, K. and Thomson, W. 1991. No-envy and consistency in economies with indivisible goods. *Econometrica* 59, 1755–67.

Tadenuma, K. and Thomson, W. 1993. The fair allocation of an indivisible good when monetary compensations are possible. *Mathematical Social Sciences* 25, 117–32.

Tadenuma, K. and Thomson, W. 1995. Refinements of the no-envy solution in economies with indivisible goods. *Theory and Decision* 39, 189–206.

Thomson, W. 1978. Monotonic allocation mechanisms: preliminary results. Mimeo, University of Rochester.

Thomson, W. 1983a. Equity in exchange economies. *Journal of Economic Theory* 29, 217–44.

Thomson, W. 1983b. The fair division of a fixed supply among a growing population. *Mathematics of Operations Research* 8, 319–26.

Thomson, W. 1987. Monotonic allocation rules. Revised 1989. University of Rochester.

Thomson, W. 1988. A study of choice correspondences in economies with a variable number of agents. *Journal of Economic Theory* 46, 237–54.

Thomson, W. 1992. Consistency in exchange economies. Mimeo, University of Rochester.

Thomson, W. 1994a. Notions of equal, and equivalent, opportunities. *Social Choice and Welfare* 11, 137–56.

Thomson, W. 1994b. Resource-monotonic solutions to the problem of fair division when preferences are single-peaked. *Social Choice and Welfare* 11, 205–23.

Thomson, W. 1994c. Consistent solutions to the problem of fair division when preferences are single-peaked. *Journal of Economic Theory* 63, 219–45.

Thomson, W. 1994d. Consistent extensions. *Mathematical Social Sciences* 28, 35–49.

Thomson, W. 1995a. Population-monotonic solutions to the problem of fair division when preferences are single-peaked. *Economic Theory* 5, 229–46.

Thomson, W. 1995b. The theory of fair allocation. Mimeo, University of Rochester.

Thomson, W. 1995c. Endowment monotonicity in economies with single-peaked preferences. Mimeo, University of Rochester.

Thomson, W. 1996. The replacement principle in classical economies with private goods. Mimeo, University of Rochester.

Thomson, W. 1997. The replacement principle in private good economies with single-peaked preferences. *Journal of Economic Theory* 76, 145–68.

Thomson, W. 1998. The replacement principle in economies with indivisible goods. *Social Choice and Welfare* 15, 57–66.

Thomson, W. 1999. Welfare-domination and preference-replacement: a survey and open questions. *Social Choice and Welfare* 16, 373–94.

Thomson, W. 2003. On monotonicity in economies with indivisible goods. *Social Choice and Welfare* 21, 195–205.

Thomson, W. 2006a. Population-monotonic allocation rules. Mimeo, University of Rochester.

Thomson, W. 2006b. Consistent allocation rules. Mimeo, University of Rochester.

Thomson, W. 2006c. Fair allocation rules. Mimeo, University of Rochester.

Thomson, W. 2007. Children crying at birthday parties; Why? *Economic Theory* 31, 501–21.

Thomson, W. and Zhou, L. 1993. Consistent allocation rules in atomless economies. *Econometrica* 61, 575–87.

Ünver, U. 2003. Market mechanisms for fair division with indivisible objects and money. Mimeo, University of Pittsburg.

van den Nouweland, A., Peleg, B. and Tijs, S. 1996. Axiomatic characterizations of the Walras correspondence for generalized economies. *Journal of Mathematical Economics* 25, 355–72.

Varian, H. 1974. Equity, envy, and efficiency. *Journal of Economic Theory* 9, 63–91.

Varian, H. 1975. Distributive justice, welfare economics, and the theory of fairness. *Philosophy and Public Affairs* 4, 223–47.

Varian, H. 1976. Two problems in the theory of fairness. *Journal of Public Economics* 5, 249–60.

Vohra, R. 1992. Equity and efficiency in non-convex economies. *Social Choice and Welfare* 9, 185–202.

Watts, A. 1999. Cooperative production: a comparison of lower and upper bounds. *Journal of Mathematical Economics* 32, 317–31.

Weller, D. 1985. Fair division of measurable space. *Journal of Mathematical Economics* 14, 5–17.

Willson, S. 2003. Money-egalitarian-equivalent and gain-maximin allocations of indivisible items with monetary compensation. *Social Choice and Welfare* 20, 247–59.

Woodall, J.R. 1980. Dividing a cake fairly. *Journal of Mathematical Analysis and Applications* 78, 233–47.

Yannelis, N. 1983. Existence and fairness of value allocation without convex preferences. *Journal of Economic Theory* 31, 283–92.

Yoshihara, N. 1998. Characterizations of public and private ownership solutions. *Mathematical Social Sciences* 35, 165–84.

Young, P. 1994. *Equity*. Princeton: Princeton University Press.

Zhou, L. 1992. Strictly fair allocations and Walrasian equilibria in large exchange economies. *Journal of Economic Theory* 57, 158–75.

How to cite this article

Thomson, William. "fair allocation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000303> doi:10.1057/9780230226203.0547

The New Palgrave Dictionary of Economics Online

falsificationism

Daniel M. Hausman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Blaug, M.; falsificationism; Friedman, M.; Popper, K.; scientific method; theory appraisal; verification

Article

Many economists would emphasize that scientific claims must be capable of *falsification*. According to Milton Friedman, an hypothesis ‘is rejected if its predictions are contradicted.… Factual evidence can never “prove” a hypothesis; it can only fail to disprove it …’ (1953, p. 9). These claims echo Karl Popper's philosophy of science, which, on one interpretation, maintains that what distinguishes scientific theories from theories that are not scientific is that scientific theories are falsifiable. A theory is falsifiable if it is logically inconsistent with some finite set of ‘basic statements’ – that is, true or false reports of observation. A true theory will not be inconsistent with any set of true basic statements, but it will still be falsifiable because it is inconsistent with (or ‘forbids’) some observation reports. In other words, logic and observation can force one to give up falsifiable theories. Popper notes that there is an asymmetry between falsification and verification: basic statements can be logically inconsistent with universal generalizations and can thereby disprove them, but they do not imply that any universal generalizations are true. In his view scientific knowledge grows exclusively from falsification. Verification and even confirmation are impossible.

Although Popper distinguishes theories that are falsifiable from theories that are not falsifiable, he is also distinguishes the ‘critical’ attitudes and norms that characterize scientists – who are willing to test theories harshly and to give up claims that do not pass the test – from the dogmatic attitudes of non-scientists, who seek supporting evidence and explain away apparently disconfirming evidence. It is this latter methodological distinction between science and non-science that is Popper's more important contribution.

To maintain that scientific theories are falsifiable is problematic, because, with very few exceptions, scientific theories are not testable or falsifiable by themselves. Observing an increase in demand for some commodity after a rise in its price does not falsify the law of demand if there has been a change in tastes, an even greater increase in the price of a close substitute, a general rise in the price level and

hence a drop in the real price, or some other complicating factor. To say that an hypothesis ‘is rejected if its predictions are contradicted’ is misleading, because hypotheses rarely have predictions of their own. Significant scientific hypotheses imply predictions only when combined with other statements. So, if one insists that scientific claims have to be testable all by themselves, virtually nothing in science counts as science. On the other hand, if one insists only that, like the law of demand, scientific claims must be falsifiable in combination with other claims, then one cannot rule out even the most blatant pseudo-sciences. When Popper criticizes the scientific credentials of Freudian psychology, he does not maintain that, coupled with other statements, it makes no predictions. His criticism is instead that, when those predictions fail, psychoanalysts never cast blame on Freud's theory.

What distinguishes sciences from pseudo-sciences is methodology: when amalgams of theories and various auxiliary hypotheses make false predictions, scientists, unlike practitioners of pseudo-science, are willing to modify or even discard their theories. However, it is difficult to specify exactly how willing scientists should be to surrender their theories. Deciding whether observations give one good reason to reject an hypothesis, like deciding whether observations give one good reason to accept an hypothesis, requires weighing alternative explanations of the data. There is no simple asymmetry between falsification and confirmation.

The significance of falsification is methodological rather than logical or linguistic – a question of the norms that should govern science. The message of falsification is that science treats its findings as subject to criticism and revision. How can one make this platitude concrete? As even Popper and his followers have recognized, some dogmatism may be a good thing. Theories are hard to come by and should not be surrendered too easily. What characterizes successful sciences is on the one hand a mixture of attitudes on the part of individual scientists, with some much more critical than others, and on the other hand an institutional structure in which criticism is not too risky to individuals, and successful criticisms are strongly rewarded.

Those commentators on economic methodology who have been most influenced by Popper have generally been critical of economists. Mark Blaug, for example, argues that economists practise ‘innocuous falsificationism’ (1976, pp. 159–60), paying lip service to the importance of falsification while in fact showing little interest in criticism.

See Also

- theory appraisal

Bibliography

Blaug, M. 1976. Kuhn versus Lakatos *or* Paradigms versus research programmes in the history of economics. In *Method and Appraisal in Economics*, ed. S. Latsis. Cambridge: Cambridge University Press.

Caldwell, B. 1991. Clarifying Popper. *Journal of Economic Literature* 29, 1–33.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago:

University of Chicago Press.

Popper, K. 1957. *The Poverty of Historicism*. New York: Harper and Row.

Popper, K. 1966. *The Open Society and Its Enemies*, vol. 2, 5th edn., Princeton: Princeton University Press.

Popper, K. 1968. *The Logic of Scientific Discovery*, revised edn. London: Hutchinson & Co.

Popper, K. 1969. *Conjectures and Refutations; The Growth of Scientific Knowledge*, 3rd edn. London: Routledge and Kegan Paul.

Popper, K. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.

Popper, K. 1976. *The Unended Quest*. La Salle, IL: Open Court.

How to cite this article

Hausman, Daniel M. "falsificationism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000297> doi:10.1057/9780230226203.0548

The New Palgrave Dictionary of Economics Online

family decision making

Shelly Lundberg and Robert A. Pollak
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The classic unitary model assumes that households maximize a household utility function and implies resource ‘pooling’ – household behaviour does not depend on individuals’ control over resources within the household. Since the 1980s, economists have modified the unitary model in ways that have theoretical, empirical and practical implications. Non-unitary alternatives based on joint decision-making by individual family members with distinct preferences broaden the range of observable behaviour consistent with economic rationality. Many non-unitary models imply that both individuals’ control over resources and ‘environmental factors’ can affect intra-household allocation. Empirical evidence has consistently rejected income pooling and, hence, the unitary model.

Keywords

altruist model of the family; bargaining; collective model of the household; consensus model of the family; cooperative bargaining model of marriage; cooperative games; Cournot–Nash equilibrium; family decision-making; gender specialization; household production and public goods; marriage and divorce; non-cooperative bargaining model of marriage; rotten kid theorem; self-enforcing agreements; separate spheres bargaining model; Slutsky symmetry; two-stage games; unitary and non-unitary models of the household

Article

Economic models of consumer demand and labour supply begin with an individual economic agent choosing actions that maximize his or her utility subject to a budget constraint. How can we reconcile this individualistic theory of the consumer with the reality that people tend to live, eat, work and play in families? Economists have dealt with a possible multiplicity of decision-makers in the family in two ways. The first, in ascendancy until the 1980s, was the unitary approach – treating the family as though it were a single decision-making agent, with a single pooled budget constraint and a single utility function that includes the consumption and leisure time of every family member. The second approach, pioneered in the early 1980s by Manser and Brown and by McElroy and Horney, was to model family behaviour as the solution to a cooperative bargaining game. Other non-unitary approaches have subsequently been developed, including the ‘collective’ model of Chiappori, extensions of the cooperative models of Manser–Brown and McElroy–Horney, and various non-cooperative models.
Most non-unitary models of family behaviour allow two decision makers – the husband and the wife; children are customarily excluded from the set of decision-making agents, though they may be recognized as consumers of goods chosen and provided by loving or dutiful parents. Bargaining models have also been used to analyse interactions between parents and adolescent or young adult children, and between elderly parents and adult children. These interactions may involve family members living in different households, and, in many of these models, who lives with whom is endogenous. As a class, non-unitary models are consistent with a wider range of behaviour than unitary models. The empirical implications of specific non-unitary models of the family depend upon their assumptions about preferences, opportunities, and the form of the game.

Unitary models

Two models provide the theoretical underpinning of the unitary, or common preference, approach to family behaviour: Samuelson's (1956) consensus model and Becker's (1974;

1981) altruist model. The consensus model was introduced by Samuelson to exhibit the conditions under which family behaviour can be rationalized as the outcome of maximizing a single utility function. Consider a two-member family consisting of a husband and a wife. Each partner has an individual utility function that depends on his or her private consumption of goods, but, by consensus, they agree to maximize a social welfare function incorporating their individual utilities, subject to a joint budget constraint that pools the income received by the two spouses. Then we can analyse the household's observed aggregate expenditure pattern as though the family were a single agent maximizing a utility function (that is, the consensus social welfare function). That is, the household maximizes $U(c^h, c^w)$, where c^h and c^w are the private consumptions of husband (*h*) and wife (*w*), subject to the budget constraint $p(c^h + c^w) = y = y^h + y^w$ which pools the individual incomes of husband and wife. This problem generates demand functions $c^i = f^i(p, y)$ that depend only on prices and total family income and that have standard properties provided the utility functions are well behaved. Thus, the comparative statics of traditional consumer demand theory apply directly to family behaviour under the consensus model. Samuelson did not, however, purport to explain how the family achieves a consensus regarding the joint welfare function, or how this consensus is maintained.

Becker's (1974; 1981) altruist model addresses these questions, and also provides an account of how resources are distributed within the family. In Becker's model, the family consists of a group of purely selfish but rational ‘kids’ and one altruistic parent whose utility function reflects his concern for the well-being of other family members. Becker argues that the presence of an altruistic parent who makes positive transfers to each member of the family is sufficient to induce the selfish kids to act in an apparently unselfish way. The altruistic parent will adjust transfers so that each ‘rotten kid’ finds it in his interest to choose actions that maximize family income. The resulting distribution is the one that maximizes the altruist's utility function subject to the family's resource constraint, so the implications of the altruist model for family demands coincide with those of the consensus model (see Bergstrom, 1989 for a discussion of the conditions under which the rotten kid theorem holds and does not hold).

Unitary models provide a simple, powerful mechanism for generating demand functions and establishing their comparative statics for use in applied problems. Since the introduction of the bargaining paradigm however, these models have been criticized on both empirical and theoretical grounds. We first discuss the theoretical criticisms, and then turn to the accumulating empirical evidence inconsistent with the unitary model.

Dissatisfaction with unitary models on theoretical grounds has been the product of serious study of marriage and divorce. Models of marriage and divorce require a theoretical framework in which agents compare their expected utilities inside marriage with their expected utilities outside marriage, but the individual utilities of husband and wife outside marriage cannot be recovered from the social welfare function that generates consumption, labour supply, fertility, and other behaviour within marriage. If the analysis of marriage and divorce is awkward, the analysis of marital decisions in the shadow of divorce is even more so. If unilateral divorce is possible, individual rationality implies that marital decisions cannot leave either husband or wife worse off than they would be outside the marriage. This individual rationality requirement, however, alters the comparative statics of the model, and destroys the correspondence between the behaviour of a single utility maximizing agent and the behaviour of a family.

Non-unitary models

Cooperative bargaining models

A viable alternative to unitary models of the family must recognize, in a non-trivial fashion, the involvement of two or more agents in determining family consumption. Bargaining models from cooperative game theory, first applied to marriage by Manser and Brown (1980) and by McElroy and Horney (1981), satisfy these conditions. A typical cooperative bargaining model of marriage begins with a family that consists of only two members, a husband and a wife. Each has a utility function that depends on his or her consumption of private goods ($U^h(c^h)$ for the husband and $U^w(c^w)$ for the wife). If agreement is not reached, then the payoff received is represented by the ‘threat point’, $(T^h(Z), T^w(Z))$ – the utilities associated with a default outcome of divorce or, in the ‘separate spheres’ model of Lundberg and Pollak (1993), a non-cooperative equilibrium within the marriage. The threat point depends, in turn, upon a set of exogenous distribution factors *Z* that influence individual well-being in the default outcome.

The Nash bargaining model provides the leading solution concept in bargaining models of marriage. Nash bargaining implies that the couple maximizes the Nash product function $N = [U^h(c^h) - T^h(Z)]\,[U^w(c^w) - T^w(Z)]$ subject to a pooled budget constraint, and this results in demand functions of the form $c^i = f^i(p, y, Z)$. Thus demands and individual utilities depend upon the distribution factors *Z*, which may include individual incomes y^h and y^w. This solution can be illustrated by a diagram in utility space (Figure 1), where AB is the utility-possibility frontier. Nash (1950) shows that a set of four axioms, including Pareto efficiency – which ensures that the solution lies on the utility-possibility frontier – uniquely characterize the Nash bargaining solution.

Figure 1

The Nash bargaining solution

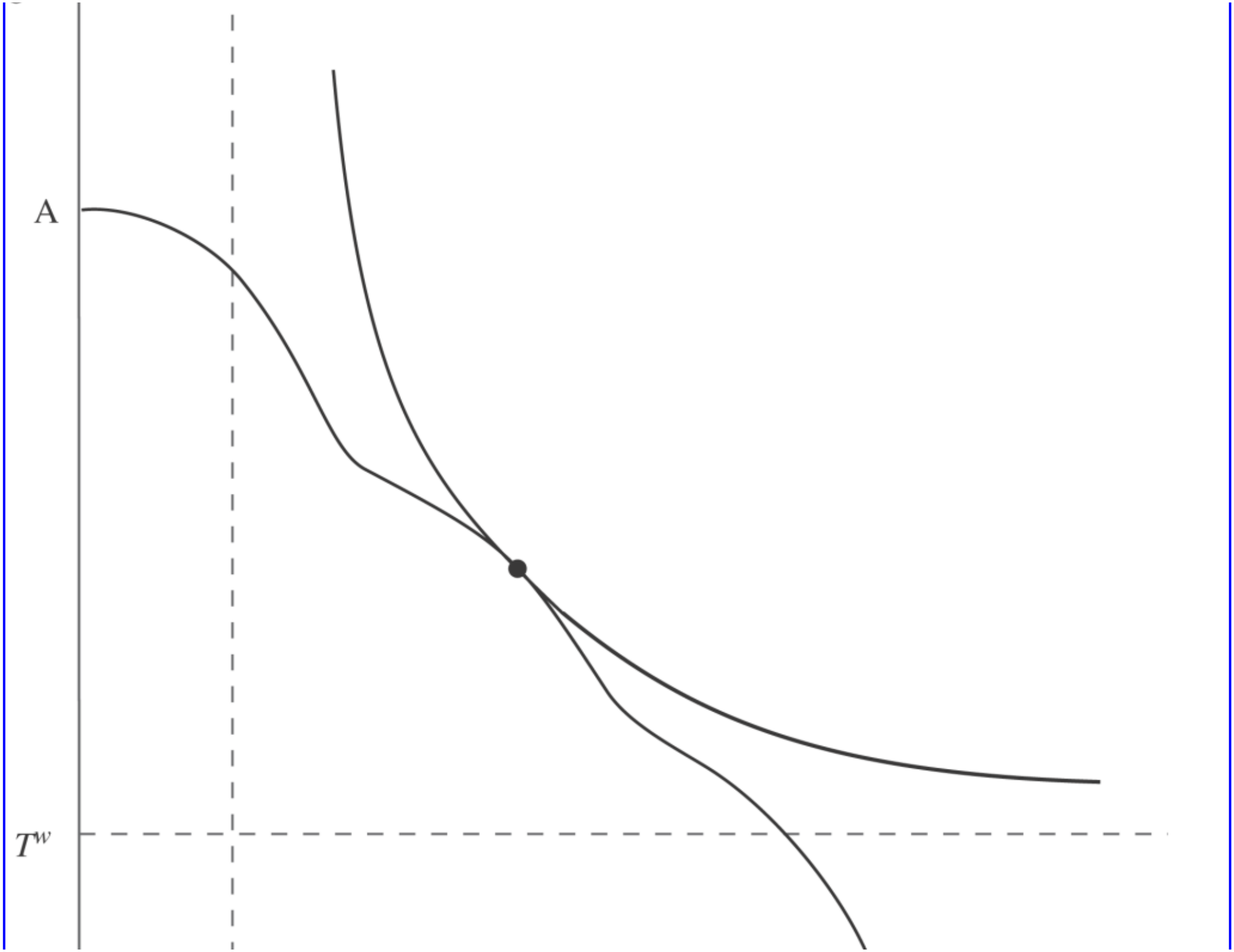
A
T^w

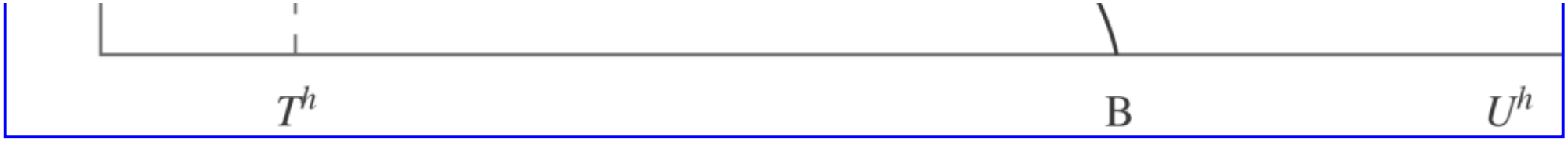

The utility received by husband or wife in the Nash bargaining solution depends upon the threat point: the higher one's utility at the threat point, the higher one's utility in the Nash bargaining solution. This dependence is the critical empirical implication of Nash bargaining models: family demands depend, not only on prices and total family income, but also on determinants of the threat point.

In divorce-threat bargaining models, the threat point is the maximal level of utility attainable outside the marriage. Hence, the threat point depends on wage rates and on the assets each spouse would take if the marriage were to end in divorce. The divorce threat point is also likely to depend on environmental factors (extra-household environmental parameters, or EEPs, in McElroy's, 1990, terminology) that do not directly affect marital utility, such as conditions in the remarriage market and the income available to divorced men and women. The family demands that result from divorce-threat marital bargaining will therefore depend upon these parameters as well.

In the separate spheres bargaining model of Lundberg and Pollak (1993), the threat point is internal to the marriage, not external as in divorce-threat bargaining models. The husband and wife settle their differences by Nash bargaining, but the alternative to agreement is an inefficient non-cooperative equilibrium within marriage. In this non-cooperative equilibrium, each spouse voluntarily provides household public goods, choosing actions that are utility-maximizing, given the actions of their partner. Divorce may be the ultimate threat available to marital partners in disagreement, but a non-cooperative marriage in which the spouses receive some benefits due to joint consumption of public goods may be a more plausible threat in day-to-day marital bargaining.

The introduction of this internal threat point has important implications, because the separate spheres model generates family demands that, under some circumstances, depend not on who receives income after divorce, but on who receives (or controls) income within the marriage. Lundberg and Pollak assume gender specialization in the non-cooperative provision of household public goods, with the husband providing one good out of his own resources, and the wife providing a separate good from her individual resources. This specialization occurs because socially prescribed gender roles provide a focal point for non-cooperative bargaining. The individual reaction functions in this game determine a Cournot–Nash equilibrium in which the public goods contributions may be inefficiently low, and may depend upon the distribution of individual incomes within the family.

As the divorce-threat and separate spheres models show, cooperative bargaining does not necessarily imply income pooling, that is, the property that demands depend only on total household income, rather than its separate components. Bargained outcomes depend upon the threat point, and the income controlled by husband and wife will affect family behaviour (and the relative well-being of men and women within marriage) if this control influences the threat point. This dependence implies that public policy (for example, taxes and transfers) need not be neutral in their effects on distribution within the family. Also, the absence of pooling and the presence of extra-household environmental parameters in family demands yield a model that can be tested against the unitary alternative. For example, changes in the welfare payments available to divorced mothers, or in the laws defining marital property and regulating its division upon divorce, should affect distribution between men and women in two-parent families through their effect on the threat point.

The ‘collective’ approach

Most models of the family either assume or conclude that family behaviour is Pareto efficient. Unitary models ensure Pareto efficiency by assuming a family social welfare function that is increasing in the utilities of all family members: when such a utility function is maximized, no member can be made better off without making another worse off. Cooperative bargaining models characterize the equilibrium distribution by means of a set of axioms, one of which is Pareto efficiency.

Pareto efficiency is the defining property of the ‘collective model’ of Chiappori (1988; 1992). Rather than applying a particular cooperative or non-cooperative bargaining model to the household allocation process, Chiappori assumes only that equilibrium allocations are Pareto efficient. He demonstrates that, given a set of assumptions including weak separability of public goods and the private consumption of each family member, Pareto efficiency implies, and is implied by, the existence of a ‘sharing rule’. Under a sharing rule, the family acts as though decisions were made in two stages: first total family income is divided between public goods and the private expenditures of each individual, and then each individual allocates his or her share among private goods. The collective model implies a set of testable restrictions on the response of household demands to ‘distribution factors’ that affect the household's sharing rule.

Non-cooperative bargaining models

The use of models that assume Pareto efficiency of outcomes relies on the judgement that information within families is relatively good (or at least not asymmetric) and that members are able to make binding, costlessly enforceable agreements. Since legal institutions do not provide for external enforcement of contracts regarding consumption, labour supply, and allocation within marriage, however, the binding-agreement assumption is unappealing.

Non-cooperative game theory focuses on self-enforcing agreements. It is possible for non-cooperative bargaining to yield Pareto efficient outcomes under certain conditions. For example, repeated non-cooperative games have multiple equilibria which are sustained by credible threats of punishment, and some of these equilibria are Pareto efficient. One of the benefits of modelling distribution within marriage as a non-cooperative game is the opportunity to treat efficiency as endogenous, potentially dependent upon the institutions and social context of marriage in a particular society and upon the characteristics of the marital partners.

The prevalence of destructive or wasteful phenomena such as domestic violence and child abuse, as well as the demand for marriage counselling and family therapy, suggests that we consider the possibility that family behaviour is sometimes inefficient. Other researchers have pointed to gender segmentation in the management of businesses or agricultural plots in many countries as evidence of an essentially non-cooperative, and possibly inefficient, family environment. One piece of evidence is provided by Udry (1996), who finds that in Burkina Faso the marginal product of land controlled by women is below the marginal product of land controlled by men and concludes that the household allocation of inputs to male- and female-controlled agricultural plots is inefficient.

Intertemporal models

In dynamic bargaining models, decisions made in one period can alter the relative bargaining power of individual family members in future periods. If family members cannot agree on rules for sharing household resources in the future, and make credible promises to obey such rules, then inefficiencies of the standard ‘hold-up’ variety will result. Lundberg and Pollak (2003) model the two-earner couple location problem as a two-stage game in which a couple must decide where to live and whether to stay together without being able to make binding commitments about allocation in the new location. Lundberg and Pollak show that the equilibrium of this two-stage game need not be efficient even if the second-stage game is conditionally efficient (that is, efficient given the location determined at the first stage).

Even if prospective spouses can make binding agreements in the marriage market, they cannot make agreements with potential spouses they have not yet met. Konrad and Lommerud (2000) show that individuals will over-invest in education prior to marriage to increase their marital bargaining power, even if they expect to bargain cooperatively once they find and marry a spouse. Models of limited commitment in marriage can also be applied to decisions about childbearing, career choice and work effort.

Empirical evidence

Recent empirical evidence suggests that the restrictions imposed on demand functions by unitary models are not well supported. Rejections of the family income pooling hypothesis, in particular, have been most influential in weakening economists’ attachment to unitary models. Unitary models imply that the fraction of income received or controlled by one family member should not influence demands, given total family income. A large number of recent empirical studies have rejected pooling, finding that earned and unearned income received by the husband or wife significantly affect demand patterns when total income or expenditure is held constant. Some studies find that children appear to do better when their mothers control a larger fraction of family resources (Thomas, 1990; Haddad and Hoddinott, 1994). These results are inconsistent with the unitary framework, but consistent with both bargaining models (provided individual incomes affect the threat point) and with the collective model (provided individual incomes are included among the ‘distribution factors’ that influence the household's sharing rule).

The collective model imposes, in addition, a proportionality restriction on the influence of distribution factors on demands. The ratio of the marginal propensities to consume any two goods must be the same for all sources of income, for example, because individual incomes affect consumption only through the sharing rule. A generalization of Slutsky symmetry in price effects can also be derived (Browning and Chiappori, 1998). A series of empirical tests have found that consumption expenditures in households reject the unitary framework but are generally consistent with the collective model (for example, Bourguignon et al., 1993; Browning and Chiappori, 1998).

Tests of the unitary model against non-unitary alternatives require a measure of husband's and wife's relative control over resources. Relative earnings would seem to be an attractive candidate for this measure, since labour income is by far the largest component of family income, and earnings data are readily available and reliably measured. Also, the earnings of wives relative to husbands have increased dramatically in the United States and many other countries, and we would like to assess the distributional consequences, if any, of this change. The difficulty with this approach is that earnings are clearly endogenous with respect to household time-allocation decisions. Earnings are the product of hours worked, a choice variable, and hourly wage rates, which measure the prices of time for husband and wife and therefore enter demand functions directly in the unitary model. This implies that households with different ratios of wife's earnings to husband's earnings are likely to face different prices and may have different preferences.

One might try to avoid these problems by testing the pooling of unearned income rather than earnings. Unearned income is not contaminated by price effects, but most unearned income sources are not entirely exogenous with respect to past or present household behaviour. Schultz (1990), who like Thomas (1990) uses unearned income to test the pooling hypothesis, points out that variations in unearned income over a cross-section are likely to be correlated with other (possibly unobservable) determinants of consumption. For example, property income reflects, to a considerable extent, accumulated savings and is therefore correlated with past labour supply and, if those who worked a lot in the past continue to do so, with current labour supply. Public and private transfers may be responsive to household distress due to unemployment or bad health, and may be related to

expenditures through the events that prompted them. *Unexpected* transfers such as lottery winnings, unexpected gifts or unexpected bequests will affect resources controlled by individuals without affecting prices, but are likely to be sporadic and unimportant for most families.

Other standard empirical proxies for the relative bargaining power of husbands and wives (or, in the terminology of the collective model, distribution factors) include the relative ages, educations, or measures of family background of husband and wife. The interpretation of these factors, however, is contaminated by assortative mating on unobserved characteristics. It would be unwise to assume that a highly educated woman married to a man with less education has relatively more control over the allocation of household resources without controlling for other personal characteristics that affected the decision of this couple to marry in the first place. The same critique applies to measures of relative assets brought to the marriage by the husband and wife, even when they maintain separate ownership of these assets during marriage and divorce.

The ideal test of the pooling hypothesis, and therefore of the unitary family model, would be based on an experiment in which some husbands and some wives were randomly selected to receive income transfers. A less-than-ideal test could be based on a ‘natural experiment’ in which some family members receive an exogenous income change, and one can study a constant population of families before and after the change. Several studies exploiting such policy changes have found evidence against income pooling, and have also supported the hypothesis that women have a greater propensity, on average, to spend on children's goods. Lundberg, Pollak and Wales (1997) examine the effects of a policy change in the United Kingdom that transferred a substantial child allowance from husbands to wives in the late 1970s. They find strong evidence that a shift towards relatively greater expenditures on women's goods and children's goods coincided with this income redistribution, and interpret this as a rejection of the pooling hypothesis. Duflo (2000) studied the effect of an extension of the South African Old Age Pension on children's health and nutrition, and found that payments to grandmothers had a substantial effect on these outcomes, especially for girls, while payments to grandfathers had no effect. These results both reject a unitary framework for multi-generation families, and support the hypothesis that children benefit from female control of household resources. Tests of pooling using PROGRESA, a public cash transfer programme in Mexico directed at women, have been more complicated. A random assignment social experiment, PROGRESA had a substantial income effect and benefits were conditional on child school enrolment. Attanasio and Lechene (2002) reject household pooling using PROGRESA data, and Rubalcava, Teruel, and Thomas (2004) find that these transfers to women were more likely to be spent on child goods, improved nutrition, and investments in small livestock than other household income.

One important implication of non-unitary models of the household is that government programmes targeted to particular individuals within households may affect the intra-household allocation. Even if, as rejections of the unitary model suggest, targeted transfers are effective in the short run, we cannot conclude that targeted transfers will be effective in the long run. Lundberg and Pollak (1993) show that the long-term effects of such policy changes on intra-household allocation may be very different from the short-term effects, as adjustments occur in the marriage market of subsequent cohorts. If prospective couples can make binding agreements when they marry, then the distributional effects of policy can be offset by subsequent generations of families. Even if such marital agreements are not possible, changes in the expected gains to marriage will affect who marries whom and who marries at all, and this will also affect the long-run distributional effects of policy. Cross-sectional studies of intra-household allocation that use state variation in policy or laws (such as divorce laws or property settlement rules) will be estimating the equilibrium effects of long-standing differences in policy, including any marital sorting effects.

Conclusion

The classic unitary model assumes that households maximize a household utility function subject to household resource and technology constraints. Unitary models imply income or resource ‘pooling’ – household behaviour does not depend on individual control over resources within the household. Since the 1980s, economists have modified the unitary model in ways that have theoretical, empirical and practical implications. Non-unitary alternatives based on joint decision-making by individual family members with distinct preferences broaden the range of observed behaviour consistent with economic rationality. Non-unitary models also permit the analysis of marriage and divorce within the same framework as household demands and the labour supply of household members. Unlike unitary models, many non-unitary models imply that both individual control over resources and ‘environmental factors’, such as divorce laws, that affect the well-being of individuals outside the household can affect intrahousehold allocation. Empirical evidence has consistently rejected income pooling and, hence, the unitary model.

See Also

- collective models of the household
- family economics
- intrahousehold welfare
- marriage and divorce

Bibliography

Attanasio, O. and Lechene, V. 2002. Tests of income pooling in household decisions. *Review of Economic Dynamics* 5, 720–48.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82, 1063–93.

Becker, G.S. 1981. Altruism in the family and selfishness in the market place. *Economica* 48, 1–15.

Becker, G.S. 1991. *A Treatise on the Family*. Enlarged edn. Cambridge, MA: Harvard University Press.

Bergstrom, T. 1989. A fresh look at the rotten kid theorem and other household mysteries. *Journal of Political Economy* 97, 1138–59.

Bourguignon, F., Browning, M., Chiappori, P.-A. and Lechene, V. 1993. Intra household allocation of consumption: a model and some evidence from French data. *Annales d'Economie et de Statistique* 29, 138–56.

Browning, M. and Chiappori, P.-A. 1998. Efficient intra-household allocations: a general characterization and empirical tests. *Econometrica* 66, 1241–78.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56, 63–89.

Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100, 437–67.

Duflo, E. 2000. Child health and household resources in South Africa: evidence from the old age pension program. *American Economic Review* 90, 393–8.

Haddad, L. and Hoddinott, J. 1994. Women's income and boy–girl anthropometric status in the Côte d'Ivoire. *World Development* 22, 543–53.

Konrad, K.A. and Lommerud, K.E. 2000. The bargaining family revisited. *Canadian Journal of Economics* 33, 471–87.

Lundberg, S.J. and Pollak, R.A. 1993. Separate spheres bargaining and the marriage market. *Journal of Political Economy* 101, 988–1010.

Lundberg, S.J. and Pollak, R.A. 2003. Efficiency in marriage. *Review of Economics of the Household* 1, 153–67.

Lundberg, S.J., Pollak, R.A. and Wales, T.J. 1997. Do husbands and wives pool their resources? Evidence from the United Kingdom child benefit. *Journal of Human Resources* 32, 463–80.

Manser, M. and Brown, M. 1980. Marriage and household decision-making: a bargaining analysis. *International Economic Review* 21, 31–44.

McElroy, M.B. 1990. The empirical content of Nash-bargained household behavior. *Journal of Human Resources* 25, 559–83.

McElroy, M.B. and Horney, M.J. 1981. Nash-bargained household decisions: toward a generalization of the theory of demand. *International Economic Review* 22, 333–49.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Rubalcava, L., Teruel, G. and Thomas, D. 2004. Spending, saving, and public transfers paid to women. California Center for Population Research. Online Working Paper Series. CCPR-024–04. Online. Available at http://www.ccpr.ucla.edu/ccprwpseries/ccpr_024_04.pdf, accessed 22 April 2007.

Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70, 1–22.

Schultz, T.P. 1990. Testing the neoclassical model of family labor supply and fertility. *Journal of Human Resources* 25, 599–634.

Thomas, D. 1990. Intra-household resource allocation: an inferential approach. *Journal of Human Resources* 25, 635–64.

Udry, C. 1996. Gender, agricultural production and the theory of the household. *Journal of Political Economy* 104, 1010–46.

How to cite this article

Lundberg, Shelly and Robert A. Pollak. "family decision making." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000290> doi:10.1057/9780230226203.0549

The New Palgrave Dictionary of Economics Online

family economics

John Ermisch
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Family economics is the application of the analytical methods of microeconomics to family behaviour. It aims to improve our understanding of resource allocation and the distribution of welfare within the family, investment in children and inter-generational transfers, family formation and dissolution and how families and markets interact. In family economics, non-market interactions are crucial for family behaviour and individual welfare.

Keywords

altruism; Becker, G.; child nutrition and mortality; collective models of the household; demographic transition; family decision-making; family economics; family planning; fertility in developed countries; fertility in developing countries; human capital; intergenerational income mobility; intergenerational transfers; intrahousehold welfare; labour supply; Malthus, T.; marriage and divorce; Rotten Kid Theorem; shadow pricing; women's work and wages

Article

Family economics is the application of the analytical methods of microeconomics to family behaviour. It aims to improve our understanding of resource allocation and the distribution of welfare within the family, investment in children and inter-generational transfers, family formation and dissolution, and how families and markets interact. Family economics lifts the lid on the ‘black box’ of the family, within which non-market interactions are crucial for family behaviour and individual welfare. It analyses how markets affect family behaviour and on how family context affects market behaviour, such as labour supply and consumer demand, thereby linking family economics with traditional fields of economics. Medical and social sciences indicate the importance of nutritional, cognitive and emotional development during childhood for a person's lifetime health and prosperity, and these developments are a product of parents’ actions, including family break-up. Acquiring a better understanding of family formation and

dissolution and of decisions within the family, particularly as these are affected by elements of people's budget constraints, is an important prerequisite for understanding how public policy can influence the family.

In a broad sense, family economics has been around for over two hundred years. Thomas Malthus believed that human fertility was determined by the age at marriage and frequency of coition during marriage. He contended that an increase in people's income would encourage them to marry earlier and have sexual intercourse more often. Modern economic theories of fertility generalized the Malthusian theory (starting with Becker, 1960), and Gary Becker subsequently developed a broader economic analysis of the family (Becker, 1981), which forms the foundation for today's family economics (Ermisch, 2003).

What influences family decisions?

Individualism needs to be the foundation of family economics if we are to analyse the impacts of public policies and technological developments on the welfare of *individuals*. In particular, decisions about marriage and divorce must make comparisons between *individual* welfare within and outside a couple. The family is best viewed as a ‘governance structure’ for organizing its activities rather than as a preference ordering augmented by home production technology, or as a set of long-term contracts (Pollak, 1985). This suggests that bargaining models, in which alternatives and ‘threat points’ affect intra-family allocation and distribution, provide a useful framework for analysing family behaviour. A bargaining approach naturally focuses on the structure of family membership and its internal organization (for example, comparing an intact nuclear family with divorced parents), and allows decisions to evolve in a flexible way.

A fruitful starting point is to assume that all individuals act to maximize their welfare as they evaluate it, given the predicted behaviour of others in the family. Some authors have adopted this non-cooperative approach to studying family choices (for example, Konrad and Lommerud, 1995). But, in many circumstances (for example, the co-resident family), cooperative behaviour is a better representation of family behaviour because of repeated interaction between family members, which facilitates information flows and monitoring. Nevertheless, family members must obtain welfare from cooperation that is at least as great as they would achieve from a non-cooperative outcome, although in some circumstances divorce may be a credible threat affecting decisions within the family (Bergstrom, 1996).

Cooperation achieves an efficient allocation of resources within the family. Individual welfare depends, in general, on individual incomes and prices and possibly other ‘distribution factors’ like marriage market conditions, divorce laws and other institutions (Browning and Chiappori, 1998), whose influence reflects bargaining between family members. For example, an increase in the mother's income may have two effects on family choices. It increases family resources, expanding welfare-enhancing opportunities for all family members. It also may increase her threat point (bargaining power), which pushes family choices in her favour, thereby increasing her welfare relative to the father's. Put differently, her income affects the position of the family's utility possibility frontier and also the position on it. If, for example, mothers’ preferences put more weight on children than fathers’ preferences do, then an increase in her share of family income would increase expenditure on children. If this is the case, then children do better when mothers control more of family resources, developments which improve women's earning

opportunities affect the distribution of welfare within families, and it is possible to target policies on individuals within families. In a dynamic setting, in which current decisions affect future bargaining power, efficiency is harder to sustain because of the difficulty of making binding commitments, but individual incomes and distribution factors still affect intra-family allocation.

One particular decision-making rule has been an important part of family economics: the family maximizes the welfare of an ‘effective altruist’. It is in fact a special case of the cooperative (efficient outcomes) framework just discussed. A person is said to be *altruistic* toward someone if his or her welfare depends on the welfare of that person. Altruism is usually defined more narrowly, by what have been called ‘caring’ preferences: the altruist's ‘social’ utility takes the form $W^A[U^A(\boldsymbol{x}_A,\boldsymbol{G}), U^B(\boldsymbol{x}_B,\boldsymbol{G})]$, where $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are vectors of private goods consumed by persons A and B respectively, $\boldsymbol{G}$ is a vector of public goods and $U^A(\cdot)$ and $U^B(\cdot)$ are ‘private’ utility indices for each person. The altruist A does not care how (in terms of $\boldsymbol{x_B}$ and $\boldsymbol{G}$) a given level of private utility is obtained by his/her beneficiary B. Caring preferences limit only the relevant range of the utility possibility frontier expressed in terms of *private* preferences. Some family decision rule is still needed to determine the point on the frontier that is chosen, but there are circumstances in which caring preferences can produce distinctive predictions. Suppose that a wife and her husband care for each other, and her share of joint income is sufficiently large that she is making transfers to him to ensure that his welfare is not too low. To use Becker's (1981) term, she is an *effective altruist*. Only joint income matters for family decisions in these circumstances. Thus, effective altruism provides partial insurance for family members and insulates the family from targeted changes in taxes and benefits. Becker's (1981) claim that effective altruism also provides incentives for the beneficiary to act in the best interests of the family and reduces intra-family conflict – the so-called Rotten Kid Theorem – is, however, valid only with very restrictive preferences (Bergstrom, 1989).

If, however, the couple's incomes are relatively similar, then neither spouse is rich enough relative to the other to make transfers to the other, and individual incomes are likely to affect family decisions. In either case, non-market interactions between family members are important for determining individual welfare, through either bargaining or intra-family transfers motivated by altruism, and these also affect market behaviour like consumer demand and labour supply.

Fertility, investments in children and security in old age

The primary reason that most men and women enter a long-term relationship is to bear and raise children. In addition to the number of children, parents’ welfare is likely to depend on the lifetime well-being of each child – ‘child quality’ for short. That is, parents receive more satisfaction from having children who are better off throughout their life, and they make monetary transfers and human capital investments to influence their children's lifetime standard of living.

If parents view child quantity and quality as substitutes and treat all their children equally, their budget constraint contains the product between the number of children and quality per child (Willis, 1973). This implies that the ‘shadow price’ of an additional child is proportional to the level of child quality, and the shadow price of raising child quality is proportional to the number of children. As a consequence, there is an important interaction between family size and child quality. For example, a higher return to human

capital increases investment per child. This raises the shadow price of children, which lowers family size and the price of child quality, thereby raising child quality further, and so on. Thus, increases in the returns to human capital investment associated with technical change may lead to simultaneous *large* reductions in fertility and increases in human capital investment in children. This is consistent with important stylized facts of economic development and links family economics with the study of economic growth and development (Rosenzweig, 1990).

The quantity–quality interaction may also produce a ‘high fertility–low child investment trap’, in which low quality produces a low price of children, high fertility and a high shadow price of child quality. Higher parents’ income increases fertility and the price of child quality, keeping child investment low. It may take some policy or technological development that alters the prices of quantity or quality independently, such as a family planning intervention or a large change in the return to human capital investment, to ‘spring the trap’. Once sprung, if the quality income elasticity exceeds the one for quantity, then the ratio of quality to the number of children rises with higher income, thereby increasing the shadow price of an additional child relative to the shadow price of child quality. The substitution effect induced by this increase may be sufficiently large to produce a decline in fertility when income increases, even though children are normal goods.

The ultimate manifestation of low child quality is a child not surviving to adulthood. Scientific advance or a policy intervention, such as better water supply or public health, increases the probability of child survival, but it has conflicting impacts on fertility. On the one hand, it reduces the price of a surviving birth, thereby encouraging higher fertility. But if parents can influence the chances that their own children survive to become adults by spending more on each child, then it is *possible* that exogenous improvements in child survival reduce fertility, provided that improvements in child survival substitute for parents’ expenditure on child health (Cigno, 1998). Such a relationship may help account for the ‘demographic transition’ – the change from a high fertility-high ‘child mortality environment to a low fertility-low mortality one.

The factors affecting the cost of children (including investment in them) are closely associated with the key role of parental time in the rearing of and investment in children. The rearing of children is usually presumed to be *time-intensive* relative to other home production activities, and mothers provide a disproportionate share of parental time in the production of child quality. Thus, the cost of children relative to the cost of the parents’ living standard is directly related to the mother's cost of time (Willis, 1973). This links the cost of children with women's educational and earning opportunities, with implications for their effects on fertility, women's labour supply and investment in children.

Fertility and child investment may also be motivated by the need for support in old age. If people do not have access to a capital market, an extended family network including three generations at different stages of life could substitute for a capital market (Cigno, 2000). In effect, it arranges ‘loans’ to its young members from its middle-aged ones and enforces repayment later when the young borrowers have become middle-aged and the middle-aged lenders have become old. People may have children only because they are needed to transfer resources through time. The opening of a capital market with a sufficiently high interest rate offers the middle-aged an alternative to this family transfer system. A threat of no support from the family in old age is no longer a deterrent, because they can make their own provision for old age through the market. In broad terms, this prediction is consistent with the observation that the growth of the financial sector, or the introduction of a state pension system, tends to coincide with a sharp decline in private transfers from the middle-aged to their elderly parents and a fall

in fertility. The fact that childbearing does not cease suggests that the demand for children is not entirely derived from the need for transfers from them to finance consumption in old age. Again we see the important role of institutions in shaping family behaviour.

In countries with well-developed capital markets and pension systems, it is often observed that financial transfers from adult children to parents are rare, but transfers in the other direction are more common. Also, children are often observed providing ‘services’ to their parents that do not have clear market substitutes, such as companionship, attention and adapting their behaviour to their parents’ wishes. Such services come at a cost to the children, and so transfers from parents to their child may be an exchange for these services (Cox, 1987). Parents may also want to help their children financially when they need it, but they want them to behave responsibly in the sense of expending sufficient effort to support themselves. How transfers from parents respond to an adult child's income depends on the balance of altruistic motives, parents’ intention to provide an incentive for high effort and the effects of parent–child bargaining on the provision of child-services.

Marriage and divorce

In addition to love and companionship, marriage offers two people the opportunity to share household public goods and benefit from the division of labour, and it facilitates risk sharing. Whom a person marries influences family behaviour (for example, fertility) and individual welfare through family resources, bargaining and costs (for example, of children). But the process of finding a spouse is one in which information is scarce, and it takes time to gather it. These market frictions affect who marries whom, the gains from each marriage and the distribution of gains between spouses (Burdett and Coles, 1999). The positive correlation between spouses in desirable attributes like education is expected to be weaker when frictions are larger. The chances of divorce, and therefore divorce laws, also affect matching in the marriage market. A higher divorce rate makes people less choosy when selecting a spouse, because it reduces the perceived benefits from waiting for a better match by making it more likely that a person will return to the single state. Poorer matches ensue, and these have a higher probability of dissolving.

Marriage market frictions may also be responsible for childbearing outside marriage. A woman who has a relationship with a man she does not wish to marry, or who will not marry her, would choose to have a child by the man if the short-run gain exceeds the long-term costs in terms of damage to her marriage prospects (Ermisch, 2003, ch. 7). Those women who expect to obtain a significant increase in welfare when they marry suffer a greater long-term cost by having a child while single than women whose marriage prospects are such that they expect to gain little from marriage. Thus, women with poorer marriage prospects are more likely to have children outside marriage.

Parents are likely to continue to care about the welfare of their children after they divorce, and so expenditure on children, such as investment in their human capital, is a public good to the parents. When living together, they choose the efficient level of this public good. But after breaking up, the mother usually obtains custody of the children and she decides the level of expenditure on children (Weiss and Willis, 1985). The father can influence it only by making transfers to the mother, and he must transfer more than a dollar to obtain a dollar's more expenditure on children, because the mother spends part of the transfer on herself. The higher effective price for child expenditure when divorced encourages him to

spend less on children after divorce (perhaps nothing), resulting in a lower, inefficient level of expenditure on children overall. This is likely to have implications for the lifetime welfare of children. The probability that a couple divorces is inversely related to this efficiency loss from divorce. Behaviour within marriage is likely to be affected by exogenous variation in the probability of divorce (for example, through legal changes). If, for example, more participation in paid employment raises future wages, the risk of divorce can encourage more paid employment by the mother during marriage and, by raising the cost of child quality, lower expenditure on children and lower fertility. These ‘defensive investments’ are undertaken to increase welfare later, when outcomes are uncertain because of the possibility of divorce. Thus, the probability of divorce affects women's wages and labour supply in the economy.

Examples have illustrated the distinctive aspects of family economics: how market prices and personal incomes affect non-market interactions between individuals in the family (through altruistic motives and bargaining), fertility and investment in children. These channels link family economics to traditional fields like growth and development, labour economics, consumer demand, savings and inter-generational transfers.

See Also

- child health and mortality
- collective models of the household
- family decision making
- fertility in developed countries
- fertility in developing countries
- marriage and divorce
- rotten kid theorem

Bibliography

Becker, G. 1960. An economic analysis of fertility. In *Demographic and Economic Change in Developed Countries*, ed. National Bureau of Economic Research. Princeton, NJ: Princeton University Press.

Becker, G. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Bergstrom, T. 1989. A fresh look at the Rotten Kid Theorem – and other household mysteries. *Journal of Political Economy* 97, 1138–59.

Bergstrom, T. 1996. Economics in a family way. *Journal of Economic Literature* 34, 1903–34.

Browning, M. and Chiappori, P.-A. 1998. Efficient intra-household allocations: a general characterization and empirical tests. *Econometrica* 66, 1241–78.

Burdett, K. and Coles, M. 1999. Long-term partnership formation: marriage and employment. *Economic Journal* 109, F307–F334.

Cigno, A. 1998. Fertility decisions when infant survival is endogenous. *Journal of Population Economics* 11, 21–8.

Cigno, A. 2000. Self-enforcing family constitutions. In *Sharing the Wealth: Intergenerational Economic Relations and Demographic Change*, ed. A. Mason and G. Tapinos. Oxford: Oxford University Press.

Cox, D. 1987. Motives for private income transfers. *Journal of Political Economy* 95, 508–46.

Ermisch, J. 2003. *An Economic Analysis of the Family*. Princeton, NJ: Princeton University Press.

Konrad, K. and Lommerud, K. 1995. Family policy with non-cooperative families. *Scandinavian Journal of Economics* 97, 581–601.

Pollak, R. 1985. A transaction cost approach to families and households. *Journal of Economic Literature* 23, 581–608.

Rosenzweig, M. 1990. Population growth and human capital investments: theory and evidence. *Journal of Political Economy* 98, S38–S70.

Weiss, Y. and Willis, R. 1985. Children as collective goods and divorce settlements. *Journal of Labor Economics* 3, 268–92.

Willis, R. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81, S14–S64.

How to cite this article

Ermisch, John. "family economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000292> doi:10.1057/9780230226203.0550

The New Palgrave Dictionary of Economics Online

famines

Cormac Ó Gráda

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Today the ultimate Malthusian check of ‘gigantic, inevitable famine’ is confined to the very poorest pockets of the globe. Economic development, medical technology and the globalization of disaster relief have reduced the size and duration of famines in the recent past. On the other hand, totalitarianism and the enhanced role of human agency produced in the 20th century some of the biggest famines ever. Topics discussed include the demography and long-run impact of famine, the role of public and private action in relieving those at risk, and how markets function during famines.

Keywords

agency problems; charitable donations; corruption; democracy; demographic transition; epidemiological transition; famines; fertility; food availability declines; food markets; global warming; health; hoarding; life expectancy; Malthus, T.R.; migration; mortality; non-governmental organizations; nutrition; poverty; public works; Sen, A.K.

Article

‘Famine’ is defined narrowly here as a food shortage leading directly to excess mortality from starvation or hunger-induced illnesses (compare Howe and Devereux, 2004). By this definition, the 20th century presents a paradox in the history of famines. On the one hand, it witnessed in China in 1959–61 the greatest famine in world history. On the other, it saw the virtual elimination of famine across most of the globe. Economic growth in the 19th century led to the disappearance of famine in Europe in peacetime and, after the 1870s, a reduction in famine intensity throughout Asia.

Today's high-profile famines are, relatively speaking, small and confined to poverty-stricken and often war-torn corners of Africa. In principle, famine prevention should be ‘easy’. Better communications, better understanding of nutritional requirements and medical remedies, and the globalization of disaster relief mean that the risks faced by the world's most underdeveloped economies should be far fewer than

those faced by equally poor countries in the past.

‘Malthusian’ famines

Famines and economic backwardness are closely related. Malthus would not have been surprised to hear of famine in Niger, probably the world's poorest economy, in 2005, or that the cross-sectional correlation between excess mortality and poverty was strong within Ireland in the 1840s and Bengal in the 1940s. And he would have deemed the extreme backwardness of the Chinese economy in the mid-1950s a contributory factor to the Great Leap Forward famine of 1959–61: Chinese real GDP per head then was less than half the African average in 2006 (Maddison 2006).

Most famine victims succumb to infectious diseases rather than to actual starvation. Poverty prevents proper medical care because the associated remedies are costly and difficult to implement in crisis conditions. Sub-Saharan Africa has yet to complete the ‘epidemiological transition’, mainly because the resources and the political capabilities to put what is available locally or obtainable from abroad to most effective use are lacking. Famines are the exception where the transition has been completed, but when they occur, as in Nazi-occupied Leningrad, Greece, and the western Netherlands during the Second World War, the diseases mainly responsible for excess mortality were very different. In these relatively developed societies, public health structures that prevented the spread of infectious disease had become part of daily routine, and continued to be so during the war (Mokyr and Ó Gráda, 2002; Maharatna, 1996, pp. 159–61; Hionidou, 2006).

Most famines strike in the wake of major crop failures, although crop failure is neither a necessary nor a sufficient condition for famine. Even the most backward economies often have the resilience to cope with once-off harvest shortfalls, so that in the past the worst famines have been the product of back-to-back shortfalls of the staple crop. Thus, the probability of back-to-back poor harvests should provide some sense of the likelihood of famine in the past. Agricultural and meteorological data imply that such back-to-back events were uncommon (Ó Gráda, 2007).

Entitlements and governance

Civil unrest and bad government can also lead to famine by limiting production and trade or failing to prevent the spread of epidemic disease. The impact of war on the supply of shipping and grain imports from abroad was an important contributory factor to famine in Bengal in 1943–44. Panics about the food supply and poorly performing food markets may exacerbate famine. In such instances factors other than crop shortfalls reduce the purchasing power or ‘entitlements’ of vulnerable sections of the population: the size of the loaf matters less than its distribution. Claims that even during famines there is adequate food for everyone are not new. Such claims, which invert the relative importance of food supply on the one hand, and human action and distribution on the other, had a particular resonance for the 20th century. On several occasions between the 1930s and the 1950s, not only did totalitarian regimes engage in policies that placed millions at risk, but they also managed to keep the consequences largely hidden from the outside world. Analyses of 20th-century famines accordingly have tended to dwell less on economic factors such as the background level of development and the extent of the crop shortfall than on the role of human agency – be it the ruthlessness of dictators or the incompetence of officialdom. Yet closer

inspection suggests that even the most notorious ‘man-made’ famines of the 20th century in the Soviet Union in 1932–33, in China in 1959–61, and even in Bengal in 1943–44, entailed what Amartya Sen (1981) has dubbed ‘food availability declines’ (FADs) (Davies and Wheatcroft, 2004: ch. 5; Tauger, 2006; Ó Gráda, 2007). The paucity of evidence for ‘pure’ entitlement famines – famines where there was no food availability decline – suggests that modern scholarship may underestimate the role of food supply in the relatively recent past.

Sen's claim that famine and democracy are incompatible (Sen, 2001) is a special case of the more general claim that democratic institutions promote economic justice and reduce inequality. Exceptions to this rule seem few: Banik's analysis of press reports of starvation deaths in Orissa in the 1990s confirms it in so far as famines are concerned, but highlights the inability of a free press and collective action to prevent mass malnutrition and ‘many, many deaths’ (Banik, 2002). It also bears noting that in poverty-stricken, ethnically divided, low-literacy economies democracy may not be sustainable. Nonetheless, the exogenous element in democratic institutions surely matters.

Markets and famines

Economists have long argued that, since crop failures are subject to spatial variation and rarely occur two years in succession, spatial and intertemporal arbitrage in food markets should help mitigate the cost of famines (Persson, 1999). However, natural obstacles (poor communications) and artificial obstacles (war, civil unrest, trade restrictions and price controls) have often impeded the scope for arbitrage. Research on Bengal in 1942–44 and Bangladesh in 1974–75 claims that food markets worked poorly in these instances, in the double sense of inadequate inter-regional trade and ‘excessive’ hoarding on the part of producers and traders (Sen, 1981; Ravallion, 1987, pp. 19, 111–13; 1997, pp. 1219–21). Formal studies of market performance during pre-20th century famines are few, although evidence from pre-industrial Europe suggests that they functioned no worse than in normal times (Ó Gráda, 2005). The asymmetry in speculators’ expectations implied by the findings of Sen and Ravallion – over-pessimism in the event of a harvest shortfall – is absent in the earlier data. That does not mean that markets worked like clockwork in pre-industrial Europe, but merely that their responses to spatial and intertemporal disequilibria were no weaker than in non-crisis times. In practice, markets may adjust too slowly to prevent famine: in the mid-19th century, for example, before the telegraph and long-distance bulk carriage by steamship could have made the difference, global grain markets could not have prevented mass mortality in Ireland and India. Nor does this mean that well-functioning, integrated markets always benefit the poor: as Sen emphasizes, they might allow inhabitants of less affected areas, endowed with the requisite purchasing power, to attract food away from famine-threatened areas. Much depends on whether such exports are used to finance cheaper imported substitutes, and on the speed with which food markets adjust. Dogmatic generalizations are not warranted.

Free markets can mitigate the impact of famines in two other respects. First, migration arguably limits the damage wrought by poor harvests, since the migrants reduce the pressure on scarce food and medical resources where the crisis is deepest. This is probably true even when the poorest lack the resources to migrate. Although migration undoubtedly exacts a cost in terms of the spread of infectious disease in host countries, on balance it saves lives.

Second, regional specialization increases aggregate output, with a resultant reduction in the risks

attendant on any proportionate harvest shortfall. Increasing commercialization also makes for more effective arbitrage in food markets. For example, the implied reduction in the cost of holding carry-over stocks and of transport greatly reduced the vulnerability of the Italian and the English poor in the early modern era (Persson, 1999; Ó Gráda, 2007).

Public and private action

Throughout history, whether out of fear or compassion, ruling elites have accepted a degree of responsibility for those at risk during famines. Most analytical attention has focused on the management rather than the extent of relief allocation. Since human interventions almost always give rise to principal–agent problems, choosing the appropriate yardstick for effective famine relief is an abiding issue. In the past, because governing elites were remote from those at risk, they often relied on sub-bureaucracies and landowners to identify deserving recipients of relief. History is full of examples of trade-offs between the degree of delegation on the one hand, and corruption and red tape on the other (see, for example, Shiue, 2004).

The choice of appropriate public action in the presence of such agency problems during famines is discussed in Drèze and Sen (1989), Besley and Coate (1992), Ravallion (1997), and elsewhere.

Transfers of food at subsidized prices may risk corruption and hoarding; hence the frequent focus on the provision of non-tradable and highly perishable food rations. Income transfers (for example, through wages paid on public work schemes) are less likely to distort food markets, though if linked to work performance they may well discriminate against those in most need. Public works schemes also risk spreading infectious diseases. A further problem with public works is that fiscal stringency or fears of distorting labour markets, as in Ireland in the 1840s and in southern India in the 1870s, may entail below-subsistence wages and consequent excess mortality.

Private charity can mitigate famine but is rarely adequate during big crises. Since the 1950s famine relief has been globalized through non-governmental organizations (NGOs) such as Oxfam. NGOs have been effective at highlighting the link between Third World poverty and the risk of famine, and at fund-raising in the wake of highly publicized crises. Nonetheless, their record in mitigating and averting famine raises several issues.

First, agencies originally founded as famine relief agencies tend to reinvent themselves as bureaucracies. Such organizations must balance the public's wish to relieve disasters as they happen with their own need for bureaucratic sustainability. This has entailed focusing more on development than on famine relief per se. Budgetary pressures have also tempted NGOs to exaggerate the risks or gravity of famine, or to claim the credit when the crisis is ‘averted’ (De Waal, 1997). Given the likely long-term costs of such tactics, and the recent increasing dependence of NGOs on public funding, independent monitoring of their activities is essential. Moreover, NGO interventions typically lag, rather than lead, media reports; instead of drawing on previously accumulated reserves, they rely on crises to solicit aid, and their over-reliance on emergency-generated funding has led them to locations where they lack the detailed expertise and connections essential for effective famine relief. Most NGOs continue to spread themselves too thin, and are too small to offer the insurance required for a rapid response against famine.

Measuring the demographic cost

Soaring food prices and poor harvests are often harbingers of famine, but are neither necessary nor sufficient conditions for one. On the one hand, appropriate relief policies may prevent famine; on the other, not all famines result from aggregate food deficits or inflated food prices. An abnormal jump in mortality is a surer signal of famine, and is usually regarded as its defining feature. For most historical famines, however, establishing excess mortality with any precision is impossible, and inferences derived from incomplete data are often controversial. Much hinges on assumptions about the under-registration of deaths at the time. Controversy still surrounds the true tolls in the Soviet Union in 1931–33, China in 1959–61, Cambodia in 1975–79, and North Korea in 1995–99.

Nonetheless, it is clear that modern famines are, relatively speaking, far less costly in terms of human lives than earlier famines. Although non-crisis death rates in Africa remain high, excess mortality from famines in recent decades has been low. In Devereux's useful listing of major 20th-century famines only two – Nigeria in 1968–70 and Ethiopia in 1983–85 – are accorded tolls nearing one million (Devereux, 2000). Elsewhere, deaths were far fewer.

Although famine had virtually disappeared from Europe by the mid-19th century, 30 million is a conservative estimate of famine mortality in India and China alone between 1870 and about 1900, and ‘fifty million might not be unrealistic’ (Davis, 2001, p. 7). One hundred million would be a conservative guess at global famine mortality during the 19th century as a whole. Given that global population rose from about 1.3 billion in 1870 to 2.5 billion in 1950, in relative terms famines were much more lethal in the 19th century than in the 20th. The late 19th century saw a reduction in famine intensity in India, due to a combination of better communications and improvements in relief policy; in Russia, too, famines became more localized. Japan, where famines were common in the 17th century, and less so in the 18th, experienced its last true famine in the 1830s.

As noted earlier, infectious diseases usually account for most famine deaths. These include deaths due to diet-related diseases brought on by impaired immunity, or to poisoning from inferior or unfamiliar foods. They also include deaths stemming from the disruption of personal life and societal breakdown attendant on famine. Disease spreads with the increased mobility of the poor and the inevitable deterioration in sanitary conditions. Famines also are associated with outbreaks of seemingly unrelated diseases such as cholera, influenza, and malaria (Mokyr and Ó Gráda, 2002).

The implications of focusing on relative rather than absolute mortality are also worth noting. In relative terms, excess mortality in China in 1959–61 was modest compared, for example, with Ireland in the 1840s or Finland in 1867–68. The lower rate matters to the extent that it affected the characteristics of the famine. But such comparisons beg the question of the appropriate denominator. Most of these famines were regionally concentrated, but the denominators refer to larger political or geographical units. Finally, most famines last a year or two at most. Ireland in the 1840s, Cambodia in the 1970s, and North Korea in the 1990s are exceptional in this respect.

Although in the past non-crisis male life expectancy usually exceeded female, the evidence for a female advantage during famines is overwhelming (for example, Hionidou, 2006, p. 165; Maharatna 1996, pp. 231–4). The main reason for this is physiological. Whether the female advantage has changed over time remains a moot point, but there is some presumption that the female advantage is greater when the main cause of death is literal starvation. Most famine victims tend to be the very young and those beyond middle age, although the greatest *proportional* increases in death rates are at ages in between. In cases where population growth of two or three per cent per annum is the norm, such age and gender biases are

unlikely to have much impact, and population growth may be expected to quickly fill the resultant demographic vacuum. Where non-crisis growth is slow, these biases may matter more, and post-famine recovery is likely to be slower.

For several reasons, the demographic consequences of famine are more complex than implied by the standard measure of excess mortality. First, that measure ignores the drop in births that usually accompanies famine. Famines almost invariably entail significant reductions in births and marriages (for example, Maharatna 1996, pp. 179–83; Hionidou, 2006, pp. 178–89). There is a case for including the births deficit in the demographic reckoning. Births lost due to the Great Irish Famine numbered about 0.4 million in a population of eight million, whereas estimates for China in the wake of the 1959–61 famine run as high as 30 million in a population of 650 million (Yao, 1999). There are several reasons for such declines in the birth rate, including lower libido, spousal separation, and weaker reproductive functioning. Famines also usually entail fewer marriages although, clearly, in most situations marriage reductions have implications only for *first* births.

Second, the excess mortality measure omits both the rebound in the birth rate and the decline in the death rate that sometimes follow once the crisis has passed. Births in China in 1962 exceeded those in any year since 1951, and in the following three years the birth rate was also higher than in any other year in the 1950s and 1960s. Therefore, to some extent at least, births ‘lost’ during the famine seem to have been merely postponed.

Third, it leaves out of account any longer-run impact on mortality and morbidity. Famines hasten the deaths of some ill and elderly people who would have died soon in any case. The ensuing impact on the demographic structure entails a reduction in the death rate in the wake of famines.

Long-term health effects

Recent medical-historical research has revealed a close link between health and nutrition *in utero* and in early childhood on the one hand, and adult health and longevity on the other (Barker, 1992). The implications for the long-term demographic and health effects of famines are obvious. Research on Russian, Dutch and Chinese data links foetal exposure to famine to increased risks in later life of diseases as varied as schizophrenia, breast cancer, arteriosclerosis, and antisocial personality disorders (Khoroshinina, 2005, p. 208). There is evidence from Leningrad that being born just before or during famines reduces expected adult height (for example, Kozlov and Samsonova, 2005, pp. 178–89; Khoroshinina, 2005, pp. 198–200). Such evidence suggests that the human cost of famines has been underestimated in the past, although it is too soon to say by how much. Finally, there is the further disturbing possibility – still unexplored – that famine-induced malnutrition *in utero* or early childhood adversely affects the mental development of those at risk.

Conclusion

Famine's range has been narrowing since Malthus's time. By 1900 Europe and its industrialized extensions, Latin America, and Japan were virtually famine-free, and today major, *prolonged* famine anywhere is conceivable only in contexts of endemic warfare or self-enforced isolation. Compared with the persistent effects of HIV/AIDS on the population of sub-Saharan Africa, the damage wrought by

famine is minimal. Moreover, given that throughout most of history land hunger has been a powerful predictor of famine, recent trends in the balance between population and food production offer room for cautious optimism about the near future. In both Asia and Latin America, food production has grown much faster than population since the 1960s. In sub-Saharan Africa the balance has been much closer, although the problem there has been very rapid population growth rather than sluggish food output growth. Moreover, some African countries such as Burkina Faso and Niger have walked a high demographic tightrope while others (such as Malawi and Zimbabwe) have performed poorly despite slower population growth.

The few remaining places still vulnerable to textbook Malthusian famine are those yet to undergo the fertility decline of the demographic transition. Those countries have experienced considerable mortality improvement in recent decades, but they lag behind in terms of fertility decline. A key issue is how fertility decline, scarcely yet under way, unfolds in such vulnerable economies. The experience of post-fertility transition economies worldwide strongly mirrors the historical pattern whereby declines in fertility were preceded by declines in mortality. However, the length of the lag and the extent of the fertility decline are clearly crucial. A guarded historical lesson for countries like Niger is that the transition, once under way, has been more rapid in latecomers than in pioneers. Africa's sluggish fertility transition, itself a function of economic underdevelopment, has increased its share of global population from only 8.8 per cent in 1950 to 14 per cent today; it is set to reach 21.7 per cent by 2050. Even though a drop in the annual growth rate from 2.5 per cent during the second half of the 20th century to 1.4 per cent during the first half of the 21st in Africa as a whole is implied, population is predicted to treble by 2050 in famine-prone countries such as Niger, Uganda and Mali. When coupled with the problem of global warming, which is likely to impact disproportionately on the productivity of arid lands limited to a short growing season, the implied threat to living standards is clear.

See Also

- Malthus, Thomas Robert
- poverty
- Sen, Amartya

Bibliography

Banik, D. 2002. Democracy, drought and starvation in India: testing Sen in theory and practice. Ph.D. thesis, Department of Political Science, University of Oslo.

Barber, J. and Dzeniskevich, A., eds. 2005. *Life and Death in Leningrad, 1941–44*. London: Palgrave Macmillan.

Barker, D.J.P., ed. 1992. *Fetal and Infant Origins of Adult Disease*. London: BMJ Publishing Group.

Besley, T. and Coate, S. 1992. Workfare vs. welfare: incentive arguments for work requirements in poverty alleviation programs. *American Economic Review* 82, 249–61.

Davies, R.W. and Wheatcroft, S.G. 2004. *The Years of Hunger: Soviet Agriculture, 1931–33*. London: Palgrave Macmillan.

Davis, M. 2001. *Late Victorian Holocausts*. London: Pluto.

Devereux, S. 2000. Famine in the twentieth century. Working Paper No. 105, Institute of Development Studies, University of Sussex.

De Waal, A. 1997. *Famine Crimes: Politics and the Disaster Relief Industry in Africa*. Oxford: James Currey.

Drèze, J. and Sen, A. 1989. *Hunger and Public Action*. Oxford: Oxford University Press.

Dyson, T. and Ó Gráda, C., eds. 2002a. *Famine Demography: Perspectives from the Past and Present*. Oxford: Oxford University Press.

Hionidou, V. 2006. *Famine and Death in Occupied Greece, 1941–1944*. Cambridge: Cambridge University Press.

Howe, P. and Devereux, S. 2004. Famine intensity and magnitude scales: a proposal for an instrumental definition of famine. *Disasters* 28, 353–72.

Khoroshinina, L. 2005. Long-term effects of lengthy starvation in childhood among survivors of the siege. In Barber and Dzeniskevich (2005).

Kozlov, I. and Samsonova, A. 2005. The impact of the siege on the physical development of children. In Barber and Dzeniskevich (2005).

Maddison, A. 2006. World Population, GDP and Per Capita GDP, 1–2003 AD (2006 update). Online. Available at http://www.ggdc.net/maddison/, accessed 31 January 2007.

Maharatna, A. 1996. *The Demography of Famines: An Indian Historical Perspective*. Delhi: Oxford University Press.

Mokyr, J. and Ó Gráda, C. 2002. What do people die of during famines? The Great Irish Famine in comparative perspective. *European Review of Economic History* 6, 339–64.

Ó Gráda, C. 2005. Markets and famines in pre-industrial Europe. *Journal of Interdisciplinary History* 26, 143–66.

Ó Gráda, C. 2007. Making famine history. *Journal of Economic Literature* 31, 3–36.

Persson, K.-G. 1999. *Grain Markets in Europe 1500–1900, Integration and Regulation*. Cambridge: Cambridge University Press.

Ravallion, M. 1987. *Markets and Famines*. Oxford: Oxford University Press.

Ravallion, M. 1997. Famines and economics. *Journal of Economic Literature* 35, 1205–42.

Sen, A. 1981. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Oxford University Press.

Sen, A. 2001. *Development as Freedom*. Oxford: Oxford University Press.

Shiue, C.H. 2004. Local granaries and central government disaster relief: moral hazard and intergovernmental finance in eighteenth- and nineteenth-century China. *Journal of Economic History* 64, 100–24.

Tauger, M. 2006. Arguing from errors: on certain issues in Robert Davies’ and Stephen Wheatcroft's analysis of the Soviet grain harvest and the Great Soviet Famine of 1931–33. *Europe–Asia Studies* 58, 973–84.

Yao, S. 1999. A note on the causal factors of China's famine in 1959–1961. *Journal of Political Economy* 107, 1365–69.

How to cite this article

Gráda, Cormac Ó. "famines." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000318> doi:10.1057/9780230226203.0551

The New Palgrave Dictionary of Economics Online

Farr, William (1807 – 1883)

R.M. Smith

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Farr, W.; human capital; mortality

Article

William Farr, born in Kenley, Shropshire on 30 November 1807, died in London on 14 April 1883, was a statistician in the General Register Office who had been appointed in 1840 as ‘compiler of abstracts’ and was two years later made Statistical Superintendent, a post he held until his retirement in 1880. He pioneered the quantitative study of morbidity and mortality and in the process became one of Victorian England's most prominent figures in the public health and reform movements (Cullen, 1975). He made major contributions in the fields of data collection, being largely responsible for the introduction of a cause of death classification which was linked with his derivation of the ‘zymotic’ theory of epidemic disease (Eyler, 1979; Pelling, 1978). As an Assistant Census Commissioner for each of the censuses of 1851, 1861 and 1871, he was largely responsible for the development of reliable procedures for the recording of occupations (McDowall, 1983). He is, however, best known as a statistical analyst, for in 1843 he constructed the first English Life Table based on deaths in 1841 linked to the census of that year. At the same time he established the formula for deriving from a rate of mortality by age *m* the probability of survival *p* at the initial age. In 1850 and 1864 Farr produced his second and third English Life Tables, the last mentioned being used as the actuarial basis for the life insurance scheme set up by the Post Office for its employees. Farr in his work on occupational mortality was the first to make extensive use of the standard mortality rate, allowing comparisons of the mortality of different groups by means of a summary statistic which took account of differences in the age structure of the groups being compared. A recurring theme in his work was the identification of variation in mortality in different urban areas of the country. Such differential mortality was viewed as an index of human welfare. For example, in 1850 one-tenth of the registration districts, those he named ‘healthy districts’, had average mortality rates not exceeding 17 per 1,000, a rate he thought indicative of the ‘natural’ mortality which, when exceeded, would indicate those deaths attributable to unnatural and preventable diseases. An underlying aim in much of his work was to discover statistical laws or numerical expressions of

regularities such as he proposed in the laws of recovery and death in smallpox, the elevation law for cholera mortality in London (Lewes, 1983) and the law of the relation between population density and mortality. He was also an early contributor to human-capital theory (Kiker, 1968) arguing, in particular, that the economic value of men varied with age as well as social class, and this he used as powerful publicity for urban reform by drawing attention to the financial losses that followed from diseases that were the causes of death and illness in society at large.

Selected works

1843a. Causes of the high mortality in town districts. *5th Annual Report of the Registrar General*, 406–35 (*Parliamentary Papers* XXI, 200–15).

1843b. English Life Table No. 1. *5th Annual Report of the Registrar General*, 354–8, 366–7 (*Parliamentary Papers* XXI, 168–71, 178) and *6th Annual Report of Registrar General*, 517–666 (*Parliamentary Papers*, 1844, XIX, 290–358).

1850. English Life Table No. 2: Males. *12th Annual Report of the Registrar General*, Appendix, 73–152.

1852a. Influence of elevation on the fatality of cholera. *Journal of the Statistical Society* 15, 155–83.

1852b. *Report on the Mortality of Cholera in England, 1848–49*. London: HMSO.

1854. Vital statistics. In *A Descriptive and Statistical Account of the British Empire: Exhibiting Its Extent, Physical Capacities, Population, Industry, and Civil and Religious Institutions*, ed. J.R. McCulloch, 4th edn. London: Longman, Brown, Green and Longmans.

1859. English Life Table No. 2: Females. *20th Annual Report of the Registrar General*, Appendix, 177–203 (*Parliamentary Papers*, 1859, sess. 2 XII).

1864. *English Life Table: Tables of Lifetimes, Annuities and Premiums, with an Introduction by William Farr, M.D., F.R.S., D.C.L.* London.

1866. Mortality of children in the principal states of Europe. *Journal of the Statistical Society* 29, 1–35.

1867–8. *Report of the Cholera Epidemic of 1866 in England: Supplement 29th Annual Report on the Registrar General* (*Parliamentary Papers* XXXVI).

1885. *Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr, M.D. D.C.L. C.B. F.R.S*, ed. N.A. Humphreys. London.

Bibliography

Cullen, M.J. 1975. *The Statistical Movement in Early Victorian Britain: The Foundations of Empirical Social Research*. Brighton: Harvester Press.

Eyler, J.M. 1979. *Victorian Social Medicine: The Ideas and Methods of William Farr*. Baltimore and London: Johns Hopkins University Press.

Kiker, B.F. 1968. *Human Capital: In Retrospect*. Essays in Economics No. 16. Columbia: Bureau of Business and Economic Research, University of South Carolina.

Lewes, F. 1983. William Farr and cholera. *Population Trends* 31(Spring), 8–12.

McDowall, W. 1983. William Farr and the study of occupational mortality. *Population Trends* 31 (Spring), 21–4.

Pelling, M. 1978. *Cholera, Fever and English Medicine*. Oxford: Oxford University Press.

How to cite this article

Smith, R. M. "Farr, William (1807–1883)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000025> doi:10.1057/9780230226203.0552

The New Palgrave Dictionary of Economics Online

Farrell, Michael James (1926 – 1975)

Christopher Bliss
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Andrews, P.; community indifference curves; competitive equilibrium; convexity; Farrell, M. J.; firm, theory of; linear programming; profit maximization; rational expectations; Stone, J. R. N.; subjective probability; yield gap

Article

M.J. Farrell was born in 1926 and read Politics, Philosophy and Economics at New College, Oxford, graduating with First Class Honours. He moved to Cambridge in 1949 to work with Richard Stone at the Department of Applied Economics. He became a Fellow of Gonville and Caius College and the University made him Lecturer in Economics and eventually Reader. He was Editor of the *Review of Economic Studies* and a Fellow of the Econometric Society. In 1957 Farrell contracted poliomyelitis which left him dependent on crutches to get about. He died in 1975.

The bibliography of Farrell's work provided by Fisher (1976) lists 25 journal papers, about one a year in a cruelly shortened academic life. The quality of these papers is remarkable. They reveal the clarity of their author's mind and an outstanding creativity. Farrell often answered questions that others had hardly considered.

As a young man Farrell was influenced by Phillip Andrews, the author of *Manufacturing Business*, and they shared a dissatisfaction concerning the prevailing theory of the firm: ‘They [economists in the 1920s and 1930s] reduced the theory of the firm to a maximization problem soluble by the most elementary application of the differential calculus …’ and ‘Unfortunately these conclusions did not fit the regrettably complex facts well …’ (Farrell, 1971, p. 10). Farrell's work on the theory of the firm displayed an acute understanding of the subtlety of profit maximization as a strategy. In (1954) he provided one of the first applications of linear programming to this field. Farrell believed that the case for profit maximization eventually depended in part on the operation of a selection process. His (1970) paper remains to this day one of the best papers ever written on that topic.

Farrell wrote on the measurement of productive efficiency, on the consumption function, and on welfare economics. On some topics he produced a single paper – his last was on social choice theory.

In (1959) Farrell made two observations which were important innovations at the time. First, he exposed what he called ‘the fallacy’. This is a confusion between sufficient and necessary conditions for competitive equilibrium to be efficient. Convexity, as Farrell neatly demonstrated, is *sufficient* for existence of equilibrium but is *necessary* neither for existence nor for efficiency. Second, ‘… concavities in individual indifference maps disappear when one aggregates over a large enough number of individuals’ (1959, p. 381).

This deep aggregation result, which gave rise to an extensive literature (see, for example, Arrow and Hahn, 1971, chs 7 and 8), is based on a simple point. To illustrate it consider consumers and let them all have the same tastes, which may be represented by $U(x)$, where x is a vector of consumptions. Suppose that $U(x_1) = U(x_2)$ and let there be N consumers. We now wish to see whether a convex combination of $N \cdot x_1$ and $N \cdot x_2$, that is $\lambda \cdot N \cdot x_1 + (1-\lambda) \cdot N \cdot x_2$, can be distributed so as to make each consumer at least as well off as with x_1 or x_2. If it can, community indifference curves will be convex even if those derived from $U(\ )$ are not.

If consumers were indefinitely divisible we could achieve this result by giving x_1 to $\lambda \cdot N$ consumers and x_2 to $(1-\lambda) \cdot N$ consumers. However, as $\lambda \cdot N$ may not be an integer this exact procedure is inadmissable. Nevertheless, as N becomes large an integer $M<N$ will eventually emerge such that M/N approximates λ to any desired degree of accuracy. Hence Farrell's result follows.

Farrell treated the often sloppily discussed question whether speculation could be destabilizing and still profitable, in (1966). His demonstration within a very general framework that linearity of demand functions is required to exclude this possibility greatly advanced the general understanding of this problem.

In (1962) Farrell considered the well-known problem of the yield gap, the observation that equities at certain times show a different rate of return from that obtained from bonds. He provided some calculations which showed that there had been yield gaps in the past even when returns were corrected for capital gains. In considering what light these *ex post* observations throw on investors’ *ex ante* decisions, Farrell asked ‘… what do we mean by perfect knowledge in a market where uncertainty is present?’ (1962, p. 835). This led him to analyse what he called ‘accurate’ expectations: ‘… an individual's expectation is “accurate” if his subjective probability distribution is the same as the hypothetical frequency distribution by which we represent the real world’ (1962, p. 836). Long before the idea of rational expectations became fashionable, Farrell saw its relevance to the analysis of securities markets. However the careful student of profit maximization and selection processes found no reason to assume that expectations would necessarily be ‘accurate’.

Selected works

1954. An application of activity analysis to the theory of the firm. *Econometrica* 22, 291–302.

1959. The convexity assumption in the theory of competitive markets. *Journal of Political Economy* 67, 377–91.

1962. On the structure of the capital market. *Economic Journal* 72, 830–44.

1966. Profitable speculation. *Economica* 33, 183–93.

1970. Some elementary selection processes in economics. *Review of Economic Studies* 37, 305–19.

1971. Philip Andrews and manufacturing business. *Journal of Industrial Economics* 20, 10–13.

Bibliography

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. Amsterdam: North-Holland.

Fisher, M.R. 1976. The economic contribution of Michael James Farrell. *Review of Economic Studies* 43, 371–82. (Includes a complete discussion of Farrell's works.)

How to cite this article

Bliss, Christopher. "Farrell, Michael James (1926–1975)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000026> doi:10.1057/9780230226203.0553

The New Palgrave Dictionary of Economics Online

Fasiani, Mauro (1900 – 1950)

Massimo Finoia
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; Consumption taxation; corporate state; Einaudi, L.; excise tax; Fasiani, M.; fiscal illusion; Fisher, I.; Fuoco, F.; labour supply; mathematical economics; Pareto, V.; production at constant costs; public debt; public finance; Puviani, A.; stabilization policy; tax incidence; tax shifting; taxation of income; taxation of saving; Viti de Marco, A. de

Article

Fasiani was born in Turin and died in Genoa. Clearly the most important Italian scholar of fiscal theory to emerge in the interwar period (Buchanan, 1960, p. 36), he taught public finance in Turin, Sassari, Trieste and, from 1934, in Genoa. His career was rapid and exclusively academic. Despite his untimely death, he left important works on fiscal theory, and also on economic theory, economic policy and the history of economic thought.

Following Pareto's theory of the ruling class and Puviani's idea of fiscal illusion, which he rediscovered, Fasiani asserts that fiscal activity is to be explained on the basis of the nature of the political entity and not in terms of economic calculus or by sacrifice theories or by the ability-to-pay principle (1932a, 1941). As taxation and public expenditure are political phenomena, it is impossible to know the laws of fiscal activity. Fiscal theory can only be built through static models reflecting the different types of political societies. To De Viti de Marco's models of the ‘monopolistic’ state, where the ruling class governs only in its own interest, and of the ‘cooperative’ state, where the ruling class governs in the interest of every member of the community, Fasiani adds the model of the ‘modern, nationalistic or corporative’ state, in which the ruling class governs in the interest of the collectivity, considered as a whole (1941).

He dealt with the duration of the process of tax shifting (1934) and with the characteristics of intermediate positions in the transition from one state of equilibrium to another (1932b); with tax shifting in conditions of constant, increasing and decreasing costs in competition and in monopoly (1941, App. I and II) and with the effects of an excise tax under conditions of industrial concentration (1942a). He analysed the different elements determining the ‘quantity of labour’ and proved the

impossibility of understanding the effects of taxation on labour supply assuming as variables only working hours and income (1942c). He devoted much research to the problem of the double taxation of saving (1926), confirming the validity of J.S. Mill's thesis in opposition to the theories of Einaudi and of Fisher. Fasiani also wrote important notes on the application of the Paretian indifference curve apparatus to the classical problem of the relative burden of income tax and consumption tax (1930) and on the analysis of the relationship between taxation and risk-taking (1935b).

In order to study the effects of taxation in a state of equilibrium, Fasiani re-examined and criticized some problems of economic theory. Among other things, he reasserted the hypothesis of production at constant costs and redefined the variables of the labour supply. Specifically he dealt with business cycles and stabilization policy, giving a decisive role to monetary policy (1935a; 1937a; 1942b).

His most important work in the history of economic thought is a very long essay on fiscal theory in Italy (1932c). In this work Fasiani critically examined the general theories of public finance formulated in Italy between 1880 and 1930, that is to say the economic theory, the political theory, the sociological theory, and also the theses on the effects of taxation and public debt on tax shifting and tax incidence. Finally, the essays on fiscal theory in the 18th century (1936) and on Francesco Fuoco (1774–1841), a forerunner of mathematical economics (1937b), are worthy of note.

Selected works

A full bibliography of Fasiani's works is contained in: *Rivista di Diritto finanziario e Scienza delle Finanze* 9 (September 1950), 216–18.

1926. Sulla teoria dell’esenzione del risparmio dall’imposta. *Memorie della Reale Accademia delle Scienze di Torino* 61, offprint.

1930. Di un particolare aspetto delle imposte sul consumo. *La Riforma Sociale* 41(January–February), 1–20.

1932a. Temi teorici ed ‘exponibilia’ finanziari. *La Riforma Sociale* 43(July–August), 383–425.

1932b. Velocità delle variazioni della domanda e dell’offerta e punti di equilibrio stabile e instabile. *Atti della Reale Accademia delle Scienze di Torino* 67, 383–425.

1932c. Der gegenwärtige Stand der reine Theorie der Finanzwissenschaft in Italien. *Zeitschrift für Nationalökonomie* 3(3), 651–91; 4(1) (1933), 79–107; 4(3), 357–88.

1934. Materials for a theory of the duration of the process of tax shifting. *Review of Economic Studies* 1 (February), 81–101; 2, February 1935, 122–37.

1935a. Fluttuazioni economiche ed economia corporativa. *Annali di Statistica e di Economia* 3, 1–70.

1935b. Imposta e rischio. In AA. VV., *Studi in onore del prof. Salvatore Ortu Carboni*. Roma:

Tipografia del Senato.

1936. Precedenti di alcune recenti teorie finanziarie. *Annali di Statistica e di Economia* 4, 195–240.

1937a. Principi generali e politiche della crisi. *Annali di Statistica e di Economia* 12, 25–108.

1937b. Note sui ‘Saggi economici’ di Francesco Fuoco. *Annali di Statistica e di Economia* 5, 1–131.

1941. *Principi di Scienza delle Finanze*, 2 vols. Turin: Giappichelli; 2nd edn, 1951.

1942a. La translazione dell’imposta in regime di concentrazione industriale. *Studi Economici Finanziari Corporativi* 2(April-September), 200–25.

1942b. Potenziale di lavoro e moneta. *Annali di Statistica e di Economia* 9–10, 65–137.

1942c. Appunti critici sulla teoria degli effetti dell’imposta sull’offerta individuale di lavoro. *Annali di Statistica e di Economia* 9–10, 139–233.

Bibliography

Buchanan, J.M. 1960. La scienza dell finanze: the Italian tradition in fiscal theory. In J.M. Buchanan, *Fiscal Theory and Political Economy, Selected Essays*. Chapel Hill: University of North Carolina Press.

Cosciani, C. 1950. Mauro Fasiani. *Economia Internazionale* 3(November), 913–19.

Einaudi, L. 1950. Mauro Fasiani. *Rivista di Diritto finanziario e Scienza delle Finanze* 9(September), 199–201.

Scotto, A. 1950. Gli scritti di Mauro Fasiani. *Rivista di Diritto finanziario e Scienza delle Finanze* 9 (September), 202–15.

How to cite this article

Finoia, Massimo. "Fasiani, Mauro (1900–1950)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000029> doi:10.1057/9780230226203.0554

The New Palgrave Dictionary of Economics Online

Faustmann, Martin (1822 – 1876)

Anthony Scott

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Faustmann, M.; forests; Ohlin, B. G.; rotation periods

Article

Faustmann was a German forester who spent much of his life working on the grand-ducal forests of Hesse. Between 1849 and 1865 he entered into controversies with other foresters concerning methods of forest valuations, his ideas eventually prevailing among that minority of forest economists who accepted the discipline of a positive rate of interest in making forest calculations. Although it has been said that his work was approved by such ‘national economists’ as Wagner and Roscher (*Allgemeine Deutsche Biographie*, 1877), it was evidently quite unknown to the more theoretically oriented German and Austrian specialists in capital and interest. Incorrect solutions to the optimum forest rotation problem were subsequently offered by such economists as Jevons, J.B. Clark and Irving Fisher, in the course of simplified expositions of the idea of the production period of a single investment. Not until the 1950s did economists working outside forestry realize that Faustmann's approach as explained to generations of resistant forestry school students contained a correct approach to the forestry question.

The economists’ discovery was sparked by F. and V. Lutz, M. Gaffney, P.H. Pearse and, a few years later, Paul Samuelson. (The literature suggests that some Scandinavian and German economists, notably Ohlin, either knew of Faustmann's formula or worked it out for themselves.)

Faustmann's formula is derived from his investigations into forest values, needed at that time to guide the allocation of landowners’ acres between trees and agriculture. His predecessors had consequently attempted to value the soil and the forest separately. In this they failed, partly because they confused stocks and flows. Faustmann cleared this up in 1849 by providing a single forward-looking approach for the present value of the next and future forest crops. As his professional readership required, his formulation also made it possible to take account of expected planting, husbanding, thinning and harvesting net costs during the life of each subsequent stand. He was able to solve his predecessors’ problem by showing that the soil value (with which agricultural values are to be compared) is the value of the forest enterprise when it is still bare land, before a crop rotation has been commenced.

Faustmann is known today by resource economists for two by-products of his original perception. First, he showed correctly how to calculate the rotation age that is optimal for the owner in the presence of all expected costs and expected subsequent harvests. Second, by including the expected net discounted returns from subsequent rotations in his value and rotation-age formulae, he took the step that later eluded 20th-century economists, such as Fisher. He included the implicit forgone rent or shadow price of the land. He showed that the effect of doing this is that a given growth-and-harvest cycle will be shorter than economists' analyses would have predicted. Shorter rotations advance the date on which the next and all subsequent rotations will be harvested, thus reducing the effect of waiting on calculated soil values.

Faustmann made subsequent contributions to professional forestry, but they are of little interest today.

Selected works

1849. Berechnung des Werthes, welchen Waldboden sowie noch nicht haubare Holzbestände für die Waldwirtschaft besitzen. *Allgemeine Forst und Jagd-Zeitung* 25, 441–55. Trans. W. Linnard and included in Gane (1968). Samuelson (1976) contains an extended bibliography.

1877. Faustmann, Martin. In *Allgemeine Deutsche Biographie*, vol. 6. Leipzig: Duncker & Humblot.

Bibliography

Bently, W.R. and Teeguarden, D. 1965. Financial maturity: a theoretical review. *Forest Science* 2 (March), 76–87.

Dickson, H. 1953. Forest rotation. *Weltwirtschaftliches Archiv* 70.

Dowdle, B., ed. 1974. *The Economics of Sustained Yield Forestry*. Seattle: College of Forestry, University of Washington.

Fernow, B.E. 1902. *Economics of Forestry*. New York: T.Y. Crowell.

Fernow, B.E. 1911. *History of Forestry*. Revised edn, Toronto: University of Toronto Press.

Gaffney, M. 1960. *Concepts of Financial Maturity of Timber and Other Assets*. Agricultural Economics Information Series No. 62. Raleigh: Department of Agricultural Economics, North Carolina State College.

Gane, M., ed. 1968. *Martin Faustmann and the Evolution of Discounted Cash Flow: Two Articles from the original German of 1849*. Trans. W. Linnard (includes translated articles by von Gehren and Faustmann). Oxford: Commonwealth Forestry Institute; Institute Paper No. 42.

Hiley, W.E. 1930. *The Economics of Forestry*. Oxford: Clarendon Press.

Lutz, F. and Lutz, V. 1951. *Theory of Investment of the Firm*. Princeton: Princeton University Press.

Pearse, P.H. 1967. The optimal forest rotation. *Forestry Chronicle* 43(June), 178–95.

Samuelson, P.A. 1976. Economics of forestry in an evolving society. *Economic Inquiry* 14, 466–92. This article evolved from the symposium edited by B. Dowdle (1974).

Scott, A. 1983. *Natural Resources and the Economics of Conservation*. 3rd edn, Ottawa: Carleton University Press. (See Appendix to ch. 3.)

How to cite this article

Scott, Anthony. "Faustmann, Martin (1822–1876)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000031> doi:10.1057/9780230226203.0555

The New Palgrave Dictionary of Economics Online

Federal Reserve System

Donald D. Hester

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Federal Reserve System was established in 1913 to provide the United States with an elastic currency. It managed security offerings to finance the First World War, and evolved from a set of 12 semi-autonomous banks to a centralized institution in the 1920s. Having failed to prevent the Depression of the early 1930s, it was substantially reorganized in 1933 and 1935. After the Second World War and a 1951 accord reached with the Treasury, it started on an odyssey of monetary policy interventions, employing many policy instruments, indicators, and powers with varying degrees of success to the present day.

Keywords

‘availability of credit’ doctrine; Accord (1951); availability of credit; bank holding companies; Bank Holiday; central banking; commercial paper; deposit turnover; Depository Institutions Deregulation and Monetary Control Act; discount rate; discount window; discretionary monetary policy; eligible paper; exchange rates; federal funds; federal reserve system; Federal Open Market Committee (FOMC); financial deregulation; financial holding company; Financial Services Modernization Act; free gold; Garn–St Germain Act; Glass–Steagall Act; gold standard; Great Depression; inflation; Meltzer, A.; money supply measures; net free reserves; open-market operations; real bills doctrine; repurchase agreements; reserve requirement; retail sweep programmes; Riefler–Burgess doctrine; unemployment

Article

The Federal Reserve System of the United States was established on 23 December 1913, when President Woodrow Wilson signed the Federal Reserve Act. The need for a new federal banking institution became clear when a severe crisis occurred in 1907. In May 1908 the Aldrich–Vreeland Act established a bipartisan National Monetary Commission that proposed establishing a National Reserve Association with 15 locally controlled branches that would ‘provide an elastic note issue based on gold and

commercial paper' (Warburg, 1930, p. 59). The proposal was not enacted, nor was a subsequent proposal for a central bank with about 20 branches that would be controlled by a centralized Federal Reserve Board, consisting largely of commercial bankers. In the debate preceding the Federal Reserve Act, banking industry domination was rejected in favour of a board that had five members appointed by the President and two ex officio members, the Secretary of the Treasury and the Comptroller of the Currency. The appointed members had staggered terms and were to represent different commercial, industrial, and geographic constituencies. A sixth appointed member representing agriculture was added in 1923. The composition of the Board and its relation to Federal Reserve banks were drastically changed in 1935. Partly because of continuing disagreements about public versus commercial bank control, the new Board's powers were left ambiguous in the act.

The act mandated that all national banks become members of the new system and stockholders of Federal Reserve banks. Because reserves were to be concentrated in 12 Federal Reserve banks, the act substantially reduced reserve requirements at national banks. State chartered banks could join if they chose to and were judged to be financially strong. The first Board was sworn in on 10 August 1914 and the system opened for business on 16 November 1914. Federal Reserve notes that were backed 100 per cent by 'eligible paper' and, additionally, 40 per cent by gold began to circulate. Eligible paper was self-liquidating, short-term paper that arose in commerce and industry. The rationalization for eligible paper was the real bills doctrine, which held that credit extended for financing only the production and distribution of goods would not lead to inflation. The doctrine is invalid because of fungibility; there is no relation between paper acquired by Federal Reserve banks and loans the commercial banks are extending. In addition, all deposits at Federal Reserve banks had to be backed at least 35 per cent by gold. Subsequent amendments to the act effectively eliminated the supra-100 per cent collateralization of notes. A June 1917 amendment to the act forced all member banks to pool required reserves at Federal Reserve banks and further reduced reserve requirements to decrease the burden of membership on national banks and attract more state-chartered banks to the system.

The early years

The early years of the Federal Reserve System were marked by struggles to define the distribution of power between Federal Reserve banks and the Board, in the context of growing US involvement in the First World War. The Board gradually assumed more powers, but was unsuccessful in controlling open-market trading, which inevitably was concentrated in New York. Benjamin Strong, the New York bank governor, managed system trading. (Until 1935 the chief executives of Federal Reserve banks were called 'governors'. After 1935 their title was changed to 'president' and members of the Board were called 'governors'.) The Federal Reserve System was made fiscal agent for the Treasury in 1920, but the Treasury dealt directly with Federal Reserve banks, not the Board. Until 1922 the Board's statistical research office was located in New York, and arguably the Board was less informed than the New York bank about money market conditions.

Federal Reserve banks immediately sought earning assets in order to pay expenses and the six per cent required dividends on member bank capital subscriptions. As they expanded their portfolios of bills, US securities, discounted commercial paper, and acceptances, the breadth and liquidity of these markets increased. In early 1915 the New York bank was buying and selling for other Federal Reserve banks. Discount rates charged by reserve banks varied across Federal Reserve districts.

In anticipation of the US declaration of war on Germany in 1917, Federal Reserve banks became responsible for issuing and redeeming short-term Treasury debt certificates before and during Liberty Loan drives. There would be four large Liberty Loans and a Victory Loan in 1919 that required extensive Federal Reserve involvement. US bonds were sold to the public on an instalment plan by member banks; the interest rate banks charged on the unpaid balance on a bond was equal to the coupon rate on the bond. Member banks, in turn, discounted short-term US debt at Federal Reserve banks at an interest rate below the yield on the debt, which allowed them to recover their costs of instalment lending. US government interest-bearing debt rose from $1.0 billion at the end of 1916 to $25.5 billion at the end of 1919, and would never again fall below $15 billion. This huge increase, and the fact that Federal Reserve banks offered preferentially low interest rates when member banks discounted government debt, had important lasting consequences on the money market. Before the war, Federal Reserve banks had schedules of discount rates that varied across the quality and maturity of discounted paper and the amount of borrowing by a member bank. Because of the low discount rate on government debt, member banks almost exclusively offered it as collateral when borrowing. The discount rate effectively became the rate charged on government debt. By 1922 each reserve bank effectively had a single discount rate, but rates still varied across Federal Reserve districts.

The November 1918 armistice brought new challenges. Continuing shortages of food and other goods in Europe and large increases in the stock of money led to inflation in the United States. The rate of inflation peaked in May 1920 and was followed by a sharp deflation in the following year of about 45 per cent in wholesale prices. In that year industrial production fell by about 30 per cent and unemployment soared. Until October 1919 Federal Reserve banks were obliged to keep the low wartime discount rates in order to allow banks and the public to absorb the 1919 Victory Loan. In November, Federal Reserve banks began raising their discount rates in an effort to combat inflation. In June 1920 four banks raised the rate to seven per cent. Amplifying the effects of the interest rate increases was an outflow of gold to Europe and a sharp reduction in discount window borrowing as Federal Reserve banks cut back on subsidizing the public's instalment purchases of US bonds.

The Boston bank lowered its rate from seven per cent to six per cent in April 1921, and was gradually followed by other reserve banks in an effort to respond to the slowdown. Deposits at all member banks reached a local maximum of $26.1 billion in the December 1919 call report and then fell to $22.8 billion in the April 1921 report. Discount window borrowings reached a year end high of $2.7 billion in December 1920 and then fell to $0.6 billion at the end of 1922 as gold flows turned positive. As gold flowed in, reserve banks lowered their discount rates to 4.5 per cent in 1923 and early 1924.

While gold inflows slackened after 1923, it became apparent that new operating guidelines were needed. Governor Strong understood that the real bills doctrine was invalid and that many countries were not acting according to the old gold-standard rules. As interest rates fell, most reserve banks were again acquiring securities to augment their income. Strong, on the other hand, had begun to sterilize the New York bank's holdings of gold by selling its securities in the open market. The Treasury was concerned that reserve bank trading was upsetting securities markets when it was buying or selling debt. In May 1922 the reserve banks established the Governors Executive Committee consisting of the governors of the Boston, Chicago, Cleveland, New York, and Philadelphia banks to manage transactions for all 12 banks. The committee executed orders on behalf of the banks in the light of Treasury plans and made recommendations, but acted only as agents and had no executive power. In April 1923 it was renamed the Open Market Investment Committee (OMIC), which had the same membership as its predecessor

but was required

> to come under the general supervision of the Federal Reserve Board; and that it be the duty of this committee to devise and recommend plans for the purchase, sale and distribution of open-market purchases of the Federal Reserve Banks in accordance with … principles and such regulations as may from time to time be laid down by the Federal Reserve Board. (Chandler, 1958: 227–8)

Strong dominated the OMIC and began to understand the way open-market operations worked. He noted in particular that the sum of reserve bank open-market purchases and gold inflows almost equalled negative changes in member bank borrowing. He developed a case for active monetary policy and argued that restrictive monetary policy should be initiated with open-market sales and followed by increases in the discount rate. This was the likely origin of member bank borrowings and nominal interest rates as indicators of monetary policy. Policy instruments were open-market operations and the discount rate. While proposals to change discount rates originated with Federal Reserve banks, they required Board approval, which may explain why Strong preferred to lead with open-market operations. Strong was sensitive to the effects of monetary policy on prices, but objected to any legislated targeting of prices. His analysis was seriously incomplete when banks were not net borrowers from the Federal Reserve, and in such circumstances so were his policy tactics. Tragically, beginning in 1916 Strong suffered from recurrent attacks of tuberculosis and would die in October 1928, before such circumstances arose.

The 1923 Board Annual Report advocated an activist policy, but continued to support the real bills doctrine. In response to pressure from the Treasury and the Board, Federal Reserve banks sold most of their government securities in 1923; yearend holdings fell from $436 million to $134 million between 1922 and 1923. Federal Reserve notes and member bank reserves backed by such assets were unjustifiable under the doctrine, and the Treasury objected to Federal Reserve banks profiting from such assets. However, at the end of 1924 the banks held $540 million, and the banks' portfolio of government securities fluctuated considerably in the following years in response to changes in the volume of discounted bills and gold flows. Discount rates at Federal Reserve banks were lowered in the latter half of 1924 and 1925 before converging on four per cent at the beginning of 1926, largely following short-term interest rates in New York. Short-term market rates fell because of a sharp recession; the Federal Reserve index of industrial production (1997=100) fell from 7.84 in May 1923 to 6.43 in July 1924. Clearly policy was active, but not because of the real bills doctrine!

The discount rate was four per cent in June, when Federal Reserve banks began to cut the rate to 3.5 per cent and to make open-market purchases. At the beginning of 1928 discount rates were increased because of developing speculation in the stock market and continued to rise to as much as six per cent in October 1929, when the stock market crashed. In part, Federal Reserve discount rates were again responding to changes in industrial production, which had been quite sluggish until the end of 1927 and then began to grow rapidly until July 1929. In part, the 1927 rate cut reflected Federal Reserve efforts to help the United Kingdom maintain sales of gold at the pre-war sterling price, which had been restored in 1925. Governor Strong and Montagu Norman, the Governor of the Bank of England, were working to re-establish a gold standard that could restore order to international finance. To help the United Kingdom in 1925, the New York bank extended the Bank of England a $200 million gold credit and attempted to

keep interest rates low in New York relative to those in London. By reopening gold sales at the pre-war price, Britain had effectively revalued the pound upward in 1925 by about ten per cent, with devastating consequences for its economy.

As Strong's health failed in 1928, a leadership vacuum developed. In an attempt to coordinate policy among all 12 reserve banks and the Board, the Board proposed in August 1928 that the five member OMIC be replaced by a new Open Market Policy Committee (OMPC) that included all 12 reserve bank governors and was chaired by the Governor of the Federal Reserve Board. This proposal was rejected by bank governors, but a modified form was adopted in January 1930. Strong had been aware of growing stock market speculation and did not object to Federal Reserve open-market sales and the increase in the discount rate. These actions were reinforced by outflows of gold. In mid-1928 gold flows reversed, apparently attracted by high and rising short-term interest rates. Federal Reserve banks continued to sell bills and government debt, forcing member banks into the discount window to the extent of about $1 billion in the second half of 1928 and in the middle of 1929. At the end, Strong was aware of the danger of restrictive monetary policy actions over an extended period on the real economy, but remained reasonably optimistic that the situation could be controlled (Chandler, 1958: 460–3). After his death the struggle for control continued between his successor at the New York bank, George L. Harrison, and the Board; the latter argued that the real bills doctrine was not dead and that reserve banks should take direct action to penalize member banks making loans that supported security speculation. The Federal Reserve index of industrial production peaked in July 1929, Bureau of Labor Statistics (BLS) wholesale and consumer price indices had been slowly falling since 1926, and in October the stock market collapsed.

The Great Depression

Led by the New York bank, the Federal Reserve flooded the money market with cash by aggressively buying government securities. Discount window borrowing by member banks fell from $1,037 million in June 1929 to $632 million in December and to $271 million in June 1930. Further, discount rates at reserve banks were rapidly reduced; at the New York bank the rate was lowered from six per cent in October to 2.5 per cent in June 1930. The monthly average Standard and Poor common stock index (1935–1939=100) began to stabilize; it was 195.6 in January 1929, 237.8 in September, 159.6 in November, and 191.1 in April 1930. However, the index of industrial production continued to fall after the open-market purchases, and the BLS index of wholesale prices was ten per cent lower in 1930 than in 1929.

In mid-1930 reserve banks sharply reduced their purchases of government securities in the belief that monetary policy was adequately expansionary. The OMPC seems to have been guided by what Meltzer (2003: 164) calls the Riefler–Burgess Doctrine: ‘If [discount window] borrowing and interest rates were low, policy was easy; if the two were high policy was tight.’ An interpretation is that if member banks wanted to lend they could have inexpensive and relatively easy access to funds; if not, there was little more that the Federal Reserve could do. While total member bank discount window borrowing was positive, many banks were holding excess reserves. Conventional wisdom has it that the reserve banks should have continued buying securities. However, it is unclear even today whether continued large open-market purchases by the Federal Reserve would have had much of an impact on real economic activity in late 1930; the experiment was never tried. Rapid expansion of reserves and member bank deposits did occur in the late 1930s, with little effect on real economic activity.

On average about 600 bank failures a year occurred between 1920 and 1930; most failing banks were small and not members of the Federal Reserve System. The number of failing banks doubled in 1930 and increased by another 70 per cent in 1931. The total deposits of failing banks between 1920 and 1930 averaged less than $200 million a year, but more than quadrupled in 1930 and doubled again in 1931. Total deposits and currency had begun to fall after December 1928 and continued to fall after the stock market crash. Currency in circulation began to rise in November 1930, as bank failures increased. Industrial production and wholesale prices were falling at an accelerating rate. The directors of the New York bank counselled Governor Harrison to continue open-market purchases in 1930, but he encountered opposition in the OMPC and little was done. Net gold inflows were offset by open-market sales because the OMPC collectively believed monetary policy was expansionary. Reserve bank discount rates and money market interest rates trended down until 21 September 1931, when the United Kingdom suspended gold payments.

The British abandonment of gold led to very large withdrawals of gold and currency from the United States that were initially partially offset by open-market purchases of bills and increased discount window borrowing, which occurred at sharply higher interest rates as recommended by Bagehot (1873). However, Federal Reserve bank credit fell from $2.2 billion in October 1931 to $1.6 billion in March 1932. During this period of rising bank failures, rapidly declining economic activity, and falling prices, Harrison argued against open-market purchases for a number of reasons, but primarily because of the possibility of a shortage of ‘free gold’, that is, gold that was not required as collateral for Federal Reserve notes and reserves. The Glass–Steagall Act of 1932 authorized the Federal Reserve banks temporarily to use US government securities as collateral for Federal Reserve notes and thus largely solved the problem of a lack of free gold. In February 1932 Federal Reserve banks began aggressive open-market purchases of government securities that more than offset continuing gold losses and allowed member bank borrowings to fall about 50 per cent by August 1932. Discount rates at the New York and Chicago banks were lowered to 2.5 per cent in June 1932, but all other banks kept their rates at 3.5 per cent until the national banking ‘holiday’ that began on 5 March 1933 when President Roosevelt closed all US banks. Net free reserves (excess reserves minus discount window borrowing) had turned positive in September and thus signalled excessive ease to some individuals on the OMPC.

Restructuring the Federal Reserve System

It was obvious that the Federal Reserve had been ineffective in combating the collapse of the banking system and responding to the Great Depression. The banking system and the Federal Reserve needed to be restructured and strengthened. The Emergency Banking Act of 9 March 1933 authorized the Treasury to license and reopen national banks that were judged to be sound; state chartered banks that were sound would receive licences from state banking commissioners. Many reopening banks received capital injections by selling preferred stock to the Reconstruction Finance Corporation. At year end 1929 there were 24,026 commercial banks of which 8,522 were members of the Federal Reserve System; at year end 1933 there were 14,440 commercial banks of which 6,011 were member banks. For a period of one year all banks, whether members or not, could borrow on acceptable collateral from Federal Reserve banks.

Many of the reforms that were adopted would survive at least until late in the 20th century. Because of a belief that the collapse lay in undisciplined stock market trading, the Glass–Steagall Act of 1933

required that commercial banks divest themselves of investment banking activities. This act introduced deposit insurance that became effective in January 1934. It also banned interest payments on demand deposits and allowed the Board to impose ceilings on interest rates that banks could pay on time and savings deposits. Finally, the act renamed the OMPC the ‘Federal Open Market Committee’ (FOMC), but as in earlier incarnations its executive committee remained the same. The Securities Exchange Act of 1934 authorized the Board to impose margin requirements on stock market trades. Federal Reserve banks were authorized to make commercial and industrial loans to non-financial firms.

Having failed to expand reserve bank credit between July 1932 and February 1933, the Board found itself under extraordinary political pressure to expand resources to the banking system. As Meltzer (2003: 435–41) explains, President Roosevelt threatened to have the Treasury issue currency in the form of greenbacks if the FOMC failed to expand sufficiently. Net free reserves turned positive in May 1933 and rose to more than $3.0 billion by January 1936. The revaluation of gold in February 1934 together with subsequent large gold inflows from Europe and hesitancy to lend by member banks contributed to this surge in excess reserves.

The reconstruction of the Federal Reserve System continued with Roosevelt's nomination of Marriner Eccles to become Governor of the Federal Reserve Board in November 1934. Eccles had argued that system power should be concentrated in the Board and that reserve banks be prevented from undertaking open-market operations on their own accounts. Eccles's initiatives were opposed by Senator Carter Glass, many reserve bank governors, and the banking industry, but he largely succeeded in achieving his goals. The reforms were in the Banking Act of 1935, which restructured the Board to consist of seven appointed governors, each with a staggered 14-year term. The FOMC was restructured to consist of the seven governors and five reserve bank presidents. Two of the governors were to be appointed for four-year terms as chairman and vice-chairman of the Board by the president, with the advice and consent of the Senate. Eligible paper was no longer restricted to being short-term paper that originated in commerce and industry. The Board was empowered to vary reserve requirements; the upper limit was twice the percentages that were specified in the 1917 amendments to the Federal Reserve Act.

Members of the renamed Board of Governors of the Federal Reserve System took office in February 1936, with Eccles as chairman. For some time the FOMC had expressed concern about the inflationary potential of large excess reserves. In particular, because excess reserves exceeded reserve bank credit, the FOMC would not be able to absorb them without an increase in reserve requirements. Employing its new policy instrument, on 14 July 1936 the Board announced an increase in reserve requirements on August 15 of 50 per cent on all deposits at member banks. The increase was expected to absorb less than half of system excess reserves and was not expected to impinge on member bank lending or the economic recovery. In part because of continuing gold inflows, excess reserves were $3.0 billion at the end of July 1936, and averaged about $2.0 billion through the end of February 1937. Because excess reserves continued to be large, the Treasury began to sterilize gold inflows in December 1936, but not to the extent desired by the Board. At the end of January the Board announced a further two-step increase in reserve requirements of one-third to take place in March and May 1937. These actions took reserve requirements to their legal maxima and reduced excess reserves to below $800 million in summer months. In August and September reserve banks reduced their discount rates to one per cent or 1.5 per cent, levels that would last until December 1941. Coinciding with the May increase, the industrial production index (1997=100) reached a high of 10.4 and then decreased to 7.0 in May 1938. Continuing gold inflows and the Treasury's February 1938 abandonment of gold sterilization allowed excess

reserves to increase to $1.5 billion in March 1938. Beginning after the Board's reduction in reserve requirements of more than ten per cent in April 1938, excess reserves began a rise to nearly $7 billion in late 1940; however, industrial production did not pass its 1937 peak until October 1939, after the Second World War had begun in Europe.

Second World War and recovery

As the war approached gold flowed into the United States, and the FOMC allowed its security holdings to fall and their maturity to lengthen. In response to inflationary pressures, the Board introduced consumer credit controls in September 1941 and again raised reserve requirements to their legal maxima in November. After the United States declared war, monetary policy was constrained to facilitate war finance. In April 1942 the FOMC set interest rate ceilings on treasury bills at 0.375 per cent and on long-term bonds at 2.5 per cent. The yield curve was upward-sloping and effectively ‘pegged’ by these two boundary conditions into the post-war period. Because capital gains could be earned by buying high coupon securities and selling as they approached maturity, the cost of intermediate term debt was higher than rates shown on the yield curve. Discount rates were lowered to one per cent by all reserve banks and were not raised again until 1948. A preferential discount rate of 0.5 per cent was charged for loans collateralized by short-term US debt. Reserve requirements for central reserve city member banks were lowered in 1942, causing interest-free reserves to disappear into interest-bearing US securities. Finally, a variety of selective credit controls were imposed during and after the war, which ended in August 1945. Yearend deposits and government securities of member banks had risen from $61.7 billion and $19.5 billion in 1941 to $129.7 billion and $78.3 billion respectively in 1945. Because of the pegging of the yield curve, Federal Reserve bank yearend ownership of US securities rose from $2.3 billion in 1941 to $24.3 billion in 1945; treasury bills were $10 million in 1941 and $14.4 billion in 1946.
The preferential discount rate was eliminated in the spring of 1946. In July 1947 the FOMC relaxed the rate ceiling on treasury bills and the rate rose to about one per cent by yearend. Reserve banks raised the discount rate to 1.25 per cent in early 1948. Eccles's long term as chairman ended in February 1948, but he continued as a member of the Board. Reserve requirements were increased in 1948 as the Board sought to control inflation, although prices were actually falling at yearend when a recession occurred. Indeed, the reserve requirement policy instrument was used many times between April 1948 and February 1951 because it was perceived not to have a direct effect on treasury interest rates. A continuing struggle between the Board and the Treasury for an independent monetary policy would not be resolved until a spurt of inflation after the start of the Korean War led to an accord signed on 4 March 1951. It effectively freed the Board from pegging interest rates. Partly because of frictions leading to the accord, a new chairman, William McChesney Martin, Jr., was appointed in April.

Resumption of discretionary monetary policy

In the Martin era of discretionary monetary policy, new operating techniques were needed. In 1953 the FOMC settled on a policy of ‘bills only’, which meant that open-market operations would be largely confined to the market for treasury bills, because it was recognized that large policy actions in thin markets could impair market efficiency. Indicators of monetary policy continued to be net free reserves and market interest rates. Because evidence was lacking that interest rates had much effect on private

sector investment, a new paradigm, the ‘availability of credit’ doctrine, was used to rationalize the transmission of policy actions to the real economy. It argued that banks rationed credit to marginal borrowers when restrictive policy led to rising interest rates or indebtedness at the discount window. With these adjustments the FOMC vigorously and unsuccessfully pursued goals of lowering inflation and combating unemployment in the turbulent decade of the 1950s. In that decade there were three business cycles, which were marked by successively rising peaks of interest rates, inflation, and unemployment. The reason for this failure was thought to be inflation-induced rising marginal rates of taxation, which were addressed by large tax cuts in the following decade.

As interest rates rose, the opportunity cost of holding excess reserves rose, which led to the reappearance of a federal funds market in which banks traded reserves. Because banks paid no interest on demand deposits, there was also rapid expansion of the market for commercial paper in which large firms with good credit ratings traded idle funds without the direct intervention of banks. Both markets had atrophied after the 1920s because of low interest rates, and served to change the relation between open-market operations and real economic activity. They were precursors of a wave of innovations that would have similar effects in the coming decade. These included large-denomination negotiable certificates of deposit, one-bank holding companies, offshore ‘shell’ branches, the Eurodollar market, and bank-related commercial paper.

Beginning in 1961, the Kennedy administration attempted to coordinate fiscal and monetary policy by proposing large tax cuts to encourage investment and economic expansion. A new problem was that the United States was experiencing large gold outflows as the world continued to recover from the world war. To cope with this new approach and problem, the FOMC was encouraged to abandon its bills-only policy and to attempt to twist the yield curve by buying long-term bonds and selling bills. As short-term rates rose the Board repeatedly raised the ceiling on interest rates that banks could pay on time and savings deposits. It was argued that lower long-term interest rates would encourage capital formation and that higher short rates would discourage foreign interests from converting dollars into gold, as they were entitled to under the Bretton Woods agreements. These efforts were not successful in discouraging gold outflows, but investment and the economy expanded strongly. In 1965 the Board introduced a Voluntary Foreign Credit Restraint programme, which discouraged banks from overseas lending that was not financing US exports. Nevertheless, gold continued to flow out and the requirement that Federal Reserve notes and reserves be backed by gold was cancelled in 1968. Large open-market purchases had been needed to offset gold losses.

Policy coordination between the Board and the new Johnson administration effectively ended in December 1965, when the Board approved an increase in the discount rate because of inflation arising from mobilizing for the Vietnamese War. Net free reserves had turned negative in 1965 and were increasingly so until late 1966. Short-term interest rates rose until October. Higher rates increased the cost of the mobilization and had devastating effects on residential construction and the savings and loan associations and mutual savings banks (hereafter thrifts) that financed it, because in September Congress passed legislation limiting interest rates that thrifts could pay on time and savings accounts. These limits meant thrifts would experience withdrawals of funds or ‘disintermediation’ because depositors switched funds to government securities, which had no limits. This policy transmission channel would soon disappear because Congress and the administration could not withstand the resulting political pressures. In 1968 the Federal National Mortgage Association was privatized and in 1970 the Federal Home Loan Mortgage Corporation was created. Both bypassed depository institutions by securitizing mortgage

loans. Banks also responded to Board policies and restrictions on innovations by opening overseas offices that were not subject to them. A ten per cent income tax surcharge in 1967 was insufficient to stop inflation, and short-term interest rates rose to new highs in January 1970, when Chairman Martin's term ended. Net free reserves averaged about a negative $1 billion between May 1969 and July 1970. A decrease in short-term interest rates followed the then largest-ever US bankruptcy of the Penn Central Transportation Company in June 1970, but led to large new capital outflows in 1971 that pressured the dollar. The FOMC responded by forcing short-term rates and net borrowed reserves up again.

Towards flexible exchange rates

The amplitude of changes in interest rates increased between 1965 and 1971, and the United States experienced a recession in 1970. As in the 1950s the Federal Reserve was unable simultaneously to achieve satisfactory unemployment, inflation, and exchange rate outcomes. Many of the Board's policy instruments, such as the discount rate, reserve requirement changes, and many regulations had effectively been disabled by innovations, so that only open-market operations were available to achieve multiple targets. For example, an increase in reserve requirements induced banks to resign from the system or to conduct more of their business overseas. One exception to this loss of powers was the 1970 amendments to the Bank Holding Company Act, which finally gave the Board regulatory authority over one-bank holding companies. In August 1971 the Nixon administration, with new Board Chairman Arthur F. Burns as an advisor, announced a 90-day freeze on prices and wages, suspension of gold sales, and several other major changes in the United States. The suspension of gold sales led to a floating exchange rate system, devaluation of the dollar, and sharp rises in dollar-denominated prices in international markets. The shift from a fixed to a floating exchange rate system is likely to have increased the potency of monetary policy, as was predicted by Mundell (1961). The FOMC responded to consequent high inflation by driving nominal short-term interest rates to very high levels in 1973 and 1974, which helped to induce a severe recession beginning in August 1973, but were inadequate because on average the real federal funds interest rate (calculated with the GDP deflator) was negative between the end of 1973 and 1978. Real estate and other durable goods prices rose relative to the GDP deflator, and the international value of the dollar fell. After the resignation of President Nixon in 1974, Congress required the Chairman to explain policy in semi-annual public hearings and report the FOMC's targets for two money stock measures: M1, a measure of transactions balances, and M2, a measure of liquid assets. Friedman and Schwartz (1963) had recommended using money as an indicator of monetary policy instead of interest rates or net free reserves.

Part of the explanation for the policy failure was continuing financial market innovation. Foreign banks operating in the United States grew rapidly and were unregulated until the 1978 International Banking Act, which placed them under Board supervision. The introductions of money market mutual funds (MMMFs) and negotiable order of withdrawal (NOW) accounts in 1972, the Chicago Board Options Exchange in 1973, and financial futures markets in 1975 again began changing the relation between financial and real markets. A more important change was the rapid expansion of repurchase agreements after 1970. In a repurchase agreement, a client's deposits are borrowed to finance a bank's or dealer's inventory of government securities, often only overnight. Large bank holdings of government securities often represented transactions balances of large corporations and state governments that could not easily be controlled.

The real federal funds rate turned distinctly positive in the third quarter of 1979 when Paul A. Volcker became chairman. In early October he announced that the FOMC would no longer limit fluctuations in short-term interest rates and would use open-market operations to control bank reserves. This was a major policy change from practices dating from the 1951 accord. Further, he imposed eight per cent marginal reserve requirements on non-deposit liabilities, that is, Eurodollar borrowing, federal funds purchased from non-member banks, and funds acquired through repurchase agreements. These vigorous actions together with large income tax cuts by the Reagan administration between 1981 and 1983 drove real short-term interest rates to levels not seen since the early 1930s and caused MMMFs to grow rapidly. In only two quarters between 1979 and 1986 was the average real federal funds less than five per cent. These high rates caused the trade-weighted value of the US dollar to appreciate by 87 per cent between July 1980 and February 1985, which savaged US exports and attracted imports with adverse consequences for US manufacturing.

Financial deregulation

The landmark Depository Institutions Deregulation and Monetary Control Act was signed by President Carter at the end of March 1980. It radically changed the Federal Reserve System by eliminating the significance of membership in the system. After an eight-year phase-in period, all depository institutions would be subject to uniform reserve requirements on demand and time deposits, although the requirement on the first $25 million of transactions deposits was less than that on other transactions deposits. The Board could vary reserve requirements. All depository institutions had access to reserve bank discount windows. This strengthened the system because banks could no longer threaten to leave it in order to get the lower requirements that many states imposed. Further, Federal Reserve banks were required to charge banks for the cost of services they provided. Before this act they had been giving away services as an inducement for banks to stay in the system. This pricing requirement in turn forced depository institutions to begin to charge their clients for services, which changed the way banking services were used. The act mandated that interest rate ceilings on time and savings accounts be eliminated after six years, increased deposit insurance, and had other important provisions that are beyond the scope of this discussion.

In late 1980 the Board announced that transfers from overseas branches to the United States could be treated as collected funds on the day they were transferred. Before then, transfers in a day were not ‘good funds’ until the following day. The expansionary effects of this change, rapidly growing repurchase agreements, and other innovations are evident in demand deposit turnover statistics that the Board reported from 1919 until August 1996. Turnover is the annualized value of all withdrawals from deposit accounts divided by aggregate deposit balances.

High interest rates were savaging thrift institutions, which had negative gaps (more fixed-rate assets than fixed-rate liabilities on most future dates), and allowed MMMFs to expand rapidly. Congress intervened in September 1982 by passing the Garn–St Germain Act, which provided temporary emergency assistance and among other changes introduced money market deposit accounts and super NOW accounts, which paid market interest rates. MMMF growth was slowed by this act, but the weakening condition of banks and thrift institutions would result in large numbers of failures as the decade wore on. Large banks also experienced large losses because the appreciating dollar had resulted in failures of sovereign states, especially in Latin America, to meet their loan obligations. Chairman Volcker was

heavily involved in negotiating solutions for these defaults.

The restrictive monetary policy resulted in the deepest recession since the Depression; the unemployment rate was 10.8 per cent at the end of 1982. At the end of Volcker's term in August 1987 the unemployment rate had fallen to six per cent and the consumer inflation rate was less than two per cent. Real interest rates had fallen from 10.5 per cent in mid-1981 to four per cent, and the trade-weighted value of the dollar fell correspondingly. Volcker's February 1987 statement of monetary policy objectives to the Congress reported that M1 was not a reliable indicator of monetary policy and would be de-emphasized.

While his successor, Alan Greenspan, inherited a much improved economy, many problems remained from a rising wave of bank failures and the collapse of thrift institutions. Real estate markets were especially disorderly when the thrift crisis was resolved beginning in 1989 and were further distorted by provisions in the Tax Reform Act of 1986, which disallowed many interest tax deductions. After 1990 interest on home loans was effectively the only deductible interest on individual income tax returns. In addition, a collapse of stock prices in October 1987, strong foreign demand for US currency associated with the collapse of the Soviet Union, and a recession at the end of 1990 presented further challenges.

The FOMC responded to these challenges by varying the real federal funds rate, defined using the contemporaneous GDP price deflator inflation rate. This rate fell sharply for two quarters after the stock market crash, rose before falling for two quarters after a second stock market dip in October 1989, and then began to fall in the fourth quarter of 1990. In July 1993 testimony before Congress, Greenspan disclosed that the FOMC was downgrading M2 as an indicator of monetary policy and, as could have been surmised from its actions, that an important guidepost was now real interest rates. The real federal funds rate averaged less than one per cent in 1993. In early 1995 it had risen to four per cent and held that value as an average until the collapse of a large hedge fund in September 1998. After the fallout from the hedge fund collapse had been resolved, the real federal funds rate was restored to an average of about four per cent in 2000. When a new recession appeared in 2001 together with a sustained large collapse in stock market prices, the real federal funds rate was lowered to near zero in the fourth quarter; the rate had averaged zero for 13 consecutive quarters as of March 2005.

Between December 1990 and April 1992 reserve requirements on time and demand deposits were reduced, which helped banks to increase net income. In January 1994 ‘retail sweep programmes’ were introduced. In these programmes, a bank shifts funds from a depositor's transactions account to a synthetic time deposit account in the depositor's name in order avoid reserve requirements, usually without the depositor's knowledge. The Board does not measure the amount of funds swept, except at the time the programme was established. The Board estimated that as of August 1997 required reserves fell by one-third because of these programmes.

In November 1999 President Clinton signed the Financial Services Modernization (Gramm–Leach–Bliley) Act, which reversed the 1933 Glass–Steagall Act's ban on combining commercial and investment banking. The ban had been eroding since 1987, when some large bank holding companies were authorized by the Board to establish subsidiaries that could underwrite state and local government revenue bonds. The new act authorized the establishment of financial holding companies, which were to be regulated by the Board and could engage in an approved list of activities that included commercial banking, insurance, securities underwriting, merchant banking, and complementary financial undertakings. In 2003 there were more than 600 financial holding companies, which resemble the universal banks that exist in other countries.

In December 2002 the Federal Reserve discarded the discount rate as a policy instrument by replacing it with an interest rate on primary credit extended by the discount window that is one per cent above the FOMC target federal funds rate. Primary credits are collateralized loans to banks in sound financial condition.

As the foregoing dramatic institutional changes suggest, the Federal Reserve System is a work in progress. Its set of policy instruments and its dimensions have radically changed. Because of offshore banking facilities and retail sweep accounts, reserve requirement changes are no longer an effective policy instrument. As noted in the preceding paragraph, the discount rate has been discarded as an instrument; it is simply a penalty rate that is related to a bank rate, as is often the practice in other countries. Regulations on the interest rates banks pay on time and savings deposits have been discarded. Open-market operations are almost the sole policy instrument that can be used to achieve the Board's target nominal and real federal funds interest rates. While the FOMC has been able to control the overnight federal funds rate, the linkage between it and real economic activity is changing. First, the combined holdings of US government securities by foreign central banks have recently exceeded those of Federal Reserve banks. Foreign central bank holdings are partly a result of their efforts to manipulate exchange rates; their holdings are likely to change when FOMC policies change. Second, repurchase agreements and offshore transactions vary considerably over time and their volumes appear to be sensitive to US economic activity. Third, the outstanding stock of securitized mortgage and other debt has been growing rapidly; such debt is a close substitute for US government debt and its amount has real economic effects. Fourth, because of decreasing required reserves and growing offshore holdings of US currency, 89 per cent of Federal Reserve liabilities were in the form of Federal Reserve notes in December 2003; the corresponding share was 34 per cent in 1941, 57 per cent in 1970, and 79 per cent in 1989. In part, the Federal Reserve recently has become an institution for collecting seigniorage from the rest of the world. Finally, over the decade ending in 2003, the share of all credit market assets held by depository institutions in the Federal Reserve's flow of funds accounts fell. In the context of the most recent 13 quarters of a zero real federal funds interest rate, more changes could be expected.

See Also

- Great Depression
- monetary and fiscal policy overview

Bibliography

Bagehot, W. 1873. *Lombard Street: A Description of the Money Market*. Homewood, IL: Richard D. Irwin, 1962.

Chandler, L. 1958. *Benjamin Strong: Central Banker*. Washington, DC: The Brookings Institution.

Chandler, L. 1970. *America's Greatest Depression 1929–1941*. New York: Harper and Row.

Crabbe, L. 1989. The international gold standard and US monetary policy from World War I to the New

Deal. *Federal Reserve Bulletin* 75, 423–40.

Dykes, S. 1989. The establishment and evolution of the Federal Reserve Board: 1913–23. *Federal Reserve Bulletin* 75, 227–43.

Friedman, M. and Schwartz, A. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press.

Goldenweiser, E. 1951. *American Monetary Policy: A Research Study for the Committee for Economic Development*. New York: McGraw-Hill.

Greider, W. 1987. *Secrets of the Temple: How the Federal Reserve Runs the Country*. New York: Simon and Schuster.

Meltzer, A. 2003. *A History of theFederal Reserve, Volume I (1913–1951)*. Chicago: University of Chicago Press.

Mundell, R. 1961. Flexible exchange rates and employment policy. *Canadian Journal of Economics* 27, 509–17.

Scott, I., Jr. 1957. The availability of credit doctrine: theoretical underpinnings. *Review of Economic Studies* 25(4), 41–8.

Warburg, P. 1930. *The Federal Reserve System: Its Origin and Growth*, vol. 1. New York: Macmillan.

How to cite this article

Hester, Donald D. "Federal Reserve System." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000045> doi:10.1057/9780230226203.0556

The New Palgrave Dictionary of Economics Online

Fel'dman, Grigorii Alexandrovich (1884 – 1958)

Michael Ellman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Fel’dman was one of the founders of the theory of economic growth, the economics of planning and development economics. His contributions were made in the USSR in the late 1920s. He developed a two-sector growth model and showed how different growth rates implied different economic structures. He derived two theorems. He is regarded as the father of the ‘heavy industry first’ strategy of economic development. A brilliant pioneer, Fel’dman's work was cut short by the Stalinists. Later analysis and international experience revealed a number of limitations of a narrowly Fel’dmanite approach to economic policy.

Keywords

development economics; Fel’dman, G.; Fel’dman's two theorems; Gosplan; growth, models of; Kalecki, M.; planning; Soviet growth record

Article

Fel’dman was one of the founders of the theory of economic growth under socialism, the economics of planning and development economics. An electrical engineer by profession, he worked in Gosplan from February 1923 to January 1931. It was in this period that his contribution to economics was made. At first he was in the department analysing and forecasting developments in the world economy (he concentrated on Germany and the USA). His first work on the theory of growth was a comparative study of the structure and dynamics of the US economy in 1850–1925 with projections of the Soviet economy between 1926/27 and 1940/41. His most important work (‘On the theory of the rates of growth of the national income’) was a report to Gosplan's committee for compiling a long-term plan for the development of the national economy of the USSR. It was published in two parts in Gosplan's journal in 1928. A year later Fel’dman published a paper (1929c) which provides a more popular presentation of how to utilize his ideas to calculate long-term plans. The ideas of Fel’dman formed the methodological basis for the preliminary draft of a long-term plan worked out by the committee, then headed by N.A. Kovalevskii. This draft was discussed at meetings of Gosplan's economic research institute in February and March 1930. Apart from this serious discussion, during 1930 Fel’dman came under public attack for his ideas. His reliance on mathematics and his lack of fanaticism did not fit in well with the political fervour of 1930. The concrete numerical work of Fel’dman and Kovalevskii in 1928–30 was much too optimistic. It treated as feasible entirely unrealizable goals. The attempt to realize them had disastrous effects on the economy. Unfortunately, the political situation in the USSR prevented Fel’dman from publishing anything on economics after 1930. Even when, in 1933, he reverted from the sensitive subject of socialist industrialization to the problems of capitalist growth, his book was not published.

As far as growth theory is concerned, Fel’dman's work was much in advance of contemporary Western work. He developed a two-sector growth model and showed how different growth rates implied different economic structures. He derived two important results, one about the ratios of the capital stocks in the two sectors, the other about the allocation of investment between the two sectors. The first result is that a high rate of growth requires that a high proportion of the capital stock be in the producer-goods sector. This is illustrated in Figure 1. Fel’dman's second theorem is that, along a steady growth path, investment should be allocated between the sectors in the same proportion as the capital stock. For example, suppose that a 20 per cent rate of growth requires a K_c / K_p of 3.7. Then, to maintain growth at 20 per cent p.a. requires that 3.7/4.7 of annual investment goes to the consumer-goods industries and 1.0/4.7 of annual investment goes to the producer goods industries.

Figure 1

Fel’dman's first theorem. *Notes*: K_c is the capital stock in the consumer goods industry, K_p is the capital stock in the producer goods industry.

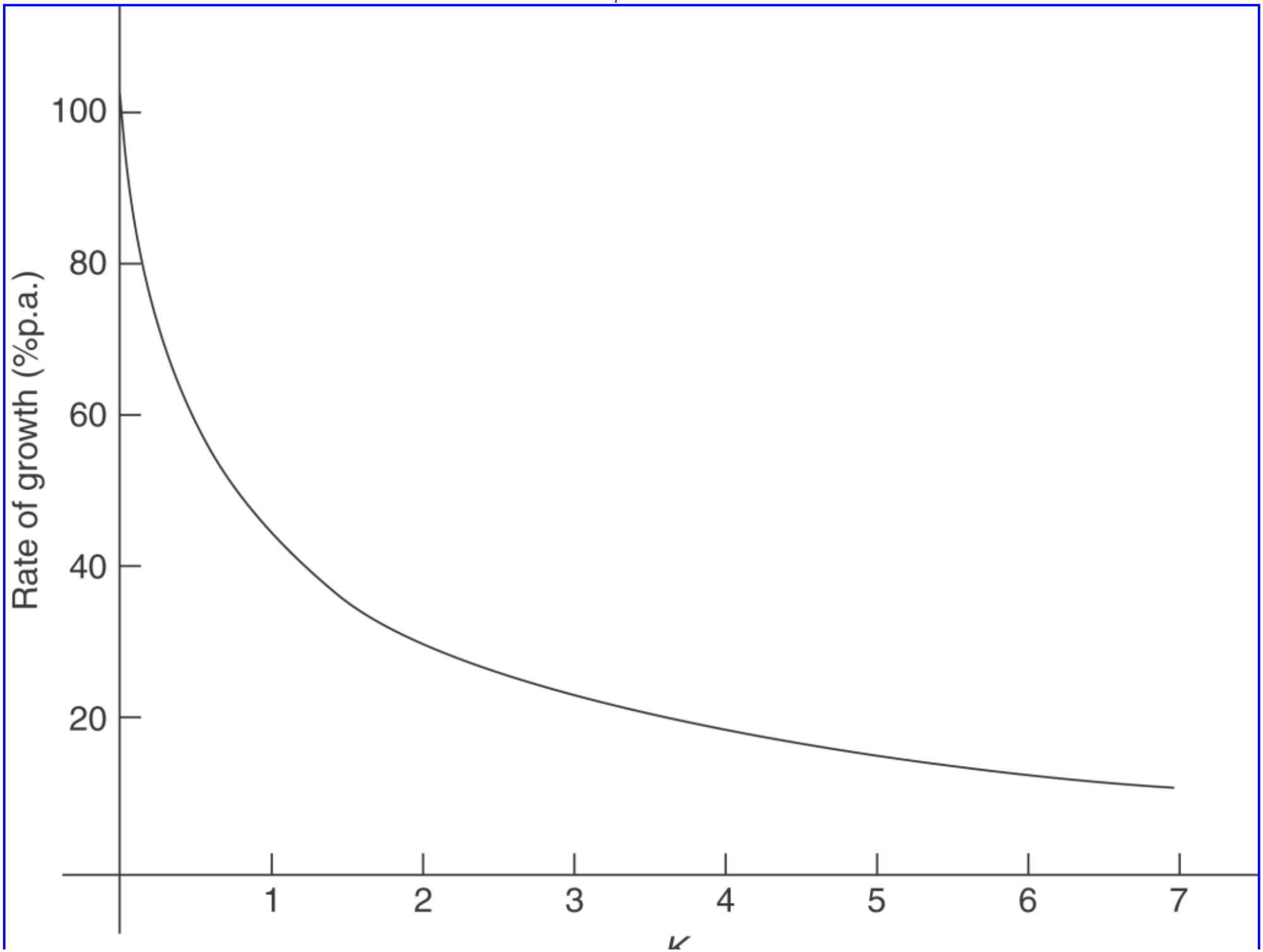

$$\frac{K_c}{K_p}$$

The interrelationship between the two theorems is shown in Table 1, in which Fel'dman explained how any desired growth rate, given the capital–output ratio, determined both the necessary sectoral composition of the capital stock and the sectoral allocation of investment.

Fel'dman's two theorems

$\frac{K_p}{K_c}$	$\frac{dy}{dt}$ (in % p.a.) (when K/Y=2.1)	$\frac{\Delta K_p}{\Delta K_c + \Delta K_p}$
0.106	4.6	0.096
0.2	8.1	0.167
0.5	16.2	0.333
1.0	24.3	0.500

Given the capital–output ratio, the higher the K_p / K_c ratio, that is, the greater the proportion of the capital stock in the producer goods sector, and correspondingly the higher the $\Delta K_p / \Delta K_c + \Delta K_p$) ratio, that is, the greater the proportion of new investment in the producer-goods sector, the higher the rate of growth. With a capital–output ratio of 2.1, to raise the growth rate from 16.2 to 24.3 per cent requires raising the proportion of the capital stock in the producer-goods sector from a third to a half, and the share of investment in the producer-goods sector from a third to a half.

The conclusion Fel'dman drew from his model was that the main tasks of the planners were to regulate the capital–output ratios in the two sectors and the ratio of the capital stock in the producer-goods sector to that in the consumer-goods sector. For the former task, Fel'dman recommended rationalization and multi-shift working; for the latter, investment in the producer-goods sector.

As far as the economics of planning is concerned, the main lesson to be learned from the Fel'dman model is that the capacity of the capital-goods industry is one of the constraints limiting the rate of growth of an economy. There may well be other constraints, such as foreign exchange, urban real wages or the marketed output of agriculture. (Indeed, it is possible that one or more of these are binding constraints and that the limited capacity of the producer-goods sector is a non-binding constraint.) Economic planning is largely concerned with the removal of constraints to rapid economic growth. Accordingly, a planned process of rapid growth may require that the planners stimulate the rapid development of the producer-goods sector.

As far as development economics is concerned, Fel'dman is important because of the argument in his 1928 paper that 'an increase in the rate of growth of income demands industrialization, heavy industry, machine building, electrification …'. When first formulated, this conclusion struck many economists as counter-intuitive and paradoxical.

Fel'dman's work, as is natural for a pioneer, suffers from serious limitations. As far as the theory of economic growth under socialism is concerned, he was an important early contributor, but his work has to be complemented by Kalecki's (1969) emphasis on the limits of growth and Kornai's (1992, ch. 9) emphasis on the behavioural regularities actually generating the growth process. As for the economics of planning, his arguments have to be complemented by a proper understanding of the role of agriculture, foreign trade and personal consumption and of the danger of an over-accumulation crisis. In development economics, experience in the USSR in the 1930s, India in the 1950s and China in the Maoist period has shown the limitations of a narrowly Fel'dmanite approach.

A brilliant pioneer, Fel'dman's work was ended after only a few years by the Stalinists. In January 1931 Fel'dman was forced out of Gosplan. He seems to have been arrested in 1937 and only released – probably from the Gulag – in 1943, but even then was forbidden to return to Moscow. He was only allowed to return to Moscow in 1953, by which time he was seriously ill.

See Also

- development economics
- Kalecki, Michal
- Soviet growth record

Selected works

1927. Soobrazheniya o strukture i dinamike narodnogo khozyaistvo SSha s 1850 po 1925g i SSSR s 1926/27 po 1940/41 gg [Reflections on the structure and dynamics of the national economy of the USA from 1850 to 1925 and of the USSR from 1926/27 to 1940/41]. *Planovoe khozyaistvo* No. 7. Also published as a booklet.

1928. K teorii tempov narodnogo khozyaistva [On the theory of the rates of growth of the national income]. *Planovoe khozyaistvo* Nos. 11 and 12. English translation in *Foundations of Soviet Strategy for Economic Growth*, ed. N. Spulber. Bloomington: Indiana University Press, 1964.

1929a. SSSR i mirovoe khozyaistvo na rubezhe vtorogo goda pyatletki [The USSR and the world economy on the eve of the second year of the five year plan]. *Na planovom fronte* No. 2.

1929b. O limitakh industrializatsii [On the limits of industrialization]. *Planovoe khozyaistvo* No. 2.

1929c. Analiticheskii metod postroeniya perspektivnykh planov [An analytical method for constructing perspective plans]. *Planovoe khozyaistvo* No. 12.

1930. Problemy electrifikatsii na novom etape [Problems of electrification at a new stage]. In *Na novom etape sotsialisticheskogo stroitel'stva*, vol. 1. Moscow: Gos. planovoe khozyaistvennoe izd-vo.

Bibliography

Soviet evaluations

Planovoe khozyaistvo. 1930. Report of the discussion in Gosplan's economic research institute of February and March 1930. No. 3, 117–211.

Vainshtein, A. and Khanin, G. 1968. Pamyati vydayushchegocya sovetskogo ekonomista-matematika G.A. Fel’dmana [In memory of the outstanding Soviet mathematical economist G.A. Fel’dman]. *Ekonomika i matematicheskie melody* 4(2).

English-language works

Allen, R. 2003. The development problem in the 1920s. Ch. 3 in *Farm to Factory*. Princeton: Princeton University Press.

Chng, M. 1980. Dobb and the Marx-Fel’dman model. *Cambridge Journal of Economics* 4, 393–400.

Dobb, M. 1967. The question of ‘Investment priority for heavy industry’. Ch. 4 of *Papers on Capitalism, Development and Planning*. London: Routledge.

Domar, E. 1957. A Soviet model of growth. In *Essays in the Theory of Economic Growth*. New York: Oxford University Press.

Erlich, A. 1978. Dobb and the Marx-Fel’dman model: a problem in Soviet economic strategy. *Cambridge Journal of Economics* 2, 203–14.

Kalecki, M. 1969. *Introduction to the Theory of Growth in a Socialist Economy*. Oxford: Blackwell.

Kornai, J. 1992. *The Socialist System: The Political Economy of Communism*. Oxford: Oxford University Press.

Mahalanobis, P. 1953. Some observations on the process of growth of national income. *Sankhya* 12, 307–12.

Mahalanobis, P. 1955. The approach of operational research to planning in India. *Sankhya* 16, 3–130.

Sen, A. and Raj, K. 1961. Alternative patterns of growth under conditions of stagnant export earnings. *Oxford Economic Papers* 13, 43–52.

Tinbergen, J. and Bos, H. 1962. *Mathematical Models of Economic Growth*. New York: McGraw.

How to cite this article

Ellman, Michael. "Fel'dman, Grigorii Alexandrovich (1884–1958)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000049> doi:10.1057/9780230226203.0557

The New Palgrave Dictionary of Economics Online

feminist economics

Julie A. Nelson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Feminist economics is a field that includes both studies of gender roles in the economy from a liberatory perspective and critical work directed at biases in the economics discipline. It challenges economic analyses that treat women as invisible, or that serve to reinforce situations oppressive to women, and develops innovative research designed to overcome these failings. Feminist economics points out how subjective biases concerning acceptable topics and methods have compromised the reliability of economics research. Topics addressed include the economics of households, labour markets, care, development, the macroeconomy, national budgets, and the history, philosophy, methodology, and teaching of economics.

Keywords

Sen, A.; American Economic Association; bargaining models; caring work; children; data quality; development; divorce; ecology; economic man; equal rights movement; feminist economics; fertility; gender; gender equity; gender roles; health care; household production; household labour; institutional economics; International Association for Feminist Economics; innate differences; intrahousehold distribution; labour market discrimination; marriage; masculinist bias; methodology of economics; occupational segregation; opportunity cost method; and household labour; Post Keynesian economics; postmodernism; race; radical economics; replacement cost method; social economics; structural adjustment; subcontracting; trade liberalization; unpaid work; women's economic status

Article

Feminist economics is a field that includes both studies of gender roles in the economy from a liberatory perspective and critical work directed at biases in the content and methodology of the economics discipline. It challenges economic analyses that treat women as invisible or that serve to reinforce situations oppressive to women, and develops innovative research designed to overcome these failings.

Feminist economics points out how subjective biases concerning acceptable topics and methods have compromised the reliability and objectivity of economics research, and explores more adequate alternatives.

The origins of feminist economics

Feminist economics in its contemporary form began in the 1970s in response to the prevailing pattern of labour market and household studies. Up until the 1960s, women and women's traditional activities had been subsumed into the ‘black box’ of the household within neoclassical economics. Neoclassical theory had been defined as the study of choices made in markets by rational, autonomous actors. A household was generally understood to be represented by its male ‘head’, whose preferences, it was assumed, determined household labour supply and consumption decisions. The household was assumed to enjoy a single utility level, and activities within the household were classified as ‘leisure’. Studies of paid labour generally focused on men only, and household production was (and is still) excluded from national accounts. Women, women's traditional activities, and the well-being of women and children were invisible.

During the 1960s, issues of labour market discrimination by race and sex began to be debated. The idea that household activities might include unpaid work as well as leisure also gained ground. The New Home Economics school sought to extend rational choice theory to intra-household decisions. Often, however, work by economists on these issues simply defended traditional sex roles in the family, women's segregation into a narrow range of paid occupations, and women's lesser earnings in the paid labour market. In general, neoclassical economists of the time argued that the prevailing patterns resulted from rational choices, with variations between men and women due only to presumably innate differences between men and women in tastes and abilities, often expressed in different choices about human capital formation. As well, circular reasoning was used: women's lesser market earnings were used to explain their specialization in household work, and women's household responsibilities were used to justify their lesser market earnings. While these works recognized women's existence, they were not feminist in that they served to rationalize rather than explore and question women's assignment to second-class status and financial dependency.

A key distinction feminist economists make is between *sex*, understood as the biological difference between males and females, and *gender*, the social beliefs that society constructs on the basis of sex. While traditional economists saw household and labour market outcomes as reflecting only sex differences, feminist economists raised the question of how much these outcomes might, instead, reflect misleading stereotypes and rigid social constraints. Some works called into question, for example, the ideas that specialization in household work would be an optimizing choice for a woman (given rising divorce rates) or that it would necessarily yield greater household well-being than other, more egalitarian, arrangements (Ferber and Birnbaum, 1977; all references given in this article are examples from larger literatures). Others emphasized the role of discrimination in limiting women's labour market opportunities (Bergmann, 1974) or the interplay of household and workplace power relations (Hartmann, 1976).

In actuality, as the equal rights movements of the 1960s and 1970s loosened many of the legal restrictions and social norms that had artificially narrowed women's educational and job choices in a

number of countries, women moved increasingly into the labour market and into formerly all-male occupations. Surveys of women's economic history, economic status, and progress towards gaining economic equality have since been undertaken for many countries and regions, along with surveys of policies related to gender equity. Recognition of the importance of social beliefs and power structures in creating gendered economic outcomes has remained a hallmark of feminist economics.

The critique of mainstream economics

While feminist were originally dissatisfied with mainstream economic scholarship because it neglected and distorted women's experiences, by the late 1980s feminists were also advancing a more thoroughgoing critique. Many feminist economists were finding that traditional formal choice-theoretic modelling and a narrow focus on mathematical and econometric methods were a Procrustean bed when it came to analysing phenomena characterized by connection to others, tradition, and relations of domination. Feminists began to raise questions about the mainstream definition of economics, its central image of ‘economic man’ and the exclusive use of a particular set of methodological tools.

Essays on this theme were brought together in a 1993 volume, *Beyond Economic Man: Feminist Theory and Economics* (Ferber and Nelson, 1993). In this volume it was suggested that economics be defined by a concern with the provisioning of life in all spheres where this occurs rather than only in markets. Investigations were undertaken into how a particular set of professional values, emphasizing culturally masculine-associated factors such as autonomy, separation, and abstraction, had come to take precedence over culturally feminine-associated factors such as interdependence, connection, and concreteness. The contributors argued that, rather than taking the former as a sign of ‘rigour’ in the discipline, the truncation of methods created by masculinist bias had weakened the discipline's ability to explain real-world phenomena. Questions were raised about mainstream economics not because it was too objective but because it was not objective enough.

A conference held in Amsterdam in 1993 further developed this theme, and contributed innovative discussions on economic methodology (Kuiper and Sap, 1995). While many feminist economists continue to make use of traditional mainstream tools, on the whole the field has come to be characterized by the inclusion of a broader range of concepts and methods. Theories of human behaviour that include a balance between individuality and relationship, autonomy and dependence, and reason and emotion are being developed (Ferber and Nelson, 1993; 2003). The use of historical studies, case studies, interviews and other qualitative data, as well as greater attention to issues such as data quality and replication in quantitative work, are being explored (Bergmann, 1989; Nelson, 1995). Feminist economists tend to find that such serious efforts to create and promote more adequate forms of economic practice lead to new insights across the board, whether or not the topic being studied is explicitly gender-related.

The formation of a field

With the publication of a number of books and articles, and gatherings at early conferences, feminist economics coalesced into an organized field in the early 1990s. The International Association for Feminist Economics was formed in 1992, and its journal, *Feminist Economics*, commenced publication

a few years later (Strassmann, 1995). The field was first described in a journal of the American Economic Association in 1995 (Nelson, 1995), an encyclopedia of feminist economics was published in 1999 (Peterson and Lewis, 1999), and a review of developments during the first ten years of feminist economics was published in 2003 (Ferber and Nelson, 2003).

International and wide-ranging in scope, feminist economics now includes work on a number of subjects, including topics in microeconomics, macroeconomics, history, and philosophy.

Labour, households and care

True to its roots, feminist economics continues to develop analyses of gender roles in labour markets and households. Many studies of women's paid labour supply, labour market discrimination, and the origins of occupational segregation have been undertaken. Some feminists make use of mainstream theories or econometric models to examine the wage gap between men and women and its possible explanations. Other feminist economists raise questions about the ability of such tools, used alone, to shed light on the underlying causes of inequality, and encourage increased investigation into the social, political and institutional structures of gender and labour markets (Bergmann, 1989; Rubery, 1998; Figart, Mutari and Power, 2002).

Studies of unpaid work within households have sought to obtain quantitative measures of this labour and to increase the attention paid to unpaid work in the design of policies (Waring, 1988; Ironmonger, 1996). The issue of valuing this work remains controversial among feminists. Some feminists endorse the use of replacement cost or opportunity cost methods of assigning dollar values to unpaid household labour. Others argue that these methods lead to understatement because the wages used in such imputations have been kept artificially low by discrimination. Still others believe that this issue serves to draw attention away from women's lack of access to real money and power.

Issues of intra-household distribution and decision-making have been investigated by many feminist economists. The dramatic effect of skewed intra-household distribution by sex in countries such as China, India and Pakistan has been brought to public attention (Sen, 1990). Bargaining models (McElroy and Horney, 1981) have been developed as one way of bringing women's agency within households to the fore. Issues concerning marriage, divorce, fertility and the well-being of children have been investigated from feminist perspectives. A number of feminist economists go beyond choice-theory-based bargaining models to examine legal, social, and psychological issues related to intra-household decision-making and well-being (Sen, 1984, ch. 16; Agarwal, 1997; Wheelock, Oughton and Baines, 2003).

Much of women's traditional work in sex-segregated occupations (such as nursing and childcare) and within households can be described as ‘caring work’. Caring work presents a challenge to mainstream economics since the traditional image of ‘economic man’ is of an autonomous, self-interested individual who neither requires care nor has any inclination to provide it. The conceptual and empirical study of work with dependency, emotional or other-regarding components has recently become a field of active investigation for many feminist economists (Folbre, 1994; Himmelweit, 1999; Folbre and Nelson, 2000; Bettio and Plantenga, 2004).

Feminist economists have developed critiques of theories and policies that assume that economic agents

are unencumbered prime-age workers, and that delve into the economic problems of elderly women, parents of young children, and lone mothers who are faced with simultaneous responsibilities for income generation and family care (MacDonald, 1998; Albelda, Himmelweit, and Humphries, 2004).

Development, macroeconomics and national budgets

Feminist economists have also made innovations in the analysis of national and global economies. Studies of the effects of including unpaid production in GDP (Wagman and Folbre, 1996) and the analysis of government budgets according to their effects on gender equity (Budlender et al., 2002) have become well-developed fields.

Feminist economists have challenged the definition of economic development in terms of industrialization and GDP growth, drawing attention instead to issues of growth in human well-being and capabilities (Elson, 1991; Benería, 2003; Agarwal, Humphries, and Robeyns, 2003). Many have studied the changes in women's status that have come about during transitions from socialism and during other forms of macroeconomic restructuring (Aslanbeigui, Pressman and Summerfield, 1994).

The effects of macroeconomic policies of structural adjustment and the liberalization of global trade and finance have been looked at from a feminist point of view (Çagatay, Elson and Grown, 1995; Grown, Elson and Çagatay, 2000). For example, programmes that prescribe macroeconomic belt-tightening through cutbacks in health care often have their most immediate impact on women, as women are expected take on, unpaid, the work of providing services no longer provided by governments. Men and women may also be affected differently, depending on the degree to which they work in subsistence or traded sectors. Women's employment in subcontracting firms has also received considerable attention (Kabeer, 2000; Balakrishnan, 2002).

History, philosophy and teaching

As well as investigating the history of women's economic activities (Humphries, 1990) and the history of economic thought in relation to women (Folbre, 1991; Pujol, 1992), feminist economists have looked at the history of women and feminists within the economics discipline itself (Dimand, Dimand and Forget, 1995). The most recent national studies indicate that women are still under-represented in the top ranks of academic economics (Booth, Burton and Mumford, 2000), receiving tenure less frequently than men even when factors such as publications and family are controlled for (Ginther and Kahn, 2004). Sexual harassment, sex discrimination, and inhospitable environments are among the barriers yet to be overcome in some departments and universities (Ginther and Kahn, 2004).

Feminist economists have also engaged in philosophical discussion concerning the epistemological and methodological foundations of economics in dialogue with postmodernist, post-colonialist, critical realist and other perspectives (Barker and Kuiper, 2003). Feminists have also explored comparisons of aims and methods with various heterodox schools of economics including institutionalist economics (Waller and Jennings, 1990), social economics (Emami, 1993), radical economics (Matthaei, 1996) and Post Keynesian economics (Danby, 2004).

Regarding the teaching of economics, feminist economists have investigated how the content of economics courses can be made less biased concerning women (Feiner, 2004), how courses can be enriched by feminist re-evaluation of theories and methods, and how pedagogy can be adapted to better reach students with diverse backgrounds and learning styles (Shackelford, 1992; Aerni and McGoldrick, 1999).

Feminism and other concerns

Feminist economists have also analysed how such factors as race and caste (Brewer, Conrad and King, 2002) and sexual preference (Badgett, 2001), in interaction with gender, affect economic outcomes. Feminist economists' scepticism about the adequacy of the image of 'economic man' has also stimulated new thinking in areas other than gender relations. The analysis of relations of power and of care, first generated by study of women's work and family relations, has been extended to the subject of interpersonal relations among people economy-wide (Nelson, 2005). Feminist explorations into ecological economics examine how natural processes, like women, have been treated as invisible and freely exploitable in traditional economic thought (Perkins, 1997).

See Also

- economic man
- gender roles and division of labour
- household production and public goods
- intrahousehold welfare
- methodology of economics
- Sen, Amartya
- women's work and wages

Bibliography

Aerni, A. and McGoldrick, K., eds. 1999. *Valuing Us All: Towards Feminist Pedagogy in Economics*. Ann Arbor: University of Michigan Press.

Agarwal, B. 1997. 'Bargaining' and gender relations: within and beyond the household. *Feminist Economics* 3(1), 1–51.

Agarwal, B., Humphries, J. and Robeyns, I. 2003. A special issue on Amartya Sen's work and ideas. *Feminist Economics* 9(2/3).

Albelda, R., Himmelweit, S. and Humphries, J. 2004. A special issue on lone mothers. *Feminist Economics* 10(2).

Aslanbeigui, N., Pressman, S. and Summerfield, G., eds. 1994. *Women in the Age of Economic Transformation*. London: Routledge.

Badgett, L. 2001. *Money, Myths and Change: The Economic Lives of Lesbians and Gay Men*. Chicago: University of Chicago Press.

Balakrishnan, R., ed. 2002. *The Hidden Assembly Line: Gender Dynamics of Subcontracted Work in a Global Economy*. Bloomfield: Kumarian Press.

Barker, D. and Kuiper, E., eds. 2003. *Toward a Feminist Philosophy of Economics*. London and New York: Routledge.

Benería, L. 2003. *Gender, Development, and Globalization*. London: Routledge.

Bergmann, B. 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* 1(2/3), 103–10.

Bergmann, B. 1989. Does the market for women's labor need fixing? *Journal of Economic Perspectives* 3 (1), 43–60.

Bettio, F. and Plantenga, J. 2004. Comparing care regimes in Europe. *Feminist Economics* 10(1), 85–113.

Booth, A., Burton, J. and Mumford, K. 2000. The position of women in UK academic economics. *Economic Journal* 110, 312–33.

Brewer, R., Conrad, C. and King, M., eds. 2002. A special issue on gender, color, caste and class. *Feminist Economics* 8 (2).

Budlender, D., Elson, D., Hewitt, G. and Mukhopadhyay, T. 2002. *Gender Budgets Make Cents*. London: Commonwealth Secretariat Publications.

Çagatay, N., Elson, D. and Grown, C., eds. 1995. Special issue on gender, adjustment and macroeconomics. *World Development* 23(11).

Danby, C. 2004. Toward a gendered Post Keynesianism. *Feminist Economics* 10(3), 55–75.

Dimand, M. A., Dimand, R. W. and Forget, E. L., eds. 1995. *Women of Value: Feminist Essays on the History of Women in Economics*. Aldershot: Edward Elgar.

Elson, D., ed. 1991. *Male Bias in the Development Process*. Manchester: Manchester University Press.

Emami, Z. 1993. Challenges facing social economics in the twenty-first century: A feminist perspective. *Review of Social Economy* 52, 416–25.

Feiner, S. 2004. There are none so blind… In *A Guide to What's Wrong with Economics*, ed. E. Fullbrook. London: Anthem Press.

Ferber, M. and Birnbaum, B. 1977. The ‘new home economics’: retrospects and prospects. *Journal of Consumer Research* 4, 19–28.

Ferber, M. and Nelson, J., eds. 1993. *Beyond Economic Man: Feminist Theory and Economics*. Chicago: University of Chicago Press.

Ferber, M. and Nelson, J., eds. 2003. *Feminist Economics Today: Beyond Economic Man*. Chicago: University of Chicago Press.

Figart, D., Mutari, E. and Power, M. 2002. *Living Wages, Equal Wages*. London: Routledge.

Folbre, N. 1991. The unproductive housewife: her evolution in nineteenth-century economic thought. *Signs* 16, 463–485.

Folbre, N. 1994. *Who Pays for the Kids?* London: Routledge.

Folbre, N. and Nelson, J. 2000. For love or money – or both? *Journal of Economic Perspectives* 14, 123–140.

Ginther, D. and Kahn, S. 2004. Women in economics: moving up or falling off the academic career ladder? *Journal of Economic Perspectives* 18(3), 193–214.

Grown, C., Elson, D. and Çagatay, N., eds. 2000. Special issue on growth, trade, finance and gender inequality. *World Development* 28(7).

Hartmann, H. 1976. Capitalism, patriarchy and job segregation by sex. *Signs* 1(3, part 2), 137–69.

Himmelweit, S. 1999. Caring labor. *Annals of the American Academy of Political and Social Science* 561, 27–38.

Humphries, J. 1990. Enclosures, common rights, and women. *Journal of Economic History* 50, 17–42.

Ironmonger, D. 1996. Counting outputs, capital inputs and caring labor. *Feminist Economics* 2(3), 37–64.

Kabeer, N. 2000. *The Power to Choose: Bangladeshi Women and Labour Market Conditions in London and Dhaka*. London: Verso.

Kuiper, E. and Sap, J., eds. 1995. *Out of the Margin: Feminist Perspectives on Economics*. London: Routledge.

MacDonald, M. 1998. Gender and social security policy. *Feminist Economics* 4(1), 1–25.

Matthaei, J. 1996. Why feminist, Marxist and anti-racist economists should be feminist-Marxist-anti-racist economists. *Feminist Economics* 2(1), 22–42.

McElroy, M. and Horney, M. 1981. Nash bargained household decisions. *International Economic Review* 22, 333–49.

Nelson, J. 1995. Feminism and economics. *Journal of Economic Perspectives* 9(2), 131–48.

Nelson, J. 2005. Interpersonal relations and economics. In *Economics and Social Interactions*, ed. B. Gui and R. Sugden. Cambridge: Cambridge University Press.

Perkins, E., ed. 1997. Women, ecology and economics. Special issue of *Ecological Economics* 20 (2).

Peterson, J. and Lewis, M. 1999. *The Elgar Companion to Feminist Economics*. Cheltenham: Edward Elgar.

Pujol, M. 1992. *Feminism and Anti-Feminism in Early Economic Thought*. Aldershot: Edward Elgar.

Rubery, J., ed. 1998. *Equal Pay in Europe?* London: Macmillan.

Sen, A. 1984. *Resources, Values, and Development*. Cambridge, MA: Harvard University Press, ch. 16.

Sen, A. 1990. More than 100 million women are missing. *The New York Review of Books*, 20 December.

Shackelford, J. 1992. Feminist pedagogy. *American Economic Review* 82, 570–76.

Strassmann, D. 1995. Editorial: creating a forum for feminist inquiry. *Feminist Economics* 1(1), 1–5.

Wagman, B. and Folbre, N. 1996. Household services and economic growth in the United States, 1870–1930. *Feminist Economics* 2, 43–66.

Waller, W. and Jennings, A. 1990. On the possibility of a feminist economics. *Journal of Economic*

Issues 24, 613–72.

Waring, M. 1988. *If Women Counted*. San Francisco: Harper & Row.

Wheelock, J., Oughton, E. and Baines, S. 2003. Getting by with a little help from your family: toward a policy-relevant model of the household. *Feminist Economics* 9(1), 19–45.

How to cite this article

Nelson, Julie A. "feminist economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000286> doi:10.1057/9780230226203.0558

The New Palgrave Dictionary of Economics Online

fertility in developed countries

Alicia Adsera
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

After completing the first demographic transition, developed countries experienced a fertility boom in the post-Second World War period. However, after the 1960s fertility rates fell dramatically and now, in 2007, stand below the replacement level of 2.1 births per woman in most of these countries. The entry of women into the workforce, economic development and changes in values and secularization are the causes of this demographic transformation.

Keywords

capital intensity; child care; demographic transition; family planning; fertility in developed countries; household production; infant mortality; labour market institutions; labour supply; second demographic transition; unemployment; wage differentials; welfare state; women's work and wages

Article

Decrease in fertility rates

Fertility rates in the developed world started to decline sharply at the end of the 19th century. In Europe fertility declined by about 40 per cent between 1870 and 1930, and the majority of Western countries experienced this transition before the Second World War (Lee, 2003). The Second World War was followed by a period of increases in fertility. After peaking during the ‘baby boom’ years of the late 1950s, the average total fertility rate in developed countries fell from an average of 2.8 births per woman in 1960 to 2.0 in 1975 and then to 1.6 in the late 1990s (and below 1.3 in southern Europe). The total fertility rate (TFR) estimates the number of children a woman would bear if she went through her childbearing years exposed to the current age-specific birth rates for women between the ages of 15–44 years. Table 1 presents the evolution of the total fertility rate since 1965 to 2004 in the most developed economies. In 1965, fertility rates were almost three children in many of these countries and even higher in Canada, Portugal, Iceland and Ireland. During the next decade, rates fell sharply in the richest countries, but they remained above replacement level in southern Europe, Ireland and Iceland. The transition to lower fertility has only occurred in southern Europe since the early 1980s but its speed and extent went beyond what previous countries had experienced. By the mid-1990s fertility rates in these countries were under 1.3, a threshold level used by some demographers to define ‘lowest-low fertility’ (Kohler, Billari and Ortega, 2002). With the exception of the United States, Iceland and Ireland, all advanced countries now have fertility rates well below the replacement rate of 2.1 (the fertility rate needed to sustain a steady level of population) though cross-national differences have remained significant.

Total fertility rate in developed countries 1965–2004

	1965	1975	1985	1995	2004
Australia	2.98	2.22	1.89	1.82	1.77
Austria	2.7	1.83	1.47	1.4	1.42

Belgium	2.62	1.74	1.51	1.56	1.64
Canada	3.15	1.85	1.61	1.62	1.53
Denmark	2.61	1.92	1.45	1.81	1.78
Finland	2.47	1.69	1.64	1.81	1.8
France	2.83	1.93	1.81	1.7	1.92
Germany	2.51	**1.45**	**1.28**	1.34	1.37
Greece	2.32	2.32	1.67	1.32	1.31
Iceland	3.71	2.65	1.93	**2.08**	2.03
Ireland	**4.03**	**3.4**	**2.5**	1.85	1.99
Italy	2.59	2.17	1.42	**1.17**	1.33
Japan	**2.14**	1.91	1.76	1.42	**1.29**
Netherlands	3.04	1.66	1.51	1.53	1.73
Norway	2.95	1.98	1.68	1.87	1.81
Portugal	3.15	2.63	1.72	1.4	1.4
Spain	2.97	2.79	1.64	1.18	1.33
Sweden	2.42	1.77	1.74	1.73	1.75
Switzerland	2.61	1.61	1.52	1.48	1.42
UK	2.86	1.81	1.79	1.71	1.77
USA	2.88	1.77	1.84	2.02	**2.05**
Average	2.84	2.05	1.68	1.61	1.64

Note: Maximum and minimum rates in bold.

Sources: National official statistics and United Nations Population Division, various years.

Synthetic indices such as total fertility rates may not provide a precise picture of fertility changes in periods when the younger cohorts of women shift the timing of their fertility to older ages. Both the age at first marriage and the age at first birth have been rising since the early 1980s across the developed world. Delayed maternity may artificially deflate total fertility rate since a larger proportion of births is bound to occur among older mothers over time, but this is not still reflected in the behaviour of women currently in their thirties. Adjustments for these ‘tempo effects’ suggest that, even though important for some countries, such as France or the Netherlands in the mid-1980s, they account for only part of the reduction in TFR (Bongaarts, 2001). Completed fertility by cohort provides an alternative and more accurate way to measure fertility changes. It computes the mean number of children born to women of a given generation at the end of their childbearing years. Recent data in completed fertility show for most countries a downward trend similar to that in TFR though with more moderate inter-country differences. Among women born in 1965, for example, whereas the Irish are projected to bear around 2.2 children, Spaniards, German and Italians are expected to bear only 1.5 children.

In any case, delayed childbearing itself is likely to imply lower completed fertility. Women who become mothers at a later age are expected to bear fewer children by the end of their fertile life because of both time and fecundity constraints (Kohler, Billari and Ortega, 2002). Still, the negative relationship between postponement and completed fertility seems to have weakened somewhat, possibly due to the improvement in reproductive techniques. Even if lifetime childlessness has risen steadily among women born between the 1940s and 1970s, particularly in German-speaking countries, demographers project it will stabilize around 15 to 20 per cent in these countries.

Quantity and quality of children

The basic microeconomic model of fertility (Willis, 1973; Becker, 1991) identifies a broad set of factors that influence fertility: household preferences over the number and quality of children, their labour supply decisions and their access to family planning. Each one of these factors has been relevant to the dramatic reduction in fertility rates across developed countries as income kept rising during the 20th century. In addition, with modernization, infant mortality decreased sharply. For parents interested in a certain number of surviving children, the increase in the likelihood of survival of any child born constituted an independent cause of the reduction in fertility.

The quantity–quality model developed by Becker and Lewis (1973) and Willis (1973) provided the first explanation of the observation that the number of children per family did not increase with income. In this model, each family maximizes a utility function depending on the quantity of children, the quality of children (expenditures on a child's well-being such as health or education) and consumption goods. Parents provide the same quality for all their children. The quality and quantity of children enter multiplicatively in the budget constraint of the household through the total expenditures on children. Overall expenditures on children tend to increase with income, which indicates that children are normal goods. An increase in the quantity of children raises the shadow price of the quality of children and vice versa. For example, an increase in the number of children raises the cost of providing more quality for each child because there are more children. This explains why the quantity and quality of children interact more closely than any other random pair of goods, even without assuming that both are close substitutes (Becker and Lewis, 1973). If the income elasticity of demand for quality of children is higher than that of quantity, rising income increases the optimal ratio of quality to quantity of children. This implies a rise in the relative cost of an additional child relative to quality and can lead both to higher quality per child and fewer children. The income effect on fertility may be offset by the substitution effect induced by the increase in the shadow price of an additional child. If a high average education in a society generates positive externalities that boost the returns to each individual's human capital, the quality–quantity trade-offs are strengthened as families invest more heavily in each of the children (Becker, Murphy and Tamura, 1990)

Labour market participation and fertility

Household production models in which both consumption and production decisions are jointly analysed provide a second major explanation for the decrease in the number of children per woman. These models spell out how labour supply decisions are related to choices in both the timing and the level of fertility.

As a result of economic development since the Industrial Revolution, capital intensity in production increased and, further, the emphasis in activities that require physical strength diminished with the gradual shift from manufacturing towards services. This technological transformation pushed upwards the demand for activities where women have a comparative advantage and increased the relative wage of women (Galor and Weil, 1996). Developed countries therefore experienced a massive entry of women into the labour market, with average female labour force participation rates climbing from 41 per cent in 1960 to 64 per cent by the late 1990s.

Childbearing is time-intensive relative to other activities and its associated opportunity cost can be measured by the potential wages of the mother. While increases in men's work mainly entail an income effect that increases the demand for children, increases in women's wages give rise to a combination of income and substitution effects as they result in an increase in the cost of a child relative to other goods (Mincer, 1963). Accordingly, women with high potential wages may restrict their fertility and trade off children for less time-demanding alternatives if the substitution effects are important (Becker, 1991). An alternative to this is the purchase of childcare services in the market. This may lessen the substitution effect and the net impact of higher wages may even turn positive (Ermisch, 1989). In Scandinavian countries, for example, where publicly provided childcare is abundant, work–family trade-offs are diminished. As women's wages have increased across the developed world, however, the cost of childcare outside the home has also risen because its provision is intensive in woman's work.

The original fertility models are static since all life-cycle choices are made at the beginning of the parent's lifetime without assuming any uncertainty. Later models emphasize the sequential nature of these decisions and incorporate stochastic shocks to the household – for example, contraceptive failure (Heckman and Willis, 1976). Moffit (1984) explores how, in addition to current wage losses, lower experience and skill depreciation from career interruption may result in permanent wage gaps between women with different childbearing patterns. (His model supports the prediction that women with strong preferences for children may self-select into occupations with low wage-growth prospects: Mincer and Polachek, 1974.) Hotz, Klerman and Willis (1997) offer a good review of dynamic models of fertility.

Family planning

The decrease in the cost of contraception is a third important factor that facilitates the family choices discussed in the previous models. Widespread access to the birth control pill since the late 1960s had two important effects on women's careers and fertility behaviour in developed nations (Goldin and Katz, 2002). First, it promoted women's career investment by reducing pregnancy risk while allowing women to remain sexually active. Second, it had an indirect impact through a social multiplier effect: the overall delay of marriage produced a thicker marriage market for career women and increased their likelihood of finding a suitable spouse later in life. Accordingly, in the United States from around 1970, more women entered professional schools and delayed marriage.

Differences across countries

As expected from the household production models, during the 1960s and the 1970s fertility was lower in the countries where women had entered the labour market more rapidly. Surprisingly, as female labour participation kept growing the cross-country negative relationship between fertility and labour force participation reversed by the mid-1980s. As shown

in Figure 1 for the mid-1990s, those countries with the lowest levels of female labour force participation – such as Spain, Italy and Greece – also portrayed the lowest fertility rates. Further, even if fertility differed substantially across countries at that time (Table 1), surveys indicated that the ideal family size was above replacement level and relatively homogenous. Hence, this positive correlation was probably related to the differential support to women from government policies as well as the flexibility and performance of their labour markets.

Figure 1

Female activity rate and total fertility rate in developed countries, 1996. *Sources*: OECD, *Employment Outlook*, and Council of Europe.

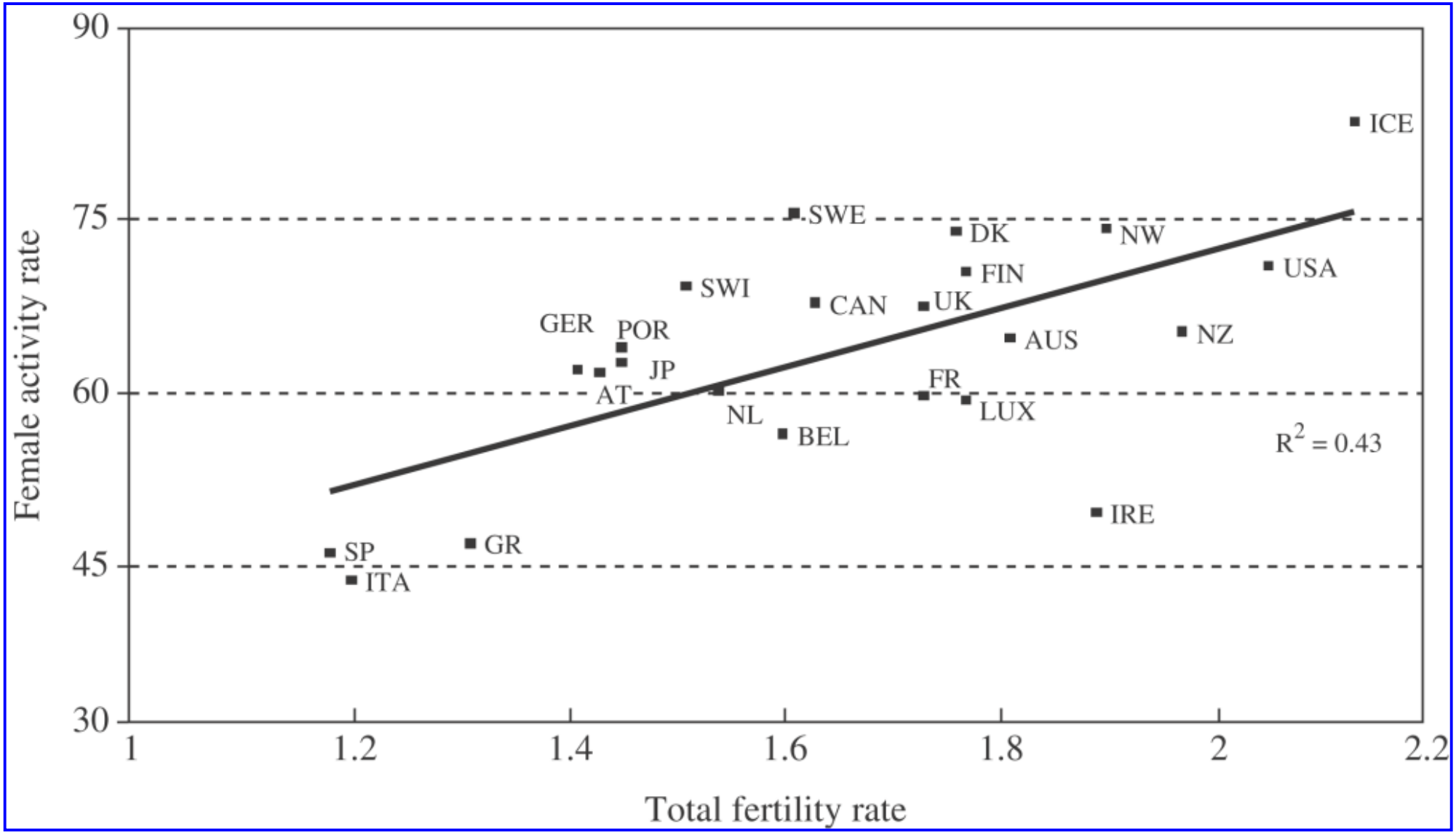

The rapid increase in persistent unemployment since the mid-1980s was contemporaneous with the sharpest fall of fertility rates and postponement of childbearing in many countries. European unemployment went up from less than three per cent before 1975 to about ten per cent in the 1990s. By 1990, around 50 per cent of those unemployed in the European Union had been out of work for more than 12 months. Within the standard microeconomic model of fertility the associated fall in current opportunity costs (in terms of forgone wages) makes unemployment spells good times for childbearing (Butz and Ward, 1979). Still, job loss impairs human capital accumulation and, with it, the future prospects of employment, particularly of young workers with low labour market experience. Among the employed, (temporary) withdrawals from the labour market associated with maternity

have a similar effect. High and long-lasting unemployment intensifies the relevance of the latter and negative income effects can reduce fertility, as happened during the Great Depression (Becker, 1991). Individuals may want to secure an adequate level of human capital (experience or education) before starting a family, and so the attractiveness of an early childbearing strategy declines. Since the 1980s, fertility postponement was more important in countries where joblessness was more prevalent and persistent – particularly among women – such as those in southern Europe (Adsera, 2005).

The extent of the negative impact of unemployment, however, is related to the labour regulation and types of contractual arrangements available in each labour market. The rapid feminization of the labour force in southern Europe, where traditionally there was low female participation, collided with rigid labour market institutions that favoured traditional full-time male employment and limited the availability of part-time positions for women (Adsera, 2005). In addition, the expansion of temporary contracts among young workers after partial labour reforms were passed in the late 1980s exacerbated those problems. By contrast, fertility rates are among the highest in countries with high female participation and either a flexible regulation and access to part-time employment (and low unemployment), such as the United States, the United Kingdom, or the Netherlands, or abundant public sector employment (mostly tenured jobs with features that make childbearing and participation compatible), as in the Nordic countries and France.

Changes in values and the ‘second demographic transition’

Changes in values as well as secularization have long been considered independent causes of recent demographic adjustments. The fall in period fertility has been coupled with a set of changes to childbearing behaviour and family formation in most Western countries that demographers characterize as a ‘second demographic transition’ (Van de Kaa, 1987). The most relevant features of the second demographic transition are a reduction in fertility, extensive use of modern contraceptive methods, increases in mean age at marriage and age at first birth, together with rises in extra-marital childbearing, cohabitation and divorce. In 2003, around one-third of births in developed countries occurred out of wedlock, but cross-country differences remained substantial. The share of births outside marriage ranged from just under five per cent in Greece to 63 per cent in Iceland and 56 per cent in Sweden. At the core of these changes are an accentuation of individual autonomy, the abandonment of traditional religious beliefs, and a decline in sentiments of religiosity among individuals (Lesthaeghe and Surkyn, 1988). This transformation, which was already under way in most of Western Europe during the 1970s, became increasingly widespread in southern Europe from the middle of the 1980s.

Future implications

These demographic transformations have progressively moved to the centre of public debate both because of their social implications and the challenge they pose to the sustainability of welfare states in Western economies. As women continue to enter the labour force, labour market institutions need to adapt to minimize the trade-offs connected with childbearing to encourage fertility. In the absence of massive migration flows, smaller future cohorts facing improved economic conditions thanks to lower pressure in labour and housing markets may increase their fertility, as predicted by Easterlin's model (1975). However, since this would take place only in the long run, fertility rates are not likely to rebound to the replacement level in the near future (Bongaarts, 2001). In the meantime, recent data from the Eurobarometer shows that the ideal number of children has been decreasing for women aged between 20 and 34 since the late 1980s across the European Union. The average is just above 2.1, but for the first time, some countries such as Germany and Austria already portray below-replacement desired fertility (Goldstein, Lutz and Testa, 2003).

See Also

- demographic transition
- Easterlin hypothesis
- family economics
- human capital, fertility and growth
- labour market institutions
- population aging

Bibliography

Adsera, A. 2005. Vanishing children: from high unemployment to low fertility in developed countries. *American Economic Review* 95, 189–93.

Becker, G.S. 1991. *A Treatise on the Family*, enlarged edn. Cambridge, MA: Harvard University Press.

Becker, G.S. and Lewis, H.G. 1973. On the interaction between quantity and quality of children. *Journal of Political Economy* 81 (Suppl.), S279–88.

Becker, G.K. Murphy and Tamura, R. 1990. Human capital, fertility and economic growth. *Journal of Political Economy* 98, S12–S37.

Bongaarts, J. 2001. Fertility and reproductive preferences in post-transitional societies. *Population and Development Review* 27 (Suppl.), 260–81.

Butz, W.P. and Ward, M.P. 1979. The emergence of countercyclical U.S. fertility. *American Economic Review* 69, 318–28.

Easterlin, R. 1975. An economic framework for fertility analysis. *Studies in Family Planning* 7, 54–63.

Ermisch, J. 1989. Purchased child care, optimal family size and mother's employment. *Journal of Population Economics* 2, 79–102.

Galor, O. and Weil, D. 1996. The gender gap, fertility and growth. *American Economic Review* 86, 374–87.

Goldin, C. and Katz, L. 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. *Journal of Political Economy* 110, 730–70.

Goldstein, J.R., Lutz, W. and Testa, M.R. 2003. The emergence of sub-replacement family size ideals in Europe. *Population Research and Policy Review* 22, 479–96.

Heckman, J. and Willis, R. 1976. Estimation of a stochastic model of reproduction: an econometric approach. In *Household Production and Consumption 40*, ed. N. Terleckyj. New York: Columbia University Press.

Hotz, V.J., Klerman, J.A. and Willis, R. 1997. The economics of fertility in developed countries. In *Handbook of Population and Family Economics*, ed. M.R. Rosenzweig and O. Stark. Amsterdam: Elsevier.

Kohler, H.P., Billari, F. and Ortega, J.A. 2002. The emergence of lowest-low fertility in Europe during the 1990s. *Population and Development Review* 28, 599–639.

Lee, R. 2003. The demographic transition: three centuries of fundamental change. *Journal of Economic Perspectives* 17(4), 176–90.

Lesthaeghe, R. and Surkyn, J. 1988. Cultural dynamics and economic theories of fertility change. *Population and Development Review* 14, 1–45.

Mincer, J. 1963. Market prices, opportunity costs and income effects. In *Measurement in Economics: Studies in Mathematical Economics in Honor of Yehuda Grunfeld*. Stanford: Stanford University Press.

Mincer, J. and Polachek, S. 1974. Family investments in human capital: earnings of women, *Journal of Political Economy* 82 (Suppl.), S76–108.

Moffit, R. 1984. Optimal Life-cycle profiles of fertility and labor supply. *Research in Population Economics* 5, 29–50.

Van de Kaa, D. 1987. Europe's second demographic Transition. *Population Bulletin* 42, 1–57.

Willis, R. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81(Suppl.), S14–S64.

How to cite this article

Adsera, Alicia. "fertility in developed countries." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000289> doi:10.1057/9780230226203.0559

The New Palgrave Dictionary of Economics Online

fertility in developing countries

T. Paul Schultz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The associations between fertility and outcomes in the family and society have been treated as causal, but this is inaccurate if fertility is a choice coordinated by families with other life-cycle decisions, including labour supply of mothers and children, child human capital, and savings. Estimating how exogenous changes in fertility that are uncorrelated with preferences or constraints affect others depends on our specifying a valid instrumental variable for fertility. Twins have served as such an instrument and confirm that the cross-effects of fertility estimated on the basis of this instrument are smaller in absolute value than their associations.

Keywords

aging and retirement; birth control; capital accumulation; child care; child labour; child mortality; collective models of the household; comparative advantage; demographic transition; economic growth; family planning; fertility in developing countries; Heckscher, E. F.; home production; human capital; life-cycle savings; Malthus, T. R.; microcredit; mortality; overpopulation; Poor Law; population growth; precautionary insurance; time use; value of time; women's work and wages

Article

Fertility is a choice by parents involving a life-cycle claim on their resources, from which they may receive satisfaction as consumers and benefit as producers from children's labour and care-giving support. In addition, fertility may be the source of externalities that affect members of society other than the decision-making parents, in which case society may view fertility as a legitimate issue for social policy. To forecast fertility and the conditions under which public policies might be justified to modify fertility, economists require a basic understanding of its determinants as well as social consequences. In approaching this topic from the perspective of low-income countries today, the ideas of Malthus remain influential. He argued that population growth caused by high fertility erodes the welfare and productivity

of workers, and thus social policy which fostered greater fertility, such as the English Poor Law, contributed to ‘overpopulation’. Before considering how these spillover effects of fertility might be identified, an overview of historical thinking about the demographic–economic system may help to indicate the context in which Malthus's thinking was relevant to pre-industrial Europe, and how modern economics has extended his thinking to fertility as a lifetime choice of parents related to their time allocation and accumulation of human and physical capital.

Malthus's framework for the pre-industrial demographic – economic equilibrium

The determinants of fertility have engaged the interest of economists for some time. Adam Smith (1776) noted families were larger in settings where labour was scarce and child labour was especially valuable to parents, as in North America with its abundant land. Smith recognized that child mortality was higher among the poor, especially among those who were dependent on charity (for example, the Poor Laws). However, Malthus (1798) viewed fertility not as an individual choice but as an outcome of social institutions, because he did not think birth control was effective. He thought fertility was governed by the economic requirements society placed on a couple before allowing them to marry. Once married, the ‘constant passion of the sexes’ would lead in unregulated fashion to fertility. Society therefore restricted entry into marriage to those with favourable prospects for a livelihood or the income and assets to support the children that were expected to follow from the union. Over his lifetime, Malthus accumulated corroborating evidence on fertility, population growth and economic growth. Historians have since added to Malthus's evidence, confirming that Europe exhibited a late median age at marriage for a woman in her mid-twenties. This delay in childbearing led European women to have four or five births over their lifetime, rather than the six or seven if they had married five years earlier. Given the short life expectancy in pre-industrial Europe of about 35–40 years, this restrained level of fertility diminished substantially the resulting rate of population growth, except at frontiers of settlement where labour was scarce, land abundant, and marriage consequently early.

Heckscher (1963) thought Malthus's framework was relevant to Sweden. With the Swedish church's good records of marriages, births and deaths, and the Swedish king's need to estimate crop yields (for the purposes of taxation), annual time series for Sweden after 1720 appear accurate and show a positive covariation in marriage and fertility with good crop years, and shortfalls in marriage and subsequently fertility following poor crop years. Temperature and rainfall data available for Sweden after 1750 allow later analysts to incorporate this exogenous variation in weather and employ vector autoregression to estimate weather-driven Malthusian cycles in wages, fertility, as well as mortality (Eckstein, Schultz and Wolpin, 1984).

Working with French and Swiss parish registries of marriage, births, and deaths, Louis Henry (1972), the demographer, found evidence that couples exhibited a ‘natural’ rate of childbearing after marriage, until they eventually began to increase the intervals between their births after later parities, if economic conditions became less favourable. The emergence of this form of parity-specific application of birth control over the life-cycle of marriages was interpreted by Coale (1973) as an indicator of the onset of the ‘demographic transition’, when cultural restraints on fertility evolved from ‘natural’ proximate determinants to controlled ‘modern’ reproductive behaviour relying primarily on birth control.

Parish registries were then sampled from England from 1541 to 1871 by Wrigley and Schofield (1981)

to further investigate the Malthusian framework. Lee (1981) found that increases in marriage and birth rates were related to good weather and resulting declines in the price of wheat, as Malthus would have expected. But only about half of the covariation in weather/prices and annual birth rates is due to the fluctuations in first births that follow in the wake of variations in marriage. The other half is explained by variation in the length of inter-birth intervals. The latter finding casts doubt on Malthus's view that in this pre-industrial period couples did not exercise fertility choices within marriage. This spacing of births in response to economic wage cycles implied that the adoption of parity-specific birth control may not have been a cultural innovation, as assumed by Coale, but a customary form of individual behaviour adopted when additional births were unwanted. Some couples in pre-industrial societies appear able and willing to practice effective birth control when motivated economically. Fertility is thus to some degree a voluntary choice variable within marriage even in pre-industrial societies.

As the Industrial Revolution progressed in Europe and real wages increased, fertility nonetheless began to decline widely by the end of the 19th century. The Malthusian framework needed to be amended further to fit this experience in Europe and be applicable to low-income countries after 1960 as new methods of family planning were disseminated in the world and fertility fell despite modern economic growth. How was the secular decline in fertility to be explained in the face of rising personal incomes? The decline in child mortality, which gathered speed after 1870, reduced the need for parents to have extra births to replace the one out of five who might have at earlier times died from childhood diseases and infections. Parents might also scale back their demand for ‘insurance’ births motivated to reduce the likelihood that a couple would sustain above average child losses (Schultz, 1981). Becker (1960) proposed that the relative price of rearing children increased over time, causing the decline in parents’ demand for children. Mincer (1963) hypothesized that an increase in women's wages increased a couple's opportunity cost of having children, raising the shadow price of children. He argued that the rise in female labour-force participation and the decline in fertility were both caused by conditions increasing women's wages relative to other consumer prices and men's wages. These empirical patterns in the United States were soon replicated in other high-income countries.

Changing the relative prices of outputs of the economy is one possible source of variation in women's wages relative to men's that could explain changes in fertility. Men's labour in European agriculture was critical for plowing and producing food grains, whereas women specialized in home production as domestic servants and wives and to some degree in animal husbandry and the production of dairy commodities. Consequently, changing scarcity of grains relative to livestock and dairy product contributed to swings in the relative wages of men and women in Europe. The secular decline in international grain prices relative to dairy and livestock prices in the latter half of the 19th century was unprecedented due to the opening of new lands at the frontiers of European settlement in the United States and Russia, and contributed along with changes in production technologies to the rise in women's agricultural wages relative to men's in northern Europe and to the decline in fertility. Swedish historical data by region document after 1860 the fall in world grain prices, the associated increase in the wages of women relative to men, and the secular fall in fertility, when other developments are controlled for (Schultz, 1985).

Another factor credited with reducing fertility is the improvement in birth control technology, which reduced the monetary and psychic cost of limiting births, and provided techniques controlled by women, which were independent of sex. The major advances in technology occurred in the 1960s with the introduction of oral steroids (the pill) and the intra-uterine device (IUD), followed by further refinements

in their delivery systems. Traditional mechanisms for population control such as abortion, infanticide, coitus interruptus, and condoms have nonetheless allowed individuals to adjust their family size and affect population growth in various periods and parts of the world, well before the advent of these modern means of birth control. Although they may have facilitated the later demographic transition, these birth control technologies do not appear to have been necessary.

Microeconomic models of fertility behaviour

Willis (1973) adapted a comparative advantage trade model to the household lifetime fertility choice problem, wherein women's education was assumed to enhance women's productivity only in the market, and thereby increase the relative price of home production and decrease their demand for fertility. In his economic treatise on the family, Becker (1981) assigns a central role to market/non-market specialization of spouses in the household, with childbearing and rearing being the dominant non-market production activity traditionally performed by women.

To place more structure on fertility choices, Becker (1960; 1981) and Willis (1973) hypothesize that parents viewed the human capital of their children (child quality) as a substitute for their number of children (child quantity). If this were the case, then by definition income-compensated cross-price effects should be positive between child quantity and quality. In other words, increasing the price of children, for example by reducing the cost of birth control, would directly decrease fertility and indirectly increase the demand for child quality (with income held constant). Conversely, increasing the wage returns to schooling in the labour market would directly increase the demand for schooling and indirectly decrease the demand for births. Becker and Lewis (1974) postulate further that the income elasticity of demand for child quality exceeded the positive income elasticity for child quantity, which could account for the paradoxical decline in fertility with growth in income, without having to assume that children (quantity) are an ‘inferior’ good for which income effects are negative, or to show increases in women's value of their time in the modern economy caused the decline in their fertility.

The decline in fertility by half in high-income countries during the 20th century brought population growth to a halt in many of these countries. The decline in fertility by more than half in low-income countries in 40 years (1965–2005) is not yet comprehensively accounted for, although demographers are agreed that these trends in fertility are irreversible and the size of the world's population will stabilize later in the 21st century. How much does each of these conceptually distinct factors economists have described explain of this remarkable decline in fertility? I do not yet find a consensus on how to weight these factors in explaining cohort fertility. What fraction is due to an exogenous decline in mortality, the decline in the relative value of child labour, the increase in the value of women's time used in child care and the related increase in their empowerment, the increase in returns to schooling children, the greater income elasticities of demand for child quality than for quantity, and finally the improvements in birth control technology?

Identifying the effect of fertility on the welfare of families and society

The policy-relevant externalities of fertility could arise at the aggregate level or in terms of substitution effects within families. Malthus assumed that fertility added to subsequent generations of workers,

which reduced their wages and also changed the age composition of the population. But empirical evidence for these aggregate effects of fertility has not led to a consensus on their importance for today's low-income countries (National Research Council, 1986). At the microeconomic level of the family, fertility is found to be closely associated with other life-cycle choices by parents, including the share of time women allocate to the market economy, the investments parents make in the human capital of each of their children, and perhaps the savings out of income they accumulate in physical capital, possibly for old age support or precautionary insurance. But to assess the magnitude of these cross-effects of fertility, researchers must first specify an exogenous factor (not a choice variable within the orbit of the family) that affects fertility but leaves other constraints on the family life-cycle choices and outcomes unaffected and is unrelated to parent preferences (Schultz, 2007). In other words, an exclusion restriction or a valid instrumental variable is needed to account for some part of the variation in fertility that is independent of parent preferences and family life-cycle economic constraints. Otherwise, these cross-effects observed at the family level may not be causal and cannot be expected to occur when population policies reduce (or increase) fertility.

Twins are proposed by Rosenzweig and Wolpin (1980; 2000) as a ‘shock’ to the quantity of children that is uncorrelated with parent preferences or unobserved determinants of other family and child outcomes. Adjustment of investment in the schooling of other children in the family due to the occurrence of twins can then test the quantity–quality substitution hypothesis. They found support for the trade-off of quantity–quality on non-twin siblings in rural Indian households observed in 1970. A larger sample of twins collected in China provides the basis for estimating the impact of a twin on the quality of earlier- or later-born siblings, providing bounds to the magnitude of the cross effects, adjusted for substitution effects between siblings (Rosenzweig and Zhang, 2006). However, when twins are an instrument for fertility, the estimated quantity–quality trade-off tends to be smaller in absolute value than when estimated by direct association, that is, ordinary least squares (OLS). This could be due to the twin instrument being weak either because it occurs for only a small fraction of births (for example, one per cent) or because the underlying causal relationship is in fact weak and appears important only in biased single-equation associations (that is, OLS). The heterogeneity in parent preferences or other unobserved determinants of behavior could inversely affect child quantity and quality (Schultz, 2007).

Other studies have exploited twins as an instrument for fertility to assess how exogenous fertility affects the mother's market labour supply. These studies in high- and low-income countries generally confirm that the twin instrumental variable estimate of the effect of a birth on the mother's market labour supply tends to be absolutely smaller (negative) than the OLS estimate. The Durbin–Wu–Hausman specification test rejects the exogeneity of fertility in the determination of the mother's allocation of time to market work (Schultz, 2007), implying that the consistent instrumental variable estimate is preferred over the OLS estimate.

This twin-based cross effect of fertility on mothers’ labour supply may help to explain how policies which reduce fertility can facilitate modern economic growth, by adding to the per capita supply of labour and increasing the human capital of future generations. Finally, if parents when they have fewer children increase life-cycle savings for their support in old age, policies that facilitate a decline in fertility could raise savings and further augment growth rates. But estimates of these three potential cross effects of fertility-reducing population policies remain currently speculative.

The other instrument commonly used to identify the consequences of fertility on the welfare of families

relies on the sex composition of births, and has serious drawbacks. This variable may significantly affect parents' decisions on whether to have further children, and it may be assumed to be approximately independent of parent preferences or family constraints if there is no sex-selective abortion or infanticide. But this variable may not satisfy the criteria for a valid instrument, because the social and economic consequences of a child's sex involve many culturally distinct costs and benefits for his or her parents, such as the provision of dowries for daughters in some parts of the world. Thus, the sex composition of early births is likely to involve lifetime wealth effects for parents, in addition to affecting fertility, giving rise to many changes in family time allocation, expenditure patterns, and life-cycle savings (Rose, 2000). Therefore, the sex composition of children is not an instrumental variable for estimating how parents respond to a change in their fertility due to a population policy, if income and other family constraints are held constant. Finally, it should be noted that population policies may on the one hand subsidize learning and use of birth control, or at the other extreme fix a birth quota, as in China. There is no reason to expect expanding voluntary choices in the first case will have the same effect as rationing choices in the other policy regime.

Conclusions and research challenges

Parents may altruistically internalize in their fertility decisions the effects of their fertility on their welfare and that of their children, including investments in child quality and lifetime savings in financial assets (Becker, 1981). These parents are typically assumed to have secure property rights to their savings and access to financial institutions that minimize credit constraints. Population policies that reduce the cost of avoiding unwanted births may also be expected to affect gender empowerment, which does not enter decisively in the unitary model of the family proposed by Becker, but emerges in various recent bargaining and collective models of the family. Women may differentially gain from improved control of reproduction, because they physically bear the health costs of having births and invest disproportionately their time in child rearing. To derive predictions on how family bargaining affects fertility or vice versa requires more context-specific assumptions. Do mothers or fathers value children more highly? Does improved birth control technology empower women to bargain for a larger share of the gains from marriage? These remain open questions for more study. Women may value children as much as men do, and use their own increases in wealth to have more. Increased unearned income owned by the wife is associated, if the husband's income is held constant, with higher fertility in Thailand but not in Brazil (Schultz, 1990). Microcredit targeted to groups of women in Bangladesh increases women's earnings and increases their later fertility (Pitt et al., 1999).

In an experimentally designed family planning and health programme started in 1977 for women in rural villages of Matlab, Bangladesh, the women in villages benefiting from the programme had one fewer child by 1996 than did comparable women in comparison villages (Joshi and Schultz, 2006). The programme is also associated with increased woman's health, as measured by their body mass index (weight divided by height squared), reduced child mortality before age five, and increased years of schooling of boys aged 9–14 and 15–29. More studies of these long-run consequences of population policies on fertility and other family outcomes will be needed to assess the within-family consequences of fertility and population policies. Recognition that fertility is endogenous to other family life-cycle choices challenges economists to measure these potentially important life-cycle causal connections, and

thereby provide a sounder basis for evaluating how population policies affects the social allocation of resources.

See Also

- child labour
- fertility in developed countries
- human capital, fertility and growth

Bibliography

Becker, G.S. 1960. An economic analysis of fertility. In *Demographic and Economic Change in Developed Countries*. Princeton, NJ: Princeton University Press and NBER.

Becker, G.S. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Becker, G.S. and Lewis, H.G. 1974. *Interactions between quantity and quality of children*. In *Economics of the Family*, ed. T.W. Schultz. Chicago: University of Chicago Press.

Coale, A.C. 1973. The demographic transition reconsidered. *Proceedings of the International Population Conference*, vol. 1. Liège: International Union for the Scientific Study of Population.

Eckstein, Z., Schultz, T.P. and Wolpin, K. 1985. Short run fluctuations in fertility and mortality in pre-industrial Sweden. *European Economic Review* 26, 295–317.

Heckscher, E.F. 1963. *An Economic History of Sweden*, trans. G. Ohlin. Cambridge, MA: Harvard University Press.

Henry, L. 1972. *On the Measurement of Human Fertility*. Amsterdam: Elsevier.

Joshi, S. and Schultz, T.P. 2006. Family planning as a long term investment in development. Discussion paper, Economic Growth Center, Yale University.

Lee, R. 1981. Short term variation: vital rates, prices, and weather. In *The Population History of England, 1541–1871*, ed. J.A. Wrigley and R.S. Schofield. Cambridge, MA: Harvard University Press.

Malthus, T.R. 1798. *Essay on the Principle of Population*, ed. A. Flew. Harmondsworth: Penguin, 1970.

Mincer, J. 1963. Market prices, opportunity costs and income effects. In *Measurement in Economics*, ed. C. Christ et al. Stanford, CA: Stanford University Press.

National Research Council. 1986. *Population Growth and Economic Development: Policy Questions*. Washington, DC: National Academy Press.

Pitt, M., Khandkar, S.R., McKernan, S.-M. and Latif, M.A. 1999. Credit programs for the poor and reproductive behavior in low income countries. *Demography* 35, 1–21.

Rose, E. 2000. Gender bias, credit constraint and time allocation in rural India. *Economic Journal* 110, 738–58.

Rosenzweig, M.R. and Wolpin, K.I. 1980. Testing the quantity–quality fertility model: the use of twins as a natural experiment. *Econometrica* 48, 227–40.

Rosenzweig, M.R. and Wolpin, K.I. 2000. Natural ‘natural’ experiments in economics. *Journal of Economic Literature* 38, 827–74.

Rosenzweig, M.R. and Zhang, J. 2006. Do population control policies induce more human capital investment? Twins, birthweight, and China's ‘one child’ policy. Discussion Paper No. 933, Economic Growth Center, Yale University.

Schultz, T.P. 1981. *Economics of Population*. Reading, MA: Addison Wesley.

Schultz, T.P. 1985. Changing world prices, women's wages, and the fertility transition. *Journal of Political Economy* 93, 1126–54.

Schultz, T.P. 1990. Testing the neoclassical model of family labor supply and fertility. *Journal of Human Resources* 25, 559–634.

Schultz, T.P. 2007. Population policies, fertility, women's human capital and child quality. In *Handbook of Development Economics*, vol. 4, ed. T.P. Schultz. Amsterdam: North-Holland.

Smith, A. 1776. *The Wealth of Nations*, ed. E. Cannan. London: Methuen, 1961.

Willis, R.J. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81(2, Part II), S14–S64.

Wrigley, J.A. and Schofield, R.S. 1981. *The Population History of England, 1541–1871*. Cambridge, MA: Harvard University Press.

How to cite this article

Schultz, T. Paul. "fertility in developing countries." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000287> doi:10.1057/9780230226203.0560

The New Palgrave Dictionary of Economics Online

Fetter, Frank Albert (1863 – 1949)

Murray N. Rothbard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; basing-point pricing; Böhm-Bawerck, E. von; business cycles; capitalization; distribution theories; Fetter, F. A.; productivity; rent; time preference; uniform pricing; value theories

Article

Fetter was born on 3 March 1863 in the town of Peru, Indiana, and died on 21 March 1949 in Princeton, New Jersey. He was educated at Indiana and Cornell Universities and received his doctorate in economics at the University of Halle in Germany in 1894; he spent most of his life teaching at Cornell (1901–11) and Princeton universities (1911–34).

In journal articles on capital, interest and rent written largely between 1900 and 1914 (Fetter, 1977), and particularly in two treatises on economic principles (Fetter, 1904 and 1915), Fetter built upon Böhm-Bawerk and the Austrian School to develop a lucid and remarkable integrated structure of economic theory. He was able to accomplish this feat by purging economics of all traces of Ricardian or other British objectivist theories of value and distribution, in particular any differential theories of rent or productivity theories of interest.

Much of Fetter's achievement rested on his insight into the ordinary language meaning of ‘rent’ as simply the price of any durable good per unit time. He was then able to show that the prices of consumer goods are determined by their marginal utilities, and that these values are imputed back to determining the rental prices of factors of production by their marginal value productivity in serving consumers. The capital value, or price of the whole good (whether land, capital goods, or, Fetter might have added, the labourer under slavery) is then determined by the sum of its expected future returns, or rents, discounted by the social rate of time preference, or rate of interest. Thus, Fetter went beyond Böhm-Bawerk by arriving at a pure time preference theory of interest. Productivity and time preference are both highly important, but they have very different functions: the former in determining rents, and the latter determining the rate of interest. Thus, future rents are discounted by the rate of time preference and summed up, or ‘capitalized’, into their present capital value. Indeed, Fetter often called his contribution the ‘capitalization theory of interest’.

Fetter presented the fullest portrayal yet attained of the time market, the market for present as against future goods, as it permeates the economic system. The time market is not only the loan market, but also exists when entrepreneurs purchase or hire discounted factors of production (future goods) in return for money (a present good) and then reap a time or interest return when the product is later sold as a present good. Entrepreneurs earn profits, or suffer losses, as they lead the economy in the direction of a general equilibrium determined by marginal utility, marginal value productivity, and time preference.
While Fetter was led by his capitalization theory to arrive independently at the Mises–Hayek theory of the business cycle in 1927 (Fetter, 1977, pp. 260–316), he virtually abandoned value and distribution theory in the last two decades of his life to concentrate on the alleged monopolistic evils of basing-point pricing. He assumed that competition requires uniform pricing of products at the mill, while uniform pricing at centres of consumption is somehow monopolistic and deserves to be outlawed (Fetter, 1931). Fetter's shift of concern, coupled with a general loss of interest in economic theory in the United States between the two world wars and the continuing dominance of neo-Ricardian Marshallian theory in Britain, gravely hindered the incorporation of Fetter's notable contributions into modern economics.

Selected works

1904. *The Principles of Economics*. New York: Century.

1915. *Economic Principles*. New York: Century.

1931. *The Masquerade of Monopoly*. New York: Harcourt, Brace.

1977. *Capital, Interest and Rent: Essays in the Theory of Distribution*, ed. M. Rothbard. Kansas City: Sheed, Andrews and McMeel.

Bibliography

Coughlan, J.A. 1965. The contributions of Frank Albert Fetter (1863–1949) to the development of economic theory. Doctoral dissertation, Catholic University of America, Washington, DC.

Hoxie, R.F. 1905. Fetter's theory of value. *Quarterly Journal of Economics* 19(February), 210–30.

How to cite this article

Rothbard, Murray N. "Fetter, Frank Albert (1863–1949)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000056> doi:10.1057/9780230226203.0561

The New Palgrave Dictionary of Economics Online

feudalism

Robert Brenner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalist property relations; class; colonization; demesne production; diversified production; Dobb, M. H.; exchange; feudalism; Marx, K. H.; peasant economy; peasants; population growth; pre-capitalist property relations; subsistence; Sweezy, P. M.; vassalage.

Article

Modern discussions of feudalism have been bedevilled by disagreement over the definition of that term. There are three main competing conceptualizations. (1) Feudalism refers strictly to those social institutions which create and regulate a quite specific form of legal relationship between men. It constitutes a relationship in which a freeman (vassal) assumes an obligation to obey and to provide, primarily military, services to an overlord who, in turn, assumes a reciprocal obligation to provide protection and maintenance, typically in the form of a fief, a landed estate to be held by the vassal on condition of fulfilment of obligations (Bloch, 1939–40). (2) Feudalism refers, more broadly, to a form of government or political domination. It is a form of rule in which political power is profoundly fragmented geographically; in which, even within the smallest political units, no single ruler has a monopoly of political authority; and in which political power is privately held, and can thus be inherited, divided among heirs, given as a marriage portion, mortgaged, and bought and sold. Finally, the armed forces involve, as a key element, a heavy armed cavalry which is secured through private contracts, whereby military service is exchanged for benefits of some kind (Strayer, 1965; Ganshof, 1947). (3) Feudalism refers to a type of socio-economic organization of society as a whole, a mode of production and of the reproduction of social classes. It is defined in terms of the social relationships by which its two fundamental social classes constitute and maintain themselves. Specifically, the peasants, who constitute the overwhelming majority of the producing population, maintain themselves by virtue of their possession of their full means of subsistence, land and tools, so require no productive contribution by the lords to survive. This possession is secured by means of the peasants' collective political organization into self-governing communities, which stand as the ultimate guardian of the individual peasants' land. As a result of the peasants' possession and their consequent economic independence,

mere ownership of property cannot be assumed to yield an economic rent; in consequence, the lords are obliged to maintain themselves by appropriating a feudal levy by the exercise of *extra-economic* coercion. The lords are able to extract a rent by extra-economic coercion only in consequence of their political self-organization into lordly groups or communities, by means of which they exert a degree of domination over the peasants, varying in degree from enserfment to mere tribute taking (Marx, 1894; Dobb, 1946).

Though often thought to be in conflict, these conceptions are not only complementary but in fact integrally related to one another. While the lords' very existence as lords was based, as Marxists correctly insist, upon their appropriating a rent from the peasantry by extra-economic coercion, their capacity actually to exert such force in the rent relationship depended upon from their ability to construct and maintain the classically political ties of interdependence which joined overlord to knightly follower and thereby constituted the feudal groups which were the ultimate source of the lords' power. Conversely, while feudal bonds of interdependence were constructed, as the Weberians emphasize, to build highly localized governments capable at once of waging warfare, dispensing justice and keeping the peace, the *raison d'être* of the mini-states thus created was to constitute the dominant class of feudal society by establishing the instruments for extracting, redistributing and consuming the wealth upon which this class depended for their maintenance and reproduction. State and ruling class were thus two sides of the same coin. The distinctive ties which bound man to man in feudal society (not only the relations of vassalage strictly speaking, but also the more loosely defined associations structured by patronage, clientage, and family) constituted the building blocks, at one and the same time, for the peculiarly fragmented, locally based and politically competitive character of the feudal ruling class and for the peculiarly particularized nature of the feudal state. It was the lords' feudal levies which provided the material base for the feudal polity. It was the parcellized character of the feudal state, itself the obverse side of the decentralized structure of lordship through which rent was appropriated from the peasantry, which thus created the basic opportunities, set the ultimate limits and posed the fundamental problems for the lords' reproduction as a ruling class.

The origins of feudalism

The rise of feudalism was conditioned by an extended process of political fragmentation within the old Carolingian Empire. This is understandable, in part, in terms of a tendency to decentralization inherent in patrimonial rule. The patrimonial lord, to maintain his following, had, paradoxically, to provide his followers with the means to establish their independence from him. He could counteract their tendency to assert their autonomy through successful warfare and conquest, in which the followers found it worth their while to continue to submit to his authority. But in the absence of such profitable aggression, the followers had every incentive to assert their independence. It was in this way that the devolution and dissolution of more centralized forms of authority took place within the Carolingian Empire during the 9th and 10th centuries, as the Franks and their followers ceased to be conquerors, following a long period in which the empire had expanded. Fragmentation was hastened by the contemporaneous invasions of the Northmen, Saracens and Magyars. Effective authority fell, successively, from the king to his princes, to the counts and, ultimately, to local castleholders and even manorial lords, as the newly emerging, highly localized rulers turned their pillaging from foreign enemies to the local population

(Weber, 1956; Duby, 1978, pp. 147 ff).

Feudalism originally took shape in the early part of the 11th century in many parts of western Europe, including much of France, northern Italy and western Germany. Feudal rule was first constituted through the formation of lordly political groups, initially organized around a castle and led by the castellan. The castellan's power was derived from his knightly followers. The knights possessed military training, fought on horseback wearing (increasingly elaborate) coats of armour, often lived in the castle, and, from around the second third of the 11th century, tended to be bound to the castellan through ties of vassalage. The castellan's hegemony was manifested in his capacity to exert the right of the ban over his district – whose outer limits were usually no more than half a day's ride from the central fortress. The right of the ban, traditionally in the hands of the early medieval kings and the direct expression of their authority, allowed the castellan, above all, to extract dues from the peasant households within his jurisdiction, as well as to dispense justice and keep the peace. Although the surrounding lesser lords were usually tied to a castellan, in some cases they retained their full independence, not only collecting feudal rents derived from their authority over their tenants, but imposing taxes and exerting justice within their manorial mini-jurisdictions. In any case, all these lords confirmed their membership in the dominant class by claiming exemption from fiscal exactions: freedom under feudalism thus took the form of privilege. The peasants' unfreedom in some cases originated from their ancestors' having formally commended themselves to their lord; that is, their having subjected themselves to his domination in exchange for his assuring their safety. But, with the crystallization of feudal domination, it simply expressed the lords' having appropriated the right to extort protection money from them. The peasants' unfreedom was thus defined and constituted precisely by their subjection to arbitrary levies (Duby, 1973; 1978).

The feudal economy was thus structured, on the one hand, by a form of pre-capitalist property relations in which the individual peasant families, as members of a village community, *individually possessed* their means of reproduction. This contrasted with other pre-capitalist property forms in which the village community itself was the possessor (or more of one). On the other hand, under feudalism, the individual lords reproduced themselves by *individually appropriating* part of the peasants' product, backed up by localized communities of lords connected by various sorts of political bond, classically vassalage. This contrasted with other pre-capitalist property systems, in which the community, or communities, of lords appropriated the peasants' product collectively (as a tax) and shared out the proceeds among the community's, or communities', members.

Feudal property relations and the forms of individual economic rationality

The fundamental feudal property relationships of peasant possession and of lordly surplus extraction by extra-economic compulsion shaped the long-term evolution of the feudal economy. This was because these relationships were systematically maintained by the conscious actions of communities of peasants and of lords and thus constituted relatively inalterable constraints under which individual peasants and lords were obliged to choose the patterns of economic activity most sensible for them to adopt in order to maintain and improve their condition. The potential for economic development under feudalism was thus sharply restricted because both lords and peasants found it in their rational self-interest to pursue individual economic strategies which were largely incompatible with, if not positively antithetical to, specialization, productive investment and innovation in agriculture.

First, and perhaps most fundamental, because both lords and peasants were in full possession of what they needed to maintain themselves as lords and peasants, they were free from the *necessity* to buy on the market what they needed to reproduce, thus freed from dependence on the market and the necessity to produce for exchange, and thus exempt from the requirement to sell their output competitively on the market. In consequence, both lords and peasants were free from the necessity to produce at the socially necessary rate so as to maximize their rate of return and, in consequence, relieved of the requirement to cut costs so as to maintain themselves, and so of the necessity constantly to improve production through specialization and/or accumulation and/or innovation. Feudal property relations, in themselves, thus failed to *impose* on the direct producers that relentless drive to improve efficiency so as to survive, which is the *differentia specifica* of modern economic growth and required of the economic actors under capitalist property relations in consequence of their subjection to production for exchange and economic competition.

Absent the necessity to produce so as to maximize exchange values and in view of the underdeveloped state of the economy as a whole, the peasants tended to find it most sensible actually to deploy their resources so as to ensure their maintenance by producing directly the full range of their necessities; that is, *to produce for subsistence*. Given the low level of agricultural productivity which perforce prevailed, harvests and therefore food supplies were highly uncertain. Since food constituted so large a part of total consumption, the uncertainty of the food market brought with it highly uncertain markets for other commercial crops. It was therefore rational for peasants to avoid the risks attached to dependence upon the market, and to do so they had to diversify rather than specialize, marketing only physical surpluses. In fact, beyond their concern to minimize the risk of losing their livelihood, the peasants appear to have found it desirable to carry out diversified production simply because they wished to maintain their established mode of life – and, specifically, to avoid the subjection to the market which production for exchange entails, and the total transformation of their existence which that would have meant.

To make possible ongoing production for subsistence, the peasants naturally aimed to maintain their plots as the basis for their existence. To ensure the continuance of their families into the future, they also sought to ensure their children's inheritance of their holdings. Meanwhile, they tended to find it rational to have as many children as possible, so as to ensure themselves adequate support in their old age. The upshot was relatively large families and the subdivision of plots on inheritance.

Like the peasants, the lords occupied a ‘patriarchal’ position, possessing all that they needed to survive and thus freed of any necessity to increase their productive capacities. Moreover, even to the extent they wished, for whatever reason, to increase the output of their estates, the lords faced nearly insuperable difficulties in accomplishing this by means of increasing the productive powers of their labour and their land. Thus, if the lords wished to organize production themselves, they had no choice but to depend for labour on their peasants, who possessed their means of subsistence. But precisely because the peasants were possessors, the lords could get them to work only by directly coercing them (by taking their feudal rent in the form of labour) and could *not* credibly threaten to ‘fire’ them. The lords were thereby deprived of perhaps the most effective means yet discovered to impose labour discipline in class-divided societies. Because the peasant labourers had no *economic* incentive to work diligently or efficiently for the lords, the lords found it extremely difficult to get them to use advanced means of production in an effective manner. They could force them to do so only by making costly unproductive investments in supervision.

In view of both the lords’ and the peasants’ restricted ability effectively to allocate investment funds to

improved means of production to increase agricultural efficiency, both lords and peasants found that the only really effective way to raise their income via productive investment was by opening up new lands. Colonization, which resulted in the multiplication of units of production on already existing lines, was thus the preferred form of productive investment for both lords and peasants under feudalism.
Beyond colonization and the purchase of land, feudal economic actors, above all feudal lords, found that the best way to improve their income was by forcefully *redistributing* wealth away from the peasants or from other lords. This meant that they had to deploy their resources (surpluses) towards building up their *means of coercion* by means of investment in military men and equipment, in particular to improve their ability to fight wars. A drive to *political accumulation*, or state building, was the feudal analogue to the capitalist drive to accumulate capital.

The long-term patterns of feudal economic development

Feudal property relations, once established, thus obliged lords and peasants to adopt quite specific patterns of individual economic behaviour. Peasants sought to produce for subsistence, to hold on to their plots, to produce large families and to provide for their families' future generations by bequeathing their plots. Both lords and peasants sought to use available surpluses funds to open new lands. Lords directed their resources to the amassing of greater and better means of coercion. Generalized on a society-wide basis, these patterns of individual economic action determined the following developmental patterns, or laws of motion, for the feudal economy as a whole:

(i) Declining productivity in agriculture (Bois, 1976; Hilton, 1966; Postan, 1966)

The generalized tendency to adopt production for subsistence on the part of the peasantry naturally constituted a powerful obstacle to commercial specialization in agriculture and to the emergence of those competitive pressures which drive a modern economy forward. In so doing, it also posed a major barrier to agricultural improvement by the peasantry, since a significant degree of specialization was required to adopt almost all those technical improvements which would come to constitute 'the new husbandry' or the agricultural revolution (fodder crops, up-and-down farming, and so on). In addition, production aimed at subsistence and the maintenance of the plot as the basis for the family's existence posed a major barrier to those rural accumulators, richer peasants and lords who wished to amass land or to hire wage labour, since the peasants would not readily part with their plots, which were the immediate bases for their existence, unless compelled to do so; nor could they be expected to work for a wage unless they actually needed to.
Further counteracting any drive to the accumulation of land and labour was the tendency on the part of the possessing peasants to produce large families and subdivide their holdings among their children. The peasants' parcellization of plots under population growth tended to overwhelm any tendency towards the build-up of large holdings in the agricultural economy as a whole, further reducing the potential for agriculture improvement.
Finally, individual peasant plots were, most often, integrated within a village agriculture which was, in critical ways, controlled by the community of cultivators. The peasant village regulated the use of the pasture and waste on which animals were raised, and the rotation of crops in the common fields. Individual peasants thus tended to face significant limitations on their ability to decide how to farm their

plots and thus, very often, on their capacity to specialize, build up larger consolidated holdings, and so forth.

To the extent that the lords succeeded in increasing their wealth by means of improving their ability coercively to redistribute income away from the peasantry, they further limited the agricultural economy's capacity to improve. Increased rents in whatever form reduced the peasants' ability to make investments in the means of production. Meanwhile, the lords' allocation of their income to military followers and equipment and to luxury consumption ensured that the social surplus was used unproductively, indeed wasted. To the extent – more or less – that the lords increased their income, the agricultural economy was undermined.

(ii) Population growth (Postan, 1966)

The long-term tendency to the decline of agricultural productivity thus conditioned by the feudal structure of property was realized in practice as a consequence of rising population. The peasants' possession of land allowed children to accede to plots and, on that basis, to form families at a relatively early age. Married couples, as noted, had an incentive to have many children, both to provide insurance for their old age and to assure that the line would be continued. The result was that all across the European feudal economy we witness a powerful tendency to population growth from around the beginning of the 12th century, which led, almost everywhere, to a doubling of population over the following of two centuries.

(iii) Colonization (Postan, 1966; Duby, 1968)

The only significant method by which the feudal economy achieved real growth and counteracted the tendency to declining agricultural productivity was by way of opening up new land for cultivation. Indeed, economic development in feudal Europe may be understood, at one level, in terms of the familiar race between the growth of the area of settlement and the growth of population. During the 12th and 13th centuries, feudal Europe was the scene of great movements of colonization, as settlers pushed eastward across the Elbe and southward into Spain, while reclaiming portions of the North Sea in what became the Netherlands. The opening of new land did, for a time, counteract and delay the decline of agricultural productivity. Nevertheless, in the long run – as expansion continued, as less fertile land was brought into cultivation, and as the man/land ratio rose – rents rose, food prices increased, and the terms of trade increasingly favoured agricultural as opposed to industrial goods. At various points during the 13th and early 14th centuries, all across Europe, population and production appear to have reached their upper limits, and there began to ensue a process of demographic adjustment along Malthusian lines.

(iv) Political accumulation or state building (Dobb, 1946; Anderson, 1974; Brenner, 1982)

Given the limited potential for developing the agricultural productive forces and the limited supply of cultivable land, the lordly class, as noted, tended to find the build-up of the means of force for the purpose of redistributing income to be the best route for amassing wealth. Indeed, the lords found themselves more or less *obliged* to try to increase their income in order to finance the build-up of their

capacity to exert politico-military power. This was, first of all, because they could not easily escape the politico-military conflict or competition that was the inevitable consequence of the individual lords' direct possession of the means of force (the indispensable requirement for their maintenance as members of the ruling class over and against the peasants) and thus of the wide dispersal of the means of coercion throughout the society. It was, secondly, because they had to confront increasingly well-organized peasant communities and, as feudal society expanded geographically, to counteract the effects of increasing peasant mobility.

In the first instance, of course, military-politico efficacy required the collecting and organizing of followers. But to gain and retain the loyalty of their followers the overlords had to feed and equip them and, in the long run, competitively reward them. Minimally, the overlord's household had to become a focus of lavish display, conspicuous consumption and gift-giving, on par with that of other overlords. But beyond this, it was generally necessary to provide followers with the means to maintain their status as members of the dominant class – that is, a permanent source of income, requiring a grant of land with associated lordly prerogatives (classically the fief). But naturally such grants tended to increase the followers' independence from the overlords, leading to renewed potential for disorganization, fragmentation and anarchy. This was the perennial problem of all forms of patrimonial rule and at the centre of feudal concerns from the beginning. The tendency to fragmentation was, moreover, exacerbated as a result of the pressure to divide lordships and lands among children. To an important degree, then, feudal evolution may be understood as a product of lordly efforts to counteract political fragmentation and to construct firmer intra-lordly bonds with the purpose of withstanding intra-lordly politico-military competition and indeed of carrying on the successful warfare that provided the best means to amass the wealth ultimately required to maintain feudal solidarity. This meant not only the development of better weapons and improved military organization, but also the creation of larger and more sophisticated political institutions, and naturally entailed increased military and luxury consumption.

Actually to achieve more effective political organization of lordly groups required political innovation. Speaking broadly, the constitution of military bands around a leading warlord for external warfare, especially conquest, most often provided the initial basis of intra-lordly cohesion. This served as the foundation for developing more effective collaboration within the group of lords for the protection of one another's property and for controlling the peasantry. As a further step in this direction, the overlord would establish his pre-eminence in settling disputes among his vassals (as in Norman England). Next, the leading lord might extend feudal centralization by establishing immediate relations with the undertenants of his vassals. One way this took place was through constructing direct ties of dependence with these rear vassals (as in 11th-century England). More generally, it was accomplished by the extension of central justice to ever broader layers of the lordly class, indeed the free population as a whole. Sometimes the growth of central justice was achieved through the more or less conscious collaboration of the aristocracy as a whole (as in 12th-century England). On other occasions it had to be accomplished through more conflicted processes whereby the leading lord (monarch, prince) would accept appeals over the heads of his vassals from their courts (as in medieval France). Ultimately, the feudal state could be further strengthened only by the levying of taxes, and this almost always required the constitution of representative assemblies of the lordly class.

This is not to say that a high level of lordly organization was always required. Nor is it to argue that state building took place as an automatic or universal process. At the frontiers of European feudal society, to

the south and east, colonization long remained an easy option, and there was relatively little (internally generated) pressure upon the lordly class to improve its self-organization. At the same time, just because stronger feudal states might become necessary did not always determine that they could be successfully constructed. Witness the failure of the German kings to strengthen their feudal state in the 12th century, and the long-term strengthening of the German principalities which ensued. The point is that, to the degree that disorganization and competition prevailed within and between groups of feudal lords, they would tend to be that much more vulnerable not only to depredations from the outside, but to the erosion of their very dominance over the peasants. The French feudal aristocracy thus paid a heavy price for their early, highly decentralized feudal organization, suffering not only significant losses of territory to the Anglo-Normans, but a serious reduction in their control over peasant communities and a consequent decline in dues. The French aristocracy's later recovery and successes may be attributed, at least in large part, to their evolution of a new, more centralized, more tightly knit form of political organization – the tax/office state, where property in office (rather than lordship/land) gave the aristocracy rights to a share in centralized taxation (rather than feudal rent) from the peasants. In sum, the economic success of individual lords, or groups of them, does seem to have depended upon successful feudal state building, and the long-term trend throughout Europe, from the 11th through to the 17th century, appears to have been towards ever more powerful and sophisticated feudal states.

Trade, towns and feudal crisis

The growing requirements of the lordly class for the weaponry and luxury goods (especially fine textiles) needed to carry on intra-feudal politico-military competition were at the source of the expansion of commerce in feudal Europe. The growth of trade made possible the rise of a circuit of interdependent productions in which the artisan-produced manufactures of the towns were exchanged for peasant-produced necessities (food) and raw materials, appropriated by the lords and sold to merchant middlemen. Great towns thus emerged in Flanders and north Italy in the 11th and 12th centuries on the basis of their industries' ability to capture a preponderance of the demand for textiles and armaments of the European lordly class as a whole.

In the first instance, the growth of this social division of labour within feudal society benefited the lords, for it reduced costs through increasing specialization, thus making luxury goods relatively cheaper. Nevertheless, in the long run it meant a growing disproportion between productive and unproductive labour in the economy as a whole, for little of the output of the growing urban centres went back into production to augment the means of production or the means of subsistence of the direct peasant producers; it went instead to military destruction and conspicuous waste. Over time, increasingly sophisticated political structures and technically more advanced weaponry meant growing costs and thus increased unproductive expenditures. At the very time, then, that the agricultural economy was reaching its limits, the weight of urban society upon it grew significantly, inviting serious disruption.

Because the growth of lordly consumption proceeded in response to the requirements of intra-feudal competition in an era of increasingly well-constructed feudal states, the lords could not take into account its effect on the underlying agricultural productive structure. All else being equal, the growth of population beyond the resources to feed it could have been expected to call forth a Malthusian adjustment, and most of Europe did witness the onset of famine and the beginning of demographic downturn in the early 14th century. Nevertheless, while the decline of population meant fewer mouths to

feed with the available resources, it also meant fewer rent-paying tenants and so, in general, lower returns to the lords. The decline in seigneurial incomes induced the lords to seek to increase their demands on the peasantry, as well as to initiate military attacks upon one another. The peasants were thus subjected to increasing rents and the ravages of warfare at the very moment that their capacity to respond was at its weakest, and their ability to produce and to feed themselves was further undermined. Further population decline brought further reductions in revenue leading to further lordly demands – resulting in a downward spiral which was not reversed in many places for more than a century. The lordly revenue crisis and the ensuing seigneurial reaction thus prevented the normal Malthusian return to equilibrium. A general socio-economic crisis, the product of the overall feudal class/political system, rather than a mere Malthusian downturn, gripped the European agrarian economy until the middle of the 15th century (Dobb, 1946; Hilton, 1969; Bois, 1976; Brenner, 1982).

In the long run, feudal crisis brought its own solution. With the decline of population, peasant cultivation drew back onto the better land, making for the potential of increased output per capita and growing peasant surpluses. Meanwhile, civil and external warfare seem to have abated, a reflection perhaps of the exhaustion of the lordly class, and the weight of ruling class exactions on the peasantry declined correspondingly, especially as the peasants were now in a far better position to pay. The upshot was a new period of population increase and expansion of the area under cultivation, of the growth of European commerce, industry and towns, and, ultimately, of the familiar outrunning of production by population. Meanwhile, lordly political organization continued to improve, feudal states continued to grow, intra-feudal competition continued to intensify, and, over the long run, lordly demands on the peasants continued to increase even as the capacity of the peasantry began, once again, to decline. By the end of the 16th century one witnesses, through most of Europe, a descent into the ‘general crisis of the 17th century’ which took a form very similar to that of the ‘general crisis of the 14th and 15th centuries’. Clearly, through most of Europe, the old feudal property relations persisted, undergirding the repetition of established patterns of feudal economic non-development.

Approaches to transition

It is an implication of the foregoing analysis that so long as feudal property relations persisted, the repetition of the same long-term economic patterns could be expected. So long as feudal property relations obtained, lords and peasants could be expected to find it rational to adopt the same patterns of individual economic behaviour; in consequence, one could expect the same long-term cyclical tendencies to declining agricultural productivity, population growth, and the opening of new land, issuing in a tendency to Malthusian adjustment but overlaid by a continuation of the secular tendency to lordly state building and growing unproductive expenditures. Generally speaking, so long as feudal property relations obtained, no inauguration of a long-term pattern of modern economic growth could be expected. From these premises, it is logical to conclude that the onset of economic development depended on the transformation of feudal property relations into capitalist property relations, and that indeed is the point of departure of a long line of theorists and historians (Marx, 1894; Dobb, 1946; Hilton, 1969; Bois, 1976).

Nevertheless, beginning with Adam Smith himself, a whole school of historically sensitive theorists have found it quite possible to ignore, or sharply to downplay, the problem of the transformation of property relations and of social relationships more generally in seeking to explain economic

development. These theorists naturally refuse to go along with the Adam Smith of *Wealth of Nations* Book I in contending that the mere application of individual economic rationality will, directly and automatically, bring economic development. They nevertheless follow the Adam Smith of *Wealth of Nations* Book III in arguing that, given the appearance of certain specific, *quite-reasonable-to-expect* exogenous economic stimuli, rational self-interested individuals can indeed be expected to take economic actions which will detonate a pattern of modern economic growth. Specifically, it is their hypothesis that the growth of commerce, an enormously widespread if not universal phenomenon of human societies, systematically has led pre-capitalist economic actors to assume capitalist motivations or goals, to adopt capitalist norms of economic behaviour, and, eventually, to bring about the transformation of pre-capitalist to capitalist property relations. It is undoubtedly because Adam Smith and his followers have believed that the growth of exchange will *in itself* sooner or later create the necessary conditions for modern economic growth that they have not greatly concerned themselves with these conditions or viewed their emergence as a problem which needs addressing.

Thus, Smith and a long line of followers, prominently including the economic historian of medieval Europe Henri Pirenne and the Marxist economist Paul Sweezy, have all produced analyses which follow essentially the same progression. First, merchants, emanating from outside feudal society, offer previously unobtainable products to lords and peasants who hitherto had produced only for subsistence. This is understood as a more or less epoch-making historical event, an original rise of trade. Next, the very opportunity to purchase these new commodities induces the individual economic actors to adopt businesslike attitudes and capitalist motivations, specifically to relinquish their norm of production for subsistence and to adopt the economic strategy of capitalists-in-embryo – viz., production for exchange so as to maximize returns by way of cost cutting. Third, since pre-capitalist property relations, marked by the producers' possession of the means of subsistence and by the lord's extraction of a surplus by means of extra-economic coercion, prevent the individual economic actors from most effectively deploying their resources to maximize exchange values, both lords and peasants move, on a unit-by-unit basis, to transform these property relations in the direction of capitalist property relations. In particular, the lords dispense with their (unproductive) military followers and military luxury expenditures; they free their hitherto dominated peasant producers; they expropriate these peasants from the land; then, finally, they enter into contractual relations with these free, expropriated peasants. This gives rise, within each unit to the installation of free, necessarily commercialized (market dependent) tenants on economic leases, who, ultimately, hire wage labourers. The end result is the establishment of capitalist property relations and capitalist economic norms in the society as a whole and the onset of economic development (Smith, 1776; Pirenne, 1937; Sweezy, 1950).

The foregoing argument of what might be called the Smithian school is designed, implicitly or explicitly, to show how the rise of exchange in a feudal setting, in itself creates the conditions under which rational economic actors will pursue self-interested action which leads, on an economy-wide basis, to modern economic growth. Nevertheless, the validity of each step in the Smithian argument can be, and has been, challenged by those who take as their point of departure the historically established property relations. It is the essence of their position that the Smithians can sustain their argument only by failing sufficiently to understand what patterns of economic activity individual lords and peasants will find it rational to adopt in response to the rise of trade, *given* the prevalence of feudal property relations (Marx, 1894; Dobb, 1946; Bois, 1976).

In the first place, although long-distance merchants may bring to feudal lords and peasants commodities

they could not previously obtain, the merchants' mere offer of these commodities cannot ensure that the lords and peasants will, in turn, put their own products on the market in order to buy them. Given the existence of feudal property relations, both lords and peasants may be assumed to have everything they need to maintain themselves. The opportunity to buy new goods may very well make it possible for the pre-capitalist economic actors to increase or enrich their consumption, but this does not mean that they will take advantage of this opportunity. The increased potential for exchange simply cannot determine that exchange will increase (Luxemburg, 1913).

Secondly, even where the appearance of new goods brought by merchants does induce the lords to try to increase their consumption by raising their output and increasing the degree to which they orient their production towards exchange, this will hardly lead them to find it in their rational self-interest to dismantle, in piecemeal fashion, the existing feudal property relations by freeing and expropriating their peasants. Given the reproduction of feudal property relations by communities of feudal lords and peasants, the individual lords can hardly find it in their rational self-interests to free their peasants, for they would lose thereby their very ability to exploit them, and thus their ability to make an income. The point is that, once freed from the lord's extra-economic domination, his *possessing* peasants would have no need to pay *any* levy to him, let alone increase the quality and quantity of their work for him. Moreover, even if the lord could, at one and the same time, free *and* expropriate his peasants, he would still lose by the resulting transformation of his unfree peasant possessors into free landless tenants and wage labourers, for the newly landless tenants or wage labourers would have no reason to stay and work for their former lord or to take up a lease from him.

To the degree, then, that lords sought to increase their output in response to trade, they appear to have found it in their rational self-interest not to transform but to intensify the pre-capitalist property relations. Because they found it, on the one hand, difficult to get their possessing peasants effectively to use more productive techniques on their estates, and, on the other hand, irrational to instal capitalist property relations within their units, they seem to have had little choice but to try to do so within the constraints imposed by feudal property relations – by increasing their levies on the direct producers in money, kind or labour. To make this possible, they had no choice but to try to strengthen their institutionalized relationship of domination over their peasants, by investing in improved means of coercion and by improving the politico-military organization of their lordly groups. It needs to be emphasized that the lords could not be sure they could succeed in this, for the peasants would likely resist, and perhaps successfully. But in so far as the lords could dictate terms, this was the route they found most promising. Witness the growth of demesne farming in response the growth of the London market in 13th-century England or, more spectacularly, the rise of a neo-serfdom throughout later medieval and early modern eastern Europe in response to the growth of trade with the west (Dobb, 1946).

Finally, it needs to be noted that the sorts of products on the market which were most likely to stimulate the exploiters to try to increase their income for the purpose of trade were goods which 'fit' their specific reproductive needs. These were not producer goods but, on the contrary, means of consumption – specifically, materials useful for building up the exploiters' political and military strength. They were certainly not luxury goods in the ordinary sense of superfluities, for they were, in fact, necessities for the exploiters. But they were luxuries in that their production involved a subtraction from the means available to the economy to expand its fundamental productive base.

Paradoxically, then, to the extent that the rise of trading opportunities, *in itself*, can be expected to affect

precapitalist economies, it is likely to bring about not the loosening but the tightening of pre-capitalist property forms, the growth of unproductive expenditure, and the quickening not of economic growth but of stagnation and decline.

From feudalism to capitalism

The onset of modern economic growth thus appears to have required the break-up of pre-capitalist property relations characterized by the peasants' possession of their means of subsistence and the lords' surplus extraction by extra-economic compulsion. Nevertheless, neither the regular recurrence of system-wide socio-economic crisis nor the widespread growth of exchange could, in themselves, accomplish this. The problem which thus emerges is how feudal property relations could ever have been transformed.

To begin to confront this question, one can advance two basic hypotheses which follow more or less directly from the central themes of this article:

1. In so far as lords and peasants, acting either individually or as organized into communities, were able to realize their conscious goals, they succeeded, in one way or another, in maintaining pre-capitalist property forms. This is to say, once again, that the patterns of economic activity that individual lords and peasants found it reasonable to pursue could not aim at transforming the feudal property structure. It is also to emphasize that, because peasants and lords organized themselves into communities for the very purpose of maintaining and strengthening, respectively, peasant possession and the institutionalized relationships required for taking a feudal rent by extra-economic coercion, lords and peasants acting as communities were unlikely to aim at undermining feudal property forms. Peasants might, through collective action, conceivably have reduced to zero the lords' levies and eliminated the lords' domination; but, even in this extreme case, they would have ended up constituting a community of peasants fully in possession of their means of subsistence, with all of the barriers to economic development entailed by that set of property relations. Were the lords, on the other hand, to have succeeded to the greatest extent conceivable in overcoming peasant resistance, they would only to that degree have strengthened their controls over the peasants and increased their rate of rent, thus tightening feudal property relations.

2. Where breakthroughs took place to modern economic growth in later medieval and early modern Europe, these must be understood as *unintended consequences* of the actions by individual lords and peasants and by lordly communities and peasant communities in seeking to maintain themselves as lords and peasants in feudal ways. In other words, the initial transitions from feudal to capitalist property relations resulted from the attempts by feudal economic actors, as individuals and collectivities, to follow feudal economic norms or to reproduce feudal property relations under conditions where, doing so, actually had the effect – for various reasons – of undermining those relations.

To give substance to these hypotheses would require a lengthy historical discussion. It is here possible only to note a broad contrast in the historical evolutions of the different European regions during the late medieval and early modern periods. Through most of pre-industrial Europe, east and west, varying processes of class formation brought, in one form or another, the reproduction of feudal property relations and, in turn, the repetition of long-term developmental patterns familiar from the medieval period. However, in a few European regions, feudal property relations dissolved themselves, giving rise, for the first time, to essentially modern processes of economic development.

Thus, through much of later medieval and early modern western Europe (France and parts of western Germany), although peasants succeeded in very much strengthening peasant possession, winning their freedom and destroying all forms of surplus extraction by extra-economic coercion by individual lords, the lords succeeded, in response, in maintaining themselves by means of constituting a new, more potent form of now-collective surplus extraction by extra-economic compulsion, the tax/office state. At the same time, throughout late medieval and early modern eastern Europe, despite the peasants' initially very powerful rights in the land and the lords' initially very weak feudal controls, the lords ended up erecting an extremely tight form of individual lordly domination and surplus extraction by extra-economic compulsion – serf-operated demesne production. The consequence of these reconsolidations of essentially feudal property relations throughout most of Europe, east and west, was the reappearance throughout most of Europe during the early modern period of the same trends towards demographically powered expansion, towards the continued build-up of larger and more sophisticated states and, ultimately, towards socio-economic crisis as had characterized the medieval period.

The evolution of property relations in late medieval and early modern England was in some contrast to that of both eastern and (most of) western Europe, with epochal consequences for the long-term pattern of economic development. During this period, English lords, unlike those in eastern Europe, failed, as did those throughout almost all of western Europe, in their attempts to maintain, let alone intensify, their extra-economic controls over their peasantry. On the other hand, the English lords, unlike those throughout much of western Europe, did ultimately succeed in maintaining their positions by means of preventing their customary tenants from achieving full property in their plots. They were able, in consequence, to consign these tenants to leasehold status, and thus to assert their own full property in the land.

The unintended consequence of the actions of English peasants and lords aiming to maintain themselves as peasants and lords in feudal ways was thus to introduce a new system of now-capitalist property relations in which the direct producers were free from the lords' extra-economic domination but also separated from their full means of reproduction (subsistence). In the upshot, tenants without direct access to their means of reproduction had no choice but to produce competitively for exchange and thus, so far as possible, to specialize, accumulate and innovate. At the same time, the landlords found themselves obliged to create larger, consolidated and well-equipped farms if they wished to attract the most productive tenants. The long-run results were epoch making. Under the pressures of competition, processes of differentiation led to the emergence of an entrepreneurial class of capitalist tenant farmers who were ultimately able to employ wage labourers. Meanwhile, the drive to cut costs in agricultural production ultimately brought about an agricultural revolution, as market-dependent farmers were obliged to adopt techniques which long had been available, but long eschewed by possessing peasants who would not intentionally take the risks of specialization, let alone make the necessary capital investments. The secular decline in food costs and the secular rise in living standards which resulted underpinned the movement of population off the land and into industry and made possible the rise of the home market. Industry and agriculture, for the first time, proved mutually supporting, rather than mutually competitive, and population increase served to stimulate economic growth rather than to undermine it. England experienced unbroken industrial and demographic growth right through the 17th and 18th centuries, which ultimately issued in the Industrial Revolution.

See Also

- Dobb, Maurice Herbert
- peasants
- Sweezy, Paul Marlor

Bibliography

Anderson, P. 1974. *Passages from Antiquity to Feudalism*. London: New Left Books.

Bois, G. 1976. *La crise du féodalisme*. Paris: Editions EHESS.

Bloch, M. 1939–40. *Feudal Society*, Trans. L.A. Manyon Chicago: University of Chicago Press, 1961.

Brenner, R. 1982. The agrarian roots of European feudalism. In *The Brenner Debate: Agrarian Class Structure and Economic Development in Preindustrial Europe*, ed. T.H. Aston, Cambridge: Cambridge University Press, 1985.

Dobb, M. 1946. *Studies in the Development of Capitalism*. London: Routledge & Kegan Paul; New York: International Publishers, 1947.

Duby, G. 1968. *Rural Economy and Country Life in the Medieval West*. Trans. C. Postan Columbia: University of South Carolina Press.

Duby, G. 1973. *The Early Growth of the European Economy*. Trans. H.B. Clarke, Ithaca: Cornell University Press, 1974.

Duby, G. 1978. *The Three Orders of Society*. Trans. T.N. Bisson, Chicago: University of Chicago Press, 1980.

Ganshof, F.L. 1947. *Feudalism*. Trans. P. Grierson, New York: Harper & Row, 1961.

Hilton, R.H. 1966. *A Medieval Society*. New York: Wiley.

Hilton, R.H. 1969. *The Decline of Serfdom*. London: Macmillan.

Luxemburg, R. 1913. *The Accumulation of Capital*. Trans. A. Schwarzschild New York: Monthly Review Press, 1968.

Marx, K. 1894. *Capital*, Vol. 3, New York: International Publishers, 1967.

Pirenne, H. 1937. *Economic and Social History of Medieval Europe*. New York: Harcourt Brace & Co.

Postan, M.M. 1966. Medieval agrarian society in its prime: England. In *The Cambridge Economic History of Europe*, vol. 1: *The Agrarian Life of the Middle Ages*, 2nd edn, ed. M.M. Postan and H.J. Habakkuk. Cambridge: Cambridge University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. R.H. Campbell, A. S. Skinner and W.B. Todd, Oxford: Clarendon Press, 1976.

Strayer, J.R. 1965. *Feudalism*. New York: Van Nostrand Reinhold.

Sweezy, P. 1950. The transition from feudalism to capitalism. *Science and Society* 14(2), 134–57.

Weber, M. 1956. Patriarchalism and patrimonialism. Feudalism, Standestaat, and patrimonialism. In *Economy and Society*, 2 vols, ed. G. Roth and C. Wittich, Berkeley: University of California Press, 1978.

How to cite this article

Brenner, Robert. "feudalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000058> doi:10.1057/9780230226203.0562

The New Palgrave Dictionary of Economics Online

fiat money

Neil Wallace

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Fiat money is an intrinsically useless object that serves as a medium of exchange. One challenge is to construct models that depict the ancient notion that a medium of exchange is beneficial. Another is to construct models in which the medium of exchange has a low rate of return. This article reviews how those challenges have been approached and argues that progress has been achieved by taking seriously some old ideas about the circumstances in which money is helpful and about the desirable properties of money: money is helpful when there are absence-of-double-coincidence difficulties that cannot be easily overcome with credit; and a good money has desirable physical properties – recognizability, portability and divisibility.

Keywords

absence of double coincidence; Arrow–Debreu model; asymmetric information; cash-in-advance models; central banking; commitment; commodity money; Cournot quantity game; credit; fiat money; Friedman rule; imperfect monitoring; incentive feasibility; incomplete markets; Infinite horizons; inside and outside money; large-family models; medium of exchange; money; Nash equilibrium; open-market operations; perfect monitoring; pooling equilibria; production functions; quantity theory of money; risk sharing; Shapley–Shubik trading posts; Walras's Law

Article

An object is often said to qualify as *money* if it plays one or more of the following roles: a unit of account, a medium of exchange, a store of value. The first and third seem insufficient. The Arrow–Debreu model with prices expressed in terms of either an abstract numeraire or one of the goods is not a model of a monetary economy. Neither is every model that contains an asset or durable good. That leaves the medium-of-exchange function: an object is a medium of exchange if it appears in many transactions – in the sense of a Clower (1967) transaction matrix.

As regards kinds of money, one distinction is between outside money, such as gold coins, and inside (private sector) money, such as demand deposits. (The quantity of outside money is unaffected by consolidation over the balance sheets of everyone in the economy, while the quantity of inside money disappears when that consolidation is performed – an inside money being someone's asset and someone else's liability.) Among outside monies, a distinction is usually made between commodity and fiat money. A commodity money is an object that has intrinsic value as a consumption good or as an input, while a fiat money does not.

One challenge is to construct models that depict the ancient notion that a medium of exchange is beneficial. (This notion goes back at least to the Roman jurist Paulus who said: ‘Since occasions where two persons can just satisfy each other's desires are rarely met, a material was chosen to serve as a general medium of exchange’ –Monroe, 1966.) Another is to construct models in which media of exchange are relatively poor stores of value, have low rates of return. And accompanying those challenges is a wide range of related policy questions. How, if at all, should inside money be regulated? How should a government monopoly on outside money be managed? Should there be country-specific outside monies?

Progress in meeting those challenges and in addressing policy questions has come about by taking seriously some old ideas: money is helpful when there are absence-of-double-coincidence difficulties that cannot be easily overcome with credit; and a good money has some desirable physical properties – recognizability, portability, and divisibility. In order to better appreciate the challenges and the progress, it is helpful to review the history of monetary theory.

The classical dichotomy

At the beginning of the 20th century, the dominant economic theory was a two-part model: a rudimentary Arrow–Debreu theory of relative prices and allocations; and a quantity-theory equation that was often interpreted as a supply-equals-demand for money equation. As was widely recognized, this model suffers from a blatant inconsistency. Everybody in the model is completely described in the theory of relative prices and allocations. Who, then, holds money, which is not one of goods in the relative price-allocation part of model? Patinkin (1951) called attention to this inconsistency by pointing out that the model fails to satisfy Walras's Law.

The model has other defects. Because the model does not describe transactions, it is silent about whether money is a medium of exchange. Whether it is or not, money is not helpful in the model because allocations are determined exactly as they would be in its absence. And, as was widely recognized, the real return on money in the model – determined entirely by the time path of the stock of money and its effect on the time path of the price level – could be less than, equal to, or greater than the real interest rate determined in the relative price part of the model. The third possibility was viewed as problematic because people would then, presumably, hold only money.

Notice, by the way, that money in the above model is implicitly fiat money and that holdings of it are minimized subject to being able to carry out transactions. Neither was an obvious feature of the economies to which the theory was applied for centuries. For most of that time, money was in fact a commodity and one that may not have been a poor store of value – if only because few alternatives were available. The distinction between commodity and fiat money may not be important because for some specifications of the intrinsic value of commodity money, the value of commodity money is determined

in the same way as the value of fiat money (see, for example, Samuelson, 1968; Sargent and Wallace, 1983). The implicit assumption that money is a poor store of value is more significant because it means that money cannot be treated as an ordinary asset.

Real balances in utility or production functions

The first models to overcome the blatant inconsistency of the classical dichotomy and to, in some way, integrate value and monetary theory were models of fiat money in which its quantity and its price were arguments of utility or production functions (see Samuelson, 1961). Such models are consistent with individual endowments of money and have equilibria in which it has value.

The models were intended to overcome the inconsistency of the classical dichotomy, while preserving as much of the relative price part of the model as possible. However, not everything was preserved. After explaining why real balances, not nominal balances, are introduced as an additional argument of utility functions, Samuelson (1961, p. 119) says, ‘This is not the only case in which economists have found it necessary to introduce prices into the indifference loci; there is also the example of goods which have snob appeal, or scarcity appeal…’ Samuelson (1968) describes the welfare consequences of his formulation: the failure of the first welfare theorem. That failure should not be surprising; putting prices into utility or production functions is a back-door way of introducing externalities. The failure gave rise to the vast literature on the so-called Friedman rule: tax to support the payment of interest on money either explicitly or through deflation.

A desirable feature of these models is that money cannot have a higher pecuniary real return than other assets. The models treat real balances like clothing or refrigerators. Such assets throw off services and, therefore, in equilibrium have lower pecuniary rates of return than assets like bonds that do not throw off services.

Cash in advance and trading posts

Utility or production functions with real balances as arguments were always regarded as *indirect* functions. If so, then there ought to be a direct or underlying model. One suggestion for the underlying model is a model in which the Arrow–Debreu budget set is replaced by separate sets which insure that money will appear in many trades (see Clower, 1967). Some goods can be purchased only with money and the sellers of those goods who receive money can use that money only in subsequent trades. Such models, dubbed cash-in-advance models, are special cases of models of *incomplete markets* (see, for example, Magill and Quinzii, 2006).

Viewed that way, cash-in-advance models depart from the Arrow–Debreu model by amending its equilibrium concept. Shubik (1973) adopts that way of modelling money, but insists that trade be modelled as an explicit game. In particular, he suggests that it be modelled using what are called Shapley–Shubik trading posts, with each post defined by the pair of objects traded at the post. In static versions of that model in which the game is modelled as the simultaneous choice of quantities (a version of a Cournot quantity game), inactivity at any subset of posts (including all posts) is a Nash equilibrium. Such inactivity has been used as a rationale for selecting a subset of posts that produces the kind of transaction matrix we observe – for example, some goods cannot be traded for anything other than

money and, in a multi-country context, some goods can be traded only for home money. However, Krishna (2005) questions the robustness of shutting down posts in which goods trade for assets that dominate money in rate of return.

Starr and Stinchcombe (1999) use a version of this model with fixed costs of operating a post to suggest that scale economies can imply that the efficient arrangement of posts when there are $n+1$ objects, *n* goods and money is a monetary structure: at each of *n* active posts, money trades for one of the goods. Howitt (2005) uses an infinite-horizon version of that model with utility-maximizing agents who operate the posts to argue that there can be equilibria with that monetary structure of posts.

Imperfect monitoring and money

A different approach to modelling money is to depart from the environment of the Arrow–Debreu model – in particular, from its assumptions about commitment and information. Implicit in the absence-of-double-coincidence rationale for money is that the two persons cannot commit to future actions and are strangers. After all, a student in a class is more likely to say to a neighbouring student ‘lend me a pencil’ than ‘sell me a pencil’. More generally, in order that absence of double coincidence be a basis for a beneficial role for money, it must be augmented by no-commitment and by informational assumptions that inhibit the use of credit in its most general sense –informational assumptions that in game theory are called *imperfect monitoring*.

One of the first discussions of the informational assumptions is in Ostroy (1973). Townsend (1989) uses imperfect monitoring in an explicit intertemporal model and Kocherlakota (1998) further formalizes it. This work treats fiat money as a mechanism whose only role is to provide evidence of previous actions that would otherwise not be known. Fiat money, a physical object, can play that role because, counterfeiting aside, others can say ‘show me’ if one tries to overstate ones holdings of it.

The potentially crucial role of imperfect monitoring can be illustrated by considering the well-known risk-sharing model in Green (1987) and the variant of it studied by Levine (1990). There is a non-atomic measure of people who have identical preferences and maximize expected discounted utility. The model is one of pure exchange with a single good at each date. At each date, each person receives an endowment realization from a two-point set (high or low), where realizations are i.i.d. among people at a date and over time and are private information. Green studies a version of this model with perfect monitoring: at each date, each person makes a report about the person's endowment realization, a report which in the future is associated with that person.

Levine (1990) studies a variant of this model, but assumes no monitoring at all. In his version, no announcement or action made by a person at a date is associated with that person in the future. Moreover, if endowments are treated as owned by individuals, then under Levine's assumption, there is a role for money even if endowment realizations are public information. If there is no way to remember in the future that a person with a high endowment surrendered some of it, then the person will not surrender it – except for something that the person can carry into the future. In a pure-exchange setting, that thing can only be fiat money.

Pairwise meetings

Absence of double coincidence is almost always described in terms of meetings between two people. That, of course, is very different from having everyone together or at least connected as in the Arrow–Debreu model. But, if the role of such pairwise meetings is only to prevent quid pro quo trade in commodities, then it is unnecessary. Such trade cannot happen in Green (1987), even in deterministic versions of it. So why bother with models of pairwise meetings?

One reason is that Paulus and others were reporting what they were seeing: namely, exchanges between two people. Another reason to study such models is to investigate their implications for transactions. Kiyotaki and Wright (1989) are the first to succeed in formulating and analysing such a model. In a world with many objects, they study the relationship between the intrinsic storage properties of objects – in particular, the (utility) cost of storing them – and their role in exchange. In order to make headway on that question, they adopt simplifying assumptions: objects are indivisible, each person can hold at most one unit of some object, and the intrinsic storage quality of an object is modelled as a utility cost which once realized does not become part of the state of the economy. Even with those simplifying assumptions, their model is not simple because the state of the economy is a distribution of holdings of the different objects. Nevertheless, they could show that there can be steady states in which objects other than the least costly-to-store object can play a medium-of-exchange role. (For the welfare properties of different equilibria in their model, see Renero, 1999.)

Still another reason for studying models with pairwise meetings is that such meetings can provide a rationale for imperfect monitoring. In a large economy, if people meet in pairs and, therefore, know only what they have experienced or what they have been told by people they meet, then imperfect monitoring emerges as an implication. This point of view is explored in non-monetary models in Kandori (1992) and in monetary models in Kocherlakota (1998) and Araujo (2004). Finally, models of pairwise meetings are attractive settings for exploring the consequences of imperfect recognizability and imperfect divisibility of money and other assets.

Models of pairwise meetings, however, also come with complications. One is the wide range of equilibrium concepts used to answer the old question: what do a pair who meet to trade do? One approach taken in the literature is descriptive – for example, the buyer and the seller make alternating offers, buyers make take-it-or-leave-it offers, or sellers commit to posted prices. Another approach explores all implementable outcomes subject either to individual defection or to such defection and cooperative defection by the pair in the meeting.

Another complication is the endogeneity of the distribution of assets. Such endogeneity also arises in models in which fiat money is the only durable object, in which people can hold more than one unit of money, and in which the meeting process gives rise to a distribution of outcomes – a person can end up buying, selling, or not trading. Obviously, in such models we do not expect to obtain simple closed-from solutions for equilibria or even steady states.

One response is to accept the endogeneity and to derive results for the model despite not having closed-form solutions (see Green and Zhou, 1998; Molico, 2006; Zhu, 2003; 2005). Another is to avoid it: by using the so-called *large-family* model (see Shi, 1997); by using a setting in which pairwise meetings alternate in some fashion with centralized meetings in which preferences are quasi-linear (see Lagos and Wright, 2005); or by using some other meeting process that lends itself to a simple or degenerate distribution of money.

Applications

New theoretical work should provide insights previously unavailable – insights about seemingly paradoxical observations or policies or both.

Outside money, credit and cashless economies

If we maintain the innocuous assumption that people cannot commit to future actions, then a model economy with perfect monitoring has no role for money, while one with no monitoring has no role for credit. Therefore, in order to find roles for both money and credit, we should study models with some, but not perfect, monitoring.

Several alternative formulations of such imperfect monitoring have been studied. Kocherlakota and Wallace (1998) use the pairwise setting in Trejos and Wright (1995) and Shi (1995), and assume that there is a lag in updating the public record of individual actions. They show that the set of implementable allocations is larger the shorter the lag – an obvious result, but one that represents the sense in which technological improvements that allow better monitoring improve trade outcomes. Cavalcanti and Wallace (1999) use the same background model, but assume that some people are perfectly monitored and others not at all. They permit each person to issue perfectly recognizable durable objects that are specific to the person, objects that are best interpreted as transferable trade-credit instruments. They show that the set of implementable outcomes in which such instruments are not valued (or are prohibited) is a strict subset of those in which such instruments issued by monitored people are valued. (Kocherlakota, 2002, shows that there is a way to support efficient allocations in such models using only spot trade with money. However, his punishment scheme would not survive allowing either the defector or the non-defector to move first in a meeting.) Aiyagari and Williamson (2000) use an environment that is close to that of Green (1987), but assume that a report to the planner can be made with some probability less than 1. Their focus is on how competitive trade in money influences what the planner can achieve.

Obviously, limiting cases of the above formulations of imperfect monitoring give rise to what can be interpreted as cashless economies. Although there are many conceptions of cashless economies, one of which is the Arrow–Debreu model, the above formulations have the desirable property that the cashless limit is a limit of a cash economy in which a medium of exchange plays a beneficial role. Moreover, because the cashless economy is achieved by taking a limit with respect to monitoring while maintaining the no-commitment assumption, the cashless limit is not an Arrow-Debreu model.

In Cavalcanti and Wallace (1999) and Cavalcanti, Erosa and Temzelides (1999), the money issued by monitored people is used by and passed around among nonmonitored people. Wallace and Zhu (2007) use that idea to offer a new interpretation of the paradox concerning banknote issue pointed out by Friedman and Schwartz (1963). Toward the end of 19th century, many countries permitted banks to issue payable-to-the-bearer notes subject to redemption on demand and to collateral restrictions. In the United States and, presumably, in other countries, those systems seemed to give rise to a failure of an arbitrage condition: the yields on eligible collateral often seemed too high to be reconciled with their use as collateral for note issue. Put differently, those systems seemed not to produce currencies that were elastic with respect to the yield on eligible collateral. The explanation offered by Wallace and Zhu has

two components. First, the profitability of note issue depends on the implied float. Second, note issuers face a menu of opportunities for issuing notes – a menu that displays an inverse association between the magnitude of possible note placements and the implied float. The paradox results from treating the observed float as if it applied to all possible uses of notes, rather than taking into account the fact that high-placement low-float opportunities – for example, in organized financial markets – are not chosen. In Wallace and Zhu, the low-placement, high-float opportunities are in pairwise meetings.

Physical properties of assets

Discussions of money have often described desirable physical properties of media of exchange: recognizability, portability and divisibility. Implicit in any such discussion is the idea that those properties are scarce, are not shared equally by all objects. However, only recently have the consequences of such scarcity been explored.

Recognizability

Freeman (1985) and Williamson and Wright (1994) use imperfect recognizability of alternatives to fiat money to produce models in which fiat money is helpful. In Freeman, the alternative to fiat money is a claim to long-lived capital. Under the assumption that such claims can be costlessly counterfeited, he argues that genuine claims cannot be traded competitively. Williamson and Wright use a model of pairwise matching *without* an absence-of-double-coincidence problem to show that imperfect recognizability of the (durable) goods is enough to make trade involving fiat money helpful.

In both of those models and many others, the holder of an asset knows more about it than at least some potential holders. (An exception is Huggett and Krasa, 1996.) Models of pairwise meetings are attractive for studying the role of such imperfect recognizability because it is in such meetings, rather than in ‘large markets’, that asymmetric information about quality ought to be important. Moreover, if, as in Freeman or Williamson and Wright, the low-quality asset is worthless, then it gets traded when subject to asymmetric information only if it masquerades as being genuine – that is, only in a *pooling* equilibrium.

However, pooling equilibria do not always exist – at least if refinements on beliefs about off-equilibrium actions are imposed. It remains to be determined whether such refinements could be used to strengthen the Freeman result. In particular, could a small difference in counterfeiting costs between two assets – between fiat money and claims to capital, or home money and foreign money, or outside money and inside money – be enough to generate trade in one of the assets and no trade in the other even if the less-costly-to-counterfeit asset, as in Freeman, has a large rate-of-return advantage?

Portability

Townsend (1989) and Smith (2002) build models based on portability of fiat money and the lack of portability of capital. However, as they emphasize, the mere lack of portability of real capital needs to be supplemented by imperfect monitoring. And when supplemented by sufficiently imperfect monitoring, such models give rise to a role for fiat money that is very similar to its role in other absence-of-double-

coincidence settings.

To see the similarity, consider a version of those models in which people meet in pairs and in which there is one good per date. When two people meet, suppose that they have available to them some amount of the good that can either be consumed or used as an input (investment) that will give output at the next date, but only at the same location. Moreover, suppose that one and only one of the two people will be at the same location at the next date. If there is no monitoring, then fiat money, despite having a lower real return than investment, can have a beneficial role – the same role it has in other absence-of-double-coincidence settings with no monitoring. That is, the stayer retains all the capital, while the leaver takes some fiat money. The absence of monitoring prevents the leaver from retaining a claim to any of the capital.

Divisibility

Historians of monetary systems and others have often noted that money was generally not available in conveniently small denominations (see, for example, Redish, 2000; Sargent and Velde, 2002). However, until recently no models described how such absence would inhibit trade. Models of pairwise meetings are an obvious candidate: if neither the buyer nor the seller has small change, then trade (even if lotteries are permitted) is inhibited. If the model is to have implications for optimal divisibility, then it should also contain something to limit divisibility. Lee, Wallace and Zhu (2005) assume that there is a direct cost of carrying monetary items that is independent of denomination (that is, carrying thousands of pennies is very costly), while Lee and Wallace (2006) assume costs of producing and maintaining the stock of money that increase with divisibility.

Concluding remarks

Why is it better to make assumptions about meeting patterns, information, and the physical characteristics of potential assets than about which markets are open or the pattern of transaction costs over objects? First, the former lends itself to standard notions of incentive feasibility, which is what we ought to mean by integrating monetary economics with the rest of economics. Second, such an approach meets the proof-of-the-pudding criterion. Compare, for example, the results about inside money that can be obtained by working with the imperfect-monitoring point of view with what can be done with a cash-in-advance model.

But is such foundational work needed to deal with the nuts and bolts of monetary policy? It is generally agreed that open-market operations matter because the medium of exchange is a low-return asset and because the central bank has a monopoly on its supply. Can it be that beneficial management of that monopoly does not depend on how we explain the low return of the medium of exchange?

Finally, can we look forward to a monetary theory that in generality rivals the Arrow–Debreu model? Probably not. A need for a medium of exchange does not arise in every conceivable economy – think of Robinson Crusoe, even after he meets Friday, or of the Arrow–Debreu model. Such a need arises when there is some absence-of-double-coincidence difficulty that cannot be overcome with credit because people cannot commit to future actions and because there is imperfect monitoring. Those features may not lend themselves to a general formulation.

See Also

- currency competition
- inside and outside money
- money
- money and general equilibrium

Bibliography

Aiyagari, S. and Williamson, S. 2000. Money and dynamic credit arrangements with private information. *Journal of Economic Theory* 91, 248–79.

Araujo, L. 2004. Social norms and money. *Journal of Monetary Economics* 51, 241–56.

Cavalcanti, R., Erosa, A. and Temzelides, T. 1999. Private money and reserve management in a random matching model. *Journal of Political Economy* 107, 929–45.

Cavalcanti, R. and Wallace, N. 1999. Inside and outside money as alternative media of exchange. *Journal of Money, Credit and Banking* 31, 443–57.

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6, 1–8.

Freeman, S. 1985. Transaction costs and the optimal quantity of money. *Journal of Political Economy* 93, 146–57.

Friedman, M. and Schwartz, A. 1963. *A Monetary History of the United States*. Princeton: Princeton University Press.

Green, E. 1987. Lending and the smoothing of uninsurable income. In *Contractual Arrangements for Intertemporal Trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

Green, E. and Zhou, R. 1998. A rudimentary random-matching model with divisible money and prices. *Journal of Economic Theory* 81, 252–71.

Howitt, P. 2005. Beyond search: fiat money in organized exchange. *International Economic Review* 46, 405–29.

Huggett, M. and Krasa, S. 1996. Money and storage in a differential information economy. *Economic Theory* 8, 191–210.

Kandori, M. 1992. Social norms and community enforcement. *Review of Economic Studies* 59, 63–80.

Kiyotaki, N. and Wright, R. 1989. On money as a medium of exchange. *Journal of Political Economy* 97, 927–54.

Kocherlakota, N. 1998. Money is memory. *Journal of Economic Theory* 81, 232–51.

Kocherlakota, N. 2002. The two-money theorem. *International Economic Review* 43, 333–46.

Kocherlakota, N. and Wallace, N. 1998. Optimal allocations with incomplete record-keeping and no-commitment. *Journal of Economic Theory* 81, 272–89.

Krishna, R.V. 2005. Non-robustness of the cash-in-advance equilibrium in the trading-post model. *Economics Bulletin* 5, 1–5.

Lagos, R. and Wright, R. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 113, 463–84.

Lee, M. and Wallace, N. 2006. Optimal divisibility of money when money is costly to produce. *Review of Economic Dynamics* 9, 541–56.

Lee, M., Wallace, N. and Zhu, T. 2005. Modeling denomination structures. *Econometrica* 73, 949–60.

Levine, D. 1990. Asset trading mechanisms and expansionary policy. *Journal of Economic Theory* 54, 148–64.

Magill, M. and Quinzii, M. 2006. *Theory of Incomplete Markets, Volume I*. Cambridge, MA: MIT Press.

Molico, M. 2006. The distribution of money and prices in search equilibrium. *International Economic Review* 47, 701–22.

Monroe, A. 1966. *Monetary Theory Before Adam Smith*. New York: Kelley.

Ostroy, J. 1973. The informational efficiency of monetary exchange. *American Economic Review* 63, 597–610.

Patinkin, D. 1951. The invalidity of classical monetary theory. *Econometrica* 19, 134–51.

Redish, A. 2000. *Bimetallism: An Economic and Historical Analysis*. Cambridge: Cambridge University Press.

Renero, J. 1999. Does and should a commodity medium of exchange have relatively low storage costs? *International Economic Review* 40, 251–64.

Samuelson, P. 1961. *Foundations of Economic Analysis*. Cambridge: Harvard University Press.

Samuelson, P. 1968. What classical and neoclassical monetary theory really was. *Canadian Journal of Economics* 1, 1–15.

Sargent, T. and Velde, F. 2002. *The Big Problem of Small Change*. Princeton: Princeton University Press.

Sargent, T. and Wallace, N. 1983. A model of commodity money. *Journal of Monetary Economics* 12, 163–87.

Shi, S. 1995. Money and prices: a model of search and bargaining. *Journal of Economic Theory* 67, 467–98.

Shi, S. 1997. A divisible search model of money. *Econometrica* 65, 75–102.

Shubik, M. 1973. Commodity money, oligopoly, credit and bankruptcy in a general equilibrium model. *Western Economic Journal* 11, 24–38.

Smith, B. 2002. Taking intermediation seriously. *Journal of Money, Credit and Banking* 35, 1319–58.

Starr, R. and Stinchcombe, M. 1999. Exchange in a network of trading posts. In *Markets, Information and Uncertainty: Essays in Economic Theory in Honor of Kenneth J. Arrow*, ed. G. Chichilnisky. Cambridge: Cambridge University Press.

Townsend, R. 1989. Currency and credit in a private information economy. *Journal of Political Economy* 97, 1323–44.

Trejos, A. and Wright, R. 1995. Search, bargaining, money and prices. *Journal of Political Economy* 103, 118–41.

Wallace, N. and Zhu, T. 2007. Float on a note. *Journal of Monetary Economics* 54, 229–246.

Williamson, S. and Wright, R. 1994. Barter and monetary exchange under private information. *American Economic Review* 84, 104–23.

Zhu, T. 2003. Existence of a monetary steady state in a matching model: indivisible money. *Journal of Economic Theory* 112, 307–24.

Zhu, T. 2005. Existence of a monetary steady state in a matching model: divisible money. *Journal of Economic Theory* 123, 135–60.

How to cite this article

Wallace, Neil. "fiat money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000059> doi:10.1057/9780230226203.0563

The New Palgrave Dictionary of Economics Online

fiducial inference

D.A.S. Fraser

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Fiducial inference introduced the pivotal inversion that is central to modern confidence theory. Initially this provided confidence bounds but later was generalized to give confidence distributions on the parameter space. For this it came in direct conflict with the then prominent Bayesian approach called inverse probability. Confidence distributions are now however widespread in modern likelihood theory. Recent results from this theory indicate that the developed fiducial confidence approach is giving a consistent statement of where the parameter is with respect to the data, and indeed is consistent with recent Bayesian approaches that allow data dependent priors.

Keywords

Bayesian inference; confidence theory; fiducial inference; frequentist school; inverse probability; likelihood; Markov chain Monte Carlo methods

Article

In a seminal paper, R.A. Fisher (1930) introduced the notion of fiducial inference as an alternative to what was then called inverse probability. The key step in fiducial inference is pivotal inversion, which is now standard in all of confidence theory. Fisher's example involved four pairs of observations with a concern for the correlation coefficient ρ between observations in a pair. He had available the distribution function $F(r;\ \rho)$ for the sample correlation coefficient r, which depends only on the population correlation ρ; and he had an observed correlation value $r^0 = .99$. He did numerical calculations with the distribution function $F(r;\ \rho)$, which he had himself previously derived. And he then reported (.765, 1) as a 95 per cent interval for ρ. This is fully in accord with current confidence interval theory. In present notation we would write

$$P(r < .99; \rho) = .95 = P(\hat{\rho}_L < \rho; \rho) = P\{\rho \text{ in } (\hat{\rho}_L, 1); \rho\},$$

where the solution of $F(r; \rho) = .95$ for ρ to obtain the parameter lower bound $\hat{\rho}_L = \hat{\rho}_L(r)$ is standard confidence or pivotal inversion applied to the pivot $u = F(r; \rho)$, which of course has a Uniform (0,1) distribution.

But Fisher (for example, 1930; 1933; 1935; 1956) went further and presented a distribution, called a fiducial distribution, for the parameter ρ, which as a density can be used for calculations such as

$$\int_{.765}^{1} f_{\text{fid}}(\rho; r^0)\, d\rho = .95,$$

and where for the example the density has the form

$$f_{\text{fid}}(\rho; r) = -(\partial / \partial \rho) F(r; \rho);$$

this density agrees with what in recent likelihood theory would be called a confidence distribution.

But Fisher went still further and spoke of fiducial probability rather than just statements for an interval such as confidence level that we would commonly use. This attribution of probability that a parameter lies in the interval (.765, 1) attracted attack from both the inverse probability community at the time and from the more conventional community that would now be called the frequentist, and includes those having philosophical persuasions. As a consequence, many have viewed fiducial probability as wrong, and strong stigmata have been attached to it. This is rather extraordinary, given that the papers by Fisher are seminal for all of confidence theory and differ only in small deviations of presentation and development.

The key aspects of fiducial that evoked criticism are (*a*) that different pivots can lead to different distributions and thus different intervals, (*b*) that marginalization of a parameter distribution to a component parameter can give a distribution that depends on data in a way different from the obvious that would come from that data, and (*c*) that constraints on the parameter can give a distribution without total probability being equal to 1.

The alternative culture when Fisher (1930) introduced fiducial inference was inverse probability (Bayes, 1763). For this, the probability at a data point y^0, given as $f(y^0; \theta)$ and now called likelihood (Fisher, 1922) and written $L(\theta; y^0)$, is adjusted by a weight function $w(\theta)$ to give the composite

$$w(\theta)L(\theta;\ y^0)$$

which is then treated as an unnormed density for the parameter. The weight function $w(\theta)$ is chosen based on properties of the model and called by various names, with default prior being the most unassuming. The present rather large community using this approach is a subgroup of the Bayesian community and the approach has come to be called default Bayesian inference rather than inverse probability analysis; it can also be viewed as a routine frequentist use of the frequentist likelihood function coupled with an ad hoc weight function.

This commonly called default Bayesian approach offers great freedom for the development of statistical techniques: take an observed likelihood $L(\theta; y^0)$ based on Fisher's (1922) proposal; attach a convenient weight function $w(\theta)$ to it; and use the composite for inference for θ. With available high-powered computers and Markov Chain Monte Carlo this leads to a wealth of possible analyses, in contrast to rather limited results from earlier frequentist approaches.

But this leads to perhaps the most influential criticism of the fiducial method (Lindley, 1958): (*d*) that a fiducial distribution is typically not an inverse probability or default Bayesian posterior.

Curiously, one finds that the default Bayesian approach is subject to precisely the same criticisms (*a*), (*b*), (*c*) that have been attached to the fiducial approach (for example, concerning (*b*), see Dawid, Stone and Zidek, 1973; see also Fraser, 1961; 1995). So the fact (d) that a fiducial analysis is not in general a default Bayesian analysis seems a rather hollow criticism by Lindley (1958). And of course default Bayes typically does not lead to intervals that have the confidence property. Moreover, a recently dominant interest within the current Bayesian community (Fraser and Reid, 2002) is to have methods that do reproduce in repeated sampling as do confidence intervals. Perhaps the default Bayesian community is rushing in where the frequentist community neglected its own likelihood function.

But perhaps Fisher and his fiducial approach should be given credit for the fundamental contribution of the pivotal inversion, and of giving rise to the universal confidence procedures. The change of name from fiducial to confidence and then the derogation of fiducial seem a rather heavy historical penalty to Fisher and his profound and seminal developments in statistics. Perhaps ‘fiducial’ did move too quickly, certainly for the times, and did neglect to develop some fine details. But the results are profound; and the default Bayesian community is finding that it cannot ignore in substance the fiducial criticisms (*a*), (*b*), (*c*); and can't avoid the repeated sampling reproducibility that is the foundation of confidence theory (*d*).

But then, how does fiducial inference work in more general contexts, particularly in the light of recent likelihood theory? For each independent coordinate, say, y_i, a pivot $z_i = h_i(y_i;\ \theta)$ is needed that describes with full deference to continuity how the coordinate y_i measures or provides information on the parameter θ; this pivot needs to be of the same dimension as the variable y_i and of course as implied by its name has a fixed distribution free of θ. If a coordinate is scalar, the pivot is necessarily equivalent to the distribution function $F_i(y_i;\ \theta)$ for that coordinate; if it is vector then the choice of pivot represents an explicit statement of how that coordinate variable affects the parameter and is taken as a given for the inference process.

Likelihood theory then shows that the full pivot can be re-expressed to third-order accuracy in the moderate deviations region by an equivalent pivot in which the parameter θ of, say, dimension *p*

appears in only *p* coordinates of the new pivot. The conditional distribution of these *p* coordinates given the remaining pivot coordinates (which are of course directly observable) gives effectively a new pivot with of course the same dimension as the parameter. This allows for the standard confidence pivotal inversion to produce confidence regions.

If inference focuses on a particular parameter component $\psi(\theta)$ of interest with dimension *d*, then the recent likelihood theory shows that the interest parameter can be isolated to third order in a *d* dimensional component of an equivalent pivot, and the marginal model for that pivot is otherwise free of the full parameter and provides third-order confidence regions for the interest parameter. For some background see Fraser and Reid (2001), Fraser, Reid and Wu (1999), and Fraser (2004).

See Also

- empirical likelihood
- Fisher, Ronald Aylmer
- maximum likelihood

Bibliography

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London* 53, 370–418. Reprinted in *Biometrika* 45 (1958), 293–315.

Dawid, A., Stone, M. and Zidek, J. 1973. Marginalization paradoxes in Bayesian and Structural Inference. *Journal of the Royal Statistical Society B* 35, 189–233.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* 222, 309–68.

Fisher, R. 1930. Inverse probability. *Proceedings of the Cambridge Philosophical Society* 26, 528–35.

Fisher, R. 1933. The concept of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Statistical Society* A 139, 343–8.

Fisher, R. 1935. The fiducial argument in statistical inference. *Annals of Eugenics* 6, 391–8.

Fisher, R. 1956. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fraser, D. 1961. The fiducial method and invariance, *Biometrika* 48, 261–80.

Fraser, D. 1995. Some remarks on pivotal models and the fiducial arguments in relation to structural models. *International Statistical Review* 64, 231–6.

Fraser, D. 2004. Ancillaries and conditional inference, with discussion. *Statistical Science* 19, 333–69.

Fraser, D. and Reid, N. 2001. Ancillary information for statistical inference. In *Empirical Bayes and Likelihood Inference*, ed. S. Ahmed and N. Reid. New York: Springer-Verlag.

Fraser, D. and Reid, N. 2002. Strong matching of frequentist and Bayesian inference. *Journal of Statistical Planning and Inference* 103, 263–85.

Fraser, D., Reid, N. and Wu, J. 1999. A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* 86, 249–64.

Lindley, D. 1958. Fiducial distribution and Bayes' theorem. *Journal of the Royal Statistical Society B* 20, 102–7.

Neyman, J. 1941. Fiducial argument and the theory of confidence intervals. *Biometrika* 32, 128–50.

How to cite this article

Fraser, D. A. S. "fiducial inference." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000064> doi:10.1057/9780230226203.0564

The New Palgrave Dictionary of Economics Online

field experiments

John A. List and David Reiley

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Field experiments have grown significantly in prominence since the 1990s. In this article, we provide a summary of the major types of field experiments, explore their uses, and describe a few examples. We show how field experiments can be used for both positive and normative purposes within economics. We also discuss more generally why data collection is useful in science, and more narrowly discuss the question of generalizability. In this regard, we envision field experiments playing a classic role in helping investigators learn about the behavioural principles that are shared across different domains.

Keywords

charitable giving; field experiments; generalizability; laboratory experiments; matching funds; testing; uniform-price auctions; Vickrey auctions

Article

Field experiments occupy an important middle ground between laboratory experiments and naturally occurring field data. The underlying idea behind most field experiments is to make use of randomization in an environment that captures important characteristics of the real world. Distinct from traditional empirical economics, field experiments provide an advantage by permitting the researcher to create exogenous variation in the variables of interest, allowing us to establish causality rather than mere correlation. In relation to a laboratory experiment, a field experiment potentially gives up some of the control that a laboratory experimenter may have over her environment in exchange for increased realism. The distinction between the laboratory and the field is much more important in the social sciences and the life sciences than it is in the physical sciences. In physics, for example, it appears that every hydrogen atom behaves exactly alike. Thus, when astronomers find hydrogen's signature wavelengths of light coming from the Andromeda Galaxy, they use this information to infer the quantity of hydrogen present there. By contrast, living creatures are much more complex than atoms and molecules, and they correspondingly behave much more heterogeneously. Despite the use of ‘representative consumer’

models, we know that not all consumers purchase the same bundle of goods when they face the same prices. With complex, heterogeneous behaviour, it is important to sample populations drawn from many different domains – both in the laboratory and in the field. This permits stronger inference, and one can also provide an important test of generalizability, testing whether laboratory results continue to hold in the chosen field environment.

We find an apt analogy in the study of pharmaceuticals, where randomized experiments scientifically evaluate new drugs to treat human diseases. Laboratory experiments evaluate whether drugs have desirable biochemical effects on tissues and proteins *in vitro*. If a drug appears promising, it is next tested *in vivo* on several species of animals, to see whether it is absorbed by the relevant tissues, whether it produces the desired effects on the body, and whether it produces undesirable side effects. If it remains with significant promise after those tests, it is then tested in human clinical trials to explore efficacy and measure any side effects.

Even after being tested thoroughly in human clinical trials and approved by regulators, a drug may sometimes reveal new information in large-scale use. For example, *effectiveness* may be different from the *efficacy* measured in clinical trials: if a drug must be taken frequently, for example, patients may not remember to take it as often as they are supposed to or as often as they did in closely supervised clinical trials. Furthermore, rare side effects may show up when the drug is finally exposed to a large population.

Much like this stylized example, in economics there are a number of reasons why insights gained in one environment might not perfectly map to another. Field experiments can lend insights into this question (see also Bohm, 1972; Harrison and List, 2004; Levitt and List, 2007; List, 2007). First, different types of subjects might behave differently; university students in the laboratory might not exhibit the same behaviour as financial traders or shopkeepers. In particular, the people who undertake a given economic activity have selected into that activity and market forces might have changed the composition of players as well; you might expect regular bidders to have more skill and interest in auctions than a randomly selected laboratory subject, for example.

A second reason why a field experiment might differ from a laboratory experiment is that the laboratory environment might not be fully representative of the field environment. For example, a typical donor asked to give money to charity might behave quite differently if asked to participate by choosing how much money to contribute to the public fund in a public-goods game (List, 2007). The charitable-giving context could provide familiar cognitive cues that make the task easier than an unfamiliar laboratory task. Even the mere fact of knowing that one's behaviour is being monitored, recorded, and subsequently scrutinized might alter choices (Orne, 1962).

Perhaps most important is the fact that any theory is an approximation of reality. In the laboratory, experimenters usually impose all the structural modelling assumptions of a theory (induced preferences, trading institutions, order of moves in a game) and examine whether subjects behave as predicted by the model. In a field experiment, one accepts the actual preferences and institutions used in the real world, jointly testing both the structural assumptions (such as the nature of values for a good) and the behavioural assumptions (such as Nash equilibrium).

For example, Vickrey (1961) assumes that in an auction there is a fixed, known number of bidders who have valuations for the good drawn independently from the same (known) probability distribution. He uses these assumptions, along with the assumption of a risk-neutral Nash equilibrium, to derive the ‘revenue equivalence’ result: that Dutch, English, first-price, and second-price auctions all yield the same expected revenue. However, in the real world the number of bidders might actually vary with the

good or the auction rules, and the bidders might not know the probability distribution of values. These exceptions do not mean that the model should be abandoned as ‘wrong’; it might well still have predictive power if it is a reasonable approximation to the truth. In a field experiment (such as Lucking-Reiley, 1999, for this example), we approach the real world; we do not take the structural assumptions of a theory for granted.

Such an example raises the natural question related to the actual difference between laboratory and field experiments. Harrison and List (2004) propose six factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the subjects operate. Using these factors, they discuss a broad classification scheme that helps to organize one's thoughts about the factors that might be important when moving from the laboratory to the field.

A first useful departure from laboratory experiments using student subjects is simply to use ‘non-standard’ subjects, or experimental participants from the market of interest. Harrison and List (2004) adopt the term ‘artefactual’ field experiment to denote such studies. While one might argue that such studies are not ‘field’ in any way, for consistency of discussion we denote such experiments as artefactual field experiments for the remainder of this article, since they do depart in a potentially important manner from typical laboratory studies. This type of controlled experiment represents a useful type of exploration beyond traditional laboratory studies.

Moving closer to how naturally occurring data are generated, Harrison and List (2004) denote a ‘framed field experiment’ as the same as an artefactual field experiment but with field context in the commodity, task, stakes, or information set of the subjects. This type of experiment is important in the sense that a myriad of factors might influence behaviour, and by progressing slowly towards the environment of ultimate interest one can learn about whether, and to what extent, such factors influence behaviour in a case-by-case basis.

Finally, a ‘natural field experiment’ is the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are participants in an experiment. Such an exercise represents an approach that combines the most attractive elements of the laboratory and naturally occurring data – randomization and realism. In this sense, comparing behaviour across natural and framed field experiments permits crisp insights into whether the experimental proclamation, in and of itself, influences behaviour.

Several examples of each of these types of field experiments are included in List (2006). Importantly for our purposes, each of these field experimental types represents a distinct manner in which to generate data. As List (2006) illustrates, these field experiment types fill an important hole between laboratory experiments and empirical exercises that make use of naturally occurring data. Yet an infrequently discussed question is: why do we bother to collect data in economics, or in any science?

First, we use data to collect enough facts to help construct a theory. Several prominent broader examples illustrate this point. After observing the anatomical and behavioural similarities of reptiles, one may theorize that reptiles are more closely related to each other than they are to mammals on the evolutionary tree. Watson and Crick used data from Rosalind Franklin's X-ray diffraction experiment to construct a theory of the chemical structure of DNA. Careful observations of the motions of the planets in the sky led Kepler to theorize that planets (including Earth) all travel in elliptical orbits around the Sun, and

Newton to theorize the inverse-square law of gravitation. After observing with a powerful telescope that the fuzzy patches called ‘spiral nebulae’ are really made up of many stars, one may theorize that our solar system is itself part of its own galaxy, and the spiral nebulae are external to our Milky Way galaxy. Robert Boyle experimented with different pressures using his vacuum pump in order to infer the inverse relationship between the pressure and the volume of a gas. Rutherford's experiments of shooting charged particles at a piece of gold foil led him to theorize that atoms have massive, positively charged nuclei.

Second, we use data to test theories’ predictions. Galileo experimented with balls rolling down inclined planes in order to test his theory that all objects have the same rate of acceleration due to gravity. Pasteur rejected the theory of spontaneous generation with an experiment that showed that microorganisms grow in boiled nutrient broth when exposed to the air, but not when exposed to carefully filtered air. Arthur Eddington measured the bending of starlight by the sun during an eclipse in order to test Einstein's theory of general relativity.

Third, we use data to make measurements of key parameters. On the assumption that the electron is the smallest unit of electric charge, Robert Millikan experimented with tiny, falling droplets of oil to measure the charge of the electron. On the assumption that radioactive carbon-14 decays at a constant rate, archaeologists have been able to provide dates for various ancient artifacts. Similarly, scientists have assumed theory to be true and designed careful measurements of many other parameters, such as the speed of light, the gravitational constant, and various atomic masses.

Field experiments can be a useful tool for each of these purposes. For example, Anderson and Simester (2003) collect facts useful for constructing a theory about consumer reactions to nine-dollar endings on prices. They explore the effects of different price endings by conducting a natural field experiment with a retail catalogue merchant. Randomly selected customers receive one of three catalogue versions that show different prices for the same product. Systematically changing a product's price varies the presence or absence of a nine-dollar price ending. For example, a cotton dress may be offered to all consumers, but at prices of 34, 39, and 44 dollars, respectively, in each catalogue version. They find a positive effect of a nine-dollar price on quantity demanded, large enough that a price of 39 dollars actually produced higher quantities than a price of 34 dollars. Their results reject the theory that consumers turn a price of 34 dollars into 30 dollars by either truncation or rounding. This finding provides empirical evidence on an interesting topic and demonstrates the need for a better theory of how consumers process price endings.

List and Lucking-Reiley (2000) present an example of a framed field experiment designed to test a theory. The theory of multi-unit auctions predicts that a uniform-price sealed-bid auction will produce bids that are less than fully demand-revealing, because such bids might lower the price paid by the same bidder on another unit. By contrast, the generalized Vickrey auction predicts that bidders will submit bids equal to their values. In the experiment, List and Lucking-Reiley conduct two-person, two-unit auctions for collectible sportscards at a card trading show. The uniform-price auction awards both items to the winning bidder(s) at an amount equal to the third-highest bid (out of four total bids), while the Vickrey auction awards the items to the winning bidder(s) for amounts equal to the bids that they displaced from winning. List and Lucking-Reiley find that, as predicted by the theory of demand reduction, the second-unit bids submitted by each bidder were lower in the uniform-price treatment than in the Vickrey treatment. The first-unit bids were predicted to be equal across treatments, but in the experiment they find that the first-unit bids were anomalously higher in the uniform-price treatment. Subsequent laboratory experiments (see, for example, Engelmann and Grimm, 2003; Porter and Vragov,

2003), have confirmed this finding.

Finally, Karlan and List (2007) is an example of a natural field experiment designed to measure key parameters of a theory. In their study, they explore the effects of 'price' changes on charitable giving by soliciting contributions from more than 50,000 supporters of a liberal organization. They randomize subjects into several different groups to explore whether solicitees respond to upfront monies used as matching funds. They find that simply announcing that a match is available considerably increases the revenue per solicitation – by 19 per cent. In addition, the match offer significantly increases the probability that an individual donates – by 22 per cent. Yet, while the match treatments relative to a control group increase the probability of donating, larger match ratios – 3:1 dollars (that is, 3 dollars match for every 1 dollar donated) and 2:1 dollar – relative to smaller match ratios (1:1 dollar) have no additional impact.

In closing, we believe that field experiments will continue to grow in popularity as scholars continue to take advantage of the settings where economic phenomena present themselves. This growth will lead to fruitful avenues, both theoretical and empirical, but it is clear that regardless of the increase in popularity, the various empirical approaches should be thought of as strong complements, and combining insights from each of the methodologies will permit economists to develop a deeper understanding of our science.

See Also

- experimental economics
- experimental labour economics
- experimental methods in environmental economics
- experimental methods in economics
- experiments and econometrics
- experimental economics, history of

Bibliography

Anderson, E.T. and Simester, D. 2003. Effects of $9 price endings on retail sales: evidence from field experiments. *Quantitative Marketing and Economics* 1, 93–110.

Bohm, P. 1972. Estimating the demand for public goods: an experiment. *European Economic Review* 3, 111–30.

Engelmann, D. and Grimm, V. 2003. Bidding behavior in multi-unit auctions—an experimental investigation and some theoretical insights. Working paper, Centre for Economic Research and Graduate Education, Economic Institute, Prague.

Harrison, G.W. and List, J.A. 2004. Field experiments. *Journal of Economic Literature* 42, 1009–55.

Karlan, D. and List, J.A. 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, forthcoming.

Levitt, S.D. and List, J.A. 2006. What do laboratory experiments measuring social preferences tell us about the real world? *Journal of Economic Perspectives* 21(2), 153–74.

List, J.A. 2006. Field experiments: a bridge between lab and naturally occurring data. *Advances in Economic Analysis & Policy* 6(2), Article 8. Abstract online. Available at: http://www.bepress.com/bejeap/advances/vol6/iss2/art8, accessed 26 May 2007.

List, J.A. and Lucking-Reiley, D. 2000. Demand reduction in a multi-unit auction: evidence from a sportscard field experiment. *American Economic Review* 90, 961–72.

Lucking-Reiley, D. 1999. Using field experiments to test equivalence between auction formats: magic on the Internet. *American Economic Review* 89, 1063–80.

Orne, M.T. 1962. On the social psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist* 17, 776–83.

Porter, D. and Vragov, R. 2003. An experimental examination of demand reduction in multi-unit versions of the uniform-price, Vickrey, and English auctions. Working paper, Interdisciplinary Center for Economic Science, George Mason University.

Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16, 8–37.

How to cite this article

List, John A. and David Reiley. "field experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000305> doi:10.1057/9780230226203.0565

The New Palgrave Dictionary of Economics Online

finance (new developments)

Jiang Wang
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This survey of some of the developments in finance since the mid-1980s begins with advances in the application of arbitrage pricing, and then expands into areas of general asset pricing under the title ‘risk and return’. Limitations in our current understanding of risk as well as more data explorations have led to the ‘discovery’ of anomalies, which challenge classic notions of market efficiency. We examine recent attempts to expand the neoclassical framework to incorporate market imperfections in asset pricing, which, in their more general forms, take centre stage in advances in corporate finance.

Keywords

adverse selection; agency problems; anchoring; anomalies; arbitrage; arbitrage pricing; asymmetric information; bankruptcy; behavioural finance; belief perseverance; bonds; capital asset pricing model; capital structure; consumption-based capital asset pricing model; contagion; control rights; corporate control; corporate finance; corporate governance; cost of capital; debt and equity; derivatives; efficient market hypothesis; equilibrium; equity premium puzzle; expected utility; finance; finance (new developments); human capital; incomplete contracts; interest rates; intertemporal capital asset pricing model; inventory cost; irreversible investment; jumps; law of one price; liquidity; market frictions; market microstructure; mean-variance efficient portfolios; moral hazard; no arbitrage; optimal contracts; option pricing; overconfidence; private information; probability; reduced form models; risk; risk aversion; risk premium; serial correlation; short-sale constraints; speculative bubbles; state-dependent preferences; stochastic discount factor; stochastic volatility models; structural models; substitutable securities; takeovers; trading costs; transaction costs; uncertainty; venture capital

Article

This article attempts to survey some of the developments in finance since the mid-1980s. By then, what

we know as neoclassical finance had taken its broad shape and become a foundation for our understanding of the central issues in finance as well as the starting point for further developments. As true for any science, the advances in finance since then are characterized by an interactive process of more rigorous testing and revision of the existing theories, more extensive exploration of the data, old and new, and further expansion of theory beyond the known territories.

Finance is concerned with how the financial market facilitates the allocation of capital or assets. In a well-functioning market, it is prices that guide the allocation. Thus, how to value financial assets or financial securities is a primary focus of finance. Since the value of an asset comes from its future payoff, which extends over time and is uncertain in nature, risk is the key element in asset valuation. Relying on a few basic principles, neoclassical finance has developed a rich set of models and tools for asset pricing, risk analysis and corporate finance.

However, much of the neoclassical finance abstracts away from market imperfections, most notably information asymmetry and market frictions. Such an abstraction draws the boundaries of the neoclassical theory. In particular, the theory's applicability softens when these imperfections are important. Moreover, the theory itself does not provide much guidance in gauging the relevance of imperfections. The omission of imperfections also leaves the neoclassic theory mostly free of institutions. Such an simplification is perhaps most stark in corporate finance, as the very existence and consequently the behaviour of firms are presumably institutional arrangements in response to imperfections, but it is similarly striking in the context of financial market, which consists of a complex and collection of institutions and intermediaries. Naturally, this limits what the neoclassical theory says about the behaviour of institutions as major participants in the market and its implications on market behaviour and capital allocation. It should be emphasized that significance of imperfections in a given context is as much, if not more, of an empirical issue as a theoretical matter. To a large extent, developments beyond the neoclassical theory involve very much the interplay between the research in these two dimensions.

Our discussion of these developments will begin with the advances in the application of arbitrage pricing, arguably the most successful area of modern finance. It then expands into areas of general asset pricing under the title ‘risk and return’. Limitations in our current understanding of risk as well as more data explorations have led to the ‘discovery’ of anomalies, which amounted to new challenges to classic notions of market efficiency. After reviewing some of this empirical evidence, we examine some of the recent attempts to expand the neoclassical framework to incorporate market imperfections. We then turn to advances in corporate finance, in which imperfections, in their more general forms, take centre stage.

This article is intended to provide an update to finance, which remains a timeless piece in capturing the spirit and the essence of neoclassical finance. We will rely on it for a more detailed review of the earlier work as well as their historical context. In order to make the two articles more integrated, we adopt a similar framework, but adjusted to reflect the current landscape. Needless to say, any survey of a broad and fast-growing field such as finance will be partial and incomplete.

Arbitrage pricing

The first principle of asset pricing is the absence of arbitrage. Here, an arbitrage refers to a set of transactions in the market based on public information that always yields net gains. The intuition for no arbitrage is straightforward: any such an opportunity would be exploited by market participants until it

disappears. When transactions of financial securities face little frictions, no arbitrage yields sharp results on the prices of securities which are close substitutes. In particular, securities with same payoffs must have the same price. Classical applications of this principle include arbitrage relations between the prices of default-free bonds with different coupons, spot and forward prices of commodities, and spot and forward exchange rates between currencies and their corresponding interest rates. The path-breaking work of Black and Scholes (1973) and Merton (1973) on option pricing greatly expanded the applications of arbitrage pricing. Vasicek (1977) and Cox, Ingersoll and Ross (1985b) demonstrated how the arbitrage method can be applied to the pricing of default-free bonds. Merton (1974) and Black and Cox (1976) applied the option pricing technique to value bonds with default risks.

Arbitrage pricing as a general methodology enjoyed unprecedented success in finance and in all of economics. Not only did earlier empirical tests find strong support from the data (see, for example, Black and Scholes, 1973), but data converged to theory as deviations disappeared quickly with the theory's dissemination. The explosive applications of the methodology by the financial industry, ranging from new products and markets, new pricing and trading technologies, to new investment and risk management practices, gave the core substance for what is now branded as financial engineering.

For a set of traded securities, let P_t denote their current prices and X_{t+1} their next period payoffs (in vector forms). In its general formulation by Ross (1976) and Harrison and Kreps (1979), the absence of arbitrage is equivalent to the existence of a set of positive state prices ϕ such that

$$P_t = \sum_{\omega} \varphi_t(\omega) X_{t+1}(\omega) = E_t^* [e^{-r_{0,t+1}} X_{t+1}] \tag{1}$$

where ω denotes a future state of the economy, $\phi_t(\omega)$ the state price at t for state ω at t+1, $r_{0,t+1}$ denotes the risk-free interest rate from t to t+1, and $E_t^*[\cdot]$ denotes the conditional expectation using normalized stated prices as probabilities, which is also referred to as the risk-neutral measure. For a set of securities with payoffs determined by the same set of future states ω, they become substitutes. Their prices will then be related by arbitrage through the corresponding state prices. If we can identify a sufficient number of such securities, then their prices will reveal the corresponding state prices, which then allow us to value other substitutes. Certain securities are natural substitutes, such as an underlying asset and its derivatives, the whole collection of fixed income securities, and a firm's bonds and its equity. Arbitrage pricing has been widely used in the valuation of these securities.

Equity options

The basic framework for option valuation was established by the pioneer work of Black and Scholes (1973) and Merton (1973) and the contributions of Cox and Ross (1976a; 1976b) and Cox, Ross and Rubinstein (1979). The huge body of work that followed has enriched this framework substantially. One

focus is to allow for more general price behaviour for the underlying asset. This is in part motivated by deviations in the observed prices from the Black–Scholes formula, which assumes a geometric Brownian diffusion for the price of the underlying asset. For example, from the observed option prices, the volatility implied from the Black–Scholes formula changes over time and differs for different strike prices, a phenomenon referred to as volatility smiles or smirks. Two natural extensions are to allow stochastic volatility and jumps. Hull and White (1987), Heston (1993) and Stein and Stein (1991) proposed models with time-varying volatility. Extending Merton (1976), Amin (1993), Scott (1997) and Bates (2000) have incorporated jumps into stochastic volatility models. Empirical analysis of the data on both options and underlying equities suggests that both stochastic volatility and jumps are helpful in explain their behaviour (see, for example, Melino and Turnbull, 1990; Bates, 1996; Bakshi, Cao and Chen, 1997; Pan, 2002). In a discrete-time setting, the distinction between stochastic volatility and jumps becomes moot. Rubinstein (1994) has suggested modifications to binomial model of Cox, Ross and Rubinstein (1979) to accommodate the effects of time-varying price dynamics.

Although most of the recent work on equity options stays within the neoclassical arbitrage pricing framework, it significantly enriches the pricing models to better fit the data. But the fit is never perfect. Is the gap eventually going to be closed with more sophisticated models or revealing something more? We are not totally sure. The arbitrage approaches works when options are truly substitutes of the underlying asset. If so, why do they appear in the first place? Market imperfections may be part of the reason, but the exact nature of this link is far from being well understood.

Default-free bonds

Bonds of no default risk are closely related to each other as their prices are all driven by interest rates. From (1), the price of a pure discount bond, which has a unit payoff at date T, is given by

$$B_t(T) = \mathrm{E}_t^*\left[e^{-\int_t^T r_{0,s}ds}\right]. \tag{2}$$

The specification of the interest rate process under the risk-neutral measure will then allow us to price bonds by computing the above conditional expectation. It is well known that bond returns share a small number of common factors (for example, Litterman and Scheinkman, 1991). A natural approach is to specify the interest rate as a function of a few state variables. Earlier work chooses the short-term interest rate as the single state variable and assumes tractable dynamics. For example, Vasicek (1977) assumes a Gaussian Markov process for the short rate and Cox, Ingersoll and Ross (1985b) assume a square-root process (the CIR model). Other candidate models include Brennan and Schwartz (1979), Cox, Ingersoll and Ross (1980), Courtadon (1982), and more recently Longstaff (1989), Chan et al.

(1992), Constantinides (1992) and Ahn, Dittmar and Gallant (2002). These models aim at tractable solutions to bond prices to capture their basic behaviour.

The empirical evaluation of these models requires the further specification of r_t under the statistical measure, that is, the true data-generating process. The transformation from the risk-neutral measure to the statistical measure effectively reflects the risk premium. Thus, the test of an arbitrage-based pricing model is really a joint test of the proposed interest rate process and the associated risk premium process. Analysis by Brown and Dybvig (1986), Chan et al. (1992) and Gibbons and Ramaswamy (1993) readily demonstrates that the parsimony of single factor models also limits their ability to fit the data. Nonparametric tests, such as Ait-Sahalia (1996) and Stanton (1997), further suggest that single-factor diffusion models are unable to capture several important features of interest rate dynamics. Following Langetieg (1980), Cox, Ingersoll and Ross (1985b) and Schaefer and Schwartz (1984), multifactor extensions became the next pursuit, notably Chen and Scott (1992), Longstaff and Schwartz (1992) and Hull and White (1994).

Despite their added flexibility, multifactor models face two challenges. On the one hand, they quickly become less tractable as more state variables are added. On the other hand, even though a small number of factors, typically three, capture a large percentage of commonality in bond returns, it is far from clear if they are enough in describing bond prices.

The first challenge has largely limited the focus to tractable models. One notable example is the so-called affine models, in which the short-term interest rate is assumed to be an affine function of a set of state variables. In addition, the vector of state variables follows a diffusion process with its drift and covariance both being its own affine functions. Closed-form solutions can be obtained for bond prices and yields under this specification. Brown and Schaeffer (1994), for example, explored an extension of the CIR model under the affine structure. Duffie and Kan (1996) expanded the scope of the affine models and Dai and Singleton (2000) provided an extensive empirical analysis of their pros and cons. They can capture many aspects of the bond price behaviour, but always leaving a few others. This situation is not unique to the affine models as it is shared by other multifactor models outside the affine class (for example, Ahn, Dittmar and Gallant, 2002). Other enrichments have also been considered, such as jumps in interest rates (for example, Johannes, 2004; Piazzesi, 2005) and regime shifts in interest rate dynamics (for example, Hamilton, 1988; Gray, 1996), to enhance the descriptive power of the arbitrage-based models.

It might be unrealistic to attempt to describe the rich behaviour of a large cross-section of bond prices by a small number of risk factors following relatively simple dynamics. One possibility is to relax the limit on the dimension of risk factors. Kennedy (1994), Goldstein (2000) and Santa-Clara and Sornette (2001) have explored models with an infinite-dimensional state vector. However, to make these models empirically implementable, restrictive structures need to be imposed.

Aside from specific modelling issues, what we confront is a more general situation. To the extent that default-free bonds are substitutes, we can rely on arbitrage methods in their valuation. While being very close, they are rarely exact substitutes. From this perspective, arbitrage results provide only an approximation. We can take comfort in the empirical success of existing models so far – the bottle is not full, but close to it. But to fill the rest is proven hard. It suggests that old tricks may be inefficient and something new is at play.

Defaultable bonds

As Black and Scholes (1973) and Merton (1973) observed, a firm's securities, its bonds and equity, all share the same risk, the risk of its asset value. Given the firm's value and its dynamics, we can then value its bonds in the same way as we value equity options. Merton (1974) and Black and Cox (1976) are among the first set of so-called structural models, which rely on the specific risk structure of corporate bonds with respect to its underlying asset to value them using arbitrage methods. More comprehensive models were further developed to incorporate richer risk structures embedded in corporate bonds. For example, Longstaff and Schwartz (1995) and Saa-Requejo and Santa-Clara (1999) allow interest rate risks, which can affect when a bond defaults and its value then. Richer debt structure, the cost of default and shareholders optimal financing and default choices are also considered by Leland (1994), Anderson and Sundaresan (1996), and Leland and Toft (1996), among others.

Structural models, coupled with specific description of the firm value dynamics, also impose certain behaviour on the event of default. The desire to have more flexibility in modelling the default event has led to the development of so-call reduced form models, for example, Jarrow and Turnbull (1995), Lando (1998) and Duffie and Singleton (1999). These models start by modelling the default process and recovery rate under the risk-neutral measure. From (1), the price of a defaultable bond with zero coupon and unit par value is given by

$$P_t(T) = E_t^*\left[e^{-\int_t^T r_{0,s}ds} 1_{\{\tau > T\}}\right] + E_t^*\left[e^{-\int_t^T r_{0,s}ds} Z_\tau 1_{\{\tau \le T\}}\right] \quad (3)$$

where τ denotes the default time, Z_τ the recovery rate in the case of default and $1_{\{.\}}$ is an indicator function. The implementation and evaluation of the reduced form models requires additional information on default events under both the risk-neutral and the statistical measures. Such information is scarce as default is relatively infrequent. However, these models become more applicable for credit derivatives when more securities related to the same default events became available.

Other derivatives

The application of arbitrage pricing finds its most fertile ground in valuing derivatives, which, together with the underlying asset, are close substitutes. Its success has fuelled the big bang of derivatives since the 1970s, which provided new areas for more applications of the theory. Commodity and financial futures, interest rate derivatives such as swaps, caps and swaptions, credit derivatives such as credit default swaps (CDS) and collateralized debt obligations (CDO) are major examples.

While no arbitrage as a theoretical principle is quite general, its usefulness in asset valuation remains a

practical matter. It applies to financial securities which are close substitutes, when market frictions are negligible. Its success in asset pricing as opposed to in pricing physical goods very much reflects the fact that financial securities are easy to trade and replicate and they are close in nature in the sense that their value all arises from their financial payoffs. Nonetheless, market frictions do exist. Moreover, except in certain instances, securities of interest are often not exact substitutes. The theory is much less definitive about how it extends to these situations.

Its limitations are evident even in areas where it was successful, such as equity options, bonds and other derivatives. Nagging deviations from arbitrage-based models, though arguably small, persist. The need to address these deviations tends to push for more complex models with more risk factors. However, a blind pursuit in this process might be losing not only the empirical ground, as data becomes less sufficient in supporting the models, but also the theoretical ground. The deviations may well reflect the influence of market frictions or other factors, which are beyond the ‘limits’ of arbitrage arguments.

Another limitation of the arbitrage approach, perhaps a more fundamental one, is that it takes the risks and their risk premia, that is, the relevant states and the corresponding state prices, as given in establishing price relations among substitutable securities. From a broader perspective, it is important to understand the economic underpinnings of different risks and their pricing implications. Such an understanding may provide the basis to further improve arbitrage pricing models. More importantly, it will allow us to value assets more broadly.

Risk and return

The broader principle in asset pricing is market equilibrium, which requires that security prices must equate supply and demand. This approach is general since it focuses on how security prices are determined by economic fundamentals, which drive supply and demand. Its application, however, faces several challenges. We need to first specify what constitute the fundamentals. We also need to determine how these fundamentals influence the supply and demand for securities and ultimately their prices. Additional structure on the fundamentals is also needed before we can arrive at useful results.

A key fundamental is the risk characteristics of asset payoffs. We start with the pricing equation (1). Suppose that given the state of the economy at t, the conditional probability for state ω at t+1 is $p_t(\omega)$. We can then rewrite (1) as

$$P_t = \sum_{\omega} P_t(\omega) \frac{\varphi_t(\omega)}{P_t(\omega)} X_{t+1} = E_t[m_{t+1} X_{t+1}], \tag{4}$$

where E_t is the expectation under the actual probability measure given the state at t and $m_{t+1} \equiv \phi_t(\omega)/p_t(\omega)$ is referred to as state price density or the stochastic discount factor (SDF). Realizing that $X_{t+1}/P_t \equiv 1 + r_{t+1}$ where R_{t+1} is the security's return from t to t+1, we can also re-express (4) as

$$1 = E_t[m_{t+1}(1 + r_{t+1})]. \tag{5}$$

A slight variation of (5) gives a commonly used expression for the pricing equation (4):

$$E_t[r_{t+1}] - r_{0,t+1} = -\mathrm{Cov}_t[m_{t+1} / E_t[m_{t+1}], r_{t+1} - r_{0,t+1}], \tag{6}$$

where Cov_t denotes the conditional covariance. Equation (6) suggests that we can decompose an asset's risk into two components:

$$r_{t+1} - r_{0,t+1} - (E_t[r_{t+1}] - r_{0,t+1}) = a_t(1 - m_{t+1}) / E_t[m_{t+1}] + u_{t+1}. \tag{7}$$

The first component, which is correlated with the SDF, will influence the asset's expected return or its price. The second component, which is uncorrelated, does not. Such a decomposition clearly reveals that risks come in two types, ‘priced’ and ‘not priced’. The amount of an asset's priced risk is measured by a_t, its covariance with the SDF, which determines its risk premium:

$$E_t[r_{t+1}] - r_{0,t+1} = a_t\sigma^2_{m,t}, \tag{8}$$

where $\sigma^2_{m,t}$ is the conditional variance of the SDF. Knowing the SDF will allow us to specify the priced risk and its premium. Thus, the goal for a general asset pricing theory is then to determine the SDF. Two differently approaches have been followed in developing models for the SDF. The first approach is to start from the primitives of the economy such as asset payoffs and investors’ preferences, derive the asset demand (and supply) and finally arrive at the equilibrium SDF. The second approach is to start from the equilibrium, rely on certain properties of the equilibrium to arrive at the SDF. The first approach has more micro-texture to it and often leads to sharper specifications of the SDF, while the second approach allows more flexibility but with less microeconomic basis.

Factor models of the SDF

The well-known capital asset pricing model (CAPM) follows the first approach. Under the mean-variance framework of Markowitz (1952) and Tobin (1958), all investors will hold mean-variance efficient portfolios. Such a commonality in investors' asset demand implies that, in equilibrium, the market portfolio, which represents the total supply of assets, must be a mean-variance efficient portfolio. This insight has led Sharpe (1964) and Lintner (1965) to identify the return on the market portfolio $r_{M,t+1}$ to be a proxy for the SDF:

$$1 - m_{t+1} / \mathrm{E}_t[m_{t+1}] = b_{M,t}(r_{M,t+1} - \mathrm{E}_t[r_{M,t+1}]),$$

which immediately leads to the CAPM:

$$\mathrm{E}_t[r_{i,t+1}] - r_{0,t+1} = \beta_{iM,t}(\mathrm{E}_t[r_{M,t+1}] - r_{0,t+1}), \tag{9}$$

where $\beta_{iM,t}$ is the market beta for asset *i*. Ross (1976a; 1976b) started directly from the risk structure of asset payoffs. By proposing a linear factor model for asset returns and requiring the absence of limiting arbitrages as an equilibrium condition, he obtained a factor representation for the SDF:

$$1 - m_{t+1} / \mathrm{E}_t[m_{t+1}] = \sum_{k=1}^{K} b_{k,t} f_{k,t+1}, \tag{10}$$

which leads to the arbitrage pricing theory (APT):

$$\mathrm{E}_t[r_{i,t+1}] - r_{0,t+1} = \sum_{k=1}^{K} \beta_{ik,t} \lambda_{k,t}, \tag{11}$$

where $\beta_{ik,t} = \sigma_{ik,t} / \sigma^2_{k,t}$ is the 'beta' of asset *i* on risk factor *k* (that is, the regression coefficient of its return on $f_{k,t+1}$), $\sigma^2_{k,t}$ the conditional variance of factor *k*, $\sigma_{ik,t}$, is its conditional covariance with the return of asset *i*, and $\lambda_{k,t}=b_{k,t}\sigma^2_{k,t}$ its risk premium. Thus, both the CAPM and the APT can be viewed as the case when the SDF has a linear factor structure. The key distinction is that the CAPM identifies the market return as the SDF while the APT allows the SDF to be spanned by multiple factors.

Earlier empirical tests found some supporting evidence for the CAPM (see, for example, Black, Jensen and Scholes, 1972; Fama and MacBeth, 1973). Due to the noise in estimating expected returns and the difficulty to actually identify the market portfolio, questions regarding the strength of the support as well as the nature of these tests have left a strong need for more tests (see, for example, Roll, 1977). Further empirical exploration also reveals evidence that is at odds with the CAPM, at least on the face of it. Banz (1981) discovered the size effect that small stocks (measured by market capitalization) yield higher average returns than large stocks, after controlling for what the CAPM predicts. Basu (1983) reported the value effect that stocks with book values higher than their market values – the book-to-market ratio – yield higher average returns than stocks with lower ratios. In a comprehensive empirical analysis, Fama and French (1992) synthesized this evidence, demonstrating the weak explanatory power of the CAMP for the cross-section of stock returns as well as certain patterns they display.

The APT allows a richer structure than the CAPM. However, the fact that the theory itself does not identify the factors poses challenges to its empirical testing (see, for example, Shanken, 1982; Dybvig and Ross, 1985). Other means need to be used, theoretical or statistical, to identify the factors. Earlier empirical tests along this route find some supportive evidence (for example, Chen, Roll and Ross, 1986) but leaves more to be desired. Relying on statistical analysis, Connor and Korajczyk (1988) use principal components and Lehmann and Modest (1988) use factor analysis to empirically identify the factors. The evidence in support of the APT is, however, mixed. Exposures to the empirically identified factor risks explain only part of the cross-sectional variation in average returns. Based on the observed average returns, Fama and French (1993) propose to use firm characteristics such as size and book-to-market ratio to form portfolios, whose returns are then used to identify risk factors in addition to the market. They show that the size factor (the difference in returns from small and large stocks), the value factor (the difference in returns from high and low book-to-market stocks), plus a broad market index can explain most of the cross-sectional variation in returns from portfolios sorted on by their loadings.

What to take away from the empirical results remains a matter of active discussion. The appeal of the CAPM has led to continuous efforts to reconcile it with the empirical evidence. All empirical implementations of the CAPM use a market index as a proxy for the market portfolio, which leaves the possibility of misidentifications. The lack of significant gains from improving market proxies based on traded assets, as Stambaugh (1982) and Shanken (1987) have shown, has led to the inclusion of non-traded assets. Using labour income growth to measure the return on human capital, Jagannathan and Wang (1996) showed that including human capital in the market proxy may help to increase the explanatory power of the CAPM.

In general the CAPM gives a conditional relation between risk, as measured by an asset's conditional beta, and its conditional risk premium, as (9) clearly indicates. However, earlier tests are mostly

unconditional, looking at the relation between assets' unconditional beta and their unconditional premia. By allowing time-varying betas and the market premium, more tests are directed at the conditional version of the CAPM, notably Harvey (1989), Shanken (1990), Jagannathan and Wang (1996), and, more recently, Wang (2003) and Petkova and Zhang (2005). How far the added flexibility from the conditional variables and their impact can lead us remains to be seen, as Lewellen and Nagel (2006) demonstrate. More fundamental questions exist about the test of conditional CAPM. For example, what are the appropriate conditioning variables implied by the model? Without knowing this, how do we distinguish the impact of conditioning variables from additional risk factors?

The added flexibility in the empirical multi-factor models like that of Fama and French (1992; 1993) also leave plenty of room for alternative interpretations. It is open to potential dangers of data snooping, as Lo and MacKinlay (1990) and Kothari, Shanken and Sloan (1995) point out. It can also be a result of misidentification of the true risk factor even when the conditional CAPM holds (see, for example, Berk, Green and Naik, 1999; Gomes, Kogan and Zhang, 2003).

The intertemporal CAPM

Merton (1973) developed an intertemporal version of the CAPM (ICAPM), which shows that time-varying market conditions give rise to dynamic risk factors in addition to the risk of the market portfolio. In particular, if we let the first factor in (10) be the return on the market portfolio, the other factors will represent the state variables driving the market conditions. The ICAPM contains the conditional CAPM as the special case when the dynamic risks carry no premium. In this sense, any test of the conditional CAPM can be viewed as a test of the ICAPM with additional restrictions on risk premia. However, in the ICAPM, the dynamic risk factors are taken as given, not derived from the theory. The pricing relation, in the form of (11), comes as an equilibrium condition under a given form of price dynamics rather than an equilibrium outcome in terms of economic primitives. In this regard, the ICAPM has more in common with APT than with the classical CAPM.

The ICAPM provides a useful framework for analysing risk and return in an intertemporal setting. Its empirical implementation has been limited until recently. Tests of the APT, which also allows multiple risk factors, were often interpreted as tests of the ICAPM, since both models allow plenty of flexibility in identifying these factors. However, such a view leaves out the additional implications from the ICAPM on the intertemporal properties of the dynamic risks. Lo and Wang (2006) have developed a version of the ICAPM in which the time-varying market risk premium captures the dynamic risk. Using the cross-sectional data on trading volume, they empirically identify the dynamic risk factor and test its power in explaining the time series and the cross section of asset returns. Ang et al. (2006) and Adrian and Rosenberg (2006) also test a version of the ICAPM in which the market volatility is a dynamic risk factor.

By identifying risk factors other than the market proxy as dynamics risks, the ICAPM gives more guidance on the properties of these additional risk factors, in particular, their correlation structure with the underlying state variables. These properties may help the empirical construction of these factors and their tests. More work along this direction is called for.

Consumption-based CAPM

The APT and the ICAPM focus on the statistical properties of risks, in particular, their correlation structure, both in cross-section and time series. But from a pricing perspective, their economic properties are particularly important. Clearly, investors' attitude towards different risks also matter. For an investor, an asset is riskier if its return co-varies more strongly with his future marginal utility. Using the same setting as Merton's ICAPM, Breeden (1979) showed that for pricing purposes, all risks can be collapsed into one, measured by assets' covariance with aggregate consumption (see also Rubinstein, 1976). As a result, the beta of an asset with respect to aggregate consumption, the consumption-beta, determines its risk premium. This is the so-called consumption-based CAPM (CCAPM).

Let us start with the principle of market equilibrium, namely, asset prices must equate demand and supply. Since demand and supply for assets are determined by the fundamentals through market participants' optimizing behaviour, so will be their equilibrium prices. Consider a representative investor in the market who has a time-separable, expected utility function $u(c_t)+\rho\, u(c_{t+1})$. The optimality of his asset holdings requires that $u'(c_t)P_t=E_t[\rho\, u'(c_{t+1})X_{t+1}]$ *or*

$$1 = E_t\left[\frac{\rho u'(c_{t+1})}{u'(c_t)}(1 + r_{t+1})\right]. \tag{12}$$

Comparing (12) with (5), it is apparent that market equilibrium imposes additional restrictions on asset prices. In particular, it relates the stochastic discount factor to the marginal utilities of the representative investor: $m_{t+1}=\rho\, u'(c_{t+1})/u'(c_t)$.

The simple structure of the CCAPM and its economic appeal has generated a lot of interest in its empirical implementation, lead by Breeden (1980), Grossman and Shiller (1981) and Hansen and Singleton (1983). The focus has been on two fronts, the behaviour of aggregate prices and the cross section of asset returns. The procedure typically involves an estimation of the consumption process and certain specification of the marginal utility function for the representative investor. Combining the two gives an estimate for the SDF, which can then be used to test its pricing implications.

From (7) and (8), Hansen and Jagannathan (1991) established the following condition on the SDF:

$$\sigma_t(m_{t+1}) / E_t[m_{t+1}] \geq \frac{E_t[r_{i,t+1}] - r_{0,t+1}}{\sigma_t(r_{i,t+1} - r_{0,t+1})}, \tag{13}$$

where i denotes any traded asset. The right-hand side is asset i's Sharpe ratio. Thus, the maximum Sharpe ratio of traded assets gives a lower bound for the volatility of the SDF. In the CCAPM, the volatility of the SDF comes from the volatility of aggregate consumption. However, the aggregate consumption data seems very ‘dull’, exhibiting close to i.i.d. growth with very low volatility. This poses certain challenges to the simple forms of the CCAPM. In order for the SDF to have the desired properties, the marginal utility function has to do all the work. Using a time-separable utility function with constant relative risk aversion, Mehra and Prescott (1985) show that the implied risk aversion has to be very high to yield a volatility of the SDF that exceeds the bounds posted by the Sharpe ratio of the market index. Such a high implied risk aversion seems inconsistent with other evidence on individual risk preferences, leaving us with what is referred as the ‘equity premium puzzle’. Applying the model in such a manner quickly led to many more ‘puzzles’. Real interest rates have been low, at least in developed markets, which is inconsistent with the observed consumption growth and a high risk aversion (for example, Weil, 1989). Also, the low variability in the SDF and the low observed variability in aggregate dividend growth are at odds with the high observed volatility in aggregate asset prices or asset returns (for example, LeRoy and Porter, 1981; Shiller, 1981; Campbell and Shiller, 1988).

Many efforts have been made to reconcile the volatility on the SDF implied by asset returns and the consumption data, mostly along three directions. The first is to improve on the measure of consumption risk. For example, Rietz (1988) suggests that the small probability events like severe drops in consumption may be important risks not fully captured by the data. Bansal and Yaron (2004) propose to use the variability in long-horizon consumption growth as a measure of risk. These explorations are useful, but also stretches the boundaries of the data (for example, a finite sample period will limit the length of the horizon).

The second direction is to allow for more general preferences. The simple form of utility function assumed for the representative investor in early studies is probably too restrictive. For example, even if the preferences of individual investors are restricted to simple forms – such as those exhibiting constant relative risk aversion – a certain degree of heterogeneity will lead to a more complex preference at the aggregate level, which depends on the relative importance of each investor (see, for example, Dumas, 1989; Wang, 1996; Chan and Kogan, 2002). As more flexibility is needed in fitting the data, different forms of state-dependent preferences were considered, notably Sundaresan (1989), Abel (1990), Constantinides (1990), Epstein and Zin (1991), and Campbell and Cochrane (1999), allowing factors like habit, aggregate consumption, and the timing of risk resolution to influence behaviour. The flexibility this approach enjoys also comes at certain costs. On the theory side, the aggregation properties of simple preferences are lost, which leads to questions about the link between what is assumed for the representative agent and the micro justifications used to motivate the preference structure. On the empirical side, there is a lack of discipline in identifying the true preference structure.

The third direction is to allow for certain forms of market imperfections. This approach opens up many possibilities but goes beyond the neoclassical setting. We will return to market imperfections in the next section.

The CCAPM has also been applied to explain the cross section of asset returns, as the other pricing models. From $m_{t+1} = \rho u'(c_{t+1})/u'(c_t)$, m_{t+1} can be approximated by

$$1 - m_{t+1} / E_t[m_{t+1}] \approx b_t(\Delta c_{t+1} - E_t[\Delta c_{t+1}]), \quad (14)$$

where Δc_{t+1} denotes the aggregate consumption growth. The cross section of asset returns are then given by a formula similar to (9) with the exception that the beta is now replaced by the consumption beta, the beta of assets' payoff with respect to aggregate consumption growth. Lettau and Ludvigson (2001) have implemented the CCAPM in this form and find that it can explain the cross-sectional pattern in portfolio returns as presented by Fama and French (1992; 1993). Bansal, Dittmar and Lundblad (2005) also find encouraging signs in using assets' betas of their long-run dividends with respect to the long-horizon consumption growth to explain the cross section of their returns.

The appeal of the consumption-based CAPM mainly comes from its simple economic structure. However, its validity relies on strong assumptions about the behaviour of market participants and the structure of the market. It is unclear how much the behaviour of major market participants, such as institutional investors and delegated money managers, is related to consumption. It is also unclear whether the existing market structure allows the kind of efficiency in risk allocation and the proper aggregation needed for the CCAPM. Market imperfections will cause deviations from these assumptions, which may well contribute to the challenges in fitting the model to data. On the empirical side, it is worth pointing out that estimates on risk premium and consumption risk are fairly rough.

Market efficiency and anomalies

The neoclassical theory of asset pricing relies on two simplifications, namely, frictions are negligible in financial markets and information is reasonably homogenous among market participants. While the second simplification is less relevant for arbitrage pricing, both are needed for equilibrium-based pricing models. The idea that market participants have similar information regarding future asset payoffs is closely related to the notion of financial markets being informationally efficient, a hallmark of neoclassical finance. The efficient market hypothesis (EMH) postulates that market prices fully reflect all the relevant information available in the market (see, for example, Fama, 1970). The intuition behind the hypothesis is simple, very much in the spirit of no arbitrage. Any available information that is not properly reflected in prices will be taken advantage of by profit-seeking market participants, sometimes referred to as arbitrageurs, until prices fully adjust for it.

The exact formulation of the hypothesis, however, involves important subtleties, including the precise definition of relevant information and its reflection in prices. Perhaps the simplest formulation of EMH is to assume the presence of arbitrageurs with unlimited risk tolerance and access to capital (see, for example, Samuelson, 1965). This then implies that current prices are unbiased forecasts of future prices (adjusted for time value). Event studies found broad support for prices reacting quickly and quite accurately on average to public news. Extensive tests of predictability found the evidence to be largely consistent. Nonetheless, deviations exist. A natural way to account for the deviations is to relax the condition of risk-neutrality and properly account for risks. After all, (imperfect) predictability does not

imply arbitrage, as apparent in Lucas (1978). But such an approach immediately leads us to the choice of a particular method of risk adjustment or an asset pricing model. Tests of EMH then becomes tests of the asset pricing model used, which complicates the matter substantially.

Perhaps as a reaction to the incredible success of the EMH, the initial empirical support was followed by the recording of an increasing number of exceptions, which are also referred to as anomalies. The earlier set of results is about the predictability on equity returns. DeBondt and Thaler (1985) studied long-horizon returns over three to five years. They examined two portfolios, a winner portfolio and a loser portfolio, which consist of stocks with higher or lower than market adjusted returns, respectively, and found that the winner portfolio yields lower returns in the following years and the loser portfolio yields higher returns. Fama and French (1988) and Poterba and Summers (1988) also found negative serial correlation in long-horizon market index returns. While it is harder to make inferences from long-horizon returns as the sample size becomes relatively small, Lo and MacKinlay (1988) and others have documented positive serial correlation in market index returns over weekly and monthly horizons. This evidence suggests that stock prices need not follow random walks as the weak form of the EMH claims. It was also documented that returns of large stocks can predict returns of small stocks on weekly and monthly basis, the so-called lead–lag phenomenon (see, for example, Lo and MacKinlay, 1990). Bernard and Thomas (1989), extending the earlier work by Ball and Brown (1968) and Jones and Litzenberger (1970), have presented convincing evidence of the under-reaction of stock prices to earnings announcements, which is later called the ‘post earnings announcement drift’. Many studies, such as Campbell and Shiller (1988) and Fama and French (1988), have also suggested that financial ratios such as dividend yield can predict aggregate market returns. These results are certainly at odds with the semi-strong form of the EMH, which requires no predictability of asset returns using public information. Jegadeesh and Titman (1993) take the winner–loser comparison to individual stocks over shorter horizons. By sorting stocks on returns over past one or two quarters, they show that winner portfolio continues to yield higher returns while loser portfolio yields lower returns, a phenomenon called ‘momentum’. If we take long positions in winners and offsetting short positions in losers, the average return can be substantial. This is particularly intriguing as we expect diversification and cancellation to greatly limit the net risk exposure of this strategy.

The search for predictability in stock returns has also gained momentum of its own. Different variables were found to have predictive power for equity returns such as trading volume (for example, Gervais, Kaniel and Mingelgrin, 2001), short interest (for example, Jones and Lamont, 2002), share repurchases (for example, Ikenberry, Lakonishok and Vermaelen, 1995), dispersion in analysts forecasts (for example, Diether, Malloy and Scherbina, 2002), and transactions of institutional investors (for example, Chan and Lakonishok, 1995). The list goes on and may continue to grow. However, several caveats always accompany these findings. First, their significance, both statistical and economical, is quite moderate. Second, their persistence over time needs further testing. Third, more work is desired to distinguish them from potential spurious findings due to data mining.

A corollary of the EMH is that news on future payoffs or the SDF move prices. Roll (1988) showed that *ex post* public news can only explain a fraction of price movements of individual stocks over daily to monthly horizons (see also Roll, 1984a). This result parallels what is observed at the aggregate level by LeRoy and Porter (1981) and Shiller (1981), that is, aggregate market indices exhibit a volatility much

higher than the volatility in aggregate dividends in the data. But it is more striking as risk considerations seem to be less important over short horizons. The inability to explain price movements even *ex post* has been viewed as a serious challenge to neoclassical theory, in particular the EMH. One possible explanation is that much of the price movement is driven by private news which is not captured by the information set used in the empirical studies. Another is the influence of time-varying risk which may contribute to movements in the discount factor.

To avoid the complication of risk, some studies have focused on more direct tests on the foundation of the EMH, the principle of no arbitrage. A longtime puzzle along this line is the significant discount on closed-end funds from their net asset values (see, for example, Malkiel, 1977; Lee, Shleifer and Thaler, 1991). In violation of the law of one price, anomalies of this nature document price differences between two seemingly identical assets. Other well-known examples include the price differences between the on-the-run and the off-the-run Treasury bonds with close to identical payoffs and shares of the same company with the same dividend streams but traded on different exchanges. A pair trade, to buy the security with lower price and sell the price with the higher price, which requires no private information and substantial capital but yields sure profits, seems to present an arbitrage opportunity. Obviously, the persistent existence of these price anomalies suggests that there is more to what meets the eye. For example, Ross (2002) has shown that management fees contribute significantly to the close-end discounts.

How to interpret the empirical anomalies, assuming their presence, requires further assessment. In the whole, as the name suggests, anomalies do not overweight the vast positive evidence in support of the EMH. Additional factors also need to be included in the consideration. First, predictable patterns in returns or deviations from the law of one price documented in the data are not equivalent to actual profitable opportunities in the market. Frictions in the market need to be taken into account (see, for example, Tuckman and Vila, 1992). Second, strategies attempted at taking advantage of these anomalies always involve certain risks in the presence of frictions. The dynamic nature of these risks make them harder to assess (see, for example, Merton, 1981; Dybvig and Ross, 1985).

Nonetheless, these anomalies, together with deviations in asset returns from neoclassical asset pricing models, do pose a challenge to our understanding of how the market works and how asset prices are determined. It is clear that the notion of market efficiency need to be examined in an equilibrium asset pricing framework, which allows for information asymmetry (and possibly market frictions). Grossman (1976) and Grossman and Stiglitz (1980) demonstrated that such a framework is much richer than the simple form of market efficiency implies, for both the behaviour of asset prices and the importance of information asymmetry in determining it. However, many of the implications of this framework needed to be fleshed out, which became a fertile ground for recent work.

Market imperfections

Limitations of the neoclassic theory have led to efforts to incorporate imperfections into our analysis, in particular, frictions and asymmetric information. Imperfections influences how the market operates at two levels. At a superficial level, imperfections directly affect why and how investors trade in the market, which ultimately determine asset prices. At a more fundamental level, imperfections also determine the institutional structure of the market itself as well as the economic characteristics of major

market participants, both of which also contribute to the actual imperfections observed in the market. Although efforts have been made in analysing imperfections at both levels, more of them focused on the former.

Market frictions

Despite the relative ease of transactions in the financial market, frictions exist. They range from simple trading costs such as commissions and bid–ask spreads to price impact, costs and restrictions on short sales, constraints on borrowing to simple inability to trade-certain claims or contracts. They also include the costs of setting up trading operations, gathering and processing information, maintaining market presence and the costs of introducing a new security, creating and maintaining a market and providing liquidity for it. Since frictions hinder the efficient allocation of capital in the market, their impact is closely related to the notion of liquidity or illiquidity.

Factoring in market frictions sheds new light on the empirical anomalies. Many of them do not provide profitable trading opportunities when trading costs are included. For example, Krishnamurthy (2002) finds that costs in financing the arbitrage between the on-the-run and off-the-run bonds are substantial and outweigh the potential gains. Lesmond, Schill and Zhou (2004), among others, show that momentum in individual stocks is not profitable after adjusting for trading costs. These results are comforting for the EMH and in many ways not surprising. But they do not settle all the questions. In particular, why are these patterns there in the first place, and how do they fit into the overall asset pricing framework?

The general impact of market frictions on asset prices is hard to analyse as they make the behaviour of market participants, the interaction among them and the equilibrium outcome very complex. Recent studies have mainly focused on how specific frictions such as transactions costs, short-sale constraints and borrowing constraints may influence three aspects of asset prices, the overall level, the cross section and dynamics.

Relying on partial equilibrium arguments, earlier work has examined how transactions costs may influence the level of asset prices. For example, Constantinides (1986) considered the equivalent price adjustment to offset the welfare loss from proportional transactions costs and found that its magnitude is of higher order of the cost and quantitatively insignificant. Amihud and Mendelson (1986) calibrated the present value of implied transactions costs using observed stock trading volume and showed it to be substantial. What is not fully incorporated in the partial equilibrium analysis is how costs affect the actual equilibrium. Vayanos (1998) considered a general equilibrium model in which investors trade for life-cycle reasons and reached the same conclusion as Constantinides. This is not surprising since life-cycle considerations generate little trading and consequently transactions costs have a limited effect. When high levels of trading are needed, as observed in the market, the situation can be different. Unable to trade frequently, investors have to bear additional risks they could otherwise unload in the market, which can significantly alter their behaviour. Allowing high frequency trading needs compatible with observed volume, Lo, Mamaysky and Wang (2004) show that moderate fixed transactions costs can have a significant impact on investors' asset demand and the resulting equilibrium prices.

In the context of consumption-based CAPM, incorporating market frictions can potentially help to reconcile a high-risk premium with a smooth consumption path (see, for example, He and Modest, 1995;

Luttmer, 1999). The mechanism is quite straightforward. In the presence of frictions, the equality in (12) is in general replaced by inequalities. For example, with proportional transaction costs κ and no short sales and borrowing, (12) becomes

$$E_t\left[\rho\frac{u'_{t+1}(c_{t+1})}{u'_t(c_t)}(1+r_{t+1})\right] \leq \frac{1+\kappa}{1-\kappa}, \tag{15}$$

which loosens the link between prices and marginal utilities. However, using an equilibrium model calibrated to the trading needs to households' heterogeneous labour income risks, Heaton and Lucas (1996) found that transactions costs have a limited effect on the equilibrium risk premium because trading is very moderate in consumption-based models, but trading restrictions such as short-sale and borrowing constraints can potentially have larger effects (see also Constantinides and Duffie, 1996; Constantinides, Donaldson and Mehra, 2002; Brav, Constantinides and Geczy, 2002).

How market frictions affect the cross section of asset returns is a challenging issue. Merton (1987) considered an extension of the CAPM in which investors invest only in a subset of assets due to the information cost of learning about them. He showed that the segmentation of the market leads to modifications to the CAPM which exhibit a complex structure, depending on investor preferences, endowments and the nature of the segmentation. Here, more empirical guidance can be helpful. Using various measures of liquidity for individual stocks, Brennan and Subrahmanyam (1996) have documented an empirical link between liquidity and average returns. Liquidity of individual assets seems to exhibit commonalities (see Chordia, Roll and Subrahmanyam, 2000). This suggests the possibility that liquidity may contain factor risks. Assuming the CAPM to hold net of costs, Acharya and Pedersen (2005) allowed the effective costs in asset trading to be correlated with market returns to help explain the deviations in observed, pre-cost returns from the CAPM. Pastor and Stambaugh (2003) directly include the market average of a liquidity measure proposed by Campbell, Grossman and Wang (1993) as an additional risk factor in the SDF and find that it can enhance the explanatory power of multifactor models. Much is needed for the theoretical basis of the connections between frictions and the cross section of asset returns.

From a theoretical point of view, market frictions can contribute to the dynamic properties of asset prices. When flow of capital is costly, for example, the risk tolerance of marginal investors may increase and become dependent on market conditions, which can lead to predictable asset returns and more volatile prices. For example, Grossman and Vila (1992) showed that borrowing constraints can force risk-neutral investors to behave in a risk-averse manner. Grossman and Miller (1988) emphasize the imperfect mobility of capital by imposing costs on maintaining market presence and demonstrate that these costs lead to limited risk tolerance in the market and mean reversion in returns when trades are not perfectly synchronized. Using return and volume to infer order imbalances in the market, Campbell,

Grossman and Wang (1993) find that they indeed generate return reversals. Pagano (1989) and Allen and Gale (1994) also argue that costly participation in the market can exacerbate price volatility driven by demand shifts over long horizons. Huang and Wang (2006a; 2006b) further point out that low capital mobility in the form of costly participation in the market can lead to endogenous order imbalances. Moreover, the endogenous order imbalances tend to be asymmetric and large when they occur, leading to market crashes, fat-tails in asset returns and return reversals. There is now growing empirical evidence suggesting the low mobility of capital (for example, Coval and Stafford, 2006; Mitchell, Pedersen and Pulvino, 2007).

Constraints can also influence asset price dynamics. E. Miller (1977) and Harrison and Kreps (1979) have shown that short-sale constraints can inflate asset prices as they prevent short positions and thus can increase asset demand. Scheinkman and Xiong (2003) further demonstrated that short-sale constraints can lead to bubbles and high volatility in prices. Basak (1995) and Grossman and Zhou (1996) show that wealth constraints on market participants can lead to positive correlation between their risk tolerance and price movements. Such a correlation can contribute to higher and more persistent price volatility and mean reversion in returns.

Frictions have also been considered in explaining many other pricing anomalies. For example, Duffie (1996) and Vayanos and Weill (2006) examine how costs in trading from searching in the market can help to explain the price premium and the specialness (that is, high borrowing cost) of on-the-run Treasury bonds. Chen, Hong and Stein (2002) attempt to associate individual stock return momentum with short-sale constraints. Short-sale constraints can also help to explain empirical findings relating short interests, volume, and dispersion of analysts' forecasts to future returns. Kyle and Xiong (2001) suggest that capital constraints can be the cause of market contagion, which refers to negative co-movements across markets in the absence of negative news affecting both markets.

Although most of the literature has focused on how frictions in the market affect asset demand and consequently prices, some work has also examined the asset supply side, in particular how frictions in firms' real investments may affect the payoffs of corporate securities and their equilibrium prices. For example, Kogan (2001) shows that irreversibility in firms' real investments can lead to time-varying stock risks, which can help to explain their returns. Zhang (2006) examines potential links between the time-varying risks from the real side and several empirical anomalies.

The empirical and theoretical work so far suggest that market frictions can be an important factor in determining asset prices. However, they are mainly indicative. The models and the phenomena they address tend to be quite specialized. A more general framework capable of providing both a qualitative characterization and a quantitative assessment of the importance of frictions on the market behaviour is still lacking. This in part reflects the complexity of the problem. In the presence of frictions, the behaviour of market participants and the interactions among them become much more involved, and the simple aggregation properties assumed in the neoclassical framework no longer hold. Whether a general theory will eventually emerge or we have to settle for a collection of specialized models to deal with each individual phenomenon remains unclear at this point.

Information asymmetry

Information is a critical force driving financial markets. As is evident from (4), it is the expectation of market participants of discounted future cash flows that determines asset prices. In general, information is asymmetric among different market participants. Under the extreme situation when the market is competitive and sufficiently complete, the price system will be efficient in aggregating and revealing the information of all participants in the market, which gives a strong form of the EMH (see, for example, Grossman, 1976; Milgrom and Stokey, 1982). However, in the presence of frictions, in particular certain forms of market incompleteness, prices fail to be a sufficient static for the information in the market (see, for example, Grossman and Stiglitz, 1980; Hellwig, 1980; Diamond and Verrecchia, 1981). While information asymmetry also contributes to the existence of frictions, most of the analysis takes certain form of frictions as given and examines the effect of information asymmetry.

Information asymmetry substantially enriches the possible behaviour of asset prices. In general, current prices do not reveal all the information in the market. This immediately implies that past prices or other public information can provide additional information over current prices (see, for example, Brown and Jennings, 1989; Grundy and McNichols, 1989). More importantly, under asymmetric information, the behaviour of market participants will depend not only on their own information but also on their perception of the information others may have. In an intertemporal equilibrium setting, Wang (1993) demonstrates that information asymmetry can have a broad impact on asset prices, ranging from increasing the risk premium and price volatility to generating rich patterns in return dynamics. Allen, Morris and Shin (2006) further show that speculation on what others think may lead to price bubbles.

While information asymmetry increases the flexibility of the theory, its impact is harder to identify empirically as private information, by its nature, is mostly unobservable. By comparing price volatility on days when the stock market is open for trading with days when it is closed, French and Roll (1986) demonstrated convincingly the important role private information plays. Wang (1994) proposed using the joint behaviour of price and volume to examine the effect of information asymmetry (see also He and Wang, 1995). Empirical work along this line, notably, Llorente et al. (2002), have found supporting evidence for this approach. Recently, more detailed data on individual investors' trading records has become available (for example, Odean, 1998; Grinblatt and Keloharju, 2000), which will allow more direct tests on the importance of information asymmetry. For example, following a segment of the market, Evans and Lyons (2002) find that order flow in the currency market contains significant amount of information.

Another challenge to the neoclassical theory is market crashes, that is, large price drops without significant macro news. If the prices before and after a crash both reflect the market's expectation of discounted future cash flows, either the discount rate or the expectation (or both) must have changed during crash. As discussed earlier, liquidity effect can cause the discount rate to vary abruptly, as shown by Huang and Wang (2006b). Alternatively, the market expectation can change, reflecting changes in the information it contains. In the absence of big exogenous news, this information must come from the private information investors already possess. Various models have been proposed to explain market crashes, including Grossman (1988), Gennotte and Leland (1990), Bikhchandani, Hirshleifer and Welch (1992) and Romer (1993). These models typically allow both information asymmetry and market frictions, such as restrictions on what and how to trade, which prevent private information from being fully reflected in market prices. An unsettling issue for some of these models is the symmetry in large

price movements they produce, that is, equal likelihood of market crashes and surges. Asymmetry in favour of crashes can arise when frictions of asymmetric nature are present, such as borrowing constraints (for example, Yuan, 2005) and short-sale constraints (for example, Bai, Chang and Wang, 2006).

With regard to the impact of information asymmetry on return cross section, we face a similar situation as with frictions. The theory loses its tractability very quickly. Using a simple setting similar to that of the CAPM, Admati (1985) demonstrated that, under information asymmetry, the behaviour of equilibrium prices becomes very complex and sensitive to the information structure. We have not moved much beyond this point.

How information influences prices is a central issue in asset pricing. Existing work points to important channels for these influences, but the analysis is far from complete. On the one hand, the models so far are quite simplistic, especially in capturing the nature of information asymmetry in the market, and richer models are needed. On the other hand, even models with simple forms of information asymmetry are easily lost in their complexity. Both empirical and methodological breakthroughs are needed here.

Market microstructure

Many frictions are endogenous. A lot of effort has been devoted to studying how certain frictions, in particular liquidity, are determined in the market through the actual trading process, which is also referred to as the market microstructure (Garman, 1976). Despite its sophistication, the trading processes in the financial market are far more complex than what is assumed in most of the theoretical models, that is, through a Walrasian auction. The trading process also differs across different markets, ranging from over-the-counter markets and centralized exchanges with specialists to electronic limit order books, and constantly evolves over time. Several questions arise. How does a particular trading process influence the ease of trading or liquidity, investors’ trading behaviour, and the properties of prices? How does it influence the efficiency of the market and overall asset valuation? What determines the form of the trading process in a given market and how it evolves?

A large body of work focuses on how market-makers, who provide liquidity by absorbing transitory order imbalances, influence effective trading costs and high-frequency price dynamics. Market-makers’ behaviour depends on the costs they face, which have two components: the cost of holding an inventory and the cost of adverse selection when trading against better informed investors. Earlier analysis emphasized the former. Attributing the inventory cost to the risk in the value of inventory, Stoll (1979) showed that the effective trading cost, as measured by the bid-ask spread, increases with competitive market makers’ risk aversion and the volatility of asset value (see also Amihud and Mendelson, 1980). Roll (1984b) developed an empirical measure of the effective bid-ask spread and found it to be nontrivial for most individual stocks. Later attention has turned to the effect of adverse selection. Glosten and Milgrom (1985) showed how the existence of informed trades contributes to the bid-ask spread. Kyle (1985) demonstrated how an insider's strategic behaviour hinders the informational efficiency of the market and reduces its liquidity. Similar analysis have been carried out for markets organized as a limit order book, such as in Copeland and Galai (1983), Rock (1990) and Glosten (1994). High frequency data on quotes and trades made it possible for extensive studies of the behaviour of trading and prices in different markets, following the intuition developed in theory, including Glosten

and Harris (1988), Hasbrouck (1991), Madhavan and Smidt (1993), Biais, Hillion and Spatt (1995), and Lyons (1995). Imperfect competition among market makers also leads to additional complexity in the supply of liquidity (see, for example, Christie and Schultz, 1994; Barclay et al., 1999; Wahal, 1997; Ellis, Michaely and O'Hara, 2002). Additional theoretical work has been directed at how market-makers behave strategically in their liquidity provision under different trading processes, notably Glosten (1989), Foucault (1999), Vayanos (1999), and Goettler, Parlour and Rajan (2005).

Although most of the theoretical analysis on microstructure has focused on centralized markets, some considers the over-the-counter (OTC) markets, which by some measures are more common. Duffie, Garleanu and Pedersen (2005) develop a search-based model for the OTC market. It was then applied to several markets such as securities borrowing (Duffie, Garleanu and Pedersen, 2002) and Treasury bonds (Vayanos and Weill, 2006).

Market microstructure effects provide new insights on market behaviour at high frequency. For example, Admati and Pfleiderer (1988) and Foster and Viswanathan (1990) considered how traders' strategic behaviour in response to the liquidity in the market can help to explain intraday variations in trading volume and price volatility (see also Hong and Wang, 2000, for alternative explanations). To the extent that market microstructure affects transactions costs in the market, it also influences asset prices in general, as we discussed earlier (see also O'Hara, 2003).

Many studies have also compared the different ways trading is organized, in particular how different market organizations may affect their liquidity provision and informational efficiency. For example, Copeland and Galai (1983) illustrated certain benefits of call auctions. Grossman (1992) examined the efficiency of upstairs market for block trades. Glosten (1994) discussed the advantage of an electronic limit order book. Seppi (1997) considered the impact of competition between a limit order book and a specialist when they coexist. Direct empirical comparisons of different trading mechanisms are difficult as they are usually adopted for different markets. But the general evidence is clear: market behaviour does vary with the actual trading process (for example, Amihud and Mendelson, 1991; Ready, 1999; Goldstein and Kavajecz, 2000; Bessembinder, 2003; Boehmer, Saar and Yu, 2005).

Although a lot has been learned about market microstructure, more remains to be learned. Many factors are at play and only a few are considered at a time, both theoretically and empirically. Their relative importance is hard to gauge empirically to allow for possible simplification. It remains a question why a given market is organized in a certain fashion. A better understanding of the precise nature and the magnitude of its impact on asset valuation and market efficiency is also needed.

From a broader perspective, there is also the question on the overall market structure (such as what securities are traded and why), which we may refer as market macrostructure. Most of the neoclassical theory takes it as given. But the dramatic evolution of the market, driven by a flood of innovations in finance and advances in technology and changes in the global economy, has forced researchers to think hard about this question. Some preliminary work has emerged in addressing this question. Allen and Gale (1988) consider the choice of firms in issuing securities when taking into account its impact on market structure and the resulting prices. Duffie and Rahi (1995) examine how exchanges decide on the derivative contracts to offer. Huang and Wang (1997) analyse how the introduction of new securities may influence the overall informational efficiency of the market. Of course, a significant number of

financial transactions are carried out through financial intermediaries rather than in the form of financial securities. Allen and Gale (2004) further explore the interplay between the two. Despite the importance of this question, the work so far is extremely primitive and in many ways merely serves to keep the question in play.

Behavioural finance

In the search for alternative explanations of asset pricing anomalies, attention has also turned to some of the basic assumptions of the neoclassical theory. The absence of arbitrage and the notion of efficient markets rely on the assumption that marginal investors in the market are not hindered by market frictions. The work on frictions has attempted to relax this assumption. Equilibrium models of asset pricing further adopt the assumption that average investors behave ‘rationally’. However, the notion of rationality is an ambiguous one. Earlier models describe rationality in the form of expected utility, where the expectation is taken under the actual probability measure, for example, in the form of von Neumann and Morgenstern (1944). This implies that an investor's belief about market behaviour is consistent with its true behaviour. In addition, the utility function is assumed to depend only on the level of consumption. In a simple form, an investor's behaviour is described by the following expected utility

$$u_t(c_t) + E_t[u_{t+1}(c_{t+1})] = u_t(c_t) + \sum_{\omega} p(\omega) u_{t+1}(c_{t+1}), \tag{16}$$

where $p(\omega)$ is the actual probability for a future state ω and c_{t+1} is the level of future consumption. For simplicity, here we assume time-separable utility function and symmetric information. (In the case of asymmetric information, $p(\omega)$ becomes the probability conditional on the investor's information.) Since its justification is more normative than positive, this form of rationality has attracted many criticisms from very earlier on, notably, Allais (1953), Ellsberg (1961), Kahneman and Tversky (1974). Deviations from this simple form of rationality have gained prominence in various attempts to explain market anomalies. Since these explanations are mostly based on various assumptions on investor behaviour, this area of research has gained the name of behaviour finance.

Most of the evidence against the simple form of rationality is from laboratory experiments on human subjects with hypothetical prospects or small-stake choices. It was documented that subjects often fail to form objective and consistent probabilistic assessments, exhibiting patterns like overconfidence (for example, Fischhoff, Slovic and Lichtenstein, 1977; Weinstein, 1980), belief perseverance, and anchoring (Kahneman and Tversky, 1974). When facing gambles with stated probabilities, subjects’ choices are incompatible with the expected utility, as documented by Kahneman and Tversky (1974). In addition, when confronted with outcomes with unknown probabilities, subjects’ choices cannot be reconciled with a consistent probabilistic assessment of the possible outcomes (for example, Ellsberg, 1961). Knight (1936) referred to this situation as uncertainty as opposed to risk, for which probabilities are known.

The richness in these behavioural variations, when used to describe investor behaviour, gives tremendous flexibility in providing possible explanations of asset price behaviour. For example, DeBondt and Thaler (1985) attribute the reversals in long-horizon market returns to investor overreaction. Using a version of prospect theory, in which investors exhibit loss aversion (that is, over weighting potential losses from a benchmark point over gains), Barberis, Huang and Santos (2001) demonstrate that it can help to reconcile the high equity premium and price volatility with smooth consumption. Daniel, Hirshleifer and Subrahmanyam (2001) interpret the cross-sectional deviations in average equity returns from the CAPM, in particular the value and size premia, as a result of overconfidence in investors' interpretation of their private information. Models based on belief perseverance and representativeness (for example, Barberis, Shleifer and Vishny, 1998), overconfidence (Daniel, Hirshleifer and Subrahmanyam, 1998), and under-reaction to information (Hong and Stein, 1999) have been used to explain anomalies like short-horizon return momentum, long-horizon return reversal, and post-earnings announcement drift. Liu, Pan and Wang (2005) assume uncertainty aversion to reconcile the high premium of options paying off in rare events with their low probabilities seen in the data.

The experimental basis of behavioural assumptions raises the question of their relevance for actual individual behaviour in real economic decisions. As data on individual investments becomes available, more direct examination of their behaviour is possible. Investors are found to invest more in stocks they are familiar with (for example, French and Poterba, 1991; Grinblatt and Keloharju, 2001), to diversify naively (for example, Benartzi and Thaler, 2001), to trade excessively (for example, Barber and Odean, 2000), and to sell winners quickly while holding on to losers (for example, Odean, 1998). These investment patterns are interpreted as being consistent with some of the behavioural assumptions. However, the presence of many other factors, ranging from taxes, information to portfolio considerations, leaves plenty of space for alternative interpretations.

Although the behavioural patterns explored in the literature are not fully described by the expected utility theory and are thus referred to as irrational, most of them can be captured by a more general form of rationality formulated by Savage (1954), which allows for subjective beliefs and state-dependent utility functions:

$$u_t(c_t) + \sum_{\omega} p^i(\omega) u_{t+1}(c_{t+1}, \omega), \tag{17}$$

where $p^i(\omega)$ denotes the subjective probability of an investor i. The subjective expected utility theory in (17) still exhibits a general form of consistency on behaviour but can accommodate rich variations in individual beliefs and preferences. In addition, within this more general notion of rationality, the distinction between beliefs and preferences become more of a formality. For example, a subjective expected utility function after the following transformation

$$u_t(c_t) + \sum_{\omega} p^i(\omega)\, u_{t+1}(c_{t+1}, \omega) = u_t(c_t) + \sum_{\omega} p(\omega)\, \tilde{u}_{t+1}(c_{t+1}, \omega),$$

where $\tilde{u}_{t+1}(c_{t+1}, \omega) = [\, p^i(\omega) \,/\, p(\omega)\,]\, u_{t+1}(c_{t+1}, \omega)$, becomes an expected utility function (under the true probability measure *p*) describing the same behaviour. From this point of view, we have three observations. First, many behavioural patterns can be obtained from state-dependent expected utility, a form of rationality slightly more general than that defined by state-independent expected utility. Second, with state dependent utility, many behaviour models are formally indistinguishable from those considered within the neoclassic framework. Third, without additional restrictions, the distinction between belief and preference is largely arbitrary.

As for the consumption-based CAPM with habits, behavioural models, as they stand now, also face major limitations. First, without additional discipline, the theory is simply too flexible. As the distance between assumption and result decreases, the multiplicity of potential explanations actually increases. Second, even taking the behavioural patterns at the individual level as given, it is less clear how they aggregate. Idiosyncratic biases at the individual level may well average out at the market level. Another critical and perhaps more important issue is to what extent deviations in individual behaviour from rationality, even if they persist at the market level, influence asset prices. Take momentum as an example. If the predictability arises from the under-reaction of some investors to new information, investors who have information and capital, also referred to as arbitrageurs, should jump in to take advantage of the predictability until it disappears. As discussed in the section on market efficiency, two factors can hold back this market force, namely, risk and frictions. De Long et al. (1990) argued that irrationality can generate sufficient risk in the market to deter the arbitrageurs. But Sandroni (2000) demonstrated that, in a perfect market, investors acting on irrational beliefs do not survive in the long run (although their price impact may persist longer, as shown in Kogan et al., 2006). For risk to matter and the impact of irrational behaviour to persist, frictions or ‘limits of arbitrage’ are essential, a point emphasized by Shleifer and Vishny (1997). However, as discussed earlier in this section, in the presence of frictions various so-called anomalies can be accounted for without relying on additional behavioural assumptions. Faced with many competing and piece-wise ‘theories’, the challenge we face is to further pin down the actual causes of observed pricing patterns within a unified, and hopefully simple theory.

Corporate finance

Guided by prices, the actual allocation of capital is achieved by transactions among market participants, mainly firms and households. Firms’ financial behaviour is of particular importance as their main function is to create value from the existing capital, while households are the ultimate owner and beneficiaries. The pricing principles from the neoclassical finance have lent powerful tools for corporate and individual financial decision making. A better understanding of the financial behaviour of firms and individuals, which drives the demand and supply of assets, is also essential to our understanding of asset prices.

Corporate finance in the neoclassical theory began with the seminal work of Modigliani and Miller (1958; 1963). Using the principle of no arbitrage, they showed that in the absence of imperfections a firm's value depends only on its investment decisions. Financing and payout decisions merely determine how payoffs from investments are split between different claims associated with each financing vehicle – for example, equity and debt. The irrelevancy results of Modigliani and Miller (MM) clearly points to the areas where corporate finance matters. Much of the work since has focused on these areas where assumptions of MM are relaxed, in particular when frictions and information problems are important. The information problems have been mostly framed in the interaction between a firm's insider who manages it and its outside investor who finances it. The insider can be an entrepreneur seeking outside capital or a manager running a mature public company. Two types of information problems were identified early on, namely, adverse selection and moral hazard. Leland and Pyle (1977) and Ross (1977) examined how the adverse-selection problem influences the firm's financial decisions when firm insiders/managers know more about firm assets. In this case, firms' actions also serve as signals to outside investors, which will influence their perception of firm value. Viewing outside investors (for example, equity and debt holders) as the principal and managers as agents, Wilson (1968) and Ross (1973) considered the moral hazard problem when managers have more information on firm assets and their own actions. Based on this type of agency theory, Jensen and Meckling (1976) and Myers (1977) examined how firm value can be influenced by the conflicts between different stakeholders, that is, investors versus managers and shareholders versus bondholders.

Recent developments in corporate finance have followed this theme. Corporate behaviour was often viewed as a manifestation of these conflicts. In order to turn this general perspective into testable theories, more structure is needed with regard to the nature and the magnitude of these imperfections. In this regard, more guidance from the data becomes critical. Our discussion starts with how a firm chooses its financing or capital structure, the focus of neoclassical theory. We then turn to how inefficiencies in financing caused by frictions and information problems influence a firm's investments. Finally, we consider the issues concerning corporate control, which looks at the problems in corporate finance from a more fundamental perspective.

Financing

Financially, a firm is about how to raise capital and how to use it. The two questions are obviously intertwined. Under MM, the two become independent. A firm's overall cost of capital, that is, the valuation of its assets, is not affected by how it is financed. An important friction omitted in this irrelevancy result is taxes. With different tax treatments on debt and equity financing, different choices of capital structure will affect the firm's tax liability and naturally its value. For example, when interest payments on corporate debt are excluded from corporate taxes, the firm can pass on higher returns to its investors by substituting debt for equity. This tax arbitrage, however, has its barbs. First, it does not account for investors' personal taxes. Miller (1977) showed that, as investors of different tax clienteles settle for different mixes of debt and equity, an equilibrium is reached when marginal investors are indifferent between the two, which determines the total amount of debt and equity but not for individual firms as their securities are substitutes. Second, it does not take into account the potential cost of using debt, which can lead to bankruptcy. When the effects of personal and corporate taxes are different (for

example, when securities of different firms are not perfect substitutes) and bankruptcy is costly, we have the so-called ‘trade-off’ theory of capital structure. Each firm is trading off the tax benefits of debt and the bankruptcy costs.

How significant these benefits and costs are remains an empirical question. Data seems to suggest that they are important. For example, Graham (2000) estimates that the effective tax rate paid by marginal investors on debt is significantly higher than that on equity, suggesting a large benefit of debt finance. The cost of bankruptcy has several sources, the direct cost of bankruptcy process (for example, Weiss and Wruck, 1998) and the indirect cost of financial distress such as conflicts between different stakeholders (for example, Asquith and Wizman, 1990), loss of business and financial counter parties (for example, Maksimovic and Titman, 1991), pressure from competitors (for example, Chevalier, 1995). A more recent study by Andrade and Kaplan (1998) estimates costs of financial distress to be in the range of 10–20 per cent of the firm value prior to distress. *Ex post* the cost of this size may seem modest since on an *ex ante* basis one has to factor in its probability. However, the fact that these costs tend to have large negative beta (negatively correlated with the market) may imply that their present value is non-trivial (for example, Almeida and Philippon, 2006).

The trade-off theory has several implications. First, each firm should have an optimal capital structure, which depends on its tax status and cost of financial distress. Although the theory does not fully specify what determines these two factors, different proxies were used empirically. For example, firms with higher business risk and more intangible assets were associated with higher distress costs and thus a low debt–asset ratio. Many empirical studies have found positive evidence on the link between these proxies and the capital structure, such as Auerbach (1985), Titman and Wessels (1988) and Rajan and Zingales (1995). But the evidence is not uniformly supportive. Wald (1999) finds that profitability has a strong negative relation with debt–asset ratios, while the trade-off theory would imply that more profitable firms should use more debt to shield their income. Second, if adjustment is costly, a firm's capital structure will be away from its optimum most of the time but always evolves towards it. As a result, the firm is more likely to issue debt when below the target and equity when above. Earlier tests found this prediction to be consistent with the data (for example, Taggart, 1977; Auerbach, 1985), but more recent tests have found mixed evidence (for example, Hovakimian, Opler and Titman, 2001). The trade-off theory is intuitive and enjoys partial empirical success, but still leaves some gaps. In particular, the significant costs of financial distress need both theoretical and empirical justification.

Based on patterns in firms’ financing choices, Myers and Majluf (1984) propose the pecking-order theory of capital structure. It starts with the premise that outsider investors have less information about a firm's use of capital. Thus, they face an adverse selection problem and will on average undervalue new shares. Equity becomes more costly as a financing vehicle than debt. This simple theory yields several predictions: (*a*) firms prefer internal to external finance and debt to equity finance; (*b*) the market reacts negatively to new share issues; (*c*) dividends are persistent; (*d*) a firm's debt–asset ratio changes with its cumulative needs for external financing. Heuristic empirical observations are surprisingly compatible with the pecking-order theory. But more extensive tests reveal some inconsistencies. For example, Jung, Kim and Stulz (1996) and Fama and French (2002) have found that small-growth firms rely heavily on equity financing. Although the pecking-order theory is very much in the spirit of Ross (1977), it relies on simplifying assumptions. In particular, no optimal contracting is considered by allowing for more

complex forms of financing and incentives to resolve the information problem.

The agency theory of capital structure focuses on the conflict of interest between managers and shareholders. This is different from the trade-off and pecking-order theories, in which managers act on behalf of current shareholders. When their information and actions are not fully observable to outsiders, which may include current shareholders, managers can benefit themselves at the cost of shareholders. When incentives through contracting fail to fully mitigate this conflict, capital structure will be influenced by investors' efforts to contain managers. Following Jensen (1986), different variations of the agency theory have been proposed, notably Harris and Raviv (1993), Stulz (1990), and Zwiebel (1996). For example, Jensen (1986) argued that debt helps to get cash out of managers' hands and is thus preferred to outside equity before bankruptcy becomes important. Although empirically leverage does curb investments (for example, Lang, Ofek and Stulz, 1996), new debt issues do not seem to increase firm value (Eckbo, 1986).

Factors like taxes, cost of financial distress, information and agency problems do matter for firms' financing decisions, as the empirical evidence suggests. But each of the theories captures only part of the picture. They are also mostly partial equilibrium by nature, taking certain aspects of the problem as given such as the contracting environment and firms' investment opportunities. A more integrated and empirically refutable theory would be desirable. On the empirical side, it remains a challenge to reconcile the financing patterns found under certain circumstances with the lack of a link between taxes, financing and market value documented in Fama and French (1998) over a large sample of firms.

Investments

Clearly, the forces driving a firm's financing choices also influence its investments, that is, the use of capital. We have identified at least two channels. The first channel is simply through the cost of capital. In the case of trade-off theory, for example, a firm's cost of capital varies with its capital structure and so will its investments. The second channel is through the behaviour of managers, who make investment decision in response to the incentives they face, which are also related to the firm's financing choices. The direct effect of cost of capital has found supportive evidence. For example, using a structural model to calibrate firms' investment opportunities, Hennessy (2004) finds that a high debt level curbs investments. In the state of distress, firms also cut down their investments, as documented in many studies, including Chevalier (1995b), Phillips (1995) and Zingales (1998).

The agency effect has attracted more interest. A variety of private benefits of managers were suggested, ranging from empire building (Williamson, 1964; Jensen, 1986) and career considerations (Narayanan, 1985; Holmstrom, 1999) to inertia (for example, Bertrand and Mullainathan, 2003). Misalignment between managers' and shareholders' interests will lead to suboptimal investment decisions. A simple prediction of this argument is that firms with more free cash in hand will make more and less desirable investments. Broad empirical evidence was found to be consistent with this prediction, such as Fazzari, Hubbard and Petersen (1988), Hoshi, Kashyap and Scharfstein (1991), and Gilchrist and Himmelberg (1995). One challenge in establishing the empirical link is the problem of endogeneity. For example, a firm's free cash is endogenous and may vary with its investments opportunities. Several studies have used 'natural experiments' to avoid the endogeneity issue. For example, Blanchard, Lopez-de-Silanes

and Shleifer (1994) find that firms' acquisition activities increase after receiving cash windfalls from legal settlements unrelated to their business.

A positive correlation between cash and investments does not prove the agency theory. It can also be consistent with the pecking-order theory. More cash relaxes the capital constraint imposed by high cost of external financing. The question is whether free cash flow leads to negative net present value (NPV) investments. Evidence such as the negative price reaction to new equity issues (for example, Asquith and Mullins, 1986) may suggest so. However, more direct evidence indicates otherwise. For example, McConnell and Muscarella (1985) documented positive market reactions to firms' capital expenditure announcements.

Firms' investment decisions are of central importance to finance. Frictions and information problems imply inefficient use of capital. Extensive evidence is suggestive of such inefficiencies, but it is far from definitive. A comprehensive empirical evaluation of the extent and the magnitude of these efficiencies and the potential forces driving them is not available yet.

Corporate control and governance

Most of the new theories in corporate finance take as given the means different parties use to resolve conflicts. For example, in the pecking-order theory or the agency theory of financing, the mixture of debt and equity is the tool available to balance the interests of managers and outside investors. But a whole set of devices can be utilized to resolve their conflicts, including incentive contracts for managers and a rich set of corporate securities beyond equity and bonds. It makes sense to think at a deeper level about the economic structure of a firm and its resulting behaviour.

Built on the ideas of Coase (1937), Grossman and Hart (1986) proposed the idea that a firm is defined by the allocation of control rights over its assets, the rights to utilize these assets. In such a setting, conflicts among different stakeholders are resolved by optimal allocation of control rights rather than extensive contracting, which is assumed to be infeasible with hard-to-specify future contingencies. Such an allocation will then determine how the firm behaves, including its investment decisions and financing arrangements. It will also determine how it is governed – for example, who takes control and when. A collection of theories on corporate behaviour was developed under this setting.

Aghion and Bolton (1992) considered the financing problem of an entrepreneur who also enjoys private benefits from running his firm (see also Hart and Moore, 1998). The optimal structure of the firm would be for him to retain the control rights of the firm (so he can enjoy the private benefits) while selling cash flow claims to outside investors. This looks very much like a mixture of equity and debt financing, except that now it is the outcome of optimal corporate control. If embedded in an intertemporal environment with uncertainty, the model also lead to implications on the dynamics of the firm's financing and investments. By looking at venture capital investments in start-up companies, Gompers (1995) and Kaplan and Stromberg (2001) have found patterns compatible with the model's predictions. This model, however, captures mostly inside equity and is less descriptive of large public firms, which involves mostly equity held by outsiders. Fluck (1998) and Myers (2000) have considered models for outside equity financing. Within a similar framework, Grossman and Hart (1980, 1988) analyse the market for corporate control in the form of takeovers (see also Harris and Raviv, 1988) when

shareholdings are diverse. Aghion and Tirole (1997) examined issues concerning corporate governance, such as the role of corporate boards, which act as shareholders' representative in exercising their control rights.

Models of incomplete contracting capture some salient features of firms and attempt to examine corporate finance issues from a more basic and integrated perspective. But they are highly simplified. Their predications are mostly qualitative and dependent on deeper parameters, such as what can or cannot be contracted. The fact that these parameters are hard to observe make it hard to empirically test the models.

Another approach is to consider the firm as a full contract among its stakeholders, including managers and outside investors, very much in the spirit of Leland and Pyle (1977) and Ross (1977) (see also Townsend, 1978; Gale and Hellwig, 1985). For example, Gertler (1992), Clementi and Hopenhayn (2006), and DeMarzo and Fishman (2007) examine optimal contracts between investors and a manager to induce optimal investments. Atkeson and Cole (2005) consider the optimal financing contract in the presence of agency problems and manager risk aversion. In contrast to the assumptions in the models based on incomplete contracts, this approach explores what optimal contracting can achieve. As shown in Dybvig and Zender (1991), under certain circumstances optimal contracting can largely resolve the information problems between managers and shareholders.

The full contracting approach avoids some of the arbitrariness in the theory of incomplete contracts. But it has its own challenges. It is quite limited in describing large public companies, which involves a large number of stakeholders, including a hierarchy of managers and a diverse set of investors. Its predictions depend on the assumptions about other frictions such as verification and enforcement costs. Realistic assumptions about these frictions are also hard to pin down. This also leads to the question of the robustness of contractual arrangements from the models.

Conclusion

Developments in finance since the mid-1980s have expanded the success of neoclassical theory, especially in the area of arbitrage pricing, as well as its boundaries. Extensive and more rigorous empirical analysis has exposed the limitations of the simple asset pricing models and the simplistic notion of market efficiency. The fact that we still don't have a satisfactory notion of risk and can't explain movements in asset prices after the fact clearly suggests the need to enrich our theory. Imperfections such as frictions and information asymmetry are part of the market reality and should be incorporated. They can very much enhance our understanding of market participants' behaviour and its impact on the market itself. A rich set of models, accompanied by empirical work, has been explored to explain the observed patterns in the financial market and in corporate finance. Liquidity and agency problems have been identified as manifestations of imperfections in the market and corporate contexts and useful lenses through which to examine their behaviour.

An unavoidable challenge in modelling imperfections is that they come in all shapes and sizes, and their impact is in general complex. Empirical evaluation of the significance of various imperfections is very much needed to arrive at a unified framework, synthesizing the important intuition from the collection of specialized models we have. After carefully collecting and studying the pieces and parts, we may be able to hope for a more general theory of finance.

See Also

- arbitrage pricing theory
- capital asset pricing model
- corporate governance
- dividend policy
- finance
- risk

Bibliography

Abel, A.B. 1990. Asset prices under habit formation and catching up with the Joneses. *American Economic Review* 80, 38–42.

Acharya, V. and Pedersen, L. 2005. Asset pricing with liquidity risk. *Journal of Financial Economics* 77, 385–410.

Admati, A.R. 1985. A noisy rational expectations equilibrium for multiple asset securities markets. *Econometrica* 53, 629–57.

Admati, A.R. and Pfleiderer, P. 1988. A theory of intraday patterns: volume and price variability. *Review of Financial Studies* 1, 3–40.

Adrian, T. and Rosenberg, J.V. 2006. Stock returns and volatility: pricing the short-run and long-run components of market risk. Staff Report No. 254, Federal Reserve Bank of New York.

Aghion, P. and Bolton, P. 1992. An incomplete contracts approach to financial contracting. *Review of Economic Studies* 59, 473–94.

Aghion, P. and Tirole, J. 1997. Formal and real authority in organizations. *Journal of Political Economy* 105, 1–29.

Ahn, D.H., Dittmar, R.F. and Gallant, A.R. 2002. Quadratic term structure models: theory and evidence. *Review of Financial Studies* 16, 243–88.

Ait-Sahalia, Y. 1996. Nonparametric pricing of interest rate derivative securities. *Econometrica* 64, 527–60.

Allais, M. 1953. Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’École américaine. *Econometrica* 21, 503–46.

Allen, F. and Gale, D. 1988. Optimal security design. *Review of Financial Studies* 1, 229–63.

Allen, F. and Gale, D. 1994. Limited market participation and volatility of asset prices. *American Economic Review* 84, 933–55.

Allen, F. and Gale, D. 2004. Financial intermediaries and markets. *Econometrica* 72, 1023–61.

Allen, F., Morris, S. and Shin, H.S. 2006. Beauty contests and iterated expectations in asset markets. *Review of Financial Studies* 19, 719–52.

Almeida, H. and Philippon, T. 2006. The risk-adjusted cost of financial distress. Working paper, New York University.

Amihud, Y. and Mendelson, H. 1980. Dealership market: market making with inventory. *Journal of Financial Economics* 8, 31–53.

Amihud, Y. and Mendelson, H. 1986. Asset pricing and the bid–ask spread. *Journal of Financial Economics* 17, 223–49.

Amihud, Y. and Mendelson, H. 1991. Volatility, efficiency, and trading: evidence from the Japanese stock market. *Journal of Finance* 46, 1765–89.

Amin, K.I. 1993. Jump diffusion option valuation in discrete time. *Journal of Finance* 48, 1833–63.

Anderson, R.W. and Sundaresan, S. 1996. Design and valuation of debt contracts. *Review of Financial Studies* 9, 37–68.

Andrade, G. and Kaplan, S.N. 1998. How costly is financial (not economic) distress? Evidence from highly leveraged transactions that become distressed. *Journal of Finance* 53, 1443–93.

Ang, A., Hodrick, R.J., Xing, Y. and Zhang, X. 2006. The cross-section of volatility and expected returns. *Journal of Finance* 51, 259–99.

Asquith, P. and Mullins, D.W. 1986. Equity issues and offering dilution. *Journal of Financial Economics* 15, 61–89.

Asquith, P. and Wizman, T.A. 1990. Event risk, covenants, and bondholder returns in leveraged buyouts. *Journal of Financial Economics* 27, 195–213.

Atkeson, A. and Harold, H. 2005. A dynamic theory of optimal capital structure and executive

compensation. Working Paper No. 11083. Cambridge, MA: NBER.

Auerbach, A.S. 1985. Real determinants of corporate leverage. In *Corporate Capital Structure in the United States*, ed. B.M. Friedman. Chicago: University of Chicago Press.

Bai, Y., Chang, E.C. and Wang, J. 2006. Asset prices and short-sale constraints. Working Paper, Massachusetts Institute of Technology.

Bakshi, G., Cao, C. and Chen, Z. 1997. Empirical performance of alternative option pricing models. *Journal of Finance* 52, 2003–49.

Ball, R. and Brown, P. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6, 159–78.

Bansal, R., Dittmar, R.F. and Lundblad, C.T. 2005. Consumption, dividend, and the cross section of equity returns. *Journal of Finance* 60, 1639–72.

Bansal, R. and Yaron, A. 2004. Risks for the long run: a potential resolution of asset pricing puzzles. *Journal of Finance* 59, 1481–510.

Banz, R.W. 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics* 9, 3–18.

Barber, B.M. and Odean, T. 2000. Trading is hazardous to your wealth: the common stock investment performance of individual investors. *Journal of Finance* 55, 773–806.

Barberis, N., Huang, M. and Santos, T. 2001. Prospect theory and asset prices. *Quarterly Journal of Economics* 116, 1–53.

Barberis, N., Shleifer, A. and Vishny, R. 1998. A model of investor sentiment. *Journal of Financial Economics* 49, 307–43.

Barclay, M.J. Christie, W.G., Harris, J.H., Kandel, E., et al. 1999. The effects of market reform on the trading costs and depths of Nasdaq stocks. *Journal of Finance* 54, 1–34.

Basak, S. 1995. A general equilibrium model of portfolio insurance. *Review of Financial Studies* 8, 1059–90.

Basu, S. 1983. The relationship between earnings’ yield, market value and return for NYSE common stocks: further evidence. *Journal of Financial Economics* 12, 129–56.

Bates, D.S. 1996. Jumps and stochastic volatility: exchange rate processes implicit in Deutsche mark options. *Review of Financial Studies* 9, 69–107.

Bates, D.S. 2000. Post-’87 Crash fears in the S&P 500 futures option market. *Journal of Econometrics* 94, 181–238.

Bessembinder, H. 2003. Trade execution costs and market quality after decimalization. *Journal of Financial and Quantitative Analysis* 38, 747–77.

Benartzi, S. and Thaler, R.H. 2001. Naive diversification strategies in defined contribution savings plans. *American Economic Review* 91, 79–98.

Berk, J.B., Green, R.C. and Naik, V. 1999. Optimal investment, growth options and security returns. *Journal of Finance* 54, 1553–607.

Bernard, V.L. and Thomas, J.K. 1989. Post-earningsannouncement drift: delayed price response or risk premium? *Journal of Accounting Research* 27, 1–36.

Bertrand, M. and Mullainathan, S. 2003. Enjoying the quiet life? Corporate governance and managerial preferences. *Journal of Political Economy* 111, 1043–75.

Biais, B., Hillion, P. and Spatt, C. 1995. An empirical analysis of the limit order book and order flow in the Paris bourse. *Journal of Finance* 50, 1655–89.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* 100, 992–1026.

Black, F. and Cox, J.C. 1976. Valuing corporate securities: some effects of bond indenture provisions. *Journal of Finance* 31, 351–67.

Black, F., Jensen, M.C. and Scholes, M. 1972. The capital asset pricing model: some empirical tests. In *Studies in the Theory of Capital Markets*, ed. M.C. Jensen. New York: Praeger.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Blanchard, O.J., Lopez-de-Silanes, F. and Shleifer, A. 1994. What do firms do with cash windfalls? *Journal of Financial Economics* 36, 337–60.

Boehmer, E., Saar, G. and Yu, L. 2005. Lifting the veil: an analysis of pre-trade transparency at the NYSE. *Journal of Finance* 60, 783–815.

Brav, A., Constantinides, G.M. and Geczy, C.C. 2002. Asset pricing with heterogeneous consumers and limited participation: empirical evidence. *Journal of Political Economy* 110, 793–842.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7, 265–96.

Breeden, D.T. 1980. Consumption risk in futures markets. *Journal of Finance* 35, 503–20.

Brennan, M.J. and Schwartz, E.S. 1979. A continuous time approach to the pricing of bonds. *Journal of Banking and Finance* 3, 133–55.

Brennan, M.J. and Subrahmanyam, A. 1996. Market microstructure and asset pricing: on the compensation for illiquidity in stock returns. *Journal of Financial Economics* 41, 341–64.

Brown, D.P. and Jennings, R.H. 1989. On technical analysis. *Review of Financial Studies* 2, 527–51.

Brown, R.H. and Schaefer, S.M. 1994. The term structure of real interest rates and the Cox Ingersoll and Ross model. *Journal of Financial Economics* 35, 3–42. Online available at http://econpapers.repec.org/article/eeejfinec/

Brown, S.J. and Dybvig, P.H. 1986. The empirical implications of the Cox, Ingersoll Ross theory of the term structure of interest rates. *Journal of Finance* 41, 617–30.

Campbell, J.Y. and Cochrane, J. 1999. By force of habit: a consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107, 205–51.

Campbell, J.Y., Grossman, S.J. and Wang, J. 1993. Trading volume and serial correlation in stock returns. *Quarterly Journal of Economics* 108, 905–39.

Campbell, J.Y. and Shiller, R.J. 1988. The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195–228.

Chan, K.C., Karolyi, G.A., Longstaff, F.A. and Sanders, A.B. 1992. An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* 47, 1209–27.

Chan, L.K. and Lakonishok, J. 1995. The behavior of stock prices around institutional trades. *Journal of Finance* 50, 1147–74.

Chan, Y.L. and Kogan, L. 2002. Catching up with the Joneses: heterogeneous preferences and the dynamics of asset prices. *Journal of Political Economy* 110, 1255–85.

Chen, J., Hong, H. and Stein, J. 2002. Breadth of ownership and stock returns. *Journal of Financial Economics* 66, 171–205.

Chen, N., Roll, R. and Ross, S.A. 1986. Economic forces and the stock market. *Journal of Business* 59, 383–404.

Chen, R. and Scott, L. 1992. Pricing interest rate options in a two-factor Cox–Ingersoll–Ross model of the term structure. *Review of Financial Studies* 5, 613–36.

Chevalier, J.A. 1995. Capital structure and product market competition: empirical evidence from the supermarket industry. *American Economic Review* 85, 415–35.

Chordia, T., Roll, R. and Subrahmanyam, A. 2000. Commonality in liquidity. *Journal of Financial Economics* 56, 3–28.

Christie, W.G. and Schultz, P.H. 1994. Why do Nasdaq market makers avoid odd-eighth quotes? *Journal of Finance* 49, 1813–40.

Clementi, G.L. and Hopenhayn, H.A. 2006. A theory of financing constraints and firm dynamics. *Quarterly Journal of Economics* 121, 229–65.

Coase, R.H. 1937. The nature of the firm. *Economica* 4, 386–405.

Connor, G. and Korajczyk, R.A. 1988. Risk and return in an equilibrium APT: application of a new test methodology. *Journal of Financial Economics* 21, 255–90.

Constantinides, G.M. 1986. Capital market equilibrium with transaction costs. *Journal of Political Economy* 94, 842–62.

Constantinides, G.M. 1990. Habit formation: a resolution of the equity premium puzzle. *Journal of Political Economy* 98, 519–43.

Constantinides, G.M. 1992. A theory of the nominal term structure of interest rates. *Review of Financial Studies* 5, 531–52.

Constantinides, G.M., Donaldson, J.B. and Mehra, R. 2002. Junior can't borrow: a new perspective on the equity premium puzzle. *Quarterly Journal of Economics* 117, 269–96.

Constantinides, G.M. and Duffie, D. 1996. Asset pricing with heterogeneous consumers. *Journal of Political Economy* 104, 219–40.

Copeland, T.E. and Galai, D. 1983. Information effects and the bid–ask spread. *Journal of Finance* 38, 1457–69.

Courtadon, G. 1982. The pricing of options on default-free bonds. *Journal of Financial and Quantitative Analysis* 17, 75–100.

Coval, J. and Stafford, E. 2007. Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics* 86, 479–512.

Cox, J.C., Ingersoll, J.E. and Ross, S.A. 1980. An analysis of variable rate loan contracts. *Journal of Finance* 35, 389–403.

Cox, J.C., Ingersoll, J.E. and Ross, S.A. 1985a. An intertemporal general equilibrium model of asset prices. *Econometrica* 53, 363–84.

Cox, J.C., Ingersoll, J.E. and Ross, S.A. 1985b. A theory of the term structure of interest rates. *Econometrica* 53, 385–407.

Cox, J.C. and Ross, S.A. 1976a. A survey of some new results in financial option pricing theory. *Journal of Finance* 31, 383–402.

Cox, J.C. and Ross, S.A. 1976b. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3, 145–66.

Cox, J.C., Ross, S.A. and Rubinstein, M. 1979. Option pricing: a simplified approach. *Journal of Financial Economics* 7, 229–63.

Dai, Q. and Singleton, K.J. 2000. Specification analysis of affine term structure models. *Journal of Finance* 55, 1943–78.

Daniel, K.D., Hirshleifer, D. and Subrahmanyam, A. 1998. Investor psychology and security market under- and over-reactions. *Journal of Finance* 53, 1839–86.

Daniel, K.D., Hirshleifer, D. and Subrahmanyam, A. 2001. Overconfidence, arbitrage, and equilibrium asset pricing. *Journal of Finance* 56, 921–65.

De Long, J.B., Shleifer, A., Summers, L.H. and Waldman, R.J. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98, 703–38.

DeBondt, W.F.M. and Thaler, R. 1985. Does the stock market overreact? *Journal of Finance* 40, 793–805.

DeMarzo, P.M. and Fishman, M.J. 2007. Agency and optimal investment dynamics. *Review of Financial Studies* 20, 151–88.

Diamond, D.W. and Verrecchia, R.E. 1981. Information aggregation in a noisy expectations economy. *Journal of Financial Economics* 9, 221–35.

Diether, K.B., Malloy, C.J. and Scherbina, A. 2002. Differences in opinion and the cross section of stock returns. *Journal of Finance* 57, 2113–41.

Duffie, D. 1996. Special repo rates. *Journal of Finance* 51, 493–526.

Duffie, D., Garleanu, N. and Pedersen, L.H. 2002. Securities lending, shorting, and pricing. *Journal of Financial Economics* 66, 307–39.

Duffie, D., Garleanu, N. and Pedersen, L.H. 2005. Over-the-counter markets. *Econometrica* 73, 1815–47.

Duffie, D. and Kan, R. 1996. A yield-factor model of interest rates. *Mathematical Finance* 6, 379–406.

Duffie, D. and Rahi, R. 1995. Financial market innovation and security design: an introduction. *Journal of Economic Theory* 65, 1–42.

Duffie, D. and Singleton, K.J. 1999. Modeling term structures of defaultable bonds. *Review of Financial Studies* 12, 687–720.

Dumas, B. 1989. Two-person dynamic equilibrium in the capital market. *Review of Financial Studies* 2, 157–88.

Dybvig, P.H. and Ross, S.A. 1985. Yes, the APT is testable. *Journal of Finance* 40, 1173–88.

Dybvig, P.H. and Zender, J.F. 1991. Capital structure and dividend irrelevance with asymmetric information. *Review of Financial Studies* 4, 201–19.

Eckbo, E.B. 1986. Mergers and the market for corporate control: the Canadian evidence. *Canadian Journal of Economics* 19, 236–60.

Ellis, K., Michaely, R. and O’Hara, M. 2002. The making of a dealer market: from entry to equilibrium in the trading of Nasdaq stocks. *Journal of Finance* 57, 2289–316.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Epstein, L.G. and Zin, S.E. 1991. Substitution, risk aversion, and temporal behavior of consumption and

asset returns: an empirical analysis. *Journal of Political Economy* 99, 263–86.

Evans, M.D.D. and Lyons, R.K. 2002. Order flow and exchange rate dynamics. *Journal of Political Economy* 110, 170–80.

Fama, E.F. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25, 383–417.

Fama, E.F. and French, K.R. 1988. Permanent and temporary components of stock prices. *Journal of Political Economy* 96, 246–73.

Fama, E.F. and French, K.R. 1992. The cross-section of expected stock returns. *Journal of Finance* 47, 427–65.

Fama, E.F. and French, K.R. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 5–56.

Fama, E.F. and French, K.R. 1998. Value versus growth: the international evidence. *Journal of Finance* 53, 1975–79.

Fama, E.F. and French, K.R. 2002. Testing trade-off and pecking order predictions about dividends and debt. *Review of Financial Studies* 15, 1–33.

Fama, E.F. and MacBeth, J.D. 1973. Risk, return and equilibrium: empirical tests. *Journal of Political Economy* 81, 607–36.

Fazzari, S.M., Hubbard, R.G. and Petersen, B.C. 1988. Financing constraints and corporate investment. *Brookings Papers on Economic Activity* 1, 141–95.

Fischhoff, B., Slovic, P. and Lichtenstein, S. 1977. Knowing with certainty: the appropriateness of extreme confidence. *Journal of Experimental Psychology* 3, 552–64.

Fluck, Z. 1998. Optimal financial contracting: debt versus outside equity. *Review of Financial Studies* 11, 383–418.

Foster, F.D. and Viswanathan, S. 1990. A theory of the intraday variations in volume, variance, and trading costs in securities markets. *Review of Financial Studies* 3, 593–624.

Foucault, T. 1999. Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets* 2, 99–134.

French, K.R. and Poterba, J.M. 1991. Investor diversification and international equity markets. *American Economic Review* 81, 222–6.

French, K.R. and Roll, R. 1986. Stock return variances: the arrival of information and reaction of traders. *Journal of Financial Economics* 17, 5–26.

Gale, D. and Hellwig, M. 1985. Incentive-compatible debt contracts: the one-period problem. *Review of Economic Studies* 52, 647–63.

Garman, M.B. 1976. Market microstructure. *Journal of Financial Economics* 3, 257–75.

Gennotte, G. and Leland, H. 1990. Market liquidity, hedging and crashes. *American Economic Review* 80, 999–1021.

Gertler, M. 1992. Financial capacity and output fluctuations in an economy with multi-period financial relationships. *Review of Economic Studies* 59, 455–72.

Gervais, S., Kaniel, R. and Mingelgrin, D.H. 2001. The high-volume return premium. *Journal of Finance* 56, 877–919.

Gibbons, M. and Ramaswamy, K. 1993. A test of the Cox, Ingersoll and Ross model of the term structure. *Review of Financial Studies* 6, 619–58.

Gilchrist, S. and Himmelberg, C.P. 1995. Evidence on the role of cash flow for investment. *Journal of Monetary Economics* 36, 531–72.

Glosten, L.R. 1989. Insider trading, liquidity, and the role of the monopolist specialist. *Journal of Business* 62, 211–35.

Glosten, L.R. 1994. Is the electronic open limit order book inevitable? *Journal of Finance* 49, 1127–61.

Glosten, L.R. and Harris, L.E. 1988. Estimating the components of the bid–ask spread. *Journal of Financial Economics* 21, 123–42.

Glosten, L.R. and Milgrom, P.R. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14, 71–100.

Goettler, R.L., Parlour, C.A. and Rajan, U. 2005. Equilibrium in a dynamic limit order market. *Journal of Finance* 60, 2149–92.

Goldstein, M.A. and Kavajecz, K.A. 2000. Eighths, sixteenths and market depth: changes in tick size

and liquidity provision on the NYSE. *Journal of Financial Economics* 56, 125–49.

Goldstein, R.S. 2000. The term structure of interest rates as a random field. *Review of Financial Studies* 13, 365–84.

Gomes, J.F., Kogan, L. and Zhang, L. 2003. Equilibrium cross section of returns. *Journal of Political Economy* 111, 693–732.

Gompers, P.A. 1995. Optimal investment, monitoring, and the staging of venture capital. *Journal of Finance* 50, 1461–89.

Graham, J.R. 2000. How big are the tax benefits of debt? *Journal of Finance* 55, 1901–41.

Gray, S.F. 1996. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* 42, 27–62.

Grinblatt, M. and Keloharju, M. 2000. The investment behavior and performance of various investor types: a study of Finland's unique data set. *Journal of Financial Economics* 55, 43–67.

Grinblatt, M. and Keloharju, M. 2001. What makes investors trade? *Journal of Finance* 56, 589–616.

Grossman, S.J. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 31, 573–85.

Grossman, S.J. 1988. An analysis of the implications for stock and futures price volatility of program trading and dynamic hedging strategies. *Journal of Business* 61, 275–98.

Grossman, S.J. 1992. The informational role of upstairs and downstairs trading. *Journal of Business* 65, 509–28.

Grossman, S.J. and Hart, O.D. 1980. Takeover bids, the free-rider problem, and the theory of the corporation. *Bell Journal of Economics* 11, 42–64.

Grossman, S.J. and Hart, O.D. 1986. The costs and benefits of ownership: a theory of vertical and lateral integration. *Journal of Political Economy* 94, 691–719.

Grossman, S.J. and Hart, O.D. 1988. One share-one vote and the market for corporate control. *Journal of Financial Economics* 20, 175–202.

Grossman, S.J. and Miller, M.H. 1988. Liquidity and market structure. *Journal of Finance* 43, 617–37.

Grossman, S.J. and Shiller, R.J. 1981. The determinants of the variability of stock market prices. *American Economic Review* 71, 222–7.

Grossman, S.J. and Stiglitz, J.E. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70, 393–408.

Grossman, S.J. and Vila, J.L. 1992. Optimal dynamic trading with leverage constraints. *Journal of Financial and Quantitative Analysis* 27, 151–68.

Grossman, S.J. and Zhou, Z. 1996. Equilibrium analysis of portfolio insurance. *Journal of Finance* 51, 1379–403.

Grundy, B. and McNichols, M. 1989. Trade and revelation of information through prices and direct disclosure. *Review of Financial Studies* 2, 495–526.

Hamilton, J.D. 1988. Rational-expectations econometric analysis of changes in regime: an investigation of the term structure of interest rates. *Journal of Economics Dynamics and Control* 12, 385–423.

Hansen, L.P. and Jagannathan, R. 1991. Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99, 225–62.

Hansen, L.P. and Singleton, K.J. 1983. Stochastic consumption, risk aversion and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249–68.

Harris, M. and Raviv, A. 1988. Corporate control contests and capital structure. *Journal of Financial Economics* 20, 55–86.

Harris, M. and Raviv, A. 1993. Differences of opinion make a horse race. *Review of Financial Studies* 6, 473–506.

Harrison, J.M. and Kreps, D. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economics Theory* 20, 381–408.

Hart, O. and Moore, J. 1998. Default and renegotiation: a dynamic model of debt. *Quarterly Journal of Economics* 113, 1–41.

Harvey, C.R. 1989. Time-varying conditional covariances in tests of asset pricing models. *Journal of Financial Economics* 24, 289–317.

Hasbrouck, J. 1991. Measuring the information content of stock trades. *Journal of Finance* 46, 179–207.

He, H. and Modest, D.M. 1995. Market frictions and consumption-based capital asset pricing. *Journal of Political Economy* 103, 94–117.

He, H. and Wang, J. 1995. Differential information and dynamic behavior of stock trading volume. *Review of Financial Studies* 8, 919–72.

Heaton, J. and Lucas, D. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104, 443–87.

Hellwig, M. 1980. On the aggregation of information in competitive markets. *Journal of Economic Theory* 22, 477–98.

Hennessy, C.A. 2004. Tobin's Q, debt overhang, and investment. *Journal of Finance* 59, 1717–42.

Heston, S.L. 1993. A close form solution for options with stochastic volatility. *Review of Financial Studies* 6, 327–43.

Holmstrom, B. 1999. Managerial incentive problems: a dynamic perspective. *Review of Economic Studies* 66, 169–82.

Hong, H. and Stein, J.C. 1999. A unified theory of underreaction, momentum trading and overreaction in asset markets. *Journal of Finance* 54, 2143–184.

Hong, H. and Wang, J. 2000. Trading and returns under periodic market closures. *Journal of Finance* 55, 297–354.

Hoshi, T., Kashyap, A. and Scharfstein, D. 1991. Corporate structure, liquidity, and investment: evidence from Japanese industrial groups. *Quarterly Journal of Economics* 106, 33–60.

Hovakimian, A., Opler, T. and Titman, S. 2001. The debt–equity choice. *Journal of Financial and Quantitative Analysis* 36, 1–24.

Huang, J. and Wang, J. 1997. Market structure, security prices and informational efficiency. *Macroeconomic Dynamics* 1, 169–205.

Huang, J. and Wang, J. 2006a. Liquidity, asset prices and welfare under costly participation. Working paper, Massachusetts Institute of Technology.

Huang, J. and Wang, J. 2006b. Liquidity and market crashes. Working paper, Massachusetts Institute of Technology.

Hull, J. and White, A. 1987. The pricing of options on assets with stochastic volatility. *Journal of Finance* 42, 281–300.

Hull, J. and White, A. 1994. Numerical procedures for implementing term structure models II: two-factor models. *Journal of Derivatives* 2, 37–48.

Ikenberry, D., Lakonishok, J. and Vermaelen, T. 1995. Market underreaction to open market share repurchases. *Journal of Financial Economics* 39, 181–208.

Jagannathan, R. and Wang, Z. 1996. The conditional CAPM and the cross-section of expected returns. *Journal of Finance* 51, 3–53.

Jarrow, R.A. and Turnbull, S.M. 1995. Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50, 53–85.

Jegadeesh, N. and Titman, S. 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance* 48, 65–91.

Jensen, M.C. 1986. Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review* 76, 323–39.

Jensen, M.C. and Meckling, W.H. 1976. Theory of firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

Johannes, M. 2004. The statistical and economic role of jumps in continuous-time interest rate models. *Journal of Finance* 59, 227–60.

Jones, C.M. and Lamont, O.A. 2002. Short sale constraints and stock returns. *Journal of Financial Economics* 66, 207–39.

Jones, C.P. and Litzenberger, R.H. 1970. Quarterly earnings reports and intermediate stockprice trends. *Journal of Finance* 25, 143–8.

Jung, K., Kim, Y. and Stulz, R.M. 1996. Timing, investment opportunities, managerial discretion, and the security issue decision. *Journal of Financial Economics* 42, 159–86.

Kahneman, D. and Tversky, A. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–31.

Kaplan, S.N. and Stromberg, P. 2001. Financial contracting theory meets the real world: an empirical analysis of venture capital contracts. *Review of Economic Studies* 70, 281–315.

Kennedy, D.P. 1994. The term structure of interest rates as a Gaussian random field. *Mathematical Finance* 4, 247–58.

Knight, F.H. 1936. The quantity of capital and the rate of interest. *Journal of Political Economy* 44, 433–63, 612–42.

Kogan, L. 2001. An equilibrium model of irreversible investment. *Journal of Financial Economics* 62, 201–45.

Kogan, L., Ross, S.A., Wang, J. and Westerfield, M.M. 2006. The price impact and survival of irrational traders. *Journal of Finance* 61, 195–229.

Kothari, S.P., Shanken, J. and Sloan, R.G. 1995. Another look at the cross-section of expected returns. *Journal of Finance* 50, 185–224.

Krishnamurthy, A. 2002. The bond/old-bond spread. *Journal of Financial Economics* 66, 463–506.

Kyle, A.S. 1985. Continuous auctions and insider trading. *Econometrica* 53, 1315–35.

Kyle, A.S. and Xiong, W. 2001. Contagion as a wealth effect. *Journal of Finance* 56, 1401–40.

Lando, D. 1998. On Cox processes and credit-risky securities. *Review of Derivatives Research* 2, 99–120.

Lang, L., Ofek, E. and Stulz, R.M. 1996. Leverage, investment, and firm growth. *Journal of Financial Economics* 40, 3–29.

Langetieg, T.C. 1980. A multivariate model of the term structure of interest rates. *Journal of Finance* 35, 71–97.

Lee, C.M.C., Shleifer, A. and Thaler, R.H. 1991. Investor sentiment and the closed-end fund puzzle. *Journal of Finance* 46, 75–109.

Lehmann, B.N. and Modest, D.M. 1988. The empirical foundations of the arbitrage pricing theory. *Journal of Financial Economics* 21, 213–54.

Leland, H.E. 1994. Risky debt, bond covenants and optimal capital structure. *Journal of Finance* 49, 1213–52.

Leland, H.E. and Pyle, D.H. 1977. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance* 32, 371–87.

Leland, H.E. and Toft, K.B. 1996. Optimal capital structure, endogenous bankruptcy and the term structure of credit spreads. *Journal of Finance* 50, 789–819.

LeRoy, S.F. and Porter, R.D. 1981. The present-value relation: tests based on implied variance bounds. *Econometrica* 49, 555–74.

Lesmond, D.A., Schill, M.J. and Zhou, C. 2004. The illusory nature of momentum profits. *Journal of Financial Economics* 71, 349–80.

Lettau, M. and Ludvigson, S. 2001. Resurrecting the (C)CAPM: a cross-sectional test when risk premia are time-varying. *Journal of Political Economy* 109, 1238–87.

Lewellen, J. and Nagel, S. 2006. The conditional CAPM does not explain asset-pricing anomalies. *Journal of Financial Economics* 82, 289–314.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 13–37.

Litterman, R. and Scheinkman, J. 1991. Common factors affecting bond returns. *Journal of Fixed Income* 1, 54–61.

Liu, J., Pan, J. and Wang, T. 2005. An equilibrium model of rare-event premia and its implication for option smirks. *Review of Financial Studies* 18, 131–64.

Llorente, G., Michaely, R., Saar, G. and Wang, J. 2002. Dynamic volume–return relation of individual stocks. *Review of Financial Studies* 15, 1005–47.

Lo, A.W. and Mackinlay, A.C. 1988. Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies* 1, 41–66.

Lo, A.W. and MacKinlay, A.C. 1990. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* 3, 175–205.

Lo, A.W., Mamaysky, H. and Wang, J. 2004. Asset prices and trading volume under fixed transactions costs. *Journal of Political Economy* 112, 1054–90.

Lo, A.W. and Wang, J. 2006. Trading volume: implications of an intertemporal capital asset pricing model. *Journal of Finance* 61, 2805–40.

Longstaff, F.A. 1989. A nonlinear general equilibrium model of the term structure of interest rates.

Journal of Financial Economics 23, 195–224.

Longstaff, F.A. and Schwartz, E.S. 1992. Interest rate volatility and the term structure: a two factor general equilibrium model. *Journal of Finance* 47, 1259–82.

Longstaff, F.A. and Schwartz, E.S. 1995. A simple approach to valuing risky fixed and floating rate debt. *Journal of Finance* 50, 789–819.

Lucas, R.E., Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

Luttmer, E.G.J. 1999. What level of fixed costs can reconcile consumption and stock returns? *Journal of Political Economy* 107, 969–97.

Lyons, R.K. 1995. Tests of microstructural hypotheses in the foreign exchange market. *Journal of Financial Economics* 39, 321–51.

Madhavan, A. and Smidt, S. 1993. An analysis of changes in specialist inventories and quotations. *Journal of Finance* 48, 1595–628.

Maksimovic, V. and Titman, S. 1991. Financial policy and reputation for product quality. *Review of Financial Studies* 4, 175–200.

Malkiel, B.G. 1977. The valuation of closed-end investment company shares. *Journal of Finance* 32, 847–59.

Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7, 77–91.

McConnell, J.J. and Muscarella, C.J. 1985. Corporate capital expenditure decisions and the market value of the firm. *Journal of Financial Economics* 14, 399–422.

Mehra, R. and Prescott, E.C. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Melino, A. and Turnbull, S.M. 1990. Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* 45, 239–65.

Merton, R.C. 1973. The theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–83.

Merton, R.C. 1974. On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* 29, 449–70.

Merton, R.C. 1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–44.

Merton, R.C. 1981. On market timing and investment performance part I: an equilibrium theory of value for market forecasts. *Journal of Business* 54, 363–406.

Merton, R.C. 1987. A simple model of capital market equilibrium with incomplete information. *Journal of Finance* 42, 483–510.

Milgrom, P. and Stokey, N. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26, 17–27.

Miller, E.M. 1977. Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32, 1151–68.

Miller, M.H. 1977. Debt and taxes. *Journal of Finance* 32, 261–76.

Mitchell, M., Pedersen, L.H. and Pulvino, T. 2007. Slow moving capital. *American Economic Review* 97, 215–20.

Modigliani, F. and Miller, M.H. 1958. The cost of capital, corporate finance and the theory of investment. *American Economic Review* 48, 261–97.

Modigliani, F. and Miller, M.H. 1963. Corporate income taxes and the cost of capital: a correction. *American Economic Review* 53, 433–43.

Myers, S.C. 1977. Determinants of corporate borrowing. *Journal of Financial Economics* 5, 147–76.

Myers, S.C. 2000. Outside equity. *Journal of Finance* 55, 1005–37.

Myers, S.C. and Majluf, N.S. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13, 187–221.

Narayanan, M.P. 1985. Managerial incentives for short-term results. *Journal of Finance* 40, 1469–84.

O’Hara, M. 2003. Presidential address: liquidity and price discovery. *Journal of Finance* 58, 1335–54.

Odean, T. 1998. Are investors reluctant to realize their losses? *Journal of Finance* 53, 1775–98.

Pan, J. 2002. The jump-risk premia implicit in options: evidence from an integrated time-series study. *Journal of Financial Economics* 63, 3–50.

Pastor, L. and Stambaugh, R. 2003. Liquidity risk and expected stock returns. *Journal of Political Economy* 113, 642–85.

Petkova, R. and Zhang, L. 2005. Is value riskier than growth? *Journal of Financial Economics* 78, 187–202.

Phillips, G.M. 1995. Increased debt and industry product markets: an empirical analysis. *Journal of Financial Economics* 37, 189–238.

Piazzesi, M. 2005. Bond yields and the Federal Reserve. *Journal of Political Economy* 113, 311–44.

Poterba, J. and Summers, L.J. 1988. Mean reversion in stock prices: evidence and implications. *Journal of Financial Economics* 22, 27–59.

Rajan, R. and Zingales, L. 1995. What do we know about capital structure? Some evidence from international data. *Journal of Finance* 50, 1421–60.

Ready, M.J. 1999. The specialist's discretion: stopped orders and price improvement. *Review of Financial Studies* 12, 1075–112.

Rietz, T.A. 1988. The equity risk premium: a solution. *Journal of Monetary Economics* 21, 117–31.

Rock, K. 1990. The specialist's order book and price anomalies. Working paper, Harvard Business School.

Roll, R. 1977. A critique of the asset pricing theory tests. Part I: on past and potential testability of the theory. *Journal of Financial Economics* 4, 129–76.

Roll, R. 1984a. Orange juice and weather. *American Economic Review* 74, 861–80.

Roll, R. 1984b. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* 39, 1127–39.

Roll, R. 1988. R^2. *Journal of Finance* 43, 541–66.

Romer, D. 1993. Rational asset-price movements without news. *American Economic Review* 83, 1112–30.

Ross, S.A. 1973. The economic theory of agency: the principals problems. *American Economic Review* 63, 134–9.

Ross, S.A. 1976. Options and efficiency. *Quarterly Journal of Economics* 90, 75–89.

Ross, S.A. 1977. The determination of financial structure: the incentive signaling approach. *Bell Journal of Economics* 8, 23–40.

Ross, S.A. 2002. Neoclassical finance, alternative finance, and the closed end fund puzzle. *European Financial Management* 8, 129–37.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7, 407–25.

Rubinstein, M. 1994. Implied binomial tress. *Journal of Finance* 49, 771–818.

Saa-Requejo, J. and Santa-Clara, P. 1999. Bond pricing with default risk. Working paper, UCLA.

Samuelson, P.A. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 41–9.

Sandroni, A. 2000. Do markets favor agents able to make accurate predictions. *Econometrica* 68, 1303–41.

Santa-Clara, P. and Sornette, D. 2001. The dynamics of the forward interest rate curve with stochastic string shocks. *Review of Financial Studies* 14, 149–85.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Schaefer, S.M. and Schwartz, E.S. 1984. A two-factor model of the term structure: an approximate analytical solution. *Journal of Financial and Quantitative Analysis* 19, 413–24.

Scheinkman, J.A. and Xiong, W. 2003. Overconfidence and speculative bubbles. *Journal of Political Economy* 111, 1183–219.

Scott, L.O. 1997. Pricing stock options in a jump-diffusion model with stochastic volatility and interest rates: applications of Fourier inversion methods. *Mathematical Finance* 7, 413–26.

Seppi, D.J. 1997. Liquidity provision with limit orders and a strategic specialist. *Review of Financial Studies* 10, 103–50.

Shanken, J. 1982. The arbitrage pricing theory? It is testable. *Journal of Finance* 37, 1129–40.

Shanken, J. 1987. Multivariate proxies and asset pricing relations: living with the Roll critique. *Journal*

of Financial Economics 18, 91–110.

Shanken, J. 1990. Intertemporal asset pricing: an empirical investigation. *Journal of Econometrics* 45, 99–120.

Sharpe, W.F. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71, 421–36.

Shleifer, A. and Vishny, R.W. 1997. The limits of arbitrage. *Journal of Finance* 52, 35–55.

Stambaugh, R.F. 1982. On the exclusion of assets from tests of the two parameter model. *Journal of Financial Economics* 10, 235–68.

Stanton, R. 1997. A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance* 52, 1973–2002.

Stein, E.M. and Stein, C.J. 1991. Stock price distributions with stochastic volatility: an analytic approach. *Review of Financial Studies* 4, 727–52.

Stoll, H.R. 1979. *Regulation of Securities markets: An Examination of the Effects of Increased Competition*. New York: Graduate School of Business, New York University.

Stulz, R. 1990. Managerial discretion and optimal financing policies. *Journal of Financial Economics* 26, 3–27.

Sundaresan, S.M. 1989. Intertemporally dependent preferences and the volatility of consumption and wealth. *Review of Financial Studies* 2, 73–88.

Taggart, R.A. 1977. A model of corporate financing decision. *Journal of Finance* 32, 1467–84.

Titman, S. and Wessels, R. 1988. The determinants of capital structure choice. *Journal of Finance* 43, 1–19.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, 65–86.

Townsend, R.M. 1978. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21, 265–93.

Tuckman, B. and Vila, J.L. 1992. Arbitrage with holding costs: a utility based approach. *Journal of Finance* 47, 1283–302.

Vasicek, O. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5, 177–88.

Vayanos, D. 1998. Transaction costs and asset prices: a dynamic equilibrium model. *Review of Financial Studies* 11, 1–58.

Vayanos, D. 1999. Strategic trading and welfare in a dynamic market. *Review of Economic Studies* 66, 219–54.

Vayanos, D. and Weill, P.O. 2006. A search-based theory of the on-the-run phenomenon. Working Paper No. 12670. Washington, DC: NBER.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. New York: Wiley.

Wahal, S. 1997. Entry, exit, market makers and the bid-ask spread. *Review of Financial Studies* 10, 871–901.

Wald, J. 1999. Capital structure with dividend restrictions. *Journal of Corporate Finance* 5, 193–208.

Wang, J. 1993. A model of intertemporal asset prices under asymmetric information. *Review of Economic Studies* 60, 249–82.

Wang, J. 1994. A model of competitive stock trading volume. *Journal of Political Economy* 102, 127–68.

Wang, J. 1996. The term structure of interest rates in a pure exchange economy with heterogeneous investors. *Journal of Financial Economics* 41, 75–110.

Wang, K.Q. 2003. Asset pricing with conditioning information: a new test. *Journal of Finance* 58, 161–96.

Weil, P. 1989. The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics* 24, 401–21.

Weinstein, N.D. 1980. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* 39, 806–20.

Weiss, L.A. and Wruck, K.H. 1998. Information problems, conflicts of interest and asset stripping: chapter 11's failure in the case of eastern airlines. *Journal of Financial Economics* 48, 55–97.

Williamson, O.E. 1964. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

Wilson, R. 1968. On the theory of syndicates. *Econometrica* 36, 119–32.

Yuan, K. 2005. Asymmetric price movements and borrowing constraints: a rational expectations equilibrium model of crises, contagion, and confusion. *Journal of Finance* 60, 379–411.

Zhang, L. 2006. Anomalies. Working Paper No. 11322. Cambridge, MA: NBER.

Zingales, L. 1998. Survival of the fittest or the fattest? Exit and financing in the trucking industry. *Journal of Finance* 53, 905–38.

Zwiebel, J. 1996. Dynamic capital structure under managerial entrenchment. *American Economic Review* 86, 1197–215.

How to cite this article

Wang, Jiang. "finance (new developments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000291> doi:10.1057/9780230226203.0567

The New Palgrave Dictionary of Economics Online

finance

Stephen A. Ross
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The neoclassical theory of finance is based on the study of (*a*) efficient markets, meaning markets that use all available information in setting prices, (*b*) the trade-off between return and risk, (*c*) option pricing and the principle of no arbitrage, and (*d*) corporate finance, that is, the structure of financial claims issued by companies. This article surveys these theories and their empirical support and it also identifies certain empirical regularities unexplained by the neoclassical theory that are being addressed by theories of asymmetric information.

Keywords

arbitrage; arbitrage pricing theory; Arrow-Debreu model; asymmetric information; Black-Scholes model; capital asset pricing model; consumption beta model; corporate finance; corporate taxation; efficiency; efficient markets hypothesis; finance; Hicks, J.; interest rates; martingales; mean variance analysis; Merton, R.; Modigliani-Miller theorem; option pricing; Pareto efficiency; principal and agent; productive efficiency; rational expectations; risk and return; risk premium; security market line equation; signalling models; Tobin, J.; von Neumann-Morgenstern utility function

Article

Finance is a subfield of economics distinguished by both its focus and its methodology. The primary focus of finance is the workings of the capital markets and the supply and the pricing of capital assets. The methodology of finance is the use of close substitutes to price financial contracts and instruments. This methodology is applied to value instruments whose characteristics extend across time and whose payoffs depend upon the resolution of uncertainty.

Finance is not terribly concerned with the problems that arise in a barter economy or, for that matter, in a static and certain world. But, once the element of time is introduced, transactions develop a dual side to them. When a loan is made, the amount and the terms are recorded to insure that repayment can be enforced. The piece of paper or the computer entry that describes and legally binds the borrower to repay the loan can now trade on its own as a ‘bearer’ instrument. It is at the point when debts were first traded that capital markets and the subject of finance began.

The study of finance is enriched by having a large body of evolving data and market lore and some powerful and, at times, competing intuitions. These intuitions are used to structure our understanding of the data and the markets which generate it. The modern tradition in finance began with the development of well-articulated models and theories to explore these intuitions and render them susceptible to empirical testing.

While the subject of finance is anything but complete, it is now possible to recognize the broad outlines of what might be called the neoclassical theory. In the discussion which follows we will group the subjects under four main headings corresponding with four basic intuitions. The first topic is efficient markets, which was also the first area of finance that matured into a science. Next come the twinned subjects of return and risk. This leads naturally into option pricing theory and the central intuition of pricing close substitutes by the absence of arbitrage. The principle of no arbitrage is used to tie together the major subfields of finance. The fourth section looks at corporate finance from its well-developed form as a consequence of no arbitrage to its current probings. A short conclusion ends the entry.

Efficient markets

The word efficient is too useful to be monopolized by a single meaning in economics. As a consequence, it has a variety of related but distinct meanings. In neoclassical equilibrium theory efficiency refers to Pareto efficiency. A system is Pareto efficient if there is no way to improve the well being of any one individual without making someone worse off. Productive efficiency is an implication of Pareto efficiency. An economy is productively efficient if it is not possible to produce more of any one good or service without lowering the output of some other.

In finance the word efficiency has taken on quite a different meaning. A capital market is said to be (informationally) efficient if it utilizes all of the available information in setting the prices of assets. This definition is purposely vague and it is designed more to capture an intuition than to state a formal mathematical result. The basic intuition of efficient markets is that individual traders process the information that is available to them and take positions in assets in response to their information as well as to their personal situations. The market price aggregates this diverse information and in that sense it ‘reflects’ the available information.

The relation between the definitions of efficiency is not obvious, but it is not unreasonable to think of the efficient markets definition of finance as being a requirement for a competitive economy to be Pareto efficient. Presumably, if prices did not depend on the information available to the economy, then it would only be by accident that they could be set in such a way as to guarantee a Pareto efficient allocation (at least with respect to the commonly held information).

If the capital market is competitive and efficient, then neoclassical reasoning implies that the return that an investor expects to get on an investment in an asset will be equal to the opportunity cost of using the funds. The exact specification of the opportunity cost is the subject of the section on risk and return, but for the moment we can observe that investing in risky assets should carry with it some additional measure of return beyond that on riskless assets to induce risk averse investors to part with their funds. For now we will defer the measurement of this risk premium, and simply represent the opportunity cost by the letter ‘*r*’.

In much of the early empirical work on efficient markets no attempt was made to measure risk premia, and the opportunity cost of investing was set equal to the riskless rate of interest. This can be justified either by assuming that there are risk neutral investors who are indifferent to risk (or, as we shall see, by assuming that the asset's risk is diversified away in large portfolios). Whatever the rationale, to focus on the topic of efficient markets rather than on the pricing of risk, we will let *r* be the riskless interest rate.

If R_t denotes the total return on the asset – capital gains as well as payouts – over a holding period from *t* to $t + 1$, then the efficient markets hypothesis (EMH) asserts that

$$E(R_t | I_t) = (1 + r_t), \tag{1}$$

where *E* is the expectation taken with respect to a given information set I_t, that is available at time *t* (and that includes r_t). An alternative formulation of the basic EMH equation is in terms of prices. For an asset with no payouts, since

$$R_t \equiv p_{t+1} / p_t,$$

we can rewrite (1) as

$$E(p_{t+1} | I_t) = (1 + r_t)\, p_t, \tag{2}$$

or, equivalently, discounted prices must follow the martingale,

$$\frac{1}{(1+r_t)}E(p_{t+1}|I_t) = p_t.$$

The EMH is given empirical content by specifying the information set that is used to determine prices. Harry Roberts (1967) first coined the terms which have come to describe the categories of information sets and, concomitantly, of efficient market theories that are employed in empirical work. Fama (1970) subsequently articulated them in the form which we now use. These categories describe a hierarchy of nested information sets. As we go up the hierarchy from the smallest to the biggest set (i.e. from coarser to finer partitions) we are requiring efficiency with respect to increasing amounts of information. At the far end of the spectrum is strong-form efficiency. Strong-form efficiency asserts that the information set, I_t, used by the market to set prices at each date t contains all of the available information that could possibly be relevant to pricing the asset. Not only is all publicly available information embodied in the price, but all privately held information as well.

A substantial notch down from strong-form efficiency is semistrong-form efficiency. A market is efficient in the semistrong sense if it uses all of the publicly available information. The important distinction is that the information set, I_t, is not assumed to include privately held information, i.e. information that has not been made public. Making this distinction precise is possible in formal models but categorizing information as publicly available or not can be subjective. Presumably, accounting information such as the income statements and the balance sheets of the firm is publicly available, as is any other information that the government mandates should be released such as the stock holdings of the top executives in the firm. Presumably, too, the true but unrevealed intention of a major stockholder would fall into the category of private information. In between these extremes is a large grey area. The tendency in the empirical literature has been to take a purist's view of semistrong efficiency, and to adopt the position that if the information was in the public domain then it was available to the public and should be reflected in prices. This ignores the cost of acquiring the information, but the intuitive justification for this position is that the costs of acquiring such public information are small compared to the potential rewards. Thus, while the government mandated and publicly reported trades of the top executives require a bit more effort to obtain in a timely fashion than some average of their past holdings, such trades, when reported, would fall squarely within the realm of publicly available information under the semistrong version of the EMH.

If the asset is traded on an organized exchange, then of all the information that is clearly available to the public, none is as accessible and cheap as its past price history. At the bottom of the ladder in the efficiency hierarchy, weak-form efficiency requires only that the current and past price history be incorporated in the information set. If there is empirical validity to the EMH then, at the very least, the market for an asset should be weak-form efficient, that is, efficient with respect to its own past price history.

Empirical testing

The empirical implications of efficiency with respect to a particular information set are that the current price of the asset embodies all of the information in that set. Since the categories of information sets are nested, rejection of any one type, say, weak-form efficiency, implies the rejection of all stronger forms.

For example, according to weak-form efficiency, the current price of an asset embodies all of the information contained in the past price history. This implies that,

$$E(R_t|R_{t-1}, R_{t-2}, \ldots) = (1 + r_2), \tag{3}$$

or, in price terms,

$$E(p_{t+1}|p_t, p_{t-1}, \ldots) = (1 + r_t)\, p_t.$$

The most dramatic consequence of the EMH and certainly the one that receives the most attention from the public, is that it denies the possibility of successful trading schemes. If, for example, the market is weak-form efficient, then an investor who makes use of the ‘technical’ information of past prices can only expect to receive a return of the opportunity cost (1 $+r_t$). No amount of clever manipulation of the past information can improve this result.

As a test of weak-form efficiency, then, we could test (although not as a simple regression) the null hypothesis that

$$H_0: E(p_{t+1}|p_t, p_{t-1}) = \beta_0 + \beta_1 p_t + \beta_2 p_{t-1}, \tag{4}$$

where

$$\beta_0 = 0\beta_1 = (1 + r_t).$$

and

$$\beta_2 = 0.$$

The important feature of this hypothesis is that it tells what information does *not* play a role (given r_t), namely the lagged price, p_{t-1}. If the coefficient β_2 should prove to be statistically significant, then this would constitute a rejection of the weak-form EMH.

The other empirical implication of the EMH that is often cited as a defining characteristic is that an efficient price series should ‘move randomly’. The precise meaning of this in our context is that price changes should be serially uncorrelated.

Consider the serial covariance between two adjacent rates of return,

$$\mathrm{cov}(R_{t+1}, R_t) \equiv E([R_{t+1} - E(R_{t+1})][R_t - E(R_t)]). = E(R_{t+1}[R_t - E(R_t)]) = E(E(R_{t+1}|R_t)[R_t - E(R_t)]) \tag{5}$$

In equation (5), since we have not specified the information set with respect to which the expectations are to be taken, they are unconditional expectations. Under weak-form efficiency, the information set will contain the past rates of return. Suppose that the (expected) opportunity cost, e.g. the interest rater r, independent of past returns on the asset or that changes are of a second order of magnitude. This would occur, for example, if we held r_t constant at r. In such a case, since weak-form efficiency implies that I_{t+1} contains R_t, we have

$$E(R_{t+1}|R_t) = E[E(R_{t+1}|I_{t+1})|R_t] = E[(1 + r_{t+1})|R_t] = E(1 + r_{t+1}), \tag{6}$$

the unconditional expectation of next period's opportunity cost. Putting (5) and (6) together yields,

$$\mathrm{cov}(R_{t+1}, R_t) = E(1 + r_{t+1})E[R_t - E(R_t)] = 0. \tag{7}$$

which is to say that rates of return are serially uncorrelated.

Tests of the EMH are legion and by and large they have been supportive. The early tests were essentially tests of the inability of trading schemes or of the random walk nature of prices, which implies that actual rates of return are serially uncorrelated. While the EMH does not imply that prices follow a random walk, such a price process is consistent with market efficiency. Alternatively, unable to specify closely the opportunity cost, some of the early tests took refuge in the view that it must be positive, which leads to a submartingale model for prices,

$$E(p_{t+1}|I_t) \geq p_t. \tag{8}$$

The lack of a specification of the opportunity cost characterizes the early tests (see Cowles (1933), Granger and Morgenstern (1962) and Cootner (1964) and see Roll's (1984) study of the orange juice futures market for a modern example of such a test). Following Fama (1970), the literature shifted to a concern for specifying the opportunity cost and, in this sense, empirical tests became joint tests of the EMH and of the correct specification of the opportunity cost and its attendant theory.

In terms of the information hierarchy, the general message that emerged from the testing is that the market does appear to be consistent with weak-form efficiency. Tests of stronger forms of efficiency, though, have produced mixed results. Fama, Fisher, Jensen, and Roll (1969) introduced a new methodology to test semistrong efficiency and applied it to stock splits. They observed that the residuals from a simple regression of a stock's returns on a market index would measure the portion of the return that was not attributable to market movements. By adding the residuals over a period of time, the resulting cumulative residual measures the total return over that period that is attributable to nonmarket movements. If a stock splits, say, 2 for 1, then under semistrong efficiency its price should split in proportion. i.e., halve for a 2 for 1 split. Using this ‘event study’ approach, Fama et al. verified that stock split data was consistent with semistrong efficiency. The event study methodology they introduced and the use of cumulative residuals (averaged over firms) has become the standard method for examining the impact of information on stock returns.

By contrast with their supportive findings, Jaffé (1974), for example, found that a rule based on the publicly released information about insider trades produced abnormal returns. These results and others like them (see the section on *Risk and Return* below) have been much debated and no final verdict on the matter is likely.

Recently a more interesting empirical challenge to the EMH has come from a different tack. Shiller (1981), has argued that the traditional statistical tests that have been employed are too weak to examine the EMH properly and, moreover, that they are misfocused. Shiller adopts the intuitive perspective that if stock prices are discounted expected dividends, then they ought not to vary over time as much as actual dividends. He argues that since the price is an expectation of the dividends and future price, what actually occurs will be this expectation plus the error in the forecast and should be more variable than the price. This leads him to formulate statistical tests of the EMH based on the volatility of stock prices which are claimed to be more powerful than the traditional (regression based) tests.

An alternative view has been taken by critics of this perspective, notably Kleidon (1986), Flavin (1983), and Marsh and Merton (1986). These critics have taken issue with Shiller's specification of the statistical tests of volatility and, more importantly, with his basic intuition. In particular, they contend that the single realization of dividends and prices that is observed is only one drawing from all of the random possibilities and that the price is based on the expectation taken over all of these possibilities. A little bit of information, then, can have an important influence on the current price. Furthermore, they argue that when the smoothing of dividends and the finite time horizon of the data samples are taken into account, volatility tests do not reject the EMH. The testing of the EMH is taking a new direction because of this work, but, at present, the results are still mixed.

Less cosmic in scope, but perhaps more worrisome is the discovery by French and Roll (1985) that the variance per unit time of market returns over periods when the market is closed (for example, from Tuesday's close to Thursday's close when the market was closed on Wednesday because of a backlog of paperwork) is many times smaller than when it is open. It is difficult to reconcile this result with the requirement that prices reflect information about the cash flows of the assets, unless the generation of fundamental information slows dramatically when the market closes – no matter why it is closed.

Theoretical formulations

The attempts to formalize the EMH as a consistent, analytical economic theory have met with less success than the empirical tests of the hypothesis. The theory can be broken into two parts. The first part is neoclassical and is largely formulated in terms of models in which investors share a common information set. Such models focus on the intertemporal aspects of the theory and the changing shape of the information set.

It has long been recognized that a competitive economy with a single risk neutral investor would lead to the traditional efficient market theories with respect to the information set employed by that investor. More interestingly, Cox, Ingersoll and Ross (1985a), and Lucas (1978) have developed intertemporal rational expectations models each of which is consistent with certain versions of the efficient market theories.

There is, however, an important sense in which these models fail to capture the essential intuition of efficient markets. In informationally efficient markets, prices communicate information to participants. Information possessed by one investor is communicated to another through the influence – however microscopic – that the first investor has on equilibrium prices. In models where investors have homogeneous information sets such information transfer is irrelevant.

A variety of attempts have been made to develop models of financial markets which can deal with such informational issues, but the task is formidable and a satisfactory resolution is not now in hand. This work parallels that of the neoclassical rational expectations view of macroeconomics. This is no accident since the rational expectations school of macroeconomics was very clearly influenced by the intuition of efficiency in finance. The original insight that prices reflect the available information lies at the heart of rational expectations macroeconomics. In this latter work aggregate prices, for example, not only provide the terms of trade for producers, they also inform producers about the aggregate state of production in the economy.

Perhaps the principal difficulty is that models with fully rational investors tend to break down. As investors apply the full scope of their analytical and reasoning talents, the result is an equilibrium in which they lack the incentive to engage in trade. (See Grossman, 1976; Grossman and Stiglitz, 1980; Diamond and Verrecchia, 1981; Milgrom and Stokey, 1982; and Admati, 1985.) The only way out of this bind seems to be to add a discomforting element of irrationality – or an alternative motive for trade from an equilibrium, such as insurance – to the model.

To understand this point, consider a risk-averse individual trading in a market where he or she receives information signals about the ultimate value of the asset being traded and where it is common knowledge that all investors are in the same position. That is not to say that all investors have the same information, rather, it only means that they all begin with the same information, have the same view of the world (Bayesian priors), and then receive signals from the same sort of information generating mechanism. In such a market, the offer to trade on the part of any one investor communicates information to other investors. In particular, it tells them that the individual, based upon his or her information, will be improved by the trade. If all investors are rational they will all feel similarly bettered by trade. But, if the market had been in an equilibrium prior to the receipt of new information, and if it is common knowledge that trade balances, then in the new equilibrium not all of them can be improved. This contradiction can only be resolved by having no further trade upon the receipt of information.

To put the matter in an equivalent form, consider an investor who possesses some special information. Presumably, it is by trading that this information is incorporated into the market price. The above argument implies that the mere announcement of a wish to trade results in a change in prices with no profits for the investor since none will trade at the original prices. If information is costly to acquire and impossible to profit from, then why bother? In other words, if the price reflects the available information possessed by the individual participants, then why gather information if one only needs to look at the price?

The resolution of this dilemma can take many forms, and research will proceed by altering the assumptions that lead to this result. For example, we can drop the assumption about a common prior and let investors come to the markets with different a priori beliefs. We could also drop the assumption that all investors are perfectly rational and introduce ‘noisy’ traders. Lastly, we could drop efficiency and complete markets or integrate insurance motives in other ways.

All of these approaches are being explored but we must leave this discussion with the theory that underlies the incorporation of asymmetric information into securities prices in an unsettled state. The traditional theory that prices reflect the available information is well understood with a representative individual. The theory with asymmetric information is not well understood at all. In short, the exact mechanism by which prices incorporate information is still a mystery and an attendant theory of volume is simply missing.

To conclude, the efficient market paradigm is the backbone of much of financial research and it continues to guide a large body of theoretical and empirical work. Its usefulness is beyond question, but its fine structure is not. In a sense, like much of economics, it remains a central intuition whose analytical representations seem less compelling than the insight itself. This presents more of a problem for theory than for empirical work, but the empirical side is also not without challenge. Although the evidence in support of the efficiency of capital markets is widespread, troublesome pockets of anomalies are growing and the power of the traditional methodology to test the theory is being seriously questioned. Nevertheless, there is currently no competitor for the basic intuition of efficient markets and few insights have proven as fruitful.

Risk and return

The theory of efficient markets leads inexorably to the second central intuition in finance, the trade-off between risk and return. It has long been recognized that risk-averse investors require additional return to bear additional risk. Indeed, this insight goes back to the earliest writings on gambling and it is as much a definition of risk aversion as it is a description of risk-averse behaviour. The contribution made by finance has been to translate this observation into a body of intuition, theory, and empirics on the workings of the capital markets.

The intuition that in a competitive market higher return is accompanied by higher risk owes at least as much to Calvin as it does to Adam Smith, but, in large part the development of capital market theory has been an attempt to explain risk premia, the difference between expected returns and the riskless interest rate. The foundations for the models that would first explain risk premia and that would become the workhorses of financial asset pricing theories were laid by Hicks (1946), Markowitz (1959), and Tobin (1958). These authors developed a rigorous micro-model of individual behaviour in a ‘mean variance’ world where investment portfolios were evaluated in terms of their mean returns and the total variance of their returns. They justified focusing on these two distributional characteristics by assuming either that investors had quadratic von Neumann-Morgenstern utility functions or that asset returns were normally distributed. In such a world, investors would choose mean variance efficient portfolios, i.e., portfolios with the highest mean return for a given level of variance. This observation reduced the study of portfolio choice to the analysis of the properties of the mean variance efficient set. Building on their work, Sharpe (1964), Lintner (1965), and Mossin (1966), all came to the fundamental insight that this micromodel could be aggregated into a simple model of equilibrium in the capital markets, the capital asset pricing model or CAPM.

The mean variance capital asset pricing model (CAPM)

In neoclassical equilibrium models, an investor evaluates an asset in terms of its marginal contribution to his or her portfolio. The decision to alter the proportion of the portfolio invested in an asset will depend on whether the cost of doing so in terms of risk is greater or less than the benefit in expected return. An individual in a personal equilibrium will find the cost at the margin equal to the benefit.

We will assume that a unit addition of an asset to the portfolio can be financed at an interest rate of r. In a mean variance model the net benefit of adding an asset to a portfolio is the additional expected return it brings, E, less the cost of financing it. Such a change, Δx, will augment the expected return on the portfolio, E_p, by the risk premium of the asset, i.e. by the difference between the expected return on the asset, E_i, and the cost of the financing, r,

$$\Delta E_p = (E_i - r)\Delta x. \tag{9}$$

The marginal cost, in terms of risk, of an increase in the holding of an asset is the addition to the total variance of the portfolio occasioned by an increase in the holding of the asset. To compute this increase, let v denote the variance of returns on the current portfolio, let var(i) stand for the variance of asset i's returns, let cov(i, p) denote the covariance between the return of asset i and that of the portfolio, p, and let Δx be the addition in the holding of asset i.

The variance of the portfolio after adding Δx of asset i will be,

$$v + \Delta v = v + 2\Delta x \operatorname{cov}(i, p) + (\Delta x)^2 \operatorname{var}(i),$$

which means the change in the variance is given by

$$\Delta v = 2(\Delta x)\operatorname{cov}(i, p) + (\Delta x)^2 \operatorname{var}(i),$$

and for a small marginal change, Δx this approximates,

$$\Delta v \approx 2(\Delta x)\operatorname{cov}(i, p).$$

The marginal rate of transformation between return and risk, then, is given by

$$\mathrm{MRT} = \frac{\Delta E_p}{\Delta v} = \frac{(E_i - r)\Delta x}{2(\Delta x)\,\mathrm{cov}(i,\ p)} = \frac{(E_i - r)}{2\,\mathrm{cov}(i,\ p)}. \tag{10}$$

An investor will be in a personal equilibrium when this trade-off is equal to his or her personal marginal rate of substitution between return and risk. But, if the portfolio *p* is an optimal one for the investor then it must also have a trade-off between return and risk that is equal to the investor's marginal rate of substitution, and this permits us to use it as a benchmark. Consider, then, the alternative possibility of changing the portfolio position not by changing the amount of asset *i* being held, but rather by changing the amount of the entire portfolio *p* being held, again financing the change by an alteration in the holding of the riskless asset. This is equivalent to leveraging the portfolio of risky assets and altering the amount of the riskless asset so as to continue to satisfy the budget constraint. Such a change will produce a trade-off between return and risk exactly analogous to the one examined above.

$$\mathrm{MRS} = \frac{E_p - r}{2\mathrm{var}(p)}, \tag{11}$$

where we have written this as the marginal rate of substitution, MRS. Since in equilibrium all of the marginal rates of transformation must equal the common marginal rate of substitution, putting these two equations together we have,

$$E_i - r = (E_p - r)\beta_{ip}, \tag{12}$$

where

$$\beta_{ip} \equiv \frac{\mathrm{cov}(i,\ p)}{\mathrm{var}(p)}, \tag{13}$$

the regression coefficient of the returns of asset *i* on the returns of portfolio, *p*. Equation (12) is the famous security market line equation, the SML. It describes the necessary and sufficient condition for a portfolio *p* to be mean variance efficient. It also provides a clear statement of the risk premium, asserting that it is proportional to the asset's beta, β_{ip}. The insight of Sharpe, Lintner and Mossin was the observation that the SML and the mean variance analysis could be aggregated almost without change to a full equilibrium in the

capital market. If we assume that all individuals have the same information and, therefore, see the same mean variance picture, then each individual's efficient portfolio will satisfy equation (12). Since the SML equation is linear in the portfolio holding, *p*, we can simply weight each individual's equation by the proportion of wealth that individual holds in equilibrium, and add up the individual SML's. The result will be an SML equation for the aggregate portfolio, *m*, that is the weighted average of the individual portfolios. In equilibrium, the weighted average of all of the individual portfolios, *m*, is the market portfolio, i.e., the portfolio of all assets held in proportion to their market valuation. In other words, each asset *i*, must lie on the SML with respect to the market,

$$E_i - r = (E_m - r)\beta_{im} \tag{14}$$

which means that the market portfolio, *m*, is a mean variance efficient portfolio.

The geometry of the mean variance analysis is illustrated in Figure 1. The set of mean variance efficient portfolios maps out a mean variance efficient frontier in the mean standard deviation space of Figure 1. Each investor will pick some point on this frontier and that point will be associated with a mean variance efficient portfolio that is suitable for the investor's particular degree of risk aversion. All such portfolios will themselves be portfolios of just two assets: the riskless asset, *r*, and a common portfolio, *p*, of risky assets. This fortunate simplification of the individual portfolio optimization problem is referred to as two fund separation. It implies that the only role for individual preferences lies in choosing the appropriate combination of the risky portfolio, *p*, and the riskless asset, *r*. As a consequence, when we aggregate, the market risk premium, $(E_m - r)/\mathrm{var}(m)$, will be an average of individual measures of risk aversion.

Figure 1

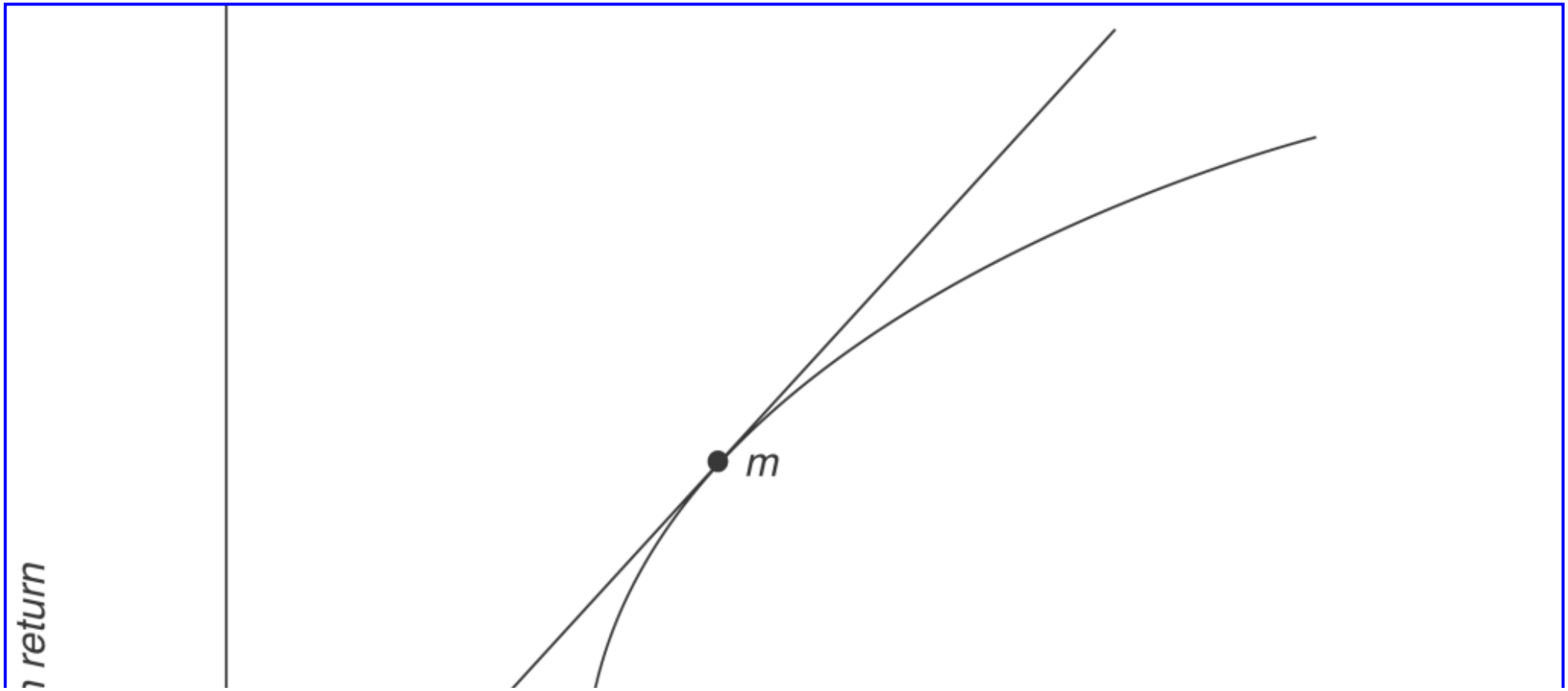

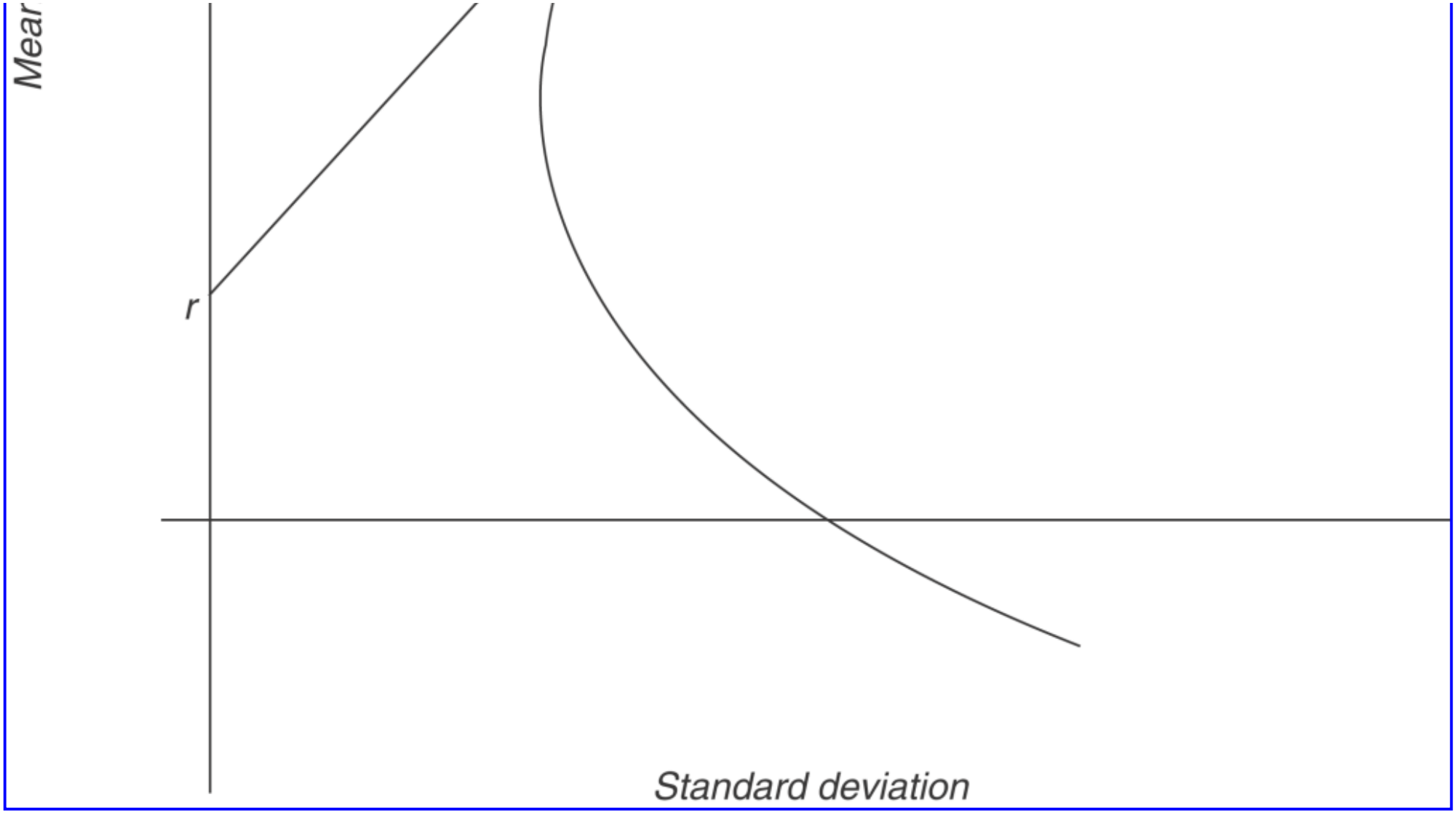

Black (1972) showed that two fund separation would still hold in the mean variance model even if there were no riskless asset. In such a case he found that an efficient portfolio orthogonal-the ‘zero beta portfolio’-to the market portfolio could be found, and that all investors would be able to find their optimal portfolios as combinations of *m* and this zero beta portfolio. In the above development of the CAPM we can simply let *r* be the expected return on a zero beta portfolio.

The necessary and sufficient conditions on return distributions for them to have this two fund separation property - for any concave utility function – were established by Ross (1978a). Ross characterized the class of distribution whose efficient frontier, i.e. the set of portfolios that *some* investor would choose, was spanned by *k* funds, and showed that it extended beyond the normal distributions in the case of *k*=2 fund separation. This work was extended by Chamberlain (1983), who found the class of distributions for which expected utility was a function of just mean and variance for any portfolio as well as for the efficient ones. Cass and Stiglitz (1970) found the conditions on investor utility functions for a similar property to hold regardless of assumptions on return distributions.

It follows immediately from two fund separation that the tangency portfolio, *p*, in Figure 1 must be the market portfolio of risky assets since all investors hold all risky assets in the same proportions. If there is no net supply of the riskless asset then *p* must be the market portfolio, *m*, itself.

The central feature of the CAPM is the mean variance efficiency of the market portfolio and the emergence of the beta coefficient on the market portfolio as the determinant of the risk premium of an asset. Those features of an asset that contribute to its variance but do not affect its covariance with the market will not influence its pricing. Only beta matters for

pricing; the idiosyncratic or unsystematic risk, i.e. that portion which is the residual in the regression of the asset's returns on the market's returns and is therefore orthogonal to the market, playing no role in pricing.

This produces some results that were at first viewed as counter-intuitive. The older view that the risk premium depended on the asset's variance was no longer appropriate, since if one asset had a higher covariance with the market than another, it would have a higher risk premium even if the total variance of its returns were lower. Even more surprising was the implication that a risky asset that was uncorrelated with the market would have no risk premium and would be expected to have the same rate of return as the riskless asset, and that assets that were inversely correlated with the market would actually have expected returns of less than the riskless rate in equilibrium.

These results for the CAPM were supposedly explicated by the twin intuitions of diversification and systematic risk. There could be no premium for bearing unsystematic risk since a large and well diversified portfolio (i.e. one whose asset proportions are not concentrated in a small subset) would eliminate it - presumably by the law of large numbers. This would leave only systematic risk in any optimal portfolio and since this risk cannot be eliminated by diversification, it has to have a risk premium to entice risk averse investors to hold risky assets. From this perspective it becomes clear why an asset that is uncorrelated with the market bears no risk premium. One that is inversely correlated with the market actually offers some insurance against the all pervasive systematic risk and, therefore, there must be a payment for the insurance in the form of a negative risk premium.

There is nothing wrong with this intuition, but it does not fit the CAPM very well. The residuals from the regression of asset returns on the market portfolios are orthogonal to the market, but they could be highly correlated with each other. In fact, they are linearly dependent since when they are weighted by the market proportions they sum to zero. This means the law of large numbers cannot be used to insure that large portfolios of residuals other than the market portfolio will be negligible. But, if that is the case, then the residuals could capture systematic risks not reflected in the market portfolio.

The CAPM was the genesis for countless empirical tests (see, e.g., Black, Jensen, and Scholes, 1972; and Fama and MacBeth, 1973). The latter paper developed the most widely used technique. The general structure of these tests was the combination of the efficient market hypothesis with time series and cross section econometrics. Typically some index of the market, such as the value weighted combination of all stocks would be chosen and a sample of firms would be tested to see if their excess returns, $E - r$, were ‘explained’ in cross-section by their betas on the index, i.e., whether the SML was rejected.

Roll (1977b, 1978) put a stop to this indiscriminate testing by calling into question precisely what was being tested. Roll's critique had two parts. First, he argued that the tests were of very low power and probably could not detect departures from mean variance efficiency. His central point, though, began by noting that tests of the CAPM were tests of the implications of the statement that the *entire* market portfolio was mean variance efficient, and were not simply tests of the efficiency of some limited index such as could be formed from the stock market. The essential role played by the market portfolio in the CAPM had been stressed by others; Ross (1977b) had shown the equivalence between the CAPM and the mean variance efficiency of the market portfolio. (Ross (1976a) had also shown that in the absence of arbitrage there was always some efficient portfolio.) Roll went beyond this simple observation, though, by stressing the essential point that the market portfolio is unmeasurable. This called into question the entire cottage industry of testing the CAPM and all of the uses to which the theory had been put, such as performance measurement.

Inter-temporal models

In the aftermath of Roll's critique, attention was turned to alternative models of asset pricing and the intertemporal nature of the theory became more important. Two separate strands of development can be traced. One essentially followed the lines of the CAPM and developed the intertemporal versions of it, the ICAPM. Merton (1973a) pioneered in this. Using continuous time diffusion analysis, Merton showed that the CAPM could be generalized to an intertemporal setting. Most interestingly, though, he demonstrated that, if the economic environment was described by a finite dimensional vector of state variables, x, and if asset prices were exogenously specified random variables, then a version of the SML would hold at all moments of time with the addition to the risk premium of a linear combination of the betas between the assets’ returns and each of the state variables, x_i.

Ross (1975) developed a similar inter-temporal extension of the CAPM, but Ross's model simplified preferences in order to close the model with an inter-temporal rationality constraint and to study equilibrium price dynamics. Along the lines being developed in the modern literature on macroeconomics, inter-temporal rationality and the efficient market theory required that the distribution of prices be determined endogenously. A discrete time Markov model with this feature was presented in Lucas (1978) and a full rational expectations general equilibrium in continuous time was developed in Cox, Ingersoll and Ross (1985a).

Cox, Ingersoll and Ross (1985b) applied their model to analyse and resolve some long-standing questions in the theory of the term structure of interest rates. The theory of the term structure is one of the most important subfields of finance, and the bond markets were one of the first areas where the EMH was applied. In an efficient market, ignoring risk aversion, forward rates should be (unbiased) predictors of future spot rates and many early theories and tests of the EMH were formulated to examine this proposition (see e.g. Malkiel, 1966). Roll (1970) integrated the EMH with the CAPM and used the resulting framework to examine empirically liquidity premia in the bond markets; the work of Cox, Ingersoll and Ross (1985b) can be considered as the logical extension of his analysis to a rational inter-temporal setting.

Merton's model was simplified markedly by Breeden (1979), who showed that, if investors had intertemporally additive utility functions, then Merton's ICAPM and its version of the SML could be collapsed back into a single beta model, the consumption beta model, with all assets being priced, that is, having their risk premiums determined, by their covariance

with aggregate consumption (see also Rubinstein (1976)). If we think of returns as relative prices between wealth today and in future states of nature, then optimizing individuals will set their marginal rates of substitution between consumption today and in future states equal to the rates of return. With continuous asset prices and additive utility functions, indirect utility functions are locally quadratic in consumption and this implies that consumption plays the role of wealth in the static CAPM. This work led to a variety of attempts to measure the ability of betas on aggregate consumption to explain risk premia (see e.g. Hansen and Singleton, 1983).

Arbitrage pricing theory (APT)

A separable but related strand of theory is the arbitrage pricing theory (APT) (see e.g. Ross, 1976a, 1976b). The CAPM and the consumption beta model share the common feature that they explain pricing in terms of endogenous market aggregates, the market portfolio, and aggregate consumption, respectively. The APT takes a different tack.

The intuition of the CAPM (or of the Consumption Beta model) is that idiosyncratic risk can be diversified away, leaving only the systematic risk to be priced. Idiosyncratic risk, though, is defined with reference to the market portfolio as the residual from a regression of returns on the market portfolio's returns. Since no further assumptions are made about the residuals, contrary to intuition a large diversified portfolio that differs from the market portfolio will not in general have insignificant residual risk. The exception is the market portfolio, but then the intuition that diversification leads to pricing by the market portfolio is circular at best.

The APT addresses this issue by assuming directly a return structure in which the systematic and idiosyncratic components of returns are defined a priori. Asset returns are assumed to satisfy a linear factor model,

$$R_i = E_i + \sum_j \beta_{ij} f_j + \varepsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

where E_i, is the expected return, f_j is a demeaned exogenous factor influencing each asset i through its beta on the factor, β_{ij}, and ε_i is an idiosyncratic mean zero term assumed to be sufficiently uncorrelated across assets that it is negligible in large portfolios. An implication of the factor structure is that the ε terms become negligible in large well diversified portfolios and, therefore, such portfolios approximately follow an exact factor structure,

$$R_i \approx E_i + \sum_j \beta_{ij} f_j, \tag{16}$$

where i now denotes the ith well diversified portfolio. In an Arrow–Debreu state space framework, equation (16) can be interpreted as a restriction on the rank of the state-space tableaux.

An exact factor structure implies that there will be arbitrage unless the expected return on each portfolio is equal to a linear combination of the beta coefficients,

$$E_i - r = \sum_j \lambda_j \beta_{ij}, \tag{17}$$

where λ_j is the risk premium associated with the jth factor, f_j. This equation is the APT version of the SML in the CAPM.

The APT is consistent with a wide variety of equilibrium models (including the CAPM if there is a factor structure) and it has been the object of much theoretical and empirical attention. In a sense, the APT can be thought of as a snapshot of any intertemporal model in which the factors represent innovations in the underlying state variables. This means that a rejection of the APT would imply a fairly wide ranging rejection of attempts to model asset markets with a finite set of state variables.

The original theoretical development of the APT (Ross, 1976a, 1976b) showed formally that, if preferences are continuous in the quadratic mean, then the returns on a sequence of portfolios which require no wealth cannot converge to a positive return with a zero variance. This, in turn, implies that the sum of squared deviations from exact APT pricing is bounded above. These results were simplified by Huberman (1982) and extended by Ingersoll (1984) and Chamberlain and Rothschild (1983), all of whom side-stepped the issue of preferences by simply assuming that there could be no sequences converging to an arbitrage situation of a positive return with no variance. By contrast, Dybvig (1983) makes assumptions on preferences and aggregate supply to obtain a tight bound on pricing. His simple order of magnitude calculation is evidence that the pricing error is too small to be of practical significance.

By modelling the capital market explicitly as responding to innovations in exogenous variables, the APT is immediately inter-temporally rational. By contrast with the CAPM and the Consumption Beta models which price assets in terms of their relation with a potentially observable and endogenous market aggregate (wealth for the CAPM and consumption for the Consumption Beta models), the APT factors are exogenous, but unspecified. Much empirical work is now under way to determine a suitable set of factors for representing systematic risk in a factor structure and to examine if they price assets successfully. (For example, see Roll and Ross, 1980; Brown and Weinstein, 1983; and Chen, Roll and Ross, 1986.)

The lack of an a priori specification for the factors has been the focus of criticism of the testability of the APT by Shanken (1982). Shanken argues that, since the factors are not pre-specified, the intuitive derivation of the APT given above can be used to verify APT pricing falsely even when it does not hold, and that to prevent this some equilibrium model, such as that proposed by Connor (1984), must be used. Shanken emphasizes that his critique applies not to the theory of the APT but rather to the way in which it has been tested. Dybvig and Ross (1985) dispute his arguments, stressing that Shanken wants to test the theory including its assumptions and approximations rather than take the positive approach of testing the model's conclusions.

Empirical testing of asset pricing models

Since Roll's critique, the methodology for testing asset pricing models has changed. There has been a retreat from testing a model per se to an explicit view that what is being tested is not the CAPM, for example, but rather whether the particular index being used for pricing is mean variance efficient. This change of focus has led to a more formal approach to the statistics of testing. Ross (1980) developed the maximum likelihood test statistic for the efficiency of a given portfolio and pointed out the analogy between this and the mean variance geometry, and Gibbons (1982) showed that the test of efficiency could be conducted by the use of seemingly unrelated regressions. These results have been extended by others. For example, Kandel (1984) and Jobson and Korkie (1982) and Gibbons, Ross and Shanken (1986) have developed and exploited an exact small sample test of the efficiency of a given index in the presence of a riskless asset. Similar tests of the APT have not yet been developed, and to date much of the testing of the APT has focused on comparisons between the APT and pricing using the value weighted index (see e.g. Chen, Roll, and Ross, 1986).

The most important empirical finding in asset pricing, though, has been the discovery of a wide array of phenomena that appear to be inconsistent with nearly any neoclassical model. Consider, first, the secular effects. Asset returns fall, on average, over the weekend and rise during the week (see French, 1980). Similarly, it has been found that asset returns behave differently in the first half of the month than they do in the second. The most attention, though, has been lavished on the ‘small firm effect’. It appears that the average returns on small firms exceed those on large firms no matter what theory of asset pricing is used to correct for differences in the risk premium between these two categories of assets. Furthermore, the bulk of the return difference is concentrated in the first few days of January. Indeed, on average, returns in January appear to be abnormally large for all stocks (see, for example, Keim, 1983, or Roll, 1981; 1983).

Potentially these sorts of anomalies can be explained by secular changes in risk premia – perhaps due to secular patterns in the release of information – but their persistence and magnitude make them serious challenges to all the asset pricing models. When evidence of this sort appears difficult to explain by any pricing model it calls into question the efficient market hypothesis itself. Tests of an asset pricing model are usually joint tests of both market efficiency and the pricing model; rejecting a wide enough range of such models is tantamount to rejecting efficiency itself.

Substitution and arbitrage: option pricing

The APT is the child of one of the central intuitions of finance, namely, that close substitutes have the same price. This intuition reached fruition in the path breaking paper by Black and Scholes (1973) on option pricing. Since then the theory has found myriad applications and has been significantly extended; see, for example, Merton (1973b), Cox and Ross (1976a; 1976b; 1976c), Rubinstein (1976), Ingersoll (1977), Cox, Ross and Rubinstein (1979), and Cox, Ingersoll and Ross (1985a). The Black-Scholes model employed stochastic calculus, but a simpler framework for option pricing was presented by Cox, Ross and Rubinstein (1979) that retained its essential features and was more flexible for computational

purposes. We will briefly outline this binomial approach and show its connections to the major theoretical features of option pricing.

The binomial model

The binomial model begins with the assumption that the price of a stock, *S*, follows a proportional geometric process:

$$S(t+1) = \begin{cases} aS(t) & \text{with probability } \pi \\ bS(t) & \text{with probability } 1-\pi \end{cases} \tag{18}$$

In addition to the stock there is also a riskless bond with a return of 1+*r*. The basic problem of option pricing theory is to determine the value of a derivative security, i.e., a security whose payoff depends only upon the value of an underlying primitive security, the stock in this case.
Let *C*(*s*, *t*) denote the value of the derivative security as a function of the price of the stock and the time, *t*. Since its value depends only upon the movement of the stock – a result that is sometimes derived as a function of other attributes such as its value at the end of some period – it will also follow a binomial process:

$$C(S, t+1) = \begin{cases} C(aS, t) & \text{with probability } \pi \\ C(bS, t) & \text{with probability } 1-\pi. \end{cases} \tag{19}$$

The time *t*+1 values are illustrated in Figure 2. At any moment of time the information structure branches into relevant states, state *a* and state *b*, defined by whether the stock goes up by *a* or *b*. As the figure is drawn, *a*>1+*r*>*b*, and clearly 1+*r* must lie between *a* and *b* to prevent the stock or the bond dominating. At this point there are two separate approaches to the analysis. The first is in the spirit of the original Black-Scholes model.
Figure 2

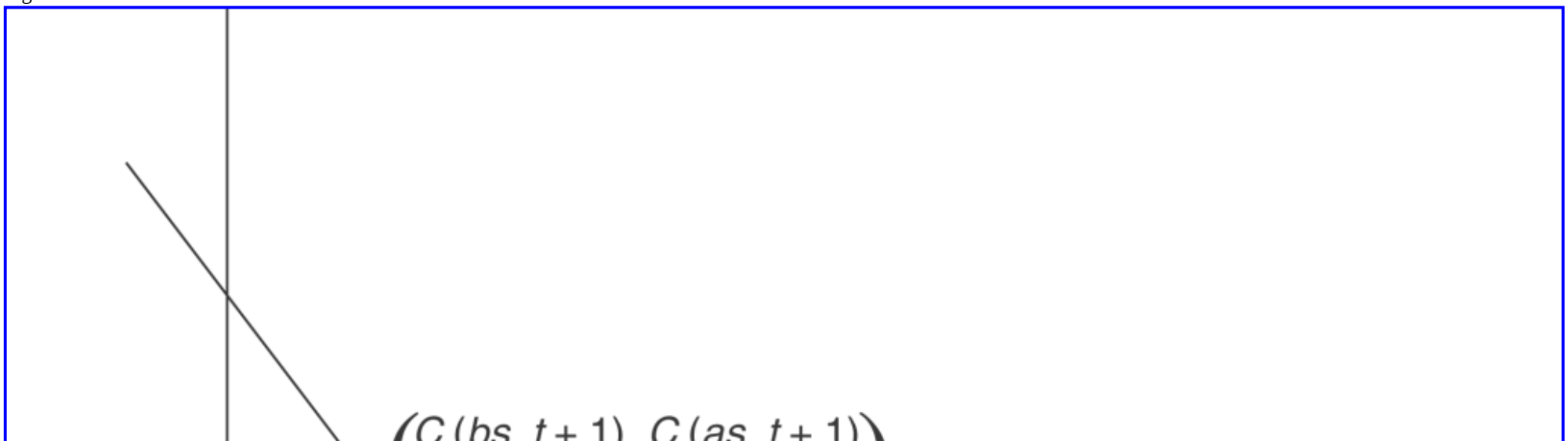

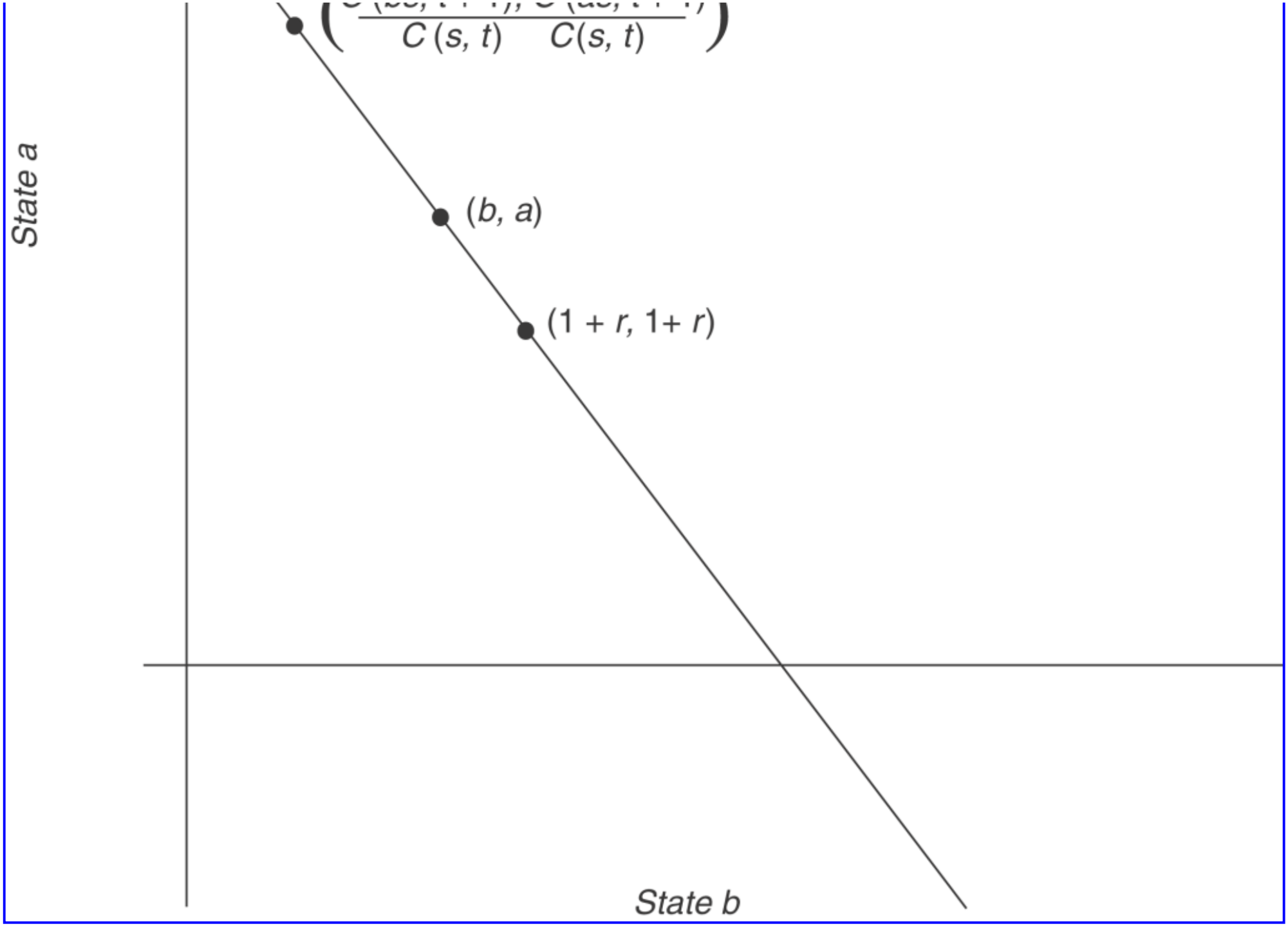

Suppose that at time *t* we form a portfolio of the riskless bond and the stock with α dollars invested in the stock and 1 – α dollars invested in the bond. We will choose the investment proportion so that the return on the portfolio coincides with the return on the derivative security in state *b*. This means choosing α so that

$$\frac{C(bS, t+1)}{C(S, t)} = \alpha b + (1-\alpha)(1+r), \tag{20}$$

which implies that

$$\alpha = \frac{(1+r) - C(bS, t+1) / C(S, t)}{(1+r) - b}. \tag{21}$$

But, since the portfolio's return matches that of the derivative security in state *b*, it must also match it in state *a*. If it did not, then either the portfolio or the derivative security would dominate the other, which would be an arbitrage opportunity. In other words, we must have,

$$\frac{C(bS, t+1)}{C(S, t)} = \alpha a + (1-\alpha)(1+r). \tag{22}$$

Putting these two equations together produces a difference equation which is satisfied by the value of the derivative security,

$$\pi^* C(aS, t+1) + (1-\pi^*) C(bS, t+1) - (1+r) C(S, t) = 0, \tag{23}$$

where

$$\pi^* \equiv \frac{(1+r) - b}{a - b}. \tag{24}$$

Perhaps the most remarkable feature of this equation is that it does not involve the original probabilities for the process, π, but rather is a function of what are called the martingale probabilities, π^*.

To solve this difference equation for the value, C, of a particular derivative security we would need only to append the contractual boundary conditions that define it. For example, a European call option is specified to have the value max ($S - E$, 0), at a specified future date, T, where E is its exercise price. Such an option gives the holder the right – but not the

obligation – to buy the stock for *E* at time *T*. The dual security is a European put option which gives the holder the right, but again not the obligation, to sell the stock for *E*, at time *T*. The problem is more difficult if the derivative security is of the American variety which means that the holder may exercise it any time up to and including the maturity date *T* and need not wait until *T*.

Soon after the Black–Scholes paper, Merton (1973b) examined a variety of option contracts and showed how extensive was the range of the technique. Notably, Merton was able to derive a number of qualitative results on option pricing that were relatively independent of the particular process being modelled. For example, he showed that an American call option on a stock that pays no dividends will never be exercised before its maturity date and, therefore, will have the same value as a similar European call. He also demonstrated that put/call parity, i.e. the equivalence between the positions of holding the stock and a put option and holding a bond and a call option, was not generally valid for American options. Ross (1976c) showed that the literature's emphasis on puts and calls was not misplaced since any derivative security could be composed of puts and calls.

A second approach to the valuation problem in our simple example illuminates why the original probabilities played no role in the analysis. Figure 2 displays what is essentially a two-state Arrow-Debreu model. In such a model if there are two pure contingent claims contracts paying one dollar in each state, then all securities can be valued as a function of their values, q_a and q_b. It follows, then, that any two securities which are not linearly dependent will span the space just as two pure contingent claims would and they can be used to value all securities in the space.

In our example, the value of the bond is 1 and it must satisfy,

$$1 = q_a(1 + r) + q_b(1 + r), \tag{25}$$

and the value of the stock must satisfy,

$$S = q_a(aS) + q_b(bS),$$

or

$$1 = q_a a + q_b b. \tag{26}$$

Solving these two equations we can find the implicit values of the state contingent claims,

$$q_a = \frac{(1 + r) - b}{(1 + r)(a - b)},$$

and

$$q_b = \frac{a - (1 + r)}{(1 + r)(a - b)}. \tag{27}$$

Notice that these prices do not depend on the original probability, π, since they are derived from the values of the stock and the bond. Whatever influence the probability, π, has on values is already reflected in the returns on the stock and the bond, and the derivative security value will just be a function of the implicit state prices. Using these prices, it is readily verified that the difference equation for the value of the derivative security, equation (23), is the same as,

$$q_a C(aS, t + 1) + q_b C(bS, t + 1) = C(S, t). \tag{28}$$

Geometrically, this means that the point,

$$\{[C(bS, t + 1) / C(S, t)], [C(aS, t + 1) / C(S, t)]\},$$

we plot on the same line as the return points for the bond and the stock, (1+*r*, 1+*r*) and (*b*, *a*). For a call option the point will be as drawn in Figure 2 indicating that the call is more volatile than the stock.

Notice from (24) and (27) that

$$\pi^* = (1 + r) q_a,$$

which means that the state space price can be interpreted as the discounted martingale probability. It is this interpretation that ties together the Cox and Ross (1976a) risk-neutral approach to solving option pricing problems and the general theory of the absence of arbitrage.

Cox and Ross (1976a) argued that since the difference equation that emerged for solving option pricing problems made no explicit use of any preference information, the resulting solution must also be independent of preferences. For example, then, the resulting solution must be the same as that which would obtain in a risk neutral world. In such a world, the state probabilities must be such that the expected returns on all assets are the same,

$$\pi^* a + (1 - \pi^*) b = 1 + r,$$

where the solution for the probability, π^*, is the same martingale probability defined above. For a European call option, then, the solution will be

$$C(S,t) = \frac{1}{(1+r)^{T-t}} E^*[\max(S_T - E, 0)] = \frac{1}{(1+r)^{T-t}} \sum_{j \geq \ln[E/Sb^{-(T-t)}]/\ln(a/b)} (\pi^*)^j \times (1-\pi^*)^{T-t-j} (Sa^j b^{T-t-j} - E),$$
(29)

where E^* is the expectation with respect to the martingale probabilities, π^* and $(1-\pi^*)$. It is easily verified that (29) is the solution to the difference equation (23) subject to the boundary condition,

$$C(S,T) = \max(S - E, 0).$$

Contrast this formula with the original Black-Scholes formula for the value of a call option in a continuous time diffusion model,

$$C(S,t) = SN(d_1) - e^{-r(T-t)} N(d_2),$$
(30)

where $N(\cdot)$ is the standard cumulative normal distribution function and,

$$d_1 \equiv \frac{\ln(S/E) + r(T-t) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{(T-t)}},$$

and

$$d_2 \equiv d_1 - \sigma\sqrt{(T-t)}.$$

Equation (30) is the solution to the Black–Scholes option pricing differential equation,

$$\frac{1}{2}\sigma^2 S^2 C_{SS} + rSC_S - rC = -C_t,$$
(31)

subject to the boundary condition,

$$C(S, T) = \max(S - E, 0).$$

The Black–Scholes differential equation (31) is derived from an analogous hedging argument to that for the binomial model, applied to the continuous log-normal stock process,

$$dS / S = \mu dt + \sigma dz$$

where z is a standard Brownian motion. In fact, as the time interval between jumps converges to zero and the jump sizes shrink appropriately, the binomial converges to the lognormal diffusion and its option pricing solution will converge to that for the lognormal diffusion. Notice, too, that in analogy with the binomial whose solution does not depend upon the state probabilities, the Black-Scholes option price (30) is independent of the expected return on the stock, μ.

The most interesting comparative statics result from these models is the observation that call or put option values increase with increasing variance, σ^2. This is a consequence of these options being convex functions of the terminal stock value, S_T (Cox and Ross, 1976b).

The general theory of arbitrage

All of the above analysis can be tied together by the general theory of arbitrage. Under quite general conditions, it can be shown that the absence of arbitrage implies the existence of a linear pricing rule that values all of the assets (see e.g., Ross, 1976a, 1978b; Harrison and Kreps, 1979). In a static model with m states of nature, this means the existence of implicit state prices, q_j, such that $q_j>0$, and such that any asset with payoffs of x_j in the states of nature will have the value,

$$p = \sum_j q_j x_j. \tag{32}$$

The intertemporal extension of this result is most neatly displayed in terms of the martingale expectation used above. The absence of arbitrage now implies the existence of a martingale measure such that, with obvious notation,

$$p = E^*\left\{\exp\left[-\int_0^T r(s)ds\right]x_T\right\}.$$

This theory permits us to tie together not only the basic results of option pricing, but also our previous analysis of asset pricing models. For example, applying it to the exact factor model,

$$R_i = E_i + \sum_j \beta_{ij} f_j,$$

yields the APT,

$$1 = E^*(R_i) = E^*\left(E_i + \sum_j \beta_{ij} f_j\right) = \frac{1}{(1+r)}\left[E_i + \sum_j \beta_{ij} E^*(f_j)\right].$$

or

$$E_i = (1+r) + \sum_j \lambda_j \beta_{ij},$$

where

$$\lambda_j \equiv -E^*(f_j).$$

Similarly, in a mean variance framework the martingale analysis can be used to prove that there is always a portfolio whose covariances are proportional to the excess returns on each asset. In other words, the absence of arbitrage implies the existence of a mean variance efficient portfolio (see Ross, 1976a; Chamberlain and Rothschild, 1983).

Empirical testing

Perhaps because the option pricing theory works so well, it has generated a surprisingly small empirical literature. Some early tests, for example, Black and Scholes (1973) and Galai (1977), focused on whether the models could be used to generate successful trading rules and found that any success was easily lost to transactions costs. Most interestingly, MacBeth and Merville (1979) found that the option formulas tended to underprice ‘in the money’ options and overprice ‘out of the money’ options, but Geske and Roll (1984) have argued that this effect disappears with a reformulation of the statistics.

Given a theory that works so well, the best empirical work will be to use it as a tool rather than to test it. Chiras and Manaster (1978), for example, show that implicit volatilities, i.e. variances computed by inverting the option formulas to obtain variance as a function of the quoted option price, have strong predictive power for explaining future realized stock variances. Patell and Wolfson (1979) use the implicit variances to examine whether stock prices are more volatile around earnings announcements.

These efforts should increase; options and option pricing theory give us an opportunity to measure directly the degree of anticipated uncertainty in the markets. Financial press terms such as ‘investor confidence’ take on new meaning when they can actually be measured.

This does not mean, however, that there are no important gaps in the theory. Perhaps of most importance, beyond numerical results (see, for example, Parkinson, 1977; or Brennan and Schwartz, 1977), very little is known about most American options which expire in finite time. The American call option on a stock paying a dividend or the American put option are both easily solved in the infinite maturity case since the optimal exercise boundary is a fixed stock value independent of time (Merton, 1973b; Cox and Ross, 1976a, 1976b). If

dividends occur at discrete points, then if the call is exercised prematurely it will only be optimal to do so just prior to a dividend payment. This permits a recursive approach to the solution of this finite maturity option (see Roll, 1977a; Geske, 1979). But, with continuous payouts, surprisingly little is known about the exercise properties of either of these options in the American case.

Despite such gaps, when judged by its ability to explain the empirical data, option pricing theory is the most successful theory not only in finance, but in all of economics. It is now widely employed by the financial industry and its impact on economics has been far ranging. At a theoretical level, we now understand that option pricing theory is a manifestation of the force of arbitrage and that this is the same force that underlies much of neoclassical finance.

The whole is the sum of the parts – corporate finance

The use of arbitrage as a serious tool of analysis coincided with the beginning of the modern theory of corporate finance. In two seminal papers on the cost of capital, Modigliani and Miller (1958, 1963) argued that the overall cost of capital and, therefore, the value of the firm would be unaffected by its financing decision. Specifically, using arbitrage arguments, Modigliani and Miller showed that the debt/equity split would not alter a firm's value and they then argued that with the investment decision held constant, the dividend payout rate of the firm would also not affect that value. These two irrelevance propositions defined the study of corporate finance in much the same way that Arrow's Impossibility Theorem defined social choice theory. At one and the same time they propounded an irreverent theory whose central feature was the irrelevance of the topic under study. This challenge, to weaken in a useful way the assumptions of their analysis, has guided research in this area ever since.

The Modigliani – Miller analysis

Since the Modigliani-Miller (henceforth MM) irrelevance propositions are developed from the absence of arbitrage, they are quite robust to alternative specifications of the economic model. To derive the Modigliani-Miller propositions we will employ the no arbitrage theory above. Consider a firm which will liquidate all of its assets at the end of the current period, and let x denote the random liquidation value of the assets. Assume that the firm has debt outstanding with a face value of F and that the remainder of the value of the firm is owned by the stockholders who have the residual claim after the bondholders.

At the end of the period, if x is large enough the stockholders will receive $x - F$ and if x falls short of F they will receive nothing. Formally, then, the terminal payment to the stockholders is

$$\max(x - F, 0),$$

which will be recognized as the terminal payment on a call option. In other words – in a tribute to the ubiquitous nature of option pricing theory – the stockholders have a call option on the terminal value of the firm, x, with an exercise price equal to the face value of the debt, F. The bondholders can claim the entire assets if x is not sufficient to cover the promised payment of F, which means that they will receive,

$$\min(x, F)$$

The current value of the firm, V, is defined to be the value of all of the outstanding claims against its assets which in this case is the value of the stocks, S, and the bonds, B. Using the no arbitrage analysis, we find that (ignoring discounting),

$$V \equiv S + B = E^*[\max(x - F, 0)] + E^*[\min(x, F)] = E^*[\max(x - F, 0) + \min(x, F)] = E^*(x),$$

which is independent of the face value, F, of the debt and, therefore, independent of the relative amounts of debt and equity. This verifies the first of the MM irrelevance propositions. To verify the irrelevance of value to the dividend payout, consider a firm about to pay a dividend, D. The current, pre-dividend, value of the stock is $p^{-}(D)$ and by the no arbitrage martingale analysis this is given by,

$$p^{-}(D) = E^{*}[D + p^{+}(D)] = D + E^{*}[p^{+}(D)],$$

where $p^{+}(D)$ is the ex-dividend price. If the investment policy of the firm has been fixed, then the only impact that the current dividend payout can have on the stockholders is through its alteration of the cash in the firm. This means that changing the dividend to, say $D+\Delta D$, would necessitate a change in current assets of $-\Delta D$. From the first MM proposition the mode of financing this change in the dividend will be irrelevant to the determination of the firm's value and to simplify the analysis we will assume that it is financed by riskless debt. At an interest rate of r this would entail, say, a perpetual outflow from the firm of $r\Delta D$. Again applying the analysis and letting x_{t+s} be the cash flow at time $t+s$ given that a dividend of D is paid now, we have,

$$p^{+}(D+\Delta D) = E^{*}\left[\int_0^{\infty} e^{-rs}(x_{t+s} - r\Delta D)\mathrm{d}s\right] = E^{*}\left(\int_0^{\infty} e^{-rs}x_{t+s}\mathrm{d}s\right) - E^{*}\left(\int_0^{\infty} e^{-rs}r\Delta D\mathrm{d}s\right) = p^{+}(D) - \Delta D.$$

Thus, we have the irrelevance proposition,

$$p^{-}(D+\Delta D) = E^{*}[(D+\Delta D) + p^{+}(D+\Delta D)] = D + \Delta D + E^{*}[p^{+}(D+\Delta D)] = D + \Delta D + E^{*}[p^{+}(D)] - \Delta D = D + E^{*}[p^{+}(D)] = p^{-}(D).$$

The MM results were startling to those who had worked in corporate finance and had taken it for granted that the way in which a firm was financed affected its value. To understand the importance of the MM results for the most practical of problems; recall that the original impetus for the study of corporate finance was the determination of the firm's opportunity cost for investments, ρ. For a marginal investment, financed by the issuance of debt and equity, the cost of capital, ρ, also known as the weighted average cost of capital, WACC, would be the weighted average cost of the debt, r, and the cost of equity, k,

$$\rho = (S/V)k + (B/V)r, \tag{33}$$

(where we have ignored tax effects).

If debt is riskless, then r is the interest rate on such debt and k, the cost of equity, will be the return required by investors for the risk inherent in the stock. Presumably k could be found by appeal to one of the asset pricing models discussed above.

Now it is tempting to think, for example, that if $k>r$, then an increase in debt relative to equity will lower ρ. If this goes too far, debt will become risky and as r rises there will be a unique optimal debt/equity ratio, (B/S)0, that minimizes the cost of capital, ρ. This would be the discount rate to use for present value calculations and it would maximize the value of the firm. This was the traditional analysis of the leverage decision before MM.

By the MM theorem, though, value, V, is unaffected by leverage. This means that ρ is unaltered, since the total (expected) return to the stockholders and the bondholders, $Sk+Br$, is

unaltered [see Equation (33)]. In terms of the WAAC, then, as the leverage (B/S) is increased by the substitution of debt for equity, the cost of equity changes.

$$k = \rho + (B/S)(\rho - r),$$

but not the WAAC.

Spanning arguments

The efforts to elude these results and to develop a meaningful theory of corporate finance have taken many forms. First, it has been argued that the analysis itself contains a hidden and critical assumption, namely that the pricing operator is independent of the corporate financial structure. The alternative is that the change in the debt/equity decision, for example, will also change the span of the marketed assets in the economy and, consequently, the operator used for pricing will change. The simplest such example would be a single firm in a two-state world. If the firm is an all equity firm and if there are no other traded assets, then individuals cannot adjust their consumption across the states of nature and must split it according to the equity payoff. If this firm now issues debt the two securities will span the two states of nature and complete the market. This, in turn, will generally alter pricing in the economy.

While this argument has generated a large literature, the problem of the determination of the corporate financial structure and the value of the firm is primarily a microeconomic question and it is difficult to believe that it will be resolved or even illuminated by assuming that firms have some monopoly power that enables them to alter pricing in the capital markets. At the microlevel the MM propositions are unlikely to be seriously affected by such general equilibrium arguments.

At the microlevel, too, the intuition behind the MM propositions and its conclusions is so robust as to be daunting. Consider the following argument. According to MM there can be no optimal, i.e. value maximizing, financial structure since value is independent of structure. Suppose that there was an optimal, say, debt/equity ratio, $(B/S)^0$. Any departure from this target $(B/S)^0$, however, could not lower the firm's value since it would immediately afford an arbitrage opportunity to buy the total firm at its lowered value and refinance it in the optimal target propositions, $(B/S)^0$. (This somewhat facetious argument gets the point across, but it really means that we have not fully specified the rules of the game, e.g., who moves first, what happens when no one moves, etc.)

Signalling models

A more promising route which formally exploits incomplete spanning, but does not argue that the pricing operator itself is altered by any one firm changing its financial structure, makes use of the theory of asymmetric information and signalling (see Ross, 1977a; Leland and Pyle, 1977; and Bhattacharya, 1979). If the managers of the firm possess information that is not held by the market then the market will make inferences from the actions of the firm and in particular, from financial decisions. Changes in its financial structure or its dividend policy will alter investors' perceptions of its risk class and, therefore, its value. While the operator, $E^*(\cdot)$, does not change, the perception of the distribution of the firm's cash flows does. In an effort to maximize their value, firms will take actions, such as taking on high debt to equity ratios, which can be imitated by lesser firms only at a prohibitive cost. This will distinguish them from lesser firms that the uninformed market erroneously puts into the same classes with them. In this fashion, a hierarchy of firm risk classes will emerge, and, in equilibrium, firms will signal their true situations and investors will draw correct inferences from their signals.

All of this has a nice ring to it, but the nagging question that remains is why firms use their financial decisions to accomplish all of this information transfer. Financial changes are cheap, but even cheaper might be guarantees or, for that matter, a system of legislation. These issues remain unresolved, but it is difficult to think that much will be explained by theories that argue that firms take on more debt just to show the world that 'they can do it'. There is a limit to macho-finance.

Taxes

Another line of attack has been to introduce more 'imperfections', especially taxes, into the models. Modigliani and Miller originally had noted that the presence of a corporate tax meant that firms would have an incentive to issue additional debt. Since interest payments on debt are excluded from corporate taxes, substituting debt for equity permits firms to pass returns to investors with a lowered tax cut to the government. At the limit, firms would be all debt if the tax authorities still recognized such debt payments as excludable from taxable corporate income. Presumably, the only brake to this expansion would be the real costs of dealing with the inevitable bankruptcies of high debt firms. This is logically possible, but at the expense of reducing corporate finance to the study of the tradeoff between the tax advantages of debt and the costs of bankruptcy.

Miller (1977) found a more profound brake to this tendency to increase debt. He argued that while the firm could lower its taxes by increasing its debt, the ability of investors to defer or offset capital gains implies that they pay higher taxes on interest income than on the returns from equities. With a rising tax schedule, an equilibrium is possible in which the marginal investor has a tax differential between ordinary income and equity returns that exactly offsets the firm's corporate tax advantage to debt. In such an equilibrium, investors in a higher tax bracket than the marginal investor would purchase only equity (or non-taxed bonds such as municipals for US investors) and those in lower tax brackets would purchase corporate bonds. There would be an equilibrium amount of debt for the corporate sector as a whole, but not for any individual firm (assuming the absence of inframarginal firm tax schedules).

Miller's analysis led to a large literature on the impact of taxes on pricing. Black and Scholes (1974) had made a related argument for the absence of a tax effect on dividends, arguing that stocks with relatively higher yields should not have higher gross returns to compensate investors for the additional tax burden since companies would then just cut their dividends to increase the stock price. Black and Scholes verified their results empirically, but, using a different methodology, Litzenberger and Ramaswamy (1982) found that gross returns were higher for stocks with higher dividends. Whether the supply side or the demand side dominates remains undecided.

Whatever the resolution of this and similar debates, the equilibrium tax argument initiated by Miller has changed much of the analysis of these issues. Miller and Scholes (1978), for example, argue that by employing a number of ‘laundering’ devices individuals can dramatically cut their taxes. Their conclusion that, in theory, taxes should be much lower than they appear to be in practice, focuses attention on the role played by informational asymmetries and the related costliness of using techniques such as investing through tax exempt intermediaries.

Agency models

The emphasis on informational asymmetries has been the cornerstone of an alternative approach to corporate finance, agency theory. Wilson (1968) and Ross (1973) developed agency models in which one party, the agent (e.g. a corporate manager) acts on behalf of another the principal (e.g. stockholders). Jensen and Meckling (1976), building on the agency theory and on Williamson's (1975) transaction cost approach, argue that corporate finance can be understood in terms of the monitoring and bonding costs imposed on stockholders and managers by such relations. The manager qua employee has an incentive to divert firm resources to his own benefit. Jensen and Meckling refer to the loss in value in restraining this incentive as the (equilibrium) agency cost of the relation.

To some extent this conflict can be resolved *ex ante* by the indenture agreements and covenants in financial contracts, but the cost of doing so rises with the monitoring requirements. Myers (1977), for example, has studied the implications for investment policy of the conflict between the stockholders and the bondholders. Stockholders own a call option on the assets of the firm and the value of a call increases with the variance of the asset value. Conversely, such increases will come at the expense of the bondholders. *Ex ante* indenture agreements can limit the ability of management and stockholders to take on additional risk, but the more precise the limits the costlier it is to write, observe and enforce them.

These trade-offs are the intuition and subject matter of the agency approach to corporate finance, but to date it is more a collection of intuitions than a well-articulated theory. The agency approach has pointed in some intriguing directions, but it fares poorly if judged by asking what it is that would be a counter observation or count as evidence against it. To the contrary, no phenomenon seems beyond the reach of ‘agency costs’ and at times the phrase takes on more of the trappings of an incantation than an analytical tool. The role of asymmetric information in corporate finance and in explaining the managerial and financial forces at work in the firm is self evident, but it remains fertile ground for theory.

Empirical evidence

The early empirical work examined the relation between the corporate financial structure and other characteristics of the firm. Hamada (1972), for example, studied whether the beta of a firm's equity was related to the beta of the firm's assets as would be predicted by the cost of capital, equation (33). There continues to be empirical work on these issues, but the attention of empiricists has shifted to the arena of corporate control.

A boom in merger and acquisition activity in the late 1970s and through to the present time has brought some striking and unexplained empirical regularities. On average, shareholders in firms that are the targets of tender offers gain significantly from such offers while the rewards to bidders are still ambiguous (Jensen and Ruback, 1983). For unsuccessful tenders the target firms appear to average an eventual loss and the bidders may, too. These results and the discrepancy between targets and bidders have been the object of close scrutiny.

If firms realize such abnormal gains as targets, and if it reflects the release of information about the value of their underlying assets, then that raises the question of why they were not priced correctly to begin with. On the other hand, if the returns for successful targets reflect synergies rather than simply a revaluation of their assets, why does the bidder get so little? Several game theoretic and bidding models have been built in an attempt to explain these results (see, e.g., Grossman and Hart, 1980), but a consensus has yet to emerge. Furthermore, some of the important empirical issues, such as whether bidders actually gain or lose on average remain unresolved.

Conclusion

For corporate finance, like the other major areas of finance, the neoclassical theory is now well established, but, like the other areas, the inadequacy of the neoclassical analysis is pushing researchers to begin the challenging but promising exploration of theories of asymmetric information. This work holds out the hope of explaining some of the deeper mysteries of finance that have eluded the neoclassical theory, from the embarrassing plethora of anomalies in capital markets to the basic questions of financial structure. Perhaps the feature that truly distinguishes finance from much of the rest of economics is this constant interplay between theory and empirical analysis. The test of these new approaches will be decided less by reference to their aesthetics and more by their usefulness in explaining financial data. At the height of the subject, these two criteria become one.

See Also

- arbitrage
- capital asset pricing model
- dividend policy
- efficient markets hypothesis
- finance (new developments)
- options

Bibliography

Admati, A. 1985. A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica* 53(3), May, 629–57.

Bhattacharya, S. 1979. Imperfect information, dividend policy and the ‘Bird in the Hand’ fallacy. *Bell Journal of Economics and Management Science* 10(1), Spring, 259–70.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45(3), July, 444–55.

Black, F., Jensen, M. and Scholes, M. 1972. The capital asset pricing model: some empirical tests. In *Studies in the Theory of Capital Markets*, ed. M.C. Jensen. New York: Praeger.

Black, F. and Scholes, M. 1972. The valuation of option contracts and a test of market efficiency. *Journal of Finance* 27(2), May, 399–417.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3), May-June, 637–54.

Black, F. and Scholes, M. 1974. The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics* 1(1), May, 1–22.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7(3), September, 265–96.

Brennan, M.J. and Schwartz, E.S. 1977. The valuation of American put options. *Journal of Finance* 32(2), May, 449–62.

Brown, S. and Weinstein, M. 1983. A new approach to testing asset pricing models: the bilinear paradigm. *Journal of Finance* 38(3), June, 711–43.

Cass, D. and Stiglitz, J. 1970. The structure of investor preferences and asset returns, and separatability in portfolio selection: a contribution to the pure theory of mutual funds. *Journal of Economic Theory* 2(2), June, 122–60.

Chamberlain, G. 1983. Funds, factors and diversification in arbitrage pricing models. *Econometrica* 51(5), September, 1305–23.

Chamberlain, G. and Rothschild, M. 1983. Arbitrage, factor structure and mean-variance analysis on large asset markets. *Econometrica* 51(5), September, 1281–304.

Chen, N., Roll, R. and Ross, S.A. 1986. Economic forces and the Stock market. *Journal of Business* 59, July, 383–403.

Chiras, D.P. and Manaster, S. 1978. The information content of option prices and a test of market efficiency. *Journal of Financial Economics* 6(2–3), June–September, 213–34.

Connor, G. 1984. A unified beta pricing theory. *Journal of Economic Theory* 34(1), October, 13–31.

Cootner, P., ed. 1964. *The Random Character of Stock Market Prices*. Cambridge, Mass: MIT Press.

Cowles, A. 1933. Can stock market forecasters forecast? *Econometrica* 1, (July), 309–24.

Cox, J.C., Ingersoll, J. and Ross, S.A. 1985a. An intertemporal general equilibrium model of asset prices. *Econometrica* 53(2), March, 363–84.

Cox, J.C., Ingersoll, J. and Ross, S.A. 1985b. A theory of the term structure of interest rates. *Econometrica* 53(2), March, 385–407.

Cox, J.C. and Ross, S.A. 1976a. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3(1–2), January–March, 145–66.

Cox, J.C. and Ross, S.A. 1976b. A survey of some new results in financial option pricing theory. *Journal of Finance* 31(2), May, 383–402.

Cox, J.C., Ross, S.A. and Rubinstein, M. 1979. Option pricing: a simplified approach. *Journal of Financial Economics* 7(3), September, 229–63.

Diamond, D. and Verrecchia, R. 1981. Information aggregation in a noisy rational expectations economy. *Journal of Financial Economics* 9(3), September, 221–35.

Dybvig, P. 1983. An explicit bound on deviations from APT pricing in a finite economy. *Journal of Financial Economics* 12(4), December, 483–96.

Dybvig, P. and Ross, S.A. 1985. Yes, the APT is testable. *Journal of Finance* 40(4), September, 1173–88.

Fama, E.F. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25(2), May, 383–417.

Fama, E.F., Fisher, L., Jensen, M. and Roll, R. 1969. The adjustment of stock prices to new information. *International Economic Review* 10(1), February, 1–21.

Fama, E.F. and MacBeth, J. 1973. Risk, return and equilibrium: empirical tests. *Journal of Political Economy* 81(3), May-June, 607–36.

Flavin, M. 1983. Excess volatility in the financial markets: a reassessment of the empirical evidence. *Journal of Political Economy* 91(6), December, 929–56.

French, K. 1980. Stock returns and the weekend effect. *Journal of Financial Economics* 8(1), March, 55–69.

French, K. and Roll, R. 1985. Stock return variances, the arrival of information and the reaction of traders. UCLA Working Paper.

Galai, D. 1977. Tests of market efficiency of the Chicago Board Options Exchange. *Journal of Business* 50(2), April, 167–97.

Geske, R. 1979. A note on an analytical valuation formula for unprotected American call options on stocks with known dividends. *Journal of Financial Economics* 7(4), December, 375–80.

Geske, R. and Roll, R. 1984. Isolating the observed biases in call option pricing: an alternative variance estimator. UCLA Working Paper, April.

Gibbons, M. 1982. Multivariate tests of financial models: a new approach. *Journal of Financial Economics* 10(1), March, 3–27.

Gibbons, M., Ross, S.A. and Shanken, J. 1986. A test of the efficiency of a given portfolio. Stanford University Working Paper No. 853.

Granger, C.W.J. and Morgenstern, O. 1962. Spectral analysis of New York stock market prices. Econometric Research Program, Princeton University, Research Memorandum, September.

Grossman, S.J. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 31(2), May, 573–85.

Grossman, S.J. and Hart, O. 1980. Takeover bids, the free-rider problem, and the theory of the corporation. *Bell Journal of Economics and Management Science* 11(1), Spring, 42–64.

Grossman, S.J. and Stiglitz, J. 1980. The impossibility of informationally efficient markets. *American Economic Review* 70, June, 393–408.

Hamada, R.S. 1972. The effect of the firm's capital structure on the systematic risk of common stocks. *Journal of Finance* 27(2), May, 435–52.

Hansen, L. and Singleton, R. 1983. Stochastic consumption, risk aversion and the temporal behavior of asset returns. *Journal of Political Economy* 91(2), April, 249–65.

Harrison, J.M. and Kreps, D.M. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20(3), June, 381–408.

Hicks, J.R. 1946. *Value and Capital*. 2nd edn, London: Oxford University Press.

Huberman, G. 1982. A simple approach to arbitrage pricing theory. *Journal of Economic Theory* 28(1), October, 183–91.

Ingersoll, J. 1977. A contingent-claims valuation of convertible securities. *Journal of Financial Economics* 4(3), May, 289–322.

Ingersoll, J. 1984. Some results in the theory of arbitrage pricing. *Journal of Finance* 39(4), September, 1021–39.

Jaffé, J. 1974. The effect of regulation changes on insider trading. *Bell Journal of Economics and Management Science* 5(1), Spring, 93–121.

Jensen, M.C. and Meckling, W.H. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3(4), October, 305–60.

Jensen, M.C. and Ruback, R.S. 1983. The market for corporate control: the scientific evidence. *Journal of Financial Economics* 11(1–4), April, 5–50.

Jobson, J.D. and Korkie, B. 1982. Potential performance and tests of portfolio efficiency. *Journal of Financial Economics* 10(4), December, 433–66.

Kandel, S. 1984. On the exclusion of assets from tests of the mean variance efficiency of the market portfolio. *Journal of Finance* 39(1), March, 63–73.

Keim, D. 1983. Size-related anomalies and stock return seasonality: further empirical evidence. *Journal of Financial Economics* 12(1), June, 13–32.

Kleidon, A. 1986. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94(5), October, 953–1001.

Leland, H. and Pyle, D. 1977. Informational asymmetries, financial structure and financial intermediation. *Journal of Finance* 32(2), May, 371–87.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, February, 13–37.

Litzenberger, R. and Ramaswamy, K. 1982. The effect of dividends on common stock prices: tax effects or information effects? *Journal of Finance* 37(2), May, 429–43.

Lucas, R.E., Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46(6), November, 1429–45.

MacBeth, J. and Merville, L. 1979. An empirical examination of the Black-Scholes call option pricing model. *Journal of Finance* 34(5), December, 1173–86.

Malkiel, B. 1966. *The Term Structure of Interest Rates: Expectations and Behavior Patterns*. Princeton: Princeton University Press.

Markowitz, H.M. 1959. *Portfolio Selection: Efficient Diversification of Investments*. New York: Wiley.

Marsh, T. and Merton, R.C. 1986. Dividend variability and variance bounds tests for the rationality of stock market prices. *American Economic Review* 76(3), June, 483–98.

Merton, R.C. 1973a. An intertemporal capital asset pricing model. *Econometrica* 41(5), September, 867–87.

Merton, R.C. 1973b. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4(1), Spring, 141–83.

Milgrom, P. and Stokey, N. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26(1), February, 17–27.

Miller, M.H. 1977. Debt and taxes. *Journal of Finance* 32(2), May, 261–75.

Miller, M.H. and Scholes, M. 1978. Dividends and taxes. *Journal of Financial Economics* 6(4), December, 333–64.

Modigliani, F. and Miller, M.H. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48, June, 261–97.

Modigliani, F. and Miller, M.H. 1963. Corporate income taxes and the cost of capital. *American Economic Review* 53, June, 433–43.

Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34(4), October, 768–83.

Myers, S. 1977. Determinants of corporate borrowing. *Journal of Financial Economics* 5(2), November, 147–75.

Parkinson, M. 1977. Option pricing: the American put. *Journal of Business* 50(1), January, 21–36.

Patell, J.M. and Wolfson, M.A. 1979. Anticipated information releases reflected in call option prices. *Journal of Accounting and Economics* 1(2), August, 117–40.

Roll, R. 1970. *The Behavior of Interest Rates: An Application of the Efficient Market Model to US Treasury Bills*. New York: Basic Books.

Roll, R. 1977a. An analytic valuation formula for unprotected American call options on stocks with known dividends. *Journal of Financial Economics* 5(2), November, 251–8.

Roll, R. 1977b. A critique of the asset pricing theory's tests. *Journal of Financial Economics* 4(2), March, 129–76.

Roll, R. 1978. Ambiguity when performance is measured by the securities market line. *Journal of Finance* 33(4), September, 1051–69.

Roll, R. 1981. A possible explanation of the small firm effect. *Journal of Finance* 36(4), September, 879–88.

Roll, R. 1983. The turn-of-the-year effect and small firm premium. *Journal of Portfolio Management* 9(2), Winter, 18–28.

Roll, R. 1984. Orange juice and weather. *American Economic Review* 74(5), December, 861–80.

Roll, R. and Ross, S.A. 1980. An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35(5), December, 1073–103.

Ross, S.A. 1973. The economic theory of agency: the principal's problem. *American Economic Review* 63(2), May, 134–9.

Ross, S.A. 1975. Uncertainty and the heterogeneous capital good model. *Review of Economic Studies* 42(1), January, 133–46.

Ross, S.A. 1976a. Return, risk and arbitrage. In *Risk and Return in Finance*, ed. I. Friend and J. Bicksler. Cambridge, Mass: Ballinger.

Ross, S.A. 1976b. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13(3), December, 341–60.

Ross, S.A. 1976c. Options and efficiency. *Quarterly Journal of Economics* 90(1), February, 75–89.

Ross, S.A. 1977a. The determination of financial structure: the incentive signalling approach. *Bell Journal of Economics and Management Science* 8(1), Spring, 23–40.

Ross, S.A. 1977b. The capital asset pricing model (CAPM), short-sale restrictions and related issues. *Journal of Finance* 32(1), March, 177–83.

Ross, S.A. 1978a. Mutual fund separation in financial theory – the separating distributions. *Journal of Economic Theory* 17(2), April, 254–86.

Ross, S.A. 1978b. A simple approach to the valuation of risky streams. *Journal of Business* 51(3), July, 453–75.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics and Management Science* 7(2), Autumn, 407–25.

Shanken, J. 1982. The arbitrage pricing theory: is it testable? *Journal of Finance* 37(5), December, 1129–40.

Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, September, 425–42.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71(3), June, 421–36.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, February, 65–86.

Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

Wilson, R.B. 1968. The theory of syndicates. *Econometrica* 36(1), January, 119–32.

How to cite this article

Ross, Stephen A. "finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000071>
doi:10.1057/9780230226203.0566

The New Palgrave Dictionary of Economics Online

financial intermediation

J.H. Boyd

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article deals with the process of financial intermediation: that is, savings and investment flows that are intermediated through organizations such as banks and insurance companies. There are five major topics: stylized facts about financial intermediary organizations and markets; the history of thought about financial intermediation; the theory of financial intermediaries, with an aside on equilibrium credit rationing; the regulation of financial intermediation; and trends in recent research and open research questions.

Keywords

asset transformation; bank runs; banking crises; banks; corporations; default risk; deflation; delegated monitoring; deposit insurance; discount window; equilibrium credit rationing; *ex post* monitoring; financial intermediaries; financial intermediation; financial markets; fractional reserve banking; Friedman rule; Great Depression; information economics; interest rate risk; liquidity; monetary policy; payment services; private information; safety net (financial); savings and loan industry

Article

1 Preliminaries and introduction

Writing an article such as this requires making some tough decisions about what to include and what not. Many deserving topics in financial intermediation have not been mentioned at all and I cannot begin to cite all the good papers that deserve reference. Primarily, I rely on two excellent survey articles, one that focuses on theory (Gorton and Winton, 2003), and another that focuses on empirics (Levine, 2005).
I received helpful comments from Doug Diamond, Jack Kareken, Ross Levine and Ed Prescott; however, they are totally absolved from any errors that remain.
It is the convention to distinguish between ‘financial markets’ and ‘financial intermediaries’. A financial

market is a market in which investors acquire direct claims against ultimate borrowers, usually in the form of debt or equity. A financial intermediary (FI) is a firm that substitutes its own liability for that of some ultimate borrower. That is, an investor lends to the FI and, in turn, the FI lends to an ultimate borrower. I adopt this standard convention even though the distinction is often imprecise. (For example, debt and equity claims are rarely traded directly between the ultimate claimants. Even these are ‘intermediated’.) Next, let us turn to some facts about FIs.

The assets of FIs are almost exclusively financial claims. FIs do not have many physical assets, except buildings and computers, and they produce no physical products; thus they are service firms. Important and easily recognizable examples of FIs would include commercial banks, savings and loan associations, credit unions, life insurance firms, property and casualty insurers, consumer finance companies, and mortgage bankers.

Banks largest

Commercial banks (hereafter *banks*) are the most important class of FIs, and this has been true for centuries. In developing economies, banks often play a dominant role and may be, essentially, ‘the only game in town’. Even in the United States, with its highly developed financial markets, banks accounted for about 14.2 per cent of financial intermediary assets, which is the largest private share, followed by mutual funds at 12.4 per cent (Board of Governors of the Federal Reserve System, 2005). This size factor helps explain why banks have been the most-studied class of FI by a wide margin. Banks are also especially important and heavily studied because they create money and thus are the conduit for monetary policy. This article follows the norm and devotes a disproportionate amount of its attention to banks.

Heavily regulated

FIs are heavily regulated relative to non-financial firms. Most of this regulation is advertised to promote ‘safety and soundness’, meaning that its stated intent is to reduce the frequency of failures and other problems in the industry. There are four basic forms of regulation: minimum capital requirements, examination by regulatory authorities, portfolio restrictions on asset holdings, and restrictions on who can own or manage an FI. In many countries, there has been a trend towards less intensive regulation of FIs since the mid-1990s, but in these four forms regulation remains obtrusive relative to most industries.

A large industry

The FI industry is relatively large. Especially in developed economies, the FI sector is a significant part of the economy, with a substantial share of measured output. In the United States, for example, the total value-added of financial intermediaries (essentially profits, wages and salaries) amounts to about 8.1 per cent of GDP. This makes the US FI sector much larger than (say) the agricultural sector, whose share of total value-added is about one per cent (Bureau of Economic Analysis, 2006). Across countries, there is a strong correlation between size and quality of the FI sector and the level of economic development. This relationship is an important topic in development economics but such issues are not considered

here. (See financial structure and economic development.)

Organizational form

In most countries the dominant form of organization for FIs is the corporation; however, there are important exceptions. In particular, many FIs are organized as ‘mutuals’ or ‘cooperatives’. With this alternative form of organization, there is no separate class of shareholders or equity owners, as would be the case in a corporation. For example, in mutual life insurance companies the policy holders are also the owners. In mutual savings and loan associations, the depositors are the owners. These alternative organizational forms are common in the United States, Europe and many other parts of the world.

Recent trends

Since the mid-1990s, the FI sector has experienced substantial change. The main trends worldwide are towards consolidation (a smaller number of larger firms), diversification (a larger set of financial activities or ‘products’ offered at the same FI), and internationalization (operating across borders). Almost every part of the world has participated in these developments, excepting sub-Saharan Africa (De Nicolò et al., 2004).

2 History of thought on financial intermediation

In the 1960s and 1970s, the economic analysis of *FI*s was largely focused on banks, and these were viewed essentially as ‘black box’ organizations that turned high-powered money (bank reserves) into money. At that time, most intellectual interest in banks derived from their role in creating money, and their being the conduits for monetary policy. In some sense, the study of banking was in those decades incidental to the study of monetary policy and macroeconomics. There had been an earlier literature on FIs that showed great depth of understanding, but in a non-mathematical, descriptive context. Scholars such as Bagehot, Goldsmith and Schumpeter wrote about, and clearly understood, information asymmetries, liquidity, and so forth. When ambitious scholars, such as Tobin (1969) or McKinnon (1973), tried to incorporate FIs into Keynesian models before the profession had invented the mathematical tools to formally model information and liquidity, the crucial intuitive insights about the role of FIs were absent from the models. Thus, finance became money, and money was simply a stock associated with real capital.

In the mid-1980s a new body of thought emerged and was largely attributable to the seminal work of Diamond (1984) and Diamond and Dybvig (1986). Other significant papers at about that same time included Williamson (1986) and Boyd and Prescott (1986). This new approach to studying financial intermediation stressed that FIs are firms that produce valuable economic services of a variety of kinds, and *explicitly modelled the nature of those services*. This literature was careful to model the profit, share price, or utility-maximizing behaviour of FIs subject to appropriate constraints, and much of this work was done in general equilibrium. More importantly, almost all this work and the large literature that followed featured environments with private information – private in the sense that different agents were endowed with different knowledge. This was a major deviation from the previously studied world of

Arrow–Debreu, in which markets are frictionless and perfectly competitive, and all relevant information is common knowledge. It was a critical innovation because in the environment of Arrow–Debreu FIs are irrelevant (cannot increase welfare). In that world FIs are just not very interesting to study in a serious way, and they weren't.

Sequence was also very important in the development of the modern FI literature. Since the post-1983 FI literature almost exclusively employed models with private information, this meant that development of the literature depended on, and naturally followed, advances in information economics thanks to the pioneering work of Akerlof, Hurwitz, Stigler and others. Most likely, this is why earlier efforts to force FIs into Keynesian macro models were a failure; the required tools simply had not yet been invented.

In the next section, I briefly review some of the modern FI models developed in the 1980s and subsequently. Later, in Section 5, I discuss some areas in financial intermediation where, in my judgment, there remain important gaps in our knowledge.

3 The theory of financial intermediation

Banks and other FIs are firms that take in funds (FI liabilities) through a hypothetical front door, and put out funds (FI assets) through a hypothetical back door. They produce no physical products. To survive, they must earn a profit, meaning that the average rate of return on their assets must exceed the average cost of their liabilities. This spread between asset returns and liability costs must be large enough to cover operating costs (primarily wages and salaries), and to earn a rate of return to equity investors. That FIs earn such positive profits has always troubled critics of the industry (of which there have always been many), who may conclude that FIs are somehow exploiting consumers or businesses. In fact, FIs are permitted to earn these positive interest rate spreads because they provide valuable economic services to the economy, and it is costly to provide these valuable economic services. Let us next consider these services.

One important function, offered by banks but not other FIs, is payment services. This is the 'creators of money' banking function that the old literature stressed, virtually to the exclusion of other FI functions. When we need to execute transactions, we use cash and coin, paper checks, credit cards, and wire transfers. All of these transaction tools are generally provided by banks and for obvious reasons they are economically important.

Another important function of FIs is that they are 'brokers' in the sense that they bring together large numbers of ultimate borrowers and lenders. When they bring these groups together, FIs substitute their own liabilities for those of ultimate borrowers, and this is what ultimately distinguishes FIs from financial markets. This process has been given many names in the literature ('asset transformation' is common) and understanding it is key to understanding what FIs actually do. Hypothetically, consider one single bank depositor, a wealthy individual, and one single bank borrower, a small business. The bank depositor might have lent directly to the small business through the stock or bond market. Instead, by assumption, he or she lends to the bank in the form of a deposit. In turn, by assumption, the small business borrows from the bank in the form of a commercial loan. The bank places itself in the middle of the exchange and becomes the counter-party to the others.

Why is this valuable? The answer is that bank liabilities typically have different attributes from ultimate borrower liability attributes, ones that are crafted to be desirable to the bank liability holders. If they are made better off, they are willing to lend at a lower rate than they would have required to lend directly.

Thus, this process of asset transformation can, and usually does, make both borrowers and lenders better off.

For banks, the general direction of such asset transformation is well understood: bank liabilities will typically have shorter maturity than bank assets, and will be more liquid and less risky. As will become apparent, a key ingredient to this process is that the banks borrow from a large number of creditors, and lends to a large number of borrowers.

Shorter maturity

Bank liabilities often have shorter average maturity (or duration) than bank assets, and *ceteris paribus* this may make bank liabilities relatively more attractive to savers. Such maturity mismatching exposes banks to an interest rate lottery and the risk that interest rates will increase, in which case they will suffer capital losses. Bank creditors are partially protected against interest rate risk by the bank's equity, at least until that is exhausted. The degree of interest rate risk exposure naturally depends on the magnitude of the asset–liability maturity mismatch, and on how volatile are interest rates. In the 1970s, the US savings and loan (S&L) industry experienced massive losses due to interest rate risk, losses so large as to bankrupt much of the industry as well as its government insurer, the Federal Savings and Loan Insurance Corp (FSLIC). The S&Ls' maturity mismatch was substantial, and interest rates had become extremely volatile by historical standards. However, the savings and loan industry should not be blamed for this sad experience. Government regulations essentially forced this industry to borrow short and lend long. Since the mid-1990 banks and other FIs have become clever in finding ways to hedge interest rate risk in the forward, futures and swap markets. (Of course, someone still has to bear the aggregate risk.) Also, there is some evidence that, in the United States at least, FIs have in recent years become less willing to expose themselves to interest rate risk. As a practical matter, however, it is difficult to accurately measure the maturity mismatch of banks, and standard duration methods may not work very well for this industry. That's because a substantial proportion of bank liabilities are in the form of demand (checking) deposits. For these liabilities, the technical maturity is instantaneous but the true maturity is much longer, depends on economic conditions, and must be empirically estimated.

More liquid

Bank liabilities, especially deposits, are more liquid than bank assets. This is another desirable form of asset transformation since, *ceteris paribus*, lenders like to hold liquid assets. The liquidity provision function has been heavily studied by scholars, and the seminal reference on the topic is Diamond and Dybvig (1986). Now, liquidity is hard to define, let alone understand, and it may help to consider a simple theoretical environment, similar in some ways to the more complicated environment studied by Diamond and Dybvig (1986). Imagine a world in which there are only two assets: gold coins and land. By assumption, gold coins are perfectly liquid and can be spent at any time but earn no rate of return. Land, on the other hand, is highly productive but illiquid. It is hard to sell land in an emergency, and possibly it can't be sold at all. All agents in this economy have a known, say one per cent, chance of an 'emergency', the occurrence of which is independent across agents. In an emergency, agents desperately want to have all their wealth immediately so they can consume it. Now, consider the problem facing

individual agents. If they put all their wealth in coins (land) they will do well 1 (99) per cent of the time; however, they will do very badly 99 (1) percent of the time. Common sense suggests that the best strategy will be to split up their holdings, and if you guessed that you would be right at least for most preferences. Even then, however, agents are not doing as well as they potentially could in either state of the world.

Next, assume a bank is organized, which offers each individual a deposit account that can be *redeemed in gold coins on demand*. Further, assume the bank puts 1 per cent of its assets in gold coins and 99 per cent in land. Now, if the bank deals with a sufficiently large number of depositors, it will have enough coins to just cover withdrawals and all the remaining can be invested in highly productive land. Everyone is better off than they could have done on their own account.

This kind of an arrangement is usually referred to as ‘fractional reserve banking’. The key to its smashing success is diversification across a large number of depositors, and the fact that depositor withdrawal demands are independent. Now, as Diamond and Dybvig are quick to point out, this idealized solution may not always work out in practice. Suppose, for example, that emergency withdrawals become correlated, perhaps because there is a war. Then the bank can easily run out of coins, fail on its obligations, and land must be inefficiently liquidated. Even worse, just a false rumour of war could send too many depositors to the bank and cause it to fail. This sort of occurrence is called a ‘bank run’ and these have been quite common both historically and in recent times. In an imperfect world where withdrawals may be correlated and bank runs are possible, every bank faces a fundamental and unpleasant trade-off: if it holds a high fraction of gold coins (reserves) risk of insolvency will be low, but the average rate of return on its assets will be low. If it holds a high fraction of land (earning assets) its average rate of return on assets will be high, but its risk of insolvency will be high. There is a large literature on this topic, much of which is referenced in Gorton and Winton (2003).

Less risky

Bank liabilities are on average less risky than bank assets, and obviously this tends to make bank liabilities *ceteris paribus* more attractive. Now, bank liabilities can be less risky than the representative bank loan for a variety of reasons. One is that banks often place some fraction of their assets in default-risk-free government securities. A second reason is that banks raise part of their funds in the form of equity, and the bank's shareholders must suffer a total loss before liability holders lose. A third reason is that banks hold portfolios of different kinds of loans that are diversified by industry and geography, so that their loan portfolio is less risky than its individual components. A fourth reason is that in most countries bank deposits are fully or partially insured by government.

In addition, banks are very good at determining to whom to lend, and in setting loan terms for those who are funded. This topic has been heavily studied in the FI literature and the reader can find many studies under the headings ‘adverse selection’, ‘sorting’ and ‘screening’ in Gorton and Winton (2003). In most of these models, some loan applicants are better credit risks than others, applicants know their own types, and are willing to misrepresent (say they’re good when they’re bad). FIs do not know the applicants’ types, although it is conventional to assume that everyone knows the underlying distribution of applicant types. The FI's objective is to accept (reject) good (bad) applicants where possible. In some but not in all cases, it is possible to adroitly choose terms of lending such that good applicants voluntarily sign up, and bad applicants withdraw. In other cases, the best strategy is simply to accept

(reject) all applicants.

Another important aspect of lending, and an aspect at which FIs excel, is monitoring borrowers after they have received the money. This '*ex post* monitoring' has also been heavily studied in the FI literature. Once they have the money, borrowers may take actions that reduce their probability of repaying, or events beyond their control may have the same effect. To protect their interests lenders normally pre-specify loan covenants that state what happens in such cases, and they monitor borrowers to enforce these covenants. An example that homeowners will understand is a residential mortgage: to protect its interests, the lender must be sure that property taxes are being paid, and that the house is fully insured. Now, it is often the case that loans are large relative to the wealth of individual agents in the economy. This naturally occurs because many production technologies exhibit economies of scale. For example, an automobile plant must be of a particular size to be efficient, and few if any agents can fund such an investment with their own wealth. Therefore, to fund a loan often requires obtaining financing from several agents simultaneously. Unless FIs are present there is a coordination problem among the several lenders, and it is a problem first studied by Diamond (1984) and Williamson (1986).

Monitoring of borrowers is costly, and no one wants to do it if they don't have to. Now, for simplicity, assume that there are just two lenders for a given loan, lender A and lender B. Now, A (B) may assume that B (A) will monitor, in which case neither lender actually does. This is obviously undesirable because their interests are not being protected. Alternatively, lender A and lender B might both be conservative, assume the other is unreliable, and monitor themselves. In that case there would be redundant monitoring which is unnecessary and wasteful. Clearly, what is needed is an arrangement in which all lenders agree to have *ex post* monitoring done by a single, efficient 'delegated monitor'. What is critical, if such an arrangement is to work, is that the delegated monitor finds it in its own interests (incentive compatible) to actually do the work as promised. Otherwise, it might be necessary to monitor the monitor, which obviously would be inefficient, too. Diamond (1984) and Williamson (1986) showed that efficient *ex post* monitoring can be achieve by a bank that pools funds from many depositors and uses the proceeds to make many loans.

Summary

In a world in which different agents have different information sets FIs earn a positive interest spread between their average asset returns and average liability costs, in return for providing valuable services. They are brokers between ultimate borrowers and ultimate lenders, and they provide payments services. They transform ultimate financial claims in the sense that their liabilities have different attributes from their assets. Typically, their liabilities are shorter in maturity, more liquid and less risky; thus, such liabilities are more desirable to savers. This process of 'asset transformation' is not without risk. FIs are exposed to interest rate risk and particularly vulnerable to unexpected interest rate increases. We discussed the case of the US savings and loan industry and its devastating exposure to interest rate increases. Due to their liquidity provision, banks are exposed to the risk of bank runs. Bank runs have been common historically, and still have occurred with some frequency in the modern wave of banking crises. Finally, all FIs are exposed to default risk when their loans or other investments do not pay off in a timely manner.

4 An aside on equilibrium credit rationing

When economists began studying intermediation environments with private information, in which agents could withhold the facts, intentionally deceive one another, and so on, all manner of new and interesting results were obtained. One seminal model of financial intermediation featured an outcome called ‘equilibrium credit rationing’ (Stiglitz and Weiss, 1981). In such cases, at the equilibrium rate of interest there is excess demand in the sense that some would-be borrowers are denied access to credit. This is quite at odds with a classical market equilibrium, and immediately raises the question, ‘why don't lenders just increase the rate of interest to a level at which demand equals supply?’ A variety of answers to this question can be found in the literature, reflecting the different environments that have been shown to produce equilibrium credit rationing. For one example, assume that credit applicants are of two types, good and bad, and that lenders take account of borrower heterogeneity in their rate setting. Then, it can be the case that for sufficiently low interest rates both good and bad will borrow, but above some threshold rate r^* good types become unwilling to borrow. In such cases lenders may find it optimal to set the rate at r^* even though there is excess demand at that rate. A second example is an environment with moral hazard in the form of a bad action that borrowers may take *ex post* (such as increasing the risk of their investment project). For some parameterizations, when rates are below a threshold r^+, borrowers will not take the bad action, but above r^+ they will. As in the case above, it may be optimal for lenders to set the rate at r^+, thus avoiding the bad action, and resorting to credit rationing.

These first two environments are with private information: however, a third one can result in equilibrium credit rationing even when all information is public. Imagine that default by borrowers results in a deadweight loss – for example, an out-of-pocket bankruptcy cost. Then, the probability of costly default directly depends on the rate of interest, and the higher that rate is the higher the default probability is. Increasing the rate of interest increases the expected rate of return to lenders in good (non-default) states, but also increases the probability of default which is costly to both parties. Depending on the distribution of possible returns facing borrowers, it may be that raising the rate beyond some threshold r^- is futile in the sense that the marginal cost exceeds the marginal benefit. In these cases, rates above r^- are harmful to both parties and will never be observed in equilibrium. Yet it may also be true that *r*-plus is too low to clear the market, and equilibrium credit rationing will again be observed (Williamson, 1986).

Arguably, equilibrium credit rationing is a topic where theory leads measurement. There has not been a lot of good empirical work on credit rationing per se, primarily because it is so hard to do right. Credit rationing equilibria are off the usual demand and supply curves that econometricians like to estimate, and they may exhibit nasty jumps, discontinuities, and so on. If the theorists are right, however, and credit rationing is popping up all over, more empirical work would be useful, especially in the area of finance and development.

5 Regulation

Banks and other FIs are, almost without exception, rather heavily regulated. This is true in virtually all countries and has been true for centuries. There are at least three reasons for this special and obtrusive regulatory treatment. First, banks are the conduit for monetary policy, and problems in banking are likely to interfere with monetary policy conduct. Second, it is widely believed that bank failures may result in negative externalities (social costs). And third, governments may find it irresistible to control a

critical industry that creates money and allocates a large fraction of investment capital. Some recent work has emphasized the importance of political economy issues for regulation, in particular arguing that it is unlikely that bank regulation can contribute positively to social welfare in economies with weak and/or corrupt governments (Barth, Caprio and Levine, 2006).

The Great Depression was a difficult time for banks in the United States and many other countries, and during the late 1920s and early 1930s there were literally thousands of bank failures worldwide. Many of these were associated with bank runs and panics. In response, many nations substantially beefed up their regulation of FIs and put in mechanisms such as deposit insurance to reduce or eliminate the prevalence of bank runs. For example, the Federal Deposit Insurance Corporation was created by US federal legislation in 1933. Beginning in the mid-1930s, the industry stabilized (at least in developed nations), and went through a period of relative calm that lasted for about three decades. Many observers believed that these policy interventions had solved the problem of instability in banking; but that was not to be. Beginning in roughly the mid-1960s, a new wave of banking crises affected well over 100 nations. Banking crises – some of them severe – have been recently experienced in developing and developed economies alike.

No one knows for sure what has caused this interesting historical sequence of events in banking, but many scholars have emphasized that *policy interventions intended to stabilize the industry may have actually had opposite effect*. In most countries, banks have access to emergency borrowing from the government (a Discount Window), and have some form of government insurance to protect depositors. Additionally, there is a common practice known as ‘too big to fail’ whereby governments will prop up their very largest FIs if they get into trouble. This package of interventions is widely referred to as ‘the safety net’, and it has been very heavily studied. Most of the literature on this topic concludes that, whatever the benefits of a safety net, it also distorts bank incentives in a perverse way. Depositors and other bank creditors don't care how much risk the bank takes (they are protected by government), and normal market risk-constraining mechanisms become ineffective.

In the presence of a safety net, banks may have an incentive to take on more risk *ceteris paribus* than otherwise; indeed, they may even become risk lovers who intentionally seek out investments with low expected returns and high variance. It's not hard to see why this is so. If an FI has very risky investments and these payoff, all the profits go to FI shareholders. If they don't payoff the FI goes broke, but the resulting losses are mostly absorbed by government. In essence, this is a ‘heads I win tails you lose’ gamble. Perhaps the most dramatic evidence of this distortion turned up during the U.S. S&L crisis. At that time, many S&Ls were obviously bankrupt but could not be closed down since their federal deposit insurer, the FSLIC, had run out of money. Many such institutions gambled for redemption by taking extreme risks. If they were lucky enough they might survive, and if not … well, they were already broke.

As of 2007, solving the problems associated with the safety net is arguably the most vexing policy issue facing FI regulators and scholars of that industry. Many regulatory interventions, such as restrictions on asset holdings, attempt to control FIs’ behaviour but do not deal with the fundamental distortion of risk incentives. Other regulatory interventions such as capital regulation are intended to reduce FIs’ distortion of risk incentives, but may not be effective (Hellmann, Murdoch and Stiglitz, 2000). FIs have a natural tendency to try to get around all these regulations, pursuing strategies that render the regulations ineffective. On the other hand, getting rid of the safety net would have its own risks, and it is far from obvious how governments could ever credibly commit to a policy of no FI bailouts. This issue is probably best described as important but unfinished business.

6 Trends in recent research, and open research questions

1. 1. As discussed earlier, the modelling of financial intermediaries has come a long way since the mid-1980s, and most modern macroeconomic models reflect that reality. Even so, there is still recent work that reflects old ways of thinking about FIs. To make this point I provide just one example: the ongoing discussions of the so-called ‘Friedman Rule’. This rule, in simplest form, calls for a monetary policy that produces a rate of *deflation* such that the real rate of return on bank reserves equals the rate of interest on real investment. Then, it is argued, banks will voluntarily hold all their assets in the form of reserves, and bank runs, crises, and so on will never happen. Bruce Smith (whose death in 2003 was a great loss to economics) makes it beautifully clear that this once-beguiling idea should be relegated to the history of economic thought (Smith, 2002). Application of the Friedman rule may indeed result in risk-free banks. However, except for the provision of payments services, it precludes banks from making any of their valuable economic contributions detailed by Diamond (1984), Diamond and Dybvig (1986) and others, and as discussed earlier.
2. 2. Boyd and Prescott (1986) have a theorem that financial intermediary coalitions composed of large numbers of agents can support allocations that cannot be supported with decentralized markets, and are efficient subject to resource and incentive constraints. As lamented by Green and Zhou (2001), virtually all subsequent theoretical research on FIs has studied decentralized (market) environments. Now, this could be just a matter of preferences amongst theorists as to the most interesting and tractable environment in which to study FIs. It's not, in my opinion, and this topic is of more than theoretical interest. Boyd–Prescott financial intermediary coalitions look (at some high level of abstraction) like mutual or cooperative FIs. It is fact that over several continents and many centuries mutual FIs seem to endogenously spring up with great regularity. When a class of arrangements is ‘revealed preferred’ so often, there is probably a good reason for it. There has been some theoretical research on this topic, but arguably not enough.
3. 3. Virtually all of our general equilibrium models with FIs force agents into discrete silos: for example, an agent must choose to become a producer (borrower), a consumer (lender), or an FI. In reality we often observe organizations that are both producers and financial intermediaries at the same time (for example, General Electric or Cargill). Moreover, we sometimes see firms radically change their blend of activities. For example, in a few years Enron evolved from a production firm to a financial intermediary. I am aware of only one study (Bhanot and Mello, 2006) that allows, in a serious way, for endogenous choice of FI *and* non-FI activities in the same organization. More work along these lines could be useful.
4. 4. As discussed, banks, even very simple ones, perform a number of economic functions *simultaneously*: brokerage, payments service provision, maturity transformation, liquidity provision, and default risk reduction. This is what we observe in reality and there is undoubtedly a reason. Yet our theoretical models tend to isolate these economic functions and look at them one at a time. Only a few studies have seriously looked at the jointness in providing even two services simultaneously (Kasyap, Rajan and Stein, 2002). This separation of functions is done for tractability, and even then our models can become complex. Putting all of these features in a

model simultaneously becomes technically daunting, but it needs to be done. There are undoubtedly interesting interactions or synergisms among these activities, and we cannot learn about those by studying them individually.

See Also

- finance
- financial structure and economic development

Bibliography

Barth, J., Caprio, G. and Levine, R. 2006. *Rethinking Bank Regulation: Till Angels Govern*. Cambridge: Cambridge University Press.

Bhanot, K. and Mello, A. 2006. Should production and trading activities be separated? Working paper, University of Wisconsin.

Board of Governors of the Federal Reserve System. 2005. *Flow of Funds Accounts*. Online. Available at http://www.federalreserve.gov/RELEASES/z1/, accessed 16 January 2007.

Boyd, J.H. and Prescott, E.C. 1986. Financial intermediary coalitions. *Journal of Economic Theory* 2, 211–32.

Bureau of Economic Analysis. 2006. *Industry Economic Accounts*. Online. Available at http://bea.gov/bea/dn2/home/annual_industry.htm, accessed 3 February 2007.

De Nicolò, G., Bartholomew, P., Zaman, J. and Zephirin, M. 2004. Bank consolidation, internationalization and conglomeration: trends and implications for financial risk. *Financial Markets, Institutions & Instruments* 13(4), 173–217.

Diamond, D. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51, 393–414.

Diamond, D. and Dybvig, P. 1986. Banking theory, deposit insurance, and bank regulation. *Journal of Business* 59, 55–68.

Gorton, G. and Winton, A. 2003. Financial intermediation. In *Handbooks in the Economics of Finance, Volume 1A: Corporate Finance*, ed. G. Constantinides, M. Harris and R. Stulz. Amsterdam: North-Holland.

Green, E. and Zhou, R. 2001. Financial intermediation regime and efficiency in a Boyd–Prescott economy. *Carnegie-Rochester Series on Public Policy* 54, 117–29.

Hellmann, T., Murdoch, K. and Stiglitz, J. 2000. Liberalization, moral hazard in banking and prudential regulation: are capital requirements enough? *American Economic Review* 90, 147–65.

Kasyap, A., Rajan, R. and Stein, J. 2002. Banks as liquidity providers: an explanation of the coexistence of lending and deposit-taking. *Journal of Finance* 57, 33–74.

Levine, R. 2005. Finance and growth: theory and evidence. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

McKinnon, R. 1973. *Money and Capital in Economic Development*. Washington, DC: Brookings Institution.

Smith, B.D. 2002. Monetary policy, banking crises and the Friedman rule. *American Economic Review* 92, 128–34.

Stiglitz, J. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 2, 461–72.

Williamson 1986. Costly monitoring, financial intermediation, and equilibrium credit rationing. *Journal of Monetary Economics* 18, 159–79.

How to cite this article

Boyd, J. H. "financial intermediation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000320> doi:10.1057/9780230226203.0568

The New Palgrave Dictionary of Economics Online

financial liberalization

Romain Rancière, Aaron Tornell and Frank Westermann
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Financial liberalization has led to financial deepening and higher growth in several countries. However, it has also led to a greater incidence of financial crises. Here, we review the empirical evidence on these dual effects of financial liberalization across different groups of countries. We then present a conceptual framework that explains why there is a trade-off between growth and incidence of crisis, and helps account for the cross-country difference in the effects of financial liberalization.

Keywords

bail-out guarantees; banking crises; boom–bust cycles; capital account liberalization; contract enforceability; credit growth; currency crises; economic growth; equity market liberalization; financial liberalization; financial openness; foreign direct investment; India; insolvency risk; international flows; investment; investment subsidies; lending booms; portfolio flows; probit models; prudential regulation; skewness; Thailand; tradable and non-tradable Sectors; trade liberalization

Article

Financial liberalization (FL) refers to the deregulation of domestic financial markets and the liberalization of the capital account. The effects of FL have been a matter of some debate. In one view, it strengthens financial development and contributes to higher long-run growth. In another view, it induces excessive risk-taking, increases macroeconomic volatility and leads to more frequent crises. This article brings together these two opposing views.

The data reveals that FL leads to more rapid economic growth in middle-income countries (MICs), but does not have the same effect in low-income countries (LICs). In MICs this process is not smooth, however: It takes place through booms and busts. Indeed, MICs that have experienced occasional financial crises have grown faster, on average, than non-liberalized countries with stable credit conditions. In LICs liberalization does not lead to higher growth because their financial systems are not sufficiently developed so as to permit significant increases in leverage and financial flows.

The contrasting experiences of Thailand and India illustrate these dual effects. Thailand, a liberalized economy, has experienced lending booms and crises, while India, a non-liberalized economy, has followed a slow but safe growth path (see Figure 1). In India GDP per capita grew by only 99 per cent between 1980 and 2002, whereas Thailand's GDP per capita grew by 148 per cent, despite a major crisis. As will be shown below in a set of data analyses, this trade-off exists more generally across MICs.

Figure 1

Safe vs. risky growth path: a comparison of India and Thailand, 1980–2002. *Note*: The values for 1980 are normalized to 1. *Sources*: International Financial Statistics (IMF) and Word Bank Development Indicators.

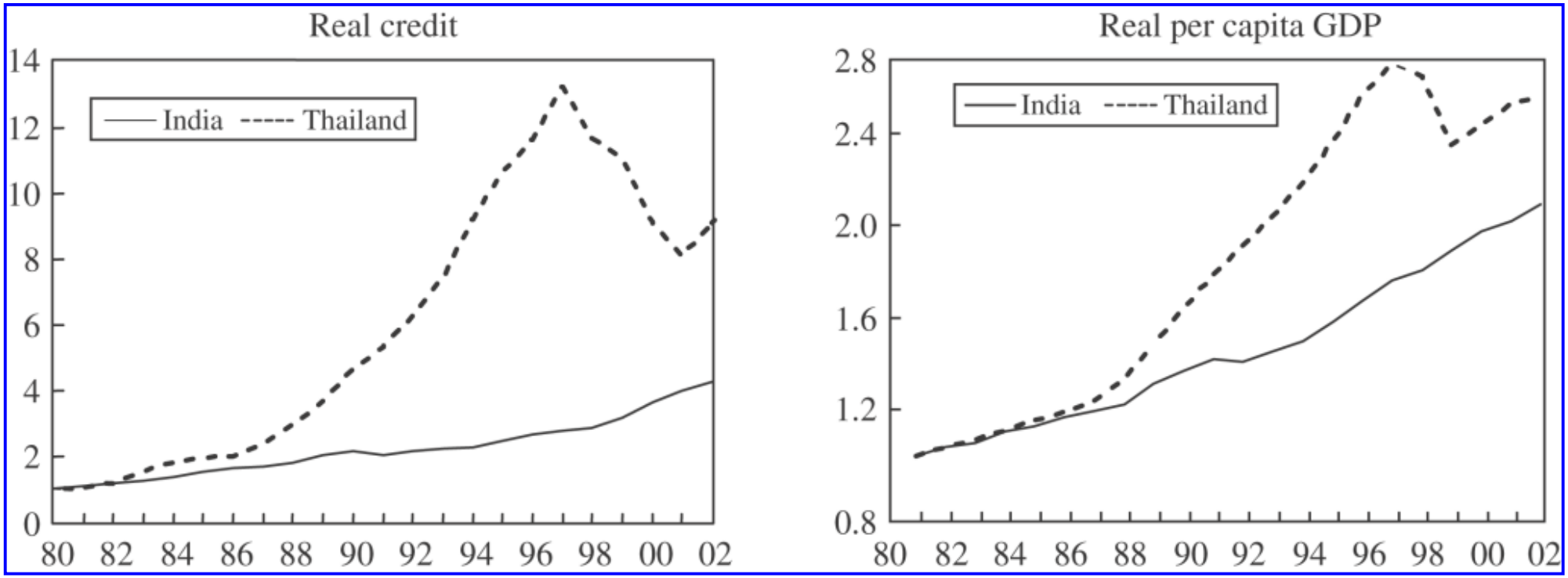

Asymmetric financial opportunities across sectors are key to understanding the effects of FL. In particular, in MICs contract enforceability problems affect the tradable (T) and non-tradable (N) sectors differently. Many T-sector firms are able to overcome these problems and gain access to international capital markets, whereas most N-sector firms are financially constrained and depend on domestic banks for their financing. Trade liberalization promotes faster productivity growth in the T-sector, but is of little direct help to the N-sector. By allowing banks to borrow on international capital markets, FL leads to an increase in investment by financially constrained firms, most of which are in the N-sector. However, while FL increases investment, it also increases borrowers' incentives to take on insolvency risk because there are implicit and explicit bail-out guarantees that cover lenders against systemic defaults. This is why greater leverage and growth is associated with aggregate financial fragility and occasional crises.

In the rest of this article we describe the ways in which FL has been measured and the empirical estimates of its effects on growth and crises. We then present a conceptual framework and a review of the policy issues. In a nutshell, any evaluation of FL must weigh its benefits against its costs. Focusing exclusively on the growth effects of liberalization during good times would miss the link between FL and crises. Focusing only on volatility and crises could lead to an excessive cautiousness about the risks of FL. The case for FL requires that its growth and welfare benefits outweigh the costs associated with more frequent financial crises.

Measuring financial liberalization

There are three classes of FL indices. First, there are *de jure* indices based on official dates of policy reforms. An example is the index based on the IMF *Annual Report on Exchange Arrangements and Exchange Restrictions* (Grilli and Milesi-Ferretti, 1995). This class of indices permits a comparison of the periods before and after liberalization. A drawback, however, is that legislated changes take time to translate into liberalization on the ground. Liberalization may even fail to materialize altogether when well-functioning domestic financial markets are absent. Bekaert, Harvey and Lundblad (BHL, 2005) overcome this problem by constructing a *de jure* indicator of equity market liberalization that records the date after which foreign investors are able to invest in domestic securities. A second class of indices uses *de facto* measures of financial openness, like the capital flows–GDP ratio used by Edison et al. (2004). The drawback is that these measures are contaminated by cyclical fluctuations and thus are imprecise indicators for dating FL. Lastly, *de facto* indices identify structural breaks in the trend of capital inflows (Tornell, Westermann and Martinez TWM, 2003). These indices combine the advantages of the two previous classes as they provide more precise FL dates based on actual, rather than merely legislated, policy reforms.

Financial liberalization and growth

BHL (2005) find that equity market liberalization leads to an increase of one percentage point in average real per-capita GDP growth. Rancière, Tornell and Westermann (RTW, 2005) find that capital account liberalization leads to a similar gain in growth. To illustrate the link between FL and growth we add liberalization dummies to a standard growth regression:

$$\Delta y_{it} = \lambda y_{i,\text{ini}} + \gamma X_{it} + \varphi_1 TL_{it} + \varphi_2 FL_{it} + \varepsilon_{jt}, \quad (1)$$

where Δy_{it} is the average growth rate of GDP per capita; $y_{i,\text{ini}}$ is the initial level of GDP per capita; X_{it} is a vector of control variables that includes initial human capital, the average population growth rate, and life expectancy; and TL_{it} and FL_{it} are the trade and financial liberalization dummies of TWM (2003), respectively. For each country and each variable, we construct ten-year averages starting with the period 1980–9 and rolling forward to the period 1990–9. Thus each country has up to ten data points in the time-series dimension. The liberalization dummies take values in the interval [0,1], depending on the proportion of liberalized years in a given window. We estimate the panel regressions using generalized least squares.

The FL dummy enters significantly at the five percent level in all regressions. Regression 1-1 of Table 1 shows that following FL growth in real GDP per-capita increases by 1.5 percentage points per year, after controlling for the standard variables. Trade liberalization increases growth by 0.8 percent per year (column 1-2). When both liberalization dummies are included (column 1–3), both enter significantly. This suggests that trade and financial liberalization have independent effects and jointly contribute to higher long-run growth.

Regressions explaining growth in GDP per capita, 1980–99[a]

Independent variable[a]	1–1	1–2	1–3	1–4	1–5[b]	1–6	1–7[b]
Mean of real credit growth rate				0.154[c]	0.170[c]	0.110[c]	0.093[c]
			(0.009)	(0.012)	(0.009)	(0.007)	
Standard deviation of real credit growth rate				–0.030[c]	–0.029[c]	–0.019[c]	–0.014[c]
			(0.003)	(0.007)	(0.004)	(0.003)	
Negative skewness of real credit growth rate				0.266[c]	0.174[c]	0.135[c]	–0.095[c]
			(0.021)	(0.069)	(0.031)	(0.053)	
Financial liberalization	1.530[c]		1.443[c]			1.811[c]	1.894[c]
	(0.191)		(0.221)			(0.163)	(0.122)
Trade liberalization		0.793[c]	0.776[c]			0.895[c]	0.838[c]
		(0.152)	(0.196)			(0.198)	(0.155)
Summary statistics:							
Adjusted R^2	0.848	0.897	0.807	0.629	0.667	0.731	0.752
No. of observations	409	430	408	424	269	408	253

Notes: [a]The estimated equations are eqs. (1) and (2) in the text; the dependent variable is the average annual growth rate of real GDP per capita. Control variables include initial per capita income, secondary schooling, population growth, and life expectancy. Standard errors are reported in parentheses and are adjusted for heteroskedasticity according to Newey and West (1987).

[b]This regression includes the group of middle-income countries only.

[c]Significance at the 5% level. The equation is estimated in an overlapping panel regression by GLS with data as ten-year averages starting with 1980–9 and rolling forward to 1990–9.

Source: Authors' regressions.

Figure 2 illustrates the link between FL and growth for individual MICs. For each country, we plot growth residuals before and after FL. Growth residuals are obtained by regressing real per capita growth on initial income per capita and population growth. Figure 2 shows clearly that for almost all countries growth has been higher in the financially liberalized period.

Figure 2

Liberalization and annual percent growth. *Note*: The country episodes are constructed using windows of different length for each country. Country episodes that are shorter than five years are excluded. Averaging over these periods, we estimate a simple growth regression by OLS in which real per capita growth is the dependent variable and which only includes the respective initial income and population growth. The figure plots the residuals from this regression, from 1980 to 1999. *Sources*: Population growth for Portugal: International Financial Statistics (IMF). All other series: World Bank Development Indicators.

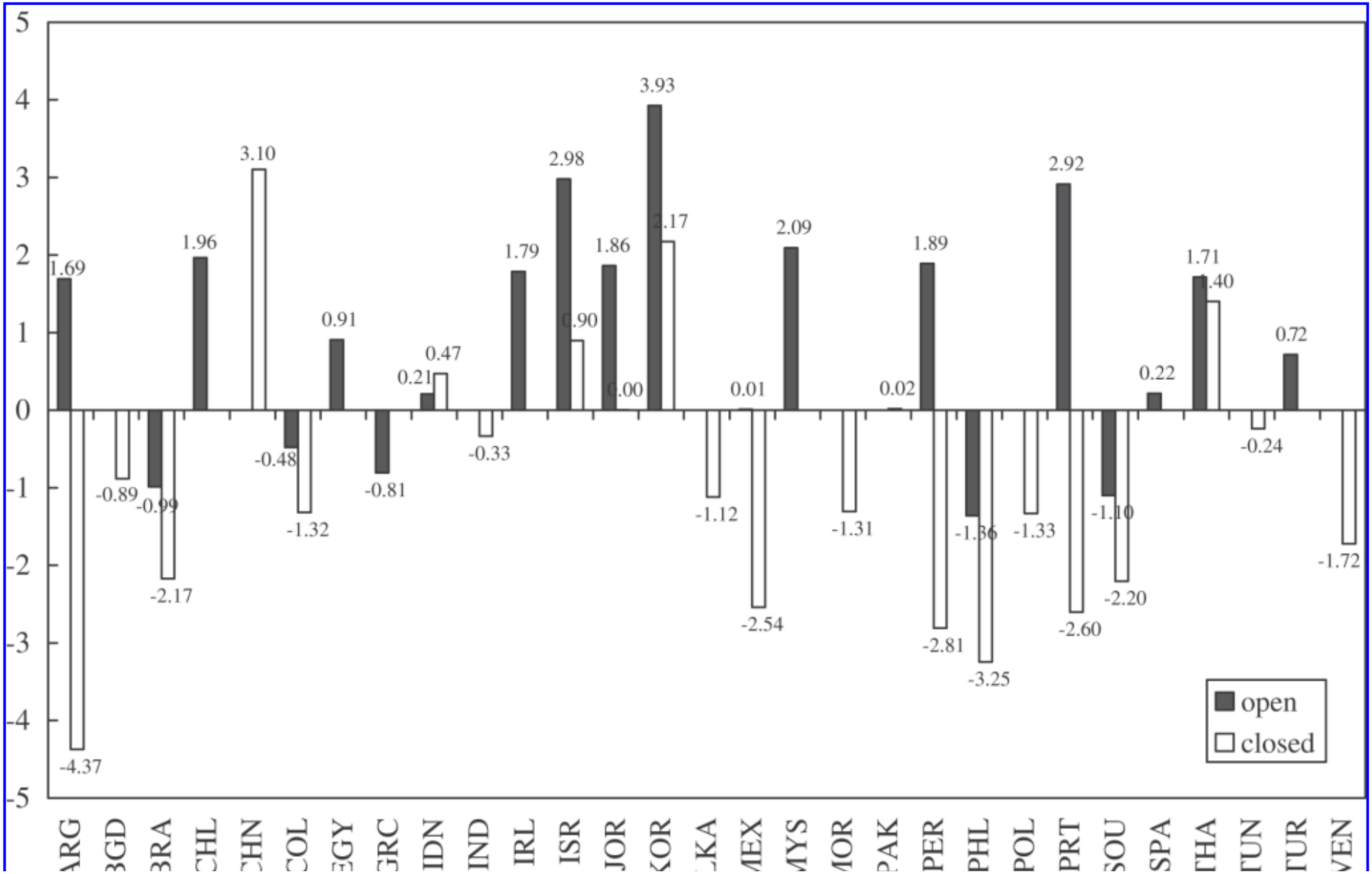

Several studies find mixed evidence on the link between financial openness and growth. This can be attributed either to the indicators of openness used or to the sample considered. First, some studies include low-income countries that do not have functioning financial markets. In these countries we do not expect the financial deepening mechanism to work. One might also expect the growth effect of FL to be smaller in high-income than in middle-income countries as the latter face more severe borrowing constraints. Hence, sample heterogeneity can create a bias against finding a linear growth effect of FL. Klein (2005) finds that FL contributes to growth among MICs but not among poor or rich countries.

Second, some studies test the effect of changes in the capital flows–GDP ratio on growth. However, because this index does not identify a specific liberalization date, it is not appropriate for comparing the behaviour of macroeconomic variables before and after liberalization. Furthermore, these measures tend to exhibit year-to-year fluctuations that do not reflect actual changes in the degree financial openness.

Financial liberalization and crises

FL is typically followed by boom–bust cycles. During the boom, bank credit expands very rapidly and excessive credit risk is undertaken. As a result, the economy becomes financially fragile and prone to crisis. Although the likelihood that a lending boom will crash in a given year is low, many booms do eventually end in a crisis. During such a crisis, new credit falls abruptly and recuperates only gradually.

The incidence of crises can be measured by analysing countries' financial histories and by codifying the occurrence of banking crises, currency crises, and sudden stops in capital inflows. Kaminsky and Reinhart (1999) use such a crisis index in a probit model to test whether banking and currency crises are crises are more likely to occur after FL.

RTW (2005) use a more parsimonious indicator of financial fragility: the *negative skewness* of credit growth. Negative skewness is a *de facto* indicator that captures the existence of infrequent, sharp and abrupt falls in credit growth. Since credit growth is relatively smooth during boom periods, and crises happen only occasionally, in financially fragile countries the distribution of credit growth rates is characterized by negative outliers in a long enough sample. These outliers correspond to the abrupt falls in credit growth that occur during the crisis or 'bust' stage of the boom–bust cycle. The advantages of this skewness measure, relative to other more complex indicators of crises, are that it is objective and comparable across countries.

In the literature variance is the typical measure of volatility. Variance, however, is not a good instrument to identify growth-enhancing credit risk because high variance reflects not only the presence of boom–bust cycles but also the presence of high-frequency shocks.

Table 2 partitions country-years into two groups: liberalized and non-liberalized. The table shows that, across MICs, the financial deepening induced by FL has not been a smooth process but has been characterized by booms and occasional busts. We can see that FL leads to an increase in the mean of credit growth of four percentage points (from 3.8 percent to 7.8 percent) and a fall in the skewness of credit growth from near zero to –1.09, and has only a negligible effect on the variance of credit growth. Notice that, across high-income countries, credit growth exhibits near-zero skewness, and both the mean and the variance are smaller than across MICs. This difference reflects the absence of severe credit market imperfections in high-income countries.

Moments of credit growth before and after financial liberalization

Moment	Liberalized country-years	Non-liberalized country-years
MICs		
Mean	0.078	0.038
Standard deviation	0.151	0.170
Skewness	–1.086	0.165
HICs		
Mean	0.025	…
Standard deviation	0.045	…
Skewness	0.497	…

Note: The sample is partitioned into two country-year groups: liberalized and non-liberalized. Before the standard deviation and skewness are calculated, the means are removed from the series and data errors for Belgium, New Zealand and the United Kingdom are corrected for. The total sample ranges from 1980 to 1999.

Source: Authors' calculations.

Growth and crises

To close the circle we show that countries with a greater incidence of crises countries have grown faster than those with smooth credit paths. We do so by adding three moments of real credit growth to growth regression (1)

$$\Delta y_{it} = \lambda y_{i,\mathrm{ini}} + \gamma X_{it} + \beta_1 \mu_{\Delta B,it} + \beta_2 \sigma_{\Delta B,it} + \beta_3 S_{\Delta B,it} + \varphi_1 TL_{it} + \varphi_2 FL_{it} + \varepsilon_{jt}, \quad (2)$$

where Δy_{it}, $y_{i,\mathrm{ini}}$, X_{it}, TL_{it}, and FL_{it} are defined as in eq. (1), and $\mu_{\Delta B,it}$, $\sigma_{\Delta B,it}$, and $S_{\Delta B,it}$ are the mean, standard deviation, and skewness of the real credit growth rate, respectively. We estimate eq. (2) using the same type of overlapping panel data regression as for eq. (1). Columns 1–4 through 1–7 of Table 1 report the estimation results. Consistent with the literature, we find that, after controlling for the standard variables, the mean growth rate of credit has a positive effect on long-run GDP growth, and the variance of credit growth has a negative effect. Both variables enter significantly at the five percent level in all regressions.

The first key point is that the financial deepening that accompanies rapid GDP growth is not smooth but, rather, takes place via booms and busts. Columns 1–4 and 1–5 show that negative skewness – a bumpier growth path – is on average associated with faster GDP growth across countries with functioning financial markets. This estimate is significant at the five percent level.

To interpret the estimate of 0.27 for skewness, consider India, which has near-zero skewness, and Thailand, which has a skewness of about minus 2. A point estimate of 0.27 implies that an increase in the bumpiness index of 2 (from zero to minus 2) increases the average long-run GDP growth rate by 0.54 of a percentage point a year. Is this estimate economically meaningful? To address this question, note that, after controlling for the standard variables, Thailand grows about two percentage points faster per year than India. Thus, about a quarter of this growth differential can be attributed to credit risk taking, as measured by the skewness of credit growth.

The second key point is that the association between skewness and growth does not imply that crises are good for growth. Crises are costly. They are the price that has to be paid in order to attain faster growth in the presence of credit market imperfections. To see this, consider column 1–6. When the FL dummy is included, bumpiness enters with a negative sign (and is significant at the five percent level). In the MIC set, given that there is FL, the lower the incidence of crises the better. We can see the same pattern when we include high-income countries in column 1–7.

Clearly, liberalization without fragility is best, but the data suggest that this combination is not available to MICs. Instead, the existence of contract enforceability problems implies that liberalization leads to higher growth because it eases financial constraints but, as a by-product, also induces financial fragility. However, because crises occur relatively rarely, FL has a positive net effect on long-run growth.

A unified approach

An alternative approach to understand the contrasting effects of FL is to combine the linear growth regression with a crisis probit model. In this way one can decompose the net effect of FL into a direct pro-growth effect and an indirect anti-growth effect, via a higher propensity to crises. Using this approach, RTW (2006) find that the direct effect of FL on growth is 1.2 percentage points and the indirect effect is minus 0.25 percentage points. In order to understand this result, one should keep in mind that even in financially liberalized countries crises are rare events. Therefore, even if crises have large output consequences, their estimated growth effect remains modest. In contrast, since FL is likely to improve access to external finance, it has a first-order impact on growth.

Conceptual framework

To analyze FL and the subsequent boom–bust cycles, consider an economy with two sectors: non-tradables (N) and tradables (T). Alternatively, one can think of ‘new-economy’ and ‘traditional’ sectors, respectively. The key is that each sector uses as input the other sector’s output.

This economy is subject to severe contract enforceability problems that generate financing constraints. While T-firms can overcome such constraints and finance themselves in bond and equity markets, most N-firms are financially constrained and bank-dependent. Since N-goods serve as intermediate inputs for both sectors, the N-sector constrains the long-run growth of the T-sector and that of GDP: there is a bottleneck.

In such an economy, FL increases GDP growth by increasing the investment of financially constrained firms. However, the easing of financial constraints is associated with the undertaking of insolvency risk because FL not only lifts restrictions that preclude risk taking but also is associated with explicit and implicit systemic bail-out guarantees that cover

creditors against systemic crises.

It is a stylized fact that, if a critical mass of borrowers is on the brink of bankruptcy, authorities will implement policies to ensure that creditors get repaid (at least in part) and thus avoid an economic meltdown. These bail-out policies may come in the form of an easing of monetary policy in response to a financial crash, the defence of an exchange rate peg in the presence of liabilities denominated in foreign currency, or the recapitalization of the financial sector.

Because domestic banks have been the prime beneficiaries of these guarantees, investors use domestic banks to channel resources to firms that cannot pledge international collateral. Thus liberalization results in biased capital inflows. T-firms and large N-firms are the recipients of foreign direct investment (FDI) and portfolio flows, whereas most of the inflows to the N-sector are intermediated through domestic banks, which enjoy bail-out guarantees. Insolvency risk often takes the form of maturity mismatch or risky debt denomination (currency mismatch).

Taking on insolvency risk reduces expected debt repayments because authorities will cover part of the debt obligation in the event of a systemic crisis. Thus the guarantee allows financially constrained firms to borrow more than they could otherwise. This increase in borrowing and investment is accompanied by an increase in insolvency risk. When many firms take on insolvency risk, aggregate financial fragility arises together with increased N-sector investment and growth. Faster N-sector growth then helps the T-sector grow faster because N-sector goods are used in T-sector production. Therefore, the T-sector will enjoy more abundant and cheaper inputs than otherwise. As a result, as long as a crisis does not occur, growth in a liberalized economy is faster than in a non-liberalized one.

Of course, financial fragility implies that a self-fulfilling crisis may occur. And during crises GDP growth falls. Crises must be rare, however, in order to occur in equilibrium – otherwise agents would not find it profitable to take on credit risk in the first place. Thus, average long-run growth is greater along a risky path than along a safe one even if there are large crisis costs. This is why FL leads both to higher long-run growth and to a greater incidence of crises. Schneider and Tornell (2004) and RTW (2003) formalize the intuitive argument we described using a general equilibrium model with rational agents.

This discussion of the mechanism through which FL affects the growth of MICs also explains why FL does little to improve the growth of LICs. LICs often do not have functioning financial markets and thus lack the infrastructure that allows the financial system to direct international funds to profitable firms. MICs, by contrast, have enough financial infrastructure to allocate funds reasonably well, even though contract enforceability problems prevent them from doing so as efficiently as high-income countries (HICs). Because of the imperfections in their financial systems, the price of fast growth in MICs is financial fragility. The contrasting experiences of Thailand and India during the period 1980–2002 illustrate this trade-off clearly. As we discussed earlier, Thailand experienced booms and busts while India did not. While Thailand experienced spectacular growth, India's growth was dismal. Recently, India has opened its economy to both trade and finance. Not surprisingly, India is currently experiencing a lending boom. It will be interesting to analyse the evolution of the Indian economy around 2015.

Economic policy

Several observers have suggested that partial liberalization is the optimal policy to reap the growth benefits of openness without having to suffer from volatility and crises. They suggest the implementation of trade liberalization but not of FL, or the restriction of capital flows to FDI, the least volatile form of capital flows. These recommendations seem impractical. First, an open trade regime is usually sustained by an open financial regime because exporters and importers need access to international financial markets. Since capital is fungible, it is difficult to insulate the financial flows associated with trade transactions. The data indicates that trade liberalization has typically been followed by FL. As Figure 3 shows, by 1999 72 percent of countries that had liberalized trade had also liberalized financial flows, bringing the share of MICs that were financially liberalized to 69 percent, from 25 percent in 1980.

Figure 3

Share of MICs that liberalized trade and financial flows, 1980–99. *Note*: The figure shows the share of countries that have liberalized relative to the total number of MICs in our sample. *Source*: Tornell, Westermann and Martinez (2003).

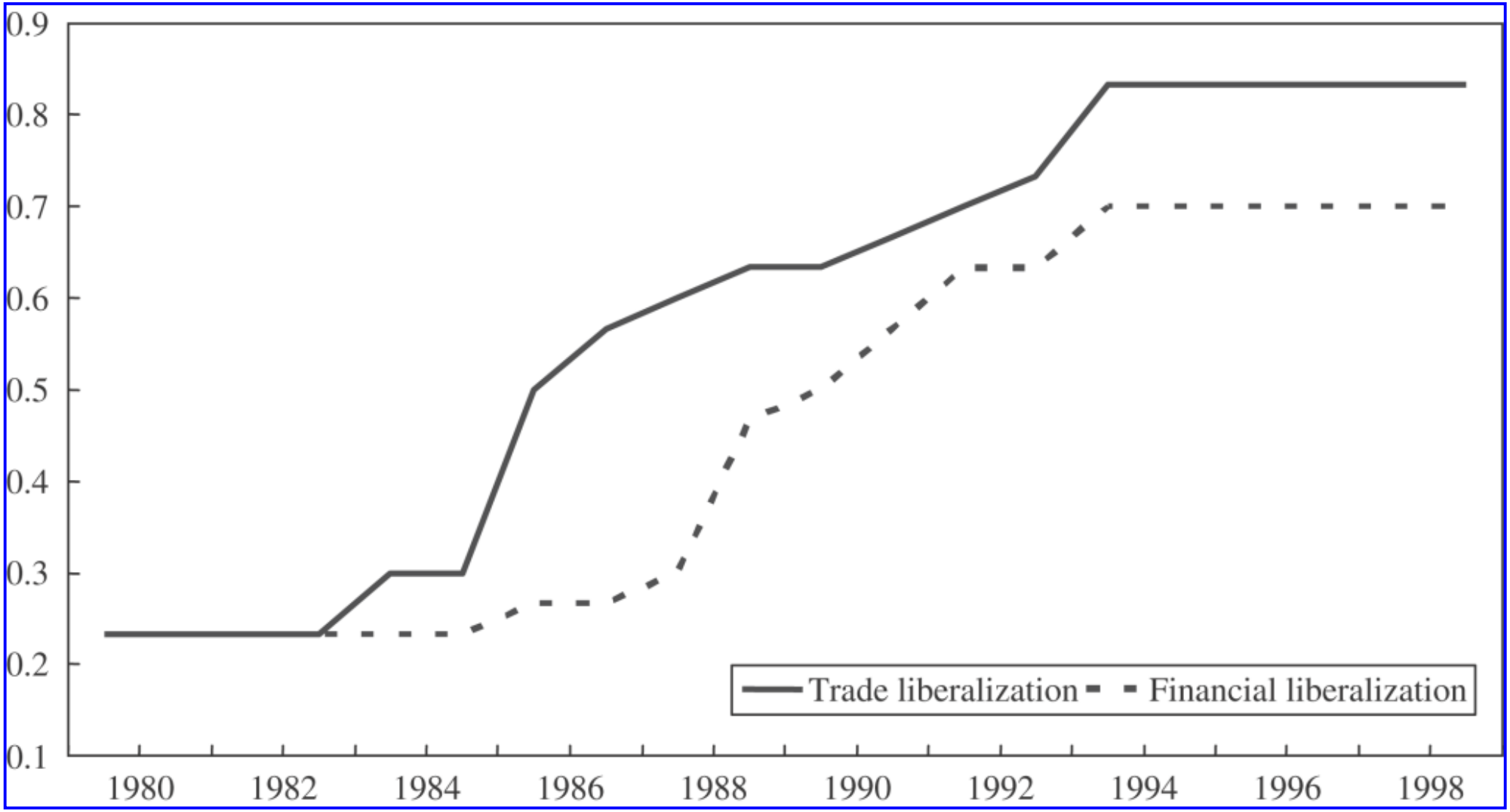

Second, FDI does not obviate the need for risky international bank flows. FDI goes mostly to financial institutions and large firms, which are mostly T-firms. Thus, bank flows are practically the only source of external finance for most N-firms (Tornell and Westermann, 2005). Curtailing such risky flows would reduce N-sector investment and generate bottlenecks that would limit long-run growth. Bank flows are hardly to be recommended, but for most firms it might be that or nothing. Clearly, allowing risky capital flows does not mean that anything goes. Appropriate prudential regulation must also be in place.

In an environment with asymmetric financial opportunities authorities may be tempted to make direct investment subsidies to constrained sectors. The historical evidence indicates that such centrally planned policies typically fail. We now know that either authorities do not possess the appropriate information or crony capitalism and rampant corruption take over. A second-best policy is to liberalize financial markets and allow banks to be the means through which resources are channelled to financially constrained firms. Here, it is important to make a distinction between ‘systemic’ and ‘unconditional’ bail-out guarantees. The former are granted only if a critical mass of agents default. The latter are granted on an idiosyncratic basis whenever there is an individual default. We have argued that, if authorities can commit to grant only systemic guarantees, and if prudential regulation works efficiently, then FL will induce higher long-run growth in a credit-constrained economy. In contrast, if guarantees are granted on an unconditional basis or there is a lax regulatory framework, the monitoring and disciplinary role of banks in the lending process will be negated. In this case, FL will simply lead to overinvestment and corruption.

One should not conclude that in order to enjoy the growth and welfare benefits of FL countries have to be exposed for ever to the risk of crises. The amelioration of contract enforceability problems, through a better legal system and other institutional reforms, is a fundamental source of higher growth and lower volatility in the long-run. However, it often takes time for these reforms to be achieved. In the meantime, countries with functioning financial markets can be made better off by liberalizing and experiencing a rapid but risky

growth path, rather than remaining closed and trapped in a safe but slow growth path.

See Also

- banking crises
- currency crises
- foreign direct investment
- international capital flows

Bibliography

Bekaert, G., Harvey, C. and Lundblad, C. 2005. Does financial liberalization spur growth? *Journal of Financial Economics* 77, 3–55.

Edison, H., Klein, M., Ricci, L. and Sløk, T. 2004. Capital account liberalization and economic performance: survey and synthesis. *IMF Staff Papers* 51, 111–55.

Grilli, V. and Milesi-Ferretti, G. 1995. Economic effects and structural determinants of capital controls. Working Papers No. 95/31. Washington, DC: IMF.

Kaminsky, G. and Reinhart, C. 1999. The twin crises: the causes of banking and balance-of-payments problems. *American Economic Review* 89, 473–500.

Klein, M. 2005. Capital account liberalization, institutional quality and economic growth: theory and evidence. Working Paper No. 11112. Cambridge, MA: NBER.

Newey, W. and West, K. 1987. A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.

Rancière, R., Tornell, A. and Westermann, F. 2003. Crises and growth: a re-evaluation. Working Paper No. 10073. Cambridge, MA: NBER.

Rancière, R., Tornell, A. and Westermann, F. 2005. Systemic crises and growth. Working Paper No. 11076. Cambridge, MA: NBER.

Rancière, R., Tornell, A. and Westermann, F. 2006. Decomposing the effects of financial liberalization: crises vs. growth. *Journal of Banking and Finance* 30, 3331–48.

Schneider, M. and Tornell, A. 2004. Balance sheet effects, bailout guarantees and financial crises. *Review of Economic Studies* 71, 883–913.

Tornell, A. and Westermann, F. 2005. *Boom-Bust Cycles and Financial Liberalization*. Cambridge, MA: MIT Press.

Tornell, A., Westermann, F. and Martinez, L. 2003. Liberalization, growth and financial crisis: lessons from Mexico and the developing world. *Brookings Papers on Economic Activity* 2, 1–112.

How to cite this article

Rancière, Romain, Aaron Tornell and Frank Westermann. "financial liberalization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000314> doi:10.1057/9780230226203.0569

The New Palgrave Dictionary of Economics Online

financial market anomalies

Donald B. Keim
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Financial market anomalies are cross-sectional and time series patterns in security returns that are not predicted by a central paradigm or theory. The focus here is on equity market anomalies including the size effect, value effect, serial correlation in returns and calendar-related patterns in returns related to month of the year and day of the week. Many of these patterns have persisted for decades, suggesting they are not evidence of market inefficiencies. Although transactions costs might preclude trading that would eliminate such patterns, it is possible that our benchmark models might be less than complete descriptions of equilibrium price formation.

Keywords

after-tax asset pricing models; bid-ask spread; capital asset pricing model; capital gains tax; cross section of stock returns; dividend yield effect; equity premium puzzle; financial market anomalies; informational efficiency; Kuhn, T.; liquidity effect; measurement error; momentum effect; risk premia; size effect; taxation of income; time series patterns in security returns; value effect; weekend effect

Article

Financial market anomalies are cross-sectional and time series patterns in security returns that are not predicted by a central paradigm or theory. This sense of the term ‘anomaly’ can be traced to Kuhn (1970). Documentation of anomalies often presages a transitional phase towards a new paradigm.

Discoveries of financial market anomalies typically arise from empirical tests that rely on a joint null hypothesis – to wit, security markets are informationally efficient *and* returns behave according to a pre-specified equilibrium model (for example, the capital asset pricing model, CAPM). If the joint hypothesis is rejected, we cannot attribute the rejection to either branch of the hypothesis. Thus, even though anomalies are often interpreted as evidence of market inefficiency, such a conclusion is inappropriate because the rejection may be due to an incorrect equilibrium model. Some have argued that, once identified by researchers, the magnitude of financial anomalies will tend to dissipate as investors seek to profitably exploit the return patterns or because their discovery was simply a sample-specific artifact. Although this has happened for some of the findings discussed below (such as the weekend effect), most of the anomalies discussed continue to persist. The fact that so many of these patterns have persisted for decades suggests that they are not evidence of market inefficiencies. Rather, our benchmark models might be less than complete descriptions of equilibrium price formation.

The number of documented anomalies is large and continues to grow. The focus here is on equity market anomalies, and on the subset whose existence has proven most robust with respect to both time and the number of stock markets in which they have been observed. We broadly classify the findings as being cross-sectional or time series in nature.

Cross-sectional return patterns

Given certain simplifying assumptions, the CAPM states that the return on a security is linearly related to the security's non-diversifiable risk (or beta) measured relative to the market portfolio of all marketable securities. If the model is correct and security markets are efficient, security returns will on *average* conform to this linear relation.

Empirical tests of the CAPM first became possible with the creation of computerized databases of stock prices in the United States in the 1960s. To implement the tests, researchers

often estimate cross-sectional regressions of the form

$$R_i = a_o + a_1\beta_i + \sum a_j c_{ij} + e_i \quad (1)$$

where β_i is the security's beta which measures its covariance with the return on the market and c_{ij} represents security-specific characteristic j (size, earnings yield, and so on) for security i. The CAPM predicts that the a_j , for j>1, are zero. Early tests supported the CAPM (for example, significant positive values for a_1, insignificant values for a_j, for j>1). The explanatory power of beta came into question in the late 1970s when researchers identified security characteristics such as the earnings-to-price ratio and market capitalization of common equity with more explanatory power than beta.

This section presents a sample of the more important contributions in this area that collectively stand as a challenge for alternative asset pricing models.

The value effect

The value effect refers to the positive relation between security returns and the ratio of accounting-based measures of cash flow or value to the market price of the security. Examples of the accounting-based measures are earnings per share and book value of common equity per share. Investment strategies based on the value effect have a long tradition in finance and can be traced at least to Graham and Dodd (1940). Ball (1978) argues that variables like the earnings-to-price ratio (E/P) are proxies for expected returns. Thus, if the CAPM is an incomplete specification of priced risk, it is reasonable to expect that E/P might explain the portion of expected return that is compensation for risk variables omitted from the tests. Basu (1977) was the first to test the notion that value-related variables might explain violations of the CAPM. He found a significant positive relation between E/P ratios and average returns for US stocks that could not be explained by the CAPM. Reinganum (1981) confirmed and extended Basu's findings. Rosenberg, Reid and Lanstein (1985), De Bondt and Thaler (1987) and many others have documented a significant positive relation between returns and the book-to-price ratio (B/P). Researchers have also identified a significant relation between security returns and value ratios that use cash flow (earnings plus accounting depreciation expense) in place of earnings in the numerator of the ratio. The value effect in its many forms has been reproduced by numerous researchers for many different sample periods and for most major securities markets around the world (see Hawawini and Keim, 2000, for a review).

Dividend yield, the ratio of cash dividend to price, has also been shown to have cross-sectional return predictability. Although similar in construction to the value ratios, the explanatory power of dividend yields is most often attributed to the differential taxation of capital gains and ordinary income as described in the after-tax asset pricing models developed by Brennan (1970) and Litzenberger and Ramaswamy (1979). Although a positive relation between stock returns and dividend yields has been documented in many studies, interpretation of the results as support for an after-tax pricing model has been controversial. Evidence on the dividend yield effect has been provided by Litzenberger and Ramaswamy (1979), Miller and Scholes (1982) and many others.

The size effect

The size effect refers to the negative relation between security returns and the market value of the common equity of a firm. Banz (1981) was the first to document this phenomenon for US stocks (see also Reinganum, 1981). In the context of eq. (1), Banz found that the coefficient on size has more explanatory power than the coefficient on beta in describing the cross section of returns. Indeed, Banz finds little explanatory power for market betas. Like the value effect, the size effect has been reproduced for numerous sample periods and for most major securities markets around the world (Hawawini and Keim, 2000).

Interpretation of the value and size effects

The separately identified value and size effects are not independent phenomena because the security characteristics all share a common variable – price per share of the firm's common stock. Indeed, researchers have shown a high rank correlation between size and price and between the value ratios and price, and others have documented a significant cross-sectional relation between price per share and average returns. To sort out the relative importance of the different variables, Fama and French (1992) (FF) estimate eq. (1) with

multiple value and size variables included as explanatory variables (see also Jaffe, Keim and Westerfield. 1989). FF find that B/P and Size provide the greatest explanatory power in describing the cross section of returns, and suggest that B/P and Size are proxies for the influence of two additional risk factors omitted from the CAPM. In this context, the value and size variables can be viewed as capturing sensitivities to the omitted factors, and the coefficients multiplying the value and size variables (a_j in eq. (1)) are estimates of the risk premia required to compensate for that exposure. (A valid question is whether a characteristic like B/P proxies for an underlying (but unknown) risk factor which is the determinant of expected returns or whether the characteristic itself is the determinant of expected returns. Daniel and Titman, 1997, address this issue and conclude that security characteristics appear to be more important than the covariance of security returns with a factor related to the characteristic.) Predicated on this interpretation, Fama and French (1993) propose a three-factor model to describe the time series behaviour of security returns:

$$R_t - r_{f,t} = \beta_0 + \beta_1(r_{m,t} - r_{f,t}) + \beta_2 SmB_t + \beta_3 HmL_t + \varepsilon_t \quad (2)$$

where R_t is the return on the asset in month t, $r_{f,t}$ is the monthly treasury bill rate, $r_{m,t}$ is the return on a value-weighted market portfolio, SmB_t is a monthly size premium (Small stock return minus Large stock return), HmL_t is a monthly value premium (High B/P return minus Low B/P return), and ε_t is the error term. As constructed, *SmB* and *HmL* are zero net investment portfolios. If these three factors span all sources of common systematic co-movement in security returns, β_0 ('alpha') will on average equal zero. The model has received much empirical confirmation and appears to explain numerous previously reported incidences of anomalous cross-sectional return patterns (that is, such effects have $\beta_0 = 0$ in eq. (2)).
As mentioned above, the mean values of the three factors in model (2) can be interpreted as the premium or compensation earned by an investment position for unit exposure to each separate factor. The relative magnitudes of these factor premia are of economic interest. The market risk premium quantifies the return, in excess of a default-risk-free return, provided for investing in a broadly diversified portfolio as represented by the value-weighted market portfolio. Over the period 1927–2005 the average equity market risk premium in (2) is 0.64 per cent per month. Utility-based asset pricing models have difficulty explaining an equity premium of this magnitude – either because the returns on default-risk-free bonds are too low, or the returns on equities are too high. This has been called the equity premium puzzle (Mehra and Prescott, 1985) and has generated an extensive literature trying to reconcile the theory and empirical evidence.
The mean risk premia associated with the size effect (*SmB*) and the value effect (*HmL*) should be zero if the CAPM is correct. Consistent with the research described above, *SmB* and *HmL* are both positive. For the period January 1927–December 2005 the monthly mean (*t*-value) is 0.25 per cent (2.01) for *SmB* and 0.48 per cent (3.64) for *HmL*, and the correlation between the two premia is 0.13. Figure 1 plots the time series of the intra-year monthly means of the two premia. The figure shows that (*a*) both premia display substantial variability over time and (*b*) the two series display a considerable common co-movement despite the low estimated correlation.
Figure 1
The value and size premia, 1927–2005

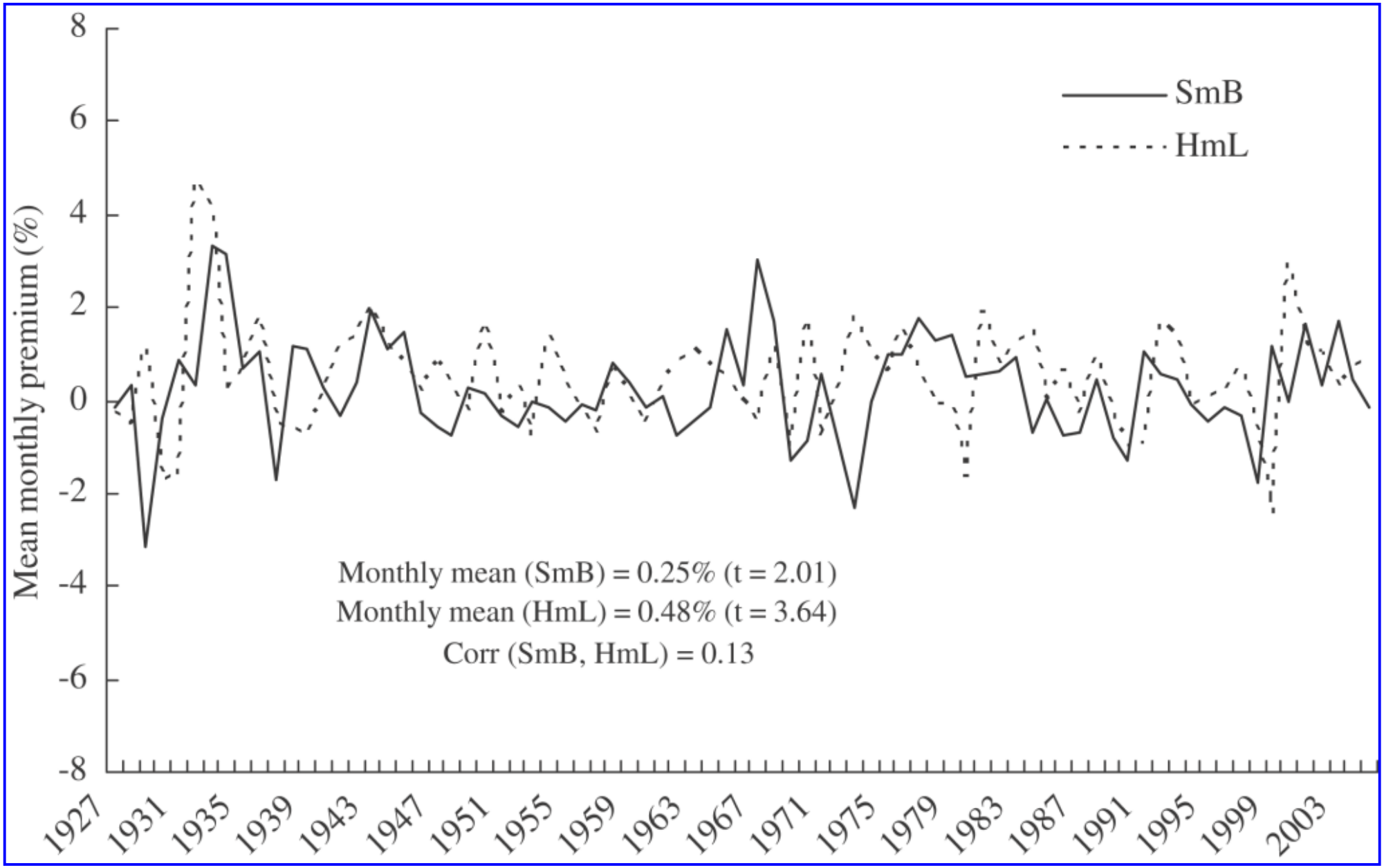

On the first point, there are extended periods when the signs of the risk premia are reversed. This is particularly evident for the size effect – for extended periods in the 1950s and the 1980s large firms outperformed small firms, in contrast to other periods (1930s, 1940s, 1970s, and post-2000) when small stocks outperformed large stocks. Because the estimated magnitudes of the effects are sensitive to the period in which they are measured, it is important to distinguish between unconditional and conditional expected values for the effects. Further, it is relevant to ask whether the 79-year sample we have for the US market (longer than in other developed equity markets) is long enough to capture the 'long-run' magnitudes of such volatile effects. (The same caveat has been raised regarding the magnitude of the equity premium.)

On the second point, the visual appearance of common co-movement between the series suggests the two effects are not entirely independent. This possibility is confirmed when the time series plots of *SmB* and *HmL* are decomposed into separate plots for January and February-to-December observations (Figure 2A and 2B). Much research has shown that the size and value effects are most pronounced in the month of January. This research is discussed in more detail in the next section. For now, we limit discussion to the difference in the

behaviour of *SmB* and *HmL* between January and February–December. First, the mean values for both premia are an order of magnitude larger in January than in February–December. Second, the correlation of 0.40 in January versus 0.06 for February–December demonstrates that the commonality between the two series in Figure 1 arises mostly from their common behaviour in January.

Figure 2A

The value and size premia – January only

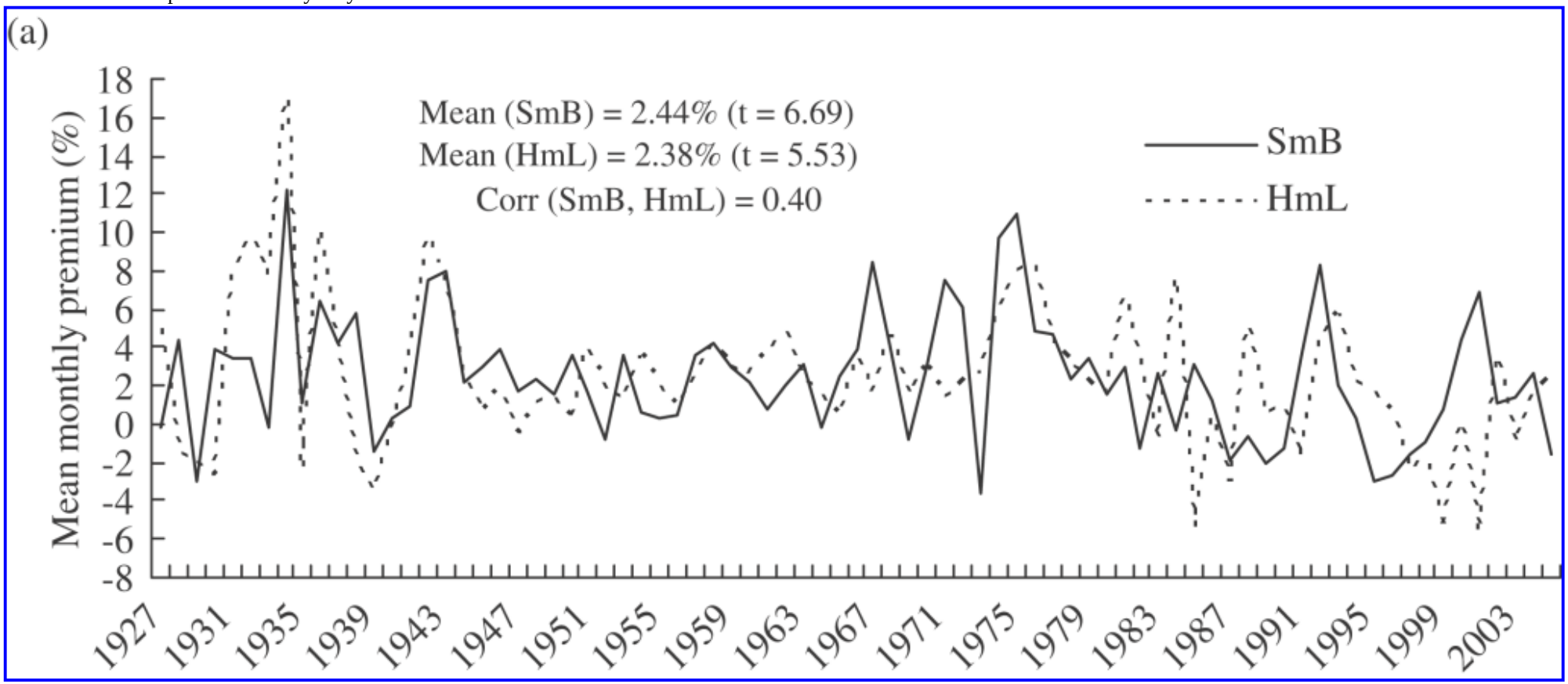

Figure 2B

The value and size premia – Feb to Dec

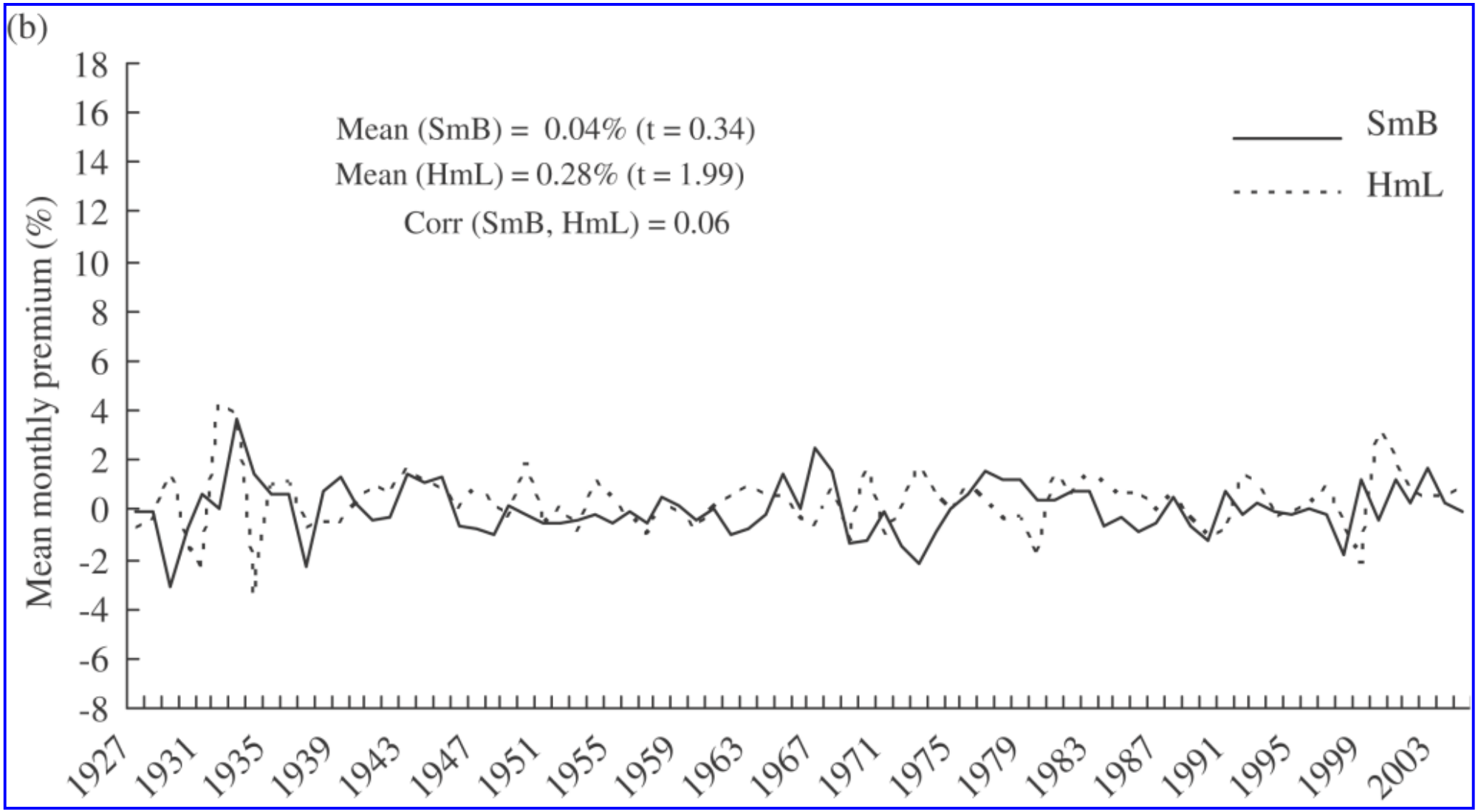

What explains the value and size effects? That both premia reflect some common element which manifests itself only in January is hard to reconcile with a risk compensation story. (Non-risk-based explanations of the January effect are discussed in the section on seasonal patterns in stock returns below.) Much recent research, nevertheless, has characterized the value premium as compensation for financial distress risk. Theoretical models have been developed in which such risk plays a central role, and value (high B/P) stocks accordingly earn higher equilibrium returns than growth (low B/P) stocks. Others have argued that the size effect is actually a liquidity effect in which small-cap stocks are less liquid than large-cap stocks and therefore provide correspondingly higher returns to offset the higher transactions costs (see, for example, Brennan, Chordia and Subrahmanyam, 1998). Still others have suggested that the size and B/P results may be due to survivor biases in the databases used by researchers (see, for example, Kothari, Shanken and Sloan, 1995).

One final hypotheses concerns measurement error in the estimated market betas used in the tests. Firms whose stocks have recently declined in price (for example, many high B/P and small-cap stocks), in the absence of a concomitant decline in the value of the debt, have become more leveraged and, other things equal, more risky in a beta sense. Traditional estimation methods produce ‘stale’ betas that underestimate ‘true’ beta risk for such firms. Thus, B/P and size may be viewed as better instruments for ‘true’ market beta risk than traditional estimates of beta, and the value and size effects are simply capturing the measurement error in the traditional beta estimates.

The prior return or momentum effect

Prior stock returns have been shown to have explanatory power in the cross section of common stock returns. Stocks with prices on an upward (downward) trajectory over a prior period of 3 to 12 months have a higher than expected probability of continuing on that upward (downward) trajectory over the subsequent 3 to 12 months. This temporal pattern in prices is referred to as momentum. Jegadeesh and Titman (1993) show that a strategy that simultaneously buys past winners and sells past losers generates significant abnormal returns over holding periods of 3 to 12 months. The abnormal profits generated by such offsetting long and short positions appear to be independent of market, size or value factors, and have persisted in the data for many years. To this end, Carhart (1997) estimates an extension of eq. (2) that includes a momentum factor (in addition to market, size and value factors) defined in the spirit of Jegadeesh and Titman as the difference in returns between a portfolio of ‘winners’ and a portfolio of ‘losers’. The coefficient on the momentum factor is positive and statistically significant, and cannot be explained by the other three factors. Finding a rational risk-related explanation for the momentum effect has proven difficult. A number of researchers have posited behavioural (psychology-based) explanations of momentum that rely on irrational market participants who underreact to news, but these models are hard to reconcile with psychology-based models of overreaction posited to explain the value premium (for example, Lakonishok, Shleifer and Vishney, 1994).

Time series return predictability

Consider a model of stock prices in which expected stock returns are constant through time (see Fama, 1976, for discussion of this model and related tests of the behaviour of stock prices). Much recent evidence suggests that expected returns are not constant, but contain a time-varying component that is predicted by past returns, *ex ante* observable variables and calendar turning points. The following subsections discuss this evidence.

Predicting returns with past returns I: individual security autocorrelations

Much research finds that autocorrelations of higher-frequency (daily, weekly) individual stock returns are negative and that the autocorrelations are inversely related to the market capitalization of the stock. The exception is that the largest market cap stocks have positive autocorrelations for daily returns. The inverse relation between individual return autocorrelations and market capitalization is due to the influence of a bid-ask bounce in high frequency stock prices that may induce ‘artificial’ serial dependencies into returns. Niederhoffer and Osborne (1966) find that successive trades tend to occur alternately at the bid and then the ask price, resulting in negative serial correlation in returns. This negative serial dependency is more pronounced for smaller stocks that have lower prices and, consequently, for which the bid-ask spread represents a larger percentage of price. Because of the high variance of individual stock returns, researchers find that past returns explain a trivial percentage of total return variability at high frequencies (typically less than one per cent). And the predictability at high frequencies is economically insignificant: profits from trading strategies attempting to exploit the predictability in individual stocks are indistinguishable from zero.

Predicting returns with past returns II: aggregate return autocorrelations

Because of variance reduction obtained from diversification, aggregated or portfolio returns provide more powerful tests of return predictability using past returns. However, this increased power may be offset by upward-biased autocorrelations caused by the infrequent trading of securities in the portfolios (Fisher, 1966). This bias is more serious for portfolios of smaller-cap stocks that contain less frequently traded stocks. In the United States and other global equity markets positive autocorrelations for high-frequency portfolio returns range from 0.4 for small-cap stocks to 0.1 for large-cap stocks. Research has shown, however, that positive portfolio autocorrelations are not due to infrequent trading of the securities in the portfolio. Indeed, many researchers have reported statistically significant positive portfolio autocorrelations for return frequencies up to one month in the United States, an interval over which virtually all securities will have traded. There is no evidence, however, of profitable trading opportunities based on daily, weekly or monthly aggregate return autocorrelations. (Lo and MacKinlay, 1990, reconcile the paradox of positive portfolio autocorrelations and negative individual stock autocorrelations: because the autocorrelation of portfolio returns is the sum of individual security autocovariances and cross-autocovariances, if the cross-autocovariances are sufficiently larger than the autocovariances – empirically, they are – then the cross-autocovariances will overshadow the contribution of the autocovariances.)

Significant predictability – both economically and statistically – has been identified in longer-horizon stock returns. As mentioned in the previous section, Jegadeesh and Titman (1993) identify profitable trading strategies based on past price momentum over 3- to 12-month intervals. De Bondt and Thaler (1985) find that New York Stock Exchange stocks identified as the biggest losers (winners) over a period of three to five years earn, on average, the highest (lowest) market-adjusted returns over a subsequent holding period of the same length of time, a phenomenon that does not seem to disappear when returns are adjusted for size and beta risk. This predictable reversal pattern is often attributed to market ‘overreaction’ in which stock prices diverge from fundamental values because of (irrational) waves of optimism or pessimism before returning eventually to fundamental values. Evidence of this longer-horizon return predictability has been reported in most equity markets around the world. But the significance of negative autocorrelation for long horizon returns is subject to the statistical problems discussed in the next subsection.

Predicting aggregate returns with predetermined observable variables

The evidence above shows that past returns contain information about expected returns, but they are a noisy signal. A more powerful test uses predetermined explanatory variables that potentially convey more precise information about expected returns. Much recent research documents such predictability using past information. An incomplete list of the variables in these studies includes expected inflation, yield spreads between long- and short-term interest rates and between low- and high-grade bonds, the dividend-to-price ratio, the earnings-to-price ratio, the book-to-price ratio, and the level of consumption relative to income. Importantly, predictability is stronger when the tests use returns measured over longer horizons, with explanatory power rising to levels of 20–40 per cent at two to four year horizons. Unfortunately, the increased explanatory power does not come without econometric problems. First, the number of independent observations decreases with the return horizon. To accommodate, researchers use overlapping observations, but the adjustments for standard errors to account for this perform poorly for the relatively small sample periods used in these tests. Second, most of the variables listed above are highly persistent (in contrast to lagged returns used in autocorrelation tests), and their innovations are correlated with return innovations, resulting in biased test statistics. Despite these shortcomings, the level of statistical significance and the robust nature of the results – across so many different explanatory variables and across so many worldwide equity markets – strongly argue for a predictable component in aggregate returns.

Patterns in daily returns around weekends

Consider an exchange where trading takes place Monday–Friday. If the process generating stock returns operates continuously, then Monday returns should be three times the returns expected on each of the other days to compensate for a three-day holding period. Call this the calendar-time hypothesis. An alternative is the trading-time hypothesis: returns are generated only during trading periods, and average returns are the same for each of the five trading days in the week. Inconsistent with both hypotheses, stock returns in many countries are negative, on average, on Monday (French, 1980). (In Australia, Korea, Japan and Singapore average returns on Tuesday are negative because of time zone differences relative to the US and European markets.)

What causes the weekend effect? That the pattern exists in so many different markets argues persuasively against many institution-specific explanations. Research has shown that the weekend effect cannot be explained by: differences in settlement periods for transactions occurring on different weekdays; measurement error in recorded prices; market maker trading activity; or systematic patterns in investor buying and selling behaviour. That an explanation has been elusive may not be important: in the post-1977 period in the United States and in numerous other markets, the weekend effect has all but disappeared (see Schwert, 2003).

Patterns in returns around the turn of the year

Keim (1983) and others document that 50 per cent of the annual size premium in the United States is concentrated in the month of January, particularly in the first week of the year. This finding has been reproduced on many equity markets throughout the world. Blume and Stambaugh (1983) subsequently demonstrated that, after an upward bias in average returns for small stocks (related to the magnitude of bid-ask spreads) had been corrected, the size premium is evident only in January.

What explains this phenomenon? Two hypotheses rely on the buying and selling behaviour of market participants to explain the turn-of-the-year size premium. The first hypothesis attributes the effect to year-end tax-related selling by taxable individual investors of stocks that have declined in price (an attribute shared by many small-cap stocks). In such trades the investor realizes a capital loss which can be used to offset realized capital gains, thereby reducing taxable income. There is much evidence that such tax-related trading occurs at the end of the tax year (which in many countries coincides with the end of the calendar year), but a clear link between such trading and stock return behaviour has not been established. A second hypothesis concerns the impact of institutional ‘window dressing’ at the end of the calendar year – selling off ‘loser’ stocks that have declined in price (again, typically small-cap stocks) so they don't appear on year-end statements sent to constituent shareholders. Although there is evidence that institutions behave in this fashion, any resulting impact on stock prices is difficult to distinguish from the impact of tax-loss selling. In the end, large bid-ask spreads and high transaction costs for small-cap stocks preclude the profitable exploitation of the short-term return differences between individual small- and large-cap stocks. As a result, the turn-of-the-year size premium continues to be positive in recent years (see Figure 2A).

Conclusion

Recent research in finance has revealed stock price behaviour that is inconsistent with the predictions of familiar models. The research on time series predictability, as a whole, is convincing evidence that expected returns are not constant through time. There are reasonable business conditions stories that can account for time variation in expected returns. However, some of the temporal patterns in returns – in particular those relating to calendar turning points – are troubling as they defy economic interpretations.

The evidence on cross-sectional anomalies poses a significant challenge to well-established asset pricing paradigms. Yet, despite mounting evidence, there is little consensus on alternative theoretical models. As such, the focus of future research should be on developing such models. Indeed, one of the most significant contributions of this strand of research has been the recognition of potential alternative sources of risk (for example, risk related to financial distress) and of the potential importance of behavioural models. Importantly, researchers must recognize that the existence of this anomalous evidence does not constitute proof that existing paradigms are 'wrong'. There is the issue of data snooping – much of the empirical research on financial market anomalies is predicated on previous research that documented similar findings with the same data. And although many of these effects have persisted for nearly 100 years, this in no way guarantees their persistence in the future. More research is necessary to resolve these issues.

See Also

- capital asset pricing model

Bibliography

Ball, R. 1978. Anomalies in relationships between securities' yields and yield-surrogates. *Journal of Financial Economics* 6, 103–26.

Banz, R. 1981. The relationship between return and market value of common stock. *Journal of Financial Economics* 9, 3–18.

Basu, S. 1977. Investment performance of common stocks in relation to their price-earnings ratio: a test of the efficient market hypothesis. *Journal of Finance* 32, 663–82.

Blume, M. and Stambaugh, R. 1983. Biases in computed returns: an application to the size effect. *Journal of Financial Economics* 12, 387–404.

Brennan, M. 1970. Taxes, market valuation, and corporate financial policy. *National Tax Journal* 23, 417–27.

Brennan, M., Chordia, T. and Subrahmanyam, A. 1998. Alternative actor specifications security characteristic, and the cross section of stock returns. *Journal of Financial Economics* 49, 345–73.

Carhart, M. 1997. On the persistence in mutual fund performance. *Journal of Finance* 52, 57–82.

Daniel, K. and Titman, S. 1997. Evidence on the characteristics of cross-sectional variation in stock returns. *Journal of Finance* 52, 1–33.

De Bondt, W. and Thaler, R. 1985. Does the stock market overreact? *Journal of Finance* 40, 793–805.

De Bondt, W. and Thaler, R. 1987. Further evidence on investor overreactions and stock market seasonality. *Journal of Finance* 42, 557–81.

Fama, E. 1976. *Foundations of Finance*. New York: Basic Books.

Fama, E. and French, K. 1992. The cross section of expected stock returns. *Journal of Finance* 47, 427–66.

Fama, E. and French, K. 1993. Common risk factors in the returns of stocks and bonds. *Journal of Financial Economics* 33, 3–56.

Fisher, L. 1966. Some new stock-market indices. *Journal of Business* 39, 191–225.

French, K. 1980. Stock returns and the weekend effect. *Journal of Financial Economics* 8, 55–69.

Graham, B. and Dodd, D. 1940. *Security Analysis: Principles and Technique*. New York: McGraw-Hill Book Company, Inc.

Hawawini, G. and Keim, D. 2000. The cross section of common stock returns: a review of the evidence and some new findings. In *Security Market Imperfections in Worldwide Equity Markets*, ed. D. Keim and W. Ziemba. Cambridge: Cambridge University Press.

Jaffe, J., Keim, D. and Westerfield, R. 1989. Earnings yields, market values and stock returns. *Journal of Finance* 45, 135–48.

Jegadeesh, N. and Titman, S. 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance* 48, 65–92.

Keim, D. 1983. Size-related anomalies and stock return seasonality: further empirical evidence. *Journal of Financial Economics* 12, 13–32.

Kothari, S., Shanken, J. and Sloan, R. 1995. Another look at the cross-section of expected stock returns. *Journal of Finance* 50, 185–224.

Kuhn, T. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakonishok, J., Schleifer, A. and Vishny, R. 1994. Contrarian investment, extrapolation and risk. *Journal of Finance* 49, 1541–78.

Litzenberger, R. and Ramaswamy, K. 1979. The effects of personal taxes and dividends on capital asset prices: theory and empirical evidence. *Journal of Financial Economics* 7, 163–95.

Lo, A. and MacKinlay, C. 1990. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* 3, 175–205.

Mehra, R. and Prescott, E. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Miller, M. and Scholes, M. 1982. Dividend and taxes: some empirical evidence. *Journal of Political Economy* 90, 1118–41.

Neiderhoffer, V. and Osborne, M. 1966. Market making and reversal on the stock exchange. *Journal of the American Statistical Association* 61, 897–916.

Reinganum, M. 1981. A misspecification of capital asset pricing: empirical anomalies based on earnings yields and market values. *Journal of Financial Economics* 9, 19–46.

Rosenberg, B., Reid, K. and Lanstein, R. 1985. Persuasive evidence of market inefficiency. *Journal of Portfolio Management* 11, 9–17.

Schwert, G. 2003. Anomalies and market efficiency. In *Handbook of the Economics of Finance*, vol. 1, ed. G. Constantinides, M. Harris and R. Stulz. Amsterdam: North Holland.

How to cite this article

Keim, Donald B. "financial market anomalies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000308> doi:10.1057/9780230226203.0570

The New Palgrave Dictionary of Economics Online

financial market contagion

Morgan Kelly
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The power of the metaphor of contagion – that beliefs, actions, and strategies spread among economic agents like pathogens among biological organisms – causes it to recur in disparate areas of economics. This article focuses on four applications of contagion to economics: social influence or memoryless learning; Bayesian social learning; strategy choice in coordination games; and the spread of crises in international financial markets.

Keywords

arbitrage; Bayesian learning; competitive exclusion; contagion; coordination games; ergodic theorem of Markov chains; financial market contagion; information cascades; overconfidence; popularity weighting; portfolio allocation; probability; social influence

Article

Social influence

The metaphor of contagion is central to the early studies of crowd psychology of Mackay (1841), Tarde (1900) and LeBon (1895); and classical early models of disease diffusion were applied to financial markets by Shiller (1984).

The modern analysis of social influence starts with Allport and Postman (1946–47) who studied the spread of wartime rumour. They identified four circumstances that facilitate the spread of rumour: two are characteristics of the rumour, two of the population. The topic of the rumour should be important to people and the rumour should be hard to verify individually; while individuals should be credulous, and going through a time of unusual stress.

Motivations for neglecting formal Bayesian learning differ between economics and sociology. Sociology emphasizes situations that do not lend themselves to Bayesian updating either through lack of time (is a bank about to fail?), or the nature of the question (what is the one true religion?). Economics, by contrast, emphasizes computational simplicity: rules of thumb make fewer cognitive demands on agents than formal

updating algorithms.

Kirman (1993) analyses a simple model of influence that is motivated by the foraging behaviour of ants, but applicable, he argues, to the behaviour of stock market investors. Faced with a choice between two identical piles of food, ants switch periodically from one pile to the other. Kirman supposes that there are *N* ants and that each switches randomly between piles with probability ε (this prevents the system getting stuck with all at one pile or the other), and imitates a randomly chosen other ant with probability δ.

By the ergodic theorem of Markov chains, there is a unique steady state distribution of ants between piles, and Kirman shows by simulation that the shape of the distribution depends on the relative magnitudes of the imitation parameter δ and the mutation parameter ε. With weak imitation and strong mutation there is a single peak at $\frac{1}{2}$, with equal numbers of ants at each pile. With stronger imitation and weaker mutation, the steady state distribution has two peaks at 0 and *N*: most ants concentrate on a single pile and switch periodically to the other – the behaviour observed among real ants and possibly stock market participants. In contrast to Bayesian learning models, the absence of martingale convergence allows society continually to flip between beliefs.

The independent work of Weidlich and Haag (1983) in quantitative sociology presents an analogous model in continuous time. Agents switch states with a logistic probability that again depends on the relative social popularity of each choice, but Weidlich and Haag also allow agents to have a personal preference for one of the choices. Again, for sufficiently strong imitative behaviour there is a steady state distribution with two peaks, but now the relative magnitude of the peaks depends on how much agents prefer each choice. Society spends most time at the choice preferred by each agent, but will spend time at the choice that is less popular with everyone, as a consequence of social influence.

Ellison and Fudenberg (1993) look at the role of popularity weighting in choosing between a superior and an inferior technology. They observe that popularity can be a useful summary of the relative past performances of the two technologies – the better technology should be more popular – but that the amount of information conveyed by popularity is diluted the more people rely on it. They therefore look at the likelihood that the better technology is adopted, allowing a fixed fraction of the population to change its choice each period, when the relative weights put on the popularity of the technology versus its performance in the last period are allowed to vary.

Ellison and Fudenberg (1993) show that there is an optimal popularity weighting that causes the system to converge to everyone's using the better technology. If popularity weighting exceeds this optimum, the system converges to a steady state where everyone uses one technology, but which technology depends on the starting number of users of each. With under-weighing of popularity, the inefficient alternative can survive indefinitely.

The competitive exclusion principle, proven in the context of ordinary differential equations by Levin (1970), states that the number of coexisting species cannot exceed the number of resources they compete for. Here there are two competing species or technologies competing for one resource, being used by people, so if the technological choice problem is recast as one of biological competition we know that only one technology will survive. This is done by Juang (2001), who uses an evolutionary selection argument to show how an Ellison–Fudenberg society can reach the optimum when different groups of agents have sufficiently different popularity weightings. In periods when the inferior technology is excessively popular, agents putting low weight on popularity receive higher payoffs and increase in number, while agents who put high weight on popularity do better in periods when the superior technology is popular.

In the popularity weighting models of Kirman (1993), Weidlich and Haag (1983) and Ellison and Fudenberg (1993), every person is equally influenced by every other member of society. In many situations however, we

are influenced more by individuals whom we know and have learned to trust than by strangers. To model the greater social influence of neighbours, the individual is put into some mathematical space, where he or she is more likely to interact with individuals close by than far away. Durlauf (1997) looks at the behaviour of agents in an Ising model (originally developed to model the flipping of magnetic poles of atoms in a crystal) where agents live on a lattice and change between two actions at a rate that depends logistically on the state of their nearest neighbours.

If the influence of neighbours lies above a critical value, the system has two steady state distributions (there are an infinite number of agents so the ergodic theorem of Markov chains does not apply) with all agents either in one state or the other. If agents have a preference for one state over the other (the physical analogue is an external magnetic field) however, the system has only one steady state with all choosing the preferred action.

In Durlauf's model, agents in each state influence each other symmetrically, affecting only their nearest neighbours. Durrett and Levin (1998) analyse a system where agents of different types can affect others over different distances. While biologically motivated – Durrett and Levin (1998) are interested in how slow-growing trees can out-compete rapidly growing grasses – this analysis suggests how propaganda and advertising can be used to cause bad ideas to drive out good ones.

Suppose that type 0 dominates type 1: an agent of type 0 converts a type 1 neighbour at rate 1, whereas a type 1 agent converts a type 0 only at rate $\delta<1$. If both types have the same radius of influence then, so long as the dominant type 0 avoids getting wiped out by an unlucky run at the start, it will take over. However, Durrett and Levin (1998) show that if the dominant type affects only neighbours in a radius of 1, whereas the dominated type affects neighbours over a large radius *R*, there is a critical value of the conversion rate $\delta_c<1$, above which the dominated type 1 takes over.

It is straightforward to demonstrate the existence of social influence empirically when individuals observe the overall popularity in society rather than among neighbours. The influence of best-seller lists on book buying is sufficiently well known for publishers to seek to manipulate them by buying books in stores known to be tracked by the lists, and a variety of examples of imitative behaviour are given by Bikchandani, Hirshleifer and Welch (1992) and Chamley (2004, pp. 59–60).

Testing for the influence of neighbours is more difficult because neighbourhood choice is frequently endogenous: one must make sure that the behaviour one is attributing to the influence of neighbours is not due to some individual factor that led the person to choose this neighbourhood over others in the first place.

The classic Ryan and Gross (1943) study, which found that the main factor influencing farmers to adopt hybrid corn was the number of nearby farmers who had adopted it, passes the exogeneity test: it is unlikely that farmers chose farms in order to be near other innovative farmers. Sacerdote (2001) uses the random allocation of roommates to incoming Dartmouth University students to show how roommates influence each others' behaviour, finding that roommates have an effect on individual academic performance, while dormitory effects influence decisions to join fraternities. Kelly and O Grada (2000) look at the behaviour of Irish immigrants, mostly housemaids and day labourers, in 1850s New York during two bank runs. Since they are immigrants it is possible to identify their social network from their place of origin in their home country: newly arrived immigrants tend to associate with people they knew at home. Kelly and O Grada (2000) found that immigrants from one set of counties in Ireland tended to close their accounts during the panics, while otherwise identical immigrants from other counties stayed put.

Bayesian learning

Bayesian models of social learning allow individuals to infer the information of other agents from their observed actions in an optimal manner rather than through ad hoc imitation. Bayesian social learning can exhibit pathologies. After the first few agents have chosen, subsequent actions convey little new information and are dominated by idiosyncratic noise. Society converges slowly to the optimal action and, in some circumstances, may become stuck on the suboptimal action. A useful textbook discussion of the literature is given by Chamley (2004).

In Bikchandani, Hirshleifer and Welch (1992) and Banerjee (1992), the world can be in either state σ_0 or σ_1. Each agent receives a signal s_0 or s_1 with symmetric precisions $P(s_0|\sigma_0)=P(s_1|\sigma_1)=p$ and must choose whether or not to invest. Agents choose in a fixed order and, before receiving his private signal, the agent investing in period t observes the history of past investments and uses this to determine their prior probability π_{It} that the state is 1.

Bikchandani, Hirshleifer and Welch (1992) start with the case where the cost of investment is $\frac{1}{2}$, the payoff in state 1 is 1, and 0 otherwise. Their expected payoff is $p\pi_{It}/(p\pi_{It}+(1-p)(1-\pi_{It}))$. After a number of moves there will be a sufficient difference between the number who has invested and those who have not for the agent's action to be determined solely by his prior belief, irrespective of his signal. Specifically, if the first agent gets a good signal, the second invests if he gets a good signal, and all subsequent agents will then invest irrespective of their signals. If the second gets a bad signal he is indifferent about investing and is assumed to invest, so the third investor again invests regardless of signal, and so on. Once there are two more investors than non-investors, the excess of positive signals outweighs any negative signal an agent might have. Everyone invests regardless of signal, leading to a cascade.

An unlucky series of wrong signals at the start of the game can lead society to fix on the wrong equilibrium. Bikchandani, Hirshleifer and Welch (1992) observe that this wrong equilibrium is fragile, being based on the observations of a handful of early agents, and vulnerable to being overturned by public information available to all agents.

A frequent criticism of cascade models is their reliance on finite signals: all signals are equal and there is no way for a huge negative signal to counteract a series of positive ones. However, the important lesson of the cascade literature is not that society can get stuck at the wrong equilibrium – which requires signals that are finite – but that Bayesian learning when individual signals are observed imperfectly is very slow to converge to the true equilibrium. Vives (1993) shows how adding noise to a Gaussian model slows down its convergence from rate t to rate $t^{1/3}$: 1,000 noisy observations are equivalent to ten clean ones.

The basic intuition of cascades models that imperfectly observed individual information is poorly incorporated into social beliefs is the basis of several other models. Bulow and Klemperer (1994) model rational frenzies in auctions where participants reveal their valuations by bidding. Bidders with high valuations are willing to pay just under the Walrasian clearing price and, being usually inframarginal, all face similar optimization problems. A bid by one agent therefore sets off a chain of bidding by other agents, leading to a pattern of booms and crashes. Caplin and Leahy (1994) look at investment where individuals have Gaussian signals. If the true state is bad, individuals continue to invest, driven by the dominating effect of past actions. Eventually, however, because signals are Gaussian, a few agents get sufficiently bad signals to induce them to stop investing, causing priors rapidly to move to a belief that the state is bad, leading to a market crash and ‘wisdom after the fact’.

While the essence of the cascade literature is that agents transmit a noisy signal of their information, Avery and Zemsky (1998) observe that this is not the case for markets obeying the efficient markets hypothesis where price reflects all publicly available information. In such markets, assuming risk-neutral agents, the price

of an asset worth 1 in the good state and 0 in the bad is the Bayesian prior π_{It}, causing agents always to trade according to their private signal. They show that cascades can still occur if extra dimensions of uncertainty are added – specifically if there is event uncertainty (agents know that something important has happened by whether it is good or bad), or compositional uncertainty (agents are uncertain how many informed traders are active in the market).

Underlying Bayesian models of cascades is the obvious but strong assumption that people are Bayesians. Probability is difficult for most people, and conditional probability especially so. Even with trivial problems of the form ‘a family has two children, one of whom is a daughter: what is the probability that the other child is a son?’ most will incorrectly answer $\frac{1}{2}$ rather than $\frac{2}{3}$. Similarly, when asked ‘one per cent of the population has a disease. A test detects the disease in 95 per cent of patients when it is present, and generates ten per cent false positives when it is absent. What is the probability that someone who tests positive has the disease?’, most will give answers slightly below 95 per cent rather than the correct 1.05 per cent.

In other words, people appear to ignore base rates, assuming that the probability of a state given a signal equals the probability of the signal given the state $P(\sigma_i|s_i) = P(s_i|\sigma_i)$ even when the probability of the state is considerably lower than the probability of the signal. Agents show overconfidence, focusing excessively on their own signal rather than the history of signals of other agents contained in the prior.

If people neglect priors in this way, cascades cannot occur when private signals are uncorrelated. However, if the signal is common, cascades can still occur. For instance if agents view market price as the signal, a run of rising prices induced by improving fundamentals (such as the good macroeconomic conditions and loose credit that Kindleberger (1978) saw as the preconditions for speculative bubbles) are treated by agents as a positive signal inducing them to buy, driving up price and inducing others to buy, and so on.

Strategies in coordination games

Kandori, Mailath and Rob (1993) considered the strategies of players in a coordination game with payoffs

	L	R
L	a, a	b, c
R	c, b	d, d

where $a>c$, $d>b$ and $(a-c)>(d-b)$ so (L, L) and (R, R) are Nash equilibria and (L, L) is the risk dominant one. With myopic, best-response play, they show that a small probability of mutation suffices for the risk dominant equilibrium to be chosen.

Ellison (1993) observed that this convergence is slow, requiring many simultaneous mutations, and showed instead that if there is local interaction of players along a line, the $\frac{1}{2}$-dominant strategy (the best response if half your neighbours adopt it) spreads rapidly, but not in two dimensions. Blume (1995) shows that non-trivial mixed long run equilibria exist in two dimensional interaction but not in one, while Morris (2000) examines the characteristics of arbitrary networks that permit the risk-dominant strategy to spread. Lee and Valentinyi (2000) look at a game without mutation but where initial strategy choice is random and show that myopic best response to strategies played by immediate neighbours on the lattice causes large populations to coordinate on the risk dominant equilibrium.

International market contagion

Large falls in asset values in one country are sometimes followed rapidly by falls in other countries. To the extent that these falls are too great to be explained by interdependence in trade or exposure to common macroeconomic factors, the process is called contagion.

Two main sources of contagion have been proposed: financial fragility and common financial linkages; and pathologies in the diffusion of information. The empirical study of Kaminsky, Reinhart and Vegh (2003) argues that three sources of fragility underlie international contagion: rapid inflows of capital; macroeconomic shocks that occur too rapidly for gradual portfolio rebalancing; and a leveraged common creditor. Allen and Gale (2000) show that if banks in different regions have claims on each other, a fall in asset values in one

region can bring banks in other regions under pressure and lead to falls in asset values in those regions. In Kyle and Xiong (2001) losses suffered by traders who arbitrage between markets dominated by fundamentalists and markets dominated by noise traders cause traders to reduce their positions in both markets, leading to returns becoming more volatile and more correlated.

Models of contagion as information transmission abstract away from agents who revise excessively optimistic forecasts of returns in all markets after a fall in one market, and concentrate on rational actors instead. Calvo and Mendoza (2000) show that if there are fixed costs to gathering and processing information specific to one country and limits to short selling in each country, the benefits of acquiring information about each country in one's portfolio fall as the portfolio expands. Agents put more weight on the behaviour of other investors, making portfolio allocation more sensitive to realized returns in each market. In Kodres and Pritsker (2002), portfolio rebalancing by informed investors can set off panics among the uninformed who misinterpret it as negative information about the market.

The empirical literature on testing for contagion has focused on increases in the correlation of returns between markets during periods of crisis. Forbes and Rigobon (2002) show the elementary weakness of simple correlation tests: with an unchanged regression coefficient, a rise in the variance of the explanatory variable reduces the coefficient standard error, causing a rise in the correlation of a regression.

The regression underlying contagion tests is of the form

$$y_{it} = \delta_i' \mathbf{z}_t + \alpha_i' \mathbf{x}_{it} + \beta_i I(y_j - c_j) + \varepsilon_{it}$$

where y_i is asset return in country i, $\mathbf{z}$ are common macroeconomic factors, $\mathbf{x}_i$ are country specific factors, and I is an indicator of a period of crisis in the originating economy j. As Pesaran and Pick (2007) observe, this is a difficult system to estimate econometrically. To disentangle contagion from interaction effects, county-specific variables have to be used to instrument foreign returns. Choosing the crisis period introduces sample selection bias, and it has to be assumed that crisis periods are sufficiently long to allow correlations to be reliably estimated. In consequence, there appears to be no strong consensus in the empirical literature as to whether contagion occurs between markets, or how strong it is.

See Also

- information cascades

Bibliography

Allen, F. and Gale, D.M. 2000. Financial contagion. *Journal of Political Economy* 108, 1–33.

Allport, G.W. and Postman, L. 1946–47. An analysis of rumor. *Public Opinion Quarterly* 10, 501–17.

Avery, C. and Zemsky, P. 1998. Multidimensional uncertainty and herd behavior in financial markets. *American Economic Review* 88, 724–48.

Banerjee, A.V. 1992. A simple model of herd behaviour. *Quarterly Journal of Economics* 107, 797–817.

Bikchandani, S., Hirshleifer, D. and Welch, I. 1992. A theory of fads, custom, and cultural change as informational cascades. *Journal of Political Economy* 100, 992–1026.

Blume, L. 1995. The statisticial mechanics of best-response strategy revision. *Games and Economic Behavior* 11, 111–45.

Bulow, J. and Klemperer, P. 1994. Rational frenzies and crashes. *Journal of Political Economy* 102, 1–23.

Calvo, G.A. and Mendoza, E.G. 2000. Rational contagion and the globalization of security markets. *Journal of International Economics* 51, 79–113.

Caplin, A. and Leahy, J. 1994. Business as usual, market crashes, and wisdom after the fact. *American Economic Review* 84, 548–65.

Chamley, C. 2004. *Rational Herds: Economic Models of Social Learning*. Cambridge: Cambridge University Press.

Durlauf, S.N. 1997. Statistical mechanics approaches to socioeconomic behavior. In *The Economy as an Evolving Complex System II*, ed. B. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison-Wesley.

Durrett, R. and Levin, S. 1998. Spatial aspects of interspecific competition. *Theoretical Population Biology* 53, 30–43.

Ellison, G. 1993. Learning, local interaction, and coordination. *Econometrica* 61, 1047–71.

Ellison, G. and Fudenberg, D. 1993. Rules of thumb for social learning. *Journal of Political Economy* 101, 612–43.

Forbes, K.J. and Rigobon, R. 2002. No contagion, only interdependence: measuring stock market comovements. *Journal of Finance* 57, 2223–61.

Juang, W.-T. 2001. Learning from popularity. *Econometrica* 69, 735–47.

Kaminsky, G.L., Reinhart, C.M. and Vegh, C.A. 2003. The unholy trinity of financial contagion. *Journal of Economic Perspectives* 17(4), 51–74.

Kandori, M., Mailath, G.J. and Rob, R. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61, 29–56.

Kelly, M. and O Grada, C. 2000. Market contagion: evidence from the panics of 1854 and 1857. *American Economic Review* 90, 1110–25.

Kindleberger, C. 1978. *Manias, Panics, and Crashes*. New York: Basic Books.

Kirman, A. 1993. Ants, rationality, and recruitment. *Quarterly Journal of Economics* 108, 137–56.

Kodres, L. and Pritsker, M. 2002. A rational expectations model of financial contagion. *Journal of Finance* 57, 769–800.

Kyle, A. and Xiong, W. 2001. Contagion as a wealth effect. *Journal of Finance* 56, 1401–40.

LeBon, G. 1895. *Psychologie des Foules*. Translated as *The Crowd: A Study of the Popular Mind*, New York: Viking, 1960.

Lee, I.H. and Valentinyi, A. 2000. Noisy contagion without mutation. *Review of Economic Studies* 67, 47–56.

Levin, S.A. 1970. Community equilibria and stability, and an extension of the competitive exclusion principle. *American Naturalist* 104, 413–23.

Mackay, C. 1841. *Extraordinary Popular Delusions and the Madness of Crowds*. London: Bentley.

Morris, S. 2000. Contagion. *Review of Economic Studies* 67, 57–78.

Pesaran, M.H. and Pick, A. 2007. Econometric issues in the analysis of contagion. *Journal of Economic Dynamics and Control* 31, 1245–77.

Ryan, B. and Gross, N.C. 1943. The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* 8, 15–24.

Sacerdote, B. 2001. Peer effects with random assignment: results for Dartmouth roommates. *Quarterly Journal of Economics* 116, 681–704.

Shiller, R.J. 1984. Stock prices and social dynamics. *Brookings Papers on Economic Activity* 1984(2), 457–98.

Tarde, G. 1900. *Les Lois de l'Imitation*. Translated as *The Laws of Imitation*, Gloucester, MA: Peter Smith, 1962.

Vives, X. 1993. How fast do agents learn? *Review of Economic Studies* 60, 329–47.

Weidlich, W. and Haag, G. 1983. *Concepts and Models of Quantitative Sociology*. New York: Springer.

How to cite this article

Kelly, Morgan. "financial market contagion." The New Palgrave Dictionary of Economics. Second Edition.

Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000319> doi:10.1057/9780230226203.0571

The New Palgrave Dictionary of Economics Online

financial structure and economic development

John Toye

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Most early development economists neglected the financial aspects of development, often restricting them to domestic taxation, the self-finance of enterprises and the negotiation of foreign credits. In the 1970s, a few economists proposed that private financial intermediation, operating with market-set interest rates, improved incentives to save and the availability of credit, and allocated savings more efficiently between borrowers. Against this, new institutional economists have argued that financial intermediation involves considerable risks since banks find it difficult to acquire skills in risk assessment. The relationship between increases in real income and the size and complexity of the financial superstructure remains loose.

Keywords

adverse selection; asymmetric information; capital controls; credit rationing; development banks; financial intermediaries; financial intermediation; financial interrelations ratio; financial liberalization; financial repression; financial structure and economic development; Goldsmith, R.; interest-rate liberalization; Kalecki, M.; microcredit; moneylenders in developing countries; moral hazard; new institutional economists; prudential regulation; reserve ratio requirement; risk assessment

Article

The question of how financial structure relates to economic development departs from a distinction between an economy's infrastructure of real wealth – its physical assets produced by human labour and natural resources – and a set of financial claims that exists side by side with it and is somehow connected with it. This set of claims consists of short-term and long-term loans and credits and equity securities. A second distinction is between two types of issuers and holders of these financial instruments: non-financial institutions, such as governments, business enterprises and households, whose assets are mainly – but not exclusively – held in physical form, and financial institutions, whose assets

and liabilities are mainly financial instruments. This second distinction divides the original question into two parts: the link between the real infrastructure and the volume of financial instruments in the economy, and the link between the real infrastructure and the volume of funds held by the financial institutions.

The US economist Raymond W. Goldsmith (1904–88) provided much of the statistical framework and empirical basis for the examination of these questions. In a lifetime of painstaking scholarly labour, he collected data that allowed comparison across countries of key ratios of real and financial assets, and also of how these ratios changed within individual countries over time. His research suggested one – now widely accepted – statistical generalization, that of the *rising financial interrelations ratio*. As expressed by Gurley and Shaw (1967, p. 257), ‘during economic development … countries usually experience more rapid growth in financial assets than in national wealth and national products’. The increase, however, does not continue without limit. This process of financial deepening has been experienced by many of the now developed capitalist countries. However, it tends to be most evident in the early and middle phases of their economic development, after which it levels off. The exceptions to this are higher ratios in periods of repressed inflations during and just after major wars. In the United States, little financial deepening has been noticeable since 1950.

It is also clear that this ratio can be influenced by strategic choices in the quest for development. Countries that adopt state-led development strategies, such as the USSR and its eastern European satellites, exhibited smaller ratios than those of countries that relied on private sector growth to drive their economic development.

As one would expect, developing countries have much lower financial interrelations ratios (between 0.6 and 1.0) than do Europe and North America (between 1.0 and 1.5). This is a reflection of the lower degree of monetization of their economies and the relative lack of separation of the functions of saving and investment. The composition of the value of total financial instruments shows a smaller share of financial institutions in the developing countries, for the same reason (Goldsmith, 1969, pp. 44–7).

Compositional changes in financial instruments

The process of development is associated with compositional change, as well as growth, in the stock of financial instruments. The start of financial development is the diffusion of fiduciary money into the economy through the banking system. This is followed by the growth of banking deposits, and then the share of banking deposits declines as new types of financial institution proliferate – building societies, mortgage companies, insurance companies and pension funds – providing financial services that are tailored to special needs.

The main thrust of early development economics neglected these financial aspects of development. Until the 1980s, the main focus of analysis was on the real economy, particularly the accumulation of real physical capital, the acquisition of new human skills and the expansion of international trade. When the problem of financing ‘real’ development was discussed, it was often restricted to the problem of domestic taxation, the self-finance of enterprises and the negotiation of foreign credits. Michal Kalecki, who greatly influenced the early development literature, explicitly argued that the volume of investment is not subject to financial limits. In the Kaleckian view, financial institutions appear, if at all, as pre-existing constraints on production that have to be removed – for example, rural moneylenders (Kalecki,

1972, pp. 145–61; FitzGerald, 1990, p. 184) – or as publicly established agencies for channelling resources to sectors of the economy that were desired to expand (see Eshag, 1983, pp. 186–8 on development banks). Thus financial development was long a secondary consideration, and it was viewed from the perspective of a government deciding which monetary institutions to create and which to destroy.

The McKinnon – Shaw view of financial intermediation

Independently of this dominant post-Keynesian approach, Ronald McKinnon, John Gurley and Ed Shaw in the 1970s elaborated a much more positive view of the role of the growth of private financial intermediaries in development. Using the context of an agricultural sector exhibiting strong technological dualism and lumpy investment, they argued that private financial intermediation, operating with market-set interest rates, improved incentives to save and the availability of credit. It did so by spreading risks and transforming the maturity structure of debt in ways more attractive to both savers and borrowers. Their claim was that financial intermediation would provide the benefits of additional savings and the more efficient allocation of those increased savings between borrowers. On this account, the growth of financial intermediation promotes both capital accumulation and the diffusion of technical progress by spreading risks more widely and in conformity with people's willingness to bear them (Gurley and Shaw, 1960; 1967; McKinnon, 1973; Shaw, 1973).

This positive view of private financial intermediaries has been criticized by post-Keynesians on both theoretical grounds – an incomplete accounting of all the incentive effects of market-determined interest rates – and empirical grounds – the absence of the predicted incentive effects in the savings, investment and interest rate data. However, the most compelling theoretical critique came from the new institutional economists. Rejecting the assumption of perfect information, they showed how various information asymmetries between the knowledge possessed by the private bankers and the knowledge possessed by their clients (savers and investors) generated a radically altered assessment of the potential benefits and dangers of private financial intermediation.

An important conclusion was that, as interest rates rose, the banks’ lending portfolios became riskier. This resulted both from adverse selection, as the marginal borrowers are more liable to default, and from moral hazard, as the marginal borrowers are more likely to invest in high-risk projects. Private banks thus have an incentive to continue to lend at less than the market-clearing rate of interest, and to borrow from depositors at an even lower rate, and then to ration credit (Stiglitz and Weiss, 1981). The benefits to be expected from private financial intermediation under asymmetric information assumptions are smaller than those derived from McKinnon–Shaw perfect information reasoning.

Evidence of financial repression

In the 1980s it became increasingly clear that many existing financial institutions in developing countries were dysfunctional. Moreover, in many cases the cause of the dysfunction was diagnosed as inappropriate government regulation. The analysis of ‘financial repression’ in Shaw (1973) was often borne out in reality. Interest rates were administered and maximum rates held very low, while reserve ratio requirements were set very high to force banks to buy and hold government debt issued at below-

market rates of interest. Banks were treated as a source of government finance, rather than providers of financial services to the private sector. Meanwhile capital controls were in place to stop the flight of private capital seeking better returns abroad.

The consequences of these widespread interventions included banks' inability to offer attractive rates to depositors; an artificially low level of deposits; a shortage of credit; the rationing of available credit; political pressures directing the allocation of credit; low repayment rates; the accumulation of bad debts; and ultimately the effective insolvency of the banks. The flourishing business of rural moneylenders and 'kerb' markets, despite the charging extortionate rates of interest, was simultaneously observed, with puzzlement, complaint or cynicism, as the observer preferred.

The move to financial liberalization

The policy response was that international organizations and national aid donors pressed for the removal of these policy-induced distortions, and the adoption of reforms aimed at financial liberalization. Interest-rate liberalization was adopted as one of the components of structural adjustment programmes in the 1980s and 1990s. Unfortunately, liberalization was not enough by itself to end the effects of financial repression. While deposits did climb as a share of GNP, there was little expansion of credit to the private sector, as many state banks remained in existence, and their habits of directing credit died hard (World Bank, 2005, pp. 207–39). Worse still, financial liberalization led to increasingly frequent financial crises. They were the result of increased competition between private banks, increased opportunities of foreign borrowing for all banks and the serious inadequacy of prudential regulation of the banking system.

As the new institutional economists had pointed out, financial intermediation involves considerable risks, and banks find it difficult to acquire skills in risk assessment, especially when that skill has not been previously salient. Hence, the possibility of miscalculation of risk is ever present. In addition, bank regulation and supervision is itself a risky business, for the by now familiar reason of asymmetric information, and can provoke banking problems as well as prevent them – and can also be vulnerable to corrupt pressures to look the other way. All of this suggests that the building of functional financial sectors is likely to remain a work in progress for the foreseeable future.

Economic and financial development: a loose reciprocal relationship

Few would be inclined to deny that there is a rough parallel between economic and financial development, if periods of several decades are the time period under consideration. As real income and wealth increase, so do the size and complexity of the financial superstructure. Yet this is a loose relationship. The *financial intermediation ratio*, the share of financial institutions' assets in the value of all financial assets, is even more loosely tied to the stock of real wealth. Rather, 'it is to a large extent the result of institutional arrangements and savers' preferences' (Goldsmith, 1983, pp. 54).

It is difficult, therefore, to interpret the causal significance of these highly aggregative ratios. It is hard to argue that a given volume or composition of financial assets is a sufficient condition for the development of real sectors of the economy – or even a necessary condition, given that rapid growth has sometimes taken place during periods of deliberate financial repression. We are not, however, obliged to

retreat to a view of finance as purely passive, accommodating growth that is driven by other means. Financial innovation has at times sparked off virtuous circles of growth in particular sectors and regions. The microcredit movement in Bangladesh (and elsewhere) in response to extortionate rural moneylending is one recent example where a new financial technology, carefully managed, has been the spur to the growth of the incomes and welfare of poor borrowers. However, if building a functioning formal sector of financial intermediaries is arduous and costly, the evolution of financial structure and real economic development may well be mutually determined, with causation flowing in both directions (Greenwood and Jovanovic, 1990).

See Also

- credit rationing
- financial liberalization
- Goldsmith, Raymond William
- micro-credit

Bibliography

Eshag, E. 1983. *Fiscal and Monetary Problems in Developing Countries*. Cambridge: Cambridge University Press.

FitzGerald, E. 1990. Kalecki on financing development: an approach to the macroeconomics of the semi–industrialised economy. *Cambridge Journal of Economics* 14, 183–203.

Goldsmith, R. 1969. *Financial Structure and Development*. New Haven, CT: Yale University Press.

Goldsmith, R. 1983. *The Financial Development of India, Japan and the United States: A Trilateral Institutional, Statistical and Analytic Comparison*. New Haven, CT: Yale University Press.

Greenwood, J. and Jovanovic, B. 1990. Financial development, growth and the distribution of income. *Journal of Political Economy* 98, 1076–107.

Gurley, J. and Shaw, E. 1960. *Money in a Theory of Finance*. Washington, DC: Brookings Institution.

Gurley, J. and Shaw, E. 1967. Financial structure and economic development. *Economic Development and Cultural Change* 15, 257–68.

Kalecki, M. 1972. *Selected Essays on the Economic Growth of the Socialist and the Mixed Economy*. Cambridge: Cambridge University Press.

Levine, R. 1997. Financial development and economic growth: views and agenda. *Journal of Economic*

Literature 35, 688–726.

Levine, R. 2003. More on finance and growth: more finance, more growth? *Federal Reserve Bank of St Louis Review*, July/August, 31–46.

McKinnon, R. 1973. *Money and Capital in Economic Development*. Washington, DC: Brookings Institution.

Shaw, E. 1973. *Financial Deepening in Economic Development*. New York: Oxford University Press.

Stiglitz, J. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

World Bank. 2005. *Economic Growth in the 1990s: Learning from a Decade of Reform*. Washington, DC: World Bank.

How to cite this article

Toye, John. "financial structure and economic development." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000104> doi:10.1057/9780230226203.0572

The New Palgrave Dictionary of Economics Online

finite sample econometrics

Aman Ullah
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

asymptotic theory (large sample) econometrics; bootstrap; Edgeworth expansion; Edgeworth, F.; empirical likelihood estimators; finite sample method in econometrics; generalized least squares estimators; generalized method of moments estimators; hypothesis testing; least squares estimators; likelihood ratio method; linear models; maximum likelihood estimators; Monte Carlo methods; quantile estimators; Rao's score; simultaneous equations models; Wald's test

Article

Economic models, which provide relationships between economic variables, are useful in making scientific predictions and policy evaluations. Well-known examples include classical linear regression models, where the explanatory variables are assumed to be non-stochastic (fixed) and the errors are normally distributed, and non-classical models, where these assumptions are violated. These non-classical models are frequently used in empirical work, and they include the simultaneous equations model, models with serial correlation and heteroscedasticity, limited dependent-variables models, panel and spatial models, non-linear models, and models with non-normal errors.

Based on sample data, econometric methods provide techniques of estimation and hypothesis testing related to these and other models. The commonly used estimators are the least squares (LS) or the generalized LS (GLS), the maximum likelihood (ML), the generalized method of moments (GMM), the empirical likelihood (EL) and the quantiles. The hypothesis-testing procedures used are Wald's (W), Rao's score (RS) and the likelihood ratio (LR) methods. Since all these are based on sample information, the statistical properties (unbiasedness, consistency, efficiency, distributions) of these procedures are of great interest for both small and large samples. This has led to the development of asymptotic theory (large sample) econometrics (White, 2001) and finite sample econometrics (Ullah, 2004).

The large sample theory properties may not imply finite sample behaviour of econometric estimators and test statistics, and they can give misleading results for small or even moderately large samples. As an example, consider a regression model

$$y_i = x_i\beta + u_i \qquad i = 1, 2\ldots, n,$$

where y_i is a univariate response, x_i is a univariate fixed regressor, β is an unknown parameter to be estimated, and u_i is an additive error assumed to be independently and identically distributed (i.i.d.) with mean zero and variance σ^2. Let b_1, and $b_2 = (1 - 1/n)b_1$ be two estimators of β, where b_1 is the LS estimator. Then, the asymptotic distributions of b_1 and b_2 are

$$\sqrt{n}(b_1 - \beta) \sim N(0, \sigma^2 / m_{xx}), \sqrt{n}(b_2 - \beta) \sim N(0, \sigma^2 / m_{xx}),$$

where $m_{xx} = \Sigma_{i=1}^{n} \frac{x_i^2}{n}$ as n tends to ∞.
Thus, asymptotically, both estimators are unbiased, and they have the same variances and distributions. But these results do not hold for finite samples (small or moderately large), since in this case $Eb_1 = \beta$, $Eb_2 = \beta(1 - 1/n)$, $V(b_1) = \sigma^2 / \Sigma_{i=1}^{n} x_i^2$, $V(b_2) = (1 - 1/n)^2 V(b_1)$, that is, while b_1 is unbiased, b_2 is biased and their variances are different. Further, the distributions of b_1 and b_2 are generally not known but, if we assume normality of errors, then both b_1 and b_2 are normally distributed.
Fisher (1921; 1922) and then the work of Cramér (1946) laid the foundations of statistical finite sample theory on the exact distributions and moments which are valid for any sample size. This exact theory on distributions and moments was brought into econometrics by the seminal work of Haavelmo (1947) and Anderson and Rubin (1949) on the exact confidence regions of structural coefficients, Hurwicz (1950) on the exact LS bias in an autoregressive model, Basmann (1961) and Phillips (1983) on the exact density and moments of the estimators in the structural model, and Ullah (2004) on the exact moments. However, these exact results are often very complicated for drawing meaningful inferences since they are expressed in terms of multivariate integrals or complex infinite series. Also, the results are not derivable for non-classical models, especially for non-linear models or models with non-normal errors. Another major development took place through the pioneering work of Nagar (1959) on obtaining the approximate moments of the k-class estimators in simultaneous equations. This was followed by Sargan (1975) and Phillips (1980), who rigorously developed the theory and applications of the Edgeworth expansions to derive the approximate distribution functions of econometric estimators. (The idea of the Edgeworth expansions originates from the fundamental work of Edgeworth, 1896.) The approximate distributions and moments provide results which can tell us how much we lose by using asymptotic results and how far we are from the exact results if they are known. Most of the contributions, however, were confined to the analytical derivation of the moments and distributions in the simultaneous equations model and the dynamic first-order autoregressive (AR (1)) model, but with i.i.d. normal

observations. These also included the finite sample results using the Monte Carlo methodology (Hendry, 1984) and advances in bootstrapping (resampling) procedures (see Efron, 1979; Hall, 1992). The analytical and bootstrap results for non-classical models, especially those that are non-linear with non-normal and non-i.i.d. observations, remain a challenging task for future development in this area of research. For the approximate analytical results some development has begun to take place (Rilstone, Srivastava and Ullah, 1996) with a non-i.i.d. extension in Ullah (2004). This provides results which can be used to evaluate the approximate bias and mean-squared error of a class of estimators (ML, LS, GMM) for linear and non-linear models with normal or non-normal errors, and the observations can be i.i.d. or non-i.i.d. In the same spirit Newey and Smith (2004) develop the properties of generalized empirical likelihood estimators. Similarly, there are developments in the bootstrapping procedures for studying the properties of the GMM and extremum estimators in various econometric models with i.i.d. as well as dependent and non-stationary observations (see Horowitz, 2001).

The progress in finite sample econometrics has indeed been ongoing. The developments described provide analytical and simulation-based procedures for finite sample analysis of econometric models. In the broad sense, the frontier of this research area has moved on. With the advances in computer technology this subject will further develop in both the analytical and the bootstrapping domains.

See Also

- bootstrap
- econometrics
- simultaneous equations models

Bibliography

Anderson, T. and Rubin, H. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equation. *Annals of Mathematical Statistics* 20, 46–63.

Basmann, R. 1961. Note on the exact finite sample frequency functions of generalized classical linear estimators in two leading overidentified cases. *Journal of the American Statistical Association* 56, 619–36.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Edgeworth, F. 1896. The asymmetrical probability curve. *Philosophical Magazine* 41, 90–9.

Efron, B. 1979. Bootstrap methods: another look at the jacknife. *Annals of Statistics* 7, 1–26.

Fisher, R. 1921. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1, 1–32.

Fisher, R. 1922. The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society* 85, 597–612.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Haavelmo, T. 1947. Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Association* 42, 105–22.

Hendry, D. 1984. The Monte Carlo experimentation in econometrics. In *Handbook of Econometrics*, vol. 2, ed. M. Intriligator and Z. Griliches. Amsterdam: North-Holland.

Horowitz, J. 2001. The bootstrap in econometrics. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Hurwicz, L. 1950. Least square bias in time series. In *Statistical Inference in Dynamic Economic Models*, ed. T. Koopmans. New York: Wiley.

Nagar, A. 1959. The bias and moments matrix of the general k-class estimators of the parameters in structural equations. *Econometrica* 27, 575–95.

Newey, W. and Smith, R. 2004. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72, 219–55.

Phillips, P. 1980. Finite sample theory and the distribution of alternative estimators of the marginal propensity to consume. *Review of Economic Studies* 47, 183–224.

Phillips, P. 1983. Exact small sample theory in simultaneous equations models. In *Handbook of Econometrics*, vol. 1, ed. M. Intriligator and Z. Griliches. Amsterdam: North-Holland.

Rilstone, P., Srivatsava, V. and Ullah, A. 1996. The second order bias and MSE of nonlinear estimators. *Journal of Econometrics* 75, 239–395.

Sargan, J. 1975. Gram-Charlier approximations applied to t ratios of k-class estimators. *Econometrica* 43, 326–46.

Ullah, A. 2004. *Finite Sample Econometrics*. New York: Oxford University Press.

White, H. 2001. *Asymptotic Theory for Econometricians*. New York: Academic Press.

How to cite this article

Ullah, Aman. "finite sample econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000313> doi:10.1057/9780230226203.0573

The New Palgrave Dictionary of Economics Online

Finley, Moses (1912 – 1986)

Isabel Raphael

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Finley, M.; slavery

Article

Sir Moses Finley had an immense influence on classical studies and particularly ancient history because he brought to them the new disciplines and techniques of the modern social sciences. He was unique among ancient historians in that his early training had been in law, economics and sociology.

Born on 20 May 1912, Finley graduated (BA) from Syracuse University at the age of 15 and from Columbia (MA) at 17, his major subjects being psychology and US constitutional law. Westermann encouraged him to try ancient history, and he taught himself Latin and Greek, financing himself with his earnings and those of his wife Mary, a school teacher whom he married in 1932. Theirs was a childless but devoted marriage, Lady Finley dying two days before her husband.

Finley worked from 1930 to 1933 on the *Encyclopedia of Social Sciences* and was much influenced by the Frankfurt Institute for Social Research; his reading of social theory made him left-wing and at least partly Marxist. He was active on behalf of the Republicans during the Spanish civil war and raised funds for Russian war relief in the Second World War. After founding the American Committee for the Defence of International Freedom against McCarthyism he was dismissed from his post as Assistant Professor of History at Rutgers University. Known by now for his lectures in England, he was given the post of lecturer in classics at Cambridge in 1955, and was a Fellow of Jesus College from 1957 to 1976. He became a British subject in 1962. He succeeded to the chair of ancient history in 1960, and in 1976 became the first Master of Darwin College. Finley's doctoral dissertation, ‘Studies in Land and Credit in Ancient Athens’ (1950), gained him an international reputation. He asked questions that had not been considered before in this field, and saw the ancient world with modern eyes. Classical scholars had used the word ‘economics’ in its ancient and particular sense, as the management of a household and hence of a state; Finley opened up the discipline to the interests of modern social sciences, dealing with matters such as property, contracts, succession, the value of goods and coin and the laws of war. He stepped aside from the traditional track to look at the exact relationship between masters and slaves, the nature of

debt bondage, the consumer society and urban and rural production. He was the first ancient historian to tackle the methodological problems implied by the new style of social history.

Finley could appear cantankerous and was famous for his feuds; he enjoyed creating shock waves in the academic world. But at his best he was a new wind blowing through an old and rather old-fashioned subject, and he changed and refreshed the classics more than any other scholar this century.

Selected works

1956. *The World of Odysseus*. London: Chatto & Windus.

1963. *The Ancient Greeks*. London: Chatto & Windus.

1970. *Early Greece: The Bronze and Archaic Ages*. London: Chatto & Windus.

1973a. *Democracy Ancient and Modern*. London: Chatto & Windus. Revised edn, 1985.

1973b. *The Ancient Economy*. London: Chatto & Windus.

1980. *Ancient Slavery and Modern Ideology*. London: Chatto & Windus.

How to cite this article

Raphael, Isabel. "Finley, Moses (1912–1986)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000113> doi:10.1057/9780230226203.0574

The New Palgrave Dictionary of Economics Online

firm boundaries (empirical studies)

Thomas N. Hubbard

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The empirical literature on the determinants of firms' boundaries examines relationships between firms' boundaries and asset specificity, especially how relationship-specific investments create 'hold-up' problems that increase the costs of competitive contracting; relationships between firms' boundaries and the contracting environment, reflecting the role of incomplete contracting in the theoretical literature and the extent to which firms subcontract downstream stages rather than input procurement; and how firms' boundaries vary with 'job design'. This literature has established that asset specificity is empirically relevant for understanding integration decisions, and that relationships between subcontracting decisions, the contracting environment, and the division of labour are subtle.

Keywords

agency costs; asset specificity; Coase, R.; contracting; division of labour; firm boundaries; hold-up problem; incomplete contracts; outsourcing; vertical integration

Article

This article discusses empirical work on the determinants of firms' boundaries, focusing on 'make-or-buy' decisions. Examples of such decisions include whether firms procure inputs (or distribute outputs) through in-house divisions or other firms. It concentrates on work that draws on Coase (1937), which depicts firms and markets as alternative means of governing transactions. The theoretical literature in the Coasean tradition is vast, and includes well-known works by Williamson (1975; 1979; 1985), Klein, Crawford and Alchian (1978), Grossman and Hart (1986), and Holmstrom and Milgrom (1994). This contrasts with the neoclassical literature, in which firms' boundaries are determined by production technology and, perhaps, market power-related issues. This other literature examines how, for example, vertical integration reflects firms' incentive to eliminate double marginalization, raise rivals' costs, or protect themselves from competitors' attempts to raise their own costs.

Firms ’ boundaries and relationship-specific assets or investments

By far the largest branch of the empirical literature examines relationships between firms’ boundaries and asset specificity. This branch is primarily motivated by Klein, Crawford and Alchian's (1978) and Williamson's (1979; 1985) analysis of how relationship-specific investments create ‘hold-up’ problems that increase the costs of competitive contracting. On the assumption that such investments do not create as severe problems when transactions take place within firms, it follows that vertical integration should be more prevalent, and outsourcing less prevalent, when transactions involve relationship-specific assets than when they do not.

Several early papers examine this proposition in procurement contexts. Monteverde and Teece (1982) and Masten (1984) examine outsourcing decisions of auto makers and an aerospace firm, respectively, and find that outsourcing is less prevalent when components are firm-specific than not. The latter finds that it is also less prevalent when co-locating production of the component with that of successive production stages is more valuable. Joskow (1985) finds that vertical integration is prevalent when coal-burning power plants are located close to coal mines, but power plants procure coal from outside firms when they are not co-located. These and other correlations uncovered by this early work provided the first evidence that asset specificity is empirically relevant for understanding integration decisions and, more broadly, that analysing firms’ boundaries from a contractual perspective could lead to new empirical insights.

This branch has since developed along several lines. Researchers have found relationships between asset specificity and vertical integration in other industrial contexts, and have explored the empirical limits of this proposition by examining the extent to which asset specificity and integration are correlated in contexts where investments are smaller and less specific than in the contexts discussed above. Still others have investigated the closely related question of how, given that vertical integration is not chosen, contractual relationships vary with asset specificity (see Joskow, 1988, and Klein, 2005, for comprehensive surveys).

There is significant debate over the theoretical interpretation of this evidence. Asset specificity is an important element of many theories in this literature, so correlations between asset specificity and vertical integration need not provide evidence in favour of any one in particular. Whinston (2003) discusses this problem at length, and concludes that, while these theories’ empirical implications are not the same, the data requirements of distinguishing tests are considerable and the existing empirical evidence is not dispositive.

Some recent papers indicate that the relationship between vertical integration and investment can be subtle. Woodruff's analysis (2002) of vertical integration between shoe manufacturers and retailers, and Acemoglu et al.'s analysis (2004) of vertical integration in British manufacturing indicate that whether vertical integration is more or less prevalent when investments are more important depends critically on the source and nature of the investment. Understanding empirical relationships between integration and investment incentives is a major focus of current research.

Firms ’ boundaries and the contracting environment

A second branch of the empirical literature examines relationships between firms' boundaries and the contracting environment. Many theories motivate this research, reflecting the essential role incomplete contracting plays throughout the theoretical literature. This branch typically examines the extent to which firms subcontract downstream stages rather than input procurement. Examples include Anderson and Schmittlein's work (1984) on whether manufacturers rely on internal or external sales representatives, Baker and Hubbard's investigations (2003; 2004) of firms' boundaries in trucking, Brickley, Linck and Smith's analysis (2003) of whether bank offices are independent entities or branches, and some of the research (see Lafontaine and Slade, 1997, for a survey) that examines whether chain outlets are company-owned or franchises. These papers exploit variation in the availability of good measures of downstream individuals' performance, which, in turn, derives from technological change or differences in the nature of the downstream individual's job. (Work exploiting the latter is classified here rather than below, when authors emphasize differences in the contractibility rather than the number or diversity of tasks.) Results from these papers generally indicate that more (downstream) integration tends to be associated with better performance measures.

These results have several implications. First, they imply that the contracting environment is not organization-neutral. Non-neutrality is not obvious. Agency problems exist between upstream and downstream entities regardless of whether the latter are employees or subcontractors. If contractual improvements reduce agency costs independently of integration-related trade-offs, they should not affect firms' boundaries. The results indicate otherwise. Second, they imply that the contracting environment affects the costs of transacting within as well as between firms. Some theories, including Coase (1937), propose that coordination takes place 'by fiat', and hence the contracting environment is irrelevant, within firms. If so, contractual improvements should always favour market transacting and thus less vertical integration. Again, the results indicate otherwise. Third, they suggest that while contracting problems exist both within and between firms, empirical variation in the availability of good performance measures tends to be related to inefficiencies associated with transacting within firms. Although this conclusion is preliminary, it implies that it is particularly productive for those researching (or making) 'make-or-buy' decisions to identify the source of these inefficiencies, because they may have more real-world 'bite'.

Firms' boundaries and the division of labour

A third, related branch examines how firms' boundaries vary with the division of labour, or 'job design'. Holmstrom and Milgrom's (1994) analysis of how multitask agency problems influence firms' boundaries motivates much of this work. This branch includes analyses of how whether in-house salesmen or sales reps are used depends on whether salesmen are also given non-selling responsibilities (Anderson, 1985), how whether pharmaceutical firms outsource clinical trials depends on whether the work involves more than just data collection (Azoulay, 2004), and how whether restaurants are company-owned or franchised depends on how much food production and service takes place at the restaurant (Yeap, 2005).

Results from these papers indicate that integration tends to be less prevalent when individuals are allocated a narrower set of responsibilities. Combined with the evidence above, they imply that relationships between subcontracting decisions, the contracting environment, and the division of labour

are subtle. The previous subsection suggests that *replacing* an easily contractible task with a less contractible one tends to make subcontracting more likely. The evidence here suggests that *adding* a less contractible task to a more contractible one tends to make subcontracting less likely.

Other work has found that the division of labour and firms' boundaries are related in horizontal contexts as well; for example, Garicano and Hubbard (2003) find that law firms' field boundaries narrow as market size increases and lawyers become more specialized.

Firms' boundaries and economic outcomes

Most of the literature investigates what determines whether firms integrate rather than what actually happens when they do, but research on the latter is important because it reveals whether integration is an economically important issue.

Some evidence has come from firm or industry case studies. Early work includes Masten, Meehan and Snyder (1991), which concludes that organizational costs make up a significant fraction of production costs in shipbuilding, and that incorrect choices with respect to integration decisions can increase organization-related costs by as much as 70 per cent. More recently, Gil (2004) investigates relationships between how long movies play at a theatre and whether the theatre is owned by the movie's distributor, and finds that movies play two weeks longer at distributor-owned theatres than other, similarly situated theatres.

Perez-Gonzalez (2004) provides cross-industry evidence. This investigates how the elimination of Mexican laws that constrained multinational firms from having majority control of affiliated enterprises affected plant-level investment and productivity. He finds that, within technology-intensive sectors, increases in integration associated with the elimination of these constraints led to significant investment increases and an approximately ten per cent increase in total factor productivity at these enterprises. The allocation of control rights, and thus vertical integration, can have a major impact on investment incentives and productivity. In short, integration decisions can matter a lot.

See Also

- contract theory
- hold-up problem
- incomplete contracts
- property rights

Bibliography

Acemoglu, D., Aghion, P., Griffith, R. and Zilibotti, F. 2004. Vertical integration and technology: theory and evidence. Working Paper No. 10997. Cambridge, MA: NBER.

Anderson, E. 1985. The salesperson as outside agent or employee: a transaction-cost analysis. *Marketing Science* 4, 234–54.

Anderson, E. and Schmittlein, D. 1984. Integration of the sales force: an empirical examination. *Rand Journal of Economics* 15, 385–95.

Azoulay, P. 2005. Capturing knowledge within and across firm boundaries: evidence from clinical development. *American Economic Review* 94, 1591–612.

Baker, G. and Hubbard, T. 2003. Make versus buy in trucking: asset ownership, job design, and information. *American Economic Review* 93, 551–72.

Baker, G. and Hubbard, T. 2004. Contractibility and ownership: on-board computers and governance in U.S. trucking. *Quarterly Journal of Economics* 119, 1443–79.

Brickley, J., Linck, J. and Smith, C. 2003. Boundaries of the firm: evidence from the banking industry. *Journal of Financial Economics* 70, 351–83.

Coase, R. 1937. The nature of the firm. *Economica* 4, 386–405.

Garicano, L. and Hubbard, T. 2003. Specialization, firms, and markets: the division of labor within and between law firms. Working Paper No. 9719. Cambridge, MA: NBER.

Gil, R. 2004. Decision rights and vertical integration in the movie industry. Mimeo. Department of Economics, University of California, Santa Cruz.

Grossman, S. and Hart, O. 1986. The costs and benefits of ownership: a theory of vertical and lateral integration. *Journal of Political Economy* 94, 691–719.

Holmstrom, B. and Milgrom, P. 1994. The firm as an incentive system. *American Economic Review* 84, 972–91.

Joskow, P. 1985. Vertical integration and long term contracts: the case of coal-burning electric generating plants. *Journal of Law, Economics, and Organization* 1, 33–80.

Joskow, P. 1988. Asset specificity and the structure of vertical relationships: empirical evidence. *Journal of Law, Economics, and Organization* 4, 95–117.

Klein, P. 2005. The make-or-buy decision: lessons from empirical studies. In *Handbook of New Institutional Economics*, ed. C. Menard and M. Shirley. Boston: Kluwer.

Klein, B., Crawford, R. and Alchian, A. 1978. Vertical integration, appropriable rents and the competitive contracting process. *Journal of Law and Economics* 21, 297–326.

Lafontaine, F. and Slade, M. 1997. Retail contracting: theory and practice. *Journal of Industrial Economics* 45, 1–25.

Masten, S. 1984. The organization of production: evidence from the aerospace industry. *Journal of Law and Economics* 27, 403–17.

Masten, S., Meehan, J. and Snyder, E. 1991. The costs of organization. *Journal of Law, Economics, and Organization* 7, 1–25.

Monteverde, K. and Teece, D. 1982. Supplier switching costs and vertical integration in the automobile industry. *Bell Journal of Economics* 13, 206–13.

Perez-Gonzalez, F. 2004. The impact of acquiring control on productivity. AFA 2005 Philadelphia Meetings.

Whinston, M. 2003. On the transaction cost determinants of vertical integration. *Journal of Law, Economics, and Organization* 19, 1–23.

Williamson, O. 1975. *Markets and Hierarchies: Analysis and Anti-Trust Implications*. Glencoe, IL: Free Press.

Williamson, O. 1979. Transaction-cost economics: the governance of contractual relations. *Journal of Law and Economics* 22, 233–61.

Williamson, O. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

Woodruff, C. 2002. Non-contractible investments and vertical integration in the Mexican footwear industry. *International Journal of Industrial Organization* 20, 1197–224.

Yeap, C. 2005. Residual claims and incentives in restaurant chains. Mimeo. University of Chicago.

How to cite this article

Hubbard, Thomas N. "firm boundaries (empirical studies)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000288> doi:10.1057/9780230226203.0575

The New Palgrave Dictionary of Economics Online

firm, theory of the

G.C. Archibald

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agency theory; bargaining; bounded rationality; comparative statics; Cournot, A. A.; creative destruction; dynamic programming; expense preference; firm, theory of the; free-rider problem; Gibrat's Law; implicit contracts; incentive compatibility; innovation; institutional rent; insurance; joint-stock companies; laws of returns; linear programming; long run and short run; long-run equilibrium; marginal revolution; Markov processes; monitoring; optimal control; optimizing models; ownership and control; principal and agent; quasi-rent; rent; representative firm; residual; returns to scale; risk; risk aversion; risk sharing; transaction costs; uncertainty.

Article

It is doubtful if there is yet general agreement among economists on the subject matter designated by the title ‘theory of the firm’, on, that is, the scope and purpose of the part of economics so titled. There is, probably, general agreement on the subject matter of economics itself: the allocation and distribution of scarce resources. (Some economists would have us add explicitly ‘and growth’ to ‘allocation and distribution’, but traditionally growth is subsumed under ‘allocation’.) Then we may take it that the purpose of the theory of the firm is to investigate the behaviour of firms as it affects allocation and distribution. We now come immediately to a fork. An economist who believes that a ‘firm’ is a profit-maximizing agent (whether by conscious, rational decision or otherwise), endowed with a known and given technology, and operating subject to a well-defined market constraint, will see no need for any special theory of the firm: the theory of the firm is nothing but the file of optimizing methods (and perhaps market structures). *If* firms maximize, *how* they do it is not of great interest or at least relevance to economics. The economist's job is simply to cultivate and apply optimizing techniques. Given this view, it is unnecessary to inquire further: to seek to ‘inquire within’ is otiose, perhaps methodologically misguided. (As we shall see, the theory of the firm has been, and perhaps still is, the battleground for some fierce methodological warfare.) Economists who doubt any of the three critical assumptions see an urgent need to inquire within, but diverge substantially thereafter (for example, managerial utility

functions, behaviourism). Later on, I shall try to exhibit a systematic tree, although this is not easy since some of the branches are sadly tangled. Before doing that, I want to show that the first fork, referred to above, was recognized a long time ago, and to sketch some of the history of our subject. First, though, I must impose more narrow limits on it.

In most of the work on the theory of the firm it is at least implicitly assumed that the agent whose behaviour is to be examined is a capitalist firm (which may or may not be a joint-stock corporation) engaged in manufacturing, processing or perhaps extraction. Thus the study of financial intermediaries, although they are firms, is conventionally relegated to some other branch of our discipline. Partnerships and cooperatives (labour-managed firms) may be usefully examined with the techniques of the theory of the firm, as may not-for-profit organizations, but their study is conventionally filed under ‘comparative systems’. For convenience and brevity, although not out of conviction, I shall respect these conventions here. It is also necessary to place some demarcation line between the theory of the firm and ‘market structure’ or ‘industrial organization’. For the moment, at least, I think it better to let this one be implicit.

We must also ask why firms exist at all. The classic – and neoclassical – answer was provided by Coase (1937): transactions costs. I call this a ‘neoclassical’ answer because part of the tradition, still embodied in much contemporary general equilibrium theory, is the assumption of constant returns to scale. Some increasingness of returns may be a very good reason for the existence of firms, or at least help to explain their size, but it is obviously vastly convenient to have a sufficient reason which is not inconsistent with constant returns. Coase suggested that the firm was an area (subset of the economy) in which allocation proceeded by direction rather than via markets, because some procedures, such as the allocation of workers to tasks, could be more cheaply done that way – coordination by command rather than by price. The word ‘command’ suggests that some monitoring, enforcement or internal incentive structure will be required, and indeed these matters have been receiving increasing attention. Alchian and Demsetz (1972), in particular, discussed the problem of monitoring, suggesting, in effect, that the need for it explained and justified the existence of the capitalist firm. They posed the question of who monitors the monitor, and suggested that the incentive problem is solved if the ultimate monitor is the residual claimant. O. Williamson (1980) reviewed alternative organizational structures. He suggested that the existence of firms economizes on explicit contracts which, given uncertainty and bounded rationality, are expensive instruments. He also found that ownership and hierarchy are only weakly related.

A recent work to emphasize the reasons for the existence of firms is Aoki's (1984). He argues that if firms exist because institutional allocation is cheaper than market allocation, reasons for which he explores thoroughly, then firms must enjoy ‘institutional rent’. Furthermore, not all the resources used within the firm will have prices uniquely determined by external markets. Thus the distribution of rewards is not uniquely determined, and there is room for bargaining. Aoki argues that this is best modelled as a cooperative *game*, the players of which are the stockholders and the workers. Managers are reduced to the role of technocratic mediators (which, in view of recent developments in agency theory, discussed below, is perhaps surprising). This approach proves to be very flexible: Aoki can handle as special cases the neoclassical model (shareholders get all the residual) and the labour-managed firm in which the workers get it all (and even, with some interpretation, managerial models).

In what follows, I shall take the existence of firms for granted and return later to the matter of incentives. The first fork, referred to above, will be familiar to any careful reader of Adam Smith (1776). He relied upon the self-interest of the butcher, the baker and the brewer to provide his dinner. The ‘firms’ in which

he had confidence were small, owner-operated (whether single owner or partnership), without limited liability. He had serious misgivings about joint-stock companies. He pointed out what has become known in this century, thanks to Berle and Means (1933), as the ‘divorce between ownership and control’. And he doubted if the managers had appropriate incentives to try to maximize the owners’ returns; that is, he raised the question of what is now called ‘incentive compatibility’. Thus, in considering the joint-stock company, Smith went unhesitatingly down what I will call the ‘troublemaker's branch’: we do have to inquire within. The joint-stock company is, of course, the predominant contemporary organization.

After Smith, there is not much that can be called ‘theory of the firm’ in classical economics. (Ricardo's firms are Smith's butchers and bakers.) The exception, as so often, is Marx, but there is not space to discuss Marx here. (J.S. Mill, 1848, in the famous chapter ‘On the Probable Futurity of the Labouring Classes’, expressed concern about both the incentive structure and morality of the capitalist form of organization, and recommended a cooperative form instead.) We must notice, however, the startlingly modern work of Cournot (1838). He wrote down a demand function and, in his famous discussion of the mineral spring, employed explicit optimizing methods (and, so far as I know, was the first to do so). Not only this, he carried out a deliberate and formal exercise in comparative statics – in 1838! In applying marginal analysis to the theory of the firm he thus thoroughly anticipated the ‘marginalists’. The ‘marginal revolution’ in due course produced a wholly desirable unification of the theories of production, allocation and distribution, creating the neoclassical branch from the fork, but with little that could be called ‘theory of the firm’. The firm was, however, central in Marshall's (1890) work, and he, characteristically, put a foot on each branch. Formal, mathematical, Marshall is strictly neoclassical, as I employ the term. The informal Marshall, concerned with growth, offered suggestive literary dynamics.

Let us consider first the more formal Marshall. His distinction between the short and long runs is essential to much of his work. This distinction is, of course, the one currently in use: in the long run all factors are variable, in the short run one at least (commonly capital) is not. This allowed him to distinguish between fixed and variable costs, and between the effects of adding more labour to a fixed-capital stock and the effects of altering the scale of operations. We now have short-run diminishing returns in industry generally, while there may be increasingness in the long run. Thus Marshall was not limited to the constant coefficients case of his classical predecessors: he was able to offer a thorough analysis of the ‘laws’ of returns. This allowed him to give a fairly complete analysis of the short-run equilibrium conditions for a firm selling in a perfect market. (There is in his analysis an even shorter ‘short run’, the market period in which the price of, say, a catch of herrings is determined. This does not appear to concern us here.) Marshall did not, of course, solve all the problems of the theory of production, costs, supply and distribution in competition. He left room for the important work of Viner (1931) and Stigler (1939).

A further and vital step was Marshall's generalization of Ricardo's theory of rent. He distinguished between a quasi-rent, which would in the long run be competed away, and a true rent, which definitionally could not be. (Both, of course, are any excess of rewards over opportunity cost.) If the quasi-rent is due to an increase in the demand for the product of specific capital equipment, then the long run in which it is competed away and the long run in which all factors are variable are, of course, identical. (That the period in which quasi-rent is competed away and that in which all factors are variable may differ is noted below.) This in turn allowed Marshall to develop the long-run equilibrium

conditions for a competitive industry: quasi-rent must be competed away (or negative profit eliminated by exit) so that the normal profit condition is satisfied. Here he seems to have followed Walras (1874).

Marshall made many other contributions to the theory of the firm. He noted that, if increasingness in returns (to scale, as we should say) is internal to the firm, competition is not viable, whence a downward-sloping competitive supply curve can only be attributed to economies external to the firm (internal to the industry; but he also considered economies external to the industry and internal, perhaps, only to the whole economy). He also offered a formal monopoly model some features of which require remark. The firm's demand curve coincides with the market demand curve for the ‘product’ (a given primitive of analysis): there is no oligopolistic interaction here. This model is still with us, although the analysis has become more elegant. In his geometry, Marshall had us finding the profit-maximizing output by looking for the biggest profit rectangle: $(AR\text{-}AC)q$. Cournot (1838) had written down the marginal revenue function in his discussion of the mineral springs case, but Marshall chose not to follow him. (The discovery of the marginal revenue curve in Cambridge in the 1930s seems to have caused great excitement.)

The less formal Marshall was concerned with growth and the intertemporal behaviour of firms. His firms were joint-stock, but otherwise rather Smithian. He had, loosely speaking, a ‘clogs to clogs in three generations’ model. The first entrepreneur would be vigorous and innovative, finding some source of quasi-rent. His son would be more passive and probably mistake the quasi-rent for rent itself. The spoiled and idle grandson would certainly make this mistake, the quasi-rent would be competed away, and the cycle would be over.

This is, of course, not a good description of the history of a typical (immortal) joint-stock company. What is important is the link between innovation, quasi-rent and economic growth. Now, of course, the period in which quasi-rent is competed away is not necessarily identical to that in which capital can be varied. It may be possible to copy an innovation very quickly, or necessary to wait for the expiry of a patent. And if the quasi-rent is due to exceptional managerial talent and vigour (really, a rent to ability), it does not get competed away at all, but eventually withers. It was, however, this link between innovation and quasi-rent that Schumpeter (1934) made explicit in his great vision of the source of growth in a capitalist economy: the incessant seeking for quasi-rent via innovation, each source of quasi-rent being in turn competed away by further innovation in the process of ‘creative destruction’. One notes, of course, that this model does not depend on the generational cycle of Marshall's family firm: widely owned joint-stock companies can continue to play Schumpeter's game so long as they are appropriately managed.

Marshall had the task of reconciling his view of the intertemporal behaviour of firms with his short-run profit-maximizing conditions and long-run industry equilibrium conditions. His device of the ‘representative firm’ appears to have been designed to do this. The representative firm would not only be in short-run profit-maximizing equilibrium but would be earning precisely normal profit when the industry as a whole was in equilibrium. This means that the definition of long-run equilibrium needs to be more carefully stated. It is not ‘all firms earn normal profit’. It is rather ‘there is no tendency for the total number of firms in the industry to alter; the representative firm earns normal profits but others may still be expanding or already withering; in any case the net change is zero.’ Here the representative firm is implicitly defined. As Newman (1960, p. 590) put it, in his discussion of Marshall's ‘statistical’ concept of long-run equilibrium, ‘Long-run equilibrium for Marshall meant the equality of long-run demand and supply; just that and no more.’ In the 1920s and 1930s there was a considerable literature on

Marshall's value theory, not discussed here (see Newman, 1960, for references). Since the work of Chamberlin (1933) and Joan Robinson (1933), the notion of the representative firm has tended to disappear from the literature. It has become usual to assume that each firm is always, by choice, in short-run equilibrium, and then to consider how Marshall's long-run competitive forces will impose industry equilibrium (normal profit for all firms simultaneously). Newman and Wolfe (1961), on the other hand, followed up the ‘statistical’ interpretation of Marshall's long-run equilibrium. They were not the first to apply Markov-chain analysis to the behaviour of an industry; but they were the first to integrate it with value theory. (Other more or less contemporary applications of Markov-chain analysis at most appeal to ‘Gibrat's Law’. Newman and Wolfe may be thought to have prepared the ground for Nelson and Winter, 1982, discussed below.)

I shall now attempt to describe some other forks and branches of the tree. To do this it is easiest to jump to the present, since so much has happened since the Second World War that needs to be allocated to its appropriate branch. (Chamberlin, 1933, and Joan Robinson, 1933, had, of course, made significant extensions of Marshall's formal models before the war. These contributions are discussed elsewhere.)

We encountered above a fork between what I call the smooth neoclassical branch and the rough and troublesome ‘other’ branch. There is another possible basis for classification, between optimizing and other models. The advantage of the first is that it gives the neoclassical model the prominence it deserves; the advantage of the second that it brings into prominence the importance of the assumptions we make about information and computational capacity. Perhaps somewhat arbitrarily, I shall classify the models to be considered here as optimizing and ‘other’. The optimizing set of models divides again, between profit maximization and the optimization of other (usually managerial) objective functions.

Let us consider some arguments concerning the classes of models we have already identified.

The advantages of an optimizing model are clear: it is analytically tractable. We have well-developed techniques to handle it, even if the economic agents considered may not. It may also be thought to have important predictive power, but this is more dubious. The programme of qualitative comparative statics (Samuelson, 1947) has been shown to be more limited than we might have hoped. The objections to optimizing models are well known, but also debatable. They are essentially two. The first is that firms, or the human beings that manage them, cannot optimize: they have neither the information nor the computational capacity, whence the most we can have is Simon's ‘bounded rationality’ (Simon, 1955; 1959; see also 1979). Nelson and Winter (1982) have recently made a major contribution to this approach, discussed below. The position here is not that we give up the fundamental Smithian assumption of purposeful, self-interested behaviour (with what would we replace it?) but rather that we abandon the optimizing model and consider instead how, in a world of uncertainty, firms (managers) may explore their environment and try to ‘make the best of it’. It is not suggested, at least by Nelson and Winter, that we ‘inquire within’ for the sake of it but rather to improve our understanding of how actual firms, seeking for profit but essentially too ignorant to optimize, may try to allocate resources. The second objection to optimizing models comes from those who have enquired within and report that firms ‘just don’t’ (see, for example, Hall and Hitch, 1939; Andrews, 1949; Cyert and March, 1963). Many critics of this behaviourist school feel that it says little more than ‘firms do what they do’, and fails to analyse the relationship between the observed behaviour reported and resource allocation.

An example may show the force of the criticism. It is no longer open to doubt that firms commonly adopt markup pricing routines. In their study of a department store, Cyert and March (1963) report their

discovery of the markup formula in use. They then congratulate themselves on being able to predict, given the wholesale price of an article, its posted price. They also notice that if profits are not satisfactory, the firm may adjust by altering its product-mix; that is, buying better (more expensive) or cheaper stock. But it is here that the important allocational decisions are taken, and this decision process is not analysed at all. (It should be noted that Cyert and March, 1963, p. 268, place on their agenda matters which do not appear to be relevant to allocation and distribution at all, and which I accordingly exclude from consideration.)

Two related arguments in favour of profit-maximizing models may usefully be noticed now. The first is the ‘biological analogy’: survival of the fittest (see Alchian, 1950; Penrose, 1952; Friedman, 1953; Machlup, 1946; 1967). It is suggested that in a competitive world a firm must maximize to survive. Thus, however decisions are taken, whatever routines are adopted, firms which in fact maximize will prosper and be able, in particular, to retain and attract capital, while those that do not will wither. There are three points to raise here. The first is: how competitive is the environment? (See below.)

The second is that to survive, one does not have to be perfect but only good enough to handle the competition. Indeed, Charles Darwin seems to have anticipated this misuse of his argument when he wrote,

> Natural selection tends only to make each organic being as perfect as, or slightly more perfect than, the other inhabitants of the same country with which it has to struggle for existence … Natural selection will not produce absolute perfection … (Darwin, 1859, pp. 201–2)

The third is that, to make effective use of the *biological* analogy, one has to offer something that can serve as a *gene*. Nelson and Winter (1982) have recently suggested a candidate (see below).

The second, and related, argument is that one can maximize without consciously trying. Thus Day and Tinney (1968) show that a firm can climb to the top of a (suitably concave) profit ‘hill’ by use of a simple feedback algorithm: if an action (change in output) succeeds (increases profit), repeat it; if not, back up. The notion that one may climb the hill ‘driving only by the rear-vision mirror’ must certainly be attractive to those who worry about the firm's information state and computational capacity. Yet obviously this simple feedback process works only if it converges ‘fast enough’ relative to the stability of the environment. Otherwise, it will be necessary to improve the algorithm to speed up convergence; for example, by adding feed-forward loops. The survival argument suggests that it will then be the firms that can do this that will survive. Then the loops (routines) are identified by Nelson and Winter as the genes in the evolutionary process. Notice, however, that this identification was made in 1982, not by those who originally proposed the biological analogy (see also Winter, 1975).

We have now distinguished between optimizing models and ‘other’. We have glimpsed the next two subdivisions, that between profit-maximizing and other optimizing models, and between behaviourism and other non-optimizing models. (We shall soon find another fork on the profit-maximizing branch, too; see below.) We have also noticed some relevant argument. We may now explore some developments along each of these branches.

Developments in and since the Second World War, some emerging from operations research, have extended the scope of optimizing models at a staggering rate. In a few short years, we had linear

programming (for economic applications, see Dorfman, Samuelson and Solow, 1958), and activity analysis (see Koopmans, 1951). Optimizing techniques were extended to inventory control (Whitin, 1953; Simon, 1952). We then had what I will call the ‘dynamic explosion’ as the techniques of optimal control and dynamic programming were increasingly applied to the firm's problems; see, for example, Lucas (1967) and Treadway (1969) on the flexible accelerator, Mortensen (1970) and Brechling (1975) on the demand for labour.

Another major development has been the extension of optimizing models of the firm to include considerations of risk. Risk had been explicitly considered by Knight (1921), who offered an unsurpassed account of the ways in which the institutions of the capital market facilitate risk sharing. Knight tried to distinguish between ‘risk’ and ‘uncertainty’ in a way that many have found unsatisfactory: ‘risk’ was insurable; ‘uncertainty’, any uninsurable residual. Profit was the reward for bearing uncertainty (since risk could be covered by insurance). He was, I believe, the first to make the point that entrepreneurs would have to be less risk-averse than others (their employees) with whom they entered into explicit contracts. Recent work does not, however, follow Knight. It took a new departure from the work of von Neumann and Morgenstern (1944); see particularly Arrow (1971), and for specific applications to the theory of the firm, see for example Sandmo (1971). The main result (Sandmo) is that the risk-averse competitive firm will produce less than a risk-neutral competitive firm or one which knew with certainty that the price was going to be equal to its expected value. Drèze (1985) has used risk as a means of introducing a more realistic model of the firm into general equilibrium theory. General equilibrium theory is beyond the scope of this essay; but we should note that he does ‘inquire within’ and that his approach has much in common with that of Aoki (1984).

This brings us to a fork on the profit-maximizing branch. The divorce between ownership and control is explicitly recognized and the theory of agency developed to deal with it. The divorce occurs whenever an owner (or principal) submits a risky operation in which he has an interest to an operator (or agent) whose conduct he cannot monitor costlessly. Thus the theory of agency, originally developed in the discussion of sharecropping (risk sharing) and other forms of tenancy (see Stiglitz, 1974) has the widest application, evidently to insurance, and, of particular interest in the present context, to the interior operations of firms, not only the relationship between owners and controllers but even between managers and teams (of employees) (see particularly Ross, 1973; Jensen and Meckling, 1976; Holmstrom, 1982; Grossman and Hart, 1983). It is commonly cheaper to give the operator (whether tenant, car-driver or executive) an incentive to good behaviour than to try to monitor him or her. This, of course, leads to less than optimal risk sharing (collision deductible in automobile insurance). Another incentive to good behaviour in the face of costly monitoring is suggested by Eaton and White (1983): this is to give an employee a bonus, a wage above his or her opportunity cost, so that, in the case that misconduct is detected, dismissal is a genuine penalty (see also Shapiro and Stiglitz, 1984). Thus both carrots and sticks have been considered. When behaviour is unobservable, incentive compatibility may require some surprising forms of contract. Thus Holmstrom has shown that the only way to avoid the free-rider problem in a team in which effort is not observable is a contract which threatens to break the budget: deliver the target, or no member gets anything (someone else takes the full value of whatever is delivered). This raises two immediate problems. First, it may pay the ‘someone else’ to bribe a member of the team to shirk (‘just a little’). Second, if achievement of the target depends on effort and some

random variable(s), how would risk-averse members of the team dare to enter into such a contract?
Above I distinguished between two approaches to the theory of the firm, that of the maximizers and of those who wished to ‘inquire within’. In agency theory we see the two converging. We are ‘within’, but not for its own sake; the agenda is still the allocation and distribution of scarce resources. We are forced within to deal, inter alia, with problems raised by Adam Smith over two centuries ago, in conjunction with our own better understanding of risk.
Let us now consider other optimizing models. These depend not merely upon the divorce between ownership and control but on the idea that there is ‘slack’ within which the controllers may play their own game without being noticed and called to account. This in turn depends on the existence of market imperfections. The usual story has been that large firms are typically in a position to make monopoly rents, and that these rents can be forgone, used up, or ploughed back at the discretion of the controllers. It is acknowledged that rents usually turn out to be quasi-rents, but suggested that the large firms (conglomerates) can, by heavy R&D expenditure, enjoy a perpetual stream of quasi-rents: while one source is being competed away, another is being developed (perhaps patented). Thus there is always some room for discretionary expenditure by the controllers. This room may in turn be limited by the perspicacity of the capital market, but it is suggested (Marris, 1964) that the power of the capital market to discipline controllers is limited by the costs of information and the fact that the supply of capital to potential takeover raiders is not infinitely elastic. Suppose, however, that capital markets were perfect. So long as the divorce between ownership and control remained, so would the problem of arranging incentive-compatible contracts for managers, whoever owned the equity.
How much scope for discretionary behaviour there actually is, then, is an empirical question to which we do not have a final answer. There is, however, no shortage of models of how managers will behave if they have the room – room to maximize their own utility functions, that is. We have Baumol (1959): maximize growth subject to a minimum profit constraint. Marris (1964) and J. Williamson (1966) offer more sophisticated versions. O.E. Williamson (1964) introduces the idea of ‘expense preference’. The controllers can dissipate the rents by padding costs in ways which increase their utility. These ideas (and there are others) have obvious application to regulated industries, at least in the case in which the regulatory standard is a profit ceiling. Marris and J. Williamson both take into account the financial structure of the firm. There is now a large literature on this subject which I shall not discuss here.
(The first formal application of utility maximization to the theory of the firm was probably Scitovsky's, 1943. I have not listed him above because I take him to be writing of a Smithian entrepreneur taking time off to play golf rather than following the ‘divorce branch’.)
The set of ‘other’ models may be seen to subdivide again, between behaviourism, and something more purposeful associated with the work of Herbert Simon (‘don't maximize, Simonize!’). To be sure, the firms in Cyert and March wanted to make a profit: they just do not seem to have been very good at it. Along the ‘Simon branch’ we have purposeful, self-interested behaviour. We may call it rational too, as long as it is understood that optimization is thought to be too difficult, and it is accordingly rational not to try. It does not follow that optimization does not occur: firms may adopt a convergent process, as in Day and Tinney (1968). In a ‘sufficiently stable’ environment, convergence might, of course, be quite common. But convergence must be proved rather than optimization assumed. It is thought rational for the firm to adopt routines or standard operating procedures that work at least ‘well enough’. The meaning of ‘innovation’ is now extended. The introduction of a new routine that successfully handles a

complicated decision that has to be taken with limited information is as much an innovation as a new product or an improvement in the technology. (From this point of view, a new legal or financial instrument that reduces transactions costs is an innovation too.)

It would not, I think, be a good use of space to catalogue all Simon's own innovations and suggestions. (For more recent discussion of bounded rationality, and related matters, see March, 1978.) Instead, I shall consider only a recent contribution on this branch, the work of Nelson and Winter (1982) already referred to. These writers are much concerned with economic growth, perhaps less in static allocational problems. They inherit from Schumpeter, and Marshall, as well as Simon, and they name Cyert and March among their intellectual ancestors, as well as Alchian (1950).

Nelson and Winter argue that firms do not know the well-defined technological choice sets of standard theory. They only know how to do what they do do, and how to make at least local searches to do other things. Thus there is no sharp distinction between the choice set and the choice, and maximization is not an appropriate concept or mode of analysis. Neither is equilibrium for either firm or industry. The configuration of an industry at any time is seen as the outcome of an evolutionary process, whence the appropriate tool is a Markov process (as in Newman and Wolfe, 1961). The ‘genes’ required for biological analogy are the firms’ routines: the standard procedures (in production, marketing, finance, and so on) that it knows how to operate. Its environment is stochastic, and the firm continually has to search for new routines (mutations). Chance enters twice. The search for a new routine may be deliberate, but its success is subject to chance. Once discovered, its application is subject to chance. Thus we have purposeful, self-interested behaviour, but success is a matter of luck. Routines are inherited, but new routines, once discovered, may also be copied by others, which allows the evolutionary process to be much faster than the biological process. There is another important point here. Nelson and Winter show that it may be more profitable to wait and to copy an innovation made by others than to incur the expenses necessary to develop it oneself. This seems to be contrary to the Schumpeterian intuition. There is also a shift in focus from the ‘firm’. For Nelson and Winter the evolution of the industry is the subject of study, and the routines are the genes in the evolutionary process. The ‘firm’, although it is assumed to adopt purposeful, self-interested conduct (to seek profit), is not itself a matter of particular interest: it is something of a transient which happens, at any moment of time, to have inherited some routines, and may or may not succeed in developing some new, successful, ones. As in the earlier biological analogies, success will be rewarded and failure punished, but this is not advanced as an argument for ‘as if’ optimizing behaviour; it is part of the evolutionary process. Indeed, Nelson and Winter offer the first formal proof that, in this process, it is the profitable firms that survive. For other problems (R&D and technological change; Schumpeterian competition), they have to rely on simulation techniques which, however well handled, always leave one a little uncertain about what has been established, or, at least, at what level of generality.

It is now time to return to the question posed at the beginning of this article: what is the scope and purpose of the theory of the firm? Indeed, is there *a* theory of the firm at all? Perhaps not. There is a file of optimizing models. We may include in this file the theory of agency and much recent work on information and incentives. (There are also inquiries into such organizational matters as integration and the divisional structure of large corporations, which I do not discuss here.) In the ‘other’ branch, profit-seeking but not optimizing, there is the recent work by Nelson and Winter, in which the focus is on the development of the industry (population), and the firm is little more than an agent (unit organism) for the transmission of genes. And there is recent work, very exciting work, exploiting the ideas of capital

commitment and credible threats, much of it in the spatial literature, on the strategic behaviour of firms in small group situations. Much of this work has been associated with developments in game theory. I shall not describe it here on the possibly dubious grounds that it is better filed as ‘Industrial organization’ or ‘theory of market structure’. Demarcation lines are not, of course, well established; it could be argued that, whenever we invoke the ubiquitous Cournot–Nash equilibrium concept, we are taking a game-theoretic approach, and some might wish to interpret theory of the firm more widely than I have done. Be that as it may, there is clearly no such thing as *a* theory of the firm. But there is a great deal in the file, subdivide it as we will, and since the Second World War we have seen great advances, on many different fronts, albeit differently motivated and with different methodological orientations.

See Also

- advertising
- competition and selection
- corporations
- entrepreneurship
- ideal output
- marginal and average cost pricing
- market structure
- monopoly
- oligopoly
- predatory pricing
- price discrimination (theory)

Bibliography

Alchian, A.A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58, 211–12.

Alchian, A.A. and Demsetz, H. 1972. Production, information costs and economic organization. *American Economic Review* 62, 777–95.

Andrews, P.W.S. 1949. *Manufacturing Business*. London: Macmillan.

Aoki, M. 1984. *The Co-operative Game Theory of the Firm*. Oxford: Clarendon Press.

Arrow, K.J. 1971. *Essays in the Theory of Risk-Bearing*. Chicago: Markham.

Baumol, W.J. 1959. *Business Behavior, Value and Growth*. New York: Macmillan.

Berle, A.A. and Means, G.C. 1933. *The Modern Corporation and Private Property*. New York: Macmillan.

Brechling, F.P.R. 1975. *Investment and Employment Decisions*. Manchester: Manchester University Press.

Chamberlin, E.H. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press; London: Oxford University Press.

Coase, R.H. 1937. The nature of the firm. *Economica*, NS 4, 386–405.

Cournot, A.A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: L. Hachette. Trans. N.T. Bacon as *Researches into the Mathematical Principles of the Theory of Wealth*, London and New York: Macmillan, 1897.

Cyert, R.M. and March, J.G. 1963. *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice Hall.

Darwin, C. 1859. *On the Origin of Species*. London: Murray.

Day, R.H. and Tinney, E.H. 1968. How to cooperate in business without really trying: a learning model of decentralized decision making. *Journal of Political Economy* 76, 583–600.

Dorfman, R., Samuelson, P.A. and Solow, R.M. 1958. *Linear Programming and Economic Analysis*. New York: McGraw-Hill.

Drèze, J.H. 1985. Uncertainty and the firm in general equilibrium theory. *Economic Journal* 95 (Supplement), 1–20.

Eaton, B.C. and White, W.D. 1983. The economy of high wages: an agency problem. *Economica*, NS 50, 175–82.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. ed. M. Friedman. Chicago: University of Chicago Press.

Grossman, S.J. and Hart, O.D. 1983. An analysis of the principal–agent problem. *Econometrica* 51, 7–46.

Grossman, S.J. and Stiglitz, J. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70, 393–408.

Hall, R.L. and Hitch, C.J. 1939. Price theory and business behaviour. *Oxford Economic Papers* 2, 12–45.

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13, 324–40.

Jensen, M.C. and Meckling, W.H. 1976. Theory of the firm: managerial behaviour, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

Knight, F. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

Koopmans, T.C., ed. 1951. *Activity Analysis of Production and Allocation*. Cowles Commission Monograph No. 13. New York>: John Wiley.

Lucas, R.E. 1967. Optimal investment policy and the flexible accelerator. *International Economic Review* 8, 78–85.

Machlup, F. 1946. Marginal analysis and empirical research. *American Economic Review* 36, 519–54.

Machlup, F. 1967. Theories of the firm; marginalist, behavioral, managerial. *American Economic Review* 57, 1–33.

March, J.G. 1978. Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics* 9, 587–610.

Marris, R. 1964. *The Economic Theory of 'Managerial' Capitalism*. London: Macmillan.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Mill, J.S. 1848. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*. London: J.W. Parker.

Mortensen, D.T. 1970. A theory of wage and employment dynamics. In *Microeconomic Foundations of Employment and Inflation Theory*, ed. E.S. Phelps. New York: W.W. Norton.

Nelson, R.R. and Winter, S.G. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA and London: Harvard University Press.

Newman, P. 1960. The erosion of Marshall's theory of value. *Quarterly Journal of Economics* 74, 587–601.

Newman, P. and Wolfe, J.N. 1961. A model for the long-run theory of value. *Review of Economic Studies* 29, 51–61.

Penrose, E.T. 1952. Biological analogies in the theory of the firm. *American Economic Review* 42, 804–19.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Ross, S. 1973. The economic theory of agency: the principal's problem. *American Economic Review* 63, 134–9.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sandmo, A. 1971. On the theory of the competitive firm under price uncertainty. *American Economic Review* 61, 65–73.

Schumpeter, J.A. 1934. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Scitovsky, T. 1943. A note on profit maximization and its implications. *Review of Economic Studies* 11, 57–60.

Shapiro, C. and Stiglitz, J.E. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Simon, H.A. 1952. On the application of servomechanism theory in the study of production control. *Econometrica* 20, 247–68.

Simon, H.A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69, 99–118.

Simon, H.A. 1959. Theories of decision making in economics. *American Economic Review* 49, 253–83.

Simon, H.A. 1979. Rational decision making in business organizations. *American Economic Review* 69, 493–513.

Smith, A. 1776. *An Enquiry into the Nature and Causes of the Wealth of Nations*. London: W. Strahan and T. Cadell.

Stigler, G.J. 1939. Production and distribution in the short run. *Journal of Political Economy* 47, 305–27.

Stiglitz, J.E. 1974. Incentives and risk sharing in sharecropping. *Review of Economic Studies* 41, 219–55.

Treadway, A.B. 1969. On rational entrepreneurial behaviour and the demand for investment. *Review of*

Economic Studies 36, 227–39.

Viner, J. 1931. Cost curves and supply curves. *Zeischrift für Nationalökonomie* 3, 23–46.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Walras, L. 1874. *Eléments d' économie politique pure*. Lausanne: L. Corbaz; Paris: Guillaumin; Basle: H. Georg.

Whitin, T.M. 1953. *The Theory of Inventory Management*. Princeton: Princeton University Press.

Williamson, J. 1966. Profit, growth and sales maximization. *Economica*, NS 33, 1–16.

Williamson, O.E. 1964. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Williamson, O.E. 1970. *Corporate Control and Business Behavior*. Englewood Cliffs NJ: Prentice-Hall.

Williamson, O.E. 1980. The organization of work: a comparative institutional assessment. *Journal of Economic Behaviour and Organization* 1, 5–38.

Winter, S.G. 1975. Optimization and evolution in the theory of the firm. In *Adaptive Economic Models*, ed. R.H. Day and T. Groves. London and New York: Academic Press.

How to cite this article

Archibald, G. C. "firm, theory of the." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000114> doi:10.1057/9780230226203.0577

The New Palgrave Dictionary of Economics Online

firm-level employment dynamics

Jeff Campbell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Firm-level employment dynamics deals with the evolution of firms' employment decisions when they face costs of creating and destroying jobs. It lies at the intersection of labour economics, industrial organization and macroeconomics. Recent contributions use rational expectations models of labour demand to match salient statistics from establishment-level employment records.

Keywords

adjustment costs; firm-level employment dynamics; job creation and destruction; structured and unstructured jobs

Article

Firm-level employment dynamics is the branch of economics that deals with the evolution of firms' employment decisions. The static analysis of labour demand equates a firm's marginal product of labour with the wage. Observation suggests that this abstracts from important considerations of employers when expanding or contracting their firms. Recruiting new employees requires effort, and preparing them for the jobs at hand might require training. Employees with substantial tenure might have legal rights that make their dismissal costly. All these realistic constraints make a firm's current employment complementary with its level at any future date. Hence, a firm's employment decisions when properly considered are *dynamic*. Firm-level employment dynamics is the area of economics that seeks to understand this decision using both theory and measurement. It lies at the intersection of industrial organization, labour economics and macroeconomics. It shares with labour economics a central concern with the employment relationship. Because firm entry and exit plays a substantial role in the evolution of total employment, it shares with industrial organization an interest in entrepreneurship. Firm-level costs of adjusting employment provide one potential source of persistence in economy-wide employment, so the area has contributed to the macroeconomics of business cycles.

Early theoretical treatments of the firm's dynamic employment choices came from the Ph.D. theses of Oi and Rosen. Oi (1962) coined the adjective 'quasi-fixed' to describe a factor of production that could be changed from its previous value at a price. He considered the labour demand of a firm facing a constant wage and interest rate that must incur recruiting and training costs when expanding employment. Denote these with W, r, and τ. Then the firm's first-order condition for labour is $P \times f'(N) = W + r \times \tau$, where the production function holding other inputs constant is $f(\cdot)$, P is the output price, and N is the firm's employment choice. Oi noted two fundamental implications of this equation. First, the marginal product of labour generally exceeds the wage, so wage-setting institutions must support such a gap if the firm is to recover its investment in job creation. Second, unexpected permanent reductions in P leave N unchanged so long as the marginal product of labour remains above the wage. Rosen (1968) extended this by noting that adjusting an employee's hours worked generally costs less than adding or dismissing workers. Hence, fluctuations in hours worked should be more important for workers with high training and recruiting costs. He examined the employment decisions of regulated railways and found that they conform to this pattern.

Oi and Rosen intuitively saw many of the fundamental theoretical implications of imposing labour adjustment costs, but the first fully dynamic treatment of the firm's labour demand curve came from macroeconomics. Sargent (1978) considered a firm with a quadratic production function maximizing profit subject to quadratic costs of adjusting employment. Given a stochastic wage, W_t, he showed that the firm's optimal labour demand curve takes the form

$$N_t = (1-\rho)N_{t-1} - \theta\mathbb{E}_t\left[\sum_{j=0}^{\infty}\lambda^j W_{t+j}\right].$$

Here, N_t is the firm's employment at date t, $\mathbb{E}$ is the expectations operator, and ρ and λ are positive parameters that depend on the interest rate, the cost of adjustment, and the production function's concavity. Intuitively, N_{t-1} influences the profit-maximizing choice of N_t because it changes the cost of achieving any given level of employment. The complementarity between current and future employment makes N_t a function of the wage and its expected value at all future dates. This rule for N_t aggregates easily: if all firms follow it, then total employment does so as well. Sargent estimated this using US data on private employment and real wages. His model considered both straight-time and overtime employment. The quarterly data did not contain substantial evidence against the model, but the response of employment to real wage fluctuations in the model was much less than that measured with a vector autoregression.

With quadratic costs of adjustment, firms smooth their employment adjustments across time. This conflicts with the casual observation that firms' employment adjustment is *lumpy*, that is, they alternate between periods with very little or no employment adjustment and others with high rates of hiring or firing. Further credence to that view came from observations of firm employment collected by national statistical agencies. Using plant-level employment observations from the Dutch economy in 1988 and 1990, Hamermesh, Hassink and van Ours (1996) showed that 28.3 per cent of firms kept employment constant over that two-year period. All these firms changed the *identities* of their employees. The average hiring rate for these firms was 11.3 per cent, so they apparently face costs of changing the jobs in the firm that are independent of the costs of changing the workers filling them.

With their book-length study of plant-level employment dynamics in the US manufacturing sector Davis, Haltiwanger and Schuh (1996) reinforced the conclusion that firm-level employment adjustment is lumpy. Their data came from the Longitudinal Research Database, an unbalanced panel of quarterly firm-level employment observations created from the surveys underlying the Annual Survey of Manufacturers and the Census of Manufacturing. These are confidential US Census records, but they may be used for approved projects that benefit the US Census at one of several regional census research data centres. Denote the employment of firm i in quarter t with N_{it}, and let $N_t = \sum_{i=1}^{M_t} N_{it}$ be the employment of the M_t firms with positive employment in quarter t. With these data, Davis, Haltiwanger and Schuh defined the job creation and destruction rates as

$$POS_t \equiv 2\times\sum_{it} I\{N_{it} > N_{it-1}\}\frac{N_{it}-N_{it-1}}{N_t+N_{t-1}} NEG_t \equiv 2\times\sum_{it} I\{N_{it} < N_{it-1}\}\frac{N_{it-1}-N_{it}}{N_t+N_{t-1}}.$$

So defined, the difference between these two rates equals the rate of employment growth. These authors refer to their sum as employment reallocation.

The examination of these statistics from 1972: IV to 1988: IV yielded the following conclusions. (1) The rates of job creation and destruction are both very large relative to total employment changes. The average annual rates of job creation and destruction equalled 9.1 and 10.3 per cent. (2) The job creation and destruction rates of the population of young and middle-aged plants (less than ten years old) are much higher than those of their older counterparts. (3) Plants' employment changes are persistent. Some 70 per cent of newly created jobs last at least one year, and 80 per cent of newly destroyed jobs fail to reappear within a year. (4) Employment adjustment is lumpy. Two-thirds of job creation and destruction occurs at plants that adjust their employment by 25 per cent or more. (5) Employment drops in a recession because job destruction increases. Job creation is relatively acyclical. Together, these facts have become the empirical touchstone for firm-level employment dynamics.

These facts motivated the creation of new models of firms' employment choices that incorporated adjustment costs and lent themselves to aggregation. A pair of papers by Campbell and Fisher (2000; 2004) develops one such model and applies it to explain the job creation and destruction facts of Davis, Haltiwanger, and Schuh. They begin with the labour demand problem of a single plant that produces a homogenous good for sale in a competitive market. The plant uses one factor of production, labour, that comes in fixed shift lengths. The per-period cost of employment measured in units of the output price is W_t, and let n_t denote employment at this plant. The plant's output in period t is $z_t n_t^{\alpha}$, where z_t is the plant's idiosyncratic productivity term and $0 < \alpha < 1$. The wage follows a Markov chain over $\{W_l \le W_h\}$ with transition probability p, and the idiosyncratic productivity shock follows a random walk with bounded innovation ε_t. The production function's strict concavity could arise from limits to a manager's effective span of control. When the plant changes its

employment, it incurs adjustment costs that are proportional to the number of jobs created or destroyed. If employment at the plant expands, the cost per job created in units of lost output is τ_c, and the analogous cost per job destroyed is τ_d.

With these primitives, the profit maximization problem for a plant manager discounting future profits with β can be represented as a dynamic programming problem with initial states n_{t-1}, z_t, and W_t. Its associated Bellman equation is

$$v(n_{t-1}, z_t, W_t) = \max_{n_t} z_t n_t^{\alpha} - W_t n_t - \tau(n_t, n_{t-1})(n_t - n_{t-1}) + \beta \mathbb{E}_t\left[v(n_t, z_{t+1}, W_{t+1})\right].$$

Here, $\tau(y, x) \equiv \tau_c \times I\{y > x\} - \tau_d \times \{y < x\}$ is the per-job adjustment cost incurred. Campbell and Fisher (2000) show that the plant's optimal employment policy has a very simple structure. There exist job creation and destruction schedules, $\underline{n}(z, W) = \underline{y}(W) z^{1/(1-\alpha)}$ and $\overline{n}(z, W) = \overline{y}(W) z^{1/(1-\alpha)}$, such that

$$n_{t+1} = \begin{cases} \underline{n}(z, W) & \text{if } n_{t-1} \le \underline{n}(z, W) \\ n_{t-1} & \text{if } \underline{n}(z, W) < n_{t-1} < \overline{n}(z, W) \\ \overline{n}(z, W) & \text{if } \overline{n}(z, W) \le n_{t-1} \end{cases}.$$

Figure 1 illustrates these policies. On its horizontal axis is ln z_t, while its vertical axis gives ln n_{t-1} and ln n_t. The three plants labelled *A*, *B*, and *C* all start with identical values of n_{t-1} but different values of z_t. The job creation and destruction schedules are both linear with slopes equal to $1/(1-\alpha)$. Plant *A* lies above the job destruction schedule, so it reduces employment. Plant *C* lies below the job creation schedule, so it creates jobs. Plant *B* lies between the two schedules. Here, the costs of job creation and destruction both exceed their associated benefits, so the plant's optimal employment is unchanged. Thus, this model automatically replicates one of Hammermesh, Hassink, and van Ours's findings: the plant's optimal employment frequently does not change.

Figure 1

Optimal employment policy in Campbell and Fisher (2000)

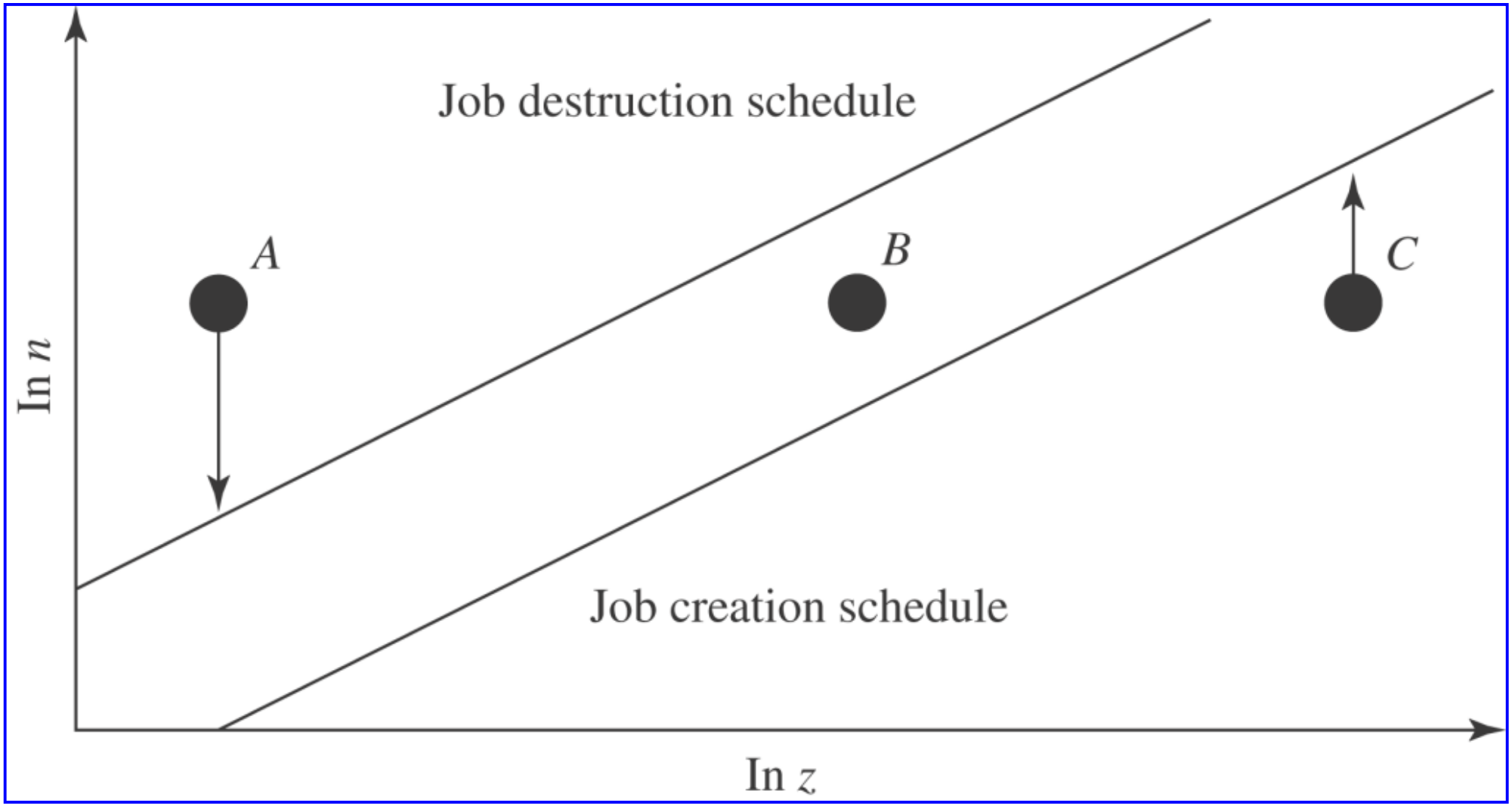

Davis, Haltiwanger, and Schuh's finding that job destruction accounts for most cyclical employment variation attracted a great deal of attention in macroeconomics. Campbell and Fisher (2000) show that this simple model can replicate that fact if employment fluctuations arise from variation in W_t. To appreciate how this can be, note that the total cost of creating a job is $W_t + \tau_c$, which has an elasticity with respect to W_t that is less than one. The total cost of destroying a job is $W_t - \tau_d$, so its elasticity with respect to W_t exceeds 1. This asymmetry in the costs of job creation and destruction translates into asymmetric responses of the job creation and destruction schedules to changes in W_t. When the model is calibrated to match the characteristics of a typical US manufacturing industry, this microeconomic asymmetry produces the observed aggregate dynamics in a large population of such plants: the variance of job destruction exceeds the variance of job creation.

Campbell and Fisher (2004) extend this model to address Davis, Haltiwanger, and Schuh's finding that the magnitude of job creation and destruction declines with a plant's age and a related fact: *aggregate* fluctuations in young and middle-aged plants' employment exceed those of employment at older plants. To do so, they incorporate a life cycle into the above model. Plants exit exogenously and are instantly replaced by new entrants. All entrants begin life in a 'volatile' state with high probability of exit and high idiosyncratic productivity variance. In each period, a plant has a constant probability of transiting to a 'stable' state with lower exit probability and idiosyncratic productivity variance. Alone, this change would (mechanically) replicate the finding that young plants display greater job creation and destruction rates than their older counterparts. To generate young plants' greater business-cycle sensitivity, Campbell and Fisher add 'unstructured' jobs. Creating and destroying these jobs is costless, but for a worker to fill such a job is less productive than filling a structured job, which is costly to create and destroy.

Intuition suggests that a plant's use of unstructured jobs depends on its position in the life cycle. Young firms face high uncertainty about their future productivity and survival, so they find unstructured jobs more attractive than their older more predictable counterparts. This is indeed the case. In the calibrated version of the model that Campbell and Fisher use, firms in the 'mature' life-cycle stage never use unstructured jobs. Their employment dynamics qualitatively mimic those in the simpler model. In contrast, young plants' greater uncertainty induces them to create fewer structured jobs. This increases the marginal product of labour and thereby makes creating unstructured jobs more attractive. Figure 2 illustrates such young firms' employment choices. As in Figure 1, the four plants labelled *A*, *B*, *C*, and *D* all have the same previous employment in *structured* jobs. Job creation and destruction schedules govern these plants' choices of structured jobs. These are the figure's solid lines. The dashed line gives the optimal employment in unstructured jobs if structured jobs were not available. Plants *A* and *B* do not use unstructured jobs, because they lie above this frictionless labour demand schedule. Plants *C* and *D* lie below it, and so they employ workers in both structured and unstructured jobs. Plants *B* and *C* both lie between the job creation and destruction schedules, so small changes in productivity induce neither of them to change their employment in structured jobs. However, only plant *B* would keep total employment constant. Plant *C* would change its employment in unstructured jobs following a small change in z_t. In this sense, the greater uncertainty young plants face leads them to choose more flexible production structures. Campbell and Fisher show in their calibrated version of this model that this greater microeconomic flexibility leads to larger aggregate responses to aggregate productivity shocks. Thus, the microeconomic differences between plants at different stages of the life cycle lead directly to the different aggregate differences in their employment dynamics.

Figure 2

Employment choices when structured and unstructured jobs are used

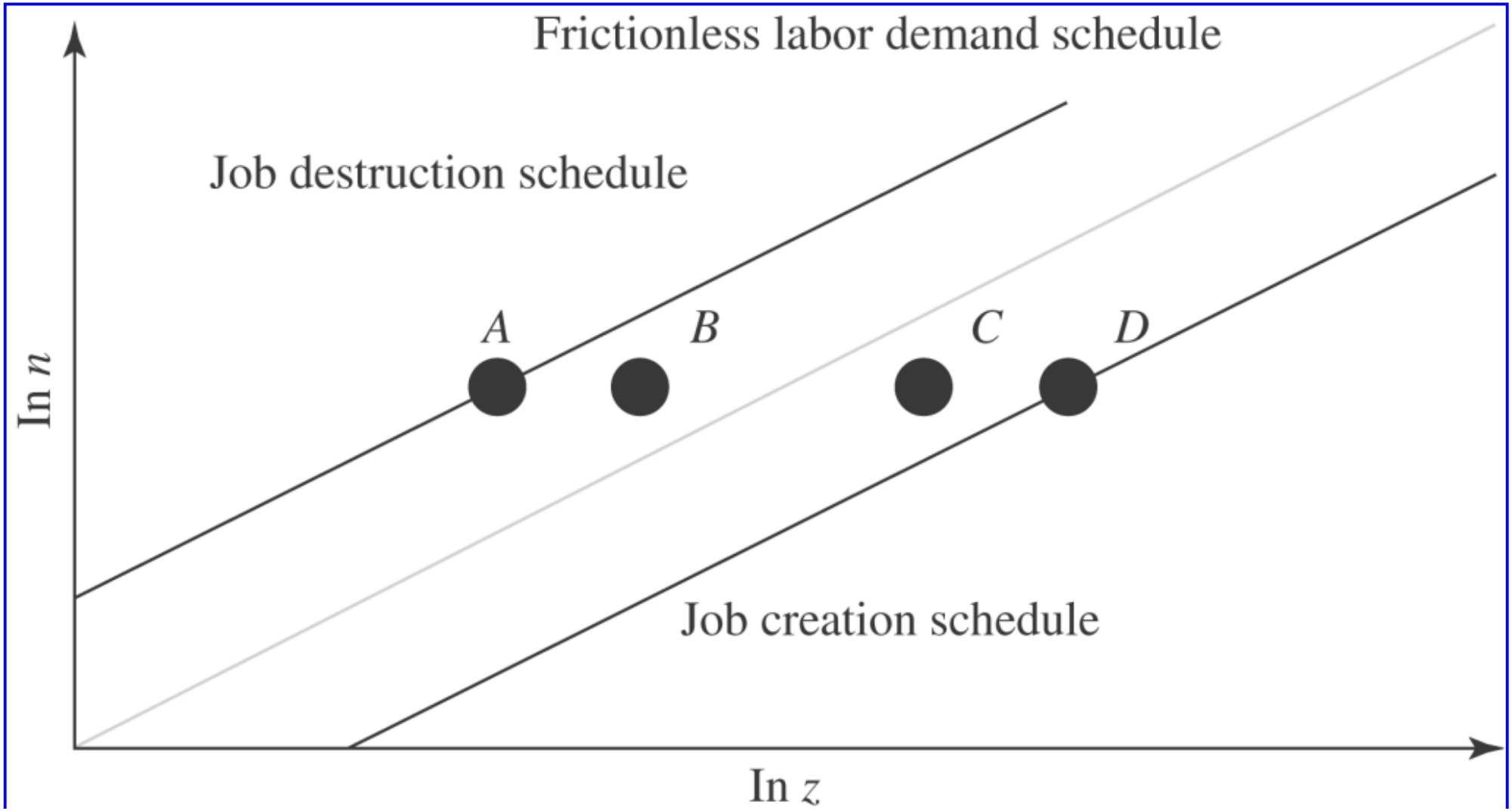

One aspect of firm-level employment dynamics not captured by Campbell and Fisher's models is the prevalence of very large employment adjustments. To generate this, Bentolila and Bertola (1990) add fixed costs of employment adjustment. This non-convexity complicates the model's analysis, but under certain conditions a plant's optimal employment policy follows a two-sided version of an (S, s) policy familiar from inventory models. Denote the gap between a plant's actual employment and its optimal value without adjustment costs using g_t. Then the firm lowers the gap to the target u by destroying jobs whenever it would otherwise exceed the trigger U, and it raises it to the target l by creating jobs whenever it would otherwise fall below the trigger L. Campbell and Fisher's model can be written in this form, where $u=U$ and $l=L$. Fixed costs of employment adjustment cause the targets to differ from their associated triggers and induce the firm to make only large employment adjustments.

Research on firm-level employment dynamics currently examines areas far removed from the initial focus on US manufacturing. Foote (1998) and Campbell and Lapham (2004) examine the dynamics of employment in service and retail industries. Foote finds that job creation dominates aggregate employment fluctuations in these industries. Consistent with this, Campbell and Lapham find that retail industries expand employment following a demand shock by increasing net entry. The importance of entrepreneurship for retail industries' employment is intuitive, and it suggests that the empirical and theoretical lessons learned from studying manufacturing industries will not apply easily to this important sector.

See Also

- adjustment costs
- aggregation (production)
- business cycle measurement
- firm boundaries (empirical studies)
- Rosen, Sherwin

Bibliography

Bentolila, G. and Bertola, S. 1990. Firing costs and labour demand: how bad is eurosclerosis? *Review of Economic Studies* 57, 381–402.

Campbell, J.R. and Fisher, J.D.M. 2000. Aggregate employment fluctuations with microeconomic asymmetries. *American Economic Review* 90, 1323–45.

Campbell, J.R. and Fisher, J.D.M. 2004. Idiosyncratic risk and aggregate employment dynamics. *Review of Economic Dynamics* 7, 331–53.

Campbell, J.R. and Lapham, B. 2004. Real exchange rate fluctuations and the dynamics of retail trade industries on the U.S. – Canada border. *American Economic Review* 94, 1194–206.

Davis, S.J., Haltiwanger, J.C. and Schuh, S. 1996. *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Foote, C.L. 1998. Trend employment growth and the bunching of job creation and destruction. *Quarterly Journal of Economics* 113, 809–34.

Hamermesh, D.S., Hassink, W.H.J. and van Ours, J.C. 1996. Job turnover and labor turnover: a taxonomy of employment dynamics. *Annales d'Economie et de Statistique* 41/42, 21–40.

Oi, W. 1962. Labor as a quasi-fixed factor. *Journal of Political Economy* 70, 538–55.

Rosen, S. 1968. Short-run employment variation on class-I railroads in the U.S., 1947–1963. *Econometrica* 36, 511–29.

Sargent, T.J. 1978. Estimation of dynamic labor demand schedules under rational expectations. *Journal of Political Economy* 86, 1009–44.

How to cite this article

Campbell, Jeff. "firm-level employment dynamics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000315> doi:10.1057/9780230226203.0576

The New Palgrave Dictionary of Economics Online

fiscal and monetary policies in developing countries

David Fielding

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Low levels of economic development constrain fiscal and monetary policy in several ways. Few developing countries are able to raise much direct tax revenue, and so must rely on other sources of funding, including seigniorage. Institutional constraints often lead to a high risk of hyperinflation and currency crises. Credible, effective institutions can be created with appropriate outside help, but there are few examples of this in practice.

Keywords

budget deficits; central bank independence; currency board; currency crises; developing countries; development economics; direct taxation; exchange rate peg; fiscal policy in developing countries; foreign aid; heterodox macroeconomics; inflation targeting; inflation; Laffer Curve; monetarism; monetary policy in developing countries; monetary unions; Phillips Curve; public expenditure; seigniorage; tariffs; Taylor Rules; time consistency

Article

Policymakers in developing (low-income, semi-industrialized) countries face particular challenges when setting taxes, interest rates and quantitative monetary instruments. All that follows should be preceded by a caveat: developing countries encompass at least as much economic diversity as the OECD. There is no such thing as a representative developing country; much harm can be (and has been) done by the incautious application of stylized models from development macroeconomics to individual countries. Nevertheless, we can identify those characteristics of developing countries that are likely to impose severe constraints on macroeconomic policy-making.

Fiscal and monetary characteristics of developing countries

The descriptive statistics in Table 1 provide some insight into the ways in which the fiscal and monetary characteristics of many developing countries differ from those of the developed world. The first row of the table contains information about fiscal structure and financial development in the United States. Subsequent rows show equivalent average figures for those low-income countries for which data are available.

Selected descriptive statistics for 2000

	Direct taxes (% of total taxes)	Import tax (% of total taxes)	Total taxes (% of GDP)	M1/M2 (%)	M2/GDP (%)
United States	61	01	20	24	60
*Countries with per capita GNI <$10K**					
Africa	27	33	19	55	33
Americas	21	10	13	23	29
Asia	16	11	09	26	30
Europe	11	02	16	30	17
*Countries with per capita GNI <$5K**	21	21	15	47	34
*Countries with per capita GNI of $5–10K**	24	11	20	37	37

*The *per capita* gross national income (GNI) figures are PPP-adjusted. The averages are constructed from those countries for which complete data are available in World Bank (2003): Algeria, Bolivia, Bulgaria, Congo Republic, Costa Rica, Cote d'Ivoire, Croatia, Dominican Republic, El Salvador, Estonia, Georgia, India, Iran, Jamaica, Jordan, Kazakhstan, Latvia, Lithuania, Madagascar, Mauritius, Mexico, Moldova, Mongolia, Nepal, Pakistan, Paraguay, Peru, Philippines, Poland, Romania, Russia, Sri Lanka, St. Vincent, Swaziland, Tajikistan, Thailand, Tunisia, Turkey, Uganda, Ukraine, Uruguay, Venezuela and Vietnam. Russia and Turkey are included in the figures for Europe.

The table indicates some of the structural differences between the developing country average and the United States:

- Direct taxation in developing countries makes up a much smaller fraction of total tax revenue, and import duties make up a larger fraction.
- In Asia and the Americas, total tax revenue makes up a substantially smaller fraction of GDP.
- In Africa and Europe, M1 makes up a much larger fraction of M2.
- M2 makes up a much smaller fraction of GDP.

All these features are more pronounced for countries with a per capita gross national income below $5,000 than for those countries in the $5,000–$10,000 range.

The low levels of direct taxation in developing countries reflect that fact that a large fraction of private sector income is non-monetized: for example, many peasant households grow subsistence crops for their own consumption. Even when income is monetized, the administrative costs of direct taxation are often

relatively high because of, for example, low levels of literacy and limited information technology. Governments are therefore forced to rely to a much greater degree on seigniorage revenue and on import duties. (High tariffs are often motivated by the need for fiscal revenue rather than by import substitution.) Inflation in developing countries is usually far higher than in the OECD. Between 1990 and 2000, the average annual inflation rate for the median developing country in Table 1 was 46 per cent; only five countries had single-digit inflation, and 14 had average inflation rates over 1,000 per cent per annum.

The low levels of broad money demand in developing countries reflect low savings rates and limited access to financial services. Commercial banks are often absent from rural areas, where low per capita income, low population density and poor transport and communication infrastructure entail high costs in financial service provision to individual customers. In many developing countries a large fraction of the total money stock is in the form of cash, and few households have access to interest-bearing assets. One consequence is that the interest elasticities of saving and money demand are often very low; another is that there is limited scope for absorption of public debt by the domestic private sector.

Monetary policy: the neoclassical perspective

Central to the monetarist approach to development macroeconomics is the argument that in developing countries high inflation is always and everywhere a fiscal phenomenon. Agénor and Montiel (1999) provide an extensive survey of this approach. In the standard formulation of the argument, which embodies many of the constraints highlighted above, there is a Cagan money demand function:

$$M / P = \exp(-\alpha \cdot \pi^{\beta}) \tag{1}$$

where M is the nominal money stock, P is the price index and $\pi = \dot{P} / P$. M is to be interpreted as narrow money. There are no interest-bearing assets and no interest elasticity of money demand, so the opportunity cost of holding money depends just on the inflation rate. There is also a fixed real budget deficit, D, financed entirely be seigniorage:

$$D = \dot{M} / P = [M / P] \cdot \mu \tag{2}$$

where $\mu = \dot{M} / M$. This reflects the government's limited access to tax revenue and domestic credit. Combining equations (1) and (2) we have:

$$D = \exp(-\alpha \cdot \pi^{\beta}) \cdot \mu \tag{3}$$

Equation (1) entails that an equilibrium with a constant π requires $\pi = \mu$, so that *M*/*P* is constant. For low enough values of *D* there will be two such equilibria, solutions to equation (3) with $\pi = \mu$. But for high values of *D* there is no equilibrium: successively higher levels of inflation lead to lower levels of real money demand, requiring higher rates of monetary expansion to finance the budget deficit, and so yet more inflation.

The first goal of macroeconomic policy is therefore to reduce the budget deficit to a level compatible with a stable inflation rate. In the absence of alternative sources of revenue this entails a reduction in public expenditure, which may have a negative impact on social and economic development. This provides a rationale for foreign aid to subsidize public expenditure in the medium term, while the country develops the institutions that will facilitate a wider fiscal base and a financial sector that will support some public debt. This approach still views the main macroeconomic function of a central bank in a developing country as generating seigniorage revenue. Policy reform is intended to reduce seigniorage, not to zero, but to a range compatible with a stable inflation rate. Indeed, a part of the neoclassical development macroeconomics literature analyses the inflation tax using concepts explicitly drawn from public finance, for example the Laffer Curve. The use of monetary policy for business cycle stabilization is at most a secondary objective.

Time consistency in monetary policy

The critique of Kydland and Prescott (1977) can readily be applied to a seigniorage model. The simple model above provides an extreme case. Consider a policymaker for whom *D* is a variable to be maximized, subject to the equilibrium condition that $\pi = \mu$. From equation (3), the optimal rate of monetary expansion is $[\alpha \cdot \beta]^{-1/\beta}$; rates higher than this will reduce revenue. But if we modify equation (1) so that current money demand is based on a predetermined *expectation* of inflation, then for a given expectation and a given level of money demand the optimum inflation rate is infinite. The rational expectation of inflation is therefore infinite, in which case money demand and revenue are zero. The policymaker's problem is how to pre-commit credibly to a rate of expansion equal to $[\alpha \cdot \beta]^{-1/\beta}$. Failure to solve this problem is one suggested reason for the failure of disinflation programmes in developing countries.

The standard solution to the time inconsistency problem in industrialized countries is to delegate control of monetary policy to a central bank governor with a contract to target a given inflation rate. The constraint facing many developing countries is the absence of a political tradition or political institutions that will give people confidence in any laws enacted to create central bank independence. Evidence on the link between central bank independence and inflation in developing countries is very weak. There is no significant correlation between historical inflation rates and historical indices of independence, which are based on the assumption that laws in developing countries have the same force as those in

industrialized countries (Cuckierman, Webb and Neyapti, 1992). One interpretation of these results is that legislation for central bank independence would of little use in many developing countries, either because independence de jure does not entail independence de facto or because underdeveloped political institutions are unable to deliver enough clarity or stability in the decision-making process to allay people's doubts.

A possible alternative to central bank independence legislation is commitment to a credible nominal exchange rate peg. For a given real exchange rate a fixed peg against, for example, the euro or US dollar delivers an inflation rate equal to that of the eurozone or the USA. However, a fixed peg will be credible in the long run only if it is accompanied by an appropriate rate of domestic monetary expansion. Excessive domestic monetary expansion will lead to persistent balance of payments deficits and a loss of official foreign exchange reserves; eventually this will cause a collapse in the demand for domestic currency.

‘First generation’ currency crisis models show that with excessive monetary expansion this collapse can happen long before official reserves are finally depleted (Flood and Garber, 1984). It would be irrational to hold on to domestic currency until reserves were finally depleted: at that point there would be a discrete fall in the value of domestic currency as the exchange rate shifted to a market value unsupported by central bank intervention, and those left holding domestic currency would make a loss. Instead, people will offload domestic currency as soon as monetary expansion has driven the implicit market value without intervention below the pegged rate.

‘Second generation’ models (Obstfeld, 1996) go a step further, explaining currency crises in cases where there is moderate monetary growth, no greater than the rate of growth of the supply of foreign currency. Private sector views on the probability of an imminent abandonment of a peg will depend on an assessment of the likely costs and benefits of the peg for the government. (One example of such a scenario is when seigniorage revenue is higher under a more flexible exchange rate regime, but such flexibility deters foreign investment.) But these views will themselves influence the current level of demand for foreign and domestic currency, and so the opportunity cost of maintaining the peg. Models of such an environment typically imply the existence of multiple equilibria. There may be an equilibrium with a low perceived probability of collapse and a low opportunity cost of maintaining the peg, but this equilibrium is unlikely to be globally stable. The feedback between the perceived probability of collapse and the true opportunity cost of the peg means that some rumour questioning the government's commitment to the peg, however small and baseless, could eventually undermine this commitment, regardless of its fiscal and monetary discipline.

Table 2 illustrates some cases in which monetary and fiscal discipline has not been sufficient for the maintenance of an exchange rate peg. In the three cases shown, the size of the budget deficit in the years prior to collapse was not an excessive fraction of GDP (Bolivia, Honduras), or else seigniorage revenue did not account for a large fraction of the deficit (Zambia).

Budget deficits and seigniorage revenue in the run-up to the abandonment of an exchange rate peg

	Bolivia (T=1982)		Honduras (T=1990)		Zambia (T=1981)	
	Deficit/GDP	Seigniorage/deficit	Deficit/GDP	Seigniorage/deficit	Deficit/GDP	Seigniorage/deficit
T–3	0.074	0.151	0.036	0.415	0.144	0.040
T–2	0.079	0.401	0.030	0.336	0.091	0.058

T–1	0.204	0.240	0.033	0.525	0.185	0.051

Note: The peg in each country was abandoned in year T.

Sources: IMF (1983; 1999)

One interpretation of such examples is that they emphasize the need for monetary institutions free from all domestic political pressures and whose commitment to monetary discipline is without doubt. Lost seigniorage revenue is not the only cost of an exchange rate peg. An appreciation of the euro or US dollar due to idiosyncratic shocks in the eurozone or the USA is likely to create a recession in any country pegging to one of these currencies. These recessions can make the peg very unpopular. If we ignore the question of whether such unpopularity is justified, one suggested route to the creation of politically independent monetary institutions is the establishment of currency boards. In a currency board system the central bank is legally required to back issue of domestic currency one-for-one with reserves in a given foreign currency. In eastern Europe currency boards have met with some success, at least in terms of maintaining a fixed peg. Recent examples are Bosnia–Herzegovina and Bulgaria pegging to the euro, and Latvia and Lithuania pegging to the US dollar. By contrast, the currency board system in Argentina met with spectacular failure, showing that currency board systems can be abandoned with almost as much ease as a conventional fixed peg.

A second suggested route to independent monetary institutions is the formation of monetary unions. A transnational central bank may well be free from many of the political pressures facing the central bank of a single country. In order to exert any political pressure on their central bank, the governments (and populations) of a monetary union would need to coordinate their actions. At any one time, conflicting economic interests are likely to undermine coordination attempts. It is always possible to secede from a monetary union, but in the absence of existing national monetary institutions this is potentially very costly. Given the ill will that secession is likely to generate among the remaining members of the union, it is also likely to be an irreversible decision, unlike the abandonment of a currency board. This irreversibility is likely to deter governments from abandoning their commitment to the monetary union.

Currently, there are three major monetary unions among developing countries. These are the East Caribbean Currency Union (ECCU: Anguilla, Antigua, Dominica, Grenada, Montserrat, St Kitts, St Lucia, St Vincent), the West African Economic and Monetary Union (UEMOA: Benin, Burkina Faso, Côte d'Ivoire, Guinea–Bissau, Mali, Niger, Senegal, Togo) and the Economic and Monetary Community of Central Africa (CEMAC: Cameroon, Central African Republic, Chad, Congo Republic, Equatorial Guinea, Gabon). The ECCU has for decades maintained a fixed peg to the US dollar with a currency board arrangement. The two African monetary unions have maintained a fixed peg against the French franc (and now the euro) since the member states' independence in the 1960s, with just one devaluation in 1994. All three monetary unions have maintained low and stable rates of inflation.

However, it is unlikely that these three monetary unions could easily be replicated elsewhere. The ECCU is a group of small island economies where tourism makes up a large fraction of GDP and the US dollar circulates freely anyway; the monetary institutions just formalize pre-existing dollarization. The African monetary unions maintain a peg in cooperation with the French government. The French treasury exchanges euros for the two African currencies at a fixed rate, so the peg does not constrain the two central banks' use of domestic monetary instruments in the short run. (There are rules to prevent excessive monetary expansion in the long run.) Moreover, the French provide overdraft facilities to the two central banks to help cushion balance of payments shocks. So when the euro appreciates because of

macroeconomic shocks specific to Europe, the African countries are not obliged to live through a recession. The feasibility of the peg has relied on an unusually strong (and arguably neo-colonial) economic commitment from the country issuing the anchor currency. Otherwise, it is likely that countries without credible domestic monetary institutions can buy a low inflation rate only at the cost of a fixed peg that periodically generates damaging recessions. Even if ‘second generation’ currency crises can somehow be averted, weak domestic monetary institutions and incomplete information about policymakers’ preferences mean that any relaxation of the exchange rate regime in times of recession will undermine the credibility of the commitment to low inflation.

Monetary policy: alternative perspectives

There is a body of literature that encompasses alternatives to the monetarist approach discussed above. This literature is often labelled ‘heterodox’ in the context of policy formation and ‘structuralist’ in the context of theoretical models, of which Cardoso (1981) is a good example. At its core is the idea that inflationary spirals can be generated by the wage- and price-setting institutions within an imperfectly competitive economy, with the supply of money responding passively to increases in prices.
Suppose for example that industrial prices (p) in a closed economy are set by monopolistic firms as a mark-up on nominal industrial wages (w):

$$p = [1 + \theta] \cdot w \tag{4}$$

Workers would like to maintain a fixed real wage, so in equilibrium the ratio of nominal wages to consumer prices will be fixed. If consumption is made up of industrial goods and non-industrial goods in fixed proportions (ϕ, $1-\phi$), then the constant real wage condition can be written as:

$$w = \eta \cdot [\varphi \cdot p + (1 - \varphi) \cdot q] \tag{5}$$

where η is the target real wage and q is the price of non-industrial goods. (The closed economy and fixed consumption share assumptions are not essential to this class of model.) Together equations (4–5) pin down relative prices:

$$p / q = \frac{1 - \varphi}{[\eta \cdot (1 + \theta)]^{-1} - \varphi} \tag{6}$$

There is a positive relationship between relative prices and the target real wage. Now if the supply of non-industrial goods depends just on relative prices, equation (6) will pin down non-industrial production and hence also, with full employment and a given level of resources, industrial production. In general, these production shares will not be equal to the consumption shares ϕ and $1-\phi$, in which case the model is overdetermined and has no equilibrium. An inflationary spiral will exist if non-industrial prices adjust to clear goods markets at a level of *p/q* less than $[1-\varphi]/\{[\eta\cdot(1+\theta)]^{-1}-\varphi\}$, which entails a real wage less than η. In such a world workers will raise wage demands, so $\dot{w}/w>0$, but from equation (4) $\dot{p}/p=\dot{w}/w$ and with non-industrial prices adjusting to maintain the initial level of *p/q* we also have $\dot{q}/q=\dot{p}/p$, so there is no change in the real wage; nominal wages and prices will rise indefinitely.

Various policy prescriptions follow from such a model. A government-imposed nominal industrial wage freeze will halt the inflationary cycle at no cost to industrial workers, since their real wages are constant for all levels of inflation. Alternatively, subsidizing consumption of the non-industrial good will raise the real wage and reduce inflationary pressure. This is the macroeconomic basis for arguments in favour of food subsidies. One criticism of subsidies is that they increase the size of the budget deficit, so in models that integrate monetarist and structuralist elements the impact of subsidies on inflation is ambiguously signed.

Evidence on the effectiveness of heterodox anti-inflation measures, compared with orthodox fiscal and monetary contraction, is very limited. Many of the high-profile programmes designed to tackle hyperinflation in the 1980s (for example, Argentina and Israel in 1985, Brazil in 1986 and Mexico in 1987) combined orthodox fiscal and monetary reforms with heterodox wage and price controls of one kind or another. It is very unclear which elements of these programmes were crucial in determining their success or failure. There is some limited evidence from reduced-form macro-econometric models on the direction of causality between wage growth, price growth and money growth (for example, Montiel, 1989). This suggests that different macroeconomic processes are at work in different countries. On the basis of current evidence, policy prescriptions for any one country should be accompanied by a large caveat.

Taylor rules in developing countries

The discussions above relate to the problems developing countries face in achieving a stable fiscal policy environment with a moderate rate of monetary growth. This has been the main focus of the theoretical and empirical literature to date. However, there is also a growing literature that extends the mainstream concerns of the monetary policy literature in OECD countries – in particular, issues surrounding the optimal policy response to exogenous macroeconomic shocks – to developing countries. Certainly, developing countries are at least as vulnerable to external shocks as OECD countries. Many developing countries are small in size, trading a relatively large fraction of their GDP and exporting a narrow range of primary commodities for which world prices are highly volatile. In these countries, yearly changes in the terms of trade can increase or reduce domestic income by several percentage

points. The average value of such changes over 1990–2000 in the developing countries in Table 1 is greater than three per cent of GDP; in the USA it is less than 0.2 per cent. Values in excess of ten per cent have been recorded for some countries in some years. These figures are large relative to the magnitude of supply shocks estimated for most OECD countries. So there is a strong case for advocating an active short-run monetary policy in those developing countries with a stable underlying monetary and fiscal regime.

The current norm in OECD countries is an institutionally independent central bank which regularly adjusts a monetary policy instrument – the quantity of short-term lending to commercial banks, or more frequently the corresponding interest rate – in order to meet an implicit or explicit medium-term inflation target. This target is usually accompanied by an injunction to avoid ‘unnecessary’ volatility in real macroeconomic indicators such as GDP growth or the unemployment rate. The relative weight to be given to the two goals is seldom explicit, but the academic literature – including the research divisions of many central banks, though never their policy statements – interprets the trade-off between output and price stability in terms of the framework introduced by Taylor (1993). That is, the optimal value of the instrument in any one period is derived from the maximization of an objective function including the deviations of inflation and output (or unemployment) from their target values, subject to a constraint embodied in a short-run supply curve (or Phillips curve). Shocks to the supply curve shift the constraint and so change the optimal value of the instrument; the magnitude of the change depends on the weights on the different targets in the objective function. Past central bank behaviour is often interpreted as such a Taylor rule plus inertia reflecting model uncertainty and a random component reflecting unquantifiable information about the economy.

In recent years, some non-OECD countries have introduced explicit inflation targeting with a degree of central bank independence and accountability. These are not countries typical of those in the study of Cuckierman, Webb and Neyapti (1992): they have relatively stable political institutions and relatively democratic governments. High-profile examples are Brazil, Chile, the Czech Republic, Poland and South Africa. There are also central banks that lack an explicit inflation target but nevertheless publish policy reports that motivate the adjustment of monetary instruments by reference to short-term movements in inflation and output, accompanied by research division working papers on the application of Taylor rules to their economy. The central banks of the two African monetary unions discussed above, the BCEAO and the BEAC, are examples of this phenomenon. Econometric studies of the evolution of monetary instruments and macroeconomic variables in these counties suggest that in most cases their monetary institutions function in a broadly similar way to those of OECD countries, although the shocks to which they are responding (often normalized in econometric analysis!) are greater in magnitude.

The creation of such institutions is surely endogenous to a country's level of political and economic development. They are the consequence rather than the cause of a stable policy environment and relatively developed financial markets. The former ensures the credibility of monetary institutions; the latter ensures an identifiable monetary transmission mechanism in which interest rate changes can be expected to impact on the economy in a consistent way. Such examples represent one tail of the distribution of institutional quality, in which institutions reduce macroeconomic instability. There are still many countries in which institutions increase macroeconomic instability, as witnessed by the hyperinflation still endemic in many parts of the world. Nevertheless, they indicate that a high level of per capita GDP is not a necessary condition for monetary institutions equally as effective as those in the

OECD.

See Also

- development economics
- exchange rate volatility
- hyperinflation
- inflation targeting
- international financial institutions (IFIs)
- Taylor rules

Bibliography

Agénor, P.-R. and Montiel, P. 1999. *Development Macroeconomics*, 2nd edn. Princeton, NJ: Princeton University Press.

Cardoso, E. 1981. Food supply and inflation. *Journal of Development Economics* 8, 269–84.

Cukierman, A., Webb, S. and Neyapti, B. 1992. Measuring the independence of central banks and its effect on policy outcomes. *World Bank Economic Review* 6, 353–98.

Flood, R. and Garber, P. 1984. Collapsing exchange rate regimes: some linear examples. *Journal of International Economics* 17, 1–13.

IMF (International Monetary Fund). 1983. *International Financial Statistics Yearbook*. Washington, DC: IMF.

IMF (International Monetary Fund). 1999. *International Financial Statistics Yearbook*. Washington, DC: IMF.

Kydland, F. and Prescott, E. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–90.

Montiel, P. 1989. Empirical analysis of high inflation episodes in Argentina, Brazil and Israel. *IMF Staff Papers* 36, 527–49.

Obstfeld, M. 1996. Models of currency crises with self-fulfilling features. *European Economic Review* 40, 1037–47.

Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on*

Public Policy 39, 195–214.

World Bank. 2003. *World Bank Indicators 2003*. Washington, DC: World Bank.

How to cite this article

Fielding, David. "fiscal and monetary policies in developing countries." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000116> doi:10.1057/9780230226203.0578

The New Palgrave Dictionary of Economics Online

fiscal federalism

David E. Wildasin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Fiscal federalism is concerned with the division of policy responsibilities among different levels of government and with the fiscal interactions among these governments. Public service provision by lower-level governments can be efficiency-enhancing, although competition for mobile resources can also interfere with efficient resource allocation both in the public and private sectors. Intergovernmental transfers affect the overall equity and efficiency properties of public policies. Global economic integration and political and economic reforms in developing and transition economies – which have institutional contexts very different from those of the mature federations – present important challenges for a ‘second generation’ of federalism research.

Keywords

decentralization; factor mobility; fiscal competition; fiscal federalism; horizontal equity; intergovernmental grants; local public goods; optimal currency areas; planning; policy coordination; stabilization policy; tax competition; tax distortions; Tiebout hypothesis

Article

Fiscal federalism is concerned with the division of policy responsibilities among different levels of government and with the fiscal interactions among these governments.

The institutional context of fiscal federalism

Fiscal federalism has long been a topic of keen interest in the United States and Canada. In both nations, subnational governments have traditionally played major roles in the provision of important public services, notably in the areas of education, health, social services, transportation, public safety, and economic development. In addition to non-tax revenues, subnational governments in both countries have

had significant sources of tax revenues, with state/provincial governments relying heavily on retail sales taxes and taxes on personal and business income and with local governments depending on property taxes. Higher-level governments (national in relation to subnational, and state/provincial in relation to local) have supported the finances of lower-level governments with extensive programmes of intergovernmental fiscal transfers in order to promote the provision of particular public goods and services, to supplement (or perhaps displace) lower-level government taxes, and to advance broad social welfare objectives. Although they are subject to constitutional, statutory, and regulatory constraints, state/provincial and local governments exercise substantial fiscal autonomy with respect to expenditures, taxation and borrowing. National and subnational fiscal policies have been developed and implemented within the context of continuously evolving but fundamentally durable market, political, and legal institutions, underpinned by stable democratic constitutional structures.

There are long-established federations (and long traditions of scholarly research on federalism) in other parts of the world as well, but interest in fiscal federalism has become particularly intense in developing and transition economies since the early 1990s, no doubt in part because of broad reform initiatives that have reduced the role of the state in economic planning and control (Wildasin, 1997a, ch. 2). In many of these countries, constitutional, economic, and political reforms have led to significant decentralization of tax, expenditure, and borrowing responsibilities, often accompanied by the development of new systems of intergovernmental fiscal transfers. In contrast to the mature North American federations, the newly (or increasingly) decentralized and liberalized economic and fiscal systems of many developing and transition economies are being implemented in the absence of the background political, legal, and market institutions found in more developed nations. The development and restructuring of federations around the world has presented many practical challenges and, for scholars, important questions regarding the design of federal systems, the implementation of fiscal reforms in such systems, and the interactions between basic social institutions and the public sector in federations.

Fiscal federalism is also a subject of increased interest and concern in the European Union. Fiscal decentralization has accompanied economic and political reforms in several European nations. In addition, the interactions of tax, expenditure, debt, and monetary policies among EU member states continuously raise questions concerning international policy coordination and the development of EU-wide supranational institutions. Controversy surrounds the issues of national sovereignty and the upward transfer of powers from national governments to EU executive, legislative, and judicial bodies. In important respects, however, the EU can be viewed as an emerging federation in which EU-level political and fiscal institutions are gradually developing within the context of an increasingly integrated and expanding system of developed and transition economies. From this perspective, the EU itself is a (so far relatively limited) higher-level government in relation to the national governments of its member states.

Fiscal federalism is thus a subject of great interest throughout the world. Wide international variation in the institutional context of federalism has stimulated what Oates (2005) calls a ‘second generation’ of fiscal federalism research, differentiated from ‘first-generation’ research by its heightened attention to political, constitutional, financial and macroeconomic institutions. For example, issues of fiscal discipline, soft budget constraints, and subnational government borrowing, little discussed within the context of traditional federalism research, have received considerable attention in recent years (Inman, 2003; Wildasin, 1997b; 2004), especially with reference to newly decentralizing fiscal systems. Because the policy issues and institutional context of federalism varies widely throughout the world, a rapidly

growing literature deals with fiscal federalism in an international context, often focusing on unique policy issues facing individual countries (see, for example, Bird and Vaillancourt, 1998; Martinez-Vasquez and Alm, 2003; and Rodden, Eskeland and Litvack, 2003, which contain many studies of federalism problems in developing and transition economies).

As the foregoing remarks suggest, problems of fiscal federalism touch upon almost all aspects of fiscal policy, in almost all nations (especially the large nations and economic regions) of the world. The subject is correspondingly very broad. The following paragraphs highlight recurring themes that have occupied researchers for many years as well as selected issues that are likely to be the subject of active enquiry in coming years. The discussion begins with fundamental issues regarding the economic functions of different levels of government, noting their implications for the organization of the public sector. The potential efficiency gains from decentralized policymaking as well as the limitations of decentralization are discussed next, emphasizing the importance of resource mobility and fiscal competition as a crucial feature of the decision-making environment facing lower-level governments. Finally, directions for new research are briefly discussed.

The organization of the public sector

What economic functions can, do, or should be performed by different levels of government? This fundamental question has been a focus of the federalism literature from its inception. There has been a broad normative consensus (Oates, 1972) that, of Musgrave's (1959) ‘three branches of the public household’, the highest-level government (normally a national government, but possibly a supranational entity like the EU) should take responsibility for stabilization functions (that is, macroeconomic and monetary policies), that allocative functions (the provision of public goods and services and correction of market failures) should be undertaken by governments whose jurisdictional boundaries are co-terminous with the geographical scope of the regions affected by these policies, and that higher-level governments should be responsible for policies that target the distribution of income. Subnational economies are comparatively more open than national economies, which means that the impacts of stabilization policies are diluted through capital, labour, and financial flows when undertaken by lower-level governments; see, for example, Mundell's (1961) classic work on optimal currency areas. Similarly, the mobility of labour and capital constrains the ability of (small, open) subnational governments to alter the net distribution of income. For example, high taxes on the rich in one jurisdiction create incentives for the rich to locate elsewhere, while the provision of generous cash or in-kind benefits for the poor attracts beneficiaries (Stigler, 1957). In addition to distorting the efficiency of resource allocation, the spatial reallocation of resources in response to local redistributive policies limits the set of feasible policies as well as their impact on net incomes. Lower-level governments may, however, serve effectively to provide public goods and services in the amounts that are most efficiently adapted to local benefits and costs, which normally vary among locations in accordance with differences in demographic composition, incomes, and technologies (Oates's ‘decentralization theorem’).

Allocative efficiency at the local level

The decentralization theorem shows that non-uniform provision of public goods, varying in accordance

with local benefits and costs, may be more efficient than uniform provision. In principle, however, an omniscient and omnipotent central planner could implement optimal non-uniform policies, obviating the need for distinct administrative units of lower-level government. Such a planner could manage all public sector functions (in fact, all economic decisions) for the entire world. A key idea in the literature of fiscal federalism, however, is that lower-level units of government may be better informed about and more responsive to local demands. The *information* needed for efficient decision-making, and the *incentives* to use this information, may differ by level of government, just as markets provide incentives guiding decentralized market decisions for households and firms in ways not achievable, in practice, by central planning mechanisms.

This idea is developed explicitly, if informally, in Tiebout (1956). Tiebout draws the analogy between consumers shopping for commodities in the marketplace and households choosing residences from among a collection of localities. Writing soon after and in response to Samuelson's classic contributions to public goods theory, Tiebout asserts that households reveal their preferences for local public goods when they choose where to reside. Different localities provide different levels of public services, as illustrated by local school districts in the United States that offer different qualities of elementary and secondary education. Households with high valuations for education can outbid others for residences in localities with good schools, thus leading to a sorting of households by demand for public services. According to Tiebout, this matching of demand and supply leads to efficient provision of local public goods.

Tiebout's paper identifies local governments as distinct economic units that can perform important allocative functions in ways that central governments cannot. Tiebout is not specific, however, about exactly how local decision-makers determine public goods levels – whether by voting or through some other mechanism. Many subsequent contributions (see, for example, Wildasin, 1986, for a survey and references), including both theoretical and empirical analyses, explore in detail the phenomenon of ‘Tiebout sorting’ and the implications of community stratification, by income, race, religion, age and other household attributes, for variation in local public expenditures. Median voter models (and variants thereof) commonly provide a theoretical starting point for empirical analyses of the demand for local public goods. Linkages between housing markets and local fiscal policies, as revealed by hedonic price relationships, suggest that local voters have incentives to support policies that preserve property values. In the extreme, these linkages may obviate altogether the need for households to participate in the collective decision-making process, by providing profit-maximizing property developers and other market participants with the information and incentives to make efficient policy choices, resulting in completely market-driven provision of public goods (Fischel, 2001, discusses land use regulation, property development and their interactions with community formation and local policymaking).

In addition to the information and incentives that may result from the mobility of households and firms, emphasized by Tiebout, decentralized policymaking may also provide a framework for experimentation and learning about policy alternatives and their consequences as well as for learning about the performance of policymakers themselves (Besley and Case, 1995).

Limits to decentralization: efficiency and distributional considerations

Tiebout's analysis and much subsequent research highlights the potential benefits, especially with

respect to the efficiency of public good provision, from competition among lower-level governments for mobile households and firms. The potential disadvantages of fiscal decentralization have long been recognized, however. For instance, the economic service areas for local public goods may not closely match jurisdictional boundaries. Local health, educational, or transportation policies may benefit residents of neighbouring localities or society at large, spillover benefits that local decision-makers may ignore. These externalities can potentially be internalized through voluntary policy coordination among neighbouring governments. Such coordination can be costly, however, resulting in inefficient decentralized public good provision. Within a federation, a higher-level government can use intergovernmental grants (generally conditional grants, especially matching grants that reduce the marginal cost of public good provision for recipient governments) in order to induce more efficient provision of externality-generating local public goods and services (Breton, 1965). If the spillover benefits from a public good are sufficiently widespread, a higher-level government may assume complete responsibility for its provision. Such centralization of a governmental function involves a trade-off between the potential benefits from internalization of externalities and the potential informational disadvantages of centralized collective decision-making for a larger and more heterogeneous population (Alesina and Spolaore, 2003).

A second possible drawback of decentralized policymaking arises if there are significant limitations on the fiscal instruments available to lower-level governments. In the competition among lower-level governments for households and businesses, taxes (or non-tax revenue instruments such as user fees or licenses) perform a ‘price like’ function by rationing access to public services. Taxes may also introduce inefficiencies of their own, however, not only through ‘classical’ tax distortions (distortion of *in situ* labour/leisure, consumption, savings, and investment decisions) but more especially through their effects on the locational choices of households and businesses. For example, subnational government income taxes may inefficiently drive profitable businesses and high-income households into low-tax jurisdictions, and retail sales taxes may encourage inefficient cross-boundary shopping. Fiscal competition for mobile factors of production or consumers may discourage taxation of these resources, changing the composition of the subnational revenue structures toward less-mobile tax bases if these are available and potentially constraining the overall level of government revenues. Underprovision of public goods may result, which, as in the case of spillover benefits, may potentially be remediated with well-designed fiscal transfers from higher-level governments (Wildasin, 2006a; Wilson and Wildasin, 2004; Wilson, 1999). On the other hand, if Leviathan governments are likely to engage in excessive spending, fiscal competition may impose useful constraints on their revenue-raising powers (Brennan and Buchanan, 1980).

A further difficulty for federalized systems arises from the fact that many public policies, by their nature, intermingle allocative and distributional impacts, so that a clean separation of allocative and redistributive functions between higher- and lower-level governments may be unattainable. Health, education, transport, economic development, and many social services involve allocative functions (service delivery for geographically limited areas) but also promote distributional goals. Particularly when competition among lower-level governments results in the formation of communities that are relatively homogeneous (with respect to income, race, age or other socioeconomic characteristics), the efficiency gains from decentralization may be realized in part precisely through increased disparities in public service provision. The demand for education, for example, is a normal good, so that stratification

of localities by income produces disparities in educational quality between rich and poor localities, as efficiency requires. In the United States, concern about the fairness of inequality in education, partly as expressed in state government constitutions, has resulted in extensive litigation leading to judicial mandates for policy reforms, notably including extensive programmes of equalizing fiscal transfers from state to local governments (Inman and Rubinfeld, 1979). More generally, the equalization of fiscal transfers from higher- to lower-level governments provides a mechanism through which to limit horizontal inequities in the fiscal treatment of households in rich and poor jurisdictions and the locational incentives to which they give rise (Boadway and Flatters, 1983).

As noted earlier, factor mobility imposes constraints on the ability of governments to redistribute incomes. The integration of capital and labour markets can improve the efficiency of factor allocations and thus raise output and welfare, an important potential benefit that underpins policy initiatives, such as economic integration within the EU, that seek to remove barriers to factor mobility. Factor mobility also affects factor prices, giving rise to potentially important first-order distributional impacts. Thus, economic integration affects not only the cost of ‘decentralized’ redistribution – which, in a global context with international factor mobility, includes redistribution by national as well as subnational governments. By affecting factor prices and the underlying distribution of income, it also may increase or decrease the benefits of redistributive policies. International capital mobility and the migration of younger workers (both skilled and unskilled) from developing and transition economies to aging developed nations thus create new policy trade-offs, particularly for the extensive redistributive systems of North America and Western Europe (Wildasin, 2006b), the consequences of which will unfold in coming decades.

Directions for future research

As noted at the outset, the challenges of policy and institutional reform throughout the world have stimulated new interest in fiscal federalism. The incentives embedded in the institutional structures of the mature federations seem to have ensured that subnational governments maintain sufficient fiscal discipline to avoid major widespread or recurring fiscal crises, while preserving their ability to exercise significant policy autonomy with respect to the level and composition of their taxes, expenditures and debts (Buettner and Wildasin, 2006; Inman, 2003; Wildasin, 2004). Such institutions cannot be taken for granted, however, and many informed observers see potential risks from fiscal decentralization in the evolving federations of the developing and transition economies, including risks from excessive (that is, inefficiently high) spending or borrowing by subnational governments. An appropriate mix of revenue and expenditure assignments, intergovernmental fiscal transfers, borrowing flexibility, and policy autonomy is needed in order to realize the potential efficiency gains from fiscal decentralization (McLure and Martinez-Vasquez, n.d.; Weingast, 2006). The interplay between the market environment (especially financial markets and institutions and capital and labour mobility), the assignment of fiscal and regulatory authorities by level of government, and the constraints that influence political decision-making is not well understood and promises to be the subject of extensive study in coming years.

The integration of national and international markets for labour and capital, of crucial importance for federalism, appears to be increasing over time, and affects the competitive pressures facing governments at all levels. The global configuration of age-imbalanced demographic structures (young poor

populations in developing countries and old rich populations in developed countries) implies that international migration incentives are unlikely to diminish in the foreseeable future. The fiscal systems of developed nations, with their extensive systems of intra- and intergenerational transfers, will face growing challenges in coming decades as a result of population aging, even as competition for capital investment and mobile high-income households may increasing constrain their capacity to finance redistribution (Wildasin, 2006c). Policy coordination, perhaps through newly developed governmental structures (for example, at the EU level), may provide opportunities for national governments to limit the degree of fiscal competition, helping them to finance the liabilities arising under existing redistributive systems. Alternatively, or in addition, national governments may explicitly or implicitly shift some expenditure responsibilities to lower-level governments as they manage growing fiscal imbalances arising from demographic change. In any case, growing fiscal imbalances are likely to form the backdrop for public finance in developed countries in coming decades, offering opportunities for fruitful analysis of the dynamics of factor mobility, factor market integration, dynamic fiscal adjustment, and institutional change within and among nations.

See Also

- intergovernmental grants
- local public finance
- public finance
- tax competition
- Tiebout hypothesis

Bibliography

Alesina, A. and Spolaore, E. 2003. *The Size of Nations*. Cambridge, MA: MIT Press.

Besley, T. and Case, A. 1995. Incumbent behavior: vote seeking, tax setting and yardstick competition. *American Economic Review* 85, 25–45.

Bird, R.M. and Vaillancourt, F. 1998. *Fiscal Decentralization in Developing Countries*. Cambridge: Cambridge University Press.

Boadway, R.W. and Flatters, F. 1983. Efficiency and equalization payments in a federal system of government: a synthesis and extension of recent results. *Canadian Journal of Economics* 15, 613–33.

Brennan, G. and Buchanan, J.M. 1980. *The Power to Tax*. Cambridge: Cambridge University Press.

Breton, A. 1965. A theory of government grants. *Canadian Journal of Economics and Political Science* 31, 175–87.

Buettner, T. and Wildasin, D.E. 2006. The dynamics of municipal fiscal adjustment. *Journal of Public Economics* 90, 1115–32.

Fischel, W.A. 2001. *The Homevoter Hypothesis: How Home Values Influence Local Government Taxation, School Finance, and Land-Use Policies*. Cambridge, MA: Harvard University Press.

Inman, R.P. 2003. Transfers and bailouts: enforcing local fiscal discipline with lessons from US federalism. In Rodden, Eskeland and Litvack (2003).

Inman, R.P. and Rubinfeld, D.L. 1979. The judicial pursuit of local fiscal equity. *Harvard Law Review* 92, 1662–750.

Martinez-Vasquez, J. and Alm, J. 2003. *Public Finance in Developing and Transition Countries: Essays in Honor of Richard Bird*. Cheltenham: Edward Elgar.

McLure, C.E. and Martinez-Vasquez, J. n.d. *The Assignment of Revenues and Expenditures in Intergovernmental Fiscal Relations*. Washington, DC: World Bank.

Mundell, R.A. 1961. A theory of optimum currency areas. *American Economic Review* 51, 509–17.

Musgrave, R.A. 1959. *Theory of Public Finance*. New York: Macmillan.

Oates, W.E. 1972. *Fiscal Federalism*. New York: Harcourt Brace Jovanovich.

Oates, W.E. 2005. Toward a second-generation theory of fiscal federalism. *International Tax and Public Finance* 12, 349–73.

Rodden, J., Eskeland, G.S. and Litvack, J. 2003. *Fiscal Decentralization and the Challenge of Hard Budget Constraints*. Cambridge, MA: MIT Press.

Stigler, G.J. 1957. The tenable range of functions of local government. Joint Economic Committee, *Federal Expenditure Policy for Economic Growth and Stability*. Washington, DC: US Government Printing Office.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Wallack, J. and Srinivasan, T.N. 2006. *Federalism and Economic Reform: International Perspectives*. Cambridge: Cambridge University Press.

Weingast, B. 2006. Second generation fiscal federalism: implications for decentralized democratic governance and economic development. Unpublished paper presented at IFIR-CESifo Conference, New

Directions in Fiscal Federalism. Lexington, KY.

Wildasin, D.E. 1986. *Urban Public Finance*. New York: Harwood Academic.

Wildasin, D.E. 1997a. *Fiscal Aspects of Evolving Federations*. Cambridge: Cambridge University Press.

Wildasin, D.E. 1997b. Externalities and bailouts: hard and soft budget constraints in intergovernmental fiscal relations. Policy Research Working Paper No. 1843. Washington, DC: World Bank.

Wildasin, D.E. 2004. The institutions of federalism: toward an analytical framework. *National Tax Journal* 62, 247–72.

Wildasin, D.E. 2006a. Fiscal competition. In *Oxford Handbook of Political Economy*, ed. B. Weingast and D. Wittman. Oxford: Oxford University Press.

Wildasin, D.E. 2006b. Global competition for mobile resources: implications for equity, efficiency, and political economy. *CESifo Economic Studies* 52, 61–111.

Wildasin, D.E. 2006c. Public finance in an era of global demographic change: fertility busts, migration booms, and public policy. Unpublished paper presented at Council on Foreign Relations conference, Skilled Migration Today: Prospects, Problems, and Policies, New York.

Wilson, J.D. 1999. Theories of tax competition. *National Tax Journal* 52, 296–315.

Wilson, J.D. and Wildasin, D.E. 2004. Capital tax competition: bane or boon? *Journal of Public Economics* 88, 1065–91.

How to cite this article

Wildasin, David E. "fiscal federalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000118> doi:10.1057/9780230226203.0579

The New Palgrave Dictionary of Economics Online

fiscal theory of the price level

Marco Bassetto

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The fiscal theory of the price level (FTPL) describes fiscal and monetary policy rules such that the price level is determined by government debt and fiscal policy alone, with monetary policy playing at best an indirect role. This theory clashes with the monetarist view that states that money supply is the primary determinant of the price level and inflation. Furthermore, many authors have argued that the fiscal rules upon which the FTPL relies are misspecified. We review the sources of disagreement, and highlight aspects upon which some consensus has emerged.

Keywords

budget deficits; commodity money; debt crises; dynamic competitive equilibrium; exogenous interest rates; fiscal theory of the price level; government budget constraint; inflation; interest rate peg; interest rate rules; intertemporal budget constraint; monetarism; monetary policy; monetization of government debt; money supply rule; multiple equilibria; nominal interest rate; price level; quantity theory of money; real money balances; seigniorage; sunspots; uniqueness of equilibrium; velocity of circulation

Article

The fiscal theory of the price level (FTPL) describes policy rules such that the price level is determined by government debt and the present and future tax and spending plans, with no direct reference to monetary policy.

In understanding the FTPL and tracing its roots, we start from two simple relations: the velocity equation and the government budget constraint.

The velocity equation defines the velocity of money in period t (V_t) as the ratio of nominal output (the price level P_t times real output Y_t) to nominal money balances (M_t):

$$V_t = \frac{P_t Y_t}{M_t}, \; t = 0, 1, \ldots \tag{1}$$

Differences across monetary models arise in the way these four economic variables are determined, and in the specification of which (if any) of these variables is to be treated as exogenous as opposed to endogenous. Prior to the introduction of the FTPL, eq. (1) was viewed as the primary determinant of the price level. As an example, the quantity theory of money states that V_t is fixed and exogenous. In this case, the price level is proportional to the money supply. High prices arise because too much money is chasing too few goods, which is the heart of the monetarist doctrine. In a more sophisticated theory, velocity is itself affected by other macroeconomic variables, chief among them the nominal interest rate. Furthermore, in general, the price level needs to be determined jointly with M_t, Y_t, and V_t by computing the entire equilibrium path of the economy. The FTPL traces its roots to an incompleteness in the monetarist view of the price level: often, the equilibrium price level fails to be uniquely determined, that is, there are many paths of P_t that satisfy (1) as well as all the other equilibrium requirements (see discussion in Kocherlakota and Phelan, 1999). This is especially true when monetary policy prescribes an exogenous interest rate; Sargent and Wallace (1975) show that the initial price level is then indeterminate, and subsequent inflation is subject to ‘sunspots’, uncertainty driven by self-fulfilling expectations. In the simplest case, an interest-rate peg determines the level of velocity (V_t), and real output and interest rates are independent of money and prices; eq. (1) then pins down *real* money balances (M_t/P_t), but it does not specify whether those balances will be attained by high or low nominal money supply and prices.

The FTPL (Woodford, 1994; Sims, 1994) determines prices from a different equation:

$$\frac{B_t}{P_t} = \text{Present value of primary fiscal surpluses as of time } t, \; t = 0, 1, \ldots, \tag{2}$$

where B_t is the nominal value of government liabilities (debt and money) at the beginning of period t. Equation (2) is the government budget constraint, in its present value form: the left-hand side represents real government liabilities, matched by assets on the right-hand side. In its simplest form, the FTPL assumes that the government commits to a fixed and exogenous present value of primary fiscal surpluses; this is a special case of what Leeper (1991) defines as an ‘active’ fiscal policy and Woodford (1995) a ‘non-Ricardian’ fiscal regime. Given an initial condition for debt, B_0, a unique price level is consistent with (2): the FTPL successfully selects a unique price level at time 0, even in the case of an interest rate peg, for which the monetarist view offered no prediction. The power of the FTPL is not limited to period 0; the possibility of sunspot equilibria is ruled out in all subsequent periods, since again

a unique level of prices is consistent with a given present value of surpluses and the nominal debt inherited from the past. Nonetheless, monetary policy does have an effect on inflation after period 0: the evolution of nominal liabilities B_t depends on the nominal interest rate, which is affected by monetary policy.

Since its inception, the FTPL has been extremely controversial. I focus here on two main areas of concern.

The value of money or the value of debt?

The price level is defined as the inverse of the value of money: how much money it takes to buy a given basket of goods. By contrast, the FTPL is about the inverse of the value of *government debt*. This is explained particularly clearly in Cochrane (2005). As Buiter (2002) points out, there is no reason in general for the value of debt and the value of money to coincide. To the extent that households anticipate a government default, they may trade government debt at a discount, without necessarily affecting the value of money. This criticism is particularly serious when the central bank adopts a monetary policy that rules out monetization of government debt. As an example, consider the case in which the central bank adopts a constant money supply rule and does not engage in open market operations. In this case, there is no link between government debt and money, and no reason why a maturing T-Bill with a face value of $1,000 should trade at par with ten $100 notes issued by the central bank. Maturing debt and money will trade at par only if *fiscal policy* is run in such a way that the government will have the appropriate amounts of money to repay its debt, independently of the price level: this requires real tax revenues to adjust to prices, violating the central assumption of the FTPL.

The same criticism does not apply when the monetary policy of the central bank allows unlimited monetization of debt, as in the case of an interest rate peg. In this case, the central bank commits to exchange arbitrary amounts of money and one-period government debt at a fixed price. This commitment is not inconsistent with a second commitment, to redeem all maturing government debt at par, offering money in exchange. Since the central bank has unlimited ability to produce money, a government default on nominal debt is now ruled out. In this case, the FTPL is simply a version of a commodity money standard; money, as well as other government liabilities, is backed by the present value of future government surpluses, just as the value of Microsoft shares is backed by the present value of Microsoft profits (this is the main example in Cochrane, 2005).

While the original treatment of the FTPL was ambiguous (in particular, Woodford, 1995, considers the case of the FTPL under a money supply rule), it is now widely agreed that the FTPL requires an implicit or explicit institutional commitment to prevent a government default (or excess repayments by the government) through an appropriate (de)monetization of debt. In this form, the FTPL bears some similarities with the ‘unpleasant monetarist arithmetic’ of Sargent and Wallace (1981). Under the monetarist arithmetic, a fiscal deficit imbalance will trigger inflation, because seigniorage revenues are necessary to prevent the government from defaulting. Even though monetization of government debt plays a central role in both theories, there are important differences. According to the unpleasant monetarist arithmetic, seigniorage revenues (which are part of the present value of surpluses in eq. (2)) will have to respond to changes in P_t to ensure that the government budget constraint holds; hence, equilibrium occurs through adjustments in the right-hand side of (2). In the FTPL, seigniorage revenues

on the monetary base play at best a minor role. Under the FTPL, it is the price level that responds to shocks to spending and taxes; its fluctuations cause the real value of debt (the left-hand side of (2)) to appreciate or depreciate to reach an equilibrium.

Government constraints and equilibrium conditions

The FTPL is based on the assumption that eq. (2) holds only at an equilibrium. The critics of the FTPL view instead (2) as a constraint that forces the government to match the real value of debt with an appropriate present value of primary surpluses, for all conceivable levels of prices. To better understand the issue, it is useful to note that (2) looks identical to the intertemporal budget constraint of any household in the economy: it is sufficient to relabel B_t as the nominal liabilities of the household, and the right-hand side as the present value of its non-asset income, net of consumption. In the case of a household, there is universal agreement that (2) should be viewed as a constraint: given any value of P_t, the household must choose a consumption/income plan that satisfies (2). The critics of the FTPL argue that the government should be no different from any other agent. Unlike the previous criticism, the heated debate that has emerged on this point has not resulted in widespread agreement. As Bassetto (2005) points out, the disagreement stems from a fundamental weakness in the tools that have been used to study this problem. Both critics and supporters of the FTPL adopt the dynamic competitive equilibrium framework. This framework is designed for environments populated by many small players; in the presence of a large and potentially strategic player, such as the government, it offers little guidance in distinguishing between equilibrium conditions and constraints that the large player(s) faces under any circumstances, even away from an equilibrium. While there are many applications for which this ambiguity is not important, a proper account of the distinction is essential to study the uniqueness or multiplicity of equilibria, which is the object of interest in the case of the FTPL.
To overcome this difficulty, Bassetto (2002) explicitly describes the economy as a game, where the actions available to all households and the government at any point in time are clearly spelt out. Bassetto shows that the basic version of the FTPL, with an unconditional commitment to a sequence of primary surpluses, is not a valid government strategy in a well-specified game, at least if the sequence includes a primary deficit at any point in time. Intuitively, a primary deficit is possible only if the government is able to raise revenues through borrowing. Since lending is voluntary (unlike payment of taxes), any plausible game includes the possibility that private agents will not lend; if this circumstance arises, the government is forced to a fiscal adjustment. Bassetto then proves that there exist other government strategies that lead to a unique equilibrium price level that is determined from taxes and spending alone. These strategies paint a very different picture of the conditions under which a FTPL arises: whereas the traditional view relies on the government setting taxes and spending exogenously, with no regard for the evolution of debt, the strategies described by Bassetto require the government to strongly react to incipient ‘debt crises’ by accumulating larger surpluses in present value.

Empirical studies

A small empirical literature (for example, Canzoneri, Cumby and Diba, 2001; Cochrane, 2001) has looked into the usefulness of (2) in accounting for the evolution of prices. The results are not very

favourable; in particular, when a government runs an unexpected deficit, the real market price of its debt increases, suggesting that households expect that the government will make up for the shortfall through increased surpluses in the future. If future surpluses were exogenous and fixed, (2) would suggest that an unexpected deficit should have its primary effect through inflation, by depressing the real market value of debt. While these observations cannot refute the central claim of the FTPL, that (2) is only an equilibrium condition, they call into question the usefulness of the FTPL in explaining actual inflationary episodes.

Conclusion

Recent research into monetary policy has looked for interest rate rules that ensure price level determinacy independently of the fiscal policy of the government; this has weakened interest in the FTPL. Though no issue as controversial as the FTPL has emerged since, this recent analysis is still open to ambiguous distinctions between policy rules, which should capture government behaviour in all possible scenarios, and equilibrium relations across the endogenous variables of an economic system. A more complete analysis awaits the development of new tools that are as simple and powerful as dynamic competitive equilibrium, and yet able to appropriately capture the special role of the government.

See Also

- commodity money
- determinacy and indeterminacy of equilibria
- government budget constraint
- monetarism
- monetary and fiscal policy overview
- multiple equilibria in macroeconomics
- quantity theory of money
- sunspot equilibrium

Bibliography

Bassetto, M. 2002. A game-theoretic view of the fiscal theory of the price level. *Econometrica* 70, 2167–95.

Bassetto, M. 2005. Equilibrium and government commitment. *Journal of Economic Theory* 124, 79–105.

Buiter, W. 2002. The fiscal theory of the price level: a critique. *Economic Journal* 112, 459–80.

Canzoneri, M., Cumby, R. and Diba, B. 2001. Is the price level determined by the needs of fiscal solvency? *American Economic Review* 91, 1221–38.

Cochrane, J. 2001. Long term debt and optimal policy in the fiscal theory of the price level. *Econometrica* 69, 69–116.

Cochrane, J. 2005. Money as stock. *Journal of Monetary Economics* 52, 501–28.

Kocherlakota, N. and Phelan, C. 1999. Explaining the fiscal theory of the price level. *Federal Reserve Bank of Minneapolis Quarterly Review* 23(4), 14–23.

Leeper, E. 1991. Equilibria under ‘active’ and ‘passive’ monetary policies. *Journal of Monetary Economics* 27, 129–47.

Sargent, T. and Wallace, N. 1975. ‘Rational’ expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83, 241–54.

Sargent, T. and Wallace, N. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5(3), 1–17.

Sims, C. 1994. A simple model for study of the determination of the price level and the interaction of monetary and fiscal policy. *Economic Theory* 4, 381–99.

Woodford, M. 1994. Monetary policy and price level determinacy in a cash-in-advance economy. *Economic Theory* 4, 345–80.

Woodford, M. 1995. Price level determinacy without control of a monetary aggregate. *Carnegie-Rochester Conference Series on Public Policy* 43, 1–46.

How to cite this article

Bassetto, Marco. "fiscal theory of the price level." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000301> doi:10.1057/9780230226203.0580

The New Palgrave Dictionary of Economics Online

Fisher, Irving (1867 – 1947)

James Tobin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Irving Fisher was born in Saugerties, New York, on 27 February 1867; he was residing in New Haven, Connecticut at the time of his death in a New York City hospital on 29 April 1947.

Fisher is widely regarded as the greatest economist America has produced. A prolific, versatile and creative scholar, he made seminal and durable contributions across a broad spectrum of economic science. Although several earlier Americans, notably Simon Newcomb, had used some mathematics in their writings. Fisher's dedication to the method and his skill in using it justify calling him America's first mathematical economist. He put his early training in mathematics and physics to work in his doctoral dissertation on the theory of general equilibrium. Throughout his career his example and his teachings advanced the application of quantitative method not only in economic theory but also in statistical inquiry. He, together with Ragnar Frisch and Charles F. Roos, founded the Econometric Society in 1930; and Fisher was its first President. He had been President of the American Economic Association in 1918.

Much of standard neoclassical theory today is Fisherian in origin, style, spirit and substance. In particular, most modern models of capital and interest are essentially variations on Fisher's theme, the conjunction of intertemporal choices and opportunities. Likewise, his theory of money and prices is the foundation for much of contemporary monetary economics.

Fisher also developed methodologies of quantitative empirical research. He was the greatest expert of all time on index numbers, on their theoretical and statistical properties and on their use in many countries throughout history. From 1923 to 1936, his own Index Number Institute manufactured and published price indexes of many kinds from data painstakingly collected from all over the world. Indefatigable and innovative in empirical research, Fisher was an early and regular user of correlations, regressions and other statistical and econometric tools that later became routine.

To this day Fisher's successors are often rediscovering, consciously or unconsciously, Fisher's ideas and building upon them. He can be credited with distributed lag regression, life cycle saving theory, the ‘Phillips curve’, the case for taxing consumption rather than ‘income’, the modern quantity theory of money, the distinction between real and nominal interest rates, and many more standard tools in economists’ kits. Although Fisher was not fully appreciated by his contemporaries, today he leads other

old-timers by wide and increasing margins in journal citations. In column inches in the *Social Sciences Citation Index* (1979, 1983), Fisher led his most famous contemporaries, Wesley Mitchell, J.B. Clark, and F.W. Taussig in that order, by rough ratios 5:3:1:1 in 1971–5 and 9:3:1:1 in 1976–80. Much more than the others, moreover, Fisher is cited for substance rather than for history of thought.

For all his scientific prowess and achievement, Fisher was by no means an ‘ivory tower’ scholar detached from the problems and policy issues of his times. He was a congenital reformer, an inveterate crusader. He was so aggressive and persistent, and so sure he was right, that many of his contemporaries regarded him as a ‘crank’ and discounted his scientific work accordingly. Science and reform were indeed often combined in Fisher's work. His economic findings, theoretical and empirical, would suggest to him how to better the world; or dissatisfaction with the state of the world would lead him into scientifically fruitful analysis and research. Fisher's search for conceptual clarity about ‘the nature of capital and income’ led him not only to lay the foundations of modern social accounting but also to argue that income taxation wrongly puts saving in double jeopardy. Fisher turned his talents to monetary theory because he suspected that economic instability was largely the fault of existing monetary institutions. His ‘debt-deflation theory of depression’ was motivated by the disasters the Great Depression visited upon the world.

Economics was not the only aspect of human and social life that engaged Fisher's reformist zeal. He was active and prolific in other causes: temperance and Prohibition; vegetarianism, fresh air, exercise and other aspects of personal hygiene; eugenics; and peace through international association of nations.

Fisher was an amazingly prolific and gifted writer. The bibliography compiled by his son lists some 2000 titles authored by Fisher, plus another 400 signed by his associates or written by others about him. Fisher's writings span all his interests and causes. They include scholarly books and papers, articles in popular media, textbooks, handbooks for students, tracts, pamphlets, speeches and letters to editors and statesmen. They include the weekly releases of index numbers, often supplemented by commentary on the economic outlook and policy, issued for thirteen years by Fisher and assistants from the Index Number Institute housed in his New Haven home.

Fisher was the consummate pedagogical expositor, always clear as crystal. He hardly ever wrote just for fellow experts. His mission was to educate and persuade the world. He took the trouble to lead the uninitiated through difficult material in easy stages. Whenever he was teaching or tutoring students, he wrote handbooks or texts for their benefit – in mathematics and science when he was still a student himself, in the principles of economics when he was the professor responsible for the introductory course. Fisher's economics text was published in 1910 and 1911. Its graceful exposition of sophisticated theoretical material will impress a modern connoisseur, but it was too difficult for widespread adoption. Some of it survived in a leading introductory text of the 1920s and 1930s, by the younger Yale economists Fairchild, Furniss and Buck (1926).

A Brief Biography

Irving Fisher grew up and attended school successively in Peace Dale, Rhode Island; New Haven, Connecticut; and St Louis, Missouri. His father, a Congregational minister, died of tuberculosis just when Irving had finished high school and was planning to attend Yale College, his father's *alma mater*. Irving was now the principal breadwinner for himself, his mother and his younger brother. He did have a

$500 legacy from his father for his college education. The family moved to New Haven, and together managed to make ends meet. Irving tutored fellow students during term and in summers.

Fisher was a great success in Yale College, ranking first in his class and winning prizes and distinctions not only in mathematics but across the board. He was also determined to make good in the extra-curricular college culture so important in those days. His efforts won him election to the most prestigious secret senior society, Skull and Bones, the ultimate reward senior campus leaders bestowed on members of the class behind them.

Awarded a scholarship for graduate study, he stayed on at Yale. Graduate Studies were not departmentalized in those days, and Fisher ranged over mathematics, science, social science and philosophy. His most important teachers were Josiah Willard Gibbs, the mathematical physicist celebrated for his theory of thermodynamics, William Graham Sumner, famous still in sociology but at the time also important in political economy, and Arthur Twining Hadley, a leading economist specializing in what is now known as Industrial Organization.

As the time to write a dissertation approached, Fisher had still not chosen his life work. Young Fisher's interests and talents were universal. In the seven years at Yale before he finished his doctorate, he had written and published poetry, political commentary, book reviews, a geometry text together with tables of logarithms, and voluminous notes on mathematics, mechanics and astronomy for the benefit of students he was teaching or tutoring. If he had specialized in anything in six years at Yale, it was mathematics, but even in his graduate years he had spent half his time elsewhere.

Summer put him on to mathematical economics, and in his third year of graduate study, he finished the dissertation that won him worldwide recognition in economic theory. Fisher's 1891 PhD was the first one in pure economics awarded by Yale, albeit by the faculty of mathematics. Although the university, thanks to Summer, Hadley and Henry W. Farnum, was strong in ‘political economy’, there was no distinct department for the subject, let alone for ‘economics’. This was generally the case in American universities. Venturing into mathematical economic theory, Fisher was very much on his own; and his route into economics was quite different from that of most American economists of his era.

The dominant tradition in American political economy was imported from the English classical economists, mainly Smith, Ricardo and John Stuart Mill; it was just beginning to be updated by Marshall. This tradition Fisher's mentors at Yale had taught him well. But the neoclassical developments on the European continent from 1870 on, the works of Warlas and Menger and Böhm-Bawerk, or even those of their English counterparts Jevons and Edgeworth, had been little noticed at Yale or elsewhere in America.

At the time, the main challenge in America to classical political economy was coming from quite a different direction. The American Economic Association was founded in 1886 by young rebels against Ricardian dogma and its *laissez-faire* political and social message. They included Richard T. Ely, J.B. Clark, Edwin R.A. Seligman and other future luminaries of American economics. Many of them had pursued graduate studies in Germany. In the German emphasis on historical, institutional and empirical studies they found welcome relief from implacable classical theory, and in the German faith in the state as an instrument of socially beneficial reform they found a hopeful antidote to the fatalism of economic competition and social Darwinism. Sumner was prominent among several elders who refused to join an Association born of such heresy; he did not relent even though the AEA very soon became sufficiently neutral and catholic to attract his Yale colleagues and other initial holdouts. Fisher, a bit younger than the founding rebels and educated solely at one American university, was not involved. It was his

reconstruction, rather than their revolution, that was destined eventually to replace the classical tradition in the mainstream of American economics.

Fisher stayed at Yale throughout his career. He started teaching mathematics, evidently even before he received his doctorate and was appointed Tutor in Mathematics. His first economics teaching was under the auspices of the mathematics faculty, an undergraduate course on ‘The Mathematical Theory of Prices’. In 1894–5 during his *Wanderjahr* in Europe, this young American star was welcomed by the leading mathematically inclined theorists in every country. On his return he became Assistant Professor of Political and Social Science and began teaching economics proper. He was appointed full Professor in 1898 and retired in 1935.

Fisher was struck by tuberculosis in 1898. He spent the first three years of his professorship on leave from Yale and from science, recuperating in more salubrious climates. His lifelong crusade for hygienic living dates from this personal struggle to regain health and vigour. The experience powerfully reinforced his determination to gain ‘a place among those who have helped along my science’ and his ambition ‘to be a *great* man’, as he wrote to his wife (I.N. Fisher, 1956, pp. 87–8). After his recovery the books and articles began flowing from his pen, never to stop until his death at the age of 80.

Fisher participated actively in teaching and in university affairs until 1920. Thereafter his writings and his myriad outside activities and crusades preoccupied him. He taught only half time and had little impact on students, undergraduate or graduate. Thus Fisher had few personal disciples; there was no Fisherian School. The student to whom Fisher was closest, personally and intellectually, was James Harvey Rogers, a 1916 PhD who returned to Yale as a professor in 1930. His career was prematurely ended by his tragic death in a plane crash in 1939 at the age of 55.

Fisher was, on top of everything else, an inventor. His most successful and profitable invention was the visible card index system he patented in 1913. In 1925 Fisher's own firm, the Index Visible Company, merged with its principal competitor to form Kardex Rand Co., later Remington Rand, still later Sperry Rand. The merger made him wealthy. However, he subsequently lost a fortune his son estimated to amount to 8 or 10 million dollars, along with savings of his wife and her sister, when he borrowed money to exercise rights to buy additional Rand shares in the bull market of the late 1920s.

More than money was at risk in the market. Fisher had staked his public reputation as an economic pundit by his persistent optimism about the economy and stock prices, even after the 1929 crash. His reputation crashed too, especially among non-economists in New Haven, where the university had to buy his house and rent it to him to save him from eviction. Until the 1950s the name Irving Fisher was without honour in his own university. Except for economic theorists and econometricians, few members of the community appreciated the genius of a man who lived among them for 63 years.

Irving Fisher's marriage to Margaret Hazard in 1893 was a very happy one for 47 years. She died in 1940. They had two daughters and one son, his father's biographer. The death of their daughter Margaret in 1919 after a nervous breakdown was the greatest tragedy of her parents’ lives. Their daughter Carol brought them two grandchildren.

General Equilibrium Theory

Fisher's doctoral dissertation (1892) is a masterly exposition of Walrasian general equilibrium theory. Fisher, who was meticulous about acknowledgements throughout his career, writes in the preface that he was unaware of Walras while writing the dissertation. His personal mentors in the literature of

economics were Jevons (1871) and Auspitz and Lieben (1889).

Fisher's inventive ingenuity combined with his training under Gibbs to produce a remarkable hydraulic-mechanical analogue model of a general equilibrium system, replete with cisterns, valves, levers, balances and cams. Thus could he display physically how a shock to demand or supply in one of ten interrelated markets altered prices and quantities in all markets and changed the incomes and consumption bundles of the various consumers. The model is described in detail in the book; unfortunately both the original model and a second one constructed in 1925 have been lost to posterity. Anyway Fisher was a precursor of a current Yale professor, Herbert Scarf (1973) and other practitioners of computing general equilibrium solutions. In his formal mathematical model-building too, Fisher was greatly impressed by the analogies between the thermodynamics of his mentor Gibbs and economic systems, and he was able to apply Gibbs's innovations in vector calculus.

Fisher expounds thoroughly the mathematics of utility functions and their maximization, and he is careful to allow for corner solutions. He uses independent and additive utilities of commodities in his first mathematical approximation and in his physical model; later he was to show how this assumption could be exploited to measure marginal utilities empirically (1927). But the general formulation in his dissertation makes the utility of every commodity depend on the quantities consumed of all commodities. At the same time, he states clearly that neither interpersonally comparable utility nor cardinal utility for each individual is necessary to the determination of equilibrium. Fisher's list of the limitations of his analysis is candid and complete. The supply side of Fisher's model is, as he acknowledges, primitive. Each commodity is produced at increasing marginal cost, but neither factor supplies and prices nor technologies are explicitly modelled.

Finally, Fisher shows his enthusiasm for his discovery of mathematical economics by appending to his dissertation as published an exhaustive survey and bibliography of applications of mathematical method to economics.

General equilibrium with intertemporal choices and opportunities

The distribution of income and wealth, and in particular the sources, determinants and social rationales, of interest and other returns to private property, were obsessive topics in economics, both in Europe and North America, at the turn of the century. One important reason, especially in Europe, was the Marxist challenge to the legitimacy of property income. Answering Marx was a strong motivation for the Austrian school, in particular for the capital theory of Böhm-Bawerk and his followers. Neoclassical economics was in a much better position than its classical precursor to respond to the Marxist challenge. The labour theory of value, which Marx borrowed from the great classical economists themselves, neither explains nor justifies functionally or ethically incomes other than wages.

These topics engaged the two leading American economists of the era, John Bates Clark and Fisher. Clark (1899) set forth his marginal productivity theory of distribution, arguing that a generalized factor of production, capital, the accumulation of past savings, has like labour a marginal product that explains and justifies the incomes of its owners.

Fisher attacked these problems in a more elegant, abstract, mathematical, general and ethically neutral manner than Clark, and than Böhm-Bawerk. At the same time, his approach was clearer, simpler and more insightful than that of Walras.

The general equilibrium system of Fisher's dissertation was a single-period model. No intertemporal

choices entered; hence the theory was silent on the questions of capital and interest. But Fisher took up these subjects soon after.

His first contribution, one that should not be underestimated, was to set straight the concepts and the accounting. This he did in (1896) and (1906) with clarity and completeness that have scarcely been surpassed. It's all there: continuous and discrete compounding; nominal versus real rates; the distinction between high prices and rising prices, and its implications for observations of interest rates; the inevitable differences among rates computed in different *numéraires*; rates to different maturities and consistency among them; appreciation, expected and unexpected; present values of streams of in- and out-payments; and so on. Schumpeter calls this work ‘the first economic theory of accounting’ and says ‘it is (or should be) the basis of modern income analysis’ (1954, p. 872).

Perhaps the most remarkable feature is Fisher's insistence that ‘income’ is consumption, including of course consumption of the services of durable goods. In principle, he says, income is psychic, the subjective utility yielded by goods and services consumed. More practically, income could be measured as the money value, or value in some other *numéraire*, of the goods and services directly yielding utility, but only of those. Receipts saved and invested, for example in the purchase of new durable goods, are not ‘income’ for Fisher; they will yield consumption and utility later, and those yields will be income. To include both the initial investment and the later yields as income is, according to Fisher, as absurd as to count both flour and bread in reckoning net output. This view naturally led Fisher to oppose conventional income taxation as double taxing of saving, and to favour consumption taxation instead. His views on these matters are loudly echoed today.

Fisher published his theory of the determination of interest rates in *The Rate of Interest* (1907). A revised and enlarged version was published in 1930 as *The Theory of Interest*. One motivation for the revision was that Fisher's many critics apparently did not understand the 1907 version. They typically concentrated on the ‘impatience’ side of Fisher's theory of intertemporal allocation and missed the ‘opportunities’ side. It was there in 1907 already; the theory is much the same in both versions.

In 1930 Fisher is at pains to label his theory the ‘impatience and opportunity’ theory. ‘Every essential part of it’, he acknowledges, ‘was at least foreshadowed by John Rae in 1834.’ He does claim originality for his concept of ‘investment opportunity’. This turns on ‘the rate of return over cost, [where] both cost and return are differences between two optional income streams’ (1930, p. ix). As Keynes acknowledged, this is the same as his own ‘marginal efficiency of capital’ (Keynes, 1936, p. 140).

In these books Fisher extended general equilibrium theory to intertemporal choices and relationships. This strategy was different from Walras. Walras tried to extend his multicommodity multi-agent model of exchange to allow for production, saving and investment. This maintained his stance of full generality but was also difficult to expound and to understand. Fisher saw that intertemporal dependences were tricky enough to justify isolating them from the intercommodity complexities that had concerned him in his doctoral thesis. Therefore he proceeded as if there were just one aggregate commodity to be produced and consumed at different dates. This simplification enabled him to illuminate the subject more brightly than Walras himself.

The methodology of Fisher's capital theory is very modern. His clarifications of the concepts of capital and income lead him to formulate the problem as determination of the time paths of consumption – that is, income – both for individual agents and for the whole economy. Then he divides the problem into the two sides, tastes and technologies, that are second nature to theorists today. One need only read Böhm-Bawerk's murky mixture of the two in his list of reasons for the agio of future over present consumption

to realize that Fisher's procedure was not instinctive in those times.

Fisher's theory of individual saving is basically the standard model to this day. Undergraduates learn the two-period ‘Fisher diagram’, where a family of indifference curves in the two commodities consumption now c_1 and consumption later c_2 confront a budget constraint $c_1+c_2/(1+r)=y_1+y_2(1+r)$, where the y's are exogenous wage incomes in the two periods and r is the (real) market interest rate. From the usual tangency can be read the consumption choices and present saving or dissaving. This is indeed a Fisher diagram, but of course he went much beyond it.

He stated clearly what we now call the ‘life cycle’ model, explaining why individuals will generally prefer to smooth their consumption over time, whatever the time path of their expected receipts. But he was not dogmatic, and he allowed room for bequests and for precautionary saving. Where Fisher differed from later theorists, and especially from contemporary model-builders, was in his unwillingness to impose any assumed uniformity on the preferences (or expectations of ‘endowments’ – the latter term was not familiar to him though the concept was) of the agents in his economies, and in his scruples against buying definite results by assuming tractable functional forms. In general, many of the advances claimed in present-day theory appear to depend on greater boldness in these respects.

On the side of technology, Fisher's approach was the natural symmetrical partner of his formulation of preferences, equally simple, abstract and general. He assumed that the ‘investment opportunities’ available to an individual (not necessarily the same for everybody) and to the society as a whole can be summarized in the terms on which consumption at any date can be traded, with ‘nature’, for consumptions at other dates. In modern language, we would say that Fisher postulated intertemporal production possibility frontiers, properly convex in their arguments, consumptions at various dates.

All that remained for Fisher, then, was to assume complete intertemporal loan markets cleared by real interest rates, count equations, and show that in principle the equalities of saving and investment at every date determine all interest rates and the paths of consumption and production for all individuals and for the society. Like hundreds of mathematical theorists since, he set the problem up so that it conformed to a paradigm he knew, in this case the Walrasian paradigm of his own doctoral dissertation. A more rigorous proof of the existence of the equilibria Fisher was looking for came much later, from Arrow and Debreu (1954). As we know, the problems of infinity, whether agents are assumed to have infinite or finite horizons, are much more troublesome than Fisher imagined.

In any event, Fisher had an excellent vantage point from which to comment on the controversies over capital and interest raging in his day. His formulation of ‘investment opportunities’ seems to allow for no factor of production one could call ‘capital’ and enter as argument in a production function. For that matter, he doesn't explicitly model the role of labour in production either, or of land. Strangely, in Fisher's insistence that interest is *not* a cost of production, he seems to say that labour is the only cost, evidently because labour and labour alone is a source of disutility, the loss of utility from leisure, the opportunity cost of the consumption afforded by work. Proceeding in the same spirit, he postulates that, from a position of equality of present and planned future consumption a typical individual will require more extra future consumption than present consumption as compensation for extra work. The difference, the agio, is interest, whether or not it is a ‘cost’. Fisher attributes the agio to ‘impatience’, at the same time scorning the notion that interest is the cost of securing the services of a factor of production called ‘abstinence’ or ‘waiting’.

In the 1890s and 1900s Knut Wicksell, discovering marginal productivity independently of Clark, was modelling production as a function of labour and land inputs with the output also depending on the lags

between those inputs and the harvests (Wicksell [1911], 1934, vol. I, pp. 144–66). This is an ‘Austrian’ formulation, akin to Böhm-Bawerk's examples of trees and wine, in which time itself appears to be productive. Fisher rightly objects to any generalization that waiting longer increases output. His own intertemporal frontiers are, to be sure, sufficiently general to encompass such technologies. They can also accommodate Leontief input-output tables and Koopmans-Dantzig activity matrices with lags, Hayekian triangular structures with inventories of intermediate goods in process, Solow technologies with durable goods and labour jointly yielding output contemporaneously or later. The only common denominator of these and other representations of technology is that they relate consumption opportunities at different dates to one another, though not necessarily always in the convex trade-off terms Fisher assumed. There does not appear to be any summary scalar measure to which the productivity of a process is generally monotonically related, whether roundaboutness, average period of production, or replacement value of existing stocks of goods.

Fisher describes himself as an advocate of ‘impatience’ as an explanation of interest, although he realizes there are two sides of the saving-investment market, and although he acknowledges that real interest rates can at times be zero or negative. He does appear to believe that in a stationary equilibrium with constant consumption streams, consumers will require positive interest, and that only those technologies and investment opportunities affording a ‘rate of return over cost’ equal to this pure time preference rate would be used. He does not face up to Schumpeter's argument in 1912 that in such a repetitive and riskless ‘circular flow’, rational consumers would not care whether a marginal unit of consumption occurs today or tomorrow (Schumpeter [1912], 1934, pp. 34–6). Like Böhm-Bawerk, Fisher appeals to the shortness and uncertainty of life as a reason for time preference. For life-cycle consumers, however, time preferences are entangled with age preferences, and it is hard to defend any generalization as to their net direction. Fair annuities take care of the uncertainty.

Monetary theory: the equation of exchange and the quantity theory

Irving Fisher was the major American monetary economist of the early decades of this century; the subject occupied him until the end of his career. Here especially Fisher combined theorizing with empirical research, both historical and statistical. The problems he encountered led him to invent statistical and econometric methods – index numbers and distributed lags in particular – to apply for the purposes at hand to the data he and his assistants compiled. (He even studied the turnover of cash and checking accounts of a sample of Yale students, professors and employees.)

Money was a big subject in American economic literature in the 19th century, before Fisher came on the scene. The monetary events of the times – the inconvertible greenbacks issued during the Civil War, their redemption in gold in 1879, the demonetization of silver, the rapidly increasing importance of banks – stimulated research and controversy. Nevertheless, monetary theory was relatively undeveloped and unsystematized, both in Europe and in America. Fisher's treatise (1911a) was an ambitious attempt to organize with the help of theory a large body of historical and institutional information.

Yet for all its theory, statistics and index numbers, *The Purchasing Power of Money* is a tract supporting Fisher's proposal for stabilizing the value of money. This came to be known as the ‘compensated dollar’, the gold-exchange standard combined with a rule mandating periodic changes in the official buying and selling prices of gold inverse to changes in a designated commodity price index. In 1911 Fisher proposed that the gold price changes be uniform and synchronous in the currencies of all countries

linked by fixed exchange parities, in proportional amounts related to an international price index. Later he was willing to accept as second best that the United States adopt the scheme on its own. Keynes proposed a similar but less formal rule for the United Kingdom (1923).

The proposal is an early example of a policy *rule*, another Fisherian idea ahead of its time, more likely to be popular among economists today than it was with Fisher's contemporaries. Indeed, some rules recently proposed are quite Fisherian, for example Hall (1985).

The ‘compensated dollar’ is but one of several proposals Fisher advanced over the years for stabilizing price levels or mitigating the effects of their unforeseen variation. In the 1911 book he also writes favourably of the ‘tabular standard’, which meant no more operationally than facilitating priceindexed indexed contracts. In the 1920s he launched a crusade for 100 per cent reserves against checkable deposits, culminating in *100% Money* (1935). This idea is also beginning to resurface in the 1980s as a preventive defence against the monetary hazards of bank failures. In Schumpeter's view, Fisher's zeal for monetary reforms lost him some of the attention and respect his scientific contributions to monetary economics deserved, and made him come across as more monetarist than his own analysis and evidence justified (Schumpeter, 1954, pp. 872–3).

The Purchasing Power of Money is a monetarist book. Fisher asserts the quantity theory as earnestly and persuasively as Milton Friedman. There are two species of quantity theories. One is a simple implication of the ‘classical dichotomy’: since only relative prices and real endowments enter commodity and factor demand and supply functions, the solution values for real variables in a general equilibrium are independent of scalar variations of exogenous nominal quantities. While Fisher mentions this implication of general equilibrium theory, he does not dwell upon it as one might expect. Anyway, it does not quite apply to a commodity money system like the gold standard, which Fisher was analysing. Fisher's theory is mainly of the second kind, based on the demand for and supply of the particular nominal assets serving as media of exchange.

Fisher is usually given credit for the Equation of Exchange, although Simon Newcomb, a celebrated figure in American astronomy as well as an economist, had anticipated him (1886, pp. 315–47). The Equation is the identity $MV=PT$, where M is the stock of money; V its velocity, the average number of times per year a dollar of the stock changes hands; P is the average price of the considerations traded for money in such transactions; and T is the physical volume per year of those considerations. It is an identity because it is in principle true by definition. Actually Fisher, of course, recognized the heterogeneity of transactions by writing also $MV=\sum p_iQ_i$, where the p_i and Q_i are individual prices and quantities. His interest in index numbers was substantially a quest for aggregate indexes P and T derived from the individual p_i and Q_i in such a way that the two forms of the equation would be consistent. Much of the book (1911a), both text and technical appendices, is devoted to this quest.

Here and in later writings, particularly (1921) and (1922), Fisher was looking for the ‘best’ index number formula. He postulated certain criteria and evaluated a host of formulas, investigating their properties both *a priori* and from applications to data. Since the criteria inevitably conflict, there can be no formula that excels on all counts. Although Fisher was mainly interested in measuring movements of the aggregate price level, naturally he wanted a price index P and a quantity index T to have the property that $P_1T_1/P_0T_0=(\sum p_1Q_1)/(\sum p_0Q_0)$, where the subscripts represent two time periods at which observations of p's and Q's are available.

This and various other desirable consistency properties are not hard to meet. The difficult question is the

choice of weights in the two indexes, especially when a whole series of consistent period-to-period comparisons is desired, not just one isolated comparison. For a price index, should the quantity weights be those of a fixed base year, yielding what we now call a ‘Laspeyres’ index $(\sum p_1Q_0)|(\sum p_0Q_0)$? Or should the weights be those of the ever-changing current period, yielding a ‘Paasche’ index $(\sum p_1Q_1)|(\sum p_0Q_1)$? The indicated correlate quantity indexes would be the opposites, respectively ‘Paasche’ and ‘Laspeyres’. In 1911 Fisher opted for the Paasche price index. He also seemed to approve the idea of chain indexes, in which the period 0 of the above formulas is not fixed in calendar time but is always the prior period, even though these violate one possible desideratum, that the relative change between two periods should be independent of the base used. He also wrote favourably of the practical advantages of an entirely different procedure, namely taking the median of an expenditure-weighed distribution of percentage price changes from one period to the next.

In 1920, however, Fisher proposed as the ‘Ideal Index’ a candidate he had not ranked high in 1911, namely the geometric mean of the Laspeyres and Paasche formulas. This formula has the pleasant property that the correlate of an Ideal price index is an Ideal quantity index. Correa Walsh, another index number expert, on whose comprehensive treatise (1901) Fisher relied heavily from the beginning of his own investigations, reached the same conclusion independently at about the same time (Walsh, 1921).

These index number issues do not seem as important to present-day economists as they did to Fisher. Knowing that they are intrinsically insoluble, we finesse them and use uncritically the indexes that government statisticians provide. But Fisher's explorations have been important to those practitioners.

In Fisher's Equation of Exchange (1911a) the T and the Q_i are measures of all transactions involving the tender of money, intermediate goods and services as well as final goods and services, old goods as well as newly produced commodities, financial assets as well as goods. The corresponding velocity is likewise comprehensive, much more so than the ‘income’ or ‘circuit’ velocity preferred by some monetary theorists, notably Alfred Marshall and his followers in Cambridge (England), who count only transactions for final goods, for example for Gross National Product.

Fisher elaborated the equation to distinguish the quantities M and M' of the two media currency and checking deposits and their separate velocities V and V' $MV+M'V'=PT$. This was a bow to the rising importance of bank deposits relative to currency as transactions media. Previous practice counted only government-issued currency as money, in modern parlance high-powered or base money, and regarded bank operations as increasing its velocity rather than adding to a money stock.

How does the quantity theory come out of the Equation of Exchange? Fisher argues that the real volume of money-using transactions T is exogenous; that the velocities are determined by institutions and habits and are independent of the other variables in the equation; that the division of the currency supply, the monetary base in current terminology, between currency and bank reserves is stable and independent of the variables in the equation; that banks are fully ‘loaned up’ so that deposits M' are a stable multiple of reserves, determined by the prudence of banks and by regulation; that exogenous changes in currency supply itself are the principal source of shocks, which, given the preceding propositions, move price level P proportionately. The many qualifications for transitional adjustments are conscientiously presented, but the monetarist message is loud and clear.

The argument is familiar to modern readers, but certain features deserve notice:

1. (1) Fisher gives the most illuminating account available of the institutions and habits that

generate the society's demand for transactions media relative to the volume of transactions. He rightly emphasizes the fact that, and the degree to which, receipts and payments are imperfectly synchronized. He seeks the determinants of velocity in such features of social and economic structure as the frequency of wage and bill payments and the degree of vertical integration of firms. His belief that these institutions change only slowly supports his contention that velocities are exogenous constants.

2. (2) Much ink has been spilled on the difference between Fisher's velocity approach to money demand and the Cambridge (England) ‘k’ formulation. The latter, like Walras's *encaisse desiré*, directs attention to agents’ portfolio decisions. To Fisher's critics that seems behavioural, while velocity is mechanical. The issue is overblown; the same phenomena can be described in either language. If the other variables in the equation are defined and measured the same way, then V and k are just reciprocals each of the other. Fisher himself discusses hoarding. Fisher's explicit attention, in discussing economy-wide demand for circulating media in distinction to other stores of value, to the fact that money ‘at rest’ soon takes ‘wing’ to fly from one agent to another seems to be a merit of his approach.
3. (3) As already noted, Fisher resolved a question current in his day, whether banks’ creation of deposit substitutes for currency should be regarded as increasing the velocity of basic money or as enlarging the supply of money. His choice of the latter course compels attention to the structure, behaviour and regulation of banks. He could not be expected to foresee that the proliferation of future candidates for designation as ‘money’ would create the monetarist ambiguities we see today.
4. (4) For the most part later writers have not followed Fisher in his preference for a comprehensive concept and measure of transactions volume. It is hard to attach meaning to the *real* volume of financial transactions, and therefore to see why a T that includes them should be a constant or exogenous term in the equation. On the other hand, modern students of money demand tend simply to forget transactions other than those on final payments.
5. (5) Fisher ignores the possibility that other liquid assets can serve as imperfect substitutes for money holdings because they can be converted into means of payment as needed, though at some cost. Partly for this reason, he ignores interest rate effects on demand for transactions media. In his day there may have been more excuse for these omissions than there was later. But they are still surprising for an author who elsewhere pays so much attention to the effects of interest rates and opportunity costs on behaviour.
6. (6) When Fisher was writing, the United States was on the gold standard; the exchange parities of the dollar with sterling and other gold-standard currencies were fixed. Fisher discusses in detail the implications of foreign transactions for the elements of the Equation of Exchange and for the quantity theory. He recognizes that tendencies towards purchasing-power parity, even though imperfect, make money supplies in any one country endogenous, tie prices to those of other countries and enhance quantity adjustments to monetary shocks in the short run. Much of the 1911 book applies, therefore, to the gold standard economies in aggregate. Indeed, Fisher finds the increase in gold production after 1896 to be the main cause of price increases throughout the world.

Macroeconomics: business fluctuations and the Great Depression

The quantity theory by no means exhausts Fisher's ideas on macroeconomics. His views were much more subtle then straightforward monetarism, but they are scattered through his writings and not systematically integrated. Consider the following non-neutralities emphasized by Fisher:

(1) Probably Fisher's principal source of fame, especially among non-economists, is his equation connecting nominal interest *i*, real interest *r* and inflation π: $i=r+\pi$. It is frequently misused. Like the Equation of Exchange, it is first of all an identity, from which, for example, an unobservable value of *r* can be calculated from observations of the other two variables. More interesting, certainly to Fisher, is its use as a condition of equilibrium in financial markets; for this purpose π must be replaced by expected inflation π^e, another unobservable. In a longer run, as Fisher recognized, steady-state equilibrium would also be characterized by equality of actual and expected inflation: $\pi = \pi^e$.

The Fisher equation is frequently cited nowadays in support of complete and prompt pass-through of inflation into nominal interest rates. Fisher's view throughout his career was quite different. For one thing, neither Fisher's theory of interest nor his reading of historical experience suggested to him that equilibrium real rates of interest should be constant. Moreover, from (1896) on he believed that adjustment of nominal interest rates to inflation takes a very long time. This he confirmed by sophisticated empirical investigations, regressions in which the formation of inflation expectations was modelled by distributed lags on actual inflation. During the transition, inflation would lower real rates; nominal rates would adjust incompletely. The effect was symmetrical; he attributed the severity of the Great Depression to the high real rates resulting from price deflation.

Moreover, Fisher was quite explicit about the effects of these movements of real interest rates on real economic variables, including aggregate production and employment. In *The Purchasing Power of Money* these transitional effects are mentioned, but minimized in the author's zeal to convince readers of the importance of stabilizing money stocks. But in Fisher's writings on interest rates, the transitions turn out to be long. In his accounts of cyclical fluctuations in business activity, and especially of the Great Depression, they play the key role.

(2) An assiduous student of price data, Fisher knew that some prices were more flexible than others, that money wages were on the sticky side of the spectrum, and that the imperfect flexibility of the price level meant that the *T* on the right-hand side of his Equation of Exchange would absorb some of the variations of the left-hand side.

In the early 1930s he came to a very modern position. Real variables like production and employment are independent of the level of prices, once the economy has adjusted to the level. But they are not independent of the rate of change of prices; they depend positively on the rate of inflation. He even calculated a ‘Phillips’ correlation between employment and inflation (1926). He was just one derivative short of the accelerationist position (Friedman, 1968); in a little more time he would have made that step, aware as he was of the difference between actual and expected inflation. Anyway, his policy conclusion was that stabilizing the price level would also stabilize the real economy.

(3) During the Great Depression, observing the catastrophes of the world around him, which he shared personally, Fisher came to quite a different theory of the business cycle from the simple monetarist version he had espoused earlier. This was his ‘debt-deflation theory of depression’ (1932), summarized in the first volume of *Econometrica*, the organ of the international society he helped to found (1933). The essential features are that debt-financed Schumpeterian innovations fuel a boom, followed by a recession which can turn into depression via an unstable interaction between excessive real debt burdens

and deflation. Note the contrast to the Pigou real balance effect, according to which price declines are the benign mechanism that restores full-employment equilibrium. The realism is all on Fisher's side. This theory of Fisher's has room for the monetary and credit cycles of which he earlier complained, and for the perversely pro-cyclical real interest rate movements mentioned above.

Fisher did not provide a formal model of his latter-day cycle theory, as he probably would have done at a younger age. The point here is that he came to recognize important non-monetary sources of disturbance. These insights contain the makings of a theory of a determination of economic activity, prices, and interest rates in short and medium runs. Moreover, in his neoclassical writings on capital and interest Fisher had laid the basis for the investment and saving equations central to modern macroeconomic models. Had Fisher pulled these strands together into a coherent theory, he could have been an American Keynes. Indeed the ‘neoclassical synthesis’ would not have had to wait until after World War II. Fisher would have done it all himself.

His practical message in the early 1930s was ‘Reflation!’ When his Yale colleagues and orthodox economists throughout the country protested against public-works spending proposals and denounced Roosevelt's gold policies, Fisher was a conspicuous dissenter. He was right. Characteristically, he crusaded vigorously for his cause – in speeches, pamphlets, letters and personal talks with President Roosevelt and other powerful policy-makers. Characteristically too, as his letters home (I.N. Fisher, 1956, p. 275) disclose, he saw clearly and unapologetically that in lobbying for what was good for the country he was also hoping to rescue the Fisher family finances.

Addressing the President of Yale shortly after Fisher's death, Joseph Schumpeter and eighteen colleagues in the Harvard economics department wrote, ‘No American has contributed more to the advancement of his chosen subject . The name of that great economist and American has a secure place in the history of his subject and of his country.’ According to his son, this is the eulogy that would have pleased Irving Fisher the most (I.N. Fisher, 1956, pp. 337–8). Today, four decades later, economists can confirm the judgement and prediction of that eulogy.

Author's Note: Fortunately Fisher's son, Irving Norton Fisher, preserved the memory of his father in two indispensable publications, a biography and a comprehensive bibliography (1956, 1961). I have also relied extensively on Professor John Perry Miller's biographical essay (1967) and Professor William Barber's account (1986) of political economy at Yale before 1900. My review of Fisher's contributions to general equilibrium theory, the theory of capital and interest, monetary theory and macroeconomics draws heavily and often literally on a recent essay of my own (Tobin, 1985).

Selected works

1892. *Mathematical Investigations in the Theory of Value and Prices*. New Haven: Connecticut Academy of Arts and Sciences, *Transactions 9*, 1892. Reprinted, New York: Augustus M. Kelley, 1961.

1896. Appreciation and interest. *AEA Publications* 3(11), August, 331–442. Reprinted, New York: Augustus M. Kelley, 1961.

1906. *The Nature of Capital and Income*. New York: Macmillan.

1907. *The Rate of Interest*. New York: Macmillan.

1910. *Introduction to Economic Science*. New York: Macmillan.

1911a. *The Purchasing Power of Money*. New York: Macmillan.

1911b. *Elementary Principles of Economics*. New York: Macmillan.

1921. The best form of index number. *American Statistical Association Quarterly* 17, March, 533–7.

1922. *The Making of Index Numbers*. Boston: Houghton Mifflin.

1926. A statistical relation between unemployment and price changes. *International Labour Review* 13, June, 785–92.

1927. A statistical method for measuring ‘marginal utility’ and testing the justice of a progressive income tax. In *Economic Essays Contributed in Honor of John Bates Clark*, ed. J.H. Hollander. New York: Macmillan.

1930. *The Theory of Interest*. New York: Macmillan.

1932. *Booms and Depressions*. New York: Adelphi.

1933. The debt-deflation theory of great depressions. *Econometrica* 1(4), October, 337–57.

1935. *100% Money*. New York: Adelphi.

Bibliography

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22(3), July, 265–90.

Auspitz, R. and Lieben, R. 1889. *Untersuchungen über die Theorie des Preises*. Leipzing Duncker & Humblot.

Barber, W.J. 1986. Yale the fortunes of political economy an environment of academic conservatism. In *Economists and American Higher Learning in the Nineteenth Century*, ed. W.J. Barber. Middletown, Conn.: Wesleyan University Press.

Clark, J.B. 1899. *The Distribution of Wealth*. New York: Macmillan.

Fairchild, F.R., Furniss, E.S. and Buck, N.S. 1926. *Elementary Economics*. 2 vols, New York: Macmillan. 5th edn, 1948.

Fisher, I.N. 1956. *My Father Irving Fisher*. New York: Comet Press.

Fisher, I.N. 1961. *A Bibliography of the Writings of Irving Fisher*. New Haven: Yale University Library.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1), 1–17.

Hall, R.E. 1985. Monetary policy with an elastic price standard. In *Price Stability and Public Policy*, Federal Reserve Bank of Kansas City, 137–60.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. New York: Harcourt, Brace.

Miller, J.P. 1967. Irving Fisher of Yale. In *Ten Economic Studies in the Tradition of Irving Fisher*, ed. William Fellner et al., New York: Wiley.

Newcomb, S. 1886. *Principles of Political Economy*. New York: Harper.

Rae, J. 1834. *The Sociological Theory of Capital* Reprinted, New York: Macmillan, 1905.

Samuelson, P.A. 1967. Irving Fisher and the theory of capital. In *Ten Economic Studies in the Tradition of Irving Fisher*, ed. William Fellner et al., New York: Wiley.

Scarf, H. (With T. Hansen.) 1973. *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Schumpeter, J.A. 1912. *Theory of Economic Development*. Trans. from the 2nd German edn of 1926 by R. Opie, Cambridge, Mass.: Harvard University Press, 1934.

Schumpeter, J.H. 1954. *History of Economic Analysis*. Ed. E.B. Schumpeter, New York: Oxford University Press.

Social Sciences Citation Index. 1979, 1983. *Five Year Cumulation*, 1971–5 and 1976–80. Philadelphia: Institute for Scientific Information.

Tobin, J. 1985. Neoclassical theory in America. *American Economic Review* 75(6), December, 28–38.

Walsh, C.M. 1901. *The Measurement of General Exchange Value*. New York and London: Macmillan.

Walsh, C.M. 1921. *The Problem of Estimation*. London: King & Sons.

Wicksell, K.M. 1911. *Lectures on Political Economy*. Trans E. Classen from the 2nd Swedish edn, London: George Routledge & Sons, 1934.

How to cite this article

Tobin, James. "Fisher, Irving (1867–1947)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000126> doi:10.1057/9780230226203.0581

The New Palgrave Dictionary of Economics Online

Fisher, Ronald Aylmer (1890 – 1962)

A.W.F. Edwards

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

ancillarity; conditional inference; fiducial probability; Fisher, R. A.; game theory; likelihood; natural selection; statistical inference

Article

R.A. Fisher was born in London on 17 February 1890, the son of a fine-art auctioneer. His twin brother was stillborn. At Harrow School he distinguished himself in mathematics, despite being handicapped by poor eyesight which prevented him working by artificial light. His teachers used to instruct by ear, and Fisher developed a remarkable capacity for pursuing complex mathematical arguments in his head. This manifested itself in later life in his ability to reach a conclusion whilst forgetting the argument; to handle complex geometrical trains of thought; and to develop and report essentially mathematical arguments in English (only for students to have to reconstruct the mathematics later).

He entered Gonville and Caius College, Cambridge, as a scholar in 1909, graduating BA in mathematics in 1912. Prevented from entering war service in 1914 by his poor eyesight, Fisher held several jobs before being appointed Statistician to Rothamsted Experimental Station in 1919. In 1933 he became Galton Professor of Eugenics at University College London, and in 1943 Arthur Balfour Professor of Genetics in Cambridge and a Fellow of Caius College. He retired in 1957 and spent his last few years in Adelaide, Australia, where he died from a post-operative embolism on 29 July 1962.

He married Ruth Eileen Guiness in 1917 and they had two sons and six daughters. He was elected a Fellow of the Royal Society in 1929 and was knighted in 1952 for services to science.

Fisher made a most profound contribution to applied and theoretical statistics and to genetics. He had been attracted to natural history, and especially the works of Darwin, at school, and he had bought Bateson's *Principles of Genetics*, with its translation of Mendel's paper, in his first term as an undergraduate. Before graduating he had already remarked on the surprisingly good fit of Mendel's data, published a paper introducing the method of maximum likelihood, and given a proof of the distribution of the '*t*' statistic which Student had only conjectured.

In 1915 Fisher published the distribution of the correlation coefficient; in 1918 the seminal work in

biometrical genetics, ‘The correlation between relatives on the supposition of Mendelian inheritance’, in which he introduced the word ‘variance’ and foreshadowed his later development of the analysis of variance; and in 1922 ‘On the Mathematical Foundations of Theoretical Statistics’, a paper which revolutionized statistical thought.

As Statistician at Rothamsted he founded the subject of experimental design based on randomization, pursued vigorously the development of statistical estimation theory and invented – or, at least, captured – the quixotic notion of fiducial probability. Moving to London the pace did not slacken, for in addition to pioneering genetical work, especially in connection with the human blood groups, Fisher's statistical explorations revealed the likelihood principle, conditional inference and the concept of ancillarity.

The Second World War found him embattled on many fronts. Unhappy at home, he found his scientific activity disrupted by wartime conditions including the evacuation of his department from London. The profundity of his work on statistical inference was ill-appreciated in America, where preoccupation with wartime problems encouraged an excessively mathematical and operational view with which Fisher had little sympathy. In mathematical genetics there were similar difficulties as the American school, starting from his ‘fundamental theorem of natural selection’, developed ideas of ‘adaptive topographies’ with false analogies to physical systems. It was not until well after his death that in both statistical inference and mathematical genetics the criticisms which he had advanced came to be appreciated.

After the war, from the relative peace of Cambridge, Fisher saw his theoretical work in both subjects suffer further temporary eclipse. He made great, but ultimately unsuccessful, efforts to establish biochemical genetics in his department and to secure for Cambridge the national laboratories for human blood-group work. When close to retirement, he was amongst the first to realize the significance of Watson and Crick's discovery of the structure of DNA (1953), and to apply the new computers to a biological problem (1950).

Perhaps embittered by his post-war experiences (though he never relaxed his scientific work), he found some consolation in the Presidency of Caius College from 1956 to 1959, a post second to the Master, and further happiness in retirement in Adelaide.

Fisher wrote five books and published a famous set of statistical tables jointly with F. Yates. An extremely informative and admirably objective biography was published by one of his daughters in 1978 (Box, 1978).

In the field of economics Fisher's name would be remembered for his contributions to statistics alone, so fully chronicled in Box's biography, but we may here draw attention to three other areas not emphasized in the biography but which are especially relevant.

First, the ‘fundamental theorem of natural selection’ (1930). Although this is specifically directed at a genetical problem, it relies on a simpler implicit theorem of widespread relevance wherever discussion centres on differential growth rates, namely ‘the rate of change in the growth-rate is proportional to the variance in growth-rates’. This precise theorem, which is easily proved mathematically, captures the notion that the growth rate of the fastest-growing sub-population (or economic sector, and so on) will come to dominate the overall growth rate.

Secondly, the modern preoccupation with ‘socio-biology’ has as one of its origins *The Genetical Theory of Natural Selection* (1930), a fact that only surprises those who have not studied the book and Fisher's other writings on human affairs in the two decades before the Second World War.

Thirdly, Fisher not only introduced the theory of games into evolutionary biology (at the suggestion of Dr Cavalli, later Professor Cavalli-Sforza), but he discovered and published the idea of a randomized or

‘mixed’ strategy as early as 1934, independently of von Neumann. The problem was the card game ‘*Le Her*’, though if Fisher had gone to the primary source (the correspondence between Montmort and Nicholas Bernoulli, published in 1713) rather than relying only on Todhunter's *History of the Mathematical Theory of Probability* (1865), he would have found that his solution had already been given by Waldegrave.

Selected works

1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 507–21.

1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399–433.

1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions. Royal Society of London, Series A* 222, 309–68.

1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

1930. *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

1938. With F. Yates. *Statistical Tables for Biological, Agricultural and Medical Research*. Edinburgh: Oliver and Boyd.

1949. *The Theory of Inbreeding*. Edinburgh: Oliver and Boyd.

1950. *Contributions to Mathematical Statistics*. New York: Wiley.

1956. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

1971–4. *Collected Papers of R.A. Fisher*, 5 vols, ed. J.H. Bennett. Adelaide: University of Adelaide.

Bibliography

Box, J.F. 1978. *R.A. Fisher: The Life of a Scientist*. New York: Wiley.

Bibliographic addendum

A celebrated discussion of Fisherian statistics is L.J. Savage, ‘On Rereading R.A. Fisher (with

discussion)', *Annals of Statistics* 4 (1976), 441–500. Other interesting writings include B. Efron, 'R.A. Fisher in the 21st Century', *Statistical Science* 31(2) (1998), 95–122, which demonstrates the continuing direct relevance of Fisher's work; C.R. Rao, 'R.A. Fisher: The Founder of Modern Statistics', *Statistical Science* 7 (1992), 34–48, an appreciation by a former student; and S. Stigler, 'Fisher in 1921', *Statistical Science* 20 (2005), 32–49, which discusses the origins of Fisher's monumental 1922 paper.

How to cite this article

Edwards, A. W. F. "Fisher, Ronald Aylmer (1890–1962)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000127> doi:10.1057/9780230226203.0582

The New Palgrave Dictionary of Economics Online

fisheries

Colin W. Clark

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Marine fisheries throughout the world continued to be severely overexploited throughout the 20th century and beyond. Even under intensive ‘scientific’ management many important fisheries have collapsed, some never to recover. Vast overcapacity of fishing fleets is also widespread. Both outcomes can be attributed to the common-pool aspect of fishery resources. One method of countering these developments, individual transferable catch quotas (ITQs), is currently in use in several countries. Provided this instrument is combined with resource taxes (royalties), an efficient and equitable management system is feasible. (Owing to lack of jurisdiction, deep-sea fisheries seem destined to continue to suffer from overfishing.)

Keywords

bionomic equilibrium; common property resources; extended fisheries jurisdiction (EFJ) zones; fisheries; individual fishing quotas (IFQs); individual transferable quotas (ITQs); maximum sustainable yield (MSY) paradigm; precautionary management; resource rents; royalities

Article

By the end of the 1970s, most of the world's coastal states had declared 200-nautical-mile zones of extended fisheries jurisdiction (EFJ zones) over marine fisheries. These zones allowed coastal states to exert full control over fishing activities. Over 90 per cent of global marine fishery landings thus came under the control of coastal states.

The need to regulate fishing activities arises from the common-pool nature of marine fish (and other living-resource) populations. In simple terms, under unregulated open-access conditions any fish stock that can be profitably harvested will in fact be exploited. Whether such exploitation eventually leads to biological depletion and reduced harvest levels depends on a number of circumstances, including demand for the product, cost of fishing, and the distribution, abundance and behaviour of the fish.

Using a simple graphical model, H.S. Gordon (1954) argued that an unregulated fishery would achieve ‘bionomic’ equilibrium, reaching a stock level at which the revenue from catching and selling fish would just balance the opportunity costs of fishing. Populations with high price–cost ratio would thus be heavily exploited, while those with low ratio would remain lightly exploited or unexploited. Countless actual examples support this prediction.

The next obvious question is whether bionomic equilibrium is undesirable and, if so, what can be done about it. Early commentators largely agreed that, because bionomic equilibrium results in the dissipation of economic rents, it is economically undesirable, independent of any biological consequences. But in many cases bionomic equilibrium also implies biological overfishing, defined as the reduction of the fish population to a level at which net annual biological productivity is below the maximum that could be generated. Indeed, in extreme cases overfishing can lead to the collapse of the fishery, with little or no recovery after fishing is terminated (Dulvy, Sadovy and Reynolds, 2003). Pauly et al. (1998), Myers and Worm (2003) and other scientists have documented the historical decline of marine fish stocks on a worldwide scale, especially over recent decades. Even many stocks within 200-mile EFJ zones have continued to be overfished. The reasons for this outcome are only beginning to be generally understood.

Fisheries management has traditionally been based on the objective of determining and achieving the maximum sustainable yield (MSY) available from each population. In its own right, this approach is beset with difficulties generated by unobservability and uncertainty pertaining to marine populations and ecosystems (Caddy and Seijo, 2005). Management difficulties also arise because the MSY paradigm overlooks all economic aspects of fishing.

To be more specific, until recently most fishery management programmes have been based almost exclusively on the total allowable catch (TAC) method. The annual TAC is calculated on the basis of an accepted model, and the fishery is managed so as to achieve this quota, usually by means of restricted annual openings of the fishery. If correctly calculated and implemented, this method can indeed prevent overfishing and produce positive economic rents – temporarily.

But *positive rents attract additional fishing effort* – this is the basis of Gordon's original theory. If annual effort is controlled through TAC-based seasonal openings, the response is that either the fishermen increase the power and capacity of their vessels, or additional vessels enter the fishery, or both. Further shortenings of the fishing season are then needed, and so on. A new regulated bionomic equilibrium is reached when the average capital costs of expanding fishing capacity are just equal to the average present value of net operating revenues. Rents in a TAC-managed fishery are then dissipated through over-expansion of fishing capacity rather than via overfishing. Extreme overcapacity of fishing fleets worldwide is today considered to be a major impediment to rational management.

It seems natural, therefore, to attempt to control fishing capacity. In cases where excess capacity has already developed, vessel buy-back programmes have often been used to reduce fleet size. However, such buy-back programmes, which can be very costly, do nothing to eliminate the incentives for additional expansion. Indeed, if buy-backs are anticipated by fishermen, they may actually induce a higher level of initial overcapacity than would otherwise occur (Clark, Munro and Sumaila, 2005).

Two possible approaches to resolving the joint problems of overfishing and overcapacity are, first, taxes or royalties, and second, individual fishing quotas (IFQs). Although these are usually considered as alternatives, they can in fact readily be used in combination. By the early 21st century IFQs and related programmes were in effect in several countries, with generally positive results (Cunningham and

Bostock, 2005; Clark, 2006).

IFQs can be envisioned as a form of quasi-property rights, an interpretation that is strengthened if the quotas are tradable (that is, individual transferable quotas, ITQs). Its owner considers an ITQ as a productive asset, whose value will be enhanced if the resource is protected and well-managed. It also seems likely (though this remains to be demonstrated in practice) that ITQ owners will favour risk-averse management strategies, such as conservative TACs and the use of marine reserves.

Various economic distortions can arise, however, if ITQs are awarded free of charge. For example, the initial recipients of the quotas may become greatly enriched. This possibility being well known to fishermen, the anticipation of a forthcoming ITQ programme may attract extra entry into the fishery, dissipating much of the future rents in advance. Besides this there is the question of social equity – why should the government award special access privileges to a publicly owned resource to a chosen few individuals? Charging significant catch royalties can reduce rent-seeking incentives, while also compensating the resource owner, namely, the general public.

During the current transitional phase from managing ocean fisheries as common-pool resources to managing them with individual quotas, royalty charges will probably remain minimal. But, once a profitable fishery develops, it seems likely that the public will expect and demand a fair share of the resource rents – as is already the case with other natural resource assets.

Whatever system is used, the management of marine fisheries will always face high levels of uncertainty. Marine ecosystems are complex, poorly observable, and subject to unpredictable, environmentally induced fluctuations. Finely tuned management, intended for example to maximize some specified objective, will remain elusive. The threat of overfishing persists, even for closely monitored and managed stocks. Also, recent experience has shown that the recovery of depleted stocks can often be slow or non-existent (Hutchings, 2000).

For these reasons it is now widely agreed that a precautionary management approach is needed (Charles, 2001). Conservative annual catch quotas are necessary to protect against inadvertent overfishing. In addition, breeding stocks need to be strongly protected, as do sea-floor and estuarine habitats, and marine ecosystems in general. Furthermore, fishing activities that damage and degrade the marine environment, leading to long-term reductions in productivity, need to be controlled or eliminated.

Marine reserves, permanently protecting substantial areas of the ocean from harvesting activities, can provide a valuable hedge against management error resulting from biological uncertainty or from imperfect control of fishing operations. Such reserves can protect breeding stocks, ensuring a continued supply of recruits even when stocks are overfished elsewhere. Reserves are not a substitute for well-designed and operated traditional management systems; rather, they need to be used in conjunction with normal management methods.

Space limitations preclude the discussion of other important issues such as: ocean pollution, aquaculture, illegal fishing and non-regulated deep-sea fisheries, and ecosystem-based management programmes.

See Also

- common property resources

Bibliography

Caddy, J. and Seijo, J. 2005. This is more difficult than we thought: the responsibility of scientists, managers and stakeholders to mitigate the unsustainability of marine fisheries. *Philosophical Transactions of the Royal Society B* 360, 59–75.

Charles, A. 2001. *Sustainable Fishery Systems*. Oxford: Blackwell Science.

Clark, C. 2006. *Worldwide Crisis in Fisheries*. Cambridge: Cambridge University Press.

Clark, C., Munro, G. and Sumaila, U. 2005. Subsidies, buybacks and sustainable fisheries. *Journal of Environmental Economics and Management* 50, 47–58.

Cunningham, S. and Bostock, T., eds. 2005. *Successful Fisheries Management*. Delft: Eburon Academic Publishers.

Dulvy, N., Sadovy, Y. and Reynolds, J. 2003. Extinction vulnerability in marine populations. *Fish and Fisheries* 4, 24–64.

Gordon, H. 1954. The economic theory of a common property resource: the fishery. *Journal of Political Economy* 62, 124–42.

Hutchings, J. 2000. Collapse and recovery of marine fishes. *Nature* 406, 882–5.

Myers, R. and Worm, B. 2003. Rapid worldwide depletion of predatory fish communities. *Nature* 423, 280–3.

Pauly, D., Christensen, V., Dalsgaard, J., Froese, R. and Torres, F. 1998. Fishing down marine food webs. *Science* 279, 860–3.

How to cite this article

Clark, Colin G. "fisheries." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000128> doi:10.1057/9780230226203.0583

The New Palgrave Dictionary of Economics Online

fixed effects and random effects

Badi H. Baltagi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Unobservable individual effects in panel data models are employed to control for heterogeneity. These can be thought of as random variables that are uncorrelated with the regressors, thus generating a random effects model. Alternatively, these random individual effects are allowed to be completely correlated with the regressors, thus generating a fixed effects model. The choice between these two alternatives is usually settled using a Hausman (1978) test. This article argues that one should interpret a rejection by the Hausman test as a rejection of the random effects model, not necessarily an endorsement of the fixed effects model.

Keywords

attrition; autocorrelation; cross-section data; fixed effects; Haavelmo, T.; heteroskedasticity; instrumental variable estimators; least squares dummy variables (LSDV) model; multicollinearity; over-identification; panel data; random effects; sample selection

Article

One of the major benefits from using panel data as compared to cross-section data on individuals is that it enables us to control for individual heterogeneity. Not controlling for these unobserved individual specific effects leads to bias in the resulting estimates. Consider the panel data regression

$$y_{it} = \alpha + X_{it}'\beta + u_{it} i = 1, \ldots, N; t = 1, \ldots, T \quad (1)$$

with i denoting individuals and t denoting time. The panel data is *balanced* in that none of the observations is missing whether randomly or non-randomly due to attrition or sample selection. α is a scalar, β is $K\times1$ and X_{it} is the itth observation on K explanatory variables. Most panel data applications utilize a one-way error component model for the disturbances, with

$$u_{it} = \mu_i + v_{it} \tag{2}$$

where μ_i denotes the *unobservable* individual specific effect and v_{it} denotes the remainder disturbance. For example, in an earnings equation in labour economics, y_{it} will measure earnings of the head of the household, whereas X_{it} may contain a set of variables like experience, education, union membership, sex, or race. Note that μ_i is time-invariant and it accounts for any individual specific effect that is not included in the regression. In this case we could think of it as the individual's unobserved ability. The remainder disturbance v_{it} varies with individuals and time and can be thought of as the usual disturbance in the regression. If the μ_i's are assumed to be *fixed parameters* to be estimated, we get the *fixed effects (FE) model*. If the μ_i's are assumed random variables independent of X_{it} and ν_{it}, for all i and t, we get the *random effects (RE) model*.
For the fixed effects model, the regression equation in (1) becomes

$$y_{it} = \alpha + X_{it}'\beta + \mu_i + v_{it} \tag{3}$$

where the μ_i's can be estimated as coefficients of dummy variables, one for each individual. This model is also known as the least squares dummy variables (LSDV) model. Note that only $(\alpha + \mu_i)$ is estimable and that is why it is sometimes denoted by α_i. For large labour or consumer panels, where N is very large, LSDV regressions like (3) may not be feasible. In this case, one is including (N–1) dummy variables in the regression and therefore inverting a huge matrix of dimension $(N + K)$ rather than (K+1) as in (1). In addition, this FE regression suffers from a large loss of degrees of freedom, since we are estimating (N–1) extra parameters, and too many dummies may aggravate the problem of multicollinearity among the regressors. In particular, this FE estimator cannot estimate the effect of any time-invariant variable like gender, race, religion which may be of prime interest for the researcher especially in attempting to estimate wage differentials among men and women or whites and non-whites, with other factors held constant. In fact, these time-invariant variables are spanned by the individual dummies in (3) and therefore any OLS regression attempting to estimate (3) will fail, signalling perfect multicollinearity. Averaging (3) over time yields

$$\bar{y}_{i.} = \alpha + \bar{X}_{i.}'\beta + \mu_i + \bar{v}_{i.}$$
(4)

Subtracting (4) from (3) gives

$$y_{it} - \bar{y}_{i.} = (X_{it} - \bar{X}_{i.})'\beta + (v_{it} - \bar{v}_{i.}).$$
(5)

One can show that the FE estimator of β (denoted by $\hat{\beta}_{FE}$) obtained from the sometimes infeasible LSDV regression in (3) can be alternatively obtained from the simpler regression given in (5). The latter regression is known as the *within*-regression since it is based on the within variation in the data. Regression (4), which is a cross-section regression, is known as the *between*-regression since it is based on the between variation in the data. If (3) is the true model, FE is the best linear unbiased estimator (BLUE) as long as the remainder disturbances (the v_{it}'s) are independent and identically distributed (i.i.d.)$(0, \sigma_v^2)$. Of course, here we are assuming that the X_{it}'s are independent of the v_{it} for all i and t. The fixed effects model is deemed appropriate when one is focusing on a specific set of N countries, states, counties, regions or firms. Inference in this case is conditional on the particular N firms, countries or states that are observed. Note that, if T is fixed and $N \to \infty$ as typical in short labour panels, then only the FE estimator of β is consistent; the FE estimators of the individual effects (α_i) are not consistent since the number of these parameters increases as N increases. This is the *incidental parameter problem* discussed by Neyman and Scott (1948) and reviewed more recently by Lancaster (2000). Note that, when the true model is fixed effects as in (3), pooled OLS on (1) yields biased and inconsistent estimates of the regression parameters. This is an omission variables bias because OLS deletes the individual dummies when in fact they are relevant. One could test the joint significance of these dummies, that is, $H_0: \mu_1 = \mu_2 = \ldots = \mu_{N-1} = 0$, by performing an F-test. This is a simple Chow test with the restricted residual sums of squares (RRSS) being that of OLS on the pooled model and the unrestricted residual sums of squares (URSS) being that of the LSDV regression in (3) or equivalently the residual sum of squares from the within-regression in (5). In this case

$$F_0 = \frac{(RRSS - URSS)\ /\ (N-1)}{URSS\ /\ (NT - N - K)} \overset{H_0}{\sim} F_{N-1, N(T-1)-K}.$$
(6)

One computational caution for those using the within-regression computed from (5). The s^2 of this regression as obtained from a typical regression package divides the residual sums of squares by $NT–K$ since the intercept and the dummies are not included. The proper s^2, say s^{*2} from the LSDV regression in (3), would divide the same residual sums of squares by $N(T–1)–K$. Therefore, one has to adjust the variances obtained from the within-regression by multiplying the variance-covariance matrix by (s^{*2}/s^2) or simply by multiplying by $[NT - K] / [N(T - 1) - K]$. For robust estimates of the standard errors for the FE model, see Arellano (1987).

For the random effects model, $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$, $\nu_{it} \sim \text{IID}(0, \sigma_\nu^2)$ and the μ_i's are independent of the v_{it}'s. In addition, the X_{it}'s are independent of the μ_i and ν_{it}, for all i and t. The random effects model is an appropriate specification if we are drawing N individuals randomly from a large population, and we have no endogeneity between the regressors and the disturbances. For household panel studies, special attention is usually taken in the design of the panel to make it ‘representative’ of the population we are trying to make inferences about. In this case, N is usually large, and a fixed effects model would lead to an enormous loss of degrees of freedom. The individual effect is characterized as random, and inference pertains to the population from which this sample was randomly drawn. But what is the population in this case? Nerlove and Balestra (1992) emphasize Haavelmo's (1944) view that the population ‘consists *not* of an infinity of individuals, in general, but of an infinity of *decisions*’ that each individual might make. They argue that the fixed effects model may be more appropriate in cases where the population is sampled exhaustively (like data from geographic regions over time), whereas the random effects model is more consistent with Haavelmo's view given above. They argue that what differentiates individuals, who make the decisions with which we are concerned, is largely historical. Taking a leaf from Knight (1921), they argue that these inheritances from the past are material goods and appliances, knowledge and skill, and morale. In a dynamic context, this means that the primary reasons for heterogeneity among individuals is the different history each one has.

The random effects model implies a homoskedastic variance $\text{var}(u_{it}) = \sigma_\mu^2 + \sigma_\nu^2$ for all i and t, and an equi-correlated block-diagonal covariance matrix which exhibits serial correlation over time only between the disturbances of the same individual. In fact,

$$\text{cov}(u_{it}, u_{js}) = \sigma_\mu^2 + \sigma_\nu^2 \text{for} i = j, t = s = \sigma_\mu^2 \quad \text{for} i = j, t \neq s$$

and zero otherwise. This also means that the correlation coefficient between u_{it} and u_{js} is

$$\rho = \text{correl}(u_{it}, u_{js}) = 1 \text{ for } i = j, t = s = \sigma_\mu^2 / (\sigma_\mu^2 + \sigma_\nu^2) \text{ for } i = j, t \neq s$$

and zero otherwise. In this case, the BLUE of the regression coefficients is GLS which can be obtained from a least squares regression of $y^*_{it} = y_{it} - \theta\bar{y}_{i.}$ on $X^*_{it} = X_{it} - \theta\bar{X}_{i.}$ and a constant (see Fuller and Battese, 1974). The GLS estimator of β for this random effects model will be denoted by $\hat{\beta}_{RE}$. Here $\theta = 1 - (\sigma_v / \sigma_1)$ and $\sigma_1^2 = T\sigma_u^2 + \sigma_v^2$. Note that (i) if $\sigma_u^2 = 0$ then $\theta = 0$ and $\hat{\beta}_{RE}$ reduces to $\hat{\beta}_{OLS}$ since y^*_{it} reduces to y_{it}; (ii) if $T \to \infty$, then $\theta \to 1$ and $\hat{\beta}_{RE}$ tends to $\ddot{\beta}_{FE}$ since y^*_{it} reduces to $\ddot{y}_{it}$. The variance components can be estimated from the between- and within-variation of the disturbances:

$$\hat{\sigma}_1^2 = T\sum_{i=1}^{N} \hat{u}_{i.}^2 / (N - K - 1) \tag{7}$$

and

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{u}_{it}^2}{[N(T-1) - K]} \tag{8}$$

where $\hat{u}_{i.}$ denotes the between-residuals from (4). Note that (7) is T times the s^2 of the between-regression obtained in (4). Also, $\ddot{u}_{it}$ denotes the FE residuals from (5). So, (8) is the s^2 of the FE regression obtained in (5). Substituting these estimates for the variance components in θ and running y^*_{it} on X^*_{it} yields a feasible GLS or RE estimator suggested by Swamy and Arora (1972). For alternative estimators of the variance components, see Baltagi (2005). These are implemented using standard econometric software, including EViews, Stata, TSP, RATS and LIMDEP, to mention a few.

After this discussion of the fixed effects and the random effects models and the assumptions underlying them, the reader is left with the daunting question: which to choose? This is not as easy a choice as it might seem. In fact, the fixed versus random effects issue has generated a hot debate in the biometrics and statistics literature, which has spilled over into the panel data econometrics literature. Economists cannot perform natural experiments of, say, the effect of fertilizer brand on crop yield controlling for the effect of land and other inputs. We have to deal with human subjects whose individual effects may be correlated with the regressors even when we randomly draw these individuals. Mundlak (1961) and Wallace and Hussain (1969) were early proponents of the fixed effects model, and Balestra and Nerlove (1966) were advocates of the random effects model. The modern econometric interpretation of the μ_i's

is that they are random variables but in the RE model the $E(\mu_i / X_{it}) = 0$. This implies that the individual effects are uncorrelated with the regressors. This is a strong assumption given economists preoccupation with endogeneity issues. For example, in an earnings equation, μ_i may denote the unobservable ability of the individual and this may be correlated with the schooling variable included as a regressor. In this case, $E(\mu_i / X_{it}) \neq 0$ and the RE estimator $\hat{\beta}_{RE}$ becomes biased and inconsistent for β. However, the within-transformation wipes out these μ_i's and leaves the FE estimator $\ddot{\beta}_{FE}$ unbiased and consistent for β. Hausman (1978) suggested comparing $\hat{\beta}_{RE}$ and $\ddot{\beta}_{FE}$, both of which are consistent under the null hypothesis H_0; $E(\mu_i / X_{it}) = 0$. In this case, the contrast $\hat{q} = \hat{\beta}_{RE} - \tilde{\beta}_{FE}$ will have $\text{plim}\, \hat{q} = 0$ under H_0. However, if H_0 is not true, $\text{plim}\, \hat{q} \neq 0$ and the Hausman test statistic is given by

$$m = \hat{q}' [\text{var}(\hat{q})]^{-1} \hat{q} \tag{9}$$

Under H_0 this is asymptotically distributed as χ^2_K where K denotes the dimension of slope vector β. For significant values of m, we reject the consistency of the RE estimator. Since $\hat{\beta}_{RE}$ is the efficient estimator under the null hypothesis H_0, one can show that the $\text{cov}(\hat{q}, \hat{\beta}_{RE}) = 0$ and that the $\text{var}(\hat{q}) = \text{var}(\tilde{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE})$. This makes the computation of (9) simple. Nevertheless, Hausman (1978) suggested an alternative asymptotically equivalent test to (9) that can be obtained from the augmented regression

$$y^* = X^* \beta + \ddot{X} \gamma + w \tag{10}$$

where $y^*_{it} = y_{it} - \theta \bar{y}_{i.}$, $X^*_{it} = X_{it} - \theta \bar{X}_{i.}$ and $\ddot{X}_{it} = X_{it} - \bar{X}_{i.}$. Hausman's test is now equivalent to testing whether $\gamma = 0$. This is a standard Wald test for the omission of the FE regressors $\ddot{X}$ from the RE regression. For an alternative variable addition test that produces a Hausman test which is robust to autocorrelation and heteroskedasticity of arbitrary form, see Arellano (1993).

Note that the FE model allows for endogeneity of the regressors and the individual effects, whereas the RE model does not. This is why the FE model is more popular among economists. Mundlak (1978) assumed that the individual effects are a linear function of the averages of *all* the explanatory variables across time, that is,

$$\mu_i = \bar{x}_i' \pi + \varepsilon_i \tag{11}$$

where $\varepsilon_i \sim \text{IIN}(0, \sigma_\varepsilon^2)$ and $\bar{x}_{i.}'$ is $1 \times K$ vector of observations on the explanatory variables averaged over time. These effects are uncorrelated with the explanatory variables if and only if $\pi = 0$. In fact, a test for $\pi = 0$ yields the Hausman (1978) test based on the contrast between the FE and the between-estimators. Mundlak (1978) shows that GLS on (3) augmented with (11) yields $\hat{\beta}_{FE}$. Only if $\pi = 0$ does it yield $\hat{\beta}_{RE}$. This all-or-nothing choice of correlation between the individual effects and the regressors prompted Hausman and Taylor (1981) to suggest a model where *some* of the regressors are correlated with the individual effects. They proposed an instrumental variable estimator, denoted by HT, which uses both the between- and within-variation of the strictly exogenous variables as instruments. More specifically, the individual means of the strictly exogenous regressors are used as instruments for the time invariant regressors that are correlated with the individual effects (see Baltagi, 2005, for more details). The over-identification conditions are testable. In fact, this is a Hausman test based upon the contrast between the FE and the HT estimators.

Most applications in economics since the 1980s have made the choice between the RE and FE estimators based upon the standard Hausman test. If this standard Hausman test rejects the null hypothesis that the conditional mean of the disturbances given the regressors is zero, the applied researcher reports the FE estimator. Otherwise, the researcher reports the RE estimator. Unfortunately, applied researchers have interpreted a rejection as an adoption of the fixed effects model and non-rejection as an adoption of the random effects model. Chamberlain (1984) showed that the fixed effects model imposes testable restrictions on the parameters of the reduced form model and one should check the validity of these restrictions before adopting the fixed effects model (see also Angrist and Newey, 1991). For the applied researcher, performing fixed effects and random effects and the associated Hausman test, it is important to carry this analysis a step further. Test the restrictions implied by the fixed effects model derived by Chamberlain (1984) before accepting the FE estimator and check whether a Hausman and Taylor (1981) specification might be a viable alternative.

See Also

- artificial regressions
- dummy variables
- Haavelmo, Trygve
- linear models

Bibliography

Angrist, J.D. and Newey, W.K. 1991. Over-identification tests in earnings functions with fixed effects. *Journal of Business and Economic Statistics* 9, 317–23.

Arellano, M. 1987. Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics* 49, 431–4.

Arellano, M. 1993. On the testing of correlated effects with panel data. *Journal of Econometrics* 59, 87–97.

Balestra, P. and Nerlove, M. 1966. Pooling cross-section and time-series data in the estimation of a dynamic model: the demand for natural gas. *Econometrica* 34, 585–612.

Baltagi, B.H. 2005. *Econometric Analysis of Panel Data*. Chichester: Wiley.

Chamberlain, G. 1984. Panel data. In *Handbook of Econometrics*, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Fuller, W.A. and Battese, G.E. 1974. Estimation of linear models with cross-error structure. *Journal of Econometrics* 2, 67–78.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement), 1–118.

Hausman, J.A. 1978. Specification tests in econometrics. *Econometrica* 46, 1251–71.

Hausman, J.A. and Taylor, W.E. 1981. Panel data and unobservable individual effects. *Econometrica* 49, 1377–98.

Knight, F.H. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

Lancaster, T. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95, 391–413.

Mundlak, Y. 1961. Empirical production function free of management bias. *Journal of Farm Economics* 43, 44–56.

Mundlak, Y. 1978. On the pooling of time series and cross-section data. *Econometrica* 46, 69–85.

Nerlove, M. and Balestra, P. 1992. Formulation and estimation of econometric models for panel data. In *The Econometrics of Panel Data: Handbook of Theory and Applications*, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Neyman, J. and Scott, E.L. 1948. Consistent estimation from partially consistent observations.

Econometrica 16, 1–32.

Swamy, P.A.V.B. and Arora, S.S. 1972. The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica* 40, 261–75.

Wallace, T.D. and Hussain, A. 1969. The use of error components models in combining cross-section and time-series data. *Econometrica* 37, 55–72.

How to cite this article

Baltagi, Badi H. "fixed effects and random effects." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000295> doi:10.1057/9780230226203.0584

The New Palgrave Dictionary of Economics Online

fixed factors

Walter Y. Oi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Small firms invest relatively less in custom-made machines and specifically trained employees. The overhead costs of fixed-capital assets are relatively larger for big firms that engage in the volume production of standardized products. Large firms also incur higher fixed employment costs to recruit and train a specialized workforce. Workers in large firms are paid higher wages designed to reduce labour turnover rates. These phenomena could not be explained without a formal analysis of fixed and quasi-fixed factors. A continuum of degrees of fixity makes for a richer theory of factor markets than a dichotomy of fixed versus variable factors.

Keywords

amortization; barriers to entry; Clark, J. M.; elasticity of substitution; firm size; firm-specific factors; fixed factors; human capital; implicit contracts; labour as a quasi-fixed factor; labour market search; labour markets contracts; monitoring costs; overhead costs; rationing; shadow price; specialization; substitutes and complements; training; wage differentials

Article

In moving from one market equilibrium to another, a firm may choose to hold fixed the rate of employment of one or more factors of production. The presence of fixed factors and their associated overhead costs will affect the firm's responses to changing market conditions. The residually determined quasi-rents which constitute the returns to the fixed factors must, in the long run, cover their overhead costs; otherwise, the inputs of fixed factors have to be contracted. The importance of fixed factors and overhead costs, which varies across firms and industries, was analysed by J.M. Clark (1923), who emphasized the first of the following three questions: (1) How do fixed factors affect the behaviour of prices, outputs and inputs of variable factors? (2) What determines whether a factor of production will be fixed or variable? (3) How do the fixed employment costs of quasi-fixed labour inputs affect

contractual arrangements in labour markets?

In the short run, certain paths of adjustment are barred to the firm. The usual assumption is that the input of one or more factors is fixed. Total unit costs, which include the outlays for fixed factors, lie above average variable costs so that price, in the short run, can remain well below the minimum long-run average cost. If fixed costs in an industry are high, they can pose a barrier to entry of new firms and could result in wide short-run fluctuations in price. Further, the upper-bound constraint on inputs of fixed factors affects the firm's demand for the remaining variable inputs in a manner analogous to the theory of rationing of consumer goods analysed by E. Rothbarth (1941). An increase in the demand for the final product raises the shadow price of the fixed factor, which increases the demand for variable factors that are substitutes for the fixed factor and decreases the demand for complementary variable factors. This result could explain the greater cyclical volatility in the demand for unskilled labour relative to skilled labour if unskilled labour is a closer substitute for the fixed factor, capital. Moreover, the smaller the elasticity of substitution of labour for capital, the steeper is the slope of the marginal cost curve, implying larger cyclical swings in product prices.

A firm will fix the input rate of a factor if (*a*) the factor is specific to the firm in the sense that employment in this firm constitutes its highest valued use, or (*b*) reallocation to some higher-valued use is precluded by some contractual agreement or by a prohibitively high transaction cost. In the former case, equipment, buildings and even labour can be specialized to fit into a firm's idiosyncratic production methods. The internal values of such specialized resources are likely to exceed their external values to outside users. These resources are more likely to be owned (rather than hired or leased), because of their specificity. Long-term contracts that account for some fixed factors occur where there are gains from risk-sharing or high costs of transferring resources to other firms.

A richer theory of factor markets can be developed if the dichotomy of fixed versus variable factors is replaced by a continuum of degrees of fixity. The discipline of labour economics has now accepted the principle that labour is a quasi-fixed factor. The cost of hiring and training workers constitutes the fixed component of the full cost of labour, while the variable component is the wage paid to the employee. In long-run equilibrium, the expected marginal value product which depends on the expected product price P^* and labour's marginal physical product f_N, is equated to the full labour cost:

$$P^* f_N = W + q, \left[q = \left(\frac{F}{r}\right)(1 - e^{-rT}) \right]$$

where W is the wage, and q is the periodic rent that amortizes the fixed employment cost F at a discount rate r over the worker's expected period of employment T. The gap between the wage and labour's marginal value product will be relatively larger, the higher is the degree of fixity which can be measured by $f = q / W + q$.

The cyclical behaviour of the labour market is characterized by an uneven incidence of unemployment, a compression of occupational wage differences in the upswing, persistent differences in labour turnover rates, hiring/firing practices that smack of discrimination. The quasi-fixity of labour goes a long way in explaining these phenomena. In the downswing, the product price falls below its long-run level P^*. If

labour is a completely variable input, meaning that $q = F = 0$, its marginal value product Pf_N will be equated to the wage in each period. Hence, when P falls, the demand for this grade of labour is contracted until f_N climbs to restore equilibrium in both factor and product markets. However, if labour is a quasi-fixed factor, the periodic amortization of the fixed cost drives a wedge between the wage and marginal value product. For a small decline in product price, the firm will not contract the demand for a quasi-fixed grade of labour as long as its short run MVP exceeds the wage, which is the variable cost of labour; that is, if $Pf_N > W$ even though $Pf_N < (W + q)$, the input of this grade of labour will not be reduced in the downswing. There is, for each quasi-fixed factor, a trigger price P_i at which the firm will choose to reduce employment. The trigger price which induces a decline in factor demand will be lower for factors with higher degrees of fixity. In the early stages of a downturn, labour with low degrees of fixity will become unemployed, while other workers will be retained until the drop in product price P is driven below P_T. At the trough of a cycle, most grades of labour satisfy a short-run equilibrium condition where labour's MVP is equated to its variable cost, $Pf_N = W$. As P rises in the recovery, a firm will increase its demand for a quasi-fixed factor if the price rise is such that labour's MVP exceeds its *full cost*; that is, employment is expanded if and only if $Pf_N > (W + q)$. In the upturn, the rightward shift in factor demand will be greater for factors with lower degrees of fixity. Employment will be more stable, and the incidence of unemployment will be lower for those workers in occupations with higher degrees of fixity.

Some firms find that it is profitable to incur the fixed employment costs of assembling a firm-specific workforce. Recruiting is the means by which an employer identifies more productive individuals and ascertains whether an applicant will meet prescribed hiring standards. Recruitment for high-wage positions usually entails higher costs because of the variability of individual productivities. Employers who have well-defined internal labour markets and who organize production around teams also incur higher recruiting costs. In an internal labour market, workers are hired at a limited number of ports of entry and are typically given on-the-job training to adapt them to the firm's idiosyncratic production methods. Larger investments in firm-specific human capital are indicative of the greater specialization of the labour input. Firm-specific training is less profitable when labour turnover rates are high due either to the high separation propensities of workers or the low survival odds of firms. Smaller firms spend less on recruiting and appear to invest less in formal training. The estimates reported by Oi (1962) and Parsons (1972) reveal that employers incurred substantially higher fixed employment costs for workers in higher skill levels. The degree of fixity, $f = q / (W + q)$, is positively related to the wage rate W, and this relation allows us to test the implications of a theory of labour as a quasi-fixed factor. Employees in high-wage occupations experience greater employment stability over the cycle. Occupational wage differentials widen in the downswing and narrow in the upswing. Labour turnover rates are lower, and recruiting costs are higher in large firms whose workforces exhibit a higher degree of fixity.

The persistence of unemployment and the failure of wages to clear labour markets call for an explanation. Some unemployed workers are in a state of pseudo-idleness while they look for work: ‘When actively searching for work, the situation is that he is really investing in himself by working on his own account without immediate remuneration. He is prospecting’ (Hutt, 1977, p. 83). The time and money spent by new entrants and disemployed workers in their search for suitable job matches constitute a fixed cost which has to be recovered over the course of the employment relation. Each job is, in a very real sense, specialized to the worker–firm attachment. In a search model, unemployment can

be efficient in two senses. First, it may be the least-cost means of finding a durable job. Second, a worker on a temporary layoff may stay in a state of availability awaiting recall rather than seeking work. Labour turnover is costly, both to the employer for whom labour is a quasi-fixed factor due to the fixed investments in hiring and training, as well as to the employee for whom this job is specific due to the fixed costs of search. Both parties have incentives to form an implicit contract that can raise the returns to these fixed employment costs by lengthening the expected period of employment.

Long-term employment contracts could be the result of risk-averse workers seeking job security. An employer can reduce his full labour costs by providing a tacit agreement in which the risks of income variability are shared. Such long-term agreements end up increasing the fixity of labour. Implicit, long-term contracts may also result from an employer's desire to discourage shirking and dishonesty. Firms will incur monitoring and enforcement costs to deter dysfunctional behaviour and malfeasance. These enforcement costs can be reduced by designing compensation packages which reward workers with separation pay and pensions if they perform in accordance with prescribed work standards. Stable and durable employment relations make sense only when there are fixed costs of forging and maintaining specific jobs defined by worker–firm attachments.

When physical or human capital is specialized to a firm, it must capture any quasi-rents that it can because the fixed investments in these specialized resources cannot be reallocated to some alternative use. Fixed, firm-specific factors only make sense in a world of heterogeneous firms. In Oi (1983) I advanced the thesis that firm-specific capital was systematically related to firm size. Small firms with low survival odds do not invest in custom-made machines and specifically trained employees. They are more likely to purchase used assets and to hire inexperienced workers with general human capital. The overhead costs of fixed-capital assets are relatively larger for big firms that engage in the volume production of standardized products. Large firms also incur higher fixed employment costs to recruit and train a specialized workforce. Workers in large firms are paid higher wages and are provided with employee compensation packages that are designed to reduce labour turnover rates. These phenomena could not be explained without a formal analysis of fixed and quasi-fixed factors.

See Also

- rent

Bibliography

Clark, J.M. 1923. *Studies in the Economics of Overhead Costs*. Chicago: University of Chicago Press.

Hutt, W.H. 1977. *The Theory of Idle Resources*. Indianapolis: Liberty Press.

Oi, W.Y. 1962. Labor as a quasi-fixed factor. *Journal of Political Economy* 70, 538–55.

Oi, W.Y. 1983. The fixed employment costs of specialized labor. In *The Measurement of Labor Costs*, ed. J.E. Triplett. Chicago: University of Chicago Press.

Parsons, D.O. 1972. Specific human capital: an application to quit rates and layoff rates. *Journal of Political Economy* 80, 1120–43.

Rothbarth, E. 1941. The measurement of changes in real income under conditions of rationing. *Review of Economic Studies* 8, 100–7.

How to cite this article

Oi, Walter Y. "fixed factors." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000134> doi:10.1057/9780230226203.0585

The New Palgrave Dictionary of Economics Online

fixed point theorems

Andrew McLennan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article gives statements of the Tarski fixed point theorem and the main versions of the topological fixed point principle that have been applied in economic theory. Pointers are given to literature concerned with proofs of Brouwer's theorem, and with algorithms for computing approximate fixed points. The topological results are all consequences of a slightly weakened version of the Eilenberg–Montgomery (1946) fixed point theorem. The axiomatic characterization of the Leray–Schauder fixed point index (which is even more powerful) is also stated, and its application to issues concerning robustness of sets of equilibria is explained.

Keywords

absolute neighbourhood retract; algebraic topology; Brouwer's fixed point th; contraction mapping th; convexity; cooperative game theory; cooperative game theory (core); Debreu–Gale–Kuhn–Nikaido lemma; Eilenberg–Montgomery th; essential sets of fixed points; excess demand; existence of equilibrium; fixed point property; fixed point theorems; homotopy methods; Hopf's th; Kakutani's th; Kinoshita's th; K–K–M–S th; Lefschetz fixed point th; Leray–Schauder fixed point index; Nash equilibrium; perfect equilibrium; Sard's th; Scarf algorithm; Schauder fixed point th; Sperner's lemma; strategic stability; Tarski's fixed point th

Article

The Brouwer (1910) fixed point theorem and its descendants are key mathematical results underlying the foundations of economic theory.

Let $f : X \to X$ be a function from a space to itself. A *fixed point* of *f* is a point $x^* \in X$ that is mapped to itself by $f : f(x^*) = x^*$. A *fixed point th* is a result asserting that, under some hypotheses, the set of fixed points of *f* is nonempty. A simple example with many applications is:

Theorem 1: **(contraction mapping th)**. *If the metric space (X, d) is complete (recall that this means that every Cauchy sequence is convergent) and there is a number c ∈ (0,1) such that* $d(f(x), f(x')) \leq cd(x, x')$ *for all* $x, x' \in X$*, then f has a unique fixed point*.

Another example illustrating the importance of the general notion of completeness, but otherwise based on quite different principles, is:

Theorem 2: **(Tarski's (1955) fixed point theorem)**. *Let (X, ≤) be a complete lattice: ≤ is a partial ordering of X and every subset of X has a greatest lower bound and a least upper bound. If* $f : X \to X$ *is monotone – that is,* $f(x) \leq f(x')$ *whenever* $x \leq x'$ *– then there are fixed points* $\underline{u}, \overline{u} \in X$ *such that* $\underline{u} \leq x$ *whenever* $x \leq f(x)$ *and* $x \leq \overline{u}$ *whenever* $f(x) \leq x$.

This result is foundational for the theory of strategic complementarities – for example, Milgrom and Shannon (1994); Echenique (2005) – and has been applied to growth theory by Hopenhayn and Prescott (1992).

The rest of our discussion is devoted to results related to Brouwer's fixed point theorem. A topological space has the *fixed point property* if every continuous map from the space to itself has a fixed point. Brouwer's theorem states that a nonempty compact convex subset of a Euclidean space has the fixed point property. This celebrated result underlies many of the advanced results of topology, and was a pivotal event in the development of algebraic topology, which has influenced many areas of mathematics. In the half-century following Brouwer's paper the theory of fixed points was extended in various directions, yielding several generalizations of Brouwer's result that are themselves famous theorems. Early in the

post-war period fixed point theorems were used by Arrow and Debreu (1954), McKenzie (1959), Nash (1950; 1951), and Debreu (1952) to prove the fundamental equilibrium existence results of theoretical economics: every economy with finitely many goods and agents has a competitive equilibrium; every finite normal form game has a Nash equilibrium. Fixed point theory continues to play an important role in the extensive body of research that grew out of these fundamental discoveries.

Useful books devoted to fixed point theory include Border (1985), which emphasizes results used in economic theory, Brown (1971), which develops the theory of the fixed point index using the methods of algebraic topology, and Dugundji and Granas (2003), which comprehensively surveys the topic from the point of view of applications to analysis and topology. The latter book features extensive historical information concerning the development, and the developers, of the subject.

Proofs and algorithms

Since Brouwer's theorem is a breakthrough result, one should expect proofs to reveal deep mathematical principles, and in fact Brouwer's work was a major stimulus to the development of the subject that is now known as algebraic topology. Eventually Sperner (1928) distilled a relatively simple combinatoric argument out of the topological ferment of that era. Although this argument is the most popular in graduate education in economics, in the author's opinion the exposition in Milnor (1965) of an argument due to Hirsch is worth whatever additional effort it entails, because the student also learns Sard's theorem, which is another fundamental result of the 20th century with important applications in economic theory. Although the substance of the argument in Milnor (1978) appears to be less useful, its brevity and elementary character are stunning. The proof of McLennan and Tourky (2005) is also relatively simple, and displays how Kakutani's theorem follows easily from the existence of Nash equilibrium for a special class of two-person games, which is one of the simplest manifestations of the fixed point principle.

Computation of approximate fixed points has many applications in economics and other fields, and is an important topic of research. Iteration of a function is guaranteed to work only when the function is a contraction, as in Theorem 1, but this method is often practical for functions that do not satisfy this condition. Other methods are derived from proofs of Brouwer's theorem. The method pioneered by Scarf (1973; Doup, 1988) is a method of moving through the simplices of a simplicial subdivision of the simplex. It is justified by a refinement of the proof of Sperner's lemma. The proof derived from Sard's theorem points towards homotopy methods, which have a huge literature (Garcia and Zangwill, 1981; Algower and Georg, 1990). The proof in McLennan and Tourky (2005) also points towards algorithms in which the equilibria of certain two-person games give rise to approximate fixed points.

Variants

We will give statements of the main forms in which the fixed point principle is applied in economic theory. Let X and Y be metric spaces. A *correspondence* $F: X \to Y$ assigns a nonempty $F(x) \subset Y$ to each $x \in X$. When $Y = X$, a point x^* is said to be a *fixed point* if $x^* \in F(x^*)$. If P is any property of sets, then F is *P valued* if each image $F(x)$ has property P. It is *upper semicontinuous* (u.s.c.) if it is compact valued and, for each $x \in X$ and each neighborhood V of $F(x)$, there is a neighborhood U of x such that $F(x') \subset V$ for all $x' \in U$. It is not hard to show that if Y is compact, then F is u.s.c. if and only if its graph

$$Gr(F) = \{(x, y) \in X \times Y: y \in F(x)\}$$

is closed. We think of a function as a singleton-valued correspondence, in which case upper semicontinuity coincides with the usual notion of continuity.

Economic models frequently give rise to sets of optimal individual choices that are convex, but may have more than one element. For this reason the most prominent fixed point theorem in economic applications is:

Theorem 3: **(Kakutani, 1941)**. *If X is a nonempty compact convex subset of a Euclidean space and* $F: X \to X$ *is a u.s.c. convex valued correspondence, then F has a fixed point.*

The following variant is tailored for applications in general equilibrium theory, where one is searching for a price vector that equates supply and demand in all markets.

Theorem 4: **(Debreu–Gale–Kuhn–Nikaido lemma)**. *Let*

$$\Delta := \left\{ p \in \mathbb{R}^n_+ : \sum_{j=1}^{n} p_j = 1 \right\}$$

be the n–1 dimensional simplex. If $Z: \Delta \to \mathbb{R}^n$ *is a u.s.c.c.v. correspondence satisfying* $p \cdot z = 0$ *for all* $p \in \Delta$ *and all* $z \in Z(p)$*, then there is a* $p^* \in \Delta$ *and* $z^* \in Z(p^*)$ *such that* $z^* \leq 0$.

The following result of Shapley (1973a; 1973b; see also Herings, 1997, and references cited therein) generalizes the famous K–K–M theorem of Knaster, Kuratowski and Mazurkiewicz (1929). It has important applications to the theory of the core and other aspects of cooperative game theory and general equilibrium theory.

Theorem 5: **(K–K–M–S th)**. *Let* $\mathcal{N} = 2^{\{1,\ldots,n\}} \setminus \varnothing$*, and for* $S \in \mathcal{N}$ *let* $\Delta^S := \{x \in \Delta : x_i = 0 \text{ for all } i \notin S\}$*. If* $\{C^S\}_{S \in \mathcal{N}}$ *is a collection of closed sets such that* $\Delta^T \subset \cup_{S \subset T} C^S$ *for all* $T \in \mathcal{N}$*, then there is* $\mathcal{B} \subset \mathcal{N}$ *and numbers* $\lambda_S \geq 0$ *for* $S \in \mathcal{B}$ *such that* $\sum_{i \in S \in \mathcal{B}} \lambda_S = 1$ *for all* $i = 1, \ldots, n$ *(such a* $\mathcal{B}$ *is called a balanced collection) and* $\cap_{S \in \mathcal{B}} C^S \neq \varnothing$.

The original K–K–M theorem is the special case in which $C^S = \varnothing$ whenever *S* has more than one element. That is, $C_1 \cap \ldots \cap C_n \neq \varnothing$ whenever $C_1, \ldots, C_n \subset \Delta$ are closed sets satisfying $\Delta^T \subset \cup_{i \in T} C_i$ for all $T \in \mathcal{N}$.

Generalizations

During the first half of the 20th century there emerged a sequence of increasingly general versions of Brouwer's theorem. Let *X* and X' be metric spaces, and let $\Phi: X \to X'$ be a homeomorphism. A point $x^* \in X$ is a fixed point of a continuous function $f: X \to X$ if and only if $\varphi(x^*)$ is a fixed point of $\varphi \circ f \circ \varphi^{-1}$, so the fixed point property is invariant under homeomorphism. Compactness and continuity are invariant properties, but the assumptions of convexity and finite dimensionality in Brouwer's theorem seem too strong, as does the assumption of convex valuedness in Kakutani's theorem. One is led to search for weaker, topological assumptions that imply the fixed point property.

Let *Y* be another metric space. A continuous function

$$h: X \times [0, 1] \to Y$$

is called a *homotopy*. For each $0 \leq t \leq 1$ let $h_t = h(\cdot, t): X \to Y$. We think of 'continuously deforming' h_0 into h_1, with the variable *t* representing time, and we say that h_0 and h_1 are *homotopic*. The space *X* is *contractible* if the identity function on *X* is homotopic to a constant function. If *X* is convex, then for any $x_0 \in X$ the function

$$h(x, t) = x_0 + (1 - t)(x - x_0)$$

is such a homotopy, so convex sets are contractible. It was conjectured that nonempty compact contractible sets have the fixed point property, but eventually counterexs were discovered by Kinoshita (1953) and others.

A *retraction* of *X* onto a subset *A* is a continuous function $r: X \to A$ whose set of fixed points is *A*, so that $r(a) = a$ for all $a \in A$. In this circumstance we say that *A* is a *retract* of *X*. One point of interest is that if *X* has the fixed point property, then so does *A*: if $g: A \to A$ is continuous, then $g \circ r: X \to A \subset X$ has a fixed point x^*, and $x^* = g(r(x^*)) = g(x^*)$ because x^* must be in *A*.

The subspace *A* is a *neighbourhood retract* if there is an open $U \supset A$ and a retraction $r: U \to A$. A continuous function $e: X \to Y$ is an *embedding* if it is injective and $e^{-1}: e(X) \to X$ is continuous, that is, *e* is a homeomorphism onto its image. A metric space *X* is an *absolute neighbourhood retract* (ANR) if *e*(*X*) is a neighbourhood retract whenever $e: X \to Y$ is an embedding of *X* in a metric space *Y*. The class of ANRs is large, encompassing many important types of spaces such as manifolds, simplicial complexes, and convex sets, and there is an extensive theory (for example, Borsuk, 1967) that cannot be described here. One may think of an ANR as a space that has bounded complexity, in a certain sense, in a

neighbourhood of each of its points. (An example of a space that is *not* an ANR is the union X of the unit circle centred at the origin in $\mathbb{R}^2$ and the set $\{(1-\theta^{-1})(\cos\theta, \sin\theta : 1 \leq \theta < \infty\}$. If X was an ANR, then there would exist a retraction of a neighbourhood $U \subset \mathbb{R}^2$ onto X, and the retraction would take small connected neighbourhoods of (1, 0) in U to small connected neighbourhoods of (1,0) in X, but small neighbourhoods of (1,0) in X are disconnected.)

Eilenberg and Montgomery (1946) gave a fully satisfactory generalization of Brouwer's theorem: F has a fixed point whenever X is a nonempty compact acyclic ANR and $F: X \to X$ is a u.s.c. acyclic valued correspondence. Acyclicity is a concept from algebraic topology that cannot be defined here; the important point for us is that contractible sets are acyclic, and that the loss of generality in passing from acyclicity to contractibility is of slight concern in economic theory.

Contractible valued correspondences that are not convex valued appear in McLennan (1989a) and Reny (2005). There are many applications in economics of the special case of the Eilenberg–Montgomery theorem in which X is convex (but possibly infinite dimensional) and F is convex valued, for which relatively simple and direct proofs were given by Fan (1952) and Glicksberg (1952). In turn this result is more general than both Kakutani's theorem and the well known Schauder (1930) fixed point theorem.

The Leray – Schauder fixed point index

Consider the fixed points of the function from [0,1] to itself shown in Figure 1. The points A and C are qualitatively similar, and qualitatively different from B. In the one-dimensional setting one can easily see that, if the function is differentiable and its graph is not tangent to the diagonal at any of its fixed points, then the number of fixed points of the first type must be one greater than the number of fixed points of the second type. In particular, the number of fixed points must be odd. These properties extend to smooth functions $f: C \to C$, where C is an n-dimensional convex set, that intersect the diagonal in the 'expected' manner: the Jacobian of $\mathrm{Id}_C - f$ is nonsingular. Debreu (1970) used Sard's theorem (for example, Milnor, 1965) to show that for an exchange economy with fixed preferences, the excess demand function generated by a 'generic' endowment vector has well-behaved equilibria, and Dierker (1972) showed that the qualitative conclusions described above hold in this circumstance. Mas-Colell (1985) summarizes the extensive literature descended from these seminal contributions.

Figure 1

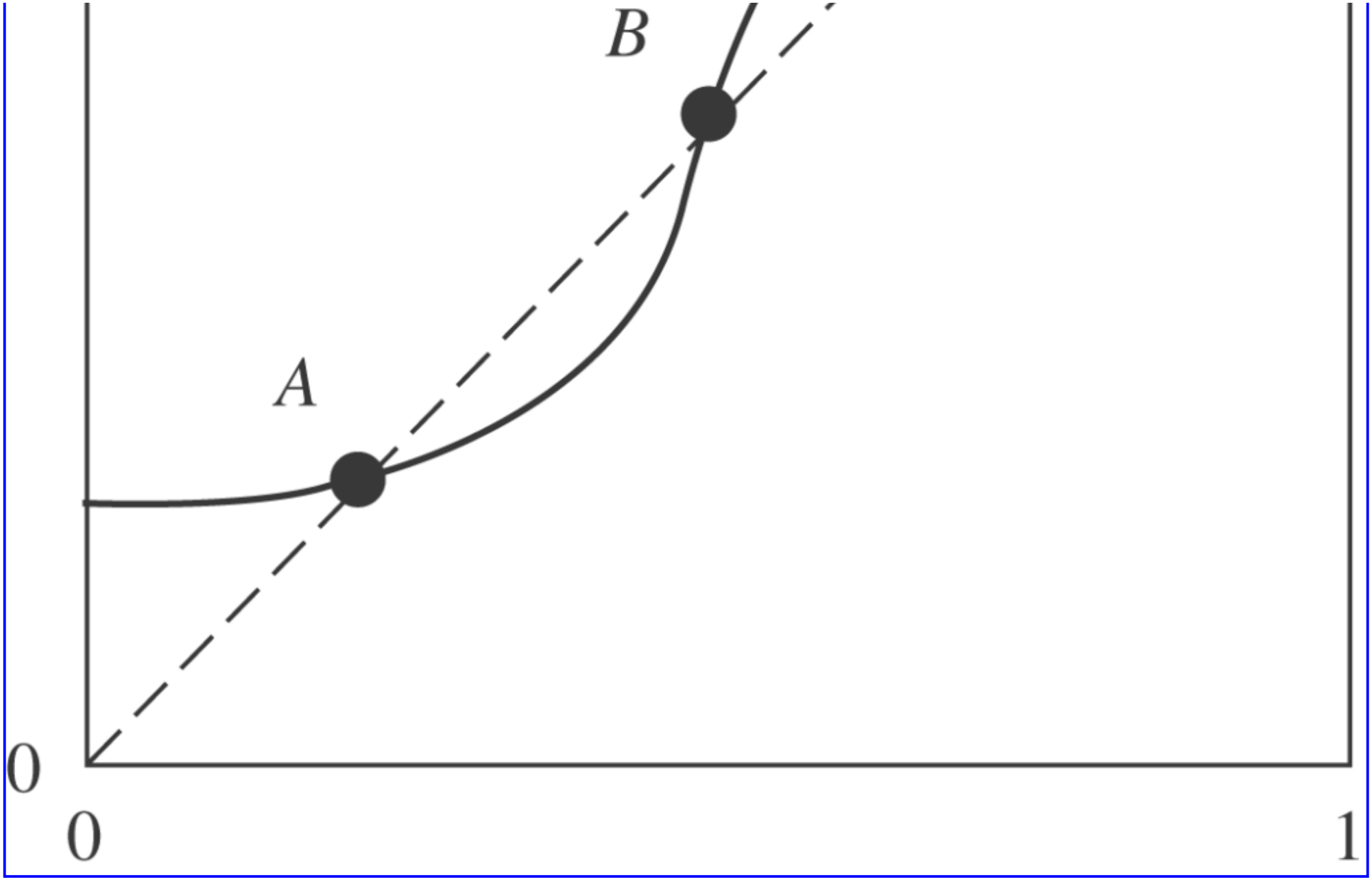

The *Leray–Schauder fixed point index* generalizes these aspects of the theory to correspondences, to sets of fixed points that are not singletons, and to general ANRs. Suppose X is a nonempty compact ANR, $U \subset X$ is open and $\bar{U}$ is its closure. A correspondence $F: \bar{U} \to X$ is *index admissible* if it is u.s.c. and does not have any fixed points in its boundary $\bar{U} \setminus U$. Let $\mathcal{F}_X$ be the set of index admissible contractible valued correspondences $F: \bar{U} \to X$ where $U \subset X$ is open. A homotopy $h: \bar{U} \times [0, 1] \to X$ is *index admissible* if each h_t is index admissible.

The next result gives an axiomatic characterization of a number $\Lambda_X(F)$. When there are finitely many fixed points the Additivity axiom allows us to think of $\Lambda_X(F)$ as the sum of their indices. When $X \subset \mathbb{R}^n$, $f: \bar{U} \to X$ is a smooth function, and x is a fixed point in the interior of X with $\mathrm{Id}_{\mathbb{R}^n} - Df(x)$ nonsingular, the index of x is +1 or –1 according to whether the determinant of $\mathrm{Id}_{\mathbb{R}^n} - Df(x)$ is positive or negative.

Theorem 6: There is a unique function $\Lambda_X: \mathcal{F}_X \to \mathbb{Z}$ *satisfying*:

1. (A) *(Normalization) If* $c: X \to X$ *is a constant function, then* $\Lambda_X(c) = 1$.
2. (B) *(Additivity) If* $F: \bar{U} \to X$ *is in* $\mathcal{F}_X$, $U_1, \ldots, U_r$ *are disjoint open subsets of U, and F has no fixed points in* $\bar{U} \setminus (U_1 \cup \ldots \cup U_r)$, *then*

$$\Lambda_X(F) = \sum_{i=1}^{r} \Lambda_X(F|_{\bar{U}_i}).$$

3. (C) *(Homotopy) If* $h: \bar{U} \times [0, 1] \to X$ *is an index admissible homotopy, then*

$$\Lambda_X(h_0) = \Lambda_X(h_1).$$

4. (D) *(Continuity) For each* $F: \bar{U} \to X$ *in* $\mathcal{F}_X$ *there is a neighborhood* $W \subset \bar{U} \times X$ *of Gr(F) such that* $\Lambda_X(F') = \Lambda_X(F)$ *for all* $F': \bar{U} \to X$ *with* $F' \in \mathcal{F}_X$ *and*

$$Gr(F') \subset W.$$

The index is closed related to the Brouwer degree of a function between manifolds of the same dimension. These ideas evolved from the time of Brouwer's work until O'Neill (1953) achieved the axiomatic expression of the concept (for functions) given above.
Theorem 1 has many important consequences. To begin with note that if $F \in \mathcal{F}_X$ has no fixed points, then Additivity implies that

$$\Lambda_X(F) = \Lambda_X(F|_{\varnothing}) = \Lambda_X(F|_{\varnothing}) + \Lambda_X(F|_{\varnothing}) = 0.$$

Therefore *F* must have a fixed point whenever $\Lambda_X(F) \neq 0$. If $f: X \to X$ is a continuous function, then $\Lambda_X(f\bullet)$ is called the *Lefschetz* number of *f*. The famous Lefschetz (1923) fixed point theorem states that *f* has a fixed point if its Lefschetz number is nonzero, and provides connections to algebraic topology that give tools for computing the Lefschetz number.
We now use the following approximation result to recover the weak version of the Eilenberg–Montgomery theorem stated above, thereby showing that Theorem 6 embodies the fixed point principle. This result generalizes Kakutani's method of passing from Brouwer's theorem to his result, and it plays an important role in one method of proving Theorem 6.
Theorem 7: **(Mas-Colell, 1974**; **McLennan, 1989b)**. *If X is a compact ANR, U, V* $\subset$ *X are open with* $\bar{V} \subset U$, $F: \bar{U} \to X$ *is a u.s.c. contractible valued correspondence, and* $W \subset \bar{U} \times X$ *is a neighbourhood of Gr(F), then there is a continuous function* $f: \bar{V} \to X$ *with Gr(f)*$\subset$*W*.
Suppose that $F: X \to X$ is a u.s.c. contractible valued correspondence. Applying the last result with $U = V = X$ and *W* as in (14), we find that there is a continuous function $f: X \to X$ with $\Lambda_X(f) = \Lambda_X(F)$. If *X* is contractible, so that there is a homotopy $h: X \times [0, 1] \to X$ with $h_0 = Id_X$ and h_1 a constant function, then $j(x, t) = f(h(x, t))$ is a homotopy with

$j_0 = f$ and j_1 a constant function, so Homotopy and Normalization imply that $\Lambda_X(f) = 1$. We conclude that $\Lambda_X(F) = 1$, and that F necessarily has a fixed point.

Recall that a subset C of a metric space Y is *connected* if there do not exist open sets $V_1, V_2 \subset Y$ with $V_1 \cap V_2 \neq \varnothing$ and $V_1 \cap C \neq \varnothing \neq V_2 \cap C$. A subset of Y is a *connected component* if it is the union of all connected sets containing some point y. Each connected component is connected, and the connected components partition Y.

Suppose that X is a compact contractible ANR, that $F: X \to X$ is in $\mathcal{F}_X$, and that the set of fixed points of F has finitely many connected components $C_1,\ldots,C_r$. Additivity implies that each component C_i has a well-defined index λ_i that depends on the restriction of F to an arbitrarily small neighbourhood of C_i. Suppose that it is possible to show that $\lambda_i = 1$ for each i. Since additivity implies that $\Sigma_i \lambda_i = \Lambda_X(F) = 1$, it follows that $r = 1$. This style of proof of uniqueness is applicable to many economic settings, but usually more elementary methods are available. At present no alternative to its application in Eraslan and McLennan (2005) is known. It is more common to use the index to prove nonuniqueness: it suffices to display a connected component whose index is different from one.

The fixed point index has two other important properties.

Theorem 8: **(Multiplication)**. *If X and Y and compact ANRs, $U \subset X$ and $V \subset Y$ are open, $F: \bar{U} \to X$ and $G: \bar{V} \to Y$ are index admissible contractible valued correspondences, and $F \times G: \bar{U} \times \bar{V} \to X \times Y$ is the correspondence that takes (x,y) to $F(x) \times G(y)$, then*

$$\Lambda_{X \times Y}(F \times G) = \Lambda_X(F) \cdot \Lambda_Y(G).$$

Theorem 9: **(Commutativity)**. *If X and Y are compact ANRs and $f: X \to Y$ and $g: Y \to X$ are continuous functions, then*

$$\Lambda_X(g \circ f) = \Lambda_Y(f \circ g).$$

There is a more general version of Commutativity for functions defined on subsets of X and Y, but its statement involves technical complications. In view of the uniqueness asserted in Theorem 6, Multiplication and Commutativity are, in principle, consequences of (A)–(D), but it is not known how to prove them in this way. In practice these properties are treated as axioms and shepherded up the ladder of generality, one rung at a time, along with everything else. In fact Commutativity (which was introduced by Browder, 1948, for this purpose) plays a critical role at one stage of this process.

Essential sets of fixed points

The two fixed points in Figure 2 are qualitatively different. Arbitrarily small perturbations of the function have no fixed point near A, but this is not the case for B. In the terminology introduced by Fort (1950) A is *inessential* while B is *essential*. Let X be a compact contractible ANR, let $F: X \to X$ be a u.s.c. contractible valued correspondence, and let C be the set of fixed points of F. Kinoshita (1952) extended Fort's ideas to correspondences, and to sets of fixed points, defining an *essential set of fixed points* of F to be a compact $C' \subset C$ such that for any neighbourhood U of C' there is a neighbourhood W of Gr(F) such that any continuous function $f: X \to X$ with $Gr(f) \subset W$ has a fixed point in U.

Figure 2

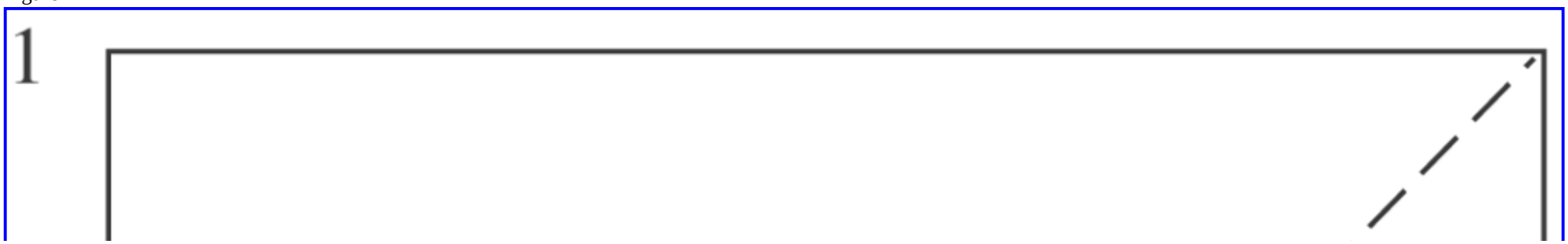

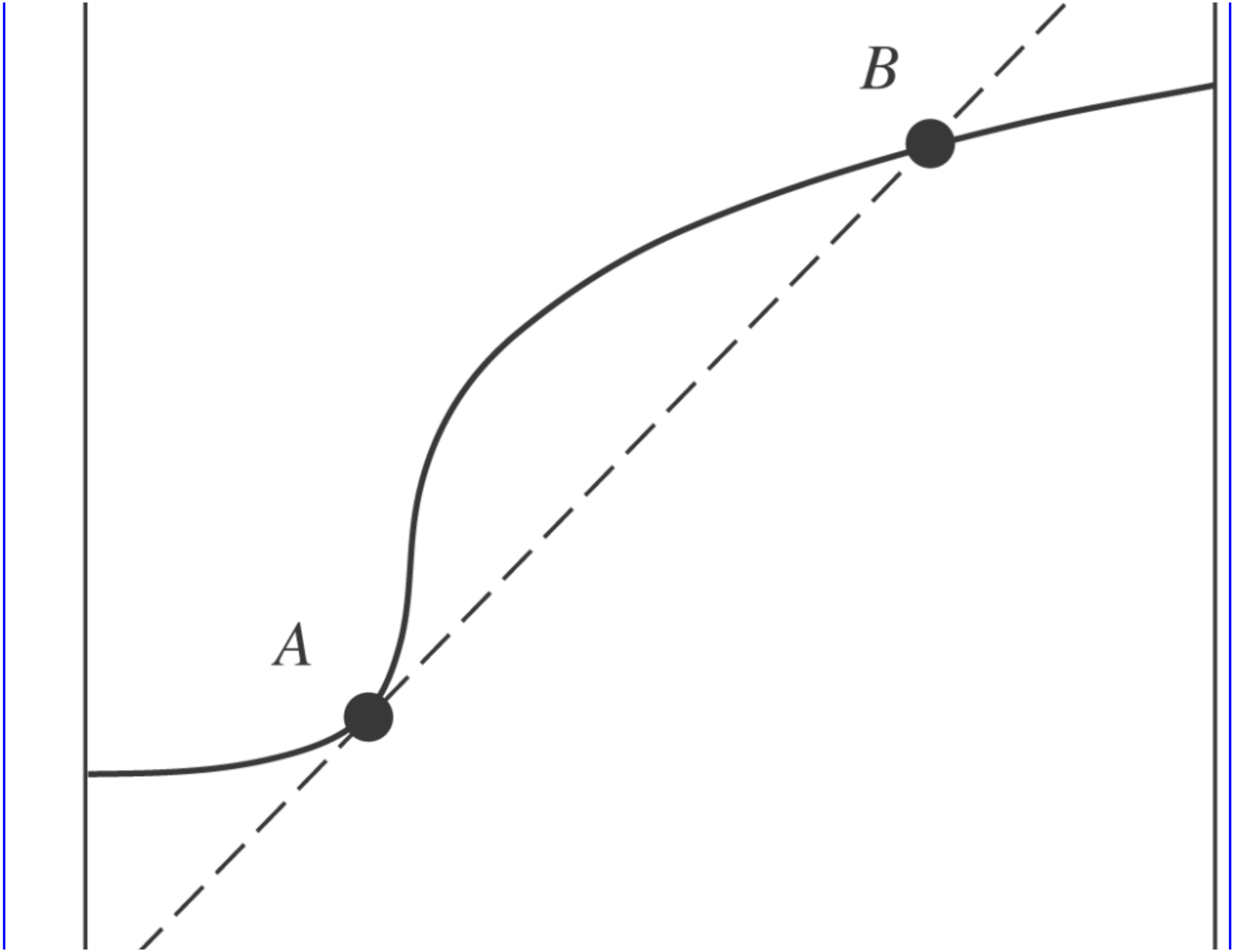
B
A

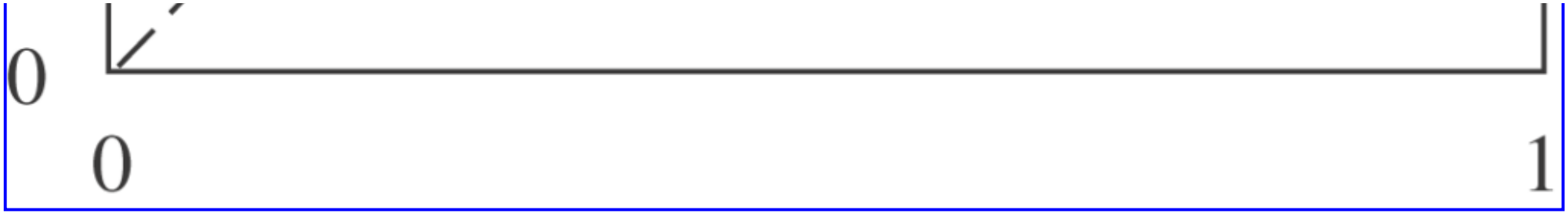

For any neighbourhood U of C we can find a neighbourhood W of Gr(F) that cannot have any fixed points outside of U, so C is essential. That is, without some additional condition, essentiality does not distinguish some fixed points from others. Following Kohlberg and Mertens (1986), one is led to consider minimal essential sets, which exist by virtue of the following argument. Let $B_1, B_2,\ldots$ be a listing of the open balls of rational radii centred at points in some countable dense subset of X. Define a sequence $K_0, K_1, K_2,\ldots$ inductively by setting $K_0 = C$ and, for $j \geq 1$, setting $K_j = K_{j-1} \backslash B_j$ if this set is essential and otherwise setting $K_j = K_{j-1}$. We claim that $K_\infty = \cap_j K_j$ is a minimal essential set. Any neighbourhood U of K_∞ contains some K_j (the accumulation points of a sequence $\{x_j\}$ with $x_j \in K_j \backslash U$ must be outside U but also in each K_j, by compactness, hence in K_∞) and each K_j is essential, so K_∞ is essential. If there was a smaller essential set there would be some j such that $K_\infty \backslash B_j \neq K_\infty$ was essential, but then $K_{j-1} \backslash B_j$ would also be essential, in which case $K_\infty \cap B_j \subset K_j \cap B_j = \varnothing$.

Kinoshita (1952) showed that minimal essential sets are connected when X is convex and F is convex valued. Otherwise one could find a minimal essential set $C_1 \cup C_2$, where C_1 and C_2 are nonempty, compact, and disjoint. Then C_1 and C_2 are inessential, so there is a perturbation of F that has no fixed points near C_1 and another such perturbation of F has no fixed points near C_2. The main idea of Kinoshita's argument is that these can be combined, by using convex combination with locally varying weights, to give a perturbation of F that has no fixed point near $C_1 \cup C_2$, thereby contradicting the assumption that $C_1 \cup C_2$ is essential.

Kinoshita's theorem is pertinent to the literature on refinements of Nash equilibrium that began with the introduction in Selten (1975) of perfect equilibrium. An important technique is to give a privileged status to those Nash equilibria that can be approximated by fixed points of certain perturbations of the given correspondence. In particular, it has important connections to the notion of strategic stability of Kohlberg and Mertens (1986).

The fixed point index also has implications for essential sets. For the sake of simplicity assume that C consists of finitely many connected components $C_1,\ldots,C_r$. (This condition holds in the application to Nash equilibrium.) Any C_i with nonzero index is essential, by Continuity. Since the sum of the indices is one, some C_i must have nonzero index, so a connected essential set exists. Harder arguments, which apply the Hopf theorem (Milnor, 1965) to ‘transport’ fixed points of perturbations to a desired location, and to eliminate pairs of fixed points of opposite index, show that any proper subset of a C_i is inessential, and that C_i is inessential if its index is zero. Thus the minimal essential sets are precisely those C_i with nonzero index.

See Also

- computation of general equilibria
- computation of general equilibria (new developments)
- existence of general equilibrium
- game theory
- mathematical economics
- mathematics and economics
- Nash equilibrium, refinements of
- non-cooperative games (equilibrium existence)

Bibliography

Algower, E. and Georg, K. 1990. *Numerical Continuation Methods*. New York: Springer Verlag.

Arrow, K. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Border, K. 1985. *Fixed Point Theorems with Applications to Economics and Game Theory*. Cambridge: Cambridge University Press.

Borsuk, K. 1967. *Theory of Retracts*. Warsaw: Polish Scientific Publishers.

Browder, F. 1948. The topological fixed point theory and its applications to functional analysis. Ph.D. thesis, Princeton University.

Brouwer, L. 1910. Uber Abbildung von Mannigfaltikeiten. *Mathematische Annalen* 71, 97–115.

Brown, R. 1971. *The Lefschetz Fixed Point Theorem*. Glenview, IL: Scott Foresman and Co.

Debreu, G. 1952. A social equilibrium existence th. *Proceedings of the National Academy of Science* 38, 886–93.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Dierker, E. 1972. Two remarks on the number of equilibria of an economy. *Econometrica* 40, 951–3.

Doup, T. 1988. *Simplicial Algorithms on the Simplotope*. Berlin: Springer-Verlag.

Dugundji, J. and Granas, A. 2003. *Fixed Point Theory*. New York: Springer-Verlag.

Echenique, F. 2005. A short and constructive proof of Tarski's fixed point th. *International Journal of Game Theory* 33, 215–18.

Eilenberg, S. and Montgomery, D. 1946. Fixed-point theorems for multivalued transformations. *American Journal of Mathematics* 68, 214–22.

Eraslan, H. and McLennan, A. 2005. Uniqueness of stationary equilibrium payoffs in coalitional bargaining. Mimeo, University of Pennsylvania.

Fan, K. 1952. Fixed point and minimax theorems in locally convex linear spaces. *Proceedings of the National Academy of Sciences* 38, 121–6.

Fort, M. 1950. Essential and nonessential fixed points. *American Journal of Mathematics* 72, 315–22.

Garcia, C. and Zangwill, W. 1981. *Pathways to Solutions, Fixed Points, and Equilibria*. Englewood Cliffs, NJ: Prentice-Hall.

Glicksberg, I. 1952. A further generalization of the Kakutani fixed point theorem with applications to Nash equilibrium. *Proceedings of the American Mathematical Society* 3, 170–4.

Herings, P. 1997. An extremely simple proof of the K–K–M–S th. *Economic Theory* 10, 361–7.

Hopenhayn, H. and Prescott, E. 1992. Stochastic monotonicity and stationary distributions for dynamic economies. *Econometrica* 60, 1387–406.

Kakutani, S. 1941. A generalization of Brouwer's fixed point th. *Duke Mathematical Journal* 8, 416–27.

Kinoshita, S. 1952. On essential components of the set of fixed points. *Osaka Mathematical Journal* 4, 19–22.

Kinoshita, S. 1953. On some contractible continua without the fixed point property. *Fundamentae Mathematicae* 40, 96–8.

Knaster, B., Kuratowski, C. and Mazurkiewicz, C. 1929. Ein Beweis des Fixpunktsatzes fur n-dimensionale Simplexe. *Fundamenta Mathematicae* 14, 132–7.

Kohlberg, E. and Mertens, J.-F. 1986. On the strategic stability of equilibria. *Econometrica* 54, 1003–38.

Lefschetz, S. 1923. Continuous transformations of manifolds. *Proceedings of the National Academy of Sciences* 9, 90–3.

Mas-Colell, A. 1974. A note on a theorem of F. Browder. *Mathematical Programming* 6, 229–33.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: A Differentiable Approach*. Cambridge: Cambridge University Press.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27, 54–71.

McLennan, A. 1989a. Consistent conditional systems in noncooperative game theory. *International Journal of Game Theory* 18, 141–74.

McLennan, A. 1989b. Fixed points of contractible valued correspondences. *International Journal of Game Theory* 18, 175–84.

McLennan, A. and Tourky, R. 2005. From imitation games to Kakutani. Mimeo, University of Minnesota.

Milgrom, P. and Shannon, C. 1994. Monotone comparative statics. *Econometrica* 62, 157–80.

Milnor, J. 1965. *Topology from the Differentiable Viewpoint*. Charlottesville: University Press of Virginia.

Milnor, J. 1978. Analytic proofs of the ‘hairy ball th’ and the Brouwer fixed-point th. *American Mathematical Monthly* 85, 521–4.

Nash, J. 1950. Non-cooperative games. Ph.D. thesis, Department of Mathematics, Princeton University.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

O’Neill, B. 1953. Essential sets and fixed points. *American Journal of Mathematics* 75, 497–509.

Reny, P. 2005. On the existence of monotone pure strategy equilibria in Bayesian games. Mimeo, University of Chicago.

Scarf, H. 1973. *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Schauder, J. 1930. Der Fixpunktsatz in Funktionalraumen. *Studia Mathematica* 2, 171–80.

Selten, R. 1975. Re-examination of the perfectness concept for equilibrium points of extensive games. *International Journal of Game Theory* 4, 25–55.

Shapley, L. 1973a. On balanced games without side payments. In *Mathematical Programming Study*, ed. T. Hu and S. Robinson. New York: Academic Press.

Shapley, L. 1973b. On balanced games without side payments: a correction. Rand Paper Series Report No. p-4910/1. Santa Monica, CA: RAND Corporation.

Sperner, E. 1928. Neuer Beweis fiir die Invarianz der Dimensionszahl und des Gebietes. *Abhandlungen aus dem Mathematischen Seminar der Hamburgischen* 6, 265–72.

Tarski, A. 1955. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics* 5, 285–309.

How to cite this article

McLennan, Andrew. "fixed point theorems." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000135> doi:10.1057/9780230226203.0586

The New Palgrave Dictionary of Economics Online

fixprice models

Joaquim Silvestre
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The general competitive theory of markets (Walras, Arrow–Debreu) presupposes that no agent has market power and that prices and wages instantaneously adjust to equilibrate price-taking supply and demand. Fixprice models follow its emphasis on the interactions across markets, but under the more realistic assumption that markets frequently operate under excess demand or supply, with prices often exceeding marginal costs because prices and wages adjust slowly, or because of market power. The original fixprice models, which adopted the short-run method with static expectations, are the precursors of neo-Keynesian dynamic macroeconomics based on market power and the stickiness of wages or prices.

Keywords

bargaining; comparative statics; dual decision hypothesis; dynamic stochastic macroeconomic models; employment; excess demand and supply; fixprice models; full employment; general equilibrium; imperfect competition; inflation; market power; microfoundations; monopolistic competition; oligopoly; oligopsony; Patinkin, D.; price control; real business cycles; rent control; second best; staggered prices; staggered wages; sticky prices; sticky wages; unemployment; voluntariness; wage control; Walrasian equilibrium; Walras–Samuelson tâtonnement

Article

The canonical fixprice model (Benassy, 1975; 1976; 1982; Drèze, 1975; Younès, 1975) was born at the interface of two extensions of general equilibrium theory: the study of out-equilibrium price dynamics, and the incorporation of price-setting behaviour by firms. Fixprice analysis aimed at providing microfoundations for macroeconomic theory and policy. Accordingly, it first generated static macroeconomic models of the interaction between the labour and the output markets at given prices and wages with or without explicit market power. Later, it exerted a diffuse influence on the more recent dynamic macroeconomic models with market power and/or wage or price stickiness.

Many wages and prices appear to change infrequently and fail to respond quickly to shocks. Casual observation suggests that the wages of many workers are fixed in nominal terms for at least several months, and do not drop quickly in response to adverse shocks in demand. This observation is well supported by quantitative studies (Taylor, 1999) as well as by the in-depth interviews of Bewley (1999). For instance, Cecchetti (1984) found that, even when the rate of inflation was high, union wages were fixed at nominal levels for an average of one year. Later researchers, such as Card and Hyslop (1997), have obtained similar results for non-union workers. Bewley (1999) finds that wage rigidity is stricter in long-term, full-time jobs (the ‘primary sector’) than in short-term, part-time ones, and emphasizes downwards rigidity over upwards rigidity, an asymmetry that is qsted by Taylor (1999).

Price rigidity has also been subject to extensive inquiry (Andersen, 1994; Taylor, 1999). Even though one may presume that prices are less rigid than wages, and many commodities are indeed sold at continuously changing prices, several studies have found that, on the average, prices may stay fixed for relatively long periods (Carlton, 1986; 1989; Cecchetti, 1986; Blinder et al., 1998). In Taylor's (1999, p. 1020) words, ‘… the studies suggest that *price changes and wage changes have about the same average frequency – about one year*’ (emphasis in original.) Later work has found shorter, but still ample, average periods (Baharad and Eden, 2004).

In addition, prices seem to systematically exceed marginal costs in many industrial markets. These observations challenge the relevance of models where wages and prices are assumed to adjust instantaneously to their Walrasian equilibrium values.

Theoretical roots of the canonical fixprice model

Out-of-equilibrium price dynamics

The Walrasian approach postulates that prices adjust very rapidly in response to excess demand or supply, so that no transactions occur before equilibrium is reached. A rigorous formulation of this idea is the Walras–Samuelson tâtonnement process. Consider an exchange economy with two commodities and two traders, Trader *i* being initially endowed with ω_{ij} units of commodity *j* (*i*, $j = 1, 2$). Let the aggregate Walrasian excess demand functions be $z^j(p_1, p_2|\omega)$, $j = 1, 2$. (Here the vector $\omega = (\omega_{ij})$ is fixed.) As long as $z^1(p_1, p_2|\omega) \neq 0$ or $z^2(p_1, p_2|\omega) \neq 0$, no transactions occur and prices adjust according to the differential equation:

$$\frac{dp_j}{dt} = z^j(p_1(t), p_2(t)|\omega), \; j = 1, 2.$$

Walrasian excess demands provide the ‘market signals’ for the adjustment of prices in the Walras–Samuelson tâtonnement. Of course, the Walrasian excess demand functions express plans made under the conjecture that any quantities can be bought and sold at the going prices. If transactions did occur at non-Walrasian prices, then such a conjecture would be falsified, since some agents would be unable to realize their plans (see Arrow, 1959). This led Patinkin to postulate that disequilibrium transactions in a market create spillover effects on others, so that, for example, ‘the pressure of excess demand in the one market affects the price movements in all other markets’ (1956, p. 157). Patinkin's formulation was imprecise (Negishi, 1965; Clower, 1965), but his search for the ‘relevant market signals’ motivates Clower's (1965) ‘dual decision hypothesis’. This idea, generalized by Barro and Grossman (1971; 1976), is central to Benassy's fixprice model.

It was discovered in the late 1950s that the Walras–Samuelson tâtonnement process fails to converge unless some restrictive assumptions are imposed, motivating the non-tâtonnement adjustment process. Here two simultaneous movements occur: the distribution of the endowments changes according to some rule for trading at non-Walrasian prices, and prices adjust in response to Walrasian excess demands at the current endowments, for example, for some rule g_{ij},

$$\frac{d\omega_{ij}}{dt} = g_{ij}(p_1(t), p_2(t)|\omega(t)),$$

$$\frac{dp_j}{dt} = z^j(p_1(t), p_2(t)|\omega(t)), \; i, \; j = 1, 2.$$

This process is hard to interpret except possibly as depicting the sequential exchange of durable goods, and, as just argued, the appeal to Walrasian excess demand in the price-adjustment equation is unjustified. But some conditions on the functions g_{ij} (Hahn and Negishi, 1962; Uzawa, 1962) originally meant to guarantee the convergence of the non-tâtonnement process inspired the fixprice model of Younès (1975).

General equilibrium with market power

The monopolistic general equilibrium analysis pioneered by Negishi (1960) led to the construction of simple models where firms or workers had price- or wage-setting capacity (Benassy 1976; 1977; 1982; 1991; Hart, 1982; Silvestre, 1990; 1993). This work revealed intimate connections between market power and the fixity or stickiness of prices and wages. First, oligopoly displays formal parallelisms to excess supply, and oligopsony to excess demand (Madden and Silvestre, 1991; 1992; Silvestre, 1986). Second, an imbalance between supply and demand gives temporary market power to agents on the short side who then face non-horizontal demand or supply curves for large enough quantities (Arrow, 1959; Negishi, 1974; 1979; Hahn, 1978; John, 1985). Third, a firm with market power experiencing frequent demand or productivity shocks may optimize by changing prices at discrete intervals even if the costs of changing prices are small (Sheshinski and Weiss, 1977; Akerlof and Yellen, 1985; Mankiw, 1985; Parkin, 1986; Caplin and Spulber, 1987; Blanchard and Kiyotaki, 1987), and the resulting stickiness is magnified by strategic complementarities among the pricing decisions of firms (Fishman and Simhon, 2005).

Fixprice allocations

Trading at non-Walrasian prices

Fixprice analysis postulates a common medium of exchange (money) in each market. Thus, there are $n+1$ goods (from 0 to *n*) in the case of *n* markets, the zero good being money. The analysis addresses two qsts. First, given a price vector *p* (normalized with respect to money), what allocations are compatible with it? Second, given a *p* and an allocation compatible with it, which is the type of disequilibrium in each market? The answers are derived from three basic principles: (*a*) voluntary trading; (*b*) absence of market frictions, and (*c*) effective demand. The last requires the explicit recognition of the interaction among markets. The first two impose conditions on the trades carried out in a market, namely, that, at the going price, (*a*) no trader may gain by trading less; (*b*) no pair formed by a buyer and a seller may gain by trading more.

The fixprice model provides a general framework (which includes Walrasian markets as a limit) for price-guided allocation mechanisms. It has several applications: (*a*) *short-run analysis*, which assumes that it takes time for prices and quantities to adjust; (*b*) *market power* (imperfect or monopolistic competition); (*c*) *price* (*wage or rent*) *controls*; this in particular motivates Drèze's formulation; and (*d*) *price* (*or wage*) *negotiation* (representatives of buyers and sellers negotiate prices that are taken as given by individual traders: see Silvestre, 1988). Fixprice analysis can be viewed as abstracting from specific features and focusing instead on basic market principles common to alternative specifications.

The definitions of fixprice equilibrium due to Bénassy, Drèze and Younès vary in form and motivation, but turn out to be equivalent under some assumptions (Silvestre, 1982; 1983). Rather than reproducing them in all generality, we exemplify the common concepts in two simple but important cases.

Differentiable exchange economies

There are $n+1$ goods, indexed 0, 1,…, *n* (i.e. *n* markets). There are *m* traders: trader *i* is endowed with an ($n+1$) dimensional vector ω_i of initial endowments and a differentiable utility function $u_i(x_{i0}, x_{i1},\ldots, x_{in})$. A net trade allocation is an *m*-tuple of *n*-dimensional net trade vectors $(z_i) = ((z_{i1}, \ldots, z_{in}))$, one for each trader, satisfying: $\Sigma_i z_i = 0$. It is understood that, for $j = 1, \ldots, n$, if $z_{ij} > 0$ (or <0) then trader *i* is buying (or selling) in market *j*. The (normalized) price vector $p = (p_1, \ldots, p_n)$ is given. The vector $\hat{x}_i(p, z_i) \equiv (\omega_{i0} - p \bullet z_i, \omega_{i1} + z_{i1}, \ldots, \omega_{in} + z_{in})$ is then the consumption vector associated with (p, z_i). Define *i*'s *marginal utility of trading in market j at the going price* as: $\mu_{ij}(p, z_i) \equiv \partial u_i / \partial x_{ij} - p_j \partial u_i / \partial x_{i0}$, with derivatives evaluated at $\hat{x}_i(p, z_i)$.

Definition 1: A net trade allocation (z_i^*) is a *fixprice equilibrium for p* if, writing $\mu_{ij}^* \equiv \mu_{ij}(p, z_i^*)$:

1. (a) *Voluntariness*: For $i = 1, \ldots, m$, $z_{ij}^* \cdot \mu_{ij}^* \geq 0$;
2. (b) *Absence of market frictions*: For $j = 1, \ldots, n$ and for any pair of consumers *i*, *h*, $\mu_{ij}^* \cdot \mu_{hj}^* > 0$.

Figure 1 illustrates the case of $n = 1$ and $m = 2$ in an Edgeworth box: point *A* represents the (unique) fixprice equilibrium at the price vector *p*: there Trader 1 is a buyer ($z_{11} > 0$). The straight line through points ω and *A* depicts the budget constraints. Allocations in the segment [ω, *A*) violate condition (b). Those in the segment [*A*,*B*) (in particular the Pareto efficient point *D*) violate condition (a) for Trader 1.

Figure 1

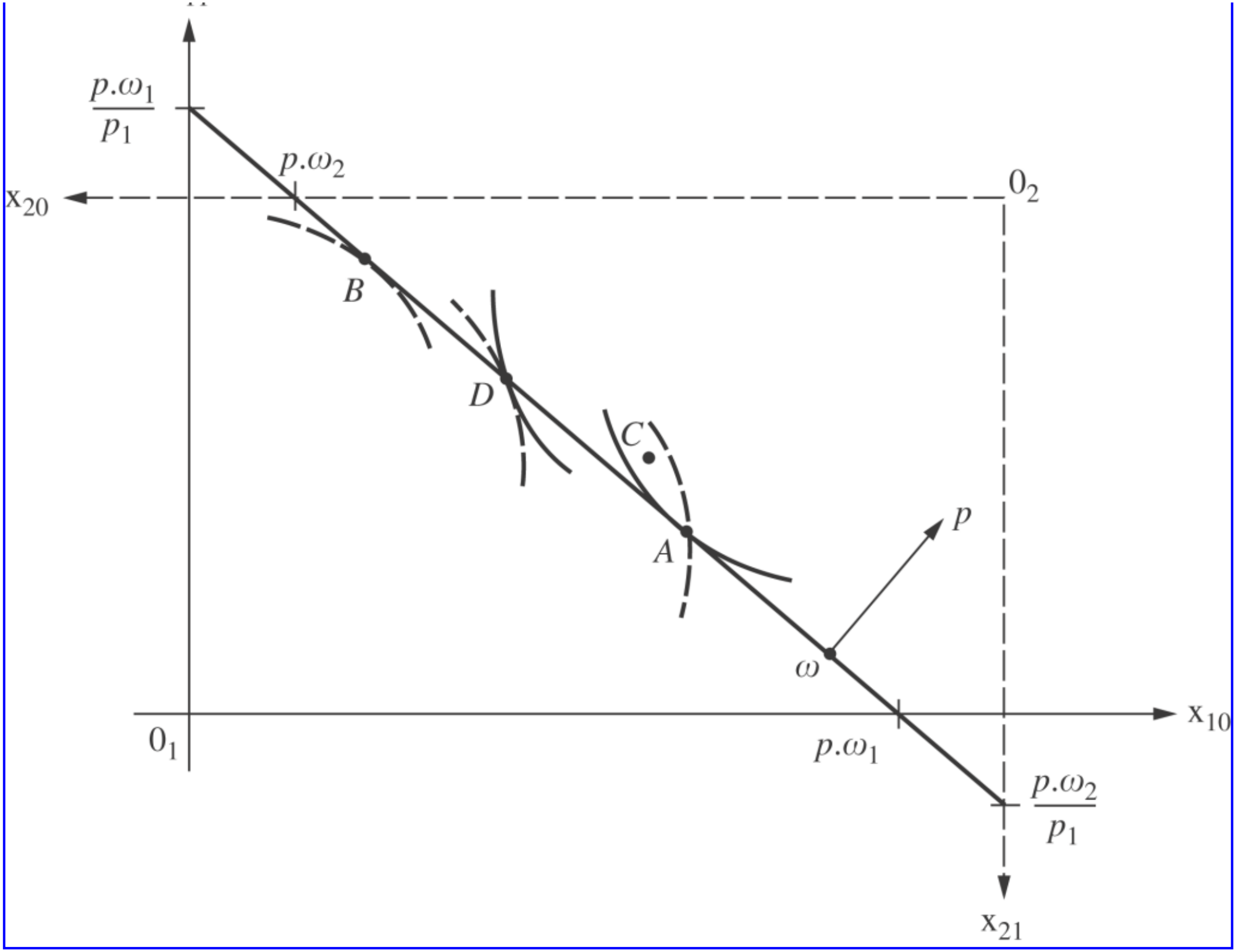
$\frac{p.\omega_1}{p_1}$
$p.\omega_2$
x_{20}
0_2
B
D
C
A
p
ω
x_{10}
0_1
$p.\omega_1$
$\frac{p.\omega_2}{p_1}$
x_{21}

The more complex case of $n = m = 2$ is illustrated in Figures 2a–d. Figure 2(a) depicts Trader 1's budget set in $\Re^3_+$. Rather than drawing a three-dimensional Edgeworth box, we graph first separately (Figure 2(b–c)) and then together (Figure 2(d)) the two-dimensional budget triangles of the traders. Figure 2(a–b) also depicts the intersections of some indifference surfaces of Trader 1 with her budget set, Q_1 being her most preferred point in the budget set. At point A she is selling in both markets (that is, $z_{11} < 0$ and $z_{12} < 0$: she gets money in exchange), with $\mu_{12} < 0$ (she would like to sell more in market 2) and $\mu_{11} = 0$. Figure 2(c) corresponds to Trader 2: at point A she is buying in both markets (i.e., $z_{21} > 0$ and $z_{22} > 0$), with $\mu_{21} > 0$ and $\mu_{22} = 0$. Figure 2(d) superimposes the two graphs (with the axes corresponding to Trader 2 reversed, and with the initial endowment points coinciding at ω). Points A in Figures 2(b–c) have been chosen so that they also coincide in Figure 2(d), i.e., $z_{2j} + z_{2j} = 0$, $j = 1, 2$. These trades constitute a fixprice equilibrium.

Figure 2

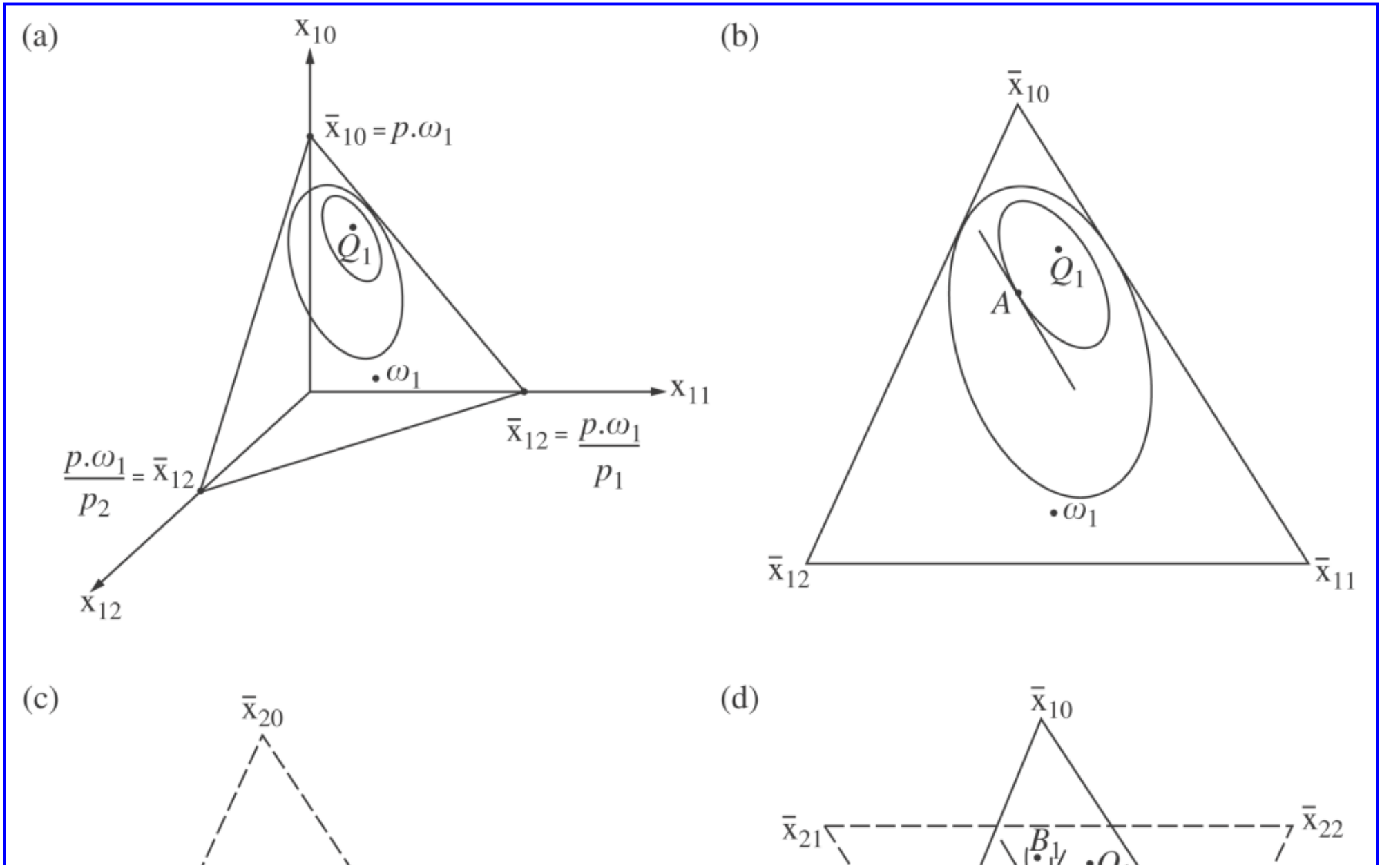

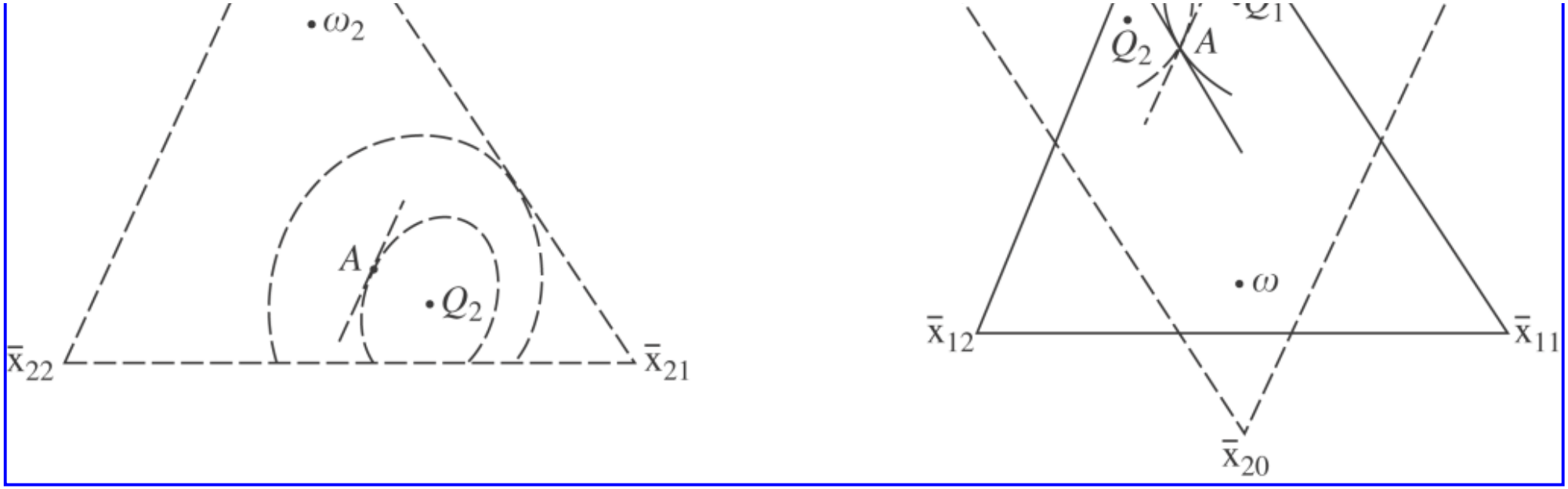

The three-good model

The model originated in Barro and Grossman (1971; 1976) and was further elaborated by Benassy (1977; 1982; 1986) and Malinvaud (1977) among others. Let there be three goods: money, denoted by M, initially available in M_0 units; labour, denoted by L, initially available in L_0 units, and output, denoted by Y, which is produced by labour according to the production function $Y = f(L)$. There are two markets, the labour market, with (nominal) wage w, and the output market, with price p. There is one firm and one consumer, with preferences represented by the homogeneous utility function $U(Y, M)$, who owns M_0 and L_0, and receives all profits. The labour supply is fixed at L_0.

Define the *marginal rate of substitution* as $V'(Y) \equiv \frac{\partial U/\partial Y}{\partial U/\partial M}$, with derivatives evaluated at (Y, M_0). Define the marginal cost curve as $C'_w(Y) \equiv w(f^{-1})'(Y)$, and the *full employment output* as $Y_0 \equiv f(L_0)$.

Definition 2: The level of output Y is a fixprice equilibrium output for the price–wage pair (p,w) if $Y = \min\{(V') - 1(p), (C'w) - 1(p), Y_0\}$.

This equality embodies in a compact way four conditions. First, $Y \leq Y_0$, that is, output cannot exceed the full employment level. Second, $Y \leq (V')^{-1}(p)$, or alternatively $p \leq V'(Y)$: the consumer cannot gain by buying less output at the going price (it is a condition of voluntary trading for the consumer). Third, $Y \leq (C'_w)^{-1}(p)$, or $p \geq C'_w(Y)$: the price cannot be lower than the marginal cost, or, in other words, profits cannot increase by selling less at the going price (it is a condition of 'voluntary trading' for the firm). Finally, at least one of these weak inequalities must be an equality: this is the condition of frictionless markets.

Figure 3 partitions the (p, w) plane according to which one of the three possible equalities determines output (solid lines). In region E (full employment), $Y = Y_0$. In region K (Keynesian unemployment), $p = V'(Y)$, and in region C (classical unemployment of full capacity), $p = C'_w(Y)$. In the boundaries between regions the two relevant equalities hold. At the Walrasian point W all three equalities hold. There is full employment in region E and unemployment outside it. The dashed lines are isoemployment loci, with the arrows indicating the directions of increasing employment.

Figure 3

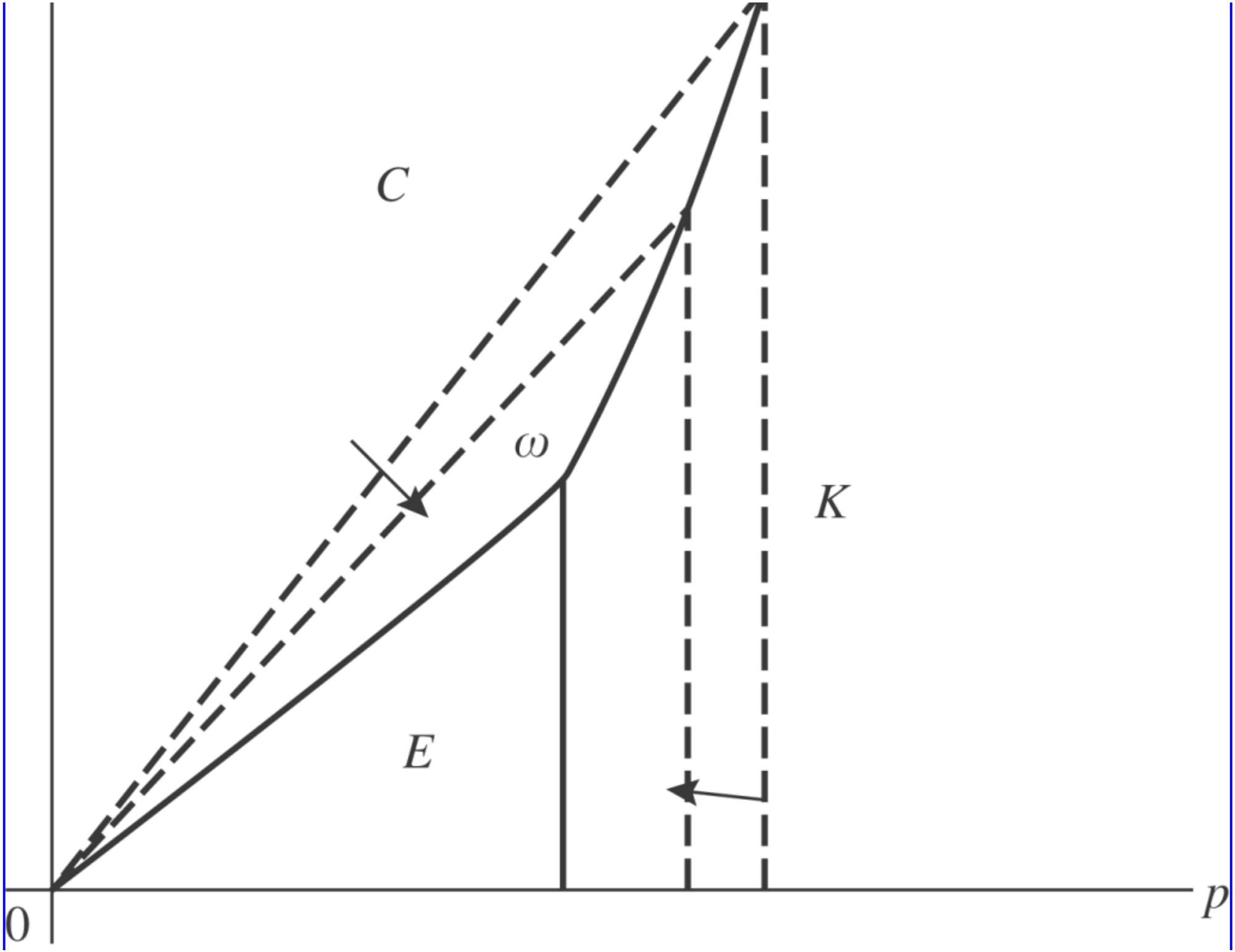
C
ω
K
E
0
p

The labour market is in excess supply (or excess demand) in the interior of regions *K* and *C* (or region *E*), and the output market is in excess supply (or demand) in the interior of region *K* (or regions *C* and *E*). At the Walrasian point *W* both markets are balanced. In the Keynesian region the condition $p = V'(Y)$ for determination of output can be rewritten in terms of the consumption function as in the textbook Keynesian multiplier model. The homogeneity of *U* implies that demand for output, as a function of *p* and wealth *I*, can be written as *h*(*p*)*I*, where the function *h*(*p*) satisfies: (*a*) $ph(p) < 1$ and (*b*) the marginal equality $(\partial U / \partial Y) / (\partial U / \partial M) = p$ whenever the consumption vector is a multiple of (*h*(*p*), 1 –*ph*(*p*)). By setting $I = M_0 + pY$ we obtain the *effective demand for output* $C(Y) = h(p)(M_0 + pY)$: this is the traditional consumption function, with marginal propensity to consume equal to $ph(p) < 1$. The satisfaction of effective demand requires $Y = C(Y)$, that is, $Y / M_0 = h(p) / [1 - ph(p)]$, which by the above marginal equality implies that $p = V'(Y)$.
The distinction between the two types of excess supply of labour has important implications for policy and for comparative statics. Output is determined in region *C* by the condition ‘price=marginal cost’. Hence, lowering wages (nominal or real) will increase employment, whereas an increase in demand will have no effect on employment. But in region *K* a decrease in the nominal wage has no effect on employment: only lowering the price or otherwise stimulating demand will work. This analysis also offers insights on the effects of different kinds of shocks (Malinvaud, 1977): a business cycle driven by demand shocks will fluctuate between Keynesian unemployment and full employment, whereas productivity shocks will yield fluctuations between the Keynesian and the classical types of unemployment.

Welfare analysis

The budget equality and the market institution impose constraints on trades. Thus, the resulting allocation may very well be Pareto dominated by other allocations that do not satisfy these constraints. The study of such inefficiencies is important for the normative analysis of the situations covered by fixprice theory (short-run market disequilibria, price controls, monopolistic market power). On the other hand, to the extent that these constraints cannot be circumvented, they are for policy purposes as effective as physical and resource constraints, motivating the study of efficiency subject to these additional constraints (‘second best’.)

Inefficiency relative to the set of physically attainable allocations

Consider Figure 1. Note that the allocation given by *A* is not Pareto efficient: both traders would be better off at *C*, but *C* cannot be reached without violating some budget constraint. A similar phenomenon may occur if there are two traders in one side of the market. Modify the example of Figure 1 by duplicating Trader 2: that is, Traders 1 and 2 are unchanged, but now there is a Trader 3 with the same preferences and endowments as Trader 2. Let $z_{21} = (-1/4)z_{11}$, and $z_{31} = (-3/4)z_{11}$. Then there are mutually beneficial reallocations between Traders 2 and 3, but they violate the budget constraint.
One can say that this type of inefficiency is caused by ‘wrong prices’. Note, however, that trade at non-Walrasian prices does not per se imply inefficiency. Point *D* in Figure 1, for instance, is Pareto efficient, and all budget constraints are satisfied there. (A general treatment of allocations of this type is given by Balasko, 1979, and Keiding, 1981.) But there is forced trading at point *D*. It is the combination of non-Walrasian prices and the voluntariness condition that implies inefficiency (Silvestre, 1985).

Inefficiencies relative to allocations satisfying the budget constraint

When there is only one market (see Figure 1), the absence of frictions guarantees that no allocation that satisfies the budget constraint is Pareto superior to a fixprice equilibrium; that is, a fixprice equilibrium is efficient relative to allocations that satisfy the budget constraints. This ceases to be true with several markets: for instance, point *B* in Figure 2(d) Pareto dominates point *A* and satisfies all budget constraints. (Note that point *B* violates voluntariness.) Such inefficiencies have been studied in Benassy (1975; 1977; 1982) and Younès (1975). A particularly striking case occurs in Keynesian allocations of the three-good model: the markets for labour and output are in excess supply, and a direct barter of labour against output would benefit both the firm and the worker, and improve welfare. This phenomenon was viewed by Clower (1965) as a failure of coordination among markets.

Undominated price – wage pairs

Suppose that, in the three-good model, wages and prices are determined by negotiation between representatives of labour and business, and then taken as given by individual firms and workers, so that a fixprice allocation results. If bargaining is efficient, any movement away from the negotiated price–wage pair (*p*, *w*) will make somebody worse off, in which case we say that (*p*, *w*) is *undominated*. Do there exist undominated price–wage pairs besides the Walrasian pair? Note that this question is different from the ones addressed in the previous paragraphs: there, we compared allocations at a given (*p*, *w*), whereas now we compare price–wage pairs.
The answer depends on the rationing of unemployment, that is, on whether unemployment falls uniformly on workers, or, on the contrary, some workers are dismissed whereas others

experience no rationing (Silvestre, 1988; 1989). In the first case, the answer is affirmative under some assumptions, in which case the set of undominated price–wage pairs is a segment of the Keynesian–classical boundary of Figure 3, implying that the output market is balanced. Non-uniform rationing of unemployment typically expands this set to a band of the Keynesian region adjoining the Keynesian–classical boundary.

The analysis is extended to a dynamic model by Jacobsen and Schultz (1990; 1991), who characterize the conditions for unemployment at undominated wages under the assumptions that unemployment is uniformly rationed, and that the output market is always balanced.

Wage and price rigidities in dynamic macroeconomic models

Dynamic stochastic macroeconomic models were first developed under the Walrasian assumptions of price taking and market clearing in models labelled ‘real business cycle’ (Kydland and Prescott, 1982; King and Plosser, 1984) aimed at mimicking business cycle regularities. Later developments have improved the fit, in particular for the persistence of real effects of monetary shocks, by introducing, singly or in combination, market power, or price or wage rigidity.

Here we focus on rigidities. (See Silvestre, 1995, for an early account of market-power, dynamic macroeconomic models; Svensson, 1986, combines market power with sticky prices in a dynamic model.) A first departure from Walrasian market clearing is the assumption that nominal wages are predetermined in the short run, before technological or monetary shocks are experienced. For instance, they may be preset at the expected Walrasian level (Gray, 1976), so that expected demand equals expected supply, whereas actual discrepancies between supply and demand are resolved in favour of demand: workers supply the amount of labour demanded by firms at the predetermined wage. (This simplifying assumption conflicts with the voluntariness condition of the canonical fixprice model, as described above. Benassy, 1995b; 2002, modifies the dynamic model of preset wages by postulating that unions maximize a utility function subject to the voluntariness condition.) This form of rigidity or stickiness yields predictions quite different from the Walrasian model: it grants monetary shocks the ability to generate large effects on employment and output, allowing for contemporary money shocks to generate countercyclical behaviour of the real wage, as well as cyclical behaviour of prices (Benassy, 1995a). The determinants of the accompanying welfare costs are studied in Cho, Cooley and Phaneuf (1997). A shortcoming of this approach is that monetary shocks show relatively little persistence, limited by the length of the period in which wages are fixed (Taylor, 1999).

This limitation, together with the observation that wage contracts are not synchronized across firms, led to the models of staggered wages or prices. In their simplest form (Taylor, 1979; 1980), wages are fixed for a given number N of dates, but in each date $1/N$ firms renew their contracts, so that at any moment the average wage is defined by the current contract plus the ones set in the last $N-1$ dates. More complex versions allow for various contract lengths. An influential formulation is that of Calvo (1983), who postulates that the contract length is stochastic: a given contract remains unchanged at each date with probability π, and terminated and reset with probability $1-\pi$. This approach typically yields propagation mechanisms and persistence characteristics capable of matching stylized facts of economic fluctuations (Benassy, 2002; 2003; Yun, 1996). Christiano, Eichenbaum and Evans (2005) show that wage staggering performs better than price staggering in generating the observed type of persistence, confirming Andersen's (1998) analysis in a model based on staggered prices or wages with fixed contract length.

See Also

- Arrow–Debreu model of general equilibrium
- dynamic models with non-clearing markets
- general equilibrium
- Keynesianism
- microfoundations
- new Keynesian macroeconomics
- second best
- sticky wages and staggered wage setting
- temporary equilibrium
- underemployment equilibria
- unemployment

Bibliography

Akerlof, G.A. and Yellen, J.L. 1985. A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100, 823–38.

Andersen, T.M. 1994. *Price Rigidity: Causes and Macroeconomic Implications*. Oxford: Clarendon Press.

Andersen, T.M. 1998. Persistence in sticky price models. *European Economic Review* 42, 593–603.

Arrow, K.J. 1959. Towards a theory of price adjustment. In *The Allocation of Economic Resources*, ed. M. Abramovitz. Stanford: Stanford University Press.

Baharad, E. and Eden, B. 2004. Price rigidity and price dispersion: evidence from micro data. *Review of Economic Dynamics* 7, 613–41.

Balasko, Y. 1979. Budget constrained Pareto-efficient allocations. *Journal of Economic Theory* 21, 359–79.

Barro, R.J. and Grossman, H.I. 1971. A general equilibrium model of income and employment. *American Economic Review* 61, 82–93.

Barro, R.J. and Grossman, H.I. 1976. *Money, Employment and Inflation*. Cambridge: Cambridge University Press.

Benassy, J.P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42, 503–23.

Benassy, J.P. 1976. The disequilibrium approach to monopolistic price setting and general monopolistic equilibrium. *Review of Economic Studies* 43, 69–81.

Benassy, J.P. 1977. A neo-Keynesian model of price and quantity determination in disequilibrium. In *Equilibrium and Disequilibrium in Economic Theory*, ed. G. Schwödiauer. Boston: Reidel.

Benassy, J.P. 1982. *The Economics of Market Disequilibrium*. New York: Academic Press.

Benassy, J.P. 1986. *Macroeconomics: An Introduction to the Non-Walrasian Approach*. New York: Academic Press.

Benassy, J.P. 1991. Microeconomic foundations and properties of a macroeconomic model with imperfect competition. In *Issues in Contemporary Economics: Markets and Welfare*, vol. 1, ed. K.J. Arrow. London: Macmillan.

Benassy, J.P. 1995a. Money and wage contracts in an optimizing model of the business cycle. *Journal of Monetary Economics* 35, 303–15.

Benassy, J.P. 1995b. Nominal rigidities in wage-setting by rational trade unions. *Economic Journal* 105, 635–43.

Benassy, J.P. 2002. *The Macroeconomics of Imperfect Competition and Nonclearing Markets: A Dynamic General Equilibrium Approach*. Cambridge, MA: MIT Press.

Benassy, J.P. 2003. Staggered contracts and persistence: microeconomic foundations and macroeconomic dynamics. *Recherches Économiques de Louvain/Louvain Economic Review* 69, 125–44.

Bewley, T.F. 1999. *Why Wages Don't Fall during a Recession*. Cambridge, MA: Harvard University Press.

Blanchard, O.J. and Kiyotaki, N. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77, 647–66.

Blinder, A.S., Canetti, E.R., Lebow, D.E. and Rudd, J. 1998. *Asking About Prices: A New Approach to Understanding Price Stickiness*. New York: Russell Sage Foundation.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Caplin, A.S. and Spulber, D.E. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102, 703–25.

Card, D. and Hyslop, D. 1997. Does inflation ‘grease the wheels of the labor market’? In *Reducing Inflation*, ed. C. Romer and D. Romer. Chicago: University of Chicago Press.

Carlton, D.W. 1986. The rigidity of prices. *American Economic Review* 76, 637–58.

Carlton, D.W. 1989. The theory and the facts of how markets clear: is industrial organization valuable for understanding macroeconomics? In *Handbook of Industrial Organization*, vol. 1, ed. R. Schmalensee and R.D. Willig. Amsterdam: North-Holland.

Cecchetti, S.G. 1984. Indexation and incomes policy: a study of wage adjustment in unionized manufacturing. *Journal of Labor economics* 5, 391–412.

Cecchetti, S.G. 1986. The frequency of price adjustment: a study of newsstand prices of magazines. *Journal of Econometrics* 31, 255–74.

Cho, J.O., Cooley, T. and Phaneuf, L. 1997. The welfare cost of nominal wage contracting. *Review of Economic Studies* 64, 465–84.

Christiano, L.J., Eichenbaum, M. and Evans, C.L. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–45.

Clower, R.W. 1965. The Keynesian counter-revolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Drèze, J. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16, 301–20.

Fishman, A. and Simhon, A. 2005. Can small menu costs explain sticky prices? *Economics Letters* 87, 227–30.

Gray, J.A. 1976. Wage indexation: a macroeconomic approach. *Journal of Monetary Economics* 2, 221–35.

Hahn, F.H. 1978. On non-Walrasian equilibria. *Review of Economic Studies* 45, 1–17.

Hahn, F.H. and Negishi, T. 1962. A theorem on non-tâtonnement stability. *Econometrica* 30, 463–9.

Hart, O. 1982. A model of general equilibrium with Keynesian features. *Quarterly Journal of Economics* 97, 109–38.

Jacobsen, H.J. and Schultz, C. 1990. A general equilibrium, macro model with wage bargaining. *Scandinavian Journal of Economics* 92, 379–98.

Jacobsen, H.J. and Schultz, C. 1991. Undominated wage rates in a unionized, overlapping generations economy. *European Economic Review* 35, 1255–75.

John, R. 1985. A remark on conjectural equilibria. *Scandinavian Journal of Economics* 87, 137–41.

Keiding, H. 1981. Existence of budget constrained Pareto-efficient allocations. *Journal of Economic Theory* 24, 393–7.

King, R.G. and Plosser, C.I. 1984. Money, credit and prices in a real business cycle. *American Economic Review* 74, 363–80.

Kydland, F.E. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Madden, P. and Silvestre, J. 1991. Imperfect competition and fixprice equilibria when goods are gross substitutes. *Scandinavian Journal of Economics* 93, 479–94.

Madden, P. and Silvestre, J. 1992. Imperfect competition and fixprice equilibria under consumer aggregation and net substitutes. *Scandinavian Journal of Economics* 94, 103–11.

Malinvaud, E. 1977. *The Theory of Unemployment Reconsidered*. Oxford: Basil Blackwell.

Mankiw, N.G. 1985. Small menu costs and large business cycles: a macroeconomic model of monopoly. *Quarterly Journal of Economics* 100, 529–39.

Negishi, T. 1960. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28, 196–201.

Negishi, T. 1965. Market clearing processes in a monetary economy. In *The Theory of Interest Rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Negishi, T. 1974. Involuntary unemployment and market imperfection. *Economic Studies Quarterly* 25, 32–46.

Negishi, T. 1979. *Microeconomic Foundations of Keynesian Macroeconomics*. Amsterdam: North-Holland.

Patinkin, D. 1956. *Money, Interest and Prices*. Evanston, IL: Row Peterson.

Parkin, M. 1986. The output–inflation tradeoff when prices are costly to change. *Journal of Political Economy* 94, 200–24.

Sheshinski, E. and Weiss, Y. 1977. Inflation and costs of price adjustment. *Review of Economic Studies* 54, 287–303.

Silvestre, J. 1982. Fixprice analysis in exchange economies. *Journal of Economic Theory* 26, 28–58.

Silvestre, J. 1983. Fixprice analysis in productive economies. *Journal of Economic Theory* 30, 401–09.

Silvestre, J. 1985. Voluntary and efficient allocations are Walrasian. *Econometrica* 53, 807–16.

Silvestre, J. 1986. The elements of fixprice microeconomics. In *Microeconomic Theory*, ed. L. Samuelson. Boston: Kluwer Nijhoff.

Silvestre, J. 1988. Undominated prices in the three good model. *European Economic Review* 32, 161–78.

Silvestre, J. 1989. Who benefits from unemployment? In *The Economics of Imperfect Competition and Employment: Joan Robinson and Beyond*, ed. G. R. Feiwel. London: Macmillan.

Silvestre, J. 1990. There may be unemployment when the labor market is competitive and the output market is not. *Economic Journal* 100, 899–913.

Silvestre, J. 1993. The market-power approach to macroeconomic policy. *Journal of Economic Literature* 31, 105–41.

Silvestre, J. 1995. Market power and macroeconomic policy: new developments. *Annales d'Économie et de Statistique* 37–38, 319–56.

Svensson, L. 1986. Sticky goods prices, flexible asset prices monopolistic competition and monetary policy. *Review of Economic Studies* 53, 385–405.

Taylor, J.B. 1979. Staggered wage setting in a macro model. *American Economic Review* 69, 108–13.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

Taylor, J.B. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of Macroeconomics*, vol. 1, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Uzawa, H. 1962. On the stability of Edgeworth's barter process. *International Economic Review* 3, 218–32.

Younès, Y. 1975. On the role of money in the process of exchange and the existence of a non-Walrasian equilibrium. *Review of Economic Studies* 42, 489–501.

Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37, 345–70.

How to cite this article

Silvestre, Joaquim. "fixprice models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000141>
doi:10.1057/9780230226203.0587

The New Palgrave Dictionary of Economics Online

Fleming, John Marcus (1911 – 1976)

S.C. Tsiang

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Fleming, J. M.; Haberler, G. H.; welfare economics

Article

Fleming was born on 13 March 1911 at Bathgate, Scotland, and died on 3 February 1976. He was educated at Edinburgh University, where he received the degrees of MA (Honours) in history in 1932 and MA (First Class Honours) in political economy in 1934. He was a graduate research fellow at the Institut Universitaire des Hautes Etudes Internationales in 1934–5, and a graduate student at the London School of Economics in 1935.

At the end of 1935, he joined the Secretariat of the League of Nations, Economic Intelligence Section, as a research economist, and assisted Gottfried Haberler in the latter's *Prosperity and Depression* (first published by the League in 1937). During the Second World War, he served with the UK Ministry of Economic Warfare from 1939 until 1942, and then joined the Economic Section of the Cabinet Office under Lord Robbins, rising eventually to the position of Deputy Director of the section. He was also a member of the UK Delegation to the San Francisco Conference in 1945; a member of the Preparatory Commission of the United Nations in 1946; a member of the International Trade Conference Preparatory Commission, 1947; and UK Representative to the Economics and Employment Commission, United Nations in 1950. From 1951 to 1954 Fleming was Visiting Professor at Columbia University, New York. He joined the International Monetary Fund in 1954 as a division chief, and in 1964 became the Deputy Director of the Research Department.

His academic contributions are mostly in the fields of welfare theory and trade and exchange policies. The most notable of his contributions was the seminal article ‘On Making the Best of Balance of Payments Restrictions on Imports’ (1951), which, in James Meade's words, was ‘the begetter of the analysis of the second best’ (Meade, 1978), which rapidly became a fashionable new topic in welfare theory during the 1950s.

Selected works

1944. (With J.E. Meade.) Price and output policy of state enterprise: a symposium. *Economic Journal* 54, 321–8, 337–9.

1951. On making the best of balance of payments restrictions on imports. *Economic Journal* 61, 48–71.

1952. A cardinal concept of welfare. *Quarterly Journal of Economics* 6, 366–84.

1961. International liquidity: ends and means. *IMF Staff Papers* 8, 439–63.

1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9, 369–79.

1964. (With R.A. Mundell.) Official intervention in the forward exchange market. *IMF Staff Papers* 11, 1–17.

1971a. *Essays in International Economics*. Cambridge, MA: Harvard University Press.

1971b. The SDR: some problems and possibilities. *IMF Staff Papers* 18, 25–47.

1977. International aspects of inflation. In *Inflation Theory and Anti-inflation Policy*, ed. E. Lundberg. London: Macmillan.

Bibliography

Haberler, G. 1937. *Prosperity and Depression*. Geneva: league of Nations.

Meade, J.E. 1978. Commemorations. In J. Marcus Fleming (posthumous), *Essays on Economic Policy*. New York: Columbia University Press.

How to cite this article

Tsiang, S. C. "Fleming, John Marcus (1911–1976)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000138> doi:10.1057/9780230226203.0588

The New Palgrave Dictionary of Eco

flypaper effect

Robert P. Inman
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The flypaper effect results when a dollar of exogenous grants-in-aid leads to significantly greater public spending than an equivalent dollar of citizen income: money sticks where it hits. Viewing governments as agents for a representative citizen voter, this empirical result is an anomaly. Four alternative explanations have been offered. First, it is a data problem; exogenous aid is mismeasured. Second, it is an econometric problem; important explanators of spending correlated with aid or income are excluded from the specification. Third, it is a specification problem; the representative citizen misperceives aid and the rational voter model misses this point. The empirical evidence suggests none of these explanations is sufficient. A fourth explanation seems most promising: it is politics. Rather than an anomaly, the flypaper effect is best seen as an outcome of political institutions and the associated incentives of elected officials.

Keywords

flypaper effect; grant aid; political spending; public funds

Article

In the late 1960s James Henderson (1968) and Edward Gramlich (1969) changed the direction of empirical research on how local governments tax and spend. While all prior work detailed the demographic and economic correlates with government budgets, Henderson and Gramlich sought an explanation for those correlations. To them as economists, the answer was clear. Citizens

demand services from their elected officials, and elected officials respond subject to the availability of government resources. Resources come from citizen incomes and from fiscal transfers given by the central government as grants-in-aid. From this perspective, Henderson and Gramlich specified and estimated demand equations based on the maximization of a representative citizen's utility subject to that citizen's ‘full income’ constraint specified as the sum of personal income and the citizen's share of the government's unconstrained fiscal transfers. So specified, personal income and the citizen's share of fiscal transfers should impact spending identically – money is money. The empirical analyses of Henderson and Gramlich revealed something unexpected, however. An extra dollar of personal income increased government spending on the order of $0.02 to $0.05, but an equivalent extra dollar of grants-in-aid increased government spending by from $0.30 to often as much as a full dollar. When Gramlich first presented his results, his colleague Arthur Okun called this larger effect of lump-sum aid on government spending a ‘flypaper effect’, noting that ‘money seems to stick where it hits’. The label stuck too, as has the puzzle of why intergovernmental transfers are so stimulative. A Google search reveals that over 3,500 research papers – excluding those studying the effects of real flypaper on insect populations – have now been written documenting and seeking to explain the flypaper effect.

Why do we care about this apparent anomaly? There are two reasons. First, as a matter of policy, understanding *how* recipient governments spend intergovernmental transfers is essential for the design of efficient fiscal policy in federal economies. Second, as a matter of science, understanding *why* governments spend citizens' incomes as they do provides valuable insights into how citizen preferences are represented in government policies. The taxation of citizen incomes and the allocation of grants-in-aid provide two ‘tracers’ as to the inner workings of political decision-making, one (taxes) that is directly observed and controlled by citizens, and the other (grants) perhaps only imperfectly so.

The benchmark for both the policy and political economy literatures is how a politically decisive citizen would like to see government resources allocated, specified by the maximization of that representative citizen's welfare over private (x) and public (g) goods, indexed by U(x, g), subject to a current period budget constraint specified as:

$$Y = \{I + h \cdot z\} = x + p_g \cdot g$$

where I is the citizen's private income (or tax base), h is the citizen's share of *unconstrained or lump-sum* intergovernmental transfers per capita (z) specified as $h = I / \bar{I}$ with $\bar{I}$ equal to the average income (or tax base) in the citizen's political jurisdiction, and p_g is the 'tax price' for government services (g) equal to $c \cdot (1-m) \cdot h$ where c is the per unit production cost of g and m is the matching rate for open-ended matching federal aid. Private goods cost $1. Y is called the citizen's 'full income'. The citizen's preferred allocations will be $x=x(1, p_g, Y)$ and $g=g(1, p_g, Y)$, where:

$$\Delta g_I = (\delta g / \delta Y) \cdot (\delta Y / \delta I) \cdot \Delta I = (\delta g / \delta Y) \cdot (\Delta I = \$1),$$

for an extra dollar of personal income and:

$$\Delta g_z = (\delta g / \delta Y) \cdot (\delta Y / \delta z) \cdot \Delta z = (\delta g / \delta Y) \cdot h \cdot (\Delta z = \$1)$$

for an extra dollar of aid, implying that estimated marginal effects of aid to income should be related as $\Delta g_z/\Delta g_I=h$. In most political jurisdictions the representative citizen has a tax base (often specified as the median tax base) less than the average tax base; thus, in most cases, if our representative citizen has had her way, then we should expect $\Delta g_z/\Delta g_I=h<1$. The overwhelming empirical evidence summarized by Gramlich (1977), Inman (1979), Fisher (1982) and Hines and Thaler (1995) shows just the opposite, however; Δg_I ranges from $0.02 to $0.05 while the companion estimates of Δg_z typically fall between $0.30 to $1.00. Income to the citizen stays with the citizen; grants to the government stay with the government. Money sticks where it hits. Why?

Four explanations have been offered. First, the answer is in the data. Researchers mismeasure intergovernmental aid by confusing matching grants that lower the marginal price of public services (p_g) with lump-sum aid (z) that shifts outward the representative citizen's budget constraint. Matching aid has a price effect, lump-sum aid an income effect. For local politics controlled by a representative citizen, consumer theory predicts that a matching grant's price effect will stimulate more government services than an equivalent dollar of lump-sum aid. If the dollar transfer received from matching aid is erroneously classified as lump-sum aid, then $\Delta g_z > \Delta g_I$ will result; see Moffitt (1984), Megdal (1987), and Baker, Payne and Smart (1999). Even after correctly classifying aid programmes and measuring p_g and z appropriately, however, the flypaper effect remains; see for example Wyckoff (1991).

The second explanation sees the anomaly as an econometric problem. Researchers may have omitted important determinants of government spending likely to be correlated with citizen income or intergovernmental aid, leading to biased estimates of Δg_I and Δg_z. Bruce Hamilton (1983) and Jonathan Hamilton (1986) attribute the flypaper effect to misspecifications of the technology or costs of providing local services. Bruce Hamilton argues that estimated demand equations omit important variables such as the citizen's talents or willingness to volunteer which are positively correlated with citizen income and also contribute to the provision of government services. If these omitted effects are substitutes for (negatively correlated with) purchased government inputs, then the estimated coefficient for income will be biased downward, perhaps sufficiently so that $\Delta g_z > \Delta g_I$. Jonathan Hamilton suggests the misspecification arises from a failure to account correctly for residential exit from high tax jurisdictions leading to a loss of tax base when specifying the price of government services. Local taxes are inefficient and the correctly specified price of local services must reflect this fact. If citizens tend to reside in localities of comparable income, and higher-income residents are more mobile, then the representative citizen's income will be positively correlated with the correct price, which is negatively correlated with government services. Again, there is a downward bias in the estimated income effect, with $\Delta g_z > \Delta g_I$ as a possible result.

Neither of the Hamiltons's biases are likely to fully explain estimated flypaper effects, however. A plausible upper estimate for Δg_I can be obtained as $\Delta g_I = (\delta g/\delta Y) = \varepsilon_{g,Y} \cdot (g/Y)$, where $\varepsilon_{g,Y}$ is the income elasticity of demand for

government services and g/Y is the average rate of spending by recipient governments. This ratio for the US state and local government sectors combined from 1970 to 2008 – the period used for most all studies – is at most 0.15. Since most state and local services are arguably necessities, $\varepsilon_{g,Y} \leq 1$ seems reasonable. If so, then $\Delta g_I \leq (g/Y)=0.15$ bounds an unbiased income effect. Since most estimates of Δg_z exceed 0.15, the flypaper effect remains. Perhaps then the explanation lies in an upward bias in the estimates of Δg_z? Here the results of four recent studies are particularly instructive. Each takes advantage of a plausibly exogenous, or 'natural experiment', change in lump-sum national aid to state or local governments. Gordon (2004) uses US federal legislation's required changes in Title I education aid caused by state-level (exogenous to the local budget) demographic changes before and after census years as her measure of exogenous aid. She finds strong evidence of a flypaper effect for local school districts in the first year after the change in Title I aid – $\Delta g_z=1.00$ – but that this effect evaporates after three years, with most of the new aid returned to voters as lower local tax revenues. In contrast, Ladd (1993) and Singhal (2008) find evidence for a significant and quantitatively large flypaper effect for US state governments, as do Dahlberg et al. (2008) in their study of national aid to municipalities in Sweden. Ladd uses windfall tax revenues to state governments following the Tax Reform Act of 1986 as her exogenous measure of aid, and estimates $\Delta g_z=0.40>\Delta g_I=0.03$. Singhal (2008) uses outside revenues received by state governments from a recent legal settlement with the tobacco industry as her measure of z, and finds $\Delta g_z=0.20$ for spending on tobacco control programmes, compared with an estimate of $\Delta g_I \approx 0$ for income's effects on the same programmes. Dahlberg et al. (2008) exploit a discontinuity in the national aid formula that gives significant additional assistance to communities that experience more than 2 per cent outmigration over the previous ten years; communities just below the threshold receive no additional aid, those just above do. The analysis includes community and time fixed effects – there is no direct estimate of Δg_I – and they find $\Delta g_z=1.00$ and no local tax relief. Ladd's, Singhal's and Dahlberg's estimated flypaper effects remain over time.

The flypaper effect appears to be a real phenomenon. As a third explanation, then, perhaps our model of citizen fiscal choice is misspecified. First, voters

may not understand the complexity of grant programmes. Both Courant, Gramlich and Rubinfeld (1979) and Oates (1979) conjecture that the representative citizen misperceives lump-sum aid's income effect as an average price effect. They conjecture that the voter uses taxes paid per unit of services received – $(p_g \cdot g - z)/g$ or $p_g - (z/g)$ – as their estimate of the true marginal tax cost of government services, p_g. If so, lump-sum aid (z) will impact spending as a price subsidy, and the estimated effect aid on spending will imply that $\Delta g_Z > \Delta g_I$. Wyckoff (1991) and Turnbull (1998) test this hypothesis by including both p_g and $[p_g - (z/g)]$ as competing explanators of local spending. They find plausible (negative) marginal price effects but implausible (positive) effects of the misperceived average price. Estimated flypaper effects are comparable to those of previous studies. From this evidence, it is unlikely that price misperception provides the explanation for the flypaper effect.

Filimon, Romer and Rosenthal (1982) and Hines and Thaler (1995) provide alternative versions of the voter ignorance hypothesis. For Filimon, Romer and Rosenthal the representative voter fails to see through the veil of government budgets; he does not know the level of aid received by the local government. For Hines and Thaler, the representative voter sees through the veil but budgets using mental accounts; there is a ‘public budget’ that is the responsibility of government officials and a ‘private budget’ that is the citizen's responsibility. Both hypotheses need a theory of public budgets to explain Δg_Z. Hines and Thaler leave this an open question, but Filimon, Romer, and Rosenthal are quite explicit: public officials are budget maximizers and therefore $\Delta g_Z = 1$. They test their theory for a sample of Oregon school districts, and cannot reject the null hypothesis that $\Delta g_Z = 1$ for state education aid.

In Romer, Rosenthal, and Munley (1992), the authors replicate their analysis for a sample of New York school districts, and here the conclusion varies by the size of the school district. Large districts (>20,000 students) show budget-maximizing behaviour and a full flypaper effect: $\Delta g_Z = 1$. In smaller districts, however, the estimated aid and income effects are about equal: $\Delta g_Z \approx h \cdot \Delta g_I$. These results parallel those from Ladd and Singhal for larger state governments and from Gordon for local school districts. Together, this evidence is sufficient to reject a strict version of the mental accounting

explanation. It leaves open, however, the question of why the flypaper effect remains for larger governments.

Here a fourth explanation for the flypaper effect seems the most promising: it is politics. This approach assumes that voters are informed and rational, but conceal their preferences when it is strategically useful to do so. Such strategic behaviours require the use of less than efficient institutions for preference revelation, such as majority rule or representative legislatures. From this perspective, the flypaper effect is a consequence of an inability of citizens to write complete ‘political contracts’ with their elected officials. Consistent with the results of Ladd, Singhal, and Romer, Rosenthal and Munley, we might expect these contracting problems to be greater, and the flypaper effect more likely, for large governments.

Chernick (1979) and Knight (2002) offer specifications of a political contract between a donor central government and recipient local governments as a way to understand the flypaper effect. Chernick (1979) specifies donor-recipient contracting as an auction. Assuming an exogenous level of federal aid, local governments bid for the right to provide aided services by offering to share the costs of provision. Beginning with the highest offer price, the central government selects recipient local governments until its grants budget is exhausted. The resulting allocation will equalize the marginal contribution of each local government to the incremental benefits from the provision of the local service. Local governments with the highest valuations will provide more services and receive more aid. Chernick offers evidence in support of this prediction from the US federal Water and Sewer Grant programme. Importantly, any reduced form estimate of Δg_z for this programme that did not account for the auction that sets aid would be biased upward and imply a strong flypaper effect.

Knight (2002) specifies and estimates a model of political contracting for grants policy that sets both the aggregate size of the aid budget and its allocation. The budget is chosen to ensure its passage and to maximize local constituent net benefits for the central government's agenda-setter. Again, the allocation process is an auction. Legislators bid to be part of the winning coalition by offering to vote for the grants budget in return for intergovernmental aid. The agenda-setter picks the smallest 51 per cent of the bids. He then sets his own grant award to maximize the net benefits to his own constituents. Those legislators whose state or local governments value the aided local service most highly make the winning offers. The result is again a positive correlation between grants awarded and local spending. Failure to

control for this correlation will lead to an upward bias in the estimate of Δg_z. For a statistically consistent estimate of Δg_z we need instruments that both predict grants (z) and are independent of constituents' demand for the aided service. Legislative institutions that select agenda-setters independent of constituent preferences will serve this purpose. Knight uses the legislators' tenures and majority party memberships as his instruments in his empirical study of highway grants and state highway spending. Least squares estimation of grants' effect on spending shows $\Delta g_z=1$; instrumental variables estimation rejects that extreme flypaper result but cannot reject a partial effect ($1>\Delta g_z>h\cdot\Delta g_I$). In a companion piece, Knight (2004) estimates that this agenda-setting process for highway grants imposes an allocative inefficiency of $0.40 per dollar of aid.

Over the first decade of the 21st century, the devolution of economic responsibilities to lower-tier governments has become increasingly important, not only in formally federal states but in unitary states as well. Central governments typically grant fiscal assistance to these local governments for the provision of those services. Knowing how grants will be spent is important for the appropriate design of central government transfer policies. Credible estimates of aid's effects on local spending requires good instrumental variables to predict aid, or ideally ‘natural experiments’ providing truly exogenous measures of central government assistance. Knowing how money is spent as it is helps us to understand allocative performance of intergovernmental transfers, given federal and local political institutions. Knowing why grant money is spent as it is, is just as important. Here the specification and estimation of structural models of central government transfer spending and local government allocations of transfer incomes are essential. This information provides a basis for reforming these important institutions, and there is perhaps no more striking example of the benefits of such structural analyses of the aid process than the work of Reinikka and Svensson (2003, 2004) on the allocation of Ugandan central government aid to local schools. Initially, only $0.15 of each centrally allocated school aid dollar found its way into the local schools; $0.85 was ‘captured’ by the district bureaucracy for its own use. The problem was inadequate information and weak local political organizations. Reforms publicized aid allocations and empowered village councils to monitor that spending. The end results was to reduce district capture to $0.15 per aid dollar – a plausible administrative cost – and to increase local school resources by $0.85 per aid dollar.

Once viewed as an anomaly, the flypaper effect should now be seen as a reality of fiscal politics, and its study as an opportunity to fashion central government transfer policies and intergovernmental fiscal institutions that better reflect citizen preferences for local public goods.

See Also

- foreign aid
- intergovernmental grants
- political institutions, economic approaches to

Bibliography

Baker, M., Payne, A. and Smart, M. 1999. An empirical study of matching grants: the ‘cap’ on CAP. *Journal of Public Economics* 72, 269–88.

Chernick, H. 1979. An economic model of the distribution of project grants. In *Fiscal Federalism and Grants-in-Aid*, ed. P. Mieszkowski and W. Oakland. Washington, DC: Urban Institute Press.

Courant, P., Gramlich, E. and Rubinfeld, D. 1979. The stimulative effects of intergovernmental grants: or why money sticks where it hits. In *Fiscal Federalism and Grants-in-Aid*, ed. P. Mieszkowski and W. Oakland. Washington, DC: Urban Institute Press.

Dahlberg, M., Mörk, E., Rattsø, J. and Ågren, H. 2008. Using a discontinuous grant rule to identify the effect of grants on local taxes and spending. *Journal of Public Economics* 92, 2320–35.

Filimon, R., Romer, T. and Howard Rosenthal, H. 1982. Asymmetric information and agenda control. *Journal of Public Economics* 17, 51–70.

Fisher, R. 1982. Income and grant effects on local expenditures: the flypaper effect and other difficulties. *Journal of Urban Economics* 12, 324–45.

Gordon, N. 2004. Do federal grants boost school spending? Evidence from Title I. *Journal of Public Economics* 88, 1771–92.

Gramlich, E. 1969. State and local governments and their budget constraint. *International Economic Review* 10, 163–82.

Gramlich, E. 1977. Intergovernmental grants: a review of the empirical literature. In *The Political Economy of Federalism*, ed. W.E. Oates. Lexington, Mass.: Lexington Books.

Hamilton, B. 1983. The flypaper effect and other anomalies. *Journal of Public Economics* 22, 347–62.

Hamilton, J. 1986. The flypaper effect and the deadweight loss from taxation. *Journal of Urban Economics* 19, 148–55.

Henderson, J. 1968. Local government expenditures: a social welfare analysis. *Review of Economics and Statistics* 50, 156–63.

Hines, J. and Thaler, R. 1995. Anomalies: the flypaper effect. *Journal of Economic Perspectives* 9, 217–26.

Inman, R. 1979. The fiscal performance of local governments: an interpretative review. In *Current Issues in Urban Economics*, ed. P. Mieszkowski and M. Straszheim. Baltimore, Md.: Johns Hopkins Press.

Knight, B. 2002. Endogenous federal grants and crowd-out of state government spending: theory and evidence from the Federal Highway Aid Program. *American Economic Review* 92, 71–92.

Knight, B. 2004. Parochial interests and the centralized provision of local public goods: evidence from congressional voting on transportation projects. *Journal of Public Economics* 88, 845–66.

Ladd, H. 1993. State responses to the TRA86 revenue windfalls: a new test of the flypaper effect. *Journal of Policy Analysis and Management* 12, 82–104.

Megdal, S. 1987. The flypaper effect revisited: an econometric explanation. *Review of Economics and Statistics* 59, 347–51.

Moffitt, R. 1984. The effects of grants-in-aid on state and local government spending: the case of AFDC. *Journal of Public Economics* 23, 279–305.

Oates, W. 1979. Lump-sum intergovernmental grants have price effects. In *Fiscal Federalism and Grants-in-Aid*, ed. P. Mieszkowski and W. Oakland. Washington, DC: Urban Institute Press.

Reinikka, R. and Svensson, J. 2003. The power of information: evidence from a newspaper campaign to reduce capture. Mimeo., World Bank, December.

Reinikka, R. and Svensson, J. 2004. Local capture: evidence from a central government transfer in Uganda. *Quarterly Journal of Economics* 93, 6562–87.

Romer, T., Rosenthal, H. and Munley, V. 1992. Economic incentives and political institutions: spending and voting in school budget referenda. *Journal of Public Economics* 4, 1–33.

Singhal, M. 2008. Special interest groups and the allocation of public funds. *Journal of Public Economics* 92, 548–654.

Turnbull, G. 1998. The overspending and flypaper effects of fiscal illusion: theory and empirical evidence. *Journal of Urban Economics* 44, 1–26.

Wyckoff, P. 1991. The elusive flypaper effect. *Journal of Urban Economics* 30, 310–28.

How to cite this article

Inman, Robert P. "flypaper effect." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_F000323> doi:10.1057/9780230226203.1901

The New Palgrave Dictionary of Economics Online

Fogel, Robert William (born 1926)

Stanley L. Engerman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A pioneer in the development of cliometrics, Robert Fogel has always focused on linking economic analysis to the study of historical problems and on the need for large-scale data collection and analysis. Based on this approach, he found that that slavery was profitable and viable even on the eve of the American Civil War; and he used information on height to draw inferences on food consumption by slaves. Fogel has since become involved in the economics of aging and longevity, the impact of the expansion of leisure time in the developed world, and the increasing burden of health care.

Keywords

aging populations; agriculture; anthropometric history; cliometrics; counterfactuals; economic history; Fogel, R.; health care; nutrition; railroads; slavery; transport

Article

Robert William Fogel is one of the pioneering figures in the development of cliometrics or the new economic history during the 1950s, a contribution for which he was awarded, with Douglass C. North, another innovator, the 1993 Nobel Prize in economic science. As of 2005 they are the only two economic historians to obtain this honor. Fogel was born in New York City on 1 July 1926, and graduated from Cornell University in 1948. Active politically in left-wing organizations for several years, he did not begin graduate work in economics until after 1956. He received a master's degree from Columbia University writing under the supervision of Carter Goodrich. He then went to the John Hopkins University, receiving a Ph.D. under the direction of Simon Kuznets in 1963. He has held teaching positions at the University of Rochester, the University of Chicago, and Harvard University, being the Charles R. Walgreen Distinguished Professor of American Institutions at the University of Chicago since 1981.

Fogel's career has always focused on the linking of economic analysis to the study of historical problems. The application of economic theory to specific historical questions has characterized his

writings since he began graduate work. Also characteristic of his work has been a concern with empirical data, at first the use of quantitative data to study specific problems, and later, with developments in computer technology, with attention given to the collection and analysis of large data-sets of economic and demographic evidence.

Railroads

Fogel's first book, *The Union Pacific Railroad: A Case in Premature Enterprise*, based on his master's thesis, was published in 1960. The basic question asked was whether the building of the Union Pacific in the 1860s, with government subsidy, led to corruption and abnormal profits earned by the railroad's builders and promoters. It had long been a staple of historical scholarship about the Union Pacific that the charges of corruption were true, and that the Union Pacific was to be viewed as part of America's late 19th-century ‘Great Barbecue’ and the ‘Gilded Age’. Fogel's extensive primary research permitted some more accurate accounting measures of the profits, and he used data on bond prices and related information to estimate the anticipated ‘risk of failure’ at the time of financing. By adjusting the accounting profits and pointing to the great measured risk in this pioneering transcontinental venture, Fogel argued that the extent of abnormal profits was overstated, but also that the mixture of public and private financing may not have been the most effective way to undertake construction. The novel historical application of economic theory here was in the measuring of the market assessment of risk on the basis of standard financial models.

Fogel's next book, *Railroads and American Economic Growth: Essays in Econometric History*, was based upon his doctoral dissertation. Published in 1964, it has become one the two early classics of cliometrics, the other being the study of the economics of American slavery by Alfred H. Conrad and John R. Meyer (1958). This work, aimed at estimating the contribution of the railroad to American economic growth in the 19th century, something that many contemporaries had discussed, led to significant debates about both economic techniques of measurement and the methodological principles of historical analysis. Fogel's book (and that published a few years later by Albert Fishlow, 1965) asked, on the basis of considerable empirical information, what the estimated difference in costs was between shipping goods by railroad and shipping them by the next-best alternative: road, canal, river, or lake. This was used to measure what Fogel called the ‘social savings’ based on the difference in costs of shipping between railroads and alternatives, and was to be the basic measure of the railroads’ contribution to economic growth in 1890. That the number came out smaller than expected led to some critiques of the analysis, but in a controversial next step Fogel argued that this was too high, since it did not allow for possible structural adjustments in the economy to the absence of a railroad, including the building up of a canal network that never existed, but seemed feasible, if necessary, and for which ample amounts of water existed. Also, following the development economics of the period, Fogel estimated the contribution of the railroad via backward linkages (for inputs) and forward linkages (for outputs). In general none made the contribution of the railroad as large as expected, a point that has been used to argue that no single innovation can itself explain much growth, and that for an economy to be successful there has to be a broad spread of productivity gains within the economy. Whatever the specific criticisms, the overall fruitfulness of Fogel's method of analysis is seen in the number of country studies undertaken using his approach for the study of the economic effects of the railroad. Some studies

indicate a small growth contribution, although in several cases (particularly Mexico) the measured effects were large due to the lack of good alternative means of transportation.

A major debate concerning the use of so-called ‘what if’ or counterfactual history in Fogel's analysis arose primarily among historians. To economists used to drawing supply and demand curves and discussing the impact of changes, this approach was rather standard and not questioned, and one might have felt that, given the form of most historical analysis, the same general acceptance would be expected among historians. This was not, however, a general view, and the explicit use of counterfactuals led to much debate. In some cases there was, as there should be, a questioning of the appropriateness of the particular counterfactuals used, since, as argued later, a counterfactual based on Napoleon using an atomic bomb is of doubtful usefulness. In other cases, however, the criticism was of the general use of the approach, with the implications that no ‘what if’ statement can be used at any time. This debate has disappeared in recent years, with apparent agreement that counterfactuals have long been part of the historian's approach to the past, and their use is a generally accepted, if not necessary, part of any historical study.

Economics of slavery

Fogel's next major project concerned one of the major issues of American historiography, the economics of slavery in the United States South. This project led to numerous publications, including several books and many articles, over a 30-year period by Fogel, his colleagues, and his students. The first major publication, in 1974, was the two-volume *Time on the Cross* (co-authored with Stanley Engerman); the first volume subtitled *The Economics of American Negro Slavery*, aimed at a large audience, and the second subtitled *Evidence and Methods: A Supplement*, which contained more detailed descriptions of data and analysis, aimed primarily at a professional, scholarly readership. These works presented findings from numerous types of primary data located in southern archives as well as census publications and manuscripts, and used many research assistants to collect and analyse the primary data, a practice then not typical in either history or economics. Earlier work on manuscript data from the federal census of 1860, prepared by William N. Parker (1970) and Robert E. Gallman (1970), was also of particular use, and the works of numerous historians and economists, in previous decades as well as that available from the then booming area of slavery studies, was important in shaping the arguments.

Both because of its heavy use of quantitative methods and also because of several of its major findings that seemed to go against some then commonly held views, *Time on the Cross* attracted an unexpected amount of attention and criticism for an academic publication, and there emerged a rather extended series of debates on many of the questions studied, leading to the publication of several books and many articles developing these disagreements. As before, some of the debate was about the nature of questions asked and some about the specifics of the substantive analysis.

The major economic findings in *Time on the Cross* were that slavery was profitable and was expected, by southerners and others, to be viable even on the eve of the Civil War. These findings were based on standard measures of profitability and price–rental ratios, but the calculation required collections of data on slave prices (by age, sex, and so on), slave productivity, slave demography, and the material consumption allowed to the slaves by their masters. While profitability and viability had not always been widely accepted, by the time the debates ended they did seem to be acceptable, suggesting that slavery would not collapse under of its own weight, that southern planters had behaved in a manner that

indicated a responsiveness to economic incentives, and, moreover, with the use of related evidence, that the South was doing quite well economically on the eve of the Civil War. Two other arguments were, and remain, still somewhat debated. A straightforward economist's measure of the relative productivity of northern and southern agriculture in 1860, used to answer a question long-discussed by contemporaries and many subsequent scholars, compared agricultural output with inputs of land, labour, and capital, and indicated that southern agriculture was more ‘efficient’ than northern agriculture and that within the South it was the larger slave-using plantations (over 16 slaves) that were more ‘efficient’ than were the small, free white farms and smaller slave farms. The concept of efficiency was interpreted by some, not as a standard concept of economic measurement, but as a measure with distinct moral overtones. The findings for the South led a discussion of economies of scale in slave plantations in the United States and in Caribbean sugar production, and the importance of scale has been seen to be significant for understanding slave societies as well as for evaluating the economic adjustment to the emancipation of slaves. A second continuing controversy concerned what was regarded as the favorable material treatment allowed slaves, based upon estimates of consumption allowed by masters, and the argued-for limited impact on slave family and cultural life. The former argument was based on demographic and related evidence. These questions have now become more important, and the ability of slaves to defeat masters’ attempts to exercise complete power over slaves is now more widely argued for in slave studies. Nevertheless, some disagreements on these issues remain.

In the aftermath of the *Time on the Cross* debates various articles by its co-authors and others were written for conference presentation and for publication. In 1989 Fogel published a new book on slavery, *Without Consent or Contract: The Rise and Fall of American Slavery*, which covered some of the earlier themes but also provided much new information on the politics of abolition in Britain and the United States. In general Fogel expanded on several discussions, particularly on cultural and demographic matters, but he still maintained most of the basic positions of his earlier writings on slavery, and the book is more of a defence than a revision of those arguments. In 1992 three edited volumes of earlier papers and notes by Fogel and others were published, adding greatly to the information and analysis of *Without Consent or Contract*. Fogel was invited to give the William Lynwood Fleming Lectures at Louisiana State University in 2001, published in 2003 as *The Slavery Debates, 1952–1990*. This brief, non-technical volume reviews the many debates on slavery, examines the trends and shifts in the study of slavery in the United States, and provides a very useful summary of changes in views over a 50-year period.

Heights and demographic history

One of the debates concerning the material conditions of slave life related to the issue of food consumption and nutrition. Only after the publication of *Time on the Cross* did Fogel and his collaborators become aware of the valuable information provided by information on height. The collection of this data from coastal shipping manifests of slaves carried in the interstate movement between 1808 and 1865, and other sources such as military records and the registrations of free blacks, turned out to be exceptionally important, both for comparisons of the heights of southern slaves with other populations, supporting the argument of basically adequate consumption by slaves, and in opening another major project for Fogel and for other economic historians, historians, and economists. There were, of course, some difficulties in making inferences about food consumption from information on

height, given differences in work regimen and disease environments and some truncations introduced by height requirements.

Nevertheless, so widespread were available data on heights in many countries over long periods of time, mainly from military records, that the study of height and its use as an alternative (or complementary) measure of welfare became widely used by economic historians in many different countries. Studies by Floud, Wachter and Annabel (1990), Komlos (1994), Steckel (1995), Goldin and Rockoff (1992), and Steckel and Floud (1997), among others, frequently utilized measures for comparative purposes across countries, as well as for studying long-term trends within specific countries. Some unexpected patterns developed, such as long-period cycles in height, rather than simply monotonic change, and periods of time in which heights and per capita incomes move in different directions. Fogel used the study of heights as a method of approaching a number of different problems, such as long-term variations in longevity and health and their contributions to economic growth, changes in diseases and patterns of aging, and the economics of the health care industry. As earlier, several of these projects were based on extensive data retired from archival sources, and required collaborative work with many scholars from different disciplines.

In 1996 Fogel gave the McArthur Lectures at Cambridge University and these were published in 2004 as *The Escape from Hunger and Premature Death, 1700–2100: Europe, America, and the Third World*. In these essays he was concerned with changes in productivity due to improvements in human capacity to perform. He focused on changes in the 20th century in health, the dramatic change in the caloric input of the French and British populations from their earlier limited available energy, and also the great increases in available leisure time. Such benefits of health and leisure have not yet occurred in much of the Third World today, where people adapt to limited energy by a smaller body size, limiting the prospective productivity in these societies relative to that in the developed world.

Fogel's most recent project, based on extensive data collection from archival sources and involving many students and scholars in collection and analysis, concerns long-term longitudinal studies of health, diseases, and the role of socio-economic and biomedical factors. The initial major data-set was based on the pension records of the Union army in the Civil War, which present very detailed medical histories of veterans from childhood until death, and run from the Civil War into the 20th century. These data, with more recent information, provide a basis for examining not only changes in life expectation and health, but also the nature of the changing pattern of diseases over time. These studies had been supplemental by other longitudinal data-sets, including sampling of births and of babies born between 1910 and 1934, to examine inter-generational factors in health and longevity. As a result of these studies Fogel has become involved in the analysis of the economics of aging and longevity, the impact of the expansion of leisure time in the developed world, and the increasing burden of health care and the complexities of achieving equity in health care in recent years, arguing that these issues reflect social and economic progress, not new difficulties. He has also estimated, based on the work of Dora Costa (2003) and others, the magnitude of the continued increase in the length of life in the 20th century.

Fogel has also made other important contributions to the study of economics and history, to the study of methodology in the social sciences and in history, as seen in his debate with Geoffrey Elton, *Which Road to the Past? Two Views of History* (1983), and to the study of the relations among religious, economic, and political changes. His 2000 book, *The Fourth Great Awakening and the Future of Egalitarianism*, studied the emergence of a religious belief in egalitarianism over time, and how the periodic bursts of awakening influenced the political and economic worlds, as well as what were the

major changes in measured inequality in the United States over time.

In addition to his numerous contributions to economics and economic history, Fogel has been a leading figure in proselytizing for cliometrics in economics and in history (including the publication of a 1971 collection of cliometric essays, *The Reinterpretation of American Economic History*, coedited with Stanley Engerman), has been influential in advocating large-scale data collection and analysis, and has been a major producer of scholars for the next generation of economic history. Honours, besides the Nobel Prize, include membership in the National Academy of Science and the American Academy of Arts and Sciences; presidencies of the Economic History Association, the Social Science History Association, and the American Economic Association; the Bancroft Prize; and the Pitt Professorship of American History and Institutions at the University of Cambridge. He was the first director of the Development of the American Economy Program of the National Bureau of Economic Research, chairman of the Committee on Mathematical and Statistical Methods in History of the Mathematical Social Science Board, and is presently the Director of the Center for Population Economics at the University of Chicago.

See Also

- anthropometric history
- cliometrics
- economic history
- population aging

Selected works

1960. *The Union Pacific Railroad: A Case in Premature Enterprise*. Baltimore, MD: Johns Hopkins Press.

1964. *Railroads and American Economic Growth: Essays in Econometric History*. Baltimore, MD: Johns Hopkins Press.

1971. (With S. Engerman, eds.) *The Reinterpretation of American Economic History*. New York: Harper and Row.

1974. (With S. Engerman.) *Time on the Cross: The Economics of American Negro Slavery*. Boston: Little, Brown.

1974. (With S. Engerman.) *Time on the Cross: Evidence and Method, A Supplement*. Boston: Little, Brown.

1983. (With G. Elton.) *Which Road to the Past? Two Views of History*. New Haven, CT: Yale University Press.

1989. *Without Consent or Contract: The Rise and Fall of American Slavery*. New York: Norton.

2000. *The Fourth Great Awakening and the Future of Egalitarianism*. Chicago: Chicago University Press.

2004. *The Escape from Hunger and Premature Death, 1700–2100: Europe, America, and the Third World*. Cambridge: Cambridge University Press.

Bibliography

Conrad, A. and Meyer, J. 1958. The economics of slavery in the Antebellum South. *Journal of Political Economy* 66, 95–130.

Costa, D. 2003. *Health and Labor Force Participation over the Life Cycle: Evidence from the Past*. Chicago: University of Chicago Press.

Fishlow, A. 1965. *American Railroads and the Transformation of the Ante-Bellum Economy*. Cambridge, MA: Harvard University Press.

Floud, R., Wachter, K. and Annabel, G. 1990. *Height, Health, and History: Nutritional Status in the United Kingdom, 1750–1980*. Cambridge: Cambridge University Press.

Gallman, R. 1970. Self-sufficiency in the cotton economy of the Antebellum South. In *The Structure of the Cotton Economy of the Antebellum South*, ed. W. Parker. Washington, DC: Agricultural History Society.

Goldin, C. and Rockoff, H., eds. 1992. *Strategic Factors in Nineteenth Century American History: A Volume to Honor Robert W. Fogel*. Chicago: Chicago University Press.

Komlos, J., ed. 1994. *Stature, Living Standards, and Economic Development: Essays in Anthropometric History*. Chicago: University of Chicago Press.

Parker, W., ed. 1970. *The Structure of the Cotton Economy of the Antebellum South*. Washington, DC: Agricultural History Society.

Steckel, R. 1995. Stature and the standard of living. *Journal of Economic Literature* 33, 1903–40.

Steckel, R. and Floud, R., eds. 1997. *Health and Welfare during Industrialization*. Chicago: University of Chicago Press.

How to cite this article

Engerman, Stanley L. "Fogel, Robert William (born 1926)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000285> doi:10.1057/9780230226203.0589

The New Palgrave Dictionary of Economics Online

Forbonnais, François V é ron Duverger de (1722 – 1800)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bimetallism; Forbonnais, F. V. G. de; Gournay, Marquis de; mathematics and economics; Physiocracy; Quesnay, F.

Article

French economist, industrialist and inspector of commerce, Forbonnais was born at Le Mans in 1722 and died in Paris in 1800. After initial employment in industry and trade in Nantes, his desire to obtain an official position in the government services (successful in 1756 when he was appointed general inspector of currency) inspired his career as a writer on economic and financial subjects. These all have a strong mercantilist flavour, and also display considerable antagonism to the Physiocrats. Forbonnais contributed a number of economic articles to the *Encyclopédie* and provided translations of some important writings on commerce. These include King's *The British Merchant* (1721) and Uztariz's *Theory and Practice of Commerce* (1724), the former translation according to Morellet (1821) inspired by Gournay.

Forbonnais' major works are his *Elémens du commerce* (1754) and his *Principes et observations oeconomiques* (1767). The *Elémens* has the distinction of being the first French work on economics using mathematical argument. This is his analysis of equilibrium conditions with respect to the rates of exchange between more than two countries and in situations of bimetallism where there are differences in the price ratios of gold and silver (Theocharis, 1961). The *Principes* is a polemical work in which the major part is devoted to criticism of Quesnay's *Tableau économique* and his *Encyclopédie* articles on Farmers and Corn after an elucidation of general principles. Forbonnais' criticism of Physiocratic analysis is noteworthy because it was directed at its empirical foundations. In the discussion of general principles he develops arguments on the interdependence of production and trade, the balance of trade, the balance of trade doctrine in relation to money supply and employment, the beneficial consequences of gradual price rises, and the advantages of paper credit.

Selected works

1754. *Eléments du commerce*. Leyden and Paris: Briasson.

1767. *Principes et observations oeconomiques*. Amsterdam and Paris: M.M. Rey.

Bibliography

King, C. 1721. *The British Merchant*. Translated freely from the English by F.V.D. de Forbonnais as *Le négociant anglais*, Dresden and Paris, 1753.

Morellet, 1’ Abbé de. 1821. *Mémoires de l’Abbé Morellet*. Paris: Librairie Française.

Theocharis, R.D. 1961. *Early Developments in Mathematical Economics*. London: Macmillan.

Uztariz, G. de 1724. *Theory and Practice of Commerce* [*Téorica y práctica del commercio*]. Translated freely from the Spanish by F.V.D. de Forbonnais as *Théorie et pratique du commerce et de la marine* Paris, 1753.

How to cite this article

Groenewegen, Peter. "Forbonnais, François Véron Duverger de (1722–1800)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000156>
doi:10.1057/9780230226203.0590

The New Palgrave Dictionary of Economics Online

forced saving

Björn Hansson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The forced saving doctrine proposes that an increase in the amount of money may be favourable to capital accumulation at the cost of a reduction in consumption of certain individuals, who have not saved voluntarily. A consensus emerged that new credit might lead to additional, at least temporary, investment even in a full employment situation via an increase in the price level, though Lindahl and Keynes did not consider the extra saving to be forced. However, it was generally thought unwise and unjust to rely on credit inflation as a means of increasing capital accumulation.

Keywords

Bentham, J.; Bullionist Controversy; business cycle; capital accumulation; credit cycle; forced saving; Hayek, F. A. von; inflation; Keynes, J. M.; Lindahl, E. R.; Malthus, T. R.; Mises, L. E. von; over-investment; Ricardo, D.; Robertson, D. H.; saving equals investment; Thornton, H.; voluntary saving; Wicksell, J. G. K.

Article

The doctrine of forced saving proposes that an increase in the amount of money may be favourable to capital accumulation at the cost of a reduction in consumption of certain individuals, but the latter have not saved voluntarily and they do not receive any immediate benefit. The doctrine was developed in the early 19th century by Thornton (1802) and Bentham (1804). They used the terms ‘defalcation of revenue’ and ‘forced frugality’ respectively. It was Mises who coined the term ‘forced saving’ (*erzwungenes Sparen*).
Thornton published his *Paper Credit* (1802) during the debate on the suspension of gold payments by the Bank of England in 1797; the debate concerned the possible existence of a natural tendency to keep the circulation of the Bank of England within the limits which would prevent a dangerous depreciation. An excessive issue of paper money could, according to Thornton, at least temporarily increase the price

level of commodities while the money wage and other fixed incomes stayed the same. This would not only lead to a general rise in prices but also to some increase in real capital, since the real consumption of the labourers and recipients of fixed incomes would be reduced, which was the meaning of ‘defalcation of revenue’.

Jeremy Bentham, in the manuscript ‘Institute of Political Economy’ of 1804, some of which had already been written in the years 1800 and 1801, analysed the effects of an increase of paper money in a situation where all hands were employed in the most advantageous manner. If the money in the first instance were used for productive expenditure, that is, buying inputs for producing capital goods, then it would add to real capital. In the second round the money would be exclusively used for consumption and only prices would be affected. The extra real capital was due to the ‘forced frugality’ of the possessors of fixed income which was engineered by the decrease in the value of money; it operated exactly like an indirect tax upon pecuniary income. But the effect of ‘forced frugality’ was probably quite small. It was also an unjust mechanism for increasing national wealth, and under normal circumstances voluntary sacrifices would be sufficient to augment the mass of real wealth. It is obvious in these early enquiries that the forced saving by receivers of fixed incomes came from a decrease in the amount of their real consumption, while the total amount of their money expenditures was kept the same and there was no change in the amount of hoarded funds.

During the course of the Bullionist Controversy, Malthus raised the issue in his 1811 review of Ricardo's *High Price of Bullion* (1810). Malthus proposed that if a new issue of notes came into the hands of the productive classes (described as a change in the distribution of the circulating medium), then capital accumulation would increase. The mechanism of forced saving worked via the increase in the price level, which reduced the share of the annual produce of those classes who were only buyers and not sellers. Ricardo replied, in an appendix to the fourth edition of *The High Price of Bullion* published in 1811, that Malthus's results were based upon the assumption that those who lived on fixed incomes must consume their whole income. In the case of money saving it was possible that the issue of banknotes and the ensuing inflation merely transferred saving from the receivers of fixed incomes to those who had borrowed from the banks. Thus Ricardo saw no reason why it should add anything to the productive classes.

Later, comments on forced saving are found in the works of J.S. Mill and Walras, but the doctrine became important once again when it was incorporated into the pre-Keynesian analysis of credit and business cycles. The analysis took off from Wicksell's brief mention that during a cumulative process rising prices might force people living on fixed money income to reduce their consumption, an ‘involuntary saving’ which could lead to the production of new real capital. Mises (1912) and later Hayek (1929) developed Wicksell's analysis, and forced saving was used to explain the upswing in the so-called ‘over-investment’ theories of cyclical movements. An overextension of credit, since the money rate of interest was too low, and the ensuing cumulative process led to a distortion of the vertical structure of production. Production of producers’ goods outstripped the production of consumers’ goods since means of production were transferred from the latter to the former. The increase in real capital took place because of forced saving, which worked through prices rising faster than disposable income of wage-earners and the rigidity of certain incomes. The intermediate result was the same as for voluntary saving. Consumers were forced to forgo what they used to consume so as to give the entrepreneurs, who had received the additional money, command over resources for the production of extra capital goods. However, no permanent increase of real capital was possible with the help of

inflationary credit expansion and forced saving, and the new capital built during the upswing would necessarily be destroyed during the downturn.

Dennis Robertson made a most detailed analysis of different forms of saving or ‘lacking’ in *Banking Policy and the Price Level* (1926). He introduced the term ‘automatic lacking’: an involuntary reduction in planned consumption, which came about when the price level increased because newly created money was added to the daily stream of money which competed for the daily stream of marketable goods.

Parts of the doctrine of forced saving were questioned with the publication of Keynes's *Treatise on Money* (1930) and his subsequent debate with Hayek and Robertson. Robertson had, according to Keynes, no distinct definition of voluntary saving, which was related to a confusion concerning the definition of income, and it implied a deficient view of the meaning of forced saving. Keynes defined saving as the difference between income or normal costs and expenditure on consumption, which could differ from investment since saving and investment were decisions taken by different agents, windfall profits and losses being the balancing figure between investment and saving. Forced saving or automatic lacking existed when investment exceeded saving and purchasing power was redistributed by the accompanying inflation; it was represented on the one hand by the increased amount of money which spenders had to pay for that part of consumption which they continued to enjoy, and on the other hand by the extra investment provided out of the windfall gains of the entrepreneurs. Hence Keynes did not challenge the fact that an increase in net investment took place via the redistribution of purchasing power, but it was not an involuntary act.

At the same time, Erik Lindahl presented a similar analysis in *The Rate of Interest and the Price Level* (1930). The rising prices during an upward cumulative process had to change the distribution in favour of those who had a strong incentive to save, until the total saving in the community corresponded to the value of real investment, which was primarily determined by the rate of interest. This saving was mainly voluntary, since an individual was free to consume as much as he liked and the only limit was his credit standing. Keynes had the same view in the *General Theory*: this type of saving was in complete agreement with the free will of the individual to save what he chose irrespective of what he or others might be investing, since no individual could be compelled to own the additional money (corresponding to the new bank-credit) unless he deliberately preferred to hold more money rather than some other form of wealth. Lindahl reserved forced saving for the possibility that the individual has to limit planned consumption out of income (defined as the rate of interest on the capital value of all capital goods including human capital) because he is not able to obtain credit, which might be explained by banking rules concerning the collateral for loans, i.e. it is not a perfect capital market.

Once the notions of *ex ante* and *ex post* were introduced all these problems could be solved. A fall in the money rate leads to an excess of planned and realized investment over planned saving (related to planned income), and the subsequent increase in prices would imply higher incomes *ex post* for the entrepreneurs, which is the same as Keynes's concept of windfall gains in the *Treatise*. This unexpected windfall, which could not be spent during the period, would contribute the extra necessary saving, since investment *ex post* had to be equal to saving *ex post*. Lindahl denoted this as ‘unintentional saving’ and he found ‘forced saving’ to be an inappropriate term. However, Keynes seemed to have changed his position slightly in *How to Pay for the War* (1940). The process could be successful only if wages lagged behind prices, for otherwise an unlimited inflation would take place. As such it was a method of compulsorily converting a part of workers’ earnings, which they do not plan to save voluntarily, into the

voluntary saving of the entrepreneurs. From an analytical point it was voluntary saving, but it was ‘a matter of taste’ whether this was a suitable name.

To sum up: there was a consensus that new credit might lead to an additional, at least temporary, investment even in a full employment situation via an increase in the price level. But the most recent contributions – for example, Lindahl and Keynes – did not consider the extra saving to be forced. At the same time almost all of them found it unwise and unjust to rely on credit inflation as a means of increasing capital accumulation. However, after Keynes's analysis in the *General Theory* the problem seems to have disappeared from the agenda.

See Also

- inflation

Bibliography

Bentham, J. 1804. Institute of Political Economy. In *Jeremy Bentham's Economic Writings*, vol. 3, ed. W. Stark. London: George Allen & Unwin, 1954.

Haberler, G. 1937. *Prosperity and Depression*. 5th edn, London: George Allen & Unwin, 1964.

Hayek, F. von. 1929. *Geldtheorie und Konjunkturtheorie*. Vienna and Leipzig: Hölder-Pichler-Tempsky. Trans. as *Monetary Theory and the Trade Cycle*, London: Jonathan Cape, 1933.

Hayek, F. von. 1931. *Prices and Production*. 2nd edn, London: George Routledge & Sons, 1935.

Hayek, F. von. 1932. A note on the development of the doctrine of forced saving. *Quarterly Journal of Economics* 47(November), 123–33.

Keynes, J.M. 1930. *A Treatise on Money*, vol. 1. In *The Collected Writings of John Maynard Keynes*, vol. 5, London: Macmillan, 1971.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. In *The Collected Writings of John Maynard Keynes*, vol. 8, London: Macmillan, 1973.

Keynes, J.M. 1940. *How to Pay for the War*. In *The Collected Writings of John Maynard Keynes*, vol. 22, London: Macmillan, 1978.

Lindahl, E. 1930. *Penningpolitikens medel*. Lund: C.W.K. Gleerup. Trans. as ‘The rate of interest and the price level’ in E. Lindahl, *Studies in the Theory of Money and Capital*, London: George Allen & Unwin, 1939.

Machlup, F. 1943. Forced or induced saving: an exploration into its synonyms and homonyms. *Review of Economics and Statistics* 25, 26–39.

Malthus, T.R. 1811. Review of Ricardo's *High Price of Bullion*. *Edinburgh Review*, February.

Mill, J.S. 1844. *Essays on Some Unsettled Questions of Political Economy*. In *Collected Works of John Stuart Mill*, vol. 4, London: Routledge & Kegan Paul, 1967.

Mises, L. von. 1912. *Theorie des Geldes undder Umlaufsmittel*. Munich: Dunker & Humblot. 2nd edn, 1924. Trans. as *The Theory of Money and Credit*, London: Jonathan Cape, 1934.

Ricardo, D. 1810. The High price of Bullion. A proof of the depreciation of bank notes. In *The Works and Correspondence of David Ricardo*, vol. 3, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Robertson, D. 1926. *Banking Policy and the Price Level*. London: P.S. King & Sons.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. Reprinted, London: George Allen & Unwin, 1939.

Walras, L. 1879. Théorie mathématique du billet de banque. Reprinted in *Etudes d'économie politique appliquée*, Lausanne and Paris, 1898.

Wicksell, K. 1935. *Lectures on Political Economy*. vol. 2, New York: Kelley. Trans. from the 3rd Swedish edn of *Föreläsningar i nationalekonomi*, vol. 2. London: George Routledge & Sons, 1929.

How to cite this article

Hansson, Björn. "forced saving." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000159> doi:10.1057/9780230226203.0592

The New Palgrave Dictionary of Economics Online

forecasting

Clive Granger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Providing timely and useful forecasts is among the most relevant tasks of economists. The choice among the many techniques and approaches depends on the variables being forecast and the length of the forecast horizon. Providing confidence intervals around the point forecasts is becoming standard practice, as are sophisticated attempts at evaluating the quality of the forecasts and the intervals. Forecasts are often combined, raising questions about the appropriate cost functions to use in the evaluation process. Economists once concentrated on forecasting the mean of a process, then moved to variance, and now consider quantities and the whole distribution.

Keywords

Akaike information criterion; ARCH models; ARMA models; Bayes information criterion; copulas; error-correction models; forecasting; Kalman filters; leading indicators; linear models; neural networks; quantiles; switching models; time series analysis; vector autoregressions

Article

Decisions in the fields of economics and management have to be made in the context of forecasts about the future state of the economy or market. As decisions are so important as a basis for these fields, a great deal of attention has been paid to the question of how best to forecast variables and occurrences of interest. There are several distinct types of forecasting situations, including event timing, event outcome, and time-series forecasts. Event timing is concerned with the question of when, if ever, some specific event will occur, such as the introduction of a new tax law, or of a new product by a competitor, or of a turning point in the business cycle. Forecasting of such events is usually attempted by the use of leading indicators, that is, other events that generally precede the one of interest. Event outcome forecasts try to forecast the outcome of some uncertain event that is fairly sure to occur, such as finding the winner of an election or the level of success of a planned marketing campaign. Forecasts are usually based on data

specifically gathered for this purpose, such as a poll of likely voters or of potential consumers. There clearly should be a positive relationship between the amount spent on gathering the extra data and the quality of the forecast achieved.

A time series x_t is a sequence of values gathered at regular intervals of time, such as daily stock market closing prices, interest rates observed weekly, or monthly unemployment levels. Irregularly recorded data, or continuous time sequences may also be considered but are of less practical importance. When at time n (now), a future value of the series, x_{n+h}, is a random variable where h is the forecast horizon. It is usual to ask questions about the conditional distribution of x_{n+h} given some information set I_n, available now from which forecasts will be constructed. Of particular importance are the conditional mean

$$f_{n,h} = E[x_{n+h}|I_n]$$

and variance, $V_{n,h}$. The value of $f_{n,h}$ is a point forecast and represents essentially the best forecast of the most likely value to be taken by the variable x at time $n+h$. With a normality assumption, the conditional mean and variance can be used together to determine an interval forecast, such as an interval within which $x_{n,h}$ is expected to fall with 95 per cent confidence. An important decision in any forecasting exercise is the choice of the information set I_n. It is generally recommended that I_n include at least the past and present of the individual series being forecast, $x_{n-j}, j \geq 0$. Such information sets are called *proper*, and any forecasting models based upon them can be evaluated over the past. An I_n that consists just of x_{n-j}, provides a univariate set so that future x_i are forecast just from its own past. Many simple time-series forecasting methods are based on this information set and have proved to be successful. If I_n includes several explanatory variables, one has a multivariate set. The choice of how much past data to use and which explanatory variables to include is partially a personal one, depending on one's knowledge of the series being forecast, one's levels of belief about the correctness of any economic theory that is available, and on data availability. In general terms, the more useful are the explanatory variables that are included in I_n, the better the forecast that will result. However, having many series allows for a confusing number of alternative model specifications that are possible so that using too much data could quickly lead to diminishing marginal returns in terms of forecast quality. In practice, the data to be used in I_n will often be partly determined by the length of the forecast horizon. If h is small, a short-run forecast is being made and this may concentrate on frequently varying explanatory variables. Short-term forecasts of savings may be based on interest rates, for example. If h is large so that long-run forecasts are required, then slowly changing, trending explanatory variables may be of particular relevance. A long-run forecast of electricity demand might be largely based on population trends, for example. What is considered short run or long run will usually depend on the properties of the series being forecast. For very long forecasts, allowances would have to be made for technological change as well as changes in demographics and the economy. A survey of the special and separate field of technological forecasting can be found in Martino (1993) with further discussion in Martino (2003).

If decisions are based on forecasts, it follows that an imperfect forecast will result in a cost to the

decision-maker. For example, if $f_{n,h}$ is a point forecast made at time n, of x_{n+h}, the eventual forecast error will be

$$e_{n,h} = x_{n,h} - f_{n,h},$$

which is observed at time $n+h$. The cost of making an error e might be denoted as $C(e)$, where $C(e)$ is positive with $C(0)=0$. As there appears to be little prospect of making error-free forecasts in economics, positive costs must be expected, and the quality of a forecast procedure can be measured as the expected or average cost resulting from its use. Several alternative forecasting procedures can be compared by their expected costs and the best one chosen. It is also possible to compare classes of forecasting models, such as all linear models based on a specific, finite information set, and to select the optimum model by minimizing the expected cost. In practice the true form of the cost function is not known for decision sequences, and in the univariate forecasting case a pragmatically useful substitute to the real $C(e)$ is to assume that it is well approximated by ae^2 for some positive a. This enables least-squares statistical techniques to be used when a model is estimated and is the basis of a number of theoretical results including that the optimal forecast of x_{n+h} based on I_n is just the conditional mean of $x_{n,h}$. Machina and Granger (2006) have considered cost functions generated by decision makers and then find implications for their utility functions. This is just one component of considerable developments in the area of evaluation of forecasts; see West (2006) and Timmermann (2006), for example.

When using linear models and a least-square criterion, it is easy to form forecasts under an assumption that the model being used is a plausible generating mechanism for the series of interest. Suppose that a simple model of the form

$$x_t = \alpha x_{t-1} + \beta y_{t-2} + \varepsilon_t$$

is believed to be adequate where ε_t is a zero-mean, white noise (unforecastable) series. When at time n, according to this model, the next value of x will be generated by

$$x_{n+1} = \alpha x_n + \beta y_{n-1} + \varepsilon_{n+1}.$$

The first two terms are known at time n, and the last term is unforecastable. Thus

$$f_{n,1} = \alpha x_n + \beta y_{n-1}$$

and

$$e_{n,1} = \varepsilon_{n+1}.$$

x_{n+2}, the following x, will be generated by

$$x_{n+2} = \alpha x_{n+1} + \beta y_n + \varepsilon_{n+2}.$$

The first of these terms is not known at time n, but a forecast is available for it, αf_n; the second term is known at time n, and the third term is not forecastable, so that

$$f_{n,2} = \alpha f_{n,1} + \beta y_n$$

and

$$e_{n,2} = \varepsilon_{n+2} + \alpha(x_{n+1} - f_{n,1}) = \varepsilon_{n+2} + \alpha\varepsilon_{n+1}.$$

To continue this process for longer forecast horizons, it is clear that forecasts will be required for y_{n+h-2}. The forecast formation rule is that one uses the model available as though it is true, asks how a future x_{n+h} will be generated, uses all known terms as they occur, and replaces all other terms by optimal forecasts. For non-linear models this rule can still be used, but with the additional complication that the optimum forecast of a function of x is not the same function of the optimum forecast of x.

The steps involved in forming a forecast include deciding exactly what is to be forecast, the forecast horizon, the data that is available for use, the model forms or techniques to be considered, the cost function to be used in the evaluation procedure, and whether just one single forecast would be produced or several alternatives. It is good practice to decide on the evaluation to be used before starting a sequence of forecasts. If there are several alternative forecasting methods involved, a weighted

combination of the available forecasts is both helpful for evaluation and can often provide a superior forecast.

The central problem in practical forecasting is choosing the model from which the forecasts will be derived. If a univariate information set is used, it is natural to consider the model developed in the field of time-series analysis. A class of models that has proved to be successful in short-term forecasting is the autoregressive (AR) model class. If a series is regressed on itself up to *p* lags, the result is an AR(p) model. These models were originally influenced by Box and Jenkins (1970) as a particularly relevant subclass of their ARMA (p, q) models, which involve moving average components. The number of lags in an AR(p) can be chosen using a selection criterion; the most used are the Bayes information criterion (BIC) and the less conservative Akaike information criterion (AIC).

The natural extension was to vector autoregressive models. Later, when it was realized that many series in macroeconomics and finance had the property of being integrated, and so contained stochastic trends, the natural multivariate form was the error-correction model. It is quite often found that error-correction models improve forecasts, but not inevitably. There are a variety of ways of building models with many predictive variables, including those with unobserved components and using special data, such as survey expectations, real-time macro data, and seasonal components.

In recent years the linear models have been joined by a variety of nonlinear forms (see Terasvita, 2006), including switching models and neural networks as well as linear models with time varying coefficients estimated using Kalman filters.

Traditionally, forecasters concentrated on the mean of the predictive distribution. Towards the end of the 20th century considerable attention was given to forecasting the variance of the distribution, particularly in the financial area, often using Engle's (1995) ARCH model or one of its many generalizations (see the survey by Andersen et al., 2006). Recently forecasts of the whole distribution have become more common in practice, both in finance and in macroeconomics: see Corradi and Swanson (2006) for a recent discussion. These forecasts will include discussions of quantiles, and the use of copulas gives a way into multivariate distribution forecasts. The topics mentioned in this paragraph are covered by chapters in Elliot, Granger and Timmermann (2006).

Bibliography

Andersen, T., Bollerslev, T., Christoffersen, P. and Diebold, F. 2006. Volatility and correlation forecasting. In Elliot, Granger and Timmermann (2006).

Box, G. and Jenkins, G. 1970. *Time Series Analysis, Forecasting and Control*. San Francisco: Holden Day.

Clements, M. and Hendry, D. 2003. *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Corradi, V. and Swanson, N.R. 2006. Predictive density evaluation. In Elliot, Granger and Timmermann 2006).

Elliott, G., Granger, C.W.J. and Timmermann, A., eds. 2006. *Handbook of Economic Forecasting*. Amsterdam: North-Holland.

Engle, R.F. 1995. *ARCH: Selected Readings*. Oxford: Oxford University Press.

Granger, C.W.J. and Newbold, P. 1987. *Forecasting Economic Time Series*. 2nd edn. New York: Academic Press.

Machina, M. and Granger, C.W.J. 2006. Forecasting and decision theory. In Elliot, Granger and Timmermann (2006).

Martino, J.P. 1993. *Technological Forecasting for Decision Making*. 2nd edition. New York: McGraw Hill.

Martino, J.P. 2003. A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change* 70, 719–33.

Terasvita, T. 2006. Forecasting economic variables with nonlinear models. In Elliott, Granger and Timmermann (2006).

Timmermann, A. 2006. Forecast combinations. In Elliott, Granger and Timmermann (2006).

West, K. 2006. Forecast evaluation. In Elliott, Granger and Timmermann (2006).

How to cite this article

Granger, Clive. "forecasting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000161> doi:10.1057/9780230226203.0591

The New Palgrave Dictionary of Economics Online

foreign aid

Tarp Finn
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Foreign aid has evolved significantly since the Second World War in response to a dramatically changing global political and economic context. This article (*a*) reviews this process and associated trends in the volume and distribution of foreign aid; (*b*) reviews the goals, principles and institutions of the aid system; and (*c*) discusses whether aid has been effective. While much of the original optimism about the impact of foreign aid needed modification, there is solid evidence that aid has indeed helped further growth and poverty reduction.

Keywords

Bretton Woods agreement; charity; Chenery, H.; corruption; dependency; Development Assistance Committee (DAC); Dutch disease; economic development; foreign aid; foreign direct investment; Harrod-Domar model; International Development Association (IDA); international public goods; Krueger, A.; Marshall Plan; micro-macro paradox; Millennium Challenge Account (MCA); Millennium Development Goals (MDG); non-governmental organizations (NGOs); Official Development Assistance (ODA); poverty; rent-seeking; Washington consensus; World Bank

Article

Foreign aid and its usefulness in promoting economic development in developing countries has been a topic of intense controversy ever since Rosenstein-Rodan (1943) advocated aid to eastern and south-eastern Europe. Early optimism and confidence in the impact of foreign aid have been tempered with time, but aid continues to loom large in the public discourse; and aid remains squarely on most policy agendas concerned with poverty and inequality in Africa and elsewhere in the developing world.
What is foreign aid? Loosely, it covers governmental transfers to poor countries that are mainly destined for developmental purposes. For a more precise definition it is useful to turn to the Development Assistance Committee (DAC) of the OECD. DAC is the principal body through which the OECD deals

with issues related to cooperation with developing countries, and DAC publishes the most comprehensive data available on foreign aid (OECD, 2004). DAC countries also account for almost 95 per cent of all aid flows. In 2002 the total amount of foreign aid disbursed by donors to developing countries and multilateral organizations reached $61.5 billion (Table 1). Multilateral organizations disbursed some 30 per cent (Table 2), and Table 3 shows that international development assistance is an important resource for many developing countries.

Net ODA disbursements by donor, 1960–2002

	2002 prices ($ billion)				Per cent of total			
	ODA per capita (2002 prices, $)				Per cent of donor GNI			
	1960–73	1992	1998	2002	1960–73	1992	1998	2002
United States	14.7	14.1	9.4	13.3	47.1	23.0	18.3	21.6
74.9	55.3	34.8	46.1	0.4	0.2	0.1	0.1	
Japan	2.5	10.5	10.4	9.3	8.0	17.1	20.2	15.1
24.5	84.4	82.2	72.8	0.2	0.3	0.3	0.2	
France	3.9	7.2	5.1	5.5	12.8	11.8	9.9	8.9
80.6	126.2	87.3	92.3	0.8	0.6	0.4	0.4	
Germany	2.8	6.6	4.9	5.3	9.1	10.7	9.5	8.7
48.0	81.4	59.5	64.5	0.4	0.4	0.3	0.3	
United Kingdom	3.2	3.6	3.8	4.9	10.2	5.8	7.4	8.0
58.0	61.3	64.7	83.5	0.5	0.3	0.3	0.3	
DK, NL, NO and SE	1.3	7.1	7.4	8.7	4.2	11.5	14.4	14.1
44.6	211.9	217.0	248.2	0.3	1.0	0.8	0.9	
Other DAC	2.6	10.9	9.4	11.3	8.5	17.8	18.2	18.4
23.0	57.9	46.0	53.6	0.3	0.4	0.3	0.3	
Non-DAC		1.4	1.0	3.2		2.2	2.0	5.2
			67.2		0.3	0.1	0.4	
Total	31.0	61.3	51.5	61.5	100	100	100	100
51.6	76.9	61.7	67.6	0.4	0.3	0.2	0.2	
Bilateral ODA	26.5	41.9	34.9	43.5	85.5	68.3	67.8	70.7
Multilateral ODA	4.9	19.1	16.6	18.0	15.6	31.1	32.2	29.3

Notes: Denmark (DK), the Netherlands (NL), Norway (NO) and Sweden (SE) reached the UN ODA target of 0.7% of GNI in respectively 1978, 1975, 1976 and 1975. Luxembourg reached the target in 2000.

Source: OECD (2004).

Multilateral aid disbursements, 1960–2002

	2002 prices, $ billion				Per cent			
	1960–73	1992	1998	2002	1960–73	1992	1998	2002
Multilateral, total *of which:*	2.8	16.3	14.4	17.0	100	100	100	100
United Nations	0.9	5.3	2.6	3.8	31.4	32.6	17.9	22.1
IMF and WB	0.8	5.3	5.0	6.0	30.0	32.7	35.0	35.1
European Commission	0.6	3.8	4.6	5.1	23.1	23.1	32.3	30.3
Regional Development Banks	0.4	1.6	1.9	1.8	15.4	10.0	13.2	10.5
Other multilateral institutions	0.0	0.3	0.2	0.4	0.0	1.6	1.6	2.1

Source: OECD (2004).

ODA by recipient, 1960–2002

	GNI in 2002 (US $ billion)	GNI per capita (2002, US$)	Total ODA receipts (2002 prices–US$ billion)				In per cent of total flows (ODA+OOF+private)			
			ODA per capita (2002 prices, US$)				In per cent of GNI			
			1960–73	1992	1998	2002	1960–73	1992	1998	2002
Developing countries, total			30.2	58.3	49.3	60.5	74.2	55.3	26.6	88.2
Least developed, countries, total			4.1	16.3	12.2	17.8	88.1	96.8	82.9	116.2
Other low-income countries, total			10.7	10.8	10.2	12.3	89.8	63.7	59.7	86.4
Low-middle-income countries, total			6.7	16.9	13.6	16.1	75.0	69.6	31.4	96.3
China	1251.1	977.1		2.8	2.4	1.5		49.2	31.6	–61.9
			2.4	1.9	1.2		0.7	0.3	0.1	
Mexico	636.1	6309.3	0.2	0.3	0.0	0.1	23.1	4.4	0.5	2.3
		3.9	3.2	0.4	1.3	0.2	0.1	0.0	0.0	
India	506.2	482.7	4.5	2.3	1.6	1.5	98.5	78.4	57.8	5359.6
		9.0	2.6	1.6	1.4	1.9	1.0	0.4	0.3	
Brazil	443.0	2538.9	0.9	–0.3	0.3	0.3	69.8	–14.9	1.4	12.3
		10.5	–2.2	1.9	1.9	0.7	–0.1	0.0	0.1	
Indonesia	164.6	777.2	1.3	1.8	1.2	1.3	87.3	33.0	18.8	8185.7
		11.8	10.0	6.1	6.2	4.4	1.6	1.4	0.8	
Israel	100.9	15365.4	0.5	2.4	1.2	0.8	62.2	64.0	31.3	139.7

		193.4	475.2	193.9	115.3	2.4	3.2	1.1	0.8	
Egypt	90.0	1355.3	0.7	3.7	1.9	1.2	79.6	217.3	47.5	63.2
		24.0	68.4	31.3	18.7	2.8	8.7	2.3	1.4	
Malaysia	88.4	3639.1	0.1	0.2	0.2	0.1	59.6	16.7	–25.0	2.5
		14.0	10.1	9.1	3.5	0.7	0.4	0.3	0.1	
Philippines	83.1	1039.9	0.4	1.7	0.6	0.6	67.2	115.8	14.8	22.8
		12.9	26.0	8.1	6.9	1.0	3.2	0.9	0.7	
Colombia	77.8	1778.5	0.4	0.2	0.2	0.4	73.6	288.3	6.8	–19.0
		18.4	6.0	4.1	10.1	1.4	0.5	0.2	0.6	
Pakistan	59.8	412.4	1.8	1.0	1.0	2.1	97.1	58.1	62.4	114.0
		33.5	8.6	7.8	14.8	4.5	2.1	1.7	3.6	
Bangladesh	49.7	366.6	0.7	1.8	1.1	0.9	98.8	93.0	87.7	102.0
		10.5	15.5	8.8	6.7	5.2	5.6	2.5	1.8	
Nigeria	36.9	278.0	0.3	0.3	0.2	0.3	59.2	250.9	61.1	6.6
		6.3	2.5	1.6	2.4	0.9	0.9	0.7	0.9	
Guatemala	23.0	1915.9	0.1	0.2	0.2	0.2	99.8	155.8	30.0	94.8
		14.5	22.8	20.9	20.7	1.1	1.9	1.2	1.1	
Sri Lanka	16.3	858.4	0.2	0.6	0.4	0.3	86.5	91.5	81.3	83.0
		13.1	37.5	23.2	18.1	1.6	6.7	2.7	2.1	
Kenya	12.2	389.5	0.3	0.8	0.4	0.4	83.3	92.1	90.0	101.4
		32.0	34.0	13.8	12.6	4.8	11.6	3.7	3.2	
Tanzania	9.3	265.0	0.2	1.3	1.0	1.2	90.8	104.0	100.5	121.0
		18.2	46.5	29.8	35.0		30.3	12.1	13.2	
Bolivia	7.6	862.3	0.2	0.7	0.6	0.7	88.8	81.3	79.4	208.6
		39.5	94.5	75.4	77.3	2.6	12.3	7.5	9.0	
Ghana	6.0	302.9	0.2	0.6	0.7	0.6	102.0	82.2	99.4	105.1
		20.9	36.6	37.9	32.6	1.9	9.8	9.6	10.8	
Ethiopia	6.0	89.6	0.1	1.1	0.6	1.3	85.2	100.7	83.4	119.5
		5.3	20.2	10.3	19.4		11.8	10.2	21.7	
Senegal	4.9	484.9	0.2	0.6	0.5	0.4	96.5	93.5	90.2	82.3
		43.7	79.8	51.6	44.5	5.7	11.4	10.9	9.2	
Mali	3.1	272.8	0.1	0.4	0.3	0.5	99.7	99.6	85.1	141.5
		18.2	45.6	31.9	41.0	7.8	15.2	13.6	15.0	
Other			17.3	33.9	33.9	44.3	72.5	53.1	28.2	99.6

Notes: OOF: Other official flows. For Israel, 1998 and 2002 are OA (official aid) flows, not ODA. Average ODA per capita is for Bangladesh (1971–73). Average ODA in percent of GNI is for Bangladesh (1973); Bolivia (1970–73); Indonesia 1967–73); Mali (196–73); Pakistan (1967–73); Senegal (1968–73).

Source: OECD (2004).

The term ‘foreign aid’ or ‘development assistance’ refers to financial flows that qualify as Official Development Assistance (ODA). ODA is defined as grants and loans to aid recipients that are: (*a*) undertaken by the official sector of the donor country, (*b*) with promotion of economic development and welfare as the main objective, (c) at concessional financial terms, where the grant element is equal to at least 25 per cent.

Conventionally the market rate of interest used to assess a loan is taken as ten per cent. Thus, while the grant element is nil for a loan carrying an interest rate of ten per cent, it is 100 per cent for a pure grant, and lies between these two limits for a soft loan. In addition to financial flows, technical cooperation costs are included in ODA; but grants, loans and credits for military purposes are excluded, and transfer payments to private individuals are in general not counted. The same goes for private charity, hard loans and foreign direct investment (FDI).

While the OECD operates with a consolidated list of recipient countries to capture all aid flows, this list is divided into two parts. Only aid to ‘traditional’ developing countries counts as ODA. For these (Part I) countries there is a long-standing United Nations (UN) target that they should receive 0.7 per cent of donors’ gross national income (GNI) as aid. Assistance to the ‘more advanced’ eastern European and developing (Part II) countries is recorded separately as ‘official aid’ (OA), which is not included as part of ODA.

Historical background

Foreign aid emerged out of the disruption that followed the Second World War. The international economic system had collapsed, and war-ravaged Europe faced a critical shortage of capital and an acute need for physical reconstruction. The response was the European Recovery Programme, commonly known as the Marshall Plan. During the peak years the United States devoted some two or three per cent of its national income to helping restore Europe. This objective was achieved on schedule, and fuelled optimistic expectations about the future effectiveness of foreign aid.

After the success of the Marshall Plan, the attention of industrialized nations turned to the developing countries, many of which became independent around 1960. Economic growth in a state-led planning tradition became a key objective during the 1950s and 1960s, and it was widely believed that poverty and inequality would eventually be eliminated through growth and modernization (‘trickle down’). A major part of the rapidly increasing bilateral flows during the 1950s came from the United States, but colonial ties remained strong, and developing regions continued to receive bilateral (country-to-country) support from the former colonial powers, notably France and the United Kingdom. Yet the 1960s was also the decade when a range of new bilateral donor agencies was established in, for example, the Nordic countries. They accounted for much of the increase in aid flows in the 1970s.

A transition toward more independent, multilateral relations began to emerge during the 1960s. Hjertholm and White (2000) argue that this created, a constituency for foreign aid, and the non-aligned

movement of developing countries gave a focus to this voice, as did the various organs of the UN, which accounted for around one-third of multilateral assistance during 1960–73. The International Bank for Reconstruction and Development (IBRD, or World Bank), established at the Bretton Woods Conference in 1944, is central in multilateral development assistance, especially following the creation of the International Development Association (IDA) in 1960. IDA channels resources to the poorest countries on ‘soft’ conditions alongside the regional development banks, formed from 1959 to 1966, and the European Commission.

The original Marshall Plan was built around support to finance general categories of imports and strengthen the balance of payments (that is, programme aid), but from the early 1950s project aid became the dominating aid modality. Some donors continued to supply programme aid (including food aid), but aid was increasingly disbursed for the implementation of specific capital investment projects and associated technical assistance to support advances in infrastructure and productive sectors.

The multilateralism of aid became somewhat more pronounced after the mid-1970s, when the UN, the World Bank and other multilateral agencies expanded their activities quite considerably; since then the share of multilateral aid in total aid has remained close to 30 per cent. The 1970s also saw an increased focus on employment, income distribution, and poverty alleviation as essential objectives of development and foreign aid. The effectiveness of trickle-down was widely questioned, and new strategies referred to as ‘basic human needs’ and ‘redistribution with growth’ were formulated. Nevertheless, the typical project aid modality remained largely unchanged.

During the 1960s and 1970s, economic progress was visible in much of the developing world. This era came to an abrupt end at the beginning of the 1980s. The international debt crisis erupted in association with macroeconomic imbalances in many countries, and it soon became evident that the downturn would be long-lasting, not temporary as in 1973. On the political scene Ronald Reagan and Margaret Thatcher came to power in the USA and UK, and at the World Bank Anne Krueger became Vice-President and Chief Economist, replacing Hollis Chenery. This change was symbolic and substantive (Kanbur, 2003).

Economic circumstances in the developing countries and the relations between the North and South had changed radically. The crisis hit hard, especially in many African countries; progress over previous decades ground to a halt, inflation got out of control and the deficit in the balance payments could not be financed on a sustainable basis. Focus in development policy shifted to internal domestic failures, and achieving macroeconomic balance (externally and internally) became widely perceived as an essential prerequisite for renewed development.

‘Rolling back the state’ turned into a rallying call in the reform efforts, and reliance on market forces, outward orientation, and the role of the private sector, including non-governmental organizations (NGOs), was emphasized by the World Bank and others. In parallel, poverty alleviation somehow slipped out of view in mainstream agendas for economic reform, but remained at the centre of attention in more unorthodox thinking, such as the ‘adjustment with a human face’ approach of the UN Children's Fund (Cornia, Jolly and Stewart, 1987).

At the same time, bilateral donors and international agencies grappled with how to channel resources to the developing world. Channelling fresh resources to developing countries in the form of discrete investment projects had become increasingly difficult. Project rates of return did not seem to justify the investments. Various kinds of quick-disbursing macroeconomic programme assistance, such as balance of payments support and sector budget support, which were not tied to investment projects and which could be justified under the headings of stabilization and adjustment, appeared to be an ideal solution to

this problem. Financial programme aid and adjustment loans (and eventually debt relief) became fashionable and policy conditionality more widespread. A rationale had been found for maintaining the flow of resources, which corresponded well to the orthodox guidelines for good policy summarized by the 'Washington *consensus*' (Williamson, 1997).

Meanwhile, total aid continued to grow steadily in real terms until the early 1990s, and more than tripled as a share of the growing national income of the donor community during 1970–90. After 1992, total aid flows started to decline in absolute terms (especially in the USA). Many reasons account for the fall in aggregate flows after 1992, including the decline of communism and the end of the cold war. Weakening patron–client relationships among the developing countries and the former colonial powers also played a role, and the traditional support for foreign aid by vocal interest groups in the industrial countries receded. Bilateral and multilateral aid institutions were subjected to criticism, and at times characterized as blunt instruments of commercial interests in the industrial world or as self-interested, rent-seeking bureaucracies. Moreover, acute awareness in donor countries of cases of bad governance, corruption, and 'crony capitalism' led to scepticism about the credibility of governments receiving aid. Aid fatigue became widespread during the second half of the 1990s.

Aid allocation

Foreign aid has over the years been justified in public policy pronouncements in widely differing ways, ranging from pure altruism to the shared benefits of economic development in poor countries and to the political ideology, foreign policy and commercial interests of the donor country. Few dispute that humanitarian sentiments have motivated donors. Action following severe natural calamities, which continue to be endemic in poor countries, is an example. Food and emergency relief also remains an important form of aid. In addition, the data available in Table 3 suggest that donors allocate relatively more ODA to the poorest countries. The broader validity of this casual observation is confirmed in cross-country econometric work (Alesina and Dollar, 2000). While studying bilateral aid only, they conclude that most donors give more aid to poorer countries, ceteris paribus. They stress as well that there is considerable variation among donors.

Emphasis on the needs of poor countries was a prominent characteristic – and the underlying economic rationale – in much of the policy literature on foreign aid in the 1950s and 1960s. Here the focus was on estimating aid requirements in the tradition of the two-gap model (Chenery and Strout, 1966). With time, development concerns have broadened. The two-gap model has become somewhat unfashionable, at least in academia, and the role of aid has changed to a much more multi-dimensional set of concerns (Thorbecke, 2000). Nevertheless, economic development in aid-receiving countries continues as a yardstick both in its own right (at least for some donors) and as a necessary condition for the realization of other development aims.

A second observation from Table 2 is that large, populous countries, such as China and India, receive relatively small amounts of aid in per capita terms. Smaller countries such as Mali, Ghana, Bolivia and Sri Lanka are given more favourable per capita treatment. This finding is confirmed econometrically by Alesina and Dollar (2000). They stress, however, the critical and complex importance of political and strategic considerations in aid allocations.

It is not news that selfish motives are critical in donor decisions. In the past, the cold war was used as a

powerful justification for providing aid to developing countries to stem the spread of communism. Similarly, aid from socialist governments was motivated to promote socialist political and economic systems. Other strategic interests play a role as well. The USA has over the years earmarked very substantial amounts of aid to Egypt and Israel; being a former colony is an important determinant in getting access to French aid; and voting behaviour in the UN can affect aid allocation both bilaterally (Alesina and Dollar, 2000) and through the multilateral system (Andersen, Harr and Tarp, 2004).

In sum, there is often a wide gap between donor rhetoric and practice when attention is on the size and allocation of foreign aid. This gap is illustrated by the fact that the donor countries are indeed very far from contributing the 0.7 per cent of their national income as ODA, which was agreed as a UN target in 1970. As shown in Table 1, only the group of Nordic countries and the Netherlands have consistently met this target, while the USA contributed around 0.1 per cent of the US GNI in 2002. Finally, Table 3 shows that total ODA, ODA per capita, ODA as a share of GNI and ODA as a share of total flows actually vary considerably in real terms in many aid-receiving countries. Economic management in general, and management of aid inflows in particular, are not easy tasks in developing countries.

The impact of foreign aid

If the economic development rationale for foreign aid is taken seriously, it is of interest to ask whether aid-receiving countries benefit from such transfers and, if so, how. What are the mechanisms through which aid works, and what are the potential negative effects associated with foreign aid? Over the past 60 years a vast amount of empirical work has (*a*) studied the impact of aid at micro-, meso- and macroeconomic level; (*b*) relied on cross-country as well as single-country data; and (*c*) included broad surveys of a qualitative and interdisciplinary nature as well as more strict quantitative econometric work. Many surveys are available; see for example Cassen (1987) and Tarp (2000).

An influential literature focused on cross-country econometric approaches to the analysis of aid effectiveness. This literature has gone through three generations (Hansen and Tarp, 2000); and from the early 1990s macro-econometric studies came to dominate the academic and public discourse. This work was motivated in part by the availability of much better data across a range of countries and in part by insights emerging from new growth theory and the rapidly increasing number of general empirical studies of growth.

The simple Harrod–Domar model (and the two-gap Chenery–Strout extension) was used extensively in the past as the analytical framework of choice for assessing aid impact. The underlying idea was simple. Assume physical capital is the only factor of production (so investment is the key constraint on growth), and assume as well that all aid is invested. Then it is straightforward to calculate the growth impact of additional aid. If aid corresponds to six per cent of the gross national product and the capital–output ratio is estimated at 3.0, then aid adds 2.0 percentage points a year to the growth rate. The impact of aid is clearly positive, and aid works by helping to fill a savings or a foreign exchange gap.

The Achilles heel in this type of calculation is, first, that it is a tall order to expect that all aid is invested. Aid is provided for many reasons. In addition, the share of aid that ends up being invested (rather than consumed) will, in even the best of circumstances, depend on the degree of fungibility of the foreign aid transfer. Yet, even if aid adds to domestic savings and investment on less than a one-to-one basis, aid does continue to have a positive impact on growth in the traditional line of thinking – as long as total

savings and investment go up.

A second line of critique of the Harrod–Domar and two-gap approach has been the argument that growth is less related to physical capital investment (including aid) than often assumed (Easterly, 2001). If the key driver of the productive impact of aid is related more to incentives and relative prices and more generally to the policy environment, then it becomes important to consider potentially distortionary effects of aid on incentives and economic policies in the aid-receiving system, and vice versa. An example is ‘Dutch disease’, and domestic demand and resource allocation can certainly be twisted in undesirable directions following an aid inflow. One concrete example is that aid donors often pay much higher wages to national experts and staff than equally important national institutions. Another illustration is change in the structure of domestic demand following the aid inflow.

Third, a large and growing literature on the political economy of aid (see Kanbur, 2003, and Gunning, 2005, for references) has argued that, if aid allows a recipient government (local elites) to pursue behaviour that is in any way anti-developmental, then the potential positive impact of aid can be undermined. There are many such examples available in practice ranging from outright misuse of aid to more subtle issues such as the potential negative impact of aid on domestic taxation (Adam and O’Connell, 1999).

The fear that foreign aid can generate undesirable aid dependency relationships persisted throughout the 1990s and into the 21st century, and gradually the perception that policy conditionality was failing to promote policy reform started to assert itself (Kanbur, 2000, and Svensson, 2003). This perception prompted a keen interest in new kinds of donor–recipient relationships. One outcome was calls for increased national ownership of aid programmes. Another was that World Bank and independent academic researchers started digging into the aid–growth relationship using modern analytical techniques.

Much of the recent debate has roots in Mosley's (1987) micro–macro paradox. He suggested that, while aid seems to be effective at the microeconomic level, identifying any positive impact of aid at the macroeconomic level is harder or even impossible. Along with the implementation of adjustment programmes during the 1980s, traditional evaluation methods such as calculating the internal rate of return of projects came under severe criticism. The perception spread that aid channelled through sovereign governments is fully fungible. The internal rate of return approach also became problematic as donors started to embrace wider social goals for aid. The wave of cross-country work during the 1990s and the later, more extensive use of randomized programme evaluation (Duflo, 2004) are ways of trying to come to grips with these issues.

The cross-country analysis by Boone (1996) suggested that aid does not work at all and is simply a waste of resources. This was followed up with an analysis by Burnside and Dollar (1997; 2000). They argue that some aid does work, and provided an attractive solution to the micro–macro paradox. Aid works, but only in countries with ‘good policy’. They based this conclusion on an aid-policy interaction term that emerged as statistically significant in their analysis of the relationship between aid and growth. Burnside and Dollar, and more recently Collier and Dollar (2001; 2002), have used the foregoing framework as a basis for suggesting that aid should be directed to ‘good policy’ countries to improve aid's impact on poverty alleviation. This recommendation is partly justified by reference to the seeming inability of aid to change policy, a finding that has emerged from other Bank-funded research (Devarajan, Dollar and Holmgren, 2001). While the Bank's Monterrey document (World Bank, 2002)

toned down these recommendations, the basic thrust in much of the international aid debate remains that macroeconomic performance evaluation and policy criteria (established by the World Bank) should play a key role in aid allocation.

The work of Burnside, Collier, and Dollar led to discussions about what constitutes good policy. In many ways these discussions are extensions of more general debates and views about development strategy and policy, and the World Bank has gradually expanded the good policy concept to include a wider and more complex set of characteristics than originally considered. Nevertheless, if the variation in aid effectiveness across countries is not policy-induced but rather a result of poor initial conditions, a different aid allocation rule would maximize the impact of foreign aid. Moreover, the empirical finding that aid is effective, but only when accompanied by good policy, turns out to be delicate. It is robust neither to alternative specifications of the regression model (Hansen and Tarp, 2001) nor to new data (Easterly, Levine and Roodman, 2004).

Clemens, Radelet and Bhavnani (2004), Dalgaard, Hansen and Tarp (2004) and Roodman (2004) offer up-to-date accounts. It emerges that the single most common result of recent empirical studies is that aid has a positive impact on per capita growth. There is also strong evidence to suggest that the importance of ‘deep’ structural characteristics is not yet fully understood. In sum, the accumulated cross-country evidence is encouraging, and Dalgaard and Hansen (2005) estimate that the aggregate real rate of return on foreign aid financed investments is in the range of 20–25 per cent. Attention should turn to how the effectiveness of aid can and should be improved rather than concentrating on whether aid works. This implies, for example, that focus should shift from aggregate aid to different forms of aid and their application in different types of aid receiving countries – modalities matter.

Future prospects

After many years when the project modality was the main vehicle for transferring aid, stabilization and broad structural reforms with associated programme aid were promoted vigorously in the early 1980s. A decisive shift from the state to the market as the key driver behind development was pursed. The East Asian financial crisis in 1997 signalled that the time had come for a rethink of the Washington *consensus*; and it is now widely agreed that quick-fix and single-actor approaches to development – focusing on either the state or the market – are not going to work. The state and the market have complementary roles to play in the struggle against poverty and inequality.

Aid fatigue is still evident in the international aid community, but it does seem that aid is gradually being rehabilitated from the low point of the mid-1990s. The empirical evidence that ‘aid works’ has been mounting steadily, and recent calls have been made for a ‘big push’ or a ‘Marshall Plan’ for Africa (World Economic Forum, 2005), and foreign aid flows seem to have picked up considerably after 2002.

The UN has established a target of halving world poverty by 2015 in the context of its Millennium Development Goals (MDG) (UN, 2002), and the USA has embarked on a $5 billion Millennium Challenge Account (MCA) meant to stimulate aid to poor countries (Bush, 2005).

All of this should not detract attention from the fact that many key challenges remain to be effectively addressed. The institutional set-up for bilateral aid delivery remains complex, uncoordinated and overburdened with many diverse tasks and aims; and calls for reform of the UN have become common. Moreover, it is far from settled where the balance between selectivity and conditionality is situated. An

underlying dilemma here is that it remains disputed how the balance between real or perceived needs on the one hand and development potential and performance on the other should be struck. Various proposals and guidelines exist (including the existing IDA aid allocation formula), but much of this relies ‘too heavily on a uniform model of what works in development policy’ (Kanbur, 2005). Past experiences provide many useful lessons about foreign aid (Robinson and Tarp, 2000), but the search for more effective answers to these kinds of questions is far from complete.

Finally, aid has gradually become a much smaller player in the world economy than private capital flows. Foreign aid decision makers are well advised to try to sharpen their implementation skills and develop complementary relationships with, among others, private capital markets and NGOs (Roland-Holst and Tarp, 2004). In an increasingly global world, possibilities and challenges are also opening up in the arena of international public goods. Foreign aid analysts would do well to explore these possibilities alongside more traditional investment and programme support activities, targeted on the provision of domestic public goods in poor countries.

See Also

- development economics
- fiscal and monetary policies in developing countries
- international financial institutions (IFIs)
- micro-credit
- third world debt

Bibliography

Adam, C. and O’Connell, S. 1999. Aid, taxation and development in Sub-Saharan Africa. *Economics and Politics* 11, 225–54.

Alesina, A. and Dollar, D. 2000. Who gives aid to whom and why? *Journal of Economic Growth* 5, 33–63.

Andersen, T. B., Harr, T. and Tarp, F. 2004. On US politics and IMF lending. Discussion Paper 04–11. Copenhagen: Institute of Economics, University of Copenhagen. Online. Available at http://www.econ.ku.dk/wpa/pink/2004/0411.pdf, accessed 22 June 2005.

Boone, P. 1996. Politics and the effectiveness of foreign aid. *European Economic Review* 40, 289–329.

Burnside, C. and Dollar, D. 1997. Aid, policies, and growth. Policy Research Working Paper 1777. Washington, DC: Development Research Group, World Bank.

Burnside, C. and Dollar, D. 2000. Aid, policies, and growth. *American Economic Review* 90, 847–68.

Bush, G. 2005. The Millennium Challenge Account. Online. Available at http://www.whitehouse.gov/infocus/developingnations/millennium.html, accessed 22 June 2005.

Cassen, R. and Associates. 1987. *Does Aid Work?* Oxford: Clarendon Press.

Chenery, H. and Strout, A.M. 1966. Foreign assistance and economic development. *American Economic Review* 56, 679–733.

Clemens, M., Radelet, S. and Bhavnani, R. 2004. Counting chickens when they hatch: the short-term effect of aid on growth. Working Paper 44. Washington, DC: Center for Global Development. Online. Available at http://www.cgdev.org/Publications/index.cfm?PubID=130, accessed 22 June 2005.

Collier, P. and Dollar, D. 2001. Can the world cut poverty in half? How policy reform and effective aid can meet the international development goals. *World Development* 29, 1787–802.

Collier, P. and Dollar, D. 2002. Aid allocation and poverty reduction. *European Economic Review* 46, 1475–500.

Cornia, G., Jolly, R. and Stewart, F. 1987. *Adjustment with a Human face: Protecting the Vulnerable and Promoting Growth*. Oxford: Clarendon Press.

Dalgaard, C.-J. and Hansen, H. 2005. The return to foreign aid. Discussion Paper 05–04. Copenhagen: Institute of Economics, University of Copenhagen. Online. Available at http://www.econ.ku.dk/wpa/pink/2005/0504.pdf, accessed 22 June 2005.

Dalgaard, C.-J., Hansen, H. and Tarp, F. 2004. On the empirics of foreign aid and growth. *Economic Journal* 114, F191–216.

Devarajan, S., Dollar, D. and Holmgren, T., eds. 2001. *Aid and Reform in Africa*. Oxford: Oxford University Press for the World Bank.

Duflo, E. 2004. Evaluating the impact of development aid programs: the role of randomized evaluation. Paper presented at the 2nd AFD-EUDN Conference, Paris, 25 November. Online. Available at http://www.eudnet.net/Download/AfD-EUDN04_Duflo.pdf, accessed 22 June 2005.

Easterly, W., Levine, R. and Roodman, D. 2004. Aid, policies, and growth: comment. *American Economic Review* 94, 774–80.

Easterly, W. 2001. *The Elusive Quest for Growth*. Cambridge, MA: MIT Press.

Gunning, J. 2005. Why give aid? Original paper presented at the 2nd AFD-EUDN Conference, Paris, 25 November. Revised paper online. Available at http://www.eudnet.net/Download/AfD-EUDN04_Gunning-revised.pdf, accessed 22 June 2005.

Hansen, H. and Tarp, F. 2000. Aid effectiveness disputed. *Journal of International Development* 12, 375–98.

Hansen, H. and Tarp, F. 2001. Aid and growth regressions. *Journal of Development Economics* 64, 547–70.

Hjertholm, P. and White, H. 2000. Foreign aid in historical perspective: background and trends. In *Foreign Aid and Development: Lessons Learnt and Directions for the Future*, ed. F. Tarp. London and New York: Routledge.

Kanbur, R. 2000. Aid, conditionality and debt in Africa. In *Foreign Aid and Development: Lessons Learnt and Directions for the Future*, ed. F. Tarp. London and New York: Routledge.

Kanbur, R. 2003. The economics of international aid. Working Paper 39. New York: Department of Applied Economics and Management, Cornell University. Online. Available at http://www.arts.cornell.edu/poverty/kanbur/HandbookAid.pdf, accessed 22 June 2005.

Kanbur, R. 2005. Reforming the formula: a modest proposal for introducing development outcomes in IDA allocation procedures. Original paper presented at the 2nd AFD-EUDN Conference, Paris, 25 November. Revised paper online. Available at http://www.arts.cornell.edu/poverty/kanbur/IDAForm.pdf, accessed 22 June 2005.

Mosley, P. 1987. *Overseas Aid: Its Defence and Reform*. Brighton: Wheatsheaf Books.

OECD (Organization for Economic Development and Cooperation). 2004. Development Assistance Committee: Disbursements and commitments of official and private flows (Table 1) and Development Assistance Committee: destination of official development assistance and official aid – disbursements (Table 2a). Online. Available at http://www.oecd.org/dac/stats/idsonline, accessed 16 February 2005.

Robinson, S. and Tarp, F. 2000. Foreign aid and development: summary and synthesis. In *Foreign Aid and Development: Lessons Learnt and Directions for the Future*, ed. F. Tarp. London and New York: Routledge.

Roland-Holst, D. and Tarp, F. 2004. New perspectives on aid effectiveness. In *Toward Pro-Poor Policies – Aid, Institutions and Globalization*, ed. B. Tungodden, N. Stern and I. Kolstad. Washington and New York: World Bank and Oxford University Press.

Roodman, D. 2004. The anarchy of numbers: aid, development, and cross-country empirics. Working paper 32. Washington, DC: Center for Global Development. Online. Available at http://www.cgdev.org/Publications/?PubID=36, accessed 22 June 2005.

Rosenstein-Rodan, P. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53, 202–11.

Svensson, J. 2003. Why conditional aid doesn't work and what can be done about it? *Journal of Development Economics* 70, 381–402.

Tarp, F., ed. 2000. *Foreign Aid and Development: Lessons Learnt and Directions for the Future*. London and New York: Routledge.

Thorbecke, E. 2000. The evolution of the development doctrine and the role of foreign aid, 1950–2000. In *Foreign Aid and Development: Lessons Learnt and Directions for the Future*, ed. F. Tarp. London and New York: Routledge.

UN (United Nations). 2002. The Millennium Development Goals and the United Nations role. Fact sheet. Online. Available at http://www.un.org/millenniumgoals/MDGs-FACTSHEET1.pdf, accessed 16 February 2005.

Williamson, J. 1997. The Washington Consensus revisited. In *Economic and Social Development into the XXI Century*, ed. L. Emmerij. Washington, DC: Inter-American Development Bank.

World Bank. 2002. *A Case for Aid: Building a Consensus for Development Assistance*. Washington, DC: World Bank.

World Economic Forum. 2005. Global leaders call for big push on aid to Africa at World Economic Forum meeting in Davos. Press Release. Online. Available at http://www.csrwire.com/article.cgi/3480.html, accessed 22 June 2005.

How to cite this article

Finn, Tarp. "foreign aid." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000165> doi:10.1057/9780230226203.0593

The New Palgrave Dictionary of Economics Online

foreign direct investment

Bruce A. Blonigen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Foreign direct investment (FDI) occurs when an individual or firm acquires controlling interest in productive assets of another country. We review the literature on FDI, which can be divided into two broad categories. The first is the inquiry into why multinational production occurs and the factors that determine the patterns of worldwide FDI. The second is the impact that FDI and multinational enterprises (MNEs) have on the parent and host countries, including economic growth, returns to factors of production, and externalities.

Keywords

agglomeration externalities; double taxation issue; efficiency wages; exchange rate volatility; factor endowments; firm, theory of; foreign direct investment; general equilibrium model; greenfield foreign direct investment; horizontal foreign direct investment; information externalities; international capital flows; intra-firm trade; knowledge-capital model of multinational enterprises; multinational enterprises (MNEs); multinational firms; ownership-location-internalization' (OLI) theory of multinational enterprises; partial equilibrium model of firm behaviour; portfolio investment; productivity spillovers; size of nations; tax competition; tax treaties; taxation of corporate profits; trade protection; transactions costs; vertical foreign direct investment; wage heterogeneity, sources of; wage spillovers

Article

Foreign direct investment (FDI) occurs when an individual or firm acquires a controlling interest (typically defined as at least ten per cent ownership) in productive assets in another country. This contrasts with portfolio investment, which includes purchases of foreign bonds, currencies, and stocks in amounts that do not provide control. The most common method of FDI is through the acquisition of a firm. Construction of a new plant is also common and typically referred to as 'greenfield' FDI. Other forms of FDI include partnerships in a foreign joint venture and earnings reinvested in an existing

foreign affiliate. Firms with affiliates in more than one country are termed ‘multinational enterprises’ (MNEs).

While real world GDP grew at a 2.5 per cent annual rate and real world exports grew by 5.6 per cent annually from 1986 through 1999, real world FDI inflows grew by 17.7 per cent over this same period (Giorgio and Venables, 2004). Additionally, Bernard, Jensen and Schott (2005) find that 90 per cent of US exports and imports flow through MNEs, with roughly 50 per cent of US trade flows occurring between affiliates of the same MNE, or what is termed ‘intra-firm trade’. While the majority of FDI flows are between developed countries, FDI accounted for the majority of capital flows to less-developed countries from 1990 to 2003 (UNCTAD, 2004).

The study of FDI can be divided into two broad categories. The first is the inquiry into why multinational production occurs and the factors that determine the patterns of worldwide FDI. The second is the impact that FDI and MNEs have on the parent and host countries, including economic growth, returns to factors of production, and externalities for innovative activity.

Understanding what motivates FDI by MNEs

Theory

Theoretical treatment of FDI and MNEs in the economics profession can be traced back to the 1970s, when researchers began to consider why some firms choose to locate production abroad rather than serve such markets through exports or licensing. A key insight is that MNEs may be distinguished by their ownership of firm-specific assets for which market failures can make exporting or licensing arrangements less attractive to the firm than FDI. For example, a foreign licensee may not offer full value in negotiations over a contract if the firm-specific asset is intangible and not fully revealed (for example, a unique production process), but the licensor firm will not want to reveal the asset fully until a contract is finalized. The costs associated with this inherent hold-up problem may then lead the firm to set up its own affiliate in the foreign market. This is termed ‘internalization’ in the literature, and forms the key element in the ‘ownership-location-internalization’ (OLI) theory of MNEs that developed out of this era and has been surveyed recently by Dunning (2001).

The OLI theory is an international business concept that was never formally represented in a mathematical model. As such, the international economics literature continued to treat FDI as simply another capital flow until the mid-1980s, even though its features and patterns differed from those of other capital flows. This changed with papers by Markusen (1984) and Helpman (1984) that developed general equilibrium models of MNEs. Both papers focused on another feature of firm-specific assets, namely, the public-goods aspect of many firm-specific assets that can be applied simultaneously in production across all plants owned by the firm. This feature of firm-specific assets makes it more attractive for a firm to build multiple plants, though something else must be added to a model to explain locating plants into foreign countries. In Helpman (1984) this is accomplished by assuming that MNEs can be separated into two types of activities: a skill-intensive headquarters that generates the firm-specific assets, and a low-skill-intensive production process. If endowment differences are sufficient across countries, MNEs will vertically separate the firm between headquarter services in the skill-abundant parent country and production in the low-skill host country. This type of model is called a

‘vertical FDI’ model. In contrast, Markusen's (1984) model generates multi-plant MNEs through the introduction of trade costs (that is, transportation costs, trade barriers, and so on) that are large enough that an MNE chooses to replicate itself in the foreign country to serve the market there. This type of model is termed ‘horizontal FDI’.

These models have become the main theoretical MNE frameworks for trade economists, as recent literature has extended these models. Brainard (1997) develops and tests hypotheses from a simplified horizontal MNE model assuming monopolistic competition. Markusen et al. (1996) develop an MNE model that blends both the horizontal and vertical models into what is termed the ‘knowledge-capital’ model. More recently, Helpman, Melitz and Yeaple (2004) have developed a model that can explain the coexistence of both exporting and MNEs in the same industry by allowing for heterogeneity across firms; other papers have developed models that formalize the role of transactions costs and theory of the firm (for example, Antras and Helpman, 2004; Feenstra and Hanson, 2005).

Empirics

Empirical work on the factors that determine FDI patterns has focused primarily on the effect of government policies and macroeconomic phenomena such as exchange rates and taxes. Most of these studies motivate their analyses with a partial equilibrium model of firm behaviour responding to these various factors. Only recently have empirical studies examined the more fundamental long-run drivers of total FDI activity, such as country size and factor endowments, as predicted by the general equilibrium modelling discussed above. Availability of micro-level data has been an issue for the literature as well. Testing theories of firm-level models with industry- or country-level data requires strong assumptions about firm characteristics. While firm-level data is being employed more often in recent work, much of the literature has examined more aggregate data.

Exchange rates

The effects of exchange rate movements on FDI are not immediately obvious. If a host country's currency depreciates relative to the parent country's currency, this lowers the price of host-country assets. However, if the asset generates returns in the host country's currency, these returns have likewise depreciated in the parent-country currency. Froot and Stein (1991) and Blonigen (1997), however, provide theoretical links that predict that host-country depreciations increase inbound FDI; and empirical evidence generally supports this. A related literature has examined how exchange rate expectations may affect FDI decisions. Campa (1993) provides theory and evidence that exchange rate uncertainty will decrease FDI, while Cushman (1985) and Goldberg and Kolstad (1995) conclude that quite opposite results can be expected and found depending on the firm's trade linkages across markets. On a final note, there has been recent work on the impact of exchange rate crises on FDI. Surprisingly, FDI is relatively stable through currency crises in host countries and, in fact, Aguiar and Gopinath (2005) show that MNEs opportunistically increase their investments in these host countries.

Taxes

Like exchange rate movements, the effect of taxes on FDI has not proven to be straightforward either. While there is an array of taxes that may affect FDI, the primary focus has been on corporate income tax rates in host countries. The natural hypothesis is that higher host-country tax rates discourage FDI, and a survey by de Mooij and Ederveen (2003) finds a median elasticity of tax rates on FDI of minus 3.3 across 25 different empirical studies. However, the literature has also shown that the effects of taxes on FDI can vary substantially depending on the type of taxes, the form of FDI (see, for example, Hartman, 1985), and the influence of government policy.

Perhaps the most explored issue in this literature has been the issue of how parent countries deal with the ‘double taxation’ issue – taxation in both host and parent countries. The common distinction is between territorial countries that do not tax any income outside of the parent country, exempting foreign-earned income from tax liability, and a worldwide tax method which considers all earned income by its parent firms potentially taxable, but may treat foreign income in a number of ways to avoid double taxation of the MNE. The standard treatment to deal with this double taxation issue is for the home country to offer a credit or a deduction of foreign tax payment made by the MNE. A number of studies of the US 1986 tax reform find mixed evidence for differences in FDI behaviour under different parent-country tax regimes (for example, Scholes and Wolfson, 1990; Swenson, 1994). Much stronger results come from work by Hines (1996) which finds that US taxation decreases FDI more for non-credit-system foreign investors than for credit-system foreign investors.

A final significant literature in this area is tax competition between countries competing for FDI (for example, Janeba, 1995) and the impact of bilateral tax treaties between countries (for example, Chisik and Davies, 2004). Hines (1999) and Gresik (2001) have excellent surveys of the FDI and taxation literature.

Other factors

A variety of other smaller literatures have investigated the effect of other factors on FDI. These include the effects of host-country institutions (Wei, 2000), trade protection policies, and agglomeration and information externalities (Head, Ries and Swenson, 1995; Blonigen, Ellis and Fausten, 2005).

Examination of general-equilibrium model predictions

More recently, empirical efforts have been made to more closely match empirical specifications of country-level FDI activity with general-equilibrium models of MNEs. Most previous empirical work uses gravity-based variations to model country-level FDI patterns where size of countries and distance between them are key regressors. Carr, Markusen and Maskus (2001) instead lay out an empirical specification based on the knowledge-capital model of MNE activity which suggests that factor endowment differences are an important control not found in gravity-based specifications. These endowment differences are important as they proxy for vertical MNE motivations. While Carr, Markusen and Maskus (2001) find that the data fit the knowledge-capital model, follow-up work has found specification issues that call into question evidence of vertical motivations for FDI (see Blonigen, Davies and Head, 2003; Braconier, Norback and Urban, 2005). Alternative approaches by Yeaple

(2003b) and Hanson, Mataloni and Slaughter (2005), however, have confirmed vertical motivations in the data, at least for certain sectors such as electronics and transportation equipment. Another concern pointed out by Yeaple (2003a) is that third country interactions may matter for FDI patterns. Recent empirical work by Baltagi, Egger and Pfaffermayr (2007) suggests that such effects are important empirically.

The economic impact of FDI and MNE activity

A second significant part of the FDI literature is the examination of FDI impacts on parent and, particularly, host countries. The primary areas of study have been on the effect of FDI on host country wages, technology spillovers, and economic growth.

Studies of FDI effects on host-country wages typically begin with the hypothesis that MNEs raise wages in the host country. Part of this is ascribed to the fact that the value of marginal product will be higher with MNEs due to productivity advantages and, thus, MNEs pay higher wages. However, an argument can also be made that MNEs need to pay higher efficiency wages than local firms to attract quality workers in an environment which they are relatively uninformed. Regardless of the explanation, the empirical evidence clearly shows that MNEs pay higher wages in both developed countries (for example, Globerman, Ries and Vertinsky, 1994) and less-developed ones (for example, Aitken, Harrison and Lipsey, 1996).

The more intriguing question is whether there are wage spillovers, in the sense that MNEs raise the wages paid by local firms as well. Spillovers are inherently difficult to identify in the data. Virtually all of the studies rely on interpreting a positive correlation between the presence of foreign firms in a local industry and the wages of local firms as evidence of spillovers. Not surprisingly, the evidence is decidedly mixed across numerous studies, as discussed by Lipsey and Sjöholm (2005). The theoretical development behind this issue is also relatively undeveloped in the literature as to when and where we should expect such wage spillovers.

A related issue is the effects of FDI on wage inequality. If MNEs have different technologies that demand different types of labour from local firms, increased FDI can lessen or exacerbate existing wage inequality. There are a number of cross-country studies that find a variety of FDI effects on wage inequality for the host country. Results for the United States using more detailed industry-level data likewise indicate little to no impact of outbound or inbound FDI on US wage inequality (Slaughter, 2000; Blonigen and Slaughter, 2001). Feenstra and Hanson (1997) provides a model to show how FDI can increase the difference between skilled and unskilled workers' wages in both host and parent countries with empirical work that finds strong impacts of US FDI activity on Mexican wage inequality.

The literature on productivity spillovers from FDI is vast compared with the one on wage spillovers, yet the evidence is decidedly mixed as well (see Görg and Strobl, 2001, for a survey). This is not surprising in many ways. First, theory is ambiguous on this issue. Foreign firms are presumably more efficient than the average local firm. Thus, FDI lowers market shares for local firms, which can lead to productivity losses for these firms, particularly if economies of scale are important. However, better technologies of foreign firms may ultimately leak to local firms through, for example, former employees or common suppliers. The second likely reason for mixed evidence is again the difficulty of identifying spillovers in the data (see Aitken and Harrison, 1999, for a discussion).

There is also a significant literature that attempts to gauge the overall impact of FDI on a host economy's economic growth. Like the trade and growth literature, this is difficult because of the obvious endogeneity issue, which is difficult to overcome. Such a question also relies on aggregate cross-country data, which is often quite poor. Most papers in the literature do not adequately control for these issues, and Carkovic and Levine (2005) points out the statistical sensitivity of these studies' results.

There are much smaller literatures on a variety of other host- and parent-country effects of FDI. This includes the impact of FDI on parent-country investment and employment (Blomström, Fors and Lipsey, 1997), the effects of FDI on host-country trade policies (Blonigen and Figlio, 1998), and differences in how MNEs adjust to local factor prices (Giorgio, Checci and Turrini, 2003).

See Also

- international capital flows
- location theory

Bibliography

Aguiar, M. and Gopinath, G. 2005. Fire-sale foreign direct investment and liquidity crises. *Review of Economics and Statistics* 87, 439–52.

Aitken, B. and Harrison, A. 1999. Do domestic firms benefit from direct foreign investment? Evidence from Venezuela. *American Economic Review* 89, 605–18.

Aitken, B., Harrison, A. and Lipsey, R. 1996. Wages and foreign ownership: a comparative study of Mexico, Venezuela, and the United States. *Journal of International Economics* 40, 345–71.

Antras, P. and Helpman, E. 2004. Global sourcing. *Journal of Political Economy* 112, 552–80.

Baltagi, B., Egger, P. and Pfaffermayr, M. 2007. Estimating models of complex FDI: are there third-country effects? *Journal of Econometrics*, 140(1), 5–51.

Bernard, A., Jensen, J. and Schott, P. 2005. Importers, exporters and multinationals: a portrait of the firms in the U.S. that trade goods. Working Paper No. 11404. Cambridge, MA: NBER.

Blomström, M., Fors, G. and Lipsey, R. 1997. Foreign direct investment and employment: home country experience in the United States and Sweden. *Economic Journal* 107, 1787–97.

Blonigen, B. 1997. Firm-specific assets and the link between exchange rates and foreign direct investment. *American Economic Review* 87, 447–65.

Blonigen, B., Davies, R. and Head, K. 2003. Estimating the knowledge-capital model of the

multinational enterprise: comment. *American Economic Review* 93, 980–94.

Blonigen, B., Ellis, C. and Fausten, D. 2005. Industrial groupings and foreign direct investment. *Journal of International Economics* 65, 75–91.

Blonigen, B. and Figlio, D. 1998. Voting for protection: does direct foreign investment influence legislator behavior? *American Economic Review* 88, 1002–14.

Blonigen, B. and Slaughter, M. 2001. Foreign-affiliate activity and U.S. skill upgrading. *Review of Economics and Statistics* 83, 362–76.

Braconier, H., Norback, P.-J. and Urban, D. 2005. Reconciling the evidence on the knowledge-capital model. *Review of International Economics* 13, 770–86.

Brainard, S. 1997. An empirical assessment of the proximity-concentration trade-off between multinational sales and trade. *American Economic Review* 87, 520–44.

Campa, J. 1993. Entry by foreign firms in the U.S. under exchange rate uncertainty. *Review of Economics and Statistics* 75, 614–22.

Carr, D., Markusen, J. and Maskus, K. 2001. Estimating the knowledge-capital model of the multinational enterprise. *American Economic Review* 91, 693–708.

Carkovic, M. and Levine, R. 2005. Does foreign direct investment accelerate economic growth? In *Does Foreign Direct Investment Promote Development?*, ed. T. Moran, E. Graham and Blomström. Washington, DC: Institute for International Economics.

Chisik, R. and Davies, R. 2004. Asymmetric FDI and tax-treaty bargaining: theory and evidence. *Journal of Public Economics* 88, 1119–48.

Cushman, D. 1985. Real exchange rate risk, expectations, and the level of direct investment. *Review of Economics and Statistics* 67, 297–308.

de Mooij, R. and Ederveen, S. 2003. Taxation and foreign direct investment: a synthesis of empirical research. *International Tax and Public Finance* 10, 673–93.

Dunning, J. 2001. The eclectic (OLI) paradigm of international production: past, present and future. *International Journal of Economics and Business* 8, 173–90.

Feenstra, R. and Hanson, G. 2005. Ownership and control in outsourcing to China: estimating the property-rights theory of the firm. *Quarterly Journal of Economics* 120, 729–61.

Feenstra, R. and Hanson, G. 1997. Foreign direct investment and relative wages: evidence from Mexico's maquiladoras. *Journal of International Economics* 42, 371–93.

Froot, K. and Stein, J. 1991. Exchange rates and foreign direct investment: an imperfect capital markets approach. *Quarterly Journal of Economics* 106, 1191–217.

Giorgio, B., Checci, D. and Turrini, A. 2003. Adjusting labor demand: multinational versus national firms: a cross-European analysis. *Journal of the European Economic Association* 1, 708–19.

Giorgio, B. and Venables, A. 2004. *Multinational Firms in the World Economy*. Princeton and Oxford: Princeton University Press.

Globerman, S., Ries and Vertinsky, I. 1994. The economic performance of foreign affiliates in Canada. *Canadian Journal of Economics* 27, 143–56.

Goldberg, L. and Kolstad, C. 1995. Foreign direct investment, exchange rate variability and demand uncertainty. *International Economic Review* 36, 855–73.

Görg, H. and Strobl, E. 2001. Multinational companies and productivity spillovers: a meta-analysis. *Economic Journal* 111, 723–39.

Gresik, T. 2001. The taxing task of taxing transnationals. *Journal of Economic Literature* 39, 800–38.

Hanson, G., Mataloni, R. and Slaughter, M. 2005. Vertical production networks in multinational firms. *Review of Economics and Statistics* 87, 664–78.

Hartman, D. 1985. Tax policy and foreign direct investment. *Journal of Public Economics* 26, 107–21.

Head, K., Ries, J. and Swenson, D. 1995. Agglomeration benefits and location choice: evidence from Japanese manufacturing investments in the United States. *Journal of International Economics* 38, 223–47.

Helpman, E. 1984. A simple theory of international trade with multinational corporations. *Journal of Political Economy* 92, 451–71.

Helpman, E., Melitz, M. and Yeaple, S. 2004. Export versus FDI with heterogeneous firms. *American Economic Review* 94, 300–16.

Hines, J., Jr. 1996. Altered states: taxes and the location of foreign direct investment in America. *American Economic Review* 86, 1076–94.

Hines, J., Jr. 1999. Lessons from behavioral responses to international taxation. *National Tax Journal* 52, 305–22.

Janeba, E. 1995. Corporate income tax competition, double taxation treaties, and foreign direct investment. *Journal of Public Economics* 56, 311–25.

Lipsey, R. and Sjöholm, F. 2005. The impact of inward FDI on host countries: why such different answers? In *Does Foreign Direct Investment Promote Development?*, ed. T. Moran, E. Graham and M. Blomström. Washington, DC: Institute for International Economics.

Markusen, J. 1984. Multinationals, multi-plant economies, and the gains from trade. *Journal of International Economics* 16, 205–26.

Markusen, J., Venables, A., Eby-Konan, D. and Zhang, K. 1996. A unified treatment of horizontal direct investment, vertical direct investment and the pattern of trade in goods and services. Working Paper No. 5696. Cambridge, MA: NBER.

Scholes, M. and Wolfson, M. 1990. The effects of changes in tax law on corporate reorganization activity. *Journal of Business* 63, S141–S164.

Slaughter, M. 2000. Production transfer within multinational enterprises and American wages. *Journal of International Economics* 50, 449–72.

Swenson, D. 1994. The impact of U.S. tax reform on foreign direct investment in the United States. *Journal of Public Economics* 54, 243–66.

UNCTAD (United Nations Conference on Trade and Development). 2004. *World Investment Report 2004: The Shift to Services*. New York and Geneva: United Nations.

Wei, S.-J. 2000. How taxing is corruption on international investors? *Review of Economics and Statistics* 82, 1–11.

Yeaple, S. 2003a. The complex integration strategies of multinationals and cross country dependencies in the structure of foreign direct investment. *Journal of International Economics* 60, 293–314.

Yeaple, S. 2003b. The role of skill endowments in the structure of U.S. outward foreign direct investment. *Review of Economics and Statistics* 85, 726–34.

How to cite this article

Blonigen, Bruce A. "foreign direct investment." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000168> doi:10.1057/9780230226203.0594

The New Palgrave Dictionary of Economics Online

foreign exchange market microstructure

Martin D.D. Evans

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Research on foreign exchange market microstructure focuses on the idea that trading is an integral part of the process whereby information relevant to the pricing of foreign currency becomes embedded in spot rates. Micro-based models of this process produce empirical predictions that find strong support in the data. Micro-based models can account for a large proportion of the daily variation in spot rates. They also supply a rationale for the apparent disconnect between spot rates and fundamentals. Micro-based models provide out-of-sample forecasting power for spot rates that is an order of magnitude above that usually found in exchange-rate models.

Keywords

arbitrage; common knowledge news; depreciation rates; exchange rate dynamics; exchange rate puzzles; financial market contagion; foreign exchange market microstructure; foreign exchange risk premium; information aggregation; order flows; spot exchange rates; stop-loss orders

Article

Models of foreign exchange (FX) market microstructure examine the determination and behaviour of spot exchange rates in an environment that replicates the key features of trading in the FX market. Traditional macro exchange-rate models pay little attention to how trading in the FX market actually takes place. The implicit assumption is that the details of trading (that is, who quotes currency prices and how trade takes place) are unimportant for the behaviour of exchange rates over months, quarters or longer. Micro-based models, by contrast, examine how information relevant to the pricing of foreign currency becomes reflected in the spot exchange rate via the trading process. According to this view, trading is not an ancillary market activity that can be ignored when one considers exchange rate behaviour. Rather, trading is an integral part of the process through which spot rates are determined and evolve. Recent micro-based FX models also differ from other areas of microstructure research in their

focus on the links between trading, asset price dynamics and the macroeconomy.

Recent research on exchange rates stresses the role of heterogeneity (for example, Bacchetta and van Wincoop, 2006; and Hau and Rey, 2006). Micro-based exchange-rate models start from the premise that much of the information about the current and future state of the economy is dispersed across agents (that is, individuals, firms and financial institutions). Agents use this information in making their everyday decisions, including decisions to trade in the FX market at the prices quoted by dealers. Dealers quote prices (for example, dollars per unit of foreign currency) at which they stand ready to buy or sell foreign currency; they will purchase foreign currency at their bid quote, and sell foreign currency at their ask quote. Agents that choose to trade with an individual dealer are termed the ‘dealer's customers’. The difference between the value of purchase and sale orders *initiated* by customers during any trading period is termed ‘customer order flow’. Importantly, order flow is different from trading volume because it conveys information. Positive (negative) order flow indicates to dealers that, on balance, their customers value foreign currency more (less) than their asking (bid) price. By tracking who initiates each trade, order flow provides a measure of the information exchanged between counterparties in a series of financial transactions.

Trading in the FX market also takes place between dealers. In direct inter-dealer trading, one dealer asks another for a bid and ask quote, and then decides whether he wishes to trade. When the dealer initiating the trade purchases (sells) foreign currency, the trade generates a positive (negative) inter-dealer order flow equal to the value of the purchase (sale). Inter-dealer trading can also take place indirectly via brokerages that act as intermediaries between two or more dealers. In recent years electronic brokerages have come to dominate inter-dealer trading, but the inter-dealer order flow generated by brokered trades plays the same informational role as the order flow associated with direct inter-dealer trading.

Micro-based exchange rate determination

At first sight, the pattern of FX trading activity seems far too complex to provide any useful insight into the behaviour of exchange rates. However, on closer examination two key features emerge. First, the equilibrium spot exchange rate does not come out of a ‘black box’. Instead, it is solely a function of the foreign currency prices quoted by dealers at a point in time. This is a distinguishing feature of micro-based exchange rate models and has far-reaching implications. Second, information about the current and future state of the economy will impact on exchange rates only when, and if, it affects dealer quotes. Dealers may revise their quotes in response to new public information that arrives via macroeconomic announcements. They may also revise their quotes based on orders they receive from customers and other dealers. This order flow channel is the means though which dispersed information concerning the economy affects dealer quotes and hence the spot exchange rate. The role played by order flow in transmitting information to dealers, and hence to their quotes, is another distinguishing feature of micro-based exchange rate models.

Micro-based models incorporate these two features of FX trading into a simplified setting. Canonical multi-dealer models, such as Lyons (1997) and Evans and Lyons (2002a), posit a simple sequence of quoting and trading. At the start of each period, dealers quote FX prices to customers. These prices are assumed to be good for any amount and are publicly observed. Each dealer then receives orders from a subset of agents, his customers. Dealers next quote prices in the inter-dealer market. These prices, too,

are good for any quantity and are publicly observed. Dealers then have the opportunity to trade among themselves. Inter-dealer trading is simultaneous and trading with multiple partners is feasible.

In this trading environment, optimal quote decisions take a simple form; all dealers quote the same FX price to both customers and other dealers. We can represent the period-*t* quote as

$$s_t = (1-b)\sum_{i=0}^{\infty} b^i E[f_{t+i}|\Omega_t^D], \tag{1}$$

where $0 < b < 1$. s_t is the log price of foreign currency quoted by all dealers, and f_t denotes exchange rate fundamentals. The form for fundamentals differs according to the macroeconomic structure of the model. For example, in Evans and Lyons (2004b) f_t includes home and foreign money supplies and household consumption. In models where central banks conduct monetary policy via the control of short-term interest rates (that is, follow Taylor rules), f_t will include variables used to set policy. More generally, f_t will include a term that identifies the foreign exchange risk premium.

While eq. (1) takes the present value form familiar from standard international macro models, here it represents how dealers quote the price for foreign currency in equilibrium. All dealers choose to quote the same price in this trading environment because doing otherwise opens them up to arbitrage, a costly proposition. (Recall that quotes are publicly observed and good for any amount, so any discrepancy between quotes would represent an opportunity for a riskless trading profit.) Consequently, the month-t quote must be a function of information known to all dealers. Equation (1) incorporates this requirement with the use of the expectations operator, $E[\cdot|\Omega_t^D]$, that denotes expectations conditioned on information common to all dealers at the start of month *t*, Ω_t^D. This is not to say that all dealers have the same information. On the contrary, the customer order flows received by individual dealers represent an important source of private information, so there may be a good deal of information heterogeneity across dealers at any one time. The important point to note from eq. (1) is that, due to the ‘fear of arbitrage’, individual dealers choose not to quote prices based on their own private information. In this trading environment dealers use their private information in initiating trade with other dealers, and, in so doing, contribute to the process through which all dealers acquire information.

The implications of micro-based models for the dynamics of spot rates are most easily seen by rewriting (1) as

$$\Delta s_{t+1} = \frac{1-b}{b}(s_t - E[f_t|\Omega_t^D]) + \varepsilon_{t+1}, \tag{2}$$

where $\Delta s_{t+1} = s_{t+1} - s_t$, and

$$\varepsilon_{t+1} = \frac{1-b}{b}\sum_{i=1}^{\infty} b^i (E[f_{t+i}|\Omega_{t+1}^D] - E[f_{t+i}|\Omega_t^D]). \quad (3)$$

Equation (2) decomposes the change in the log spot rate (that is, the depreciation rate for the home currency) into two components: the expected change $E[\Delta s_{t+1}|\Omega_t^D]$ identified by the first term, and the unexpected change, $\varepsilon_{t+1} = s_{t+1} - E[s_{t+1}|\Omega_t^D]$, shown in eq. (3). Both terms contribute to exchange rate dynamics in micro-based models. In equilibrium, dealers' period-*t* quote must be based on expectations, $E[\Delta s_{t+1}|\Omega_t^D]$, that match the risk-adjusted returns on different assets. This means that variations in the interest differential between home and foreign bonds can contribute to the volatility of the depreciation rate via the first term in (2). The second term, ε_{t+1}, identifies the impact of new information received by all dealers between the start of periods *t* and $t+1$. Equation (3) shows that new information impacts on the FX price quoted in period $t+1$ to the extent that it revises forecasts of the present value of fundamentals based on dealers' common information.

As an empirical matter, depreciation rates are very hard to forecast, so the dynamics of spot rates are largely attributable to the effects of news. Here micro-based models have a big advantage over their traditional counterparts because their trade-based foundations provide detail on how news affects spot rates. In particular, as eq. (3) indicates, micro-based models focus on how new information about the fundamentals reaches dealers and induces them to revise their FX quotes.

News concerning fundamentals can reach dealers either directly or indirectly. Common knowledge (CK) news operates via the direct channel. CK news contains unambiguous information about current and/or future fundamentals that is simultaneously observed by all dealers and immediately incorporated into the FX price they quote. In principle, macroeconomic announcements (for example, on GDP, industrial production or unemployment) could be a source for CK news, but in practice they rarely contain much unambiguous new information. In fact, CK news events appear rather rare. The indirect channel operates via order flow and conveys dispersed information about fundamentals to dealers. Dispersed information comprises micro-level information on economic activity that is correlated with fundamentals. Examples include the sales and orders for the products of individual firms, market research on consumer spending, and private research on the economy conducted by financial institutions. Dispersed information first reaches the FX market via the customer order flows received by individual dealers. These order flows have no immediate impact on dealer quotes because they represent private information to the recipient dealer. The information in each customer flow will impact on quotes only when it is known to all dealers. Inter-dealer order flow is central to this process. Individual dealers use their private information to trade in the inter-dealer market. In so doing, information on their customer orders is aggregated and

spread across the market. This process is known as ‘information aggregation’. Dispersed information is incorporated into dealer quotes once this process is complete.

Empirical evidence

The appeal of micro-based models is not solely based on their theoretical foundations. In marked contrast with traditional exchange-rate models, micro-based models have enjoyed a good deal of empirical success. Evans and Lyons (2002a) first demonstrated their empirical power when studying the relation between depreciation rates and inter-dealer order flow at the daily frequency. In particular, they show that aggregate inter-dealer order flow from trading in the spot dollar/deutschmark market on day d accounts for 64 per cent of the variation in the depreciation rate, Δs_{d+1}, between the start of days *d* and $d+1$. This is a striking result because macro models can account for less than one per cent of daily depreciation rates. It is also readily explained in terms of eqs. (2) and (3). Aggregate inter-dealer order flow during day *d* trading provides a measure of the market-wide information flow that dealers use to revise their quotes between the start of days *d* and $d+1$. This contemporaneous relationship between depreciation rates and inter-dealer order flows appears robust. It holds for many different currencies and for different currency-order flow combinations (for example, Evans and Lyons, 2002b; Payne, 2003; and Froot and Ramadorai, 2005). It is also worth emphasizing that order flow's impact on spot rates is very persistent. There is very little serial correlation in the daily depreciation rates for major currencies, so the order flow impact on current FX quotes persists far into the future.

While consistent with the idea that dispersed information is impounded into spot exchange rates via inter-dealer order flow, these results do not provide direct evidence on the ultimate source of exchange rate dynamics. According to micro-based models, the analysis of customer order flows should provide the evidence. In particular, if inter-dealer order flows measure the market-wide flow of information concerning fundamentals originally motivating customer orders, customer orders should also have explanatory power for depreciation rates. This is indeed the case. Evans and Lyons (2004b) show that a significant contemporaneous relationship exists between depreciation rates and the customer order flows of a single large bank. Moreover, the strength of this relationship increases as we move from a one-day to a one-month horizon. This, too, is consistent with micro-based models: At longer horizons, customer flows from a single bank should be a better proxy for the market-wide flow of information driving spot rates.

Micro-based models also make strong empirical predictions about the relationship between order flows and fundamentals. According to eq. (1), dealers are forward-looking when quoting FX prices, so spot rates embody their forecasts for fundamentals based on common information, Ω_t^D. One empirical implication of this observation is that spot exchange rates should have forecasting power for fundamentals. While there is some evidence that this is true for variables that comprise fundamentals in many models (Engel and West, 2005), the forecasting power is rather limited. Micro-based models also have implications for the forecasting power of order flows. If order flows convey information about fundamentals that is not yet common knowledge to all dealers (that is, not in Ω_t^D), then they should have incremental forecasting power for fundamentals, beyond the forecasting ability of any variable in Ω_t^D.

This is a strong prediction: it says that order flow should add to the forecasting power of all other variables in Ω_t^D, including the history of spot rates and the fundamental variable itself. Nevertheless, Evans and Lyons (2004b) find ample support for this prediction using customer order flows and candidate fundamental variables such as output, inflation and money supplies. These findings provide direct evidence on the information content of customer order flows, and provide a new perspective on the link between exchange rates and fundamentals.

Dispersed information concerning fundamentals need not come only from the activities of individuals, firms and financial institutions. Scheduled announcements on macroeconomic variables (for example, GDP, inflation or unemployment) can also be a source of dispersed information. If agents have different views about the mapping from the announced variable to fundamentals, then the news contained in any announcement, while simultaneously observed, will not be common knowledge. For example, two firms may interpret the same announcement on last quarter's GDP as having different implications for future GDP growth. Differing interpretations about the implications of commonly observed news will be a source of customer order flows because they imply heterogeneous views about future returns, which in turn induces portfolio adjustment. Thus, micro-based models raise the possibility that the exchange rate effects of macro announcements operate via both a direct channel (that is, when the announcement contains CK news) and an indirect channel. Love and Payne (2003) and Evans and Lyons (2003; 2005b) find evidence that both channels are operable. Evans and Lyons estimate that roughly two-thirds of the effect of a macro announcement is transmitted indirectly to the dollar/deutschmark spot rate via order flow, and one-third directly into quotes. With both channels operating, macro news is estimated to account for more than one-third of the variance in daily depreciation rates. This level of explanatory power far surpasses that found in earlier research analysing the impact of macro news on exchange rates (for example, Andersen et al., 2003). It also further cements the link between spot rates and the macro variables comprising fundamentals.

Order flows, returns and the pace of information aggregation

The process by which the information contained in the customer flows becomes known across the market, and hence embedded into FX quotes, is complex. The individual customer and inter-dealer orders received by each dealer contain some dispersed information about the economy, but extracting the information from each order constitutes a difficult inference problem. Under some circumstances the inference problems are sufficiently simple for every dealer to learn all there is to know about fundamentals in a few rounds of inter-dealer trading. In this case, the pace of information aggregation is very fast, so that new information concerning fundamentals is quickly reflected in dealer quotes whether the news is initially dispersed or common knowledge. The resulting dynamics for exchange rates over weeks, months or quarters will be indistinguishable from the predictions of macro models. Under other circumstances, the inference problem facing individual dealers is sufficiently complex to slow down the pace of information aggregation. Here it takes many rounds of inter-dealer trading before the dispersed information concerning fundamentals becomes known across the market. This scenario is much more likely from a theoretical perspective. Evans and Lyons (2004a) show that the conditions needed for fast information aggregation are quite stringent. Of course, because inter-dealer trading takes places continuously, dispersed information could be completely embedded in FX quotes in a short period of

calendar time (for example, a day), even if the pace of information aggregation is slow. In principle, dealers might be able to learn a good deal from the multitude of orders they receive in a typical day, even if individual orders are relatively uninformative. The question of whether it takes significant amounts of calendar time before dispersed information is embedded in FX quotes can be answered only empirically.

If the pace of information aggregation is slow, customer order flows across the market contain information that will become known to all dealers only at a later date. So, if the customer orders received by an individual bank are representative of the market-wide flows, they should have forecasting power for the future market-wide flow of information that drives quote revision. Recent empirical findings support this possibility. Evans and Lyons (2005b; 2005c) show that customer order flows have significant forecasting power for future depreciation rates both in and out of sample. These results are qualitatively different from the contemporaneous empirical link between order flows and depreciation rates discussed above. In the context of eqs. (2) and (3), the market-wide flow of information from period-*t* trading impacts on the deprecation rate, Δs_{t+1}, via ε_{t+1}. The contemporaneous link arises because period-*t* inter-dealer order flows measure the market-wide information flow, ε_{t+1}. In contrast, the forecasting power of customer flows for the depreciation rate arises because ε_{t+1} contains information that was originally in the customer orders received by individual banks *before* period-*t* trading.

These forecasting results are surprising in terms of both their horizon and strength. In particular, out-of-sample forecasts based on customer flows from month $t-1$ can account for roughly 16 per cent of the variation in next month's depreciation rate, Δs_{t+1}. This finding suggests that the pace of information aggregation is far, far slower than was previously thought; it seems to take weeks, not minutes, for dispersed information to be fully assimilated across the market. The level of forecasting power is also an order of magnitude above that usually found in exchange rate models. For example, the in-sample forecasting power of interest differentials for monthly depreciation rates is only in the two to four per cent range.

The slow pace of information aggregation may shed light on one of the long-standing puzzles in exchange rate economics; the disconnect between spot exchange rates and fundamentals over short and medium horizons (Meese and Rogoff, 1983). The idea is quite simple. If changes in fundamentals are reflected in spot rates only when information concerning the change is recognized by dealers across the market, the slow pace of information aggregation will mask the link between the depreciation rate and the change in fundamentals over short horizons, because the latter is a poor proxy for the market-wide flow of information. Simulations in Evans and Lyons (2004a) show that this masking effect can be quite substantial. Fundamentals account for only 50 per cent of variation in spot rates at the two-year horizon even though information aggregation takes at most four months.

One factor that might contribute to the slow pace of information aggregation is the presence of price-contingent order flow generated by feedback trading. Stop-loss orders, for example, represent a form of positive feedback trading in which a fall in the FX price triggers negative order flow from customers wishing to insure their portfolios against further losses. Feedback trading of a known form does not complicate the inference problem facing dealers because the orders it generates are simply a function of old market-wide information. However, when the exact form of the feedback is unknown it makes inferences less precise and so slows down the pace of information aggregation. Osler (2005) argues that

feedback trading will be an important component of order flow when quotes approach the points at which stop-loss orders cluster. A fall in FX quotes at these points can trigger a self-reinforcing price cascade where causation runs from quotes to order flow.

Some economists argue that the early empirical findings linking order flow and the depreciation rate reflected the presence of positive feedback trading rather than the transmission of dispersed information. Indeed, there is no way to tell whether intra-day causation runs from order flows to quotes or vice verse from just the contemporaneous correlation between order flow and the deprecation rate measured in daily data. However, the new evidence on the forecasting power of order flow for both depreciation rates and fundamentals firmly points to order flow as the conveyor of dispersed information. This is not to say that feedback trading is absent. Portfolio insurance and other price-contingent trading strategies (such as liquidity provision) undoubtedly contribute to order flows, and their presence may actually explain why the pace of information aggregation is so slow.

Future research

Exchange rate research based on micro-based models is still in its infancy. The past few years have seen a rapid advance in theoretical modelling and some surprising empirical results. Advances on the empirical side will be spurred by the greater availability of trading data. On the theoretical side, micro-based modelling may provide new insights into the determinants of the foreign-exchange risk premium, the efficacy of foreign exchange intervention, and the anatomy of financial contagion.

See Also

- exchange rate dynamics
- exchange rate volatility
- information aggregation and prices

I thank Richard Lyons for valuable discussions and gratefully acknowledge the financial support of the National Science Foundation.

Bibliography

Andersen, T., Bollerslev, T., Diebold, F. and Vega, C. 2003. Micro effects of macro announcements: real–time price discovery in foreign exchange. *American Economic Review* 93, 38–62.

Bacchetta, P. and van Wincoop, E. 2006. Can information heterogeneity explain the exchange rate determination puzzle? *American Economic Review* 96, 552–76.

Engel, C. and West, K. 2005. Exchange rates and fundamentals. *Journal of Political Economy* 113, 485–517.

Evans, M. and Lyons, R. 2002a. Order flow and exchange rate dynamics. *Journal of Political Economy* 110, 170–80.

Evans, M. and Lyons, R. 2002b. Informational integration and FX trading. *Journal of International Money and Finance* 21, 807–31.

Evans, M. and Lyons, R. 2003. How is macro news transmitted to exchange rates? Working Paper No. 9433. Cambridge, MA: NBER.

Evans, M. and Lyons, R. 2004a. A new micro model of exchange rates. Working Paper No. 10379. Cambridge, MA: NBER.

Evans, M. and Lyons, R. 2004b. Exchange rate fundamentals and order flow. Workingnpaper. Online. Available at http://www.georgetown.edu/faculty/evansm1/, accessed 16 November 2007.

Evans, M. and Lyons, R. 2005a. Do currency markets absorb news quickly? *Journal of International Money and Finance* 24, 197–217.

Evans, M. and Lyons, R. 2005b. Meese–Rogoff Redux: micro-based exchange rate forecasting, *American Economic Review, Papers and Proceedings*, 95, 405–14.

Evans, M. and Lyons, R. 2005c. Exchange rate fundamentals and order flow. Mimeo, Georgetown University. Online. Available at http://www.georgetown.edu/faculty/evansm1/, accessed 6 June 2006.

Froot, K. and Ramadorai, T. 2005. Currency returns, intrinsic value, and institutional–investor flows. *Journal of Finance* 60, 1535–65.

Hau, H. and Rey, H. 2006. Exchange rates, equity prices, and capital flows. *Review of Financial Studies* 19, 273–317.

Love, R. and Payne, R. 2003. Macroeconomic news, order flows, and exchange rates. Discussion Paper No. 475. Financial Markets Group, London School of Economics.

Lyons, R. 1997. A simultaneous trade model of the foreign exchange hot potato. *Journal of International Economics* 42, 275–98.

Meese, R. and Rogoff, K. 1983. Empirical exchange rate models of the seventies, *Journal of International Economics* 14, 3–24.

Osler, C. 2005. Stop-loss orders and price cascades in currency markets. *Journal of international Money and Finance* 24, 219–41.

Payne, R. 2003. Informed trade in spot foreign exchange markets: an empirical investigation. *Journal of International Economics* 61, 307–29.

How to cite this article

Evans, Martin D. D. "foreign exchange market microstructure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000299> doi:10.1057/9780230226203.0595

The New Palgrave Dictionary of Economics Online

foreign exchange markets, history of

Marcello de Cecco

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Foreign exchange transactions, known in classical antiquity, developed into markets in the Middle Ages. Italian dealers dominated the market until the 16th century, when they started being replaced by the Dutch and English. The City of London has been the centre of world forex markets since the 18th century and remains dominant even today. Transaction modes have been revolutionized by information technology. Volume has also grown enormously. But personal contact is still important, hence financial centres persist. The arrival of the euro has had consequences for the forex market are discussed, as will the emergence of China.

Keywords

arbitrage; bills of exchange; China, economics in; derivatives; euro; European Central Bank; European Monetary Union; fiat money; foreign exchange controls; foreign exchange markets; information technology; interest; specie-flow mechanism; speculation; usury

Article

‘Foreign exchange markets’ is an expression that people normally associate with foreign currency transactions, whether in notes or coins. That association is correct, but foreign exchange markets trade in all transactions concerning debt instruments denominated in foreign currencies. This is not a modern development, even if it is true that debt instruments and the transactions associated with them have multiplied as economies have become more complex and more open to one another.

Origins and causes

Trade in coins and debt instruments denominated in foreign currency is an ancient activity. Reference to it is found in ancient literatures and inscriptions belonging to many different cultures. From what one

can glean from these ancient texts, it was always a type of trade organized by dealers, who were sometimes only brokers but more often than not traded for their own account, and often mixed foreign exchange dealing with merchandizing and lending.

In the development of foreign exchange activities, it is impossible to exaggerate the importance exercised by the Aristotelian prohibition of usury, which the Koran and scholastic doctrine perpetuated in Muslim and Christian lands. Aristotle thought that only living beings could bear fruit. Money, not a living being, was by its nature barren, and any attempt to make it bear fruit (*tokos*, the Greek for ‘bearing fruit’, also means ‘interest’) was a crime against nature.

The need for intertemporal planning of economic activities requires the use of lending and its remuneration. Human ingenuity discovered, very soon after the Aristotelian prohibition, that while lending gave rise to interest (which was against nature), the sale of one asset against another, including coins, was a legitimate activity. Hence, the price at which that sale occurred could very appropriately hide a lending transaction. There followed an enormous diffusion of asset sales–purchases, which, after the break-up of the Roman Empire in the fifth century ad and the fragmentation of the Roman currency area into many smaller zones, often became foreign exchange transactions. The fluctuation of exchange rates between currencies provided a convincing case of risk associated with foreign exchange activities, and further reduced the possibility of transactors being accused of usury.

Raymond De Roover (1954) attributes to the Aristotelian prohibition the redirection of banking towards foreign exchange transactions that occurred from the early Middle Ages onwards. Since lending and borrowing at interest were outlawed, they had to be hidden inside more and more imaginatively devised foreign exchange transactions. This is a perfect case of financial innovation spurred by legal prohibition, which acquires a momentum of its own, generating a huge crop of by-products. Most of these by-products, and foreign exchange contracts and practices devised in the Middle Ages, are still present in today's markets, often even keeping their original names.

The most typical case is that of the bill of exchange, which is a transaction between two or more agents, giving rise to an exchange of foreign currency to be effected in different places at different times. The multiplicity of transactors, and of the contract's attributes, allows the fashioning of the contract in a remarkable number of different ways, following the needs of the transactors and the development of commercial and banking habits.

The fact that foreign exchange transactions are sales–purchases of assets denominated in foreign currencies, and that for a long time what could easily have been transacted in one currency had to be hidden behind a foreign exchange transaction, contributed from early on to the weaving of foreign exchange theory into an intricate web, as trade flows were recognized to be just one of the factors determining foreign exchange rates. Asset transactions obviously contributed at least as much to their determination. But while trade was visible, asset sales–purchases were not easily detected and recorded, and it was much more difficult to attribute exchange rate oscillations to their influence. This was especially so if, as we have already noted, a great number of such foreign exchange transactions, giving rise to a large volume of bills of exchange, actually hid domestic lending activity.

Foreign exchange in the Middle Ages: the rise of Italian market supremacy

The fragmentation of the Roman Empire gave rise in Italy to a fragmentation of monetary sovereignty and to the accompanying early specialization of Italian merchants in foreign exchange transactions. The

fact that the papacy was also seated in Italy made the adherence to religious prohibitions of usury superficially stricter; but, with the help of scholastic doctors, merchants were able to devise ways to circumvent the prohibitions.

All this ended up in helping Italian merchants to develop a vast body of knowledge about foreign exchange banking, which they tried to keep to themselves for as long as they could. Thus they became specialists in the transfer of funds from one place to another. Difficulty and danger connected with travel discouraged the physical transportation of metallic money, which was a scarce good anyway, at least until the diffusion of fiat currency in the 19th and especially the 20th centuries. Whoever could transfer titles to assets between geographically distant places stood to gain a great deal of money and power. Italians became masters of these arcane practices. First Florentine and Venetian bankers, then the Genoese, practically cornered this market for several centuries. They developed an enormous clearing network, encompassing most relevant trading places, where they kept agents and correspondents. As a result they could effect transfers everywhere. Sovereign rulers, who had to transfer vast sums because of their military operations in foreign lands, were the Italian bankers' best and worst clients. They tried to escape from the bankers' clutches, and to foster competition, but more often than not they were forced back into the bankers' hands by the superiority of the Italians' skills and by the bankers' monopoly of power. Philip II of Spain confided in a letter his dismay at not being able to understand foreign exchange problems. He had tried to get rid of the Genoese, but had to accept soon after that only they were able to transfer his American treasure from Spain to Flanders, and thus circumvent the maritime power of the English.

The market shift to Atlantic Europe

With the decline of the religious condemnation of usury, and the shift of trade from the Mediterranean to the Atlantic, the Italians' tight monopoly on the foreign exchange market faded away and was transferred first to Belgium and the Low Countries and then to Great Britain, or, more precisely, to London. It is quite remarkable how this skill always managed to bypass France, despite its being the richest country in Europe. Champagne fairs were dominated by foreign merchants, who monopolized exchange transactions. The same was true in Lyons. In fact, even the transfer of foreign exchange transactions to the shores of the North Sea and the Atlantic should be seen largely as a physical relocation of foreign exchange specialists to the places where trade had flourished. Foreign exchange transactions have remained a footloose activity, practised by a close-knit coterie of specialists who can move their show to where conditions are favourable, decamping without much ado from places where regulators have become too nosy or fiscal requests too oppressive. This is true even in this day of huge national banks and powerful central banks. It was even more apparent when those institutions were in their infancy and international bankers roamed the world free, holding sovereign rulers in their power.

The City of London's market supremacy

The monopoly that the Italians held over foreign exchange transactions was reproduced in more modern times by the City of London, where even today the largest concentration of such transactions takes place. British bankers have presided over most of the innovations that have taken place in this market because of the development of modern technologies. Everybody has heard of the homing pigeon

informing the House of Rothschild of the outcome of Waterloo – such was the state of information transmission at the beginning of the 19th century. In the second half of that century, however, technical progress in this field advanced by leaps and bounds, revolutionizing foreign exchange technology. Distance between markets and the slow flow of information had meant that interest rate differentials between different financial markets could remain open for months before being noticed and closed by foreign financial flow. Arbitrage activity had thus been linked, more than to anything else, to seasonal patterns, as one easily discovers by reading contemporary treatises. It was noticed that money was recurrently scarce in one particular month or season in one specific market. Merchants would contribute to fill the gap, if enough profit was expected from transferring money from other places. Alternatively, or concurrently, the Humean specie-flow mechanism would intervene to transform this money scarcity into increased exports and imports. With faster flow of information made possible first by the steam engine, then by the telegraph, then by the international and intercontinental cable, and finally by the radio and telephone, the arbitrage margins between different financial markets came to be closed at speeds that could not be compared with earlier times. This became particularly apparent from the end of the 19th century. However, the vast increase in the speed and volume of foreign exchange transactions which has accompanied innovation in information technology appears to have given just as much chance to foreign exchange speculation, linking together asset markets that had previously remained purely domestic, and by mixing speculation in foreign exchange with commodity speculation in a volume that could not have been attained in previous times.

Inception of foreign exchange controls

Given the advances in information technology, the prevalence of speculation over arbitrage could have generated major international financial crises and so endangered the work of the international economy as much as the advances in information technology had enhanced it. The realization of these dangers, and the palpable loss of monetary sovereignty which the linking of financial markets brought in its train, convinced economic authorities in the period between the two world wars to try to isolate their respective national financial markets by foreign exchange restrictions. Although they were practised with great fervour and severity in Britain too, after the Second World War the City of London managed to persuade the authorities to get rid of them and give the City a chance to go back to its earlier domination of the commodities and exchange markets. In spite of the emergence of New York, Tokyo, and Frankfurt as prime financial markets, the hold British bankers have managed to keep over commodities and exchange transactions is indeed remarkable, and can be considered equal in length of time, breadth and intensity only to that previously exercised on the same activities by the Italian bankers.

Persistence of London's supremacy

This persistence, in the face of the obvious decline of British and previously Italian economic power, is extremely interesting. The commodities and foreign exchange markets seem to have successfully ridden, and to have used to their benefit, the momentous advances in information technology which came in waves in the 19th and 20th centuries. It was expected for these advances to enhance the diffusion of such transactions, by de-concentrating and de-monopolizing them. This of course has happened, but not nearly to the extent that was expected. Technical innovations have also been used to reinforce market

power. Having the dollar as a reserve currency has not seemed to help New York become the home of the commodities and foreign exchange markets. Nor does the demotion of sterling seem to have unduly penalized the City as far as those markets are concerned.

It is obvious that some of the reasons that have brought merchants to congregate in certain places ever since early times persist even in the age of global real-time transactions. Physical proximity and cultural affinity are still powerful enhancers of smooth and successful transactions, as is the confidence that the government will not disturb operations with crippling regulations or with oscillatory behaviour, which destroys certainty. It is perhaps this unique mix of factors that makes for the permanence of foreign exchange markets in certain places. Other pillars of economic power, like a great industrial structure, seem to be somewhat inimical to the permanence of commodities and foreign exchange markets in a given place. Industry certainly generates exports and foreign exchange transactions; but it soon also develops credit needs of its own, and possibly protectionism, both of which work against the permanence of a foreign exchange market. Governments are asked by industry to adopt policies that go against the total freedom that commodities and exchange markets require in order to thrive. Their adoption of such policies induces the community of foreign exchange dealers to pitch its tents elsewhere.

Market growth since the 1980s

This plea for continuity in history must not, however, be to the detriment of realism – and realism imposes a thorough appreciation of the huge increase the foreign exchange market has experienced since the 1980s. As we have already noticed, computer power increased prodigiously in the 1980s and permitted the real-time connection of forex markets across time zones, in a temporal and geographical continuum. Computer power also allowed ever more sophisticated forex contracts to be priced in real time and thus to be executed very rapidly. Among the more exotic contracts, so-called derivatives must be mentioned, which further contributed to increasing the size of forex markets. The size of the market in 2007 is estimated by the Bank for International Settlements (BIS) to be around two trillion dollars.

As we noted above, continuity remains a feature of this huge market. London is still the place where more than 25 per cent of all transactions are processed, with New York coming a distant second, and Tokyo an even more distant third.

And, in spite of the huge size of the market, a few giant international banks concentrate a remarkable percentage of total transactions. The ten largest dealers account for 70 per cent of total transactions. Six of them are commercial banks and four are investment banks.

How the market looks today

At the turn of the millennium, the euro was introduced, an important novel type of currency, not the expression of a sovereign state, but issued by the European Central Bank on behalf of the European Union. This innovation profoundly changed the forex market, as it marked the disappearance of all transactions denominated in the currencies of the European Monetary Union member states, and it meant the arrival of a dominant currency pair the euro/US dollar pair, which in 2004 already accounted, according to the BIS (2004), for 28 per cent of all forex transactions, followed by the US dollar/Japanese yen pair, which accounted for 18 per cent of all transactions. Remarkably, in 2004 14 per cent of all

transactions were still taking place between the US dollar and the British pound. Since 1985 most British merchant banks have been swallowed up by foreign financial institutions, mostly commercial banks, but the British pound, and London, remain foreign exchange favourites.

The present situation in the forex market thus bears an important echo of past power, in the persistence of the British pound, a testimony of recent world economic events, with the arrival and very rapid establishment of the euro as dominant instrument for forex transactions, and of the Japanese yen as the third most important currency. Almost no trace is yet to be seen in the forex market of the meteoric rise of China on the world economic scene. The Chinese currency has recently gained some current account convertibility, but it will be years before it becomes fully convertible. Until then, it will not be able to form important currency pairs with the other dominant currencies. This should come as no surprise if we remember how many years it took the yen to establish itself in the position it now enjoys in the forex market. It should also constitute a final and conclusive piece of evidence in favour of what was noted above on the international foreign exchange community's susceptibility to national fetters and regulations.

See Also

- foreign exchange market microstructure
- gold standard

Bibliography

BIS (Bank for International Settlements). 2004. *Triennial Central Bank Survey 2004*. Basel: BIS.

De Roover, R. 1954. New interpretations of the history of banking. *Journal of World History* 4, 38–76.

Kindleberger, C.P. 1984. *A Financial History of Western Europe*. London: Allen & Unwin.

Mandich, G. 1953. *Le Pacte de Ricorsa et le Marché Italien de Changes un XVIIe Siècle*. Paris: Armand Colin.

How to cite this article

de Cecco, Marcello. "foreign exchange markets, history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000172> doi:10.1057/9780230226203.0596

The New Palgrave Dictionary of Economics Online

Foster, William Trufant (1879 – 1950)

Robert W. Dimand

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Catchings, W.; Foster, W.; Great Depression; growth, models of; Harrod–Domar growth theory; Hoover Plan; monetary policy rules; paradox of thrift; Pollak Foundation; recessions; underconsumption

Article

The educator and heterodox monetary economist William Trufant Foster was born in Boston, Massachusetts, on 18 January 1879, and died in Winter Park, Florida, on 18 October 1950. After his father's early death, Foster worked his way through high school and Harvard University, graduating first in his class in 1901. After teaching at Bates College in Lewiston, Maine, he returned to Harvard to take an A.M. in English in 1904, followed by a Ph.D. from Teachers College of Columbia University. His exceptional success as a teacher of rhetoric and a textbook author, and the vision of an ‘ideal college’ presented in his doctoral dissertation (published in 1911), led to his remarkably early promotion from instructor to full professor at Bowdoin College in Brunswick, Maine, in 1905, and his appointment at the first president of Reed College in Portland, Oregon, in 1910. Foster served as an inspector with the American Red Cross in France after US entry into the First World War. Health problems from overwork, together with controversy over his pacifism, led Foster to resign from Reed College in December 1919. He then became director of the Pollak Foundation for Economic Research, founded in Newton, Massachusetts, by his Harvard classmate Waddill Catchings, an investment banker.
The Pollak Foundation was a vehicle for expounding the heterodox monetary theories of Foster and Catchings, and, through Houghton Mifflin, published their books on *Money* (1923), *Profits* (1925), and *Business without a Buyer* (1927). They held that recessions, such as that of 1920–1, happen because a monetary economy does not automatically generate enough consumption to buy potential output. Saving, which enriches the individual saver, contributes to recessions both by reducing consumption and, through investment, by adding to the potential output to be purchased. Because of this paradox of thrift and their support for counter-cyclical public works, Foster and Catchings had been considered as possible forerunners of Keynesian macroeconomics, while their emphasis on a steadily increasing rate of

investment as a prerequisite for stable growth has been related to later Harrod–Domar growth theory (see Gleason, 1959; Carlson, 1962). Their support for the proposal by Carl Snyder of the Federal Reserve Bank of New York that the volume of currency and credit be increased by a steady four per cent a year has been suggested as an anticipation of monetarist policy rules (Tavlas, 1976). In late 1928, with President-elect Hoover's endorsement and with Foster as his expert witness, Governor Ralph Brewster of Maine submitted to the annual governors' conference a plan for standby credit authorization for $3 billion of federal, state and local public works to be undertaken once a federal board certified the imminence of a recession (Dorfman, 1959; Barber, 1985). Although Foster and Catchings promoted this as the 'Hoover Plan', once the Depression hit President Hoover felt that budget deficits precluded such large-scale counter-cyclical public works.

The Pollak Foundation offered a $5,000 prize for the best adverse criticism of Foster and Catchings's *Profits*, with the competition judged by the two most recent presidents of the American Economic Association, Wesley Mitchell and Allyn Young, and by Owen Young of General Electric. The competition attracted 431 submissions, and the four winning essays were published with a reply by Foster and Catchings as *Pollak Prize Essays* (1927). Also in 1927, the magazine *World's Work* offered a $1,000 prize for the best essay on a series of articles by Foster and Catchings in the magazine. These prizes brought Foster and Catchings considerable professional attention, as did the Pollak Foundation's publication of substantial studies of index numbers by Irving Fisher and of real wages by Paul Douglas (who had been Foster's student at Bowdoin and junior colleague at Reed). Foster and Catchings also found a more popular audience: *The Road to Plenty* (1928), presented as a conversation aboard a train, sold 58,000 copies, while *Progress and Plenty* (1930) reprinted 206 of their 400 two-minute talks on economic problems distributed by the McClure Newspaper Syndicate in 1929 and 1930.

Financial difficulties forced Catchings to withdraw from active participation in the Pollak Foundation during the Depression. Foster continued to direct the foundation, and for three years in the 1930s wrote a syndicated daily newspaper column on economics for the layperson. He served on the Consumers Advisory Board of the National Recovery Administration from 1933 to 1935 (recommended by Paul Douglas) and was an economic adviser at the International Labor Conference in Geneva in 1938.

See Also

- Catchings, Waddill
- monetary cranks
- underconsumptionism

Selected works

1911. *Administration of the College Curriculum*. New York: Columbia University Press.

1923. (With W. Catchings.) *Money*. Boston: Houghton Mifflin.

1925. (With W. Catchings.) *Profits*. Boston: Houghton Mifflin.

1927. (With W. Catchings.) *Business without a Buyer*. Boston: Houghton Mifflin.

1927. (With W. Catchings.) *Pollak Prize Essays*. Newton, MA: Pollak Foundation for Economic Research.

1928. (With W. Catchings.) *The Road to Plenty*. Boston: Houghton Mifflin.

1930. (With W. Catchings.) *Progress and Plenty*. Boston: Houghton Mifflin.

Bibliography

Barber, W. 1985. *From New Era to New Deal: Herbert Hoover, the Economists, and American Economic Policy, 1921–1933*. Cambridge: Cambridge University Press.

Carlson, J. 1962. Foster and Catchings: a mathematical reappraisal. *Journal of Political Economy* 70, 400–2.

Dorfman, J. 1959. *The Economic Mind in American Civilization. Volumes 4 and 5: 1918–1933*. New York: Viking.

Gleason, A. 1959. Foster and Catchings: a reappraisal. *Journal of Political Economy* 67, 156–72.

Tavlas, G. 1976. Some further observations on the monetary economics of Chicagoans and non-Chicagoans. *Southern Economic Journal* 42, 685–92, with comment by J. Davis and reply by Tavlas, 45 (1979), 919–31.

How to cite this article

Dimand, Robert W. "Foster, William Trufant (1879–1950)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000304> doi:10.1057/9780230226203.0597

The New Palgrave Dictionary of Economics Online

fractals

Laurent E. Calvet

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Fractals have become increasingly useful tools for the statistical modelling of financial prices. While early research assumed invariance of the return density with the time horizon, new processes have recently been developed to capture nonlinear changes in return dynamics across frequencies. The Markov-switching multifractal (MSM) is a parsimonious stochastic volatility model containing arbitrarily many shocks of heterogeneous durations. MSM captures the outliers, volatility persistence and power variation of financial series, while permitting maximum likelihood estimation and analytical multi-step forecasting. MSM compares favourably with standard volatility models such as GARCH(1,1) both in- and out-of-sample.

Keywords

Bayesian filtering; Brownian motion; continuous time valuation; fractals; fractional Brownian motion; Lévy-stable processes; long memory models; Mandelbrot, B.; Markov-switching multifractal (MSM); maximum likelihood; multifractal model of asset returns (MMAR); multifractality; regime switching; self-similarity processes

Article

The word ‘fractal’ was coined by the French mathematician Benoît Mandelbrot (1982) to characterize a wide class of highly irregular scale-invariant objects. It originates from the Latin adjective *fractus*, meaning ‘broken’ or ‘fragmented’. The defining characteristic of fractals is that their degree of irregularity remains the same at all scales. This invariance permits parsimonious modelling of complex objects, and has been useful for analysing a wide variety of natural phenomena. The entry reviews the use of fractals in economics and finance, and more specifically their application in the statistical modelling of asset returns, which has been a remarkably active field since the early 1960s.
Consider the price $P(t)$ of a financial asset, such as a stock or a currency, and let $p(t)$ denote its

logarithm. The process $p(t)$ is said to be *self-similar* if there exists a constant $H > 0$ such that for every set of instants $t_1 \leq \ldots \leq t_k$ and for every $\lambda > 0$, the vector $\{p(\lambda t_1), \ldots, p(\lambda t_k)\}$ has the same distribution as $\lambda^H \{p(t_1), \ldots, p(t_k)\}$, that is,

$$\{p(\lambda t_1), \ldots, p(\lambda t_k)\} \stackrel{d}{=} \{\lambda^H p(t_1), \ldots, \lambda^H p(t_k)\}. \quad (1)$$

The constant H is called the *self-similarity* index.

Three classes of self-similar processes have been widely used in finance: the Brownian motion, Lévy-stable processes and the fractional Brownian motion, which are successively discussed. The Brownian motion (Bachelier, 1900), with self-similarity index $H = 1/2$, pervades modern financial theory and notably the Black–Merton–Scholes approach to continuous time valuation. Its lasting success arises from several appealing properties, including tractability and consistency with the financial concepts of no-arbitrage and market efficiency.

The stable processes of Paul Lévy (1924) are characterized by thicker tails than the Brownian motion. They are thus more likely to accommodate the outliers exhibited by financial series, as was pointed out by Mandelbrot in a series of seminal papers (for example, 1963). The increments of Lévy-stable processes are stationary and have stable distributions, where stability refers to invariance under linear combinations (see Samorodnitsky and Taqqu, 1994). Tails are Paretian:

$$\mathbb{P}\{p(\Delta t) > x\} \sim c\, x^{-\alpha} \text{ as } x \to +\infty,$$

with index $\alpha = 1/H \in (0; 2)$. The variance of a Lévy-stable process is infinite, which is at odds with both empirical evidence and mean-variance asset pricing. Furthermore, stable processes have independent increments and thus cannot account for volatility clustering.

The fractional Brownian motion (Kolmogorov, 1940; Mandelbrot, 1965; Mandelbrot and Van Ness, 1968) with $H > 1/2$ is a self-similar process with strongly dependent returns. Increments are stationary, correlated, and normally distributed. Their autocorrelation declines at the hyperbolic rate

$$Cov[r(t); r(t+n)] \sim c\,(2H-1)\, n^{2H-2} \text{ as } n \to \infty,$$

where $r(t) = p(t) - p(t - \Delta t)$ denotes the return on a time interval of fixed length Δt. Hyperbolic

autocorrelation is the defining property of long-memory processes, whose use in economics was advanced by the discrete-time fractional integration approach of Granger and Joyeux (1980). While research on long memory has generally been very fruitful in economics (see Baillie, 1996, for a review), the fractional Brownian motion rarely represents a practical model of asset prices. Specifically, long memory in returns is both empirically inaccurate in most markets (Lo, 1991) and inconsistent with arbitrage-pricing in continuous time (Maheswaran and Sims, 1993). There is, however, abundant evidence of long memory in the *volatility* of returns (for example, Dacorogna et al., 1993; Ding, Granger and Engle, 1993).

In all the above self-similar processes, returns observed at various frequencies have identical distributions up to a scalar renormalization:

$$p(t+\lambda\Delta t) - p(t) \stackrel{d}{=} \lambda^{H} p(\Delta t).$$

Most financial series, however, are not exactly self-similar, but have thicker tails and are more peaked in the bell at shorter horizons. This observation is consistent with the economic intuition that high-frequency returns are either large if new information has arrived, or close to zero otherwise. Thus, self-similar processes do not capture in a single model the most salient features of asset returns.

A partial solution to these difficulties is provided by the multifractal model of asset returns (MMAR; Calvet, Fisher and Mandelbrot, 1997; Calvet and Fisher, 2002a). This approach builds on multifractal measures (Mandelbrot, 1974), which are constructed by the iterative random reallocation of mass within a time interval. The MMAR extends multifractals from measures to diffusions. The asset price is specified by compounding a Brownian motion with an independent random time-deformation:

$$p(t) = B[\theta(t)],$$

where θ is the cumulative distribution of a multifractal measure $\theta(t) = \mu[0, t]$. Returns are uncorrelated and the price *p* is a martingale in MMAR, which precludes arbitrage. The time deformation induces sharp outliers in returns and long memory in volatility. The MMAR also captures nonlinear changes in the return density with the time horizon (Lux, 2001).

The price *p* inherits highly heterogeneous time-variations from the multifractal measure. Its sample paths are continuous but can be more irregular than a Brownian motion at some instants. Specifically, the local variability of a sample path at a given date *t* is characterized by the local Hölder exponent $\alpha(t)$, which heuristically satisfies

$$|p(t+dt) - p(t)| \sim c_t (dt)^{\alpha(t)} \text{ as } dt \to 0.$$

Traditional jump diffusions impose that $\alpha(t)$ be equal to 0 at points of discontinuity, and to 1/2 otherwise. In a multifractal process, however, the exponent $\alpha(t)$ takes a *continuum* of values in any time interval.

Asset returns at different frequencies satisfy the moment-scaling rule:

$$\mathbb{E}[|p(\Delta t)|^q] = c_q(\Delta t)^{\tau(q)+1},$$

which holds for every (finite) moment q and time interval Δt. These moment restrictions represent the basis of estimation and testing (Calvet, Fisher and Mandelbrot, 1997; Calvet and Fisher, 2002a; 2002b; Lux, 2004). The MMAR provides a well-defined stochastic framework for the analysis of moment-scaling, which has generated extensive interest in econophysics (for example, LeBaron, 2001). The multifractal model is also related to recent econometric research on power variation, which interprets return moments at various frequencies in the context of traditional jump-diffusions (for examples, Andersen et al., 2001; Barndorff-Nielsen and Shephard, 2004).

Despite its appealing properties, the MMAR is unwieldy for econometric applications because of two features of the underlying measure: (*a*) the recursive reallocation of mass on an entire time-interval does not fit well with standard time series tools; and (*b*) the limiting measure contains a residual grid of instants that makes it non-stationary.

The Markov-switching multifractal (MSM) resolves these difficulties by constructing a fully stationary volatility process that evolves stochastically through time (Calvet and Fisher, 2001; 2004). MSM builds a bridge between multifractality and regime-switching, which permits the application of Bayesian filtering and maximum likelihood estimation to a multifractal process. Volatility is driven by the first-order Markov state vector $M_t = (M_{1,t}; M_{2,t}; \ldots; M_{\bar{k},t}) \in \mathbb{R}_+^{\bar{k}}$, whose components have unit mean and heterogeneous persistence levels. In discrete time, returns are specified as

$$r_t = \sigma(M_{1,t}M_{2,t}\ldots M_{\bar{k},t})^{1/2}\varepsilon_t, \tag{2}$$

where σ is a positive constant and $\{\varepsilon_t\}$ are independent standard Gaussians.

Volatility components follow independent Markov processes that are identical except for time scale. Given the volatility state M_t, the next-period multiplier $M_{k,t+1}$ is drawn from a fixed distribution M with probability γ_k, and is otherwise left unchanged.

Components differ in their transition probabilities γ_k but not in their marginal distribution *M*. The transition probabilities are tightly specified by $\gamma_k = 1 - (1 - \gamma_1)^{(b^{k-1})}$, which is approximately geometric at low frequency: $\gamma_k \sim \gamma_1 b^{k-1}$. In empirical applications, a unique scalar m_0 typically determines the distribution *M*. The return process (2) is then specified by the four parameters $(m_0, \sigma, b, \gamma_1)$. Since the number of frequencies $\bar{k}$ can be arbitrarily large, MSM provides a tight specification of a high-dimensional state space. The approach conveniently extends to continuous time (Calvet and Fisher, 2001) or a multivariate setting (Calvet, Fisher and Thompson, 2006).

When *M* has a discrete distribution, the state space is finite and MSM defines a stochastic volatility model with a closed-form likelihood. It then bypasses the estimation problems of traditional stochastic volatility settings based on smooth autoregressive transitions. On the other hand when *M* has a continuous (for example, lognormal) distribution, estimation can proceed by simulated method of moments (Calvet and Fisher, 2002b), generalized method of moments (Lux, 2004), or simulated likelihood via a particle filter (Calvet, Fisher and Thompson, 2006).

MSM tends to substantially outperform traditional models both in and out of sample. Calvet and Fisher (2004) thus report considerable gains in exchange rate volatility forecasts at horizons of 10 to 50 days as compared with GARCH-type processes. Lux (2004) obtains similar results with lognormal MSM using linear predictions. Furthermore, bivariate MSM compares favourably with multivariate GARCH under criteria such as the likelihood function, integral transforms and value-at-risk (Calvet, Fisher and Thompson, 2006).

The integration of multifrequency models into asset pricing is now at the forefront of current research. Calvet and Fisher (2005a) thus introduce a parsimonious equilibrium set-up in which regime shifts of heterogeneous durations affect the volatility of dividend news. The resulting return process is endogenously skewed and has significantly higher likelihood than the classic Campbell and Hentschel (1992) specification. Calvet and Fisher (2005b) similarly illustrate the potential of MSM for building parsimonious multifrequency jump-diffusions.

See Also

- regime switching models

Bibliography

Andersen, T., Bollerslev, T., Diebold, F. and Labys, P. 2001. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96, 42–55.

Bachelier, L. 1900. Théorie de la spéculation. *Annales Scientifiques de l'Ecole Normale Supérieure* 17, 21–86.

Baillie, R. 1996. Long memory processes and fractional integration in econometrics. *Journal of Econometrics* 73, 5–59.

Barndorff-Nielsen, O. and Shephard, N. 2004. Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* 2, 1–37.

Calvet, L. and Fisher, A. 2001. Forecasting multifractal volatility. *Journal of Econometrics* 105, 27–58.

Calvet, L. and Fisher, A. 2002a. Multifractality in asset returns: theory and evidence. *Review of Economics and Statistics* 84, 381–406.

Calvet, L. and Fisher, A. 2002b. Regime-switching and the estimation of multifractal processes. Working Paper. Harvard University and University of British Columbia.

Calvet, L. and Fisher, A. 2004. How to forecast long-run volatility: regime-switching and the estimation of multifractal processes. *Journal of Financial Econometrics* 2, 49–83.

Calvet, L. and Fisher, A. 2005a. Multifrequency news and stock returns. Working Paper No. 11441. Cambridge, MA: NBER.

Calvet, L. and Fisher, A. 2005b. Multifrequency jump diffusions: an equilibrium approach. Working Paper. HEC School of Management and University of British Columbia.

Calvet, L., Fisher, A. and Mandelbrot, B. 1997. A multifractal model of asset returns. Discussion Papers 1164–1166. Cowles Foundation, Yale University.

Calvet, L., Fisher, A. and Thompson, S. 2006. Volatility comovement: a multifrequency approach. *Journal of Econometrics* 131, 179–215.

Campbell, J. and Hentschel, L. 1992. No news is good news: an asymmetric model of changing volatility in stock returns. *Journal of Financial Economics* 31, 281–318.

Dacorogna, M., Müller, U., Nagler, R., Olsen, R. and Pictet, O. 1993. A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance* 12, 413–38.

Ding, Z., Granger, C. and Engle, R. 1993. A long memory property of stock returns and a new model. *Journal of Empirical Finance* 1, 83–106.

Granger, C. and Joyeux, R. 1980. An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 15–29.

Kolmogorov, A. 1940. Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen raum. *Comptes Rendus de l'Académie des Sciences de l'URSS* 26, 115–18.

LeBaron, B. 2001. Stochastic volatility as a simple generator of apparent financial power laws and long memory. *Quantitative Finance* 1, 621–31.

Lévy, P. 1924. Théorie des erreurs: la loi de Gauss et les lois exceptionnelles. *Bulletin de la Société Mathématique de France* 52, 49–85.

Lo, A. 1991. Long memory in stock market prices. *Econometrica* 59, 1279–313.

Lux, T. 2001. Turbulence in financial markets: the surprising explanatory power of simple cascade models. *Quantitative Finance* 1, 632–40.

Lux, T. 2004. The Markov-switching multifractal model of asset returns: GMM estimation and linear forecasting of volatility. Working paper. Kiel University.

Maheswaran, S. and Sims, C. 1993. Empirical implications of arbitrage-free asset markets. In *Models, Methods and Applications of Econometrics*, ed. P. Phillips. Oxford: Blackwell.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Mandelbrot, B. 1965. Une classe de processus stochastiques homothétiques à soi. *Comptes Rendus de l'Académie des Sciences de Paris* 260, 3274–77.

Mandelbrot, B. 1974. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics* 62, 31–58.

Mandelbrot, B. 1982. *The Fractal Geometry of Nature*. New York: Freeman.

Mandelbrot, B. and Van Ness, J. 1968. Fractional Brownian motion, fractional noises and applications. *SIAM Review* 10, 422–37.

Samorodnitsky, G. and Taqqu, M. 1994. *Stable Non-Gaussian Random Processes*. New York: Chapman and Hall.

How to cite this article

Calvet, Laurent E. "fractals." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/

article?id=pde2008_F000296> doi:10.1057/9780230226203.0598

The New Palgrave Dictionary of Economics Online

France, economics in (after 1870)

Alain Béraud and Philippe Steiner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

After 1870, classical liberals gradually lost their influence. Political economy began to be taught in university faculties of law, and also in some of the engineering schools. This laid the foundations for a long-standing divide between two groups of economists. Professors of political economy in the law faculties often inclined to an institutionalist approach, and opposed the mathematical approach to political economy that economic engineers and some mathematicians adopted. This antagonism abated after the Second World War as French economists strengthened their relations with foreign colleagues.

Keywords

Aftalion, A.; Allais, M.; Aupetit, A.; Austrian School; Bachelier, L.; Balasko, Y.; Barrère, A.; Baudin, L.; Benassy, J. -P.; Bertrand, J.; birth rate; Boiteux, M.; Borel, E.; Boyer, R.; Braudel, F.; Brownian motion; budgetary equilibrium; Cartan, H.; Chamberlin, E.; classical liberalism; Colson, C.; Condillac, E.; consumer price index; contract theory; Convention School; correlation analysis; Cournot, A.; Courtin, R.; creative destruction; Currency School; D'Aspremont, C.; Debreu, G.; decisions under uncertainty; demand for money; differentiability; disequilibrium theory; Divisia, F.; Dos Santos, R.; Drèze, J.; econometrics; economic growth; efficient markets; excess labour supply; existence of equilibrium; expectations; family planning; general equilibrium; Gérard-Varet, L. -A.; German Historical School; Gide, C.; golden rule; Grandmont, J. -M.; Gruson, C.; Guesnerie, R.; Hicks, J.; identification; indicative planning; inflation; innovation; Institut de Sciences Économiques Appliquées; institutionalism; intertemporal choice; IS–LM model; Jevons, W.; Juglar, C.; Keynes. J. M.; Labrousse, E.; Lacaillon, J.; Laffont, J. -J.; Landry, A.; Lenoir, M.; Leroy-Beaulieu, P.; Lescure, J.; Lindahl, E.; local equilibrium; Malinvaud, E.; March, L.; Marchal, A.; Marchal, J.; Marjolin, R.; Marxism; Massé, P.; mathematical economics; method of moments; minimax theorem; Modigliani, F.; Molinari, G. de; monetary theory of crises; money supply; mortality; Nogaro, B.; oligopolistic competition; Orléan, A.; overlapping generations model; partial equilibrium; Pearson, K.; Perroux, F.; probability; purchasing

power parity; quantity theory of money; regulation; Regulation School; risk; Rist, C.; Roy, R.; Rueff, J.; Saillard, Y.; Say, J. -B.; Say's Law; shadow prices; Simiand, F.; system of curves; tâtonnement; Tirole, J.; transaction costs; transfer problem; unemployment; Villey, D.; von Neumann, J.; wage determination; Walras, L.; Younès, Y.

Article

The publication of Léon Walras's *Éléments d'économie politique pure* in 1874 marks an important turning point in the history of economic analysis. But for many years his ideas remained misunderstood. Recognition of the importance of his work on the part of French economists followed a lengthy and difficult period, in which the publication of Maurice Allais's *À la recherche d'une discipline économique, l'économie pure* in 1943 marks a vital stage. Allais introduced the analysis of risk and intertemporal choice to the theory of general equilibrium and in this way posed new questions to which Gérard Debreu, Marcel Boiteux, Edmond Malinvaud and many others would respond. Nonetheless, many French economists had considerable reservations about the theory of general equilibrium. They favoured an emphasis upon the role of institutions, and the need to integrate the various elements of the social sciences – economics, sociology and history – if economic phenomena were to be understood.

From 1870 to 1943

In the years after 1870 the domination of the liberal school was increasingly questioned, and this was largely the consequence of institutional developments (Le Van-Lemesle, 2004). The teaching of political economy was introduced into the faculties of law in 1877, but the professors in law in charge of this teaching progressively became scientifically independent. In 1887 they founded the *Revue d'Économie Politique* so that the new political economy might be more widely diffused, and this quickly became far more influential than the liberal *Journal des Économistes*.

Classics liberals and institutionalists

The best known of the last classical liberals, Gustave de Molinari and Paul Leroy-Beaulieu, sought to defend very different positions. Liberals had maintained that the state should limit itself to the provision of individual security but de Molinari (*L'évolution politique et la révolution*, 1884) argued that it was necessary to go much further. All branches of production, including the judiciary, the police and defence, should be freed from state control. If a need for security exists and if the state does not foresee it, then this need will be met by private initiative, and so much the better. Leroy-Beaulieu did not challenge the principle that a state had its prerogatives and that it would exercise them. However, while Molinari defended the classical theory of distribution, Leroy-Beaulieu (*Essai sur la répartition des richesses*, 1881) thought it necessary to abandon this theory. The consequences which it foresaw – a fall in the rate of profit, an increase in the rate of rents, and a reduction of wages to subsistence levels – were refuted by factual evidences: wage rates were increasing, and rents were diminishing in proportion. Institutionally Leroy-Beaulieu belonged to the group of older classical liberals, but he abandoned the propositions basic to this school.

Charles Gide occupied a leading place among the professors of the law faculties. He was a staunch eclectic, which led him to reject extreme theses in favour of an intermediate synthesis. In studying prices and distribution he made use of ideas borrowed from Jevons and Walras, but played down their contribution. If Jevons's analysis of value was ingenious, it was nonetheless not new; Condillac had long before made clear that the utility of an object determined its value. Gide was somewhat reluctant to make use of the notion of marginal productivity, since he did not consider that the distribution of revenues to be solely determined by economic factors, and he argued (*Principes d'économie politique*, 1901) that social relations among agents also played a part.

Adolphe Landry and François Simiand were part of a very small group of philosophers educated at the École Normale who chose to become economists. In his *Révolution démographique* (1934) Landry distinguished three types of regulation as of importance to the study of demographic development. First, under the *ancien régime*, parents did not concern themselves with the consequences of the birth of children. Mortality played the principal role in regulating the population. Second, during the transitional phase, men and women chose their age of marriage so that they might maintain the standard of living to which they had become accustomed, and there was no voluntary birth control in marriage. Third, in modern times, on the contrary, the timing and number of births had become a matter of choice. Landry used this argument to persuade parliament to vote through, in 1932, 1939 and 1946, the three laws which determine the allocations of family support: for if the birth rate is the product of choice, then one can hope to end demographic decline with the aid of a system of financial incentives.

French positive economics developed with the work of François Simiand (*La méthode positive en science économique*, 1912). He rejected both the approach of the German Historical School as well as what he termed 'orthodox' economics, referring in this way to French liberals, the Austrian School and mathematical economics. The German Historical School, he suggested, lacked principles and had produced nothing but an empty accumulation of knowledge. 'Orthodox' economists constructed theories that were poorly founded, since they drew upon incomplete or implicit observations. Simiand, by contrast, made use of long statistical series, analysing them in terms of models that described the behaviour of social groups. He applied this method to the study of the development of wages and prices in his major works of the 1930s (*Recherches anciennes et nouvelles sur le mouvement des prix du 16ème au 19ème siècle*, 1932, et *Le salaire, l'évolution sociale et la monnaie*, 1932). In these works he argued that variations in the money supply drove the cycle and that cyclical fluctuation was a necessary part of economic progress. This approach influenced Ernest Labrousse (*Esquisse du mouvement des prix et des revenus au 18ème siècle*, 1933) who, on the basis of meticulously constructed statistical series, put forward a simple theory of the crisis of the *ancien régime* as engendered by the agricultural cycle: bad harvests brought about a rise in the price of wheat, consumers spent an increasing proportion of their revenues on agricultural goods and so the crisis was transmitted to industry.

Albert Aftalion (*Les crises périodiques de surproduction*, 1913) and Jean Lescure (*Des crises générales et périodiques de surproduction*, 1906) took their inspiration from Say and their analysis of crises from Juglar. They retained Say's Law of Markets. From Juglar they drew three lessons. Their analysis rested upon study of empirical data. They used price movements to determine the phases of the cycle. The crisis was defined as the point at which prices ceased rising, inevitably followed by a fall in prices – it was only one phase of the cycle. But whereas Juglar put forward a monetary theory of crises, Aftalion and Lescure proposed a real theory. At the bottom of a recession production had difficulty satisfying needs. The marginal utility of consumer goods and their prices would thus rise. To meet this demand,

machinery is needed. The price of machinery rises and in turn stimulates production. However, when the new production goods come into service consumer goods become over-abundant. Their final utility and value collapse and this has repercussions for the price of machinery. The crisis becomes almost, or entirely, general. Lescure placed the emphasis on the role of profits and on the interdependence between activities. At the end of an expansionary phase, costs rise faster than prices and new enterprises that have paid a high price for their means of production face losses. Their insolvency brings about the crisis, which spreads from one branch to another. The crisis is not general, but generalized.

Over a lengthy period, French economists had criticized the version of the quantity theory of money advocated by partisans of the Currency School, and this continued after 1870. Bertrand Nogaro (*Contribution à une théorie réaliste de la monnaie*, 1906) noted that money was the object neither of demand nor supply; the general price level is not determined, as the quantity theory supposed, by the relation between the money stock and desired cash holdings, but by global demand for goods, or as argued by Aftalion (*Monnaie, prix et change*, 1927), by the relationship between monetary revenue and the volume of production. The consequences of a variation in the stock of money depended for its effect upon the demand and supply of goods, and hence on the way that it is introduced into the system. Nogaro and Aftalion rejected the idea that variations in the price of goods explained variations in the exchange rate. The direction of causality was not necessarily from prices to exchange rates. The current exchange rate depended upon the expected future rate and, since it affected producer costs and agents' revenues, domestic prices are determined by psychological factors.

Walras, the mathematicians and the statisticians

For many years both mathematicians and engineers had reservations about the idea of general equilibrium. They considered partial equilibrium quite adequate for the study of most problems. Walras's use of mathematics seemed quite superfluous. Even when the importance of Walras's work gradually became more generally accepted, his successors remained critical of his methodology. Instead they shared Pareto's view that the criterion of a theory's truth lies in its correspondence to reality. They did not attempt to resolve the theoretical difficulties presented by the Walrasian construct. Instead, they were interested in understanding the instruments which permitted the analysis of facts while using economic theory. The procedure followed by Albert Aupetit, the leading disciple of Walras, is quite typical. His dissertation, *Essai sur la théorie génerale de la monnaie* (1901), presents itself both as a development of Walrasian monetary theory and as verification of its empirical relevance.

The tradition of engineer–economists continued with Clément Colson. His works (*Cours d'économie politique*, 1901–7) drew more on Dupuit's analysis than on Walras's, but he encouraged François Divisia, René Roy and Jacques Rueff to study Walrasian theory since he was aware of the importance of the interdependence of markets. It was not possible to study the determination of wages independently of that of the rate of interest. Since labour and capital are substitutes, the proportions in which they should be employed depended both upon the wage rates and interest rates. Here one can see at work the fundamental idea that had driven Walras to use mathematics and make use of models of general equilibrium.

Divisia's analysis of monetary phenomena illustrates this connection of theory to empirical research. It had sometimes been thought that the quantity equation implies that prices vary with the quantity of money. Divisia rejected this idea, arguing that the transactions equation is an identity. Appealing to

statistical observation for verification is an absurdity, but it does allow the definition of what should be an indicator of prices. Weights are quantities of goods and services exchanged, not quantities produced or consumed. Divisia (*L'indice monétaire et la théorie de la monnaie*, 1925–6) explained that it is not possible to set these weights; the index should be a chain index. In order to determine the value of money in 1900 relatively to its value in 1800, it is not enough to know the quantities of goods and services bought in 1800 and 1900, all the intermediate values should also be known. René Roy followed the same line of argument. He introduced (*De l'utilité, contribution à une théorie des choix*, 1942) the idea of the indirect utility function to demonstrate that the consumer price index is the number by which primary prices have to be multiplied to render the satisfaction of an individual (under the assumption of constant monetary income) equal to his satisfaction at current prices.

Even while invoking Walras, Rueff appeared above all to be the defender of classical arguments against attack by Institutionalists and by Keynes. Contrary to Nogaro, he argued (*Théorie des phénomènes monétaires*, 1926) that price variations are determined by effective holdings of cash relative to desired holdings. He based his arguments on a reformulation of the theory of purchasing power parity in dealing with the problem of transfers. Contrary to Keynes, he maintained that the sole levy that would enable the Germans to pay reparations to France would be a rise in taxes. Of course, in the flexible exchange rate regime that was then prevailing, the D-Mark would depreciate and the wage rates of German workers expressed in foreign currency would diminish; but the price of German products would diminish in proportion, so that real wages remained unchanged. It was, however, his analysis of unemployment that made him famous. Following the First World War, unemployment rose in Great Britain and changed in nature: instead of being cyclical, it became permanent. Drawing upon the relation he had put forward between unemployment and the real wage rate, Rueff suggested that this development followed from the emergence of a system of unemployment relief which checked the fall in the money wages despite the existence of an excess labour supply.

The establishment of a more direct link between theory and empirical research involved the development of statistics. Lucien March was the first Frenchman to make Karl Pearson's work known, and he took (*Les principes de la méthode statistique*, 1930) from Pearson three fundamental techniques: the method of moments, the system of curves, and correlation analysis. Marcel Lenoir's 1913 doctoral dissertation (*Etudes sur la formation et le mouvement des prix*), which dealt with price formation and price movements, marked the beginning of econometrics. He not only made careful use of correlation and regression, but he posed, and resolved, the problem of identification. If one had a time series of quantities exchanged and their prices it was possible to plot a path on a graph, but not to interpret this graph as a supply or a demand curve. Lenoir, using moving averages, plotted the long-run trend of cyclical fluctuations. He then calculated regression coefficients and interpreted his results by introducing the idea that short-run variations in prices reflected shifts of the demand curve, while long-term variations were more indicative of shifts in the supply curve and the influence of monetary factors.

Apart from the engineers, French mathematicians took hardly any interest in political economy. Two of them however, Louis Bachelier and Émile Borel, did, at the beginning of the 20th century, make fundamental contributions to the development of economic science. The arguments advanced in Bachelier's *Théorie de la speculation* (1900) lie at the origins of the mathematical analysis of finance: here can be found the essentials of the theory of efficient markets and the premises of the notion of Brownian motion which he developed in 1913. Borel's point of departure is the analysis made by Joseph Bertrand of the game of baccarat in his *Calcul des probabilités* (1889). Bertrand highlighted the

existence of a strategic interdependence between the players similar to that which, he suggested, Cournot had wrongly ignored in his analysis of duopoly. But Borel in turn accused Bertrand of overlooking the case where players determined their strategy by drawing lots. He argued that, if one were to reveal the psychological mechanism governing choices, then it had to be connected to the notion of probability: at each moment, each player chooses his or her strategy with a given probability. The player's mathematical hope of gain depends on the way in which the probabilities are allocated to each alternative. In a symmetric game no information can provide one of the players with the certainty of the gain advantage. The best strategy is to distribute probabilities so that one does not lose whatever the opponent does. Borel demonstrated in *La théorie du jeu et les équations intégrales à noyau symétrique* (1921) that a solution exists for a game in which two players could choose between three ways of playing. Nonetheless, it was von Neumann who in 1928 demonstrated at a general level the theorem of the minimax. Jean Ville suggested in 1938 a more simple demonstration, and showed that the result applied to continuous variables.

From 1943 to the present day

The publication in the early 1940s of books by Robert Marjolin (*Prix, monnaie et production*, 1941), Maurice Allais (*À la recherche d'une discipline économique*, 1943), François Perroux (*La valeur*, 1943) and by Jacques Rueff (*L'ordre social*, 1945) all testify to a shift in the analyses of French economists. But if they were all certain of the need for a break with traditional liberalism, their work led in different, even contradictory, directions.

Liberals, Keynesians and Institutionalists

If, despite the efforts of Daniel Villey and Louis Baudin, the heritage of French classical liberalism was fading, after 1940 liberalism experienced a renaissance, but it was a liberalism quite different from that of Molinari and Leroy-Beaulieu. Its most typical representatives, Rueff and Rist, admired Walras for the manner in which he showed that variations in prices always led to equilibrium, since they continued up to the point where they stabilized. René Courtin took up exactly this point in his *Cours de théorie économique* (1950) when he accused Keynes of having assumed absolute rigidity of prices, and of nominal wages in particular. If such a rigidity exists (a doubtful interpretation of Keynes's book), it is never absolute, for while it is capable of explaining unemployment in the short run, it cannot explain its persistence. According to Rueff, the modern social order rests on two institutions: property rights which prevent appropriation by violence, and the market, with its characteristic flexibility of prices which mutually adjust to the point where equilibrium is reached. A property right should be understood as a pool of value, of known volume, which can be filled with whatever wealth offered on the market at the behest of its owner. In so far as the value of this pool corresponds to the value of the goods that it contains, one can say that the right is a real one. But if this is not so, then the right is false. Rights of this sort can be introduced in a number of ways. The simplest example is that of a budget deficit financed by the creation of money. The state, by buying goods or leasing services, creates rights for its creditors. When these expenditures are covered by taxes the rights are real; but if they are not so covered then they are false rights – state creditors hold paper claims to wealth which does not exist. Inevitably, policies of this kind lead to inflation. And in so conducting itself the government weakens the judicial system that

protects the social order. Some individuals are not able to provide the rights which they hold with the volume of their choice. The unconditional character of the law is irremediably compromised.

Soon after the publication of the *General Theory*, several works inspired by Keynes appeared, in particular the works of Marjolin (*Prix, monnaie et production*, 1941), Claude Gruson (*Esquisse d'une théorie générale de l'équilibre économique*, 1949) and Alain Barrère (*Théorie économique et impulsion keynésienne*, 1952). They touched on Keynes's work in a very specific manner. Their common problem was the construction of dynamic analysis. They had doubts about the analysis that Keynes had developed in the *General Theory*, but his book had the merit of addressing – even if not fully consciously – the economic problems of growth, and the most fundamental economic policy issue, that of growth coordinated by deliberate and conscious policy. They showed little interest in the models that Modigliani and Hicks had introduced to analyse short-term monetary and budgetary policy. The IS–LM model was for many years neither taught nor discussed in France.

The majority of university economists remained distanced from both liberal arguments and Keynesian ideas. They argued that it was barely possible to understand economic choices without studying its social, cultural and institutional determinants. They argued for a concrete and positive economics closely linked to other social sciences such as sociology and history. The will to renew the link to positive economics was expressed with the foundation in 1950 of the *Revue Économique*, which quickly became the most important of French academic journals. Aftalion was among the founders, alongside historians such as Braudel and Labrousse. This conception of economic science led them to place the study of structure, defined as an ensemble of relations characteristic of a social and economic system – following the example of André Marchal's *Systèmes et structures* (1959) – at the centre of their studies. This method was applied in particular to the analysis of distribution (as in Jean Marchal and Jacques Lecaillon, *La répartition du revenu national*, 1958–70), production structures, spatial organization and the relationships between national economies.

François Perroux played an important role after the Second World War. He created and directed the Institut de Sciences Économiques Appliquées, which for many years was the leading centre for economic research in France. He became a professor at the Collège de France, the most prestigious French scientific institution. Perroux was open to different influences, and which sometimes appeared to conflict. His first works, in particular his book *La valeur*, revealed the influence Austrian marginalists had played in his thinking. *Économie appliqué*, the journal that he edited, was one of the important channels for the diffusion of Keynes' thinking in France. But his masters were Chamberlin and Schumpeter. He admired Schumpeter as the theorist of innovation, and of creative destruction. What interested him about Chamberlin was the detailed criticism of hypotheses regarding pure and perfect competition. He proposed a general theory of the impact of domination at the level of enterprise, industry and national economy. He saw in this analysis a first and indispensable step towards a much larger synthesis between a theory of the economy and a theory of force, power and of constraints.

And so following the Second World War French economists sought to reconnect with the tradition of positive economics founded with Aftalion and Simiand. This institutionalist project collapsed at the end of the 1960s when the new generation turned to either Marxism or the theory of general equilibrium. Nonetheless, institutionalism has remained an active force within French political economy up to the present day with the Convention School (André Orléan, *Analyse économique des conventions*, 1994) and the theory of regulation (Robert Boyer, *La théorie de la régulation: une analyse critique*, 1986; Boyer and Saillard, *Théorie de la régulation: l'état des savoirs*, 1995). In both schools there is agreement that

political economy has to collaborate with other social sciences, history and sociology. The conventionalists are interested in situations where existing prices are insufficient to coordinate the activity of agents on account of uncertainty concerning the future and the quality of products. It is necessary to take account of conventions, understood as legitimate routines of interpretation on the part of agents. The theory of regulation has much larger ambitions: the development of an economic theory which presents an alternative to orthodox theory. Its key concept is the mode of regulation, that is, the manner in which several institutions (the financial system, the wage relation, forms of competition) join together to form a system. Hence the Fordist mode of regulation is characterized by oligopolistic competition, the development of credit, the growth of productivity in mass production and the indexation of wages to gains in productivity. The theory of regulation addresses itself to the description and explanation of different forms of regulation and the specificity of the crises which characterize it.

Reformulations of general equilibrium theory

Divisia and Roy had not profoundly modified the basic framework of Walrasian analysis. In 1943 Allais had put forward some new directions for research by introducing intertemporal economies, where each good is defined by the location and date at which it becomes available, and in which there exist markets for all future goods. He demonstrated, making use of Walrasian *tâtonnement*, that the equilibrium was stable. He established the two propositions fundamental to the theory of welfare. In 1947, in *Économie et Intérêt*, he developed a synthesis combining the theory of interest, prices and money. He put forward the first proof of the golden rule. He noted that the existence of transaction costs explained why agents hold money rather than stocks and shares. On this basis he showed that the demand for money is a function of income and of the rate of interest. To illustrate the influence of basic elements of the theory of interest, he introduced a model of overlapping generations. The third fundamental contribution by Allais was the development of a theory of decisions in a state of uncertainty. He showed in *Le comportement de l'Homme rationnel devant le risque* (1953) that, if one wants to account for the behaviour of agents, it is necessary to take account of characteristics of the index of utility other than its average. Finally, in his *La théorie générale des surplus* (1981), Allais put forward a complete modification of the frame of reference: in place of the Walrasian market model he put forward a model of markets founded upon the decentralized search for realizable surpluses.

Debreu was trained as a mathematician; he had been the pupil of Henri Cartan and through him had come under the influence of the Bourbaki group which had an axiomatic approach to mathematics. It was through the study of Allais's book *À la recherche d'une discipline économique* that he was initiated into the theory of general equilibrium. If Debreu found in his reading of Allais the point of departure for his own studies, the reorientation is significant. Up to that point economic analysis consisted in maximizing differentiable functions and deriving the characteristics of maxima from first-order conditions. Debreu abandoned this approach; differential calculus gave way to topological arguments which quite clearly increased the generality and simplicity of theory. But it was not only the mathematical tools that changed. Allais had maintained that 'in the last analysis it was experience, and only experience, which could determine whether a theory had merit or whether it must be rejected' (Allais, 1943, p. 116). In the work of Debreu, the concern for rigour dominates: he stipulated that the axiomatic form of analysis or of theory was, strictly speaking, logically entirely disconnected from its interpretations. In his *Théorie de la valeur* (1954) Debreu took up the analytical framework

employed by Allais in 1943. He demonstrated the existence of an equilibrium and established the two theorems of welfare through the use of convex sets. But he refrained from discussing the problem of stability which was central to Allais's preoccupations. The uniqueness of equilibrium posed a problem. At the end of the 1960s it became evident that the hypotheses under which the uniqueness of equilibrium could be established were too restrictive and that it was necessary to make do with an analysis of local equilibrium. Debreu (1970) demonstrated that, using the hypothesis of differentiability, the number of economies that did not have a local equilibrium was ‘negligible’, that is, ‘contained in a closed set of Lebesgue measure zero’. This result, gained by using the concepts and techniques of differential topology, was the origin of the theory of regular economies that Yves Balasko in particular developed.

Following the Second World War the problems of reconstruction, of developing a system of indicative planning, and the management of public enterprises lent Allais, Pierre Massé and their pupils occasion to apply the theoretical propositions that they had elaborated. Among the contributions that French economists made during this period to the theory of the efficient allocation of resources and to the study of public policy, Jacques Drèze (1964) underlined the importance of two themes: the management of public enterprises and the analysis of the conditions under which the accumulation of capital is socially effective.

Edmond Malinvaud (1953) explicitly introduced time into the model of general equilibrium. From this he derived an analysis of the determination of the rate of interest and the meaning that it gives to the proposition that the rate of interest is equal to the marginal productivity of capital. One can only regret that the economists who became involved in the controversy that led to the theory of capital did not always record the results that they arrived at.

Marcel Boiteux (1956) suggested a new approach to the management of public monopolies constrained by budgetary equilibrium. He sought to define a rule for the management of public monopolies by adding to natural connections a new constraint: the budgetary equilibrium. He then defined the shadow prices which were the solution to the problem. Public monopolies should maximize their profits in terms of these shadow prices. The gap between real prices and shadow prices is proportional to the inverse of the price elasticity of compensated demand. While Dupuit and Colson referred to marginal costs, Boiteux took account of shadow marginal costs and prices.

What remains to be determined is whether the enterprise or the regulator is the better at determining tariffs. Jean Tirole and Jean-Jacques Laffont analysed systematically this type of problem by using the theory of contracts. The central idea is that information at the disposal of the managers of a public monopoly is greater than that available to the regulator. It is therefore necessary to determine the nature of the contract which the regulator is able to propose to the enterprise to minimize the costs of production of the good which it produces, while explicitly taking account of the capacity of the agent to manipulate the information.

In Debreu's model, all agents have, *ab initio*, access to a complete system of forward markets and adjustments are made solely by price. All contracts are concluded on the starting date; there is no incentive to reopen markets at a later date. The model is essentially atemporal; the role of money cannot be explained, nor the existence of a market for stocks nor the underemployment of resources. Lindahl and Hicks suggested that a partial equilibrium framework was appropriate for dealing with this kind of problem. Michel Grandmont, in a series of articles published in the course of the 1970s, took up and then systematically developed this notion by assuming that agents formed, at every moment, expectations of the future states of the economy that were not necessarily realized. It was in this

framework that, in the 1970s, Jean-Pascal Benassy, Drèze, Malinvaud and Yves Younès built their theory of disequilibrium. More recently, this framework was used to study the relations between value and money (Grandmont, *Money and value*, 1983), between competition and underemployment (Claude D'Aspremont, Louis Gérard-Varet, Rodolphe Dos Santos, *On Monopolistic Competition and Involuntary Unemployment*, 1990, and Benassy, *The Economics of Imperfect Competition and Underemployment*, 2002) and rational expectations (Roger Guesnerie, *Assessing Rational Expectations*, 2001).

Until the 1970s, French economics had a flavour of its own with engineer–economists interested in planning and the management of public enterprises, and with many professors still following the French institutionalist tradition. Thereafter, this distinctiveness disappeared and, with the exception of the Regulation School, French economists became thoroughly integrated into an international economics profession.

See Also

- Allais, Maurice
- Allais paradox
- Aupetit, Albert
- Bachelier, Louis
- Bertrand, Joseph Louis François
- Braudel, Fernand
- Debreu, Gerard
- Divisia, François Jean Marie
- divisia index
- Gérard-Varet, Louis-André
- Gibrat, Robert Pierre Louis
- Juglar, Clément
- Laffont, Jean-Jacques
- Leroy-Beaulieu, Pierre-Paul
- Massé, Pierre
- Perroux, François
- Roy, René François Joseph
- Sauvy, Alfred
- Walras, Léon

Bibliography

Allais, M. 1943. *À la recherche d'une discipline économique. Première partie, l'économie pure*. Saint-Cloud, chez l'Auteur, deuxième édition sous le titre *Traité d'économie pure*. Paris, Imprimerie Nationale, 1952; troisième édition, Paris: Clément Juglar, 1994.

Aréna, R. 2000. Les économistes français en 1950. *Revue économique* 51, 969–1007.

Boiteux, M. Sur la gestion des monopoles publics astreins à l'équilibre budgétaire. *Econometrica* 24, 22–40.

Debreu, G. 1954. *Theory of Value*. New York: John Wiley.

Debreu, G. 1970. Economies with a finite set of equilibria, *Econometrica* 38, 387–92.

Divisia, F. 1951. *Exposés d'économique, Introduction générale. L'apport des ingénieurs français aux sciences économiques*. Paris: Dunod.

Dockès, P., Frobert, L., Klotz, G., Potier, J.-P. and Tiran, A. 2000. *Les traditions économiques françaises*. Paris: CNRS éditions.

Drèze, J. 1964. Some postwar contributions of French economists to theory and pubic policy: with special emphasis on problems of resource allocation. *American Economic Review* 54(4), Part 2: Supplement: Survey of foreign postwar developments in economic thought, 1–64.

Faccarello, G., ed. 1998. *Studies in the History of French Political Economy, from Bodin to Walras*. London and New York: Routledge.

Greffe, X., Lallement, J. and De Vroey, M., eds. 2002. *Dictionnaire des grandes œuvres économiques*. Paris: Dalloz.

Le Van-Lemesle, L. 2004. *Le Juste ou le Riche. L'enseignement de l'économie politique, 1815–1950*. Paris: Comité pour l'histoire économique et financière de la France.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21, 233–68.

Steiner, P. 2000. La Revue Économique, 1950–1980: la marche vers l'orthodoxie économique. *Revue Économique* 51, 1009–58.

Zylberberg, A. 1990. *L'économie mathématique en France, 1870–1914*. Paris: Economica.

How to cite this article

Béraud, Alain and Philippe Steiner. "France, economics in (after 1870)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan,

2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000307> doi:10.1057/9780230226203.0600

The New Palgrave Dictionary of Economics Online

France, economics in (before 1870)

Alain Béraud and Philippe Steiner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

From the late 17th century onwards, French economists were major contributors to the rise of economic liberalism, developing many of the analytical tools of political economy. After the Revolution, their major concern was the growth and stability of what they called ‘industrial society’; and a distinction arose between those who claimed that such a society needed to be regulated (the Saint-Simonians) and those in favour of a more decentralized and market-oriented system. After 1848, French economists became deeply involved in the struggle against socialism, and devoted a great deal of energy to the diffusion of sound principles of political economy.

Keywords

animal economy; Arrow, K. J.; Bastiat, F.; Batbie, A.; Bicquilley, C. -F.; Blanc, L.; Boisguilbert, P.; calculus; cameralism; Canard, N.; Cantillon, R.; Chevalier, M.; Child, J.; Cobden–Chevalier Treaty; Colbert, J. -B.; Condillac, E. B.; Condorcet, M. de; d’Alembert, C.; decreasing returns to capital in agriculture; Diderot, D.; division of labour; Dupuit, A. -J.; economic governance; entrepreneurship; fiduciary money; Forbonnais, F. V. de; France, economics in; Garnier, J.; Gournay, V. de; impossibility theorem; industrialism; Isnard, A. -N.; Law, J.; Le Mercier de la Rivière, P. -P.; Mably, Abbé de; mercantilism; Mirabeau, Comte de; Molinari, G. de; Montchrestien, A. de; Montesquieu, Comte de; mutualism; natural price; Necker, J.; net product; overproduction; Physiocracy; political economy; probability; proportionate price; protectionism; Proudhon, P. -J.; public debt; Quesnay, F.; Ricardo, D.; right of association; Roederer, P. -L.; Rossi, P.; Rousseau, J. -J.; Saint-Simonians; Say, J. -B.; single tax; Sismondi, J. C.; social mathematics; socialism; subjective theory of value; taxation in kind; Turgot, A. R. J.; Vauban, S.

Article

From the end of the 17th century to 1755

The term *économie politique* first appeared in French in Antoine de Montchrestien's *Traicté de l'œconomie politique* of 1615. However, during the 17th century there was no French counterpart to English mercantilist thought, nor the kind of economic administration formed on cameralist principles found in Austria and Germany, despite Colbert's attempt to promote the wealth and power of the monarchy through the regulation of commerce. Censorship and the weakness of the French merchants as a class could explain this situation. By the end of the 17th century reflection on economic matters was just beginning, and the monarchy was increasingly conscious of the gravity of the problems that recurrent dearth and high levels of debt represented. This created the conditions for the questioning of economic policy with respect to both the provisioning of markets and taxation.

Vauban argued in his *Dixme royale* (1707) that the principal cause of the monarchy's economic distress was the way its fiscal system was organized. Taxation, he wrote, should be raised in kind as a proportion of the gross yield from the annual harvest. Such a tax would therefore be proportional to agricultural wealth. For commerce and industry he anticipated light taxes that could be passed on in trade.

Boisguilbert's proposal (*Le détail de la France*, 1695) goes much further, even though his attention was likewise directed to taxation. His theory of markets derived from Jansenist moral philosophy, according to which a society in which behaviour was founded upon interests would also be ordered in the same way as a society composed of charitable and pious people. Boisguilbert endorsed laissez-faire as the sole condition permitting the emergence of the 'proportionate price', a price at which each gained from participating in exchange and in which each party to the exchange adhered to his budget constraint. He argued that both good and bad harvests disrupted economic activity because they would bring about violent price changes if, as was then the case in France, free competition were absent. Since the wheat market determined the level of agents' revenues (the remuneration of agricultural capital, the payment of rents), variations in the price of wheat affected other markets. Moreover, the price of wheat was vital to the subsistence of populations. Expectations on the part of agents, whether justified or not, disturbed the economy, and government intervention was not capable of stabilizing the market since such intervention was in turn perceived to be the sign of an even more serious crisis.

After the death of Louis XIV in 1715, the regent accepted John Law's arguments concerning financial policy. According to Law (*Considérations sur le commerce et l'argent*, 1720), France's poor economic performance was due to an inadequate money supply. In 1716, he founded a bank which had the creation of paper money as its principal function; this paper money was supposed to substitute for coins and to permit a refinancing of government debt. Here Law's ideas were at variance with those of Boisguilbert, but Law also went on to argue that money could also be backed by land or by shares, that is, by productive capital. These ideas were given shape with the formation of a commercial company that was granted an exclusive right to trade with Louisiana. The company's shares could be purchased only with *billets d'Etat* (government securities) at their face value instead of being discounted about 70 per cent, but the public could hope for capital gains if the company's trade was well managed. The company gained in this way an exclusive right to the exploitation of vast wealth, and the state transformed its floating debt into long-term debt with a lower interest rate. The merging of the company and the bank permitted monetary expansion and at the same time boosted the value of the company's own shares. At the end of 1719 Law became Comptroller General of Finance – money issue was strong (around a million *livres*) and the rate of interest touched a low point of two per cent. The price of the company's

own shares was stabilized by an office which intervened in the market. The system collapsed as soon as agents sought to exchange their shares and securities for cash. Law's collapse had a lasting impact. The chance of modernizing the public finances had been missed, and for the entire 18th century the collapse weighed heavily on the capacity of the French monarchy to finance its military conflicts with Britain. In addition, a marked suspicion of fiduciary money and banking prevailed right up to the Revolution.
Discussion of monetary matters and Law's system continued in the early part of the 18th century, but gave way to an interest in commerce from the perspective of the legislator, as for instance in Richard Cantillon's *Essai sur la nature du commerce en général* (written around 1728–30 and published in 1755) and Jean-François Melon's *Essai politique sur le commerce* (1736). Cantillon's text is the more notable of the two on account of his theory of price (measured in land) and his general theory of the circulation of goods founded upon the behaviour of the entrepreneur. The theory of the balance of trade is modified by taking account of the value in land of the products exchanged, and Cantillon associates with it an automatic equilibrating mechanism mediated by modifications to the expenditures of landed proprietors.
The science of *commerce politique* was given a decisive boost in 1751 with Vincent de Gournay's accession to the post of Supervisor of Commerce. The intention was that France should follow the example of England in supporting mercantile activity, but Gournay's economic thinking was not that original: it remained close to the brand of mercantilism advanced by Josiah Child and which saw in a low rate of interest the best way of promoting commerce. His significance, rather, lay in the fact that he gathered around himself young administrators (such as Véron de Forbonnais, the abbé Morellet, and Turgot) who would be influential up to the time of the Revolution.
The science of commerce that crystallized in de Gournay's writings and those of his group, or in Montesquieu's *Spirit of Laws*, can be characterized by four features. First, trade is composed of flows of goods between nations which exchange their surplus thanks to the practical knowledge of traders. Second, trade depends upon self-interested behaviour, and it implies that the trader has an interest, both economic and symbolic, in keeping to his particular station in life rather than in achieving nobility. Third, trade is the most important form of economic activity. And fourth, the particular interest of the trader could be opposed to that of the state.

1756 – 1789: From Physiocratic Philosophie é conomique to Condorcet's social mathematics

From 1750, economic publications multiplied and this growth accelerated in the years leading up to the Revolution. New contributors to the genre emerged with François Quesnay and the Physiocrats during a troubled political period including the Seven Years' War (1756–63) and the Treaty of Paris, under which a large part of the French colonial empire was lost.
Diderot and d'Alembert's *Encyclopédie* presented Forbonnais with the opportunity of writing a series of entries which were then brought together in his influential *Eléments du commerce* (1754). This publicized the views of the group around de Gournay on the importance of monetary flows and a low rate of interest. But there were two other important contributors to the *Encyclopédie*. Rousseau argued that the General Will was the first principle of political economy and the basic rule of government. This proposition opposed republican virtue to wealth and interested behaviour. The abbot Mably took this argument up in criticism of the Physiocrats (*Doutes présentés sur l'ordre légal et essentiel des sociétés politiques*, 1767). The same argument was revived during the Revolution, when the most radical of the Montagnards reclaimed for themselves ancient republican egalitarianism in order to promote the right of

property and economic development through the market.

Quesnay came to political economy from medicine. There he had encountered the then contemporary notion of animal economy – economy understood as a harmonious organization of diverse phenomena which came together in one coherent whole: the body. He transferred this notion, as was fashionable at the time, to the level of the state so that he was able to talk of economic government, a concept vital to the presentation of his ideas. The task of economic government was to administer resources – men, land, money – in such a way that the nation would enjoy abundance; under-employment of resources was not to be attributed to individuals, but to the errors of economic government. According to Quesnay (*Grains*, 1757), economic government should leave the decision of what is best in matters of culture or trade to the interested behaviour of men. It should limit itself to providing an institutional context favourable to interested behaviour; commercial freedom and a predictable tax levied upon the net product (and not on the gross product as in Vauban) so that productive capital might be maintained. The latter was later elevated to the status of the central variable in the economy since the amount of the net product is always fixed as a proportion of farmers' circulating capital.

Quesnay elaborated the advantages of free trade in the market for wheat in arguments that Dupont de Nemours and Turgot then adopted. He explained how free trade blunted brutal market fluctuations – a phenomenon noted by Gregory King and elaborated by Charles Davenant (*Essay upon the Probable Methods of Making a People Gainers in the Balance of Trade*, 1699) in the 17th century – by allowing compensating adjustments between nations. The consumer enjoyed the benefits of more stable prices. The producer who would benefit from a better price will be prompted to produce more – so long as the price did not fall too far as a consequence of a good harvest. These interests conjoin those of the consumer (in the security of provision) and those of the state (enhanced wealth and increased fiscal returns).

In 1758 Quesnay converted to his camp Count Mirabeau, whose *L'ami des hommes* (1758) on population and commerce had been well received. Their subsequent close collaboration led to the major doctrinal publications of Physiocracy – *Théorie de l'impôt* (1760) and *La philosophie rurale* (1763) – in which Quesnay elaborated his idea of the single tax payable by sole proprietors on the grounds that they were the sole recipients of agricultural rent. But the theoretical landmark of this period is the *Tableau Economique*, which appeared in different versions between 1758 and 1767. There are echoes in the *Tableau* of Cantillon's approach, his text having circulated in manuscript before 1755. Flows between rural and urban classes are conceived at the highest level of abstraction so that the relation of these classes to each other might be clearly demonstrated. The key difference is that Cantillon was interested in monetary phenomena and commercial uncertainty, matters neglected by Quesnay.

In the initial versions of the *Tableau*, Quesnay showed how landowners' expenditures made possible the circulation of the wealth produced by farmers and artisans. The later versions, more 'macroeconomic' in form, showed under what conditions the monetary expenditure of a society restricted to three classes (farmers, landowners and artisans) allowed the reproduction of the conditions of agricultural wealth at an optimal level. This final version of the *Tableau* also allowed the impact on the amount of the net product of accrued luxury expenditures, or of indirect taxation, to be studied; it hence made possible an estimation of their importance to the nation as a whole.

The Physiocratic School gained in importance during the 1760s and played a role in economic administration. In 1764–5 the Comptroller General, Bertin, liberalized trade in wheat and in flour; together with Turgot, Inspector in Limousin, and Pierre Paul Le Mercier de la Rivière, Inspector in the

Antilles, the highest reaches of administration opened up to Physiocracy. The doctrine spread abroad: to Baden, Austria, Poland, Russia and Sweden. However, a succession of poor harvests in the later 1760s put an end to those tentative efforts at trade liberalization. Quesnay lost interest in political economy; the baton was taken up by a small number of writers, among whom Turgot was pre-eminent.

Turgot is close to Physiocracy, but he differs in theoretical points and practical matters. He was close in so far as he was a strong advocate of a complete freedom of trade, distancing himself from Gournay's slogan ‘liberty and protection’; and he adopted Quesnay's analysis of the price of wheat in respect of the theory of the net product and the single tax. But Turgot never made use of the *Tableau Economique*; he was, he said, happy to employ its metaphysics, meaning the competitive process upon which it was founded.

Turgot's originality is evident from his *Réflexions sur la formation et la distribution des richesses*, published in 1766 in the Physiocratic journal *Ephémérides du citoyen*, and can also be appreciated from many of his writings of this period that were either never completed, or remained unpublished, such as his essay *Valeur et monnaie*. His approach is based upon sensualist philosophy, and this orients him to a subjective theory of value and utility. The economic thought of abbé de Condillac, the principal theorist of sensualism in France, was similar in this respect, for in his *Le gouvernement et le commerce considérés relativement l’un à l’autre* (1776) Condillac defined value in terms of judgement and opinion made in respect of the scarcity and utility of a good – combining this with a more thorough study of the competitive process.

This led Turgot to a number of significant findings: the formation of markets upon the foundation of mutual interest between buyers and sellers constrained by transport costs (*Foires et marchés*, 1757); a theory of price (estimated value) proceeding from a discussion of the scarcity (quoted value) of a good for parties to an exchange – although Turgot stopped at two agents and two goods (*Valeur et monnaie*, 1769); the justification for interest upon loans and its determination according to market forces (*Mémoire sur le prêt à intérêt*, 1770); a theory of the formation of a uniform rate of profit, or a stable hierarchy of such rates (*Réflexions*, 1766). If one adds to this list the discovery of the principle of decreasing returns to capital in agriculture it is clear that Turgot's theoretical contribution was a considerable one, especially in view of the fact that he had heavy responsibilities in his various administrative posts – as Inspector in Limoges (1761–74), then Navy Minister (1774), and finally Comptroller General for Louis XVI (1774–6).

In this last appointment, together with a small number of loyal supporters (Dupont de Nemours, Condorcet) Turgot worked to re-establish the freedom of trade in grain, which gave rise to a dispute with Jacques Necker (*Sur la législation et le commerce des grains*, 1775), Necker opposing to Turgot's liberalism a more flexible and pragmatic conception of the administration of trade for which the anticipations and beliefs of agents were vital, a factor neglected by Turgot.

Political economy thus assumed an explicitly political dimension. For Quesnay and Le Mercier de la Rivière (*L’ordre naturel et essentiel des sociétés politiques*, 1767) the community of economic interests shared by different groups secured the harmony of the social body, provided that the legislator surrounded himself with experts in the science of economics. Mirabeau and Turgot considered that landed proprietors represented the general interest and should determine the level of taxation in local assemblies. This connection between property, taxation and the citizenry would play an essential role in the course of the Revolution. This connection is also the basis upon which a general science of the social was conceived (the moral or political sciences according to the abbé Baudeau, and social science as in

Sièyes, Condorcet or Roederer) in which political economy took its place alongside ethics, politics and jurisprudence. It was in this form that political economy was first institutionalized in the classes on moral and political sciences at the Institut (1795).

We should also take note of a specific development owed to the presence of Condorcet, a mathematician of the first rank, in Turgot's entourage. Condorcet's interest in public affairs during the Revolution gave rise to his essays on social mathematics which inserted calculus and the theory of probability into social science with respect to issues such as insurance or the rate of interest on loans. Quite remarkable is the result obtained by Condorcet in respect of the determination of truth on the part of a jury or assembly when there are several votes and more than two choices. Condorcet formulated the result which Kenneth Arrow demonstrated in 1951 as the ‘impossibility theorem’. But for the time being, this avenue remained undeveloped, apart from the work of isolated scholars like Achylle-Nicolas Isnard (*Traité des richesses*, 1781), Nicolas Canard (*Principes d’économie politique*, 1801) or Charles-François Bicquilley (*Théorie élémentaire du commerce*, 1804). Say rejected it quite explicitly.

1800 – 30: Say, the Saint-Simonians and the industrial order

Physiocracy continued to play a role during the revolutionary period. A number of followers had been shaped by this doctrine, and this remained true even of those who had distanced themselves on central points, such as the abbé Sièyes, Roederer or Condorcet. However, the diffusion of the *Wealth of Nations* profoundly altered the way in which the economy was conceived in France. Two authors symbolize this progression: Jean-Baptiste Say (*Traité d’économie politique*, 1803) and Jean-Charles Simonde de Sismondi (*De la richesse commerciale*, 1803). Despite their evident indebtedness to Quesnay and Turgot, many traces from these authors remaining in their writings, they founded their political economy upon the *Wealth of Nations*, Germain Garnier's influential translation being published in 1802. For Say and Sismondi, Smith had highlighted two salient points. The first was that the industrial producer acquired his social independence thanks to the market. He no longer depended upon a person of influence (such as a rich landed proprietor) but on a collection of purchasers. The second was that the level of economic activity did not depend on expenditures, but on the quantity of capital. In this respect the social and political dimension of political economy came to the fore in a conception of a new type of society which Say called ‘industrial society’.

Say's political economy is characterized by the manner in which he orders his material by the tripartite schema of production, distribution and consumption. He did more than simply put Smith's ideas in order; he modified both Smith's ideas and those of his British interpreters. Say followed the tradition of Turgot and Condillac. His theory of value is based on utility, not labour. He thus rejected the opposition of natural to market price, in the last editions of the *Traité* considering only market price. Say's theory of production minimizes the role of the division of labour. He argues that the progress of wealth arises from the introduction of new machines incorporating scientific knowledge which places at the disposal of producers the free forces of nature, thereby reducing costs of production. The theory of distribution is entirely based on relations between supply and demand among different categories of the suppliers of productive services, including those of entrepreneurs.

Say's name is firmly linked to two fundamental contributions: his formulation of the law of markets and his analysis of the role of the entrepreneur. The latter played a significant part in his theory. The entrepreneur coordinates the employment of productive services within an enterprise and links different

markets (for final goods and for productive services). In this respect Say's entrepreneur, as in Cantillon, is the economic agent who confronts the uncertainties involved in market transactions.

Say argues that value depends upon utility and is the measure of wealth. From 1815 Say encountered criticism on these two points from Ricardo, and never managed satisfactorily to meet the criticism that the fall in value of a good consequent upon technical progress cannot at the same time indicate that the society is richer (a given amount of utility being obtained at a lower cost) and also poorer (since value has diminished). In this debate Say had trouble in defining a theoretically founded position which was not a reformulation of the Ricardian theory, including here the theory of rent. The difference in method is certainly here more marked and on this point Say received support from Sismondi (*Nouveaux principes d'économie politique*, 2nd edition, 1826). But they were not in agreement on the implications of the law of market opportunities and on the interpretation of the English industrial crisis of 1825: for Say, it resulted from excessive credit being extended by banks, while Sismondi saw it as a crisis of overproduction originating in the growth of production exceeding that of consumption.

In France the debate on value took a distinctive course. Rossi, the successor to Say at the Collège de France, rapidly abandoned the position of his predecessor and moved nearer that of Ricardo. He also elaborated a methodological synthesis which distinguished between a pure and abstract economics in the fashion of Ricardo and an applied political economy influenced by institutional and political context. Most importantly, however, following on from Rossi, Dupuit criticized Say's position: the value of a good was not measured by its utility, instead one might measure utility by the maximum sacrifice a purchaser was prepared to make to obtain it.

Beyond these theoretical debates, the political economy of Say and his successors bore upon the nature of society. The doctrine of industrialism expressed the idea that modern society depended upon the mastery of man over nature thanks to science and technology on the one hand, and social science on the other. Industrialism endorsed and promoted industry, the social independence produced by the market and the reconfiguration of the political sphere, where the state played a diminished role, permitting agents to decide what was best for themselves while it also assigned a greater role to industrial classes in the representation of the citizenry. During the 1820s this doctrine divided into two paths: the liberal industrialism of Say, Charles Dunoyer and Charles Comte separated from the organized industrialism of Henri-Saint-Simon, Auguste Comte and the Saint-Simonians. This latter tendency argued that the market was not an institution adequate to the effective redistribution of resources, as periodic economic crises showed. It was the same with the hereditary transmission of property; in its place, industrialism envisaged a centralized and rational organization of economic activity. In addition, it asserted that industrial society could not be based simply on selfish interest and the doctrine of utility, but had need of a moral or religious dimension. Here we are already approaching socialist theses that flourished during the 1840s.

This opposition assumed particular force with the link that developed between organized industrialism and a new social category, that of the engineer. Since the 18th century France had provided itself with a corps of engineers charged with the provision and maintenance of infrastructure (bridges, roads and canals), mines and defence. These engineers were selected for their abilities in mathematics, and they employed this in a profession placed between technology and economy. 'Engineer economists' (Etner, 1987) created a link between political economy and mathematics in economic calculation. This is evident in the work of Dupuit, who calculated the utility of infrastructure, and expounded the principle that a tariff should be charged according to the gain that a user enjoyed. Antoine-Augustin Cournot

(*Recherches sur les principes mathématiques de la théorie des richesses*, 1838) was himself a pure mathematician. While he developed an economic approach in respect of theses expounded by Rossi on the value of exchange, he remained, as a writer, isolated in his use of mathematics and also on account of his critique of free trade. His work was hardly read by his contemporaries.

1830 – 70: The French classical liberal school, socialism and the teaching of political economy

Say dedicated much of his life to teaching political economy: at the Athénée royal (1815–19), at the Conservatoire des arts et métiers (1819–32) and finally at the Collège de France (1830–2). The importance that Say attached to teaching political economy derived mainly from his adherence to Enlightenment philosophy, according to which human misfortune resulted from ignorance of the laws of nature and of society, and from the ascendancy of doctrines which prevented individuals from daring to think for themselves. It also followed from his own economic theory, for he maintained that scientific knowledge was among the productive services that the entrepreneur had to bring together so that he might serve the public effectively.

This perspective came to be of importance in the debate with Ricardo. Say did not neglect theory, and he sought to develop it (the law of markets, the theory of value, the theory of productive services and so on), but he considered that the essentials were already understood. Republican in outlook, Say saw in political economy the means to bring about a more efficient society, one in which there was greater justice because it was more egalitarian. The diffusion of a liberal credo favourable to commercial freedom, free trade and reduced taxation was therefore important. Agreement among economists provided a secure foundation for the production of a body of ideas appropriate for public instruction. Ricardo's theoretical refinements, which he did not himself think had practical consequences, brought about disagreements which alienated readers from political economy and its applications, as shown by the jibes against economists of François Ferrier, a customs official and defender of the balance of trade (*Du gouvernement considéré dans ses rapports avec le commerce, ou de l'administration commerciale opposée aux économistes du 19ème siècle*, 1804 and 1822.)

After the death of Say in 1832 this conception of political economy was epitomized in the various institutions around which liberals organized themselves. In 1832 François Guizot re-established the Académie des sciences morales et politiques that Bonaparte had suppressed; in 1842 economists founded the Society for Political Economy so that they might there discuss theory and policy; the publisher Guillaumin saw that their work was published (the *Collection des principaux économistes* in 1842 and then, in 1852–3, the remarkable *Dictionnaire de l'économie politique*). Finally, liberal economists founded a journal, the *Journal des économistes*, which was published from 1841 right up to the French military collapse in 1940.

The initial aim of the *Journal des économistes* was the diffusion of economic theory, thought to be already complete, so that it might lead to practical ends. The contemporary problem appeared to relate to the forms of association between workers and capitalists, and support for a spirit of enterprise that had not brought about all its anticipated benefits. Frédéric Bastiat led a powerful campaign on behalf of free trade, seeking to create in France a movement which was the equivalent of Cobden's Anti-Corn Law League. The struggle against socialism was not therefore a priority for liberal economists in dialogue with 'social reformers', notably with Pierre-Joseph Proudhon who, thanks to his relationship with Joseph Garnier, then director of the *Journal*, was regarded a part of the circle of economists and published his

Contradictions économique with the publisher Guillaumin. It is true that he was similar to them on one count – the defence of freedom – and which he sought to reflect in mutualism, one of the forms of association in question. Matters quickly changed in 1848; the suppression by the new authorities of the chair of political economy at the Collège de France profoundly upset economists, who set to work in support of its re-establishment; and they opposed many projects developed at this time, such as the 'right to work' and national workshops, which generally promoted the centralized regulation of economic activity. Besides writing in support of property and social order, the economists (especially Michel Chevalier and Joseph Garnier) opposed the ideas of Louis Blanc: remunerating work independently of its productive contribution, as in the national workshops, created a problem with incentives. Nevertheless, the *Journal des économistes* saw its principal adversaries as ignorance of the principles of political economy, protectionist prejudices, and socialist illusions. Bastiat developed this idea on his *Sophismes économiques* (1845). Socialism and protectionism were conceived as equivalent, for both involved despoliation, an involuntary transfer of resources which impoverished society to the advantage of one particular section of that society.

The creation of the Empire in 1851 opened up a cleavage among the economists. The most liberal among them, such as Gustave de Molinari, left the country, while others furthered their industrial ideas and political careers, like Chevalier, who became a Privy Councillor and personal Councillor to Napoleon III. From this position he was able to promote the central idea of liberal economics with the signature of the Cobden–Chevalier Treaty on free trade in 1860. During the Empire period there were additional measures that conformed to liberal ideas, such as restoring the right of association to workers in 1864 and furthering education in political economy. Hitherto it had been taught only in several specialized institutions (the Conservatiore, the Collège de France, and the Ponts et Chaussée), but from 1860 public education in political economy began in the provinces, and in Paris in the law faculty with a course given by Anselme Batbie. However, the development of teaching in political economy really began to develop only with the reform of the teaching of law in 1877.

See Also

- Boisguilbert, Pierre le Pesant, Sieur de
- Canard, Nicolas-François
- Cantillon, Richard
- Chevalier, Michel
- Colbert, Jean-Baptiste
- Condillac, Etienne Bonnot de, Abbé de Mureau
- Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de
- Dupuit, Arsene-Jules-Emile Juvenal
- Forbonnais, François Véron Duverger de
- Isnard, Achylle Nicolas
- Law, John
- Mercier De La Rivière, Pierre-Paul (Mercier or Lemercier)
- physiocracy

- Quesnay, François
- Saint-Simon, Claude-Henri
- Say, Jean-Baptiste
- Sismondi, Jean Charles Leonard Simonde de
- Turgot, Anne Robert Jacques, Baron de L'Aulne

Bibliography

Béraud, A. and Steiner, P., eds. 2004. L'économie politique néo-smithienne sur le Continent: 1800–1848. Special issue of *Œconomia* 34.

Breton, Y. and Lutfalla, M., eds. 1991. *L'économie politique en France au 19ème siècle*. Paris: Economica.

Charles, L., Lefèvre, P. and Théré, C., eds. 2007. *Commerce, population et société autour de Vincent de Gournay*. Paris: Ined.

Dockès, P. et al., eds. 2000. *Les traditions économiques françaises: 1848–1939*. Paris: Cnrs éditions.

Etner, F. 1987. *Histoire du calcul économique en France*. Paris: Economica.

Faccarello, G., ed. 1998. *Studies in the History of French Political Economy*. London: Routledge.

Faccarello, G. 1999. *The Foundations of Laissez-faire: The Economics of Pierre de Boisguilbert*. London: Routledge.

Hollander, S. 2005. *Say and the Classical Canon in Economics: The British Connection in French Classicism*. London: Routledge.

Kaplan, S. 1976. *Bread, Politics and Political Economy in the Reign of Louis XV*. The Hague: Martinus Nijhoff.

Le Van-Lemesle, L. 2004. *Le juste et le riche: L'enseignement de l'économie politique 1815–1950*. Paris: Comité pour l'histoire économique et financière de la France.

Perrot, J.-C. 1992. *Une histoire intellectuelle de l'économie politique (XVII-XVIII siècle)*. Paris: Ehess.

Potier, J.-P. and Tiran, A., eds. 2003. *Say: nouveaux regards sur son œuvre*. Paris: Economica.

Steiner, P. 1998. *La 'Science nouvelle' de l'économie politique*. Paris: Presses universitaires de France.

Steiner, P. 1998. *Sociologie de la connaissance économique: Essai sur les rationalisations de la connaissance économique (1750–1850)*. Paris: Presses universitaires de France.

Vatin, F. 1998. *Economie politique et économie naturelle chez Cournot*. Paris: Presses universitaires de France.

Whatmore, R. 2000. *Republicanism and the French Revolution: An Intellectual History of Say's Political Economy*. Oxford: Oxford University Press.

How to cite this article

Béraud, Alain and Philippe Steiner. "France, economics in (before 1870)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000306> doi:10.1057/9780230226203.0599

The New Palgrave Dictionary of Economics Online

franchising

Francine Lafontaine
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Franchising typically refers to contractual relationships between legally independent firms, where one firm pays the other for the right to operate under the latter's brand, or sell its product, in a given location and time period. Franchised firms account for a large portion of commerce in the United States and around the world. The economics literature on franchising has focused mostly on why and how firms franchise, emphasizing incentive or opportunism issues on the part of franchisees and franchisors to explain various aspects of the relationships. Empirical findings have confirmed the importance of such issues in shaping these contractual relationships.

Keywords

antitrust; business-format and traditional franchising; chain structures; contract enforcement; franchising; industrial organization; principal and agent; risk; royalties; sharecropping; vertical integration

Article

Franchising typically refers to contractual relationships between legally independent firms under which one of the firms, the franchisee, pays the other firm, the franchisor, for the right to sell the franchisor's product and/or the right to use its trademarks and business format in a given location and for a specified period of time.
According to the *American Heritage Dictionary of the English Language*, the word ‘franchise’ comes from the old French word *franche*, meaning free or exempt. In medieval times, a franchise was a right or privilege granted by a sovereign power – king, Church, or local government – to engage in activities such as building roads, holding fairs, organizing markets, or to maintain civil order and collect taxes, in a particular location and for a certain period of time. The grantee was typically required to pay a share of its product or profit to the sovereign power for this right or privilege. That payment was called a *royalty*,

a term we still use today.

Governments still grant franchises in certain industries, such as the cable television industry (see for example Zupan, 1989; Prager, 1990) and highway construction projects (see Engel, Fisher and Galetovic, 2001). The word ‘franchise’ is used also in the sports industry to refer to the right to operate a team in a particular locale. Most commonly, however, the term refers to the type of ongoing business relationships defined above.

In the United States, the Federal Trade Commission (FTC) has jurisdiction over federal disclosure rules for franchisors. It requires three conditions for a business relationship to be deemed a franchise and thus subject to these rules. First, the franchisor must license a trade name and trademark that the franchisee operates under, or the franchisee must sell products or services identified by this trademark. Second, the franchisor must exert significant control over the operation of the franchisee or provide significant assistance to the franchisee. Third, the franchisee must pay at least 500 dollars to the franchisor at any time before or within the first six months of operation (see Disclosure Requirements and Prohibitions concerning Franchising and Business Opportunity Ventures, CFR, Title 16, Part 436). Authorities outside the United States, including Australia, Canada, and the European Union, typically rely on similar criteria.

Franchise agreements take one of two forms: business-format franchises, where the relationship ‘includes not only the product, service, and trademark, but the entire business format itself – a marketing strategy and plan, operating manuals and standards, quality control, and continuing two way communication’ (U.S. Department of Commerce, 1988, p. 3) and product and trade name or traditional franchising, where franchised dealers ‘concentrate on one company's product line and to some extent identify their business with that company’ (1988, p. 1). The latter include car dealerships, petrol stations, and bottlers. Several countries, however, exclude these from their franchise statistics.

In 2001, the revenues of franchised chains in the United States were estimated at 1.37 trillion dollars or 13.6 per cent of GDP (Blair and Lafontaine, 2005, p. 26). In retailing, it is estimated that about one-third of each dollar of sales is achieved via franchised chains. Three-quarters of these sales occur in traditional franchise outlets. Business-format franchising, however, accounts for the majority of jobs and outlets: of the more than 750,000 franchised establishments in the United States in 2001, 620,000 were associated with the 2,500–3,000 business-format franchisors in the economy. Thus business-format franchising accounted for 4.3 times as many establishments, and employed four times as many workers, as traditional franchising did in 2001 (Price Waterhouse Coopers, 2004, p. 1).

While the United States franchising sector remains the largest in the world, franchising is increasingly a global phenomenon. Several large US-based franchisors have expanded abroad aggressively. With the development of many home-grown franchise companies, this has led to franchising sectors of many developed countries now rivalling that in the United States. According to Arthur Andersen & Co. (1995), countries such as Canada, Japan and Australia have more franchisees per capita than the United States. Still, the extent of franchising continues to vary significantly across countries.

The interest of industrial organization economists in the study of franchising emerged in the 1970s. Going back at least to Caves and Murphy (1976), Rubin (1978) and Klein (1980), economists have formulated theories about why franchising exists and why the contracts take the form they do. The economic significance of franchising in itself would easily justify this interest. However, much of the research on franchising has been carried out with a much broader goal in mind, namely to understand

how firms organize their activities generally, with franchising viewed as an exemplar of the types of long-term, contract-based organizations that stand between spot market interactions and complete vertical integration, and thus a context in which to test agency and transaction cost theory. As Caves and Murphy note (1976, p. 572), ‘The franchise relation raises fundamental questions concerning the nature of the firm and the extent of its integration.’

Caves and Murphy introduced many of the issues that have remained central themes in the literature, noting in particular the scale differential that gives rise to chain structures, the need to price franchise rights to give incentives to franchisees, and the factors that lead firms to rely to varying degrees on franchising rather than company ownership. Regarding the latter, the authors emphasized the franchisor's initial need for capital, the importance of owner operators in some industry segments, and the possibility that franchisees might, through various activities and spillover effects, damage the brand. Mathewson and Winter (1985) formalized many of these ideas. Rubin (1978) pointed out the role that franchisors play in developing and maintaining the value of their brands, thereby noting explicitly that franchisor incentives also matter. Based on this idea, Bhattacharyya and Lafontaine (1995) developed a model to explain some remaining puzzling facts about the contracts, namely the degree of uniformity and stability of the financial terms in these contracts. Finally, a separate but complementary approach to explaining various aspects of franchise contracts, which focuses on self-enforcement, was proposed in part by Rubin but developed most explicitly by Klein (1980; 1995).

Perhaps what distinguishes franchising the most from other contractual contexts, however, is the amount of empirical work that has been conducted on the subject. This empirical literature has established several facts. First, it has shown that incentive issues on the franchisee's and the franchisor's side play a central role in franchise contracting (see Lafontaine and Slade, 2007 for a review). It has also shown that franchisees’ local profit-maximizing behaviour – or opportunism – can be a problem for franchisors. Consequently, the relationships are designed with self-enforcement in mind (see for example Brickley, Dark and Weisbach, 1991, and Kaufmann and Lafontaine, 1994, on the role of contract termination and the presence of ongoing rent respectively).

In some cases, the theories and the facts have not matched so well. For example, franchising, like sharecropping, tends to be positively associated with risk (see for example Allen and Lueck, 1995, on sharecropping). This is inconsistent with the typical agency argument that risk-averse agents should be insured more when the environment is more volatile. Lafontaine and Bhattacharyya (1995) and Prendergast (2002) explain this empirical ‘anomaly’ by noting that franchisees choose their effort level in ways that exacerbate the high and low demand signals they receive, which in turn makes the variance of outcomes – measured risk – larger for franchised than company outlets. Prendergast (2002), moreover, argues that principals will need to delegate more, and thus give higher powered incentives to agents, in uncertain environments. Ackerberg and Botticini (2002) propose instead that this anomalous effect of risk reflects an endogenous matching problem.

Finally, the literature on franchising has found that incentive requirements and mechanisms interact in important ways within a given relationship or contract (see notably Slade, 1996; Bradach, 1997; Brickley, 1999; Lafontaine and Raynaud, 2002). Moreover, competition or antitrust policy, as well as franchise-specific laws, constrain the set of contract terms franchisors can rely on. Another important – and underdeveloped – segment of the literature examines the effect of franchising on economic outcomes. The need for exogenous variation in organizational form has made this type of work difficult,

but results suggest that prices, for example, are somewhat higher under franchising (see Lafontaine and Slade, 2007, for a review). Much more work is needed, however, in both these areas.

See Also

- antitrust enforcement
- contract theory
- moral hazard
- principal and agent (i)
- principal and agent (ii)
- sharecropping
- vertical integration

Bibliography

Ackerberg, D.A. and Botticini, M. 2002. Endogenous matching and the empirical determinants of contractual form. *Journal of Political Economy* 110, 564–91.

Allen, D.W. and Lueck, D. 1995. Risk preferences and the economics of contracts. *American Economic Review* 85, 447–51.

Arthur Andersen & Co. 1995. Worldwide franchising statistics.

Bhattacharyya, S. and Lafontaine, F. 1995. Double-sided moral hazard and the nature of share contracts. *RAND Journal of Economics* 26, 761–81.

Blair, R.D. and Lafontaine, F. 2005. *The Economics of Franchising*. Cambridge: Cambridge University Press.

Bradach, J.L. 1997. Using the plural form in the management of restaurant chains. *Administrative Science Quarterly* 42, 276–303.

Brickley, J.A. 1999. Incentive conflicts and contracting: evidence from franchising. *Journal of Law and Economics* 42, 745–74.

Brickley, J.A., Dark, F.H. and Weisbach, M.S. 1991. The economic effects of franchise termination laws. *Journal of Law and Economics* 34, 101–32.

Caves, R.E. and Murphy, W.F. 1976. Franchising: firms, markets, and intangible assets. *Southern Economic Journal* 42, 572–86.

Engel, E.M.R.A., Fisher, R.D. and Galetovic, A. 2001. Least-present-value-of-revenue auctions and highway franchising. *Journal of Political Economy* 109, 993–1020.

Kaufmann, P.J. and Lafontaine, F. 1994. Costs of control: the source of economic rents for McDonald's franchisees. *Journal of Law and Economics* 37, 417–54.

Klein, B. 1980. Transaction cost determinants of ‘unfair’ contractual arrangements. *American Economic Review* 70, 356–62.

Klein, B. 1995. The economics of franchise contracts. *Journal of Corporate Finance* 2, 9–37.

Lafontaine, F. and Bhattacharyya, S. 1995. The role of risk in franchising. *Journal of Corporate Finance* 2, 39–74.

Lafontaine, F. and Raynaud, E. 2002. Residual claims and self enforcement as incentive mechanisms in franchise contracts: substitutes or complements. In *The Economics of Contract in Prospect and Retrospect*, ed. E. Brousseau and J.M. Glachant. Cambridge: Cambridge University Press.

Lafontaine, F. and Slade, M.E. 2007. Vertical integration and firm boundaries: the evidence. *Journal of Economic Literature* 45(3), 631–687.

Mathewson, F. and Winter, R.A. 1985. The economics of franchise contracts. *Journal of Law and Economics* 28, 503–26.

Prager, R.A. 1990. Firm behavior in franchise monopoly markets. *RAND Journal of Economics* 21, 211–25.

Prendergast, C. 2002. The tenuous trade-off between risk and incentives. *Journal of Political Economy* 110, 1071–102.

Price Waterhouse Coopers. 2004. Economic impact of franchised businesses.

Rubin, P. 1978. The theory of the firm and the structure of the franchise contract. *Journal of Law and Economics* 21, 223–33.

Slade, M.E. 1996. Multitask agency and contract choice: an empirical assessment. *International Economic Review* 37, 465–86.

U.S. Department of Commerce. 1988. *Franchising in the Economy*, prepared by Andrew Kostecka. Washington, DC: U.S. Department of Commerce.

Zupan, M.A. 1989. Cable franchise renewals: do incumbent firms behave opportunistically? *RAND Journal of Economics* 20, 473–82.

How to cite this article

Lafontaine, Francine. "franchising." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000200> doi:10.1057/9780230226203.0601

The New Palgrave Dictionary of Economics Online

free banking era

Arthur J. Rolnick and Warren E. Weber

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the free banking era entry into banking was virtually unrestrained, banks could issue their own currency and governments did not insure banks; many banks closed and many noteholders reportedly suffered. An early view of this period is that free entry led to banks over-issuing notes, resulting in large losses for noteholders. More recent research has shown that this is incorrect. Although such failures and losses did occur, these were generally due to the capital losses banks suffered when the prices of the state bonds backing their notes fell, rather than to note over-issuance or fraudulent banking practices.

Keywords

free banking; free banking era; free banking laws; wildcat banks

Article

Imagine the US economy without Federal Reserve notes, that is, without a uniform currency. Instead, imagine that the currency consists of notes issued by privately owned banks and that are redeemable in specie on demand. And imagine that to enter the banking business is relatively easy, so that the notes of hundreds of banks exist. And imagine as you travel around the country, notes of out-of-town banks are not readily accepted as means of payment at par because the solvency of such banks is difficult to ascertain.

How well would such a banking system function? In particular, with free entry into banking, would banks not have an incentive to over-issue their notes, leaving the public holding worthless pieces of paper when the banks failed? And would trade not be difficult without the existence of a uniform currency? Indeed, a reading of historical accounts of the so-called free banking era – the 26 years from 1837 to 1863, a period when entry into banking was relatively free and banks issued their own notes – would lead to this conclusion. The prevailing view of this period, at least until the mid-1970s, was that allowing such freedom in banking was a mistake. However, a more recent examination of the era reveals that while the free banking system was not without its problems, free banks and their noteholders fared

much better than has often been portrayed.

The beginning of free banking

Prior to 1837, to establish a bank in the United States was a very cumbersome, and at times political, process. Individuals who wanted to start a bank had to obtain a charter from the legislature of the state in which they wanted to operate. Beginning in 1837, some states reformed their bank-chartering systems so that entry into the banking industry would be easier. States tempered the goal of easy entry with another goal: to provide the public with a safe bank currency. Most states attempted to reach these goals by enacting what were called *free banking laws*.

The first free banking law was proposed in New York. Its provisions openly aimed at both easy entry and safety. The law allowed anyone to operate a bank as long as two basic requirements were met: (*a*) all notes the bank issued had to be backed by state bonds deposited at the state auditor's office and (*b*) all notes had to be redeemable on demand at par, or face, value. If the bank failed to redeem notes presented for payment, however, the auditor would close the bank, sell the bonds, and pay off the noteholders. If the bond sale did not generate enough specie to redeem the bank's notes at par, noteholders had additional protection by having first legal claim to the bank's other assets. Thus, free banking meant free *entry* into banking; it did not mean laissez-faire banking.

New York's proposed free banking law became the basic blueprint for the free banking laws in other states. (Michigan actually passed a free banking law modelled on the New York proposal a year before the legislation was passed there.) Table 1 shows which states passed free banking laws and when the laws passed. Note that of the states that passed such legislation, most did so in the 1850s.

US states with and without free banking laws by 1860

States with free banking laws	Year law passed	States without free banking laws
Michigan	1837[a]	Arkansas
Georgia	1838[b]	California
New York	1838	Delaware
Alabama	1849[b]	Kentucky
New Jersey	1850	Maine
Illinois	1851	Maryland
Massachusetts	1851[b]	Mississippi
Ohio	1851[c]	Missouri
Vermont	1851[b]	New Hampshire
Connecticut	1852	North Carolina
Indiana	1852	Oregon
Tennessee	1852[b]	Rhode Island
Wisconsin	1852	South Carolina
Florida	1853[b]	Texas

Louisiana	1853	Virginia
Iowa	1858[b]	
Minnesota	1858	
Pennsylvania	1860[b]	

[a]Michigan prohibited free banking after 1839 and then passed a new free banking law in 1857.

[b]According to Rockoff, very little free banking was done under the laws of these states.

[c]In 1845, Ohio passed a law that provided for the establishment of ‘independent banks’ with a bond-secured note issue.

Source: Rockoff (1975, pp. 3, 125–30).

The experience

One effect of the free banking laws was to increase the number of banks. In Michigan, for example, the number of banks rose from ten before the law was passed in March 1837 to 33 one year later. In New York the number of banks rose from 97 before the law was passed in March 1838 to 162 three years later. And Indiana, Illinois, and Wisconsin, which each had only one bank in existence when their free banking laws were passed, saw 13, 41, and 15 new banks established respectively within two years. Minnesota had no banks when its free banking law was passed; it saw 16 banks established within one year. In total, of the almost 2,300 banks that existed in the United States prior to the Civil War, slightly more than three-eighths were established or operated under a free banking law (Weber, 2006).

Free banking, however, must also be judged by the laws’ second objective – by how many banks survived and provided their communities with a stable source of banking services, especially a safe currency. Measured by this criterion, free banking is generally considered a failure.

Michigan's disastrous experience with free banking is probably the most famous. By the end of 1839, less than two years after its free banking law was passed, all but four of Michigan's free banks closed (Rockoff, 1975, p. 96). Although explicit loss data do not exist, it has been estimated that the total loss to Michigan's noteholders was as high as four million dollars. This would have been nearly 45 per cent of Michigan's annual income in 1840 (Rockoff, 1975, pp. 17–48). Other states’ experiences with free banking, while not as famous as Michigan's, were almost as bad. Of the 16 free banks that opened under Minnesota's 1858 law, for example, 11 closed by 1863. And many that closed left their noteholders with very little.

However, some states had positive experiences with free banking. New York had very few free bank failures and noteholder losses after 1843. Indiana had much the same record after 1854. And all the failures and losses experienced by Wisconsin free banks occurred in 1861 after the Civil War had begun and the bonds issued by Southern states had greatly depreciated in value.

Free banking was not wildcat banking

According to some historians and economists writing about this period (see, for example, Hammond, 1985, p. 618; Knox, 1903, p. 747; and Luckett, 1980, p. 242), the losses experienced under free banking

were due to fraudulent banking practices by so-called wildcat banks. These were banks that purportedly located redemption offices in remote areas, issued notes far in excess of what they planned to redeem, and then disappeared, leaving the public with notes worth considerably less than their original value. Although some wildcat banking may have occurred, this explanation is not appropriate for most free banking experience because the data do not support it. Wildcat banks supposedly stayed in business for only a few months, after which time their noteholders sustained losses. However, in New York, Indiana, Wisconsin, and Minnesota – four states that were supposed to have had many wildcats – free banks were generally not short-lived.

Most losses to the holders of free bank notes were due not to fraud, but to capital losses suffered by the banks because of several substantial drops in the prices of the state bonds that were required to back the notes they issued. Moreover, while these declines in bond prices may have been induced by any number of economic developments, they were not induced by wildcat banks.

Summary and conclusion

The free banking era was a time when entry into banking was virtually unrestrained, when banks could issue their own currency and when the government did not insure banks. It was also a time when many banks closed and many noteholders reportedly suffered. An early view of this period is that free entry led to banks over-issuing notes, resulting in large losses for noteholders. More recent research has shown that this view is not correct. Although free bank failures and noteholder losses did occur, these were generally due to capital losses banks suffered when the prices of the state bonds backing their notes fell. In general, they were not due to note over-issuance or fraudulent banking practices.

See Also

- banking crises
- Banking School, Currency School, Free Banking School
- monetary economics, history of

Bibliography

Hammond, B. 1985. *Banks and Politics in America*. Princeton: Princeton University Press.

Knox, J.J. 1903. *A History of Banking in the United States*. New York: Bradford Rhodes.

Luckett, D.G. 1980. *Money and Banking*. New York: McGraw-Hill.

Rockoff, H. 1974. The free banking era: a reexamination. *Journal of Money, Credit and Banking* 6, 141–67.

Rockoff, H. 1975. *The Free Banking Era: A Re-examination*. Dissertations in American History. New

York: Arno Press. Ph.D. dissertation, University of Chicago, 1972.

Rolnick, A.J. and Weber, W.E. 1983. New evidence on the free banking era. *American Economic Review* 73, 1080–91.

Rolnick, A.J. and Weber, W.E. 1984. The causes of free bank failures: a detailed examination. *Journal of Monetary Economics* 14, 267–91.

Weber, W.E. 2006. Early state banks in the United States: how many were there and when did they exist? *Journal of Economic History* 66, 433–55.

How to cite this article

Rolnick, Arthur J. and Warren E. Weber. "free banking era." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000205> doi:10.1057/9780230226203.0602

The New Palgrave Dictionary of Economics Online

free disposal

Theodore C. Bergstrom

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

competitive equilibrium; excess demand; free disposal; Kakutani's fixed point theorem; monotonicity

Article

> ‘I should like to buy an egg, please’ she said timidly. ‘How do you sell them?’
> ‘Fivepence farthing for one – twopence for two,’ the Sheep replied.
> ‘Then two are cheaper than one?’ Alice said, taking out her purse.
> ‘Only you must eat them both if you buy two,’ said the Sheep.
> ‘Then I’ll have one please’, said Alice, as she put the money down on the counter. For she thought to herself, ‘They mightn't be at all nice, you know.’
> Lewis Carroll, *Through the Looking-Glass*

If I dislike a commodity, you may have to pay to get me to accept it. But so long as some otherwise non-sated consumer finds this commodity to be desirable, or at least harmless, it could not have a negative price in competitive equilibrium. Likewise, if some firm can dispose of arbitrary amounts of a commodity without using any other inputs or producing any other (possibly noxious) outputs, its price in competitive equilibrium cannot be negative. Therefore competitive equilibrium analysis can be confined to the case of non-negative prices if every commodity is either harmless to someone or freely disposable. If a commodity is not freely disposable and is a ‘bad’ in the sense that everyone prefers less of it to more, it is possible to redefine the ‘commodity’ as the absence of the bad. The commodity so defined can then be treated as a good with a positive price. More generally, it might be possible to choose some alternative coordinate system in which to measure commodity bundles so that in the new coordinate system either there is free disposability or more is preferred to less. But if people are willing to pay a positive sum for a small amount of a commodity and less for a large amount, then the question of whether that commodity will have a positive or negative price in competitive equilibrium cannot be decided in advance. The sign of the equilibrium price will in general depend on supplies of this and other goods and on the detailed configuration of preferences in the economy.

Sometimes a noxious by-product of production or consumption can be transformed into a useful output if sufficient other resources are used. Then the equilibrium price for the by-product may be either positive or negative, depending on the prices of the other inputs and of the output into which it is transformed. This is particularly evident when commodities are distinguished by location. Garbage located in the centre of a city is undesirable to everyone. To bury or incinerate it is costly and generates no valuable outputs. Therefore, if garbage is disposed of in this way, its equilibrium price must be negative. But the garbage could be transported to the country, boiled and fed to pigs. Depending on the costs of this process and the price of pork, it may turn out that converting garbage to pig feed is profitable even when garbage at the city centre has a zero or positive price. Both the ultimate disposition of garbage and the sign of its price have to be determined endogenously in the competitive process.

Early proofs of the existence of competitive equilibrium (Arrow and Debreu, 1954; Gale, 1955; Debreu, 1959) assumed that all commodities are freely disposable or, equivalently, defined equilibrium so as to allow the possibility that in equilibrium some goods might be in excess supply but have zero price. Debreu (1956) shows how the assumptions of free disposal and monotonicity can be greatly relaxed. McKenzie (1959) and Debreu (1962) present general theorems on the existence of equilibrium in which free disposal is not assumed. Rader (1972), Hart and Kuhn (1975), Bergstrom (1976) and Shafer (1976) suggest further generalizations and simplifications in dealing with negative prices in equilibrium.

The formal treatment of negative prices in existence proofs presents an interesting mathematical problem. Most of the standard existence proofs apply the Kakutani fixed-point theorem to a correspondence that maps the set of possible equilibrium prices into itself in such a way that a fixed point for the mapping is a competitive equilibrium price vector. The Kakutani theorem applies to an upper hemicontinuous mapping from a closed bounded convex set to its compact, convex subsets. If the only prices to be considered are non-negative, then the domain for this correspondence can be chosen to be the unit simplex. If all price vectors, positive and negative, must be considered, then an obvious candidate for the domain of this mapping would be the unit sphere $\{p \in R^n | p \cdot p = 1\}$. But this is not a convex set. The closed unit ball $\{p \in R^n | p \cdot p \leq 1\}$ is a convex set, but it contains the vector zero, at which point the excess demand mapping is not upper hemicontinuous.

Debreu (1956) solved this problem neatly in a brief, elegant paper that has received less attention than it deserves. The existence proofs that assume free disposability of all goods had shown that there exists a non-negative price vector at which the excess demand vector is either zero or belongs to the negative orthant. Debreu generalized this result to show that if there is free disposability on any convex cone which is not a linear subspace, then a price vector can be found at which excess demand is either zero or belongs to the cone of free disposability. Furthermore, this price vector gives a non-positive value to every activity in the cone of free disposability. In particular, consider the case where one good is assumed to be freely disposable. Then, from Debreu's theorem, it follows that there exists some price vector at which excess demand for all goods other than the freely disposable good is zero, and at which there is either zero or negative excess demand for the freely disposable good. From Walras's Law and the fact that that there exists some price vector at which excess demand for all other goods is zero, it follows that the price of the freely disposable good can be positive only if excess demand is zero. Therefore this price vector is a competitive equilibrium. Thus Debreu weakened the free disposability assumption from ‘all goods are freely disposable’ to ‘at least one good is freely disposable’.

We can take Debreu's argument one step further and eliminate the assumption of even one freely

disposable good. Nowhere in Debreu's proof is it necessary to assume that the freely disposable good is desirable to anyone. This suggests that the existence of a freely disposable good is not likely to be essential for the existence of equilibrium. For suppose that there is an economy with no freely disposable goods. A fictional good could be introduced which is freely disposable but totally useless and totally harmless to everyone. For the augmented economy found by adding this fictional good to the original economy, by Debreu's theorem there would exist a competitive equilibrium. In this new economy it turns out that the equilibrium price of the useless, freely disposable good must be zero and the vector of equilibrium prices for the other goods can serve as a competitive equilibrium price vector for the original economy.

The approach taken by Bergstrom (1976) is equivalent to introducing a useless and harmless fictional good into Debreu's model. Taking the formal steps of this argument directly without intermediary fictions leads to an upper hemicontinuous mapping from the unit ball into itself for which there is a fixed point on the boundary of the unit ball. This fixed point turns out to be a competitive equilibrium price vector. An interesting alternative approach was taken by Rader (1972) and by Hart and Kuhn (1975). Instead of the Kakutani theorem, they use a theorem about fixed and antipodal points of a continuous mapping from the unit sphere into itself, and are thereby able to deal with all prices on the unit sphere as potential equilibrium prices.

The first and second welfare theorems and the theorems about the equivalence between the core and the set of competitive equilibria apply straightforwardly when there is not free disposal. For example, in order to prove the Pareto optimality of competitive equilibrium in an exchange economy, we simply argue along the usual lines that if any allocation is Pareto superior to a competitive equilibrium, then at the original competitive prices, the aggregate value of consumption in the proposed Pareto superior allocation must exceed the aggregate value of initial endowments. But if the proposed allocation is feasible, then the aggregate consumption vector in the proposed allocation must equal the aggregate initial endowment vector. It follows, whether prices are positive, negative or zero that if the two vectors are equal they must have the same value at the competitive price vector. Therefore there cannot be a feasible allocation which is Pareto superior to a competitive equilibrium. Similar arguments apply to the core theorem. The only matter in which a bit of care must be taken is in defining the activities available to a potential blocking coalition so as to exclude the possibility of dumping undesirable commodities. This simply amounts to the assumption that a blocking coalition must exactly equalize its total consumption of all goods to its total endowment.

See Also

- fixed point theorems
- general equilibrium

Bibliography

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Bergstrom, T. 1976. How to discard free disposability – at no cost. *Journal of Mathematical Economics* 3, 131–4.

Debreu, G. 1956. Market equilibrium. *Proceedings of the National Academy of Sciences of the USA* 42, 876–8.

Debreu, G. 1959. *The Theory of Value*. New York: Wiley.

Debreu, G. 1962. New concepts and techniques for equilibrium analysis. *International Economic Review* 3, 257–73.

Gale, D. 1955. The law of supply and demand. *Mathematica Scandinavica* 3, 155–69.

Hart, O. and Kuhn, H. 1975. A proof of the existence of equilibrium without the free disposal assumption. *Journal of Mathematical Economics* 2, 335–43.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27, 54–71.

Rader, T. 1972. *Theory of General Economic Equilibrium*. New York: Academic Press.

Shafer, W. 1976. Equilibrium in economies without ordered preferences or free disposal. *Journal of Mathematical Economics* 3, 135–7.

How to cite this article

Bergstrom, Theodore C. "free disposal." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000210> doi:10.1057/9780230226203.0603

The New Palgrave Dictionary of Economics Online

Friedman, Milton (1912 – 2006)

Original article by Alan Walters, updated by the editors
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Milton Friedman is widely regarded as one of the most important economists of the 20th century. He is famous for his rehabilitation of money as a major determinant of macroeconomic outcomes. For many academic economists, *A Theory of the Consumption Function* (1957) is his greatest work. Friedman showed that the Keynesian concept of household behaviour was fundamentally flawed, arguing that people adjusted their consumption to variations in their long-term expected (‘permanent’) income. As such, his theory foreshadows the approach to microfoundations that is the cornerstone of modern macroeconomics. His advocacy of economic freedom and market solutions to various socio-economic problems made him a leading policy thinker.

Keywords

assumptions controversy; Bernoulli, D.; choice under uncertainty; consumption function; econometrics; economic freedom; excise taxes; expected utility hypothesis; flexible exchange rates; Friedman, M.; Hotelling, H.; inflation; Keynesianism; Knight, F. H.; monetarism; monetary approach to the balance of payments; monetary transmission mechanism; money; money supply; natural rate of unemployment; nominal income; non-accelerating inflation rate of unemployment (NAIRU); permanent income hypothesis; Phillips curve; prediction; price theory; quantity theory of money; Robbins, L. C.; sequential sampling; unemployment; variance

Article

Early years

Born on 31 July 1912 in New York City, Milton Friedman was the son of a poor immigrant dry-goods merchant, who died when Friedman was 15. Friedman was clearly outside the East Coast establishment of the United States, although he did spend a year in graduate studies at an Ivy League school,

Columbia. He graduated (BA) at Rutgers University in 1932 and completed his AM at Chicago in the following year. After a fellowship at Columbia in 1933–4, he returned to Chicago as a research assistant to Henry Schultz to work on demand analysis, until in 1935 he joined the staff of the National Resources Committee. From 1937 he started a long association with the National Bureau of Economic Research (NBER), which continued until 1981. From 1938 he began another long association – with Rose Director, his wife, which produced, *inter alia*, a son and a daughter.

In 1940 there followed a brief period as visiting professor of economics at Wisconsin. Then, after a two-year stint (1941–3) in the Treasury in the division of tax research, he became associate director of the statistical research group in the division of war research at Columbia University, which lasted until the end of the Second World War. He then spent a year as associate professor at the University of Minnesota, before returning to Chicago as professor of economics in 1946, the year in which he received a Ph.D. from Columbia. His teachers at Rutgers were Homer Jones and Arthur Burns; at Chicago, Frank Knight, Lloyd Mints and Jacob Viner; and at Columbia, Harold Hotelling, J.M. Clark and Wesley Mitchell.

Superficially this record does not seem impressive. Yet it encompasses what some scholars, particularly statisticians, would regard as Friedman's most impressive contributions. Inspired by Hotelling's work on the rank correlation coefficient, his first seminal contribution (1937) was the development of the use of rank order statistics to avoid making the assumption of normality in the analysis of variance. After 70 years this article is still regarded as one of the two or three critical papers in the development of nonparametric methods in the analysis of variance, and it was followed by a discussion of the efficiency of tests of significance of ranked data. It is not surprising that these papers have been of considerable practical use, since they were largely a development of Friedman applying his mind to the practical problems he encountered in analysing incomes and consumer expenditure at the NBER and in Washington. Even at this early stage his work bears the imprint that readily identifies all his subsequent work: it is seemingly ‘simple’, eschewing complexities and complications, concentrating on essentials, and all combined into a lucid exposition.

The detailed analysis of data on incomes and expenditures was Friedman's main occupation during these years. With the exception of Kuznets, Mitchell and Burns, it is difficult to find any eminent economist who acquired such a grounding in the basic empirical material of economics. It is characteristic of all his work that the organization of such data would suggest theoretical developments and new ways of arranging the material, and above all new insights into the economic process. His first published article (1934) was on a method of using the separability of the utility function to measure price elasticities from budgetary data. This exploration of new insights into old data was particularly evident in his book (1945; with Kuznets as joint author) on incomes from private professional practice; there one sees the first signs of the permanent income hypothesis and, indeed, the perceptive reader may guess what is likely to follow. In this book, which Friedman submitted as a doctoral thesis, he argued that the process of state licensure enabled the medical profession more effectively to limit entry into their profession and so enabled them to exploit their patients, keeping fees high and competitors out. The fact that the argument was tightly constructed and buttressed with convincing evidence generated the most vehement opposition and animosity from that proud profession, which appears unabated seven decades later.

Wartime service in the statistical research group, although an interlude in Friedman's basic work on incomes and expenditures, generated one of the most remarkable advances in statistical theory since the seminal contributions of Sir Ronald Fisher. The group was a galaxy, consisting of Abraham Wald, Allen

Wallis, Jacob Wolfowitz, Harold Hotelling and many other distinguished statisticians. The sampling inspection of wartime production of munitions and so forth was a tedious process of selecting a sample of a given size and testing to see the fraction of good ones in the batch. Friedman, together with Allen Wallis and Captain Schuyler, observed that testing a given size of sample was clearly wasteful. The process of testing itself gave information that enabled one to determine the degree of confidence achieved. Thus instead of continuing to test up to a fixed size of sample, the testing could be halted whenever a predetermined level of confidence in the decision had been reached. Friedman formulated the basic idea of what later came to be called ‘sequential sampling’ and caught the interest and imagination of Wald, who developed and proved the theorem underlying the probability ratio test and eventually produced the influential book Sequential Analysis in 1947. These ideas were adapted very rapidly, and sequential analysis became the standard method of quality control inspection. Like so many of Friedman's contributions, in retrospect it seems remarkably simple and obvious to apply basic economic ideas to quality control; that, however, is a measure of his genius.

At the end of the Second World War, Friedman could have continued his work as a statistician. He would have achieved a stature probably as great as that of his most influential teacher, Harold Hotelling. Alternatively he had all the basic qualifications to take the lead in developing the burgeoning field of econometrics, with its great emphasis on the adaptation of statistical theory to modelling economic phenomena. He chose neither. His excursions into statistics were utilitarian rather than speculative, and he could see little to be gained by the endless sharpening of statistical knives, which was the stuff of econometrics during those years following the Second World War. In this decade, his contributions to statistics were even more intimately linked with his strong belief, implanted largely by Mitchell, that economics could acquire plausibility only by being subjected to empirical verification. In spite of the predilections of many economists, Friedman believed that economics should be viewed as an empirical science.

1946 – 1955

This decade at Chicago, much influenced by the wisdom of Frank Knight, witnessed the rapid development of economics as a positive science with its own methodology. The prevailing view of economic theory, as developed by Lionel (later Lord) Robbins, was that the veracity of theory could be tested primarily by the correspondence between assumptions and facts. In his ‘Methodology of Positive Economics’ (in *Essays in Positive Economics*, 1953), Friedman argued per contra that even if one could specify empirical correlates for the assumptions (and this cannot be done in cases where the assumptions are ‘ideal types’ such as homo economicus), that is irrelevant for judging the usefulness of the theory. Only by the correspondence of the predictions and facts should theories be provisionally accepted or rejected. Results, not assumptions, should be the main focus of our scientific activity in understanding the real world. This approach applied the new philosophy of science, developed by Karl Popper, to economics and by implication to associated social sciences. To countless students, Friedman provided an agenda for what Imre Lakatos later called a progressive research programme. The simplicity of a theory in its ability to explain a lot in exchange for a little input and the degree of ‘surprise’ in the prediction were the hallmarks of the new approach to theory. But it was in the efficacy and power of the empirical tests that substantial progress was to be made.

In subsequent years the ‘Methodology’ has been the subject of enormous controversy. There is general

agreement that in applying the theory one cannot dismiss the factual basis of the assumptions in quite such a cavalier manner. Furthermore, no one would be rash enough to declare a (refutable) theory discredited if there were a single or a few counter-examples to contradict the predictions. Such absolutism has given way to more subtle interpretations depending, as Lakatos argued, on the new and surprising insights to be obtained. Most theories coexist with small subsets of anomalous results that strictly should discredit them, and yet they remain useful theories and superior to any suggested alternative. But there is no doubt that the substance of Friedman's ‘Methodology’ has not merely stood the test of time but has also had a profound and lasting effect on the profession.

The application of this methodological approach reached its apotheosis in what most academic economists would regard as Friedman's greatest work, *A Theory of the Consumption Function* (1957). The fundamental proposition that emerged from Keynes's *General Theory* (1936) was that households expanded their consumption spending by an amount less than the increase in their current income, and that this relationship was sufficiently stable to form the basis for the multiplier through which an increase in autonomous expenditure at the macro level generated a considerably larger increase in real aggregate demand. Since the regularity and predictability of the consumption function was central for the Keynesian control of the economy, it was with trepidation that many observers found that there were considerable inconsistencies between the patterns of household behaviour, particularly from the cross-section data of household surveys and the time series of the historical record. It certainly appeared that the data were quite inconsistent with the Keynesian consumption function. Friedman showed that the Keynesian concept of household behaviour was fundamentally flawed, and that the statistical results suffered from the regression fallacy. People adjusted their consumption with respect to variations in their long-term expected (or ‘permanent’) income, and paid little heed to transitory variations. This basis idea was not new – indeed it can be found in the 18th-century writings of Bernoulli – but Friedman's development showed his genius for simplicity and for the insights of thinking concretely.

But the main quality of *A Theory of the Consumption Function* was the incomparable amassing, organization and interpretation of the evidence. The relatively low propensities to consume evident in the cross-section data were shown to be entirely consistent with the much higher propensities that emerged from analyses of aggregate time series, when both sets of figures were interpreted in the form of the permanent income hypothesis. Because of the transitory component in the cross-section samples of households, the variance of measured income exceeded the variance of permanent income, and so the slope of the regression of consumer spending on income was much lower than in the aggregate time series regressions, where the transitory component was trivially small. The permanent income hypothesis adequately passed the acid test of using little to explain much.

The integrity of scholarship was demonstrated by the diligent search to find evidence that would discredit the permanent income hypothesis. It was not, and is not, normal practice to scour the literature and statistical evidence for material that might discredit a theory. But Friedman used the hypothesis in the most imaginative way to forecast, for example, the values of regression coefficients for different groups with varying fractions of transitory to permanent income. And he left instructions for other researchers to guide them in tests to be made with further analyses of different data. One of the great contributions of this book was to give a new standard for empirical economics generally. Clearly this was how it *should* be done. The second important effect was the introduction of the concept of permanent income into virtually every field of applied economics, such as monetary economics, housing, transport and international trade. It was a new way of thinking about chance variations and

people's decisions in the real world.

A particularly fruitful theoretical approach to the utility analysis of risk and the measurement of utility, based on the work of von Neumann and Morgenstern, appeared in two papers with L. J. Savage (1948; 1952). Using axioms that most observers would regard as acceptable and reasonable, these papers showed that choice under conditions of uncertainty could be represented as a simple process of maximizing expected utility. Thus the utilities of each of the chance outcomes were weighted by the probability of that outcome, and the sum gave an index of expected utility which, given the axioms, would be maximized by choosing from the alternative uncertain prospects. Again the basic idea was not new (it was developed originally by Bernoulli in solving the St Petersburg paradox), but Friedman and Savage discovered new insights and implications, with wide-ranging applications. Apart from rationalizing the widespread practice of simultaneously gambling and insuring, the hypothesis had a profound effect on the theory and practice of portfolio selection. For the pure economic theorist it offered the attractive proposition that, up to an arbitrary linear transformation of origin and scale, utility should be regarded as a cardinal magnitude.

Subsequent discussion (particularly by Maurice Allais) suggested that one of the axioms (the so-called ‘strong independence axiom’ which asserted that the preference order would not be affected by mixing these outcomes with equiprobable alternative outcomes) was clearly implausible and violated frequently in practical decisions. Research suggested also that in some fields, for example in air passenger insurance, the expected utility hypothesis was discredited. Nevertheless, the hypothesis still forms a cornerstone of all work – and particularly practical work – in choice among risky alternatives. With some minor exceptions, these papers mark the last contributions of Friedman to the pure theory of statistics and decision-making. Many statisticians regard the diversion of such a fertile mind from its natural field as a great shame and loss.

The gain to empirical economics – and during these years, particularly to the theory of price – was, one suspects, worth the loss. The reformulation of Marshallian demand theory as a practical instrument of analysis (1949) was an exercise in meticulous scholarship in the history of thought, but one which also argued for approaching demand analysis as a positive rather than a normative discipline, an approach which he attributed to Marshall. But the analysis of economic policy, and particularly a critique of the logical structure of the arguments and the empirical evidence adduced to support proposals on economic policy, became increasingly important. Thus the critique of the arguments showing the inferiority of excise taxes compared with alternative income taxes (1952) exposed basic methodological weaknesses in what were the standard treatments of the day.

The demonstration of the uses, as well as some abuses, of the theory of price was one of the highlights of Friedman's lectures of 1946 to 1976 (with a gap from 1963 to 1973), at the graduate school of the University of Chicago. The exploitation of demand and supply as an ‘engine of discovery’ reached out well beyond those conventionally defined limits of the subject. In these lectures Friedman gave full rein to his persistence and determination to fearlessly pursue the argument, with subtlety and imagination, wherever it led. To the students it opened up new vistas – such as the theory of human capital – and exciting ways of unravelling puzzles and resolving problems. In his hands, economics had both power and point, reality and relevance (for example, 1962). As distinct from much economic work, where complicated ideas are developed in a simple way, Friedman showed how to interpret simple ideas in a most sophisticated way.

This quality characterized his work on money, which, with the inauguration of his monetary workshop

in 1951, began to be a major interest for Friedman himself and the distinguished students and faculty that he inspired. The motivations for studying money were firmly implanted when Friedman was at the Treasury dealing with wartime inflation management, but the immediate incentive was the request from the NBER to contribute a study on money for Wesley Mitchell's project on long-term business cycles. Monetary policy as a main tool of macroeconomic management was consistent with a wide degree of free unfettered enterprise and so had an obvious appeal to the liberal (which will be used here in the 19th-century sense) Friedman. The prevailing Keynesian orthodoxy, with its emphasis on expanding the public sector, appeared to threaten liberal society. The Post Keynesian contempt for money was a tempting target that was difficult to resist. But undoubtedly Friedman's imagination had been challenged by the Chicago School's preference (particularly by Knight and Simons) for rules rather than authorities in macroeconomic as well as microeconomic policy. The uncertainties of the economic environment would be much reduced if the Federal Reserve Board followed simple rules. Friedman first suggested (1948) a countercyclical rule of financing recession-induced increases in the federal budget deficit by money creation and correspondingly by retiring money during a boom-induced surplus. The empirical evidence that he explored in subsequent years, however, led him to formulate the rule of a fixed and known expansion of the money stock, rather than indulging in countercyclical operations in vain attempts to stabilize the economy. Whatever his motives, however (and one should note that motives are quite irrelevant in judging substantive propositions), for the next 30 years Friedman's work was focused on money. At last monetary economics was to be interpreted as part of the central corpus of price theory; it was to be integrated into economics.

The monetary revolution and the rise of monetarism, 1956 – 1975

In the late 1950s, to anyone subjected to the Anglo-Saxon schools of economics during the previous two decades any attempt to revive monetary economics appeared to be foolhardy, like flogging a decomposing horse. The Radcliffe Committee, advised by the most eminent economists, had reported in 1959 that the quantity of money was of little or no interest since the velocity of circulation had no limits. The quantity theory of money was subject to particular scorn as a mere identity without content. As Friedman was to point out, however, all theory consists of tautologies; all that theory does is to rearrange the implications of the axioms to produce interesting, even surprising, consequences. But they remain empty and devoid of substantive as distinct from speculative content, until they have been tested against a wide body of facts.

Of course, for many years the quantity theory of money had been tested against experience and data and over several critical periods of change. The most distinguished exponents of such tests had included Irving Fisher and Keynes himself, as well as the irrepressible Clark Warburton. Yet the methodology was murky, the statistics slim, and interrelationships between data and theory obscure. In *Studies in the Quantity Theory of Money* (1956), Friedman and his co-authors redefined the quantity theory in terms of statements specifying a degree of stability in the demand for money. It was proposed that the demand for money by the individual household would be a stable function of its money income (later thought to be permanent income or wealth) and the cost of holding money represented by the rate of interest and the expected rate of inflation.

Friedman's presentation of the theory of the demand for money in the first essay in *Studies* is one of his most widely quoted papers, primarily because it is thought to show that in presenting the money demand

function as a portfolio decision with respect to alternative assets, rather than a demand related to the flow of transactions and income, Friedman was a closet Keynesian. Substantively this was a side issue; the main point was the stability of demand, particularly with respect to nominal income or wealth. Unfortunately, this first essay was not one of Friedman's better expositions. The other essays in *Studies*, particularly that of Cagan on hyperinflations and Selden on velocity, however, established the value of examining nominal income and inflation in the context of the demand for money. The quantity theory in its new reborn Chicago form had passed its first tests.

The unknowns, however, remained legion. The vexed question of the nature of the regime controlling the supply of money, and how to interpret the problem of identifying the demand function in the data were to persist, in the eyes of many critics, as the major weakness in such studies. Was the stock of money reacting passively to changes in nominal income (or wealth) or were prices and output responding to endogenous changes in the supply of money? The Chicago workshop averred that the answer to such questions could be obtained only by painstaking research into the history of the monetary process. Undoubtedly there were occasions when the money stock passively responded to changes in nominal income, but equally obvious were instances where the money supply changed for reasons quite independent of past or contemporaneous movements in money incomes. The role of the balance of payments and the exchange rate regime was clearly recognized, and it is not difficult to discover the genesis of the monetary theory of the balance of payments in ‘Real and Pseudo Gold Standards’ (1961) and other essays in *Dollars and Deficits* (1968).

Although the detailed development of the history of the money supply process and the relationships with gold and exchange rates were to appear in the monumental *A Monetary History of the United States, 1867–1960* (1963), Friedman had already made it perfectly clear that a stable growth of the money supply was unlikely to be feasible under a regime of fixed exchange rates. His advocacy of flexible exchange rates (in 1953) followed logically on his views of the efficacy of free markets. Friedman was one of the very few economists (Gottfried Haberler and Egon Sohmen were among them) who clearly showed that the ambient dollar shortage was merely a consequence of fixed exchange rates and divergent monetary policies. His analysis was amply justified when by the 1960s, due to the change in monetary policies, the dollar shortage had turned into a dollar glut.

Yet in spite of the increasing attention paid to the balance of payments and the money supply process generally, the prime focus of Friedman's work remained the examination of the effects of monetary variations on nominal income, prices and output. The main questions were: (*a*) what was the relative importance of monetary compared with fiscal variations (the Keynes versus Monetarist debate); (*b*) what was the time pattern of adjustment; and (*c*) could expansionary financial or fiscal policies affect real output in the short or long run? The answers that evolved from Friedman's analysis were: to (*a*), although an increased fiscal deficit had an impact effect on nominal income this soon disappeared, whereas after a lag the increased rate of money growth permanently augmented the rate of inflation; to (*b*), the adjustment of nominal income to an increased rate of monetary growth involves lags that are ‘long and variable’; and to (*c*), in the long run additional monetary growth affects only the rate of inflation and has virtually no effect on either the level of output or its growth rate. In essence Friedman found that variations in the rate of growth of the money supply had short-run effects – sometimes, as in 1931 of a devastating magnitude – on real output as well as on prices; but in the long run (more than three years) the only substantial effect was on prices.

Over the 1960s and 1970s the results of Friedman's research for the long run were widely accepted. The

logic as well as the data were appealing: nominal variations (in money) have nominal effects (on prices) and no real effects (on output). But such agreement did not readily extend to his short-run claims of, first, the impotence of fiscal policy in countering cyclical oscillations and shocks; and, secondly, the large but unpredictable effects of monetary variation on real output and employment. The claims of Keynesian economists for the stability and size of the fiscal multipliers continued, but it is noteworthy that estimates of the size of the multipliers, except for those produced by the Cambridge (England) School, were substantially reduced in the 1980s. (One is not able to determine whether the economists or the economies have become less Keynesian and more monetarist.)

One of the abiding criticisms of Friedman's work on money (much of it in joint authorship with Anna Schwartz) is that it has no theoretical structure – or more charitably that such theoretical structure as exists is implicit rather than explicit. Processes of monetary transmissions as he describes them are alleged to be ‘black boxes’ with no precise specification of the way in which money works its magic. Friedman attempted to produce a theoretical underpinning for his approach to research in *Milton Friedman's Monetary Framework* (1974) by producing a seven-equation basic model of the (closed) economy. The critical difference between the Keynesian and the classical models was the choice of the last equation; the Keynesians chose to specify the price level as fixed by exogenous forces and the level of output as a variable determined by the level of aggregate demand, whereas the classical economists held that the level of real output was fixed by technology, skill and so on, and that the price level was determined by the model. With this simple model, Friedman was able to highlight the differences of method and approach as primarily different views about the size and stability of the coefficients of the system. In principle, at least, such issues could be resolved by appeals to the evidence. *The Framework* did not, however, make substantial progress in providing a sound analytical basis for the dynamics of the adjustment, through output, price and interest rate effects, to the new long-run equilibrium. The transmission mechanism and dynamics remain enshrouded in the gloom of a black box.

Yet in spite of what many theoretical economists considered to be drastic limitations for sound theoretical developments, in the most important and influential paper in macroeconomics in the post-war years, his presidential address to the American Economic Association, Friedman showed that the view of macroeconomic policy as a trade-off between unemployment and inflation was fundamentally flawed (1968). In the long run there was no such trade-off, while in the short run the trade-off took place only during the adjustment to the new inflationary environment, and then only because people were temporarily surprised by the new environment. The overriding objective of contractual arrangements was to fix real wages and prices. Money served as a veil, sometimes seductive but always obscuring underlying reality. The so-called Phillips curve was a short-term temptation rather than a long-term choice.

Friedman caught opinion at ebb and turned it into a flood. Throughout the 1960s the trade-off between unemployment and inflation appeared more and more illusory. Unemployment went up but inflation did not go down; it also increased. Into the 1970s and particularly during the great inflationary recession of 1974/75, when both inflation and unemployment reached new highs in most Organisation for Economic Co-operation and Development (OECD) countries, it appeared that only Friedman's view made any sense. Like Keynes's *General Theory*, it was one of the very few contributions that changed both the approach of professional economists and the policies adopted by finance ministers. Some time during the 1970s most governments recognized that the road to fuller employment did not lie over the high sierra of soaring inflation. Doctrinally, economists took into their toolbox the Friedman concept of a

‘natural rate’ of unemployment where inflation would neither accelerate nor decelerate. (The word ‘natural’, which was usually considered either normative or even desirable, was generally eschewed in favour of the term ‘non-accelerating inflation rate of unemployment’ or NAIRU.)

The natural level of unemployment was held to be determined by the nature of labour markets, such as the conventions of wage contracts, the degree of mobility, the level of unemployment benefits, the marginal utility of income, and many other ‘structural’ factors that are independent of the rate of inflation. As in the case of the permanent income hypothesis, to which it is distantly related, the concept had applications in fields far from labour markets. At the same time it provided one of the many missing links between the macroeconomics of aggregate output and inflation and the microeconomics of industrial adjustment and resource allocation. Again, in retrospect it all seems obvious; but that merely measures the magnitude of the contribution.

By any standards – even those of Keynes and the *General Theory* – Friedman's contribution to monetary analysis and policy must be ranked very high. Every economist, finance minister and banker felt his influence. But, as an accomplishment of the intellect, one suspects that most of Friedman's peers would still regard his work on the consumption function as the maximum maximorum of his contributions to economics. Friedman's monetary analysis did not have that sense of comprehensiveness and structural balance that are the hallmarks of his work on consumer spending. One closed *A Theory of the Consumption Function*, not with the feeling that nothing more need be said, but that whatever was discovered in the future must fit neatly into this superb and satisfying framework. The architecture could accommodate, and indeed so far has shaped and absorbed, all new contributions. The *Monetary History* and the *Framework*, however, although probably more influential in doctrine and policy, did not provide the commodious and harmonic form of the Consumption Function. A number of awkward corners left one wondering what to do. And since the theoretical plans were left obscure, sometimes there were questions whether the superstructure would really hold up. But this does not belittle the *Monetary History* so much as praise the *Consumption Function*.

1975 – 2006

The award of the Nobel Prize for Economics, long overdue in 1977, at last recorded that Friedman's great contributions had even penetrated the Swedish academies. Inevitably Friedman's rise to stardom had given many more opportunities to persuade electorates through the medium of the popular press (highlights include his columns in *Newsweek* from 1966 to 1984) and television (in the popular PBS and BBC series *Free to Choose* in 1980). His contributions to persuasive journalism delighted many, infuriated some, and made all his serious readers, if not wiser, then certainly better informed. In all these popular articles the high professional standards of integrity were maintained. But at the same time Friedman continued with his scholarly work on monetary analysis; examples include Friedman (1988; 1990; 1992). Perhaps the main output, after more than 20 years of effort, was his book with Anna J. Schwartz, *Monetary Trends in the United States and the United Kingdom, Their Relation to Income, Prices and Interest Rates, 1867–1975* (1982).

The main methodological decision lying behind this study was that, since there was too much inexplicable variation in short-run variations in income and money, it was best to ignore these and concentrate on comparing the cyclical phase averages. These would screen out the short-term effect and would enable an analysis to be made of the underlying long-term money–income–interest relationships.

Even for this team of Friedman and Schwartz, the treatment of the data and the integrity of their analysis reached new heights of meticulous scholarship.

Yet, considering the enormous value of the input of time and energy, the results are, as the authors confess, hardly worth the cost. For the most part the study confirms, and demonstrates with comparative data for the United States and the United Kingdom, the basic propositions on velocity, real income, prices and interest rates that had emerged in the *History*.

In his final decades, it may be claimed that Friedman had fallen prey to the same temptations that affected Alfred Marshall. For many years of his mature professional career, Marshall spent much of his time revising and refining his great *Principles*. In retrospect it seemed to be a great loss to scholarship that Marshall did not leave the *Principles* well alone and turn to his projected study of the economics of the state. The opportunity was missed. It would be, however, a travesty to draw a close parallel between Friedman and Marshall in their mature years. Perhaps with the example of Marshall in mind, Friedman had generally launched his studies on the profession and then left them largely to fend for themselves. (The only exception is the textbook *Price Theory*: *A Provisional Text*, 1962, which was revised in 1976.)

Yet there is a sense in which Friedman, trapped by his immense success in monetary economics, had been prevented from deploying his mind in scholarly work in other fields of economics.

The possibilities are revealed in Friedman's more popular writings on issues such as public spending, price and rent control, taxation, and many issues in microeconomics. Characteristic flashes of insight and phrase, together with the innovations of approach – especially the simplifications – give the professional reader a tantalizing taste of what might have been yet another great contribution to economic science. Many economists have always believed that, in spite of his great strides in money, Friedman's relative advantage was always in the study of price theory and its manifest applications. There is the measure of the man.

The public image of Friedman

The conventional view of Friedman is that he is one of the most ardent and most effective advocates of free enterprise and monetarist policies over the six decades 1945 to 2006. If far short of his wishes, the success of his advocacy has by any objective standard been enormous. Opinion in Western countries, even among the clerisy, has moved decisively in its preference for those economic freedoms that he has so eloquently advocated.

It is not possible to parcel out any neat attribution of influence on these great changes in attitude and policy. Friedman himself would probably give by far the largest weight to the experience of the 1970s, particularly the disappointments over failure to restrain the growth of government spending and the great inflation from 1965 to 1981. The explanation of these events and the development of an alternative strategy, with institutions that would ensure individual economic liberty and freedom from inflation, have been, in the public perception, Friedman's great contribution to the reforms. In his appearances in the various media he was a great persuader, and he played a critical role in promoting such ideas as an all volunteer army, the voucher schemes for education and health, and indexing income tax. In effectiveness, breadth and scope, his only rival among the economists of the 20th century is Keynes.

See Also

- monetarism
- quantity theory of money

Selected works

1934. Professor Pigou's method for measuring elasticities of demand from budgetary data. *Quarterly Journal of Economics* 1, 151–63.

1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32, 675–701.

1940. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics* 11, 86–92.

1945. (With S. Kuznets.) *Income from Independent Professional Practice*. New York: NBER.

1948a. (With L.J. Savage.) The utility analysis of choices involving risk. *Journal of Political Economy* 56, 279–304.

1948b. (With H.A. Freeman, F. Mosteller, and W. Allen Wallis.) *Sampling Inspection*. New York: McGraw-Hill.

1949. The Marshallian demand curve. *Journal of Political Economy* 57, 463–95. Repr. in (1953).

1952. (With L.J. Savage.) The expected utility hypothesis and the measurability of utility. *Journal of Political Economy* 60, 463–74.

1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

1956. *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

1957a. (With G.S. Becker.) A statistical illusion in judging Keynesian models. *Journal of Political Economy* 65, 64–75.

1957b. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

1959. *A Program for Monetary Stability*. New York: Fordham University Press.

1962a. *Capitalism and Freedom*. Chicago: University of Chicago Press.

1962b. *Price Theory: A Provisional Text*. Chicago: Aldine.

1963a. (With A.J. Schwartz.) *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press for the NBER.

1963b. (With D. Meiselman.) The relative stability of monetary velocity and the investment multiplier in the United States, 1897–1958. In *Stabilization Policies*. Englewood Cliffs, NJ: Prentice-Hall.

1967. (With R.V. Roosa.) *The Balance of Payments: Free versus Fixed Exchange Rates*. AEI Rational Debate Seminar. Washington, DC: American Enterprise Institute.

1968a. The role of monetary policy. Presidential Address, American Economic Association, 29 December 1967. *American Economic Review* 58, 1–17. Repr. in (1969).

1968b. *Dollars and Deficits: Inflation, Monetary Policy and the Balance of Payments*. Englewood Cliffs, NJ: Prentice-Hall.

1969. *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

1970. (With A.J. Schwartz.) *Monetary Statistics of the United States*. New York: Columbia University Press for the NBER.

1972a. (With Wilbur J. Cohen.) *Social Security: universal or selective?* AEI Rational Debate Seminar. Washington, DC: American Enterprise Institute.

1972b. *An Economist's Protest: Columns on Political Economy*. Glen Ridge, NJ: Thomas Horton.

1973. *Money and Economic Development*. Horowitz Lectures of 1972. New York: Praeger.

1974. *Milton Friedman's Monetary Framework: A Debate with His Critics*, ed. R.J. Gordon. Chicago: University of Chicago Press.

1976. *Price Theory*. Chicago: Aldine. (Revised and enlarged version of the 1962 edn).

1978. *Tax Limitation, Inflation and the Role of Government*. Dallas: Fisher Institute.

1980. (With R. Friedman.) *Free to Choose*. New York: Harcourt Brace Jovanovich.

1982. (With A.J. Schwartz.) *Monetary Trends in the United States and the United Kingdom*. Chicago: University of Chicago Press, for the NBER.

1984. (With R. Friedman) *Tyranny of the Status Quo*. San Diego, New York and London: Harcourt Brace Jovanovich.

1988. Money and the stock market. *Journal of Political Economy* 96, 221–45.

1990. Bimetallism revisited. *Journal of Economic Perspectives* 4(4), 85–104.

1991. (With A. Schwartz.) Alternative approaches to analyzing economic data. *American Economic Review* 81, 39–49.

1992. Do old fallacies ever die? *Journal of Economic Literature* 30, 2129–32.

1992. *Money Mischief: Episodes in Monetary History*. San Diego and London: Harcourt Brace Jovanovich.

1998. (With R. Friedman.) *Two Lucky People: Memoirs*. Chicago: University of Chicago Press.

Bibliography

Hammond, J.D. 1999. *The Legacy of Milton Friedman as Teacher*. Cheltenham: Edward Elgar.

Hammond, J.D. 2005. *Theory and Measurement: Causality Issues in Milton Friedman's Monetary Economics*. Cambridge: Cambridge University Press.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. New York: Macmillan.

Wald, A. 1947. *Sequential Analysis*. New York: John Wiley.

Wood, J. 1990. *Milton Friedman; Critical Assessments*. London: Routledge.

How to cite this article

Walters, Alan. "Friedman, Milton (1912–2006)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000222> doi:10.1057/9780230226203.0604

The New Palgrave Dictionary of Economics Online

Frisch, Ragnar Anton Kittel (1895 – 1973)

P. Nørregaard Rasmussen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

acceleration principle; Clark, J. M.; demand analysis; Econometric Society; econometrics; Frisch, R. A. K.; input–output analysis; Institute of Economics (Norway); macroeconomics, theory of; Marshall, A.; methodology of economics; national accounting; production, theory of; statistics and economics

Article

Frisch lived a long, varied and extremely productive life. He graduated in economics at the University of Oslo in 1919 (although as a son of a goldsmith he ‘supplemented’ this by finalizing his apprenticeship as a goldsmith in 1920!). He studied in France from 1921 to 1923 and in Britain in 1923; was an associate at the University of Oslo from 1925 and received his doctorate in 1926 in mathematical statistics (Frisch, 1926a). Further studies abroad in the USA, France and Italy (1927–8) were followed by an associate professorship at the University of Oslo (1928) and a full professorship in 1931. Frisch was head of the (newly established) Institute of Economics in Oslo from 1932 to his retirement in 1965. He was also chief editor of *Econometrica* (1933–55), followed by his chairmanship of the editorial board. He was one of the founders (1930) and, in fact, the driving force behind the creation of the Econometric Society. He was a member of a number of national and international expert committees and adviser on several occasions to developing countries (India 1954–5 and Egypt several times over the years 1957–64). He received honorary doctorates from a number of universities (*inter alia* Stockholm, Copenhagen, Cambridge, Birmingham) and was – together with Jan Tinbergen – the first (1969) to receive the Alfred Nobel Memorial Prize in Economics. In addition he received (as the first recipient) the Schumpeter Prize (1955), the Feltrinelli Prize (1956) and three *Festschriften*. He was a visiting professor or guest lecturer to a number of universities – Yale, Minnesota, Paris, Pittsburgh, for example – and he was a very active participant at numerous international meetings of economists, statisticians and mathematicians. In the late 1940s there was a joke among Norwegian students that he was also a ‘visiting’ professor in Oslo. This was unfair. In particular during the 1930s he put a lot of effort into his teaching and was writing a series of lecture notes, most of them seminal, though many remained unpublished. The impressive list of his publications (Haavelmo, 1973) and activities could be continued because he was a genius, cutting

through problems like a warm knife through butter, and because his working power was extraordinary. To survey his contributions is not easy for the simple reason that there is scarcely any area of economics Frisch has not been into and left his imprint. To cooperate with him was not always easy. He was too strong, as shown by the fact that he seldom had co-authors. The list of his published and printed works comprises about 160 items. But to this should be added a long series of mimeographed contributions – many will recall the ‘Memoranda fra Socialkonomisk Institutt’ – from about 1946 onwards. They amounted altogether to 6,500 pages, and most of them are still awaiting publication (though Frisch himself argued that ‘for editing it needs a very good man, and if he is good enough, he should write himself’).

Frisch began his academic work in the theory of mathematical statistics. This profession today acknowledges his early contributions and regrets his departure from it, though admitting that in terms of the more applied theory of statistics he made noticeable contributions later on. His years in Paris, where he concentrated on mathematics, were not in vain.

It is, however, in economics that Frisch made his name, He was at most of the centres and many of the corners of the subject. One may, however, also argue that his most significant contribution is in economic methodology. This comes out not only in his applications of methods but also in their general presentation. A very good example was written overnight in a hotel room at Colmar, after a day's discussions at a meeting of the Econometric Society. The article (Frisch, 1936b) is a classic, clearing the ground about the very meaning of static versus dynamic analysis. This is by now elementary, but it is elementary because of Frisch. In his principal works on methodology, he used and unified the tools he had mastered so well: economic theory, mathematics and statistics. It is no accident that he invented the word ‘econometrics’, for in general he enriched our methodological vocabulary by a number of precise concepts: macro- versus micro-analysis, statics versus dynamics, exogenous versus endogenous variables, the concept of autonomous relations, the problem of identification of relations, confluent relations, decision models, conjectural behaviour (of firms) – a complete list would be very long.

Few would hesitate to agree that Frisch ‘created’ econometrics in the modern sense of the word. It is much more notable that he warned again and again against misuses of the new tools. In the first issue of *Econometrica* in 1933 he wrote: ‘The policy of *Econometrica* will be as heartily to denounce futile playing with mathematical symbols in economics as to encourage their constructive use.’ In Frisch (1970), he argued that ‘the econometric army has now grown to such proportions that it cannot be beaten by the silly arguments that were used against us previously. This imposes on us a *social and scientific responsibility* of high order in the world of today’ (p. 153). But in the very same article he also stressed (p. 163) that ‘I have insisted that econometrics must have relevance to concrete realities – otherwise it degenerates into something which is not worthy of the name econometrics, but ought rather to be called playometrics’.

Always underlying Frisch's contributions to methodology were his consistent efforts to turn economics into a precise science, quantifying the variables and the structures. This is different from the traditional ‘on the one hand and on the other’, where on balance the answer is left in the air. But this ‘aggressive’ view also presents new challenges. The economist must be prepared for the troublesome work of gathering data, to face the difficulties in estimating structures and in the end to attempt a balanced interpretation of the outcome. Frisch saw this and contributed to this debate throughout his career; illustrations might be Frisch (1933a; 1934b; 1936a; 1939). These and many other contributions had a profound influence and wide applications in pre-war as well as post-war econometrics. Again, and sadly

enough, one could also refer to a number of unpublished papers, though these were influential as contributions to scientific gatherings. A supreme example is a paper on ‘Statistical versus Theoretical Relations in Macrodynamics’, a contribution to a conference sponsored by the League of Nations in 1938 to discuss Jan Tinbergen's work for the League of Nations on the trade cycle.

One may wonder why Frisch left so many path-breaking contributions without taking the trouble to publish them. I think there is a double answer. On the one hand Frisch was an impatient man: if he had given the gist of the solution to a problem, he tended to go on to new problems. On the other hand, he was extremely careful: a publication going to the printer had to be perfect and finalized – a troublesome process which he often tended to avoid. Haavelmo (1973), reports that Frisch often argued that proofreading is one of the most difficult and important tasks of a scientist.

The general assessment above can be verified by considering Frisch's contributions in the field of demand analysis, the theory of production and the theory of macroeconomics.

In demand analysis he began as early as 1926 (Frisch, 1926b), by formulating a number of basic axioms and from these to deduce a theory of demand. Thus utility functions were not postulated but were derived from more basic axioms, all of these being formulated as being, in principle, open to testing. It may be fair to say that his work in this field culminated in Frisch (1959). It is a tribute to his work that it has in fact been used in practice, for example in Norwegian planning.

In the theory of production Frisch was a forerunner, formulating the theory in a strict mathematical form but also applying it on concrete problems. An example is Frisch (1935). However, most of his works were in the form of mimeographed lecture notes in the 1930s and remained unpublished until 1962 and later (Frisch, 1962;1963). The main results, however, were internationally known through the works of Schneider and Carlson, who at times were research associates in Oslo and very much influenced by Frisch.

Also in the theory of macroeconomics, Frisch was at the front, even, it can be argued, ahead of Keynes. Anybody reading his booklet, Frisch (1933b), and the subsequent articles in *Econometrica* (1934a), might be willing to argue that Frisch made it first. He shows convincingly how a capitalist economy may go into a deadlock when, to put it in a simple way, the tailor cannot sell to the shoemaker because the shoemaker cannot sell to the tailor:

> … the cause of great depressions, such as the one we are actually in, is … fundamentally connected with the fact that modern economic life has been divided into a number of regions or groups.
>
> Under the present system, the blind ‘economic laws’ will under certain circumstances, create a situation where these groups are forced mutually to undermine each other's position. Each group is forced to curtail the use goods produced and services rendered by the other groups, which, in turn, will cause a still further contraction of the demand for its own products, and so on. (Frisch, 1934a, pp. 259f.)

He also, in the 1934 articles, outlined (a couple of years before Leontief) an input–output analysis. His contributions in these areas were not appreciated at the time, but from a historical perspective they are path-breaking. This also holds for his contribution to the (famous) Cassel *Festschrift* (Frisch, 1933c), where a dynamic system for the economy as a whole was outlined and where he distinguished in a sharp

and fruitful way between the impulses and the propagation mechanism. In this context one may also, as an illustration of his interest in the development of economic theory, make a reference to his excellent analysis of Marshall (Frisch, 1950).

There is a direct line from here to his systems of national accounts (first published in Frisch, 1939) which had a profound influence on the planning in Norway and elsewhere after the Liberation. In the context of macroeconomics it is illustrative to mention Frisch's discussion with J.M. Clark over the acceleration principle (Frisch, 1931, 1932). In an amazingly simple way Frisch cleared up the issue, that is, the interplay between the pure acceleration principle and the reinvestment cycle, as the following quotation shows:

Let *z* be consumer-taking [in present day language this is simply consumption] per unit of time, *w* capital production per unit of time, and *W* the capital stock that exists at any moment of time. All the three magnitudes *z*, *w*, and *W* are, of course, functions of time. In practice they would be represented by time series.

Let us, for simplicity, make the following two assumptions: A. Consumer-taking *z* is the same as the production of the consumer good, and this again is at any time proportional to the existing capital stock *W*. In other words, we have

$$W = kz, \tag{1}$$

where *k* is a constant independent of time. B. The depreciation per unit of time *u*, that is to say, the capital production that is needed for replacement purposes, is proportional to the existing capital stock. In other words, we have

$$u = hW, \tag{2}$$

where *h* is a constant independent of time. Now, the rate of change with respect to time of the capital stock is equal to

$$\dot{W} = w - u \tag{3}$$

By virtue of (1) we have, however,

$$\dot{W} = k\dot{z}, \tag{4}$$

where $\dot{z}$ is the rate of change of consumer-taking. Inserting this into (3), and expressing *u* in terms of *z* by (2) and (1), we get

$$k\dot{z} = w - khz.$$

So that we finally have

$$w = k(hz + \dot{z}). \tag{5}$$

The rate of change with respect to time of capital production is thus equal to

$$\dot{w} = k(h\dot{z} + \ddot{z}). \tag{6}$$

Formula (5) indicates the two parts of which total capital production is made up. In the first place we have the part *khz* that represents capital production for replacement purposes. This part is (under our simplified assumption) proportional to the *size* of consumer-taking. In the second place, we have the part $k\dot{z}$ representing capital production for expansion purposes. This part is (under the present simplified assumption) proportional to the *rate of change* of consumer-taking. Thus there are two forces that act upon total capital production. If consumer-taking is increasing, but at a constantly decreasing rate, the first of these two forces tends to increase, and the second tends to slow down capital production. Which one of the two forces shall have the upper hand depends on the *manner* in which the increase in consumer-taking slows down, and it depends also on *the rate of depreciation* (Frisch, 1931, pp. 647f.). As will be seen, it is all so simple, *provided* the problem is *formulated* clearly. And formulating problems in a fruitful way was one of his secrets.

What is a genius? It might be argued that Frisch up till now was one of the ten in our profession in the 20th century. Not that he cannot be criticized. On occasion he used his brains more or less in vain, for

example, on unimportant calculating schemes. Long after electronic calculators were on the market, he used time and effort on inventing schemes for inverting a matrix on a simple desk calculator. His various methods for the solution of programming problems – ‘the logarithmic potential method’, ‘the multiplex method’ and ‘the nonplex method’ (for example, Frisch, 1956; 1957; 1961a; 1961b) – are still disputable, taking present-day techniques into account. In other words, he might have had a weak point in not always being able to evaluate the importance of a problem; that is, he might now and then have used his immense working power on issues where his opportunity costs were too high.

Even so, his life work is impressive. And so was the man himself. His political attitude was rather to the left than to the right – while at the same time he was a devout Christian. He felt a strong social responsibility, as proved through his work on the problems of the 1930s as well as, and perhaps even more so, by his consciousness towards the less developed countries. He could at times be a bit harsh on colleagues who did not live up to his own standards for serious work. At the same time he was extremely kind and helpful to students doing their best. He never failed to encourage. And few will forget when his strong blue eyes were shining with joy.

Selected works

1926a. Sur les semi-invariants et moments employés dans I’étude des distributions statistiques. *Skrifter utgitt av det Norske Videnskaps-Akademi i Oslo II. Hist.-Filos. Klasse 1926*, No. 3.

1926b. Sur un problème d’économie pure. (An attempt at developing an axiomatic foundation of utility as a quantitative notion and at measuring statistically the variation in the marginal utility of money on the basis of data from the Paris Cooperative Society.) *Norsk Matematisk Forenings Skrifter* Series I, No. 16, 1–40.

1931, 1932. The Interrelation between capital production and consumer taking. *Journal of Political Economy* 39(5) (1931), 646–54. A rejoinder, 40(2) (1932), 253–5. A final word, 40(5) (1932), 694.

1933a. *Pitfalls in the Statistical Construction of Demand and Supply Curves*. Leipzig: H. Buske.

1933b. *Sparing og Cirkulasjonsregulering*. Oslo.

1933c. Propagation problems and impulse problems in dynamic economics. In *Economic Essays in Honor of Gustav Cassel*. London: Allen & Unwin. Reprinted in *Readings in Business Cycles*, ed. R.A. Gordon and L.R. Klein. London: Allen & Unwin, 1966.

1934a. Circulation planning. *Econometrica*, Pt I, 2(3), 258–336; Pt II, 2(4), 422–35.

1934b. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: University Institute of Economics.

1935. The principle of substitution. An example of its application in the chocolate industry. *Nordisk*

Tidsskrift for Teknisk Økonomi, September, 12–27.

1936a. Annual survey of general economic theory: the problem of index numbers. *Econometrica* 4(1), 1–38.

1936b. On the notion of equilibrium and disequilibrium. *Review of Economic Studies* 3(2), 100–105.

1939. Nasjonalregnskapet. *Transactions of the Nordisk Statistiker Møte*, Oslo.

1943. *Økosirk-systement. 1943. Ekonomisk Tidskrift* 45, 106–21.

1950. Alfred Marshall's theory of value. *Quarterly Journal of Economics* 64, 495–524.

1952. Frisch on Wicksell. In *The Development of Economic Thought*, ed. H.W. Spiegel. New York: John Wiley & Sons.

1956. La résolution des problèmes de programme linéaire par la méthode du potential logarithmique. In *Séminaire économétrique*, ed. R. Roy. Paris: CNRS.

1957. The multiplex method for linear programming. *Sankhya* 18(3–4), 329–60.

1959. A complete scheme for computing all direct and cross demand elasticities in a model with many sectors. *Econometrica* 27(2), 177–96.

1961a. Mixed linear and quadratic programming by the multiplex method. In *Money, Growth and Methodology, Fesiskrift til Johan Akerman*, ed. H. Hegeland. Lund: Gleerup.

1961b. Quadratic programming by the multiplex method in the general cases where the quadratic form may be singular. *Bulletin of the International Statistical Institute* 38(4) (Tokyo), 283–332.

1962. *Innledning til Produksjonsteorien*. Oslo: Universitets Forlaget.

1963. *Lois techniques et économiques de la production*. Paris: Dunod. Trans. as *Theory of Production*, Dordrecht: D. Reidel; Chicago: Rand McNally, 1965.

1970. Econometrics in the world of today. In *Induction, Growth and Trade; Essays in Honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F. Scott and J.N. Wolfe. London: Clarendon Press.

Bibliography

Arrow, K.J. 1960. The work of Ragnar Frisch, econometrician. *Econometrica* 28(April), 175–92.

Edvardsen, K.N. 1970. A Survey of Ragnar Frisch's contributions to the science of economics. *Economist* 118(2), 174–96.

Edvardsen, K.N. 2001. *Ragnar Frisch: An Annotated Bibliography*. Report 4/2001. Ragnar Frisch Centre for Economic Research, University of Oslo. Online. Available at http://www.frisch.uio.no/pdf/rapp01_04.pdf, accessed 9 November 2006.

Haavelmo, T. 1973. Minnetale over Professor, dr. philos. Ragnar Frisch. *Arbok del Norske Videnskabs-Akademi i Oslo*.

Johansen, L. 1969. Ragnar Frisch's contributions to economics. *Swedish Journal of Economics* 71(4), 302–24.

How to cite this article

Nørregaard Rasmussen, P. "Frisch, Ragnar Anton Kittel (1895–1973)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000225> doi:10.1057/9780230226203.0605

The New Palgrave Dictionary of Economics Online

full and limited information methods

Thomas J. Rothenberg

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Econometricians have developed a number of alternative methods for estimating parameters and testing hypotheses in simultaneous equations models. Some of these are limited information methods that can be applied one equation at a time and require only minimal specification of the other equations in the system. In contrast, the full information methods treat the system as a whole and require a complete specification of all the equations.

The distinction between limited and full information methods is, in part, simply one of statistical efficiency. As is generally true in inference problems, the more that is known about the phenomena being studied, the more precisely the unknown parameters can be estimated with the available data. In an interdependent system of equations, information about the variables appearing in one equation can be used to get better estimates of the coefficients in other equations. Of course, there is a trade-off: full information methods are more efficient, but they are also more sensitive to specification error and more difficult to compute.

Statistical considerations are not, however, the only reason for distinguishing between limited and full information approaches. Models of the world do not come off the shelf. In any application, the choice of which variables to view as endogenous (i.e. explained by the model) and which to view as exogenous (explained outside the model) is up to the analyst. The interpretations given to the equations of the model and the specification of the functional forms are subject to considerable discretion. The limited information and full information distinction can be viewed not simply as one of statistical efficiency but one of modelling strategy.

The simultaneous equations model can be applied to a variety of economic situations. In each case, structural equations are interpreted in light of some hypothetical experiment that is postulated. In considering the logic of econometric model building and inference, it is useful to distinguish between two general classes of applications. On the one hand, there are applications where the basic economic question involves a single hypothetical experiment and the problem is to draw inferences about the parameters of a single autonomous structural equation. Other relationships are considered only as a means for learning about the given equation. On the other hand, there are applications where the basic economic question being asked involves in an essential way an interdependent system of experiments. The goal of the analysis is to understand the interaction of a set of autonomous equations.

An example may clarify the distinction. Consider the standard competitive supply demand model where price and quantity traded are determined by the interaction of consumer and producer behaviour. One can easily imagine situations where consumers are perfectly-competitive price takers and it would be useful to know the price elasticity of market demand. One might be tempted to use time-series data and regress quantity purchased on price (including perhaps other demand determinants like income and prices of substitutes as additional explanatory variables) and to interpret the estimated equation as a demand function. If it could plausibly be assumed that the omitted demand determinants constituting the error term were uncorrelated over the sample period with each of the included regressors, this interpretation might be justified. If, however, periods where the omitted factors lead to high demand are also the periods where price is high, then there will be simultaneous equations bias. In order to decide whether or not the regression of quantity on price will produce satisfactory estimates of the demand function, the mechanism determining movements in price must be examined. Even though our interest is in the behaviour of consumers, we must consider other agents who influence price. In this case a model of producer behaviour is needed.

This example captures the essence of many econometric problems: we want to learn about a relationship defined in terms of a hypothetical isolated experiment but the data we have available were in fact generated from a more complex experiment. We are not particularly interested in studying the process that actually generated the data, except in so far it helps us to learn about the process we *wish* had generated the data. A simultaneous equations model is postulated simply to help us estimate a single equation of interest.

Some economic problems, however, are of a different sort. Again in the supply–demand set-up, suppose we are interested in learning how a sales tax will affect market price. If tax rates had varied over our sample period, a regression of market price on tax rate might be informative. If, however, there had been little or no tax rate variation, such a regression would be useless. But, in a correctly specified model, the effects of taxes can be deduced from knowledge of the structure of consumer and producer decision making in the absence of taxes. Under competition, for example, one needs only to know the slopes of the demand and supply curves. Thus, in order to predict the effect of a sales tax, one might wish to estimate the system of structural equations describing market equilibrium.

The distinction between these two situations can be summarized as follows: in the one case we are interested in a structural equation for its own sake; in the other case our interest is in the reduced-form of an interdependent system. If our concern is with a single equation, we might prefer to make few assumptions about the rest of the system and to estimate the needed parameters using limited information methods. If our concern is with improved reduced-form estimates, full-information approaches are natural since specification of the entire system is necessary in any case. A further discussion of these methodological issues can be found in Hood and Koopmans (1953, chs 1 and 6).

Limited information methods

Consider a single structural equation represented by

$$y = Z\alpha + u \tag{1}$$

where y is a T-dimensional (column) vector of observations on an endogenous variable, Z is a $T \times n$ matrix of observations on n explanatory variables, α is an n-dimensional parameter vector, and u is a T-dimensional vector of random errors. The components of α are given a causal interpretation in terms of some hypothetical experiment suggested by economic theory. For example, the first component might represent the effect on the outcome of the experiment of a unit change in one of the conditions, other things held constant. In our sample, however, other conditions varied across the T observation. The errors represent those conditions which are not accounted for by the explanatory variables and are assumed to have zero mean.

The key assumption underlying limited-information methods of inference is that we have data on K predetermined variables that are unrelated to the errors. That is, the error term for observation t is uncorrelated with each of the predetermined variables for that observation. The $T \times K$ matrix of observations on the predetermined variables is assumed to have rank K and is denoted by X. By assumption, then, $E(X' u)$ is the zero vector. Some of the explanatory variables may be predetermined and hence some columns of Z are also columns of X. The remaining explanatory variables are thought to be correlated with the error term and are considered as endogenous. Implicitly, equation (1) is viewed as part of a larger system explaining all the endogenous variables. The predetermined variables appearing in X but not in Z are assumed to be explanatory variables in some other structural equation. Exact specification of these other equations is not needed for limited information analysis.

In most approaches to estimating α it is assumed that nothing is known about the degree of correlation between u and the endogenous components of Z. Instead, the analysis exploits the zero correlation between u and X. The simplest approach is the method of moments. Since $X' u$ has mean zero, a natural estimate of α is that vector a satisfying the vector equation $X'(y-Za)=0$. This is a system of K linear equations in n unknowns. If K is less than n, the estimation method fails. If K equals n, the estimate is given by $(X' Z)^{-1}X' y$, as long as the inverse exists. The approach is often referred to as the method of instrumental variables and the columns of X are called instruments.

If K is greater than n, any n independent linear combinations of the columns of X can be used as instruments. For example, for any $n \times K$ matrix D, α can be estimated by

$$(D'X'Z)^{-1}D'X'y, \tag{2}$$

as long as the inverse exists. *Often D* is chosen to be a selection matrix with each row containing zeros except for one unit element; that is, n out of the K predetermined variables are selected as instruments and the others are discarded. If Z contains no endogenous variables, it is a submatrix of X, least squares can then be interpreted as instrumental variables using the regressors as instruments.

The estimator (2) will have good sampling properties if the instruments are not only uncorrelated with the errors but also highly correlated with the explanatory variables. To maximize that correlation, a

natural choice for D is the coefficient matrix from a linear regression of Z on X. The instruments are then the predicted values from that regression. These predicted values (or projections) can be written as NZ where N is the idempotent projection matrix $X(X'X)^{-1}X'$; the estimator becomes

$$(Z'NZ)^{-1}Z'Ny \tag{2'}$$

Because $N=NN$, the estimator (2) can be obtained by simply regressing y on the predicted values NZ. Hence, this particular instrumental variables estimator is commonly called *two-stage least squares*. The two-stage least-squares estimator is readily seen to be the solution of the minimization problem

$$\min (y-Za)'N(y-Za). \tag{3}$$

As an alternative, it has been proposed to minimize the ratio

$$\frac{(y-Za)'N(y-Za)}{(y-Za)'M(y-Za)} \tag{4}$$

where $M=1-N$ is also an idempotent projection matrix. This yields the *limited-information maximum-likelihood* estimator. That is, if the endogenous variables are assumed to be multivariate normal and independent from observation to observation, and if no variables are excluded a priori from the other equations in the system, maximization of the likelihood function is equivalent to minimizing the ratio (4). This maximum likelihood estimate is also an instrumental variable estimate of the form (2). Indeed, the matrix D turns out to be the maximum likelihood estimate of the population regression coefficients relating Z and X. Thus the solutions of (3) and (4) are both instrumental variable estimates. They differ only in how the reduced-form regression coefficients used for D are estimated.

The sampling distribution of the instrumental variable estimator depends, of course, on the choice of D. The endogenous variables in Z are necessarily random. Hence, the estimator behaves like the ratio of random variables; its moments and exact sampling distribution are difficult to derive even under the assumption of normality. However, large-sample approximations have been developed. The two-stage least-squares estimate and the limited information maximum-likelihood estimate have, to a first order of

approximation, the same large-sample probability distribution. To that order of approximation, they are optimal in the sense that any other instrumental variable estimators based on X have asymptotic variances at least as large. The asymptotic approximations tend to be reasonably good when T is large compared with K. When $K - n$ is large, instrumental variable estimates using a subset of the columns of X often outperform two-stage least squares. Further small-sample results are discussed by Fuller (1977).

Full information methods

Although limited-information methods like two-stage least squares can be applied to each equation of a simultaneous system, better results can usually be obtained by taking into account the other equations. Suppose the system consists of G linear structural equations in G endogenous variables. These equations contain K distinct predetermined variables which may be exogenous or values of endogenous variables at a previous time period. The crucial assumption is that each predetermined variable is uncorrelated with each structural error for the same observation.

Let $y_1,\ldots, y_G$ be T-dimensional column vectors of observations on the G endogenous variables. As before, the $T \times K$ matrix of observations on the predetermined variables is denoted by X and assumed to have rank K. The system is written as

$$y_i = Z_i\alpha_i + u_i (i = 1, \ldots, G) \tag{5}$$

where Z_i is the $T \times n_i$, matrix of observations on the explanatory variables, u_i is the error vector, and α_i is the parameter vector for equation i. Some of the columns of Z_i, are columns of X; the others are endogenous variables.

Again, estimates can be based on the method of moments. Consider the set of GK equations

$$X'(y_i - Z_i a_i) = 0 \quad (i = 1, \ldots, G) \tag{6}$$

If, for any i, K is less than n_i, the corresponding parameter α_i cannot be estimated; we shall suppose that any equation for which this is true has already been deleted from the system so that G is the number of equations whose parameters are estimable. If $n_i=K$ for all i, the solution to (6) is obtained by using limited information instrumental variables on each equation separately. If, for some i, $n_i<K$, the system (6) has more equations than unknowns. Again, linear combinations of the predetermined variables can be used as instruments. The optimal selection of weights, however, is more complicated than in the limited-information case and depends on the pattern of correlation among the structural errors.

If the structural errors are independent from observation to observation but are correlated across equations, we have the specification

$$E(u_i u_j') = \sigma_{ij} I (i, j = 1, \ldots, G)$$

where the σ's are error covariances and *I* is a *T*-dimensional identity matrix. As a generalization of (3), consider the minimization problem

$$\min \sum_i \sum_j (y_i - Z_i a_i)' N (y_i - Z_j a_j) a^{ij}$$

(7)

where the σ^{ij} are elements of the inverse of the matrix [σ_{ij}]. For given σ's, the first-order conditions are

$$\sum_j Z_i' N (y_i - Z_j a_j) a^{ij} = 0 \; (i = 1, \ldots, G)$$

(8)

which are linear combinations of the equations in (6). It can be demonstrated that the solution to (8) is an instrumental variables estimator with asymptotically optimal weights. In practice, the σ's are unknown but can be estimated from the residuals of some preliminary fit. This approach to estimating the α's is called *three-stage least squares* since it involves least-squares calculations at three stages, first to obtain the projections NZ_j, again to obtain two-stage least-squares estimates of the σ's, and finally to solve the minimization problem (7). For details, see Zellner and Theil (1962).

If the structural errors are assumed to be normal, the likelihood function for the complete simultaneous equations system has a relatively simple expression in terms of the reduced-form parameters. However, since the reduced form is nonlinear in the structural parameters, analytic methods for maximizing the likelihood function are not available and iterative techniques are used instead. Just as in the limited-information case, the maximum-likelihood estimator can be interpreted as an instrumental variables estimator. If in (8) the least-squares predicted values NZ_i are replaced by maximum-likelihood predictions and if the σ's are replaced by their maximum-likelihood estimates, the resulting solution is the (full-information) maximum-likelihood estimate of the α's. See Malinvaud (1970, ch. 19) for details.

At one time full-information methods (particularly those using maximum likelihood) were computationally very burdensome. Computer software was almost non-existent, rounding error was hard to control, and computer time was very expensive. Many econometric procedures became popular simply because they avoided these difficulties. Current computer technology is such that computational burden is no longer a practical constraint, at least for moderate-sized models. The more important constraints at the moment are the limited sample sizes compared with the number of parameters to be estimated and limited confidence we have in the orthogonality conditions that must be imposed to get any estimates at all.

See Also

- econometrics
- identification
- instrumental variables
- simultaneous equations models
- two-stage least squares and the *k*-class estimator

Bibliography

Fuller, W. 1977. Some properties of a modification of the limited information estimator. *Econometrica* 45(May), 939–53.

Hood, W. and Koopmans, T., eds. 1953. *Studies in Econometric Method*. Cowles Foundation Monograph No. 14, New York: Wiley.

Malinvaud, E. 1970. *Statistical Methods of Econometrics*. 2nd edn, Amsterdam: North-Holland.

Zellner, A. and Theil, H. 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica* 30(January), 54–78.

How to cite this article

Rothenberg, Thomas J. "full and limited information methods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000228> doi:10.1057/9780230226203.0606

The New Palgrave Dictionary of Economics Online

Fullarton, John (1780? – 1849)

Roy Green
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Banking School; convertibility; Currency School; Fullarton, J.; law of reflux; Ricardo, D.; Tooke, T.

Article

John Fullarton shared at least one characteristic with his great predecessor, Ricardo: he also seemed, in the words of Lord Brougham, ‘as if he had dropped from another planet’. Although Fullarton is described in the *Dictionary of National Biography* as a ‘traveller and writer on the currency’, travel occupied by far the greater proportion of his life, along with a keen interest in the world of art and literature. Yet the single published work on which his considerable reputation as an economist is based had an impact comparable with that of Ricardo's intervention in the Bullion Controversy at the turn of the century.

In his early twenties, Fullarton became a surgeon in India and found time to edit a Calcutta newspaper. There he subsequently made a fortune in banking and began the first of his extensive tours through ‘our eastern possessions’, as the *Dictionary of National Biography* endearingly calls them. On this tour, Fullarton collected vast amounts of information and made many notes of his observations, but these were never published. In 1823, having returned to England to live, he contributed articles to the *Quarterly Review* on the reform crisis; however, it was not long before he resumed his travels, this time around Britain and the continent in a coach specially fitted with a library. In 1833, as a Fellow of the Royal Asiatic Society, Fullarton went again to India, and, in the following year, to China; but his zeal evaporated along with his fortune as a result of the failure of his bankers, and he moved back to London permanently.

It was in 1844, during the passage of the Bank Charter Act through the House of Commons, that Fullarton published his major work, *On the Regulation of Currencies*, subtitled ‘an examination of the principles on which it is proposed to restrict, within certain fixed limits, the future issues on credit of the Bank of England, and of the other banking establishments throughout the country’. It was immediately hailed as a formidable challenge to the Currency School orthodoxy, whose support for the Bank Charter Act had overwhelmed Tooke's lonely opposition in the opening round of the ‘currency-banking debate’.

Indeed, according to Gregory, Fullarton's ‘penetrating tract’ was ‘perhaps the most subtle and able production emanating from the Banking School’ (introduction to Tooke, 1838/57, p. 81).

Fullarton's aim was a simple one: to bolster Tooke's case against what they both saw as ill-conceived banking legislation; in doing so, however, he not only improved its presentation, but also developed the theoretical basis of the argument in a number of important respects, taking the opportunity to lament the fact that ‘Mr. Tooke himself has been exceedingly slow in following out his original conclusions on the subject of price to all their consequences’ (1844, p. 18).

The Currency School had asserted that convertibility would not be a sufficient safeguard against the overissue of bank notes and their consequent depreciation; and that the quantity of notes in circulation would have to be regulated in accordance with the movement of bullion across the foreign exchanges. The response of Fullarton and the Banking School took three main lines. First, starting from the assumption that legal convertibility necessarily implied economic convertibility, they pointed out that any discrepancy between the note issue and a purely metallic system arose from the Currency School's erroneous theory of metallic circulation rather than from the supposed autonomy of the notes. Second, any effect on prices attributed to bank notes could not be denied to a range of financial assets excluded by the Currency School from their definition of money. Third, bank notes were in any case not money but credit, and therefore never could be overissued, though the credit structure as a whole might be extended beyond the limits of real accumulation by speculation. It was in this context that Fullarton developed the famous ‘law of reflux’, which he called ‘the great regulating principle of the internal currency’ (1844, p. 68).

Tooke, in turn, warmly welcomed Fullarton's analysis in the subsequent volume of his massive *History of Prices*, and gave some indication of the surprise he must have experienced upon its publication:

> [L]est his estimate of the value of my contributions to an extension of the knowledge of this subject, should be ascribed to the bias of friendship, I think it right to state that the distinguished author was unknown to me, except by name and reputation, till after the publication of his treatise, and that I had not the slightest knowledge of such a work being in preparation. (1838/57, vol. 4, pp. x–xi)

Tooke then paid Fullarton the compliment of quoting extensively from his work, repeatedly praising the ‘wonderful clearness and vigour which distinguish his writings’ (vol. 5, p. 537). Nor was Fullarton above self-promotion: it appears that he had a hand in a *Quarterly Review* article, ‘The Financial Pressure’, which saw the crisis of 1847 as confirming the warnings of ‘Mr Fullarton's masterly treatise’ (see Fetter, 1965, p. 212).

It is certainly true that Fullarton's work ‘enjoyed, in England and on the Continent, a persistent success such as few contributions to an ephemeral controversy have ever enjoyed’ (Schumpeter, 1954, p. 725). Marx, for example, included Fullarton among ‘the best writers on money’ (1867, p. 129); in his view, ‘the economic literature worth mentioning since 1830 resolves itself mainly into a literature on currency, credit, and crises’ (1894, pp. 492–3). Hilferding, too, drew heavily on Fullarton (Hilferding, 1910); and even Keynes was impressed with his ‘most interesting’ contribution to monetary thought (Keynes, 1936, p. 364 n.). Many of Fullarton's arguments later resurfaced in the Radcliffe Report of 1959, and are still today being ‘rediscovered’. As Fullarton himself pointed out (1844, p. 5), ‘this is a subject on which

there never can be any efficient or immediate appeal to the public at large. It is a subject on which the progress of opinion always has been, and always must be, exceedingly slow.'

See Also

- Banking School, Currency School, Free Banking School

Selected work

1844. *On the Regulation of Currencies*. London: John Murray.

Bibliography

Fetter, F.W. 1965. *Development of British Monetary Orthodoxy, 1797–1875*. Fairfield: Kelley, 1978.

Hilferding, R. 1910. *Finance Capital: A Study of the Latest Phase of Capitalist Development*. London: Routledge & Kegan Paul, 1981.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Marx, K. 1867. *Capital*, vol. 1. Moscow: Progress Publishers, 1959.

Marx, K. 1894. *Capital*, vol. 3. Moscow: Progress Publishers, 1971.

Radcliffe (Lord) (Chairman). 1959. *Committee on the Working of the Monetary System*. London: HMSO.

Schumpeter, J. 1954. *History of Economic Analysis*. London: George Allen & Unwin, 1955.

Tooke, T. 1838/57. *History of Prices, and of the State of the Circulation from 1792 to 1856*, 6 vols. London: P.S. King & Son, 1928.

How to cite this article

Green, Roy. "Fullarton, John (1780?–1849)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000234> doi:10.1057/9780230226203.0607

The New Palgrave Dictionary of Economics Online

functional analysis

Leonid Kantorovich and Victor Polterovich

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A branch of mathematics mainly concerned with infinite-dimensional vector spaces and their maps, functional analysis is so called because elements (points) of certain important specific spaces are functions. The necessity of considering infinite-dimensional models arises in economics in many problems, including assessment of random effects in a situation with an infinite number of natural states; study of effects arising from a ‘very large’ number of participants; problems of spatial economics; study of economic development in continuous time, in particular, with due regard for lags; economic growth on an infinite time interval; and the influence of commodity differentiation on exchange processes.

Keywords

competition models; competitive equilibrium; economic growth in the very long run; extension principle; fixed-point theorems; functional analysis; global analysis; hyperplanes; infinite-dimensional models; Kakutani theorem; mathematical economics; measure theory; monopolistic competition; openness principle; product differentiation; separation theorems; spatial economics; spectral analysis; uniform boundedness principle

Article

Functional analysis is a branch of mathematics mainly concerned with infinite-dimensional vector spaces and their maps. Elements (points) of certain important specific spaces are functions, hence the term ‘functional analysis’.
An important role in the development of functional analysis was played by set theory, abstract algebra and axiomatic geometry. General topology, measure theory, differential equations and some other branches of mathematics evolved in close contact with functional analysis, so that it is difficult to indicate where these disciplines end and functional analysis begins.
The fundamental ideas of functional analysis appeared at the turn of the 20th century; by the 1920s it

had already evolved into an autonomous discipline. Among its founders were Banach, Fréchet, Hadamard, Hilbert, von Neumann, Riesz and Volterra.

The creation of functional analysis resulted in basic changes in the approach to many mathematical problems. The study of individual functions and equations was replaced by that of families of such objects. Abstract forms of investigation ensured a unified approach to questions which seemed distant at first glance; they were instrumental in finding more general, yet deeper and more concrete relationships. From the outset, the development of functional analysis was stimulated by the intrinsic requirements of mathematics, as well as by applications, especially to quantum mechanics. Today the language of functional analysis is actually used in all of continuous mathematics. Its methods have become the foundation of a whole series of new branches of research, both theoretical and applied, such as the theory of random processes, differential topology, dynamic systems, optimal control theory, mathematical programming, and so on. Functional methods penetrate deeper and deeper into theoretical physics and into different engineering disciplines. These methods find more and more widespread applications in mathematical economics.

Spaces studied in functional analysis usually belong to the class of linear (vector) topological spaces, that is, linear spaces supplied with a topology (a system of open sets and hence a notion of limit), for which the linear operations are continuous. A narrower class of spaces is metric vector spaces, for which distance between points is defined. The distance is given by a function (the metric, assigning a non-negative number to each pair of vectors) which possesses certain specific properties of ordinary distance. The topology in such spaces is naturally induced by the metric.

An important subclass of metric spaces is normed spaces, that is, linear spaces in which to each element *x* a non-negative number $\|x\|$, called the norm of *x*, is assigned, and the following conditions are satisfied:

1. (1) $\|x\| = 0$ if and only if $x = 0$;
2. (2) $\|\lambda x\| = |\lambda|\|x\|$ for any scalar λ (homogeneity);
3. (3) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

The norm is an abstraction of the notion of ‘vector length’. The function $d(x, y) = \|x - y\|$ is the metric in normed spaces. It is said that a sequence x_t of elements converges to the element *x* in the strong topology, if $\|x_t - x\| \to 0$ as $t \to \infty$. A normed space is said to be a Banach space if it is complete; this means that any of its fundamental sequences (that is, such that $\|x_t - x_s\| \to 0$ as $t, s \to \infty$) has a limit. Banach spaces often appear in applications.

A Banach space *X* is said to be a Hilbert space if it is supplied with a numerical function (*x*,*y*), called scalar product of vectors $x, y \in X$, related to the norm by the identity $\|x\|^2 = (x, x)$ and satisfying the conditions:

1. (1) (*x*, *y*) and (*y*, *x*) are complex conjugates (in particular, for real vector spaces, $(x, y) = (y, x)$);
2. (2) $(\lambda_1 x_1 + \lambda_2 x_2, y) = \lambda_1 (x_1, y) + \lambda_2 (x_2, y)$;
3. (3) $(x, x) \geq 0$ and $(x, x) = 0$ only if $x = 0$.

The scalar product makes it possible to characterize the ‘angle between vectors’ and, in particular, to introduce the notion of orthogonal vector. As a result, the geometry of Hilbert spaces is close to Euclidean geometry.

Let us present some examples of specific spaces. The space $l_p (1 \leq p < \infty)$ of all numerical sequences $x = (\alpha_n)$ with the norm

$$||x|| = \left(\sum_{n=1}^{\infty} |\alpha_n|^p \right)^{1/p}$$

is a Banach space. For p=2 it is a Hilbert space if the scalar product is defined by the formula

$$(x, y) = \sum_{n=1}^{\infty} \alpha_n \bar{\beta}_n, \; x = (\alpha_n), \; y = (\beta_n),$$

where $\bar{\beta}_n$ is the complex number conjugate to β_n. The space $L_2(a, b)$ of all real functions defined on the closed interval $[a, b]$, square integrable in the Lebesgue sense, is a Hilbert space (functions which differ on a set of zero measure are identified) if the scalar product is defined by the formula

$$(x, y) = \int_a^b x(t)\, y(t) \mathrm{d}t .$$

$L_2(a, b)$ is a particular case of the Banach spaces $L_p (1 \leq p \leq \infty)$ of functions defined on so-called measure spaces. The theory of the spaces L_p is part of the foundations of probability theory, where the functions from L_p are interpreted as random variables. For $p \neq 2$ the spaces l_p and L_p are not Hilbert spaces.

Another important example is the Banach space $C(S)$ – the collection of all continuous scalar functions on the compact space S, with the norm

$$||x|| = \max_{s \in S} ||x(s)||.$$

All the spaces listed above are infinite dimensional, that is, contain an infinite subset of linearly independent vectors (the notion of linear independence here is the same as in linear algebra). A finite dimensional vector space may be transformed into a Banach space in many different ways by appropriate choices of norms, but the convergence in any norm will be equivalent to the coordinate one. Although many facts of classical analysis can be generalized to Banach spaces, the infinite dimensional theory is essentially different from the finite dimensional one in many ways. One of the reasons is that a bounded sequence (with respect to norm) in a Banach space does not necessarily contain any fundamental subsequences and therefore may have no limiting points; such is the sequence $l_n, n = 1, 2, \ldots$ in l_2, whose *n*th element l_n is the vector all of whose coordinates are zero, except the nth, which equals 1.

A function from one space into another is often said to be an operator. Operators with scalar values are called functionals. The operators most thoroughly studied are the linear ones. An operator *T* from the vector space *X* to the vector space *Y* is called linear if

$$T(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 T(x_1) + \lambda_2 T(x_2)$$

for all $x_1, x_2 \in X$ and arbitrary scalars λ_1, λ_2. In particular, the derivation and integration operations determine linear operators for appropriate choices of the spaces *X*, *Y*. If *X* and *Y* are finite dimensional, linear operators from *X* to *Y* are determined by matrices.

The theory of linear operators in Banach spaces is one of the most developed sections of functional analysis. It is a far-reaching generalization of linear algebra and, in particular, of matrix theory. However, the purely algebraic approach is insufficient in the infinite dimensional case. One of the reasons is the necessity of distinguishing continuous and discontinuous linear operators (continuity is not an algebraic notion), while for operators in finite dimensional space linearity implies continuity.

For a linear operator from one Banach space to another to be continuous, it is necessary and sufficient that it be bounded, that is, that it map bounded sets into bounded sets.

The set *B*(*x*, *y*) of continuous linear operators from *X* to *Y* is a linear space with respect to the natural operations of addition and multiplication by scalars. This set becomes a Banach space if the norm $\|T\|$ of the operator *T* is defined by the formula

$$\|T\| = \sup_{\|x\| \le 1} \|T(x)\|.$$

In the particular case when *Y* is the set of scalars, we get the Banach space X^* of all linear continuous functionals on *X*, which is called adjoint to *X*. The study of adjoint spaces is not only of intrinsic interest but is also needed to obtain deeper results about the initial space *X*.

The adjoint space of an *n*-dimensional space is also *n*-dimensional. The space adjoint to l_p coincides, in a certain sense, with the space l_q, where $1/q + 1/p = 1$ (a similar statement holds for L_p). A complete description of linear continuous functionals has been obtained for many specific spaces. We only mention F. Riesz's famous theorem describing the general form of a linear continuous functional on the space *C(S)* of continuous functions. In the particular case when *S* is the closed interval [*a, b*] on the numerical line, any element $f \in C^*(a, b)$ can be represented in the form

$$f(x) = \int_a^b x(t)\,\mathrm{d}\varphi(t),$$

where ϕ is a function of bounded variation.

The operation of taking adjoint spaces can be iterated, yielding a sequence of Banach spaces X, X^*, X^{**}, … each of which is adjoint to the previous one. Each vector $x \in X$ can be viewed as an element of the second adjoint space X^{**} by putting $x(f) = f(x)$ for any $f \in X^*$; the functional thus defined is linear, continuous and its norm coincides with $\|x\|$. If all the elements of X^{**} can be represented in this way, the initial Banach space X is called reflexive.

In certain aspects reflexive spaces have more resemblance to finite dimensional ones than do non-reflexive spaces.

A sequence x_n in a Banach space X is said to converge weakly to $x \in X$ if $f(x_n) \to f(x)$ as $n \to \infty$ for any functional $f \in X^*$. This definition implicitly supplies X with the weak topology which differs, as a rule, from the original one. The consideration of different versions of convergence on the same linear space and the study of their relationships is typical of functional analysis.

Among the numerous facts of Banach space it is customary to single out three theorems which, because of their importance and manifold applications, are known as the main principles of linear analysis.

The extension principle (Hahn–Banach Theorem) states that every continuous linear functional defined on a subspace of a normed space can be extended to the entire space, preserving norm. Using this principle it is possible to prove so-called separation theorems, which claim that under appropriate conditions two non-intersecting convex sets in a Banach space may be separated by a hyperplane, that is, a set of the form $\{x \mid f(x) = \alpha\}$, where f is a non-zero continuous linear functional and α is a scalar. Separation theorems make possible the wide use of geometric ideas in the study of Banach spaces.

The uniform boundedness principle (Banach–Steinhaus Theorem) states that a sequence of linear continuous operators $T_n \in B(X, Y)$ is pointwise convergent, that is, $T_n(x) \to T(x)$ as $n \to \infty$ for all $x \in X$ if and only if the two following conditions hold:

1. (1) such a convergence takes place on a set of arguments whose linear envelope is dense in X;
2. (2) the norms of all the T_n are uniformly bounded with respect to n.

According to the openness principle (Banach Theorem), any continuous linear operator from one

Banach space to another sends open sets into open sets.
The development of the theory of linear operators, especially at its initial stage, was stimulated by the problem of solving linear operator equations.

$$T(x) = y \tag{1}$$

where *x*, *y* are elements of infinite dimensional spaces.
The similarity between linear functionals and algebraic equations, previously noted for linear differential equations, turned out to be just as productive for integral equations, whose foundations were laid at the beginning of the century by Fredholm, Hilbert, Noether and Volterra.
An exhaustive theory has only been constructed for certain classes of equations (1). In particular, the case when $T = I + K$ where *I* is the identity operator and *K* is compact (that is, maps bounded sets into sets with compact closure) has been conclusively studied. Compact operators often appear in applications and are very similar to finite dimensional ones.
In the study of operator equations and in many applications of operator theory a leading role is played by the notion of spectrum. The spectrum of a continuous linear operator *T* defined in a complex Banach space is by definition the set of all scalars λ for which the operator $T - \lambda I$ has no inverse, that is, $T - \lambda I$ is either not injective (one-to-one) or not surjective (onto). Non-zero solutions of the equation $T(x) = \lambda x$ are called eigen-vectors of the operator *T*, while the values of λ for which such solutions exist are its eigen-values. All the eigen-values are contained in the spectrum, but, unlike the finite dimensional case, the spectrum may also contain other values. A compact operator has a spectrum containing a finite or countable number of distinct numbers; in the latter case they converge to zero. Spectral analysis – the branch of functional analysis studying the properties of operator spectra – has achieved penetrating advances in the theory of Banach and operator algebras (Gelfand, von Neumann).
A linear operator *T* in Hilbert space is called self-adjoint if $(T(x), y) = (x, T(y))$ for all *x*, *y*. A compact self-adjoint operator has properties similar to that of a symmetric matrix; for example, there exists an orthonormal basis consisting of its eigen-vectors (Hilbert–Schmidt Theorem).
Among the branches of functional analysis beyond the framework of the theory of Banach spaces, the theory of distributions (or ‘generalized functions’), initially developed (by Sobolev and Schwartz) as a rigorous foundation for formal operations with δ-functions used in physics, should be mentioned.
In many theoretical and applied problems – in particular, in mathematical economics – it is necessary to consider semi-ordered vector spaces, characterized by the fact that some of their elements are involved in a comparison relation. The most important are those semi-ordered spaces for which every bounded (in the sense of the order relation) subset possesses a least upper bound. The foundations of the theory of such spaces were developed in the 1930s by Kantorovich and are called Kantorovich spaces (K-spaces). For example, the spaces l_p and L_p have a natural partial order relation: one sequence is greater than another, if all the coordinates of the first are greater than the corresponding coordinates of the second; the function *x* is greater than *y* if $x(t)$ is greater than $y(t)$ for almost all *t*. A somewhat wider class is constituted by vector lattices, in which the existence of l.u.b. is guaranteed only for finite sets. In semi-

ordered spaces the notion of positive (not necessarily linear) operator can be introduced in a natural way; this notion has been used to generalize the theory of positive matrices.

Positive operators are an important class of maps studied in non-linear functional analysis. Another important class – the monotone operators – includes operators in Hilbert space satisfying the inequality

$$(T(x) - T(y), x - y \leq 0) \text{ for all } x, y.$$

A third example is that of contraction operators, i.e. operators such that

$$\|T(x) - T(y)\| < \alpha\|x - y\| \text{ for some } \alpha < 1.$$

For those (and some other) classes of non-linear operators, conditions for the existence and uniqueness of operator equation solutions have been obtained in global terms. But, just as in classical analysis, the most universal means of studying nonlinear problems is the differential calculus. Many facts of classical differential calculus (in particular, Taylor expansions and the implicit function theorem) have been generalized to Banach spaces.

Among the main instruments of mathematical economics, convex analysis and fixed-point theorems should be noted. Both are in essence branches of functional analysis. The recent extremely rapid advances in convex analysis have been stimulated by the requirements of the theory of extremal problems in abstract spaces (mathematical programming and optimal control). A typical extremal problem is to find the maximum of the functional $f(x)$ defined on the subset G of the space X under the constraints $T(x) \geq 0$, $x \in G$ where T is an operator from X to a linear topological space Y supplied with the partial order $\geq$. As in the finite-dimensional situation, here the necessary and sufficient conditions for the existence of an extremum (under appropriate assumptions) may be stated in terms of saddle points of the Lagrange function

$$L(x, y^*) = f(x) + y^*(T(x)),$$

where the Lagrange multiplier y^* is an element of the space Y^* adjoint to Y. In deducing this condition, separation theorems, the differential calculus and theorems on the representation of linear functionals play a fundamental role.

In order to solve functional equations and extremal problems in functional spaces, various computational procedures have been developed. In particular, generalizations of gradient methods and Newton's method have been obtained (the first results here are due to Kantorovich); the Newton–Kantorovich

method also turned out to be a powerful means of proving existence and uniqueness of solutions.
Another approach to computational problems is based on the approximation of the given functional equation by a simpler one. The application of functional analysis methods leads to a general theory of such approximation methods within whose framework the rate of convergence is studied and error estimates are given for a series of computational procedures.
In certain cases approximate solutions may be obtained by computer in analytic rather than numerical form (‘deductive computations’).
The necessity of considering infinite-dimensional models arises in economics in many problems, among which the following may be distinguished: (1) assessment of random effects in a situation with an infinite number of natural states; (2) study of effects arising from a ‘very large’ number of participants (competition models); (3) problems of spatial economics; (4) study of economic development in continuous time, in particular, with due regard for lags; (5) economic growth on an infinite time interval; (6) influence of commodity differentiation on exchange processes. This list is not exhaustive.
As a rule, it is possible in principle to use a finite dimensional model and then pass to the limit if necessary. However, the ‘infinite dimensional’ statement of the problem is often easier to study because a more powerful analytic apparatus may be applied.
The concept of adjoint (dual) spaces mentioned above is of fundamental importance in economics. In a typical case the elements of the given space are interpreted as utilized and produced goods, while elements of the adjoint space (continuous linear functionals) are prices; the value of the functional on the given product vector determines its cost (expenditures, profits, and so on). Then semi-ordered vector spaces, expressing the ‘greater than’ relationship for certain pairs of expenditure and production vectors and taking into consideration the positivity of prices, turn out to be a natural instrument.
In the use of functional analysis methods, a very delicate question is that of choosing the functional space into which the model should be ‘embedded’; it is closely related to the chosen estimate of economic and social values.
As an example let us consider a problem of type (5). In stating dynamical optimal planning problems considerable difficulties are involved in the choice of a plan horizon and objectives for the end of a planning period. However, in many cases the initial interval of the optimal trajectory depends very weakly on these parameters and is close to the corresponding interval of the optimal (in a certain sense) infinite trajectory. This is one of the reasons growth on an infinite time interval is worth studying.
For a wide class of models it is possible to show that any optimal trajectory is the result of maximizing integral profits calculated in appropriately chosen prices. An effective way of studying this question is the following. Let us embed the set of all admissible trajectories of economic growth (that is, trajectories satisfying technological and resource constraints) in an appropriate Banach space X so that the adjoint space X^* is interpreted as the space of prices; the value of a continuous linear functional on a vector $x \in X$ may be interpreted as the integral of the profits obtained in motion along the trajectory x. The set of trajectories which are better than the optimal one does not intersect the set of admissible trajectories. Under appropriate conditions these two sets may be separated by a hyperplane. The corresponding continuous linear functional will determine the required price trajectory. Using this approach, it is possible to investigate the relationship between competitive equilibrium and optimum for an infinite time interval.
Another example of productive application of functional analysis concerns the influence of commodity differentiation on market processes, a problem occupying an important place in the theory of

monopolistic competition. In the simplest case, product differentiation is characterized by a scalar parameter assuming values in the closed interval [*a*, *b*]. Each consumer may choose any finite number of different goods (that is, a finite number of points t_i on the interval) and acquire them in arbitrary quantities x_i as long as he satisfies his budget restrictions for the given prices. It is natural to assume that the price $p(t)$ depends continuously on the characteristic of the product $t \in [a, b]$, i.e. $p(t) \in C(a, b)$. The result of a consumer's choice is a finite set of pairs x_i, t_i which determines a continuous linear functional in the price space $C(a, b)$ according to the rule

$$z(p) = \sum_i x_i p(t_i);$$

then $z \in C^*(a, b)$. But $C(a, b)$ can be identified with a subset of its second adjoint space (see above). Thus, as usual, price is a continuous linear functional of the space of collections of goods $C^*(a, b)$. The fact that this space is adjoint to a certain Banach space considerably facilitates its study, since adjoint spaces possess useful topological properties. The analysis of models based on this construction yields conditions under which a market with differentiated commodities and ‘small’ participants, similar to contemporary competitive markets, ensures an optimal distribution of resources (Mas-Colell, 1975).

The proof of the existence of competitive equilibrium in the finite dimensional case is based on fixed-point theorems. Several such theorems, including the Kakutani theorem, are also valid for Banach spaces; however, in this case their application becomes more difficult because of the essential trait of infinite dimensional spaces mentioned previously – the non-compactness of the unit sphere. Another trait of infinite dimensional spaces is that special conditions are required for the separability of non-intersecting convex sets. Both of these circumstances considerably complicate the study of economic models.

In discussing the economic applications of functional analysis, two other disciplines closely related to it – measure theory and global analysis – should be mentioned. The first is widely used in the study of probabilistic models, as well as in models with a continuum of participants or products (see Hildenbrand, 1974; Mas-Colell, 1975). Global analysis, introduced into mathematical economics by Debreu and Smale, allowed us to understand the deeper structures of the sets of equilibrium states and to advance to the solution of equilibrium stability problems (see Smale, 1981).

Above we mentioned some applications of functional analysis to economics. In their turn, the problems of economics have influenced the development of mathematics. This is natural since economics is a vast field of research, differing in principle from those classical physical and mathematical disciplines on the basis of which functional analysis developed. The theory of systems of linear inequalities developed a hundred years later than the theory of linear equations, and precisely because of the needs of economics. Another interesting and important example is the transportation problem, which was first studied under the name of mass shifting problem by Kantorovich in 1942. The metric introduced in its study (interpreted as the expenditures required to shift a unit mass) has found numerous applications in functional analysis and some other fields. Many mathematical problems from functional analysis

originating in economics still await their solution. In particular, the functional equations describing macroeconomic dynamics taking into account the differentiation of funds according to their time of creation have not been exhaustively studied (for example, see Kantorovich, Zhiyanov and Khovansky, 1978). It can be expected that further advances in the mathematical analysis of economics will become an even more powerful source in the development of mathematical methods, including functional analysis.

See Also

- calculus of variations
- non-standard analysis
- Pontryagin's principle of optimality
- Roos, Charles Frederick

Bibliography

Dunford, N. and Schwartz, J.T. 1958. *Linear Operators*. New York: Interscience Publishers.

Ekeland, I. and Temam, R. 1976. *Convex Analysis and Variational Problems*. Vol. 1 of *Studies in Mathematics and Its Applications*. Amsterdam: North-Holland.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Kantorovich, L.V. and Akilov, G.P. 1984. *Functional Analysis*. London: Pergamon Press.

Kantorovich, L.V., Zhiyanov, V.I. and Khovansky, A.G. 1978. The principle of differential optimization as applied to a singleproduct dynamical economic model. *Sibirski matematicheskii zhurnal* 19, 1053–64.

Kutateladze, S.S. 1983. *Foundations of Functional Analysis*. Novosibirsk: Nauka.

Mas-Colell, A. 1975. A model of equilibrium with differentiated commodities. *Journal of Mathematical Economics* 2, 263–95.

Schaefer, H.H. 1971. *Topological Vector Spaces*. New York: Springer.

Smale, S. 1981. Global analysis and economics. In *Handbook of Mathematical Economics*, vol. 1, ed. K. J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

How to cite this article

Kantorovich, Leonid and Victor Polterovich. "functional analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000236> doi:10.1057/9780230226203.0608

The New Palgrave Dictionary of Economics Online

functional central limit theorems

Werner Ploberger

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Functional limit theorems are generalizations of classical central limit theorems. They allow us not only to approximate the distributions of sums of random variables, but also describe their temporal evolution. The necessary mathematical concepts as well as some sufficient conditions for convergence to a random walk are discussed.

Keywords

central limit theorems; convergence; functional limit theorems; general limit theorems; Gordin's th; invariance principle; likelihood; Lindeberg condition; martingale differences; random walk; separability; Skorohod metric

Article

Central limit theorems guarantee that the distributions of properly normalized sums of certain random variables are approximately normal. In many cases, however, a more detailed analysis is necessary. When testing for structural constancy in models, we might be interested in the temporal evolution of our sums. So for random variables X_i we are interested in analysing the behaviour of

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{t} X_i \tag{1}$$

as a function of t for $t \leq N$. It is convenient to normalize the time, too, and consider for $0 \leq z \leq 1$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{Tz} X_i.$$
(2)

Another popular application is the asymptotic behaviour of the empirical distribution function or its multivariate generalizations, though we will only briefly discuss it.
‘Functional limit ths’ are generalizations of the classical central limit theorem (CLT). Instead of analysing random variables with values in **R**, I deal with random variables in more general spaces. Here I discuss only one specific example, namely the analysis of the properly normalized partial sums of random variables. In order to do so, I will first sketch the necessary concepts concerning the topology of the spaces involved. In particular, I want to demonstrate the necessity of using spaces and metrics which, at the first glance, may not look that plausible. The results are well known and can be found in many textbooks. A classical reference is Billingsley (1999). Another introduction in this field, more geared towards econometricians, is Davidson (1994).

Foundations: metric spaces and convergence in distribution

A common framework, allowing us to formulate more general limit theorems, assumes that our ‘random variables’ take values in so-called ‘Polish spaces’, which are just metric spaces which are separable and complete. So let us assume that we have given such a space E, with a metric $d(.,.)$ on it, so that there exists a countable dense subset and that the space is complete (that is, every Cauchy sequence converges). Examples are the finite-dimensional spaces with the usual distance. The space $C[0,1]$ of all continuous functions from [0,1] to **R** (the set of real numbers), endowed with the metric

$$d_M(x, y) = \max_{0 \le t \le 1} |x(t) - y(t)|.$$
(3)

Let us assume that we have random variables X_n, X with values in E: Then we define convergence in distribution of X_n to X

$$X_n \to^D X$$
(4)

if and only if for all bounded, continuous functions ϕ from E to **R**,

$$E\varphi(X_n) \to E\varphi(X). \tag{5}$$

We can easily see that, in the special case of the space E being the set **R**, our definition here is a generalization of the familiar concept of convergence in distribution. An 'invariance principle' is simply a statement convergence in distribution of random variables in a complex space.
If we have given a statement like (4), then for continuous ϕ and large n we can approximate the distribution of $\phi(X_n)$ by the distribution of $\phi(X)$. As an example, assume our underlying space is $C[0,1]$ (defined above), and our distance is given by (3). Suppose we have $X_n \to^D X$. We can easily see that the functions attaching each $z \in C[0,1]$ $\max_{0 \le t \le 1} z(t)$ or $\int_0^1 z(t)^2 dt$ are continuous with respect to our metric.
Hence we can immediately conclude that

$$\max_{0 \le t \le 1} X_n(t) \to \max_{0 \le t \le 1} X(t) \tag{6}$$

or

$$\int_0^1 X_n(t)^2 dt \to \int_0^1 X(t)^2 dt, \tag{7}$$

where '$\to$' stands for the usual convergence in distribution of real-valued random variables. Sometimes it is, however, burdensome to establish continuity for some functionals, or we might even be forced to consider discontinuous functionals. In this kind of situation the following theorem is helpful. Since we only work in separable, metric spaces a function ϕ defined on a general metric space E is continuous at a point $x \in E$ if for all $x_n \to x$ $\phi(x_n) \to \phi(x)$. Otherwise the ϕ is called discontinuous in x, and let D_ϕ be the set of all points where ϕ is discontinuous. Now assume we have some random elements X_n, X and

$$X_n \to^D X.$$

Then we have the following theorem.
Theorem 1: Suppose that

$$P([X \in D_\varphi]) = 0.$$

Then

$$\varphi(X_n) \to^D \varphi(X).$$

If the discontinuities of ϕ are a null set with respect to the limiting distribution, the distributions of $\phi(X_n)$ can be approximated better and better by the distribution of $\phi(X)$.
In any case, the usefulness of functional limit theorems depends on the set of continuous functions associated with our space. On the one hand, a metric with 'many' continuous functions will allow us to establish many limiting relationships like (6) or (7). On the other hand, it will be harder to establish convergence, since we have to show the relation (5) for *more* functions ϕ. Hence we have to compromise.

The space D[0,1]

The first and most important application of functional limit theorems is the analysis of partial sums. When dealing with normalized sums like (1), (2) we encounter the first problem: we can easily let the time t or z be a continuous variable. but then the sum (1),(2) is a discontinuous function. Hence we have to look at spaces more general than $C[0,1]$. One such space is the space $D[0,1]$, defined to be the space of all bounded functions f which have only 'jumps' as discontinuities: at every time z the limits to the right and left of f ($f(z+0)$ and $f(z-0)$ exist).
Next we have to define a distance between the functions f, g from $D[0,1]$. The first candidate, namely the supremum-norm (3), has the disadvantage that the corresponding space is not separable: consider for each $a \in (0,1)$ the functions f_a defined as

$$f_a(z) = \begin{cases} 0 \text{ if } z < a \\ 1 \text{ if } z \geq a. \end{cases}$$

(8)

Then we can easily see that in the supremum norm (3), the distance between f_a and f_b is equal to 1. Since we have more than countable real numbers in (0,1), we cannot have a countable dense subset.
A distance better suited to this space is the so-called Skorohod metric. Let us first define the set Λ to be the set of all functions from [0,1] to [0,1] which are monotonically increasing, continuous, and map 0 and 1 into 0 and 1, respectively. Then define

$$d_S(f, g) = \inf_{\lambda \in \Lambda} \sup_z (|f(z) - g(\lambda(z))| + |z - \lambda(z)|).$$

The Skorohod distance is related to the maximal distance. The main difference, however, is that we do not compare the functions *f* and *g* for the same values. The Skorohod metric allows us to 'bend' the argument a little. This rather small modification has enormous consequences. The corresponding space is separable: that is, there exists a countable dense subset. The metric itself is not complete (that is, there exist Cauchy sequences which do not converge). There exists, however, an equivalent metric (that is, a metric which determines the same open sets, neighbourhoods, convergent subsequences, continuous functions,…) which is complete. Moreover, we can easily see that

$$d_S(f, g) \leq d_M(X, Y),$$

so convergence in the maximum distance implies convergence in the Skorohod metric.
The next question is the set of continuous functions. We can easily see that some of the usual candidates, like for example the functional mapping each *f* to $\sup_{0 \leq z \leq 1} f(z)$, are continuous. The functional mapping *f* to *f*(*z*), however, is for 0<*z*<1 *not* continuous. Hence th 1 will come in handy.
The most important types of limiting processes will all have continuous trajectories. Hence, the class of functionals covered by th 1 contains all functionals which are continuous in *C*[0,1]. For establishing this continuity, we have an interesting criterion.
Theorem 2: Suppose we have $f \in C[0,1]$, *and* $f_n \in D[0,1]$ *so that* $f_n \to f$ *in the Skorohod metric. Then we have convergence in the supremum metric* (3), *too*.
This result may explain the usefulness of *D*[0,1]. On the one hand, the metric on *D*[0,1] is weak enough to allow for separability. This has, however, the drawback that it is hard to establish continuity of a function in the general case.
If the limiting random element lies with probability 1 in *C*[0,1], however, it is easy to check the requirements of th 1 for a function $\varphi: D[0, 1] \to \mathbb{R}$. One only has to show that $\varphi(f_n) \to \varphi(f)$ if $f_n \to f$ *uniformly*, which is much easier to handle.

Examples for limit theorems

In this section, I want to bring some examples of functional limit theorems. Together with the discussion above, they can be used as ‘building blocks’ for the derivation of general limit theorems.
The first functional limit theorem is one of the most important, namely, the functional limit theorem for martingale differences. This theorem is of utmost importance in many statistical applications: the scores of the conditional likelihood functions are martingale differences. Furthermore, the theorem is quite general. It only assumes a Lindeberg condition (which is quite similar to the case of the classical central limit th) and some kind of normalization condition. The role of the standard normal distribution is played by the ‘standard random walk’ W. W is a random element with values in $C[0,1]$ (that is, a random function) with the following properties:

- $W(0)=0$.
- W is ‘Gaussian’. All finite-dimensional marginal distributions $(W(z_0),\ldots,W(z_k))$ are Gaussian with expectation 0.
- The covariance of $W(z_1)$ and $W(z_2)$ is $\min(z_1, z_2)$.

A quite tedious but well known proof shows that there exists such a random element, and that its distribution (the induced probability measure on $C[0,1]$) is unique. Moreover, it is easy to show that W has all the properties associated with a random walk: its increments are independent from past values.
Theorem 3: (McLeish, 1974): *Suppose we have given a triangular array of random variables* $X_{i,n}$. $1 \leq i \leq n$, *together with some adapted* σ*-algebras* $\mathfrak{F}_{i,n}$ *so that*

$$E(X_{i,n} / \mathfrak{F}_{i-1,n}) = 0.$$

Furthermore assume that the following two conditions are satisfied:

1. 1. *The ‘norming condition’ is satisfied*:

$$\sum_{i \leq nz} E(X_{i,n}^2 / \mathfrak{F}_{i-1,n}) \to z$$

   uniformly in probability as $n \to \infty$.
2. 2. *The ‘conditional Lindeberg condition’ is fulfilled: for all* $\varepsilon > 0$

$$\sum_{i \le n} E(X_{i,n}^2 I[|X_{i,n}| > \epsilon] / \mathcal{F}_{i-1,n}) \to 0$$

in probability as $n \to \infty$.
Then let us introduce the random elements S_n, *defined by*

$$S_n(z) = \sum_{i \le nz} X_{i,n}$$

for $0 \le z \le 1$. *Then the* S_n *converge in distribution to a standard random walk* W.

Another important class of processes are stationary processes X_n, $n \in \boldsymbol{Z}$. In general, we will not even have a CLT. If, however, the conditional expectation of X_n given the $X_0, X_{-1}, \ldots$ decreases sufficiently fast, we will have a functional limit theorem, analogous to Gordin's theorem. Let us define the normalized, partial sums S_n by

$$S_n(z) = \frac{1}{\sqrt{n}} \sum_{i \le nz} X_i.$$

Furthermore we will use the $\boldsymbol{L}_2$-norm of random variables: Define $\|X\| = \sqrt{EX^2}$.
Theorem 4: (Peligrad and Utev, 2005): *Assume that we have given a stationary process* X_i *so that*

$$\sum \frac{1}{n} \|E(S_n(1) / X_0, X_{-1}, \ldots)\| < \infty,$$

for all i

$$\frac{1}{n} \sum X_t X_{t-1} \to E(X_t X_{t-i})$$

and

$$\sigma^2 = \sum_{i \in \mathbf{Z}} E(X_t X_{t-i}) < \infty .$$

Then

$$\frac{1}{\sigma} S_n$$

converges in distribution to a standard random walk W.
These two theorems should only act as illustrations for functional limit theorems. Especially for stationary processes, more general theorems are available. A good survey about recent results can be found in Merlevede, Peligrad and Utev (2006).

Conclusion

This short introduction article should serve only as an introduction to functional limit theorems. Over the years, a rich theory has developed unifying many aspects of the limiting behaviour of functions of random variables. In particular, I would like to mention the limiting theorems for empirical distribution functions and their generalizations (see for example van der Vaart and Wellner, 1996, for a survey, and Andrews and Pollard, 1994, for dependent random variables). These results can be used to derive ‘uniform’ central limit theorems.

See Also

- central limit theorems

Bibliography

Andrews, D.W.K. and Pollard, D. 1994. An introduction to functional central limit theorems for dependent stochastic processes. *Revue internationale de statistique* 62, 119–32.

Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd edn. New York: Wiley-Interscience.

Davidson, J. 1994. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.

McLeish, D.L. 1974. Dependent central limit theorems and invariance principles. *Annals of Probability* 2, 620–8.

Merlevede, F., Peligrad, M. and Utev, S. 2006. Recent advances in invariance principles for stationary sequences. *Probability Surveys* 3, 1–36.

Peligrad, M. and Utev, S. 2005. A new maximal inequality and invariance principle for stationary sequences. *Annals of Probability* 33, 789–815.

van der Vaart, A. and Wellner, J.A. 1996. *Weak Convergence and Empirical Processes*. Berlin: Springer.

How to cite this article

Ploberger, Werner. "functional central limit theorems." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000311> doi:10.1057/9780230226203.0609

The New Palgrave Dictionary of Economics Online

functional finance

David Colander
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘functional finance’ was created by Abba Lerner to contrast with sound finance. It involves making decisions about the deficit and the money supply with regard to their functionality, not some abstract moralistic premise. While it seems to play no role in the dynamic stochastic general equilibrium model prevalent in macroeconomics today, it does play a potential role in a more complex model where heterogeneous agents with limited information interact in a model with many different aggregate equilibria. Yet Lerner's functional finance theoretical model is far too simple to be acceptable, even as a rough guide for policy.

Keywords

budget deficits; coordination problems; functional finance; general equilibrium; Great Depression; incomes policy; inflation; Keynesianism; Lerner, A.; macroeconomic externalities; money supply; multiple equilibria in macroeconomics; national debt; optimal taxation; sound finance; stagflation

Article

In the debate about how to pull economies out of the Great Depression, Abba Lerner created a steering wheel metaphor to contrast his ‘economics of control’ approach to policy with the then prevailing ‘laissez-faire’ policy. He argued that the laissez-faire approach was similar to driving a car without a steering wheel, the natural result of which was that the economy continually crashed, veering off the road first in one direction and then in another. It was time, he argued, for the government to adopt a Keynesian ‘economics of control’ approach in which the government used an explicit steering wheel – functional finance – to keep the economy running smoothly.
To complement that distinction between economics of control and laissez-faire, he contrasted the laissez-faire policy of sound finance with the economics-of-control policy of functional finance. Sound finance involved a set of rules – always balance the budget except in wartime, and do not increase the money

supply at a rate greater than the growth rate of the economy. The problem, for Lerner (1944; 1951), was that these rules of sound finance were not analysed; they were simply accepted as being right. Lerner argued that, when governments understood how the macroeconomy actually operated, they would adopt an alternative ‘functional finance’ set of rules. Under the rules of functional finance, decisions about the deficit and the money supply would be made with regard to their functionality – their effect on the economy – and not with regard to some abstract moralistic premise that deficits, debt and expansionary monetary policy are inherently bad.

The rules of functional finance

Functional finance consists of the following three rules (Lerner, 1941).

1. 1. The government shall maintain a reasonable level of demand at all times. If there is too little spending and, thus, excessive unemployment, the government shall reduce taxes or increase its own spending. If there is too much spending, the government shall prevent inflation by reducing its own expenditures or by increasing taxes.
2. 2. By borrowing money when it wishes to raise the rate of interest, and by lending money or repaying debt when it wishes to lower the rate of interest, the government shall maintain that rate of interest that induces the optimum amount of investment.
3. 3. If either of the first two rules conflicts with the principles of ‘sound finance’, balancing the budget, or limiting the national debt, so much the worse for these principles. The government press shall print any money that may be needed to carry out rules 1 and 2.

In proposing these rules of functional finance, Lerner's purpose was to shift thinking about government finance from principles of sound finance that make sense for individuals – such as running a balanced budget – to functional finance principles that make sense for the aggregate economy. Functional finance principles used the budget balance as a steering wheel: deficits increased economic activity, surpluses decreased economic activity. The budget balance had these effects because, in the Keynesian model, government spending and taxing decisions directly affected levels of economic activity. These effects had to be considered because, in the aggregate, the secondary effects of spending decisions and savings decisions, which Lerner and I (Colander, 1979) called macro externalities, had to be taken into account, whereas in individual decisions they did not.

Lerner's stark presentation of these rules of functional finance caused much stir in the 1940s and 1950s, when most Keynesians, including Keynes himself, were politically more circumspect about what came to be known as Keynesian ideas for government fiscal policy than they became in the 1960s (Colander and Landreth, 1996). Lerner's rules specifically ruled out worrying about the size of a country's budget deficit or national debt.

In the 1950s and 1960s, Lerner's functional finance rules became both the basis of most textbook presentations of Keynesian economics and the basis of textbook macroeconomic policy discussions. It became what was generally considered Keynesian policy. This could occur because Keynes's *General Theory* contained almost no discussion of policy; it did not mention fiscal policy, and yet there were strong political forces pushing for its use. Thus, when ‘Keynesian policy’ was attacked in the late 1960s

and early 1970s, it was primarily the idea of Lerner's policy of functional finance that most people were attacking (see Colander, 1984, for a discussion).

That attack on ‘Keynesian policy’ intensified through the 1970s and 1980s, and by the 1990s textbook presentations of Keynesian policies had faded away. As they did so, so too did the concept of functional finance, and by the early 2000s few economists under the age of 50 had heard of it.

While the term ‘functional finance’ has disappeared from the macroeconomic textbooks, its influence continues among macro policy economists. The rhetoric of policy-oriented macro economists and their reaction to recessions is now quite different from what it was in pre-Keynesian times. When presenting fiscal policy to voters, governments are far less likely to talk about balanced budgets. Today, the potential benefits of government deficits in a recession are recognized. Similarly, policy-oriented macroeconomists discuss fiscal policy generally in terms of debt-carrying capacity such as represented by deficits as a percentage of GDP, not the need for a balanced budget, as was the case with sound finance. Even when a policy of functional finance is not used, the functional-finance role of fiscal policy is still seen as important since the expectation that government functional-finance policy will be adopted when crises occur can reassure agents and provide stability to the economy.

Why functional finance lost favour

Functional finance lost favour for a variety of reasons. First, Lerner's discussion of functional finance did not consider the politics of government finance; it assumed that the government could change taxes and spending according to the needs of the macroeconomy. In reality, both spending and taxing are difficult political issues, and the needs of politics generally trump the needs of stabilization. Second, the lags between recognition of a problem and implementation of a policy were significant, and the policy would often go into effect long after the situation had changed. In Lerner's automobile metaphor, it was as if the steering wheel and the wheels were connected with a 30-second lag, and the windshield was opaque. Third, functional finance is built upon an assumption that the government knows what functional finance policy is best to follow – in inflationary times, increase taxes and decrease the money supply; in recessionary times, decrease taxes and increase the money supply. In the 1970s, when both inflation and recession occurred simultaneously, the functional finance rules seemed to give contradictory advice. These practical problems with implementing functional finance eliminated much, if not all, of the benefit of the steering wheel.

The reaction of Keynesian economists to the practical and informational problems was to limit the use of the deficit as a tool for fine-tuning the economy; the fiscal policy tool was a sledge, not a ball-peen hammer. The economics profession's reaction to stagflation was to accept a high rate of unemployment as the trigger for implementing an expansionary policy. Lerner did not follow the profession. His reaction to the stagflation problem was to argue that much inflation was not the result of excess demand but was instead what he called sellers’ inflation. Sellers’ inflation operated quite apart from demand pressures. Depending on how sellers’ inflation was dealt with, there could be either high full employment or low full employment (Lerner, 1972).

Lerner saw sellers’ inflation as so important that, beginning in the 1960s, he changed his research programme to centre on finding cures for sellers’ inflation. He developed a market-based incomes policy in which property rights in prices are established, and individuals have to buy the right to change prices from others who change their price in the opposite direction (Lerner and Colander, 1980). Under a

market-based incomes policy, rights in value-added prices would be tradable, so that any firm wanting to change its nominal price would have to make a trade with another firm that wanted to change its nominal price in the opposite direction. Thus, by law, the average price level would be constant, but relative prices would be free to change. With inflation controlled by such a plan, the rules of functional finance would once more become relevant (Colander, 1979). Politically, in the early 2000s such policies had little chance of even being considered by governments and had faded from economists' radar screen.

Macro theory and functional finance

It was not only the practical problems of functional finance that led to its demise. It was also that the profession essentially dropped the theoretical model upon which the concept was based. Functional finance was based on a coordination-failure model of macroeconomics – when individuals spent or saved, they did not take into account the effect of that decision on the aggregate level of spending; thus the economy needed some mechanism to internalize the spending complementarity and thereby determine the aggregate level of spending.

Today, among theoretical macroeconomists macro policy is thought of in a dynamic stochastic general equilibrium framework, and fiscal policy is discussed within an optimal taxation framework that assumes a representative agent is optimizing over a long-term horizon. The intuition behind such models is that the effect of any government deficit is mitigated by compensatory changes in the representative agent's spending decisions. This occurs because the agent will be responsible for paying off that deficit in the future. In the now prevalent modern macroeconomic theoretical approach, the possible existence of macro externalities is essentially ruled since the representative individual is assumed to take all the indirect effects of spending into account.

Assessment of functional finance

So what should one make of functional finance? My view is that, theoretically, it remains important. The fact that much modern macroeconomic theory does not allow for the possible existence of macro externalities is, in my view, a problem of modern macro theory, not a problem with functional finance. The probability that the unique equilibrium, perfect rationality, perfect foresight, representative agent model underlying much of modern macroeconomics has much relevance to the real-world macro problems that we face is exceedingly small.

The macroeconomic theory problem seems more appropriately described as a coordination problem in which heterogeneous agents with limited information interact in a model in which many different aggregate equilibria are possible due to enormous strategic complementarities among agents. With multiple equilibria and coordination problems, there is no presumption of global optimality of the equilibrium chosen by the market. Everyone can know of the existence of a preferable equilibrium, but may not be able to achieve it by private actions. We can say something about that question only when we have a theory of equilibrium selection mechanisms. Currently we have none. Thus, in a multiple equilibrium economy with coordination failures, there should be no general presumption that the private economy, given its institutions, arrives at an equilibrium preferable to one achieved with government guidance.

That said, the functional finance theoretical model of Lerner is far too simple to be acceptable, even as a rough guide for policy. To say that individuals have limited information and do not fully take account of future effects of policy is not to say that they take no account of them. Private institutions develop which do precisely that, and any meaningful theoretical macro model must integrate such forward-looking private institutions into its structure. Doing so will involve highly complex models in which model selection by agents, agent interdependency, and social interaction by multiple agents are taken seriously. We are a long way from making such models tractable, so any formal macro model incorporating usable rules of functional finance is long in the future.

See Also

- budget deficits
- cost-push inflation
- Keynesianism
- multiple equilibria in macroeconomics

Bibliography

Colander, D. 1979. Rationality, expectations and functional finance. In *Essays in Post Keynesian Inflation*, ed. J. Gapinski. Cambridge, MA: Ballinger.

Colander, D. 1984. Was Keynes a Lernerian? *Journal of Economic Literature* 22, 1572–5.

Colander, D. and Landreth, H. 1996. *The Coming of Keynes to America*. Cheltenham: Edward Elgar.

Lerner, A. 1941. The economic steering wheel. *University Review*, June, 2–8.

Lerner, A. 1944. *The Economics of Control*. New York: Macmillan.

Lerner, A. 1951. *The Economics of Employment*. New York: McGraw Hill.

Lerner, A. 1972. *Flation*. New York: Penguin Books.

Lerner, A. and Colander, D. 1980. *MAP: A Market Anti-Inflation Plan*. New York: Harcourt Brace Jovanovich.

How to cite this article

Colander, David. "functional finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000294> doi:10.1057/9780230226203.0610

The New Palgrave Dictionary of Economics Online

Furtado, Celso (1920 – 2004)

Mauro Boianovsky
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Celso Furtado (1920–2004) was one of the most influential Latin American economists of the 20th century. He was head of the development division of the United Nations Commission for Latin America in the 1950s, where he helped to formulate the structuralist approach to economics. His *Formação Economica do Brasil* (1959) is the classic interpretation of the economic history of Brazil. In 1961 he published a collection of essays about the notion of underdevelopment and development as interdependent phenomena. Furtado's last contribution was his careful discussion in the 1970s of the concept of cultural and economic dependence in underdeveloped countries.

Keywords

balance of payments constraint; balanced growth; Baran, P.; big push; Brazil; centre–periphery system; dependency theory; Furtado, C.; Gerschenkron, A.; import substitution; industrialization; inflation; Kaldor, N.; Latin American development; Lewis, W. A.; Nurkse, R.; Presbisch, R.; Robinson, J.; Rosenstein-Rodan, P.; Rostow, W.; structuralism; surplus; underdevelopment

Article

Celso Furtado was born on 26 July 1920 in Pombal (in the state of Paraiba, north-east of Brazil), and died on 20 November 2004 in Rio de Janeiro. Together with the Argentinean Raúl Prebisch, Furtado was the most widely read and influential Latin American economist of the second half of the 20th century. A prolific writer, he published more than 20 books on the economic history of Brazil and Latin America, and on the theory and policy of economic development, many of them translated into English, French and other languages.
He graduated at Universidade do Brasil (Rio) in 1944 and received his doctorate from the Sorbonne (Paris) in 1948; his thesis was about the Brazilian colonial economy. Maurice Byé was his supervisor, but it was François Perroux who impressed him most at the time. Upon his return to Brazil in that same

year, Furtado was invited to join the staff of the new United Nations Economic Commission for Latin America (ECLA) in Santiago. From 1950 to 1957 he was head of the development division of ECLA. He then left Santiago to spend an academic year at Cambridge University working with Nicholas Kaldor and Joan Robinson with a Rockefeller Foundation scholarship. In 1958 he was appointed director of the Brazilian National Development Bank, where he conceived the project that led to the creation of SUDENE (Development Agency of the Northeast of Brazil) in 1959, of which Furtado was the first director (Hirschman 1963, chapter 1). In 1962 he also became Brazil's first minister of planning, charged with drafting a national economic plan, a position he held until 1963. Deprived of his political rights following the military coup in 1964, he left Brazil to take up appointments at American and European universities. Furtado went back to Paris and became the first foreign professor to be appointed at the Sorbonne, where he taught development economics from 1965 to 1985. After Brazil returned to democracy he was appointed Minister of Culture (1986–88), and elected to the Brazilian Academy of Letters and to the Brazilian Academy of Sciences in 1997 and 2003 respectively. (Furtado's autobiography, originally published in three volumes between 1985 and 1991, was collected in 1997; the first volume, with recollections from the 1950s, his most productive period, was translated into French in 1987.)

Structuralism and economic history

Together with Prebisch and other economists at ECLA in the 1950s, Furtado was one of the formulators of structuralism in Latin American economics. His main contributions can be found in two books, both available in English. In his 1961 volume on economic development, which collected essays written during the 1950s, Furtado provided the most elaborate exposition of the structuralist analysis in the literature at the time. In his 1959 classic *Formação Economica do Brasil*, written in Cambridge in 1957–8 and based on Furtado (1950; 1952; 1954), the structuralist approach was applied for the first time to the interpretation of the economic history of a Latin American country, an exercise Furtado would expand to the whole region in his 1969 book. Furtado's methodological innovation was the use of historical investigation to identify factors that are specific to each structure through time: ‘bring history near to economic analysis, get from the latter precise questions and find answers in history’ (1997, vol. 1, p. 205). In *Formação* he pioneered the use of modern income analysis to deal with historical phenomena by introducing macroeconomic models into the analysis of each phase of Brazilian economic development from the 16th century to the 1950s (see also Furtado, 1963, for a brief account). Furtado's role in the historiography of the industrialization process of Brazil in particular and Latin American in general may be compared to Alexander Gerschenkron's well-known interpretation (1952) of the late industrialization of Russia and other continental countries. Like Gerschenkron, Furtado examined industrialization from the point of view of history. Both rejected Walt Rostow's (1960) view that the economic development of different countries goes through a succession of phases to which a single analytical framework can be applied.

The main feature of the 1959 book is the argument that the economic history of Brazil (and other Latin American countries as well) must be based on an open growth model with international trade treated as an endogenous variable, since these countries’ economies evolved as suppliers of raw materials to the world market. Furthermore, the income-distribution profile is a main determinant of the economic

growth process through its effect on the level and structure of domestic demand in different historical phases. Furtado shows that throughout the four centuries from 1530 to 1930 the Brazilian economy depended on external demand to provide stimulus to higher productivity without previous capital accumulation, with three long-period cycles – sugar exports (1530–1650), gold mining (1700–80) and the expansion of the world market for coffee (1840–1930) – and intervening periods of relative stagnation. That phase came to an end in the economic crisis of 1929, when the collapse of export-commodity prices cut the country's import-purchasing power in half. According to Furtado, the policy adopted by the Brazilian government at the time to maintain the coffee price – that is, buying the unmarketable coffee and burning it – had the effect of an unwitting ‘Keynesian’ anti-cyclical deficit-financing policy. This contributed to maintaining domestic demand and, together with the diminished capacity to import, pushed up domestic prices of imported goods and stimulated investments in import-substituting industrial consumer goods. That process marked the beginning of a new phase in the development of Brazil, based on internal demand and import-substituting industrialization. Brazilian late industrialization – as compared with that of the United States – is explained in part by the differences between the productive structure of Brazil's export agriculture and the small agricultural properties in the English colonies of North America. The Brazilian internal market was much thinner due to the concentration of income and property, which served to maintain its stagnant colonial structure. Moreover, whereas the United States participated in the first wave of the Industrial Revolution as exporter of a key raw material (cotton), the main cause of the relative backwardness of the Brazilian economy in the first half of the 19th century, according to Furtado, was the damming up of its exports and the increase of the subsistence sector with lower productivity. Also in contrast with the late industrialization of continental European countries in the second half of the 19th century studied by Gerschenkron, the import-substitution process in Latin America did not lead to an intensive development of producer goods industries or changes in international trade (exports of manufactured goods and imports of raw materials). The evolution of trade patterns in Latin American countries during their industrialization after 1930 was quite the opposite: exports were still based on a few commodities and imports concentrated on goods whose production required huge investment and/or advanced technology.

The concept of economic underdevelopment

It was in attempting to explain the backwardness of Brazil that Furtado hit upon the idea that underdevelopment and development are two interdependent phenomena which appear simultaneously as part of the evolution of industrial capitalism. The theme was elaborated in his 1961 book, where Furtado put forward concepts of economic underdevelopment and development that have been largely accepted in the literature. An underdeveloped structure is one in which ‘full utilization of available capital is not a sufficient condition to complete absorption of the working force at a level of productivity corresponding to the technology prevailing in the dynamic sector of the economy’ (1961b, p. 141). Underdeveloped economies (as distinct from simply backward ones) are hybrid structures characterized by technological heterogeneity of the various sectors. This comes from the historical fact that the import-substituting industrialization process in those economies led entrepreneurs to adopt a technology compatible with a cost and price structure similar to that prevailing abroad. Technology becomes, therefore, an independent variable in economies where industrialization is induced from outside. Whereas

industrialization in underdeveloped economies was determined by demand, the formation process of capitalist European economies in the 18th and 19th centuries was dominated by supply factors, which led Furtado to define economic development as the introduction of new combinations of production factors which increase labour productivity. Underdevelopment is regarded as a permanent feature of the centre–periphery system, not as a stage on the road to development. Those ideas originally appeared in an essay written as the first critical comment on Ragnar Nurkse's notion of ‘balanced growth’ (advanced in Nurkse's 1950 Rio lectures), where Furtado pointed out that the dynamics of demand (internal and external) in underdeveloped economies should be studied in tandem with the process of accumulation. According to Furtado, underdeveloped countries lack incentives to save (because of the consumer habits of higher-income groups), not to invest. The accumulation process should be examined from the point of view of changes in the process of generation, utilization and appropriation of the economic surplus, especially as affected by foreign trade. Furtado first developed these ideas in an essay originally written in Portuguese in 1955 (two years before Paul Baran made the concept of surplus a central notion of his own approach to development) and further elaborated it as part of a comment on Paul Rosenstein-Rodan's theory of ‘big push’, made at the International Economic Association conference on economic development held in Rio in 1957, and in his 1967 and 1980 textbooks.

Foreign trade and dependency

One of the main aspects of the industrialization process of Latin American countries, as discussed by Furtado in 1958 and 1960, is the persistent tendency towards balance of payment crises and inflationary pressures. Anticipating some elements of the two-gap model later developed by Chenery and Bruno (1962), Furtado showed in a two-sector model featuring a modern and a backward sector how balance of payment disequilibrium could constraint the economic growth process under the assumption that the coefficient of imports in the investment sector is larger than in the consumption sector, as is typically the case in underdeveloped countries. Such chronic disequilibrium has structural (not monetary) causes and may lead to the ‘strangulation’ of economic growth. Another obstacle to growth is that, after the end of the ‘easy’ phase of the substitution of imported consumer goods, as industrialization advances to the production of intermediate and capital goods the rate of profit falls because of the higher capital output ratio accompanied by increasing income concentration and lower aggregate demand. This was an essential element of Furtado's (1965; 1970) interpretation of the slowdown of economic growth in Latin America in the early 1960s, but, as the Brazilian economy recovered in the late 1960s and early 1970s, Furtado's stagnationist argument was criticized by economists in Brazil (see Tavares and Serra 1973). Furtado (1972; 1974; 1978) eventually concluded that, after the two earlier periods of economic growth – determined respectively by comparative advantages and import-substitution – the Latin American economy had entered a new dynamic path in which consumption demand by high-income groups could under certain conditions become the leading factor of the system. This led him to explore in detail a theme that had often come up in his writings in the 1950s: dependency theory.

Furtado argued that underdeveloped economies feature cultural dependence, that is, consumption patterns are historically transplanted from developed countries by the upper strata of the underdeveloped areas as a result of their appropriation of the economic surplus generated through comparative advantages in foreign trade. Such modernized component of consumption brings dependence into the technological sphere by making it part of the production structure. Dependent structures are also

dualistic systems with unlimited supply of labour at a subsistence wage, as first described by Furtado (1950) in his investigation of the dynamics of the labour market in Brazilian economic history. This is close to W. Arthur Lewis's (1954) classic model, but, in contrast with Lewis, Furtado's conclusion is that industrialization within a dualist dependent structure reproduces this dualism and does not bring about a homogeneous system with real wages increasing in tandem with the average productivity of the economy. The relationship between the centre and the periphery in the world economy is defined not just by the unequal sharing of the benefits of development and technical progress (as in Prebisch's terms-of-trade argument) but by dependence involving domination and control of access to modern technology by transnational corporations. In Furtado's view, economic growth does not entail economic development in dependent and reflex economies, since it implies an aggravation of both external and internal exploitation, and thereby tends to make underdevelopment even more acute.

See Also

- dependency
- Gerschenkron, Alexander
- Prebisch, Raúl
- structural change

Selected works

1950. Características gerais da economia brasileira. *Revista Brasileira de Economia* 4, 7–37.

1952. Capital formation and economic development (trans. J. Cairncross). In *The Economics of Underdevelopment*, ed. A. Agarwala and S. Singh. Oxford: Oxford University Press, 1958.

1954. *A Economia Brasileira*. Rio: A Noite.

1955. O desenvolvimento econômico – ensaio de interpretação histórico-analitica. *Econômica Brasileira* 1, 3–24.

1958. The external disequilibrium in the underdeveloped countries. *Indian Journal of Economics* 38, 403–10.

1959. *The Economic Growth of Brazil – A Survey from Colonial to Modern Times*, trans. R. Aguiar and E. Drysdale. Berkeley: University of California Press, 1963.

1960. Industrialization and inflation (trans. P. Schwartz and E. Henderson). *International Economic Papers* 12(1967), 101–119.

1961a. Comments on Professor Rosenstein-Rodan's paper. In *Economic Development for Latin America*,

ed. H. Ellis and H. Wallich. London: Macmillan.

1961b. *Development and Underdevelopment – A Structural View of the Problems of Developed and Underdeveloped Countries*, trans. R. Aguiar and E. Drysdale. Berkeley: University of California Press, 1964.

1963. The development of Brazil. *Scientific American* 209, 208–20.

1965. Development and stagnation in Latin America: a structuralist approach. *Studies in Comparative International Development* 1, 159–75.

1967. *Théorie du développement économique*, trans. A. Silva and J. Peffau. Paris: Presses Universitaires de France, 1970.

1969. *Economic Development of Latin America – A Survey from Colonial Times to the Cuban Revolution*, trans. S. Macedo. Cambridge: Cambridge University Press, 1970.

1970. *Obstacles to Development in Latin America*, trans. C. Ekker. New York: Anchor Books.

1971. Dependencia Externa y Teoria Economica. *El Trimestre Economico* 38, 335–49.

1972. *Analyse du 'modèle' brésilien*, trans. E. Treves. Paris: Anthropos, 1974.

1974. *Le mythe du développement économique*, trans. E. Treves. Paris: Anthropos, 1976.

1978. *Accumulation and Development*, trans. S. Macedo. Oxford: Martin Robertson, 1983.

1980. *Brève introduction au développement*, trans. A. Ahmed. Paris: Publisud, 1989.

1985. *La fantaisie organisée*, trans. E. Bailby. Paris: Publisud, 1987.

1987. Underdevelopment: to conform or reform. In *Pioneers in Development – Second Series*, ed. G. Meier. New York: Oxford University Press for the World Bank.

1997. *Obra Autobiografica de Celso Furtado*, 3 vols. São Paulo: Paz e Terra.

2000. Celso Furtado (born 1920). In *A Biographical Dictionary of Dissenting Economists*, 2nd edn, ed. P. Arestis and M. Sawyer. Cheltenham: Edward Elgar.

Bibliography

Arndt, H. 1987. *Economic Development – The History of an Idea*. Chicago: University of Chicago Press.

Baer, W. 1969. Furtado on development: a review essay. *Journal of Developing Areas* 3, 270–80.

Baran, P. 1957. *The Political Economy of Growth*. New York: Monthly Review.

Bielschowsky, R. 1988. *Pensamento Economico Brasileiro – O ciclo ideologico do desenvolvimentismo*. Rio: IPEA.

Cardoso, E. 1981. The Great Depression and commodity-exporting LDCs: the case of Brazil. *Journal of Political Economy* 89, 1239–50.

Chenery, H. and Bruno, M. 1962. Development alternatives in an open economy: the case of Israel. *Economic Journal* 72, 79–103.

Dobb, M. 1965. Review of C. Furtado, *Development and Underdevelopment – A Structural View of the Problems of Developed and Underdeveloped Countries* (1964). *Economica* 32, 460–1.

Furtado, C. 1952. Capital formation. In *International Economic Papers*. 1954 no. 4: 124–44.

Gerschenkron, A. 1952. Economic backwardness in historical perspective. In *The Progress of Underdeveloped Areas*, ed. B. Hoselitz. Chicago: University of Chicago Press.

Hirschman, A. 1963. *Journeys Toward Progress*. New York: The Twentieth Century Fund.

Hunt, D. 1989. *Economic Theories of Development: An Analysis of Competing Paradigms*. New York: Harvester Wheatsheaf.

Kay, C. 1989. *Latin American Theories of Development and Underdevelopment*. London: Routledge.

Lewis, W. 1954. Economic development with unlimited supplies of labour. *Manchester School* 22, 139–91.

Lowe, J. 1996. *Crafting the Third World: Theorizing Underdevelopment in Rumania and Brazil*. Stanford: Stanford University Press.

Mallorquin, C. 2005. *Celso Furtado – Um Retrato Intelectual*. Rio: Contraponto.

Mueller, H. 1963. Review of C. Furtado, *The Economic Growth of Brazil – A Survey from Colonial to Modern Times* (1963). *Journal of Economic History* 23, 359–60.

Nurkse, R. 1953. *Problems of Capital Formation in Underdeveloped Countries*. Oxford: Blackwell.

Prebisch, R. 1949. The economic development of Latin America and its principal problems. *Economic Bulletin for Latin America* 7 (1962), 1–22.

Rostow, W. 1960. *The Stages of Economic Growth: A Non-Communist Manifesto*. Cambridge: Cambridge University Press.

Szmrecsányi, T. 2005. The contributions of Celso Furtado (1920–2004) to development economics. *European Journal of the History of Economic Thought* 12, 691–702.

Tavares, M. and Serra, J. 1973. Beyond stagnation: a discussion on the nature of recent development in Brazil. In *Latin America: From Dependence to Revolution*, ed. J. Petras. New York: John Wiley & Sons.

How to cite this article

Boianovsky, Mauro. "Furtado, Celso (1920–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000293> doi:10.1057/9780230226203.0611

The New Palgrave Dictionary of Economics Online

futures markets, hedging and speculation

David M. Newbery

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Futures markets provide partial income risk insurance to producers whose output is risky, but very effective insurance to commodity stockholders at remarkably low cost. Speculators absorb some of the risk but hedging appears to drive most commodity markets. The equilibrium futures price can be either below or above the (rationally) expected future price (backwardation or contango). The various effects futures markets can have on market and income stability are discussed. Rollover hedges can extend insurance from short-horizon contracts over longer periods.

Keywords

arbitrage; backwardation; capital asset pricing model; cobweb models; commodity stabilization scheme; contango; electricity markets; expectations; forward contracts; futures markets; futures markets, hedging and speculation; hedging; income stability; information aggregation; information sharing; liquidity; portfolio choice; price discovery; price stability; rational expectations; risk aversion; risk premium; risk sharing; rollover hedges; speculation; subjective probability; vertical integration; Working, H.

Article

Futures markets for grain emerged in Chicago in the middle of the 19th century and spread rapidly to other commodities and centres. Forward contracts, in which two agents agree on the details of a transaction for delivery at a specified future date, must date back to the beginnings of commerce itself, but the distinctive feature of a futures market is that the contracts are standardized, transactions costs are minimized, and liquidity is high, so that contracts can be, and typically are, bought and sold many times during their lifetime, in contrast to most forward contracts. The standard explanation for the role of futures markets is that they help to spread and hence reduce risks, and to motivate the collection and dissemination of relevant information. Forward markets provide the same risk-sharing opportunities, but the greater transparency and liquidity of futures markets makes the latter far more potent institutions for

‘price discovery’.
The question of how well futures markets (and securities markets more generally) perform this role of collecting, aggregating and disseminating information is a large and important topic, best handled under the wider heading of ‘information’. If we assume agents have rational expectations and share common information, then the price-discovery role of futures markets can be ignored and remaining issues of risk-sharing studied in isolation. In this case there is little conceptual difference between futures and forward markets, and we can concentrate attention on the two characteristic modes of behaviour exhibited by these markets – speculation and hedging.

Speculation and hedging in commodity markets

Speculation is the purchase (or temporary sale) of goods for later resale (repurchase), rather than use, in the hope of profiting from the intervening price changes. In principle, any durable good could be the subject of speculative purchase, but, if carrying costs are high, or the good is illiquid, then the margin between the buying and selling price will be large, and speculation in that good will be normally be unattractive. Liquidity in this context means that there exists a perfect, or near-perfect, market in which the good can be sold immediately for a well-defined price, and this requirement severely limits the range of assets available for large-scale speculation. There are two types of assets – commodities traded on organized futures markets, and financial assets (bonds, shares) whose properties lend themselves particularly to speculation. Hedging, on the other hand, typically refers to a transaction on a futures markets undertaken to reduce the risks arising from some other risky activity, whether producing the commodity, storing it, or processing it for final sale.
Thus a risk-averse wheat farmer may hedge his future harvest by selling October wheat futures in January, in which case he is ‘long’ in actuals and ‘short’ in futures. A risk-averse miller who anticipates being short of wheat may hedge by buying futures now, in which case he will be a ‘long’ hedger. Speculators may be on the long or short end of any transaction, but in aggregate their position must offset any net imbalance in the long and short hedgers’ positions.
It might appear from this that hedging consists in shifting the price risk onto the speculators in return for a risk premium. This view of speculation, advanced by Keynes (1923) and Hicks (1946), has been challenged by Working (1953; 1962), who denies any fundamental difference between the motivations of hedgers and those of speculators. One danger with looking exclusively at the price risk is that it ignores the more fundamental quantity risks that give rise to the price risks. Once this is appreciated, it is possible to formulate a simple theoretical model in which all agents are alike in attempting to maximize their expected utility but differ in the risks to which they are exposed, and these differences motivate trade on futures markets. While the activities of speculators are quite well defined, those of ‘hedgers’ are in general a mixture of insurance and speculation, as we shall see.
The simplest model of speculation and hedging has just two time periods. In the first period farmers plant their wheat, and the futures market opens. In the second period the wheat is harvested, sold, and the futures contracts expire. There are only three types of agents – farmers, who produce wheat but do not consume it; speculators, who neither produce nor consume wheat; and consumers, who neither produce wheat nor trade on futures markets. All agents are assumed to have beliefs about the relevant variables, which can be described by (subjective) probability distributions, and their behaviour is described by the theory of expected utility maximization. There are *n* farmers, and for the moment

suppose that they have no choice over the amount of wheat to plant, but only over the size of their sales on the futures market. In the first period farmer *i* believes that his second period output will be $\tilde{q}_i$ (a random variable), and that the market clearing price will be $\tilde{p}^i$, also a random variable. In particular, he believes that $\tilde{q}_i$ and $\tilde{p}^i$ are jointly normally distributed. The price of futures is *f*, observable now, and he sells z_i futures, so that he believes his second period income will be

$$\tilde{y}_i = \tilde{p}^i \tilde{q}_i + z_i(f - \tilde{p}^i), \tag{1}$$

a random variable. The farmer's utility function exhibits constant absolute risk aversion, A_i, and takes the form $U^i(y) = -k_i \exp(-A_i \tilde{y})$, where $\tilde{y}$ is the random component of his income. (Any non-random components can be absorbed into the constant, k_i.) This particular form has the property that maximizing expected utility is equivalent to maximizing

$$W = Ey - \frac{1}{2} A \operatorname{Var} y, \tag{2}$$

where *Ey* is the expected value of income, Var *y* is its variance, provided, as in the case here, that *y* is normally distributed. (These are the standard assumptions of the capital asset pricing model for portfolio choice, and can be viewed as second-order approximations to more general utility functions; see Newbery and Stiglitz, 1981.) If eq. (1) is substituted in (2), and if z_i can be positive (futures sales) or negative (purchases), then the value of z_i that maximizes *W* is

$$z_i = \frac{\operatorname{Cov}(\tilde{p}^i, \tilde{p}^i \tilde{q}_i)}{\operatorname{Var} \tilde{p}^i} - \frac{E\tilde{p}^i - f}{A_i \operatorname{Var} \tilde{p}^i}. \tag{3}$$

Speculator *j* has no risky production, so for him $\tilde{q}_j$ is zero, and the first terms in (1) and (3) vanish. Thus the second term in (3) can be identified as the speculative term, and is readily interpreted. The perceived

riskiness of the futures contract is measured by Var $\tilde{p}^i$, and the cost of this risk as $1/2A_i\mathrm{Var}\,\tilde{p}^i$. The expected return to selling a futures contract is $f - Ep^i$. In order to persuade a risk-averse speculator to *buy* futures and accept the risk, the return to *selling* must be negative, hence *f* must be below the expected spot price, $E\tilde{p}^i$ – a situation of *normal backwardation*. The first term in (3) is the pure hedging term, for if the futures market appears *unbiased* (that is, $f = E\tilde{p}^i$) then there is no expected speculative profit, and the only motive for trade is the income insurance offered by the price insurance. The quality of income insurance depends on how well income *pq* and price risks are correlated; that is, on the ratio of the covariance to the variance. If output is perfectly certain, then income and price are perfectly correlated, the first term will be equal to q_i, and the farmer would sell his entire crop on the futures market if he believed it to be unbiased. In general, though, he will not believe it to be unbiased, and he will wish to speculate in addition to hedging. His net futures trade will reflect the balance of the desire to insure and the returns to speculating.

The futures market clears, so that the sum of z_i across all participants must be zero, and this condition will yield a value for the futures price. What this implies for the value of *f* and its relation for the subsequent spot price, *p*, depends on beliefs, as well as preferences. If agents hold *rational expectations*, and have full information about the nature of all production and demand risks, then they will agree on the common values of the expected spot price, *Ep*, and its variance, Var•*p*. In such a case the only motive for trading on the futures market is to share risk, and speculators will be willing to absorb some of the risk in return, on average, for some profit. If all farmers face perfectly correlated production risk, and if the coefficient of variation of output is σ_q, of price is σ_p, and the correlation coefficient between price and output is *r*, then market clearing on the futures market gives the bias as

$$\frac{Ep - f}{Ep} = \frac{\bar{Q} \cdot Ep\sigma_p^2(1 + r\sigma_q / \sigma_p)}{\Sigma 1 / A_i} \tag{4}$$

and a farmer's futures sales will be

$$\frac{z_i}{Eq_i} = \beta_i(1 + r\sigma_q / \sigma_p), \beta_i \equiv 1 - \frac{\bar{Q}}{Eq_i A_i \Sigma_j 1 / A_j}, \tag{5}$$

where $\bar{Q} = \Sigma Eq_i$ is average total output (see Newbery and Stiglitz, 1981, p. 186). Thus β_i is a measure of the extent to which the farmer is more risk-averse than the average (the term in A_i) and more exposed to

risk ($\bar{q}_i / \bar{Q}$). If there are *n* identical farmers and *m* identical speculators, all with the same coefficient of absolute risk aversion, *A*, then $\beta = m/(n + m)$. If there is no output risk, so $\sigma_q = 0$, then, while a farmer would sell his entire crop forward on an unbiased futures market, here he would only sell a fraction β representing the fraction of the total risk which the speculators are willing to bear. If the only source of risk is supply variability, then $r = -1$, $\sigma_q / \sigma_p = \varepsilon$, the elasticity of demand, and the farmer will sell a fraction of his crop $\beta(1 - \varepsilon)$ on the futures market, possibly negative.

What lesson can be drawn from this very simplified model? First, futures markets allow speculators to bear some of the farmer's risks. The more highly correlated income and price risks, the better the market is at insuring farmers, but in general it will provide only partial insurance. It is, however, much better suited to providing insurance to stockholders who store the commodity after the harvest until needed for consumption or processing, and it is not surprising that most hedging is done by stockholders rather than farmers. Second, the greater the agreement over the expected spot price, and the less risk-averse are the speculators, the smaller will be the average perceived bias, and the larger will be the fraction of hedging to speculative sales by producers (or stockholders). Third, the greater the degree of agreement on the expected spot price, the more will speculation be a response to the demand for hedging services. The greater the disagreement on the expected spot price, the more likely it is that speculation, in the form of gambling over the expected spot price, will dominate the market. In a masterly series of studies, Holbrook Working showed that most commodity futures markets depend primarily on hedging for their existence, that the size of the open interest follows closely the demand for hedging of seasonal storage, with speculators standing ready to assume the risks offered by the hedgers (Working, 1962). The cost of these hedging services (that is, the return to the speculators) was quite remarkably small. Thus for cotton traders, the *gross* profit per dollar of sales over a sample of some 3,000 trades was 0.023 of one per cent with the traders making losses on 15 out of 43 trading days. (Net profits after paying commissions and expenses were substantially less; Working, 1953).

The issue of bias turns out to be more complex than the simple Keynes–Hicks risk-premium view, for even in a bilateral market of farmers and speculators the bias can go either way. Once stockholders and processors are brought into the picture, the relative demands for long and short hedges will change yet again, and in turn influence the direction of speculation (long or short) and hence of the risk premium, or bias. Hirshleifer (1988) examines the determinants of bias in a market with primary producers subject to output risk (growers) and intermediate producers (processors). He finds that processors tend to hedge long, but, if transaction costs are low, there is a downward bias in futures prices (backwardation). If transaction costs are high, growers are differentially driven from the futures market, and could reverse the bias to contango.

Effect of speculators on stability

Several important questions can be asked about the role of speculators. Do they tend to destabilize the spot market and/or the futures market? Do they improve efficiency? Do they have adverse macroeconomic effects? To the layman the association of speculative activity with volatile markets is often taken as proof that speculators are the cause of the instability, though the body of informed opinion is that the volatility creates a demand for hedging or insurance, which is met by the willingness of speculators to bear the risk. It is hard to test the proposition that speculation is stabilizing, for speculative

activity (notably, stockholding) can take place without futures markets. In practice, the usual question is: do futures markets, which, by lowering transaction costs, greatly facilitate speculative behaviour, improve the stability of the spot market? Even this question is not straightforward. Futures markets provide an incentive to collect information about the future market-clearing spot price, though, as often with information gathering, there are public-good problems associated with its use. Much theoretical effort has been devoted to the question of whether futures prices perfectly reveal the relevant information available to participants, and, if so, what incentives would remain for its collection. It now appears that, except in special cases, the information is only partially revealed in the market, leaving incentives for its collection, but nevertheless improving the forecasts of otherwise uninformed traders. If so, and if the spot market is intrinsically volatile (because of variations in supply caused by weather, or demand caused by the trade cycle), then better forecasts of future spot prices will tend to elicit compensating supply responses – if prices are expected to be high tomorrow, then it will pay to produce more, and to carry more stocks forward, tending to reduce, or stabilize, price fluctuations. To the extent that futures markets reduce storage risks, storage becomes cheaper, and this will tend to stabilize supplies and prices directly. On the other hand, anticipated disturbances will have a more immediate effect on current prices, and will tend to make them more responsive to news. A frost in Brazil expected to affect next year's coffee production is likely to have a more rapid effect on current coffee prices in the presence of a futures market than in its absence. Nevertheless, it improves the efficiency of the current market if it does respond to this relevant information.

The clearest example of the stabilizing effect of futures market is provided by cobweb models, in which producers base current production decisions on last year's realized price, with consequent self-sustaining fluctuations in output without any exogenous shocks. If a futures market is set up, then producers initially planning to expand production in response to last year's high price, and selling futures, would cause the futures price to fall to the predicted spot price, and would lead them to revise their incorrect production plans, hence eliminating the cobweb and stabilizing the market.

Two other factors bear on the question of market stability. It is clear that much hinges on the nature of expectations. Speculation without hedging is a zero-sum game, and, if two speculators, each holding different views of the future price, $E\tilde{p}$, trade with each other, one will gain while the other will lose. If they are rational, and risk-averse, they should not be willing to engage in such swaps. On this view, speculators who are more successful at forecasting the future price will make money, and those who are less successful will lose, and be forced to leave the market, until only the good forecasters are left, and they make money only in the course of moving futures prices towards the forecast spot price. However, it is possible that a steady supply of less good speculators, who add noise to the system, lose money and exit, to be replaced by others. Their presence may worsen the predictive power of the futures price or, by increasing the returns to information gathering by the informed speculators, may actually improve the predictive power of the futures prices (Anderson, 1984a; Kyle, 1984). Depending on the direction of the net effect of uninformed speculators, the presence of a futures market (which provides them with the opportunity to gamble) may improve or worsen the efficiency of the spot market.

The other possibility is that futures markets will provide opportunities for market manipulation, by the better informed at the expense either of the less well informed (corners, squeezes) or of the larger at the expense of the smaller. It is easy to show that the futures price has an effect on production decisions by extending the model of eq. (1) to allow producers to choose inputs. In the case of pure demand risk (no output uncertainty) it can be shown that the producer will base his production decisions solely on the

future price. Large producers (Brazil for coffee, OPEC for oil, and so on) may then find it profitable to intervene in the futures market to influence the production decisions of their competitors in the spot market, and in extreme cases may find it profitable to increase price instability, though the extent to which this is feasible will be limited by the supply of and risk tolerance of other speculators in the futures market (Newbery, 1984). This is true even if all agents hold rational expectations, and share full information (except about the actions of the large producers). If some agents use naive forecasting rules to guide their futures trading, and if these rules are known to other agents who possess market power, then it may pay the large rational agents to destabilize the price and exploit the irrationalities in the forecasting behaviour of the naive agents (Hart, 1977).

Although speculation may stabilize prices, it is quite possible for it to make prices more unstable, even if all agents have equal information and hold rational expectations. Compare two possible arrangements. In the first, futures markets are prohibited, the commodity is perishable, so there is no scope for speculative storage or speculation on the futures market. The commodity can be produced by two methods, one perfectly safe, the other risky, but on average more profitable (for example, two varieties of irrigated rice, one higher-yielding but susceptible to rust in certain weather conditions). Farmers allocate their land between the two production techniques but, in the absence of the futures market, find the risky technique relatively unattractive and so produce little. In the second arrangement, futures markets are permitted and speculators are willing to trade for a very low risk premium. Farmers are now able to sell the crop forward, and are therefore more willing to produce the risky crop, whose supply is very variable. Total supply variability increases, and hence the spot price becomes more variable.

It is quite possible that destabilizing speculation of this type yields higher potential social welfare, for yields are higher, if riskier, and the risks are borne at relatively low cost. It is also perfectly possible for speculation on a futures market to be stabilizing (by reducing the costs of storage and therefore improving arbitrage between crop years) and yet make everyone worse off (see, for example, Newbery and Stiglitz, 1981). We now know that, if the market structure is incomplete, creating additional markets can make matters worse. Speculation, which creates a market in price risks, does not thereby complete the market structure because quantity risks may remain imperfectly insured. The reason is that the market in price risks causes changes in the market equilibrium which affects the degree to which the other risks (income and quantity risks) are effectively insured. In particular, if prices are stabilized but quantities remain unstable, incomes may be less stable than if prices were free to move in response to the quantity changes.

Finally, there remains the old Keynesian question of whether speculation which succeeds in stabilizing prices will exacerbate income fluctuations. The argument, due to Kaldor (1939), is straightforward. Speculators undertake or assume the risks for storage, which then responds to mismatches in supply and demand. These stocks, or inventories of goods, will fluctuate markedly and will have the same macroeconomic effect as fluctuations in investment, tending, through the multiplier, to have a magnified effect on national income. Whether these speculative stock movements are stabilizing or destabilizing then turns on whether they offset or amplify the fluctuations in income associated with the mismatch in demand and supply that caused the stock change. Kaldor's view was that stock changes caused by supply shocks would tend to stabilize total income, while those caused by demand stocks would be destabilizing, but much will depend on the commodity price elasticities of demand and the nature of the various transmission mechanisms, particularly the lag structure. Nevertheless, the OPEC oil shocks have demonstrated that commodity supply shocks can cause significant macroeconomic disturbances, while

the increasing ease of currency speculation as restrictions are removed and transaction costs lowered has reawakened the fear that speculation may, in some cases, destabilize income and impose needless costs.

Commodity stabilization schemes and longer-term insurance

At various times governments and international agencies have argued that primary commodity price variability is costly to vulnerable, often poor, primary exporters, and that therefore some form of commodity stabilization scheme should be implemented. Such schemes are often poorly designed to minimize the cost of reducing risk and have a doubtful record (Newbery and Stiglitz, 1981). One might also expect that, in the presence of the kind of market failure suggested by this costly risk, alternative institutions might emerge to reduce risk, and that is indeed the case, with futures markets being the most obvious solution to commodity price risk. If primary exporting countries can hedge the export commodity price variability, then their risk will be reduced, and would seem to be eliminated if all the risk arose from price variability, with no variability in output. This would be true if there were no serial correlation in prices from year to year, but, as Deaton and Laroque (1992) found, there is considerable serial correlation for the 24 commodities they studied over the period 1900–87. Their results suggest that about one-quarter of price shocks are permanent, that three-quarters or more of the price shock will persist for at least a year, and even after two years typically 60 per cent of the price shock will persist. If countries (or producers) hedge only for the coming year, their income will still vary considerably from year to year. If they could hedge for many years ahead this problem would be reduced.

Most futures markets extend only a relatively short period ahead and, even when they extend out several years, active trading and hence liquidity is mostly confined to the near-term future, measured in months rather than years. Apart from primary exporters having to deal with serial correlation (or persistence in price shocks), producers making large, irreversible sunk investment decisions (for example, in an oil refinery, offshore oil exploitation, LNG liquefaction and regasification facilities, aluminium smelters, nuclear power stations) would make better investment decisions knowing future prices (of inputs and outputs). They would be able to borrow more cheaply if risk were reduced by contracts or hedging, reducing the cost of capital-intensive products.

Liquid futures extending out ten years would clearly help, but are lacking. In their absence, companies may prefer to vertically integrate down the supply chain to provide an implicit (if partial) hedge. Electricity and gas liberalization has been premised on separating out natural monopoly pipes and wires from potentially competitive services supplied over the networks, regulating the former and creating wholesale and retail markets for the latter. Vertical unbundling (particularly of generation and transmission) appears critical to delivering the efficiency benefits of competition (Newbery, 2005), but increases risk as wholesale electricity and fuel markets are so volatile. Forward and futures markets for electricity (and fuels such as gas) exist but *basis* risk (the difference between the price of the product traded and that of interest to the contractor) is high and markets are very illiquid. Vertical integration between generation and supply (or retailing) reduces spot price risk but makes the market less contestable.

Nevertheless, it is possible to use a sequence of short-term futures markets to hedge longer-term risks through a sequence of rollover hedges. Kletzer, Newbery and Wright (1992) show how to compute an *n*-year rollover hedge for a commodity with serially correlated price risk, no output risk but supply

responsive to expected price. The way the rollover works is to sell more futures initially than needed for one-period hedging, and then use the surplus futures sales to finance the next year's futures transactions. This is not perfect, for the amount of hedging required next year will depend on production, and that will depend on the futures price prevailing next year, not as yet known. Consequently, despite the absence of production risk, future output cannot be perfectly hedged, and there remains some residual risk (as there would be if there were output risk). Nevertheless, because the costs of risk increase with the square of the deviation, reducing the risk by a given fraction reduces the cost of risk by more than that fraction and can be worthwhile. The further forward the hedge extends, the lower is the extra risk benefit provided, until the extra costs outweigh the benefit, so there is an optimal length of such a hedge.

The idea of using rollover hedging and portfolios of futures of different maturities to reduce risk has proved powerful both in theory and in hedging practice. Ross (1997) considers a world in which commodity prices are determined by many factors, and that, given enough different futures contracts and sufficiently precise knowledge of the underlying model determining prices, it would be possible to devise a perfect hedge, although in practice any such hedge would be imperfect. Neuberger (1999) develops this approach to identify an optimal hedging strategy using futures of different maturities and thus hedge long-term exposures with a combination of short-term futures. Neuberger tests his model on crude oil futures traded on NYMEX from 1986 to 1994. He asks how well one can hedge a forward commitment to deliver oil in five years' time using two futures contracts of not more than nine months to maturity. The annualized volatility of the five-year contract is 26 per cent and that of the hedged portfolio is less than one per cent, with a hedge of short 2.89 seven-month contracts (of 1,000 barrels) and long 3.93 nine-month contracts, for each contract to deliver in five years' time. In a model in which a trader wishes to hedge for delivery in 36 months' time, if the portfolio is balanced monthly, 488 contracts are traded per contract delivered, although this can be cut to fewer than 60 with bimonthly rebalancing (and at lower risk).

The fact that rollover hedges allow one to reduce risk over a longer time horizon than the duration of current futures offered in the market has a number of interesting implications. It can explain why near-term futures are more popular and liquid than longer-term contracts, for they may provide a substitute for the latter at lower cost. It also explains why the volume of liquid futures can so greatly exceed the underlying physical trade, often by factors of 10 to 20. Rollovers require both a greater ratio of futures to physicals and a higher rate of trading to rebalance the portfolio over time, contributing to volume, liquidity and hence cost reduction.

Rollovers are, however, not perfect, and they may tempt traders to take imprudently large risks. One such famous case was the near-bankruptcy of Metallgesellschaft (MG), whose losses were estimated at DM 4 billion and whose survival was ensured only by a major rescue operation (Wahrenburg, 1996). At one time MG was reportedly holding short-term positions equivalent to 160 million barrels of oil or 80 times the daily output of Kuwait (Hilliard, 1999).

The case became celebrated as a test of whether MG had adopted a sound or imperfect hedging strategy. Some writers such as Culp and Miller (1995) argued that MG was following 'a textbook hedging strategy which was not properly understood by MG's supervisory board and house banks' (Wahrenburg, 1996, S29). Others, such as Edwards and Canter (1995), Mello and Parsons (1995), and Verleger (1999) argue that MG was excessively exposed in the wrong products. Wahrenburg argues that the MG's hedging strategy could indeed significantly reduce risk, but not completely, and that MG's equity capital

was insufficient to cover the remaining risk.

See Also

- arbitrage
- hedging
- information aggregation and prices
- options
- options (new perspectives)
- present value

Bibliography

Anderson, R. 1984a. The industrial organization of futures markets: a survey. Ch. 1 of Anderson (1984b).

Anderson, R. (ed.) 1984b. *The Industrial Organization of Futures Markets*. Lexington, MA: Lexington Books.

Culp, C. and Miller, M. 1995. Metallgesellschaft and the economics of synthetic storage. *Journal of Applied Corporate Finance* 7, 6–21.

Deaton, A. and Laroque, G. 1992. On the behaviour of commodity prices. *Review of Economic Studies* 59, 1–24.

Edwards, F. and Canter, M. 1995. The collapse of Metallgesellschaft: unhedgeable risks, poor hedging strategy or just bad luck? *Journal of Futures Markets* 15, 211–64.

Hart, O. 1977. On the profitability of speculation. *Quarterly Journal of Economics* 91, 579–97.

Hilliard, J. 1999. Analytics underlying the Metallgesellschaft hedge: short term futures in a multi-period environment. *Review of Quantitative Finance and Accounting* 12, 195–219.

Hirshleifer, D. 1988. Risk, futures pricing, and the organization of production in commodity markets. *Journal of Political Economy* 96, 1206–20.

Hicks, J. 1946. *Value and Capital*, 2nd edn. Oxford: Oxford University Press.

Kaldor, N. 1939. Speculation and economic stability. *Review of Economic Studies* 7(1), 1–27. Reprinted in N. Kaldor, *Essays on Economic Stability and Growth*. London: Duckworth, 1960.

Keynes, J.M. 1923. Some aspects of commodity markets. *Manchester Guardian Commercial, Reconstruction Supplement* 29, March.

Kletzer, K., Newbery, D. and Wright, B. 1992. Smoothing primary exporters' price risks: bonds, futures, options and insurance. *Oxford Economic Papers* 44, 641–71.

Kyle, A. 1984. A theory of futures market manipulation. Ch. 5 of Anderson (1984b).

Mello, A. and Parsons, J. 1995. Maturity structure of a hedge matters: lessons from the Metallgesellschaft debacle. *Journal of Applied Corporate Finance* 8, 86–105.

Neuberger, A. 1999. Hedging long-term exposures with multiple short-term futures contracts. *Review of Financial Studies* 12, 429–59.

Newbery, D. 1984. The manipulation of futures markets by a dominant producer. Ch. 2 of Anderson (1984b).

Newbery, D. 2005. Electricity liberalization in Britain: the quest for a satisfactory wholesale market design. *Energy Journal* 26, 43–70. Special Issue on European Electricity Liberalisation, ed. D. Newbery.

Newbery, D. and Stiglitz, J. 1981. *The Theory of Commodity Price Stabilization*. Oxford: Clarendon Press.

Ross, S. 1997. Hedging long-run commitments: exercises in incomplete market pricing. *Economic Notes* 26, 385–419.

Verleger, P. 1999. Was Metallgesellschaft's use of petroleum futures part of a rational corporate strategy? *Journal of Energy Finance and Development* 4, 89–115.

Wahrenburg, M. 1996. Hedging oil price risk: lessons from Metallgesellschaft. *Chicago Board of Trade Research Symposium Proceedings* Part 2, S29–S47.

Working, H. 1953. Futures trading and hedging. *American Economic Review* 43, 314–43.

Working, H. 1962. New concepts concerning futures markets and prices. *American Economic Review* 52, 432–59.

How to cite this article

Newbery, David M. "futures markets, hedging and speculation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan,

2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000247> doi:10.1057/9780230226203.0612

The New Palgrave Dictionary of Economics Online

Galbraith, John Kenneth (1908 – 2006)

Lester C. Thurow
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

corporatism; countervailing power; Galbraith, J. K.; invisible hand; planning; price control; technical change.

Article

John Kenneth Galbraith was a paradox. Born in Canada in 1908, he began his professional career armed with a Ph.D. in agricultural economics from the University of California. During the Second World War he was in charge of price controls and immediately after was director of the Strategic Bombing Survey. Later he was in charge of economic affairs in the occupied countries and was awarded the Medal of Freedom for his efforts. He became a Professor of Economics at Harvard, a President of the American Economic Association, and an advisor to presidents and presidential candidates, the latter leading to his appointment as ambassador to India during the Kennedy Administration.

Yet throughout this distinguished career the economics profession moved steadily towards more formal mathematizable models and exhibited less and less interest in old-fashioned political economy, while Galbraith himself never moved an iota in either direction. In the spirit that one might expect from a former editor of *Fortune* magazine, his books were written always in the form of verbally persuasive economic tracts, without a hint of mathematics. His interests were always those of political economy, with political considerations ranking at least as high as, and most often higher than, those of economics. Perhaps because of his writings on the causes and consequences of the Great Depression in *The Great Crash* (1961) and his successful experience as a price controller during the Second World War, he was never a believer in the wisdom of the invisible hand. If there is an essential theme in his economic writings, it is that the government has a role to play in successful economies.

The Affluent Society (1955) documents the tendency of the invisible hand to promote private splendour and public squalor. Others have made that case (before and since; analytically and verbally), but no one has ever grabbed the public's attention with vivid examples as he did. There were other forces also at work, but much of the effort to improve the quality of the public sector during the 1960s can be traced to his writings. Of course, Galbraith would have wanted the government to go much farther and returned to

this theme in *The Culture of Contentment* (1992) and *The Good Society: A Humane Agenda* (1996) which describe a range of interventions to address contemporary problems.

Planning, however, was not just something for the public sector. Planning was essential to the smooth functioning of the private sector. As a result large firms had an important role to play in the private economy. They were not just actual or potential anti-trust threats. In many ways *American Capitalism* (1952b) and its doctrines of countervailing power have come to be the accepted wisdom. Big is no longer automatically bad. Major new government antitrust cases have almost disappeared. That said, *The Economics of Innocent Fraud* (1994a) emphasized Galbraith's concerns late in life about the power of corporate managers to shape society.

In the Galbraith view in *The New Industrial State* (1967a), large firms are essential since they finance much of the research and development that leads to the technical innovations that are necessary to secure a rising standard of living. Technical change had traditionally stood outside of economics as an exogenous force, although with the advent of the new growth economics this is no longer the case. Galbraith placed it where it should be at the centre of his analysis and it led to very different conclusions regarding the role of the large firm. Today it is fashionable to point to the many formerly small firms that have become technological leaders, but Galbraith would reply that most of these firms can be shown to have sprung from the laboratories of some large firm or university.

The invisible hand systematically leads to too few resources for the public sector, too few resources for research and development, and poor coordination between firms, but it also, in Galbraith's view, leads to too few resources for the poor. In *Economic Development* (1962), *The Nature of Mass Poverty* (1979a) and *The Voice of the Poor* (1983a) he has systematically argued for public actions to redress the imbalances produced by the market in the distribution of income. He was never a believer in the virtues of ‘trickle down’. And as the percentage of total income going to the bottom 40 per cent of the population fell in the mid-1980s under the impact of America's current experiment with benign neglect, he could claim vindication for his earlier arguments, as he could have in the last decade of his life.

The result was an economist out of the mainstream of economic thought, but in the mainstream of economic events.

Selected works

1952a. *A Theory of Price Control*. Cambridge, MA: Harvard University Press.

1952b. *American Capitalism*. Boston: Houghton Mifflin.

1955. *The Affluent Society*. Boston: Houghton Mifflin.

1961. *The Great Crash*. Boston: Houghton Mifflin.

1962. *Economic Development*. Cambridge, MA: Harvard University Press.

1967a. *The New Industrial State*. Boston: Houghton Mifflin.

1967b. *How to Get Out of Viet Nam*. New York: New American Library.

1969a. *Ambassador's Journal*. Boston: Houghton Mifflin.

1969b. *How to Control the Military*. New York: Doubleday.

1973a. *A China Passage*. Boston: Houghton Mifflin.

1973b. *Economics and the Public Purpose*. Boston: Houghton Mifflin.

1975. *Money, Whence it Came, Where it Went*. Boston: Houghton Mifflin.

1977. *The Age of Uncertainty*. Boston: Houghton Mifflin.

1979a. *The Nature of Mass Poverty*. Cambridge, MA: Harvard University Press.

1979b. *Annals of an Abiding Liberal*. Boston: Houghton Mifflin.

1981. *Life in our Times*. Boston: Houghton Mifflin.

1983a. *The Voice of the Poor*. Cambridge, MA: Harvard University Press.

1983b. *The Anatomy of Poverty*. Boston: Houghton Mifflin.

1987. *Economics in Perspective: A Critical History*. Boston: Houghton Mifflin.

1992. *The Culture of Contentment*. Boston: Houghton Mifflin.

1993. *A Short History of Financial Euphoria*. Boston: Houghton Mifflin.

1994a. *The Economics of Innocent Fraud*, Boston: Houghton Mifflin.

1994b. *A Journey through Economic Time*. Boston: Houghton Mifflin.

1996. *The Good Society: A Humane Agenda*. Boston: Houghton Mifflin.

2005. A comprehensive biography by R. Parker, *John Kenneth Galbraith: His Life, His Politics, His Economics*, New York: Farrar, Straus and Giroux.

How to cite this article

Thurow, Lester C. "Galbraith, John Kenneth (1908–2006)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000004> doi:10.1057/9780230226203.0613

The New Palgrave Dictionary of Economics Online

Gale, David (1921 – 2008)

Joel Sobel
From The New Palgrave Dictionary of Economics Online, 2008,
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the research of David Gale, who made lasting contributions to game theory, general equilibrium theory, and growth theory. In addition to his influence on the development of economic theory, his work has had important implications for many branches of mathematics and on mathematical education.

Keywords

assignment problem; competitive equilibrium; convexity; Debreu, G.; existence of competitive equilibrium; Gale, David; Gale–Shapley algorithm; game theory; global univalence; Kuhn, H.; linear inequalities; matching; Mathsite; Nikaido, H.; Ramsey, F.; Shapley, L.; Tucker, A.W.; von Neumann, John; zero-sum game

Article

David Gale was born in New York on 13 December 1921, and died in Berkeley, California on 7 March 2008. He received an undergraduate degree from Swarthmore and a Master's degree from the University of Michigan before earning a Ph.D. in Mathematics at Princeton. It was at Michigan, under the influence of Professor Norman Steenrod, that Gale decided to give up his study of physics and pursue a Ph.D. in mathematics. He taught at Brown University from 1950 through 1965 and then joined the faculty at the University of California, Berkeley. His principal appointment was in the Mathematics Department, but he maintained affiliations with the departments of Economics and Industrial Engineering.
Gale won wide recognition for his research. His awards included a Fulbright research fellowship, two Guggenheim fellowships, elections to the American Academy of Arts and Science and the National Academy of Science, the Lester Ford Prize (for outstanding mathematical exposition), the John von Neumann Theory Prize (for fundamental contributions to operations research), and the Pirelli Internetional Award (for the Internet Mathematics Museum ‘MathSite’).

Mukul Majumdar (1992) edited the volume *Equilibrium and Dynamics: Essays in Honour of David Gale*. The *International Journal of Game Theory*, Volume 36, Numbers 3–4, March 2008 contains a collection of papers dedicated to David Gale on the occasion of his 85th birthday. This volume was edited by Marilda Sotomayor, who had also organized a scientific day in David's honour during the 18th Summer Festival on Game Theory in Stony Brook, 12/13 July 2007. Special issues of *Games and Economic Behavior* and *The Mathematical Intelligencer* are in preparation.
Gale lived in Berkeley, California and Paris, France with his partner Sandra Gilbert, a renowned feminist literary scholar and poet. Her 2000 book of poetry *Kissing the Bread* included a section of poems she wrote for Gale called ‘When she was kissed by the mathematician’. He had three daughters and two grandsons. Julie Gale, his former wife and the mother of his daughters, died in February 2008.

Linear inequalities

As a graduate student at Princeton, David Gale worked with classmate Harold Kuhn on a research project supervised by Professor Albert Tucker. At the time, there was considerable excitement about the new fields of zero-sum game theory and linear programming, but the mathematics of linear inequalities had not been developed. Existing proofs of the minimax theorem of zero-sum game theory required fixed-point arguments and did not make the relationship between the theory and linear inequalities explicit. The project led to important results that identified the deep connections between the two new areas. Gale, Kuhn and Tucker (1951) contained the first complete proof of the duality theorem of linear programming, and used the theorem to prove the minimax theorem of zero-sum, two-person game theory. This paper uses convex analysis rather than fixed-point arguments to prove the minimax theorem, and implicitly provided a computational foundation for equilibrium points in zero-sum games.
Gale's book *The Theory of Linear Economic Models* (1960) contains central results on the theory of linear inequalities, including Gale, Kuhn and Tucker (1951) and Gale's extension of von Neumann's model of an expanding economy (1956a). It discusses Dantzig's simplex algorithm and gives an economic interpretation to canonical problems. The book also contains a concise and elementary introduction to the theory of linear inequalities (including a proof of the separating hyperplane theorem for convex polytopes), a chapter containing essential results on non-negative matrices, and a clean treatment of dynamic linear models of growth.
Largely in recognition of their joint work, Gale, Kuhn and Tucker won the 1980 von Neumann prize for work they began in the late 1940s. Their citation stated that they ‘played a seminal role in laying the foundations of game theory, linear and non-linear programming – work that continues to be of fundamental importance to modern operations research and management science’.

Infinite games

Early work on non-cooperative game theory concentrated on two-player, zero-sum games. When players had finitely many pure strategies, these games were well understood. All two-player finite zero-sum games have a value, and by playing to maximize their minimum payoff, a player could guarantee this value independent of his or her opponent's strategy. Gale and Stewart (1953) studied a class of infinite zero-sum games and demonstrated that the minmax theorem need not hold in this more general setting.

The paper examines the simplest possible infinite zero-sum game. In the two-player game of perfect information, the players take turns naming binary digits, which can be thought of as the binary expansion of a number between zero and one. The first player wins if the expansion is an element of a pre-specified set. Otherwise, the second player wins. Gale and Stewart show that basic results from finite games hold if the prespecified set is closed, but that in general the game does not have a value. The class of infinite games introduced by Gale and Stewart has had broad implications for mathematics. Gale and Stewart's result led to research that identified a general set of games that do have values, culminating in a theorem of Martin (1975). The fact that some games do not have values led to developments in set theory (Mycielski, 1964).

Growth

Gale (1956a and 1960,chapter 8, section 5) generalizes and simplifies the von Neumann (1937/46) model of an expanding economy in what in now called the von Neumann–Gale model of growth. Gale made two essential additions to the original model. He substituted von Neumann's requirement that each production process involves each good in the economy (as either input or output), with a weaker, more plausible condition.

Gale (1967) provides the definitive treatment of the multi-good Ramsey problem, which is a generalization of the linear von Neumann–Gale model. An agent starts with a given endowment, which must be allocated between immediate consumption and investment. What the agent invests is transformed, via a given technology, into the next period's endowment, which again may be allocated between immediate consumption and investment. The process continues indefinitely. The agent cares about the (undiscounted) sum of utility received from consumption. The problem is to find the appropriate definition for optimality and to characterize optimal consumption paths. Gale presents an appropriate optimality condition, provides conditions under which optimal paths exist, and characterizes these paths in terms of ‘turnpike’ properties. Roughly speaking, we can construct an optimal program with two phases: a bounded initial transition phase in which the state is built up to (approximate) a sustainable optimal steady state, followed by a program that approximates the best steady state consumption.

General equilibrium

Gale made several important contributions to the foundations of general equilibrium theory. Indeed, he made basic contributions to the three central issues of the theory: existence, uniqueness and stability. Gale (1955) contains a result known as the Gale–Debreu–Nikaido Lemma (Debreu, 1956; Nikaido, 1956) which contains the essential mathematical result needed to prove the existence of market equilibrium. Gale and Nikaido (1965) proves a theorem on the global univalence of differentiable mappings on $\mathbb{R}^n$. When translated into a general equilibrium context, the theorem gives sufficient conditions for equilibrium prices to be unique (see, for example, Arrow and Hahn, 1971, chapter 9).

Gale (1963) provides an early robust example of global instability of the tâtonnement process in general equilibrium.

Gale and Mas-Colell (1975) provides an existence theorem in an economy without ordered preferences.

College admissions and the stability of marriage

Gale's paper with Lloyd Shapley on the stable marriage problem (1962) is his most cited, and probably most influential, work. Detailed overviews of the research appear in Knuth (1976/1997), Gale (2001), Roth (2008) and Roth and Sotomayor (1990).

The short, deceptively simple paper is important for several reasons. The motivation for the problem comes from the real world. Gale (2001) describes how the idea for the problem came from thinking about the college application process. The translation of the practical problem into mathematics captures many important considerations but remains extraordinarily simple. The solution to the problem is not obvious, but is easy to understand. The framework lends itself to modifications that lead to insight into more complicated practical problems.

The basic problem is how to create an assignment of items from one group to items from another. The groups can be men and woman (the marriage problem), workers to jobs (labour-market matching), or students to universities (college admissions). For concreteness, consider the marriage problem, in which it is natural to impose the constraint that there are equal numbers of men and women and the desired matching is one to one. Assume everyone has preferences over potential partners (so each man can order the women from the most preferred to least preferred marriage partner, and likewise each woman can order the men). Finding a match is simple. One can order them by age and match the youngest man to the youngest woman, the second youngest man to the second youngest woman, and so on. Gale and Shapley looked for a matching in which there is no unmatched man and woman who prefer each other to their current partners. If this property failed, you would expect the matching to be unstable. Gale and Shapley show that stable matchings exist, and present a simple algorithm that constructs stable matches. Starting in 1951, 11 years before the publication of the Gale–Shapley paper, the National Intern Matching Program used an essentially equivalent algorithm to match graduating medical students to hospital residency programmes (see Roth and Sotomayor, 1990, pp. 169–70, and Roth, 2008, appendix for a discussion of the independent development of the matching algorithm). Practical problems in a wide variety of areas (from school assignment to kidney exchange) continue to stimulate the development of matching theory.

Assignment markets and auctions

Gale's work shows sensitivity to computational issues. Knowing the connection between zero-sum, two-person game theory and the theory of linear programming combined with computational methods (like the simplex method) provides a tractable method for computing equilibria in two-player games. Demonstrating the equivalence between an equilibrium and the solution to a well-behaved optimization problem is the reason that equilibria in linear economies studied by Eisenberg and Gale (1959) can be found efficiently. Gale's work on markets with indivisible goods is another example of a situation in which Gale adds just enough structure to a general model to obtain strong results.

Shapley and Shubik (1971) introduce the assignment model, a market equilibrium model with indivisible goods. Their model has the structure of a matching game with the added feature that agents can exchange a divisible commodity, money. Demange and Gale (1985) show that this market inherits

many properties from the college admissions problem. Demange, Gale, and Sotomayor (1986) apply the framework to the study of multi-unit auctions and show how to define variations of the Vickrey auction to the multi-good setting.

Other contributions

Gale (1956b) contains a lasting contribution to the study of convex polyhedra, introducing what are now known as ‘Gale transforms’ and ‘Gale diagrams’ (see Grünbaum, 2003).
Gale (2009) describes the board games invented by Gale and his contemporaries at Princeton. Gale's article provides an introduction to Bridg-It (or, as Martin Gardner called it, the ‘Game of Gale’) and also John Nash's game of Hex. Gale (1974) invented the game of Chomp, a simple two-player game of perfect information in which it is easy to show that one player has a winning strategy, but the winning strategy is hard to find in general.

Mathematical explorations

Gale made examples of beautiful mathematical arguments accessible to a broad audience. Between 1991 and 1996 he wrote a column entitled ‘Mathematical explorations’ for *The Mathematical Intelligencer*. The columns, collected in a book titled *Tracking the Automatic Ant* (Gale, 1998), are in the tradition of Martin Gardner's long-running ‘Mathematical games’ column in *Scientific American*. He also developed MathSite, a pedagogic website that uses interactive exhibits to illustrate important mathematical ideas. MathSite won the 2007 Pirelli Internetional Award for Science Communication in Mathematics.

Acknowledgements

I thank Harold Kuhn and Bernhard von Stengel for helpful comments. Sobel (2009) is a more detailed overview of David Gale's research contributions.

See Also

- auctions (applications)
- convex programming
- existence of general equilibrium
- game theory
- linear programming
- matching
- Ramsey model

Selected works

1951. (With Harold W. Kuhn, and Albert W. Tucker.) Linear programming and the theory of games. In *Activity Analysis of Production and Allocation*, ed. Tjallings C. Koopmans, ch. 19, pp. 287–97. New York: Wiley.

1953. (With Frank M. Stewart.) Infinite games with perfect information. In *Contributions to the Theory of Games*, ed. Harold W. Kuhn and Albert W. Tucker, vol. 28 of Annals of Mathematics Studies, pp. 245–66. Princeton: Princeton University Press.

1955. The law of supply and demand. *Mathematica Scandinavica* 3.

1956a. A closed linear model of production. In *Linear Inequalities and Related Systems*, ed. Harold W. Kuhn and Albert W. Tucker, vol. 38 of Annals of Mathematics Studies, ch. 18, pp. 285–303. Princeton: Princeton University Press.

1956b. Neighboring vertices on a convex polyhedron. In *Linear Inequalities and Related Systems*, ed. Harold W. Kuhn and Albert W. Tucker, vol. 38 of Annals of Mathematics Studies, ch. 15, pp. 255–63. Princeton: Princeton University Press.

1959. (With Edmund Eisenberg.) Consensus of subjective probabilities: the pari-mutuel method. *Annals of Mathematical Statistics* 30, 165–8.

1960. *The Theory of Linear Economic Models*. New York: McGraw-Hill.

1962. (With Lloyd S. Shapley.) College admissions and the stability of marriage. *American Mathematical Monthly* 69, 9–15.

1963. A note on global instability of competitive equilibrium. *Naval Research Logistics Quarterly* 10, 81–7.

1965. (With Hukune Nikaido.) The Jacobian matrix and global univalence of mappings. *Mathematische Annalen* 2, 81–93.

1967. On optimal development in a multi-sector economy. *Review of Economic Studies* 34, 1–18.

1974. A curious Nim-type game. *American Mathematical Monthly* 81, 876–9.

1975. (With Andreu Mas-Colell.) An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2, 9–15.

1985. (With Gabrielle Demange.) The strategy structure of two-sided matching markets. *Econometrica* 53, 873–88.

1986. (With Gabrielle Demange and Marilda Sotomayor.) Multi-item auctions. *Journal of Political Economy* 94, 863–72.

1998. *Tracking the Automatic Ant*. New York: Springer.

2001. The two-sided matching problem. origin, development and current issues. *International Game Theory Review* 3, 237–52, June–September.

2009. Topological games at Princeton, a mathematical memoir. *Games and Economic Behavior*.

Bibliography

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Debreu, G. 1956. Market equilibrium. *Proceedings of the National Academy of Sciences* 42, 876–8.

Grünbaum, B. 2003. *Convex Polytopes*, 2nd edn. New York: Springer-Verlag.

Knuth, D.E. 1976. *Mariages Stables et leurs Relations avec d'Autres Problèmes Combinatoires: Introduction à l'Analyse Mathématique des Algorithmes*. Montréal: Presses de l'Université de Montréal. (Trans. as *Stable Marriage and its Relation to Other Combinatorial Problems: An Introduction to the Mathamatical Analysis of Algorithms*, vol.10 of Crm Proceedings and Lecture Notes. American Mathematical Society, 1997.)

Majumdar, M. (ed.) 1992. *Equilibrium and Dynamics: Essays in Honour of David Gale*. London: Macmillan.

Martin, D.A. 1975. Borel determinacy. *Annals of Mathematics*, 2nd series, 102, 363–71, September.

Mycielski, J. 1964. On the axiom of determinateness. *Fundamenta Mathematicae* 53, 205–24.

Neumann, J. von. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktzatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, Wien. [Trans. as A model of general economic equilibrium. *Review of Economic Studies* 13, 1–9, 1945–6.]

Nikaido, H. 1956. On the classical multilateral exchange problem. *Metroeconomica* 8, 135–45.

Roth, A.E. 2008. Deferred acceptance algorithms: history, theory, practice, and open questions. *International Journal of Game Theory* 36, 537–69, March.

Roth, A.E. and Sotomayor, M. 1990. *Two-Sided Matching: A Study in Game-Theoretic Modeling and*

Analysis. Econometric Society Monograph Series. Cambridge: Cambridge University Press.

Shapley, L.S. and Shubik, M. 1971. The assignment game I: The core. *International Journal of Game Theory* 1, 111–30, December.

Sobel, J. 2009. Regale: Some memorable results. *Games and Economic Behavior* (forthcoming).

How to cite this article

Sobel, Joel. "Gale, David (1921–2008) ." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000216> doi:10.1057/9780230226203.1884

The New Palgrave Dictionary of Economics Online

Galiani, Ferdinando (1728 – 1787)

Filippo Cesarano
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balance of payments adjustment; fiduciary monetary systems; Galiani, F.; Hume, D.; inflationary expectations; invisible hand; labour theory of value; Law, J.; laws of nature; money supply; paradox of value; quantity theory of money; scarcity; targets; time preference theory of interest; utility; value, theory of

Article

Galiani was born at Chieti, Italy, on 2 December 1728 and died in Naples on 30 October 1787. At the age of seven he was sent to Naples, where he received a classical education under the supervision of his uncle Celestino Galiani, chief almoner to the king. The young Galiani was in close touch with the cultural circles of the time and was soon introduced to the study of economics. In 1744 he translated some of Locke's writings on money. One year later he took religious orders. His extensive monetary studies culminated in the publication of *Della moneta* (1751), his main work. In 1759 he was appointed secretary of the Neapolitan embassy in Paris where he lived, almost without interruptions, for about ten years. At the end of his stay he wrote the *Dialogues sur le commerce des bléds* (1770). After his return to Naples, Galiani held several high positions in the civil service and published other essays on policy issues and in fields outside economics (Galiani, 1974; 1975).

Most of Galiani's theoretical work can be found in his *Della moneta* (1751), which appeared when he was 22. Despite the variety of topics addressed in the book, the basic contributions concern value and monetary theory. Having defined value as a relationship of subjective equivalence between a quantity of one commodity and a quantity of another, Galiani argues that value depends on utility (*utilitá*) and scarcity (*raritá*) (1751, pp. 36–56). Utility is the property of commodities to procure welfare or happiness. Man does not wish to satisfy only primary wants – like eating, drinking, and sleeping – because, once the latter have been satisfied, several others emerge so that full satisfaction is not attainable. Thus, a non-satiation postulate is assumed to hold. Scarcity refers to the quantity of goods available in the market. Although the interdependence between price and quantity in the determination of market equilibrium is clearly explained by Galiani (1751, pp. 53–4), together with the concept of

demand elasticity with respect to wealth, he states that the value of commodities is given by the quantity of labour. Galiani's stress is on value as a relative notion, not related to the intrinsic properties of commodities (1751, p. 119). This theoretical framework allows him to offer a lucid explanation of the so-called paradox of value; according to Schumpeter (1954, p. 300), he ‘carried this analysis to its 18th-century peak’.

The main subject of Galiani's 1751 book, however, is money. In order to analyse the properties of a monetary economy, he inquires into the feasibility of dispensing with the use of money altogether, as in religious communities (1751, pp. 87–91). In a large society, goods could be deposited in public warehouses where each producer would be given a receipt (*bullettino*) stating the quantity of commodities deposited so that he would be entitled to withdraw an equivalent amount of commodities. Relative prices would be fixed by the prince. Yet these receipts are nothing but money; money is the means by which everyone's product is represented. Galiani's analysis foreshadows a basic idea shown by recent research (Ostroy, 1973), that is, that money is a mechanism to avoid inconsistent claims on commodities on the part of individuals who are motivated by self-interest (1751, p. 90). This analysis notwithstanding, Galiani vigorously rejects a fiduciary monetary system, likely under the influence of events related to John Law's experience in France. These results provide the basis for his theory of the origin of money (1751, pp. 74–81). Media of exchange were not deliberately introduced by man but emerged because some goods had properties that let them be used as means of payment. Galiani's important insight – that the commodities performing monetary functions should be of uniform quality and easily recognizable in order to bring about the reduction of transaction costs and the production of information – can be found in recent work on the subject (Jones, 1976, p. 775).

The validity of the quantity theory of money is taken for granted by Galiani. There is, however, a dynamic process through which equilibrium is attained and during this adjustment period changes in money supply affect the economy (1751, pp. 187–9). The same argument was advanced by David Hume in a celebrated passage (1752, pp. 37–8), one year after the publication of *Della moneta*. Although the inefficacy of expected inflation is clearly stated by Galiani (1751, p. 189), an unexpected increase in prices is thought to bring about benefits and costs. Both are discussed at length, but the analysis is rather poor and marred by inconsistencies. However, Galiani clearly understands that inflation is a concealed way of levying taxes (1751, pp. 198–9, 203–4, 208) and favours the recourse to such a policy in a critical situation when the benefits will more than offset the eventual costs (Cesarano, 1976; 1983).

As regards the international aspects of monetary economics, several passages in *Della moneta* show the basic principles of the theory of balance of payments adjustment, pointing out that money flows should not be tampered with by laws or regulations. Galiani views the balance of payments as an essentially monetary phenomenon and payments imbalances as a necessary event which should never be meddled with. Finally, the rate of interest is defined as the relative price of goods dated at different points in time (1751, pp. 290–1), stressing the role of different degrees of risk. In this analysis, an anticipation of the time preference theory of interest may be found.

Galiani places full trust on the laws of nature which regulate economic phenomena. These laws have universal validity and, like physical laws, can never be violated. Hence, the implementation of policy actions is constrained by the existence of natural laws (Cesarano, 1976, section 1). The economic process is guided by a ‘supreme Hand’ (1751, p. 57) which is the religiously biased counterpart (Galiani was an abbot) of Adam Smith's ‘invisible hand’ a quarter of a century later. This methodological

standpoint can also be found in his later book *Dialogues sur le commerce des bléds* (1770), a discussion of the 1764 French law liberalizing corn exports. The theoretical contributions of this work are not as remarkable as those of *Della moneta*. Nevertheless, the *Dialogues* are to be noted for the rather modern treatment of the principles of economic policy. The latter (1770, pp. 319–23) centres upon the fixing of a target and the choice of the means to achieve it. Galiani stresses the need to avoid abrupt changes in policy and to consider the institutional and political setting before following a specific policy. Although natural laws cannot be violated in the long run and so impose a constraint on policy actions, the latter can be effective in the short run.

Galiani's work on economics reveals a large number of contributions putting him far ahead of his time. Concerning the theory of value, Schumpeter stated:

> … he [Galiani] displayed sure-footed mastery of analytical procedure and, in particular, neatness in his carefully defined conceptual constructions to a degree that would have rendered superfluous all the 19th-century squabbles – and misunderstandings – on the subject of value had the parties to these squabbles first studied his text, *Della moneta*, 1751. (1954, pp. 300–1)

His analysis of the subject of money embodies a rather coherent theoretical structure showing the basic principles upon which classical monetary theory is built.

Selected works

1751. *Della moneta*. Introduction by G. Caracciolo, ed. A. Merola. Milan: Feltrinelli, 1963. Reprinted 1780 with the addition of a foreword, 35 notes and an epilogue. Other editions include: P. Custodi, ed., in *Scrittori classici italiani di economia politica*, Parte Moderna, vols. 3 and 4, Milan: Destefanis, 1803; F. Nicolini, ed., Bari: Laterza, 1915. English translation by P.R. Toscano, as *On Money*, Ann Arbor: University Microfilms International, 1977. A French partial translation has been edited by G.H. Bousquet and J. Crisafulli, *De la monnaie*, Paris: Riviére, 1955. An English translation of the main passages can be found in A.E. Monroe, *Early Economic Thought*, Cambridge, MA: Harvard University Press, 1924.

1770. *Dialogues sur le commerce des bléds*. Ed. F. Nicolini. Naples: Ricciardi, 1959.

1974. *Nuovi saggi inediti di economia*. Introduction by G. Demaria ed A. Agnati, Padua: Cedam.

1975. *Opere*, ed. F. Diaz and M. Guerci. Naples: Ricciardi.

Bibliography

Accademia Nazionale dei Lincei. 1975. *Ferdinado Galiani*. Quaderno N. 211. Rome: Accademia Nazionale dei Lincei.

Cesarano, F. 1976. Monetary theory in Ferdinando Galiani's *Della moneta*. *History of Political Economy* 8, 380–99.

Cesarano, F. 1983. The rational expectations hypothesis in retrospect. *American Economic Review* 73, 198–203.

Einaudi, L. 1953. Galiani economista. In *Saggi bibliografici e storici intorno alle dottrine economiche*. Rome: Edizioni di Storia e Letteratura.

Hume, D. 1752. Of money. In *Writings on Economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.

Jones, R.A. 1976. The origin and development of media of exchange. *Journal of Political Economy* 84, 757–75.

Ostroy, J.M. 1973. The information and efficiency of monetary exchange. *American Economic Review* 63, 597–610.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Tagliacozzo, G., ed. 1937. *Economisti napoletani dei sec. XVII e XVIII*. Bologna: Cappelli.

How to cite this article

Cesarano, Filippo. "Galiani, Ferdinando (1728–1787)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000005> doi:10.1057/9780230226203.0614

The New Palgrave Dictionary of Economics Online

game theory and biology

Olof Leimar

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Darwinian evolutionary dynamics and learning dynamics provide the foundation for game theory in biology. The theory is used to analyse interactions between individuals. Animal fighting behaviour, cooperative interactions and signalling interactions are examples of important areas of application. The payoffs to strategies in biological games represent Darwinian fitness, viz. survival and reproductive success. The strategies can be behaviour patterns, but also choices of phenotypic properties such as becoming a male or a female. The evolutionary analysis of allocation to male and female function is one of the most successful applications of game theory in biology.

Keywords

alternative reproductive strategies; class structured populations; cooperation; deterministic evolutionary dynamics; division of labour; evolutionarily stable strategies; evolutionary theory; game theory and biology; learning and evolution in games; life-history strategy; mate choice; mixed strategy equilibria; natural selection; playing the field; Prisoner's Dilemma; producer–scrounger game; reciprocity; reproductive value; sex ratio theory; signalling; tit for tat

Article

In biology, game theory is a branch of evolutionary theory that is particularly suited to the study of interactions between individuals. The evolution of animal fighting behaviour was among the first applications and it was in this context that Maynard Smith and Price (1973) developed the concept of an evolutionarily stable strategy (ESS) (see learning and evolution in games: ESS). Cooperative interactions (Trivers, 1971) and signalling interactions (Grafen, 1991), such as when males signal their quality to females, are examples of other important areas of application. There is an overlap of ideas between economics and biology, which has been quite noticeable since the 1970s and, in a few instances, earlier (Sigmund, 2005). In the early 21st century, the interchange takes the form of a joint

exploration of theoretical and empirical issues by biologists and economists (Hammerstein and Hagen, 2005).

Strategies in games inspired by biology can represent particular behaviour patterns, including rules about which behaviour to perform in which circumstance. Other aspects of an individual's phenotype can also be viewed as the result of strategic choice. A life-history strategy specifies choices that have major impact on an individual's course of life, for instance, whether to become a male or a female or, for certain insects, whether or not to develop wings. Interactions between individuals are modelled as games where the payoffs represent Darwinian fitness. Random matching of players drawn from a large population is one common game model, which was used to study fighting behaviour (Maynard Smith and Price, 1973; Maynard Smith and Parker, 1976). ‘Playing the field’ (Maynard Smith, 1982) is a more general modelling approach, where the payoff to an individual adopting a particular strategy depends on some average property of the population (cf. population games in deterministic evolutionary dynamics). Game theory is needed for situations where payoffs to strategies depend on the state of a population, and this state in turn depends on the strategies that are present. For matching of players drawn from a population, the distribution of opposing strategies is of course given by the population distribution, but there are other reasons why the distribution of strategies influences expected payoffs. A ‘playing-the-field’ example is the choice by an individual, or by its mother, to develop into a male or a female. The two sexes occur in roughly equal proportions in many species. This observation intrigued Darwin, who was unable to provide a satisfactory explanation, writing that ‘I formerly thought that when a tendency to produce the two sexes in equal numbers was advantageous to the species, it would follow from natural selection, but I now see that the whole problem is so intricate that it is safer to leave its solution to the future’ (Darwin, 1874, p. 399). The solution to the problem was found by Düsing (1884; see also Edwards, 2000, and Fisher, 1930), and rested on the principle that, in diploid sexual organisms, the total genetic contribution to offspring by all males in a generation must equal the contribution by all females in the same generation. This gives a reproductive advantage to the rarer sex in the passing of genes to future generations. The payoffs to a mother from producing a son or a daughter must then depend on the population sex ratio, and this dependence can result in an evolutionary equilibrium at an even sex ratio (see below). The idea arose before the development of the concept of an ESS by Maynard Smith and Price (1973), but it can be regarded as the first instance of game-theoretical reasoning in biology.

Payoffs, reproductive value, and evolutionary dynamics

Class-structured populations in discrete time (Caswell, 2001) are often used as settings for evolutionary analysis. The classes or states are properties like female and male, and time might be measured in years. Let $n_i(t)$ denote the number of individuals in state i at time t. We can write a deterministic population dynamics as $n(t+1) = An(t)$, where n is the vector of the n_i and A is the so-called population projection matrix. The elements a_{ij} of A can depend on n and on the strategies that are present in the population. They represent per capita genetic contributions of individuals in state j to state i, in terms of offspring or individual survival. A common evolutionary analysis is to determine a stationary n for the case where all individuals use a strategy x and to examine whether rare mutants with strategy x' would increase in number in such a population.

Let us apply this scheme to the just mentioned sex ratio problem. Suppose that a mother can determine the sex of her offspring and that females in the population produce a son with probability *x* and a daughter with probability 1 – *x*. For non-overlapping generations, the dynamics $n(t+1) = An(t)$ can be written as

$$\begin{pmatrix} n_f(t+1) \\ n_m(t+1) \end{pmatrix} = \begin{pmatrix} 0.5(1-x)b & 0.5(1-x)bq \\ 0.5xb & 0.5xbq \end{pmatrix} \begin{pmatrix} n_f(t) \\ n_m(t) \end{pmatrix},$$

where *b* is the reproductive output (number of offspring) of a female, *bq* is the reproductive output of a male, and the factor 0.5 accounts for the genetic shares of the parents in their offspring. Because the reproductive output of all males must equal that of all females, it follows that $q = n_f(t) / n_m(t)$ and thus that $q = (1-x)/x$. In a stationary population, *b*=1/(1 – *x*) must hold, which could come about through a dependence of *b* on the total population size. Introducing the matrix

$$B(x', x) = \frac{1}{2} \begin{pmatrix} (1-x')/(1-x) & (1-x)/x \\ x'/(1-x) & 1 \end{pmatrix},$$

the population projection matrix for a stationary population is $A = B(x, x)$ and the stationary $n = (n_f, n_m)$ is proportional to the leading eigenvector, $w = (1-x, x)$, of this *A*. Suppose a mutant gene causes a female to adopt the sex ratio strategy x', but has no effect in a male. As long as the mutant gene is rare, only the strategy of heterozygous mutant females needs to be taken into account, and the dynamics of the mutant sub-population can be written as $n'(t+1) = A'n'(t)$ with $A' = B(x', x)$. The mutant can invade if $\lambda(x', x) > 1$ holds for the leading eigenvalue $\lambda(x', x)$ of $B(x', x)$. Direct computation of the leading eigenvalue shows that a mutant with $x' > x$ can invade if $x<0.5$ and one with $x'<x$ can invade if $x' > 0.5$, resulting in an evolutionary equilibrium at $x=0.5$.

The reproductive value of state *i* is defined as the *i*th component of the leading 'left eigenvector' *v* of the stationary population projection matrix $A = B(x, x)$, that is, *v* is the leading eigenvector of the transpose of *A*. It is convenient to normalize *v* so that its scalar product $v \cdot w$ with the leading eigenvector *w* equals 1. For our sex ratio problem we have $v = \frac{1}{2}(1/(1-x), 1/x)$. The reproductive value of state *i* can be interpreted as being proportional to the expected genetic contribution to future generations of an individual in state *i*. The eigenvectors *v* and *w* can be used to investigate how the leading eigenvalue depends on x' near $x' = x$. It is easy to show that $\partial\lambda(x', x) / \partial x' = \partial(v \cdot B(x', x)w) / \partial x'$ holds at $x' = x$ (for example, Caswell, 2001), and this result can be used to identify evolutionary equilibria. If a

mutation has an effect in only one of the states, like the females in our example, there is further simplification in that only one column of $B(x', x)$ depends on x'. It follows that evolutionary change through small mutational steps in the sex ratio example can be described as if females were selected to maximize the expected reproductive value per offspring, given by $V(x', x) = \frac{1}{2}(1 - x') / (1 - x) + \frac{1}{2} x' / x$. Payoff functions having this form were introduced by Shaw and Mohler (1953), in what may have been the first worked out game-theoretical argument in biology. As we have seen, analysis of such payoff functions corresponds to an analysis of mutant invasion in a stationary population.

The concept of reproductive value was introduced by Fisher (1930) and plays an important role in the very successful field of sex ratio theory (Charnov, 1982; Pen and Weissing, 2002), as well as in evolutionary theory in general (McNamara and Houston, 1996; Houston and McNamara, 1999; Grafen, 2006). The concept is useful to represent payoffs in games played in populations in stationary environments. Reproductive value can be regarded as a Darwinian representation of the concept of utility in economics. For populations exposed to large-scale environmental fluctuations, as well as for those with limit-cycle or chaotic attractors of the population dynamics, concepts similar to reproductive value have proven less useful. In such situations, one needs the more general approach of explicitly considering evolutionary dynamics for populations of players of strategies. There are several influential approaches to the study of evolutionary dynamics in biology (Nowak and Sigmund, 2004), ranging from replicator dynamics (see deterministic evolutionary dynamics) and adaptive dynamics (Metz, Nisbet and Geritz, 1992; Metz et al., 1996; Hofbauer and Sigmund, 1998) to the traditional modelling styles of population genetics and quantitative genetics (Rice, 2004). These approaches make different assumptions about such things as the underlying genetics and the rate and distribution of mutation. Recent years have seen an increasing emphasis on explicitly dynamical treatments in evolutionary theory.

Are there mixed strategies in nature?

Biologists have wondered how individuals, as players of a game, come to play one strategy or another. For life-history strategies, involving choices between alternative phenotypes, a population containing a mixture of phenotypes could be the result of randomization at the level of an individual, which corresponds to a mixed strategy, or there could be a genetic polymorphism of pure strategies (Maynard Smith, 1982). These two possibilities can be contrasted with a third, where individuals (or their parents) use information about themselves or their local environment to make life-history choices, which could correspond to a conditional strategy in a Bayesian game. The general question is related to the issue of purification of mixed strategy equilibria in game theory (see purification). When observing populations that are mixtures of discrete phenotypes, biologists have tried to establish if one of the above three possibilities applies and, if so, what the evolutionary explanation might be. This question has been asked, for instance, about the phenomenon of alternative reproductive strategies (Gross, 1996; Shuster and Wade, 2003), like the jack and hooknose males in coho salmon (Gross, 1985) or the winged and wingless males in fig wasps (Hamilton, 1979). Since there are likely to be a number of factors that

influence the relative success of reproductive alternatives and could be known to a developing individual – for instance, its juvenile growth rate and thus its potential adult size – one might expect some form of conditional strategy to evolve. This expectation agrees with the observation that conditional determination is common (Gross, 1996). There are also instances of genetic determination of reproductive alternatives (Shuster and Wade, 2003) but, somewhat surprisingly, there is as yet no empirically confirmed case of a mixed strategy of this kind. Perhaps it has been difficult for evolution to construct a well-functioning randomization device, leaving genetic polymorphism as a more easily achieved evolutionary outcome.

Evolution of cooperation

Among the various applications of game theory in biology, the evolution of cooperation is by far the most studied issue. This great interest is based on the belief that cooperation has played a crucial role in the evolution of biological organization, from the structure of chromosomes, cells and organisms to the level of animal societies. An extreme form of cooperation is that of the genes operating in an organism. Several thousand genes coordinate and direct cellular activities that in the main serve the well-being of their organism. Kin selection (Hamilton, 1964), which predicts that agents have an evolutionary interest in assisting their genetic relatives, cannot be the main explanation for this cooperation, since the different genes in an organism are typically not closely related by descent (except for a given gene in one cell and its copies in other cells). It is instead division of labour that is the principle that unites the parts of an organism into a common interest, of sufficient strength to make it evolutionarily unprofitable for any one gene to abandon its role in the organism for its own advantage. There are of course exceptions, in the form of selfish genetic elements, but these represent a minority of cases (Burt and Trivers, 2006).

Trivers (1971) and Axelrod and Hamilton (1981) promoted the idea that many of the features of the interactions between organisms would find an explanation in the give and take of direct reciprocation. In particular, the strategy of tit for tat (Axelrod and Hamilton, 1981) for the repeated Prisoner's Dilemma game was thought to represent a general mechanism for reciprocity in cooperative interactions and received much attention from biologists. On the whole, this form of direct reciprocity has subsequently failed to be supported by empirical observation. Two reasons for this failure have been proposed (Hammerstein, 2003). One is that the structure of real biological interactions differs in important ways from the original theoretical assumptions of a repeated game. The other is that the proposed strategies, like tit for tat, are unlikely to be reached by evolutionary change in real organisms, because they correspond to unlikely behavioural mechanisms. In contrast to reciprocity, both the influence of genetic relatedness through kin selection (Hamilton, 1964) and the presence of direct fitness benefits to cooperating individuals have relatively strong empirical support. Division of labour and the direct advantages of the trading of benefits between agents are likely to be crucial ingredients in the explanation of cooperation between independent organisms. The idea of a market, where exchanges take place, is thus relevant in both biology and economics (Noe and Hammerstein, 1994).

Evolution of signalling

Signals are found in a wide variety of biological contexts, for instance in aggressive interactions, parent–offspring interactions, and in connection with mate choice. There is now a fairly well developed set of theories about biological signals (Maynard Smith and Harper, 2003). One of the most influential ideas in the field is Zahavi's handicap principle (Zahavi, 1975). It states that a signal can reliably indicate high quality of the signaller only if the signal is costly, to the extent that it does not pay low-quality individuals to display the signal. The idea can be seen as a non-mathematical version of Spence's signalling theory (Spence, 1973; 1974), but, because biologists, including Zahavi, were unaware of Spence's work in economics, Zahavi's principle remained controversial in biology until Grafen (1991) provided a game-theoretical justification. The turn of events illustrates that biologists might have benefited from being more aware of theoretical developments in economics.

An example where Zahavi's handicap principle could apply is female mate choice in stalk-eyed flies (David et al., 2000). Males of stalk-eyed flies have long eye stalks, increasing the distance between their eyes, which is likely to be an encumbrance in their day-to-day life. A high level of nutrition, but also the possession of genes for high phenotypic quality, cause males to develop longer eye stalks. Female stalk-eyed flies prefer to mate with males with eyes that are far apart, and in this way their male offspring have a greater chance of receiving genes for long eye stalks. Female choice will act to reduce genetic variation in males, but if a sufficiently broad range of genetic loci can influence eye-stalk length, because they have a general effect on the phenotypic quality of a male, processes like deleterious mutation could maintain a substantial amount of genetic variation. In this way, signalling theory can explain the evolution of elaborate male ornaments, together with a mating preference for these ornaments in females, illustrating the power of game-theoretical arguments to increase our understanding of biological phenomena.

Learning

Viewing strategies as genetically coded entities on which natural selection operates, with evolutionarily stable strategies as endpoints of evolutionary change, is not the only game-theoretical perspective that is of relevance in biology. For many categories of behaviour, learning or similar adjustment processes are important in shaping the distribution of strategies in a population. For instance, when animals search for food or locate suitable living quarters, they may have the opportunity to evaluate the relative success of different options and to adjust their behaviour accordingly. A well-studied example is the so-called producer–scrounger game, for which there are experiments with birds that forage in groups on the ground (Barnard and Sibly, 1981; Giraldeau and Caraco, 2000). The game is played by a group of foragers and consists of a number of rounds. In each round an individual can choose between two behavioural options. Producers search for and utilize new food sources, and scroungers exploit food found by producers. The game presupposes that the activities of producing and scrounging are incompatible and cannot be performed simultaneously, which is experimentally supported (Coolen, Giraldeau and Lavoie, 2001). The payoffs to the options, measured as the expected food intake per round, depend on the frequencies of the options in the group. For instance, scrounging is most profitable to an individual if no one else scrounges and yields a lower payoff with more scrounging in the group. By specifying the details of the model, one can compute an equilibrium probability of scrounging at which the payoffs to producing and scrounging are equal. This equilibrium is influenced by parameters

like expected search times and the amount of food found in a new location. It has been experimentally verified that groups of spice finches converge on such an equilibrium over a period of a few days of foraging (Mottley and Giraldeau, 2000; Giraldeau and Caraco, 2000). It is not known precisely which rules are used by individuals in these experiments to modify their behaviour, but such rules are likely to play an important role in shaping behaviour in many animals, including humans (see learning and evolution in games: adaptive heuristics). The study of these kinds of adjustments of behaviour could therefore represent an important area of overlap between biology and economics.

See Also

- deterministic evolutionary dynamics
- learning and evolution in games: adaptive heuristics
- learning and evolution in games: ESS
- mixed strategy equilibrium
- purification
- utility

Bibliography

Axelrod, R. and Hamilton, W.D. 1981. The evolution of cooperation. *Science* 211, 1390–6.

Barnard, C.J. and Sibly, R.M. 1981. Producers and scroungers: a general model and its application to captive flocks of house sparrows. *Animal Behaviour* 29, 543–50.

Burt, A. and Trivers, R. 2006. *Genes in Conflict: The Biology of Selfish Genetic Elements*. Cambridge: Harvard University Press.

Caswell, H. 2001. *Matrix Population Models*, 2nd edn. Sunderland: Sinauer.

Charnov, E.L. 1982. *The Theory of Sex Allocation*. Princeton: Princeton University Press.

Coolen, I., Giraldeau, L.-A. and Lavoie, M. 2001. Head position as an indicator of producer and scrounger tactics in a ground-feeding bird. *Animal Behaviour* 61, 895–903.

Darwin, C. 1874. *The Descent of Man and Selection in Relation to Sex*, 2nd edn. London: Murray.

David, P., Bjorksten, T., Fowler, K. and Pomiankowski, A. 2000. Condition-dependent signalling of genetic variation in stalk-eyed flies. *Nature* 406, 186–8.

Düsing, C. 1884. *Die Regulierung des Geschlechtsverhältnisses*. Jena: Fischer.

Edwards, A.W.F. 2000. Carl Düsing (1884) on the regulation of the sex-ratio. *Theoretical Population Biology* 58, 255–7.

Fisher, R.A. 1930. *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

Giraldeau, L.-A. and Caraco, T. 2000. *Social Foraging Theory*. Princeton: Princeton University Press.

Grafen, A. 1991. Biological signals as handicaps. *Journal of Theoretical Biology* 144, 517–46.

Grafen, A. 2006. A theory of Fisher's reproductive value. *Journal of Mathematical Biology* 53, 15–60.

Gross, M.R. 1985. Disruptive selection for alternative life histories in salmon. *Nature* 313, 47–8.

Gross, M.R. 1996. Alternative reproductive strategies and tactics: diversity within sexes. *Trends in Ecology & Evolution* 11, 92–8.

Hamilton, W.D. 1964. The genetical evolution of social behaviour, I, II. *Journal of Theoretical Biology* 7, 1–52.

Hamilton, W.D. 1979. Wingless and fighting males in fig wasps and other insects. In *Reproductive Competition, Mate Choice and Sexual Selection in Insects*, ed. M.S. Blum, and N.A. Blum. New York: Academic Press.

Hammerstein, P. 2003. Why is reciprocity so rare in social animals? A protestant appeal. In *Genetic and Cultural Evolution of Cooperation*, ed. P. Hammerstein. Cambridge: MIT Press.

Hammerstein, P. and Hagen, E.H. 2005. The second wave of evolutionary economics in biology. *Trends in Ecology & Evolution* 20, 604–9.

Hofbauer, J. and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Houston, A.I. and McNamara, J.M. 1999. *Models of Adaptive Behaviour: An Approach Based on State*. Cambridge: Cambridge University Press.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, J. and Harper, D. 2003. *Animal Signals*. Oxford: Oxford University Press.

Maynard Smith, J. and Parker, G.R. 1976. The logic of asymmetric contests. *Animal Behaviour* 24, 159–

75.

Maynard Smith, J. and Price, G.R. 1973. The logic of animal conflict. *Nature* 246, 15–8.

McNamara, J.M. and Houston, A.I. 1996. State-dependent life histories. *Nature* 380, 215–21.

Metz, J.A.J., Nisbet, R.M. and Geritz, S.A.H. 1992. How should we define ‘fitness’ for general ecological scenarios? *Trends in Ecology & Evolution* 7, 198–202.

Metz, J.A.J., Geritz, S.A.H., Meszéna, G., Jacobs, F.J.A. and van Heerwaarden, J.S. 1996. Adaptive dynamics, a geometrical study of nearly faithful reproduction. In *Stochastic and Spatial Structures of Dynamical Systems. Proceedings of the Royal Dutch Academy of Science (KNAV Verhandelingen)*, ed. S. J. van Strien, and S.M. Verduyn Lunel. Amsterdam: North-Holland.

Mottley, K. and Giraldeau, L.-A. 2000. Experimental evidence that group foragers can converge on predicted producer-scrounger equilibria. *Animal Behaviour* 60, 341–50.

Noe, R. and Hammerstein, P. 1994. Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology* 35, 1–11.

Nowak, M.A. and Sigmund, K. 2004. Evolutionary dynamics of biological games. *Science* 303, 793–9.

Pen, I. and Weissing, F.J. 2002. Optimal sex allocation: steps towards a mechanistic theory. In *Sex Ratios – Concepts and Research Methods*, ed. I.C.W. Hardy. Cambridge: Cambridge University Press.

Rice, S.H. 2004. *Evolutionary Theory – Mathematical and Conceptual Foundations*. Sunderland: Sinauer.

Shaw, R.F. and Mohler, J.D. 1953. The selective significance of the sex ratio. *American Naturalist* 87, 337–42.

Shuster, S.M. and Wade, M.J. 2003. *Mating Systems and Strategies*. Princeton: Princeton University Press.

Sigmund, K. 2005. John Maynard Smith and evolutionary game theory. *Theoretical Population Biology* 68, 7–10.

Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87, 355–74.

Spence, M. 1974. *Market Signaling*. Cambridge: Harvard University Press.

Trivers, R.L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46, 35–57.

Zahavi, A. 1975. Mate selection - a selection for a handicap. *Journal of Theoretical Biology* 53, 205–14.

How to cite this article

Leimar, Olof. "game theory and biology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000214> doi:10.1057/9780230226203.0616

The New Palgrave Dictionary of Economics Online

game theory in economics, origins of

Robert Leonard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Game theory entered economics with the publication in 1944 of the *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern. The authors were, respectively, a Hungarian mathematician and an Austrian economist. Paying attention to the scientific and cultural context, this article discusses the creation, content and impact of that work.

Keywords

Austrian economics; cardinal utility; coalitional game; coalitions; coopperative games; Courant, R.; Cournot, A. A.; Debreu, G.; dominance; evolutionary biology; existence of equilibrium; experimental economics; game theory; game theory in economics, history of; general equilibrium; Hilbert, D.; mathematics and ecoonmics; Menger, K.; minimax theorem; Morgenstern, O.; Nash equilibrium; Nash, J.; non-cooperative games; Shapley value; Shapley. L.; side payments; strategic equivalence; Ville, J.; von Neumann, J.; Weyl, H.

Article

> Johnny called me; he likes my manuscript …. I am very happy about this. After all, it wasn't easy for me to simplify his mathematical theory, and to represent it correctly. He is working continuously without a break; it is nearly eerie.
> Oskar Morgenstern, Diary, 7 August 1941 (author's translation)

Thus confided Oskar Morgenstern to his wartime diary at Princeton, while working on the introductory chapter of what would become the *Theory of Games and Economic Behavior* (1944). Like other private reflections written at the time, it speaks to the distance that lay between him, the Viennese economist, and ‘Johnny’ von Neumann, the Hungarian mathematician. The two exiles had come to Princeton by quite different paths, and, in 1941, had known each other well for only two years. A product of the

Austrian school of economics, Morgenstern had strong critical faculties and epistemological interests, but limited mathematical training. Von Neumann, on the other hand, was a Hungarian mathematician of first rank, with little time for philosophical speculation but boundless confidence in the application of mathematics across the scientific domain. Yet, differences notwithstanding, they managed to forge a fruitful partnership, writing a landmark 600-page book on mathematical social science that would mark the creation of game theory and link them thereafter in the public eye.

Von Neumann (1903–1957), the privately tutored mathematical prodigy, came from a prosperous banking family of assimilated Hungarian Jews. It is not insignificant for the present subject that, during his formative years, he was witness to political upheaval, including not only the First World War but also, in Hungary, the 1919 Communist revolution of Bela Kun and its subsequent brutal suppression. Indeed, the Kun regime saw the von Neumann family temporarily flee Budapest. He also watched the growth of anti-semitism in Hungary, which would increasingly restrict the opportunities available to even well-integrated Jews such as himself. In the mid-1920s, he completed degrees in mathematics and chemical engineering at Budapest and Zurich, during that time writing several papers, mainly in the areas of axiomatic set theory and the consistency of mathematics. In 1926, he became postdoctoral fellow at the University of Göttingen, then a world centre in mathematics, whose rich environment allowed him to work close to not only its leader David Hilbert but other luminaries such as Richard Courant and Hermann Weyl. During this period, he continued working on set theory and foundations and, in particular, the mathematical theory of quantum mechanics (see von Neumann, 1932). In these works, there emerge features that would characterize von Neumann's use of mathematics in social science, including an emphasis on achieving axiomatic description of the field under study and, perhaps inherited from quantum mechanics, a belief in the inherently probabilistic nature of the world.

Both of these features surfaced in another of von Neumann's Göttingen papers, one which stood apart from his main interests. This was his 1928 ‘Zur Theorie der Gesellschaftsspiele’, the theory of parlour games, first presented when von Neumann was 23 years old. Were space restrictions unimportant here, we could explore the rich background to the mathematical treatment of games. Chess held a great place in the Jewish culture of *Mitteleuropa* at the turn of the 20th century: from the psychological investigation of the thought processes of the grandmasters to the use of the game as source of inspiration in novelists' fiction. Legendary chess champion and mathematician, Emanuel Lasker, drew on the game for inspiration as he wrote about the workings of social life. Other mathematician contemporaries of his wondered whether so human an activity as chess could be made amenable to formal treatment. The key figures here were Ernest Zermelo in 1913 and then, in the 1920s, von Neumann's Hungarian contemporaries, Dénes König and László Kalmár. Independently, in Paris, again in the 1920s, French mathematician, Emile Borel, drew on his experience as a player of cards rather than chess to begin constructing a mathematical analysis of games involving strategy and to probe the question of equilibrium play. Von Neumann's paper may be regarded as the crowning contribution of these mathematical investigations.

His paper is a brilliant description of the generic two-person, zero-sum game, that is, in which the interests of the two players are directly opposed. He defines the game by the strategies available to both players and their associated payoffs, and, never too concerned about elegance in mathematics, gives a tortuously difficult proof of the existence of a minimax equilibrium. This is a preferred way to play, possibly requiring that strategies be chosen in a probabilistic manner, that allows each player to minimize the amount ceded to the other. The paper, which brought the discussion of the existence of

such an equilibrium to a close, finished with preliminary suggestions about how to extend the analysis to games of three or more players, in terms of coalitions and their winnings. But it went no further than that, and neither did von Neumann. Apart from some unpublished work at the time, in which he showed how the strategy of ‘bluffing’ in a simple two-person poker corresponded to the mathematically rational way to play, he essentially put the theory of games aside. Following a period as *Privatdozent* at the University of Berlin, he spent six months of 1930 at Princeton University's Department of Mathematics. The following year, gauging that his opportunities in Europe were limited, he moved permanently to the United States, where, along with Albert Einstein, he became one of the first members of the newly founded Institute for Advanced Study, close to Princeton University.

Morgenstern (1906–76) was part of the thriving interwar economics community in Vienna constituted by such figures as Ludwig von Mises, Hans Mayer, Friedrich Hayek, Fritz Machlup and Gottfried Haberler. Having obtained his *Habilitation* in 1928, he became lecturer at the University of Vienna and then director of the Rockefeller-financed *Institut für Konjunkturforschung* (Business Cycle Institute). Although an Austrian economist, he was not as ardent an advocate of laissez-faire liberalism as Mises and Hayek, being closer to von Wieser and Hans Mayer, both of whom were more accepting of public intervention and strong government. As Institute director, Morgenstern also had to accommodate himself to the authoritarian Christian Social government that governed Austria between 1934 and 1938. Like his fellow members of the Austrian School, not least Hayek, Morgenstern was critical of general equilibrium theory, and particularly of what the Viennese viewed as its lack of precision in treating the knowledge, beliefs and expectations of forward-looking economic actors (see Morgenstern, 1935).

Unlike many of his Viennese colleagues, however, and notwithstanding his own limited training in the subject, Morgenstern learned to see the further application of mathematics as a means by which to improve the rigour of economic theory. In this, he differed from Mises and Hayek, for example, who were quite sceptical about what was to be gained by applying mathematics to the non-mechanical realm of human action. Here, Morgenstern was influenced by his contact with mathematician Karl Menger, who was son of the founder of the Austrian School of economics, was close to the Vienna Circle, and was leader of that small coterie of mathematical economists in interwar Vienna, which included Karl Schlesinger, Abraham Wald and a young Franz Alt. Menger's friendship, activities and writings, including his 1934 book on ethics and social compatibility, were of fundamental importance to Morgenstern at this time (see Menger, 1934; Leonard, 1998).

Like von Neumann, Morgenstern's career was shaped in part by social and political upheaval. In 1927, Vienna was the theatre of political violence between the Austrian Right and the Socialists. In 1934, there was a civil war in Austria, as the conservative Chancellor Dollfuss crushed the Left, and in 1938 the country was annexed by Germany. This marked the demise of one of the most intellectually and culturally active cities of interwar Europe. Although not Jewish, Morgenstern found himself ousted from his Institute and, leaving Austria in 1938, he took a position at the department of economics at Princeton University. The latter was then very much a sleepy gentlemen's college, so that, for sophisticated intellectual company, Morgenstern found himself turning towards the mathematicians and physicists at the Institute of Advanced Study.

By this time, von Neumann was already returning to game theory. Throughout the 1930s his correspondence shows him to be a very astute observer from afar of the political situation in Europe, and it was against this background of irrationality and social instability that he returned, at the close of the decade, to the development of a mathematics of alliances and coalitions. In late 1940 and 1941, quite

independent of Morgenstern, he extended his 1928 theory to the treatment of three, four and more players, culminating in an analysis of the general n-person, non-zero-sum game. These ideas on games attracted Morgenstern, who saw in them not so much a theory of the social order as a response to his Viennese concerns about how to model the interaction between economic agents. The ensuing collaboration with von Neumann in the period 1941–3 resulted in the *Theory of Games and Economic Behavior*, the entire technical apparatus of which was provided by von Neumann, and the introduction and general orientation by Morgenstern (see Leonard, 1995; 2008).

Contents of Theory of Games and Economic Behavior

That introductory chapter is the most accessible, and therefore most widely read and cited, part of the *Theory of Games*. It is at once a defence of the use of mathematics in social science and a critique of the prevailing state of mathematical economics. Von Neumann's influence, and almost religious faith in the supremacy of mathematical formalism, is clear throughout. There is nothing intrinsically different about social science, it is claimed, that renders it inaccessible to mathematical treatment. Natural phenomena, whether or not they concern human behaviour, are potential repositories of mathematics, the richness of which is likely to be correlated with the empirical prominence of the field. Social and economic activity is of such great worldly importance that it is likely to, so to speak, generate a mathematics of its own. The most prominent treatment of the area, general equilibrium theory, is merely the imitative grafting of physical science methods onto social science. This brings with it assumptions about underlying continuity of change, whereas the social domain likely requires attention to discretely separate structures, and thus the use of a different mathematics. General equilibrium theory has also failed to account for the properly interactive nature of social behaviour, particularly that which is manifest in situations involving ‘small’ numbers of agents, be they involved in the exchange of goods or in the distribution of gains through the formation of social and political groups. Throughout the book, von Neumann's preference for ‘modern’, discrete mathematics (that is, set theory and combinatorics) over the differential and integral calculus is evident. Several pages are devoted to defending the use of cardinal, or numerical, utilities, with the axiomatic proof of the existence of a cardinal utility function being included in an Appendix to the second edition, published in 1947.

Chapter 2 lays out the notion of a game, introducing the mathematical concepts of sets and partitions, and showing how the game may be described axiomatically in these terms. The whole is presented as a piece of modern, axiomatic mathematics in the spirit of Hilbert, which is to say that, although the axioms are stimulated by the common-sense features of games, the latter are soon let recede into the background and the theory pursued in a spirit of relative abstraction. While the mathematics is being followed through, the empirical is held at arm's length and everyday terms are introduced in inverted commas – hence, ‘class’, ‘discrimination’, ‘exploitation’, and so on. Only during periodic returns to the heuristics is the vocabulary of the everyday re-invoked, and the ‘common sense’ meaning of the results discussed. The minimax theorem is proved in the next chapter, using, not von Neumann's earlier proof, but a modification of the elementary 1938 proof by Borel's student Jean Ville, based on the theory of convex sets. From here on, chapter by chapter, von Neumann systematically goes through the zero-sum game for three, four and more players, exploring their combinatorial possibilities for coalition-formation and compensations (side payments). Each game is described in terms of its characteristic function, which shows the maximal payoff available to each possible coalition of the game, assuming that the

coalition plays minimax against its complement and that utility is transferable between players. In Chapter 9, the concept of strategic equivalence is introduced to show how the move from the zero-sum restriction to a constant sum retains the basic features of the game, thus allowing it to be solved by the same means. In the eleventh chapter, von Neumann drops the zero (or constant) sum restriction, moving to the ‘general game’.

The central theoretical part of the *Theory of Games* is von Neumann's solution to coalitional games, the stable set. The solution is a ‘complicated combinatorial catalogue’, indicating the minimum each participant can get if he behaves rationally. He may, of course, get more if the others behave ‘irrationally’, that is, make mistakes. Were the solution to consist of a single imputation – a vector of amounts to be received by each player – then the ‘structure of the society under consideration would be extremely simple: There would exist an absolute state of equilibrium in which the quantitative share of every participant would be precisely determined’ (1944, p. 34). However, such a unique solution does not generally exist – a given society can be organized in various ways – so the notion needs to be broadened. The solution is thus a *set* of possible imputations.

> Any particular alliance describes only one particular consideration which enters the minds of the participants when they plan their behavior. Even if a particular alliance is ultimately formed, the division of the proceeds between the allies will be decisively infuenced by the other alliances which each one might alternatively have entered … It is, indeed, this whole which is the really significant entity, more so than its constituent imputations. Even if one of these is actually applied, i.e., if one particular alliance is actually formed, the others are present in a ‘virtual’ existence: Although they have not materialized, they have contributed essentially to shaping and determining the actual reality. (1944, p. 36)

In an n-person game, therefore, a ‘solution should be a system of imputations possessing in its entirety some kind of balance and stability the nature of which we shall try to determine. We emphasize that this stability – whatever it may turn out to be – will be a property of the system as a whole and not of the single imputations of which it is composed’ (p. 36).

This stability is based on the notion of ‘domination’. One imputation, x , is said to dominate another, y , ‘when there exists a group of participants each one of which prefers his individual situation in x to that in y, and who are convinced that they are able, as a group – i.e. as an alliance – to enforce their preferences’ (p. 38). Dominance, which is not a transitive ordering since the demurring coalition may be different in each case, forms the basis for game solutions. Von Neumann defines the solution to an n-person game as a set of imputations, S, with the following characteristics:

> No imputation y contained in S is dominated by an imputation x contained in S.
> Every y not contained in S is dominated by some x contained in S.

A solution is thus not a single imputation but a set of possible imputations, and such a set is stable in so far as its member imputations do not dominate each other and every imputation outside the set is dominated by at least one imputation inside. Further, not only is a solution comprised of possibly many imputations, linked by these stability criteria, but a given game may have many solutions. To take a simple example, consider the zero–sum game in which a ‘pie’ of value 1 has to be divided amongst three

people. It has the following four solutions:

1. 1. (1/2, 1/2, 0)•(1/2, 0, 1/2)•(0, 1/2, 1/2)
2. 2. (x_1, x_2, c)•where $0 \leq c \leq 1/2$ and $x_1+x_2+c=1$
3. 3. (c, x_2, x_3)•where $0 \leq c \leq 1/2$ and $x_2+x_3+c=1$
4. 4. (x_1, c, x_3)•where $0 \leq c \leq 1/2$ and $x_1+x_3+c=1$

Here, not only are there multiple solutions, but three of those actually admit an infinite number of possible imputations. Note also that the observation of a given imputation, such as (1/2, 1/2, 0), says little about which solution obtains, as that imputation could occur in any of the four solutions above. The question of which solution will obtain in a given situation, the authors say, can be broached only by considering ‘standards of behaviour’, the various rules, customs or institutions governing social organization at the time. These are extra-game considerations, not contained in the information provided by the characteristic function. To understand the analogy, von Neumann and Morgenstern advise the reader to ‘temporarily forget the analogy with games and think entirely in terms of social organization’ (p. 41, n. 1):

> Let the physical basis of a social economy be given, – or to take a broader view of the matter, of a society. According to all tradition and experience human beings have a characteristic way of adjusting themselves to such a background. This consists of not setting up one rigid system of apportionment, i.e. of imputation, but rather a variety of alternatives, which will probably express some general principles but nevertheless differ among themselves in many particular respects. This system of imputations describes the ‘established order of society’ or ‘accepted standard of behavior’.

Thus, in the above game, in solution 2, player 3 is held to an amount, *c*, that may be as small as zero, or as high as 1/2. The actual value of *c* would reflect the norms governing that player's social standing. Depending on tradition, the marginal member might or might not be completely exploited. As von Neumann and Morgenstern write: ‘A theory which is consistent at this point cannot fail to give a precise account of the entire interplay of economic interest, influence and power’ (p. 43).
When one considers the personal context in which von Neumann developed this social theory, his letters of the time dwelling on European instability, the strategic alliances involving the Germans and the Allies, the plight of the Hungarian Jews, which directly affected his family, and one then reads the *Theory of Games*, with its emphasis on stability and its pervasive reference to norms, discrimination and power, the book appears as his attempt, not simply to replace general equilibrium theory, but to achieve a mathematical description of social organization, broadly defined. And, to the end of his life, von Neumann spoke of it in these terms. For example, in 1955, at a Princeton conference on game theory, when mathematician John Nash raised the problem of the great multiplicity of solutions to cooperative games, von Neumann replied ‘that this result was not surprising in view of the correspondingly enormous variety of observed stable social structures; many differing conventions can endure, existing today for no better reason than that they were here yesterday’ (Wolfe, 1955, p. 25).

Role and impact of the book

The initial influence of the *Theory of Games* was felt, however, not in the area of economic or social theory per se, but in that of military strategy. During the early 1940s, while Morgenstern remained at Princeton struggling with the technical draft chapters, von Neumann was out in the world, increasingly heavily involved as mathematical advisor to various branches of the American military forces. Through his influence on the work of mathematicians at the Princeton branch of the Statistical Research Group and at the Anti-submarine Warfare Operations Research Group, game theory became an element in mathematical models of military engagements such as submarine-search and bombing strategy. Such military models involved the application of a very small part of the mathematics – usually centred on minimax theorem – to specific, confined problems. Thus, game theory qua operations research was far removed from the ambitious, abstract representation of the social order that von Neumann had pursued in the *Theory of Games*. Be that as it may, it was the perceived success of operations research during the Second World War that provided the impetus for the Army Air Corps' creation of the RAND Corporation in the late 1940s, where models of this kind continued to be developed, and for the next decade game theory was given strong institutional support. As it happens, there is little evidence that, throughout the 1950s at RAND, these game theoretic models were of anything other than very limited influence in quantitatively shaping particular strategic decisions. It is incontestable, however, that the language, terminology and ‘thought framework’ of game theory became important to the strategic mindset that dominated the Cold War period, helping shape such books as Herman Kahn's *Thinking the Unthinkable* and Thomas Schelling's *The Strategy of Conflict*.

The *Theory of Games* also set new standards for mathematical rigour in the field of economic theory. For example, before leaving France to move to the United States, economic theorist Gerard Debreu read the book in Salzburg, Austria, at a summer-school run by Harvard University. Though Debreu would never work on game theory, the book shaped his thinking greatly. His pathbreaking *Theory of Value* (1959), an axiomatic treatment of Walrasian general equilibrium theory, refers to the outstanding influence of von Neumann and Morgenstern (1944) ‘which freed mathematical economics from its traditions of differential calculus and compromises with logic’ (1959, p. x). Debreu's stance, too, on the relationship between the mathematics of general equilibrium and the empirical economic substrate is exactly that of von Neumann on games: ‘the theory, in the strict sense, is logically entirely disconnected from its interpretations’ (p. x). This austere, formal view shaped an entire generation of economic ‘high theorists’ from the 1950s till the 1980s, during which general equilibrium theory was the pinnacle of intellectual achievement in the discipline. That von Neumann's game theory should have ended up providing sustenance for Walrasian general equilibrium is one of the many historical ironies in the intellectual history of our field.

It was also in the post-war military–academic milieu that a new generation of game theorists proper came of age. Whether at Princeton or the RAND Corporation, or alternating between the two, young mathematicians such as Lloyd Shapley and John Nash took game theory and made it their own. Shapley, a towering influence in the game theory community from the 1950s onwards, produced, amongst other things, the Shapley value, which described the solution to a coalitional game in terms of the amount brought by each player to an average, randomly formed coalition. For his Ph.D. thesis, Nash sought to provide for n-person games a solution that was as well-defined as von Neumann's minimax for the two-person game. He made a conceptual division of games into cooperative, in which coalitions are

permitted, and non-cooperative, in which players act in isolation. For the latter, he proved the existence, under specific conditions, of what he called an ‘equilibrium point’, later the Nash equilibrium (see Nash, 1950a; 1950b; 1951). That von Neumann found this non-cooperative approach to be rather trivial is understandable in the light of the ambitious social theory he was pursuing, but that leaves unchanged the fact of his influence. It was also at this time that the work of Augustin Cournot was redisovered and reinterpreted in the light of the Nash equilibrium (see Leonard, 1994; Dimand and Dimand, 1996). Subsequent work on non-cooperative game theory by Harsanyi, Selten, Aumann, Kreps and others resulted in a veritable transformation of the microeconomic canon and shaped modelling in industrial organization, international trade and a range of areas (see Dimand, 2000). The field of experimental economics, currently in rapid expansion, owes it existence in part to the appearance of game theory. Although von Neumann himself voiced scepticism as to the ability of laboratory experimentation to shed light on the stable set, game theory did provide a structured basis for empirically testing the theory of individual decision, via its utility axioms, and various solution concepts, both cooperative and non-cooperative. This experimentation, too, began at the RAND Corporation (see Kalisch et al., 1954). It should also be mentioned that, under the influence on John Maynard Smith, the theory of games has had a great impact on the field of evolutionary biology (see Maynard Smith, 1988).

In short, although it quickly attained the status of a classic, which is to say that it was cited by many but read by few, the *Theory of Games and Economic Behavior* set in motion developments that, in ways sometimes quite unintended by the its authors, gradually reshaped the warp and weft of the economics discipline. From the recasting of the economic agent as a strategic player to the reshaping of entire fields of economics; from the introduction to general equilibrium and social welfare theory of axiomatic methods and discrete mathematics to the rise of experimental economics, the direct and indirect effects of von Neumann and Morgenstern's wartime book have been profound and long-lasting.

See Also

- game theory
- Morgenstern, Oskar
- Nash, John Forbes
- von Neumann, John

Bibliography

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New Haven and London: Yale University Press.

Dimand, M.A. and Dimand, R.W. 1996. *A History of Game Theory, Vol. I. From the Beginnings to 1945*. London and New York: Routledge.

Dimand, R. 2000. Strategic games: from theory to application. In *The History of Applied Economics*, ed. R.E. Backhouse and J. Biddle. Annual Supplement to *History of Political Economy* 32, 199–225.

Kahn, H. 1962. *Thinking about the Unthinkable*. New York: Horizon Press.

Kalisch, G., Milnor, J.W., Nash, J. and Nering, E.D. 1954. Some experimental n-person games. In *Decision Processes*, ed. R.M. Thrall, C.H. Coombs and R.L. Davis. New York: Wiley.

Leonard, R. 1994. Reading Cournot, reading Nash: the creation and stabilisation of the Nash equilibrium. *Economic Journal* 104, 492–511.

Leonard, R. 1995. From parlor games to social science: von Neumann, Morgenstern, and the creation of game theory, 1928–1944. *Journal of Economic Literature* 33, 730–61.

Leonard, R. 1998. Ethics and the excluded middle: Karl Menger and social science in interwar Vienna. *Isis* 89(1), 1–26.

Leonard, R. 2008. *Von Neumann, Morgenstern and The Creation of Game Theory, 1900–1960*. New York and Cambridge: Cambridge University Press.

Maynard Smith, J. 1988. *Games, Sex and Evolution*. New York and Toronto: Harvester-Wheatsheaf.

Menger, K. 1934. *Moral, Wille und Weltgestaltung. Grundlegung zur Logik der Sitten*. Vienna: Julius Springer. Trans. as Morality, Decision and Social Organization: Towards a Logic of Ethics. Dordrecht: Reidel, 1974.

Morgenstern, O. 1935. Vollkommene Voraussicht und wirtschaftliches Gleichgewicht. *Zeitschrift für Nationalökonomie* 6, 337–57 Trans. F. Knight, mimeo, University of Chicago. Repr. in *Selected Economic Writings of Oskar Morgenstern*, ed. A. Schotter. New York: NYU Press, 1976.

Morgenstern, O. Diary. Oskar Morgenstern Papers, Special Collections Library, Duke University, USA.

Nash, J. 1950a. Equilibrium Points in N-Person Games. *Proceedings of the National Academy of Science* 36, 48–9.

Nash, J. 1950b. Non-cooperative games. Ph.D. thesis, Princeton University.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

Schelling, T. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320 Trans. S. Bargmann as ‘On the Theory of Games of Strategy’, in *Contributions to the Theory of Games*,

vol. 4, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press, 1959.

von Neumann, J. 1932. *Mathematische Grundlagen der Quantenmechanik*. Berlin: J. Springer. Trans. R. Beyer as *Mathematical Foundations of Quantum Mechanics*. Princeton: Princeton University Press, 1955.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd edn. Princeton: Princeton University Press.

Wolfe, P., ed. 1955. Report of an informal conference on recent developments in the theory of games. Mimeo, Logistics Research Project, Department of Mathematics, Princeton University.

How to cite this article

Leonard, Robert. "game theory in economics, origins of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000193> doi:10.1057/9780230226203.0617

The New Palgrave Dictionary of Economics Online

game theory

R.J. Aumann
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Game theory concerns the behaviour of decision makers whose decisions affect each other. Its analysis is from a rational rather than a psychological or sociological viewpoint. It is indeed a sort of umbrella theory for the rational side of social science, where ‘social’ is interpreted broadly, to include human as well as non-human players (computers, animals, plants). Its methodologies apply in principle to all interactive situations, especially in economics, political science, evolutionary biology, and computer science. There are also important connections with accounting, statistics, the foundations of mathematics, social psychology, law, business, and branches of philosophy such as epistemology and ethics.

Keywords

asymmetric information; Aumann, R. J.; Axelrod, R.; axiomatics; bargaining; bounded rationality; Brouwer's fixed-point theorem; coalitional games; coalitions; common knowledge assumption; competitive equilibrium; consistency; continuous games; cooperative game theory; cooperative games; cores; correlated equilibria; cost allocation; differential games; distributed computing; duality theorems; dynamic games; ethics; evolutionary economics; expected utility theory; extensive form games; fixed threats; folk theorem; game theory; games of incomplete information; general equilibrium; Gillies, D.; Harsanyi, J.; impossibility theorem; imputations; individual rationality; Kakutani's fixed point theorem; kernel; Kuhn, H.; Lemke–Howson algorithm; linear inequalities; linear programming; Luce, R.; market games; mathematical economics; mathematical programming; Milnor, J.; minimax theorem; mixed strategy game; monopoly; monopsony; Morgenstern, O.; Nash equilibrium; Nash program; Nash, J.; net worth; non-cooperative games; nucleolus; oligopoly; perfect information; prisoner's dilemma; probability distributions; Raiffa, H.; Ramsey, F.; randomization; refinements of Nash equilibrium; repeated games; Savage, L.; Selten, R.; shadow pricing; Shapley value; Shapley, L.; Shubik, M.; small worlds; Maynard Smith, J.; solution concepts; stable set theory; statistical decision theory; stochastic games; strategic equilibrium; strategic games; strictly competitive games; strictly determined games; subgame perfection; subgames; superadditivity; testing; tit for tat; transferable utility; Tucker, A.; uncertainty; utility functions; value equivalence principle; von Neumann, J.; voting games; weighted voting game; Zermelo's theorem; zero-sum games

Article

Introduction

‘Interactive decision theory’ would perhaps be a more descriptive name for the discipline usually called game theory. This discipline concerns the behaviour of decision makers (*players*) whose decisions affect each other. As in non-interactive (one-person) decision theory, the analysis is from a rational, rather than a psychological or sociological viewpoint. The term ‘game theory’ steams from the formal resemblance of interactive decision problems (*games*) to parlour games such as chess, bridge, poker, monopoly, diplomacy or battleship. The term also underscores the rational, ‘cold’, calculating nature of the analysis.

The major applications of game theory are to economics, political science (on both the national and international levels), tactical and strategic military problems, evolutionary biology, and, most recently, computer science. There are also important connections with accounting, statistics, the foundations of mathematics, social psychology, and branches of philosophy such as epistemology and ethics. Game theory is a sort of umbrella or ‘unified field’ theory for the rational side of social science, where ‘social’ is interpreted broadly, to

include human as well as non-human players (computers, animals, plants). Unlike other approaches to disciplines like economics or political science, game theory does not use different, ad-hoc constructs to deal with various specific issues, such as perfect competition, monopoly, oligopoly, international trade, taxation, voting, deterrence, and so on. Rather, it develops methodologies that apply in principle to *all* interactive situations, then sees where these methodologies lead in each specific application. Often it turns out that there are close relations between results obtained from the general game-theoretic methods and from the more ad-hoc approaches. In other cases, the game-theoretic approach leads to new insights, not suggested by other approaches.

We use a historical framework for discussing some of the basic ideas of the theory, as well as a few selected applications. But the viewpoint will be modern; the older ideas will be presented from the perspective of where they have led. Needless to say, we do not even attempt a systematic historical survey.

1910 – 1930

During these earliest years, game theory was preoccupied with *strictly competitive* games, more commonly known as *two-person zero-sum* games. In these games, there is no point in cooperation or joint action of any kind: if one outcome is preferred to another by one player, then the preference is necessarily reversed for the other. This is the case for most two-person parlour games, such as chess or two-sided poker; but it seems inappropriate for most economic or political applications. Nevertheless, the study of the strictly competitive case has, over the years, turned out remarkably fruitful; many of the concepts and results generated in connection with this case are in fact much more widely applicable, and have become cornerstones of the more general theory. These include the following:

(i) The *extensive*(or *tree*) *form* of a game, consisting of a complete formal description of how the game is played, with a specification of the sequence in which the players move, what they know at the times they must move, how chance occurrences enter the picture, and the *payoff* to each player at the end of play. Introduced by von Neumann (1928), the extensive form was later generalized by Kuhn (1953), and has been enormously influential far beyond zero-sum theory.

(ii) The fundamental concept of *strategy* (or pure strategy) of a player, defined as a complete plan for that player to play the game, as a function of what he observes during the course of play, about the play of others and about chance occurrences affecting the game. Given a strategy for each player, the rules of the game determine a unique outcome of the game and hence a payoff for each player. In the case of two-person zero-sum games, the sum of the two payoffs is zero; this expresses the fact that the preferences of the players over the outcomes are precisely opposed.

(iii) The strategic (or matrix) form of a game. Given strategies $s^1, \ldots, s^n$ for each of the *n* players, the rules of the game determine a unique outcome, and hence a payoff $H^i(s^1, \ldots, s^n)$ for each player *i*. The *strategic* form is simply the function that associates to each profile $s := (s^1, \ldots, s^n)$ of strategies, the *payoff profile*

$$H(s) := (H^1(s), \ldots, H^n(s)).$$

For two-person games, the strategic form often appears as a matrix: the rows and columns represent pure strategies of Players 1 and 2 respectively, whereas the entries are the corresponding payoff profiles. For zero-sum games, of course, it suffices to give the payoff to Player 1. It has been said that the simple idea of thinking of a game in its matrix form is in itself one of the greatest contributions of game theory. In facing an interactive situation, there is a great temptation to think only in terms of ‘what should I do?’. When one writes down the matrix, one is led to a different viewpoint, one that explicitly takes into account that the other players are also facing a decision problem.

(iv) The concept of *mixed* or *randomized* strategy, indicating that rational play is not in general describable by specifying a single pure strategy. Rather, it is often non-deterministic, with specified probabilities associated with each one of a specified set of pure strategies. When randomized strategies are used, payoff must be replaced by expected payoff. Justifying the use of expected payoff in this context is what led to expected utility theory, whose influence extends far beyond game theory (see *1930–1950*, viii).

(v) The concept of ‘individual rationality’. The *security level* of Player *i* is the amount max min $H^i(s)$ that he can guarantee to himself, independent of what the other players do (here the max is over *i*'s strategies, and the min is over $(n-1)$-tuples of strategies of the players other than *i*). An outcome is called *individually rational* if it yields each player at least his security level. In the game tic-tac-toe, for example, the only individually rational outcome is a draw; and indeed, it does not take a reasonably bright child very long to learn that ‘correct’ play in tic-tac-toe always leads to a draw.

Individual rationality may be thought of in terms of pure strategies or, as is more usual, in terms of mixed strategies. In the latter case, what is being ‘guaranteed’ is not an actual payoff, but an expectation; the word ‘guarantee’ means that this level of payoff can be attained in the mean, regardless of what the other players do. This ‘mixed’ security level is always at least as high as the ‘pure’ one. In the case of tic-tac-toe, each player can guarantee a draw even in the stronger sense of pure strategies. Games like this – i.e. having only one individually rational payoff profile in the ‘pure’ sense – are called *strictly determined*.

Not all games are strictly determined, not even all two-person zero-sum games. One of the simplest imaginable games is the one that game theorists call ‘matching pennies’, and children call ‘choosing up’ (‘odds and evens’). Each player privately turns a penny either heads up or tails up. If the choices match, 1 gives 2 his penny; otherwise, 2 gives 1 his penny. In the pure sense, neither player can guarantee more than –1, and hence the game is not strictly determined. But in expectation, each player can guarantee 0, simply by turning the coin heads up or tails up with 1/2 – 1/2 probabilities. Thus (0, 0) is the only payoff profile that is individually rational in the mixed sense. Games like this – i.e. having only one individually rational payoff profile in the ‘mixed’ sense – are called *determined*. In a determined game, the (mixed) security level is called the *value*, strategies guaranteeing it *optimal*.

(vi) *Zermelo's theorem*. The very first theorem of Game Theory (Zermelo, 1913) asserts that chess is strictly determined. Interestingly, the proof does not construct ‘correct’ strategies explicitly; and indeed, it is not known to this day whether the ‘correct’ outcome of chess is a win for white, a win for black, or a draw. The theorem extends easily to a wide class of parlour games, including checkers, go, and chinese checkers, as well as less well-known games such as hex and gnim (Gale, 1979, 1974); the latter two are especially interesting in that one can use Zermelo's theorem to show that Player 1 can force a win, though the proof is non-constructive, and no winning strategy is in fact known. Zermelo's theorem does not extend to card games such as bridge and poker, nor to the variant of chess known as kriegsspiel, where the players cannot observe their opponents' moves directly. The precise condition for the proof to work is that the game be a two-person zero-sum game of *perfect information*. This means that there are no simultaneous moves, and that everything is open and ‘above-board’: at any given time, all relevant information known to one player is known to all players.

The domain of Zermelo's theorem – two-person zero-sum games of perfect information – seems at first rather limited; but the theorem has reverberated through the decades, creating one of the main strands of game theoretic thought. To explain some of the developments, we must anticipate the notion of *strategic equilibrium* (Nash, 1951; see *1950–1960*, i). To remove the two-person zero-sum restriction, H.W. Kuhn (1953) replaced the notion of ‘correct’, individually rational play by that of equilibrium. He then proved that *every n-person game of perfect information has an equilibrium in pure strategies*.

In proving this theorem, Kuhn used the notion of a *subgame* of a game; this turned out crucial in later developments of strategic equilibrium theory, particularly in its economic applications. A subgame relates to the whole game like a subgroup to the whole group or a linear subspace to the whole space; while part of the larger game, it is self-contained, can be played in its own right. More precisely, if at any time, all the players know everything that has happened in the game up to that time, then what happens from then on constitutes a subgame.

From Kuhn's proof it follows that every equilibrium (not necessarily pure) of a subgame can be extended to an equilibrium of the whole game. This, in turn, implies that every game has equilibria that remain equilibria when restricted to any subgame. R. Selten (1965) called such equilibria *subgame perfect*. In games of perfect information, the equilibria that the Zermelo–Kuhn proof yields are all subgame perfect.

But not all equilibria are subgame perfect, even in games of perfect information. Subgame perfection implies that when making choices, a player looks forward and assumes that the choices that will subsequently be made, by himself and by others, will be rational; i.e. in equilibrium. Threats which it would be irrational to carry through are ruled out. And it is precisely this kind of forward-looking rationality that is most suited to economic applications.

Interestingly, it turns out that subgame perfection is not enough to capture the idea of forward-looking rationality. More subtle concepts are needed. We return to this subject below, when we discuss the great flowering of strategic equilibrium theory that has taken place since 1975, and that coincides with an increased preoccupation with its economic applications. The point we wished to make here is that these developments have their roots in Zermelo's theorem.

A second circle of ideas to which Zermelo's theorem led has to do with the foundations of mathematics. The starting point is the idea of a game of perfect information with an infinite sequence of stages. Infinitely long games are important models for interactive situations with an indefinite time horizon – i.e. in which the players act as if there will always be a tomorrow.

To fix ideas, let A be any subset of the unit interval (the set of real numbers between 0 and 1). Suppose two players move alternately, each choosing a digit between 1 and 9 at each stage. The resulting infinite sequence of digits is the decimal expansion of a number in the unit interval. Let G_A be the game in which 1 wins if this number is in A, and 2 wins otherwise. Using Set Theory's ‘Axiom of Choice’, Gale and Stewart (1953) showed that Zermelo's theorem is false in this situation. One can choose A so that G_A is not strictly determined; that is, against each pure strategy of 1, Player 2 has a winning pure strategy, and against each pure strategy of 2, Player 1 has a winning pure strategy. They also showed that if A is open or closed, then G_A is strictly determined.

Both of these results led to significant developments in foundational mathematics. The axiom of choice had long been suspect in the eyes of mathematicians; the extremely anti-intuitive nature of the Gale–Stewart non-determinateness example was an additional nail in its coffin, and led to an alternative axiom, which asserts that G_A is strictly determined for every set A. This axiom, which contradicts the axiom of choice, has been used to provide an alternative axiomatization for set theory (Mycielski and Steinhaus, 1964), and this in turn has spawned a large literature (see Moschovakis, 1980, 1983). On the other hand, the positive result of Gale and Stewart was successively generalized to wider and wider families of sets A that are ‘constructible’ in the appropriate sense (Wolfe, 1955; Davis, 1964), culminating in the theorem of Martin (1975), according to which G_A is strictly determined whenever A is a Borel set.

Another kind of perfect information game with infinitely many stages is the *differential game*. Here time is continuous but usually of finite duration; a decision must be made at each instant, so to speak. Typical examples are games of pursuit. The theory of differential games was first developed during the 1950s by Rufus Isaacs at the Rand Corporation; his book

on the subject was published in 1965, and since then the theory has proliferated greatly. A differential game need not necessarily be of perfect information, but very little is known about those that are not. Some economic examples may be found in Case (1979).

(vii) *The minimax theorem*. The minimax theorem of von Neumann (1928) asserts that every two-person zero-sum game with finitely many pure strategies for each player is determined; that is, when mixed strategies are admitted, it has precisely one individually rational payoff vector. This had previously been verified by E. Borel (e.g. 1924) for several special cases, but Borel was unable to obtain a general proof. The theorem lies a good deal deeper than Zermelo's, both conceptually and technically.

For many years, minimax was considered the elegant centrepiece of game theory. Books about game theory concentrated on two-person zero-sum games in strategic form, often paying only desultory attention to the non-zero sum theory. Outside references to game theory often gave the impression that non-zero sum games do not exist, or at least play no role in the theory.

The reaction eventually set in, as it was bound to. Game theory came under heavy fire for its allegedly exclusive concern with a special case that has little interest in the applications. Game theorists responded by belittling the importance of the minimax theorem. During the fall semester of 1964, the writer of these lines gave a beginning course in Game Theory at Yale University, without once even mentioning the minimax theorem.

All this is totally unjustified. Except for the period up to 1928 and a short period in the late Forties, game theory was never exclusively or even mainly concerned with the strictly competitive case. The forefront of research was always in n-person or non-zero sum games. The false impression given of the discipline was due to the strictly competitive theory being easier to present in books, more ‘elegant’ and complete. But for more than half a century, that is not where most of the action has been.

Nevertheless, it is a great mistake to belittle minimax. While not the centrepiece of game theory, it *is* a vital cornerstone. We have already seen how the most fundamental concepts of the general theory – extensive form, pure strategies, strategic form, randomization, utility theory – were spawned in connection with the minimax theorem. But its importance goes considerably beyond this.

The fundamental concept of non-cooperative n-person game theory – the strategic equilibrium of Nash (1951) – is an outgrowth of minimax, and the proof of its existence is modelled on a previously known proof of the minimax theorem. In cooperative n-person theory, individual rationality is used to define the set of *imputations*, on which much of the cooperative theory is based. In the theory of repeated games, individual rationality also plays a fundamental role.

In many areas of interest – stochastic games, repeated games of incomplete information, continuous games (i.e. with a continuum of pure strategies), differential games, games played by automata, games with vector payoffs – the strictly competitive case already presents a good many of the conceptual and technical difficulties that are present in general. In these areas, the two-person zero-sum theory has become an indispensable spawning and proving ground, where ideas are developed and tested in a relatively familiar, ‘friendly’ environment. These theories could certainly not have developed as they did without minimax.

Finally, minimax has had considerable influence on several disciplines outside of game theory proper. Two of these are statistical decision theory and the design of distributed computing systems, where minimax is used for ‘worst case’ analysis. Another is mathematical programming; the minimax theorem is equivalent to the duality theorem of linear programming, which in turn is closely related to the idea of shadow pricing in economics. This circle of ideas has fed back into game theory proper; in its guise as a theorem about linear inequalities, the minimax theorem is used to establish the condition of Bondareva (1963) and Shapley (1967) for the non-emptiness of the core of an n-person game, and the Hart-Schmeidler (1988) elementary proof for the existence of correlated equilibria.

(viii) *Empirics*. The correspondence between theory and observation was discussed already by von Neumann (1928), who observed that the need to randomize arises endogenously out of the theory. Thus the phenomenon of bluffing in poker may be considered a confirmation of the theory. This kind of connection between theory and observation is typical of game theory and indeed of economic theory in general. The ‘observations’ are often qualitative rather than quantitative; in practice, we do observe bluffing, though not necessarily in the proportions predicted by theory.

As for experimentation, strictly competitive games constitute one of the few areas in game theory, and indeed in social science, where a fairly sharp, unique ‘prediction’ is made (though even this prediction is in general probabilistic). It thus invites experimental testing. Early experiments failed miserably to confirm the theory; even in strictly determined games, subjects consistently reached individually irrational outcomes. But experimentation in rational social science is subject to peculiar pitfalls, of which early experimenters appeared unaware, and which indeed mar many modern experiments as well. These have to do with the motivation of the subjects, and with their understanding of the situation. A determined effort to design an experimental test of minimax that would avoid these pitfalls was recently made by B. O'Neill (1987); in these experiments, the predictions of theory were confirmed to within less than one per cent.

1930 – 1950

The outstanding event of this period was the publication, in 1944, of the *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern. Morgenstern was the first economist clearly and explicitly to recognize that economic agents must take the interactive nature of economics into account when making their decisions. He and von Neumann met at Princeton in the late Thirties, and started the collaboration that culminated in the *Theory of Games*. With the publication of this book, game theory came into its own as a scientific discipline.

In addition to expounding the strictly competitive theory described above, the book broke fundamental new ground in several directions. These include the notion of a cooperative game, its coalitional form, and its von Neumann–Morgenstern stable sets. Though axiomatic expected utility theory had been developed earlier by Ramsey (1931), the account of it given in this book is what made it ‘catch on’. Perhaps most important, the book made the first extensive applications of game theory, many to economics.

To put these developments into their modern context, we discuss here certain additional ideas that actually did not emerge until later, such as the core, and the general idea of a solution concept. At the end of this section we also describe some developments of this period not directly related to the book, including games with a continuum of strategies, the computation of minimax strategies, and mathematical advances that were instrumental in later work.

(i) *Cooperative games*. A game is called *cooperative* if commitments – agreements, promises, threats – are fully binding and enforceable (Harsanyi 1966, p. 616). It is called *non-cooperative* if commitments are not enforceable, even if pre-play communication between the players is possible. (For motivation, see *1950–1960*, iv.)

Formally, cooperative games may be considered a special case of non-cooperative games, in the sense that one may build the negotiation and enforcement procedures explicitly into the extensive form of the game. Historically, however, this has not been the mainstream approach. Rather, cooperative theory starts out with a formalization of games (the coalitional form) that abstracts away altogether from procedures and from the question of how each player can best manipulate them for his own benefit; it concentrates, instead, on the possibilities for agreement. The emphasis in the non-cooperative theory is on the individual, on what strategy he should use. In the cooperative theory it is on the group: What coalitions will form? How will they divide the available payoff between their members?

There are several reasons that cooperative games came to be treated separately. One is that when one does build negotiation and enforcement procedures explicitly into the model, then the results of a non-cooperative analysis depend very strongly on the precise form of the procedures, on the order of making offers and counter-offers, and so on. This may be appropriate in voting situations in which precise rules of parliamentary order prevail, where a good strategist can indeed carry the day. But problems of negotiation are usually more amorphous; it is difficult to pin down just what the procedures are. More fundamentally, there is a feeling that procedures are not really all that relevant; that it is the possibilities for coalition forming, promising and threatening that are decisive, rather than whose turn it is to speak.

Another reason is that even when the procedures are specified, non-cooperative analyses of a cooperative game often lead to highly non-unique results, so that they are often quite inconclusive.

Finally, detail distracts attention from essentials. Some things are seen better from a distance; the Roman camps around Metzada are indiscernible when one is in them, but easily visible from the top of the mountain. The coalitional form of a game, by abstracting away from details, yields valuable perspective.

The idea of building non-cooperative models of cooperative games has come to be known as the *Nash program* since it was first proposed by John Nash (1951). In spite of the difficulties just outlined, the programme has had some recent successes (Harsanyi, 1982; Harsanyi and Selten, 1972; Rubinstein, 1982). For the time being, though, these are isolated; there is as yet nothing remotely approaching a general theory of cooperative games based on non-cooperative methodology.

(ii) A *game in coalitional form*, or simply *coalitional game* is a function *v* associating a real number *v*(S) with each subset S of a fixed finite set *I*, and satisfying $v(\varnothing) = 0$ ($\varnothing$ denotes the empty set). The members of *I* are called *players*, the subsets S of *I* *coalitions* and *v*(S) is the *worth* of S.

Some notation and terminology: The number of elements in a set S is denoted |S|. A *profile* (of strategies, numbers, etc.) is a function on I (whose values are strategies, numbers, etc.). If *x* is a profile of numbers and S a coalition, we write $x(S) := \Sigma_{i \in S} x^i$. An example of a coalitional game is the *three-person voting game*; here $|I| = 3$, and $v(S) = 1$ or 0 according as to whether $|S| \square 2$ or not. A coalition S is called *winning* if $v(S) = 1$, *losing* if $v(S) = 0$. More generally, if *w* is a profile of non-negative numbers (*weights*) and *q* (the *quota*) is positive, define the *weighted voting game v* by $v(S) = 1$ if $w(S) \square q$, and $v(S) = 0$ otherwise. An example is a parliament with several parties. The players are the parties, rather than the individual members of parliament, w^i is the number of seats held by party *i*, and *q* is the number of votes necessary to form a government (usually a simple majority of the parliament). The weighted voting game with quota *q* and weights w^i is denoted [*q; w*]; e.g., the three-person voting game is [2; 1, 1, 1].

Another example of a coalitional game is a *market game*. Suppose there are *l* natural resources, and a single consumer product, say ‘bread’, that may be manufactured from these resources. Let each player *i* have an endowment e^i of resources (an *l*-vector with non-negative coordinates), and a concave production function u^i that enables him to produce the amount $u^i(x)$ of bread given the vector $x = (x_1, \ldots, x_l)$ of resources. Let *v*(S) be the maximum amount of bread that the coalition S can produce; it obtains this by redistributing its resources among its members in a manner that is most efficient for production, i.e.

$$v(S) = \max\left\{\sum_{i \in S} u^i(x^i) : \sum_{i \in S} x^i = \sum_{i \in S} e^i\right\}$$

where the x^i are restricted to have non-negative coordinates.

These examples illustrate different interpretations of coalitional games. In one interpretation, the payoff is in terms of some single desirable physical commodity, such as bread; v(S) represents the maximum total amount of this commodity that the coalition S can procure for its members, and it may be distributed among the members in any desired way. This is illustrated by the above description of the market game.

Underlying this interpretation are two assumptions. First, that of *transferable utility* (TU): that the payoff is in a form that is freely transferable among the players. Second, that of *fixed threats*: that S can obtain a maximum of v(S) no matter what the players outside of S do.

Another interpretation is that v(S) represents some appropriate index of S's strength (if it forms). This requires neither transferable utility nor fixed threats. In voting games, for example, it is natural to define $v(S) = 1$ if S is a winning coalition (e.g. can form a government or ensure passage of a bill), 0 if not. Of course, in most situations represented by voting games, utility is not transferable.

Another example is a market game in which the x^i are consumption goods rather than resources. Rather than bread, $\Sigma_{i \in S} u^i(x^i)$ may represent a social welfare function such as is often used in growth or taxation theory. While v(S) cannot then be divided in an arbitrary way among the members of S, it still represents a reasonable index of S's strength. This is a situation with fixed threats but without TU.

Von Neumann and Morgenstern considered strategic games with transferable payoffs, which is a situation with TU but without fixed threats. If the profile s of strategies is played, the coalition S may divide the amount $\Sigma_{i \in S} H^i(s)$ among its members in any way it pleases. However, what S gets depends on what players outside S do. Von Neumann and Morgenstern defined v(S) as the maxmin payoff of S in the two-person zero-sum game in which the players are S and $I \setminus S$, and the pay off to S is $\Sigma_{i \in S} H^i(s)$; i.e., as the expected payoff that S can assure itself (in mixed strategies), no matter what the others do. Again, this is a reasonable index of S's strength, but certainly not the only possible one.

We will use the term *TU coalitional game* when referring to coalitional games with the TU interpretation.

In summary, the coalitional form of a game associates with each coalition S a single number v(S), which in some sense represents the total payoff that that coalition can get or may expect. In some contexts, v(S) fully characterizes the possibilities open to S; in others, it is an index that is indicative of S's strength.

(iii) *Solution concepts*. Given a game, what outcome may be expected? Most of game theory is, in one way or another, directed at this question. In the case of two-person zero-sum games, a clear answer is provided: the unique individually rational outcome. But in almost all other cases, there is no unique answer. There are different criteria, approaches, points of view, and they yield different answers.

A *solution concept* is a function (or correspondence) that associates outcomes, or sets of outcomes, with games. Usually an ‘outcome’ may be identified with the profile of payoffs that outcome yields to the players, though sometimes we may wish to think of it as a strategy profile.

Of course a solution concept is not just any such function or correspondence, but one with a specific rationale; for example, the strategic equilibrium and its variants for strategic form games, and the core, the von Neumann–Morgenstern stable sets, the Shapley value and the nucleolus for coalitional games. Each represents a different approach or point of view.

What will ‘really’ happen? Which solution concept is ‘right’? None of them; they are indicators, not predictions. Different solution concepts are like different indicators of an economy; different methods for calculating a price index; different maps (road, topo, political, geologic, etc., not to speak of scale, projection, etc.); different stock indices (Dow Jones, Standard and Poor's NYSE, etc., composite, industrials, utilities, etc.); different batting statistics (batting average, slugging average, RBI, hits, etc.); different kinds of information about rock climbs (arabic and roman difficulty ratings, route maps, verbal descriptions of the climb, etc.); accounts of the same event by different people or different media; different projections of the same three-dimensional object (as in architecture or engineering). They depict or illuminate the situation from different angles; each one stresses certain aspects at the expense of others.

Moreover, solution concepts necessarily leave out altogether some of the most vital information, namely that not entering the formal description of the game. When applied to a voting game, for example, no solution concept can take into account matters of custom, political ideology, or personal relations, since they don't enter the coalitional form. That does not make the solution useless. When planning a rock climb, you certainly want to take into account a whole lot of factors other than the physical characteristics of the rock, such as the season, the weather, your ability and condition, and with whom you are going. But you also do want to know about the ratings.

A good analogy is to distributions (probability, frequency, population, etc.). Like a game, a distribution contains a lot of information; one is overwhelmed by all the numbers. The median and the mean summarize the information in different ways; though other than by simply stating the definitions, it is not easy to say how. The definitions themselves do have a certain fairly clear intuitive content; more important, we gain a feeling for the relation between a distribution and its median and mean from experience, from working with various specific examples and classes of examples over the course of time.

The relationship of solution concepts to games is similar. Like the median and the mean, they in some sense summarize the large amount of information present in the formal description of a game. The definitions themselves have a certain fairly clear intuitive content, though they are not predictions of what will happen. Finally, the relations between a game and its core, value, stable sets, nucleolus, and so on is best revealed by seeing where these solution concepts lead in specific games and classes of games.

(iv) *Domination, the core and imputations*. Continuing to identify ‘outcome’ with ‘payoff profile’, we call an outcome y of a game *feasible* if the all-player set I can achieve it. An outcome x *dominates* y if there exists a coalition S that can achieve at least its part of x, and each of whose members prefers x to y; in that case we also say that S can *improve upon* y. The *core* of a game is the set of all feasible outcomes that are not dominated.

In a TU coalitional game *v*, feasibility of *x* means $x(I) \square v(I)$, and *x* dominating *y* via S means that $x(S) \square v(S)$ and $x^i > y^i$ for all *i* in *S*. The core of *v* is the set of all feasible *y* with $y(S) \square v(S)$ for all *S*.

At first, the core sounds quite compelling; why should the players be satisfied with an outcome that some coalition can improve upon? It becomes rather less compelling when one realizes that many perfectly ordinary games have empty cores, i.e. every feasible outcome can be improved upon. Indeed, this is so even in as simple a game as the 3-person voting game.

For a coalition S to improve upon an outcome, players in S must trust each other; they must have faith that their comrades inside S will not desert them to make a coalition with other players outside S. In a TU 3-person voting game, $y := (1/3, 1/3, 1/3)$ is dominated via {1, 2} by $x := (1/2, 1/2, 0)$. But 1 and 2 would be wise to view a suggested move from *y* to *x* with caution. What guarantee does 1 have that 2 will really stick with him and not accept offers from 3 to improve upon *x* with, say, (0, 2/3, 1/3)? For this he must depend on 2's good faith, and similarly 2 must depend on 1's.

There are two exceptions to this argument, two cases in which domination does not require mutual trust. One is when *S* consists of a single player. The other is when $S = I$, so that there is no one outside S to lure one's partners away.

The requirement that a feasible outcome *y* be undominated via one-person coalitions (*individual rationality*) and via the all-person coalition (*efficiency or Pareto optimality*) is thus quite compelling, much more so than that it be in the core. Such outcomes are called *imputations*. For TU coalitional games, individual rationality means that $y^i \square v(i)$ for all i (we do not distinguish between *i* and {*i*}), and efficiency means that $y(I) = v(I)$. The outcomes associated with most cooperative solution concepts are imputations; the imputations constitute the stage on which most of cooperative game theory is played out.

The notion of core does not appear explicitly in von Neumann and Morgenstern, but it is implicit in some of the discussions of stable sets there. In specific economic contexts, it is implicit in the work of Edgeworth (1881) and Ransmeier (1942). As a general solution concept in its own right, it was developed by Shapley and Gillies in the early Fifties. Early references include Luce and Raiffa (1957) and Gillies (1959).

(v) *Stable sets*. The discomfort with the definition of core expressed above may be stated more sharply as follows. Suppose we think of an outcome in the core as ‘stable’. Then we should not exclude an outcome *y* just because it is dominated by *some* other outcome *x*; we should demand that *x* itself be stable. If *x* is not itself stable, then the argument for excluding *y* is rather weak; proponents of *y* can argue with justice that replacing it with *x* would not lead to a more stable situation, so we may as well stay where we are. If the core were the set of all outcomes not dominated by any element of the core, there would be no difficulty; but this is not so.

Von Neumann and Morgenstern were thus led to the following definition: A set K of imputations is called *stable* if it is the set of all imputations not dominated by any element of K. This definition guarantees neither existence nor uniqueness. On the face of it, a game may have many stable sets, or it may have none. Most games do, in fact, have many stable sets; but the problem of existence was open for many years. It was solved by Lucas (1969), who constructed a ten-person, TU coalitional game without any stable set. Later, Lucas and Rabie (1982) constructed a fourteen-person TU coalitional game without any stable set and with an empty core to boot.

Much of the *Theory of Games* is devoted to exploring the stable sets of various classes of TU coalitional games, such as 3- and 4-person games, voting games, market games, compositions of games, and so on. (If *v* and *w* have disjoint player sets I and J, their *composition u* is given by $u(S) := v(S \cap I) + w(S \cap J)$.) During the 1950s many researchers carried forward with great vigour the work of investigating various classes of games and describing their stable sets. Since then work on stable sets has continued unabated, though it is no longer as much in the forefront of game-theoretic research as it was then. All in all, more than 200 articles have been published on stable sets, some 80 per cent of them since 1960. Much of the recent activity in this area has taken place in the Soviet Union.

It is impossible here even to begin to review this large and varied literature. But we do note one characteristic qualitative feature. By definition, a stable set is simply a set of imputations; there is nothing explicit in it about social structure. Yet the mathematical description of a given stable set can often best be understood in terms of an implicit social structure or form of organization of the players. Cartels, systematic discrimination, groups within groups, all kinds of subtle organizational forms spring to one's attention. These forms are endogenous, they are not imposed by definition, they emerge from the analysis. It is a mystery that just the stable set concept, and it only, is so closely allied with endogenous notions of social structure.

We adduce just one, comparatively simple example. The TU three-person voting game has a stable set consisting of the three imputations (1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2). The social structure implicit in this is that all three players will *not* compromise by dividing the payoff equally. Rather, one of the three 2-person coalitions will form and divide the payoff equally, with the remaining player being left ‘in the cold’. Because any of these three coalitions can form, competition drives them to divide the payoff equally, so that no player will prefer any one coalition to any other.

Another stable set is the interval $\{(\alpha, 1-\alpha, 0)\}$, where α ranges from 0 to 1. Here Player 3 is permanently excluded from all negotiations; he is ‘discriminated against’. Players 1 and 2 divide the payoff in some arbitrary way, not necessarily equally; this is because a coalition with 3 is out of the question, and so competition no longer constrains 1 and 2 in bargaining with each other.

(vi) *Transferable utility*. Though it no longer enjoys the centrality that it did up to about 1960, the assumption of transferable utility has played and continues to play a major role in the development of cooperative game theory. Some economists have questioned the appropriateness of the TU assumption, especially in connection with market models; it has been

castigated as excessively strong and unrealistic.

This situation is somewhat analogous to that of strictly competitive games, which as we pointed out above (*1930–1950*, vii), constitute a proving ground for developing and testing ideas that apply also to more general, non-strictly competitive games. The theory of NTU (non-transferable utility) coalitional games is now highly developed (see *1960–1970*, i), but it is an order of magnitude more complex than that of TU games. The TU theory is an excellent laboratory or model for working out ideas that are later applied to the more general NTU case.

Moreover, TU games are both conceptually and technically much closer to NTU games than strictly competitive games are to non-strictly competitive games. A very large part of the important issues arising in connection with non-strictly competitive games do not have any counterpart at all in strictly competitive games, and so simply cannot be addressed in that context. But by far the largest part of the issues and questions arising in the NTU theory do have counterparts in the TU theory; they can at least be addressed and dealt with there. Almost every major advance in the NTU theory – and many a minor advance as well – has had its way paved by a corresponding advance in the TU theory. Stable sets, core, value, and bargaining set were all defined first for TU games, then for NTU. The enormous literature on the core of a market and the equivalence between it and competitive equilibrium (c.e.) in large markets was started by Martin Shubik (1959a) in an article on TU markets. The relation between the value and c.e. in large markets was also explored first for the TU case (Shapley, 1964; Shapley and Shubik, 1969b; Aumann and Shapley, 1974; Hart, 1977a), then for NTU (Champsaur, 1975, but written and circulated circa 1970; Aumann, 1975; Mas-Colell, 1977; Hart, 1977b). The same holds for the bargaining set; first TU (Shapley, 1984), then NTU (Mas-Colell, 1988). The connection between balanced collections of coalitions and the non-emptiness of the core (*1960–1970*, viii) was studied first for TU (Bondareva, 1963; Shapley, 1967), then for NTU (Scarf, 1967; Billera, 1970b; Shapley 1973a); this development led to the whole subject of Scarf's algorithm for finding points in the core, which he and others later extended to algorithms for finding market equilibria and fixed points of mappings in general. Games arising from markets were first abstractly characterized in the TU case (Shapley and Shubik, 1969a), then in the NTU case (Billera and Bixby, 1973; Mas-Colell, 1975). Games with a continuum of players were conceived first in a TU application (Milnor and Shapley, 1979, but written and circulated in 1960), then NTU (Aumann, 1964). Strategic models of bargaining where time is of the essence were first treated for TU (Rubinstein, 1982), then NTU (Binmore, 1982). One could go on and on.

In each of these cases, the TU development led organically to the NTU development; it isn't just that the one came before the other. TU is to cooperative game theory what *Drosophila* is to genetics. Even if it had no direct economic interest at all, the study of TU coalitional games would be justified solely by their role as an outstandingly suggestive research tool.

(vii) *Single play*. Von Neumann and Morgenstern emphasize that their analysis refers to ‘one-shot’ games, games that are played just once, after which the players disperse, never to interact again. When this is not the case, one must view the whole situation – including expected future interactions of the same players – as a single larger game, and it, too, is to be played just once.

To some extent this doctrine appears unreasonable. If one were to take it literally, there would be only one game to analyse, namely the one whose players include all persons ever born and to be born. Every human being is linked to every other through some chain of interactions; no person or group is isolated from any other.

Savage (1954) has discussed this in the context of one-person decisions. In principle, he writes, one should ‘envisage every conceivable policy for the government of his whole life in its most minute details, and decide here and now on one policy. This is utterly ridiculous …’ (p. 16). He goes on to discuss the *small worlds* doctrine, ‘the practical necessity of confining attention to, or isolating, relatively simple situations …’ (p. 82).

To a large extent, this doctrine applies to interactive decisions too. But one must be careful, because here ‘large worlds’ have qualitative features totally absent from ‘small worlds’. We return to this below (*1950–1960*, ii, iii).

(viii) *Expected utility*. When randomized strategies are used in a strategic game, payoff must be replaced by expected payoff (*1910–1930*, iv). Since the game is played only once, the law of large numbers does not apply, so it is not clear why a player would be interested specifically in the mathematical expectation of his payoff.

There is no problem when for each player there are just two possible outcomes, which we may call ‘winning’ and ‘losing’, and denominate 1 and 0 respectively. (This involves no zero-sum assumption; e.g. all players could win simultaneously.) In that case the expected payoff is simply the probability of winning. Of course each player wants to maximize this probability, so in that case use of the expectation is justified.

Suppose now that the values of i's payoff function H^i are numbers between 0 and 1, representing win probabilities. Thus, for the ‘final’ outcome there are still only two possibilities; each pure strategy profile s induces a random process that generates a win for i with probability $H^i(s)$. Then the payoff expectation when randomized strategies are used still represents i's overall win probability.

Now in any game, each player has a most preferred and a least preferred outcome, which we take as a win and a loss. For each payoff h, there is some probability p such that i would as soon get h with certainty as winning with probability p and losing with probability $1 - p$. If we replace all the h's by the corresponding p's in the payoff matrix, then we are in the case of the previous paragraph, so use of the expected payoff is justified.

The probability p is a function of h, denoted $u^i(h)$, and called i's von Neumann–Morgenstern *utility*. Thus, to justify the use of expectations, each player's payoff must be replaced by its utility.

The key property of the function u^i is that if h and g are random payoffs, then i prefers h to g iff $\mathrm{E}u^i(h) > \mathrm{E}u^i(g)$, where E denotes expectation. This property continues to hold when

we replace u^i by a linear transform of the form $\alpha u^i + \beta$, where α and β are constants with $\alpha > 0$. All these transforms are also called utility functions for *i*, and any one of them may be used rather than u^i in the payoff matrix.

Recall that a strictly competitive game is defined as a two-person game in which if one outcome is preferred to another by one player, the preference is reversed for the other. Since randomized strategies are admitted, this condition applies also to 'mixed outcomes' (probability mixtures of pure outcomes). From this it may be seen that a two-person game is strictly competitive if and only if, for an appropriate choice of utility functions, the utility payoffs of the players sum to zero in each square of the matrix.

The case of TU coalitional games deserves particular attention. There is no problem if we assume fixed threats and continue to denominate the payoff in bread (see ii). But without fixed threats, the total amount of bread obtainable by a coalition S is a random variable depending on what players outside S do; since this is not denominated in utility, there is no justification for replacing it by its expectation. But if we do denominate payoffs in utility terms, then they cannot be directly transferred. The only way out of this quandary is to assume that the utility of bread is linear in the amount of bread (Aumann, 1960). We stress again that no such assumption is required in the fixed threat case.

(ix) *Applications*. The very name of the book, *Theory of Games and Economic Behavior*, indicates its underlying preoccupation with the applications. Von Neumann had already mentioned *Homo Economicus* in his 1928 paper, but there were no specific economic applications there.

The method of von Neumann and Morgenstern has become the archetype of later applications of game theory. One takes an economic problem, formulates it as a game, finds the game-theoretic solution, then translates the solution back into economic terms. This is to be distinguished from the more usual methodology of economics and other social sciences, where the building of a formal model and a solution concept, and the application of the solution concept to the model, are all rolled into one.

Among the applications extensively treated in the book is voting. A qualitative feature that emerges is that many different weight-quota configurations have the same coalitional form; [5; 2, 3, 4] is the same as [2; 1, 1, 1]. Though obvious to the sophisticated observer when pointed out, this is not widely recognized; most people think that the player with weight 4 is considerably stronger than the others (Vinacke and Arkoff, 1957). The Board of Supervisors of Nassau County operates by weighted voting; in 1964 there were six members, with weights of 31, 31, 28, 21, 2, 2, and a simple majority quota of 58 (Lucas, 1983, p. 188). Nobody realized that three members were totally without influence, that [58; 31, 31, 28, 21, 2, 2] = [2; 1, 1, 1, 0, 0, 0].

In a voting game, a winning coalition with no proper winning subsets is called *minimal winning* (mw). The game [*q; w*] is *homogeneous* if $w(S) = q$ for all minimal winning S; thus [3; 2, 1, 1, 1] is homogeneous, but [5; 2, 2, 2, 1, 1, 1] is not. A *decisive* voting game is one in which a coalition wins if and only if its complement loses; both the above games are decisive, but [3; 1, 1, 1, 1] is not. TU decisive homogeneous voting games have a stable set in which some mw coalition forms and divides the payoff in proportion to the weights of its members, leaving nothing for those outside. This is reminiscent of some parliamentary democracies, where parties in a coalition government get cabinet seats roughly in proportion to the seats they hold in parliament. But this fails to take into account that the actual number of seats held by a party may well be quite disproportional to its weight in a homogeneous representation of the game (when there is such a representation).

The book also considers issues of monopoly (or monopsony) and oligopoly. We have already pointed out that stable set theory concerns the endogenous emergence of social structure. In a market with one buyer (monopsonist) and two sellers (duopolists) where supply exceeds demand, the theory predicts that the duopolists will form a cartel to bargain with the monopsonist. The core, on the other hand, predicts cut-throat competition; the duopolists end up by selling their goods for nothing, with the entire consumer surplus going to the buyer.

This is a good place to point out a fundamental difference between the game-theoretic and other approaches to social science. The more conventional approaches take institutions as given, and ask where they lead. The game theoretic approach asks how the institutions came about, what led to them? Thus general equilibrium theory takes the idea of market prices for granted; it concerns itself with their existence and properties, calculating them, and so on. Game Theory asks, *why* are there market prices? How did they come about? Under what conditions will all traders trade at given prices?

Conventional economic theory has several approaches to oligopoly, including competition and cartelization. Starting with any particular one of these, it calculates what is implied in specific applications. Game theory proceeds differently. It starts with the physical description of the situation only, making no institutional or doctrinal assumptions, then applies a solution concept and sees where it leads.

In a sense, of course, the doctrine is built into the solution concept; as we have seen, the core implies competition, the stable set cartelization. It is not that game theory makes no assumptions, but that the assumptions are of a more general, fundamental nature. The difference is like that between deriving the motion of the planets from Kepler's laws or from Newton's laws. Like Kepler's laws, which apply to the planets only, oligopoly theory applies to oligopolistic markets only. Newton's laws apply to the planets and also to apples falling from trees; stable sets apply to markets and also to voting.

To be sure, conventional economics is also concerned with the genesis of institutions, but on an informal, verbal, ad-hoc level. In game theory, institutions like prices or cartels are outcomes of the formal analysis.

(x) Games with a *continuum of pure strategies* were first considered by Ville (1938), who proved the minimax theorem for them, using an appropriate continuity condition. To guarantee the minimax (security) level, one may need to use a continuum of pure strategies, each with probability zero. An example due to Kuhn (1952) shows that in general one cannot guarantee anything even close to minimax using strategies with finite support. Ville's theorem was extended in the fifties to strategic equilibrium in non-strictly competitive games.

(xi) *Computing* security levels, and strategies that will guarantee them, is highly non-trivial. The problem is equivalent to that of linear programming, and thus succumbed to the simplex method of George Dantzig (1951a, 1951b).

(xii) The major advance in relevant mathematical methods during this period was *Kakutani's fixed point theorem* (1941). An abstract expression of the existence of equilibrium, it is the vital active ingredient of countless proofs in economics and game theory. Also instrumental in later work were Lyapounov's theorem on the range of a vector measure (1940) and von Neumann's selection theorem (1949).

1950 – 1960

The 1950s were a period of excitement in game theory. The discipline had broken out of its cocoon, and was testing its wings. Giants walked the earth. At Princeton, John Nash laid the groundwork for the general non-cooperative theory, and for cooperative bargaining theory; Lloyd Shapley defined the value for coalitional games, initiated the theory of stochastic games, co-invented the core with D.B. Gillies, and, together with John Milnor, developed the first game models with continua of players; Harold Kuhn worked on behaviour strategies and perfect recall; Al Tucker discovered the prisoner's dilemma; the Office of Naval Research was unstinting in its support. Three Game Theory conferences were held at Princeton, with the active participation of von Neumann and Morgenstern themselves. Princeton University Press published the four classic volumes of *Contributions to the Theory of Games*. The Rand Corporation, for many years to be a major centre of game theoretic research, had just opened its doors in Santa Monica. R. Luce and H. Raiffa (1957) published their enormously influential *Games and Decisions*. Near the end of the decade came the first studies of repeated games.

The major applications at the beginning of the decade were to tactical military problems: defense from missiles, Colonel Blotto, fighter-fighter duels, etc. Later the emphasis shifted to deterrence and cold war strategy, with contributions by political scientists like Kahn, Kissinger, and Schelling. In 1954, Shapley and Shubik published their seminal paper on the value of a voting game as an index of power. And in 1959 came Shubik's spectacular rediscovery of the core of a market in the writings of F.Y. Edgeworth (1881). From that time on, economics has remained by far the largest area of application of game theory.

(i) An *equilibrium* (Nash, 1951) of a strategic game is a (pure or mixed) strategy profile in which each player's strategy maximizes his payoff given that the others are using their strategies. See the entry on Nash equilibrium, refinements of.

Strategic equilibrium is without doubt the single game theoretic solution concept that is most frequently applied in economics. Economic applications include oligopoly, entry and exit, market equilibrium, search, location, bargaining, product quality, auctions, insurance, principal-agent, higher education, discrimination, public goods, what have you. On the political front, applications include voting, arms control, and inspection, as well as most international political models (deterrence, etc.) Biological applications of game theory all deal with forms of strategic equilibrium; they suggest a simple interpretation of equilibrium quite different from the usual overt rationalism (see *1970–1986*, i). We cannot even begin to survey all this literature here.

(ii) *Stochastic and other dynamic games*. Games played in stages, with some kind of stationary time structure, are called *dynamic*. They include stochastic games, repeated games with or without complete information, games of survival (Milnor and Shapley, 1957; Luce and Raiffa, 1957; Shubik, 1959) or ruin (Rosenthal and Rubinstein, 1984), recursive games (Everett, 1957), games with varying opponents (Rosenthal, 1979), and similar models.

This kind of model addresses the concerns we expressed above (*1930–1950*, vii) about the single play assumption. The present can only be understood in the context of the past and the future: ‘Know whence you came and where you are going’ (Ethics of the Fathers III:1). Physically, current actions affect not only current payoff but also opportunities and payoffs in the future. Psychologically, too, we learn: past experience affects our current expectations of what others will do, and therefore our own actions. We also teach: our current actions affect others' future expectations, and therefore their future actions.

Two dynamic models – stochastic and repeated games – have been especially ‘successful’. *Stochastic* games address the physical point, that current actions affect future opportunities. A strategic game is played at each stage; the profile of strategies determines both the payoff at that stage and the game to be played at the next stage (or a probability distribution over such games). In the strictly competitive case, with future payoff discounted at a fixed rate, Shapley (1953a) showed that stochastic games are determined; also, that they have optimal strategies that are stationary, in the sense that they depend only on the game being played (not on the history or even on the date). Bewley and Kohlberg (1976) showed that as the discount rate tends to 0 the value tends to a limit; this limit is the same as the limit, as $k \to \infty$, of the values of the k-stage games, in each of which the payoff is the mean payoff for the k stages. Mertens and Neyman (1981) showed that the value exists also in the undiscounted infinite stage game, when payoff is defined by the Cesaro limit (limit, as $k \to \infty$, of the average payoff in the first k stages). For an understanding of some of the intuitive issues in this work, see Blackwell and Ferguson (1968), which was extremely influential in the modern development of stochastic games.

The methods of Shapley, and of Bewley and Kohlberg, can be used to show that non-strictly competitive stochastic games with fixed discounts have equilibria in stationary strategies, and that when the discount tends to 0, these equilibria converge to a limit (Mertens, 1982). But unlike in the strictly competitive case, the payoff to this limit need not correspond to an equilibrium of the undiscounted game (Sorin, 1986b). It is not known whether undiscounted non-strictly competitive stochastic games need at all have strategic equilibria.

(iii) *Repeated* games model the psychological, informational side of ongoing relationships. Phenomena like cooperation, altruism, trust, punishment, and revenge are predicted by the

theory. These may be called ‘subjective informational’ phenomena, since what is at issue is information about the behaviour of the players. Repeated games of incomplete information (*1960–1970*, ii) also predict ‘objective informational’ phenomena such as secrecy, and signalling of substantive information. Both kinds of informational issue are quite different from the ‘physical’ issues addressed by stochastic games.

Given a strategic game *G*, consider the game G^{∞} each play of which consists of an infinite sequence of repetitions of *G*. At each stage, all players know the actions taken by all players at all previous stages. The payoff in G^{∞} is some kind of average of the stage payoffs; we will not worry about exact definitions here.

The reader is referred to the entry on repeated games. Here we state only one basic result, known as the *Folk Theorem*. Call an outcome (payoff profile) *x feasible* in *G* if it is achievable by the all-player set when using a correlated randomizing device; i.e. is in the convex hull of the ‘pure’ outcomes. Call it *strongly individually rational* if no player *i* can be prevented from achieving x^i by the other players, when they are randomizing independently; i.e. if $x^i \square \min \max H^i(s)$, where the max is over *i*'s strategies, and the min is over $(n-1)$-tuples of mixed strategies of the others. The Folk Theorem then says that the equilibrium outcomes in the repetition G^{∞} *coincide with the feasible and strongly individually rational outcomes in the one-shot game* G.

The authorship of the Folk Theorem, which surfaced in the late Fifties, is obscure. Intuitively, the feasible and strongly individually rational outcomes are the outcomes that could arise in cooperative play. Thus the Folk Theorem points to a strong relationship between repeated and cooperative games. Repetition is a kind of enforcement mechanism; agreements are enforced by ‘punishing’ deviators in subsequent stages.

(iv) The *Prisoner's Dilemma* is a two-person non-zero sum strategic game with payoff matrix as depicted in Figure 1. Attributed to A.W. Tucker, it has deservedly attracted enormous attention; it is said that in the social psychology literature alone, over a thousand papers have been devoted to it.

Figure 1

	f	g
f	3,3	0,4
g	4,0	1,1

One may think of the game as follows: Each player decides whether he will receive $1000 or the other will receive $3000. The decisions are simultaneous and independent, though the players may consult with each other before deciding.

The point is that ordinary rationality leads each player to choose the $1000 for himself, since he is thereby better off *no matter what the other player does*. But the two players thereby get only $1000 each, whereas they could have gotten $3000 each if both had been ‘friendly’ rather than ‘greedy’.

The universal fascination with this game is due to its representing, in very stark and transparent form, the bitter fact that when individuals act for their own benefit, the result may well be disaster for all. This principle has dozens of applications, great and small, in everyday life. *People who fail to cooperate for their own mutual benefit are not necessarily foolish or irrational*; they may be acting perfectly rationally. The sooner we accept this, the sooner we can take steps to design the terms of social intercourse so as to encourage cooperation. One such step, of very wide applicability, is to make available a mechanism for the enforcement of voluntary agreements. ‘Pray for the welfare of government, without whose authority, man would swallow man alive’ (Ethics of the Fathers III:2). The availability of the mechanism is itself sufficient; once it is there, the players are naturally motivated to use it. If they can make an *enforceable* agreement yielding (3, 3), they would indeed be foolish to end up with (1, 1). It is this that motivates the definition of a cooperative game (*1930–1950*, i).

The above discussion implies that (*g*, *g*) is the unique strategic equilibrium of the prisoner's dilemma. It may also be shown that in any finite repetition of the game, all strategic equilibria lead to a constant stream of ‘greedy’ choices by each player; but this is a subtler matter than the simple domination argument used for the one-shot case. In the infinite repetition, the Folk Theorem (iii) shows that (3, 3) is an equilibrium outcome; and indeed, there are equilibria that lead to a constant stream of ‘friendly’ choices by each player. The same holds if we discount future payoff in the repeated game, as long as the discount rate is not too large (Sorin, 1986a).

R. Axelrod (1984) has carried out an experimental study of the repeated prisoner's dilemma. Experts were asked to write computer programmes for playing the game, which were matched against each other in a ‘tournament’. At each stage, the game ended with a fixed (small) probability; this is like discounting. The most successful program in the tournament turned out to be a ‘cooperative’ one: Matched against itself, it yields a constant stream of ‘friendly’ choices; matched against others, it ‘punishes’ greedy choices. The results of this experiment thus fit in well with received theoretical doctrine.

The design of this experiment is noteworthy because it avoids the pitfalls so often found in game experiments: lack of sufficient motivation and understanding. The experts chosen by Axelrod understood the game as well as anybody. Motivation was provided by the investment of their time, which was much more considerable than that of the average subject, and by the glory of a possible win over distinguished colleagues. Using computer programmes for strategies presaged important later developments (*1970–1986*, iv).

Much that is fallacious has been written on the one-shot prisoner's dilemma. It has been said that for the reasoning to work, pre-play communication between the players must be forbidden. This is incorrect. The players can communicate until they are blue in the face, and agree solemnly on (*f*, *f*); when faced with the actual decision, rational players will still choose *g*. It has been said that the argument depends on the notion of strategic equilibrium, which is open to discussion. This too is incorrect; the argument depends only on strong domination, i.e. on the simple proposition that people always prefer to get another $1000. ‘Resolutions’ of the ‘paradox’ have been put forward, suggesting that rational players will play *f* after all; that my choosing *f* has some kind of ‘mirror’ effect that makes you choose it also. Worse than just nonsense, this is actually vicious, since it suggests that the prisoner's dilemma does not represent a real social problem that must be dealt with.

Finally, it has been said that the experimental evidence – Axelrod's and that of others – contradicts theory. This too is incorrect, since most of the experimental evidence relates to repeated games, where the friendly outcome is perfectly consonant with theory; and what evidence there is in one-shot games does point to a preponderance of ‘greedy’ choices. It is true that in long finite repetitions, where the only equilibria are greedy, most experiments nevertheless point to the friendly outcome; but fixed finite repetitions are somewhat artificial, and besides, this finding, too, can be explained by theory (Neyman, 1985; see *1970–1986*, iv).

(v) We turn now to cooperative issues. A model of fundamental importance is the *bargaining problem* of Nash (1950). Formally, it is defined as a convex set C in the Euclidean plane, containing the origin in its interior. Intuitively, two players bargain; they may reach any agreement whose payoff profile is in C; if they disagree, they get nothing. Nash listed four *axioms* – conditions that a reasonable compromise solution might be expected to satisfy – such as symmetry and efficiency. He then showed that there is one and only one solution satisfying them, namely the point x in the non-negative part of C that maximizes the product x^1x^2. An appealing economic interpretation of this solution was given by Harsanyi (1956).

By varying the axioms, other authors have obtained different solutions to the bargaining problem, notably Kalai-Smorodinski (1975) and Maschler-Perles (1981). Like Nash's solution, each of these is characterized by a formula with an intuitively appealing interpretation.

Following work of A. Rubinstein (1982), K. Binmore (1982) constructed an explicit bargaining model which, when analyzed as a non-cooperative strategic game, leads to Nash's solution of the bargaining problem. This is an instance of a successful application of the 'Nash program' (see *1930–1950*, vi). Similar constructions have been made for other solutions of the bargaining problem.

An interesting qualitative feature of the Nash solution is that it is very sensitive to risk aversion. A risk loving or risk neutral bargainer will get a better deal than a risk averse one; this is so even when there are no overt elements of risk in the situation, nothing random. The very willingness to take risks confers an advantage, though in the end no risks are actually taken.

Suppose, for example, that two people may divide $600 in any way they wish; if they fail to agree, neither gets anything. Let their utility functions be $u^1(\$x) = x$ and $u^2(\$x) = \sqrt{x}$, so that 1 is risk neutral, 2 risk averse. Denominating the payoffs in utilities rather than dollars, we find that the Nash solution corresponds to a dollar split of $400-$200 in favour of the risk neutral bargainer.

This corresponds well with our intuitions. A fearful, risk averse person will not bargain well. Though there are no overt elements of risk, no random elements in the problem description, the bargaining itself constitutes a risk. A risk averse person is willing to pay, in terms of a less favourable settlement, to avoid the risk of the other side's being adamant, walking away, and so on.

(vi) The *value* (Shapley, 1953b) is a solution concept that associates with each coalitional game v a unique outcome ϕv. Fully characterized by a set of axioms, it may be thought of as a reasonable compromise or arbitrated outcome, given the power of the players. Best, perhaps, is to think of it simply as an index of power, or what comes to the same thing, of social productivity (see Shapley value).

It may be shown that Player i's value is given by

$$\varphi^i v = (1/n!)\sum v^i(S^i_R),$$

where Σ ranges over all $n!$ orders on the set I of all players, S^i_R is the set of players up to and including i in the order R, and $v^i(S)$ is the *contribution* $v(S) - v(S\backslash i)$ of i to the coalition S; note that this implies linearity of ϕv in v. In words, $\phi^i v$ is i's mean contribution when the players are ordered at random; this suggests the social productivity interpretation, an interpretation that is reinforced by the following remarkable theorem (Young 1985): Let ψ be a mapping from games v to efficient outcomes ψv, that is symmetric among the players in the appropriate sense. Suppose $\psi^i v$ depends only on the 2^{n-1} contributions $v^i(S)$, and monotonically so. Then ψ must be the value ϕ. In brief, if it depends on the contributions only, it's got to be the value, even though we don't assume linearity to start with.

An intuitive feel for the value may be gained from examples. The value of the 3-person voting game is (1/3, 1/3, 1/3), as is suggested by symmetry. This is not in the core, because {1, 2} can improve upon it. But so can {1, 3} and {2, 3}; starting from (1/3, 1/3, 1/3), the players might be well advised to leave things as they are (see 1930–1950, iv). Differently viewed, the symmetric stable set predicts one of the three outcomes (1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2). Before the beginning of bargaining, each player may figure that his chances of getting into a ruling coalition are 2/3, and conditional on this, his payoff is 1/2. Thus his 'expected outcome' is the value, though in itself, this outcome has no stability.

In the homogenous weighted voting game [3; 2, 1, 1, 1], the value is (1/2, 1/6, 1/6, 1/6); the large player gets a disproportionate share, which accords with intuition: '1'union fait la force.'

Turning to games of economic interest, we model the market with two sellers and one buyer discussed above (*1930–1950*, ix) by the TU weighted voting game [3; 2, 1, 1]. The core consists of the unique point (1, 0, 0), which means that the sellers must give their merchandise, for nothing, to the buyer. While this has clear economic meaning—cutthroat competition – it does not seem very reasonable as a compromise or an index of power. After all, the sellers do contribute something; without them, the buyer could get nothing. If one could be sure that the sellers will form a cartel to bargain with the buyer, a reasonable compromise would be (1/2, 1/4, 1/4). In fact, the value is (2/3, 1/6, 1/6), representing something between the cartel solution and the competitive one; a cartel is possible, but is not a certainty.

Consider next a market in two perfectly divisible and completely complementary goods, which we may call right and left gloves. There are four players; initially 1 and 2 hold one and two left gloves respectively, 3 and 4 hold one right glove each. In coalitional form, $v(1234) = v(234) = 2$, $v(ij) = v(12j) = v(134) = 1$, $v(S) = 0$ otherwise, where $i = 1, 2$, and $j = 3, 4$. The core consists of (0, 0, 1, 1) only; that is, the owners of the left gloves must simply give away their merchandise, for nothing. This in itself seems strange enough. It becomes even stranger when one realizes that Player 2 could make the situation entirely symmetric (as between 1, 2 and 3, 4) simply by burning one glove, an action that he can take alone, without consulting anybody.

The value can never suffer from this kind of pathological breakdown in monotonicity. Here $\varphi v = (1/4, 7/12, 7/12, 7/12;)$, which nicely reflects the features of the situation. There *is* an oversupply of left gloves, and 3 and 4 do benefit from it. Also 2 benefits from it; he always has the option of nullifying it, but he can also use it (when he has an

opportunity to strike a deal with both 3 and 4). The brunt of the oversupply is thus born by 1 who, unlike 2, cannot take measures to correct it.

Finally, consider a market with 2,000,001 players, 1,000,000 holding one right glove each, and 1,000,001 holding one left glove each. Again, the core stipulates that the holders of the left gloves must all give away their merchandise, for nothing. True, there *is* a slight oversupply of left gloves; but one would hardly have imagined so drastic an effect from one single glove out of millions. The value, too, takes the oversupply into account, but not in such an extreme form; altogether, the left-glove holders get about 499,557 pairs, the right about 500,443 (Shapley and Shubik, 1969b). This is much more reasonable, though the effect is still surprisingly large: The short side gains an advantage that amounts to almost a thousand pairs.

The value has many different characterizations, all of them intuitively meaningful and interesting. We have already mentioned Shapley's original axioms, the value formula, and Young's characterization. To them must be added Harsanyi's (1959) dividend characterization, Owen's (1972) fuzzy coalition formula, Myerson's (1977) graph approach, Dubey's (1980) diagonal formula, the potential of Hart and Mas-Colell (1986), the reduced game axiomatization by the same authors, and Roth's (1977) formalization of Shapley's (1953b) idea that the value represents the utility to the players of playing a game. Moreover, because of its mathematical tractability, the value lends itself to a far greater range of applications than any other cooperative solution concept. And in terms of general theorems and characterizations for wide classes of games and economies, the value has a greater range than *any* other solution concept, bar none.

Previously (*1930–1950*, iii), we compared solution concepts of games to indicators of distributions, like mean and median. In fact the value is in many ways analogous to the mean, whereas the median corresponds to something like the core, or to core-like concepts such as the nucleolus (*1960–1970*, iv). Like the core, the median has an intuitively transparent and compelling definition (the point that cuts the distribution exactly in half), but lacks an algebraically neat formula; and like the value, the mean has a neat formula whose intuitive significance is not entirely transparent (thought through much experience from childhood on, many people have acquired an intuitive feel for it). Like the value, the mean is linear in its data; the core, nucleolus, and median are not. Both the mean and the value are very sensitive to their data: change one datum by a little, and the mean (or value) will respond in the appropriate direction; neither the median nor the core is sensitive in this way: one can change the data in wide ranges without affecting the median (or core) at all. On the other hand, the median can suddenly jump because of a moderate change in just one datum; thus the median of 1,000,001 zeros and 1,000,000 ones is 0, but jumps to 1 if we change just one datum from 0 to 1. We have already seen that the core may behave similarly, but the mean and the value cannot. Both the mean and the value are mathematically very tractable, resulting in a wide range of applications, both theoretical and practical; the median and core are less tractable, resulting in a narrower (though still considerable) range of applications.

The first extensive applications of the value were to various voting games (Shapley and Shubik, 1954). The key observation in this seminal paper was that the value of a player equals his probability of *pivoting* – turning a coalition from losing to winning – when the players are ordered at random. From this there has grown a very large literature on voting games. Other important classes of applications are to market games (*1960–1970*, v) and political-economic games (e.g. Aumann and Kurz, 1977; Neyman, 1985b).

(vii) *Axiomatics*. The Shapley value and Nash's solution to the bargaining problem are examples of the axiomatic approach. Rather than defining a solution concept directly, one writes down a set of conditions for it to satisfy, then sees where they lead. In many contexts, even a relatively small set of fairly reasonable conditions turn out to be self-contradictory; there is no concept satisfying all of them. The most famous instance of this is Arrow's (1951) impossibility theorem for social welfare functions, which is one of the earliest applications of axiomatics in the social sciences.

It is not easy to pin down precisely what is meant by ‘the axiomatic method’. Sometimes the term is used for any formal deductive system, with undefined terms, assumptions, and conclusions. As understood today, all of game theory and mathematical economics fits that definition. More narrowly construed, an axiom system is a small set of individually transparent conditions, set in a fairly general and abstract framework, which when taken together have far-reaching implications. Examples are Euclid's axioms for geometry, the Zermelo–Fraenkel axioms for set theory, the conditions on multiplication that define a group, the conditions on open sets that define a topological space, and the conditions on preferences that define utility and/or subjective probability.

Game theoretic solution concepts often have both direct and axiomatic characterizations. The direct definition applies to each game separately, whereas most axioms deal with relationships between games. Thus the formula for the Shapley value ϕv enables one to calculate it without referring to any game other than v. But the axioms for ϕ concern relationships between games; they say that if the values of certain games are so and so, then the values of certain other, related games must be such and such. For example, the additivity axiom is $\varphi(v+w) = \varphi v + \varphi w$. This is analogous to direct vs. axiomatic approaches to integration. Direct approaches such as limit of sum work on a single function; axiomatic approaches characterize the integral as a linear operator on a *space* of functions. (Harking back to the discussion at (vi), we note that the axioms for the value are quite similar to those for the integral, which in turn is closely related to the mean of a distribution.)

Shapley's value and the solutions to the bargaining problem due to Nash (1950), Kalai and Smorodinsky (1975) and Maschler and Perles (1981) were originally conceived axiomatically, with the direct characterization coming afterwards. In other cases the process was reversed; for example, the nucleolus, NTU Shapley value, and NTU Harsanyi value were all axiomatized only years after their original direct definition (see *1960–1970)*. Recently the core, too, has been axiomatized (Peleg, 1985, 1986).

Since axiomatizations concern relations between different games, one may ask why the players of a given game should be concerned with other games, which they are not playing. This has several answers. Viewed as an indicator, a solution of a game doesn't tell us much unless it stands in some kind of coherent relationship to the solutions of other games. The ratings for a rock climb tell you something if you have climbed other rocks whose ratings you know; topographic maps enable you to take in a situation at a glance if you have used them before, in different areas. If we view a solution as an arbitrated or imposed outcome, it is natural to expect some measure of consistency from an arbitrator or judge. Indeed,

much of the law is based on precedent, which means relating the solution of the given ‘game’ to those of others with known solutions. Even when viewing a solution concept as a norm of actual behaviour, the very word ‘norm’ implies that we are thinking of a function on classes of games rather than of a single game; outcomes are largely based on mutual expectations, which are determined by previous experience with other games, by ‘norms’.

Axiomatizations serve a number of useful purposes. First, like any other alternative characterization, they shed additional light on a concept, enable us to ‘understand’ it better. Second, they underscore and clarify important similarities between concepts, as well as differences between them. One example of this is the remarkable ‘reduced game property’ or ‘consistency principle’, which is associated in various different forms with just about every solution concept, and plays a key role in many of the axiomatizations (see *1970–1986*, vi). Another example consists of the axiomatizations of the Shapley and Harsanyi NTU values. Here the axioms are exact analogues, except that in the Shapley case they refer to payoff profiles, and in the Harsanyi case to 2^n-tuples of payoff profiles, one for each of the 2^n coalitions (Hart, 1985a). This underscores the basic difference in outlook between those two concepts: The Shapley value assumes that the all-player coalition eventually forms, the intermediate coalitions being important only for bargaining chips and threats, whereas the Harsanyi value takes into account a real possibility of the intermediate coalitions actually forming.

Last, an important function of axiomatics relates to ‘counter-intuitive examples’, in which a solution concept yields outcomes that seem bizarre; e.g. the cores of some of the games discussed above in (vi). Most axioms appearing in axiomatizations do seem reasonable on the face of it, and many of them are in fact quite compelling. The fact that a relatively small selection of such axioms is often categoric (determines a unique solution concept), and that different such selections yield different answers, implies that all together, these reasonable-sounding axioms are contradictory. This, in turn, implies that any one solution concept will necessarily violate at least some of the axioms that are associated with other solution concepts; thus if the axioms are meant to represent intuition, counter-intuitive examples are inevitable.

In brief, axiomatics underscores the fact that a ‘perfect’ solution concept is an unattainable goal, a *fata morgana*; there is something ‘wrong’, some quirk with every one. Any given kind of counterintuitive example can be eliminated by an appropriate choice of solution concept, but only at the cost of another quirk turning up. Different solution concepts can therefore be thought of as results of choosing not only which properties one likes, but also which examples one wishes to avoid.

1960 – 1970

The Sixties were a decade of growth. Extensions such as games of incomplete information and NTU coalitional games made the theory much more widely applicable. The fundamental underlying concept of common knowledge was formulated and clarified. Core theory was extensively developed and applied to market economies; the bargaining set and related concepts such as the nucleolus were defined and investigated; games with many players were studied in depth. The discipline expanded geographically, outgrowing the confines of Princeton and Rand; important centres of research were established in Israel, Germany, Belgium and the Soviet Union. Perhaps most important was the forging of a strong, lasting relationship with mathematical economics and economic theory.

(i) *NTU coalitional games and NTU value*. Properly interpreted, the coalitional form (*1930–1950*, ii) applies both to TU and to NTU games; nevertheless, for many NTU applications one would like to describe the opportunities available to each coalition more faithfully than can be done with a single number. Accordingly, define a game in *NTU coalitional form* as a function that associates with each coalition S a set V(S) of S-tuples of real numbers (functions from S to $\mathbb{R}$). Intuitively, V(S) represents the set of payoff S-tuples that S can achieve. For example, in an exchange economy, V(S) is the set of utility S-tuples that S can achieve when its members trade among themselves only, without recourse to agents outside of S. Another example of an NTU coalitional game is Nash's bargaining problem (1950–1960, iii), where one can take $V(\{1, 2\}) = C$, $V(1) = \{0\}$, $V(2) = \{0\}$.

The definitions of stable set and core extend straightforwardly to NTU coalitional games, and these solution concepts were among the first to be investigated in that context (Aumann and Peleg, 1960; Peleg, 1963a; Aumann, 1961). The first definitions of NTU value were proposed by Harsanyi (1959, 1963), but they proved difficult to apply. Building on Harsanyi's work, Shapley (1969) defined a value for NTU games that has proved widely applicable and intuitively appealing.

For each profile λ of non-negative numbers and each outcome *x*, define the *weighted outcome* λx by $(\lambda x)^i = \lambda^i x^i$. Let v_λ (S) be the maximum total weight that the coalition S can achieve,

$$v_\lambda(S) := \max\left\{\sum_{i\in S}\lambda^i x^i,\ x \in V(S)\right\}.$$

Call an outcome *x* an *NTU value* of V if $x \in V(N)$ and there exists a weight profile λ with $\lambda x = \varphi v_\lambda$; in words, if *x* is feasible and corresponds to the value of one of the coalitional games v_λ.

Intuitively, $v_\lambda(S)$ is a numerical measure of S's total worth and hence $\phi^i v_\lambda$ measures *i*'s social productivity. The weights λ^i are chosen so that the resulting value is feasible; an infeasible result would indicate that some people are overrated (or underrated), much like an imbalance between supply and demand indicates that some goods are overpriced (or underpriced).

The NTU value of a game need not be unique. This may at first sound strange, since unlike stability concepts such as the core, one might expect an 'index of social productivity' to be unique. But perhaps it is not so strange when one reflects that even a person's net worth depends on the prevailing (equilibrium) prices, which are not uniquely determined by the exogenous description of the economy.

The Shapley NTU value has been used in a very wide variety of economic and political-economic applications. To cite just one example, the Nash bargaining problem has a unique NTU value, which coincides with Nash's solution. For a partial bibliography of applications, see the references of Aumann (1985).

We have discussed the historical importance of TU as pointing the way for NTU results (*1930–1950*, vi). There is one piquant case in the reverse direction. Just as positive results are easier to obtain for TU, negative results are easier for NTU. Non-existence of stable sets was first discovered in NTU games (Stearns 1967), and this eventually led to Lucas's famous example (1969) of non-existence for TU.

(ii) *Incomplete information*. In 1957, Luce and Raiffa wrote that a fundamental assumption of game theory is that 'each player … is fully aware of the rules of the game and the utility functions of each of the players … this is a serious idealization which only rarely is met in actual situations' (p. 49). To deal with this problem, John Harsanyi (1967) constructed the theory of games of incomplete information (sometimes called differential or asymmetric information). This major conceptual breakthrough laid the theoretical groundwork for the great blooming of information economics that got under way soon thereafter, and that has become one of the major themes of modern economics and game theory.

For simplicity, we confine attention to strategic form games in which each player has a fixed, known set of strategies, and the only uncertainty is about the utility functions of the other players; these assumptions are removable. Bayesian rationality in the tradition of Savage (1954) dictates that all uncertainty can be made explicit; in particular, each player has a personal probability distribution on the possible utility (payoff) functions of the other player. But these distributions are not sufficient to describe the situation. It is not enough to specify what each player thinks about the other's payoffs; one must also know what he thinks they think about his (and each others') payoffs, what he thinks they think he thinks about their payoffs, and so on. This complicated infinite regress would appear to make useful analysis very difficult.

To cut this Gordian knot, Harsanyi postulated that each player may be one of several *types*, where a type determines both a player's own utility function and his personal probability distribution on the types of the other players. Each player is postulated to know his own type only. This enables him to calculate what he thinks the other players' types – and therefore their utilities – are. Moreover, his personal distribution on their types also enables him to calculate what he thinks they think about his type, and therefore about his utility. The reasoning extends indefinitely, and yields the infinite regress discussed above *as an outcome*.

Intuitively, one may think of a player's type as a possible state of mind, which would determine his utility as well as his distribution over others' states of mind. One need not assume that the number of states of mind (types) is finite; the theory works as well for, say, a continuum of types. But even with just two players and two types for each player, one gets a non-trivial infinite string of beliefs about utilities, beliefs about beliefs, and so on.

A model of this kind – with players, strategies, types, utilities, and personal probability distributions – is called an *I-game* (incomplete information game). A *strategic equilibrium* in an I-game consists of a strategy for each *type* of each player, which maximizes that type's expected payoff given the strategies of the other players' types.

Harsanyi's formulation of I-games is primarily a device for thinking about incomplete information in an orderly fashion, bringing that wild, bucking infinite regress under conceptual control. An (incomplete) analogy is to the strategic form of a game, a conceptual simplification without which it is unlikely that game theory would have gotten very far. Practically speaking, the strategic form of a particular game such as chess is totally unmanageable, one can't even begin to write it down. The advantage of the strategic form is that it is a comparatively simple formulation, mathematically much simpler than the extensive form; it enables one to formulate and calculate examples, which suggest principles that can be formulated and proved as general theorems. All this would be much more difficult – probably unachievable – with the extensive form; one would be unable to see the forest for the trees. A similar relationship holds between Harsanyi's I-game formulation and direct formulations in terms of beliefs about beliefs. (Compare the discussion of perspective made in connection with the coalitional form (*1930–1950*, i). That situation is somewhat different, though, since in going to the coalitional form, substantive information is lost. Harsanyi's formulation of I-games loses no information; it is a more abstract and simple – and hence transparent and workable – formulation of the same data as would be contained in an explicit description of the infinite regress.)

Harsanyi called an I-game *consistent* if all the personal probability distributions of all the types are derivable as posteriors from a single prior distribution *p* on all n-tuples of types. Most applications of the theory have assumed consistency. A consistent I-game is closely related to the ordinary strategic game (*C-game*) obtained from it by allowing 'nature' to choose an n-tuple of types at random according to the distribution *p*, then informing each player of his type, and then playing the I-game as before. In particular, the strategic equilibria of a consistent I-game are essentially the same as the strategic equilibria of the related C-game. In the cooperative theory, however, an I-game is rather different from the related C-game, since binding agreements can only be made after the players know their types. Bargaining and other cooperative models have been treated in the incomplete information context by Harsanyi and Selten (1972), Wilson (1978), Myerson (1979, 1984), and others.

In a repeated game of incomplete information, the same game is played again and again, but the players do not have full information about it; for example, they may not know the others' utility functions. The actions of the players may implicitly reveal private information, e.g. about preferences; this may or may not be advantageous for them. We have seen

(*1950–1960*, iii) that repetition may be viewed as a paradigm for cooperation. Strategic equilibria of repeated games of incomplete information may be interpreted as a subtle bargaining process, in which the players gradually reach wider and wider agreement, developing trust for each other while slowly revealing more and more information (Hart, 1985b).

(iii) *Common knowledge*. Luce and Raiffa, in the statement quoted at the beginning of (ii), missed a subtle but important point. It is not enough that each player be fully aware of the rules of the game and the utility functions of the players. Each player must also be aware of this fact, i.e. of the awareness of all the players; moreover, each player must be aware that each player is aware that each player is aware, and so on ad infinitum. In brief, the awareness of the description of the game by all players must be a part of the description itself. There is evidence that game theorists had been vaguely cognizant of the need for some such requirement ever since the late Fifties or early Sixties; but the first to give a clear, sharp formulation was the philosopher D.K. Lewis (1969). Lewis defined an event as *common knowledge* among a set of agents if all know it, all know that all know it, and so on ad infinitum.

The common knowledge assumption underlies all of game theory and much of economic theory. Whatever be the model under discussion, whether complete or incomplete information, consistent or inconsistent, repeated or one-shot, cooperative or non-cooperative, the model itself must be assumed common knowledge; otherwise the model is insufficiently specified, and the analysis incoherent.

(iv) *Bargaining set, kernel, nucleolus*. The core excludes the unique symmetric outcome (1/3, 1/3, 1/3) of the three-person voting game, because any two-person coalition can improve upon it. Stable sets (*1930–1950*, v) may be seen as a way of expressing our intuitive discomfort with this exclusion. Another way is the bargaining set (Davis and Maschler, 1965). If, say, I suggests (1/2, 1/2,•0) to replace (1/3, 1/3, 1/3), then 3 can suggest to 2 that he is as good a partner as 1; indeed, 3 can even offer 2/3 to 2, still leaving himself with the 1/3 he was originally assigned. Formally, if we call (1/2, 1/2, 0) an *objection* to (1/3, 1/3, 1/3), then (0, 2/3, 1/3) is a *counter objection*, since it yields to 3 at least as much as he was originally assigned, and yields to 3's partners in the counter-objection at least as much as they were assigned either originally or in the objection. In brief, the counter-objecting player tells the objecting one, ‘I can maintain my level of payoff and that of my partners, while matching your offers to players we both need.’ An imputation is in the core if there is no objection to it. It is in the *bargaining set* if there is no *justified* objection to it, i.e. one that has no counter-objection.

Like the stable sets, the bargaining set includes the core (dominating and objecting are essentially the same). Unlike the core and the set of stable sets, the bargaining set is for TU games never empty (Peleg, 1967). For NTU it may be empty (Peleg, 1963b); but Asscher (1976) has defined a non-empty variant; see also Billera (1970a).

Crucial parameters in calculating whether an imputation x is in the bargaining set of v are the *excesses* $v(S) - x(S)$ of coalitions S w.r.t. x, which measure the ability of members of S to use x in an objection (or counter-objection). Not, as is often wrongly assumed, because the initiator of the objection can assign the excess to himself while keeping his partners at their original level, but for precisely the opposite reason: because he can parcel out the excess to his partners, which makes counterobjecting more difficult.

The excess is so ubiquitous in bargaining set calculations that it eventually took on intuitive significance on its own. This led to the formulation of two additional solution concepts: the *kernel* (Davis and Maschler, 1965), which is always included in the bargaining set but is often much smaller, and the *nucleolus* (Schmeidler, 1969), which always consists of a single point in the kernel.

To define the nucleolus, choose first all those imputations x whose maximum excess (among the 2^n excesses $v(S) - x(S)$) is minimum (among all imputations). Among the resulting imputations, choose next those whose second largest excess is minimum, and so on. Schmeidler's theorem asserts that by the time we have gone through this procedure 2^n times, there is just one imputation left.

We have seen that the excess is a measure of a coalition's ‘manoeuvring ability’; in these terms the greatest measure of stability, as expressed by the nucleolus, is reached when all coalitions have manoeuvring ability as nearly alike as possible. An alternative interpretation of the excess is as a measure of S's total dissatisfaction with x, the volume of the cry that S might raise against x. In these terms, the nucleolus suggests that the final accommodation is determined by the loudest cry against it. Note that the *total* cry is determining, not the average cry; a large number of moderately unhappy citizens can be as potent a force for change as a moderate number of very unhappy ones. Variants of the nucleolus that use the average excess miss this point.

When the core is non-empty, the nucleolus is always in it. The nucleolus has been given several alternative characterizations, direct (Kohlberg, 1971, 1972) as well as axiomatic (Sobolev, 1975). The kernel was axiomatically characterized by Peleg (1986), and many interesting relationships have been found between the bargaining set, core, kernel, and nucleolus (e.g. Maschler, Peleg and Shapley, 1979). There is a large body of applications, of which we here cite just one: In a decisive weighted voting game, the nucleolus constitutes a set of weights (Peleg 1968). Thus the nucleolus may be thought of as a natural generalization of ‘voting weights’ to arbitrary games. (We have already seen that value and weights are quite different: see *1950–1960*, vi.)

(v) *The equivalence principle*. Perhaps the most remarkable single phenomenon in game and economic theory is the relationship between the price equilibria of a competitive market economy, and all but one of the major solution concepts for the corresponding game (the one exception is the stable set, about which more below). By a ‘market economy’ we here mean a pure exchange economy, or a production economy with constant returns.

We call an economy ‘competitive’ if it has many agents, each individual one of whom has too small an endowment to have a significant effect. This has been modelled by three approaches. In the *asymptotic approach*, one lets the number of agents tend to infinity, and shows that in an appropriate sense, the solution concept in question – core, value, bargaining set, or strategic equilibrium – tends to the set of competitive allocations (those corresponding to price equilibria). In the *continuum approach*, the agents constitute a (non-

atomic) continuum, and one shows that the solution concept in question actually equals the set of competitive allocations (see the entry on large economies). In the *non-standard* approach, the agents constitute a non-standard model of the integers in the sense of Robinson (1974), and again one gets equality. Both the continuum and the non-standard approaches require extensions of the theory to games with infinitely many players; see vi.

Intuitively, the equivalence principle says that the institution of market prices arises naturally from the basic forces at work in a market, (almost) no matter what we assume about the way in which these forces work. Compare (*1930–1950*, ix).

For simplicity in this section, unless otherwise indicated, the terms ‘core’, ‘value’, etc., refer to the limiting case. Thus ‘core’ means the limit of the cores of the finite economies, or the core of the continuum economy, or of the non-standard economy.

For the core, the asymptotic approach was pioneered by Edgeworth (1881), Shubik (1959) and Debreu and Scarf (1963). Anderson (1986) is an excellent survey of the large literature that ensued. Early writers on the continuum approach included Aumann (1964) and Vind (1965); the non-standard approach was developed by Brown and Robinson (1975). Except for Shubik's, all these contributions were NTU. See the entry on CORE. After a 20-year courtship, this was the honeymoon of game theory and mathematical economics, and it is difficult to convey the palpable excitement of those early years of intimacy between the two disciplines.

Some early references for the value equivalence principle, covering both the asymptotic and continuum approaches, were listed above (*1930–1950*, vi). For the non-standard approach, see Brown and Loeb (1976). Whereas the core of a competitive economy equals the set of *all* competitive allocations, this holds for the value only when preferences are smooth (Shapley, 1964; Aumann and Shapley, 1974; Aumann 1975; Mas-Colell, 1977). Without smoothness, every value allocation is competitive, but not every competitive allocation need be a value allocation. When preferences are kinky (non-differentiable utilities), the core is often quite large, and then the value is usually a very small subset of the core; it gives much more information. In the TU case, for example, the value is always a single point, even when the core is very large. Moreover, it occupies a central position in the core (Hart, 1980; Tauman, 1981; Mertens, 1988); in particular, when the core has a centre of symmetry, the value is that centre of symmetry (Hart, 1977a).

For example, suppose that in a glove market (*1950–1960*, vi), the number (or measure) of left-glove holders equals that of right-glove holders. Then at a price equilibrium, the price ratio between left and right gloves may be anything between 0 and ∞ (inclusive!). Thus the left-glove holders may end up giving away their merchandise for nothing to the right-glove holders, or the other way around, or anything inbetween. The same, of course, holds for the core. But the value prescribes precisely equal prices for right and left gloves.

It should be noted that in a finite market, the core contains the competitive allocations, but usually also much more. As the number of agents increases, the core ‘shrinks’, in the limit leaving only the competitive allocations. This is not so for the value; in finite markets, the value allocations may be disjoint from the core, and a fortiori from the competitive allocations (*1950–1960*, vi).

We have seen (*1930–1950*, iv) that the core represents a very strong and indeed not quite reasonable notion of stability. It might therefore seem perhaps not so terribly surprising that it shrinks to the competitive allocations. What happens, one may ask, when one considers one of the more reasonable stability concepts that are based on domination, such as the bargaining set or the stable sets?

For the bargaining set of TU markets, an asymptotic equivalence theorem was established by Shapley and Shubik in the mid-Seventies, though it was not published until 1984. Extending this result to NTU, to the continuum, or to both seemed difficult. The problems were conceptual as well as mathematical; it was difficult to give a coherent formulation. In 1986, Shapley presented the TU proof at a conference on the equivalence principle that took place at Stony Brook. A. Mas-Colell, who was in the audience, recognized the relevance of several results that he had obtained in other connections; within a day or two he was able to formulate and prove the equivalence principle for the bargaining set in NTU continuum economies (Mas-Colell, 1988). In particular, this implies the core equivalence principle; but it is a much stronger and more satisfying result.

For the strategic equilibrium the situation had long been less satisfactory, though there were results (Shubik, 1973; Dubey and Shapley, 1980). The difficulty was in constructing a satisfactory strategic (or extensive) model of exchange. Very recently Douglas Gale (1986) provided such a model and used it to prove a remarkable equivalence theorem for strategic equilibria in the continuum mode.

The one notable exception to the equivalence principle is the case of stable sets, which predict the formation of cartels even in fully competitive economies (Hart, 1974). For example, suppose half the agents in a continuum initially hold 2 units of bread each, half initially hold 2 units of cheese, and the utility functions are concave, differentiable, and symmetric (e. g., $u(x, y) = \sqrt{x} + \sqrt{y}$). There is then a unique price equilibrium, with equal prices for bread and cheese. Thus each agent ends up with one piece of bread and one piece of cheese; this is also the unique point in the core and in the bargaining set, and the unique NTU value. But stable set theory predicts that the cheese holders will form a cartel, the bread holders will form a cartel, and these two cartels will bargain with each other as if they were individuals. The upshot will depend on the bargaining, and may yield an outcome that is much better for one side than for the other. Thus at each point of the unique stable set with the full symmetry of the game, each agent on each side gets as much as each other agent on that side; but these two amounts depend on the bargaining, and may be quite different from each other.

In a sense, the failure of stable set theory to fall into line makes the other results even more impressive. It shows that there isn't some implicit tautology lurking in the background, that the equivalence principle makes a substantive assertion.

In the *Theory of Games*, von Neumann and Morgenstern (1944) wrote that

> when the number of participants becomes really great, some hope emerges that the influence of every particular participant will become negligible . These are, of course, the classical conditions of ‘free competition’ … The current assertions concerning free competition appear to be very valuable surmises and inspiring anticipations of results. But they are not results, and it is scientifically unsound to treat them as such.

One may take the theorems constituting the equivalence principle as embodying precisely this kind of ‘result’. Yet it is interesting that Morgenstern himself, who died in 1977, never became convinced of the validity of the equivalence principle; he thought of it as mathematically correct but economically wrongheaded. It was his firm opinion that economic agents organize themselves into coalitions, that perfect competition is a fiction, and that stable sets explain it all. The greatness of the man is attested to by the fact that though scientifically opposed to the equivalence principle, he gave generous support, both financial and moral, to workers in this area.

(vi) *Many players*. The preface to *Contributions to the Theory of Games* I (Kuhn and Tucker, 1950) contains an agenda for future research that is remarkable in that so many of its items – computation of minimax, existence of stable sets, *n*-person value, NTU games, dynamic games – did in fact become central in subsequent work. Item 11 in this agenda reads, ‘establish significant asymptotic properties of *n*-person games, for large *n*’. We have seen ((v)) how this was realized in the equivalence principle for large economies. But actually, political game models with many players are at least as old as economic ones, and may be older. During the early Sixties, L.S. Shapley, working alone and with various collaborators, wrote a series of seven memoranda at the Rand Corporation under the generic title ‘Values of Large Games’, several of which explored models of large elections, using the asymptotic and the continuum approaches. Among these were models which had both ‘atoms’ – players who are significant as individuals – and an ‘ocean’ of individually insignificant players. On example of this is a corporation with many small stockholders and a few large stockholders; see also Milnor and Shapley (1978). ‘Mixed’ models of this kind – i.e. with an ocean as well as atoms – have been explored in economic as well as political contexts using various solution notions, and a large literature has developed. The core of mixed markets has been studied by Drèze, Gabszewicz and Gepts (1969), Gabszewicz and Mertens (1971), Shitovitz (1973) and many others. For the nucleolus of ‘mixed’ voting games, see Galil (1974). Among the studies of values of mixed games are Hart (1973), Fogelman and Quinzii (1980), and Neyman (1987).

Large games in which *all* the players are individually insignificant – *non-atomic* games – have also been studied extensively. Among the early contributions to value theory in this connection are Kannai (1966), Riker and Shapley (1968), and Aumann and Shapley (1974). The subject has proliferated greatly, with well over a hundred contributions since 1974, including theoretical contributions as well as economic and political applications.

There are also games with infinitely many players in which *all* the players are atoms, namely games with a denumerable infinity of players. Again, values and voting games loom large in this literature. See, e.g., Shapley (1962), Artstein (1972) and Berbee (1981).

(vii) *Cores of finite games and markets*. Though the core was defined as an independent solution concept by Gillies and Shapley already in the early Fifties, it was not until the Sixties that a significant body of theory was developed around it. The major developments centre around conditions for the core to be non-empty; gradually it came to be realized that such conditions hold most naturally and fully when the game has an ‘economic’ rather than a ‘political’ flavour, when it may be thought of as arising from a market economy.

The landmark contributions in this area were the following: the Gale–Shapley 1962 paper on the core of a marriage market; the work of Bondareva (1963) and Shapley (1967) on the balancedness condition for the non-emptiness of the core of a TU game; Scarf's 1967 work on balancedness in NTU games; the work of Shapley and Shubik (1969a) characterizing TU market games in terms of non-emptiness of the core; and subsequent work, mainly associated with the names of Billera and Bixby (1973), that extended the Shapley–Shubik condition to NTU games. Each of these contributions was truly seminal, in that it inspired a large body of subsequent work.

Gale and Shapley (1962) asked whether it is possible to match *m* women with *m* men so that there is no pair consisting of an unmatched woman and man who prefer each other to the partners with whom they were matched. The corresponding question for homosexuals has a negative answer: the preferences of four homosexuals may be such that no matter how they are paired off, there is always an unmatched pair of people who prefer each other to the person with whom they were matched. This is so, for example, if the preferences of *a*, *b*, and *c* are cyclic, whereas *d* is lowest in all the others' scales. But for the heterosexual problem, Gale and Shapley showed that the answer is positive.

This may be stated by saying that the appropriately defined NTU coalitional game has a non-empty core. Gale and Shapley proved not only the non-emptiness but also provided a simple algorithm for finding a point in it.

This work has spawned a large literature on the cores of discrete market games. One fairly general recent result is Kaneko and Wooders (1982), but there are many others. A fascinating application to the assignment of interns to hospitals has been documented by Roth (1984). It turns out that American hospitals, after fifty years of turmoil, finally developed in 1950 a method of assignment that is precisely a point in the core.

We come now to general conditions for the core to be non-empty. Call a TU game *v superadditive at* a coalition U if $v(U) \geq \sum_j v(S_j)$ for any partition of U into disjoint coalition S_j. This may be strengthened by allowing partitions of U into disjoint ‘part-time’ coalitions θS, interpreted as coalitions S operating during a proportion θ of the time $(0 \leq \theta \leq 1)$. Such a partition is therefore a family $\{\theta_j S_j\}$, where the total amount of time that each player in U is employed is exactly 1; i.e., where $\sum_j \theta_j \chi_{S_j} = \chi_U$, where χ_S is the indicator function of S. If we think of $v(S)$ as the revenue that S can generate when operating full-time, then the part-time coalition θS generates $\theta v(S)$. Superadditivity at U for part-time coalitions thus means that

$$\sum_j \theta_j x_{S_j} = x_U \text{ implies } v(U) \geq \sum_j \theta_j v(S_j)$$

A TU game *v* obeying this condition for $U = I$ is called *balanced*; for all U, *totally balanced*.

Intuitively, it is obvious that a game with a non-empty core must be superadditive at I; and once we have the notion of part-time coalitions, it is only slightly less obvious that it must be balanced. The converse was established (independently) by Bondareva (1963) and Shapley (1967). Thus *a TU game has a non-empty core if and only if it is balanced*.

The connection between the core and balancedness (generalized superadditivity) led to several lines of research. Scarf (1967) extended the notion of balancedness to NTU games, then showed that every balanced NTU game has a non-empty core. Unlike the Bondareva–Shapley proof, which is based on linear programming methods, Scarf's proof was more closely related to fixed-point ideas. Eventually, Scarf realized that his methods could be used actually to prove Brouwer's fixed-point theorem, and moreover, to develop effective algorithms for approximating fixed points. This, in turn, led to the development of algorithms for approximating competitive equilibria of economies (Scarf, 1973), and to a whole area of numerical analysis dealing with the approximation of fixed points (see computation of general equilibria).

An extension of the Bondareva-Shapley result to the NTU case that is different from Scarf's was obtained by Billera (1970a).

Another line of research that grew out of balancedness deals with characterizing markets in purely game-theoretic terms. When can a given coalitional game *v* be expressed as a market game (*1930–1950*, ii)? The Bondareva–Shapley theorem implies that market games have non-empty cores, and this also follows from the fact that outcomes corresponding to competitive equilibria are always in the core. Since a subgame of a market game is itself a market game, it follows that *for v to be a market game, it is necessary that it and all its subgames have non-empty cores*, i.e., that the game be totally balanced. (A *subgame* of a coalitional game *v* is defined by restricting its domain to subcoalitions of a given coalition *U*.) Shapley and Shubik (1969a) showed that *this necessary condition is also sufficient*. Balancedness itself is not sufficient, since there exist games with non-empty cores having subgames with empty cores (e.g., $|I| = 4$, $v(S) := 0$, 0, 1, 1, 2 when $|S| = 0, 1, 2, 3, 4$, respectively).

For the NTU case, characterizations of market games have been obtained by Billera and Bixby (1973), Mas-Colell (1975), and others.

Though the subject of this section is finite markets, it is nevertheless worthwhile to relate the results to non-atomic games (where the players constitute a non-atomic continuum, an ‘ocean’). The total balancedness condition then takes on a particularly simple form. Suppose, for simplicity, that *v* is a function of finitely many measures, i.e., $v(S) = f(\mu(S))$, where $\mu = (\mu_1, \ldots, \mu_n)$, and the μ_j are non-atomic measures. Then *v* is a market game iff *f* is concave and 1-homogeneous $(f(\theta x) = \theta(fx))$ when $\theta \square 0$). This is equivalent to saying that *v* is superadditive (at all coalitions), and *f* is 1-homogeneous (Aumann and Shapley, 1974).

Perhaps the most remarkable expression of the connection between superadditivity and the core has been obtained by Wooders (1983). Consider coalitional games with a fixed finite number *k* of ‘types’ of players, the coalitional form being given by $v(S) = f(\mu(S))$, where $\mu(S)$ is the profile of type sizes in S, i.e. it is a vector whose *i*'th coordinate represents the number of type *i* players in S. (To specify the game, $\mu(I)$ must also be specified.) Assume that *f* is superadditive, i.e. $f(x + y) \square f(x) + f(y)$ for all *x* and *y* with non-negative integer coordinates; this assures the superadditivity of *v*. Moreover, assume that *f* obeys a ‘Lipschitz’ condition, namely that $|f(x) - f(y)| / \|x - y\|$ is uniformly bounded for all $x \neq y$, where $\|x\| := \max_j |x_j|$. Then for each $\varepsilon > 0$, when the number of players is sufficiently large, the ε-core is non-empty. (The ε-core is defined as the set of all outcomes *x* such that $x(S) \square v(S) - \varepsilon|S|$ for all S.) Roughly, the result says that the core is ‘almost’ non-empty for sufficiently large games that are superadditive and obey the Lipschitz condition. Intuitively, the superadditivity together with the Lipschitz condition yield ‘approximate’ 1-homogeneity, and in the presence of 1-homogeneity, superadditivity is equivalent to concavity. Thus *f* is approximately a 1-homogeneous concave function, so that we are back in a situation similar to that treated in the previous paragraph. What makes this result so remarkable is that other than the Lipschitz condition, the only substantive assumption is superadditivity.

Wooders (1983) also obtained a similar theorem for NTU; Wooders and Zame (1984) obtained a formulation that does away with the finite type assumption.

1970 – 1986

We do not yet have sufficient distance to see the developments of this period in proper perspective. Political and political economic models were studied in depth. Non-cooperative game theory was applied to a large variety of particular economic models, and this led to the study of important variants on the refinements of the equilibrium concept. Great strides forward were made in almost all the areas that had been initiated in previous decades, such as repeated games (both of complete and of incomplete information), stochastic games, value, core, nucleolus, bargaining theory, games with many players, and so on (many of these developments have been mentioned above). Game Theory was applied to biology, computer science, moral philosophy, cost allocation. New light was shed on old concepts such as randomized strategies.

Sociologically, the discipline proliferated greatly. Some 16 or 17 people participated in the first international workshop on game theory held in Jerusalem in 1965; the fourth one, held

in Cornell in 1978, attracted close to 100, and the discipline is now too large to make such workshops useful. An international workshop in the relatively restricted area of repeated games, held in Jerusalem in 1985, attracted over 50 participants. The *International Journal of Game Theory* was founded in 1972; *Mathematics of Operations Research*, founded in 1975, was organized into three major 'areas', one of them Game Theory. Economic theory journals, such as the *Journal of Mathematical Economics*, the *Journal of Economic Theory*, *Econometrica*, and others devoted increasing proportions of their space to game theory. Important centres of research, in addition to the existing ones, sprang up in France, Holland, Japan, England, and India, and at many Universities in the United States.

Gradually, game theory also became less personal, less the exclusive concern of a small 'in' group whose members all know each other. For years, it had been a tradition in game theory to publish only a fraction of what one had found, and then only after great delays, and not always what is most important. Many results were passed on by word of mouth, or remained hidden in ill-circulated research memoranda. The 'Folk Theorem' to which we alluded above (*1950–1960*, iii) is an example. This tradition had both beneficial and deleterious effects. On the one hand, people did not rush into print with trivia, and the slow cooking of results improved their flavour. As a result, phenomena were sometimes rediscovered several times, which is perhaps not entirely bad, since you understand something best when you discover it yourself. On the other hand, it was difficult for outsiders to break in; non-publication caused less interest to be generated than would otherwise have been, and significantly impeded progress.

Be that as it may, those days are over. There are now hundreds of practitioners, they do not all know each other, and sometimes have never even heard of one another. It is no longer possible to communicate in the old way, and as a result, people are publishing more quickly. As in other disciplines, it is becoming difficult to keep abreast of the important developments. Game theory has matured.

(i) *Applications to biology*. A development of outstanding importance, whose implications are not yet fully appreciated, is the application of game thory to evolutionary *biology*. The high priest of this subject is John Maynard Smith (1982), a biologist whose concept of *evolutionarily stable strategy*, a variant of strategic equilibrium, caught the imagination both of biologists and of game theorists. On the game theoretic side, the theme was taken up by Reinhard Selten (1980, 1983) and his school; a conference on 'Evolutionary theory in biology and economics', organized by Selten in Bielefeld in 1985, was enormously successful in bringing field biologists together with theorists of games to discuss these issues. A typical paper was tit for tat in the great tit (Regelmann and Curio, 1986); using actual field observations, complete with photographs, it describes how the celebrated 'tit for tat' strategy in the repeated prisoners' dilemma (Axelrod, 1984) accurately describes the behaviour of males and females of a rather common species of bird called the great tit, when protecting their young from predators.

It turns out that ordinary, utility maximizing rationality is much more easily observed in animals and even plants than it is in human beings. There are even situations where rats do significantly better than human beings. Consider, for example, the famous probability matching experiment, where the subject must predict the values of a sequence of i.i.d. random variables taking the values L and R with probabilities 3/4 and 1/4 respectively; each correct prediction is rewarded. It is of course optimal always to predict L; but human subjects tend to match the probabilities, i.e. to predict L about 3/4 of the time. On the other hand, while rats are not perfect (i.e. do not predict L *all* the time), they do predict L significantly more often than human beings.

Several explanations have been suggested. One is that in human experimentation, the subjects try subconsciously to 'guess right', i.e. to guess what the experimenter 'wants' them to do, rather than maximizing utility. Another is simply that the rats are more highly motivated. They are brought down to 80 per cent of their normal body weight, are literally starving; it is much more important for them to behave optimally than it is for human subjects.

Returning to theory, though the notion of strategic equilibrium seems on the face of it simple and natural enough, a careful examination of the definition leads to some doubts and questions as to why and under what conditions the players in a game might be expected to play a strategic equilibrium. See the entry on Nash equilibrium, refinements of.

Evolutionary theory suggests a simple rationale for strategic equilibrium, in which there is no conscious or overt decision making at all. For definiteness, we confine attention to two-person games, though the same ideas apply to the general case. We think of each of the two players as a whole species rather than an individual; reproduction is assumed asexual. The set of pure strategies of each player is interpreted as the locus of some gene (examples of a locus are eye colour, degree of aggressiveness, etc.); individual pure strategies are interpreted as alleles (blue or green or brown eyes, aggressive or timid behaviour, etc.). A given individual of each species possesses just one allele at the given locus; he interacts with precisely one individual in the other species, who also has just one allele at the locus of interest. The result of the interaction is a definite increment or decrement in the fitness of each of the two individuals, i.e. the number (or expected number) of his offspring; thus the payoff in the game is denominated in terms of fitness.

In these terms, a mixed strategy is a distribution of alleles throughout the population of the species (e.g., 40% aggressive, 60% timid). If each individual of each species is just as likely to meet any one individual of the other species as any other one, then the probability distribution of alleles that each individual faces is precisely given by the original mixed strategy. It then follows that a given pair of mixed strategies is a strategic equilibrium if and only if it represents a population equilibrium, i.e. a pair of distributions of characteristics (alleles) that does not tend to change.

Unfortunately, sexual reproduction screws up this story, and indeed the entire Maynard Smith approach has been criticized for this reason. But to be useful, the story does not have to be taken entirely literally. For example, it applies to evolution that is cultural rather than biological. In this approach, a 'game' is interpreted as a *kind* of confrontational situation (like shopping for a car) rather than a specific instance of such a situation; a 'player' is a role ('buyer' or 'salesman'), not an individual human being; a pure strategy is a possible kind of behaviour in this role ('hard sell' or 'soft sell'). Up to now this is indeed not very different from traditional game theoretic usage. What is different in the evolutionary interpretation is that pure or mixed strategic equilibria do not represent conscious rational choices of the players, but rather a population equilibrium which evolves as the result of how successful

certain behaviour is in certain roles.

(ii) *Randomization as ignorance*. In the traditional view of strategy randomization, the players use a randomizing device, such as a coin flip, to decide on their actions. This view has always had difficulties. Practically speaking, the idea that serious people would base important decisions on the flip of a coin is difficult to swallow. Conceptually, too, there are problems. The reason a player must randomize in equilibrium is only to keep others from deviating. For himself, randomizing is unnecessary; he will do as well by choosing any pure strategy that appears with positive probability in his equilibrium mixed strategy.

Of course, there is no problem if we adopt the evolutionary model described above in (i); mixed strategies appear as population distributions, and there is no explicit randomization at all. But what is one to make of randomization within the more usual paradigm of conscious, rational choice?

According to Savage (1954), randomness is not physical, but represents the ignorance of the decision maker. You associate a probability with every event about which you are ignorant, whether this event is a coin flip or a strategic choice by another player. The important thing in strategy randomization is that the *other* players be ignorant of what you are doing, and that they ascribe the appropriate probabilities to each of your pure strategies. It is not necessary for you actually to flip a coin.

The first to break away from the idea of explicit randomization was J. Harsanyi (1973). He showed that if the payoffs to each player *i* in a game are subjected to small independent random perturbations, known to *i* but not to the other players, then the resulting game of incomplete information has *pure* strategy equilibria that correspond to the mixed strategy equilibria of the original game. In plain words, nobody really randomizes. The appearance of randomization is due to the payoffs not being exactly known to all; each player, who does know his own payoff exactly, has a unique optimal action against his estimate of what the others will do.

This reasoning may be taken one step further. Even without perturbed payoffs, the players simply do not know which strategies will be chosen by the other players. At an equilibrium of ‘matching pennies’, each player knows very well what he himself will do, but ascribes 1/2 –1/2 probabilities to the other's actions; he also knows that the other ascribes those probabilities to his own actions, though it is admittedly not quite obvious that this is necessarily the case. In the case of a general n-person game, the situation is essentially similar; the mixed strategies of *i* can always be understood as describing the uncertainty of players other than *i* about what *i* will do (Aumann, 1987).

(iii) *Refinements of strategic equilibrium*. In analysing specific economic models using the strategic equilibrium – an activity carried forward with great vigour since about 1975 – it was found that Nash's definition does not provide adequately for rational choices given one's information at each stage of an extensive game. Very roughly, the reason is that Nash's definition ignores contingencies ‘off the equilibrium path’. To remedy this, various “refinements’ of strategic equilibrium have been defined, starting with Selten's (1975) ‘trembling hand’ equilibrium. Please refer to our discussion of Zermelo's theorem (1930–1950, vi), and to Section IV of the entry on nash equilibrium, refinements of.

The interesting aspect of these refinements is that they use *irrationality* to arrive at a strong form of rationality. In one way or another, all of them work by assuming that irrationality cannot be ruled out, that the players ascribe irrationality to each other with a small probability. True rationality requires ‘noise’; it cannot grow in sterile ground, it cannot feed on itself only.

(iv) *Bounded rationality*. For a long time it has been felt that both game and economic theory assume too much rationality. For example the hundred-times repeated prisoner's dilemma has some $2^{2^{100}}$ pure strategies; all the books in the world are not large enough to write this number even once in decimal notation. There is no practical way in which all these strategies can be considered truly available to the players. On the face of it, this would seem to render statements about the equilibrium points of such games (*1950–1960*, iv) less compelling, since it is quite possible that if the sets of strategies were suitably restricted, the equilibria would change drastically.

For many years, little on the formal level was done about these problems. Recently the theory of automata has been used for formulations of bounded rationality in repeated games. Neyman (1985a) assumes that only strategies that are programmable on an automaton of exogenously fixed size can be considered ‘available’ to the players. He then shows that even when the size is very large, one obtains results that are qualitatively different from those when all strategies are permitted. Thus in the n-times repeated prisoner's dilemma, only the greedy-greedy outcome can occur in equilibrium; but if one restricts the players to using automata with as many as $e^{o(n)}$ states, then for sufficiently large *n*, one can approximate in equilibrium any feasible individually rational outcome, and in particular the friendly–friendly outcome. For example, this is the case if the number of states is bounded by any fixed polynomial in *n*. In unpublished work, Neyman has generalized this result from the prisoner's dilemma to arbitrary games; specifically, he shows that a result similar to the Folk Theorem holds in any long finitely repeated game, when the automaton size is limited as above to subexponential.

Another approach has been used by Rubinstein (1986), with dramatically different results. In this work, the automaton itself is endogenous; all states of the automaton must actually be used on the equilibrium path. Applied to the prisoner's dilemma, this assumption leads to the conclusion that in equilibrium, one cannot get anywhere near the friendly–friendly outcome. Intuitively, the requirement that all states be used in equilibrium rules out strategies that punish deviations from equilibrium, and these are essential to the implicit enforcement mechanism that underlies the folk theorem. See the discussion at (*1950–1960*, iii) above.

(v) *Distributed computing*. In the previous subsection (iv) we discussed applications of computer science to game theory. There are also applications in the opposite direction; with the advent of distributed computing, game theory has become of interest in computer science. Different units of a distributed computing system are viewed as different players, who must communicate and coordinate. Breakdowns and failures of one unit are often modelled as malevolent, so as to get an idea as to how bad the worst case can be. From the point of view of computer tampering and crime, the model of the malevolent player is not merely a fiction; similar remarks hold for cryptography, where the system must be made proof against purposeful attempts to ‘break in’. Finally, multi-user systems come close to being games in the ordinary sense of the word.

(vi) *Consistency* is a remarkable property which, in one form or another, is common to just about all game-theoretic solution concepts. Let us be given a game, which for definiteness we denote *v*, though it may be NTU or even non-cooperative. Let *x* be an outcome that 'solves' the game in some sense, like the value or nucleolus or a point in the core. Suppose now that some coalition S wishes to view the situation as if the players outside S get their components of *x* so to speak exogenously, without participating in the play. That means that the players in S are playing the 'reduced game' v_x^S, whose all-player set is S. It is not always easy to say just how v_x^S should be defined, but let's leave that aside for the moment. Suppose we apply to v_x^S the same solution concept that when applied to *v* yields *x*. Then the consistency property is that *x*|S (*x* restricted to S) is the resulting solution. For example, if *x* is the nucleolus of *v*, then for each *v*, the restriction *x*|S is the nucleolus of v_x^S.

Consistency implies that it is not too important how the player set is chosen. One can confine attention to a 'small world', and the outcome for the denizens of this world will be the same as if we had looked at them in a 'big world'.

In a game theoretic context, consistency was first noticed by J. Harsanyi (1959) for the Nash solution to the *n*-person bargaining game. This is simply an NTU game V in which the only significant coalitions are the single players and the all-player coalition, and the single players are normalized to get 0. The Nash solution, axiomatized by Harsanyi (1959), is the outcome *x* that maximizes the product $x^1x^2 \bullet \ldots \bullet x^n$. To explain the consistency condition, let us look at the case $n = 3$, in which case V({1, 2, 3}) is a subset of 3-space. If we let $S = \{1, 2\}$, and if x_0 is the Nash solution, then 3 should get x_0^3. That means that 1 and 2 are confined to bargaining within that slice of V{(1, 2, 3}) that is determined by the plane $x^3 = x_0^3$. According to the Nash solution for the two-person case, they should maximize x^1x^2 over this slice; it is not difficult to see that this maximum is attained at (x_0^1, x_0^2), which is exactly what consistency requires.

Davis and Maschler (1965) proved that the kernel satisfies a consistency condition; so do the bargaining set, core, stable set, and nucleolus, using the same definition of the reduced game v_x^S as for the kernel (Aumann and Drèze, 1974). Using a somewhat different definition of v_x^S, consistency can be established for the value (Hart and Mas-Colell, 1986). Note that strategic equilibria, too, are consistent; if the players outside S play their equilibrium strategies, an equilibrium of the resulting game on S is given by having the players in S play the same strategies that they were playing in the equilibrium of the large game.

Consistency often plays a key role in axiomatizations. Strategic equilibrium is axiomatized by consistency, together with the requirement that in one-person maximization problems, the maximum be chosen. A remarkable axiomatization of the Nash solution to the bargaining problem (including the 2-person case discussed at *1950–1960*, v), in which the key role is played by consistency, has been provided by T. Lensberg (1981). Axiomatizations in which consistency plays the key role have been provided for the nucleolus (Sobolev, 1975), core (Peleg, 1985, 1986), kernel (Peleg, 1986), and value (Hart and Mas-Colell, 1986). Consistency-like conditions have also been used in contexts that are not strictly game-theoretic, e.g. by Balinski and Young (1982), W. Thomson, J. Roemer, H. Moulin, H.P. Young and others.

In law, the consistency criterion goes back at least to the 2000-year old Babylonian Talmud (Aumann and Maschler, 1985). Though it is indeed a very natural condition, its huge scope is still somewhat startling.

(vii) The fascination of *cost allocation* is that it retains the formal structure of cooperative game theory in a totally different interpretation. The question is how to allocate joint costs among users. For example, the cost of a water supply or sewage disposal system serving several municipalities (e.g. Bogardi and Szidarovsky, 1976); the cost of telephone calls in an organization such as a university or corporation (Billera, Heath and Raanan, 1978); or the cost of an airport (Littlechild and Owen, 1973, 1976). In the airport case, for example, each 'player' is one landing of one airplane, and *v*(*S*) is the cost of building and running an airport large enough to accommodate the set *S* of landings. Note that *v*(*S*) depends not only on the number of landings in *S* but also on its composition; one would not charge the same for a landing of a 747 as for a Piper, for example because the 747 requires a longer runway. The allocation of cost would depend on the solution concept; for example, if we are using the Shapley value ϕ, then the fee for each landing *i* would be $\phi^i v$.

The axiomatic method is particularly attractive here, since in this application the axioms often have rather transparent meaning. Most frequently used has been the Shapley value, whose axiomatic characterization (see Shapley value) is particularly transparent (Billera and Heath, 1982).

The literature on the game theoretic approach to cost allocation is quite large, probably several hundred items, many of them in the accounting literature (e.g. Roth and Verrecchia, 1979).

Concluding remarks

(i) *Ethics*. While game theory does have intellectual ties to ethics, it is important to realize that in itself, it has no moral content, makes no moral recommendations, is ethically neutral. Strategic equilibrium does not tell us to maximize utility, it explores what happens when we do. The Shapley value does not recommend dividing payoff according to power, it simply measures the power. Game Theory is a tool for telling us where incentives will lead. History and experience teach us that if we want to achieve certain goals, including moral and ethical ones, we had better see to the incentive effects of what we are doing; and if we do not want people to usurp power for themselves, we had better build institutions that spread power as thinly and evenly as possible. Blaming game theory – or, for that matter, economic theory – for selfishness is like blaming bacteriology for disease. Game theory studies

selfishness, it does not recommend it.

(ii) *Mathematical methods*. We have had very little to say about mathematical methods in the foregoing, because we wished to stress the conceptual side. Worth noting, though, is that mathematically, game theoretic results developed in on context often have important implications in completely different contexts. We have already mentioned the implications of two-person zero-sum theory for the theory of the core and for correlated equilibria (*1910–1930*, vii). The first proofs of the existence of competitive equilibrium (Arrow and Debreu, 1954) used the existence of strategic equilibrium in a generalized game (Debreu, 1952). Blackwell's 1956 theory of two-person zero-sum games with vector payoffs is of fundamental importance for n-person repeated games of complete information (Aumann, 1961) and for repeated games of incomplete information (e.g. Mertens, 1982; Hart, 1985b). The Lemke–Howson algorithm (1962) for finding equilibria of two-person non-zero sum non-cooperative games was seminal in the development of the algorithms of Scarf (1967, 1973) for finding points in the core and finding economic equilibria.

(iii) *Terminology*. Game theory has sometimes been plagued by haphazard, inappropriate terminology. Some workers, notably L.S. Shapley (1973b), have tried to introduce more appropriate terminology, and we have here followed their lead. What follows is a brief glossary to aid the reader in making the proper associations.

Used here	*Older term*
Strategic form	Normal form
Strategic equilibrium	Nash equilibrium
Coalitional form	Characteristic function
Transferable utility	Side payment
Decisive voting game	Strong voting game
Improve upon	Block
Worth	Characteristic function value
Profile	*n*-tuple
1-homogeneous	Homogeneous of degree 1

See Also

- exchange
- Shapley value

Bibliography

Anderson, R.M. 1986. Notions of core convergence. In Hildenbrand and Mas-Colell (1986), 25–46.

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: John Wiley.

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Artstein, Z. 1972. Values of games with denumerably many players. *International Journal of Game Theory* 3, 129–40.

Asscher, N. 1976. An ordinal bargaining set for games without side payments. *Mathematics of Operations Research* 1, 381–9.

Aumann, R.J. 1960. Linearity of unrestrictedly transferable utilities. *Naval Research Logistics Quarterly* 7, 281–4.

Aumann, R.J. 1961. The core of a cooperative game without side payments. *Transactions of the American Mathematical Society* 98, 539–52.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometerica* 32, 39–50.

Aumann, R.J. 1975. Values of markets with a continuum of traders. *Econometerica* 43, 611–46.

Aumann, R.J. 1985. On the non-transferable utility value: a comment on the Roth-Shafer examples. *Econometrica* 53, 667–77.

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55, 1–18.

Aumann, R.J. and Drèze, J.H. 1974. Cooperative games with coalition structures. *International Journal of Game Theory* 3, 217–38.

Aumann, R.J. and Kurz, M. 1977. Power and taxes. *Econometrica* 45, 1137–61.

Aumann, R.J. and Maschler, M. 1985. Game theoretic analysis of a bankruptcy problem from the Talmud. *Journal of Economic Theory* 36, 195–213.

Aumann, R.J. and Peleg, B. 1960. Von Neumann-Morgenstern solutions to cooperative games without side payments. *Bulletin of the American Mathematical Society* 66, 173–9.

Aumann, R.J. and Shapley, L.S. 1974. *Values of Non-Atomic Games*. Princeton: Princeton University Press.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Balinski, M.L. and Young, H.P. 1982. *Fair Representation*. New Haven: Yale University Press.

Berbee, H. 1981. On covering single points by randomly ordered intervals. *Annals of Probability* 9, 520–28.

Bewley, T. and Kohlberg, E. 1976. The asymptotic theory of stochastic games. *Mathematics of Operations Research* 1, 197–208.

Billera, L.J. 1970a. Existence of general bargaining sets for cooperative games without side payments. *Bulletin of the American Mathematical Society* 76, 375–9.

Billera, L.J. 1970b. Some theorems on the core of an n-person game without side payments. *SIAM Journal of Applied Mathematics* 18, 567–79.

Billera, L.J. and Bixby, R. 1973. A characterization of polyhedral market games. *International Journal of Game Theory* 2, 253–61.

Billera, L.J. and Heath, D.C. 1982. Allocation of shared costs: a set of axioms yielding a unique procedure. *Mathematics of Operations Research* 7, 32–9.

Billera, L.J., Heath, D.C. and Raanan, J. 1978. Internal telephone billing rates – a novel application of non-atomic game theory. *Operations Research* 26, 956–65.

Binmore, K. 1982. Perfect equilibria in bargaining models. *ICERD Discussion Paper No*. 58. London: London School of Economics.

Blackwell, D. 1956. An analogue of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics* 6, 1–8.

Blackwell, D. and Ferguson, T.S. 1968. The big match. *Annals of Mathematical Statistics* 39, 159–63.

Bogardi, I. and Szidarovsky, F. 1976. Application of game theory in water management. *Applied Mathematical Modelling* 1, 11–20.

Bondareva, O.N. 1963. Some applications of linear programming methods to the theory of cooperative games [in Russian]. *Problemy kibernetiki* 10, 119–39.

Borel, E. 1924. Sur les jeux où interviennent l'hasard et l'habilité des joueurs. In *Eléments de la theorie des probabilités*, ed. J. Hermann. Paris: Librairie Scientifique, 204–24.

Braithwaite, R.B., ed. 1950. F.P. Ramsey, *The Foundations of Mathematics and Other Logical Essays*. New York: Humanities Press.

Brams, S.J., Lucas, W.F. and Straffin, P.D., Jr., eds. 1983. *Political and Related Models*. New York: Springer.

Brown, D.J. and Loeb, P. 1976. The values of non-standard exchange economies. *Israel Journal of Mathematics* 25, 71–86.

Brown, D.J. and Robinson, A. 1975. Non standard exchange economies. *Econometrica* 43, 41–55.

Case, J.H. 1979. *Economics and the Competitive Process*. New York: New York University Press.

Champsaur, P. 1975. Cooperation vs. competition. *Journal of Economic Theory* 11, 394–417.

Dantzig, G.B. 1951a. A proof of the equivalence of the programming problem and the game problem. In Koopmans (1951), 330–38.

Dantzig, G.B. 1951b. Maximization of a linear function of variables subject to linear inequalities. In Koopmans (1951), 339–47.

Davis, M. 1964. Infinite games with perfect information. In Dresher, Shapley and Tucker (1964), 85–101.

Davis, M. 1967. Existence of stable payoff configurations for cooperative games. In Shubik (1967), 39–62.

Davis, M. and Maschler, M. 1965. The kernel of a cooperative game. *Naval Research Logistics Quarterly* 12, 223–59.

Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences of the United States* 38, 886–93.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 236–46.

Dresher, M.A., Shapley, L.S. and Tucker, A.W., eds. 1964. *Advances in Game Theory*. Annals of Mathematics Studies Series 52, Princeton: Princeton University Press.

Dresher, M.A., Tucker, A.W. and Wolfe, P., eds. 1957. *Contributions to the Theory of Games III*. Annals of Mathematics Studies Series 39, Princeton: Princeton University Press.

Drèze, J.H., Gabszewicz, J. and Gepts, S. 1969. On cores and competitive equilibria. In Guilbaud (1969), 91–114.

Dubey, P. 1980. Asymptotic semivalues and a short proof of Kannai's theorem. *Mathematics of Operations Research* 5, 267–70.

Dubey, P. and Shapley, L.S. 1980. Non cooperative exchange with a continuum of traders: two models. *Technical Report of the Institute for Advanced Studies*, the Hebrew University of Jerusalem.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Everett, H. 1957. Recursive games. In Dresher, Tucker and Wolfe (1957), 47–78.

Fogelman, F. and Quinzii, M. 1980. Asymptotic values of mixed games. *Mathematics of Operations Research* 5, 86–93.

Gabszewicz, J.J. and Mertens, J.F. 1971. An equivalence theorem for the core of an economy whose atoms are not ‘too’ big. *Econometrica* 39, 713–21.

Gale, D. 1974. A curious nim-type game. *American Mathematical Monthly* 81, 876–79.

Gale, D. 1979. The game of hex and the Brouwer fixed-point theorem. *American Mathematical Monthly* 86, 818–27.

Gale, D. 1986. Bargaining and competition, Part I: Characterization; Part II: Existence. *Econometrica* 54, 785–806; 807–18.

Gale, D. and Shapley, L.S. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69, 9–15.

Gale, D. and Stewart, F.H. 1953. Infinite games with perfect information. In Kuhn and Tucker (1953), 245–66.

Galil, Z. 1974. The nucleolus in games with major and minor players. *International Journal of Game Theory* 3, 129–40.

Gillies, D.B. 1959. Solutions to general non-zero-sum games. In Luce and Tucker (1959), 47–85.

Guilbaud, G.T., ed. 1969. *La décision: aggrégation et dynamique des orders de préférence*. Paris: Editions du CNRS.

Harsanyi, J.C. 1956. Approaches to the bargaining problem before and after the theory of games: a critical discussion of Zeuthen's, Hicks' and Nash's theories. *Econometrica* 24, 144–57.

Harsanyi, J.C. 1959. A bargaining model for the cooperative n-person game. In Tucker and Luce (1959), 325–56.

Harsanyi, J.C. 1963. A simplified bargaining model for the n-person cooperative game. *International Economic Review* 4, 194–220.

Harsanyi, J.C. 1966. A general theory of rational behavior in game situations. *Econometrica* 34, 613–34.

Harsanyi, J.C. 1967–8. Games with incomplete information played by ‘Bayesian’ players, parts I, II and III. *Management Science* 14, 159–82, 320–34, 486–502.

Harsanyi, J.C. 1973. Games with randomly disturbed payoffs: a new rationale for mixed strategy equilibrium points. *International Journal of Game Theory* 2, 1–23.

Harsanyi, J.C. 1982. Solutions for some bargaining games under the Harsanyi-Selten solution theory I: Theoretical preliminaries; II: Analysis of specific games. *Mathematical Social Sciences* 3, 179–91; 259–79.

Harsanyi, J.C. and Selten, R. 1972. A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science* 18, 80–106.

Harsanyi, J.C. and Selten, R. 1987. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Hart, S. 1973. Values of mixed games. *International Journal of Game Theory* 2, 69–86.

Hart, S. 1974. Formation of cartels in large markets. *Journal of Economic Theory* 7, 453–66.

Hart, S. 1977a. Asymptotic values of games with a continuum of players. *Journal of Mathematical Economics* 4, 57–80.

Hart, S. 1977b. Values of non-differentiable markets with a continuum of traders. *Journal of Mathematical Economics* 4, 103–16.

Hart, S. 1980. Measure-based values of market games. *Mathematics of Operations Research* 5, 197–228.

Hart, S. 1985a. An axiomatization of Harsanyi's nontransferable utility solution. *Econometrica* 53, 1295–314.

Hart, S. 1985b. Non zero-sum two-person repeated games with incomplete information. *Mathematics of Operations Research* 10, 117–53.

Hart, S. and Mas-Colell, A. 1986. The potential: a new approach to the value in multi-person allocation problems. Harvard University Discussion Paper No. 1157.

Hart, S. and Schmeidler, D. 1988. Correlated equilibria: an elementary existence proof. *Mathematics of Operations Research*.

Hildenbrand, W., ed. 1982. *Advances in Economic Theory*. Cambridge: Cambridge University Press.

Hildenbrand, W. and Mas-Colell, A. 1986. *Contributions to Mathematical Economics in Honor of G. Debreu*. Amsterdam: North-Holland.

Hu, T.C. and Robinson, S.M., eds. 1973. *Mathematical Programming*. New York: Academic Press.

Isaacs, R. 1965. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. New York: John Wiley.

Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8, 457–9.

Kalai, E. and Smorodinsky, M. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43, 513–18.

Kaneko, M. and Wooders, M. 1982. Cores of partitioning games. *Mathematical Social Sciences* 3, 313–27.

Kannai, Y. 1966. Values of games with a continuum of players. *Israel Journal of Mathematics* 4, 54–8.

Kohlberg, E. 1971. On the nucleolus of a characteristic function game. *SIAM Journal of Applied Mathematics* 20, 62–6.

Kohlberg, E. 1972. The nucleolus as a solution to a minimization problem. *SIAM Journal of Applied Mathematics* 23, 34–49.

Koopmans, T.C., ed. 1951. *Activity Analysis of Production and Allocation*. New York: Wiley.

Kuhn, H.W. 1952. *Lectures on the Theory of Games*. Issued as a report of the Logistics Research Project, Office of Naval Research, Princeton University.

Kuhn, H.W. 1953. Extensive games and the problem of information. In Kuhn and Tucker (1953), 193–216.

Kuhn, H.W. and Tucker, A.W., eds. 1950. *Contributions to the Theory of Games I*. Annals of Mathematics Studies Series 24, Princeton: Princeton University Press.

Kuhn, H.W. and Tucker, A.W., eds. 1953. *Contributions to the Theory of Games II*. Annals of Mathematics Studies Series 28, Princeton: Princeton University Press.

Lemke, L.E. and Howson, J.T. 1962. Equilibrium points of bimatrix games. *SIAM Journal of Applied Mathematics* 12, 413–23.

Lensberg, T. 1981. The stability of the Nash solution. Unpublished.

Lewis, D.K. 1969. *Convention*. Cambridge, Mass.: Harvard University Press.

Littlechild, S.C. 1976. A further note on the nucleolus of the ‘airport game’. *International Journal of Game Theory* 5, 91–5.

Littlechild, S.C. and Owen, G. 1973. A simple expression for the Shapley value in a special case. *Management Science* 20, 370–72.

Lucas, W.F. 1969. The proof that a game may not have a solution. *Transactions of the American Mathematical Society* 137, 219–29.

Lucas, W.F. 1983. Measuring power in weighted voting systems. In Brams, Lucas and Straffin (1983), ch. 9.

Lucas, W.F. and Rabie, M. 1982. Games with no solutions and empty core. *Mathematics of Operations Research* 7, 491–500.

Luce, R.D. and Raiffa, H. 1957. *Games and Decisions, Introduction and Critical Survey*. New York: John Wiley.

Luce, R.D. and Tucker, A.W., eds. 1959. *Contributions to the Theory of Games IV*. Annals of Mathematics Studies Series 40, Princeton: Princeton University Press.

Lyapounov, A.A. 1940. On completely additive vector-functions (in Russian, abstract in French). *Akademiia Nauk USSR Izvestiia Seriia Mathematicheskaia* 4, 465–78.

Martin, D.A. 1975. Borel determinacy. *Annals of Mathematics* 102, 363–71.

Maschler, M., ed. 1962. *Recent Advances in Game Theory*. Proceedings of a Conference, privately printed for members of the conference, Princeton: Princeton University Conferences.

Maschler, M., Peleg, B. and Shapley, L.S. 1979. Geometric properties of the kernel, nucleolus, and related solution concepts. *Mathematics of Operations Research* 4, 303–38.

Maschler, M. and Perles, M. 1981. The superadditive solution for the Nash bargaining game. *International Journal of Game Theory* 10, 163–93.

Mas-Colell, A. 1975. A further result on the representation of games by markets. *Journal of Economic Theory* 10, 117–22.

Mas-Colell, A. 1977. Competitive and value allocations of large exchange economies. *Journal of Economic Theory* 14, 419–38.

Mas-Colell, A. 1988. An equivalence theorem for a bargaining set. *Journal of Mathematical Economics*.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Mertens, J.F. 1982. Repeated games: an overview of the zero-sum case. In Hildenbrand (1982), 175–82.

Mertens, J.F. 1988. The Shapley value in the non-differentiable case. *International Journal of Game Theory* 17, 1–65.

Mertens, J.F. and Neyman, A. 1981. Stochastic games. *International Journal of Game Theory* 10, 53–66.

Milnor, J.W. 1978. Values of large games II: Oceanic games. *Mathematics of Operations Research* 3, 290–307.

Milnor, J.W. and Shapley, L.S. 1957. On games of survival. In Dresher, Tucker and Wolfe (1957), 15–45.

Moschovakis, Y.N. 1980. *Descriptive Set Theory*. New York: North-Holland.

Moschovakis, Y.N., ed. 1983. *Cabal Seminar 79–81: Proceedings, Caltech-UCLA Logic Seminar 1979–81*. Lecture Notes in Mathematics 1019, New York: Springer-Verlag.

Mycielski, J. and Steinhaus, H. 1964. On the axiom of determinateness. *Fundamenta Mathematicae* 53, 205–24.

Myerson, R.B. 1977. Graphs and cooperation in games. *Mathematics of Operations Research* 2, 225–9.

Myerson, R.B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47, 61–74.

Myerson, R.B. 1984. Cooperative games with incomplete information. *International Journal of Game Theory* 13, 69–96.

Nash, J.F., Jr. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Nash, J.F., Jr. 1951. Non-cooperative games. *Annals of Mathematics* 54, 289–95.

Neyman, A. 1985a. Bounded complexity justifies cooperation in the finitely repeated prisoner's dilemma. *Economics Letters* 19, 227–30.

Neyman, A. 1985b. Semivalues of political economic games. *Mathematics of Operations Research* 10, 390–402.

Neyman, A. 1987. Weighted majority games have an asymptotic value. *Mathematics of Operations Research*.

O'Neill, B. 1987. Non-metric test of the minimax theory of two-person zero-sum games. *Proceedings of the National Academy of Sciences of the United States* 84, 2106–9.

Owen, G. 1972. Multilinear extensions of games. *Management Science* 18, 64–79.

Peleg, B. 1963a. Solutions to cooperative games without side payments. *Transactions of the American Mathematical Society* 106, 280–92.

Peleg, B. 1963b. Bargaining sets of cooperative games without side payments. *Israel Journal of Mathematics* 1, 197–200.

Peleg, B. 1967. Existence theorem for the bargaining set $M_1^{(i)}$. In Shubik (1967), 53–6.

Peleg, B. 1968. On weights of constant-sum majority games. *SIAM Journal of Applied Mathematics* 16, 527–32.

Peleg, B. 1985. An axiomatization of the core of cooperative games without side payments. *Journal of Mathematical Economics* 14, 203–14.

Peleg, B. 1986. On the reduced game property and its converse. *International Journal of Game Theory* 15, 187–200.

Pennock, J.R. and Chapman, J.W., eds. 1968. *Representation*. New York: Atherton.

Ramsey, F.P. 1931. Truth and probability. In Braithwaite (1950).

Ransmeier, J.S. 1942. *The Tennesee Valley Authority: A Case Study in the Economics of Multiple Purpose Stream Planning*. Nashville: Vanderbilt University Press.

Regelmann, K. and Curio, E. 1986. How do great tit (Parus Major) pair mates cooperate in broad defence? *Behavior* 97, 10–36.

Riker, W.H. and Shapley, L.S. 1968. Weighted voting: a mathematical analysis for instrumental judgements. In Pennock and Chapman (1968), 199–216.

Robinson, A. 1974. *Non-Standard Analysis*. Amsterdam: North-Holland.

Rosenthal, R.W. 1979. Sequences of games with varying opponents. *Econometrica* 47, 1353–66.

Rosenthal, R.W. and Rubinstein, A. 1984. Repeated two player games with ruin. *International Journal of Game Theory* 13, 155–77.

Roth, A.E. 1977. The Shapley value as a von Neumann–Morgenstern utility. *Econometrica* 45, 657–64.

Roth, A.E. 1984. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of Political Economy* 92, 991–1016.

Roth, A.E. and Verrecchia, R.E. 1979. The Shapley value as applied to cost allocation: a reinterpretation. *Journal of Accounting Research* 17, 295–303.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50, 97–109.

Rubinstein, A. 1986. Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory* 39, 83–96.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: John Wiley.

Scarf, H.E. 1967. The core of an n-person game. *Econometrica* 35, 50–69.

Scarf, H.E. 1973. *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Schelling, T.C. 1960. *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

Schmeidler, D. 1969. The nucleolus of a characteristic function game. *SIAM Journal of Applied Mathematics* 17, 1163–70.

Selten, R.C. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. *Zeitschrift für die gesammte Staatswissenschaft* 121, 301–24; 667–89.

Selten, R.C. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4, 25–55.

Selten, R.C. 1980. A note on evolutionary stable strategies in asymmetric animal conflicts. *Journal of Theoretical Biology* 84, 101–101.

Selten, R.C. 1983. Evolutionary stability in extensive two-part games. *Mathematical Social Sciences* 5, 269–363.

Shapley, L.S. 1953a. Stochastic games. *Proceedings of the National Academy of Sciences of the United States* 39, 1095–100.

Shapley, L.S. 1953b. A value for n-person games. In Kuhn and Tucker (1953), 305–17.

Shapley, L.S. 1962. Values of games with infinitely many players. In Maschler (1962), 113–18.

Shapley, L.S. 1964. Values of large games, VII: a general exchange economy with money. *RAND Publication* RM–4248, Santa Monica, California.

Shapley, L.S. 1967. On balanced sets and cores. *Naval Research Logistics Quarterly* 14, 453–60.

Shapley, L.S. 1969. Utility comparison and the theory of games. In Guilbaud (1969), 251–63.

Shapley, L.S. 1973a. On balanced games without side payments. In Hu and Robinson (1973), 261–90.

Shapley, L.S. 1973b. Let's block ‘block’. *Econometrica* 41, 1201–2.

Shapley, L.S. 1984. Convergence of the bargaining set for differentiable market games. Appendix B in Shubik (1984), 683–92.

Shapley, L.S. and Shubik, M. 1954. A method for evaluating the distribution of power in a committee system. *American Political Science Review* 48, 787–92.

Shapley, L.S. 1969a. On market games. *Journal of Economic Theory* 1, 9–25.

Shapley, L.S. 1969b. Pure competition, coalitional power and fair division. *International Economic Review* 10, 337–62.

Shitovitz, B. 1973. Oligopoly in markets with a continuum of traders. *Econometrica* 41, 467–501.

Shubik, M. 1959a. Edgeworth market games. In Luce and Tucker (1959), 267–78.

Shubik, M. 1959b. *Strategy and Market Structure*. New York: John Wiley.

Shubik, M., ed. 1967. *Essays in Mathematical Economics in Honor of Oskar Morgenstern*. Princeton: Princeton University Press.

Shubik, M. 1973. Commodity, money, oligopoly, credit and bankruptcy in a general equilibrium model. *Western Economic Journal* 11, 24–36.

Shubik, M. 1982. *Game Theory in the Social Sciences, Concepts and Solutions*. Cambridge, Mass.: MIT Press.

Shubik, M. 1984. *A Game Theoretic Approach to Political Economy*. Cambridge, Mass.: MIT Press.

Sobolev, A.I. 1975. Characterization of the principle of optimality for cooperative games through functional equations (in Russian). In Vorobiev (1975), 94–151.

Sorin, S. 1986a. On repeated games of complete information. *Mathematics of Operations Research* 11, 147–60.

Sorin, S. 1986b. An asymptotic property of non-zero sum stochastic games. *International Journal of Game Theory* 15(2), 101–7.

Tauman, Y. 1981. Value on a class of non-differentiable market games. *International Journal of Game Theory* 10, 155–62.

Ville, J.A. 1938. Sur le théorie générale des jeux où intervient I'habilité des joueurs. In *Traité du calcul des probabilités et de ses applications*, Vol. 4, ed. E. Borel. Paris: Gauthier-Villars, 105–13.

Vinacke, W.E. and Arkoff, A. 1957. An experimental study of coalitions in the triad. *American Sociological Review* 22, 406–15.

Vind, K. 1965. A theorem on the core of an economy. *Review of Economic Studies* 32, 47–8.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320.

von Neumann, J. 1949. On rings of operators. Reduction theory. *Annals of Mathematics* 50, 401–85.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Vorobiev, N.N., ed. 1975. *Mathematical Methods in Social Science* (in Russian). Vipusk 6, Vilnius.

Wilson, R. 1978. Information, efficiency, and the core of an economy. *Econometrica* 46, 807–16.

Wolfe, P. 1955. The strict determinateness of certain infinite games. *Pacific Journal of Mathematics* 5, 841–7.

Wooders, M.H. 1983. The epsilon core of a large replica game. *Journal of Mathematical Economics* 11, 277–300.

Wooders, M.H. and Zame, W.R. 1984. Approximate cores of large games. *Econometrica* 52, 1327–50.

Young, H.P. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14, 65–72.

Zermelo, E. 1913. Über eine Anwendung der Mengenlehre auf die theorie des Schachspiels. *Proceedings of the Fifth International Congress of Mathematicians* 2, 501–4.

How to cite this article

Aumann, R. J. "game theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000007> doi:10.1057/9780230226203.0615

The New Palgrave Dictionary of Economics Online

games in coalitional form

Ehud Kalai

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

How should a coalition of cooperating players allocate payoffs to its members? This question arises in a broad range of situations and evokes an equally broad range of issues. For example, it raises technical issues in accounting, if the players are divisions of a corporation, but involves issues of social justice when the context is how people behave in society.
Despite the breadth of possible applications, coalitional game theory offers a unified framework and solutions for addressing such qsts. This article presents some of its major models and proposed solutions.

Keywords

additive games; auction game; Aumann–Shapley prices; axiomatic characterizations; balanced games; bargaining; bargaining sets; coalition; coalition formation; coalitional game; coalitional monotonicity; communication graph; consistency; cooperation; cooperative game theory; core; cost allocation; dummy player; egalitarian solution; egalitarian value; equivalence th; flow game; folk th; game theory; games in coalitional form; grand coalition; Harsanyi value; imputation; independence of irrelevant alternatives; individual monotonicity; individual rationality; Kalai–Smorodinsky solution; kernel; large market games; law of large numbers; linear programming game; majority game; market games; Maschler–Perles

The New Palgrave Dictionary of Economics Online

Geary, Robert Charles (1896 – 1983)

J.R.N. Stone

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Geary, R. C.; instrumental variables; Ireland, economics in; linear expenditure system; methodology of econometrics

Article

Geary was born on 11 April 1896 in Dublin, and died on 8 February 1983, also in Dublin. He was educated at University College, Dublin, from 1913 to 1918, and at the Sorbonne from 1919 to 1921. In 1922 he became assistant lecturer in mathematics at University College, Southampton. He held the post of statistician in the Department of Industry and Commerce in Dublin from 1923 to 1949. From 1946 to 1947 he was a Senior Research Fellow in the Department of Applied Economics in Cambridge. He held the position of First Director in the Central Statistical Office of Eire from 1949 to 1957. In 1957 he was appointed Head of the National Accounts Branch of the United Nations Statistical Office, which post he held until 1960. From 1960 to 1966 he was director of the newly founded Economic (and Social) Research Institute in Dublin, remaining attached to it as consultant from 1966 until his death.

Geary was a mathematical statistician of international standing. His statistical writings cover a wide range of topics, among them testing for normality, the distribution of ratios, parameter estimation, and so on, some of which are relevant to econometric methodology. Indeed, much of his work has an explicitly economic content: he was probably responsible for the excellent report prepared by the Department of Industry and Commerce containing the first national accounts of Eire (Eire, Minister for Finance, 1946); he wrote many papers on the determination of relationships between variables, notably Geary (1948; 1949), in the second of which he uses instrumental variables; he derived the form of the utility function underlying the linear expenditure system (Geary, 1950–51); he was part-author of a monograph on linear programming applied to economics (Geary and McCarthy, 1964); and he built a model of the Irish economy based on an accounting framework (Geary, 1963–4). His published output numbers 112 titles, more than half of which appeared after his 65th birthday. A full bibliography is appended to Spencer (1976; 1983).

Selected works

1946. Eire, Minister for Finance. *National Income and Expenditure* 1938–1944. Dublin: Stationery Office.

1948. Studies in relations between economic time series. *Journal of the Royal Statistical Society*, Series B 10(1), 140–58.

1949. Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica* 17, 30–58.

1950–51. A note on ‘A constant utility index of the cost of living’. *Review of Economic Studies* 18, 65–6.

1964. (With M.D. McCarthy.) *Elements of Linear Programming with Economic Applications*. London: Griffin. 2nd edn (with J.E. Spencer), 1973.

1963–4. Towards an input–output decision model for Ireland. *Journal of the Statistical and Social Inquiry Society of Ireland* 21, Pt 2, 67–119.

Bibliography

Spencer, J.E. 1976. The scientific work of Robert Charles Geary. *Economic and Social Review* 7(3), 233–47.

Spencer, J.E. 1983. Robert Charles Geary – an appreciation. *Economic and Social Review* 14(3), 161–4.

How to cite this article

Stone, J. R. N. "Geary, Robert Charles (1896–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000018> doi:10.1057/9780230226203.0619

The New Palgrave Dictionary of Economics Online

Gee, Joshua (fl.1725 – 1750)

Douglas Vickers
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Child, J.; consumers' expenditure; employment of the poor; Gee, J.; labour supply; protectionism; workhouses

Article

Joshua Gee's place in the history of economics rests on his contributions to the protectionist literature during the early decades of the 18th century. He collaborated with Henry Martin among others in publishing the *British Merchant* that argued the protectionist case in 1713 and 1714 against the Treaty of Commerce proposed at Utrecht, and he published an extensive discussion of England's foreign trade, together with strong protectionist sentiments, in *The Trade and Navigation of Great Britain Considered* in 1729.

Addressing the current decline in the English export trades, the high level of imports of certain commodities, the demand for which, particularly in the case of French fashion goods, could be met by home produced import substitutes, the declining health of the woollen industry, and the currently widespread unemployment, Gee made a number of proposals for government regulation of trade and manufacturing. These proposals were directed principally to the need for ‘finding effectual ways for employing the poor’, thereby aligning his work with the widespread employment argument of the time. To the same end he advocated also a wider development of workhouses. Following Josiah Child, Gee proposed that trade with the colonial plantations should be regulated in such a way as not only to encourage their production of the materials needed for English manufacturing industries, thereby ‘employing all the poor’, but also to facilitate ‘supplying our plantations with everything they want and all manufactured within ourselves’.

Gee's descriptive essays exhibited an understanding of the interdependence of economic activities and processes, ‘one employment depending on another’, and of ‘the circulation of commerce that must infuse riches into every part’. He argued that higher domestic commodity prices would induce workers to increase their supply of labour, leading to higher incomes and also higher discretionary consumption expenditures. But such potentially important notions were not accorded any systematic or analytical

development.

See Also

- mercantilism

Selected works

1729. *The Trade and Navigation of Great Britain Considered*. London.

1742. *An Impartial Enquiry into the Importance and Present State of the Woollen Manufactures of Great Britain*. London.

1742. *The Grazier's Advocate, or Free Thoughts of Wool and the Woollen Trade*. London.

How to cite this article

Vickers, Douglas. "Gee, Joshua (fl.1725–1750)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000019> doi:10.1057/9780230226203.0620

The New Palgrave Dictionary of Economics Online

gender differences (experimental evidence)

Catherine C. Eckel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Laboratory experiments find differences between women and men in three main areas: altruism, risk aversion and competition. The types of experiments and findings are described, and findings summarized. These results parallel similar findings in other social sciences, and are consistent with observed differences in the field.

Keywords

altruism; altruism in experiments; charitable giving; competition; cooperation; dictator game; experimental economics; gender differences; gender gap; overconfidence; preference elicitation; risk aversion; social dilemma; ultimatum game; women's work and wages

Article

Henry Higgins famously enquires, ‘Why can't a woman be more like a man?’ While others might phrase it differently, the question of how and why women and men differ is the subject of hundreds of books and articles every year. Experimental economists have investigated gender differences in at least three areas: cooperation and altruism, attitudes toward risk, and preferences for engaging in competitive activities. While this research has proceeded largely independently of other social sciences, the results across fields are parallel.

Most experimental research on gender differences is motivated by an interest in the persistent gender gap in earnings, with women earning significantly less than men even after adjusting for productivity related differences in education, experience, choice of employment, and so on (Weichselbaumer and Winter-Ebmer, 2005). While this gender gap has diminished since the 1970s it has not disappeared. Attention is also directed at the fact that women are underrepresented in leadership positions. Within economics, the Committee on the Status of Women in the Economics Profession of the American Economic Association keeps tabs on the progress of women. Their most recent survey shows that

women continue to lag behind men in their progress towards higher academic ranks. While women earned 20 per cent of Ph.D.s in economics in the 1980s and 25 per cent in the 1990s, only 8.3 per cent of full professors in Ph.D. granting departments are women (see Committee on the Status of Women, 2007).

Differences in the behaviour of women and men are extensively documented in research in psychology and sociology. An overview of this work, which covers differences in ability, personality, leadership styles, aggression, competitiveness and so on, can be found in Rhoads (2004) and Maccoby (1998) among many others. Experiments have examined gender differences in situations involving salient monetary incentives since Rapoport and Chammah (1965), who explored variations of the Prisoner's Dilemma (PD) game. Early experimental work in psychology and sociology tended to involve this game, or related social dilemma (SD) games, with mixed results. In games with this incentive structure – where each player has a dominant strategy to free ride, but group payoffs are maximized by choosing a cooperative strategy – many studies have found that women are more cooperative, and many that they are less so.

Experimental research in economics focuses on examining the types of preferences that might be related to the gender gap: those that relate to cooperating, taking risks and competing. Compared to the stereotypical male person, the stereotypical female person is more altruistic and cooperative, and more averse to risk and competition. Partial surveys of research in experimental economics on gender differences are provided by Eckel and Grossman (2008a), which focuses on altruism and cooperation, and (2008b), which surveys studies of risk aversion, and a more comprehensive review is contained in Croson and Gneezy (2007).

Cooperation

If there is no systematic difference between the sexes in their play of PD and SD games, can we abandon this element of the stereotype and conclude that women are no more cooperative than men when money is at stake? Eckel and Grossman (1998) were the first to point out that these games confound two possible differences in the preferences of women and men. Suppose that, true to stereotype, women are both more altruistic and more risk averse. Altruistic preferences imply that women will be more likely to choose a cooperative strategy in PD and SD games. However, risk aversion implies just the opposite. The cooperative strategy is also the risky strategy; a cooperator risks being exploited, with corresponding low earnings. The best choice for an altruistic, risk-averse person would depend on the parameters of the game, that is, the trade-off between the gain to cooperation and the penalty if one is betrayed. Thus the games that have been most commonly used to measure cooperation may be confounded by risk aversion.

Eckel and Grossman's (1998) strategy was to separate altruism from risk aversion. In a double-anonymous dictator game, where there is no financial (or social) risk, they report that women give about twice as much to an anonymous partner. This result has not always been replicated by subsequent studies, and behaviour can vary with the characteristics of the recipients when they are known, but overall it is rare to find a situation where men are more altruistic. In more complex experiments (Andreoni and Vesterlund, 2001; Dickinson and Tiefenthaler, 2002), subjects make a series of dictator decisions in tokens, where the tokens have different exchange rates for each of the players. In these games men tend to maximize efficiency, allocating more to the partner with the better exchange rate,

while women tend to try to equalize earnings. Thus men appear more altruistic at exchange rates that benefit the recipient. At equal exchange rates, women give more than men. Moreover, in studies where subjects can give to a charitable organization, Eckel and Grossman consistently find that women give more than men (for example, Eckel and Grossman, 2003).

Another experimental environment that has received a great deal of attention is the ultimatum game, which suffers from a similar problem. In that game, a person might make a generous offer because of altruism or because of risk aversion; similarly, a person might accept a low offer for multiple reasons. The greater altruism and risk aversion attributed to women implies more generous ultimatum offers by women. However, results in this game are mixed (Eckel and Grossman, 2001; Solnick, 2001). Women and men make similar offers on average, but, more importantly, both make lower offers to women than to men, suggesting a commonly held belief that women will accept lower offers (because they are more altruistic?). On the respondent side, the results of these two studies are contradictory. In general, however, results indicate that women are more likely to accept an offer of a given size than the reverse. A higher degree of altruism is consistent with lower wages, with more altruistic persons both requesting and accepting lower wage offers. It is worth pointing out that (to my knowledge) no studies have tested the external validity of these measures of altruism; that is, economists have little or no knowledge of how well laboratory-elicited preferences ‘predict’ how people behave as they go about their daily lives. While lab decisions are real in the sense that there are resource consequences to decisions, the context is very different from field decisions. However, there are several studies that explicitly examine cultural context, and find a positive relationship between how groups play a public goods game and how they harvest natural resources (Carpenter and Seki, 2006). There is also some evidence that the gender gap in earnings is smaller in the nonprofit sector, where altruistic preferences might be especially valuable (Leete, 2000).

Risk aversion

Like cooperation, gender differences in risk aversion have been much studied in fields outside economics. In most situations, greater risk taking by men is well documented (Byrnes, Miller and Schafer, 1999). Economists have a rather narrow way of thinking about risk aversion compared with other social scientists; we view preferences as represented by a utility function that evaluates alternatives across all decision-making domains. Diminishing marginal utility of income or wealth produces risk aversion: the expected value of a gamble always has higher utility than the gamble itself. (Of course constant marginal utility implies risk neutrality, and increasing marginal utility risk seeking.) This view of risk aversion implies that any task that measures the curvature of the utility function in money should give a good measure of risk attitudes that is then applicable across all situations. Experimental economists have developed different games that do just that.

Like their counterparts in the other social sciences, economists tend to find women more risk averse than men, though both are surprisingly risk averse considering the level of stakes in our games. Though the difference is not always statistically significant, it is rare that it goes the other way. However, there is a potential problem with commonly used measures that might distort the gender difference. The experiments tend to be complicated, requiring a relatively high level of mathematical ability to be clearly understood. This is not a big problem if any resulting ‘noise’ does not bias the measure.

Unfortunately, there is some indication that difficult tasks cause low-ability subjects to make systematically different choices. To the extent that mathematical ability is correlated with gender, this could bias inferences about differences in risk aversion.

A popular initial experiment used a risky version of the Becker, DeGroot and Marschak (1964) (BDM) preference elicitation procedure. This mechanism elicits subjects' valuations for gambles with various probabilites of winning a particular prize. To make it incentive compatible, this mechanism requires two stages. In the first the subject writes down a minimum selling price for a gamble that pays off X dollars with probability p. In the second a random price is drawn from a uniform distribution between 0 and X dollars. If the drawn price is above the elicited price, the subject sells the gamble, and if not the subject plays the gamble. This mechanism has been criticized for its complexity, and for the low incentive for accuracy with low-probability gambles. Subjects with low maths ability may be more likely to overvalue these low probability gambles because they are confused by the second stage of the game. For example, consider a 0.05 probability of winning 10 dollars; any value drawn in the second stage is likely to be far above the expected value of the gamble, so it may not be worthwhile for the subject to bother calculating his reservation value, and he is likely to err on the high side. This would tend to make subjects look less risk averse. Indeed, BDM studies tend to find fewer risk-averse and more risk-seeking subjects.

A cleverly designed game developed by Holt and Laury (2002) has subjects choose between pairs of lotteries that are constructed so as to easily ‘back out’ a coefficient of relative risk aversion for a specified utility functional form. This game is easily comprehended by college students, and produces intuitively appealing results in educated populations (Andersen et al., 2006). However, there is some evidence it is less successful for less literate populations, limiting its usefulness in the field (Dave et al., 2007). Like the BDM procedure, failure to account for differences in mathematical ability may distort estimates of gender differences.

A third type of game involves fewer choices among simpler, 50/50 gambles (Binswanger, 1980; Eckel and Grossman, 2002). A subject chooses her favourite from among a set of 50–50 gambles that vary in risk and expected return. The experiment allows categorization of subjects into ordered categories, from most to least risk averse. There is some evidence that this experiment is easier to comprehend for populations with low mathematical literacy, although the trade-off is that the measure is coarser than the others described above. One troubling result even for educated groups is that different measures of risk aversion completed by the same set of people tend to exhibit low correlations across measures, suggesting that our underlying construct may need some work.

The experiments above all involve individual decisions. Additional indirect evidence of risk aversion can be found in a market environment. Women are more likely than men to overbid in first-price auctions, behaviour that can be caused by risk aversion. Chen, Katuscak and Ozdenoren (2005) find that women tend to overbid, but women's bids are most like men's when oestrogen levels are lowest, suggesting a biological mechanism driving greater risk aversion.

Gender differences are typically, but not always, found across all experiments designed to measure risk aversion. Women are more risk averse across environments. Several studies have begun to examine external validity of the measures. In general, lab measures of risk attitudes have low (though sometimes statistically significant) correlations with decisions in other lab experiments, and low correlations with risky field behaviours, such as buying an extended warranty for an automobile or computer (Moore, 2002). Risk attitudes also are related to a person's willingness to borrow to finance higher education

expenditure (Eckel et al., 2007). Many current studies will further examine external validity of experimental preference measures. As with altruism, to my knowledge, no study has related experimental risk measures to employment earnings. However, field studies tend to confirm gender stereotypes, with women investing in more conservative portfolios (Sunden and Surette, 1998), more likely to buy warranties (Moore, 2002), and more likely to negotiate contracts with larger salary components and smaller performance-related components (Chauvin and Ash, 1994). The outstanding question is whether experimentalists can measure risk attitudes in experimental games in a way that meaningfully predicts risky choices in field and, in particular, employment settings.

Competition

Women do not like competition. Psychologists have long known that girls are less competitive than boys, that they play different games and avoid competitive situations. For example, Maccoby (1998) quotes many such studies, including one showing that, in same-sex groups of fourth and sixth graders, boys spontaneously engaged in competitive activities 50 per cent of the time, while girls engaged in such play only one per cent of the time (1998, p. 39). Men do not merely like competition, they also do better when a situation is more competitive. Rhoads (2004) surveys work in this area and gives dozens of examples. Some authors have used these differences to argue that women are inherently ill-suited to the workplace (Browne, 2002), and others that women have an advantage because competition does not get in the way of making the best decisions (Helgesen, 1990). The taste for competition is no doubt related to men's higher levels of confidence; overconfidence can also interfere with profit-maximization, as Barber and Odean (2001) show in a study of online stock trading.

Experimental economists have discovered this, too: Gneezy, Niederle and Ructichini (2003) show in a lab experiment that introducing competition makes men, but not women, more productive in solving mazes. The study compares work performance under two types of compensation: piece rates, where workers are paid by the maze, and winner-take all, tournament rate, where only the highest producer is paid. Women work about the same under the two schemes, while men work significantly harder for the tournament payment. This result spurred two additional studies where women and men choose their preferred compensation rate. In the first, Gupta, Poulsen and Villeval (2005) again use mazes and find that 60 per cent of men and 34 per cent of women choose the tournament rate. Niederle and Vesterlund (2007) are careful to choose a task where women and men perform the same under piece and tournament rates – solving easy maths problems. Here again, men are more likely to choose the tournament (73 per cent compared to 35 per cent). This effect remains after controlling for subjects' measured ability as well as their own perceptions of their abilities; thus the result is not due to overconfidence. Men sacrifice earnings in this game because low-ability men choose the tournament, but women lose more and so earn less than men because high-ability women shy away from the tournament.

If women avoid competition, this, too, may have consequences for earnings. If a preference to avoid competition transfers from the lab to the field, then it is likely to affect the earnings of women. As with cooperation and risk, more study is needed to verify the external validity of the lab-based measures of aversion to competition.

Conclusion

Laboratory experiments show a collection of preferences that differ, on average, between the sexes. Women tend to be more altruistic, risk averse, and competition averse. This pattern of preferences could lead to patterns of behaviour that result in lower wages for women, such as accepting low offers or avoiding competitive situations. For example, Babcock and Laschever (2003) find that lower average starting salaries for women public policy graduates are the result of differences in the way men and women treat job offers. Women tend to accept the best offer they receive from potential employers; men, by contrast, respond to an offer by asking for more. This behaviour seems very much like that observed in the lab, and suggests that altruism, risk aversion, and competition aversion may play a role in explaining this.

The results of economics lab experiments are largely consistent with research from the other social sciences, and psychology in particular. Economics experiments are conducted in settings where payoffs are salient, and where there is no deception. This work has not only confirmed but also legitimized research on gender differences for economists. The importance of the work is to show that individual differences in preferences, whether by nature or nurture, can be substantial, and are correlated with observable characteristics of individuals. Decision-making in the workplace occurs in a much more complex environment, making it difficult or impossible to sort out the effects of the various dimensions of preferences. However, lab experiments allow a much higher degree of control over the environment so that variability in specific aspects of preferences can be isolated.

See Also

- gender roles and division of labour
- social preferences
- women's work and wages

Bibliography

Andersen, S., Harrison, G.W., Lau, M.I. and Rutstrom, E.E. 2006. Eliciting risk and time preferences. Working Paper 05-24, Department of Economics, College of Business Administration, University of Central Florida.

Andreoni, J. and Vesterlund, L. 2001. Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics* 116, 293–312.

Babcock, L. and Laschever, S. 2003. *Women Don't Ask: Negotiation and the Gender Divide*. Princeton, NJ: Princeton University Press.

Barber, B.M. and Odean, T. 2001. Boys will be boys: gender, overconfidence and common stock investing. *Quarterly Journal of Economics* 116, 261–92.

Becker, G.M., DeGroot, M.H. and Marschak, J. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9, 226–32.

Binswanger, H.P. 1980. Attitudes toward risk: experimental measurement in rural India. *American Journal of Agricultural Economics* 62, 395–407.

Browne, K. 2002. *Biology at Work: Rethinking Sexual Equality*. New Brunswick, NJ: Rutgers University Press.

Byrnes, J., Miller, D.C. and Schafer, W.D. 1999. Gender differences in risk taking: a meta analysis. *Psychological Bulletin* 125, 367–83.

Carpenter, J. and Seki, E. 2006. Competitive work environments and social preferences: field experimental evidence from a Japanese fishing community. Contributions to *Economic Analysis & Policy* 5, Article 2. Abstract online. Available at http://www.bepress.com/bejeap/contributions/vol5/iss2/art2, accessed 16 June 2007.

Chauvin, K.W. and Ash, R.A. 1994. Gender earnings differentials in total pay, base pay, and contingent pay. *Industrial and Labor Relations Review* 47, 634–49.

Chen, Y., Katuscak, P. and Ozdenoren, E. 2005. Why can't a woman bid more like a man? Working paper, University of Michigan.

Committee on the Status of Women. 2007. Winter newsletter. Online. Available at http://www.cswep.org/newsletters/CSWEP_nsltr_Winter2007.pdf, accessed 12 June 2007.

Croson, R. and Gneezy, U. 2007. Gender differences in preferences. Working paper, University of California, San Diego.

Dave, C., Eckel, C.C., Johnson, C.A. and Rojas, C. 2007. Eliciting risk preferences: heterogeneity, stability and precision. Working paper, University of Texas at Dallas.

Dickinson, D. and Tiefenthaler, J. 2002. What is fair? Experimental evidence. *Southern Economic Journal* 69, 414–28.

Eckel, C.C. and Grossman, P.J. 1998. Are women less selfish than men? Evidence from dictator games. *Economic Journal* 108, 726–35.

Eckel, C.C. and Grossman, P.J. 2001. Chivalry and solidarity in ultimatum games. *Economic Inquiry* 39, 171–88.

Eckel, C.C. and Grossman, P.J. 2002. Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior* 23, 281–95.

Eckel, C.C. and Grossman, P.J. 2003. Rebates versus matching: does how we subsidize charitable contributions matter? *Journal of Public Economics* 87, 681–701.

Eckel, C.C. and Grossman, P.J. 2008a. Differences in the economic decisions of men and women: experimental evidence. In *Handbook of Experimental Economics Results*, ed. C. Plott and V.L. Smith. New York: Elsevier, forthcoming.

Eckel, C.C. and Grossman, P.J. 2008b. Men, women and risk aversion: experimental evidence. In *Handbook of Experimental Economics Results*, ed. C. Plott and V.L. Smith. New York: Elsevier, forthcoming.

Eckel, C.C., Johnson, C.A., Montmarquette, C. and Rojas, C. 2007. Debt aversion and the demand for loans for postsecondary education. *Public Finance Review* 35, 233–62.

Gneezy, U., Niederle, M. and Ructichini, A. 2003. Performance in competitive environments: gender differences. *Quarterly Journal of Economics* 118, 1049–74.

Gupta, N.D., Poulsen, A. and Villeval, M.-C. 2005. Do (wo)men prefer (non)competitive jobs? Working paper, Institute for the Study of Lavoe (IZA).

Helgesen, S. 1990. *The Female Advantage: Women's Ways of Leadership*. New York: Doubleday.

Holt, C.A. and Laury, S.K. 2002. Risk aversion and incentive effects. *American Economic Review* 92, 1644–55.

Leete, L. 2000. Wage equity and employee motivation in nonprofit and for-profit organizations. *Journal of Economic Behavior & Organization* 43, 423–46.

Maccoby, E. 1998. *The Two Sexes*. Cambridge, MA: Harvard University Press.

Moore, E. 2002. An investigation into the demand for service contracts. Ph.D. thesis, Department of Economics, Virginia Polytechnic Institute and State University.

Niederle, M. and Vesterlund, L. 2007. Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics* 122(3), forthcoming.

Rapoport, A. and Chammah, A.M. 1965. Sex differences in factors contributing to the level of cooperation in the prisoner's dilemma game. *Journal of Personality and Social Psychology* 2, 831–8.

Rhoads, S.E. 2004. *Taking Sex Differences Seriously*. San Francisco: Encounter Books.

Solnick, S. 2001. Gender differences in the ultimatum game. *Economic Inquiry* 39, 189–200.

Sunden, A.E. and Surette, B.J. 1998. Gender differences in the allocation of assets in retirement savings plans. *American Economic Review* 88, 207–11.

Weichselbaumer, D. and Winter-Ebmer, R. 2005. A metaanalysis of the international gender wage gap. *Journal of Economic Surveys* 19, 479–511.

How to cite this article

Eckel, Catherine C. "gender differences (experimental evidence)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000199> doi:10.1057/9780230226203.0621

The New Palgrave Dictionary of Economics Online

gender roles and division of labour

Joyce P. Jacobsen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

All human societies exhibit some degree of division of labour by gender. These divisions continue to exist as participation in paid work has increased over time. Gender divisions occur between household tasks, between unpaid and paid work, and within paid work. Economists have explained these divisions through reliance on essentialist arguments and/or the fundamental economic concepts of efficiency of specialization and division of labour, and investment in human capital. However, gender discrimination can also cause division of labour, and the feedback effects of such discrimination make it difficult to untangle the causes of the gender division of labour.

Keywords

affirmative action; Becker, G.; capitalism; gender roles and division of labour; household economy; human capital; intrahousehold welfare; labour market discrimination; marriage markets; non-market work; occupational segregation; patriarchy; social norms; socialism; technical change; women's work and wages; work–leisure trade-off

Article

All human societies have a gender-related division of labour, although the particulars of division vary across time and culture. It is generally agreed to be a pre-capitalist phenomenon, based on anthropological and historical information, and related to the widespread existence of patriarchy, that is, male-controlled and male-favouring social systems. Let us start by considering the arguments for why there would be gender-related task specialization in a non-market setting, that is, in a pre-capitalist society.

Gender roles and task specialization

Most of the arguments posited for why a gender-related division of labour exists have been essentialist. The general analytical approach has been to posit a specific biological sex-related difference and then show that it leads to gender task specialization. The factor can be a sex difference in ability, including in some models the ability to develop or learn particular types of skills – and thus specialization leads to efficiency in production. Or the determining factor can be a sex difference in preference or taste – and thus specialization leads directly to utility maximization. In the first category, analysts have posited differences in childbirth and child-raising abilities, fecundity, physical strength, aggression/dominance/coalition-building/risk-taking, or cognitive differences (Fausto-Sterling, 1985; Duley, Sinclair and Edwards, 1986; Becker, 1991; Siow, 1998). Gender differences in ability need not imply unbiased gains for both sexes. Indeed, if patriarchy arises because of one or more of these differences, then it is also plausible that the gender division of labour favours men, or at least some men. In the second category, analysts have posited such factors as sex differences in preferences for spouse's age (Elul, Silva-Reus and Volij, 2002), differences in caring for children and others (Folbre, 1995), and different preferences for meaningful work and other job characteristics over money (Brown and Corcoran, 1997).

A smaller set of analysts has presented non-essentialist arguments, where the general approach has been to posit that a division of labour is efficient and that some specialized human capital must be acquired initially in order to improve efficiency further; in addition, human capital in the form of specialized experience can be developed through continued application to the specific task. Therefore, in order to maximize output, societies should train some people to do one type of task, and others to do something different, and leave them in those roles for extended periods of time. Formal models of this type then link the labour market with the marriage market to consider the coordination problem and societal output maximization in arguing why it is important that people of one type, for example, sex, be assigned to one sort of task (Hadfield, 1999; Engineer and Welling, 1999; Baker and Jacobsen, 2006). If some types of output can be produced and traded only with the household, then it is important to match people of different types within marriages. Thus men should do one type of task and women another to reduce the coordination problem. The dilemma for these models is to explain why particular tasks are assigned to men and others assigned to women, particularly if some tasks are preferable on some dimension (for example, they have higher prestige or portability) and also tend empirically to be assigned to men. Thus these models often have to fall back on an essentialist starting point in order to determine initial assignment. They can then argue that societal dynamics in determining future gender assignments are affected by the initial assignment and by technological change.

The non-essentialist arguments do a better job than the essentialist arguments of explaining why societies have prescribed gender roles rather than allowing for flexibility of task assignment based on actual individual abilities. A strong essentialist argument would broach no conflict between biological sex and gender roles, yet we see deviations from gender roles by individuals, weakening the essentialist argument. Thus essentialists need a more nuanced approach in which biological sex (whether chromosonal or hormonal) leads to different probabilities of particular outcomes, or different distributions of traits. Then an argument based on efficiency of division of labour, along with the need to make specialized human capital investments early on in children's lives (Becker, 1991), leads society to assign gender roles based on average or modal outcomes by sex.

The question then arises of how to deal with deviations from gender roles. People can generally articulate gender norms, that is, roles that are considered sex-appropriate, and know when they or others

are violating them. Gender roles can relate also to age, and also may have caste or class or racial/ethnic aspects, so tasks may be assigned differentially based on these other dimensions, too. Societies deal with deviations in different ways, including complete proscription; allowing people to change later on at an efficiency loss; allowing for exceptions only if people show particular deviant traits early on; and laissez-faire. Akerlof and Kranton (2000) posit that gender identity appears in the utility function and that deviations from sex-appropriate gender identity cause utility loss for individuals. Badgett and Folbre (2003) discuss the potential penalty that one may face in the marriage market for being in a gender-non-conforming occupation. Changes in social norms regarding the social construction of sexuality may have had some effect in reducing these losses (Matthaei, 1995). Deviation tends to appear only within a dualist system; indeed, the prominence of gender duality in most, but not all, cultures is notable. In cultures with a third gender role, such persons either are assigned to specific reserved occupations or must conform to the cross-gender role if sex is not aligned with gender (Jacobsen, 2006).

Gender division of labour between non-market and market work

With the development of markets for paid labour, division additionally occurs between unpaid (non-market) and paid (market) work. It is a general observation across societies that women are more likely to specialize in non-market work and men in market work, or women to divide time between non-market and market work, and men to specialize in market work. As a first step in explaining this pattern, the neoclassical approach posits that division of labour is efficient. The household is considered the nexus for production and consumption of non-market commodities. Thus division of labour within the household occurs, with some members specializing in market work, others in non-market work. In models of the modern household, children are generally treated as consumption goods, or sometimes as investments, when previously they were thought of as additional suppliers of market or non-market labour. In order to motivate the particular gender division of labour, writers fall back generally on one or more essentialist arguments for why women do non-market work, in particular the relation to bearing and raising children. If the division of labour between non-market and market falls along gender lines, the marriage market may then be conceptualized as the market for non-market, or spousal, labour (Grossbard-Shechtman, 1993).

While the argument that specialization and division of labour is more efficient might hold in a static framework, it is not obvious that this is necessarily optimal in a dynamic framework, at least not for both parties. Specialization in non-market labour is the less desirable specialty as it limits the market for one's services by definition; indeed, if there is no marriage, there is no market for one's services at all. Thus models have explicitly linked the household division of labour to the operation of the marriage market, whether these concerns are taken into account in settling on a division of output before entering into marriage (Lundberg and Pollak, 1994) or within marriages through ongoing negotiation over distribution of the household's product.

It is also problematic to argue that specialization decisions are made solely on the basis of relative productivity. If there is discrimination in the labour market such that women are paid below their actual marginal product, then women's comparative advantage is more likely to lie in household work. Lower wages can also lead to intermittent labour force attachment, which leads back to lower wages (Gronau, 1988). Becker (1991) argues that effort as well as time must be allocated across types of labour; if

women expend more effort on household work, then they have less effort to exert on market work, thus receiving lower wages. Also, if women train to do household work, whether or not they have a greater aptitude for it, then they are more likely to have comparative advantage in household work (and the opposite for men and market work). Thus the efficiency arguments can be self-fulfilling.

One way to attempt to untangle these feedback mechanisms is to see what happens when society experiences changes through technology, political upheaval, or other factors. But the processes of political and economic change have had mixed effects on the gender division of labour, even as they have had large effects in changing the nature of work and the mix between household and market work. Some hold the view that capitalism accentuates the gender division of labour through accentuating the division between household and market work, while others think capitalism is useful in reducing the gender division of labour. Some argue that patriarchy and capitalism are mutually reinforcing (Hartmann, 1976; Humphries, 1991) and that socialism needs to include the overturning of both systems (Engels, 1884) including reassignment or eradication of patriarchy-enforcing property rights (Braunstein and Folbre, 2001). Cases of transition from capitalism to socialism – and for some countries back again – have provided mixed evidence; in practice, socialism appears to have increased women's total work time, increasing their paid work without decreasing their unpaid work (Jacobsen, 2006).

In practice, most people in modern societies do both paid and unpaid work, whether in a given time period or across the life cycle. Technological change affects gender work assignments over time whenever it is non-neutral with respect to initially gendered task assignments. It appears that the particular form in which technological change has occurred has made capital complementary to women's market work (Galor and Weil, 1996) and substitutable for women's non-market work (Greenwood, Seshadri and Yorukoglu, 2005). Real wages have been rising for women over the past century. The net effect has been to reduce women's time spent in non-market work and to increase their time spent in market work.

Gender segregation in market work

Even as paid labour has become more extensive and women have increasingly participated in paid work, extensive gender segregation persists across time and space in labour markets (Anker, 1998; Jacobsen, 2006). Market work for women often still emulates their areas of traditional female-dominated non-market work, such as child care and teaching of young children, nursing and eldercare, and food preparation and service. Men still dominate the occupations that have required more physical strength, and in industrialized societies are more likely to work in outdoor occupations. What is harder to explain along essentialist or traditionalist lines is why there would be gender segregation for other types of occupations that have arisen later in economic development, such as various types of professions.

Economists have advanced various explanations for occupational gender segregation. Again, many rely on essentialist arguments regarding differences in abilities and/or preferences to explain why women and men would choose different paid work. One approach is to argue that some jobs are more compatible with non-market duties. Thus, if women are doing most of the non-market work (which begs the question of why they are doing it), they must choose jobs that allow for this balance, including those that allow for part-time work. A dynamic version of this argument is that women know they will be balancing paid work with non-market work, particularly during their childbearing years, and thus choose

occupations that are potentially more compatible with this lower level of attachment to paid work. Notably, even in occupations where women have increased their representation substantially, there is within-occupation gender segregation along various lines, including sub-specialties, firm size, employer type (for-profit, non-profit, government), and so on. These patterns appear in many cases to be consistent with an argument that women prefer more flexible employment, that is, jobs that require less travel and less overtime work, and allow for part-time and/or flexible hours.

To switch from a supply-side to a demand-side focus, other economists have argued that gender segregation is driven by employers' choosing whom to hire, not by employees' choosing where to work. Re Becker (1971), employers may get utility directly from discriminating or simply from maintaining social norms. Employer discrimination can occur if there is insufficient competition from non-discriminating employers to drive out discriminating employers. Segregation can occur without loss of profits in Becker's employee and customer discrimination models. Male-dominated unions and other professional organizations can keep women out of particular occupations by denying them training (Fawcett, 1892). Statistical discrimination is another potential explanation of gender segregation, with the usual chicken-and-egg problem: employers don't see women doing the non-traditional task, so cannot tell whether women are good at it (Lundberg and Startz, 1983). Some have also called into question the implicit assumption of exogeneity of gender-linked preferences if pre-labour market treatment of girls and boys is different (Corcoran and Courant, 1987).

There is interesting evidence regarding the instability of gender integration from studying workers like clerks, bank tellers, and schoolteachers, whose occupations have tipped from being male-dominated to female-dominated, thus re-segregating quite rapidly (Reskin and Roos, 1990). In addition, jobs can vary in their gender assignment from society to society (Jacobsen, 2006). Thus the maintenance of gender segregation in and of itself appears to be even more fundamental than essentialist arguments regarding differential ability and/or job preferences can explain.

A notable pattern is that female-dominated jobs tend to pay less, even to the men in them, than do ‘comparable’ male-dominated jobs. Thus occupational segregation is linked with lower pay for women. This relationship could arise through various mechanisms. If women are crowded into a smaller set of occupations through hiring discrimination, crowding will lead to lower pay for women if labour demand does not adjust across occupations. If women are willing to trade off pay for working conditions, crowding into the more desirable occupations means lower pay by choice. However, reducing occupational segregation is neither necessary nor sufficient to raise pay for women; some countries (such as Sweden and Australia) with higher relative earnings for women also exhibit greater occupational segregation than countries with lower relative earnings for women (such as Japan and the United States) (Jacobsen, 2006).

Policies affecting the gender division of labour

In post-industrial societies, many public and business policies affect the gender division of labour and occupational segregation. Any policy affecting the net wage rate, including taxes on earnings, the deducibility of childcare expenses, or means-tested government benefits, affects the market work-non-market work–leisure trade-off. In general, the asymmetry between taxable income and non-taxable household production produces a bias towards household production. The net effect on behaviour of the

large number of relevant policies is unclear.

Few policies have directly aimed at reducing occupational segregation. Affirmative action has had much more notable effects on racial employment than on gender employment patterns. Educational access policies, such as the opening of college and postgraduate programmes to women, have been more important, particularly for increasing women's participation in the professions. However, the focus on access to formal education has not encouraged women to enter those jobs traditionally learned through apprenticeships, such as the crafts and trades; these areas continue to be among the most male-dominated of occupations. Meanwhile, few policies have encouraged men to enter female-dominated occupations such as nursing and childcare, even as shortages of caring labour appear.

Lack of explicit gender-desegregation public policy reflects the continued ambivalence in society regarding the desirability of gender desegregation. This stands in notable contrast to stated beliefs regarding racial desegregation, where separatism has become increasingly spurned. Gender segregation occurs in other social spaces such as sports and schooling. Most amateur sports teams continue to be gender-segregated even as US Title IX legislation and similar actions in other countries increases access to sports for high school and college women. Single-sex schooling persists in a wide range of societies and is even encouraged up through high school, although colleges are mainly co-educational except in the most gender-segregated societies such as Saudi Arabia. This segregation is often couched in terms of improving women's (and sometimes men's) outcomes (that is, through arguing that women perform better in single-sex systems), yet still constitutes an argument for separate spheres. In addition, ambivalence continues towards men raising children, the desirability of outsourced childcare, and thus the desirability of mothers working full-time. Economists and other social scientists have performed a useful service in documenting the extent and nature of gender segregation, but have not yet led a full public debate as to its desirability.

See Also

- Becker, Gary S.
- family economics
- marriage markets
- social norms
- women's work and wages

Bibliography

Akerlof, G. and Kranton, R. 2000. Economics and identity. *Quarterly Journal of Economics* 115, 715–53.

Anker, R. 1998. *Gender and Jobs: Sex Segregation of Occupations in the World*. Geneva: International Labour Office.

Badgett, M. and Folbre, N. 2003. Job gendering: occupational choice and the marriage market.

Industrial Relations 42, 270–98.

Baker, M. and Jacobsen, J. 2006. Marriage, specialization, and the gender division of labor. *Journal of Labor Economics* 24 (forthcoming).

Becker, G. 1971. *The Economics of Discrimination*, rev. edn. Chicago: University of Chicago.

Becker, G. 1991. *A Treatise on the Family*, enlarged edn. Cambridge, MA, and London: Harvard University Press.

Braunstein, E. and Folbre, N. 2001. To honor and obey: efficiency, inequality, and patriarchal property rights. *Feminist Economics* 7, 25–44.

Brown, C. and Corcoran, M. 1997. Sex-based differences in school content and the male/female wage gap. *Journal of Labor Economics* 15, 431–65.

Corcoran, M. and Courant, P. 1987. Sex-role socialization and occupational segregation: an exploratory investigation. *Journal of Post Keynesian Economics* 9, 330–46.

Duley, M., Sinclair, K. and Edwards, M. 1986. Biology versus culture. In *The Cross-Cultural Study of Women: A Comprehensive Guide*, ed. M. Duley and M. Edwards. New York: Feminist Press, City University of New York.

Elul, R., Silva-Reus, J. and Volij, O. 2002. Will you marry me? A perspective on the gender gap. *Journal of Economic Behavior and Organization* 49, 549–72.

Engels, F. 1884. The Origin of the Family, Private Property, and the State. Hottingen-Zurich.

Engineer, M. and Welling, L. 1999. Human capital, true love, and gender roles: is sex destiny? *Journal of Economic Behavior and Organization* 40, 155–78.

Fausto-Sterling, A. 1985. *Myths of Gender: Biological Theories About Women and Men*. New York: Basic Books.

Fawcett, M. 1892. Mr. Sidney Webb's article on women's wages. *Economic Journal* 2, 173–6.

Folbre, N. 1995. ‘Holding hands at midnight’: the paradox of caring labor. *Feminist Economics* 1, 73–92.

Galor, O. and Weil, D. 1996. The gender gap, fertility, and growth. *American Economic Review* 86, 374–87.

Greenwood, J., Seshadri, A. and Yorukoglu, M. 2005. Engines of liberation. *Review of Economic Studies* 72, 109–33.

Gronau, R. 1988. Sex-related wage differentials and women's interrupted labor careers – the chicken or the egg. *Journal of Labor Economics* 6, 277–301.

Grossbard-Shechtman, S. 1993. *On the Economics of Marriage: A Theory of Marriage, Labor, and Divorce*. Boulder and Oxford: Westview Press.

Hadfield, G. 1999. A coordination model of the sexual division of labor. *Journal of Economic Behavior and Organization* 40, 125–53.

Hartmann, H. 1976. Capitalism, patriarchy, and the division of labor. *Signs* 1(3), pt. 2, 137–69.

Humphries, J. 1991. The sexual division of labor and social control: an interpretation. *Review of Radical Political Economics* 23, 269–96.

Jacobsen, J. 2006. *The Economics of Gender*, 3rd edn. Malden, MA: Blackwell.

Lundberg, S. and Pollak, R. 1994. Noncooperative bargaining models of marriage. *American EconomicReview* 84, 132–7.

Lundberg, S. and Startz, R. 1983. Private discrimination and social intervention in competitive labor markets. *American Economic Review* 73, 340–7.

Matthaei, J. 1995. The sexual division of labor, sexuality, and lesbian/gay liberation: towards a marxist-feminist analysis of sexuality in U.S. capitalism. *Review of Radical Political Economics* 27, 1–37.

Reskin, B. and Roos, P. 1990. *Job Queues, Gender Queues: Explaining Women's Inroads into Male Occupations*. Philadelphia: Temple University Press.

Siow, A. 1998. Differential fecundity, markets, and gender roles. *Journal of Political Economy* 106, 334–54.

How to cite this article

Jacobsen, Joyce P. "gender roles and division of labour." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000185> doi:10.1057/9780230226203.0622

The New Palgrave Dictionary of Economics Online

general equilibrium (new developments)

William Zame

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

General equilibrium theory is the theory of mass markets. The foundations of general equilibrium theory were laid in the late 19th and early 20th centuries by Walras and Edgeworth. The modern formulation was conceived in the 1950s by Arrow, Debreu and McKenzie, who also established the fundamental results: existence of competitive equilibrium, Pareto optimality of equilibrium allocations (the First Welfare Theorem), and supportability of Pareto optimal allocations as equilibria with transfers (the Second Welfare Theorem). The ideas of general equilibrium theory are widely used in models of markets of all kinds, including in finance, international trade and macroeconomics.

Keywords

adverse selection; aggregate excess demand function; Arrow–Debreu model of general equilibrium; asymmetric information; bargaining; commodity space; competitive equilibrium; Convexity; core convergence; core equivalence; default; degree theory; Edgeworth, F. Y.; efficient markets hypothesis; equity premium; existence of equilibrium; general equilibrium; implicit function theorem; incentive-compatible core; incomplete markets; Kakutani fixed point theorem; law of demand; Lipschitz functions; market power; moral hazard; multiple equilibria; perfect competition; pooling; private core; private information; rational expectations equilibrium; revealed preference; Sard's theorem; separation theorem; tâtonnement; transferable utility; transversality conditions; uniqueness of equilibrium; Walras's Law; Walrasian expectations equilibrium

Article

The fundamental ideas, results and applications of general equilibrium theory are well described in general equilibrium, which was originally written for the first (1987) edition of *The New Palgrave*. However, there has been a great deal of notable work since – too much to adequately survey in the limited space available here. The discussion addresses only a few topics on which research has been

especially active:

- determinacy of equilibrium
- perfect competition and justification of the assumption of price-taking
- equilibration
- infinitely many commodities
- incomplete markets
- hidden information and hidden actions.

Determinacy

For many purposes, it is not enough to know simply that competitive equilibrium exists; we would like to know how equilibrium varies when the underlying parameters of the model vary. Such comparative statics analysis is simplest and most convincing when equilibrium is unique and depends nicely on the underlying parameters. However, it has been known for a long time that even some very simple economies admit multiple equilibria and that conditions on the primitives of an economy that guarantee uniqueness of equilibrium must necessarily be unpleasantly strong. For many purposes, however, it is enough to know that equilibria are locally unique, and locally depend nicely on underlying parameters. Competitive equilibrium prices are the solutions of the system of equations asserting, for each good, that demand equals supply (equivalently, that aggregate excess demand is zero). In an economy with L consumption goods, this is a system of L equations with L unknowns; taking account of the price normalization and Walras's Law (aggregate expenditure equals aggregate income), this reduces to a system of L–1 equations in L–1 unknowns. Because the number of equations equals the number of unknowns, heuristic considerations (linear approximations, for example) suggest that local uniqueness might not be too much to hope for. However, local uniqueness does not *always* obtain: it is easy to exhibit simple Edgeworth box (two persons, two goods) exchange economies for which the set of equilibrium prices is a continuum.

Debreu (1970; 1972) showed that, if preferences are sufficiently smooth and indifference surfaces are not flat and do not intersect the boundary (for example, if preferences arise from utility functions that are twice continuously differentiable and differentiably strictly concave, and exhibit infinite marginal utility for consumption at zero levels of consumption), then almost all specifications of initial endowments lead to a finite number of equilibria, and those equilibria depend locally smoothly on endowments. Debreu's method was to first use the assumptions on preferences to show that the aggregate excess demand mapping is continuously differentiable, then to rely on Sard's theorem (which guarantees that for almost every $y \in Y$ every point in the inverse image $f^{-1}(y)$ is regular), and finally to appeal to the implicit function theorem.

Two limitations of Debreu's analysis are of particular note. The first is the assumption that indifference surfaces do not intersect the boundary. An implication of this assumption is that, at equilibrium, every agent consumes every commodity. This certainly seems false-to-fact, and this assumption would be objectionable in many applied models. Boundary consumptions create problems because they lead (almost necessarily) to an aggregate excess demand function that is not differentiable. Shannon (1994) extends Debreu's results, obtaining generic determinacy while accommodating boundary consumptions,

by showing that the aggregate excess demand function, although not differentiable, is Lipschitz (that is, $|f(x) - f(y)| \bullet C\,|x - y|$ for some constant C) and that the required implications of smooth analysis remain valid for Lipschitz functions. Blume and Zame (1993) follow a different approach, based on real algebraic geometry and particularly useful for applied models, treating only utility functions that are sufficiently smooth (roughly, piecewise real-analytic), but accommodating both boundary consumptions and utility functions that are not strictly concave.

A second, more subtle, limitation of Debreu's analysis is that the set of boundary endowments is of measure zero. Hence, saying that ‘almost all specifications of initial endowments lead to a finite number of equilibria’ says nothing at all about an environment in which some agents are not endowed with strictly positive amounts of all commodities. If it is true that most economic agents do not consume all goods, it is even more true that most economic agents are endowed with only a few goods – perhaps even with their own labour and nothing else. A more satisfactory specification would allow for the possibility that some agents' endowments of some goods are constrained to be zero, and to ask for determinacy for generic specifications of *other* goods. Surprisingly, Minehart (1997) finds that such specifications are compatible with robust indeterminacy. Mas-Colell (1985) and Anderson and Zame (2001) show that generic determinacy is restored if we interpret genericity in the sense of preferences (or utility functions) as well as endowments.

For economies with an infinite dimensional space of commodities, Debreu's arguments are typically inapplicable because the commodity space and the price space are different and demand functions are almost never continuously differentiable (Araujo, 1987). However, the same outline can be applied, not to the aggregate excess demand mapping, but to the aggregate excess spending mapping. (Given a vector $\lambda = (\lambda_i)$ of utility weights for the agents, find the (unique) allocation (x_i) that maximizes the weighted sum $\sum \lambda_i u_i(x_i)$ of agent utilities and the price p that supports this allocation. The value of the excess spending map S at weights (λ_i) is the vector $S(\lambda) = (p \cdot x_i - p \cdot e_i)$. The price p and allocation (x_i) constitute a competitive equilibrium if $S(\lambda) = 0$.) Under the assumptions that utility is separable (across time or across states of the world) and that the underlying felicity functions satisfy Debreu's assumptions (at each date or in each state of the world), Kehoe and Levine (1985) show that the excess spending map is smooth and hence that equilibria are generically finite in number and depend nicely on parameters. In the general case, Shannon (1999) and Shannon and Zame (2002) identify conditions on utility functions implying that the excess spending map is Lipschitz; an application of Lipschitz analysis again yields generic determinacy.

Perfect competition and price-taking

The definition of competitive equilibrium rests on the assumption of perfect competition; that is, that agents are price-takers. This assumption is clearly untenable if some agents are large, in the sense of controlling resources that are a significant fraction of the social total (although competitive equilibrium may serve as a useful benchmark even in such environments). A large and important literature attempts to understand when the assumption of price-taking behaviour is sensible.

One of the central themes in this literature seeks to justify price-taking behaviour by showing that cooperative outcomes are competitive – or almost competitive – when the population is large and agents

are small. The largest portion of this literature is inspired by Edgeworth, who gave an informal argument that, for economies with two commodities and two consumers, replication shrinks the core of the economy (the set of feasible allocations on which no coalition can improve using only its own resources) to the set of competitive allocations. Debreu and Scarf (1963) give a formal statement of Edgeworth's assertion and show that it holds for economies with any (finite) number of commodities and consumers (assuming that consumers have strictly convex preferences). Formally: as the number of consumers grows the core coincides to the set of competitive allocations. Aumann (1964; 1966) constructs a formal limit model (with a continuum of consumers) and establishes (under quite general assumptions on preferences) that the core coincides with the set of competitive allocations.

Although these results are exceedingly elegant, the ‘real’ economy is finite and that replication and strict convexity of preferences are strong assumptions; hence neither Aumann's core equivalence theorem nor Debreu and Scarf's core convergence theorem applies directly to the ‘real’ economy. The results of Keiding (1974), Dierker (1975) and Anderson (1978) go a long way to removing this objection, showing that for *every* finite economy every core allocation can be approximately decentralized by some price, and that the deviation from exact decentralization (by some measures) is small provided the number of consumers is large and that no consumer is endowed with more than a small fraction of the social total of any good. For an excellent survey of the state of the art in the late 1980s, see Anderson (1993).

Although these (and related) results are usually accepted as cooperative justification for price-taking, more recent work reveals surprising subtleties.

1. 1. Aumann's core equivalence theorem for continuum economies assumes only that preferences are locally non-satiated (not necessarily monotone). It had been widely assumed that convergence theorems for finite economies should obtain under the same assumption. (The results of Debreu and Scarf, Keiding, Dierker, and Anderson assume that preferences are strictly monotone.) However, Manelli (1991a; 1991b) shows that core convergence may fail if preferences are not monotone, and Hara (2005) shows that further problems may arise if some commodities are bads.
2. 2. Most of the work on core convergence treats only exchange economies. Xiong and Zheng (2005) show that the validity of core convergence for production economies depends in a subtle way on smoothness of preferences, the presence or absence of boundary allocations, and especially on the interpretation of firm shares as control rights.
3. 3. The core is a cooperative solution notion based on blocking; the various *bargaining sets* use core logic, but impose more stringent requirements for blocking. (An allocation fails to be in the core there is a coalition C and an allocation g for C such that all members of C prefer g to f; we might say the coalition C has an *objection* to f. An allocation f fails to be in the *classical bargaining set* if there is a coalition C and an allocation g for C such that all members of C prefer g to f – so C has an objection to f – and in addition there is no coalition D and allocation h for D such that all members of D prefer h to f and all members of $C \cap D$ prefer h to g – so no coalition has a *counter-objection* to g.) The bargaining sets are larger than the core, so convergence of bargaining sets to the set of competitive allocations is a more stringent test of competition than is convergence of the core. Anderson (1998) establishes a convergence result for the classical bargaining set and a variant, but Anderson, Trockel and Zhou (1997) show that core convergence can fail for other bargaining sets. Convergence of the core is sometimes interpreted as the

inability of groups to manipulate the outcome; non-convergence of (some) bargaining sets casts doubt on this interpretation.

4. 4. If the number of commodities is fixed and finite and the number of traders is large, then there are many potential buyers and sellers of each good; markets are *thick*. Recent work on the core in economies with a continuum of agents and an *infinite* number of commodities highlight the importance of thick markets for perfect competition. If the commodity space is separable, then there are ‘many more agents than commodities’, and Aumann's core equivalence theorem obtains (Rustichini and Yannelis, 1991), but if the commodity space is sufficiently large then there can be ‘more commodities than agents’, and Aumann's theorem can fail (Tourky and Yannelis, 2001). Ostroy and Zame (1994) focuses on the extent to which commodities are good substitutes, which can be interpreted as ‘economic thickness’. If markets are economically thick then again Aumann's theorem obtains, but if markets are economically thin then Aumann's theorem can fail.
5. 5. When the commodity space is infinite dimensional and the number of traders is finite, the situation is even subtler. The Debreu–Scarf core convergence theorem holds (Aliprantis, Brown and Burkinshaw, 1985), but most of the obvious analogues of the general core convergence theorems of Keiding, Dierker and Anderson fail (Anderson and Zame, 1997). Thus there is a substantial difference between replica economies and general large finite economies; there can also be a substantial difference between large finite economies and continuum economies. Anderson and Zame (1997) identifiy two reasons for these differences: the first is that the integrability assumptions inherent in the continuum model, usually viewed as economically innocuous in the finite dimensional setting, impose economically serious restrictions in the infinite dimensional setting; the second is that various compactness properties that are inherent in the finite dimensional setting no longer obtain in the infinite dimensional setting. A different consequence of the latter fact is that there are well-behaved continuum economies for which the core is empty and no competitive equilibrium exists (Zame, 1986).

The work discussed above seeks to give cooperative justifications for price-taking behaviour; a different literature seeks to give non-cooperative justifications. For exchange economies, Rubinstein and Wolinsky (1985) propose a search model in which agents enter the market, meet and trade at random, leave the market to consume, and are replaced. They argue that the steady-state outcome of the bargaining game may differ from the competitive outcome. Rubinstein and Wolinsky focus on a setting in which the number of potential buyers is different from the number of potential sellers, but posit a replacement process in which agents who match and leave the market are replaced with exact duplicates. As a result, agents on the short side of the market are unable to exercise their market power. Gale (1986a; 1986b; 1987; 2000) offers different models, using different replacement processes, and shows that outcomes of the search/bargaining game (both in the steady state and not) *do* coincide with competitive outcomes.

For production economies, Allen and Hellwig (1986a; 1986b) argue that Bertrand price competition leads to approximately competitive prices and outcomes if there is a large number of firms supplying a competitive consumption sector. Cheng (2002) argues that, if firms are risk averse and costs are uncertain, then Cournot quantity competition leads to approximately competitive prices and outcomes

but Bertrand price competition may not.

Ostroy (1980; 1981) suggests a different approach to perfect competition, based more directly on the ability of individuals to influence prices or to favourably manipulate outcomes. Favourable manipulation will certainly be impossible from any outcome for which each individual already extracts his or her marginal product. A formal definition is easiest for finite economies with transferable utility. For each sub-coalition *S* of the set *N* of all agents, let *v*(*S*) be the maximal total utility obtainable by redistribution of the total endowment of *S*. (The assumption of transferable utility guarantees that this is economically sensible.) The marginal contribution of agent *i* to society is thus $v(N) - v(N \setminus i)$. If *x* is the vector of utilities of a feasible allocation, then agent *i* extracts his/her marginal product if $x_i = v(N) - v(N \setminus i)$; an allocation at which each agent extracts his or her marginal product is said to satisfy the *no-surplus* property. The appropriate extensions of these definitions to continuum economies use limits of small coalitions as proxies for individuals and derivatives as proxies for individual marginal products. For continuum economies with a finite number of goods, no-surplus is generic – but not universal. No-surplus is closely related to perfect elasticity of demand and supply (Ostroy, 1984) and to the inability of agents to manipulate; for some environments, these notions are coincident (Gretsky, Ostroy and Zame, 1999). Makowski and Ostroy (2001) provide an overview, bibliography and applications to innovation and mechanism design; Makowski (2004) gives a striking application to non-contractible investment and the ‘hold-up’ problem. See also perfect competition; Shapley value.

Equilibration

The definition of competitive equilibrium identifies a particular state of the economy but provides no clue as to the process by which the economy is to reach this state. Without such a process, competitive equilibrium may retain its usefulness as a benchmark (normative solution), but is in doubt as a description of reality (positive solution). Unfortunately, no such process has been described.

Walras suggested a simple and appealing process that he called *tâtonnement*: from any given price system *p*, adjust prices in proportion to excess demands. However, for some economies, the tâtonnement process does not converge (Scarf, 1960). Indeed, in view of the fact that, aside from the necessity of satisfying Walras's law, the excess demand function of an economy is essentially arbitrary (Sonnenschein, 1973; Mantel, 1974; Debreu, 1974), the tâtonnement process may follow an essentially arbitrary dynamic: converge from some initial prices, diverge from others, and cycle from still others. Although more complicated adjustment processes have been proposed, none seems economically sensible. Moreover, any adjustment process that is universally convergent must of necessity use an enormous amount of information: not just the excess demand of each good at each price, but the derivatives of excess demand of each good with respect to own price *and* the prices of other goods as well (Saari and Simon, 1978; Saari, 1985). An additional difficulty with tâtonnement is that, because prices adjust without trade, it does not seem to describe any process we see in the real world. (Walras imagined a fictitious ‘auctioneer’ who sets a tentative price, receives tentative demands, adjusts the tentative price, and so on, with the process continuing until excess demand for all goods is zero, at which point trade takes place.)

Keisler (1996) offers a suggestive model of price adjustment *with* trade. Consider a large finite

population of traders and a single warehouse. The warehouse manager announces an initial price p_0. At each time t thereafter, a single consumer, chosen at random, comes to the warehouse, trades at the current price p_t, and leaves the economy. (Thus, the trader has no incentive to misrepresent or to wait.) The warehouse manager adjusts prices in a direction proportional to the net trade of the most recent consumer, leading to a new price p_{t+1}, and the process continues. Keisler shows that, if the population is large, the price adjustments (the constant of proportionality) are small, and the initial price p_0 is in the basin of attraction of some price p^* that is a stable equilibrium for the Walrasian tâtonnement process, then with probability close to 1 the price path will reach, and eventually stay in, a small neighbourhood of p^* and most trades will take place at prices near p^*. (Because the warehouse manager adjusts prices following every trade, prices do not converge – but they do not leave a small neighbourhood of p^*.)

Given that the Walrasian tâtonnement process need have no stable equilibria, one might ask why Keisler's result is of interest. It should be kept in mind, however, that the Debreu–Mantel–Sonnenschein theorem describes only the *theoretical possibilities* for the aggregate excess demand function of an economy; it does not describe the aggregate excess demand function of any real economy. If the aggregate excess demand function of the real economy is – always or frequently – well-behaved, then Keisler's result provides hope for a sensible process that converges to equilibrium.

Of course it is not possible to observe the aggregate excess demand function of the real economy. Failing that, it seems natural to ask whether there are reasonable conditions on preferences and endowments – and especially on the *distribution* of preferences and endowments – that are compatible with empirical observation and also guarantee that the aggregate excess demand function is well-behaved (stable or locally stable for Walrasian tâtonnement or some other natural adjustment process). A sufficient condition for this to be true is that the economy admit a representative consumer, in the sense that the demand function of the economy is the demand function of a one-agent economy. (The tâtonnement process follows the differential equation $dp/dt=K[w - D(p)]$, where w is the aggregate endowment and $D(p)$ is aggregate demand at the price p, given endowment w. Fix any equilibrium price p^*; by definition, $D(p^*)=0$. At any non-equilibrium price p, w is affordable but not chosen, so w is dis-preferred to $D(p)$. By assumption, D is the demand function of a single agent, so revealed preference applies; hence $D(p)$ cannot be affordable at the equilibrium price p^*; that is, $p^*\cdot D(p)>0$. Walras's Law guarantees that $p\cdot D(p)=p\cdot w$ for all prices p. Taken together, these are enough to guarantee that the tâtonnement process converges from any initial price to the equilibrium price p^*.)

One promising approach focuses on the distribution of preferences and endowments and shows that aggregate market demand D obeys the *law of demand*; that is, $(p - p')\cdot(D(p) - D(p'))<0$. (Existence of a representative consumer is not enough to not guarantee the law of demand in the aggregate.) For example, if all agents have the same demand function, income is independent of price, the income density is decreasing and the smallest incomes are sufficiently small, then the law of demand will hold in the aggregate (Hildenbrand, 1983). These assumptions are strong but can be significantly weakened (Chiappori, 1985; Quah, 1997). Alternatively, if all agents have the same income, then sufficient heterogeneity of demand functions will also imply the law of demand in the aggregate (Grandmont, 1987); again, these assumptions are strong, but can be significantly weakened (Grandmont, 1992; Quah, 2002). See also aggregation (theory).

Incomplete markets

The standard Arrow–Debreu–McKenzie general equilibrium model posits a market for every commodity. Since the description of a commodity includes the date and state of nature in which it will be delivered, this entails markets for all claims to all goods in all future dates and states of the world. Radner (1972), building on a earlier model of Arrow (1953), offers an alternative model in which at each date and state of the world there are spot markets for commodities available at that date and state of the world and for *assets* or *securities* (state-contingent claims to wealth at future dates), but not for commodities at other dates and states. In this model, the transfer of wealth across time or across states of nature can be accomplished only by trading available assets. If all state- and time-dependent wealth transfers can be accomplished by trading available assets, then asset markets are *complete*, and the model reduces to the Arrow–Debreu–McKenzie model; in the alternative case, asset markets are *incomplete*.

In principle, asset payoffs (or *dividends*) in a given date-event may depend arbitrarily on commodity spot prices in that date-event, and even on the prices of other assets. Two particular kinds of securities are of special interest. *Financial assets* or *nominal assets* are those whose dividends are independent of prices; such assets are abstractions of real-world instruments such as treasury bills. *Real assets* are those whose dividends in a given date-event are the value at commodity spot prices of a specified bundle of commodities; such assets are abstractions of real-world instruments such as commodity forward contracts. (Most real-world forward contracts are *marked-to-market*; that is, they promise to deliver the *value* of a particular bundle of commodities rather than the physical bundle itself. In a perfectly competitive market, the distinction is unimportant, but in a real-world market the distinction can be significant if, as sometimes happens, the physical good promised is in sufficiently short supply that the promised quantity cannot be delivered.)

Radner (1972) establishes the existence of an asset market equilibrium (for either nominal or real assets), assuming an exogenously given bound on short sales. Such a constraint is unsatisfactory because if the constraint is binding then the equilibrium depends on an arbitrarily given, perhaps not economically meaningful, bound. For economies in which all assets are nominal, Cass (1984), Werner (1985) and Duffie (1987) show that short sale bounds are not necessary. The case of real assets is more subtle, because the possibilities for wealth transfer may depend on commodity spot prices. Suppose, for example, that trading takes place today and tomorrow, that there are two possible states – rainy and sunny – of the world tomorrow, and that there are only two assets, one promising delivery of (the value of) one bushel of wheat in each state, the other promising delivery of (the value of) one bushel of corn in each state. If the ratio of the price of wheat to the price of corn is different in the rainy state than in the sunny state, then the dividends of these assets are linearly independent, and the market is complete; if the ratio of the price of wheat to the price of corn is the *same* in the rainy state as in the sunny state, then these assets are collinear and the market is incomplete. In the former case, the space of wealth patterns that can be achieved by trading assets is two-dimensional; in the latter case, it is one-dimensional. In particular, consumers' budget sets are *discontinuous* functions of commodity spot prices. (For financial assets, dividends are by definition independent of prices, so this phenomenon cannot arise.) As Hart (1975) shows, this phenomenon leads to examples in which no equilibrium exists. However, Duffie and Shafer (1985; 1986) show that equilibrium does exist for *generic* values of the parameters. (The proof

uses degree theory, because familiar arguments based on the Kakutani fixed point theorem are not applicable. Elegant fixed point proofs were later discovered by Husseini, Lasry and Magill, 1990, and Hirsch, Magill and Mas-Colell, 1990. Geanakoplos and Shafer, 1990, gave an equally elegant homotopy argument.) For more complicated assets, such as options, equilibrium may fail to exist for an open set of parameters (Ku and Polemarchakis, 1990). The volume by Magill and Quinzii (1996) presents an excellent extended discussion and bibliography.

The incomplete markets model is attractive in part because it provides a framework in which to model and address many interesting economic phenomena and questions. For instance:

- When asset markets are incomplete, equilibrium commodity allocations need not be Pareto optimal. Indeed, Geanakoplos and Polemarchakis (1986) show that equilibrium commodity allocations will typically fail to be even constrained optimal: a social planner could improve welfare of all participants, even if constrained to use only existing asset markets to transfer wealth.
- Enlarging the set of available assets increases trading opportunities, so it is tempting to believe that it improves welfare. A surprising example due to Hart (1975) shows that enlarging the set of available assets may make *everyone* worse off; Elul (1995) and Cass and Citanna (1998) show that the possibility of such Pareto worsening is a robust phenomenon.
- By definition, the dividends of financial assets are independent of commodity spot prices – but the purchasing power of these dividends may depend on price levels. In a market with only financial assets, there is nothing to connect price levels in one date-event to price levels in another, so equilibrium asset prices and purchasing power are generally indeterminate; this leads to robust indeterminacy of equilibrium prices and consumptions as well (Balasko and Cass, 1989; Geanakoplos and Mas-Colell, 1989). For real assets, the purchasing power of dividends is independent of price levels, and equilibrium prices and consumptions are generically determinate (Geanakoplos and Polemarchakis, 1986).
- Default is suggestive of disequilibrium and inefficiency. Dubey, Geanakoplos and Shubik (2005) show, to the contrary, that default is compatible with equilibrium and may in fact promote welfare. Zame (1993) uses a similar framework to underscore the positive role played by default in expanding the effective span of assets when markets are incomplete.

A recurring theme in the study of asset markets is the importance of re-trading long-lived assets. The power of frequent trading underlies both the celebrated option-pricing formula of Black and Scholes (1973) and portfolio insurance (Leland, 1980). (Cox, Ross and Rubinstein, 1979, present a more easily understood discrete time version.) In a general discrete time model, Kreps (1982) argues that, if the number of long-lived assets is at least as great as the degree of uncertainty from one date-event to the next, then it will generically be possible to replicate any wealth pattern by frequent trading of a few long-lived assets. In particular, asset market equilibrium will, in this circumstance, coincide with complete markets equilibrium. Duffie and Huang (1985) identify an appropriate of Kreps's spanning condition in the continuous time setting (the setting most used in finance) and proves the corresponding generalization of Kreps's dynamic completeness result.

In an infinite-horizon setting, the existence of equilibrium when markets are incomplete requires ruling out the possibility of perpetual borrowing to pay each period's debts; the ‘right’ conditions (which may be expressed as transversality conditions, as limits on short sales, or as debt constraints) were identified (independently) by Magill and Quinzii (1994), Levine and Zame (1996) and Hernandez and Santos (1996).

The infinite-horizon setting is of particular interest because of its connection with asset-pricing. Mehra and Prescott (1985) show that historical returns on equity (stocks) cannot be reasonably explained within a complete-markets, infinite-horizon, asset-pricing model in the style of Lucas (1978). US data, (real) returns on safe assets (such as Treasury bills) are about one per cent and returns on equity are about seven per cent. With reasonable choices for time preference and risk aversion, the model suggests that returns on safe assets should be about two to three per cent and that returns on equity should be about three to four per cent; even extreme specifications of risk aversion do not yield an equity premium (that is, a rate of return on equity in excess of the return on safe assets) above two per cent. For overviews see Kocherlakota (1996) and Cochrane (2001).

A plausible objection to the complete markets assumption made by Mehra and Prescott is that labour income is not readily tradable; however, computationally tractable models with plausible parametrizations of untradable labour income do not appear to deliver a substantially higher equity premium (Telmer, 1993; Lucas, 1994; Heaton and Lucas, 1996). On the other hand, Constantinides and Duffie (1996) show that an *arbitrary* equity premium can be generated if enough income is untradable. More precisely, given an aggregate income process and *any* system of asset prices that are consistent with time preference, there is a distribution of untradable income for which the given prices constitute an equilibrium. The argument of Constantinides and Duffie relies on individual shocks that are permanent; whether that assumption is necessary and whether the true distribution of labour income (or other untradable income) is sufficient to generate the observed equity premium is a subject of considerable interest (Levine and Zame, 2002; Cogley, 2002; De Santis, 2005).

An interesting alternative explanation for the divergence of observed asset prices from theoretical predictions is that markets are complete but participation is not. Alvarez and Jermann (2000), building on a model of Kehoe and Levine (1993), explore the implications for asset pricing of an environment in which imperfect enforcement generates endogenously incomplete participation. See also incomplete markets.

Infinitely many commodities

The description of a commodity includes its physical characteristics and the date, location and state of the world at which it will delivered. If time is modelled as continuous or the horizon is modelled as infinite, if uncertainty is modelled by the use of an infinite state space, or if commodities are modelled as having a continuous range of possible characteristics, then the number of commodities will be infinite. Examples include:

- the use (especially in macroeconomics and asset pricing) of ℓ^{∞} (the space of bounded sequences) to model of consumption and trade over an infinite time horizon (Bewley, 1972;

Lucas, 1978; Mehra and Prescott, 1985);

- the use (especially in finance) of $L^2(\Omega, F, P)$ (the space of random variables with finite mean and variance on some probability space (Ω, F, P)) (and related spaces) to model choice and asset trading under uncertainty (Black and Scholes, 1973; Merton, 1973; Duffie and Huang, 1985);
- the use of $C[0, T]$ (the space of continuous functions on the interval $[0, T]$) or $L^\infty(\mathbb{R}_+)$ (the space of bounded functions on the non-negative real numbers) to model consumption and trade in a continuous-time framework (Gabszewicz, 1968; Bewley, 1972); and
- the use of $M(K)$ (the space of (signed) measures on some space K of commodity characteristics) to model finely differentiated commodities (Mas-Colell, 1975; Dixit and Stiglitz, 1977; Hart, 1979; 1985a; 1985b; Jones, 1984).

The Arrow–Debreu–McKenzie framework is powerful because it can be applied in many different economic environments, so it is natural to look for an extension of this framework which applies in the models above (and hopefully in many others).

In looking for such an extension, several problems arise. The most obvious problem is that neither budget sets nor feasible sets need be compact (in the given topology); thus the existence of optimal choices is immediately in doubt. An approach that avoids this difficulty, and is quite generally applicable, is to make use of the fact that most infinite dimensional vector spaces admit many topologies. For example, $L^\infty = L^\infty(\mathbb{R}_+)$ admits a norm topology (where $\|f\|$ is the essential supremum of $|f|$), but it also admits two weaker topologies that arise from viewing L^∞ as the space of continuous linear functionals on $L^1 = L^1(\mathbb{R}_+)$ (the space of integrable functions on $\mathbb{R}_+$): $\sigma(L^\infty, L^1)$ (the *weak topology*) is the weakest vector space topology on L^∞ for which L^1 is the space of continuous linear functionals, and $\tau(L^\infty, L^1)$ (the *Mackey topology*) is the strongest vector space topology for which L^1 is the space of continuous linear functionals. The weak topology is weaker than the Mackey topology which is in turn weaker than the norm topology (Dunford and Schwartz, 1957). The applicability to equilibrium analysis comes from three facts:

1. (i) in the weak topology, closed and bounded sets are compact;
2. (ii) convex sets that are closed in the Mackey topology are also closed in the weak topology;
3. (iii) for preference relations, continuity in the Mackey topology can be interpreted as impatience.

In view of (iii), Mackey continuity of preferences is an economically meaningful and natural assumption; in view of (ii), preferences that are convex and Mackey continuous are also weakly upper hemi-continuous; in view of (i), such preferences admit optimal choices whenever feasible sets are closed and bounded. (This interplay between topologies is a familiar functional-analytic theme.)

The second problem that arises is that sensible-looking preference relations may not admit supporting prices. When the commodity space is finite-dimensional and preferences are convex, the separation theorem guarantees that a supporting price at a consumption bundle x can be constructed as a linear functional separating x from the set $p(x)$ of bundles strictly preferred to x. When the commodity space is infinite dimensional, the separation theorem only guarantees the existence of a linear functional

separating x from $p(x)$ in case $p(x)$ has non-empty interior. In many spaces, this is problematical because the positive cone (the most natural consumption set) has empty interior, whence, *a fortiori*, the strictly preferred set also must have empty interior.

Mas-Colell (1986) shows by example that supporting prices need not exist, and identifies a class of preferences for which supporting prices *do* exist. Say that a preference relation defined on the positive cone X_+ of some commodity space X is *uniformly proper* if there is some vector $v \in X_+$ and some open cone C containing v such that for each $x \in X_+$ every bundle in $(x - C) \cap X_+$ is strictly dis-preferred to x. (The non-existence of supporting prices may be interpreted as unboundedness of marginal rates of substitution; uniform properness may be interpreted as a bound on marginal rates of substitution.) For commodity spaces that are topological vector lattices (ordered topological vector spaces in which every pair of elements have an infimum and a supremum and in which the lattice operations are continuous), Mas-Colell shows that uniform properness is the crucial additional assumption needed to guarantee the existence of competitive equilibrium for exchange economies with a finite number of agents. Mas-Colell (1986) and Zame (1987) extend the existence theorem to include production economies, introducing (necessary) additional conditions that bound marginal rates of transformation as well as marginal rates of substitution. Mas-Colell and Richard (1991) offer a different proof that makes weaker assumptions by focusing more on the lattice structure of the price space and less on the lattice structure of the commodity space; this is important in a number of applications. Mas-Colell and Zame (1991) and Aliprantis, Brown and Burkinshaw (1989) provide good surveys of the existence theorems, including discussion of a number of different proof strategies.

A surprising aspect of the analysis in the infinite dimensional setting is that it relies heavily on the order structure of the commodity and prices spaces and on the assumption that consumption sets coincide with the positive cone of the commodity space, which play no role in the finite dimensional setting. That the order structures should play an important role is suggested by Aliprantis and Brown (1982), but a foreshadowing can already be seen in Bewley (1972). In that paper, the role of the order structure is not to guarantee the existence of equilibrium, but to guarantee the existence of equilibrium with economically meaningful prices (that is, prices in ℓ^1 rather than in the full dual of ℓ^∞). Bewley's argument rests on the possibility of decomposing a socially feasible bundle dominated by the social endowment into a sum of individually feasible bundles dominated by individual endowments. That this is possible is a consequence of the *Riesz decomposition property*, which holds when the commodity space is a vector lattice and consumption sets are the positive cone, but not for general commodity spaces or consumption sets. The arguments used by Mas-Colell to construct prices that support a Pareto optimal allocation, and by Yannelis and Zame (1986) to provide estimates on supporting prices, make similar use of the Riesz decomposition property and so again require that the commodity space be a lattice. For explorations of equilibrium theory when assumptions on the order structure are relaxed, see Aliprantis, Tourky and Yannelis (2001) and Aliprantis, Florenzano and Tourky (2005). See also functional analysis.

Hidden information and hidden actions

The standard model of competitive markets treats an environment in which all agents are equally

informed about economically relevant parameters, and has nothing to say about environments in which agents are asymmetrically informed – but even casual observation suggests that the latter environments are more common than the former.

That asymmetric information can have an enormous impact on market outcomes is pointed out forcefully in Akerlof's (1970) seminal discussion of used-car markets. Because sellers typically have private information about their own cars, such markets display *adverse selection*: low-quality cars are offered for sale more readily than are high-quality cars, so the quality distribution of cars offered for sale at a given price will be skewed downward in comparison with the overall distribution of cars in the market. As a result, market outcomes may be less efficient than in the case where all information is public; in extreme situations, only autarkic outcomes (no trade) may obtain, even though every potential buyer values every car more than its original owner.

Adverse selection arises in many other markets as well. Potential borrowers may know more about their creditworthiness than do potential lenders, owners/operators of a productive firm may know more about its future profitability than potential investors, and potential buyers of insurance may know more about their accident or health risks than do potential sellers of insurance. In the insurance context, Rothschild and Stiglitz (1976) argue that adverse selection may become so important that equilibrium does not exist. (But Rothschild and Stiglitz use a mixed, and not strictly price-taking, notion of equilibrium.)

The work of Akerlof makes it clear that asymmetric information may matter for market outcomes; the work of Rothschild and Stiglitz makes it clear that asymmetric information may matter for the way we model markets as well. Following these seminal contributions, a large literature has sought to integrate asymmetric information with general equilibrium modelling of competitive markets. Central issues in this literature include: which models are appropriate for which kinds of asymmetric information? Does the operation of the market reveal information? If so, how? And how much?

Radner (1968) develops a model of an environment in which agents learn their private information (modelled as an information partition over true states of the world) and make contingent trades *before* the true state of the world occurs. In this model, private information acts as a constraint on choices: each agent can choose only among state-contingent consumption bundles that are measurable with respect to his/her private information; call such bundles *private*. A *Walrasian expectations equilibrium* consists of economy-wide prices and private consumption bundles for each agent such that each agent's bundle is optimal among private, budget-feasible consumptions, and markets clear in each state of the world. If free disposal is permitted, standard conditions guarantee that Walrasian expectations equilibrium exists. For these environments, Walrasian expectations equilibrium can be justified as a descriptive theory in much the same way that Walrasian equilibrium is justified for symmetric information environments: as the limit of a cooperative solution. For these asymmetric information environments, the appropriate cooperative notion, the *private core*, consists of private allocations (vectors of private consumption choices) which have the property that no coalition can construct an improving private allocation, using only its own resources (Yannelis, 1991; Allen, 1994; 2003. Aliprantis, Tourky and Yannelis (2001), Einy, Moreno and Shitovitz (2003) and Hervés-Beloso, Moreno-Garcia and Yannelis (2005) show that versions of Debreu and Scarf's (1963) core convergence theorem and Aumann's (1964) core equivalence theorem obtain for the private core and Walrasian expectations equilibrium.

For environments in which trade takes place *after* the true state of nature occurs but before it is publicly known, agents may use their own private information *and* draw inferences from market activities, such

as prices (Radner, 1979). A *rational expectations equilibrium* consists of a (state-dependent) price function and an allocation such that each agent's state-dependent consumption bundle is measurable with respect to the join of his or her own information and the information in prices, each agent optimizes among measurable budget-feasible bundles, and markets clear in each state of the world. A simple example (Kreps, 1979) shows that some simple economies do not admit any rational expectations equilibrium. However, for generic specifications of economic parameters (endowments, information and preferences or utility functions) there is always a rational expectations equilibrium in which all information is fully revealed (Allen, 1981). In the presence of noise, such as uncertainty about aggregate endowments, equilibrium may be partially, but not fully, revealing Admati, 1991). Rational expectations equilibrium forms the theoretical basis of a cornerstone of finance, the *efficient markets hypothesis*, which asserts that all information is revealed in prices.

An alternative view of rational expectations equilibrium, as a rest point of a rational, but imperfect, learning process, is offered by Anderson and Sonnenschein (1985). In their framework, each agent has an exogenously given (linear) model of the world and the economy, and chooses parameters of the model to best fit the observed data. At equilibrium (a rest point of the process of fitting parameters to data), each agent's model is best fitting but not necessarily correct. At such an equilibrium, agents may not have learned everything, but they have learned all that is possible for them to learn, given their models. Bossaerts (2002) addresses the econometric implications of this kind of rational, but imperfect, learning process.

A criticism of rational expectations equilibrium is that it does not address the mechanism through which agents obtain their private information. This seems an important omission because, if all information were to be revealed by prices, there would seem to be no incentive for agents to acquire information in the first place, especially if acquiring information is costly (Grossman and Stiglitz, 1980). A second criticism of rational expectations equilibrium is that extracting information from prices seems to require agents to have a great deal of information about the economy (including information about other agents); when equilibrium is fully revealing, for example, agents must be able to invert the map from states of the world to equilibrium prices. Perhaps the most serious criticism of rational expectations equilibrium is that it provides no process by which information gets into prices. If agents use information in prices in forming their demands, how do those demands influence prices? If demands do not influence prices, where do prices come from?

A very fruitful approach to the revelation of private information takes as its starting point the observation that private information gives rise to incentive problems: if private information is valuable, agents may not wish to take actions that reveal that private information. However, incentive problems may not matter much if agents are *informationally small*. (A new-car variation of Akerlof's familiar used-car market provides an intuitive idea of what it means to be informationally small. Suppose that all cars have the *same* true quality, unknown to both buyers and sellers, but that sellers receive noisy signals of this common quality. If there are many sellers, and sellers' signals are conditionally independent, then the marginal amount of information revealed by the signal of a given seller is small. Put differently: in an economy with many sellers, each seller's signal has little effect on the true posterior information about quality.) This idea can be formalized in a number of different ways. For instance, Gul and Postlewaite (1993) and McLean and Postlewaite (2002) describe classes of environments for which allocations close to competitive allocations are incentive compatible in large replications. McLean and

Postlewaite (2005) define an *incentive-compatible core* and show that incentive compatible core allocations in replica economies are close to competitive allocations of the full information economy provided that agents are informationally small. Forges, Heifetz and Minelli (2001) prove a related convergence result in a different formulation that includes lotteries.

Prescott and Townsend (1984a; 1984b) pioneered an approach that treats certain kinds of private information and incentive issues within the standard general equilibrium paradigm. Prescott and Townsend model shocks (hidden information) as affecting preferences. Incentive problems (moral hazard) arise in guaranteeing that agents who experience a particular shock should have no incentive to misrepresent themselves as having experienced a different shock; these incentive constraints are incorporated into consumption sets. Two unusual aspects of the model is that objects of choice are *lotteries* over consumption bundles and that prices are linear in probabilities but not necessarily in consumption. Competitive equilibria exist and are Pareto optimal within the class of allocations that satisfy the incentive constraints. (Allowing for lotteries guarantees that consumption sets and preferences are convex, so that familiar Arrow–Debreu existence results can be applied. It does not seem commonly observed that, in the continuum of agents framework Prescott and Townsend adopt, convexity of consumption sets and preferences is not necessary to guarantee the existence of an equilibrium.)

If enforcement of contractual arrangements is imperfect, then issues of moral hazard and adverse selection arise in financial markets as well. Borrowers may choose to default on promises (moral hazard) and borrowers who are poor risks or less affected by sanctions are more likely to default than are borrowers who are good credit risks (adverse selection). Surprisingly, neither of these interferes with the existence equilibrium, provided that deliveries on asset promises are *pooled* (Dubey, Geanakoplos and Shubik, 2005). The most familiar pooled financial instruments are collateralized mortgage obligations, which are pools of individual mortgages. Pooling mortgage deliveries spreads the default risk across all lenders; absent pooling, each lender would face the idiosyncratic risk that individual borrowers default against them. Although default is suggestive of deadweight losses, allowing for default may be Pareto-improving when markets are incomplete, because it expands the effective span of available assets (Dubey, Geanakoplos and Shubik, 2005; Zame, 1993).

Bisin et al. (2002) argue that, if *all trade* – even trade for commodities – is carried out through contracts, and deliveries on all contracts are pooled, then both adverse selection and moral hazard can be accommodated within almost standard general equilibrium models. (The assumption that deliveries on commodity contracts are pooled would seem natural for orange juice, but not for used cars; the assumption that deliveries on financial contracts are pooled would seem natural for collateralized mortgage obligations but not for individual mortgages.) Dubey and Geanakoplos (2002) use a similar idea to reformulate the insurance economy of Rothschild and Stiglitz (1976) and argue that an equilibrium always exists.

All of the work described above considers either economies in which individuals consume by themselves and production (if any) is of the standard Arrow–Debreu–McKenzie type. A growing literature treats economies in which individuals consume and/or produce in small groups (teams or firms). Prescott and Townsend (2006), Rahman (2005) and Song (2006) treat general equilibrium models with team production. Output is a function of observable investment and unobservable effort, which creates a moral hazard problem. Working in the tradition of Prescott and Townsend, these papers

find institutional arrangements that permit the decentralization of incentive-efficient configurations. When production is deterministic and utility is transferable, the requisite institutions include contract arbitrageurs and Lindahl (personalized) prices for team membership; when production is stochastic and utility is not transferable, public randomization devices and assets whose payoffs depend on the distribution of idiosyncratic uncertainty are required as well.

Zame (2005) takes a different approach, adding hidden information and hidden actions to the clubs framework of Ellickson et al. (1999; 2006). In that model, firm output and individual utility depend on skills and actions of other agents, which are unobservable and uncontractable; thus there is scope for both adverse selection and moral hazard. Moreover, because output within firms depends on action profiles, agents are subject to idiosyncratic risk. The set of firms that form and the contractual arrangements that appear, the assignments of agents to firms, the prices faced by firms for inputs and outputs, and the incentives to agents are all determined endogenously at equilibrium. Agents choose consumption – but they also choose which firms to join, which roles to occupy in those firms, and which actions to take in those roles. Agents interact anonymously with the (large) market, but strategically within the (small) firms they join. The model accommodates moral hazard, adverse selection, signalling and insurance. Equilibrium allocations may be incentive efficient and even Pareto ranked.

See Also

- adverse selection
- aggregation (theory)
- efficient markets hypothesis
- functional analysis
- incomplete markets
- perfect competition
- Shapley value

Bibliography

Admati, A. 1991. A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica* 53, 629–58.

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Aliprantis, C.D. and Brown, D.J. 1982. Equilibria in markets with a Riesz space of commodities. *Journal of Mathematical Economics* 11, 189–207.

Aliprantis, C.D., Brown, D.J. and Burkinshaw, O. 1985. Edgeworth equilibria. *Econometrica* 55, 1109–38.

Aliprantis, C.D., Brown, D.J. and Burkinshaw, O. 1989. *Existence and Optimality of Competitive Equilibria*. New York: Springer.

Aliprantis, C.D., Florenzano, M. and Tourky, R. 2005. General equilibrium analysis in ordered topological vector spaces. Working Paper, CERSEM.

Aliprantis, C.D., Tourky, R. and Yannelis, N.C. 2001. A theory of value with non-linear prices: equilibrium analysis beyond vector lattices. *Journal of Economic Theory* 100, 22–72.

Allen, B. 1981. Generic existence of completely revealing equilibria for economies with uncertainty when prices convey information. *Econometrica* 49, 1173–99.

Allen, B. 1994. Market games with asymmetric information: the core with finitely many states of the world. In *Models and Experiments in Risk and Rationality*, ed. B. Munier and M.J. Machina. Dordrecht: Kluwer.

Allen, B. and Hellwig, M. 1986a. Price-setting firms and the oligopolistic foundations of perfect competition. *American Economic Review* 76, 387–92.

Allen, B. and Hellwig, M. 1986b. Bertrand–Edgeworth oligopoly in large markets. *Review of Economic Studies* 53, 175–204.

Allen, B. 2003. Incentives in market games with asymmetric information: the core. *Economic Theory* 21, 527–44.

Allen, B. and Yannelis, N. 2001. Differential information economies: introduction. *Economic Theory* 18, 263–73.

Alvarez, F. and Jermann, U.J. 2000. Efficiency, equilibrium, and asset pricing with risk of default. *Econometrica* 68, 775–97.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46, 1483–7.

Anderson, R.M. 1993. The core in perfectly competitive economies. In *Handbook of Game Theory with Economic Applications*, vol. I, ed. R. Aumann and S. Hart. Amsterdam: North-Holland.

Anderson, R.M. 1998. Convergence of the Aumann–Davis–Maschler and Geanakoplos bargaining sets. *Economic Theory* 11, 1–37.

Anderson, R.M. and Sonnenschein, H. 1985. Rational expectations equilibrium with econometric models. *Review of Economic Studies* 52, 359–69.

Anderson, R.M., Trockel, W. and Zhou, L. 1997. Nonconvergence of the Mas-Colell and Zhou bargaining sets. *Econometrica* 65, 1227–39.

Anderson, R.M. and Zame, W.R. 1997. Edgeworth's conjecture with infinitely many commodities: L^1. *Econometrica* 65, 225–74.

Anderson, R.M. and Zame, W.R. 1998. Edgeworth's conjecture with infinitely many commodities: differentiated commodities. *Economic Theory* 11, 331–77.

Anderson, R.M. and Zame, W.R. 2001. Genericity with infinitely many parameters. *Advances in Theoretical Economics* 1, 1–62.

Araujo, A. 1987. The non-existence of smooth demand in general Banach spaces. *Journal of Mathematical Economics* 17, 1–11.

Arrow, K. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie, Colloques Internationaux du C.N.R.S.* 40, 41–7.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Aumann, R.J. 1966. Equilibrium in markets with a continuum of traders. *Econometrica* 34, 1–17.

Balasko, Y. and Cass, D. 1989. The structure of financial equilibrium with exogenous yields: the case of incomplete markets. *Econometrica* 57, 135–62.

Bennardo, A. and Chiappori, P.A. 2003. Bertrand and Walras equilibria under moral hazard. *Journal of Political Economy* 111, 785–817.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 43, 514–40.

Bisin, A., Geanakoplos, J., Gottardi, P., Minelli, E. et al. 2002. Markets and contracts. Discussion paper, Cowles Foundation.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 837–54.

Blume, L. and Zame, W.R. 1993. The algebraic geometry of competitive equilibrium. In *Essays in General Equilibrium and International Trade: In Memoriam Trout Rader*, ed. W. Neuefeind. New York: Springer.

Bossaerts, P. 2002. *The Paradox of Asset Pricing*. Princeton: Princeton University Press.

Cass, D. 1984. Competitive equilibria in incomplete financial markets. Working Paper No. 84-09, CARESS, University of Pennsylvania.

Cass, D. and Citanna, A. 1998. Pareto improving financial innovation in incomplete markets. *Economic Theory* 11, 467–94.

Cheng, H.C. 1996. Values of perfectly competitive economies. In *Handbook of Game Theory with Economic Applications*, ed. R.J. Aumann and S. Hart. New York: Elsevier North-Holland.

Cheng, H. 2002. Bertrand vs. Cournot equilibrium with risk averse firms and cost uncertainty. *Economic Theory* 20, 555–77.

Chiappori, P.-A. 1985. Distribution of income and the ‘law of demand’. *Econometrica* 53, 109–28.

Cochrane, J. 2001. *Asset Pricing*. Princeton: Princeton University Press.

Cogley, T. 2002. Idiosyncratic risk and the equity premium: evidence from the consumer expenditure survey. *Journal of Monetary Economics* 49, 309–34.

Cole, H., Mailath, G. and Postlewaite, A. 2001. Efficient non-contractible investments in large economies. *Journal of Economic Theory* 101, 333–73.

Cole, H. and Prescott, E.C. 1997. Valuation equilibrium with clubs. *Journal of Economic Theory* 74, 19–39.

Constantinides, G. and Duffie, J.D. 1996. Asset pricing with heterogeneous traders. *Journal of Political Economy* 104, 219–40.

Cox, J.C., Ross, S.A. and Rubinstein, M. 1979. Option pricing: a simplified approach. *Journal of Financial Economics* 7, 229–63.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40, 603–15.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–23.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

De Santis, M. 2005. Interpreting aggregate stock market behavior: how far can the standard model go? Working paper, Dartmouth College.

Dierker, E. 1975. Equilibria and core of large economies. *Journal of Mathematical Economics* 2, 155–69.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Dubey, P. and Geanakoplos, J. 2002. Competitive pooling: Rothschild-Stiglitz reconsidered. *Quarterly Journal of Economics* 117, 1529–70.

Dubey, P., Geanakoplos, J. and Shubik, M. 2005. Default and punishment in general equilibrium. *Econometrica* 73, 1–38.

Duffie, J.D. 1987. Stochastic equilibria with incomplete financial markets. *Journal of Economic Theory* 41, 405–16. Corrigendum, 1989. *Journal of Economic Theory* 49, 384.

Duffie, J.D. and Huang, C.-F. 1985. Implementing Arrow–Debreu equilibria by continuous trading of few long-lived securities. *Econometrica* 53, 1337–56.

Duffie, J.D. and Shafer, W. 1985. Equilibrium in incomplete markets I: a basic model of generic existence. *Journal of Mathematical Economics* 14, 285–300.

Duffie, J.D. and Shafer, W. 1986. Equilibrium in incomplete markets II: generic existence in stochastic economies. *Journal of Mathematical Economics* 15, 199–216.

Dunford, N. and Schwartz, J.T. 1957. *Linear Operators*. New York: Interscience.

Einy, E., Moreno, D. and Shitovitz, B. 2003. Competitive and core allocations in large economies with differential information. *Economic Theory* 18, 321–32.

Ellickson, B., Grodal, B., Scotchmer, S. and Zame, W.R. 1999. Clubs and the market. *Econometrica* 67, 1185–217.

Ellickson, B., Grodal, B., Scotchmer, S. and Zame, W.R. 2006. The organization of production, consumption and learning. In *Institutions, Equilibria and Efficiency: Essays in Honor of Birgit Grodal*, ed. C. Schultz and K. Vind. Berlin: Springer.

Elul, R. 1995. Welfare effects of financial innovation in incomplete markets economies with several consumption goods. *Journal of Economic Theory* 65, 43–78.

Forges, F., Heifetz, A. and Minelli, E. 2001. Incentive compatible core and competitive equilibria in differential information economies. *Economic Theory* 18, 349–65.

Gabszewicz, J. 1968. Coeurs et allocations concurrentielles dans les économies d'exchange avec un continu de bains. Librarie Universitaire, Université Catholique de Louvain.

Gale, D. 1986a. Bargaining and competition. Part I: characterization. *Econometrica* 54, 785–806.

Gale, D. 1986b. Bargaining and competition. Part II: existence. *Econometrica* 54, 807–18.

Gale, D. 1987. Limit theorems for markets with sequential bargaining. *Journal of Economic Theory* 43, 20–54.

Gale, D. 2000. *Strategic Foundations of General Equilibrium: Dynamic Matching and Bargaining Games*. Cambridge: Cambridge University Press.

Geanakoplos, J. and Mas-Colell, A. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47, 22–38.

Geanakoplos, J. and Polemarchakis, H. 1986. Existence, regularity and constrained suboptimality of competitive allocations when the asset market is incomplete. In *Uncertainty, Information and Communication: Essays in Honor of K.J. Arrow*, vol. 3, ed. W. Heller, R. Starr and D. Starrett. Cambridge: Cambridge University Press.

Geanakoplos, J. and Shafer, W. 1990. Solving systems of simultaneous equations in economics. *Journal of Mathematical Economics* 19, 69–93.

Geanakoplos, J. and Zame, W.R. 2002. Collateral and the enforcement of intertemporal contracts. Working paper, UCLA.

Ghosal, S. and Polemarchakis, H.M. 1997. Nash–Walras equilibria. *Ricerche Economiche* 51, 31–40.

Grandmont, J.-M. 1987. Distribution of preferences and the ‘law of demand’. *Econometrica* 55, 155–61.

Grandmont, J.-M. 1992. Transformations of the commodity space, behavioral heterogeneity, and the aggregation problem. *Journal of Economic Theory* 57, 1–35.

Gretsky, N., Ostroy, J.M. and Zame, W.R. 1999. Perfect competition in the continuous assignment model. *Journal of Economic Theory* 88, 60–118.

Grossman, S. and Stiglitz, J. 1980. On the impossibility of informationally efficient markets. *American*

Economic Review 70, 393–408.

Gul, F. and Postlewaite, A. 1993. Asymptotic efficiency in large exchange economies with asymmetric information. *Econometrica* 60, 1273–92.

Hara, C. 2005. Existence of equilibria in economies with bads. *Econometrica* 73, 647–58.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11, 418–43.

Hart, O. 1979. Monopolistic competition in a large economy with differentiated commodities. *Review of Economic Studies* 46, 1–30.

Hart, O. 1985a. Monopolistic competition in the spirit of Chamberlin: a general model. *Review of Economic Studies* 52, 529–46.

Hart, O. 1985b. Monopolistic competition in the spirit of Chamberlin: special results. *Economic Journal* 95, 889–908.

Hart, S. and Mas-Colell, A. 1996. Bargaining and value. *Econometrica* 64, 357–80.

Heaton, J. and Lucas, D.J. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104, 433–67.

Hernandez, D.A. and Santos, M.S. 1986. Competitive equilibria for infinite-horizon economies with incomplete markets. *Journal of Economic Theory* 71, 102–30.

Hervés-Beloso, C., Moreno-Garcia, E. and Yannelis, N.C. 2005. An equivalence theorem for a differential information economy. *Journal of Mathematical Economics* 41, 844–56.

Hildenbrand, W. 1983. On the ‘law of demand’. *Econometrica* 51, 997–1020.

Hirsch, M., Magill, M. and Mas-Colell, A. 1990. A geometric approach to a class of equilibrium existence problems. *Journal of Mathematical Economics* 19, 95–106.

Husseini, S.Y., Lasry, J.M. and Magill, M. 1990. Existence of equilibrium with incomplete markets. *Journal of Mathematical Economics* 19, 39–67.

Jones, L. 1984. A competitive model of commodity differentiation. *Econometrica* 52, 507–30.

Kehoe, T.J. and Levine, D.K. 1985. Comparative statics and perfect foresight in infinite horizon

economies. *Econometrica* 53, 433–52.

Kehoe, T.J. and Levine, D.K. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60, 885–8.

Kehoe, T.J., Levine, D.K., Mas-Colell, A. and Zame, W.R. 1989. Determinacy of equilibrium in large-square economies. *Journal of Mathematical Economics* 18, 231–63.

Keiding, H. 1974. A limit theorem on the core of large but finite economies. Working paper, University of Copenhagen.

Keisler, J. 1996. Getting to a competitive equilibrium. *Econometrica* 64, 29–49.

Kocherlakota, N. 1996. The equity premium: it's still a puzzle. *Journal of Economic Literature* 34, 42–71.

Kreps, D. 1979. Three essays on capital markets. Technical Report No. 298. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Kreps, D. 1982. Multiperiod securities and the efficient allocation of risk: a comment on the Black–Scholes option pricing model. In *The Economics of Uncertainty and Information*, ed. J. McCall. Chicago: University of Chicago Press.

Ku, B.-I. and Polemarchakis, H. 1990. Options and equilibrium. *Journal of Mathematical Economics* 19, 107–12.

Leland, H. 1980. Who should buy portfolio insurance? *Journal of Finance* 35, 581–94.

Levine, D.K. and Zame, W.R. 1996. Debt constraints and equilibrium in infinite horizon economies with incomplete markets. *Journal of Mathematical Economics* 26, 103–31.

Levine, D.K. and Zame, W.R. 2002. Does market incompleteness matter? *Econometrica* 70, 1805–39.

Lucas, D.J. 1994. Asset pricing with undiversifiable income risk and short sales constraints: deepening the equity premium puzzle. *Journal of Monetary Economics* 34, 325–41.

Lucas, R.E. 1978. Asset pricing in a pure exchange economy. *Econometrica* 46, 1429–45.

Magill, M.J.P. and Quinzii, M. 1994. Infinite horizon incomplete markets. *Econometrica* 62, 853–80.

Magill, M.J.P. and Quinzii, M. 1996. *Theory of Incomplete Markets*, vol. 1. Cambridge: MIT Press.

Magill, M.J.P. and Quinzii, M. 1997. Which improves welfare more: a nominal or an indexed bond? *Economic Theory* 10, 1–37.

Makowski, L. 2004. Pre-contractual investment with the fear of holdups: the perfect competition connection. Working paper, UC Davis.

Makowski, L. and Ostroy, J.M. 2001. Perfect competition and the creativity of the market. *Journal of Economic Literature* 39, 479–535.

Manelli, A. 1991a. Monotonic preferences and core equivalence. *Econometrica* 59, 123–38.

Manelli, A. 1991b. Core convergence without monotonic preferences or free disposal. *Journal of Economic Theory* 55, 400–15.

Mantel, R. 1974. On the characterization of aggregate excess demand functions. *Journal of Economic Theory* 7, 348–53.

Mas-Colell, A. 1975. A model of equilibrium with differentiated commodities. *Journal of Mathematical Economics* 2, 263–95.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: A Differentiable Approach*. Cambridge: Cambridge University Press.

Mas-Colell, A. 1986. The price equilibrium existence theorem in topological vector lattices. *Econometrica* 54, 1039–54.

Mas-Colell, A. 1986. Valuation equilibrium and Pareto optimum revisited. In *Contributions to Mathematical Economics*, ed. W. Hildenbrand and A. Mas-Colell. New York: North-Holland.

Mas-Colell, A. and Richard, S.F. 1991. A new approach to the existence of equilibria in vector lattices. *Journal of Economic Theory* 53, 1–11.

Mas-Colell, A. and Zame, W.R. 1991. Equilibrium theory in infinite-dimensional spaces. In *Handbook of Mathematical Economics*, vol. IV, ed. W. Hildenbrand and H. Sonnenschein. New York: Elsevier.

McLean, R. and Postlewaite, A. 2002. Informational size and incentive compatibility. *Econometrica* 70, 2421–54.

McLean, R. and Postlewaite, A. 2005. Core convergence with asymmetric information. *Games and Economic Behavior* 50, 58–78.

Mehra, R. and Prescott, E.C. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 14, 145–61.

Merton, R.C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41, 867–87.

Minehart, D.F. 1997. Generic finiteness of the set of equilibria in a finite exchange economy. *Mathematical Social Sciences* 34, 75–80.

Minelli, M. and Polemarchakis, H.M. 2000. Nash–Walras equilibria of a large economy. *Proceedings of the National Academy of Sciences* (*USA*) 97, 5675–8.

Ostroy, J.M. 1980. The no-surplus condition as a characterization of perfectly competitive equilibrium. *Journal of Economic Theory* 22, 183–207.

Ostroy, J.M. 1981. Differentiability as convergence to perfectly competitive equilibrium. *Journal of Mathematical Economics* 8, 59–73.

Ostroy, J.M. 1984. A reformulation of the marginal productivity theory of distribution. *Econometrica* 52, 599–630.

Ostroy, J.M. and Zame, W.R. 1994. Non-atomic economies and the boundaries of perfect competition. *Econometrica* 62, 593–633.

Prescott, E.C. and Townsend, R.M. 1984a. Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica* 52, 21–45.

Prescott, E.C. and Townsend, R.M. 1984b. General competitive analysis in an economy with private information. *International Economic Review* 25, 1–20.

Prescott, E.S. and Townsend, R.M. 2006. Clubs as firms in Walrasian economies with private information. *Journal of Political Economy* 114, 644–71.

Quah, J. 1997. The law of demand when income is price dependent. *Econometrica* 65, 1421–42.

Quah, J. 2002. The monotonicity of individual and market demand. *Econometrica* 68, 911–30.

Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 31, 31–58.

Radner, R. 1972. Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica* 40, 289–303.

Radner, R. 1979. Rational expectations equilibrium: generic existence and the information revealed by prices. *Econometrica* 47, 655–78.

Rahman, D. 2005. Contractual pricing with incentive constraints. Working paper, UCLA.

Rothschild, M. and Stiglitz, J. 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90, 630–49.

Rubinstein, A. and Wolinsky, A. 1985. Equilibrium in a market with sequential bargaining. *Econometrica* 53, 1133–50.

Rustichini, A. and Siconolfi, P. 2002. General equilibrium in economies with adverse selection. Working paper, University of Minnesota.

Rustichini, A. and Yannelis, N. 1991. Edgeworth's conjecture in economies with a continuum of agents and commodities. *Journal of Mathematical Economics* 20, 307–26.

Saari, D. 1985. Iterative price mechanisms. *Econometrica* 53, 1117–32.

Saari, D. and Simon, C. 1978. Effective price mechanisms. *Econometrica* 46, 1097–125.

Scarf, H. 1960. Some examples of global instability of the competitive equilibrium. *International Economic Review* 1, 157–71.

Shannon, C. 1994. Regular nonsmooth equations. *Journal of Mathematical Economics* 23, 147–66.

Shannon, C. 1999. Determinacy of competitive equilibria in economies with many commodities. *Economic Theory* 14, 29–87.

Shannon, C. and Zame, W.R. 2002. Quadratic concavity and determinacy of equilibrium. *Econometrica* 70, 631–62.

Shapley, L. 1969. Utility comparison and the theory of games. In *La Décision*. Paris: éditions du CNRS.

Song, J. 2006. Contractual matching: limits of decentralization. Working paper, UCLA.

Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6, 343–54.

Telmer, C.I. 1993. Asset-pricing puzzles and incomplete markets. *Journal of Finance* 48, 1803–33.

Tourky, R. and Yannelis, N.C. 2001. Markets with many more agents than commodities: Aumann's ‘hidden’ assumption. *Journal of Economic Theory* 101, 189–221.

Werner, J. 1985. Equilibrium in economies with incomplete financial markets. *Journal of Economic Theory* 36, 110–9.

Xiong, S. and Zheng, C. 2005. Core equivalence theorem with production. Working paper, Northwestern University.

Yannelis, N.C. 1991. The core of an economy with differential information. *Economic Theory* 1, 183–98.

Yannelis, N.C. and Zame, W.R. 1986. Equilibria in Banach lattices without ordered preferences. *Journal of Mathematical Economics* 15, 85–110.

Zame, W.R. 1986. Economies with a continuum of traders and infinitely many commodities. Working paper, SUNY–Buffalo.

Zame, W.R. 1987. Competitive equilibria in production economies with an infinite dimensional commodity space. *Econometrica* 55, 1075–108.

Zame, W.R. 1993. Efficiency and the role of default when security markets are incomplete. *American Economic Review* 83, 1142–64.

Zame, W.R. 2005. Incentives, contracts and markets: a general equilibrium theory of firms. Working paper, UCLA.

How to cite this article

Zame, William. "general equilibrium (new developments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000187> doi:10.1057/9780230226203.0624

The New Palgrave Dictionary of Economics Online

general equilibrium

Lionel W. McKenzie

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Unlike partial equilibrium theory, general equilibrium theory treats as constant only non-economic influences and embraces all sales and purchases of all agents involved in exchanges. It implies that all subsets of agents are in equilibrium and that all individual agents are in equilibrium. The development of a formal general equilibrium theory in mathematical terms was initiated in the 19th century by Walras, who moved from a model of an exchange economy to an equilibrium with production. It was completed in the 1950s by McKenzie, who formalized Walrasian theory, and by Arrow and Debreu, who formalized Hicksian theory.

Keywords

Arrow, K.; Arrow–Debreu model; competitive equilibrium; convexity; Cournot, A.; Debreu, G.; equilibrium of production; equilibrium over time; existence of general equilibrium; general equilibrium; Hicks, J.; Jevons, W.; Marshall, A.; Menger, C.; Mill, J. S.; monetary equilibrium; overlapping generations model of general equilibrium; Pareto efficiency; Pareto, V.; partial equilibrium; Poinsot, L.; rational expectations; Samuelson, P.; Smith, A.; temporary equilibrium; von Neumann's Law; Wald's Law; Walras, L.; Walras's Law; Walras's theory of investment

Article

General equilibrium theory is in contrast with partial equilibrium theory where some specified part of an economy is analysed while the influences impinging on this sector from the rest of the economy are held constant. In general equilibrium the influences which are treated as constant are those which are considered to be noneconomic and thus beyond the range of economic analysis. Of course, this does not guarantee that these influences will in fact remain constant when the economic factors change, and the usefulness of economic analysis for predictive purposes may depend on to what degree influences treated as noneconomic are really independent of the economic variables.

The institution whose phenomena are the primary subject matter of economic analysis is the market, made up of a group of economic agents who buy and sell goods and services to one another. In partial equilibrium theory the group of agents may be confined to those who are involved in one industry, either buying or selling its product or buying or selling the materials and productive services used in making its product. However, in general equilibrium theory all the agents involved in exchanges with each other should ideally be included and all their sales and purchases should be allowed for. However, it may happen that the activities of many agents are only treated in the aggregate and the list of goods and services may be reduced by aggregation. The aggregation of agents and commodities into a few categories is especially important when general equilibrium theory is applied to special areas of public policy such as the government budget, money and banking, or foreign trade. Much of the theory developed for these subjects is general equilibrium theory in aggregated form.

The general equilibrium implies that all subsets of agents are in equilibrium and in particular that all individual agents are in equilibrium. The conscious development of a formal general equilibrium theory stated in mathematical terms seems to have been inspired by a formal theory of the equilibrium of the individual consumer faced with a given set of trading opportunities or prices. This theory was developed by the marginal utility, or neo-classical, school of economists in the third quarter of the 19th century, independently, by Gossen (1854), Jevons (1871), and Walras (1874–7), who used mathematical notations, and by Menger (1871) who did not. The step was taken in the most effective way by Walras.

The equilibrium of an exchange economy

Walras assumed that the utility derived from the consumption of a good was given as a function of the amount of that good alone that was consumed and independent of the amounts consumed of other goods. He also assumed that the first derivative of the utility function was positive and decreasing up to a point of satiation when one exists. He then gave a rigorous derivation of the demand for a good by a consumer from the maximization of utility subject to a budget constraint. The demand functions give the equilibrium quantities traded by the consumer as a function of market prices. As Walras saw, this is a crucial step in the development of a general equilibrium theory for an economy. It has remained in a generalized form the cornerstone of general equilibrium theory since Walras.

The simplest problem of general equilibrium arises in the theory of the exchange economy without production. In this economy the budget constraint of the trader is established by his initial stocks and the list of prices. Then the individual demand function represents the equilibrium of the single trader in face of a given price system. The market demand function is the sum of the individual demand functions, and the equilibrium of the market occurs at a price for which the sum of demands, including offers as negative demands, is equal to 0 for each good, or, if free disposal is allowed, is not positive for any good. This idea was expressed in classical economic theory by the equality of supply and demand in each market, but its expression in a set of equations to be satisfied by the list of equilibrium prices was due to Walras, although Cournot (1838) had foreshadowed the Walrasian analysis in his discussion of the international flow of money and Mill (1848) in his discussion of foreign trade.

Suppose there are n goods to be traded and there are m traders. Let w_i^h be the quantity of the ith good held initially by the hth trader. Let $u^h(x)$ where $x=(x_1•,\ldots,•x_n)$ be the utility to the hth trader of possessing

the quantities $x_1,\ldots,x_n$, of the n goods traded. Then the hth trader is in equilibrium at the prices $p=(p_1\bullet, \ldots,\bullet p_n)$ and the quantities x^h if $u^h(x)$ is a maximum at x^h over all values of x which satisfy $\Sigma_1^n p_i x_i \leq \Sigma_1^n p_i w_i^h$. If smoothness and concavity conditions are met by the utility function, and the goods are divisible, the maximizing x will be unique and will define a function $f^h(p)$ over an appropriate price domain. Since the set of commodity bundles x at which the utility function is maximized does not change when the prices p are multiplied by a positive scalar, this function will satisfy $f^h(p) = f^h(\alpha p)$ for $\alpha > 0$.

The market demand function is $f(p) = \Sigma_1^m f^h(p)$. Then the market equilibrium for a trading economy is given by a price vector p and an allocation of goods $(x^1,\ldots, x^m)$ such that $x^h = f^h(p)$ and $\Sigma_1^m x^h = \Sigma_1^m w^h$, or, assuming free disposal, $\Sigma_1^m x^h \leq \Sigma_1^m w^h$. The first condition expresses the equilibrium of the individual trader and the second condition is the equality of supply and demand. Thus there are n scalar equations $\Sigma_{h=1}^m f_i^h\left(p = \Sigma_{h=1}^m w_i^h\right)$ to determine the n equilibrium prices p_i. The given data are the consumer tastes, expressed in the utility functions u^h, and the initial stocks of goods w^h.

It is clear that the market demand function satisfies the homogeneity condition $f(\alpha p) = f(p)$ for $\alpha \geq 0$. Thus equilibrium prices are only determined up to multiplication by a positive number. This reflects the fact that the equilibrium of the consumer is not affected if prices are multiplied by α and market equilibrium is the simultaneous equilibrium of all consumers at the same prices. It is often convenient to adopt some normalization of prices. Walras chooses a good whose price is known to be positive in equilibrium and gives this good, which he calls the numeraire, the price 1. Another convention which is useful when free disposal is assumed, so that prices are necessarily non-negative, is to choose p such that $\Sigma_1^n p_i = 1$. Then the domain of definition for the demand functions may be taken to be all p such that $p_i \geq 0$ and $\Sigma_1^n p_i = 1$.

There is an analogy between the equilibrium of the trading economy and the equilibrium of mechanical forces. Indeed, one of the inspirations for the theory of Walras appears to have been a treatise on statics by Poinsot (1803, 1842). According to the principle of virtual work an infinitesimal displacement of a mechanical system, which is at equilibrium under the stress of forces and subject to constraints, does no work. In the economy at equilibrium an infinitesimal displacement of the allocation of goods $(x^1,\ldots, x^m)$ cannot increase the utility of one trader unless it reduces the utility of another. This is an easy implication of the fact that utility is maximized over the budget constraint, provided no one is saturated. This means that a new allocation to a trader cannot preserve his utility level if its value at the equilibrium prices falls. On the other hand, the utility level of a trader cannot increase unless his allocation becomes more valuable at the equilibrium prices. But then the new allocations $x^{h\prime}$ would satisfy

$$\sum_{h=1}^{m}\sum_{i=1}^{n} p_i x_i^{h\prime} > \sum_{h=1}^{m}\sum_{i=1}^{n} p_i w_i^h$$

which is impossible since the total allocation cannot exceed the total supply of goods. Indeed, if each trader holds all goods in his equilibrium allocation and the utility functions are differentiable, which implies that goods are divisible, an infinitesimal reallocation would have no effect on utility levels if it has no effect on the levels of individual budgets. This property of market equilibrium was first recognized by Pareto (1909), and an allocation of goods with the property that no displacement of it can benefit one consumer unless it harms another is said to be Pareto optimal. The implication from competitive equilibrium to Pareto optimality requires that no consumer be locally satiated. It is also true that a Pareto optimal allocation may be realized as a competitive equilibrium given an appropriate distribution of initial stocks but the conditions are more severe. The first general theorems were proved by Arrow (1951).

Equilibrium with production

The next step in developing the general equilibrium of an economy is to introduce production under the condition that the output matures without a lapse of time. This step was taken by Walras, who introduced linear activities which list the quantities of productive services required to produce one unit of a good. There may be many alternative activities for the production of any given good and a choice is made among them in order to minimize the cost of production at given market prices. Let $z=(z_1, \ldots, z_r)$ be a list of quantities of productive services and let $g^i(z)$, $i=1, \ldots, n$, be production functions for the n goods. Since linear activities are assumed, the production functions will satisfy $\alpha g^i(z) = g^i(\alpha z)$. In particular, we may consider the unit isoquant or the set A_i such that $g^i(z) = 1$ for z in A_i. Then the activities which minimize cost at given prices q are represented by production coefficients $a^i(q)$, contained in A_i, where $q' a^i(q) \leq q' z$ for z in A_i. Equilibrium in the production sector is given by price vectors p and q and activity vectors $a^i(q)$ where $p_i \leq \Sigma_{j=1}^r q_j a_j^i(q)$ for all i and equality holds if the ith good is produced.

In an equilibrium of the production sector any quantities y of outputs may be produced provided quantities z of productive services are available where $z_j = \Sigma_{i=1}^n y_i a_j^i(q)$. In order to include the productive sector in a market equilibrium the utility functions of consumers must be extended to include productive services among their arguments. They may be written $u^h(x,\bullet z)$. If we reinterpret x_i as the quantity of a good traded rather than the quantity consumed, the initial stocks may be suppressed. This is convenient since it is not clear how initial stocks of labour services can be specified. Then the individual consumer is in equilibrium given prices p and q for goods and productive services when the quantities traded (x^h, z^h) maximize $u^h(x,\bullet z)$ over all $(x,\bullet z)$ such that $\Sigma_1^n p_i x_i - \Sigma_1^r q_i z_i \leq 0$. The maximizing quantities need not be unique in general, so it is necessary to represent demand by a correspondence that takes a set of trades as its value and write $(x^h, z^h) \in f^h(p, q)$ when (x^h, z^h) is a maximizer given prices p and q.

As before, market equilibrium is achieved when all economic agents are in equilibrium at the same prices and supply is equal to demand. Since risk is not present in this economy, the productive services

involved in organizing production need not be given a distinguished role. Activities may be treated as conducted by the whole set of owners of the productive services involved in them. Then if it should happen that $p_i > \Sigma_{j=1}^{r} q_j a_j^i(q)$ for the *i*th good, there will be an opportunity for some owners of productive services to earn larger returns producing the *i*th good than those prevailing generally as given by *q*. Thus productive services will leave other activities and flow to this activity, so equilibrium does not obtain for owners of productive services. This equilibrium now requires, on the one hand, equilibrium of each economic agent as consumer of goods and provider of productive services, that is, $(x^h, z^h) \in f^h(p, q)$, and, on the other hand, equilibrium of each economic agent as a participant in production, that is, $p_i \leq \Sigma_{j=1}^{r} q_j a_j^i(q)$, with equality if the *i*th good is produced. However, market equilibrium also requires that $\Sigma_{h=1}^{m} z_j^h = \Sigma_{h=1}^{m} \Sigma_{i=1}^{r} x^h a_j^i(q)$, that is, the supply of productive services must equal the quantities needed to produce the quantities of goods demanded. As before, if surplus productive services may be freely disposed of, the equality in the last equation may be replaced by an inequality.

The demand functions $f_i^h(p, q)$ and the supply functions $f^h_{n+j}(p,\bullet q)$ express the equilibrium of the household sector. Therefore, the relation $\Sigma_{i=1}^{n} p_i f_i^h(p, q) = \Sigma_{j=1}^{r} q_j f_{n+j}^h(p, q)$ holds for all values of *p* and *q* in the price domain. Let x_i be the amount of the *i*th good produced and let z_j be the amount of the *j*th factor used in production. Then equilibrium in the production sector implies that

$$\sum_{i=1}^{n} p_i x_i = \sum_{j=1}^{r} \sum_{i=1}^{n} q_j x_i a_j^i = \sum_{j=1}^{r} q_j z_j.$$

Let $f(p, q) = \Sigma_{h=1}^{m} f^h(p, q)$. Then household equilibrium implies $\Sigma_{i=1}^{n} p_i f_i(p, q) = \Sigma_{j=1}^{n} q_j f_{n+j}(p, q)$. Let excess demand for a good be $e_i(p, q) = f_i(p, q) - x_i$, and excess demand for a productive service be $e_{n+j}(p, q) = z_j - f_{n+j}(p, q)$. Then equilibrium in the production and household sectors together implies that $\Sigma_{i=1}^{n} p_i e_i(p, q) + \Sigma_{j=1}^{n} q_j e_{n+j}(p, q) = 0$, or the value of excess demand is zero whatever price system is set. This relation is referred to as Walras's Law.

If there is free disposal, prices must be non-negative. Otherwise, disposal would be profitable. Also with free disposal the condition for equilibrium of the market is $e(p, q) \leq 0$. Then Walras's Law immediately implies $p_i e_i(p, q) = q_j e_{n+j}(p, q) = 0$, and if any good or productive service is in excess supply in equilibrium, its equilibrium price must be 0. This might be termed Wald's Law, since he made crucial use of it in the first rigorous proof that equilibrium exists in a competitive economy (Wald, 1935, 1936a).

A production sector composed of activities with single outputs is the model used by Walras, who was responsible for the first fully developed general equilibrium theory. The natural generalization of this

model is to introduce more than one output. Then the *k*th activity is represented by an output vector $b^k = \left(b_1^k, \ldots, b_n^k\right)$ and an input vector for productive services $a^k = \left(a_1^k, \ldots, a_r^k\right)$. Assume that activities may be replicated and are independent of each other. Then if (a^k, b^k) is a possible input-output combination for the *k*th activity, so is (αa^k, αb^k) where α is any non-negative integer. Indeed, if all inputs and outputs are divisible it is possible for α to take as its value any real number.

This model of the production sector which embraces the transformation of productive services into goods and services is due to Walras in the context of a theory of general equilibrium. It is convenient to think of the market as held periodically to arrange for the delivery of goods and services over a certain basic period of time. This view of the market, which is also a device of Walras, leads to a theory of temporary equilibrium. The theory was further elaborated by Hicks (1939) and in recent years by other authors. In order to explain the demand and supply of products and productive services in the periodic market it is necessary to introduce some assumptions on the formation of expectations for the prices which will prevail in future markets. The simplest assumption is that the prices arrived at in one market are expected to prevail in future markets. This type of expectation formation is sometimes referred to as static expectations. Walras usually appears to assume static expectations. Hicks introduced a notion of elasticity of expectations to allow expectations of future prices to depend on the change of prices from one temporary equilibrium to another. In recent work analysis has proceeded upon more general assumptions, using various formal properties of dependencies between past prices and expected prices. A quite different approach to expectations which enjoys much current popularity is to assume that expectations are correct, at least in a stochastic sense. The rationale of this approach is that any persistent bias in forecasts of future prices implies that there are unexploited opportunities for profit from further trading which eventually should be recognized.

The model of the production sector as a set of potential linear activities was subsequently used by Cassel (1918) in a simplified Walrasian model which preserved the demand functions and the production coefficients but which did not deduce the demand functions from utility functions or preferences. The model was generalized to allow joint production in a special context by von Neumann (1937). It was given a thorough elaboration and analysis in a model where intermediate products are introduced explicitly by Koopmans (1951). In the Walrasian picture intermediate products were eliminated through the combination of activities so that activities were described as transforming productive services directly into final products whether consumer goods and services or capital goods. However, such a description of the economy depends for its relevance on prices which do not change from one temporary equilibrium to another, so that the choice of activities is not changing.

In the general linear model of production it is no longer adequate to treat the choice of activities as a process of cost minimization given the price vectors *p* and *q*. Cost minimization must be replaced by the condition that no activity may offer a profit and no activity which is used in competitive equilibrium may suffer a loss. This is exactly the condition *‘ni benefice ni perte’* which Walras used to define equilibrium in production, initially in a model with fixed coefficients of production. However, this condition was first used in a general production model by von Neumann, so it might be termed von Neumann's Law for an activities model of production. Koopmans explored the relation between efficient production and von Neumann's Law. He established an equivalence between the proposition that an output is efficient and the proposition that prices exist such that von Neumann's Law is satisfied when the activities used are those needed to produce this output. Moreover, if each good or service is either

desired in unlimited quantities or freely disposable the prices must be non-negative. Thus under these demand and supply conditions any competitive equilibrium must include an efficient output from the production sector. The activities approach to the production sector of a competitive economy was used by Wald and then by McKenzie (1954) in proofs of existence for competitive equilibrium. It was also used by Scarf (1973) in an algorithm for finding a competitive equilibrium given the technology, the resources, and the demand functions.

An alternative model of the production sector emphasizes the productive organization or firm rather than the activities or technology. A set of actual or potential firms is given and each firm is endowed with its own set of possible input-output combinations. The set of possible input-output combinations achievable by the economy, independently of resource availabilities, is the sum of the sets of input-output combinations achievable by the firms. The condition for equilibrium in the production sector is that each firm maximizes its profits, that is, the value of the input-output combination over its production possibility set, given the prices of inputs and outputs. This view of production was explicit in a partial equilibrium context in Cournot. It was at least implicit in the work of Marshall (1890) and Pareto, and became quite explicit in a general equilibrium context in the work of Hicks (1939) and Arrow and Debreu (1954).

In the Hicksian model a firm is associated with each economic agent who is a consumer and who may be a worker and owner of resources, but who also may be an entrepreneur. As an entrepreneur he owns a possible production set based on his personal characteristics and perhaps some other non-marketed resources. Of course, most of these individual enterprises will be inactive. A difficulty with this model is that it seems unrealistic to treat the entrepreneur as a profit maximizer unless all the resources which he himself supplies have market prices so that they could equally well be bought by him from the market or sold by him to the market. But if that is the case we are back to the concept of the entrepreneur used by Walras and it seems more realistic to refer to activities, which are impersonal, rather than to individual enterprises.

In the model of Arrow and Debreu, which is the first complete general equilibrium model in which the existence of equilibrium was rigorously proved, the production sector is made up of firms which are described as joint stock companies. Each firm has a production possibility set based on resources which it owns and the ownership of the firm is spread in a prescribed way over a set of consumers. The production sector is in equilibrium when each firm has chosen an input-output combination from its production possibility set which maximizes profit at the market prices. Since the outputs of one firm may be inputs of another and the resort to integrated activities which convert productive services directly into products is not available in a model based on firms, it is convenient to distinguish inputs from outputs by signs rather than by lists. Let Y_j denote the production possibility set of the jth firm, and let $y=(y_1 \bullet,\ldots,\bullet y_n)$ denote an element of this set. There are n goods and services in the economy, and $y_i<0$ denotes an input, while $y_1>0$ denotes an output. Let y^j be the input-output vector of the jth firm. Then equilibrium in the production sector requires that the condition $p \cdot y^j \geq p \cdot y$ for all $y \in Y_j$ holds for all j, where j indexes the set of firms, and p is the market price vector.

The Arrow–Debreu approach to the production sector involves a major difficulty. It is not well adapted to handle the formation of new firms and the dissolution of old ones. If firms are based on the assembly of a set of resources jointly owned by the shareholders, it becomes critical to give the principle which underlies such an assembly. If the firm's resources are priced and traded, so the firm's production may be

treated like an activity, there is no difficulty since von Neumann's Law may be applied. Otherwise, the rules governing the entry and exit of firms are unclear. The problem is similar to the general problem of coalition formation in the theory of cooperative games.

A formal model

A formal model of the competitive economy, presented in the form of a series of axioms, was developed in the 1950s. It was intended that the axioms should be interpretable to apply to real economic systems, albeit in some approximate sense. However, as a formal mathematical model the implications of the axioms could be developed independently of the applications. The selection of axioms was influenced by the possibility of making useful interpretations, but also by the facility with which results can be derived.

Two closely related sets of assumptions were developed. One, developed primarily by McKenzie (1959), is a formalization of the Walrasian theory and uses a linear model of production. The other, developed primarily by Arrow and Debreu, is a formalization of the Hicksian theory where the production sector is described as an assembly of firms. On the side of consumers and the market there are no significant differences at a fundamental level, although there are sometimes differences of approach. A history of the problem of existence of equilibrium for the formal models may be found in Weintraub (1983).

In the fully developed McKenzie model (see McKenzie, 1981) two assumptions are made for the consumption sector, two for the production sector, and two assumptions relate the consumption and production sectors. On the consumption side there is a finite number m of consumers indexed by h, and each consumer has a set X_h of trades which are feasible for him. There are n goods and the sets X_h are contained in R^n, the n dimensional Euclidean space. The convention is used that quantities supplied by consumers are negative and quantities received by consumers are positive. The consumer has preferences defined on X_h by a correspondence P_h. The preference correspondence P_h takes as its value at $x \in X_h$ the subset of X_h each of whose members is preferred to x. This subset may be empty. The assumptions on the consumers which hold for all h, are.

1. (1) X_h is convex, closed and bounded below.
2. (2) P_h is open valued relative to X_h and lower semi-continuous. Also x is not in convex hull P_h (x).
   Convexity of X_h implies that a good is divisible if someone can consume it in more than one quantity. X_h bounded below means that the consumer is not able to supply an indefinite quantity of any good.
   Closedness and boundedness are needed to provide compact feasible sets.
   On the production side there is an activities model with no limitation on the number of activities. The activities are linear and give rise to a possible production set Y contained in R^n. If $y \in Y$, the negative components of y denote quantities of inputs and the positive components denote quantities of outputs. The assumptions on Y are.
3. (3) Y is a closed convex cone.

4. (4) $Y \cap R^n_+ = \{0\}$. R^n_+ is the set of non-negative vectors in R^n.
   That Y is a convex cone is equivalent to the production set being generated by linear activities. It means that if y and y' are producible, that is, elements of Y, then $\alpha y + \beta y'$ is also producible, that is, an element of Y, for any non-negative numbers α and β. Thus producible goods are divisible. Closedness is needed for the compactness of the feasible set. Assumption (4) is not restrictive. It is a recognition that goods which are never scarce are irrelevant to problems of economizing.
   Finally two assumptions relate the consumption sector and the production sector. Let X be the total possible consumption set, that is, $X = \Sigma_{h=1}^m X_h$. The first relation is
5. (5) relative interior $X \cap$ relative interior $Y \neq \varnothing$.
   Here the relative interior of a set is relative to the smallest linear subspace that contains it. This assumption insures that someone has income at any price vector which is consistent with equilibrium in the production sector, that is, satisfies von Neumann's Law. The second relation is an assumption that the economy is irreducible. Let I_1 and I_2 refer to nonempty subsets of consumers such that $I_1 \cup I_2$ includes all consumers and $I_1 \cap I_2 = \varnothing$. Let $X^1 = \Sigma X_h$ for $h \in I_1$, and similarly for I_2. Let $\bar{X}_h$ be the convex hull of X_h, and the origin of R^n. The irreducibility assumption is
6. (6) however I_1 and I_2 may be selected, if $x^1 = y - x^2$ with $x^1 \in X^1$, $y \in Y$ and $x^2 \in \bar{X}^2$, then there is also $\tilde{y} \in Y$ and $w \in \bar{X}^2$, such that $\tilde{x}^1 = \tilde{y} - x^2 - w$ and $\tilde{x} \in P(x^h)$ for all $h \in I_1$.

Assumption (6) guarantees that everyone has income if anyone has income. The meaning of having income is that the consumer is able to reduce his spending at the market price vector below the cost of his allocation and remain within his possible consumption set X_h.

Competitive equilibrium is defined by a price vector p, an output vector y, and vectors $x^1, \ldots, x^m$ of consumer trades. There is equilibrium in the production sector if von Neumann's Law holds, that is:

1. (I) $y \in Y$ and $p \cdot y = 0$, and for any $y' \in Y$, $p \cdot y' \leq 0$.
   When y satisfies (I) it is not possible for the owners of inputs to withdraw them from activities where they are being used and employ them in other activities, whether in use or not, so that the receipts from the resulting outputs allow some inputs to earn larger returns while none of them earns less. This is the same condition for equilibrium in production that was given by Walras, or, for that matter, by Adam Smith (1776).
   There is equilibrium in the consumer sector if the x^h satisfy
2. (II) $x^h \in X_h$ and $p \cdot x^h \leq 0$, and $p \cdot z > 0$ for any $z \in P_h(x^h)$, $h = 1, \ldots, m$.
   When x^h satisfies condition (II), there is no preferred bundle of goods, including goods or services that are supplied by the consumer, which is available to him under his budget constraint. This is essentially the same condition used by Walras, except that he assumed that preferences could be represented by a strictly concave utility function. Thus he is able to refer to

maximization of the utility function over the budget set uniquely at x^h.
Finally, there is market equilibrium when

3. (III) $\Sigma_{h=1}^{m} x^h = y$.

This is the condition that markets clear which was used by Walras.
If there is free disposal, Wald's Law may be derived directly from equilibrium in the production sector. The possibility of free disposal is recognized by the inclusion of disposal activities in the production cone, that is, an activity y^i for $i = 1, \ldots, n$ which has $y_i^i = -1$ and $y_j^i = 0$ for $j \neq i$. The condition $p \cdot y^i \leq 0$ implies that $p_i \geq 0$ must hold. Then if disposal occurs the condition $p \cdot y^i = 0$ implies that $p_i = 0$.
On the basis of Assumptions 1 through 6 it is possible to prove that a competitive equilibrium exists. This was first achieved in a model with assumptions for the demand sector put directly on preferences, in the manner of Walras, by Arrow and Debreu. At the same time McKenzie proved existence for a model with assumptions put on the demand functions rather than directly on preferences. Also McKenzie assumed a linear technology rather than a set of firms. This was a generalization of a model of Wald in which joint production was absent and the very special assumption was made that the market demand functions satisfied the weak axiom of revealed preference. The weak axiom says if x is demanded at p and x' at p', then $p \cdot x' \leq p \cdot x$ implies that $p' \cdot x < p' \cdot x$. This is a consistency requirement on choice under budget constraints. Wald's assumption was a deep insight. He anticipated the statement of this principle by Samuelson (e.g. 1947) who applied it to the demand of the individual consumer to derive most of the propositions of demand theory. Wald showed that the weak axiom assumed for the market leads to uniqueness of equilibrium. Subsequently it was shown by Arrow and Hurwicz (1958) that the weak axiom is implied by the assumption that all goods are gross substitutes. They also proved that the weak axiom confined to a comparison of choices between the equilibrium prices and other prices implies the global stability of a process of price adjustment in which the prices of goods are increased if excess demand exists and lowered if excess supply exists. Wald (1936b) wrote another paper on equilibrium in an exchange market which used assumptions closer to those of Arrow and Debreu, but this paper unfortunately was lost.
The only important distinction between the approach of Arrow and Debreu (see Debreu, 1962) and the approach expressed in Assumptions 1 through 6 is the use of a set of firms rather than a set of activities to generate the production set. Mathematically, through the introduction of entrepreneurial factors the approaches can be reconciled. However, the intentions of the two approaches are quite different. The linear model is intended to represent free entry into any line of production by cooperating factors, however organized in a legal sense, where economies of scale are sufficiently small to allow approximate linearity to be achieved by the multiplication of producing units. The lumpiness which is present is compared to that resulting from goods which are in fact indivisible, although they are treated as divisible. This leads to a reasonable approximation to real markets only if units are small compared with the levels of trade. This view of the competitive economy is consistent with the analysis of Marshall as well as Walras. Of course, it has to be recognized that in real economies some sectors cannot be approximated in this way. However, when linearity becomes a bad approximation to the production sector, convexity has in all likelihood become an equally bad approximation to the

production sets of firms.

Recently an explicit modelling of the approach of the firms economy to the activities economy has been given by Novshek and Sonnenschein (1980). They use the model of quantity adjusting firms developed in a partial equilibrium context by Cournot to find an equilibrium for the firms economy. Then they let the firm size shrink and show that the Walrasian equilibrium of an activities economy is approached in the limit.

Two interpretations of the formal model

Two basic interpretations of the general equilibrium model were described by Hicks and referred to as the spot economy and the futures economy. The spot economy is a market held on ‘Monday’ at which all transactions are arranged that involve delivery during the ‘week’. This is the economy described by Walras. The equilibrium of the spot economy is called temporary equilibrium in the modern literature. Some effort has been devoted to an analysis of the path followed by such an economy through a succession of temporary equilibria. The role of expectations in the spot economy is critical, as Hicks recognized.

The futures economy on the other hand has a single market in which all future transactions are negotiated at once. Hicks does not treat this economy in detail, but turns to a sequence of spot markets with trading that is guided by expectations. In the futures economy goods available in different periods would be treated as different goods, so that the number of goods would be finite only if the economic horizon is finite. If there is perfect foresight the futures economy is a reasonable alternative and there is not reason why markets should reopen. However, when the future is uncertain and the available futures contracts are for sure delivery, or at least do not exist in sufficient variety to take account of all contingencies, there is no assurance that the contracts entered into will remain desirable or indeed can be executed. For this reason Hicks chose to do a dynamic analysis of a sequence of temporary equilibria in the main body of his work.

In order to avoid the problem of the feasibility of plans and the need to reopen markets, Debreu (1959) following a lead of Arrow (1953) introduced a specification of goods by the event in which they are made available. The set of events would have to discriminate all the circumstances that might make delivery impossible or undesirable, so there would be no motive for traders to reopen markets. Despite this complexity, it is a consistent model which may have relevance to the real world. In order to keep the set of goods finite they assume a finite horizon and a finite set of events, in addition to assuming a finite list of goods in terms of location and physical characteristics.

With this interpretation of the formal model there is no room for borrowing and lending since payments are cleared only once, at the beginning of time. Uncertainty is present since there is no assumption that the event realized at any future time is known. Rather it will be revealed when the time arrives. There is no reason for spot markets to arise since the transactions which have been made for the future event that is revealed are the ones each trader desired at the prices paid in those circumstances. Thus if a spot market were opened no transactions would take place.

Of course it is idealization to suppose that all relevant events could be described in advance, or, if they could, that it would be feasible to establish markets discriminating between them. An alternative is to use a succession of markets in which temporary equilibria are established while some trading in futures contracts takes place. However, the limiting cases of the pure spot economy or the pure futures economy

have an analytical tractability that the mixed cases lack and for this reason they remain of great importance.

Temporary equilibrium

Once a sequence of markets is contemplated, rather than a single comprehensive market, plans for future trades become relevant and, therefore, expectations of the prices at which they can be made. Also money stocks and loans become useful in making financial preparations for the trading that is planned. Also, if there may be forward trading as well as spot trading, arbitrage is possible, and speculative trading arises which expresses disagreement among consumers about probable price levels on future spot and forward markets.

These complications were handled by Walras without an explicit analysis of demand by consumers for goods in the future using utility functions in which these goods appear. Rather he reduces the demand for future goods to a demand for assets in general which would provide the means for future purchases. On the other hand, he carefully distinguishes between stocks of goods and their services, and the investments of the consumer are treated as if they were made directly in the stocks of goods whose services are sold to the entrepreneurs, or directly to consumers in the case of services of consumer goods. The spirit of this analysis is to choose a period short enough that it is not too great a distortion of reality to suppose that all trades for this period can be concluded in advance as in the Arrow–Debreu model for the entire horizon, but the forms of industrial organization are abstracted from, so that attention may be concentrated on the productive activities and the ultimate beneficial owners of the resources whose services are used in them. Also to give the future some role in the decisions of the consumers but not a role requiring detailed analysis, Walras assumed that present market prices are expected to persist. In contrast, Hicks and Arrow–Debreu deal explicitly with intertemporal planning by firms and consumers. In a succession of markets this allows Hicks to analyse the effects of changes in expectations on the present market prices and the plans of agents.

The theory of Walras provides the most complete and detailed model of temporary general equilibrium that has ever been given, an impressive performance since it was also the first formal model of general equilibrium. He was able to deal with money, production, lending, and capital accumulation, and in his model an interest rate, price levels, and prices of capital goods and their services are all determined. He showed that the system was not overdetermined, and probably not underdetermined either, in that the number of independent functional relationships and the number of economic quantities to be determined are equal. He was not able to give a proof that an equilibrium in non-negative real variables exists for his model. However, proofs have since been given for simplified versions of it.

A fundamental difference between temporary equilibrium and equilibrium over a horizon is that part of the consumers’ demand for goods in the temporary equilibrium is intended for investment rather than for consumption within the period while in the economy of the classical existence theorem consumers’ demand is entirely aimed at consumption within the horizon. This raises two problems. One is to distinguish between resources devoted to this period's consumption and resources reserved for the support of consumption in future periods. The other is to explain how the decision to reserve a certain quantity of resources for future use is made. Walras went further to make the distinction between current and future use than any of his successors. They, on the other hand, have done much more analysis of the relation between investment and expectations. The Walrasian assumption on expectations was usually to

project the prices arrived at in the current market into the future. This assumption is only appropriate for a stationary, or a steadily progressive, state of the economy. Of course, it has often been remarked that it is only in these conditions that expectations are likely to be correct.

Walras distinguished between consumption goods and services which are consumed in one use and consumption goods which are in effect capital goods providing consumer services, that is, having more than one use. Among the consumption goods which serve as capital goods he included consumption goods which are held in stocks to provide, as Walras put it, services of availability. Thus part of a person's income for a period may be invested in new stocks of consumer goods as well as in capital goods which are intended for use in productive activities. By the same token some of the productive activities which occur may occur in the household rather than in the factory, and these should satisfy the same profit conditions as the productive activities that occur in the firms.

The Walrasian approach to temporary equilibrium is entirely appropriate only to steady states where underlying circumstances, technology, tastes, and resources are constant, perhaps with capital stocks and population expanding at uniform rates. Then the comparative statics that can be done is a comparison of different steady states. On the other hand, in the Hicksian model where expectations of price changes are allowed, it is possible to consider the effect on the temporary equilibrium of changes in price expectations which need not duplicate changes in current prices. However, the approach of Walras allows him to ignore the consumer's portfolio problem and treat the consumer as only making a saving decision, since all assets of equal value are treated as indifferent with equal rates of return after allowing for depreciation and insurance costs. When there is uncertainty, the treatment of all assets as indifferent in this fashion is not justified even by the mean variance theory of portfolio selection. The variances and covariances of asset returns must be taken into account. Thus Walras's theory of investment requires that expectations be held with certainty, although he only explicitly assumes certainty within the horizon of a single period, after allowing for fully insurable risks.

There are two features of the Walrasian theory of investment which are quite effective, even by modern standards. One is the analysis of the demand for money. Money is needed during the period to make payments which are planned in advance and the cost of this money service is simply the interest on a loan of that amount for the period. This is very close to the treatment of the demand for money for transactions purposes in modern theory. The demand for money as an asset is merged with the general demand for assets, since any net money balance at the end of the period will be expected to be lent at the current interest rate for the next period, either to others or implicitly to oneself. This represents a cash balance approach to monetary theory where cash balances are only wanted for transactions purposes. It leads to a strict quantity theory of the price of money in terms of other goods in comparisons between steady states.

The second effective feature of Walras's theory of investment is the recognition that the cost of investment goods will depend on the level of investment, since in the general equilibrium high levels of investment will raise the prices of the productive services needed to produce investment goods and thus the prices of the investment goods themselves. In this way the Walrasian theory takes account of the distinction between the marginal efficiency of investment and the marginal efficiency of capital familiar in the Keynesian literature, as well as the modern notion of the cost of adjustment resulting from an increase in the level of investment.

The two main deficiencies of the Walrasian theory of temporary equilibrium are its lack of an analysis of the demand for assets in general in terms of the future consumption streams that the assets are expected

to support and the expected utility they promise to yield, and its lack of an analysis of the demand for particular assets in terms of the distribution of their expected returns.

The neglect of future plans for consumption in determining current demand was addressed by Hicks. He did not suppose that consumers make detailed plans but that they form vague plans and expectations of future prices, which still allow some comparative statics methods to be applied in estimating the effect on current demand of changes in current or future expected prices.

Since firms are recognized explicitly in Hicks's model, they are also represented as making plans for future inputs and outputs in the light of price expectations, which in his case can be identified with the expectations of individuals who become entrepreneurs. The equilibrium of such a model in one period is a set of prices for all the goods and services traded in the market of that period such that the demand for each good or service, including any contract for future delivery that happens to be traded, equals the supply.

Hicks assumes that each consumer and each firm in its planning applies actual or expected interest rates to discount expected future prices to the present so that the problem of maximizing utility for the consumer, or present value for the firm, does not differ, in principle, from the static problem. However, he must assume that agents are risk neutral or in any case that distributions of prices may be replaced by single prices, or certainty equivalents. Thus he is no more able than Walras to analyse how the value of an asset is influenced by the distribution of its returns. But he is able to consider how changes in current prices influence expected future prices, when expected future prices do not necessarily change by the same amounts. This may be the most significant advance made by Hicks beyond Walras, together with the corollary of planning by firms and consumers for a future that involves expected prices changes.

Expectations in temporary equilibrium

A natural way to generalize the Hicksian model and one which has been followed in recent years, for example, by Grandmont, is to impute to each trader an expectation function which gives a probability distribution over future prices, and perhaps over other relevant variables, both market and environmental, as functions of previous values taken by the same variables. Then assuming that each trader has a criterion by which he can choose an optimal trade plan given his expectations, he will determine an excess demand as a function of current prices. Then equilibrium is achieved if there is market clearing at the current prices. Since in the Walrasian or Hicksian model there are two kinds of traders, consumers and entrepreneurs or firms, criteria must be found for each kind of trader.

The criterion for the consumer is rather easily arrived at. It is assumed that each consumer has a von Neumann-Morgenstern utility function, so that any current trade can be evaluated in terms of the expected utility which it makes possible. The utility in turn is derived from the utility of the various possible consumption streams multiplied by their probabilities of occurrence. Of course, these consumption streams and their probabilities logically underlie the expected utilities but they cannot be known to the consumer in detail. The probability distribution on consumption streams is induced by the probability distribution on prices and environmental variables, together with the current trade of the consumer and his plans for future trades, which are in turn contingent on the prices and environmental variables realized in the future. As Hicks points out the consumer may only try to plan levels of spending and certain large expenditures for the future. Particular price expectations will affect these plans and current spending, in total as well as on specific items. What is needed for the theory is to

express consumer's demand finally as a function of current prices so that the condition of market clearing will characterize equilibrium prices. The logic of this analysis is entirely compatible with the methods of Walras, given stationary conditions for tastes, technology, and resources. In simple models it can be spelled out in detail.

On the other hand, there is little agreement on an appropriate criterion for the firm. The difficulty arises that the firm is usually owned by many consumers whose preferences and probability beliefs differ. The consumer does not own capital goods directly but only stock in firms. Moreover, the firms make investment plans and plan their dividend streams in considerable independence of their owners. Walras abstracts from these difficulties in his formal development by two means. First, he treats the consumer as the owner of capital goods which are rented to the entrepreneur. Second, he values the capital goods on the assumption that prices of productive services, interest rates, depreciation rates, and insurance rates will be constant in the future. Given the prices of the productive services arbitrage in the market for capital goods results in a uniform ratio between the net rental of the capital goods, or the prices of their productive services less depreciation and insurance charges, and the prices of the capital goods. In Walras's notation $P_k = p_k / (i + \mu_k + v_k)$ where k indexes capital goods, P_k is the price of the capital good, p_k is the price of its service, i is the interest rate per period, $\mu_k P_k$ is the depreciation change per period, and $v_k P_k$ is the insurance charge per period. In equilibrium the consumer will be indifferent between capital goods in making investments since they all promise equally attractive returns. This also applies in a similar way to investments in circulating capital or in loans.

Hicks adapts the Walrasian viewpoint to a model in which expectations are point valued but not static by imputing to the entrepreneur, who now owns the capital goods, a plan of inputs, including initial stocks, and outputs, including terminal stocks, whose values are discounted back to the present. Then the entrepreneur chooses a plan with the largest discounted value. In this case the firm achieves maximum value in the eyes of its owner. Radner (1972) adapts the Hicksian viewpoint to a model in which point estimates of future prices are not a sufficient basis for decisions. In a temporary equilibrium model his approach imputes to each firm a von Neumann-Morgenstern utility function over alternative dividend streams. This would imply an expected utility for alternative investments in the current period in the same way that the utility of alternative consumption plans implies expected utility for current spending by the consumer.

On the other hand, by use of the stock market it is possible to bring consumers into the decision-making of firms. The firm's criterion is then to choose a plan of production and investment which leads to a maximum value for its shares on the stock market. It can be argued that if the firm chooses a plan which fails to maximize its value in the stock market the stock market will not be in equilibrium, since there is a profitable arbitrage opportunity for someone to buy controlling interest in the firm and revise its planning.

Existence theorems for temporary equilibrium have been proved in many special cases, particularly for trading economies where production does not enter and the number of periods is taken to be finite. Typically the method of proof parallels a method of proof for the model with complete markets, that is, appropriate continuity properties for individual, and thus market, excess demand functions are proved for the goods and services, and the futures contracts, if any, which are traded in the current period. The application of a fixed point theorem completes the proof that a price system exists which results in market clearing, that is, puts each excess demand function equal to zero. However, some special

problems do arise.

Consider a market at the start of period 1 when there are two periods and a second market will be held at the start of period 2. There is uncertainty about the endowment of period 2 and about the spot prices of the second market. All goods are perishable. Suppose there is trading in contracts for current delivery and in forward contracts for delivery in the second period. Let x_1^h, x_2^h be the vectors of goods and services delivered to the *h*th consumer in periods 1 and 2 respectively. Denote by w_1^h and w_2^h the vectors of endowments for the *h*th consumer in periods 1 and 2 respectively. Let $\psi^h(p_1, q_1)$ be the expectation function of the *h*th consumer, that is, the value of ψ^h is a probability distribution of (w_2^h, p_2), where p_1 and p_2 are the vectors of spot prices in periods 1 and 2, while q_1 is the vector of forward prices in period 1 for sure delivery in period 2. There is a finite set of goods and services in each period and a finite number of consumers each of whom holds positive initial stocks in the first period. The possible consumption sets are $X_1^h = R_+^{n_1}$ and $X_2^h = R_+^{n_2}$, the positive orthants of the respective commodity spaces. The following assumptions are made for the consumer.

1. (1) There is a concave and monotone utility function u^h of von Neumann-Morgenstern type, that is, preferences over trades in the first period may be determined by taking the sum of the utilities of the resulting consumption vectors weighted by these probabilities of occurrence.
2. (2) The expectation function $\psi^h(p_1, q_1)$ is continuous in an appropriate sense.
3. (3) For every (p_1, q_1), $\psi^h(p_1, q_1)$ gives probability 1 to the set of (w_2^h, p_2) for which p_2 is positive.
4. (4) The support of ψ^h is independent of (p_1, q_1). The convex hull of the projection of the support of ψ^h on the second period price space has a non-empty interior Π^h.

With these assumptions a necessary and sufficient condition for the existence of competitive equilibrium is that the intersection Π of the Π^h not be empty. In other words there must be an open set of spot prices in the second period which all traders believe to have a positive probability of occurrence. Then, if the forward prices q_1, lie in Π and p_1 and q_1 are positive, excess demand is well defined. Let *D* be the set of (p_1, q_1), satisfying these conditions. As (p_1, q_1) converges to the boundary of *D*, excess demand diverges to ∞. This happens because preferences are monotone and for q_1 outside Π unlimited arbitrage becomes profitable to some trader. These results were reached by Green (1973). It should be noted that point expectations are not consistent with the assumption that Π is not empty, unless all traders expect the same prices next period. However, Π might not be needed to bound short sales if other considerations limit the commitments that will be accepted in view of the likelihood that they can be fulfilled.

Money in temporary equilibrium

There is little difficulty in introducing money into the temporary equilibrium model. It must be

recognized that money serves in at least two capacities, to facilitate exchange, and as an asset with its own prospects for losing or gaining value relative to other goods. In addition it may serve as a numeraire, in terms of which prices are stated. In its capacity as an asset in a market with uncertainty, money may contribute to a diversified portfolio. On the other hand, in its capacity to facilitate exchange money balances will affect the cost of making transactions and thus the stream of consumption which is realizable from given resources. Given his context, where risks are assumed to be insurable, Walras is particularly clear in his treatment of money. If some good other than money serves as numeraire, the price of the service of availability of money is written by Walras as p_m, and the price of money itself as P_m. Then as for any asset the ratio of the net rental to the asset price is equal to the interest rate or $p_m/P_m=i$. Thus if money serves as the numeraire, $P_m=1$ and $p_m=i$. Although his analysis seems somewhat artificial because uninsurable risks are absent, Walras indicates clearly how cash balances may contribute to productive efficiency and to consumer utility.

If attention is concentrated on the asset role of money, so that the transaction role is neglected, it may be shown that the assumption of static expectations may lead to the absence of equilibrium for the current period. Static expectations imply that the relative prices of present and future goods cannot be changed. Therefore, price changes leading to inter-temporal substitution are prevented. Only the wealth effects of price changes have free play since price level decreases raise the value of the money stock and conversely for increases. However, as Grandmont (1983) has demonstrated, these real balance effects may be insufficient to equate supply and demand. For example, if there is excess demand for current goods, this excess demand may not be eliminated by increases in the current price level which are accompanied by equally large increases in the future price level. In a trading economy the effect of the price increases is to reduce the wealth of the traders toward the endowment point ($w(1)$, $w(2)$) in a two period model. Suppose there is only one good, which is perishable, and money is the only store of value. Then if the marginal utility of the current endowment exceeds the marginal utility of the second period endowment for all traders, the price of the good cannot rise high enough to reduce current demand to the current endowment. The same dilemma may arise when the Hicksian elasticity of expectations is equal to one, even though expected prices do not equal current prices.

Grandmont considers a model of this type where trading in futures contracts is excluded so that point expectations do not cause difficulties. It is a trading economy in which consumers receive an endowment of perishable goods in each period of their lives and an initial money stock in the first period. In the current period they maximize a utility function of consumption over the remaining periods of life (assuming the life span to be known) subject to budget constraints of the form $p_t x_t + m_t = p_t w_t + m_{t-1}$, where future prices p_t are equal to functions ψ_t of present prices p_1. He assumes

1. (1) The utility function $u^h(x_1,\ldots, x_{n(h)})$ is continuous, increasing, and strictly quasi-concave for every h.
2. (2) The endowments w_t^h are positive for all h and t, $1 \leq t \leq n(h)$.
3. (3) Total money stock $M = \Sigma_h m_h$ is positive.

He then proves that the temporary monetary equilibrium exists, that is, money prices are well defined, if

every agent's price expectations ψ_t^h are continuous and, for at least one agent, who will be living in the next period and who has a positive money stock, price expectations are bounded away from 0 and ∞. In Grandmont's opinion this result leaves the existence of temporary equilibrium ‘somewhat problematic’. However, it seems quite inappropriate to deal with a money which has no role to play in facilitating transactions. Grandmont and Younes (1972) have studied general equilibrium in a model similar to the model just described except that lifetimes are taken to be infinite and utility functions are separable by time period, that is $u^h(x_1, \ldots) = \Sigma_{t=1}^{\infty} \delta^t u^h(t)$ for $0< \delta <1$. Also money is now assigned a role in transactions, that is, only part of the proceeds of sales in the current period can be used to finance purchases in this period. Thus in each period there is both a budgetary constraint, as before and, in addition, a liquidity constraint, which may be written in simplest form as

$p_t(x_t - w_t)^+ \leq m_t + k p_t(x_t - w_t)^-$, where for any vector *z* we write $z_i^+ = \max(z_i, 0)$ and $z_i^- = \max(-z_i, 0)$, and $0<k<1$. Thus the fraction k of receipts from sales can be used to buy goods in the current period. This fraction could be allowed to vary by consumer and by good. The constraint on purchases is entirely in the spirit of Walras. It is an explicit modelling of a need for liquidity that he left implicit in his account.

In order to prove that a monetary equilibrium exists an assumption to bound expected prices is made which is very similar to the previous assumption for this purpose, and also very similar to the assumption made by Green to obtain existence of temporary equilibrium in a non-monetary economy with futures trading. The assumption is that the set of expected prices, over a finite planning horizon, that result from all possible choices of current prices, which are assumed positive, lie in a compact subset of the set of positive future prices. Then if all consumers have continuous expectations which satisfy this assumption, and the assumptions of the previous model are also met, there will exist a temporary equilibrium in this case also. Indeed, the case $k=1$, where the liquidity motive is lacking, can be allowed.

In the second model where money has a transactions role expectations are described as depending on past prices as well as current prices, which leads inelastic price expectations to be more plausible. It also gives plausibility to correct foresight in states of stationary equilibrium over sequences of periods. Grandmont and Younes (1973) prove that the stationary equilibria of the model are not Pareto optimal. However, they can be made Pareto optimal by use of a lump sum tax to reduce the quantity of money by a factor equal to the discount factor for utility. It is then proved that a continuum of such equilibria exists to sustain any Pareto optimal allocation, since the price level falls by the same factor, and it is not worthwhile to reduce a money stock, even if it is in excess of transaction requirements. Moreover, if the tax rate is set slightly too high, the consumer will always wish to increase his real balances and no stationary equilibrium will exist. Grandmont and Younes are not able to prove that an exact stationary equilibrium exists for a fixed money stock, although a near equilibrium exists if the discount factor is near 1.

In addition to proofs of existence and non-optimality for monetary equilibria, Grandmont and Younes show that the quantity theory holds between stationary equilibria, that is, if p and m_h, $h = 1, \ldots, m$, provide a stationary equilibrium, then λp and λm_h also provide one. This is the conclusion of Walras as well. On the other hand, the stationary equilibria of a monetary economy will differ from the stationary equilibria of a barter economy unless $\delta=1$. This is apparent from the fact that the barter

economy's equilibria are Pareto optimal and the monetary economy's equilibria are not, unless $\delta = 1$. Thus the simple ‘classical dichotomy’ does not hold.

Equilibrium over time

In addition to temporary equilibrium Hicks considered the possibility of equilibrium over time, in the sense that the expectations held by traders in one market about prices on future markets are realized when those future markets are held. However, when there is uncertainty it is not clear what is meant by the realization of expectations. If expectations take the form of a non-atomic probability measure over future prices, any vector of prices within the support of the measure is as likely as any other, that is, it has zero probability. Nor does the Hicksian trick of replacing the probability distribution by a representative price, depending on the trader, avoid the difficulty, since the representative price is not typically a statistic of the price distribution, such as the mean or the mode. Thus even if all traders held the same expectations in the sense of a probability distribution for prices, they would not have the same representative prices except by the chance that their circumstances and their risk preferences also coincide.

A way to resolve this dilemma was provided by Radner (1972). His solution is a type of perfect foresight. All traders hold the same point expectations for prices with certainty, contingent on the event in which the market is held. Only a finite number of dates are allowed and only a finite number of events may occur in each. From the viewpoint of a given market the relevant elementary events are the possible sequences of states of nature that may occur up to the horizon. For any such sequence the traders expect correctly a corresponding sequence of prices. This does not lead to a grand initial market in which all future exchanges are arranged because the set of forward commitments which are actually available in the market is a small subset of all those associated with future events. For example, it may be that most commodities are traded for sure delivery and only one commodity (money or the numeraire) is traded on a contingent basis (insurance). It should be noted that this construction does not depend on any agreement between traders on the probabilities of the alternative events. Thus the expectation functions which were introduced in the discussion of temporary equilibrium would not be likely to be the same for different traders.

In this setting the trader plans a sequence of consumptions contingent on the events in which they occur and also a sequence of trades on the markets which are open. Spot markets are open for all commodities at all dates but only a small subset of the possible markets in forward contracts may be open at any particular date. In any case since the number of dates and states of nature and thus of elementary events is finite, only finitely many prices will arise.

Let X_h be the consumption set of the hth consumer. Let M be the set of elementary date-events pairs. A consumption-trade plan for the hth consumer is a pair (x^h, z^h) where x_m^h is the consumption planned for $m \in M$ and z_m^h is the trade planned for $m \in M$. Let $\Gamma_h(p)$ be the set of feasible plans for h, given prices p. In particular, (x^h, z^h) in $\Gamma_h(\mathrm{p})$ implies that consumption x_m^h plus net deliveries z_m^h due at m are not greater than resource endowments w_m^h for each m and the budget constraint $p_m z_m^h$ holds at each $m \in M$.

Let $\gamma^h(p)$ be the set of plans in $\Gamma_h(p)$ which are optimal for h. An equilibrium of plans and price

expectations (including current prices which are known) is given by plans (x^h, z^h) and expected prices p such that (x^h, z^h) is in $\gamma^h(p)$ for each h, that is, the plans are preferred at the expected prices, and the sum $\Sigma_h \Sigma_m^h$ of commitments at each m is non-negative, and the value of commitments $p_m \Sigma_h z_m^h = 0$ at each m, that is, Walras's Law holds. In such a purely trading economy for perishable goods with a finite set of dates and events and under assumptions of the usual kind on preferences, and positive endowments which lie in the interior of consumption sets, Radner proves that an equilibrium exists.

It is not difficult to bring production into this setting if firms are introduced with fixed production plans and with shares which are traded on a stock exchange. The ownership of a share of a firm can be equated to the ownership of a share of its output, including the end of the period capital stock. The output of a firm at any date would depend on the event, and the function relating this output to the events would be known by traders, just as future prices of goods are known, contingent on events. Now, in addition to goods prices, share prices are foreseen in each event at each date with certainty. As before the number of dates and events is finite.

A feature of this model not present in the trading model is that consumers do not own the resources of the firms as individual goods but as proportions of the batch of goods that firms hold. The consumer can buy and sell goods forward by means of long and short positions in the stock market but the trade he arranges by these means for one event at the next date determines his trade for all other events at that date. Thus spot markets still may offer useful alternatives, quite aside from the practical difficulties of physically dissolving the firm. Of course, given the presence of spot markets, dissolution of the firms is not needed if the value of the firm equals or exceeds the value of its resources.

If one tries to go further to specify how the production and trading plans are arrived at, a major problem arises of setting the objectives of the firm. Hicks solves this problem by assuming that the production plan chosen would have the maximum discounted value among those available. This value could be calculated since expectations were single-valued and interest rate, actual or expected, could be used in arriving at present values. Moreover, firms were treated like single proprietorships. In the modern literature firms have sometimes been assigned utility functions defined on the streams of profits. Another suggestion is to suppose that the firm adopts the plan that maximizes the value of its shares on the stock market. This would seem to be the approach most in accord with other parts of general equilibrium theory. However, it encounters the difficulty that the judgement of the management and the judgement of the market on the probability of different events may not coincide. If this difference of judgement exists, the market solution would be for the firm to be purchased through a takeover by those who value its potential most highly and the management displaced. Markets which work in this way would correspond quite well to the original Walrasian model.

Various results on the existence of a general equilibrium have been reached with special models of production by firms. One theorem of Radner extends the existence of an equilibrium of plans and price expectations to this context. His assumptions are:

1. (1) Consumers satisfy the usual conditions on convexity, non-satiation, and positive endowments.
2. (2) Consumers own the shares of firms and each consumer owns shares in every firm.
3. (3) Producers have closed, convex production sets with free disposal. The total production set satisfies the condition that the negative of a producible vector of commodities is not producible.
4. (4) Each firm has a continuous, strictly concave utility function on profit streams.

With these assumptions he does not achieve a full existence theorem because the model is not well adapted to handle the entry and exit of firms. What may happen is that some firms show an excess supply of shares in some events and dates. Then since the firms are treated like partnerships with unlimited liability, negative share prices might be justified at this point. In any case the questions of entry and exit of firms is one that the Arrow–Debreu model also fails to deal with. The theorem proved by Radner only finds a ‘pseudo-equilibrium’ where the value of total excess supply (of shares) is minimized.

In the foregoing discussion it has been assumed that only a subset, possibly small, of the potential Arrow–Debreu markets is open. It is possible to justify the selection of markets which are open by postulating costs for carrying out transactions. If the markets which are open are given, the previous equilibria may be supported by assigning infinite transactions costs to the lost markets and zero costs to the open ones. Otherwise the open markets will be endogenous to the general equilibrium. In the analysis of markets with transaction activities which consume resources the same convexity or linearity assumptions have been used as for the production technology. Then it is not difficult to prove existence of equilibrium under assumptions of the usual sort.

Rational expectations

It has been implicitly assumed in the preceding discussion of temporary equilibrium that the traders have the same information available. If this is not the case the complication arises that the equilibrium price may convey information. For example, in the market for umbrellas if some traders have the benefit of weather forecasts and some do not, a high price based on the demand of informed traders will signal to uninformed traders that rain is expected. Then all traders are informed and an equilibrium price must be consistent with fully informed demand.

A difficulty arises if it happens that the utilities of consumers depend on events in contrary ways, that is, uninformed consumers use umbrellas to ward off sun and informed traders to ward off rain. Then price will be higher if rainy weather is expected by informed traders but if uninformed traders perceive this and become informed, the high price may not appear and a fully informed market may not show a price difference depending on the weather forecast. But then no information is transmitted so the weather forecast cannot be read out of equilibrium prices. The conclusion is that no equilibrium is possible. However, the result requires an exact balance in the effects of rain and sun on the two sets of traders, so it is unlikely to hold. More robust examples of nonexistence were given by Green (1977) and Kreps (1977). The idea of the discontinuity was first proposed by Radner (1967).

A rational expectations equilibrium is said to exist if there is a function ϕ mapping states of the world into equilibrium prices which is invertible, that is, ϕ^{-1} exists, mapping prices, from a normalized set, into states of the world. It is clear that such a function will exist if the equilibrium price which appears when all traders are fully informed is uniquely determined by the elementary event, and the relation is one to one. It is also clear, given a finite set of elementary events, that the correspondence of prices to elementary events will be one to one in all but exceptional cases. Then the equilibrium is said to be revealing. But the price function of a revealing full information equilibrium is a price function that provides a rational expectations equilibrium. This observation is due to Grossman (1981).

The situation is more complicated when the possibility is recognized that spending resources will allow

more information to be gathered. The information that is disseminated free of charge by prices will discourage the use of resources to gather information and thus prevent the attainment of a Pareto optimum. In welfare terms a suboptimal amount of resources will be devoted to information activities.

An infinite horizon

In the Arrow–Debreu model of general equilibrium there are a finite number of periods, a finite number of locations, a finite number of events, and a finite number of commodity types, so the number of distinct goods when all these grounds for distinguishing goods have been recognized is still finite. The principal objection to the restriction to a finite number of goods is that it requires a finite horizon and there is no natural way to choose the final period. Moreover, since there will be terminal stocks in the final period there is no natural way to value them without contemplating future periods in which they will be used. The finiteness of the number of locations and commodity types is achieved by making a discrete approximation to a continuum, and perhaps the finiteness of the number of states of nature can also be viewed in this light. But in the case of time, a discrete approximation by periods still leaves a denumerable infinity of dates.

There are two principal models in which an infinite number of goods appear. In one model there is a finite number of infinitely lived consumers. Such a consumer may be considered to represent a series of descendants stretching into the indefinite future, so that consumers alive in the present period have an interest in the goods of all periods. The other model has an infinite number of consumers, but only a finite number of them are alive in any period. This model is called the overlapping generations model. It was first proposed and explicitly analysed by Samuelson (1958).

A model of general competitive equilibrium with a finite number of consumers and an infinite number of commodities was first presented in rigorous form by Peleg and Yaari (1970). They assumed the number of commodities to be denumerable. This is a basic case since a noncompact but separable commodity space can be approximated arbitrarily closely with a denumerable set of commodities in the same sense that a compact commodity space can be approximated by a finite set of commodities. This assumes that a sensible neighbourhood system can be defined in the commodity space, as Debreu does for the dimensions of location and time with places and periods.

Peleg and Yaari present a trading model without production. The commodity space s is the space of all real sequences. In order to discuss continuity the space must be given a topology, in this case, the product topology. Thus a sequence of points converges if it converges in every coordinate, that is, $x^s \to x,\ s = 1, 2, \ldots$, if $x^s(i) \to x(i)$ for $i = 0, 1, \ldots$. The space is presented as a sequence of real numbers but by grouping terms it may equally well represent a sequence of vectors, for example, commodity bundles occurring in successive time periods. The hth trader has an initial stock w_h, where $w_h \in s$ and a preference relation $\succsim_h$, which is reflexive, transitive, and complete on s_+, the set of non-negative sequences. Strict preferences $\succ_h$ is defined by $x \succsim_h y$ if $x \succsim_h y$ and not $y \succsim_h x$.

Peleg and Yaari prove an existence theorem for this economy on the following assumptions.

1. (1) Desirability. If $x \geq y$, then $x \succsim_h y$.
2. (2) Strong convexity. If $x \neq y$ and $x \succsim_h y$, then $\alpha x + (1-\alpha) y \succ_h y$ for $0 < \alpha < 1$.
3. (3) Continuity. The two sets $\{y \mid y \succsim_h x\}$ and $\{y \mid x \succsim_h y\}$ are closed.

4. (4) Positivity of total supply. Let $w = \Sigma_{h=1}^{m} w_h$. Then $w > 0$.

A price system is a real sequence $\pi > 0$ which satisfies $\Sigma_{i=0}^{\infty} \pi(i) w_h(i) < \infty$, that is, the value of the initial bundles is finite. This implies that $\pi(i)w_h(i)$ converges to 0 as $i \to \infty$. A competitive equilibrium is given by $(x_1,\ldots,\ldots x_m;\ldots \pi)$ such that π is a price system, $\Sigma_{i=0}^{\infty} \pi(i) x_h(i) \leq \Sigma_{i=0}^{\infty} \pi(i) w_h(i)$, for each h, and $\Sigma_{i=0}^{\infty} \pi(i) x(i) \leq \Sigma_{i=0}^{\infty} \pi(i) w_h(i)$ implies $x_h \succsim_h x$. Peleg and Yaari prove that a competitive equilibrium exists.

It is clear from their discussions, and it has become even clearer in subsequent work, that the use of a topology such that, in the context of an infinite horizon interpretation of the model, impatience is implied by continuity of preferences is the crucial assumption for a proof of existence. That the product topology implies impatience may be seen in the following way. If $x \succsim_h y$ then by continuity there is a neighbourhood U of x such that $z \in U$ implies $z \succ_h y$. However, a neighbourhood U is defined by $|z(i) - x(i)| \langle\in\rangle 0$ for a finite number of coordinates where the remaining coordinates are free. Thus given $y \succsim_h x$, there must exist $N > 0$ such that $z(i) = x(i)$ for $i \leq N$ and $z(i) = 0$ for $i > N$, and $z \succ_h y$. These conditions are met if the preference order is representable by a separable utility function which is the sum of periodwise utilities discounted back to the present at a constant rate per period, and these utilities are continuous and uniformly bounded. Such a utility function is a common way of expressing impatience.

A model of general competitive equilibrium which allows for production where there is an infinite number of commodities was first presented in a rigorous form by Bewley (1972). A preference relation $\succsim_h$ is assumed for each consumer as in Peleg and Yaari. We will describe Bewleys model for the case of a sequence of periods with an infinite horizon where N_1 is the finite set of commodities available in the tth period. Then the set of all commodities is

$$M = \bigcup_{t=1}^{\infty} N_t.$$

It is assumed that $M = M_c \cup M_p$ where M_c and M_p are disjoint and M_c contains the consumption goods. Bewley confines attention to the commodity space l_∞ of bounded sequences of real numbers. Let $K_c=\{x \in l_\infty | x(i)=0$ for not $i \in M_c$, $x(i) \geq 0$ for $i \in M_c\}$, and similarly for K_p. Let $\check{K}_c$ have the same definition as K_c except that $x(i) > \varepsilon > 0$ for all $i \in M_c$ for some given ε. Bewley's existence theorem holds for a weaker notion of continuity than that of componentwise convergence, but we will stay with the definition used by Peleg and Yaari for the sake of simplicity. Then the assumptions on the consumer sector are

1. (1) The consumption sets $X_h = K_c - w_h$ where w_h is the endowment of the hth consumer.

2. (2) The sets $\{y \mid y \succsim_h x\}$ and $\{y \mid x \succ_h y\}$ are closed. Also $\{y \mid y \succsim_h x\}$ is convex.
3. (3) M_c is not empty and for each h, if $x \in X_h$ and $y \in \check{K}_c$, then $x + y \succsim_h x$.

   The production sector is defined by means of production sets Y_t which convert inputs belonging to N_{t-1} into outputs belong to N_t. The $Y = \Sigma_{t=1}^{\infty} Y_t$. The assumptions on the production sector are:
4. (4) Y is a convex closed cone with vertex at 0.
5. (5) If $w \in l_\infty$, then $Y + w \cap l_\infty$ is bounded.
6. (6) If $y \in Y$, then $y^n \in Y$ where $y_t^n = y_t$ for $t = 0, \ldots, n$, and $y_t^n = 0$ for $t>n$.
7. (7) $-K_p \subset Y$

   Assumption (4) means that each Y_t is a linear activities model as Walras assumed. Assumption (5) excludes unbounded production from given inputs. Assumption (6) allows production to end at any time with free disposal of the final outputs. Assumption (7) allows free disposal of all goods other than consumption goods.

   In addition there is one assumption which relates the consumption sector and the production sector.
8. (8) For each consumer h, there exists $\bar{x}_h \in X_h$ and $\bar{y}_h \in Y$ such that $\bar{y}_h(i) - \bar{x}_h(i) > \in > 0$ for all i and some $\in>0$.

Assumption (8) protects consumer income in the sense that the consumer is not reduced to the subsistence level in equilibrium. That is to say, there are cheaper consumption bundles within this consumption set at equilibrium prices. An equilibrium is an allocation $(x_1\ldots,\ldots,\ldots x_m,y)$ and a price sequence $\pi = (\pi(0), \pi(1), \ldots)$ where $\pi(i)$ is non-negative for all i but different from zero for some i, which satisfy the conditions:

1. (I) $y \in Y$ and $\pi y = 0$, $\pi z \leq 0$ for all $z \in Y$. The profit condition.
2. (II) $x_h \in X_h$ and $\pi x_h = 0$, all h, and $z \succ_h x_h$ implies that $\pi z > \pi x_h$. The demand condition.
3. (III) $\Sigma_{h=1}^{m} x_h = y$. The balance condition.

On the basis of the assumptions Bewley is able to prove that an equilibrium exists where the price system $\pi \in l_1$, that is, $\Sigma_{i=0}^{\infty} \pi(i)<\infty$. This represents a generalization of the classical existence theorem in the form given by McKenzie to the case of denumerably many commodities, retaining the assumption of a finite number of consumers. The argument is stated in terms of an infinite horizon and a finite number of goods in each period, but the original theorem is more general and applies to the case of uncertainty with an infinite number of events as well as to models with a continuum of commodities. The continuum of commodities may arise from a variation in the physical properties of the goods and services.

Overlapping generations

In the overlapping generations model of general equilibrium the number of consumers as well as the number of commodities is infinite. However, at any given time the number of both is finite. While the model with a finite number of infinitely lived consumers treats the consumers who are living as if their lives were extended into the indefinite future by the lives of their descendants, in the classical overlapping generations model bequests are neglected and each generation is assumed to be interested only in its own consumption.

The first rigorous analyses of an overlapping generations model in a general equilibrium setting were done by Balasko, Cass and Shell (1980) and by Wilson (1981). They treat an exchange model in which all goods perish in each period, and each consumer receives an endowment in each period. They assume that each consumer lives for two periods. However, this assumption is not essential. What is essential is that lifetimes are finite in length and some of the people alive at any date have lifetimes which overlap the lifetimes of some people who are born later than they.

The formal model makes these assumptions.

1. (1) In each period t (t=1, 2,…) there is an arbitrary, finite number of perishable commodities $n^t \geq 1$.
2. (2) Each consumer h=1, 2, … lives for two periods. At the start of period t an arbitrary but finite number of consumers is born with indices $h \in G^t$.
3. (3) Consumption sets $X_h = R_+^{n(0)}$ for $h \in G^0$ the consumers alive when the economy begins and $X_h = R_+^{n(t)} \times R_+^{n(t+1)}$ for $h \in G^t$, $t \geq 1$. Write $x_h = x_h'$ for $h \in G^0$ and $x_h = (x_h(t), x_h(t+1))$ for $h \in G^t$.
4. (4) Each consumer has a utility function, $u_h(x(1))$ for $h \in G^0$ and $u_h(x(t), x(t+1))$ for $h \in G^t$. Utility functions u_h are continuous, quasi-concave, and without local maxima.
5. (5) Each consumer receives an endowment, $w_h = w_h(1)$ for $h \in G^0$ and $w_h = (w_h(t), w_h(t+1))$, for $h \in G^t$. For each h, $w_h \geq 0$ and $w_h \neq 0$.
6. (6) The economy is inter-temporally irreducible. Let $I(t) = \{h | h \in G^s \text{ for } 0 \leq s \leq t\}$. Then there exists a sequence $t_\mu \to \infty$ with the following property. Given any allocation $x=(x_1, x_2,\ldots)$ and $I_1(t_\mu)$ and $I_2(t_\mu) \neq \varphi$, with $I_1(t_\mu) \cap I_2(t_\mu) = \varphi$, and $I_1(t_\mu) \cup I_2(t_\mu) = I(t_\mu)$, there exist $y_h \geq 0$ for $h \in I_1(t_\mu)$ and $x_h' \geq 0$ for $h \in I_2(t_\mu)$, such that $\sum_{h \in I_1(t_\mu)} y_i(t) = 0$ when $\sum_{h \in I_1(t_\mu)} w_{hi}(t) = 0$, for $1 \leq i \leq n^t$, $1 \leq t \leq t_\mu + 1$, and

$$\sum_{h \in I_2(t_\mu)} x_h' \leq \sum_{h \in I_1(t_\mu)} (w_h + y_h) + \sum_{h \in I_2(t_\mu)} w_h .$$

Moreover, $u_h(x'_h) \geq u_h(x_h)$ for all $h \in I_2(t_\mu)$ with the strict inequality for some h.
Assumption (6) is the irreducibility assumption of McKenzie adapted to economies made up of the consumers born by the period t_μ. It says that it is always possible to increase the welfare of the second subgroup if the scale of the endowment of the first subgroup is increased.
Let $p=(p(1), p(2),\bullet\ldots)$ where $p(t) \in R^{n(t)}$. Then the pair (x, p) is a competitive equilibrium if

1. (I) For all h, $u_h(x_h)$ is maximal over all z_h such that
   $p(t)z_h(t) + p(t+1)z_h(t+1) \leq p(t)w_h(t) + p(t+1)w_h(t+1)$ if $h \in G^t$, $t \geq 1$, and
   $p(1)z_h(1) \leq p(1)w_h(1)$ if $h \in G^0$, where $z_h \geq 0$.
2. (II) $\Sigma_h x_{hi}(t) \leq \Sigma_h w_{hi}(t)$ with equality if $p_i(t) > 0$, where the summation is over
   $h \in G^{t-1} \cup G^t$, $1 \leq i \leq M(t)$ and $t \geq 1$.

Condition (I) is the usual demand condition and condition (II) is the balance condition. Balasko, Cass and Shell (1980) prove that the six assumptions listed imply the existence of a competitive equilibrium. They show that the artificial assumptions on birthdates and lifetimes are irrelevant by a redefinition of the period. They also conjecture that the introduction of production and consumption sets of the usual classical type, which are closed, convex, and bounded below, would cause no major difficulties.
Wilson (1981) treats an economy which may contain both finite lived and infinite lived consumers and which may be specialized to either. He also allows intransitive preferences. He uses a somewhat simpler version of irreducibility and proves existence in an exchange economy where the number of goods in each period is finite in two circumstances (1) when the consumers are all finite lived and (2) when a finite subset of infinite lived consumers own a positive fraction of the endowment in all but a finite number of periods. If preferences are transitive and strictly convex, the competitive equilibrium is also Pareto optimal. Thus Wilson's results contain the theorems on existence of Bewley and Balasko, Shell, and Cass as special cases while also providing conditions in the model sufficient for Pareto optimality.
A striking difference between the competitive equilibria of economies where the number of consumers is finite, and the competitive equilibria of economies with overlapping generations and an infinite horizon, where the number of consumers is infinite, is that with perfect foresight the former equilibria are also Pareto optima while the latter need not be. This is the major point emphasized by Samuelson in his initial paper. The most general theorem proving that competitive equilibria are Pareto optima even when the number of commodities is infinite provided that the number of consumers is finite is due to Debreu (1954). Under some additional smoothness conditions on utility and boundedness conditions on prices and allocations Balasko and Shell (1980) prove that the allocation x of a competitive equilibrium is Pareto optimal if and only if $\Sigma_t(1 / \|p_t\|) = \infty$. This is a condition which had already been shown to characterize efficiency in neoclassical production economies by Cass. It is clear that $\lim \inf(\|p_{t+1}\| / \|p_t\|) = r \leq 1$ implies that the condition for Pareto optimality is satisfied since the sums dominate $\Sigma_t(1 / r^t)$ which diverges. Intuitively, for a stationary economy if the interest rates are asymptotically non-negative, the competitive equilibria will be Pareto optimal, or if the economy is growing, if the interest rates exceed the growth rate, Pareto optimality follows.

Limitations of the analysis

As mentioned in the beginning the claim of the theories described as general equilibrium theories to be ‘general’ is qualified by the set of conditions considered to be constant. Walras as well as most subsequent theorists classified the constant factors as tastes, technology, and resources, including population. However, all three of these categories have been treated by some economists as responding, in ways amenable to analysis, to market variables. These studies have usually been confined to a few variables and have usually been partial equilibrium in character, although the classical school of economists included population as a major variable in models of economic development. Their models are comprehensive but lack the market equilibrium analysis of the general equilibrium theories, whose inspiration appears to have been found in the marginal utility theory of consumer demand. Similarly, tastes have sometimes been modelled to depend on past consumption or advertising, and technology has been modelled to depend on research and development spending and on the rewards to innovation. Also natural resources, in terms of resources known to exist, are often treated as responding to prices.

From this perspective general equilibrium theory is a partial theory of economic affairs with a special set of *ceteris paribus* assumptions. The variables which are left free are chosen because they lend themselves to a particularly elegant theory in terms of consumer demand under budget constraints and producer supplies with profit conditions where these constraints and conditions are established by prices equating demand and supply. This was the vision of Walras, perhaps guided by the theory of static equilibrium of mechanical forces which he found in Poinsot.

Another direction of abstraction in general equilibrium theory in its classic expressions has been to ignore the effects of processes which do not pass through the market. In particular each consuming unit is described as interested only in its own consumption in the theory of Pareto optimality and as uninfluenced in its choices by the choices made by other households. Similarly, the production possibilities of one firm or process are treated as independent of the productive activities of other firms. Some attempts have been made to incorporate these effects in the general equilibrium models but not with complete success. In particular there is not a good theory of existence when consumer possibility sets or production sets are affected by levels of consumption and production.

The convexity assumptions which have appeared in general equilibrium models from the time of Walras are often not good approximations of reality though they are depended on for many of theorems of the subject, such as the theorems on existence and Pareto optimality. However, there is a theory of approximate equilibria and of limiting results as the size of the market increases relative to the participants which does something to bridge the gap between theory and fact.

Finally, the assumption that the market participants take prices as independent of their actions fails to describe many markets, and describes very few exactly. Nonetheless, this assumption may be useful for a theory that embraces all markets, whose special features cannot be described in detail. It may, that is, give a good approximation to the working of the economy as a whole. Also it is useful for its implications for optimality, a point which was perceived, albeit through a glass darkly, by Walras. The proper notion was later found by Pareto.

Just as the model does not accommodate monopoly easily, government does not fit in well. A chief difficulty arises from its compulsory features which allow it to extract resources by force rather than by voluntary agreement. Government is not easily described either as a producer selling services, or as a

voluntary organization performing acts of collective consumption, though in ways it resembles both. Voluntary societies also do not fit perfectly in the scheme of producers and households though the disparity is less, since they must meet their expenses from contributions by the membership who will not contribute unless the services of the society to them are worth the dues they pay.

Properties of general equilibrium

Walras set the major objectives of general equilibrium theory as they have remained ever since. First, it was necessary to prove in any model of general equilibrium that the equilibrium exists. Then its optimality properties should be demonstrated. Next it should be shown how the equilibrium would be attained, that is, the stability of the equilibrium and its uniqueness should be studied. Finally, it should be shown how the equilibrium will change when conditions of demand, technology, or resources are varied, the subject now called comparative statics. He contributed to all these lines of research.

Walras's arguments for existence are not conclusive but he did contribute a basic principle, that the model should be neither underdetermined nor overdetermined. That is, the number of independent equations to be satisfied and the number of variables to be determined should be equal. Some critics saw right away that this equality did not ensure a meaningful solution to the equation system, for example, that the solution to such an equation system is not guaranteed to be real. The question was not taken up seriously until the 1930s and the first rigorous treatment was given by Wald (1935, 1936b). Then in the 1950s more complete solutions on neo-classical assumptions were found by Arrow and Debreu (1954), McKenzie (1954) and Nikaido (1956).

In the discussion of models of general equilibrium that have been given above, the first requirement has been a set of assumptions from which existence could be inferred. This approach to the subject was begun in the papers of Wald and von Neumann, presented to the colloquium of Karl Menger (mathematician and son of Carl Menger, the neoclassical economist) in Vienna in the 1930s.

The optimality that Walras claimed for competitive equilibrium, under conditions of certainty, except for insurable risk, did not seem to go beyond individual maximization of utility in face of an equilibrium price system. However, Pareto gave a genuinely social definition that the allocation of goods and services in a competitive equilibrium is such that no reallocation is possible with some consumer better off unless some consumer is made worse off. In fact, Walras seemed to be groping for the same definition and his arguments may be slightly extended to establish Pareto's proposition.

As noticed in the earlier discussion of markets with certainty, Pareto optimality is implied by maximization of preference under budget constraints and von Neumann's law, or maximization of profit given the technology. The former implies that an allocation which improves one consumer's position and harms none must, given local non-satiation for all consumers, be more valuable at equilibrium prices while the latter implies that no more valuable allocation is achievable. This argument depends on the finiteness of the value of the goods in the economy. Otherwise the impossibility of a more valuable allocation is not meaningful. Thus when the horizon is infinite and the discount factor is too large, for example, equal to 1 if the economy is stationary, or in general greater than or equal to the reciprocal of the growth rate, Pareto optimality may fail in competitive equilibrium, as Samuelson showed. Also there is no reason to expect Pareto optimality, in an exact sense, when some markets are missing, a very likely eventuality when there is uncertainty and goods must be traded on every possible contingency to provide

complete markets.

A second theorem on Pareto optimality asserts that any Pareto optimum can be realized as a competitive equilibrium. This theorem requires assumptions which are similar to those leading to existence, in particular, assumptions providing local non-satiation for some consumers and convexity of the preferred sets and the feasible set. Moreover, when the number of goods is infinite as in the case of an infinite horizon an additional condition is needed to give the existence of the prices. This condition may be that the sum of consumers' preferred sets has an interior or that the production set has an interior. In the case of the product topology and free disposal by consumers the preferred sets will have interiors if the periodwise utility functions are continuous and bounded (see Debreu, 1954). Finally it was shown by Arrow (1953) that in order for the Pareto optimal allocation to maximize preference over the budget set rather than only to minimize the cost of achieving a given preference level, it is useful to assume that x_i, the consumption set of the ith consumer, contains a point which is cheaper than the allocation he receives, for $i=1,\ldots,m$.

The stability theory for general equilibrium has been largely devoted to the stability of the Walrasian tâtonnement, or process of groping for equilibrium prices through a process of price revision according to excess demand. That is, prices rise or fall depending on whether excess demand is positive or negative. In the tâtonnement there is no trading until equilibrium prices have been reached. The most convincing theorems concern local stability and the dominant assumption leading to local stability is that the market excess demand function satisfies the weak axiom of revealed preference between the equilibrium price and any other price in a sufficiently small neighbourhood of the equilibrium price. That is, if $\bar{p}$ is an equilibrium price and e is the excess demand function, $p \cdot e(\bar{p}) - p \cdot e(p) \leq 0$ implies $\bar{p} \cdot e(\bar{p}) - \bar{p} \cdot e(p) < 0$. Since $\bar{p}$ is an equilibrium price, $e(\bar{p}) = 0$, and $p \cdot e(p) = 0$ by Walras's Law. Therefore, the condition holds and we may conclude that $\bar{p} \cdot e(p) > 0$. The weak axiom for the market may be expected to hold if the net income effect of price changes is small.

Consider the price revision process given by $dp_i/dt = \dot{p}_i = e_i(p),\ i = 1, \ldots, n-1$, where the nth good is numeraire so $\dot{p}_n \equiv 0$. Then consider the function $|p(t) - \bar{p}|^2$, the square of the distance from the equilibrium price vector to the price vector at time t. We derive

$$d/dt(|p(t) - \bar{p}|)^2 = 2\sum_1^n (p_i - \bar{p}_i)\dot{p}_i = 2\sum_1^n (p_i - \bar{p}_i)e_i < 0,$$

using the weak axiom of revealed preference and Walras's Law. Thus the distance of $p(t)$ from $\bar{p}$ constantly falls, or $p(t) \to \bar{p}$ as $t \to \infty$. Since locally the rate of price change can be equated to excess demand for any continuous tâtonnement by choice of units, this is a general argument. Since the assumption of gross substitutes ($e_{ij}<0$ for $i \neq j$ and $e_{ij}=\partial e_i(p)/\partial p_j$) implies the weak axiom, and the assumption of a negative definite Jacobian $[e_{ij}]$, $i, j=1,\ldots, n-1$, at equilibrium is equivalent to the weak axiom locally, the weak axiom is a dominant condition for local stability. All global stability results are very special and relatively unconvincing.

A rigorous treatment of the stability problem for the tâtonnement was given by Arrow and Hurwicz (1958) and Arrow, Block and Hurwicz (1959). A stability theory which allows for trading was given by Hahn and Negishi (1962). These theories do not allow for speculative trading although profitable arbitrage opportunities would be likely to exist for any speculator who correctly inferred what the price revision process was. The stability of the tâtonnement was conjectured by Walras, to be the normal case for economies with many goods and essentially correct arguments were given by Walras for the case of exchange economies with two goods. He recognized and illustrated the case of locally unstable equilibria in the two goods case.

Finally, as Walras saw, it may be possible through a general equilibrium analysis to determine the effect of changes in the exogenous factors, resources, technology, or tastes, on the economic variables in equilibrium. This is analogous to the effect of a change in the constraints on the equilibrium of mechanical forces, an analogy with which Walras would have been familiar from the book of Poinsot. In the case of the exchange of two commodities Walras derives some simple and correct results for comparative statics just as he does for stability. He observes that an increase in the marginal utility of a good or a reduction in its supply will raise its price. In drawing this inference from his demand and offer curves he confines himself to stable equilibria as the only equilibria of interest.

Hicks used the comparative static result of Walras in a market with many goods to define stability of equilibrium. Samuelson (1947) pointed out that stability of equilibrium, where stability is given a dynamic interpretation as in a continuous tâtonnement, may imply comparative static results as a general principle. However, the straightforward generalization of Walras is the use of conditions which are sufficient to imply stability as a basis for deriving theorems on comparative statics. The most interesting theorem may be that derived from the revealed preference assumption at equilibrium.

Suppose that $e(\bar{p}) = 0$ but excess demand changes so that the new excess demand function $e_i'(\bar{p}) = e_i(\bar{p})$ for $i \neq 1$ or *n* and $e_1'(\bar{p}) = \delta_1 < 0$ while $e_n'(\bar{p}) = \delta_n > 0$. Let *n* be numeraire. This change can be arranged by taking δ_n of the *n*th good from some holder and compensating him with $\delta_1 = \delta_n / \bar{p}_1$ of the first good. Suppose that the new equilibrium price is *p*, or $e'(p) = 0$. By Walras's Law $\bar{p} \cdot e'(\bar{p}) = 0$ and by the assumption of revealed preference $p \cdot e'(\bar{p}) > 0$. Thus $(\bar{p} - p) \cdot e'(\bar{p}) < 0$, or $(\bar{p}_1 - \bar{p}_1)\delta_1 < 0$, or $\bar{p}_1 - p_1 > 0$. Any good falls in price when the excess demand for the numeraire rises at the expense of that good (see Allingham, 1975).

A type of stability has been proved for competitive equilibrium over time which concerns the path of equilibrium prices over real time rather than the path of disequilibrium prices over virtual time, that is, the time of the tâtonnement. It was shown by Negishi (1960) that there is a social welfare function associated with a competitive equilibrium which is maximized in the equilibrium over feasible allocations. Suppose each consumer has a concave utility function which is given by a discounted sum of periodwise utilities. Then the social welfare function which is maximized is also a discounted sum of periodwise utilities equal to a weighted sum of the individual utilities. Then using results from turnpike theory for optimal capital accumulation it has been shown by Bewley (1982) that the competitive equilibrium allocations converge over time to the allocations of a stationary competitive equilibrium whose capital stocks and allocations are the same as those of the unique optimal stationary path of capital accumulation given the social welfare function. The utility functions and the production

functions are assumed to be strictly concave and the discount factors are the same for all consumers and sufficiently near 1. However, these conditions may be relaxed.

Comparative static and comparative dynamic results have been derived from stability conditions in the context of optimal capital accumulation, which is equivalent to competitive equilibrium over time with a representative consumer. We may say that an optimal stationary path of capital is regular if an increase in the discount factor implies an increase in the value of capital stocks at initial prices. Then there are sufficient conditions for local stability of the optimal stationary path which imply that the path is regular. Similar dynamic results may be achieved for non-stationary paths as well (see Araujo and Scheinkman, 1979). It may be possible to extend these results to Bewley-type economies.

See Also

- Arrow–Debreu model of general equilibrium
- existence of general equilibrium
- mathematical economics
- overlapping generations model of general equilibrium
- uncertainty and general equilibrium

Bibliography

Allingham, M. 1975. *General Equilibrium*. New York: Wiley.

Araujo, A.P. de and Scheinkman, J.A. 1979. Notes on comparative dynamics. In *General Equilibrium, Growth, and Trade*, ed. J.R. Green and J.A. Scheinkman, New York: Academic Press.

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium*, ed. J. Neyman, Berkeley: University of California Press.

Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie*, Paris: Centre National de la Recherche Scientifique. Trans. as ‘The role of securities in the optimal allocation of risk-bearing’, *Review of Economic Studies* 31, (1964), 91–6.

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Arrow, K.J. and Hurwicz, L. 1958. On the stability of the competitive equilibrium I. *Econometrica* 26, 522–52.

Arrow, K.J., Block, H.D. and Hurwicz, L. 1959. On the stability of the competitive equilibrium II. *Econometrica* 27, 82–109.

Balasko, Y. and Shell, K. 1980. The overlapping generations model, I: the case of pure exchange without money. *Journal of Economic Theory* 23, 281–306.

Balasko, Y., Cass, D. and Shell, K. 1980. Existence of competitive equilibrium in a general overlapping generations model. *Journal of Economic Theory* 23, 307–22.

Bewley, T.F. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4, 514–40.

Bewley, T.F. 1982. An integration of equilibrium theory and turnpike theory. *Journal of Mathematical Economics* 10, 233–68.

Cassel, G. 1918. *Theoretische Sozialökonomie*. 5th German edn, trans. as *The Theory of Social Economy*, New York: Harcourt Brace, 1932.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. Trans. as *Researches into the Mathematical Principles of the Theory of Wealth*, New York: Kelley, 1960.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40, 588–92.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Debreu, G. 1962. New concepts and techniques for equilibrium analysis. *International Economic Review* 3, 257–73.

Gossen, H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs*. 3rd edn, Berlin: Prager, 1927.

Grandmont, J.M. 1973. On the efficiency of a monetary equilibrium. *Review of Economic Studies* 40, 149–65.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45, 535–72.

Grandmont, J.M. 1983. *Money and Value*. New York: Cambridge University Press.

Grandmont, J.M. and Younes, Y. 1972. On the role of money and the existence of a monetary equilibrium. *Review of Economic Studies* 39, 355–72.

Green, J.R. 1973. Temporary general equilibrium in a sequential trading model with spot and future transactions. *Econometrica* 41, 1103–23.

Green, J.R. 1977. The nonexistence of informational equilibria. *Review of Economic Studies* 44, 451–63.

Grossman, S.J. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48, 541–60.

Hahn, F.H. and Negishi, T. 1962. A theorem on non-tâtonnement stability. *Econometrica* 30, 463–9.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan; 5th edn, New York: Kelley and Millman, 1957.

Koopmans, T.C. 1951. Analysis of production as an efficient combination of activities. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans, Wiley: New York.

Kreps, D.M. 1977. A note on fulfilled expectations equilibria. *Journal of Economic Theory* 14, 32–43.

McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22, 147–61.

McKenzie, L.W. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27, 54–71.

McKenzie, L.W. 1981. The classical theorem on existence of competitive equilibrium. *Econometrica* 49, 819–41.

Marshall, A. 1890. *Principles of Economics*. 8th edn, London: Macmillan, 1920.

Menger, C. 1871. *Grundsätze der Volkswirtschaflehre*. Vienna. Trans. as *Principles of Economics*, Glencoe, Ill.: Free Press, 1950.

Mill, J.S. 1848. *Principles of Political Economy*. London: Parker. New edn, London: Longmans, 1909.

Negishi, T. 1960. Welfare economics and existence of an equilibrium for a competitive economy. *Metroeconomica* 12, 92–7.

Neumann, J. von. 1937. Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, 73–83. Trans. in *Review of Economic Studies* 13, (1945), 1–9.

Nikaido, H. 1956. On the classical multilateral exchange problem. *Metroeconomica* 8, 135–45.

Novshek, W. and Sonnenschein, H. 1980. Small efficient scale as a foundation for Walrasian equilibrium. *Journal of Economic Theory* 22, 243–55.

Pareto, V. 1909. *Manuel d'économie politique*. Paris. Trans. from 1927 edn as *Manual of Political Economy*, New York: Kelley, 1971.

Peleg, B. and Yaari, M.E. 1970. Markets with countably many commodities. *International Economic Review* 11, 369–77.

Poinsot, L. 1803. *Eléments de statique*. 8th edn, Paris, 1842.

Radner, R. 1967. Equilibre des marchés à terme et au comptant en cas d'incertitude. *Cahiers d'Econométrie*, Paris: CNRS, 4, 35–52.

Radner, R. 1972. Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica* 40, 289–303.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Samuelson, P.A. 1958. An exact consumption – loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66, 467–82.

Scarf, H.F. (With T. Hansen) 1973. *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 5th edn, ed. E. Cannan, London: Methuen, 1906.

Wald, A. 1935. Über die eindeutige positive Losbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines mathematischen Kolloquiums* 6, 12–20.

Wald, A. 1936a. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalkönomie* 7, 637–70. Trans. as 'On some systems of equations of mathematical economics', *Econometrica* 19, 1951, 368–403.

Walras, L. 1874–7. *Elements d'économie politique pure*. Lausanne: Corbaz. Trans. by W. Jaffé as *Elements of Pure Economics*, London: George Allen & Unwin, from the 1926 definitive edition, 1954.

Weintraub, E.R. 1983. On the existence of a competitive equilibrium: 1930–1954. *Journal of Economic*

Literature 21, 1–39.

Wilson, C.A. 1981. Equilibrium in dynamic models with an infinity of agents. *Journal of Economic Theory* 24, 95–111.

How to cite this article

McKenzie, Lionel W. "general equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000023> doi:10.1057/9780230226203.0623

The New Palgrave Dictionary of Economics Online

general purpose technologies

Peter L. Rousseau
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists have come to use the term ‘general purpose technology’ (GPT) to describe technological advances that pervade many sectors, improve rapidly, and spawn further innovations. This article addresses the concept of a GPT by example, showing the extent to which electricity and information technology might qualify as members of this special class of inventions, as opposed to more ordinary ones.

Keywords

diffusion of technology; electricity; general purpose technologies; growth, models of; information technology and economic growth; innovation; intertemporal substitution; patents; productivity growth; skill premium; technical change

Article

Economists have long been interested in how technological change affects long-run growth and aggregate fluctuations, yet it remains most often treated as incremental in nature, adding only a trend to standard growth models. History tells us, however, that such change can appear in bursts, with flurries of innovative activity following the introduction of a new core technology. This observation leads economists to reserve the term ‘general-purpose technology’ (GPT) to describe fundamental advances that drive these flurries, which in turn transform both household life and the ways in which firms conduct business. Over the past 200 years or so, steam, electricity, internal combustion, and information technology (IT) seem to have served as GPT-type technologies. They affected entire economies. Earlier, the very ability to communicate in writing and later to disseminate written information via the printed page also appears to fit well into the idea of a GPT.
The notions that GPTs differ from the more incremental refinements that occur in between their arrivals and that they represent real-side shocks that permanently change the nature of production and preferences provide the basis of a potentially useful way to organize thinking about long-run economic fluctuations and growth. But to support such a view with anything more than casual observation, it is necessary to establish criteria for determining just what features a technology must possess in order to be a GPT rather than a more ordinary invention. This article defines GPTs in terms of a number of tangible criteria, and then uses two candidate GPTs, electrification and IT, to demonstrate how identification of a GPT might proceed. Attention then turns to other indicators that may signal the start of a GPT era.

Dating a GPT's arrival

Associating a point in time with a GPT's ‘arrival’ depends on what exactly one means by this term. If defined with a measure such as, in the case of electrification, attaining a one per cent share of horsepower in the manufacturing sector, then some time around 1895 might be appropriate. This coincides roughly with the start-up of the world's first large scale hydroelectric power facility at Niagara Falls, New York, in 1894. It would be reasonable to argue, however, that electricity arrived earlier, perhaps in 1882 when Thomas Edison brought the first centralized electricity system online at the Pearl Street station in lower Manhattan. For IT, it is true that mainframe computers had existed for two decades before the invention of the 4004 chip in 1971, and had even been used to project the winner of the 1952 US presidential election. Yet, if measured by the attainment of a one per cent share in the

industrial sector's stock of equipment, 1971 remains the most likely candidate for dating IT's ‘arrival’.

Whether electricity and IT arrived in 1895 and 1971, respectively, or some time prior to these dates, one characteristic noted by David (1991) is that neither delivered productivity gains immediately. Indeed, productivity growth as measured by output per man-hour seems to have been relatively high in the 1870s, when steam was the dominant power source for industry, but fell as electrification arrived in the 1880s and 1890s. It was only in the period after 1915, which also saw the diffusion of secondary motors and the widespread establishment of centralized power distribution systems, that measured productivity numbers began to rise. (This can be seen in the series for output per man-hour in the non-farm business sector from US Census Bureau, 1975, Series D684, p. 162.) Further, Intel's 1971 invention of the 4004 microprocessor (the key component in the first generation of personal computers), if taken to be the start of the IT era, did not reverse the decline in productivity growth that had begun more than a decade earlier.

Identification of a new core technology as a GPT

Once the arrival date of a new technology has been established, identification of that technology as a GPT can proceed by considering characteristics associated with its diffusion. One set of criteria, proposed by Bresnahan and Trajtenberg (1995), suggests that a GPT should have the following three characteristics:

1. 1. *Pervasiveness:* the GPT should spread to most sectors.
2. 2. *Improvement:* the GPT should get better over time and, hence, should keep lowering the costs of its users.
3. 3. *Innovation spawning:* the GPT should make it easier to invent and produce new products or processes.

Most technologies possess each of these characteristics to some degree, and therefore a GPT cannot differ qualitatively from them. But the extent to which technologies have all three characteristics should determine which ones are likely to be GPTs.

For example, both electrification and IT were pervasive, and so might qualify as GPTs under the first criterion, yet had quite different absorption paths across sectors. Figure 1 shows the shares of total horsepower electrified in manufacturing sectors at ten-year intervals from 1889 to 1954 in percentile form, with the shaded area highlighting the period of electricity's most rapid diffusion. Figure 2 shows the spread of IT, measured as the share of IT equipment in the capital stock at the two-digit standard industry classification level. The striking difference between the two figures is that electricity diffused uniformly across sectors while the adoption of IT was not as widespread. On this count, then, electricity would be the stronger GPT candidate.

Figure 1

Shares of electrified horsepower by manufacturing sector in percentiles, 1890–1954. *Source*: DuBoff (1964, Tables E-ll and E-12–12e).

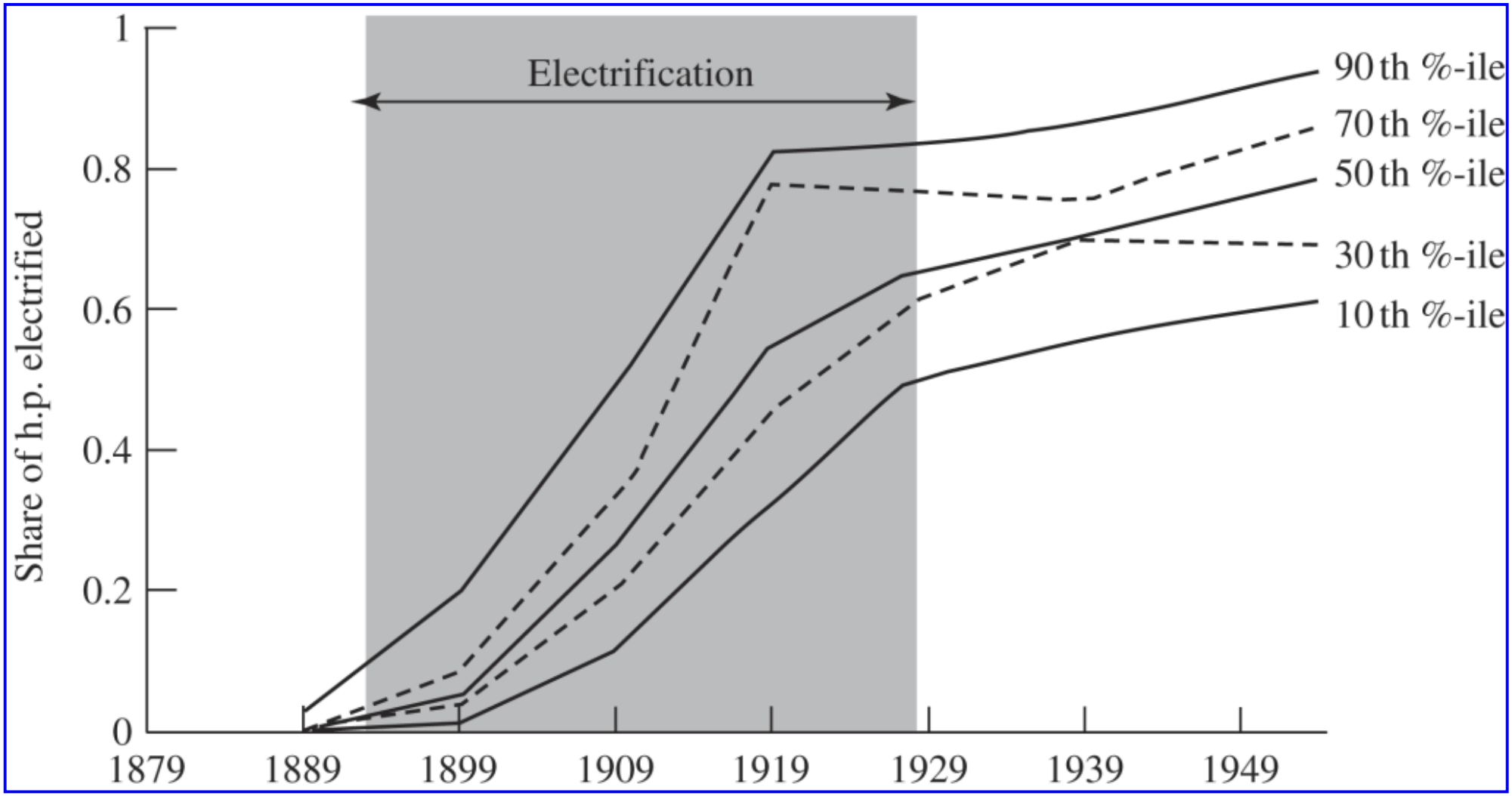

Figure 2
Shares of IT equipment and software in the capital stock by sector in percentiles, 1960–2001. *Source*: Detailed non-residential fixed asset tables in fixed 1996 dollars made available by the US Bureau of Economic Analysis (2004).

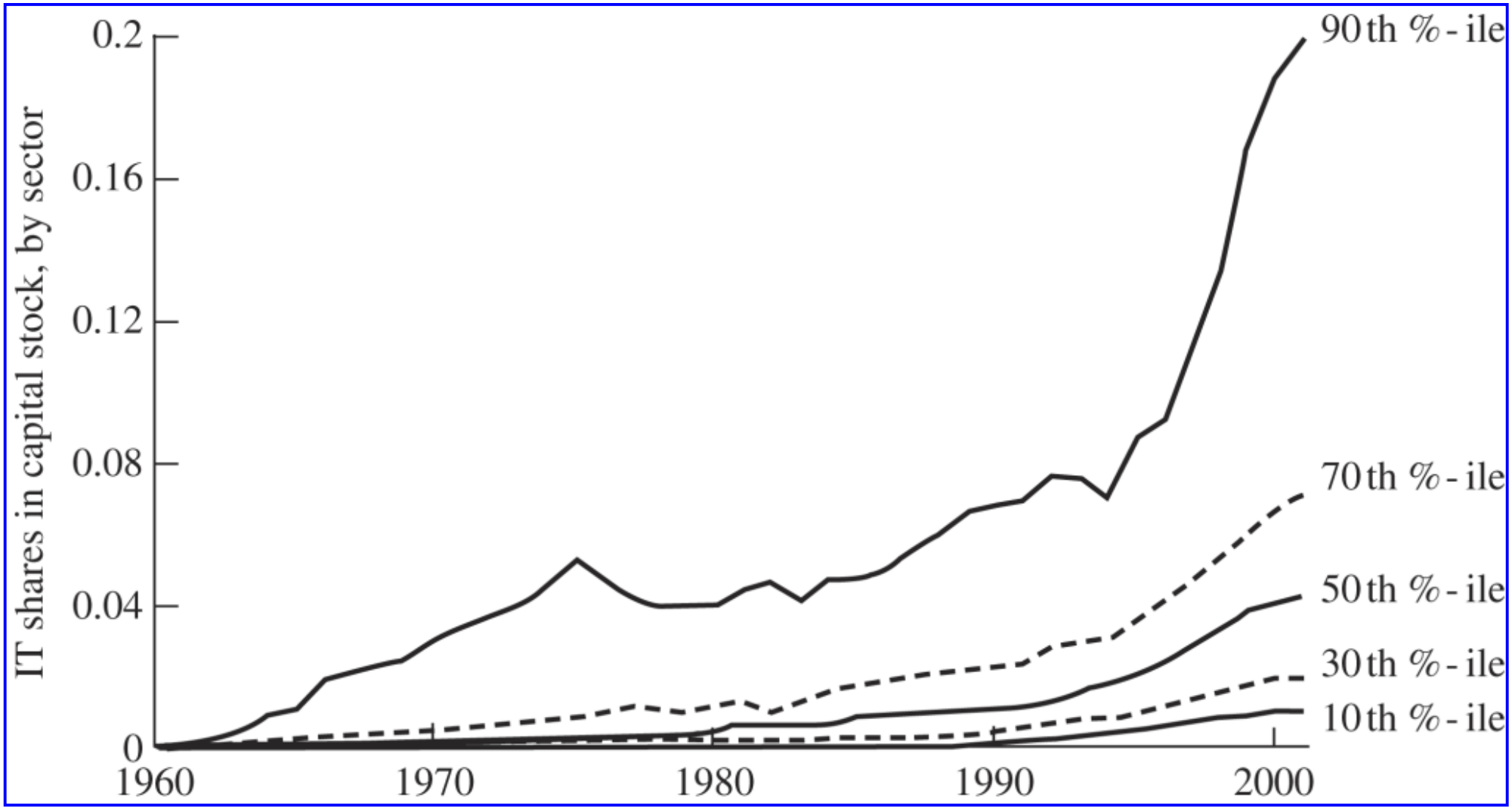

Presumably, the second characteristic – improvement – would show up in a decline in prices associated with the technology, an increase in quality, or both. How much a GPT improves can therefore be measured by how much cheaper a unit of quality gets over time. If the new technology is embodied in capital and begins to account for an increasing share of the net capital stock, capital should on the whole be getting cheaper faster during a GPT era, but especially capital that is tied to the new technology.

Figure 3 plots the price of the components of the aggregate capital stock tied to the two GPTs. Because deflators for electrically powered capital are not available in the first half of the 20th century, the figure compares the declines in relative price of electricity itself with the quality-adjusted price of computers, both relative to the consumption price index. The use of the left-hand scale for electricity and the right-hand scale for computers underscores the extraordinary decline in computer prices since 1960 relative to electricity. While electricity prices fall by a factor of 10, the computer price index falls by a factor of 10,000!

Figure 3

Price indices for products of two ‘GPT eras’, 1895–2000. *Sources*: The quality-adjusted price index for IT is formed by joining the ‘final’ price index for computer systems from Gordon (1990, Table 6.10, col. 5, p. 226) for 1960–78 with the pooled index developed for desktop and mobile personal computers by Berndt, Dulberger and Rappaport (2000, Table 2, col. 1, p. 22) for 1979–99. Electricity prices are averages of all electric energy services in cents per kilowatt hour from US Census Bureau (1975, series S119, p. 827) for 1903, 1907, 1917, 1922, and 1926–70, and from the US Census Bureau, *Statistical Abstract of the United States*, for 1971–89. For 1990–2000, prices are US city averages (June figures) from the US Bureau of Labor Statistics. Both indices are set to 1,000 in the first years of the samples (that is, 1903 and 1960).

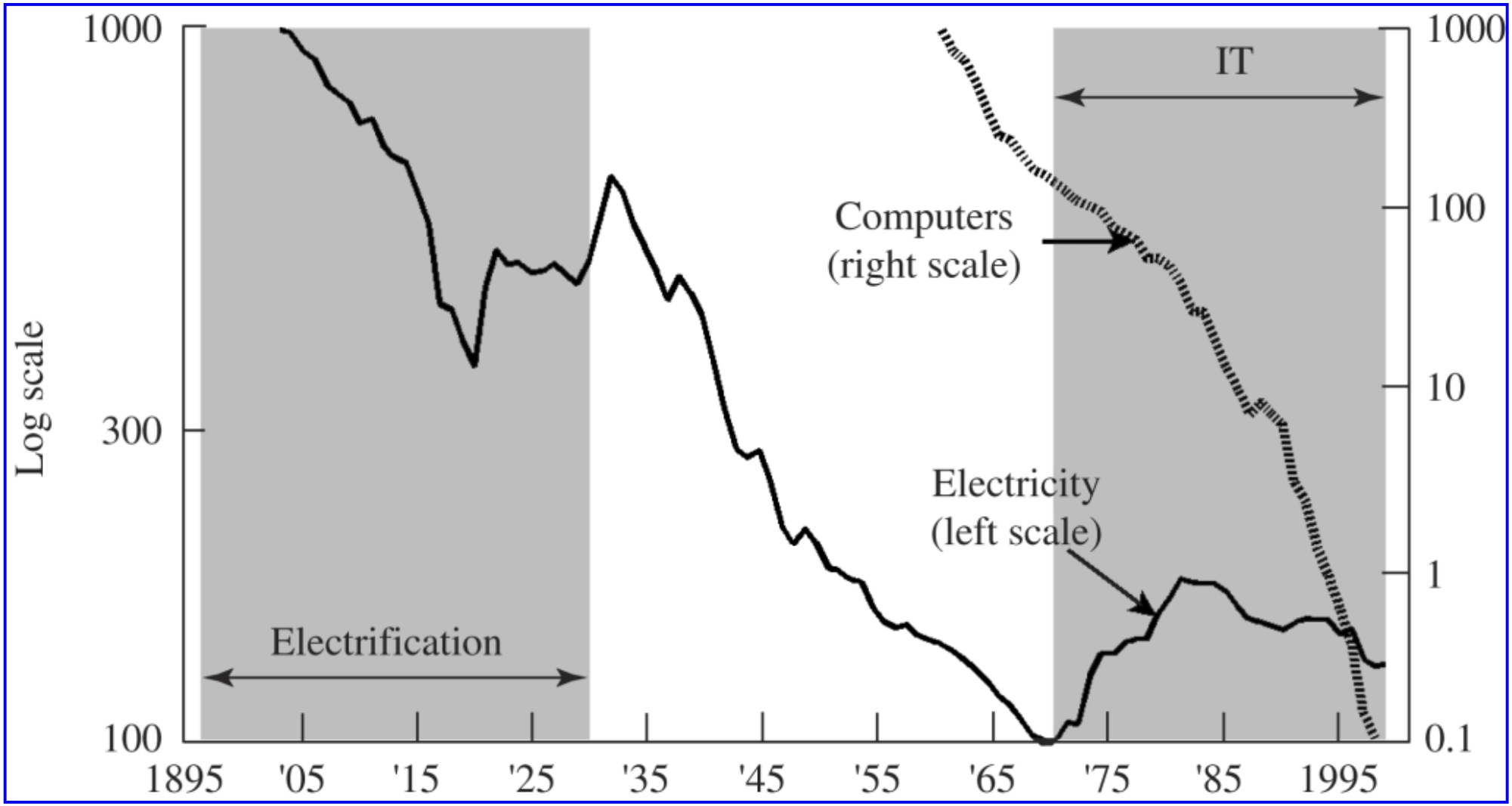

It can be said that the electricity index, being the price of a kilowatt hour, understates the accompanying technological change because it does not account for improvements in electrical equipment, and especially improvements in the efficiency of electrical motors. Based on the price evidence in Figure 3, however, both electricity and computers might qualify as GPTs, with computers clearly more revolutionary.

With respect to the ability to generate further innovation, it is reasonable to assume that any GPT will affect all sorts of production processes, including those for invention and innovation. Some GPTs will be biased towards helping to produce existing products, others towards inventing and implementing new ones. Electricity and IT have both helped reduce the costs of making existing products, and they both spawn innovation. The 1920s especially saw a wave a new products powered by electricity, and the computer is now embodied in many new products as well. But the evidence suggests that IT has contributed more to furthering innovation.

In particular, patenting should be more intense after a GPT arrives and while it is spreading due to the introduction of related new products. US patent data confirm this, showing two surges in the annual number of invention patents issued per capita from 1890 to 2000 – one between 1900 and 1930, and the other after 1977. At the same time, the surge during the IT period was stronger than that observed during electrification. Interestingly, the slow rate of patenting during the Second World War years and the acceleration immediately thereafter suggests that there is some degree of intertemporal substitution in the release of new ideas away from times when they might be more difficult to popularize and towards times better suited for the entry of new products.

Of course, patent data may reflect fluctuations in the number of actual inventions or may simply reflect changes in the law that raise the propensity to patent. The distinction is important because, over longer periods of time, patents may reflect policy rather than invention. Kortum and Lerner (1998) analyse this question and find that the surge of the 1990s

was worldwide, but not systematically related to country-specific policy changes, and they conclude that technology was the cause of the surge.

Other characteristics of GPTs

In addition to the three basic qualities of a GPT, there are other, less direct signals implied by various theoretical models that deal with GPTs. These models predict the following:

1. 1. *New ideas should come to market faster*. If a new technology has the potential for large productivity gains, firms will spend less time perfecting ideas associated with the new technology in order to realize the gains sooner (see, for example, Jovanovic and Rousseau, 2001).
2. 2. *Entry, exit and mergers should rise*. New technologies may require some relocation of assets from firms that are unable to adopt them effectively to others with managements better equipped for their deployment (see, or example, Jovanovic and Rousseau, 2002).
3. 3. *Young and small firms should do better*. The ideas and products associated with the GPT will often be brought to market by new firms. The market share and market value of young firms should therefore rise relative to old firms.
4. 4. *Stock prices should initially fall*. The value of old capital should fall in anticipation of the new and more productive technology. How fast it falls depends on the way that the market learns of the GPT's arrival (see, for example, Hobijn and Jovanovic, 2001).
5. 5. *Interest rates and the trade deficit should be affected*. The rise in desired consumption relative to output should cause interest rates to rise or the trade balance to worsen.
6. 6. *The skill premium should rise*. If the GPT is not user-friendly at first, skilled people will be in greater demand when the new technology arrives, and their earnings should rise compared with those of the unskilled.

The available evidence suggests that predictions 1–3 hold for both the electrification and IT eras, but that a stock market decline (4) occurred only at the start of the IT period. Interest rates (5) rose in both eras, but the electrification period was associated with a trade surplus due to the First World War. It also appears that the skill premium (6) has risen over the IT period, but evidence of a rise in the electrification era is weaker.

To sum up, based upon the criteria chosen and the available evidence, both electricity and IT were pervasive, improving, and innovation-spawning, and thus seem to qualify as GPTs. At the same time, electricity was more pervasive, affecting sectors faster and more evenly than IT, while IT improved more dramatically, with computer prices falling more than 100 times faster than the price of electricity. IT also seems to have generated more innovation than electricity, and the initial productivity slowdown was also deeper in the IT era. All this would lead one to regard IT as the more ‘revolutionary’ GPT.

This is not to say that the differences between electrification and IT, or indeed between any two candidate GPTs, are unimportant. At the same time, the GPT paradigm emphasizes the commonalities, namely, that technological progress is uneven, that it does entail the episodic arrival of new core technologies, and that these GPTs bring on turbulence and lower growth early on and higher growth and prosperity later. Interestingly, the IT era has already outlasted that of electrification, but even six decades after what Field (2003) has called the ‘most technologically progressive decade of the century’ (that is, the 1930s), electricity has yet to become obsolete. Given the multitude of firms and households that have not quite yet adopted IT, its continuing price decline and the widespread increases in computer literacy among children and adults worldwide suggest that perhaps the most productive period of this GPT still lies ahead.

See Also

- diffusion of technology
- electricity markets
- information technology and the world economy
- technical change

The author acknowledges financial support from the National Science Foundation.

Bibliography

Berndt, E.R., Dulberger, E.R. and Rappaport, N.J. 2000. Price and quality of desktop and mobile personal computers: a quarter century of history. Working paper. Cambridge, MA:

Sloan School of Management, MIT.

Bresnahan, T.F. and Trajtenberg, M. 1995. General purpose technologies: engines of growth? *Journal of Econometrics* 65, 83–108.

David, P.A. 1991. Computer and dynamo: the modern productivity paradox in a not-too-distant mirror. In *Technology and Productivity: The Challenge for Economic Policy*. Paris: OECD.

DuBoff, R.B. 1964. Electric power in American manufacturing, 1889–1958. Ph.D. thesis, University of Pennsylvania.

Field, A. 2003. The most technologically progressive decade of the century. *American Economic Review* 93, 1399–414.

Gordon, R.J. 1990. *The Measurement of Durable Goods Prices*. Chicago: University of Chicago Press.

Hobijn, B. and Jovanovic, B. 2001. The IT revolution and the stock market: evidence. *American Economic Review* 91, 1203–20.

Jovanovic, B. and Rousseau, P.L. 2001. ‘Why wait? A century of life before IPO. *American Economic Review Papers and Proceedings* 91, 336–41.

Jovanovic, B. and Rousseau, P.L. 2002. The Q-theory of mergers. *American Economic Review Papers and Proceedings* 92, 198–204.

Kortum, S. and Lerner, J. 1998. Stronger protection or technological revolution: what is behind the recent surge in patenting? *Carnegie-Rochester Conference Series on Public Policy* 48, 247–304.

US Bureau of Economic Analysis. 2004. *Survey of Current Business*. Washington, DC: Government Printing Office.

US Census Bureau. 1975. *Historical Statistics of the United States, Colonial Times to 1970*. Washington, DC: Government Printing Office.

US Census Bureau (various years). *Statistical Abstract of the United States*. Washington, DC: Government Printing Office.

How to cite this article

Rousseau, Peter L. "general purpose technologies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000205> doi:10.1057/9780230226203.0625

The New Palgrave Dictionary of Economics Online

generalized method of moments estimation

Lars Peter Hansen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Generalized method of moments estimates econometric models without requiring a full statistical specification. One starts with a set of moment restrictions that depend on data and an unknown parameter vector to be estimated. When there are more moment restrictions than underlying parameters, there is family of such estimators. The tractable form of the large sample properties of this family facilitates efficient estimation and statistical testing. This article motivates the method, presents some of the underlying statistical properties and discusses implementation.

Keywords

calibration; central limit theorems; Gauss–Markov theorem; generalized method of moments; identification; instrumental variables; Lagrange multipliers; law of large numbers; likelihood; martingales; maximum likelihood; rational expectations models; sequential estimation; statistical inference; stochastic discount factor models; Wald test

Article

1 Introduction

Generalized method of moments (GMM) refers to a class of estimators constructed from the sample moment counterparts of population moment conditions (sometimes known as orthogonality conditions) of the data generating model. GMM estimators have become widely used, for the following reasons:

1. 1. GMM estimators have large sample properties that are easy to characterize. A family of such estimators can be studied simultaneously in ways that make asymptotic efficiency comparisons easy. The method also provides a natural way to construct tests which take account of both sampling and estimation error.
2. 2. In practice, researchers find it useful that GMM estimators may be constructed without specifying

the full data generating process (which would be required to write down the maximum likelihood estimator.) This characteristic has been exploited in analysing partially specified economic models, studying potentially misspecified dynamic models designed to match target moments, and constructing stochastic discount factor models that link asset pricing to sources of macroeconomic risk.

Books with good discussions of GMM estimation with a wide array of applications include: Cochrane (2001), Arellano (2003), Hall (2005), and Singleton (2006). For a theoretical treatment of this method see Hansen (1982) along with the self-contained discussions in the books. See also Ogaki (1993) for a general discussion of GMM estimation and applications, and see Hansen (2001) for a complementary article that, among other things, links GMM estimation to related literatures in statistics. For a collection of recent methodological advances related to GMM estimation, see the journal issue edited by Ghysels and Hall (2002). While some of these other references explore the range of substantive applications, in what follows we focus more on the methodology.

2 Set-up

As we will see, formally there are two alternative ways to specify GMM estimators, but they have a common starting point. Data are a finite number of realizations of the process $\{x_t \bullet : \bullet t=1, 2,\ldots\}$. The model is specified as a vector of moment conditions:

$$Ef(x_t, \beta_0) = 0$$

where f has r coordinates and β_0 is an unknown vector in a parameter space $\mathbb{P} \subset \mathbb{R}^k$. To achieve identification we assume that on the parameter space $\mathbb{P}$

$$Ef(x_t, \beta) = 0 \text{ if, and only if } \beta = \beta_0. \tag{1}$$

The parameter β_0 is typically not sufficient to write down a likelihood function. Other parameters are needed to specify fully the probability model that underlies the data generation. In other words, the model is only partially specified.
Examples include:

1. (a) linear and nonlinear versions of instrumental variables estimators as in Sargan (1958; 1959), and Amemiya (1974);
2. (b) rational expectations models as in Hansen and Singleton (1982), Cumby, Huizinga and Obstfeld

(1983), and Hayashi and Sims (1983);

3. (c) security market pricing of aggregate risks as described, for example, by Cochrane (2001), Singleton (2006) and Hansen et al. (2007);
4. (d) matching and testing target moments of possibly misspecified models as described by, for example, Christiano and Eichenbaum (1992) and Hansen and Heckman (1996).

Regarding example (a), many related methods have been developed for estimating correctly specified models, dating back to some of the original applications in statistics of method-of-moments-type estimators. The motivation for such methods was computational. See Hansen (2001) for a discussion of this literature and how it relates to GMM estimation. With advances in numerical methods, the fully efficient maximum likelihood method and Bayesian counterparts have become much more tractable. On the other hand, there continues to be an interest in the study of dynamic stochastic economic models that are misspecified because of their purposeful simplicity. Thus moment matching remains an interesting application for the methods described here. Testing target moments remains valuable even when maximum likelihood estimation is possible (for example, see Bontemps and Meddahi, 2005).

2.1 Central limit theory and martingale approximation

The parameter dependent average

$$g_N(\beta) = \frac{1}{N}\sum_{t=1}^{N} f(x_t, \beta)$$

is featured in the construction of estimators and tests. When the law of large numbers is applicable, this average converges to the $Ef(x_t, \beta)$. As a refinement of the identification condition:

$$\sqrt{N}\, g_N(\beta_0) \Rightarrow \text{Normal}(0, V) \tag{2}$$

where $\Rightarrow$ denotes convergence in distribution and V is a covariance matrix assumed to be nonsingular. In an iid data setting, V is the covariance matrix of the random vector $f(x_t, \beta_0)$. In a time series setting:

$$V = \lim_{N \to \infty} NE\left[g_N(\beta_0) g_N(\beta_0)'\right], \tag{3}$$

which is the long-run counterpart to a covariance matrix.
Central limit theory for time series is typically built on martingale approximation (see Gordin,1969; Hall and Heyde, 1980). For many time series models, the martingale approximators can be constructed directly and there is specific structure to the V matrix. A leading example is when $f(x_t, \beta_0)$ defines a conditional moment restriction. Suppose that x_t, t=0,1,… generates a sigma algebra $\mathcal{F}_t$, $E[|f(x_t, \beta_0)|^2] < \infty$ and

$$E\left[f(x_{t+\ell}, \beta_0)|\mathcal{F}_t\right] = 0$$

for some $\ell \geq 1$. This restriction is satisfied in models of multi-period security market pricing and in models that restrict multi-period forecasting. If $\ell = 1$, then g_N is itself a martingale; but when $\ell > 1$ it is straightforward to find a martingale m_N with stationary increments and finite second moments such that

$$\lim_{N \to \infty} E\left[|g_N(\beta_0) - m_N(\beta_0)|^2\right] = 0,$$

where |·| is the standard Euclidean norm. Moreover, the lag structure may be exploited to show that the limit in (3) is

$$V = \sum_{j=-\ell+1}^{\ell-1} E\left[f(x_t, \beta_0) f(x_{t+j}, \beta_0)'\right]. \tag{4}$$

(The sample counterpart to this formula is not guaranteed to be positive semidefinite. There are a variety of ways to exploit this dependence structure in estimation in constructing a positive semidefinite estimate. See Eichenbaum, Hansen and Singleton, 1988, for an example.) When there is no exploitable structure to the martingale approximator, the matrix V is the spectral density at frequency zero.

$$V = \sum_{j=-\infty}^{\infty} E\left[f(x_t, \beta_0) f(x_{t+j}, \beta_0)'\right].$$

2.2 Minimizing a quadratic form

One approach for constructing a GMM estimator is to minimize the quadratic form:

$$b_N = \arg\min_{\beta \in \mathbf{P}} g_N(\beta)' W g_N(\beta)$$

for some positive definite *weighting matrix W*. Alternative weighting matrices *W* are associated with alternative estimators. Part of the justification for this approach is that

$$\beta_0 = \arg\min_{\beta \in \mathbf{P}} Ef(x_t, \beta)' W Ef(x_t, \beta).$$

The GMM estimator mimics this identification scheme by using a sample counterpart.
There are a variety of ways to prove consistency of GMM estimators. Hansen (1982) established a uniform law of large numbers for random functions when the data generation is stationary and ergodic. This uniformity is applied to show that

$$\sup_{\beta \in \mathbf{P}} |g_N(\beta) - E[f(x_t, \beta)]| = 0$$

and presumes a compact parameter space. The uniformity in the approximation carries over directly the GMM criterion function $g_N(\beta)' W g_N(\beta)$. See Newey and McFadden (1994) for a more complete catalogue of approaches of this type.
The compactness of the parameter space is often not ignored in applications, and this commonly invoked result is therefore less useful than it might seem. Instead, the compactness restriction is a substitute for checking behaviour of the approximating function far away from β_o to make sure that spurious optimizers are not induced by approximation error. This tail behaviour can be important in practice, so a direct investigation of it can be fruitful. For models with parameter separation:

$$f(x, \beta) = Xh(\beta)$$

where *X* is an *r*×*m* matrix constructed from *x* and *h* is a one-to-one function mapping $\mathbb{P}$ into subset of $\mathbb{R}^m$,

there is an alternative way to establish consistency (see Hansen, 1982 for details). Models that are either linear in the variables or models based on matching moments that are nonlinear functions of the underlying parameters can be written in this separable form.

The choice of $W=V^{-1}$ receives special attention, in part because

$$N g_N(\beta)' V^{-1} g_N(\beta) \Rightarrow \chi^2(r).$$

While the matrix V is typically not known, it can be replaced by a consistent estimator without altering the large sample properties of b_N. When using martingale approximation, the implied structure of V can often be exploited as in formula (4). When there is no such exploitable structure, the method of Newey and West (1987b) and others can be employed that are based on frequency-domain methods for time series data.

For asset pricing models there are other choices of a weighting matrix motivated by considerations of misspecification. In these models with parameterized stochastic discount factors, the sample moment conditions $g_N(\beta)$ can be interpreted as a vector of pricing errors associated with the parameter vector β. A feature of $W=V^{-1}$ is that, if the sample moment conditions (the sample counterpart to a vector pricing errors) happened to be the same for two models (two choices of β), the one for which the implied asymptotic covariance matrix is larger will have a smaller objective. Thus there is a *reward* for parameter choices that imply variability in the underlying central limit approximation. To avoid such a reward, it is also useful to compare models or parameter values in other ways. An alternative weighting matrix is constructed by minimizing the least squares distance between the parameterized stochastic discount factor and one among the family of discount factors that correctly price the assets. Equivalently, parameters or models are selected on the basis of the maximum pricing error among constant weighted portfolios with payoffs that have common magnitude (a unit second moment). See Hansen and Jagannathan (1997) and Hansen, Heaton and Luttmer (1995) for this and related approaches.

2.3 Selection matrices

An alternative depiction is to introduce a *selection matrix A* that has dimension $k \times r$ and to solve the equation system:

$$A g_N(\beta) = 0$$

for some choice of β, which we denote b_N. The selection matrix A reduces the number of equations to be solved from r to k. Alternative selection matrices are associated with alternative GMM estimators. By relating estimators to their corresponding selection matrices, we have a convenient device for studying simultaneously an entire family of GMM estimators. Specifically, we explore the consequence of using alternative subsets of moment equations or more generally alternative linear combinations of the moment

equation system. This approach builds on an approach of Sargan (1958; 1959) and is most useful for characterizing limiting distributions. The aim is to study simultaneously the behaviour of a family of estimators. When the matrix *A* is replaced by a consistent estimator, the asymptotic properties of the estimator are preserved. This option expands considerably the range of applicability, and, as we will see, is important for implementation.

Since alternative choices of *A* may give rise to alternative GMM estimators, index alternative estimators by the choice of *A*. In what follows, replacing *A* by a consistent estimator does not alter the limiting distribution. For instance, the first-order conditions from minimizing a quadratic form can be represented using a selection matrix that converges to a limiting matrix *A*. Let

$$D = E\left[\frac{\partial f(x_t, \beta_0)}{\partial \beta}\right].$$

Two results are central to the study of GMM estimators:

$$\sqrt{N}(b_N - \beta_0) \approx -(AD)^{-1}A\sqrt{N}g_N(\beta_0) \tag{5}$$

and

$$\frac{1}{\sqrt{N}}g_N(b_N) \approx \left[I - D(AD)^{-1}D\right]\sqrt{N}g_N(\beta_0). \tag{6}$$

Both approximation results are expressed in terms of $\sqrt{N}g_N(\beta_0)$, which obeys a central limit theorem, see (2). These approximation results are obtained by standard local methods. They require the square matrix *AD* to be nonsingular. Thus, for there to exist a valid selection matrix, *D* must have full column rank *k*. Notice from (6) that the sample moment conditions evaluated at b_N have a degenerate distribution. Pre-multiplying by *A* makes the right-hand side zero. This is to be expected because linear combinations of the sample moment conditions are set to zero in estimation.

In addition to assess the accuracy of the estimator (approximation (5)) and to validate the moment conditions (approximation (6)), Newey and West (1987a) and Eichenbaum, Hansen and Singleton (1988) show how to use these and related approximations to devise tests of parameter restrictions. (Their tests imitate the construction of the likelihood ratio, Lagrange multiplier and the Wald tests familiar from likelihood inference methods.)

Next we derive a sharp lower bound on the asymptotic distribution of a family of GMM estimators indexed by the selection matrix *A*. For a given *A*, the asymptotic covariance matrix for a GMM estimator constructed using this selection is:

$$cov(A) = (AD)^{-1}AVA'(D'A')^{-1}.$$

A selection matrix in effect over-parameterizes a GMM estimator, as can be seen from this formula. Two such estimators with selection matrices of the form *A* and *BA* for a nonsingular matrix *B* imply

$$cov(BA) = cov(A)$$

because the same linear combinations of moment conditions are being used in estimation. Thus without loss of generality we may assume that *AD=I*. With this restriction we may imitate the proof of the famed Gauss–Markov Theorem to show that

$$D'V^{-1}D \leq cov(A) \tag{7}$$

and that the lower bound on left is attained by any $\tilde{A}$ such that $\tilde{A} = BD'V^{-1}$ for some nonsingular *B*. The quadratic form version of a GMM estimator typically satisfies this restriction when W_N is a consistent estimator of V^{-1}. This follows from the first-order conditions of the minimization problem.

To explore further the implications of this choice, factor the inverse covariance matrix V^{-1} as $V^{-1} = \Lambda'\Lambda$ and form $\Delta = \Lambda D$. Then

$$V^{-1}D(D'V^{-1}D)^{-1}D'V^{-1} = \Lambda'[\Delta(\Delta'\Delta)^{-1}\Delta']\Lambda.$$

The matrices $\Delta(\Delta'\Delta)^{-1}\Delta'$ and $I - \Delta(\Delta'\Delta)^{-1}\Delta'$ are each idempotent and

$$\begin{bmatrix} [I-\Delta(\Delta'\Delta)^{-1}\Delta'] \\ \Delta(\Delta'\Delta)^{-1}\Delta' \end{bmatrix} \sqrt{N}\Lambda g_N(\beta_0) \to \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I-\Delta(\Delta'\Delta)^{-1}\Delta' & 0 \\ 0 & \Delta(\Delta'\Delta)^{-1}\Delta' \end{bmatrix}\right).$$

The first coordinate block is an approximation for $\sqrt{N}\Lambda g_N(b_N)$ and the sum of the two coordinate blocks is $\sqrt{N}\Lambda g_N(\beta_o)$. Thus we may decompose the quadratic form

$$N[g_N(\beta_0)]'V^{-1}g_N(\beta_0) \approx N[g_N(b_N)]'V^{-1}g_N(b_N) + N[g_N(\beta_0)]'V^{-1}D(D'V^{-1}D)^{-1}D'V^{-1}g_N(\beta_0). \quad (8)$$

where the two terms on the right-hand side are distributed as independent chi-square. The first has *r* degrees of freedom and the second one has *r–k* degrees of freedom.

3 Implementation using the objective function curvature

While the formulas just produced can be used directly using consistent estimators of *V* and *D* in conjunction with the relevant normal distributions, looking directly at the curvature of the GMM objective function based on a quadratic form is also revealing. Approximations (5) and (6) give guidance on how to do this. For a parameter vector β let $V_N(\beta)$ denote an estimator of the long-run covariance matrix. Given an initial consistent estimator b_N, suppose that $V_N(b_N)$ is a consistent estimator of *V* and

$$D_N = \frac{1}{N}\sum_{t=1}^{N}\frac{\partial f(x_t, b_N)}{\partial \beta}.$$

Then use of the selection $A_N = D_N'[V_N(b_N)]^{-1}$ attains the efficiency bound for GMM estimators. This is the so-called two-step approach to GMM estimation. Repeating this procedure, we obtain the so-called iterative estimator. (There is no general argument that repeated iteration will converge.) In the remainder of this section we focus on a third approach, resulting in what we call the continuous-updating estimator. This is obtained by solving:

$$\min_{\beta\in\mathbf{P}} L_N(\beta)$$

where

$$L_N(\beta) = N[g_N(\beta)]'[V_N(\beta)]^{-1}g_N(\beta).$$

Let b_N denote the minimized value. Here the weighting matrix varies with β.
Consider three alternative methods of inference that look at the global properties of the GMM objective $L_N(\beta)$:

1. (a) $\{\beta \in \mathbb{P}: L_N(\beta) \leq C\}$ where C is a critical value from a $\chi^2(r)$ distribution.
2. (b) $\{\beta \in \mathbb{P}: L_N(\beta) - L_N(b_N) \leq C\}$ where C is a critical value from a $\chi^2(k)$ distribution.
3. (c) Choose a *prior* π. Mechanically, treat $-\frac{1}{2}L_N(\beta)$ as a log-likelihood and compute

$$\frac{\exp\left[-\frac{1}{2}L_N(\beta)\right]\pi(\beta)}{\int \exp\left[-\frac{1}{2}L_N(\beta)\right]\pi(\tilde{\beta})\,d\tilde{\beta}}.$$

Method (a) is based on the left-hand side of (8). It was suggested and studied in Hansen, Heaton and Luttmer (1995) and Stock and Wright (2000). As emphasized by Stock and Wright, it avoids using a local identification condition (a condition that the matrix D have full column rank). On the other hand, it combines evidence about the parameter as reflected by the curvature of the objective with overall evidence about the model. A misspecified model will be reflected as an empty confidence interval.
Method (b) is based on the second term on right-hand side of (8). By translating the objective function, evidence against the model is netted out. Of course it remains important to consider such evidence because parameter inference may be hard to interpret for a misspecified model. The advantage of (b) is that the degrees of freedom of the chi-square distribution are reduced from r to k. Extensions of this approach to accommodate nuisance parameters were used by Hansen and Singleton (1996) and Hansen, Heaton and Luttmer (1995). The decomposition on the right-hand side of (8) presumes that the parameter is identified locally in the sense that D has full column rank, guaranteeing that the $D'V^{-1}D$ is nonsingular. Kleibergen (2005) constructs an alternative decomposition based on a weaker notion of identification that can be used in making statistical inferences.
Method (c) was suggested by Chernozhukov and Hong (2003). It requires an integrability condition which will be satisfied by specifying a uniform distribution π over a compact parameter space. The resulting histograms can be sensitive to the choice of this set or more generally to the choice of π. All three methods explore the global shape of the objective function when making inferences. (The large sample justification remains local, however.)

4 Backing off from efficiency

In what follows we give two types of applications that are not based on efficient GMM estimation.

4.1 Calibration-verification

An efficient GMM estimator selects the best linear combination among a set of moment restrictions. Implicitly a test of the over-identifying moment conditions examines whatever moment conditions are not used in estimation. This complicates the interpretation of the resulting outcome. Suppose instead there is one set of moment conditions for which we have more confidence and are willing to impose for the purposes and calibration or estimation. The remaining set of moment conditions are used for the purposes of verification or testing. The decision to use only a subset of the available moment conditions for purposes of estimation implies a corresponding loss in efficiency. See Christiano and Eichenbaum (1992) and Hansen and Heckman (1996) for a discussion of such methods for testing macroeconomic models.
To consider this estimation problem formally, partition the function *f* as:

$$f(x, \beta) = \begin{bmatrix} f^{[1]}(x, \beta) \\ f^{[2]}(x, \beta) \end{bmatrix}$$

where $f^{[1]}$ has r_1 coordinates and $f^{[2]}$ has $r-r_1$ coordinates. Suppose that $r_1 \geq k$ and that β is estimated using an *A* matrix of the form:

$$A = [A_1 \quad 0],$$

and hence identification is based only on

$$A_1 E f^{[1]}(x_t, \beta) = 0.$$

This is the so-called calibration step. Let b_N be the resulting estimator.

To verify or test the model we check whether $g_N^{[2]}(b_N)$ is close to zero as predicted by the moment implication:

$$Ef^{[2]}(x_t, \beta_0) = 0.$$

Partition the matrix D of expected partial derivatives as:

$$D = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}$$

where D_1 is r_1 by k and D_2 is $r-r_1$ by k. Here we use limit approximation (6) to conclude that

$$\sqrt{N} g_N^{[2]}(b_N) \approx \begin{bmatrix} -D_2(A_1 D_1)^{-1} A_1 & I \end{bmatrix} \sqrt{N} g_N(\beta_0),$$

which has a limiting normal distribution. A chi-square test can be constructed by building a corresponding quadratic form of $r-r_1$ asymptotically independent standard normally distributed random variables. (When r_1 exceeds k it is possible to improve the asymptotic power by exploiting the long-run covariation between $f^{[2]}(x_t, \beta_o)$ and linear combination of $f^{[1]}(x_t, \beta_o)$ not used in estimation. This can be seen formally by introducing a new parameter $\gamma_o = E[f^{[2]}(x_t, \beta)]$ and using the GMM formulas for efficient estimation of β_o and γ_o.)

4.2 Sequential estimation

Sequential estimation methods have a variety of econometric applications. For models of sample selection see Heckman (1976), and for related methods with generated regressors see Pagan (1984). For testing asset pricing models, see Cochrane (2001, chs 12 and 13).
To formulate this problem in a GMM setting, partition the parameter vector as

$$\beta = \begin{bmatrix} \beta^{[1]} \\ \beta^{[2]} \end{bmatrix}$$

where $\beta^{[1]}$ has k_1 coordinates. Partition the function f as:

$$f(x, \beta) = \begin{bmatrix} f^{[1]}\left(x, \beta^{[1]}\right) \\ f^{[2]}(x, \beta) \end{bmatrix}$$

where $f^{[1]}$ has r_1 coordinates and $f^{[2]}$ has $r-r_1$ coordinates. Notice that the first coordinate block only depends on the first component of the parameter vector. Thus the matrix d is block lower triangular:

$$D = \begin{bmatrix} D_{11} & 0 \\ D_{21} & D_{22} \end{bmatrix}$$

where

$$D_{ij} = E\left[\frac{\partial f^{[i]}(x_t, \beta_0)}{\partial \beta^{[j]}}\right].$$

A sequential estimation approach exploits the triangular structure of the moment conditions as we now describe. The parameter $\beta_0^{[1]}$ is estimable from the first partition of moment conditions. Given such an estimator, $b_N^{[1]}$, $\beta_0^{[2]}$ is estimable from the second partition of moment conditions. Estimation error in the first stage alters the accuracy of the second stage estimation, as I now illustrate.
Assume now that $r_1 \geq k_1$. Consider a selection matrix that is block diagonal:

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

where A_{11} has dimension k_1 by r_1 and A_{22} has dimension $k-k_1$ by $r-r_1$. It is now possible to estimate $\beta_0^{[1]}$ using the equation system:

$$A_{11} g_N^{[1]}(\beta^{[1]}) = 0$$

or a method that is asymptotically equivalent to this. Let $b_N^{[1]}$ be the solution. This initial estimation may be done for simplicity or because these moment conditions are embraced with more confidence. Given this estimation of $\beta_0^{[1]}$, we seek an estimator $b_N^{[2]}$ of $\beta_0^{[2]}$ by solving:

$$A_{22} g_N^{[2]}\left(b_N^{[1]}, \beta^{[2]}\right) = 0.$$

To proceed, we use this partitioning and apply (5) to obtain the limiting distribution for the estimator $b_N^{[2]}$. Straightforward matrix calculations yield,

$$\sqrt{N}\left(b_N^{[2]} - \beta_0^{[2]}\right) \approx -(A_{22}D_{22})^{-1} A_{22}\left[-D_{21}(A_{11}D_{11})^{-1}A_{11} \quad I\right]\sqrt{N} g_N(\beta_0). \tag{9}$$

This formula captures explicitly the impact of the initial estimation of $\beta_0^{[1]}$ on the subsequent estimation of $\beta_0^{[2]}$. When D_{21} is zero an adjustment is unnecessary.

Consider next a (second-best) efficient choice of selection matrix A_{22}. Formula (9) looks just like formula (5) with A_{22} replacing A, D_{22} replacing D and a particular linear combination of $g_N(\beta_0)$. The matrix used in this linear combination ‘corrects’ for the estimation error associated with the use of an estimator $b_N^{[1]}$ instead of the unknown true value $\beta_0^{[1]}$. By imitating our previous construction of an asymptotically efficient estimator, we construct the (constrained) efficient choice of A_{22} given A_{11}:

$$A_{22} = B_{22}(D_{22})'\left(\left[-D_{21}(A_{11}D_{11})^{-1}A_{11} \quad I\right] V \begin{bmatrix} -\left[D_{21}(A_{11}D_{11})^{-1}A_{11}\right]' \\ I \end{bmatrix}\right)^{-1}$$

for some nonsingular matrix B_{22}. An efficient estimator can be implemented in the second stage by solving:

$$\min_{\beta^{[2]}} g_N^{[2]}\left(b_N^{[1]}, \beta^{[2]}\right)' W_N g_N^{[2]}\left(b_N^{[1]}, \beta^{[2]}\right)$$

for $V_N^{[2]}$ given by a consistent estimator of

$$V^{[2]} = \left(\left[-D_{21}(A_{11}D_{11})^{-1}A_{11} \quad I\right] V \begin{bmatrix} -\left[D_{21}(A_{11}D_{11})^{-1}A_{11}\right]' \\ I \end{bmatrix}\right)^{-1}$$

or by some other method that selects (at least asymptotically) the same set of moment conditions to use in estimation. Thus we have a method that adjusts for the initial estimation of $\beta^{[1]}$ while making efficient use of the moment conditions $Ef^{[2]}(x_t, \beta) = 0$.

As an aside, notice the following. Given an estimate $b_N^{[1]}$, the criterion-based methods of statistical inference described in Section 3 can be adapted to making inferences in this second stage in a straightforward manner.

5 Conditional moment restrictions

The bound (7) presumes a finite number of moment conditions and characterizes how to use these conditions efficiently. If we start from the conditional moment restriction:

$$E\left[f(x_{t+\ell}, \beta_0) | \mathcal{F}_t\right] = 0,$$

then in fact there are many moment conditions at our disposal. Functions of variables in the conditioning information set can be used to extend the number of moment conditions. By allowing for these conditions, we can improve upon the asymptotic efficiency bound for GMM estimation. Analogous conditional moment restrictions arise in cross-sectional settings.

For a characterizations and implementations appropriate for cross-sectional data, see Chamberlain (1986) and Newey (1993), and for characterizations and implementations in a time series settings see Hansen

(1985; 1993), and West (2001). The characterizations are conceptually interesting but reliable implementation is more challenging. A related GMM estimation problem is posed and studied by Carrasco and Florens (2000) in which there is a pre-specified continuum of moment conditions that are available for estimation.

6 Conclusion

GMM methods of estimation and inference are adaptable to a wide array of problems in economics. They are complementary to maximum likelihood methods and their Bayesian counterparts. Their large sample properties are easy to characterize. While their computational simplicity is sometimes a virtue, perhaps their most compelling use is in the estimation of partially specified models or of misspecified dynamic models designed to match a limited array of empirical targets.

See Also

- Bayesian methods in macroeconometrics
- rational expectations models, estimation of
- simulation-based estimation

I greatly appreciate comments from Lionel Melin, Monika Piazzesi, Grace Tsiang and Francisco Vazquez-Grande. This material is based upon work supported by the National Science Foundation under Award Number SES0519372.

Bibliography

Amemiya, T. 1974. The nonlinear two-stage least-squares estimator. *Journal of Econometrics* 2, 105–10.

Arellano, M. 2003. *Panel Data Econometrics*. New York: Oxford University Press.

Bontemps, C. and Meddahi, N. 2005. Testing normality: a GMM approach. *Journal of Econometrics* 124, 149–86.

Carrasco, M. and Florens, J.P. 2000. Generalization of GMM to a continuum of moment conditions. *Econometric Theory* 20, 797–834.

Chamberlain, G. 1986. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34, 305–34.

Chernozhukov, V. and Hong, H. 2003. An MCMC approach to classical estimation. *Journal of Econometrics* 115, 293–346.

Christiano, L.J. and Eichenbaum, M. 1992. Current real business cycle theories and aggregate labor market

fluctuations. *American Economic Review* 82, 430–50.

Cochrane, J. 2001. *Asset Pricing*. Princeton: Princeton University Press.

Cumby, R.E., Huizinga, J. and Obstfeld, M. 1983. Two-step two-stage least squares estimation in models with rational expectations. *Journal of Econometrics* 21, 333–5.

Eichenbaum, M.S., Hansen, L.P. and Singleton, K.J. 1988. A time series analysis of representation agent models of consumption and leisure choice under uncertainty. *Quarterly Journal of Economics* 103, 51–78.

Ghysels, E. and Hall, A. 2002. Editors' Introduction to *JBES* twentieth anniversary issue on generalized method of moments estimation. *Journal of Business and Economic Statistics* 20, 441.

Gordin, M.I. 1969. The central limit theorem for stationary processes. *Soviet Mathematics Doklady* 10, 1174–6.

Hall, A.R. 2005. *Generalized Method of Moments*. New York: Oxford University Press.

Hall, P. and Heyde, C.C. 1980. *Martingale Limit Theory and Its Application*. Boston: Academic Press.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Hansen, L.P. 1985. A method for calculating bound on asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics* 30, 203–38.

Hansen, L.P. 1993. Semiparametric efficiency bounds for linear time-series models. In *Models, Methods and Applications of Econometrics: Essays in Honor of A.R. Bergstrom*, ed. P.C.B. Phillips. Cambridge, MA: Blackwell.

Hansen, L.P. 2001. Method of moments. In *International Encyclopedia of the Social and Behavior Sciences*. New York: Elsevier.

Hansen, L.P., Heaton, J. and Luttmer, E. 1995. Econometric evaluation of asset pricing models. *Review of Financial Studies* 8, 237–74.

Hansen, L.P. and Heckman, J.J. 1996. The empirical foundations of calibration. *Journal of Economic Perspectives* 10(1), 87–104.

Hansen, L.P. and Jagannathan, R. 1997. Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52, 557–90.

Hansen, L.P. and Singleton, K.J. 1982. Generalized instrumental variables of nonlinear rational

expectations models. *Econometrica* 50, 1269–86.

Hansen, L.P. and Singleton, K.J. 1996. Efficient estimation of linear asset pricing models with moving average errors. *Journal of Business and Economic Statistics* 14, 53–68.

Hansen, L.P., Heaton, J.C., Lee, J. and Roussanov, N. 2007. Intertemporal substitution and risk aversion. In *Handbook of Econometrics*, vol. 6A, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Hayashi, F. and Sims, C. 1983. Nearly efficient estimation of time-series models with predetermined, but not exogenous, instruments. *Econometrica* 51, 783–98.

Heckman, J.J. 1976. The common structure of statistical methods of truncation, sample selection, and limited dependent variables and a simple estimator of such models. *Annals of Economic and Social Measurement* 5, 475–92.

Kleibergen, F. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73, 1103–23.

Newey, W. 1993. Efficient estimation of models with conditional moment restrictions. In *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao, and H.D. Vinod. Amsterdam: North-Holland.

Newey, W. and McFadden, D. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

Newey, W.K. and West, K.D. 1987a. Hypothesis testing with efficient method of moments estimation. *International Economic Review* 28, 777–87.

Newey, W.K. and West, K.D. 1987b. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.

Ogaki, M. 1993. Generalized method of moments: econometric applications. In *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao and H.D. Vinod. Amsterdam: North-Holland.

Pagan, A.R. 1984. Econometric issues in the analysis of models with generated regressors. *International Economic Review* 25, 221–47.

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.

Sargan, J.D. 1959. The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society: Series B* 21, 91–105.

Singleton, K.J. 2006. *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment*. Princeton: Princeton University Press.

Stock, J.H. and Wright, J.H. 2000. GMM with weak identification. *Econometrica* 68, 1055–96.

West, K.D. 2001. On optimal instrumental variables estimation of stationary time series models. *International Economic Review* 42, 1043–50.

How to cite this article

Hansen, Lars Peter. "generalized method of moments estimation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000206> doi:10.1057/9780230226203.0626

The New Palgrave Dictionary of Economics Online

generational accounting

Jagadeesh Gokhale

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Many government programmes transfer resources between different population groups. Programmes to provide retirement and health security levy taxes on workers to finance transfers to retirees. Initiating or expanding such programmes often redistributes wealth across generations by altering their lifetime tax burdens. Although standard budget measures such as national debt and deficits do not fully reflect them, such public intergenerational redistributions could substantially affect different generations' economic choices. Generational accounting measures the size of prospective net tax burdens facing different generations under current government tax and expenditure policies. It also analyses how those fiscal burdens would change under alternative policies.

Keywords

aging populations; budget deficits; consumption; fiscal burden; fiscal policy; generational accounting; generational balance; gifts; government intertemporal budget constraint; inheritance and bequests; intergenerational transfers; income taxes; labour productivity; labour supply; labour-force participation; lifetime net tax rates; national debt; redistribution of income and wealth; risk; saving; sensitivity analysis; social insurance; wealth

Article

Before the 1990s, studies of the distributional impact of fiscal policies distinguished between groups according to their income, wealth or consumption at a point in time but not according to their life-cycle stage. Feldstein (1974) first pointed out the possibility of implementing large resource transfers across generations even under balanced government budgets. Nevertheless, notions about the impact of fiscal policies across generations remained limited to a presumed positive association between larger budget deficits and larger tax burdens on future generations.
Auerbach, Gokhale, and Kotlikoff (1991) developed generational accounting, a method for estimating

the economic impact of fiscal policy on different cohorts – including future ones – distinguished by birth year and gender. With rapidly ageing populations in developed countries and growing costs of social insurance programmes that redistribute resources from younger to older generations, the demand for evaluating the intergenerational effects of government fiscal policies increased considerably. As a result, generational accounting is now used as a fiscal-analysis tool in dozens of countries.

Generational accounting (GA) is a method of estimating *prospective per capita lifetime net tax burdens* that different cohorts would face under existing fiscal policies. ‘Prospective’ means that fiscal burdens are evaluated over cohorts’ remaining lifetimes; ‘net tax’ means that government transfers are subtracted from taxes; and ‘lifetime’ indicates that future dollar flows are actuarially discounted back to the present and aggregated into a summary measure of the fiscal burden in present value. Changes in the GAs of different cohorts arising from changes in government tax and spending policies measure fiscal policy-induced changes in those cohorts' lifetime resources.

Generational accounting method

Under current (year t) policies, the present discounted value of the government's projected purchases of goods and services (PVG_t) must be paid for out of the government's current net financial wealth (NW_t), the present value of net tax payments by living generations (PVL_t), and the present value net tax payments by future-born cohorts (PVF_t). In this government intertemporal budget constraint,

$$PVG_t = NW_t + PVL_t + PVF_t, \tag{1}$$

NW_t is calculated as the sum of past budget surpluses – which would be negative if past budgets mostly accrued deficits. The government's real assets, such as land, roads, buildings and public parks, are not included because that would require inclusion of a compensating term on the left-hand-side of eq. (1) – the rental cost of the services those real assets provide.

For calculating PVL_t, official government projections of annual aggregate taxes and transfers are first distributed across officially projected populations using profiles of tax payments and transfer receipts by age and gender obtained from the latest available micro-data surveys. Per capita taxes and transfers for years beyond the government projection horizon are obtained by growing the terminal year's per capita values at the labour productivity growth rate underlying official aggregate projections.

Next, each living cohort's GA is calculated by actuarially discounting its projected net taxes per capita using cohort-specific mortality projections and an assumed rate of discount. Because fiscal dollar flows are more volatile than returns on government bonds but less volatile than private capital returns, an intermediate rate of interest is used. Multiplying each cohort's GA by its year-t population and aggregating across all cohorts yields PVL_t.

PVG_t is calculated by projecting government purchases of goods and services – such as administrative and judicial services, defence, and infrastructure – at current levels per capita using official population

projections, and discounting those amounts back to year *t*. The term PVF_t in eq. (1) is calculated as a residual.

Both PVL_t and PVG_t are calculated by projecting fiscal flows under unchanged policies. PVL_t equals the present value of net taxes that cohorts alive in year *t* would pay collectively if their fiscal treatment remained unchanged throughout their lifetimes. PVG_t indicates the size of the bill in present value for providing public goods and services at current levels for ever. To maintain the current fiscal treatment of living generations and current public goods and service levels for ever, the present value cost that future generations must pay equals $PVG_t - PVL_t - NW_t$.

Thus, generational accounting reveals the fiscal burden that future generations collectively face under current government fiscal policies. That burden does not necessarily equal the government's outstanding debt: $-NW_t$.

Estimating per capita fiscal burdens facing future-born generations requires knowing how it would be distributed among them. Generational accounting assumes, hypothetically, an equal distribution of the residual fiscal burden except for an adjustment for productivity growth. If we ignore gender differences for simplicity, the GA facing those born in year $t+1$ is calculated as

$$GA_{t+1} = \frac{(PVG_t - PVL_t - NW_t)(1+r)}{\sum_{s=t+1}^{\infty} N_s[(1+g)/(1+r)]^{s-(t+1)}} \quad (2)$$

Here, *r* represents the discount rate; *g* represents labour productivity growth; *s* represents future cohorts' birth years; and N_s represent their population sizes. In eq. (2), the residual fiscal burden in present value as of period $t+1$ is divided by the weighted sum of the population of future-born persons with weights based on *r* and *g*. The discount rate, *r*, is included in the weighting scheme to account for the differences in the timing of net tax payments by different future-born cohorts. Such weighting ensures that people born in period $s > t+1$ pay lifetime net taxes that are $(1+g)^{s-(t+1)}$ times larger than those paid by persons born in period $t+1$.

Generational accounts for the United States

By using projections from the Budget of the US government for fiscal year 2005 (with *t*=fiscal year 2004), applying a five per cent discount rate, and calculating US dollar amounts in constant 2004 dollars, PVG_t is estimated to be \$26.8 trillion; NW_t equals –\$4.4 trillion; and PVL_t equals 4.9 trillion. That leaves future generations to collectively pay \$26.3 trillion.

Table 1 shows GAs for selected US male and female cohorts with *t*=fiscal year 2004. They exhibit a standard life-cycle pattern: older cohorts face negative GAs – they receive benefits on net – and younger ones face positive GAs. Younger women have smaller GAs than men because of their lower labour-force participation and earnings. Very young cohorts with many years to go before paying taxes face

considerably smaller GAs because of discounting. Older women receive larger net benefits in present value than older men despite their lower prior labour-force activity because they live longer and receive social insurance benefits based on their male spouses' earnings. The GA for those born in 2005 (year *t* +1) equals $333,200 per capita – considerably larger than that for 2004-newborns.

Generational accounts for the United States (thousands of constant 2004 dollars)

Year of birth	Age in 2004	Male	Female
2005 (future-born)	–1	333.2	26.0
2004 (newborn)	0	104.3	8.1
1989	15	185.7	42.0
1974	30	201.3	30.2
1959	45	67.8	–54.1
1944	60	–162.6	–189.4
1929	75	–171.1	–184.1
1914	90	–65.0	–69.2

Source: Author's calculations based on data from Gokhale and Smetters (2006).

Lifetime net tax rates and generational balance

Alternatively, fiscal burdens can be represented as *lifetime net tax rates* (*LNTR*) that different generations would face under the given assumptions. For future generations, $LNTRf=GA_s/PVE_s$, for all $s > t$, where PVE_s represents the present value as of period *s* of projected (pre-tax) labour earnings per capita for the cohort born in period *s*. Future labour earnings per capita are projected in a manner similar to that used for projecting taxes and transfers. Equation (2)'s distribution rule implies that both lifetime net taxes and lifetime earnings grow at the same rate for successive cohorts, implying that *LNTRf* applies to all future cohorts.

An important generational accounting concept is that of *generational balance*. It is derived by comparing the lifetime net tax rate facing year-*t* newborns, $LNTR_t = GA_t / PVE_t$, with *LNTRf*. Note that $LNTR_t$ is based on current tax and transfer policies extended throughout the lifetime of year-*t* newborns whereas *LNTRf* is a hypothetical rate imputed for future generations based on an equal growth-adjusted distribution of the residual fiscal burden across future-born cohorts. A finding of $LNTR_t<LNTRf$ would show current policy as being generationally out-of-balance – one that levies a smaller *LNTR* on current newborns than would be required of future ones on average to balance the government's books. Thus, a policy that is generationally out of balance is also unsustainable.

Calculations based on the GAs shown in Table 1 reveal that US fiscal policy is considerably out of generational balance as of fiscal year 2004. The present value of lifetime earnings for males born in 2004 is estimated to be $562,000, making $LNTR_{2004}$ equal to 18.5 per cent. For future-born cohorts, *LNTRf* equals 58.2 per cent. Continuing existing tax and spending laws for living generations would require future generations to bear fiscal burdens that are more than three times larger on average.

If current policy is out of generational balance (that is, if $LNTR_t<LNTRf$), GA machinery can also be used to calculate alternative policy changes that would restore generational balance. This exercise reveals the policy trade-offs involved in moving from a generationally out-of-balance policy to one that is balanced.

A large initial generational imbalance requires a large fiscal adjustment. Restoring generational balance to US fiscal policy via income tax hikes would require average income tax rates to be 39 per cent larger. That is, federal income tax revenues that according to the US Congressional Budget Office (2006) amounted to 8.6 per cent of GDP in 2004 would have to be immediately and permanently increased to 11.9 per cent of GDP. Alternatively, federal discretionary outlays would have to be reduced immediately and permanently by 67 per cent.

Criticisms of generational accounting

Generational accounting has been subject to several criticisms. First, it measures the direct net costs of taxes and transfers but excludes the benefits derived from government public goods and service purchases. If the benefits from some purchases accrue much later, the average GA facing future generations may not accurately reflect their fiscal treatment under current policies. Second, generational accounting does not factor in the costs and benefits from government insurance provision.

These two criticisms indicate that generational accounting is not a ‘utility measure’ of the impact of fiscal policies on different generations. However, dynamic simulation studies suggest that changes in GAs correspond reasonably well to welfare gains and losses arising from policy changes.

Third, generational accounting ignores dynamic economic responses when estimating policy adjustments for restoring generational balance. However, its ‘static’ estimates constitute lower bounds of the required adjustments. For example, increasing income taxes would normally reduce labour supply and require a larger tax hike to achieve generational balance.

Fourth, to qualify as ‘budget concepts’ fiscal measures must show the implications of keeping policies unchanged. However, the generational balance measure employs a hypothetical policy for future generations. Gokhale and Smetters (2003) provide alternative fiscal and generational imbalance measures that do not involve hypothetical policies.

Fifth, generational accounting discounts future fiscal flows using a common discount rate whereas taxes and transfers may be subject to different degrees of policy and economic uncertainties. And sixth, it may be appropriate to use different discount rates for different cohorts because they face different risks. However, generational accounting studies include sensitivity analyses under alternative assumptions, including alternative discount rates.

Final remarks

It is important to note that generational accounting tracks only the redistributive impact of government fiscal policies. It does not include the impact of private bequests and *inter vivos* gifts. In theory, private intergenerational transfers may substantially or fully offset government transfers. However, the weight of evidence, at least for the United States, suggests that such offsets are quite small.

A chief lesson from the generational accounting literature is that the frequently cited aggregate cash-

flow measures of fiscal policy – such as the size of national debt and annual budget deficits – are uninformative and, indeed, may mislead policymakers about the true distributional and economic implications of current fiscal policies and policy changes.

To the extent that traditional deficit and debt measures miss significant policy-induced intergenerational redistributions – with potentially large effects on agents' economic choices such as consumption and labour supply – generational accounting calculations can provide useful information to policymakers and the public.

Generational accounting is also likely to prove useful in further economics and public-policy research. For example, generational accounts could be combined with other elements of wealth – human, non-human and private pension wealth – on a cohort basis to estimate whether changes over time in the cohort-distribution of resources are related to changes in cohort saving and labour force participation. Generational accounts could also be used to calculate changes in the degree of cohort wealth annuitization for examining the extent of insurance against uncertain longevity.

In many countries, government programmes for providing insurance to the public against various types of economic risks are financially unsustainable. Uncertainty about prospective changes in taxes and transfers for correcting those fiscal imbalances constitute a major source of risk for households. Analyses using generational accounting may help in better understanding the extent to which government fiscal policies mitigate or exacerbate the economic risks facing different generations.

See Also

- population agesng
- public debt
- public finance

Bibliography

Auerbach, A., Gokhale, J. and Kotlikoff, L. 1991. Generational accounts: a meaningful alternative to deficit accounting. *Tax Policy and the Economy* 5. Cambridge, MA: MIT Press and the NBER.

Auerbach, A., Gokhale, J. and Kotlikoff, L. 1994. Generational accounting: a meaningful way to evaluate fiscal policy. *Journal of Economic Perspectives* 8(1), 73–94.

Congressional Budget Office. 2006. Historical Budget Data, 26 January, Table 4. Online. Available at http://www.cbo.gov/budget/historical.pdf, accessed 28 June 2006.

Diamond, P. 1996. Generational accounting and generational balance: an assessment. *National Tax Journal* 49, 597–607.

Feldstein, M. 1974. Social Security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82, 905–26.

Gokhale, J., Kotlikoff, L. and Sabelhaus, J. 1996. Understanding the postwar decline in U.S. saving: a cohort analysis. *Brookings Papers on Economic Activity* 1996(1), 315–407.

Gokhale, J. and Smetters, K. 2003. *Fiscal and Generational Imbalances: New Budget Measures for New Budget Priorities*. Washington, DC: American Enterprise Institute Press.

Gokhale, J. and Smetters, K. 2006. Fiscal and generational imbalances: an update. *Tax Policy and the Economy*, vol. 20, ed. J. Poterba. Cambridge, MA: MIT Press and the NBER.

Kotlikoff, L. and Fehr, H. 1996/97. Generational accounting in general equilibrium. *Finanzarchiv* 53, 1–27.

Kotlikoff, L. 1997. Reply to Cutler's and Diamond's views on generational accounting. *National Tax Journal* 50, 303–14.

How to cite this article

Gokhale, Jagadeesh. "generational accounting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000203> doi:10.1057/9780230226203.0627

The New Palgrave Dictionary of Economics Online

Genovesi, Antonio (1712 – 1769)

Angelo Bertolini

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Genovesi, A.; mercantilism; population growth; scarcity; socialists of the chair; value; wealth

Article

Genovesi was born near Salerno and died at Naples: he took holy orders in 1736. In 1741 he taught metaphysics at the University of Naples. He was intimately acquainted with Bartolomeo Intieri, who induced him to follow Broggia and Galiani in the study of economics; and when, in 1754, by the advice of Intieri and with funds' liberally supplied by him, the teaching of economics, then termed mechanics and commerce, was established at Naples, Genovesi was called to the chair. He was ‘the most distinguished and the most moderate of all Italian mercantilists.•…•Commerce was for him not an end only, but also a means by which the products of industry at large were brought to the right market. He, moreover, distinguished between useful commerce which exported manufactured goods and brought back in return raw material, and harmful commerce which exported raw material and imported foreign goods; he also insisted that useful commerce calls rather for liberty than for protection, while upon harmful commerce the strictest embargo should be laid, or at least it should as far as possible be bound hand and foor’ (Cossa, *Introduction to Political Economy*, translation, p. 235).

These ideas, neither new nor original even in his time, were maintained by Genovesi in many of his works, and brought together, but without any systematic order, in his *Lezioni di Commercio ossia di Economia Civile* (Napoli, 1765, e. ii. ediz. 1768–70, 2 vols). Though the *Lezioni* do not form a regular treatise, they contain the author's opinions on the mercantilist system and the most important principles of economics, which he terms *Civile ‘la scienza che abbraccia le regole per rendere la sotto-posta nazione popolata, potente, saggia, polita’* (the science which embraces the laws which make a nation populous, powerful, wise, and cultured), limiting thus the science to the increase of population and the production of wealth.

As to population, Genovesi follows the mistaken principle of his times, exaggerating the advantage of a large population, proposing that government should encourage marriages by granting privileges and honours. He says that the population ought not only to be numerous but supplied with comforts, and he

sees the relation between population and means of subsistence or production of wealth.
As a writer he is a mercantilist, though he does not regard money as the only form of riches; he says that the wealth of a nation is quite apart from the quantity of money treasured up.
He derives the idea of value from demand, distinguishing different degrees of demand according to their abstract importance in several categories, maintaining that a thing which satisfies a want repeatedly has a higher value than what satisfies only a few wants or the same only sometimes (*puo soddisfare ad un bisogno più volte, ha maggior prezzo che non quella, la quale o non puo soddisfare che pochi bisogni o al medesimo qualche volta*). What is able to satisfy a great want is of more value than what satisfies a small want (*una cosa fatta a soddisfare il maggior bisogno si apprezza più che quella la quale non é fatta che a soddisfare ad un minore*); and further he asserts that the quality of things influences the value. Graziani (*Storia della teoria del valore in Italia*, Milano, 1889, p. 108) justly remarks that in this Genovesi approaches the important question which Galiani answered: namely, why do luxuries generally cost more than necessaries? In this he is obliged to have recourse to the element of scarcity, a line of argument which he does not know how to reconcile with those previously mentioned. Genovesi's want of originality is obvious, as F. Ferrara has shown (*Bibl. dell' Econo*.,1^{a}. S. vol. iii. Introduz.) in contradistinction to the exaggerated opinion which Bianchini held respecting him (*La scienza del ben vivere sociale*), since the Socialists of the Chair persist, erroneously, in considering him as a precursor of their opinions. This tendency is also attributed to Genovesi, as well as to Beccaria, Verri, and Romagnosi by the French socialist B. Malon; which is a further example of the errors of the socialists in their historical criticism of political economy.

See Also

- mercantilism

How to cite this article

Bertolini, Angelo. "Genovesi, Antonio (1712–1769)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000027> doi:10.1057/9780230226203.0628

The New Palgrave Dictionary of Economics Online

George, Henry (1839 – 1897)

Mason Gaffney
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advances; barriers to exchange; city planning; Clark, J. B.; common property; excise taxes; Fabianism; George, H.; Hotelling, H.; Irish land question; land markets; land tax; libertarianism; Lorenz curve; marginal cost pricing; market failure; mixed economy; public utility pricing; scarcity; specialization; speculation; supply side economics; taxation of income; urban economics; urban planning; urban transportation economics; Veblen, T.; Vickrey, W. S.; wages fund; wages tax; Wicksteed, P. H.

Article

Henry George was by turns sailor, prospector, printer, reporter, San Francisco newspaper editor and publisher, orator and political activist before closeting himself to write on political economy. His *Progress and Poverty* (1879) electrified reformers, catapulted him to fame and began a worldwide movement for land reform and taxation, opening to George an extraordinary career in radical politics. Returning from Ireland as reporter for *The Irish World of* New York he was lionized by Irish-New Yorkers for his stand on the Irish land question. With ethnic, union and socialist backing he formed the United Labour Party and ran for mayor of New York in 1886, nearly winning.

He toured Britain and won over the Radical-Liberals, and then toured Australia as a folk hero. At home he was courted by Democrat and later by Populist leaders. He died in 1897 while running again for New York mayor, but his followers rose in and helped shape the Progressive movement which dominated the next 20 years. His name has become a byword for ideas and policies he espoused.

George is best known today for *Progress and Poverty* (1879). Eloquent, timely and challenging, it soon became and remains the all-time best-seller on economic theory and policy.

George defines ‘The Problem’ as increase of want with increase of wealth. Dismissing Malthusian fatalism as merely a device to rationalize privilege, George attributes low wages and unemployment rather to artificial scarcity of land and barriers to free exchange. Artificial scarcity results from unequal dispensation of public lands, concentration and ‘speculation’. George's speculation is pervasive market failure endemic to land, which failure he attributes to holding for the unearned increment.

George proposed to raise the *ad valorem* property tax rate on bare land (broadly defined as all natural

opportunities), thus socializing rent without excess burden. He would remove other taxes, calling them barriers to commerce, employment and capital formation. The cash drain of the *ad valorem* tax, while neutral at the margin, would move and lubricate the land market as a whole, forcing land into full use. Observation persuaded him that otherwise speculation overrode the incentives to use land fully. Release of hoarded lands would open wider opportunities for both labour and man-made capital. His overriding concern was for labour, but he saw capital mainly as a form of labour, produced by labour, complementing labour. So in an era before payroll taxes it was actually capital and commerce he sought to untax for the benefit of labour – a preview of ‘business Keynesianism’ in W. Heller's production of Camelot.

George did not see investment employing labour, but labour producing capital, a difference that to him was more than a nuance. While admiring Quesnay he never absorbed the Physiocratic idea of *‘avances’*. Instead he attacked its English derivative, the wages-fund theory with its advances of subsistence, a concept he rejected as condescending to labour. He developed no concept of economic circulation, of either capital or spending. He lacked a good capital theory, belittling Austrian interest theory and botching his own. These faults narrowed the effective scope of his otherwise seminal work and ultimately limited its influence, which is still wide and sustained but mainly outside the macroeconomic field it addressed.

His programme would level barriers to exchange and specialization and production and synergy. These include spatial barriers forced by land speculation (for example, scattered settlement and urban sprawl); fiscal barriers like excise and wage taxes; and social barriers from unequal wealth and contempt for workmanship, which he (like Veblen) traced to the influence of privilege and unearned wealth. This ‘true free trade’ would unleash technological, scientific, cultural and spiritual development in a more egalitarian and moral society organized around a perfected market mechanism.

George drew on earlier thinkers: Quesnay, Smith, Ricardo, Spencer, and Mill. And he contributed much to later thinking.

George was system-minded and sought to unify the laws of production and distribution in a coordinated harmonious system. His theoretical framework is an early adumbration of the marginal productivity theory of wages, which he integrates with Ricardo's rent law. J.B. Clark was a nemesis, and P. Wicksteed a friend, but both were formalizing insights from George.

Although best known as a deductive thinker, the journalist was also an observer with statistical intuition. In debate with Francis Walker on ‘The March of Concentration’ in farming, George anticipated Lorenz's method of analysing size distributions and goaded the US Census into publishing farm size data in that form.

George wanted radical redistribution but without revolution. He pioneered the idea that taxation, properly crafted, can redistribute wealth without damage to the market. His influence on Fabianism was early and wide; also on American reformers like Tom L. Johnson, Upton Sinclair, John R. Commons and Norman Thomas. The modern ‘mixed economy’ is in the Georgist spirit of reform within traditional forms.

Continued heavy reliance on real estate taxation in Canada and the United States, with separate assessment of land value, reflects George's influence, as do the inclusion of land rents and gains in the income tax base, and the efforts of Lloyd George, Asquith and Snowden to introduce national land taxes in Britain.

Free provision of public goods, social dividends, and marginal-cost pricing for urban mass transit and

utilities are vintage Henry George. H. Hotelling and W. Vickrey have acknowledged their debt.
The optimistic ‘economics of abundance’ idea owes much to George. The prevailing ‘dismal’ economics was a science of choice where all the choices were bad and leaders could only call for more sacrifices. George promised full employment at higher wages by unlocking natural opportunities now held in speculation. Needed capital would be formed in the very process of making jobs, an idea pervading Keynes. Social synergy would produce a surplus that spills over into higher land rents, a ‘free lunch’ that government may tap in lieu of taxes that penalize and abort useful activity.
George lives too in urban economics and city planning. George's emphasis on the synergistic gains from urban linkages, and the wastes of sprawl caused by failure of the land market anticipates much of planning doctrine. Ebenezer Howard is an obvious link: his ‘Garden City’ presupposed Georgist taxation to move the land market.
The idea that environment is a common heritage for future generations in pure Georgism. ‘Spaceship Earth’, common property, and rights of the unborn are his very phrases.
As to economic development, the economists are legion who have recommended a ‘dose of Henry George’ to help LDCs take off, and some, like Taiwan, belatedly following the counsel of the Georgist Dr Sun Yat-sen, have taken the dose with good results.
On the conservative side, George was a pioneer of tax limitation, insisting that land rent set an upper limit on government spending. The resurgence of libertarianism and supply-side economics may set a new stage for George, whose programme was mainly oriented to increasing production in the private sector. Religion in politics should not threaten George, who unabashedly presented economic policy as an implementation of religious ideals.
George's blend of radicalism and conservatism can puzzle one until it is seen as a reconciliation of the two. The system is internally consistent but defies conventional stereotypes.

See Also

- land tax

Selected works

Complete Works. Garden City, NY: The Fels Fund; also other publishers.

Bibliography

Andelson, R., ed. *Critics of Henry George*. London and Cranbury, NJ: Associated University Presses.

Barker, C.A. 1955. *Henry George*. New York: Oxford University Press.

Cord, S.B. 1965. *Henry George: Dreamer or Realist?* Philadelphia: University of Pennsylvania Press.

de Mille, A.G. 1950. *Henry George: Citizen of the World*. Chapel Hill: University of North Carolina

Press.

Geiger, G.R. 1933. *The Philosophy of Henry George*. New York: Macmillan Co.

George, H., Jr. 1900. *The Life of Henry George*. Reprinted, New York: Robert Schalkenbach Foundation, 1943.

Lawrence, E.P. 1957. *Henry George in the British Isles*. East Lansing: University of Michigan Press and Vanguard Press.

Post, L.F. 1930. *The Prophet of New York*. New York: Vanguard Press.

Sawyer, R.A. 1926. *Henry George and the Single Tax: A Catalogue of the Collections in the New York Public Library*. New York: New York Public Library.

The Standard (weekly). 1887–92. New York.

How to cite this article

Gaffney, Mason. "George, Henry (1839–1897)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000028> doi:10.1057/9780230226203.0629

The New Palgrave Dictionary of Economics Online

Georgescu-Roegen, Nicholas (1906 – 1994)

Stefano Zamagni
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Georgescu-Roegen, N.; integrability; transitivity; non-substitution theorem; input–output analysis; separating hyperplane theorem; regime switching; bioeconomics; entropy law.

Article

Born in Constanza, Rumania, on 4 February 1906, Georgescu-Roegen obtained his first degree in mathematics in 1926 from the University of Bucharest. He then went to Paris where, under the supervision of E. Borel and G. Darmois, he received in 1930 the doctorate in mathematical statistics. In October of the same year he moved to London to pursue further research with K. Pearson. By 1932 Georgescu was Professor of Statistics at the University of Bucharest. His life was inextricably bound up with the social and political events of his country, which explains the emergence of his interest in economics and his consequent decision to spend a two-year ‘apprenticeship’ (1934–6) at Harvard where he was able to work closely with Schumpeter. In 1937 he returned to Rumania, where he combined an active academic career with increasing responsibilities in public institutions. In February 1948 he fled from his country and, after a short stay at Harvard, was appointed professor at Vanderbilt University, where he remained until his retirement in 1976.

Georgescu-Roegen's scientific work is notable for an early phase centred around consumer theory, input–output analysis and production theory at large, and a later phase mainly devoted to growth modelling, methodological issues and the ambitious attempt to develop a ‘bioeconomic’ approach to economic thinking. The early phase is well represented by his 1936 classic article on consumer theory and his 1954 famous paper on ‘Choice, Expectations and Measurability’. In the former article, which deals with the ‘mysterious’ problem of integrability in the theory of demand, one finds two major results: the demonstration that the integral varieties do not necessarily coincide with the indifference varieties – whence the distinction between mathematical integrability and economic integrability – and the demonstration that the two kinds of varieties come to the same thing in the presence of the postulate of transitivity of preferences. The latter essay, focusing on the non-existence of the indifference map of the consumer as a consequence of the pervasiveness of lexicographic ordering of preferences, allowed him

to prove what he called the ‘ordinalist fallacy’ and to inquire about the origin and implications of probabilistic preferences, a subject that is at the very frontiers of economics even today.

On the other front, three contributions are particularly noteworthy. In Georgescu (1951a) we find the first and the most general statement of the celebrated non-substitution theorem: justifying the separation of scale and composition in linear multisectoral models, the theorem provides a theoretical underpinning and analytic rationale for the consistency of input–output analysis. The (1951b) paper offers the first ‘geometric’ proof of the existence of a von Neumann's equilibrium by using the separating hyperplane theorem – a theorem that was to enter the toolbox of the economist. In his (1951c) essay, Georgescu challenged the two most intractable problems in macrodynamics – nonlinearities and discontinuities – providing, on the basis of an innovative application of the theory of relaxation oscillations, a fundamental result for investigations of regime switching.

The later phase begins with the 1966 famous methodological essay containing Georgescu-Roegen's critique of standard economics for having reduced the economic process to a mechanical analogue and a proposal of a new alliance between economic activity and the natural environment – what later would become his ‘bioeconomic programme’. The key to such a project is found in the entropy law (‘the most economical of physical laws’), which brought Georgescu to inquiry on the fundamental relation between mankind's existence and its environmental dowry. This problem prompted him to step over the fence of economics into thermodynamics, where he formulated a new law (the ‘fourth law’): the impossibility of the perpetual motion of the third kind defined as a *closed* system that could perform work at a constant rate indefinitely. The implications for economics of this line of thinking and in particular of his strong rebuttal of the ‘energetic dogma’ (‘only energy matters’) are nicely developed in his 1971 and 1976 books. In this last book, Georgescu lays the foundations of a new approach to production theory: the ‘flow-fund’ model as a radical alternative to both the production function model and the activity analysis model, models whose main drawback lies in their inability to tackle properly the time element in the productive process.

The long introductory essay (145 pages) that Georgescu wrote in 1983 for the English edition of Gossen's *The Law of Human Relations* is not simply a splendidly written intellectual biography, showing the depth and breadth of his economic culture, but it contains also a restatement in modern analytical terms and an expansion of Gossen's theory of economic behaviour. Georgescu-Roegen was one of those rare scientists able to couple a remarkable expertise in their specific field with a philosophical bent of mind. In this sense he was a true Renaissance man, which perhaps helps to explain the generalized *fin de non recevoir* of the profession with respect to his critical message, the message of a scholar who cannot be identified with any single school of economic thought and whose intellectual endeavour is best seen as a major contribution to the shifting of the frontiers of economic theory and methodology.

Selected works

1936. The pure theory of consumers' behavior. *Quarterly Journal of Economics* 50, 545–93.

1951a. Some properties of a generalized Leontief model. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans et al. New York: Wiley.

1951b. Relaxation phenomena in linear dynamic models. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans et al. New York: Wiley.

1951c. The aggregate line as production function and its applications to Von Neumann's economic model. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans et al. New York: Wiley.

1954. Choice, expectations and measurability. *Quarterly Journal of Economics* 68, 503–34.

1966. Some orientation issues in economics. In N. Georgescu-Roegen, *Analytical Economic Issues and Problems*. Cambridge, MA: Harvard University Press.

1971. *The Entropy Law and the Economic Process*. Cambridge, MA: Harvard University Press.

1976. *Energy and Economic Myths: Institutional and Analytical Economic Essays*. Oxford: Pergamon Press.

1983. Hermann Heinrich Gossen: his life and work in historical perspective. In H.H. Gossen, *The Laws of Human Relation and the Rules of Human Action Derived Therefrom*. Trans. from the original German edn of 1854 by R.C. Blitz. Cambridge, MA: MIT Press.

How to cite this article

Zamagni, Stefano. "Georgescu-Roegen, Nicholas (1906–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000029> doi:10.1057/9780230226203.0630

The New Palgrave Dictionary of Economics Online

G é rard-Varet, Louis-Andr é (1944 – 2001)

Rodolphe Dos Santos Ferreira
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adverse selection; art, economics of; competitive toughness index; endogenous fluctuations; enforcement mechanisms; expected externality mechanism; Farkas lemma; general equilibrium; Gérard-Varet, L.-A.; imperfect competition; industrial organization; involuntary unemployment; Lagrange multipliers; market share; market size; mechanism design; moral hazard; oligopolistic competition; overlapping generations; P-equilibrium; pricing schemes; truth revelation

Article

Born on 30 June 1944 at Auxerre, Gérard-Varet studied economics and sociology at the University of Dijon, and prepared his doctorate partly at CORE (Louvain). He taught, as full professor, in Strasbourg, Toulouse and, for most of his academic career, Marseille, where he died on 31 January 2001. He played a quite important role, not only through his teaching and his scientific production, but also as an organizer, in particular as long-term director of his research centre (GREQAM, Marseille) and president of national and international economic associations.

The first set of the theoretical contributions of Gérard-Varet concerns mechanism design, a field in which he began to collaborate in 1973 with Claude d'Aspremont, in the context of a project on cross-border pollution. Starting from the Vickrey–Clarke–Groves mechanism, which ensures that truth revelation by each agent is a dominant-strategy equilibrium, but not that the budget is balanced, they introduced the *expected externality* (or *AGV*) *mechanism*, which is both truthfully implementable as a Bayesian equilibrium and budget-balanced. The mechanism was explicit, while requiring independence of agents' beliefs. This condition was considerably generalized by switching from a constructive to an existence proof, based on the Farkas lemma (1979; 1990a). Restrictions on beliefs under adverse selection were also shown to be transposable to stochastic outcome functions in team moral hazard, and to be applicable to two kinds of enforcement mechanisms: enforcement through transfer schemes and enforcement through repetition (1998). More generally, the work of the mid-1970s opened the way to a lifelong research programme.

The second set of the theoretical contributions of Gérard-Varet concerns oligopolistic competition (in

partial, general and macroeconomic equilibrium). In 1980, he engaged in the analysis of the macroeconomic effects of significant output market power. He first showed the possibility of so-called involuntary unemployment (in Keynes's sense of persistent unemployment at an arbitrarily low money wage). This results either from the non-existence of a full employment equilibrium (because marginal revenue eventually becomes negative as wages decrease) or, when equilibria are multiple, from a failure to coordinate on that equilibrium (1990b). The early static results were extended to overlapping generation economies and linked to the emerging literature on markup variability as a source of endogenous fluctuations (1995a). The need for a unified treatment of different standard varieties of imperfect competition motivated the formulation of the *P-equilibrium* concept, first applied to an industry or a group of industries (1991), then to the whole economy, in a general equilibrium approach (1997). Strategic agents simultaneously choose price signals in order to manipulate market prices through some *pricing scheme P*, and quantities, required to satisfy market realization constraints. A distinct but related concept of oligopolistic equilibrium was designed later (2007), where producers of elements of a composite good choose price–quantity pairs under two constraints, on market share and on market size. The associated Lagrange multipliers are used to build an index of *competitive toughness*, parameterizing the set of equilibria and appearing as a foundation to the ‘conjectural variations’ parameter of the empirical industrial organization studies.

The preceding themes by no means exhaust the list of subjects on which Gérard-Varet has made theoretical and applied contributions, often combining features of public economics and industrial organization, of which the economics of visual arts is a good example (1995b).

See Also

- mechanism design

Selected works

1979. (With C. d'Aspremont.) Incentives and Incomplete Information. *Journal of Public Economics* 11, 25–45.

1990a. (With C. d'Aspremont and J. Crémer.) Incentives and the existence of Pareto-optimal revelation mechanisms. *Journal of Economic Theory* 51, 233–54.

1990b. (With C. d'Aspremont and R. Dos Santos Ferreira.) On monopolistic competition and involuntary unemployment. *Quarterly Journal of Economics* 105, 895–919.

1991. (With C. d'Aspremont and R. Dos Santos Ferreira.) Pricing schemes and Cournotian equilibria. *American Economic Review* 81, 666–73.

1995a. (With C. d'Aspremont and R. Dos Santos Ferreira.) Market power, coordination failures and endogenous fluctuations. In *The New Macroeconomics: Imperfect Markets and Policy Effectiveness*, ed. H. Dixon and N. Rankin. Cambridge: Cambridge University Press.

1995b. On pricing the priceless: comments on the economics of the visual art market. *European Economic Review* 39, 509–18.

1997. (With C. d'Aspremont and R. Dos Santos Ferreira.) General equilibrium concepts under imperfect competition: a Cournotian approach. *Journal of Economic Theory* 73, 199–230.

1998. (With C. d'Aspremont.) Linear methods to enforce partnerships under uncertainty: an overview. *Games and Economic Behavior* 25, 311–36.

2007. (With C. d'Aspremont and R. Dos Santos Ferreira.) Competition for market share or for market size: oligopolistic equilibria with varying competitive toughness. *International Economic Review* 48(3) (forthcoming).

How to cite this article

Ferreira, Rodolphe Dos Santos. "Gérard-Varet, Louis-André (1944–2001)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000202> doi:10.1057/9780230226203.0631

The New Palgrave Dictionary of Economics Online

German economics in the early 19th century

K. Tribe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

During the 18th century ‘cameralism’ became a recognized part of the introductory curriculum of (mainly northern, Protestant) German universities. The French Revolution, the Napoleonic occupation, and Kant's new Critical Philosophy together swept established doctrine aside during the final decade of the century, allowing a reoccupation of the university curriculum by ‘modern economics’. Jean-Baptiste Say had more direct influence on this than Adam Smith's *Wealth of Nations*. When combined with a post-Critical emphasis on human needs, this directed German writings away from the English emphasis on value and distribution and laid the foundation for a new ‘marginalist’ economics in the 1870s.

Keywords

cameralism; Engels, F.; German economics in the early 19th century; Hermann, F.; Hufeland, G.; Jakob, L.; Marx, K.; Menger, C.; Musgrave, R.; *Nationalökonomie* ; Physiocrats; Rau, K.; Sartorius, G.; Say, J.-B.; Schumpeter, J.; Smith, A.; Steuart, J.; von Soden, J.

Article

Nationalökonomie emerged in early 19th century Germany as a new doctrine which took human needs and their satisfaction to be the first principle of economic analysis.
Cameralism had become in mid-18th century Germany a regular part of the university curriculum in German universities, taught in faculties of philosophy to future state officials. By the 1820s this function had been modified, with lectures on economics becoming a compulsory part of the law curriculum, as elsewhere in Continental Europe. This institutional development coincided with the emergence of a new economics which displaced the older *Cameralwissenschaften*, together with their focus on state finance and national wealth. The older teaching was pushed to one side, and its place in the lecture room taken by the new principles and doctrine of *Nationalökonomie*. This joint development was not a direct outcome of the Revolutionary and Napoleonic Wars (although these certainly caused a great deal of

physical disruption to universities), nor a consequence of the new republican ideas fostered by these wars and by the French Revolution itself (although these certainly played some role). Nor did it follow directly from the diffusion of Adam Smith's teaching, although Smith did in the early 1800s become a major point of reference in discussion of economic principles. This internal transformation in the teaching of economics in German universities resulted instead from an assault upon the older, eudaemonistic natural law tradition by converts to Critical (Kantian) Philosophy. This process was just gaining momentum around 1795; by 1805 the transformation was all but complete. Although textbooks in the cameralistic tradition still appeared, and, it can be assumed, professors continued as always to read out their old lectures, the teachings of Smith and Say now found a definite place within the university, as part of a new *Fach*, *Nationalökonomie*.

This new science drew on a range of sources, but remained quite distinct from the political economy being developed at the time in Britain by James Mill, Robert Malthus and David Ricardo. It shared neither their emphasis upon distribution between agents defined by the process of production, nor the peculiarly English preoccupation with value and its measurement. As we shall see, German writers were, directly or indirectly, influenced by the work of Jean-Baptiste Say rather than by any English political economist. Especially important was Say's tripartite schema of production, distribution and consumption, and his argument that production produced neither ‘things’ nor value, but *utilities*.

A direct line can therefore be traced from this early *Nationalökonomie* to the work of the early Austrian economists. When in 1871 Carl Menger published his *Grundsätze der Volkswirthschaftslehre*, he defined as ‘goods’ ‘utilities …related to the satisfaction of human needs’ (1871, p. 2). A long footnote was appended to this statement, beginning with Aristotle's conception of goods, proceeding on through Forbonnais, Le Trosne and Say; and listing as the first relevant German authors Soden, Jakob and Hufeland. These three writers were the principal architects of the new *Nationalökonomie*. Among them, Jakob's definition of a good is the most pithy: ‘Everything that serves the satisfaction of human needs’ (1805, §. 23). Jakob was also Say's German translator (1807), his own textbook of 1805 clearly following the organization and argument of Say's *Traité d’économie politique*. And, like Schlözer (1805, §. 12), Jakob made a clear distinction between state and economy, or politics and economics.

The definition of National***ö***konomie

‘German economics’ has always been primarily an academic rather than a public or popular discourse, but during the 18th century it had been not uncommon for teachers of cameralism to be shared with the practical world of state administration. Soden was one of the last representatives of this tradition – he was never a university teacher, but spent his working life in state administration before retiring in 1796 to his estate so that he might devote himself to literary pursuits. Prompted by what he saw as the lack of system in Adam Smith's *Wealth of Nations*, a new translation of which had just appeared (1794–6), Soden set out to provide a more systematic basis for the new science, defining it as:

> …the *Natural Law* of *sociable* mankind with respect to the maintenance and promotion of its physical welfare, and in the same way that the Law of Nations outlines the laws according to which *nations*, in the reciprocal condition of co-existence, must adhere in every respect; so *Nazional-Oekonomie* provides the principles which … must be adhered

> to, such that every member of every nation achieves the highest possible degree of physical welfare, and maintains this position. (1805, pp. v, vi)

The leading principle of *Nationalökonomie* was then described as '… the highest perfection of the physical condition of sociable mankind' (1805, p. 14), underscoring the impact of the contemporary intellectual fashion for Critical Philosophy beyond the confines of philosophy and ethics.
Gottlieb Hufeland, the third writer named by Menger, was, like Jakob, a professor of law who had become a 'Kantian' – and both, like Soden, found their way to Adam Smith via Kant's new philosophy. Hufeland's *Neue Grundlegung der Staatswirtschaftskunst* began with a review of the relative merits of James Steuart and Adam Smith as systematic theorists of economic life. Smith's *Wealth of Nations* had quickly been translated into German, but found at first little resonance among university professors, although, like the Physiocrats, Smith was more widely read by lay members of local literary and economic societies than by university scholars. The German 'Smith reception' proper began with the publication of the second translation in the mid-1790s, coinciding with the developing wave of enthusiasm for Critical Philosophy. And as Hufeland noted in his introductory remarks, it was only towards the end of the 1790s that there were clear signs that Smith had been read *and* understood (1807, n.p.).
Soden and Jakob, observed Hufeland, had dubbed their new field of study *Nationalökonomie*, and, while he did not find this very objectionable, he suggested that it would be 'better and clearer' to use a German expression, *Volkswirthschaft* – which is indeed the root term that became generally accepted about a century later. This latter expression was more suitable, he thought, because it expressed a clear distinction with respect to *Staatswirthschaft*, the generic term that cameralists had used to describe the domain of economic life in an intellectual system where no distinction was made between 'state' and 'society'. The problem with the German word *Wirthschaft*, he noted, was that it implied a governing person – invoking the Aristotelian head of household on the one hand, and the more down-to-earth figure of the farmer or inn-keeper on the other (*Landwirt*, or simply *Wirt*). Such a figure was absent from the *Volkswirthschaft* 'where many thousands pursue their economic life' (*wirthschaften*)' (1807, p. 14). This he later clarified as a 'sphere of goods', goods being defined as any medium for the realization of human purposes; hence, the 'sphere of goods' was a domain of autonomous human economic activity independent of state action (1807, pp.17–18, 116).

Jakob and the architecture of National*ö*konomie

Of the writers introduced so far, Ludwig Heinrich Jakob was the most influential in recasting German economics around the conception of human need. He began teaching a course on 'Political Economy and State Economy according to Sartorius' in 1801. Up to this time he had been preoccupied with the creation of a new natural law based on critical principles, exemplified by his *Philosophische Rechtslehre oder Naturrecht* of 1795. We can better see how this new natural law contributed to the reshaping of cameralism if we consider the structure of the book that Jakob first used as a textbook, Sartorius's *Handbuch der Staatswirthschaft zu Gebrauche bey akademischen Vorlesungen* (1796). It was a standard requirement that professors select a textbook to which their lectures were directed, and quite normal for a new course of lectures to be developed as a commentary on this text. It had also become established

practice that each lecturer found the existing texts in some way unsuited to his purposes, and so from this commentary there would develop a new text which became in turn the assigned text. If simple enthusiasm for the writings of Adam Smith had been sufficient to bring about the demise of cameralism and its replacement by *Nationalökonomie*, then we might reasonably expect Sartorius to have played a key role in this, for he was one of the first and most articulate Smithians. But while many late cameralistic writers recognized that Smith's *Wealth of Nations* was an important work, none of them contributed significantly to the new conception of human need, economic activity and welfare that was to survive as the core of German economics for more than a century.

Sartorius's textbook carries the subtitle *Nach Adam Smiths Grundsätzen ausgearbeitet*, for it is principally a condensed version of *Wealth of Nations*, about 40,000 words long, based on lectures that Sartorius had delivered in Göttingen since 1791 as a Privatdozent in the philosophy faculty. This, therefore, gives us an idea of what Sartorius taught, and also what argumentative opportunities this text presented to Jakob in his own initial lectures. Sartorius's presentation is brisk: Smith's ‘Introduction and Plan of the Work’ is dealt with in 16 lines, and by the fifth page we have already arrived at Book I, Ch. V. Books I and II of *Wealth of Nations* are dispatched in 90 pages of summary, each paragraph corresponding to a chapter or a part of a chapter in the original. Discursive sections of *Wealth of Nations* are reduced to bare propositions; importantly, the argument concerning the human propensity to exchange is suppressed entirely. Once the summary reaches the end of Smith's Book II, Sartorius inserts his own section which summarizes the principles of Books III–V as if they were part of a cameralistic treatise on *Staatswirtschaft*: the title runs ‘Of State Economy, or the Rules which the Government of a State must Pursue, so that Individual Citizens might be placed in the Position of being able to Create for Themselves a Sufficient Income, as well as Providing the Same for Public State Expenditures’. Here, although freedom is the means, a eudaemonistic conception of welfare is the objective. This shift towards an older German conception of the state, its tasks and objectives is continued into the treatment of public finances.

For all his admiration of Smith during the 1790s, Sartorius presented a brisk précis of *Wealth of Nations* rather than a summary of, or commentary upon, its leading principles. The characteristic emphasis that we have seen in Jakob, Hufeland, Schlözer and Soden upon human needs, upon goods as means for the satisfaction of such needs, and upon the economy as the domain within which human individuals sought to maximize their satisfaction of need finds no place here. Two years later Say's *Traité* was published, and this would prove a far more promising avenue through which these concerns, springing from Critical Philosophy, could be brought to bear upon economizing activity.

Jakob notes in the Preface to his *Grundsätze* that he had used Sartorius's *Handbuch* for some years, but that he had come to the view that some of Smith's ideas were obscured by the form of presentation adopted. He proceeds to redefine the state and its affairs in a manner that denies it a decisive role in the formation and distribution of wealth, analogously to the manner in which Say had clearly separated politics from political economy. The expression ‘State’, Jakob argues, can be used to refer only to *public* affairs; state property is therefore merely a part of national property (*Volksvermögen*), separate from it and for use in pursuit of public and common ends.

> *Staatswirthschaftslehre* can be in fact nothing other than *financial science* or *Policey*, insofar as care for public order is part of good public economy. (1805, p. vi)

How does this intent translate into the principles of *Nationalkönomie*, and what relationship is established to political economy on the one hand and the cameralistic sciences on the other?
The most striking initial feature of Jakob's book is its plan of organization. At the end of the ‘Introduction’, he states that *Nationalökonomie* deals with three principal issues:

1. 1. The formation and increase of national wealth.
2. 2. The principles of the most advantageous distribution of national wealth among the members of society.
3. 3. The consumption of national property and the various effects of the same. (1805, p. 12 §. 20)

Or, in other words, the trinity of production, distribution and consumption introduced by Say in the second edition of his *Traité* (1814). Say had already stated in his first edition that Smith distinguished politics as the science of legislation from a political economy dealing with the formation, distribution and consumption of wealth (1803, vol. 1, pp. i–ii); but this first edition was divided into five books, dealing in turn with production, money, value, revenue and consumption. Jakob, by contrast, not only stated that *Nationalökonomie* dealt with the production, distribution and consumption of wealth; his book is divided up in this way too. Viewed from this perspective, the sequence of chapters in Jakob's 1805 textbook resembles more closely the order in which material is treated in Say than it is in Smith.

Karl Heinrich Rau and the systematization of National*ö*konomie

Jakob, Soden, Hufeland and Schlözer, more or less simultaneously and independently, created a new conception of economic life separate from the work of state administration, which had hitherto been thought to provide a necessary framework for the orderly conduct of production and consumption. This did, however, remain very general in outline, and in many cases it was simply taught alongside the traditional practical areas of economic administration, such as agriculture and forestry, finance, and botany. In 1826 Karl Heinrich Rau, since 1822 a professor at Heidelberg, published the first volume of a textbook that was to end this equivocal state of affairs. Three volumes of his new *Lehrbuch der politischen Oekonomie* – Rau chose to revert to a name for the subject generally accepted outside Germany and more recognizable to French or English readers – were published between 1822 and 1837. The work ran into many editions, the last revised edition appearing in 1876. Here again there is a clear line of connection to Austrian economics, for Menger drafted the first outline of his *Grundsätze* in the margins of his 1863 edition of Rau (Menger, 1963).
The first volume of the *Lehrbuch* deals with ‘Die Volkswirthschaftslehre’ – ‘those characteristic laws which can be perceived in the economic activities of peoples regardless of the intervention of government’ (1826, p. x). It begins by making a clear distinction between private and public economics. ‘Private economics’ is composed of the rules governing the optimum satisfaction of needs through the acquisition, maintenance, and use of material goods. ‘Public economics’ by contrast deals with the satisfaction of needs by the allocation of material goods on the part of the state – it has a strictly redistributive character, recirculating goods produced in the ‘private economy’ using revenues derived from taxation. Whereas the *Volkswirthschaft* is conceptualized by the individual pursuit of self-interest, there are general aims of the state that the individual cannot attain unaided, and so the role of

government in the economy is to intervene to ensure that these general aims are secured. Rau's account does not have the kind of internal theoretical structure that readers of English political economy might anticipate. Although he announces his intention to develop a theory of economic forces based upon natural laws, the volume merely enumerates economic objects without regard to their mutual relationship in production, distribution and consumption. The second volume, first published in 1828, is devoted to economic welfare. Although this has affinities with the older cameralistic teaching, Rau here consistently distinguishes between the wealth of an individual and the wealth of a people, only the latter being the proper object of economic analysis. The function of the state is strictly limited to the facilitation of individuals' desire to better their conditions, through educational provision and the promotion of commercial enterprise. Again, however, once past the initial principles the work becomes a review of specific measures fostering enterprise or removing hindrances to individual enterprise. Thus, we read under the heading ‘Promotion of Exchange or Encouragement of Trade’ about newspapers, fairs, weights and measures, money, roads, railways, canals, and bridges. Later we can read that savings and insurance are to be promoted, and gambling restricted. The proper employment of state expenditure and the relation of taxation to such employments are dealt with in the third volume, devoted to ‘financial science’. The first edition appeared in 1832, and Rau was revising the text for a sixth edition at the time of his death in 1870. Shortly beforehand he had suggested to his family that the work be taken over by Adolph Wagner, who published in 1872 his own revised version of Rau's treatment of financial science, which in successive editions became the standard textbook on finance up to and beyond the turn of the century.

Schumpeter's *History of Economic Analysis* (1954, p. 503 n. 2) devotes no more than a dozen lines to Karl Heinrich Rau, dismissing his textbook as adequate for teaching but of little further interest. But, as noted above, it was with this textbook that Menger started to draft his *Grundsätze*, while the link to Wagner reaches on to Richard Musgrave's work on public finance (he completed his *Diplom Volkswirt* in Heidelberg in 1933) and thence to Buchanan's conception of public goods. And Rau is also linked to the final writer considered here, Friedrich Benedict Hermann, who graduated from the University in Erlangen in 1823 only one year after Rau left for the chair in Heidelberg.

Towards post-classicism

Hermann's *Staatswirthschaftliche Untersuchungen* of 1832 sketched a clear relation between the supply of and demand for economic goods as formative of market prices. His introductory discussion identifies the level of profit and the relation of profit to wages as the most difficult area of economic analysis, given a rigorous treatment by Ricardo but still in need of much refinement. This immediately establishes an approach far more theoretical in character than was at the time usual for German writers on economics But, while it is true that Hermann's work looks forward to the kind of discussion of price formation that we later find elaborated in Mangoldt (1863), it is clear that Hermann shares the preconceptions of all the writers outlined above. His account of basic principles begins with the definition of a good as anything satisfying human need, an ‘economic good’ being acquired through sacrifice of labour or money; and the book ends with the statement that consumption is destruction of use value, a conception drawn from Jean-Baptiste Say.

The initial discussion of need and its satisfaction reviews their treatment in James Steuart's 1767

Principles and Say's 1828 *Cours*, arguing that use value is the main feature of a good because of its capacity to satisfy needs. This does not, however, prevent Hermann from developing an analysis of price formation in which the price level for a particular good is made dependent upon the relation of demand and supply or, what is much the same thing, the relation between the number of sellers and the number of buyers, which echoes Jakob's account of prices and effective demand, with the addition of the term ‘equilibrium’ to describe the point where ‘…goods are demanded and supplied in the same quantities’ (1832, p. 67). Given a basic cost which includes the usual rate of interest and entrepreneurial profit, he suggests that if the price falls below cost then capital and talent will move elsewhere; conversely, where the price prevails above cost then new entrepreneurs will be attracted, in turn leading to a steady reduction in the price until once more prices and costs are equalized (1832, pp. 4–5, 67–81). But even elementary arguments with this degree of clarity are rare in the literature of the period.

Conclusion

This account of German economics in the early 19th century has emphasized the way in which economic discourse had long been a part of university teaching. Left out of the above account are those lacking a university background and whose work thus falls outside this tradition, but who have since become noted as important parts of the early 19th century German context. First among these is Adam Müller, whose Dresden lectures of 1808–9 countered the new idea of society as a self-organizing system with a romantic, organic conception of man and the state (1922). This found no general resonance in contemporary economic writing, and was rediscovered later by early 20th century cultural critics of capitalism. The principal contemporary influence attributed to Müller connects him to Friedrich List, but there is little evidence for this supposition. List for his part did briefly teach at Tübingen during 1818 and 1819, but in ‘administrative practice’, not ‘economics’ in even the widest contemporary sense. The arguments that he developed during the 1830s and early 1840s, and for which he has been remembered, developed economic ideas he had discovered in American writing during the 1820s, and have no direct relation to German economics in this period. The reputation of Karl Marx, who as a law student during the late 1830s in Berlin would presumably have been exposed to some lectures in economics, likewise owes nothing to contemporary German economic writings, since his own interest in political economy was first stimulated by Friedrich Engels's enthusiasm for (English) Owenite ideas, was later developed as a critique of the classical economics of Mill and Ricardo, and betrays little knowledge of contemporary German writing in politics and economics.

See Also

- cameralism
- Musgrave, Richard Abel

Bibliography

Hermann, F. 1832. *Staatswirthschaftliche Untersuchungen*. Munich: Anton Weber.

Hufeland, G. 1807. *Neue Grundlegung der Staatswirthschaftskunst*, vol. 1 Gießen: Tasche und Müller.

Jakob, L. 1805. *Grundsätze der National-Oekonomie oder National-Wirthschaftslehre*. Halle: Ruffsche Verlagshandlung.

Mangoldt, H. von. 1863. *Grundriß der Volkswirthschaftslehre*. Stuttgart: J. Engelhorn.

Menger, C. 1871. *Grundsätze der Volkswirthschaftslehre*. Gesammelte Werke, Bd.1. Tübingen: J.C.B. Mohr (Paul Siebeck), 1968.

Menger, C. 1963. *Carl Mengers erster Entwurf zu seinem Hauptwerk 'Grundsätze'*. Tokyo: Library of Hitotsubashi University.

Müller, A. 1922. *Die Elemente der Staatskunst*, 2 vols. Jena: Gustav Fischer.

Rau, K. H. 1826–37. *Lehrbuch der politischen Oekonomie*, 3 vols. Heidelberg: C. F. Winter.

Rau, K. H. 1872. *Lehrbuch der politischen Oekonomie*, Bd. III 1, 2, 6th edn, ed. A. Wagner. Leipzig: C. F. Winter.

Sartorius, G. 1796. *Handbuch der Staatswirthschaft zu Gebrauche bey akademischen Vorlesungen, nach Adam Smith's Grundsätzen ausgearbeitet*. Berlin: J. F. Unger.

Say, J.-B. 1803. *Traité d'économie politique*, 2 vols. Paris: Deterville.

Say, J.-B. 1807. *Abhandlung über die Nationalökonomie*, 2 vols, trans. L. Jakob. Halle: Ruffsche Buchhandlung.

Say, J.-B. 1814. *Traité d'économie politique*, 2 vols, 2nd edn. Paris: A.-A. Renouard.

Schlözer, C. von. 1805. *Anfangsgründe der Staatswirthschaft oder die Lehre von dem Nationalreichthume*, vol. 1. Riga: C. J. G. Hartmann.

Schumpeter, J. 1954. *A History of Economic Analysis*. London: George Allen and Unwin.

Smith, A. 1794–6. *Untersuchung über die Natur und Ursachen des Nationalreichthums*, 4 vols., trans. C. Garve and A. Dörrien. Breslau: Wilhelm Gottlieb Korn.

Soden, J. von. 1805. *Die Nazional-Oekonomie*, vol. 1. Leipzig: Johann Ambrosius Barth.

Streissler, E. 1994. German predecessors of the Austrian School. In *The Elgar Companion to Austrian*

Economics, ed. P. Boettke. Aldershot: Edward Elgar.

Tribe, K. 1988. *Governing Economy: The Reformation of German Economic Discourse 1750–1840*. Cambridge: Cambridge University Press, chs. 8 and 9.

Tribe, K. 1998. Natural law and the origins of *Nationalökonomie*: L. H. von Jakob. In *The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850*, J. Heilbron L. Magnusson and B. Wittrock. Dordrecht: Kluwer.

Tribe, K. 2002. The German reception of Adam Smith. In *A Critical Bibliography of Adam Smith*, ed. K. Tribe. London: Pickering and Chatto.

Tribe, K. 2003. Continental political economy from the Physiocrats to the marginal revolution. In *The Cambridge History of Science*, vol. 7, ed. T. Porter and D. Ross. Cambridge: Cambridge University Press.

How to cite this article

Tribe, K. "German economics in the early 19th century." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000194> doi:10.1057/9780230226203.0632

The New Palgrave Dictionary of Economics Online

German hyperinflation

Theo Balderston

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The erratic development of inflation in Germany during the First World War into the hyperinflation of 1922–3 has served as a major test-bed of monetary theory ever since. This article charts contemporary and modern explanations of its genesis, stabilization and effects. Modern analysis focuses on the interaction between fiscal and accommodating monetary policy and the expectations of financial asset holders; it disagrees, as did contemporaries, over the degree of agency of the government in determining its own deficit. After an optimistic ‘Keynesian’ assessment of its effects in growth, more recent scholarship has relapsed into pessimism as to its effects on investment.

Keywords

bubble; German hyperinflation; Germany, economics in the 20th century; gold standard; hyperinflation; inflation expectations; labour hoarding; money supply; public debt; quantity theory of money; rational expectations; seigniorage; velocity of circulation

Article

German hyperinflation after the First World War originated in the decision of July/August 1914 to suspend the gold convertibility of the mark and associated gold-reserve requirements. As with other hyperinflations, this one was irregular. German wholesale prices slightly more than doubled during the First World War. By February 1920 the ratio to 1913 prices was about 17, but then fell, irregularly, to a ratio of 13 in May 1921. After May 1921 inflation resumed and between then and June 1922 average monthly inflation was 13.5 per cent; in the following 12 months it reached 60 per cent (including a short cessation in early 1923 as the Reichsbank temporarily pegged the exchange rate), and 32,700 per cent or about 20 per cent per day between June and November 1923. The mark was stabilized in later November 1923 at one million millionth of its 1913 dollar exchange rate. Although only the period from June 1922 was ‘hyperinflationary’ (above 50 per cent per month), this period cannot be studied independently of

the preceding inflationary history (Holtfrerich, 1986).

Contemporary explanation was highly politicized (Kindleberger, 1984a). The ‘quantity theory’ was adopted, especially by the French, to prove the agency of the German authorities in causing the inflation, allegedly in order to undermine the reparations regime. The official German counter-explanation was a variant of the ‘quantity theory’ known as the ‘balance of payments’ theory, whereby a budget deficit and its monetization followed inexorably from the exchange-rate collapse, which they blamed on the Treaty of Versailles and its reparations demands (see Williams, 1922). The quantity theory presumed a constant velocity of circulation, which was at variance with the facts (Graham, 1930; Bresciani-Turroni, 1931); an intellectually satisfying resolution of this puzzle awaited Cagan's (1956) embodiment of ‘expected inflation’ as an argument in the demand-for-money function. The rational expectations' revolution, however, argued that Cagan's formulation of price expectations as a weighted average of past inflation was rational only if the money supply were endogenously determined (Sargent and Wallace, 1973).

The question whether German hyperinflation was a ‘bubble’ divorced from monetary ‘fundamentals’ continues to be discussed, but the evidence remains inconclusive (for example, Chan, Lee and Woo, 2003). The centrality of fiscal policy and seigniorage to the generation of the German hyperinflation is generally agreed. It is the starting point of Webb's (1989) analysis. The Reichsbank, considering Germany still effectively in a state of war, subordinated its monetary policy to the financing of the Reich's expenditure. Though scarcely stable, a real deficit persisted throughout the inflation, albeit with some tendency to decline as inflation accelerated. The private sector's real investment in debt diminished as its belief weakened in the sufficiency of future budget surpluses to meet the state's contractual debt-servicing obligations (including reparations). The private sector inferred from this insufficiency that prices would rise to reduce the real value of this debt-servicing, and converted its non-monetary debt into money and money into goods. This forced greater monetization of the budget deficit; and the conjuncture of the declining real demand for money with rising nominal supply made the public expectation of inflation self-realizing. ‘Unpleasant monetarist arithmetic’ would probably have produced an analogous result even with Reichsbank independence (Holtferich, 1986, pp. 172 ff.).

Frenkel (1977) sought direct evidence of inflationary expectations from the forward discount on the mark in the London foreign exchange market; but, awkwardly from an analytical point of view, until July 1922 the mark sold at a forward *premium*. Webb argued that this reflected the animal spirits of – mainly foreign – speculators with their diversified portfolios, rather than inflationary expectations; these he inferred from the rate of shrinkage of the real value of government debt. On this basis he could link the major shifts in the rate of inflation with announcements of fiscal ‘news’ that prompted state debt-holders into revising their previous estimates of future real budget surpluses. Plausible connexions of this sort can be made for November 1918 (the Armistice), May 1919 (publication of the Treaty of Versailles), May 1921 (announcement of the Allies' London Reparations Plan) and June 1922 (refusal of a bankers' committee headed by J.P. Morgan Jr. to recommend a loan to Germany except on the – at that point unlikely – condition of a reduction in Allied reparations claims).

Webb explained the sudden cessation of inflation in March 1920 by a conjectural calculation that the expected revenues from the new federal direct taxation introduced by Finance Minister M. Erzberger in 1919 now harmonized with debt obligations (though the reparations obligation was still undefined). He explained the stabilization in November 1923 with reference to the cessation, in late September, of state-subsidized ‘passive resistance’ against the Franco-Belgian occupation of the Ruhr; to the imposition of

indexed tax liabilities from October (see Franco, 1990); to the appointment of the Dawes Committee to propose a temporary rescheduling of reparations; possibly to awareness that the Reichsbank was at last threatening to use the independence granted it in May 1922 to cease monetizing the deficit from the end of 1923; and to the successful pegging of the exchange rate against the dollar in mid-November. These developments have to be assumed to have influenced the minds of state debt-holders more than the evidence of the disintegration of the Reich, the collapse of the majority coalition on 3 November, and the lack of clarity, in the hour of France's triumph, over what level of reparations' revision would actually be agreed. Perhaps, after the trauma of hyperinflation, the ‘credibility bar’ over which stabilization policy had to jump was much lowered (Horsman, 1988, p. 33).

The ‘Structural School’ (Kindleberger, 1984b; see Alesina and Drazen, 1991) argues that domestic social conflict, especially on the labour market and partly operating through non-budgetary channels, was central to the hyperinflation. Burdekin and Burkitt (1996) focus on the hugely increased discounting of private-sector bills at the Reichsbank from mid-1922, in order (in their view) to pre-finance inflationary wage settlements. Prior to this, foreign speculation in the mark had financed bank lending to business at negative real rates of interest, so that domestic distributional conflicts could be assuaged out of the wealth of foreigners (Holtfrerich, 1986, pp. 279 ff.). However, once the forward exchange rate flipped over to discount in July 1922, in the absence of Reichsbank accommodation business would have had to pay positive real interest rates, with a correspondingly deflationary effect.

Webb (1989, p. 42) denied that inflation was deliberate government policy. The only reason that the stabilization after March 1920 did not ‘stick’ was that the Allies' ‘London Plan’ of May 1921 derailed it; without this element, the Erzberger fiscal reforms were propelling the budget towards surplus. It was irrational to operate in a hyperinflationary zone when, according to the theoretical consensus, real seigniorage revenues would have been greater at a lower rate of inflation. Webb also accepted the ‘structural’ case that parliamentary conditions and civil-service wage pressure prevented further fiscal reform before autumn 1923 (see Kunz, 1986). Cukierman (1988), however, argued for government agency in the inflationary process on the grounds that, due to increasing lags of inflationary expectations behind actual inflation, it could temporarily increase its seigniorage by increasing inflation, even if at the expense of lower seigniorage in the longer run. The foreshortened time preference of the Reich during its acute diplomatic crisis with the Allies made this rational. Only when expected inflation entered the zone where seigniorage revenues declined – partly due to substitution of other currencies (Bernholz, 1995) – did the government stabilize. Cukierman combines this with an argument that the government and the electorate in any case preferred lower long-run seigniorage revenues as these curbed the reparations rapacity of the Allies.

Holtfrerich (1986, pp. 203–05) argued that the inflation counterfactually raised output by neutralizing the effects of the global post-war slump, and equalized income and wealth (but see Kindleberger, 1994). However, the ultra-low unemployment of the period was also partly due to vast labour hoarding by public enterprises, dating from the demobilization, and to a trough in participation rates. Bresciani-Turroni (1931, pp. 197–203, 403) argued that the inflation caused misallocation of investment; but Holtfrerich argued (1986, pp. 205–06) that not this misallocation but the deflationary gold-standard regime from 1924 caused the low-capacity utilization of the later 1920s. However, Lindenlaub's (1985) archival investigation concluded that, except for industries receiving government compensation for

treaty losses, real fixed investment was minimal (see Fischer, Sahay and Végh, 2002).
Sargent (1986, pp. 40 ff.) argued that the credibility of the German stabilization made it virtually costless. But Dornbusch (1987) regarded the willingness to make monetary policy hurt from November 1923 to June 1924 as necessary to establishing credibility. The ‘stabilization boom’ of the second half of 1924 and the delayed but sharp year-long recession from June 1925 may roughly replicate recent high-inflationary experience (Fischer, Sahay and Végh, 2002).

See Also

- Germany, economics in (20th century)
- hyperinflation
- inflation expectations
- quantity theory of money
- rational expectations

Bibliography

Alesina, A. and Drazen, A. 1991. Why are stabilizations delayed? *American Economic Review* 81, 1170–88.

Bernholz, P. 1995. Currency competition, inflation, Gresham's law and exchange rate. In *Great Inflations of the 20th Century. Theories, Polices, Evidence*, ed. P. Siklos. Aldershot: Edward Elgar.

Bresciani-Turroni, C. 1931. *The Economics of Inflation: A Study of Currency Depreciation in Post-War Germany*. London: Allen & Unwin, 1937.

Burdekin, R.C.K. and Burkitt, P. 1996. *Distributional Conflict and Inflation: Theoretical and Historical Perspectives*. Basingstoke: Macmillan.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Chan, H.L., Lee, S.K. and Woo, K.-Y. 2003. An empirical investigation of price and exchange-rate bubbles during the interwar European hyperinflations. *International Review of Economics* 12, 327–44.

Cukierman, A. 1988. Rapid inflation: deliberate policy or miscalculation? *Carnegie-Rochester Series on Public Policy* 29, 11–76.

Dornbusch, R. 1987. Lessons from the German inflation experience of the 1920s. In *Macroeconomics and Finance. Essays in Honour of Franco Modigliani*, ed. R. Dornbusch, S. Fischer and J. Bossons.

Cambridge, MA: MIT Press.

Fischer, S., Sahay, R. and Végh, C.A. 2002. Modern hyper- and high inflations. *Journal of Economic Literature* 40, 837–80.

Franco, G.H.B. 1990. Fiscal reforms and stabilisation: four hyperinflation cases examined. *Economic Journal* 100, 176–87.

Frenkel, J.A. 1977. The forward exchange rate, expectations, and the demand for money: the German hyperinflation. *American Economic Review* 67, 653–70.

Graham, F.D. 1930. *Exchange, Prices, and Production in Hyper-Inflation: Germany 1920–23*. Princeton: Princeton University Press.

Holtfrerich, C.-L. 1986. *The German Inflation 1924–23: Causes and Effects in International Perspective*. Berlin: De Gruyter. German original, 1980.

Horsman, G. 1988. *Inflation in the Twentieth Century: Evidence from Europe and North America*. Hemel Hempstead: Harvester Wheatsheaf.

Kindleberger, C.P. 1984a. *A Financial History of Western Europe*. London: Allen & Unwin.

Kindleberger, C.P. 1984b. A structural view of the German inflation. In *The Experience of Inflation*, ed. G.D. Feldman, C.-L. Holtfreich and G.A. Ritter. Berlin: De Gruyter.

Kindleberger, C.P. 1994. Review: the great disorder: a review of the book of that title by Gerald D. Feldman. *Journal of Economic Literature* 32, 1216–25.

Kunz, A. 1986. *Civil Servants and the Politics of Inflation in Germany 1914–1924*. Berlin: De Gruyter.

Lindenlaub, D. 1985. *Maschinenbauunternehmen in der deutschen Inflation*. Berlin: De Gruyter.

Sargent, T.J. 1986. *Rational Expectations and Inflation*. New York: Harper & Row.

Sargent, T.J. and Wallace, N. 1973. Rational expectations and the dynamics of hyperinflation. *International Economic Review* 14, 328–50.

Webb, S.B. 1989. *Hyperinflation and Stabilization in Weimar Germany*. New York: Oxford University Press.

Williams, J.H. 1922. German foreign trade and reparations payments. *Quarterly Journal of Economics*

36, 482–503.

How to cite this article

Balderston, Theo. "German hyperinflation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000211> doi:10.1057/9780230226203.0633

The New Palgrave Dictionary of Economics Online

German reunification, economics of

Jennifer Hunt
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

German reunification in 1990 posed the challenge of introducing markets to an economy with none. For citizens of the formerly Communist East Germany, the transition brought an immediate increase in political freedom and living standards, yet also a deep trough in output and persistent unemployment. I examine the reasons for the output trough and the subsequent labour market difficulties, analyse the impact of reunification on West Germany and Europe, and draw lessons for transition and economics generally.

Keywords

‘big bang’ reform; active labour-market programmes: in Germany; Communism; Czech Republic; European Monetary Union (EMU); Exchange Rate Mechanism (ERM) (EU); German monetary union; German reunification, economics of; investment subsidies: in Germany; labour productivity: in Germany; migration and labour markets; privatization: in Germany; trade unions: in Germany; transition; unemployment insurance: in Germany; unemployment: in Germany; wage floor theory; welfare state: in Germany

Article

On 3 October 1990 the formerly Communist German Democratic Republic joined the Federal Republic of Germany, thereby reunifying Germany and posing the challenge of introducing markets to an economy with none. German reunification was part of the dramatic demise of Communism in Europe, an event as significant for economic as for political reasons. For citizens of the former German Democratic Republic (henceforth East Germany), the transition brought an immediate increase in both political freedom and living standards, yet also a large rise in economic uncertainty, manifested not least through the sudden emergence of high unemployment. Although markets and institutions were successfully introduced, they have not led to the rapid economic convergence of the two parts of Germany for which some had hoped, and unemployment has remained high. The enormous costs of reunification have proved a burden for West Germany, which prior to unification had been the economic engine of Europe. The shock of unification and the subsequent slow growth in West Germany have in turn affected the rest of Europe.

Historical and contemporary factors ought to have ensured the best outcomes of any transition economy. Before the Second World War, East German GDP per capita was slightly above the German average (Sinn and Sinn, 1992), and both at that time and under Communism East Germany was richer than (other) eastern European countries. East Germany's relatively small population – 20 per cent of unified Germany – made feasible the large financial transfers from its rich cousin, West Germany. East Germany has benefited from West German institutions, know-how and investment. Yet the Czech Republic had a GDP per capita only 13 per cent lower than that of East Germany in 2004 (OECD, 2005), and, if post-1999 trends continue, the Czech Republic will converge with West Germany before East Germany does.

In this article, I note the successful introduction to East Germany of markets, institutions, democracy and rule of law, and assess why the short-term cost in terms of output and employment was so high. I examine the reasons for the subsequent labour market difficulties, analyse the impact of reunification on West Germany and Europe, and draw lessons for transition and economics generally.

Chronology of unification

The process culminating in the unification of Germany was set in motion when the Hungarian government began allowing East German citizens to leave Hungary via Austria in May 1989. This occurred against the backdrop of reforms in the Soviet Union by President Michael Gorbachev. By August, large numbers of East Germans were reaching West Germany via Hungary, Czechoslovakia and Poland, and in September anti-government demonstrations began in East German cities. On the night of 9 November 1989, a combination of government weakness and confusion led to a crowd being permitted to breach the wall dividing Berlin. The ensuing mass migration to the West removed the power of the East German government to threaten its citizens: five per cent of the eastern population emigrated in 1989–1990.

The East German government organized free elections for March 1990. The victory of the counterpart of the western Christian Democrat Party was seen as a mandate for rapid reunification. Monetary, economic and social union occurred on 1 July 1990. Political union followed on 3 October 1990. As East Germany was formally joining the Federal Republic of Germany, all western institutions were transferred, and only a small number were subject to a transition period. The western systems of justice, regulation, industrial relations, banking, education and social security and welfare were all transplanted, to a large degree by experts from the west.

Faced with the task of integrating a region with decrepit infrastructure, outdated technology and no capitalist experience, the West German government confronted a number of important decisions in 1990. These included: the exchange rate at which to effect monetary union; how to privatize eastern firms; how to spend money in the east, especially how to spend on consumption versus investment (and infrastructure) and on capital versus labour, and the amounts and details of these expenditures; and whether to raise the money through taxes or debt. Important early decisions by other actors included the decision of labour unions to follow a high-wage strategy.

The financial implications of the government's decisions were colossal. From 1991 to 2003 the west spent four to five per cent of its GDP yearly on the east, including transfers within the social welfare system (Ragnitz, 2000, and updated numbers provided by Ragnitz to the author). This spending represented more than 50 per cent of eastern GDP in 1991, and stabilized at about 33 per cent in 1995.

Economic progress of East Germany

Table 1 documents the evolution of various indicators in east and west. Reunification precipitated a disastrous collapse in real eastern GDP, with falls of 15.6 per cent in 1990 and 22.7 per cent in 1991, cumulating to a one-third decline. Meanwhile, West Germany experienced two boom years with growth rates of over five per cent. From 1992, East Germany experienced four years of recovery followed by stagnation. Growth in the west has also been lacklustre since 1992.

Percent change in real GDP, productivity, capital and population, 1990–2004

Year	GDP		Productivity		Capital stock		Population	
	East	West	East	West	East	West	East	West
1990	**–15.6**	**5.7**	–	–	–	–	**–2.5**	**1.6**
1991	**–22.7**	**5.1**	–	–	–	–	**–1.5**	**1.2**
1992	6.2	1.7	18.3	0.7	6.3	2.9	**–0.7**	**1.2**
1993	8.7	–2.6	11.0	–1.4	7.1	2.5	**–0.6**	**0.7**
1994	8.1	1.4	6.3	2.0	7.4	2.1	**–0.4**	**0.4**
1995	3.5	1.4	2.1	1.5	7.4	2.0	**–0.4**	**0.5**
1996	1.6	0.6	2.6	0.7	6.8	1.8	**–0.3**	**0.4**
1997	0.5	1.5	1.9	1.4	6.5	1.7	**–0.4**	**0.2**
1998	0.2	2.3	0.1	0.9	5.8	1.7	**–0.5**	**0.1**
1999	1.8	2.1	1.4	0.7	4.9	1.9	**–0.5**	**0.3**
2000	1.2	3.1	1.7	0.8	4.3	2.0	**–0.6**	**0.3**
2001	–0.6	1.1	0.6	0.3	3.7	1.9	–0.6	0.4
2002	0.2	0.1	1.8	0.4	2.6	1.7	–0.7	0.4
2003	–0.3	–0.1	0.9	0.9	–	–	–0.6	0.2
2004	1.3	1.6	1.0	1.2	–	–	–0.6	0.1

Notes: Berlin is included in the eastern statistics, except for the figures in boldface, where east and west Berlin are included in eastern and western statistics respectively. West Berlin is about 13% of the population of the ‘greater east’. Productivity is measured as GDP per worker. The change in the eastern population in 1989 was –2.5%.

Sources: GDP, productivity, capital stock, population from 2001: Statistische Ämter des Bundes und der Länder. GDP growth in 1990, 1991: Burda and Hunt (2001). Population until 2000: Statistisches Bundesamt.

Labour productivity growth in the east was very rapid through 1994, but has since been modest, although higher than in the west. The eastern capital stock, on the other hand, grew at almost six per cent per year or more through 1998, and has continued to grow faster than the western stock since then. Emigration and a plunge in fertility (a 54 per cent fall between 1988 and the 1994 trough) have caused the eastern population to decline each year since unification. Meanwhile, the western population grew quickly in 1990–2 with the arrival of East Germans and immigrants from ex-Communist countries other than East Germany.

Table 2 represents key indicators as the ratio of east to west. Eastern GDP per capita improved from 49 per cent of the western level in 1991 to 66 per cent in 1995, since when convergence has stalled. Because many of the transfers from the west have been to consumption, disposable income per capita has reached a considerably higher plateau, at 81–3 per cent. Capital per worker has continued to converge gradually where other measures have stalled, reaching 84 per cent of the western level in 2002. Compensation per worker rose rapidly from 34 per cent in 1990 to 56 per cent in 1991 and 68 per cent in 1992, and then stabilized at 79 per cent in 1995.

Measures of convergence – East–West ratios, 1990–2004

Year	GDP			Disposable income	Capital	Compensation	
	per capita	per worker	per hour	per capita	per worker	per worker	per hour
1990	–	–	–	–	–	**34**	–
1991	49	51	–	63	47	56	–
1992	53	60	–	67	54	68	–
1993	60	68	–	74	57	74	–
1994	64	70	–	77	59	77	–
1995	66	71	–	81	61	79	–
1996	67	72	–	83	64	80	–
1997	67	73	–	83	68	80	–
1998	66	72	67	82	72	81	74
1999	66	72	69	83	75	81	74
2000	66	73	68	82	79	81	75
2001	65	73	68	81	82	81	76
2002	66	74	71	82	84	82	76
2003	67	74	71	82	–	82	73
2004	67	74	–	–	–	82	–

Notes: East as a percent of west. Berlin is included in the eastern statistics, except for the figures in boldface, where east and west Berlin are included in eastern and western statistics respectively. 1990 figures are for the first quarter, seasonally adjusted. Productivity is measured as GDP per worker or GDP per hour worked.

Sources: Statistische Ämter des Bundes und der Länder; author's calculations. For 1990, German Institute for Economic Research, Berlin; data on GDP, employment and compensation in East Germany (without West Berlin) from 1989 to 1998 no longer available on the Institute's website.

Reunification might be considered a success in terms of standard of living were it not for problems in the labour market. The left panel of Figure 1 shows the share of the labour force registered as unemployed soared to 20 per cent (from officially zero at the start of 1990), while the western rate has also ratcheted up to a higher level than in 1990. The lack of a search requirement for registering as unemployed means these rates are overstated by several percentage points. The eastern rate is nevertheless very high, especially as some of the many active labour market programme participants would have been unemployed had they not been in the programme. The German Socio-Economic Panel data for the mid-1990s indicate that 15 per cent of the eastern female population and ten per cent of the male population were unemployed (searching and available). The right panel of Figure 1 shows the plunge in the eastern employment rate.

Figure 1

Unemployment and employment rates, 1990–2004

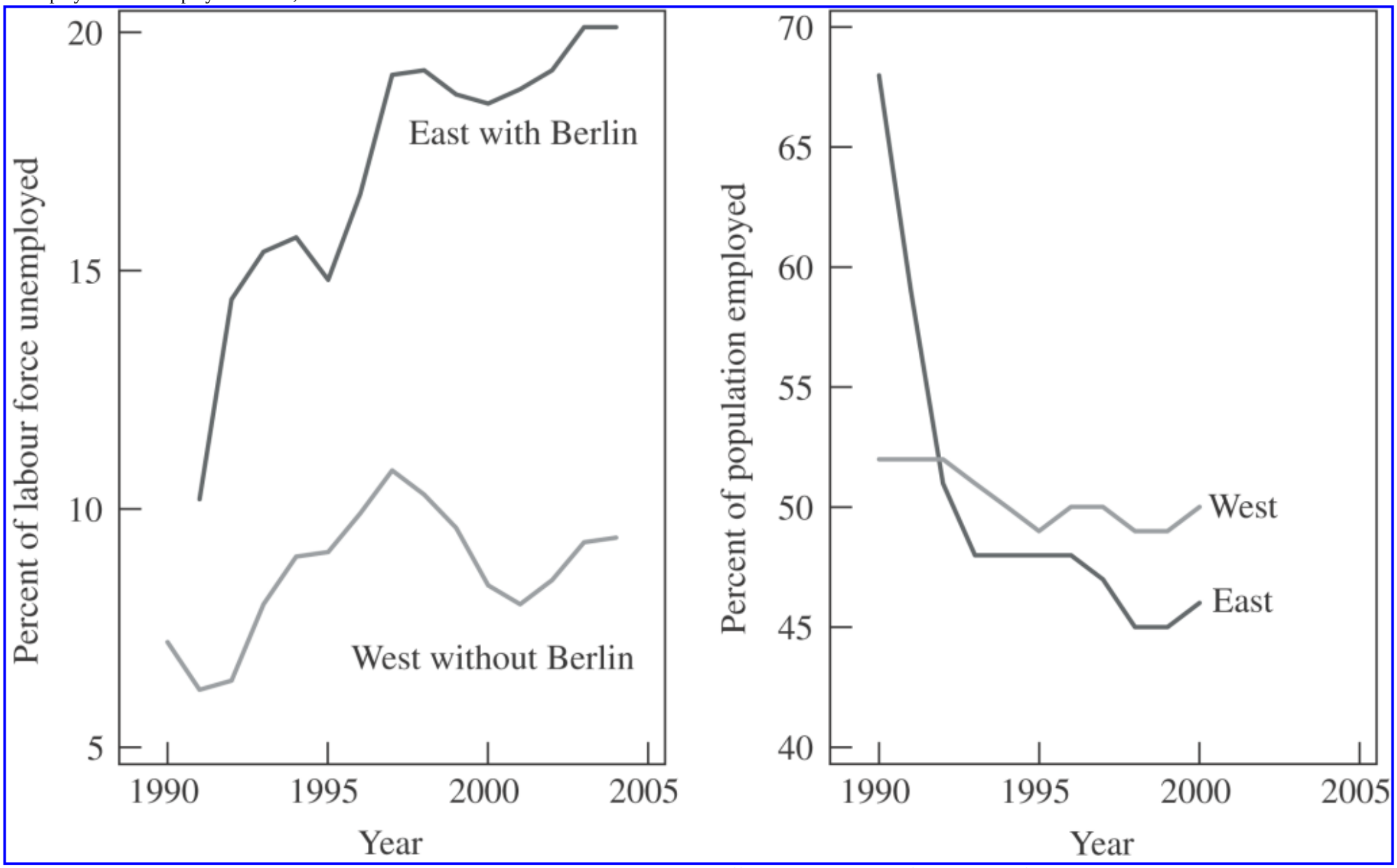

Splitting the east into its constituent states changes the picture little. The unemployment rates differ little across the six federal states of East Germany. Furthermore, with the exception of unified Berlin, the differences in GDP per capita across the six eastern states are small compared with the east–west gap. This may be seen in Figure 2, which plots real GDP per capita for Lower Saxony (Niedersachsen), the poorest western state in 2004, Saxony (Sachsen), the richest eastern state in 2004, and Mecklenburg–Vorpommern, the poorest eastern state in 2004.

Figure 2

Real GDP per capita for selected states

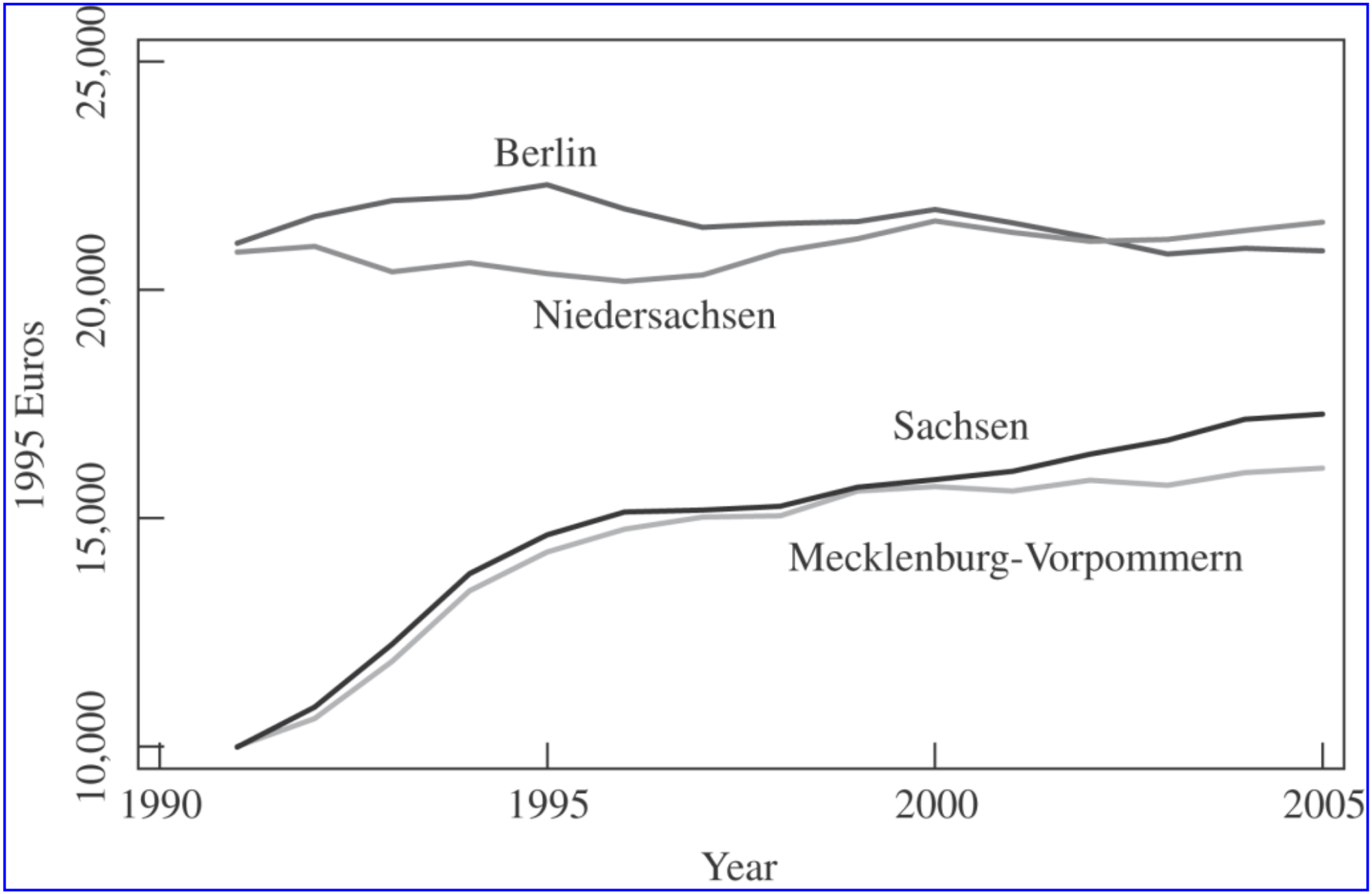

Even the fastest-growing states of Sachsen and Sachsen–Anhalt are growing considerably more slowly than the Czech Republic, as may be seen with the aid of Figure 3. While East German employment languishes at 60 per cent of its 1989 level, and real GDP has barely risen above its 1989 level, Czech GDP is 20 per cent above its 1990 level, and, while Czech employment has not recovered from liberalization, it fell much less than East German employment.

Figure 3

Czech and East German comparisons

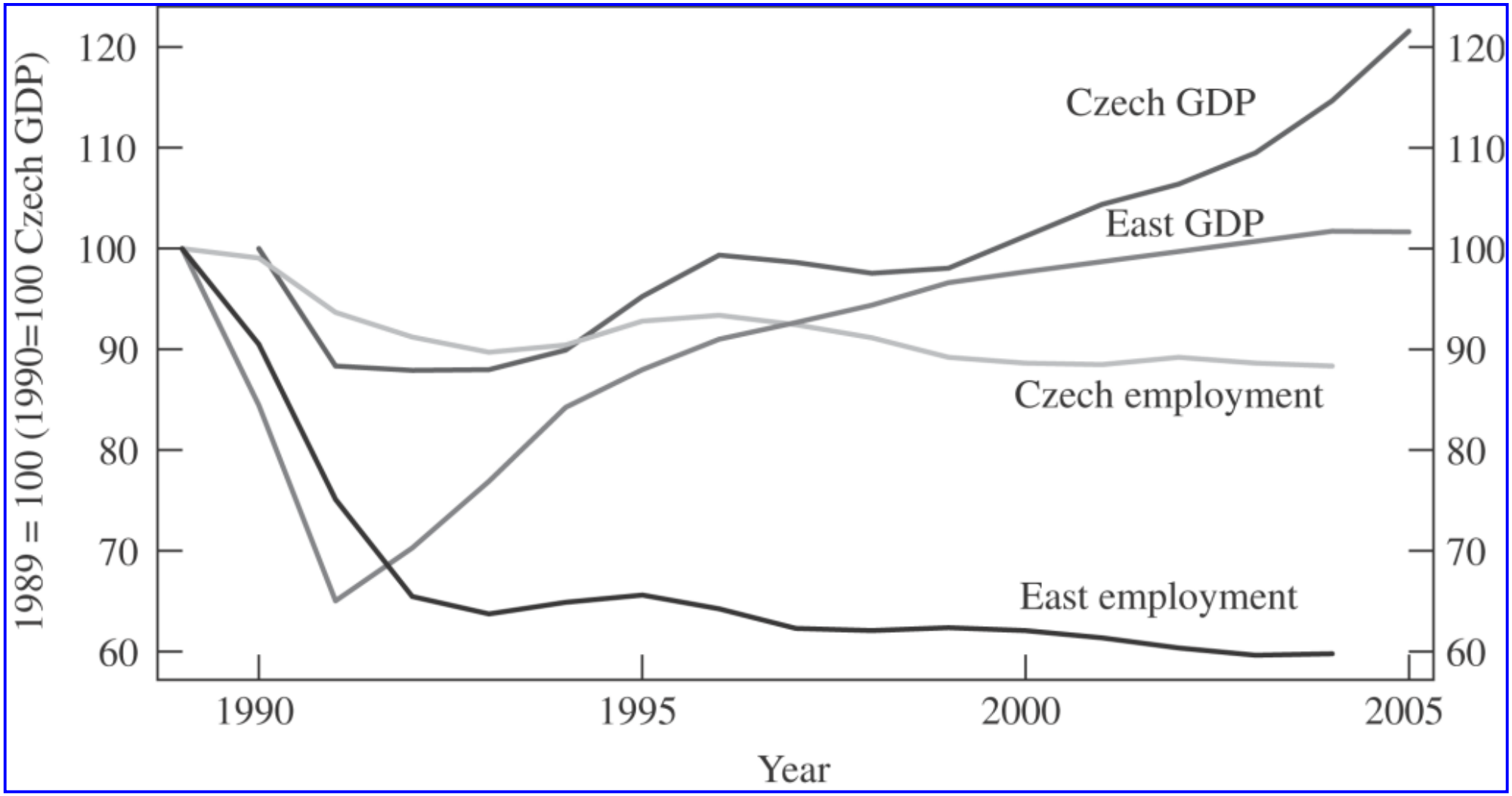

Explaining the initial collapse of GDP and employment

All former Communist countries except China experienced output declines following price liberalization, and many countries of the former Soviet Union had larger and longer output falls than East Germany. Roland (2000) examines why price liberalization depressed output, emphasizing theories of disruption of supply chains and the need to identify new business partners before investing. The main other potential culprits for the GDP and employment declines in East Germany are a reduction in labour supply, substitution to western products, the exchange rate chosen for monetary union, the increase in wages, and the privatization process.

Reduction in labour supply

Some of the output decline could have been caused by the employment decline rather than the reverse. Employment declined by 3.3 million people from 1989 to 1992. The government paid one million people to stop working by offering early retirement onto the western pension benefits implied by easterners' years of work experience. A further one million people emigrated to the west in 1989–91 (Hunt, 2006, draws lessons from eastern emigration).

Among the prime-aged remaining in the east, women experienced a particularly large employment decline, a fact often explained by the dismantling of the Communist day-care

system. However, Hunt (2002) shows that the employment rate of women with small children fell by no more than that of other women.

Substitution to western goods

Immediately upon monetary union, eastern shops filled with western goods. Easterners wanted to consume western products, and at this time ‘one couldn't sell an East German egg’ (personal communication from eastern state politician; see also Sinn and Sinn, 1992). Economists agree that this caused a sudden fall in demand for eastern goods, and hence a fall in output.

Monetary union exchange rate

For political reasons, the (western) government decided to choose a one-to-one exchange rate between the eastern Ostmarks and the western Deutschmarks. Early studies, in particular, argued that an overvalued exchange rate had made the eastern products uncompetitive with western products, leading to an output decline. With hindsight, it seems unlikely that the exchange rate was an important contributor to the output decline, as eastern prices and wages subsequently rose, rather than falling to correct the real exchange rate.

Unions and the wage increase

Although it is possible that the rapid rise in wages was the result of factor price equalization across regions, there is a consensus that labour unions were the driving force behind the rise. Unions acquired great power at a time when employers had little, and were not acting only in the interests of eastern workers. Western unions established themselves in the east in 1990, and were very successful in recruiting new members. The new eastern unions were led by westerners, who were concerned with east–west equity and eastern welfare but also with western wages and the perceived threat to them of mass east–west migration. The unions pushed for rapid wage convergence with the west, believing this was just, would prevent mass migration, and would enable eastern workers who were laid off to receive higher unemployment benefits (these being tied to the pre-layoff wage). At this time, most firms had no owners, and the unions were bargaining either with managers, who had no incentive to resist wage increases, or even with members of the western employers’ federations, whose incentive was to prevent undercutting of western prices by eastern firms.

Most economists believe this rapid rise in wages represented a classic textbook wage floor that reduced employment, led to high unemployment, and made East German firms unviable, thereby leading to the output collapse (for example, Akerlof et al., 1991; Sinn and Sinn, 1992; Sachverständigenrat, 2004).

Privatization

Small, mostly service firms were privatized separately from large industrial firms. As in eastern European countries, this privatization was rapid and successful, and was completed by March 1992 (Sinn and Sinn, 1992). Large industrial firms were privatized by a politically independent body known as the Treuhandanstalt (THA). Its initial portfolio was 8,500 previously state-owned enterprises containing 44,000 plants and 45 per cent of the workforce (Carlin, 1994).

The THA closed the unviable firms and plants, reduced employment at the viable plants, and sought buyers for the remaining core businesses. The THA's aim, at which it was successful, was to match firms with western management expertise in the same industry (Dyck, 1997). Weighted by employment, 74 per cent of sales were to West German firms or families, six per cent were to non-German firms, and only 20 per cent were to eastern buyers. Privatization thus created subsidiaries of western companies (Carlin, 1994). By 31 December 1994, the THA had finished its privatization with net losses of DM 193 billion (about 95 billion euros or 120 billion US dollars; Brada, 1996).

The THA destroyed many jobs in the short run, with the aim of curtailing inefficient production and promoting faster medium-run employment and output growth than would otherwise have occurred. Most economists studying privatization believe that the THA carried out its mandate well, leaving a legacy of viable and well-run companies. However, Roland (2000) believes that the employment reduction necessitated by the mandate caused a depression in the short run and retarded transition in the medium run.

Explaining the persistent labour market problems

Even observers who did not expect faster GDP convergence than has occurred are dismayed at the state of the labour market. Most explanations for the initial employment collapse apply to the short run only. Even labour union power has been severely weakened: while unions controlled wages from 1990 to 1993, a subsequent employer revolt allowed wages to be determined more freely. The share of workers whose employer belonged to an employer federation, which determines whether workers are paid the union wage, declined from 76 per cent in 1993 to 45 per cent in 1998 and 29 per cent in 2003 (Brenke, 2004).

Either the causes of the initial collapse have had lasting effects – for example, perhaps it is hard to reduce real wages in a low-inflation environment – or there must be other explanations. The leading one is the introduction of the western social welfare system. Others are investment subsidies, the wholesale transfer of western regulations, ineffectual active labour market programmes and impediments to the optimal allocation of resources across sectors.

Social welfare and wage floors

Many economists (for example, von Hagen and Strauch, 1999) stress the disincentives of the social welfare system as a cause of low employment in East Germany. After a very brief transition period, benefits were set at western levels, which in some cases made them higher relative to wages than in the west. This was the case in particular with *Sozialhilfe*, or social assistance (welfare), and with pensions. Unemployment insurance benefits are a fraction of the previous wage, so, to the extent that unemployment insurance is a greater problem than in the west, it is related to wages being too high. A generous social safety net sets a floor under wages, similar to a union wage, though affecting labour supply rather than labour demand.

The wage floor theory implies that wages at the bottom of the distribution should have risen the most, while employment of the least skilled should have fallen the most. Employment rates indeed fell more for the less skilled than the skilled. However, wage growth for the skilled was equal to or greater than that of the unskilled (Burda and Hunt, 2001). Furthermore, by 1999 wage inequality and the wage structure more generally were very similar to those in the west. Patterns of unemployment duration were also similar (Hunt, 2004). These results are inconsistent with the effect of a wage floor for the less skilled, which appears to rule out the social welfare theory. However, it is possible that a wage floor was too simple a model for the effects of the unions, who indeed appeared to aim to raise the wages of all members.

Investment subsidies

At least with hindsight, subsidizing capital (investment) in the face of grave labour market difficulties seems not obviously a good idea. Indeed, the capital–labour ratio in manufacturing is now higher in the east than the west (Sachverständigenrat, 2004). Furthermore, many have criticized the subsidies as being skewed towards structures at the expense of equipment (for example, Burda and Hunt, 2001). Finally, the subsidies were designed as tax breaks, and were hence attractive only to profitable, that is, established western, companies. The funds for investment subsidies appear not to have been spent optimally.

Active labour market programmes

Easterners are well educated, and the return to eastern schooling was not reduced by transition (Krueger and Pischke, 1995). The post-unification fall in the return to experience indicated that the human capital lacking was experience working in capitalist firms. Off-the-job training and make-work jobs were therefore unlikely to be very helpful, despite the large number of participants: in 1994 there were 259,000 participants in public training programmes and 280,000 participants in jobs whose wage was paid by the government, compared with 1,142,000 registered unemployed.

The best-documented effect of training programmes has been that of keeping participants out of the labour force for the duration of the sometimes long programmes (Lechner, Miquel and Wunsch, 2007). Meanwhile, participants in public jobs had no incentive to look for another job, as they received 100 per cent of the union wage (90 per cent from 1994 on). While some groups have benefited from some public programmes, the gains are unlikely to have justified the large expenditures (Eichler and Lechner, 2002; Lechner, Miquel and Wunsch, 2007).

Sectoral allocation

Various factors may have intervened to prevent an optimal allocation of resources across sectors. Brada (1996) observes that the THA requirement that buyers continue operating the firm in the same industry as before may have delayed sectoral restructuring. Unions, bargaining at the industry level, may have chosen the wrong wage structure across sectors, reducing incentives for restructuring (Burda and Hunt, 2001; Hunt, 2001). A further complicating factor has been the boom and subsequent bust of the construction industry. Many observers believe the manufacturing sector is too small, at 15 per cent of employment in 2004 compared with 22 per cent in West Germany and 30 per cent in the Czech Republic. Yet manufacturing in the United States employed a smaller share of the workforce than in East Germany in 2004, so East Germany may simply have leapfrogged West Germany in this regard.

Effect of unification on the west

In the short term, reunification was a positive aggregate demand shock for the west, leading to the boom seen in Table 1. The leap in the demand for capital for investment in the east, combined with the reduction in the money supply to contain inflation, raised the interest rate. As the cost of reunification became clear, the government was forced to raise taxes, but debt rose from 41.8 per cent of GDP in 1989 to 64.2 per cent in 2003. The budget went from surplus in 1989 to a 3.1 per cent deficit in 1991, and has been close to or above three per cent since then.

It is unclear to what degree the western stagnation that has followed the 1993 end of the boom can be attributed to reunification. While exports have recovered, domestic demand has remained weak (Sachverständigenrat, 2005). This could possibly be the result of government debt leading consumers to revise their wealth downwards, depressing consumption and growth (Carlin and Soskice, 2006). The increase in western unemployment, seen in Figure 1, could be caused in part by increases in payroll taxes to finance reunification. On the other hand, Siebert (2005) emphasizes that before reunification West Germany had already had problems with sluggish growth, rising unemployment, and funding social security.

Posen (2005) considers that approximately 1.4 per cent of German GDP per year is paid in transfers to the east that are for neither investment nor infrastructure, nor part of the unified social welfare system. He calculates the opportunity cost of this money (that could have been invested and received a return), the increase in interest payments on other debt (owing to a higher interest rate caused by higher debt), and the deadweight loss from increased taxes. He concludes that the burden of these transfers is (at most) 0.7 per cent of German GDP per year, a large sum.

Reunification has affected the West German labour market through the weakening of labour unions caused by the collapse of eastern unions. The impact of eastern immigrants and commuters is not known. The impact, if any, would have been in addition to that of the concomitant and similarly sized immigration of ethnic Germans from other formerly Communist countries.

Effect of unification on Europe

The rise in the German interest rate had important consequences for Europe, as it led to a crisis in the European Exchange Rate Mechanism (ERM) that preceded European Monetary Union. The higher German interest rate meant that the Deutschmark required a revaluation within the ERM, or, equivalently, the devaluation of other ERM currencies. France and other countries attempted to maintain the existing exchange rates, fearful of a loss of deflationary credibility. But in 1992 speculative attacks forced several countries to devalue, while the United Kingdom and Italy left the ERM.

The crisis was not all bad in the long run: for the United Kingdom, which had joined the ERM at an unsuitable exchange rate, leaving the ERM proved to be an economic boon (Carlin and Soskice, 2006). However, Germany may have entered monetary union at a rate that would prove overvalued once the reunification shock to interest rates had passed (Sinn, 1999), thus requiring a later depreciation. The difficulty of price and wage adjustments within monetary union may currently be preventing such a depreciation from occurring, slowing German, and therefore European, growth (Carlin and Soskice, 2006).

Lessons learned

Because East Germany joined the well-functioning and larger Federal Republic of Germany, it could feasibly and credibly have an institutional ‘big bang’, immediately importing a coherent set of institutions generally suitable for the region. This provided confidence and familiarity to western investors. The institution that obviously made a poor transition was the industrial relations system: because labour unions were established before employer federations, labour unions were initially unnaturally strong, possibly with lasting consequences.

Some economists believe the social welfare system made an equally poor transition; yet the nature of reunification meant that there was politically no alternative to transferring the system fairly rapidly. Siebert (2005) bemoans the transfer of product regulation and taxation. Yet firms may have complied with western constraints even had they not been imposed on the east, either in the expectation of their being imposed later or for fear of disgruntlement in their western works council. For example, Volkswagen applied the western prohibition on female night work to its eastern plant although the east was exempt (Turner, 1998).

The feasibility of an institutional big bang made feasible an economic big bang. Price liberalization and macro stabilization were flawless. The privatization process was speedy and had many merits, although it may have led to an excessive employment decline, and was too expensive for most countries to countenance. However, Koreans should note that even an unusual transition that satisfies both the ‘Washington consensus’ economists, who emphasize speed of economic reform, and the ‘evolution-institutionalist’ economists, who stress the necessity of establishing institutions before economic reform, can leave in its wake a difficult regional convergence problem.

For economists interested in unemployment, East Germany is both a validation of textbook models and a puzzle. Surely the collapse in employment and output in 1990–2 must have been strongly influenced by high union wages. Yet, now that labour unions have much less influence and the wage structure is similar to that of the west, why has unemployment

remained so high? Good education and high emigration are not enough to control unemployment.

See Also

- privatization
- total factor productivity
- transition and institutions
- unemployment
- unemployment insurance

I am very grateful to Michael Burda, Wendy Carlin, Adam Posen and Harald Uhlig for helpful discussions, and to Karl Brenke, Michaela Kreyenfeld, Joachim Ragnitz and Werner Smolny for generously and quickly providing me with data.

Bibliography

Akerlof, G., Rose, A., Yellen, J. and Hessenius, H. 1991. East Germany in from the cold: the economic aftermath of currency union. *Brookings Papers on Economic Activity* 1991(1), 1–87.

Brada, J. 1996. Privatization is transition – or is it? *Journal of Economic Perspectives* 10(2), 67–86.

Brenke, K. 2004. Ostdeutsche Industrie: Weitgehende Abkehr von der kollektiven Lohnfindung. *DIW-Wochenbericht* 13, 5–8.

Burda, M. and Hunt, J. 2001. From reunification to economic integration: productivity and the labor market in Eastern Germany. *Brookings Papers on Economic Activity* 2001(2), 1–92.

Carlin, W. 1994. Privatization, distribution and economic justice: efficiency in transition. In *Privatization in Central and Eastern Europe*, ed. S. Estrin. London: Longman.

Carlin, W. and Soskice, D. 2006. *Macroeconomics: Imperfections, Institutions and Policies*. Oxford: Oxford University Press.

Dyck, I. 1997. Privatization in Eastern Germany: management selection and economic transition. *American Economic Review* 87, 565–97.

Eichler, M. and Lechner, M. 2002. An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics* 9, 143–86.

Hunt, J. 2001. Post-unification wage growth in East Germany. *Review of Economics and Statistics* 83, 190–5.

Hunt, J. 2002. The transition in East Germany: when is a ten point fall in the gender wage gap bad news? *Journal of Labor Economics* 20, 148–69.

Hunt, J. 2004. Convergence and determinants of non-employment durations in Eastern and Western Germany. *Journal of Population Economics* 17, 249–66.

Hunt, J. 2006. Staunching emigration from East Germany: age and the determinants of migration. *Journal of the European Economic Association* 4(5), 1014–37.

Krueger, A. and Pischke, J.-S. 1995. A comparative analysis of East and West German labor markets: before and after unification. In *Differences and Changes in Wage Structures*, ed. R. Freeman and L. Katz. Chicago: University of Chicago Press.

Lechner, M., Miquel, R. and Wunsch, C. 2007. The curse and blessing of training the unemployed in a changing economy: the case of East Germany after unification. *German Economic Review* 9(1), 468–905.

OECD (Organisation for Economic Co-operation and Development). 2005. *OECD in Figures, 2005*. Paris: OECD.

Posen, A. 2005. Much ado about little: the fiscal impact of German economic unification. *CESifo Forum* 4, 33–6.

Ragnitz, J. 2000. Was kostet die Einheit? Zur Bewertung der Transferleistungen für Ostdeutschland. In *Nutzen und Kosten der Wiedervereinigung*, ed. D. Brümmerhof. Baden Baden: Nomos.

Roland, G. 2000. *Transition and Economics: Politics, Markets, and Firms*. Cambridge, MA: MIT Press.

Sachverständigenrat zur Begutachtung der gesamtwirtschaflichen Entwicklung. 2004. *Erfolge im Ausland – Herausforderungen im Inland. Jahresgutachten 2004/05*. Reutlingen: Servicecenter Fachverlage. Online. Available at http://www.sachverstaendigenrat-wirtschaft.de, accessed 3 May 2006.

Sachverständigenrat zur Begutachtung der gesamtwirtschaflichen Entwicklung. 2005. *Die Chance nutzen – Reformen mutig voranbringen. Jahresgutachten 2005/06*. Reutlingen: Servicecenter Fachverlage. Online. Available at http://www.sachverstaendigenrat-wirtschaft.de, accessed 3 May 2006.

Siebert, H. 2005. *The German Economy*. Princeton: Princeton University Press.

Sinn, G. and Sinn, H.-W. 1992. *Jumpstart: The Economic Unification of Germany*. Cambridge, MA: MIT Press.

Sinn, H.-W. 1999. International implications of German unification. In *The Economics of Globalization: Policy Perspectives from Public Economics*, ed. A. Razin and E. Sadka. Cambridge: Cambridge University Press.

Turner, L. 1998. *Fighting for Partnership: Labor and Politics in Unified Germany*. Ithaca, NY: Cornell University Press.

von Hagen, J. and Strauch, R. 1999. Tumbling giant: Germany's experience with the Maastricht fiscal criteria. In *From EMS to EMU: 1979 to 1999 and Beyond*, ed. D. Cobham and G. Zis. London: Macmillan.

How to cite this article

Hunt, Jennifer. "German reunification, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000266> doi:10.1057/9780230226203.0634

The New Palgrave Dictionary of Economics Online

Germany, economics in (20th century)

Harald Hagemann

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The development of German economics in the 20th century is characterized by an interaction between internal scientific factors and external political factors. The dominance of the Historical School ended with the death of Schmoller and the First World War. The pressing economic problems of the young Weimar Republic stimulated macroeconomic research and national income accounting, whereas the Nazis' rise to power caused an important intellectual emigration from which German economics recovered only slowly after 1945. After an early period of ordoliberalism, as in other countries the development of economics has increasingly reflected a process of internationalization dominated by American economics.

Keywords

Arndt, H.W.; Aumann, R.; Bombach, G.; Brunner, K.; Bruno, M.; Burchardt, F.; business cycles; Christaller, W.; Colm, G.; Corden, W.M.; Cowles Commission; creative destruction; economic liberalism; Eltis, W.; entrepreneurship; Erhard, L.; Eucken, W.; Föhl, C.; Freiburg School; game theory; general equilibrium; German Historical School; Germany, economics in; Giersch, H.; Great Depression; Hahn, F.; Harsanyi, J.C.; Hayek, F.A.; Hilferding, R.; innovation; Jaffé, E.; Jessen, J.; Keynesianism; Krelle, W.; Launhardt, C.F.W.; Lautenbach, W.; Lederer, E.; Leontief, W.; Liefmann, R.; Lösch, A.; Lowe, A.; Mandelbaum, K.; Marginal Utility School; Marschak, J.; Marx, K. H.; Meidner, R.; Menger, C.; Menger, K.; *Methodenstreit* ; Mises, L. von; monetarism; Morgenstern, O.; Musgrave, R.; Myrdal, G.; Nash, J.; national income accounting; Neisser, H.; neoclassical synthesis; ordoliberalism; Peter, H.; Preiser, E.; price control; protectionism; Röpke, W.; Roscher, W.; Rosenstein-Rodan, P.; Schelling, T.; Schmoller, G. von; Schneider, E.; Schumpeter, J.A.; Selten, R.; Singer, H.; social market economy; Sombart, W.; spatial economics; Spiethoff, A.A.K.; Stackelberg asymmetric duopoly; Stackelberg, H. von; stagflation; Streeten, P.; Tisch, C.; wage controls; Wagemann, E.; Weber, A.; Weber, M.; *Werturteilsstreit*

Article

1900 – 1918: dominance of the Historical School

Until the beginning of the 20th century there were relatively few economics students in Germany and their training heavily depended on other disciplines. Not until the 1880s was a special Ph.D. option opened for economists, and it was not until 1923 that a special curriculum was set up that led to a diploma degree. Nevertheless, in the second half of the 19th century a growing professionalization process took place, with the emergence of scholarly journals and the foundation of economic societies, the most important of which was the Verein für Sozialpolitik in 1872 (Hagemann, 2001). The driving force behind, and chairman of, the Verein from 1890 until his death was Gustav Schmoller (1838–1917), the undisputed leader of the 'Younger' German Historical School and the most influential economist in imperial Germany, particularly in Prussia, at the beginning of the 20th century. Schmoller favoured a historical and ethical approach to economics and the inductive method of collecting large amounts of statistical material, rather than the abstract axiomatic–deductive method of the classical economists and his neoclassical contemporaries. From 1883 he was involved in a famous dispute on method with Carl Menger; however, although a major issue in German literature, this *Methodenstreit* did not play a significant role at the meetings of the Verein. In remarkable contrast, the question of whether economists or other social scientists should make normative judgements had been a heated issue since the 1880s. This *Werturteilsstreit*, in which Max Weber, with his strong plea for a clear separation of social *science* from social *policy*, was a key figure, escalated at the 1909 Vienna meeting. 'The controversy about norms and values' raged until the outbreak of the First World War, at which time it was still unsettled.

1918 – 33: the Weimar Republic and the new complexity

After the end of the First World War, and with the death of Schmoller in 1917, the Historical School lost the dominant position it had acquired in the German Empire from 1871, although it remained influential among many economists. In the early years of the Weimar Republic urgent policy problems, such as the socialization of production, reparation payments and particularly hyperinflation, dominated. From the mid-1920s onwards, theoretical and empirical research on business cycles became the most important issue, and a new generation of more theoretically oriented young economists came to the fore. This is best reflected in the 1928 Zurich meeting of the Verein für Sozialpolitik which focused on the explanation of business cycles. Major contributions were delivered by brilliant young economists including Adolph Lowe (born 1893), Friedrich August Hayek (born 1899), Wilhelm Röpke (born 1899) and Oskar Morgenstern (born 1902).

The long-run development of the capitalist economy and the study of the crises and economic fluctuations surrounding it had been a major issue since the mid-19th century. This holds for representatives of the 'Older' Historical School such as Wilhelm Roscher as well as for Karl Marx. At the beginning of the 20th century leading economists like Arthur Spiethoff and Werner Sombart became widely known with their studies on business cycles, with Joseph A. Schumpeter as the towering figure. In his *Theory of Economic Development* (Schumpeter, 1911/12) the idea of creative destruction by innovation and the notion that bank credit is the prerequisite of innovation and of the foundation of new

enterprises are of central importance. According to Schumpeter, capitalist development can proceed only in a wavelike form, and pioneering entrepreneurs are the key agents for this development.

More systematic empirical research on business cycles became established with the foundation of the German Institute for Business Cycle Research in Berlin in 1925, with Ernst Wagemann as director (he was also President of the Statistisches Reichsamt, the German Statistical Office). Two years later the Austrian Institute for Research on Business Cycles was founded in Vienna on the initiative of Ludwig von Mises. Hayek, the first director, was succeeded by Morgenstern in 1931 after Hayek moved to the London School of Economics. Furthermore, in 1926 the Kiel Institute of World Economics engaged Adolph Lowe to build up a new department for statistical world economics and international trade cycles, which soon developed a most promising research programme. Lowe managed to recruit a group of highly talented young economists including Gerhard Colm, a leading expert on public finance and financial statistics (which was especially important for the payment of reparations), who later became the chief architect of West Germany's successful currency reform of 20 June 1948; the monetary theorist Hans Neisser; Fritz (later Frank) Burchardt, who after the Second World War became director of the Oxford Institute of Statistics; Wassily Leontief (1927–28, 1930–31) who, while at Kiel, wrote his Berlin Ph.D. thesis on the economy as a circular flow but mainly worked on the statistical analysis of supply and demand curves; and Jacob Marschak (1928–30). Lowe's Kiel habilitation thesis, ‘How is business cycle theory possible at all?’ (Lowe, 1926), which raised the fundamental methodological problem of the (in)compatibility of business cycle theory with the theory of general economic equilibrium, also exerted a major influence on Hayek, as can be seen from the latter's 1928 Vienna habilitation thesis, published in English as *Monetary Theory and the Trade Cycle* (1933, ch. 1).

The pressing economic problems of the early Weimar Republic and the mass unemployment and deflation of the Great Depression in the closing years of the republic, which required practical solutions, can also explain why German economics in the interwar period was not completely dominated by academic economists. Practitioners at public and private research institutions as well as leading bureaucrats played a key role. This is reflected, for example, in the core members of the Kiel group, Lowe, Colm and Neisser, who all had worked for the Weimar government and/or the Statistisches Reichsamt, or in the personage of Wilhelm Lautenbach, a leading member of the Ministry of Economics whose practical proposals during the Great Depression earned him the nickname ‘the German Keynes’. Due to governmental needs the Weimar Republic invested heavily in macroeconomic research, which brought it to the forefront of statistical innovation and the development of national income accounting. Wagemann, as the key figure, provided the subsequent Nazi government with the statistical tools for economic planning (Tooze, 2001).

1933 and after: dismissal, expulsion, and emigration of German-speaking economists

The political events of 1933 marked an important turning point for the economics profession in Germany. Shortly after its rise to power the new Nazi government passed the Restoration of Civil Service Act, which formed the basis for the dismissal of ‘disagreeable’ persons from public services either for racist or for political reasons. By the winter of 1934/35 about 14 per cent of the faculty at German universities had been dismissed, but in economics the figure was 24 per cent. However, the dispersion was significant. Whereas some years later in Austria the great majority of dismissals and expulsions were concentrated on the University of Vienna, in Nazi Germany the three universities of

Berlin, Breslau (today's Wroclaw in Poland) and Frankfurt had to cope with the greatest losses. Breslau in Silesia had a large and recognized Jewish community – it was home, for example, to Fritz Haber, the Nobel Prize-winner in chemistry, and the historian Fritz Stern (1999). Neisser, the development economist Heinz Wolfgang Arndt, the trade theorist Warner Max Corden and the later architect of the Swedish workers' investment funds Rudolf Meidner, all came from Breslau.

In economics, the highest losses were suffered by the faculties of Heidelberg (which lost seven of 11 members), Kiel (five of ten) and Frankfurt (13 of 33). Heidelberg had been a blossoming centre of liberal intellectual discourse in the Weimar Republic and characterized by multidisciplinarity, as expressed in Emil Lederer's analysis of the ‘new middle classes’. A key role there was played by the Institute for Social and State Sciences, founded in 1924, with Max Weber's younger brother Alfred as the director until 1933 when he ordered students to remove the swastika flag from the main campus building. (Alfred Weber can be regarded as one of the very few scholars for whom the notion of ‘internal emigration’ really fits.) The new Goethe University in Frankfurt had developed into a leading centre of the social sciences within the short period of the Weimar Republic, as can be seen by looking at the long list of outstanding scholars dismissed in 1933, including Franz Oppenheimer, Karl Pribram, Lowe, Fritz Neumark, Karl Mannheim and his research associate Norbert Elias, the theologian Paul Tillich, among others. When Gunnar and Alva Myrdal visited the Institute of World Economics in Kiel in the summer of 1933 on behalf of the Rockefeller Foundation they diagnosed a deteriorated scientific reputation because by then almost all the best scholars had emigrated.

The long-term loss of quality and of international reputation of German economics caused by the political events of 1933 and thereafter is also reflected in the evolution of scholarly journals. German-language journals not only lost most of the émigré economists as contributors, but also most foreign economists stopped writing in the German language or publishing in German (and from 1938 onwards also in Austrian) journals. The obverse of this was the increased number of articles written by émigré economists in the leading Anglo-Saxon journals. With the exception of Spiethoff, who stayed in office as the editor of *Schmollers Jahrbuch*, journal editors were replaced after the Nazis' rise to power (Hagemann, 1991). From 1904 the *Archiv für Sozialwissenschaft und Sozialpolitik* had been the leading journal in economics and the social sciences, with a run of fine editors, Max Weber, Werner Sombart and Edgar Jaffé, followed in 1922 by Emil Lederer and his two associates, Joseph Schumpeter and Alfred Weber. But this journal too had to cease publication under Nazi rule. For a short period in the 1930s the Vienna-based *Zeitschrift für Nationalökonomie* became the outstanding scholarly journal in economics in the German language, under the experienced editorship of Oskar Morgenstern. It published important articles particularly on capital theory, the role of time in economics or general equilibrium analysis, but after Hitler's invasion of Austria the quality of the journal collapsed. In the wake of the *Anschluss* the Vienna Institute for Research on Business Cycles became a branch office of the Berlin Institute.

The group of dislocated German and Austrian economists who had acquired academic degrees comprises 253 scholars, of whom 148 were dismissed from universities, 57 from private research institutes, and 28 from other public employment, and 20 were young economists who just had completed their studies, students like Richard Musgrave who had gained his diploma degree at the University of Heidelberg in May 1933 (Hagemann and Krohn, 1999). Some 221 (87 per cent) emigrated. Of the remaining 32 several were killed in the Holocaust, concentration camps, or Gestapo prisons, these including Rudolf Hilferding, Robert Liefmann and Cläre Tisch. The intellectual loss included 75

members of the so-called ‘second generation’, young students or pupils who emigrated with their parents and later made a career as economists, such as, for example, Robert Aumann, Michael Bruno, Otto Eckstein, Walter Eltis and Frank Hahn.

1933 – 45: German economics in the Nazi period

German economics in the Nazi period moved far away from economic liberalism, and state interventions and regulations such as price controls and wage freezes played a major role (Janssen, 2000). The 1932 Dresden meeting of the Verein für Sozialpolitik, organized by Mises, focused on problems of value theory and was dominated by the representatives of the Marginal Utility School, but was also clearly marked by heated debates on protection and the question of autarky. This meeting turned out to be the last before the outbreak of the Second World War because in December 1936 the majority of the members decided to dissolve the Verein in order to escape *Gleichschaltung* by the Nazis.

This decision shows that the group of dedicated National Socialists, which included Feder and Wiskemann, was remarkably small. More important in numbers and influence were the two groups of fellow travellers including Gottl-Ottlilienfeld, Predöhl, Wagemann and national and conservative opportunists including Sombart and Spann respectively (Rieter and Schmolz, 1993, p. 95), who largely followed the historical–holistic approach and also introduced ideas from contemporary German philosophy into economics. A fourth group consisted of ‘renegades’, former Nazis who later changed their views. Prominent members of this latter group include Jens Jessen – who in 1933 succeeded Harms as the director of the Kiel Institute and, as a result of his involvement in the failed attempt to assassinate Hitler in July 1944, was executed some months later – and Heinrich von Stackelberg, who probably only escaped the same fate because of his move to Spain in 1943. A fifth group consisted of opponents of the Nazi regime who either passively distanced themselves from the regime, thereby ending their professional careers, or actively fought against it, like the members of the Freiburg School.

The first subgroup of opponents included Hans Peter, a very able mathematical economist and theorist of the circular flow, who together with Erich Schneider and Stackelberg from 1935 to 1942 edited the new journal *Archiv für mathematische Wirtschafts- und Sozialforschung*. Due to his defence of his liberal socialist convictions in the Nazi period Peter obtained a full professorship at the University of Tübingen only in 1947. August Lösch, a brilliant economist of great personal integrity, received his habilitation from the University of Bonn in 1939 with his *The Economics of Location*, in which he applied general equilibrium theory to the space dimension. (Since the days of Thünen, then via Launhardt, Alfred Weber and Christaller to Lösch, spatial economics has been an area of economics where German economists have made important contributions.) Lösch himself, however, did not have a successful professorial career: because of his outspoken anti-Nazi views he survived fascism only by becoming a researcher at the Kiel Institute. He died tragically from scarlet fever a few weeks after the end of the war.

Although in international terms German economics fell behind in the Nazi period, there were nevertheless some significant contributions. Stackelberg was one of the most gifted theoretical economists. His *Marktform und Gleichgewicht* (1934), with its creation of the Stackelberg asymmetric duopoly, went beyond the work of Chamberlin and Joan Robinson in the depth of its theoretical analysis and in its mathematical rigour. However, although he was one of the very few German economists whose analysis was deeply embedded in the Anglo-Saxon approach to price and cost theory, and

although the book was immediately reviewed by top theorists such as Hicks, Kaldor, Lange, Leontief and Zeuthen, it failed to make an impact. Because its author was a well-known supporter of the Nazis and the book ended with a short paean to the corporate state, the book's theoretical achievements were overlooked and even today it has not been translated into English.

Keynes was a central figure of reference in theoretical and policy debates in interwar Germany ever since his opposition to the reparation stipulations of the Versailles Treaty; not surprisingly, the first translation of his *General Theory* was into German (with a special Preface published in the same year 1936) and was extensively reviewed and debated. Leaving aside the complex question of German anticipations of the *General Theory* (Bombach et al., 1981), Carl Föhl's *Geldschöpfung und Wirtschaftskreislauf*, published in 1937 but already completed in December 1935, exhibits striking parallels with Keynes's theories (despite its using a different conceptual apparatus), and was the outstanding achievement in contemporary German literature on macroeconomics.

The post-1945 development of economics in Germany: ordoliberalism and the social market economy

As is well known, after the Second World War, economic order and economic policy in the new Federal Republic of Germany were decisively influenced by the ordoliberal thinking of Walter Eucken and the Freiburg School and the principles of the social market economy (Watrin, 1979). However, the roots of ordoliberalism go back to the years 1938–45, with opposition to National Socialism based on Christian convictions (Rieter and Schmolz, 1993). Eucken's main work, *Foundations of Economics*, was published in 1940. Although he was a well-known critic of the Historical School, his taxonomic approach, which focuses on reconciling economic with legal, institutional and social factors, is clearly embedded in the German tradition of state sciences. Thus in the competitive order he perceived that the state plays a substantially stronger role than his Anglo-American colleagues would allow, not to speak of Austrian contemporaries such as Mises and Hayek. This is particularly visible in Eucken's ‘regulating principles’ – that is, monopoly control, social policy, external effects – that supplement the ‘constituting principles’ of the market economy, that is, private ownership, competition in open markets, freedom of contract, a functioning price mechanism, monetary stability and consistency in economic policy.

German economics, which went through a laborious catching-up process after 1945 without being able to compensate fully in the following decades for the loss of qualified personnel in the Nazi period, suffered a further blow by the untimely deaths of Lösch (1945), Stackelberg (1946) and Eucken (1950). The economists who were driven out of Nazi Germany had meanwhile contributed significantly to innovative research in their host countries. This holds particularly for the United States, which was the final destination for about 60 per cent of the émigré economists, but also for the UK where the new field of development economics had been shaped by Paul Rosenstein-Rodan, Kurt Mandelbaum (Martin), Arndt, Hans Singer, Paul Streeten and others, all émigrés from German-speaking countries. In the United States economists at the New School for Social Research, where the graduate faculty had been founded as the ‘University in Exile’ in autumn 1933, gained a certain influence in the period of the Roosevelt and Truman presidencies (Krohn, 1993), but in the long run those who kept their distinctly German scholarly identity were less influential.

Among the important new developments in economics which were transferred to Germany after 1945, if with some delay, were game theory and econometrics, developed in particularly at the Cowles

Commission in Chicago after Marschak became the Research Director there in 1943. This work was later mirrored in Wilhelm Krelle's research centre in econometrics at the University of Bonn in the 1960s and 1970s. The foundations of modern game theory were laid by John von Neumann and Oskar Morgenstern's *Theory of Games and Economic Behavior* (1944), which was the fruit of their cooperation at the Institute for Advanced Study in Princeton in the years 1939–43. Their work together came about as a result of exile, although the Budapest-born mathematician von Neumann had made his habilitation at the University of Berlin in 1927 and presented his famous paper on the general economic equilibrium of an expanding economy at Karl Menger's mathematical colloquium at the University of Vienna in 1936. The award of the Nobel Prize in economics in 1994 to John F. Nash, the Hungarian-born John C. Harsanyi, and Reinhard Selten as the first German economist, as well as in 2005 to Robert Aumann and Thomas Schelling, also reflects the great role of German, Austrian and Hungarian scholars in the development of game theory.

The development of modern public finance by Musgrave reflects the cross-fertilization of the more theoretically oriented and rigorous Anglo-Saxon tradition of public finance with the German tradition of *Finanzwissenschaft*, including its institutional, historical and legal aspects. With his division of the public sector into three branches, which besides allocation also include stabilization and distribution as a fiscal concern, Musgrave indicates some German influences in the émigré's baggage (Musgrave, 1996). The German translation of Musgrave's *Theory of Public Finance* (1959) was used at most universities for more than two decades as the standard textbook. Distributional theory and wealth formation among workers were also key issues for some leading German economists – Erich Preiser, Gottfried Bombach and Krelle – in the 1960s.

1967 – 74: the high years of Keynesianism

At German universities the Keynesianism of the Hicks–Samuelson neoclassical synthesis had become the dominant position since the late 1950s. This can largely be attributed to Erich Schneider, whose three-volume *Introduction into Economic Theory*, originally published in 1946–52, became the dominant textbook in the 1950s and 1960s, going through many editions. Schneider, who had made his habilitation with Schumpeter in Bonn in 1932 and became Professor in Aarhus in 1936, came back from Denmark in 1946 to become professor in Kiel, where he also directed the Institute of World Economics from 1961–9 (when he was succeeded by Herbert Giersch). From 1963 to 1966 Schneider was chairman of the Verein für Socialpolitik, which had been re-founded in 1948. When the influential Theoretical Committee was re-established shortly afterwards, Schneider exercised his power as chairman in the direction of a more mathematically oriented approach, which at the beginning had to overcome strong resistance (Schefold, 2004).

Until the recession of 1966–7, however, economic policy was still dominated by ordoliberal ideas. As a consequence, Ludwig Erhard, who had been a successful Minister of Economics from 1949 to 1963, lost his job as Chancellor in December 1966, when the first ‘Grand Coalition’ of Christian and Social Democrats was formed. With the Social Democrats' entry into government and the ratification of the Stability and Growth Act in June 1967, Keynesianism gained a relatively late admission into Germany. According to Article 1 of the Act, the federal and state governments

> have to respect the requirement of macroeconomic equilibrium in their economic and financial policy measures which have to be taken in a way that they contribute, within the scope of a market economy, to simultaneously achieve stability of the price level, a high level of employment, and external equilibrium together with steady and appropriate growth.

These four macroeconomic goals appeared in the statutes of the German Council of Economic Advisers (CEA), which was founded in August 1963 and from autumn 1964 presented its annual report. The German council differs from the American in being an external and independent committee for policy consultation rather than part of the government.

In the public eye Karl Schiller's term of office as economics minister from the end of 1966 to the summer 1972 is remembered as the heyday of Keynesian economic policy in Germany. This is due to Schiller's remarkable ability to coin phrases such as '*Globalsteuerung*' ('macroeconomic demand management'), and his charismatic interpretation of economic policy which contributed to a widespread belief in the government's management power of macro variables, before the first oil price shock and the new phenomenon of stagflation shook that confidence. However, it should not be overlooked that Schiller had always followed a synthesis of Keynesianism and ordoliberal ideas. This is expressed most clearly in his influential article on economic policy in the *Handwörterbuch der Sozialwissenschaften* (Handbook of Social Sciences) (Schiller, 1962), in which he formulated his famous credo: 'competition to the extent possible, planning to the extent necessary', with 'planning' understood in the sense of Keynesian demand management. Through his homage to Eucken, that is, in supplementing process policy with *Ordnungspolitik*, Keynesian policies took on a distinctly German tinge. This came against the background of discrediting the interventionist policies of the Nazi period, policies that were being pursued in Stalinist East Germany, and the need to safeguard the market economy against Marxist policies that were finally given up by the Social Democratic Party (SPD) only in its Godesberg programme adopted in 1959. In line with Giersch and the majority of the CEA, Schiller also advocated flexible exchange rates in the final years of the Bretton Woods system and in debates in the German 1969 election campaign when currency flexibility was heavily opposed by the Christian Democrats and the German export industry. The strong revaluation of the Deutschmark thereafter contributed to a dampening of inflation in Germany.

From 1974 to the present: is there a German economics?

Even in the short period when Keynesian influence on policy was at its peak, the Bundesbank was a powerful institution that followed its own policy of securing price stability, thereby constraining the implementation of Keynesian full employment policies. After the second oil price shock, when the German economy ran into a current account deficit in 1979–81 and there was unusual pressure to devalue the Deutschmark, the Bundesbank reacted by raising interest rates, a restrictive monetary policy that led to a major controversy with Chancellor Schmidt. After December 1974, when for the first time it had announced a target for the growth of the money supply, the Bundesbank followed an explicit monetarist policy, in line with many of the major central banks in the Western world. The Constance seminar on monetary theory and policy, initiated by Karl Brunner, which from 1970 had brought together American and German economists as well as practitioners from the Bundesbank and

commercial banks, was the main vehicle for the breakthrough of monetarist ideas in theory and practice. Since 1976 the CEA, the majority of whose members had favoured moderate wage policies – which repeatedly led them into controversy with the trade unions – explicitly propagated supply side policies. With German unification in 1990 another German *Sonderweg* ended. University economics faculties in East Germany, where from 1945 to 1989 economics had been mainly reduced to Marxism-Leninism and a narrowly defined socialist business administration, were now completely restructured, as also happened in law and the social sciences. The government tried to restore the pre-1933 prestige of the Humboldt University in Berlin, formerly Germany's leading academic institution, which produced many Nobel Prize-winners in physics, chemistry and medicine. Its economics faculty entered the club of the leading faculties, which included Bonn, Mannheim and Munich.

The post-1945 development of economics is characterized by a growing process of internationalization combined with an increasing dominance of American economics (Coats, 1997; 2000). The triumphant ascent of American economics after the Second World War is the consequence of the economic and political leadership of the United States, the benefits of the importation of scholars from Hitlerian and Stalinist Europe – Scherer (2000, p. 622) calculates the citations received by the German-speaking émigré economists in the Social Science Citation Index (SSCI) for the period 1960–4 as ‘roughly equivalent to the adjusted citation output of the first-ranked Harvard and second-ranked MIT plus the 19th-ranked University of Illinois economics departments’ – and a national style of economic research characterized by the early introduction of graduate studies at the leading universities, with pressure to acquire the necessary mathematical and econometric tools for a specialized theoretical and applied work. The consequential international convergence process and the ‘professional *Gleichschaltung*’ (Peacock in Frey and Frey, 1995, pp. 267–71) associated with it led to increasing debates of the type ‘Is there a European economics?’ (Frey and Frey, 1995). In that sense, at the end of the 20th century there is no recognizable ‘German’ economics, as there was at the beginning of the century. German economists participate in international networks as do economists from other European countries or other parts of the globe, with English as the lingua franca. This is also reflected in the fact that since 2000 the Verein für Socialpolitik has published the international *German Economic Review* instead of the *Zeitschrift für Wirtschafts- und Sozialwissenschaften*, formerly *Schmollers Jahrbuch*, which had been revived under its old name but is no longer linked to the Verein, which in 2000 had 2,928 individual members from more than 20 countries.

Bibliography

Bombach, G., Netzband, K.-B., Ramser, H.-J. and Timmermann, M., eds. 1981. *Der Keynesianismus III. Die geld- und beschäftigungstheoretische Diskussion in Deutschland zur Zeit von Keynes*. Berlin, Heidelberg, New York: Springer.

Coats, A.W. 1997. *The Post-1945 Internationalization of Economics*. Durham, NC: Duke University Press.

Coats, A.W. 2000. *The Development of Economics in Western Europe since 1945*. London: Routledge.

Eucken, W. 1940. *The Foundations of Economics*. London: William Hodge, 1950.

Föhl, C. 1937. *Geldschöpfung und Wirtschaftskreislauf*. Berlin: Duncker & Humblot.

Frey, B.S. and Frey, R.L. 1995. Is there a European economics? *Kyklos* 48, 185–311.

Hagemann, H. 1991. Learned journals and the professionalization of economics: the German language area. *Studi Economici* 1, 33–57.

Hagemann, H. 2001. The Verein für Sozialpolitik from its foundation (1872) until World War I. In *The Spread of Political Economy and the Professionalisation of Economists: Economic Societies in Europe, America and Japan in the Nineteenth Century*, ed. M.M. Augello and M.E.L. Guidi. London and New York: Routledge.

Hagemann, H. and Krohn, C.-D. 1999. *Biographisches Handbuch der deutschsprachigen wirtschaftswissenschaftlichen Emigration nach 1933*, 2 vols. Marburg: Metropolis.

Hayek, F.A. 1929. *Monetary Theory and the Trade Cycle*. London: J. Cape, 1933.

Janssen, H. 2000. *Nationalökonomie und Nationalsozialismus. Die deutsche Volkswirtschaftslehre in den dreißiger Jahren*. Marburg: Metropolis.

Krohn, C.-D. 1993. *Intellectuals in Exile: Refugee Scholars and the New School for Social Research*. Amherst, MA: University of Massachusetts Press.

Lösch, A. 1939. *The Economics of Location*. New Haven: Yale University Press, 1954.

Lowe, A. 1926. How is business cycle theory possible at all? *Structural Change and Economic Dynamics* 8(1997), 245–70.

Musgrave, R.A. 1996. Public finance and Finanzwissenschaft traditions compared. *Finanzarchiv* 53, 145–93.

Rieter, H. and Schmolz, M. 1993. The ideas of German Ordoliberalism 1938–45: pointing the way to a new economic order. *European Journal of the History of Economic Thought* 1, 87–114.

Schefold, B. 2004. Wissenschaft als Gegengabe – Neugründung und Aktivitäten des Theoretischen Ausschusses im Verein für Socialpolitik von 1949–1973. *Schmollers Jahrbuch* 124, 579–608.

Scherer, F.M. 2000. The emigration of German-speaking economists after 1933. *Journal of Economic Literature* 38, 614–26.

Schiller, K. 1962. Wirtschaftspolitik. *Handwörterbuch der Sozialwissenschaften*, vol. 12. Stuttgart: Gustav Fischer.

Schneider, E. 1946–52. *Einführung in die Wirtschaftstheorie*, 3 vols. Tübingen: Mohr-Siebeck.

Schumpeter, J.A. 1911/12. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Cambridge, MA: Harvard University Press, 1934.

Stackelberg, H.von. 1934. *Marktform und Gleichgewicht*. Vienna and Berlin: Julius Springer.

Stern, F. 1999. *Einstein's German World*. Princeton, NJ: Princeton University Press.

Tooze, A. 2001. *Statistics and the German State, 1900–1945*. Cambridge: Cambridge University Press.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Watrin, C. 1979. The principles of the Social Market Economy: its origins and early history. *Zeitschrift für die gesamte Staatswissenschaft* 135, 405–25.

How to cite this article

Hagemann, Harald. "Germany, economics in (20th century)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000192> doi:10.1057/9780230226203.0635

The New Palgrave Dictionary of Economics Online

Gerschenkron, Alexander (1904 – 1978)

Albert Fishlow
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

backwardness; capital accumulation; economic history; Gerschenkron, A.; Industrial Revolution; industrialization; Laspeyres index; latecomer development; Modernity; Paasche index; stages of growth; structural change

Article

Gerschenkron was born in Odessa in 1904 and died in Cambridge, Massachusetts, in 1978. He left Russia in 1920 and settled in Austria. In 1938, a decade after receiving the degree of *doctor rerum politicarum* from the University of Vienna, he emigrated to the United States and spent the next six years at Berkeley. After a short period at the Federal Reserve Board, he went to Harvard in 1948 to teach both economic history and Soviet studies. His passion for the former dominated, and he flourished there as the doyen of economic history in the United States. He influenced a generation of Harvard economists through his required graduate course in economic history and attracted several to his seminar and the field. His erudition and breadth were legendary, and defined an indelible, if unattainable, standard of scholarship for his colleagues and students.

Gerschenkron's principal contribution to economics was the elaboration of a model of latecomer economic development. Its central hypothesis is the positive role of relative economic backwardness in inducing systematic substitution for supposed prerequisites for industrial growth. State intervention could, and did, compensate for the inadequate supplies of capital, skilled labour, entrepreneurship and technological capacity found in follower countries. Thus the German institutional innovation of the ‘great banks’ provided access to needed capital for industrialization, even while greater Russian backwardness required a larger and more direct state role.

Gerschenkron's analysis is consciously anti-Marxian: it rejected the English Industrial Revolution as the normal pattern of economic development, and deprived the original accumulation of capital of much of its conceptual force. Elements of modernity and backwardness could survive side by side, and did in a systematic way. Apparent disadvantageous initial conditions of access to capital could be overcome. Success was rewarded with proportionately more rapid growth, signalled by a decisive spurt in industrial

expansion.

This model, first presented in 1952 in an essay entitled *Economic Backwardness in Historical Perspective* (reprinted in 1962), underlay Gerschenkron's extensive research into the specific developmental experiences of Russia, Germany, France, Italy, Austria and Bulgaria. Out of those historical studies emerged a comparative, all-encompassing European picture. ‘In this fashion, the industrial history of Europe is conceived as a unified and yet graduated pattern’ (Gerschenkron, 1962, p. 1). In turn, his hypotheses became progressively more precise. They may be summarized as follows:

1. 1. Relative backwardness creates a tension between the promise of economic development, as achieved elsewhere, and the reality of stagnancy. Such a tension motivates institutional innovation and promotes locally appropriate substitution for the absent preconditions of growth.
2. 2. The greater the degree of backwardness, the more interventionist was the successful channelling of capital and entrepreneurial guidance to nascent industries. Also, the more coercive and comprehensive were the measures to reduce domestic consumption.
3. 3. The more backward the economy, the more likely were: an emphasis upon producers' goods rather than consumers' goods; use of capital-intensive rather than labour-intensive methods of production; emergence of larger rather than smaller units of both plant and enterprise; dependence upon borrowed, advanced technology rather than indigenous techniques.
4. 4. The more backward the country, the less likely was the agricultural sector to provide a growing market to industry through rising productivity, and the more unbalanced the resulting productive structure of the economy.

The considerable and continuing appeal of the Gerschenkron model derives from its logical and consistent ordering of the process of European development, the conditional nature of its predictions and its generalizability to the experience of the late latecomers of the present Third World. His formulation rises above other theories which emphasize stages of growth both because of its attention to historical detail and its insistence upon the special attributes of latecomer development that cause differential evolution. In Gerschenkron's own hands, his propositions afforded an opportunity to blend ideology, institutions and the historical experience of industrialization, especially that of Russia, in a dazzling fashion. For others, his approach has proved a useful starting point for the discussion of non-European latecomers, including Japan and the newly industrializing countries.

The model is, of course, not without its limitations. History, even of Europe alone, does not in every detail bear easily the weight of such a grand design. In other parts of the world, as might be expected from a concept rooted in the special features of the historical European experience, larger amendments are frequently required. And somewhat surprisingly, in view of Gerschenkron's own path-breaking essay in political economy, *Bread and Democracy in Germany* (1943), there is too little attention to the domestic classes and groups whose interests the interventionist state must adequately incorporate if it is to play the central role required. Backwardness too easily becomes an alternative, technologically rooted explanation, distracting attention from the state rather than focusing upon its opportunities and constraints.

Still, the concept of relative backwardness, and Gerschenkron's always insightful and rich elaborations in so many national contexts, represent a brilliant and original contribution to economic history for which he is justly celebrated. It is not the only one. The ‘Gerschenkron effect’, arising from the

difference between calculated Paasche and Laspeyres indexes of Soviet machinery output (1951), also commemorates him. Current price weights will tend to underestimate the extent of growth because prices and quantities are negatively correlated, just as base year weights exaggerate it. The larger is the difference between the alternatively constructed quantity indexes, the greater is the degree of structural change. Again, divergence rather than uniformity is the source of useful information about historical processes.

Alexander Gerschenkron has few peers, past or present, in his command of comparative economic history. Scholarly interest in contemporary economic development has brought him an increasing following. His insights thus continue to influence a new generation of scholars and guarantee him a central place in any assessment of the evolution of the discipline of economic history.

Selected works

1943. *Bread and Democracy in Germany*. Berkeley: University of California Press.

1951. *A Dollar Index of Soviet Machinery Output*. Santa Monica: Rand Corporation.

1962. *Economic Backwardness in Historical Perspective*. Cambridge, MA: Harvard University Press.

1968. *Continuity in History and Other Essays*. Cambridge, MA: Harvard University Press.

1970. *Europe in the Russian Mirror*. London: Cambridge University Press.

1977. *An Economic Spurt That Failed*. Princeton: Princeton University Press.

How to cite this article

Fishlow, Albert. "Gerschenkron, Alexander (1904–1978)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000033> doi:10.1057/9780230226203.0636

The New Palgrave Dictionary of Economics Online

Gervaise, Isaac (fl. 1680 – 1720)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

free trade; Gervaise, I.; Law, J.; specie-flow mechanism

Article

Merchant and economist of French Huguenot extraction. Gervaise was born in the second half of the 17th century, probably in Paris, and migrated with his family to London in 1681. With his father he was associated with the Royal Lustring Company (1688–1720) engaged in the manufacture of a fine, light, black, glossy silk under patent granted by parliament. Ironically, the year the company lost its charter saw the publication of Gervaise's 34-page pamphlet, *The System or Theory of the Trade of the World* with its attack on exclusive companies. Foxwell (1940, p. 167) described it as ‘one of the earliest formal systems of political economy … stating one of the most forcible practical arguments for free trade’. Quite unlike much contemporary writing on trade, Gervaise's pamphlet is tersely written and especially noted for its peculiar terminology and highly abstract argument. Gervaise is presumed to have died in London by 1739.

The real discoverer of Gervaise's work, Viner (1937, pp. 79–80), has described it as ‘an elaborate and close reasoned exposition of the nature of international equilibrium and of the self-regulating mechanism whereby specie obtained its “natural” or proper international distribution’. The novelty of Gervaise's treatment of the specie mechanism is his emphasis on the role of income rather than prices, in strong contrast with subsequent treatments by Cantillon, Vanderlint and Hume. The starting point for the analysis is the proposition that the equilibrium bullion stock of any nation is proportioned to its output in terms of labour and that such a stock also maintains the balance between consumption and production, exports and imports. Excess bullion breaks these balances by raising consumption and reducing production, thereby lowering exports and raising imports, hence bullion will be exported and the balances will be resorted. An inadequate bullion stock leads to specie inflow by raising production relative to consumption, and exports to imports. Gervaise treats credit as if it were bullion; oversupply or deficiency is self-correcting via the balance of trade, though the adjustment process with credit is more rapid through its additional income effects of interest payment to suppliers of credit, whom he sees as

consumers rather than producers. Hence ‘credit is of pernicious consequences to that Nation that uses or encourages it beyond nature’ (Gervaise, 1720, p. 14) – a comment perhaps not unrelated to contemporary developments with Law's system in France. War, capital consumption or export, and restrictions on trade may prevent or postpone attainment of monetary equilibrium. For this reason, and for the resource misallocation potential flowing from encouragement of specific manufactures through companies, laws or taxes on imports, Gervaise (1720, pp. 17–18) concluded that ‘Trade is never in a better condition, than when it's natural and free’. Gervaise also pointed out that the ‘natural proportion’ of bullion for specific countries was influenced by their situation, particularly as regards proximity to water transport, and that implementing policies of debasing the currency had similar effects on trade and the balance of consumption and production as credit oversupply. Although less elegantly written than Hume's later account of the specie mechanism, the emphasis on adjustment through income rather than price effects, though not always clearly explained, makes Gervaise's short and penetrating contribution to the subject more modern than most of its successors in the century and a half which followed.

Selected works

1720. *The System or Theory of the Trade of the World*. Reprinted with a biographical introduction by J. M. Letiche and a foreword by J. Viner, Baltimore: Johns Hopkins Press, 1954.

Bibliography

Foxwell, H.S. 1940. Comment reproduced in *Catalogue of the Kress Library of Business and Economics*, Baker Library, Harvard Graduate School of Business Administration, Boston.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper and Brothers.

How to cite this article

Groenewegen, Peter. "Gervaise, Isaac (fl. 1680–1720)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000034> doi:10.1057/9780230226203.0637

The New Palgrave Dictionary of Economics Online

Gesell, Silvio (1862 – 1930)

Victoria Chick

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

depressions; Gesell, S.; hoarding; land nationalization; stamped money principle

Article

Gesell was born in Germany but emigrated to the Argentine in 1886, where he was so successful as an importer that he retired to Switzerland in 1900 to farm and to continue to write. The ‘retirement’ included a return to the Argentine to manage his late brother's business and an involvement in Bavarian politics at their most chaotic. As a deposed minister of finance he was tried for high treason, and acquitted.

His prolific writing began in Argentina, provoked by the economic chaos of the late 1880s there. But his fame rests on *The Natural Economic Order*, originally published in two parts in 1906 and 1911. It was translated into English in 1929. Rent-free land and interest-free money characterize that Order. Land would be nationalized, its owners compensated by the issue of state bonds. Through the device of stamped money, which would remain current only if a stamp, obtained at a cost set by government, was regularly affixed, the rate of interest on these bonds and other lending instruments would eventually be driven to zero. With no income diverted to rent or interest the worker would receive the full value of his output. Mothers were to receive income from annuities based on the nationalized land, since their ‘output’, the population, was the source of demand for land and hence rent.

Gesell attributed depressions to inadequate investment and the latter to the fall in the expected rate of return as investment continued, coupled with a money rate of interest which was prevented from falling by the alternative opportunity of hoarding. This analysis substantially anticipates Keynes's (1936), as Keynes amply acknowledges (pp. 353–8). Gesell suggested adjusting the stamp duty on money to force down the rate of interest.

The stamped money principle was three times applied on a local scale in the 1930s: in Bavaria, in the Austrian Tyrol and in Alberta, Canada. In each case the scheme successfully raised demand and employment, but the money was soon banned by the authorities.

Though theoretical inadequacies and practical difficulties are claimed against Gesell's theory, its *aim* is

probably more responsible for its eclipse. But it lives on furtively, below the surface, in the underworlds of Keynes's *General Theory* and Fisher's *Booms and Depressions*.

Selected works

1891a. *Currency Reform as a Bridge to the Social State*. Buenos Aires. Trans. P. Pye, typescript, 1951.

1891b. *Nervus Rerum: Continuation of 'Currency Reform as a Bridge to the Social State'*. Buenos Aires. Trans. P. Pye, typescript, 1951.

1929. *The Natural Economic Order: A Plan to Secure an Uninterrupted Exchange of The Products of Labour*. Trans. P. Pye, Berlin: Neo-Verlag. Online. Available at http://www.systemfehler.de/en/neo/, accessed 10 November 2006.

Bibliography

Fisher, I. 1933. *Booms and Depressions*, Appendix 7. London: George Allen & Unwin.

Gaitskell, H. 1969. *Four Monetary Heretics*. Christchurch: Lyn Christie & Son.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Wise, L. n.d. *Silvio Gesell*. London: Holborn Publishing Co.

How to cite this article

Chick, Victoria. "Gesell, Silvio (1862–1930)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000035> doi:10.1057/9780230226203.0638

The New Palgrave Dictionary of Economics Online

ghettoes

Robert A. Margo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A ghetto is an area of a city in which a minority group is highly segregated and kept there by social, legal, or political forces. This article describes the history of Jewish ghettos of medieval Europe through the Nazi period. The African-American experience is also described, along with the effects of segregation on racial differences in economic outcomes. Examples of ghettos in Japan (the Burakumin), Australia (aborigines), and South Africa (apartheid) are mentioned, along with the related concept of an ethnic enclave.

Keywords

apartheid; Burakumin; ghettoes; housing markets; human capital; immigration; internal migration; neighbours and neighbourhoods; racial discrimination; racial segregation; restrictive covenant; urban segregation

Article

A ghetto is a specific area of a city in which a racial, ethnic, or other minority group constitutes the overwhelming majority of residents, and which is maintained as such by social, political, or legal pressures.
In early medieval Europe local authorities granted Jews living quarters in exchange for services as traders, moneylenders, or tax collectors. With the Crusades, these voluntary arrangements turned compulsory, first in Spain and Portugal and later in Germany. The Jews of early 15th Venice were forced to live in the *ghetto Nuevo*, an abandoned foundry located on an island isolated from the general population (Curiel, Cooperman and Arici, 1990). Walls were erected and guards stood watch by the foundry gates. By day Jews could continue to conduct business but they were required to wear special clothing and were subject to punishment if they were found outside the ghetto at night. Within the walls of the typical Jewish ghetto an entirely separate world of institutions – courts, educational and cultural

institutions, retail establishments, and recreational facilities – emerged. Overcrowding, however, was the norm because geographic expansion at the periphery of the ghetto was severely limited, if it existed at all. By the 19th century the extant Jewish ghettos were seen as outmoded relics of an earlier age. In continental Europe the last ghetto to be abolished was the Roman ghetto in the late 19th century. The so-called Pale of Settlement in the far West of Russia survived until the October Revolution of 1917. A few decades later, Jewish ghettos were revived in German-occupied Europe with a vengeance by the Nazis, who used them primarily as holding pens prior to extermination (Browning, 1986).

In the early 20th century social scientists began to refer loosely to any area of a city in which a particular racial, ethnic, or religious group was very highly segregated as a ‘ghetto’ (Wirth, 1928). Later in the century the usage spread to other types of segregation (such as in the phrase ‘pink ghettos’, referring to the concentration of women in certain occupations). Numerical indices were devised, primarily by sociologists, to measure the extent of segregation (Massey and Denton, 1993; Echenique and Fryer, 2004).

Wacquant (2004) argues that the following are necessary, if not always sufficient, characteristics of a ghetto in the looser sense. First, the minority group must be readily identifiable. Second, the group must be kept physically isolated by the majority, which has the power to enforce the isolation. Third, institutions emerge within the ghetto to provide services that ghetto residents cannot otherwise obtain from the outside world. A distinct ghetto culture may also emerge, elements of which might persist long after the specific mechanisms that kept the affected group ‘in its place’ are abolished.

The word ‘ghetto’ has often been used to describe the experience of African Americans in urban areas in the 20th century (Drake and Clayton, 1945; Clark, 1965; Osofsky, 1971; Hirsch, 1983; Massey and Denton, 1993; Cutler, Glaeser, and Vigdor, 1999). In the late 19th century African Americans were concentrated in the rural southern United States, where they faced very high levels of racial discrimination. In the northern United States the African-American population was a small share of the total and of the urban population; and only in one city (Norfolk, Virginia) was the extent of racial segregation great enough to plausibly justify claims that a ghetto existed (Cutler, Glaeser and Vigdor, 1999).

Although a steady trickle of African Americans to urban areas had occurred since the end of the civil war, the trickle became a flood during the First World War. European immigration from Europe was largely cut off and, buoyed by wartime demand, urban manufacturers turned to southern blacks as a new labour supply. By 1940, virtually all major cities in the North had black ghettos, some known by specific names – ‘Harlem’ in New York, ‘Bronzeville’ in Chicago. Unlike the Jewish ghettos of the medieval Europe, the boundaries of the black ghettos were not set in stone, and whites living nearby adopted a variety of tactics, legal and otherwise, to limit the ghetto's geographic expansion. One example was the *restrictive covenant*, a clause contained in a deed of sale that enjoined a white owner from selling to a black family. Such covenants were in widespread use until rendered unenforceable by the United States Supreme Court in 1948. the Second World War had a much greater effect on black migration from the rural South than the First World War had, and segregation levels continued to increase, reaching ‘staggering levels’ by 1970 (Cutler, Glaeser, and Vigdor, 1999, p. 470). Since 1970 racial segregation in American cities has declined, although levels remain relatively high compared with the late 19th century.

Economists have considered how ghettos affect economic and social outcomes of African Americans. In the model developed by Cutler and Glaeser (1997), there are three groups: skilled whites, skilled blacks,

and unskilled blacks. The groups occupy a city which is divided into three neighbourhoods fixed in geographic size. Economic decisions are made by parents, and parental utility is a positive function of their children's human capital, which itself depends on the human capital of the parent and the average level of human capital in the community where the household resides. Utility is a negative function of housing costs and, in addition, blacks (whites) must pay a fixed entry cost to live in a white (black) neighbourhood.

Cutler and Glaeser consider an initial equilibrium in which whites live in one neighbourhood while skilled and unskilled blacks are distributed across the remaining two neighbourhoods. An increase in the cost to blacks of entering the white neighbourhood reduces the number of skilled blacks living in the white neighbourhood. White utility increases because housing costs fall and average skill levels are unaffected. As more skilled blacks choose to live in one of the black neighbourhoods, housing costs in the neighbourhood rise, which hurts unskilled blacks living there. However, average human capital in the neighbourhood increases, which benefits the human capital production of the children of unskilled blacks. The net effect on the location decisions of unskilled blacks is indeterminate but, depending on the overall effect, it is possible for welfare of unskilled blacks to decrease.

Cutler and Glaeser conduct an empirical analysis of the relationship between segregation and economic outcomes for African Americans, using census data for 1990. They show that an increase in residential segregation reduces educational attainment and income and increases the incidence of motherhood at younger ages. Collins and Margo (2000) replicate the empirical analysis for earlier census years, finding that black ghettos turned increasingly bad on all fronts after 1970. Collins and Margo (2003) study the evolution of racial differences in the values of owner-occupied housing since 1940. They demonstrate that, in the country as a whole, the racial gap in housing values was narrowing prior to 1970. However, in central cities the racial gap widened in the 1970s, and the extent of widening was greater in cities that were initially heavily segregated in 1970.

If high levels of racial segregation turned increasingly bad for African Americans in central cities after 1970, there was no shortage of possible causes. Massey and Denton (1993) emphasize deindustrialization, whereas Wilson (1987) hypothesizes that decreases in housing market discrimination enabled middle- and upper-class Americans to leave central cities ghettos, thereby removing effective role models from the ghetto. Although the United States has a long (and terrible) history of race-related civil disturbances, the number and geographic extent of those occurring in the 1960s were unprecedented. Collins and Margo (2004a; 2004b) demonstrate that the occurrence of a severe riot between 1960 and 1970 was associated with deteriorating employment, earnings, and housing values for urban blacks, an association that persisted at least through the 1970s.

European Jews and African Americans are far from the only groups whose histories include ghettos or ghetto-like conditions. One such group is the Burakumin of Japan. The Burakumin were descendants of the lowest of four castes that existed during the feudal era of Japanese history. Viewed as ‘untouchable’ by the religious majority (Buddhists and Shinto), the Burakumin were faced with similar constraints to those imposed on Jews, such as restrictions on intermarriage and geographic mobility. Like Jews, the Burakumin had to wear special clothing and behave in a deferential manner towards the majority. Although ‘emancipated’ by edict in 1871, a combination of formal and informal institutions nevertheless resulted in the formation of numerous impoverished and crime-ridden ghettos of Burakumin throughout Japanese cities that have persisted (DeVos and Wagatsuma, 1966; Hane, 1982). Brock (1993) describes

the emergence, persistence, and functioning of three ‘outback’ ghettos (Poonindle, Koonidule, and Nepabunna) of Aboriginal people in Australia. Residential segregation was a key aspect of apartheid in South Africa as well as other colonial regimes (Abu-Lughod, 1980; Western, 1981). Large-scale international migrations have frequently resulted in ghetto-like conditions in the receiving countries (see Wirth, 1928, for a classic discussion). Such ‘ethnic enclaves’ may acquire names – ‘Chinatown’ (see Zhou, 1992) or ‘Little Italy’ – but are fundamentally different because the newcomers eventually assimilate into the broader society. The techniques developed for measuring racial segregation in the United States are now being routinely applied to other countries; see, for example, Peach (1996), and also Johnston, Forest, and Poulsen (2002), who shows that overall levels of ethnic segregation in British cities around 1990 were considerably lower than levels of racial segregation in American cities.

See Also

- housing policy in the United States
- residential segregation

Bibliography

Abu-Lughod, J. 1980. *Rabat: Urban Apartheid in Morocco*. Princeton, NJ: Princeton University Press.

Brock, P. 1993. *Outback Ghettos: A History of Aboriginal Institutionalism and Survival*. New York: Cambridge University Press.

Browning, C. 1986. Nazi ghettoization policy in Poland, 1939–1941. *Central European History* 19, 343–68.

Clark, K. 1965. *Dark Ghetto: Dilemmas of Social Power*. New York: Harper.

Collins, W. and Margo, R. 2000. Residential segregation and socioeconomic outcomes: when did ghettos go bad? *Economics Letters* 69, 239–43.

Collins, W. and Margo, R. 2003. Race and the value of owner-occupied housing, 1940–1990. *Regional Science and Urban Economics* 33, 255–86.

Collins, W. and Margo,R. 2004a. The Labor Market Effects of the 1960s Riots. In *Brookings-Wharton Papers on Urban Affairs 2004*, ed. W. Gale and J. Pack. Washington DC: Brookings Press.

Collins, W. and Margo, R. 2004b. The Economic Impact of the 1960s Riots: Evidence from Property Values. Working Paper No. 10493. Cambridge, MA: National Bureau of Economic Research.

Curiel, R., Cooperman, R. and Arici, G. 1990. *The Venetian Ghetto*. New York: Rizzoli.

Cutler, D. and Glaeser, E. 1997. Are ghettos good or bad? *Quarterly Journal of Economics* 112, 827–72.

Cutler, D., Glaeser, E. and Vigdor, J. 1999. The rise and decline of the American ghetto. *Journal of Political Economy* 107, 455–506.

DeVos, G. and Wagatsuma, H., eds. 1966. *Japan's Invisible Race: Caste in Culture and Personality*. Berkeley: University of California Press.

Drake, S. and Clayton, H. 1945. *Black Metropolis: A Study of Negro Life in a Northern City*. Chicago: University of Chicago Press.

Echenique, F. and Fryer, R. 2004. On the measurement of segregation. Working paper. Cambridge, MA: Department of Economics, Harvard University.

Hane, M. 1982. *Peasants, Rebels, and Outcasts: The Underside of Modern Japan*. New York: Pantheon.

Hirsch, A. 1983. Making the Second Ghetto: Race and Housing in Chicago, 1940–1960. New York: Cambridge University Press.

Johnston, R., Forrest, J. and Poulson, M. 2002. Are there ethnic enclaves/ghettos in English cities? *Urban Studies* 39, 591–618.

Massey, D. and Denton, N. 1993. American Apartheid: Segregation and the Making of the Underclass. Cambridge, MA: Harvard University Press.

Osofsky, G. 1971. Harlem: The Making of a Ghetto – Negro New York, 1890–1930. New York: Harper and Row.

Peach, C. 1996. Does Britain have ghettos? Transactions of the Institute of British Geographers 21, 216–35.

Wacquant, L. 2004. Ghetto. In International Encyclopedia of Social and Behavioral Sciences, ed. N. Smelser and P. Baltes. London: Pergamon Press.

Western, J. 1981. Outcast Cape Town. Minneapolis: University of Minnesota Press.

Wilson, W. 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. Chicago: University of Chicago Press.

Wirth, L. 1928. The Ghetto. Chicago: University of Chicago Press.

Zhou, M. 1992. Chinatown: The Socioeconomic Potential of an Urban Enclave. Philadelphia: Temple University Press.

How to cite this article

Margo, Robert A. "ghettoes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000183> doi:10.1057/9780230226203.0639

The New Palgrave Dictionary of Economics Online

Gibrat, Robert Pierre Louis (1904 – 1980)

David E.R. Gay

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Gibrat, R. P.L.; Gibrat's Law; Econometric Society

Article

Gibrat was born on 23 March 1904 in Lorient, France, and died on 13 May 1980 in Paris. He studied at Saint Louis de Paris, as well as in Rennes, Lorient and Brest, and in 1922 he entered the Ecole Polytechnique to become a mining engineer. He received a bachelor's degree in science and a doctorate in law from the University of Paris. He was a technical consultant in private firms before being named director of electricity in the Ministry of Public Works, 1940–42. He became Secretary of State for Communications under the Laval government but resigned after the Allied invasion of North Africa. After the Liberation he was chief engineer of mines. He was consulting engineer for French Electric on tidal energy (1945–68) and served as Director General for atomic energy (Indatom, 1955–74), president of the scientific and technical committee of Euratom (1962), and as a consulting engineer for Central Thermique, 1942–80. He taught at Ecole des Mines from 1936 to 1968. He served as President of the French Society of Electricians, Vice President and President of the Civil Engineers of France, President of the Statistical Society of Paris (1966), President of the French Statistical Society, President and Honorary President of the Technical Committee for the Hydrotechnical Society of France, President of the French Meteorological Society (1969), Honorary President of the World Federation of Organizations of Engineers, and President of the French Section of the American Nuclear Society. Gibrat was the author of reports to the Academy of Sciences, some 100 professional articles and two books (on economics and tidal energy), and a Knight of the Legion of Honour.
His major contribution to economics is known as Gibrat's Law. This states that the expected growth rate for a firm is independent of its size. Gibrat's Law has been successfully tested by French and American investigators, among others. His famous economics work, *Les inégalités économiques*, was published in 1931. For his contributions to economics and mathematical economics he was elected a Fellow of the Econometric Society in 1948.

Selected works

1930. Une loi des répartitions économiques: l'effet proportionnel. *Bulletin de la Statistique générale de la France et du Service d'observations des prix* 19, July–September, 469.

1931a. *Les inégalités économiques*. Paris: Sirey.

1931b. Les inégalités économiques. *Revue de l'industrie minérale*, 15 July.

1935. La science économique. Méthodes et philosophie. *Actualités scientifiques et industrielles*, Paris.

How to cite this article

Gay, David E. R. "Gibrat, Robert Pierre Louis (1904–1980)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000037> doi:10.1057/9780230226203.0640

The New Palgrave Dictionary of Economics Online

Gibrat's Law

José Mata

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Gibrat's law states that a great deal of the evolution of firm size distribution over time is due to the action of chance. While chance is still seen as an important driver of the size of firms, recent studies call for chance to be supplemented by some more structured models in order to explain the observed patterns of firm size evolution.

Keywords

concentration; firm growth; firm size; firm size distribution; Gibrat's law; heteroskedasticity

Article

In one of the first studies on firm size distribution (FSD) Gibrat (1931) observed that the size of firms followed the lognormal distribution very closely, from which he concluded that firms' rate of growth ought to be a random process. In particular, he reasoned that growth should not depend on the initial size of firms, as such a process would inevitably produce a lognormal distribution. This assertion became known as Gibrat's law.

If we denote the size of the firm at time t by S_t and the proportional growth between t and $t-1$ by ε_t

$$(S_t - S_{t-1}) / S_{t-1} = \varepsilon_t$$

$$S_t = S_{t-1}(1 + \varepsilon_t) = S_0(1 + \varepsilon_1)(1 + \varepsilon_2)\ldots(1 + \varepsilon_t),$$

taking logs and using the approximation $\log(1+\varepsilon_t) \approx \varepsilon_t$ leads to

$$\log S_t = \log S_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t.$$

As $t \to \infty$, $\log S_0$ becomes less important relative to $\log S_t$ and, if ε_t is drawn from a normal distribution with mean μ and variance σ^2, $\log S_t$ can be approximated by a normal distribution with mean μt and variance $\sigma^2 t$. The result that the variance of the FSD is bound to increase over time due to the sole action of chance is probably responsible for the popularity of Gibrat's law in industrial organization, as it provides a nice explanation for the observed empirical regularity that industrial concentration increased over time. Other fields in which the application of Gibrat's law has been discussed include those of distributions of income (Sahota, 1978) and the size of cities (Eeckhout, 2004).

It is not surprising that soon after the publication of Gibrat's book, different studies tried to test the law empirically (see Sutton, 1997, for a survey). Some proceeded to compare the observed firm size distribution with the lognormal distribution, others analysed the relationship between firm size and firm growth. By and large, the results coincided. The firm size distribution seemed to conform well to the lognormal distribution, and firm growth seemed to be largely independent of firm size.

The studies in this first wave typically used data that were readily available in public sources, that is, data on the largest firms in the economy. In an influential study published in the early 1960s, however, Mansfield (1962) collected data on ‘practically all firms’ in three American industries in different time periods and analysed the relationship between the size and the growth of firms. He suggested that different interpretations regarding the extent to which Gibrat's law was applicable were possible, and tested the validity of the law according to these different interpretations. According to the first interpretation Gibrat's law would hold for all firms including those that exit; with this interpretation, a negative relationship between initial size and growth was discovered in the majority of the samples that were considered. The second interpretation posited that the law would hold only for those firms that had not exited; a significant relationship showed up only in a minority of samples. The third interpretation stated that the law would hold only for those firms whose size exceeded a given threshold. Setting this threshold as the minimum efficient scale in the industry, Mansfield found no significant relationship in any of the samples. Yet, even with this restricted interpretation, Mansfield's samples failed to pass a second test of Gibrat's law, that the variance of growth would be independent of size.

The topic did not attract much attention during the rest of 1960s and the 1970s. When it again came under scrutiny in the mid- to late 1980s, new data sources had became available. These new data sources covered many more firms than before, providing a much better coverage of the smallest firms in the economy. Furthermore, their longitudinal dimension, which allowed researchers to follow firms over time, led to the discovery that the entry, exit and growth movements that were taking place in most industries of developed countries were of a previously unsuspected large magnitude (see Caves, 1998,

for a survey).

Concerning Gibrat's law, the major consequence was that attention was this time mostly drawn to the relationship between firm growth and size, and problems of sample selection became routinely addressed using econometric techniques that had meanwhile become available. Whether or not sample selection was taken into account, however, the results of the studies using these recently developed databases suggested that firm growth was not independent of firm size, smaller firms growing faster than their larger counterparts (for example, Evans, 1987). Another systematic concern of this literature was heteroskedasticity: most studies found that large firms display a less variable pattern of growth than do smaller units. Although consistent with the idea that the diversification associated with size reduces risk, this is a pattern that does not conform well to Gibrat's postulate.

A negative relationship between size and growth does not imply that concentration does not increase. This point can be clearly seen by imagining that there are firms of only two sizes, small and large, present in equal numbers in an economy. If those firms that are small in one period become large in the next period and vice versa, the overall distribution remains constant despite the obvious relationship between growth and size. There have been few studies in this new wave that have specifically examined the FSD. McCloughan (1995) simulated the effect of different violations of Gibrat's postulates upon the development of market structure and concluded that the nature of the size–growth relationship (in contrast to the effect considering entry and exit) was the most important determinant of the evolution of concentration.

In one of the few studies that have used these new comprehensive data-sets to analyse the actual development of the FSD, Cabral and Mata (2003) found that the FSD is considerably more skewed than the lognormal in the earliest years, but gradually approaches it as firms get older. Convergence to the lognormal is what would be expected from a Gibrat process, and the fact that the lognormal is approached from a more skewed distribution may seem to be unimportant from the strict standpoint of Gibrat's law, as the law posits that the starting point does not matter in the long run. The finding, however, creates an additional challenge: if we are to rely on random forces to explain the evolution of the FSD, what can possibly explain its starting position? How can this be part of a theory of the evolution of firm size, and how can it coexist with models such as Jovanovic's (1982), in which skewedness emerges gradually as firms learn about their abilities? One possibility is that the size of firms at start-up is the minimum of two sizes: a size to be achieved in the long run – determined by the ability of the entrepreneur in the spirit of Lucas (1978) – and a short-run size, given by some constraint. Cabral and Mata suggested that this constraint could to be a financial one, but other constraints might do the job as well.

An unaddressed question is the extent to which the FSD converges to a position that depends on a pre-existing distribution of abilities and to what extent abilities are a product of explicit decisions made by firms as to the learning process (Ericson and Pakes, 1995). Another question pertains to the appropriate level of analysis. Machado and Mata (2000) report that failure to control for industry-specific conditions leads to a significantly greater departure from the lognormal than when these conditions are controlled for. It is also not obvious that the FSD should be governed by the same forces and evolve along the same lines irrespectively of the specific competitive conditions in the industry (Sutton, 1997), but little work has been done on how these conditions affect the FSD. Perhaps the streams of the literature that has given more attention to the evolution of industries is that following the work by Klepper and Graddy

(1990), which shows that, over their life cycle, industries exhibit significant variation, namely, with respect to the changes in the number of firms and their patterns of growth. The implications of these changes to the evolution of the FSD of industries are still under-explored.

See Also

- firm-level employment dynamics
- growth and learning-by-doing
- lognormal distribution

Bibliography

Cabral, L. and Mata, J. 2003. On the evolution of the firm size distribution: facts and theory. *American Economic Review* 93, 1075–90.

Caves, R.E. 1998. Industrial organization and new findings on the turnover and mobility of firms. *Journal of Economic Literature* 36, 1947–82.

Eeckhout, J. 2004. Gibrat's law for (all) cities. *American Economic Review* 94, 1429–51.

Ericson, R. and Pakes, A. 1995. Markov-perfect industry dynamics: a framework for empirical work. *Review of Economic Studies* 62, 53–82.

Evans, D. 1987. Tests of alternative theories of firm growth. *Journal of Political Economy* 95, 657–74.

Gibrat, R. 1931. *Les Inegalités economiques: applications aux inégalités des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle: la loi de l'effect proportionnel*. Paris: Sirey.

Jovanovic, B. 1982. Selection and evolution of industry. *Econometrica* 50, 649–70.

Klepper, S. and Graddy, E. 1990. The evolution of new industries and the determinants of market structure. *Rand Journal of Economics* 21, 27–44.

Lucas, R.E. 1978. Size distribution of business firms. *Bell Journal of Economics* 9, 508–23.

Machado, J. and Mata, J. 2000. Box-Cox quantile regression and the distribution of firm sizes. *Journal of Applied Econometrics* 15, 253–74.

Mansfield, E. 1962. Entry, Gibrat's law, innovation, and the growth of firms. *American Economic*

Review 52, 1023–51.

McCloughan, P. 1995. Modified Gibrat growth, entry, exit and concentration development. *Journal of Industrial Economics* 43, 405–33.

Sahota, G.S. 1978. Theories of personal income distribution: a survey. *Journal of Economic Literature* 16, 1–55.

Sutton, J. 1997. Gibrat's legacy. *Journal of Economic Literature* 35, 40–59.

How to cite this article

Mata, José. "Gibrat's Law." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000038> doi:10.1057/9780230226203.0641

The New Palgrave Dictionary of Economics Online

Giffen, Robert (1837 – 1910)

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Giffen, R.; measurement; Royal Economic Society; Royal Statistical Society; statistical comparisons

Article

Robert Giffen's name seems likely to be known by students of economics for generations to come in relation to the famous result in the theory of consumer demand which bears his name but about which, so far as can be determined, he had nothing to say. Marshall originated the tradition when he associated the result with Giffen's name in the third edition of his *Principles* in 1895 (p. 208).

Giffen was born in Lanarkshire. At the age of 13 he was apprenticed to a solicitor in Strathaven, and continued in the same vocation until 1860 (though during the last seven years of this period he resided and worked in Glasgow). Still only 23 years old, Giffen struck out on a career in journalism – in which he was to be successful in establishing his reputation in economic circles of the day. He begun as a sub-editor for the *Stirling Journal*, moved to London in 1862 to work at the *Globe*, transferred to the *Fortnightly Review* in 1866, and in 1868 became assistant editor at *The Economist* – a post at which he remained until his next change of vocation in 1876. He was also city editor at the *Daily News* between 1873 and 1876. Giffen's third and final career was as a professional civil servant, first as chief of the statistical department at the Board of Trade, and then in 1882 as its Assistant Secretary. He retired from the civil service at the age of 60. Giffen served on numerous royal commissions (including the Gold and Silver Commission of 1886–8); he was editor of the *Journal of the Royal Statistical Society* (1876–91), President of that society (1882–4), twice presided over the economics section of the British Association (1887 and 1901), and was one of the founders of the Royal Economic Society. In short, he was one of those figures encountered frequently in British economics whose not inconsiderable power and prestige appears to be disproportionate to their actual contribution to economic science.

In so far as he was primarily a statistician, Giffen's work did attempt to alert economists to the dangers of theory without measurement. His presidential address to the Royal Statistical Society in 1882 was devoted to the subject, and in 1901 as president of Section F of the British Association (his second term in that office) he returned to the same theme (see 1904, vol. 2, chs 13 and 28). Indeed, according to

Higgs in his edition of this *Dictionary* (1925), Giffen's statistical prowess was one of the factors which helped to secure the respect of theorists. His article on international statistical comparisons in the *Economic Journal* (1892b), for example, can be singled out for special mention since it treats for the first time a problem which has still not been adequately resolved. Of course, it was not always the case that Giffen's careful mustering of the statistical evidence allowed him, any more than the theorists, to avoid the pitfalls of making predictions which subsequent experience has proven to be silly – witness his claim that the whole protectionist school would die out within a decade (1898, p. 16).
However, in the final analysis it is in Giffen's attempts to provide reasonably accurate measurements of indicators like wage rates, economic growth (see 1884), and national product (1889) that one should isolate his main contribution. While it is true that subsequent work in this field has advanced well beyond Giffen's early efforts, he remains one of the pioneers of applied economics in its modern sense.
It seems that Giffen was also a strong supporter of a Channel tunnel: not for one between England and France, but between Ireland and England. He died on 12 April 1910 and is buried in Strathaven.

Selected works

1872. (With B. Cracroft.) *American Railways as Investment*. London.

1873. The production and movement of gold since 1848. In Giffen (1880).

1877. *Stock Exchange Securities*. London: G. Bell & Sons.

1880. *Essays in Finance*. First Series, vol. 1. London: G. Bell & Sons.

1884. *The Progress of the Working Class in the Last Half Century*. London: G. Bell & Sons.

1886. *Essays in Finance*. Second Series, vol. 2. London: G. Bell & Sons.

1887. The recent rate of material progress in England. Address as President of Section F of the British Association. *Journal of the Royal Statistical Society* 50, 615–47. Reprinted in Giffen (1904), vol. 2.

1889. *The Growth of Capital*. London: G. Bell & Sons.

1892a. *The Case Against Bimetallism*. London: G. Bell & Sons.

1892b. On international statistical comparisons. *Economic Journal* 2(June), 209–38.

1898. Protection for manufactures in new countries. *Economic Journal* 8(March), 3–16.

1904. *Economic Inquiries and Studies*, 2 vols. London: G. Bell & Sons.

How to cite this article

Milgate, Murray. "Giffen, Robert (1837–1910)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000041> doi:10.1057/9780230226203.0642

The New Palgrave Dictionary of Economics Online

Giffen's paradox

John Nachbar
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

comparative statics; general equilibrium; Giffen good; Giffen preferences; Giffen, R.; Giffen's paradox; income effect; law of demand; preference externalities; revealed preference; Slutsky equation; substitution effect

Article

Giffen's paradox refers to the possibility that standard competitive demand, with nominal wealth held constant, can be upward sloping, violating the law of demand. From the Slutsky equation, Giffen's paradox arises if and only if a good is inferior and the income effect is larger than the absolute value of the substitution effect. A *Giffen good* is a good for which Giffen's paradox can arise. *Giffen preferences* are preferences that can exhibit Giffen's paradox. For explicit examples of Giffen preferences, see Moffatt (2002) and Sorensen (2005).

The term ‘Giffen's paradox’ originates in a passage in Marshall (1920), which credits the statistician Robert Giffen (1837–1910) with observing a failure of the law of demand in the market for bread. The widespread association of Giffen's paradox with potatoes during the Irish potato famine of the 1840s may have originated in Samuelson (1964). A number of authors have since argued that potatoes were not, in fact, a Giffen good for Irish potato farmers (for example, Dwyer and Lindsay, 1984; Rosen, 1999). For more on the tortured intellectual history of Giffen's paradox, see Walker (1987). For empirical evidence that Giffen goods may exist, see Baruch and Kannai (2001) and Jensen and Miller (2002).

In thinking about Giffen's paradox, bear in mind four points. First, the budget constraint forces a crude form of compliance with the law of demand: as the price of a good goes to infinity, consumption of that good must go to zero. Conversely, under standard assumptions on preferences (such as monotonicity), as the price of a good goes to zero, demand for the good becomes large. Thus, under standard assumptions, the graph of demand for a Giffen good, with price on the vertical axis, is roughly Z-shaped.

Second, whether Giffen's paradox arises for aggregate demand, summed across consumers, depends on

the wealth distribution as well as on individual preferences. This point is a variation on the preceding one. For any given consumer, if prices are held fixed, consumption of any good must fall for large enough drops in nominal wealth, which implies that no good is inferior for all wealth levels. Therefore, even if consumers have the same Giffen preferences, if consumers have different wealth levels, then, over any given price interval, the law of demand may hold for some consumers even if it is violated for others. As a consequence, even if all consumers have the same Giffen preferences, aggregate demand may obey the law of demand. A striking example is due to Hildenbrand (1983): if there is a continuum of consumers all of whom have the same preferences and if nominal wealth is uniformly distributed on an interval containing zero, then aggregate demand cannot exhibit Giffen's paradox.

Third, in a general equilibrium (GE) model, with endogenous wealth, the comparative statics of Giffen goods can run counter to standard textbook intuition. Consider an exchange economy (no production) with two goods and only one consumer. In the equilibrium of this trivial economy, the consumer eats her own endowment. Equilibrium relative prices (which are well defined even though there is no trade) are given by the slope of the consumer's indifference curve through her consumption/endowment point. Naive textbook supply and demand analysis predicts that, if demand for good 1 is upward sloping, then an increase in the endowment of good 1 results in a higher equilibrium price (if we assume that the price of good 2 is constant). One can easily show, however, that in this economy an increase in the endowment of good 1 can increase its equilibrium price only if the good is *normal*, hence not Giffen. In fact, if the good is Giffen then the endowment increase causes the price to fall by so much that nominal wealth falls (causing the demand curve to shift *out*, since the good is inferior), even though the endowment increase makes the consumer better off.

The critical feature of this example is that individual demand automatically obeys the weak axiom of revealed preference, and in exchange economies the weak axiom implies the law of demand for *endogenous* wealth demand near equilibrium. Thus, if preferences are Giffen, endogenous wealth demand slopes down at equilibrium even though fixed wealth demand slopes up, and it is endogenous wealth demand that drives GE comparative statics. This analysis generalizes to multi-consumer economies, with active trade, provided aggregate, endogenous wealth demand satisfies the weak axiom (see Nachbar, 2002); the basic intuition goes back to Hicks (1939). Note, however, that in multi-consumer economies aggregate demand does not automatically satisfy the weak axiom.

For other work that investigates the behavior of Giffen goods in environments richer than those usually considered in textbook treatments, see Barzel and Suen (1992) and Rosen (1999).

Finally, in standard economics usage, Giffen's paradox refers both to a phenomenon – failure of the law of demand for standard competitive demand – and to a particular mechanism underpinning that phenomenon – income effects. Giffen's paradox is, however, sometimes conflated with similar phenomena arising from quite different mechanisms, often based on preference externalities of some form. A classic citation is Leibenstein (1950). For an interesting application, see Pesendorfer (1995).

See Also

- comparative statics
- law of demand

Bibliography

Baruch, S. and Kannai, Y. 2001. Inferior goods, Giffen goods, and shochu. In *Economics Essays: A Festschrift for Werner Hildenbrand*, ed. G. Debreu, W. Neuefeind and W. Trockel. Heidelberg: Springer Verlag.

Barzel, Y. and Suen, W. 1992. The demand curves for Giffen goods are downward sloping. *The Economic Journal* 102, 896–905.

Dwyer, Jr., G. and Lindsay, C. 1984. Robert Giffen and the Irish potato. *American Economic Review* 74, 188–92.

Hicks, J. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51, 997–1018.

Jensen, R. and Miller, N. 2002. Giffen behavior: theory and evidence. Working Paper No. RWP02-014. Kennedy School of Government, Harvard University.

Leibenstein, H. 1950. Bandwagon, snob, and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64, 283–07.

Marshall, A. 1920. *Principles of Economics*. 8th edn. London: Macmillan.

Moffatt, P. 2002. Is Giffen behaviour compatible with the axioms of consumer theory? *Journal of Mathematical Economics* 27, 259–67.

Nachbar, J. 2002. General equilibrium comparative statics. *Econometrica* 70, 2065–74.

Pesendorfer, W. 1995. Design innovation and fashion cycles. *American Economic Review* 85, 771–92.

Rosen, S. 1999. Potato paradoxes. *Journal of Political Economy* 107, 294–313.

Samuelson, P. 1964. *Economics*, 5th edn. New York: McGraw-Hill.

Sorensen, P. 2005. Simple utility functions with Giffen demand. Working paper. Department of Economics, University of Copenhagen.

Walker, D. 1987. Giffen's paradox. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate, and P. Newman. Basingstoke: Palgrave.

How to cite this article

Nachbar, John. "Giffen's paradox." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000042> doi:10.1057/9780230226203.0643

The New Palgrave Dictionary of Economics Online

Gilman, Charlotte Perkins (1860 – 1935)

B. Berch

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Gilman, C. P.; women's work and wages

Article

Gilman was born on 3 July 1860 in Hartford, Connecticut, and died on 17 August 1935 in Pasadena, California. Known worldwide as a feminist theorist and a generally iconoclastic social critic, Gilman was a major intellectual force in America at the turn of the 20th century. Largely self-educated, problems with her first marriage led her to separate from her first husband and begin an unconventional freelance life based in California, earning her living from her lecturing and writing. *Women and Economics* (1898) was her first book-length exposition of her theory of the evolution of gender relations. Influenced by the ideas of Edward Bellamy, Lester Frank Ward, Darwin, the Webbs and G. Bernard Shaw, she explained that human institutions (like the species itself) has evolved over time, favouring the survival of the best adapted. A major exception, however, was the definition of ‘women's place’. Here social development had been frozen by Tradition. Women were confined to households which were no longer the locus of any socially productive activity, since now the factory produced the needed consumption goods, and children were better raised in schools, by professionals. The role of full-time housewife and mother had become anachronistic, reducing women to the state of social parasites. As she also argued in her 1903 classic *The Home* that, for their own progress and for the progress of human civilization overall, women would have to leave these domestic prisons and take up socially useful work in the larger world of production. In the articles and didactic fiction that she wrote for her monthly magazine the *Forerunner*, she developed a wide range of startlingly rational ideas for social reorganization.

Selected works

1892. The yellow wall-paper. *New England Magazine*. January. Reprinted in *The Charlotte Perkins Gilman Reader*, ed. A. Lane, New York: Pantheon, 1980.

1898. *Women and Economics*, ed. C. Degler, New York: Harper & Row, 1966.

1903. *The Home*. Introduction by W. O'Neill, Urbana: University of Illinois Press, 1972.

1915. Herland. *The Forerunner* 6. Reprinted, with introduction by A. Lane, as *Herland: A Lost Feminist Utopian Novel*, New York: Pantheon, 1978.

How to cite this article

Berch, B. "Gilman, Charlotte Perkins (1860–1935)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000045> doi:10.1057/9780230226203.0644

The New Palgrave Dictionary of Economics Online

Gini ratio

Camilo Dagum
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

In a national economy, the price system determines both resource allocation and the income distribution. The imputation to the factors of production of the mass of income associated with an economy's output determines its distribution by factor shares, or *functional* income distribution. This mainstream of research follows Ricardo's (1817) contribution. Another mainstream of research was initiated by Pareto (1895, 1897), and deals with the distribution of a mass of income among the members of a set of economic units (family, household, individual), considering either the total income of each economic unit or its disaggregation by source of income, such as wages and salaries, property income, self-employment income, transfers, etc. This type of inquiry deals with distribution by size of income, or *personal* income distribution, and the quantitative assessment of the relative degree of income inequality among the members of a given set of economic units. Such inquiries provide basic quantitative information in support of a comprehensive research strategy on income distributions, including causal explanations for social welfare and policy.

It is of interest to remark that Pareto's research on income distribution was motivated by the polemic he engaged in with French and Italian socialists concerning the ways and means of achieving a less unequal distribution. Thus, the actual *measurement* of inequality was brought to the fore, with its main purposes the assessment of (i) the evolution of inequality in a given country or region, and (ii) the relative degree of inequality between countries or regions.

In a series of methodological and applied contributions Corrado Gini (1955) enriched this field of research. In 1910 he corrected the interpretation of Pareto's inequality parameter and, in 1912, proposed a new measure of income inequality, the Gini ratio.

Pareto (1896, 1897) specified three versions of this model of income distribution. The most widely used model is Pareto Type I

$$S(x) = 1 - F(x) = (x / x_0)^{-\alpha},\ 0 < x_0 < x,\ \alpha > 1, \tag{1}$$

where $S(x)=P(X>x)$ is the survival distribution function (SDF) of the income variable X, $F(x)$ is the cumulative distribution function (CDF), x_0 is the minimum value of X, α is a scale-free inequality parameter, and the mathematical expectation of income is

$$\mu = E(X) = \alpha x_0^{\alpha} \int_{x_0}^{\infty} x^{-\alpha} dx = \alpha x_0 / (\alpha - 1). \tag{2}$$

Pareto seems to have assumed that income growth implies less income inequality. This assumption, together with eqn (2), led him to the conclusion that income inequality is an

increasing function of α. Gini (1910) reversed this interpretation, proving that, given model (1), income inequality is a decreasing function of α. Gini's rationale was as follows: given n units with incomes $x_1 \leq x_2 \leq x_3 \bullet, \ldots, \bullet \leq x_n$, the average of the last $m (m \leq n)$ income units $\Sigma_{i=0}^{m-1} x_{n-i} / m$ is greater than or equal to the average income $\mu = \Sigma_{i=0}^{n} x_i / n$ of the population, hence, there exists a $\delta \geq 1$ such that

$$\left(\sum_{i=0}^{m-1} x_{n-i} / \sum_{i=1}^{n} x_i \right)^{\delta} = m / n, \ \delta \geq 1, \tag{3}$$

Equation (3) is known as the Gini model. Gini (1910) interpreted the scale-free parameter δ as a measure of income inequality and called it a *concentration ratio* because it is an increasing function of the concentration of income in the upper income groups. For this reason, Gini called eqn (3) a *concentration curve*, where the abscissa represents the CDF $F(x_m) = m/n$ and the ordinate the income share $\Sigma_{i=1}^{m} x_i / \Sigma_{i=1}^{n} x_i$, $m = 1, 2, \ldots, n$, δ being an unknown parameter that has to be estimated.
Using the CDF $F(x)$ and the Lorenz curve $L(x)$ (also called the Lorenz-Gini curve since it was independently introduced by both authors), eqn (3) takes the form

$$1 - F(x) = [1 - L(x)]^{\delta}, \ \delta \geq 1, \tag{4}$$

where

$$L(y) = (1 / \mu) \int_0^y x \mathrm{d}F(x) . \tag{5}$$

Replacing $F(x)$ from model (1) into eqns (4) and (5), Gini (1910) proved that $\delta = \alpha / (\alpha - 1)$ and thus reversed Pareto's interpretation of α. In fact, when $\alpha \to \infty$, $\delta \to 1$ and $F(x) = L(x)$, and the mass of income is equally distributed.
Gini (1912) specified the Gini mean difference with and without replacement. The latter is by definition

$$\Delta = \sum_{j=1}^{n} \sum_{i=1}^{n} |x_j - x_i| / n(n-1), \ 0 \leq \Delta \leq 2\mu, \tag{6}$$

and using the Riemann–Stieltjes integral, which covers, as particular cases, both discrete and continuous distributions, we have

$$\Delta = \int_0^\infty \int_0^\infty |y - x| dF(x) dF(y), \tag{7}$$

where X and Y are identically and independently distributed variables. When $x_1=x_2=\ldots=x_n$, $\Delta=0$, and when $x_1=x_2=\ldots=x_{x-1}=0$ and $x_n=n\mu$ (the total income), $\Delta=2\mu$. Since Δ is a monotonic increasing function of the degree of income inequality, Gini (1912) specified

$$G = \Delta / 2\mu,\ 0 \le G \le 1 \tag{8}$$

as an income inequality measure. Equation (8) is known as the Gini ratio or Gini index and it is widely used in theoretical and applied research on income and wealth distributions. Gini (1914) proved the important theorem that $G=\Delta/2\mu$ is equal to twice the area between the equidistribution line $F(x)=L(x)$ and the Lorenz curve $L(x)$ (see Figure 1). Moreover,

$$G = \Delta / 2\mu = 2\int_0^1 (F - L)\, dF = (2/\mu)\int_0^\infty x\left[F(x) - \frac{1}{2}\right] dF(x) = (2/\mu)\int_0^\infty x\left[\frac{1}{2} - S(x)\right] dF(x). \tag{9}$$

Figure 1
Lorenz curve *L(x)* and Gini ratio *G*.

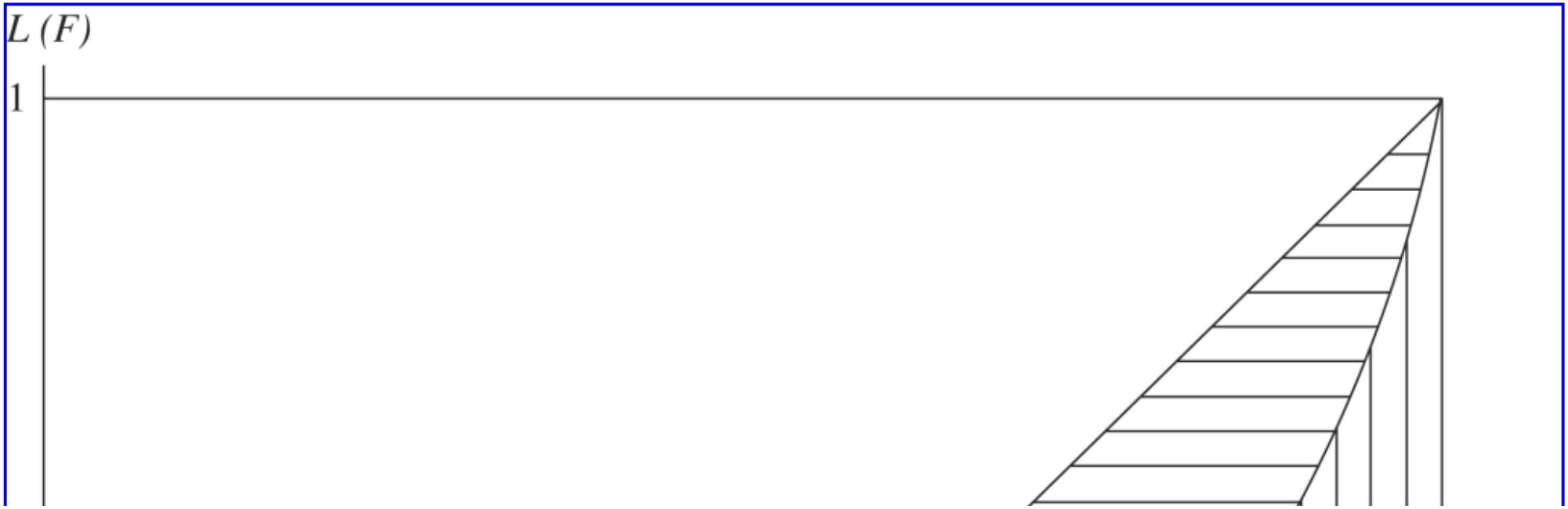

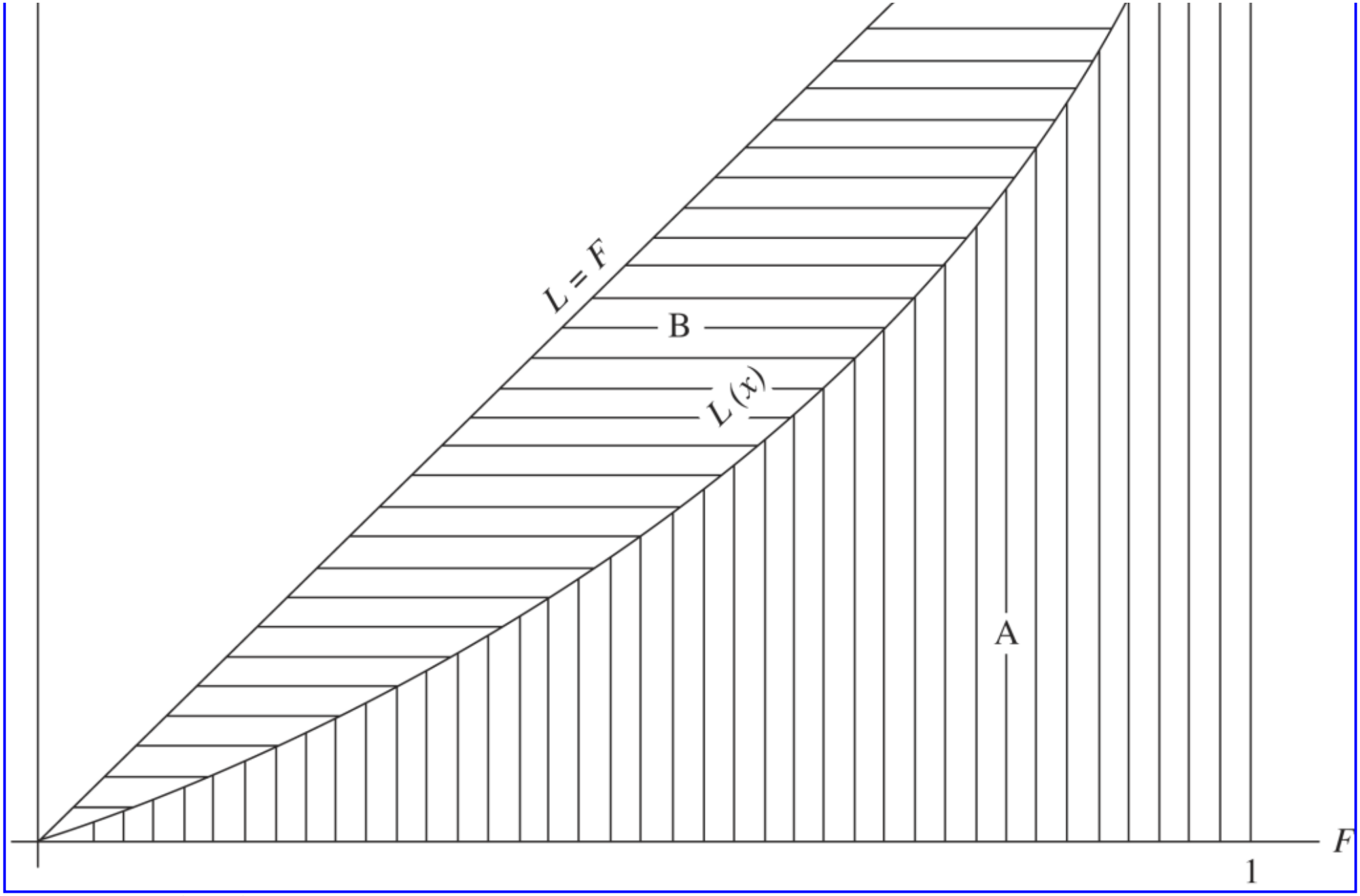

For the discrete case, it follows from eqns (6) and (8), that

$$(F(x)) = (1/\mu)\int_0^x y\,dF(y)\,G = 2B = 1 - 2A = 1 - 1\int_0^1 L\,dF$$

$$G = [2 / n(n-1)\mu] \sum_{k=1}^{n} k x_k - (n+1) / (n-1) = (n+1)(n-1) - [2 / n(n-1)\mu] \sum_{k=1}^{n} (n-k+1) x_k, \tag{10}$$

showing that the welfare function underlying the Gini ratio is a rank-order-weighted sum of the economic units' income shares.

The properties that an income inequality measure must fulfil were first discussed by Dalton (1920). It can be shown (Dagum, 1983, pp. 34–5) that G fulfils the properties of (i) transfer, (ii) proportional addition to incomes, (iii) equal addition to incomes, (iv) proportional addition to persons, (v) symmetry, (vi) normalization, and (vii) operationality.

The Gini ratio is sensitive to transfers to all income levels. In fact, it follows from eqn (10), that a transfer of h dollars from the richer j to the poorer i theorem, without modifying their income ranks, is

$$\Delta G(j, i, h) = -2(j-i)h / n(n-1)\mu > 0, \quad j > i, \tag{11}$$

therefore $-\Delta G$ is an increasing function of $j-i=F(x_j)-F(x_i)$ and a decreasing function of both n and μ. The maximum reduction of G is achieved when $h=(x_j-x_i)/2$, and is not necessarily given by eqn (11) unless the transfer fulfils certain conditions with respect to the original income ranking of the population.

Often, the Gini ratio is misinterpreted when it is incorrectly claimed that it attaches more weight to transfers to income near the mode of the distribution than at the tails. In particular, the misinterpretation arises when eqn (11) instead of eqn (9) is applied to unimodal distributions when assessing the relative sensitivity of G to income transfers. Consequently, the assumptions supporting the mathematical structure of eqn (11) are ignored.

It follows from eqn (10) that the Gini ratio fulfils the duality principle between the representation of an inequality measure (I) satisfying the principle of transfer, i.e. $I=E[V(x)]$, and that of a social welfare (SW) function, i.e. $SW=E[-V(x)]$, where $-V(x)$ is concave, or more generally, S-concave (Berge, 1966). It follows from eqn (9), that two equivalent forms of $V(x)$ in $G=E[V(x)]$ are

$$V(x) = 2xF(x) / \mu - 1, \quad \text{and} \quad V(x) = x[2F(x) - 1] / \mu. \tag{12}$$

Sen (1974) introduced an axiomatic system for the SW interpretation of the Gini ratio based on the individual income ranking of the population suggested by the structure of eqn (10). Following Sen's ideas, Kakwani (1980, pp. 77–9) presented a SW interpretation of the Gini ratio as a function of income. Both approaches can be presented in a compact form by making use of the SDF $S(x)=1-F(x)$ and the first moment survival distribution function $S_1(x)=1-L(x)$. In fact, specifying the SW function

$$SW(X) = E[Xv(x)] \tag{13}$$

where $v(X)$ is a decreasing and differentiable function of X, and making $v(X)=2S(X)=2(1-F(X))$, i.e., twice the frequency of economic units with income greater than X, we deduce

$$SW(X) = 2\int_0^\infty xS(x)\,dF(x) = \mu(1-G), \tag{14}$$

which proves Sen's (1974, p. 410) theorem that the SW function (14) ranks a set of distributions of a constant total income and population in precisely the same way as the negative of the Gini ratio of the respective distributions, i.e. in reverse order from that by the cardinal value of the Gini ratio. On the other hand, making $v(X)=bS_1(X)=b[1-L(X)]$, $b>0$ and $\int_0^\infty v(x)\,dF(x) = 1$, where $S_1(x)$ is the income share of the economic units with income greater than x, we deduce

$$b\int_0^\infty [1-L(x)]\,dF(x) = b(1+G)/2 = 1, \text{ and } SW(X) = [2/(1+G)]\int_0^\infty x[1-L(x)]\,dF(x) = \mu/(1+G), \tag{15}$$

which also states that the SW function (15) is a decreasing function of the Gini ratio. The result obtained in eqn (14) supports Sen's (1976, p. 384) cogent statement that ‘one might wonder about the significance of the debate on the non-existence of any additive utility function which ranks income distributions in the same order as the Gini ratio’.
The Gini ratio stimulated important contributions such as:

1. (i) The construction of a confidence interval for G. Given a random sample of size n, eqn (10) is an unbiased estimator of G. However, income distribution data are presented by class intervals, hence Gini (1914) proposed the formula $G_L=1-2A$, where A is the area under the Lorenz curve (Figure 1) estimated by application of the trapezoidal approximation to

$$\int_0^1 L\,dF,$$

   thus underestimating Gbecause the trapezoidal rule implies that within each interval, income is equally distributed. Gastwirth (1972)derived an upper bound G_u, by maximizing the spread within each income interval, and proposed (G_L, G_u) as a confidence interval within which a parametric estimate of Gshould fall. Dagum (1980a)proved that his confidence interval is a necessary but not sufficient condition to assess a model goodness of fit.
2. (ii) The Gini ratio gives a welfare ranking (weak ordering) of a set of income distributions of a constant mass of income and over a constant population, and a strict partial ordering among the subset of income distributions with non-intersecting Lorenz curves. This conclusion is further supported by eqns (14) and (15).
3. (iii) The welfare ranking of income distributions with equal and different means can be obtained via a decision function $R(G, D)$, where the ratio G states the preference for less inequality (inequality aversion) regardless of the mean income (so that the partial derivative $R_G<0$), and the relative economic affluence D (Dagum, 1980b, 1987) states the preference for more income (poverty aversion), so that the partial derivative $R_D>0$.
4. (iv) Research on the economics of poverty led Sen (1976) to the specification of an axiomatic structure of a new poverty measure as a function of (a) the relative frequency of the poor members of the population, (b) a weighted average of the poverty gap, i.e. the aggregate shortfall from the poverty line of the poor population, and (c) the Gini ratio of

the income distribution of the subpopulation with incomes below the poverty line.

5. (v) Gini (1932) introduced a new coordinate system taking as the abscissa the egalitarian line $F=L$ and as the ordinate the distance between the Lorenz curve and the egalitarian line. Gini thoroughly analysed this new coordinate system and its relation to the G ratio. Kakwani (1980, ch. 7) worked with a similar transformation.
6. (vi) Analysing consumer behaviour in India, Mahalanobis (1960) extended and generalized the Lorenz curve and the Gini ratio with the introduction of the concentration curve and ratio, respectively. Other authors such as Kakwani (1980, chs 8–14) made further contributions and dealt with the relationships among the distribution of several economic variables such as expenditures and income after tax, and investigated the degree of tax and public expenditure progressivity or regressivity. If $y=g(x)$ is the function of income that is the object of inquiry, $g(x)$ must be non-negative. For the particular case of $g(x)=x$, the concentration curve and ratio are identical to the Lorenz curve and Gini ratio, respectively. Moreover, if $g(x)$ is an increasing and differentiable function of x, i.e. $g'(x)>0$, then the concentration ratio is equal to the Gini ratio for the function g(x).
7. (vii) The decomposition approach disaggregates a population according to some relevant socio-economic attributes and analyses the equality within each subpopulation and between them, and assesses the contribution of each subpopulation to overall inequality. This approach also disaggregates the income variable by source of income such as wages and salaries, self-employment, pension and government transfers. Bhattacharya and Mahalanobis (1967) were the first to deal with the decomposition of the Gini ratio. Several authors made further contributions to this topic, among them Pyatt (1976) and Shorrocks (1983).

Bibliography

Berge, C. 1966. *Espaces topologiques, fonctions multivoques*. 2nd edn, Paris: Dunod.

Bhattacharya, N. and Mahalanobis, P.C. 1967. Regional dispartities in household consumption in India. *Journal of the American Statistical Association*62, 143–61.

Dagum, C. 1980a. The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquée*33(2), 327–67.

Dagum, C. 1980b. Inequality measures between income distributions with applications. *Econometrica*48(7), November, 1791–803.

Dagum, C. 1983. Income inequality measures. In *Encyclopedia of Statistical Sciences*, Vol. IV, ed. S. Kotz and N.L. Johnson. New York: John Wiley & Sons, 34–40.

Dagum, C. 1987. Measuring the economic affluence between populations of income receivers. *Journal of Business and Economic Statistics*5(1), 5–12.

Dalton, H. 1920. The measurement of the inequality of incomes. *Economic Journal*30(September), 348–61.

Gastwirth, J.L. 1972. The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics*54, 306–16.

Gini, C. 1910. Indici di concentrazione e di dependenza. *Atti delta III Riunione delta Società Italiana per il Progresso delle Sciencze*, in Gini(1955), 3–120.

Gini, C. 1912. Variabilitàe mutabilità. *Studi economico-giuridici, Universit*à *di Cagliari* III, 2a, in Gini (1955), 211–382.

Gini, C. 1914. Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti*, in Gini (1955), 411–459.

Gini, C. 1932. Intorno alle curve di concentrazione. *Metron* 9(3–4), in Gini (1955), 651–724.

Gini, C. 1955. *Memorie di metodologia statistica*. Vol. 1: *Variabilità>concentrazione*, ed. E. Pizetti and T. Salvemini, Rome: Libreria Eredi Virgilio Veschi.

Kakwani, N. 1980. *Inequality and Poverty: methods of estimation and policy applications*. Oxford: Oxford University Press.

Mahalanobis, P.C. 1960. A method of fractile graphical analysis. *Econometrica* 28(2), April, 325–351.

Pareto, V. 1895. La legge della domanda. *Giornale degli Economisti*, 59–68.

Pareto, V. 1896. Ecrits sur la courbe de la répartition de la richesse. In *Oeuvres complètes de Vilfredo Pareto*, ed. Giovanni Busino, Geneva: Librairie Droz, 1965.

Pareto, V. 1897. *Cours d'éeconomie politique*, ed. G.H. Bousquet and G. Busino, Geneva: Librairie Droz 1964.

Pyatt, G. 1976. On the interpretation and disaggregation of Gini coefficients. *Economic Journal* 86(June), 243–55.

Ricardo, D. 1817. Principles of Political Economy. In *Works and Correspondence of David Ricardo*, Vol. I, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Sen, A.K. 1974. Information bases of alternative welfare approaches. *Journal of Public Economics*3, 387–403.

Sen, A.K. 1976. Poverty: an ordinal approach to measurement. *Econometrica*44(2), March, 219–31.

Shorrocks, A.F. 1983. The impact of income components on the distribution of family income. *Quarterly Journal of Economics*98(2), May, 311–26.

How to cite this article

Dagum, Camilo. "Gini ratio." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000047> doi:10.1057/9780230226203.0646

The New Palgrave Dictionary of Economics Online

Gini, Corrado (1884 – 1965)

Camilo Dagum

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

demography; distribution of income and wealth; Gini Coefficient; Gini identity; Gini mean difference; Gini, C.; human capital; index numbers; inequality (measurement); internal migration; international migration; statistics and economics

Article

Gini, perhaps best known to economists because of the Gini Coefficient, was born in Motta di Livenza, Italy and died in Rome. He studied at the University of Bologna; his doctoral thesis *Il sesso dal punto di vista statistico* (1908), defended in 1905, was awarded the Vittorio Emanuele prize for social sciences. Gini distinguished himself as a teacher and a researcher. In 1909 he was appointed an assistant professor of the University of Cagliari, becoming full professor a year later. Gini won a chair at the University of Padova in 1913, then joined the University of Rome in 1925, where in 1955 he was awarded the distinction of emeritus professor. Social scientist and statistician, Gini taught economics, statistics, sociology and demography, making path-breaking contributions to these highly related disciplines. Among them we mention the neo-organicist theory (Gini, 1909; 1924a) that presents a dynamic theory of society in which demographic factors (differential birth rates among social classes and social mobility) play a basic role. In this theory, Gini introduced and analysed self-conservation, self-regulative and self-re-equilibrating mechanisms, thus offering a well-structured anticipation of Wiener's cybernetics, von Bertalanffy's general system theory and modern disequilibrium economics. He provided new insights to the analysis of inter- and intra-national migrations (Gini, 1948) and demographic dynamics (Gini, 1908; 1909; 1912a; 1931). He developed a methodology to evaluate the income and wealth of nations (Gini, 1914a; 1959) including a discussion of human capital, already present in his research on the causes and consequences of international migrations. In this context he specified a model of income and wealth distributions and a measure of income and wealth inequalities (Gini, 1909; 1912b; 1914b; 1955). Gini's research interests motivated important contributions to statistics and economics, such as the Gini identity (1921; 1924b) on price index numbers, the Gini mean difference (1912b), the transvariation theory (Gini, 1916; 1960), the index of dissimilarity (Gini, 1914c) and the Gini

Coefficient. Gini founded several scientific journals, such as *Metron* and *Genus*, and academic institutions, such as the Institute and Faculty of Statistics, Demography and Actuarial Sciences of the University of Rome; and was the organizer and first president (1926–32) of the Istituto Centrale di Statistica. An extraordinarily prolific writer and thinker, endowed with powerful new ideas that he developed in more than 70 books and 700 articles, Gini was in the 20th century a true Renaissance man.

Selected works

1908. *Il sesso dal punto di vista statistico*. Milan: Sandrom.

1909. Il diverso accrescimento delle classi sociali e la concentrazione della ricchezza. *Giornale degli Economisti* 38, January, 27–83.

1912a. *I fattori demografici dell'evoluzione delle nazioni*. Turin: Bocca.

1912b. Variabilità e mutabilità. Reprinted in Gini (1955).

1914a. *L'ammontare e la composizione della ricchezza delle nazioni*. 2nd edn, Turin: UTET, 1962.

1914b. Sulla misura della concentrazione e della variabilità dei caratteri. Reprinted in Gini (1955).

1914c. Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche. *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti* 74, Part II.

1916. Il concetto di transvarizione e le sue prime applicazioni. Reprinted in Gini (1960).

1921. Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali. *Metron* 1, 63–82.

1924a. *Patologia economica*. 5th edn, Turin: UTET, 1954.

1924b. Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues. *Metron* 4, 3–162.

1931. *Le basi scientifiche della politica della popolazione*. Catania: Studio Editoriale Moderno.

1948. Apparent and real causes of American prosperity. *Banca Nazionale del Lavoro Quarterly Review* 6, July–September.

1955. *Memorie di metodologia statistica*. Vol. 1: *Variabilità e concentrazione*, ed. A. Pizzetti and T. Salvemini. Rome: Veschi.

1959. *Ricchezza e reddito*. Turin: UTET.

1960. *Transvariazione*, ed. G. Ottaviani. Rome: Libreria Goliardica.

How to cite this article

Dagum, Camilo. "Gini, Corrado (1884–1965)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000048> doi:10.1057/9780230226203.0645

The New Palgrave Dictionary of Economics Online

GIS data in economics

Henry G. Overman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Geographical information systems (GIS) are used for inputting, storing, managing, analysing and mapping spatial data. This article considers the role each of these functions can play in economics. GIS can map economic data with a spatial component, generate additional spatial data as inputs to statistical analysis, calculate distances between features of interest and define neighbourhoods around objects. GIS also introduce economics to new data. For example, remote sensing provides large amounts of data on the earth's surface. These data are of inherent interest, but can also provide an exogenous source of variation and allow the construction of innovative instrumental variables.

Keywords

geographical information systems (GIS) and economics; spatial data; hedonic analysis

Article

Geographical information systems (GIS) are used for inputting, storing, managing, analysing and mapping spatial data. In this article, we consider each of these functions to help assess the role that GIS can play in economic analysis. Of course, a wide range of software can provide similar functions for quantitative data, so it is the geographical, or spatial, element that separates GIS. That spatial dimension is the focus here. One important aspect of GIS that is not covered is the choice of software. Standard texts, such as Longley et al. (2005, ch. 7), consider the question of appropriate software in some depth.

At the outset, note that, while GIS are widely used in business, government and a range of academic disciplines, their application in economics has to date been more limited. The most frequent application in economics is the use of GIS to visualize or map economic data with a spatial component. Most entry-level courses in econometrics begin with a plea to ‘plot the data’ at an early stage of the analysis to help identify trends, outliers and so forth. Much the same could be said of the role of mapping spatial data, and GIS provide a simple and efficient way to do this.

Less common, but arguably more interesting, is the use of GIS storage and management functions to

generate additional data as inputs into further statistical analysis. In the simplest case this will involve using GIS to manage spatial data from a variety of sources. Many of these sources – for example, sampling and census data – will be familiar to economists, others, such as aerial photography and remote sensing data, less so. The spatial nature, or format, of the data will depend on the geographical data model used. The two most common models are raster format (assigning a code to each cell on a regular grid) and vector format (assigning a code to, and providing coordinates for, irregular polygons). GIS provide tools for moving between these different geographical data models. While the methods used to do this are rather intuitive, the devil is in the detail. As with the implementation of pre-packed econometric routines, one should understand the underlying basis of these transformations before proceeding. These issues are covered at depth in most, if not all, of the standard references, and we do not consider them further here.

More generally, it is the ability of GIS to reconcile spatial data from different sources that allows the creation of new data-sets. In the simplest case, this may involve combining socio-economic data from different spatial units – for example, population data from US census tracts with employment data for US zip codes. Many economists will be used to using ready-made concordances (that is, mappings from one set of spatial units to another) for undertaking such data merges. GIS bring the flexibility of allowing users to define their own concordances between different geographical units of observation when faced with data from different sources.

The construction of more ambitious data-sets is possible if one is willing to draw on a range of analytical functions available in the more advanced GIS. GIS can be used to identify whether observations occur at particular locations and, if so, to identify the characteristics of observations at those locations. For example, one of the most frequent applications of GIS in economics to date has been to identify and characterize properties for use in hedonic analysis (see Bateman et al., 2002, for a review). At its most basic, this will simply involve the merging of different data-sets as described above. However, much more complex analysis is possible. Given that GIS data are spatial, a natural use is to measure the distance between observations or between observations and other features of interest. These distances could be physical distances or network distances (for example, along a transport network), or involve some more general concept of social distance.

Observation-to-observation distance calculations have been widely applied in the fields of biology and biomedical sciences through the statistical analysis of spatial point patterns (see Diggle, 2003). Knowing the distance between observations is useful if we think that there may be interactions between them and that the strength of these interactions is mitigated by distance. For example, in industrial organization models of spatial competition the intensity of competition may depend on the distance between firms. Observation-to-observation distances are also useful when the underlying entities are free to choose their location and we want to assess whether there are systematic patterns to those location decisions. For example, the study of localization asks whether firms in a particular industry tend to be spatially concentrated relative to overall economic activity. If they do, one would expect the observation-to-observation distances to be less for firms in that industry than for a randomly chosen set of firms from the economy at large. The increasing availability of geo-referenced economic data suggests that the application of appropriately adapted procedures will become more common in economics (see Duranton and Overman, 2005). Hedonic analysis again provides the most frequent application of GIS to calculate distances of observations from other features. For example, in their study valuing rail access Gibbons

and Machin (2005) use GIS to measure the proximity of properties to rivers, coasts, woodlands, roads, railway lines and airports.

In addition to the calculation of distances, GIS can be used to construct measures of area or to define neighbourhoods (or ‘buffers’) around objects. For example, Burchfield et al. (2006) in their study of urban sprawl use GIS to calculate the percentage of the urban fringe – defined as a 20-kilometer buffer around existing development – that lies above water-yielding aquifers.

These examples cover the main types of spatial analyses that are undertaken to construct spatial data in economic applications, but others are possible and should be covered in any of the standard texts. It should be noted that in advanced GIS these operations can be done both interactively and automatically using batch files (that is, where the user writes a sequence of commands in a file that the computer implements one by one). Both approaches, but particularly the latter, involve fairly large fixed costs in terms of both purchasing software and learning how to implement the relevant procedures. There are other methods for conducting many of these analyses that do not imply the use of GIS. For example, great circle distances can easily be calculated on the basis of latitude and longitude (see Overman and Ioannides, 2004). Whether the fixed cost investment is worthwhile will depend on individual circumstances. The benefits can be substantial. In many circumstances, GIS calculations should be more accurate than short cuts implemented with the use of non-spatial software, and some analysis such as the calculation of areas and buffers is much easier to implement in GIS.

GIS also introduce economics to new sources of data. In particular, remote sensing from either satellite or aerial photography, or digitized geological maps, can provide a huge amount of data on the earth's surface. Early applications using these kinds of data tended to focus on issues arising from natural resource management such as valuing timber yields from forested areas. However, data on land cover and land use (that is, the physical features that cover the land and what those features are used for), soil type, geological and landscape features, elevation and climate are opening up new avenues of research. These data sources allow the description of different features of the economic landscape that one might seek to explain. For example, Burchfield et al. (2006) use remote sensing data to track the evolution of land use on a grid of 8.7 billion 30×30 metre cells covering the conterminous United States and then seek to explain differences in land development patterns across cities. Another example is Rappaport's (2006) study of the role that weather plays in explaining population changes in US counties. The meteorological data that he uses comes from 6,000 meteorological stations and covers 20 winter, summer and precipitation variables. GIS analysis by the Spatial Climate Analysis Surface at Oregon State University applied to this meteorological data allows the construction of weather variables for a two-kilometre grid covering the continental United States.

GIS data also have the potential to contribute to a range of established areas of study, particularly because data on the earth's surface can provide an exogenous source of variation and thus allow researchers to construct instrumental variables using GIS. Some examples should help to make this idea concrete. Hoxby (2000) is interested in whether competition among public schools improves school outcomes. That is, do cities with more school districts have better public schools and less private schooling? The problem that the analysis needs to confront is that, for a city of a given size, better public schools and fewer private schools in a city should imply more school districts. That is, the number of districts is endogenous to public school quality. What is needed is an instrument that should determine the supply of school districts but that is independent of the local public school quality. Hoxby argues

that the number of streams in a metropolitan area provides such an instrument. Cities with a large number of streams end up with more school districts for reasons that are surely nothing to do with public school quality. Hoxby provides a well-known example of the strategy, although not of the use of GIS, as her work is based on the study of detailed paper maps.

Rosenthal and Strange (2005) provide an example of the use of GIS to implement such an instrumental variables strategy. They are interested in whether density of employment helps determine wages. The problem is that higher wages should attract more workers and lead to higher employment densities. That is, density may be caused by wages and not vice versa. Rosenthal and Strange argue that the density of employment will be partly determined by the height of buildings in a location. They point out that the height of buildings is, in turn, partly dependent on the underlying geology of the site. Given that geology should not determine wages directly (they are studying cities, not agricultural production), the underlying geology can be used as an instrument. Locations with a suitable underlying geology can have higher buildings and higher employment density, and should thus have higher wages. Rosenthal and Strange use GIS data on the type of underlying bedrock, seismic and landslip hazard as instruments for the density of employment in their regressions of wages on employment density. Such examples suggest a potentially important role in future work for GIS data as a component in novel instrumentation strategies.

This piece has only skimmed the surface of the potential applications of GIS in economics. As spatially referenced socio-economic data becomes more widely available, it is to be expected that the scope for applications can only increase.

See Also

- location theory
- new economic geography
- spatial econometrics
- spatial economics

Bibliography

Bateman, I. and Lovett, A. 1998. Using Geographical Information Systems (GIS) and large area databases to predict yield class: a study of Sitka spruce in Wales. *Forestry* 71, 147–68.

Bateman, I., Jones, A., Lovett, A., Lake, I. and Day, B. 2002. Applying Geographical Information Systems (GIS) to environmental and resource economics. *Environmental and Resource Economics* 22, 219–69.

Burchfield, M., Overman, H., Puga, D. and Turner, M. 2006. The determinants of sprawl: a portrait from space. *Quarterly Journal of Economics* 135, 587–633.

Diggle, P. 2003. *Statistical Analysis of Spatial Point Patterns*. London: Arnold.

Duranton, G. and Overman, H. 2005. Testing for localisation using micro-geographic data. *Review of Economic Studies* 72, 1077–106.

Gibbons, S. and Machin, S. 2005. Valuing rail access using transport innovations. *Journal of Urban Economics* 57, 148–69.

Hoxby, C. 2000. Does competition among public schools benefit students and taxpayers? *American Economic Review* 90, 1209–38.

Longley, P., Goodchild, M., Maguire, D. and Rhind, D. 2005. *Geographical Information Systems and Science*, 2nd edn. Chichester, UK: John Wiley and Sons.

Overman, H. and Ioannides, Y. 2004. The spatial evolution of the US urban system. *Journal of Economic Geography* 4, 131–56.

Rappaport, J. 2006. Moving to nice weather. Working Paper No. 03–07. Research Division, Federal Reserve Bank of Kansas City.

Rosenthal, S. and Strange, W. 2005. The attenuation of agglomeration economies: a Manhattan skyline approach. Working paper. University of Toronto.

How to cite this article

Overman, Henry G. "GIS data in economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000184> doi:10.1057/9780230226203.0647

The New Palgrave Dictionary of Economics Online

global analysis in economic theory

Steve Smale

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Global analysis in economics puts the main results of classical equilibrium theory into a global calculus context. The advantages of this approach are: (*a*) the proofs of existence of equilibrium are simpler (the main tool is the calculus of several variables); (*b*) comparative statics is integrated into the model in a natural way; (*c*) the calculus approach is closer to the older traditions of the subject; and (*d*) as far as possible the proofs of equilibrium are constructive.

Keywords

Arrow–Debreu theory; calculus of variations; comparative statics; existence of equilibrium; global analysis in economic theory; inverse function th; proofs of equilibrium; Sard's th; Walras's Law

Article

The goal here is to illustrate ‘global analysis in economics’ by putting the main results of classical equilibrium theory into a global calculus context. The advantages of this approach are fourfold:

1. 1. The proofs of existence of equilibrium are simpler. Kakutani's fixed point theorem is not used, the main tool being the calculus of several variables.
2. 2. Comparative statics is integrated into the model in a natural way, the first derivatives playing a fundamental role.
3. 3. The calculus approach is closer to the older traditions of the subject.
4. 4. In so far as possible the proofs of equilibrium are constructive. These proofs may be implemented by a speedy algorithm, which is Newton's method modified to give global convergence. On the other hand, the existence proofs are sufficiently powerful to yield the generality of the Arrow–Debreu theory.

Only two references are given at the end of this entry, each containing an extensive bibliography with historical notes. The two references themselves give detailed, expanded accounts of the subject of global analysis in economic theory.

Let us proceed to an account of this model. The basic equation of equilibrium theory is ‘supply equals demand’, or in symbols, $S(p) = D(p)$. Since we are in a situation of several markets, there are several variables in this equation. Equilibrium prices are obtained by setting the excess demand $z = D - S$ equal to zero and solving. Consider this function *z* on a more abstract level.

Suppose that, given an economy of *l* markets, or of *l* commodities with corresponding prices written as $p_1,\ldots, p_l$, the excess demand for the *i*th commodity is a real valued function $z_i = z_i, (p_1, \ldots, p_l)\, p_j \geq 0$ and we form the vector $z = (z_l, \ldots, z_1)$. Thus the excess demand can be interpreted as a map, which we take to be sufficiently differentiable, from $\mathbb{R}^l_+$ to $\mathbb{R}^l$, where $\mathbb{R}^l$ is Cartesian *l*-space and $\mathbb{R}^l_+ = \left\{p \in \mathbb{R}^l / \; p_l \geq 0\right\}$. An economic equilibrium is a set of prices $p = (p_1, \ldots, p_l)$, for which excess demand is zero, that is, $z(p) = 0$.

Economic theory imposes some conditions on the function *z* which go as follows.

First and foremost is Walras's Law, which is expressed simply by $p \cdot z(p) = 0$ (inner product). Written out, this is

$$\sum_{i=1}^{l} p_i \cdot z_i(p_1, \ldots, p_l) = 0$$

and states that the value of the excess demand is zero. This is a budget constraint which asserts that the excess demand is consistent with the total assets of the economy. It can be proved from a reasonable microeconomic foundation, as can be seen below.

Second is the homogeneity condition $z(\lambda p) = z(p)$ for all $\lambda > 0$. Changing all prices by the same factor does not affect excess demand. This condition reflects the fact that the economy is self-contained; prices are not based on anything lying outside the model.

The final condition is the boundary condition that $z_i(p) \geq 0$ if $p_i = 0$. This may be interpreted as: if the good is free, then there will be a non-negative excess demand for it.

The following result and its generalizations and ramifications lie at the heart of economic theory.

Existence th

Suppose that an excess demand *z* satisfies Walras's Law, homogeneity, and the boundary condition. Then there is a price equilibrium.

We will give the proof under the additional mild non-degeneracy condition that the derivative of *z* is non-singular somewhere on the boundary of $\mathbb{R}^l_+$. This proof is based on Sard's theorem and the inverse function theorem, two basic theorems of global analysis.

Consider a differentiable map *f* from a set *U* contained in $\mathbb{R}^k$ to $\mathbb{R}^n$. A vector $y \in \mathbb{R}^n$ is said to be a *regular* value if at each point $x \in U$ with $f(x) = y$, the derivative $Df(x): \mathbb{R}^k \to \mathbb{R}^n$ is surjective. A subset of $\mathbb{R}^n$ is of *full measure* if its complement has measure zero.

Sard's theorem

If a map $f: U \to \mathbb{R}^n$ is of class (sufficient differentiability Cr, $r > k - n$, 0) then the set of regular values of *f* has full measure.
A subset *V* of $\mathbb{R}^n$ is called a *k*-dimensional *submanifold* if for each point, there is a neighbourhood *U* in *V* and a change of coordinates of $\mathbb{R}^n$ which throws *U* into a coordinate subspace of dimension *k*.

Inverse function theorem

If $y \in \mathbb{R}^n$ is a regular value of a smooth map $f: U \to \mathbb{R}^n$, $U \subset \mathbb{R}^k$, then either $f^{-1}(y)$ is empty or it is a submanifold of dimension *k–n*.
Let us sketch out the proof of the existence theorem. Define the space of normalized prices by

$$\Delta_1 = \left\{ p \in \mathbb{R}^l_+ \,/\, \Sigma p_l = 1 \right\}.$$

A space auxiliary to the commodity space is defined by

$$\Delta_0 = \left\{ z \in \mathbb{R}^l \,/\, \Sigma z_l = 0 \right\}.$$

From the excess demand map $z: \mathbb{R}^l_+ - 0 \to \mathbb{R}^l$, define an associated map $\varphi: \Delta_1 \to \Delta_0$ by $\varphi(p) = Z(p) - \varphi z^i(p)\, p$. Note that $\phi(p)$ is well-defined (i.e. $\varphi(p) \in \Delta_0$) and also smooth. Note moreover that if $\varphi(p) = 0$, then $z(p) = 0$, and that *p* is a price equilibrium. This follows from Walras's Law as follows. If $\varphi(p) = 0$, then $z(p) = \Sigma z^i(p)\, p$ and so $p \cdot z(p) = \Sigma z^i(p)\, p \cdot p = 0$. Therefore $\Sigma z^i(p) = 0$, since $p \cdot p \neq 0$. By the previous equation *z*(*p*) must be zero.
The boundary condition on *z* implies that ϕ satisfies a similar boundary condition. That is, if $p_i = 0$ then $\varphi_i(p) = z_i(p) \geq 0$.
It is now sufficient to show that $\varphi(p) = 0$ for some $p \in \Delta_1$. The argument for this proceeds by defining yet another map $\hat{\varphi}$ by

$$\hat{\varphi}(p) = \frac{\varphi(p)}{\|\varphi(p)\|}$$

where $E = \varphi^{-1}(0)$ and S^{1-2} is the set of unit vectors in Δ_0.
By definition the set E is the set of price equilibria, which is to be shown not empty.
Let p_0 be a price vector on the boundary of Δ_1 where the derivative $D\phi$ (p_0) is non-degenerate (our special hypothesis implies the existence of this p_0). One applies Sard's theorem to obtain'a regular value y of $\hat{\varphi}$ in S^{1-2} near $\hat{\varphi}(p_0)$, where $\hat{\varphi}^{-1}(y)$ is non-empty.
From the inverse function theorem it follows that $\hat{\varphi}^{-1}(y)$ a smooth curve in Δ_1, (a 1-dimensional submanifold). From the boundary condition and a short argument which we omit, it follows that this curve cannot leave Δ_1.
Since the curve $\hat{\varphi}^{-1}(y)$ is a closed set in $\Delta_1 - E$ and has no end points (the inverse function theorem implies that) it must tend to E. In particular, E is not empty and therefore the existence theorem is proved.
The above proof is ‘geometrically” constructive in that a curve $\gamma = \hat{\varphi}^{-1}(y)$ is constructed which leads to a price equilibrium. This picture can be made analytic by showing that y is a solution of the ordinary differential equation ‘Global Newton’, $d\varphi / dt = \lambda D\varphi(p)^{-1}\varphi(p)$, where λ is +1 or –1 determined by the sign of the determinant of $D\phi$ (p). As a consequence the Euler method of approximating the solution of an ordinary differential equation can be used to obtain a discrete algorithm for locating a price equilibrium. By an appropriate choice of steps, ±1, this discrete algorithm near that equilibrium is Newton's Method; thus the appellation ‘Global Newton’ for the differential equation.
One would like to understand the process of convergence to equilibrium in terms of decentralized mechanics of price adjustment. Unfortunately the situation in this respect is unclear.
Next we give a brief picture of how global analysis relates to a pure exchange economy. This will allow a microeconomic derivation of the excess demand function discussed above, so that the existence theorem just proved will imply an existence theorem for a price equilibrium of a pure exchange economy. Continuing in this framework one can prove Debreu's theorem on generic finiteness of price equilibria, by putting the structure of a differentiable manifold on the big set of price equilibria. The equilibrium manifold is a natural setting for comparative statics.
A trader's preferences will be supposed to be represented by a smooth utility function $u: P \to \mathbb{R}$, where $P = \{x \in \mathbb{R}^l, x_i > 0\}$, is commodity space. The indifference surfaces are those $u^{-1}(c) \subset P$. We make strong versions of classical hypotheses on this function.
Monotonicity. The gradient, grad $u(x)$, has positive coordinates.
Convexity. The second derivative D2u(x) is negative difinite on the tangent space at x of the corresponding indifference surface.

Boundary condition. The indifference surfaces are closed sets in $\mathbb{R}^l$ (not just P).

From the utility function, one defines for the individual trader a demand function. $f : \mathbb{R}^l_+ \times \mathbb{R}_+ \to P$ of prices $p \in \mathbb{R}^l_+$ and wealth $w > 0$. For this, consider the budget set $B_{p.w} = \{x \in P \mid p \cdot x = w\}$. Then *f*(*p*, *w*) is the maximum of *f*. on $B_{p \cdot w}$.

One can prove:

Proposition: . The demand function *f* satisfies

1. (a) grad $u(f(p, w) = \lambda p$, for some $\lambda > 0$
2. (b) $p \cdot f(p, w) = w$
3. (c) $f(\lambda p, \lambda w) = f(p, w)$ any $\lambda > 0$
4. (d) *f* is smooth.

A pure exchange economy will be a set of *m* traders, each with preferences as discussed above, associated to utility fuctions u_i $i = 1, \ldots, m$ defined on the same commodity space *P*. Also associated to the *i*th trader is an endowment vector $e_i \in P$. At prices *p*, this trader's wealth is the value of his endowment $p \cdot e_i = w_i$. A *state* is an allocation (x_1,…,x_m), $x_i \in P$ and a price system $p \in \mathbb{R}_+$.

Feasibility is the condition:

$$(F) \sum x_i = \sum e_i$$

A kind of satisfaction condition of a state is

(*S*) For each *i*, x_i *maximizes* u_i *on the budget set*

$$B = \{y \in P \mid p \,.\, y = p \,.\, e_i\},$$

An economic equilibrium of a pure exchange economy (e_1,…,e_m, u_1,…,u_m) is a state [x_1, …, x_m), *p*] satisfying ((*F*)) and (*S*).

Theorem.: There exists a price equilibrium of every pure exchange economy.

The proof goes by applying the previous existence theorem above. Define the excess demand $Z = D - S$ as follows:

$$S(p) = \sum e_i,\ D(p) = \sum f_i(p,\ p \cdot e_i),$$

where f_i is the above defined demand of the ith trader. One then shows Walras's Law:

$$p \cdot Z(p) = p \cdot D(p) - p \cdot S(p) = \sum p \cdot f_i(p, p \cdot e_i) - p \cdot \sum e_i = 0$$

using (b) of the proposition above.
Use (c) of the proposition to confirm the homogeneity of z. The use of the boundary condition is more technical. But under the rather strong hypotheses, this gives a fairly complete existence proof for a price equilibrium of a pure exchange economy.
This existence proof extends to prove the Arrow–Debreu theorem in the generality of the latter's *Theory of Value*.

See Also

- general equilibrium
- mathematical economics
- regular economies

Bibliography

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: A Differentiable Approach*. Cambridge: Cambridge University Press.

Smale, S. 1981. Global analysis and economics. In *Handbook of Mathematical Economics*, vol. 1, ed. K. J. Arrow and M. D. Intriligator. Amsterdam: North-Holland.

How to cite this article

Smale, Steve. "global analysis in economic theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000160> doi:10.1057/9780230226203.0648

The New Palgrave Dictionary of Economics Online

global games

Stephen Morris
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Global games are a class of incomplete information games where small uncertainty about payoffs implies a significant failure of common knowledge. This allows strategic uncertainty to play a crucial and natural role in pinning down equilibrium play. Introduced in the context of two player, two action games by Carlsson and van Damme (1993), global games have inspired tractable modelling frameworks that have been used in a wide variety of applications. This article reviews the key ideas that have played a role in theoretical analysis and creating useful applied tools.

Keywords

comparative statics; complete information games; currency crises; equilibrium selection; global games; incomplete information games; limit dominance; limit uniqueness; multiple equilibria; noise; noise independent selection; regime change game; risk dominance; robustness herding; signalling; strategic complementarities; sunspot equilibrium; supermodularity; threshold strategies

Article

Complete information games often have multiple Nash equilibria. Game theorists have long been interested in finding a way of removing or reducing that multiplicity. Carlsson and van Damme (1993) (CvD) introduced an original and attractive approach to doing so. A complete information model entails the implicit assumption that there is common knowledge among the players of the payoffs of the game. In practice, such common knowledge will often be lacking. CvD suggested a convenient and intuitive way of relaxing that common knowledge assumption: suppose that, instead of observing payoffs exactly, payoffs are observed with a small amount of continuous noise; and suppose that – before observing their signals of payoffs – there was an *ex ante* stage where any payoffs were possible. Based on the latter feature, CvD dubbed such games ‘global games’. It turns out that there is a unique equilibrium in the game with a small amount of noise. This uniqueness remains no matter how small the noise is and is independent of the distribution of the noise. Since complete information, or common knowledge of payoffs, is surely always an idealization anyway, the play selected in the global game with small noise can be seen as a prediction for play in the underlying complete information game.
The following example illustrates the main idea. There are two players each of whom must decide whether to invest or not invest. Action ‘not invest’ always gives a payoff of 0. Action ‘invest’ always gives a payoff of θ; but there are strategic complementarities, and if the other player does not invest, then the player loses 1. Thus the payoff matrix is:

	Invest	Not invest
Invest	θ, θ	$\theta-1, 0$
Not invest	$0, \theta-1$	$0, 0$

Let us first examine the Nash equlibria of this game when θ is common knowledge. If $\theta<0$, then 'invest' is a strictly dominated action for each player, and thus 'not invest, not invest' is the unique Nash (and dominant strategies) equilibrium. If $\theta>1$, then 'not invest' is a strictly dominated action for each player, and 'invest, invest' is the unique Nash (and dominant strategies) equilibrium. The multiplicity case arises if $0<\theta<1$. In this case, there are two strict Nash equilibria (both not invest and both invest) and there is also a strictly mixed Nash equilibrium.

But suppose the players do not exactly observe θ. Suppose for convenience that each player believes that θ is uniformly distributed on the real line (thus there is an 'improper' prior with infinite mass: this does not cause any technical or conceptually difficulties as players will always condition on signals that generate 'proper' posteriors). Suppose that each player observes a signal $x_i = \theta + \sigma\varepsilon_i$, where each ε_i is independently normally distributed with mean 0 and standard deviation 1.

In this game of incomplete information, a pure strategy for player *i* is a mapping $s_i: \mathbb{R} \to \{\text{Invest, Not invest}\}$. Suppose player 1 was sure that player 2 was going to follow a 'threshold' strategy where she invested only if her signal were above *k*, so

$$s_2(x_2) = \begin{cases} \text{Invest, if } x_2 > k \\ \text{Not invest, if } x_2 \le k \end{cases}$$

What is player 1's best response? First, observe that his expectation of θ is x_1. Second, note that (under the uniform prior assumption) his posterior on θ is normal with mean x_1 and variance σ^2, and thus his posterior on x_2 is normal with mean x_1 and variance $2\sigma^2$. Thus his expectation that player *j* will not invest is $\Phi(\frac{1}{\sqrt{2}\sigma}(k-x_1))$, where Φ is cumulative distribution of the standard normal. Thus his expected payoff is

$$x_1 - \Phi\left(\frac{1}{\sqrt{2}\sigma}(k-x_1)\right), \quad (1)$$

and player 1 will invest if and only if (1) is positive. Now if we write $b(k)$ for the unique value of x_1 setting (1) equal to 0 (this is well defined since (1) is strictly increasing in x_1), the best response of player 1 is then to follow a cut-off strategy with threshold equal to $b(k)$. Observe that as $k \to -\infty$ (player 2 always invests), (1) tends to x_1, so $b(k) \to 0$. As $k \to \infty$ (player 2 never invests), (1) tends to x_1-1, so $b(k) \to 1$. Also observe that if $k = \frac{1}{2}$, then $b(k) = \frac{1}{2}$, since if player 1 observes signal $\frac{1}{2}$, his expectation of θ is $\frac{1}{2}$ but he assigns probability $\frac{1}{2}$ to player 2 not investing. Finally, observe that (by total differentiation)

$$b'(k) = \frac{1}{1 + \frac{\sqrt{2\pi}}{\varphi\left(\frac{1}{\sqrt{2}\sigma}(k - x_1)\right)}} \in (0, 1),$$

so $b(k)$ is strictly increasing in k and we can immediately conclude that there is a unique 'threshold' equilibrium where each player uses a threshold of $\frac{1}{2}$.

The strategy with threshold $\frac{1}{2}$ is in fact the unique strategy surviving iterated deletion of (interim) strictly dominated strategies. In fact, a strategy s survives n rounds of iterated deletion of strict dominated strategies if and only if

$$s(x) = \begin{cases} \text{Invest, if } x > b^n(1) \\ \text{Not invest, if } x < b^n(0) \end{cases}$$

where $b^n(k) = \overbrace{b(b(\ldots b(k)))}^{n \text{ times}}$.

The key intuition for this example is that the uniform prior assumption ensures that each player, whatever his signal, attaches probability $\frac{1}{2}$ to his opponent having a higher signal and probability $\frac{1}{2}$ to him having a lower signal. This property remains true no matter how small the noise is, but breaks discontinuously in the limit: when noise is zero, he attaches probability 1 to his opponent having the same signal.

In this article, I will first report how Carlsson and van Damme's (1993) analysis can be used to give a complete general analysis of two player two action games. I will then report in turn theoretical extensions of their work and a literature that has used insights from global games in economic applications. This dichotomy is somewhat arbitrary (many 'applied' papers have significant theoretical contributions) but convenient.

1 Two-player, two-action games

Let the payoffs of a two-player, two-action game be given by the following matrix:

	A	B
A	θ_1, θ_2	θ_3, θ_4
B	θ_5, θ_6	θ_7, θ_8

Thus a vector $\theta \in \mathbb{R}^8$ describes the payoffs of the game and is drawn from some distribution. For a generic choice of θ, there are three possible configurations of Nash equilibria.

1. 1. There is a unique Nash equilibrium with both players using strictly mixed strategies.

2. 2. There is a unique strict Nash equilibrium with both players using pure strategies.
3. 3. There are two pure strategy strict Nash equilibria and one strictly mixed strategy Nash equilibrium.

In the last case, Harsanyi and Selten (1988) proposed the criterion of risk dominance to select among the multiple Nash equilibria. Suppose that (*A*, *A*) and (*B*, *B*) are strict Nash equilibria of the above game (that is, $\theta_1 > \theta_5$, $\theta_7 > \theta_3$, $\theta_2 > \theta_4$ and $\theta_8 > \theta_6$). Then (*A*, *A*) is a risk dominant equilibrium if

$$(\theta_1 - \theta_5)(\theta_2 - \theta_4) > (\theta_7 - \theta_3)(\theta_8 - \theta_6).$$

Generically, exactly one of the two pure Nash equilibria will be risk dominant.
Now consider the following incomplete information game $G(\sigma)$. Each player *i* observes a signal $x_i = \theta + \sigma\varepsilon_i$, where the ε_i are eight-dimensional noise terms. Thus we have an incomplete information game parameterized by $\sigma \geq 0$. A strategy for a player is a function from possible signals $\mathbb{R}^8$ to the action set {*A*, *B*}. For any given strategy profile of players in the game $G(\sigma)$ and any actual realization of the payoffs θ, we can ask what is the distribution over action profiles in the game (averaging across signal realizations).
Theorem: For any sequence of games $G(\sigma^k)$ where $\sigma^k \to 0$ and any sequence of equilibria of those games, average play converges at almost all payoff realizations to the unique Nash equilibrium (if there is one) and to the risk dominant Nash equilibrium (if there are multiple Nash equilibria).
This is shown by the main result of Carlsson and van Damme (1993) in cases (2) and (3) above. They generalize the argument from the example described above to show that, if an action is part of a risk dominant equilibrium or a unique strict Nash equilibrium of the complete information game θ, then – for sufficiently small σ – that action is the unique action surviving iterated deletion strictly dominated strategies. Kajii and Morris (1997) show that, if a game has a unique correlated equilibrium, then that equilibrium is ‘robust to incomplete information’, that is, will continue to be played in some equilibrium if we change payoffs with small probability. This argument can be extended to show the theorem for case (1) (the extension is discussed in Morris and Shin, 2003).

2 Theoretical extensions; many players and many actions

Carlsson and van Damme (1993) dubbed their perturbed games for the two player, two action case ‘global games’ because all possible payoff profiles were possible. They showed that there was a general way of adding noise to the payoff structure such that, as the noise went to zero, there was a unique action surviving iterated deletion of (interim) dominated strategies (a ‘limit uniqueness’ result). And they showed that the action that got played in the limit was independent of the distribution of noise added (a ‘noise independent selection’ result). Their result does not extend in general to many player many action games. In discussing known extensions, we must carefully distinguish which of their results extend.
Frankel, Morris and Pauzner (2003) consider games with strategic complementarities (that is, supermodular payoffs). Rather than allow for all possible payoff profiles, they restrict attention to a one-dimensional set of possible payoff functions, or states, which are ordered so that higher states lead to higher actions. The idea of ‘global’ games is captured by a ‘limit dominance’ property: for sufficiently low values of θ, each player has a dominant strategy to choose his lowest action, and that for sufficiently high values of θ, each player has a dominant strategy to choose his highest action. Under these restrictions, they are able to present a complete analysis of the case with many players, asymmetric payoffs and many actions. In particular, a limit uniqueness result holds: if each player observes the state with noise, and the size of noise goes to zero, then in the limit there is a unique strategy profile surviving iterated deletion of strictly dominated strategies. Note that while Carlsson and van Damme required no strategic complementarity and other monotonicity properties, when there are multiple equilibria in a two-player, two-action game – the interesting case for Carlsson and van Damme’s analysis – there are automatically strategic complementarities.
Within this class of monotonic global games where limit uniqueness holds, Frankel, Morris and Pauzner (2003) also provide sufficient conditions for ‘noise independent selection’. That is, for some complete information games, which action gets played in the limit as noise goes to zero does not depend on the shape of the noise. Frankel, Morris and Pauzner (2003) show that a generalization of the potential maximizing action profile is sufficient for noise independent selection. This sufficient condition encompasses the risk dominant selection in two player binary action games; the selection of the ‘Laplacian’ action (a best response to a uniform distribution over others’ actions) in many player, binary action games (Morris and Shin, 2003). It also yields unique predictions in the continuum player currency crisis of Guimaeres and Morris (2004) and in two-player, three-action games with symmetric payoffs. Morris and Ui (2005) give further sufficient conditions for equilibria to be ‘robust to incomplete information’ in the sense of Kajii and Morris (1997), which will also ensure noise independent section.

However, Frankel, Morris and Pauzner (2003) also provide an example of a two-player, four-action, symmetric payoff game where noise independent selection fails. Thus there is a unique limit as the noise goes to zero, but the nature of the limit depends on the exact distribution of the noise. Carlsson (1989) gave a three-player, two-action example in which noise independent selection failed. Corsetti et al. (2004) describe a global games model of currency crises, where there is a continuum of small traders and a single large trader. This is thus a many-player, two-action game with *asymmetric* payoffs. The equilibrium selected as noise goes to zero depends on the relative informativeness of the large and small traders' signals. This is thus an application where noise-independent selection fails.

More limited results are available on global games without supermodular payoffs. In many applications – such as bank runs – there are some strategic complementarities but payoffs are not supermodular everywhere: conditional on enough people running on the bank to cause collapse, I am better off if I run if few people run and share in the liquidation of the bank's assets. An important paper of Goldstein and Pauzner (2005) has shown equilibrium uniqueness for 'bank run payoffs' – satisfying a single crossing property – with uniform prior and uniform noise. This analysis has been followed in a number of applications. They establish that there is a unique equilibrium in threshold strategies and there are no non-threshold equilibria. However, their analysis does not address the qst of which strategies survive iterated deletion of strictly dominated strategies. Morris and Shin (2003) discuss how the existence of a unique threshold equilibrium can be established more generally under a signal crossing property on payoffs and a monotone likelihood ratio property on signals (not required for global games analysis with supermodular payoffs); however, these arguments do not rule out the existence of non-monotonic equilibria. Results of van Zandt and Vives (2007) can be used more generally to establish the existence of a unique monotone equilibrium under weaker conditions than supermodularity.

The original analysis of Carlsson and van Damme (2003) relaxed the assumption of common knowledge of payoffs in a particular way: they assumed that there was a common prior on payoffs and that each player observes a small conditionally independent signal of payoffs. This is an intuitively small perturbation of the game and this is the perturbation that has been the focus of study in the global games literature. However, when the noise is small one can show that types in the perturbed game are close to common knowledge types in the product topology on the universal type space: that is, for each type t in the perturbed game, there is a common knowledge type t' such that type t and t' almost agree in their beliefs about payoffs, they almost agree about their beliefs about the opponents' beliefs, and so on up to any finite level. Thus the 'discontinuity' in equilibrium outcomes in global games when noise goes to zero is illustrating the same sensitivity to higher order beliefs of the famous example of Rubinstein (1989). Now we can ask: how general is the phenomenon that Rubinstein (1989) and Carlsson and van Damme (1993) identified? That is, for which games and actions is it the case that, under common knowledge, the action is part of an equilibrium (and thus survives iterated deletion of strictly dominated strategies) but for a type 'close' to common knowledge of that game, that action is the unique action surviving iterated deletion of strictly dominated strategies. Weinstein and Yildiz (2007) shows that this is true for every action surviving iterated deletion of strictly dominated strategies in the original game. This observation highlights the fact that the selections that arise in standard global games arise not just because one relaxes common knowledge, but because it is relaxed in a particular way: the common prior assumption is maintained and outcomes are analysed under that common prior, and the noisy signal technology ensures particular properties of higher-order beliefs, that is, that each player's beliefs about how other players' beliefs differ from his is not too dependent on the level of his beliefs.

3 Applications; public signals and dynamic games

Complete information models are often used in applied economic analysis for tractability: the complete information game payoffs capture the essence of the economic problem. Presumably there is not in fact common knowledge of payoffs, but if asymmetries of information are not the focus of the economic analysis, this assumption seems harmless. But complete information games often have multiple equilibria, and policy analysis – and comparative statics more generally – are hard to carry out in multiple equilibrium models. The global games analysis surveyed above has highlighted how natural relaxations of the common knowledge assumptions often lead to intuitive selections of a unique equilibrium. This suggests these ideas might be useful in applications. Fukao (1994) and Morris and Shin (1995) were two early papers that pursued this agenda. The latter paper – published as Morris and Shin (1998) – was an application to currency crises, where the existing literature builds on a dichotomy between 'fundamentals-driven' models and multiple equilibrium or 'sunspot' equilibria views of currency crises. This dichotomy does not make sense in a global games model of currency crises: currency attacks are 'self-fulfilling' – in the sense that speculators are attacking only because they expect others to do so – but their expectations of others' behaviour may nonetheless be pinned down by higher order beliefs (see Heinemann, 2000, for an important correction of the equilibrium characterization in Morris and Shin, 1998). Morris and Shin (2000) laid out the methodological case for using global games as a framework for economic applications. Morris and Shin (2003) surveys many early applications to currency crises, bank runs, the design of international institutions and asset pricing, and there have been many more since. Rather than attempt to survey these applications, I will highlight two important methodological issues – public signals and dynamics – that have played an important role in the developing applied literature.

To do this, it is useful to consider an example that has become a workhorse of the applied literature, dubbed the 'regime change' game in a recent paper of Angeletos, Hellwig and Pavan (2007). The example comes from a 1999 working paper on 'Coordination Risk and the Price of Debt' presented as a plenary talk at the 1999 European meetings of the Econometric Society, eventually published as Morris and Shin (2004). A continuum of players must decide whether to invest or not invest. The cost of investing is c. The payoff to investing is one if the proportion investing is at least $1-\theta$, 0 otherwise. If there is common knowledge of θ and $\theta \in (0, 1)$, there are multiple Nash equilibria of this continuum

player complete information game: ‘all invest’ and ‘all not invest’. But now suppose that θ is normally distributed with mean *y* and standard deviation τ. Each player in the continuum population observes the mean *y* (which is thus a public signal of θ). But in addition, each player *i* observes a private signal x_i, where the private signals are distributed in the continuum population with mean θ and standard deviation σ (that is, as in the example at the beginning of this article). Morris and Shin (2004) show that the resulting game of incomplete information has a unique equilibrium if and only if $\sigma \leq \sqrt{2\pi}\tau^2$, that is, if private signals are sufficiently accurate relative to the accuracy of public signals. This result is intuitive: we know that if there is common knowledge of θ, there are multiple equilibria. A very small value of τ means that the public signal is very accurate and there is ‘almost’ common knowledge.

This result makes it possible to conduct comparative statics within a unique equilibrium not only in the uniform prior, no ‘public’ information, limit but also with non-trivial public information. A distinctive comparative static that arises is that the unique equilibrium is very sensitive to the public signal *y*, even conditioning on the true state θ (see Morris and Shin, 2003; 2004; Angeletos and Werning, 2006). This is because, for each player, the public signal *y* becomes a more accurate prediction of others’ behaviour than his private signal, even if they are of equal precision.

But the sensitivity of the uniqueness result to public signals also raises a robustness qst. Public information is endogenously generated in economic settings, and thus a qst that arises in many dynamic applications of global games in general and the regime change game in particular is when endogenous information generates enough public information to get back multiplicity (Tarashev, 2003; Dasgupta, 2007; Angeletos, Hellwig and Pavan, 2006; 2007; Angeletos and Werning, 2006; Hellwig, Mukherji and Tsyvinski, 2006). This literature has highlighted the importance of endogenous information revelation and the variety of channels through which such revelation may lead to multiplicity or enhance uniqueness. In addition, these and other dynamic applications of global games raise many other important methodological issues, such as the interaction between the global game uniqueness logic and ‘herding’ – informational externalities in dynamic settings without payoff complementarities – and ‘signalling’ – biasing choices from static best responses in order to influence opponents’ beliefs in the future.

See Also

- coordination problems and communication
- currency crises
- purification
- quantal response equilibria

Bibliography

Angeletos, G.-M., Hellwig, C. and Pavan, A. 2006. Signalling in a global game: coordination and policy traps. *Journal of Political Economy* 114, 452–84.

Angeletos, G.-M., Hellwig, C. and Pavan, A. 2007. Dynamic global games of regime change: learning, multiplicity and timing of attacks. *Econometrica* 75, 711–56.

Angeletos, G.-M. and Werning, I. 2006. Crises and prices: information aggregation, multiplicity and volatility. *American Economic Review* 96, 1720–36.

Carlsson, H. 1989. Global games and the risk dominance criterion. Unpublished manuscript, Department of Economics, University of Lund.

Carlsson, H. and van Damme, E. 1993. Global games and equilibrium selection. *Econometrica* 61, 989–1018.

Corsetti, G., Dasgupta, A., Shin, H.S. and Morris, S. 2004. Does one Soros make a difference? The role of a large trader in currency crises. *Review of Economic Studies* 71, 87–114.

Dasgupta, A. 2007. Coordination and delay in global games. *Journal of Economic Theory* 134, 195–225.

Frankel, D., Morris, S. and Pauzner, A. 2003. Equilibrium selection in global games with strategic complementarities. *Journal of Economic Theory* 108, 1–44.

Fukao, K. 1994. Coordination failures under incomplete information and global games. Discussion Paper No. A299, Institute of Economic Research, Hitotsubashi University.

Goldstein, I. and Pauzner, A. 2005. Demand-deposit contracts and the probability of bank runs. *Journal of Finance* 60, 1293–327.

Guimaeres, B. and Morris, S. 2004. Risk and wealth in a model of self-fulfilling currency attacks. Discussion Paper No. 1433R, Cowles Foundation.

Harsanyi, J. and Selten, R. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Heinemann, F. 2000. Unique equilibrium in a model of self-fulfilling currency attacks: comment. *American Economic Review* 90, 316–8.

Hellwig, C., Mukherji, A. and Tsyvinski, A. 2006. Self-fulfilling currency crises: the role of interest rates. *American Economic Review* 96, 1769–87.

Kajii, A. and Morris, S. 1997. The robustness of equilibria to incomplete information. *Econometrica* 65, 1283–309.

Morris, S. and Shin, H.S. 1995. Informational events that trigger currency attacks. Working Paper No. 95-24, Federal Reserve Bank of Philadelphia.

Morris, S. and Shin, H.S. 1998. Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review* 88, 587–97.

Morris, S. and Shin, H.S. 2000. Rethinking multiple equilibria in macroeconomics. In *NBER Macroeconomics Annual 2000*, ed. B.S. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Morris, S. and Shin, H.S. 2003. Global games: theory and applications. In *Advances in Economics and Econometrics (Proceedings of the Eighth World Congress of the Econometric Society)*, ed. M. Dewatripont, L. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

Morris, S. and Shin, H.S. 2004. Coordination risk and the price of debt. *European Economic Review* 48, 133–53.

Morris, S. and Ui, T. 2005. Generalized potentials and robust sets of equilibria. *Journal of Economic Theory* 124, 45–78.

Rubinstein, A. 1989. The electronic mail game: strategic behavior under almost common knowledge. *American Economic Review* 79, 385–91.

Tarashev, N. 2003. Speculative attacks and informational role of the interest rate. *Journal of the European Economic Association*.

van Zandt, T. and Vives, X. 2007. Monotone equilibria in Bayesian games of strategic complementarities. *Journal of Economic Theory* 134, 339–60.

Weinstein, J. and Yildiz, M. 2007. A structure theorem for rationalizability with application to robust predictions of refinements. *Econometrica* 75, 365–400.

How to cite this article

Morris, Stephen. "global games." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000179> doi:10.1057/9780230226203.0649

The New Palgrave Dictionary of Economics Online

globalization and labour

Richard B. Freeman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The entry of China, India, and the ex-Soviet countries into the world trading system in the 1990s has made globalization an increasingly important driver of labour outcomes across the world. Through trade, capital flows, the spread of technology and education, the world has begun to move towards a truly global labour market. Still, the dispersion of wages for similar work across countries remains high and immigration is the least developed part of globalization, leaving considerable scope for national labour markets, policies, and institutions to affect wages and worker well-being into the foreseeable future.

Keywords

brain drain; child labour; comparative advantage; cost of capital; diffusion of technology; factor endowments; factor mobility; factor price equalization; fair trade; first-mover advantage; foreign direct investment; foreign portfolio investment; globalization; globalization and labour; Heckscher–Ohlin trade theory; higher education; inequality (global); international migration; international trade; labour standards; North–South economic relations; occupational health and safety; price dispersion; product life cycle; production possibility frontier; purchasing power parity; Ricardian trade theory; transfer of technology; wage dispersion

Article

Globalization – the export and import of goods and services, international capital mobility, labour mobility, and technical knowledge across national borders – connects economies and influences the economic well-being of workers worldwide. Imports reduce the demand for workers in a country by substituting foreign labour, whose work is embodied in the imports, for domestic labour. Exports increase the demand for workers by selling what workers produce to other countries. Since the wages paid for labour differ among countries, firms have sizable incentives to offshore some jobs to foreign countries, including many service sector jobs that have been historically non-tradable. Capital mobility

changes the capital stock with which workers operate, raising or lowering demand for labour. Immigration, business trips, international study and tourism affect the supply and demand for labour. And, most important of all, the flow of knowledge across borders allows countries to improve their technical and economic prowess and operate along the global production possibility frontier even when they lack the scientific base to expand the frontier. Although economic analyses generally treat trade in goods and services, capital flows, labour flows, and the transfer of knowledge separately, these four facets of globalization have feedbacks and connections that help determine their impact on the economy and on the work force.

At the end of the 20th century globalization became a more powerful driver of labour market outcomes than ever before. The collapse of Soviet Communism, China's shift to market capitalism and India's market reforms and entry into the global trading system produced a single economic world based on capitalism and markets. Before those changes, the global economy encompassed roughly half of the world's population – the advanced countries, Latin America, the Caribbean, Africa, and some parts of Asia – while the other half lived in separate economic spaces. Workers in the United States and other higher income countries and in market-oriented developing countries did not face competition from low-wage Chinese or Indian workers nor from workers in the Soviet empire. The entry of these economies into the world trading system in the 1990s increased the global labour pool from approximately 1.46 billion workers to 2.93 billion workers – ‘the great doubling’ (see Freeman, 2005a; 2005b).

As documented in Freeman (2006, pp. 150–1) and data given on the International Monetary Fund (IMF) website, all aspects of globalization grew at the turn of the 21st century. World trade increased relative to world GDP so that world exports rose to 27 per cent of world GDP in 2005 compared to just 12 per cent of world GDP in 1970. Foreign direct investment, which had been 2–3 per cent of global gross capital formation in the 1970s rose to 7–20 per cent of gross capital formation in the 1990s–2000s. The share of foreign equities in investors’ equity portfolios rose from negligible numbers to about 15 per cent in the early 2000s. Immigration from developing countries to advanced countries increased so that in 2000 8.7 per cent of the population in the high income countries had been born elsewhere. The single biggest recipient of immigrants was the United States, where the share of immigrants nearly tripled from 1970 to 2005 and where roughly one in five workers aged 25–39 was foreign-born. As for the transfer of knowledge, university enrolments grew rapidly worldwide and multinationals moved production to developing countries. China, in particular, made rapid gains in measures of technological prowess. According to the Georgia Technology Policy and Assessment Center, between 1993 and 2005, China more than tripled its rating in technological standing (Porter et al., 2006, Table 3).

A comparison of the different facets of globalization indicates that the ratio of immigrants to the world work force is lower than the ratio of trade to goods production and international capital flows to activity in capital markets, which suggests that immigration is the least developed part of globalization. To some extent, this may reflect the greater personal cost in moving from one country to another than to ship goods or capital across borders. But there is a political economy reason as well. Even countries committed to freer trade and capital mobility do not allow for free immigration.

Economists use two types of models to analyse the effects of globalization on economic performance and the well-being of workers. They use the basic Heckscher–Ohlin model of comparative advantage to analyse trade between advanced countries and developing countries. This model takes country factor endowments (labour skills, natural resources, capital) as given and examines how these differences affect trade, capital flows, and labour flows, and through them prices, wages, and returns to capital. In

this model trade and factor mobility are substitute ways to reduce the economic effects of differing factor endowments and thus to reduce price and wage differences across countries. Restrictions of trade induce capital or labour flows that substitute for the restricted trade, and conversely restrictions on factor mobility induce trade (Mundell, 1957).

To analyse trade among countries with similar levels of economic development, economists use Ricardian models of trade. These models treat differences in technology as the fundamental determinant of trade and factor flows and examine how investments in technology create comparative advantage. Factor mobility magnifies differences in factor endowments because labour and capital move to economies where the technological advantage creates greater demand for them. Trade and factor mobility are complements in the sense that a technologically advantaged sector which uses, say highly skilled labour, will attract highly skilled immigrants to help it expand.

Because factor endowments and technology differ across countries and change over time, both sets of models are needed to make sense of globalization and labour.

When factor endowments differ

If one identifies skilled labour, unskilled labour, capital, and natural resources as the relevant factors of production, trade patterns between advanced and developing countries fit the Heckscher–Ohlin model to a first approximation (Debaere, 2003, gives a favourable reading of the empirical validity of this model, while Trefler, 1995, is more critical). Countries with abundant skilled labour, such as the United States, export goods produced by skilled workers, and import goods made by low-skill labour, while countries with natural resources export those resources and import goods and services made with other inputs. But Heckscher–Ohlin models are silent on the huge volumes of trade among advanced countries with similar factor endowments and on the huge volumes of trade within industries (see Ruffin, 1999).

In addition, the pattern of factor flows is not consistent with the Heckscher–Ohlin model. Unskilled labour migrates from developing countries, where it is relatively abundant, to advanced countries, where it is relatively scarce, as the model predicts, but skilled labour also migrates to advanced countries, while it should move in the other direction. The brain drain, which is a sizable part of immigration, magnifies differences in factor proportions across countries, and thus creates a problem for analyses that view factor flows as responses to factor endowments.

The model also has predictions about the impact of globalization on factor prices that do not fit reality. It predicts that trade and factor flows will lower the relative returns to scarce factors and reduce the returns to abundant factors. This implies that globalization should increase wage differentials and inequality in advanced countries and reduce wage differentials and inequality in less advanced countries. Globalization is associated with rising inequality in advanced countries, which is consistent with the model, but it is also associated with rising inequality in many developing countries, which is inconsistent with the model. (For a review of studies showing rising wage differentials in developing countries associated with globalization, see Goldberg and Pavcnik, 2007.) One explanation for this is that the skilled workers in the developing countries are more comparable in their skills to the unskilled workers in the advanced countries, so that when the developing countries export products previously made by the unskilled workers in advanced countries, demand for skilled workers in developing countries is increased. But other factors may be at work as well (see Zhu and Trefler, 2005), so there is no clear

resolution to this surprising pattern.

Wage and factor price equalization

Goods and factor flows motivated by national differences in factor endowments should reduce the cross-country dispersion of prices, the cost of capital and the wages of comparable workers. The factor price equalization theorem predicts that under specified conditions, trade alone will equalize factor prices. While some trade theorists dismiss factor price equalization as a theoretical curiosum, the logic of globalization dictates market pressures towards equality of wages as well as other prices across country lines.

In fact, the prices of many goods and services differ only moderately across countries. For instance, in 2004 the price of McDonald's Big Mac sandwich showed a narrow distribution across countries. The 80th percentile of Big Mac prices among 65 countries was 2.65 dollars while the 20th percentile of Big Mac prices was 1.40 dollars – a 1.9:1 spread (Freeman, 2006, p. 151). Similarly, estimates of international differences in the cost of capital show a ratio of costs at the top 25th percentile of countries to costs at the bottom 25th percentile of 1.43. This averages estimates from five different sources from Hail and Leuz (2004, Table 1). By contrast, the variation of wages in the same occupation is much greater. The 1998–2002 occupational wages around the world data file shows that wages for the country at the top 20 per cent point of the earnings distribution of countries for a *given narrowly defined* occupation are about 12 times the wages in the country at the bottom 20 per cent point of earnings distribution, if one uses exchange rates to compare currencies, and four to five to one if one compares currencies with purchasing power parity price indices – that is, price indices of different currencies based on a given basket of good that differ from exchange rates in part due to the different prices of non-tradables across countries (Freeman and Oostendorp, 2001). While part of the cross-country variation in wages for workers in the same occupation reflects differences in the education and skill of workers in the same occupation in advanced and developing countries, this cannot explain the wide variation in the earnings of, say, barbers in low-income countries and in high-income countries. The offshoring of computer programming and of call-centre work to India in the 2000s highlighted the fact that in some occupations workers in low-income countries have similar skills to those in more advanced countries. What differs are the wages paid across countries.

That the cross-country dispersion of wages is greater than the cross-country dispersion of prices or of the cost of capital suggests that globalization has had a smaller impact on the price of labour than on the prices of other factors. One possible reason for this is that, as noted earlier, international migration is a quantitatively smaller facet of globalization than trade or the international flow of capital. This explanation requires that the direct effect of trade on the prices of goods and services and the direct effect of capital flows on the cost of capital is greater than their indirect effect on wages.

The labour standards debate

Globalization has made labour standards – workplace safety, freedom from discrimination, rights to unionize, hours and wage regulations – in developing countries a major issue for the international community. Human rights activists in advanced countries campaign to get multinational firms to

implement better labour conditions in their plants and in those of their subcontractors in developing countries. The activists contend that consumers are willing to pay for the higher standards through higher prices and will avoid products made under bad conditions. There is indeed evidence that consumers will pay a bit more for ‘fair trade’ products and will shun products made under poor conditions (see Elliot and Freeman, 2003; Hiscox and Smyth, 2005). In response to activist pressures, many multinationals have developed and implemented codes of conduct for their operations in developing countries. Although activists fear that low standards in developing countries will produce a global race to the bottom in standards, labour standards have risen in advanced countries during the period of rapidly increased trade with developing countries and in many developing countries as well. One indication of this is that advanced and developing countries have signed on to more International Labor Organization conventions during the period of globalization than ever before. Even the poorest countries have sought to reduce the use of child labour (see Elliot and Freeman, 2003).

Some advocates of free trade regard the activist pressure for improved labour standards in developing countries as disingenuous protectionism. ‘The talk of “exploitation”, failure to pay a “living wage”•…• (is) little more than cynical manipulation of our moral instincts and an obfuscation of the reality to pursue our economic interest’ (Bhagwati, 2000). ‘The demand for linkage between trading rights and the observance of standards with respect to the environment and labour would seem to arise largely from protectionist motivation’ (Srinivasan, 1994, p. 36). But the activists are not rival producers of imported products who aim to move production from developing countries to advanced countries. Rather, they are students, consumers, and members of non-governmental organizations who seek to organize retail markets so that consumers pay higher prices for items made under better conditions. Their motivation is intrinsic, not pecuniary interest.

What troubles free trade advocates is the danger that standards will impair the comparative advantage of developing countries. Motives aside, even policies intending to help workers in developing countries could harm them if those policies were so costly that they reduced the cost advantage of developing countries to expand in some low-wage labour-intensive sectors. However, the huge gap in labour costs between countries suggests that improved standards cannot threaten comparative advantage. In any case, part of the cost of standards falls on workers who prefer higher standards to lower standards; and part will be paid by consumers who want products made under good conditions. Some standards that raise costs to firms, moreover, benefit developing economies over the long run. Child labour laws, and school attendance laws, for instance, increase human capital formation; while occupational safety regulations reduce injuries and fatalities that may burden a country's medical or welfare system. As long as countries have flexible exchange rates, moreover, they can buy whatever labour standards they want without suffering economic disaster. If Brazil chooses to spend more on occupational health and safety than China, Brazilian firms will be at a competitive disadvantage *at a given exchange rate*. But the Brazilian currency will depreciate in relation to the Chinese currency, and all Brazilians will bear the cost of the health and safety standards through the higher cost of imports. Brazilian industries that spend a lot more to meet health and safety standards will contract as Brazil's comparative advantage shifts to industries that do not need to spend much more. Thus, globalization does not restrict national choices in labour standards or in other areas of social choice.

Globalization when technology differs

In a truly global labour market the same worker would earn roughly comparable real pay in different countries, as measured in the purchasing power parity price indices that give a more realistic comparison of living standards across countries than do comparisons of earnings based on exchange rates. The labour market at the outset of the 21st century was far from a single global market. Workers from a low-income country could make more than six times the earnings in their home country by immigrating to an advanced country (see Freeman, 2006). Why? Because the advanced country had higher capital–labour ratios, superior infrastructure, greater legal protections of property and persons, and more *advanced technology*, which raised productivity compared to productivity in the immigrant's native country. The differences in technology that affect earnings should be thought of in broad terms as including differences in organizational structure and business practices as well as differences in engineering or scientific technology. They include the economies of scale that give ‘first mover’ advantages to the firm or country that produces a good first. Countries with a comparative advantage in a sector – higher productivity in relation to other sectors compared to trading partners – will export output of that sector and import goods and services from sectors in which its trading partner has a comparative advantage. The sector with comparative advantage will expand, raising the wages of the factors that it uses the most and attracting those factors from the country and the rest of the world. If an advanced country has comparative advantage in, say, high tech that uses many computer scientists, persons with computer science degrees will immigrate to the country and strengthen its advantage in that sector. Trade and mobility magnify differences in factor endowments. Since countries will shift resources towards sectors in which they have comparative advantage, world output will rise, and so too will the wages of workers.

The ‘North–South’ model provides a platform for analysing trade between advanced countries (North) and developing countries (South) when investment in technology creates comparative advantage. In this model the North's advantage is in innovative high-tech products because it has many scientists, engineers, and other high-skilled workers, while the South's advantage is in producing standard products that use less skilled labour. The wages of ordinary workers in the North exceed those of workers in the South because the North earns a monopoly rent on technological innovation. The wage advantage is higher the greater the rate of technological innovation in relation to the rate of knowledge transfers to the South. The result is an industry or product life cycle that begins with an innovation in the North and ends with production in the South. Krugman (1979) gives a clear exposition of this model.

When technology creates comparative advantage, the productivity advances in one country can affect the economy of a trading partner positively or negatively depending on whether the advance occurs in goods or services that the trading partner exports or goods/services that it imports. If a trading partner improves productivity in an import, this will reduce the cost of production and the price of the import, which benefits the country that imports the good as well as (in most cases) the exporter. But if a trading partner improves technology in an export, this can harm the exporting country, just as an improved technology in a competitor can harm a firm. The increased supply of an exported good will drive down its price and thus the income of the country that originally dominated the production.

As a result, when countries make their comparative advantage by investing in skills or technology rather than having comparative advantage set by factor endowments there can be situations of ‘conflicting national interests’ in trade, as stressed by Gomory and Baumol (2000). If a foreign competitor gains comparative advantage in industries that have desirable attributes – that employ large numbers of highly educated and skilled workers or offer great opportunities for rapid technological advance – the lead

economy will have to shift resources to less desirable sectors – and lose some of the advantages it had gained from trade. Applying these analyses to debates over offshoring and technological transfer in the United States, Paul Samuelson reminded trade economists that while the spread of technology around the world raises world output and productivity it need not be in the interest of the technological leader (Samuelson, 2004).

Globalization/labour debates: human resource leapfrogging

The rapid growth of higher education in populous developing countries, notably China and India, challenges the assumption that advanced countries inevitably have comparative advantage in high-tech sectors. The share of scientific papers from Asia has risen substantially, due largely to increased scientific activity in China, which is moving to the forefront of science and technology. Digitalization of work has led to offshoring computer-related work, particularly to India. To take advantage of low-priced scientists and engineers in these countries, multinational firms have established research centres there. While the South has far fewer scientists and engineers per capita than the North, it can compete in high tech because success at the technological frontier depends on the *absolute* number of scientists and engineers in an area, not simply on the number in relation to the total work force. A country like China or India can have proportionately fewer scientists, engineers, and entrepreneurs per capita than an advanced country but still have absolutely more of these workers available at lower wages than the advanced country. By producing numerous graduate scientists, engineers, and other university specialists and deploying them in the high-tech innovative sectors that the advanced countries had viewed as their birthright, the populous developing countries can move to the technological frontier through ‘human resource leapfrogging’.

This does not mean that advanced countries lose when developing countries raise their technological prowess and economic competitiveness. The increased supply of highly educated workers around the world should expand the world's production possibility frontier rapidly, which will benefit all countries. In addition, the lower prices of high-tech goods and services produced in developing countries, such as PCs from China and call centre technical advice from India, benefit all consumers. But increased competition from low-wage countries in sectors where advanced countries have had comparative advantage can reduce or eliminate their advantages in those sectors. Comparative advantage in high-tech or most other sectors is not the birthright of any country. Globalization of technological progress will raise world output and income. It is likely to benefit workers in developing countries more than those in advanced countries, reducing global income inequality. It could lower the living standards or rate of growth of the living standards of advanced country workers for whom workers in low-income countries are good substitutes.

Does globalization rule the roost? Will it dominate labour outcomes in the future?

In 1995 I posed the question ‘Are your wages set in Beijing?’ to direct attention towards the impact of globalization on wages in advanced countries (Freeman, 1995). My answer then, and now, is negative. The dispersion of wages for similar work around the world documented earlier shows that globalization does not rule labour markets. National labour markets, and the policies and institutions that unions,

firms, and countries use to regulate those markets, affect wages and worker well-being independent of what happens in other countries. But the pressures of globalization on wage setting around the world will rise as the highly populous economies of China and India increase their share of the global economy. Globalization makes what happens in Beijing•…•and Calcutta•…•and Rio•…•and Warsaw and so on, important drivers of labour market developments worldwide. Still, the persistence of variation in labour market outcomes across the states in the United States, where there are no restrictions on goods, factor flows, or knowledge flow, suggests that, even though global economic forces are likely to increase their impact on wages and other outcomes, they will not 'rule the roost'. There will remain space for variation in labour markets among countries just as there is for regional and local markets within countries.

See Also

- globalization
- Heckscher–Ohlin trade theory
- labour economics
- purchasing power parity
- Ricardian trade theory
- trade, technology diffusion and growth

Bibliography

Bhagwati, J. 2000. Why Nike is on the right track. *Financial Times*, 1 May.

Debaere, P. 2003. Relative factor abundance and trade. *Journal of Political Economy* 111, 589–610.

Elliot, K. and Freeman, R. 2003. *Can Labor Standards Improve Under Globalization?* Washingon, DC: Institute for International Economics.

Freeman, R.B. 1995. Are your wages set in Beijing? *Journal of Economic Perspectives* 9(3), 15–32.

Freeman, R. 2005a. The great doubling: America in the new global economy. Usery Lecture, Georgia State University, 8 April.

Freeman, R. 2005b. What really ails Europe (and America): the doubling of the global workforce. *The Globalist*, 3 June.

Freeman, R. 2006. People flows. *Journal of Economic Perspectives* 20(2), 145–70.

Freeman, R.B. and Oostendorp, R.H. 2001. The occupational wages around the world data file. *International Labour Review* 140, 379–401.

Goldberg, P.K. and Pavcnik, N. 2007. Distributional effects of globalization in developing countries. *Journal of Economic Literature* 45, 39–82.

Gomory, R. and Baumol, W. 2000. *Global Trade and Conflicting National Interests*. Cambridge, MA: MIT Press.

Hail, L. and Leuz, C. 2004. International differences in the cost of equity capital: do legal institutions and securities regulation matter? Working Paper 04–06, Wharton Financial Institutions Center, University of Pennsylvania. Online. Available at http://fic.wharton.upenn.edu/fic/papers/04/0406.pdf, accessed 12 June 2007.

Hiscox, M. and Smyth, N. 2005. Is there consumer demand for improved labor standards? Evidence from field experiments in social labeling. Harvard University. Online. Available at http://www.courses.fas.harvard.edu:9095/~gov3009/Calendar/SocialLabeling_2.pdf, accessed 17 June 2007.

International Monetary Fund. 2007. World Economic Outlook database 2004. Online. Available at http://www.imf.org/external/pubs/ft/weo/2004/01/data/index.htm, accessed 12 June 2007.

Krugman, P. 1979. A model of innovation, technology transfer, and the world distribution of income. *Journal of Political Economy* 87, 253–66.

Mundell, R. 1957. International trade and factor mobility. *American Economic Review* 47, 321–35.

Porter, A.L, Roessner, J.D. Newman, N., Jin, X.-Y. and Johnson, D. 2006. *High Tech Indicators: Technology-Based Competitiveness of 33 Nations: 2005 Final Report*. Atlanta: Technology Policy and Assessment Center, Georgia Institute of Technology.

Ruffin, R.J. 1999. The nature and significance of intra-industry trade. *Economic and Financial Review*. Federal Reserve Bank of Dallas. Online. Available at www.dallasfed.org/research/efr/1999/efr9904a.pdf, accessed 12 June 2007.

Samuelson, P. 2004. Where Ricardo and Mill Rebut and confirm arguments of mainstream economists supporting globalization. *Journal of Economic Perspectives* 18(3), 135–46.

Srinivasan, T.N. 1994. International labor standards once again! In *International Labor Standards and Global Economic Integration: Proceedings of a Symposium*, ed. G.K. Schoepfle and K.A. Swinnerton. Washington, DC: Bureau of International Labor Affairs, US Department of Labor.

Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85, 1029–46.

Zhu, S.C. and Trefler, D. 2005. Trade and inequality in developing countries: a general equilibrium analysis. *Journal of International Economics* 65, 21–48.

How to cite this article

Freeman, Richard B. "globalization and labour." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000253> doi:10.1057/9780230226203.0651

The New Palgrave Dictionary of Economics Online

globalization and the welfare state

Heinrich W. Ursprung
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Most scholarly investigations do not support the often heard claim that globalization impairs the welfare state: observed cuts in welfare programmes appear to be mainly driven by domestic factors. The empirical evidence supporting the converse claim – that globalization gains are used to compensate the losers from global economic integration – is, however, also inconclusive. In order to disentangle the multifaceted and potentially inconsistent globalization effects on the plethora of welfare state activities, future research will have to adopt a more explicit micro-orientation, better econometric techniques, and an empirical research strategy that is more firmly based on political economic theories.

Keywords

compensation effect of globalization; deindustrialization; diffusion processes; efficiency effect of globalization; globalization; globalization and the welfare state; labour market institutions; labour market risk; political economy; risk sharing; social insurance; trade openness; welfare state

Article

Globalization and the welfare state are commonly considered to be related; more precisely, many people believe that globalization exerts a negative influence on the size and scope of the welfare state. This contention has been examined in an impressive number of scholarly investigations. Since globalization has far-reaching effects on income distribution, this issue has, however, attracted not only social scientists but also all kinds of political entrepreneurs: well-meaning public figures concerned with the globalization-induced social dynamic, political demagogues vying for political support, and even street rioters.

The worries of the well-meaning objectors to global economic integration originate in the conviction that globalization will bring about a loss of power of the nation states in general. They argue that liberalization of international transactions renders tax bases increasingly footloose, which induces a global tax race to the bottom and, as a consequence, jeopardizes the nation states' ability to finance welfare state activities. This downward pressure on the supply side of public welfare programmes, depending on the viewpoint of the observer, reduces the *efficiency* of benevolent governments and/or *disciplines* egoistic governments that transform discretionary power into political support. At any rate, the so-called efficiency or discipline effect of globalization tends to reduce the size and scope of government welfare programs.

The opponents of globalization ignore, however, the demand side of the political market, which derives from governments' political support maximization motives to redistribute the gains from globalization; that is, the losers from globalization, in particular workers who become exposed to higher labour market risks, will to some extent be *compensated* via an extension of social welfare programmes. Reviving Ruggie's (1982) notion of ‘embedded liberalism’, Rodrik (1998) interprets this kind of compensation as an exchange of social insurance in return for public support for openness. The so-called ‘compensation’ effect thus counteracts the ‘efficiency’ effect, implying that, from a theoretical point of view, the total effect of globalization on the welfare state remains ambiguous.

The interaction of the two effects is summarized in Figure 1. The marginal benefit (in terms of political support) of welfare state expenditures decreases, whereas the marginal cost increases. Political support is maximized at the level where the MB and the MC curves intersect. A deepening or widening of economic integration now increases the marginal cost of supplying social welfare programmes and also increases the marginal benefit via increased demand, thereby shifting the two curves upwards (MC_0 to MC_1 and MB_0 to MB_1). Whether the resulting *efficiency effect* of globalization dominates the *compensation effect* or vice versa is a matter that can be resolved only by empirical research.

Figure 1
Effects of globalization on social welfare expenditures

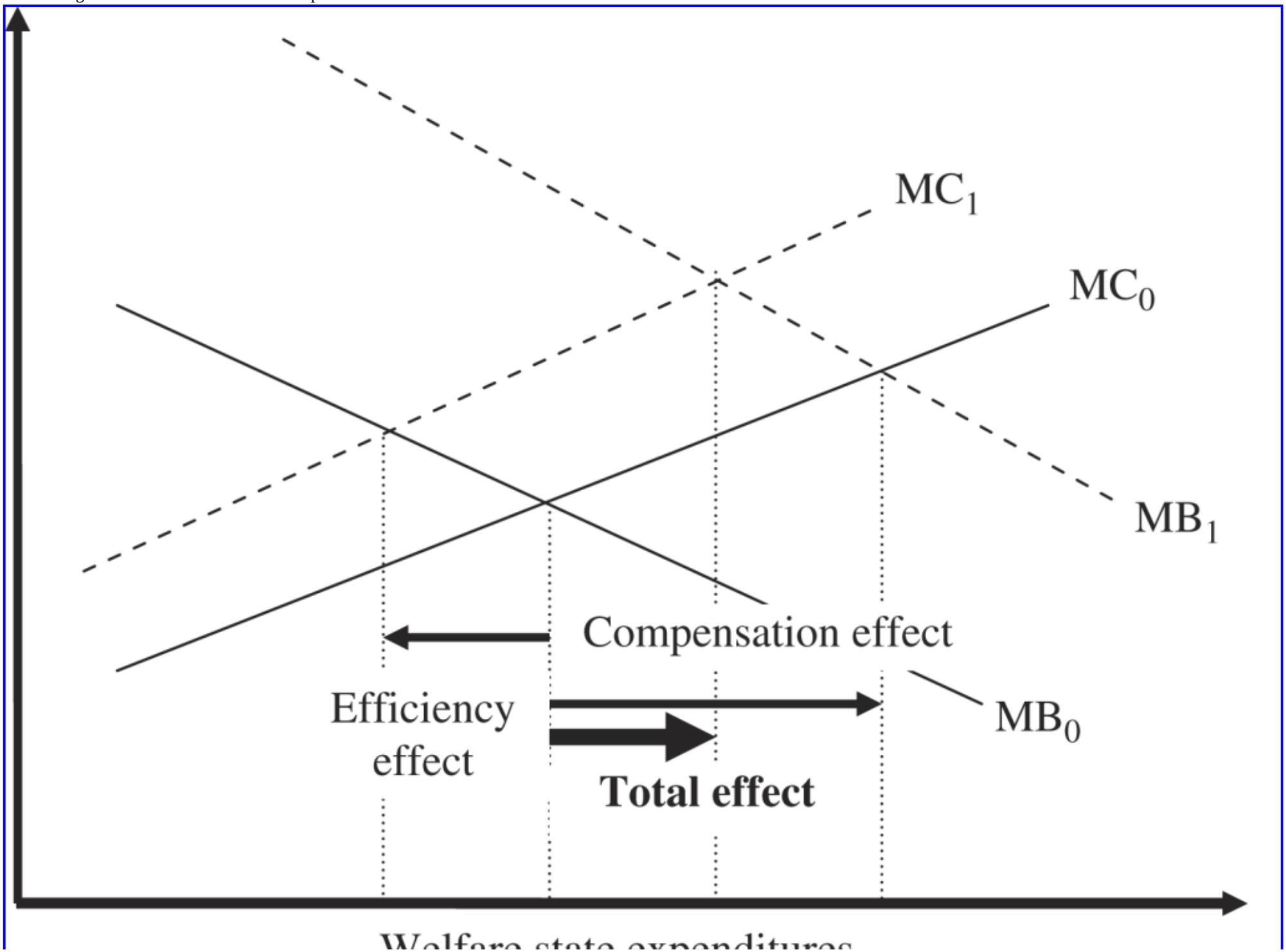

Welfare state expenditures

The first-generation literature on the globalization–welfare state nexus which appeared in the 1990s focused on this task of estimating globalization's net influence on government spending. Surveying this literature, Schulze and Ursprung (1999, pp. 345–7) arrive at the following conclusion:

> The general picture drawn by the few econometric studies available thus far does not lend any support to any alarmist view. At an aggregate level, many of these studies find no negative relationship between globalization and the nation states' ability to conduct independent fiscal policies.•…•no strong evidence points to a significant globalization-induced change of the level of public spending. But also accustomed expenditure patterns do not appear to have changed in the course of globalization. This may be due, however, to a lack of studies using strongly disaggregated public expenditure data.

In the meantime, many scholars have indeed taken up this implicit challenge and have used more disaggregated data; others have analysed specific groups of countries, have refined the empirical methods, or have investigated non-economic routes of influence. To foreshadow the main result of the second-generation literature, no unambiguous consensus has thus far emerged from these investigations. Instead, the new approaches have painted a multi-faceted picture that does not lend itself to a straightforward overall interpretation. Common to all empirical investigations is, however, the general research strategy which implies – usually in the framework of a panel data-set – the regression of some measure of welfare state activities on a set of explanatory variables which include globalization-related determinants, domestic economic and/or demographic determinants, and variables describing the domestic political–institutional setting.

As far as the dependent variable is concerned, one observes that measures of the size of the government sector (such as the ratio of total government expenditures to GDP) have been replaced by variables that better describe the size or scope of welfare policies; examples are expenditures on health, education, and social security (usually as a share of GDP) or net replacement rates of unemployment insurance. In addition, it has been argued that year-to-year policy adjustments and discontinuous policy shifts may be governed by different forces. Hicks and Zorn (2005), for example, identify welfare retrenchment events in OECD countries and find that they are not induced by direct globalization effects. They do, however, dampen subsequent globalization-induced increases of social spending. Factors that promote welfare policies may therefore (by doing so excessively and in ways that aggravate the policy's negative by-products) build up pressures for sudden policy reversals. Another novel way of looking at the welfare-state implications of globalization proposed by Hays, Ehrlich and Peinhardt (2005) bridges the micro–macro divide in the traditional literature by investigating the extent to which trade policy stances of voters depend on the welfare policies conducted in their respective home countries. It transpires that protectionist sentiments of workers in import-competing industries can indeed be reduced with welfare policies that provide some kind of insurance against the labour market risk that is supposed to be associated with international exposure.

The crucial independent variables capturing the influence of globalization have also become more precisely tailored over time. Traditional measures of globalization, such as total international trade (imports plus exports) as a share of GDP, financial deregulation indices, and foreign direct investment (FDI) as a share of GDP, are not sufficiently focused on the dark side of globalization which gives rise to compensatory demands. Better suited for this purpose are, for example, imports as a share of GDP, imports from low-wage countries as a share of total imports, and outward FDI flows as a share of GDP. A further refinement of these measures is based on the argument that flow shares are biased with respect to country size because small countries are by nature more open than larger ones. Bretschger and Hettich (2002) have therefore decided to work with a measure of trade openness which controls for country size. A more substantial refinement consists in taking non-economic dimensions of globalization into account. Dreher (2006) derives indices of three dimensions of globalization – namely, economic, political and social globalization – and arrives at the result that none of the three dimensions appears to have a significant impact on overall and social expenditures. Jahn (2006), in a similar attempt, introduces diffusion processes in order to capture globalization forces that transcend purely economic channels of influence. Diffusion mechanisms allow for learning and emulation processes which, in a setting of deepening globalization, become increasingly important since political-economic integration allows all political agents to better compare domestic policy efficiency with policy efficiency pursued in other countries. The diffusion variable employed in Jahn's study measures – for each country – the weighted average of the dependent variable of the other countries, where the weights represent closeness in terms of the respective bilateral trade intensity. This diffusion variable has a statistically significant influence on the social expenditure behaviour in the 16 OECD countries analysed, which may be interpreted to imply that globalization, via political yardstick competition, facilitates the adoption of best-practice policies.

Even though most of the second-generation studies do not support the efficiency hypothesis, there are notable exceptions. Garrett and Mitchell (2001), in particular, have claimed that ‘year to year’ increases in total trade are associated with less total government spending, less government consumption and lower security benefits as a share of GDP. This result has been construed to imply a preponderance of the efficiency effect over the compensation effect even though most of the financial globalization indicators employed in this particular study actually do not point into this direction and the share of low-wage country imports even appears to have a positive, compensatory, effect. More momentous than the daring interpretation of the study's results is, however, the critique that has been levelled against the econometric modelling. A whole series of recent studies have taken this controversial study as a starting point for scrutinizing the appropriate specification of panel data regression models of welfare state development (Kittel and Winner, 2005; Plümper, Manow and Troeger, 2005; and Podesta, 2006). What has unambiguously emerged from this deliberation is that panel data inferences react in a very sensitive manner to the chosen model specification and that the controversial Garrett–Mitchell estimates are driven by misspecifications. Re-estimating the relationship using the original variables but statistically better-

behaved models reveals that changes in government spending are primarily driven by the state of the domestic economy.

This finding is in line with the influential study by Iversen and Cusack (2000), who do not find any relationship between globalization and the level of labour-market risk, presumably because trade not only increases this risk via specialization, but at the same time also diversifies it across a larger market, implying that the net effect of increased trade exposure remains ambiguous. According to Iversen and Cusack, most of the uncertainties in modern societies originate from dislocations caused by technologically induced structural transformations, and it is this transformation towards deindustrialization that has spurred electoral demands for welfare state compensation and risk sharing. Demand for welfare state extensions thus appear to be largely home-made.

Most studies investigating the nexus between globalization and the welfare state focus on OECD countries, that is, countries that were blessed with democratic government in the period under review. Only recently have scholars begun to concern themselves with developing countries, which requires that one squarely address the crucial role of the political regime and the labour market institutions in accommodating the demand for welfare policies. After all, defending welfare benefits under the pressures of globalization is likely to be much easier in countries honouring civil and political rights than in countries in which the losers of globalization are politically powerless and poorly organized. Most studies indeed confirm that democracies are more responsive to compensation demands than autocratic regimes. Avelino, Brown and Hunter (2005), to name a prominent study on Latin America, arrive at a somewhat more differentiated conclusion: they confirm that democracies have a strong positive influence on social spending, but also point out that it is questionable whether this influence is reinforced by globalization-induced social insurance demands. Further analysis is needed to obtain a better understanding of how globalization and regime effects interact.

Even though the globalization–welfare state nexus has been the subject of intense empirical research since the early 1990s, the underlying relationship has remained rather elusive. To be sure, the second-generation studies corroborate the conclusions that have been drawn from the earlier studies, that is, the evidence certainly does not point towards an alarming globalization-induced ‘race to the bottom’. The available signs of compensatory welfare policy measures are, however, rather weak and inconclusive. What does this mean in the final analysis? Some scholars believe that globalization simply does not matter. Swank (2002, pp. 119–20), for example, squarely declares that ‘the conventionally hypothesized globalization dynamics are absent. Internationalization has no systematic impact on welfare policy change.’ On the other hand, one could argue that globalization gives rise to significant efficiency and compensation effects which, however, neutralize each other. In order to discriminate between these two views the macro-perspective that characterizes the bulk of the relevant literature does not seem to be helpful. Future research, therefore, will have to undergo a reorientation.

The leading scholars in the field no longer believe that globalization and the welfare state represent low-dimensional phenomena which are linked by a simple causal relationship. The ambiguous empirical results are often attributed to the fact that the nexus between globalization and the welfare state is more complex than the mechanism illustrated in Figure 1 suggests. This view implies that future research needs to focus even more on the interaction at the level of very specific globalization forces and on social policy responses which are just as narrowly specified. Nobody expects, however, that these micro-level interdependencies will add up to a sufficiently consistent macroeconomic response that would allow ostentatious claims of the sort that stimulated the early literature.

A second aspect that is likely to influence future research concerns econometric methods. To be sure, much progress has been made in perfecting empirical methods, but the struggle for better methods has certainly not yet come to an end. This is not meant to imply that the ambiguity of the results will disappear when proper empirical methods are applied across the board. On the other hand, it cannot be denied that methodological shortcomings may well be responsible for some of the observed discrepancies between studies that analyse closely related issues.

A final point concerns the fact that the empirical models employed almost invariably rest on ad hoc postulates. Even though political economy has become an integral part of mainstream economic thinking, theoretical arguments concerning the identity and the stakes of the involved political agents and a detailed investigation of how the strategic interaction of these interests shapes the ongoing political process have not found their way into the predominantly empirical literature on welfare state development. Investigations which are firmly based on political economy arguments would not only guide the empiricist in identifying worthwhile associations but also shift the focus of the discussion from addressing questions associated with political recriminations back to issues which are more securely rooted in the traditional scholarly discourse.

See Also

- globalization
- welfare state

Bibliography

Avelino, G., Brown, D. and Hunter, W. 2005. The effects of capital mobility, trade openness, and democracy on social spending in Latin America, 1980–1999. *American Journal of Political Science* 49, 625–41.

Bretschger, L. and Hettich, F. 2002. Globalisation, capital mobility and tax competition: theory and evidence for OECD countries. *European Journal of Political Economy* 18, 695–716.

Dreher, A. 2006. The influence of globalization on taxes and social policy – an empirical analysis for OECD countries. *European Journal of Political Economy* 22, 179–201.

Garrett, G. and Mitchell, D. 2001. Globalization, government spending and taxation in the OECD. *European Journal of Political Research* 39, 145–77.

Hays, J., Ehrlich, D. and Peinhardt, C. 2005. Government spending and public support for trade in the OECD: an empirical test of the embedded liberalism thesis. *International Organization* 59, 473–94.

Hicks, A. and Zorn, C. 2005. Economic globalization, the macro economy, and reversals of welfare: expansion in affluent democracies, 1987–94. *International Organization* 59, 631–62.

Iversen, T. and Cusack, T. 2000. The causes of welfare state expansion: deindustrialization or globalization? *World Politics* 52, 313–49.

Jahn, D. 2006. Globalization as ‘Galton's Problem’: the missing link in the analysis of diffusion patterns in welfare state development. *International Organization* 60, 401–31.

Kittel, B. and Winner, H. 2005. How reliable is pooled analysis in political economy? The globalization–welfare state nexus revisited. *European Journal of Political Research* 44, 269–93.

Plümper, T., Manow, P. and Troeger, V. 2005. Pooled data analysis in the comparative political economy of the welfare state: a note on methodology and theory. *European Journal of Political Research* 44, 327–54.

Podesta, F. 2006. Comparing time series cross-section model specifications: the case of welfare state development. *Quality & Quantity* 40, 539–59.

Rodrik, D. 1998. Why Do More Open Economies Have Bigger Governments? *Journal of Political Economy* 106, 997–1032.

Ruggie, J.G. 1982. International regimes, transactions, and change: embedded liberalism in the postwar economic order. *International Organization* 36, 379–415.

Schulze, G. and Ursprung, H.W. 1999. Globalization of the economy and the nation state. *World Economy* 22, 295–352.

Swank, D. 2002. *Global Capital, Political Institutions, and Policy Change in Developed Welfare States*. Cambridge: Cambridge University Press.

How to cite this article

Ursprung, Heinrich W. "globalization and the welfare state." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000212> doi:10.1057/9780230226203.0652

The New Palgrave Dictionary of Economics Online

globalization

William Easterly
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The passion that surrounds the vague term ‘globalization’ is best seen as a proxy for the long-standing debate about free-market capitalism. The zero-sum mindset, the difference between Pareto superiority and common norms of fairness, and the belief that all outcomes are caused by an intentional agent often cause communication problems between non-economists and free-market economists, who themselves often exaggerate what ‘free-market reforms’ can accomplish and endorse overly ambitious programmes of change (‘shock therapy’), underestimating problems of transition and the second best. Economists could try to understand the protests against ‘globalization’ rather than dismissing them out of hand.

Keywords

anti-capitalism; Asian miracle; banking crises; business networks; Calhoun, J. C.; Carlyle, T.; contract enforcement; creative destruction; development economics; dismal science; economic growth; financial liberalization; financial regulation; gains from trade; globalization; inequality; international trade; invisible hand; Lenin, V. I.; Outsourcing; poverty; poverty alleviation; reform consultants; second best; shock therapy; slavery; spontaneous order; structural adjustment; stylized facts; total factor productivity; Washington Consensus

Article

‘Globalization’ is a word that gets both its proponents and opponents very agitated. But what exactly is it? What is the globalization debate really about?
The answer is that the globalization debate is about a surprisingly large number of issues, including some that lie outside of economics. A non-exhaustive list of issues derived from a reading of the writings of both economists and non-economists (see a very partial list of references in the bibliography) follows:

1. 1. Liberalization versus regulation of international trade, capital movements, and migration.
2. 2. Market imperfections that arise with (either domestic or international) goods markets, capital markets, privatization, macroeconomic crises, intellectual property rights, and so on.
3. 3. Evaluation of the performance of the International Monetary Fund (IMF) and the World Bank, including in particular their policy prescriptions (the ‘Washington Consensus’, ‘shock therapy’, or ‘structural adjustment’).
4. 4. Effects of freer trade and capital movements on rich country workers (‘out-sourcing’) and on poor country workers (‘sweatshops’).
5. 5. Extreme world inequality and poverty.
6. 6. Capitalism (‘neoliberalism’) versus alternative systems.
7. 7. Westernization/Americanization versus local culture.
8. 8. Unequal distribution of political power between the West (both Western governments and corporations) and the Rest.
9. 9. Effect of global economic growth on the environment.
10. 10. Western imperialism and military intervention in the rest of the world.

Arguably, the vagueness of a term that includes at least ten separate debates has done a disservice to economic and political debate, causing many ‘globalization’ debate participants to think they disagree with people with whom they really agree, or to think they agree with people with whom they really disagree. It also explains some of the difficulties in communication between economists and non-economists about globalization, because the two groups really have different debates in mind. Economists (including those identified as ‘globalization critics’) have focused largely on issues 1–5, while the non-economists – though not ignoring 1–5 – seem to have something else in mind like 6–10.

For example, Dani Rodrik (1997) and Joseph Stiglitz (2002), who have both acquired a reputation as globalization critics by focusing mainly on issues 1–3, are embraced eagerly by some ‘globalization protesters’ whose main issue is really 6: the critique of capitalism (sometimes called ‘neoliberalism’). This is not meant as a criticism of Rodrik and Stiglitz; rather, it highlights the confusion that exists when two prominent mainstream economists who are talking about tinkering with and fine-tuning capitalist markets are seen as allies by those who are opposed to free market capitalism.

This article can hardly do justice to the complexity of all of these debates, nor is there much hope of getting everyone to discontinue the almost criminally vague use of the term ‘globalization’ in debate. The article argues that most of the energy in the debate indeed comes from the clash of attitudes – enthusiastic and antipathetic – towards capitalism and free markets.

This article thus focuses on two key themes about the globalization debate. First, I give some intellectual history of the debate about capitalism (issue 6), which will place in perspective some of today's globalization debate including that by the non-economists. This has the objective of dispelling some of the puzzlement that many economists feel about the sound and fury surrounding globalization, through realization that it is partly just another manifestation of a long intellectual debate about capitalist free markets, which economists have been engaged in for decades if not centuries. Second, the article tries to place the antipathy towards free markets in contemporary perspective by discussing whether overly simplistic models and unrealistic promises of quick and sizeable results from ‘globalization’ for poor countries have further fuelled this antipathy. I consider at the same time whether the zeal of the globalizers may have led them to endorse counterproductive and unrealistic attempts at wholesale social transformation, which generate an even more severe backlash.

Let's start with the long-standing debate about capitalism. Intellectual history makes clear how the gains from trade (in goods, finance, and labour services) under capitalism amount to such a revolutionary idea that economists are often its lone proponents in the wilderness. There are three major habits of thinking that create difficulty in communication between economists and non-economists on gains from trade. One is the mindset that holds that economic interactions are zero-sum games (a partially understandable mindset when capitalists have such skeletons in the closet as military conquest, colonization, slavery, predatory behaviour by firms, and so on). The second is the difference between economists' notion of Pareto-superior outcomes and common social norms of fairness. The third barrier to communication is the difficulty of accepting the economists' notion of the invisible hand that creates spontaneous outcomes not designed by anyone, where the common habit of thinking is that a good or bad outcome must be the result of intentional action by a good or bad agent.

To start with the zero-sum mindset, one early father of Christianity, St. Jerome, thought any wealth was automatically ‘unjust riches’, since ‘no one can possess them except by the loss and ruin of others’. St. Augustine put it more tersely: ‘If one does not lose, the other does not gain’ (quoted by Muller, 2002, p. 6).

Centuries later, even after Adam Smith and the Industrial Revolution, both sides of the political spectrum still often thought in zero-sum terms. Friedrich Engels wrote that ‘the consequences of the factory system’ were ‘oppression and toil for the many, riches and wealth for the few’ (quoted by Muller, 2002, p. 180). Henry Adams saw capitalism as a system that divided humanity ‘into two classes, one which steals, the other which is stolen from’ (quoted by Herman, 1997, p. 160).

We are so used to thinking of conservatives as pro-market that it surprises us that some on the Right in the 19th century also attacked free market economics (see Levy, 2001, for a fine narrative). The Right's attack on the laissez-faire Left (how things have changed!) was that the latter were hypocritical advocating both capitalism and the end of slavery, because capitalism made ‘free’ workers no better than slaves. For example, Thomas Carlyle (the man who disliked economists so much that he coined the phrase ‘dismal science’) told workers: ‘you are fallen captive to greedy sons of profit-and-loss; to bad and ever to worse … Algiers, Brazil or Dahomey hold nothing in them so authentically *slave* as you are’ (Carlyle, 1850). This is zero-sum thinking in the extreme!

Similarly, on the American 19th-century Right, John C. Calhoun defended American slavery in 1828 by claiming that industrial capitalism in the North was no better; it caused wages to ‘sink more rapidly than the prices of the necessaries of life, till the operatives … portion of the products of their labor … will be barely sufficient to preserve existence’ (quoted by Muller, 2002, p. 177). Ironically the great African-American intellectual W. E. B. Du Bois, reached similar conclusions to Calhoun's about industrial capitalism, as he observed it in the South after the Civil War:

> [The] men who have come to take charge of the industrial exploitation of the New South…thrifty and avaricious Yankees•…•. For the laborers as such, there is in these new captains of industry neither love nor hate, neither sympathy nor romance; it is a cold question of dollars and dividends. Under such a system all labor is bound to suffer.•…•The results among them, even, are long hours of toil, low wages, child labor, and lack of protection against usury and cheating. (Du Bois, 1903)

Lenin famously linked zero-sum Western imperialism and non-zero-sum trade and capital flows. Profits for the companies that follow in the wake of the imperialists are high in the ‘backward countries’, where the capitalists relocate their capital because ‘the price of land is relatively low, wages are low, raw materials are cheap’ (Lenin, 1917). Lenin may have

been the first 20th-century critic of outsourcing.

Today, we see similar zero-sum thinking in globalization critics on the Left and the Right. Oxfam GB (2004, p. 12) identifies such products as Olympic sportswear as forcing labourers into ‘working ever-faster for ever-longer periods of time under arduous conditions for poverty-level wages, to produce more goods and more profit’ (Statements like this come from an organization that is actually much friendlier to free trade than most non-governmental organizations.)

Global Policy Forum, a popular globalization website elaborates: ‘trade is inherently unequal and poor countries seldom experience rising well-being but increasing unemployment, poverty, and income inequality.’ Former Tanzanian President Julius Nyerere summarized the zero-sum mindset back in a 1975 state visit to Britain: ‘I am poor because you are rich’ (quoted in Lindsey, 2001, p. 105).

On the Right, there is still today concern about free markets creating winners at the expense of losers. Patrick Buchanan claimed in a 1998 book that free trade causes ‘broken homes, uprooted families, vanished dreams, delinquency, vandalism, crime’ (quoted in Micklethwait and Wooldridge, 2000, p. 282). Edward Luttwak claimed that global capitalism requires ‘harsh laws, savage sentencing, and mass imprisonment’ to deal with ‘disaffected losers’ (1999, p. 236). Although of course the Right in general is today more sympathetic to free market capitalism than the Left, the persistence of this thinking shows how the zero-sum mindset is an independent force from political ideology.

Of course, capitalism/globalization does create losers as well as winners, unleashing gales of creative destruction. Since losers tend to be more vocal than winners, it is easy to understand the perception that the losers outnumber the winners, which then reinforces the already ingrained habit of thinking in zero-sum terms. To complicate matters further, some poorly conceived attempts at rapid transition from non-capitalism to capitalism (for example, ‘shock therapy,’ to be discussed below) can in fact create more losers than winners.

The second source of communication breakdown about globalization is the difference between economists' general enthusiasm for Pareto improvements and common norms of fairness (see Aisbett, 2005, for a provocative discussion). Following Aisbett, let's say for example that a multinational firm opens a factory in a low-income country. Suppose that the new investment enables the firm to double its profits and the newly employed workers in the factory to double their previous incomes. Suppose the workers were formerly part of the extreme poor (conventionally measured as an income of a dollar a day), so that now they have escaped extreme poverty. Who can argue with such a Pareto improvement?

From another perspective, however, what is happened is that a very poor person has gained a dollar (in a ‘sweatshop’) while a captain of industry previously making, say, $1,000 a day has gained another $1,000. It violates many norms of fairness (abundantly confirmed in the laboratory by experimental economics), when someone already far better off gains 1,000 times more than the less fortunate person from this transaction. To point this out doesn't lead to any obvious conclusions – most economists will say the transaction is still worth doing to relieve absolute poverty, while critics will protest that a more fair division of the gains should be possible (but even if the worker gets a fivefold increase in income – an amazing escape velocity from poverty – while the capitalist just doubles his income, the capitalist's gains are still 250 times larger).

The third barrier to constructive communication about globalization is the common assumption that an outcome must result from an intentional action by an identifiable agent. This couldn't contrast more with the economists' notion of the invisible hand. The intentionality mindset is that ‘globalization’ represents someone's agenda, and it is to blame for the tragedies of world poverty. To give an illustrative example of this kind of anti-globalization rhetoric (not necessarily representative): the ‘transnational corporations …expand, invest and grow, concentrating ever more wealth in a limited number of hands. They work in coalition to influence local, national and international institutions’. ‘Corporate elites…forge common agendas outside the formal institutions of democracy.’ They use forums such as the Trilateral Commission, the International Chamber of Commerce, the World Economic Forum, trade associations, and the many national and international business and industrial roundtables’ (IFG, 2002, p. 140). The participants at such ‘posh gatherings…chart the course of corporate globalization in the name of private profits …’ (IFG, 2002, p. 4). Searching for whom to blame for the miseries of the poor, the rich multinational corporations make for natural villains (alternative villains are the IMF and World Bank). These are not only villains, but foreign villains! This mindset is further strengthened when corporations (who of course really are self-interested profit-seekers) get caught doing something like despoiling the environment in a poor country or doing shady deals with the local kleptocrats. The economists' idea of a spontaneous system of myriads of uncoordinated agents, with nobody in charge, generating outcomes that are not intended (or even forseeable) by anyone is a lot harder sell.

With such fundamental differences in thinking, perhaps we can understand why there is little prospect of a constructive conversation between advocates and opponents of globalization/free market capitalism/neoliberalism. The World Social Forum, the counterpoint annual meeting to that of the capitalist globalizers at the World Economic Forum in Davos, says in its charter that it is ‘an open meeting place for reflective thinking, democratic debate of ideas, free exchange of experiences’, except that the debate is limited to ‘groups and movements of civil society that are opposed to neoliberalism’. A similar spirit seems to inform the complaint that the case for capitalism arises from ‘rationalist constructions of knowledge’ featured in such reunions of the ‘global managerial class’ as ‘AEA conventions’ (Global Policy Forum, 2006). Of course, mainstream economists probably do not seem to their critics much more open to debate on ‘neoliberalism’!

Things are made even worse by the second major theme of this article, the overselling of globalization. Simplistic models and promises of quick and sizeable results create expectations, and when these expectations are disappointed (even when the results are gradually and increasingly positive), there is a backlash against globalization.

A classic example of the overselling of globalization is the World Bank (2002) report *Globalization, Growth, and Poverty*. The following graph (Figure 1), the first one shown in the report, is prominently displayed in the overview (2002, p. 5):

Figure 1

Divergent paths of developing countries in the 1990s

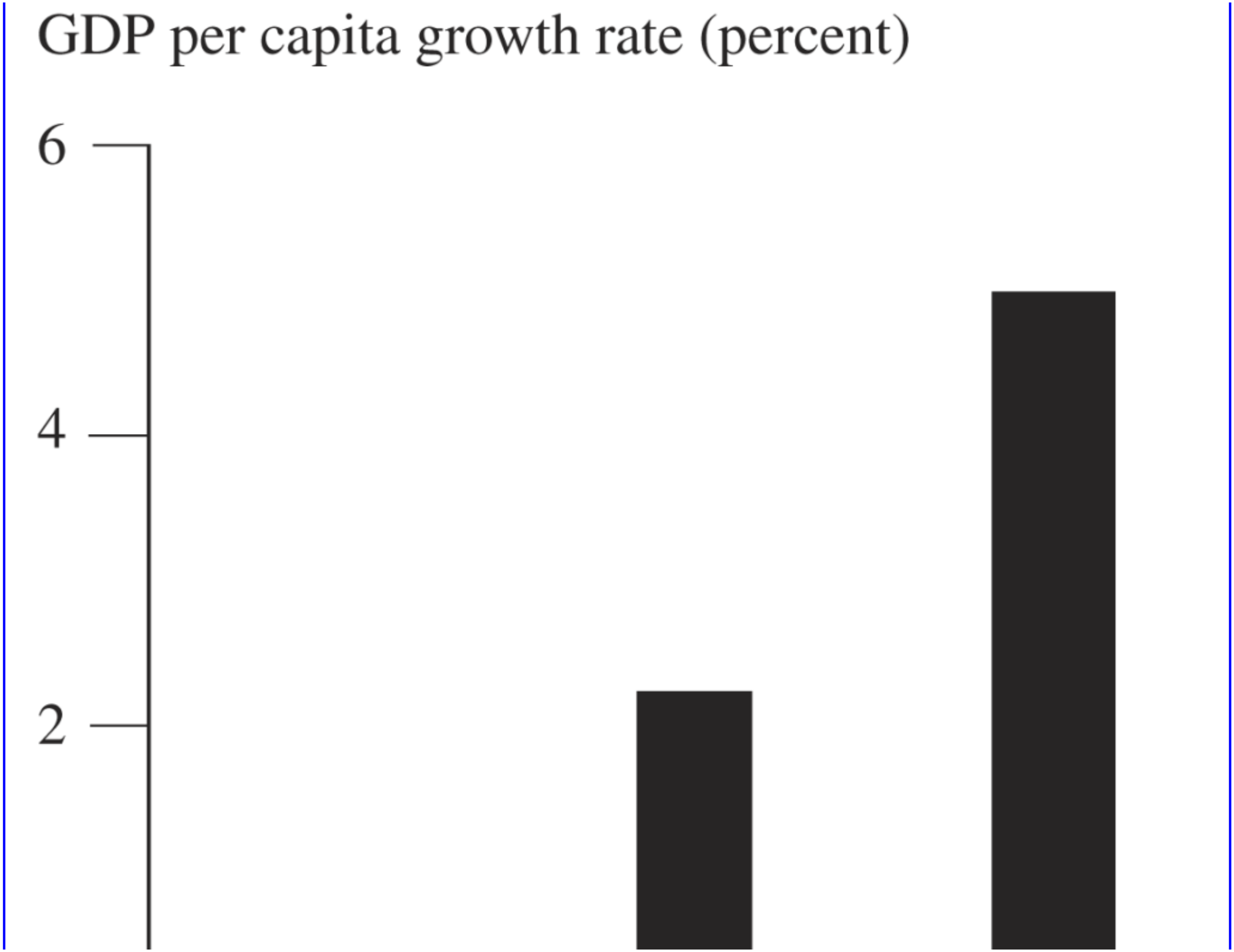
GDP per capita growth rate (percent)
6
4
2

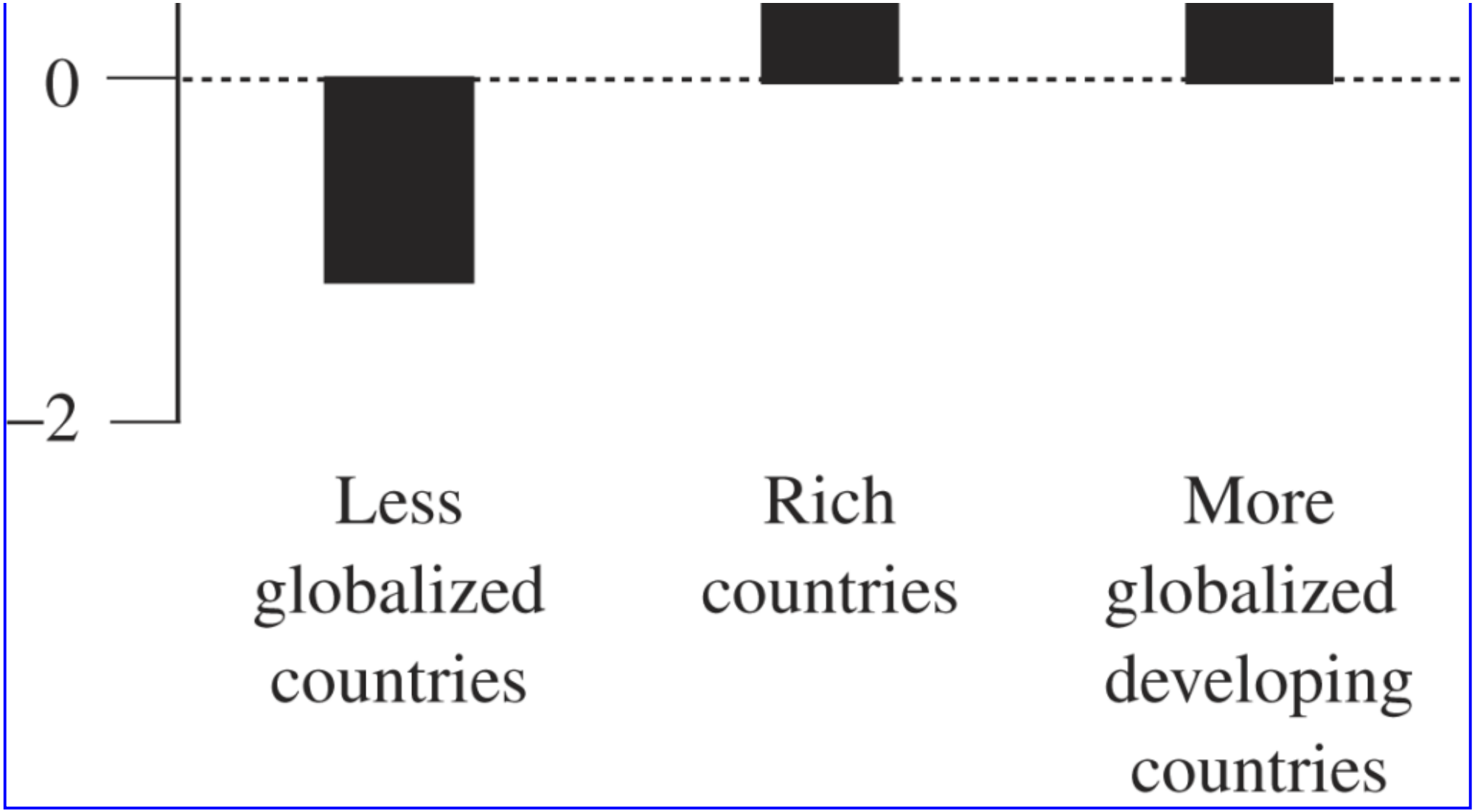

Although never explicitly stated and some caveats are expressed, the impression left with many readers is that being more globalized makes the difference between five per cent per capita growth and minus one per cent per capita growth, which is an amazingly strong claim for the effects of globalization. World Bank researchers reinforce this kind of claim with statements promising that world poverty can be cut in half with policy reforms: ‘Poverty reduction – in the world or in a particular region or country – depends primarily on the quality of economic policy. Where we find in the developing world good environments for households and firms to save and invest, we generally observe poverty reduction’ (Collier and Dollar, 2001). (I have to admit with some embarrassment that this statement was based on one of my own unpublished growth regressions, which eventually showed up in published form in Easterly 2001 making the opposite point – that the growth response to policy reform was disappointing. Regressions can be dangerous!)

The IMF likewise has a standard set of policies that it advocates (together, the IMF's and the World Bank's notion of ‘good policies’ form what is often called the ‘Washington Consensus’), many of which are oriented towards creating freer markets (more ‘globalization’). The IMF also claimed that ‘Where [good] policies have been sustained, they have raised growth and reduced poverty’ (2000).

This is speculation, but some of the World Bank/IMF belief in policy reforms to explain good outcomes may ironically stem from the same intentionality impulse that makes critics blame the World Bank and IMF for bad outcomes. People find it more comfortable to attribute success to the action of a few heroic policy reformers or technocrats (or strong leaders implementing good policies, like Singapore's Lee Kuan Yew), rather than to some more mysterious bottom-up process of many spontaneous individual entrepreneurs.

Unfortunately, there is little evidence for strong growth effects of policy changes that involve anything less than getting rid of self-destructive extremes (like moving from autarchy to

allow some trade, or from hyperinflation to moderate inflation), and even then hardly six percentage points of permanent change in growth, as documented in Easterly (2005).

Contrary to the impression conveyed in the foregoing statements, the economics profession actually knows very little about how to raise economic growth over the short to medium run with policy changes in the range in which most countries are operating (see for example the survey in Kenny and Williams, 2001). (Besides this, the methodological problem is that countries that are more or less globalized are not defined in terms of policies that promote free trade, free capital movements, free migration, or some other policy measure that features in the debates on globalization. The 'more globalized' countries are defined in terms of outcomes: it is those that are in the top third of countries in terms of the increase in their trade-to-GDP ratios. Defining globalization in terms of one endogenous measure of success that is likely related to other endogenous measures of success – like the GDP growth rates being explained – is rather unfortunate.) Growth in developing countries is extremely volatile (on average 75 per cent of a country's deviation from the global mean per capita growth in a five- or ten-year period disappears in the following period, as pointed out in Easterly et al., 1993, since replicated with more recent data.) Overeager growth-watchers are too quick to proclaim 'growth miracles' and the lessons that allegedly follow from them. As Dixit (2006, p. 23) says,

> At any time, some country is doing well, and academic as well as practical observers are tempted to generalize from its choices and recommend the same to all countries. After a decade or two, this country ceases to do so well, some other country using some other policies starts to do well, and becomes the new star that all countries are supposed to follow.

The success of China and the earlier successes of the East Asian miracles (all associated with great success in global markets) are often used by promoters of globalization to bolster their case. Unfortunately, the implicit promise that such unusually rapid growth (on the order of five per cent per capita) is available to all 'globalizers' rests on very shaky ground. First, such rapid growth is very rare – 1.7 per cent of countries registered five per cent per capita growth or more over 1950–2001, and only 0.7 percent of all half-century country per capita growth episodes since 1820 surpassed five per cent (Maddison, 2003). Most rich countries today (usually agreed to have been globalized and capitalist for quite some time) got to be rich by registering something on the order of two per cent per capita growth for one or two centuries.

Things are made worse when casual empiricism is married to simplistic theory. In the simplest textbook model, freedom of trade and capital movements each promotes poverty alleviation, that is, rapid catching up of wages or incomes in capital-scarce, labour-abundant poor countries to wages or incomes in rich countries. According to the model, free trade allows unskilled wages in poor countries to rise rapidly through labour-intensive exports, while free capital movements allow high investment in poor countries to remedy the gap in capital per worker between rich and poor countries. A side effect would be that inequality within the poor country (driven mainly by the differences between labour and capital earnings) should decrease. Even aside from the fact that a vast trade literature does not support most predictions of the first story, the growth and development literature has pointed out that total factor productivity differences between countries are a much more plausible explanation of income differences between countries than differences in capital per worker (Hall and Jones, 1999; Klenow and Rodriguez-Clare, 1997; Easterly and Levine, 2001; Hsieh, 2002). Stylized facts on trade, inequality, and poverty do not support the predictions of the simple textbook model where income differences are due to differences in capital per worker. (See Easterly, 2006, for a more extensive discussion of these points.)

These false expectations of very rapid growth through globalization (or 'free markets' in general) have arguably done a lot of damage, creating fertile ground for an anti-capitalist backlash in places as diverse as Thailand, South Africa, Russia, Bolivia, Venezuela, Peru, Argentina, Ecuador and Mexico. The critics of globalization can all the more easily seize upon any growth setbacks (such as the Mexico crisis of 1994/95, the East Asian crisis of 1997/98, or the Argentine crisis of 2001, or disappointing growth in Latin America in general since market liberalization in the 1980s), whatever their cause (usually hard to explain anyway in the volatile pattern described earlier), to say 'see, globalization/neoliberalism doesn't work'.

The backlash has been made all the worse because of the overconfidence of IMF and World Bank policymakers (and freelance 'reform consultants') 'globalizing' whole societies that start out with many different barriers to efficient free markets. The economics profession can demonstrate fairly convincingly that some long-run market-friendly policies and institutions are most conducive to prosperity, but it knows very little about the sequencing and the transitional paths of reforms to get from initial conditions to that ideal state. (Lipsey and Lancaster's, 1956–57, theory of the second best recognized this problem, but it seems that each new generation must discover it afresh.) It is obvious that different kinds of reforms are complementary to each other – for example, financial market liberalization works well only if there is sufficient transparency of banks to depositors, and a good regulatory and supervisory framework to ensure that banks don't cheat (Barth, Caprio and Levine, 2006). Otherwise, financial liberalization often leads to bad loans, enrichment of insiders, and subsequent banking system crises, as abundant experience has already demonstrated.

Yet the usual answer to policy complementarity – 'do everything at once' 'structural adjustment,' or 'shock therapy' – doesn't really escape the curse of the second-best. Policymakers neither know what 'everything' is nor have the ability to change 'everything' at once (or any time soon). The choice is really between large-scale partial reforms (which shock therapy mislabels 'comprehensive reform') and small-scale partial reforms. Any economy is a complex system of informal networks, social norms, relationships, trades, and formal institutions, many of which lie outside the control of the policymaker. As Dixit (2004) points out, an existing network under the current system of rules can at least enforce contracts in that it can threaten to expel any member who cheats another member. Drawing up a brand-new set of rules overnight (like moving abruptly from an interventionist economy to a free-market economy) can have perverse impacts in the short run. It can mean that people can choose to exit the old network (cheating their old partners) because they now have the

option of operating under the new system of rules and the new networks generated by the new rules. The net effect can be to disrupt the functioning of the old economy much more than it facilitates the creation of the new economy. This is theoretical speculation at this point, but it does illustrate the potential pitfalls of promising rapid results from rapid reforms. To think that economists could re-engineer the whole society and economy looks in retrospect like the worst kind of intellectual hubris (see McMillan, 2007, for a great discussion).

Attempts at rapid ‘comprehensive’ reform of poor countries look a lot like what Karl Popper long ago decried as ‘utopian social engineering’ versus what has worked for most rich countries to attain prosperity: ‘piecemeal democratic reform’.

The most notorious case of this hubris was the attempt to reform the former Soviet Union with ‘shock therapy’. Murrell (1992; 1993) – a long-time scholar of centrally planned economies – argued against shock therapy as utopian social engineering. His objections are all the more compelling because they were *ex ante* rather than *ex post*. History vindicated his scathing description of shock therapy at the time:

> There is complete disdain for all that exists … History, society, and the economics of present institutions are all minor issues in choosing a reform program… Establishment of a market economy is seen as mostly involving destruction… shock therapists assume that technocratic solutions are fairly easy to implement… One must reject all existing arrangements•…•(Murrell, 1993)

Murrell was quick to realize the relevance of Popper for what was later half-jokingly called a Leninist push for free markets in Russia. His quote from Popper in 1992 is a perfect prediction of how Russian reform would fail: ‘It is not reasonable to assume that a complete reconstruction of our social system would lead at once to a workable system’ (quoted in Murrell, 1992). After the former Soviet republics experienced some of the greatest depressions in economic history, the prescience of such viewpoints became apparent.

For its part, IMF- and World Bank-supported ‘structural adjustment’ was also uncomfortably like ‘utopian social engineering’, and produced a similar debacle in Africa and Latin America. The resulting anti-market/anti-globalization backlash (in the former Soviet Union as well as Africa and Latin America) was all the more severe because the reforms involved some IMF/World Bank coercion through conditional loans. One can hardly think of a better formula for an anti-capitalist backlash in poor countries than to introduce overambitious, oversold programmes of large-scale ‘globalization’ reforms imposed by foreigners!

In conclusion, economists are unlikely to find the term ‘globalization’ a precise enough concept to advance most research agendas. Instead, it mainly seems to point to the long-standing debate about economists' traditional embrace of free-market capitalism. Perhaps some progress in these debates can be made by understanding some of the traditional mindsets that make a system of spontaneous gains from trade such a revolutionary concept. It also would help a great deal if policymakers and the economists advising them did not pursue overambitious attempts at rapid wholesale transformation of the economy and society, and did not exaggerate the likely size and speed of the gains for the economy from such programmes.

Articulate arguments of the case for capitalism/globalization continue to be made in books such as Lindsey (2001) and Wolf (2004). Mishkin (2006) has recently made a fascinating case for the kind of globalization that opponents find most frightening (and even many economists shy from), financial globalization. Despite such eloquent statements, the discomforts caused by the spectre of globalization are unlikely to abate any time soon. Economists can arguably contribute more to the debate by seeking to understand the discomfort rather than dismissing it out of hand.

Bibliography

Aisbett, E. 2005. Why are the critics so convinced that globalization is bad for the poor? Working Paper No. 11066. Cambridge, MA: NBER.

Barth, J.R., Caprio, G. and Levine, R. 2006. *Rethinking Bank Regulation: Till Angels Govern*. Cambridge: Cambridge University Press.

Carlyle, T. 1850. The present time. In *Latter-Day Pamphlets*. London: Chapman & Hall.

Collier, P. and Dollar, D. 2001. Can the world cut poverty in half? How policy reform and international aid can meet international development goals. *World Development* 29, 1767–1802.

Dixit, A.K. 2004. *Lawlessness and Economics: Alternative Modes of Governance*. Princeton, NJ: Princeton University Press.

Dixit, A.K. 2006. Evaluating recipes for development success. Policy Research Working Paper No. 3859. Washington, DC: World Bank.

Du Bois, W.E.B. 1903. Of the sons of master and man. Ch. 9 in *The Souls of Black Folk*. New York: Bartleby.com, 1999.

Easterly, W. 2001. The lost decades: explaining developing countries' stagnation in spite of policy reform 1980–1998. *Journal of Economic Growth* 6, 135–57.

Easterly, W. 2005. National policies and economic growth: a reappraisal. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North Holland.

Easterly, W. 2006. Globalization, poverty, and all that: factor endowment versus productivity views. In *Globalization and Poverty*, ed. A. Harrison. Chicago: University of Chicago Press.

Easterly, W. and Levine, R. 2001. It's not factor accumulation: stylized facts and growth models. *World Bank Economic Review* 15(2), 177–219.

Easterly, W., Kremer, M., Pritchett, L. and Summers, L. 1993. Good policy or good luck? Country growth performance and temporary shocks. *Journal of Monetary Economics* 32, 459–83.

Global Policy Forum. 2006. Political struggle will determine better globalization. 15 March. Online. Available at http://www.globalpolicy.org/globaliz/define/2006/03scholte.htm, accessed 2 January 2007.

Global Policy Forum. 2007. International trade and development. Online. Available at http://www.globalpolicy.org/socecon/trade/index.htm, accessed 2 January 2007.

Hall, R.E. and Jones, C.L. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114, 83–116.

Herman, A. 1997. *The Idea of Decline in Western History*. New York: Free Press.

Hsieh, C.T. 2002. What explains the industrial revolution in East Asia? Evidence from the factor markets. *American Economic Review* 92, 502–26.

IFG (International Forum on Globalization). 2002. *Alternatives to Economic Globalization: A Better World is Possible*. San Francisco: Berrett-Koehler Publishers.

IMF (International Monetary Fund). 2000. *Policies for Faster Growth and Poverty Reduction in Sub-Saharan Africa and the Role of the IMF*. Issues Brief. Washington, DC.

Kenny, C. and Williams, D. 2001. What do we know about economic growth? Or, why don't we know very much? *World Development* 29(1), 1–22.

Klenow, P. and Rodriguez-Clare, A. 1997. The neoclassical revival in growth economics: has it gone too far? *NBER Macroeconomics Annual 1997* 12, 73–103.

Lenin, V.I. 1917. *Imperialism, The Highest Stage of Capitalism*. Online. Available at http://www.marxists.org/archive/lenin/works/1916/imp-hsc/ch04.htm, accessed 2 January 2007.

Levy, D.M. 2001. *How the Dismal Science Got its Name: Classical Economics and the Ur-text of Racial Politics*. Ann Arbor: University of Michigan Press.

Lindsey, B. 2001. *Against the Dead Hand: The Uncertain Struggle for Global Capitalism*. New York: Wiley.

Lipsey, R.G. and Lancaster, K. 1956–1957. The general theory of second best. *Review of Economic Studies* 24(1), 11–32.

Luttwak, E. 1999. *Turbo-Capitalism: Winners and Losers in the Global Economy*. New York: HarperCollins.

Maddison, A. 2003. *The World Economy: Historical Statistics*. Paris: OECD.

McMillan, J. 2007. Avoid hubris. In *Reinventing Foreign Aid*, ed. W. Easterly. Boston, MA: MIT Press.

Micklethwait, J. and Wooldridge, A. 2000. *A Future Perfect: The Challenge and Hidden Promise of Globalization*. New York: Crown Business.

Mishkin, F.S. 2006. *The Next Great Globalization: How Disadvantaged Nations Can Harness their Financial Systems to Get Rich*. Princeton: Princeton University Press.

Muller, J.Z. 2002. *The Mind and the Market: Capitalism in Modern European Thought*. New York: Alfred A. Knopf.

Murrell, P. 1993. What is shock therapy? What did it do in Poland and Russia? *Post-Soviet Affairs* 9(2), 111–40.

Murrell, P. 1992. Conservative political philosophy and the strategy of economic transition. *East European Politics and Societies* 6(1), 3–16.

Oxfam, GB. 2004. *Play Fair at the Olympics*. Oxford: Oxfam BG.

Rodrik, D. 1997. *Has Globalization Gone Too Far*? Washington, DC: Institute for International Economics.

Stiglitz, J. 2002. *Globalization and its Discontents*. New York: Norton.

Wolf, M. 2004. *Why Globalization Works*. New Haven, CT: Yale University Press.

World Bank. 2002. *Globalization, Growth, and Poverty: A Policy Research Report*. Washington, DC: World Bank.

World Social Forum. 2007. Charter. Online. Available at http://www.portoalegre2002.org/default.asp, accessed 30 January 2007.

How to cite this article

Easterly, William. "globalization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000178>
doi:10.1057/9780230226203.0650

The New Palgrave Dictionary of Economics Online

gold standard

Lawrence H. Officer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The world has had two experiences with gold standards: the classical gold standard and the interwar gold standard. The ‘rules of the game’, government policies to preserve the gold standard, were rarely followed. Rather, government responsible policy and credible commitment to the standard, private stabilizing arbitrage and speculation, and stable political and economic environment made the classical gold standard a success. The absence of these elements and the presence of the Great Depression combined to make the interwar gold standard a failure.

Keywords

arbitrage; Bank of France; bank rate; Bank Restriction Period; banking crises; bimetallism; bullion; bullion standard; central banking; commitment; convertibility; current account deficits; deflation; exchange controls; Federal Reserve System; fixed exchange rates; floating exchange rates; gold standard; gold-exchange standard; Great Depression; monetary base; money multiplier; money supply; orthodox metallism; Reichsbank; reserve ratio; silver standard; specie-flow mechanism; speculation; sterilization; sticky prices; sticky wages

Article

The classical gold standard (which ended in 1914) and the interwar gold standard are examined within the same framework, but their experiences are vastly different.

Types of gold standard

All gold standards involve (*a*) a fixed gold content of the domestic monetary unit, and (*b*) the monetary authority both buying and selling gold at the mint price (the inverse of the gold content of the monetary unit), whereupon the mint price governs in the marketplace. A ‘coin’ standard has gold coin circulating

as money. Privately owned bullion (gold in form other than domestic coin) is convertible into gold coin, at (approximately) the mint price, at the government mint or central bank. Private parties may melt domestic coin into bullion – the effect is as if coin were sold to the monetary authority for bullion. The authority could sell gold bars directly for coin, saving the cost of coining.

Under a pure coin standard, gold is the only money. Under a mixed standard, there are also notes issued by the government, central bank, or commercial banks, and possibly demand deposits. Government or central-bank notes (and central-bank deposit liabilities) are directly convertible into gold coin at the fixed price on demand. Commercial-bank notes and demand deposits are convertible into gold or into gold-convertible government or central-bank currency. Gold coin is always exchangeable for paper currency or deposits at the mint price. Two-way transactions again fix the currency price of gold at the mint price.

The coin standard, naturally ‘domestic’, becomes ‘international’ with freedom of international gold flows and of foreign-exchange transactions. Then the fixed mint prices of countries on the gold standard imply a fixed exchange rate (mint parity) between their currencies.

A ‘bullion’ standard is purely international. Gold coin is not money; the monetary authority buys or sells gold bars for its notes. Similarly, a ‘gold-exchange’ standard involves the monetary authority buying and selling not gold but rather gold-convertible foreign exchange (the currency of a country on a gold coin or bullion standard).

For countries on an international gold standard, costs of importing and exporting gold give rise to ‘gold points’, and therefore a ‘gold-point spread’, around the mint parity. If the exchange rate, number of units of domestic per unit of foreign currency, is greater (less) than the gold export (import) point, arbitrageurs sell (purchase) foreign currency at the exchange rate and also obtain (relinquish) foreign currency by exporting (importing) gold. The domestic-currency cost of the transaction per unit of foreign currency is the gold export (import) point; so the ‘gold-point arbitrageurs’ receive a profit proportional to the exchange-rate/gold-point divergence. However, the arbitrageurs’ supply of (demand for) foreign currency returns the exchange rate to below (above) the gold export (import) point. Therefore perfect arbitrage would keep the exchange rate within the gold-point spread. What induces gold-point arbitrage is the profit motive and *the credibility of the monetary-authorities’ commitment* to (*a*) the fixed gold price and (*b*) freedom of gold and foreign-exchange transactions.

A country can be effectively on a gold standard even though its legal standard is bimetallism. This happens if the gold–silver mint-price ratio is greater than the world price ratio. In contrast, even though a country is legally on a gold standard, its government and banks could ‘suspend specie payments’, that is, refuse to convert their notes into gold; so that the country is in fact on a ‘paper standard’.

Countries on the classical gold standard

Britain, France, Germany and the United States were the ‘core countries’ of the gold standard. Britain was the ‘centre country’, indispensable to the spread and functioning of the standard. Legally bimetallic from the mid-13th century, Britain switched to an effective gold standard early in the 18th century. The gold standard was formally adopted in 1816, ironically during a paper-standard regime (Bank Restriction Period). The United States was legally bimetallic from 1786 and on an effective gold standard from 1834, with a legal gold standard established in 1873–4 – also during a paper standard (the greenback period). In 1879 the United States went back to gold, and by that year not only the core

countries but also some British dominions and non-core western European countries were on the gold standard. As time went on, a large number of other countries throughout the globe adopted gold; but they (along with the dominions) were in ‘the periphery’ – acted on rather than actors – and generally (except for the dominions) not as committed to the gold standard.
Almost all countries were on a mixed coin standard. Some periphery countries were on a gold-exchange standard, usually because they were colonies or territories of a country on a coin standard.
In 1913, the only countries not on gold were traditional silver-standard countries (Abyssinia, China, French Indochina, Hong Kong, Honduras, Morocco, Persia, Salvador), some Latin American paper-standard countries (Chile, Colombia, Guatemala, Haiti, Paraguay), and Portugal and Italy (which had left gold but ‘shadowed’ the gold standard, pursuing policies as if they were gold-standard countries, keeping the exchange rate relatively stable).

Elements of instability in classical gold standard

Three factors made for instability of the classical gold standard. First, the use of foreign exchange as official reserves increased as the gold standard progressed. While by 1913 only Germany among the core countries held any measurable amount of foreign exchange, the percentage for the rest of the world was double that for Germany. If there were a rush to cash in foreign exchange for gold, reduction of the gold of reserve-currency countries would place the gold standard in jeopardy.
Second, Britain was in a particularly sensitive situation. In 1913, almost half of world foreign-exchange reserves was in sterling, but the Bank of England had only three per cent of gold reserves. The Bank of England's ‘reserve ratio’ (ratio of ‘official reserves’ to ‘liabilities to foreign monetary authorities held in London financial institutions’) was only 31 per cent, far lower than those of the monetary authorities of the other core countries. An official run on sterling could force Britain off the gold standard. Private foreigners also held considerable liquid assets in London, and could themselves initiate a run on sterling.
Third, the United States was a source of instability to the gold standard. Its Treasury held a high percentage of world gold reserves (in 1913, more than that of the three other core countries combined). With no central bank and a decentralized banking system, financial crises were more frequent and more severe than in the other core countries. Far from the United States assisting Britain, gold often flowed from the Bank of England to the United States, to satisfy increases in US demand for money. In many years the United States was a net importer rather than exporter of capital to the rest of the world – the opposite of the other core countries. The political power of silver interests and recurrent financial panics led to imperfect credibility in the US commitment to the gold standard. Indeed, runs on banks and on the Treasury gold reserve placed the US gold standard near collapse in the 1890s. The credibility of the Treasury's commitment to the gold standard was shaken; twice the US gold standard was saved only by cooperative action of the Treasury and a bankers’ syndicate, which stemmed gold exports.

Automatic force for stability: price specie-flow mechanism

The money supply is the product of the money multiplier and the monetary base. The monetary authority alters the monetary base by changing its gold holdings and domestic assets (loans, discounts, and securities). However, the level of its domestic assets is dependent on its gold reserves, because the authority generates demand liabilities (notes and deposits) by increasing its assets, and convertibility of

these liabilities must be supported by a gold reserve. Therefore the gold standard provides a constraint on the level (or growth) of the money supply.

Further, balance-of-payments surpluses (deficits) are settled by gold imports (exports) at the gold import (export) point. The change in the money supply is the product of the money multiplier and the gold flow, providing the monetary authority does not change its domestic assets. For a country on a gold-exchange standard, holdings of foreign exchange (a reserve currency) take the place of gold.

A country experiencing a balance-of-payments deficit loses gold and its money supply decreases *automatically*. Money income contracts and the price level falls, thereby increasing exports and decreasing imports. Similarly, a surplus country gains gold, exports decrease, and imports increase. In each case, balance-of-payments equilibrium is restored via the current account, the ‘price specie-flow mechanism’. To the extent that wages and prices are inflexible, movements of real income in the same direction as money income occur; the deficit country suffers unemployment, while the payments imbalance is corrected.

The capital account also acts to restore balance, via interest-rate increases in the deficit country inducing a net inflow of capital. The interest-rate increases also reduce real investment and thence real income and imports. The opposite occurs in the surplus country.

Rules of the game

Central banks were supposed to reinforce (rather than ‘sterilize’) the effect of gold flows on the monetary base, thereby enhancing the price specie-flow mechanism. A gold outflow decreases the international assets of the central bank and the money supply. The central-bank's ‘proper’ response is: (1) decrease lending and sell securities, thereby decreasing domestic assets and the monetary base; (2) raise its ‘discount rate’, which induces commercial banks to adopt a higher reserves–deposit ratio, thereby reducing the money multiplier. On both counts, the money supply is further decreased. Should the central bank increase its domestic assets when it loses gold, it engages in sterilization of the gold flow, violating the ‘rules of the game’. The argument also holds for gold inflow, with sterilization involving the central bank decreasing its domestic assets when it gains gold.

Monetarist theory suggests the ‘rules’ were inconsequential. Under fixed exchange rates, gold flows adjust money supply to money demand; the money supply is not determined by policy. Also, prices, interest rates, and incomes are determined worldwide. Even core countries can influence these variables domestically only to the extent that they help determine them in the global marketplace. Therefore the price-specie flow and like mechanisms cannot occur. Historical data support this conclusion: gold flows were too small to be suggestive of these processes; and, at least among the core countries, prices, incomes, and interest rates moved closely in correspondence, contradicting the specie-flow mechanism and rules of the game.

Rather than rule (1), central-bank domestic and international assets moving in the same direction, the opposite behaviour – sterilization – was dominant, both in core and non-core European countries. The Bank of England followed the rule more than any other central bank, but even so violated it more often than not!

The Bank of England did, in effect, manage its discount rate (‘Bank Rate’) in accordance with rule (2). The Bank's primary objective was to maintain convertibility of its notes into gold, and its principal tool was Bank Rate. When the Bank's ‘liquidity ratio’ (ratio of gold reserves to outstanding note liabilities)

decreased, it usually increased Bank Rate. The increase in Bank Rate carried with it market short-term interest rates, inducing a short-term capital inflow and thereby moving the exchange rate away from the gold-export point. The converse also held, with a rise in the liquidity ratio generating a Bank Rate decrease. The Bank was constantly monitoring its liquidity ratio, and in response altered Bank Rate almost 200 times over 1880–1913.

While the Reichsbank also generally moved its discount rate inversely to its liquidity ratio, other central banks often violated rule (2). Discount-rate changes were of inappropriate direction, or of insufficient magnitude or frequency. The Bank of France kept its discount rate stable, choosing to have large gold reserves, with payments imbalances accommodated by fluctuations in its gold rather than financed by short-term capital flows. The United States, lacking a central bank, had no discount rate to use as a policy instrument.

Reason for stability: credible commitment to convertibility

From the late 1870s onward, there was absolute private-sector credibility in the commitment to the fixed domestic-currency price of gold on the part of Britain, France, Germany, and other important European countries. For the United States, this absolute credibility applied from about 1900. That commitment had a contingency aspect: convertibility could be suspended in the event of dire emergency; but, after normal conditions were restored, convertibility and honouring of gold contracts would be re-established at the pre-existing mint price – even if substantial deflation was required to do so. The Bank Restriction and greenback periods were applications of the contingency. From 1879, the ‘contingency clause’ was exercised by none of these countries.

The absolute credibility in countries’ commitment to convertibility at the existing mint price implied that there was zero ‘convertibility risk’ (Treasury or central-bank notes non-redeemable in gold at the established mint price) and zero ‘exchange risk’ (alteration of mint parity, institution of exchange control, or prohibition of gold export).

Why was the commitment to credibility so credible?

1. 1. Contracts were expressed in gold; abandonment of convertibility meant violation of contracts – anathema to monetary authorities.
2. 2. Shocks to economies were infrequent and generally mild.
3. 3. The London capital market was the largest, most open, most diversified in the world, and its gold market was also dominant. A high proportion of world trade was financed in sterling, London was the most important reserve-currency centre, and payments imbalances were often settled by transferring sterling assets rather than gold. Sterling was an international currency – a boon to other countries, because sterling involved positive interest return, and its transfer costs were much less than those of gold. Advantages to Britain were the charges for services as an international banker, differential interest return on its financial intermediation, and the practice of countries on a sterling (gold-exchange) standard of financing payments surpluses with Britain by piling up short-term sterling assets rather than demanding Bank gold.
4. 4. ‘Orthodox metallism’ – authorities’ commitment to an anti-inflation, balanced-budget, stable-money policy – reigned. This ideology implied low government spending, low taxes, and limited monetization of government debt. Therefore, it was not expected that a country's price level

would get out of line with that of other countries.

5. 5. Politically, gold had won over paper and silver, and stable-money interests (bankers, manufacturers, merchants, professionals, creditors, urban groups) over inflationary interests (farmers, landowners, miners, debtors, rural groups).
6. 6. There was a competitive environment and freedom from government regulation. Prices and wages were flexible. The core countries had virtually no capital controls, Britain had adopted free trade, and the other core countries had only moderate tariffs. Balance-of-payments financing and adjustment were without serious impediments.
7. 7. With internal balance an unimportant goal of policy, preservation of convertibility of paper currency into gold was the primary policy objective. Sterilization of gold flows, though frequent, was more ‘meeting the needs of trade’ (passive monetary policy) than fighting unemployment (active monetary policy).
8. 8. The gradual establishment of mint prices over time ensured that mint parities were in line with relative price levels; so countries joined the gold standard with exchange rates in equilibrium.
9. 9. Current-account and capital-account imbalances tended to be offsetting for the core countries. A trade deficit induced a gold loss and a higher interest rate, attracting a capital inflow and reducing capital outflow. The capital-exporting core countries could stop a gold loss simply by reducing lending abroad.

Implications of credible commitment

Private parties reduced the need for balance-of-payments adjustment, via both gold-point arbitrage and stabilizing speculation. When the exchange rate was outside the spread, gold-point arbitrage quickly returned it to the spread. Within the spread, as the exchange value of a currency weakened, the exchange rate approaching the gold-export point, speculators had an ever greater incentive to purchase domestic with foreign currency (a capital inflow). They believed that the exchange rate would move in the opposite direction, enabling reversal of their transaction at a profit. Similarly, a strengthened currency involved a capital outflow. The further the exchange rate moved toward a gold point, the greater the potential profit opportunity in betting on a reversal of direction; for there was a decreased distance to that gold point and an increased distance from the other point. This ‘stabilizing speculation’ increased the exchange value of depreciating currencies, and thus gold loss could be prevented. Absence of controls meant such private capital flows were highly responsive to exchange-rate changes.

Government policies that enhanced stability

Specific government policies enhanced gold-standard stability. First, by the turn of the 20th century, South Africa – the main world gold producer – was selling all its gold output in London, either to private parties or to the Bank of England. Thus the Bank had the means to replenish its gold reserves. Second, the orthodox-metallism ideology and the leadership of the Bank of England kept countries’ monetary policies disciplined and in harmony. Third, the US Treasury and the central banks of the other core countries manipulated gold points, to stem gold outflow. The cost of exporting gold was artificially increased (for example, by increasing selling prices for bars and foreign coin) and/or the cost of importing gold artificially decreased (for example, by providing interest-free loans to gold importers).

Fourth, central-bank cooperation was forthcoming during financial crises. The precarious liquidity position of the Bank of England meant that it was more often the recipient than the provider of financial assistance. In crises, the Bank would obtain loans from other central banks, and the Bank of France would sometimes purchase sterling to support that currency. When needed, assistance went from the Bank of England to other central banks. Also, private bankers unhesitatingly made loans to central banks in difficulty.

Thus, ‘virtuous’ interactions were responsible for the stability of the gold standard. The credible commitment to convertibility of paper money at the established mint price, and therefore to fixed mint parities, were both a cause and an effect of the stable environment in which the gold standard operated, the stabilizing behaviour of arbitrageurs and speculators, and the responsible policies of the authorities – and these three elements interacted positively among themselves.

Experience of periphery

An important reason for periphery countries to join and maintain the gold standard was the fostering of access to core-countries’ capital markets. Adherence to the gold standard connoted that the peripheral country would follow responsible macroeconomic policies and repay debt. This ‘seal of approval’, by reducing the risk premium, involved a lower interest rate on the country's bonds sold abroad, and very likely a higher volume of borrowing, thereby enhancing economic development.

However, periphery countries bore the brunt of the burden of adjustment of payments imbalances with the core (and other western European) countries. First, when the gold-exchange-standard periphery countries ran a surplus (deficit), they increased (decreased) their liquid balances in the United Kingdom (or other reserve-currency country) rather than withdraw gold from (lose gold to) the reserve-currency country. The monetary base of the periphery country increased (decreased), but that of the reserve-currency country remained unchanged. Therefore, changes in domestic variables – prices, incomes, interest rates, portfolios – that occurred to correct the imbalance were primarily in the periphery. Second, when Bank Rate increased, London drew funds from France and Germany, which attracted funds from other European countries, which drew capital from the periphery. Also, it was easy for a core country to correct a deficit by reducing lending to, or bringing capital home from, the periphery. While the periphery was better off with access to capital, its welfare gain was reduced by the instability of capital import. Third, periphery-countries’ exports were largely primary products, sensitive to world market conditions. This feature made adjustment in the periphery take the form more of real than financial correction.

The experience of adherence to the gold standard differed among periphery groups. The important British dominions and colonies successfully maintained the gold standard. They paid the price of serving as an economic cushion to the Bank of England's financial situation; but, compared with the rest of the periphery, gained a stable long-term capital inflow. In southern Europe and Latin America, adherence to the gold standard was fragile. The commitment to convertibility lacked credibility, and resort to a paper standard occurred. Many of the reasons for credible commitment that applied to the core countries were absent. There were powerful inflationary interests, strong balance-of-payments shocks, and rudimentary banking sectors. The cost of adhering to the gold standard was apparent: loss of the ability to depreciate the currency to counter reductions in exports. Yet the gain, in terms of a steady capital inflow from the core countries, was not as stable or reliable as for the British dominions and colonies.

Breakdown of classical gold standard

The classical gold standard was at its height at the end of 1913, ironically just before it came to an end. The proximate cause of the breakdown of the classical gold standard was the First World War. However, it was the gold-exchange standard and the Bank of England's precarious liquidity position that were the underlying cause. With the outbreak of war, a run on sterling led Britain to impose extreme exchange control – a postponement of both domestic and international payments – making the international gold standard inoperative. Convertibility was not suspended legally; but moral suasion, legalistic action, and regulation had the same effect. The Bank of England commandeered gold imports and applied moral suasion to bankers and bullion brokers to restrict gold exports.

The other gold-standard countries undertook similar policies – the United States not until 1917, when it adopted extra-legal restrictions on convertibility and restricted gold exports. Commercial banks converted their notes and deposits only into currency. Currency convertibility made mint parities ineffective; floating exchange rates resulted.

Return to the gold standard

After the First World War, a general return to gold occurred; but the interwar gold standard differed institutionally from the classical gold standard. First, the new gold standard was led by the United States, not Britain. The US embargo on gold exports was removed in 1919, and currency convertibility at the pre-war mint price was restored in 1922. The gold value of the dollar rather than pound sterling was the typical reference point around which other currencies were aligned and stabilized. The core now had two central countries, the United Kingdom (which restored gold in 1925) and the United States.

Second, for many countries there was a time lag between stabilizing the currency in the foreign-exchange market (fixing the exchange rate or mint parity) and resuming currency convertibility. The interwar gold standard was at its height at the end of 1928, after all core countries were fully on the standard and before the Great Depression began. The only countries that never joined the interwar gold standard were the USSR, silver–standard countries (China, Hong Kong, Indochina, Persia, Eritrea), and some minor Asian and African countries.

Third, the ‘contingency clause’ of convertibility conversion, that required restoration of convertibility at the mint price that existed prior to the emergency (the First World War), was *broken* by various countries, and even core countries. While some countries (including the United States and United Kingdom) stabilized their currencies at the pre-war mint price, others (including France) established a gold content of their currency that was a fraction of the pre-war level: the currency was devalued in terms of gold, the mint price was higher than pre-war. Still others (including Germany) stabilized new currencies adopted after hyperinflation.

Fourth, the gold coin standard, dominant in the classical period, was far less prevalent in the interwar period. All four core countries had been on coin in the classical gold standard; but only the United States was on coin interwar. The gold-bullion standard, non-existent pre-war, was adopted by the United Kingdom and France. Germany and most non-core countries were on a gold-exchange standard.

Instability of interwar gold standard

The interwar gold standard was replete with forces making for *instability*.

1. 1. The process of establishing fixed exchange rates was piecemeal and haphazard, resulting in disequilibrium exchange rates. Among core countries, the United Kingdom restored convertibility at the pre-war mint price without sufficient deflation, and had an overvalued currency of about ten per cent. France and Germany had undervalued currencies.
2. 2. Wages and prices were less flexible than in the pre-war period.
3. 3. Higher trade barriers than pre-war also restrained adjustment.
4. 4. The gold-exchange standard economized on total world gold via the gold of the United Kingdom and United States in their reserves role for countries on the gold-exchange standard and also for countries on a coin or bullion standard that elected to hold part of their reserves in London or New York. However, the gold-exchange standard was unstable, with a conflict between (*a*) the expansion of sterling and dollar liabilities to foreign central banks, to expand world liquidity, and (*b*) the resulting deterioration in the reserve ratio of US and UK authorities. This instability was particularly severe, for several reasons. First, France was now a large official holder of sterling, and France was resentful of the United Kingdom. Second, many more countries were on the gold-exchange standard than pre-war. Third, the gold-exchange standard, associated with colonies in the classical period, was considered a system inferior to a coin standard.
5. 5. In the classical period, London was the one dominant financial centre; in the interwar period it was joined by New York and, in the late 1920s, Paris. Private and official holdings of foreign currency could shift among the two or three centres, as interest-rate differentials and confidence levels changed.
6. 6. There was maldistribution of gold. In 1928, official reserve-currency liabilities were much more concentrated than in 1913, British pounds accounting for 77 per cent of world foreign-exchange reserves and French francs less than two per cent (versus 47 and 30 per cent in 1913). Yet the United Kingdom held only seven per cent of world official gold and France 13 per cent. France also possessed 39 per cent of world official foreign exchange. The United States held 37 per cent of world official gold.
7. 7. Britain's financial position was even more precarious than in the classical period. In 1928, the gold and dollar reserves of the Bank of England covered only one-third of London's liquid liabilities to official foreigners, a ratio hardly greater than in 1913. UK liquid liabilities were concentrated on stronger countries (France, United States), whereas UK liquid assets were predominantly in weaker countries (Germany). There was ongoing tension with France, which resented the sterling-dominated gold-exchange standard and desired to cash in its sterling holding for gold, to aid its objective of achieving first-class financial status for Paris.
8. 8. Internal balance was an important goal of policy, which hindered balance-of-payments adjustment, and monetary policy was influenced by domestic politics rather than geared to preservation of currency convertibility.
9. 9. Credibility in authorities’ commitment to the gold standard was not absolute. Convertibility risk and exchange risk could be high, and currency speculation could be destabilizing rather than stabilizing. When a country's currency approached or reached its gold-export point, speculators

might anticipate that currency convertibility would not be maintained and that the currency would be devalued.

10. 10. The ‘rules of the game’ were violated even more often than in the classical gold standard. Sterilization of gold inflows by the Bank of England can be viewed as an attempt to correct the overvalued pound by means of deflation. However, the US and French sterilization of their persistent gold inflows reflected exclusive concern for the domestic economy and placed the burden of adjustment (deflation) on other countries.
11. 11. The Bank of England did not provide a leadership role in any important way, and central-bank cooperation was insufficient to establish credibility in the commitment to currency convertibility. The Federal Reserve had three targets for its discount-rate policy: strengthen the pound, combat speculation in the New York stock market, and achieve internal balance – and the first target was of lowest priority. Although, for the sake of external balance, the Bank of England kept Bank Rate higher than internal considerations would dictate, it was understandably reluctant to abdicate Bank Rate policy entirely to the balance of payments, with little help from the Federal Reserve. To keep the pound strong, substantial international cooperation was required, but was not forthcoming.

Breakdown of interwar gold standard

The Great Depression triggered the unravelling of the gold standard. The depression began in the periphery. Low export prices and debt-service requirements created insurmountable balance-of-payments difficulties for gold-standard commodity producers. However, US monetary policy was an important catalyst. In 1927 the Federal Reserve favoured easy money, which supported foreign currencies but also fed the New York stock-market boom. Reversing policy to tame the boom, higher interest rates attracted monies to New York, weakening sterling in particular. The crash of October 1929, while helping sterling, was followed by the US depression. This spread worldwide, with declines in US trade and lending. In 1929 and 1930 a number of periphery countries –both dominions and Latin American countries – either formally suspended currency convertibility or restricted it so that currencies violated the gold-export point.

It was destabilizing speculation, emanating from lack of confidence in authorities’ commitment to currency convertibility, which ended the interwar gold standard. In May 1931 there was a run on Austria's largest commercial bank, and the bank failed. The run spread to other eastern European countries and to Germany, where an important bank also collapsed. The countries’ central banks lost substantial reserves; international financial assistance was too late; and in July 1931 Germany adopted exchange control, followed by Austria in October. These countries were definitively off the gold standard.

The Austrian and German experiences, as well as British budgetary and political difficulties, were among the factors that destroyed confidence in sterling, which occurred in mid-July 1931. Runs on sterling ensued, and the Bank of England lost much of its reserves. Loans from abroad were insufficient, and in any event taken as a sign of weakness. The gold standard was abandoned in September, and the pound quickly and sharply depreciated on the foreign-exchange market, as overvaluation of the pound would imply.

Following the UK abandonment of the gold standard, many countries followed, some to maintain their

competitiveness via currency devaluation, others in response to destabilizing capital flows. The United States held on until 1933, when both domestic and foreign demands for gold, manifested in runs on US commercial banks, became intolerable. ‘Gold bloc’ countries (France, Belgium, Netherlands, Switzerland, Italy, Poland), with their currencies now overvalued and susceptible to destabilizing speculation, succumbed to the inevitable by the end of 1936.

The Great Depression was worsened by the gold standard: gold-standard countries hesitated to inflate their economies, for fear of suffering loss of gold and foreign-exchange reserves, and being forced to abandon convertibility or the gold parity. The gold standard involved ‘golden fetters’, which inhibited monetary and fiscal policy to fight the Depression. As countries left the gold standard, removal of monetary and fiscal policy from their ‘gold fetters’ enabled their use in expanding real output, providing the political will existed.

In contrast to the interwar gold standard, the classical gold standard functioned well because of a confluence of ‘virtuous’ interactions, involving government policies, credible commitment to the standard, private arbitrage and speculation, and fostering economic and political environment. We will not see its like again.

See Also

- banking crises
- Bank of England
- bimetallism
- Bretton Woods system
- commodity money
- silver standard
- specie-flow mechanism

Bibliography

Bayoumi, T., Eichengreen, B. and Taylor, M.P., eds. 1996. *Modern Perspectives on the Gold Standard*. Cambridge: Cambridge University Press.

Bordo, M.D. and Kydland, F.E. 1995. The gold standard as a rule: an essay in exploration. *Explorations in Economic History* 32, 423–64.

Bordo, M.D. and Rockoff, H. 1996. The gold standard as a ‘Good Housekeeping Seal of Approval’. *Journal of Economic History* 56, 389–428.

Bordo, M.D. and Schwartz, A.J., eds. 1984. *A Retrospective on the Classical Gold Standard, 1821–1931*. Chicago: University of Chicago Press.

De Macedo, J.B., Eichengreen, B. and Reis, J., eds. 1996. *Currency Convertibility: The Gold Standard*

and Beyond. London: Routledge.

Eichengreen, B. 1992. *Golden Fetters: The Gold Standard and the Great Depression, 1919–1939*. New York: Oxford University Press.

Eichengreen, B. and Flandreau, M. 1997. *The Gold Standard in Theory and History*, 2nd edn. London: Routledge.

Gallarotti, G.M. 1995. *The Anatomy of an International Monetary Regime: The Classical Gold Standard, 1880–1914*. New York: Oxford University Press.

Officer, L.H. 1996. *Between the Dollar–Sterling Gold Points*. Cambridge: Cambridge University Press.

Officer, L.H. 2001. Gold standard. *EH. Net Encyclopedia*, ed. R. Whaples. Online. Available at http://eh.net/encyclopedia/article/officer.gold.standard, accessed 20 October 2006.

How to cite this article

Officer, Lawrence H. "gold standard." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000204> doi:10.1057/9780230226203.0653

The New Palgrave Dictionary of Economics Online

golden rule

Edmund S. Phelps

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The so-called golden rule (of capital accumulation) is a proposition about the consequences for national welfare possibilities of alternative paths of national wealth, and hence of national saving, in a closed economy. It states that the steady-growth state that gives the maximum path of consumption is the one along which national consumption equals the national wage bill and thus national saving equals ‘profits’. The basic significance of the golden rule is as a warning against national policies of over-saving or counterproductive austerity.

Keywords

capital accumulation; capital–output ratio; golden rule; maximin; national consumption function; natural rate of growth; Phelps, E. S.; reciprocity; Robinson, J. V.; saving–output ratio; Schumpeter, J. A.; social optimum; social rate of return; Solow, R. M.; steady-growth state; Swan, T. W.; technological progress; utilitarianism

Article

The so-called golden rule, or golden rule of capital accumulation, is a proposition about the consequences for national consumption – more broadly, for national welfare possibilities – of alternative paths of national wealth, and hence of national saving, in a closed economy. It developed out of the dynamic models of capital accumulation and output growth, generally in a setting of steady technical progress and demographic increase, begun in 1956 by R.M. Solow and T.W. Swan after some early explorations by Harrod, Domar and Robinson. Solow and Swan had shown that, provided diminishing returns to capital set in strongly enough (whether or not smoothly), there exists a state of steady growth corresponding to each possible value of the saving–output ratio, and, interestingly enough, this steady-state growth rate is independent of the value of the saving ratio – and often called the natural rate of growth. Hence, an upward shift of saving at each level of output, through increased private thrift or else

higher taxes or lower spending by the government, cannot have a permanent, or non-vanishing, effect on the growth rate of output, only a transient effect.

If this theorem was the first law of the new ‘growth economics’, the golden rule of accumulation was the second law of growth economics. It states that the steady-growth state that gives the maximum path of consumption – the path layered on top of all the other steady-growth consumption tracks – is the one along which national consumption equals the national wage bill and thus national saving equals ‘profits’ (gross of interest in the present use of the term). Equivalently, the consumption-maximizing steady-growth path is the steady state along which the competitive rate of interest, which is the social rate of return to investment and to saving, is equal to the natural rate of growth. (To see the equivalence, divide profits and saving by capital.) Hence, a country (with any given history) that now plans for ever to equate saving to profits could not hope to achieve a sustainable increase in the consumption path by some date in the future through a shift of policy towards increased saving; very possibly, even a temporary increase of consumption would not result. The reason is that despite a boost to future output brought by greater accumulation, the increase in future investment would eat up the increase in future output – and then some.

The arrival *circa* 1960 of the golden rule result was a classic case of multiple discoverers. And discovery seems the apt word, since the golden rule theorem was just a simple insight about a set or sets of equations in existence for several years that was waiting to be noticed, not a creative vision of the world springing from an independent empirical sense; accordingly, many or most of the discoverers were fledgling, pre-flight theorists still working on the ground of existing models. The earliest publishers of the result were Phelps (1961), Robinson (1962), and Swan (1963). However, it quickly became apparent that there were also discoveries on the Continent by von Weizsäcker and Allais, and even within the tiny space of the Cowles Foundation at Yale there were additional independent discoveries by Beckmann and Srinivasan. Robinson coined the proposition ‘the neoclassical theorem’, but eventually Phelps's coinage ‘the golden rule’ became the standard. This was not a case of bad money driving out good, as will be explained.

The term ‘golden rule’ was something of a play on words. Mrs Robinson had dubbed states of steady growth as ‘golden ages’, so a proposition (if not exactly a maxim) about choosing among golden ages was natural to call a golden rule. In addition, there was also an allusion in the term to the biblical golden rule, do unto others as you would have them do unto you. The sense of that maxim, presumably, is that if one asserts a right to a certain policy, or treatment, from others, then in one's own treatment of others one must accord them the right to the same policy; so the choice of the rights to assert is subject to a reciprocity or cost constraint, which is a useful thing, for otherwise one would demand the most extreme sacrifices of others. Of course, this precept – the national saving policy, or national consumption function, that a future generation would have preceding ones follow, in view of its self-interests, it must likewise adopt on behalf of succeeding generations – does not by itself determine the just policy of national saving. Yet the golden rule perspective serves to alert us that there will be a limit to the austerity that future generations would ask of the present generation if they are obliged to practice the same austerity that they choose to preach. To make this effective, it should be noted, the saving policies from which society is to choose must be linear-homogeneous, and thus expressed in terms of saving as a ratio to output or profits or some related variable. (Otherwise, a future generation could piously call for lower consumption only at the present, comparatively low level of national income – and thus travel as a free rider.)

The meaning and indeed the significance of the golden rule becomes quite transparent in the special case of an economy in which the technology and the (working-age) population are constant, so that the natural rate of growth of the economy is zero. In this case the golden rule state is the Schumpeterian zero-interest stationary state; and since the net rate of return to investment is zero, gross profits are simply depreciation allowances and equal to gross investment, which is entirely replacement investment in a stationary state. Here it is abundantly clear why an alternative stationary state with a constant negative rate of interest would actually yield a lower path of consumption: the extra replacement investment would more than eat up the extra gross output, leaving an actual diminution of the net national product and consumption (to which NNP is equal). It is also perfectly clear that a society is not required as a matter of efficiency to aim for the Schumpeterian state; if the initial rate of return to investment is positive, it would take *lower* consumption in the present than would otherwise be possible on a sustainable basis (simply by consuming income) in order to move to the Schumpeterian state so that in the future a higher consumption level could be sustained than would otherwise be possible. Neither is such a move required as a matter of justice. From the utilitarian side there are economists who cheerfully discount the utilities of future people, and from the ‘maximin’ perspective it is obvious that present people would not optimally sacrifice to make better-off those who were not worse-off than they to begin with. The basic significance of the golden rule, then, is as a warning against national policies of over-saving, or counterproductive austerity. The golden rule theorem is simply a generalization to a growing economy of these observations.

Further results on the inefficiency entailed by exceeding, so to speak, the golden rule in certain respects were later obtained. It was shown with the help of T.C. Koopmans that keeping the capital–output ratio indefinitely in excess (by a non-vanishing amount) of the golden rule level would be dominated in terms of consumption, and thus utility, by another path, feasible from the same initial conditions, along which the capital–output ratio is always ‘epsilon’ smaller (Phelps, 1965; 1966). A much more general analysis came later from D. Cass in 1972 in which the borderline between efficient and dynamically inefficient paths is systematically examined.

‘But the golden rule path could be the social optimum, couldn't it? Certainly it is very beautiful, and not obviously unjust!’ There was a tendency among some to regard it as the optimum at least provisionally, for working purposes. However, any budding claims that may have existed for the ‘optimality’ of the golden rule path met with an objection by I.F. Pearce (1962). Start there at T_1, Pearce said, and end there at T_2. Then, if there is steady population growth, so the golden rule interest rate is positive (being equal to the population growth rate), it will increase the integral of total utility to save more now and less later, causing the capital stock to arch over its golden-rule track, since more saving will increase output. There could be no denying this, although some utilitarians prefer to sum the per capita utility of people over time (or the utility of per capita consumption), which suggests there is a maximin impulse in their otherwise utilitarian hearts; from this angle it would not be preferable to deviate along Pearce's arching detour. Then, using the per capita utility version of utilitarianism, P.A. Samuelson (1967) took up the cudgels with a revision of the Pearce argument: if there is steady *technical progress*, so the golden rule interest rate exceeds the population growth rate, it will increase the integral of per capita utility to cause the capital stock to arch above its golden-rule track since more saving will increase per capita output – as long as the interest rate remains above Samuelson's ‘biological’ level, which is the population growth rate. Again, it is *nolo contendere* from the golden rule side. Yet maximin advocates might object that if

the per capita utility received by succeeding generations is rising, and unavoidably so, so that the oldest generation extant is always the worst-off, the detour from the golden rule path proposed by Samuelson would presumably entail some belt-tightening by the oldest generation along with the others in order to produce the consumption splurge for the benefit of some younger or future generations – hence a reduction in *minimum* per capita utility across generations. That cannot be a maximin improvement and is indeed a maximin worsening. Thus goes the maximin rejoinder to the turnpike ‘refutation’ of the golden rule.

See Also

- neoclassical growth theory
- Ramsey model
- Robinson, Joan Violet

Bibliography

Cass, D. 1972. On capital overaccumulation in the aggregative neoclassical model of economic growth: a complete characterization. *Journal of Economic Theory* 4, 200–23.

Pearce, I.F. 1962. The end of the golden age in Solovia. *American Economic Review* 52, 1088–97.

Phelps, E.S. 1961. The golden rule of accumulation: a fable for growthmen. *American Economic Review* 51, 638–43.

Phelps, E.S. 1965. Second essay on the golden rule of accumulation. *American Economic Review* 55, 793–814.

Phelps, E.S. 1966. *Golden Rules of Economic Growth*. New York: Norton; Amsterdam: North-Holland.

Robinson, J. 1962. A neo-classical theorem. *Review of Economic Studies* 29, 219–26.

Samuelson, P.A. 1967. A turnpike refutation of the golden rule in a welfare-maximizing many-year plan. In *Essays on the Theory of Optimal Economic Growth*, ed. K. Shell. Cambridge, MA: MIT Press.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Swan, T.W. 1963. Of golden ages and production functions. In *Economic Development with Special Reference to East Asia*, ed. K.E. Berrill. London: Macmillan.

How to cite this article

Phelps, Edmund S. "golden rule." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000074> doi:10.1057/9780230226203.0654

The New Palgrave Dictionary of Economics Online

Goldfeld, Stephen (1940 – 1995)

Richard E. Quandt
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

computational economics; disequilibrium models; econometrics; Goldfeld, S.; GRADX algorithm; homoscedasticity; money demand function; nonlinear models in econometrics; Sargan distribution; Suits agricultural model; voice

Article

Goldfeld was born in Bronx, New York, and received his undergraduate training in mathematics at Harvard University. He obtained a Ph.D. in economics from MIT in 1963. Except for visiting appointments at Center for Operations Research and Econometrics (CORE) at the Université Catholique de Louvain, the University of California, Berkeley, and the Technion in Haifa, and a year as a member of the US Council of Economic Advisors, he spent his professional life at Princeton University, where he served as Provost from 1993 until his death.

Goldfeld largely divided his professional interests between monetary economics and econometric methodology. In the former category are his early book (1966a) and two definitive papers (1973a) and (1976) that are models of careful empirical econometrics. They address the question of whether the money demand function is basically stable in the short run. In estimating the money demand function for the post-war period, Goldfeld exhaustively analyses specifications differing from one another in terms of the extent of aggregation, the type of lag structure and the types of variables included in the demand function. In (1973a), he finds no substantial short-run instability in the demand function. He also concludes that disaggregation by sector probably helps explain money demand. But the (1976) paper attacks the poor predictive performance of the money demand equation in the period following 1973: the equation consistently over-predicts money demand. An even more extensive and thorough analysis aimed at fixing this problem makes him suggest that a more fundamental rethinking of money demand is indicated. He also studied savings and loan associations and their behaviour with respect to rate setting and allocational efficiency (1970), forecasting and policy evaluation in large-scale econometric models (1971) and numerous related topics.

His methodological interests began with the construction of a new test for homoscedasticity (1965),

subsequently called the Goldfeld–Quandt test, followed in short order by an abiding interest in nonlinear estimation (1968) and computational econometrics (1966b). This led to the invention of the GRADX algorithm, which solves the maximization problem for non-concave functions by embedding them in a concave analogue. He then turned to the case of switching regressions: problems in which the parameters of a regression equation switch from one set of values to another, but at an unknown point in time (1973b) (hence making the Chow test inappropriate). The superficial similarity of these models to disequilibrium models, in which the observed quantity represents the short side of the market, led him to an abiding interest in models of the latter type, sometimes within the framework of financial institutions (1975; 1980). A reconsideration of the Suits agricultural model may contain the first proof of the unboundednes of a likelihood function for a disequilibrium model when sample separation is not known a priori (1975). Because of the relatively intractable nature of the econometrics when error terms are normally distributed in these models, he explored the properties of a new distribution called the Sargan distribution (1981).

In the last few years of his life, Goldfeld was much interested in the microtheoretic problems of socialist enterprises operating in the presence of a soft budget constraint, a concept derived from János Kornai (1979; 1980; 1982). In the models, firms can undertake ‘whining’ when they face losses and bail-outs permit them to continue operating. These models confirm Kornai's conjecture that the expectation of bail-outs will lead to larger input demand than in the standard competitive case (1988; 1990). Embedding the possibility of ‘whining’ in a multi-firm case in which inputs are rationed and whining is intended to induce a more generous input allotment leads to the determination of a Nash equilibrium in whining (1990).

Goldfeld had enormous range and creativity and was a precise, careful and methodical worker as well as a much-loved teacher of both undergraduates and graduate students. His premature death was a great loss to the profession.

See Also

- command economy
- money
- numerical optimization methods in economics

Selected works

1965. (With R. Quandt.) Some tests for homoscedasticity. *Journal of the American Statistical Association* 60, 539–47.

1966a. *Commercial Bank Behavior and Economic Activity: A Structural Study of Monetary Policy in the Postwar United States*. Contributions to Economic Analysis Series No. 43. Amsterdam: North-Holland.

1966b. (With R. Quandt and H. Trotter.) Maximization by quadratic hill-climbing. *Econometrica* 34, 541–51.

1968. (With R. Quandt.) Nonlinear simultaneous equations: estimation and prediction. *International Economic Review* 91, 113–36.

1970. (With D Jaffee.) The determinants of deposit-rate setting by savings and loan associations. *Journal of Finance* 25, 615–32.

1971. Forecasting and policy evaluation using large scale econometric models: comment. In *Frontiers of Quantitative Economics*, ed. M. Intriligator. Amsterdam: North-Holland.

1972. (With R. Quandt.) *Nonlinear Methods in Econometrics*. Contributions to Economic Analysis Series No. 77. Amsterdam: North-Holland.

1973a. The demand for money revisited. *Brookings Papers on Economic Activity* 1973: 3, 577–638.

1973b. (With R. Quandt.) The estimation of structural shifts by switching regressions. *Annals of Economic and Social Measurement* 2, 475–85.

1974a. (With D. Aigner.) Estimation and prediction from aggregate data when aggregates are measured more accurately than their components. *Econometrica* 42, 113–34.

1975. (With R. Quandt.) Estimation in a disequilibrium model and the value of information. *Journal of Econometrics* 3, 325–48.

1976. The case of the missing money. *Brookings Papers on Economic Activity* 1976: 3, 683–730.

1980. (With D. Jaffe and R. Quandt.) A Model of FHLBB advances: rationing or market clearing. *Review of Economics and Statistics* 62, 339–47.

1981. (With R. Quandt.) Econometrics modelling with nonnormal disturbances. *Journal of Econometrics* 17, 141–55.

1988. (With R. Quandt.) Budget constraints, bailouts and the firm under central planning. *Journal of Comparative Economics* 12, 502–20.

1990. (With R. Quandt.) Output targets, the soft budget constraint and the firm under central planning. *Journal of Economic Behavior and Optimization* 14, 205–22.

1994. (With R. Quandt.) The competition for rationed resources. *Journal of Economic Behavior and Organization* 25, 53–71.

Bibliography

Kornai, J. 1979. Resource-constrained versus demand-constrained systems. *Econometrica* 47, 801–20.

Kornai, J. 1980. *Economics of Shortage*. Amsterdam: North-Holland.

Kornai, J. 1982. *Growth, Shortage and Efficiency*. Amsterdam: North-Holland.

How to cite this article

Quandt, Richard E. "Goldfeld, Stephen (1940–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000201> doi:10.1057/9780230226203.0655

The New Palgrave Dictionary of Economics Online

Goldsmith, Raymond William (1904 – 1988)

E. Denison

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

financial intermediation ratios; financial interrelations ratios; Goldsmith, R. W.; national accounting; national income; perpetual inventory method; reproducibility

Article

Born in Brussels, Goldsmith studied in Berlin (Ph.D., 1927). He emigrated to the United States and served on the US Securities and Exchange Commission (1934–41), the War Production Board (1942–7) and as a member of the Senior Staff, National Bureau of Economic Research (1953–78). He was a Professor at New York University (1956–61) and at Yale University (1962–73; Emeritus, 1973 onwards). He pioneered the measurement of national and sector saving, investment, wealth, and balance sheets.

In the 1930s Goldsmith initiated Securities and Exchange Commission estimates of saving, obtained by a new method: subtracting additions to liabilities from additions to assets. During 1947–68 the Commerce Department national income accounts reported personal saving derived from the Securities and Exchange Commission series as an alternative to estimates obtained as income less consumption and as investment less corporate and government saving. The Federal Reserve Board then absorbed the Securities and Exchange Commission work into its flow-of-funds accounts. These had been started by Copeland in research partly paralleling Goldsmith's (Copeland, 1952, p. xvi).

A still more important breakthrough, in 1950, was Goldsmith's perpetual inventory method of estimating the stock and consumption of durable capital. The United States and other countries use it to obtain official national income and capital stock series. Goldsmith's monumental *Study of Saving* followed, providing balance sheets, saving, and wealth, for the nation and sectors. It featured broad scope, great detail, and series running back to the 1890s (to 1805 for reproducible wealth; in Goldsmith, 1951b). Goldsmith subsequently updated these series and introduced similar ones for other countries. Lifelong adherence to the principle of reproducibility, which calls for sufficient description of all estimates to enable the reader to reproduce them, increases the usefulness of Goldsmith's data.

Goldsmith wrote extensively about financial institutions and capital markets. He introduced the financial

interrelations and financial intermediation ratios (Goldsmith, 1966; 1969). In his later years Goldsmith studied pre-industrial economies (Goldsmith 1987).

Selected works

1939. (Assisted by W. Salant.) The volume and components of saving in the United States 1933–1937. *Studies in Income and Wealth*, vol. 3. New York: NBER.

1951a. A perpetual inventory of national wealth. *Studies in Income and Wealth*, vol. 14. New York: NBER.

1951b. The growth of reproducible wealth of the United States of America from 1805 to 1950. In *Income and Wealth of the United States: Trends and Structure*, ed. S. Kuznets. Cambridge: Bowes and Bowes.

1955–6. *A Study of Saving in the United States*. 3 vols. Princeton: Princeton University Press.

1958. *Financial Intermediaries in the American Economy Since 1900*. Princeton: Princeton University Press.

1962. *The National Wealth of the United States in the Postwar Period*. Princeton: Princeton University Press.

1963. (With R. Lipsey and M. Mendelson.) *Studies in the National Balance Sheet of the United States*. 2 vols. Princeton: Princeton University Press.

1966. *The Determinants of Financial Structure*. Paris: OECD.

1969. *Financial Structure and Development*. New Haven: Yale University Press.

1982. *The National Balance Sheet of the United States, 1953–1980*. Chicago: University of Chicago Press.

1984. An estimate of the size and structure of the national product of the Early Roman Empire. *Review of Income and Wealth* 30, 263–88.

1985. *Comparative National Balance Sheets: A Study of Twenty Countries, 1688–1978*. Chicago: University of Chicago Press.

1987. *Premodern Financial Systems: A Historical Comparative Study*. Cambridge: Cambridge University Press.

Bibliography

Copeland, M. 1952. *A Study of Money-flows in the United States*. New York: NBER.

How to cite this article

Denison, E. "Goldsmith, Raymond William (1904–1988)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000075> doi:10.1057/9780230226203.0656

The New Palgrave Dictionary of Economics Online

goods and commodities

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

characteristics; commodities; demand theory; exchange value; goods and commodities; Jevons, W. S.; Johnson, H. G.; Lancaster, K.; marginal revolution; Marshall, A.; Marx, K. H.; Menger, C.; Mill, J. S.; produced and scarce commodities; Ricardo, D; Robbins, L. C.; scarcity; separable utility functions; Smith, A.; subjective theory of value; water–diamonds paradox

Article

Towards the end of the 1950s, Harry Johnson produced the following theorem on value theory:

> Define a good as an object or service of which the consumer would choose to have more. Then the collection of goods he chooses when he has more money to spend (prices being constant) must represent more goods than that he chooses when he has less money to spend (since he could have had more of each separate good).
>
> 1. i. If his income rises, he buys more goods; this implies a presumption that normally the income effect is positive.
> 2. ii. If he chooses collection B when he could have had collection A for the same money (i.e. $\Sigma p_b q_b = \Sigma p_b q_a$), he does not choose A if he could have had B for less money, because that would mean collection B represented less goods than collection A, and conflict with the definition of goods. Hence, when A is chosen B must be at least as expensive (i.e. $\Sigma p_a q_b \geq \Sigma p_a q_a$). This establishes that the substitution effect is non-negative (by subtraction, $\Sigma (p_b - p_a)(q_b - q_a) \leq 0$).
>
> Hence we derive both parts of the law of demand from the definition of goods. The hypothesis from which we have deduced it is that goods are goods. (1958, p. 149)

The idea that definition of goods carried with it the whole of the theory of demand, that the explanation of the determination of the exchangeable value of ‘things’ was intimately bound up with the definition of the ‘things’ themselves, probably struck many of Harry Johnson's readers as amusing. Indeed, Johnson may even have had this end in mind – no doubt many in the profession were beginning to wonder whether the frequency with which reconsiderations of something apparently so obvious as the theory of demand were being undertaken was entirely necessary. However, it did not strike at least one of his readers as being just another amusing aphorism: Kelvin Lancaster, upon whose review of Hicks's *Revision of Demand Theory* Johnson was commenting when he produced his theorem, took it seriously. Eight years later in the *American Economic Review* Lancaster advanced the so-called characteristics theory of demand. The argument was a simple corollary of the Johnson theorem: if it is the aim of the theory of demand to determine the prices of goods, then one ought to specify as clearly as possible the goods which are being demanded. After all, on this line of reasoning one demands not just physical objects, but the qualities with which they are endowed; it is to their *characteristics* that the potential purchaser first turns his attention.

The interesting features of this little episode, however, are not exhausted in a consideration of the ideas to which it gave rise. Quite as important are the implications which follow upon the recognition of the fact that the kinds of questions which lie behind Johnson's theorem had been debated before in contexts where certain useful results were generated. At least since the time of Adam Smith, economists have struggled to be clear about what it is in the nature of the things which are daily exchanged on markets that gives rise to exchangeable value. When Smith discussed the famous water–diamonds paradox, and drew from it (however perilously) the conclusion that the theory of exchangeable value should focus upon what may be called the objective conditions of production of things, rather than upon the subjective conditions of their consumption, he was engaged in just such an endeavour.

Smith was followed in this project by Ricardo. In the opening passages of the *Principles*, by establishing a clear line of demarcation between scarce and reproducible commodities, Ricardo reached Smith's conclusion by a different route. Marx praised this passage from Ricardo and focused his attention exclusively upon what he termed the commodity form. Moreover, this was not exclusively a classical preoccupation. Later writers, to whom modern economics seems to owe much more, also took the question very seriously indeed. Having returned to Smith's original paradox, they applied the distinction between total and marginal utility and to their satisfaction resolved it. This deprived Smith's original conclusion of its validity and allowed neoclassical writers to rebuild the theory of exchangeable value upon the basis of the subjective conditions of consumption of goods. Marshall was very clear about this at the beginning of the second chapter of Book II of his *Principles*.

The questions that Johnson's theorem prompts, therefore, include also those which were raised in these widely publicized and not insignificant debates of the 19th century over the distinction in the theory of value between those physical objects whose main characteristic is that they can be said to be in short supply, and those whose quantity may be increased by reproduction on an extended scale. To what extent, if at all, the choice of terminology by earlier writers reflects these differences is the subject matter of an investigation into goods and commodities.

Etymological preliminaries

In English the word *good* derives from the Old English word *g ō d*. It is related also to the Old Frisian

g ō *d*, the Old High German *guot*, the Old Saxon *g* ō *d*, and the Old Norse *gòdr*. The word is defined in the *Oxford Dictionary of English Etymology* as ‘the most general adjective of commendation’. The substantive plural form, *goods*, while sharing the origins, seems not to have appeared in English until the 13th century with a meaning much as it has today: objects or things which confer some advantage or produce some desirable effect upon their owner. Two further points may be noted. The first is that although there exists a genitive singular in Old Norse, no Teutonic language seems to have possessed a substantive plural form. Its usage in this manner probably derives from the Latin *bona*. The second is that despite the standard *O.E.D*. classification of *goods* as indeclinable, a substantive singular form has become common among economists.

In both modern French and German, the adjectives *bien* and *gut* share the meaning and sense as *good* in English (the German sharing the same Old Teutonic origins). The substantive plurals *biens* and *Güter* likewise share meaning and sense, together with the partial Latin origin of the English.

James Bonar's definition of the term *goods* in the original edition of this *Dictionary* – that ‘by the plural (Goods) is denoted concrete embodiments of usefulness’ – suggests that nothing of substance had been altered in the definition of the word even after it had been co-opted into the formal terminology of economic theory. Furthermore, Bonar's statement that *goods* are the physical embodiment of the metaphysical quality of *good*, seems to apply across all three languages. Of course, given that a substantive singular form is now in common usage, that part of Bonar's definition which went on to argue that the substantive singular *commodity* ‘is employed by economists to represent the missing singular of goods’, must be abandoned.

The word *commodity* is of entirely different origin and meaning. Its roots are in Latin and it is defined in the *Oxford English Dictionary* as ‘a thing produced for use or sale, an article of commerce, an object of trade’.

Commodities

Questions as to the essential properties of the things exchanged in a market economy, though they had arisen in the work of earlier economists, took on an entirely new dimension with the commencement of the systematic study of exchangeable value in the last half of the 18th century. Following immediately upon the definition of wealth as the ‘annual production’ of the system, and the analysis of the effects of progress in the division of labour on the ‘proportion between annual production and consumption’, Adam Smith had confronted the issue of the valuation of this ‘quantity of commodities annually circulated’. The problem was to establish, in the first instance, the sphere within which exchangeable value was to be examined. His answer, though failing to take into account conditions of relative scarcity, illustrates just as clearly as Johnson's theorem how an apparently neutral choice of language may be the bearer of certain theoretical precepts upon which an entire argument rests.

Perhaps even more importantly, Smith seems to have established not only the formal framework for the theory of value, but also the very language in which it was transmitted in orthodox circles right down to the time of Ricardo. That argument is sufficiently familiar not to have to be rehearsed here – the essential ingredient that is relevant to us is its rejection of the notion that the ‘utility of some particular object’ has anything to do with determining exchangeable value and that, instead, the exchangeable value of an object is to be explained in terms of what Smith variously called the ‘toil and trouble of obtaining it’ or its ‘difficulty and facility of production’.

These objects are quite consistently called by Smith *commodities* and not *goods*. The terminology and the theoretical construct seem to match quite well. If one is to follow Smith into an investigation of the relationship between conditions of production and relative prices, the usage of the term *goods* would be less than apposite. This, of course, is not to say that the familiar word *goods* does not crop up from time to time in the *Wealth of Nations* (opening the book at random would quickly disprove such a strong assertion). Nor is it to claim that Smith even bothered to take the time to explain his pattern of usage. But a determinate pattern there surely is.

Consider, for example, the discussion of the water–diamonds paradox, a passage where *goods* appears twice. This very short passage is followed by a carefully constructed paragraph setting out in a quite formal and purposeful way the project for the remainder of Book One. In that particular place, the term *commodities* is used exclusively. Indeed, the following three chapters, on real and nominal price, the component parts of price, and natural and market price, adhere fairly rigidly to this pattern of formal usage. The index, which was added to the original in its third edition of 1784, contains a lengthy entry under *commodities* but not one for *goods*. What is also interesting is that as between the two words, *goods* usually appears in those more discursive passages of the *Wealth of Nations*, whereas *commodities* is reserved for passages of a more formal, theoretical kind.

A remarkable parallel is to be found in the third edition of Ricardo's *Principles*. There, in the first paragraphs of the chapter on value, Ricardo makes a significant attempt to define just what it is that is important in the nature of those objects whose prices are determined on markets. At the same time, it should be noted, he replaces Smith's argument as to why exchangeable value is to be investigated in the sphere of production. The argument is pure Ricardo. Utility is ‘essential to exchangeable value’, objects which contribute in no way towards ‘gratification’ would be ‘destitute of exchangeable value’, but it does not determine it. Two conditions then remain to determine exchangeable value – Smith's difficulty and facility of production and, what Smith had passed over, scarcity. There follows Ricardo's famous twofold classification of commodities: those which are currently reproduced (produced commodities) and those which are fixed in quantity (scarce commodities). The exchangeable value of the former, when competition operates without restraint, is to be investigated in terms of the available methods of production. Relative prices of the latter depend upon the ‘wealth and inclinations of those who are desirous of possessing them’. Ricardo restricts the investigation to produced commodities and his use of terms resembles the pattern one discerns in the *Wealth of Nations*.

This particular argument was taken up subsequently by two writers who stand in contrasting positions with respect to this classical conception of the framework for the analysis of exchangeable value – John Stuart Mill and Karl Marx. Both built quite self-consciously on the work of Smith and Ricardo. But as it happens, while Mill was effectively to put in place ideas (which admittedly had been in the air for some time) that were quite dramatically to modify the classical position, Marx was to revivify it. How closely terminological conventions reflect these factors is a question of some importance in the present context. To begin with, let us turn to the German language, and to Marx, who claimed that Ricardo's argument for restricting the domain of the theory of value and distribution to the sphere of produced commodities had been ‘formulated and expounded in the clearest possible manner’ (1859, p. 60). Quite unlike Ricardo, Marx not only consistently avoided the use of term *Güter* (*goods*) – it is hardly possible to forget that the first chapter of *Capital* bears the title *Wares* – but actually considered the theoretical consequences of these terminological conventions in the *Contribution to the Critique of Political Economy*. Though not especially satisfying in itself, the theoretical argumentation of the *Critique* is

simple enough: ‘use-value as such, since it is independent of the determinate economic form, lies outside the sphere of political economy’ (1859, p. 28).

Goods

As the basis of the theory of value shifted away from the old classical idea of production as a circular process, towards the newer and different idea of an economic process resembling a one-way street – from ‘factors of production’ to ‘goods’ – there began simultaneously a retreat from the examination of exchangeable value in terms of the objective conditions surrounding the production of commodities, and an advance towards a theory of value grounded in the subjective conditions surrounding the consumption of goods. Of course, this was not entirely an unprecedented idea (the work of Lauderdale and Bailey comes to mind), what is different is the fact that these notions are now placed on a firmer theoretical footing than had hitherto been the case and that they come to form the mainstream of the discipline.

The orientation thereby imparted to the theory of exchangeable value by the economists in the vanguard of this change took as its starting point precisely those passages of the *Wealth of Nations* which had been so important in establishing the conceptual apparatus of the earlier classical economists. But the lesson that was drawn from them was not that which had been drawn by Smith. They, too, were keenly interested in the properties of the actual objects of market exchange, but from Smith's water–diamonds paradox they did not reach the classical conclusion, but rather one that held that the joint conditions of scarcity and utility would act to determine relative prices. As Pareto was eloquently to put it, economics became the study of equilibrium between man's tastes and the obstacles to satisfying them. Exchangeable value, to borrow Jevons's terminology, would be determined by the final degree of utility.

The 1870s were, of course, the years in which the basic provisions of the new constitution of economics were laid down almost simultaneously in Britain, France and Germany. Precursors had been sought out and honoured by those in the vanguard of the new theory, and the battle against the ‘noxious influence of authority’, as Jevons put it, already promised success – even the sterner opposition of the historical school was beginning to seem less formidable. Yet despite these quite rapid developments, in the initial years of the marginal revolution the language and usage in English economics seems to have remained essentially as it had been in the classical period. An example may serve to highlight the point.

William Stanley Jevons, who by the second edition of his *Theory of Political Economy* in 1879 had succeeded in substituting ‘economics’ for the older term ‘political economy’ in everything but the title of his book, retained the substantive *commodities* even though it was to their want-satisfying qualities that he wished to defer in his explanation of exchangeable value. Usage of the term *goods*, which conveys with greater accuracy the theoretical conceptions at the base of this new approach to the theory of value, appears to have been consciously avoided by Jevons. There is a particularly interesting passage from the *Theory of Political Economy* that illustrates the degree to which Jevons grappled with the language in which to express his theory:

> It will be allowable … to appropriate the good English word *discommodity*, to signify any substance or action which is the opposite of *commodity*, that is to say, *anything which we desire to get rid of* … Discommodity is, indeed, properly an abstract form signifying inconvenience or disadvantage. (1871, p. 114, italics in original)

It is impossible to resist the temptation to add that ‘the good English word’ *discommodity* is of Latin origin, and that the formal introduction of the simple Old English word *goods* at this juncture would have relieved Jevons of the need to conduct such linguistic exercises (see also, Jevons, 1882, p. 11).

Nevertheless, the example is sufficient to indicate that in the English language at least, *goods* was not at this time in formal use in theoretical economics.

Despite this, however, the appearance of the term goods in the formal literature of economics is inextricably linked with the rise of the neoclassical theory of exchange and demand. But to see how this is so, it is necessary to turn to the writings of German economists of the new school.

In the German neoclassical literature, quite precise definitions were given for the formal usage of the substantives *Gut* and *Güter*. Carl Menger's *Grundsätze der Volkswirtschaftslehre* (1871) provides a particularly striking example of this:

> Diejenigen Dinge, welche die Tauglichkeit haben in Causalzusammenhang mit der Befriedigung menschlicher Bedürfnisse gesetzt zu werden, nennen wir Nützlichkeiten, wofern wir diesen Causalzusammenhang aber erkennen und es zugleich in unserer Macht haben, die in Rede stehenden Dinge zur Befriedigung unserer Bedürfnisse tatsächlich heranzuziehen, nennen wir sie Güter. (1871, pp. 1–2)

Menger, in fact, went so far as to devote an exceedingly long footnote (printed as an appendix to the *Werke* edition of the *Grundsätze*) to the history of the usage of this term in this sense.

How and when the equivalent term entered the formal language of English economics – and when it might be said to have established itself – is not a difficult question to answer. Alfred Marshall's *Principles* (1890) seems to be the innovator.

In a passage from the second edition of the *Principles* dating from 1891, Marshall remarked that lacking any short term in common use to represent all desirable things, that is ‘things that satisfy human wants’, he proposed ‘to use the term *Goods* for that purpose’ (1961, I, p. 54, italics in original). In the second edition Marshall appended a footnote to the effect that he intended to replace the singular *commodity* with the term *good*, and gives as explicit justification for this the correspondence between his usage and that of the German economists (see 1961, II, p. 185e). This appears to be the first systematic application of the term *goods* in the formal terminology of economic theory – what is more, its usage is derived from the German. Note that a substantive singular form also appears.

It would seem reasonable to conclude, therefore, that it is from this source (that is, from Marshall's *Principles*) that the term *goods* gained wide circulation in economic theory. The date by which it might be reckoned to have established itself would appear to be around the mid-1890s, as it was in the fourth edition of the *Principles* in 1898 that Marshall chose to delete the footnote alluded to in the previous paragraph. This would accord broadly with the date at which the original edition of this *Dictionary* appeared containing James Bonar's entry under the heading *goods*. It is interesting to note that by the end of the 1920s the term had been so fully absorbed into the language of economic theory that Robbins chose to omit from the English edition of Wicksell's *Lectures* a paragraph where the question of this terminology is discussed. According to Robbins's editorial note, this paragraph was ‘of no interest to English readers’ (Wicksell, 1934, vol. 1, p. 15 n.l).

The originators of the modern theory of exchange and demand, nevertheless, had established a terminology through which to convey one of the basic tenets of their argument. Not only did they appreciate that goods are goods, but they expended a considerable amount of time and energy establishing that the subjective conditions of consumption were the appropriate place to locate the analysis of exchangeable value.

Characteristics

This brings us back to Harry Johnson's theorem. The grounding of a theory of exchangeable value upon the notion of *goods* was taken in certain to require a closer specification of the want-satisfying qualities of the 'things' which are daily exchanged on markets – since these are, in the final analysis, the *goods* which form the subject of the examination. When one contemplates the kinds of developments in the theory of exchange which might contribute towards the fulfilment of this requirement, nearly all of them seem to entail a widening of the gap between the actual 'things' which are exchanged on markets, and their want-satisfying characteristics which are the real subjects of demand. This, of course, is the direction in which the characteristics theory of demand has already taken us. Its problematic, of course, is to establish a transformation from characteristics to the actual objects through which these characteristics are transmitted. In the language of Lancaster, what is required is a well-defined mapping from the characteristics space to the goods space – since in the end the prices that are thrown up on markets are attached to actual objects and not to their characteristics. The theory of separable utility functions has been of immense assistance in this regard.

However, if the classification of these characteristics could be rendered sufficiently fine, then a concomitant implication would seem to be that the idea of securing a theory of the prices attached to actual objects exchanged on markets would need to be sought in some other direction. Otherwise, we should be left with a theory of exchange and demand which made no contact, even at an abstract theoretical level, with the material realities of market exchange in modern economies. We should certainly have a theory of *goods*, but to what form of economic organization such a theory could be held to apply, if any, is not at all obvious.

So that such speculations should not be thought to be idle, it is interesting to note that language is not only a vehicle for the transmission of theoretical conceptions; it is often the vehicle through which a whole array of structural and cultural data about social interaction is conveyed. Exchange in different kinds of societies frequently embodies these complex social relations – so much so that the familiar idea of economists of the modern school that a universally applicable analysis of exchange is somehow desirable, or even possible, would seem to be fraught with pitfalls.

See Also

- exchange

Bibliography

Jevons, W.S. 1871. *The Theory of Political Economy*, ed. R.D.C. Black. Harmondsworth: Penguin, 1970.

Jevons, W.S. 1882. *The Principles of Economics*, ed. H. Higgs. New York: Kelley, 1965.

Johnson, H. 1958. Demand theory further revisited or goods are goods. *Economica* 25, 149.

Lancaster, K.J. 1966. Change and innovation in the technology of consumption. *American Economic Review, Papers and Proceedings* 56(1–2), 14–23.

Maitland, J. (Earl of Lauderdale). 1804. *An Inquiry into the Nature and Origin of Public Wealth*. New York: Kelley, 1962.

Malthus, T.R. 1827. *Definitions in Political Economy*. New York: Kelley, 1963.

Marshall, A. 1961. *Principles of Economics*. 9th variorum edition. London: Macmillan. (1st edn, 1890.)

Marx, K. 1857. *Grundrisse*. Harmondsworth: Penguin, 1973.

Marx, K. 1859. *A Contribution to the Critique of Political Economy*. London: Lawrence & Wishart, 1971.

Marx, K. 1867. *Capital*. 4th edn, 1909. 3 vols, New York: International Publishers, 1967.

Mauss, M. 1925. *The Gift*. Ed. E.E. Evans-Pritchard. London: Routledge & Kegan Paul, 1970.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. In *The Collected Works of Carl Menger*, vol. 1. London: LSE Reprints No. 17, 1934.

Mill, J.S. 1848. *Principles of Political Economy*. Peoples edn. London: Longmans, 1873.

Palgrave, R.H.I., ed. 1894–9. *Dictionary of Political Economy*, 3 vols. London: Macmillan.

Roscher, W. 1854. *Principles of Political Economy*. Trans. from the 13th German edn, Chicago: Callaghan and Coy, 1882.

Say, J-B. 1803. *A Treatise on Political Economy*. Trans. C.R.C. Prinsep, New York: Kelley. Reprinted, 1971.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vols, ed. E. Cannan. London: Metheun, 1961.

Sraffa, P. (ed. with the collaboration of M.H. Dobb) 1951–73. *The Works and Correspondence of David*

Ricardo. Cambridge: Cambridge University Press.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Wicksell, K. 1934. *Lectures on Political Economy*, 2 vols, ed. L. Robbins. London: Routledge & Kegan Paul.

How to cite this article

Milgate, Murray. "goods and commodities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000080> doi:10.1057/9780230226203.0657

The New Palgrave Dictionary of Economics Online

Gordon, Robert Aaron (1908 – 1978)

B.G. Hickman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; business cycles; Cowles Commission; econometric modeling; Gordon, R. A.; National Bureau of Economic Research; price theory

Article

Gordon was born 26 July 1908 in Washington, DC, and died 7 April 1978 in Berkeley, California. He was a policy-oriented economist whose research style was quantitative but not econometric, and whose influence was felt not only through his basic research but also as a dedicated and tireless public servant. After his formative years as graduate student and instructor at Harvard University, 1929–38, he accepted a position at the University of California (Berkeley), where he remained until his retirement from teaching in 1976.

His early work in industrial organization culminated in the influential volume on *Business Leadership in the Large Corporation* (1944), which is noteworthy for its pioneering use of empirical data in the field. Another major strand is his work on unemployment, its structural and cyclical causes, and the goal of full employment (1962; 1967).

His lifelong research interests were primarily focused, however, on business cycles and their causes. He championed the quantitative-historical method of cycle analysis over the National Bureau of Economic Research (Burns–Mitchell) and econometric modelling (Cowles Commission) approaches (1949), and he devoted much of his career to implementing his approach in studies of the business fluctuations of the interwar period (1951) and before and after the Second World War (1955a; 1969; 1974). His eclectic analysis of the causes of business cycles emphasized the Schumpeter–Hansen distinction between major and minor business cycles (1956) and attributed the major cycles to the appearance and exhaustion of investment opportunities in particular industries (1955b).

He combined a formidable talent for economic analysis with a sense of historical and institutional relevance and was impatient with the tendency he discerned in the profession at large to favour rigour over relevance in economic theorizing. This impatience was expressed at an early stage with regard to price theory (1948), reiterated in his presidential address to the American Economic Association on

‘Rigor and Relevance in a Changing Institutional Environment’ (1976), and repeated in his last piece on ‘A Skeptical Look at the “Natural Rate” Hypothesis’ (1978).

His contributions to the public weal were many and lasting, but two in particular may be cited. In 1956–59 Gordon undertook a massive study of business education jointly with James E. Howell for the Ford Foundation, and their 1959 report provided the stimulus for a radical reorientation of MBA programmes in graduate business schools toward the use of analytical methods drawn from economics, statistics, and the behavioural sciences. He also served as chair of the President's Committee to Appraise Statistics on Employment and Unemployment in 1961–62, which led to important reforms in the nation's statistics in this vital area.

Selected works

1944. *Business Leadership in the Large Corporation*. Washington, DC: Brookings Institution.

1948. Short-period price determination in theory and practice. *American Economic Review* 38, 265–88.

1949. Business cycles in the interwar period: the ‘quantitative-historical’ approach. *American Economic Association, Papers and Proceedings* 39(May), 47–63.

1951. Cyclical experience in the interwar period: the investment boom of the twenties. In Universities–National Bureau Committee, *Conference on Business Cycles*, New York: NBER.

1955a. Investment opportunities in the US before and after World War II. In *The Business Cycle in the Post-War World*, ed. Erik Lundberg. New York: Macmillan.

1955b. Investment behavior and business cycles. *Review of Economics and Statistics* 37, 23–34.

1956. Types of depression and programs to combat them. In Universities–National Bureau Committee, *Policies to Combat Depression*. Princeton: Princeton University Press.

1959. (With James E. Howell.) *Higher Education for Business*. New York: Columbia University Press.

1962. (Co-author.) *Measuring Employment and Unemployment*. Report of the President's Committee on Employment and Unemployment Statistics. Washington, DC: Government Printing Office.

1967. *The Goal of Full Employment*. New York: Wiley.

1969. The stability of the U.S. economy. In *Is the Business Cycle Obsolete*? ed. Martin Bronfenbrenner. New York: Wiley.

1974. *Economic Growth and Instability: The American Record*. New York: Harper.

1976. Rigor and relevance in a changing institutional setting. *American Economic Review* 66, 1–14.

1978. A skeptical look at the ‘natural rate’ hypothesis. In *Economic Theory for Economic Efficiency: Essays in Honor of Abba P. Lerner*, Boston: MIT Press.

How to cite this article

Hickman, B. G. "Gordon, Robert Aaron (1908–1978)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000082> doi:10.1057/9780230226203.0658

The New Palgrave Dictionary of Economics Online

Gorman, W.M. (Terence)

Charles Blackorby, Daniel Primont and R. Robert Russell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

W.M. (Terence) Gorman was a theorist's theorist who nonetheless thought himself to be very practical. His contributions were brilliantly original – both in conception and in technical execution – but they often were a difficult read. We sketch below two of his lifelong interests – aggregation over agents and aggregation over commodities – and discuss briefly a few of his other contributions. We hope this brief outline will encourage others to peruse his works in more detail.

Keywords

aggregation over agents; aggregation over commodities; Cambridge School; capital aggregation; characteristics; closed form representation; compensation principle; concavity; diminishing marginal utility; duality; Econometric Society; Engel curve; equivalence scales; expected utility theory; expenditure function; general equilibrium; Gorman, W. M.; Hotelling's lemma; indirect utility functions; interpersonal utility comparisons; Le Chatelier principle; marginal product of capital; price aggregation; profit functions; Roy's theorem; separability; two-stage budgeting

Article

1 Introduction

W.M. Gorman lived from 1923 to 2003. He graduated from Trinity College, Dublin, in 1948 in economics and a year later in mathematics. He held posts at the University of Birmingham, Oxford University, and the London School of Economics. Among many honours, he was the president of the Econometric Society in 1972, a fellow of the British Academy, and an honorary foreign member of the American Academy of Arts and Sciences and of the American Economic Association. (For a more detailed discussion of Gorman's career and work, see Honohan and Neary, 2003.)
Gorman was interested in economics interpreted very broadly. He read history, philosophy, mathematics, and statistics. Nevertheless, most of his work was very abstract and many readers found it difficult to follow. We begin here by discussing his long-standing interest in two kinds of aggregation, over agents and over commodities, before considering some of his other contributions.

2 Aggregation across agents

Gorman's first published paper (Gorman, 1953) provided necessary and sufficient conditions for the existence of a representative consumer. That is, given that each consumer in society has well-behaved preferences and demand functions,

$$x^h = f^h(p, m_h) \text{ for } h = 1, \ldots, H, \tag{2.1}$$

when can aggregate demand, $\Sigma_h x^h$, be generated by the demands of an aggregate agent so that

$$x = \sum_h x^h = f(p, m), \tag{2.2}$$

where $m = \Sigma_h m_h$? He demonstrated that a necessary and sufficient condition is that consumers have affine parallel Engel curves,

$$x^h = a(p) m_h + b^h(p), \tag{2.3}$$

so that the demands of the aggregate consumer can be written as

$$x = a(p) m + b(p), \tag{2.4}$$

where $b(p) = \Sigma_h b^h(p)$. In a later paper, Gorman (1961) provided a closed form representation of the preferences of consumer with parallel affine Engle curves: the indirect utility functions can be written as

$$V^h(p, m_h) = A(p) m_h + B^h(p) \text{ for } h = 1, \ldots, H, \tag{2.5}$$

and the indirect utility function of the representative consumer can be written as

$$V(p, m) = A(p) m + B(p), \tag{2.6}$$

where $B(p) = \Sigma_h B^h(p)$. Using Roy's theorem, the reader can easily check that (2.3) and (2.4) are derived from (2.5) and (2.6) respectively. The intuition behind this result is straightforward: if we take a dollar away from consumer 1 and give it

to consumer 2, aggregate demand cannot change, because total income has not changed; the two consumers therefore must have identical marginal propensities to consume. There have been many attempts to generalize this notion of a representative agent (see, for example, Muellbauer, 1976) but all suggest that aggregate demand for the *i*th commodity can be written as

$$x_i = \sum_{j=1}^{J} a_{ij}(p) b_j(m, z), \tag{2.7}$$

where *z* is a vector of characteristics. Gorman (1981) demonstrates that utility maximization requires that the rank of the matrix with elements a_{ij} be less than or equal to 3. Note that the idea of a representative consumer given in (2.3) entails that the matrix have rank 2. This proved that there is not much more that could be done to generalize this idea.

Over the course of the 1960s a battle raged between the US Cambridge and the English Cambridge over – to put it crudely – whether the marginal product of capital was a meaningful aggregate concept (see Fisher and Monz, 1992, for an extensive discussion of this issue). Gorman (1968a) approached this problem in the dual, using profit functions. His argument in favour of using the dual was that it was important to model problems in the appropriate variables. When thinking about capital aggregation he noted that the individual agents all face the same prices. Hence he thought the problem should be attacked using profit rather than production functions.

In the simplest terms, suppose that each firm has a profit function given by

$$\pi^f(p, k^f) = \max_{y^f} \left\{ p \cdot y^f \middle| y^f \in T^f(k^f) \right\} \text{for } f = 1, \ldots, F, \tag{2.8}$$

where k^f is a vector of the quasi-fixed factors of firm *f* and $T^f(k^f)$ is the technology set of the firm for each fixed value of the k^f. In this context, the marginal product of capital only makes sense if there exists an aggregate profit function that can be written as

$$\pi(p, k) = \max_{y} \left\{ p \cdot y | y \in T(k) \right\}, \tag{2.9}$$

where *k* is a scalar measure of the aggregate capital stock. Gorman showed that a necessary and sufficient condition for (2.9) is that

$$\pi^f(p, k^f) = \alpha(p)\varphi^f(k^f) + \beta^f(p) \text{for } f = 1, \ldots, F, \tag{2.10}$$

so that the aggregate capital stock is given by

$$k = \sum_f \varphi^f(k^f),$$
(2.11)

thus demonstrating that the marginal product of capital is not likely to be a meaningful concept very often. (For the implied restrictions on the technology sets and production functions, see Blackorby and Schworm, 1984.)

The intuition for this result is a little less straightforward. Think of $\varphi^f(k^f)$ as a measure of bolted-down capital of firm *f*. Then, $\alpha(p)$ is clearly the marginal product of capital in firm *f* and (2.10) requires this be the same in every firm so that the aggregate profit function can be written as

$$\pi(p, k) = \alpha(p)k + \beta(p)$$
(2.12)

where *k* is given by (2.11) and $\beta(p) = \Sigma_f \beta^f(p)$.

3 Aggregation across commodities

Gorman's best-known paper in this area (1959a) was written in response to a paper by Strotz (1957) on two-stage budgeting. Suppose that a consumer (or any organization with a well-behaved objective function and an expenditure constraint) wants to simplify its purchasing decisions as follows: first it wants to allocate funds optimally to broad categories of commodities and then later make the detailed calculations of how to spend the funds within any particular category. For example, a consumer might first decide how much money to budget for food and then later decide exactly how to allocate the food budget to particular commodities. The latter decision turns out to be equivalent to the separability of the commodities in each category, whereas the former hinges on a notion of price aggregation. (For more on two-stage budgeting and separability, see separability.)

Separability seemed a natural assumption to Gorman; it allowed the researcher to focus on a particular group of commodities without having to worry very much about anything else. In Gorman (1968b), however, he showed that this could have previously unknown implications. Suppose that one assumed that two groups of commodities were separable from their complements so that the utility function could be written as

$$U(x) = \bar{U}(U^A(x^A), x^C) \text{ and } U(x) = \hat{U}(U^B(x^B), x^D).$$
(3.1)

Suppose in addition that some of the commodities in group A are not in group B, some are in both A and B, and the others are in B but not in A. Gorman showed that this implied that the utility function could in fact be written as

$$U(x) = \overset{*}{U}(U^a(x^a) + U^b(x^b) + U^c(x^c)), \tag{3.2}$$

where the variables in *a* are those that are in A and not in B, those in *b* are in B but not in A, and those in *c* are in both groups. This surprising result can be taken in two ways: (1) perhaps too much separability is a dangerous thing but (2) on the other hand this theorem provides a fundamental and clean way of modelling additivity.

Ever attuned to different ways of looking at things, Gorman (1970), in a paper published only in 1995, noted that if a quasi-concave utility function could be written additively, as would often be required to study intertemporal or uncertain events,

$$U(x) = \overset{*}{U}\left(\sum_{s=1}^{S} U^s(x^s)\right), \tag{3.3}$$

then at most one of the functions U^s could fail to be concave. Thus, for example, if the U^s functions were the same as in expected utility theory (that is, linear transformations of one another), then quasi-concavity would imply concavity. Again this can be taken two ways – as an easy way to obtain concavity or as the danger of too much separability.

4 General interests

Gorman was well known for having widely circulated and widely cited but unpublished papers. Perhaps, the best known of these was ‘A Possible Procedure for Analysing Quality Differentials in the Egg Market’, a 1956 working paper of the Iowa Agricultural Experiment Station. (This paper was eventually published as Gorman, 1980. A more complete discussion can be found in Honohan and Neary, 2003.) Suppose that it is not the commodities that consumers want but rather the characteristics embodied in them. Thus type *i* egg would contain a_i of characteristic *A*, b_i of characteristic *B*, and so on. Further, Gorman assumed that if one bought x_1 of type 1 egg and x_2 of type 2, then the total amount of characteristic A would be $a_1x_1+a_2x_2$. Thus, at arbitrary prices only two types of eggs would be purchased if there were only two characteristics that were relevant and three types if three characteristics were relevant, except in the degenerate case where relative prices were just right. Then the consumer would be indifferent between two or three or more types of eggs. Gorman argued that equilibrium prices should not contain any arbitrage opportunities and hence the degenerate case would be the normal one. This then suggested an agenda for empirical work that he and students pursued over the years. (Not very much of this research actually surfaced, although a 1959 University of Birmingham working paper entitled ‘Demand for Fish: an Application of Factor Analysis’ and his 1972 presidential address to the Econometric Society, ‘A Sketch for the Demand for Related Goods’, were widely cited for some time.)

Although Gorman wrote and published little in welfare economics, what he did write demonstrated a profound understanding of the issues involved. In order to avoid the known problem with the Kaldor (1939) compensation principle (namely, that situation B could be preferred to situation A by the Kaldor compensation criterion and that A could be preferred to B by the same criterion), Scitovsky (1941) had proposed a new test that required that B should be preferred to A if B was preferred by the Kaldor compensation test and A was not preferred to B by the same test. In Gorman (1955) he demonstrates in a very elegant manner that the Scitovsky criterion is intransitive; that is, that B could be preferred to A by the Scitovsky criterion, C preferred to B by the same criterion, and then that A could be preferred to C by the same test.

In a rather more philosophical vein, Gorman (1959b) presents a series of arguments that might lead one to think that social indifference curves are convex. In this he was clearly aware of the problems of interpersonal comparisons of utility and the idea of diminishing marginal utility for each individual. Even now this makes for a thought-provoking read.

Although Gorman did very little research in general equilibrium, when preparing a paper for a Festschrift in honour of Ivor Pierce he felt that only a general equilibrium paper would be appropriate. Gorman (1984) prepared an analysis of the Le Chatelier principle in general equilibrium. To give the reader the flavour of his elegant argument, let $\Pi(p)$ and $\pi(p)$ be the long-run and short-run profit functions of a firm. It seems natural to assume that

$$\Pi(p) \geq \pi(p). \tag{4.1}$$

Suppose that at $p = \bar{p}$ we are at a long-run equilibrium, so that

$$\Pi(\bar{p}) = \pi(\bar{p}). \tag{4.2}$$

Thus, the price vector $p = \bar{p}$ minimizes the difference between long-run and short-run profits, $\Pi(p) - \pi(p)$. The second-order condition for this minimization problem is given by

$$\sum_i \sum_j \Pi_{ij}(\bar{p})\,\theta_i\theta_j \geq \sum_i \sum_j \pi_{ij}(\bar{p})\,\theta_i\theta_j. \tag{4.3}$$

By Hotelling's lemma, this implies immediately that

$$\frac{\partial \bar{X}_i}{\partial p_i} \geq \frac{\partial \bar{x}_i}{\partial p_i}; \tag{4.4}$$

that is, the long-run response of commodity *i* to a change in its price is greater than or equal to the short-run response – the Le Chatelier principle.

Technically, much of Gorman's work was difficult; he frequently employed transformations of variables and functions with such speed that the reader felt, if not lost, at least dizzy. In Gorman (1976) he wrote what he thought of as a reasonable introduction to some of these tools and called the paper ‘Tricks with Utility Functions’. We conclude with a discussion of one of these tricks and the lesson that Gorman thought could be learned from it.

Gorman begins his discussion of equivalent adult scales (equivalence scales) with a quote from a former schoolmaster who said ‘When you have a wife and a baby, a penny bun costs threepence’. Consider a family of type $a=(a_1, \ldots, a_n)$ whose utility function could be written – after Barten (1964) – as

$$u_a = U^a(x) = U(x^a),$$

(4.5)

where the adjusted consumption vector,

$$x^a = \left(\frac{x_1}{a_1}, \ldots, \frac{x_N}{a_N}\right),$$
(4.6)

corrects for the number of equivalent adults. Note that the second equal sign implies that all households have the same utility function but defined on the adjusted variables. The expenditure function dual to U^a in (4.5) is given by

$$E^a(u_a, p) = \min_x \left\{ p_1 x_1 + p_2 x_2 + \ldots + p_N x_N \middle| U^a(x) \geq u_a \right\} = \min_{x^a} \left\{ a_1 p_1 x_1^a + \ldots + a_N p_N x_N^a \middle| U(x^a) \geq u_a \right\} = E(u_a, p^a),$$
(4.7)

where $p^a=(a_1p_1, \ldots, a_Np_N)$ and E is the expenditure function dual to U in (4.5). Thus, for a family of three bread-equivalent adults ‘a penny bun costs threepence’. The adjusted compensated demand for good i is

$$x_i^a = E_i(u_a, p^a),$$
(4.8)

so that the ordinary compensated demand is

$$x_i = a_i E_i(u_a, p^a).$$
(4.9)

From this, the compensated demand elasticities can be written as

$$\varepsilon_{ij} = \delta_{ij} + \alpha_{ij},$$
(4.10)

where δ_{ij} is the Kronecker delta and

$$\alpha_{ij} = \frac{\partial \ln x_i^a}{\partial \ln a_j} \tag{4.11}$$

at a fixed u_a is the compensated elasticity with respect to family size. It is easy from here to calculate the uncompensated elasticities as well. From this Gorman concludes that ‘Were the theory true, and were the sample to include a great enough variety of family types, we could use them to calculate the price elasticities from survey data. As long as everyone faces the same prices, we need not even know what they are’. If one had tried to do this analysis, as Barten did, working only with (4.5) the simplicity of this model would not have been exposed.

This paper had dozens of such ‘tricks’, techniques that use either duality or separability arguments, and we recommend them to the reader.

See Also

- duality
- Engel curve
- equivalence scales
- indirect utility function
- separability

Selected works

1953. Community preference fields. *Econometrica* 51, 63–80.

1955. The intransitivity of certain criteria used in welfare economics. *Oxford Economic Papers* 11, 23–55.

1959a. Separability and aggregation. *Econometrica* 27, 469–481.

1959b. Are social indifference curves convex? *Quarterly Journal of Economics* 73, 485–96.

1961. On a class of preference fields. *Metroeconomica* 13, 53–6.

1968a. Measuring the quantities of fixed factors. In *Value, Capital and Growth: Papers in Honour of Sir John Hicks*. Edinburgh: Edinburgh University Press.

1968b. The structure of utility functions. *Review of Economic Studies* 32, 369–90.

1970. The concavity of additive utility functions. Manuscript, University of North Carolina. In Gorman (1995).

1976. Tricks with utility functions. In *Essays in Economic Analysis*, ed. M. Artis and R. Nobay. Cambridge: Cambridge University Press.

1980. The demand for related goods: a possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies* 47, 843–56.

1981. Some Engel curves. In *Essays in the Theory and Measurement of Demand in Honour of Sir Richard Stone*, ed. A. Deaton. Cambridge: Cambridge University Press.

1984. Le Chatelier and general equilibrium. In *Demand, Equilibrium and Trade: Essays in Honour of Ivor F. Pearce*, ed. A. Ingham and A.M. Ulph. London: Macmillan.

1995. *The Collected Works of W.M. Gorman*, vol. 1, ed. C. Blackorby and A. Shorrocks Oxford: Oxford University Press.

Bibliography

Barten, A. 1964. Family composition, prices, and expenditure patterns. In *Econometric Analysis for National Economic Planning*, ed. P.E. Hart, G. Mills and J.K. Whitaker. London: Butterworths.

Blackorby, C. and Schworm, B. 1984. The structure of economies with aggregate measures of capital: a complete characterization. *Review of Economic Studies* 51, 633–50.

Fisher, F.M. and Monz, J. 1992. *Aggregate Production Functions and Related Topics*. Cambridge, MA: MIT Press.

Honohan, P. and Neary, J.P. 2003. W.M. Gorman. *Economic and Social Review* 34, 195–209.

Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49, 549–52.

Muellbauer, J. 1976. Community preference fields and the representative consumer. *Econometrica* 44, 979–99.

Scitovsky, T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* 9, 77–88.

Strotz, R. 1957. The empirical implications of a utility tree. *Econometrica* 25, 269–80.

How to cite this article

Blackorby, Charles, Daniel Primont and R. Robert Russell. "Gorman, W.M. (Terence)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000197> doi:10.1057/9780230226203.0659

The New Palgrave Dictionary of Economics Online

Gossen, Hermann Heinrich

Jürg Niehans
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Gossen, H. H.; imputation; intermediate products; invisible hand; libertarianism; marginal rate of substitution; marginal revolution; marginal utility; optimal allocation; optimal saving; rent; time preference; value theory

Article

Gossen was born in Düren (between Aachen and Cologne) on 7 September 1810; he died in Cologne on 13 February 1858. Little is known about his life, partly because the inconspicuous bachelor did not attract attention, partly because most of those who had known him were dead by the time he became famous, partly also because his literary remains, scant as they must have been, are lost. The principal biographical source is the essay by Walras (1885). The available facts are admirably surveyed by Georgescu-Roegen (1983), on whose masterly introduction to the English translation of Gossen's book the following life sketch is mostly based.

Gossen's father was a tax collector under Napoleon and subsequently the Prussian administration; later he managed his wife's estate near Godesberg. Hermann obtained a good high-school education, showing ability in ‘elementary mathematics’, but his mathematical training never went beyond that level. Since his father insisted on a government career in the tradition of his forebears, his university studies in Bonn and Berlin concentrated on law and government.

In 1834, Gossen entered the civil service as a ‘Referendar’ (junior law clerk) in Cologne. While he seems to have been a well-mannered young man, the performance of his duties left much to be desired. He simply had no interest in a government career and loved the good things in life. There were complaints and reprimands, and the promotion to the rank of ‘Regierungsassessor’ came rather later than usual. Finally, in 1847, though his superiors seem to have shown considerable sympathy, he had no choice but to resign.

The transition to a new career was perhaps eased by his father's death, which spared him recriminations about his failure and provided him with the means for a new start. Gossen went to Berlin, where he seems to have sympathized with the liberal revolution, and then returned to Cologne as a partner in a

new accident insurance firm. He soon withdrew from the firm, but continued to devise grandiose insurance projects.

Living with his two sisters, Gossen now devoted most of his energies to developing the unorthodox ideas he had expressed in his civil service examination papers into his magnum opus. The preface suggests that he hoped this would not only make him the Copernicus of the social universe but also open the door to an academic career. In 1853 an attack of typhoid fever undermined his health, and the disappointment about the fate of his book depressed him. Death came from pulmonary tuberculosis. He seems to have been an amiable, sincere and idealistic human being with broad interests, including music and painting. Brought up a Catholic, he developed into an enthusiastic hedonist. Dreaming of reforming the world, he lacked the force to conquer it.

The *Entwickelung der Gesetze des menschlichen Verkehrs* was published in 1854 at Gossen's expense by the publisher Vieweg in Brunswick. Very few copies were sold and the book remained unnoticed for years. Shortly before his death, Gossen withdrew it from circulation and the unsold copies were returned to him. After the author had become famous, Vieweg's successor, Prager, bought this stock from Gossen's nephew, a professor of mathematics by the name of Hermann Kortum, and put it on the market again with a new title page, as a ‘second edition’, in 1889. There is an Italian translation by Tullio Bagiotti and there is now, since 1983, a careful English translation by Rudolph C. Blitz, nicely divided into chapters. The manuscript of a French translation by Walras was apparently lost.

The first known references to Gossen's book were by Julius Kautz (1858/60), but they only show that their author did not understand the problems Gossen had solved. Slightly more understanding was shown by F.A. Lange, but again in no more than a footnote. Fortunately, Kautz's reference was seen by Robert Adamson, who was able to get hold of a copy and reported its content to Jevons. In the second edition of *The Theory of Political Economy* (1879) Jevons included a generous acknowledgement of Gossen's priority ‘as regards the general principles and method of the theory of Economics’, which became the ignition point of Gossen's posthumous fame. Though Gossen's name became famous, his book remains largely unread to this day.

At the level of individual behaviour, Gossen's basic theoretical problem concerns optimization with limited resources (references are to the 1889 edition; they are followed by the corresponding references to the English translation, marked T). Resources are first visualized as time (p. 1 f.; T ch. 1). The given lifetime has to be allocated to enjoyable activities in such a way that lifetime enjoyment or, in modern terminology, utility, is maximized.

For a given activity, marginal utility is assumed to be a declining function of the time spent on it. In Gossen's words, ‘The magnitude of a given pleasure decreases continuously if we continue to satisfy this pleasure without interruption until satiety is ultimately reached’ (p. 4f.; T p. 6). This is the postulate Wilhelm Lexis (1895) christened ‘Gossen's First Law’. *In itself*, it was neither new nor profound. W.F. Lloyd had expressed it 20 years earlier just as clearly, it had a long ancestry reaching back to Bentham, the French ‘subjectivists’, Daniel Bernoulli, and the scholastics, and it is essentially commonplace. To simplify, Gossen assumes marginal utility curves to be linear. It is important to note that Gossen's curves do *not* describe the decline in the marginal utility of a good as its quantity increases, but the decline in the utility from the marginal unit of resources as the quantity of resources is increased. While this facilitated the analysis in some respects, it became a crucial handicap in others. Gossen realized that each of these marginal utility functions must be thought of as being derived by solving a suboptimization problem, inasmuch as time allocated to a given activity must be spent in the most

enjoyable way, probably with interruptions. However, his analysis of this difficult sub-problem, though original and suggestive, remained incomplete and unsatisfactory, leaving much to do for future research on the allocation of time.

Gossen recognized at once that a necessary condition for the optimal allocation of resources is the equality of the marginal utilities in different activities. This is ‘Gossen's Second Law’, which he had printed in heavy type: ‘The magnitude of each single pleasure at the moment when its enjoyment is broken off shall be the same for all pleasures’ (p. 12; T p. 14). This theorem is Gossen's principal claim to fame. In it he had no forerunners. It was the key that opened the door to a fruitful analytical use of the First Law and thus initiated the ‘marginal revolution’ in the theory of value.

The third stage of Gossen's analysis is reached with the introduction of exchange (p. 80 f.; T ch. 7). Gossen begins with the bilateral case. He immediately perceives that there are many different opportunities for mutually beneficial exchange, but his discussion of these possibilities is, understandably, inconclusive. As a necessary condition for optimal exchange he postulates that the marginal utilities must be equalized between individuals for each product. While this formulation requires both cardinality and interpersonal comparability of utility, its economic substance, since it can be expressed in terms of ‘marginal rates of substitution, is independent of these assumptions. The concept of a ‘contract curve’, however, is not used. The statement that each individual would usually be willing to forego a portion of what he receives suggests some notion of consumer's surplus.

The analysis is finally extended to market exchange, where each individual can exchange goods and effort at parametrically given prices, expressed in a common numéraire called money. We thus end up with the optimization problem that became the banner of the ‘marginal revolution’. The ‘Second Law’ can then be expressed by the condition that ‘the last atom of money creates the same pleasure in each pleasurable use’ (p. 93f.; T p. 109).

The solution to this problem determines the individual's market demand and supply for each product and effort. Gossen also shows how the value of intermediate products can sometimes be derived from that of the final goods, thereby foreshadowing Menger's theory of ‘imputation’, but he is careful to note that the market mechanism works even where imputation fails (p. 24f.; T p. 28f). If prices are specified at random, aggregate demand and supply will generally differ. Gossen explains how this exerts pressure on prices until all markets are cleared. Prices are thus endogenously determined by general equilibrium. This argument, though concise, is presented in verbal form only. The mathematical formulation of general equilibrium, foreshadowed by Cournot, had to wait for Walras.

In the fourth stage, Gossen introduces rent (p. 102 f; T chs 8–12). If the profundity of an economist can be gauged by his treatment of rent, he comes out near the top. The worker is assumed to own a specific piece of land. Suppose he is now offered the use of land at a superior location, owned by another individual. This does not affect his utility curves, but for the amount of effort for which the marginal utility of effort is just zero he can now earn a higher income. At the same time the marginal disutility curve becomes flatter because the same total enjoyment is now spread over a higher income. In the absence of rent, the superior location would, of course, promise higher income. However, moves to superior locations are not free, but cost rent. This means that at the new location the individual has to earn a certain amount before he can even begin to buy commodities.

What is the maximum rent an individual is willing to pay for a superior location? This ‘warranted rent’ is reached at the point where total utility at the superior location is equal to the total utility at the original location. Gossen shows algebraically that with rent at the warranted level, superior locations are

associated with higher earned income and higher consumption.

While Gossen developed a novel and fruitful way to incorporate rent into a general equilibrium framework, his theory of rent is less rich than von Thünen's, published 28 years before. Gossen was no better in reading his predecessors than the later ‘marginalists’ were in reading Gossen.

The fifth stage introduces capital and interest (p. 114; T ch. 13). The basic question concerns the highest amount of present utility that could be sacrificed for a piece of land with a given annual rent, continuing into the distant future. Gossen finds the answer by discounting the utility of each future rent payment at the appropriate rate of psychological time preference (as we would call it), reflecting uncertainty of expectations (pp. 30, 115; T pp. 35, 134). This promising idea is not successfully exploited, however, and the adaptation of the land paradigm to capital goods remains sketchy. Gossen thinks in terms of land and labour, while capital goods are played down (p. 172; T p. 194). He also makes an effort to determine the optimal amount of saving by the condition that the highest price the individual is willing to pay for a source of rent should be equal to the market price, but he seems to confuse average and marginal concepts and the sense of his argument remains obscure.

In an effort to interpret everyday observations in the light of his theory, Gossen offers an elaborate discussion of the effect of price changes on demand and expenditure. This discussion anticipates a lot of later work on demand elasticities, but it is also cumbersome. The reason is that Gossen's analytical engine, while permitting a brilliantly simple determination of the optimal budget at given prices, is ill suited for the analysis of price change. Since Gossen's curves, as observed above, relate the marginal utility of expenditure to expenditure, they have to be redrawn after each price change. The insights which Marshall's apparatus made so easy to communicate, remained virtually incommunicable for Gossen. This may be one of the main reasons why his achievement, though at the highest intellectual level, remained sterile. If he had read Cournot, his fate might have been different.

The second part of Gossen's book is largely devoted to social philosophy and policy. It shows its author as a passionate libertarian. Through free markets, mankind would succeed without effort where all socialist planning must fail, namely in reaching the highest possible happiness. Abhorring all forms of protection, Gossen was in favour of free trade, the protection of property rights and a liberal education for both sexes. To prevent fluctuations in the value of money, he advocated a metallic currency and the abolition of paper money. That he also asked for restrictions on child labour and government sponsorship of credit unions seems to indicate that he knew externalities and market imperfections when he saw them. Competitive equilibrium was for him much more than an economic theory or an ideology; it was the gospel, revealing the perfection of a benevolent creator. For him, the ‘invisible hand’ was not a didactic metaphor, but religion itself. Today, this apotheosis of competition, in language closer to a revival meeting than to scientific discourse, strikes one as bizarre.

Major sources of inefficiency, Gossen thought, were distortions in the allocation of land, preventing land from actually being used by the potentially most efficient user. To correct this defect, he proposed that the government use borrowed money to buy land on the free market and then lease it to the highest bidder (p. 250 f.; T ch. 23). Since governments differ from individuals by (1) being immortal, (2) having a higher credit rating, and (3) a lower time preference, such a scheme, he argued, would actually improve government wealth, and the initial debt could eventually be repaid out of rising rent income. For a given year, the scheme would be viable if the price paid by government for a piece of land did not exceed the sum of the rent and the annual increase in the value of land, capitalized at the market rate of interest. However, Gossen was not a ‘land socialist’; he was not concerned about ‘land monopoly’ and

the ‘socialization of rent’. His objective was the correction of a market imperfection without any limitation of property rights.

Gossen, though perhaps not quite a genius, had a brilliant, original and precise mind. With his one book, he moved constrained optimization into the centre of the theory of value and allocation, where it has since remained. With respect to economic content, his was probably the greatest single contribution to this theory in the 19th century. He failed, however, to develop the basic principle into a usable analytical engine. As a consequence, the so-called ‘founders’ of the modern theory of value had to rediscover those principles before they could proceed with their engineering work.

Selected works

1854. *Entwickelung der Gesetze des menschlichen Verkehrs, und der daraus fliessenden Regeln für menschliches Handeln*. Braunschweig: Vieweg. Reprinted, Amsterdam: Liberac, 1967. 2nd edn, Berlin: Prager, 1889. 3rd edn (with introduction by F.A. Hayek), Berlin: Prager, 1927. Italian translation by T. Bagiotti as *Ermanno Enrico Gossen, Sviluppo delle leggi del commercio umano*, Padua: CEDAM, 1950. English translation by R.C. Blitz (with introductory essay by N. Georgescu-Roegen) as *The Laws of Human Relations and the Rules of Human Action Derived Therefrom*, Cambridge, MA: MIT Press, 1983.

Bibliography

Bagiotti, T. 1957. Reminiszenzen anlässlich des hundertsten Jahrestages des Erscheinens des Buches von Gossen. *Zeitschrift für Nationalökonomie* 17, 39–54.

Bousquet, G.-H. 1958. Un centenaire: l'oeuvre de H.H. Gossen (1810–1858) et savéritable structure. *Revue d’économie politique* 68, 499–523.

Edgeworth, F.Y. 1896. Gossen, Hermann Heinrich. In *Dictionary of Political Economy*, vol. 2, ed. R.H. Inglis Palgrave. London: Macmillan.

Georgescu-Roegen, N. 1983. Introduction to H.H. Gossen, *The Laws of Human Relations and the Rules of Human Action Derived Therefrom*, trans. R.C. Blitz. Cambridge, MA: MIT Press.

Jevons, W.S. 1879. *The Theory of Political Economy*. 2nd edn. London: Macmillan.

Kautz, J. 1858, 1860. *Theorie und Geschichte der National-Oekonomik*. 2 vols. Vienna: Gerold.

Krauss, O. 1910. Gossen, Hermann Heinrich. In *Allgemeine Deutsche Biographie*, vol. 55. Leipzig: Duncker & Humblot.

Lange, F.A. 1875. *Die Arbeiterfrage. Ihre Bedeutung für Gegenwart und Zukunft*. 3rd edn. Winterthur: Bleuler-Hausheer.

Lexis, W. 1895. Art. Grenznutzen. *Handwörterbuch der Staatswissenschaften*, vol. 1 (Supplement). Jena: Fischer.

Liefmann, R. 1910. Hermann Heinrich Gossen und seine Lehre. Zur hundertsten Wiederkehr seines Geburtstages am 7. September 1910. *Jahrbücher für Nationalökonomie und Statistik* 40.

Neubauer, J. 1931. Die Gossenschen Gesetze. *Zeitschrift für Nationalökonomie* 2, 733–53.

Pantaleoni, M. 1889. *Principii di economia pura*. Florence: G. Barbéra. Trans. by T.B. Bruce as *Pure Economics*, London: Macmillan, 1898.

Riedle, H. 1953. *Hermann Heinrich Gossen 1810–1858. Ein Wegbereiter der modernen ökonomischen Theorie*. Winterthur: Keller.

Walras, L. 1885. Un économiste inconnu: Hermann-Henri Gossen. *Journal des Economistes*. Reprinted in L. Walrus, *Etudes d'économie sociale*, Lausanne: Rouge, 1896.

How to cite this article

Niehans, Jürg. "Gossen, Hermann Heinrich." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000084> doi:10.1057/9780230226203.0660

The New Palgrave Dictionary of Economics Online

Gournay, Jacques Claude Marie Vincent, Marquis de (1712 – 1759)

Peter Groenewegen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cantillon, R.; Forbonnais, F. V. D.; free trade; Gournay, Marquis de; laissez faire; Morellet, Abbe de; Physiocracy; Trudaine, D.; Tucker, J.; Turgot, A. R. J.

Article

French economist, merchant and government official, Gournay was born at St. Malo in 1712. After a long career as merchant, spent largely in Cadiz (1729–44), his partner's death in 1746 permitted his retirement two years later from active trade and his entry into public life and more serious research into economics. Gournay has been traditionally associated with the propagation in France of free trade ideas such as deregulation of colonial trade, abolition of the guilds and of the system of government inspection of manufactures, aspects of his work illustrated by the important place generally assigned to him in the history of the phrase, *laissez faire, laissez passer* (Schelle, 1897, pp. 214–17). Turgot (1759, pp. 30–2) has noted, however, that his free trade position should be qualified and in addition that, unlike the Physiocrats, he accorded an important role in economic development to industry and trade as well as agriculture. He has therefore sometimes been described as the founder of a separate non-Physiocratic free trade school, whose members, among others, included Turgot, Morellet and Trudaine. Apart from *Observations sur l'agriculture, le commerce et les arts de Bretagne* (1757), only his notes accompanying the translation of Child (1754), now edited by Tsuda (1983), appear to have survived. His long friendship with Turgot exerted some influence on the latter's economics, partly because Turgot accompanied Gournay on his tours of inspection of industry between 1753 and 1756. Gournay's most important contribution to French economics seems to have been the encouragement he gave to the study of English economics literature. With Butel-Dumont he had himself translated Child and Culpeper (1754), he encouraged Forbonnais to abridge King's *The British Merchant*, Turgot to translate one of Tucker's pamphlets and, most importantly, may have been responsible for the publication of Cantillon's *Essay* in 1755 (Morellet, 1821, pp. 36–7). His death in 1759 provided the occasion for Turgot's eulogy on which much of the information about his life and work is based, though as Ashley (1900, p. 306)

warns, there are reasons for being hesitant in accepting Turgot's eulogy (1759) ‘as evidence of Gournay's opinions’.

Bibliography

Ashley, W.J. 1900. Gournay. In W.J. Ashley, *Surveys, Historic and Economic*. London: Longman & Co.

Child, J. 1754. *Traités sur le commerce et sur les avantages qui résultent de la reduction de l'intéret de l'argent par Josias Child … avec un petit traité contre l'usure par Thomas Culpeper*. Traduit de l'Anglois [by Gournay and Butel-Dumont], Amsterdam and Berlin.

Morellet, l'Abbé. 1821. *Mémoires de l'Abbé Morellet*. Paris: Librairie Française.

Schelle, G. 1897. *Vincent de Gournay*. Paris: Guillaumin.

Tsuda, T. 1983. *Josiah Child, Traités sur le commerce de Josiah Child, avec les remarques inédites de Vincent de Gournay*. Tokyo: Kinokuniya Company.

Turgot, A.R.J. 1759. In praise of Gournay. In *The Economics of A.R.J. Turgot*, ed. and trans. P.D. Groenewegen, The Hague: Nijhoff, 1977.

How to cite this article

Groenewegen, Peter. "Gournay, Jacques Claude Marie Vincent, Marquis de (1712–1759)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000085> doi:10.1057/9780230226203.0661

The New Palgrave Dictionary of Economics Online

government budget constraint

Eric M. Leeper and James M. Nason

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The government budget constraint is an accounting identity linking the monetary authority's choices of money growth or nominal interest rate and the fiscal authority's choices of spending, taxation, and borrowing at a point in time and across time. The intertemporal links create a rich set of possible outcomes from standard macro policy experiments. Taking the government budget constraint seriously can overturn some widely held beliefs about policy effects.

Keywords

Barro, R.; bond–money ratio; endowment economy; Euler equation; fiscal policy; fiscal theory of the price level; Fisher relation; government budget constraint; household budget constraint; inflation; Markov processes; monetary policy; money supply; Ricardian equivalence; Ricardo, D.

Article

The government budget constraint is an accounting identity linking the monetary authority's choices of money growth or nominal interest rate and the fiscal authority's choices of spending, taxation, and borrowing at a point in time. Whenever borrowing is the source of some fiscal financing, the government budget constraint also serves to link current monetary and fiscal choices to expected future monetary and fiscal policy variables. This intertemporal dimension creates a rich set of possible impacts of routine macro policy actions, as current or future policies can be expected to adjust to satisfy the government budget, along with other equilibrium conditions. Taking the government budget constraint seriously can overturn some widely held beliefs about policy effects.

The notion that current government policy has intertemporal implications goes back to Barro (1974), who revived ideas associated with Ricardo (1821). Traditional Keynesian models, in contrast, mostly ignored the impact of the government budget constraint on allocations and prices until the work of Christ (1967), 1968 see Sims (1998) for a review and extensions). Hansen, Roberds and Sargent (1991)

show that identification of the responses of allocations and prices to changes in the government budget constraint require specification of the economic primitives of preferences, technology and market structure.

The modern treatment of the government budget constraint begins with Sargent and Wallace (1981). They show that, when the primary fiscal surplus is fixed, an open-market sale of debt, and contraction of base money, produces higher future inflation. This stunning result arises because, with fiscal policy fixed, faster money supply growth is the only policy expected to balance future government budget constraints. A related but different mechanism by which the government budget constraint can restrict the equilibrium price level, namely, the ‘fiscal theory of the price level’, is developed by Leeper (1991), Sims (1994), Woodford (1995); 2001 and Cochrane (1999), among others. That theory demonstrates that, under certain assumptions on policy behaviour, debt-financed cuts in lump-sum taxes can stimulate aggregate demand, in apparent contradiction of Ricardian equivalence.

This article uses endowment and growth economies to study the restrictions the government budget constraint imposes on the intertemporal trade-offs between current and future monetary and fiscal policies. The endowment economy allows us to depict the policy trade-offs associated with a bond-financed tax cut, holding government spending fixed. We show that the effects of policy changes depend on current and expected future monetary and fiscal policies that are consistent with the government budget constraint at each date. Although we illustrate these points for a bond-financed tax cut, analogous results hold for an open-market operation. Implicit in the analyses is that the expected discounted value of real government debt has no value at the infinite horizon; that is, a transversality condition for government debt holds at the infinite horizon. This is a sufficient condition for an equilibrium to exist.

Model primitives

The models are variations of Sidrauski (1967) and share the following features: perfect foresight, a representative, infinitely lived household with utility defined over consumption, c_t, and real balances, M_t / P_t, $U(c_t, M_t / P_t) = u(c_t) + v(M_t / P_t)$, and nominal one-period government bonds, *B*, paying net nominal interest of i. The models also have in common two equilibrium conditions that stem from optimal household choices: a portfolio balance expression

$$\frac{v'(M_t / P_t)}{u'(c_t)} = \frac{i_t}{1 + i_t}, \tag{1}$$

and optimality of bond choices, a Fisher relation, represented by the Euler equation

$$1 = \beta(1 + i_t)\left[\frac{u'(c_{t+1})}{u'(c_t)}\frac{P_t}{P_{t+1}}\right],$$
(2)

where $0 < \beta < 1$ is the household's discount factor.
The structure of the government balance sheet, revenue sources, and expenditure process is also common across the models we examine. The government chooses sequences of $\{M_t, B_t, T_t, z_t\}$ to finance purchases of goods and services, g_t, and transfer payments, z_t, to satisfy the government budget constraint

$$g_t + z_t = T_t + \frac{M_t - M_{t-1}}{P_t} + \frac{B_t - (1 + i_{t-1})B_{t-1}}{P_t},$$
(3)

where T_t denotes total tax revenues. Government spending is specified as shares of output: $g_t = s_t^g Y_t$ and $z_t = s_t^z Y_t$. The government budget constraint (3) has the present value form

$$\frac{B_{t-1}}{P_{t-1}} = E_t\sum_{j=0}^{\infty}\prod_{l=0}^{j}\left(\frac{\pi_{t+l}}{1 + i_{t-1+l}}\right)\left[T_{t+j} - g_{t+j} - z_{t+j} + s_{t+j}\right],$$
(4)

where $s_t = (M_t - M_{t-1}) / P_t$ is seigniorage revenues. To arrive at (4), the infinite-horizon transversality condition for debt from the household's optimization problem has been imposed: $\lim_{q\to\infty} E_t\beta^q u'(c_{t+q})\prod_{j=0}^{q}\left(\frac{\pi_{t+j}}{1+i_{t+j+1}}\right)\frac{B_{t+q}}{P_{t+q}} = 0$. This is the relevant sufficient condition because it forces the household to expect that it cannot postpone consumption, hold government bonds for ever, and raise lifetime utility (see Becker and Boyd, 1997, for good economic intuition). It is important to note that in stochastic models the transversality condition need not hold always and everywhere along equilibrium paths, as it does in perfect foresight equilibria. Rather, it holds only in *expectation* (see Kamihigashi, 2005, for discussion and examples).

Endowment economy

It is useful to study an endowment economy because it draws out the role of the government budget constraint in macroeconomic analyses. The household budget constraint is

$$c_t + \frac{M_t + B_t}{P_t} \leq y_t + z_t + \frac{M_{t-1} + (1 + i_{t-1})B_{t-1}}{P_t}, \tag{5}$$

where *y* is the endowment of goods each period and we have set $T_t = 0$, for all *t*, so $z^t>0$ (<0) represents lump-sum transfers (taxes). Output and government purchases are constant, so $y_t = y$ and $g_t = s^g y$, which implies that in equilibrium consumption is a constant share of GDP, $c_t = c = (1 - s^g) y$. Thus, the equilibrium real interest rate equals a constant, $1 / \beta$, the Fisher relation reduces to $1 + i_t = \beta^{-1} \pi_{t+1}$ (where $\pi_{t+1} = P_{t+1} / P_t$), and money demand varies only with the nominal interest rate, $v'(M_t / P_t) = u'(c)[i_t / (1 + i_t)]$.

We focus on circumstances in which the economy is in a stationary equilibrium at dates $s > t$, but starts from a different equilibrium at date *t*. Denote money growth by $\rho_t = M_t / M_{t-1}$. Assume tax and monetary policies are fixed in the future stationary equilibrium: $s_s^z = s^z$ and $\rho_s = \rho$ for $s > t$; at date *t*, however, policies may be different: $s_t^z \neq s^z$ and $\rho_t \neq \rho$.

In the stationary equilibrium with constant real money balances, inflation depends only on money growth, $\pi = \rho$, which implies the Fisher relation is $1 + i_s = \beta^{-1} \rho_{s+1}$, $s \geq t$. Stationary real money balances become $M_s / P_s = h(\rho_{s+1})$, for dates $s \geq t$.

We derive two versions of the government budget constraint that describe the trade-offs among current and future monetary and fiscal policies that arise in equilibrium. By imposing equilibrium prices on the government budget constraint (3), we obtain

$$\frac{h(\rho)}{y}\left[1 - \frac{1}{\rho_t} + \frac{B_t}{M_t} - \frac{1 + i_{t-1}}{\rho_t}\frac{B_{t-1}}{M_{t-1}}\right] = s^g + s_t^z. \tag{6}$$

For given future expected policies, expression (6) reports the feasible trade-offs among current (date *t*) policies, when initial liabilities are $(M_{t-1}, (1 + i_{t-1})B_{t-1})$. On the assumption that future policy is anticipated (i.e., $1 + i = \beta^{-1}\rho$), the government budget constraint is

$$\frac{h(\rho)}{y}\left[1-\frac{1}{\rho}+\left(1-\frac{1}{\beta}\right)\frac{B}{M}\right]=s^{g}+s^{z}, \tag{7}$$

along the equilibrium path for dates $s > t$, given $B_t / M_t = B / M$. Note that the bond–money ratio is constant in the stationary equilibrium. Conditional on the state of government indebtedness, equation (7) describes the trade-offs among future policies that are consistent with equilibrium.

Policy analysis

In the policy experiments we consider, government purchases, s^g, are held fixed. The experiments take the form of an initial cut in taxes at date *t* (negative s_t^z becomes larger in absolute value), which is financed by sales of nominal bonds. We consider three alternative responses of current and future policies that satisfy (6) and (7). The analysis traces the effects of each specification of policy behaviour on the price level and inflation.

Policy 1

For policy experiment 1, suppose current and future money growth, (ρ_t, ρ), are held fixed. This policy, together with the money demand relation, (1), and Fisher relation, (2), peg the nominal interest rate at $1 + i = \rho / \beta$ and fix equilibrium real balances at $h(\rho)$. Neither the initial price level, P_t, nor the stationary inflation rate, π, changes. A reduction in taxes today is consistent with equilibrium if nominal debt expands to satisfy the government budget constraint (6) with fixed money growth. This raises B_t / M_t, which, by the government budget constraint (7), forces future taxes to rise sufficiently to service the new, higher level of government indebtedness. This mix of policies yields Ricardian equivalence: the timing of taxes and debt is irrelevant for equilibrium allocations and prices. The policies also imply monetary policy is independent of fiscal considerations, as the quantity theory of money maintains. Of course, as this exercise illustrates, the quantity theory requires specific fiscal behaviour.

Policy 2

In the second experiment, the central bank credibly pegs the nominal interest rate by fixing future money growth, ρ, and the fiscal authority credibly fixes future taxes. Can this be an equilibrium? With future policies fixed, the anticipated budget constraint (7) implies current policies cannot alter government indebtedness in the future, which is summarized by B / M. Since the expansion in nominal debt cannot be transformed into future higher real debt, P_t must rise in proportion to B_t. However, a pegged nominal interest rate fixes real money balances. The result is that the current money stock must expand in proportion to the increase in prices, which ensures B_t / M_t is unchanged in the date *t* budget constraint (6).

The central bank loses control of the current money stock and the price level in this experiment. Changes in these variables are governed by fiscal needs that are beyond the central bank's direct control. A pegged nominal rate subordinates current monetary policy to fiscal needs, but this is not ‘monetization of deficits’ in the usual sense of printing money to purchase newly issued government debt. Instead, the expansion in money is a passive adjustment of the money supply to clear the money market at the prevailing interest rate and price level. The monetary expansion is given by $dM_t = dB_t / (B_t / M_t)$, making clear that monetary accommodation varies inversely with the level of indebtedness. This exercise corresponds to the fiscal theory of the price level as described by Leeper (1991), Sims (1994), and Woodford (1995). The precise result relies on government debt being sold at par, as Cochrane (2001) observes. If government debt is sold at a discount, bond prices may absorb some of the adjustment to equilibrium, which pushes the price level effects into the future.

Policy 3

The third experiment has the central bank fix current money growth, ρ_t, while the fiscal authority continues to hold future taxes, s^T, constant. It remains feasible for current policy to imply more debt in the future because the anticipated increase in debt service forces future money growth and inflation to rise. The date *t* response is seen in a higher nominal interest rate and reduced real money balances driven by an increase in P_t to clear the money market, which follows from a fixed M_t. Beyond date *t*, debt service is financed by higher inflation and seigniorage – ‘inflation tax’ on nominal assets – revenues. Again, with future net-of-interest fiscal deficits held fixed at $s^g + s^T$, monetary policy is constrained by fiscal needs. In this case, the central bank loses control of future inflation. Sargent and Wallace (1981) employ these assumptions about policy in their classic ‘unpleasant monetarist arithmetic’ example.

A growth economy

A growth model with elastic capital supply, inelastic labour supply, and a distorting income tax extends the analysis by adding interesting intertemporal margins. The model consists of a representative household, a firm that produces the single consumption good, and a government (see , Gordon and Leeper (2006) for related analysis). Assume physical capital depreciates completely after one period. Output is allocated to consumption, capital, k_t, and government purchases of goods, with the technology $f(k_{t-1})$ generating output, Y_t, where $f(0) = 0$, $f'(k_{t-1}) > 0$, and $0 \geq f''(k_{t-1})$. Capital share's of production is denoted by σ. The economy is closed with the aggregate resource constraint

$$c_t + k_t + g_t = f(k_{t-1}). \tag{8}$$

A competitive firm rents capital at rate *r* from the household and pays taxes levied against sales of

goods, which determine the profit function, $D_t = (1 - \tau_t) y_t - (1 + r_t) k_{t-1}$. Profit maximization yields the after-tax factor price $1 + r_t = (1 - \tau_t) f'(k_{t-1})$.
The household supplies labour inelastically, owns the firm, and receives factor payments. Subject to the budget constraint

$$c_t + k_t + \frac{M_t + B_t}{P_t} \leq (1 + r_t) k_{t-1} + D_t + z_t + \frac{M_{t-1} + (1 + i_{t-1}) B_{t-1}}{P_t}, \tag{9}$$

the household maximizes the expected discounted value of its infinite horizon utility function, given P_t, i_t, τ_t, and the initial conditions $(k_{-1} > 0,\ M_{-1} + (1 + i_{-1}) B_{-1} > 0)$. Government behaviour is unchanged from the endowment economy, but tax revenue is $T_t = \tau_t (1 + r_t) k_{t-1}$.

Equilibrium

We recover an explicit characterization of the model's equilibrium with $u(c_t) = \ln(c_t)$ and $v(M_t / P_t) = \ln(M_t / P_t)$. After imposing transversality conditions for capital, debt and money, equate the supply and demand for capital to find the solution

$$k_t = \left(1 - \frac{1}{\eta_t}\right)(1 - s_t^g) f(k_{t-1}), \tag{10}$$

where $\eta_t \equiv E_t \Sigma_{i=0}^{\infty} (\sigma\beta)^i \Pi_{j=0}^{i-1} \left(\frac{1 - \tau_{t+j+1}}{1 - s_{t+j+1}^g} \right)$. Money market equilibrium sets money supply to money demand to yield

$$\frac{M_t}{P_t} = c_t \mu_t, \tag{11}$$

where $\mu_t \equiv E_t \Sigma_{i=0}^{\infty} \beta^i \Pi_{j=0}^{i-1} \frac{1}{\rho_{t+j+1}}$. Note that μ and η completely summarize what agents need to

know to form rational expectations. Since eqs. (8) and (10) imply a decision rule for consumption, equilibrium real balances can be expressed in terms of their opportunity cost, $1/\mu_t = i_t/(1+i_t)$, the transactions they help finance, $c_t + k_t$, and expected fiscal policies:

$$\frac{M_t}{P_t} = \frac{1}{\eta_t}\left(\frac{i_t}{1+i_t}\right)^{-1} c_t + k_t. \tag{12}$$

With $c + k$ serving as a scale variable, expression (12) is a conventional money demand function except for the dependence on expected fiscal policies. Expectations about future fiscal policies are essential to tie down the equilibrium. This is a key to the dynamics of the growth model and the impacts of fiscal policy on the current equilibrium.

Equilibrium requires that current and future policies satisfy the government's budget constraint and that agents' expectations of policy are consistent with equilibrium. This creates interactions between current and future policies. As before, we distill the analysis down to two periods – now and the future. Fix current and future government spending shares, $\{s_t^g, s_t^z\}$, for all *t*, and assume future money growth and tax rates are constant $\rho_s = \rho, \tau_s = \tau, s > t$. Current policies, however, may differ: $\rho_t \neq \rho, \tau_t \neq \tau$.

The government budget constraint can be expressed entirely in terms of current and expected policies. In period *t*, the constraint is

$$\left[\frac{\rho_t - 1}{\rho_t} + \frac{B_t}{M_t} - \frac{1+i_{t-1}}{\rho_t}\frac{B_{t-1}}{M_{t-1}}\right]\frac{\mu_t}{\eta_t} = \frac{s_t^g + s_t^z - \tau_t}{1 - s_t^g}. \tag{13}$$

Given policy expectations are embedded in μ_t/η_t and initial government indebtedness is summarized by $(1+i_{t-1})B_{t-1}/M_{t-1}$, expression (13) reports the equilibrium trade-offs among current policies. Equilibrium trade-offs between current and future policies are given by the state of government indebtedness. We use the budget constraint (13) to develop this idea for the growth model. Shift the timing of eq. (13) forward one period and assume future interest liabilities are correctly anticipated at *t* by substituting the expression for the equilibrium nominal return $1 + i_t$. Given the bond–money ratio is constant at $B/M = B_t/M_t$ in the stationary equilibrium, there can be no net additions to debt in the future. Dropping the time subscript for variables dated $t+1$ and imposing equilibrium yields

$$\frac{\mu}{\eta}\left[\left(1-\frac{1}{\beta}\right)\frac{B}{M}+\left(\frac{\rho-1}{\rho}\right)\right]=\frac{s^{g}+s^{I}-\tau}{1-s^{g}}. \tag{14}$$

Equation (14) describes the trade-offs among future policies that are consistent with fixed μ_t / η_t being an equilibrium. The trade-offs represented by eqs. (13) and (14) tie together current policies and expectations of future policies. Any change in policy at date *t* that requires a change in μ_t / η_t must be accompanied by a change in policy in the future that is consistent with revised values of μ_t / η_t, conditional on the level of government debt B / M.

Policy analysis

As for the endowment economy, we study the current and future responses of fiscal and monetary policy to a date *t* debt finance tax cut in the analysis that follows.

Policy 1

Hold current and future money growth fixed at (ρ_t, ρ). This policy pegs the nominal interest rate by fixing μ_t but it does not fix real money balances unless η_t is also constant. Since new debt issued to finance the tax reduction raises B_t / M_t, a higher level of debt is carried into the future. To clear the government budget constraint in the future, budget constraint (14) implies future taxes must rise. Higher taxes reduce the return on capital (a lower η) and induce substitution from real to nominal assets, which includes money. Equilibrium in the money market requires the current price level to fall. The source of the non-Keynesian reduction in inflation is the link between current policy (that is, the fiscal expansion) and the expectation that future policy will expand government debt.

Policy 2

Fix both future money growth and future taxes at (ρ, τ). By assumption, all future policies are constant in the face of the current tax cut. Current policies must adjust to ensure the real value of debt in the future is unchanged, as was true when this policy was applied to the endowment economy. The real value of debt remains unchanged because the current money stock rises by the amount that the current budget constraint (13) dictates is needed to maintain the pre-tax cut level of B_t / M_t. The monetary expansion necessary to maintain equilibrium is sufficient to produce additional future seigniorage (that is, the level of the money supply rises). Since the fixed rate of money growth is just enough to pay for the increased debt service and with equilibrium real money balances fixed by constant future policies, (12) predicts the current price level rises in proportion to the increase in M_t. Gordon and Leeper (2006) label this ‘the canonical fiscal theory exercise’.

The implications of the fiscal theory contrast with the tax cut of policy 1. The bond-financed tax cut is pure fiscal policy in the sense that it is *independent* of the path of the money stock. It also reduces

nominal spending and the price level. An essential aspect of the fiscal theory is that the current money stock adjusts passively to clear the money market, raising nominal demand and the price level. If the policy authorities peg the nominal interest rate and fix future taxes without reference to the rest of the economy, higher prices are inevitable consequences of a tax cut. This is an illustration of the fiscal theory.

Policy 3

The fiscal authority holds future taxes constant and the central bank fixes current money growth. If future money growth rises sufficiently to generate the seigniorage revenue to service the new debt, an expansion in current debt can be carried into the future. Expected inflation increases, which lowers the expected return on money (that is, μ falls), decreases money demand, raises the price level, and contributes to higher future inflation. The change in future money growth depends, of course, on future B/M, which drives the change in debt service.

Concluding remarks

The equilibria described in this article can easily be couched in terms of arbitrary sequences of policy variables. It has become increasingly popular, however, to endow policy authorities with simple rules that make the policy instrument a time-invariant function of only a few variables that are not directly related to the actions of other policy institutions (that is, the interest rate–monetary policy rules studied in Taylor, 1999). Although this approach has the advantages of being interpretable and tractable, it runs the risk of oversimplifying policy behaviour. For example, it is difficult to square simple time-invariant policy functions with the observation that policy regimes can, and do, change, sometimes because of the interactions of different policymakers.

A natural extension of simple rules allows feedback parameters to take on finitely many values (‘regimes’) whose evolution is governed by a Markov process. Relative to simple rules, this extension produces a far richer set of expectations of future policy variables, a generalization that can overturn some of the principles guiding macro policy research that have been obtained from simple rules (see Davig, Leeper and Chung, 2004, and Davig and Leeper, 2005).

Markov switching of policy rules has also generalized the test of the long-run sustainability of fiscal policy proposed by Hamilton and Flavin (1986). Davig (2005) finds expansionary and contractionary regimes in US government debt that nonetheless yield a stochastic process for discounted debt with an unconditional expected value equal to zero in the long run.

Theoretical work that takes seriously the restrictions imposed by the government budget constraint has established some important and surprising results. In light of these theoretical findings, it is remarkable how little applied work on monetary and fiscal policy treats the government budget constraint with equal seriousness. This is an open area of research.

See Also

- fiscal theory of the price level

- hyperinflation
- inflation dynamics
- monetary and fiscal policy overview
- optimal fiscal and monetary policy (with commitment)
- optimal fiscal and monetary policy (without commitment)

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1095–117.

Becker, R. and Boyd, J., III 1997. *Capital Theory, Equilibrium Analysis, and Recursive Utility*. Malden, MA: Blackwell.

Christ, C. 1967. A short-run aggregate-demand model of the interdependence and effects of monetary and fiscal policies with Keynesian and classical interest elasticities. *American Economic Review* 57, 434–43.

Christ, C. 1968. A simple macroeconomic model with a government budget restraint. *Journal of Political Economy* 76, 53–67.

Cochrane, J. 1999. A frictionless view of U.S. inflation. In *NBER Macroeconomics Annual 1998*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Cochrane, J. 2001. Long-term debt and optimal policy in the fiscal theory of the price level. *Econometrica* 69, 69–116.

Davig, T. 2005. Periodically expanding discounted debt: a threat to fiscal policy sustainability? *Journal of Applied Econometrics* 20, 829–40.

Davig, T. and Leeper, E. 2005. Fluctuating macro policies and the fiscal theory. Working Paper No. 11212. Cambridge, MA: NBER.

Davig, T., Leeper, E. and Chung, H. 2004. *Monetary and fiscal policy switching. Working Paper No. 10362*. Cambridge, MA: NBER.

Gordon, D. and Leeper, E. 2006. The price level, the quantity theory of money, and the fiscal theory of the price level. *Scottish Journal of Political Economy* 53, 4–27.

Hamilton, J. and Flavin, M. 1986. On the limitations of government borrowing: a framework for empirical testing. *American Economic Review* 76, 808–19.

Hansen, L., Roberds, W. and Sargent, T. 1991. Time series implications of present value budget balance and of martingale models of consumption and taxes. In *Rational Expectations Econometrics*, ed. L. Hansen and T. Sargent. Boulder, CO: Westview Press.

Kamihigashi, T. 2005. Necessity of the transversality condition for stochastic models with bounded or CRRA utility. *Journal of Economic Dynamics and Control* 29, 1313–29.

Leeper, E. 1991. Equilibria under ‘active’ and ‘passive’ monetary and fiscal policies. *Journal of Monetary Economics* 27, 129–47.

Ricardo, D. 1821. *On the Principles of Political Economy and Taxation*. 3rd edn. London: John Murray.

Sargent, T. and Wallace, N. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5, 1–17.

Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review Papers and Proceedings* 57, 534–44.

Sims, C. 1994. A simple model for study of the determination of the price level and the interaction of monetary and fiscal policy. *Economic Theory* 4, 381–99.

Sims, C. 1998. Econometric implications of the government budget constraint. *Journal of Econometrics* 83, 9–19.

Taylor, J. 1999. *Monetary Policy Rules*. Chicago: University of Chicago Press.

Woodford, M. 1995. Price-level determinacy without control of a monetary aggregate. In *Carnegie-Rochester Conference Series on Public Policy*, ed. B. McCallum and C. Plosser, 43, 1–46.

Woodford, M. 2001. Fiscal requirements for price stability. *Journal of Money, Credit, and Banking* 33, 669–728.

How to cite this article

Leeper, Eric M. and James M. Nason. "government budget constraint." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000161> doi:10.1057/9780230226203.0662

The New Palgrave Dictionary of Economics Online

Graham, Frank Dunstone (1890 – 1949)

Randall Hinshaw
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

ceiling velocity; commodity-reserve monetary standard; flexible exchange rates; German hyperinflation; Graham, F. D.; international trade theory; Mill, J. S.; protection; tariffs; Taussig, F. W.; terms of trade; velocity of circulation

Article

Graham was born in Halifax, Nova Scotia, and died in Princeton, New Jersey. He is known mainly for his work in the theory of international trade, and especially for his attack on classical and neoclassical trade theory. He received his doctorate from Harvard, where he came under the influence of Taussig. After teaching at Rutgers and Dartmouth, he joined the Princeton faculty in 1921, becoming a full professor in 1930. In addition to undergraduate teaching, he taught the Princeton graduate courses in international trade and in monetary theory.

In a path-breaking article (1923b), Graham argued that J.S. Mill, by using a two-country, two-commodity model, had reached erroneous conclusions concerning the effect of changes in international demand on the commodity terms of trade. Mill had reasoned that, within the limits set by comparative cost (limits which he – but not Graham – regarded as improbable cases), an increase in a country's demand for imports would worsen the country's terms of trade. Retaining Mill's assumptions of free trade, costless transportation and constant cost per unit of output, Graham concluded that, when a given commodity is produced by more than one country, the cost structures of the affected countries are locked together and that, therefore, changes in international demand do not affect the equilibrium terms of trade so long as the same commodities continue to be produced by the same countries; instead, within possibly wide limits, international adjustment takes place through shifts in output, the limits occurring when commodities disappear from, or are added to, national production schedules.

To illustrate these points, Graham devised a multi-country, multi-commodity model, with all variables expressed in real terms. Operating with assumed national opportunity-cost ratios, national productive capacities and national demand functions, he was able to derive, by a trial-and-error process using simple arithmetic, an equilibrium solution specifying the commodity terms of trade and each country's

consumption, production (if any) and exports or imports of each commodity. In his final work, *The Theory of International Values* (1949), he developed these ideas at length, using illustrations with as many as ten countries and ten commodities. Because of the assumption of costless transportation, domestic (non-traded) goods do not appear in the trade model, but Graham examined their role in international adjustment in his earliest article (1922) and in his 1949 treatise.

Although the Graham model, which assumes full employment, can be used to demonstrate that national and world real output are maximized under free trade, Graham was not a doctrinaire free trader. In an early article (1923a), he made a case for permanent protection for decreasing-cost industries. The article was attacked on various grounds by Knight (1924) and others, but Graham retained the argument in his book, *Protective Tariffs* (1934), which, while critical of most arguments for tariffs, included a chapter on ‘Rational Protection’.

In the field of money, Graham's major work was his treatise (1930) on the German hyperinflation after the First World War. Perhaps his most significant conclusion was the concept of ‘ceiling velocity’. He found that, in the German case, monetary velocity reached an upper limit which was about 25 times the pre-war normal; thereafter, the German price level rose at approximately the same rate as the German money supply.

Graham had a passionate interest in economic policy. He was an early advocate of flexible exchange rates (on a managed basis), and during the Great Depression he devised various plans to promote recovery. Later, he advocated a commodity-reserve monetary standard as a means of achieving price-level stability and full employment.

An iconoclast with a caustic wit, Graham was an unusually stimulating teacher and had a profound influence on his students, two of whom – T.M. Whitin and L.W. McKenzie – extended his work on the trade model. In a 1953 article which illustrated the model geometrically, Whitin concluded that Graham's work ‘anticipated linear programming models by many years’, and McKenzie, in a powerful 1954 article employing a theorem from topology, demonstrated what Graham firmly believed but was never able to prove: that his trade model yields an equilibrium for any continuous demand functions and that this solution is unique for the demand functions which Graham actually used.

Selected works

1922. International trade under depreciated paper: the United States, 1862–79. *Quarterly Journal of Economics* 36, February, 220–73.

1923a. Some aspects of protection further considered. *Quarterly Journal of Economics* 37, February, 199–227.

1923b. The theory of international values re-examined. *Quarterly Journal of Economics* 38, November, 54–86.

1930. *Exchange, Prices, and Production in Hyperinflation: Germany, 1920–1923*. Princeton: Princeton University Press.

1932a. The theory of international values. *Quarterly Journal of Economics* 46, August, 581–616.

1932b. *The Abolition of Unemployment*. Princeton: Princeton University Press.

1934. *Protective Tariffs*. New York: Harper & Bros. Reprint edn, Princeton: Princeton University Press, 1942.

1942. *Social Goals and Economic Institutions*. Princeton: Princeton University Press.

1949. *The Theory of International Values*. Princeton: Princeton University Press.

Bibliography

Knight, F.H. 1924. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38 (August), 582–606.

McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22, 147–61.

Metzler, L.A. 1950. Graham's theory of international values. *American Economic Review* 40, 301–22.

Whitin, T.M. 1953. Classical theory, Graham's theory, and linear programming in international trade. *Quarterly Journal of Economics* 67(November), 520–44.

Whittlesey, C.R. 1952. Frank Dunstone Graham, 1890–1949. *Economic Journal* 62, 440–45.

How to cite this article

Hinshaw, Randall. "Graham, Frank Dunstone (1890–1949)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000094> doi:10.1057/9780230226203.0663

The New Palgrave Dictionary of Economics Online

Granger, Clive W. J.

Timo Teräsvirta
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cointegration; conditional variance models; degree of integration; econometrics; forecasting; fractional integration; GARCH models; Granger representation theorem; Granger, C.W. J.; Granger-causality; Haavelmo, T.; long memory models; maximum likelihood; nonlinear time series; non-stationary variables; reduced rank regression; spectral analysis; spurious regressions; time series analysis; vector autoregressions

Article

Granger, Clive William John, born 4 September 1934, Swansea, Wales. British citizen, knighted in the 2005 New Year's Honours. Emeritus Professor of Economics, University of California, San Diego. Degrees: BA, University of Nottingham 1955, Ph.D., University of Nottingham 1959. Career: Lecturer, then Professor of Applied Statistics and Econometrics, University of Nottingham, 1956–73; Professor of Economics, University of California, San Diego, 1974–2003. Honours and awards: Fellow, Econometric Society 1972; Fellow, American Academy of Arts and Sciences 1994; Fellow, International Institute of Forecasters 1996; Foreign member, Finnish Society of Sciences and Letters 1997; Corresponding Fellow, British Academy 2002; Distinguished Fellow of American Economic Association 2002; Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel 2003 ‘for methods of analyzing economic time series with common trends (cointegration)’. Honorary degrees: University of Nottingham, Universidad Carlos III de Madrid, Stockholm School of Economics, University of Loughborough, Aarhus University, Aristotle University.

Clive Granger is one of the best-known time-series econometricians of our time. He has contributed to many areas in econometrics. They include the analysis of non-stationary time series, causal relations between economic variables, long memory, nonlinearity, forecasting economic time series, modelling stock prices and volatility, and price formation.

The most important concept that Granger has introduced to econometrics is *cointegration*. It can be viewed as an extension to non-stationary time series of Nobel Laureate Trygve Haavelmo's formulation of an economy as a system of simultaneous stochastic relationships that laid the foundations to modern

time-series econometrics. The origins of the concept may be traced to a paper in which Granger and his associate Paul Newbold consider regressing a *random walk* series on another independent random walk series (Granger and Newbold, 1974). They pointed out that the classical t-test in such a regression may often suggest a statistically significant relationship between variables although none exists. This result suggested that many relationships found between non-stationary economic variables in static econometric models of the time could in fact have been spurious. It called for a more careful analysis of non-stationary economic time series, and Granger and Newbold also strongly emphasized the importance of dynamic models instead of static ones that did not take the dynamic properties of economic variables into account.

A solution to this spurious regression problem was to model the relationships between economic variables in first differences. This, however, created a problem because these relationships were generally expressed in levels, not differences. Granger's solution to this problem (Granger, 1981) may be illustrated by the following regression equation:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \tag{1}$$

where y_t is the dependent variable, x_t the single exogenous regressor, and $\{\varepsilon_t\}$ a white-noise, mean-zero sequence. Granger used the concept of *degree of integration* of a variable. If variable z_t can be made approximately stationary by differencing it *d* times, it is called integrated of order *d*, or $I^{(d)}$. Weakly stationary random variables are thus I(0). Many macroeconomic variables can be regarded as I(1) variables: if $z_t \sim$ I(1), then $\Delta z_t \sim$ I(0). Assume now that both $x_t \sim$ I(1) and $y_t \sim$ I(1) in eq. (1). Then generally the linear combination $y_t - \beta x_t \sim$ I(1) as well. There is, however, one important exception. Many macroeconomic variables can be regarded as I(1) variables: if $z_t \sim$ I(1), then $\Delta z_t \sim$ I(0). It has to do with the fact that for an equation such as (1) to be meaningful it has to be *balanced*, a concept Granger employed in his work. An equation is balanced when its right-hand and left-hand sides are of the same order of integration. Rewrite (1) as

$$y_t - \beta x_t = \alpha + \varepsilon_t.$$

If $\varepsilon_t \sim$ I(0), then $y_t - \beta x_t \sim$ I(0), that is, the linear combination $y_t - \beta x_t$ has the same statistical properties as an I(0) variable. There exists only one such combination so that coefficient β is unique. In this special case, variables x_t and y_t are called *cointegrated*. This notion generalizes to more than two variables.

The importance of cointegration in the modelling of non-stationary economic series becomes clear in the Granger representation theorem that was first formulated in Granger and Weiss (1983). Consider the following bivariate vector autoregressive (VAR) model of order *p*:

$$x_t = \sum_{j=1}^{p} \gamma_{1j} x_{t-j} + \sum_{j=1}^{p} \delta_{1j} y_{t-j} + \varepsilon_{1t}$$

$$y_t = \sum_{j=1}^{p} \gamma_{2j} x_{t-j} + \sum_{j=1}^{p} \delta_{2j} y_{t-j} + \varepsilon_{2t},$$

where x_t and y_t are I(1) and cointegrated, and ε_{1t} and ε_{2t} are white noise. The Granger representation theorem states that in this case the system can be written as:

$$\Delta x_t = \alpha_1 (y_{t-1} - \beta x_{t-1}) + \sum_{j=1}^{p-1} \gamma_{1j}^{*} \Delta x_{t-j} + \sum_{j=1}^{p-1} \delta_{1j}^{*} \Delta y_{t-j} + \varepsilon_{1t}$$

$$\Delta y_t = \alpha_2 (y_{t-1} - \beta x_{t-1}) + \sum_{j=1}^{p-1} \gamma_{2j}^{*} \Delta x_{t-j} + \sum_{j=1}^{p-1} \delta_{2j}^{*} \Delta y_{t-j} + \varepsilon_{2t},$$

(2)

where at least one of parameters α_1 and α_2 deviates from zero. Both equations of the system are balanced since $y_{t-1} - \beta x_{t-1} \sim$ I(0). System (2) is now in *error-correction form* where $y_t - \beta x_t = 0$ defines a dynamic equilibrium relationship between the two economic variables, *y* and *x*. While the system consists of two equations, it only has a single equilibrium relationship. More generally, if the system has *n* variables, the number of these *cointegrating relationships* is less than *n*. System (2) in disequilibrium but has a built-in tendency to adjust itself towards the moving equilibrium. The coefficients α_1 and β_2 represent the relative strength of this adjustment at any given time. In applications, the equilibrium or long-run relationship represents an economic theory proposition,

whereas the remaining variables and parameters describe the short-term dynamic behaviour of the system.

It may be mentioned that linear combinations of non-stationary variables had appeared in dynamic econometric equations prior to Granger's work. Phillips (1957), who coined the term ‘error correction’, and Sargan (1964) had employed them, but they did not, however, consider the statistical implications of introducing such components into their models.

A first test for cointegration appeared in Granger and Weiss (1983). The idea is best seen from eq. (1). Variables x_t and y_t are cointegrated if “ $\varepsilon_t \sim I(0)$. Estimate parameter β by ordinary least squares and apply a standard unit root test to the residuals. A rejection of the null hypothesis suggests cointegration.

A method for estimating parameters of systems with cointegrated variables was still needed to make the concept applicable. Granger, working jointly with Robert Engle, developed a two-stage estimation method for VAR models with cointegration (Engle and Granger, 1987). Consider the following *n*-dimensional VAR model of order *p*:

$$\Delta \mathbf{x}_t = \alpha\beta' \mathbf{x}_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta \mathbf{x}_{t-j} + \varepsilon_t (t = 1, \ldots, T) \tag{3}$$

where $\mathbf{x}_t$ is an $n\times 1$ vector of I(1) variables, $\alpha\beta'$ is an $n\times n$ matrix such that the $n\times r$ matrices α and β have rank r, $\Gamma_j, j = 1, \ldots, p-1,$ are $n\times n$ parameter matrices, and ε_t is an $n\times 1$ vector of white noise with a positive definite covariance matrix. If $0 < r < n$, the variables in $\mathbf{x}_t$ are cointegrated with r cointegrating relationships $\beta' \mathbf{x}_t$. If the variables in $\mathbf{x}_t$ are cointegrated, the parameters of (3) can be estimated in two stages. First estimate β or, more precisely, the cointegrating space (β up to a multiplicative constant) using a form of least squares. Then, holding that estimate fixed, estimate the remaining parameters by maximum likelihood. The estimators of α and $\Gamma_j, j = 1, \ldots, p-1,$ are consistent and asymptotically normal. This solution is based on the fact that the least squares estimator of β is superconsistent: its rate of convergence is faster than that of the estimators of the other parameters. Engle and Granger (1987) also contains a rigorous proof of the Granger representation theorem.

This paper by Engle and Granger is one of the most cited papers in time series econometrics (200 citations annually since its appearance), and it was followed by a flood of applications. Cointegration strongly contributed to the popularity of VAR models suggested by Sims (1980) as a convenient tool for modelling economic variables without strong assumptions originating from economic theory. The ultimate refinement of the statistical theory of cointegrated variables was provided by Søren Johansen (see Johansen, 1995, for a summary) who derived the maximum likelihood estimator of β or, more precisely, the space spanned by the r cointegrating vectors in (3), using *reduced rank regression*. He also considered tests for determining the cointegration rank r. It may be mentioned that Granger originally (Granger, 1981) defined cointegration using tools from *spectral analysis*. In fact, in 1964 he wrote an

early book on the topic jointly with Michio Hatanaka. The book that became a Citation Classic was based on the work carried out when Granger was visiting Princeton University on a Harkness fellowship in the early 1960s.

Granger also undertook very influential research concerning causal relationships between variables. He moved away from deterministic causality (‘if A then B’) to stochastic one in which time plays a decisive role. Granger's causality definition (Granger, 1969) is based on the prediction accuracy of a stationary variable *y*. If *y* can be predicted more accurately with a set of other variables than without them, then these variables are said to cause. This is an operational definition that makes it possible to test the null hypothesis of no causality between economic variables with statistical methods. This may be done in either a single-equation or a system framework. For this reason this concept of causality, nowadays called Granger-causality, has become very popular in applied work. Most of the available tests, however, test for in-sample predictability, whereas Granger has always emphasized the fact that his definition pertains to out-of-sample forecasting and the (non)existence of causality should be tested accordingly. Granger-causality has become an important tool in economic research and policymaking and is also being used in other areas than economics. For more information, see Hendry and Mizon (1999). Granger (1986) established a link between causality and cointegration. If *y* and *x* are cointegrated then there is Granger-causality at least in one direction between these variables.

Granger has also been instrumental in starting the econometric research on processes with *long memory*. They have the property that their autocorrelations as a function of the lag length decay at a slower rate than autocorrelations of a linear autoregressive–moving average process in which the decay rate is exponential. Granger and Joyeux (1980) defined a new concept, *fractional integration*, and showed how fractionally integrated processes, stationary or non-stationary, can have long memory. A stochastic variable y_t is fractionally integrated series of order *d*, I(*d*), where *d* need not be an integer, if x_t in

$$x_t = (1 - L)^d y_t$$

where *L* is the lag operator: $Lyt=y_{t-1}$, is an I(0) variable. Choosing d=1 yields the standard case where y_t ~ I(1). Time series models based on fractional integration have since become popular in econometrics, and long memory in volatility has received plenty of attention in financial econometrics.

Granger is one of the first econometricians interested in *nonlinear time series* and wrote (1978, with Allan Andersen) a book on bilinear time series models. The bilinear model has not found much application in economics, but the book has stimulated more research in this area. Granger has since expanded his interests in nonlinear models, among other things, by generalizing cointegration, originally a linear concept, into nonlinear cointegration. He has also written a book on nonlinear econometric modelling (1993, with Timo Teräsvirta).

Economic forecasting has been one of Granger's main interests throughout his career. He observed (Bates and Granger, 1969) that combining forecasts from different models often improves the forecast accuracy compared with forecasts from individual models, and proposed forecast weighting schemes for this purpose. This work has prompted a large literature that is still growing. He wrote a book (1970, with

Oskar Morgenstern) on forecasting stock markets before that became a topic of broad interest, and a classic textbook with Paul Newbold on economic forecasting (Granger and Newbold, 1977). Another forecasting topic that has attracted Granger's interest is the evaluation of forecasts and the role of the forecaster's cost function in both parameter estimation and model evaluation, and he has made important contributions in this area.

One may argue that Granger's research is principally focused on conditional mean models, but he has also contributed to the analysis of conditional variance models. He has extended the standard model of *generalized autoregressive heteroskedasticity* (GARCH) model into the power GARCH model (Ding, Granger and Engle, 1993), intended to improve the modelling of volatility in financial time series such as return series of sufficiently high (intra-daily, daily, weekly) frequency. He has also indicated the potential for statistical modelling of the decomposition of stock return series into a sign process with little or no autocorrelation and an absolute return one with strong dependence structure. He has considered forecasting volatility of financial return series, a topic that is of great importance to investors who consider volatility to be a measure of risk (Poon and Granger, 2003).

A representative collection of Granger's scientific papers can be found in Ghysels, Swanson and Watson (2001).

See Also

- cointegration
- forecasting
- Granger–Sims causality
- spurious regressions
- vector autoregressions

Selected works

1964. (With M. Hatanaka.) *Spectral Analysis of Economic Time Series*. Princeton: Princeton University Press. Also in French.

1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 423–38.

1969. (With J. M. Bates.) The combination of forecasts. *Operations Research Quarterly* 20, 451–68.

1970. (With O. Morgenstern.) *Predictability of Stock Market Prices*. Lexington. MA: D.C. Heath.

1974. (With P. Newbold.) Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.

1977. (With P. Newbold.) *Forecasting Economic Time Series*. New York: Academic Press. 2nd edn, 1986.

1978. (With A.P. Andersen.) *Introduction to Bilinear Time Series Models*. Göttingen: Vandenhoeck & Ruprecht.

1980. (With R. Joyeux.) An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 15–30.

1981. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–30.

1983. (With A.A. Weiss.) Time series analysis of error-correction models. In *Studies in Econometrics, Time Series And Multivariate Statistics in Honor of T.W. Anderson*, ed. S. Karlin, T. Amemiya and L.A. Goodman. San Diego: Academic Press.

1986. Developments in the study of cointegrated variables. *Oxford Bulletin of Economics and Statistics* 48, 213–28.

1987. (With R. F. Engle.) Cointegration and error correction: representation, estimation and testing. *Econometrica* 55, 251–76.

1993. (With T. Teräsvirta.) *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

1993. (With Z. Ding and R.F. Engle.) A long memory property of stock market returns and a new model. *Journal of Empirical Finance* 1, 83–106.

2003. (With S.-H. Poon.) Forecasting volatility in financial markets: a review. *Journal of Economic Literature* 41, 478–539.

Bibliography

Ghysels, E., Swanson, N.R. and Watson, M.W., eds. 2001. *Essays in Econometrics. Collected papers of Clive W.J. Granger*, vols 1 and 2. Cambridge: Cambridge University Press.

Hendry, D.F. 2004. The Nobel Memorial Prize for Clive W.J. Granger. *Scandinavian Journal of Economics* 106, 187–213.

Hendry, D.F. and Mizon, G.E. 1999. The pervasiveness of Granger causality in econometrics. In *Cointegration, Causality and Forecasting. Festschrift in Honour of Clive W.J. Granger*, ed. R.F. Engle and H. White. Oxford: Oxford University Press.

Johansen, S. 1995. *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Phillips, A.W. 1957. Stabilization policy and the time forms of lagged responses. *Economic Journal* 67, 265–77.

Phillips, P.C.B. 1997. The ET interview: Professor Clive Granger. *Econometric Theory* 13, 253–303.

Sargan, J.D. 1964. Wages and prices in the United Kingdom a study in econometric methodology. In *Econometric Analysis for National Economic Planning*, ed. P.E. Hart, G. Mills and J.K. Whitaker. London: Butterworths.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Teräsvirta, T. 1995. Professor Clive W.J. Granger: an interview for the *International Journal of Forecasting*. *International Journal of Forecasting* 11, 585–90.

How to cite this article

Teräsvirta, Timo. "Granger, Clive W. J.." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000177> doi:10.1057/9780230226203.0664

The New Palgrave Dictionary of Economics Online

Granger – Sims causality

G.M. Kuersteiner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The concept of Granger–Sims causality is discussed in its historical context. There follows a review of the subsequent literature that explored conditions under which the definitions of Granger and Sims are equivalent. The relationship to the potential outcomes framework is explored in light of recent developments in the literature.

Keywords

block recursive structure; causality in economics and econometrics; conditional independence; conditional probability; covariance stationary processes; equivalence relationships; Granger non-causality; Granger, C.; Granger–Sims causality; Hume, D.; impulse response analysis; Mill, J. S.; monetary policy rules; observational studies; potential outcomes; prediction error variance; Rubin causal model; Simon, H.; Sims non-causality; structural innovations; structural vector autoregressions; white noise

Article

Granger–Sims causality is based on the fundamental axiom that ‘the past and present may cause the future, but the future cannot cause the past’ (Granger, 1980, p. 330). A variable x then is said to cause a variable y if at time t the variable x_t helps to predict the variable y_{t+1}. While predictability in itself is merely a statement about stochastic dependence, it is precisely the axiomatic imposition of a temporal ordering that allows us to interpret such dependence as a causal connection. The reason is that correlation is a symmetric concept with no indication of a direction of influence, while ‘the arrow of time imposes the structure necessary’ (Granger, 1980, p. 349) to interpret correlations in a causal way.

1 Definitions

A more precise definition of Granger causality can be given as follows. Assume that all relevant current information is measured in a vector $X_t = (y_t, x_t, z_t)$ which is observed at equally spaced discrete points in time *t*. (The assumption of discrete time can be relaxed but is maintained here for expositional convenience. Almost all empirical work is in the context of discrete time models.) Denote by Y_s^t all information contained in $y_j,\ j = \{s, \ldots, t\}$ with equivalent definitions for X_s^t and Z_s^t. (In more formal terms, the expression ‘information contained in’ can be replaced by ‘sigma field generated by’; see Florens and Mouchart, 1982, for precise definitions.) Then x_t does not Granger cause y_{t+1} if

$$P\left(y_{t+1} \in A \middle| Y_{-\infty}^t, X_{-\infty}^t, Z_{-\infty}^t\right) = P\left(y_{t+1} \in A \middle| Y_{-\infty}^t, Z_{-\infty}^t\right)$$
(CI)

for all *t* and for any set *A* for which the conditional probabilities are well defined. It is worth noting that no assumptions of stationarity are needed for this definition. It is common to use the shorthand notation $y_{t+1} \perp X_{-\infty}^t | Y_{-\infty}^t, Z_{-\infty}^t, \forall t$, which states that y_{t+1} and $X_{-\infty}^t$ are independent conditional on $Y_{-\infty}^t, Z_{-\infty}^t$. This form of the definition, which imposes a conditional independence restriction on the joint distribution of the process χ_t (see Dawid, 1979, for a formal definition and alternative equivalent representations), is similar to the definition used by Granger (1980). It is more general than the original formulations of Granger (1963; 1969), which are based on prediction error variances: Let $E(y_{t+1}|X_{-\infty}^t)$ denote the optimal predictor of y_{t+1} based on $X_{-\infty}^t$, $\varepsilon_{t+1} = y_{t+1} - E(y_{t+1}|X_{-\infty}^t)$ is the prediction error and $\sigma^2(y_{t+1}|X_{-\infty}^t)$ the variance of ε_{t+1}. Then, according to Granger (1969), x_t does not cause y_{t+1} if

$$\sigma^2\left(y_{t+1} \middle| Y_{-\infty}^t, X_{-\infty}^t, Z_{-\infty}^t\right) = \sigma^2\left(y_{t+1} \middle| Y_{-\infty}^t, Z_{-\infty}^t\right).$$
(PV)

The conditional independence definition (CI) has the advantage that it does not depend on a particular risk function and is easier to relate to other definitions of causality in stochastic environments such as Suppes (1970) and Rubin (1974). The formulation based on conditional independence was later used and refined in theoretical work by Chamberlain (1982), Florens and Mouchart (1982; 1985), Bouissou, Laffont and Vuong (1986) and Holland (1986). The advantage of the prediction error variance definition, (PV) which goes back to Wiener (1956), on the other hand, is that it is easier to implement in

statistical tests and has consequently received considerable attention in applied work.
In an influential paper, Sims (1972) showed in the context of covariance stationary processes and restricted to linear predictors that in the case of a bivariate system $X_t' = (x_t, y_t)$ the definitions of Granger (1963; 1969) are equivalent to parameter restrictions of the moving average or distributed lag representations of χ_t. When χ_t is covariance stationary it can be represented as

$$x_t = \sum_{j=0}^{\infty} a_j u_{t-j} + \sum_{j=0}^{\infty} b_j v_{t-j}$$

$$y_t = \sum_{j=0}^{\infty} c_j u_{t-j} + \sum_{j=0}^{\infty} d_j v_{t-j}$$

where a_j, b_j, c_j and d_j are constants and u_t and v_t are mutually uncorrelated white noise processes. Sims (1972) shows that the condition 'x_t does not Granger cause y^t+1' is equivalent to c_j or d_j being chosen identically zero for all j. Furthermore, if χ_t has an autoregressive representation, then $x_t = \sum_{j=0}^{\infty} \pi_j y_{t-j} + \varepsilon_t$ where the π_j are parameters and ε_t is an unobservable innovation. Strict exogeneity of y_t in the context of this model is defined as the condition that y_t and ε_s are independent for all values of t and s. Sims then shows that the residuals from a projection of x_t onto y_{t-j}, $j \geq 0$ are uncorrelated with all past and future y_t if an only if x_t does not Granger cause y_{t+1}. In other words, it follows that the condition 'x_t does not Granger cause y_{t+1}' is equivalent to the condition that y_t is strictly exogenous. The relationship between Granger non-causality and strict exogeneity, first discovered by Sims (1972), is further discussed by Engle, Hendry and Richard (1983). Hosoya (1977) shows that this relationship continues to hold in a bivariate setting when processes are not necessarily stationary and have deterministic components, a situation not originally consider by Granger (1969). The strict exogeneity restrictions discussed in Sims (1972) have become known in the literature as 'Sims non-causality'. In more general terms, and in the context of the process χ_t, the absence of Sims causality of x_t for y_{t+1} is defined as a conditional independence restriction where

$$P\left(x_t \in A \middle| Y_{-\infty}^{\infty}, X_{-\infty}^{t-1}, Z_{-\infty}^{t}\right) = P\left(x_t \in A \middle| Y_{-\infty}^{t}, X_{-\infty}^{t-1}, Z_{-\infty}^{t}\right)$$

(S)

for all t or equivalently as $Y_{t+1}^{\infty} \perp x_t | Y_{-\infty}^{t}, X_{\infty}^{t-1}, Z_{-\infty}^{t}, \forall t$. This definition appears in Florens (2003) and Angrist and Kuersteiner (2004) and is closely related to Chamberlain (1982) except that here additional information contained in $Z_{-\infty}^{t}$ is allowed for.

The equivalence between the Granger and Sims notions of causality extends beyond the prediction criterion for covariance stationary processes to definitions of non-causality based on conditional independence restrictions, but, as discussed below, in models with time-varying covariates z_t Sims causality is more appealing and easier to link to the potential-outcomes causality concepts widely used to analyse randomized trials and quasi-experimental studies. Sims (1977, p. 30) emphasizes the use of strict exogeneity restrictions for model identification in structural models and notes that these restrictions can often be related to decision rules of economic agents. Granger non-causality, on the other hand, is a restriction of the reduced form. Structural vector autoregressions (VARs) provide an example of this difference where Sims non-causality imposes restrictions on the impulse response function of structural innovations (Sims, 1986, p. 9, discusses the nature of structural innovations and mentions explicitly that they may include random fluctuations of policies) while Granger non-causality imposes restrictions on the reduced form VAR representation of the model. That the two are in general not equivalent will be discussed in Section 3.

2 Motivation and interpretation of Granger – Sims causality

In order to understand the significance of the original definition of Granger–Sims causality in Granger (1963; 1969) and Sims (1972), it is useful to briefly review the preceding debate in the literature. Simon (1953) defines causality as ‘properties of a scientist's model’. Asymmetric functional relationships between variables are given causal interpretations but Simon emphasizes that no notion of time is needed for this definition. In the context of linear systems of equations Simon's definition of causal relationships is equivalent to a block recursive structure of the equations. Simon (1953, p. 65) emphasizes the need for a priori knowledge about the system to identify the block recursive structure. Wold (1954) and Strotz and Wold (1960) strengthen this definition of causality to require a model to be fully recursive. If the model has three variables *a*, *b* and *c* then recursiveness means that *a* can be solved without knowing *b* and *c*, the solution for *b* generally depends on *a*, and the solution for *c* depends on both *a* and *b*. Such a system then is interpreted to have a causal relationship where *a* causes *b* and *a* and *b* both cause *c*. Wold (1954, p. 166) writes: ‘The relationship is then defined as causal if it is theoretically permissible to regard the variables as involved in a fictive controlled experiment.’

The idea is that, because *a* does not depend on *b* an *c*, it can in principle be controlled or changed by an experimenter. In the terminology of Wold (1954, p. 166), *a* is the cause and *b* and *c* are the ‘effect variables’. Wold (1954) discusses the distinction between randomized experiments in the sense of R. A. Fisher and non-experimental observations. In the latter case causal interpretations depend on ‘subject-matter theory’ (Wold, 1954, p. 170) to identify recursive structures. In other words, a priori assumptions about the structure of the model replace randomized experiments as the source of identification. This approach remains popular to this day for the identification of structural VARs where recursive relationships based on a casual appeal to economic theory are imposed. Orcutt (1952) discusses causal

chains or triangular systems but also mentions the possibility of using temporal structures to identify causal links. Basmann (1963) criticizes the recursive identification schemes of Wold and argues in favour of more general structural economic models to identify causal relationships.

Both Granger and Sims voice scepticism about identifying restrictions which can in general not be tested but have to be accepted a priori. Granger (1980, p. 335) writes,

> [i]f these assumptions are correct or can be accepted as being correct, these definitions may have some value. However, if the assumptions are somewhat doubtful, these definitions do not prove to be useful.

And Sims (1972, p. 544; see also 1977, p. 33) notes,

> [i]f one is willing to identify causal ordering with Wold's causal chain form for a multivariate model, and if enough identifying restrictions are available in addition to those specifying the causal chain form, one can test a particular causal ordering as a set of overidentifying restrictions. The conditions allowing such a test are seldom met in practice, however.

The great advantage of Granger's definition of causality is that it is directly testable from observed data. Granger (1969) gives operational definitions and discusses testable parameter restrictions in linear time series specifications. The strict exogeneity restrictions derived by Sims (1972) are particularly revealing of the power of the flow of time as the identifying force behind uncovering causal links between time series. On the assumption that x_t is strictly exogenous, then, if x_t does not cause y_t it must be conditionally independent of future outcomes y_{t+j}. On the other hand, if conditional correlations between future outcomes y_t and current values of x_t are detected, they must be due to a causal influence of x_t on y_t, because, by assumption, the possibility of a causal link between events determining y_{t+j} that lie in the future and current observations of x_t have been excluded and thus cannot be the source of the observed correlation.

As Granger (1963; 1980) notes, the notion of causality has a long and controversial tradition in philosophy. Some treatments discussing relationships to econometric and statistical practice include Pearl (2000) and Hoover (2001). Holland (1986) discusses the causality definitions of David Hume, John Stuart Mill and Suppes (1970) in the context of Rubin's (1974) causal model. Holland points out that Hume's criteria for causality include the axiom of temporal precedence as well as the requirement of a ‘constant conjunction’ between cause and effect. (See Hoover, 2001, p. 8, for more discussion of Hume's concept of ‘constant conjunction’.) Suppes (1970) proposes a probabilistic theory of causality where he replaces constant conjunction with the requirement that

> one event is the cause of another if the appearance of the first event is followed with high probability by the appearance of the second, and there is no third event that we can use to factor out the probability relationship between the first and second events. (Suppes, 1970,

p. 10)

The definition of Suppes has some parallels to Granger's definition, notably the requirement of temporal succession, the fact that there are no restrictions on what can be a cause and the fact that causes are defined through their effect on the conditional distribution of the effect variable. Finally, Holland attributes the idea of identifying causal effects through experimentation to Mill. Experimentation has since played a central role in the statistical analysis of causality, although Granger (1980, p. 329) mentions it only in passing and does not rely on it in his definition of causality. An important consequence of an experimental concept of causality is that, as Holland (1986, p. 954) writes, ‘causes are only those things that could, in principle, be treatments in experiments’. As discussed in Section 4, this is a critical difference to the concept of Granger causality which does not restrict possible causes.
Feigl (1953) discusses various aspects of the definition of causality that have appeared historically in philosophy and attempts to extract what he calls a ‘purified’ definition of causality. Zellner's (1979) critique of the concept of Granger causality is centered on Feigl's definition according to which ‘[t]he clarified (purified) concept of causation is defined in terms of predictability according to a law’ (Feigl, 1953, p. 408). Feigl (1953, p. 417) continues to note that ‘[p]rediction may be analyzed as a form of deductive inference from inductive premises (laws, hypotheses, theories) with the help of descriptions or existential hypotheses’. Zellner (1979, p.12) writes: ‘predictability without a law or set of laws, or as econometricians might put it, without theory, is not causation.’ In other words, the causality concept put forward by Feigl is based on a priori theoretical assumptions used to generate predictions, while the Granger–Sims notion of causality replaces these a priori restrictions with the axiom of temporal priority.
Feigl (1953, p. 417) notes that causal relationships can be defined even when cause and effect are contemporaneous. According to Feigl, a more important distinction between cause and effect lies in the controllability of the cause as opposed to the effect, which leads Feigl to recommend experimental methods as the best way to identify causal factors. In light of Feigl's work Zellner's main critique of Granger–Sims causality is that it is not based on economic theory to identify causes. (Leamer, 1985, also strongly rejects the idea of conducting causal inference without relying on a priori theory.) In a reply to Zellner, Sims (1979, p. 105) notes that Feigl's definition is at least as ambiguous as the term ‘causality’ itself and that it is so general that it encompasses many other definitions of causality.
Zellner also criticizes three more specific features of Granger causality. First, the requirement that the information set needed to define relevant conditional distributions contain all available information makes the definition non-operational. Zellner (1979, p. 33) writes, ‘Granger does not explicitly mention the important role of economic laws in defining the set of “all relevant information”’ and emphasizes that additional assumptions beyond statistical criteria are necessary to implement tests for Granger non-causality. Second, the limitation to stochastic phenomena and the assumption or axiom of temporal priority of a cause is unnecessarily restrictive compared with other definitions of causality, such as the one of Feigl (1953), which does not rely on these restrictions. And finally, the use of the prediction error variance as a criterion for predictability and the reliance on an optimal predictor, which according to Zellner may both not be well defined, is too restrictive. As far as this last point is concerned it should be noted that more general definitions of Granger–Sims causality proposed by Granger (1980), Chamberlain (1982) and Florens and Mouchart (1982), which are based on conditional independence

restrictions of the joint distribution, do not have the problems that Zellner mentions because these distributional restrictions can be formulated for any process with well-defined conditional distributions. Zellner further points out that economic theory can play a role in providing overidentifying restrictions that allow directions of causality to be imposed. Sims (1979) objects to this last suggestion on the grounds that a test for causality based on overidentifying restrictions is always a joint test of the correctness of such restrictions and the hypothesis of interest and is thus never conclusive. On the other hand, strict exogeneity and thus Granger non-causality, as pointed out by Sims (1977, p. 30, 33) provides overidentifying restrictions that can be tested for. The scepticism about untestable identifying restrictions is forcefully expressed in Sims (1980).

The role of economic theory in identifying parameters of interest in empirical studies remains one of the most controversial issues in econometrics and empirical economics to date. The debate over the correct definition of causality hinges on what individual researchers are prepared to assume a priori, be it restrictions on the temporal direction of cause and effect or fundamental structures that govern the interaction between economic variables. Granger (1980) does not dispute the potential usefulness of a priori theoretical restrictions in identifying causal relationships but emphasizes the potential for misleading inference should these restrictions turn out to be incorrect.

The problem of specifying the correct information set is recognized by Granger (1969), where it is suggested to restrict the set of all available information to the set of relevant information. Granger (1980) discusses a number of examples that illustrate the sensitivity of a causal relationship between two variables to additional information in the conditioning set. This problem is also mentioned by Holland (1986). It seems, however, that, even though specification issues are of great importance in applied work, this is not a fundamental limitation of the causality concept put forward by Granger. Moreover, correct specification of relevant conditioning variables is a common problem in most statistical procedures applied to economic data and thus not specific to procedures testing for causality. At the same time, the argument in favour of guidance from economic theory when designing such procedures is probably strongest when it comes to selecting the relevant variables that need to be included in the analysis, a point elaborated in more detail below.

Further problems of interpretation are discussed in Granger (1980). Simultaneity occurs if x_t causes y_{t+1} and y_t causes x_{t+1}. In a bivariate system of equations this form of feedback, as Granger (1969) defines it, typically leads to other inferential problems as discussed in Sims (1972). In particular, the lack of exogeneity in this case invalidates conventional regression methods and complicates the interpretation of reduced form parameters such as in VARs. Furthermore, Granger causality is not a transitive relationship: if x_t causes y_{t+1} and y_t causes z_{t+1} then x_t does not necessarily cause z_{t+1}. Granger (1980, p. 339) gives the following example. Assume that ε_t and η_t are mutually independent i.i.d. sequences and that $x_t = \varepsilon_t$, $y_t = \varepsilon_{t-1} + \eta_t$ and $z_t = \eta_{t-1}$. Then, because $y_t = x_{t-1} + \eta_t$ it is clear that x_t causes y_{t+1}. In the same way, y_t causes z_{t+1} but x_t does not cause z_{t+1} if the conditioning set contains only $X_{-\infty}^{t}$ and $Z_{-\infty}^{t}$, which is the typical assumption in bivariate statements of causality. At the same time, x_t does cause z_{t+1} in this example when the information set is enlarged to $(X_{-\infty}^{t}, Y_{-\infty}^{t}, Z_{-\infty}^{t})$ because now the innovation η_{t-1}can be recovered from $\eta_{t-1} = y_{t-1} - x_{t-2}$. This example shows that the concept of Granger–Sims causality can be sensitive to the specification of

conditioning information.

3 Equivalence and non-equivalence of Granger and Sims causality

Since the original contributions of Granger (1963; 1969) and Sims (1972), which were mostly cast in terms of forecast error criteria, there has been a sizeable literature concerned with extensions of the basic definition and establishing a number of equivalence relationships. While the conditional independence formulation of Granger causality goes back at least to Granger (1980, p. 330), a formal analysis of the equivalence with a corresponding definition of Sims causality was first obtained by Chamberlain (1982) and Florens and Mouchart (1982). It turns out that the condition for Granger non-causality, which in its more general form can be stated as $y_{t+1} \perp X_{-\infty}^{t} | Y_{-\infty}^{t}, \forall t$, does imply the generalized form of Sims non-causality formulated as $Y_{t+1}^{\infty} \perp x_t | Y_{-\infty}^{t}, \forall t$, but the reverse implication does not hold generally. Florens and Mouchart (1982) give a counter-example of a nonlinear process where Sims non-causality holds but Granger non-causality does not hold. As Florens and Mouchart (1982) point out, the two conditional independence relationships are equivalent for Gaussian processes where lack of covariance is equivalent to independence. Chamberlain (1982) shows, on the other hand, that a generalized form of Sims non-causality, stated as $Y_{t+1}^{\infty} \perp x_t | Y_{-\infty}^{t}, X_{-\infty}^{t-1}$, is equivalent to $y_{t+1} \perp X_{-\infty}^{t} | Y_{-\infty}^{t}$ under a mild regularity condition limiting temporal dependence. Florens and Mouchart (1982) obtain a very similar result for slightly different definitions of the conditioning sets. General statements of this result can also be found in Bouissou, Laffont and Voung (1986). These authors define additional causality relationships: global non-causality (C) is defined as $Y_{t+1}^{\infty} \perp X_{-\infty}^{t} | Y_{-\infty}^{t}, \forall t$, Granger non-causality of order *k* (G_k) is defined as $Y_{t+1}^{t+k} \perp X_{-\infty}^{t} | Y_{-\infty}^{t}, \forall t$ and Sims non-causality of order *k* (S_k) is defined as $Y_{t+1}^{\infty} \perp X_{t-k+1}^{t} | Y_{-\infty}^{t}, X_{t-k}, \forall t$ where X_{t-k} is any subset of $X_{-\infty}^{t-k}$. It is then shown that (G_k), (S_k) and (C) are all equivalent for all *k*, a result that is also stated in Florens and Mouchart (1982, p. 580). Pierce and Haugh (1977) propose an alternative definition of Granger causality in the context of linear processes. If $H_{\{Y_0^t, U\}}(y_{t+1})$ is the linear projection of y_{t+1} on $\{Y_0^t, U\}$ where $\{Y_0^t, U\}$ is the closed linear span of all the variables generating Y_0^t and the initial conditions *U*, then x_t does not Granger cause y_{t+1} if the innovations $y_{t+1} - H_{\{Y_0^t, U\}}(y_{t+1})$ and $x_{p+1} - H_{\{X_0^p, U\}}(x_{p+1})$ are uncorrelated for all $p \leq t$. Florens and Mouchart (1985) show that this definition is equivalent to covariance-based definitions of Granger and Sims causality under some additional regularity conditions. Generally speaking, the results of this early literature show that Granger causality between two processes *x* and *y* is equivalent to appropriate definitions of Sims causality not only in a mean squared prediction error sense but more generally in terms of restrictions on appropriate conditional distributions of the joint process. It should also be noted that these equivalence results continue to hold when both *x* and *y* are vector valued processes.

The situation changes, however, quite markedly when an additional set of covariates *z* is added to the analysis. In this situation Granger non-causality defined as $y_{t+1} \perp X_{-\infty}^{t} | Y_{-\infty}^{t}, Z_{-\infty}^{t}, \forall t$ in general is not equivalent to Sims non-causality defined as $Y_{t+1}^{\infty} \perp x_t | Y_{-\infty}^{t}, X_{\infty}^{t-1}, Z_{-\infty}^{t}, \forall t$. This result seems to have been first obtained by Dufour and Tessier (1993) for the linear case and also appears in Florens (2003) and Angrist and Kuersteiner (2004) and in the biostatistics literature in Robins, Greenland and Hu (1999). Simple examples can be constructed where *x* Granger-causes *y* but does not Sims cause *y* as well as cases where *x* Sims causes *y* but does not Granger cause *y*. In related work Lütkepohl (1993) and Dufour and Renault (1998) show that, in general, (G_1) does not imply (C) if the information set contains *z*.

To illustrate the result of Dufour and Tessier (1993), assume that $X_t = (y_t, x_t, z_t)$ and that χ_t can be represented as linear functions of present and past structural innovations e_t. To simplify the exposition assume that for $C(L) = A(L)^{-1}$ where $A(L)$ is a matrix of lag polynomials of finite order and L is the lag operator, it holds that $X_t = C(L) e_t$. Also assume that the diagonal blocks of $C(0)$, partitioned according to (y_t, x_t, z_t), are full rank. The reduced form VAR representation of χ_t then is $\pi(L) X_t = u_t$ with $u_t = C(0) e_t$ and $\pi(L) = C(0) A(L)$. As was discussed before, Sims non-causality imposes zero restrictions on off-diagonal blocks of $C(L)$ while Granger non-causality imposes zero restrictions on corresponding off-diagonal blocks of $\pi(L)$. Now note that when χ_t contains only y_t and x_t, the partitioned inverse formula implies that off-diagonal blocks of $A(L)$ are zero if and only if corresponding blocks of $C(L)$ are zero. Because the latter can hold only if corresponding blocks of $C(0)$ are zero, it follows that $\pi(L)$ has zero off-diagonal blocks if and only if corresponding blocks of $C(L)$ are zero. This is the result of Sims (1972). On the other hand, when $X_t = (y_t, x_t, z_t)$ the partitioned inverse formula for matrices partitioned into three blocks shows that $C(L)$ having off-diagonal blocks no longer implies that corresponding blocks of $A(L)$ are necessarily zero as well. Thus the equivalence between Sims and Granger causality no longer holds when additional time varying covariates are included in the analysis. In Section 4, applications in monetary economics are discussed where this situation arises naturally.

4 The connection between Granger – Sims causality and potential outcomes

The notion of causality that has become standard in micro-econometrics is based on Rubin's (1974) concept of potential outcomes, which at its core uses experimental variation to identify causal relationships. The potential outcomes model has been extended to and applied in observational studies. Observational studies are situations where no experimental assignments of actions were used. Examples are medical trials where experiments might be unethical or many economic policy questions where experiments may be unethical or too expensive to carry out. The importance of experimental evidence was recognized in econometrics dating back to Haavelmo (1944). Wold (1954) similarly discusses controlled experiments as a way to uncover causal relationships. Orcutt's (1952, p. 305) notion of causality is closely related to the idea of potential outcomes and is defined in terms of consequences of actions:

> Thus when we say that *A* is the cause of *B*, we often mean that if *A* varies, then *B* will be different in a specified way from what it *would have been* if *A* had not varied.

Orcutt (1952, p. 309) goes on to discuss policy actions as a substitute for unavailable experimental evidence to identify causal relationships in observational data, an idea explored in more detail below. For expositional purposes, assume that a certain action or treatment *D* can be either given or not given to individual *i*. The causal question in this context is whether the treatment has an effect on an outcome variable of interest measured by *y*. It is convenient to define $D_i = 1$ if the treatment is given and $D_i = 0$ if the treatment is not given. The potential outcome $y_i(0)$ is defined as the outcome for individual *i* that would have occurred if the treatment had not been given and $y_i(1)$ as the outcome that would have occurred in the case the treatment had been given. The absence of causality of *D* for *y* then is defined as the situation where $y_i(0) = y_i(1)$. This condition is referred to as the 'strong null hypothesis of no causal effect'. Usually this condition cannot be directly tested because $y_i(0)$ and $y_i(1)$ are not both observed for the same individual. Instead, the observed measurement takes the form

$$y_i = D_i y_i(1) + (1 - D_i)\, y_i(0) \qquad \text{(PO)}$$

Potential outcomes may depend on a list of covariates z_i. Covariates capture characteristics of the outcome variable that are not directly related to the experiment but that need to be taken into account when assessing the outcome. An identification condition is needed to proceed to testable restrictions. Formally one imposes the condition $y_i(0), y_i(1) \perp D_i | z_i$, which is sometimes referred to as selection on observables. This condition is automatically satisfied if D_i is randomly assigned in an experiment. In observational studies the condition essentially states that actions by individuals or policymakers cannot be based on unobservable information. The 'selection on observables' condition implies that

$$P(y_i(j) \in A | z_i) = P(y_i(j) \in A | D_i, z_i)$$

for all *j* and the null hypothesis implies that

$$P(y_i \in A | D_i, z_i) = P(y_i \in A | z_i)$$

which is identical to Granger's condition of no causal effect, a result that is discussed in Holland (1986).

The power of the identifying restriction lies in the fact that it is formulated independently of the null hypothesis of no causal relationship. To be more precise, the identifying restriction imposes conditional independence of D from $y_i(0)$ and $y_i(1)$ but not from y_i. The latter holds only when the null is true. This is an important difference from Sims (1977, p. 30), who writes that 'Causality is an important identifying restriction on dynamic behavioral relations'. In other words, Sims imposes the null to identify certain structural models. Another important difference between this model of causality and the less specific definition of Granger–Sims causality is that the form of the causal link between D and y is of a simple functional form specified in (PO). This particular structure is what allows the interpretation of measured correlations as causal links.

Identification conditions thus lead to testable implications of Rubin's potential outcome framework that are identical to the Granger–Sims definition of non-causality. Nevertheless, causality in Rubin's context is closely related to experiments and counterfactuals: causal effects of a treatment are measured by comparing unobservable counterfactuals under treatment and non-treatment. On the other hand, Granger–Sims causality does not rely on the notion of treatment. It has been applied to studying such phenomena as the temporal link between interest rates and inflation, variables that are endogenously determined and where it is hard to imagine that an experiment or even a policy intervention is available for causal inference. Orcutt's idea of using policy variation can, however, still be implemented if instead of market interest rates one focuses (in the United States) on the federal funds target rate, a variable that is directly set by the policymaker. Under the additional assumption that all systematic aspects of the policy depend on observable information, it is possible to generate pseudo-experimental variation even in the interest rate example. These ideas are now explored in more detail.

The potential outcomes framework in its original form is in many ways too limited to be directly applicable to macroeconomic questions of causality where the Granger–Sims concept of causality has been mostly applied. The two main limitations of the potential outcomes approach are that it does not allow for dynamic treatments or policies and that usually the stable unit treatment assumption of Rubin (1980) is imposed. The latter rules out general equilibrium effects of treatments and is not satisfied in a macroeconomic context. Angrist and Kuersteiner (2004) propose an extension of the potential outcomes framework that overcomes these limitations (also related is the work of White, 2006). Consider an economy that is described by $X_t = (y_t, D_t, z_t)$ where y is a vector of outcome variables, D is a vector of policy variables and z is a vector of relevant covariates not already included in y. Potential outcomes $y_{t,\bullet j}\bullet(d)$ are defined as values the outcome variable $y_{+,\bullet j}$ would have taken if at time t the policy had been set to $D_t = d$. It is probably useful to discuss the nature of the potential outcome $y_{t,\bullet j}\bullet(d)$ in a context where one has a dynamic general equilibrium model describing the evolution of the process χ_t as a system of stochastic difference equations. Then $y_{t,\bullet j}\bullet(d)$ has to be thought of as a specific solution of that model indexed against a specific decision rule d of some policymaker. It is helpful, but not necessary, to assume that the model has a strong solution, in the sense of stochastic process theory, such that the strong null of no causal effect can be represented as the restriction $y_{t,j}(d) = y_{t,j}(d')$ for all possible values of d. It should be clear from this description that $y_{t,\bullet j}\bullet(d)$ is a possibly highly complex function of all the inputs that go into the model, including policy decisions taken at times different from t. Solving for $y_{t,\bullet j}(d)$ explicitly is not necessary for the definition of a causal link between D and y, a feature that is

very much in the spirit of Granger and Sims. All that is needed is an identifying restriction that allows us to interpret observed correlations as causal links. A sufficient condition is

$$Y_{t,j}(d) \perp D_t | Y_{-\infty}^{t}, D_{-\infty}^{t-1}, Z_{-\infty}^{t}, \forall t, \quad \forall j > 0.$$
(ID)

Under the sharp null of no causal effect where $Y_{t,j}(d) = Y_{t+j}$ it then follows immediately that $Y_{t+j} \perp D_t | Y_{-\infty}^{t}, D_{-\infty}^{t-1}, Z_{-\infty}^{t}$. This is the same as the condition for Sims non-causality. As discussed earlier, it is generally not equivalent to Granger non-causality. The form of the testable restriction depends critically on the form of (ID). At least in cases where *D* is a decision variable of a policymaker, this restriction leading to Sims causality seems plausible. This can be seen easily in the context of linear models where the identification assumption leading to Sims non-causality is identical to the restriction that policy innovations are independent of all future innovations affecting the outcome variables.

In order to better understand the nature of the identifying assumption (ID) it is useful to consider a specific example. The notion of Granger causality was applied to the question of a causal link between monetary policy and real economic activity, starting with Sims (1972), and thus has a long tradition in the empirical macro literature. Most of the early empirical literature has investigated this question using linear regressions of some measure of monetary aggregates or interest rates and various measures of real economic activity. In an important methodological contribution, Romer and Romer (1989) use information from the minutes of the Federal Open Markets Committee to classify US Federal Reserve policy into times of purely anti-inflationary monetary tightening and other periods. Times of tight monetary policy are called Romer dates. The idea then is to measure average economic activity following Romer dates and to compare these measurements to average economic activity at other times. While the argument that Romer dates are exogenous has been criticized (see for example Hoover and Perez, 1994; Shapiro, 1994; Leeper, 1997), the basic premise of the approach of Romer and Romer (1989) remains valid. It is to use a behavioural theory or policy rule for a policymaker to construct policy innovations which serve as exogenous variation that can be used to evaluate the effectiveness of the policy in question (Jorda, 2005, emphasizes the importance of exogenous variation to identify causal relationships). Angrist and Kuersteiner (2004) analyse the consequences of allowing for additional covariates *z* in the policy model to capture information about nominal macroeconomic variables such as the inflation rate. These variables are clearly relevant for policy decisions of the Federal Reserve and thus constitute relevant conditioning information in the sense of Granger (1969). At the same time these nominal variables are not part of the null hypothesis of no causal effects on real economic activity and thus cannot be subsumed into the *y*-process. As discussed earlier, under these circumstances reduced form regressions based on Granger's notion of non-causality cannot be used to test for Sims non-causality.

Generally speaking, a model of the policymaker, in this case the Federal Reserve, is a conditional

probability distribution $P(D_t \in A | Y_{-\infty}^{t}, D_{-\infty}^{t-1}, Z_{-\infty}^{t})$. An example of a policy model for the Federal Reserve in the context of the Romer and Romer data was developed by Shapiro (1994). The fundamental identifying assumption then is that this model is correct, especially in the sense that the conditional probability of D_t does not depend on $y_{t,\bullet j}(d)$. This condition will be satisfied when two criteria are met. All relevant information that the policymaker used to decide on the policy D_t is included in the model and the problem at hand is of a nature where the policymaker does not foresee the future. This is the way Granger's fundamental axiom of ‘the arrow of time’ plays a role to provide identifying assumptions in this setting. If these conditions are met then all random deviations of the policy D_t from what is predicted by the model are conditionally independent of $y_{t,\bullet j}(d)$. Random deviations could be due to the variation over time in policymakers’ beliefs about the workings of the economy, decision-makers’ tastes and goals, political factors, and the temporary pursuit of objectives other than changes in the outcomes of interest (for example, monetary policy that targets exchange rates instead of inflation or unemployment), and finally harder-to-quantify factors such as the mood and character of decision-makers. It is then precisely these random deviations from prescribed policies that help to identify causal links. In other words, it is not the systematic or predictable policy changes that are helpful to answer causal questions but the deviations from prescribed rules. The reason for this is that the causal model used here does not provide enough structure to disentangle causal links from endogenously varying policy variables. The situation is quite similar to the analysis of impulse response functions in structural VAR models where identification is driven by the independence of structural innovations. Impulse response analysis can thus be viewed as a special case of the potential outcomes model when χ_t is a linear process.

The potential outcome framework has the advantage that it focuses on exogenous variation and puts the identification discussion at the centre of causal inference. It helps to clarify the source of identifying variation in an analysis of Granger–Sims causality. A priori arguments for the identifying exogeneity restrictions can be based on institutional settings such as the introduction of new legislation, on procedural details as in the Romer and Romer (1989) example, or on behavioural models derived from economic theory. At the same time the potential outcome approach to identification is limited in the sense that its most natural areas of application lie in the analysis of policy effectiveness. It is less suited to analyse causality between variables that are jointly determined in equilibrium.

A point made by Granger (1980) is also relevant here. The analysis of causality is not necessarily relevant for the analysis of controlled processes. To illustrate the issue, consider the linearized Lucas model of McCallum (1984), where in an overlapping generations framework price setting happens in isolated markets based on local information. McCallum shows that random innovations to the money stock affect unemployment because agents cannot completely distinguish between real price changes in their markets and price changes due to variation in the supply of money. A test of conditional independence between money and employment for data generated by this model would find evidence of a causal relationship in the sense of Sims. At the same time, any attempt by the monetary authority to systematically exploit this relationship through a systematic policy rule would fail in this model because agents fully incorporate predictable actions of the policymaker and do not respond to nominal changes in prices. This example shows that a statistical definition of causality may indicate the existence of a causal relationship that does not lend itself to policy intervention and control. Whether individual researchers

are willing to call such a finding a causal relationship hinges upon their notion of causality and is likely to be controversial.

The situation is reversed in some models where the monetary authority can fully control output through appropriate policy rules. For the purpose of illustration consider the model by Rudebusch and Svensson, which ‘consists of an aggregate supply equation that relates inflation to an output gap and an aggregate demand equation that relates output to a short term interest rate’ (1999, p. 205). Monetary policy is conducted by setting the nominal interest rate, and affects output and inflation with a one period lag. In this model it is possible for the monetary authority to fully stabilize output such that deviations from a fixed steady state level are serially uncorrelated. On the assumption that the policy rule is augmented with an independently distributed policy innovation (this assumption is necessary for statistical identification of test procedures, see Sims, 1977, p. 39), it follows that a test for Granger causality will not be able to reject the null of no causal relationship running from interest rates to output. At the same time, a test for Granger non-causality of output for interest rates will be rejected because the interest rate setting rule depends on past output. In this example, the direction of causality in the statistical sense of Granger goes in the opposite direction of what the model indicates.

On the other hand, a test of the conditional independence restriction (S) for Sims non-causality of interest rates for output will be rejected, thus revealing the influence of the policymaker on output. Sims (1977, p. 36) considers a similar reversal of the direction of Granger causality in models where a policy variable is controlled. While Sims considers a bivariate model where both Granger and Sims causality are equivalent, the model of Rudebusch and Svensson considered here has three equations, which explains why the concepts of Granger and Sims causality do not lead to the same conclusions.

A test of (S) in this model is thus able to identify the direction of causality even when variables are controlled, at least when the test is based on the assumption that the policy model is correctly specified and the policy innovation is thus identified. However, even in this case, the measured causal effects are those of random deviations from the policy rule. As discussed earlier, attempts to exploit these effects with systematic policy actions may not be feasible due to the reactions of rationally forward-looking agents.

A related issue is the problem of analysing causal effects of systematic changes in the policy rule, a problem discussed in Sims (1977, p. 30). Without additional structure such questions seem to be hard to address, and it remains an open question to what extent evidence gained from causal inference based on notions of Granger–Sims causality can be used to investigate them.

5 Summary

This article explores the notion of Granger–Sims causality as a concept of statistical predictability. The definition is appealing because it does not require a priori theoretical restrictions but rather is formulated in terms of a directly testable implication on the distribution of observed data. The simplicity of this approach to causality has led to extensive applications in areas such as macro-econometrics, where notions of causality that rely on the possibility of carrying out experiments are difficult to apply. Difficulties with a purely statistical concept of causality, however, arise when it comes to interpreting the nature of detected causal relationships. Without additional assumptions regarding the exogeneity of one or more input variables, it seems difficult to link the statistical causality concept with the more

fundamental distinction between cause and effect. The latter distinction is fundamental to the analysis of controllability of outcome variables and thus central to many questions in the social sciences. As discussed above, there is clearly a distinction between a causal link between two variables and the possibility of controlling an output by manipulating certain inputs. Equilibrium effects which are at the core of economic analysis may, for example, pre-empt policy changes through the agent's rational anticipation of just these policy changes. Perceived causal relationships thus may not be exploitable for policy purposes even if they can be reliably identified in the history of an economy. The analysis of causality and controllability in dynamic equilibrium models thus remains a central topic of research.

See Also

- causality in economics and econometrics
- continuous and discrete time models
- Granger, Clive W.J.
- Hume, David
- propensity score
- Rubin causal model
- Simon, Herbert A.
- structural vector autoregressions

Bibliography

Angrist, J. and Kuersteiner, G. 2004. Semiparametric causality tests using the policy propensity score. Working Paper No. 10975. Cambridge, MA: NBER.

Basmann, R. 1963. The causal interpretation of non-triangular systems of economic relations. *Econometrica* 31, 439–48.

Bouissou, M., Laffont, J.-J. and Voung, Q. 1986. Tests of noncausality under Markov assumptions for qualitative panel data. *Econometrica* 54, 395–414.

Chamberlain, G. 1982. The general equivalence of Granger and Sims causality. *Econometrica* 50, 569–81.

Dawid, A. 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, Series B 41, 1–31.

Dufour, J.-M. and Renault, E. 1998. Short run and long run causality in time series: theory. *Econometrica* 66, 1099–125.

Dufour, J.-M. and Tessier, D. 1993. On the relationship between impulse response analysis, innovation

accounting and Granger causality. *Economics Letters* 42, 327–33.

Engle, R., Hendry, D. and Richard, J. 1983. Exogeneity. *Econometrica* 51, 277–304.

Feigl, H. 1953. Notes on causality. In *Readings in the Philosophy of Science*, ed. H. Feigl and M. Brodbeck. New York: Appleton-Century-Crofts, Inc.

Florens, J.-P. 2003. Some technical issues in defining causality. *Journal of Econometrics* 112, 127–8.

Florens, J.-P. and Mouchart, M. 1982. A note on noncausality. *Econometrica* 50, 583–91.

Florens, J.-P. and Mouchart, M. 1985. A linear theory for noncausality. *Econometrica* 53, 157–76.

Granger, C. 1963. Economic processes involving feedback. *Information and Control* 6, 28–48.

Granger, C. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.

Granger, C. 1980. Tests for causation – a personal viewpoint. *Journal of Economic Dynamics and Control* 2, 329–52.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12 (supplement), iii–vi, 1–115.

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–60.

Hoover, K. 2001. *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K. and Perez, S. 1994. Post hoc ergo propter once more: an evaluation of ‘does monetary policy matter?’ in the spirit of James Tobin. *Journal of Monetary Economics* 34, 47–73.

Hosoya, Y. 1977. On the Granger condition for non-causality. *Econometrica* 45, 1735–6.

Jorda, O. 2005. Estimation and inference of impulse responses by local projections. *American Economic Review* 95, 162–82.

Leamer, E. 1985. Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy* 22, 255–303.

Leeper, E. 1997. Narrative and VAR approaches to monetary policy: common identification problems.

Journal of Monetary Economics 40, 641–57.

Lütkepohl, H. 1993. Testing for causation between two variables in higher dimensional VAR models. In *Studies in Applied Econometrics*, ed. H. Schneeweiss and K. Zimmerman. Heidelberg: Springer-Verlag.

McCallum, B. 1984. A linearized version of Lucas's neutrality model. *Canadian Journal of Economics* 17, 138–45.

Orcutt, G. 1952. Actions, consequences, and causal relations. *Review of Economics and Statistics* 34, 305–13.

Pearl, J. 2000. *Causality*. Cambridge: Cambridge University Press.

Pierce, D. and Haugh, L. 1977. Causality in temporal systems. *Journal of Econometrics* 5, 265–93.

Robins, J., Greenland, S. and Hu, F. 1999. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 94, 687–700.

Romer, C. and Romer, D. 1989. Does monetary policy matter? A new Test in the spirit of Friedman and Schwartz. In *NBER Macroeconomics Annual 1989*, ed. O. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* 66, 688–701.

Rubin, D. 1980. Randomization analysis of experimental data: the Fisher randomization test. Comment. *Journal of the American Statistical Association* 75, 591–93.

Rudebusch, G. and Svensson, L. 1999. Policy rules for inflation targeting. In *Monetary Policy Rules*, ed. J. Taylor. Chicago: University of Chicago Press.

Shapiro, M. 1994. Federal Reserve policy: cause and effect. In *Monetary Policy*, ed. G. Mankiew. Chicago: University of Chicago Press.

Simon, H. 1953. Causal ordering and identifiability. In *Studies in Econometric Method*, ed. W. Hood and T. Koopmans. Cowles Commission Monograph No. 14. New York: John Wiley.

Sims, C. 1972. Money, income and causality. *American Economic Review* 62, 540–52.

Sims, C. 1977. Exogeneity and causal ordering in macroeconomic models. In *New Methods in Business*

Cycle Research: Proceedings of a Conference, ed. C. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Sims, C. 1979. A comment on the papers by Zellner and Schwert. *Carnegie-Rochester Conference Series on Public Policy* 10, 103–8.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Sims, C. 1986. Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review* 10(1), 2–16.

Strotz, R. and Wold, H. 1960. Recursive versus nonrecursive systems: an attempt at synthesis. *Econometrica* 28, 417–27.

Suppes, P. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

White, H. 2006. Time-series estimation of the effects of natural experiments. *Journal of Econometrics* 135, 527–66.

Wiener, N. 1956. The theory of prediction. In *Modern Mathematics for the Engineer, Series 1*, ed. E. Beckenback. New York: McGraw-Hill.

Wold, H. 1954. Causality and econometrics. *Econometrica* 22, 162–77.

Zellner, A. 1979. Causality and econometrics, policy and policy making. *Carnegie-Rochester Conference Series on Public Policy* 10, 9–54.

How to cite this article

Kuersteiner, G. M. "Granger–Sims causality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000190> doi:10.1057/9780230226203.0665

The New Palgrave Dictionary of Economics Online

graph theory

A.P. Kirman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Graphs are used in economics to depict situations in which agents are in direct contact with each other. The use of graph theory enables one to understand the basic properties of the communication network in an economy or market. Typical questions include: how does the structure of a network affect economic outcomes and the welfare of the individuals involved? What happens if agents can choose those with whom they interact? How will networks evolve over time? Theoretical results, economic applications and empirical examples are given.

Keywords

clusters; coalitions; complete and incomplete information; connectivity; cores; first order stochastic domination; graph theory; matching; neighbourhoods; network formation; networks; operations research; power laws; probability; small worlds; spatial economics; spillover effects; stochastic graphs; technological shocks

Article

At first sight it might seem that the rather abstract mathematical theory of graphs would be of little relevance for economics. That this is not entirely the case is largely due to developments since the early 1980s. Economists have become interested in the structure of the relations between individuals, firms and groups and their importance for economic activity. These relations can be viewed as a graph, and the properties of the particular graph will have specific implications for the economic outcomes in the situation modelled. A simple example may help. Consider a competitive economic market. Agents receive prices from a central source and do not communicate with each other. This can be represented as a *star*. The centre which emits the prices is often thought of as the ‘Walrasian auctioneer’. However, in many economic situations the organization may be very different, and it is of interest to know what consequences this may have. For example, in an ordinary *n* person game everyone is conscious of what

every other player is doing, and this situation could be represented by a complete graph in which there is a link between every pair of players.

While accounts of the use of networks in economics can be found elsewhere (see Goyal, 2006; Jackson, 2004 for excellent surveys), the question for this article is the extent to which graph theory has helped economists answer the questions they analyse.

Basic concepts in graph theory

First, a few definitions are necessary to set the stage. Think of a set V of *nodes*, (or economic entities), each pair of the nodes will be linked $v(ij)=1$ or not $v(ij)=0$, by an *edge*. If there are N nodes then a *simple graph* can be represented by its *adjacency matrix* with each element $v(i)j$ being 1 or 0 depending on whether i and j are linked. Obviously, other considerations could be included. For example, if the graph is *directed*, $v(ij)=1$ does not imply $v(ji)=1$; an obvious example of this would be an input–output system where a good i enters into the production of j but not vice versa. An *undirected graph*, on the other hand, has a symmetric adjacency matrix, and here a simple example would be that of co-authors of papers.

The *degree*, $d(v)$, of a vertex v is the number of edges with which it is incident. Two vertices are *adjacent* if they are incident to a common edge. The set of *neighbours*, $N(v)$, of a vertex v is the set of vertices which are adjacent to v. The *degree* of a vertex is the number of neighbours it has, or formally, the cardinality of its neighbour set.

A *path* is an alternating sequence of vertices and edges, with each edge being incident to the vertices immediately preceding and succeeding it in the sequence and with no repeated vertices.

The *length* of a path is the number of edges in the sequence defining the walk. If u and v are vertices, the *distance* from u to v, written $d(u,v)$, is the minimum length of any path from u to v. In an undirected graph, this is obviously a metric. The *diameter*, $diam(G)$ of the graph G is the maximum value of $d(u,v)$, where u and v are allowed to range over all of the vertices of the graph.

A graph is *complete*, or a *clique*, if every pair of vertices are adjacent. The typical n person game can be thought of as involving a *complete graph* since all players are in contact with each other. This is the other extreme from the star of the competitive model.

Economic applications

With these concepts in mind we can now ask how they may help in analysing economic problems. If we think about situations in which agents or firms are linked through a network, then there are three basic questions. How does the structure of a network affect economic outcomes, which structures are '*efficient*', and which graph structures or topologies are likely to be found? In a number of applications very different graph structures may govern economic interaction. A first and important distinction is between *exogenous* and *endogenous* graphs. In the former the agents are taken as already linked, and one examines the consequences of the way in which they are linked. One can think of the distinction often made in spatial economics between agents situated on a line or on a lattice. In the latter case the number of neighbours would be four or eight depending on how one constructs the links. In more general networks, the sort of question asked is: how fast will information travel, or how quickly will an epidemic of opinion or innovation grow? The two important characteristics here will be the *connectivity*

of a graph and a measure of the maximum distance between two nodes, the *diameter* of the graph. Another important notion is that of a *neighbourhood* and the burgeoning literature on externalities and spillover effects has come to pay attention to the structure of neighbourhoods and the idea that individuals in certain neighbourhoods may become isolated from the global network (see Durlauf, 2004). The important thing here is that neighbourhoods are not necessarily geographical but are determined by the network of interpersonal relations. The architecture of the graph will have important consequences for the allocation of resources and distribution of wealth in an economy.

One can also pose such questions as: how vulnerable is a network to the removal of links? This too is closely tied to the *connectivity* of the graph. This sort of problem has been extensively looked at in the operations research literature, where the vulnerability of a railroad or communications network to a bombing attack was studied over a long period after the Second World War (see, for example, Frank, 1967). The same techniques have been applied to the problems of power outages in electricity networks, but this sort of analysis has, for the moment, made few inroads into the economics literature. This problem has also been studied in the context of stochastic networks, as discussed below.

The notion of *vertex degree* is obviously applicable to matching problems and many matching problems can be thought of as finding a bipartite graph satisfying certain efficiency criteria. Wherever there is a limit to the number of agents who can interact, possibly for institutional reasons, this translates into the idea that the degree of a node is constrained.

While these observations give some idea of the use of graph theoretic concepts in economics, there are rather few examples of contributions which draw directly on graph theory. Indeed, it would be fair to say that most of the relevant contributions use concepts from this theory for notational rather than for analytical reasons. An example of a contribution which makes direct appeal to results from graph theory is that of Corominas-Bosch (2004). She looks at buyers and sellers linked by an exogenous network. The outcomes will depend on who is linked to whom, and players choose strategies depending on those chosen by those that they are linked with. She then uses a theorem on the partitions of bipartite graphs in order to analyse the sort of networks of connections that will produce the Walrasian outcome. What is the competitive outcome in this framework? If buyers outnumber sellers, all the surplus goes to the sellers and, conversely, if sellers outnumber buyers all the surplus goes to the buyers. When the numbers are equal the surplus is split ‘evenly’ between buyers and sellers. She shows that, given certain conditions, either the whole group of players will implement the competitive solution or the group will partition into groups of players in sub-graphs, each of which will implement the competitive solution. In other words, the graph will partition into sub-graphs, each reflecting the overall relationship between buyers and sellers, and the competitive solution will be implemented in each sub-graph. Thus, what Corominas-Bosch shows is the relation between the architecture of the graph of relations and the efficiency of the outcome.

Another example is that of Anderlini and Ianni (1996), who study the long-run properties of learning from neighbours. In their model, at each step the players play a game with one of their m neighbours, who is chosen at random with probability $1/m$. This is repeated and the final result described. Where is graph theory useful here? The idea that one can, in fact, draw at random at each point in time a one-to-one matching of players is simple, but it is not obvious that the structure of the exogenously given links will permit this. The authors wish to limit the structure of relations between individuals to guarantee that it can be done. The situations in which each member of a group is linked to another single member of

the same group are considered. Sub-graphs of this type are called *1-factors*. Obviously, the number of such matchings depends on the underlying graph structure, which is taken as exogenous. Recall that Anderlinni and Ianni (1996) require, in the original graph, that the vertex degree of all agents shall be *m*. However, it is known from graph theory that certain graphs of this sort may permit no such matching, that is, may have no 1-factor. To handle this, Anderlini and Ianni assume that their graph is the product of *m* 1-factors, so as to guarantee that they can randomly rematch their players at each stage. Once they can do this they can show whether the learning process used by the agents converges. Here, managing a rather natural, but technical, problem in a rematching game involves graph-theoretic tools.

Much of the literature in economics has focused on *undirected graphs*, although in some situations such as input–output matrices the direction is clearly important, since production processes are not reversible. Evstigneev and Taskar (1995), although working in a stochastic context, have modelled some economic equilibrium problems using *directed graphs* but, as they indicate, the task is made difficult by the fact that some of the mathematical tools which make the undirected case tractable are not available for directed graphs.

Up to this point we have considered the agents as located in a given graph. However, one of the most interesting challenges in examining an economy is to include the evolution of the network structures themselves. If one wants to proceed to a theory of endogenous network formation, a first step might be to find out which organizations of individuals are stable. Thus one would look for ‘rest points’ of a dynamic process of network evolution. Such rest points would be arrangements, or networks, which would not be subject to endogenous pressures to change them. This, as it stands, is not a well-formulated concept. More of the rules under which agents operate have to be spelled out. The dynamics of the system will be defined by the way in which links are formed. Good surveys of the literature on this subject (see network formation), are available and, once again, although the terms from graph theory are widely used, there is little evidence of the use of graph theoretic results, at least in the deterministic case, to prove economic propositions. In many cases, well-known architectures such as the *star* turn out to be efficient, and it is not difficult to see why. With *n* agents such a graph has *diameter* 2 and only *n*–1 links. Thus it is an efficient way of organizing links if links have any costs. The problem is to show that such a structure will emerge as a result of a well-specified interaction process.

Stochastic graphs

Paradoxically, it is in the case where the economic graph structure is considered to be *stochastic* that more appeal has been made to graph-theoretic results, such as those developed by Erdos and Renyi (1960).

There are different ways of defining a stochastic graph. The simplest one is that used by Erdos and Renyi themselves in which the 1's and the 0's of the adjacency matrix are replaced by probabilities, so that the edge $e(ij)$ exists with probability $p(ij)$, and so the adjacency matrix is now composed of probabilities and if the graph is undirected the matrix will be *symmetric*, that is, $p(ij)=p(ji)$. (This tool was developed in economics by Kirman, Oddou and Weber, 1986; Durlauf, 1990; and Ioannides, 1990.)

An alternative and more general possibility is to have a probability distribution over all possible adjacency matrices.

An interesting and useful fact is that the graph representing the links through which interaction takes

place may become surprisingly highly connected as the number of agents increases, provided that the probability that any two individuals are connected does not go to zero too fast. To understand what is meant by this, consider a result of Bollobas (1985). He shows that, if the probability that any two agents know each other in a graph *G* with *n* nodes, $p^n(ij)$ is greater than $\frac{1}{\sqrt{n}}$ for all *i* and *j* and for all *n*, then as *n* becomes large it becomes certain that the *diameter* of the graph, $D(G^n(p^n(ij)))$ will be at most 2. More formally,

$$\lim_{n \to \infty} \text{Prob}\left(D\left(\Gamma^n\left(p_{ij}^n\right)\right) \leq 2\right) = 1$$

In other words, any two individuals will be sure to have a ‘common friend’ if the graph is large enough. Thus, as was observed in Kirman, Oddou and Weber (1986), one should say on encountering someone with whom one has a common friend, ‘it's a large world’. This somewhat surprising result suggests that, as sociologists have long observed empirically, relational networks are likely to be much more connected than one might imagine.

It was this sort of result which led to the first use of stochastic graphs in economics in Kirman (1983). The application there was very simple. The *core* of an exchange economy is defined as that set of allocations that no coalition can improve upon, in the sense that no coalition could redistribute its own resources and do better than in the allocation proposed. A classic result, originally due to Edgeworth, shows that, if the set of agents becomes large – that is, we consider a sequence of economies with each economy having more agents than the previous one – then the core shrinks to the competitive equilibria. A standard objection to this is that it is highly implausible that all coalitions should form. The question then is: what may happen if agents communicate only with a certain probability, and only coalitions of agents who are closely linked can form? Suppose that there are *n* agents and that they can communicate with a fixed probability, so that $p_n(ij) = p_n$. Furthermore, for reasons which are obvious, assume that the probability does not decrease too fast with *n*, that is, $p_n \geq \frac{1}{\sqrt{n}}$. Now we know that it suffices to have large coalitions form for Edgeworth's result to hold. The question that remains is: if *n* becomes large, will the large coalitions be able to form if we place some restriction on how closely the agents have to be linked to form a coalition? Suppose that we allow coalitions to form only if the maximal distance between two agents (the *diameter* of the sub-graph defined by the coalition) is less than or equal to 2. Now, since the links are drawn at random, the set of allocations is itself a random variable. What we can hope for is that the probability that the core of such a random economy will be different from the set of competitive equilibria converges to zero. As should be clear, it is enough, given our assumptions, to use Bollobas's result directly to obtain this result.

In general, the property of increasing connectivity is of interest, as noted, in economic models, since the connectivity determines how fast information or a ‘technological shock’ diffuses and how quickly an epidemic of opinion or behaviour will occur. It is important to note that the result just evoked depends crucially on the fact that the actors in the network are linked with uniform probability or, slightly more

generally, that the pair of agents with the lowest probability of being linked should still be above the lower bound mentioned.

Emerging graphs

The dynamic evolution of the state of the individuals linked in a graph-like structure is particularly interesting since the stable configurations of states, if there are any, will depend on the graph in question, and some of the results from other disciplines (see Weisbuch, 1990) can be evoked in the context of economic models.

This ‘small world’ problem discussed above is related to, but rather different from, that studied by Watts (2000), since he looked at networks which evolve as the links are modified stochastically. In other words, he does not consider a realization of a graph drawn from a family of graphs with given probability but, rather, starts with a given graph and shows how structure and, in particular, connectivity may emerge as links are replaced by other links. What he examines is a situation in which agents have a fixed number of links with others, that is, the *vertex degree* of the graph is a constant. For example, they might all be situated on a circle and be linked only with their immediate neighbours. Then one of the existing links is drawn at random and replaced with a new link. This may be to any agent, in particular one to whom the distance was great in the original graph. Adding such links drastically increases the connectivity of the graph. This procedure is repeated and what emerges is a typically clustered structure in which closely linked individuals in a small group are linked to other groups through one or two links. In a pure random graph of the Erdos-Renyi (E-R) (1960) type, distances are short, as we have seen, but there is almost no clustering. However, in the sort of small world graphs studied by Watts the situation is different: distances are still short but there is a great deal of clustering. Here the few long links between different groups keep the distance down but most of the interaction is within small groups. This sort of clustering is observed in many empirical situations in economics. In this case the pure random E-R graph is used as a benchmark, but the analysis is self-contained and does not exploit results developed in stochastic graph theory itself.

An example of the direct use of stochastic graph theory concerns the *degree distribution* which, in the case where the set of agents is finite, just specifies the proportion of agents who have each vertex degree k, and which we can denote by $P(k)$. In the E-R case this distribution is Poisson, but more general degree distributions have been considered by Newman, Strogatz and Watts (2001), for example. There are two large classes of stochastic graphs. The first is composed of those which have a degree distribution which peaks at a particular k and then declines exponentially for large k. The classic E-R model and the Watts ‘small world’ model fall into this category. They are relatively homogeneous. The other class is characterized by a degree distribution which decays as a power law, that is,

$$P(k) = k^{-\gamma}.$$

In this class the probability that a node has a very large number of links is much higher than for the

previous class. The graphs in this class are referred to as ‘scale free’ networks.

A first reason for being interested in this division is spelled out by Albert, Jeong and Barabasi (2000). They show that scale-free networks are very robust to error but very vulnerable to attack. The reason for this is simple. Since in such networks the majority of nodes have a very small number of links, random failure of nodes will have rather little impact. The measure of the impact of the failure of nodes is taken here to be the change in the *diameter* of the graph. However, in E-R graphs, since all nodes have essentially the same number of links, each node contributes equally to the diameter. Thus removing any node will have the same impact. The diameter of the remaining graph will increase progressively and faster than if the same thing is done in the scale-free class. The vulnerability to attack is now clearly in inverse relation to the robustness to errors. Since in the scale-free networks there are nodes with a very large number of links which, as a result, contribute much to the connectivity of the network, an ‘enemy’ will simply choose one of these nodes as his target and have much more effect than he would on an E-R network. As an illustration, Albert, Jeong and Barabasi (2000) show that the removal at random of 2.5 per cent of the nodes in the World Wide Web, whose degree distribution is well fitted by a power law (see Faloutsos, Faloutsos and Faloutsos, 1999), has no significant effect on its diameter. However, the removal of 2.5 per cent of the most connected nodes in an E-R graph increases the diameter by a factor of 6.

The *degree distribution* is of interest for other reasons. It permits one to work out the average number of neighbours at *distance m* from the agents in a graph. This number can be calculated for each *m* and, if this sequence converges, there cannot be a *giant component* and all the agents will belong to small connected components in the graph. If this number diverges, however, there will always be a giant component containing many agents and smaller connected sets. There is a crucial level of the ratio of the average number of neighbours at distance 1 to the average number at distance 2 which determines which of these situations applies. This value will depend on the *degree distribution*. Such threshold values play an important role in Erdos and Renyi's work. Newman's interest has tended to focus on networks of co-authors and citations, and the obvious interest here is the extent to which there is a large disciplinary, or interdisciplinary, network, and the extent to which there are fragmented self-referential networks. It is clear, however, that the same ideas could be applied to trading groups or to firms and sectors.

From the empirical evidence one can form an estimation of the degree distribution in a particular network, and one can then examine the distribution of different sized components and pose the question as to the relation between the two and its consistency with the asymptotic predictions.

The degree distribution has been used in another context by Galeotti et al. (2006). They study the results for games played on networks and where payoffs depend on the players' own actions and those of their neighbours, and examine the influence of three features (whether the games involve strategic substitutes or complements, negative or positive externalities, and incomplete or complete information). It is the latter aspect that is of interest here. They say that an agent has *incomplete information* if a player knows only his own degree and the degree distribution, and *complete information* if he knows the degree of every player. Under incomplete information they show, for example, in games with positive, (negative) externalities, expected payoffs from a game are increasing (decreasing) in degree. This is in contrast with earlier results of Bramoullé and Kranton (2007), where in a public goods game players with higher degrees earn worse payoffs. The key difference from the result just mentioned is that the implicit assumption in their paper is that there is *complete information*.

The assumptions underlying the model developed by Galeotti et al. (2006) are that the players all believe that the graph in which they are situated is drawn from a family of networks that has two properties:

1. (i) the probability of any node having *k* links is *P*(*k)*; and
2. (ii) the degrees $k_i(g)$ and $k_j(g)$ of any two players *i* and *j* are stochastically independent.

This is interesting since it harks back to the 'configuration' model described in Bollobas (1985). (Ioannides, 2006, describes the case where the degrees of neighbouring agents are no longer independent, and develops the theory of Markov random graphs, which allow for dependence between neighbouring nodes.)
Another idea developed by Galleoti et al. (2006) is that of looking at the results for games played on different graphs, in particular where the degree distribution of one is more or less 'spread out' than the other. Slightly more formally, one looks at whether one distribution first order stochastically dominates (FOSD) the other or vice versa. In such cases for particular games, they are able to show a relationship between the payoff for a player with a given vertex degree when a particular type of game is played on a graph *g* and the payoff when the game is played on another graph g'. If *g* FOSD another graph g', then the payoff is higher under *g* than under g'. Unfortunately, no such unambiguous results are available under more general changes in the degree distribution.

Conclusion

Economists are often accused of the wholesale borrowing of results from various branches of mathematics and using them even in inappropriate contexts. The growing literature on the importance of networks in economics seems to provide a counter-example. Although many of the concepts and even the notation of mathematical graph theory are used, rather little of the formal structure has been imported, and most of that has been concentrated on stochastic graphs. Many results in the economic literature have, however, been proved *ex nihilo*, and one could argue that economists have in this modest way added to the graph theory literature.

See Also

- externalities
- network formation
- network goods (empirical studies)
- network goods (theory)

Bibliography

Albert, R., Jeong, H. and Barabasi, A.-L. 2000. Error and attack tolerance of complex networks. *Nature* 406, 378–82.

Anderlini, L. and Ianni, A. 1996. Path dependence and learning from neighbours. *Games and Economic Behavior* 13, 141–77.

Bollobas, B. 1985. *Random Graphs*. London: Academic Press.

Bramoullé, Y. and Kranton, R. 2007. Public goods in networks. *Journal of Economic Theory* 135, 478–94.

Corominas Bosch, M. 2004. Bargaining in a network of buyers and sellers. *Journal of Economic Theory* 115, 35–77.

Durlauf, S.N. 1990. Locally interacting systems, coordination failure, and the behavior of aggregate activity. Working paper, Dept of Economics, Stanford University.

Durlauf, S.N. 2004. Neighborhood effects. In *Handbook of Regional and Urban Economics*, ed. J.V. Henderson and J.F. Thisse. Elsevier: North-Holland.

Erdos, P. and Renyi, A. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 17–60.

Evstigneev, I.V. and Taksar, M. 1995. Stochastic equilibria on graphs. *Journal of Mathematical Economics* 24, 383–406.

Faloutsos, M., Faloutsos, P. and Faloutsos, C. 1999. On power-law relationships of the internet topology, ACM SIGCOMM '99. *Computers and Communication Review* 29, 251–63.

Frank, H. 1967. Vulnerability of communication networks. *IEEE Transactions on Communications* 15, 778–9.

Galeotti, A., Goyal, S., Jackson, M., Vega-Redondo, F. and Yariv, L. 2006. Network games. Mimeo, University of Essex and California Institute of Technology.

Goyal, S. 2006. Learning in games. In *Group Formation in Economics, Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Ioannides, Y.M. 1990. Trading uncertainty and market form. *International Economic Review* 31, 619–38.

Ioannides, Y.M. 1997. Evolution of trading structures. In *The Economy as an Evolving Complex System II*, ed. W.B. Arthur, S.N. Durlauf and D.A. Lane. Reading, MA: Addison Wesley.

Ioannides, Y.M. 2006. Random graphs and social networks. Discussion Paper No. 518, Department of

Economics, Tufts University.

Jackson, M. 2004. A survey of models of network formation: stability and efficiency. In *Group Formation in Economics; Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Kirman, A. 1983. Communication in markets: a suggested approach. *Economics Letters* 12, 101–8.

Kirman, A.P., Oddou, C. and Weber, S. 1986. Stochastic communication and coalition formation. *Econometrica* 54, 129–38.

Newman, M., Strogatz, S.H. and Watts, D.J. 2001. Random graphs with arbitrary degree distributions and their applications. *Physical Review E* 64, 026118.

Watts, D. 2000. *Small Worlds*. Princeton, NJ: Princeton University Press.

Weisbuch, G. 1990. *Complex System Dynamics*. Redwood City, CA: Addison-Wesley.

How to cite this article

Kirman, A. P. "graph theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000098> doi:10.1057/9780230226203.0666

The New Palgrave Dictionary of Economics Online

graphical games

Michael Kearns

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Graphical games and related models provide network or graph-theoretic means of succinctly representing strategic interaction among a large population of players. Such models can often have significant algorithmic benefits, as in the NashProp algorithm for computing equilibria. In addition, several studies have established relationships between the topological structure of the underlying network and properties of various outcomes. These include a close relationship between the correlated equilibria of a graphical game and Markov network models for their representation, results establishing when evolutionary stable strategies are preserved in a network setting, and a precise combinatorial characterization of wealth variation in a simple bipartite exchange economy.

Keywords

computation of equilibria; correlated equilibrium; dynamic programming; evolutionary game theory; evolutionary stable strategies; exchange economies; graphical economics; graphical games; network structure

Article

Graphical games are a general parametric model for multi-player games that is most appropriate for settings in which not all players directly influence the payoffs of all others, but rather there is some notion of ‘locality’ to the direct strategic interactions. These interactions are represented as an undirected graph or network, where we assume that each player is identified with a vertex, and that the payoff of a given player is a function of only his or her own action and those of his or her immediate neighbours in the network. Specification of a graphical game thus consists of the graph or network, along with the local payoff function for each player. Graphical games offer a number of representational and algorithmic advantages over the normal form, have permitted the development of a theory relating network topology to equilibrium properties, and have played a central role in recent results on the

computational complexity of computing Nash equilibria. They have also been generalized to exchange economies, evolutionary game theory, and other strategic settings.

Definitions

A graphical game begins with an undirected graph or network $G=(V,E)$, where V is the set of players or vertices, and E is a set of edges or unordered pairs of vertices/players. The assumed semantics of this graph are that the payoffs of players are determined only by their local neighbourhoods. More precisely, if we define the neighbour set of a player u as $N(u)=\{v: (u,v)$ is in $E\}$, the payoff of u is assumed to be a function not of the joint action of the entire population of players, but only the actions of u itself and the players in $N(u)$. Complete specification of a graphical game thus consists of the graph G, and the local payoff functions for each player. Note that at equilibrium, it remains the case that the strategy of a player may be indirectly influenced by players arbitrarily distant in the network; it is simply that such influences are effected by the propagation of the local, direct payoff influences.
In the case that G is the complete network, in which all pairs of vertices have an edge between them, the graphical game simply reverts to the multi-player normal form. However, in the interesting cases the graph may exhibit considerable asymmetry and structure, and also be much more succinct than the normal form. For instance, if
$|N(u)| <= d$ for all players u, then the total number of parameters of the graphical game grows exponentially only in the degree bound d, as opposed to exponentially in n for the normal form. Thus when d is much smaller than n (a reasonable expectation in a large-population game with only local interactions), the graphical game representation is exponentially more parsimonious than the normal form. Qualitatively, one can think of graphical games as a good model for games in which there may be many players, but each player may be directly and strongly influenced by only a small number of others. Graphical games should be contrasted with other parametric models such as congestion and potential games, in which each player has global influence, but often of a highly specific and weak form. They can also be viewed as a natural generalization of more specific network-based games studied in the game theory and economics literature (Jackson, 2007).

Computational properties

In addition to the aforementioned potential for representational parsimony, graphical games permit a family of natural and sometimes provably efficient algorithms for the computation of Nash and other equilibria. It should be emphasized here that by ‘efficient’ we mean an algorithm whose running time is a ‘slowly’ growing function of *the number of parameters of the graphical game representation* (which may be considerably more challenging than a slowly growing function of the number of parameters in the normal form representation, which may be much larger). As is standard in computer science, ‘slowly’ growing typically means a polynomial function (ideally of low degree).
For instance, in the special case that the graph structure is a tree (or can be modified to a tree via a small number of standard topological operations involving, for instance, the merging of vertices), there is an algorithm running in time polynomial in the number of parameters that computes approximations to (all) Nash equilibria of the given graphical game (Kearns, Littman and Singh, 2001.) This algorithm is based

on dynamic programming, and is decentralized in the sense that communication need take place only between neighbouring vertices in the network. In even more restrictive topologies, efficient algorithms for computing exact Nash equilibria are known (Elkind, Goldberg and Goldberg, 2006). For non-tree topologies, a generalization of this algorithm known as NashProp (Ortiz and Kearns, 2002) has been developed that is provably convergent, but has weaker guarantees of computational efficiency. Provably efficient algorithms have also been developed for computing correlated equilibria (again in the sense of the computation time being polynomial in the number of parameters) for general graphical games (Papadimitriou, 2005; see also Kakade et al., 2003). These algorithmic results are in sharp contrast to the status of computing equilibria for games represented in the normal form, where the results are either negative or remain unresolved. Graphical games have also proven valuable in establishing computational barriers to computing Nash equilibria efficiently, and certain classes of graphical games have been shown to be just as hard as the normal form in this regard (Daskalakis and Papadimitriou, 2005; 2006; Daskalakis, Goldberg and Papadimitriou, 2006; Schoenebeck and Vadhan, 2006).

Extensions of the model

Since the introduction of graphical games, a number of related models have been introduced and studied. In each case, the model again begins with an undirected network in which the edges represent the pairs of participants that are permitted to interact directly in some strategic or economic setting.

For instance, the model known as *graphical economics* (Kakade, Kearns and Ortiz, 2004) provides a network generalization of the classical exchange economies studied by Arrow and Debreu and others. As in the classical models, each consumer has an initial endowment over k commodities, and a subjective utility function describing his or her preferences over bundles of commodities. However, unlike the classical model, not all pairs of consumers may engage in trade. Instead, each consumer or vertex u may only trade with his or her neighbours $N(u)$, and there is no resale permitted. Equilibria in prices and consumption plans can still be shown to always exist, but now the equilibrium prices may need to be *local*, in the sense that two consumers may charge different prices per unit for the same commodity, and these prices may depend strongly on the network topology. This introduces variation in equilibrium wealth dependent on a consumer's position in the overall network (see below). As with graphical games, the graphical economics model permits efficient computation of equilibria under certain topological restrictions on the network.

More recently, a network version of evolutionary game theory (EGT) has also been examined (Kearns and Suri, 2006). In classical EGT, there are random encounters between all pairs of organisms; in the network generalization, such encounters are restricted to the edges of an undirected network. Thus the evolutionary fitness of an organism represented by vertex u is once again determined only by the strategies of its neighbours $N(u)$. More than one reasonable generalization of the evolutionary stable strategy (ESS) of classical EGT is possible in the network setting.

As with graphical games, both graphical economics and the network EGT model revert to their classical counterparts in the special case of the complete graph over all participants, and thus represent strict generalizations, but which are most interesting in cases where the underlying graph has some non-trivial structure.

Network structure and equilibrium properties

Aside from the algorithmic properties discussed above, one of the most interesting aspects of the various models under discussion is their ability to permit the study of how equilibrium properties are influenced by the structure or topology of the underlying network, and indeed there is a growing body of results in this direction.

In the case of graphical games, a tight connection can be drawn between the topology of the underlying graph *G* of the game and the structure of any minimal (in a certain natural technical sense) correlated equilibrium (CE), regardless of the details of the local payoff functions. Among other consequences, this result implies that CE in graphical games can be implemented using only local, distributed sources of randomization throughout the network in order effect the needed coordination, rather than the centralized randomization of classical CE. The result also provides broad conditions under which the play of two 'distant' players in the network may be conditionally independent – for instance, at CE, the play of vertices *u* and *v* is always conditionally independent given the pure strategies of any vertex cutset between them (Kakade et al., 2003).

In the graphical economics model, for certain simple cases one can give precise relationships between equilibrium wealth, price variation and network structure. For instance, in the case of an arbitrary bipartite network for a simple two-commodity buyer-seller economy with symmetric endowments and utilities (thus deliberately rendering network position the only source of asymmetry between consumers), there is no price or wealth variation at equilibrium if and only if the underlying network has a perfect matching sub-graph between buyers and sellers. More generally, a purely structural property of the network characterizes the ratio between the greatest and least consumer wealth at equilibrium. These structural results have been applied to analyse price and wealth variation in certain probabilistic models of buyer–seller network formation. For instance, it has been shown that whereas truly random networks with a certain minimum number of edges generally exhibit no variation in prices or wealth, those generated by recent models of social network formation such as preferential attachment lead to a power-law distribution of wealth (Kakade et al., 2004).

In the networked EGT setting, it has been proven that even networks with rather sparse connectivity (in which each organism directly interacts with only a small fraction of the total population), but in which the connections are formed randomly, classical ESS are always preserved, even if the initial locations of the invading population are arbitrary. Alternatively, if the network is arbitrary but the initial locations of the invading population are selected randomly, classical ESS are again preserved (Kearns and Suri, 2006). Related network models include those of Blume (1995) and Ellison (1993).

See Also

- computation of general equilibria
- learning and evolution in games: ESS
- mathematics of networks
- stochastic adaptive dynamics

Bibliography

Blume, L.E. 1995. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior* 11, 111–45.

Daskalakis, C. and Papadimitriou, C. 2005. Computing pure Nash equilibria via Markov random fields. In *Proceedings of the 6th ACM Conference on Electronic Commerce*.

Daskalakis, C. and Papadimitriou, C. 2006. The complexity of games on highly regular graphs. In the 13th Annual European Symposium on Algorithms.

Daskalakis, C., Goldberg, P.W. and Papadimitriou, C.H. 2006. The complexity of computing a Nash equilibrium. In *Proceedings of the 38th ACM Symposium on Theory of Computing*.

Elkind, E., Goldberg, L.A. and Goldberg, P.W. 2006. Nash equilibria in graphical games on trees revisited. In *Proceedings of the 7th ACM Conference on Electronic Commerce*.

Ellison, G. 1993. Learning, local interaction, and coordination. *Econometrica* 61, 1047–71.

Jackson, M. 2007. The study of social networks ineconomics. In *The Missing Links: Formation and Decay of Economic Networks*, ed. J. Podolny, and J.E. Rauch. Russell Sage Foundation.

Kakade, S., Kearns, M. and Ortiz, L. 2004. Graphical economics. In *Proceedings of the 17th Conference on Computational Learning Theory*.

Kakade, S., Kearns, M., Langford, J. and Ortiz, L. 2003. Correlated equilibria in graphical games. In *Proceedings of the 4th ACM Conference on Electronic Commerce*.

Kakade, S., Kearns, M., Ortiz, L., Pemantle, R. and Suri, S. 2004. Economic properties of social networks. In *Neural Information Processing Systems 18*.

Kearns, M. and Suri, S. 2006. Networks preserving evolutionary stability and the power of randomization. In *Proceedings of the 7th ACM Conference on Electronic Commerce*.

Kearns, M., Littman, M. and Singh, S. 2001. Graphical models for game theory. In *Proceedings of the 17th Conference Uncertainty in Artificial Intelligence*.

Ortiz, L. and Kearns, M. 2002. Nash propagation for loopy graphical games. In *Neural Information Systems Processing 16*.

Papadimitriou, C. 2005. Computing correlated equilibria in multi-player games. In *Proceedings of the*

37th ACM Symposium on the Theory of Computing.

Schoenebeck, G. and Vadhan, S. 2006. The computational complexity of Nash equilibria in concisely represented games. In *Proceedings of the 7th ACM Conference on Electronic Commerce*.

How to cite this article

Kearns, Michael. "graphical games." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000588> doi:10.1057/9780230226203.0667

The New Palgrave Dictionary of Economics Online

Graunt, John (1620 – 1674)

R.M. Smith

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bills of Mortality; fertility; Graunt, J.; life tables; mortality; Petty, W.; population growth

Article

Graunt was born on 24 April 1620 in Hampshire and died on 18 April 1674 in London. At the age of 16 he was apprenticed to his father as a haberdasher of small wares, but remarkably little is known of his life before he published his *Natural and Political Observations Made upon the Bills of Mortality* in 1662. Graunt had formed a friendship with William Petty, who came from a social and economic background similar to his own and who may have drawn Graunt's attention to the data in the London Bills of Mortality. Six months after the publication of the *Natural and Political Observations* Graunt was made a fellow of the Royal Society, in whose foundation Petty had been greatly involved. The publication of this volume had an immediate impact; a second edition was published the same year and two others in 1665. Graunt subsequently fell into disgrace following his conversion to Catholicism and he died in poverty despite generous help from Petty. The latter's own work, especially that on urban growth, owed much to methodologies initiated by Graunt.

It has been claimed that the *Natural and Political Observations* ‘created the subject of demography’ (Glass, 1963) as it involved the first truly analytic study of births and deaths within a population precisely situated in space and time. To do this Graunt employed the Bills of Mortality (Greenwood, 1948) and the records of christenings in 17th-century London in order to investigate mortality and population growth in the city. The study's most outstanding qualities are revealed by the search for regularities and configurations in mortality and fertility along with a critical and very insightful appreciation of the quality of the data. He was greatly concerned to establish mortality rates by age and through this interest he came remarkably close to constructing the first formal life table. Using the cause of death evidence in the Bills of Mortality, Graunt without any information whatsoever on ages proceeded to estimate the extent of mortality in infancy and childhood by selecting those causes of death which he guessed would only affect children ‘under four or five years old’. To these he added half of the deaths from smallpox and measles which he thought would fall upon children under six, along

with slightly less than one third of the plague victims. From these assumptions he derived an estimate suggesting that 36 per cent of all deaths occurred to children under six years. Similar principles were employed to estimate that seven per cent of deaths through ‘ageing’ could be attributed to those over 70 years. He then proceeded to take these not implausible estimates of mortality at young and old ages to construct an elementary life table which unfortunately was flawed by the unrealistic assumption that above age six the deaths in each specified age period amounted to about three-eighths of the survivors at the beginning of the period. Nonetheless, what Graunt had evolved was an outstanding innovation and was very soon taken up and developed by others. Graunt's interests in the *Natural and Political Observations* were innovative in other respects; he attempted to calculate the size and rate of growth of 17th-century London, the incidence of plague mortality (Sutherland, 1963; 1972), and sex ratios at birth and levels of maternal mortality, and he made some highly believable estimates of the levels of immigration to London that were needed to sustain the city's remarkable growth in the 17th century.

Selected works

1662. *Natural and Political Observations … upon the Bills of Mortality*. There was a further edition later in 1662, and the plague of 1665 called forth a third edition in the early summer and a fourth edition in November (printed in Oxford) of that year.

1938. *Natural and Political Observations*, ed. Walter F. Wilcox, Baltimore: Johns Hopkins University Press.

1973. Natural and Political observations made upon the Bills of Mortality. In *The Earliest Classics*, ed. P. Laslett. Farnborough, Hampshire: Gregg International Publishers.

Bibliography

Glass, D.V. 1963. John Graunt and his ‘Natural and political observations’. *Royal Society of London Notes and Records* 19, 63–100.

Greenwood, M. 1948. *Medical Statistics from Graunt to Farr*. Cambridge: Cambridge University Press.

Sutherland, I. 1963. John Graunt: a tercentenary tribute. *Proceedings of the Royal Society*, Series A 126, 537–56.

Sutherland, I. 1972. When was the Great Plague? In *Population and Social Change*, ed. D.V. Glass and R. Revelle. London: Edward Arnold.

How to cite this article

Smith, R. M. "Graunt, John (1620–1674)." The New Palgrave Dictionary of Economics. Second Edition.

Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000100> doi:10.1057/9780230226203.0668

The New Palgrave Dictionary of Economics Online

gravity equation

Robert C. Feenstra

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The gravity equation explains the amount of trade between countries based on their economic sizes and the distance between them. While it has been in use since the 1960s, its theoretical foundation has been known for a much shorter period, and recent years have seen an large amount of research on its derivation and estimation. We review the theoretical and empirical literature on the gravity equation. In addition to explaining the amount of trade, this equation has been applied to foreign direct investment, the volatility of prices, and the impact of currency unions and free trade areas.

Keywords

border effects; constant-elasticity-of-substitution (CES) preferences; currency unions; distance; elasticity of substitution; estimation; euro; fixed effects; foreign direct investment; gravity equation; international trade; monopolistic competition; Poisson distribution; product differentiation; selection equation; specialization; transport costs; World Trade Organization

Article

The importance of countries' economic size in explaining trade patterns was recognized in an equation first proposed by Tinbergen (1962). Tinbergen proposed that the volume of trade between countries would be similar to the force of gravity between objects. Suppose that two objects each have mass M_1 and M_2, and they are located distance d apart. Then, according to Newton's universal law of gravitation, the force of gravity F_g between these two objects is:

$$F_g = G = \frac{M_1 M_2}{d^2},$$

where G is the gravitational constant, M_1 and M_2 the mass of the two objects, and d the distance between them. The larger each of the objects are, or the closer that they are to each other, then the greater is the force of gravity between them.

The equation proposed by Tinbergen to explain trade between countries is very similar to Newton's law of gravity, except that instead of the mass of two objects we use the gross domestic product (GDP) of two countries, and instead of predicting the force of gravity we are predicting the amount of trade between them. So the gravity equation in trade is:

$$\text{Trade} = A\frac{\text{GDP}_1\text{GDP}_2}{d^{\gamma}}, \tag{1}$$

where trade is the amount of trade (that is, imports, exports, or their sum total) between two countries, GDP_1 and GDP_2 their gross domestic products, d the distance between them, and A a constant. We use the exponent γ on d^{γ} rather than d^2 as in Newton's law of gravity, because we are not sure of the economic relationship between distance and trade.

While Tinbergen (1962) showed that the gravity equation worked well empirically, it was some years before the theoretical foundation of this equation was known. The earliest contribution is Anderson (1979), followed by Bergstrand (1985; 1989) and Helpman (1987). All these authors suppose that *countries produce different goods*. That result might be due to the underlying assumption of monopolistic competition, in which case every firm produces a different product variety from every other firm. It follows that firms in different countries also produce different product varieties (which we label as different goods). But that is not the only model where countries produce different goods. That result also applies in a perfectly competitive model where there are more goods than factors, and some differences in factor prices or technologies across countries (Bhagwati, 1972; Davis, 1995).

Along with country specialization in different goods, let us add the assumption that the utility function of the representative consumer is homothetic, so that the income elasticities of demand are unity (that is, the share of demand spent on each good does not vary with income). In that case, we can follow Helpman (1987) and derive a simple gravity equation. Let i, j=1, …, C denotes countries, and let k=1, …, N denotes goods. Let V_k^i denote country i's production of good k. Assume that prices are the same across all countries due to free trade and no transport costs, and normalize the prices at unity, so V_k^i actually measures the *value* of production. Total GDP in each country is $Y^i = \Sigma_{k=1}^{N} V_k^i$, and world GDP is $Y^W = \Sigma_{i=1}^{C} Y^i$.

Let s^j denote country j's share of world expenditure. If we assume that trade is balanced in each country, then s^j also denotes country j's share of world GDP, so that $s^j = Y^j / Y^W$. Then with all countries

producing different goods, and with identical and homothetic demand across countries, the *exports from country i to county j of product k* are given by:

$$X_k^{ij} = s^j V_k^i. \tag{2}$$

Summing over all products *k*, we obtain:

$$X^{ij} = \sum_k X_k^{ij} = s^j \sum_k V_k^i = s^j Y^i = \frac{Y^j Y^i}{Y^W}. \tag{3}$$

So this simple model gives us a gravity equation like (1), where $A = 1 / Y^W$. The gravity equation in (3) omits a distance term, however. Notice that, if instead we solve for X^{ji} then we get $Y^j Y^i / Y^W$, which is exactly the same as (3). So in this very simple model, bilateral exports equal bilateral imports (that is, trade is balanced between every pair of countries).

The key limitation of the gravity equation (3) is that it assumes no transport costs. Since transport costs will depend on the distance between countries, by introducing them we will also introduce a distance term into the gravity equation. But allowing for transport costs means that the prices of goods will differ across countries, since a distant country will have higher c.i.f. prices (that is, including cost, insurance and freight charges). In that case we need to re-derive the gravity equation while allowing for different prices. To do so requires us to introduce an explicit demand structure into the model. A commonly used demand within the monopolistic competition model arises from constant-elasticity-of-substitution (CES) preferences. In that case the demand for a good *i* exported by that country-to-country *j* is:

$$c^{ij} = (p^{ij} / P^j)^{-\sigma} (Y^j / P^j), \tag{4}$$

where $\sigma > 1$ is the elasticity of substitution, p^{ij} the c.i.f. price of the good exported from *i* to *j*, and P^j refers to country *j*'s overall price index, defined as:

$$p^j = \left(\sum_{i=1}^{C} N^i (p^{ij})^{1-\sigma}\right)^{1/(1-\sigma)}, \tag{5}$$

where N^i is the number of goods produced in country *i* and exported to *j*.
To obtain total exports from country *i* to *j*, we add up over all N^i goods produced in country *i* and exported to *j*, obtaining $X^{ij} = N^i p^{ij} c^{ij}$, or from (4) and (5),

$$X^{ij} = N^i Y^j \left(\frac{p^{ij}}{P^j}\right)^{1-\sigma}. \tag{6}$$

Equation (6) is not quite a gravity equation because it does not have the GDP of the exporting country *i*. There are several methods that can be used to introduce that variable into (6).
One method is to solve for the number of products N^i in the exporting country, by using a free entry condition from the monopolistic competition model. In the simplest monopolistic competition model, with one factor of production and CES preferences, the number of products will be proportional to GDP, so that $N^i \propto Y^i$. We can also introduce transport costs into (6) by writing the c.i.f. prices as:

$$p^{ij} = T^{ij} p^i, \tag{7}$$

where p^i is the f.o.b. (free on board) price from country *i* and $T^{ij} \geq 1$ is the transport costs from country *i* to *j*. By substituting these relations into (6), and taking natural logs, we obtain:

$$\ln(X^{ij}) = a + \ln(Y^i Y^j) + (1-\sigma)\ln T^{ij} + (1-\sigma)\ln(p^i / P^j). \tag{8}$$

Baier and Bergstrand (2001) estimate a gravity equation like (8), using price data for the term

$\ln(p^i / P^j)$, and splitting the term $\ln T^{ij}$ into tariffs and transport costs. They use data for Organisation for Economic Co-operation and Development (OECD) countries, and take differences between the averages in 1958–60 and 1986–88 to express all the variables as changes over time. The resulting linear regression has an R^2 of 0.40, so they are able to explain nearly one-half of the growth in exports between OECD countries.

A second method to convert (6) into a gravity equation is to solve for the f.o.b. prices p^i in each exporting country. Anderson and van Wincoop (2003) argue that an implicit solution for the f.o.b. prices is:

$$\tilde{p}^i \equiv \left(s^i / N^i\right)^{1/(1-\sigma)} / \tilde{P}^i, \tag{9}$$

in which case the price indexes are solved as:

$$(\tilde{P}^j)^{1-\sigma} = \sum_{i=1}^{C} s^i \left(T^{ij} / \tilde{P}^i\right)^{1-\sigma}. \tag{10}$$

If we substitute (9) into (6), we obtain the gravity equation:

$$X^{ij} = s^i Y^j \left(\frac{T^{ij}}{\tilde{P}^i \tilde{P}^j}\right)^{1-\sigma} = \left(\frac{Y^i Y^j}{Y^W}\right)\left(\frac{T^{ij}}{\tilde{P}^i \tilde{P}^j}\right)^{1-\sigma}. \tag{11}$$

So we obtain a gravity equation like (1), where the transport costs T^{ij} can depend on distance d^{ij}. The 'constant' A in the gravity equation (11) is $A = (\tilde{P}^i \tilde{P}^j)^{\sigma-1} / Y^W$, which depends on the price indexes of the exporting and importing country. So the important lesson from Anderson and van Wincoop is that the 'constant' term must vary with the importer and exporter (and if multiple years are used in the data, then it should also vary with time).

Anderson and van Wincoop apply their version of the gravity equation to Canada–US trade. McCallum (1995) originally estimated a gravity equation for trade between ten Canadian provinces, and trade

between those provinces and 30 US states, using 1988 data. An updated version using 1993 data, which also includes trade between US states, is as follows (with standard errors in parentheses):

$$\ln X^{ij} = \underset{(0.11)}{2.75\ \text{Canada}} + \underset{(0.05)}{0.40\ \text{US}} + \underset{(0.02)}{1.13}\ln(Y^i) + \underset{(0.02)}{0.97}\ln(Y^j) - \underset{(0.03)}{1.11}\ln(d^{ij}), \tag{12}$$

$$R^2 = 0.85,\ \ N = 1511.$$

The first variable appearing on the right of (12), ‘Canada’, is an indicator variable equal to 1 for trade between Canadian provinces, and zero otherwise. The second variable, ‘US’, is an indicator variable equal to 1 for trade between US states, and zero otherwise. The remaining variables are GDPs of the exporting province or state, both of which have coefficients close to unity, and distance, with a coefficient close to minus 1.

Since the variables in (12) are in natural logs, we can take the exponential of the indicator coefficients to obtain $e^{2.75}$=16 and $e^{0.4}$=1.5. Thus, the estimates imply that cross-provincial trade within Canada is 16 times greater than cross-border trade with the United States, whereas cross-state trade within the United States is only 1.5 times greater than cross-border trade. (Using data from 1988, McCallum had found that cross-provincial trade within Canada was 22 times greater than cross-border trade.) The very large magnitude of the ‘border effect’, leading to much more trade *within* Canada than across the border, is very surprising!

Before we try to explain where this border effect might come from, we should check that the estimates in (12) are reliable. In particular, this estimate does not incorporate the price indexes $\tilde{P}^i$ and $\tilde{P}^j$ that appear in (11). If we incorporate these terms, the estimate for 1993 from Anderson and van Wincoop (2003) is:

$$\ln(X^{ij}/Y^iY^j) = -\underset{(0.08)}{1.65\ \text{Border}} - \underset{(0.03)}{0.79}\ln(d^{ij}) + \ln(\tilde{P}^i)^{\sigma-1} + \ln(\tilde{P}^j)^{\sigma-1}, \tag{13}$$

$$N = 1511.$$

Notice that Anderson and van Wincoop (2003) keep the GDPs on the left of (13), by dividing both sides of (11) by (Y^iY^j). On the right, they include an indicator variable 'Border' equal to 1 for cross-border trade, and zero otherwise, rather than the separate Canada and US indicator variables. In addition, they include the price indexes $(\tilde{P}^i)^{\sigma-1}$ and $(\tilde{P}^j)^{\sigma-1}$ on the right. Those price indexes are computed from the formula in (10), where the transport costs used in that formula are:

$$(1-\sigma)\ln T^{ij} = -1.65\,\text{Border} - 0.79\ln(d^{ij}). \tag{14}$$

Thus, we use the estimated coefficients of the border and distance from (13), substitute these into (14) and (10) to compute the price indexes $(\tilde{P}^i)^{\sigma-1}$ and $(\tilde{P}^j)^{\sigma-1}$, use these price indexes in the regression (13) to get a new coefficients of the border effect and distance, use those again in (14) and (10) to get new price indexes, and iterate on this procedure until it converges.

The estimated 'border effect' from (13) is $e^{1.65}$=5.2. So, on average, the Canada–US border leads to five times more trade within the United States and Canada than cross-border trade. This procedure does not directly give an estimate of the *separate* Canada and US border effects, but that can be obtained from an extra calculation using (10) and computing what trade within each country would be with and without the border effect. Anderson and van Wincoop (2003) conclude that trade *within* Canada is 10.5 times higher than cross-border trade due to the border effect, which is smaller than the estimate of 16 times obtained from regression (12) (or the original estimate of 22 times from McCallum, 1995). In addition, trade *within* the United States is 2.6 times higher than cross-border trade due to the border effect.

This procedure due to Anderson and van Wincoop requires custom programming to compute the price indexes from (10). Feenstra (2004, ch. 5) notes that a linear regression can be used instead, by estimating the price indexes in (13) using fixed effects, that is, using indicator variables for each exporting region and each importing region whose coefficients are the *estimated* price indexes $(\tilde{P}^i)^{\sigma-1}$ and $(\tilde{P}^j)^{\sigma-1}$. That approach gives results very similar to (13), and is easier to compute, so using fixed effects for importing and exporting regions can be considered the preferred estimation method.

The use of fixed effects in the gravity equation has now become common practice, and makes a difference. For example, Rose (2000) estimates a gravity equation across a broad sample of countries, some of which belong to a currency union. The indicator variables for members of the currency union take on a surprisingly large value, implying that a currency union increases trade between its members by three times, or 200 per cent! That result seems too large. Rose and Wincoop (2001) include fixed

effects for exporting and importing countries, and subsequent papers have also included the fixed effects while allowing them to vary over time (as needed when multiple years are included in the estimation). Baldwin (2006a; 2006b) surveys these papers, and concludes that a currency union increases trade by nearly two times, or 100 per cent. But the impact of the euro on trade is much lower – in the range of eight to 15 per cent.

The use of fixed effects does not necessarily lower the indicator variable coefficients. In another application, Rose (2004) estimates the impact of World Trade Organization (WTO) membership on trade between countries, and finds that it is surprisingly small. He added together exports and imports, using $\ln(X^{ij}+X^{ji})$ as the dependent variable (which is appropriate only in the simplest gravity model without price effects, as in (3)), and using a single set of country fixed effects. But Subramanian and Wei (2003) argue that, when the $\ln X^{ij}$ is used as the dependent variable and both exporter and importer fixed effects are included, the WTO has a substantial effect on imports, especially for the industrial countries.

Along with included fixed effects in the estimation (and allowing them to vary over time), another very important estimation issue is how ‘zero's’ are treated in the data. Notice that we use the natural log of exports as the dependent variable in the gravity equation, so when exports are zero the natural log cannot be computed. Common practice has been to omit those observations, as was done in (12) and (13). In order to incorporate the zero trade flows, Silva and Tenreyro (2006) recommend that estimation be performed as if the dependent variable had a Poisson probability distribution. The Poisson distribution is ordinarily applied to ‘count’ data, that is, the non-negative integers 0, 1, 2, 3, … But it can still be used as the estimation method for a continuous, non-negative dependent variable like X^{ij}. In that case the zero trade values can be included.

An alternative approach to incorporate zero trade values is advocated by Helpman, Melitz and Rubenstein (2007). They first model the *reason* why countries might not trade with each other, that is, because of fixed costs of exporting to another country. That model leads to a selection equation for whether countries trade or not, followed by a gravity equation when trade is positive. This two-equation system is estimated using a modified Heckman procedure (that is, one equation for whether a country has positive trade or not, and a second equation explaining the amount of trade). They argue that the estimates from the two-equation system are quite different from a single equation approach that ignores the zero trade flows.

We conclude by noting that the gravity equation is more general than we have indicated so far, in several respects. First, it can be derived even for trade in homogeneous goods (so that countries produce the same good). Eaton and Kortum (2002) derive a gravity equation in a general model with homogeneous goods and many countries, while Evenett and Keller (2002) and Feenstra, Markusen and Rose (2001) also obtain this equation in more specialized settings. Second, the gravity equation can also be used with dependent variables other than trade. For example, Eaton and Tamura (1994) and Head and Ries (2005) derive and estimate a gravity equation that uses foreign direct investment as the dependent variable. Engel and Rogers (1996) estimate a gravity equation that use the variance in prices across cities as the dependent variable, and find that crossing the Canada–US border adds as much to the volatility of prices as adding 2,500 miles between cities. These examples illustrate the many applications of the gravity equation, which will continue to be a widely used empirical tool in international economics.

See Also

- border effects
- currency unions
- international trade theory

The author would like to thank Keith Head for his very helpful comments.

Bibliography

Anderson, J.A. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69, 106–16.

Anderson, J.A. and van Wincoop, E. 2003. Gravity with gravitas: a solution to the border puzzle. *American Economic Review* 93, 170–92.

Baier, S. and Bergstrand, J.H. 2001. The growth of world trade: tariffs, transport costs, and income similarity. *Journal of International Economics* 53, 1–27.

Baldwin, R. 2006a. The euro's trade effects. Working Paper No. 594, European Central Bank.

Baldwin, R. 2006b. *In or Out: Does it Make a Difference? An Evidence Based Analysis of the Trade Effects of the Euro*. London: CEPR.

Bergstrand, J.H. 1985. The generalized gravity equation, monopolistic competition, and the factor-proportions theory in international trade. *Review of Economics and Statistics* 71, 143–53.

Bergstrand, J.H. 1989. The gravity equation in international trade: some microeconomic foundations and empirical evidence. *Review of Economics and Statistics* 67, 474–81.

Bhagwati, J.N. 1972. The Heckscher–Ohlin theorem in the multi-commodity case. *Journal of Political Economy* 80, 1052–5.

Davis, D.R. 1995. Intra-industry trade: a Heckscher–Ohlin–Ricardo approach. *Journal of International Economics* 39, 201–26.

Eaton, J. and Kortum, S. 2002. Technology, geography and trade. *Econometrica* 70, 1741–80.

Eaton, J. and Tamura, A. 1994. Bilateral and regionalism in Japanese and U.S. trade and direct foreign investment patterns. *Journal of the Japanese and International Economies* 8, 478–510.

Engel, C. and Rogers, J.H. 1996. How wide is the border? *American Economic Review* 86, 1112–25.

Evenett, S. and Keller, W. 2002. On theories explaining the success of the gravity equation. *Journal of Political Economy* 110, 281–315.

Feenstra, R. 2004. *Advanced International Trade*. Princeton: Princeton University Press.

Feenstra, R.C., Markusen, J.R. and Rose, A.K. 2001. Using the gravity equation to differentiate among alternative theories of trade. *Canadian Journal of Economics* 34, 430–47.

Head, K. and Ries, J. 2005. Judging Japan's FDI: the verdict from a dartboard model. *Journal of the Japanese and International Economies* 19, 215–32.

Helpman, E. 1987. Imperfect competition and international trade: evidence from fourteen industrial countries. *Journal of the Japanese and International Economies* 1, 62–81.

Helpman, E., Melitz, M. and Rubinstein, Y. 2007. Trading partners and trading volumes. Working Paper No. 12927. Cambridge, MA: NBER.

McCallum, J. 1995. National borders matter. *American Economic Review* 85, 615–23.

Rose, A.K. 2000. One money, one market: estimating the effect of common currencies on trade. *Economic Policy* 30, 9–45.

Rose, A.K. 2004. Do we really know that the WTO increases trade? *American Economic Review* 94, 98–114.

Rose, A.K. and van Wincoop, E. 2001. National money as a barrier to international trade: the real case for currency union. *American Economic Review* 91, 386–90.

Silva, J.S. and Tenreyro, S. 2006. The log of gravity. *Review of Economics and Statistics* 88(4), 641–658.

Subramanian, A. and Wei, S.-J. 2003. The WTO promotes trade, strongly but unevenly. Working Paper No. 10024. Cambridge, MA: NBER.

Tinbergen, J. 1962. *Shaping the World Economy*. New York: Twentieth Century Fund.

How to cite this article

Feenstra, Robert C. "gravity equation." The New Palgrave Dictionary of Economics. Second Edition.

Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000213> doi:10.1057/9780230226203.0669

The New Palgrave Dictionary of Economics Online

gravity models

Pierre-Philippe Combes

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Basic gravity models state that economic interactions between two geographically defined entities are proportional to the size of these entities and inversely related to the distance between them. They have great empirical explanatory power. The impact of distance is strong and not diminishing over time. Extended gravity models incorporate borders and contiguity effects and more sophisticated interaction cost measures. They can be theory grounded, which makes each country's location vis-à-vis the rest of the world play a role in the bilateral relationship. Various empirical approaches have been proposed to tackle the econometric issues at stake in these more sophisticated frameworks.

Keywords

borders; business and social networks; comparative advantage; distance; equity flows; fixed-effects; foreign direct investment; gravity models; heteroskedasticity bias; imperfect competition; information costs; market potential; maximum likelihood; microfoundations; migration; monopolistic competition; nonlinear estimation; patents; preference bias; proximity; trade agreements; reverse causality; selection bias; social interaction; social networks; spatial interaction; trade; trade costs; transport costs

Article

The simplest gravity model states that economic or social interactions between two geographically defined economic entities are proportional to the size of these entities and inversely related to the distance between them. The system of interactions that results from these bilateral relationships shapes the spatial organization of the global economy. Initially borrowed from the universal law of gravitation that Newton established in 1687 for heavenly bodies, this model has undergone many refinements in economics, in particular in order to better match some underlying theoretical models and data. Some of the first applications were proposed by social physicists such as Ravenstein (1885) for migration flows, and Reilly (1931) for consumers' shopping behaviour. Stewart (1947), an astronomer, suggested that the

gravity law could be applied to a very wide class of social interactions. Tinbergen (1962) initiated what continued to be the main application of gravity models, namely, the study of the determinants of trade.

The basic framework

Typically, country *i*'s exports to country *j*, F_{ij}, are modelled as a function of the distance between these countries, d_{ij}, and of their economic mass, (M_i, M_j), which is most often proxied by their GDPs. Thus a basic gravity trade model estimates parameters α, β, and δ (expected to be positive) such that

$$F_{ij} = G\frac{M_i^{\alpha} M_j^{\beta}}{d_{ij}^{\delta}}\varepsilon_{ij}, \quad (1)$$

where G is a constant and ε_{ij} an error term capturing what is left unexplained by the model. M_i can be interpreted as the supply of the good traded and M_j the demand, while d_{ij}^{δ} captures trade costs, which encompass all costs incurred in transferring goods. These costs add to the price of goods when they are not sold locally and are assumed to increase with spatial distance. Alternatively, if F_{ij} is the number of migrants from *i* to *j*, regional populations are often more relevant as measures of M_i and M_j, while d_{ij}^{δ} reflects a moving cost. Similar interpretations can be proposed for other kinds of flows.

According to gravity models, proximity is the main engine of trade, of migration or of any precisely defined social interaction between spatially distinct economic entities. This could appear as an obsolete view of the world if one believes in the ‘death of distance’, as touted by popular accounts. However, Disdier and Head's (2008) meta-analysis over 1,467 estimations of δ on trade flows indicates an average value around 0.9: halving distance increases trade by 45 per cent. These authors report even larger δ's for recent periods, which means that the distance decay effect has actually increased in recent years.

In addition to the reliable estimates of the impact of distance they lead to, the success of gravity models is due to their great explanatory power for flows, and this holds true whatever the geographical scale (countries, large or small regions), the period of study or the goods considered. This makes gravity one of the most stable relationships in economics and a useful predictive tool. It can be also used for obtaining predictors of variables used in a second stage to explain income, productivity, or growth dispersion across space (for example, Redding and Venables, 2004).

A wide range of applications

Gravity models are applicable to many other endogenous quantities in addition to trade flows. We have

mentioned migration. Urban planners use them for traffic forecasting. There is also evidence of gravity effects in explaining foreign direct investment. More novel are the estimations on equity flows. In that case, the explanatory power of the model is as great as it is for trade, and halving distance increases flows by 25 per cent (Portes and Rey, 2005). In a supposedly financially integrated world, this is high. The impact of distance is still significant, although around six times lower, for flows of ideas, identified for instance as the citations of patents in a country that have been taken out in another country (Peri, 2005).

McCallum (1995) shows that borders also matter a lot, as well as distance. Trade between two Canadian regions was found to be 20 times larger than trade between a Canadian region and a US state of the same size and at the same distance, even if this drops to around six times once some statistical problems are removed. Discrete gaps in trade flows are also systematically observed between areas that are contiguous, relative to those that are not. Such effects suggest that the impact of distance on trade is not log-linear, or even smooth, and that a wider class of spatial proximity measures is necessary to fully encompass the effects of space.

Proximity matters for spatial interactions because it proxies for many of their determinants. Transport costs are the most obvious one. Clearly, the energy consumption or the time spent in transport, which results in opportunity costs, increase with distance. For migration, moving costs (both monetary and psychological) increase with distance and jump upward once borders are crossed. International trade flows are clearly reduced by trade policies, but trade agreements are typically first established between nearby countries. More original is the idea that preferences and tastes may be biased towards local goods, which may result from better information about them. Information costs are also critical for firms that want to access distant markets. They need to find local retailers, and then to work with them (which also clearly matters for foreign direct investment). Prior to this, they must assess market size, find out about local tastes and possibly adapt their products and marketing. All of this may explain not only the impact of distance but also the role of additional proximity measures. Moreover, on top of a possible composition effect (relatively more goods that are less easy to trade are traded), this suggests another explanation for the recently increasing impact of distance. While transport costs and trade barriers have clearly strongly diminished, preference biases or information costs may have risen, possibly due the increasing number and complexity of the goods available.

Simple gravity models have been extended to control for economic factors that would directly capture trade costs. Large data-sets on trade agreements or on transport costs are now available. The fraction of the population sharing the same language is used to capture closeness of tastes or reduced information costs. More generally, the positive role of business and social networks (among migrants from the same country or among firms belonging to the same business groups) on trade is currently being studied. These measures reduce the impact of distance, borders, and contiguity, without completely eliminating them. Naturally, not all the reasons why space matters for interactions have yet been identified. Extensive references on trade costs proxies and the various applications of gravity models can be found in Anderson and van Wincoop (2004) and in Combes, Mayer and Thisse (2008, ch. 5).

Extended frameworks and the role of the rest of the world

The bilateral nature of the gravity model is somewhat surprising. One would expect the actual location

of the respective economies vis-à-vis the rest of the world to matter also. Trade between Australia and New Zealand is certainly much larger than it would be if these countries were not so isolated. But nothing in basic gravity models takes this into account. Similarly, if the relationship results from equilibrium between supply and demand, the absence of the goods prices in the model is also striking. Moreover, economists have long been challenged by the strong empirical support of gravity models and, simultaneously, their lack of theoretical foundations. Therefore, from Anderson (1979) on, a number of approaches have been proposed to derive the gravity relationship from fully specified theoretical models. Anderson and van Wincoop (2004) show that for trade flows the class of models that can be used is fairly large. They are based on either comparative advantage or on imperfect competition; the most popular uses monopolistic competition. They all lead to specifications more complex than (1) but such microfoundations significantly improve the understanding of gravity models and of their underlying mechanisms. Their main feature is that the interactions between two economic entities, such as regions, do depend on their location relative to other areas because of interesting price effects. Indeed, in an economy with more than two regions and costly trade, the supply, demand and price of goods traded between two given economies depend on the relative costs of all firms and consumers to access all the markets. For instance, if Australia and New Zealand reduced their trade costs to all other developed countries, the price of the goods they exchange bilaterally would increase relative to the price of all other goods, which would reduce their bilateral trade. On the other hand, the overall saving in trade costs and the increase in competition it induces would also imply a greater purchasing power, which would increase trade with all partners, including bilateral trade. Interestingly, in these models local incomes can be shown to be function of the area's market potential. This potential takes a form similar to the economic version proposed by Harris (1954), which is reminiscent of gravitational or electric potentials, which must be enriched by price effects again. This is the sum of all regions' income discounted by trade costs and weighted by complex price effects.

Econometric issues

Unfortunately, such formulations modify (1) in ways that make estimation more cumbersome. Typically the equation becomes nonlinear in some unknown parameters (such as price elasticity) that must be estimated simultaneously with the impact of trade costs. Various approaches, detailed in Feenstra (2004, ch. 5), have been proposed to deal with that difficulty. The first one consists of using nonlinear estimation procedures. Another solution takes as the left-hand-side variable the ratio of the bilateral flow to the flow to a reference destination, which makes the right-hand-side variables depend again on the origin and destination only. People sometimes use real price data, but these rarely match their exact definitions in the theoretical model. The least data-demanding strategy, which remains compatible with a large number of theory-grounded approaches, consists in estimating fixed effects for each origin and destination. However, the impact of the determinants of the fixed effects, the mass of the economies and their locations for instance are no longer identified.

Other econometric problems remain. First, the fact that each country trades with many destinations induces correlations between error terms, and therefore possible heteroskedasticity biases. More problematic are the zero trade flows towards a large number of destinations that often characterize many countries, which may induce selection biases. Santos Silva and Tenreyro (2006) propose a pseudo-

maximum likelihood estimator to deal with both issues. The literature based on heterogeneous firms and the presence of fixed export costs might help provide more theory-grounded approaches to these problems (Helpman, Melitz and Rubinstein, 2008). The fact that economic masses and prices are simultaneously determined with trade flows might also bias the estimations. Appropriate instrumental procedures should help, however. Last, the trade cost proxies might themselves be endogenous. For instance, new infrastructure can be built in response to an increase of trade flows, trade agreements may be signed preferably between privileged trade partners and networks may emerge once trade is large. This is more difficult to handle and these possible reverse-causality biases have yet to be really investigated.

Hence, the long history of gravity models does not prevent them from stimulating a lot of current research, both theoretical and empirical. Microfounded frameworks move the model further and further away from Newton's law, adding a lot to the understanding of the mechanisms shaping spatial interactions. Additional challenges lie ahead.

See Also

- GIS data in economics
- international trade and heterogeneous firms
- new economic geography
- regional development, geography of
- spatial economics
- systems of cities
- tradable and non-tradable commodities
- trade costs

Bibliography

Anderson, J. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69, 106–16.

Anderson, J. and van Wincoop, E. 2004. Trade costs. *Journal of Economic Literature* 42, 691–751.

Combes, P.-P., Mayer, T. and Thisse, J.-T. 2008. *Economic Geography*. Princeton: Princeton University Press.

Disdier, A.C. and Head, K. 2008. The puzzling persistence of the distance effect on bilateral trade. *Review of Economics and Statistics* 90, 37–48.

Feenstra, R.C. 2004. *Advanced International Trade*. Princeton: Princeton University Press.

Harris, C. 1954. The market as a factor in the localization of industry in the United States. *Annals of the*

Association of American Geographers 64, 315–48.

Helpman, E., Melitz, M. and Rubinstein, Y. 2008. Estimating trade flows: trading partners and trading volumes. *Quarterly Journal of Economics* 123, 441–487.

McCallum, J. 1995. National borders matter: Canada-US regional trade patterns. *American Economic Review* 85, 615–23.

Peri, G. 2005. Determinants of knowledge flows and their effects on innovation. *Review of Economics and Statistics* 87, 308–22.

Portes, R. and Rey, H. 2005. The determinants of cross-border equity flows. *Journal of International Economics* 65, 269–96.

Ravenstein, E.G. 1885. The laws of migration. *Journal of the Royal Statistical Society* 48, 167–227.

Redding, S. and Venables, A. 2004. Economic geography and international inequality. *Journal of International Economics* 62, 53–82.

Reilly, W.J. 1931. *The Law of Retail Gravitation*. New York: Pilsbury.

Santos Silva, J. and Tenreyro, S. 2006. The log of gravity. *Review of Economics and Statistics* 88, 641–58.

Stewart, J.Q. 1947. Suggested principles of ‘social physics’. *Science* 106, 179–80.

Tinbergen, J. 1962. *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: Twentieth Century Fund.

How to cite this article

Combes, Pierre-Philippe. "gravity models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000103> doi:10.1057/9780230226203.0670

The New Palgrave Dictionary of Economics Online

Great Depression (mechanisms)

Lee E. Ohanian
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes the theoretical framework and the diagnostic procedures that economists use to construct and test theories of depressions and booms, and also summarizes recent applications of these procedures to well-known depressions.

Keywords

balanced growth; banking crises; Euler equations; Great Depression; growth models; optimal growth theory; real business cycles.

Article

Depressions – prolonged periods in which output and employment fall 15 per cent or more below their long-run trend levels – are pathological. These episodes are particularly bizarre in economies like that of the United States, in which aggregate variables are almost always within a couple of percentage points of their long-run trend values (Leamer, 2004). Economists have long recognized that these abnormal episodes are strongly at variance with standard economic theory. For this reason, economists have not used equilibrium models, or, for that matter, any optimizing framework to investigate these episodes. And the reason why optimizing theories have been eschewed seems straightforward – what could equilibrium models tell us about episodes that appear to defy equilibrium reasoning? Prescott (2002) refers to the omission of theory from studies of depressions as a virtual ‘taboo.’

The omission of theory from analysis of depressions and crises comes at a cost, as it limits the tools with which economists can investigate these pathologies and thus limits the extent to which we can understand them. Since the late 1990s, however, macroeconomists have begun to use theory to investigate depressions, with a focus on the application of optimal growth theory developed by Cass (1965) and Koopmans (1965). Obviously, the steady state growth path of the Cass–Koopmans model – by definition – fails to reproduce any depression episode. But economists are beginning to learn about

the pathology of depressions as deviations from standard economic theory, much as a physician learns about illness by assessing deviations of a patient's vital signs from normality. The deviations of depressions from optimal growth theory are providing valuable diagnostic information that are used as the first step in developing and testing theories of depression. The remainder of this article describes these ‘depression diagnostics’, their application, and the new theories of depression that are being developed as a result of the use of these diagnostics.

The motivation behind the approach is that abnormal periods of macroeconomic activity – whether depressions or booms – lead to a proliferation of possible theories, and thus far there has been no systematic approach in the literature to shed light on which theories are the most promising. The diagnostic approach summarized here provides a simple method for identifying promising classes of theories. The idea of using optimal growth theory to diagnose depressions was initially used in Cole and Ohanian (1999), and the approach was further developed by Chari, Kehoe and McGrattan (2002), Cole and Ohanian (2002), and in particular by Chari, Kehoe and McGrattan (2007). The diagnostic method documents the deviations of the standard growth model, and then turns those deviations on their head to direct researchers to particularly promising classes of models.

To summarize the procedure, I begin with the following deterministic optimal growth model, which is given by:

$$\max \sum_{t=0}^{\infty} \beta^t \{\ln(C_t) + \varphi \ln(1 - L_t)\} N_t$$

where β is the household discount factor, C is consumption, L is hours worked, and N is population. The maximization is subject to the resource constraint, and the deterministic laws of motion for technology (X_t) which grows at the constant rate γ, and population (N_t) which grows at the constant rate n:

$$Y_t = F(K_t, (X_t L_t) + (1 - \delta) K_t \geq C_t + K_{t+1}, K_0 \text{ given}$$

$$X_t = (1 + \gamma)^t X_0, X_0 \text{ given}$$

$$N_t = (1 + n)^t N_0, N_0 \text{ given}.$$

While this example uses logarithmic preferences, one can use other functional forms for utility. To induce stationarity, all variables are divided by population, and all variables that grow along the steady state growth path are detrended by the exogenous productivity factor $(1+\gamma)^t$. The stationary, per-capita variables are denoted with lower-case variables. Standard dynamic programming techniques can be used to solve for the first-order conditions. The equations that characterize the planner's optimum, and that form the basis of the diagnostic procedure, are given by:

$$\varphi / (1 - l_t) = F_{lt} / c_t$$

$$(1 + \gamma)(1 + n) c_{t+1} / c_t = \beta [F_{kt+1} + 1 - \delta]$$

$$y_t = c_t + (1 + \gamma)(1 + n) k_{t+1} - (1 - \delta) k_t$$

$$y_t = F(k_t, x_0 l_t)$$

The first of the four equations listed above governs the household's allocation of time between market and non-market activities. The second equation, often called the Euler equation, governs the household's allocation of income between consumption and savings. The third equation is the resource constraint, and the fourth equation is the production function. Given parameter values, a functional form for the technology, and time series observations on output, consumption, labour, and investment, we construct the following percentage deviations between theory and data that will form the basis of the diagnostics:

$$\varepsilon_{lt} = \{\varphi / (1 - l_t) / (F_{lt} / c_t)\} - 1$$

$$\varepsilon_{kt} = \left\{(c_{t+1} / c_t) / (\beta [F_{kt+1} + 1 - \delta])\right\} - 1$$

$$\varepsilon_{yt} = \frac{y_t - c_t - i_t}{y^* - c^* - i^*} - 1,$$

where starred variables are steady state values, and

$$\varepsilon_{zt} = [y_t / F(k_t, x_0 l_t)] - 1$$

I denote these as the labour deviation, (ε_{lt}) the Euler deviation (ε_{kt}), the National Income and Product Accounts (NIPA) deviation (ε_{yt}), and the productivity deviation (ε_{zt}). With the exception of the NIPA equation, each of these deviations is constructed as the percentage difference between the left-hand side and the right-hand side of each equation. The NIPA deviation is normalized relative to a steady state value. Note that, along the steady state growth path, all of these deviations are equal to zero by construction.

The next step is to choose parameter values and functional forms. Regarding parameterization, it is often convenient to choose values so that the deviations are equal to zero immediately before the episode of interest. Thus, given values for consumption, hours worked, and the technology, the parameter value for ϕ is chosen so that the ε_{lt}=0 prior to the episode. Similarly, given values for consumption, capital, hours worked, and the depreciation rate, the parameter value for β is chosen so that (ε_{kt}) is zero prior to the period of interest. Regarding functional forms, the application here uses log preferences over consumption and leisure, though other forms can certainly be used. For the production function, it is common to use Cobb–Douglas.

For heuristic purposes, I have described the procedure in a deterministic economy. The extension to a stochastic economy is fairly straight forward, and is presented in detail in Chari, Kehoe and McGrattan (2007). In summary, the procedure is modified for stochastic environments so that the measured deviations are modelled as a vector stochastic process, which is typically specified as a vector autoregression (VAR) and which for simplicity is represented here as a first-order process. The vector *W* is a 4×1 vector containing the four deviations previously defined above. The VAR is given by:

$$W_t = \varphi W_{t-1} + \varepsilon_t, \; E(\varepsilon\varepsilon') = \Omega$$

The labour, NIPA, and productivity deviation are measured exactly the same way as in the deterministic case. Measuring the Euler deviation, however, requires evaluating the expectation of the right-hand side of the Euler equation. This can be accomplished by log-linearizing the planner's conditions, and then solving the resulting linear system. The linearized model can then be used to forecast the right-hand side of the Euler equation, which implicitly defines the stochastic Euler deviation.

Chari, Kehoe and McGrattan (2007) show that, given this procedure of measuring the deviations, the model economy is able to reproduce (up to numerical solution error) any observed sequences of fluctuations of the endogenous variables, given the sequences of these deviations and an initial value of the capital stock, population, and technology. Consequently, *any* fluctuation of an actual economy from its balanced growth path value is *entirely* accounted for by one or more of these deviations within the growth model. This insight transforms the growth model into an accounting framework, and as such the procedure can document the relative importance of each deviation for understanding fluctuations. To do

this, the investigator calculates the solution of the model using just one deviation at a time, or a subset of the four deviations. Chari, Kehoe and McGrattan (2007) conducted this type of analysis to show that the labour deviation and the productivity deviation account for virtually all of the Great Depression through 1933.

It is important to recognize that the interpretation of the deviations at this stage of analysis differs from that in the business cycle literature. For example, real business cycle models typically identify the Solow residual, or some variant of that residual, as a primitive shock. The diagnostic framework summarized here does not necessarily place this type of identifying interpretation on these deviations. Rather, the focus here is to provide clues to researchers for the class of models to consider. In the case of a productivity deviation, the key point of the diagnostic is that it informs researchers that a successful theory will be one that can be mapped into a growth model with this feature. A shift in the aggregate technology set is one interpretation of this deviation, but there are other interpretations as well, including theories based on the mismatch of resources across plants which impacts the Solow residual (see Restuccia and Rogerson, 2007), a distortion in relative prices that leads firms to shift their input mix, which also impacts the Solow residual(see Chari, Kehoe and McGrattan, 2007), or changes in government regulations (see Hansen and Prescott, 1993).

To concretely illustrate the use of this diagnostic approach further, suppose an investigator calculated these deviations, and found they were all roughly zero with the exception of the labour deviation. This information narrows the class of theories so that it would be admissible to include only those that feature some mechanism that changes the rate at which households value their time with respect to the measured wage, but leaves all other margins within the growth model unchanged. Models in this class include those with time-varying taxes on labour income, time-varying subsidies to non-market time, such as changes in unemployment benefits, changes in the incentive to accumulate human capital, and changes in union or monopoly power. The diagnostics presented in this hypothetical example also exclude several classes of models, including models of productivity shocks, models with government spending shocks, and models with time-varying taxes on capital income or investment, as the margins on which these factors operate are not distorted.

This example of a large labour distortion, but no other large distortions, not only illustrates how the method can be applied, but also happens to be the outcome of the procedure when the tools are applied to two well known and puzzling depressions, in the United States between 1933 and 1939 and the United Kingdom between 1921 and 1929. These episodes have long been considered puzzling because of their long duration, and because standard economic fundamentals were reasonably healthy during this period. In particular, the US money supply grew quickly after 1933, productivity growth was rapid after 1933, and there were no bank runs. All of these factors should have fostered a rapid recovery in the United States, yet hours worked in the United States recovered very little after the 1933 trough. In the United Kingdom, productivity growth was at its trend level during the 1920s, and the economy should have been poised for a significant post-First World War recovery.

During both of these episodes ε_{lt} was very large, but other deviations were small. Specifically, the marginal rate of substitution fell well below the wage in both of these depressions. In the United States, the real wage was as much as 100 per cent above the marginal rate of substitution, while total factor productivity was near trend in both episodes, and the intertemporal consumption-savings Euler equation was undistorted. The diagnostic thus establishes that the key depressing factors in these episodes severely affected the labour market but did not depress productivity, nor did they distort the household's

intertemporal first-order condition governing the consumption savings decision.

Cole and Ohanian (1999, p. 6) used this diagnostic information to focus on theories which could substantially distort the household's time allocation decision. Perhaps the most obvious factor that distorts this decision is changes in labour and consumption tax rates, as these taxes change the incentive to work and thus impact this first-order condition. The authors ruled this factor out based on empirical grounds, because changes in labour and consumption taxes were relatively small during this period in both countries.

For the United States, the next factor they considered was government policies that impact the labour market, and major labour market policies were adopted in both countries just prior to these episodes. In the United States, President Roosevelt introduced the National Industrial Recovery Act in 1933 that permitted industrial firms to cartelize and raise prices provided that they also raised wages. Relative prices and real wages in these sectors jumped significantly after the adoption of these policies, and employment and output remained low throughout the 1930s. This led Cole and Ohanian (2004) to develop a dynamic insider–outsider model in which firms were able to collude provided they reached a wage agreement with their workers. The model accounted for about 60 per cent of the post-1933 depression, and was consistent with the behaviour of wages and prices in both the industrial and non-industrial sectors of the economy. The insider–outsider model developed by Cole and Ohanian maps into a standard growth model with a large labour deviation, but the other deviations are roughly zero.

For the United Kingdom, Cole and Ohanian (2002), following work by Benjamin and Kochin (1974), noted that the United Kingdom adopted a very generous unemployment policy after First World War. Initially, benefits were available to those who worked for only one day in total, could be received indefinitely, and provided generous payments. At one point, the benefits paid were equal to about four per cent of GNP. Cole and Ohanian developed a model in which the policy reduced hours worked by about 15 per cent and distorted the household's first-order condition governing time allocation, but did not affect the other margins in the growth model.

In both of these episodes, the substantive finding from the diagnostic procedure led to the development of theories in which government policies distorted labour markets and significantly reduced hours worked, but did not significantly distort the incentive to save. The most provocative issue that has arisen in this diagnostic literature regards the importance of distortions to the capital market as a source of depression in the United States in the 1930s. Bernanke (1983), in a very influential paper, argued that banking panics led to a longer and deeper depression, using regression analysis that demonstrated that banking variables were statistically significant in an output equation that also included a measure of monetary shocks. Chari, Kehoe, and McGrattan (2007) show that some optimizing models that feature the financial intermediation channel emphasized by Bernanke, including Carlstrom and Fuerst (1997), map into a growth model with a substantial Euler equation deviation. However, the empirical Euler deviation is small during this period, leading Chari, Kehoe, and McGrattan to conclude that financial frictions theories that operate through the specific channel emphasized in Carlstrom and Fuerst are not quantitatively important for the Great Depression. This finding was a surprise to many economists, and is leading to new research in this area (see Christiano and Davis, 2007; Primaceri, Schaumber and Tambaloti, 2006).

It is likely that this procedure will be used not only in business cycles, but in a variety of applications. Lu (2007) uses the procedure to study Taiwan since the 1950s, and finds a very large Euler deviation.

Given this finding, and the fact that the spread between deposit and lending rates in Taiwan declines from about 12 per cent in 1950 to about two per cent in 2003, she develops a model of technological advances in financial intermediation efficiency that can account for the decline in the saving–lending spread and the Euler deviation, and finds that this development led to a significant increase in Taiwanese per-capita income.

See Also

- Great Depression
- Great Depression, monetary and financial forces in
- growth and cycles

Bibliography

Benjamin, D. and Kochin, L. 1974. Searching for an explanation of unemployment in interwar Britain. *Journal of Political Economy* 87, 441–78.

Bernanke, B. 1983. Nonmonetary effects of the financial crisis in the propagation of the great depression. *American Economic Review* 73, 257–76.

Carlstrom, C. and Fuerst, T. 1997. Agency costs, net worth, and business fluctuations: a computable general equilibrium analysis. *American Economic Review* 87, 893–910.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32, 233–40.

Chari, V.V., Kehoe, P. and McGrattan, E. 2002. Accounting for the Great Depression. *American Economic Review* 92, 22–7.

Chari, V.V., Kehoe, P. and McGrattan, E. 2007. Business cycle accounting. *Econometrica* 75, 781–836.

Christiano, L. and Davis, J. 2007. Two flaws in business cycle accounting. Discussion paper, Department of Economics, Northwestern University.

Cole, H. and Ohanian, L.E. 1999. The Great Depression in the United States from a neoclassical perspective. *Federal Reserve Bank of Minneapolis Quarterly Review* 23(1), 2–24.

Cole, H. and Ohanian, L.E. 2002. The U.S. and U.K. Great Depressions through the lens of neoclassical growth theory. *American Economic Review* 92, 28–32.

Cole, H. and Ohanian, L.E. 2004. New Deal policies and the persistence of the Great Depression.

Journal of Political Economy 112, 779–816.

Hansen, G. and Prescott, E. 1993. Did technology shocks cause the 1990–91 recession? *American Economic Review* 83, 280–6.

Koopmans, T. 1965. *On the Concept of Optimal Economic Growth*. New York: Rand McNally.

Leamer, E. 2004. The truth about GDP growth. *Harvard Business Review*, October, 24.

Lu, S.-S. 2007. Understanding growth miracles: the case of Taiwan. Ph.D. thesis, UCLA.

Prescott, E. 2002. Prosperity and depression. *American Economic Review* 92, 1–15.

Primaceri, G., Schaumburg, E. and Tambalotti, A. 2006. Intertemporal disturbances. Working Paper No. 12243. Cambridge, MA: NBER.

Restuccia, D. and Rogerson, R. 2007. Policy distortions and aggregate productivity with heterogeneous plants. Working paper, Department of Economics, University of Toronto.

How to cite this article

Ohanian, Lee E. "Great Depression (mechanisms)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000200> doi:10.1057/9780230226203.0672

The New Palgrave Dictionary of Economics Online

Great Depression, monetary and financial forces in

Satyajit Chatterjee, P. Dean Corbae

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We survey papers that seek model-based answers to the following questions regarding the Great Depression. What caused the worldwide collapse in output from 1929 to 1933? Why was the recovery from the trough of 1933 so protracted for the United States? How costly are Depression-like episodes in terms of welfare? Was the decline in output preventable? The papers point to: an important, but not exclusive, role of monetary factors in causing the decline; counterproductive labour market interventions in making the recovery slow; uninsured risk of unemployment in making Depression-like episodes costly; timely provision of liquidity as a preventive policy.

Keywords

confidence; debt-deflation hypothesis; depressions; dynamic stochastic general equilibrium (DSGE) model; financial intermediation; gold standard; Great Depression; liquidity preference; monetary and financial forces in the Great Depression; monetary base; money multiplier; multiple equilibria; sticky wages; total factor productivity

Article

What caused the worldwide collapse in output from 1929 to 1933? Why was the recovery from the trough of 1933 so protracted for the United States? How costly was the decline in terms of welfare? Was the decline preventable? These are some of the questions that have motivated economists to study the Great Depression.

Cole and Ohanian (1999) document that US per capita GNP fell 38 per cent below its long-run trend path (of two per cent per annum growth) from 1929 to 1933. Real per capita non-durables consumption fell nearly 30 per cent, durables consumption fell over 55 per cent, and business investment fell nearly 80 per cent. On the input side, total employment fell 24 per cent and total factor productivity (TFP) fell 14 per cent. On the nominal and financial side, the GNP deflator fell 24 per cent; per capita M1

(currency plus deposits) fell 30 per cent; M1 velocity fell 32 per cent; the per capita monetary base rose 9 per cent; the currency–deposit ratio rose over 160 per cent (Friedman and Schwartz, 1963, Table B3); the loan–deposit ratio fell 30 per cent (Bernanke, 1983, Table 1); and *ex post* real commercial paper rates rose from six per cent in 1929 to a peak of 13.8 per cent in 1932.

What caused the Depression? For the United States, Friedman and Schwartz (1963, p. 300) argued that it was the decline in the stock of M1 – a consequence of Fed tightening and of a fall in the money multiplier induced by banking panics. According to Eichengreen (1992), international adherence to the gold standard transmitted the US monetary contraction to other industrialized countries. Specifically, high interest rates and low prices in the United States attracted foreign inflows of gold (in 1932 the United States and France held over 70 per cent of the world gold reserves), which the Fed largely sterilized (that is, sold domestic government debt and bought money). The outflow of gold from foreign countries implied that gold-backed money supplies of those countries had to decline in order to meet their cover ratios. Further evidence (see Bernanke and James, 1991, Table 4) of the importance of the gold standard in transmitting the contraction comes from the experience of countries like Britain, which suspended the gold standard in 1931 and recovered by 1932; from Spain, which never was on it and had a much less severe contraction than those on the gold standard; and from France, which was one of the last major countries to leave it and still faced declining industrial production past the 1933 trough. As Bernanke (1995, p. 3) puts it: ‘The new gold-standard research allows us to assert with considerable confidence that *monetary factors played an important causal role*, both in the worldwide decline in prices and output and in their eventual recovery.’

However, much of this evidence is problematic in that it is in the nature of correlations between *endogenous* variables – a fact that makes it challenging to establish causality. Did the decline in M1 *cause* the decline in aggregate output or – as Temin (1976) argued early on – did M1 and aggregate output decline in response to some other common shock? If the ‘monetary-cum-exchange- rate-policy’ explanation is indeed correct, we ought to be able to demonstrate its correctness in a reasonably calibrated, dynamic stochastic general equilibrium (DSGE) model. To paraphrase Lucas (1993, p. 271): ‘If we know what a depression is, we ought to be able to *make* one.’ The challenge of ‘making’ a depression has been taken up by various researchers and constitutes a noteworthy recent development in depression research.

The conventional explanation of why money affected output is sticky nominal wages – goods prices fell as a result of the monetary contraction but nominal wages adjusted slowly and the ensuing increase in the real wage depressed the demand for labour. One significant contribution to evaluating this conventional explanation is by Bordo, Erceg and Evans (2000). They calibrate a one-sector stochastic macro model with four-quarter nominal wage rigidity and find that 70 per cent of the output decline from 1929 to 1933 can be accounted for by feeding in the negative innovations to the actual M1 money supply process during that period.

Although the findings of Bordo, Erceg and Evans are striking, there are some unresolved issues. One is that the real-wage rise in the model was chosen to mimic the actual real-wage rise in the manufacturing sector while there is some indirect evidence that non-manufacturing real wages actually fell during the 1929–33 downturn. Cole and Ohanian (2000) re-examine the sticky-wage hypothesis in a multisector model and find much less support for it.

A second unresolved issue is that Bordo, Erceg and Evans do not take into account the evidence on

aggregate labour productivity and TFP, both of which declined between 1929 and 1933. Ohanian (2002) argues that only about a third of the decline in labour productivity and/or TFP can be plausibly accounted for by mismeasurement of factor inputs. By itself, a decline in TFP could account for a substantial fall in aggregate output, consumption and investment. Unless a decline in TFP can be viewed as an endogenous response to the monetary shock (through, for example, aggregate increasing returns), the decline leaves less scope for a purely monetary explanation. Using a DSGE model where money is non-neutral due to imperfect information, Cole, Ohanian and Leung (2005) show that the decline in M1 accounts for only one-third of the decline in output from 1929 to 1933, while the effect of an exogenous decline in TFP accounts for two-thirds. They use a misperceptions model of monetary non-neutrality because such a model generates less of a counterfactual movement in labour productivity than a model with nominal wage rigidities.

Sticky wages and monetary misperceptions are not the only mechanisms through which money can affect real output. Irving Fisher (1933) pointed out that the unanticipated fall in prices during 1929–33 led to bankruptcies because it increased the real value of nominal debt of households, firms, and financial intermediaries. This ‘debt-deflation’ hypothesis was analysed by Mishkin (1978) for households and formalized by Bernanke and Gertler (1989) for firms. More generally, Bernanke (1983) argued that the reduction in borrower net worth increased the cost of obtaining external finance, while bank failures and tightened credit standards hampered the efficient allocation of capital. However, a quantitative DSGE model featuring this mechanism has yet to be implemented for the Great Depression. Such a model holds out the promise of explaining some portion of the puzzling decline in TFP during 1929–33 as an endogenous response to a misallocation of capital.

One of the most striking facts of the Depression was the reduction in the money multiplier from 1929 to 1933 associated with the flight from bank deposits to currency. Cooper and Corbae (2002) construct a model in which households have the option of saving in the form of currency or bank deposits, and in which bank deposits ultimately fund working capital for businesses. Because of increasing returns in the intermediation technology associated with fixed verification costs, their model admits multiple equilibria. In the good equilibrium the return on bank deposits is high, households hold small amounts of currency, and output is high. In the bad equilibrium, the return on bank deposits is low, households substitute into currency, and output is low. A shift from the good to the bad equilibrium replicates many of the salient nominal changes that occurred between 1929 and 1933. Although not quantitative, their work formalizes the idea that output, credit and money supply responded negatively to a loss in confidence – much as Irving Fisher (1933, p. 343) suggested it did.

Why was the recovery from the trough of 1933 so protracted for the United States? As noted by Cole and Ohanian (1999), aggregate US output was still below trend in 1939. The answer cannot be the gold standard or M1 because the United States left the gold standard in 1933 and the US money stock recovered rapidly thereafter. One explanation offered is that the National Industrial Recovery Act (NIRA) encouraged businesses to accept high real wages of industrial workers. Cole and Ohanian (2004) embed labour bargaining into a DSGE model and quantitatively explore the effect of the NIRA, giving more weight to workers in the bargaining process post 1933. Their model is reasonably successful in producing a slow recovery. Adverse labour market interventions also appear to have played a role in other industrialized countries such as Germany, France, the UK and Italy (Kehoe and Prescott, 2002).

How costly was the Depression in terms of welfare? Real per capita consumption of non-durables fell 30

per cent in the United States but it is not known how this decline was distributed across households. Chatterjee and Corbae (2006) analyse how households that can self-insure against uninsured earnings losses would fare through a depression. They found that the welfare cost of living in a world with a small likelihood of a Depression-like event is quite large – somewhere between one and seven per cent of consumption in perpetuity depending on the completeness of asset markets. Much of this cost is associated with the increased variability of individual consumption streams.

Was the Depression preventable? First, if the 'monetary-cum-exchange-rate-policy' explanation is correct, the right monetary policy could have prevented the decline. Christiano, Motto and Restagno (2003) estimate a DSGE model with many shocks but find that a liquidity preference shock inducing households to hold currency instead of deposits played the most important role in the contraction phase of the Depression. They then specify a policy rule that raises the monetary base as a function of liquidity shocks, and run a counterfactual experiment where they find that output would have declined only six per cent if such a reaction function had been in place. Second, if a portion of the decline in output was the result of a banking collapse stemming from a shock to confidence, then – as shown by Cooper and Corbae (2002) – an announcement by the monetary authority that it stands ready to supply liquidity to the banking system might have moderated the decline. Finally, with regard to the slow recovery in the United States, the only credible explanation offered is adverse labour market intervention. If this explanation is correct, we know what *not* to do to prolong a severe decline in output.

See Also

- Great Depression
- Great Depression (mechanisms)
- monetary business cycle models (sticky prices and wages)
- real business cycles

Bibliography

Bernanke, B. 1983. Nonmonetary effects of the financial crisis in the propagation of the Great Depression. *American Economic Review* 73, 257–76.

Bernanke, B. 1995. The macroeconomics of the Great Depression: a comparative approach. *Journal of Money, Credit and Banking* 27, 1–28.

Bernanke, B. and Gertler, M. 1989. Agency costs, net worth, and economic performance. *American Economic Review* 79, 14–31.

Bernanke, B. and James, H. 1991. The gold standard, deflation, and financial crisis in the Great Depression: an international comparison. In *Financial Markets and Financial Crises*, ed. R. Glenn Hubbard. Chicago: University of Chicago Press.

Bordo, M., Erceg, C. and Evans, C. 2000. Money, sticky wages, and the Great Depression. *American Economic Review* 90, 1447–63.

Chatterjee, S. and Corbae, D. 2007. On the aggregate welfare cost of Great Depression unemployment. *Journal of Monetary Economics* 54, 1529–1544.

Christiano, L., Motto, R. and Rostagno, M. 2003. The Great Depression and the Friedman–Schwartz hypothesis. *Journal of Money, Credit, and Banking* 35, 1119–97.

Cole, H. and Ohanian, L. 1999. The Great Depression in the United States from a neoclassical perspective. *Federal Reserve Bank of Minneapolis Quarterly Review* 23(1), 25–31.

Cole, H. and Ohanian, L. 2000. Re-examining the contributions of money and banking shocks to the U. S. Great Depression. *NBER Macroeconomics Annual* 15(1), 183–227.

Cole, H. and Ohanian, L. 2004. New Deal policies and the persistence of the Great Depression: a general equilibrium analysis. *Journal of Political Economy* 112, 779–816.

Cole, H. Ohanian, L. and Leung, R. 2005. Deflation and the international Great Depression: a productivity puzzle. Working Paper No. 11237. Cambridge, MA: NBER.

Cooper, R. and Corbae, D. 2002. Financial collapse: a lesson from the Great Depression. *Journal of Economic Theory* 107, 159–90.

Eichengreen, B. 1992. *Golden Fetters: The Gold Standard and the Great Depression 1919–1939*. New York: Oxford University Press.

Fisher, I. 1933. The debt–deflation theory of Great Depressions. *Econometrica* 1, 337–57.

Friedman, M. and Schwartz, A. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press.

Kehoe, T. and Prescott, E.C. 2002. *Great Depressions of the Twentieth Century*. Special Issue of the *Review of Economic Dynamics* 5(1).

Lucas, R. 1993. Making a miracle. *Econometrica* 61, 251–72.

Mishkin, F. 1978. The household balance sheet and the Great Depression. *Journal of Economic History* 38, 918–37.

Ohanian, L. 2002. Why did productivity fall so much during the Great Depression? *Federal Reserve*

Bank of Minneapolis Quarterly Review 26(2), 12–17.

Temin, P. 1976. *Did Monetary Forces Cause the Great Depression?* New York: W.W. Norton and Company.

How to cite this article

Chatterjee, Satyajit and P. Dean Corbae. "Great Depression, monetary and financial forces in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000398> doi:10.1057/9780230226203.0673

The New Palgrave Dictionary of Economics Online

Great Depression

Peter Temin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The world depression of the 1930s was the greatest peacetime economic catastrophe in history. There had been hard times before, but never without war, natural disaster or pestilence. The massive and long-lasting unemployment and hardship of the 1930s was a pathology of industrial society, caused by a malfunctioning of the economic system. Adherence to gold-standard policies led to a set of currency crises in 1931 that turned a bad recession into the Great Depression.

Keywords

banking crises; beggar-my-neighbour; currency crises; deflation; demand shock; devaluation; Federal Reserve System; Friedman, M.; German hyperinflation; gold standard; Great Depression; Kindleberger, C.; real wage rates; Schwartz, A.; specie-flow mechanism; transfer problem; unemployment; Young Plan

Article

Magnitude

Figure 1 shows the fall in industrial production during the Great Depression in the four largest national economies at that date. Industrial production declined by almost half in the United States and Germany. It fell more slowly and continuously in France, and paused rather than fell in Great Britain. National incomes did not fall as far as industrial production since services did not contract as much, but they decreased sharply; real per-capita GNP in the United States fell by one-third. National experiences in the depression varied greatly, but very few countries in the world escaped the economic hardship of the 1930s. One task for any account of the Great Depression is to explain its worldwide impact.
Figure 1
Industrial production, 1920–1939. *Source*: Temin (1989).

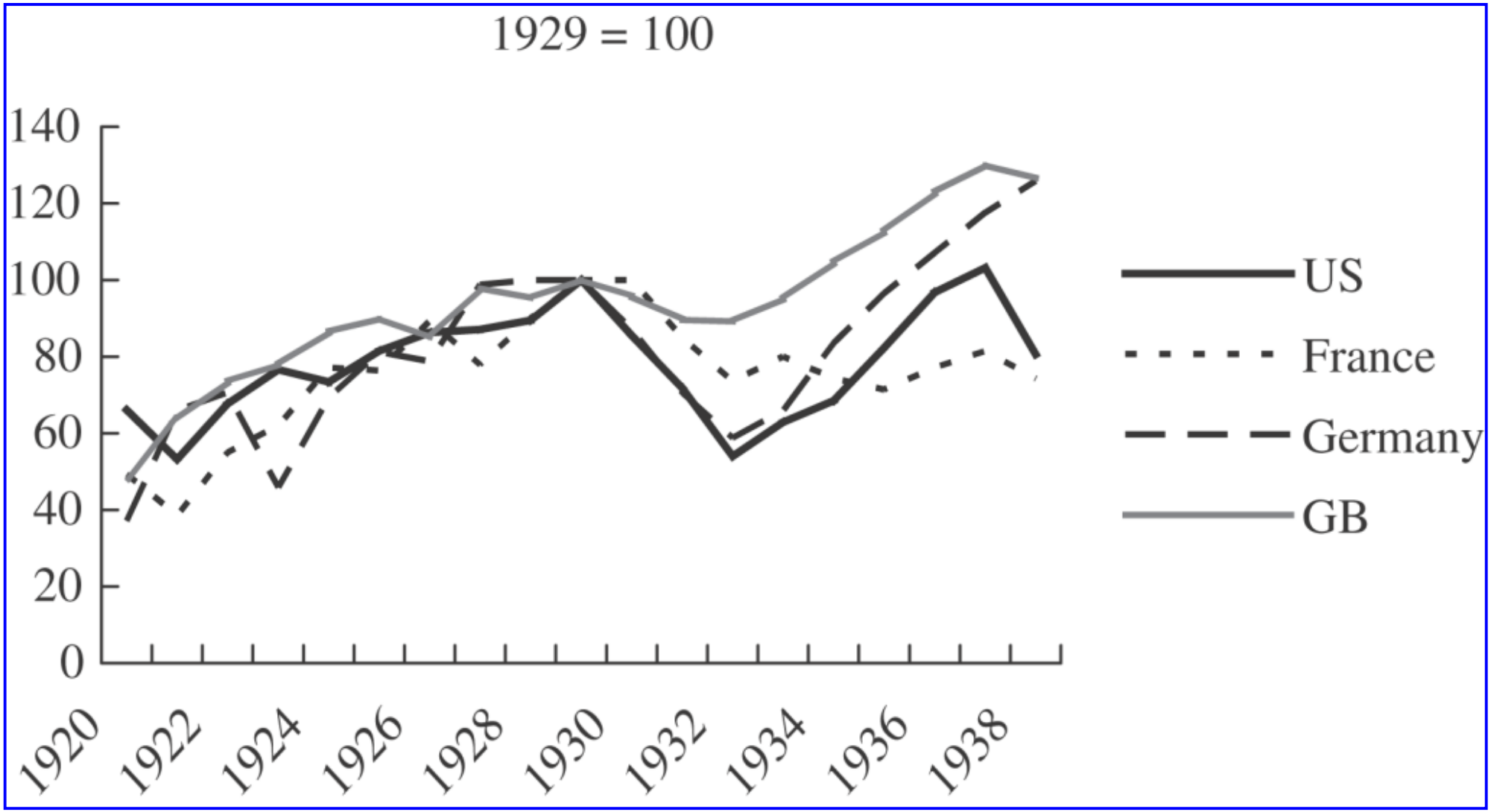

Figure 2 shows the fall in wholesale prices for the same four countries. Prices fell at the same time as production, by the same amount or more. Unemployment grew dramatically in almost all countries. Rates for the four largest economies are shown in Table 1. Only in the United Kingdom were unemployment rates approximately as high in the 1920s as in the 1930s, due to depressed conditions in Britain during the 1920s and a mild depression in the 1930s. Other countries for which we have data fit the more common pattern of higher unemployment in the 1930s.

Industrial unemployment rates, 1921–1938

Country	1921–29	1930–38	Ratio
France	3.8	10.2	2.7
Germany	9.2	21.8	2.4
United Kingdom	12	15.4	1.3
United States	7.9	26.1	3.3

Source: Temin (1989).

Figure 2
Wholesale prices, 1924–1939. *Source*: Temin (1989).

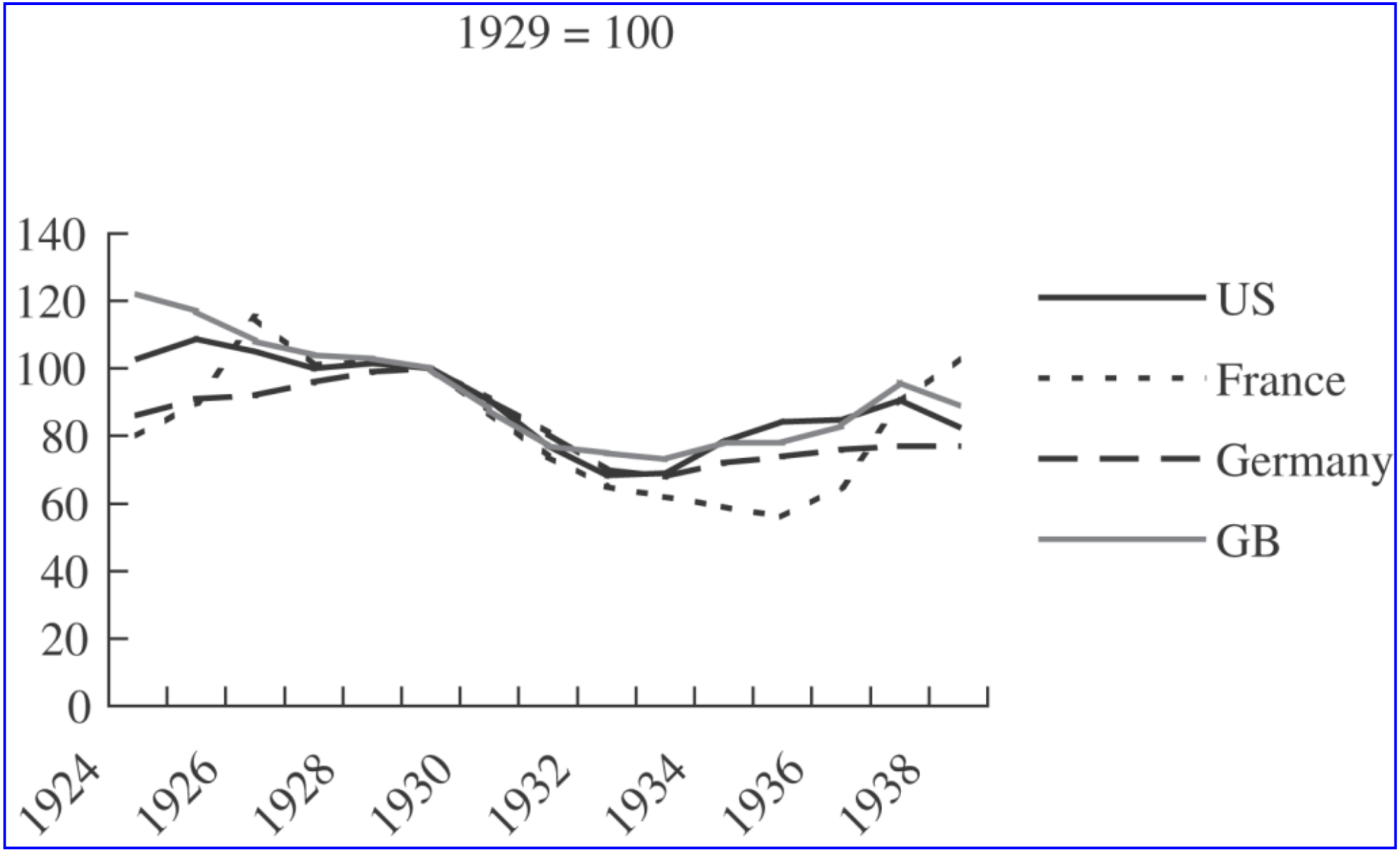

Unemployment meant distress in the 1930s, most visible in Europe and North America. Diets in Europe became very monotonous despite the presence of home-grown vegetables in some areas. Families ate meat only rarely, starches were the basis of most diets, and sugar frequently was replaced by cheaper saccharine. Even this poor diet consumed almost all the family income. Families with children bought milk, most families bought coal for heat, but there was little money left over for clothes and other expenses. Shoes in particular were a problem. Families typically could not afford to replace shoes that had worn out, and so they were patched and patched again. Some families even restricted the activities of their children to save the wear and tear on their shoes.

While spending was channelled into food, and food into bread and coffee, personal travel was reduced to journeys to local neighbourhoods and villages. Trips to towns and town

centres had been increasing during the 1920s, to go to the theatre, do Christmas shopping, or attend school. With unemployment, the money to undertake these journeys vanished. Even tram and train fares became a burden, and people relied more heavily on their bicycles. The isolation of rural villages, alleviated by the railways and the prosperity after the First World War, reappeared in the depression.

Unemployed men were exceedingly idle; an increase of apathy reduced all forms of recreational activity. Men passed their time doing essentially nothing; when asked, they could not even recall what they had done during the day. They sat around the house, went for walks – walking slowly – or played cards and chess. Most men went to bed early; there simply was no reason to stay awake. Women were far more active. They spent time cooking, mending clothes to make them last longer, and managing their budgets. Men contributed less to the running of the household than before, sometimes not even turning up on time for meals, and women had the full responsibility for maintaining the household. Even though women previously had struggled to complete their housework after working, they uniformly would have preferred being back at work.

Sociologists observed that most European unemployed families were resigned to their condition. Such families were hanging on, preserving as much of their life and family as they could on their meagre budgets. All their activity was dedicated to getting by; no thought was given to the future. Some families still planned as before, but others collapsed entirely into mental and physical neglect and conflict.

Beyond Europe and North America, the story of destitution was the same, although the workers' issues typically were more related to physical survival. Rural families in Asia and Africa suffered from the low prices that their crops received in the depressed world markets. They do not seem to have lapsed into idleness like unemployed urban workers, but rather continued to produce crops in the hope of increasing their incomes. Consumers in India, no longer able to afford imported cloth, gave a boost to domestic, beleaguered handloom weavers. Workers in Latin America retreated from cities and organized agriculture back into the countryside, and little is known of their living conditions. Latin American governments divided into active states that tried to insulate their economies from the outside world and passive states that waited for better times. Governments were surprisingly stable under this economic stress, but they collapsed in some countries, ranging from Germany to Burma.

Analysis

The first question to ask about this contraction is whether the shocks that produced it were demand or supply shocks. The simultaneous fall in production and prices indicates that the shocks were demand shocks, that the economies of the world were moving down along their upward-sloping aggregate supply curves in response to downward shifts of aggregate demand curves. The apathetic reaction to unemployment in the Great Depression confirms the hypothesis that the depression was due to a demand shock. Had it been due to a supply shock, families would have been unemployed by choice, happy with their extra leisure. The psychological depression also put great strains on the social structure, and even the political structure in some countries. It was in soil such as this that the noxious weed of National Socialism grew in Germany.

A second question about the Great Depression is how so many countries could have had negative demand shocks at the same time. The answer is that all these countries were adopting deflationary policies according to the dictates of the gold standard. The gold standard was characterized by the free flow of gold between individuals and countries, the maintenance of fixed values of national currencies in terms of gold and therefore each other, and the absence of an international coordinating or lending organization such as the International Monetary Fund. Under these conditions, the adjustment mechanism for a deficit country was deflation rather than devaluation – that is, a change in domestic prices instead of a change in the exchange rate. Lowering prices and possibly production as well would reduce imports and increase exports, improving the balance of trade and attracting gold or foreign exchange. (This is the price-specie-flow mechanism first outlined by Hume in 1752.)

A recession began at the end of the 1920s in the United States and Germany. Both countries began to contract economically, at least partly as a result of central bank pressure. The initial downturns appear to be independent in each country, but their economies were connected, and it is hard to be sure about this. In any case, it was gold-standard policies that transformed the downturn into the Great Depression and pulled the rest of the world down. The choice of deflation over devaluation was the most important factor determining the depth of the Great Depression. The choice was seen clearly and supported by contemporaries in all industrial countries who insisted that the way out of depression was to cut wages and thereby the costs of production and the prices of goods and services. Devaluation was not a respectable option.

Less developed countries were less likely to be on the gold standard than those in Europe or North America. They suffered from the depression nonetheless because of their ties to gold standard countries. As industrial countries reduced their demand for imports, exports from less developed countries declined. As industrial countries stopped exporting capital, less developed countries found their balance of payments deteriorating further. A few countries, such as Spain and Japan, devalued their currencies early and avoided the worst of the depression, but many more countries were not in a position to do this or where it would have had a large effect.

A third question that economists ask about the Great Depression is why the fall in demand was not absorbed entirely in falling prices. In other words, why did prices not fall more and production less than shown in Figures 1 and 2? The relative stability of wages caused production and employment to fall; falling prices and wages did not absorb the full brunt of the fall in demand. Falling prices also put pressure on financial institutions, whose failures reduced production as well.

Governments and central banks could not easily deflate their economies in the aftermath of the First World War. Workers, who had borne the burdens of international stability mutely in the past, expected and even demanded a voice in policy after their sacrifices during that war. The inability of economic policymakers to force wages down rapidly created the conditions for the Great Depression. The political strains generated by attempts to lower wages caused investors to fear for the stability of the gold standard even as policymakers

struggled to maintain it. One reason the gold standard worked well before 1914 was that labour had no voice. The spread of democracy both cast doubt on the monetary authorities' commitment to the gold standard and reduced price flexibility.

Banks failed right and left in the midst of deflation and currency crises. Widespread banking failures were restricted to countries on the gold standard, showing that the strain of the gold standard was the principal cause of bank distress. All banks suffered as economic activity and prices declined, but the diversion of central banks from the support of commercial banks to the defence of the currency made the difference between banks in difficulty and banking crises. The German government took over the country's great banks in June and July 1931; American banks were allowed to fail continuously as economic decline continued. It seems that a slow crisis was more destructive of economic activity than a rapid one, though there are not enough observations to test this hypothesis.

Narrative

The narrative of the Great Depression properly begins with the First World War. The dislocations of the war and the peace agreements meant that many adjustments had to be made in the international economy. Strains were evident in the immediate aftermath of the war, resulting in hyperinflations in several countries, most notably Germany. The response was to return to the gold standard in the mid-1920s in the hopes of regaining pre-war stability. Alas, the cure proved worse than the disease.

Federal Reserve policy became contractionary at the start of 1928 in order to combat speculation in the New York stock market and to arrest a gold outflow begun in part by previous financial ease. The gold outflow was a prominent determinant of the policy change, even though it was tiny relative to US reserves. The Federal Reserve's primary aim in 1928 and 1929 was to curb speculation on the stock exchange while not depressing the economy. Even though this policy did not impede stock-market speculation, it reduced the rate of growth of monetary aggregates and caused the price level to turn down. The monetary stringency was even tighter than it seems from examining the aggregate stock of money because the demand for money to effect stock-market transactions rose, leaving less for other activities.

The German economy was heavily dependent on imported capital in the 1920s. Popular history regards the capital imports as a necessary offset to Germany's outflow of war reparations payments; they were needed to solve the transfer problem. The reality was quite different. Germany managed to avoid paying reparations by a variety of economic and political manoeuvres that succeeded in postponing its obligations until they could be repudiated entirely. The capital inflow therefore represented a net increase in the resources available to the German economy. The Reichsbank paradoxically worried that this capital inflow was unhealthy and acted to curtail it, sharply reducing the amount of credit available on the German market at the end of the 1920s. The capital flow from the United States to Germany ceased at the end of the 1920s, but the downturn in Germany preceded this fall and derived largely from German economic policies.

At its inception, the Great Depression was transmitted internationally by a gold-standard ideology, a mentality that decreed that external economic relations were primary and that speculation like the booming stock markets in New York and Berlin was dangerous. As the American, British and German economies contracted, they depressed other economies through the mechanism of the gold standard. These countries reduced their imports as they contracted, reducing exports from other countries. They also reduced their capital exports or increased their capital imports in response to the tight credit conditions at the end of the 1920s.

A bad recession turned into the Great Depression in the summer and autumn of 1931. A series of currency crises led both to what we now regard as perverse policy responses and to failures of financial institutions. A warning came in May 1931 when the main bank of Austria, the Credit Anstalt, failed, taking the Austrian schilling with it. This was a preview of things to come, but not a cause of them. The German mark had been under pressure since the German recession began in the late 1920s and the Weimar government began to run increasingly large deficits. They were covered by foreign lending, of which the American Young Plan was the most famous. The Weimar government, however, scared its foreign creditors by a series of statements for domestic consumption about a customs union with Austria and a possible repudiation of First World War reparations. The Reichsbank lost reserves precipitously in late May, and free trading in the mark was suspended in July 1931.

The British government found itself in similar trouble as its deficits followed Germany's. The Bank of England, unwilling to raise the bank rate above six per cent and further depress the domestic economy, abandoned the gold standard, floated the pound, and devalued in September 1931. The Federal Reserve, facing similar problems and adverse speculation, chose to raise its discount rate by 200 basis points in October 1931. This dramatic measure saved the dollar but killed the domestic economy. It was, however, loudly applauded by the American financial community as the correct gold-standard action.

The effects of fixed exchange rates can be seen in a comparison of Figures 1 and 2. Figure 1 shows that industrial production in four major countries declined at quite different rates. Figure 2 shows that the rate of decline in prices in the same four countries was strikingly similar. The fixed exchange rates of the gold standard led to uniform changes in prices even though other factors affected the change in production. The standard deviation of price changes was smaller than the standard deviation of production changes for 21 countries on the gold standard in 1930–2, as shown in Table 2. The standard deviation of price changes was smaller than the standard deviation of changes in the industrial production index in each year, even though the standard deviation of both series rose in 1932 as some countries abandoned gold. The final row of Table 2 shows the standard deviations in 1932 for seven countries that stayed on gold in 1932. Even though data for these countries are indistinguishable from the rest of the sample in 1930 and 1931, they are far more uniform in 1932.

Standard deviation of changes in 21 gold-standard countries, 1930–1932

Year	Prices	Industrial production
1930	0.037	0.081
1931	0.055	0.078
1932	0.090	0.123
1932*	0.035	0.039

*Seven countries still on gold in 1932.

Sources: Bernanke and James (1991); Temin (1993).

No country on the gold standard, however large, could escape the discipline of this harsh regime in the depression. In almost all cases, deflation was accompanied by depression as declining aggregate demand moved countries down upward-sloping aggregate supply curves. Banking systems in many gold-standard countries collapsed under this deflationary pressure, further reducing economic activity. The Federal Reserve sharply raised the US discount rate in October 1931 in response to a threatened outflow of gold, even though the US economy was contracting rapidly and had massive gold reserves. The primary transmission channel of the Great Depression was the gold standard.

It follows that abandoning the gold standard was the only way to arrest the economic decline. Going off gold severed the connection between the balance of payments and the domestic price level. Countries could lower interest rates or expand production without precipitating a currency crisis. Changes in the exchange rate rather than changes in domestic prices could eliminate differences between the level of domestic and foreign demand without a painful deflation. Any single devaluation could beggar neighbours under some conditions, but universal devaluation would have increased the value of world gold reserves and allowed worldwide economic expansion.

Great Britain abandoned the gold standard in September 1931 after a speculative attack on the pound prompted by bad budgetary news and by contagion from the German currency crisis of July 1931. Great Britain and the countries that followed Britain off gold were not large enough for their actions to arrest the world decline, and they were criticized at the time for abandoning gold; but the world would have been far better off if others had followed them off gold.

Even in the United States, with its vast economic resources and gold reserves, going off gold was a necessary prerequisite for economic expansion. Great Britain avoided the worst of the Great Depression by going off gold in 1931, as shown in Figure 1. Spain avoided the depression by never being on the gold standard; Japan by a massive devaluation in 1932. At the other extreme, the members of the gold bloc led by France endured contractions that lasted into 1935 and 1936. The single best predictor of the severity of the depression in different countries is how long they stayed on gold. The gold standard was a Midas touch that paralysed the world economy.

Real wages stayed high in countries on the gold standard. Macroeconomic policies to preserve the value of the currency reduced prices faster than wages, and real wages stayed high or even rose. Bank failures also were widespread in gold-standard countries, further depressing production. Both high real wages and bank failures show up as explanatory variables for low incomes around 1935, and the prevalence of financial crises in countries on gold suggests that a counterfactual with more rapid deflation and no devaluations would not have resulted in the maintenance of something close to full employment.

Complications

The influence of the gold standard determined the spread and the depth of the Great Depression, but the story has many dimensions not captured in this stark description. The literature can be contentious, although apparently competing views may represent elements in a more comprehensive view.

One view of the Great Depression sees it as an American contraction that was transmitted to the rest of the world. In *A Monetary History of the United States, 1867–1960* (1963), Milton Friedman and Anna Schwartz argued that the Federal Reserve System in the United States acted with such ineptness that it plunged the world into depression. They attributed this incompetence to the death of Benjamin Strong (president of the Federal Reserve Bank of New York) in 1928, and they describe several alternative monetary policies that they argue would have eased or even eliminated the economic contraction.

Even their story cannot separate the United States from the rest of the world, however. The Federal Reserve raised interest rates in October 1931 to defend the dollar, as noted above, even though the economy was contracting. Friedman and Schwartz characterized this action as an inept mistake, but they acknowledged the power of the gold standard to unite the financial community behind this perverse policy. This contractionary policy in the midst of rapid economic decline was the classic central bank reaction to a gold-standard crisis.

Charles Kindleberger put forward a more international explanation in *The World in Depression* (1986). He argued that the lack of central bank leadership in the operation of the restored gold standard was key to the spread of the Great Depression: the proposition summed up in the phrase ‘no longer London, not yet Washington’. The diminished financial status of Great Britain meant that London was unable to act as sole conductor of the international orchestra – or, in more modern terminology, to operate as the ‘hegemon’ – while the United States was not yet willing to take over this role despite the enormous improvement in its international economic standing.

Another factor which has been put forward as the primary explanation for the problems of the interwar period is the absence of international cooperation between the United States, Britain, France and Germany. Barry Eichengreen, in *Golden Fetters* (1992), identified this behaviour as a central feature of the period, manifest particularly in the attempt of each of

the main powers to secure for itself a disproportionate share of the world's limited stocks of monetary gold. Prior to the collapse of the gold standard in 1931 their non-cooperative behaviour involved the imposition of tight monetary policies not only by countries in deficit, but also by those – notably the United States and France – which were in surplus. This added to the deflationary pressures on the world economy and increased the vulnerability of the weak currencies, such as the pound and the mark, to speculative attack.

Recovery

The world began to recover from the contraction in 1933, when the United States and Germany both abandoned the policies of the gold standard, but the Great Depression was far from over. Unemployment continued to be high in most countries, as indicated in Table 1. The world economy split up into competing currency and trading blocs, and domestic policies to combat the hardships of depression changed the role of government.

Unemployment continued to be high in most countries throughout the 1930s. Measures designed to help workers often perpetuated unemployment. The National Industrial Recovery Act of 1933 in the United States attempted to bring order to industries and income to workers by allowing industries to enforce codes of conduct that raised both prices and wages. Rising wages impeded the extension of employment, trading off the benefits to the unemployed for benefits to those working. Germany under the Nazis expanded government spending and, apparently, decreased unemployment dramatically. France and other members of the gold bloc continued to maintain contractionary policies in an effort to retain the convertibility of their currencies into gold. Only when France devalued in 1936 could its recovery begin.

Recovery, however slow and halting, did not approach the status quo ante. The world economy fragmented in the 1930s, and recovery took place within relatively isolated currency and trading blocs. The United States began the process of reducing world trade with the Smoot–Hawley tariff of 1930. The United Kingdom abandoned its tradition of free trade in 1932 in favour of protection for the British Commonwealth. Germany under the Nazis adopted a complex set of bilateral trading arrangements that reoriented its trade towards south-eastern Europe. International trade was much reduced, and international capital flows virtually disappeared.

Countries were changed internally as well. Governments became active in the economy as they attempted to reduce unemployment or ease the condition of the unemployed. Unions grew in many countries, helped both by legislation and by unemployment. Regulation grew as governments substituted direct controls for those of the market, and the world war that followed the Great Depression caused governments to take control even more firmly of their economies. The mixed economies and large governments that were typical of the last half of the 20th century were the legacy of the Great Depression and its aftermath.

It is not possible to separate the long-run effects of the Depression from those of the Nazis and the Second World War, but it is instructive to ask whether the Great Depression could have been avoided. There were indeed stresses on the world economy at the end of the 1920s, and the control mechanisms used in earlier times were not in good shape. The downturns in the United States and Germany would have produced a serious recession in the early 1930s in any case. The currency crises of 1931 then turned this recession into the Great Depression. If Germany and the United States had abandoned gold after Britain had chosen devaluation over further contraction, the world economy would have begun to recover two years earlier and before unbearable strain had been put on economic and political institutions.

Historians today debate how much freedom policymakers had in 1931. The German cabinet discussed devaluation after Britain left gold, but the memory of hyperinflation less than a decade before inhibited – if it did not preclude – an expansionary policy such as devaluation. The United States was not under the same economic pressure as Germany, but the Federal Reserve nonetheless raised interest rates sharply in late 1931 in response to gold outflows following Britain's devaluation. The Federal Reserve was following the dictates of the gold standard in actions that were applauded by the local financial community. It was a world tragedy – one that escalated from economics to politics and war – that the hold of the gold standard was so strong in the early 1930s that policymakers in the major economies chose to continue deflationary economic policies long after the need for expansionary measures was clear.

See Also

- gold standard
- Kindleberger, Charles P

Bibliography

Balderston, T. 1993. *The Origins and Course of the German Economic Crisis*. Berlin: Haude and Spener.

Bernanke, B. 1983. Nonmonetary effects of the financial crisis on the propagation of the Great Depression. *American Economic Review* 73, 257–76.

Bernanke, B. 1995. The macroeconomics of the Great Depression: a comparative approach. *Journal of Money, Credit and Banking* 27, 1–28.

Bernanke, B. and James, H. 1991. The gold standard, deflation, and financial crisis in the Great Depression: an international comparison. In *Financial Markets and Financial Crises*, ed. R.G. Hubbard. Chicago: University of Chicago Press.

Borchardt, K. 1991. *Perspectives on Modern German Economic History and Policy*. Cambridge: Cambridge University Press.

Bordo, M., Goldin, C. and White, E.N. 1998. *The Defining Moment: The Great Depression and the American Economy in the Twentieth Century*. Chicago: University of Chicago Press.

Calomiris, C.W. 1993. Financial factors in the Great Depression. *Journal of Economic Perspectives* 7(2), 61–85.

Calomiris, C.W. and Mason, J.R. 2003. Consequences of bank distress during the Great Depression. *American Economic Review* 93, 937–47.

Choudhri, E.U. and Kochin, L.A. 1980. The exchange rate and the international transmission of business cycle disturbances: some evidence from the Great Depression. *Journal of Money, Credit, and Banking* 12, 565–74.

Eichengreen, B. 1992. *Golden Fetters: The Gold Standard and the Great Depression*. New York: Oxford University Press.

Eichengreen, B. 2004. Still fettered after all these years. *Canadian Journal of Economics* 37, 1–27.

Eichengreen, B. and Sachs, J. 1985. Exchange rates and economic recovery in the 1930s. *Journal of Economic History* 12, 925–46.

Eichengreen, B. and Temin, P. 2000. The gold standard and the Great Depression. *Contemporary European History* 9, 183–207.

Feinstein, C.H., Temin, P. and Toniolo, G. 1997. *The European Economy between the Wars*. New York: Oxford University Press.

Field, A.J. 1984. Asset exchanges and the transactions demand for money, 1919–29. *American Economic Review* 74, 43–59.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press.

Hamilton, J.D. 1988. The role of the gold standard in propagating the Great Depression. *Contemporary Policy Issues* 6, 67–89.

James, H. 1986. *The German Slump: Politics and Economics, 1924–1936*. Oxford: Oxford University Press.

Kindleberger, C.P. 1986. *The World in Depression, 1929–1939*. Rev. edn. Berkeley: University of California Press.

Margo, R.A. 1993. Employment and unemployment in the 1930s. *Journal of Economic Perspectives* 7(2), 41–59.

Mouré, K. 1991. *Managing the Franc Poincaré: Economic Understanding and Political Constraint in French Monetary Policy, 1928–1936*. Cambridge: Cambridge University Press.

Romer, C.D. 1990. The Great Crash and the onset of the Great Depression. *Quarterly Journal of Economics* 105, 597–624.

Romer, C.D. 1993. The nation in depression. *Journal of Economic Perspectives* 7(2), 19–39.

Rothermund, D. 1996. *The Global Impact of the Great Depression, 1929–1939*. London: Routledge.

Temin, P. 1989. *Lessons from the Great Depression*. Cambridge MA: MIT Press.

Temin, P. 1993. Transmission of the Great Depression. *Journal of Economic Perspectives* 7(2), 87–102.

How to cite this article

Temin, Peter. "Great Depression." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000194> doi:10.1057/9780230226203.0671

The New Palgrave Dictionary of Economics Online

Great Divide

Erik Berglöf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

commitment; compensation; crowding out; fiscal responsibility; Great Divide; policy reform; privatization; rule of law; transition

Article

The expression ‘Great Divide’ refers to the differences in institutional development apparent in the first decade of transition among the countries of central and eastern Europe and former Soviet Union following the collapse of Communism (Berglöf and Bolton, 2002). Figure 1 compares these countries in 1996, 2000 and 2005 using one measure of rule of law derived through questionnaires to businesses operating in a broad range of countries.
Figure 1
Rule of law (world, 2005). *Source*: Kaufmann, Kraay and Mastruzzi (2006).

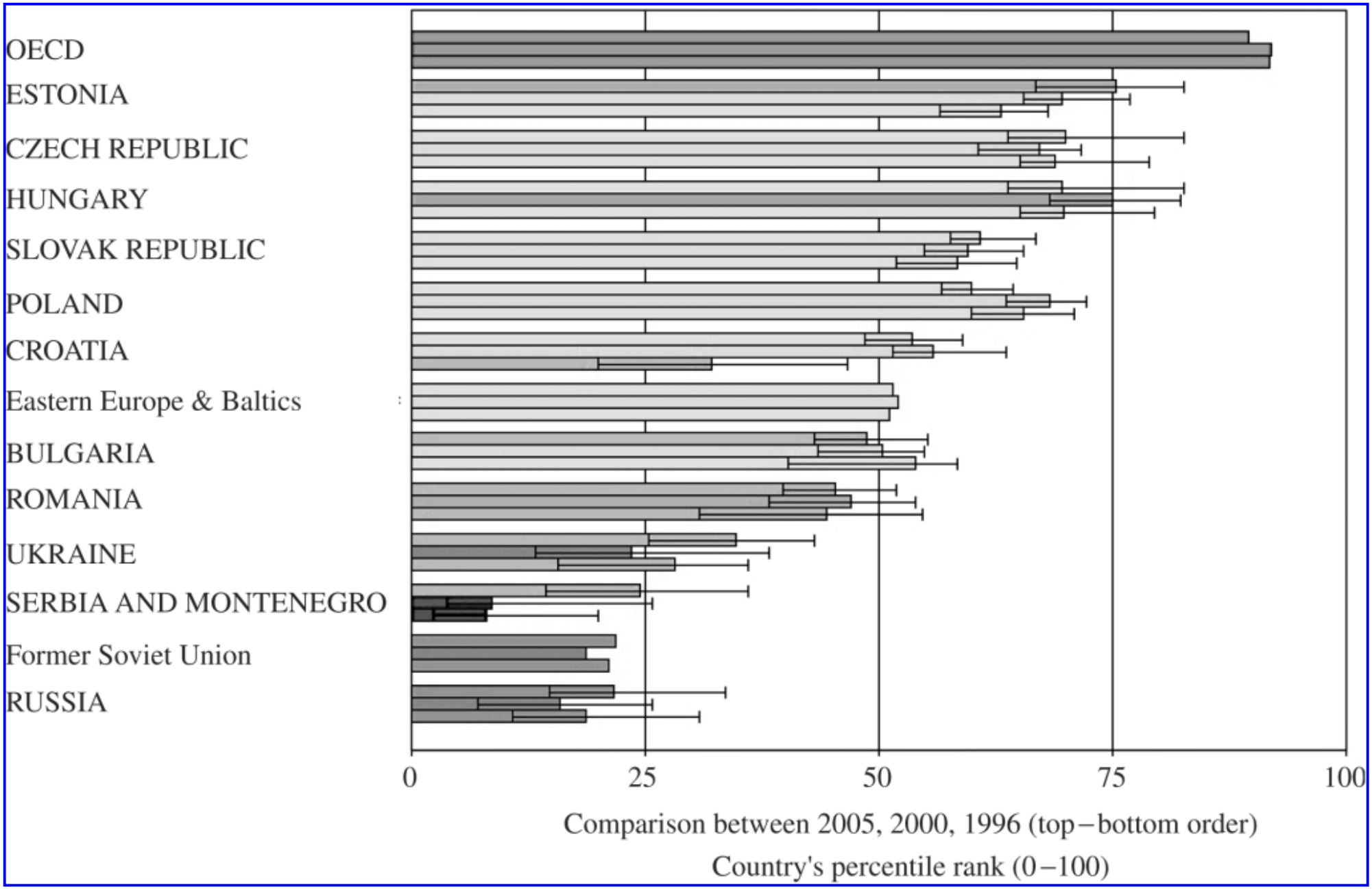

By the end of the first decade of transition in 2000, the Great Divide was visible in almost every measure of economic performance: gross domestic product (GDP) growth, investment, government finances, growth in inequality, general institutional infrastructure, and in measures of financial development. Gradually growth has picked up and macro-stability improved in the Commonwealth of Independent States countries, in large part thanks to favourable terms-of-trade changes as a result of price increases in energy and other raw materials. As indicated by Figure 1, institutional development has also progressed in most of these countries but still lags significantly on most dimensions, and there are examples of institutional regression, particularly in political institutions.

The Great Divide does not primarily refer to the depth and duration of the initial transitional recession, but rather to the more long-term institutional backwardness observed in parts of the region. Some countries (for example, Estonia, Latvia and Lithuania) had a deeper recession, but a less protracted turnaround accompanied by rapid institutional transformation.

Other countries (for example, Belarus, Turkmenistan, and Uzbekistan) did not attempt full-scale reform, and thus experienced neither initial decline nor substantial development in their institutions. The non-European transition countries China and Vietnam exhibit yet another pattern where broad, if incremental and partial, economic reform did not result in initial output decline.

The Great Divide represents one of the key puzzles of transition in that it cannot be immediately traced to the policies pursued. In fact, differences in policies among the more successful countries in Central and Eastern Europe were often more pronounced than those between some of these countries and those of former Soviet Union. A prominent, but by no means the only, example would be privatization policies, where the Czech model had more in common with that of Russia than with that of Poland.

Another, perhaps less puzzling, observation is the remarkable institutional convergence of economic systems in Central and Eastern Europe despite the diversity in terms of policies pursued. The emerging model of ownership and control of large firms is one with an owner with a large controlling share and a strong presence in day-to-day management of the firm. The financial systems are strongly dominated by commercial banks, increasingly foreign-owned, whereas stock markets on the whole remain volatile and illiquid. This convergence has been taking place even though the countries differ markedly in terms of policies when it comes to areas such as enterprise and bank privatization, bank recapitalization, stock market policies, and entry and exit of firms.

Both these observations, the emergence and persistence of the Great Divide and the convergence of economic systems in the front-runner countries in central and eastern European countries, suggest that initial conditions matter greatly for institutional development and economic growth. The relative importance of different initial conditions has been estimated by a large number of studies, but the influences of individual factors are often hard to disentangle. Broadly speaking, the Soviet legacy (the degree of integration into the economic and political system of the Soviet Union) and the prospect of membership in the European Union stand out as key in shaping the development of economic and political institutions.

Another key to understanding the origin of the Great Divide is to look at when and under what conditions it emerged. Typically, the differences in institutional development first became visible when the governments were faced with demands from different groups to be compensated for the adjustments in relative prices following pricing reforms. The ability of governments of transition economies to achieve fiscal and monetary responsibility, together with a commitment to refrain from bailing out failing banks or loss-making enterprises, determined whether economic and financial development took off.

Fiscal responsibility promotes both financial development and economic growth through two important channels: it limits the extent of crowding out of private investment by government borrowing and it makes credible the commitment of the government to maintain the macro-stability essential for private investment. In addition, it provides some guarantees that the returns from investment are not going to be taxed away in the future by excessively profligate governments desperately seeking tax revenues where they can find them. Of course, specific initial conditions and underlying country characteristics facilitate the emergence of fiscally sound governments capable of enforcing the rule of law.

See Also

- emerging markets
- financial structure and economic development
- privatization
- transition and institutions

Bibliography

Berglöf, E. and Bolton, P. 2002. The Great Divide and beyond – financial architecture in transition. *Journal of Economic Perspectives* 16(1), 77–100.

Kaufmann, D., Kraay, A. and Mastruzzi, M. 2006. *Governance Matters V: Governance Indicators for 1996–2005*. Washington, DC: World Bank.

How to cite this article

Berglöf, Erik. "Great Divide." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000189> doi:10.1057/9780230226203.0674

The New Palgrave Dictionary of Economics Online

green national accounting

Sjak Smulders

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Extending conventional national product measures, green national accounting provides better indicators of economic welfare and of the sustainability of welfare levels. The main theoretical result shows that in an undistorted economy net national product is proportional to welfare, provided some rather stringent conditions are met. With appropriately used shadow prices, the welfare effects of externalities and world market changes can be accounted for and sustainable income – the hypothetical level of consumption that can be sustained into the future – can be calculated. Practical approaches have been proposed to adjust conventional national income figures roughly in the spirit of the theoretical results.

Keywords

air quality; climate change; contingent valuation; ecological footprint; education; environment; genuine savings; green national accounting; green net investment; green net national product; hedonic prices; household work; index of sustainable economic welfare; inequality of income; intergenerational income distribution; interpersonal utility comparisons; land use; leisure; maximin preferences; national accounting; natural resources; non-renewable resources; opportunity cost; other-regarding preferences; pollution; representative agent; resource depletion; shadow prices; social welfare function; sustainability; sustainability accounting; sustainability gap; technical change; travel costs; utilitarian preferences; utility aggregation; value of time; wealth

Article

Green national accounting extends conventional national product measures to provide better indicators of economic welfare, as well as indicators of the degree to which welfare levels can be sustained. Conventional national accounts measure the size of the market or commercial activities, but do not necessarily measure very well (*a*) how these activities translate into welfare and (*b*) how non-marketed activities goods contribute to the welfare of citizens. A big part of the greening of national accounts

concerns issues related to the environment. For example, production of certain goods that generate market value contributes to national income, but, if the production generates undesirable pollution as a by-product, the contribution to welfare might be actually negative. Conventional national accounts ignore the reduction in air quality since air quality is not traded on markets and is left out from conventional national accounts. As another example, consider the depletion of oil reserves. Oil companies' profits contribute to conventional national income, but the fact that fewer reserves will be available in future is left unaccounted, even though this may affect the welfare of future generations.
The aim of measuring economic welfare by a simple number is without doubt challenging, since it is inevitably related to the formal economic concept of individual utility and the theoretical problems of aggregating utility and interpersonal utility comparisons. The theoretical approaches to green accounting circumvent these problems by assuming a social welfare function and representative agents in highly stylized models. Because of lack of data and various other constraints, the more applied approaches to green accounting are only loosely rooted in formal theory and sometimes include issues of (intergenerational) income distribution on an ad hoc basis.
The focus of green accounting is much more on dynamic and intergenerational aspects than on intra-generational issues. It typically tries to measure to what degree the lifetime utility of the representative agent in a country increases over time, and to what degree it is higher than in other countries. It sometimes also tries to measure to what degree intra-temporal levels of utility of the representative agent in a country can be maintained over time, whether the economy is investing enough to maintain non-decreasing utility levels, and how much a country can consume more than another country when both of them would ensure utility levels of their inhabitants are not declining over time. The latter type of questions is often associated with ‘sustainability accounting’ as a particular branch of green accounting.
Green accounting starts from a broad concept of economic welfare, which goes beyond welfare depending on just marketed produced goods. Thus, welfare is allowed to depend on health, environmental amenities, pollution levels, or availability of natural resources. Even altruistic preferences are allowed: utility levels of future generations may matter for the welfare objective of current generations. Society might care in particular for the utility levels of those generations that are worse off in future. The extreme case of this implies *maximin preferences*: only the generation with lowest utility levels gets a positive weight in the social welfare function and reductions in other generations' welfare do not count. This contrasts to *utilitarian preferences*, in which every generation gets a weight, and utility levels of generations further in future often get a lower weight because of a positive utility discount rate.

Welfare in an undistorted economy

A fundamental theoretical result concerns the measurement in undistorted competitive (and hence by construction welfare-maximizing) economies (Weitzman, 1976; 2003), which we will refer to as the *Weitzman principle*. Consider a society that manages to maximize its own social welfare function, either because there are no externalities or because all existing externalities are internalized by appropriate policy. In such an economy, green net national product (NNP) can be calculated and this measure is proportional to total welfare. Moreover, green net investment can be calculated and a positive (negative) value of this measure always implies an increase (decrease) in instantaneous welfare.
To start with, we consider the simplest model economy (as in Weitzman, 1976), in which a single

consumption good is produced from a single capital good, representative agents maximize utility discounted at a constant rate, and the social welfare function is the sum of individual utilities. Then the sum of the value of consumption and net investment, valued against market prices, is proportional to welfare. Note that this sum is equal to the conventional NNP number for this economy. In this model economy, NNP reflects intertemporal welfare: consumption reflects instantaneous utility, whereas investment reflects how current economic activity contributes to future utility.

Extending conventional income to ‘green’ income is needed if the welfare function has ‘unconventional’ arguments like environmental quality and health (see, for example, Asheim and Buchholz, 2004). These arguments can be seen as alternative forms of consumption, not consumption of conventionally produced goods but of natural resource services, health services, and so on. Then, in a perfect economy, according to the Weitzman principle, green NNP should be calculated as the value of all ‘consumption’ activities that matter for utility plus the value of net investment in all ‘capital’ stocks that matter for production capacity. Both ‘consumption’ and ‘capital’ are broad comprehensive measures here, with the former including for example consumption of environmental resource services and the latter including any variable that determines the production capacity for the generation of comprehensive consumption. Accordingly, the relevant capital goods, or assets, include not only physical capital and resource stocks, but also all kind of other state variables like public health (in economies that have a preference for health or in which health determines workers productivity), atmospheric pollution stocks, physical characteristics of the soil determining absorption of pollution and regeneration of nature, institutional capital and social norms subject to erosion and development.

As a special result of the Weitzman principle, the *Hartwick rule* can be derived (Hartwick, 1977). The rule says that, if society wants to maintain a constant utility level over time, it has to invest the returns to non-renewable resources in other assets such that total green net investment (the comprehensive measure of investment) is zero. Hence, in accordance with the Weitzman principle, in such an economy green NNP equals the sustainable level of utility (as well as green consumption, since green net investment is zero). As a result it is a measure of sustainability: comparing two different economies that are both undistorted and maximize maximin preferences, we can say that the country with higher green NNP can maintain indefinitely a higher welfare level than the other.

Caveats

The above results must be interpreted with care and are more limited than might seem at first sight. The important caveat is that an economy with zero green net investment is not necessarily able to maintain a constant utility level for its representative agent over time and can thus be unsustainable (Asheim, Buchholz and Withagen, 2003). This may be the case, for example, in an economy that is dependent on a non-renewable resource and that is maximizing a utilitarian welfare function with constant discount rate rather than a maximin welfare function. Such an economy might *optimally* consume growing amounts initially, but eventually consume declining amounts, which of course is always inefficient for maximin preferences. The key insight from the Hartwick rule is that a necessary, but not sufficient, condition for maintaining constant welfare over time is sufficient investment (Pezzey, 2004).

The Weitzman principle applies only when changes in production capacity depend solely on investment choices. Alternatively, production capacity, that is, the possibility to generate consumption, might

change over time due to events beyond control of the economy. Three important examples are exogenous technological change, world price changes, and geological or climatic changes. If technology improves over time or the world market price of export goods increases, an economy's welfare can improve even when green net investment (as defined above) is negative. Formally, welfare is now proportional to the sum of comprehensive consumption and net investment, augmented with a term capturing the (properly valued and discounted) benefits from improved technology and world market prices that will accrue to the economy in future (Sefton and Weale, 1996). The latter term can be labelled the ‘value of time’, whereas the sum of green NNP and the value of time is ‘augmented NNP’ (Pezzey and Toman, 2002). In the case of global warming or other negative environmental developments because of purely geological or climatic reasons, there might be a negative time premium and sufficiently positive green net investment is needed to keep welfare constant.

Externalities and sustainable income

The Weitzman principle is derived for welfare-maximizing economies. How can we measure welfare in economies that do not actually maximize welfare? Similarly, how can we measure sustainable income (consumption) levels in economies that do not actually sustain constant consumption levels?
One theoretically possible way is to construct a hypothetical income figure that represents the level of welfare or sustainable consumption, respectively, that would arise if the economy made the switch from being distorted to being welfare maximizing or sustainable, respectively. This requires a measurement of the total *wealth* of the economy, consisting of all assets valued at their corresponding shadow prices, which depend on the exact welfare function.
The problem when putting this into practice is that actual prices observed in the distorted economy have no direct relation to the shadow prices needed to calculate wealth. In the presence of externalities, market prices do not reflect certain social costs and benefits, so that the sum of consumption and net investment at market value misses some contributions to welfare. As an alternative to calculating wealth against shadow prices as indicated above, one may augment the market value of comprehensive consumption and net investment with the net present value of the marginal externality along the competitive path (Aronsson, Löfgren and Backlund, 2004). Obviously, such an augmentation term in green accounting is also hard to calculate in practice.

Green national accounting in practice

Given these theoretical results, how feasible is welfare and sustainability measurement in practice? The results suggest that we need (*a*) comprehensive accounting of consumption and investment activities, (*b*) the right (shadow) prices, and (*c*) additional forward-looking augmentation terms to capture exogenous or uninternalized developments over time. It has been concluded that (*b*) and (*c*) are insurmountable impediments to practical green accounting: fully correct green accounting is impossible and any method ignoring the problems associated with (*b*) and (*c*) is bound to produce biased numbers. At the other side of the debate it has been argued that national accounting has always been imperfect and indicative only (Cairns, 2002). According to this view the task is to focus on making national accounts more comprehensive – and satisfy at least requirement (*a*) – carefully delineating consumption and net

investment, and applying corrections to prices where reasonable and feasible. Furthermore, the value of non-marketed activities has to be imputed rather than observed. A host of methods is available to impute prices, which use hedonic pricing, contingent valuation methods, and travel cost approaches.

The resulting practical approaches differ with respect to what to include in national accounts and how to determine values and prices. Among them ‘genuine savings’ (GS), a comprehensive net investment measure, is the best-known and closest to theory (Pearce and Atkinson, 1993). Most GS correct conventional measures of investment for consumption of resources, damages from pollution, and investment in education. GS is often found to be positive in Europe and Japan (thanks to high savings and investment in education) and negative for Africa and oil-producing countries (due to the depletion of oil reserves). The latter results are quite sensitive to the way resource depletion is accounted for.

GS figures should be interpreted with care: since the GS calculation ignores ‘value of time’ terms and uses market prices rather than shadow price, GS is not a true measure of welfare increases. Persistent negative rates of GS are likely to result in decreases in welfare (unless exogenous technological change is significant), but with positive rates of GS nothing definitive can be said. So GS can be used to measure only unsustainability.

Another well-known indicator is the index of sustainable economic welfare (ISEW, initiated by Daly and Cobb, 1989). It is an extended measure of green NNP (and therefore aimed at measuring welfare) that starts from conventional income, adds changes in environmental quality, imputes value of non-marketed activities (particularly household work), subtracts consumption expenditures that do not directly contribute to welfare (such as health and pollution abatement expenditures), and weighs on an ad hoc basis remaining expenditures by a measure of income inequality. For different components different valuation methods are used so that consistency is not always guaranteed. Shadow prices or opportunity costs are rarely used to value damages. The large number of adjustment to NNP also raises the question why certain components are still omitted (for example, the value of leisure time is not included while household work is, and investment in education and technological change are omitted). Calculations of ISEW for richer countries show that ISEW has grown considerable more slowly than conventional NNP over recent decades. This result is not robust, however, for changes in the specific composition of the index and the valuation assumptions (Neumayer, 2003).

Because of the main problem of determining correct prices for non-marketed goods and for dealing with externalities, but also because of the limited substitution between natural resources and conventional man-made inputs, it is often argued that purely physical indicators are useful to measure the state of the environment and economic welfare and to supplement (rather than adjust) more conventional measures. Since no attempts are made to monetize and the common denominator is missing, different physical indicators cannot be easily aggregated to overall welfare: improvement in one indicator cannot be compared with gains elsewhere and the costs of securing improvements cannot be determined.

The *sustainability gap* indicator is an example (Ekins and Simon, 1999). Since welfare is critical dependent on air quality, a minimum level of air quality can be defined that is needed to maintain welfare at a reasonable level, and it can be measured how far society is from this standard; the exercise can be repeated for other critical natural resources.

Another popular example is ‘ecological footprint’, which measures the amount of land that is needed to generate the consumption of a country, including the land needed to assimilate the waste generated and undo climatic change from carbon dioxide emissions by means of carbon sequestration (Wackernagel

and Rees, 1996). If for the world as a whole the ecological footprint exceeds available land, the economy is said to be unsustainable. A problem here is how to aggregate over different land uses. Similar measures, with a similar aggregation problem, keep track of varieties of material resource flows. It is unlikely that the theory of green accounting can in the end be fully applied. Instead, a combination of different indicators and imperfect theory-based measures of welfare could – together with the caveats from theory – provide a useful information system to put conventional national income systems into proper perspective.

See Also

- externalities
- intertemporal equilibrium and efficiency
- shadow pricing
- social discount rate
- sustainability

Bibliography

Aronsson, T., Löfgren, K. and Backlund, K. 2004. *Welfare Measurement in Imperfect Markets: A Growth Theoretical Approach*. Cheltenham and Northampton: Edward Elgar Publishing.

Asheim, G. and Buchholz, W. 2004. A general approach to welfare measurement through national income accounting. *Scandinavian Journal of Economics* 106, 361–84.

Asheim, G., Buchholz, W. and Withagen, C. 2003. The Hartwick rule: myths and facts. *Environmental and Resource Economics* 25, 129–50.

Cairns, R.D. 2002. Green accounting using imperfect, current prices. *Environment and Development Economics* 7, 207–14.

Daly, H. and Cobb, J. 1989. *For the Common Good: Redirecting the Economy Toward Community, the Environment and a Sustainable Future*. Boston, MA: Beacon.

Ekins, P. and Simon, S. 1999. The sustainability gap: a practical indicator of sustainability in the framework of the national income accounts. *International Journal of Sustainable Development* 2, 24–58.

Hartwick, J. 1977. Intergenerational equity and investing rents from exhaustible resources. *American Economic Review* 67, 972–74.

Neumayer, E. 2003. *Weak Versus Strong Sustainability: Exploring the Limits of Two Opposing Paradigms*, 2nd rev. edn. Cheltenham and Northampton: Edward Elgar Publishing.

Pearce, D. and Atkinson, G. 1993. Capital theory and the measurement of sustainable development: an indicator of ‘weak’ sustainability. *Ecological Economics* 8, 103–8.

Pezzey, J. 2004. One-sided sustainability tests with amenities, and changes in technology, trade and population. *Journal of Environmental Economics and Management* 48, 613–31.

Pezzey, J. and Toman, M. 2002. Progress and problems in the economics of sustainability. In *The International Yearbook of Environmental and Resource Economics 2002/2003*, ed. T. Tietenberg and H. Folmer. Aldershot: Edward Elgar Publishing.

Sefton, J. and Weale, M. 1996. The net national product and exhaustible resources: the effects of foreign trade. *Journal of Public Economics* 61, 21–47.

Wackernagel, M. and Rees, W. 1996. *Our Ecological Footprint: Reducing Human Impact on the Earth*. Gabriola Island, BC: New Society Publishers.

Weitzman, M. 1976. On the welfare significance of national product in a dynamic economy. *Quarterly Journal of Economics* 90, 156–62.

Weitzman, M. 2003. *Income, Capital and the Maximum Principle*. Cambridge, MA: Harvard University Press.

How to cite this article

Smulders, Sjak. "green national accounting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000196> doi:10.1057/9780230226203.0675

The New Palgrave Dictionary of Economics Online

Gresham, Thomas (c1519 – 1579)

Eleanor G. Powell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Gresham T.

Article

The second son of Sir Richard Gresham, merchant, Sir Thomas Gresham was educated at Gonville Hall, Cambridge, apprenticed to his uncle Sir John Gresham, also a merchant, and admitted a member of the Mercers' company in 1543. In 1551 or 1552 he became royal agent or king's factor at Antwerp, in which post he received 20 shillings a day, and which he retained with few intervals during three reigns until 1574, employed in spite of his Protestant views even by Mary. His business was to negotiate royal loans with Flemish merchants, to buy arms and military stores, and to smuggle into England as much bullion as possible. He succeeded in raising the rate of exchange from 16•s. to 22•s. in the £, and is said to have saved in this way 100,000 marks to the crown and 300,000 to the nation. His operations greatly benefited English trade and credit, though the government could not be induced to pay its debts as punctually as Gresham would have liked. He did not hesitate to remonstrate with and advise Elizabeth and Cecil; but he was so useful and trustworthy that he was never seriously out of favour, except just after Mary's accession. On Mary's death he advised Elizabeth to restore the base money, to contract little foreign debt, and to keep up her credit, especially with English merchants. Later he taught her how to make use of these English merchants when political troubles in the Netherlands curtailed her foreign resources; at his suggestion the Merchant Adventurers and Staplers were forced by detention of their fleets to advance money to the state; but as they obtained interest at 12 per cent instead of the legal maximum of 10, and the interest no longer went abroad, the transaction proved advantageous to all parties and increased Gresham's favour. His journeys to and from Antwerp were very frequent, but in his later years he entrusted most of his public work to his agent, and is not known to have been at Antwerp after 1567. In 1554 he was sent to Spain to procure bullion, a very difficult task in which he was only partially successful; and in 1559 he was employed as ambassador to the Duchess of Parma, regent of the Netherlands; it was on this occasion that he was knighted.
In addition to his public services he continued throughout his life to do the work of ‘the greatest

merchant in London’. He was, in the language of the day, a banker and goldsmith, with a shop in Lombard Street, as well as a mercer; but he was a considerable country gentleman besides, with estates, chiefly in Norfolk, where his father held considerable property, and with several country houses besides the house in Bishopsgate which he built and bequeathed to London as Gresham College. He twice entertained Queen Elizabeth as his guest. His wealth was mainly earned by his private business, but he cannot be acquitted of enriching himself at the public expense by at least one dishonourable manœuvre; and he habitually forwarded his schemes by bribery. The money so gained he applied to public uses, his only son having died young: the foundation of the royal exchange, of Gresham College, and of eight almshouses, and the establishment of the earliest English paper-mills on his estate at Osterley, show the breadth of his interests, his liberality, his charity, his culture, and his commercial enterprise.

Bibliography

Anon. 1883. *A Brief Memoir of Sir Thomas Gresham, with an abstract of his will, and of the act of parliament for the foundation of Gresham's College*. London.

Bindoff, S.T. 1973. *The Fame of Sir Thomas Gresham*. London: Jonathan Cape.

Boorstin, D.J. 1980. *Gresham's Law: Knowledge or Information?* Washington, DC: Library of Congress.

Burgon, J.M. 1839. *Life and Times of Sir Thomas Gresham*, 2 vols. London.

de Roover, R.A. 1949. *Gresham on Foreign Exchange*. Cambridge, MA: Harvard University Press.

Fox Bourne, H.R. 1866. *English Merchants*. London.

Fuller, T. 1662. *Worthies of England*. London.

Hall, H. 1886. *Society in the Elizabethan Age*. London.

Holinshed, R. 1807–8. *Chronicles of England, Scotland and Ireland*, 6 vols. London.

Salter, R.F. 1925. *Sir Thomas Gresham*. London: Parsons.

Ward, J. 1740. *Life of Sir Thomas Gresham*. London.

How to cite this article

Powell, Eleanor G. "Gresham, Thomas (c1519–1579)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_G000113> doi:10.1057/9780230226203.0676

The New Palgrave Dictionary of Economics Online

Gresham's Law

François R. Velde
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article recounts the origin of Gresham's Law, discusses its theoretical and empirical problems, and presents several refinements that have been proposed.

Keywords

assignats; asymmetric information; bimetallism; exchange rate indeterminacy; good and bad money; Giffen, R.; Gresham, T.; Gresham's Law; legal tender; seigniorage

Article

Gresham's Law, which holds that ‘bad money drives out good money’, is as problematic as it is well-known. It predicts that, when two monies are in use but one is of lower quality or lower intrinsic value than the other, the former will be used as medium of exchange to the exclusion of the latter. The law implicitly relies on the monies circulating for the same value in spite of their intrinsic differences. The key question is: why would they?

Origin of Gresham's Law

Gresham's Law is one of the ‘laws’ bequeathed to us by 19th-century political economists eager to uncover laws of nature just like physicists. Henry Dunning Macleod (1855–56, vol. 2, p. xxxvi; 1858, p. 477; 1896, p. 38), the man who named it, described it successively ‘an unerring law of nature’, ‘a fundamental and universal law in Economies, which has been found to be true in all countries and ages’, and ‘this great Law, which is as well and firmly established as the Law of Gravitation’.
Perhaps it should have been named Macaulay's Law, for it was Lord Macaulay (1850–61, vol. 4, p. 620) who gave the law its familiar form: ‘where good money and bad money are thrown into circulation together, the bad money drives out the good money.’ Macaulay also noted that the phenomenon had

been described in the fifth century bc by the playwright Aristophanes, who in *The Frogs* (718–733) compared the prevalence of bad politicians with the replacement in use of high-quality gold and silver coinage by inferior copper coins. Macaulay dismissed the playwright's explanation based on preferences but thought the verses worth quoting in the original Greek. MacLeod quoted Aristophanes, too, albeit in English, but he decided to credit the first person he thought to have explained the law properly: Sir Thomas Gresham, a financier in 16th-century England. Gresham's letter to Queen Elizabeth I on her accession, urging her to restore a good quality coinage after the debasements of her predecessors, had recently been published by Burgon (1839, vol. 1, p. 483).

Wolowski (1864, p. lix) soon drew attention to earlier formulations of the law in monetary tracts of Nicholas Copernicus (written between 1519 and 1528) and Nicole Oresme (written between 1355 and 1357), although the relevant passage in Oresme's tract was later shown to be an anonymous addition of the late 15th century (Bridrey, 1906, pp. 263–5). Later, Fetter (1932) pointed out that Gresham never stated anything remotely approaching his law: the famous letter to Queen Elizabeth I merely stated that the price of the English pound sterling abroad had declined after Henry VIII's debasement of the silver currency and that gold had been exported on that occasion. The somewhat arid search for the first person to state the law is often led astray towards descriptions of the phenomenon in some particular instance (such as the aftermath of a debasement), or of a related phenomenon (such as clipping, culling, and exporting of good coins). On this score, Copernicus's claim of priority is solid, since he wrote as a general proposition that ‘introducing a new, worse money while the old, better one remains not only depreciates the latter, but, I would say, expels it’ (Wolowski, 1864, p. 56).

Be that as it may, MacLeod's coinage popularized by Jevons (1875) soon gained universal currency and, being taken at face value, this ‘law’ has been driving critical thinking out of circulation ever since.

The problems with Gresham's Law

In the usual laconic formulation of the law, ‘bad money drives out good money’, every single word cries out for clarification. What is ‘money’, what is ‘good’, what is ‘bad’, and what does ‘driving out’ mean? The nature of the law, empirical regularity or theoretical proposition, is just as uncertain.

The law, being so vague, has been applied to a wide variety of pairs of monies: freshly minted and worn or full-bodied and clipped versions of a given coin; original and debased version of a given coin; a coin and a similar but lighter coin of the same metal; coins of different metals; metallic and paper money; paper monies of different issuers; securities of varying risk characteristics; and so on. The process of ‘driving out’ is taken to mean that one money replaces the other in some monetary function either completely or partially, but also that the coins circulate alongside but not at par (Fetter, 1932). In formal models, Gresham's Law is often invoked to compare outcomes across equilibria (for some parameter values, two monies circulate, and for others only the bad one does) rather than the dynamic process suggested by the words ‘drive out’.

As a theoretical proposition, the law seems to run against basic economic intuition. Bad apples do not drive out good apples, they fetch a lower price. Why the good money could not circulate at a premium is a puzzle that the mere invocation of the law occults. Friedman and Schwartz (1963, p. 27 n.) note that the law is often misused because it ‘applies only when there is a fixed rate of exchange between the two [monies]’. This raises the question: what fixes the exchange rate? In this version, the law relies on a

postulate about prices, which economists are in the business of explaining.

As an empirical regularity, the law would have been more compelling if counter-examples had not been ignored as carefully as examples have been collected. One need only go to the original inventors of the law to find instances of selective empirics. Lord Macaulay invoked it as a general proposition to explain the state of the English coinage in 1695, a time when high-quality milled money was driven out by worn and clipped hammered money. He does not explain, however, why the law failed to apply in the previous 30 years throughout which milled and hammered coins coexisted (see Sargent and Velde, 2002, ch. 16). Sir Thomas Gresham himself was baffled by the fact that, after Henry VIII's debasements, testoons containing 40 grains of silver circulated alongside testoons containing 20 grains, and at the same value (De Roover, 1949, p. 93). Other counter-examples are easily found. High-quality currencies have dominated lower-quality competitors in international trade since the days of the Florentine florin and the Venetian ducat. Rolnick and Weber (1986) have documented numerous other violations of the law. Finally, the vast literature documenting collapses in real balances during hyperinflations surely testifies to the fact that (very) bad money will be driven out by almost anything else.

Refinements of Gresham's Law

Several refinements of the law have been proposed, and three are considered here.

One force that might fix the exchange rate is legal-tender laws. A law conferring legal tender status on a money states that a debtor is discharged of his debt by tendering that money (in the correct amount) to the creditor. It is up to the creditor to accept or refuse, but should he refuse he has no further legal recourse against the debtor. The argument that legal-tender laws can be sufficient to uphold Gresham's Law runs as follows. Suppose that two currencies, one intrinsically worth less than the other, are given equal legal tender by enforceable laws: say, one dollar coin is worth 90 per cent of the other in intrinsic content. Debtors will then discharge their debts with the bad money, and reserve the good money for other purposes. The bad money displaces the good one, at least in repayment of debts.

A first difficulty is that legal-tender laws typically apply to the discharge of debts, not to circulation in general. But – if we set this aside – while the law forces the creditor to take the bad dollar in payment, it usually does not force the debtor to tender only bad dollars, or the creditor to accept good dollars at face value. The debtor owing $100 could offer the creditor a choice between 100 bad dollars and 90 good dollars (on the assumption that the transactions costs of such an offer are not too high, as they might be if the relative price is an inconvenient fraction; see Rolnick and Weber 1986).

To shore up Gresham's Law requires unusually strong legal tender laws, such as the two examples given by Selgin (2003). In the first example, on 11 January 1776 the Continental Congress resolved to protect the paper money it was issuing, the continental, by declaring that

> if any person shall hereafter be so lost to all virtue and regard for his country, as to refuse to receive said bills in payment, or obstruct or discourage the currency or circulation thereof, and shall be duly convicted […] such person shall be deemed, published, and treated as an enemy of his country, and precluded from all trade or intercourse with the inhabitants of these colonies.

In the second example, the French revolutionary government defended its paper currency, the assignat, by decreeing on 11 April 1793 that trading specie for the assignat at anything but par, and setting or offering different prices for payment in assignats or specie, would be punished by six years of imprisonment. A law of 5 September 1793 added that such acts committed with counter-revolutionary intent were punishable by death.

Such examples, however, are too infrequent to sustain Gresham's Law with any kind of generality. They also tend to occur in periods of turmoil when the force of the legal-tender laws is questionable. The Continental Congress's resolution, which Nussbaum (1950, p. 567) finds ‘of dubious legal significance’, did not prevent the continental from quickly depreciating below par. In the French case, the assignat was already trading at less than half the price of gold coins, according to the Treasury's own records, and never traded above thereafter. These strong laws succeeded in propping up the demand for these paper currencies, but certainly not at anything like par.

A second refinement was proposed by Walker (1877, pp. 193–5) following Ricardo, and was endorsed by Giffen (1891) and by Palgrave (1894–99, vol. 2, p. 262). Bad money will drive out good money only when the sum of the two is in excess of the wants of trade. This refinement can be made more precise as follows (see Sargent and Smith, 1997, for a formalization of these ideas).

Suppose that the objects used as money have alternative uses: for example, a coin could be melted down and consumed as metal. Monetary objects may be worth more as money than in their next-best use if their supply is restricted in quantity (by government control over the issue), or unrestricted but at a markup over their alternative value (unlimited minting with a seigniorage charge). Furthermore, these monetary objects, although differing in their alternative values, may have the same value as money if they provide exactly the same monetary services, that is, their relative exchange rate is constant but the level indeterminate. Here, the force that keeps the exchange rate fixed is the Kareken and Wallace (1981) result on indeterminacy. The common value of these monetary objects, as money, will depend on the overall demand for monetary services (the ‘needs of trade’), and the total supply of these objects. These conditions may change so as to drive down the value of the objects in monetary use. For example, additions of some monetary objects to the money supply will, if the money demand remains constant, drive down the value of all monetary objects at the same time. If the value falls low enough, the objects with the most valuable alternative use will be the first to be removed from monetary use: the ‘best money’ will be driven out.

Three points should be noted about this qualified version of Gresham's Law. The foregoing reasoning described one possible equilibrium, but does not rule one where Gresham's Law would fail (such as one money circulating and all the others in their alternative use). Second, the best money can be driven out by the addition of any monetary object to the money supply, not necessarily the worst. Finally, this version depends crucially on the expectations that holders of the monetary objects may have about future rates of return on the different objects: if the value of money is expected to fall further, why would agents persist in holding balances in those objects whose value will fall further than others?

The third refinement arose from the growing importance of asymmetric information in economics. In his famous paper, Akerlof (1970, p. 489) noted the analogy between ‘lemons’ driving out good cars from the market and Gresham's Law, although he thought that, in the latter case, ‘both buyer and seller can tell the difference between good and bad money.’ Velde, Weber and Wright (1999) and Dutu, Nosal and Rocheteau (2005) have explored the role of asymmetric information about coinage. Gresham's Law can

make an appearance in such models because of a ‘lemons effect’. Good and bad coins will fetch the same price in trades when the seller cannot recognize or be convinced of the quality of the coin being traded, and this may lead to fewer or no trades involving good coins. When both parties do know the difference, they will trade both good and bad coins, with the latter at a premium. This argument, however, rationalizes Gresham's Law only when good and bad coins are hard to distinguish, for example during medieval debasements in which debased coins were almost identical to the original ones (and thus hard to distinguish) but contained less silver. In most of the cases for which Gresham's Law has been cited, such as milled and hammered money, gold and silver, metal and paper, the monies were clearly distinguishable, and this refinement of Gresham's Law would not apply.

It should now be apparent that trying to salvage Gresham's Law in its lapidary and universal form is hopeless. The phrase would be more useful, not as a general law that excuses the user from providing an explanation, but as a class of outcomes that may or many not obtain, depending on the model's assumptions or the historical episode's circumstances.

Gresham's Law and bimetallism

Since Gresham's Law owes its fame to its role as an argument against bimetallism during the extensive debates of the late 19th century, a word on this topic is in order. Bimetallism – that is, the concurrent circulation of gold and silver currency at a constant exchange rate – would always be defeated by Gresham's Law: whichever metal was cheaper on the market than at the legal rate would always displace the other. Its application in this context is misplaced. As has been shown (Walras, 1977, p. 339; Velde and Weber, 2000), bimetallism is a stable monetary system in which exogenous fluctuations (in supply, monetary or non-monetary demand) are accommodated by fluctuations in the relative shares of the two metals in circulation. An increase in the supply of gold need not result in any change in the relative price of the metals, as long as enough gold can be taken out of, and enough silver added to, non-monetary uses, keeping the ratio of marginal utilities constant. This process of displacement of one money by the other takes place precisely as long as the two monies are substitutes at a constant relative price, meaning that neither one can be said to be ‘worse’ than the other.

See Also

- bimetallism
- commodity money
- money

Bibliography

Akerlof, G.A. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Bridrey, É. 1906. *La Théorie de la monnaie au XIVe siècle. Nicole Oresme; étude d'histoire des*

doctrines et des faits économiques. Paris: V. Giard & E. Brière.

Burgon, J.W. 1839. *The Life and Times of Sir Thomas Gresham, Knt., Founder of the Royal Exchange*, 2 vols. London: Effinghman Wilson.

De Roover, R.A. 1949. *Gresham on Foreign Exchange: An Essay on Early English Mercantilism with the Text of Sir Thomas Gresham's Memorandum: For the Understanding of the Exchange*. Cambridge, MA: Harvard University Press.

Dutu, R., Nosal, E. and Rocheteau, G. 2005. On the recognizability of money. Working paper No. 512, Federal Reserve Bank of Cleveland.

Fetter, F.W. 1932. Some neglected aspects of Gresham's Law. *Quarterly Journal of Economics* 46, 480–95.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press.

Giffen, R. 1891. The Gresham Law. *Economic Journal* 1, 304–6.

Jevons, W.S. 1875. *Money and the Mechanism of Exchange*. London: H.S. King & Co.

Kareken, J.H. and Wallace, N. 1981. On the indeterminacy of equilibrium exchange rates. *Quarterly Journal of Economics* 96, 207–22.

Macaulay, T.B., Baron. 1850–61. *The History of England from the Accession of James II*, 5 vols. London: Longman, Brown, Green and Longmans.

Macleod, H.D. 1855–56. *The Theory and Practice of Banking: With the Elementary Principles of Currency, Prices, Credit, and Exchanges*, 2 vols. London: Longman, Brown, Green, and Longmans.

Macleod, H.D. 1858. *The Elements of Political Economy*. London: Longman, Brown, Green, Longmans, and Roberts.

Macleod, H.D. 1896. *The History of Economics*. New York: Putnam.

Nussbaum, A. 1950. *Money in the Law: National and International*. Brooklyn, NY: Foundation Press, Inc.

Palgrave, R.H.I, ed. 1894–99. *Dictionary of Political Economy*, 3 vols. London and New York: Macmillan and Co.

Rolnick, A.J. and Weber, W.E. 1986. Gresham's Law or Gresham's fallacy? *Journal of Political Economy* 94, 185–99.

Sargent, T.J. and Smith, B.D. 1997. Coinage, debasements, and Gresham's Laws. *Economic Theory* 10, 197–226.

Sargent, T.J. and Velde, F.R. 2002. *The Big Problem of Small Change*. Princeton, NJ: Princeton University Press.

Selgin, G. 2003. Gresham's Law. *EH. Net Encyclopedia*, ed. R. Whaples. 10 June. Online. Available at http://eh.net/encyclopedia/article/selgin.gresham.law, accessed 24 November 2006.

Velde, F.R., Weber, W.E. and Wright, R. 1999. A model of commodity money, with applications to Gresham's Law and the debasement puzzle. *Review of Economic Dynamics* 2, 291–323.

Velde, F.R. and Weber, W.E. 2000. A model of bimetallism. *Journal of Political Economy* 108, 1210–34.

Walker, F.A. 1877. *Money*. New York: Henry Holt and Company.

Walras, L. 1977. *Elements of Pure Economics*. Fairfield, NJ: Augustus M. Kelley.

Wolowski, L. 1864. *Traictie de la première invention des monnoies de Nicole Oresme, textes français et latin d'après les manuscrits de la Bibliothèque impériale et Traité de la monnoie de Copernic, texte latin et traduction française*. Paris: Guillaumin.

How to cite this article

Velde, François R. "Gresham's Law." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000114> doi:10.1057/9780230226203.0677

The New Palgrave Dictionary of Economics Online

Griliches, Zvi (1930 – 1999)

Ernst R. Berndt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Born in Lithuania and a survivor of the Dachau concentration camp, Zvi Griliches was one of the most important and influential empirical economists of the second half of the 20th century. Griliches' lifelong research focus involved detailed analyses on the role of technological change as a principal driver of productivity and long-run economic growth. His research contributions were wide-ranging and seminal, and through his students, collaborators and colleagues he greatly affected the conduct of empirical research in economics.

Keywords

American Economic Association; capital measurement; diffusion of technology; discrete count data; distributed lags; education and economic growth; Griliches, Z.; hedonic prices; inflation; input–output analysis; measurement error models; National Bureau of Economic Research; patents; pharmaceutical industry; price measurement; production functions; productivity growth; quality–price relationship; research and development; technical change; unobservable variables

Article

Zvi Griliches was one of the most important and influential empirical economists of the second half of the 20th century. His research contributions were wide-ranging and seminal, and through his students, collaborators and colleagues he greatly affected the conduct of empirical research in economics.
Zvi Griliches was born in Lithuania to a well-educated Jewish family. At age 11, along with his parents and sister, Griliches was moved into the ghetto in German-occupied Kaunas, where he remained until 1944, when he and his father were separated from his mother and sister and were sent to Dachau concentration camp. His father died there from starvation, and his mother died from typhus in the Stutthof concentration camp. In May 1945 Griliches was liberated by General Patton's 3rd US Army. After spending a year in Germany, he attempted to go to Palestine, but was prevented by the British

from entering, and was held for nine months in an internment camp on Cyprus. He arrived in Palestine in September 1947.

Griliches then spent three years in several kibbutzim, focusing much of his spare time on learning English and mathematics. Although he never went to high school, he taught himself sufficiently well to pass an external high-school equivalence examination in 1950. After studying history for a year at the Hebrew University, Griliches applied for and was awarded a scholarship by the University of California at Berkeley, where he chose agricultural economics as his major field of study. Griliches completed Berkeley's undergraduate degree requirements in two years, took his first econometrics course from George Kuznets (Simon's brother), and earned his Master's degree in 1954, at which time he transferred to the University of Chicago, where he completed his Ph.D. in 1957, writing a seminal dissertation on the economics of the diffusion of hybrid corn. After 15 years at Chicago, in 1969 Griliches moved to Harvard. In the late 1970s, he became the first Director of the National Bureau of Economic Research's Program on Technological Progress and Productivity Measurement. Elected to the National Academy of Sciences in 1975 and President of the American Economic Association in 1993, Griliches also served as co-editor of *Econometrica* for a decade, and won the John Bates Clark Medal at age 35. He died from cancer in 1999. Further details on his life are found in Lerner (2004) and Trajtenberg and Berndt (2001).

Griliches' lifelong research focus involved detailed analyses on the role of technological change as a principal driver of long-run economic growth, and included examination of the determinants of the diffusion of new technologies, the measurement of physical, human and R&D capital, the role of education, and the contribution of R&D to productivity growth. Griliches devoted a great deal of attention throughout his career to properly measuring various inputs and outputs, and adjusting prices for quality change. Although economic growth theory in the late 1950s emphasized the role of disembodied technological progress and ‘manna from heaven’, Griliches (1957; 1958; 1960a) instead developed the view that technological progress is itself an economic phenomenon amenable to economic analysis.

In formulating this theme and building supportive empirical evidence, Griliches and his collaborators enlarged considerably the set of econometric tools and procedures now commonly employed by empirical economic researchers. These econometric tool innovations included the use of distributed lags (1961b; 1967a; 1984a), procedures for dealing with measurement error and unobservable variables (1974; 1975; 1977; 1978), and with discrete count data (1984b).

Much of Griliches' early empirical research focused on measuring inputs, outputs and productivity in agriculture (1960a; 1964), but later on it extended to other sectors of the economy, particularly the service sectors (1992; 1994a; 2000a). The theoretical framework integrating these measurement issues involved use of the production function (1964; 1967b; 1969; 1970; 1971). He initially focused on measurement of physical capital inputs (1963; 1966; 1984a), but subsequently on issues involving measurement of labour input that generated several influential strands of research. One important literature involved establishing relationships between quality-adjusted labour inputs, education, and rates of return to schooling (1975; 1979; 1986a; 1997; 2000b). A related literature examined complementarity between physical capital and skilled labour (1969; 1977; 1994b).

In addition to examining the relationships among outputs and labour and capital inputs, a great deal of Griliches' research focused on the special role of R&D, differences between private and social returns to R&D, and the impact of R&D on productivity growth (1958; 1964; 1980; 1994; 1998). This research then led to a more detailed analysis of R&D, including a host of important studies that examined the extent to which patents served as a useful indicator of inventive activity generated by R&D (1986b;

1986c; 1987; 1990).

Another of Griliches' seminal contributions involved reviving intellectual and policy interest in the use of hedonic multivariate regression techniques to adjust observed prices for changes in observed and unobserved quality over time. In large part this research reflected Griliches' scepticism about measuring productivity growth as the ‘residual’, which in its simplest form simply subtracted the growth of traditionally measured inputs from the growth of outputs. To what extent, he reasoned, could what goes into that residual (‘a measure of our ignorance’) reflect instead measurement and specification errors, rather than technological and other quality change? This led Griliches to undertake empirical analyses initially linking prices of new automobiles to observed characteristics (1961a), as well as prices of used automobiles to operating cost characteristics, such as fuel efficiency and gasoline prices (1986d). Subsequent research examined the extent to which traditionally constructed government price statistics for certain high-technology goods such as personal computers overstated price inflation (or understated price deflation) by failing to incorporate fully the quality attribute improvements that were embodied in new goods (1993a; 1995). Yet another strand of this hedonic price research extended into prescription pharmaceuticals (1996), where problems of a new goods bias and oversampling of older goods were particularly important (1993b; 1994c). Adjusting measures of medical price inflation for changes in outcomes from medical treatments was a focus of Griliches' work with collaborators just prior to his death in 1999 (2000a).

Griliches' interest in price measurement led to his being appointed a member of both the 1960–61 Stigler Commission (resulting in 1961a), and the Boskin Commission of 1995–96, each of which provided influential and thoughtful recommendations on price measurement issues facing the U.S. Bureau of Labor Statistics. Although he was also named a member of the National Academy of Science Panel on the Conceptual, Measurement and Other Statistical Issues in Developing Cost-of-Living Indexes (1999–2001), because of his ill health and subsequent death in late 1999 he was unable to contribute directly to the final report.

See Also

- agricultural economics
- distributed lags
- econometrics
- ghettoes
- growth accounting
- hedonic prices
- level accounting
- measurement error models
- production functions
- technology

Selected works

1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25, 501–22.

1958. Research cost and social returns: hybrid corn and related innovations. *Journal of Political Economy* 66, 419–31.

1960a. Measuring inputs in agriculture: a critical survey. *Journal of Farm Economics* 42, 1411–33.

1960b. Hybrid corn and the economics of innovation. *Science* 132, 275–80.

1961a. Hedonic price indexes for automobiles: an econometric analysis of quality change. In *The Price Statistics of the Federal Government*, General Series No. 73. New York: NBER.

1961b. A note on serial correlation bias in estimates of distributed lags. *Econometrica* 29, 65–73.

1963. Capital stock in investment functions: some problems of concept and measurement. In *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, ed. C. Christ et al. Stanford, CA: Stanford University Press.

1964. Research expenditures, education, and the aggregate agricultural production function. *American Economic Review* 54, 961–74.

1966. (With D. Jorgenson.) Sources of measured productivity change: capital input. *American Economic Review* 56(2), 50–61.

1967a. Distributed lags: a survey. *Econometrica* 35, 16–49.

1967b. (With D. Jorgenson.) The explanation of productivity change. *Review of Economic Studies* 34, 249–83.

1969. Capital skill complementarity. *Review of Economics and Statistics* 51, 465–8.

1970. Notes on the role of education in production functions and growth accounting. In *Education, Income, and Human Capital*, NBER Studies in Income and Wealth, vol. 35, ed. W. Hansen. New York: Columbia University Press.

1971. (With V. Ringstad.) Economies of Scale and the Form of the Production Function. Amsterdam: North-Holland.

1974. Errors in variables and other unobservables. *Econometrica* 42, 971–8.

1975. (With G. Chamberlain.) Unobservables with a variance components structure: ability, schooling,

and the economic success of brothers. *International Economic Review* 16, 422–9.

1977. Estimating the returns to schooling: some econometric problems. *Econometrica* 45, 1–22.

1978. (With B. Hall and J. Hausman.) Missing data and self-selection in large panels. *Annales de l'INSEE* 30/31, 137–76.

1979. Sibling models and data in economics: beginnings of a survey. *Journal of Political Economy* 87(5, Part II), S37–S64.

1980. R&D and the productivity slowdown. *American Economic Review Papers and Proceedings* 70, 343–8.

1984a. (With A. Pakes.) Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs. *Review of Economic Studies* 51, 243–62.

1984b. (With J. Hausman and B. Hall.) Econometric models for count data with an application to the patent R&D relationship. *Econometrica* 52, 909–38.

1986a. (With J. Bound and B. Hall.) Wages, schooling and IQ of brothers and sisters: do the family factors differ? *International Economic Review* 27, 77–105.

1986b. Economic data issues. In *Handbook of Econometrics*, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

1986c. (With B. Hall and J. Hausman.) Patents and R&D: is there a lag? *International Economic Review* 27, 265–83.

1986d. (With M. Ohta.) Automobile prices and quality change: did the gasoline price increases change consumer tastes in the U.S.? *Journal of Business and Economic Statistics* 4, 187–98.

1987. (With A. Pakes and B. Hall.) The value of patents as indicators of inventive activity. In *Economic Policy and Technological Performance*, ed. P. Dasgupta and P. Stoneham. Cambridge: Cambridge University Press.

1990. Patent statistics as economic indicators: a survey. *Journal of Economic Literature* 28, 1661–707.

1992. (Ed.) *Output Measurement in the Service Sectors*. NBER Studies in Income and Wealth, vol. 56. Chicago: University of Chicago Press.

1993a. (With E. Berndt.) Price indexes for microcomputers: an exploratory study. In *Price Measurements and their Uses*, ed. M. Foss M. Manser and A. Young. NBER Studies in Income and Wealth, vol. 57. Chicago: University of Chicago Press.

1993b. (With E. Berndt and J. Rosett.) Auditing the producer price index: micro evidence from prescription pharmaceutical preparations. *Journal of Business and Economic Statistics* 27, 131–52.

1994a. Productivity, R&D and the data constraint. *American Economic Review* 84, 1–23.

1994b. (With E. Berman and J. Bound.) Changes in the demand for skilled labor within U.S. manufacturing industries: evidence from the Annual Survey of Manufacturing. *Quarterly Journal of Economics* 109, 367–98.

1994c. (With I. Cockburn.) Generics and new goods in pharmaceutical price indexes. *American Economic Review* 84, 1213–32.

1995. (With E. Berndt and N. Rappaport.) Econometric Estimates of price indexes for personal computers in the 1990s. *Journal of Econometrics* 68, 243–68.

1996. (With E. Berndt and I. Cockburn.) Pharmaceutical innovations and market dynamics: tracking effects on price indexes for antidepressant drugs. *Brookings Papers on Economic Activity: Microeconomics* 1996(2), 133–88.

1997. Education, human capital and growth: a personal perspective. *Journal of Labor Economics* 15(1, Part II), S330–S344.

1998. *R&D and Productivity: The Econometric Evidence*. Chicago: University of Chicago Press.

2000a. (With E. Berndt, D. Cutler, R. Frank, J. Newhouse and J. Triplett.) Medical care prices and output. In *Handbook of Health Economics*, vol. 1A, ed. J. Newhouse and A. Culyer. Amsterdam: Elsevier Science.

2000b. *R&D, Education and Productivity: A Retrospective*. Cambridge, MA: Harvard University Press.

Bibliography

Trajtenberg, M. and Berndt, E. 2001. In Memoriam: Zvi Griliches (1930–1999). *Journal of Economic and Social Measurement* 27, 93–7.

Lerner, B. 2004. From the particular to the aggregate: the story of Zvi Griliches. In *The Triumph of Wounded Souls: Seven Holocaust Survivors' Lives*. Notre Dame, IN: University of Notre Dame Press.

How to cite this article

Berndt, Ernst R. "Griliches, Zvi (1930–1999)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000117> doi:10.1057/9780230226203.0678

The New Palgrave Dictionary of Economics Online

gross substitutes

Lionel W. McKenzie
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The gross substitute assumption is used to establish the existence and uniqueness of an equilibrium and to prove the equilibrium to be stable for a dynamic adjustment system for prices. The gross substitute assumption also implies results of comparative statics, that is, results on the displacement of equilibrium that follows from shifts in demand or changes in initial stocks. The concept of gross substitutes was introduced by Mosak (1944) in the context of a pure trading model. A definition with wider application is the one used by Morishima (1964).

Keywords

comparative statics; excess demand; existence of equilibrium; global stability theorems; gross substitutes; income effect; revealed preference; stability; substitution effect; tâtonnement; temporary equilibrium; uniqueness of equilibrium; Walras's Law; weak gross substitutes

Article

The assumption that goods are gross substitutes is applied to a set of excess demand functions $e_i(p_1,\ldots, p_n)$, $i = 1, \ldots, n$, where p_i is the price of the ith good and e_i is the excess demand for the ith good. The concept was introduced by Mosak (1944) in the context of a pure trading model. However, his definition required that $\partial e_i|\partial p_j$ have the same sign as the substitution term in the Slutsky equation as well as be of positive sign. That is, the income effect should not overbalance the substitution effect. At about the same time Metzler (1945) said simply that the jth good is a gross substitute for the ith good if $e_{ij}=\partial e_i(p)|\partial p_j>0$ holds, and this has been the meaning used in later papers when the functions e_i have been assumed to be differentiable.
A definition with wider application is the one used by Morishima (1964). By this definition the jth good

is a gross substitute for the *i*th good if $e_i(p)<e_i(p')$ whenever $p \leq p', p_j < p'_j, p'_k = p_k$ for $k \neq j$. We will say that the assumption of gross substitutes holds if $e_i(p) < e_i(p'), i = 1, \ldots, n$, for all *p* and *p* ′ such that $p \leq p', p \neq p'$, and $p_i = p'_i$. If *e*(*p*) is differentiable, this assumption implies that the $n \times n$ matrix [$e_{ij}(p)$] has positive off-diagonal elements for all *p* in the interior of the domain of *e*(*p*). A matrix with this property and a negative diagonal is often referred to as a Metzler matrix. We will say that the assumption of weak gross substitutes holds if the condition $e_i(p) < e_i(p')$ is replaced by the weak inequality $e_i(p) \leq e_i(p')$. In the case of differentiable *e*(*p*) it is implied by the assumption of weak gross substitutes that [$e_{ij}(p)$] has non-negative off-diagonal elements. That these assumptions are not empty is shown by the case of excess demands defined by Cobb–Douglas utility functions of the form $U(x) = \Pi_{i=1}^{n} x_i^{\alpha_i}$, where $\alpha_i>0$ and $\Sigma \alpha_i = 1$. The excess demand function for a consumer in a pure exchange economy, holding initial stocks $\bar{x}$, is $e_i(p) = (\alpha_i \Sigma_{k=1}^{n} p_k \bar{x}_k / p_i) - \bar{x}_i$, so $e_{ij}(p) = \alpha_i \bar{x}_j / p_i \geq 0$ for $i \neq j$. If the initial stock of every good is positive for the whole market, the assumption of gross substitutes is satisfied.

A price vector *p* is said to be an equilibrium of the set of demand functions if $e(p) \leq 0$. The gross substitute assumption is used to establish the existence and uniqueness of an equilibrium and to prove the equilibrium to be stable for a dynamic adjustment system for prices. The gross substitute assumption also implies results of comparative statics, that is, results on the displacement of equilibrium that follows from shifts in demand or changes in initial stocks.

The following assumptions will be made on *e*(*p*).

- (B) *e*(*p*) is defined for all *p*>0, and $p^s \to p$ where $p_i = 0$ for $i \in I$, $p \neq 0$ implies $\Sigma_{i \in I} |e_i(p)| \to \infty$. Also *e*(*p*) is bounded below.
- (C) *e*(*p*) is single-valued and continuous for $p > 0$.
- (H) $e(p) = e(\lambda p)$ for any $\lambda > 0$, that is *e*(*p*) is positively homogeneous of degree 0.
- (W) $\Sigma_{i=1}^{n} p_i e_i(p) = 0$, that is, *e*(*p*) satisfies Walras's Law.

The example of Arrow–Hahn (1971, pp. 29–30), where demand is derived from the utility function $u(x) = x_1^{1/2} + x_2^{1/2} + x_3^{1/2}$ where *x* is the consumption bundle, shows that assumption B cannot easily be improved.

Uniqueness

The existence of a positive equilibrium under assumptions B, C, H, and W does not require an assumption of gross substitutes (see Debreu, 1970). However, if there should exist an equilibrium price *p* when gross substitutes is assumed, *p*>0 must hold. This follows from the fact that $p_i = 0$ for some *i*, and $p \neq 0$, is inconsistent with assumption H in that case, since it must at the same time be true, for λ >1, that $e_i(\lambda p) = e_i(p)$ and $e_i(\lambda p) > e_i(p)$. Thus *e*(*p*) cannot be defined at such a *p*.

Make the assumption that all goods are gross substitutes. Assume that two equilibria exist, say p and p' where $p' \neq \lambda p$ for any $\lambda > 0$. By assumption H we may chose p and p' so that $p_i = p_i'$ for some i and $p_j \leq p_j'$ for $j \neq i$. Then by gross substitutes $e_i(p) < e_i(p')$. But p and $p' > 0$ and assumption W implies that $e_i(p) = e_i(p') = 0$, which is a contradiction. Thus an equilibrium price vector is unique up to multiplication by a positive number.

Consider a partition of the goods into two non-empty subsets I and J. Say that the excess demand functions are connected if for any such partition $p_i = p_i'$ for all $i \in I$ and $p_j < p'_j$ for all $j \in J$ implies that $e(p) \neq e(p')$. Then by a similar argument, uniqueness of equilibrium may be seen to hold when weak gross substitutes and connectedness are assumed. Strictly speaking, connectedness need only hold at an equilibrium point.

If weak gross substitutes is assumed without connectedness, uniqueness may fail. However, it may be shown that the set of equilibrium price vectors is convex (McKenzie, 1960). Arrow and Hurwicz (1960) proved that the weak axiom of revealed preference holds between any equilibrium price vector p and any non-equilibrium price vector p' when weak gross substitutes is assumed. This means that $0 = p'e(p) = p'e(p')$ implies $pe(p') > pe(p) = 0$. In other words, $pe(p') > 0$. Suppose p and p' are both equilibria and consider $p'' = \alpha p + (1-\alpha)p'$ for $0 < \alpha < 1$. If p'' is not an equilibrium $p''e(p'') = \alpha pe(p'') + (1-\alpha)p'e(p'') > 0$ which conradicts assumption W, Walras's Law. Thus p'' is an equilibrium and the set of equilibria is convex.

Comparative statics

The modern approach to the comparison of equilibria after a shift of demand was begun by Hicks (1939). The fact that the Hicksian theorems hold locally when the excess demand functions satisfy the gross substitute assumption was proved by Mosak (1944). A global treatment of comparative statics in this context was given by Morishima (1964). Assume weak gross substitutes. Let demand shift from the ith good to the jth good. Let p and p' be the old and new equilibrium prices, e and e' the old and new excess demand functions. Then

$$p_i'e_i'(p) + p_j'e_j'(p) = \sum_{i=1}^{n} p_k'e_k'(p) > 0, \qquad \text{by the Weak Axiom}. \tag{1}$$

On the other hand,

$$p_i e_i'(p) + p_j e_j'(p) = \sum_{i=1}^{n} p_k e_k'(p) = 0, \qquad \text{by Walras' Law.}$$
(2)

Multiply (1) by p_i and (2) by p_i' and subtract to obtain $p_i p_j' - p_i' p_j > 0$, or $p_j' / p_j > p_i' / p_i$. Thus the price of the *j*th good increases relative to the price of the *i*th good.

Assume that all goods are gross substitutes. Then a shift in demand from the *i*th good to the *j*th good raises the price of the *j*th good relative to all other goods and lowers the price of the *i*th good relative to all other goods. These results are immediate from the fact that the good that rises in price relative to some goods and falls relative to none must experience a fall in demand and the good that falls in price relative to some goods and rises relative to none must experience a rise in demand. But only the *j*th good can absorb a rise and only the *i*th good can absorb a fall and still have demand equal to 0 after the shift has occurred. All other goods have zero excess demands at the old equilibrium prices after the shift, and the excess demands at the new equilibrium prices must also be zero. The same results follow from weak gross substitutes if any subset of the excess demand functions with $n - 1$ members is connected.

If the $e_i(p)$ are assumed to be continuously differentiable, the local theory of comparative statics is equivalent to determining the sign pattern of the inverses of the principal submatrices of order $n - 1$ of the Jacobian $[e_{ij}]\, i, j = 1, \ldots, n$. The gross substitutes assumption implies that $e_{ij} > 0$ for $i \neq j$. Then either assumption H or W implies that the inverses of these submatrices have all elements negative. Choose the *n*th good as numeraire, and choose units so equilibrium prices $p_i = 1$, all *i*. Then $dp_i / d\alpha = -([e_{ij}]_{nn})^{-1}_{ih}$ if $\partial e_h(p, \alpha) / \partial \alpha = 1$. This is minus the *i*th element of the *h*th column of the inverse matrix of the submatrix where the *n*th row and column are omitted. Thus $dp_i / d\alpha > 0$, and it may be shown that $dp_h / d\alpha > dp_i / d\alpha$ for $i \neq h$ or *n*. If weak gross substitutes is assumed but the Jacobian and its $n - 1$ principal minors are indecomposable the same conclusions follow.

The local results may be extended to the case where there is a numeraire and the goods other than the numeraire may be partitioned into two non-empty subsets with indices in *I* and *J*, such that $e_{ij} > 0$ for $i \neq j$ and *i* and *j* in the same subset, while $e_{ij} < 0$ for *i* and *j* in different subsets. If the principal minors of $[e_{ij}]$ of order $n - 1$ have dominant diagonals, at equilibrium with the equilibrium prices as multipliers, the shifts of demand raise the price of the good to which demand has shifted and also the prices of all other goods in the subset of the partition to which it belongs, while lowering the prices of the goods in the other subset. Also the beneficiary of the demand shift has the largest change in equilibrium price in absolute value. These results are seen to follow from those for gross substitutes by considering the matrix formed by pre- and post-multiplying a principal minor of $[e_{ij}]$ by a diagonal matrix *D* with $d_{ii} = 1$ for $i \in I$ and $d_{jj} = -1$ for $j \in J$. This case was first analysed by Morishima (1952). The gross substitute case and the Morishima case may be shown to be the only sign patterns for a Jacobian matrix of the demand functions with all elements non-zero which allow the inverse matrix to be signed without quantitative information (see Bassett, Habibagahi and Quirk, 1967).

Stability

Since weak gross substitutes implies the weak axiom of revealed preference which in turn implies local stability of the tâtonnement for both the usual price adjustment models, with and without a numeraire (Arrow and Hurwicz, 1958), there is no special advantage for gross substitutes in local analysis of stability. However, for global results on stability the gross substitute assumptions are the only ones known with much plausibility. In order to use the weak axiom the adjustment must be, after proper choice of units,

$$\text{(I)} \qquad \dot{p}_i = e_i(p),$$

for all goods other than the numeraire, if any, where $\dot{p}_i = \partial p_i / \partial t$, and (I) must hold globally. Excess demand is now assumed to be continuously differentiable. In order to use the other major possibility, a dominant diagonal for the matrix of demand elasticities, assuming a numeraire, the adjustment process

$$\text{(II)} \qquad \dot{p}_i = p_i e_i(p),$$

is used for the non-numeraire goods (see Arrow and Hahn, 1971, p. 293). While (II) may be more reasonable than (I) for a global adjustment rule it is also very special.

On the other hand, when weak gross substitutes is assumed, stability may be proved for the adjustment rule

$$\text{(III)} \qquad \dot{p}_i = h_i(p),$$

for all goods other than the numeraire, if any. The only special requirements placed on $h_i(p)$ are that it should be continuously differentiable and have the sign of $e_i(p)$. The adjustment rule III was proposed by McKenzie (1960) and global stability was proved using as a Lyapunov function the value of positive excess demand, $V[p(t)] = \Sigma_{i \in P} p_i e_i(p)$, $P = \{i | e_i(p) \geq 0\}$. The tâtonnement is shown to converge to the convex set of equilibrium prices. If the excess demand functions are connected, the equilibrium is unique.

Arrow and Hurwicz (1962) proved that the convergence of process III with weak gross substitutes is, in fact, to a particular equilibrium price vector, which will depend on the initial prices. This is clear once it is recognized that the goods whose prices are highest relative to some equilibrium price cannot be in

excess demand and their prices cannot rise, and *mutatis mutandis* for the prices which are lowest. Thus the prices during tâtonnement cannot retreat from any equilibrium price vector. In the case of gross substitutes this line of argument is very simple and effective, since the prices actually fall and rise respectively.

In the theory of tâtonnement prices are revised according to excess demand but no trading occurs until equilibrium is reached. However, the stability of the adjustment process for a pure exchange economy is not lost if trading occurs at market prices, so long as the excess demand that drives the tâtonnement is determined by the maximization of utility by each trader under a budget equal to the value of the stocks he currently holds (Negishi, 1961). The crucial fact is that trading at market prices has only second-order effects on excess demand. However, the price to which the process converges now depends on initial conditions and the course of trading.

It was pointed out by Rader (1972) that the production sector of the economy is unlikely to satisfy the gross substitute assumption in the demand for factors. As a consequence it seemed that the range of application of the gross substitutes assumption was effectively confined to pure trading economies. However, Rader was able to prove a local stability theorem assuming gross substitutes only for households. The production sector is made up of a finite number of firms with strictly convex production possibility sets. The key to the argument is that $p[e_{ij}^{H}] = 0$ at equilibrium, where $e^{H}(p)$ is household excess demand. This is established by differentiating Walras's Law $pe^{F}(p) + pe^{H}(p) = 0$ to give

$$e^{F}(p) + e^{H}(p) + p[e_{ij}^{F}] + p[e_{ij}^{H}] = 0. \tag{3}$$

Since the first two terms sum to 0 at equilibrium, and the third term is 0 by profit maximization, (3) implies $p[e_{ij}^{H}] = 0$. Then the Jacobian of the adjustment system I with a numeraire is negative definite at equilibrium and local stability follows.

Generalizations

Mukherji (1972) pointed out that some gross substitute theorems carry over if the weak gross substitute pattern is established in a transformed goods space. In particular, if there exists a matrix S such that $S^{-1}[e_{ij} + e_{ji}]S$ is indecomposable with off-diagonal elements non-negative, the tâtonnement is locally stable for the process I, with or without a numeraire. Also a rise in demand for the ith good causes the ith equilibrium price to rise. Ohyama (1972) shows that similar results follow if there is a stochastic matrix G which is positive definite and $G[e_{ij}]$ or $G[e_{ij} + e_{ji}]$ satisfies the conditions above. Uzawa (1960) proved that a discrete tâtonnement defined by $p_i(t+1) = \max\{0, p_i(t) + f_i(t)\}$ where $f_i(t) = \beta_i\{e_i[p(t)]\}$ is globally stable when the Weak Axiom holds and $\beta_i > 0$ is sufficiently small. He

assumes a numeraire and some additional differentiality and nonsingularity conditions.
Howitt (1980) defines a generalized gross substitute notion where $e(p)$ is allowed to be a convex valued correspondence rather than a function. This assumption of generalized gross substitutes holds if

1. (a) for all p, $p' > 0$, if there is a partition of indices for goods into non-empty subsets I and J where $p_i' = p_i$ for all $i \in I$ and $p_j' > p_j$ for all $j \in J$ than $\Sigma_{i \in I} p_i x_i' \geq \Sigma_{i \in I} p_i x_i$ for all $x \in e(p)$, $x' \in e(p')$;
2. (b) strict inequality holds in (a) if p is an equilibrium.

Howitt proves that the equilibrium price vector is unique and globally stable for the price adjustment process

$$\text{(IV)} \qquad \dot{p}_i \in e(p),$$

under generalized gross substitutes, when assumptions C, W, and H hold and an equilibrium exists. Excess demand $e(p)$ is assumed to be upper semi-continuous. He applies his result to the linear economy described by Gale (1976) and shown to satisfy gross substitutes by Cheng (1979).
Arrow and Hurwicz (1962) extended the adjustment process III to include expected prices. Let q represent expected prices. Then their process with adaptive expectations is

$$\dot{p}_i = h_i(p, q), \quad \text{if} \quad p_i > 0 \quad \text{and} \quad h_i(p, q) > 0, \; = 0, \quad \text{otherwise}$$

$$\text{(V)} \qquad \text{sign}\, h_i(p, q) = \text{sign}\, e_i(p, q), \quad \text{or} = 0 \quad \text{if } i \text{ is a numeraire}$$
$$\dot{q}_i = a_i(q_i - p_i)$$

They prove global stability for this process under weak gross substitutes with some auxiliary assumptions. Arrow and Hahn (1971) give a proof of global stability with an assumption of gross substitutes and assumptions B, C, H, and W. By the gross substitute assumption in this context is meant that $\partial e_i(p, q) / \partial p_j > 0$ for $i \neq j$, and $\partial e_i(p, q) / \partial q_j > 0$ for all j. The adjustment function h_i is assumed to be continuously differentiable.

Arrow and Hahn also prove a global stability theorem for a model in which expected prices q_i are given as functions $q_i(p)$ of current prices. This is in accord with models of temporary equilibrium. They prove global stability for adjustment process II with a numeraire, assuming gross substitutes for $e(p, q)$ and the Hicksian elasticity of substitution $\varepsilon_j = d \log q_j / d \log p_j \leq 1$, which is consistent with Hicks's presumption when the strict inequality holds (see Hicks, 1939, p. 251).

In this model of the tâtonnement for temporary equilibrium $q(p)$ is presumably the expected price on the assumption that p is an equilibrium price. It is not so clear how to justify adaptive expectations in the tâtonnement setting.

See Also

- substitutes and complements

Bibliography

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Arrow, K.J. and Hurwicz, L. 1958. On the stability of the competitive equilibrium, I. *Econometrica* 26, 522–52.

Arrow, K.J. and Hurwicz, L. 1960. Competitive stability under weak gross substitutability: the ‘Euclidean distance approach’. *International Economic Review* 1, 38–49.

Arrow, K.J. and Hurwicz, L. 1962. Competitive stability under weak gross substitutability: nonlinear price adjustment and adaptive expectations. *International Economic Review* 3, 233–55.

Bassett, L., Habibagahi, H. and Quirk, J. 1967. Qualitative economics and Morishima matrices. *Econometrica* 35, 221–33.

Cheng, H.-C. 1979. Linear economies are ‘gross substitute’ systems. *Journal of Economic Theory* 20, 110–17.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Gale, D. 1976. The linear exchange model. *Journal of Mathematical Economics* 3, 205–59.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Howitt, P. 1980. Gross substitutability with multi-valued excess demand functions. *Econometrica* 48, 1567–75.

McKenzie, L.W. 1960. Stability of equilibrium and the value of positive excess demand. *Econometrica* 28, 606–17.

Metzler, L. 1945. Stability of multiple markets: the Hicks conditions. *Econometrica* 13, 277–92.

Morishima, M. 1952. On the laws of change of price-system in an economy which contains complementary commodities. *Osaka Economic Papers* 1, 101–13.

Morishima, M. 1964. *Equilibrium, Stability and Growth*. Oxford: Clarendon Press.

Mosak, J.L. 1944. *General Equilibrium Theory in International Trade*. Bloomington, Indiana: Principia Press.

Mukherji, A. 1972. On complementarity and stability. *Journal of Economic Theory* 4, 442–57.

Negishi, T. 1961. On the formation of prices. *International Economic Review* 2, 122–6.

Ohyama, M. 1972. On the stability of generalized Metzlerian systems. *Review of Economic Studies* 39, 193–204.

Rader, T. 1972. General equilibrium theory with complementary factors. *Journal of Economic Theory* 4, 372–80.

Uzawa, H. 1960. Walras' tâtonnement in the theory of exchange. *Review of Economic Studies* 27, 182–94.

How to cite this article

McKenzie, Lionel W. "gross substitutes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000119> doi:10.1057/9780230226203.0679

The New Palgrave Dictionary of Economics Online

Grossman, Herschel I. (1939 – 2004)

Enrico Spolaore
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

excess demand and supply; Grossman, H.; Haavelmo, T.; Keynesianism; labour demand; labour markets; non-clearing markets; Patinkin, D.

Article

Herschel I. Grossman was born in Philadelphia on 6 March 1939. He obtained a BA from the University of Virginia in 1960, a B.Phil. from Oxford in 1962, and a Ph.D. from Johns Hopkins in 1965. He joined Brown University's economics faculty in 1964. He died suddenly while attending a conference in Marseilles on 9 October 2004.

In the early 1970s Grossman, in cooperation with Robert J. Barro, produced path-breaking research on the foundations of Keynesian macroeconomics. Barro and Grossman's main joint contribution was the article ‘A General Disequilibrium Model of Income and Employment’ (1971a). This was the first formalization linking the labour market and the output market in a general-disequilibrium setting with exogenous wage rate and price level, in which labour demand is constrained by output sales while the demand for goods is constrained by sales in the labour market. Generalized excess supply implied a Keynesian demand-multiplier effect. The analysis also shed light on cases of generalized excess demand, such as in socialist economies with artificially low prices. This work unified complementary strands of the literature – particularly, Patinkin (1965, ch. 13) and Clower (1965) – within a coherent choice-theoretic framework, and for many years was the most cited article ever published in the *American Economic Review*. Barro and Grossman summarized and extended their analysis in the landmark book *Money, Employment, and Inflation* (1976). Independent appraisals of the non-market-clearing paradigm and its limits were given by Barro (1979) and Grossman (1979). See also Grossman's entry on Monetary Disequilibrium and Market Clearing in the first (1987) edition of *The New Palgrave: A Dictionary of Economics*.

In his subsequent work Grossman became increasingly interested in understanding the foundations of economic policy and the effects of conflict on the economy, and made innovative contributions to political economy and the economics of appropriation (see Kolmar, 2005, for a survey). Building on

Haavelmo (1954), Grossman modelled conflict as an economic activity by agents allocating resources over various uses, including the uses for appropriative conflict itself. His contributions in this area comprise theories of governments as kleptocracies (1990; 1994a; 1999), insurrections (1991), appropriation and land reform (1994b), effective property rights (1995), anarchy, predation and the state (2002), and many others.

See Also

- defence economics
- fixprice models
- Haavelmo, Trygve
- Keynesianism
- Patinkin, Don

Selected works

1971a. (With R. Barro.) A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

1971b. Money, interest, and prices in market disequilibrium. *Journal of Political Economy* 79, 943–61.

1974. (With R. Barro.) Suppressed inflation and supply multiplier. *Review of Economic Studies* 41, 87–104.

1976. (With R. Barro.) *Money, Employment, and Inflation*. New York: Cambridge University Press.

1978. Risk shifting, layoffs, and seniority. *Journal of Monetary Economics* 4, 661–86.

1979. Why does aggregate employment fluctuate? *American Economic Review* 69, 64–9.

1987. Monetary disequilibrium and market clearing. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. Basingstoke: Palgrave.

1988. (With B. Diba.) Explosive rational bubbles in stock prices? *American Economic Review* 78, 520–30.

1988. (With J. Van Huyck.) Sovereign debt as a contingent claim. excusable default, repudiation, and reputation. *American Economic Review* 78, 1088–97.

1990. (With S. Noh.) A theory of kleptocracy with probabilistic survival and reputation. *Economics and*

Politics 2, 157–71.

1991. A general equilibrium model of insurrections. *American Economic Review* 81, 912–21.

1994a. (With S. Noh.) Proprietary public finance and economic welfare. *Journal of Public Economics* 53, 187–204.

1994b. Production, appropriation, and land reform. *American Economic Review* 84, 705–12.

1995. (With M. Kim.) Swords or plowshares? A theory of the security of claims to property. *Journal of Political Economy* 103, 1275–88.

1999. Kleptocracies and revolutions. *Oxford Economic Papers* 51, 267–83.

2002. ‘Make us a king’: anarchy, predation, and the state. *European Journal of Political Economy* 18, 31–46.

Bibliography

Barro, R. 1979. Second thoughts on Keynesian economics. *American Economic Review* 69, 54–9.

Clower, R. 1965. The Keynesian counterrevolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Haavelmo, T. 1954. *A Study in the Theory of Economic Evolution*. Amsterdam: North-Holland.

Kolmar, M. 2005. The contribution of Herschel I. Grossman to political economy. *European Journal of Political Economy* 21, 802–14.

Patinkin, D. 1965. *Money, Interest and Prices*, 2nd edn. New York: Harper and Row.

How to cite this article

Spolaore, Enrico. "Grossman, Herschel I. (1939–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000204> doi:10.1057/9780230226203.0680

The New Palgrave Dictionary of Economics Online

Grotius (de Groot), Hugo (1583 – 1645)

P.G. Stein

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

combinations; contract theory; Grotius (De Groot), H.; natural law; private property; usury

Article

Legal theorist, philosopher and theologian, Grotius was born in Delft on 10 April 1583 and died in Rostock on 28 August 1645. An infant prodigy, Grotius entered the University of Leyden at the age of 11, and at 15 was hailed by Henry IV of France as ‘the miracle of Holland’. Deciding on a legal career, he had become Advocate General of Holland, Zealand and West Friesland by the age of 24. In this period he wrote a treatise on the Law of Prize, of which the part dealing with freedom of the seas (*Mare Liberum*) was published in 1609. Because of his support for the moderate Arminians against the Calvinists, he was in 1619 imprisoned in Loevestein castle, and while there wrote an introduction to the law of Holland (*Inleidinge tot de Hollandsche Rechtsgeleertheyd*, published in 1631) and a tract on the truth of the Christian religion, the first of many theological writings. After two years, his wife arranged his escape in a chest ostensibly holding books, and thereafter he lived mainly in France, where he served for ten years as ambassador of Sweden.

Grotius' greatest work is *De iure belli ac pacis* (On the law of war and peace), published in 1625 and widely translated (six editions appeared in English before 1750). Written during the upheavals of the Thirty Years War, it laid down certain fundamental principles of law which purported to have the certainty of mathematics and absolute validity in all times and in all places. These principles both provided a standard for measuring the validity of the positive law of any state and also formed the basis for governing the relations between states. The work had enormous influence on the ethical and legal thought of the 17th and 18th centuries and is regarded as the beginning of ‘the law of nature and of nations’, the forerunner of modern international law.

Grotius built on the learning of late scholastic writers on natural law, such as Suarez, but he tried to make it independent of theological doctrine, so that amid the factionalism of the Reformation its principles would be unaffected by conflicting religious views. For him these principles could be proved in two ways: a priori, by logically deducing them from the rational and social nature shared by all

mankind, and a posteriori, by showing that they were generally accepted by the consensus of writers – at least in more civilized nations – through the ages. For when many writers ‘at different times and different places affirm the same thing as certain, that ought to be referred to a universal cause’, which must be either correct conclusion drawn from the principles of nature or common consent (Prolegomena, sec. 40). Grotius concentrated on the latter approach, and dealt particularly with property and contract, the area of law of most concern to market societies and to nation states dealing with each other at arm's length.

Relying on the Bible narrative and on accounts of American Indians, he envisaged a primitive state of nature in which everything was held in common. When primitive simplicity gave way to specialization in agriculture and cattle raising, the conflicts that arose led first to division of lands among nations and then to division among families; thus community of property was replaced by private property. ‘This happened not by a mere act of will … but rather by a kind of agreement, either expressed, as by a division, or implied, as by occupation’ (*De iure belli* II. 2, 21–5).

His doctrine of contracts was loosely based on Aristotle. He tolerated monopolies but condemned combinations to raise prices or to prevent the movement of goods by fraud or force. Although the law of nature did not forbid usury, divine positive law forbade it for Christians. However, Grotius adopted the canonist distinction between usury, which was forbidden, and receiving interest, which was permissible, if the rate was reasonable as in the positive law of Holland.

Bibliography

The most convenient modern edition of *De iure belli ac pacis* is in Classics of International Law Series, Oxford, 1925 (vol. 1: the Latin text of 1646; vol. 2: English translation by F.W. Kelsey). General surveys are in W.S.M. Knight, *Life and Works of Hugo Grotius*, London: Sweet & Maxwell, 1925, and E. Dumbauld, *Life and Writings of Hugo Grotius*, Norman: University of Oklahoma Press, 1969. Full bibliographies are in the annual volumes of *Grotiana* (New Series), Assen, Netherlands, 1980 onwards.

How to cite this article

Stein, P. G. "Grotius (de Groot), Hugo (1583–1645)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000121> doi:10.1057/9780230226203.0681

The New Palgrave Dictionary of Economics Online

group selection

Arthur J. Robson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Group selection’ is the biological term for the possibility that a characteristic that is beneficial to the group but possibly costly to the individual is evolutionarily successful. Although logically possible, it is generally viewed with scepticism by biologists. It is not problematic that group selection would favour an equilibrium whose payoffs dominated those of another, because there is then no conflict with individual selection within each group. Group selection might reject inefficient equilibria in a repeated game, for example. Since human societies can support rather arbitrary outcomes as equilibria, group selection could play a role in human evolution.

Keywords

altruism; assortative matching; cooperation and its evolution; cultural transmission; Darwin, C.; evolution; exchange; group selection; hunter–gatherer economies; individual selection; kin selection; natural selection; other-regarding preferences; prisoner's dilemma; punishment; repeated games; trust

Article

It is uncanny how close Darwin (1859) came to the modern view of biological evolution, given that detailed understanding of the mechanics of genetic inheritance lay in the future. Darwin understood how mutations might arise randomly rather than in response to circumstances. Further, he espoused the modern dogma concerning the separation of the germ line (sex cells) and the somatic line (all other bodily cells). Under this dogma, which explicitly contradicts the earlier biologist Lamarck, only those characteristics present in the germ line, not those acquired during the individual's lifetime, are genetically inherited by offspring. If a germ line characteristic produces a somatic or behavioural attribute that is best suited to the ecological or social circumstances of the individual (yielding the most offspring), then the attribute and the underlying genetic characteristic will become more common. Darwin typically emphasized that a particular variation would spread if this variation led to greater

reproductive success for individuals and were inherited by their descendants. However, Darwin also strayed occasionally into what would now be called ‘group selection’, especially when he was considering the implications of his theory for humans (Darwin, 1871, p. 166). Thus, he thought an individual human might engage in behaviour that is beneficial to the survival of a group, even if this behaviour had a fitness cost to the individual. It is of interest that Darwin refers to humans in particular, since it is still sometimes argued that group selection might be important for our own species.

The group selection debate in biology

There is a ‘folk wisdom’ appeal to group selection in biology. This mechanism was once invoked in popular accounts of natural selection. For example, the idea that a predator species is doing a prey species a favour by eliminating its weakest members involves an egregious form of group selection. The English experimental biologist Wynne-Edwards provided an especially explicit manifesto on group selection and became thereby a favourite target for those wishing to argue against it. In Wynne-Edwards (1962), for example, he argued that birds limit the size of their clutches of eggs to ensure that the size of the population does not exceed the comfortable carrying capacity of the environment.

These particular group selection arguments were effectively devastated by Williams (1966). If a new type of individual does not so obligingly limit her clutch, why would this more fertile type not take over the population, without regard for the standard of living? Indeed, there are compelling arguments why it is in the interests of the individual to limit her clutch size. For example, it might well be that, beyond a certain point, an increase in the number of eggs reduces the expected number of offspring surviving to maturity, because each egg then commands a reduced share in parental resources. A finite optimum for clutch size is then to be expected.

Dawkins (1976 and 1982, for example) has been even more insistent than Williams on rejecting group selection, going further in arguing for the primacy of the gene as a still lower-level unit of selection. There certainly are phenomena best understood at the level of the gene. Consider, for example, sickle-cell anemia. At the relevant locus, there exists a particular variant gene, a particular allele, that is. If one of the two alleles that are present at this locus is this variant gene, the individual has improved resistance to malaria. However, if both alleles are of this variant type, the red blood cells have a characteristic sickle shape. Such cells malfunction by not carrying adequate oxygen, implying increased mortality. Under sexual reproduction, there is no way of maintaining only the individuals who have exactly one copy of the variant gene. Rather, both alleles are maintained as a stable mixture, where each allele is present in individuals who have quite different fitnesses.

There are presumably a fair number of cases where the interests of the gene and the individual do not conflict. In any case, it is often difficult to give concrete form to the gene as the unit of selection, given our ignorance of the details of the transformation of genes into individuals, particularly for complex behavioural characteristics. (Grafen, 1991, advocates finessing such detailed questions on the genetic basis of individual variation. This is his so-called ‘phenotypic gambit’.) Hence, despite the theoretical primacy of the gene, we restrict attention here to the comparison between the individual level and the group level of selection.

In order to fix ideas, consider the following outline of the classic model that addresses the issue of individual selection versus group selection.

The haystack model

The following is a simplified account of Maynard Smith (1964). There are a number of haystacks in a farmer's field, each of which is home to two mice. Each pair of mice now play the Prisoner's Dilemma game, with the usual two choices: cooperate or defect. Payoffs for each individual take the concrete form of the number of offspring, but reproduction is asexual, with offspring inheriting their mother's choice of strategy. There are a number of subsequent stages of play, where the mice in each haystack are paired at random to play the Prisoner's Dilemma game. The number of individuals within the haystack choosing each strategy then grows in an endogenous fashion, as does the overall size of the group. Every so often, once a year, say, the haystacks are removed, and the mice are pooled into a single large population. Now pairs of mice are selected at random from the overall population to re-colonize the next set of haystacks, and excess mice die.

Maynard Smith's intention here was to give the devil his due, by building a model in which group selection might well have an effect. At the same time, he wished to show that the parametric assumptions needed to make group selection comparable in strength to individual selection would be unpalatable. In order for group selection to be effective there must be a mechanism that insulates the groups from one another. Only then can a cooperative group be immune to infection by a defecting individual, and maintain its greater growth rate. Even with the temporary insulation of each haystack in this model, cooperation will evolve only if there are sufficient rounds of play within each haystack, so that defectors from a particular haystack are likely to be eliminated when the groups are reformed.

A loose description of the problem with group selection is that it relies too heavily upon a group becoming extinct as a likely consequence of a choice that is bad for the group. There is clearly scope, in reality, for individual selection, since individuals die frequently; group selection is less plausible, since there may not be enough extinction of groups.

An example of group selection?

Despite the disfavour into which group selection has fallen in biology, there remain examples of cases that are challenging to explain otherwise. One of these concerns the interaction of the myxoma virus and European rabbits in Australia. (Lewontin, 1970, proposed a group selection interpretation of this case. Sober and Wilson, 1999, pp. 45–50, give a – somewhat partisan – view of the subsequent debate.)

English rabbits were introduced into Australia in a misguided attempt to recreate the English countryside in Australia, but their numbers grew out of control, causing massive economic damage to farms there. The myxoma virus was first identified in South American forest rabbits, where it was only mildly virulent. When this South American strain was originally tested on Australian rabbits, however, it seemed an ideal solution to the rabbit infestation there, since it killed nearly the entire sample. Unfortunately, in the long run, after the South American virus had been present in the Australian rabbit population for a while, the rabbit mortality rate fell dramatically.

Why? Perhaps the most obvious explanation is that the rabbit population had been selected to have greater resistance to the virus, consistent with individual selection of rabbits. That this was true to some extent was demonstrated by the finding that the new Australian strain of the virus had a greater effect on laboratory-bred Australian rabbits than on the feral stock. However, this effect was not sufficient to

explain the entire drop in rabbit mortality. Indeed, both laboratory and wild strains of Australian rabbits were less susceptible to the new Australian strain of the virus than they were to the original strain imported from South America. The virus had evidently evolved to be less virulent as a result of its interaction with the Australian rabbit population.

This situation might be roughly mapped onto the haystack model as follows. Consider a group of viruses to be those infecting a given rabbit. The evolutionary success of this group might best be promoted by settling for a moderate level of mortality for the host rabbit. Whatever the other advantages to the virus of strategies that induce high mortality in the rabbit, the early death of the current host makes transmission to a new host difficult. However, prolonging the life of the host is a public good from the point of view of the infecting viruses. A strain of virus with a strategy that was more lethal to the host could then grow as a fraction of the group of viruses. As in the haystack model, however, if there were enough generations of the virus within each rabbit's lifetime, this conflict between the group and the individual might be resolved in favour of the group, as suggested by the data.

Selection among equilibria

When does group selection matter in biology? In theory, it could lead to different results than would individual selection, as in the Prisoner's Dilemma, and as it may in above example. In practice, the above example is atypical, and group selection is usually a rather weak force. There is, however, one compelling scenario in which group selection would operate robustly, in any species. This is as a mechanism to select among equilibria (Boyd and Richerson, 1990). Consider a population that is divided into various sub-populations, which are largely segregated from one another, so that migration between sub-populations is limited. Each sub-population plays the same symmetric game, which has several symmetric equilibria. Play of this game involves a random matching of the members of each sub-population. Individual selection ensures that some equilibrium is attained, within each sub-population. But group selection is then free to operate in a leisurely fashion to select the Pareto-superior equilibrium, since there is no tension here between the two levels of selection.

Group selection and economics

Why does group selection matter in economics? Group selection is the most obvious mechanism for generating preferences in humans that might make them behave in the social interest rather than that of the individual. At stake, then, is nothing less than the basic nature of human beings. As an economist, one should be sceptical of the need to suppose that individuals are motivated by the common good. Economic theory has done well in explaining a wide range of phenomena on the basis of selfish preferences, and so the view of the individual as the unit of selection is highly congenial to economists. Furthermore, to the extent that armchair empiricism suggests that non-selfish motivations are sometimes present, these seem as likely to involve malice as to involve altruism. For example, humans seem sometimes motivated by relative economic outcomes, which involve such a negative concern for others. Group selection is a blunt instrument that might easily ‘explain’ more than is true.

There are, nevertheless, some aspects of human economic behaviour that are tempting to explain by group selection. For example, we have a proclivity for trade that may go beyond myopic self-interest. As

Darwin, an astute observer of both human beings and the natural world, observed on one of his visits to Tierra del Fuego,

> Some of the Fuegians plainly showed that they had a fair notion of barter. I gave one man a large nail (a most valuable present) without making any signs for a return; but he immediately picked out two fish, and handed them up on the point of his spear. (Darwin, 1845, ch. 10)

That is, human beings are often willing to trade with strangers they will likely never see again, as might be analogous to cooperating in the one-shot Prisoner's Dilemma. There is no shortage of reliable data showing that human beings are capable of such apparently irrationally cooperative behaviour, in appropriate circumstances. Whatever the underlying reasons for this, it is clearly significant, and may even help account for the prodigious economic and biological success of humans.

Furthermore, identifying the underlying reasons would help shape a theory of such behaviour that remains falsifiable; such a theory might also predict what alternative circumstances might induce non-cooperative or antagonistic behaviour. Group selection is an obvious avenue to explore in this connection.

It is sometimes argued, in particular, that the structure of hunter–gatherer societies helps account for cooperative behaviour. Hunter–gatherer societies were composed of a large number of relatively small groups, and individuals within each group were often genetically related. Perhaps, so the argument goes, we acquired an inherited psychological inclination towards conditional cooperation in such a setting, partly perhaps as a result of group selection. These inclinations then carried over into modern societies, despite genetic relatedness now being essentially zero on average. Seabright (2004), for example, argues eloquently that human societies cannot function on myopic self-interest alone, but also that the trust needed for exchange in market economies sprang from adaptations to archaic small groups. It is hard to believe, however, that hunter–gatherers never encountered strangers. If there were good reasons to condition on this distinction, why would corresponding different strategies not have evolved?

A phenomenon that looms large in the case of human beings is culture, by which is meant the non-genetic transmission of behaviour, by imitation of peers, for example. Many attempts to derive cooperative behaviour have focused then on group selection in models of cultural transmission. We now turn to this literature.

Cultural group selection and economics

A spectacular and famous example of cultural group selection features the Nuer and Dinka, who lived as neighbouring ethnic groups in 18th century southern Sudan (Kelly, 1985). Despite the similarity of their environment, these two groups differed in various economic and political respects. Perhaps the key difference was that Dinka lived in small groups, the size of which was related to the needs of their economic activity. The Nuer, on the other hand, organized their society according to a patrilineal system that bound many such smaller units together, over a greater geographic area. Over a period of 100 years, the Nuer expanded their territory at the expense of the Dinka, killing or assimilating their rivals. Nuer culture, that is, was selected over that of the Dinka.

Despite the apparently strong military advantages of the Nuer political system over that of the Dinka, it seems plausible that any individual – Dinka or Nuer – would have had the incentive to play the appropriate role within their society. It would not have been possible for an individual Dinka to shift unilaterally to Nuer behaviour.

Human societies have the capacity to render a wide range of behaviour optimal for the individual. If a society wishes to adopt a rather arbitrary rule, it may well have adequate sanctions to enforce this (Boyd and Richerson, 1992). As described above, group selection can then be relied upon to select between various equilibria. Boyd et al. (2003) make the important additional point that, because punishment is needed only infrequently near full cooperation, the weak force of group selection can support cooperation as an equilibrium, without the usual need for punishment of those who fail to punish, and so on.

Group selection is uncontroversial as a mechanism for selecting an efficient outcome within each group. But this does not directly explain observations, such as Darwin's, of our apparent willingness to trade with strangers. The difficulty is stark: defection is a strictly dominant strategy in the one-shot Prisoner's Dilemma game.

One stark option is that cooperation is hard-wired. Bowles, Choi and Hopfensitz (2003) argue that, if behaviour is directly genetically controlled, cooperative behaviour may then be sustained in the presence of culturally maintained institutions. These institutions, food-sharing for example, serve to reduce the negative impact on individuals of cooperative behaviour. Group selection arises from conflict between groups, with more cooperative groups emerging victorious in such conflict.

However, human strategic behaviour is genetically mediated in a complex and poorly understood way, and is not always cooperative. Even if we did somehow acquire a genetic inclination to cooperate in archaic societies, shouldn't we now be in the process of losing this inclination in modern large and anonymous societies?

A recent revival?

Sober and Wilson (1999) push energetically for a rehabilitation of group selection within biology. They argue that group selection is closely related to other well-accepted phenomena. For example, they argue that kin selection – the widely accepted notion that individuals are selected to favour their relations – should be regarded as a special case of group selection.

In its simplest form, kin selection is the argument that individuals should undertake actions that benefit a relation if this benefit, when deflated by the degree of relatedness, exceeds the cost to the first individual. (This is ‘Hamilton's rule’ as in Hamilton, 1964.) The empirical evidence in favour of kin selection is overwhelming – mothers, human and non-human, routinely make large sacrifices in favour of offspring. Even economists would exempt such interactions from the general presumption of selfish behaviour.

Sober and Wilson certainly make the case that these phenomena can be viewed in an integrated fashion. Indeed, it is a consequence of the ‘Price equation’ (Price, 1970) that what matters most fundamentally is the likelihood that altruistic individuals will be preferentially matched with other altruistic individuals. In the case of kin selection, this could be ensured by a mechanism to directly recognize relations – by smell, for example – or by indirect but reliable methods, such as, for example, assuming that those who

are found in proximity to one's mother are relations.

Bergstrom (2002) provides the best introduction to the literature on group selection for economists. He presents a unified and intuitive treatment of the literature, and also stresses the key role of assortative matching. Thus, for example, if the subgroups in the haystack model are dispersed after one round of the game, and then randomly recombined, there is no force to group selection. Only if the subgroups remain together for repeated play of the game is there effective assortative matching. Such a structure seems less compelling for non-relations than it is for relations.

Despite the formal analogies between kin selection and group selection, acceptance of the former does not then require acceptance of the latter. In the end, a sceptical but not dogmatic view of the importance of group selection to human economic behavior seems warranted.

See Also

- game theory and biology
- hunting and gathering economies
- learning and evolution in games: an overview

I received helpful comments from Ted Bergstrom, Lawrence Blume, Sam Bowles, Steven Durlauf and Peter Sozou. I thank them without blaming them.

Bibliography

Bergstrom, T.C. 2002. Evolution of social behavior: individual and group selection. *Journal of Economic Perspectives* 16(2), 67–88.

Bowles, S., Choi, J.-K. and Hopfensitz, A. 2003. The coevolution of individual behaviors and social institutions. *Journal of Theoretical Biology* 223, 135–47.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P.J. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* 100, 3531–5.

Boyd, R. and Richerson, P. 1990. Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology* 145, 331–42.

Boyd, R. and Richerson, P. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13, 171–95.

Darwin, C. 1845. *Journal of Researches into the Natural History and Geology of the Various Countries Visited by H. M. S. Beagle*. London: Henry Colburn. Online. Available at http://www.galapagos.to/TEXTS/J-OF-R.HTM, accessed 23 October 2006.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection*. London: John Murray. Online. Available at http://darwin-online.org.uk/EditorialIntroductions/Freeman_OntheOriginofSpecies.html, accessed 8 November 2006.

Darwin, C. 1871. *The Descent of Man and Selection in Relation to Sex*. London: John Murray. Online. Available at http://darwin-online.org.uk/EditorialIntroductions/Freeman_TheDescentofMan.html, accessed 8 November 2006.

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Dawkins, R. 1982. *The Extended Phenotype: The Gene as the Unit of Selection*. Oxford: Oxford University Press.

Grafen, A. 1991. Modelling in behavioural ecology. In *Behavioral Ecology: An Evolutionary Approach*, ed. J.R. Krebs and N.B. Davies. 3rd ed, Oxford: Blackwell Scientific.

Hamilton, W.D. 1964. The genetical evolution of social behavior. *Journal of Theoretical Biology* 7, 1–16.

Kelly, R.C. 1985. *The Nuer Conquest: The Structure and Development of an Expansionist System*. Ann Arbor: University of Michigan Press.

Lewontin, R.C. 1970. The units of selection. *Annual Reviews of Ecology and Systematics* 1, 1–18.

Maynard Smith, J. 1964. Group selection and kin selection. *Nature* 201, 1145–7.

Price, G. 1970. Selection and covariance. *Nature* 277, 520–1.

Seabright, P. 2004. *The Company of Strangers*. Princeton: Princeton University Press.

Sober, E. and Wilson, D.S. 1999. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Williams, G.C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton: Princeton University Press.

Wynne-Edwards, V.C. 1962. *Animal Dispersion in Relation to Social Behavior*. Edinburgh: Oliver and Boyd.

How to cite this article

Robson, Arthur J. "group selection." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000195> doi:10.1057/9780230226203.0682

The New Palgrave Dictionary of Economics Online

growth accounting

Francesco Caselli

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Growth accounting consists of a set of calculations resulting in a measure of output growth, a measure of input growth, and their difference, most commonly referred to as total factor productivity (TFP) growth. It can be performed at the level of the plant, firm, industry, or aggregate economy. This article discusses the theoretical interpretation of the growth-accounting exercise, problems of measurement, and main empirical results. It concludes with a (very selective) history of the field.

Keywords

growth accounting; input–output analysis; measurement error; national accounts; production functions; research and development; specification error; technical change; total factor productivity; vintage capital

Article

Growth accounting consists of a set of calculations resulting in a measure of output growth, a measure of input growth, and their difference, most commonly referred to as total factor productivity (TFP) growth. It can be performed at the level of the plant, firm, industry, or aggregate economy.

Current growth-accounting practice tends to rely on the theoretical construct of the production function both as a guide for measurement and as a source of interpretation of the results. Apart from the existence of a production function linking inputs and outputs, the main assumption is that factors of production are rewarded by their marginal product. In continuous time, this permits a representation of output growth as a weighted sum of the growth rates of the inputs, and an additional term that captures shifts over time in the production function. The weights for the input growth rates are the respective shares in total input payments. Since data on the growth of output and individual input quantities cover discrete periods of time, a discrete-time approximation to the weights is required. Current practice tends to use simple averages of the input shares at the beginning and the end of each period. In the special case that the production function is of the translogarithmic form, this procedure actually results in an exact decomposition; otherwise it can be interpreted as a second-order approximation.

It is customary to group inputs into broad categories. When output is measured as value-added the broad categories are labour and capital. When output is total production one has to add materials, which are occasionally further broken down with further entries for energy and services (giving rise to the so-called KLEMS accounting framework). This kind of grouping allows one to speak of, for example, the ‘contribution of labour (capital, materials) to output growth’. However, this grouping masks an enormous heterogeneity of the underlying inputs. This heterogeneity is the source of a large share of the measurement problems in growth accounting. These problems are most severe in the measurement of the growth of capital input.

Capital inputs are heterogeneous within vintages (for example, tractors versus personal computers) and across vintages (computers produced in 2006 versus computers produced in 2007). Heterogeneity within vintages is best addressed by having as fine a disaggregation of capital types as the data will allow. The most important data constraint on disaggregation of capital types occurs in the construction of type-specific shares in total capital income, as these require type-specific estimates of rental rates, which in turn require type-specific estimates of depreciation rates, capital gains and tax treatment. Heterogeneity across vintages, also known as embodied technical change, or quality change, poses even more difficult problems. Most practitioners' ideal solution to this problem is to put the measurement of the stocks of different types of capital on a constant-quality basis, by applying appropriate deflators reflecting quality change to the corresponding investment series. However, the availability and/or accuracy of such deflators, whose construction generally requires hedonic methods, is currently limited for most countries, industries and capital types. As a result there is a presumption that the growth rate of (the efficiency units embodied in) the capital stock is often understated.

Construction of indices of labour input growth have conceptually similar problems. However, aggregation across types (for example, female, white, high-school graduates, of age 40 to 45 versus male, black, college graduate of age 35 to 40) is simpler as average rental rates (that is, hourly wages) for reasonably fine categories are reasonably well observed (while in the case of capital goods they must be estimated). The vintage problem is typically bypassed by assuming that there is no quality change within narrowly defined categories.

Another difficult problem is how to turn the growth in input stocks into growth in the flow of input services, that is, how to account for variation in the rate of utilization of labour and capital. Measuring labour in hours is helpful, but an issue of utilization still remains if effort per hour is not constant, as is likely. For capital, various adjustments based on proxies for utilization have been proposed, a classic one being a measure of electricity consumption. But this approach to the problem of measuring utilization creates a deeper problem of interpretation, or at least a conflict with the estimate of rental rates. This is because the latter are constructed in a way that assumes them to be invariant to the rate of utilization. But in this case the opportunity cost of setting the utilization rate to 100 per cent all the time is nil, and there should be no variation in utilization. Some more systematic adjustment to the theoretical framework, such as endogenous depreciation or limited opportunity for substitution between capital and other inputs, is therefore required to fully solve the measurement and interpretation challenges posed by variable utilization.

At the plant, firm, and industry levels a choice can be made between accounting for total production or value added. The total production approach is attractive, because after all it is total production that ‘comes out’ of the production process. Furthermore, the conditions for existence of a well-defined production function for total output are far less stringent than the conditions for existence of a function

linking valued added to capital and labour inputs. On the other hand, the results of growth-accounting exercises based on total output are very sensitive to the degree of vertical integration, and this causes severe problems of interpretation.

At the country level value added is obviously the only meaningful concept of output, no matter how stringent the conditions for an aggregate valued-added function. Because of the well-known shortcomings of standard measures of valued added as indicators of the ‘want-satisfying’ capacity of the economy, some attempts have been made to augment such measures by estimates of non-market outputs, chiefly the output of the education sector. Accounting for the effects of economic activity on the natural environment is very probably the next frontier.

Among the outputs of the growth-accounting calculation the one to receive most attention is usually the difference between output and input growth. This is somewhat surprising because the interpretation of this quantity is fraught with difficulties, as underscored by the multitude of phrases used to refer to this difference: besides TFP growth, ‘multi-factor productivity’ growth, ‘(Solow) residual’, ‘measure of our ignorance’, ‘rate of technical change’, and growth in ‘output per unit of (total) input’, among others.

What is sometimes misunderstood is the relationship between the difference and technical change. An economic unit can use additions to its capital and labour either to directly produce more output, or to devise ways to rearrange the existing capital and labour so as to produce more (constant-quality units of) output, the latter being the definition of research and development (R&D). If it does so by equating the marginal products of labour and capital between direct production of output and indirect production of output through R&D, the extra output produced thanks to R&D will be fully ‘accounted for’ by the measured growth in capital and labour inputs. Hence, TFP growth does not really measure technical change as this term is commonly understood. Furthermore, failures of TFP growth to accelerate in periods/industries/firms experiencing increases in R&D spending do not need to be puzzling.

For the same reasons, TFP growth can be identified neither with disembodied nor with embodied technical change. Embodied technical change in capital-using industries is a reflection of disembodied technical change in capital-producing industries, but neither need necessarily show up in the TFP numbers, as long as R&D costs have been properly accounted for.

So what does show up? Under the maintained theoretical assumptions, the cleanest interpretation – apart from weather shocks, and costless, instantaneous flashes of inspiration (if they were not instantaneous an opportunity cost of time would have to be imputed) or innovations stumbled upon by luck, none of which seems susceptible to vary much over time and space, or with government policy – is R&D externalities. If the units performing R&D fail to capture all the social return from it, other units will experience costless growth in output per input, and this will be detected by TFP growth. Under this interpretation, a link may indeed be found between R&D and TFP growth, and if so it would be possible to use the framework to advocate policies to encourage R&D. Other forms of externalities may also give rise to positive TFP growth.

But since TFP is a residual, it also picks up, as all residuals, errors of specification and measurement. We have already discussed mismeasured input growth, chiefly in terms of incomplete adjustment for quality change. A failure to account for quality change in output will push TFP growth in the opposite direction. Note that mismeasured quality change in capital results in lower TFP growth in capital-producing industries, higher TFP growth in capital-using industries, and somewhat ambiguous effects on TFP at the aggregate level, though the net effect is usually deemed to be positive.

Many economies are likely to be characterized by frictions to the efficient allocation of resources among

economic units, implying that marginal products of homogeneous inputs are not equalized. In these cases improvements in the allocation of resources will also result in positive TFP growth.

It is impossible to overestimate the interest that growth-accounting calculations have elicited. There must be very few industries and countries for which some kind of input–output data exists that has not been used for performing a growth accounting exercise. Indeed, several national statistical agencies explicitly include the output of growth-accounting calculations, including TFP growth, into the national accounts.

I am unable to provide here an overview of this immense body of work, and the reader will have to refer to the country/industry/period of interest on a case-by-case basis. However there are a couple of broad lessons that can be distilled here. First, over the medium to long term, the residual accounts for a relatively minor portion of overall growth in output. For example for the United States it is possible to explain about two-thirds of growth in (market) output per worker over the post-war period by changes in the quality and quantity of inputs. For countries experiencing exceptionally high growth rates, such as the Asian Tigers between 1960 and 1990, this share is even higher. To the extent that the residual picks up measurement and specification errors, this is tantamount to saying that the performance of the growth-accounting methodology is very good by the standards of empirical work in economics. This interpretation is reinforced by the fact that, again by and large, the role of the residual tends to be systematically smaller in studies deploying better quality data.

Over shorter horizons, however, TFP growth is harder to underplay. For example a slowdown in TFP growth ‘accounts’ for a large fraction of the slowdown in output growth observed between the mid-1970s and the mid-1990s. Not coincidentally, the root causes of that slowdown remain as mysterious as ever.

While growth-accounting calculations can be performed at various levels of aggregation, and their interpretation is perhaps easier the smaller the unit of analysis, the origins of growth accounting are macroeconomic. The earliest growth-accounting exercises (Stigler, 1947; Schmookler, 1952; Abramowitz, 1956 – the latter also coining the expression ‘measure of our ignorance’) were a direct byproduct of the development of US aggregate national account data. One exception was agriculture, for which early growth-accounting experiments date to 1948 (Barton and Cooper) and 1951 (Kendrick and Jones). Kendrick (1956; 1961) compiled the first large-scale growth-accounting calculations broken down by many industries. He also introduced the phrase ‘total factor productivity’.

Solow (1957) laid out the theoretical foundations of growth accounting (a previous contribution in this direction by Tinbergen, 1942, with attendant calculations, was discovered by the English-language literature only subsequently). Solow (1960) and Jorgenson (1966) worked out the implications of embodied technical change. Denison (1962) introduced corrections for changes in the composition of the labour force. Griliches and Jorgenson (1966) and Jorgenson and Griliches (1967) put aggregation of inputs and outputs on a solid theoretical basis, particularly by showing how to correctly estimate rental rates. They also pioneered empirical approaches to quality change and variable utilitization. This programme was further refined by Christensen and Jorgenson (1969; 1970) for the aggregate economy, and Fraumeni, Gollop and Jorgenson (1987) for a broad set of industry-level calculations which has shaped the way US national accounts are now constructed, and whose methods are widely accepted to be the gold standard for the purposes of productivity measurement.

Christensen, Jorgenson and Lau (1973) developed the translogarithmic production frontier, and Diewert

(1976) showed that with translog production functions discrete-time approximations are no longer approximations. Jorgenson and Fraumeni (for example, 1992) attempted accounting for the output of the education sector. Young (1995) performed an influential growth-accounting exercise for the East Asian tigers.

See Also

- level accounting

Bibliography

Abramowitz, M. 1956. Resource and output trends in the United States since 1870. *American Economic Review* 46, 5–23.

Barton, G. and Cooper, M. 1948. Relation of agricultural production to inputs. *Review of Economics and Statistics* 30, 117–26.

Christensen, L. and Jorgenson, D. 1969. The measurement of U.S. real capital input, 1929–1967. *Review of Income and Wealth* 15, 293–320.

Christensen, L. and Jorgenson, D. 1970. U.S. real product and real factor input, 1929–1967. *Review of Income and Wealth* 16, 19–50.

Christensen, L., Jorgenson, D. and Lau, L. 1973. Trascendental logarithmic production frontiers. *Review of Economics and Statistics* 55, 28–45.

Denison, E. 1962. *The Sources of Economic Growth in the United States and the Alternatives before Us*. New York: Committee for Economic Development.

Diewert, E. 1976. Exact superlative index numbers. *Journal of Econometrics* 4, 115–45.

Fraumeni, B., Gollop, F. and Jorgenson, D. 1987. *Productivity and U.S. Economic Growth*. Cambridge, MA: Harvard University Press.

Griliches, Z. and Jorgenson, D. 1966. Sources of measured productivity change: capital input. *American Economic Review* 56, 50–61.

Jorgenson, D. 1966. The embodiment hypothesis. *Journal of Political Economy* 74, 1–17.

Jorgenson, D. and Griliches, Z. 1967. The explanation for productivity change. *Review of Economic*

Studies 34, 249–83.

Jorgenson, D. and Fraumeni, B. 1992. Investment in education and U.S. economic growth. *Scandinavian Journal of Economics* 94(Suppl.), 51–70.

Kendrick, J. 1956. Productivity trends: capital and labor. *Review of Economics and Statistics* 38, 248–57.

Kendrick, J. 1961. *Productivity Trends in the United States*. Princeton: Princeton University Press.

Kendrick, J. and Jones, C. 1951. Gross national farm product in constant dollars, 1910–50. *Survey of Current Business* 31, 13–19.

Schmookler, J. 1952. The changing efficiency of the American economy, 1869–1938. *Review of Economics and Statistics* 34, 214–31.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

Solow, R. 1960. Investment and technical progress. In *Mathematical Methods in the Social Sciences, 1959*, ed. K.J. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

Stigler, G. 1947. *Trends in Output and Employment*. Cambridge, MA: NBER.

Tinbergen, J. 1942. Zur Theorie der Langfristigen Wirtschaftsentwicklung. *Weltwirtschaftliches Archiv* 55, 511–49.

Young, A. 1995. The tyranny of numbers: confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 110, 641–80.

How to cite this article

Caselli, Francesco. "growth accounting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000126> doi:10.1057/9780230226203.0683

The New Palgrave Dictionary of Economics Online

growth and civil war

Paul Collier

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capital flight; civil war; economic growth; foreign aid; human capital; predation; social cost; spillovers

Article

Civil war is obviously damaging for both society and the economy. The social consequences are often difficult to measure: for example, people die as a result of disease and are traumatized through rape or experience as child soldiers. However, the consequences for the economy are much more amenable to quantification, and a key economic research issue has been to try to classify and quantify the economic damage.

There are three consequences of civil war for economic growth. The most obvious is the loss of growth experienced by the country during the war. War is destructive of both the capital stock and normal economic activity. In addition to inflicting purposive destruction, rebel forces typically prey on economic activity within their military reach, since they have no official sources of finance. Government forces may adopt the same tactics: typically, lines of command are too loose to prevent the diversion of military force into decentralized predation. In such circumstances citizens with assets radically reduce their investment and resort to capital flight abroad, and poorer citizens may similarly retreat into subsistence activities that are less vulnerable. The consequences of civil war for the growth of gross domestic product (GDP) have been estimated from time series regressions, a typical result being that the annual growth rate is reduced by around 2.3 percentage points (Collier, 1999). Since the typical civil war lasts around seven years, by the end of the conflict the country is around 15 per cent poorer than it would have been had it remained at peace.

The second consequence for growth is the legacy of war – effects that arise after the war has ended. Peace does not usually enable the economy to rebound swiftly to its previous growth path. Even if the post-conflict peace is secure the economy will take several years to rebuild its capital stock, and some costs, such as a loss of human capital, may be irreversible. More typically, the legacy of civil war creates major problems for economic recovery. The peace itself may be insecure: around 40 per cent of post-conflict situations revert to conflict within a decade. The breakdown in social order during civil war

allows opportunistic behaviour to become more prevalent. Both the high macro-risk of conflict reversion and the high micro-risk of being the victim of opportunism inhibit investment, and capital flight typically continues. These effects dampen and can potentially prevent post-conflict recovery. Often, post-conflict growth exceeds normal growth by around 1.1 per cent, so that after a war the economy needs about double the length of time of the civil war before it rejoins its long-term path (Collier and Hoeffler, 2004a). Until then the country is poorer than it would have been without the war, and so this shortfall should properly be counted as a cost. Because recovery is so slow, most of the costs of a civil war accrue after it is over. Aid accelerates growth during recovery and has been found to be particularly effective in post-conflict situations. This was indeed the original rationale for aid: the World Bank was initially going to be named the ‘International Bank for Reconstruction’.

The third consequence for growth is the effect of civil war on other countries. Typically, the adverse effects of a civil war spread far beyond the borders of the afflicted country: even countries that are not direct neighbours suffer significant reductions in growth (Murdoch and Sandler, 2002). Although the reduction in the growth rate of neighbours is much lower than that experienced by the country itself, because many countries are affected, the overall ‘spillover’ costs to neighbours arising from this loss of growth tend to exceed those experienced by the country itself.

Taken together, these consequences for growth have two important implications. One is that the true economic cost of civil war is massive even before one takes into account the social costs. By applying a conventional discount rate to the lost growth and allowing only for costs to direct neighbours, Collier and Hoeffler (2004b) estimate the value of the typical cost of a civil war in a low-income country at an astounding 64 billion dollars. The other implication is that most of these losses are externalities to the people taking the decisions at the time of the conflict: the losses accrue to neighbours and to a future generation. Hence, decisions to start civil wars are unlikely to reflect a true social calculus of the probable consequences of war.

Bibliography

Collier, P. 1999. On the economic consequences of war. *Oxford Economic Papers* 51, 168–83.

Collier, P. and Hoeffler, A. 2004a. Aid, policy and growth in post-conflict societies. *European Economic Review* 48, 1125–45.

Collier, P. and Hoeffler, A. 2004b. Conflict. In *Global Crises, Global Solutions*, ed. B. Lomborg. Cambridge: Cambridge University Press.

Murdoch, J.C. and Sandler, T. 2002. Economic growth, civil wars and spatial spillovers. *Journal of Conflict Resolution* 46, 91–110.

How to cite this article

Collier, Paul. "growth and civil war." The New Palgrave Dictionary of Economics. Second Edition. Eds.

Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000175> doi:10.1057/9780230226203.0684

The New Palgrave Dictionary of Economics Online

growth and cycles

Gadi Barlevy

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There is a long tradition in macroeconomics of treating growth and cycles as distinct phenomena. However, various economists have also recognized the virtue of incorporating the two forces into a single framework and to study the way they are related. This article reviews this literature, with emphasis on attempts not only to integrate growth and cycles into a single framework but also to endogenize growth, cycles, or both.

Keywords

capital accumulation; endogenous growth; growth and cycles; innovation; Kydland, F.; Prescott, E.; Ramsey model; research and development; risk; technical change; precautionary savings

Article

Growth and cycles are two key features that characterize real output per capita in most industrialized countries. Real output per capita grows systematically over time; and the rate at which it grows tends to fluctuate over time.

A long tradition in macroeconomics treats these two features as distinct. On the one hand, economists who study why output per capita consistently grew in most countries during the 20th century often ignore the fact that growth in any given country was uneven over time. Underlying this approach is the assumption that temporary fluctuations in economic growth are transitory and have no consequences for long-run growth. On the other hand, economists interested in cyclical fluctuations often abstract from long-run economic growth. In particular, various business cycle models have been devised in which output fluctuates around a constant level of output rather than a path that grows over time. This approach again reflects the view that long-run growth is driven by forces that are independent of the factors that drive booms and busts in economic activity. On this assumption, we can analyse why output deviates from its long-run trend without bothering to model the trend itself.

While this dichotomy has proven useful for exploring certain questions, economists have become

increasingly critical of this approach. Various attempts at integrating these two phenomena can be found in work on growth and business cycles from the late 1960s and early 1970s. Richard Goodwin's entry for growth and cycles in the first edition of this dictionary surveys some of this work (Goodwin, 1987, pp. 574–6).

Arguably, however, the article that contributed most to advancing the view that growth and business cycles should be analysed within a single model is Kydland and Prescott (1982). They argued that business cycles were driven not by short-run variations in aggregate demand, as most previous work had assumed, but by fluctuations in the same force that drives long-run growth, namely, technological progress. They started with the Ramsey (1928) growth model in which long-run growth is due to labour-augmenting technical change. But rather than assuming a constant rate of technical change, they allowed it to vary over time. This captures the notion that new ideas arrive sporadically, so productivity growth is inherently random. Households react to these shocks by varying their capital accumulation and labour supply.

Kydland and Prescott went on to argue that technology shocks could account for most of the volatility in US output during the post-war period. This claim remains controversial. However, even those who were sceptical of the claim that productivity shocks were responsible for business cycles were forced to acknowledge that temporary shocks could affect decisions relevant for long-run growth, such as capital accumulation, and conversely that the forces which shape long-run growth could have important short-run consequences. This implies that treating growth and cycles as distinct processes might overlook important connections between the two phenomena.

While Kydland and Prescott's paper was influential in promoting the view that growth and cycles should be modelled jointly, their model offered only limited insight into the connection between the two. This is because they modelled both long-run growth and fluctuations as exogenous: output per capita in their model grows because the economy is assumed to undergo technical change, and it grows in cycles because technical change is assumed to occur in cycles. As such, their model does not explain what drives technical change, why it should be inherently volatile, or whether growth and cycles might affect one another.

For example, Kydland and Prescott's model cannot tell us whether business cycles affect the rate of long-run growth. Are entrepreneurs more reluctant to undertake activities that lead to technical change in the face of macroeconomic volatility? Addressing this question requires us to model growth as an endogenous process rather than as the outcome of exogenous technical change. As another example, Kydland and Prescott asserted that technical change is inherently volatile. While this is undoubtedly true for any individual sector, it is not obvious why this volatility does not cancel out in the aggregate, resulting in a constant rate of technical change for the economy as a whole. Addressing this question requires us to model the underlying fluctuations in the rate of technical change as an endogenous outcome rather than as the result of an exogenous process. Fortunately, economists have since developed models in which either long-run growth or fluctuations, or both, are endogenous.

One line of research endogenizes growth while maintaining exogenous fluctuations. This approach allows us to study the effects of cyclical fluctuations on long-run growth. One of the first papers to tackle this question was Leland (1974), who built on previous work by Levhari and Srinivasan (1969). The latter studied the problem of a household deciding between consumption and saving given uncertain returns on its savings. Leland showed that this model could be reinterpreted as a representative

household economy with a linear technology for producing output from capital. Growth in this model was driven by capital accumulation, so shocks to productivity – the analogue of uncertain returns – affected growth by affecting average investment.

Leland showed that the effect of cycles on growth depended on household attitudes towards risk. If the coefficient of relative risk aversion among households exceeded one, they would engage in more precautionary savings in the face of macroeconomic volatility, accumulating capital more rapidly. When relative risk aversion is below unity, macroeconomic volatility would induce households to accumulate less capital, leading to a slower rate of growth. Thus, whether cycles encourage or discourage long-run growth is ambiguous from a theoretical standpoint.

Ramey and Ramey (1995) provided empirical evidence on the relationship between growth and cycles using cross-country evidence. They found that volatility is associated with slower growth. At the same time, they found that more volatile countries do not have lower investment rates, contradicting Leland's analysis on how volatility ought to affect growth. This contradiction was resolved by Ramey and Ramey (1991) and Barlevy (2004), who argued that volatility affects growth not by changing average investment but by making investment less volatile; more volatile investment lowers long-run growth because growth is concave in investment. Barlevy (2004) in particular argued that this channel implies that exogenous cyclical fluctuations would be associated with very large welfare costs.

A separate line of research proceeded in the opposite direction: it assumed long-run growth was exogenous, and examined whether fluctuations in the economy-wide rate of technical change could arise endogenously. For example, Shleifer (1986) developed a multi-sector model where in each period innovators in a fixed fraction of sectors develop more productive technologies. They could use these to earn profits for a limited period, after which rivals in their sector could copy the technology and drive profits to zero. If innovators implemented their ideas as soon as they came up with them, the rate of aggregate technical change would be constant. But Shleifer allowed firms to delay implementation, and showed that there exist equilibria where technical change occurs in spurts: innovators wait until there is enough innovation in other sectors before they implement their own ideas, so growth would be concentrated rather than spread out evenly over time.

Shleifer's result emerges because in his model implementing new technologies exhibits strategic complementarities: when one firm implements a new technology, its owners earn excess profits which they use to purchase goods in other sectors. Firms that come up with a new technology might therefore prefer to wait until others come up with new ideas. Even though the economy arrives at new ideas at a constant rate, growth proceeds at an uneven rate in equilibrium.

A third line of research has sought to endogenize both long-run growth and fluctuations. For example, Francois and Lloyd-Ellis (2003) consider a modification of the Shleifer model where innovators choose how much research to undertake, rather than assuming the rate at which new ideas arrive is fixed exogenously. This allows them to examine whether implementation cycles can affect long-run growth. Since implementation cycles emerge endogenously, the connection between cycles and growth may be different from the way growth responds to exogenous shocks as in Leland's analysis.

Francois and Lloyd-Ellis find that the equilibrium with cycles involves unambiguously higher average growth than the equilibrium in which innovators implement their new ideas immediately. This is because innovators earn higher profits when they coordinate implementation, providing them with more incentive to engage in research. However, welfare turns out to be lower in the presence of cycles, so

faster but more uneven growth is less desirable. Lastly, Francois and Lloyd-Ellis show that, if countries differ in research productivity, we would observe a negative correlation between growth and cycles across countries; countries that are less productive in research will grow more slowly and exhibit longer and larger deviations from average growth. This helps to reconcile their results with Ramey and Ramey's evidence, and points out an important caveat for interpreting the cross-country evidence on growth and cycles.

Other authors have used models where both growth and cycles arise endogenously to explore whether technical change occurs in spurts not because of implementation delays but because of fluctuations in innovation. That is, even if innovators implement their new ideas immediately, they might still choose to concentrate their research activity in particular periods. Examples include Bental and Peled (1996), Walde (2002), and Matsuyama (1999). All three describe models in which the economy alternates between capital accumulation and innovation. In the first two papers, successful innovation raises the marginal product of capital, inducing a shift towards capital accumulation until the return to capital is low enough for innovation to turn profitable again. Matsuyama develops a model in which the economy grows as the variety of goods produced increases. Profits depend on the ratio of capital to the number of goods, so successful innovation reduces the profitability of innovation rather than increase the returns to physical capital. But all three models imply that the amount of innovation, and thus the rate of technical change, fluctuates along the equilibrium path.

A central feature of these models is that the economy fluctuates between innovation and capital accumulation. However, empirical evidence suggests research and development activity is high when capital accumulation is high. Recent work by Comin and Gertler (2006) and Barlevy (2005) examines why research activity might vary positively with capital accumulation. However, both assume cycles are due to exogenous shocks rather than that they arise endogenously in equilibrium. It remains a question for future research whether innovation might fluctuate endogenously but still co-vary with capital accumulation.

See Also

- endogenous growth theory
- Kydland, Finn Erling
- Prescott, Edward Christian
- real business cycles
- technical change
- welfare costs of business cycles

Bibliography

Barlevy, G. 2004. The cost of business cycles under endogenous growth. *American Economic Review* 94, 964–90.

Barlevy, G. 2005. On the timing of innovation in stochastic Schumpeterian growth models. Working

paper. Chicago: Federal Reserve Bank of Chicago.

Bental, B. and Peled, D. 1996. The accumulation of wealth and the cyclical generation of new technologies: a search theoretic approach. *International Economic Review* 37, 687–718.

Comin, D. and Gertler, M. 2006. Medium term business cycles. *American Economic Review* 96(3), 523–51.

Francois, P. and Lloyd-Ellis, H. 2003. Animal spirits through creative destruction. *American Economic Review* 93, 530–50.

Goodwin, R. 1987. Growth and cycles. In *The New Palgrave: A Dictionary of Economics*, vol. 2. ed. J. Eatwell, M., Milgate and P. Newman. Basingstoke: Palgrave, 1998.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Leland, H. 1974. Optimal growth in a stochastic environment. *Review of Economic Studies* 41, 75–86.

Levhari, D. and Srinivasan, T. 1969. Optimal savings under uncertainty. *Review of Economic Studies* 36, 153–63.

Matsuyama, K. 1999. Growing through cycles. *Econometrica* 67, 1617–31.

Ramey, G. and Ramey, V. 1991. Technology commitment and the cost of economic fluctuations. Working Paper No. 3755. Cambridge, MA: NBER.

Ramey, G. and Ramey, V. 1995. Cross-country evidence on the link between volatility and growth. *American Economic Review* 85, 1138–51.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Shleifer, A. 1986. Implementation cycles. *Journal of Political Economy* 94, 1163–90.

Walde, K. 2002. The economic determinants of technology shocks in a real business cycle model. *Journal of Economic Dynamics and Control* 27, 1–28.

How to cite this article

Barlevy, Gadi. "growth and cycles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/

article?id=pde2008_G000127> doi:10.1057/9780230226203.0685

The New Palgrave Dictionary of Economics Online

growth and inequality (macro perspectives)

Vincenzo Quadrini

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides a review of studies that examine the relationship between economic growth and inequality. These studies are divided into two groups. The first emphasizes the channels through which inequality affects growth, while the second emphasizes the opposite channel, where economic growth affects inequality. Although several empirical studies find a significant correlation between inequality and growth, it is still an open question as to whether the correlation is driven by the first or the second channel.

Keywords

constitutions, economic approach to; creative destruction; distortionary taxes; economic growth; education; endogenous growth theory; entrepreneurship; equity–efficiency trade-off; expropriation; growth and inequality; human capital; income smoothing; increasing returns; inequality; innovation; Kuznets, S.; policy reform; redistribution of income and wealth; skill-biased technical change; wage inequality

Article

The study of ‘growth and inequality’ has a long tradition. One well-known relationship in economic development is the Kuznets curve. Kuznets observed that in the early stage of human development, when agriculture was the main economic activity, inequality in the distribution of income was relatively low. As the economy industrialized and the workforce moved towards industry and away from agriculture, the distribution of income tended to widen. At some critical point in the economy's development, this tendency reversed. Although more recent evidence does not support the Kuznets curve hypothesis (see, for example, the widening income inequality observed in the United States starting in the early 1980s), the original empirical finding of Kuznets (1955) stimulated a large body of research activity.

Other evidence of the relationship between inequality and growth comes from more recent cross-country observations. Data collected during the 1980s and 1990s show that there is a great deal of variation across countries in the degree of income inequality and economic growth. Are the different growth rates related to the degree of inequality of each individual country? Several studies find, indeed, that there is a cross-country negative correlation between inequality and growth, that is, countries with greater income inequality tend to experience slower growth; see Benabou (1996) and Perotti (1996) for a review of the empirical studies. But correlation does not imply causation, and there are good reasons to think that the causation can go in both directions. In other words, slow growth could generate greater inequality and equality could lead to faster growth.

Inequality affecting growth

One of the channels through which inequality affects growth is through the political and institutional system. A new series of studies in the 1980s, pioneered by Romer (1986) and Lucas (1988), developed a new class of models in which government policies could have a significant impact on the long-term growth of the economy (endogenous growth models). Given the importance of government policies for long-term growth, it becomes important to understand the forces and mechanisms underlying the choice of policies. This work stimulated a new series of studies in political economy. These studies start from the observation that, in a democratic society, the fundamental mechanism underlying the choice of policies is the electoral system. Therefore, in order to understand how policies are selected, we need to study the policy preferences of the population and how these preferences are translated into voting preferences.

Many factors affect the voting preferences of a society. However, for policies that have a clear redistributive content, the position of the voter in the distribution of income or wealth plays an important role. If a person is poor, his or her tax payments are smaller than the benefits he or she receives from government expenditures. Consequently, his attitude towards redistributive policies is more favourable than someone at the top of the income distribution (he or she has to pay more taxes than the received benefits). If the distribution of income is very unequal, then there will be many voters favouring larger governments. Of course this would not be a problem for efficiency if taxes were not distortionary. But, in a standard endogenous growth model, taxes have a negative impact on investment and growth. Therefore, the main conclusion of this literature is that inequality impairs the economic potential of a country because voters will demand more redistribution through distortionary taxes; Persson and Tabellini (1994), Alesina and Rodrik (1994), Krusell and Rìos-Rull (1996), Krusell, Quadrini and Rìos-Rull (1996) are some examples.

These studies also demonstrate the importance of the institutional system. Although greater inequality implies a greater demand for redistributive policies, the way political preferences are aggregated and the way policies are ultimately chosen depend on the particular institutional framework. For example, whether the representative democracy works through a parliamentary or a presidential system could lead to different sizes of government and, through distortionary taxes, to different levels of economic growth; see Persson and Tabellini (2005) for an analysis of the economic effects of constitutions.

The predictions of the politico-economic literature are consistent with several empirical studies as they find a negative relation between inequality and growth. However, a deeper empirical investigation of

this channel poses some doubts. More specifically, the politico-economic channel can be divided into two sub-channels: a positive relation between ‘inequality’ and ‘redistributive policies’ and a negative relation between ‘redistributive policies’ and ‘growth’. Perotti (1996) shows that the negative effect of redistributive policies on growth is not a robust feature of the data. On the contrary, redistributive policies may even be positively associated with economic growth. How is this possible?

Several theories envision a beneficial effect of redistributive taxes. The key ingredient is the presence of financial constraints. Let us take the Shumpeterian view that entrepreneurship is central to economic growth. However, due to financial constraints and the lack of insurance markets, entrepreneurial investment is suboptimal. Under these conditions, redistribution may provide extra resources to constrained entrepreneurs and could facilitate more investments in growth-enhancing activities. At the same time, a redistributive system provides an implicit system of income smoothing (a person pays high taxes when he or she earns high profits but receives payments in case of losses), and therefore, it provides insurance. If entrepreneurs are risk averse, this encourages more investment. The issue of whether redistributive taxes increase or decrease entrepreneurial investment is still an open area of research.

A similar story applies to investment in education or human capital. If education is important for economic growth, but because of financial constraints households choose sub-optimal levels of education, then government transfers may allow for greater investment and growth. A more direct effect could be generated by financing public education, as in Glomm and Ravikumar (1992). Examples of studies that emphasize the importance of inequality for growth in the presence of financial constraints are Galor and Zeira (1993), Banerjee and Newman (1993) and Aghion and Bolton (1997).

Another group of studies emphasizes social conflict and expropriation. Greater inequality means that a larger group of individuals is at the bottom of the distribution and faces poor economic conditions compared to the rest of the population. Faced with poor economic conditions, people have strong incentives to expropriate either by ‘stealing’ or through ‘revolutions’. The risk of expropriation has two negative effects. First, it acts as an investment tax that discourages investment. Second, more resources are devoted to protect property rights, which detracts from resources devoted to productive and growth enhancing activities. An example of this kind of theory is Benhabib and Rustichini (1996).

Another theory of inequality affecting growth is that developed in Murphy, Shleifer and Vishny (1989). This theory assumes that there are technologies with increasing returns. These technologies become profitable only if the domestic market is sufficiently large, that is, there is a large demand for the goods produced with the new technologies. If wealth is highly concentrated the domestic market remains small (since there are not enough consumers who can afford these goods). As a result, these growth-enhancing technologies will not be implemented. However, the theory finds weak support in the data (see Benabou, 1996).

Growth affecting inequality

If we take the view that growth requires innovative risky activities and these activities cannot be easily insured, we would expect that faster growth is associated with greater *ex post* inequality. At the same time, a faster rate of innovation implies greater destructions of monopoly positions (creative destruction). This would generate lower inequality because the monopoly positions, which are the source

of high-income revenues, last for a shorter period of time. Therefore, it is not obvious whether faster innovation and growth create greater inequality. However, within this environment, faster growth generates higher mobility due to a higher turnover in the holding of monopoly positions. Therefore, even if growth leads to greater inequality, it also creates a healthier social environment. Long-term growth requires technological innovation and there is no doubt that new technologies affect different groups in different ways. Therefore, growth and inequality are intrinsically related. Since 1980, wage inequality among different education groups has been widening in almost all industrialized countries. Katz and Murphy (1992) show that this increase is due to a raising demand of skilled labour. Krusell et al. (2000) propose an explanation for the increasing demand of skilled labour based on the introduction and development of new technologies that are more complementary to skilled labour (skill-biased technologies).

Suppose that there are two types of workers, skilled and unskilled. The stocks of skilled and unskilled workers change slowly over time. Now suppose that there is the introduction of skill-biased technologies, that is, technologies that require more skilled labour than unskilled labour. This will lead to an increase in the demand for skilled workers. Given the limited increase in the supply, the wages of skilled workers will increase. On the other hand, the demand for unskilled workers will decline, which leads to a fall in the wages of these workers.

This is a compelling explanation for the increasing wage premium started at the beginning of the 1980s. However, it raises the question of why the ratio of skilled versus unskilled workers has not increased that much during this period, certainly not as much as we would expect given the size of the wage premium change.

The technological innovations introduced in the 1970s seem to have affected the economy in other respects. Greenwood and Jovanovic (1999) and Hobijn and Jovanovic (2001) believe that new information technologies required a level of restructuring that incumbent firms could not face. As a result, their stock market value dropped. This is another form of redistribution in the sense that the owners of incumbent firms lose market value to the owners of the new firms.

Policy considerations

Whether we concentrate on the first channel of causation–in which growth affects inequality–or to the second–in which inequality affects growth–there are no obvious policy recommendations. If we think that inequality has a negative impact on growth because society demands more redistributive policies (as in the standard political economy literature), then the constitutional system of electoral representation becomes central. Changing the constitutional system could lead to different political outcomes. However, changing the constitutional system is not easy. We could also think of reallocating resources once and for all to change the initial distribution. Although this is possible in theory, it is difficult from a political point of view.

If we concentrate on the opposite channel, in which growth impacts on inequality, and we are concerned about having an excessively unequal distribution of incomes, then we may consider possible redistributive policies. However, these policies may also have undesired effects on efficiency. If the tax system keeps the after-tax skill premium low, the incentive to acquire skills will be lower. But, because of skilled-biased technologies, more skills are required. This could also discourage the introduction of these technologies, which would impact negatively on growth. The equity–efficiency trade-off becomes

central to the analysis. The positioning of a society in this trade-off will depend on society preferences about the degree of inequality that is socially acceptable. These preferences are based on individual beliefs that are likely to depend on individual experiences and they change very slowly over time. The relationship between personal experience and beliefs is formalized in Piketty (1995); see also Quadrini (1999).

See Also

- economic growth
- inequality (international evidence)
- policy reform, political economy of
- wage inequality, changes in

Bibliography

Aghion, P. and Bolton, P. 1997. A theory of trickle-down growth and development. *Review of Economic Studies* 64, 151–72.

Alesina, A. and Rodrik, D. 1994. Distributive politics and economic growth. *Quarterly Journal of Economics* 109, 465–90.

Banerjee, A.V. and Newman, A.F. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101, 274–98.

Benabou, R. 1996. Inequality and growth. In *NBER Macroeconomics Annual*, ed. B.S. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Benhabib, J. and Rustichini, A. 1996. Social conflict and growth. *Journal of Economic Growth* 1, 129–46.

Galor, O. and Zeira, J. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60, 35–52.

Glomm, G. and Ravikumar, B. 1992. Public versus private investment inhuman capital: endogenous growth and income inequality. *Journal of Political Economy* 100, 818–34.

Greenwood, J. and Jovanovic, B. 1999. The information technology revolution and the stock market. *American Economic Review* 89, 116–22.

Hobijn, B. and Jovanovic, B. 2001. The it revolution and the stock market: evidence. *American*

Economic Review 91, 1203–20.

Katz, L. and Murphy, K. 1992. Changes in relative wages, 1963–1987: supply and demand factors. *Quarterly Journal of Economics* 107, 35–78.

Krusell, P., Ohanian, L., Rìos-Rull, J.-V. and Violante, G. 2000. Capital–skill complementarity and inequality: a macro economic analysis. *Econometrica* 68, 1029–53.

Krusell, P., Quadrini, V. and Rìos-Rull, J.-V. 1996. Are consumption taxes really better than income taxes? *Journal of Monetary Economics* 37, 475–503.

Krusell, P. and Rìos-Rull, J.V. 1996. Vested interests in a positive theory of stagnation and growth. *Review of Economic Studies* 63, 301–31.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45, 1–28.

Lucas, R.E. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Income distribution, market size and industrialization. *Quarterly Journal of Economics* 104, 537–64.

Perotti, R. 1996. Growth, income distribution, and democracy: what the data say. *Journal of Economic Growth* 1, 149–87.

Persson, T. and Tabellini, G. 1994. Is inequality harmful for growth? *American Economic Review* 84, 600–21.

Persson, T. and Tabellini, G. 2005. *The Economic Effects of Constitutions*. Cambridge, MA: MIT Press.

Piketty, T. 1995. Social mobility and redistributive politics. *Quarterly Journal of Economics* 110, 551–84.

Quadrini, V. 1999. Growth, learning and redistributive policies. *Journal of Public Economics* 74, 263–97.

Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–36.

How to cite this article

Quadrini, Vincenzo. "growth and inequality (macro perspectives)." The New Palgrave Dictionary of

Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000222> doi:10.1057/9780230226203.0686

The New Palgrave Dictionary of Economics Online

growth and institutions

Daron Acemoglu
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Institutions are often viewed as a key determinant of economic growth. Much research inquires whether the institutions that influence economic outcomes are themselves determined by other factors. European colonization of the world provides a laboratory in which to investigate these issues since it exogenously imposed different institutions on otherwise identical societies. Colonies where Europeans settled had institutions that protected property rights, and have since prospered, while other colonies were given centralized repressive states that extracted resources from the population and have largely remained relatively poor. Choice of institutions reflects the distribution of political power in a society.

Keywords

colonization; commitment; growth and governance; growth and institutions; mortality; North, D.; political institutions, economic approaches to; property rights

Article

A central question of economics is to understand why some countries are much poorer than others. Economists have long recognized that this relates to the fact that some countries have much less human capital, physical capital and technology than others, and use their existing factors and technologies much less efficiently. Nevertheless, these differences are only *proximate* causes in the sense that they pose the next question of *why* some countries have less human capital, physical capital and technology and make worse use of their factors and opportunities. This has motivated economists and social scientists more broadly to look for potential *fundamental* causes, which may be underlying these proximate differences across countries.

Institutions have emerged as a potential fundamental cause, contrasting, for example, with geographical differences or cultural factors. While geographic characteristics of countries and regions may lead to differences in the technology available to individuals or make their investments in physical and human capital more difficult, institutional differences, associated with differences in the organization of society, shape economic and political incentives and affect the nature of equilibria via these channels. There is vibrant research, both empirical and theoretical, attempting to understand the importance of institutions for economic outcomes. Since it is impossible to do justice to this burgeoning field in such a short article, my purpose here is not to survey the literature but to present some of the main conceptual issues that are useful for future work.

What are institutions?

Douglass North (1990, p. 3) offers the following definition: ‘Institutions are the rules of the game in a society or, more formally, are the humanly devised constraints that shape human interaction.’ Three important features of institutions are apparent in this definition: (*a*) they are ‘humanly devised’, which contrasts with other potential fundamental causes like geographic factors, which are outside human control; (*b*) they are ‘the rules of the game’ setting ‘constraints’ on human behavior; and (*c*) their major effect will be through incentives (see also North, 1981).

The notion that incentives matter is second nature to economists, and institutions, if they are a key determinant of incentives, should have a major effect on economic outcomes, including economic development, growth, inequality and poverty. But do they? Are institutions key determinants of economic outcomes or secondary arrangements that respond to

other, perhaps geographic or cultural, determinants of human and economic interactions?

Much empirical research attempts to answer this question. Before we discuss some of this research, it is useful to emphasize an important point: ultimately, the aim of the research on institutions is to pinpoint *specific* institutional characteristics that are responsible for economic outcomes in specific situations (for example, the effect of legal institutions on the types of business contracts). However, the starting point is often the impact of a broader notion of institutions on a variety of economic outcomes. This broader notion, in line with Douglass North's conception, incorporates many aspects of the economic, political and social organization of society. Institutions can differ between societies because of their formal methods of collective decision-making (democracy versus dictatorship) or because of their economic institutions (security of property rights, entry barriers, the set of contracts available to businessmen). They may also differ because a given set of formal institutions is expected to, and does, function differently; for example, they may differ between two societies that are democratic because the distribution of political power lies with different groups or social classes, or because in one society democracy is expected to collapse while in the other it is consolidated. This broad definition of institutions is both an advantage and a curse. It is an advantage, since it enables us to get started with theoretical and empirical investigations of the role of institutions without getting bogged down by taxonomies. It is a curse, since, unless we can follow it up with a better understanding of the role of specific institutions, we have learned little.

The impact of institutions

There are tremendous cross-country differences in the way that economic and political life is organized. A voluminous literature documents large cross-country differences in economic institutions, and a strong correlation between these institutions and economic performance. Knack and Keefer (1995), for instance, look at measures of property rights enforcement compiled by international business organizations, Mauro (1995) looks at measures of corruption, Djankov et al. (2002) compile measures of entry barriers across countries, while many studies look at variation in educational institutions and the corresponding differences in human capital. All of these authors find substantial differences in these measures of economic institutions, and significant correlation between these measures and various indicators of economic performance. For example, Djankov et al. find that, while the total cost of opening a medium-size business in the United States is less than 0.02 per cent of GDP per capita in 1999, the same cost is 2.7 per cent of GDP per capita in Nigeria, 1.16 per cent in Kenya, 0.91 per cent in Ecuador and 4.95 per cent in the Dominican Republic. These entry barriers are highly correlated with various economic outcomes, including the rate of economic growth and the level of development.

Nevertheless, this type of correlation does not establish that the countries with worse institutions are poor *because* of their institutions. After all, the United States differs from Nigeria, Kenya and the Dominican Republic in its social, geographic, cultural and economic fundamentals, so these may be the source of their poor economic performance. In fact, these differences may be the source of institutional differences themselves. Consequently, evidence based on correlation does not establish whether institutions are important determinants of economic outcomes.

To make further progress, one needs to isolate a source of exogenous differences in institutions, so that we approximate a situation in which a number of otherwise identical societies end up with different sets of institutions. European colonization of the rest of the world provides a potential laboratory in which to investigate these issues. From the late 15th century, Europeans dominated and colonized much of the rest of the globe. Together with European dominance came the imposition of very different institutions and social power structures in different parts of the world.

Acemoglu, Johnson and Robinson (2001) document that in a large number of colonies, especially those in Africa, Central America, the Caribbean and South Asia, European powers set up ‘extractive states’. These institutions (again broadly construed) did not introduce much protection for private property, nor did they provide checks and balances against the government. The explicit aim of the Europeans in these colonies was extraction of resources, in one form or another. This colonization strategy and the associated institutions contrast with the institutions Europeans set up in other colonies, especially in colonies where they settled in large numbers: for example, the United States, Canada, Australia and New Zealand. In these colonies the emphasis was on the enforcement of property rights for a *broad cross section* of the society, especially smallholders, merchants and entrepreneurs. The term ‘broad cross section’ is emphasized here since, even in the societies with the worst institutions, the property rights of the elite are often secure, but the vast majority of the population enjoys no such rights and faces significant barriers to participation in many economic activities. Although investments by the elite can generate economic growth for limited periods, for sustained growth property rights for a broad cross section seem to be crucial (Acemoglu, Johnson and Robinson 2002; Acemoglu, 2003).

A crucial determinant of whether Europeans chose the path of extractive institutions was whether they settled in large numbers. In colonies where Europeans settled, the institutions were developed for their own future benefits. In colonies where Europeans did not settle, their objective was to set up a highly centralized state apparatus, and other associated institutions, to oppress the native population and facilitate the extraction of resources in the short run. Based on this idea, Acemoglu, Johnson and Robinson (2001) suggest that, in places where the disease environments made it easy for Europeans to settle, the path of institutional development should have been different from areas where Europeans faced high mortality rates.

In practice, during the time of colonization, Europeans faced widely different mortality rates in colonies because of differences in the prevalence of malaria and yellow fever. These mortality rates provide a possible candidate for a source of exogenous variation in institutions. The mortality rates should not directly influence output today but, by affecting the settlement patterns of Europeans, they may have had a first-order effect on institutional development. Consequently, these potential settler mortality rates can be used as an *instrument*

for broad institutional differences across countries in an instrumental-variables estimation strategy.

The key requirement for an instrument is that it should have no direct effect on the outcome that is the object of interest (other than its effect via the endogenous regressor). There are a number of channels through which potential settler mortality could influence current economic outcomes or may be correlated with other factors influencing these outcomes. Nevertheless, there are also good reasons why, as a first approximation, these mortality rates should not have a direct effect. Malaria and yellow fever were fatal to Europeans who had no immunity, and thus had a major effect on settlement patterns, but they had much more limited effects on natives who, over centuries, had developed various types of immunities. The exclusion restriction is also supported by the death rates of native populations, which appear to be similar between areas with very different mortality rates for Europeans (see, for example, Curtin, 1964).

The data also show that there were major differences in the institutional development of the high-mortality and low-mortality colonies. Moreover, consistent with the key idea in Acemoglu, Johnson and Robinson (2001), various measures of broad institutions – for example, measures of protection against expropriation – are highly correlated with the death rates Europeans faced more than a century ago and with early European settlement patterns. They also show that these institutional differences induced by mortality rates and European settlement patterns have a major (and robust) effect on income per capita. For example, the estimates imply that improving Nigeria's institutions to the level of those in Chile could, in the long run, lead to as much as a sevenfold increase in Nigeria's income. This evidence suggests that, once we focus on potentially exogenous sources of variation, the data point to a large effect of broad institutional differences on economic development.

Naturally, mortality rates faced by Europeans were not the only determinant of Europeans' colonization strategies. Acemoglu, Johnson and Robinson (2002) focus on another important aspect, namely, how densely different regions were settled before colonization. They document that in more densely settled areas Europeans were more likely to introduce extractive institutions because it was more profitable for them to exploit the indigenous population, either by having them work in plantations and mines or by maintaining the existing system and collecting taxes and tributes. This suggests another source of variation in institutions that may have persisted to the present, and Acemoglu, Johnson and Robinson (2002) show similar large effects from this source of variation.

Another example that illustrates the consequences of difference in institutions is the contrast between North Korea and South Korea. The geopolitical balance between the Soviet Union and the United States following the Second World War led to separation along the 38th parallel. The North, under the dictatorship of Kim II Sung, adopted a very centralized command economy with little role for private property. In the meantime, South Korea, though far from a free-market economy, relied on a capitalist organization of the economy, with private ownership of the means of production and legal protection for a range of producers, especially those under the umbrella of the chaebols, the large family conglomerates that dominated the South Korean economy. Although not democratic during its early phases, the South Korean state was generally supportive of rapid development and is often credited with facilitating, or even encouraging, investment and rapid growth in Korea.

Under these two highly contrasting regimes, the economies of North and South Korea diverged. While South Korea grew rapidly under capitalist institutions and policies, North Korea has experienced minimal growth since 1950 under communist institutions and policies.

Overall, a variety of evidence paints a picture in which broad institutional differences across countries have had a major influence on their economic development. This evidence suggests that to understand why some countries are poor we should understand why their institutions are dysfunctional. But this is only part of a first step in the journey towards an answer. The next question is even harder: if institutions have such a large effect on economic riches, why do some societies choose, end up with and maintain these dysfunctional institutions?

Modelling institutional differences

As a first step in modelling institutions, let us consider the relationship between three institutional characteristics: (*a*) economic institutions, (*b*) political power, and (*c*) political institutions.

As already mentioned, economic institutions matter for economic growth because they shape the incentives of key economic actors in society; in particular, they influence investments in physical and human capital and technology, and the organization of production. Economic institutions determine not only the aggregate economic growth potential of the economy but also the distribution of resources in the society, and herein lies part of the problem: different institutions will be associated not only with different degrees of efficiency and potential for economic growth, but also with different distributions of the gains across different individuals and social groups.

How are economic institutions determined? Although various factors play a role here, including history and chance, ultimately economic institutions are produced by collective choices of the society. And because of their influence on the distribution of economic gains, not all individuals and groups typically prefer the same set of economic institutions. This leads to a *conflict of interest* among various groups and individuals over the choice of economic institutions; and the *political power* of the different groups will be the deciding factor. The distribution of political power in society is also endogenous. To make more progress here, let us distinguish between two components of political power; de jure (formal) and de facto political power (see Acemoglu and Robinson, 2006). De jure political power refers to power that originates from the *political institutions* in society. Political institutions, similar to economic institutions, determine the constraints on and the incentives of the key actors, but this time in the political sphere. Examples of political institutions include the form of government – for example, democracy versus dictatorship or autocracy – and the extent of constraints on politicians and political elites.

A group of individuals, even if they are not allocated power by political institutions, may possess political power; for example, they can revolt, use arms, hire mercenaries, co-opt the military, or undertake protests in order to impose their wishes on society. This type of de facto political power originates from both the ability of the group in question to solve its collective action problem and from the economic resources available to the group (which determine their capacity to use force against other groups).

This discussion highlights the fact that we can think of political institutions and the distribution of economic resources in society as two *state variables*, affecting how political power will be distributed and how economic institutions will be chosen. An important notion is that of *persistence*; the distribution of resources and political institutions are relatively slow-changing and persistent. Since, like economic institutions, political institutions are collective choices, the distribution of political power in society is the key determinant of their evolution. This creates a central mechanism of persistence: political institutions allocate de jure political power, and those who hold political power influence the evolution of political institutions, and they will generally opt to maintain the political institutions that give them political power. A second mechanism of persistence comes from the distribution of resources: when a particular group is rich relative to others, this will increase its de facto political power and enable it to push for economic and political institutions favorable to its interests, reproducing the initial disparity. Despite these tendencies for persistence, the framework also emphasizes the potential for change. In particular, ‘shocks’ to the balance of de facto political power, including changes in technologies and the international environment, have the potential to generate major changes in political institutions, and consequently in economic institutions and economic growth.

Acemoglu, Johnson and Robinson (2005b) summarize this framework in Figure 1.

Figure 1

$$\left.\begin{array}{lcl} \text{political institutions}_t & \Rightarrow & \text{de jure political power}_t \\ & & \& \\ \text{distribution of resources}_t & \Rightarrow & \text{de facto political power}_t \end{array}\right\} \begin{array}{c} \Rightarrow \text{economic institutions}_t \\ \\ \Rightarrow \text{political institutions}_{t+1} \end{array} \Rightarrow \left\{\begin{array}{c} \text{economic performance}_t \\ \& \\ \text{distribution of resources}_{t+1} \end{array}\right.$$

Institutions in action

As a brief example, consider the development of property rights in Europe during the Middle Ages. Lack of property rights for landowners, merchants and proto-industrialists was detrimental to economic growth during this epoch. Since political institutions at the time placed political power in the hands of kings and various types of hereditary monarchies, such rights were largely decided by these monarchs. The monarchs often used their powers to expropriate producers, impose arbitrary taxation, renege on their debts, and allocate the productive resources of society to their allies in return for economic benefits or political support. Consequently, economic institutions during the Middle Ages provided little incentive to invest in land, physical or human capital, or technology, and failed to foster economic growth. These economic institutions also ensured that the monarchs controlled a large fraction of the economic resources in society, solidifying their political power and ensuring the continuation of the political regime.

The 17th century, however, witnessed major changes in the economic and political institutions that paved the way for the development of property rights and limits on monarchs' power, especially in England after the civil war of 1642–6 and the Glorious Revolution of 1688, and in the Netherlands after the Dutch revolt against the Hapsburgs. How did these major institutional changes take place? In England until the 16th century the king also possessed a substantial amount of de facto political power, and, if we leave aside civil wars related to royal succession, no other social group could amass sufficient de facto political power to challenge the king. But changes in the English land market (Tawney, 1941) and the expansion of Atlantic trade in the 16th and 17th centuries (Acemoglu, Johnson and Robinson, 2005a) gradually increased the economic fortunes, and consequently the de facto power, of landowners and merchants opposed to the absolutist tendencies of the Kings.

By the 17th century, the growing prosperity of the merchants and the gentry, based on both internal and overseas (especially Atlantic) trade, enabled them to field military forces capable of defeating the king. This de facto power overcame the Stuart monarchs in the English civil war and Glorious Revolution, and led to a change in political institutions that stripped the king of much of his previous power over policy. These changes in the distribution of political power led to major changes in economic institutions, strengthening the property rights of both landowners and capital owners and spurring a process of financial and commercial expansion. The consequence was rapid economic growth, culminating in the industrial revolution, and a very different distribution of economic resources from that in the Middle Ages.

This discussion poses, and also gives clues about the answers to, two crucial questions. First, why do the groups with conflicting interests not agree on the set of economic institutions that maximize aggregate growth? Second, why do groups with political power want to change political institutions in their favour? In the context of the example above, why did the gentry and merchants use their de facto political power to change political institutions rather than simply implement the policies they wanted? The issue of *commitment* is at the root of the answers to both questions.

An agreement on the efficient set of institutions is often not forthcoming because of the complementarity between economic and political institutions and because groups with political power cannot commit to not using their power to change the distribution of resources in their favour. For example, economic institutions that increased the security of property rights for landowners and capital owners during the Middle Ages would not have been credible as long as the monarch monopolized political power. He could promise to respect property rights, but then at some point renege on his promise, as exemplified by the numerous financial defaults by medieval kings. Credible secure property rights necessitated a reduction in the political power of the monarch. Although these more secure property rights would foster economic growth, they were not appealing to the monarchs, who would thereby lose their rents from predation and expropriation as well as various other privileges associated with their monopoly of political power. This is why the institutional changes in England as a result of the Glorious Revolution were not simply conceded by the Stuart kings. James II had to be deposed for the changes to take place.

The reason why political power is often used to change political institutions is related. In a dynamic world, individuals care about not only economic outcomes today but also those in the future. In the example above, the gentry and merchants were interested in their profits and therefore in the security of their property rights, not only in the present but also in the future. Therefore, they would have liked to use their (de facto) political power to secure benefits in the future as well as the present. However, commitment to future allocations (or economic institutions) is in general not possible because decisions in the future are made by those who hold political power at the time. If the gentry and merchants had been certain to maintain their de facto political power, this would not have been a problem. However, de facto political power is often *transient*, for example because the collective action problems that are solved to amass this power are likely to resurface in the future, or other groups, especially those controlling de jure power, can become stronger in the future. Therefore, any change in policies and economic institutions that relies purely on de facto political power is likely to be reversed in the future. In addition, many revolutions are followed by conflict among the revolutionaries. Recognizing this, the English gentry and merchants strove not just to change economic institutions in their favour following their victories against the Stuart monarchy, but also to alter political institutions and the future allocation of de jure power. Using political power to change political institutions then emerges as a useful strategy to make gains more durable. Consequently, political institutions and changes in political institutions are important as ways of manipulating future political power, and thus indirectly shaping future, as well as present, economic institutions and outcomes. Acemoglu and Robinson (2000; 2006) and Acemoglu, Johnson and Robinson (2005b) provide more detailed models and discuss further applications, including the creation and consolidation of electoral democracies in the West and in Latin America.

See Also

- class
- growth and civil war
- political institutions, economic approaches to

Bibliography

Acemoglu, D. 2003. The form of property rights: oligarchic versus democratic societies. Working Paper No. 10037. Cambridge, MA: NBER.

Acemoglu, D., Johnson, S. and Robinson, J. 2001. The colonial origins of comparative development: an empirical investigation. *American Economic Review* 91, 1369–401.

Acemoglu, D., Johnson, S. and Robinson, J. 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 118, 1231–94.

Acemoglu, D., Johnson, S. and Robinson, J. 2005a. The rise of Europe: Atlantic trade, institutional change and economic growth. *American Economic Review* 95, 546–79.

Acemoglu, D., Johnson, S. and Robinson, J. 2005b. Institutions as the fundamental cause of long-run growth. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Acemoglu, D. and Robinson, J. 2000. Why did the West extend the franchise? Democracy, inequality and growth in historical perspective. *Quarterly Journal of Economics* 115, 1167–99.

Acemoglu, D. and Robinson, J. 2006. *Economic Origins of Dictatorship and Democracy*. Cambridge: Cambridge University Press.

Curtin, P. 1964. *The Image of Africa*. Madison: University of Wisconsin Press.

Djankov, S., LaPorta, R., Lopez-de-Silanes, F. and Shleifer, A. 2002. The regulation of entry. *Quarterly Journal of Economics* 117, 1–37.

Knack, S. and Keefer, P. 1995. Institutions and economic performance: cross-country tests using alternative measures. *Economics and Politics* 7, 207–27.

Mauro, P. 1995. Corruption and growth. *Quarterly Journal of Economics* 110, 681–712.

North, D. 1981. *Structure and Change in Economic History*. New York: W.W. Norton & Co.

North, D. 1990. *Institutions, Institutional change, and Economic Performance*. New York: Cambridge University Press.

Tawney, R. 1941. The rise of the gentry, 1558–1640. *Economic History Review* 11, 1–38.

How to cite this article

Acemoglu, Daron. "growth and institutions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000174> doi:10.1057/9780230226203.0687

The New Palgrave Dictionary of Economics Online

growth and international trade

James Rauch

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International trade is believed to promote growth for countries at the technological frontier by expanding the market over which to exploit new ideas. For countries behind the technological frontier, three views of the relationship between growth and international trade are described and assessed empirically: trade hampers growth for natural resource-abundant countries that specialize in the export of technologically stagnant primary products; trade acts as the ‘handmaiden’ of growth by improving the quality of investment and slowing the tendency of its return to fall; and trade acts as the engine of growth by providing a conduit for technology transfer.

Keywords

comparative advantage; economies and diseconomies of scale; foreign direct investment (FDI); global commodity chains; growth and international trade; Heckscher-Ohlin-Samuelson model; international trade and technology; learning by doing; less developed countries; primary product exports; research and development; Ricardo-Viner model; Rybczynski effect; size of economies; technological progress; technology transfer

Article

When examining the interaction between international trade and economic growth, it is necessary to distinguish between countries that are at the technological frontier and those that are substantially behind it. The reason is that long-run growth of productivity and ultimately of per capita income is constrained by technological progress in the former group of countries but not in the latter group.

Growth and international trade for countries at the technological frontier

Rivera-Batiz and Romer (1991) argue that a larger economic size of the world promotes worldwide

technological progress, in two ways. First, current research and development (R&D) builds upon the existing stock of ideas or knowledge (standing on the shoulders of giants). Because ideas are non-rival, a larger world implies a larger stock of knowledge that facilitates R&D. International trade is implicated in this effect only to the extent that it promotes exchange or sharing of ideas among countries. Second, a larger economic size of the world raises the rents generated by monopoly holders of patented ideas (or ideas that are too costly to imitate) by providing a larger market for the goods based on the ideas. International trade is needed to exploit this large market and therefore provides greater incentives for innovation, increasing economic growth through this 'Schumpeterian' mechanism. Rivera-Batiz and Romer concentrate on these scale effects by modelling integration between similar, developed countries, thereby abstracting from the comparative advantage, resource reallocation effects of trade.

This theoretical prediction of a positive effect of international trade on the worldwide rate of technological progress needs to be qualified when international trade occurs between dissimilar countries. For example, trade can be expected to increase the relative price of skilled labour in the most skilled labour-abundant country. This could raise the cost of R&D relative to the cost of goods production enough to offset the positive effects of international trade on R&D for this country. If the most skilled labour-abundant country drives the worldwide rate of technological progress, international trade could reduce world economic growth. For a much fuller discussion of the interactions between international trade, R&D, and economic growth, see Grossman and Helpman (1991).

In so far as technological progress is transmitted equally to all countries at the worldwide technological frontier, there is no room for cross-country variation in the impact of international trade on growth from this source. In other words, the prediction that trade increases the worldwide rate of technological progress (absent strong, offsetting comparative advantage effects), and hence long-run economic growth of countries at the technological frontier, is inherently untestable using cross-sectional data because we have only one world. Since falling transportation and communication costs have tended to make countries more open to international trade over time, this prediction is consistent with time-series evidence that seems to show that the worldwide rate of growth is increasing over the very long run, but this increase is also predicted by the increasing size of the world – standing on the shoulders of more giants.

Growth and international trade for countries behind the technological frontier

The effect of international trade on the economic growth of less developed countries (LDCs) has long been one of the most passionately debated subjects in economics. Here I set out three views that span the range from negative to positive.

Trade as the enemy of growth

The case for trade as the enemy of growth can be made succinctly using the open economy version of the model of Matsuyama (1992). In his model all productivity increase takes place through learning by doing in the manufacturing sector, which is perceived as an external economy by any individual manufacturing firm. (In Matsuyama's model, unlike in the classic infant-industry argument, the potential for learning by doing is never exhausted – the manufacturing sector never catches up to international

best practice. One way this can be justified is by assuming that international best practice is a receding target.) Productivity in agriculture (or, more broadly, in the primary product sector) is constant by assumption. Now consider a country that is well endowed with natural resources relative to labour (and also relative to human and physical capital, if these are assumed to be used more intensively in manufacturing). Comparative advantage will lead this country to export primary products and import manufactures. Exports of primary products draw workers out of manufacturing, both directly and indirectly by generating rents that create demand for services, and thereby reduce productivity growth both in manufacturing and in the aggregate (since there is no productivity growth in primary products). (A variant of this argument is that trade causes ‘lagging’ economies to specialize in goods whose learning potential has been exhausted: Young, 1991.) In this way trade reduces growth in per capita income in countries with abundant natural resources. It is interesting that the sociological literature on ‘dependency’ and ‘world systems’ comes to the same conclusion that development of ‘peripheral’ countries is hindered by their exports of primary products to ‘core’ industrialized countries. For a summary of these arguments and a review of empirical studies see Crowly et al. (1998).

This case for trade as the enemy of growth does not depend on the assumption that productivity in the primary product sector is constant. Obviously learning-by-doing and other forms of productivity increase occur in this sector in the real world. What is crucial is only that this productivity increase tends to be substantially less rapid or less sustainable for a long period than in manufacturing. It is also possible that productivity growth in the primary product exportable sector worsens the terms of trade for the exporting country rather than raising its income, as modelled by Lewis (1969). In this instance the country in question must be large enough in its primary product speciality to influence its world price, or else we must treat the countries to which the argument applies as a bloc rather than individually.

Recently it has become fashionable to introduce inequality and ‘institutions’ into the argument linking trade to slow growth through a comparative advantage in primary products. In this view the problem with exports of primary products, especially minerals and tropical cash crops, is not that they are associated with low productivity growth but rather that they are associated with high inequality between owners of large mines or plantations and their employees. This inequality leads to the adoption of educational and political institutions that tend to exclude and disenfranchise the masses, making these economies less capable of realizing the potential offered by new technologies. This argument has been most thoroughly articulated by Engerman and Sokoloff (2002).

Several studies (for example, Sachs and Warner, 2001) find that the ratio of primary product exports to GDP is strongly negatively associated with per capita income growth. This negative association remains even after many other potential determinants of economic growth are controlled for, such as climate, geography, economic policies, political institutions, and external shocks. Indeed, the ratio of primary product exports to GDP is considered one of the most robust determinants of growth in cross-country growth regressions (Sala-i-Martin, 1997). This cross-country evidence cannot be considered decisive, however. Easterly (2001) points out that the primary product export share for less developed countries has tended to decline since 1960 as resources have been depleted and population has grown, yet per capita income growth rates have not tended to rise.

Trade as the ‘handmaiden’ of growth

The phrase ‘trade as handmaiden of growth’ is taken from the title of an article by Kravis (1970). He writes (1970, p. 869),

> The term ‘engine of growth’ is not generally descriptive and involves expectations which cannot be fulfilled by trade alone; the term ‘handmaiden of growth’ better conveys the notion of the role that trade can play. One of the most important parts of this handmaiden role for today's developing countries may be to serve as a check on the appropriateness of new industries by keeping the price and cost structures in touch with external prices and costs.

This supportive role can be usefully compared to that of financial development. As discussed in Levine (1997), a well-developed financial system increases the efficiency of investment by helping to channel savings to the most profitable projects. One way that trade can increase the efficiency of investment is by helping to ensure that the most privately profitable projects are also the most socially profitable ones. Foreign competition discourages investors from attempting to establish monopoly positions in small domestic markets and from producing substandard goods. Other ways in which trade can increase the efficiency of investment are enabling producers to realize economies of scale through exporting, and relieving bottlenecks that might reduce the returns to well-conceived downstream investments or divert resources from them. (Openness to international trade can also generate static and dynamic economies of scale – the latter through learning-by-doing spillovers, for example – by promoting specialization. Weinhold and Rauch, 1999, find that productivity growth in the manufacturing sector in less developed countries is higher when production is more specialized.)

Along the same lines, trade can slow or suspend the tendency of the return on investment to fall as physical (or human) capital accumulates. This link between growth and international trade was originally made by Ricardo, who argued that repeal of the Corn Laws in Britain would increase imports of grain, reduce the competition for labour between agricultural landlords and manufacturers, and thereby raise the return to investment in reproducible physical capital. His implicit model was what we now call the Ricardo–Viner model. Here we consider an extension of this model by Deardorff (1984), in which a small open economy produces an agricultural good using land and labour and, potentially, a number of manufactured goods using (reproducible) capital and labour under conditions of constant returns to scale and perfect competition. Let the manufactured goods be ranked unambiguously by capital intensity. Now consider a less developed country with an endowment of capital relative to labour and land that is so small that its manufacturing sector is completely specialized in production of the least capital-intensive manufactured good. As the country accumulates capital its return (marginal product) will fall. However, this reduction in the return to capital allows the country to become internationally competitive in the next least capital-intensive manufactured good, so that its manufacturing sector becomes incompletely specialized. At this point the return to capital (and the wage and, by extension, the rent on land) becomes fixed by international goods prices, as in the standard Heckscher–Ohlin–Samuelson model. Further accumulation of capital then causes both capital and labour resources within the manufacturing sector to be reallocated from the less to the more capital-intensive good (the Rybczynski effect). This forestalls any fall in the return to capital until the manufacturing sector becomes completely specialized in the more capital-intensive good, after which the return to capital falls

until production of the next most capital-intensive good is introduced, and so on.

If the view that international trade is the handmaiden of economic growth is correct, the large empirical literature investigating ‘causality’ between trade and growth (for example, Jung and Marshall, 1985) is somewhat beside the point. There should, however, be a strong cross-country correlation between openness to international trade and the rate of growth. Many studies have found such a correlation, but its robustness has been called into question (Rodriguez and Rodrik, 2001). Part of the problem may be that most studies try to include all countries for which there are reliable data, yet we have seen that there is no clearly predicted relationship between openness to trade and economic growth for countries at the technological frontier (the so-called industrialized or rich countries) and that trade may reduce growth for countries whose exports are concentrated in primary products. The positive correlation between openness to trade and growth should be most robust for the intermediate group of countries between least and most developed, a group sometimes labelled the ‘semi-industrialized’ countries. Given the complexity of the handmaiden view, however, it is probably best investigated by theoretically informed case studies like the classic NBER volumes (1974–8) supervised by Jagdish N. Bhagwati and Anne O. Krueger.

Trade as the engine of growth

The view of trade as the engine of growth takes technological progress rather than investment to be the ultimate source of growth, and sees imported ideas as the main determinant of technological progress in less developed countries. In other words, trade with more technologically advanced countries acts as a vehicle for the flow of knowledge from them and thereby drives growth in less advanced countries. Foreign direct investment (FDI) from more to less developed countries plays the same role. (This contemporary view of trade as the engine of growth must be distinguished from the older view, in which growth is driven by expansion of land devoted to production of technologically stagnant primary products to meet the demand of industrialized countries; see, for example, Caves, 1965.) The emphasis on imported ideas is associated with the work of Romer (for example, 1993). His work, however, leaves open the question of the specific mechanisms through which firms in less developed countries absorb knowledge from contact with technologically advanced countries.

Economists have typically modelled technology transfer as an arm's-length phenomenon. Firms are not *taught* the new technology. Rather, they engage in purposive imitative activity on their own (see, for example, Grossman and Helpman, 1991, ch. 11), employ machinery and equipment that embodies foreign knowledge (for example, Coe, Helpman and Hoffmaister, 1997), license the new technology, and so on. In reality, however, it is difficult to learn new technology from a distance. Keller (2004, p. 756) writes, ‘Only the broad outlines of technological knowledge are codified – the remainder remains “tacit”. non-codified knowledge is usually transferred through person-to-person demonstrations and instructions.’ There is a growing body of evidence that, for less developed country firms in particular, a major and perhaps predominant source of technology transfer (and transfer of managerial know-how) is instruction by developed country buyers: producers seeking cheaper suppliers of inputs and distributors seeking cheaper suppliers of final goods.

One example of such evidence is a study by Egan and Mody, who surveyed US buyers operating in LDCs, including ‘manufacturers, retailers, importers, buyers' agents, and joint venture partners’ (1992,

p. 322). They found:

> Buyers also render long-term benefits to suppliers in the form of information on production technology. This occurs principally through various forms of in-plant training. The buyer may send international experts to train local workers and supervisors … Buyers may also arrange short-term worker training in a developed country plant. (1992, p. 328)

Rhee, Ross-Larson and Pursell surveyed Korean exporters of manufactures. Their findings were similar to those of Egan and Mody:

> The relations between Korean firms and the foreign buyers went far beyond the negotiation and fulfillment of contracts. Almost half the firms said they had directly benefited from the technical information foreign buyers provided: through visits to their plants by engineers or other technical staff of the foreign buyers, through visits by their engineering staff to the foreign buyers… (1984, p. 61)

This process of learning foreign technology can be thought of as taking place within international production networks or ‘global commodity chains’ (Gereffi, 1994; 1999). This theoretical framework predicts that, once LDC firms are incorporated into the ‘bottoms’ of the chains, their learning will continue by movement up the chains. There are two types of chains: ‘producer-driven’ and ‘buyer-driven’ (Gereffi, 1994, p. 97). In the former, large manufacturers play the central roles in coordinating the production networks. Producer-driven chains are typical in capital- and technology-intensive industries such as automobiles, aircraft, computers, semiconductors, and heavy machinery. In the latter, large retailers, branded marketers, and branded manufacturers play the coordinating roles. Buyer-driven commodity chains are typical in labour-intensive, consumer goods industries such as garments, footwear, toys, housewares, and consumer electronics. Profitability is highest at the tops of the chains where barriers to entry are greatest: scale and technology in producer-driven chains, design and marketing expertise in buyer-driven chains.

In buyer-driven commodity chains, one mode through which learning is predicted to continue is *organizational succession*: from assembler to original equipment manufacturer (OEM) to original brand-name manufacturer (OBM), which is from more subordinate, competitive, and low-profit positions to more controlling, oligopolistic, high-profit positions. In the apparel industry, Gereffi (1999) finds that LDC firms that have parts provided to them for assembly learn how to find on their own the parts needed to make the product according to the design specified by the buyer (and may then subcontract the assembly); firms that have reached this level learn how to design and sell their own merchandise, becoming branded manufacturers (and may then subcontract the production, becoming branded marketers). Additional study is needed to determine whether this pattern of learning is common in other consumer goods industries. At the same time, work is needed to reconcile the kind of findings discussed here with econometric analyses (surveyed in Rodrik, 1999, ch. 2) which conclude that more productive firms export, but exporting does not make firms more productive.

In producer-driven commodity chains, one mode of learning is through ‘vertical linkages’ established between foreign subsidiaries of the large manufacturers that coordinate the production networks and host

country suppliers. Saggi (2002, p. 213) writes:

> Mexico's experience with FDI is illustrative of how such a process works. In Mexico, extensive backward linkages resulted from FDI in the automobile industry. Within five years of investments by major auto manufacturers there were 300 domestic producers of parts and accessories, of which 110 had annual sales of more than $1 million (Moran, 1998). Foreign producers also transferred industry best practices, zero defect procedures, and production audits to domestic suppliers, thereby improving their productivity and the quality of their products.

Javorcik (2004) finds econometric evidence that in Lithuania upstream suppliers to foreign subsidiaries experienced increases in productivity.

Conclusions

Here I highlight what I feel are the most important challenges facing the exponents of each of the three views of trade and growth described in the previous section. Those who see trade as the enemy of growth for natural resource-abundant countries need to do more than merely assert that the primary product export sector is incapable of rapid productivity growth. This assumption has recently been challenged by Dolan, Harris-Pascal and Humphrey (1999) and others who provide evidence that the fresh vegetable export sector in sub-Saharan Africa can realize the kind of learning benefits and investment opportunities associated with manufacturing by upgrading quality and presentation. Those who see trade as the handmaiden of growth must formulate their views more precisely if they are to be used to guide policy or subjected to rigorous empirical testing. Finally, those who see trade as the engine of growth must resolve the contradictions between case studies and econometric results regarding the benefits from exporting. Are the case studies unrepresentative, or does the statistical estimation suffer from measurement problems (as suggested by Katayama, Lu and Tybout, 2003)? Perhaps the case studies and the surveys that collect the data for econometric analysis need to be coordinated to make sure that the right questions are being asked.

See Also

- growth and learning-by-doing
- international trade theory
- Schumpeterian growth and growth policy design

Bibliography

Caves, R. 1965. ‘Vent for surplus’ models of trade and growth. In *Trade, Growth, and the Balance of Payments*, ed. R. Baldwin et al. Chicago: Rand McNally.

Coe, D., Helpman, E. and Hoffmaister, A. 1997. North-South R&D spillovers. *Economic Journal* 107 (440), 134–49.

Crowly, A., Rauch, J., Seagrave, S. and Smith, D. 1998. Quantitative cross-national studies of economic development: a comparison of the economics and sociology literatures. *Studies in Comparative International Development* 33 (Summer), 30–57.

Deardorff, A. 1984. An exposition and exploration of Krueger's trade model. *Canadian Journal of Economics* 17, 731–46.

Dolan, C., Harris-Pascal, C. and Humphrey, J. 1999. Horticulture commodity chains: the impact on the UK market of the African fresh vegetable industry. Working Paper No. 96. Brighton: Institute of Development Studies.

Easterly, W. 2001. The lost decades: developing countries' stagnation in spite of policy reform 1980–1998. *Journal of Economic Growth* 6, 135–57.

Egan, M.-L. and Mody, A. 1992. Buyer-seller links in export development. *World Development* 20, 321–34.

Engerman, S. and Sokoloff, K. 2002. Factor endowments, inequality, and paths of development among New World economies. *Economia* 3 (Fall), 41–88.

Gereffi, G. 1994. The organization of buyer-driven global commodity chains: how U.S. retailers shape overseas production networks. In *Commodity Chains and Global Capitalism*, ed. G. Gereffi and M. Korzeniewicz. Westport, CT: Praeger.

Gereffi, G. 1999. International trade and industrial upgrading in the apparel commodity chain. *Journal of International Economics* 48, 37–70.

Grossman, G. and Helpman, E. 1991. *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.

Javorcik, B. 2004. Does foreign direct investment increase the productivity of domestic firms? In search of spillovers through backward linkages. *American Economic Review* 94, 605–27.

Jung, W. and Marshall, P. 1985. Exports, growth, and causality in developing countries. *Journal of Development Economics* 18, 1–12.

Katayama, H., Lu, S. and Tybout, J. 2003. Why plant-level productivity studies are often misleading, and an alternative approach to inference. Working Paper No. 9617. Cambridge, MA: National Bureau of

Economic Research.

Keller, W. 2004. International technology diffusion. *Journal of Economic Literature* 42, 752–82.

Kravis, I. 1970. Trade as a handmaiden of growth: similarities between the nineteenth and twentieth centuries. *Economic Journal* 80, 850–72.

Levine, R. 1997. Financial development and economic growth: views and agenda. *Journal of Economic Literature* 35, 688–726.

Lewis, W. 1969. *Aspects of Tropical Trade, 1883–1965*. Stockholm: Almqvist and Wicksell.

Matsuyama, K. 1992. Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory* 58, 317–34.

Moran, T. 1998. *Foreign Direct Investment and Development*. Washington, DC: Institute for International Economics.

NBER (National Bureau of Economic Research). 1974–78. *Foreign Trade Regimes and Economic Development*. New York: Columbia University Press.

Rhee, Y., Ross-Larson, B. and Pursell, G. 1984. *Korea's Competitive Edge: Managing the Entry into World Markets*. Baltimore, MD: Johns Hopkins University Press.

Rivera-Batiz, L. and Romer, P. 1991. Economic integration and endogenous growth. *Quarterly Journal of Economics* 106, 531–55.

Rodriguez, F. and Rodrik, D. 2001. Trade policy and economic growth: a skeptic's guide to the cross-national evidence. In *NBER Macroeconomics Annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Rodrik, D. 1999. *Making Openness Work: The New Global Economy and the Developing Countries*. Washington, DC: Overseas Development Council.

Romer, P. 1993. Idea gaps and object gaps in economic development. *Journal of Monetary Economics* 32, 543–73.

Sachs, J. and Warner, A. 2001. The curse of natural resources. *European Economic Review* 45, 827–38.

Saggi, K. 2002. Trade, foreign direct investment, and international technology transfer: a survey. *World Bank Research Observer* 17, 191–235.

Sala-i-Martin, X. 1997. I just ran two million regressions. *American Economic Review* 87, 178–83.

Weinhold, D. and Rauch, J. 1999. Openness, specialization, and productivity growth in less developed countries. *Canadian Journal of Economics* 32, 1009–27.

Young, A. 1991. Learning by doing and the dynamic effects of international trade. *Quarterly Journal of Economics* 106, 369–405.

How to cite this article

Rauch, James. "growth and international trade." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000129> doi:10.1057/9780230226203.0688

The New Palgrave Dictionary of Economics Online

growth and learning-by-doing

Paul Beaudry
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Arrow, K.; economic growth theory; learning by doing; Lucas, R.; productive efficiency

Article

Learning by doing refers to improvements in productive efficiency arising from the generation of experience obtained by producing a good or service. The formal modelling of learning by doing was initiated in Arrow (1962) and was motivated by two main factors. The first motivating factor was empirical: several studies of wartime production found that input requirements decreased as a result of production experience. For example, Searle (1945) studied productivity changes in the Second World War shipbuilding programmes. During the Second World War, US production of ships increased dramatically, from 26 vessels in 1939 to 1,900 ships in 1943, an almost fiftyfold increase. Searle (1945) noticed that unit labour requirements decreased at a constant rate for a given percentage increase in output. On average, a doubling of output was associated with declines of 16 to 22 per cent in the number of man-hours required to build Liberty ships, Victory ships, tankers and standard cargo vessels. Alchian (1963) studied the relationship between the amount of direct labour required to produce an airframe and the number of airframes produced in the United States during the Second World War. He found that a doubling of production experience decreased labour input by approximately one-third. Other empirical studies of learning by doing include Rapping (1965), Irwin and Klenow (1994) and Thornton and Thompson (2001).

The second motivating factor behind the work of Arrow (1962) was a search for a theory of economic growth which did not rely on exogenous change in productivity as a driving force. In particular, Arrow's contribution and its extensions in Levhari (1966a; 1966b) were to show how economic growth could be sustained in a market with perfect competition. Arrow's original model is quite sophisticated, but the main insight can be derived in a simpler setting, as shown in Sheshinski (1967) and presented here. Consider a one good economy, where the production of the good requires capital and labour input according to the constant returns to scale production function:

$$Y = F(K, AL),\ F(\lambda K, A\lambda L) = \lambda F(K, AL).$$

In this specification of the production technology, *A* represents the efficiency of labour in producing the good. The main idea in the learning by doing literature is that *A* is a function of past experience. Arrow assumed that experience can be measured by cumulative investment or, in other words, the capital stock. The form of the relationship between *A* and the capital stock is posited to be:

$$A_t = (K_t)^{\alpha},\ 0 < \alpha < 1$$

where the assumption that $0 < \alpha < 1$ is motivated by the empirical studies. In order to close the system, assume that the labour force grows exponentially at the rate η and let capital accumulation be driven by a constant saving rate out of incomes, *s* where, in the absence of depreciation, this implies

$$\dot{K} = sY$$

In this environment, on the assumption that the change in *A* is an unintended consequence of production, it can be shown that a balanced growth path exists where per-capita income and per-capita capital grow at the rate

$$\alpha \frac{\eta}{1 - \alpha}$$

The two important aspects to note about the resulting growth rate is that it is positive if $\eta > 0$ and it is independent of the savings rate *s*. The additional property – that the rate of growth of income is tied to a positive rate of population growth – is generally seen as a weakness of this type of model. This property can be partially remedied, as shown in Romer (1986), if one assumes that $\alpha = 1$. In this case, even in the absence of labour force growth there exists a balanced growth path where the rate of growth is given by

$$sF(1, L)$$

The drawback of this specification ($\alpha = 1$) is that the growth rate now depends on the size of the labour force, which is referred to as a ‘scale effect’. The attractive feature of this specification is that the growth rate can be modified by an economic decision variable such as the savings rate. An alternative way of modifying Arrow's original model is to posit, as in Lucas (1988), that *A* depends on the per-capita value of the capital stock instead of on the level of the capital stock. This assumption is justified in Lucas (1988) on the grounds that *A* reflects the knowledge of the average worker with respect to how best to operate the technology. In the case where the relationship is given by $A = \frac{K}{L}$, the steady growth rate of per-capita output is given by $sF(1, 1) - n$. This formulation has the attractive property that it is positive even if $n = 0$, and it does not exhibit a scale effect. Accordingly it offers a succinct theory of economic growth. Lucas conjectured that the assumption of constant returns to learning (that is, $\alpha = 1$) could be justified in a model where there is bounded learning in any one good but where there is continual entry of new goods over time. This idea is formally studied in Stokey (1988) and Young (1993). There is also a large literature that discusses how learning by doing can interact with international trade and potentially give rise to income divergence across countries; see for example Lucas (1993) and Young (1991).

Bibliography

Alchian, A. 1963. Reliability of progress curves in airframe production. *Econometrica* 31, 679–93.

Arrow, K. 1962. The economic implications of learning by doing. *Review of Economic Studies* 29, 155–73.

Irwin, D. and Klenow, P. 1994. Learning by doing spillovers in the semiconductor industry. *Journal of Political Economy* 102, 1200–27.

Levhari, D. 1966a. Further implications of learning by doing. *Review of Economic Studies* 33, 31–8.

Levhari, D. 1966b. Extensions of Arrow's ‘Learning by Doing’. *Review of Economic Studies* 33, 117–31.

Lucas, R., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Lucas, R., Jr. 1993. Making a miracle. *Econometrica* 61, 251–72.

Rapping, L. 1965. Learning and World War II production functions. *Review of Economics and Statistics* 47, 81–6.

Romer, P. 1986. Increasing returns and long run growth. *Journal of Political Economy* 94, 1002–37.

Searle, A. 1945. Productivity of labour and industry. *Monthly Labor Review* 61, 1132–47.

Sheshinski, E. 1967. Optimal accumulation with learning by doing. In *Essays on the Theory of Economic Growth*, ed. K. Shell. Cambridge, MA: MIT Press.

Stokey, N. 1988. Learning by doing and the introduction of new goods. *Journal of Political Economy* 96, 701–17.

Thornton, R. and Thompson, P. 2001. Learning from experience and learning from others: an exploration of learning and spillovers in wartime shipbuilding. *American Economic Review* 91, 1350–68.

Young, A. 1991. Learning by doing and the dynamic effects of international trade. *Quarterly Journal of Economics* 106, 369–405.

Young, A. 1993. Invention and bounded learning by doing. *Journalof Political Economy* 101, 443–72.

How to cite this article

Beaudry, Paul. "growth and learning-by-doing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000181> doi:10.1057/9780230226203.0689

The New Palgrave Dictionary of Economics Online

growth models, multisector

W.A. Brock and W.D. Dechert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Multisector growth models have been increasingly used since the 1980s. The duality between growth models and dynamic general equilibrium models renders the multisector growth model ideal for the analysis of efficient intertemporal resource allocation. This includes renewable and non-renewable natural resources, produced resources such as capital, and land and labour resources. Growth models have been widely used in business cycle theory and in asset pricing theory. They have also been applied to the optimal management of dynamic ecological systems that have an economic component as a part of a complex systems model.

Keywords

asset pricing model; bequest motive; business cycles; central limit theorem; computation; concavity; convergence; decentralization; dynamic macroeconomic theory; equity premium puzzle; equivalence theorem; general equilibrium; indirect utility function; infinite horizons; law of large numbers; multisector growth models; New Keynesian macroeconomics; optimal growth models; optimal planning models; overtaking ordering; rational expectations equilibrium; real business cycles; recursive intertemporal general equilibrium models; representative agent; separating hyperplane theorem; single-sector growth models; turnpike theorems

Article

Multisector growth models are basic building blocks not only for optimal planning models (Majumdar, 1987; McKenzie, 1986) but also for recursive general equilibrium models (McKenzie, 2002; Stokey and Lucas, 1989), and for econometrically tractable models for business cycle research (Cooley, 1995) and general macroeconomics (Sargent, 1987). Majumdar (1987) has already covered some basic theory, some efficiency and decentralization analysis, as well as some optimization concepts. We attempt to fill

in the space between Majumdar (1987) and the current research frontier as well as to outline applications not treated by Majumdar.

Before we begin, we wish to stress that the style of this article is to point the reader towards surveys of the subject in order to economize on references to the many researchers who have contributed to this rather large area, and to paint, in broad strokes, the overall structure of this research area, especially its impact on empirical work, in order to illuminate directions where the research frontier might go.

Dynamic macroeconomic theory has made much use of the stochastic one-sector growth model (Cooley, 1995; Altug and Labadie, 1994; Sargent, 1987; Stokey and Lucas, 1989), for two primary reasons. First, it is a classical result that optimal growth models can be viewed as general equilibrium models by use of the separating hyperplane theorem in an appropriate space to construct the support prices. See Becker and Boyd (1997) for this general result, which they call the ‘equivalence theorem’. It is closely related to the use of decentralization prices in Majumdar (1987) and the general treatment of decentralization in Majumdar (1992).

The basic idea of the class of the ‘equivalence theorem’ of Becker and Boyd is as follows. Consider an infinite horizon intertemporal general equilibrium model with a representative infinitely lived consumer who faces intertemporal prices as given. Then it is a classical result that the rational expectations equilibrium of such a model is the same as the optimal solution of a planning problem where the planner has the same preferences as the representative consumer. Technical issues arise from the infinite horizon such as the necessity and sufficiency of transversality conditions at infinity (that is, the present discounted mathematical expectation of value of any stocks ‘left over’ at infinity should be zero, much as in a finite horizon case with no bequest motive). But the general ideas behind this type of result are much the same as in the well-known finite dimensional cases. See Becker and Boyd (1997) for the details.

Second, infinite horizon stochastic multisector models are also basic in constructing econometrically tractable models to use in analysing data. Here, especially, is where stochastic versions of the turnpike theorem (explained below) are used. For example, it is used to justify use of laws of large numbers and central limit theorems in econometric time-series applications.

A key property of the one-sector model that promotes its use in real business cycle applications as well as intertemporal general equilibrium asset pricing applications is the stochastic analog of the turnpike theorem. This theorem states that optimal capital stock and optimal consumption converge in a stochastic sense to a unique stochastic limit under standard assumptions of concavity of the payoff function (for example, the planner's preferences) and of the production function and modest assumptions on the structure of the stochastic shocks. It is much more difficult to obtain such results for general multisector stochastic models (Arkin and Evstigneev, 1987; Marimon, 1989) and even for deterministic versions of those models (McKenzie, 1986; 2002).

However, one can show that if the discount rate on the future is small enough there are results available in the literature that locate useful sufficient conditions on payoffs and technology such that stochastic convergence occurs (Marimon, 1989) and deterministic convergence occurs (McKenzie, 1986; 2002).

Results for stochastic multisector growth models in both discrete time settings and continuous time settings are also contained in the papers in Dechert (2001).

The basic idea behind these results, called ‘turnpike’ results, is to first observe that, if the discount rate on the future is zero, the dynamic optimization problem will attempt to maximize a long-run ‘static’

objective in order to avoid infinite ‘value loss’ if it failed to do so. Making this intuition mathematically precise requires introduction of a partial ordering called the ‘overtaking ordering’ and making assumptions on the objective function and the dynamics so that avoidance of infinite value loss results in convergence of the optimal quantities to a unique long-run limit (see Arkin and Evstigneev, 1987, and the papers in Dechert, 2001, for stochastic cases and McKenzie, 1986; 2002, for deterministic cases.) Once one has results well in hand for the case of zero discounting on the future, intuition suggests that there should be a notion of ‘continuity’ that would enable one to prove that, if the discount rate is close enough to zero, convergence would still hold. Unfortunately, turning such intuition into precise mathematics turns out to be rather difficult (see McKenzie, 1986; 2002, for deterministic literature and Arkin and Evstigneev, 1987, the papers in Dechert, 2001, and Marimon, 1989, for the stochastic case). We attempt to give the reader a brief idea of how the mathematical arguments work in a sketch of the arguments used to prove turnpike theorems for the deterministic case below. Let preferences of a planner be given by

$$\max_{\{x_t\}} \sum_{t=0}^{\infty} \beta^t \left[u(x_t, x_{t-1}) - u\left(x_\beta^*, x_\beta^*\right) \right] \tag{1}$$

where u: $R^{2n} \rightarrow R$ is a twice continuously differentiable function (typically an indirect utility or payoff function), β is a discount factor, $0 < \beta \leq 1$, and x_β^* is an optimal steady state which solves the first-order necessary conditions of the optimization in (1):

$$D_1 u(x_t, x_{t-1}) + \beta D_2 u(x_{t+1}, x_t) = 0 t \geq 1 \tag{2}$$

D_i denotes partial derivative with respect to the ith argument of u, and x_0 is given. We assume that u is jointly concave in its arguments and use eq. (2) evaluated at the optimal steady state x_β^* to rewrite the sum in (1). To simplify the notation, we let $u_\beta^* = u(x_\beta^*, x_\beta^*)$ and $D_i u_\beta^* = D_i u(x_\beta^*, x_\beta^*)$. Also, define

$$d_t = -\left[u(x_t, x_{t-1}) - u_\beta^* - \left(D_1 u_\beta^*\right)\left(x_t - x_\beta^*\right) - \left(D_2 u_\beta^*\right)\left(x_{t-1} - x_\beta^*\right) \right]$$

which is positive by the concavity of u. With this notation,

$$\sum_{t=1}^{T} \beta^t \left[u(x_t, x_{t-1}) - u_\beta^* \right] = \beta^T \left(D_1 u_\beta^* \right) \left(x_T - x_\beta^* \right) + \left(D_2 x_\beta^* \right) \left(x_0 - x_\beta^* \right) - \sum_{t=1}^{T} \beta^t d_t \tag{3}$$

Equation (3) immediately suggests that a good strategy to construct candidate optimal programs $\{x_t\}$ is to choose a program $\{x_t\}$ to solve

$$\min \sum_{t=1}^{\infty} \beta^t d_t. \tag{4}$$

This strategy works for all $\beta \in (0, 1]$. Following McKenzie (1986, and his references to David Gale) for $\beta = 1$, classify a program $\{x_t\}$ as good (bad) if the series $\sum_{t=1}^{T} d_t$ converges (diverges) and note that all programs $\{x_t\}$ are either good or bad. Solve (4) over good programs to get a top candidate for an optimum. By defining an appropriate partial ordering of programs that is a total ordering on the set of good programs, this top candidate turns out to be optimum. Since the series $\{d_t\}$ converges to 0 for all good programs, this forces $\{x_t\}$ to converge to a unique x^* which is the maximizer of $u(x, x)$ under the assumption that u is strictly concave. We call this analytical strategy the 'value loss' strategy.

There are basically two analytical strategies used for the case β is less than but close to 1. It is beyond the scope of this article to discuss them here; see McKenzie (1986; 2002) for the details.

All three of these analytical strategies can be generalized to stochastic cases where the indirect utility u contains stochastic shocks provided that Markovian type conditions are assumed on the stock process; $\{x_t\}$ is replaced by a sequence of random variables $\{X_t\}$; and x_β^* is replaced by a certain stationary ergodic stochastic process, X_β^*, that plays the role of the optimal stochastic steady state. This is not simple but we hope that our outline of one of the analytical strategies makes that one, at least, intuitively plausible (see, for example, Arkin and Evstigneev, 1987; Marimon, 1989; and the papers by Brock and Majumdar, 1978, Brock and Mirman, 1972, and Brock and Magill, 1979, reprinted in Dechert, 2001).

Our sketch of the above results has been deliberately brief since excellent survey treatments are readily available in the literature that we have cited. We wish to discuss here applications of multisector models to the following areas of economics: (*a*) a general vision of how the economy works; (*b*) asset pricing; (*c*) coupled ecological/economic dynamical systems.

General vision

It is no exaggeration to say that classical general equilibrium theory is analytically organized around existence of equilibrium, the core and equilibria, the two welfare theorems, as well as the ‘anything goes’ theorem of Sonnenschein, Mantel and Debreu (SMD) as the subject is expounded in McKenzie (2002). The SMD result requires users to place restrictions on the consumers and producers that populate general equilibrium models in order to use the theory for empirical work. In intertemporal economics a most popular way of doing this is to restrict oneself to recursive intertemporal general equilibrium models, and that restriction (via the ‘equivalence theorem’) places us in the domain of multisector growth models (Becker and Boyd, 1997).

Black (1995), stimulated by general equilibrium theory, sketches with broad strokes a vision of the economy that is basically operating close enough to a complete set of markets so that the device of generating equilibria by maximizing a weighted sum of utilities can be applied (McKenzie, 2002). Analytically, this device puts us in the domain of a large multisector model viewed as general equilibrium via a generalization of the ‘equivalence theorem’ in Becker and Boyd (1997). As McKenzie (2002) shows, turnpike theory could be extended to recursive intertemporal general equilibrium models with heterogeneous consumers provided markets are complete. Black (1995) proposes adding various elements to received intertemporal recursive general equilibrium models (that is, multisector growth models) not only to fill in gaps in the existing literature up to the mid-1990s but also to make the models match up better to data.

The book by Altug, Chadha, and Nolan (2003) might be viewed as an example of a realization of Black's vision. It shows the power of variations on uses of single-sector and multisector growth models as building blocks for closed- and open-economy macro models. We give some specific examples below. The examples are chosen because current cutting-edge work is being done in these areas and because the subject is moving fast in the directions of these chosen areas.

Asset pricing

Use of the ‘equivalence theorem’ rapidly lead to development of recursive econometrically tractable intertemporal general equilibrium asset pricing models based upon multisector stochastic optimal growth models (Becker and Boyd, 1997, and the papers in Dechert, 2001). The confrontation with data has not been all positive. Three main directions in which these models failed when confronted with data came to be known as the equity premium family of puzzles. But Weitzman (2004, p. 1) has shown that ‘…the subjective distribution of the future growth rate has its mean and variance calibrated to average past values. This paper shows that using the Bayesian posterior estimates of these parameters can go a very long way toward eliminating simultaneously all three puzzles.’ A major point of Weitzman is that, once the uncertainty inherent in the fact that there is estimation uncertainty in key parameters that the agents living in the model must take into account in addition to the shocks inherent in the model, then the puzzles tend to vanish.

Akdeniz and Dechert (2007) show that a single-sector stochastic asset pricing model with production and with heterogeneous firms can go a long ways toward removing the puzzles without having to

introduce Weitzman's Bayesian modification of the underlying basic model. Work like that of Akdeniz and Dechert is now possible due to advances in computational technology. Jog and Schaller (1994) have shown that a modification of the basic model for liquidity-constrained firms can account for patterns of mean reversion observed in returns data across size classes of firms.

Macroeconomics

We have already mentioned the real business cycle literature (Cooley, 1995; Altug, Chadha and Nolan, 2003) as macroeconomic applications of multisector growth models and their decentralization analysis.

A major recent development in macroeconomics is to replace the representative consumer agent and competitive firms in such models with a representative agent facing a set of differentiated products, each produced by a differentiated products monopolist who faces a stochastic process that gives it realizations of periods when it is allowed to change prices. This strategic modeling device allows one to add an analytically tractable theory of price setting which can be grafted onto the existing analytical apparatus of recursive multisector models to produce a model where a unification of the ‘real side’ and the ‘monetary side’ of macroeconomics can take place. Various devices are used to produce a demand for money balances in the model that include real balance services in the indirect utility function and cash in advance constraints. This modeling strategy has produced a new generation of very fruitful ‘New Keynesian’ macro models which has allowed treatment of key issues of monetary policy as well as better fit to data especially data resulting from interactions between the real side and the monetary side of an economy. See Altug, Chadha and Nolan (2003) and, especially, Woodford's treatise (2003) for this genre.

The real world has distortions such as taxes, inflation and other government activities such as production of public goods which require modifications of the basic structure of intertemporal recursive general equilibrium theory. Fortunately the analytical core can be quite readily modified to include these elements (Turnovsky, 1995).

Much of the literature on multisector optimal growth theory assumes convex technology and concave payoff (that is, concave utility) so that the indirect utility $u(x_t, x_{t-1}, S_t)$ is jointly concave in (x_t, x_{t-1}) for each value of the stochastic shock S_t. We believe much activity in the future will involve generalizations to models of coupled ecological and economic dynamic systems where such concavity does not hold. Some analytical work in this area has already appeared (Becker and Boyd, 1997; Majumdar, 1992) and as computational technology progresses we expect to see more developments that use a combination of analytics and computation.

See Also

- intertemporal equilibrium and efficiency
- rational expectations
- stochastic optimal control

Bibliography

Akdeniz, L. and Dechert, W. 2007. The equity premium in Brock's asset pricing model. *Journal of Economic Dynamics and Control* 31, 2263–92.

Altug, S., Chadha, J. and Nolan, C. 2003. *Dynamic Macroeconomic Analysis: Theory and Policy in General Equilibrium*. Cambridge: Cambridge University Press.

Altug, S. and Labadie, P. 1994. *Dynamic Choice and Asset Markets*. New York: Academic Press.

Arkin, V. and Evstigneev, I. 1987. *Stochastic Models of Control and Economic Dynamics*. New York: Academic Press.

Becker, R. and Boyd, R. 1997. *Capital Theory, Equilibrium Analysis and Recursive Utility*. Oxford: Blackwell.

Black, F. 1995. *Exploring General Equilibrium*. Cambridge, MA: MIT Press.

Brock, W.A. and Magill, M.J.P. 1979. Dynamics under uncertainty. *Econometrica* 47, 843–68.

Brock, W.A. and Majumdar, M. 1978. Global asymptotic stability results for multi-sector models of optimal growth under uncertainty when future utilities are discounted. *Journal of Economic Theory* 18, 225–43.

Brock, W.A. and Mirman, L. 1972. Optimal economic growth and uncertainty: the discounted case. *Journal of Economic Theory* 4, 479–513.

Cooley, T., ed. 1995. *Frontiers of Business Cycle Research*. Princeton: Princeton University Press.

Dechert, W., ed. 2001. *Growth Theory, Nonlinear Dynamics, and Economic Modelling: Scientific Essays of William Allen Brock*. Cheltenham: Edward Elgar.

Jog, V. and Schaller, H. 1994. Finance constraints and asset pricing: evidence on mean reversion. *Journal of Empirical Finance* 1, 193–209.

Majumdar, M. 1987. Multisector growth models. In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Majumdar, M., ed. 1992. *Decentralization in Infinite Horizon Economies*. Boulder, CO: Westview Press.

Marimon, R. 1989. Stochastic turnpike property and stationary equilibrium. *Journal of Economic Theory* 47, 282–306.

McKenzie, L. 1986. Optimal economic growth, turnpike theorems, and comparative dynamics. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M. Intriligator. Amsterdam: North-Holland.

McKenzie, L. 2002. *Classical General Equilibrium Theory*. Cambridge, MA: MIT Press.

Sargent, T. 1987. *Dynamic Macroeconomic Theory*. Cambridge, MA: Harvard University Press.

Stokey, N. and Lucas, R. 1989. *Recursive Methods in Economic Dynamics*. Cambridge, MA: MIT Press.

Turnovsky, S. 1995. *Methods of Macroeconomic Dynamics*. Cambridge, MA: MIT Press.

Weitzman, M. 2004. The Bayesian equity premium. Working paper, Department of Economics, Harvard University.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Brock, W. A. and W.D. Dechert. "growth models, multisector." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000276> doi:10.1057/9780230226203.0690

The New Palgrave Dictionary of Economics Online

growth take-offs

Matthias Doepke
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Following a phase of near-constant living standards lasting from Stone Age until the onset of the Industrial Revolution, a large number of countries have experienced growth takeoffs, in which stagnation gives way to sustained economic growth. What causes some countries to enter a growth takeoff while others remain poor? We discuss three mechanisms that can trigger a growth takeoff in a country previously trapped in poverty: fertility decline, structural change, and accelerating technological progress.

Keywords

child labour; demographic transition; economic growth; fertility; fixed factors; growth takeoffs; human capital; income–population feedback; Industrial Revolution; land; Malthus's theory of population; mortality; nutrition and development; population growth; productivity growth; skill-intensive technology; stagnation; structural change; technological progress; women's work

Article

Viewed on a historical timescale, economic growth in the world economy is characterized by a long phase of stagnation in living standards, followed in many, but not all, countries by a growth take-off, that is, a transition to steady and sustained economic growth.
Figure 1 illustrates the basic facts. Before 1800, GDP per capita was low and near-constant in all world regions, with little cross-country variation in income levels. The first country to experience a growth take-off was Britain with the start of the Industrial Revolution, closely followed by other west European countries and the ‘Western Offshoots’ such as the United States. More recently, a number of Asian and Latin American countries have undergone a transition to rapid economic growth as well. In much of Africa, however, income per capita continues to stagnate. What causes some countries to enter a growth takeoff while others remain poor?
Figure 1
The evolution of income per capita across world regions, years 1500–2001. *Note*: The ‘Western Offshoots’ are defined as the United States, Canada, Australia, and New Zealand. ‘Asia’ excludes Japan. *Source*: Maddison (2003, Table 8c).

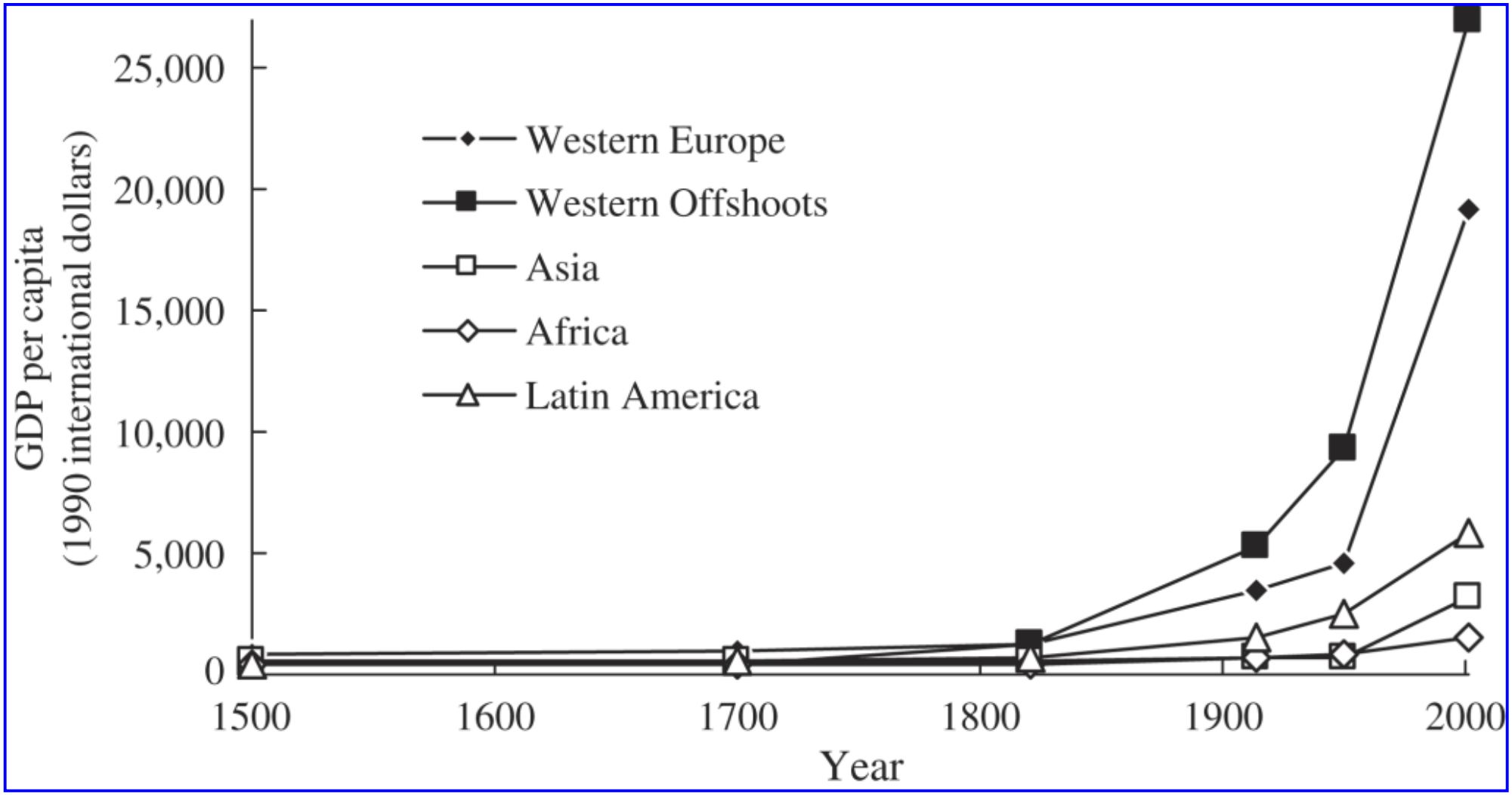

Explaining stagnation

Before one can account for a growth takeoff after a phase of stagnation, it is essential to understand why economies stagnated in the first place. The explanation suggested by one of the earliest writers on the subject, British economist Thomas Malthus in his *Essay on the Principle of Population* of 1798, is widely accepted to the present day. The Malthusian model relies on two key ingredients: an agricultural production function that uses the fixed factor of land, and an income–population feedback where the population growth rate is an increasing function of income per capita.

Consider an aggregate production function of the form

$$Y_t = A_t N_t^{\alpha} Z^{1-\alpha}, \tag{1}$$

where Y_t denotes output in period t, A_t is productivity, N_t is the size of the population, and Z is the fixed amount of land. (The results outlined below can be generalized to the case where physical capital also enters production.) In what follows, we use lower-case letters to denote per capita variables ($y_t = Y_t / N_t$, and so on), and the growth rate of a variable x is written as $\gamma(x)$. Output per capita is given by $y_t = A_t z_t^{1-\alpha}$, so that its growth rate $\gamma(y_t)$ satisfies:

$$\gamma(y_t) = \gamma(A_t) + (1-\alpha)\gamma(z_t).$$

Since land Z is constant, we have $\gamma(z_t) = \gamma(Z) - \gamma(N_t) = -\gamma(N_t)$. Using this relationship, the growth equation can be rewritten as

$$\gamma(y_t) = \gamma(A_t) - (1-\alpha)\gamma(N_t). \quad (2)$$

Growth in income per capita is thus an increasing function of productivity growth and a decreasing function of population growth. The negative effect of population growth reflects the fact that land is a fixed factor: when the size of the population increases, there is less land for each person to work with, which lowers income per capita.

To turn the growth equation (2) into a theory of stagnation, one needs to specify how productivity A_t and population N_t evolve over time. Assume for now that productivity growth is constant, $\gamma(A_t) = \bar{\gamma}_A$. The main assumption underlying the Malthusian theory of stagnation is that population growth is an increasing function of income per capita y_t:

$$\gamma(N_t) = f(y_t), \quad (3)$$

where $f'(y_t) > 0$. A number of different justifications can be given for this relationship. One possibility is that children enter the utility function of parents as normal goods. A rise in income would then increase the demand for children, leading to higher population growth. Alternatively, the mechanism could also work through mortality. If higher income leads to better nutrition and, as a consequence, lower mortality rates, a positive relationship between income per capita and population growth follows. As an empirical matter, the assumption of a positive relationship appears to fit the experience of most pre-industrial economies rather well.

Using (3), the growth equation (2) reads:

$$\gamma(y_t) = \bar{\gamma}_A - (1-\alpha) f(y_t). \quad (4)$$

According to this equation, the *growth rate* of income per capita is a decreasing function of its *level*. If the detrimental effect of population growth is sufficiently strong, this mechanism leads to stagnation as the only possible long-run outcome. In a country where income per capita is initially rising, population growth will accelerate until it fully offsets productivity growth, $(1-\alpha) f(y_t) = \bar{\gamma}_A$, resulting in stagnation.

The Malthusian model is remarkably successful in terms of explaining economic growth (or the lack thereof) until Industrial Revolution. However, we now know that ultimately many countries managed to escape from the Malthusian trap. In these countries, living standards today are far superior to what almost any human alive before 1800 could have

experienced. How can this drastic change in the economic fate of countries be explained?

Endogenous population growth

Given the growth equation (4), one scenario that could lead to a growth takeoff is a reversal of the income–population feedback. If the positive relationship described by the equation $\gamma(N_t) = f(y_t)$ breaks down, and subsequent population growth is low, growth will ensue. Consider, for example, the case where population growth ceases altogether, $\gamma(N_t) = 0$. According to eq. (2), growth in output per capita is then equal to productivity growth. Thus, as long as productivity keeps increasing, income per capita will grow indefinitely.
Historically, the Malthusian relationship between income and population growth did indeed break down in every single country that experienced a growth take-off. In a pattern known as the *demographic transition*, the high fertility and mortality rates of the pre-industrial era gave way to a new regime in which fertility, mortality, and population growth are low. In modern data, the relationship between income per capita and population growth is negative (both in a cross section of countries and in the time series for most rich countries), which is the opposite of what the Malthusian model assumes.
Figure 2 illustrates the demographic transition by comparing population growth in western Europe (the first region to experience a take-off) with Asia and Africa (the regions that stagnated the longest). In western Europe, population growth reached a peak at the end of the 19th century and has been declining since, despite rapid growth in income per capita. In Asia and Africa, in contrast, population growth has accelerated since the mid-19th century, and is now much higher than in western Europe.
Figure 2
The evolution of population growth across world regions, years 1500–2001. *Note*: ‘Asia’ excludes Japan. *Source*: Maddison (2003, Table 8a).

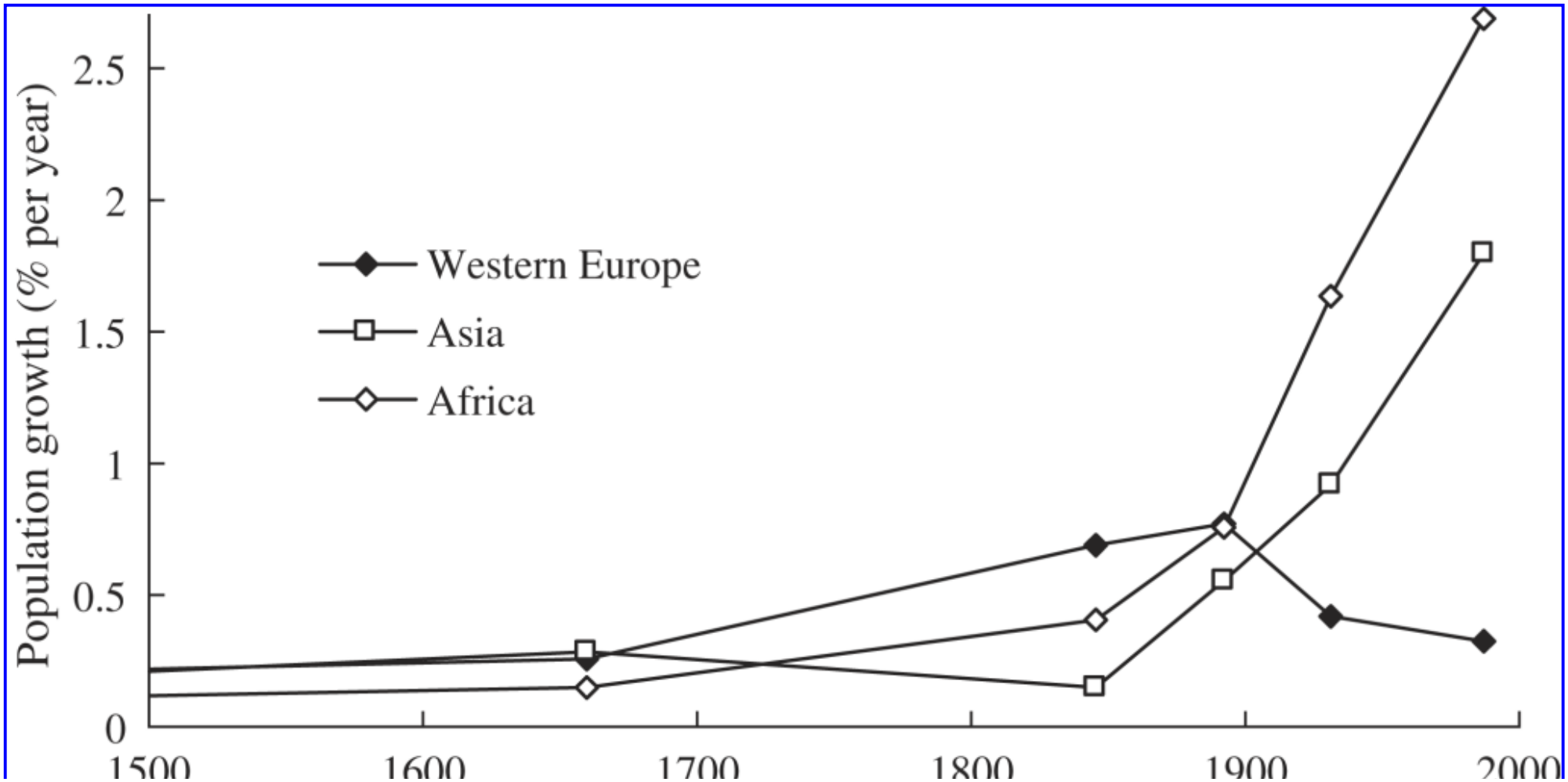

A number of authors have developed theories that integrate models of economic growth and the demographic transition to explain growth take-offs. In this literature, fertility decline is usually interpreted as a substitution of child ‘quantity’ (a large number of children) by child ‘quality’ (fewer children in which parents invest in terms of education or human capital). As an example of a model capturing this trade-off, consider the decision problem of a parent with preferences

$$u(c, n, h) = (1 - \beta)\log(c) + \beta[\log(n) + \gamma\log(h)]$$

over consumption *c*, the number of children *n*, and the children's human capital *h*, where $\beta > 0$ and $0 < \gamma < 1$. The parent has to spend fraction ϕ of its time to raise each child, and can choose to spend an additional per-child fraction *e* on educating the children. The total child-rearing time is then given by $(\varphi + e)n$, and the budget constraint for the parent is $c = (1 - (\varphi + e)n)wH$, where *H* is the parent's human capital, *w* is the wage per unit of human capital, and the time endowment is normalized to one. A child's human capital depends on the parent's human capital *H* and education time *e*:

$$h = 1 + \mu He,$$

where μ is the productivity of the education technology. Notice that a child receives at least one unit of human capital even if education *e* equals zero, which represents basic productive skills (such as physical strength) that do not rely on education. Lastly, the parent also has to observe a subsistence consumption constraint, $c \geq \bar{c}$, where $\bar{c}$ is the minimum amount of consumption required for survival.

In this model, the relationship between income and fertility depends on whether the optimal choices for education and consumption are at a corner. Assume that, initially, the wage *w* and the education productivity μ are so low that the subsistence constraint is binding and the parent chooses zero education ($e = 0$). The number of children is then constrained by the need to earn at least $\bar{c}$ units of consumption:

$$n = \frac{1}{\varphi}\left(1 - \frac{\bar{c}}{wH}\right).$$

Under this regime, the relationship between income *wH* and fertility *n* is positive, as assumed by the Malthusian model.

The outcome changes substantially if, through an increase in the wage *w* and the education productivity μ, the economy enters a regime where the subsistence constraint is no longer binding, and education is positive: $e > 0$. Under this regime, parents spend a fixed fraction of their time on child rearing. The balance between child quality and quantity depends on parental human capital *H*. The optimal decision rules are:

$$n = \frac{\beta}{\varphi + e} \text{ and}$$

(5)

$$e = \frac{1}{1-\gamma}\left(\gamma\varphi - \frac{1}{\mu H}\right). \quad (6)$$

Equation (5) captures the trade-off between child quality and quantity: the number of children is a decreasing function of education *e*. Intuitively, investing a lot in each child renders children expensive, which reduces demand. Education *e*, in turn, depends positively on parental human capital *H*. An increase in income per capita (through a rise in *H*) therefore lowers fertility *n*, the opposite of the Malthusian assumption.

Given these results, an escape from the Malthusian trap is possible if some change in the economy generates increased investment in child quality. The literature proposes different candidates for the underlying cause of such an event. In Galor and Weil (2000), the take-off is ultimately a consequence of technological progress. Accelerating productivity growth increases the return to education (the parameter μ in the model outlined above), which eventually triggers the quantity–quality substitution and the growth take-off. Galor and Moav (2002), in contrast, suggest that evolving parental preferences (through an increase in the parameter γ) are the driving force behind fertility decline. Yet other authors have emphasized the role of declining mortality rates (Boucekkine, de la Croix and Licandro, 2002; Cervellati and Sunde, 2005; Doepke, 2005; Kalemli-Ozcan, 2002; Lagerlöf, 2003a; Soares, 2005), increasing female labour-force participation (Galor and Weil, 1996; Lagerlöf, 2003b), changes in the provision of old-age security (Boldrin and Jones, 2002), changes in child-labour and education laws (Doepke and Zilibotti, 2005), and the introduction of skill-intensive production technologies that raise the return to education (Doepke, 2004).

Structural change

Apart from endogenous population growth, the Malthusian model also relies on the presence of the fixed factor of land to generate stagnation. A second potential trigger for a growth take-off is therefore *structural change* that decreases the role of land. In pre-industrial economies, agriculture was the main mode of production. In contrast, in modern industrial economies the share of agriculture in output is small, and consequently land is less important. Translated into the growth equation (4), structural change amounts to a shift in the parameter α. In particular, an increase in α lowers the detrimental effect of population growth on income per capita. In the limit case of $\alpha = 1$, income per capita is independent of the size or growth rate of the population, and is solely driven by productivity growth.

In Hansen and Prescott (2002), a decline of the role of land is generated endogenously in an environment where two competing technologies can be used for production. (Related contributions include Matsuyama, 1992; Laitner, 2000; Kögel and Prskawetz, 2001; Gollin, Parente and Rogerson, 2002; Ngai, 2004). In addition to the production function (1) above, an ‘industrial’, constant-returns technology is also available:

$$Y_t^I = A_t^I N_t^I,$$

where Y_t^I is industrial output, A_t^I is productivity, and N_t^I is the amount of labour employed in the industrial sector. Productivity A_t^I is assumed to grow at a constant rate. The total amount of labour is allocated optimally between the traditional sector and the industrial sector. Given the linear production technology, output per worker in the modern sector is given by A_t^I. Early in development, when A_t^I is still low, it is optimal to allocate all workers to the traditional sector. During this phase the economy behaves just like a Malthusian economy where the modern technology does not exist at all.

Ultimately, however, the modern technology becomes sufficiently productive to be introduced. If w_M is the (constant) marginal product of a worker in the Malthusian regime, the technology will be introduced once $A_t^I > w_M$. From this point on, population growth no longer affects output per worker, since land is not used in the industrial sector. Output per worker therefore starts to grow at the rate of technological progress. Viewed through the lens of the Hansen–Prescott model, what initially appears as a structural break in economic history is merely the outcome of an optimal sectoral allocation decision in an otherwise stable economic environment.

Endogenous technological progress

Starting once again from the growth equation (4), a third potential trigger for a growth takeoff is a sustained increase in productivity growth that is large enough to 'outrun' population growth. Clearly, population growth cannot increase indefinitely, as there are physiological constraints on child bearing. Let $\bar{\gamma}_N$ be an upper bound for population growth that cannot be exceeded for biological reasons. If now productivity growth satisfies

$$\gamma_A > (1-\alpha)\bar{\gamma}_N,$$

even at maximum population growth the detrimental effect of increasing population density does not suffice to negate productivity improvements, and improving living standards ensue.

A potential cause for accelerating productivity growth is scale effects in the production of ideas. An increase in world population implies that there are more people who might invent new, productive technologies. An increase in world population should therefore imply an acceleration of productivity growth. Scale effects of this kind underlie the takeoff models of Kremer (1993), Jones (2001), and Tamura (2002).

Conclusions

The three potential triggers for a growth take-off presented here should not be regarded as mutually exclusive alternatives, but rather as complementary explanations for a joint phenomenon. From an empirical perspective, there is little doubt that all three explanations are relevant: every country that underwent a growth takeoff also experienced a demographic transition, a sectoral shift from agriculture to industry and services, and an acceleration of productivity growth. Reflecting these observations, many papers in the literature already incorporate more than one of the mechanisms. For example, a number of authors propose models where accelerating endogenous productivity growth triggers a fertility transition. This is true, for example, of the seminal paper of Galor and Weil (2000) and, in a framework driven by human-capital externalities, for de la Croix and Doepke (2003). Similarly, Greenwood and Seshadri (2002) and Doepke (2004) integrate models of structural change with theories of fertility decline.

Building on the different mechanisms behind growth take-offs that have been proposed in recent years, a major challenge for future research is to understand why in many countries these mechanisms fail to work to the present day. Conceivably, a better understanding of the mechanisms that allowed some countries to overcome economic stagnation two centuries ago might help us learn how the same feat could be accomplished in poverty-stricken developing countries today.

See Also

- demographic transition
- economic growth in the very long run
- Industrial Revolution
- Malthus, Thomas Robert
- population and agricultural growth
- poverty traps

Bibliography

Boldrin, M. and Jones, L.E. 2002. Mortality, fertility, and saving in a Malthusian economy. *Review of Economic Dynamics* 5, 775–814.

Boucekkine, R., de la Croix, D. and Licandro, O. 2002. Vintage human capital, demographic trends, and growth. *Journal of Economic Theory* 104, 340–75.

Cervellati, M. and Sunde, U. 2005. Human capital formation, life expectancy and the process of development. *American Economic Review* 95, 1653–72.

de la Croix, D. and Doepke, M. 2003. Inequality and growth: why differential fertility matters. *American Economic Review* 93, 1091–113.

Doepke, M. 2004. Accounting for fertility decline during the transition to growth. *Journal of Economic Growth* 9, 347–83.

Doepke, M. 2005. Child mortality and fertility decline: does the Barro–Becker model fit the facts? *Journal of Population Economics* 18, 337–66.

Doepke, M. and Zilibotti, F. 2005. The macroeconomics of child labor regulation. *American Economic Review* 95, 1492–524.

Galor, O. and Moav, O. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117, 1133–91.

Galor, O. and Weil, D.N. 1996. The gender gap, fertility, and growth. *American Economic Review* 86, 374–87.

Galor, O. and Weil, D.N. 2000. Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. *American Economic Review* 90, 806–28.

Gollin, D., Parente, S. and Rogerson, R. 2002. The role of agriculture in development. *American Economic Review* 92, 160–4.

Greenwood, J. and Seshadri, A. 2002. The U.S. demographic transition. *American Economic Review* 92, 153–9.

Hansen, G.D. and Prescott, E.C. 2002. Malthus to Solow. *American Economic Review* 92, 1205–17.

Jones, C.I. 2001. Was an Industrial Revolution inevitable? Economic growth over the very long run. *Advances in Macroeconomics* 1(2), Article 1. Online. Available at http://www.bepress.com/bejm/advances/vol1/iss2/art1, accessed 5 October 2006.

Kalemli-Ozcan, S. 2002. Does the mortality decline promote economic growth? *Journal of Economic Growth* 7, 411–39.

Kögel, T. and Prskawetz, A. 2001. Agricultural productivity growth and the escape from the Malthusian trap. *Journal of Economic Growth* 6, 337–57.

Kremer, M. 1993. Population growth and technological change: One Million b.c. to 1900. *Quarterly Journal of Economics* 108, 681–716.

Lagerlöf, N. 2003a. From Malthus to modern growth: can epidemics explain the three regimes? *International Economic Review* 44, 755–77.

Lagerlöf, N. 2003b. Gender equality and long-run growth. *Journal of Economic Growth* 8, 403–26.

Laitner, J. 2000. Structural change and economic growth. *Review of Economic Studies* 67, 545–61.

Maddison, A. 2003. *The World Economy: Historical Statistics*. Paris: OECD.

Malthus, T.R. 1798. *Essay on the Principle of Population*. Harmondsworth: Penguin, 1970.

Matsuyama, K. 1992. Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory* 58, 317–34.

Ngai, R. 2004. Barriers and the transition to modern growth. *Journal of Monetary Economics* 51, 1353–83.

Soares, R. 2005. Mortality reductions, educational attainment, and fertility choice. *American Economic Review* 95, 580–601.

Tamura, R. 2002. Human capital and the switch from agriculture to industry. *Journal of Economic Dynamics and Control* 27, 207–42.

How to cite this article

Doepke, Matthias. "growth take-offs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000278> doi:10.1057/9780230226203.0691

The New Palgrave Dictionary of Economics Online

Haavelmo, Trygve (1911 – 1999)

Agnar Sandmo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balanced budget multiplier; capital theory; econometrics; Haavelmo, T.; inequality (global); investment theory; neoclassical theory of capital accumulation; probability

Article

Haavelmo was born in Skedsmo, Norway. He graduated from the University of Oslo in 1933 and joined Ragnar Frisch's newly created Institute of Economics as a research assistant. He spent the war years working for the Norwegian government in the United States. After a year's stay at the Cowles Commission at the University of Chicago, he returned to Norway in 1947, becoming professor of economics at the University of Oslo in 1948. He retired from his chair in 1979. In 1989 he was awarded the Nobel Memorial Prize in Economics, the Nobel citation referring to ‘his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures’.
Haavelmo first made his name by a series of path-breaking contributions to the theory of econometrics, most of which were written during his years in the United States. His 1943 article in *Econometrica* was the first to consider the statistical implications of simultaneity in economic models. This paper was one of the main sources of inspiration for the extensive work carried out in this area over the next decade, particularly at the Cowles Commission. Haavelmo developed his ideas further in the famous 1944 supplement to *Econometrica*; the main general contribution of this work was to base econometrics more firmly on the foundations of probability theory.
After his return to Norway, Haavelmo turned away from econometrics to economic theory as his main field of interest. In his 1957 presidential address to the Econometric Society (published the next year) he emphasized the need for a more solid theoretical foundation for empirical work as well as the need for theory to be inspired by empirical research.
Haavelmo's *Study in the Theory of Economic Evolution* (1954), is a broad exploration of the contributions that analytical economics can make to the understanding of global economic inequality. As an early contribution to growth theory it is less notable for simple models and precise theorems than for its imaginative and experimental attitude towards hypotheses concerning population growth, education,

migration and the international struggle for redistribution. The open-mindedness of the approach is very characteristic of the author.

Similar remarks apply to his 1960 book, *A Study in the Theory of Investment*. Its main objective is to provide a firmer microeconomic foundation for the macroeconomic theory of investment demand. To this end Haavelmo probes deeply into capital theory, emphasizing strongly, however, that a theory of optimum capital use does not in itself provide a theory of investment. This insight, and his clear statement of what has since been known as the neoclassical theory of capital accumulation, has been a major influence on late work in this area, both theoretical and applied.

Of Haavelmo's other contributions to economic theory, special mention should be made of his 1945 analysis of the balanced budget multiplier. The expansionary effect in a Keynesian unemployment situation of a balanced increase of public expenditure and taxes had been pointed out before, but Haavelmo was the first to provide a rigorous theoretical analysis of it.

Haavelmo was also been very active as a teacher. His lecture notes on a wide range of topics in economic theory exerted a formative influence on generations of Norwegian economists.

Selected works

1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11(January), 1–12.

1944. The probability approach in econometrics. Supplement to *Econometrica* 12(July), S1–115.

1945. Multiplier effects of a balanced budget. *Econometrica* 13(October), 311–18.

1947a. Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Society* 42(237), 105–22.

1947b. (With M.A. Girshick.) Statistical analysis of the demand for food: examples of simultaneous estimation of structural equations. *Econometrica* 15(April), 79–110.

1954. *A Study in the Theory of Economic Evolution*. Amsterdam: North-Holland.

1958. The role of the econometrician in the advancement of economic theory. *Econometrica* 26(July), 351–7.

1960. *A Study in the Theory of Investment*. Chicago: University of Chicago Press.

1970. Some observations on welfare and economic growth. In *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F.G. Scott and J.N. Wolfe. Oxford: Clarendon Press.

How to cite this article

Sandmo, Agnar. "Haavelmo, Trygve (1911–1999)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000001> doi:10.1057/9780230226203.0692

The New Palgrave Dictionary of Economics Online

Habakkuk, John Hrothgar (1915 – 2002)

Ravi Mirchandani
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Britain, economics in; capital-intensive technology; falling rate of profit; Habakkuk, J. H.; labour supply; technical change; United States, economics in

Article

Born in Wales in 1915, Habakkuk graduated from Cambridge in 1936, where he was a Fellow of Pembroke College from 1938 until 1950. He held the Oxford chair of economic history from 1950 to 1967, when he became Principal of Jesus College, Oxford. He retired in 1984. As a member of the Advisory Council on Public Records (1958–70), the Royal Commission on Historic Manuscripts, and the British Library Organizing Committee, amongst other bodies, he was active in the field of public records; he was knighted in 1974.

His major contribution was to the study of the rates of technological change in Britain and America in the 19th century and the reasons for the much more rapid development and use of manufacturing technology in the United States. In his book, *American and British Technology in the Nineteenth Century* (1962), American industrial development is roughly divided into two important stages, the period before the first wave of immigration in the 1840s, which laid the ground for future development, and the period after 1870 when abundant natural resources and rapid growth of market demand provided the stimulus for growth.

Habakkuk argues that American technological development in the early period, by contrast with Britain, was stimulated by the high cost of labour relative to capital and the relative inelasticity of labour supply. The expanding manufacturer, to avoid a falling marginal rate of profit, was more likely than his British counterpart to look to capital-intensive and labour-saving technology. Though Habakkuk was also keen to stress the importance of social factors, the suspicion of British employers and the hostility of British labour to new techniques, his explanation of the disparity is grounded in economic relationships.

Habakkuk's thesis has come under considerable scrutiny; recent research has tended to suggest that there was considerable diversity, both on a regional basis and between different industries, in development on both sides of the Atlantic. Economic historians have also questioned the timing of significant

development in the States and chosen to put greater stress on non-economic explanations. Habakkuk also made notable contributions to the debates on British population growth in the late 18th century and on the changing pattern of landholding as smaller holdings gave way to larger units in the same period.

Selected works

1953. English population in the eighteenth century. *English Historical Review* 6, 117–33.

1962. *American and British Technology in the Nineteenth Century: The Search for Labour-Saving Inventions*. Cambridge: Cambridge University Press.

1968. *Industrial Organisation since the Industrial Revolution*. Southampton: University of Southampton Press.

1971. *Population Growth and Economic Development since 1750*. Leicester: Leicester University Press.

1979–81. The rise and fall of English landed families, 1600–1800. *Transactions of the Royal Historical Society* I: 29, 187–207; II: 30, 199–221; III: 31, 195–217.

How to cite this article

Mirchandani, Ravi. "Habakkuk, John Hrothgar (1915–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000002> doi:10.1057/9780230226203.0693

The New Palgrave Dictionary of Economics Online

Haberler, Gottfried (1900 – 1995)

John S. Chipman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Barone, E.; business cycles; comparative costs; cost-of-living index; factor mobility; fixed exchange rates; free trade; Haberler, G.; homothetic preferences; indirect utility functions; inflation; international trade theory; Laspeyres index; Paasche index; Pigou effect; production-possibility frontier; purchasing power parity; real-balance effect; specific factor; stagflation; transfer problem

Article

Gottfried Haberler was born on 20 July 1900 in Purkersdorf, near Vienna. He studied economics at the University of Vienna under Friedrich von Wieser and Ludwig von Mises, where he received doctorates in law (1923) and economics (1925). After two years in the United States and Britain he returned to Vienna, received his habilitation in 1928, and was appointed lecturer, later professor, of economics, at the University of Vienna, from 1928 to 1936. He was appointed professor at Harvard University in 1936, where he remained until his retirement in 1971, after which he was a resident scholar at the American Enterprise Institute, Washington, DC. He was President of the International Economic Association (1950–1), the National Bureau of Economic Research (1955), and the American Economic Association (1963). In 1980 he was awarded the Antonio Feltrinelli prize.

Haberler's first major work was his habilitation thesis (1927), *The Meaning of Index Numbers*, summarized in Koo (1985, pp. 546–9). This work stimulated a great deal of subsequent research on the theory of the price or cost-of-living index. Haberler defined the ‘true change in the price level’ as ‘the ratio of the money income in the first period to the money income in the second period that would leave the individual indifferent’ (Koo, 1985, p. 547). Haberler's main concern was to find conditions under which this ‘true price index’ would be bounded by the Laspeyres and Paasche price indices. Some of the difficulties with this approach (and with the similar, earlier approach of Konüs, 1924) were discussed by Bortkiewicz (1928), Neisser (1929), Staehle (1935), and Frisch (1936). Frisch remarked (p. 25) that Haberler's definition of the ‘true change of the price level’ involved an implicit assumption of expenditure proportionality (homothetic preferences), and attributed this point (but apparently without

justification) to Bortkiewicz; he also interpreted Haberler (1929) in his reply to Neisser and Bortkiewicz as accepting this point. In terms of contemporary concepts we may say that homothetic preferences characterize indirect utility functions of the form $Y/C(p)$ where Y is income and $C(p)$ is a homogeneous-of-degree-1 function of prices.

Undoubtedly Haberler's most significant contribution was his reformulation of the theory of comparative costs (Haberler, 1930a), which revolutionized the theory of international trade. Prior to this paper, the Ricardian theory still held sway, but had been so amended with ill-defined concepts such as ‘real cost’ and ‘units of productive power’ taking the place of labour allocation that it had lost all its simplicity and elegance. Haberler introduced the production ‘substitution curve’ (now usually known as the production-possibility frontier), allowing for several factors of production, and taken to be concave to the origin as a result of diminishing returns. This laid the foundations for Ohlin's theory, as well as Lerner's and Samuelson's. True, as recently brought to light by Maneschi and Thweatt (1987), a footnote contained in the posthumous edition of Barone's *Principi* (1936, pp. 170–3), depicting a (non-concave) production-possibility frontier and a community indifference curve, was actually present in the first (1908) edition – but not subsequent ones; hence Barone must be accorded priority. But Haberler's independent discovery – and the use to which he put it – is what transformed the theory of international trade. Haberler also introduced the concept of a ‘specific factor’ – one that is completely immobile among industries – and used this concept with great effect in Haberler (1950) to illustrate the proposition that the gains from trade do not depend on the assumption of factor mobility.

Haberler made numerous other contributions to international economics, including (1) his synthesis and clarification of the Keynes–Ohlin debate on the transfer problem (Haberler, 1930b); (2) his judicious use of purchasing-power-parity calculations to set exchange rates (Haberler, 1945); (3) his introduction of the concept of supply and demand schedules for foreign exchange (1936) and his subsequent use of them (Haberler, 1949) in qualified support of the proposition that a devaluation in a pegged-rate regime could improve a country's balance of payments – but subject to the important proviso (1949, p. 213) that it would, through monetary expansion, likely shift these schedules; (4) his advocacy of free trade as the best policy for developing countries (Haberler, 1959); (5) numerous contributions to past and current history of international economic relations (cf. Koo, 1985).

The third area in which Haberler made major contributions is business-cycle theory (Haberler, 1937; 1942). His classic synthesis, notably in the third edition of *Prosperity and Depression* (1941), introduced the important ‘real-balance effect’, initially called the ‘Pigou effect’ by Patinkin (1948), although Patinkin in his 1951 revision acknowledged Haberler's priority over Pigou (1943). In the 1970s and 1980s Haberler furnished trenchant analyses of the phenomenon of worldwide inflation and the political economy of stagflation (cf. Koo, 1985), displaying the unique combination of clarity and wisdom that are characteristic of his writings.

Information on Haberler's life and work may be found in Schuster (1979), Chipman (1982), Baldwin (1982), Officer (1982), and Willett (1982). A complete bibliography of his writings is contained in Koo (1985).

Selected works

1927. *Der Sinn der Indexzahlen*. Tübingen: J.C.B. Mohr (Paul Siebeck).

1929. Der volkswirtschaftliche Geldwert und die Preisindexziffern. *Weltwirtschaftliches Archiv* 30 (July), 6**–14**.

1930a. Die Theorie der komparativen Kosten und ihre Auswertung für die Begründung des Freihandles. *Weltwirtschaftliches Archiv* 32(July), 350–70. Trans. as ‘The Theory of Comparative Costs and its Use in the Defense of Free Trade’ in Koo (1985).

1930b. Transfer und Preisbewegung. *Zeitschrift für Nationalökonomie* 1, 547–54; 2, 100–2. Trans. as ‘Transfer and Price Movements’ in Koo (1985).

1933. *Der internationale Handel. Theorie der weltwirtschaftlichen Zusammenhänge sowie Darstellung und Analyse der Aussenhandelspolitik*. Berlin: Julius Springer. Translated (revised by the author) as The Theory of International Trade with its Applications to Commercial Policy, London: William Hodge & Co., 1936.

1937. *Prosperity and Depression*. Geneva: League of Nations. 3rd edition enlarged by Part III, 1941. 5th and 6th editions, Cambridge, MA: Harvard University Press, 1958, 1964.

1942. *Consumer Instalment Credit and Economic Fluctuations*. New York: NBER.

1945. The choice of exchange rates after the war. *American Economic Review* 35, 308–18.

1949. The market for foreign exchange and the stability of the balance of payments. *Kyklos* 3(3), 193–218. Reprinted in Koo (1985).

1950. Some problems in the pure theory of international trade. *Economic Journal* 60, 223–40. Reprinted in Koo (1985).

1955. *A Survey of International Trade Theory*. Special Papers in Economics No. 1. Princeton: International Finance Section, Princeton University. Revised and enlarged edn, 1961. Reprinted in Koo (1985).

1959. *International Trade and Economic Development*. Cairo: National Bank of Egypt. Reprinted in Koo (1985).

Bibliography

Baldwin, R.E. 1982. Gottfried Haberler's contributions to international trade theory and policy. *Quarterly Journal of Economics* 97, 141–59.

Barone, E. 1908. *Principi di economia politica*. Rome: Tipografia Nazionale di Giovanni Bertero e C. German translation from the 3rd (1913) edn, *Grundzüge der theoretische Nationalökonomie*, Bonn: Kurt Schroeder, 1927. Posthumous edition, *Le opere economiche*, vol. 2, Bologna: Nicola Zanichelli Editore, 1936.

Bortkiewicz, L. von. 1928. Review of *Der Sinn der Indexzahlen* by Gottfried Haberler. *Magazin der Wirtschaft* 4(11), 427–9.

Chipman, J.S. 1982. Salute to Gottfried Haberler on the occasion of his 80th birthday. *Journal of International Economics* [Supplement], January, 25–30.

Frisch, R. 1936. Annual survey of general economic theory: the problem of index numbers. *Econometrica* 4, 1–38.

Konüs, A.A. 1924. The problem of the true index of the cost of living. *Economic Bulletin of the Institute of Economic Conjuncture* 9–10, October, 64–71. Translated from the Russian in *Econometrica* 7(1939), 10–29.

Koo, A.Y.C., ed. 1985. *Selected Essays of Gottfried Haberler*. Cambridge, MA: MIT Press.

Maneschi, A. and Thweatt, W.O. 1987. Barone's 1908 representation of an economy's trade equilibrium and the gains from trade. *Journal of International Economics* 22, 375–82.

Neisser, H. 1929. Der volkswirtschaftliche Geldwert und die Preisindexziffern. *Weltwirtschaftliches Archiv* 29(Part I), 6**–18**. Schlusswort, Part II, July, 14**–17**.

Officer, L.H. 1982. Prosperity and depression – and beyond. *Quarterly Journal of Economics* 97, 149–59.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38, 543–64. Revised version in *Readings in Monetary Theory*, ed. F.A. Lutz and L.W. Mints, Philadelphia: Blakiston, 1951.

Pigou, A.C. 1943. The classical stationary state. *Economic Journal* 53, 343–51.

Schuster, H., ed. 1979. Univ.-Prof. Dr. Gottfried Haberler. In *Österreicher, die der Welt gehören*. Vienna: Mobil Oil Austria AG.

Staehle, H. 1935. A development of the economic theory of price index numbers. *Review of Economic Studies* 2(June), 163–88.

Willett, T.D. 1982. Gottfried Haberler on inflation, unemployment, and international monetary economics: an appreciation. *Quarterly Journal of Economics* 97, 161–9.

How to cite this article

Chipman, John S. "Haberler, Gottfried (1900–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000003> doi:10.1057/9780230226203.0694

The New Palgrave Dictionary of Economics Online

habit persistence

Stephanie Schmitt-Grohé and Martin Uribe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the concept of habit persistence and its application in macroeconomics and finance. Special attention is given to the role of habit persistence in explaining the equity premium puzzle, observed business-cycle fluctuations and inflation dynamics, and in generating a theory of counter-cyclical markups of prices over marginal costs.

Keywords

asset pricing; business-cycle fluctuations; currency pegs; deep habits; equity premium puzzle; Euler equation; habit persistence; imperfect competition; inflation dynamics; investment adjustment costs; partial equilibrium; risk aversion; sticky prices; variable capacity utilization

Article

Habit persistence, or ‘habit formation’ in its most common representation, is a preference specification according to which the period utility function depends on a quasi-difference of consumption. Specifically, if the utility function without habit formation is given by $\sum_{t=0}^{\infty}\beta^{t}U(c_t)$, where c_t denotes consumption in period *t*, *U* denotes the period utility function, and $\beta\in(0,1)$ denotes the subjective discount factor, then the utility function with habit persistence is given by $\sum_{t=0}^{\infty}\beta^{t}U(c_t-\alpha c_{t-1})$. The parameter $\alpha\in(0,1)$ denotes the intensity of habit formation and introduces non-separability of preferences over time. Under habit persistence, an increase in current consumption lowers the marginal utility of consumption in the current period and increases it in the next period. Intuitively, the more the consumer eats today, the hungrier he wakes up tomorrow. It is in this sense that this type of preferences captures the notion of habit formation.
In the habit-forming preferences given above, past consumption represents the consumer's stock of habit in period *t*. More general specifications allow for the stock of habit to be a function of possibly all past

consumptions. In this case, the period utility function is given by $U(c_t - \alpha S_{t-1})$, where $S_{t-1} = S(c_{t-1}, c_{t-2}, \ldots)$ denotes the stock of habit in period *t*. Often, the stock of habit is assumed to follow an autoregressive law of motion of the form $S_t = (1-\delta)S_{t-1} + \lambda c_t$. The parameter δ governs the rate of depreciation of the stock of habit, and the parameter λ measures the sensitivity of the stock of habit to current consumption.

A common variant of the habit persistence model is to treat habits as external to the consumer. When habits are external, the stock of habit depends on the history of aggregate past consumption as opposed to the consumer's own past consumption. Early formulations of the habit formation model, for example Pollak (1970), were cast in the external form. Since the work of Abel (1990), external habit formation has become known as ‘catching up with the Joneses’. The external form of habit persistence simplifies the optimization problem of the consumer because the evolution of the stock of habit is taken as exogenous by the individual.

Another variation of the habit formation model is relative habit persistence, which features a quasi-ratio of consumption rather than a quasi-difference of consumption, as the argument of the period utility function (Duesenberry, 1949; Abel, 1990).

Habit persistence and the equity premium puzzle

Habit persistence has been proposed in financial economics as a possible solution to the equity premium puzzle first identified in the seminal work of Mehra and Prescott (1985). The equity premium puzzle is that, under the assumption of power utility and no habit persistence, observed excess returns of stocks over less risky assets, such as commercial paper, are too high to be consistent with actual consumption behaviour unless households are assumed to be extremely risk averse. At the heart of the equity premium puzzle lies the low volatility of observed consumption growth. To see this, note that a risky asset commands a high rate of return if it provides poor insurance against consumption fluctuations by paying plenty in periods of high consumption growth and little in periods of low consumption growth. If fluctuations in consumption growth are small (as is observed in the data), then high returns on risky assets can be supported only if one assumes that even minute consumption fluctuations are very painful to consumers. In other words, one must assume that consumers are extraordinarily risk averse.

With this intuition in mind, one can readily see why habit persistence has the potential to solve the equity premium puzzle. Habit-forming consumers dislike variations in habit-adjusted consumption, $c_t - \alpha S_{t-1}$, rather than variations in consumption itself, c_t. A given percentage change in consumption produces a much larger percentage change in habit-adjusted consumption than in consumption itself. In this way, small fluctuations in consumption growth can generate large variations in habit-adjusted consumption growth and hence explain sizable excess returns on risky assets even for moderate values of the degree of risk aversion. Early studies of the ability of habit persistence to resolve the equity premium puzzle include Sundaresan (1989), Abel (1990), and Constantinides (1990). Subsequent work has refined the habit-formation model to account for additional asset-pricing puzzles, such as the risk-free-rate puzzle and the forecastability of excess returns (see, for example, Campbell and Cochrane, 1999).

Habit persistence and the business cycle

In the asset-pricing literature, most applications of habit persistence are conducted within the context of partial equilibrium settings, in which private consumption is assumed to be exogenous. This assumption is not inconsequential. Indeed, it has been shown that, once a general equilibrium approach is adopted, in which consumption decisions are endogenous, the ability of habit persistence to reconcile the behaviour of asset prices and consumption is diminished. This is because habit formation induces excess smoothness in consumption expenditure (Lettau and Uhlig, 2000). Boldrin, Christiano and Fisher (2001) show that habit formation can help explain salient aspects of asset prices and business cycles only in combination with severe inflexibilities in factor markets.

Habit persistence features prominently in the literature devoted to the estimation of medium-scale macroeconomic models (for example, Christiano, Eichenbaum and Evans 2005; Smets and Wouters, 2004). The goal of this literature is to build dynamic general equilibrium models capable of explaining the observed behaviour at business-cycle frequency of a large number of macroeconomic variables. To this end, this literature has brought together in a single model most of the theoretical advances in business-cycle theory since the mid-1980s. Thus, these models include not only habit persistence but also other rigidities such as investment adjustment costs, variable capacity utilization, sticky product and factor prices, money demand by households and firms, and imperfect competition in product and factor markets. In the data, the response of consumption to expansionary shocks of various natures is hump-shaped, with the peak response occurring several quarters after the innovation. Such a response is hard to replicate in the absence of habit formation. For in this case consumption has a tendency to peak immediately after the shock and then to decline to its long-run level.

In the applications of habit formation discussed thus far, it matters little whether habits are of the internal or external type. The distinction is of importance in situations in which the consumer expects a regime shift of some nature in the future. A case in point is the consumption dynamics associated with temporary exchange-rate-based inflation stabilization programmes. Exchange-rate-based stabilization programmes, or currency pegs, constitute the most commonly used policy to control high inflation in emerging-market countries. It is well documented that, in response to the announcement of a currency peg, consumption rises initially, reaches a peak and then declines. Importantly, the observed eventual decline in consumption typically takes place before the currency peg is abandoned. Habit formation, be it of the internal or external type, can explain the observed gradual increase in consumption after the implementation of the stabilization plan (Uribe, 2002). However, Uribe shows that the observed contraction in consumption that begins well before the collapse of the stabilization programme can be rationalized with internal habit formation but not with the external form of habits. In effect, maintaining a high consumption habit after the collapse of the stabilization programme is expensive because high inflation acts as a tax on consumption expenditures. When consumers internalize the habitual nature of consumption they start reducing their stock of habit – by cutting back consumption – before the price of consumption increases to mitigate the transition to a lower stock of habit. By contrast, when consumers do not internalize the habitual nature of their consumption, they continue to take advantage of the temporarily low price of consumption until the last day of the stabilization programme.

Deep habits

All of the models of habit persistence discussed thus far in this article assume that habits are formed at the level of a single aggregate consumption good. An important consequence of this assumption is that the introduction of habit formation alters the propagation of macroeconomic shocks only in so far as it modifies the consumption Euler equation and possibly the household's labour supply schedule. Ravn, Schmitt-Grohé and Uribe (2006), hereafter RSU, propose a general equilibrium model of habit formation on a good-by-good basis. They refer to this type of habit formation as ‘deep habits’. They have in mind environments in which consumers can form habits separately over narrowly defined categories of goods, such as clothing, vacation destinations, music, and cars.

The assumption that agents can form habits on a good-by-good basis has two important implications for macroeconomic dynamics. First, the demand side of the macroeconomy – in particular the consumption Euler equation – is indistinguishable from that pertaining to an environment in which agents form habits over a single aggregate good. Second, and more significantly, the assumption of deep habit formation alters the supply side of the economy in fundamental ways. Specifically, when habits are formed at the level of individual goods, firms take into account the fact that the demand they will face in the future depends on their current sales. This is because higher consumption of a particular good in the current period makes consumers, all other things equal, more willing to buy that good in the future through the force of habit. Thus, when habits are deeply rooted, the optimal pricing problem of the firm becomes dynamic.

RSU embed the deep-habit-formation assumption in an economy with imperfectly competitive product markets. This combination results in a model of endogenous, time-varying markups of prices over marginal cost. A central result of RSU's work is that in the deep habit model markups behave counter-cyclically in equilibrium. In particular, expansions in output driven by demand shocks are accompanied by declines in markups. This implication of the deep-habit model is in line with the existing empirical literature. In addition, RSU show that, because of the strong counter-cyclical movements of markups, the deep-habit theory is capable of explaining increases in wages and consumption in response to a positive demand shock as is observed in the data. This latter empirical regularity has proved difficult to explain with standard models of the transmission of demand shocks.

In the deep-habit model it is of great consequence whether habits are internal or external. RSU show that the firm's pricing problem is time consistent under external habit persistence but time inconsistent under internal habit persistence.

See Also

- consumption externalities

Bibliography

Abel, A. 1990. Asset prices under habit formation and catching up with the Joneses. *American Economic Review* 80, 38–42.

Boldrin, M., Christiano, L. and Fisher, J. 2001. Habit persistence, asset returns, and the business cycle. *American Economic Review* 91, 149–66.

Campbell, J. and Cochrane, J. 1999. By force of habit: a consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107, 205–51.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–45.

Constantinides, G. 1990. Habit formation: a resolution of the equity premium puzzle. *Journal of Political Economy* 98, 519–43.

Duesenberry, J. 1949. *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.

Lettau, M. and Uhlig, H. 2000. Can habit formation be reconciled with business cycle facts? *Review of Economic Dynamics* 3, 79–99.

Mehra, R. and Prescott, E. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Pollak, R. 1970. Habit formation and dynamic demand functions. *Journal of Political Economy* 78, 745–63.

Ravn, M., Schmitt-Grohé, S. and Uribe, M. 2006. Deep habits. *Review of Economic Studies* 73, 195–218.

Smets, F. and Wouters, R. 2004. Shocks and frictions in US business cycles: a Bayesian DSGE approach. Mimeo. European Central Bank.

Sundaresan, S. M. 1989. Intertemporally dependent preferences and the volatility of consumption and wealth. *Review of Financial Studies* 2, 73–89.

Uribe, M. 2002. The price-consumption puzzle of currency pegs. *Journal of Monetary Economics* 49, 533–69.

How to cite this article

Schmitt-Grohé, Stephanie and Martin Uribe. "habit persistence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000004> doi:10.1057/9780230226203.0695

The New Palgrave Dictionary of Economics Online

Hadley, Arthur Twining (1856 – 1930)

Robert B. Ekelund, Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; cartels; economic evolution; exhaustible resources; Hadley, A. T.; Monopoly; period analysis; price discrimination; profit maximization; property rights; state capture; transport markets

Article

American economist, educator and public servant, Hadley was educated at Yale and at the University of Berlin, where he studied under German historicists. In a remarkable career, Hadley was, in turn, a freelance writer and lecturer on railway economics, a professor of political economy at Yale (1891–9), president of the American Economic Association, president of Yale University (1899–1921), chairman of the Railroad Securities Commission providing the Hadley Report on Railway finances in 1911, and was widely sought after as a political candidate for high political office in the United States. An inveterate traveller, Hadley died aboard ship in Kobe harbour in 1930.

Hadley was an extremely prolific and eclectic writer, but the bulk of his important work in economics was completed before the turn of the 20th century. His reputation rests essentially on two works, *Railway Transportation* (1885) and a basic text, *Economics: An Account of the Relations between Private Property and Public Welfare* (1896), which received high praise from his friend and colleague, Irving Fisher.

In *Railway Transportation* Hadley revealed himself as the most creative railway economist of the day through an integration of sophisticated (certainly for the time) economic analysis with the problems of railway organization. Among other theoretical insights Hadley formalized a theory of monopoly and price discrimination; developed, in the mathematical terms of Cournot, a marginal rule for profit maximization; and anticipated the period analysis of Marshall's *Principles*. More importantly, perhaps, he developed a modern and complete theory of cartels, showing that, in the presence of open competition, such unsanctioned behaviour on the part of railroads, would lead to the benefits of competition without the attendant disadvantages. In another perspicacious insight Hadley correctly characterized railway regulation as resulting from the capture, by the industry, of legal sanctions to

obtain rate stability. In the main, Hadley viewed regulation as representing a low-cost cartel enforcement device.

In *Economics* Hadley went further than Marshall by explicitly developing the interrelations between property rights, economic evolution and economic efficiency. Hadley utilized the real world examples of the fisheries and mining to demonstrate the impact of ill-defined property rights on depletable resources, emphasizing the necessity of altered systems to obtain optimal resource use and allocation. This contribution, along with his prophetic analyses of transport market structure, establishes Hadley as one of the most inventive pre–20th-century American economists.

Selected works

1885. *Railway Transportation: Its History and its Laws*. New York.

1890. The prohibition of railroad pools. *Quarterly Journal of Economics* 4(January), 158–71.

1896. *Economics: An Account of the Relations between Private Property and Public Welfare*. New York.

Bibliography

Cross, M.L. and Ekelund, R.B., Jr. 1980. A.T. Hadley on monopoly theory and railway regulation: an American contribution to economic analysis and policy. *History of Political Economy* 12, 214–33.

Cross, M.L. and Ekelund, R.B., Jr. 1981. A.T. Hadley: the American invention of the economics of property rights and public goods. *Review of Social Economy* 39, 37–50.

Fisher, I. 1930. Obituary: Arthur Twining Hadley. *Economic Journal* 40, 526–33.

Locklin, D.P. 1933. The literature on railway rate theory. *Quarterly Journal of Economics* 47(February), 167–230.

How to cite this article

Ekelund, Robert B., Jr. "Hadley, Arthur Twining (1856–1930)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000005> doi:10.1057/9780230226203.0696

The New Palgrave Dictionary of Economics Online

Hagen, Everett Einar (1906 – 1993)

Robert R. Nathan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic development; Keynesianism; New Deal; population growth; technical change; unemployment

Article

Hagen was born in Holloway, Minnesota. He graduated from St Olaf College (BA, 1927) and the University of Wisconsin (MA, 1932; Ph.D., 1941). After a short period at the University of Illinois (1948–51) he became professor of economics at the Massachusetts Institute of Technology (1953–72); from 1970 to 1972 he was Director of the Center for International Studies at MIT.

Since the Second World War, developing nations have received unprecedented attention from economists and large financial resources from the industrialized world. Dr Hagen was an important contributor to analysing key problems and processes of economic development.

Before concentrating on economic development, Hagen served in the Bureau of the Budget as a close associate of Gerhard Colm in the application of Keynesian principles to US fiscal policies. His firm commitment to Keynes's concepts was a factor in his transfer to the MIT from the University of Illinois, where more traditionalist faculty and top officialdom were hostile to the views of Keynes and of the New Deal.

In his book *On the Theory of Social Change* (1962), Hagen correctly concluded that economics alone could not provide the theoretical or policy directions for economic development. He studied deeply the role of human behaviour based on studies of anthropologists, sociologists and political scientists. Hagen's multidisciplinary approach provided invaluable insights for formulating development plans and policies.

In his fourth edition of *The Economics of Development* (1986), Hagen continued to elaborate on theoretical aspects as well as policies and implementation processes essential for development progress. Hagen updates the most promising lessons from successful nations replicable in the lagging nations. Hagen disputes the common view that high population growth rates are a major deterrent to development. He also documents the thesis that protectionism is helpful to the developing world. He sets forth a strong case for attributing considerable unemployment to technological change. These somewhat

unorthodox views are persuasively articulated and documented.
Of major importance are Hagen's conceptual formulations, his analyses based on personal experiences, and his challenges to economists and members of other disciplines to work jointly to overcome the persistent barriers to significant progress in the lagging nations.

Selected works

1962. *On the Theory of Social Change*. Homewood, IL: Dorsey Press.

1963., ed. *Planning Economic Development*. Homewood, IL: Richard D. Irwin.

1968. *The Economics of Development*. Homewood, IL: Richard D. Irwin. Revised, 1980, 1986.

How to cite this article

Nathan, Robert R. "Hagen, Everett Einar (1906–1993)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000006> doi:10.1057/9780230226203.0697

The New Palgrave Dictionary of Economics Online

Hamilton, Alexander (1755 – 1804)

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

credit; Hamilton, A.; implied powers; Jefferson, T.; Protection; public credit; public debt; subsidies

Article

One of the founding fathers of the United States and Secretary of the Treasury in President Washington's cabinet, in which Thomas Jefferson served as Secretary of State. The two great men differed widely in their views about the destiny of the young nation. Jefferson wanted to preserve the position of the states and assign to the national government not much more than authority over foreign affairs. Hamilton favoured a strong and active central government. Jefferson was eager to preserve the rural economy in which he had grown up in Virginia. Hamilton proposed to promote economic development, especially manufacture, and vest in the national government the function of actively fostering such development. Jefferson took a dim view of public debts, paper money and financial institutions. Hamilton favoured them all. Jefferson was more of an egalitarian and had greater faith in the common man than Hamilton, who placed his trust in an alliance of government and the aristocracy of wealth: neither could flourish without the support of the other. Hamilton died in a duel with a political adversary during Jefferson's presidency, but his ideas were strong enough to survive him. The exigencies of the time caused Jefferson himself to adopt a number of Hamiltonian policies.

Thus Hamilton became the architect of what in *The Federalist* (1787, No. XI) he had called ‘the great American system’, later to be buttressed by such economic writers as the Careys, Daniel Raymond and Frederick List, and by Henry Clay in politics. He set forth his economic ideas in a series of state papers, published under his name when serving as Secretary to the Treasury. These papers are the first and second *Report on the Public Credit* (1790a; 1795), the *Report on a National Bank* (1790b), and the *Report on Manufactures* (1791). The state papers are justly famous, not as repositories of economic analysis, but as a masterly presentation of a case of which Hamilton, who had been trained in the law, was an eloquent advocate.

The apotheosis of credit found in Hamilton's reports refers to public and private credit as well. According to Hamilton, credit is a substitute of capital almost as useful as gold and silver. It has, and

Hamilton has no doubt about this, a tendency to lower the interest rate. If the public credit is in a bad state, it can only have deleterious effects on private credit. The preservation of a healthy credit system is thus an important task. Foreign creditors should enjoy the same protection as domestic ones. Domestic holders of the public debt should be protected against the imposition of taxes on the public funds, as foreign creditors should be protected from repudiation or expropriation.

With the help of the public debt it will be possible to promote the economic development of the country. Scrupulous attention must be paid to the rights of the creditors, both for the sake of public expediency and as a moral obligation. The public debt of the United States should be funded, that is, arrangements should be made for the service of the debt by putting aside funds for the payment of interest and principal. A funded debt has great benefits. It will facilitate the use of instruments of debt as money and bring about lower interest rates, and will result in an increase in land values, which have declined in consequence of the scarcity of money.

Hamilton also proposed that the Union assume responsibility for the debts of the states, and that the funding of the debt should be financed in part from new duties on imported spirits and taxes on domestic ones and on stills. These proposals met considerable opposition because of the windfall gains that would accrue to speculators who had purchased instruments of the debt at low prices. To obtain Jefferson's support for this measure, Hamilton had to agree that the future capital of the nation would be located in the South, that is, in what is now Washington, DC.

The national bank, which Hamilton proposed to establish, was designed to aid in the expansion of the money supply, thereby facilitating the payment of taxes, the reduction of interest rates, the fulfilment of public functions, and the development of the national economy. The bank was to be under private rather than public direction, with the government playing the role of a minor shareholder. When the question was raised whether the Constitution granted the federal government the authority to establish a bank, Hamilton resolved it by referring to ‘implied powers’, that is, the power to employ suitable means to pursue constitutional ends. This solution was to have far-reaching consequences for the future development of constitutional law.

The *Report on Manufactures* goes into considerable detail examining the relative merits of agriculture and industry. Hamilton underlines the merits of both and the benefits which each derives from the other. He stresses that both are productive, a point that had to be made, and made forcefully, in view of the teachings of the Physiocrats. Hamilton demonstrates great ingenuity in enumerating the factors that are responsible for favourable effects of industrial development on the national income. Among these factors he mentions the division of labour, the more extensive use of machinery, the utilization of manpower that is not suited for agricultural pursuits, the promotion of immigration, the widening of opportunities for the exercise of entrepreneurial talent, and the strengthening of demand for agricultural products.

As far as international trade is concerned, Hamilton holds that the benefits from free trade are more imaginary than real because of the obstacles which foreign countries place in the way of United States exports. Moreover, foreign governments support domestic industries in various ways, and the United States should adopt similar policies by imposing protective and prohibitive duties, granting subsidies to domestic industries, and promoting internal improvements that facilitate the flow of commerce. Subsidies are liable to be abused, but their advantages outweigh the disadvantages. Lastly, Hamilton proposes that a board be established to promote economic development by bringing in skilled workers from abroad, rewarding useful improvements and inventions, paying premiums to importers of

machinery, and similar means.

Selected works

1790a. *Report on the Public Credit*. Washington, DC: Government Printing Office, 1908.

1790b. *Report on a National Bank*. Washington, DC: Government Printing Office, 1908.

1791. *Report on Manufactures*. Washington, DC: Government Printing Office, 1913.

1795. *Report on the Public Credit*.

1934. *Papers on Public Credit, Commerce and Finance*, ed. S. McKee, Jr. New York: Columbia University Press.

Bibliography

Dorfman, J. 1946. *The Economic Mind in American Civilization 1606–1865*, vol. 1. New York: Viking.

Spiegel, H.W. 1960. *The Rise of American Economic Thought*. Philadelphia: Chilton.

Bibliographic addendum

R. Chernow, *Alexander Hamilton*, New York: Penguin, 2004, has established itself as a standard biography. F. McDonald, *Alexander Hamilton*, New York: W. W. North, 1979, is a spirited defence of Hamilton's vision for American development. See also E. J. Ferguson, *The Power of the Purse: A History of American Public Finance*, Chapel Hill: University of North Carolina Press, 1961.

How to cite this article

Spiegel, Henry W. "Hamilton, Alexander (1755–1804)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000009> doi:10.1057/9780230226203.0698

The New Palgrave Dictionary of Economics Online

Hamiltonians

Karl Shell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Hamiltonian dynamics arises not only in economic optimization problems but also in descriptive economic models in which there is perfect foresight about asset prices. Hamiltonian dynamics applies in discrete time as well as in continuous time. In discrete time, the system of differential equations is replaced by a closely related system of difference equations. The theory accommodates differential correspondences or difference correspondences, which naturally arise in economics. The Hamiltonian approach through the Hamiltonian function has proved remarkably successful in establishing sufficient conditions for the saddle-point property and related stability questions in a class of optimal economic growth models.

Keywords

continuous and discrete time models; duality; Hamilton, W. R.; Hamiltonian dynamical system; Hamiltonian function; Hamiltonians; Lyapunov functions; optimal-growth theory; overlapping generations models; Poincaré, J. H.; Pontryagin's maximum principle; saddle-point property; transversality conditions

Article

The laws of motion for a perfect-foresight economy, whether centrally planned or competitive, can be described by a Hamiltonian dynamical system or by a simple perturbation thereof. The Hamiltonian dynamical system and the Hamiltonian function which generates it are named for their inventor, the great Irish mathematician William Rowan Hamilton (1805–1865).

Hamilton's differential equations serve as the basic mathematical tool of classical particle mechanics (including celestial mechanics). Let $x(t) = (x_1(t), \ldots, x_i(t), \ldots, x_m(t))$ and $y(t) = (y_1(t), \ldots, y_i(t), y_m(t))$ be *m*-vectors dependent on time *t*. Let *H* be a continuous, differentiable function of *x*, *y*, and *t*, $H: R^m \times R^m \times R \to R$. Think of *H* as the Hamilton's function (HF) which

generates Hamilton's differential equations,

$$dx_i(t) / dt = \partial H(x(t), y(t), t) / \partial y_i(t)$$

and

$$dy_i(t) / dt = \partial H(x(t), y(t), t) / \partial x_i(t)$$

for $i = 1, \ldots, m$. If the Hamiltonian function H depends on time only through the variables $x(t)$ and $y(t)$, i. e., $\partial H / \partial t \equiv 0$, then the corresponding Hamiltonian dynamical system (HDS) is said to be *autonomous*. These differential equations are frequently interpreted in physics as solutions to some extremization problem. In mechanics for example, HDS is implied by the principle of least action. Since economic planning and many other economic problems involve maximization or minimization over time, it is unsurprising that the Hamiltonian formalism has substantial application in economics. Its appeal to economists goes much further than this. There is a duality (conjugacy, in the language of mechanics) between $x_i(t)$ and $y_i(t)$ which allows us to interpret one as a (primal) economic flow and the other as a (dual) economic price. Given this point of view, the Hamiltonian function (HF) itself has important economic interpretations. Hamiltonian dynamics not only arises in economic optimization problems but it also arises in descriptive economic models in which there is perfect foresight about asset prices. Hamiltonian dynamics applies in discrete time as well as in continuous time. In discrete time, the system of differential equations is replaced by a closely related system of difference equations. The right side of the equations describing Hamilton's law of motion need not be single-valued. The theory accommodates differential correspondences or difference correspondences, which naturally arise in economics.
Consider first the application of Hamiltonian approach to the theory of economic growth; see, for example, the Cass–Shell (1976a) volume. A large class of economic growth models can be described by simple laws of motion based on the instantaneous production set T, with feasible production satisfying.

$$(c, z, -k, -l) \in T \subset \{(c, z, -k, -l) | (c, k, l) \geq 0\},$$

where c denotes the vector of consumption-goods outputs, z the vector of net investment-goods outputs, k the vector of capital-goods inputs, and l the vector of primary-goods inputs. There is an equivalent representation of static technological opportunities that is better suited to dynamic analysis: the representation of the static technology by its Hamiltonian function H.
Let p be the vector of consumption-goods prices and q be the vector of investment-goods prices. Define

the Hamiltonian function $H(p, q, k, l)$ by

$$H(p, q, k, l) = \max_{(c', z')} \left\{ pc' + qz' \mid (c', z', -k, -l) \in T \right\},$$

H is defined on the non-negative orthant and can be interpreted as the maximized value of net national product at the output prices (p, q) given input endowments (k, l).

Obviously, if we know the set T, then we know precisely the function H. If T is closed, convex, and permits free disposal, then H is continuous, convex and homogeneous of degree one in the output prices (p, q), and concave in the input stocks (k, l). If H is a function of (p, q, k, l) which is continuous, convex and homogeneous of degree one in (p, q), and concave in (k, l), then H corresponds to a unique T among closed, convex technologies permitting free disposal. In many dynamic applications, it is only the H representation which matters. Relax, for example, the free-disposal assumptions on T. For a given function H, the set T might be unique, but the dynamics would be independent of the particular set T which generated the function H. Relax, as another example, the assumption that T is convex. Given an H which is convex in (p, q), and concave in (k, l), the set T will not be unique, but the continuous dynamics (HDS) will not be altered in an essential way.

Representation of the static technology by the Hamiltonian function permits one to describe the economic laws of motion as a Hamiltonian dynamical system. In continuous time, the motion is described by

$$\dot{k}(t) \in \partial H(p(t), q(t), k(t), l(t)) / \partial q(t)$$

(HDS)

$$\dot{q}(t) \in -\partial H(p(t), q(t), k(t), l(t)) / \partial k(t)$$

where $\dot{k}(t)$ and $\dot{q}(t)$ are vectors of time derivatives and $(\partial H/\partial q)$ and $(\partial H/\partial k)$ are gradients (derivatives when H is differentiable). The first line of (HDS) is immediate from the definition of net investment since it reduces to $\dot{k}(t) = z(t)$, where $z(t)$ is the vector of net investment. The second line is an equal-asset-return condition which reduces to $\dot{q}(t) + r(t) = 0$, where $r(t)$ is the dual vector of shadow rental rates.

For discrete time, the Hamiltonian dynamical system is

$$k_{t+1} \in k_t + \partial H(p_t, q_t, k_t, l_t) / \partial q_t$$

(HDS) $'$

$$q_{t+1} \in q_t - \partial H(p_{t+1}, q_{t+1}, k_{t+1}, l_{t+1}) / \partial k_{t+1}.$$

Line 1 is equivalent to $k_{t+1} = k_t + z_t$ and line 2 is equivalent to $q_{t+1} - q_t - r_{t+1} = 0$, where z_t is the time $r(t)$ gross investment vector and r_{t+1} is the dual vector of shadow capital-goods rental rates in period $(t+1)$.

For openers, let us analyse the case where H is autonomous. This occurs if $p(t) = \bar{p}$ and $l(t) = \bar{l}$ for (HDS) or $p_t = \bar{p}$ and $l_t = \bar{l}$ for (HDS) $'$. Let (q^*, k^*) be a rest point to (HDS) or (HDS) $'$. Hence, we have

$$0 \in \partial H(\bar{p}, q^*, k^*, \bar{l}) / \partial q, 0 \in \partial H(\bar{p}, q^*, k^*, \bar{l}) / \partial k.$$

Consider the linear approximations about (q^*, k^*) of (HDS) and (HDS) $'$ (taken, for example, as if H were quadratic). Study the characteristic roots to the linearized systems. A simple but remarkable theorem due to Poincaré tells us that if λ is a root for the linearized, *autonomous* version of (HDS) then so is $-\lambda$. For the linearized, *autonomous* version of (HDS) $'$, we have if λ is a root, then so also is $1/\lambda$. If for (HDS), we could rule out pure imaginaries ($\mathrm{Re}\,\lambda \neq 0$), then we would have: The dimension of the manifold in (q, k) – space of solutions tending to (q^*, k^*) as $t \to \infty$ is equal to the dimension of the manifold of solutions tending to (q^*, k^*) as $t \to -\infty$ This is the *saddle-point property*, where the manifold of forward solutions and the manifold of backward solutions each have dimension equal to half the total dimension of the space. Similarly, we would have the saddle-point property for (HDS) $'$, if the modulus $|\lambda|$ is unequal to unity.

Poincaré's result nearly gives us the saddle-point property. In the autonomous cases, the saddle-point property can be assured if the geometry of the Hamilton function is correct. We need to add very little to the convexity–concavity assumption (see Cass and Shell, 1976b, and Rockafellar, 1976). Strict convexity in q and strict concavity in k will do the trick. So will a weaker uniform Hamiltonian steepness condition, which reduces to a value-loss condition; see, for example, McKenzie (1968), and Cass–Shell (1976b).

What about non-autonomous systems, such as optimal economic growth with the constant, positive

discount rate ρ ? Here $c(t)$ or c_t, is a scalar called felicity and usually denoted in optimal-growth problems by $u(t)$ or u_t. In this case, present prices must satisfy

$$-\dot{p}(t) / p(t) = \rho$$

or

$$-(p_t - p_{t-1}) / p_t = \rho.$$

For simplicity, allow only for a single fixed factor and adopt the convention $l(t)=1$, or $l_t=1$.
It is natural then to re-express the systems (HDS) and (HDS)$'$ in terms of current prices $Q \equiv q / p$, rather than in terms of present prices q. We then have

$$\dot{k} \in \partial H(Q, k) / \partial Q$$

(PHDS)

$$\dot{Q} \in -\partial H(Q, k) / \partial k + \rho Q$$

and

$$k_{t+1} \in k_t + \partial H(Q_t, k_t) / \partial Q_t$$

(PHDS)$'$

$$Q_{t+1} \in Q_t - \partial H(Q_{t+1}, k_{t+1}) \partial k_{t+1} + \rho Q_t.$$

The systems (PHDS) and (PHDS)′ are *perturbed* Hamiltonian dynamical systems. We no longer have Poincaré's root-splitting theorems in pure form: the roots split but not about 0 for (HDS) nor 1 for (HDS)′. The trick here is to strengthen the geometry of *H* to give a saddle-point property or something like it.

This is the basics of the approach taken by Cass and Shell (1976b), Rockafellar (1976) and Brock and Scheinkman (1976). Conditions are found on *H* which assure that either (PHDS) or (PHDS)′ along with transversality conditions defines a globally stable system. It suffices to strengthen the convexity–concavity of *H* by an amount dependent on *p* or (weaker) to strengthen the steepness of *H* by an amount dependent on *p*. (The Lyapunov function which does the trick is $V = (Q - Q^*)(k - k^*)$ in the continuous-time model.)

The Hamiltonian approach *through the Hamiltonian function* has proved remarkably successful in establishing sufficient conditions for the saddle-point property and related stability questions in a class of optimal economic growth models. The parallel programme of using the Hamiltonian formalism in optimal-growth theory to yield sufficient conditions for cycling or other dynamic configurations has not yet been pursued in a systematic fashion but should prove equally successful when applied. The success of the Hamiltonian approach in decentralized and descriptive growth theory has so far been very limited; see Cass and Shell (1976b, Section 4). This has been disappointing. I still hope to see the Hamiltonian approach playing a pivotal technical role in, say, the dynamical analysis of overlapping-generations models, but there has not been much tangible encouragement for this hope.

Many of us first met Hamiltonian dynamical systems as necessary conditions for intertemporal maximization in the form of Pontryagin's maximum principle; see Pontryagin et al. (1962). See Shell (1967) for applications to economics and references.

Samuelson and Solow (1956) were probably the first in economics to mention the Hamiltonian formalism. For some of the history of Hamiltonian dynamics, in economic, mathematics, and physics, and for some of the classical references, see Magill (1970).

Bibliography

Brock, W.A. and Scheinkman, J.A. 1976. Global asymptotic stability of optimal economic systems with applications to the theory of economic growth. In Cass and Shell (1976a).

Cass, D. and Shell, K. 1976a. *The Hamiltonian Approach to Dynamic Economics*. New York: Academic Press. [Reprinted from the *Journal of Economic Theory* 12, February 1976, Symposium: ‘Hamiltonian dynamics in economics’.]

Cass, D. and Shell, K. 1976b. The structure and stability of competitive dynamical systems. In Cass and Shell (1976a).

Magill, M.J.P. 1970. *On a General Economic Theory of Motion*. Berlin: Springer-Verlag.

McKenzie, L.W. 1968. Accumulation programs of maximum utility and the von Neumann facet. In *Value, Capital and Growth*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

Pontryagin, L.S., Boltyanskii, G., Gamkrelidze, R.V. and Mischenko, E.F. 1962. *The Mathematical Theory of Optimal Processes*. New York: Interscience.

Rockafellar, R.T. 1976. Saddlepoints of Hamiltonian systems in convex Lagrange problems having nonzero discount rate. In Cass and Shell (1976a).

Samuelson, P.A. and Solow, R.M. 1956. A complete model involving heterogeneous capital goods. *Quarterly Journal of Economics* 70, 537–62.

Shell, K., ed. 1967. *Essays on the Theory of Optimal Economic Growth*. Cambridge, MA: MIT Press.

Shell, K. 1969. Applications of Pontryagin's maximum principle to economics. In *Mathematical Systems Theory and Economics, I*, ed. H.W. Kuhn and G.P. Szegö. Berlin: Springer-Verlag.

How to cite this article

Shell, Karl. "Hamiltonians." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000011> doi:10.1057/9780230226203.0699

The New Palgrave Dictionary of Economics Online

Hannan, Edward J. (1921 – 1994)

P.M. Robinson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We review the research of the late Edward J. Hannan. Hannan contributed deeply and influentially to the development of econometric time series analysis, both in the elegance and incisiveness of his technical work and in his invention of new methodology.

Keywords

ARMA models; ARMAX models; band-spectrum regression; econometrics; errors in variables; Hannan, E.; Hannan-efficient estimation; heteroscedasticity; maximum likelihood; semi-parametric estimation; statistics and economics; time series analysis; Whittle estimation

Article

Ted Hannan, who died in 1994 at the age of 72, made contributions to statistical time series analysis of considerable depth and originality. His research by no means focused entirely on problems with econometric motivation, and he stopped publishing in econometric journals in the 1970s. However, Hannan introduced important econometric methodology and theory, and some of his ideas anticipated themes that later became important in econometrics. I focus on his most econometric-related contributions rather than attempt to survey the breadth of his research (of which an account can be found in Robinson, 1994, on which we draw upon here).

Hannan in fact started out as an economic researcher at the Reserve Bank of Australia, in Sydney, following an undergraduate commerce degree. While visiting the Australian National University (ANU), Canberra, in 1953 he was ‘discovered’ by the then Professor of Statistics P. A. P. Moran, who encouraged him to undertake doctoral research in statistics. Hannan quickly completed a Ph.D. and within a few years achieved professorial status at the ANU, retiring in 1986 but continuing to be very productive in research up to his death. Altogether he published over 130 articles and five books (all of which are definitely in the research monograph category).

Hannan's intellectual development was unusual in that his mathematical abilities and taste for abstraction increased throughout his career, suggesting that a different early training might well have led to a career in pure mathematics. This partly explains why it is the earlier part of his career in which he did most of his econometric-related work.

In the early 1950s testing for zero autocorrelation was a major theme of time series research. Two of Hannan's first papers, published in 1955 in *Biometrika*, concerned exact tests for autocorrelation. However he quickly realized the limitations of finite-sample theory, and began the research on asymptotic theory which he developed with such originality and power during the rest of his years. His first published contribution to asymptotics, in 1956, concerned Pitman efficiencies.

Soon thereafter he wrote an early, and widely uncredited, contribution to a topic that has, since the mid-1980s, been actively pursued in econometrics, so-called ‘heteroscedasticity-and-autocorrelation consistent variance estimation’. It had already been noticed by Grenander and Rosenblatt that the variance of the sample mean of a weakly dependent time series is approximately proportional to the spectral density at zero frequency. On the other hand, Parzen and others had recently developed consistent smoothed estimation of nonparametric spectral densities. Hannan put these ideas together in a 1957 paper in the *Journal of the Royal Statistical Society*. The awareness he displayed here of the importance of bandwidth choice was notable for the time. Other early contributions included bias-correction in spectrum estimation for detrended data.

One major innovation for which Hannan does receive credit is what econometricians know as ‘Hannan-efficient estimation’. The problem is one of efficiently estimating regression coefficients when the disturbances have nonparametric autocorrelation, that is, to do as well asymptotically as if one correctly assumed the disturbance followed a particular parametric model, such as an autoregression (as Cochrane and Orcutt had assumed). Based on the property of unitary transformation of a stationary time series to a heteroscedastic, approximately uncorrelated, one, Hannan, in a paper published in the 1963 Brown Symposium proceedings, proposed a frequency-domain generalized least squares procedure involving inverse weighting by the disturbance spectral density. Moreover, he established its asymptotic normality and efficiency. This was perhaps the first instance of justifying smoothed nonparametric estimation in a semi-parametric model. The technical difficulties here, of establishing parametric convergence rates despite a slowly converging nonparametric nuisance function, are now familiar, but Hannan was perhaps the first to solve them. He and others subsequently developed the approach in more general models, but it is most notable that his 1963 paper preceded by over 20 years work in the analogous problem of adapting to heteroscedasticity of nonparametric form, and by over ten years work in adapting to distribution of unknown form in location and regression models, though papers on these topics rarely mention his work. A related paper, also published in 1963, in *Biometrika*, is also insufficiently cited. There, Hannan introduced what subsequently became known as ‘band-spectrum regression’, omitting from a frequency-domain formulation of the least squares estimate non-degenerate bands of frequencies, with the aim of reducing errors-in-variables bias.

Around the same time Hannan introduced new ideas in the seasonal analysis of time series, using operators in estimating seasonality in the presence of trend and stationary noise, and modelling the seasonal component by a cosinusoid whose coefficients form stationary processes. He developed this model in a 1967 *Journal of Applied Probability* paper, allowing the coefficients to have roots on the unit circle, years before unit roots became the focus of so much econometric interest. Indeed, Hannan's model resembles the random component models that subsequently became popular. One non-time series

contribution to econometrics was his work on the relation between canonical correlation and simultaneous equations estimation, which led to a 1967 *Econometrica* paper.

Hannan had published a short but influential 1960 monograph on time series analysis, and in 1970 he developed this into the major work *Multiple Time Series*. This constituted a detailed and rigorous account of time series, mainly in a multivariate setting, and covering continuous-time as well as discrete-time processes. It has been an invaluable reference for researchers, stimulating new research on a number of aspects.

Like the earlier book, *Multiple Time Series* put considerable stress on the frequency domain, but not exclusively and certainly not focusing particularly on nonparametric spectral methods. The frequency domain is sometimes misguidedly identified with nonparametric spectral estimation, but, just as nonparametric time series analysis can be considered in the time domain, so frequency-domain inference on parametric models is possible, and indeed the frequency domain is basic in much theoretical discussion of time series. Hannan's work showed how combining time- and frequency-domain assumptions can lead to an incisive theory, and also demonstrated immense resourcefulness in using techniques from Fourier analysis and other areas of mathematics in his proofs.

These qualities stand out in Hannan's work on linear time series models, which were the focus of much of his effort from around 1970 on. The publication of Box and Jenkins’ 1970 book had greatly increased interest, especially among econometricians, in autoregressive moving average (ARMA) modelling. Hannan had already become interested in the topic, as *Multiple Time Series* shows. There, and in 1969 and 1971 papers, he addressed the difficult problem of identification in multivariate ARMA models, which he built on in several subsequent pieces of work.

In a 1973 *Journal of Applied Probability* paper, Hannan gave a rather definitive treatment of the asymptotic theory of various forms of Whittle estimation of scalar ARMA processes. This paper is notable for several aspects, which typify much of his work. First, while ARMA models are the linear time series models of leading practical importance, he showed that models with much stronger autocorrelation can be handled. Indeed, his proof of (strong) consistency actually covered long-range dependent processes, though these had not really been identified as a class at that time. Moreover, he showed that second moments suffice not only for consistency but for asymptotic normality (for parameters describing only autocorrelation). Another feature was his use of martingale difference rather than independence assumptions on innovations. From a methodological viewpoint, Hannan proposed a discrete-frequency version of Whittle estimation, which has computational advantages over Gaussian maximum likelihood or continuous-frequency Whittle, in that it makes nice use of the neat, explicit form of the spectral density for ARMA and some other models, and makes direct use of the fast Fourier transform algorithm. Later, in a 1976 paper with Dunsmuir, Hannan extended the 1973 paper to multivariate ARMA models, and in a 1980 paper with Deistler and Dunsmuir, to models with lagged explanatory variables (‘ARMAX’ models). From 1979 onwards, Hannan made several major contributions to the practically important problem of order determination in ARMA and ARMAX models. His 1988 book with Deistler, *The Statistical Theory of Linear Systems* collects much of his work on linear time series models.

Ted Hannan is mainly thought of as an outstanding technician, with the ability to elegantly solve highly challenging problems under conditions that are at the same time incisive and comprehensible, but he also repeatedly demonstrated a keen practical sense, inventing new methodology, involving interesting tricks, and well understanding computational issues. His influence in econometrics has been profound

and lasting, but had he chosen to devote much more of his brilliance and energy to econometric problems it would be difficult to overstate the further benefits to the econometric profession that could have accrued.

See Also

- econometrics
- law(s) of large numbers
- serial correlation and serial dependence
- statistics and economics
- time series analysis

Selected works

1955. Exact tests for serial correlation. *Biometrika* 42, 133–42.

1955. An exact test for correlation between time series. *Biometrika* 42, 316–26.

1956. The asymptotic powers of certain tests based on multiple correlations. *Journal of the Royal Statistical Society B* 18, 227–33.

1957. The variance of the mean of a stationary process. *Journal of the Royal Statistical Society B* 19, 282–5.

1960. *Time Series Analysis*. Methuen: London. (Published in Russian with an Appendix by Yu. B. Rozanov in 1964; published in Japanese.)

1963. Regression for time series. In *Time Series Analysis*, ed. M. Rosenblatt. New York: Wiley.

1963. Regression for time series with errors of measurement. *Biometrika* 50, 293–302.

1967. The measurement of a wandering signal amid noise. *Journal of Applied Probability* 4, 90–102.

1967. Canonical correlation and multiple equation systems in economics. *Econometrica* 35, 123–38.

1969. The identification of vector, mixed autoregressive-moving average, systems. *Biometrika* 56, 223–5.

1970. *Multiple Time Series*. New York: Wiley. (Published in Russian in 1974.)

1971. The identification problem for multiple equation systems with moving average errors. *Econometrica* 39, 751–65.

1973. The asymptotic theory of linear time series models. *Journal of Applied Probability* 10, 130–45.

1976. (With W. Dunsmuir.) Vector linear time series models. *Advances in Applied Probability* 8, 339–64.

1979. (With B. Quinn.) The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41, 190–5.

1980. (With W. Dunsmuir and M. Deistler.) Estimation of vector ARMAX models. *Journal of Multivariate Analysis* 10, 275–95.

1988. (With M. Deistler.) *The Statistical Theory of Linear Systems*. New York: Wiley.

Bibliography

Robinson, P. 1994. Edward J. Hannan, 1921–1994. *Journal of Time Series Analysis* 15, 563–76.

How to cite this article

Robinson, P. M. "Hannan, Edward J. (1921–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000178> doi:10.1057/9780230226203.0700

The New Palgrave Dictionary of Economics Online

Hansen, Alvin (1887 – 1975)

R.A. Musgrave
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Bretton Woods System; Commons, J. R.; Council of Economic Advisers; Ely, R. T.; fiscal expansion; full employment; Great Depression; Hansen, A.; International Monetary Fund; Keynesianism; multiplier–accelerator interaction; neoclassical synthesis; saving and investment; stabilization; stagnation

Article

Alvin Hansen grew up in Viborg, South Dakota, a small rural community with a one-room school house and traditional values. Preferring academic pursuits to farm work, he proceeded to Sioux Falls for his high school education and then to Yankton College for his BA degree. Several years of high school teaching followed, with rapid advancement to principal and superintendent. The financial basis for his graduate work thus laid, Hansen entered the University of Wisconsin in 1914, where John R. Commons and R.T. Ely were to impress him with the importance of data and their institutional setting. In 1919 he moved to Brown University as assistant professor. There he completed his dissertation, later published as *Cycles of Prosperity and Depression* (1921). He then accepted a position at the University of Minnesota, where he remained for nearly 20 years. His major works during the 1920s included a solid *Principles* text, co-authored with F.B. Garver (1928), and an historical study of *Business Cycle Theory* (1927). Ranging from Malthus to Spiethoff and Hawtrey, stress was on structural shifts in investment rather than on monetary factors, and special attention was given to the interaction of short cycles with longer waves of economic development.

A Guggenheim fellowship in 1928 permitted extensive travel abroad, an experience that he continued to cherish and renew in later years. The early 1930s also brought a growing policy involvement outside the campus, activities that in subsequent years were to claim an increasing share of his time. Such early activities included that of Director of Research for the Committee of Inquiry on International Economic Relations (1933–34) and service as adviser on trade agreements to Secretary Cordell Hull.

In 1936 Harvard University had received a grant to establish the Littauer School of Public Administration, and Hansen was appointed as its first Lucius S. Littauer Professor of Political Economy.

As fortune had it, his arrival at Harvard in the fall of 1937 closely followed the appearance of Keynes's *General Theory*. Hansen, distressed by the wastes of the Great Depression, soon (though with some initial hesitation) adopted the Keynesian approach. With Harvard's Fiscal Policy Seminar as his base, Hansen became the leading analyst and expositor of Keynesian economics in the United States. Driven by his enthusiasm for new ideas, his determination to find policy solutions, and his eagerness to learn as well as to teach, the seminar left a deep impact on the course of macroeconomics. The still obscure components of the Keynesian system had to be sorted out and new tools, such as the concept of governmental net contribution, the multiplier-accelerator model and the balanced budget theorem, were forged. With the application of these new tools to the setting of the US economy as its challenge, the seminar thus became the training ground for a generation of US policy economists.

The output of these years may be traced in Hansen's writings, beginning with the two key volumes of *Full Recovery and Stagnation* (1938) and *Fiscal Policy and Business Cycles* (1941). Other volumes followed, including *Business Cycles and National Income* (1951) and his widely used *A Guide to Keynes* (1953). The persistent theme was that of unemployment, caused by a failure of private investment to match the level of saving at a full employment income. With the effectiveness of monetary policy reduced by inelastic investment and high liquidity preference in a sluggish economy, the required level of aggregate demand would have to be provided by fiscal expansion responded to in the private sector by a multiplier-accelerator process. The need for expansionary fiscal policy, however, would not be one of pump-priming only. Linking back to his earlier interest in the long waves of the cycle, the weakness of the economy was seen as the downward phase of a long wave, with the declining population growth the most depressing factor. The stagnation thesis, offered in Hansen's presidential address before the American Economic Association (1937), placed the Keynesian model in a historical perspective and once more emphasized the strategic role of expansionary budget policy. Events, to be sure, proved different. The Second World War generated massive budgetary expansion and a strengthened post-war economy called for a correction, combining the traditional role of monetary policy with that of fiscal controls. Hansen the pragmatist welcomed the neoclassical synthesis of the mid-1960s.

While macro policy and stabilization remained his major concern, his activities during the Harvard years covered a much wider range. As a member of the Advisory Council on Social Security in 1937–38, he helped to shape the Social Security System. During 1941–43 he served as Chairman of the US–Canadian Joint Economic Commission, and from 1940 to 1945 he acted as Economic Advisor to the Federal Reserve Board. At the close of the war he participated in the monetary reconstruction of Bretton Woods and the birth of IMF. At the same time, he played a strategic role in the creation of the Full Employment Act of 1946 and the Council of Economic Advisers. After retiring from Harvard in 1956, Hansen remained in Belmont, Massachusetts, until 1972, when he joined his daughter in Virginia. He died there in 1975.

Throughout Hansen's work, the goal of full employment was central, as was the need for fiscal action to achieve it. His social philosophy was expressed ‘in the democratic ideal of providing for all individuals a reasonable approach to equality of opportunity’. Beyond this, he was pragmatic and non-ideological in approach. For him, economics – as James Tobin put it when presenting him with the Walker Medal at his 80th birthday – was a science for the service of mankind.

Selected works

1921. *Cycles of Prosperity and Depression in the United States, Great Britain and Germany: A Study of Monthly Data 1902–1908*. Madison: University of Wisconsin Press.

1927. *Business-Cycle Theory: Its Development and Present Status*. Boston: Ginn.

1928. (With F.B. Garver.) *Principles of Economics*. Boston: Ginn.

1938. *Full Recovery or Stagnation?* New York: Norton.

1941. *Fiscal Policy and Business Cycles*. New York: Norton.

1951. *Business Cycles and National Income*. New York: Norton.

1953. *A Guide to Keynes*. New York: McGraw-Hill.

How to cite this article

Musgrave, R. A. "Hansen, Alvin (1887–1975)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000014> doi:10.1057/9780230226203.0701

The New Palgrave Dictionary of Economics Online

happiness, economics of

Carol Graham
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economics of happiness assesses welfare by combining economists' and psychologists' techniques, and relies on more expansive notions of utility than does conventional economics. The research highlights factors other than income that affect well-being. It is well suited to informing questions in areas where revealed preferences provide limited information – for example, the welfare effects of inequality and of inflation and unemployment. Despite the potential contributions for policy, a note of caution is necessary because of the potential biases in survey data and the difficulties in controlling for unobservable personality traits.

Keywords

behavioural economics; bounded rationality; capabilities-base approach to poverty; democracy; Easterlin paradox; Easterlin, R.; expectations; expressed preferences; gender; income; vs. utility; inequality; interdependent utility; misery index; order bias; poverty; procedural utility; psychologists; revealed preferences; race; Sen, A.; status; utility maximization; Walras, L.

Article

The economics of happiness is an approach to assessing welfare which combines the techniques typically used by economists with those more commonly used by psychologists. While psychologists have long used surveys of reported well-being to study happiness, economists only recently ventured into this arena. Early economists and philosophers, ranging from Aristotle to Bentham, Mill, and Smith, incorporated the pursuit of happiness in their work. Yet, as economics grew more rigorous and quantitative, more parsimonious definitions of welfare took hold. Utility was taken to depend only on income as mediated by individual choices or preferences within a rational individual's monetary budget constraint.

Even within a more orthodox framework, focusing purely on income can miss key elements of welfare. People have different preferences for material and non-material goods. They may choose a lower-paying but more personally rewarding job, for example. They are nonetheless acting to maximize utility in a classically Walrasian sense.

The study of happiness or subjective well-being is part of a more general move in economics that challenges these narrow assumptions. The introduction of bounded rationality and the establishment of behavioural economics, for example, have opened new lines of research. Happiness economics – which represents one new direction – relies on more expansive notions of utility and welfare, including interdependent utility functions, procedural utility, and the interaction between rational and non-rational influences in determining economic behaviour.

Richard Easterlin was the first modern economist to revisit the concept of happiness, beginning in the early 1970s. More generalized interest took hold in the late 1990s (see, among others, Easterlin, 1974; 2003; Blanchflower and Oswald, 2004; Clark and Oswald, 1994; Frey and Stutzer, 2002a; Graham and Pettinato, 2002; Layard, 2005).

The approach

The economics of happiness does not purport to replace income-based measures of welfare but instead to complement them with broader measures of well-being. These measures are

based on the results of large-scale surveys, across countries and over time, of hundreds of thousands of individuals who are asked to assess their own welfare. The surveys provide information about the importance of a range of factors which affect well-being, including income but also others such as health, marital and employment status, and civic trust.

The approach, which relies on expressed preferences rather than on revealed choices, is particularly well suited to answering questions in areas where a revealed preferences approach provides limited information. Indeed, it often uncovers discrepancies between expressed and revealed preferences. Revealed preferences cannot fully gauge the welfare effects of particular policies or institutional arrangements which individuals are powerless to change. Examples of these include the welfare effects of inequality, environmental degradation, and macroeconomic policies such as inflation and unemployment. Sen's capabilities-based approach to poverty, for example, highlights the lack of capacity of the poor to make choices or to take certain actions. In many of his writings, Sen (1995) criticizes economists' excessive focus on choice as a sole indicator of human behaviour. Another area where a choice approach is limited and happiness surveys can shed light is the welfare effects of addictive behaviours such as smoking and drug abuse.

Happiness surveys are based on questions in which the individual is asked, 'Generally speaking, how happy are you with your life' or 'How satisfied are you with your life', with possible answers on a four-to-seven point scale. Psychologists have a preference for life satisfaction questions. Yet answers to happiness and life satisfaction questions correlate quite closely. The correlation coefficient between the two – based on research on British data for 1975–92, which includes both questions, and Latin American data for 2000–1, in which alternative phrasing was used in different years – ranges between .56 and .50 (Blanchflower and Oswald, 2004; Graham and Pettinato, 2002).

This approach presents several methodological challenges (for a fuller description of these, see Bertrand and Mullainathan, 2001; Frey and Stutzer, 2002b). To minimize order bias, happiness questions must be placed at the beginning of surveys. As with all economic measurements, the answer of any specific individual may be biased by idiosyncratic, unobserved events. Bias in answers to happiness surveys can also result from unobserved personality traits and correlated measurement errors (which can be corrected via individual fixed effects if and when panel data are available). Other concerns about correlated unobserved variables are common to all economic disciplines.

Despite the potential pitfalls, cross-sections of large samples across countries and over time find remarkably consistent patterns in the determinants of happiness. Many errors are uncorrelated with the observed variables, and do not systematically bias the results. Psychologists, meanwhile, find validation in the way that people answer these surveys based in physiological measures of happiness, such as the frontal movements in the brain and in the number of 'genuine' – Duchenne – smiles (Diener and Seligman, 2004).

Micro-econometric happiness equations have the standard form: $W_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$, where W is the reported well-being of individual i at time t, and X is a vector of known variables including socio-demographic and socioeconomic characteristics. Unobserved characteristics and measurement errors are captured in the error term. Because the answers to happiness surveys are ordinal rather than cardinal, they are best analysed via ordered logit or probit equations. These regressions typically yield lower R-squares than economists are used to, reflecting the extent to which emotions and other components of true well-being are driving the results, as opposed to the variables that we are able to measure, such as income, education, and marital and employment status. (Cross-section work also typically yields low R-squares.)

The availability of panel data in some instances, as well as advances in econometric techniques, are increasingly allowing for sounder analysis (Van Praag and Ferrer-i-Carbonell, 2004). The coefficients produced from ordered probit or logistic regressions are remarkably similar to those from OLS regressions based on the same equations. While it is impossible to measure the precise effects of independent variables on true well-being, happiness researchers have used the OLS coefficients as a basis for assigning relative weights to them. They can estimate how much income a typical individual in the United States or Britain would need to produce the same change in stated happiness that comes from the well-being loss resulting from, for example, divorce ($100,000) or job loss ($60,000) (Blanchflower and Oswald, 2004).

The Easterlin paradox

In his original study, Easterlin revealed a paradox that sparked interest in the topic but is as yet unresolved. While most happiness studies find that *within* countries wealthier people are, on average, happier than poor ones, studies across countries and over time find very little, if any, relationship between increases in per capita income and average happiness levels. On average, wealthier countries (as a group) are happier than poor ones (as a group); happiness seems to rise with income up to a point, but not beyond it. Yet even among the less happy, poorer countries, there is not a clear relationship between average income and average happiness levels, suggesting that many other factors – including cultural traits – are at play (see Figure 1).

Figure 1

Happiness and income per capita, selected countries, 1990s

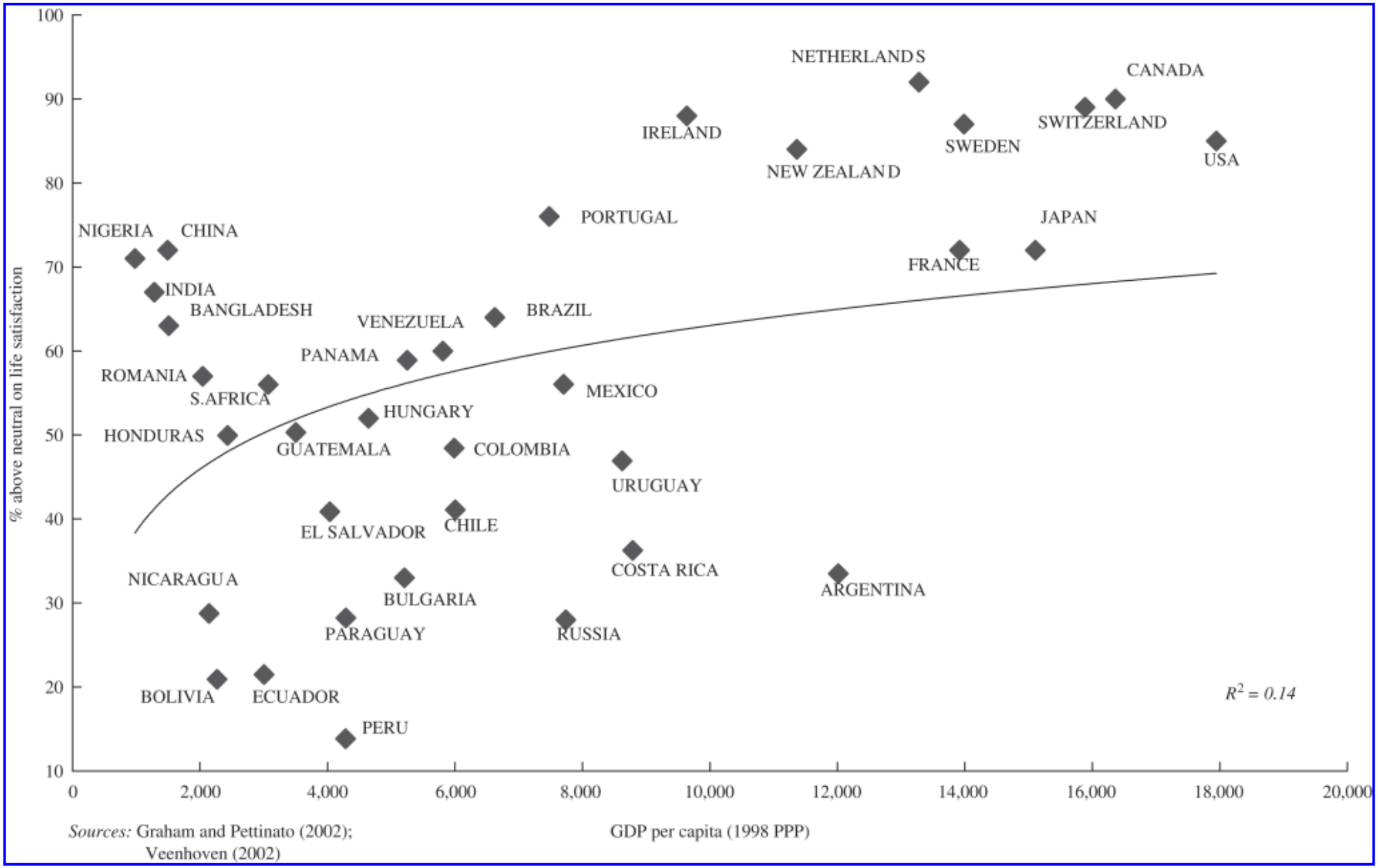

Within countries, income matters to happiness (Oswald, 1997; Diener et al., 2003, among others). Deprivation and abject poverty in particular are very bad for happiness. Yet after basic needs are met other factors such as rising aspirations, relative income differences, and the security of gains become increasingly important, in addition to income. Long before the economics of happiness was established, James Duesenberry (1949) noted the impact of changing aspirations on income satisfaction and its potential effects on consumption and savings rates. Any number of happiness studies have since confirmed the effects of rising aspirations, and have also noted their potential role in driving excessive consumption and other perverse economic behaviours (Frank, 1999).

Thus, a common interpretation of the Easterlin paradox is that humans are on a ‘hedonic treadmill’: aspirations increase along with income and, after basic needs are met, relative rather than absolute levels of income matter to well-being. Another interpretation of the paradox is the psychologists’ ‘set point’ theory of happiness, in which every individual is presumed to have a happiness level that he or she goes back to over time, even after major events such as winning the lottery or getting divorced (Easterlin, 2003). The implication of

this theory for policy is that nothing much can be done to increase happiness.

Individuals are remarkably adaptable, no doubt, and in the end can get used to most things, and in particular to income gains. The behavioural economics literature, for example, shows that individuals value losses more than gains (see Kahneman, Diener and Schwarz, 1999, among others). Easterlin argues that individuals adapt more in the pecuniary arena than in the non-pecuniary arena, while life changing events, such as bereavement, have lasting effects on happiness. Yet, because most policy is based on pecuniary measures of well-being, it overemphasizes the importance of income gains to well-being and underestimates that of other factors, such as health, family, and stable employment.

There is no consensus about which interpretation is most accurate. Yet numerous studies which demonstrate that happiness levels can change significantly in response to a variety of factors suggest that the research can yield insights into human well-being which provide important, if complementary, information for policymakers. Even under the rubric of set point theory, happiness levels can fall significantly in the aftermath of events like illness or unemployment. Even if levels eventually adapt upwards to a longer-term equilibrium, mitigating or preventing the unhappiness and disruption that individuals experience for months, or even years, in the interim certainly seems a worthwhile objective for policy.

Selected applications of happiness economics

Happiness research has been applied to a range of issues. Since a comprehensive review cannot be undertaken here, a selection of some of the issues the surveys can inform is provided. These include the relationship between income and happiness, inequality and poverty, the effects of macro-policies on individual welfare, and the effects of public policies aimed at controlling addictive substances.

Some studies have attempted to separate the effects of income from those of other endogenous factors, such as satisfaction in the workplace. Studies of unexpected lottery gains find that these isolated gains have positive effects on happiness, although it is not clear that they are of a lasting nature (Gardner and Oswald, 2001). Other studies have explored the reverse direction of causality, and find that people with higher happiness levels tend to perform better in the labour market and to earn more income in the future (Diener et al., 2003; Graham, Eggers and Sukhtankar, 2004).

A related question, and one which is still debated in economics, is how income inequality affects individual welfare. Interestingly, the results differ between developed and developing economies. Most studies of the United States and Europe find that inequality has modest or insignificant effects on happiness. The mixed results may reflect the fact that inequality can be a signal of future opportunity and mobility as much as it can be a sign of injustice (Alesina, Di Tella and MacCulloch, 2004). In contrast, recent research on Latin America finds that inequality is negative for the well-being of the poor and positive for the rich. In a region where inequality is much higher and where public institutions and labour markets are notoriously inefficient, inequality signals persistent disadvantage or advantage rather than opportunity and mobility (Graham and Felton, 2006).

Happiness surveys also facilitate the measurement of the effects of broader, non-income components of inequality, such as race, gender, and status, all of which seem to be highly significant (Graham and Felton, 2006). These results find support in work in the health arena, which finds that relative social standing has significant effects on health outcomes (Marmot, 2004).

Happiness research can deepen our understanding of poverty. The set point theory suggests that a destitute peasant can be very happy. While this contradicts a standard finding in the literature – namely, that poor people are less happy than wealthier people within countries – it is suggestive of the role that low expectations play in explaining persistent poverty in some cases. The procedural utilities and capabilities approaches, meanwhile, emphasize the constraints on the choices of the poor.

What is perceived to be poverty in one context may not be in another. People who are high up the income ladder can identify themselves as poor, while many of those who are below the objective poverty line do not, because of different expectations (Rojas, 2004). In addition, the well-being of those who have escaped poverty is often undermined by insecurity and the risk of falling back into poverty. Income data does not reveal the vulnerability of these individuals, yet happiness data shows that it has strong negative effects on their welfare. Indeed, their reported well-being is often lower than that of the poor (Graham and Pettinato, 2002).

Happiness surveys can be used to examine the effects of different macro-policy arrangements on well-being. Most studies find that inflation and unemployment have negative effects on happiness. The effects of unemployment are stronger than those of inflation, and hold above and beyond those of forgone income (Di Tella, MacCulloch and Oswald, 2001). The standard ‘misery index’, which assigns equal weight to inflation and unemployment, may be underestimating the effects of the latter on well-being (Frey and Stutzer, 2002b).

Political arrangements also matter. Much of the literature finds that both trust and freedom have positive effects on happiness (Helliwell, 2004; Layard, 2005). Research based on variance in voting rights across cantons in Switzerland finds that there are positive effects from *participating* in direct democracy (Frey and Stutzer, 2002b). Research in Latin America finds a strong positive correlation between happiness and preference for democracy (Graham and Sukhtankar, 2004).

Happiness surveys can also be utilized to gauge the welfare effects of various public policies. How does a tax on addictive substances, such as tobacco and alcohol, for example, affect well-being? A recent study on cigarette taxes suggests that the negative financial effects may be outweighed by positive self-control effects (Gruber and Mullainathan, 2005).

Policy implications

Richard Layard (2005) makes a bold statement about the potential of happiness research to improve people's lives directly via changes in public policy. He highlights the extent to which people's happiness is affected by status – resulting in a rat race approach to work and to income gains, which in the end reduces well-being. He also notes the strong positive role of security in the workplace and in the home, and of the quality of social relationships and trust. He identifies direct implications for fiscal and labour market policy – in the form of taxation on excessive income gains and via re-evaluating the merits of performance-based pay.

While many economists would not agree with Layard's specific recommendations, there is nascent consensus that happiness surveys can serve as an important complementary tool for public policy. Scholars such as Diener and Seligman (2004) and Kahneman et al. (2004) advocate the creation of national well-being accounts to complement national income accounts. The nation of Bhutan, meanwhile, has introduced the concept of ‘gross national happiness’ to replace gross national product as a measure of national progress.

Despite the potential contributions that happiness research can make to policy, a sound note of caution is necessary in directly applying the findings, both because of the potential biases in survey data and because of the difficulties associated with analysing this kind of data in the absence of controls for unobservable personality traits. In addition, happiness surveys at times yield anomalous results which provide novel insights into human psychology – such as adaptation and coping during economic crises – but do not translate into viable policy recommendations.

One example is the finding that unemployed respondents are happier (or less unhappy) in contexts with higher unemployment rates. The positive effect that reduced stigma has on the well-being of the unemployed seems to outweigh the negative effects of a lower probability of future employment (Clark and Oswald, 1994; Stutzer and Lalive, 2004; and Eggers, Gaddy and Graham, 2006). (Indeed, in Russia even *employed* respondents prefer higher regional unemployment rates. Given the dramatic nature of the late 1990s crisis, respondents may adapt their expectations downwards and are less critical of their own situation when others around them are unemployed.) One interpretation of these results for policy – raising unemployment rates – would obviously be a mistake. At the same time, the research suggests a new focus on the effects of stigma on the welfare of the unemployed.

Happiness economics also opens a field of research questions which still need to be addressed. These include the implications of well-being findings for national indicators and economic growth patterns; the effects of happiness on behaviour such as work effort, consumption, and investment; and the effects on political behaviour. In the case of the latter, surveys of unhappiness or frustration may be useful for gauging the potential for social unrest in various contexts.

In order to answer many of these questions, researchers need more and better quality well-being data, particularly panel data, which allows for the correction of unobserved personality traits and correlated measurement errors, as well as for better determining the direction of causality (for example, from contextual variables like income or health to happiness versus the other way around). These are major challenges in most happiness studies. Hopefully, the combination of better data and increased sophistication in econometric techniques will allow economists to better address these questions in the future.

See Also

- Sen, Amartya
- utilitarianism and economic theory
- wage inequality, changes in

I would like to thank Henry Aaron, Richard Easterlin, Joshua Epstein, Andrew Felton, Andrew Oswald, and George Perry for helpful comments.

Bibliography

Alesina, A., Di Tella, R. and MacCulloch, R. 2004. Inequality and happiness: are Europeans and Americans different? *Journal of Public Economics* 88, 2009–42.

Bertrand, M. and Mullainathan, S. 2001. Do people mean what they say? Implications for subjectivesurvey data. *American Economic Review* 91, 67–72.

Blanchflower, D. and Oswald, A. 2004. Well-being over time in Britain and the USA. *Journal of Public Economics* 88, 1359–87.

Clark, A. and Oswald, A. 1994. Unhappiness and unemployment. *Economic Journal* 104, 648–59.

Diener, E. and Seligman, M. 2004. Beyond money: toward an economy of well-being. *Psychological Science in the Public Interest* 5(1), 1–31.

Diener, E. et al. 2003. The relationship between income and subjective well-being: relative or absolute? *Social Indicators Research* 28, 195–223.

Di Tella, R., MacCulloch, R. and Oswald, A. 2001. Preferences over inflation and unemployment: evidence from surveys of happiness. *American Economic Review* 91, 335–41.

Duesenberry, J. 1949. *Income, Savings, and the Theory of Human Behavior*. Cambridge, MA: Harvard University Press.

Easterlin, R. 1974. Does economic growth improve the human lot? Some empirical evidence. In *Nations and Households in Economic Growth*, ed. P. David and M. Reder. New York: Academic Press.

Easterlin, R. 2003. Explaining happiness. *Proceedings of the National Academy of Sciences* 100(19), 11176–83.

Eggers, A., Gaddy, C. and Graham, C. 2006. Well-being and unemployment in Russia in the 1990s: can society's suffering be individuals' solace? *Journal of Socio-economics*, vol. 35, pp. 209–42.

Frank, R. 1999. *Luxury Fever: Money and Happiness in an Era of Excess*. Princeton, NJ: Princeton University Press.

Frey, B. and Stutzer, A. 2002a. *Happiness and Economics*. Princeton, NJ: Princeton University Press.

Frey, B. and Stutzer, A. 2002b. What can economists learn from happiness research? *Journal of Economic Literature* 40, 401–35.

Gardner, J. and Oswald, A. 2001. Does money buy happiness? A longitudinal study using data on windfalls. Mimeo. Coventry: University of Warwick.

Graham, C. and Felton, A. 2006. Inequality and happiness: insights from Latin America. *Journal of Economic Inequality*, vol. 4, pp. 107–22.

Graham, C., Eggers, A. and Sukhtankar, S. 2004. Does happiness pay? An initial exploration based on panel data from Russia. *Journalof Economic Behavior and Organization* 55, 319–42.

Graham, C. and Pettinato, S. 2002. *Happiness and Hardship: Opportunity and Insecurity in New Market Economies*. Washington, DC: Brookings Institution.

Graham, C. and Sukhtankar, S. 2004. Does economic crisis reduce support for markets and democracy in Latin America? Some evidence from surveys of public opinion and well-being. *Journal of Latin American Studies* 36, 349–77.

Gruber, J. and Mullainathan, S. 2005. Do cigarette taxes make smokers happier? *Advances in Economic Analysis & Policy* 5, 1412.

Helliwell, J. 2004. Well-being and social capital: does suicide pose a puzzle? Working Paper No. 10896. Cambridge, MA: NBER.

Kahneman, D., Diener, E. and Schwarz, N. 1999. *Well-being: The Foundations of Hedonic Psychology*. New York: Russell Sage.

Kahneman, D., Krueger, A., Schkade, D., Schwarz, N. and Stone, A. 2004. Toward national well-being accounts. *AEA Papers and Proceedings* 94, 429–34.

Layard, R. 2005. *Happiness: Lessons from a New Science*. New York: Penguin Press.

Marmot, M. 2004. *The Status Syndrome: How Social Standing Affects Our Health and Longevity*. London: Bloomsbury Press.

Oswald, A. 1997. Happiness and economic performance. *Economic Journal* 107, 1815–31.

Rojas, M. 2004. Well-being and the complexity of poverty. Research Paper No. 2004/29. Helsinki: World Institute for Development Research.

Sen, A. 1995. Rationality and social choice. *American Economic Review* 85, 1–24.

Stutzer, A. and Lalive, R. 2004. The role of social work norms in job searching and subjective well-being. *Journal of the European Economic Association* 2, 696–719.

Van Praag, B. and Ferrer-i-Carbonell, A. 2004. *Happiness Quantified: A Satisfaction Calculus Approach*. Oxford: Oxford University Press.

Veenhoven, R. 2002. World database of happiness. Online. Available at http://www2.eur.nl/fsw/research/happiness/index.htm, accessed 17 June 2005.

How to cite this article

Graham, Carol. "happiness, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000015> doi:10.1057/9780230226203.0702

The New Palgrave Dictionary of Economics Online

Harris, Seymour Edwin (1897 – 1975)

John Kenneth Galbraith

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Federal Reserve System; Harris, S. E.; Keynesianism; New Deal

Article

Harris was born in Brooklyn, New York, and graduated from Harvard University, where he also took his doctorate. His career, apart from the Second World War period and a few post-retirement years at the University of California at La Jolla, was spent at Harvard. In the Second World War, he was in charge of the pricing of exported and imported products and various liaison tasks for the Office of Price Administration. Throughout his life he undertook numerous regional and developmental assignments in New England and was one of the founders of the highly successful Massachusetts community college system.

Harris's early academic work, including a major history of the Federal Reserve System, was competent, orthodox and, as he would later view it, uninspired. Upward progress in his academic career at Harvard was also gradual and unspectacular, a circumstance related at the time to his Jewish origins. In later years he emerged as one of the most highly regarded members of the Cambridge (USA) economic and university community. He became a highly respected chairman of the Harvard economics department, and was the editor of the *Review of Economics and Statistics* and of numerous essay collections by fellow economists. He did not entirely escape criticism from his more relaxed colleagues for his prodigious work and publication schedule. President John F. Kennedy, shortly before he was killed, told of his intention of making Harris his next appointment to the Board of Governors of the Federal Reserve System.

From his earlier orthodox, even conservative, tendencies Harris was released by Keynes and the New Deal. His work came to reflect a strong commitment to Keynesian economics and policy and to the broad welfare measures of the Roosevelt, Kennedy and Johnson years. He was not a compelling writer; in his books, however, this was more than compensated for by the solid competence of his research and preparation, his strongly compassionate views on welfare issues and his very evident desire to extend knowledge on a great range of subject matter. On the economics of health care, education, social

security, international monetary policy, central-bank policy, monetary history and literally a dozen other topics, he provided the basic source material from which legislators learned what could be done, what should be done and how it might be done. A full listing of his works would be among the longest in this Dictionary. Among the prominent later examples are those listed below.

Selected works

1947. *The New Economics: Keynes' Influence on Theory and Public Policy*. New York: A.A. Knopf.

1948a. *How Shall We Pay For Education? Approaches to the Economics of Education*. New York: Harper.

1948b. *Saving American Capitalism: A Liberal Economic Program*. New York: A.A. Knopf.

1949. *The Market for College Graduates, and Related Aspects of Education and Income*. Cambridge, MA: Harvard University Press.

1952. *The Economics of New England: Case Study of an Older Area*. Cambridge, MA: Harvard University Press.

1953. *National Health Insurance, and Alternative Plans for Financing Health*. Foreword by Alfred Baker Lewis. New York: League for Industrial Democracy.

1955. *John Maynard Keynes, Economist and Policy Maker*. New York: Scribner.

1962. *Higher Education: Resources and Finance*. New York: McGraw-Hill.

1964. *The Economics of American Medicine*. New York: Macmillan.

How to cite this article

Galbraith, John Kenneth. "Harris, Seymour Edwin (1897–1975)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000021> doi:10.1057/9780230226203.0703

The New Palgrave Dictionary of Economics Online

Harris – Todaro hypothesis

M. Ali Khan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Harris–Todaro hypothesis replaces the equality of wages by the equality of ‘expected’ wages as the basic equilibrium condition in a segmented, but homogeneous, labour market, and in so doing generates an equilibrium level of urban unemployment when a mechanism for the determination of urban wages is specified. This article reviews work in which the Harris–Todaro hypothesis is embedded in canonical models of trade theory in order to investigate a variety of issues in development economics. These include the desirability (or the lack thereof) of foreign investment, the complications of an informal sector and the presence of clearly identifiable ethnic groups.

Keywords

development economics; gains from trade; Harris–Todaro hypothesis; Harris–Todaro model; Heckscher–Ohlin–Samuelson trade model; immiserizing growth; informal urban sector; international trade theory; Ricardo–Viner model; rural–urban migration; Rybczynski theorem; Stolper–Samuelson theorem; trade unions; underemployment; urban unemployment; urban wage determination; urban wage subsidies; wage differentials

Article

The replacement of the equality of wages by the equality of ‘expected’ wages as the basic equilibrium condition in a segmented, but homogeneous, labour market has proved to be an idea of seminal importance in development economics. Attributed originally to Todaro (1968; 1969) and Harris and Todaro (1970), and commonly referred to as the Harris–Todaro hypothesis, the idea was very much in the air around the late 1960s, as can be seen from the contemporaneous writings of Akerlof and Stiglitz (1969), Blaug, Layard and Woodhall (1969) and Harberger (1971), among others.
The motivation for the Harris–Todaro hypothesis lies in an attempt to explain the persistence of rural to urban migration in the presence of widespread urban unemployment, a pervasive phenomenon in many,

so-called less-developed, countries (but also see Suits, 1985; Partridge and Rickman, 1997). It is natural to ask why such unemployment does not act as a deterrent to further migration. According to the Harris–Todaro hypothesis, the answer lies in the migrant leaving a secure rural wage w_r for a higher expected urban wage w_u^e even though the latter carries with it a non-zero probability of urban unemployment. The expected wage is computed by using the rate of urban employment as an index for the probability of finding a job. Thus

$$w_u^e = w_u \frac{L_u}{L_u + U} + 0 \frac{U}{L_u + U} = w_u \frac{1}{1 + \lambda}, \tag{1}$$

where w_u is the urban wage, L_u is the number of urban employed, U the number of urban unemployed and $\lambda = (U / L_u)$ the rate of urban unemployment. Thus, the Harris–Todaro hypothesis is precisely formulated by the equilibrium condition

$$w_r = w_u^e \Leftrightarrow w_u = w_r(1 + \lambda). \tag{2}$$

Since the Harris–Todaro hypothesis introduces a further unknown, namely, the rate of unemployment, a model in which the hypothesis is embedded must be buttressed by a theory of urban wage determination. The simplest setting is the one originally adopted by Harris–Todaro and subsequently by Bhagwati and Srinivasan (1971; 1973; 1974). This setting assumes the urban wage to be an exogenously given constant and typically rationalizes it as a consequence of government fiat.

In the 1970s, however, several theories of endogenous urban wage determination were simultaneously proposed. Foremost among these is the work of Stiglitz, who provides a microfoundation for the urban wage in terms of labour turnover (Stiglitz, 1974), or in terms of biological efficiency considerations (Stiglitz, 1976). One may also mention in this context the work of Calvo (1978), who sees the equilibrium urban wage as an outcome of trade union behaviour (also Quibria, 1988; Chau and Khan, 2001; and Calvo and Wellisz, 1978, who see a higher urban wage as a consequence of costly supervision). At this stage of the development of the literature, each theory of urban wage determination led to a particular version of the Harris–Todaro model, and the common structural similarities were obscured.

In Khan (1980a), the elementary observation is made that all these variants of the Harris–Todaro model could be studied under one rubric if the Harris–Todaro hypothesis is embedded in the Heckscher–Ohlin–Samuelson (HOS) two-sector, so-called general equilibrium model (see Jones, 1965; Johnson, 1971), and the determination of urban wages is seen in a somewhat more abstract way, that is,

$$w_u = \Omega(w_r, \lambda, R, \tau), \tag{3}$$

where R is the rental on capital and τ a shift parameter. This led to a model whose importance lay, not so much in synthesizing the several variants of urban wage determination, but in emphasizing its points of contact with the trade theory literature. In particular, when (3) collapses to

$$w_u = w_r, \tag{4}$$

that is, when the elasticity of the omega function $\Omega(\cdot)$ with respect to w_r is unity, and those with respect to R and λ are zero, we obtain the HOS model.
This point deserves further articulation. Let a stylized economy consist solely of an urban and a rural sector, indexed by u and r respectively, and be endowed with positive amounts of labour L and capital K. Let the ith sector produce a commodity i in amount X_i in accordance with a production function

$$X_i = F_i(L_i, K_i), \; i = u, r, \tag{5}$$

which is assumed to exhibit constant returns to scale and is twice continuously differentiable and concave. The allocation of labour and capital, L_i and K_i, is determined through marginal productivity pricing. Thus, we have

$$p_r F_r^K = R = p_u F_u^K, \quad p_r F_r^L = w_r \text{ and } p_u F_u^L = w_u, \tag{6}$$

where F_i^j is the derivative of $F_i (i = u, r)$ with respect to $j (j = L_i, K_i)$. The economy is considered too small to influence the positive international prices of the two commodities, p_u and p_r. On rewriting the equilibrium condition (2) in the slightly more general form,

$$w_u = \rho w_r(1 + \lambda); \quad \rho \text{ a shift parameter,} \tag{7}$$

(3), (5), (6) and (7), along with the material balance equations below, complete the specification of the model.

$$K_r + K_u = K \text{ and } L_r + L_u(1 + \lambda) = L. \tag{8}$$

The first point to be noticed about this model is a *decomposability property* whereby the factor prices, w_u, w_r, R and the unemployment rate λ are all independent of the endowments of labour and capital and depend solely on p_u, p_r and the shift parameters τ and ρ. This can be seen most easily if we subsume the marginal productivity conditions (6) into price-equal-unit-cost equations

$$p_i = C_i(w_i, R), \quad i = u, r. \tag{9}$$

This allows one to decompose the model into a subsystem comprising eqs. (7) and (3) along with (9). This basic observation leads to several interesting characteristics of the equilibria of the model. First, the market rural wage and market rental correctly measure the social opportunity cost of labour and capital if we use the international value of GNP as the relevant measure of social welfare. Second, despite the presence of a distorted labour market, there is no possibility of immiserizing growth. Third, an increase in capital (labour) increases the output of the capital- (labour-) intensive commodity provided the of the labour- (capital-) intensive commodity provided the intensities are measured in employment adjusted terms, that is

$$\frac{k_u}{1 + \lambda} = \frac{K_u}{L_u(1 + \lambda)} > (\text{or} <) \ \frac{K_r}{L_r} = k_r. \tag{10}$$

This third property is an analogue of the Rybczynski property of the HOS model. Not surprisingly, we

also obtain an analogue of the Stolper–Samuelson property whereby the effect of changes in international prices on factor returns depends on factor intensities, provided these are now measured in elasticity adjusted terms. The urban sector is said to be capital intensive in elasticity adjusted terms if

$$\theta_{rL}(\theta_{uK}(1 - e_\lambda) + \theta_{uL}e_R) - \theta_{uL}\theta_{rK}(e_w - e_\lambda) > (\text{or} <)0, \tag{11}$$

where θ_{ij} is the share of the jth factor $(j = K, L)$ in the ith sector $(i = u, r)$, and e_i is the elasticity of the $\Omega(\cdot)$ function with respect to the relevant variable. In the setting where e_w equals unity and e_R and e_λ are all zero, (10) and (11) collapse to the conventional physical and value intensities of Magee (1976) and Jones (1971) for the HOS model with proportional wage differentials. Under the further specialization that ρ in (7) equals unity, there is no difference between these two kinds of intensities and a perfect correspondence between the Rybczynski and Stolper–Samuelson theorems.

This reappearance of the divergence of the physical and value intensities of the wage-differential model leads us to inquire into the possibility of downward-sloping supply curves of X_r and X_u. This is indeed a possibility, and a sharp generalization is available in the result that there are perverse price–output responses in the model if and only if the employment-adjusted factor intensities do not conflict with the elasticity-adjusted intensities; see Khan (1980b) for details. Another direct consequence of the decomposability property of the model is a generalization of the Bhagwati (1968), Johnson (1971), Brecher–Alejandro (1977) paradox. This states that capital inflow in the presence of a tariff and with full repatriation of its earnings is immiserizing if and only if the imported commodity is capital intensive in employment-adjusted terms. This result is independent of the various mechanisms for the determination of urban wages; see Khan (1982a) for details, and also subsequent work by Beladi-Naqvi (1988), Grinols (1991), Chao and Yu (1994; 1995c), Chaudhuri and Mukhopadhyay (2002), Chaudhuri (2001) and Sen, Ghosh and Barman (1997). Both of these results have a trade-theoretic flavour, and one question that has remained in the forefront of analytical work on the Harris–Todaro hypothesis relates to the effect of urban wage subsidies on urban unemployment and urban output. (As emphasized above, this question could indeed be seen as the *raison d'être* for the introduction of the hypothesis.) A seminal result here is the Corden–Findlay (1975) paradox, which draws attention to the fact that urban employment and urban output could rise if the urban wage is increased. This question has been readdressed by Neary (1981) and completely resolved in the context of endogenous urban wage determination by Khan (1980b).

So far we have focused on the comparative-static properties of the Harris–Todaro equilibrium. It is also worth emphasizing that the actual existence of the Harris–Todaro equilibrium cannot be taken for granted and must be proved. In the original Harris–Todaro model with an exogenously given rigid wage, equilibrium exists if and only if the rural sector is more capital intensive in employment-adjusted terms; see Khan (1980a) and Basu (1991) for an application of the geometric technique. Furthermore, once the 'isomorphism' with the HOS model is established and understood, one can follow Neary's (1978) lead

and ask for ‘reasonable’ adjustment processes under which the Harris–Todaro equilibrium is locally asymptotically stable. It can be shown that an adjustment process of the Marshallian type leads to a stable equilibrium if and only if the employment-adjusted factor intensities do not conflict with the elasticity-adjusted intensities; see Khan (1980b) for details. Since the elasticity-adjusted intensities of (11) collapse to $\theta_{rL}\theta_{uK}$ in the Harris–Todaro model with a rigid wage, we have the satisfying result that the criteria for the existence of equilibrium and its stability coincide; also see Neary (1981) for this special case.

The entry on this subject in the first (1987) edition of *The New Palgrave: A Dictionary of Economics* was furnished under the title Harris–Todaro hypothesis, and the model presented above referred to as the ‘generalized Harris–Todaro’ (GHT) model. This is somewhat misleading in that any model in which the Harris–Todaro hypothesis is embedded has a justifiable claim to the title of a Harris–Todaro model. Indeed, unlike the case of the HOS model where capital is intersectorally mobile, the hypothesis can be embedded in the Ricardo–Viner model, a setting with three factors, or under an alternative interpretation, one where capital can be viewed as non-shiftable (for details on this and other basic constructions of classical trade theory, see, for example, Caves and Jones, 1985). In many ways, this case of a two-sector model with sector-specific capital is more difficult and also more interesting; see Khan (1982a; 1982b) and Bhatia (2002) for details. And there is at least one example in the literature where a particular Harris–Todaro model has been exported to international trade theory rather than imported from it: Jones and Marjit (1992) investigate a multi-sectoral setting of Khan (1991) by stripping it of the Harris–Todaro hypothesis.

This updated entry would be seriously incomplete if it did not note a criticism of the Harris–Todaro hypothesis centering on the urban unemployed living on a zero wage, and a corresponding generalization of the hypothesis. This criticism also dovetails into an issue that has received increasing attention from sociologists and development economists since the early 1990s: the existence of a dynamic informal urban sector, and the possibility of the urban unemployed being incorporated in it; see Portes et al. (1989) and Fields (1975; 2005b) and their references. This has led to a reformulation of (1) and (2) to

$$w_u^e = w_u\frac{L_u}{L_u + U} + w_i\frac{U}{L_u + U} = \frac{w_u + \lambda w_i}{1 + \lambda} \Rightarrow w_r = w_u^e \Leftrightarrow w_u + \lambda w_i = w_r(1 + \lambda), \tag{12}$$

where w_i is the wage in the informal sector. Again, as in the original Harris–Todaro hypothesis, this generalized hypothesis can be embedded in alternative production structures to yield a variety of models tailored to the purpose the investigator has in mind; see Chandra (1991) and Chandra and Khan (1993) for a more detailed elaboration of this point of view. The subject continues to receive attention; see Stiglitz (1982), Fields (1990; 1997), Rauch (1991), Gupta (1993; 1997a; 1997b), Bandyopadhyay and Gupta (1995), Kar and Marjit (2001), Yabuuchi and Beladi (2001), Yabuuchi, Beladi and Wei (2005); and Chaudhuri (2003).

We conclude this article with a partial list of some other issues in trade and development that have been discussed in the context of urban–rural migration: gains from trade, now depending on the asymmetric nature of the model and on whether the rural or the urban commodity is being exported, as in Khan and Lin (1982), Chao and Yu (1993; 1997; 1999) and Choi and Yu (2006); underemployment or educated unemployment as in Bhagwati and Srinivasan (1977) or in Chaudhuri and Khan (1984) and Chaudhuri and Mukhopadhyay (2003); public inputs as in Chao, Lafargue and Yu (2006); variable returns to scale as in Panagariya and Succar (1986), Beladi (1988) and Choi (1999); growth and technical progress as in Bourguignion (1990), Chau and Yu (1995a) and Chow and Zeng (2001); foreign enclaves as in Gupta-Gupta (1998); capital markets, distorted or otherwise, as in Khan and Naqvi (1983) and Chao and Yu (1992); interaction of ethnic groups as in Khan (1979; 1991) and Khan and Chaudhuri (1985); risk and uncertainty as in Beladi and Ingene (1994); environmental issues, as in Chao, Kerkviliet and Yu (2000) and Chao and Yu (2003); cost–benefit analyses as in Srinivasan and Bhagwati (1975), Stiglitz (1977; 1982), Gupta (1988) and Chao and Yu (1995b); poverty and income inequality as in Moene (1992) and Rauch (1993). In summary then, the Harris–Todaro hypothesis is a versatile and useful analytic instrument for investigating a variety of questions arising in international and development economics where urban unemployment is a prominent issue.

See Also

- development economics
- Heckscher–Ohlin trade theory
- unemployment

Bibliography

Agesa, R. 2000. The incentive for rural to urban migration: A re-examination of the Harris–Todaro model. *Applied Economics Letters* 7, 107–10.

Akerlof, G. and Stiglitz, J.E. 1969. Capital, wages and structural employment. *Economic Journal* 79, 269–81.

Bandyopadhyay, M. and Gupta, M.R. 1995. Development policies in the presence of an informal sector: a note. *Journal of Economics* 61, 301–15.

Basu, A. 1991. Locational choice for free trade zones: a comment. *Journal of Development Economics* 50, 381–7.

Beladi, H. 1988. Variable returns to scale, urban unemployment and welfare. *Southern Economic Journal* 55, 412–23.

Beladi, H. and Naqvi, N. 1988. Urban unemployment and non-immiserizing growth. *Journal of Development Economics* 28, 365–76.

Beladi, H. and Ingene, C.A. 1994. A general equilibrium analysis of rural-urban migration under uncertainty. *Journal of Regional Science* 34, 91–103.

Bhagwati, J.N. 1968. Distortions and immiserizing growth. *Review of Economic Studies* 35, 481–5.

Bhagwati, J.N. and Srinivasan, T.N. 1971. The theory of wage differentials: production response and factor price equalization. *Journal of International Economics* 1, 19–35.

Bhagwati, J.N. and Srinivasan, T.N. 1973. The ranking of policy interventions under factor market imperfections: the case of sector-specific sticky wages and unemployment. *Sankhya*, Series B 35, 405–20.

Bhagwati, J.N. and Srinivasan, T.N. 1974. On reanalyzing the Harris–Todaro model: policy rankings in the case of sector-specific sticky wages. *American Economic Review* 64, 502–8.

Bhagwati, J.N. and Srinivasan, T.N. 1977. Education in a job ladder model and the fairness–in–hiring rule. *Journal of Public Economics* 7, 1–22.

Bhatia, K. 2002. Specific and mobile capital, migration and unemployment in a Harris–Todaro model. *Journal of International Trade and Economic Development* 11, 207–22.

Blaug, M., Layard, P.R.G. and Woodhall, M. 1969. *The Causes of Graduate Unemployment in India*. London: Allen Lane.

Bourguignion, F. 1990. Growth and inequality in a dual model of development: the role of demand factors. *Review of Economic Studies* 64, 502–08.

Brecher, R.A. and Diaz-Alejandro, C.F. 1977. Tariffs, foreign capital and immiserizing growth. *Journal of International Economics* 7, 317–22.

Calvo, G.A. 1978. Urban unemployment and wage determination in LDC's: trade unions in the Harris–Todaro model. *International Economic Review* 19, 65–81.

Calvo, G.A. and Wellisz, S. 1978. Supervision, loss of control and the optimum size of the firm. *Journal of Political Economy* 86, 943–52.

Caves, R.E. and Jones, R.W. 1985. *World Trade and Payments*, 4th edn. Boston: Little, Brown & Co.

Chakravarty, S.R. and Dutta, B. 1990. Migration and welfare. *European Journal of Political Economy* 6, 119–38.

Chandra, V. 1991. The informal sector in developing countries: a theoretical analysis. Ph.D. thesis, Johns Hopkins University.

Chandra, V. and Khan, M. Ali. 1993. Foreign Investment in the presence of an informal sector. *Economica* 60, 79–103.

Chao, C.C. and Yu, E.S.H. 1990. Urban unemployment, terms of trade and welfare. *Southern Economic Journal* 56, 743–51.

Chao, C.C. and Yu, E.S.H. 1992. Capital markets, urban unemployment and land. *Journal of Development Economics* 38, 407–13.

Chao, C.C. and Yu, E.S.H. 1993. Content protection, urban unemployment and welfare. *Canadian Journal of Economics* 26, 481–92.

Chao, C.C. and Yu, E.S.H. 1994. Foreign capital inflows and welfare in an economy with imperfect competition. *Journal of Development Economics* 45, 141–54.

Chao, C.C. and Yu, E.S.H. 1995a. Urban growth, externality and welfare. *Regional Science and Urban Economics* 24, 565–76.

Chao, C.C. and Yu, E.S.H. 1995b. The shadow price of foreign exchange in a dual economy. *Journal of Development Economics* 46, 195–202.

Chao, C.C. and Yu, E.S.H. 1995c. International capital mobility, urban unemployment and welfare. *Southern Economic Journal* 61, 486–92.

Chao, C.C. and Yu, E.S.H. 1997. Trade liberalization in oligopolistic competition with unemployment: a general equilibrium analysis. *Canadian Journal of Economics* 30, 479–96.

Chao, C.C. and Yu, E.S.H. 1999. Export promotion, unemployment and national welfare. *International Economic Journal* 13, 17–34.

Chao, C.C. and Yu, E.S.H. 2003. Jobs, production linkages and the environment: export promotion, unemployment and national welfare. *Journal of Economics* 79, 113–22.

Chao, C.C., Kerkviliet, J. and Yu, E.S.H. 2000. Environmental preservation, sectoral unemployment, and trade in resources. *Review of Development Economics* 4, 39–50.

Chao, C.C., Lafargue, J. and Yu, E.S.H. 2006. Public inputs, urban unemployment and welfare in a developing economy. *Asia Pacific Journal of Accounting and Economics* 13, 141–51.

Chau, N.H. and Khan, M.A. 2001. Optimal urban employment policies: notes on Calvo and Quibria. *International Economic Review* 42, 557–68.

Chaudhuri, S. 2001. Foreign capital inflow, non-traded intermediary, urban unemployment, and welfare in a small open economy. *Pakistan Development Review* 40, 225–35.

Chaudhuri, S. 2003. How and how far to liberalize a developing economy with informal sector and factor market distortions. *Journal of International Trade and Economic Development* 12, 403–28.

Chaudhuri, T.D. and Khan, M.A. 1984. Educated unemployment, educational subsidies and growth. *Pakistan Development Review* 23, 395–409.

Chaudhuri, S. and Mukhopadhyay, U. 2002. Removal of protectionism, foreign investment and welfare in a model of informal sector. *Japan and the World Economy* 14, 101–16.

Chaudhuri, S. and Mukhopadhyay, U. 2003. Free education policy and trade liberalization: consequences on child and adult labour markets in a small open economy. *Journal of Economic Integration* 18, 336–59.

Choi, J. 1999. Factor growth, urban unemployment and welfare under variable returns to scale. *International Economic Journal* 14, 17–34.

Choi, J. and Yu, E.S.H. 1993. Technical progress terms of trade and welfare in a mobile capital Harris–Todaro model. In *Economic Theory and International Trade: Essays in Memoriam of J. Trout Rader*, ed. W. Neuefeind and R. Reizman. New York: Springer-Verlag.

Choi, J. and Yu, E.S.H. 2006. Industrial targeting and non-shiftable capital in the Harris–Todaro model. *Review of International Economics* 14, 1–12.

Chow, Y. and Zeng, J. 2001. Foreign capital in a neoclassical model of growth. *Applied Economics Letters* 8, 613–15.

Corden, W.M. and Findlay, R. 1975. Urban unemployment, intersectional capital mobility and development policy. *Economica* 42, 59–78.

Feldmann, D.H. 1989. The trade-off between GNP and unemployment in a dual economy. *Southern Economic Journal* 56, 46–55.

Fields, G.S. 1975. Rural-urban migration, urban unemployment and job-search activity in LDCs. *Journal of Development Economics* 2, 165–87.

Fields, G.S. 1989. On-the-job search in a labor market model: ex-ante choices and ex-post outcomes. *Journal of Development Economics* 30, 159–78.

Fields, G.S. 1990. Labour market modelling and the urban informal sector: theory and evidence. In *The Informal Sector and Evidence Revisited*, ed. D. Turnham. Paris: OECD.

Fields, G.S. 1997. Wage floors and unemployment: a two-sector analysis. *Labour Economics* 4, 85–92.

Fields, G.S. 2005a. A welfare economic analysis of labor market policies in the Harris–Todaro model. *Journal of Development Economics* 76, 127–46.

Fields, G.S. 2005b. A guide to multisector labor market models. Social Protection Discussion Paper No. 0505. Washington, DC: World Bank.

Grinols, E.L. 1991. Unemployment and foreign capital: the relative opportunity costs of domestic labor and welfare. *Economica* 62, 59–78.

Grossman, G.M. 1983. Partially mobile capital: a general approach to two sector trade theory. *Journal of International Economics* 15, 1–17.

Gupta, M.R. 1988. Migration, welfare, inequality and the shadow-wage. *Oxford Economic Papers* 40, 477–86.

Gupta, M.R. 1993. Rural–urban migration, informal sector and development policies: a theoretical analysis. *Journal of Development Economics* 41, 137–51.

Gupta, M.R. 1995. Tax on foreign capital income and wage subsidy to the urban sector in the Harris–Todaro model. *Journal of Development Economics* 47, 469–79.

Gupta, M.R. 1997a. Foreign capital and the informal sector: comments on Chandra and Khan. *Economica* 64, 353–63.

Gupta, M.R. 1997b. Informal sector and informal capital market in a small open less-developed economy. *Journal of Development Economics* 52, 409–28.

Gupta, K. and Gupta, M.R. 1998. Foreign enclaves and economic development: a theoretical analysis. *Journal of Economics* 67, 317–36.

Harberger, A.C. 1971. On measuring the social opportunity cost of labour. *International Labour Review* 103, 559–79.

Harris, J.R. and Todaro, M. 1970. Migration, unemployment and development: a two sector analysis. *American Economic Review* 40, 126–42.

Jha, R. and Whalley, J.R. 2003. Migration and pollution. Working Paper No. 20034, Department of Economics, University of Western Ontario.

Johnson, H.G. 1971. *The Two-Sector Model of General Equilibrium*. Yrjö Jahnsson Lectures.Chicago: Aldine–Atherton.

Jones, R.W. 1965. The structure of simple general equilibrium models. *Journal of Political Economy* 73, 557–72.

Jones, R.W. 1971. Distortions in factor markets and the general equilibrium model of production. *Journal of Political Economy* 79, 437–59.

Jones, R.W. and Marjit, S. 1992. International trade and endogenous production structures. In *Economic Theory and International Trade: Essays in Memoriam of J. Trout Rader*, ed. W. Neuefeind, and R. Reizman. New York: Springer-Verlag.

Kar, S. and Marjit, S. 2001. Informal sector in general equilibrium: welfare effects of trade policy reforms. *International Review of Economics and Finance* 10, 289–300.

Khan, M. Ali. 1979. A multisectoral model of a small open economy with non-shiftable capital and imperfect labor mobility. *Economic Letters* 2, 369–75.

Khan, M. Ali. 1980a. The Harris–Todaro hypothesis and the Heckscher–Ohlin–Samuelson trade model: a synthesis. *Journal of International Economics* 10, 527–47.

Khan, M. Ali. 1980b. Dynamic stability, wage subsidies and the generalized Harris–Todaro model. *Pakistan Development Review* 19, 1–24.

Khan, M. Ali. 1982a. Social opportunity costs and immiserizing growth: some observations on the long run returns versus and short. *Quarterly Journal of Economics* 96, 353–62.

Khan, M. Ali. 1982b. Tariffs, foreign capital and immiserizing growth with urban unemployment and specific factors of production. *Journal of Development Economics* 10, 245–56.

Khan, M. Ali. 1991. Ethnic groups and the Heckscher–Ohlin–Samuelson trade model. *Economic Theory*

1, 355–71.

Khan, M. Ali. 1992. On measuring the social opportunity cost of labour in the presence of tariffs and an informal sector. *Pakistan Development Review* 31, 535–62.

Khan, M. Ali. 1993. Trade and development in the presence of an informal sector: a four factor model. In *Capital Investment and Development*, ed. K. Basu, M. Majumdar and T. Mitra. Oxford: Basil Blackwell.

Khan, M. Ali. and Chaudhuri, T.D. 1985. Development policies in LDCs with several ethnic groups – a theoretical analysis. *Zeitschrift für Nationalökonomie* 45, 1–19.

Khan, M. Ali. and Lin, P. 1982. Sub-optimal tariff policy and gains from trade with urban unemployment. *Pakistan Development Review* 21, 105–26.

Khan, M. Ali. and Naqvi, S.N.H. 1983. Capital markets and urban unemployment. *Journal of International Economics* 15, 367–85.

Magee, S.P. 1976. *International Trade and Distortions in Factor Markets*. New York and Basle: Marcel-Dekker.

Marjit, S. 1991. Agro-based industry and rural–urban migration: a case for an urban employment subsidy. *Journal of Development Economics* 35, 393–98.

Marjit, S. 2003. Economic reform and informal wage – a general equilibrium analysis. *Journal of Development Economics* 72, 371–78.

Marjit, S. and Beladi, H. 2003. Possibility or impossibility of paradoxes in the small country Harris–Todaro framework. *Journal of Development Economics* 72, 379–85.

Moene, K.O. 1988. A reformulation of the Harris–Todaro mechanism with endogenous wages. *Economics Letters* 27, 387–90.

Moene, K.O. 1992. Poverty and land ownership. *American Economic Review* 82, 52–64.

Neary, J.P. 1978. Dynamic stability and the theory of factor market distortions. *American Economic Review* 68, 672–82.

Neary, J.P. 1981. On the Harris–Todaro model with intersectoral capital mobility. *Economica* 48, 219–34.

Panagariya, A. and Succar, P. 1986. The Harris–Todaro model and economies of scale. *Southern Economic Journal* 52, 986–98.

Partridge, M.D. and Rickman, D.S. 1997. Has the wage-curve nullified the Harris–Todaro model? Further US evidence. *Economics Letters* 54, 277–82.

Portes, A., et al., eds. 1989. *The Informal Economy: Studies in Advanced and Less Developed Countries*. Baltimore, MD: Johns Hopkins University Press.

Quibria, M.G. 1988. The Harris–Todaro model, trade unions and the informal sector: a note on Calvo. *International Economic Review* 29, 557–63.

Rauch, J. 1991. Modeling the informal sector formally. *Journal of Development Economics* 35, 33–47.

Rauch, J. 1993. Economic development, urban underemployment and income inequality. *Canadian Journal of Economics* 26, 901–18.

Sato, Y. 2004. Migration, frictional unemployment, and welfare-improving labor policies. *Journal of Regional Science* 44, 773–93.

Sen, P., Ghosh, A. and Barman, A. 1997. The possibility of welfare gains with capital inflows in a small tariff-ridden economy. *Economica* 64, 345–52.

Srinivasan, T.N. and Bhagwati, J. 1975. Alternative policy rankings in a large open economy with sector-specific minimum wages. *Journal of Economic Theory* 11, 356–71.

Srinivasan, T.N. and Bhagwati, J. 1978. Shadow prices for project selection in the presences of distortions: effective rates of protection and domestic resource costs. *Journal of Political Economy* 86, 91–116.

Stiglitz, J.E. 1974. Alternative theories of wage determination and unemployment in LDC's: the labor-turnover model. *Quarterly Journal of Economics* 88, 194–227.

Stiglitz, J.E. 1976. The efficiency wage hypothesis, surplus labor, and the distribution of income in the LDCs. *Oxford Economic Papers* 28, 185–207.

Stiglitz, J.E. 1977. Some further remarks on cost–benefit analysis. In *Project Evaluation*, ed. H. Schwartz and R. Berney. Washington, DC: Inter-American Development Bank.

Stiglitz, J.E. 1982. The structure of labor markets and shadow prices in LDCs. In *Migration and the Labor Market in Developing Countries*, ed. R.H. Sabot. Boulder, CO: Westview Press.

Suits, D.B. 1985. US farm migration: an application of the Harris–Todaro model. *Economic Development and Cultural Change* 34, 815–28.

Todaro, M.P. 1968. An analysis of industrialization: employment and unemployment in LDCs. *Yale Economic Essays* 8, 329–492.

Todaro, M.P. 1969. A model of labor migration and urban unemployment in less developed countries. *American Economic Review* 59, 138–48.

Yabuuchi, S. and Beladi, H. 2001. Urban unemployment, informal sector and development policies. *Journal of Economics* 74, 301–14.

Yabuuchi, S., Beladi, H. and Wei, G. 2005. Foreign investment, urban unemployment, and informal sector. *Journal of Economic Integration* 20, 123–138.

How to cite this article

Khan, M. Ali. "Harris–Todaro hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000022> doi:10.1057/9780230226203.0704

The New Palgrave Dictionary of Economics Online

Harrod, Roy Forbes (1900 – 1978)

Walter Eltis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

acceleration principle; cost-push inflation; Domar, E. D.; dynamic theory; effective demand; employment multiplier; firm, theory of; Harrod, R. F.; Harrod–Domar growth model; Harrod-neutral technical progress; imperfect competition; inflation; international finance institutions; Keynesianism; knife-edge; marginal revenue curve; market share; monopolistic competition; natural rate of growth; neoclassical growth theory; warranted rate of growth

Article

Roy Harrod was born in February 1900 and died in 1978. His father, Henry Dawes Harrod, was a businessman and author of two historical monographs. His mother, Frances (née Forbes-Robertson) was a novelist, and sister of the notable Shakespearean actor-manager, Sir Johnson Forbes-Robertson. Henry Harrod's business failed in 1907, but Roy won a scholarship to St Paul's School in 1911 and a King's Scholarship to Westminster in 1913. He became Head of his House, and in 1918 won a scholarship in history to New College, Oxford, his father's college. He enlisted in September 1918 and was commissioned in the Royal Field Artillery, but the war ended before his training was completed.
He went up to Oxford in early 1919 and first read Literae Humaniores (Classical Literature, Ancient History and Philosophy). He might well have devoted his career to academic philosophy, and he valued his publications in that subject more highly than his seminal contributions to economics. He has remarked that significant economic problems have only attracted the attention of profound thinkers for about 200 years, and interest in them might well disappear in another 200. In contrast, deep thought has been devoted to the great philosophical problems (such as the validity of inductive methods of thought) for more than 2,000 years and new contributions will be read for so long as civilized life remains. But his philosophy tutor at New College, H.W.B. Joseph, deterred him from devoting his life to that subject, by reacting extremely negatively to his essays. Harrod has left an account of a seminar on Einstein's theory of relativity in Oxford in 1922 where Joseph drew attention to a few terminological problems and believed this had undermined the theory. Einstein's theory of relativity survived, but Harrod was persuaded not to pursue a career in academic philosophy. In later years he published in the distinguished

philosophical journal, *Mind*, and his *Foundations of Inductive Logic* (1956a) has received serious critical attention from philosophers as distinguished as A.J. Ayer (1970), but his main scholarly work was not to be in philosophy.

He followed his first class honours in Literae Humaniores in 1922 with a first class in modern history just one year later, and in 1923 Christ Church, Oxford, elected him to a Tutorial Fellowship (confusingly described as a studentship in that college) to teach the novel subject, economics, which was to be part of Oxford's new Honour School of Politics, Philosophy and Economics.

Harrod was allowed two terms away from Oxford so that he could learn enough economics to teach it, and it was suggested that he might spend this time in Europe, but he first went to Cambridge where he attended a wide range of lectures and wrote weekly essays on money and international trade for John Maynard Keynes. He was equally fortunate when he returned to Oxford, for while he was critically discussing the economics essays of Christ Church's undergraduates he was himself writing weekly microeconomic essays for the Drummond Professor of Political Economy, Francis Ysidro Edgeworth.

In addition to his new academic work Harrod took a notable part in the administration of his college (where he was Senior Censor in 1929–31, the most responsible office a student of Christ Church can be called upon to discharge), and also the university where he was elected to Oxford's governing body (the Hebdomadal Council) in 1929 before he was 30. In the university and in Christ Church, he fought powerful campaigns on behalf of Professor Lindemann (subsequently Lord Cherwell) who held Oxford's Chair of Experimental Philosophy (Physics), and became principal scientific adviser to Winston Churchill's wartime government and a member of his post-war cabinet.

By 1930 his economics had developed to the point where he was able to publish his first important and original contribution, ‘Notes on Supply’, in which he was the first 20th-century economist to derive the marginal revenue curve. This should have appeared in 1928 to produce a claim for international priority, but Keynes, the editor of the *Economic Journal*, sent the article to Frank Ramsey, who first believed there were difficulties with the argument. He subsequently appreciated that his objections rested on a misunderstanding, but Harrod's new contribution was less startling in 1930 than it would have been in 1928. He followed this initial contribution to the imperfect competition literature with an important article, ‘Doctrines of Imperfect Competition’ (1934), in which he summarized the essential elements of the new theories of Edward Chamberlin and Joan Robinson.

During the 1930s Harrod frequently stayed with Keynes and he was increasingly drawn into the group of brilliant young economists which included Richard Kahn and Joan Robinson, who were helping him develop the new theories which culminated in *The General Theory of Employment, Interest and Money*. Harrod had written a number of important and influential articles in the press advocating new reflationary policies in the early 1930s, and these together with his extension of Kahn's employment multiplier to international trade in his *International Economics* (1933b) prompted Joseph A. Schumpeter to write in 1946 in his obituary article on Keynes, ‘Mr Harrod may have been moving independently toward a goal not far from that of Keynes, though he unselfishly joined the latter's standard after it had been raised’.

Shortly after the *General Theory* appeared, Harrod published *The Trade Cycle* (1936a) in which he developed some of the dynamic implications of the new theory of effective demand. The conditions where output would grow were a central theme in Adam Smith's *The Nature and Causes of the Wealth of Nations*, and it had been much analysed in the great 19th-century contributions of Malthus, Ricardo, Mill and Marx, but the long-term dynamic implications of immediate changes to particular economic

variables received virtually no attention in the neoclassical work that followed the marginal revolution. In the *General Theory* Keynes mostly went no further than to work through completely the immediate effects *on a formerly stationary economy* of a variety of disturbances such as an excess of the saving which would occur at full employment over the investment businessmen considered it prudent to undertake. Harrod went a vital step further and showed what could be expected to occur if saving was *permanently high* in relation to *the long-term opportunity to invest*. In 1939 he followed *The Trade Cycle* with ‘An Essay in Dynamic Theory’ (1939c), and after the war he developed his growth theory further in the book, *Towards a Dynamic Economics* (1948a). Important articles followed including a ‘Second Essay in Dynamic Theory’ (1960a), and ‘Are Monetary and Fiscal Policies Enough?’ (1964a). It is almost certainly because of Harrod's rediscovery of growth theory in the 1930s and his notable contributions to it that Assar Lindbeck, the Chairman of the Nobel Prize Committee, chose to state that he was among those who would have been awarded a Nobel Prize in economics if he had lived a little longer. The nature of Harrod's original contribution and the gradual evolution of his theory from 1939 to 1964 are set out in the second part of this article. The detailed technical characteristics of Harrod's growth model are the subject of Eltis (1987).

In the Second World War Harrod's friendship with Lindemann and his increasing distinction as an economist led to an invitation to join the Statistical Department of the Admiralty (S Branch) which Churchill set up when he again became First Lord in 1939. This moved to Downing Street when Churchill became Prime Minister in 1940, but Harrod did not have a particular talent for detailed statistical work and he developed an increasing interest in the international financial institutions, the International Monetary Fund and the World Bank, which would need to be set up as soon as the war was won, and from 1942 onwards he pursued this work in Christ Church. In the immediate post-war years he took a strong interest in national politics, and stood for Parliament unsuccessfully as a Liberal in the general election of 1945 and for a time he was a member of that party's Shadow Cabinet. He had served on Labour Party committees before the war, and in the 1950s with Churchill's support he unsuccessfully sought adoption as a Conservative parliamentary candidate: his economic advice was warmly welcomed by Harold Macmillan, Conservative Prime Minister in 1957–63. Harrod received the honour of knighthood in 1959 in recognition of his public standing and his notable academic achievements in the pre-war and post-war decades.

He had succeeded Keynes as editor of the *Economic Journal* in 1945, and in partnership with Austin Robinson (who looked after the book reviews) he sustained its reputation and quality until his retirement from the editorship in 1966.

His own post-war academic work included important contributions in three areas. In addition to the continuing development and refinement of his pre-war work on dynamic theory, he published extensively on the theory of the firm and on international monetary theory which had been his particular concern during the war.

The Oxford Economists’ Research Group had begun to meet prominent British industrialists before the war. A group of Oxford economists which generally included Harrod invited individual industrialists to dine in Oxford, and after dinner they were questioned extensively on the considerations which actually influenced their decisions. This led to the publication of a number of much cited articles and the book, *Oxford Studies in the Price Mechanism* (Wilson and Andrews, 1951) to which Harrod himself did not contribute. Propositions which emanated from these dinners included the notion that businessmen took little account of the rate of interest in their investment decisions, and that they did not seek to profit

maximize, but priced instead by adding a margin they considered satisfactory to their average or ‘full’ costs of production. In his important articles, ‘Price and Cost in Entrepreneurs’ Policy’ (1939b) and ‘Theories of Imperfect Competition Revised’ (1952a), Harrod set out a theoretical account of how firms price in which industrialists follow something like these procedures. Their object is especially to achieve a high market share, and by setting prices low enough to deter new entry they actually succeed in maximizing their long-run profits and avoid the excess capacity that Chamberlin and Joan Robinson had considered an inevitable consequence of monopolistic or imperfect competition. This attempt to reconcile the ‘rules of thumb’ that the businessmen revealed with the propositions of traditional theory was more highly regarded outside Oxford than some of the books and articles in the new tradition.

His work on the world's international monetary problems occupied a good deal of his time and attention in the post-war decades. Keynes himself had considered the breakdown in international monetary relations a crucial element in the collapse of effective demand in so many countries in the 1930s, and he devoted much of the last years of his life to the creation of new institutions which would avoid a repetition of these disasters. Harrod believed he was continuing this vital work when he devoted much thought and energy to these questions. He arrived at the conclusion that there was bound to be some inflation in a world which was successfully pursuing Keynesian policies, and that the liquidity base of the world's financial system was bound to become inadequate if the price of gold failed to rise with other prices. He believed that underlying world liquidity which rested on gold in the last resort must be allowed to rise in line with the international demand for money. He therefore came to focus on the price of gold, and in his book, *Reforming the World's Money* (1965), he proposed that a substantial increase in the price of gold would be needed if subsequent international monetary crises were to be avoided. Harry Johnson (1970) has summarized his contribution to this debate.

Harrod took a great interest in actual developments in the United Kingdom economy, and published seven books and collections of articles in the first two post-war decades which were directly concerned with the policies Britain should follow. There was in addition an immense range of articles in the academic journals, the bank reviews and the press on these questions, not to mention monthly stockbrokers’ letters for Phillips and Drew. Harrod argued strongly and powerfully that nothing was to be gained by running the economy below full employment, which meant an unemployment rate of less than two per cent in the 1950s and the 1960s. In the late 1950s he was deeply concerned that the removal of import controls would render it increasingly difficult for Britain to pursue such Keynesian policies, and he was a vigorous opponent of European Common Market entry. He attached more significance than some distinguished Keynesians to holding down inflation but he published statistics in *Towards a New Economic Policy* (1967a) to show that in Britain this had tended to be faster when the economy was in recession than when output was allowed to expand. He argued therefore that deflationary policies could play no useful role in policies to control the rate of cost inflation, which he considered the essential element in inflation in Britain. Policy swung sharply away from this Keynesian tradition in the last years of his life, and he wrote a final letter to *The Times* on 21 July 1976 in which he praised the economics of Tony Benn and Peter Shore for their opposition to the Labour government's public expenditure cuts, for, ‘To cut public spending when there is an undesirably high rate of unemployment is crazy’.

His advocacy of import controls and his adverse reaction to deflationary policies at all times might suggest that he was an economist of the Left, but his willingness to support each of the British political parties at various times underlines how his approach to economic and social problems cannot be

typecast. The lines of policy he supported always followed directly from his understanding of the significance of the major interrelationships, and it was his belief that Keynesian theory (which he had so notably helped to refine and develop) provided the appropriate tools for the analysis of Britain's economic problems that led him towards the expansionist policies he so consistently advocated. But further theoretical and empirical relationships which he believed were equally well founded led him to advocate a series of social policies to which very right-wing labels can be attached.

Just before the 1959 election his article, ‘Why I Shall Vote Conservative’, in *The Sunday Times*, put forward the startlingly unfashionable argument that only the Conservatives would allow more money to go to the better off who had most to contribute to the future of Britain. Harrod's strong belief in the importance of the *quality* of the country's population stock (which, he held, mattered no less than the physical capital stock) lay behind this article. Harrod thought the quality of the population would be bound to deteriorate if the middle classes continued to have fewer children than the poor. He was a strong believer in the inheritance of every kind of ability, and a provocative conversational conclusion he drew was that in an ideal world one-third of Christ Church's much sought-after undergraduate places should be sold to the rich. Their children often had insufficient academic ability to perform well in examinations, but they had inherited abilities of other kinds which would take them to the highest positions, so they should go to Oxford first. Harrod's reasoning on the inheritance of ability and its implications is set out in detail in the Memorandum he submitted to the Royal Commission on Population in 1944. There he suggested that a difficulty in finding servants was one reason why the middle classes had fewer children. Among his suggestions to remedy this state of affairs was that Diplomas in Domestic Service should be established, and that it should become common practice for servants to have latch-keys and the same rights as their mistresses to enjoy social lives with no questions asked. His Memorandum reads strangely nowadays when it is widely regarded as unacceptable that any practical conclusions may be drawn from the proposition that human abilities are inherited. Harrod never hesitated to carry his arguments to their limits, and he always went where his reasoning took him, irrespective of the predictable reactions of others.

The unselfconsciousness of both his academic and his public writing comes out especially in his two biographical volumes, the official life of Keynes (commissioned by the executors) which he published in 1951 and *The Prof* (1959a), his personal sketch of Lord Cherwell. As well as providing magnificent accounts of their subjects from the standpoint of one who had known them intimately (and who profoundly understood the economic problems Keynes wrestled with), these books contain extensive autobiographical passages which will enable later generations to know more of Harrod than any biographer can begin to convey.

He ceased to lecture in Oxford in 1967 upon reaching the statutory retirement age of 67, but as a Visiting Professor he continued to teach in several distinguished North American Universities. He died in his Norfolk home in 1978 eleven years after his Oxford work came to an end.

Harrod's revival of growth theory and his contribution to Keynesian macroeconomics

Harrod was intimately involved in the origins and development of Keynesian economics. As the galley proofs of the *General Theory* emerged from the printers from June 1935 onwards, copies were sent to Harrod, to Kahn and to Joan Robinson and with their assistance, Keynes rewrote extensively for final publication. Harrod helped to clarify the relationship between Keynes's new theory of the rate of interest

and the then ruling neoclassical theory where this depended upon the intersection of *ex ante* saving and investment schedules. In the course of their correspondence, Harrod showed Keynes how well he understood the essence of the *General Theory* by setting out its novelty and its principal elements in ten lines on 30 August 1935:

Your view, as I understand it is broadly this:-

Volume of investment determined by	marginal efficiency of capital schedule rate of interest
Rate of investment determined by	liquidity preference schedule quantity of money
Volume of employment determined by	volume of investment multiplier
Value of multiplier determined by	propensity to save

Keynes responded, ‘I absolve you completely of misunderstanding my theory. It could not be stated better than on the first page of your letter.’

Almost immediately after the appearance of the *General Theory*, Harrod published *The Trade Cycle* (1936a) which contained for the first time in the Keynesian literature the concept of an economy growing at a steady rate. Keynes wrote of it to Joan Robinson on 25 March 1937, ‘I think he has got hold of some good and important ideas. But, if I am right, there is one fatal mistake’, and to Harrod himself on March 31, ‘I think that your theory in the form in which you finally enunciate it is not correct, being fatally affected by a logical slip in the argument.’ Harrod replied devastatingly on April 6th, ‘There is no slip … The fact is that you in your criticism are still thinking of once over changes and that is what I regard as a static problem. My technique relates to steady growth.’ Harrod's slip was in fact the first step towards the reinstatement of growth theory into mainstream economic analysis.

Harrod convinced Keynes, who on 12 April congratulated him for ‘having invented so interesting a theory’, but with the reservation, ‘I should doubt whether any reader who has not talked or corresponded with you could be aware that the whole of the last half of the book was intended to be in relation to a moving base of steady progress.’ Keynes added that it was vital that Harrod carry his ideas further and restate them more comprehensibly.

Harrod made important progress in the next 15 months, and on 3 August 1938 he sent Keynes a preliminary draft of the article, ‘An Essay in Dynamic Theory’, and wrote in his accompanying letter,

> my re-statement of the dynamic theory … is, I think, a great improvement on my book … I have been throwing out hints in a number of places of the possibility of formulating a simple law of growth and I want to substantiate the claim. It is largely based on the ideas

> of the general theory of employment; but I think it gets us a step forward.

A lengthy correspondence then developed between Harrod and Keynes in which the two most original elements in Harrod's contribution which later excited much interest and controversy in the economics profession were extensively discussed.

Harrod's principal innovation was the invention of a *moving equilibrium growth path* for the economy, and he described this as the ‘warranted’ line of growth. Harrod had perceived before he wrote *The Trade Cycle* that there was a fundamental contradiction between the assumptions prevalent in the microeconomic theory of the firm and industry, to which he had made notable contributions, and the new Keynesian macroeconomics. In the theory of the firm, long-term investment was zero, for firms had no motivation to undertake further investment once they were in long-period equilibrium. But the new Keynesian macroeconomics required that there be net investment by firms or the government whenever there was any net saving in the macroeconomy. A theory compatible with both macro and microeconomic equilibrium therefore required that firms invest all the time, so that they can continually absorb total net saving. Harrod's formulation of the warranted rate of growth, his novel discovery, was an attempt to set out this necessary equilibrium growth path that industrial and commercial investment decisions must all the time follow in order to achieve a complete economic equilibrium.

Harrod's moving equilibrium or warranted growth path required that saving (of s per cent of the national income) be continually absorbed into investment, so he asked the question: at what rate of growth will firms all the time choose to invest the s per cent of the national income, which equilibrium growth requires? To answer this question, he made use of the acceleration principle or ‘the relation’, as he called it, that firms need say C_r units of additional capital to produce an extra unit of output. It follows from these premises that the warranted rate of growth of output will be s/C_r per cent per annum. Since each rise in output by 1 unit entails that C_r extra units be invested, a rise in output by s/C_r per cent of the national income will call for an equilibrium investment of C_r times this, which is precisely s per cent of the national income, the ratio of *ex ante* saving in the national income. In Harrod's examples at this time, he suggested a typical s of 10 per cent of the national income and a C_r of 4, to produce a warranted rate of growth of 2.5 per cent.

This idea that if there is continual saving, then equilibrium entails a continual geometric growth in production came as a considerable surprise to Keynes and the other members of the ‘circus’. As Harrod had already explained in April 1937,

> The static system provides an analysis of what happens where there is no increase [in output] which entails (as in Joan Robinson's long-period analysis) that saving=0. Now I was on the lookout for a steady rate of advance, in which the rates of increase would be mutually consistent.

But Harrod's second discovery had equally radical implications. Suppose the actual growth of output is marginally above the equilibrium or warranted rate of growth. In Harrod's numerical example with s 10 per cent and C_r 4, it can be supposed that output actually grows 0.1 per cent faster than the warranted rate, that is by 2.6 per cent instead of 2.5 per cent. Then with 2.6 per cent output growth, the acceleration principle or relation will entail that 4 times 2.6 per cent be added to the capital stock, so that *ex ante*

investment is 10.4 per cent of the national income. With *ex ante* saving limited to 10.0 per cent, the 0.1 per cent excess of actual growth over warranted growth then produces an excess in *ex ante* investment over *ex ante* saving of 0.4 per cent of the national income. Any excess in *ex ante* investment over *ex ante* saving will be associated with extra expansion of the national income according to the economics of the *General Theory*. Thus, if the actual rate of growth exceeds the warranted rate of s/C_r per cent, the tendency will be for actual growth to rise and rise, for as soon as actual growth rises from 2.6 to say 3 per cent, required investment will rise further to 4 times 3 per cent which equals 12 per cent and so exceed the 10 per cent savings ratio by a still greater margin. Conversely, when actual growth comes out at a rate just short of the warranted 2.5 per cent, *ex ante* investment will be below the 10 per cent savings ratio, which will cause the rate of growth to decline. This second discovery, which became known as Harrod's knife-edge, was therefore that any rate of growth in excess of the equilibrium or warranted path he had discovered would set off a continual acceleration of growth, while any shortfall would set off deceleration. He wrote to Keynes of this discovery on 7 September 1938:

> If in static theory producers produce too little, they will be well satisfied with the price they get and feel happy; but this is not taken to be the *right* amount of output; they will be stimulated to produce more. The equilibrium output is taken to be that which *just* satisfies them and induces them to go on as before. Similarly the warranted rate [of growth] is that which just satisfies them and leaves them going on as before. The difference between the warranted rate and the old equilibrium (i.e. the difference between dynamic and static theory) is, in my view, that if they produce above the warranted rate, they will be more than satisfied and be stimulated, and conversely, while in the case of equilibrium in static conditions the opposite happens. The ‘field’ round the [static] equilibrium contains centripetal, that round the warranted centrifugal forces.

It took Keynes time to absorb Harrod's startling discovery. On 19 September he proposed a counter-example in which C_r was merely one-tenth, while s was also one-tenth. With this counter-example, a deviation of output by a small amount from the warranted path, say by δx, which would raise planned investment above the level at which it would otherwise be by $C_r \delta x$ would merely raise this by $0.10\ \delta x$, which would equal the rise in planned saving of $s \delta x$, which would also come to $0.10\ \delta x$, so there would be no tendency towards an explosive growth in effective demand. This would grow explosively if C_r was one-ninth (in which case planned investment would rise by $0.11\ \delta x$ and saving by only $0.10\ \delta x$) but the further growth of output would be damped if C_r was merely one-eleventh, so, Keynes insisted, ‘neutral, stable or unstable equilibrium’ are equally likely.

Harrod protested on 22 September, ‘it is absurd to suppose extra capital required [C_r] only $\frac{1}{10}$ of annual output, when the capital required in association with the pre-existent level of incomes in England today is 4 or 5 times annual output’. The probability that C_r would exceed s so that *ex ante* investment would rise by more than *ex ante* saving in order to produce instability was therefore overwhelming.

But several qualifications emerged. In comparing the increase in *ex ante* investment to the increase in *ex ante* saving following a small deviation of output from the warranted rate:

1. 1. The relevant marginal capital coefficient (C_r) which determines how much planned investment will rise is the net new requirement of *induced* investment. In so far as investment decisions are autonomous of short-term fluctuations in output, the relevant C_r will be lower than the economy's overall capital output ratio.
2. 2. The relevant coefficient which determines the increase in planned saving is the *marginal* and not the average propensity to save. Planned saving will rise more where output deviates upward from the warranted rate, the greater is the marginal propensity to save in relation to the average propensity.

The circumstances that could produce a stable upward deviation of growth from the warranted rate and the avoidance of Harrod's knife-edge are therefore a very high marginal propensity to save in combination with a situation where most investment is autonomous so that the induced investment coefficient, C_r, is considerably less than 1. In ‘An Essay in Dynamic Theory’, Harrod covered this possibility with the caveat, ‘when long-range capital outlay is taken into account … the attainment of a neutral or stable equilibrium of advance may not be altogether improbable in certain phases of the cycle’. The possibility he had in mind here is that in the early stages of a cyclical recovery there may be so much excess industrial capacity that C_r will be quite low for a time, and therefore quite possibly lower than the marginal propensity to save. But in general any deviation of growth from the warranted line of advance would raise *ex ante* investment by a greater margin than *ex ante* saving with the result that the rate of growth would deviate further.

In addition to establishing the existence of the warranted line of advance and its instability, Harrod had to define the equilibrium investment behaviour by businesses which would actually lead to expansion at the requisite rate. In his 1939 article he omitted to offer any behavioural rule but simply asserted that the warranted rate was ‘that rate of growth which, if it occurs, will leave all parties satisfied that they have produced neither more nor less than the right amount’. That is no more than a description of equilibrium growth, and much the same can be said of his definition of the warranted rate in *Towards a Dynamic Economics* (1948a) as ‘that over-all rate of advance which, if executed, will leave entrepreneurs in a state of mind in which they are prepared to carry on a similar advance’. It was only in the article ‘Supplement on Dynamic Theory’ (1952b) that Harrod arrived at a behavioural assumption that matched his algebraic formulation of the warranted rate:

> Let the representative entrepreneur on each occasion of giving an order repeat the amount contained in his order for the last equivalent period, adding thereto an order for an amount by which he judges his existing stock to be deficient, if he judges it to be deficient, or subtracting therefrom the amount by which he judges his stock to be redundant, if he does so judge it.

With that assumption an economy which once achieves growth at the warranted rate will sustain it, while any upward or downward deviations will lead to still greater deviations wherever C_r exceeds the marginal propensity to save.

But it emerged by 1964, when Harrod published ‘Are Monetary and Fiscal Policies Enough?’, that even that assumption fails to define growth at the warranted rate, for it must also be assumed that the

representative entrepreneur will expand at a rate of precisely s/C_r when he judges his capital to be neither deficient nor redundant. This requires an expectation by the representative entrepreneur that his market will grow at a rate of precisely s/C_r. Hence the full requirement for growth along Harrod's warranted equilibrium path is that entrepreneurs expect growth at this rate and expand and continue to expand at that rate so long as their capital stock continues to grow in line with their market so that it is neither deficient nor redundant. They will of course increase their rate of expansion if their capital should prove deficient, and curtail it if part of their stock becomes redundant.

The warranted rate of growth and its instability were Harrod's great innovations. From 1939 onwards he contrasted this equilibrium rate with the natural rate of growth, ‘the rate of advance which the increase of population and technological improvements allow’, which was entirely independent of the warranted rate. Harrod defined the rate of technical progress more precisely in 1948 as the increase in labour productivity ‘which, at a constant rate of interest, does not disturb the value of the capital coefficient’. This then entered the language of economics as Harrod-neutral technical progress, which, together with growth in the labour force, determines the natural rate of growth, that is, the rate at which output can actually be increased in the long run. This raised few theoretical problems in 1939, and there was nothing novel in the proposition that long-term growth must depend on the rate of increase of the labour force and technical progress. Keynes himself had said as much several years earlier in ‘Economic Possibilities for our Grandchildren’ (1930). But the contrast between this natural rate and Harrod's innovatory warranted rate offered entirely new insights.

If the warranted rate exceeds the feasible natural rate, the achievement of equilibrium growth must be impractical because the economy cannot continue to grow faster than the natural rate. It must deviate downwards from the warranted rate towards the natural rate far more than it deviates upwards with the result that ‘we must expect the economy to be prevailingly depressed’. If the natural rate is greater, output will tend to deviate upwards towards the natural rate with the result that the economy should enjoy ‘a recurrent tendency towards boom conditions’.

Keynes's own reaction to the dichotomy between the warranted and natural rates was characteristically (his letter to Harrod on 26 September 1938) that the warranted rate always exceeded the natural:

> In actual conditions … I suspect the difficulty is, not that a rate in excess of the warranted is unstable, but that the warranted rate itself is so high that with private risk-taking no one dares to attain it …
>
> I doubt if, in fact, the warranted rate – let alone an unstable excess beyond the warranted – has ever been reached in USA and UK since the war, except perhaps in 1920 in UK and 1928 in USA. With a stationary population, peace and unequal incomes, the warranted rate sets a pace which a private risk-taking economy cannot normally reach and can never maintain.

That is characteristic Keynes, but Harrod had persuaded him to express his familiar analysis in the language of his new theory of growth. In the immediate post-war decades when full employment and creeping inflation prevailed, it was widely argued that the natural rate had come to exceed the warranted. The richness of Harrod's model is demonstrated by its ability to illuminate both kinds of situation.

Evsey Domar's growth model which has a good deal in common with Harrod's was published seven

years after ‘An Essay in Dynamic Theory’, and a considerable literature emerged in the next 15 years on the stability conditions and other important features of what came to be known as the Harrod–Domar growth model. This is elegantly summarized by Frank Hahn and Robin Matthews in their celebrated 1964 survey article.

The development of neoclassical growth theory in the 1950s led to an increasing realization that the warranted and natural growth rates could be equated by an appropriate rate of interest. If the warranted rate was excessive so that over-saving led to slump conditions, a lower interest rate which raised C_r sufficiently would bring it down to the natural rate. Conversely the inflationary pressures that resulted from an insufficient warranted rate would be eliminated if higher interest rates reduced C_r sufficiently. If the real rate of interest and C_r responded in this helpful way, s/C_r, the warranted rate could always be brought into equality with the natural rate.

Harrod's response included his ‘Second Essay in Dynamic Theory’ (1960a), a title which underlines its significance. He proposed that there was an optimum real rate of interest r_n which would maximize utility, with a value of G_p/e, G_p being the economy's long-term rate of growth of labour productivity and e the elasticity of the total utility derived from real per capita incomes with respect to increases in these. If a one per cent increase in real per capita incomes raises per capita utility 0.5 per cent, e will be 0.5, and r_n the optimum rate of interest which maximizes utility will be $G_p/0.5$, viz. twice the rate of growth of labour productivity. If the marginal utility of income does not fall at all as real per capita incomes rise, per capita utility will grow one per cent when incomes rise one per cent so that e is unity, and r_n equals G_p. The more steeply the marginal utility of incomes fall, the more e will fall below unity, and the more the optimum real rate of interest, G_p/e, will exceed the rate of growth of labour productivity.

If a society actually seeks to establish the optimum rate of interest determined in this kind of way, the value of C_r will depend upon this optimum rate of interest, so it will not also be possible to use the rate of interest to equate the natural and warranted rates of growth in the manner the neoclassical growth models of, for instance, Robert Solow (1956) and Trevor Swan (1956) propose. There will therefore still be difficulties because the warranted rate of growth with real interest rates at their optimum level will not in general be equal to the natural rate. Therefore, as Harrod suggested in the final articles he published in 1960 and 1964, governments will have to run persistent budget deficits or surpluses if they are to avoid the difficulties inherent in discrepancies between the natural and the warranted rates of growth.

So Harrod remained a convinced Keynesian who continued to believe that a long-term imbalance between saving, the main determinant of the warranted rate, and investment opportunity would call for persistent government intervention. When that approach to economic policy again becomes fashionable, economists may learn a good deal from Harrod's later articles which have not yet received the same attention from the economics profession as his seminal work in the 1930s and the 1940s.

Selected works

The ‘Bibliography of the Works of Sir Roy Harrod’, in *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, ed. W.A. Eltis, M.FG. Scott and J.N. Wolfe, Oxford: Oxford University Press, 1970, includes all the articles he published in books, journals and magazines from 1928 to 1969, and some of

his most influential newspaper articles. The present list of selected works is confined to his books and academic articles in books and journals.

1930a. Notes on supply. *Economic Journal* 40, 232–41.

1930b. Progressive taxation and equal sacrifice. *Economic Journal* 40, 704–7.

1931. The law of decreasing costs. *Economic Journal* 41, 566–76. Addendum: 42, 490–2.

1933a. A further note on decreasing costs. *Economic Journal* 43, 337–41.

1933b. *International Economics*. Cambridge: Cambridge University Press. 1st revised edn, 1939; 2nd revised edn, 1957; 3rd revised edn, mainly rewritten, 1974.

1934a. Professor Pigou's theory of unemployment. *Economic Journal* 44, 19–32.

1934b. Doctrines of imperfect competition. *Quarterly Journal of Economics* 48, 442–70.

1934c. The equilibrium of duopoly. *Economic Journal* 44, 335–7.

1934d. The expansion of credit in an advancing economy. *Economica* NS 1, 287–99. Rejoinders:1, 476–8; 2 (1935), 82–4.

1936a. *The Trade Cycle: An Essay*. Oxford: Oxford University Press.

1936b. Utilitarianism revised. *Mind* 45, 137–56.

1936c. Imperfect competition and the trade cycle. *Review of Economics and Statistics* 18(May), 84–8.

1937. Mr Keynes and traditional theory. *Econometrica* 5, 74–86.

1938. Scope and method of economics. *Economic Journal* 48, 383–412.

1939a. Modern population trends. *Manchester School of Economics and Social Studies* 10(1), 1–20. Rejoinder: April (1940), 47–58.

1939b. Price and cost in entrepreneurs’ policy. *Oxford Economic Papers* 2(May), 1–11.

1939c. An essay in dynamic theory. *Economic Journal* 49, 14–33. Errata: 49, 377.

1939d. Value and Capital by J.R. Hicks. *Economic Journal* 49, 294–300.

1942. Memory. *Mind* 51, 47–68.

1943. Full employment and security of livelihood. *Economic Journal* 53, 321–42.

1945. Memorandum to the *Royal Commission on Equal Pay for Men and Women*. Appendix IX in the Fourth Volume of Memoranda of Evidence. London: HMSO.

1946. *A Page of British Folly*. London: Macmillan.

1946b. Price Flexibility and Employment. By Oscar Lange. *Economic Journal* 56, 102–7.

1946b. Professor Hayek on individualism. *Economic Journal* 56, 435–42.

1947a. A comment on R. Triffin's ‘National Central Banking and The International Economy’. *Review of Economic Studies* 14(2), 95–7.

1947b. *Are These Hardships Necessary?* London: Rupert Hart-Davis.

1948a. *Towards a Dynamic Economics: Some Recent Developments of Economic Theory and Their Application to Policy*. London and New York: Macmillan.

1948b. The economic consequences of atomic energy. In *The Atomic Age*. Sir Halley Stewart Lectures. London: Allen & Unwin.

1948c. The fall in consumption. *Bulletin of the Oxford University Institute of Statistics* 10, 162–7. Rejoinder: 10, 235–44; 10, 290–3.

1950. Memoranda (Submitted in August and December 1944). *Papers of the Royal Commission on Population*, vol. 5. London: HMSO.

1951a. *The Life of John Maynard Keynes*. London and New York: Macmillan.

1951b. Notes on trade cycle theory. *Economic Journal* 61, 261–75.

1951c. *And So It Goes On: Further Thoughts on Present Mismanagement*. London: Rupert Hart-Davis.

1952a. Theory of imperfect competition revised. In *Economic Essays*, ed. R.F. Harrod. London and New York: Macmillan.

1952b. Supplement on dynamic theory. In *Economic Essays*, ed. R.F. Harrod. London and New York:

Macmillan.

1952c. Currency appreciation as an anti-inflationary device: comment. *Quarterly Journal of Economics* 66, 102–16.

1952d. *Economic Essays*. London and New York: Macmillan.

1952e. *The Pound Sterling*. Princeton Essays in International Finance No. 13. Princeton: Princeton University Press.

1953a. *The Dollar*. London and New York: Macmillan. 2nd edn with new introduction, New York: The Norton Library, 1963.

1953b. Imbalance of international payments. *International Monetary Fund Staff Papers* 3(April), 1–46.

1953c. Foreign exchange rates: a comment. *Economic Journal* 63, 294–8.

1953d. Sir Hubert Henderson, 1890–1952. *Oxford Economic Papers* NS 5(supplement), June, 59–64.

1953e. Full capacity vs. full employment growth: a comment on Pilvin. *Quarterly Journal of Economics* 67(November), 553–9.

1955. Les relations entre l'investissement et la population. *Revue économique* 6, 356–67.

1956a. *Foundations of Inductive Logic*. London and New York: Macmillan.

1956b. The British boom, 1954–55. *Economic Journal* 66, 1–16.

1956c. Walras: a re-appraisal. *Economic Journal* 66, 307–16.

1957a. The Common Market in perspective. *Bulletin of the Oxford Institute of Statistics* 19, 51–5.

1957b. Review of *International Economic Policy* by J.E. Meade. *Economic Journal* 67, 290–95.

1957c. Clive Bell on Keynes. *Economic Journal* 67, 692–9.

1958a. *The Pound Sterling, 1951–58*. Princeton Essays in International Finance. Princeton: Princeton University Press.

1958b. The role of gold today. *South African Journal of Economics* 26(March), 3–13. Rejoinder: 27, March, 16–22.

1958c. *Policy against Inflation*. London and New York: Macmillan.

1958d. Questions for a stabilization policy in primary producing countries. *Kyklos* 11(2), 207–11.

1958e. Factor-price relations under free trade. *Economic Journal* 68, 245–55.

1959a. *The Prof: A Personal Memoir of Lord Cherwell*. London: Macmillan.

1959b. Domar and dynamic economics. *Economic Journal* 69, 451–64.

1959c. Inflation and investment in underdeveloped countries. In *Economi, Politik, Samhälle: en bok Tillagnad Bertil Ohlin*, ed. J. Bergvall. Stockholm: Bokförlaget Folk och Samhälle.

1960a. Second essay in dynamic theory. *Economic Journal* 70, 277–93. Comment: 70, 851. Rejoinder: 72, 1009–10.

1960a. New arguments for induction: reply to Professor Popper. *British Journal for the Philosophy of Science* 10(40), February, 309–12.

1960b. Keynes' attitude to compulsory military service. *Economic Journal* 70, 166–7.

1960c. Evidence submitted to the Radcliffe Committee on the Working of the Monetary System, May 1958. *Principal Memoranda of Evidence*, vol. 3. London: HMSO.

1961. The dollar problem and the gold question. In *The Dollar in Crisis*, ed. S.E. Harris. New York: Harcourt, Brace and World.

1961a. *Topical Comment: Essays in Dynamic Economics Applied*. London: Macmillan; New York: St Martin's Press.

1961b. The general structure of inductive argument. *Proceedings of the Aristotelian Society, 1960–61*, 41–56.

1961c. Real balances: a further comment. *Economic Journal* 71, 165–6.

1961d. A plan for increasing liquidity: a critique. *Economica* NS 28, 195–202.

1961e. The 'neutrality' of improvements. *Economic Journal* 71, 300–4.

1961f. Review of Sraffa's Production of Commodities by Means of Commodities. *Economic Journal* 71,

783–7.

1962a. Economic development and Asian regional cooperation. *Pakistan Development Review* 2, 1–22.

1962b. Dynamic theory and planning. *Kyklos* 15(3), 68–79.

1963. *The British Economy*. New York: McGraw Hill.

1963. Themes in dynamic theory. *Economic Journal* 73(September), 401–21. Corrigendum: December 1963, 792.

1963a. Desirable international movements of capital in relation to growth of borrowers and lenders and growth of markets. In *International Trade Theory in a Developing World*, ed. R.F. Harrod and D.C. Hague. London and New York: Macmillan.

1963b. Liquidity. In *World Monetary Reform*, ed. H.C. Grubel. Stanford: Stanford University Press.

1964a. Are monetary and fiscal policies enough? *Economic Journal* 74(December), 903–15.

1964b. *Plan to Increase International Monetary Liquidity*. Brussels and London: European League for Economic Co-operation.

1964c. Comparative analysis of policy instruments. In *Inflation and Growth in Latin America*, ed. W. Baer and I. Kerstenetzky. Homewood, IL: Richard Irvin.

1964d. Retrospect on Keynes. In *Keynes' General Theory*, ed. R. Lekachman. New York and London: Macmillan.

1965. *Reforming the World's Money*. London: Macmillan. New York: St Martin's Press.

1966a. International liquidity. *Scottish Journal of Political Economy* 13, 189–204.

1966b. Optimum investment for growth. In *Problems of Economic Dynamics and Planning: Essays in Honour of Michael Kalecki*. Oxford: Pergamon Press.

1967a. *Towards a New Economic Policy*. Manchester: Manchester University Press.

1967b. Increasing returns. In *Monopolistic Competition Theory: Studies in Impact: Essays in Honour of Edward H. Chamberlin*, ed. R.E. Kuenne. New York: Wiley.

1967c. Methods of securing equilibrium. *Kyklos* 20(1), 24–33.

1967d. World reserves and international liquidity. *South African Journal of Economics* 35(June), 91–103.

1967e. Assessing the trade returns. *Economic Journal* 77, 499–511.

1968. What is a model? In *Value, Capital and Growth: Papers in Honour of Sir John Hicks*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

1969. *Money*. London: Macmillan; New York: St Martin's Press.

1970a. *Sociology, Morals and Mystery*. Chichele Lectures, All Souls College, Oxford. London: Macmillan.

1970b. Reassessment of Keynes's views on money. *Journal of Political Economy* 78, 617–25.

1970c. Replacements, net investment, amortisation funds. *Economic Journal* 80, 24–31.

1972. Imperfect competition, aggregate demand and inflation. *Economic Journal* 82, 392–401.

1973. *Economic Dynamics*. London: Macmillan; New York: St Martin's Press.

Bibliography

Ayer, A.J. 1970. Has Harrod answered Hume? In *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F.G. Scott and J.N. Wolfe. Oxford: Oxford University Press.

Blake, R. 1970. A personal memoir. In *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F.G. Scott and J.N. Wolfe. Oxford: Oxford University Press.

Domar, E. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14, 137–47.

Domar, E. 1947. Expansion and employment. *American Economic Review* 37, 34–55.

Eltis, W. 1987. Harrod–Domar growth model. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Hahn, F.H. and Matthews, R.C.O. 1964. The theory of economic growth: a survey. *Economic Journal* 74, 779–902.

Johnson, H.G. 1970. Roy Harrod on the price of gold. In *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F.G. Scott and J.N. Wolfe. Oxford: Oxford University

Press.

Keynes, J.M. 1930. Economic possibilities for our grandchildren. In *The Collected Writings of John Maynard Keynes*, Vol. IX: *Essays in Persuasion*, London: Macmillan, 1972.

Keynes, J.M. 1973. *The General Theory and After* (Correspondence and Articles). Vols XIII and XIV of *The Collected Writings of John Maynard Keynes*, London: Macmillan.

Lindbeck, A. 1985. The Prize in Economic Science in memory of Alfred Nobel. *Journal of Economic Literature* 23, 37–56.

Phelps-Brown, H. 1980. Sir Roy Harrod: a biographical memoir. *Economic Journal* 90, 1–33.

Schumpeter, J.A. 1946. John Maynard Keynes 1883–1946. *American Economic Review* 36, 495–518.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Wilson, T. and Andrews, P.W.S. 1951. *Oxford Studies in the Price Mechanism*. Oxford: Oxford University Press.

Bibliographic addendum

Besomi, D. 1999. *The Making of Harrod's Dynamics*. London: Macmillan. New York: St. Martin's Press.

Rampa, G., Stella, L. and Thirlwall, A., eds. 1998. *Economic Dynamics, Trade and Growth: Essays on Harrodian Themes*. London: Macmillan. New York: St. Martin's Press.

Young, W. 1989. *Harrod and his Trade Cycle Group: The Origins and Development of the Growth Research Programme*. London: Macmillan. New York: St. Martin's Press.

How to cite this article

Eltis, Walter. "Harrod, Roy Forbes (1900–1978)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000023> doi:10.1057/9780230226203.0705

The New Palgrave Dictionary of Economics Online

Harsanyi, John C. (1920 – 2000)

Roger B. Myerson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

John Harsanyi worked to extend the general theoretical framework of economic analysis. He established the modern basis for utilitarian ethics. He developed a general bargaining solution to that included the Nash bargaining solution and the Shapley value as special cases. He became a leading advocate of non-cooperative game theory as the general framework for analysis of social interactions among rational individuals. He developed the tracing procedure to select among multiple equilibria of games. He showed how to interpret mixed-strategy equilibria in game theory. His general model of Bayesian games with incomplete information became a cornerstone of information economics.

Keywords

bargaining; Bayesian decision theory; Bayesian game; cardinal utility; cooperative game theory; game theory; Harsanyi, J.; incomplete information in game theory; information economics; interpersonal utility comparisons; Morgenstern, O.; Nash bargaining solution; Nash equilibrium; Neumann–Morgenstern utility axioms; non-cooperative game theory; private information; rationality; rule utilitarianism; Schelling, T.; Shapley value; von Neumann, J.; welfare economics

Article

John Harsanyi extended the theoretical framework of economic analysis with major contributions to game theory and welfare economics. His general approach to social theory was based on a fundamental assumption that people are rational decision-makers who share a basic understanding of the things that they value in the world. His personal experiences made him profoundly sceptical of theories that try to justify social systems from other assumptions, without respecting the values, the rationality, and the intelligence of all individuals in society. He understood that social institutions and policies should be evaluated by carefully analysing their impact on individuals' welfare. From his training in philosophy he appreciated the basic importance of general unified frameworks in social theory. He recognized the

foundations of such a framework in Bayesian decision theory, with its compelling axiomatic characterizations. So he devoted his career to the development of a general framework for economic analysis based on these principles. His best-known contribution is the general model of Bayesian games with incomplete information, which became a cornerstone of information economics.

Harsanyi grew up in Budapest, Hungary, where his distinction as a student was marked in 1937 by his winning the first prize in Hungary's national mathematics competition. But at the university he chose to study pharmacy so that he could share his father's business, as other options were then clouded by the threat of war. He was forced into hiding by Nazi racial policies during the last months of the German occupation. After the war, he studied philosophy and earned a doctorate from the University of Budapest in 1947, but his intellectual independence led to political difficulties with the Communist regime, which forced him out of the university. In 1950 he fled Hungary and found refuge in Australia.

He began studying economics at Sydney University, earning an MA in 1953. He then held a lectureship at the University of Queensland, where he began to read game theory. In 1956, when he already had published articles on welfare economics, he enrolled as a student at Stanford University, earning his second doctorate in 1959. He then held faculty positions at the Australian National University, Wayne State University, and, from 1965, in the School of Business Administration at the University of California, Berkeley.

In his early contributions to welfare economics, Harsanyi established the modern basis for utilitarian ethics. Von Neumann and Morgenstern (1947) had shown axiomatically that a rational individual should choose among risky alternatives by maximizing the expected value of a cardinal utility function, but some economists doubted whether this cardinal utility, defined for individual risk analysis, had any relevance for social welfare analysis. Harsanyi (1953) argued that, in ethical decision-making, to avoid any dependence on our particular roles in society, we must imagine ourselves in an initial position before social roles have been assigned, when we could only anticipate getting the role of someone drawn at random from the whole population. Thus, ethical decision-making involves an essential element of risk, and we naturally get a social welfare function equal to the average utility of all members of society. This average requires interpersonally comparable utility scales, assessed by sympathetically comparing the prospect of being in one person's position or another's. Harsanyi argued, as his ‘similarity postulate’, that such comparable utilities for all individuals can be generated by a common utility function, based on shared human values, once the factors that cause apparent differences among individuals’ tastes are included as parameters of the function. Harsanyi (1955) showed that, even without this similarity postulate, the Neumann–Morgenstern utility axioms (applied to individual and social decision-making) and the Pareto welfare axiom (that social preferences should be consistent with any unanimity of individual preferences) together imply that social utility can be defined only as some linear function of individual utility values.

In his later work on welfare economics, Harsanyi (1977a, ch. 4; 1977b; 1977c), argued that ethical analysis should be used to evaluate general social rules or institutions rather than specific acts. That is, we may consider ethical rules that prescribe people's behaviour in a wide range of situations, recognizing that behaviour in other situations could be determined by self-interest according to some Nash equilibrium. Then, as rule utilitarians, we should advocate rules that yield the highest average of expected utilities for all individuals.

Harsanyi began working on cooperative game theory in the mid-1950s, when many different cooperative solution theories were being studied. But most of these theories could yield multiple solutions or no

solutions for a game, or could not even be defined without some special structures like transferable utility. Harsanyi's view of the field was clarified by his insistence that a good solution concept should yield one well-defined solution to any game. At that time, there were only two cooperative solution concepts that yielded unique solutions to broad classes of games: the Shapley (1953) value for games with transferable utility, and the Nash (1950) bargaining solution for two-person games without transferable utility. Harsanyi (1956) showed that the Nash bargaining solution could be derived from an earlier theory of Zeuthen (1930). Then Harsanyi (1963) developed a general bargaining solution that included the Nash bargaining solution and the Shapley value as special cases.

In the mid-1960s, Harsanyi shifted from cooperative to non-cooperative game theory. The basic definition of non-cooperative equilibrium had been introduced by Nash (1951). But there was little further development of non-cooperative theory until Schelling (1960) analysed bargaining processes as games with multiple equilibria, where any cultural or environmental factor that focuses the players' attention on one equilibrium can become a self-fulfilling prophecy. Harsanyi (1961) argued that the distribution of power that is measured by a cooperative solution could be the focal factor that selects among the many non-cooperative equilibria of a bargaining game. But then Harsanyi began to recognize the force of Nash's early arguments for the greater generality of the non-cooperative approach, which is based on a precise specification of each player's individual decision problem, which is lacking in cooperative models. Thus Harsanyi became a leading advocate of non-cooperative game theory.

Harsanyi understood that the non-cooperative approach could not become a standard methodology for applied economic analysis without some refinements of Nash's equilibrium concept, because it can yield very large sets of equilibria for many games. So he began a search for theoretical criteria to select among multiple equilibria, which culminated in his book with Selten (1988). Their selection theory is based on Harsanyi's (1975) tracing procedure, which can select a unique equilibrium from a given initial hypothesis about the players' strategic behaviour. For each number t between 0 and 1, we define a t-auxiliary game that differs from the original game in that each player thinks that the other players have probability $1 - t$ of behaving according to the initial hypothesis; otherwise, with probability t, they choose their strategies rationally. The tracing procedure finds a continuous path of equilibria for these auxiliary games, starting from the trivial 0-auxiliary game and ending at a unique equilibrium of the original game when $t=1$.

Harsanyi's work on incomplete information in games began (1962) with the problems of extending Nash's bargaining solution to situations where players do not know each others' payoffs. In this work, he began to recognize the problems of modelling players' beliefs about each others' beliefs in a game. Harsanyi (1967–8) confronted these modelling problems at the most general and fundamental level, showing how the basic definition of normal-form games should be modified to analyse situations where individuals have different information.

The early development of game theory was based on von Neumann's (1928) argument that any dynamic game in extensive form can be represented by a conceptually simpler one-stage game in normal form. In this normal-form game, each player chooses a strategy that is a complete contingent plan of action, specifying what the player would do at each stage of the dynamic game as a function depending on any information that the player might learn during the game. In normal-form analysis, we assume that the players choose their strategies simultaneously and independently at the start of the game, before anyone gets any private information, and thereafter their behaviour in the dynamic game can be determined

mechanically by their strategies. Thus, questions about the players' private information are suppressed in normal-form analysis.

Harsanyi (1967–8) showed how to correct this deficiency by developing a more general game model that allows players to have different initial information, without losing the analytical simplicity of the normal form. Each player's private information at the start of the game is represented by a random variable that is called the player's *type*. Harsanyi defined a *Bayesian game* to be a mathematical model that specifies (*a*) the set of players, (*b*) the set of feasible actions for each player, (*c*) the set of possible types for each player, (*d*) each player's expected payoff for every possible combination of all players' actions and types, and (*e*), for each possible type of each player, a probability distribution over the other players' possible types, which describes what each type of each player would believe about the others' types.

The beliefs in a Bayesian game are said to be *consistent* if the players' type-contingent beliefs can all be derived by Bayes's rule from some common prior distribution over types. Although not analytically essential, this assumption of consistent beliefs has been regularly used in applied economic analysis, because it allows that differences in players' beliefs may be explained by different previous experiences.

To represent dynamic extensive-form games by games in Bayesian form, each player's action in a Bayesian game may be interpreted as a plan that describes what the player would do in any situation after the beginning of the game, as a function of what the player may learn during the game. A player's strategy, in von Neumann's original sense, would then be a function that specifies a feasible action for each of the player's possible types. But each player is assumed to know his type already when the game begins, and so Harsanyi worked to avoid the fiction of strategic decision-making by players who have not yet learned their types. It would be better, he argued, to imagine that a player's different possible types correspond to different agents, one of whom will be randomly selected to be active in the game. The point is that each player's optimal decisions will maximize his conditional expected payoff given his actual type, and there is no significance to any expected value that is not conditioned on such type information.

Harsanyi emphasized that games must be analysed from the perspective of someone who only knows the information common to all players, which is summarized in the Bayesian game model. Game-theoretic analysis requires us to deny ourselves any knowledge of any player's actual type, so that we can appreciate the uncertainty of the other players who do not know it. The actual type of each player, being private information, must be treated as an unknown quantity or random variable in our analysis. So an equilibrium of a Bayesian game specifies a feasible action for every possible type of every player, such that the specified action for each type of each player maximizes his conditional expected payoff, given his type, given his beliefs about the others' types, and given the type-contingent actions of the other players according to this equilibrium.

Applications of Bayesian games developed quickly. Harsanyi and Selten (1972) defined a generalization of Nash's bargaining solution for Bayesian games, where players have incomplete information about each other. By embedding normal-form games in the larger space of Bayesian games, Harsanyi (1973) showed how to interpret mixed-strategy equilibria in non-cooperative game theory. Such equilibria had seemed to imply paradoxically that rational players should base their decisions on randomizing devices like roulette wheels, but this apparent paradox was a consequence of the normal-form assumption that players choose strategies before they get any private information. By letting each player have some minor private information that changes payoffs only slightly, Harsanyi could transform any mixed-strategy equilibrium into a Bayesian equilibrium where each type chooses an optimal action without

randomization.

Harsanyi's Bayesian games have become the standard economic model for analysing transactions among individuals who have different information. Before 1967, the lack of a general framework for informational problems had inhibited economic inquiry about markets where people do not share the same information. The unity and scope of modern information economics were found in Harsanyi's framework.

See Also

- expected utility hypothesis
- interpersonal utility comparisons (new developments)
- Nash program
- Savage's subjective expected utility model
- Shapley value
- utilitarianism and economic theory
- von Neumann, John

Selected works

1953. Cardinal utility in welfare economics and the theory of risk-taking. *Journal of Political Economy* 61, 434–5.

1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

1956. Approaches to the bargaining problem before and after the theory of games. *Econometrica* 24, 144–57.

1961. On the rationality postulates underlying the theory of cooperative games. *Journal of Conflict Resolution* 5, 179–96.

1962. Bargaining in ignorance of the opponent's utility function, *Journal of Conflict Resolution* 6, 29–38.

1963. A simplified bargaining model for the n-person cooperative game. *International Economic Review* 4, 194–220.

1967–8. Games with incomplete information played by Bayesian players. *Management Science* 14, 159–82, 320–34, 486–502.

1972. (With R. Selten.) A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science* 18(5), 80–106.

1973. Games with randomly disturbed payoffs, a new rationale for mixed strategy equilibrium points. *International Journal of Game Theory* 2, 1–23.

1975. The tracing procedure, a Bayesian approach to defining a solution for n-person noncooperative games. *International Journal of Game Theory* 4, 61–94.

1977a. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. New York: Cambridge University Press.

1977b. Rule utilitarianism and decision theory. *Erkenntnis* 11, 25–53.

1977c. Morality and the theory of rational behavior. *Social Research* 44, 623–56.

1988. (With R. Selten.) *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Bibliography

John C. Harsanyi Memorial Issue. 2001. *Games and Economic Behavior* 36(1), 1–56.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Nash, J. 1951. Non-cooperativegames. *Annals of Mathematics* 54, 286–95.

Neumann, J. von. 1928. Zur Theories der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320.

Neumann, J. von. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*. 2nd edn. Princeton, NJ: Princeton University Press.

Schelling, T. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Shapley, L. 1953. A value for n-person games. In *Contributions to the Theory of Games*, vol. 2, ed. H. Kuhn and A. Tucker.Princeton, NJ: Princeton University Press.

Weymark, J. 1995. John Harsanyi's contributions to social choice and welfare economics. *Social Choice and Welfare* 12, 313–18.

Zeuthen, F. 1930. *Problems of Monopoly and Economic Warfare*. London: Routledge.

How to cite this article

Myerson, Roger B. "Harsanyi, John C. (1920–2000)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000158> doi:10.1057/9780230226203.0706

The New Palgrave Dictionary of Economics Online

Hart, Albert Gailord (1909 – 1997)

Peter Earl

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

decision theory; expectations; Hart, A. G.; uncertainty

Article

Born in Oak Park, Illinois, Hart received his BA from Harvard in 1930 and his Ph.D. from the University of Chicago in 1936. Most of his career – from 1946 until his retirement in 1979 – was spent as Professor of Economics at Columbia University. Much of his noteworthy work concerned the implications of uncertainty for policymakers, but he should also be remembered as having worked with Kaldor and Tinbergen (1964) to produce an ingenious proposal for a commodity reserve currency: this would serve to improve international liquidity simultaneously with providing a means of protecting incomes of primary producers against shrinkage in times of depression.

Hart's work on uncertainty included a monograph (1940), one notable feature of which was an attempt to analyse how decision makers can judge their success or failure, and thence reformulate their expectations, in the light of partial knowledge of performance distributions. From 1936 onwards, he emphasized the rationality, in situations of uncertainty, of choosing flexible production technologies which, though they might not be perfectly adapted to any specific output rate, would not be disastrously expensive to run over a range of outputs. This idea, which was also promoted by his Chicago contemporary Stigler (1939), led Hart to be critical of much writing on decision theory. He felt it misleading to theorize as if firms assign probabilities to rival hypothetical outputs, aggregate these weighted values and then build their plans around the weighted average of probable output rates (1942). Hart was also irritated by Keynes's tendency to speak of expectations in terms of certainty equivalents, and he warned that, ‘generally speaking, the business policy appropriate to a complex of uncertain anticipations is different in kind from that appropriate for any set of certain expectations’ (1947, p. 422). Hart carried this theme into work critical of deterministic macroeconomic model-building and fiscal policy formulation (1945), and into a distinctive approach to monetary theory (1948, especially part II). In the latter, he introduced the ‘margin of safety’ motive for holding liquid assets, arguing that the structure of economic affairs is such that risks are usually linked: a single disappointment is prone to

cause many other things to go wrong in consequence. Hart's concern with surprise, flexibility, and structural linkages in many ways foreshadows themes that emerged in the 1980s in the business policy literature on scenario planning and strategic choices. However, he is not usually credited as the pioneer of this kind of thinking: having been largely ignored by mainstream writers, his ideas were sufficiently poorly known to end up being reinvented.

Selected works

1940. *Anticipations, Uncertainty and Dynamic Planning*. Chicago: University of Chicago Press.

1942. Risk, uncertainty and the unprofitability of compounding probabilities. In *Studies in Mathematical Economics and Econometrics*, ed. O. Lange, F. McIntyre and T.O. Yntema. Chicago: University of Chicago Press.

1945. ‘Model-building’ and fiscal policy. *American Economic Review* 35, 531–58.

1947. Keynes's analysis of expectations and uncertainty. In *The New Economics: Keynes's Influence on Theory and Public Policy*, ed. S.E. Harris. New York: Knopf.

1948. *Money, Debt and Economic Activity*. New York: Prentice-Hall.

1964. (With N. Kaldor and J. Tinbergen.) The case for an international commodity reserve currency. Paper submitted to UNCTAD, Geneva, March–June 1964. In *Essays on Economic Policy*, vol. 2, ed. N. Kaldor. London: Duckworth.

Bibliography

Stigler, G.J. 1939. Production and distribution in the short run. *Journal of Political Economy* 47, 305–27.

How to cite this article

Earl, Peter. "Hart, Albert Gailord (1909–1997)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000026> doi:10.1057/9780230226203.0707

The New Palgrave Dictionary of Economics Online

Hawkins – Simon conditions

Hukukane Nikaido

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

In a Leontief system of interindustrial input–output relationships consisting of n sectors of industry, each of which produces a single good, without joint products, under constant returns to scale, and using n goods as input in fixed proportions, the balance of demand for and supply of goods is represented by a system of linear equations

$$x_i = \sum_{j=1}^{n} a_{ij}x_j + c_i, \ (i = 1, 2, \ldots, n),$$

where a_{ij} are non-negative input coefficients of the jth sector, x_j is the level of output of the jth sector and c_i is the level of final demand for the ith good (i, j=1,…, n).
With the input coefficient matrix A having a_{ij} in the ith row and the jth column, the output vector x having x_j in the jth component, and the final demand vector c having c_i in the ith component the system is represented in matrix form by the equation

$$x = Ax + c.$$

The system is productive enough to give positive net output over input, if x_j non-negative units of output of the jth sector (j=1,…, n) are produced to meet a bill of positive final demand c_i (i=1,…, n).
The *productivity* of the system, which is equivalent to the condition that the n-dimensional square matrix $I - A$, where I is the identity matrix, have an inverse matrix $(I - A)^{-1}$ having all the elements non-

negative, hinges on and is completely determined by the magnitudes of the input coefficients. A necessary and sufficient condition for such productivity, stated in terms of inequalities constraining the magnitudes of the input coefficients and referred to as the Hawkins-Simon conditions, after the names of its discoverers (Hawkins and Simon, 1949), is that all the principal minor determinants of the matrix *I-A* be positive. This is equivalent to the seemingly weaker conditions that the *n* principal minor determinants located in the ascending order on the upper left corner of the matrix *I-A* be positive

$$\Delta_k = \begin{vmatrix} 1-a_{11} & -a_{12} & \dots & -a_{1k} \\ -a_{21} & 1-a_{22} & \dots & -a_{2k} \\ \dots & \dots & \dots & \dots \\ -a_{k1} & -a_{k2} & \dots & 1-a_{kk} \end{vmatrix} > 0,\ (k = 1, \dots, n)\ .$$

As a mathematical result the equivalence of the Hawkins–Simon conditions to productivity is very easy to prove, as can readily be shown by transforming the equation (*I*–*A*) *x*=*c* through Gaussian elimination to a triangular form

$$b_{11}x_1 + b_{12}x_2 + \dots + b_{1n}x_n = d_1 b_{22}x_2 + \dots + b_{2n}x_n = d_2 \dots\dots\dots\dots\dots\dots\dots\dots b_{nn}x_n = d_n,$$

where

$$b_{ij} \leqq 0 (i < j),\ d_i \geqq 0 (i = 1, \dots, n)$$

and

$$\Delta_k = b_{11}b_{22}\dots b_{kk} (k = 1, \dots, n).$$

Since the Hawkins-Simon conditions ensure the productivity of the system, they are a primary prerequisite for the Leontief system, and enlarged systems involving it as a built-in subsystem, to be well-behaved. They also make the Leontief system dynamically well-behaved. In the multiplier process over discrete time,

$$x_i(t+1) = \sum_{j=1}^{n} a_{ij}x_j(t) + c_i, \quad (i = 1, \ldots, n)$$

the solution converges to the equilibrium output levels supplying net output equal to the final demand c_i (i=l, …, n), if and only if the Hawkins–Simon conditions are satisfied. This stability is equivalent to the convergence of the matrix geometric progression

$$I + A + A^2 + \ldots + A^t + \ldots$$

to the inverse matrix $(I - A)^{-1}$. In the multiplier process over continuous time,

$$\mathrm{d}x_i / \mathrm{d}t = \alpha_i \left(\sum_{j=1}^{n} a_{ij}x_j + c_i - x_i \right), \quad (i = 1, \ldots, n)$$

the Hawkins–Simon conditions are necessary and sufficient as well for the convergence of the solution to the same equilibrium output levels, which is equivalent to the condition that the real parts of all the eigenvalues of the matrix A-I be negative.

See Also

- linear models
- Perron–Frobenius theorem

Bibliography

Hawkins, D. and Simon, H.A. 1949. Note: some conditions of macroeconomic stability. *Econometrica* 17(July–October), 245–8.

How to cite this article

Nikaido, Hukukane. "Hawkins–Simon conditions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000027> doi:10.1057/9780230226203.0708

The New Palgrave Dictionary of Economics Online

Hawley, Frederick Barnard (1843 – 1929)

Mauro Boianovsky

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Frederick Barnard Hawley (1843–1929) advanced the ‘risk theory of profit’: profit is the reward entrepreneurs get to relieve the other productive factors from risk in competitive conditions. The normal rate of profit is determined by the expectation of profit that just covers the marginal entrepreneur's subjective valuation of risk. The current rate of profit will converge to its normal value because of the operation of the ‘readjustment period’, when income contraction brings about a fall in aggregate supply larger than the reduction in aggregate demand. This is explained by Hawley's concept of the consumption function.

Keywords

consumers’ expenditure; entrepreneurship; Hawley, F.; Keynes, J. M.; Knight, F.; paradox of thrift; Patinkin, D.; readjustment period; risk aversion; risk theory of profit; uncertainty; underemployment equilibria

Article

Frederick Barnard Hawley was born on 5 February 1843 in Albany (New York State), and died on 31 May 1929 in New York City. After spending his freshman year at Harvard University, he went to Williams College (Massachusetts) in 1861, where he graduated three years later. Returning to Albany, he took up the study of law, but gave it up after a year to go into the family's lumber business. In 1876 he became a cotton broker and merchant in New York City, a position he held until his retirement in 1926.

A couple of years after his move to New York Hawley published his first articles, advancing an approach to aggregate economic fluctuations based on his new conception of the saving–investment process. Those articles were expanded in 1882 into his book *Capital and Population*. In the 1890s and early 1900s Hawley published several articles in the *Quarterly Journal of Economics*, where he put

forward his ‘risk theory of profit’. Hawley's contributions to economics are contained in his 1907 book *Enterprise and the Productive Process*, in which he put together and elaborated ideas he had developed since the late 1870s. Feeling that his proposed new framework, based on the key role of the entrepreneur, had not been widely discussed, Hawley wrote 20 years later an article in the *American Economic Review* summing up his theoretical system. He was a member of the American Economic Association from 1888 to his death, and served as its treasurer (1892–95) and vice-president (1909). Hawley was one of the main protagonists in the long and intense controversy about the theory of profit that took place in American economics from the end of the 19th century to the beginning of the 20th century and culminated with the publication of Frank Knight's classic 1921 volume. Hawley enunciated the fundamental principle that there would be no profits in a competitive market in which the course of future events was entirely foreseen, since all factor services would be paid at rates fixed in advance, with changes in their productivity during the period of contract taken into consideration. Prices and costs would converge; there would be no ‘residue’. In actual economies, subject to future unforeseen influences, the function of the entrepreneur is to relieve others of risk. The entrepreneur bargains with the workers, capitalists and landlords for the use of their services, paying them not with any share of the product itself but with stipulated amounts of purchasing power. The actual product is owned by the entrepreneur, who must assume the responsibility of the enterprise and convert the output into purchasing power at the market price (cf. J. M. Keynes's 1933 similar distinction between a cooperative economy and an entrepreneur economy). According to Hawley, the entrepreneur is the dominant active element in the productive process, combining the three subsidiary passive productive factors. Since the incomes of individuals are necessarily composite, factors must be associated with functions, not with individuals. The entrepreneur's profit is a residual, non-contractual income whose amount is determined only after the output is sold. Hawley assumes that, in order to be relieved of a risk, agents are willing to pay more than the risk, calculated according to the laws of probability, is worth, since they ‘prefer a certainty to an uncertainty’. Entrepreneurs perform a service worth more to its recipients than the price they have to pay, and yet worth less to themselves than they get for it. Hence, the assumption of risk by the entrepreneur creates value by rendering a service in transferring risks from those to whom their subjective value is great to those to whom their subjective value is less, a mutually advantageous exchange of ‘certain goods’ for ‘uncertain goods’. Profit is the reward entrepreneurs get for performing that service in competitive conditions, and, by that, a component of the prices of commodities in general. Entrepreneurs are deemed less risk averse than other economic agents, except for ‘gamblers’ and ‘speculators’, who are not risk averse but do not take part in the productive process. The entrepreneurs’ subjective value of risk – the ‘irksomeness of being exposed to risk’ – is higher than its actuarial value, which means that industrial risks will not be assumed without the expectation of compensation in excess of their actuarial value. Since, under Hawley's assumption, entrepreneurs are on average and in the long run correct in their estimates, they pay productive factors less than the product will probably sell for, and absorb as profits a considerable portion of the annual flow of purchasing power. The ‘normal rate of profit’ is defined as the expectation of profit that just covers the marginal entrepreneur's subjective valuation of risks.

Hawley's theory of profit was taken up by Knight, who, however, criticized Hawley for ignoring the distinction between (known) risk and (unknown) uncertainty and overlooking the fact that the former is insurable. Although it is true that Hawley used the words ‘risk’ and ‘uncertainty’ interchangeably, it should be noted that he did pay careful attention to the implications of insurance for his argument. In the

first place, the act of insurance does not imply that the risk or its reward are extinguished, but only that the entrepreneur transfers to the insurer a corresponding part of its expected profits. Moreover, the risks of ownership – the most substantial part of risk – cannot be shifted by insurance, but only by a sale. Entrepreneurs can, to some extent, protect themselves from risks arising from price fluctuations by entering into hedging operations with speculators, as Hawley was aware from his experience as a cotton merchant. Although forward markets cannot completely eliminate the risks influencing selling prices, to the extent in which entrepreneurs hedge themselves they will be forced by competition to forgo their reward for risk bearing and lower their prices accordingly.

Whereas Hawley's theory of profit attracted some attention at the time, his contributions to macroeconomics went largely unnoticed by the profession, probably because their import became clear only after the Keynesian revolution. Fluctuations in aggregate economic activity are explained, according to Hawley, by changes in the saving–investment relation throughout the business cycle. Investment, described by the act of subjecting capital to the uncertainties inherent in actual ownership of capital goods, is naturally connected with entrepreneurs, not with capitalists or savers. The demand for new capital goods depends essentially on the entrepreneurs' profit expectations, which are subject to violent changes due to 'unforeseeable and incalculable causes' that affect the subjective valuation of risks. The treatment of savers' behaviour is based on Hawley's path-breaking concept of consumption as a function of income. He argued that expenditure changes less then income at all levels of income, since consumers keep close to the standard of living they have once adopted. More specifically, consumption plans are determined by expected income, measured by the average income of a series of years (that is, the mathematical expectation). This implies that a sudden increase of income will yield a larger percentage for saving than a gradual one of equal extent and, furthermore, that the proportion of saving out of profits is higher because it is more variable and uncertain than other sources of income.

Hawley used his new hypothesis about the consumption function to investigate the dynamics of the economy when the current rate of profit differs from its normal value. Periods of depression are characterized by a rate of profit lower than normal, associated to excess saving in the goods market. 'What can enterprisers do, by varying the character of supply, to protect themselves against this attack of the saving class upon their chances of profit?' asked Hawley (1907a, p. 224). The answer is the 'readjustment period', Hawley's main contribution to macroeconomic theory, which set him apart from the rest of the pre-Keynesian business cycle literature: a decline of output caused by excess aggregate supply will reduce supply more than demand (because of the consumption function) and bring the economy to equilibrium at less than full employment. The equilibrating effect of the contraction in aggregate income, identified by Don Patinkin (1982) as the core of the Keynesian principle of effective demand, can be found already in Hawley's writings. Some corollaries of the idea of 'readjustment period' were also pointed out by Hawley, such as the notion that an increase of the saving flow for a given investment level will bring about a contraction of income and a return of saving to its initial amount, so that in the end 'national parsimony defeats itself' – an early formulation of the 'paradox of thrift' usually associated with J. M. Keynes.

Whether he had any influence on Keynes is a moot point in the history of macroeconomics. Despite the fact that the English economist never referred to Hawley, that possibility cannot be disregarded, especially in view of the similarity of the wording of some key passages.

See Also

- entrepreneurship
- Keynes, John Maynard
- Knight, Frank Hyneman

Selected works

1879a. The ratio of capital to consumption. *National Quarterly Review* 79, 95–116.

1879b. The rationale of panics. *National Quarterly Review* 79, 277–92.

1882. *Capital and Population*. New York: A. M. Kelley, 1972.

1892. The fundamental error of ‘Kapital und Kapitalzins’. *Quarterly Journal of Economics* 6, 280–307.

1893. The risk theory of profit. *Quarterly Journal of Economics* 7, 459–79.

1900. Enterprise and profit. *Quarterly Journal of Economics* 15, 75–105.

1907a. *Enterprise and the Productive Process*. New York and London: G. P. Putnam's Sons.

1907b. Capital (Chapter 7 of *Enterprise and the Productive Process*). In *Business Cycle Theory – Selected Texts 1860–1939*, vol. 7, ed. M. Boianovsky. London: Pickering and Chatto, 2005.

1927. The orientation of economics on enterprise. *American Economic Review* 17, 409–28.

Bibliography

Bigelow, K. 1932. Hawley, Frederick Barnard. In *Encyclopaedia of Social Sciences*, vol. 7, ed. E. Seligman. London: Macmillan.

Boianovsky, M. 1996. Anticipations of the *General Theory*: The case of F. B. Hawley. *History of Political Economy* 28, 371–90.

Davis, R. 1953. Frederick B. Hawley's income theory. *Journal of Political Economy* 61, 117–26.

Dorfman, J. 1949. *The Economic Mind in American Civilization*, vol. 3. New York: Viking.

Hopkins, W. 1933. Profit in American economic theory. *Review of Economic Studies* 1, 60–6.

Keynes, J. M. 1933. The distinction between a co-operative economy and an entrepreneur economy. In *The Collected Writings of John Maynard Keynes*, vol. 29, ed. D. Moggridge. London: Macmillan, 1979.

Keynes, J. M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Knight, F. 1921. *Risk, Uncertainty and Profit*. New York: A. M. Kelley, 1964.

Patinkin, D. 1982. *Anticipations of the General Theory? And Other Essays on Keynes*. Oxford: Basil Blackwell.

Williams College. 1930. *Obituary Record of the Society of Alumni*, April, 3–4. Williamstown, MA: Williams College.

How to cite this article

Boianovsky, Mauro. "Hawley, Frederick Barnard (1843–1929)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000172> doi:10.1057/9780230226203.0709

The New Palgrave Dictionary of Economics Online

Hawtrey, Ralph George (1879 – 1975)

R.J. Bigg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cambridge School; credit; effective demand; expectations; gold standard; Hawtrey, R. G.; investment; Keynes, J. M.; labour supply; marginal productivity of capital; money demand; multiplier analysis; natural rate and market rate of interest; public works; quantity theory of money; rate of profit; Robertson, D. H.; Robinson, J. V.; savings; short-hoards; stocks; trade cycle; Wicksell, J. G. K.

Article

Hawtrey was born in Slough, near London, and went up to Trinity College, Cambridge, from Eton in 1898. Three years later he graduated 19th Wrangler in the Mathematical Tripos. Hawtrey remained at Cambridge for a further period to read for the civil service examinations, as was quite common at that time. This latter study included some economics with lectures largely by G.P. Moriarty and J.H. Clapham. In 1903 he entered the Admiralty, but in 1904 he transferred to the Treasury, where he was to remain until retirement in 1947 (his official retirement at 65 was in 1944). Hawtrey's only academic appointments in economics were in 1928–9, when he was given special leave from the Treasury to lecture at Harvard (as a visiting professor) and after his retirement, when he was elected Price Professor of International Economics at the Royal Institute of International Affairs (1947–52). Hawtrey served as President of the Royal Economic Society between 1946 and 1948.

Hawtrey was not, therefore, directly a part of the ‘Cambridge School’ of economics. Marshall took no immediate part in Hawtrey's economic education which was, for the most part, acquired in the Treasury. Nonetheless he had close contacts with the Cambridge economists. Away from economics he was involved with both the Apostles and with Bloomsbury, whilst within the subject he was a visitor to Keynes's Political Economy Club at Cambridge and his major work, *Currency and Credit* (1919a) became a standard work in Cambridge in the 1920s. Furthermore, although there were differences in approach between Hawtrey and the Cambridge School in some areas, Keynes himself noted in reviewing *Currency and Credit* the similarities between Hawtrey's approach to the theory of money and that of the Cambridge School – though Keynes remarked that Hawtrey had reached his results independently (Keynes, 1920).

I. Hawtrey was primarily a monetary economist; his major contributions related to the quantity theory and the trade cycle. He was one of the first English economists to stress the primacy of credit-money rather than metallic legal tender. Furthermore his income-based approach, like that of the Cambridge School, led to a closer integration of the theories of money and output. For Hawtrey, money income determines expenditure, expenditure determines demand and demand determines prices.
Hawtrey summarized his aims in monetary theory in the preface to *Currency and Credit*:

> Scientific treatment of the subject of currency is impossible without some form of the quantity theory•…•but the quantity theory by itself is inadequate, and it leads up to the method of treatment based on what I have called the consumers' income and the consumers' outlay – that is to say, simply the aggregates of individual incomes and individual expenditures. (1919, p. v)

Investment (the result of saving) is included in consumers' outlays, since it is spent on fixed capital. Consumers' balances are then the difference between outlays and income and thus consist only of accumulated cash balances (including money held in bank accounts). In addition there is a similar demand for money balances by traders related to their turnover. Of course individual agents may hold both consumers' and traders' balances – Hawtrey notes that the true income of traders is the profits of the business and that this is included in consumers' income.
The 'unspent margin', or total money balances, consists of the consumers' and traders' balances taken together. From this Hawtrey derives a form of the quantity theory. Hawtrey argues that traders' balances are relatively stable, and thus the operational relationships are concerned with the supply of money (in a wide sense taken to include credit) and consumers' income and outlay. It is worth noting that compared to the Cambridge income-based approach Hawtrey's places greater emphasis on the demand for nominal balances rather than real balances. It is also interesting to note that Keynes used a similar balances approach to the quantity theory in the period after 1925 leading up to the theory presented in the *Treatise on Money* (1930), where he distinguishes first between investment and cash deposits and later between income, business and savings deposits.
The demand for money is also analysed in terms of motives. Hawtrey identifies a transaction demand, a precautionary demand, and a residual demand which reflects a gradual accumulation of savings balances or what Joan Robinson has called short-hoards (Robinson, 1938). Hawtrey envisages agents as saving gradually but investing only larger sums periodically. In the meantime these short-hoards act as a buffer stock. The main costs of holding money balances is the interest forgone, and thus Hawtrey points to a balancing process between costs and advantages in determining desired balances. The introduction of a banking system into the model allows agents to substitute borrowing power for money balances (Hawtrey, 1919a, pp. 36–7).
II. Hawtrey also introduces a concept of effective demand:

> The total effective demand for commodities in the market is limited to the number of units of money of account that dealers are prepared to offer, and the number they are prepared to offer over any period of time is limited according to the number they hope to receive. (1919a, p. 3)

Later, in *Trade and Credit* (1928) Hawtrey points to a flaw in the theory of an elastic supply of labour based on marginal utilities (or disutilities) of product and effort. He argues that whilst a difference between the marginal utility of the product and the disutility of effort may prompt an additional supply of labour ‘in the simple case of a man working on his own account’ (1928, p. 148), this is not the general case since: ‘the decision as to the output to be undertaken is in the hands of a limited number of employers, and the workmen in the industry are passively employed by them for the customary hours at the prevailing rates of wages’ (1928, p. 149). In this case output decisions are based not on the gross proceeds, but on the net profit margin.

The factor of expectations is also present in Hawtrey's analysis of fluctuations. Hawtrey suggests that during a downturn in activity money balances will be reduced more quickly than they are replenished in an upswing. This is because as income drops initially consumers will draw on their balances to maintain their outlay.

There is then a further level of adjustment as changes in consumers’ outlays impacts on traders. Consider an upswing: the increase in consumers’ outlays will increase the nominal receipts of traders and reduce their physical stocks. Traders, finding their balances have increased can either order more stock from manufacturers or reduce their bank indebtedness. Prices will tend to rise as traders find they are unable to replenish their stocks fast enough. For Hawtrey quantity adjustments occur *before* price adjustments, indeed often the price movements result from the quantity movements. Thus ‘the rise of prices, when it occurs, is caused by the activity; it is a sign that production cannot keep pace with demand’ (1928, p. 156). The role of stocks in Hawtrey's theory is pivotal, in general it is quantity signals rather than price signals which are the more effective. The existence of traders’ stocks means that it is nearly always possible to meet the demand for increased consumption in the short term, which implies that at least in the short term a naive proportionality between increases in the money supply and prices does not hold. Furthermore the model opens the possibility of short-run quantity adjustments in disequilibrium. Thus, argues Hawtrey:

> It is only in times of equilibrium, when the quantity of credit and money in circulation is neither increasing nor decreasing, that the relation of prices and money values to that quantity of credit and money is determined by the individual's considered choice of the balance of purchasing power appropriate to his income. … In practice it seldom, perhaps never, happens that a state of equilibrium is actually reached. (1919a, p. 46)

Nonetheless Hawtrey's theory of the trade cycle is money-driven. It is the fluctuations in money and credit which stimulate and support the price and quantity movements. Hawtrey argued that the periodic nature of the trade cycle was solely due to monetary factors. Traders stocks are viewed as being highly interest elastic since they are held on borrowed funds, investment in fixed capital is also interest elastic (based on a marginal efficiency of capital analysis).

Thus an increase in the rate of interest will tend to reduce the demand for credit due to a lower demand for stocks and a reduced level of new investment. If the increased rate of interest is itself the result of a decreased supply of credit then there may also be some quantitative restrictions of borrowing. To reduce their stocks traders will stop giving new orders to manufacturers, leading to a drop in the level of output which will further diminish the demand for credit, as well as the level of income and demand. Traders

may reduce prices to stimulate sales to accelerate the destocking process. There is thus a tendency to a cumulative decline in output, credit and prices until the banks find themselves with excess reserves and believe it to be profitable to reduce the interest rate and expand credit. For Hawtrey, macroeconomic disequilibrium was defined in terms of monetary disequilibrium.

The solution was also therefore monetary, and in particular the short-term rate of interest (the long-term rate of interest was seen as relatively ineffective as a means of control because of its relatively slow impact on investment). Hawtrey viewed the psychological factors in the trade cycle as secondary, arguing that no amount of good news or bad could seriously affect the cycle if monetary factors were not accommodating. He also opposed the public works solution to a slump in output along similar lines – and in this respect is associated with the ‘Treasury View’ (see Hawtrey, 1925). In later life Hawtrey did acknowledge that public works could play a role in severe depressions, but as Haberler (1939, p. 23) points out Hawtrey viewed those occasions when cheap money would fail to stimulate a revival as generally very rare – although he accepts that this was the case in the 1930s.

III. For Hawtrey, investment decisions were made on a Marshallian marginal productivity of capital basis. In a perfectly competitive market, the marginal return on capital employed would be equalized across every industry. In these circumstances Hawtrey identifies the ‘ratio of labour saved per annum to the labour expended on first cost’ as ‘a physical property of the capital in use’ (1913, p. 66) and as a ‘natural rate’ of interest. Under stable monetary conditions and in the absence of a banking system this natural rate is equal to the market rate of interest or the profit rate, as in the standard marginal efficiency of capital analysis. But changes in monetary conditions will generate changes in prices and thus profits; hence the market rate will diverge from the natural rate in the same direction as the movement in prices. With the addition of a banking system, the actual rate of interest will depend on the behaviour of the banks, and in particular their reserve position. Thus the interest rate will diverge from the profit rate. There is a three-way equilibrium condition, relating the physical return on capital, the profit rate and the balance position of banks, that is, $N=p=r$ where N is the natural rate, p is the profit rate and r is the interest rate. An increase in the supply of money will cause a rise in prices and the availability of credit; thus $N<p$ at the same time the banks will find themselves with excess reserves and thus interest rates will tend to be lower than otherwise to stimulate borrowing, that is, $p>r$. This will be generally expansive, demand, investment and output will all tend to rise—but the seeds of the eventual slump are already present. The rising prices and relatively low rate of interest will encourage firms to over-invest, expecting returns greater than those actually accruing. On the downward cycle $N>p$ and $p<r$.

It is worth briefly considering the relationship between Hawtrey's natural rate and that associated with Wicksell. In his early work Wicksell took the natural rate as that prevailing if loan transactions were made in kind, but he later revised this to equate the natural rate with the rate of profits received in the form of money (see Lindahl, 1939, p. 261; Lindahl also discusses a physical return on capital ‘natural rate’ similar to Hawtrey's). Thus Wicksell's natural rate can be seen as closer to Hawtrey's profit rate. Wicksell, like Hawtrey, also associates the natural rate with an equilibrium between savings and investment and stability in the price level.

Hawtrey does not place great stress on this natural rate analysis, concentrating more on the relationship of the profit rate and the interest rate. There are also considerable practical problems in determining Hawtrey's natural rate, particularly in imperfect capital markets (see the discussion in Lindahl, 1939, and Haberler, 1939).

IV. Hawtrey, like most of the inter-war Cambridge economists, had a fundamental belief in the self-

adjusting nature of the economic system, even though much of the analysis of the period would suggest otherwise. Hawtrey believed that the system was continually approaching or seeking an equilibrium, though in practice the next shock would come before the adjustment process was complete. However, Hawtrey's theoretical approach was to concentrate on the processes of adjustment to monetary disequilibrium.

The income/inventories approach to the trade cycle is mirrored in Hawtrey's analysis of savings and investment. For Hawtrey, savings were directed into investment opportunities by securities dealers who acted like traders, holding stocks of securities financed by bank borrowing, intermediating between the savers and investors. In the early 1930s Hawtrey developed this analysis into a model where an imbalance between savings and investment results in an unanticipated change in physical stocks of goods as a result of changes in consumers' incomes (and outlays).

Savings are the excess of consumers' income over desired consumption and are represented by investment; an increase in money balances; or purchases of goods. Net investment is defined as the total of securities sold less those bought by securities dealers. Clearly the price of securities (and by implication the long-term rate of interest) will move to achieve an equilibrium between the net amount of investment and capital raised, but planned savings can exceed the resources seeking investment, in which case the excess must flow into additional money balances or additional consumption – or vice versa. In either case an expansion or contraction of demand is set in motion. Both Saulnier (1938) and Haberler (1939) note the similarity of this analysis with that of D.H. Robertson. This aspect of Hawtrey's theory is also reviewed by Davis (1981).

Hawtrey's disequilibrium analysis where unintended changes in stocks bring about an equality of actual savings and investment, but a further chain of adjustment if intended savings and investment are not equal, is remarkably close to the modern textbook presentation of the Keynesian equilibrium adjustment process. It is interesting therefore to briefly examine the discussions between Keynes and Hawtrey leading up to the *General Theory*. Indeed in Hawtrey's comments on the drafts of the *Treatise* he is often more 'Keynesian' than Keynes himself! (see Keynes, 1973, pp. 138–69). At this stage Keynes envisages:

1. (1) A decline in fixed investment relatively to saving.
2. (2) A fall of prices …
3. (3) A fall of output, as a result of the effect of falling prices and accumulating stocks on the minds of entrepreneurs (Letter to R.G. Hawtrey, 28 November 1930; Keynes, 1973, p. 143).

The fall in output leads to a disinvestment in working capital, and eventually to a situation where total investment and prices fall too far. Once output stops declining this leads to a slight rise in prices, and, given the low level of stocks at this point, so starts the upturn. Hawtrey, on the other hand, sees a direct effect on output from the contraction in demand at unchanged prices, and criticizes Keynes for only taking account of the reduction in prices relative to costs in his fundamental equations (Keynes, 1973, pp. 151–2). Hawtrey argues that 'the change in prices when it does occur is not by itself an adequate measure of the departure from equilibrium' (Keynes, 1973, p. 151). And later comments that: 'A manufacturer restricts output, not because he believes that prices are about to fall, but because he cannot secure sufficient sales at the existing price' (Letter from R.G. Hawtrey, 6 December 1930; Keynes,

1973, p. 165).

Prices are reduced only gradually in an attempt to boost orders, but Hawtrey also points out that it is the level of retail prices which will determine the ultimate level of sales – and this will depend on how quickly retailers pass on the manufacturers' reductions. Both Hawtrey and Keynes realize that the decline in output will rebound on savings, but do not appear to treat this as the main equilibrating factor (as in the later Keynesian theory).

V. The high point of Hawtrey's official career came with the Genoa International Financial Conference in 1922. The conference was concerned with the problems relating to a general return to the international gold standard after the First World War. In particular there was concern that the quantity of gold might be insufficient for a return to the system at the old pre-war parities, other concerns centred on problems relating to fluctuations in demand for monetary gold. The result was greater interest in a joint Sterling–gold standard along the lines of the gold exchange standard operated earlier by India and other countries. Hawtrey's main suggestions adopted by the Genoa conference related to greater cooperation between central banks to manage the demand for monetary gold and to regulate credit so as to stabilize the purchasing power of gold. However, the Genoa Resolutions were never acted on, largely as a result of US scepticism, and the failure of other central banks to participate in the planned follow-up conference (see Davis, 1981).

At the Treasury Hawtrey had argued that there were two primary considerations for monetary policy: the stabilization of internal prices and the stabilization of the foreign exchanges. Given the UK's status as a financial centre he argued that exchange instability was particularly damaging and would make the covering of trade finance offered through London increasingly difficult. This predisposed him towards the gold standard as the de facto most practical means of achieving exchange stability.

Though Hawtrey was aware of possible deflationary problems associated with the return to gold, he appears to have believed that the exchange rate would return to par naturally, and that the necessary adjustments would come from American inflation rather than UK deflation (see the discussion in Moggridge, 1972, pp. 71–2, 91).

VI. Despite a long and active life, Hawtrey's main theoretical contributions to economics came largely in the interwar period. His first book, *Good and Bad Trade*, was published in 1913 and sets out a view of the trade cycle which received a more rigorous theoretical treatment in *Currency and Credit* (1919a), but which remained little changed thereafter, although the debates surrounding Keynes's *Treatise* prompted some refinements and revisions, as did the experience of the 1930s depression. The last major contemporary studies of his work were Saulnier (1938), which also reviewed the theories of D.H. Robertson, F.A. von Hayek and J.M. Keynes, and Haberler (1937; 1939). Interest in Hawtrey revived in the later 1970s following his death (for example, Davis, 1977; 1981; 1983; see also Deutscher, 1990). Particular attention has been given to Hawtrey's role in the development of multiplier analysis; see Dimand (1997).

In the 1920s innovative monetary theory in England was largely associated with the Cambridge School and in particular D.H. Robertson and Keynes. Hawtrey with his close Cambridge contacts contributed to this work, as the correspondence with Keynes now reprinted in the *Collected Works* shows. The three were often working along similar lines in this period and their work reflects (to varying degrees) an increasing failure of conventional theory to match the problems of the age.

Selected works

1913. *Good and Bad Trade*. London: Constable.

1919a. *Currency and Credit*. London: Longmans.

1919b. The Gold Standard. *Economic Journal* 29, 428–42.

1921. *The Exchequer and the Control of Expenditure*. London: World of Today.

1922. The Genoa Resolutions on currency. *Economic Journal* 32, 290–304.

1923. *Monetary Reconstruction*. London: Longmans.

1924. Discussion on monetary reform. *Economic Journal* 34, 155–76.

1925. Public expenditure and the demand for labour. *Economica* 5, 38–48.

1926. *The Economic Problem*. London: Longmans.

1927. *The Gold Standard in Theory and Practice*. London: Longmans.

1928. *Trade and Credit*. London: Longmans.

1930. *Economic Aspects of Sovereignty*. London: Longmans.

1931. *Trade Depression and the Way Out*. London: Longmans.

1932. *The Art of Central Banking*. London: Longmans.

1933. Saving and hoarding. *Economic Journal* 43, 701–8.

1934. Monetary analysis and the investment market. *Economic Journal* 44, 631–49.

1937. Alternative theories of the rate of interest: three rejoinders. *Economic Journal* 47, 436–43.

1938. *A Century of Bank Rate*. London: Longmans.

1939. *Capital and Employment*. London: Longmans.

1944. *Economic Destiny*. London: Longmans.

1946. *Economic Rebirth*. London: Longmans.

1946. *Bretton Woods: For Better or Worse*. London: Longmans.

1949. *Western European Union. Implications for the UK*. London: Royal Institute of International Affairs.

1950. *Balance of Payments and the Standard of Living*. London: Royal Institute of International Affairs.

1954. *Towards the Rescue of Sterling*. London: Longmans.

1955. *Cross Purposes in Wage Policy*. London: Longmans.

1961. *The Pound at Home and Abroad*. London: Longmans.

1965. *An Incomes Policy*. Economic Papers No. 4. London: Department of Economics and Management, Woolwich Polytechnic.

1967. *Incomes and Money*. London: Longmans.

Bibliography

Cambridge University, 1917. *Historical Register of the University of Cambridge to 1910*. Cambridge: Cambridge University Press.

Davis, E.G. 1977. The economics of R.G. Hawtrey. Carleton Economic Papers 77(12), Department of Economics, Carleton University.

Davis, E.G. 1981. R.G. Hawtrey, 1879–1975. In *Pioneers of Modern Economics in Britain*, ed. D.P. O'Brien and J.R. Presley. London: Macmillan.

Davis, E.G. 1983. The macro-models of R.G. Hawtrey. Carleton Economic Papers 83(4), Department of Economics, Carleton University.

Deutscher, P. 1990. *R. G. Hawtrey and the Development of Macroeconomics*. Ann Arbor: University of Michigan Press.

Dimand, R. 1997. Hawtrey and the multiplier. *History of Political Economy* 29, 549–59.

Haberler, G. 1937. *Prosperity and Depression*. Geneva: League of Nations, Economic & Financial 1936, II.A.24.

Haberler, G. 1939. *Prosperity and Depression*. 2nd edn. Geneva: League of Nations, Economic & Financial 1939, II.A.4.

Hutchison, T.W. 1953. *A Review of Economic Doctrines 1870–1929*. Oxford: Oxford University Press.

Keynes, J.M. 1920. Review of Hawtrey's *Currency and Credit*. *Economic Journal* 30, 362–5.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London: Macmillan. Reprinted as *The Collected Writings of John Maynard Keynes*, vol. 4, London: Macmillan, 1971.

Keynes, J.M. 1930. *A Treatise on Money*, 2 vols. London: Macmillan. Reprinted as *The Collected Writings of John Maynard Keynes*, vols. 5 and 6, London: Macmillan, 1971.

Keynes, J.M. 1973. *The Collected Writings of John Maynard Keynes*, vol. 13, London: Macmillan.

Lindahl, E. 1939. *Studies in the Theory of Money and Capital*. London: George Allen & Unwin. Reprinted, New York: Augustus M. Kelley, 1970.

Moggridge, D.E. 1972. *British Monetary Policy 1924–1931*. Cambridge: Cambridge University Press.

Moggridge, D.E., ed. 1973. *The General Theory and After*. Part I: Preparation. In *The Collected Writings of John Maynard Keynes*, vol. 13, London: Macmillan.

Robinson, J.V. 1938. The concept of hoarding. *Economic Journal* 48, 231–6.

Rouse-Ball, W.W. and Venn, J.A., eds. 1913. *Admissions to Trinity College, Cambridge*. Vol. 5: *1851 to 1900*. London: Macmillan.

Saulnier, R.J. 1938. *Contemporary Monetary Theory*. Columbia University Studies in the Social Sciences No. 443. New York: Columbia University Press.

The Times. 1975. Sir Ralph Hawtrey CB (Obituary). 22 March.

How to cite this article

Bigg, R. J. "Hawtrey, Ralph George (1879–1975)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_H000028> doi:10.1057/9780230226203.0710

The New Palgrave Dictionary of Economics Online

Hayek, Friedrich August von (1899 – 1992)

Bruce Caldwell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the major intellectual contributions of the Austrian-born Nobel laureate Friedrich Hayek. Within economics, Hayek made contributions to many areas, among them monetary theory, trade cycle theory, and capital theory. His ‘knowledge-based’ critique of socialism and subsequent work on ‘the knowledge problem’ are widely viewed as seminal contributions to economics. Hayek also did substantial work in such fields as political theory, the methodology of the social sciences, psychology and intellectual history. Finally, his writings on spontaneous orders and his ‘theory of complex phenomena’ anticipated later developments in such areas as complexity theory and agent-based modelling.

Keywords

Austrian business cycle theory; Austrian economics; average period of production; behaviourism; Böhm-Bawerk, E. von; capital theory; Chicago School; common law; complements; complex phenomena; constitutionalism; constructivism; currency competition; general equilibrium; Hayek, F. A. von; information aggregation and prices; intertemporal equilibrium; Keynesian Revolution; knowledge problem; liberty; majoritarianism; markets; Mises, L. E. von; money; Mont Pèlerin Society; natural rate and market rate of interest; neutrality of money; new institutional economics; planning; public choice; roundabout methods of production; rule of law; scientism; social justice; socialism; socialist calculation debate; special interests; spontaneous order; substitutes; tacit knowledge; Wicksell, J. G. K.

Article

Born on 8 May 1899, the polymath economist and social theorist Friedrich August von Hayek had the good fortune to be repeatedly in the right place at the right time, crossing paths with some of the century's most brilliant economists and thinkers. He grew up in *fin de siècle* Vienna, a place and time of extraordinary intellectual vitality. Through his maternal grandfather, Franz von Juraschek, a professor of

civil law and civil servant, he gained an introduction to the academic world in Vienna, and through his father, August, a medical doctor and devoted botanist, a love of biology and the sciences as well as an acquaintance with another extended community of scholars. As a student at the University of Vienna his major professor was Friedrich von Wieser, and among his classmates were Oskar Morgenstern, Gottfried Haberler, and Fritz Machlup. After finishing his studies Hayek spent 15 months in the United States where, armed with letters of introduction from Joseph Schumpeter, he encountered most of the major American economists, both those contributing to the Marginalist School as well as the leading institutionalist and business cycle analyst Wesley Clair Mitchell. When he returned he joined the *Miseskreis*, Ludwig von Mises's study circle.

In the later 1920s he published an article in German that was read by Lionel Robbins, a newly appointed professor at the London School of Economics (LSE). This led to an invitation to present some lectures, and ultimately, in 1932, to Hayek being appointed to the Tooke Chair of Economic Science and Statistics. While at the LSE Hayek would engage in debates on the leading issues in economics with some of the discipline's most important members: John Maynard Keynes and Piero Sraffa over monetary theory, Frank Knight and Nicholas Kaldor over capital theory, Oskar Lange and Evan Durbin over socialism. He was also instrumental in bringing the philosopher of science Karl Popper to the LSE. Hayek remained at the LSE until 1950, when he moved to the Committee on Social Thought at the University of Chicago. There he counted among his colleagues Milton Friedman, Aaron Director, and George Stigler. Retiring in 1962, Hayek had successive appointments at the University of Freiburg and the University of Salzburg, returning again to Freiburg in 1977. In 1974 he was awarded, with Gunnar Myrdal, the Bank of Sweden Nobel Prize in Economic Sciences, and in 1991 the Presidential Medal of Freedom. Hayek died in Freiburg on 23 March 1992.

If Hayek was in the right place at the right time, it was usually with the wrong ideas, at least from the perspective of most of his contemporaries. He was a sharp critic of Keynes well before the onset of the Keynesian Revolution. Though he helped introduce English-speaking economists to general equilibrium theory, he claimed that a preoccupation with static equilibrium analysis would mislead economists about the true nature of a dynamic market process. He attacked socialism when most members of the intelligentsia viewed it as a preferred middle way between an apparently failed capitalist system and totalitarianisms of the communist and fascist varieties; for Hayek such thinking was ‘the muddle of the middle’. When most Western democracies were embracing some form of the welfare state, he criticized the concept of social justice that provided its philosophical foundations. While most of the social sciences were moving towards more and more specialized studies, his work was increasingly integrative and multidisciplinary. The views Hayek embraced over most of his career were almost systematically out of step.

From the perspective of the early 21st century, history would judge Hayek's legacy more kindly than did many of his contemporaries. He lived to witness the collapse of the Soviet bloc, which many took as vindication of his and Ludwig von Mises's early critique of central planning. His view that a competitive market system with freely adjusting prices is an essential mechanism for coordinating social action in a world of dispersed knowledge is taken by economists as a fundamental insight. His insistence that markets be embedded in a host of other social and political institutions for their proper functioning provides a jumping off point for such diverse movements within economics as experimental investigations of market institutions, public choice and constitutional analysis, and the new institutional economics. Philosophers of mind, evolutionary biologists, and neuroscientists have been attracted to his

‘connectionist’ approach for understanding the development and functioning of the brain. His theory of complex phenomena and work on spontaneous orders has clear analogues in complexity theory and agent-based computational modelling (Caldwell, 2004, ch. 14). If Hayek remains a controversial figure in some quarters, even his critics acknowledge the breadth and depth of his contributions. One pundit, writing in the *New Yorker* in 2000, even went so far as to call the 20th century ‘the Hayek century’ (Cassidy, 2000, p. 45). Considering that this was only about two decades after the British Labour politician Michael Foot had referred to him as a ‘mad professor’, the reputational turnabout has been substantial.

Early work

Hayek's first trip to the United States took place in 1923–24. While there he studied new work on monetary policy and the control of the business cycle; he also witnessed the policy experiments being undertaken under the auspices of the then only recently established Federal Reserve System. Hayek subsequently wrote a paper on US monetary policy in which he criticized the goal of stabilizing the general price level (Hayek, 1926). According to the Austrian theory of the cycle, relative price movements play an essential role in the unfolding of the cycle, so that any policy prescription that focused solely on aggregates was judged deficient for ignoring such movements.
Hayek spelled out the Austrian approach in more detail in his first book, published in 1929, an English translation of which appeared in 1933 as *Monetary Theory and the Trade Cycle*. There he argued for a monetary approach to the origins of the cycle. Hayek claimed, first, and contra both the American institutionalists and German historical economists, that any adequate explanation of the cycle must be *theoretical*, and, further, that it must be consistent with, and presuppose the validity of, the standard equilibrium theory of the day. This poses a problem, however, for if one accepts the results of standard equilibrium theory, where prices adjust to clear markets, a question immediately arises: how can a disproportionality between the production of capital goods and consumer goods that occurs during the boom phase of the cycle occur? For Hayek, money provided the answer. Though the use of money confers substantial benefits, most evidently to facilitate trade, and thereby to encourage specialization and growth, it is also a ‘loose joint’ in the system of exchange: ‘Money being a commodity which, unlike all others, is incapable of finally satisfying demand, its introduction does away with the rigid interdependence and self-sufficiency of the ‘closed’ system of equilibrium’ (Hayek, 1933, p. 44).
Another significant piece in this period was Hayek's paper ‘Intertemporal Price Equilibrium and Movements in the Value of Money’ (Hayek, 1928), which is widely acknowledged as an early important contribution to the theory of intertemporal equilibrium.

Hayek comes to the LSE

Hayek's lectures in early 1931 at the LSE were published as *Prices and Production*, a book in which he completed the task begun in *Monetary Theory and the Trade Cycle* by tracing out the effects of monetary disturbances on the economy. Using a framework developed by Knut Wicksell (1906) and further adapted by Ludwig von Mises (1924), Hayek posited a *natural rate of interest* that, in the absence of monetary factors, would just equalize the demand for capital and the supply of savings.

When households save, they forgo present for future consumption. The funds are borrowed by firms for investment in more ‘roundabout’ methods of production which allow firms to produce more goods in the future, thereby satisfying the desires of consumers. The natural rate of interest, then, is a relative price that coordinates a community's preferences regarding present and future consumption with the production processes that create the goods.

However, in the crisis stage of the cycle, an excess of capital goods (relative to consumers’ preferences) are created. This occurs because of a divergence between *the natural* and *the market rate of interest*, caused by bank lending activity. Specifically, a lowering of the market rate of interest below the natural rate leads firms to move to more roundabout methods of production, just as they would have done had there been a reduction in the natural rate. However, in this case, because there has been no change in consumers’ preferences, the lengthening of production processes is not sustainable. At some point before the completion of the transition, prices for consumer goods begin to rise, which signals to firms that they have made errors. As they begin to abandon the more roundabout methods, a cyclical downturn is initiated.

Hayek's theory carried the unfortunate policy implication that there was little that policymakers could do once an economy was in a recession. Recessions were avoidable only if one could make money ‘neutral’ by keeping the natural rate equal to the market rate of interest. Unfortunately, no one knows what the natural rate is; only the market rate is observable. The downturn, painful as it is, is actually the system returning to equilibrium, correcting for past errors. As such, policies that attempt to address a recession by injecting money only further encourage firms to persist in their mistaken behaviours, making the ultimate downturn even more severe.

Hayek's book had a tumultuous reception. In late 1930 John Maynard Keynes published his own analysis of the problems of a monetary economy, *A Treatise on Money* (Keynes, 1930), in which he also used the Wicksellian framework. Hayek's critical review of Keynes's book drew a heated response from Keynes, who also took Hayek's *Prices and Production* to task, noting famously that ‘It is an extraordinary example of how, starting with a mistake, a remorseless logician can end up in bedlam’ (Keynes, 1931, p. 154). For a while, as John Hicks later recounted, the burning question of the day for economists was, ‘Which was right, Keynes or Hayek?’ (Hicks, 1967, p. 203).

Others entered the fray, and the weight of the combined criticisms ultimately led both Keynes and Hayek to revise their theories. Keynes finished first, publishing *The General Theory of Employment, Interest and Money* in 1936. Hayek's initial plan was to construct a dynamic theory of a capital-using monetary economy. He worked on the book in starts and stops for the rest of the decade, finally publishing it as *The Pure Theory of Capital* in 1941. There Hayek abandoned the simplifying Böhm-Bawerkian notion of an ‘average period of production’, and in its place systematically explored a variety of possible relations between inputs (both those available at a given point in time and over a continuous period) and outputs (whose availability might likewise vary over time). He examined the effects of substitutability and complementarity, of the introduction of new ‘inventions’, both in cases in which they are foreseen and when they are not, and of whether decisions are made by a single individual or within a competitive system. A key theme of the book is that the capital structure is constantly evolving as the market continually provides new information. In that evolution, capital is rarely either so malleable as to be instantaneously transformable, or so permanent as to be incapable of being applied in a different production process.

Hayek's book made important advances in capital theory, but he never was able to accomplish his larger

goal. After seven years of labour he could only provide in the closing three chapters of the book a sketch of how to integrate his capital theory into a monetary framework. As he later once put it, once you get beyond Böhm-Bawerk's simplifying assumption of an average period of production, ‘things become so damn complicated it's almost impossible to follow it’ (Hayek, 1994, p. 141). Meanwhile Keynes's victory in the area of macroeconomics quickly became complete.

Socialist calculation and the knowledge problem

In the 1920s, the British economy went through wrenching structural adjustments, and with the depression of the 1930s many among the intelligentsia came to view socialist planning as the only acceptable alternative system. Economists, some of them colleagues of Hayek's at the LSE, began issuing proposals for how to organize such a system. In 1935, Hayek entered the discussion with the publication of *Collectivist Economic Planning*, a collection of translations of essays from an earlier debate that had been initiated by Ludwig von Mises. Hayek included his mentor's essay, in which Mises argued that rational planning was ‘impossible’ under socialism. His point was that a monetary economy with freely adjusting market prices reveals relative scarcities among factors of production. When the means of production are state-owned, there are no prices for factors of production, and hence no signals to help socialist managers allocate resources rationally (Mises, 1920).

Some socialists (for example, Dickinson, 1933) responded by invoking Paretian general equilibrium theory, which they argued disproved Mises's thesis. They noted that any economic system could be represented by a system of equations, so that the only difference between a planned and a free market system lay in who was responsible for ‘solving’ the equations, socialist managers or private entrepreneurs. If some of the prices that the socialist managers chose were wrong, gluts or shortages would appear, signalling them to adjust the prices up or down, just as in a free market. Through such a trial and error procedure, a socialist economy could mimic the efficiency of a competitive free market system, while avoiding its many problems: wasteful competition, the market failures that attend monopoly and externalities, and an unjust income distribution (Lange, 1938).

Hayek challenged this vision in a series of contributions (Hayek, 1937; 1945; 1968) to what has since come to be called ‘the knowledge problem’. In ‘Economics and Knowledge’ (1937) he pointed out that the standard equilibrium theory of his day assumed that all agents have full and correct information. In the real world, however, different individuals have different bits of knowledge, and furthermore, some of what they believe is wrong. In that world, the key question is how it comes about that the actions of individuals ever get coordinated, a question that equilibrium analysis with its full information assumption brushes aside.

Hayek posited the market as a key coordinating institution. He described the market process as operating in a world of constant change, in which freely adjusting prices are formed as the result of decisions, typically forward-looking, of literally millions of market participants. Their decisions are based in part on the vast array of prices that they confront in the market, prices that give them information about relative scarcities. But in addition, agents act on the basis of localized knowledge, knowledge of particular circumstances of time and place, some of which is tacit – that is, they cannot say why they are acting on it. Their market activity also reflects this localized knowledge, and by acting their knowledge becomes embedded in the array of market prices. In short, market activity is both price-determined (prices shape what people do) and price-determining (what people do, based on local knowledge,

determines what prices are). Market prices coordinate the specific knowledge of time and place possessed by millions of market agents. Socialist schemes that involve price fixing, as many of the proposals did, would keep the communication system from working. Hayek also doubted that trial and error price adjustment methods could ever mimic the speed of adjustment produced by markets, where errors to be corrected are simultaneously profit opportunities for alert entrepreneurs. Finally, Hayek criticized the profession's focus on standard equilibrium analysis which, by concentrating on equilibrium states, obscures the competitive process by which knowledge about relative scarcities becomes known: that theory ‘starts from the assumption of a “given” supply of scarce goods. But which goods are scarce goods, or which things are goods, and how scarce or valuable they are – these are precisely the things that competition has to discover’ (Hayek, 1968, p. 181). In short, market competition provides a discovery procedure. Hayek developed these ideas in a series of papers, the most famous of which, ‘The Use of Knowledge in Society’, is still widely cited by traditional general equilibrium theorists as well as economists working in the economics of information (Hayek, 1945).

The abuse of reason project and the road to serfdom

Though Hayek felt he had launched a telling attack against socialism, few in the late 1930s were persuaded by his economic reasoning. Hayek began to realize that the attractiveness of socialism went far beyond economics. Socialists promised a society that was not only more efficient than capitalism, but also one that was more just, where individuals have more self-determination and greater political freedom, and in which scientific reasoning would be used to improve upon a host of outdated social institutions. If he were successfully to challenge these utopian visions, economic arguments were not enough. He would need to develop political, historical and ethical arguments against them as well. During the Second World War Hayek began doing just that, in a massive piece of work that he called the ‘Abuse of Reason’ project. His overarching goal was to show how a number of then-popular doctrines and beliefs, doctrines with which he disagreed, had a common origin in some fundamental misconceptions about the proper methods for studying social phenomena. Central to his argument was the critique of *scientism*, which he defined as the ‘slavish imitation’ of the methods of the natural sciences in the study of social phenomena (Hayek, 1942–44, p. 24). He criticized the objectivism, historicism and collectivism of the ‘scientistic prejudice’, and contrasted these with his own preferred approach, one that was subjectivist, theoretical, and individualist. In the essay ‘Scientism and the Study of Society’ (Hayek, 1942–44) he also articulated a fundamental thesis about the limitations of our knowledge in the social sciences: that rather than make precise predictions often the best we can do is to make a pattern prediction, or alternatively to provide an explanation of the principle by which some social phenomenon came into being.

Hayek never completed the Abuse of Reason project, although sections of it were published separately during and after the war. One of these became his most famous book, *The Road to Serfdom*. As noted above, many advocates of socialism had promised that socialism would bring greater political freedom. In *The Road to Serfdom* Hayek countered that planning of the economy would soon lead to increasing political control as well. One of the virtues of a market economy is that it allows people with very different tastes to express them, and (for those with the means) to get them satisfied, through the market. In a planned economy, socialist managers must decide which goods, and in what quantities, get produced. Invariably some people will not like the decisions they make, and will protest. A change in

the mix will cause others to protest. If any progress is to be made, even democratically elected socialist regimes will at some point be forced simply to make the decisions for the people. This is much easier to do if political dissension is suppressed. Hayek's claim was that, to run a fully socialized planned economy successfully, its socialist managers ultimately must secure control of the political process as well.

Hayek's book was only one of many at the time to address the issues of planning versus markets and other issues related to the shape of the post-war economic and political order. Its fame, and in some quarters notoriety, was due to its being condensed in the pages of *Reader's Digest* in April 1945, appearing just as the war in Europe was coming to an end. *Reader's Digest* then had a circulation of almost nine million, and in addition, a Book of the Month Club reprint was made available that added another million readers. As a result, Hayek's little book, and the even smaller condensed version, gained widespread attention and iconic status among both its supporters and critics.

Besides fame, the publication of the book brought with it other unintended consequences. On a publicity trip to the United States, Hayek made a number of contacts, people who shared his views regarding the merits of a liberal democratic market order. In 1947 he organized the first meeting of the Mont Pèlerin Society, which brought together like-minded people from America and Europe to discuss and debate questions concerning the appropriate economic, political, legal and social institutional framework for a free society. Participants included Milton Friedman, Aaron Director and George Stigler, who would over the course of the next decade form the Chicago School of economics.

The sensory order

From 1945 until he joined the faculty at Chicago, Hayek took on yet another wholly different subject, theoretical psychology. Building on a student paper he had completed in 1920, he titled the resulting book *The Sensory Order* (Hayek, 1952a).

This book is probably best viewed as an outgrowth of his earlier attack on scientism. Two ‘objectivist’ doctrines that he criticized in the ‘scientism’ essay were physicalism, a view espoused by the logical positivist philosopher Otto Neurath, and behaviourist psychology. The doctrines were related: physicalism insists that all truly scientific statements make reference only to observables, and behaviourist psychology likewise insists that scientific psychology should eschew all reference to mental states and deal only with observable behaviour. By eliminating all reference to subjective states and interpretations, the objectivity of science is guaranteed.

Hayek posited two orders, the sensory order that we experience, and the underlying natural order that natural science has revealed: atoms, molecules, electromagnetic waves and the like. The question arises: why are these two orders different? Hayek's answer was that the sensory order is in fact a product of our brain. He characterized the brain as a highly complex but self-ordering, hierarchical classification system, a huge network of connections. A given stimulus triggers an extensive set of neuronal firings that gives rise to our experience of a sensation. The richness of our sensory experience is due to the sheer vastness and hierarchical nature of the classifier system. As he once noted, ‘During a few minutes of intense cortical activity the number of interneuronic connections actually made (counting also those that are actuated more than once in different associational patterns) may well be as great as the total number of atoms in the solar system (that is, 10^{56})’ (Hayek, 1964, p. 25).

If Hayek's description was right it posed problems for behaviourists, who did not even recognize the

existence of the two orders, taking the sensory order as fundamental. Furthermore, the supposedly uninterpreted sensory experience so vital to the behaviouralist was itself simply a product of our minds; it was itself an interpretation. Hayek's book went virtually unnoticed when published, but subsequent neuroscientific research broadly supports his principal claims.

Political theory

J.M. Keynes read Hayek's *The Road to Serfdom* on a boat going to the Bretton Woods conference, later writing to Hayek that ‘morally and philosophically I find myself in agreement with virtually the whole of it; and not only in agreement with it, but in a deeply moved agreement’ (Keynes, 1944, p. 385). Keynes went on to say, though, that

> You admit here and there that it is a question of knowing where to draw the line. You agree that the line has to be drawn somewhere, and that the logical extreme is not possible. But you give us no guidance whatever as to where to draw it. (1944, p. 386)

Hayek evidently took the criticism to heart, for in the coming years he would make two further important contributions to political philosophy that would refine and extend the arguments made in *The Road to Serfdom*.

In *The Constitution of Liberty* Hayek defined *liberty* as a condition ‘in which coercion of some by others is reduced as much as possible in society’ (Hayek, 1960, p. 11). This poses a dilemma, because the best way to avoid coercion is to set up a coercive power that is strong enough to suppress it. Liberal constitutionalism attempts to solve the problem by defining a private sphere of acceptable individual activity, granting the state a monopoly on coercive powers, then constitutionally limiting the power of the state to those instances where it is required to prevent coercion. The state's coercive actions are limited by the rule of law: its laws made in protection of the private sphere must be prospective, known, certain, and equally enforced (Hayek, 1960, pp. 205–10). He contrasted these with laws that seek particular outcomes within the private sphere, for example, price-fixing to help certain groups, or social legislation whose intent is to create or preserve a particular pattern of redistribution. Hayek linked his discussion with his perennial concern for problems caused by dispersed knowledge by noting how liberty enables individuals to make the best use of local knowledge:

> The rationale of securing to each individual a known range within which he can decide on his actions is to enable him to make the fullest use of his knowledge, especially of his concrete and often unique knowledge of the particular circumstances of time and place. The law tells him what facts he may count on and thereby extends the range within which he can predict the consequences of his action. (Hayek, 1960, pp. 156–7)

In the last third of the book Hayek outlined the specific sorts of government policies that are consistent with constitutional liberalism.

Soon after completing this book he felt the need to readdress some of the same questions, ultimately producing the trilogy *Law, Legislation and Liberty* (1973–79). There Hayek lamented how Western

democracies were increasingly circumventing the constitutional constraints outlined in his earlier book. Because the ideals of constitutionalism had failed to take root, the rule of law was weakening. Governments were increasingly passing coercive legislation, typically under the guise of achieving social justice, that in reality typically served well-organized coalitions of special interests. Coercive legislation was gradually replacing the rule of law.

Hayek began by contrasting spontaneous, self-generating orders (what the Greeks called a *kosmos*) with organizations that are constructed, created orders (what the Greeks referred to as a *taxis*). Agents in organizations aim at accomplishing specific goals, and do so by following explicit commands. Grown orders tend to be much more complex. They do not aim at specifiable outcomes, and agents interact in them by following abstract rules. Hayek applied these ideas to the development of the law, or *nomos*, in which rules of just conduct eventually become codified into law. He contrasted this common law heritage with legislation, the rules for organizing government, also known as *thesis*. Under the influence of various rationalist constructivist doctrines (Hayek identifies utilitarianism and legal positivism as particularly noxious), legislation to achieve particular ends began to replace the grown law, which itself does not aim at specific outcomes but instead provides a stable ordered environment in which individuals are able to employ their knowledge to make decisions.

In developing these contrasts, Hayek argued that though the concept of justice provides the foundation for notions of just conduct and ultimately of the law itself, the idea of *social justice* only has meaning within the context of a *taxis*. Only human conduct by individuals or organizations, not states of affairs or outcomes, can be called just or unjust. One must be able to hold someone responsible to apply the term. Rationalist constructivists make a fundamental error, a category mistake, to argue that it can also be applied to the outcomes of a spontaneous process, which has no specific purpose other than to allow millions of agents to pursue their own purposes. Hayek ended his trilogy with the pessimistic view that majoritarian democratic governments operating under the errors of constructivism and the guise of achieving greater social justice were increasingly replacing grown law with legislation, most of which served powerful special interests, with dire consequences for the persistence of the grown order. In the final chapter he proposed a unique political reform that aimed at increasing the independence of legislators from the influence of special interests, thereby strengthening the ideal of liberal constitutionalism. Interestingly, about the same time Hayek (1978a) also proposed an equally provocative scheme for the competing currencies that he dubbed the denationalization of money.

His final major contribution was *The Fatal Conceit* (Hayek, 1988), the conceit being socialism – for Hayek the ultimate form of rationalist constructivism. The book had its origins in the late 1970s, when he tried to arrange a debate between socialists and advocates of markets on the merits of their respective systems. Though the debate never came off, the project led him to begin work on a final wide-ranging critique of socialism and constructivism. Hayek worked on the book during the early 1980s, but when his health began to fail in 1985 the philosopher W.W. Bartley III (who was also the general editor of *The Collected Works of F.A. Hayek*) stepped in to assist him. Questions have been raised about how much of the book should be attributed to Bartley and how much to Hayek, but one fundamental Hayekian claim is that the moral rules, norms, ethical precepts and practices that have led to the development of the extended market order have emerged through a process of cultural evolution. Many of these rules go against the ‘natural morality’ that allowed earlier humans to function successfully in small hunter-gather groups. Furthermore, because they were seldom consciously adopted and their effects are often difficult to identify, they tend to chafe against human reason, as well. Many of our moral beliefs, then, lie

between, and fit uneasily with, both our instinct and our reason. This is why humans instinctively rebel against the market order, and seek to use their reason to construct an alternative.

A theme that runs throughout Hayek's work is an emphasis on the limits of our reason, and the role of rule-following in allowing us to deal successfully in a world in which knowledge is dispersed. In field after field Hayek identified spontaneous complex orders that form as the result of agents following rules. The price system represents one such an order, and, as his work on capital theory showed, if one extends the system in time it can also serve as a mechanism for the intertemporal coordination of human action. The brain is another example of a self-organizing complex order: vast networks of neuronal firings give rise to the larger phenomenon of consciousness. Within political theory, the common law tradition (as opposed to legislation) and the requirement that we follow the rule of law and obey constitutional rules are yet another manifestation of our discovering procedures that allow us to deal more successfully with the limits of our reason.

It is unfortunate that Hayek remains in some quarters a controversial figure, but it is also probably inevitable, given that so many of his key insights were formed within a context of intense political debate, and that it is difficult to separate them from that context. Even so, one hopes that his contributions on knowledge and its limits, on the role of grown institutions in helping us to overcome our ignorance, and on the workings of hierarchical networks and spontaneous self-organizing complex orders, will continue to stimulate future research.

See Also

- Austrian economics
- Friedman, Milton
- Mises, Ludwig Edler von
- Robbins, Lionel Charles
- socialist calculation debate

Bibliographical note: the book series *The Collected Works of F.A. Hayek*, published jointly by the University of Chicago Press and Routledge, is currently in process of production. The series consists of annotated versions of all of Hayek's major works, as well as supplementary material. Each volume contains an extensive editorial introduction to set the work in context. In the list of selected works below, references are made to those volumes of the *Collected Works* edition that have already appeared, otherwise to the original edition.

Selected works

1926. Monetary policy in the United States after the recovery from the crisis of 1920. In *Good Money*, Part I, ed. S. Kresge; vol. 5 of *Collected Works*, 1999.

1928. Intertemporal price equilibrium and movements in the value of money. In *Good Money*, Part I, ed. S. Kresge; vol. 5 of *Collected Works*, 1999.

1931. *Prices and Production*, 2nd edn. London: Routledge, 1935.

1933. *Monetary Theory and the Trade Cycle*. New York: Kelley, 1966.

1935. *Collectivist Economic Planning: Critical Studies on the Possibilities of Socialism*. London: Routledge.

1937. Economics and knowledge. Repr. in Hayek (1948).

1941. *The Pure Theory of Capital*. Chicago: University of Chicago Press.

1942–44. Scientism and the study of society. Repr. in Hayek (1952b).

1944. *The Road to Serfdom*. In *The Road to Serfdom: Text, Documents*, ed. B. Caldwell, vol. 2 of *Collected Works*, 2007.

1945. The use of knowledge in society. Repr. in Hayek (1948).

1948. *Individualism and Economic Order*. Chicago: University of Chicago Press.

1952a. *The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology*. Chicago: University of Chicago Press.

1952b. *The Counter-Revolution of Science: Studies on the Abuse of Reason*. Glencoe, IL: Free Press.

1960. *The Constitution of Liberty*. Chicago: University of Chicago Press.

1964. The theory of complex phenomena. Repr. in Hayek (1967).

1967. *Studies in Philosophy, Politics and Economics*. Chicago: University of Chicago Press.

1968. Competition as a discovery procedure. Repr. in Hayek (1978b).

1973–9. *Law, Legislation and Liberty*, 3 vols. Chicago: University of Chicago Press.

1978a. The denationalisation of money. In *Good Money*, Part II, ed. Stephen Kresge, vol. 6 of *Collected Works*, 1999.

1978b. *New Studies in Philosophy, Politics, Economics and the History of Ideas*. Chicago: University of Chicago Press.

1988–. *The Collected Works of F.A. Hayek*, Chicago and London: University of Chicago Press and Routledge.

1988. *The Fatal Conceit: The Errors of Socialism*. In *Collected Works*, vol. 1, ed. W.W. Bartley III, 1988.

1994. *Hayek on Hayek: An Autobiographical Dialogue*, ed. S. Kresge and L. Wenar. Chicago: University of Chicago Press.

Bibliography

Caldwell, B. 2004. *Hayek's Challenge: An Intellectual Biography of F.A. Hayek*. Chicago: University of Chicago Press.

Cassidy, J. 2000. The price prophet. *New Yorker*, 7(February), 44–51.

Dickinson, H.D. 1933. Price formation in a socialist economy. *Economic Journal* 43, 237–50.

Hicks, J.R. 1967. The Hayek story. In *Critical Essays in Monetary Theory*. Oxford: Clarendon Press.

Keynes, J.M. 1930. *A Treatise on Money*. Repr. as vols 5 and 6 of *The Collected Writings of John Maynard Keynes*. London: Macmillan, 1971.

Keynes, J.M. 1931. *The Pure Theory of Money: A Reply to Dr. Hayek*. Repr. in *Contra Keynes and Cambridge: Essays, Correspondence*, ed. Bruce Caldwell, vol. 9 of *The Collected Works F.A. Hayek*, 1995.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. Repr. as vol. 7 of Keynes (1971–89), 1973.

Keynes, J.M. 1944. *Letter, J.M. Keynes to F.A. Hayek, June 28, 1944*. Repr. in *Activities 1940–46. Shaping the Post-War World: Employment and Commodities*, vol. 27 of Keynes (1971–89), ed. D. Moggridge, 1980.

Keynes, J.M. 1971–89. *The Collected Writings of John Maynard Keynes*, 30 vols, ed. A. Robinson and D. Moggridge. London: Macmillan for the Royal Economic Society.

Lange, O. 1938. On the economic theory of socialism. In *On the Economic Theory of Socialism*, ed. B. Lippincott. Minneapolis: University of Minnesota Press.

Mises, L.V. 1920. *Economic Calculation in the Socialist Commonwealth*, tr. S. Adler. In Hayek (1935).

Mises, L.V. 1924. *The Theory of Money and Credit*, tr. H.E. Batson. London: Jonathan Cape, 1934.

Wicksell, K. 1906. *Lectures on Political Economy*, vol. 2, tr. E. Classen. London: Routledge, 1935.

How to cite this article

Caldwell, Bruce. "Hayek, Friedrich August von (1899–1992)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000030> doi:10.1057/9780230226203.0711

The New Palgrave Dictionary of Economics Online

hazardous waste, economics of

Hilary Sigman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Public policies for hazardous waste address both current waste management and the clean-up of a legacy of contamination. Policies for current waste management should provide incentives for waste generators that are sensitive to the varying hazards posed by waste. Although conventional regulations have difficulty accomplishing this variation, alternative incentive-based policies show promise empirically. Policies for clean-up of contamination often fail to strike an appropriate balance between hazards posed by the contamination and costs of clean-up. In addition, relying on legal liability to fund these clean-ups has consequences for the costs of clean-up and possibly for the redevelopment of contaminated land.

Keywords

brownfield sites; compensation; environmental economics; environmental equity; environmental externalities; environmental liability; Environmental Liability Directive (EU); hazardous waste management; hazardous waste regulation; hazardous waste, economics of; hedonic property values; legal liability; strict liability; Superfund (USA); waste-end taxes

Article

Hazardous waste has become a major focus of environmental regulation. In the United States, spending on hazardous waste rose from only about two per cent of the compliance cost of all federal environmental regulations in 1985 to a projected 17 per cent in 2000 (U.S. EPA, 1990); its share of environmental expenses in Europe may have been comparable in 2000 (European Commission, 2000). The costs arise from the management of waste from current activities and from the clean-up of a legacy of contamination. This article addresses, first, current hazardous waste management and, second, clean-up of past contamination.

Hazardous waste management

A wide range of industrial processes create hazardous wastes; they are also generated by commercial activities (for example, used oil and batteries from automobile repair shops) and by households (for example, used electronics). Once generated, wastes are managed in one of three ways: disposal, which usually involves placing wastes in landfills or injecting liquid wastes into underground wells; treatment (for example, incineration or stabilization), which renders the wastes less hazardous, but rarely eliminates the need to dispose some hazardous residuals; and recycling or reuse. Most hazardous waste is managed on-site by the small number of plants that generate vast quantities of waste; however, most generators create smaller quantities and use off-site commercial waste management.

The risk posed by a waste depends not only on the nature of the hazardous chemicals but also on their concentration and mobility and on the way the waste is managed. Waste generators control these variables through their output decisions, production processes, and handling of wastes, so the challenge for public policy is to create optimal incentives in all these dimensions. An efficient policy would correct many different environmental externalities, such as air pollution from incineration and groundwater and surface water contamination from land disposal. Such a policy might use taxes on environmental releases or, where feasible, impose legal liability for harms. With a competitive waste management market, these policies would not only affect waste management but also send the correct signals to waste generators to choose how much and which sorts of waste to generate.

Actual public policy for hazardous waste in developed countries tends to regulate waste management, with relatively few direct rules about the quantities or nature of wastes generated. Regulations use a traditional command-and-control approach, often requiring specific technologies (for example, specifying the thickness of liners required for hazardous waste landfills). These approaches do not provide much flexibility in tailoring the management methods to the characteristics and risks of the waste in question. For example, the United States requires that wastes be treated (often incinerated) before land disposal. These ‘land disposal restrictions’ probably account for most of the cost of hazardous waste regulations, yet economic assessments by the U.S. Environmental Protection Agency (EPA) strongly suggest that their costs greatly exceed their benefits (Sigman, 2000). The land-disposal restrictions founder on their absolute nature; although the restrictions would pass a cost–benefit test for some wastes, they also apply to many wastes that are not easily treated or pose low hazards.

Some jurisdictions also impose taxes on hazardous waste. Sometimes called ‘waste-end’ taxes, these taxes vary with the quantity of waste and may depend on how waste is managed. Sigman (1996) reports empirical evidence that generators respond to waste-end taxes by reducing waste and altering their management methods. Levinson (1999) shows that the waste-end taxes levied by the United States have altered the geography of waste management. Although this evidence demonstrates the potential of incentive-based environmental policies to motivate private decisions, current taxes in the United States do not seem efficient. Indeed, Levinson (2003) argues that states practise destructive competition, specifically an inefficient ‘race to the top’, through these taxes.

Enforcement of both command-and-control waste regulation and waste-end taxes is difficult. Unlike air and water pollutants, hazardous waste is easily transported away from its source, giving rise to the possibility of illegal disposal (known as ‘midnight dumping’ in the United States and ‘fly tipping’ in the UK). Sigman (1998a) finds that rules requiring recycling or reuse of waste raise the frequency of illegal dumping. If the elasticity of illegal disposal to legal waste management costs is high enough, public policies may be counterproductive because the environmental harm from illegal disposal can be much

greater than that from legal disposal.

A response to this enforcement problem is to use a deposit-refund or similar tax–subsidy combination (Fullerton and Kinnaman, 1995). The policy would tax inputs or products that give rise to waste and give a refund for legal waste management that may vary with the external costs of the waste management method. For example, it might refund a modest portion of the initial tax for land disposal and a larger share for incineration. Such a system could mimic waste-end taxes without the incentives for illegal disposal. The effective tax for illegal disposal is the forfeit of the deposit; for all other activities, the effective tax is the difference between the deposit and appropriate refund. Deposit-refunds (with the refund equal to the deposit) are common internationally for products such as used batteries, electronics, and lubricating oil (OECD/EEA, 2006).

The location of facilities that manage waste also raises issues. Local communities often reject these facilities because of their perceived hazards. Economists have sought to design policies that create optimal incentives in siting facilities through compensation for host communities (for example, Minehart and Neeman, 2002). In assessing ‘environmental equity’, numerous studies examine whether poor people and members of minority groups disproportionately live near hazardous waste facilities, with mixed conclusions for developed countries (Hamilton, 2005). Concern about the incidence of waste management costs may also be behind the Basel Convention of 1989, which restricts international shipment of hazardous wastes between developed and developing countries.

Clean-up of contaminated sites

Land disposal of hazardous wastes and other activities, such as storage of toxic substances for industrial processes, have created a legacy of contaminated sites that may cause damage and require clean-up. The U.S. federal policy for clean-up of abandoned contaminated sites is the Comprehensive Emergency Response Compensation and Liability Act (CERCLA), commonly known as Superfund. Superfund clean-up is mostly funded by legal liability imposed on the generators and transporters of waste and past and present owners of contaminated land. Liable parties must either undertake clean-up themselves or reimburse the government for clean-up by paying into a dedicated trust fund, which also received some tax financing in earlier years. European countries have similar programmes; a 2004 EU Environmental Liability Directive imposed additional requirements.

The appropriate level of clean-up (or ‘how clean is clean?’) has been the subject of a long-running debate. In the early years, public policies called for sites to be rendered completely clean; however, as costs have grown, greater (or at least more explicit) balancing of benefits and costs has become common. Still, decisions often fall well short of the economist's ideal. For example, the Superfund programme sets goals that reflect biases in risk perception and political objectives as well as costs and risks to the exposed population (Hamilton and Viscusi, 1999).

Quantifying the benefits of clean-up can be difficult. A substantial literature uses hedonic property value methods to evaluate the welfare effects of proximity to contaminated sites and finds large effects. In a literature survey by the U.S. EPA (2005), the studies on average find that house prices are seven per cent (or more) lower near contaminated sites. Studies also find that discovery of a site lowers local house prices. Although these results suggest that households perceive harm from contaminated sites, the much smaller literature that looks at whether clean-ups improve prices finds disappointing results; for

example, Greenstone and Gallagher (2005) conclude that Superfund clean-up had minimal effect on house prices.

Another debate concerns the desirability of funding clean-up through legal liability. This funding source has advantages and disadvantages relative to use of general government revenues or a dedicated tax. Liability may create desirable incentives for current waste management, reducing the need for *ex ante* regulation of land disposal, such as the command-and-control regulation discussed above. However, unless the required clean-up spending is optimal, these incentives may not be efficient. In addition, much liability is retroactive (applying to contamination from before the clean-up law) and thus does not directly affect active waste management.

Liability also helps privatize clean-ups because liable parties may dispatch their responsibility by undertaking government-approved clean-ups. Relative to the government, private parties may have stronger incentives to control costs, better knowledge of the contamination, and greater ability to limit disruption to current economic activity at the site. However, in an effort to lower their costs, private parties may drag their feet and use political and other means to make the government choose less extensive clean-up remedies (Sigman, 1998b; 2001).

Many policymakers fear that environmental liability deters redevelopment of ‘brownfields’, sites with potential contamination from their past use. Brownfields are a concern because they are a source of urban blight and because firms may develop relatively pristine land as a substitute for old industrial land. A buyer of a contaminated site may find itself partially or fully liable for clean-up costs; in the United States, CERCLA lists current landowners as among the potentially responsible parties. The government may choose to pursue a new owner, for example, if it has deeper pockets than the previous owner or is a private rather than a public entity. However, it is unclear that such liability deters redevelopment because it may be capitalized into land prices, as empirical research suggests (McGrath, 2002).

A number of distortions may make price adjustments insufficient to restore efficient incentives for redevelopment of brownfields. Segerson (1993) argues that a distortion can arise when sellers are judgment-proof (sheltered from liability by the option of declaring bankruptcy), so a sale may increase collective private clean-up costs by exposing a buyer's deeper pockets to the government. Other studies point to adverse selection, imperfect enforcement of liability, and the effects of joint and several liability as sources of a disincentive for redevelopment of brownfields (Boyd, Harrington and Macauley, 1996; Chang and Sigman, 2005). Empirical research does suggest higher vacancy rates for urban industrial land where expected liability is higher (Sigman, 2006). Numerous public policies address brownfields, for example, by providing liability protections and direct subsidies to new owners of land.

Finally, liability incurs substantial legal costs, as the government sues liable parties and liable parties sue each other and their insurance carriers. Based on surveys of liable parties, Dixon (1995) estimates that as much as 30 per cent of private spending on Superfund will be transaction costs. However, these costs may be similar to the transaction costs of tort liability generally and the excess burden of the taxes that might replace liability as a funding source.

The specific form of environmental liability has also been controversial. In the European Union (EU), a debate on strict liability preceded the adoption of the 2004 Environmental Liability Directive. Under strict liability, parties are liable for any harm whereas under an ‘at fault’ or negligence rule parties are liable only when they fail to exercise appropriate care. The EU settled on a mixed regime in which certain hazardous activities are subject to strict liability. An extensive literature in law and economics

compares these liability regimes. For hazardous waste, empirical studies have suggested that strict liability reduces accidental toxic spills and violations of hazardous waste laws (Alberini and Austin, 2002; Stafford, 2003).

Rules for apportioning liability with multiple defendants have been even more controversial. In the United States, courts have interpreted Superfund to impose ‘joint and several’ liability, meaning that the government may sue any liable party for the entire cost of clean-up at the site, regardless of that party's contribution to the contamination; the party initially held liable may then recover cost shares from other defendants. Most European countries also rely on joint and several environmental liability, but some have begun restricting its use. Joint and several liability strengthens the government's hand and increases its ability to collect ‘orphan shares’, the share of costs attributable to parties that are bankrupt or substantially judgment-proof. It also affects incentives for parties to settle rather than litigate, which may be favourable depending on empirical conditions (Kornhauser and Revesz, 1994; Chang and Sigman, 2000), and incentives for *ex ante* precaution in managing hazardous substances (Tietenberg, 1989). Although some of its effects may be desirable, joint and several liability is often decried as unfair.

See Also

- cost–benefit analysis
- environmental economics
- liability for accidents
- Pigouvian taxes
- value of life

Bibliography

Alberini, A. and Austin, D. 2002. Accidents waiting to happen: liability policy and toxic pollution releases. *Review of Economics and Statistics* 84, 729–41.

Boyd, J., Harrington, W. and Macauley, M. 1996. The effects of environmental liability on industrial real estate development. *Journal of Real Estate Economics and Finance* 12, 37–58.

Chang, H. and Sigman, H. 2000. Incentives to settle under joint and several liability: an empirical analysis of Superfund litigation. *Journal of Legal Studies* 29, 205–36.

Chang, H. and Sigman, H. 2005. The effect of joint and several liability under Superfund on brownfields. Working Paper No. 11667. Cambridge, MA: NBER.

Dixon, L. 1995. The transactions costs generated by Superfund's liability approach. In *Analyzing Superfund: Economics, Science and Law*, ed. R. Revesz and R. Steward. Washington, DC: Resources for the Future.

European Commission. 2000. *Study on Investment and Employment Related to EU Policy on Air, Water and Waste*. EC 4739/M/11452–0. Online. Available at http://europa.eu.int/comm/environment/enveco/industry_employment/investment_and_employment.htm, accessed 14 April 2006.

Fullerton, D. and Kinnaman, T. 1995. Garbage, recycling, and illicit burning or dumping. *Journal of Environmental Economics and Management* 29, 78–91.

Greenstone, M. and Gallagher, J. 2005. Does hazardous waste matter? Evidence from the housing market and the Superfund program. Working Paper No. 11790. Cambridge, MA: NBER.

Hamilton, J. 2005. Environmental equity and the siting of hazardous waste facilities in OECD countries: evidence and policies. *International Yearbook of Environmental and Resource Economics 2005/2006*, ed. H. Folmer and T. Tietenberg. Cheltenham: Edward Elgar.

Hamilton, J. and Viscusi, W. 1999. *Calculating Risks? The Spatial and Political Dimensions of Hazardous Waste Policy*. Cambridge, MA: MIT Press.

Kornhauser, L. and Revesz, R. 1994. Multidefendant settlements: the impact of joint and several liability. *Journal of Legal Studies* 23, 41–76.

Levinson, A. 1999. State taxes and interstate hazardous waste shipments. *American Economic Review* 89, 666–77.

Levinson, A. 2003. Environmental regulatory competition: a status report and some new evidence. *National Tax Journal* 56, 91–106.

McGrath, D. 2002. Urban industrial land redevelopment and contamination risk. *Journal of Urban Economics* 47, 414–42.

Minehart, D. and Neeman, Z. 2002. Effective siting of waste treatment facilities. *Journal of Environmental Economics and Management* 43, 303–24.

OECD (Organisation for Economic Co-operation and Development)/EEA (European Environment Agency). 2006. Database on Economic Instruments and Voluntary Approaches Used in Environmental Policy and Natural Resources Management. Online. Available at http://www2.oecd.org/ecoinst/queries/index.htm, accessed 12 April 2006.

Segerson, K. 1993. Liability transfers: an economic assessment of buyer and lender liability. *Journal of Environmental Economics and Management* 25, S64–S65.

Sigman, H. 1996. The effects of hazardous waste taxes on waste generation and disposal. *Journal of*

Environmental Economics and Management 30, 199–217.

Sigman, H. 1998a. Midnight dumping: public policies and illegal disposal of used oil. *RAND Journal of Economics* 29, 157–78.

Sigman, H. 1998b. Liability funding and Superfund clean-up remedies. *Journal of Environmental Economics and Management* 35, 205–24.

Sigman, H. 2000. Hazardous waste and toxic substance policies. In *Public Policies for Environmental Protection*, 2nd edn, ed. P. Portney and R. Stavins. Washington, DC: Resources for the Future.

Sigman, H. 2001. The pace of progress at superfund sites: policy goals and interest group influence. *Journal of Law and Economics* 44, 315–44.

Sigman, H. 2006. Environmental liability and the reuse of old industrial land. Working paper. Rutgers University.

Stafford, S. 2003. Assessing the effectiveness of state regulation and enforcement of hazardous waste. *Journal of Regulatory Economics* 23, 27–41.

Tietenberg, T. 1989. Indivisible toxic torts: the economics and joint and several liability. *Land Economics* 65, 305–19.

U.S. EPA (Environmental Protection Agency). 1990. *Environmental Investments: The Cost of a Clean Environment*. Washington, DC: EPA.

U.S. EPA. 2005. *Superfund Benefits Analysis* (Draft). Online. Available at http://www.epa.gov/superfund/news/benefits.htm, accessed 12 April 2006.

How to cite this article

Sigman, Hilary. "hazardous waste, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000262> doi:10.1057/9780230226203.0712

The New Palgrave Dictionary of Economics Online

health economics

Barbara Wolfe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Health care expenditures form an ever-increasing burden in most developed countries, especially the United States, where they accounted for 16.0 per cent of GDP in 2004, up from 5.1 per cent of GDP in 1960. These cost increases alone suggest that health economics is a dynamic field of economic research, but the importance and the interest of the field are driven by broader considerations. This article delineates important market failures and research issues in health and health care, the relationship between income and health, methodological issues in the measurement of health, and quality issues in the measurement of health care.

Keywords

absolute deprivation–health hypothesis; income–health hypotheses; adverse selection; asymmetric information; crowding out; deadweight loss; diffusion of technology; disability, economics of; gradient effects; health capital; health economics; health insurance; health maintenance organizations (HMOs); hospital economics; managed competition; market failures in health care; Medicaid (USA); medical practice licensing; Medicare (USA); moral hazard; mortality; non-profit organizations; nutrition; payroll tax; pharmaceutical industry; RAND Health Insurance Experiment; tax expenditures; well-being

Article

Health care expenditures form an ever-increasing burden in most developed countries. Between 1960 and 2002, expenditures as a percentage of GDP rose from 3.8 to 9.7 per cent in France, from 5.4 to 9.6 per cent in Canada, and from 4.9 to 11.2 per cent in Switzerland. (Cross-country comparisons are hindered to some extent by the differential inclusion of components of health care and social services in different countries.) In the United States, health care expenditures are far greater; they accounted for 16.0 per cent of GDP in 2004 (a per capita expenditure of $6,280), up from 5.1 per cent of GDP in 1960. By themselves these increases, projected to rise to 20 per cent or more of GDP by 2015, suggest that

health economics is a dynamic field of economic research, but the importance of the field is driven by far more than the costs of care.

In many ways, health economics mimics the broader field of classical economics in its areas of research specialization – there are theoretical studies, micro and macro studies, industrial organization studies, public economic studies, and labour studies, among others. But health economics has a unique quality, identified in one of the earliest papers in the field, by Kenneth Arrow (1963). That is the large role played by *market failures*, which make it likely that resources will be allocated inefficiently if market outcomes alone prevail.

Market failures

Market failures take several forms.

Failures related to information

Individuals tend to be poor judges of the care they need and the quality of care they obtain. This ignorance works both *ex ante* and *ex post*. Consumers do not know whether they might benefit from medical care. They tend to lack information about the appropriate type, amount, quality, and price of care. They also lack information about the counterfactual. would alternative care, or even no care at all, be equally or more or less beneficial and cost-effective? The rapid introduction of new technology and the need to make decisions under stress tend to result in ineffective information searches.

Failures related to the role of supplier agents

Providers acting for the patient are rarely perfect agents. they do not fully understand individual patient preferences, their own earnings are influenced by their advice (conflict of interest), and their training tends to impel them to do all that is technologically possible (that is, provide care until expected marginal benefits equal zero). The usual remedy for this kind of failure is public sector interventions in the health care market. These include licensing of providers to assure a minimum level of competence; licensing of facilities and new technologies (including pharmaceuticals) to assure quality; reimbursement schemes to minimize conflicts of interest; subsidies for certain types of care (for example, those with external benefits, such as vaccinations); and subsidies for the purchase of insurance.

Failures related to uncertainty

The combination of uncertain need for care and the high expense associated with a major health problem leads naturally to a demand for pooling risk and insurance. But, as in most types of insurance, willingness to pay for coverage based on individual assessments of one's own (differing) risks may lead to adverse selection (see below) and incomplete coverage. For those insured, the reduction in the price of care may lead to increased demand for care, less attention to the price of care, and less attention to avoiding the need for care (‘moral hazard’), making it very difficult to estimate optimal amounts of care. Insured individuals may demand care beyond the point where the marginal expected benefit is equal to

the true cost of the resource, or substitute care for other health-preserving options such as exercise or diet. When insurance is combined with reimbursement schemes that pay providers for each service (‘fee for service’), the likely result is demand and provision of care in excess of expected benefits. (The idea of providers recommending care to the patient, knowing that the expected benefit to the patient is less than the cost, has been termed ‘physician-induced demand’. Though numerous articles have been written on the topic, proof of such behaviour remains elusive.) But reimbursement schemes that pay providers a fixed or capitated amount for all services may lead to too little care, based on a comparison of marginal cost to marginal benefits. Much recent work has focused on designing reimbursement schemes that mitigate overuse or underuse.

The demand for health

Michael Grossman (1972) was the first to emphasize that the outcome of interest in health economics is health, not medical care. The demand for medical care, in other words, is derived from the demand for good health. Grossman's model and all the work that derives from it are essentially extensions of the household production literature. In brief, Grossman argues that health capital is a form of human capital which changes over time because of depreciation and investment. His model begins with a utility function where health is an argument in addition to utility gained from the consumption of goods and services. Health investments include time (exercise and sleep) and medical care, subject to health endowments such as genetic traits or environmental factors that are known to the individual or family. Without investment in health, health deteriorates. Net investment in health equals gross investment minus depreciation, the rate of which is assumed to be exogenous and to increase with age, so that it becomes more and more expensive to obtain good health. There is also an education efficiency parameter. Higher health stock increases healthy time, leading to higher income (increased productivity and time to work), making income endogenous in the model. All individual choices are subject to both a time and money budget constraint. The time constraint requires that the total amount of time available in any period must be exhausted by all possible uses, which include time spent working, producing health, lost to illness, and spent in leisure activities. The income constraint explicitly includes expenditures on medical care.

Empirical estimates have used both reduced form and structural versions of this model to answer questions such as how much a tax on cigarettes will reduce low birthweight (reduced form), or how much a change in maternal smoking might influence birthweight (structural estimates of the marginal product). The models underlie some studies of unhealthy behaviours (such as alcohol abuse) and have extensive applications in environmental economics. Although there are always issues of endogeneity (for example, income, medical care) and hence of identification in empirical applications of this model, the model's improved focus on the nature of the demand for medical care and the important role of time allocation have had a major impact on the field. (Gerdtham et al., 1999, conducted one of the best empirical tests of the Grossman model using Swedish individual data.)

Income and health. estimating the relationship

Consistent with the empirical evidence, the Grossman model suggests that income is positively

associated with health, but the direction of causation is not clear. Those with better health earn more and hence have higher incomes, whereas those with higher incomes can invest more in better health, suggesting that the observed income–health gradient may best be modelled as a simultaneous system.

The idea that income is associated with health goes back a long way in the health economics literature, and a number of hypotheses have been advanced to explain the relationship. Samuel Preston (1975) observed that the impact of additional income on health (as measured by mortality) is greater for those with lower income than for those with higher income. This observation of diminishing marginal productivity is called the ‘absolute income’ hypothesis. In its simplest form, it argues that, if income is all that matters to individual health, a community with more equal income will tend to have better average health than a community with more inequality, even if the two communities have the same average income. In an international context, Angus Deaton (2002) points out that, according to the absolute income hypothesis, redistribution can improve health even if average income is not increased, and that redistribution from rich to poor countries would in principle improve worldwide average health. A related concept is the ‘absolute deprivation’ or poverty hypothesis, which suggests that those with the lowest incomes face poorer health and a greater risk of mortality owing to inadequate nutrition, poor-quality health care, exposure to physical hazards, and heightened stress. According to this hypothesis, a dollar redistributed from rich to poor would improve the health of the poor and improve the average health of the entire population.

The ‘relative income’ hypothesis focuses on an individual's income relative to others in his or her group. If the incomes of all members but one in a group increase, that one person's health is expected to deteriorate. A related, ‘relative position’ hypothesis holds that one's rank (occupation or education) in society is tied to health outcomes. Research in the United States and the United Kingdom has demonstrated an association between socio-economic position and health (Mullahy, Robert and Wolfe, 2004, review the evidence for this). Referred to as a ‘gradient effect’, this hypothesis implies that psycho-social and other factors that remain unevenly distributed all the way up the income scale perpetuate income inequalities in health. Perceptions of being relatively deprived (‘keeping up with the Joneses’), stress, and other non-material factors may play a role in perpetuating income inequalities in health at the upper income levels. The distinction between absolute and relative income effects has important policy implications; if the income of everyone were to increase or decrease, no change in health would be expected under a relative income model, but change would be expected under an absolute income model.

The hypothesis that focuses most directly on the tie between inequality in both health and income has two versions, one ‘strong’ and one ‘weak’. According to the strong version of the income inequality hypothesis, if the average income of the society is held constant, societies with greater inequality produce worse health among their citizens. Those in the most unequal communities may fear for their lives and property whether they are poor or wealthy, or the stress of keeping up with the Joneses may reduce time allocated to producing health. The weaker version argues that those with incomes below the mean will be negatively influenced by greater income inequality, perhaps through higher residential density and the associated increases in crime and contagious disease. A related issue for research is the extent to which these observed ties between income and health can be ascribed to race or ethnicity, through the systematically differing average incomes of racial and ethnic groups in a country like the United States. But if groups also differ in diet and genetics, health may be causally linked to these other factors rather than to the observed income gradient.

In general, there have been two empirical approaches to examining the hypotheses described above. Research into the absolute deprivation, relative income, and relative position hypotheses has usually relied upon individual-level data on income and health or mortality to examine the existence and shape of the income–health relationship among individuals; research on race and ethnicity follows a similar approach. In contrast, research examining the income inequality hypothesis has employed aggregate data exclusively, at least in measuring income inequality.

Insurance

As noted above, uncertainty in the need for medical care and the potentially very large costs of care lead to a demand for health insurance. (Nyman, 1999, has suggested an additional motivation for demanding health insurance. having insurance permits consumers to consume very expensive care that would otherwise be beyond their budget constraints.) But costs of operation, inability of the insurer to accurately discern risks (information asymmetry), and the potential for increased expenditures on the insured mean that insurance may not be supplied at a price reflecting an individual's actuarially fair cost. A traditional market failure may occur, in which only those with high expected medical expenditures are willing to buy coverage (adverse selection). Or there may be complete failure of the insurance market because no (risk-neutral) supplier is willing to offer coverage at a price that any individual would willingly pay. This has led to publicly provided coverage in many countries and subsidies towards the purchase of insurance in others, such as the United States. It has also led to incomplete insurance, in which deductibles, co-payments, and co-insurance attempt to reduce moral hazard. (A deductible requires that some initial level of expenditures is covered directly by the consumer; a co-payment is usually a fixed dollar payment per specified unit of service; and co-insurance is a fixed percentage payment. In general, although co-payments are quite common – for example, in pharmaceutical coverage – they have poorer incentive effects than coinsurance.)
The insurance market raises several interesting issues.

1. 1. *The role of secondary insurance*. Secondary insurance may reduce the cost of public coverage if privately financed care is sufficiently substituted for publicly financed care, or may raise the cost of public coverage if it primarily pays for the cost-sharing components of publicly financed care. In the United States, for example, ‘Medigap’ insurance covers deductibles and co-payments required under Medicare, the system covering those aged 65 or more and the significantly disabled; it thereby increases demand for those services primarily paid for by Medicare.
2. 2. *The efficiency and equity of subsidizing the purchase of insurance through the tax system*. The US system provides the highest subsidies to those with the highest marginal tax rates (those with high incomes) and offers little or no subsidy to those with low incomes.
3. 3. *The incentive effects of income-conditioned eligibility for public insurance*. To be eligible for the Medicaid programme in the United States, persons have to meet state-specified eligibility requirements linked to income, assets and family structure. There is an all-or-nothing eligibility requirement, such that a person with income or assets a dollar above the cut-off is ineligible for Medicaid. Three potential consequences are less work effort (reduced earnings) by individuals who wish to become or remain eligible, increased numbers without private insurance among the near-eligible who are in effect ‘insured’ for costly care since they could become eligible if they

needed such care, and reduced savings among those eligible and near-eligible.

4. 4. *The welfare loss from public subsidies for private insurance or public provision of medical care*. This is potentially higher the more services that are covered, for example, covering all new technology rather than only that which passes some benefit–cost analysis, covering all pharmaceuticals compared with only the least costly drug within a category, or covering all types of counselling rather than that tied to a diagnosis of severe mental illness.
5. 5. *Issues tied to optimal breadth of coverage*. Such issues include, for example, whether universally provided or mandated insurance increases welfare and the optimal depth of that coverage, life-threatening emergency care but not cosmetic surgery, semi-private hospital rooms but not private rooms, treatment for cancer but not dementia.
6. 6. *The labour market implications of using a payroll tax to subsidize or directly provide health insurance coverage*. This policy potentially increases the costs to employers of hiring additional workers, especially older workers or those who have a chronically ill family member. Most modern economies struggle with how best to design employer-based taxation to fund health care insurance.
7. 7. *How to minimize crowd-out in countries with both public and private health insurance systems*. In the United States the issue is designing public insurance coverage to minimize incentives to turn down private coverage (see issue 3 above). In the Netherlands the issue is balancing payroll taxes against private insurance premiums so that younger and healthier earners will not find ways to join the private system in lieu of the public system.
8. 8. *Designing policies to increase take-up of insurance among eligible populations within the constraints of equity and efficiency*, thus avoiding unnecessary subsidies for those already enrolled.
9. 9. *How to effectively cover the treatment of mental illnesses*. A particular variant of this in the United States is the role and design of mental health parity laws.

The demand for medical care

Much research has focused on empirically estimating the elasticity of demand for medical care. This question gets renewed attention whenever there are proposed changes to insurance coverage, since the cost of such policy changes lies in the elasticity of demand. (An additional or second-order cost of any expansion of insurance via public policy is the magnitude of the social welfare loss – deadweight loss – associated with the moral hazard effect. The most-cited original contribution on this is Mark Pauly, 1968.) The simplest empirically estimated models (most of which use number of visits or units of care as the dependent variable) include the price of care, income, simple demographic factors, and the price of alternative goods and services. But accurately measuring the marginal cost of care is not a simple proposition for individuals with insurance coverage that includes deductibles, co-payments or co-insurance, maximums per episode, and otherwise incomplete coverage. More satisfactory models include the value of time (including time spent in care, time spent in transit, and time spent waiting). Most empirical research differentiates the demand for hospital stays from physician services. A number of studies narrow the question still further, to the level of the market for hospital or physician services or for individual physicians. As expected, demand elasticities for individual physicians are quite high in large markets (suggesting a competitive market), whereas elasticities for physicians as a whole or for

hospitals tend to be far lower, suggesting their considerable market power.

All non-experimental estimates of the demand for care using US populations suffer from the endogeneity of the marginal price of care because the demand for insurance is endogenous to the demand for medical care; that is, individuals or families with the highest expected medical expenses will seek out relatively generous coverage. In a large-scale experiment, the RAND Health Insurance Experiment (see Newhouse, 1994), individuals were randomly assigned to various plans ranging from full coverage (free care) to care with a 95 per cent coinsurance and a maximum dollar contribution. The design of the experiment was such that researchers could more accurately assess elasticities of demand and the marginal price of care. Participants were observed for 3–5 years. The study had shortcomings (some by design). it excluded those with the highest demand (the elderly and disabled), experienced attrition especially among those with the least generous plans, and made the decision to pay participants a lump sum to be sure all families were made no worse off by participation in the experiment. The experiment convincingly established that individuals do respond to price, even for hospital care. (The experiment found that 86.7 per cent of those with full coverage used care in a given year compared with 68 per cent of those with 95 per cent coverage. medical expenditures for the year, in 1984 dollars, were $777, or 32.8 per cent, versus $534, or 27.4 per cent, respectively.)

Substantial econometric issues in estimating demand elasticities remain and have spawned much methodological research in health economics. Four main issues are. (*a*) the highly skewed nature of utilization and expenditure distributions in populations of interest (a multipart – usually two-or four-part – model in which first the probability of any use, or particular use such as outpatient care, and then the level of use conditional on any use has been frequently used in the literature. These models may not be readily identifiable, however; Duan et al., 1983.); (*b*) the episodic nature of care; (*c*) whether to use quantity of care as the outcome of interest (and, if so, how to include dimensions of the duration, extent, and quality of care) or to use expenditures (and, if so, how to measure actual expenditures rather than billed amounts); and (*d*) how to capture the nature of demand, much of which is based on ill health, which is difficult to measure accurately.

Measuring health outcomes

Given the prominence of market failure in health economics, accurate measures of health to capture the effectiveness of medical care are crucial. (Defining health is a major problem that lies largely outside the domain of the health economist. Perhaps the most often used concept is that promoted by the World Health Organization – a complete state of well-being. This is not very useful for the measurement of health.) Mortality is the ‘health’ measure most commonly discussed – and arguably most precisely measured – in this literature, but the relationship of mortality to care may be quite distant in time as well as ‘coarse’ or ‘noisy’. Other measures of health, which may have more proximate temporal relationships to medical care, are necessary. Such non-mortality measures of health (‘biological well-being’) may be of many types. cellular or molecular (such as measles antibodies or titres); clinical (for example, body mass index); functional (for example, indices such as Activities of Daily Living scores); self-rated (for example, on a scale from excellent to poor); medical providers’ diagnosis of particular physical or mental health diseases, or tied to activities (such as days of school missed). Many measures are straightforward while others attempt to capture ‘utility’ by asking individuals to compare a state of illness to days or years of total healthiness to create indices such as quality adjusted life years (QALYs).

The measures selected must be appropriate for the purpose. For instance, some measures of health may be largely determined by genetic factors (such as risk for schizophrenia), others closely related to opportunity costs (for example, days of school missed because of parental work schedules). A measure appropriate for the analysis of labour force participation (for example, poor or fair health versus good or better health) may not be suitable for evaluating a targeted intervention. And, as with other empirical work, measures should detect changes (vary), measure what they intend to measure (be valid), and be free from error (be reliable).

Measurement issues in evaluating policy

Although the RAND Experiment established that consumers are responsive to the price of care, their results did not establish the value or effectiveness of care in influencing health. Nor did they determine how we might socially evaluate the benefits from a change in policy (such as an increase in the proportion of people eligible for public coverage) or the safety and efficacy trade-offs involved in designing regulation of new technologies. A large ‘industry’ exists to try to determine whether individual well-being improves, deteriorates, or remains largely unchanged after some change in policy or intervention. Critical issues include how to measure and value health changes and how to account for individual differences.

These issues tie health economics to core questions in public finance centring on the evaluation of policy changes using cost–benefit analysis, cost-effectiveness analysis, and multi-attribute utility analysis. Health economics has a further question, however. Is only the social perspective relevant, or is the perspective of a more narrowly defined group (payers, patients, providers) relevant also? And which population or individuals should be included in the analysis – only those currently alive, the next generation, or those who might survive only because of the intervention? The perspective adopted may determine whether to provide an intervention or invest in a new technology. Measuring costs is also very complex; calculation is rendered difficult by skewed cost distributions, interdependent costs, and economies. In many countries the pricing of new technologies and drugs is an additional and growing concern in policy design. (A related issue is the rate of change of technology, including pharmaceuticals. Decisions made on adoption and pricing modify the incentives for private investment in developing new drugs and equipment. Challenging issues include. Who should take the risk, particularly at early stages of development? How to encourage investment in new technology for diseases that have limited markets? How to share the benefits of new technology with those with limited incomes (including those in poor countries)?

Quality issues in evaluating care

Research concerning health care providers ranges from the role of licensing and malpractice to the use of quality indexes both broad (for example, hospital report cards) and narrow (for example, risk-adjusted mortality indices for interventions such as coronary artery bypass grafts or high-risk deliveries). The concept of ‘pay for performance’ is gaining credibility among both private and public payers, but the implications for quality and distribution of care and for overall expenditures need further research. Preliminary studies suggest that both practices – quality indicators and pay for performance – pose a real

danger of cream skimming; the ability to obtain high-quality care may diminish for those with the most to gain from such care. The medical malpractice system aims to signal the appropriate amount and quality of care, but to work properly it requires that all who suffer significant injury through medical care bring a legal action, that rewards or settlements reflect true utility-based losses, and that rewards are not paid where the cost of prevention exceeds the full loss. Economists continue to debate whether the system leads to defensive medicine (too much care) and to evaluate the implications of tort reform for quality of care.

Supply side issues

Providers

Do we have enough medical providers? Restrictions on entry into the profession make this an interesting question. In order to practise medicine one must be licensed, and in order to be licensed one must have a medical degree from an accredited institution. The supply of medical schools and student enrolment are regulated in nearly all countries. As Eli Ginzberg (1989, p. 88) noted, ‘Neither the restrictive policies of the first four decades of [the 20th] century, nor the expansionary policies of the postwar era were formulated and implemented on the basis of demand and supply of physician services’. Rate of return calculations indicate, in general, that rates are high for specialists and lower for primary care doctors, but, because entry into specialties is limited by available residencies, such analysis cannot fully answer the question of supply. Current research on the adequacy of supply extends beyond numbers per capita to include primary versus specialty care, geographic distribution, design of public subsidy, and repayment schemes. The substitutability of other health professionals including nurses, nurse practitioners, and social workers for medical doctors (or psychologists for psychiatrists) is tied to two main issues. whether we have sufficient medical providers and whether competition and lower reimbursement may increase the efficiency of some types of medical care, many of them routine and predictable. A sufficient supply of nurses is from time to time a particularly acute issue. Issues of interest in the market for nurses include the monopsony power of hospitals, working conditions including work hours and locations, and childcare during working hours, which tend to fall outside the standard 8 a.m. to 5 p.m. workday.

Hospitals

Much earlier work in health economics focused on hospitals, memorably described by Baumol and Bowen (1966, p. 497, referring to non-profit organizations in general) as ‘bottomless receptacles into which limitless funds can be poured’. Various alternative reimbursement schemes have been designed to limit these expenditures, but all have difficulties. Hospitals paid on the basis of *fee-for-service* have an incentive to provide care wherever marginal benefits are positive (especially if patients are fully insured.) From an efficiency perspective, this leads to too much care, because costs are not considered. Hospitals paid according to a *fee schedule* have an incentive to over-provide services for which the fee is greater than or equal to the marginal cost but to skimp on other services. Hospitals paid a *per diem* have an incentive to extend patient stays, especially since marginal costs tend to be far lower later in a stay.

Hospitals paid by *diagnosis* (such as in Medicare's diagnosis-related groups or DRGs) have the incentive to serve the healthiest of those who seek care (cream skimming), avoiding those for whom expected costs are greater than expected payments (dumping). Finally, hospitals paid on the basis of *capitation*, or that are part of a fully owned health maintenance organization (HMO) with a capitation-based income, face incentives to provide cost-effective care, but perhaps not all care with positive net benefits.
Understanding hospital behaviour is important for designing policies that influence their behaviour. Many, but far from all, hospitals are non-profit, and use their non-profit status to convey the message to patients that quality is not compromised by the desire for profits. They thus aim to generate trust that reduces the need for complicated contracts between the hospital and the consumer. (Hospitals are increasingly adding for-profit components to their array of services, thus masking the difference between for-profit and non-profit institutions.) The non-profit nature of most hospitals has led to a variety of models of hospital behaviour. One model views hospitals as two organizations in one. There is first the hospital staff, which provides resources to the physicians for the care of their patients. The physician staff want sufficient resources to treat their patients without delays and prefer some excess capacity; they want the latest technology, however expensive. Hospitals provide such technology in order to compete for physicians and their patients. The result is duplication and rapid diffusion of the newest technology. A second model is a utility maximization model of hospitals, in which hospital managers get utility from the increased quantity (size or number of beds) and quality of care provided. They can expand in both dimensions more easily within non-profits and with fewer binding constraints than within for-profit hospitals. Again the result is duplication. (A related version of this is a quantity maximization model.)
Two other models of hospitals suggest (*a*) that physicians control hospitals, behaving in ways that maximize their own incomes, or (*b*) that hospitals can be thought of as physician cooperatives that act to maximize their own well-being but are inefficient if the hospital grows too large. Some research on forms of ownership and hospital behaviour suggests that competition matters far more than form of ownership where price is concerned.

Managed competition

First put forward by Alain Enthoven (1978), this approach changed the incentives facing providers to improve efficiency and quality. The plan called for multi-specialty group practices that would provide a specified, comprehensive set of medical care services in exchange for a per capita prospective payment covering a defined period of time. Individuals could choose an HMO plan (usually a closed panel or a limited set of providers) or traditional fee-for-service; all bidders would be required to offer a plan that covered the specified set of services. Employers would offer a broad set of plans but would pay only a fixed dollar amount towards the premium; consumers would pay the full difference between that contribution and the actual premium. The lower cost and more comprehensive benefits of the prepaid plans would lead consumers to choose those plans. Information on the quality of care under these plans would be systematically collected and shared with consumers, who would have an annual open enrolment period.
Empirical evidence on the effectiveness of managed competition suggests that it generates one-time savings but that the forces driving toward new technology, the desire among consumers for point-of-service choice, and adverse selection have limited its role. Reform, analysis and experimentation with

variants of managed competition continue. The issue of how to reduce the rate of increase in the cost of medical care continues to be important in all developed countries. Managed competition is but one approach. Others include. limiting the number of providers who can practise in a jurisdiction (Canada and the United Kingdom, for example); increasing the co-payments required of consumers; modifying reimbursement of providers; using waiting lists to reduce access; providing free telephone advice to improve efficiency of demand (Australia and parts of the United Kingdom); regulating insurance coverage; setting a budget for a fixed period of time; and rationing care on the basis of age. Designing these approaches and evaluating their success is a continuing challenge.

The economics of disability

In this area, health economists have been concerned with the proper design of public policy, in particular the efficiency of disability-based benefits, including their work and health insurance incentives. Measurement (in this case, of disability) is again a major impediment to the quality of the research. Health economists are attempting to better understand the determinants of chronic health problems, including those such as obesity, asthma, and diabetes, which are on the increase. Models from other fields of economics (such as intergenerational mobility, time use, and consumer demand models) are being applied to understand the determinants of the increase and to identify interventions to stem the trend.

This short overview suggests that problems for exploration and opportunities to influence the design of policies from prevention, through insurance design, to reimbursement and regulation are likely to expand as the costs of health care continue to rise. Health economics approaches range from theory through empirical analysis to policy reform. They can include the development of new models or improved policy analysis and design. They can be country-specific or comparative; sector-specific (hospital care) or more comprehensive; and tied to labour, public, micro theory, or international studies.

See Also

- adverse selection
- crowding out
- health insurance, economics of
- market failure
- tax expenditures
- technology
- uncertainty

Thanks to Rashi Fein, Robert Haveman, John Mullahy, Nathan Tefft and Jason Fletcher for helpful suggestions.

Bibliography

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53, 941–73.

Baumol, W. and Bowen, W. 1966. *Performing Arts –The Economic Dilemma*. New York: Twentieth Century Fund.

Deaton, A. 2002. Policy implications of the gradient of health and wealth. *Health Affairs* 21(2), 13–30.

Duan, N., Manning, W., Morris, C. and Newhouse, J. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics* 1, 115–26.

Enthoven, A. 1978. A consumer choice health plan: a national health insurance proposal based on regulated competition in the private sector. *New England Journal of Medicine* 298, 650–8 and 709–20.

Gerdtham, U.-G., Johannesson, M., Lundberg, L. and Isacson, D. 1999. The demand for health: results from new measures of health capital. *European Journal of Political Economy* 15, 501–21.

Ginzberg, E. 1989. Physician supply in the year 2000. *Health Affairs* 8(2), 84–90.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80, 223–55.

Mullahy, J., Robert, S. and Wolfe, B. 2004. Health, income and inequality. In *Social Inequality*, ed. K. Neckerman. New York: Russell Sage Foundation.

Newhouse, J. and Health Insurance Experiment Group. 1994. *Free for All? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

Nyman, J. 1999. The value of health insurance: the access motive. *Journal of Health Economics* 18, 141–52.

Pauly, M. 1968. The economics of moral hazard: comment. *American Economic Review* 58, 531–7.

Preston, S. 1975. The changing relation between mortality and level of economic development. *Population Studies* 29, 231–48.

How to cite this article

Wolfe, Barbara. "health economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/

article?id=pde2008_H000031> doi:10.1057/9780230226203.0713

The New Palgrave Dictionary of Economics Online

health insurance, economics of

Joseph P. Newhouse
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Health care finance has been dominated by moral hazard, potential rents and the deadweight loss from financing them, and adverse selection. Public health services and insurance tend to be universal, solving the selection problem. Private health insurance markets and public schemes that offer a choice of insurance plans generally exhibit selection. Research has found strong evidence of responsiveness of demand to insurance coverage. In health insurance markets information is asymmetric among patients, providers, and insurers, and principal–agent relationships abound. Actual health insurance and health care financing institutions have adapted to these features.

Keywords

adverse selection; Akerlof, G.; Arrow, K.; asymmetric information; deadweight loss; employment-based medical insurance; group insurance; health insurance: private vs. public; insurance contracts; managed care insurance plans; medical care, state provision of; Medicare; moral hazard; principal-agent relationship; rent; risk aversion

Article

As the capabilities and the associated expense of medicine advanced during the 20th century, the demand for financial protection against the risk of large medical spending grew. The result of the increased demand has been widespread health insurance or direct public provision of medical care, or both, in every developed country.
Both health insurance and the public provision of medical care heavily subsidize that care at the point of service, meaning that the user bears only a fraction (usually a small one) of the cost. As a result, insurance induces moral hazard and potentially rents in factor prices as well, which in turn induces deadweight loss through the taxes needed to finance any public insurance. Both private and public insurers, however, may combat moral hazard and rents through the nature of their contracts with

providers, and sometimes through command-and-control type intervention.

In addition to moral hazard and potential rents, a voluntary health insurance market exhibits adverse selection. To combat selection, voluntary insurance contracts may contain provisions for medical underwriting (exclusion of an individual from a small group insurance plan), exclusions of pre-existing conditions from coverage, or exclusion of certain services from coverage altogether. Such features may prevent willing buyers and sellers from contracting, as well as hamper the efficiency of labour markets with employment-based insurance, as employees may not leave jobs because of the inability to obtain comparable insurance (Gruber, 2000).

These three features – moral hazard, potential rents and the deadweight loss from financing them, and adverse selection – have influenced countries' health care financing institutions, as first suggested in the seminal paper by Kenneth Arrow (1963). To combat selection, some countries, for example the United Kingdom and southern Europe, deliver health services through a public health service. In this case health insurance markets are supplemental to the public health service. Other countries, such as Canada and many northern European countries, combat selection by offering public or quasi-public insurance with no choice of insurance plan; private health insurance markets are again supplemental. Because public health services and public health insurance tend to be universal (though in some countries the affluent can opt out), the selection problem is solved, potentially at the expense of a poorly performing monopoly.

By contrast, private health insurance markets and public schemes that offer a choice of insurance plans generally do exhibit selection. Among developed countries the United States relies most on private health insurance markets, although because of selection American private health insurance is primarily organized through employment rather than through an individual insurance market. Again in part because of selection, many Americans without an employment connection, most notably the elderly, are insured through public insurance. But not all American employers offer insurance, not all those employees offered insurance purchase it, and not all those without a labour market connection are eligible for public insurance. As a result, the United States has a much higher proportion of its population with no health insurance than other developed countries. This group tends to receive care through subsidized direct delivery systems.

In this article I first discuss selection and the demand for insurance. Then I discuss moral hazard and the demand for medical care conditional on insurance. Finally, I discuss the nature of the contract between the insurer and the provider, such as the physician or hospital, and its relation to the provider's supply of services. The material covered in this entry is treated more extensively in Cutler and Zeckhauser (2000).

Selection

Health insurance was used as an example of selection by two of the classic papers on asymmetric information (Akerlof, 1970; Rothschild and Stiglitz, 1976). The models in those papers showed that an equilibrium may not exist in competitive insurance markets if insurers could not identify high risks (for example hypochondriacs and possibly the chronically ill). If a competitive market pooled high and low risks (that is both risk types buying the same policy), insurers would offer products that differentially appealed to low risks, thus breaking the pooling equilibrium. Under certain conditions a separating equilibrium (that is each risk type buying a different policy) might also be impossible.

Later papers showed that under different assumptions an equilibrium might exist (Dubey and Geanakoplos, 2002; Newhouse, 1996; Wilson, 1977), but selection behaviour seems pervasive in individual health insurance markets, and even a separating equilibrium is a form of market failure given the almost universal nature of annual contracts in private insurance markets. That is because a low risk this year has the risk of becoming a high risk next year, for example by contracting a chronic disease. But the higher premium facing a high risk is uninsurable with annual premiums. Notice that in a family insurance context this risk includes having a high-risk child born into the family, and, when the child becomes an adult, extends also to the child (typically the child can no longer be covered under a parent's policy after a certain age).

Cochrane (1995) pointed out that lifetime contracts would solve this problem, but lifetime contracts are not observed in the private market, for several reasons. First, the rapid rate of technological change in medicine and the associated cost increase (Cutler, 2004; Newhouse, 1992) is a non-diversifiable risk for a given cohort. Second, there are large economies to group insurance, in part because of lower marketing costs. To avert selection, however, groups must be formed primarily for reasons other than obtaining health insurance, which is why health insurance in many countries forms around the place of employment. But even here selection can be a problem, most obviously for the self-employed, who are effectively in an individual market, but also for small employers.

Empirical work has confirmed the importance of selection. I show three examples. The first is the variation in insurance premiums by generosity of the insurance. Table 1 orders insurance policies by generosity, where a higher percentile indicates more complete coverage. The column labelled ‘Premium’ is the premium charged for the policy; the column labelled ‘Actuarial value’ is the estimated spending among a standardized population for each policy, reflecting the increase in demand for services when the insurer covers a greater portion of any medical spending. (One might ask how actuarial values are known. They can come from similar policies from groups where selection is minimal, such as employees of large firms with no choice of insurance plan, or from other evidence on how the demand for care varies with insurance generosity; one form of such evidence is discussed below in conjunction with moral hazard.)

Cost and actuarial value of insurance policies

Percentile	Individual policy ($)		Family policy ($)	
	Premium	Actuarial value	Premium	Actuarial value
10	1,220	1,740	2,760	4,220
25	1,670	1,910	3,950	4,600
50	2,100	2,100	5,070	5,070
75	2,620	2,260	6,090	5,450
90	3,220	2,440	7,670	5,890
Difference 90–10	164%	40%	178%	40%

Source: Cutler (1994, Table 2).

Whereas premiums between the 90th and 10th percentile plans differ by a factor of about 2.7, the actuarial value of the two plans differs only by a factor of 1.4. The difference between these values indicates that high risks are disproportionately choosing the 90th percentile plan, and low risks are

disproportionately choosing the 10th percentile plan.

A second form of evidence comes from the US Medicare programme, the near-universal insurance programme for individuals over the age of 65. Medicare gives its beneficiaries choice between a traditional indemnity insurance plan that allows free choice of physician, and a prepaid plan, which can restrict choice of physician but in return charges a smaller premium or covers certain additional services. Through 2005 individuals have been allowed to change between these two types of plans monthly (under current law this is to change in 2006).

The data suggest the traditional indemnity plan has been more appealing to high-risk individuals. Although spending and use data have not been available for those in prepaid plans, one can compare use among those in the traditional plan who subsequently enrol in a prepaid plan with that of those who do not. Adjusted for age and sex and a few other covariates, spending among those switching from the traditional plan to a prepaid plan in the 12 months before they switched was 23 per cent less than among those who remained in the traditional plan (Medicare Payment Advisory Commission, 2000). Because the individuals had the same insurance plan when this difference was observed, the group opting to change to the prepaid plan appeared to be in considerably better health. Similarly, adjusted for age and sex, mortality rates among those enrolled in prepaid plans were 15 per cent less than among those in the traditional plan, a difference much too large to be plausibly related to any difference in benefits or care. Consistent with selection, the mortality rate difference was largest, 21 per cent, among those who had switched to the prepaid plan within the previous 12 months, and then steadily narrowed as individuals remained in the prepaid plan, a form of regression to the mean.

Finally, selection can give rise to so-called premium death spirals. Cutler and Reber (1998) studied a natural experiment among Harvard University employees that gave rise to one such spiral. Harvard allowed its employees to choose among insurance plans of varying generosity. Initially it subsidized a constant percentage of the premium (between 75 and 85 per cent, depending on the employee's earnings), but it subsequently changed the subsidy to a lump sum. (I note in passing that the rationale for an employer subsidy is to combat selection. In effect, such a subsidy makes it attractive for low risks to pool with high risks within the employment group.)

With a percentage-of-premium subsidy, the employee bore only 15–25 per cent of the premium difference among plans, but with a lump sum subsidy the employee bore the full incremental cost of more generous plans. For the employee who only marginally favoured a more generous plan with a percentage-of-premium subsidy (that is, the better risk within the group choosing the more generous plan), it became attractive with a lump sum to choose a less generous plan (the relative price to the employee of more generous plans rose by a factor of four or more). But, as the better risks within the more generous plans opted out, the premium necessary to cover the medical cost of those remaining rose. This in turn set off another round of plan changing, which raised the premium in the more generous plans still more. At that point the most generous plan was withdrawn from the market.

Moral hazard

Insurance creates a trade-off between risk aversion and moral hazard, or the failure to take actions that would lessen the probability of or the severity of or damage from an adverse event (Zeckhauser, 1970). In the context of health insurance, the focus has been on the costliness of the event rather than on the

likelihood of the event. That is because there are enough unpleasant uninsured consequences around illness and injury, such as pain and discomfort, that the extent of insurance for medical care probably changes the incentive to avoid illness or injury rather little. In fact, individuals when randomly assigned to better insurance do not change their lifestyle habits (Newhouse and Group, 1993).

The RAND Health Insurance Experiment randomized 2000 American families to health insurance plans that varied the portion of medical bills they had to pay, from nothing (all services were free to the family) to approximately a large family deductible of $1,000 in late 1970s dollars (Manning et al., 1987; Newhouse and Group, 1993). The deductible was scaled down for low-income families. The families were followed for either three or five years (the period was randomly assigned), and both their medical care use and health outcomes were observed. Over the course of the experiment families assigned to the large deductible plan used around 30 per cent fewer services than those assigned to the plan in which services were free. They made about two fewer visits to physicians during the year, and they were admitted to the hospital about 20 per cent less.

The average family's health was little changed by the additional medical services consumed when care was free to them, although those who were both sick and poor had better outcomes, primarily because of better control of hypertension (high blood pressure). There was thus ample confirmation of moral hazard. For a review of studies of moral hazard see Zweifel and Manning (2000).

The supply of medical care

Insurers or health care services either contract with or employ health care providers, most notably physicians, to deliver medical services. The need for this is most apparent if the insurance policy covers all of the patient's medical costs in full; because the patient has no incentive to search for a lower-cost provider, if the insurer passively reimbursed medical bills providers could in theory bill an infinite amount. The same incentives apply if patients bear a modest fixed charge for each, say, physician visit. The terms of the insurer's contracts with medical providers can have important effects on the services delivered. I discuss here two features of such contracts: provider networks and so called supply-side cost sharing. (For further discussion of these issues see Chalkley and Malcolmson, 2000; McGuire, 2000; Newhouse, 2002; and Pauly, 2000.)

Before the 1980s the usual model of American health insurance allowed ‘free’ choice of provider, meaning that to a first approximation the patient's out-of-pocket payment was unaffected by the physician(s) he or she sought care from. Many non-American models still allow this. These arrangements began to change in the 1980s with the advent of managed care insurance plans, which sought to establish networks of preferred physicians and hospitals. In-network physicians contracted with the insurer; at a minimum the contract specified a discount off a usual fee. In return, patients were given financial incentives to use in-network physicians. By increasing the elasticity of demand facing a physician, networks began to reduce rents in physician fees (Cutler, McClellan and Newhouse, 2000).

Many insurance plans went further than simply asking for fee discounts, and gave physicians financial incentives to reduce utilization. For example, instead of being paid a fee for each narrowly defined service (such as a visit or a laboratory test), a primary care physician might receive a capitation payment for each insured who selected that physician as a personal physician. In this case, the proximate marginal revenue the physician earned from delivering any additional services was zero, although there

might have been an indirect effect of reducing services on patient retention. A less high-powered incentive than pure capitation was a ‘risk corridor’ around a target utilization rate. For example, the physician and the insurer might share deviations from the target rate fifty-fifty up to a certain size deviation, and above or below that amount all risk was on the insurer.

Such arrangements, also used in the British National Health Service and in Denmark, were termed ‘supply-side cost sharing’, in contrast to the demand-side cost sharing described above that was paid by the patient. Supply-side cost sharing created two incentives. First, and most obviously, it created an incentive for the physician to treat less intensively than with pure fee-for-service reimbursement, thereby potentially addressing moral hazard (Chalkley and Malcolmson, 1998; 2000; Ellis and McGuire, 1986). Second, if the capitation were only for the primary care physician's own services, the most common arrangement, it created an incentive for that physician to refer the patient to another physician, a form of unbundling. In the American context the insurer would simply pay for the services of the other physician; in the British context the patient would be referred to a salaried physician at a hospital that might have a lengthy queue.

Several pieces of evidence support the view that physicians respond to the type of contract they face. A near-universal finding of American managed care plans is that they reduce the use of services relative to traditional indemnity plans, which passively paid a fee-for-service; that is, they indemnified the patient against any incurred medical bills (Glied, 2000). As described above, however, a characteristic of managed care plans is that they did not passively reimburse whatever service a physician chose to deliver.

Other data show the effects of particular contracts. A natural experiment in Denmark showed that physicians whose services were only partly at risk delivered more services than those who were fully at risk, although the increase was not sustained (Krasnik et al., 1990). A small-scale experiment in the United States randomized pediatricians to be paid by either a fee-for-service system or a salary; the salaried pediatricians earned no additional income for treating more intensively (Hickson, Altmeier and Perrin, 1987). The physicians paid with a fee-for-service method delivered more than 20 per cent more services, with the difference almost entirely in well-child visits (preventive care). Most likely, mothers brought sick children in, but brought well children in only with some effort from the pediatrician.

Finally, numerous studies have shown that physicians respond to the level of payment (McGuire, 2000; Newhouse, 2002). This is particularly relevant to insurers that administratively set prices on a take-it-or-leave-it basis and to insurers that negotiate fees for all physicians in an area.

In sum, health insurance markets violate numerous assumptions of the introductory textbook model of perfectly competitive markets with full or at least symmetric information. In particular, information is asymmetric among patients, providers, and insurers, and principal–agent relationships abound. Actual health insurance and health care financing institutions have adapted to these features.

See Also

- agent-based models
- contract theory
- market competition and selection

- risk sharing
- social insurance
- tragedy of the commons

Bibliography

Akerlof, G. 1970. The ‘market for lemons’: qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics* 74, 488–500.

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53, 941–73.

Chalkley, M. and Malcolmson, J. 1998. Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics* 17, 1–19.

Chalkley, M. and Malcolmson, J. 2000. Government purchasing of health services. In *Handbook of Health Economics*, 1A, ed. A. Culyer and J. Newhouse. Amsterdam, North-Holland.

Cochrane, J. 1995. Time consistent health insurance. *Journal of Political Economy* 103, 445–73.

Cutler, D. 1994. A Guide to health care reform. *Journal of Economic Perspectives* 8(3), 13–29.

Cutler, D. 2004. *Your Money or Your Life: Strong Medicine for America's Health Care System*. New York: Oxford University Press.

Cutler, D., McClellan, M. and Newhouse, J. 2000. How does managed care do it? *RAND Journal of Economics* 31, 526–48.

Cutler, D. and Reber, S. 1998. Paying for health insurance: the tradeoff between competition and adverse selection. *Quarterly Journal of Economics* 113, 433–66.

Cutler, D. and Zeckhauser, R. 2000. The anatomy of health insurance. In *Handbook of Health Economics*, 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Dubey, P. and Geanakoplos, J. 2002. Competitive pooling: Rothschild–Stiglitz reconsidered. *Quarterly Journal of Economics* 117, 1529–70.

Ellis, R. and McGuire, T. 1986. Provider behavior under prospective reimbursement. *Journal of Health Economics* 5, 129–51.

Glied, S. 2000. Managed care. In *Handbook of Health Economics*, 1A, ed. A. Culyer and J. Newhouse.

Amsterdam: North-Holland.

Gruber, J. 2000. Health insurance and the labor market. In *Handbook of Health Economics*, 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Hickson, G., Altmeier, W. and Perrin, J. 1987. Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study. *Pediatrics* 80, 344–50.

Krasnik, A. et al. 1990. Changing remuneration systems: effects on activity in general practice. *British Medical Journal* 300 (30 June), 1698–701.

Manning, W. et al. 1987. Health insurance and the demand for medical care: results from a randomized experiment. *American Economic Review* 77, 251–77.

McGuire, T. 2000. Physician agency. In *Handbook of Health Economics*, 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Medicare Payment Advisory Commission. 2000. *Report to the Congress: Improving Risk Adjustment in Medicare*. Washington, DC: Medicare Payment Advisory Commission.

Newhouse, J. 1992. Medical care costs: how much welfare loss? *Journal of Economic Perspectives* 6(3), 3–21.

Newhouse, J. 1996. Reimbursing health plans and health providers: selection versus efficiency in production. *Journal of Economic Literature* 34, 1236–63.

Newhouse, J. 2002. *Pricing the Priceless: A Health Care Conundrum*. Cambridge, MA: MIT Press.

Newhouse, J. and the Insurance Experiment Group. 1993. *Free for All? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

Pauly, M. 2000. Insurance reimbursement. In *Handbook of Health Economics*, 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Rothschild, M. and Stiglitz, J. 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90, 629–50.

Wilson, C. 1977. A model of insurance markets with incomplete information. *Journal of Economic Theory* 16, 167–207.

Zeckhauser, R. 1970. Medical insurance: a case study of the tradeoff between risk spreading and

appropriate incentives. *Journal of Economic Theory* 2, 10–26.

Zweifel, P. and Manning, W. 2000. Moral hazard and consumer incentives in health care. In *Handbook of Health Economics*, 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

How to cite this article

Newhouse, Joseph P. "health insurance, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000228> doi:10.1057/9780230226203.0714

The New Palgrave Dictionary of Economics Online

health outcomes (economic determinants)

Christopher J. Ruhm

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes the conclusions of research examining how health is affected by income inequality and temporary changes in macroeconomic conditions. In both cases, recent analyses have raised serious doubt about the ‘conventional wisdom’ derived from earlier studies using less adequate analytical methods. Specifically, the latest studies question the hypothesis that inequality has a large independent effect on population health, and suggest that economic downturns improve rather than worsen physical well-being.

Keywords

absolute income hypothesis; counter-cyclical; fixed effects; health outcomes; income inequality hypothesis; inequality of income; life expectancy; macroeconomic conditions; morbidity; mortality; pro-cyclical; relative income hypothesis; unemployment

Article

The health of individuals and populations is affected by a variety of economic factors. Other dictionary entries (health economics; population health, economic implications of) consider how health is related to income, education, infrastructural investments, prices and insurance. This article focuses on the roles of income inequality and of temporary changes in macroeconomic conditions in wealthy industrialized countries, which have been the subjects of considerable debate and empirical analysis. In both cases, the recent use of more sophisticated analytical approaches has raised serious doubt about the ‘conventional wisdom’ derived from earlier research using techniques less able to account for possible sources of bias. Income and health are positively correlated. Although there are questions about the extent to which higher incomes *cause* better health (Smith, 1999), most analysts believe that there is some causal effect and the discussion below presumes this is so. Similarly, *permanent* economic progress is assumed to improve most aspects of health.

Income inequality and health

Conceptual issues

There is a widespread belief that average health would improve if inequality could be reduced by redistributing income from richer to poorer households, without changing its average level. This is supported by the predictions of theory and empirical evidence, since at least Preston (1975), indicating that the health benefits of income exhibit diminishing returns. A direct consequence is that the health reductions resulting from lowering incomes of the well-off are more than offset by gains to the less advantaged, improving average health. Wagstaff and van Doorslaer (2000) call this the *absolute income hypothesis* (AIH). One important implication of AIH is that income inequality will be negatively related to average health in the cross section, but that this correlation will disappear if individual incomes are controlled for.

Far more controversial is the proposition that inequality has negative effects on health, with individual (or household) income held constant. This is the *income inequality hypothesis* (IIH). Under IIH, an individual living in a country with high inequality will be in worse health than a counterpart with the same income but residing in a nation with a more equal distribution. The main mechanism for this is hypothesized to be that relative income (or position) matters – this is the relative income hypothesis (RIH). For instance, being higher in the income distribution might allow access to goods that promote health and provide individuals with more control over their lives; conversely, low status might raise stress and reduce social trust or cohesion (Wilkinson, 1997).

The existence of RIH is not a sufficient condition for the income inequality hypothesis. An additional requirement is that the negative health effects of low rank exceed the gains (if any) accruing to high relative position. On the other hand, since some deleterious effects of increasing inequality (for example, loss of social capital) have adverse effects throughout the distribution, heightened inequality could worsen health of the wealthy as well as the poor. Direct tests of the relative income hypothesis are rare, partly due to the difficulty in defining an appropriate reference group. Advocates therefore often cite animal studies indicating that low-status primates have greater stress levels than their higher-status counterparts. Recently, however, some research provides direct evidence on the role of relative status. For instance, Eibner and Evans (2005) show high rates of mortality, morbidity, and body mass index, and poor self-reported health for persons whose incomes are low relative to a reference group defined by location, race, education and age.

Empirical evidence

Much interest in the income inequality hypothesis stems from Wilkinson's (1992) influential study. His most important finding was that more equal incomes were strongly positively correlated with life expectancy in nine industrialized countries, while average incomes had little effect. Subsequent research initially focused on cross-national comparisons but, starting in mid-1990s, was increasingly conducted for geographic regions within countries, particularly the United States.

The early studies were criticized on both technical and methodological grounds. For instance, Judge,

Mulligan and Benzeval (1998) discuss problems with the data used by Wilkinson (1992), and provide evidence that the results are sensitive to the choice of inequality measures. Moreover, as mentioned, ‘ecological studies’ using aggregate data will generally reveal a negative correlation between inequality and health (with average incomes controlled for) if income has diminishing benefits, and so cannot distinguish between the absolute income and income inequality hypotheses (Gravelle, 1998).

A potential solution is to use micro-data, since IIH predicts that the inequality relationship will persist after individual income is controlled for. Such research has proliferated in recent years, beginning with Fiscella and Franks (1997). These investigations do not, however, fully address the concern that cross-sectional inequality–health relationships may be confounded by omitted factors. Particularly significant, in this regard, is evidence that the association disappears or diminishes greatly when covariates are included for education, census region or per cent black (Mellor and Milyo, 2002; Deaton and Lubotsky, 2003). Other researchers (for example, Subramanian and Kawachi, 2003) argue that one or more of these variables may be caused by inequality and so are not appropriate to control for, or present evidence that a (diminished) correlation persists after including them. Nevertheless, the sensitivity of results suggests the vexing difficulty in accounting for all relevant causal factors when using cross-sectional data.

Whereas early studies provided strong evidence favouring IIH, the conclusions of more recent research, generally using better data and more sophisticated techniques, are far more mixed. An indication of this can be obtained from the comprehensive literature review by Lynch et al. (2004). They classified 98 peer-reviewed studies according to whether they supported IIH, had mixed findings, or obtained no associations or positive estimated effects of inequality on health. Overall, 40 studies contained strongly favourable evidence, 25 had mixed findings, and 33 were not supportive. However, 24 of the 37 studies published after 2001 failed to obtain results consistent with IIH, and just five were strongly favourable. The evidence is also generally less supportive when individual rather than aggregate data are used, particularly for recent analyses. For instance, of the 18 such studies published after 2001 and reviewed by Lynch et al., 12 obtained negative findings, two had mixed results, and only four strongly supported IIH.

After carefully reviewing the literature, Angus Deaton states, ‘The stories about income inequality affecting health are stronger than the evidence’ (Deaton, 2003, p. 150). This conclusion seems reasonable. There may be some causal effect but it is almost certainly weaker than that suggested by the early research, and is probably confined to a limited set of health outcomes (such as homicides).

Macroeconomic conditions and health

Conceptual issues

Health is conventionally believed to improve during economic expansions and deteriorate during downturns. Psycho-social determinants are usually focused upon, with recessions postulated to harm physical and mental health by increasing stress and risk taking (Brenner and Mooney, 1983). However, economic factors could also matter if, for example, incomes fall or medical care becomes more expensive because health insurance becomes less available or comprehensive.

However, there are at least four reasons why health instead might improve in bad times. First, the

opportunity cost of time declines, making it less expensive to undertake time-intensive health investments such as exercise and consumption of a healthy diet. Second, health is an input into the production of goods and services, implying that hazardous working conditions, job-related stress and some environmental risks (like pollution) may decrease. Third, some external sources of death may fall. For instance, traffic fatalities are likely to decrease due to reductions in driving. Fourth, migration may fall if individuals have fewer opportunities to move into areas with robust economic conditions. This could reduce social isolation, with especially beneficial effects on the young and old (Eyer, 1977).

The health effects of temporary and permanent changes in economic conditions may be quite different. An important distinction is that transitory growth is usually produced through more intensive use of existing inputs, whereas lasting improvements require some combination of technical innovation or increases in productive capital that have the potential to raise all types of consumption, including good health.

Time-series evidence

Most empirical research, until recently, analysed aggregate time-series data for a single geographic area. Particularly influential were a series of investigations by M. Harvey Brenner providing evidence that recessions (and other sources of macroeconomic instability) raise overall mortality, specific sources of death, and other health problems. For instance, using data from England and Wales for 1936–76, Brenner (1979) found that unemployment rates (growth in per capita income) were positively (negatively) related to total and age-specific mortality. However, researchers (such as Wagstaff, 1985) have pointed out serious technical flaws in Brenner's methods, and studies correcting the problems (for example, McAvinchey, 1988) failed to replicate his findings.

A key issue is that any lengthy time series may contain omitted variables that are spuriously correlated with economic conditions and have a causal effect on health. Potential confounders include changes in lifestyles, the public health infrastructure or medical technologies. Given this fundamental shortcoming, it is no surprise that the results of aggregate time-series analyses are sensitive to the countries, time periods, and proxies for health examined. After reviewing 16 such studies, Ruhm concludes: ‘with the exception of Brenner's analyses, the majority of the time series evidence suggests that the contemporaneous effect of economic downturns is to improve health or reduce mortality’ (Ruhm, 2006, p. 5). Interestingly, such ‘counterintuitive’ findings are not new. Research undertaken as early as the 1920s identifies a positive association between macroeconomic activity and mortality (Ogburn and Thomas, 1922).

Estimates using pooled data

A potential solution to the aforementioned shortcoming is to conduct ‘a more refined ecological analysis … taking advantage of local and regional variations in the business cycle as well as in disease rates’ (Kasl, 1979, p. 787). Research using such strategies has become increasingly common beginning with Ruhm's (2000) analysis of state-specific mortality rates for 1972–91. The key advantage of using multiple geographic areas and periods is that time effects can be included to account for potential time-varying confounders that have common impacts across locations, and location ‘fixed effects’ can be

added to control for unobserved factors that differ across geographic areas but remain constant within them. Researchers also often include area-specific time trends, which hold constant some other time-varying omitted determinants. Although most analyses have utilized aggregate data, the techniques can easily be adapted for use with individual data, and some research has begun to do so.

Mortality rates are the most commonly studied health outcome, and area unemployment rates are the typical proxy for macroeconomic conditions. Although data from the United States was first examined, analysis has recently been conducted for several European nations (for example, Neumayer, 2004; Tapia Granados, 2005), as well as using international data on multiple countries. Ruhm (2006) reviews seven such studies published between 2000 and 2004, and concludes that there is strong evidence of a *pro-cyclical* variation in total mortality, infant deaths and fatalities from traffic accidents, cardiovascular disease and influenza or pneumonia. The results are mixed for other types of mortality, and it is noteworthy that some studies uncover a counter-cyclical variation in suicides. This raises the possibility that people become ‘healthier but not happier’ when economic conditions deteriorate. Data restrictions have severely limited analyses of morbidities, although the majority of evidence from the few studies available (for example, Ruhm, 2003) indicates a counter-cyclical variation in health.

Lifestyle changes appear to explain some of the health improvements observed during economic downturns. Most available research suggests that alcohol use and problem drinking, smoking, severe obesity and physical inactivity all decline when the economy weakens. However, direct evidence on the role of work hours is mixed and disparate results are sometimes obtained for specific population groups or countries other than the United States. Also, the improvements in physical health appear to occur despite reductions in incomes and decreased use of medical care.

See Also

- health economics
- population health, economic implications of

Bibliography

Brenner, M.H. 1979. Mortality and the national economy. *The Lancet* 314, 568–73.

Brenner, M.H. and Mooney, A. 1983. Unemployment and health in the context of economic change. *Social Science Medicine* 17, 1125–38.

Deaton, A. 2003. Health, inequality and economic development. *Journal of Economic Literature* 41, 113–58.

Deaton, A. and Lubotsky, D. 2003. Mortality, inequality and race in American cities and states. *Social Science and Medicine* 56, 1139–53.

Eibner, C. and Evans, W.N. 2005. Relative deprivation, poor health habits, and mortality. *Journal of*

Human Resources 40, 591–620.

Eyer, J. 1977. Prosperity as a cause of death. *International Journal of Health Services* 7, 125–50.

Fiscella, K. and Franks, P. 1997. Poverty or income inequality as predictor of mortality: longitudinal cohort study. *British Medical Journal* 314, 1724–7.

Gravelle, H. 1998. How much of the relationship between population mortality and unequal distribution of income is a statistical artifact? *British Medical Journal* 316, 382–5.

Judge, K., Mulligan, J.-A. and Benzeval, M. 1998. Income inequality and population health. *Social Science and Medicine* 45, 567–79.

Kasl, S.V. 1979. Mortality and the business cycle: some questions about research strategies when utilizing macro-social and ecological data. *American Journal of Public Health* 69, 784–88.

Lynch, J., Smith, G.D., Harper, S., Hillemeier, M., Ross, N., Kaplan, G.A. and Wolfson, M. 2004. Is income inequality a determinant of population health? Part 1. A systematic review of the literature. *Milbank Quarterly* 82, 5–99.

McAvinchey, I.D. 1988. A comparison of unemployment, income and mortality interaction for five European countries. *Applied Economics* 20, 453–71.

Mellor, J.M. and Milyo, J. 2002. Income inequality and health status in the United States: evidence from the Current Population Survey. *Journal of Human Resources* 37, 510–39.

Neumayer, E. 2004. Recessions lower (some) mortality rates. *Social Science & Medicine* 58, 1037–47.

Ogburn, W.F. and Thomas, D.S. 1922. The influence of the business cycle on certain social conditions. *Journal of the American Statistical Association* 18, 324–40.

Preston, S.H. 1975. The changing relationship between mortality and the level of development. *Population Studies* 29, 249–57.

Ruhm, C.J. 2000. Are recessions good for your health? *Quarterly Journal of Economics* 115, 617–50.

Ruhm, C.J. 2003. Good times make you sick. *Journal of Health Economics* 22, 637–58.

Ruhm, C.J. 2006. Macroeconomic conditions, health and mortality. In *Elgar Companion to Health Economics*, ed. A.M. Jones. Cheltenham, UK: Edward Elgar.

Smith, J.P. 1999. Healthy bodies and thick wallets: the dual relationship between health and economic status. *Journal of Economic Perspectives* 13, 145–66.

Subramanian, S.V. and Kawachi, I. 2003. The association between state income inequality and worse health is not confounded by race. *International Journal of Epidemiology* 32, 1022–28.

Tapia Granados, J. 2005. Recessions and mortality in Spain, 1980–1997. *European Journal of Population* 21, 393–422.

Wagstaff, A. 1985. Time series analysis of the relationship between unemployment and mortality: a survey of econometric critiques and replications of Brenner's studies. *Social Science and Medicine* 21, 985–96.

Wagstaff, A. and van Doorslaer, E. 2000. Income inequality and health: what does the literature tell us. *Annual Review of Public Health* 21, 543–67.

Wilkinson, R.G. 1992. Income distribution and life expectancy. *British Medical Journal* 304, 165–9.

Wilkinson, R.G. 1997. Health inequalities: relative or absolute material standards. *British Medical Journal* 314, 591–95.

How to cite this article

Ruhm, Christopher J. "health outcomes (economic determinants)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000186> doi:10.1057/9780230226203.0715

The New Palgrave Dictionary of Economics Online

Hearn, William Edward (1826 – 1888)

M. White

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

competition; Hearn, William Edward; income distribution; prices

Article

Hearn was born in County Cavan, Ireland and died in Melbourne, Australia. Educated at Trinity College Dublin, he was appointed professor of political economy and other subjects at the University of Melbourne in 1854. Subsequently a member of the Legislative Council of the State of Victoria and contributor to the local press, Hearn is known to economists principally for his *Plutology* (1863). *Plutology* explains increasing wealth as a result of the competitive exchange of services. The analysis owes a good deal to Herbert Spencer and Frederic Bastiat. Competition is held to have three general results. It is: beneficent, since prices reflect the minimum cost of procuring a service; just, because recompense is in proportion to merit; and equalizing, since no recompense permanently reflects the effects of chance. As an ‘unfailing rule’, the pursuit of self-interest means services are produced in ‘order of their social importance’. Competition results in a natural social order, ordained by Providence, in which the principles of Darwin's natural selection are applied to industry (ch. 19).

The price of any service, determined by the extent of demand and supply, oscillates towards the minimum cost of production. The upper price limit is set where the purchaser equates desire for a service with the sacrifice necessary to either directly produce or obtain it from another source. The minimum price must cover any outlays and provide the ‘average’ reward for the vendor's type of service (ch. 14). The discussion of price formation is not conducted in marginalist terms and owes a good deal, via J.S. Mill, to de Quincey.

The distribution of income is explained according to the general principles of exchange. The manager of an enterprise, for example, contracts with the vendors of labour and ‘capital’ for their services at a fixed price. Discounting all costs and gross returns, the manager then has full title to the output, assuming responsibility for losses and receiving net gains. If capital is supplied in ‘commodity’ form (machinery, buildings), rent is paid; if it is supplied in money form (loans, insurance), interest accrues. Directly following Bastiat's *Harmonies*, Hearn argues that ground rent cannot be a gratuitous gift of nature as

land has a price only if labour is bestowed on it (ch. 18).
The role of a central government in an ‘advanced’ nation is thus basically a nightwatchman, although it may undertake some limited regulation. It is acknowledged, however, that the accumulation path will be impeded to some extent. The most serious problems result from enterprises mistaking market demand and engaging in speculative ventures. Still, fluctuations in output and investment have relatively little importance. ‘Failures, poverty, suffering and privation’ are not part of the ‘ordinary course of events’, any ‘ravages’ are soon repaired and objects destroyed in commercial fluctuations would have mainly been consumed rather than invested. Any ‘disturbances’ are thus ‘incidental’ to the natural laws of economic organization (ch. 24).
Marshall and Edgeworth bestowed high praise on *Plutology*, while Jevons considered its arguments were ‘nearly identical’ to those in his *Theory*. Subsequent commentary has noted Hearn's dogmatism and plagiarism, especially from John Rae and Bastiat.

Selected works

1851. *The Cassell Prize Essay on the Condition of Ireland*. London.

1863. *Plutology: or the Theory of the Efforts to Satisfy Human Wants*. Melbourne: Robertson; London: Macmillan, 1864.

1867. *The Government of England, its Structure and its Development*. Melbourne: Robertson; London: Longmans Green, Reader & Dyer.

1878. *The Aryan Household, its Structure and its Development*. Melbourne: Robertson; London: Longmans, Green & Co., 1879.

1883. *The Theory of Legal Duties and Rights*. Melbourne: Government Printer; London: Trubner & Co.

Bibliography

Copland, D.A. 1935. *W.E. Hearn: First Australian Economist*. Melbourne: Melbourne University Press.

La Nauze, J.A. 1949. *Political Economy in Australia*. Melbourne: Melbourne University Press.

How to cite this article

White, M. V. "Hearn, William Edward (1826–1888)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000032> doi:10.1057/9780230226203.1858

The New Palgrave Dictionary of Economics Online

heavy-tailed densities

Rustam Ibragimov
From The New Palgrave Dictionary of Economics Online, , 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews several frameworks commonly used in modelling heavy-tailed densities and distributions in economics, finance, risk management, econometrics and statistics. The results and conclusions discussed in the article indicate that the presence of heavy tails can either reinforce or reverse the implications of a number of models in these fields, depending on the degree of heavy-tailedness.

Keywords

heavy-tailed densities; heavy-tailed distributions; power laws; stable distributions; mean; variance; moments; semi-heavy tails; diversification; portfolio choice; value at risk; linear estimators; sample mean; efficiency; regression; robustness; dependence; α-symmetric distributions; models with common shocks

Article

Several notions and classes of heavy-tailed densities and distributions are available in the economic, financial, statistical and probability literature. A unifying property common to such densities and distributions is that their tails are heavier than in the Gaussian case, either in the sense of faster decay to zero or in the sense of comparisons of heavy-tailedness measures, such as kurtosis.

1 Heavy-tailed models

In models involving a heavy-tailed random variable (rv) X it is usually assumed that the distribution of X has power tails, so that

$$P(|X| > x) \approx \frac{C}{x^{\alpha}}, \quad \alpha > 0, C > 0, \quad as \quad x \to +\infty$$

(1)

(here and throughout the article, $f(x) \bullet g(x)$ as $x \to +\infty$ means that $\lim_{x \to +\infty} \frac{f(x)}{g(x)} = 1$). The parameter α in (1) is referred to as the tail index, or the tail exponent, of the distribution of X. An important property of rvs X satisfying (1) is that the absolute moments of X are finite if and only if their order is less than the tail index α : $E|X|^p < \infty$ if $p < \alpha$ and $E|X|^p = \infty$ if $p \geq \alpha$.

Examples satisfying (1) include Pareto distributions with densities $f(x) = \alpha x_0^{\alpha} / x^{\alpha+1}$ for $x > x_0 > 0$; $f(x) = 0$ for $x \leq x_0$. In addition, (1) is satisfied for Student's t – distributions with densities $f(x) = \frac{\Gamma((\alpha+1)/2)}{\sqrt{\alpha\pi}\Gamma(\alpha/2)}(1 + x^2/\alpha)^{-(\alpha+1)/2}$, $x \in \mathrm{R}$, where $\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$, $z > 0$, denotes the Gamma function. Relation (1) also holds for the important class of stable distributions that are closed under portfolio formation.

In addition to distributions that follow power laws (1), several other frameworks for modelling heavy-tailed phenomena have been proposed in the literature, including distributions with finite moments of any order and semi-heavy tails. Such tails are thinner than in the case of any power law (1) but much heavier than those of normal distributions. Semi-heavy tails in this sense are exhibited, for instance, by normal inverse Gaussian and, more generally, generalized hyperbolic distributions (see section 3.2 in McNeil, Frey and Embrechts, 2005, and references therein), as well as by the important case of log-normal distributions.

2 Empirical results

Numerous studies in economics, finance, risk management and insurance have indicated that distributions of many variables of interest in these fields exhibit deviations from Gaussianity, including those in the form of heavy tails (1) (see, among others, the discussion and reviews in Embrechts, Klüppelberg and Mikosch, 1997; Rachev, Menn and Fabozzi, 2005). This stream of literature goes back to Mandelbrot (1963) (see also Fama, 1965, and the papers in Mandelbrot, 1997), who pioneered the study of heavy-tailed distributions in economics and finance.

The following is a sample of estimates of the tail index α in distributions satisfying (1) for returns on various stocks and stock indices: $3 < \alpha < 5$ (Jansen and de Vries, 1991), $2 < \alpha < 4$ (Loretan and Phillips, 1994), $1.5 < \alpha < 2$ (McCulloch, 1997), $0.9 < \alpha < 2$ (Rachev and Mittnik, 2000), $\alpha \approx 3$ (Gabaix et al., 2003). Power laws (1) with $\alpha \approx 1$ (Zipf laws) have been found to hold for firm sizes and city sizes (see Gabaix, 1999a, 1999b; Axtell, 2001). As discussed by Nešlehova, Embrechts and Chavez-Demoulin (2006), tail indices less than 1 are observed for empirical loss distributions of a number of operational risks. Silverberg and Verspagen (2007) report the tail indices α to be significantly less than 1 for

financial returns from technological innovations. The analysis in Ibragimov, Jaffee and Walden (2008) indicates that the tail indices may be considerably less than 1 for economic losses from earthquakes and other natural disasters. Anderson (2006) discusses the heavy-tailedness paradigm in many modern economic and financial markets transformed by the Internet and the development of technology.

3 Stable distributions

Canonical examples of power laws (1) are given by stable distributions. For $0<\alpha\leq 2$ and $\sigma>0$, the symmetric stable distribution $S_\alpha(\sigma)$ is the distribution of an rv X with the characteristic function (cf) $E(e^{ixX})=\exp\{-\sigma^\alpha|x|^\alpha\}$, $i^2=-1$, $x\in\mathrm{R}$. Throughout the article, we write $X\sim S_\alpha(\sigma)$, if the rv X has the distribution $S_\alpha(\sigma)$. Given two rvs X and Y, the notation $X=^dY$ means that the distributions of X and Y are the same.

The parameters α and σ are referred to as the characteristic exponent (index of stability) and the scale parameter of the symmetric stable distribution $S_\alpha(\sigma)$. In general, stable distributions also depend on the skewness parameter β and the location parameter μ. Symmetric stable distributions $S_\alpha(\sigma)$ correspond to the case $\mu=\beta=0$.

A closed form expression for the density $f(x)$ of a stable distribution is available only in the following cases: normal densities that correspond to the case $\alpha=2$; Cauchy densities $f(x)=\sigma/(\pi(\sigma^2+(x-\mu)^2))$, $x\in\mathrm{R}$, with $\alpha=1$; and the densities $f(x)=(\sigma/(2\pi))^{1/2}exp(-\sigma/(2x))x^{-3/2}$, $x>0$; $f(x)=0$, $x\leq 0$, of Lévy distributions with $\alpha=1/2$ and their shifted and reflected versions. While normal and Cauchy distributions are symmetric about the location parameter μ, Lévy distributions are concentrated on the positive semi-axis $(0,\infty)$.

The index of stability α characterizes heaviness of tails of the distribution $S_\alpha(\sigma)$. If $X\sim S_\alpha(\sigma)$, then X satisfies power law (1). Thus, the absolute moments $E|X|^p$ of an rv $X\sim S_\alpha(\sigma)$, $\alpha\in(0,2)$ are finite if $p<\alpha$ and are infinite otherwise. The same conclusions hold for skewed stable distributions. In particular, second and higher moments are infinite for all non-Gaussian stable distributions with $\alpha<2$. Cauchy distributions with $\alpha=1$ have infinite first and higher absolute moments. If the rv X has a Lévy distribution with $\alpha=1/2$, then $E|X|^p=\infty$ for all $p\geq 1/2$.

The distributions of stable rvs X with $\alpha>1$ are moderately heavy-tailed in the sense that they have finite first absolute moments: $E|X|<\infty$. In contrast, the distributions of stable rvs X with $\alpha<1$ are extremely heavy-tailed in the sense that their first absolute moments are infinite: $E|X|=\infty$.

The scale parameter σ is a generalization of the standard deviation; it coincides with the standard deviation for normal distributions with $\alpha=2$. For $\alpha>1$, the location parameter μ of a stable distribution coincides with its mean: in particular, $EX=0$ for symmetric stable rvs $X\sim S_\alpha(\sigma)$ with $\alpha\in(1,2]$.

Stable distributions are closed under portfolio formation. In particular, if $X_i\sim S_\alpha(\sigma)$, $\alpha\in(0,2]$, are iid symmetric stable risks, then, for all portfolio weights $w_i\geq 0$, $i=1,\ldots,n$,

$$\sum_{i=1}^{n} w_i X_i \overset{d}{=} \left(\sum_{i=1}^{n} w_i^{\alpha}\right)^{1/\alpha} X_1, \tag{2}$$

or, equivalently, $\sum_{i=1}^{n} w_i X_i \sim S_\alpha(\tilde{\sigma})$, where $\tilde{\sigma} = \sigma\left(\sum_{i=1}^{n} w_i^{\alpha}\right)^{1/\alpha}$ (see Zolotarev, 1986; Embrechts et al., 1997; Rachev and Mittnik, 2000, for a review of properties of stable distributions).

Multivariate extensions of the stable family such as α – symmetric distributions allow one to model frameworks with a wide range of heavy-tailedness in marginals and dependence among them (see Fang, Kotz and Ng, 1990 and the discussion in Ibragimov, 2007, and Ibragimov and Walden, 2007). The class of α – symmetric distributions contains models with common shocks affecting all heavy-tailed risks as well as spherical distributions. Spherical distributions, in turn, include such examples as Kotz type, multinormal, logistic and multivariate α – stable distributions. In addition, they include a subclass of mixtures of normal distributions as well as multivariate *t* – distributions that were used in the literature to model heavy-tailedness phenomena with dependence and finite moments up to a certain order.

4 Robustness of economic models

Heavy-tailedness has important implications for robustness of many economic models, leading, in a number of settings, to reversals of conclusions of these models to the opposite ones.

This may be illustrated, for instance, by the properties of value at risk (VaR) models and the analysis of diversification and portfolio choice in VaR frameworks under heavy-tailedness (see Embrechts, McNeil and Straumann, 2002; ch. 12 in Bouchard and Potters, 2004; Ibragimov, 2005a, 2005b, and references therein).

Given a loss probability $q \in (0,1)$ and an rv (risk) X we denote by $VaR_q(X)$ the VaR of X at level q, that is, its $(1-q)$–quantile: $VaR_q(X) = \inf\{z \in R: P(X>z) \leq q\}$ (in what follows, we interpret the positive values of X as a risk holder's losses).

Throughout the article, $\mathcal{S}_n = \{w = (w_1, \ldots, w_n): w_i \geq 0, i = 1, \ldots, n, \sum_{i=1}^{n} w_i = 1\}$. For $w = (w_1, \ldots, w_n) \in \mathcal{S}_n$, denote by Z_w the return on the portfolio of risks $X_1,\ldots,X_n$ with weights w. Denote $\underline{w} = (1/n, 1/n, \ldots, 1/n) \in \mathcal{S}_n$ and $\bar{w} = (1, 0, \ldots, 0) \in \mathcal{S}_n$. The expressions $VaR_q(Z_{\underline{w}})$ and $VaR_q(Z_{\bar{w}})$ are thus the VaRs of the portfolio with equal weights and of the portfolio consisting of only one return (risk). It is natural to think about the portfolio with weights $\underline{w}$ as the most diversified and about the portfolio with weights $\bar{w}$ as the least diversified among all the portfolios with weights $w \in \mathcal{S}_n$.

A simple example where diversification is preferable is provided by the standard case with normal risks. Let $n \geq 2$, $q \in (0,1/2)$, and let $X_1,\ldots,X_n \sim S_2(\sigma)$ be iid symmetric normal rvs. For the portfolio of X_i's with the equal weights $\underline{w} = (1/n, 1/n, \ldots, 1/n)$ we have $Z_{\underline{w}} = (1/n)\sum_{i=1}^{n} X_i \overset{d}{=} (1/\sqrt{n}) X_1$. Consequently, by positive homogeneity of the VaR,

$VaR_q(Z_{\underline{w}}) = (1/\sqrt{n})VaR_q(X_1) = (1/\sqrt{n})VaR_q(Z_{\bar{w}}) < VaR_q(Z_{\bar{w}})$. Thus, the VaR of the most diversified portfolio with equal weights $\underline{w}$ is less than that of the least diversified portfolio with weights $\bar{w}$ consisting of only one risk Z_1.

Using (2), one can also show (see Ibragimov, 2005a, 2005b) that diversification is preferable in the VaR framework for all iid moderately heavy-tailed risks $X_i \sim S_\alpha(\sigma)$ with $\alpha \in (1,2]$ in the sense that $VaR_q(Z_{\underline{w}}) \le VaR_q(Z_w) \le VaR_q(Z_{\bar{w}})$ for all $q \in (0,1/2)$ and all weights $w \in \mathcal{S}_n$.

The settings where diversification is suboptimal in the VaR framework may be illustrated as follows. Let $q \in (0,1)$ and let $X_1,\ldots,X_n$ be iid positive risks with a Lévy distribution with the tail index $\alpha = 1/2$ and the density $f(x) = (\sigma/(2\pi))^{1/2}\exp(-\sigma/(2x))x^{-3/2}$. Similar to symmetric stable distributions, the portfolios of the risks X_i satisfy (2) with $\alpha = 1/2$. Using (2) for equal weights $w_i = 1/n$, we get $Z_{\underline{w}} = (1/n)\sum_{i=1}^{n} X_i \overset{d}{=} nX_1$. Consequently, $VaR_q(Z_{\underline{w}}) = nVaR_q(X_1) = nVaR_q(Z_{\bar{w}}) > VaR_q(Z_{\bar{w}})$.

Thus, the VaR of the least diversified portfolios with weights $\bar{w}$ that consists of only one risk is less than the VaR of the most diversified portfolio with equal weights $\underline{w}$.

Relation (2) further implies (see Ibragimov, 2005a, 2005b) that the results on diversification suboptimality in the VaR framework continue to hold for all iid extremely heavy-tailed risks $X_i \sim S_\alpha(\sigma)$ with $\alpha \in (0,1)$. Namely, for such risks, $VaR_q(Z_{\bar{w}}) \le VaR_q(Z_w) \le VaR_q(Z_{\underline{w}})$ for all $q \in (0,1/2)$ and all weights $w \in \mathcal{S}_n$.

The results in Ibragimov (2005b) provide portfolio VaR comparisons for convolutions of stable distributions with different tail indices and their extensions to dependence, skewness and heterogeneity, including convolutions of α – symmetric distributions and models with common shocks. Ibragimov and Walden (2007) and Ibragimov, Jaffee and Walden (2008) show that the (non-)diversification results in heavy-tailed value at risk models continue to hold for bounded risks. Ibragimov et al. (2008) further use these conclusions to develop a simple model for markets for catastrophic risk in which nondiversification traps may arise.

5 Implications for statistical and econometric methods

Similar to the portfolio VaR analysis, heavy-tailedness presents a challenge for applications of standard statistical and econometric methods. In particular, as pointed out by Granger and Orr (1972) and in a number of more recent studies (see, among others, ch. 7 in Embrechts et al., 1997, and references therein), many classical approaches to inference based on variances and (auto)correlations such as regression and spectral analysis, least squares methods and autoregressive models may not apply directly in the case of heavy-tailed observations with infinite second or higher moments.

An important simple illustration is provided by the failure of the law of large numbers (LLN) for observations with infinite first moments and variances. When more information about the structure of heavy-tailedness is available, one can obtain more refined results that point out to crucial differences between moderately heavy-tailed and extremely heavy-tailed populations.

Consider the problem of estimating the parameter μ in the simple model

$$X_i = \mu + \eta_i, \tag{3}$$

where η_i are iid errors with an absolutely continuous symmetric distribution. Given a random sample $X_1,\ldots,X_n$ that follows (3), denote by $\hat{\theta}_n(w)$ the linear estimator $\hat{\theta}_n(w) = \sum_{i=1}^{n} w_i X_i$.

It is well known that, if $E\eta_i^2 < \infty$, then the sample mean $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ is the best linear unbiased estimator (BLUE) of the population mean $\mu = EX_i$. That is, $\bar{X}_n$ is the most efficient estimator of μ among all unbiased linear estimators $\hat{\theta}_n(w)$ in the sense of variance comparisons: $Var[\bar{X}_n] \le Var[\hat{\theta}_n(w)]$ for all $w \in \mathcal{S}_n$.

The definition of efficiency based on variance breaks down in the case of heavy-tailed populations with infinite second moments. A natural approach to comparison of performance of estimators under heavy-tailedness is to order them by likelihood of observing their large deviations from the true population parameter. This approach relies on the concept of peakedness of rvs and leads to the following definition. Let $\hat{\theta}(v)$ and $\hat{\theta}(w)$ be two linear estimators of the parameter μ in model (3). The estimator $\hat{\theta}(v)$ is said to be more efficient than $\hat{\theta}(w)$ in the sense of peakedness (P-more efficient than $\hat{\theta}(w)$ for short) if $P(|\hat{\theta}(v) - \mu| > \varepsilon) \le P(|\hat{\theta}(w) - \mu| > \varepsilon)$ for all $\varepsilon \ge 0$, with strict inequality whenever the two probabilities are not both 0 or both 1.

The results in Ibragimov (2007) for general dependent settings such as convolutions of α – symmetric distributions and models with common shocks imply that the sample mean $\bar{X}_n$ is the best linear unbiased estimator of the population mean μ in the sense of P-efficiency for moderately heavy-tailed errors $\eta_i \sim S_\alpha(\sigma)$ with $\alpha > 1$. However, if the errors $\eta_i \sim S_\alpha(\sigma)$ are extremely heavy-tailed with $\alpha < 1$, then P-efficiency of the sample mean is smallest among all linear estimators $\hat{\theta}(w)$ of the population centre μ with weights $w \in \mathcal{S}_n$.

The conclusions in Ibragimov (2005a) show that, similar to the portfolio VaR analysis and the efficiency properties of linear estimators, many models in economics and related fields are robust to heavy-tailedness assumptions provided the distributions entering these assumptions are moderately heavy-tailed. However, the implications of these models are reversed for distributions with sufficiently heavy-tailed densities.

6 Conclusion

The results reviewed in this article and those obtained in the recent literature imply that the presence of heavy-tailedness can either reinforce or reverse the implications of many models in economics, finance, econometrics, statistics and risk management, depending on the degree of heavy-tailedness. Typically, the standard implications of the models continue to hold for moderately heavy-tailed distributions. However, these implications may become the opposite ones under sufficient heavy-tailedness. Therefore, the models should be applied with care in heavy-tailed settings, especially in the case of the

tail indices close to the value $\alpha = 1$, which in many cases provides the critical robustness boundary.

See Also

- lognormal distribution
- Pareto distribution
- power laws

Bibliography

Anderson, C. 2006. *The Long Tail*. New York: Hyperion.

Axtell, R.L. 2001. Zipf distribution of U.S. firm sizes. *Science* 293, 1818–1820.

Bouchard, J.-P. and Potters, M. 2004. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*, 2nd edn. Cambridge: Cambridge University Press.

Embrechts, P., Klüppelberg, C. and Mikosch, T. 1997. *Modelling Extremal Events for Insurance and Finance*. New York: Springer.

Embrechts, P., McNeil, A. and Straumann, D. 2002. Correlation and dependence in risk management: properties and pitfalls. In *Risk Management: Value at Risk and Beyond*, pp. 176–223, ed. M.A.H. Dempster. Cambridge: Cambridge University Press.

Fama, E. 1965. The behavior of stock market prices. *Journal of Business* 38, 34–105.

Fang, K.T., Kotz, S. and Ng, K.W. 1990. *Symmetric Multivariate and Related Distributions*, Vol. 36 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Gabaix, X. 1999a. Zipf's law and the growth of cities. *American Economic Review* 89, 129–32.

Gabaix, X. 1999b. Zipf's law for cities: an explanation. *Quarterly Journal of Economics* 114, 739–67.

Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H.E. 2003. A theory of power-law distributions in financial market fluctuations. *Nature* 423, 267–70.

Granger, C.W.J. and Orr, D. 1972. Infinite variance and research strategy in time series analysis. *Journal of the American Statistical Association* 67, 275–85.

Ibragimov, R. 2005a. *New Majorization Theory in Economics and Martingale Convergence Results in*

Econometrics. Ph.D. dissertation, Yale University.

Ibragimov, R. 2005b. Portfolio diversification and value at risk under thick-tailedness. Harvard University Research Discussion Paper 2086. Available at: <http://www.economics.harvard.edu/pub/hier/2005/HIER2086.pdf>, accessed 13 January 2009. Forthcoming in *Quantitative Finance*.

Ibragimov, R. 2007. Efficiency of linear estimators under heavy-tailedness: convolutions of α–symmetric distributions. *Econometric Theory* 23, 501–17.

Ibragimov, R., Jaffee, D. and Walden, J. 2008. Nondiversification traps in catastrophe insurance markets. *Review of Financial Studies*. Available at: <http://0-rfs.oxfordjournals.org.library.lemoyne.edu/cgi/content/abstract/hhn021v1>, accessed 13 January 2009.

Ibragimov, R. and Walden, J. 2007. The limits of diversification when losses may be large. *Journal of Banking and Finance* 31, 2551–69.

Jansen, D.W. and de Vries, C.G. 1991. On the frequency of large stock returns: putting booms and busts into perspective. *Review of Economics and Statistics* 73, 18–32.

Loretan, M. and Phillips, P.C.B. 1994. Testing the covariance stationarity of heavy-tailed time series. *Journal of Empirical Finance* 1, 211–48.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Mandelbrot, B. 1997. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*. New York: Springer-Verlag.

McCulloch, J.H. 1997. Measuring tail thickness to estimate the stable index alpha: a critique. *Journal of Business and Economic Statistics* 15, 74–81.

McNeil, A.J., Frey, R. and Embrechts, P. 2005. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton, N.J.: Princeton University Press.

Nešlehova, J., Embrechts, P. and Chavez-Demoulin, V. 2006. Infinite mean models and the LDA for operational risk. *Journal of Operational Risk* 1, 3–25.

Rachev, S.T., Menn, C. and Fabozzi, F.J. 2005. *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing*. Hoboken, N.J.: Wiley.

Rachev, S.T. and Mittnik, S. 2000. *Stable Paretian Models in Finance*. New York: Wiley.

Silverberg, G. and Verspagen, B. 2007. The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance. *Journal of Econometrics* 139, 318–39.

Zolotarev, V.M. 1986. *One-Dimensional Stable Distributions*. Providence, R.I.: American Mathematical Society.

How to cite this article

Ibragimov, Rustam. "heavy-tailed densities." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_H000191> doi:10.1057/9780230226203.1893

The New Palgrave Dictionary of Economics Online

Heckman, James (born 1944)

Richard Blundell, Lars Peter Hansen and Derek Neal

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

James Heckman has made fundamental contributions to the development of methods that allow economists to estimate models of economic behaviour using data on individual decisions. He has also produced numerous important empirical results that advanced understanding of how government policies that regulate labour markets and influence educational opportunity affect economic inequality among individuals and groups.

Keywords

additive regression models; black–white labour market inequality in the United States; childcare; competing risks model; consumer theory; depreciation; difference-in-difference estimators; discrete and continuous decision variables; duration models; Econometric Society; econometrics; employment spells; fixed effects; General Educational Development (GED) test; Heckman, J.; heterogeneity; human capital; identification; instrumental variables; internal migration; labour supply; leisure; life-cycle labour supply; marginal treatment effects; marginal utility of wealth; matching; maximum likelihood; multiple regime models; non-cognitive skills; principle assumption; programme evaluation; Roy model; selection; selection bias; state dependence; training; treatment effect; two-step estimator; uncertainty; wage differentials

Article

1 Labour supply and selection

In Heckman's early work on labour supply we see at least three related contributions. First, he integrated consumer theory and the theory of labour supply. Second, he developed an empirical life-cycle setting for labour supply. Third, he provided an economically coherent framework for the statistical analysis of participation, labour supply and market wages. Heckman's work on labour supply, which originated in

the early to mid-1970s, set the scene for the development of his research on selection, on labour market dynamics and on programme evaluation. They are all empirically oriented but with a keen eye on the identification and estimation of structural economic parameters from micro data.

Heckman's initial aim was to estimate the parameters of indifference curves for leisure and consumption. Given these, one could measure the welfare cost of some tax or welfare intervention and also simulate the impact of new policies. There were at least three key unresolved issues in the literature at that time: the econometric problem of non-participation in the presence of childcare costs; the need for a reasonably flexible functional form that could capture variation in hours worked among participants; and the lack of information concerning wage offers among those who do not participate. Heckman successfully addressed all of these issues in two remarkable papers (1974a; 1974b). He recognized that a simple least squares analysis of hours of work, wages and participation would not, by itself, identify preference parameters and that the standard Tobit model alone was also insufficient to deal with the problem. As an alternative, Heckman developed an estimation procedure that allowed the work decision to be based on interrelated choices over hours of work and the use of formal childcare, each with its own separate source of stochastic variation. This approach is the forerunner of many microeconometric developments in this area and continues to set the standard by which models are judged. Indeed, Heckman's development of a likelihood that captures the sampling information on participation and wages can be seen as the beginning of the analysis of endogenously selected samples.

Yet the contributions of this work go beyond the insights concerning selection. His marginal rate of substitution specification for preferences turned out to be a highly innovative way of dealing with non-participation while allowing flexible but heterogeneous preferences, and the endogenous choice of formal and informal childcare jointly with hours of work and participation provided a basis for the analysis of multiple regime models.

In this ‘static’ labour supply analysis, we find repeated references to the potential importance of a more dynamic setting. In fact, Heckman conducted this work alongside his development of a life-cycle framework for labour supply. The origins of this work are in the first essay of his 1971 Princeton University doctoral thesis. Heckman began by noting that both income and consumption appear to follow a similar hump-shaped path over the life cycle that is out of line with the most basic consumption-smoothing model. Heckman (1974c) provides a beautifully simple, yet complete, integration of intertemporal consumption and labour supply theory and shows that a model with labour supply and uncertainty can easily explain these empirical phenomena.

Heckman extended this life-cycle analysis in two different directions. In Heckman (1976), he incorporated human capital investment and showed how earnings functions that ignored life-cycle labour supply tended to overestimate rates of depreciation. In other work, he developed an empirically implementable form of the intertemporal substitution model for labour supply. Here, he pointed out that given standard neoclassical assumptions the marginal utility of wealth is constant over time for an individual but differs across individuals and is clearly correlated with wages. Since labour supply choices could be written in terms of current wages and the marginal utility of wealth, Heckman's observation pointed to a perfect application of a fixed effects estimator for panel data, and in Heckman and MaCurdy (1980) he applied such an estimator to the panel data analysis of female labour supply. Heckman's model could also be adjusted to account for uncertainty and so became the prototypical intertemporal model of labour supply. It directly recovered the intertemporal substitution elasticity for labour supply and showed immediately the relationship of this intertemporal elasticity with the standard

Hicksian and Marshallian elasticities, thereby tying together the ‘static’ and life-cycle approach to labour supply analysis.

Heckman's empirical investigations of individual labour supply behaviour stimulated further analysis of their statistical implications, and at least two major innovations in econometrics came out of this work. This work yielded new methods for analysing selected samples and also for estimating simultaneous multivariate choice models in which outcomes are a mixture of discrete and continuous decision variables. It is easy to see how important work on labour supply led to progress in these areas. Labour market participation is a choice based, in part, on wage offers that are observed only among those who participate, and household choices concerning participation, hours, and childcare present a mixture of both discrete and continuous decision variables. While the links to labour supply analysis are clear, the applicability of these two developments grew far beyond the study of labour supply.

Heckman's (1979) selection model is one of the most renowned econometric models since the mid-20th century. This work laid the foundation for the subsequent work on returns to training, the study of union wage differentials and to many other microeconometric problems. His approach was innovative but also simple. Starting with an additive regression model, Heckman noted that for normal distributions the conditional mean for a selected sample involves a single additional term that is itself a function of the selection probability. This term or ‘control function’ may be estimated in a first step from the choice probability model, and thus a computationally convenient two-step estimator is available for the analysis of selected samples.

Future work demonstrated that the selection model and the two-step estimator are more generally applicable in cases where the normality assumption fails. Heckman (1990) and others developed semiparametric extensions for the additively separable model. Heckman and Honoré (1990) derived the general nonparametric identification of the Roy model – a two-regime generalization of the additively separable selection model. Both Heckman and Sedlaceck (1985) and Heckman and Scheinkman (1987) provide empirical analyses of aggregate and sectoral wage distributions when individuals self-select into the labour market and into sectors of the economy.

If actions are a mixture of discrete and continuous decision variables that are simultaneously determined, then there will be a further condition on the econometric model to guarantee it provides a coherent statistical relationship between inputs and response. Heckman labelled this condition the ‘principle assumption’, and in Heckman (1978) he derives the conditions required for a coherent econometric framework. This work has influenced econometric work in industrial organization. Heckman's condition concerning a jump parameter in the mean of the latent variable underlying a discrete choice is easily mapped into analyses of entry decisions when fixed costs are present.

2 Panel data and state transitions

During the 1970s and 1980s economists gained much greater access to panel data-sets, and this development greatly shaped the research agenda in labour economics and other applied fields. Economists began to focus their attention on the sequence of decisions that firms and individuals make over time and attempted to model and understand the patterns of correlation over time in these decisions. The current choices of individuals may be correlated with their past choices because of persistent unobserved individual differences (heterogeneity) or because current preferences or opportunities depend on past actions (state dependence). Thus, heterogeneity and state dependence can easily be

confounded. Heckman (1981a; 1981b) formally characterizes this identification problem, discusses pitfalls with simple or naive attempts at measuring one or the other effects, and shows how panel data can be used constructively to sort out competing explanations.

Heckman and Singer (1984) confront formally the statistical aspects of unobserved heterogeneity in the context of duration models. Duration models are used in studying job search, mortality, labour supply, marriage and other phenomena. Fully parametric models of duration include precise parameterizations of the unobservable (to an econometrician) attributes of individuals that influence the optimal duration of spells. Heckman and Singer show that these parameterizations can contaminate estimates of the economic parameters that pin down the structural relationship between spell duration and the rate of spell completions. As an alternative they propose, formally justify and implement a maximum likelihood estimator where the unobserved heterogeneity is modelled nonparametrically. Interestingly, in their application to data on unemployment spells, their nonparametric estimator chooses a small number of support points (individual types), and their Monte Carlo results in the same paper show that their estimates of the distribution of unobserved types are never close to the truth. However, their nonparametric treatment of unobserved heterogeneity allows them to choose points of support in a flexible way that avoids contaminating the estimated parameters of interest. The empirical results presented by Heckman and Singer strongly suggest that many of the somewhat puzzling results in the previous literature on the determinants of unemployment spells were the result of researchers trying to simultaneously estimate models with parametric assumptions concerning both the true underlying model of duration times and the distribution of unobserved individual heterogeneity.

The competing risks model has been used in many disciplines. In this model, observed outcomes reflect the minimum realized transition time over a discrete set of possible state transitions. Heckman and Flinn (1982) develop a structural interpretation of the competing risks model in the context of labour markets and apply it to study employment spells. They investigate the identification of the underlying economic parameters of interest as restrictions are removed from the auxiliary statistical specification. In a related paper, Heckman and Honoré (1989) explore identification of the competing risks model of failure times, and they show how introducing regressors into the competing risks model overturns previously established non-identification results. In the previous section, we mentioned Heckman and Honoré's (1990) work on requirements for identification is a generalized Roy model. Here, too, they demonstrated precisely how identification could be achieved through either restrictions on the shape of underlying skill distributions or sufficient variation in prices or exogenous regressors.

3 Estimation of treatment effects

In Section 1 we described how Heckman's early work on labour supply led to developments in the analysis of selected samples. Over the years, Heckman began to demonstrate that numerous problems other than labour supply are actually problems where missing data are the key challenge for empirical investigators. The work on identification in general versions of the Roy model is part of this research agenda. In the 1990s Heckman produced a series of related empirical and methodological papers that grew out of sustained research on methods of programme evaluation (see Heckman, Ichimura and Todd, 1997; see also Ichimura, Heckman, Smith and Todd, 1998, and Clements, Heckman, and Smith, 1997). Heckman emphasized that the key impediment to measuring the return to any investment in training or education is the inability to see what those who receive treatment would have experienced in the absence

of training.

Heckman's work in this area helped clarify several important points. First, Heckman produced much evidence consistent with the proposition that the treatment effects associated with various training and education programmes vary greatly among participants, even among those who are similar with respect to observed demographic characteristics. Second, outcomes from experiments involving random assignment to treatment and control status do permit straightforward estimates of the average gain from treatment in a given sample, but more structure is required in order to draw inferences concerning the distribution of gains and losses from treatment. Third, the performance of non-experimental methods such as matching or control function models can be greatly improved when researchers take care to balance treated and non-treated samples with respect to the probability of treatment. Careful attention to the support of this probability in each sample as well as its density can greatly improve the performance of non-experimental estimators. For example, Ichimura, Heckman, and Todd (1997) demonstrate clearly that the performance of matching estimators improves when samples are re-weighted so that the density of the probability of treatment is the same among the treated and the untreated. Finally, difference-in-difference estimators may be an attractive strategy in situations where researchers have before and after data on treated and non-treated samples, but only in cases where there is evidence that the selection bias associated with treatment takes the form of a subject fixed effect that is constant over time.

Heckman has continued working in this area while devoting special attention to the mapping between various estimators in the literature and the precise set of counterfactual questions that they can address under various assumptions concerning how the data are generated. Heckman and Vytlacil (2005) present results that reflect the culmination of much of this research. In this paper, they explain how numerous estimators employed in the estimation of treatment effects may all be written as weighted averages of the marginal treatment effects (MTE) in the population. The marginal treatment effect is defined as the expected gain from treatment with the observed and unobserved determinants of participation held constant at particular values. Heckman and Vytlacil not only demonstrate how other estimators, such as instrumental variables, can be expressed in terms of the distribution of MTE, but they also demonstrate how MTE can be used as a building block in the construction of estimators that capture the expected treatment effects of specific policies on particular populations. This paper also clarifies an asymmetry in the way that heterogeneity enters models of selection and treatment. Agents may exhibit heterogeneity in terms of what they gain from treatment, but changes in their environment must affect each individual's likelihood of receiving treatment in a similar manner. Heckman goes on to spell out the challenges that researchers face if they wish to estimate models in which selection equations involve random coefficients.

4 Empirical work on inequality

To this point, we have focused almost exclusively on contributions that involve the development and implementation of new methods that seek to overcome some problem involving missing data. However, Heckman has also made noteworthy contributions that involve no methodological innovation but rather the use of simple methods to establish important facts about the sources of economic inequality and their relationship to government policy. His work in the areas of black–white inequality and the economics of education are ripe with examples of this type of work.

During the 1970s economists in the United States devoted considerable attention to the question of

whether or not the Civil Rights Act had actually improved the economic well-being of blacks. Butler and Heckman (1978) made an important early contribution to this literature by pointing out that declining labour-force participation rates among less skilled blacks during the post-Civil Rights era could create the impression of economic progress among blacks even if none existed. To measure changes in the distribution of potential real wages facing blacks, researchers needed to address the fact that an increasing fraction of less skilled workers did not report market work and thus did not report wages in most surveys.

Butler and Heckman (1978) greatly influenced future work on black–white economic progress, but Heckman's most important contributions to this literature are summarized in a 1991 *Journal of Economic Literature* piece with John Donohue that cataloged evidence supporting the hypothesis that the Civil Rights Act of 1964 did serve as a catalyst for a discrete episode of black economic progress that was not simply a continuation of existing trends.

Donohue and Heckman (1991) acknowledge that long-term improvements in access to schools and school quality contributed to secular progress for blacks in the labour market in decades before and after the Civil Rights Act. However, they also show that federal government intervention served as an important catalyst for black progress during the civil rights era. They demonstrate that black relative earnings rose during the 1960s primarily because of gains in the South, where civil rights laws were imposed on local communities by the federal government. Further, they cite previous work by Heckman and his co-authors that demonstrates how the Civil Rights Act broke down de facto and de jure occupational segregation in the South (see Heckman and Payner, 1989, and Butler, Payner, and Heckman, 1989. Donohue, Heckman, and Todd, 2002, show how private philanthropy and legal activism served as catalysts for improvements in black educational opportunity that pre-date the civil rights era).

Finally, they note a drastic decrease in the rate of net migration among blacks from the South to the North around 1965, a development that strongly suggests the Civil Rights Act did expand opportunity for blacks in the South.

Around the same time, Cameron and Heckman (1993) documented another set of important results concerning outcomes associated with a government programme. Cameron and Heckman demonstrated that persons who receive a high-school diploma by taking the General Educational Development (GED) test do not enjoy employment and earnings outcomes as adults that are in any way equivalent to those observed among high-school graduates. In fact, male GED recipients look quite similar to high-school dropouts with respect to many labour market outcomes.

Since the mid-1990s a growing literature has examined the cost and implied benefits of obtaining a GED for various types of individuals. Heckman and Rubinstein (2001) have taken this literature in a new direction by demonstrating that the difference between GED recipients and persons who finish high school takes the form of differences in non-cognitive skills related to self-esteem, work habits, and other personal traits. This discovery provides a natural explanation for the facts documented by Cameron and Heckman (1993). While the GED does certify that a young person has certain basic skills that are valued by employers, the combination of these skills and the failure to finish high school is a signal to employers that these youth are deficient in other areas.

Conclusion

Taken as a whole, the breadth and depth of Heckman's contributions to economics are stunning. His work on selection models, transition dynamics, and the estimation and identification of treatment effects has changed the way that economists analyse micro-data. At the same time, his empirical work in the economics of education and inequality produced numerous results that shape our understanding of modern labour markets and future research agendas for theorists and empiricists.
He has also shaped the profession as a teacher and advisor of students. During his career, Heckman has served as primary thesis adviser for scores of graduate students, and a significant number of his students have earned tenure in top economics departments, served as editors of journals, and helped produce yet another generation of scholars who take seriously the task of using economics to guide empirical investigations of important questions.

See Also

- black–white labour market inequality in the United States
- competing risks model

Selected works

1974a. (With O. Ashenfelter.) The estimation of income and substitution effects in a model of family labor supply. *Econometrica* 42, 73–86.

1974b. Shadow prices, market wages and labor supply. *Econometrica* 42, 679–94.

1974c. Life cycle consumption and labor supply: an explanation of the relationship between income and consumption over the life cycle. *American Economic Review* 64, 188–94.

1976. A life cycle model of earnings, learning and consumption. *Journal of Political Economy* 84(2, part II), S11–S44.

1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46, 931–59.

1978. (With R. Butler.) The impact of the government on the labor market status of black Americans: a critical review. In *Equal Rights and Industrial Relations*, ed. L. Hausman. Madison, WI: Industrial Relations Research Association.

1979. Sample selection bias as a specification error. *Econometrica* 47, 153–61.

1980. (With T. MaCurdy.) A life cycle model of female labour supply. *Review of Economic Studies* 47, 47–74.

1981a. Heterogeneity and state dependence. In *Studies in Labor Markets*, ed. S. Rosen. Chicago:

University of Chicago Press.

1981b. Statistical models for discrete panel data. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. Manski and D. McFadden. Cambridge, MA: MIT Press.

1982. (With C. Flinn.) New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics* 18, 115–68.

1984. (With B. Singer.) A method for minimizing the impact of distributional assumption in econometric models for duration data. *Econometrica* 52, 271–320.

1985. (With G. Sedlacek.) Heterogeneity, aggregation and market wage functions: an empirical model of self-selection in the labor market. *Journal of Political Economy* 93, 1077–125.

1987. (With J. Scheinkman.) A method for minimizing the impact of distributional assumption of earnings. *Review of Economic Studies* 54, 243–55.

1989. (With B. Payner.) Determining the impact of federal anti-discrimination policy on the economic status of blacks: a study of South Carolina. *American Economic Review* 79, 138–77.

1989. (With R. Butler and B. Payner.) The impact of the economy and the state on the economic status of blacks: a study of South Carolina. In *Markets and Institutions*, ed. D. Galenson. Cambridge: Cambridge University Press.

1989. (With B. Honoré.) The identifiability of the competing risks model. *Biometrika* 76, 325–30.

1990. (With B. Honoré.) The empirical content of the Roy model. *Econometrica* 58, 1121–49.

1990. Varieties of selection bias. *American Economic Review* 80, 313–8.

1991. (With J. Donohue.) Continuous vs. episodic change: the impact of affirmative action and civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29, 1603–43.

1993. (With S. Cameron.) The nonequivalence of high school equivalents. *Journal of Labor Economics* 11, 1–47.

1997. (With H. Ichimura and P. Todd.) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies* 64, 605–54.

1997. (With J. Smith, J. and N. Clements.) Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Review of Economic Studies* 64, 487–

535.

1998. (With H. Ichimura, J. Smith and P. Todd.) Characterizing selection bias using experimental data. *Econometrica* 66, 1017–98.

2001. (With Y. Rubinstein.) The importance of noncognitive skills: lessons from the GED testing program. *American Economic Review* 91, 145–9.

2002. (With J. Donohue and P. Todd.) The schooling of southern blacks: the roles of legal activism and private philanthropy, 1910–1960. *Quarterly Journal of Economics* 117, 225–68.

2002. (With Y. Rubinstein and J. Hsee.) The GED is a ‘mixed signal’: the effect of cognitive and noncognitive skills on human capital and labor market outcomes. Working paper, University of Chicago.

2005. (With E. Vytlacil.) Structural equations, treatment, effects and econometric policy evaluation. *Econometrica* 73, 669–738.

How to cite this article

Blundell, Richard, Lars Peter Hansen and Derek Neal. "Heckman, James (born 1944)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000159> doi:10.1057/9780230226203.0716

The New Palgrave Dictionary of Economics Online

Heckscher, Eli Filip (1879 – 1952)

C.G. Uhr

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic history; Heckscher, E. F.; Heckscher–Ohlin trade theory; imperfect competition; pure public goods

Article

Born into a Jewish family in Stockholm, Heckscher studied history under Hjärne and economics under Davidson at Uppsala University from 1897. In 1907 he became a docent at Stockholm University College of Commerce, and from 1909 to 1929 he was professor of economics and statistics. Then, because of his great research productivity, the college authorities changed his position to research professor, lightened his teaching duties and made him director of the newly established Institute of Economic History. Heckscher continued in this position until he retired in 1945. He succeeded in establishing economic history as a subject of graduate study in Sweden's universities.

In 1950, the Ekonomisk-historiska Institutet, Stockholm, through Bonniers Co., published the *Eli F. Heckscher bibliografi 1897–1979* (123 pp.). It contains 1148 entries for his 36 books, 174 articles in professional journals, his chapters in government reports, and the more than 700 short articles he wrote for the weekend issues of Stockholm's leading newspapers. Only a few of his books and articles have been translated and will be referred to by their English titles; other works will be mentioned only by the English translation of their original titles and identified by their numbers as entries in the Heckscher bibliography.

By 1929, when he was able to specialize in economic history, Heckscher had already written a dozen books on such diverse subjects as *Economic Principles* (1910, No. 158), *The Continental System* (1918, No. 443, later republished, Oxford, 1922) and *Economics and History* (1922, No. 478). As a result of his teaching, his contributions to economics are a blend of innovations in economic theory and a new methodology for economic history research, an approach to quantitative research very different from that used by leaders in his field such as Schmoller, Cunningham and Sombart.

Heckscher's most significant contributions to economic theory may be found in two articles. ‘Effects of Foreign Trade on Distribution of Income’ (1919) is the origin of the modern Heckscher–Ohlin factor

proportions theory of international trade, developed further in Ohlin (1933).

‘Intermittently Free Goods’ (1924) presents a theory of imperfect competition nine years ahead of that by Joan Robinson and Edward Chamberlin, and a discussion of collective goods not priced by the market. Heckscher observed that significant new products are introduced by firms with investment in plants that have a capacity which far exceeds initial demand for their products. The latter are sold at prices which barely cover unit variable costs, and so, for a time, the services of the fixed investment are provided as ‘free goods’. Then, as weaker firms are eliminated, demand shifts to the remaining larger firms who use up their production capacity. By and by these firms expand, enjoy economies of scale, differentiate their products and become prosperous oligopolies dividing a mass market into more or less definite shares.

A situation of the opposite kind arises when the smallest feasible production facility has a production capacity which suffices for a growing and indefinitely large demand without affecting the costs and service life of the production unit. This is the case with many so called ‘pure’ public or collective goods. Heckscher used street illumination as an example of a collective good, which can be used simultaneously by few or many persons, a service that cannot be priced per unit of individual use. The costs of providing this service, then, are usually met by an increase in local government taxes. In that case, and in contrast to that of intermittently free goods, the citizens pay the full-cost price of the service from the outset in their current taxes. Then, as activities in and use of lighted streets increase over time, the citizens derive increased utility per tax dollar spent for street lighting.

At the Institute of Economic History Heckscher's first work, one of his major and most widely known treatises, was *Mercantilism* (1931). His other major work, the fruit of many years of pioneering research devoted to his own country, was *Sweden's Economic History from the Reign of Gustav Vasa* (vols 1 and 2, 1935–6, No. 878; vols 3 and 4, 1950, No. 1146). He also wrote a popular version of this work, *Life and Work in Sweden from Medieval Times to the Present* (1941), No. 1014, republished as *An Economic History of Sweden* (1954). Among his other books of particular interest are *Materialist and Other Interpretations of History* (1944, No. 1052), *Industrialism, Its Development from 1750 to 1914* (1946, No. 1123) and *Studies of Economic History* (1936, No. 918). It was in this work he presented a new methodology he proposed for economic history research in his essay ‘The Aspects of Economic History’, pp. 9–69. This was reinforced in his articles, ‘A Plea for Theory in Economic History’ (1929) and ‘Quantitative Measures in Economic History’ (1939).

For the analysis of any epoch of economic history – as distinct from factual description of a chronologically arranged body of heterogeneous source materials – Heckscher proposed consideration of a succession of its ‘economic aspects’, to introduce order and inject economic theory into the interpretation of that epoch. Unlike ‘periods’, ‘aspects’ are *not necessarily* time dependent. They are theoretical and imply hypotheses that are, within limits, testable against the given data. A series of aspects, for instance of (*a*) the exchange processes; (*b*) natural resources and technologies; (*c*) labour force and capital; (*d*) forms of enterprise organization; and (*e*) extent and composition of demand, form an economic model of the epoch. This done, the function of the economic historian is to provide a synthetic overview and explanation of the relations between the aspects of the model.

Thus Heckscher bridged the gap between economic history and theory by addressing broad questions or hypotheses to the source materials for intensive and critical study. He always preferred to present his finding supported by statistical data. That done, he was not satisfied until he had explained and illuminated these by economic analysis, that is, by applying cognate principles of economic theory to

their interpretation.

See Also

- Heckscher–Ohlin trade theory
- Stockholm School

Selected works

1918. *Kontinentalsystemet*. Stockholm: P.A. Norstedt & Söner. Trans. as *The Continental System*, ed. H. Westergaard, Oxford: Clarendon Press, 1922.

1919. Effects of foreign trade on distribution of income. *Ekonomisk Tidskrift*. Reprinted in AEA, *Readings in the Theory of International Trade*, ed. H.S. Ellis and L. Metzler, Philadelphia: Blakiston, 1949.

1924. Intermittently free goods. *Ekonomisk Tidskrift* 553. Reprinted in German in *Ein Beitrag zur Sozialwissenschaft und Sozialpolitik* 59 (1928).

1929. A plea for theory in economic history. *Economic Journal*, Historical Supplement No. 4, 523–54.

1931. *Merkatilismen*, 2 vols. Stockholm: P.A. Norstedt & Söner. Authorized trans. by M. Shapiro as *Mercantilism*, London: G. Allen & Unwin, 1935. Revised edn, New York: Macmillan; London: G. Allen & Unwin, 1955.

1935–6, 1949. *Sveriges Ekonomiska Historia från Gustav Vasa*. Vols. 1 and 2, 1935–6; vols. 3 and 4, 1949, Stockholm: Bonnier.

1939. Quantitative measures in economic history. *Quarterly Journal of Economics* 53, 167–93.

1941. *Svenskt arbete och liv från medeltiden till nutiden*. Stockholm. Trans. G. Ohlin as *An Economic History of Sweden*, Cambridge, MA: Harvard University Press, 1954.

1950. *Eli F. Heckscher bibliograft 1897–1949*. Stockholm: Bonnier.

Bibliography

Ohlin, B. 1933. *Interregional and International Trade*. Cambridge, MA: Harvard University Press.

How to cite this article

Uhr, C. G. "Heckscher, Eli Filip (1879–1952)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000033> doi:10.1057/9780230226203.0717

The New Palgrave Dictionary of Economics Online

Heckscher – Ohlin trade theory

Ronald W. Jones

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Heckscher–Ohlin trade theory consists of four principal theorems, viz. the Heckscher–Ohlin trade theorem whereby relatively capital-abundant countries export relatively capital-intensive commodities, the factor-price equalization theorem whereby trade in goods may serve to equalize wage rates between countries, the Stolper–Samuelson theorem whereby an increase in the price of the relatively labour-intensive commodity unambiguously improves the real wage rate, and the Rybczynski theorem stating that an increase in capital endowment by itself must cause some output to fall if prices are held constant. The article discusses the nature and fate of these theorems.

Keywords

autarky; comparative advantage; dimensionality; distribution of income; factor-intensity reversal; factor-price equalization theorem; free trade; general equilibrium; Heckscher–Ohlin theorem; Hecksher–Ohlin trade theory; international trade theory; labour–capital ratio; Leontief paradox; Metzler tariff paradox; new trade theory; reciprocity relationship; relative factor abundance; relative factor intensity; Rybczynski theorem; specialization; Stolper–Samuelson theorem; tariffs; terms of trade

Article

Eli Heckscher (1919) and Bertil Ohlin (1933) laid the groundwork for substantial developments in the theory of international trade by focusing on the relationships between the composition of countries' factor endowments and commodity trade patterns as well as the consequences of free trade for the functional distribution of income within countries. From the outset general equilibrium forms of analysis were utilized in these developments, which gradually came to be sorted out into four 'core propositions' (Ethier, 1974) in the pure theory of international trade.

The four theorems

Although all four of the propositions to be discussed are an outgrowth of the seminal work of Heckscher and Ohlin, only one of these propositions bears their name explicitly. The *Heckscher–Ohlin theorem* states that countries export those commodities which require, for their production, relatively intensive use of those productive factors found locally in relative abundance. The twin concepts of relative factor intensity and relative factor abundance are most easily defined in the small dimensional context in which the basic theory is usually developed. Two countries are engaged in free trade with each producing the same pair of commodities in a purely competitive setting, supported by constant returns to scale technology that is shared by both countries. Each commodity is produced separately with inputs of two factors of production that, in each country, are supplied perfectly inelastically. (For a throrough analysis of having endowments respond endogenously, see Findlay, 1995). Following the Ricardian distinction, commodities are freely traded but productive factors are internationally immobile.

Although one country may possess a larger endowment of each factor than another, the presumed absence of returns to scale guarantees that only relative factor endowments are important. The home country is said to be relatively labour abundant if the ratio of its endowment of labour to that (say) of capital exceeds the corresponding proportion abroad. This is known as the physical version of relative factor abundance. An alternative involves a comparison of autarky relative factor prices in the two countries: the home country can be defined to be relatively labour-abundant if its wage rate (compared with capital rentals) is lower before trade than is the foreign wage (relative to foreign capital rentals). Since autarky factor prices are determined by demand as well as supply conditions, these two versions need not correspond. In particular, if the home country is, in the physical sense, relatively labour abundant it might nonetheless have its autarky wage rate relatively high if taste patterns at home are strongly biased towards the labour-intensive commodity compared with tastes abroad. In such a case the trade pattern reflects the autarky factor–price comparison: the home country exports the physically capital-intensive commodity. As discussed below, the link between commodity price ratios (the proximate determinant of trade flows) and factor price ratios is more direct than that between commodity price ratios and physical factor endowments. Thus the Heckscher–Ohlin theorem is more likely to hold if relative factor abundance is defined in terms of relative factor prices prevailing before trade. The procedure typically followed in the literature is to assume that both countries share identical and homothetic taste patterns. Such an assumption, in conjunction with the presumed identity of technology at home and abroad (with an even stronger version of homotheticity–linear homogeneity) helps to isolate the separate influence of physical factor supplies and makes the validity of the Heckscher–Ohlin theorem with the physical definition of factor abundance as likely as with the autarky factor price definition.

These assumptions are less than sufficient to guarantee the Heckscher–Ohlin theorem, even in the simple context of two-country, two-factor, two-commodity trade. The potential stumbling block is the fact that even though countries share the same technology, the commodity that is produced by relatively labour-intensive techniques at home may be produced by relatively capital-intensive techniques abroad. This is the phenomenon of factor-intensity reversal. If production processes are independent of each other, there is nothing (other than bald assumption) to rule out its appearance. The bald assumption would assert that regardless of factor endowments one industry always employs a relatively higher ratio of labour to capital than does the other industry, where techniques are chosen with reference to the wage/rental ratio common to both industries. If this is not the case, and if the commodity that is relatively labour-intensive

at home is produced by relatively capital-intensive techniques abroad, the phrasing of the Heckscher–Ohlin theorem that explicitly states ‘each country exports the commodity that is produced in that country making relatively intensive use of the factor found in relative abundance in that country’ is fatally flawed. The reason? If the relatively labour-abundant country exports its labour-intensive commodity, it must do so in exchange for the commodity that, in the relatively capital-abundant foreign country, is produced by labour-intensive techniques. Thus if one country satisfies the theorem, the other country cannot (Jones, 1956).

In the event of factor-intensity reversal, it must be the case that, whatever the commodity exported by the labour-abundant home country, the ratio of labour to capital employed in its production must exceed the labour/capital intensity adopted in foreign exports. However, this observation is of little value if one wishes to infer from an intensity comparison between exportables and import-competing goods within a given country whether that country is more labour abundant than some foreign country. Such an inference lay behind the celebrated study of Leontief (1953) on United States trade patterns. This research, the conclusions of which came to be known as the Leontief paradox (American exportables are produced by more labour-intensive techniques than are import-competing goods) provided the major stimulus to developing and defining the meaning and conditions supporting the Heckscher–Ohlin theorem.

Earlier work in Heckscher–Ohlin trade models was focused on the pricing relationships embodied in Heckscher–Ohlin theory. Ohlin (1933) stressed the effect which free trade would tend to have on the distribution of income within countries, viz. relative factor prices would move in the direction of equality between trading countries which share the same technology. Ohlin's mentor, Heckscher, went even further in his pioneering 1919 article. Absolute factor-price equalization was purported to be ‘an inescapable consequence of trade’. (For recent appraisals of each of these economists see Jones, 2002; 2006a). Nonetheless, Ohlin's view of partial equalization seems to have dominated, with the exception of Lerner's unpublished 1933 manuscript (which surfaced after Samuelson's articles), until the statement of the factor-price equalization theorem in articles by Samuelson in 1948 and 1949. Rejecting his earlier tacit acceptance of the Ohlin thesis of partial equalization (in the Stolper–Samuelson article, which appeared in 1941), Samuelson proved that within the traditional confines of the 2×2×2 model (with no factor-intensity reversals and each country incompletely specialized), free trade would drive wage rates to absolute equality in the two countries (and, as well, would equate returns to capital) despite the assumption that labour (and capital) are assumed to be immobile between countries.

The logic of the argument for the simple 2×2 case can be stated briefly. In a competitive equilibrium unit cost equals price if the commodity is produced. Thus let A represent the matrix of input–output coefficients, aij, w the vector (pair) of factor prices, and p the vector (pair) of commodity prices. Techniques need not be constant; in general they depend upon prevailing factor prices so that $A=A(w)$. Therefore the competitive profit conditions if both goods are actually produced dictate that:

$$A(w) \cdot w = p. \tag{1}$$

If we assume no factor-intensity reversals, $A(w)$ is non-singular. Therefore if countries share the same technology and face the same pair of free-trade commodity prices, they must face exactly the same set of factor prices if each country produces both goods.

This approach may suggest that the crucial issue in the factor-price equalization argument is the unique dependence of factor price vector w on commodity price vector p, and an extensive literature has developed which focuses on this issue. In the 2×2 case uniqueness is a simple question – it depends on factor intensities differing between sectors and not reversing. But from the outset Samuelson pointed out that this was not the only issue. The question of uniqueness involves properties of technology alone, whereas under appropriate circumstances two countries in free trade will have factor prices equalized only if factor endowments are reasonably similar. For, if factor endowments are too dissimilar, it will be impossible for both countries to produce both commodities, in which case the equalities in (1) cannot universally hold.

These ideas can be made more precise by considering a concept due to McKenzie (1955), which Chipman (1966) called the ‘cone of diversification’. For any factor price vector, w, there is determined a pair of techniques (labour/capital ratios) for the two commodities. Both factors can be fully employed only if the country's endowment vector is contained within the cone spanned by these techniques. Suppose two countries face a common free-trade commodity price vector, p, and that the commonly shared technology associates a unique factor price w corresponding to this p. Then if the endowment vectors of both countries lie within the cone of diversification, their factor prices must be equalized (McKenzie, 1955).

Some seven years prior to Samuelson's first factor-price equalization essay there appeared the article by Stolper and Samuelson (1941), which must be ranked a classic not only for its discussion of what became known as the *Stolper–Samuelson theorem*, but because it is one of the first concrete developments of the ideas of Heckscher and Ohlin in the explicit format of a two-factor, two-commodity, general equilibrium model. (This theorem became so widely cited that on the golden anniversary of its appearance a conference was held at Stolper's university, the University of Michigan. See Alan Deardorff and Robert Stern, 1994.) Their argument supposedly concerns the effect of protection on real wages, and in the course of the argument they assume that a tariff does not change the terms of trade so that locally import prices rise. Subsequently, in what has become known as the ‘Metzler tariff paradox’, Metzler (1949) showed that with sufficiently inelastic demand a tariff might so improve a country's terms of trade that the relative domestic price of imports falls. If so, the Stolper–Samuelson contention that a tariff yields an increase in the real return to a country's relatively scarce factor would be reversed. However, it is now commonly agreed that the Stolper–Samuelson theorem refers to the general phenomenon whereby an increase in the relative domestic price of a commodity (whether brought about by a tariff increase, decrease, or some other reason) must unambiguously raise the real return to the factor of production used relatively intensively in the production of that commodity. Introducing the production-box diagram technique (for a single country), Stolper and Samuelson illustrate how an increase in the relative price of labour-intensive watches attracts resources from capital-intensive wheat. To clear factor markets, both sectors must then use labour more sparingly. That is, the ratio of capital to labour utilized in each sector rises, which implies an unambiguous increase in labour's marginal productivity measured either in watches or in wheat. Thus regardless of workers’ taste pattern, protection has increased the real wage.

The logic of the Stolper–Samuelson argument rests heavily upon the presumed absence of joint production. It takes labour and capital to produce watches, and, in a separate activity, a higher capital/labour ratio is used to produce wheat. In competitive settings any change in a commodity's price must reflect an average of factor price changes so that unit costs change as much as do prices. Therefore one factor price must rise relatively more than either commodity price. Which factor gains depends only upon the factor-intensity ranking. If the price of watches rises, and that of wheat does not, the wage rate must increase by relatively more. And this result follows even if techniques are frozen so that no resources can be transferred between sectors (as they can be in the Stolper–Samuelson discussion) and marginal products are not well defined (Jones, 1965).

To round out the quartet of theorems, the Rybczynski theorem (1955) deals with the same model but focuses on the relationship between factor endowments and commodity outputs. Suppose commodity prices are kept fixed in the 2×2 setting and an economy is incompletely specialized. Then by the factor-price equalization theorem, factor prices are determined and fixed as well, which implies also that techniques of production remain constant. If the economy's endowment of one factor increases, while its endowment of the other factor remains constant, the economy must in some sense grow (the transformation schedule shifts out). However, this growth is strongly asymmetric: one output actually falls. The factor-intensity ranking selects the loser – the commodity that uses intensively the factor that is fixed in overall supply must decline. The reasoning is simple. As one factor expands, it must be absorbed in producing the commodity using it intensively. But with techniques frozen (since prices are assumed fixed), the expanding sector must be supplied with doses of the non-expanding factor as well. The only source for this factor is the other industry that must, perforce, contract.

Relationships among the theorems

All four propositions are based on the same ‘mini-Walrasian’ general equilibrium model of trade and there are some interesting relationships and distinctions among them. Perhaps most importantly, both the Heckscher–Ohlin theorem and the factor-price equalization theorem refer explicitly to a comparison between (two) countries, whereas the Stolper–Samuelson and Rybczynski propositions are involved with relationships within a single country. This distinction implies that the assumption that countries share an identical technology is not necessary for the latter two propositions. Thus, for example, a country could protect the factor used intensively in its import-competing sector in real terms (according to Stolper and Samuelson) regardless of the level or type of technology adopted by other countries.

The factor-price equalization theorem is a razor's-edge type of result. Should the technology available to two countries differ only slightly, any presumption of exact factor-price equalization in the absence of explicit international factor markets disappears. The Heckscher–Ohlin theorem is a little more robust in this regard. In general, trade patterns depend on all those variables that influence prices: tastes, technology, and factor endowments (not to mention taxes or other distortions). If tastes are identical (and homothetic) but factor endowments are not, the latter difference will tend to dominate the trading pattern even if technologies differ as long as this difference is ‘less important’. At issue is a weighing of endowment differences with the Ricardian emphasis on technology differences. A particular variation of the factor-price equalization theorem is more general, and does not need to assume that technologies are identical between countries. It concerns the *dependence* of factor prices only upon commodity prices. It follows as long as the country produces both commodities.

Two versions of the Heckscher–Ohlin theorem have been cited, depending on which definition of relative factor abundance is selected. If the physical factor intensity ranking is chosen as the criterion, the basis for the Heckscher–Ohlin theorem resides in the kind of link between endowment patterns and outputs for a single economy exemplified by the Rybczynski theorem. An extension of this theorem allows a comparison of the transformation schedules for two economies with similar technologies. The relatively labour abundant (physical definition) country will produce relatively more of the labour-intensive commodity at common commodity prices (Jones, 1956). Therefore, unless taste differences are sufficiently biased to counter this effect, the labour-intensive good will, in autarky, be cheaper in the labour-abundant country and, with trade, will be exported. The Stolper–Samuelson theorem is closely linked to the alternative form of the Heckscher–Ohlin theorem. Suppose there are no factor-intensity reversals. Then if both goods are produced there is a monotonic relationship between the wage/rent ratio and the relative price of the labour-intensive good such that a rise in the latter is associated with a greater than proportionate increase in the former. Thus the relatively low wage country must, in autarky, have been the relatively cheap producer of the labour-intensive commodity. As mentioned earlier, no caveat must be added about tastes, since these are already incorporated in the autarky factor-price comparison.

Although a comparison of factor endowments between countries is crucial in considering both the Heckscher–Ohlin theorem and the factor-price equalization theorem, such a comparison works in opposite directions for these two propositions. Thus if factor endowment proportions are sufficiently *dissimilar*, trade patterns suggested by the Heckscher–Ohlin theorem *must* hold (aside from the possibility of factor-intensity reversals) whereas free trade *cannot* bring about factor-price equalization. Sufficiently different factor endowments entail one country's transformation schedule being everywhere flatter than the other country's. At least one country must be specialized with trade. By contrast, the factor-price equalization result holds if factor endowments are similar enough so that international differences in the composition of outputs are capable of absorbing these endowment differences at the same set of techniques (and factor prices). If endowments are this close, it would always be possible for demand differences to be so biased that the physically labour-abundant country exports the capital-intensive commodity. Indeed, if such a demand reversal of the Heckscher–Ohlin theorem takes place, free trade must result in factor-price equalization (Minabe, 1966).

Samuelson's name occurs so frequently in the literature on Heckscher–Ohlin trade theory that it is often appended to the other two names. One of his results not cited heretofore is the reciprocity relationship (Samuelson, 1953). This states that in any general equilibrium model the effect of an increase in a commodity price (say p_j) on a factor return (say w_i) is the same as the effect of an increase in the corresponding factor endowment (V_i) on the output of commodity j. Of course, in each case some other set of variables is being held constant. Thus:

$$\frac{\partial w_i}{\partial p_j} = \frac{\partial x_j}{\partial V_i} \tag{2}$$

with all other commodity prices and all endowments held constant in the left-hand derivative and all other endowments and all commodity prices held constant in the right-hand derivative. This relationship is easy to prove (see, for example, Jones and Jose Scheinkman, 1977). It also reveals the *dual* nature of the Stolper–Samuelson and Rybczynski theorems. If an increase in the price of watches lowers capital returns, then an increase in the endowment of capital (at constant prices) would lower the output of watches. In each case it is the presumed labour-intensity of watches that is operative.

In the 2×2 setting both the Stolper–Samuelson and Rybczynski theorems reflect the ‘magnification effects’ (Jones, 1965) that stem directly from the assumed lack of joint production. With a ‘^’ over a variable designating relative changes, if watches are labour intensive and wheat capital intensive and if the relative price of watches rises,

$$\hat{w} > \hat{p}_{wa} > \hat{p}_{wh} > \hat{r}. \tag{3}$$

In addition, should an economy grow, but with labour (L) growing more rapidly than capital (K),

$$\hat{x}_{wa} > \hat{L} > \hat{K} > \hat{x}_{wh}. \tag{4}$$

Inequality ranking (3) shows commodity price changes trapped between factor-price changes (since two factors are required to make a single good), while inequality (4) shows that in order to absorb endowment changes, the composition of outputs (each of which uses both factors) must change more drastically. Stolper and Samuelson stressed the first inequality in (3), while Rybczynski focused on the last inequality in (4), assuming $\hat{K}$ equals zero.

Higher dimensions

International trade theory generally, and Heckscher–Ohlin trade theory in particular, has frequently been criticized for its restriction to the low dimensionality represented by two commodities, two factors, and two countries. In fairness to both Heckscher and Ohlin it should be stressed that their discussions typically were not so confined. But neither were their conclusions as precise as those subsequently developed by Samuelson and others in the 2×2×2 versions of the four core propositions. And in the years following Samuelson's pioneering work on factor-price equalization, scores of articles have indeed appeared dedicated to the question of robustness of these results in higher-dimensional contexts. A highly detailed discussion of the issue of dimensionality is provided in Ethier (1984), and an earlier critique of the limitations imposed by small numbers of goods and factors is found in Ethier (1974) and

Jones and Scheinkman (1977).

Part of the difficulty embedded in the move to higher dimensions lies in the ambiguity involved in what the propositions should state for cases beyond 2×2. The one proposition for which this is not the case is the factor-price equalization theorem. Consider the case of equal numbers of factors and produced commodities, with all goods traded and factors immobile internationally. The uniqueness of a factor price vector, w, corresponding to a given commodity price vector, p, is not guaranteed even in the 2×2 case; a factor-intensity reversal could lead to two (or more) values of w consistent with a given p. For the $n \times n$ case Gale and Nikaido (1965) provided conditions sufficient to guarantee global univalence of the factor price–commodity price relationship: the $A(w)$ matrix of input–output coefficients should be a 'P-matrix', that is a matrix with all positive principal minors. These conditions have been slightly weakened by Andreu Mas-Colell (1979), and earlier a fundamental interpretation of the conditions was supplied by Yasuo Uekawa (1971). It remains the case, however, that this condition on technology alone is somewhat remote from the issue of factor-price equalization. Just as in the 2×2 case, two countries sharing a common technology and each capable of producing the same set of n commodities (at the same traded-goods prices) with n productive factors, will, if techniques of production are the same, have their factor prices driven to equality if their factor endowments are sufficiently close. The concept of the 'cone of diversification' within which both endowment vectors must lie for factor-price equalization is as meaningful and relevant in n dimensions as it is in two.

Although the factor-price equalization theorem has an unambiguous meaning in higher dimensions, it is a theorem that cannot be expected to hold if the number of productive factors exceeds the number of freely traded commodities. The reasoning is basic, and can be linked to eq. (1). These competitive profit conditions supply n links between factor prices and traded commodity prices, where n is the number of traded commodities. If r, the number of factors, should exceed n, the relationships in eq. (1) are insufficient in number to provide a solution for the vector w for given p. Other conditions are required, and these are provided by the full employment conditions, one for each productive factor. Thus a nation's endowment bundle, V, becomes a determining variable affecting factor prices that is additional to the commodity price vector, p. For example, in the simple three-factor, two-commodity 'specific-factor' model (Jones, 1971; Samuelson, 1971), suppose a country faces a given world price vector, p, and experiences a slight increase in its endowment of a factor 'specific' in its use in the first industry. The intensity with which factors are utilized depends upon factor prices, and if these do not change, there is no way in which outputs can adjust to clear all factor markets. The return to the factor specific to the first industry must fall so as to encourage the further use of that factor. Two countries of this type with different endowments will generally have different sets of factor prices with trade, even if they share a common technology. It may be interesting to note that Heckscher's (1919) discussion of the necessity of factor-price equalization is focused on a three-factor, two-commodity numerical illustration (Jones, 2006a). As just suggested, such a 3×2 setting in general does *not* lead to factor-price equalization.

The Heckscher–Ohlin model with two factors but many commodities available in world markets provides a useful scenario in which to re-examine the Heckscher–Ohlin theorem concerning the pattern of trade. The strong influence of factor endowments on production and trading patterns is revealed by considering two countries sharing the same technology but with different endowment ratios. Suppose commodity prices for traded goods are determined in a world market composed of a number of different countries with potentially a wide variation in technologies. Given world prices, any pair of countries

with the same technologies shares a Hicksian composite unit-value isoquant for all traded goods (Jones, 1974), made up of strictly bowed-in sections (where only one commodity is produced) alternating with flats (where a pair of commodities is produced). Regardless of the number of commodities, each country engaged in trade need produce only one or two (in the two-factor case), and these commodities will be the ones requiring factors in proportions close to that country's endowment ratio. (Not explicitly considered here is that the activity of exporting a commodity may require factor proportions different from those required in production; see Jones, Beladi and Marjit, 1999.) In this setting the spirit of the Heckscher–Ohlin theorem is that each country concentrates its resources on a small range of commodities whose factor requirements mirror closely that country's endowment base; the country exports some or all commodities in this set and imports commodities that are more labour-intensive than these goods as well as those that are more capital-intensive. Two countries whose endowments are fairly similar *may* produce the same pair of goods and thus achieve factor-price equalization with trade. Countries further apart in endowment composition will have disparate sets of factor prices and may produce completely different bundles of commodities (see also Krueger, 1977).

With many factors and many commodities a different approach can be taken. The ability of autarky commodity price comparisons to predict trade patterns item by item is severely questioned, so that little hope remains of linking endowment differences to the detailed composition of trade. But statements about aggregates or ‘correlations’ between trade patterns and autarky prices can be made (Deardorff, 1980; Dixit and Norman, 1980). A nation's net imports, M, are positively correlated with the comparison of its autarky commodity price vector, p^{A}, and the vector of free-trade commodity prices, p^{T}. Thus:

$$(p^{A} - p^{T})M \geq 0 \tag{5}$$

(see Ethier, 1984, p. 139). This idea can be extended to the further relationship between autarky commodity prices and the vector of autarky factor prices (as in (1)) to establish that countries possess a comparative advantage, on average, in commodities using intensively factors that are relatively cheap in autarky. (See Deardorff, 1982, and Ethier, 1984, for more details.)

The reciprocity relationship expressed in (2) is quite general in terms of dimensionality and thus serves to link the Rybczynski theorem in a dual relationship to the Stolper–Samuelson theorem. However, when the number of factors exceeds the number of produced commodities, differences between the two types of theorems do appear. This basic asymmetry is linked to the failure of the factor-price equalization theorem when factors exceed commodities in number.

Major efforts have been made to generalize the Stolper-Samuelson and Rybczynski theorems from the 2×2 settings to the $n \times n$ setting, and a pair of earlier efforts met with only limited success. Murray Kemp and Leon Wegge (1969) searched for conditions on the original activity matrix, A, or distributive share matrix, θ, that would be sufficient to ensure what is known as the *strong* form of the Stolper–Samuelson theorem: Each factor is associated with a unique commodity such that if that commodity price (alone) increases, the return to the associated factor increases by a relatively greater extent and all

other factor returns fall. The conditions they tried are stated by the inequalities in (6):

$$\theta_{ii} / \theta_{ki} > \theta_{ij} / \theta_{kj} \text{ for all } i,\ j \neq i, \text{ and } k \neq i. \tag{6}$$

For each factor, *i*, the ratio of its distributive share in the industry positively associated with that factor (*i*) to that of any other factor's share in industry *i*, exceeds the corresponding ratio of these two factors in any other industry. These strong conditions do indeed lead to the desired strong result on factor returns (that is the inverse of the distributive share matrix has positive diagonal terms, greater than unity, and negative off-diagonal elements) for the 3×3 case. However, the authors provided a counter-example for the 4×4 case and that was that. Even stronger conditions for sufficiency are required, and these were supplied by Jones, Marjit, and Mitra (1993). These conditions are, in a sense, suggested by the statement of the theorem that for any price change all factor returns except one must fall. That is, they must have a relatively similar fate, suggesting fairly similar intensity use. The inequality that suffices is shown in (7):

$$\theta_{ii} / \theta_{ki} - \theta_{ij} / \theta_{kj} > \Sigma_{s \neq k, i, j} \left| \theta_{si} / \theta_{ki} - \theta_{sj} / \theta_{kj} \right| \text{for all } i,\ j,\ k \neq i. \tag{7}$$

That is, condition (6) is not strong enough; the difference between the two terms in (6) must exceed the absolute value of similar differences in all the unintensive factors (whose returns all must fall). As occasionally happens, an article by John Chipman (1969) in the same issue of the same journal provided a condition sufficient for a weaker result, namely that the elements along the diagonal all be positive and exceed unity, regardless of signs off the diagonal. His condition met the same fate – sufficient for the 3×3 case but not higher. Mitra and Jones (1999) provided a sufficient condition for the *n*×*n* case.

It is possible to argue that these conditions are so strong as to suggest the Stolper–Samuelson and Rybczynski theorems really do not generalize. However, there is a form of the Stolper–Samuelson theorem that does generalize to higher dimensions with relatively little structure and, arguably, captures the essence of the original 1941 result. Stolper and Samuelson addressed the question of a particular government policy on real wages – the imposition of a tariff. But consider a more general question. Suppose an arbitrary factor of production seeks government aid sufficient to have its *real return* improved in a *non-transparent* fashion, that is, without a direct payment (out of tariff or any other source of government revenue). It is to be done by changes in taxation or government demand that would affect commodity prices. What would be required? Very little, as shown in Jones (1985; 2006b). Suppose there is little or no joint production (to be discussed below), and that there is a sufficient number of commodities (at least equal to the number of factors). These conditions suffice to ensure that there exists a subset of commodities that, if their prices are raised by the same relative amount with no other

commodity price changes, the real return to the particular factor is guaranteed to increase. This result should have pride of place in the field of political economy and represents a significant generalization of the original Stolper–Samuelson result. The kind of detailed requirements shown by (6) and (7) is not necessary as long as a single commodity is not by itself required to do the job and as long as (along with Chipman, 1969) it is not required that all other factors lose. Indeed, the favoured factor might well appreciate not standing nakedly as the only winner.

There are a few special cases of the $n \times n$ Heckscher–Ohlin setting that deserve mention. The most important might be the contribution of Roy Ruffin (1988). Ruffin redefines the Ricardian setting in which each country has a distinct labour force whose productivities in producing a number of goods differ from those in other countries. Instead of having each type of labour restricted to a single country, Ruffin suggests letting each country be populated by a wide variety of labour types, with the relative supplies of each type differing from country to country. This shifts the focus to relative endowment differences among countries, with the same technologies (a single type of labour is the same no matter where located). The key feature of such a model is that there is not only no joint production of outputs, there is no need for any single factor to have to work jointly with any others to produce commodities. As a consequence, factor prices are always equalized by free trade in commodities. Furthermore, each country's transformation surface looks just like that of a *world* transformation surface in the Ricardian model. In the two-commodity case this is a broken, bowed out, join of the two linear schedules for each country. In higher dimensions there are various dimensional ‘facets’ down to those of zero-dimension, that is, points at which each type of labour is fully employed in a different commodity. Except for the relative size of these facets, each country's transformation curve looks like that of any other country. At given commodity prices the common ‘price plane’ is ‘tangent’ to each surface such that each labour type is assigned to the same commodities in any country. At free-trade prices the relative production pattern in any country exactly mirrors the relative labour supplies and productivity of labour in that country.

Another special version, one that does give the Kemp–Wegge strong results, is the ‘produced mobile factor’ structure introduced by Jones and Marjit (1985; 1991). Imagine an $\{(n+1) \times n\}$ specific-factors structure, with n specific factors and a single factor mobile between sectors. This is often taken to be labour, but instead, suppose it is a mobile input that is produced by all the specific factors. (That is, each ‘specific’ factor produces a particular commodity and, in addition, joins with other factors to produce the mobile factor.) This reduces to an $n \times n$ model with strong Stolper–Samuelson properties.

Fred Gruen and Max Corden (1970) introduced a simple model in discussing the possibility that a country such as Australia might, in levying a tariff, worsen its terms of trade. There are two sectors in the economy, manufacturing and agriculture. The manufacturing sector consists of a single commodity produced by labour and capital. Agriculture has two commodities, wheat and wool, each using labour and land. Thus, this is a special form of 3×3 model. As developed by Jones and Marjit (1992), it is possible to consider the $n \times n$ version of the Gruen–Corden model in which $(n - 2)$ sectors of the economy each use mobile labour and a type of capital specific to that sector to produce a single distinct commodity. In another sector labour and a specific type of capital produce a pair of commodities just as in the original Gruen–Corden case. An application of the Gruen and Corden model is also found in Findlay (1993).

The point of each of these special settings is that Heckscher–Ohlin models in the $n \times n$ case need not be difficult to analyse. However, the most popular two-factor model in the many commodity case may well

be more valuable in that it focuses attention on which good or pair of goods a country produces in an international setting. Trade allows a great degree of specialization, and this version of the Heckscher–Ohlin model allows something that the special $\{n \times n\}$ models, as well as the $\{(n+1) \times n\}$ specific-factors model, do not, *viz*. treating the pattern of production as *endogenous* (see also Jones, 2007a; 2007b).

Joint production

Both the Stolper–Samuelson theorem and the Rybczynski theorem are essentially reflections of the asymmetry between factors and commodities. This asymmetry is characterized by the assumption that productive activities are non-joint: in the non-degenerate cases more than one input is required to produce, separately, each output. Thus each commodity price change is a positive weighted average of the changes in rewards to factors used to produce that commodity. This implies that regardless of the ranking of commodity price changes, there is some factor reward that would rise relative to any commodity price rise and at least one factor reward which would rise by relatively less (or fall by more) than any commodity price change. Allowing joint production potentially destroys this asymmetry and thus the basis for the magnification effects.

There is a small literature dealing with this issue (Jones and Scheinkman, 1977; Chang, Ethier and Kemp, 1980; Uekawa, 1984). Much depends on the range of output proportions in any productive activity compared with the range of input proportions. For example, in the 2×2 case suppose one activity produces primarily the first commodity, but also a small amount of the second, while the other activity reverses these proportions. Furthermore, suppose this ‘output’ cone of diversification contains the standard ‘input’ cone of diversification. In this case traditional magnification effects underlying the Stolper–Samuelson and Rybcznyski theorems remain valid. New results emerge if these cones intersect or the input cone contains the output cone. (Cones can be made comparable by using distributive shares of inputs and outputs in activities.)

Joint production does not, by itself, interfere with the status of the factor-price equalization theorem (Jones, 1992; but see Samuelson, 1992, for an alternative view). However, joint production does suggest an alteration of the Heckscher–Ohlin theorem. Instead of concentrating on the link between factor endowments and the location of commodity outputs (and therefore trading patterns), the focus is on the location of productive activities. Each activity requires, as before, an array of inputs, and the allocation of endowment bundles among countries helps to determine where these activities are located. The pattern of commodity trade must then reflect, as well, the output composition of these activities.

Concluding remarks

The theory of international trade that has developed from the seminal writings of Heckscher and Ohlin is fundamentally based on the twin observations that countries differ from each other in the composition of their factor endowments and that productive activities are distinguished by the different relative intensities with which factors are required. As this theory has been developed four core propositions have served to summarize its content. The strict validity of each of these propositions has been seen to depend upon further specification of the technology (for example, ruling out factor intensity reversals,

joint production, and non-constant returns to scale, and imposing, for some results, that countries share the same technology), demand (for example, requiring all individuals to possess identical homothetic taste patterns), or dimensionality (for example, requiring a small number of factors and commodities, or a matching number of both). To conclude this discussion of the core propositions it is possible to point out the less precise, broad message of each.

The Heckscher – Ohlin theorem

Production patterns reflect different compositions of endowments and, unless demand differences are significant, so will patterns of trade. International trade encourages specialization in production in those activities requiring factors in proportions similar to the endowment bundle and allows a country to import commodities whose factor requirements are far from proportions found at home. In some of the writings on ‘new trade theory’, assumptions are made that all varieties of a certain type of product are produced using the same factor proportions. By assumption this rules out the Heckscher–Ohlin theorem as an explanation of trade patterns. However, if varieties differ in quality, each variety could differ in factor requirements as well, serving to re-establish the relevance of the Heckscher–Ohlin theorem.

The factor-price equalization theorem

Even if the international mobility of factors of production is ruled out by national frontiers, free trade in commodities helps to even out disparities in demand relative to supply of factors and to diminish the discrepancy between factor returns among countries. Two or more countries sharing the same technology will find that free trade brings factor returns to absolute equality if their endowments are sufficiently similar and they produce in common a sufficient number of commodities (at least equal to the number of distinct productive factors).

The Stolper – Samuelson theorem

Changes in relative commodity prices, such as those brought about by trade or interferences in trade, have strong asymmetric effects on factor rewards. If no joint production prevails, some factors find their real rewards unambiguously raised and other rewards are unambiguously lowered by relative price changes. If, further, the number of factors equals the number of produced commodities, as in the original 2×2 setting, and production is non-joint, relative commodity price changes can be constructed which, without the aid of any direct subsidies, will raise the real reward of any particular factor regardless of its taste pattern.

The Rybczynski theorem

Unbalanced growth in factor supplies tends, at given commodity prices, to lead to stronger asymmetric changes in outputs. If the numbers of factors and commodities are evenly matched and production is non-joint, this asymmetry entails that growth in some, but not all, factors (when commodity prices are given) serves to force an actual reduction in one or more outputs. By similar reasoning, differences in the

composition of endowments among countries with similar technologies results in stronger asymmetries in production patterns when all face free trade commodity prices. If tastes are somewhat similar, these endowments differences are apt to support the trading patterns described by the Heckscher–Ohlin theorem.

See Also

- factor price equalization (historical trends)
- general equilibrium
- international trade theory
- Leontief paradox

Bibliography

Chang, W., Ethier, W. and Kemp, M. 1980. The theorems of international trade with joint production. *Journal of International Economics* 10, 377–94.

Chipman, J. 1966. A survey of the theory of international trade, Part 3. *Econometrica* 34 , January, 18–76.

Chipman, J. 1969. Factor price equalization and the Stolper–Samuelson Theorem. *International Economic Review* 10, 399–406.

Deardorff, A. 1980. The general validity of the law of comparative advantage. *Journal of Political Economy* 88, 941–57.

Deardorff, A. 1982. The general validity of the Heckscher–Ohlin theorem. *American Economic Review* 72, 683–94.

Deardorff, A. and Stern, R., eds. 1994. *The Stolper–Samuelson Theorem: A Golden Jubilee*. Ann Arbor: University of Michigan Press.

Dixit, A. and Norman, V. 1980. *Theory of International Trade*. Cambridge: Cambridge University Press.

Ethier, W. 1974. Some of the theorems of international trade with many goods and factors. *Journal of International Economics* 4, 199–206.

Ethier, W. 1984. Higher dimensional issues in trade theory. In *Handbook of International Economics*, vol. 1, ed. R. Jones and P. Kenen. Amsterdam: North-Holland.

Findlay, R. 1993. Wage dispersion, international trade and the services sector. In *Trade, Growth and*

Development: The Role of Politics and Institutions, ed. G. Hansson. London: Routledge.

Findlay, R. 1995. *Factor Proportions and Growth*. Cambridge, MA: MIT Press.

Flam, H. and Flanders, M.J., eds. 1991. *Heckscher–Ohlin Trade Theory*. Cambridge, MA: MIT Press.

Gale, D. and Nikaido, H. 1965. The Jacobian matrix and the global univalence of mappings. *Mathematische Annalen* 159, 81–93.

Gruen, F. and Corden, W. 1970. A tariff that worsens the terms of trade. In *Studies in International Economics*, ed. I. McDougall and R. Snape. Amsterdam: North-Holland.

Heckscher, E. 1919. The effect of foreign trade on the distribution of income. *Ekonomisk Tidskriff*, 497–512. Translated as chapter 13 in American Economic Association, *Readings in the Theory of International Trade*. Philadelphia: Blakiston, 1949, and a new translation is provided in Flam and Flanders (1991).

Jones, R. 1956. Factor proportions and the Heckscher–Ohlin theorem. *Review of Economic Studies* 24, 1–10.

Jones, R. 1965. The structure of simple general equilibrium models. *Journal of Political Economy* 73, 557–72.

Jones, R. 1971. A three-factor model in theory, trade, and history. In *Trade, Balance of Payments, and Growth*, ed. J. N. Bhagwati et al. Amsterdam: North-Holland.

Jones, R. 1974. The small country in a many-commodity world. *Australian Economic Papers* 13, 225–36.

Jones, R. 1985. Relative prices and real factor rewards: a reinterpretation. *Economic Letters* 19, 47–9.

Jones, R. 1992. Jointness in production and factor-price equalization. *Review of International Economics* 1, 10–18.

Jones, R. 2002. Heckscher–Ohlin trade models for the new century. In *Bertil Ohlin: A Centennial Celebration (1899–1999)*, ed. R. Findlay, L. Jonung and M. Lindahl. Cambridge, MA: MIT Press.

Jones, R. 2003. Joint output and real wage rates. *International Review of Economics and Finance* 12, 513–16.

Jones, R. 2006a. Eli Heckscher and the holy trinity. In *Eli Heckscher, International Trade, and*

Economic History, ed. R. Findlay et al. Cambridge, MA: MIT Press.

Jones, R. 2006b. ‘Protection and real wages’: the history of an idea. *Japanese Economic Review* 57, 457–66.

Jones, R. 2007a. Key international trade theorems and large shocks. *International Review of Economics and Finance*.

Jones, R. 2007b. Specific factors and Heckscher–Ohlin: an intertemporal blend. *Singapore Economic Review* 52, 1–6.

Jones, R., Beladi, H. and Marjit, S. 1999. The three faces of factor intensities. *Journal of International Economics* 48, 413–20.

Jones, R. and Marjit, S. 1985. A simple production model with Stolper–Samuelson properties. *International Economic Review* 19, 565–7.

Jones, R. and Marjit, S. 1991. The Stolper–Samuelson theorem, the Leamer triangle, and the produced mobile factor structure. In *Trade, Policy, and International Adjustments*, ed. A. Takayama, M. Ohyama and H. Otah. New York: Academic Press.

Jones, R. and Marjit, S. 1992. International trade and endogenous production structures. In *Economic Theory and International Trade: Essays in Memoriam J. Trout Rader*, ed. W. Neuefeind and R. Riezman. Berlin and New York: Springer-Verlag.

Jones, R., Marjit, S. and Mitra, T. 1993. The Stolper–Samuelson theorem: links to dominant diagonals. In *General Equilibrium, Growth and Trade II: Essays in Honor of Lionel W. McKenzie*, ed. R. Becker, R. Jones and W. Thomson. New York: Academic Press.

Jones, R. and Scheinkman, J. 1977. The relevance of the two-sector production model in trade theory. *Journal of Political Economy* 85, 909–35.

Kemp, M. and Wegge, L. 1969. On the relation between commodity prices and factor rewards. *International Economic Review* 10, 407–13.

Krueger, A. 1977. *Growth, Distortion, and Patterns of Trade Among Many Countries*. Princeton Studies in International Finance, No. 40. Princeton: Princeton University Press.

Leontief, W. 1953. Domestic production and foreign trade: the American capital position re-examined. *Proceedings of the American Philosophical Society* 97, 332–49.

Lerner, A. 1933. Factor prices and international trade. Mimeo. Published in *Economica* 19 (1952), 1–15.

Mas-Colell, A. 1979. Two propositions on the global univalence of systems of cost functions. In *General Equilibrium, Growth, and Trade*, ed. J. Green and J. Scheinkman. New York: Academic Press.

McKenzie, L. 1955. Equality of factor prices in world trade. *Econometrica* 23, 239–57.

Metzler, L. 1949. Tariffs, the terms of trade, and the distribution of national income. *Journal of Political Economy* 57, 1–29.

Minabe, N. 1966. The Heckscher–Ohlin theorem, the Leontief paradox, and patterns of economic growth. *American Economic Review* 56, 1193–211.

Mitra, T. and Jones, R. 1999. Factor shares and the Chipman condition. In *Trade, Welfare and Econometrics: Essays in Honor of John S. Chipman*, ed. J. Melvin, J. Moore and R. Riezman. New York: Routledge.

Ohlin, B. 1933. *Interregional and International Trade*. Cambridge, MA: Harvard University Press, 1966.

Ruffin, R. 1988. The missing link: the Ricardian approach to the factor endowment theory of trade. *American Economic Review* 78, 759–72.

Rybczynski, T.M. 1955. Factor endowments and relative commodity prices. *Economica* 22, 336–41.

Samuelson, P. 1948. International trade and the equalization of factor prices. *Economic Journal* 58, 163–84.

Samuelson, P. 1949. International factor-price equalization once again. *Economic Journal* 59, 181–97.

Samuelson, P. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21, 1–20.

Samuelson, P. 1971. Ohlin was right. *Swedish Journal of Economics* 73, 365–84.

Samuelson, P. 1992. Factor-price equalization by trade in joint and non-joint production. *Review of International Economics* 1, 1–9.

Stolper, W. and Samuelson, P. 1941. Protection and real wages. *Review of Economic Studies* 9, 58–73; also in Deardorff and Stern (1994).

Uekawa, Y. 1971. Generalization of the Stolper–Samuelson theorem. *Econometrica* 39, 197–213.

Uekawa, Y. 1984. Some theorems of trade with joint production. *Journal of International Economics* 16, 319–33.

How to cite this article

Jones, Ronald W. "Heckscher–Ohlin trade theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000034> doi:10.1057/9780230226203.0718

The New Palgrave Dictionary of Economics Online

hedging

Gregory Connor

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Hedging is defined with a state-space model of risky outcomes. Full and partial hedging are compared, and the feasible set of hedging positions related to the available collection of traded assets. Three types of counter-parties for hedging trades are distinguished. The risk premium for a hedging asset is defined, and its relationship to economy-wide risk factors explained. The case of mean-variance preferences provides a useful formula for the optimal hedge position. Corporations undertake many hedging transactions, even though the shareholders of the corporation do not typically benefit from any risk reduction. Some explanations of corporate hedging are set out.

Keywords

adverse selection; asset pricing; bankruptcy; bid ask spread; Brownian motion; corporate hedging; differential information; futures markets; hedge portfolio; hedging; Keynes, J. M.; mean-variance preference model; moral hazard; normal backwardation; portfolio insurance; progressive taxation; risk aversion; risk premium; speculation; state space models; taxation of corporate profits

Article

Hedging is the purchasing of an asset or portfolio of assets in order to insure against wealth fluctuations from other sources. A hedge portfolio is any asset or collection of assets purchased by one or more agents for hedging. A grain dealer may hedge against losses on an inventory of grain by selling grain futures; a Middle Eastern businessman may hedge against political turmoil (and the resulting losses) by buying gold; a pension fund may hedge against capital losses on its equity portfolio by buying stock index put options.

1 A competitive equilibrium model of hedging

The fundamental concepts of hedging can best be described in the state space model. Consider a one-period economy with *M* agents and one end-of-period consumption good. For simplicity, assume that there is no consumption at the beginning of the period. Each agent possesses a real asset which produces a random amount of the consumption good at the end of the period. Agents have homogeneous beliefs. There are *N* possible states of nature, with probabilities $Pr(1), \ldots, Pr(N)$. The agents have concave, possibly state-dependent utility functions and wish to maximize the expected utility of end-of-period consumption. Let $U_j(C_j, \theta_i)$ denote the end-of-period utility of agent *j* given that his consumption is C_j and the state of nature is θ_i.

A *financial asset* is a claim to a random amount of end-of-period output which is traded between agents at the beginning of the period. A *hedge portfolio* is a particular type of financial asset or collection of financial assets which protects an agent against some particular risky outcome(s).

The analysis is simplest if we assume that the hedge portfolio consists of a mixed asset-liability with positive payoffs in some states and negative payoffs in other states, balanced so as to give a competitive equilibrium price of zero. Under this formulation a hedge portfolio is a portfolio which pays off positively in states where the agent would otherwise have a high marginal utility of consumption (that is, ‘bad’ states) and negatively in states where he would otherwise have a low marginal utility of consumption. If the agent's marginal utility is equalized across the relevant states after purchasing the hedge portfolio, then he is fully hedged; if the hedge position lowers but does not eliminate the disparity, then he is partially hedged.

Who takes the other side of the hedging transaction? There are three possibilities. First, if there exist two agents who have real asset cash flows which vary inversely, then they can trade in a way which allows both to hedge simultaneously. For example, the grain dealer who holds an inventory of grain may be able to sell a futures contract to a bread producer who has committed himself to using grain at a later stage of his production process. Both parties consider themselves as hedging. Second, one agent may be less risk-averse towards certain states of nature than another. The less risk-averse may be willing to sell the hedge asset to the more risk-averse at a price which produces mutual gains in expected utility. Third, the hedging agent may be able to trade small quantities of the hedge asset with many agents, who can then eliminate all or most of the risk of the trade by combining the asset with many others (that is, by diversifying away the risk). For example, insurance companies can sell fire insurance policies to many individuals and leave very little risk to be absorbed by the company's shareholders.

Let the number of distinct types of assets be *K* and let *Y* denote the *NxK* matrix of their payoffs in the *N* possible states of nature. The set of available trades is *span*(*Y*•) where *span*(·) denotes the subspace spanned by the matrix. In an economy without frictions, agents will create new financial assets until all mutually beneficial trade opportunities are in *span*(*Y*•). All mutually beneficial trades have been consummated if there exist positive scalars $\lambda_1 \ldots \lambda_M$ such that

$$\lambda_j U_j'(C_{ij}, \theta_i) = \lambda_h U_h'(C_{ih}, \theta_i) \quad i = 1, \ldots, N; \; j, h = 1, \ldots, M \tag{1}$$

where U' denotes the first derivative with respect to consumption. The invisible hand drives agents towards creating all the types of financial assets which can lead to mutually beneficial trades. However, there are many external factors which can offset this tendency. If agents have some control over outcomes, then moral hazard problems may limit hedging opportunities. For example, agents may not be able to hedge against changes in labour income if work requires imperfectly observable effort. If agents have special knowledge, then adverse selection can similarly limit trade. If a car owner knows more about its quality than a prospective buyer, then the owner cannot sell his car at a reasonable price when he experiences financial distress. The administrative costs of trade can also limit hedging before the full efficiency condition (1) is fulfilled.

The model described above is static. In an intertemporal model, dynamic strategies increase the set of hedging opportunities beyond the linear span of the matrix of asset payoffs. Agents can create a rich set of payoff claims by dynamically varying the proportions invested in the individual assets. With continuous trading, this process reaches its natural limit: if an asset price follows Brownian motion, then the continuously adjusted portfolio consisting of only the risky asset and riskless asset can be constructed which replicates the payoff to any put or call option on the risky asset.

The proliferation of complex financial assets, such as options on futures and interest rate and currency options, and the increased sophistication of traders has led to a bewildering array of dynamic hedging strategies, especially by large institutional investors. *Portfolio insurance* provides a good example of the kind of sophisticated new hedging instrument which can be created with a dynamic trading strategy. Consider a pension fund with a large equity portfolio and an aversion to large capital losses on this portfolio. The portfolio insurance strategy can put a floor on the random rate of return to the pension fund's portfolio. The return floor can be any rate lower than the available riskless rate (it can be a negative net return, so that the fund bounds its losses rather than assuring itself a small gain). The strategy works as follows. At the starting date of the insurance strategy, the fund has most of its money invested in equities and a small proportion in a riskless asset (that is, government notes). If the equities fall in price, the fund sells some of the equities and places the cash in the riskless asset. If equity prices continue to fall, the fund increases the proportion of investment in the riskless asset. If there is a sustained fall in equity prices, the fund will end the insurance programme invested entirely in the riskless asset. It will have earned a rate equal to the pre-chosen minimally acceptable return. The fund makes a ‘soft landing’ at this minimal value: the proportion of money invested in the riskless asset approaches 1 as the value of the portfolio approaches the minimally acceptable level.

Portfolio insurance is not a free lunch. In exchange for the return floor, the pension fund sacrifices some of the upside potential of pure equity investment. For example, if the equity market declines sharply and then rises, the fund will miss the upturn, since it will have defensively decreased its position in equities before the upturn.

There are numerous other dynamic hedging strategies, not only in equity markets but in fixed income, currency and options markets. In terms of the volume of trade, hedging in financial markets now greatly outpaces activity in commodities futures markets, the original and classic example of markets often used for hedging.

2 Risk premia and hedging

An economically interesting question is whether agents ‘pay a premium’ to hedge. Assume again that

the current price of the hedge portfolio is set to zero by appropriate balancing of the asset and liability sides of the hedge (a futures contract is a natural example). If the expected cash flow is negative (positive) next period, then the hedge portfolio carries a positive (negative) implicit asset pricing risk premium. If the expected cash flow is zero, then the implicit asset pricing risk premium is zero. If we used returns rather than prices, then the expected return premium would have the opposite sign from the asset pricing premium.

Much of the early literature on hedging was centred on hedging in commodities futures contracts. One of the key questions was whether agents who sold futures pay a positive risk premium. Keynes (1930) considers this problem for the case of commodities futures contracts. He argues that the natural supply of short sellers (sellers of futures contracts) outnumbers long hedgers (buyers) in this market. Therefore, the implicit risk premium for holding a futures contract should be negative, in order to induce other agents (henceforth called speculators) to absorb the excess hedging demand of short hedgers. This will be true if the futures price increases on average over the life of the contract, so that the expected cash flow from holding the contract is positive.

The empirical evidence for this positive trend (sometimes called *normal backwardation*) in commodities futures market prices is weak at best. Keynes's analysis implicitly assumes that the commodity futures market is isolated from other assets so that hedgers must pay other agents (the speculators) a premium to induce them to take a position in the market. In an integrated set of asset markets, hedgers need not pay any premium to induce other agents to trade. Rather, the existence of a risk premium depends upon the covariance between the payoffs to the hedge asset and the economy-wide risks faced by all the agents. If the hedge asset is uncorrelated with market-wide risks, then it will carry no risk premium, even though it may have a high value to a particular hedger due to his specific income stream. The hedge asset which protects against market risk will carry a risk premium.

There is another source of return to speculators, which is not captured in the competitive pricing model. Speculators may charge an explicit or implicit bid ask spread when trading with hedgers. If hedgers buy and hold for a long period, then this is equivalent to the asset pricing risk premium described above. However, if hedgers trade frequently, then the bid ask spread can lower their realized returns, and raise the realized returns of speculators, without affecting the observed long-run return premium of the hedge asset as reflected in transactions prices. This may explain the lack of empirical evidence for normal backwardation in commodities futures market. This effect of the bid ask spread was not recognized in most of the early literature on commodity futures markets.

The bid ask spread need not be explicit. Even open-floor markets will contain a set of implicit bid ask spreads, to the extent that traders’ strategies reflect a greater willingness to sell at higher prices and to buy at lower ones. One can view the feverish activity which is common in floor trading as speculators searching for transactions at the outer edges of an implicit bid ask spread. Hedgers, who are off the floor and are more anxious to complete a particular trade, take the losing side of the implicit spread.

3 The role of hedgers in a market with heterogeneous information

The model in sections 1 and 2 assumes homogeneous information across agents. If agents have differential information about the payoffs to assets, then the trading strategies of rational agents cannot take this simple competitive form. Agents must treat trade opportunities as signals of the information of other agents about the value of the trade.

The presence of differential information can lead to fewer hedging opportunities and/or raise the expected cost of hedging. Milgrom and Stokey (1982) show that rational agents will not trade solely because they have different beliefs about the value of an asset. If agents are distinguished merely by their differential information, then they will refuse all trades, since the willingness of the other agents to trade signals that the terms are unfavourable. This means that a financial asset market will fail to open in the absence of other motives for trade. This is a market failure due to adverse selection. The need of some agents for hedging can provide an additional reason for trade which overcomes the adverse selection problem and eliminates the market failure. Hedgers will be willing to trade even if they suspect that the other party to the transaction has superior information. Informed agents will be gaining at the expense of hedgers, but they will also be providing an insurance or liquidity service to hedgers, and so hedgers may be willing to trade with them despite their informational disadvantage. This in turn has the side effect of permitting superior information to be reflected in market prices.

I follow Glosten and Milgrom (1985) and assume that there exists a costless, competitive, risk-neutral market maker who intervenes in all trades. This is for analytical convenience and is not necessary to the basic model. Suppose that certain agents (hedgers) have a strong preference for a given asset, that is, their preference is such that they will buy (and sell) some non-zero amount at a price higher (lower) than the market-clearing equilibrium price. This implies that they are willing to trade even if they must pay a bid ask spread around the equilibrium price. Informed agents, henceforth speculators, will also trade despite a bid ask spread as long as the expected profit from their superior information is larger than the bid ask spread. The market maker's resulting equilibrium bid ask spread will allow the hedgers to trade at an expected loss and the speculators at an expected gain, leaving the market maker with an expected profit of zero (the equilibrium condition). The market maker will respond to the net demands of all traders (which partially reveals the net demand of speculators) to adjust the bid and ask prices, and so (partially) capture in the market price the superior information of speculators.

One interesting feature of this model is the symbiotic roles of speculators and hedgers. Without speculators (informed traders), the hedgers would lose liquidity; without hedgers, speculators would lose the opportunity to profit from their superior information. Without both speculators and hedgers, the price in the market would no longer provide a useful signal for agents making production and consumption decisions. Kyle (1984) develops a model in which the symbiotic relationship is made clear and describes the effects of more or fewer speculators or hedgers on the informational efficiency and liquidity of the market. Some of the results are counter-intuitive: for instance, increasing the number of hedgers, who are uninformed, can increase the informational efficiency of prices.

4 Hedging in a mean variance model

The mean variance preference model provides a useful framework for empirical and applied analysis of hedging. Suppose that an investor's utility is given by the expected value of his random end-of-period endowment, *X*, minus some multiple of the variance of this endowment:

$$U(X) = E[X] - a\,var[X]$$

Let the hedging instrument have a current price of zero and random end-of-period payoff *Y*. It is easy to show that the investor's optimal position in the hedging instrument is

$$w = \frac{E[Y]}{(2a\,var[Y])} - \frac{cov[X, Y]}{var[Y]}.$$

(See Rolfo, 1980, for the derivation). The first additive part of this optimal hedging position, $E[Y] / (2a\,var[Y])$, is called the *speculative hedge*. The second part, $-cov[X, Y] / var[Y]$, is called the *pure hedge*. Some analysts (such as Duffie, 1989) argue that in practice uninformed hedgers should ignore the speculative hedge and set the hedge position equal to the pure hedge. The justification is that the speculative hedge requires predictions about the expected payoff on the hedging instrument, whereas the pure hedge uses only the covariance of the hedging instrument and the random endowment. In many cases, covariances are more stable over time and more precisely estimated than expected payoffs. Setting the hedge position equal to the pure hedge is equivalent to minimizing variance instead of optimizing over a mean variance criterion.

The mean variance preference model provides a useful link to empirical analysis of hedging. Suppose that we observe a sample of realized random endowments and hedging instrument payoffs. Consider an ordinary least squares regression of the random endowment on the payoffs to the hedging instrument:

$$X = \alpha + \beta Y + \varepsilon.$$

The coefficient β estimates (minus) the pure hedge. The R-squared from this regression estimates the proportion of endowment variance which is eliminated by setting the hedge position equal to the pure hedge.

5 Risk premia on hedge portfolios and general equilibrium pricing

In section 2, I described two types of hedge portfolios – those with and those without risk premia – and how the distinction between them depends on the covariance between the hedge portfolio returns and the market-wide risks in the economy. In this section, I describe the relationship between hedge portfolios which protect against market-wide risks and the general equilibrium pricing of assets.

Let Q_t denote the discounted expected utility of lifetime consumption for some agent at time *t*:

$$Q_t = E_t[\sum_{\tau=t+1}^{\infty} \rho^{\tau} U(C_\tau, \theta_\tau)]$$

where ρ is the agent's discount factor and $U(.,.)$ is his utility function. Let f_t denote the change in discounted expected utility given the change in the agent's time t wealth:

$$f_t = \frac{\partial Q_t}{\partial W_t}$$

where W_t is his wealth at time t. Note that, at time $t-1$, f_t is a random variable. Let r_{it} denote the return from time $t-1$ to t of the ith financial asset. If the agent holds an equilibrium amount of this asset then the following first-order condition is satisfied:

$$E_{t-1}[r_{it} f_t] = \frac{\partial U(C_{t-1}, \theta_{t-1})}{\partial C_{t-1}}$$

which can be rewritten (using $E[ab] = E[a]E[b] + cov[a, b]$) as:

$$E_{t-1}[r_{it}] = r_{0t} + \left(\frac{1}{\gamma}\right) cov_{t-1}[r_{it}, f_t]$$
(2)

where $\gamma = \frac{\partial U(C_{t-1}, \theta_{t-1})}{\partial C_{t-1}}$ and r_{0t} is the expected return on an asset with a riskless payoff at time t. Suppose that, at time $t-1$, f_t equals a sum of a set of K uncorrelated random variables $Z_{1t}, \ldots, Z_{Kt}$:

$$f_t = Z_{1t} + \ldots + Z_{Kt}$$
(3)

The variables $Z_{1t}, \ldots, Z_{Kt}$ describe the *K* random shocks which affect the agent's marginal utility. They could be interest rate movements, output shocks, inflation shocks, and so on. Assume that there exists a set of *K* portfolios with returns $r^*_{1t}, \ldots, r^*_{Kt}$ such that the jth portfolio has perfect negative correlation with Z_{jt}:

$$cov_{t-1}[r^*_{jt}, Z_{jt}] = (var_{t-1}[r^*_{jt}]var_{t-1}[Z_{jt}])^{\frac{1}{2}}. \tag{4}$$

These portfolios are potential hedges against the *K* types of risks which affect the investor. (The agent will short sell the portfolio to hedge since the portfolio return varies inversely with marginal utility.) I call $r^*_{1t}, \ldots, r^*_{Kt}$ an indexing set of hedge portfolios since the portfolios index the random shocks to the agent's marginal utility. Using (3) and (4) we can rewrite (2) as:

$$E_{t-1}[r_{it}] = r_{0t} + \beta_{i1t}\pi_{1t} + \ldots + \beta_{iKt}\pi_{Kt} \tag{5}$$

where

$$\beta_{ijt} = \frac{cov_{t-1}[r_{it}, r^*_{jt}]}{var_{t-1}[r^*_{jt}]}$$

and

$$\pi_{jt} = E_{t-1}[r^*_{jt} - r_{0t}].$$

Equation (5) is an asset pricing relationship: it says that the expected return on any asset equals the riskless return plus a linear combination of the covariances of the asset's return with an indexing set of hedge portfolios.

6 Explaining corporate hedging

The analysis in the previous sections considers hedging from the perspective of an individual investor and consumer. In practice, most financial hedging activity is undertaken by corporations rather than by individuals. If corporations issue widely held common stock, then hedging activity between them appears unnecessary. The shareholders of the corporation are its ultimate risk bearers, and they will not benefit from hedging activity at the corporate level. So, for example, an investor with shares in both an oil-producing industry and an oil-consuming industry will not want firms in the two industries to hedge with oil futures, taking opposite hedge positions (one long oil futures and the other short oil futures). From the shareholder's perspective, the cash flows from these offsetting hedges will be diversified away at the portfolio level. The ultimate shareholder pays the transactions costs of hedging by the corporations, but does not experience any aggregate risk reduction in his portfolio.

There are several explanations for the prevalence of corporate hedging. One explanation relies on the costs of financial distress. By hedging, the corporation lowers the probability of bankruptcy or near-bankruptcy, and so increases its average market value. Another explanation relates corporate hedging to the agency costs of hiring managers to run the firm. The firm's managers have a non-diversified exposure to the profitability of the firm; hedging by the corporation more closely aligns the interests of the shareholders and managers, and also allows the shareholders to pay the managers less on average since the managers’ risk exposure is reduced. A third explanation relies on the progressiveness of the corporate tax system, which encourages corporations to hedge so as to smooth taxable earnings and thereby lower their average tax bill.

See Also

- capital asset pricing model
- futures markets, hedging and speculation
- mean-variance analysis
- risk aversion
- state space models

Bibliography

Arrow, K. 1964. The role of securities in the optimal allocation of risk bearing. *Review of Economic Studies* 31, 91–6.

Breeden, D. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7, 265–96.

Breeden, D. 1984. Futures markets and commodity options: hedging and optimality in incomplete markets. *Journal of Economic Theory* 32, 275–300.

Brennan, M. 1958. The supply of storage. *American Economic Review* 48, 50–72.

Cootner, P. 1967. Speculation and hedging. *Food Research Institute Studies, Supplement* 7, 65–105.

Duffie, D. 1989. *Futures Markets*. Englewood Cliffs, NJ: Prentice-Hall.

Duffie, D. and Huang, C. 1985. Implementing Arrow–Debreu equilibrium by continuous trading of a few long-lived securities. *Econometrica* 53, 1337–56.

Glosten, L. and Milgrom, P. 1985. Bid, ask, and transaction prices in the specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14, 1–00.

Grauer, F. and Litzenberger, R. 1979. The pricing of commodity futures contracts, nominal bonds and other risky assets under commodity price uncertainty. *Journal of Finance* 34, 69–84.

Gray, R. 1961. The search for a risk premium. *Journal of Political Economy* 69, 250–60.

Grossman, S. 1976. On the efficiency of competitive stock markets when traders have diverse information. *Journal of Finance* 31, 573–85.

Hoffman, G. 1954. Past and present theory regarding futures trading. *Journal of Farm Economics* 19, 1–11.

Houthakker, H. 1957. Can speculators forecast prices? *Review of Economics and Statistics* 39, 143–52.

Keynes, J.M. 1930. A Treatise on Money. Volume 2: The Applied Theory of Money. Reprinted in The Collected Writings of John Maynard Keynes, vol. 7. London: Macmillan, 1971.

Kyle, A. 1984. Market structure, information, futures markets, and price formation. In *International Agricultural Trade: Advanced Readings on Price Formation, Market structure, and Price Instability*, ed. G. Storey, A. Schmitz and A. Sarris. London: Westview.

Leland, H. 1980. Who should buy portfolio insurance? *Journal of Finance* 35, 581–94.

Malinvaud, E. 1972. The allocation of individual risk in large markets. *Journal of Economic Theory* 4, 312–28.

Merton, R. 1973. An intertemporal capital asset pricing model. *Econometrica* 41, 867–87.

Milgrom, P. and Stokey, N. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26, 17–27.

Rolfo, J. 1980. Optimal hedging under price and quantity uncertainty: the case of a cocoa producer. *Journal of Political Economy* 88, 100–16.

Smith, C. and Stulz, R. 1985. The determinants of firms' hedging policies. *Journal of Financial and Quantitative Analysis* 20, 391–405.

Working, H. 1953a. Futures trading and hedging. *American Economic Review* 43, 314–43.

Working, H. 1953b. Hedging reconsidered. *Journal of Farm Economics* 35, 544–61.

Working, H. 1967. Tests of the theory concerning for trading on commodity exchanges. *Food Research Institute Studies, Supplement* 7, 5–48.

How to cite this article

Connor, Gregory. "hedging." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000035> doi:10.1057/9780230226203.0719

The New Palgrave Dictionary of Economics Online

hedonic prices

Lars Nesheim
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Hedonic price functions describe the equilibrium relationships between characteristics of products and their prices. They are used to predict prices of new goods, to adjust for quality change in price indexes, and to measure consumer and producer valuations of differentiated products. They emerge as market outcomes from both competitive and non-competitive markets. The functional form is determined by the distribution of buyers and their preferences, the distribution of sellers and their costs, and the structure of competition in the market.

Keywords

arbitrage; assortative matching; bundling; cost-of-living indexes; hedonic preferences; hedonic price functions; household production; imperfect competition; Laspeyres index; misspecification; nonparametric methods; Paasch index; pure strategy Nash equilibrium; willingness to pay

Article

A hedonic price function describes the equilibrium relationship between the economically relevant characteristics of a product or service (or bundle of products) and its price. For example, in a simple labour economics model the hedonic wage function might describe how the wages of a worker depend on education, experience and skill. In a simple housing economics model, the hedonic house price might describe how the price of a house depends on geographic location, size, and quality. In each case, the hedonic price function describes equilibrium (not necessarily competitive) valuations of the economically relevant characteristics of the product.
In empirical applications, statistical estimates of hedonic price functions have primarily been used to calculate quality adjusted price indexes for goods and to measure consumer valuations or producer costs of product characteristics. They have been used to study markets for agricultural products, automobiles, labour, houses, computers, and myriad other differentiated commodities. They have been used to measure quality change in private goods markets and to measure consumer valuations of changes in public goods such as clean air, schools or transport infrastructure. In all these applications, hedonic methods are crucial because the goods in question are not homogenous and their value to buyers and sellers varies systematically with characteristics.
Key questions to be answered when developing a hedonic model to analyse a product market are what are the economically relevant characteristics of the product and what is the market environment that generates the hedonic equilibrium price. Given answers to these questions, a key theoretical goal of hedonic analysis is to determine the theoretical relationship between these market equilibrium prices and underlying structural features of the economy such as producer costs and consumer preferences. Two key empirical goals of hedonic analysis are to understand when statistical estimates of hedonic relationships provide good out-of-sample predictions of prices and to understand what structural information these statistical relationships provide about costs and preferences.

1 General hedonic demand

Hedonic models make various assumptions about whether the space of feasible characteristics is discrete or is a continuum, and whether the characteristics embodied in different products can be bundled or unbundled. This section discusses a general model of hedonic demand that encompasses these special cases. The supply side of the market and various

notions of equilibrium are discussed in section 2.

Each consumer who participates in the hedonic market derives utility from a vector of characteristics $z \in Z_m \subseteq \mathbf{R}^{n_z}$. The bundle z is obtained either by buying a single product that embodies z or by buying a set of products that together produce z. In either case the hedonic cost or price is $p(z)$. The set Z_m is the feasible set given current market conditions. The set Z_m could be a finite set or it could be a continuum. Each consumer also has the option not to participate in the hedonic market, in which case they obtain reservation utility u_0.

Assume that characteristics are defined so that utility is increasing in each element of z. Also, assume that utility is decreasing in $p(z)$.

Every consumer is represented by a type $x \in X \subseteq \mathbf{R}^{n_x}$. The space X is the space of all consumer types. The vector x is a vector of consumer characteristics (such as income, education or preference parameters) that affects utility. Consumer heterogeneity is an important feature of hedonic models.

Given hedonic price $p(z)$, consumer x chooses $z \in Z_m$ to maximize utility $u(x, z, p(z)$.) That is, they solve

$$\max_{\{z \in Z_m\}} \left\{ u(x, z, p(z)) \right\}. \tag{1}$$

The solution $z = d(x)$ is the hedonic demand function (or correspondence) for consumer x.

Several features of the model are important. First, z is a complete list of the product characteristics that both affect consumer utility and are known to the consumer at time of purchase. In the housing market example, z could measure geographic location, age of the dwelling, lot size, number of rooms, size of the yard, and so on. Second, there may be additional characteristics of the good that affect *ex post* utility but that are not known to the consumer at time of purchase. In such cases, the utility function should be interpreted as the expected utility from purchasing a good with known characteristics z. Third, buyer utility depends on x and on z. Two consumers, x_1 and x_2, with $x_1 \neq x_2$, will generally choose different bundles $(z_1, p(z_1))$ and $(z_2, p(z_2))$ and will obtain different levels of utility.

1.1 Continuous choice version

To specialize to the case where Z_m is a compact convex subset of $\mathbf{R}^n$, both u and p are differentiable and the consumer maximization problem has an interior solution, the first-order condition describing the consumer's hedonic demand is

$$\frac{\partial u(x, z, p(z))}{\partial z} + \frac{\partial u(x, z, p(z))}{\partial p} \frac{\partial p(z)}{\partial z} = 0 \tag{2}$$

which can be rewritten as

$$\frac{\partial p(z)}{\partial z} = -\left(\frac{\partial u(x, z, p(z))}{\partial z} \Big/ \frac{\partial u(x, z, p(z))}{\partial p} \right). \tag{3}$$

The marginal price at z equals the marginal rate of substitution of the consumer x who chooses z. In the quasi-linear utility case $u(x, z, p(z)) = u(x, z) - p(z)$ and eq. (3) becomes

$$\frac{\partial p(z)}{\partial z} = \frac{\partial u(x, z)}{\partial z}. \tag{4}$$

These results are the basis for the intuition that the slope of the hedonic price function measures consumers' marginal willingness to pay. Figure 1 illustrates. Consumers x_1 and x_2 optimally choose bundles z_1 and z_2 respectively. At z_1, the marginal price equals the marginal willingness to pay of consumer x_1. However, it is less than the marginal willingness to pay of consumer x_2. At z_2, the marginal price equals the marginal willingness to pay of x_2 but is greater than the marginal willingness to pay of x_1.

Figure 1

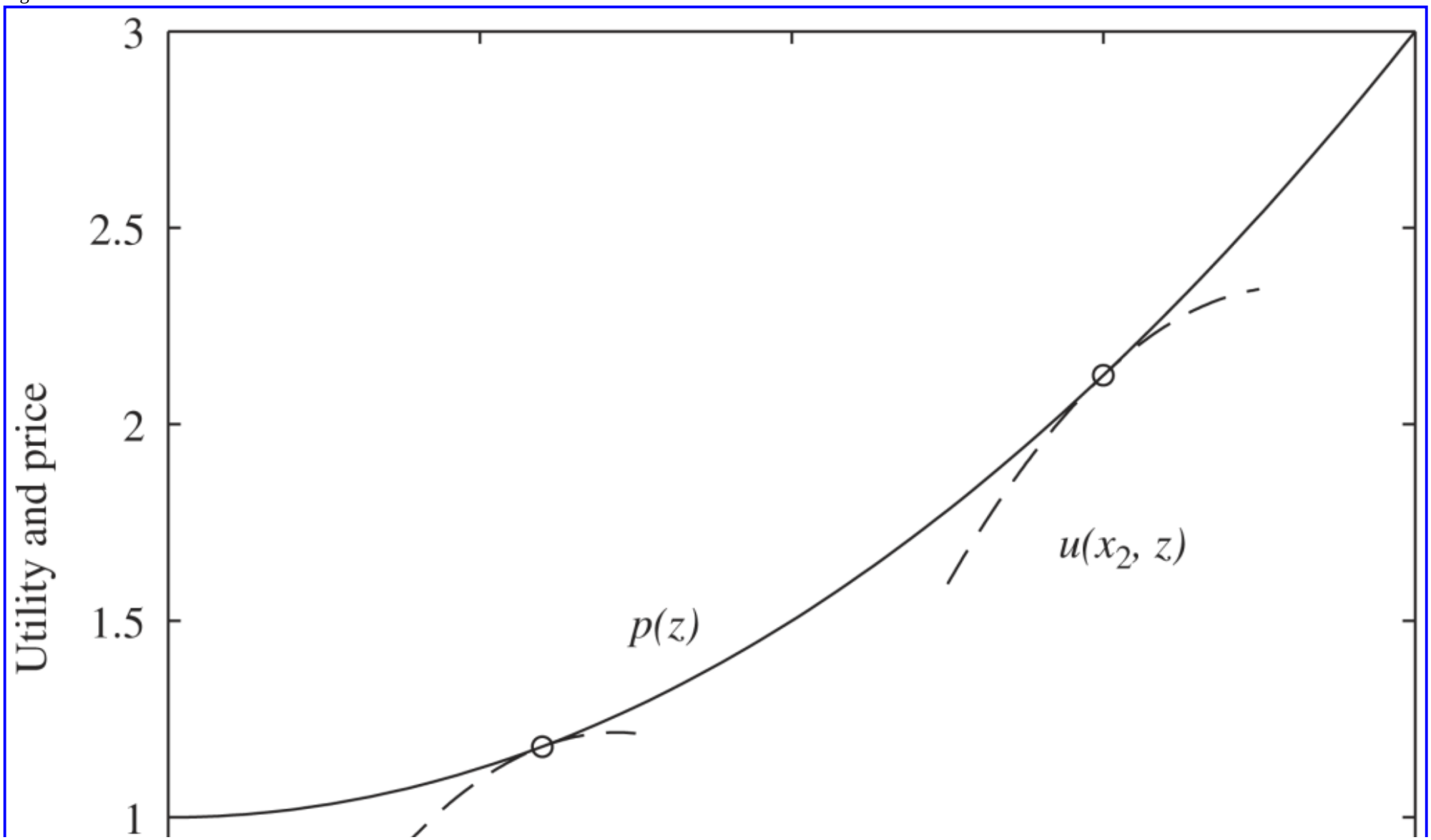

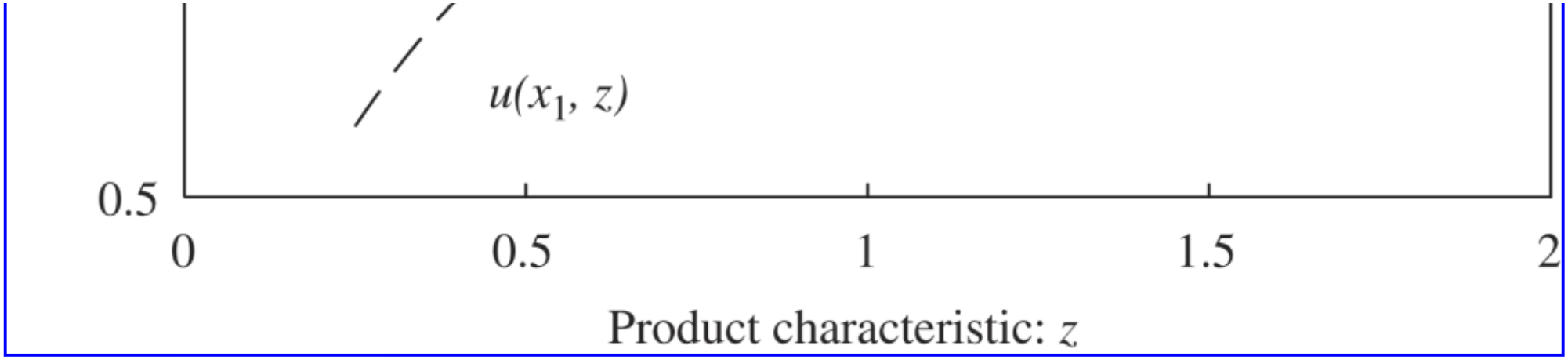

The hedonic price function reveals precise information about consumers x_1 and x_2 at points z_1 and z_2 respectively. At all other person-location pairs, it reveals only bounds on willingness to pay. It also reveals very little about how consumers x_1 and x_2 will react to large changes in the shape of the price function. More precise information requires the estimation of consumer preferences.

1.2 Discrete choice version

If the marginal conditions in (3) and (4) are replaced by inequalities, the qualitative interpretations above apply equally to economies in which Z_m is finite. Suppose there are J elements in Z_m. Let z_j be the j'th element in Z_m and let $p_j = p(z_j)$ for $j = 1, \ldots, J$. In the quasi-linear case, if consumer x chooses z_j, then

$$u(x, z_j) - p_j \geq u(x, z_k) - p_k$$

for all $k \in \{1, \ldots, J\}$.
Consider the set of consumers who choose z_j and for whom

$$u(x, z_j) - p_j = u(x, z_k) - p_k \tag{5}$$

for some $k \neq j$. These consumers are indifferent between bundle z_j at price p_j and bundle z_k at price p_k. The difference in prices between z_j and z_k exactly compensates for the difference in utilities. For these indifferent consumers, willingness to pay for z_j over z_k is

$$p_j - p_k = u(x, z_j) - u(x, z_k).$$

This is the discrete analog of the marginal willingness to pay.
Equation (5) only holds for those who are indifferent between *j* and *k*. For those who are not indifferent, the willingness to pay for z_j over z_k is strictly larger than the price. That is

$$u(x, z_j) - u(x, z_k) > p_j - p_k.$$

When the set of available alternatives Z_m is finite, the hedonic price function provides a precise measure of willingness to pay for consumers who are indifferent between options and provides bounds on willingness to pay for consumers who strictly prefer one option to others.

1.3 Single product demand version

In single product demand models, the vector *z* measures the characteristics of the unique product type that is chosen. These models assume that households cannot buy two separate products with characteristics z_1 and z_2 and combine their characteristics to obtain some other bundle z_3 (Rosen, 1974). These models do allow consumers to choose both a product type *z* and a quantity. To see this, rewrite the utility function in (1) as

$$u(x, z, p(z)) = \max_{\{q\}} \left\{ \hat{u}(x - qp(z), z, q) \right\}$$

where *q* is the quantity of product type *z* and *x* is income. This is the primary model used to study location choices and demand for land in urban economic models. See Fujita (1991).

1.4 Home production version

Home production models assume that consumers purchase a vector of goods in quantities $q \in \mathbf{R}^n_+$ at market prices $\pi \in \mathbf{R}^n_+$ and produce the bundle *z* from the goods purchased. See Gorman (1980), Lancaster (1966), and Muellbauer (1974). In home production hedonic models, consumers have a technology $f : Z \times \mathbf{R}^n \to \mathbf{R}^m$ describing the production possibility frontier. Given purchases of *q* units of market goods, any bundle *z* that satisfies the restriction $f(z, q) = 0$ is feasible.
Given market prices π and technology *f*, the cost of obtaining the bundle *z* is

$$p(z) = \min_{\{q\}} \left\{ \pi \cdot q \ \text{ subject to } \ f(z, q) = 0 \right\} \tag{6}$$

Thus, the hedonic price *p*(*z*) is the minimum cost of obtaining bundle *z* given market prices π and technology *f*. Given *p*(*z*), consumers maximize the utility given in (1). The single-product demand model is a special case of the home production model.
In the Gorman–Lancaster version of the model, the technology is linear and $f(z, q) = z - Aq$ where *A* is a $n_z \times n_q$ matrix. Each market good contains a fixed quantity of characteristics. The total amount available for consumption is the sum of characteristics across all goods purchased.

1.5 Hedonic cost of living index

In each of these models, one can calculate various hedonic cost of living indexes. See Pollak (1989) for details of many alternatives. This section discusses one alternative.
Consider a consumer who purchases a vector of quantities of homogenous goods q with linear prices π and a single differentiated product with characteristics z and hedonic price $p(z)$. When prices are (π, p), the cost of obtaining utility level u_0 is

$$c(\pi, p, u_0) = \min_{\{q, z\}} \left\{ \pi \cdot q \text{ subject to } u(q, z) \geq u_0 \right\} \tag{7}$$

If prices change from (π_0, p_0) to (π_1, p_1), then the constant utility hedonic cost of living index is

$$\frac{c(\pi_1, p_1, u_0)}{c(\pi_0, p_0, u_0)}.$$

This cost index hold utility constant and allows consumers to alter consumption of q and z in response to changing prices. When consumer preferences are unknown, this theoretical index cannot be calculated. With data on prices and quantities, empirical alternatives include the Laspeyres index and the Paasch index.
Let (q_0, z_0) solve (7) when prices are (π_0, p_0) in period zero. Let prices in period one be (π_1, p_1). Then a hedonic Laspeyres index is

$$L(q_1, p_1, q_0, p_0, x_0, z_0) = \frac{\pi_1 \cdot q_0 + p_1(z_0)}{\pi_0 \cdot q_0 + p_0(z_0)} \geq \frac{c(\pi_1, p_1, u_0)}{c(\pi_0, p_0, u_0)}.$$

This index holds the consumption bundle (q_0, z_0) constant at initial levels. Like the standard Laspeyres index, it is an overestimate of the cost of living index because it ignores a consumer's ability to alter consumption in response to changing prices. If some components of z are exogenous (for example, public goods like air quality or public safety), alternative indexes can be defined by including the time varying exogenous elements of z as arguments in the cost function.
One major problem with the index is that the set of available products often changes rapidly over time. If product z_0 is not traded in period 1, then $p_1(z_0)$ will not be observed. Pakes (2003) shows that an estimate of $p_1(z_0)$ based on observed prices is an upper bound under certain circumstances. A better option is to calculate the virtual price $p_1^V(z_0)$ that makes the household indifferent between purchasing z_0 at price $p_1^V(z_0)$ and purchasing z_1 (the product actually chosen in period 1) at price $p_1(z_1)$. The virtual price satisfies

$$p_1^V(z_0) = p_1(z_1) - (u(x, z_1) - u(x, z_0)).$$

Data on prices and quantities can be used to bound the virtual price. Precise results require estimation of consumer preferences.

Another major problem is that statistical authorities, as discussed in section 3, do not observe the elements of z that enter consumer preferences. A third major problem is that time constraints and cost constraints place severe limitations on data collection and analysis for use in practical price index calculations. Triplett (2004) provides a comprehensive overview of these issues.

2 Market equilibrium

Hedonic prices emerge as equilibrium outcomes from a market environment. They might emerge from a purely competitive environment in which neither buyers nor sellers have power to influence prices or they might emerge from an imperfectly competitive environment in which either buyers or sellers have market power. They may be observed in arms-length transactions or unobserved as in black-market wage contracts or implicit marriage contracts.

In general, the hedonic price function in a market is a nonlinear function of the characteristics z. Its functional form is determined by the distribution of buyers and their preferences, by the distribution of sellers and their costs, and by the type of equilibrium in the market. Special cases exist where more can be said. If bundles of characteristics can be unbundled, arbitrage leads to a linear hedonic price (Rosen, 1974). In the Gorman–Lancaster model, the hedonic price function is piece-wise linear (see Pollak, 1983, or Heckman and Scheinkman, 1987). In the Tinbergen (1956) model, the hedonic price is quadratic. When both buyer utility and seller costs depend on z only through an index $q(z)$, the hedonic price function satisfies $p(z) = \tilde{p}(q(z))$.

2.1 Competitive hedonic equilibrium

Consider a one-dimensional Tinbergen–Rosen model in which consumers of type $x \in \mathbf{R}_+$ choose $z \in \mathbf{R}_+$. Assume that consumer utility is $u(x, z) = x\tilde{u}(z)$ where $x\frac{\partial \tilde{u}(z)}{\partial z} > 0$. Note that $\frac{\partial^2 u(x,z)}{\partial z \partial x} = \frac{\partial \tilde{u}(z)}{\partial z} > 0$. Assume that the distribution of consumer types is described by the distribution function $F_x(x)$ with density function $f_x(x)$ and support $\mathbf{R}_+$.

Treat the supply side symmetrically. Assume that firms of type $y \in \mathbf{R}_+$ have costs of producing one unit of product z of $c(y, z) = \frac{\tilde{c}(z)}{y}$ where $(\frac{1}{y})\frac{\partial \tilde{c}(z)}{\partial z} > 0$. Note that $\frac{\partial^2 c(y,z)}{\partial z \partial y} = \left(\frac{-1}{y^2}\right)\frac{\partial \tilde{c}(z)}{\partial z} < 0$. The distribution function describing the distribution of firms is $F_y(y)$ with density $f_y(y)$ and support $\mathbf{R}_+$.

Given a differentiable price, consumers solve

$$\max_{\{z\}} \left\{ x\tilde{u}(z) - p(z) \right\}.$$

Assume there is a unique interior optimizer. The consumer first order condition is

$$x\frac{\partial \tilde{u}(z)}{\partial z} - \frac{\partial p(z)}{\partial z} = 0.$$

This equation implicitly defines the buyer demand function $z = d(x)$ and the inverse demand function $x = \tilde{d}(z) = \left(\frac{\partial p(z)}{\partial z} / \frac{\partial \tilde{u}z}{\partial z}\right)$. Note that the consumer second order condition implies that $\frac{\partial \tilde{d}(z)}{\partial z} > 0$. As a result, the distribution function describing the distribution of demand is $F_x(\tilde{d}(z)) = F_x\left(\frac{\partial p(z)}{\partial z} / \frac{\partial \tilde{u}(z)}{\partial z}\right)$.

By the same reasoning, the firms' first-order conditions define the inverse supply function $y = \tilde{S}(z) = \left(\frac{\partial \tilde{c}(z)}{\partial z} / \frac{\partial p(z)}{\partial z} \right)$ which also is monotonic. As a result the distribution of supply can be written $F_y \left(\frac{\partial \tilde{c}(z)}{\partial z} / \frac{\partial p(z)}{\partial z} \right)$.

An equilibrium hedonic price function is one that equates the distributions of supply and demand. Formally, a function $p(z)$ is an equilibrium price function if it satisfies the differential equation

$$F_x = \left(\frac{\partial p(z)}{\partial z} / \frac{\partial \tilde{u}(z)}{\partial z} \right) = F_y \left(\frac{\partial \tilde{c}(z)}{\partial z} / \frac{\partial p(z)}{\partial z} \right) \tag{8}$$

for almost all $z \in Z_m$ and if $p(z_{\min})$ ensures that all buyers and sellers obtain at least their reservation utilities.

Some simple conclusions stem from this analysis. First, since $\frac{\partial^2 u(x,z)}{\partial z \partial x} > 0$ and $\frac{\partial^2 c(y,z)}{\partial y \partial z} < 0$, the equilibrium involves positive assortative matching between buyers and sellers. Second, the equilibrium price depends on u, the preferences of buyers, c, the costs of sellers, and on F_x and F_y, the distributions of both types of agents. Third, the price function is the envelope of seller cost and buyer utility.

In more general cases and in cases of higher dimension, the differential equation (8) often does not have nice numerical properties. However, one can solve the equilibrium problem by solving the associated social welfare maximization problem which is an optimal transportation problem (an infinite dimensional linear programming problem with special structure). Recent results in this area include Gretsky, Ostroy and Zame (1999) and Chiappori, McCann, and Nesheim (2006).

2.2 Oligopoly hedonic equilibrium

When there is imperfect competition in hedonic markets, firms set prices to maximize profits. Assume individual demand is derived from the discrete choice model in section 1.2. Let $p = (p_1, \ldots, p_l)$ and $z = (z_1, \ldots, z_l)$. Given p and z, let $D_j(p, z, x) \in [0, 1]$ be the demand of consumer x for product j. Let $f_x(x)$ be the density of consumer types with support X. Aggregate demand for good j is

$$q_j(p, z) = \int_X D_j(p, z, x) f_x(x) dx.$$

Given the strategies of all firms $k \neq j$, firm j solves

$$\max_{\{z_j, p_j\}} \left\{ p_j q_j(p, z) - c(j, q_j, z_j) \right\}.$$

The first order conditions are

$$q_j + p_j \frac{\partial q_j}{\partial p_j} - \frac{\partial c_j}{\partial q} \frac{\partial q_j}{\partial p_j} = 0$$
(9)

$$p_j \frac{\partial q_j}{\partial z_j} - \frac{\partial c}{\partial q_j} \frac{\partial q_j}{\partial z_j} - \frac{\partial c}{\partial z_j} = 0.$$
(10)

A pure strategy Nash equilibrium is a set of strategies (z_j, p_j) for each firm $j = 1, \ldots, J$ such that each firm maximize profits given the strategies of its competitors. In a Nash equilibrium, the equilibrium hedonic prices p and characteristics z are determined by the distribution of buyers and their preferences, the costs of the competitors and by the competitive structure of the market. Buyers preferences u and the distribution f_x determine the structure of demand. This demand structure combined with the costs of competitors and the number of competitors determine the fierceness of competition. See, for example, Berry, Levinsohn, and Pakes (1995).

3 Estimating hedonic prices

3.1 Ideal case: z is perfectly observed

The theory of hedonic prices places no restrictions on the hedonic price functional form. The lack of theoretical predictions has led to controversy about functional form in empirical hedonic price work. Different researchers have used linear models, log linear models, Box–Cox models, and fixed-effect models. To estimate hedonic quality adjustments for use in price indexes, many statistical authorities adopt the even more restrictive ‘time-dummy’ model in which the hedonic price function takes the form

$$p_t = \beta_0 + \beta_1 z_{1t} + \beta_2 z_{2t} + \beta_3 \cdot D_t + \varepsilon_t$$
(11)

where D_t is a vector of time dummies. See Triplett (2004) for a detailed discussion. This version restricts the hedonic price function to be linear in characteristics and to have coefficients that are constant over time. The time-dummy model is rarely theoretically justified and the constant coefficient restriction is usually rejected in empirical tests. Nevertheless, Triplett (2004) argues that in many cases of interest to statistical authorities the restriction works as an approximation and does not make much empirical difference for estimates of hedonic price indexes.

There is no theoretical justification for restrictive parametric empirical models of hedonic prices unless prior knowledge of the market and the products traded exists to support the restrictions. When data-sets are large and the dimension of z is small, there is little empirical justification for parametric models either. In such cases, hedonic price functions should be estimated nonparametrically unless prior knowledge sufficient to restrict the model exists. Such nonparametric regressions can be easily estimated on desktop computers.

When sample size is small or the dimension of z is large, however, then unrestricted nonparametric methods are often impractical. In these cases, prior information should first be used to impose structure on the hedonic relationship. In some cases, it is then feasible to use semiparametric methods to estimate the hedonic relationship without imposing further structure. In many (if not most) cases, however, there is no choice but to impose further structure that is supported neither by data nor by theory. If the primary use of the method is to predict prices out-of-sample, then goodness of fit and stability with respect to changing market conditions can be useful criteria to choose functional form. If the primary use is to

estimate marginal willingness to pay in some dimension, then semiparametric methods that allow for flexibility in the dimension of interest might be of most use. Tests for robustness should be implemented and interpretations of results should consider potential misspecification biases.

3.2 Practical case: z is imperfectly observed

Empirical estimates of hedonic price functions may be biased due to omitted variables or mismeasured variables. Assume the goal is to estimate the hedonic price $p(z)$ and that the methods used will rely on estimation of conditional expectations. Discussion of estimation of $\ln(p(z))$ or methods based on other statistics such as the median would proceed along similar lines.

Let $z = (z_1, z_2)$ be the set of all hedonic characteristics and let $\tilde{z} = (\tilde{z}_1, \tilde{z}_2)$ be the set of variables that the econometrician observes. Assume that z_1 is observed without error so that $\tilde{z}_1 = z_1$. Assume that $\tilde{z}_2$ is a vector of proxy variables (or instrumental variables) and that $z_2 = g(\tilde{z}_2, \varepsilon_2)$ where ε_2 is a vector of unobservables. Let $p(z_1, z_2)$ be the theoretical hedonic price function. Observed prices $\tilde{p}$ satisfy

$$\tilde{p} = p(\tilde{z}_1, g(\tilde{z}_2, \varepsilon_2)) + \eta \tag{12}$$

where η is measurement error, $E(\eta) = 0$, and η is assumed independent of $(\tilde{z}_1, \tilde{z}_2, \varepsilon_2)$. The unobserved characteristic case, is the case where $g(\tilde{z}_2, \varepsilon_2) = \varepsilon_2$ and $f_{\varepsilon_2}(\varepsilon_2|\tilde{z}_1, \tilde{z}_2) = f_{\varepsilon_2}(\varepsilon_2|\tilde{z}_1)$. Then ε_2 is the unobserved characteristic of the product.

Under these assumptions, the expectation of $\tilde{p}$ conditional on $(\tilde{z}_1, \tilde{z}_2)$ is

$$E(\tilde{p}|\tilde{z}_1, \tilde{z}_2) = \int p(\tilde{z}_1, g(\tilde{z}_2, \varepsilon_2)) f_{\varepsilon_2}(\varepsilon_2|\tilde{z}_1, \tilde{z}_2) d\varepsilon_2 = h(\tilde{z}_1, \tilde{z}_2) \tag{13}$$

where $f_{\varepsilon_2}(\varepsilon_2|\tilde{z}_1, \tilde{z}_2)$ is the density of ε_2 conditional on $(\tilde{z}_1, \tilde{z}_2)$. This is the best predictor (in the integrated squared error sense) of $\tilde{p}$ given data on $(\tilde{z}_1, \tilde{z}_2)$. However, in general $h(\tilde{z}_1, \tilde{z}_2) \neq p(\tilde{z}_1, \tilde{z}_2)$ and little can be said about the relationship between the two without more information.

Researchers have employed instrumental variables techniques or prior information that places structure on g, on p, or on f_{ε_2} to cope with this problem. See Chay and Greenstone (2005) and Bajari and Benkhard (2005) for examples.

4 Estimating hedonic preferences

In most cases, the full set of consumer characteristics that affect choices is not observed. The econometrician observes only a subset of consumer characteristics such as education, income, age, and household structure. For example, suppose the consumer has two characteristics (x, ε) and x is observed while ε is not. Recall the consumer first order condition

$$\frac{\partial p(z)}{\partial z} = \frac{\partial u(x, \varepsilon, z)}{\partial z}. \tag{14}$$

This equation defines the hedonic demand function $z = d(x, \varepsilon)$.
When data on (x, z, p) are available, u cannot be estimated directly using (14) because z is an endogenous variable. As in Figure 1 where households with different values of x choose different value of z, households with different values of ε will choose different values of z.

Additional restrictions can help identify u. Ekeland, Heckman, and Nesheim (2004) show that the utility function can be identified nonparametrically if $\frac{\partial u}{\partial z}$ is additively separable. That is if,

$$\frac{\partial u(x, \varepsilon, z)}{\partial z} = u_0(x) + u_1(z) + \varepsilon$$

where u_0 and u_1 are arbitrary nonparametric functions.
More generally, Heckman, Matzkin and Nesheim (2005) prove that the demand function $d\bullet(x, \varepsilon)$ can be estimated using data on (z, x) alone if ε is statistically independent of x. They further show that the function u is not identified with data from a single market unless prior information is used to restrict u. For example, if marginal utility is weakly separable so that $\frac{\partial u(x, \varepsilon, z)}{\partial z} = v(q(z, x), \varepsilon)$ where q is a known function, then the function v can be estimated.
Heckman, Matzkin and Nesheim (2005) also show how to use multi-market data to estimate the unrestricted equation (14). Because cross-market variation in prices is tied to cross-market variation in the distributions of buyers and sellers, it is functionally independent of within market variation in z and x. As a result, this cross-market variation in prices can then be used to identify and estimate the function u.

See Also

- compensating differentials
- household production and public goods
- inflation measurement
- location theory
- nonparametric structural models

Bibliography

Bajari, P. and Benkhard, C. 2005. Demand estimation with heterogeneous consumers and unobserved product characteristics: a hedonic Approach. *Journal of Political Economy* 113, 1239–76.

Berry, S., Levinsohn, J. and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–90.

Chay, K. and Greenstone, M. 2005. Does air quality matter? Evidence from the housing market. *Journal of Political Economy* 113, 376–424.

Chiappori, P.-A., McCann, R. and Nesheim, L. 2006. Nonlinear hedonic pricing and matching models: finding equilibria through linear programming. Manuscript.

Ekeland, I., Heckman, J. and Nesheim, L. 2004. Identification and estimation of hedonic models. *Journal of Political Economy* 112, S60–S109.

Fujita, M. 1991. *Urban Economic Theory: Land Use and City Size*. Cambridge: Cambridge University Press.

Gorman, T. 1980. A Possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies* 47, 843–56.

Gretsky, N., Ostroy, J. and Zame, W. 1999. Perfect competition in the continuous assignment model. *Journal of Economic Theory* 88, 60–118.

Heckman, J., Matzkin, R. and Nesheim, L. 2005. Nonparametric estimation of nonadditive hedonic models. Working Paper No. CWP03/05. London: CeMMAP, Institute for Fiscal Studies.

Heckman, J. and Scheinkman, J. 1987. The importance of bundling in a Gorman–Lancaster model of earnings. *Review of Economic Studies* 54, 243–55.

Lancaster, K. 1966. A new approach to consumer theory. *Journal of Political Economy* 74, 132–56.

Muellbauer, J. 1974. Household production theory, quality, and the hedonic technique. *American Economic Review* 64, 977–94.

Pakes, A. 2003. A reconsideration of hedonic price indexes with an application to PC's. *American Economic Review* 93, 1578–96.

Pollak, R. 1983. The treatment of 'quality' in the cost-of-living index. *Journal of Public Economics* 20, 25–53.

Pollak, R. 1989. *The Theory of the Cost-of-Living Index*. New York: Oxford University Press.

Rosen, S. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* 82, 34–55.

Tinbergen, J. 1956. On the theory of income distribution. *Weltwirtschaftliches Archiv* 77, 155–73.

Triplett, J. 2004. Handbook on hedonic indexes and quality adjustments in price indexes: special application to information technology products. OECD Science Technology and Industry Working Papers, 2004/09, OECD Publishing. doi: 10.1787/643587187107.

How to cite this article

Nesheim, Lars. "hedonic prices." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000169> doi:10.1057/9780230226203.0720

The New Palgrave Dictionary of Economics Online

Heilbroner, Robert L. (1919 – 2005)

William Milberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robert Heilbroner was among the most popular historians of economic thought in the 20th century and a prominent critic of neoclassical economics and free-market capitalism. His *The Worldly Philosophers* explained how the great economists struggled to understand Western capitalism's rapid economic growth and accompanying inequities and social tensions. Heilbroner's probing ‘scenarios’ of capitalism's future drew mainly from the works of Smith, Marx and Schumpeter. His insistence that economic issues are integrally tied to moral and psychological concerns gave his work a rare depth and spoke to the political nature of all social thought.

Keywords

budget deficits; capital accumulation; capital budgeting; capitalism; economic determinism; globalization; Heilbroner, R.; history of economic thought; labour power; Lowe, A.; Marx's analysis of capitalist production; neoclassical economics; private property; profit; Schumpeter, J.; Smith, A.; socialism; wage relation

Article

One of the most prominent critics of the economics profession and of free-market capitalism, Robert Heilbroner was also responsible for motivating generations of college students to become economists. *The Worldly Philosophers: The Life, Times and Ideas of the Great Economists*, Heilbroner's classic treatment of the history of economic thought, captivated generations of readers with its elegantly written, witty, and probing discussions of how these thinkers struggled to understand Western capitalism's rapid economic growth, and industrialization and its accompanying inequities and social tensions. First published in 1953, *The Worldly Philosophers* is in its seventh edition, has been translated into 22 languages, and remains one of the best-selling books on economics of all time. Heilbroner went on to publish 25 books and over 100 articles on the history of economics and the future of capitalism,

focusing at various times on the role of the state, big business, technology, morality, psychology, private property and power. His influence went well beyond the academy, as his books were written in an accessible style and elucidated issues of concern to a broad public. He was a regular contributor to *The New Yorker* and *The New York Review of Books*, and for years served on the editorial board of the interdisciplinary journals *Dissent* and *Social Research*.

Robert Heilbroner was born in New York City in 1919, attended Horace Mann School for Boys and graduated summa cum laude from Harvard University in 1940. He worked briefly for the Office of Price Administration in Washington, D.C., before serving in the army as an interpreter in Japan in the Second World War. After the war Heilbroner came back to New York and worked as a freelance writer while he studied at the New School for Social Research. Invited to join the economics faculty at the New School, Heilbroner was granted a doctorate from the New School for his already published book *The Making of Economic Society*. Heilbroner spent his entire career at the New School for Social Research, where he helped build the programme in political economy that remains to this day one of the few Ph.D. programmes in the United States which emphasizes heterodox economics and the history of economic thought.

Heilbroner was a democratic socialist, as critical of authoritarian Soviet socialism as of dogmatic, free-market capitalism. In lucid prose, Heilbroner conveyed the consequences for everyday life of the deep and seemingly abstract economic forces which create and distribute income and wealth. His identification of these forces as embedded in politics and culture reinforced their everyday relevance.

Economics in context

The purpose of economics, Heilbroner wrote, was ‘to give meaning to economic life’. Such meaning, he argued, is necessarily forward looking: ‘There is a deep human need to be situated with respect to the future … to rescue us from a conception of social existence as all contingency and chance’ (1990a, p. 1112). Heilbroner believed that any effort to understand contemporary society required a serious consideration of the history of ideas and societies. Like his mentor Adolph Lowe, Heilbroner relied heavily on the insights of Smith, Marx and Schumpeter in his own efforts to analyse such large questions as the prospects for socialism, the viability of capitalism, and such problems as the trend of dangerous environmental degradation or the inequalities raised by the globalization of production and finance. He described these three economists as ‘great scenarists’, not because any of their long-run predictions proved right – mostly, he admitted, they were wrong – but because they provided ‘a plausible framework within which to face that most fearsome of psychological necessities – looking into the future’. Heilbroner considered these scenarios to be ‘the most significant accomplishment of economics’ (1995, pp. 5–6).

Heilbroner insisted on understanding capitalism as a particular stage in the long history of human efforts to solve the ‘economic problem’ of material provisioning and social reproduction. Knowledge of how different societies have confronted these problems gives crucial perspective to our own efforts to do so today. Thus the starting point for understanding contemporary economic life is to identify the distinguishing features of the current economic system: capitalism. Modern economics, Heilbroner argued, has largely avoided this first crucial step, ignoring rather than illuminating the rich array of social, psychological and moral forces that propel capitalist societies. ‘[B]ehind the veil of conventional economic rhetoric’, he wrote in a short autobiographical essay (2000, p. 287), ‘we can easily discern an

understructure of traditional behavior – trust, faith, honesty, and so on – as a necessary moral foundation for a market system to operate, as well as a concealed superstructure of power.' Heilbroner noted with outrage that even the word 'capitalism' had disappeared from the economics textbooks. He saw economics as an 'explanation system' of capitalism, and insisted on the relevance of economics to large questions of political economy – the role of the state, the sustainability of environmental health, the problem of world poverty and the danger of nuclear war – rather than small questions of optimal allocation under conditions of scarcity.

Capitalism's nature and logic

Rejecting neoclassicism, Heilbroner turned to Smith and Marx for the central building blocks of social analysis, since both identify a logic of capitalist development which explains capitalism's endurance and its inherent limitations. Marx is especially important because of his focus on the particularity of the capitalist drive for the expansion of wealth and the exigencies of power, politics and psychology brought on by this accumulation drive. As Heilbroner elaborated in a series of books written in the 1980s – *The Nature and Logic of Capitalism*, *Marxism For and Against*, and *Behind the Veil of Economics* – all efforts to solve the economic problem of material provisioning, be they organized by tradition, command or markets, are aimed at the production of a material surplus above the needs of subsistence. Only in capitalism, however, does this take a general form – self-expanding value. The commodity is but a way station towards the accumulation of value in 'a never-ending metamorphosis of M-C-M′′' (1988, p. 37). This circuit of capital hinges on the institution of private property in the means of production, which 'organizes and disciplines' society and serves as an instrument of power 'because its owners can establish claims on output as their quid pro quo for permitting access to their property' (1988, p. 39). Profit goes to owners of capital and not only validates the activities of particular owners but perpetuates the M-C-M′ circuit. As Heilbroner writes, 'Profit is for capitalism what victory is for a regime organized on military principles…' (1988, p. 41).

The wage relation is crucial to understanding capitalism's uniqueness. Workers are free to offer their labour power for wages, unlike forced labour in many traditional and especially feudal and slave societies. But private property relations keep workers from retaining the full value of their efforts. A second unique feature of capitalism is the distinctiveness of its private and public realms, each relying on the other for its sustenance. Capitalism thus has a unique political agenda in that the precise role and scope of the state vis-à-vis the private sector is constantly contested and debated. Despite the freedom embodied in the wage relation and the reliance of the private sector on the state, capitalism has functioned under both democratic and anti-democratic political systems. Heilbroner himself was an outspoken advocate for an active role for government in creating a decent society and productive economy. In *The Debt and the Deficit: False Alarms, Real Possibilities* (1989, co-authored with Peter Bernstein) Heilbroner argued for Keynesian deficit spending and capital budgeting by the US government.

For all its identifiable deep structures and logic, capitalism for Heilbroner is constantly changing, buffeted by other social forces. This is partly the result of history's dialectical nature: as problems are resolved through social change, the new conditions present a new set of problems. In *The Future as History* and *An Inquiry into the Human Prospect* Heilbroner focused on various implications of this unsettling aspect of social reality, in particular long-run environmental consequences of economic

development. Capitalism is all the time and everywhere contingent on independent ideas, political struggles and ethical dilemmas, and these have resulted in a variety of capitalisms around the world. For Heilbroner, theories of economic determinism – be they Marxist or neoclassical – that reduce capitalism to a system of markets cannot adequately explain social change, since economics, politics and morality are linked: '[T]he engines of history do not draw all their energies from economic drives and institutions. If socialism failed, it was for political, more than economic reasons; and if capitalism is to succeed it will be because it finds the political will and means to tame its economic forces' (1996, p. 195).

The paradox of progress

If it was Marx who best articulated the 'nature and logic' of capitalism, it was Smith who provided the most important insights into the psychology of individuals in capitalist society. Heilbroner considered Smith's *Theory of Moral Sentiments* to be as important as *The Wealth of Nations* and insisted that the socialized individual of *Theory of Moral Sentiments* was not only consistent with but necessary to the successful working of the nascent capitalism described in *The Wealth*. Smith's writings on empathy and, most importantly, subservience ('the principle of authority') and the drive for self-betterment are the psychological foundations of the 'society of perfect liberty'; that is, they are the psychological dynamics that make capitalism function. Like the other classical economists, Smith also emphasized capitalism's dark side. Capitalism's advance brings unprecedented wealth creation and the possibility of 'perfect liberty'. It also brings stagnation, poverty, inefficiency, systemic corruption and moral decay. The result was what Heilbroner termed 'the paradox of progress'. For Heilbroner these insights were important not only for students of intellectual history but also for those seeking to understand the prospects for capitalism today. 'Capitalism's uniqueness in history', he wrote in *Twenty-First Century Capitalism*, 'lies in its continuously self-generated change, but it is this very dynamism that is the system's chief enemy' (1993, p. 130). Both of these insights – the embeddedness of the economy in a broader social, political and psychological fabric, and the inherent problems in capitalist development – Heilbroner attributes to Adam Smith, although they are not part of the canonical reading of Smith as a proponent of laissez-faire.

Smith's influence on Heilbroner went beyond the issue of the psychology of individual agents in capitalism and into the existential question of the purpose of theory itself. In his *Essays on Astronomy* (published in 1758 and excerpted in Heilbroner, 1986), Smith wrote that '[T]he repose and tranquility of the imagination is the ultimate end of philosophy…Philosophy, by representing the invisible chains which bind together all these disjointed objects, endeavors to introduce order in this chaos of jarring and discordant appearances, to ally this tumult of the imagination.' Discussing this passage, Heilbroner wrote that 'We theorize … to restore our peace of mind' (1986, p. 16).

Analysis and vision in economics

Heilbroner's embrace of the classicals and rejection of the neoclassicals hinged on the Schumpeterian distinction between 'analysis' and 'vision'. Schumpeter (1954, p. 41) defined vision as the 'preanalytic cognitive act' that is inevitable and ideological. Analysis is the largely deductive process that follows

from the theory's foundations. As intellectual historian, Schumpeter separated economic analysis from its vision, leading to the *History of Economic Analysis*, published posthumously. Heilbroner embraced the Schumpeterian categories, and especially the notion that vision is an inevitable part of the process of theorizing, since ‘All systems of thought that describe or examine societies must contain their political character, knowingly and explicitly, or unknowingly and in disguise.’ (1990b, p. 109) But Heilbroner resisted Schumpeter's separation of vision and analysis, since connecting the two allowed a greater appreciation of how economic scenarios are formed. *The Worldly Philosophers* was enormously popular not only because it included juicy biographical details about the early economists but because Heilbroner revealed the lively imagination and political engagement of the ‘great scenarists’. Vision, for Heilbroner, embodied much of the creativity that informs economic problem-solving and modelling. And it is through the vision that ethical and epistemological principles are brought into theory. Vision is the expression ‘of the inescapable need to infuse ‘meaning’ – to discover a comprehensive framework – in the world’ (1990a, p. 1112). For Heilbroner, it was precisely the persistent denial of the role of vision that leaves modern economics so limited as a tool for understanding social life. In *The Crisis of Vision in Modern Economic Thought*, Heilbroner (and co-author William Milberg) developed this theme in the context of contemporary debates in macroeconomics.

In the final chapter of the seventh edition of *The Worldly Philosophers*, Heilbroner wrote of ‘the end of economics’, playing on the dual meaning of end as both purpose and termination. For all its technical sophistication, modern economics has largely failed to accomplish the purpose of a worldly philosophy: to give meaning to economic life. Heilbroner saw the narrowness of modern economics as an abandonment of the grand aspiration for social thought that Smith, Ricardo, Malthus, Marx, Mill, Keynes and Schumpeter each held in their day. Heilbroner lamented, ‘The new vision is Science, the disappearing one capitalism’ (1953, p. 314). Heilbroner was an important public intellectual of the second half of the 20th century. While he remained to his last days a severe critic of modern economics, his personal warmth, his kindness, his humaneness, his commitment to equality, opportunity and democracy, and his love of deep and serious debate on pressing social issues endeared him to a broad group of professional economists, social scientists, students and a socially-concerned public.

See Also

- capitalism
- Marx's analysis of capitalist production
- Schumpeter, Joseph Alois
- Smith, Adam
- wealth

Selected works

1953. *The Worldly Philosophers: The Lives and Times of the Great Economists*, 7th edn. New York: Simon and Schuster, 1999.

1959. *The Future as History*. New York: Harper and Brothers.

1963. (With W. Milberg.) *The Making of Economic Society*, 12th edn. Englewood, NJ: Prentice-Hall, 2006.

1974. *An Inquiry Into the Human Prospect*, rev. edn. New York: W.W. Norton, 1980.

1980. *Marxism: For and Against*. New York: W.W. Norton.

1982. The socialization of the individual in Adam Smith. *History of Political Economy* 143, 427–39.

1985. *The Nature and Logic of Capitalism*. New York: W.W. Norton.

1986. *The Essential Adam Smith*. New York: W. W. Norton.

1988. *Behind the Veil of Economics*. New York: W.W. Norton.

1989. (With P. Bernstein.) *The Debt and the Deficit: False Alarms/Real Possibilities*. New York: W.W. Norton.

1990a. Analysis and vision in the history of modern economic thought. *Journal of Economic Literature* 28, 1097–114.

1990b. Economics as ideology. In *Economics as Discourse*, ed. W. Samuels. Boston: Kluwer Academic.

1993. *Twenty-First Century Capitalism*. New York: W.W. Norton.

1995. (With W. Milberg.) *The Crisis of Vision in Modern Economic Thought*. New York: Cambridge University Press.

1996. *Teachings from the Worldly Philosophy*. New York: W.W. Norton.

2000. Robert Heilbroner. In *A Biographical Dictionary of Dissenting Economists*, 2nd edn, ed. P. Arestis and M. Sawyer. Aldershot: Edward Elgar.

Bibliography

Schumpeter, J. 1954. *History of Economic Analysis*. New York: Oxford University Press.

How to cite this article

Milberg, William. "Heilbroner, Robert L. (1919–2005)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000170> doi:10.1057/9780230226203.0721

The New Palgrave Dictionary of Economics Online

Heller, Walter Perrin (1942 – 2001)

Ross M. Starr
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Arrow, K.; Baumol–Tobin money demand model; consumption function; coordination failure; Council of Economic Advisers; demand for money; general equilibrium; growth, models of; Heller, W.P.; money and general equilibrium; non-clearing markets in general equilibrium; Stanford Institute for Theoretical Economics; Walter, W.W.

Article

Walter Perrin Heller was a leading 20th-century economic theorist, and an early member of the University of California, San Diego, faculty (from 1974 to his death in 2001). He annually taught the UCSD graduate core microeconomic theory course on welfare economics.

Heller came from an academic family distinguished in the economics discipline: his father, Walter W. Heller, was Professor of Economics at the University of Minnesota and served as chairman of the President's Council of Economic Advisers in the Kennedy and Johnson US presidential administrations. Walter P. Heller's undergraduate education took place at Oberlin College and at the University of Minnesota, particularly under the guidance at Minnesota of Professor Leonid Hurwicz (1990 recipient of the US National Medal of Science). Heller's intellectual home was Stanford University. He received his Ph.D. there in 1970 with the dissertation advice of Nobel Prize winner Kenneth J. Arrow. For three decades he participated in the Stanford summer economic theory workshop at the Institute for Mathematical Studies in the Social Sciences (IMSSS) and its successor, the Stanford Institute for Theoretical Economics (SITE). Prior to joining the UCSD faculty, he was on the economics faculty of the University of Pennsylvania.

Heller served as an associate editor of the *Journal of Economic Theory* and on the executive committee of the American Economic Association. His research treated the stability of economic growth, microeconomic foundations of macroeconomics and of the demand for money, and resource allocation under conditions of market failure due to incompleteness or monopoly. In the late 1980s and the 1990s, the research focused on a fundamental issue in the theory of unemployment, namely, *coordination failure*, or the inability – even of complete markets in price equilibrium – successfully to match supply

and demand, workers and employers. Kenneth Arrow remarked at Heller's memorial service at Stanford on 16 July 2001:

> Economic theory backed by serious mathematical reasoning was just beginning to be recognized when Walt started his graduate work … Walt was one of the leaders in using new ways – not merely for clarification – but for changing the way the economy was considered. He contributed to many aspects of [economic] theory … His long-standing project of studying the coordination failures of the economic system brought out, in an essentially novel way, the previously unclarified meaning of Keynesian insights. This work … is a vital continuing part of modern economic thought …

Stability of economic growth: A growth model over time in general competitive equilibrium (at each instant) may nevertheless be on an inter-temporally inefficient path (Hahn, 1966; 1968; Malinvaud, 1953). Further, an efficient path may be unstable (Samuelson and Solow, 1956). Heller (1971; 1975) demonstrated that inefficiency and instability depend on myopia; in the presence of complete inter-temporal capital markets (futures markets for capital), stability and efficiency of the growth path are established.

Demand for money: Heller was among the first to apply the full formal structure of an Arrow–Debreu model to the analysis of a monetary economy (1972; 1974; 1976 with R. Starr). The Baumol–Tobin money demand model with transaction costs (Tobin, 1956) is shown to be consistent with full general competitive equilibrium.

Foundations of macroeconomics: The Keynesian consumption function was long recognized anecdotally to be a result of capital market imperfection, but Heller and Starr (1979b) represents the first mathematical formalization of this notion. Unemployment equilibrium was long thought inconsistent with Walrasian general equilibrium pricing; Heller and Starr (1979a) demonstrate that expectations of uncleared markets may be self-fulfilling in equilibrium even at competitive equilibrium prices.

Coordination failure: When the formation of markets is itself a resource using activity, then some markets may not form or announce prices in equilibrium (1986; 1992; 1999) with resulting inefficiency and unemployed resources. In a model with a non-competitive (oligopoly or monopoly) sector, even with a full set of markets, there may be multiple Pareto ranked equilibria (1998).

Heller's work is elegantly written so that the underlying intuition is clear and is supported by mathematical structure. The Walter P. Heller Prize for excellence in research – instituted by Heller's colleagues – is awarded annually to a UCSD graduate student.

See Also

- Arrow, Kenneth Joseph
- general equilibrium
- Heller, Walter Wolfgang
- money and general equilibrium
- non-clearing markets in general equilibrium

Selected works

1971. Disequilibrium dynamics of competitive growth paths. *Review of Economic Studies* 38, 385–400.

1972. Transactions with set-up costs. *Journal of Economic Theory* 4, 465–78.

1974. The holding of money balances in general equilibrium. *Journal of Economic Theory* 7, 93–108.

1975. Tatonnement stability of infinite horizon models with saddle-point instability. *Econometrica* 43, 65–80.

1976. (With R. Starr.) Equilibrium with non-convex transactions costs: monetary and non-monetary economies. *Review of Economic Studies* 43, 195–215.

1979a. (With R. Starr.) Unemployment equilibrium with myopic complete information. *Review of Economic Studies* 46, 339–59.

1979b. (With R. Starr.) Capital market imperfection, the consumption function, and the effectiveness of fiscal policy. *Quarterly Journal of Economics* 93, 455–63.

1986. Coordination failure under complete markets with applications to effective demand. In *Essays in Honor of Kenneth J. Arrow. Volume 2, Equilibrium Analysis*, ed. W. Heller, R. Starr and D. Starrett. New York: Cambridge University Press.

1992. Underemployment as a coordination problem with savings and increasing returns. In *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn*, ed. P. Dasgupta. et al. Cambridge MA: MIT Press.

1998. (With A. Edlin and M. Epelbaum.) Is perfect price discrimination really efficient?: Welfare and existence in general equilibrium. *Econometrica* 66, 897–922.

1999. Equilibrium market formation causes missing markets. In *Markets, Information, and Uncertainty: Essays in Economic Theory in Honor of Kenneth J. Arrow*, ed. G. Chichilnisky. New York: Cambridge University Press.

Bibliography

Hahn, F. 1966. Equilibrium dynamics with heterogeneous capital goods. *Quarterly Journal of Economics* 80, 633–40.

Hahn, F. 1968. On warranted growth paths. *Review of Economic Studies* 35, 175–84.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21, 233–68.

Samuelson, P. and Solow, R. 1956. A complete capital model involving heterogeneous capital goods. *Quarterly Journal of Economics* 70, 537–62.

Tobin, J. 1956. The interest elasticity of the transactions demand for cash. *Review of Economics and Statistics* 38, 241–7.

How to cite this article

Starr, Ross M. "Heller, Walter Perrin (1942–2001)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000128> doi:10.1057/9780230226203.0722

The New Palgrave Dictionary of Economics Online

Heller, Walter Wolfgang (1915 – 1987)

Joseph A. Pechman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Council of Economic Advisers; full-employment surplus; Germany, economics in; Heller, W. W.; potential gross national product; progressive taxation; public finance; revenue sharing; tax expenditures

Article

Heller was born in Buffalo on 27 August 1915. He grew up in Seattle and Milwaukee, and graduated from Oberlin College. He received a doctorate in economics from the University of Wisconsin, where he studied with Harold M. Groves, who greatly influenced a generation of public finance scholars. He spent his entire academic career as professor of economics at the University of Minnesota.

Heller made important scholarly contributions to the study of public finance, but his major claim to fame was his highly successful term as chairman of the Council of Economic Advisers under Presidents John F. Kennedy and Lyndon B. Johnson from 1961 to 1964. After leaving the government, he was influential as a consultant and adviser to presidents, Congress and business. He wrote widely on current economic developments, tax policy, and state-local finance, and was also known as a stimulating lecturer and commentator on economic policy issues. In 1974, he served as president of the American Economic Association.

Heller began his professional career as an expert on state and local taxation. He wrote his doctoral dissertation on the administration of state income taxes, and later originated the idea of federal revenue sharing with the states and local governments. The details of revenue sharing were developed by a task force appointed by President Johnson, but it was enacted by Congress only after it was recommended by President Richard M. Nixon in 1972. The revenue sharing legislation was extended until the end of September 1986.

During the Second World War, Heller moved to the Treasury Department, where he contributed to the development of tax policy to finance the war. In 1947–8, he was tax adviser to the US Military Government in Germany, where he played an important role in designing the currency and fiscal reforms that helped launch the post-war German economic revival. He also served as a consultant to the

Treasury Department during the late 1940s and early 1950s. He has been a strong advocate of progressive taxation and was one of the first to recognize that unnecessary deductions and tax preferences narrow the income tax base, require higher marginal tax rates to raise the necessary revenues, and distort economic decisions.

As chairman of the Council of Economic Advisers, Heller supported innovative macroeconomic policies to promote economic growth and stability. He persuaded President Kennedy to propose a major tax cut to stimulate demand, advocated the enactment of an investment tax credit and liberalized depreciation allowances to increase investment incentives. His Council developed the first, and most successful, voluntary wage–price guidelines to help contain inflationary pressures as the economy moved to full employment.

Heller's Council pioneered fiscal analysis based on the concepts of potential gross national product – the output the economy would produce at full employment – and the full-employment surplus. It is also noted for its advocacy of the neoclassical Keynesian synthesis of fiscal and monetary policies required to achieve full employment and increase economic growth. To reach full employment, it proposed the use of stimulating budget and monetary policies. To increase growth at full employment, it stressed the need for a full-employment surplus and monetary ease to support private investment in plant and equipment, combined with public investments in education, research, and development. It also urged the dismantling of barriers to free trade among nations to achieve the benefits of international specialization and exchange.

As a result of the policies pursued by the Kennedy and Johnson administrations, the nation enjoyed a long period of economic growth and prosperity without inflation. From the fourth quarter of 1960 to the fourth quarter of 1964 (when Heller left his CEA post), US real GNP grew at an average annual rate of 4.9 per cent, consumer prices rose 1.2 per cent a year, and long-term federal bond yields never exceeded 4.2 per cent.

Heller combined his advocacy of sound economic policies with an understanding of the need to help the disadvantaged and underprivileged. He helped to persuade President Johnson to design and implement an anti-poverty programme to provide economic opportunities for low-skilled workers and a decent income for those who cannot earn their own livelihood. ‘We cannot relax our efforts to increase the technical efficiency of economic policy’, he wrote in 1966. ‘But it is also clear that its promise will not be fulfilled unless we couple with improved techniques of economic management a determination to convert good economics and a great prosperity into a good life and a great society.’

Selected works

1952. Limitations of the federal individual income tax. *Journal of Finance* 7(2), 185–202.

1959. (With C. Penniman.) *State Income Tax Administration*. Madison: University of Wisconsin Press.

1966. *New Dimensions of Political Economy*. Cambridge, MA: Harvard University Press.

1968. A sympathetic reappraisal of revenue sharing. In *Revenue sharing and the City*, ed. H.S. Perloff and R.P. Nathan. Baltimore: Johns Hopkins University Press.

1969. *Monetary vs. Fiscal Policy* (a dialogue with Milton Friedman). New York: W.W. Norton.

1975. What's right with economics? *American Economic Review* 65, 1–26.

1976. *The Economy: Old Myths and New Realities*. New York: W.W. Norton.

1982. Kennedy economics revisited. In *Economics in the Public Service: Papers in Honor of Walter W. Heller*, ed. J.A. Pechman and N.J. Smiler. New York: W.W. Norton.

How to cite this article

Pechman, Joseph A. "Heller, Walter Wolfgang (1915–1987)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000041> doi:10.1057/9780230226203.0723

The New Palgrave Dictionary of Economics Online

Hennipman, Pieter (1911 – 1994)

Arnold Heertje
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Chamberlin, E. H.; Hennipman, P.; *homo economicus* ; interpersonal utility comparisons; monopoly; Pareto optimal redistribution; Pierson, N. G.; Robbins, L. C.; Tinbergen, J.; welfare economics; unanimity; Wicksell, J. G. K.

Article

Dutch economist born in Leiden, 12 September 1911, who died in Amsterdam on 3 July 1994. Hennipman belongs to the three most important economists of the Netherlands, the two others being Nicolaas Gerard Pierson (1839–1909) and Jan Tinbergen (1903–1994). He studied at the Faculty of Economics of the University of Amsterdam, and was taught economic theory by H. Frijda and economic history by N.W. Posthumus. He took his Master's degree in 1934, and in 1938 became reader in economics at the University of Amsterdam, next to his beloved teacher Frijda. He continued his work on his dissertation and received his doctorate in July 1940, in time to enable Frijda, who soon after had to flee from the Nazis, to act as his director of his thesis. Of Hennipman's impressive work on economic motive and economic principle a much-enlarged edition appeared after the Second World War in 1945. The book presents a detailed historical-critical survey of the manifold varieties of *homo economicus*, concluding that the scope of economics is not restricted to the behaviour of such an animal. It is argued that the concept of economic welfare is subjective and devoid of specific content and that economics cannot be normative. His work shows the influence of the Austrian subjectivist way of thinking and Lionel Robbins' *Essay* (1932). In 1945 Hennipman became Professor of Economics at the University of Amsterdam.
From 1945 to 1972 Hennipman was managing editor of the Dutch Journal *De Economist*, nowadays published in English. Many articles have appeared which reveal evidence of his vast knowledge of the literature and demonstrate his ability to encourage authors to improve their manuscripts by his constructive and well-founded comments. In 1951, invited by E.H. Chamberlin, Hennipman participated in a conference held by the International Economic Association on monopoly, competition and their regulation. Hennipman's paper 'Monopoly: Impediment or Stimulus to Economic Progress?' received

great praise at the conference from J.M. Clark, G. Haberler, F.H. Knight and F. Machlup (Hennipman, 1954). In 1962 he published his essay on the theory of economic policy, of which a shortened version in English is published in Hennipman's book *Welfare Economics and the Theory of Economic Policy* (Hennipman, 1995). The analysis builds further on his dissertation by applying to the theory of economic policy the principles set out in his work on economic motive and economic principle. This essay is without doubt one of the highlights of non-mathematical economic literature. He contributed to the publication of the Walras correspondence, edited by W. Jaffé in 1965 (Jaffé, 1965).

Following his retirement in 1973 Hennipman was very active on methodology, the history of economic thought and, in particular, welfare economics. Publications during the last decade of his life mainly concern welfare economics: for example, a pair of articles exploring the historical and analytical relations between Pareto optimality and Wicksellian unanimity. A major theme in Hennipman's work is the contention that welfare economics is a non-normative theory, as he convincingly spelled out in major debates with Ezra Mishan and Mark Blaug (Hennipman, 1995). Interpersonal comparisons of utility and Pareto optimal redistribution are discussed by Hennipman from this point of view.

It is only due to his incredible and miraculous modesty that the international audience of economists had to wait until after his death for an accessible publication of his work in English. This event also explains that his influence on the development of international economics literature fell short of what would have been justified by the high quality and relevance of his contributions, which are innocent of mathematics and do not reflect empirical research. He influenced both students and professors by allowing them access to his vast knowledge of almost all areas of economic theory and his analytical insights. There is no doubt that he became the leading Dutch economist, in particular since the war, albeit still in the shadow of Jan Tinbergen, who had built his international reputation during the years of the Great Depression of the 1930s.

See Also

- Chamberlin, Edward Hastings
- Pierson, Nichlaas Gerard
- Robbins, Lionel Charles
- Tinbergen, Jan

Selected works

1945. *Economisch Motief en Economisch Principe*. Amsterdam: North-Holland.

1954. Monopoly: impediment or stimulus to economic progress? In *Monopoly and Competition and Their Regulation*, ed. E.H. Chamberlin. London: Macmillan.

1995. *Welfare Economics and the Theory of Economic Policy*, ed. D. Walker, A. Heertje and H. van den Doel, with an introduction by D. Walker. Aldershot: Edward Elgar.

Bibliography

Jaffé, W. 1965. *Correspondence of Léon Walras and Related Papers*, 3 vols. Amsterdam: North-Holland.

Robbins, L.C. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

How to cite this article

Heertje, Arnold. "Hennipman, Pieter (1911–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000188> doi:10.1057/9780230226203.0724

The New Palgrave Dictionary of Economics Online

Hermann, Friedrich Benedict Wilhelm von (1795 – 1868)

Mark Blaug
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cameralism; German economics in the early 19th century; life tables; supply and demand

Article

Hermann was born in Dinkelsbuhl, Germany. His career spanned the half-century or more in which German economics came to terms with English classical political economy, first welcoming it and then rejecting it, particularly in its Ricardian variety. After teaching mathematics in a secondary school, Hermann was appointed to the chair in what was still called *Kameralwissenschaften* [Cameralism] – an old title soon to be discarded – at the University of Munich in 1827. He made his reputation with *Staatswirthschaftliche Untersuchungen* [Investigations into Political Economy] (1832), a book which owed much to *The Wealth of Nations* but little to the writings of either Malthus or Ricardo. The book was organized around the simple but appealing idea that all economic variables are the outcome of the forces of demand and supply, so that economic analysis consists essentially of an investigation of the factors lying behind demand and supply. The book revelled in endless definitions and classifications of types of goods, wants, costs, capitals, and so on, but did not clutter the analysis with endless attacks on the deductive method of the English school. Together with Rau (1792–1870), Hermann thereby laid the foundations on which Mangoldt (1824–68) and Thünen (1783–1850) were soon to build a German brand of classical economics. No wonder Marshall much admired ‘Hermann's brilliant genius’ and frequently quoted Hermann's treatise in his own *Principles of Economics* (1890).
Hermann became a Director of the Bavarian Statistical Bureau in 1839 and organized the first official life table covering an entire German state. As a member of the Frankfurt Parliament in 1848, he advocated the unification of all German states.

How to cite this article

Blaug, Mark. "Hermann, Friedrich Benedict Wilhelm von (1795–1868)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan,

2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000047> doi:10.1057/9780230226203.0725

The New Palgrave Dictionary of Economics Online

Herskovits, Melville Jean (1895 – 1963)

George Dalton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

conventional economics; economic anthropology; formalism; Herskovits, M. J.; Knight, F. H.; Marxism; non-market economies; Polanyi, K.; substantivism

Article

Herskovits was born in Bellefontaine, Ohio, and died in Evanston, Illinois. He studied history at the University of Chicago (BA, 1920) and anthropology at Columbia University (Ph.D., 1923) as a student of Franz Boas. He taught at Columbia and Howard universities before going to Northwestern in 1927, where he spent the rest of his academic career. Herskovits did anthropological fieldwork in West Africa, the Caribbean and Brazil, and was among the first American anthropologists to specialize in African societies as well as blacks in the Caribbean and the United States. He started the first Program of African Studies in the United States, at Northwestern.

Herskovits was an early contributor to the field of study now established as economic anthropology. The first edition of his book on this topic was called *The Economic Life of Primitive Peoples* (1940), the revised edition being *Economic Anthropology* (1952).

Herskovits is best remembered by economic anthropologists for his views on a theoretical issue of importance that arose in his controversy with Frank Knight, who reviewed the 1940 edition of Herskovits's book. In the 1940 edition, Herskovits criticized the conventional economics of Marshallian microtheory for its uselessness to anthropologists trying to understand the underlying principles which explain the working of primitive economies – such as African tribal economies not yet changed by European colonial rule – primitive economies lacking capitalism's core attributes of machine technology, modern money, and market organization for the transaction of inputs and outputs. In his book review, Frank Knight criticized Herskovits for misunderstanding the ‘abstract’ and ‘intuitive’ nature of economic theory. (I doubt that Knight's portrayal of economics, as stated there, would be shared today by many economists.) Knight's review, together with a rejoinder by Herskovits, are reprinted in *Economic Anthropology* (1952).

The relevance of conventional economic theory to the analysis of pre-industrial, non-capitalist

economies remains an unresolved issue to this day. It is an issue much more important today because of the much greater interest now in the study of early and primitive economies, and in the study of the large, diverse set of developing economies in the Third World. This inability to agree on the relevance of conventional economics to the analysis of non-market economies finds expression in economic anthropology's literature of acrimonious theoretical dispute and in the existence side by side of three radically different theoretical systems all employed by archaeologists, anthropologists and historians to analyse non-capitalist economies: formalism (that is, conventional microeconomic theory); Marxism; and substantivism (that is, Karl Polanyi's system of analysis described in his *Trade and Market in the Early Empires*, 1957, and his *Primitive, Archaic, and Modern Economies*, 1971).

See Also

- economic anthropology
- Knight, Frank Hyneman
- Polanyi, Karl

Selected works

1940. Anthropology and economics. In *The Economic Life of Primitive Peoples*. New York: Knopf.

1941. Economics and anthropology: a rejoinder. *Journal of Political Economy* 49, 269–78.

1952. *Economic Anthropology*. New York: Knopf.

Bibliography

Dalton, G. 1961. Economic theory and primitive society. *American Anthropologist* 63(February), 1–25.

Dalton, G. 1969. Theoretical issues in economic anthropology. *Current Anthropology* 10(February), 63–102.

Knight, F. 1941. Anthropology and economics. *Journal of Political Economy* 49, 247–68.

Martin, J. and Knapp, K., eds. 1967. *The Teaching of Development Economics*. Chicago: Aldine Press.

Simpson, G.E. 1973. *Melville J. Herskovits*. New York: Columbia University Press.

How to cite this article

Dalton, George. "Herskovits, Melville Jean (1895–1963)." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000048> doi:10.1057/9780230226203.0726

The New Palgrave Dictionary of Economics Online

heterodox economics

Frederic S. Lee

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Although ‘heterodox economics’ is a widely used term, precisely what it means is debated. I argue that heterodox economics refers to a body of economic theories that holds an alternative position vis-à-vis mainstream economics; to a community of heterodox economists who identify themselves as such and embrace a pluralistic attitude towards heterodox theories without rejecting contestability and incommensurability among heterodox theories; and to the development of a coherent economic theory that draws upon various theoretical contributions by heterodox approaches which stand in contrast to mainstream theory.

Keywords

Association for Heterodox Economics; Austrian economics; capitalism; class; evolutionary economics; feminist economics; heterodox economics; institutional economics; interdependencies; Marxian economics; Post Keynesian economics; radical economics; scarcity; social economics; Sraffian economics; pluralism; political economy; mainstream economics; social provisioning

Article

‘Heterodox economics’ refers to economic theories and communities of economists that are in various ways an alternative to mainstream economics. It is a multi-level term that refers to a body of economic theories developed by economists who hold an irreverent position vis-à-vis mainstream economics and are typically rejected out of hand by the latter; to a community of heterodox economists who identify themselves as such and embrace a pluralistic attitude towards heterodox theories without rejecting contestability and incommensurability among heterodox theories; and to the development of a coherent economic theory that draws upon various theoretical contributions by heterodox approaches which stand in contrast to mainstream theory. Thus, the article is organized as follows. The first section outlines the emergence of ‘heterodox economics’ in the sense of a body of heterodox theories; the second section

deals with heterodox economics as a pluralist community of heterodox economists; the third section situates heterodox economics relative to mainstream economics; and the fourth section delineates heterodox economics in terms of theory and policy.

Heterodox economics as a group of heterodox theories

Heterodox as an identifier of an economic theory and/or economist that stands in some form of dissent relative to mainstream economics was used within the Institutionalist literature from the 1930s to the 1980s. Then, in 1987, Allan Gruchy used ‘heterodox economics’ to identify Institutional as well as Marxian and Post Keynesian theories as ones that stood in contrast to mainstream theory. By the 1990s, it became obvious that there were a number of theoretical approaches that stood, to some degree, in opposition to mainstream theory. These heterodox approaches included Austrian economics, feminist economics, Institutional-evolutionary economics, Marxian-radical economics, Post Keynesian and Sraffian economics, and social economics; however, none of the names of the various heterodox approaches were suitable as a general term that could represent them collectively. While terms such as ‘non-traditional’, ‘non-orthodox’, ‘non-neoclassical’ and ‘non-mainstream’ were used to collectively represent them, they did not have the right intellectual feel or a positive ring. Moreover, some thought that ‘political economy’ (or ‘heterodox political economy’) could be used as the collective term, but its history of being another name for Marxian-radical economics (and its current reference to public choice theory) made this untenable. Therefore, to capture the commonality of the various theoretical approaches in a positive light without prejudicially favouring any one approach, a descriptive term that had a pluralist ‘big-tent feel’ combined with being unattached to a particular approach was needed. Hence, ‘heterodox’ became increasingly used throughout the 1990s in contexts where it implicitly and/or explicitly referred to a collective of alternative theories vis-à-vis mainstream theory and to the economists who engaged with those theories.
The final stage in the general acceptance of heterodox economics as the ‘official’ collective term for the various heterodox theories began c. 1999. First there was the publication of Philip O'Hara's comprehensive *Encyclopedia of Political Economy* (1999), which explicitly brought together the various heterodox approaches. At the same time, in October 1998, Fred Lee established the Association for Heterodox Economics (AHE); and to publicize the conference and other activities of the AHE as well as heterodox activities around the world, he also developed from 1999 an informal ‘newsletter’ that eventually became (in September 2004) the *Heterodox Economics Newsletter*, now received by over 1,600 economists worldwide (see http://www.heterodoxnews.com). These twin developments served to establish ‘heterodox economics’ as the preferred terminology by which these groups of economists referred to themselves.

Heterodox economics as a community of heterodox economists

‘Heterodox economics’ also denotes a community of heterodox economists, which implies that the members are not segregated along professional and theoretical lines. The segregation of professional engagement has not existed among heterodox associations, with the exception of two instances in the mid-1970s. For example, from their formation in 1965–70, the three principal heterodox associations in

the United States, AFEE, ASE, and URPE (see Table 1 for full names), opened their conferences to Institutionalist, social economics, radical-Marxian, and Post Keynesian papers and sessions; appointed and/or elected heterodox economists to the editorial boards of their journals and to their governing bodies who also were members of other heterodox associations or engaged with Post Keynesian economics; and had members who held memberships in other heterodox associations, engaged with Post Keynesian economics, and subscribed to more than one heterodox economics journal. Moreover, a number of heterodox associations formed since 1988, such as AHE, EAEPE, ICAPE, SDAE and SHE, have adopted an explicitly pluralistic approach towards their name, membership and conference participation: for a list of heterodox associations, dates formed and primary country or region of activity, see Table 1. Finally, the informal and explicit editorial policies of heterodox journals have, from their formation, accepted papers for publication that engage with the full range of heterodox approaches; and this tendency strengthened since the mid-1990s as heterodox economics became more accepted. To illustrate this point, from 1993 to 2003 the eight principal English-language generalist heterodox journals – *Cambridge Journal of Economics*, *Capital and Class*, *Feminist Economics*, *Journal of Economic Issues*, *Journal of Post Keynesian Economics*, *Review of Political Economy*, *Review of Radical Political Economics*, and *Review of Social Economy* – cited each other so extensively that no single journal or subset of journals was isolated; hence they form an interdependent body of literature where all heterodox approaches have direct and indirect connections with each other. Thus, in terms of professional engagement since the mid-1990s, the heterodox community is a pluralistic integrative whole.

Heterodox economics associations (currently active)

Name	Date established	Country or region of primary activity
Association for Evolutionary Economics (AFEE)	1965	United States
Association for Heterodox Economics (AHE)	1998	United Kingdom & Ireland
Association for Institutionalist Thought (AFIT)	1979	United States
Association for Social Economics (ASE)	1970	United States
Association pour le Développement des Etudes Keynesiennes	2000	France
Belgian-Dutch Association for Institutional And Political Economy	1980	The Netherlands & Belgium
Conference of Socialist Economists (CSE)	1970	United Kingdom
European Association for Evolutionary Political Economy (EAEPE)	1988	Europe
International Association for Feminist Economics (IAFFE)	1992	World
International Confederation of Associations For Pluralism in Economics (ICAPE)	1993	United States/World
Japan Association for Evolutionary Economics (JAFEE)	1996	Japan

Japan Society of Political Economy (JSPE)	1959	Japan
Korean Social and Economic Studies Association	1987	Korea
L'Association d'Economie Politique	1980	Canada
Progressive Economics Forum (PEF)	1998	Canada
Society for the Advancement of Socio-Economics (SABE)	1989	United States
Society for the Development of Austrian Economics (SDAE)	1996	United States
Society for Heterodox Economics (SHE)	2002	Australia
Union for Radical Political Economics (URPE)	1968	United States
US Society for Ecological Economics (USSEE)	2000	United States

Theoretical segregation involves the isolation of a particular theoretical approach and its adherents from all other approaches and their adherents; that is to say, theoretical segregation occurs when there is no engagement across different theoretical approaches. However, it does not exist within heterodox economics currently, nor has it existed in the past among the various heterodox approaches. From the 1960s to the 1980s heterodox economists engaged, integrated or synthesized Institutional, Post Keynesian and Marxist-radical approaches, Institutional and Post Keynesian approaches, Post Keynesian and Marxian-radical approaches, Post Keynesian and Austrian, Austrian and Institutional, feminist and Marxist-radical approaches, Institutional and Marxist-radical approaches, Institutional and social economics, ecological and Marxian-radical approaches, and social and Marxian economics. Thus by 1990 many heterodox economists could no longer see distinct boundaries between the various approaches. Moreover, from the 1990s to the present day heterodox economics has continued the past integration efforts of engaging across the various heterodox approaches. Hence, it is clear that the heterodox community is not segregated along theoretical lines, but rather there is cross-approach engagement to such an extent that the boundaries of the various approaches do not simply overlap – they are, in some cases, not there at all. The ensuing theoretical messiness of cross-approach engagement is evidence, to detractors, of the theoretical incoherence of heterodox economics, whereas to supporters of progress it is evidence of a more theoretically coherent heterodox economics – a glass half-empty of coherence as opposed to a glass half-full of coherence.

Heterodox critique of mainstream economics

Mainstream economics is a clearly defined theoretical story about how the economy works; but this story is theoretically incoherent. That is, mainstream theory is comprised of a core set of propositions – such as scarcity, equilibrium, rationality, preferences, and methodological individualism and derivative beliefs, vocabulary, symbols and parables – while there is a range of heterogeneous theoretical developments beyond the core that do not call into question the core itself in totality. As a result, critiques of the theory vary in that they can deal with the internal coherence and/or empirical grounding of the theory; they can be directed at the theory at a particular point in time or at specific components of theory (such as methodology, concepts qua vocabulary, parables qua stories and symbols); and they can

be initiated from a particular heterodox approach. What emerges is a varied but concatenation of particular and extensive critiques that generate an emergent encompassing rejection of mainstream theory, although any one particular critique may not go that far.

Although the internal critiques and critiques of models that tell theoretical stories show that the theory is incoherent, they do not by themselves differentiate mainstream from heterodox theory. This, however, can be dealt with in terms of specific critiques of the core propositions. That is, each of the heterodox approaches has produced critiques of particular core propositions of the theory, while each core proposition has been subject to more than one critique; in addition, the multiple heterodox critiques of a single proposition overlap in argumentation. To illustrate this point, consider the critiques of the concept of scarcity. The Post Keynesians argue that produced means of production within a circular production process cannot be characterized as scarce and that production is a social process, while Institutionalists reject the view that natural resources are not 'produced' or socially created to enter into the production process, and the Marxists argue that the concept is a mystification and misspecification of the economic problem – that it is not the relation of the isolated individual to given resources, but the social relationships that underpin the social provisioning process. The three critiques are complementary and integrative and generate the common conclusion that the concept of scarcity must be rejected as well as the mainstream definition of economics as the science of the non-social provisioning process analysed through the allocation of scarce resources among competing ends given unlimited asocial wants of asocial individuals. Other critiques of the core propositions exist and arrive at similar conclusions. Together, the three critiques – internal, story qua model and core propositions – form a concatenated structured heterodox critique that rejects and denies the truth and value of mainstream theory.

Heterodox economics: theory and policy

Since the intellectual roots of heterodox economics are located in traditions that emphasize the wealth of nations, accumulation, justice, social relationships in terms of class, gender, and race, full employment, and economic and social reproduction, the discipline of economics, from its perspective, is concerned, not with prediction per se, but with explaining the actual process that provides the flow of goods and services required by society to meet the needs of those who participate in its activities. That is, economics is the science of the *social* provisioning process, and this is the general research agenda of heterodox economists. The explanation involves human agency in a cultural context and social processes in historical time affecting resources, consumption patterns, production and reproduction, and the meaning (or ideology) of market, state, and non-market/state activities engaged in social provisioning. Thus heterodox economics has two interdependent parts: theory and policy. *Heterodox economic theory* is an empirically grounded theoretical explanation of the historical process of social provisioning within the context of a capitalist economy. Therefore it is concerned with explaining those factors that are part of the process of social provisioning, including the structure and use of resources, the structure and change of social wants, structure of production and the reproduction of the business enterprise, family, state, and other relevant institutions and organizations, and distribution. In addition, heterodox economists extend their theory to examining issues associated with the process of social provisioning, such as racism, gender and ideologies and myths. Because their economics involves issues of ethical values and social philosophy and the historical aspects of human existence, heterodox economists make ethically based *economic policy* recommendations to improve human dignity, that is, recommending

ameliorative and/or radical, social and economic policies to improve the social provisioning and hence well-being for all members of society and especially the disadvantaged members. To do this properly, their economic policy recommendations must be connected to heterodox theory which provides an accurate historical and theoretical picture of how the economy actually works – a picture that includes class and hierarchical domination, inequalities, and social-economic discontent.

Given the definition of economics as the science of the social provisioning process and the structure of the explanation of the process combined with the pluralistic and integrative proclivities of heterodox economists, there has emerged a number of elements that have come to constitute the provisional theoretical and methodological core of heterodox theory. Some elements are clearly associated with particular heterodox approaches, as noted by O'Hara (2002, p. 611):

> The main thing that social economists bring to the study [of heterodox economics] is an emphasis on ethics, morals and justice situated in an institutional setting. Institutionalists bring a pragmatic approach with a series of concepts of change and normative theory of progress, along with a commitment to policy. Marxists bring a set of theories of class and the economic surplus. Feminists bring a holistic account of the ongoing relationships between gender, class and ethnicity in a context of difference … And post-Keynesians contribute through an analysis of institutions set in real time, with the emphasis on effective demand, uncertainty and a monetary theory of production linked closely with policy recommendations.

However, other provisional elements, such as critical realism, non-equilibrium or historical modelling, the gendering and emotionalizing agency, the socially embedded economy, and circular and cumulative change, emerged from a synthesis of arguments that are associated only in part with particular heterodox approaches.

The core methodological elements establish the basis for constructing heterodox theory. In particular, the methodology emphasizes realism, structure, feminist and uncertain agency qua individual, history, and empirical grounding in the construction of heterodox theory, which is a historical narrative of how capitalism works. The theory qua historical narrative does not simply recount or superficially describe actual economic events, such as the exploitation of workers; it does more in that it analytically explains the internal workings of the historical economic process that, say, generates the exploitation of workers. Moreover, because of its historical nature, the narrative is not necessarily organized around the concepts of equilibrium/long period positions and tendencies towards them. Because the narrative provides an accurate picture of how capitalism actually works and changes in a circular and cumulative fashion, economists use their theory to suggest alternative paths that future economic events might take and propose relevant economic policies to deal with them. In constructing the narrative, they have at the same time created a particular social-economic-political picture of capitalism.

The core theoretical elements generate a three-component structure–organization–agency economic theory. The first component of the theory consists of three overlapping interdependencies that delineate the structure of a real capitalist economy. The first interdependency is that the production of goods and services requires goods and services to be used as inputs. Hence, with regard to production, the overall economy (which includes both market and non-market production) is represented as an input–output matrix of material goods combined with different types of labour skills to produce an array of goods and

services as outputs. Many of the outputs replace the goods and services used up in production and the rest constitute a physical surplus to be used for social provisioning, that is for consumption, private investment, government usage and exports. A second interdependency is the relation between the wages of workers, profits of enterprises, and taxes of government and expenditures on consumption, investment, and government goods as well as non-market social provisioning activities. The last interdependency consists of the overlay of the flow of funds or money accompanying the production and exchange of the goods and services. Together, these three interdependencies produce a monetary input–output structure of the economy where transactions in each market are a monetary transaction; where a change in price of a good or the method by which a good is produced in any one market will have an indirect or direct impact on the entire economy; and where the amount of private investment, government expenditure on real goods and services, and the excess of exports over imports determines the amount of market and non-market economic activity, the level of market employment and non-market labouring activities, and consumer expenditures on market and non-market goods and services. These elements of course have parallels in non-heterodox economics, but the ideas are developed differently.

The second component of heterodox theory consists of three broad categories of economic organization that are embedded in the monetary input–output structure of the economy. The first category is micro market-oriented, hence particular to a set of markets and products. It consists of the business enterprise, private and public market organizations that regulate competition in product and service markets and the organizations and institutions that regulate the wages of workers. The second is macro market-oriented and hence is spread across markets and products, or is not particular to any market or product. It includes the state and various subsidiary organizations as well as particular financial organizations, that is, those organizations that make decisions about government expenditures and taxation, and the interest rate. Finally, the third category consists of non-market organizations that promote social reproduction and include the family and state and private organizations that contribute to and support the family. The significance of organizations is that they are the social embeddedness of agency qua the individual, the third component of heterodox theory. That is, agency, which are decisions made by individuals concerning the social provisioning process and social well-being, takes place through these organizations. And because the organizations are embedded in both instrumental and ceremonial institutions, such as gender, class, ethnicity, justice, marriage, ideology, and hierarchy qua authority, agency qua the individual acting through organizations affect both positively and negatively but never optimally the social provisioning process.

Conclusion

If mainstream economics suddenly disappeared, heterodox economics would be largely unaffected. It would still include the various heterodox traditions; there would still be an integrated professional and theoretical community of heterodox economists; and its heterodox research agenda would still be directed at explaining the social provisioning process in capitalist economies and argue for economic policies that would enhance social well-being. In this regard, heterodox economics is not out to reform mainstream economics. Rather, it is an alternative to mainstream economics: an alternative in terms of explaining the social provisioning process and suggesting economic policies to promote social well-being. Since the mid-1990s the community of heterodox economics has grown, diversified and

integrated. The previously isolated are now part of a community, heterodox associations exist in countries where previously no heterodox associations had existed, and developments in heterodox theory and policy are occurring at breakneck speed. In short, heterodox economics is now an established feature on the disciplinary landscape and the progressive future of economics.

See Also

- pluralism in economics

Bibliography

Bortis, H. 1997. *Institutions, Behaviour and Economic Theory*. Cambridge: Cambridge University Press.

Dow, S.C. 2000. Prospects for the progress of heterodox economics. *Journal of the History of Economic Thought* 22, 157–70.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* 91, 481–510.

Gruchy, A.G. 1987. *The Reconstruction of Economics: An Analysis of the Fundamentals of Institutional Economics*. New York: Greenwood Press.

Keen, S. 2001. *Debunking Economics: The Naked Emperor of the Social Sciences*. New York City: St Martin's Press.

Lavoie, M. 2006. Do heterodox theories have anything in common? A Post-Keynesian point of view. *Intervention: Journal of Economics* 3, 87–112.

Lawson, T. 1997. *Economics and Reality*. London: Routledge.

Lee, F.S. 2001. Conference of socialist economists and the emergence of heterodox economics in post-war Britain. *Capital and Class* 75, 15–39.

Lee, F.S. 2002. The association for heterodox economics: past, present, and future. *Journal of Australian Political Economy* 50, 29–43.

Lee, F.S. 2004. To be a heterodox economist: the contested landscape of American economics, 1960s and 1970s. *Journal of Economic Issues* 38, 747–63.

Lee, F.S., Cohen, S., Schneider, G. and Quick, P. 2005. *Informational Directory for Heterodox Economists: Journals, Book Series, Websites, and Graduate and Undergraduate Programs*, 2nd edn.

Kansas City: Department of Economics, University of Missouri-Kansas City. Online. Available at http://l.web.umkc.edu/leefs/htnf/HeterodoxDirectory.pdf, accessed 12 January 2007.

Lee, F.S. and Keen, S. 2004. The incoherent emperor: a heterodox critique of neoclassical microeconomic theory. *Review of Social Economics* 62, 169–99.

Matthaei, J. 1984. Rethinking scarcity: neoclassicism, neoMalthusianism, and neoMarxism. *Review of Radical Political Economics* 16, 81–94.

O'Hara, P.A. (ed.) 1999. *Encyclopedia of Political Economy*. London: Routledge.

O'Hara, P.A. 2000. *Marx, Veblen, and Contemporary Institutional Political Economy: Principles and Unstable Dynamics of Capitalism*. Cheltenham: Edward Elgar.

O'Hara, P.A. 2002. The role of institutions and the current crises of capitalism: a reply to Howard Sherman and John Henry. *Review of Social Economy* 60, 609–18.

Power, M. 2004. Social provisioning as a starting point for feminist economics. *Feminist Economics* 10, 3–19.

White, G. 2004. Capital, distribution and macroeconomics: ‘core’ beliefs and theoretical foundations. *Cambridge Journal of Economics* 28, 527–47.

Wrenn, M.V. 2004. What is heterodox economics? Ph.D. thesis, Colorado State University.

How to cite this article

Lee, Frederic S. "heterodox economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000175> doi:10.1057/9780230226203.0727

The New Palgrave Dictionary of Economics Online

heteroskedasticity and autocorrelation corrections

Kenneth D. West
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Many time series studies, including in particular those estimated by generalized method of moments, involve disturbances that are serially correlated and, possibly, conditionally heteroskedastic. The serial correlation and heteroskedasticity often are of unknown form. Corrections for serial correlation and heteroskedasticity are required for inference and efficient estimation. This article surveys procedures to implement such corrections.

Keywords

autocovariances; Bartlett kernel; generalized method of moments; heteroskedasticity; heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimation; kernel weights; long-run variance; moving average processes; Newey–West estimator; quadratic spectral kernel; serial correlation; spectral density estimation; truncated estimators; vector autoregressions

Article

Heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimation refers to calculation of covariance matrices that account for conditional heteroskedasticity of regression disturbances and serial correlation of cross products of instruments and regression disturbances. The heteroskedasticity and serial correlation may be of unknown form. HAC estimation is integral to empirical research using generalized method of moments (GMM) estimation (Hansen, 1982). In this article I summarize results relating to HAC estimation, with emphasis on practical rather than theoretical aspects.

The central issue is consistent and efficient estimation of what is called a ‘long-run variance’, subject to the constraint that the estimator is positive semidefinite in finite samples. Positive semidefiniteness is desirable since the estimator will be used to compute standard errors and test statistics. To fix notation, let h_t be a $q \times 1$ stationary mean zero random vector. Let Γ_j denote the $q \times q$ autocovariance of h_t at lag q, $\Gamma_j \equiv Eh_t h_{t-j}'$; of course, $\Gamma_j = \Gamma_{-j}'$. The long run variance of h_t is the $q \times q$ matrix

$$S = \sum_{j=-\infty}^{\infty} \Gamma_j = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma_j + \Gamma_j'). \quad (1)$$

Apart from a factor of 2π, the symmetric matrix S, which I assume to be positive definite, is the spectral density of h_t at frequency zero. As discussed below, techniques for spectral density estimation are central to HAC estimation. (For an arbitrary stationary process, the sum in the right-hand side of (1) may not converge, and may not be positive definite even if it

does converge. But here and throughout I assume unstated regularity conditions. As well, I use formulas that allow for relatively simple notation, for example assuming covariance stationarity even when that assumption can be relaxed. The cited papers may be referenced for generalizations and for technical conditions.)

To illustrate how estimation of S figures into covariance matrix estimation, consider the following simple example. As in Hansen and Hodrick (1980), let us suppose that we wish to test the 'rationality' of a scalar variable x_t as an n period ahead predictor of a variable y_{t+n+1}, for $n \geq 0$: the null hypothesis is $E_t y_{t+n+1} = x_t$, where E_t denotes expectations conditional on the information set used by market participants. The variable x_t might be the expectation of y_{t+n+1} reported by a survey, or it might be a market determined forward rate. Let u_t denote the expectational error: $u_t = y_{t+n+1} - E_t y_{t+n+1} = y_{t+n+1} - x_t$. (The expectational error u_t, which is not realized until period $t+n+1$, is dated t to simplify notation.)

One can test one implication of the hypothesis that x_t is the expectation of y_{t+n+1} by regressing y_{t+n+1} on a constant and x_t, and checking whether the coefficient on the constant term is zero and that on x_t is 1:

$$y_{t+n+1} = \beta_0 + \beta_1 x_t + u_t \equiv X_t{}'\beta + u_t; \quad H_0: \beta = (0, 1)'. \tag{2}$$

Under the null, $EX_t u_t = 0$, so least squares is a consistent estimator. As well, $X_t u_t$ follows a moving average process of order n. Thus the asymptotic variance of the least squares estimator of β is $(EX_t X_t{}')^{-1} S (EX_t X_t{}')^{-1}$, where $S = \Gamma_0 + \Sigma_{j=1}^{n} (\Gamma_j + \Gamma_j{}'), \quad \Gamma_j \equiv EX_t u_t (X_{t-j} u_{t-j})'$. This example maps into the notation used in (1) with $h_t = X_t u_t, \quad q = 2$ and a known upper bound to the number of non-zero autocovariances of h_t. Clearly one needs to estimate $EX_t X_t{}'$ and S to conduct inference. A sample average of $X_t X_t{}'$ can be used to estimate $EX_t X_t{}'$. If $n = 0$, so that h_t is serially uncorrelated, $S = EX_t u_t (X_t u_t)'$ and estimation of S is equally straightforward; White's (1980) heteroskedasticity consistent estimator can be used. The subject at hand considers ways to estimate S when h_t is serially correlated. I note in passing that one cannot sidestep estimation of S by applying generalized least squares. In this example and more generally, generalized least squares is inconsistent. See Hansen and West (2002).

To discuss estimation of S, let us describe a more general set-up. In GMM estimation, h_t is a $q \times 1$ orthogonality condition used to identify a k-dimensional parameter vector β. The orthogonality condition takes the form

$$h_t = Z_t u_t \tag{3}$$

for a $q \times \ell$ matrix of instruments Z_t and an $\ell \times 1$ vector of unobservable regression disturbances u_t. The vector of regression disturbances depends on observable data through β, $u_t = u_t(\beta)$. In the example just given, $q = 2, \quad \ell = 1, \quad Z_t = X_t, \quad u_t(\beta) = y_{t+n+1} - X_t{}'\beta$. The example just given is overly simple in that the list of instruments typically will not be identical to right-hand side variables, and the model may be nonlinear. For a suitable $k \times q$ matrix D, the asymptotic variance of the GMM estimator of β takes the form DSD' (for example, $D = (EX_t X_t{}')^{-1}$ in the example just given). In an overidentified model (that is, in models in which the dimension of the orthogonality condition q is greater than the number of parameters k) the form D takes depends on a certain weighting matrix. Let $h_{t\beta}$ be the $q \times k$ matrix $\partial h_t / \partial \beta$. When the weighting matrix is chosen optimally, $D = (Eh_{t\beta}{}' S^{-1} Eh_{t\beta})^{-1} Eh_{t\beta}{}' S^{-1}$ and the asymptotic variance DSD'

simplifies to $(Eh_{t\beta}{}'S^{-1}Eh_{t\beta})^{-1}$. The optimal weighting matrix is one that converges in probability to *S*, and thus the results about to be presented are relevant to efficient estimation as well as to hypothesis testing. In any event, the matrix $Eh_{t\beta}$ typically is straightforward to estimate; the question is how to estimate *S*. This will be the focus of the remainder of the discussion.

We have sample of size *T* and sample counterparts to u_t and h_t, call them $\hat{u}_t = u_t(\hat{\beta})$ and $\hat{h}_t = h_t(\hat{\beta})$. Here, $\hat{\beta}$ is a consistent estimate of β. In the least squares example given above, $\hat{u}_t$ is the least squares residual, $\hat{u}_t = y_{t+n+1} - X_t{}'\hat{\beta}$, and $\hat{h}_t = X_t\hat{u}_t = X_t(y_{t+n-1} - X_t{}'\hat{\beta})$. One path to consistent estimation of *S* involves consistent estimation of the autocovariances of h_t. The natural estimator is a sample average,

$$\hat{\Gamma}_j = T^{-1}\sum_{t=j+1}^{T}\hat{h}_t\hat{h}_{t-j}{}' \quad \text{for} \quad j \geq 0. \tag{4}$$

For given *j*, (4) is a consistent ($T\rightarrow\infty$) estimator of Γ_j.

I now discuss in turn several possible estimators, or classes of estimators, of *S*: (1) the truncated estimator; (2) estimators applicable only when h_t follows a moving average (MA) process of known order; (3) an autoregressive spectral estimator; (4) estimators that smooth autocovariances; (5) some recent work, on estimators that might be described as extensions or modifications of ones the estimators described in (4).

1 The truncated estimator

Suppose first that it is known a priori that the autocovariances of h_t are zero after lag *n*, as is the case in the empirical example above. A natural estimator of *S* is one that replaces population objects in (1) with sample analogues. This is the *truncated* estimator:

$$\hat{S}_{TR} = \hat{\Gamma}_0 + \sum_{j=1}^{n}(\hat{\Gamma}_j + \hat{\Gamma}_j{}'). \tag{5}$$

In the more general case in which $\Gamma_j \neq 0$ for all *j*, the truncated estimator is consistent if the truncation point $n\rightarrow\infty$ at a suitable rate. Depending on exact technical conditions, the rate may be $n/T^{1/2}\rightarrow 0$ or $n/T^{1/4}\rightarrow 0$ (Newey and West, 1987). The truncated estimator need not, however, yield a positive semidefinite estimate. With certain plausible data generating processes, simulations indicate that it will not be p.s.d. in a large fraction of samples (West, 1997). Hence this estimator is not used much in practice.

2 Estimators applicable only when h_t follows an MA process of known order n

Such a process for h_t holds in studies of rationality (as illustrated above) and in the first order conditions from many rational expectations models (for example, Hansen and Singleton, 1982).

Write the Wold representation of h_t as $h_t = e_t + \Theta_1 e_{t-1} + \cdots + \Theta_n e_{t-n}$. Here, e_t is the $q\times 1$ innovation in h_t. Let Ω denote the $q\times q$ variance covariance matrix of e_t. Then it is well known (for example, Hamilton, 1994, p. 276) that

$$S = (I + \Theta_1 + \cdots + \Theta_n)\Omega(I + \Theta_1 + \cdots + \Theta_n)' \tag{6}$$

Suppose that one fits an MA(n) process to $\hat{h}_t$, and plugs the resulting estimates of the Θ_i and Ω into the formula for S. Clearly the resulting estimator is $T^{1/2}$ consistent and positive semidefinite. Nevertheless, to my knowledge this estimator has not been used, presumably because of numerical difficulties in estimating multivariate moving average processes.

Two related estimators have been proposed that impose a smaller computational burden. Hodrick (1992) and West (1997) suggest an estimator that requires fitting an MA(n) to the vector of regression residuals $\hat{u}_t$ (or, in Hodrick's, 1992, application, using MA coefficients that are known a priori). The computational burden of such MA estimation will typically be considerably less than that of MA estimation of the h_t process, because the dimension of u_t is usually much smaller than that of h_t. For example, $\hat{u}_t$ will be a scalar in a single equation application, regardless of the number of orthogonality conditions captured in h_t. Write the estimated MA process for $\hat{u}_t$ as $\hat{u}_t = \widehat{\in}_t + \widehat{\Psi}_1\widehat{\in}_{t-1} + \cdots + \widehat{\Psi}_n\widehat{\in}_{t-n}$, where the $\widehat{\Psi}_j$ are $\ell \times \ell$. (Note that $\in_t$, the $\ell \times 1$ innovation in u_t, is not the same as e_t, the q×1 innovation in h_t.) Then a $T^{1/2}$ consistent and positive semidefinite estimator of S is

$$\hat{S}_{MA-\ell} = T^{-1}\sum_{t=1}^{T-n}\hat{d}_{t+n}\hat{d}_{t+n}', \quad \hat{d}_{t+n} = (Z_t + Z_{t+1}\widehat{\Psi}_1 + \cdots + Z_{t+n}\widehat{\Psi}_n)\widehat{\in}_t, \tag{7}$$

where, again, Z_t is the $q \times \ell$ matrix of instruments (see eq. (3)).

Eichenbaum, Hansen and Singleton (1988) and Cumby, Huizanga and Obstfeld (1983) propose a different strategy that avoids the need to estimate a moving average process for either u_t or h_t. They suggest estimating the parameters of $\hat{h}_t$'s autoregressive representation, and inverting the autoregressive weights to obtain moving average weights. Call the results $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_n$, with $\widehat{\Omega}$ the estimate of the innovation variance–covariance matrix. The resulting estimator $\hat{S} = (I + \widehat{\Theta}_1 + \cdots + \widehat{\Theta}_n)\widehat{\Omega}(I + \widehat{\Theta}_1 + \cdots + \widehat{\Theta}_n)'$ is positive semidefinite by construction. The rate at which it converges to S depends on the rate at which the order of the autoregression is increased.

3 Autoregressive estimators

Den Haan and Levin (1997) propose and evaluate an autoregressive spectral estimator. Suppose that h_t follows a (possibly) infinite-order vector autoregression (VAR)

$$h_t = \sum_{j=1}^{\infty}\Phi_j h_{t-j} + e_t, \quad Ee_te_t' = \Omega. \tag{8}$$

Then (Hamilton, 1994, p. 237)

$$S = \left(I - \sum_{j=1}^{\infty} \Phi_j\right)^{-1} \Omega \left(I - \sum_{j=1}^{\infty} \Phi_j\right)^{-1\prime}.$$
(9)

The idea is to approximate this quantity via estimates from a finite-order VAR in $\hat{h}_t$. Write the estimate of a VAR in $\hat{h}_t$ of order *p* as

$$\hat{h}_t = \widehat{\Phi}_1 \hat{h}_{t-1} + \cdots + \widehat{\Phi}_p \hat{h}_{t-p} + \hat{e}_t, \ \widehat{\Omega} = T^{-1} \sum_{t=p+1}^{T} \hat{e}_t \hat{e}_t{}'.$$
(10)

Then the estimator of $\hat{S}$ is

$$\hat{S}_{AR} = \left(I - \sum_{j=1}^{p} \widehat{\Phi}_j\right)^{-1} \widehat{\Omega} \left(I - \sum_{j=1}^{p} \widehat{\Phi}_j\right)^{-1\prime}.$$
(11)

Den Haan and Levin (1997, Section 3.5) conclude that if *p* is chosen by BIC, and some other technical conditions hold, then this estimator converges at a rate very near $T^{1/2}$ (the exact rate depends on certain characteristics of the data). A possible problem in practice with this estimator (as well as with the estimator described in the final paragraph of Section 2, which also requires estimates of a VAR in $\hat{h}_t$) is that it may require estimation of many parameters and inversion of a large matrix. Den Haan and Levin therefore suggest judiciously parametrizing the autoregressive process, for example by using the BIC criterion equation-by-equation for each of the *q* elements of $\hat{h}_t$.

4 Estimators that smooth autocovariances

In practice, the most widely used class of estimators is one that relies on smoothing of autocovariances. Andrews (1991), building on the literature on estimation of spectral densities, established a general framework for analysis. Andrews considers estimators that can be written

$$\hat{S} = \widehat{\Gamma}_0 + \sum_{j=1}^{T-1} k_j (\widehat{\Gamma}_j + \widehat{\Gamma}_j{}')$$
(12)

for a series of kernel weights $\{k_j\}$ that obey certain properties. For example, to obtain a consistent estimator, we need k_j near

zero (or perhaps identically zero) for values of j near $T-1$, since autocovariances at large lags are estimated imprecisely, while $k_j \to 1$ for each j is desirable for consistency. We would also like the choice of k_j to ensure positive definiteness.
The two most commonly used formulas for the kernel weights are:

$$\text{Bartlett: for some } m \geq 0: k_j = 1 - [j/(m+1)] \text{ for } j \leq m, k_j = 0 \text{ for } j > m. \tag{13a}$$

$$\text{Quadratic spectral (QS): for some } m > 0, \text{ and with } x_j = j/m:$$
$$k_j = [25/12\pi^2 x_j^2] \times \left\{[\sin(6\pi x_j/5)/(6\pi x_j/5)] - \cos(6\pi x_j/5)\right\}. \tag{13b}$$

If we let $z_j = 6\pi x_j/5$, the QS formula for k_j can be written in more compact form as $(3/z_j^2)\left\{[\sin(z_j)/z_j] - \cos(z_j)\right\}$. Call the resulting estimators $\hat{S}_{BT}$ and $\hat{S}_{QS}$. For example,

$$\hat{S}_{BT} = \hat{\Gamma}_0 + \sum_{j=1}^{m} [1 - j/(m+1)](\hat{\Gamma}_j + \hat{\Gamma}_j{}'). \tag{14}$$

The vast literature on spectral density estimation suggests many other possible kernel weights. For conciseness, I consider only the Bartlett and QS kernels.
To operationalize these estimators, one needs to choose the lag truncation parameter or bandwidth m. I note that for both kernels, consistency requires $m \to \infty$ as $T \to \infty$, even if h_t follows an MA process of known finite order, as in the example given above. Thus one should not set m to be the number of non-zero autocovariances. Subject to possible problems with positive definiteness, setting $m=n$ is fine for the truncated estimator (5) but not for estimators that use nontrivial weights $\{k_j\}$.
Andrews shows that maximizing the rate at which $\hat{S}$ converges to S requires that m increase as a suitable function of sample size, with the 'suitable function' varying with kernel. For the Bartlett and QS, the maximal rates of convergence are realized when

$$\text{Bartlett: } m = \gamma T^{1/3} (\text{or } m = (\text{integer part of } \gamma T^{1/3})) \text{ for some } \gamma \neq 0, \text{QS: } m = \gamma T^{1/5} \text{ for some } \gamma \neq 0, \tag{15}$$

in which case $\hat{S}_{BT}$ converges to S at rate $T^{1/3}$ and the mean squared error in estimation of S goes to zero at rate $T^{2/3}$; the comparable figures for QS are $T^{2/5}$ and $T^{4/5}$. Since both estimators are nonparametric, they converge at rates slower than $T^{1/2}$; since faster convergence is better, the QS rate is preferable to that of the Bartlett. Indeed, Andrews (1991), drawing on

Priestley (1981), shows that for a certain class of kernel weights $\{k_j\}$, the mean squared error of QS rate is optimal in the following sense: a $T^{4/5}$ rate on the asymptotic mean squared error is the fastest that can be achieved if one wants to ensure a positive definite $\hat{S}$, and within the class of kernels that achieve the $T^{4/5}$ rate, the QS has the smallest possible asymptotic mean squared error.

As a practical matter, the formulas in (15) have merely pushed the question of choice of m to one of choice of γ; putting arbitrary γ in (15) yields convergence that is as fast as possible, but different choices of γ lead to different asymptotic mean squared errors. The choice of γ that is optimal from the point of view of asymptotic mean squared error is a function of the data (Hannan, 1970, p. 286). Let $S^{(0)} = \Sigma_{j=-\infty}^{\infty}\Omega_j (= S)$; $S^{(1)} = \Sigma_{j=-\infty}^{\infty}|j|\Omega_j$; $S^{(2)} = \Sigma_{j=-\infty}^{\infty} j^2\Omega_j$. For scalar ($q$=1) S optimal choices are:

$$\text{Bartlett}: \gamma = 1.1447[S^{(1)}/S^{(0)}]^{2/3}; \quad \text{QS}: \gamma = 1.3221[S^{(2)}/S^{(0)}]^{2/5}. \tag{16}$$

(See Andrews, 1991, for the derivation of these formulas.)

Andrews (1991), Andrews and Monahan (1992) and Newey and West (1994) proposed feasible data dependent to procedures to estimate γ, for vector as well as scalar h_t. Rather than exposit the general case, I will describe two ‘cookbook’ procedures that have been offered as reasonable starting points in empirical work. One procedure relies on Andrews (1991) and Andrews and Monahan (1992), and assumes the QS kernel and estimation of γ via parametric models. The second relies on Newey and West (1994), and assumes a Bartlett kernel and nonparametric estimation of γ. I emphasize that both papers present more general results than are presented here; both allow the researcher to (for example) use any one of a wide range of kernels.

Let there be a $q\times1$ vector of weights $w=(w_1, w_2, \ldots, w_q)'$ whose elements tells us how to weight the various elements of S with respect to mean squared error. The weights might be sample dependent, and den Haan and Levin (1997) argue that there are benefits to certain sample-dependent weights, but a simple choice proposed by both papers is: w_i=0 if the corresponding element of h_t is a cross product of a constant term and a regression disturbance, otherwise w_i=1. Andrews’s loss function is the normalized expectation of $\Sigma_{i=1}^{q} w_i(S_{ii} - \hat{S}_{ii})^2$, while Newey and West’s loss function is the normalized expectation of $[w'(\hat{S} - S)w]^2$; the normalization is $T^{4/5}$ for QS and $T^{2/3}$ for Bartlett.

Both procedures begin with using a vector autoregression to prewhiten, and end with re-colouring. The basic justification for prewhitening and re-colouring is that simulation evidence indicates that this improves finite sample performance.

- 1. Prewhitening: Estimate a vector autoregression in $\hat{h}_t$, most likely of order 1. Call the residuals $\hat{h}_t^{\dagger}$

$$\hat{h}_t = \hat{A}\hat{h}_{t-1} + \hat{h}_t^{\dagger}, \quad \hat{A} = \sum_{t=2}^{T}\hat{h}_t\hat{h}_{t-1}'\left(\sum_{t=2}^{T}\hat{h}_{t-1}\hat{h}_{t-1}'\right)^{-1}. \tag{17}$$

- 2. Let $\hat{\Gamma}_j^{\dagger}$ denote the jth autocovariance of the VAR residual $\hat{h}_t^{\dagger}$, $\hat{\Gamma}_j^{\dagger} = (T-1)^{-1}\Sigma_{t=2+j}^{T}\hat{h}_t^{\dagger}\hat{h}_{t-j}^{\dagger\prime}$. Using $\{\hat{\Gamma}_j^{\dagger}\}$ (rather than $\{\hat{\Gamma}_j\}$ [the autocovariances of $\hat{h}_t$]), and choosing m optimally as described in steps 2a or 2b below,

construct an estimate of the long run variance of the residual of the VAR just estimated. Call the result $\hat{S}^{\dagger}$.

- 2a. Andrews and Monahan (1992): Fit a univariate AR(1) to each of the q elements of $\hat{h}_t^{\dagger}$. Call the resulting estimate of the AR coefficient and variance of the residual $\hat{\rho}_i$ and $\hat{\sigma}_i^2$. Compute

$$\hat{s}_2 = \sum_{i=1}^{q} w_i(4\hat{\rho}_i^2\hat{\sigma}_i^4) / (1-\hat{\rho}_i)^8, \quad \hat{s}_0 = \sum_{i=1}^{q} w_i\hat{\sigma}_i^4 / (1-\hat{\rho}_i)^4, \hat{\gamma}_{QS} = 1.3221[\hat{s}_2/\hat{s}_0]^{1/5}, \quad \hat{m}_{QS} = \hat{\gamma}_{QS}T^{1/5}. \tag{18}$$

Then plug $\hat{m}_{QS}$ into formula (13b). Call the result $\hat{k}_j$. Compute $\hat{S}^{\dagger} = \hat{\Gamma}_0^{\dagger} + \Sigma_{j=1}^{T-1}\hat{k}_j(\hat{\Gamma}_j^{\dagger} + \hat{\Gamma}_j^{\dagger\prime})$.

- 2b. Newey and West (1994): Set n=integer part of $12(T/100)^{2/9}$. Compute

$$\hat{s}^{(1)} = w'\hat{\Gamma}_0 w + 2\sum_{i=1}^{n} iw'\Gamma_i^{\dagger} w, \hat{s}^{(0)} = 2\sum_{i=1}^{n} w'\hat{\Gamma}_i^{\dagger} w, \hat{\gamma}_{BT} = 1.1447[\hat{s}^{(1)}/\hat{s}^{(0)}]^{2/3}, \hat{m}_{BT} = \text{integer part of } \hat{\gamma}_{BT}T^{1/3}. \tag{19}$$

Then compute $\hat{S}^{\dagger}$ according to (14), using $\hat{m}_{BT}$.

- 3. Re-colouring: compute $\hat{S} = (I-\hat{A})^{-1}\hat{S}^{\dagger}(I-\hat{A})^{-1\prime}$.

These two recipes for estimates of S can serve as a starting point for experimentation for alternative choices of m and alternative kernels.

What is the simulation evidence on behaviour of these and other proposed estimators? In answering this question, I focus on sizing of test statistics and accuracy of confidence interval coverage: accuracy in estimation of S is desirable mainly insofar as it leads to accuracy of inference using the relevant variance–covariance matrix. The simulations in papers cited in this article suggest the following. First, no one estimator dominates others. This means in particular that the rate of convergence is not a sufficient statistic for performance in finite samples. The truncated estimator often and the autoregressive estimator sometimes perform more poorly than the slower converging QS estimator, which in turn sometimes performs more poorly than the still slower converging Bartlett estimator. Second, given that one decides to use QS or Bartlett, performance generally though not always is improved if one prewhitens and uses a data-dependent bandwidth as described in the recipes above. Third, the QS and Bartlett estimators tend to reject too much in the presence of positive serial correlation in h_t, and have what I read as a DGP dependent rejection rate (sometimes over-reject, sometimes under-reject) in the presence of negative serial correlation in h_t. The truncated estimator is much likelier to fail to be positive semidefinite in the presence of negative than positive serial correlation. Finally, the performance of all estimators leaves much to be desired. Plausible data-generating processes and sample sizes can lead to serious mis-sizing of any given estimator. Nominal 0.05 tests can have empirical size as low as 0.01 and higher than 0.25.

5 Some recent work

Because simulation studies have yielded disappointing performance, ongoing research aims to develop better estimators. I close by summarizing a few of many recently published papers.

1. I motivated my topic by observing that consistent estimation of S is a natural element of consistent estimation of the variance–covariance matrix of a GMM estimator. Typically we estimate the variance–covariance matrix because we wish to

construct confidence intervals or conduct hypothesis tests. A recent literature has evaluated inconsistent estimators that lead to well-defined test statistics, albeit statistics with non-standard critical values. These estimators set lag truncation (or bandwidth) equal to sample size. For example, for the Bartlett estimator, these estimators set $m=T-1$ (see Kiefer, Vogelsang and Bunzel, 2000; Kiefer and Vogelsang, 2002). Simulation evidence indicates that the non-standard statistics may be better behaved than standard statistics. Jansson (2004) provides a theoretical rationale for improved performance in a special case, with more general results in Kiefer and Vogelsang (2005). Phillips, Sun and Jin (2006; 2007) propose a related approach, which under some assumptions will yield statistics with standard critical values.

2. Politis and Romano (1995) propose what they call a ‘trapezoidal’ kernel. A trapezoidal kernel is a combination of the truncated and Bartlett kernels. For given truncation lag m, let $x_j=j/(m+1)$. Then for some c, $0<c<1$, the trapezoidal weights satisfy: $k_j=1$ if $0 \le x_j \le c$, $k_j=(x_j-1)/(c-1)$ for $c < x_j \le 1$. Thus for $0 \le j \le c(m+1)$, the autocovariances receive equal weight, as in the truncated kernel; for $c(m+1) < j \le m+1$, the weights on the autocovariances decline linearly to zero, as in the Bartlett kernel. Such kernels have the advantage that, like the truncated kernel, their convergence is rapid (near $T^{1/2}$). They share with the truncated kernel the possibility of not being positive semidefinite. The authors argue, however, that these kernels are better behaved in finite samples than is the truncated kernel.

3. Xiao and Linton (2002) propose ‘twicing’ kernels. Operationally, one first computes an estimate such as one of those described in Section 4. One also constructs a multiplicative bias correction by smoothing periodogram ordinates via a ‘twiced’ kernel. For a properly chosen bandwidth and kernel, the mean squared error of the estimator is of order $T^{8/9}$ (versus $T^{4/5}$ for the QS and $T^{2/3}$ for the Bartlett, absent any corrections). As well, Hirukawa’s (2006) version of the Xiao and Linton estimator is positive semidefinite by construction. (The rate results for this estimator and that described in the previous paragraph do not contradict Andrews’s, 1991, optimality result for the QS kernel, because these procedures fall outside the class considered by Andrews.)

See Also

- rational expectations models, estimation of
- Euler equations
- generalized method of moments estimation
- spectral analysis
- time series analysis

I thank Steven Durlauf, Masayuki Hirukawa and Tim Vogelsgang for helpful comments, and the National Science Foundation for financial support.

Bibliography

Andrews, D.W.K. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–58.

Andrews, D.W.K. and Monahan, J.C. 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 60, 953–66.

Cumby, R.E., Huizanga, J. and Obstfeld, M. 1983. Two step, two-stage least squares estimation in models with rational expectations. *Journal of Econometrics* 21, 333–55.

den Haan, W.J. and Levin, A.T. 1997. A practitioner’s guide to robust covariance matrix estimation. In *Handbook of Statistics: Robust Inference*, vol. 15, ed. G. Maddala and C. Rao. New York: Elsevier.

Eichenbaum, M.S., Hansen, L.P. and Singleton, K.J. 1988. A time series analysis of representative agent models of

consumption and leisure choice under uncertainty. *Quarterly Journal of Economics* 103, 51–78.

Hamilton, J. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Hannan, E.J. 1970. *Multiple Time Series*. New York: Wiley.

Hansen, B.E. and West, K.D. 2002. Generalized method of moments and macroeconomics. *Journal of Business and Economic Statistics* 20, 460–9.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Hansen, L.P. and Hodrick, R.J. 1980. Forward exchange rates as optimal predictors of future spot rates: an econometric analysis. *Journal of Political Economy* 96, 829–53.

Hansen, L.P. and Singleton, K.J. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–86.

Hirukawa, M. 2006. A modified nonparametric prewhitened covariance estimator. *Journal of Time Series Analysis* 27, 441–76.

Hodrick, R.J. 1992. Dividend yields and expected stock returns: alternative procedures for inference and measurement. *Review of Financial Studies* 5, 357–86.

Jansson, M. 2004. The error in rejection probability of simple autocorrelation robust tests. *Econometrica* 72, 937–46.

Kiefer, N.M. and Vogelsang, T.J. 2002. Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica* 70, 2093–5.

Kiefer, N.M. and Vogelsang, T.J. 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory* 21, 1130–64.

Kiefer, N.M., Vogelsang, T.J. and Bunzel, H. 2000. Simple robust testing of regression hypotheses. *Econometrica* 68, 695–714.

Newey, W.K. and West, K.D. 1987. A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–08.

Newey, W.K. and West, K.D. 1994. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies* 61, 631–54.

Phillips, P.C.B., Sun, Y. and Jin, S. 2006. Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation. *International Economic Review* 47, 837–94.

Phillips, P.C.B., Sun, Y. and Jin, S. 2007. Long run variance estimation and robust regression testing using sharp origin kernels with no truncation. *Journal of Statistical Planning and Inference* 137, 985–1023.

Politis, D.N. and Romano, J.P. 1995. Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis* 16, 67–103.

Priestley, M.B. 1981. *Spectral Analysis and Time Series*. New York: Academic Press.

West, K.D. 1997. Another heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Journal of Econometrics* 76, 171–91.

White, H. 1980. A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–38.

Xiao, Z. and Linton, O. 2002. A nonparametric prewhitened covariance estimator. *Journal of Time Series Analysis* 23, 215–50.

How to cite this article

West, Kenneth D. "heteroskedasticity and autocorrelation corrections." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000182> doi:10.1057/9780230226203.0728

The New Palgrave Dictionary of Economics Online

Hicks, John Richard (1904 – 1989)

Christopher Bliss

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This biographical review of the life and works of John Hicks covers his contributions to numerous fields, and in each case assesses the particular contributions for which he was responsible. The fields concerned are The Theory of Wages, Value Theory, Welfare Economics, The Keynesian Revolution, Monetary Theory, Growth and Capital Theory, and Other Topics. An extensive bibliography of Hicks's writings is provided. Two points that are stressed are the unusual departure point for Hicks's thought in the general equilibrium ideas of European economists, and the radical effect on Hicks of Keynes's ideas.

Keywords

Austrian economics; capital measurement; cardinal utility; causality in economics; Chamberlin, E. H.; compensation test; consumer surplus; consumption function; economic dynamics; economic history; elasticity of expectations; elasticity of substitution; excess demand; firm, theory of the; fixprice versus flexiprice economy; general equilibrium; growth theory; Hicks, J. R.; Hicks, U.K.; Hicksian stability condition; Hicksian substitution effect; income measurement; indirect taxation; induction; inflation; intertemporal equilibrium; IS–LM model; Kaldor, N.; Keynes, J. M.; marginal and average cost pricing; marginal efficiency of capital; methodology of econometrics; monetary theory; multiplier–accelerator interaction; neoclassical synthesis; new welfare economics; nominal income; ordinal utility; price theory; probability; public expenditure; Robinson, J. V.; Samuelson, P. A.; Scitovsky, T.; simultaneous differential equation system; Slutsky, E.; social accounting; social income; social welfare function; temporary equilibrium; traverse; underconsumptionism; value judgements; value theory; von Neumann model of capital accumulation causality; wage flexibility; wages theory; waiting; welfare economics.

Article

1 Biography and intellectual development

Hicks was born in Warwick. He studied at Oxford (1922–6) and taught at the London School of Economics (1926–35). He was Professor at Manchester University (1935–46), from where he moved to Oxford, first as Fellow of Nuffield College, and from 1952 until he retired, from teaching but not from writing, as Drummond Professor of Political Economy and Fellow of All Souls College. In 1935 he married Ursula Webb, a distinguished public finance specialist, and he collaborated with her in the preparation of numerous works on public finance, its theory and its application to various countries. Ursula Hicks, as she was subsequently known, died in 1985. John Hicks was a member of the Royal Commission on the Taxation of Profits and Income in 1951. He became a Fellow of the British Academy in 1942, a Knight in 1964, and was awarded the Nobel Prize in Economics (jointly with Kenneth J. Arrow) in 1972. He died in 1989.

Hicks was the product of a generation which was the last to produce in abundance all round economic theorists – economists who could turn their minds to almost any theoretical problem. Its leading lights, among whom Hicks is certainly to be counted, left their marks on most of the major new branches and issues of economics as these in turn attracted the interest of themselves and their contemporaries. Hicks's powerful and original mind first made itself felt in what is now called microeconomics, particularly in *The Theory of Wages* (1932, 2nd edition 1963) and with R.G.D. Allen, ‘A Reconsideration of the Theory of Value’ (*Economica*, 1934) and in welfare economics. However his best-known work, *Value and Capital* (1939), goes beyond microeconomics to offer an economic dynamics and discussion of monetary theory which reaches into the new macroeconomics.

Before Keynes's *General Theory* fundamentally altered the way in which economists viewed their subject, the theory of value, including the theory of the firm, shared the field with monetary theory. Hicks was first a value theorist, but he never neglected monetary theory, and it was an area to which he was frequently to return. It was a value theorist with an interest in monetary economics who provided in ‘Mr Keynes and the “Classics”•’ (*Econometrica*, 1937) an exposition of Keynes's *General Theory* that was probably more directly influential than the original. There followed work on the trade cycle, *A Contribution to the Theory of the Trade Cycle* (1950); on growth, *Capital and Growth* (1965); and an unusual approach to capital theory, *Capital and Time: A Neo-Austrian Theory* (1973).

Each decade of Hicks's life seemed to find him more eclectic and innovative than the last. Indeed, his willingness to speculate about and write on areas in which he had not seeped himself as a specialist was a notable feature of his later writing. Striking examples are *A Theory of Economic History* (1969), in which Hicks undertook the risks inherent in proposing a grand theory of economic history, and *Causality in Economics* (1979), in which he entered ground normally reserved for philosophers and statisticians. These works can be criticized, but as their author always commands a well-provisioned base camp in the economics which is his own, they are never merely amateurish. Hicks is an economist of outstanding breadth and erudition.

With hindsight it is remarkable that the author of such a formidable theoretical corpus should write (‘Commentary’ in the 1963 edition of *The Theory of Wages*, p. 306): ‘… at first I regarded myself as a labour economist, not a theoretical economist at all’. Lionel Robbins is given the credit for interesting Hicks in theory: ‘… he moved me from Cassel to Walras and Pareto, to Edgeworth and Taussig to Wicksell and the Austrians – with all of whom I was more at home at that stage than I was with Marshall and Pigou’ (p. 306). It would be foolish to attempt to explain why Hicks became the distinctive economist that he was to become. However the above snatches of autobiography probably go some way to explaining why *Value and Capital* turned out to be a book like no other that an English economist had

written before.

Hicks's huge output (for the papers see the three-volume *Collected Essays on Economic Theory*, 1981–3) is all the more remarkable when one considers that he seldom simply reacted to the work of others. There are no papers by Hicks pointing out mistakes by other writers, and none which embody minor changes to or extensions of existing models. Naturally Hicks produced work which follows paths opened up by others. However when he did so, as in *A Revision of Demand Theory* (1956), or with the famous IS–LM model, his approach was so distinctive that the commentary is recognizably a contribution of Hicks. Other writers feature mainly in footnotes and even such a powerful contribution as Samuelson's treatment of Walrasian stability earns no more than two pages in the Second Edition of *Value and Capital*. There is a streak of self-centredness and parochialism in Hicks which mirrors that to be found in other English economists of his generation and those before. It would be insufferable in an economist less gifted and genuinely self-critical.

2 The theory of wages

Writing later (1963) of the first edition of *The Theory of Wages* its author remarks that '… there has been no date this century to which the theory that I was putting out could have been more inappropriate.' However, Hicks was careful not to attribute the shortcomings of his first book to the misfortune of publishing in the worst year of the depression and a few years ahead of the reassessment of the theory of the firm brought about by the writings of Chamberlin and Joan Robinson and, worse fortune still, ahead of the *General Theory*. In this he was right. *The Theory of Wages* set out to examine the determination of wages under supply and demand in a competitive market. This admittedly limited task is important, and had it been perfectly accomplished it would not be sensible to criticize the resulting work for not solving other problems, such as wages under imperfect competition or the consequences of nominal wage bargaining, weighty though those problems might be. However the truth is that there were shortcomings in Hicks's treatment even given its chosen emphasis. It was not as good a book as Hicks was later to show that he could write, though it was surely a better book than the later Hicks's embarrassment at its shortcomings allowed him to admit.

G.F. Shove (whose fairly hostile review Hicks reprinted in the Second Edition) identified a number of the shortcomings. Notable among these is the relatively weak treatment of the supply side of labour markets and the consequently limited ability to treat unemployment. Shove also seems to accuse Hicks of failing to provide a treatment of the general equilibrium of many labour markets, which must be counted a rather common failing among labour economists. Shove, not surprisingly, was clear on minimum cost and the adding-up problem where Hicks's account needed improvement – it was after all Shove's bread and butter at the time. A point which Shove missed is that Hicks always discussed differences in the productivity of different workers as equivalent to differences in the quantity of effective labour provided per hour of work. In other words, like Marx before him, he fudged the problem of aggregating different types of labour.

These legitimate criticisms apart, there were very considerable merits. By concentrating on the long-run determinants of wage rates Hicks was able to examine some of the most interesting influences at work. He saw changes in the demand for labour as consisting of two components quite analogous to the income and substitution effects in demand that he was to investigate later. A lower wage rate leads to an expansion of output, because the cost curve has fallen, which induces a higher demand for labour. In

addition a lower wage rate induces the adoption of more labour intensive methods of production, which increases the demand for labour for a given output. The analysis of this last effect lead to the discovery of the new concept of the elasticity of substitution, not quite as neat in Hicks's formulation as in Joan Robinson's later presentation, but this was the original. In general, Hicks's definition of the elasticity of substitution is different from Joan Robinson's, but the two are equivalent in the two-factor case. Many topics discussed only briefly and not deeply analysed were far ahead of their time. There is the idea that because capital tends to accumulate faster than labour, technical progress tends to be labour saving – the induced bias of technical progress as we would now say. There is the first ever attempt to model a labour dispute which may culminate in a strike, and more besides.

In a passing discussion in *The Theory of Wages* its author records a fascinating fact. Many wage rates in inter-War Britain were tied to the value of the output concerned, and for that reason were automatically flexible. Once account is taken of such arrangements, the remaining pure flexibility of money wages is exceedingly small. This provided an opportunity, not taken, to bring Hicks's analysis to bear on an event that must have impressed itself on the young Oxford undergraduate: the 1926•miners' strike that lead to the failed General Strike. Britain restored Sterling convertibility in 1925 at the pre-war rate of $4 to the pound. The resulting over-valuation of Sterling made much British industrial activity internationally uncompetitive. At the time the world price of coal in dollars had fallen sharply, with the consequence that British coal was worth less in dollars, and even less in over-valued Sterling. The coal miners' contracts required sharp cuts in their wages, for which reason they went on strike. Tying miners' wages to the price of coal implied too much wage flexibility in these circumstances. Britain's coal-mining sector needed to contract, which should have raised the marginal product of labour in terms of coal, where the existing contracts held that number constant.

3 Value theory

This area and welfare economics are fields to which Hicks contributed the writings that would have made him a great economist if he had done nothing else. In making the 1972 Nobel Prize award to Hicks jointly with Arrow the Committee mentioned 'general equilibrium and welfare economics'. The reference in Hicks's case was clearly to *Value and Capital* on the one hand, and to the various papers which established the Kaldor–Hicks criterion in welfare economics on the other.

Hicks's paper with R.G.D. Allen, 'A Reconsideration of the Theory of Value' (1934) was written when both authors were at the London School of Economics, but its pedigree goes back to Slutsky, who had discovered the income and substitution effects in demand as early as 1915. However Slutsky's work was almost entirely unknown to economists in the West, and this included, as Hicks informs us, himself and Allen ('… I never saw Slutsky's work until my own was very far advanced, and some time after the substance of these chapters had been published in *Economica* by R.G.D, Allen and myself' (1939, p. 19).

Value and Capital is a work so rich in ideas that a short account of it cannot hope to do it justice. It showed that the basic results of consumer theory could be obtained from ordinal utility; it expounded what became known as the 'Hicksian substitution effect', obtained by varying income as relative prices changed so as to maintain an index of utility constant; it developed the parallel results for production theory; and it popularized among English speaking economists the notion of a general equilibrium of markets. Unlike Arrow, his fellow Nobel laureate, Hicks did not take the existence argument beyond equation and variable counting. There was about the Walrasian approach, Hicks concluded, '… a certain

sterility' (1939, p. 60). The way to overcome this was to consider the 'laws of change' of a general equilibrium system. This lead Hicks to the first ever attempt to analyse the stability of a system of multiple exchange.

It is fascinating that both Hicks and Samuelson, working entirely independently, both came up with the idea that dynamics might rescue general equilibrium theory from emptiness. Paul Samuelson in various papers of the 1940s and in his *Foundations of Economic Analysis* (1947) adopted an entirely different approach from that of Hicks. Consider a system of M markets with prices $p_1, p_2, \ldots, p_M$ and excess demands for the goods $X_1, X_2, \ldots, X_M$. Making the dependence of excess demands on all prices explicit, this system can be written as:

$$\begin{aligned} X_1(p_1, p_2, \ldots\ldots\ldots, p_M) &= 0 \\ X_2(p_1, p_2, \ldots\ldots\ldots, p_M) &= 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X_M(p_1, p_2, \ldots\ldots\ldots, p_M) &= 0 \end{aligned} \tag{3.1}$$

In equilibrium prices are such that all excess demands are zero. Now consider one good, which may be taken without loss of generality to be good 1. Select any value for p_1 and suppose that there are unique values of the remaining prices such that the excess demands for goods 2 to M are zero. If the excess demands for the other goods are always maintained at zero by changes in their prices, all other prices become implicit functions of p_1. The Hicks stability condition is then the one that would be required of a single market – X_1 should decrease with p_1. Full stability requires that this condition should be satisfied for each good in turn.

At first sight the condition appears to be asymmetrical but as the condition must be satisfied by all goods, there is no genuine asymmetry involved. However each test does involve a certain kind of asymmetry, and this is what Samuelson objected to. When we look at good 1 we implicitly assume that prices in other markets react more rapidly to disequilibrium than does the price of good 1. When we look at good 2 we make the same implicit assumption for the price of good 2, and so on. What Samuelson did was to make the time rate of change of each price a function of the excess demand in its own market hence arriving at the system of simultaneous differential equations:

$$\begin{aligned} dp_1/dt &= X_1(p_1, p_2, \ldots\ldots\ldots, p_M) \\ dp_2/dt &= X_2(p_1, p_2, \ldots\ldots\ldots, p_M) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ dp_M/dt &= X_M(p_1, p_2, \ldots\ldots\ldots, p_M) \end{aligned} \tag{3.2}$$

The Hicksian stability condition can be shown to be neither necessary nor sufficient for the stability of (3.2). Hicks however defended his own approach, on the ground that it answers a different but interesting question, in the Second Edition of *Value and Capital* (Additional note C).

Parts III and IV of *Value and Capital* record the effect of a road-to-Damascus-like change of vision by Hicks. It seems that while preparing his great work on price theory, Hicks read Keynes, and, to borrow a modern term, it blew his mind. He could no longer find any real satisfaction in the static formalism of Walrasian equilibrium theory, and what he then did shows the full extent of his originality. In these later Parts of the book that eventually resulted he adapted the static theory of the earlier parts to create an economic dynamics which borrowed equally from the Marshallian–Keynesian tradition of the short period and the Walras–Wicksell tradition of long-period equilibrium. The key idea was the concept of temporary equilibrium – an equilibrium of current markets in which future markets make their influence felt indirectly, through the expectations held by agents, which influence their behaviour in current markets. From this emerged the concept of the elasticity of expectations, an idea which proved to be crucial in much later work on macroeconomic theory.

4 Welfare economics

Hicks's writings on welfare economics are largely accounted for by work on four closely connected fields of interest: the foundations of welfare economics, including the famous compensation test; the valuation of social income; the definition and measurement of consumer surplus; and, lastly, the measurement of capital.

Hicks was one of the pioneers of the ‘new welfare economics’, an approach which owed its inception to Kaldor's ‘Welfare propositions in economics and interpersonal comparisons of utility’ (Economic Journal, 1939). The problem at issue is inescapable and fundamental to the justification of the recommendations of economists. By the time the debate arose, cardinal utility was no longer generally accepted and the need was felt to differentiate between ‘scientific’ propositions and ‘value judgements’. The notion of a ‘Pareto improvement’ – a change that would make no individual worse off , and at least one better off – was familiar but was seen to be limited as a basis for recommendations, as nearly all actual changes made at least one person or group worse off. In Robbins's telling example, economists could not state scientifically that the abolition of the Corn Laws was a good thing because this reform made landlords worse off.

Hicks's suggested solution to the difficulty was the same as that proposed by Kaldor – a compensation test. A reform should be counted an improvement if the gainers could afford to compensate the losers and still be better off. In ‘The Foundations of Welfare Economics’ (*Economic Journal*, 1939), Hicks discussed the question of whether compensation must be paid for the improvement to count without a sense of how crucial this question was to prove to be. It was of course central to the issue posed by the Scitovsky example, which showed that the Kaldor–Hicks rule could lead to contradictory recommendations if compensation were not paid. A well-argued solution to this problem was proposed by I.M.D. Little (1950), but this required explicit value judgements concerning whether income distribution had improved or not in a movement from one position to another, hence negating the original intention of the exercise, which had been to remove value judgements from welfare economics.

Hicks seemed to see these developments as fairly unimportant qualifications to the original idea. In ‘The Measurement of Real Income’ (1958), he writes of the ‘new welfare economists’; ‘They were indeed over-confident in their belief that they had found a means of direct comparison which will always work. But I still maintain that they did find a means of direct comparison which will often work’ (reprinted in *Collected Essays*, Vol. I, p. 168). For a statement of Hicks's mature views on these questions see ‘The scope and status of welfare economics’ (1975). Perhaps the most interesting thing to notice about Hicks's long involvement with the foundations of welfare economics is that he seems never to have wholly accepted the conclusion upon which the majority of economists have been willing to settle. Briefly put, this view says that value judgements are an inescapable element in welfare evaluations and this should be accepted and the judgements made explicit. Hence the design of policy by the means of the maximization of an explicit social welfare function – the welfare weights of cost-benefit analysis – never engaged Hicks's interest.

It is evident that the problem of the measurement of income is closely allied to the issue of welfare improvements and Hicks, as would be expected, contributed to this area as well. Hicks discussed social accounting in his text book *The Social Framework* (1942), and the valuation of social income in a paper of that title in *Economica* (1940).

Hicks concluded that the measurement of income could mean measurement in terms of utility or measurement in terms of cost, and that the two measures were in general different. The most interesting issue to which this gave rise was the problem of how to treat indirect taxation and government expenditure on goods and services in the valuation of social income. This led Hicks into controversy with Kuznets (*Economica*, 1948; see also Essay 7 in Volume I of the *Collected Essays*). The usual practice is to measure prices at factor cost and to value public services at cost. Hicks's original position may be briefly summarized as follows:

(i) As there is no market test where public goods are concerned the taxation which pays for them is not a reliable measure of their value to the consumer; and (ii) even if consumers were to be regarded as implicitly choosing public expenditure exactly as they choose private expenditure, the appropriate price weights would not be average costs but marginal costs. For a mature statement, see the Addendum to Essay 7 in Volume I of the *Collected Essays*.

Between 1941 and 1946 Hicks published a number of papers on consumer surplus in the *Review of Economic Studies* that did much to revive interest in a concept which had seemed to lose its validity when measurable utility went out of fashion. His most important contribution to the controversial question of the measurement of capital, significantly entitled ‘Measurement of Capital in Relation to the Measurement of Other Economic Aggregates’, is in F.A. Lutz and D.C. Hague (1961).

5 The Keynesian revolution and the theory of money

Hicks's first response to the *General Theory* is described in detail in ‘Recollections and documents’ (*Economica*, 1973, included in *Economic Perspectives*, 1977). However the response for which he is best known was an expository piece ‘Mr Keynes and the “Classics”•’ (1937) that perfectly fulfilled the innate demand for a more readily accessible account of the essentials of Keynes's argument. It is important to make clear that what was provided was more than an haut vulgarization of Keynes, because the paper has been widely criticized for vulgarization and still more for seriously misrepresenting what the General Theory is about. This case has never been rigorously argued and it is

hard to see how it could succeed. Hicks reproduced rather faithfully Keynes's various specifications, but by working with a two-sector model produced a framework which resulted in a simple diagram – the IS–LM diagram – which became to macroeconomic textbooks what the benzene ring diagram is to textbooks of organic chemistry. It is no surprise therefore that Keynes on reading the paper wrote to Hicks that he had '… next to nothing to offer by way of criticism'. Certainly there is more in the General Theory than just the IS–LM model. In particular there is the idea of a long-term under-consumption problem, no less worrying for being loosely formalized. Nevertheless, the IS–LM framework is there, as is what Samuelson later called the neoclassical synthesis, however much Keynes's latter-day disciples may dislike it.

In fact Hicks's way of presenting the argument is in some ways superior to that adopted in the *General Theory* because the original IS–LM model brings out very clearly how the relative price of capital and consumption goods enters into the determination of the solution – a point which is somewhat obscure in Keynes. How ironic therefore that one of the arguments later advanced against the IS–LM model, admittedly with simpler versions than Hicks's in mind, was that it omitted an essential feature of Keynes – relative prices of capital and other goods.

Hicks's IS curve is based on the striking observation that if the capital stocks in the two sectors of the economy are given, and if the money wage is known, then outputs in the two sectors depend on the nominal prices of their products through short-term profit maximization conditions. Given these outputs and prices, the value of nominal total income follows. The output of the investment sector depends on the rate of interest through the marginal efficiency of capital relation. Then, given the rate of interest, the nominal price of the investment good follows and the part of income generated in that sector. Now choose an arbitrary value, which can be thought of as a guess at the level of total nominal income. As the part of nominal income generated in the investment goods sector is known, given the rate of interest, the guess implies a certain level of nominal income to be generated in the consumption good sector. We now have a value of total income and a value of total consumption, both in nominal terms. If these values are consistent with the consumption function our guess for the value of total income was correct and we have discovered the level of income on the IS curve for the rate of interest with which we were working.

We have discussed only the IS curve but the LM curve is relatively uncomplicated – there is less going on behind it. The beauty of this elegant and lucid way of expounding Keynes's model is that it brings out clearly the vital role played in the model by aggregation assumptions which have the effect that the model decomposes, so that parts of it can be dealt with in partial isolation from the complete system. The simple specifications of the determinants of investment and the consumption function produce this result. The role played by income and working in terms of nominal values – which are equivalent to wage units, as the nominal wage has been taken as given – are all brought out clearly.

In the hands of others the IS–LM model often became merely a model of an economy with all prices fixed and was often misused, as when it was applied to long-run questions for which it is not suitable. However it made the *General Theory* intelligible to a whole generation, not because it left out the subtleties, it was never intended to substitute for the text, but because it perfectly captured the part of Keynes's message which is most amenable to formalization.

A Contribution to the Theory of the Trade Cycle (1950) provides an example of the type of model that explains cycles as the outcome of the interaction between the multiplier and the accelerator. These systems are linear in their simplest formulations when they lead to cycles which are almost certainly

either damped or anti-damped. Three different ideas have been proposed to yield an outcome in conformity to the stylized model of a capitalist economy with regular cycles of constant amplitude. The underlying solution may be anti-damped and buffers, in the form of a floor on or a ceiling to the level of economic activity, may be added to keep the solution within bounds. The system may be made non-linear, which is equivalent to buffers which make their influence felt continuously rather than abruptly. Finally, the underlying solution may be damped, in which case the cycle will have to be kept alive by the frequent intervention of random shocks. Hicks's main model embodies the last type of approach.

From 1937 Hicks continued to write regularly on questions of macroeconomics. Volume II of his *Collected Essays* contains a selection of his best work in this vein. Essay 18, ‘Methods of dynamic analysis’, proposes the distinction between the fixprice and the flexiprice economy which was to be developed in *Capital and Growth*. In his Yrjö Jahnsson lectures, *The Crisis in Keynesian Economics* (1974), Hicks offers reflections on the Keynesian theory and particularly on the impact of inflation on a Keynesian model.

Hicks never remained far from monetary theory. *Critical Essays in Monetary Theory* (1967), shows the richness of his early writings on monetary economics, while Essay 19 in Volume II of the *Collected Essays* gives a good indication of his later work. It is tempting to say that if Hicks had written nothing but his work on monetary economics he would be counted a considerable economist. However the truth is that he could not have written on monetary economics as he did write had he not been the broad economic theorist that he was. Hicks always placed monetary theory centrally in equilibrium theory. This was the distinctive idea of his first paper on the subject, ‘A Suggestion for Simplifying the Theory of Money’ (*Economica*, 1935), and it is a theme which he was to carry through all his later work.

6 Growth and capital theory

Hicks's two other books with ‘Capital’ in their titles, *Capital and Growth* (1965) and *Capital and Time: A Neo-Austrian theory* (1973a), have little else in common. *Capital and Growth* was Hicks's response to the frantic interest in growth theory which infected the 1960s. It was a characteristically personal response in which Hicks tried to apply the framework for dynamic analysis that he had developed in *Value and Capital* to the construction of a growth model.

The analogue of the static problem of Part I of *Value and Capital* was now the steady state growth path, but once again Hicks found the most interesting question to be the dynamic adjustment to equilibrium, and once again he attacked this problem with an approach which was all his own. The ‘traverse’ was the history of the movement of an economy from one steady state to another. This approach to growth theory was not very influential and the reason was not so much that the new interest in growth had extinguished interest in equilibrium theory. Rather it was that equilibrium theory and its sister economic dynamics had moved on a great deal since *Value and Capital*. Hicks, who had taught a generation how to do general equilibrium economics, was no longer talking a language that most economic theorists found congenial.

Capital and Time was not the product of the latest fashion in economic theory but was surely the result of long meditation starting from that wonderfully fruitful comparative ignorance of Marshall and Pigou as against the Austrians and other continentals, noted above. Hicks always conceded a place to the old classical idea that capital accumulation means more ‘waiting’. In *Value and Capital* (1939a, pp. 197–8) however, he pointed out that the conclusion that the rate of interest is the marginal product of waiting is

a special case of more general rules which apply to an intertemporal equilibrium. This conclusion, that Austrian models of capital are special cases of the more general von Neumann model of capital accumulation, remains valid. However special cases permit of special results, and Hicks's analysis of the Austrian model was remarkably successful in showing how that framework permits some strong and definite conclusions to be drawn.

7 Other topics

We consider only *A Theory of Economic History* (1969) and *Causality in Economics* (1979), as these constitute the most audacious of Hicks's expeditions far from the mainstream of economic theory. A longer review of Hicks's work would have to find space to discuss his writings on economic policy (for a sample of which see *Essays in World Economics*, 1959) and on the history of economic thought (for some of which see Volume III of the Collected Essays), but here we merely note that these are serious omissions from the present survey.

We begin with *Causality in Economics*. This book was not the eventual product of long years of mental rumination, but the result, its author tells us frankly, of dissatisfaction with the 1974 International Economic Association conference on ‘The Micro-foundations of Macroeconomics’ which Hicks attended. It is book of interesting ideas on economics which are reluctantly regimented by a Sergeant-Major called ‘causality’. This gentleman turns out to be only remotely related to the ‘causation’ of Aristotle or Kant. Hicks's definition of causality is reminiscent of Hume, but without the idea that the validity of induction is importantly involved. Causation is seen as conjunction of events, possibly in a complex form. This idea is an old one and was very effectively criticized by the Cartesians but their contribution is not considered. As an essay in philosophy *Causality in Economics* cannot be taken seriously. The economics of course is of a higher standard. The last chapter provides a statement of Hicks's views on the meaning of probability and on econometric methodology. These are *obiter dicta*, not the fruits of profound investigation.

A Theory of Economic History is as ambitious a sortie into foreign territory as *Causality in Economics*, but is the product of more thought and reading and must be regarded as much more successful. The main idea, that economic history is tied up with the development of the market, is one that few would question. However most historians would be tempted to take cover behind a safe position according to which developments of ideas, knowledge, social institutions, etc., would all be seen as progressing in parallel with the development of the market, which consequently would enjoy no special status as a motive force. Put simply, Hicks's account gives a much more leading role to the market, although he does not of course go so far as to argue that the market drives history.

Such a strong argument could not fail to attract criticism, particularly from professional historians. A long book would have done the same but a very short book was a particularly provocative target. As the argument gave a lot of attention to the ancient world this proved to be a contested area. However while *A Theory of Economic History* was criticized it received respectful criticism. It may be only a way of looking at economic history but it was generally judged to be a good way. Hicks's reply to his critics may be found in *Economic Perspectives* (1977, pp. 181–4).

8 Retrospect

Schumpeter argued that the ideas of a great economist are more or less in place by the age of 40 – the rest is nurturing and polishing. At first glance Hicks appears to be an exception. He was 65, for example, when his theory of economic history was announced to the world. Yet probably on closer examination he will be seen to conform to the Schumpeter pattern. In the case of the *A Theory of Economic History* he tells us in the foreword that he had nursed the idea for years. There is indeed a powerful sense of direction to Hicks's intellectual journey. He often returns to old themes and new themes are examined from older perspectives. Probably after 40 Hicks was only nurturing and polishing, but it is no contradiction of that claim to say that the second half of his life produced some of his most creative work.

It remains to mention some particular qualities of Hicks the man. First, he wrote beautifully, in a style that is very correct from the formal point of view, yet almost conversational in its flow and ease. Secondly, his greatness justified a little vanity, and he was not wholly free of that minor vice. That said, he was always approachable, and he never attempted to win an argument by pulling rank or flaunting his formidable distinction.

Selected works

1932. *The Theory of Wages*. London: Macmillan 2nd edn, London: Macmillan, 1963.

1934(With R.G.D. Allen.) A reconsideration of the theory of value. Pts I–II. *Economica* I, Pt I, February, 52–76; Pt II, May, 196–219.

1935. A suggestion for simplifying the theory of money. *Economica* 2 February, 1–19.

1937. Mr Keynes and the ‘classics’. *Econometrica* 5 April, 147–59.

1939a. *Value and Capital*. Oxford: Clarendon Press.

1939b. The foundations of welfare economics. *Economic Journal* 49, 696–712.

1940. The valuation of the social income. *Economica* 7 May, 105–24.

1942. *The Social Framework*. Oxford: Clarendon Press.

1948. The valuation of the social income: a comment on Professor Kuznets’ reflections. *Economica* 15 August, 163–72.

1950. *A Contribution to the Theory of the Trade Cycle*. Oxford: Clarendon Press.

1956. *A Revision of Demand Theory*. Oxford: Clarendon Press.

1958. The measurement of real income. *Oxford Economic Papers* 10 June, 125–62.

1959. *Essays in World Economics*. Oxford: Clarendon Press.

1961. Measurement of capital in relation to the measurement of other economic aggregates. In *The Theory of Capital*, ed. F.A. Lutz and D.C. Hague. London: Macmillan; New York: St Martin's Press.

1965. *Capital and Growth*. Oxford: Clarendon Press.

1967. *Critical Essays in Monetary Theory*. Oxford: Clarendon Press.

1969. *A Theory of Economic History*. Oxford: Clarendon Press.

1973a. *Capital and Time: A Neo-Austrian theory*. Oxford: Clarendon Press.

1973b. Recollections and documents. *Economica* 40 February, 2–11.

1974. *The Crisis in Keynesian Economics*. Oxford: Basil Blackwell.

1975. The scope and status of welfare economics. *Oxford Economic Papers* 27, 307–26.

1977. *Economic Perspectives*. Oxford: Clarendon Press.

1979. *Causality in Economics*. Oxford: Basil Blackwell.

1981–3. *Collected Essays on Economic Theory*. Oxford: Basil Blackwell.

Bibliography

Kaldor, N. 1939. Welfare propositions in economics and interpersonal comparisons of utility. *Economic Journal* 49, 549–52.

Little, I.M.D. 1950. *A Critique of Welfare Economics*. Oxford: Clarendon Press.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

How to cite this article

Bliss, Christopher. "Hicks, John Richard (1904–1989)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000052> doi:10.1057/9780230226203.0729

The New Palgrave Dictionary of Economics Online

Hicks, Ursula Kathleen (1896 – 1985)

Alan Peacock
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

direct and indirect taxation; federal finance; Hicks, U. K.; local public finance; public finance; tax incidence; taxation of expenditure; taxation of income

Article

An Irish-born economist specializing in public finance, Lady Hicks's long career spanned teaching and research at the London School of Economics and Political Science, University of Liverpool, and the University of Oxford (latterly as Foundation Fellow of Linacre College), as well as the holding of many visiting posts in foreign universities and service as a member of advisory missions on fiscal matters, notably in the Caribbean, India and Africa.

She made three significant contributions to her specialism, the theory and practice of public finance. Her paper ‘The Terminology of Tax Analysis’ (1946) questioned the usefulness of the distinction between direct and indirect taxes and argued persuasively for distinguishing between taxes on income and taxes on expenditure (outlay), the dichotomy now used in national accounting. She also explored the difference between the formal incidence of taxes (the liability to pay taxes) and the effective incidence (the determination of tax burdens). Second, in collaboration with her husband, Sir John Hicks, she endeavoured to produce coherence between what the aims of government should be and how fiscal institutions should be organized to achieve them (1947, Part 3). Third, she applied a unique knowledge of fiscal systems to the study of federal and local finance particularly in developing countries (for example, 1961).

No account of her contribution would be complete without mentioning her immense influence as a teacher of students of public finance from all parts of the world and her part in the foundation of the *Review of Economic Studies* (together with Abba Lerner and Paul Sweezy), of which she was Managing Editor from 1933 to 1961.

Selected works

1946. The terminology of tax analysis. *Economic Journal* 56(March), 38–50.

1947. *Public Finance*. London: Nisbet & Co.; Cambridge: Cambridge University Press.

1961. *Development From Below*. Oxford: Clarendon Press.

Bibliography

David, W.L., ed. 1973. Introduction. In *Public Finance, Planning and Economic Development Essays in Honour of Ursula Hicks*, ed. W.L. David. London: Macmillan.

How to cite this article

Peacock, Alan. "Hicks, Ursula Kathleen (1896–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000053> doi:10.1057/9780230226203.0730

The New Palgrave Dictionary of Economics Online

Hicksian and Marshallian demands

Eugene Silberberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Soon after the presentation of demand in Alfred Marshall's *Principles of Economics* in 1890, a debate ensued concerning whether money income or some sort of real income should be held constant as the price of the good changed. By the mid-20th century, these two conceptions of a demand function became known as the Marshallian and Hicksian functions, respectively. The issue is critical to the interpretation of the area to the left of the demand curve between two prices as some sort of consumer surplus, that is, the gain from purchasing a good at the lower price.

Keywords

compensated demand; constrained maximum problem; consumer surplus; Edgeworth, F.; envelope theorem; Hicksian and Marshallian demand curves; homogeneity; income effect; law of demand; marginal utility of income; marginal utility of money; Palgrave, R.; Pollack, R.; Roy's identity; Samuelson, P.; Shephard's lemma; Slutsky equation; substitution effect; uncompensated demand

Article

Although earlier writers had formulated the concept of a (downward sloping) demand curve, the analysis took on much great refinement with the publication of Alfred Marshall's *Principles of Economics* in 1890 (and continuing until 1920 with the eighth edition). In Chapter III, Marshall derived the law of demand from a postulate of diminishing marginal (cardinal) utility. He measured utility in terms of money, constantly reminding us, however, of the necessity to assume that the marginal utility of money remained constant. Although it would be reasonable to conclude that the demand function he had in mind is the standard formulation $x_i = x_i^*(p_1, \ldots, p_n, M)$, where the p_i's are the prices of the n goods and M is money income, Marshall never wrote out an expression such as the above. Although he was very clear that the demand curve represented diminishing marginal values of the good to an individual, he never specified the *ceteris paribus* conditions we are now familiar with.
However, as early as 1894 in the original *Dictionary of Political Economy* edited by R. Palgrave, Edgeworth gave the now current interpretation, stated above. (See the interesting footnote 5 in Friedman, 1949). However, Marshall's suggestion that individuals purchase additional quantities only if the additional utility they gain is at least as great as the price paid suggests that the demand price represents the maximum the individual will pay for an additional unit. In that case, it would be utility, rather than money income, that was being held constant along the demand curve. To obscure things further, in Chapter VI of the same book, ‘Consumer's surplus’, Marshall insists that the adding up of demand prices to generate the consumer surplus, or net benefits of all the units purchased, is valid only when the marginal utility of income is constant, or the same across individuals within a market demand curve. These remarks about marginal utility spawned an industry of economists working on consumer surplus for the first 75 years of the 20th century. The matter finally came to an end in the 1970s when the derivations of the demand functions with either money income or utility as an argument were clarified.

The Marshallian demand functions

There are two main threads motivating the entire literature on Hicksian and Marshallian demands: first and foremost, consumer's surplus, and second, providing a rigorous discussion of the pure substitution term in the Slutsky equation. For convenience I limit the discussion to the case of two goods. The Marshallian demand functions are the solutions to the constrained maximum problem

$$\text{maximize } U = U(x_1, x_2)$$
$$\text{subject to } p_1x_1 + p_2x_2 = M$$

where, of course, x_1 and x_2 are two goods; their prices p_1 and p_2 and income M are assumed exogenous. The Lagrangian for this model is $L = U(x_1, x_2) + \lambda(M - p_1x_1 - p_2x_2)$. Differentiating partially with respect to x_1, x_2 and λ yields the necessary first-order conditions (NFOC)

$$L_1 = U_1(x_1, x_2) - \lambda p_1 = 0 \quad (1a)$$

$$L_2 = U_2(x_1, x_2) - \lambda p_2 = 0 \quad (1b)$$

$$L_\lambda = M - p_1x_1 - p_2x_2 = 0 \quad (1c)$$

The sufficient second-order condition (SSOC) is that the bordered Hessian determinant of L be positive:

$$H = \begin{vmatrix} U_{11} & U_{12} & -p_1 \\ U_{21} & U_{22} & -p_2 \\ -p_1 & -p_2 & 0 \end{vmatrix} > 0 \quad (2)$$

(In the case of n goods, the border-preserving principal minors of H alternate in sign. See, for example, Silberberg and Suen, 2000.) On the assumption that the SSOC holds, the NFOC can be solved simultaneously for the demand functions with money income as a argument, now universally termed the Marshallian demand functions

$$x_1 = x_1^M(p_1, p_2, M) \tag{3a}$$

$$x_2 = x_2^M(p_1, p_2, M) \tag{3b}$$

and the Lagrange multiplier

$$\lambda = \lambda^M(p_1, p_2, M) \tag{3c}$$

In the parlance of intermediate microeconomics texts, when a price changes ‘money income is held constant’. But this is just an imprecise way of stating that the demand for any good is a function of the price of that good, the prices of all other relevant goods, and, in particular, money income.

Substituting the Marshallian demand functions (3a) and (3b) into the utility function yields the maximum utility for given prices and money income, $U^*(p_1, p_2, M)$. This is the indirect objective (utility) function for this model. By the envelope theorem (see Silberberg and Suen, 2000)

$$U^*_{p_i} = L_{p_i} = -\lambda^M x_i^M \; i = 1, 2 \tag{4a}$$

$$U^*_M = L_M = \lambda^M \tag{4b}$$

Equation (4b) reveals that the Lagrange multiplier is the marginal utility of income. On the assumption that the constraint is preventing the consumer from gaining a higher utility, $\lambda^M > 0$. Equation (4a), known as Roy's identity, shows that (maximum) utility varies inversely with price, as previously indicated, since consumption levels are assumed positive.

The traditional comparative statics of this model proceeds by substituting eqs. (3) into the NFOC and differentiating with respect to M and, say, p_1. Since p_1 enters *two* of the first-

order equations, two terms are produced in the expression for $\partial x_1^M / \partial p_1$. In 1916, Slutsky identified these terms as a substitution effect (which is always negative) and an income term. Rather than replicate these somewhat tedious calculations, we proceed to the more modern analysis.

The Hicksian demand functions

Consider now an alternative formulation of consumer behaviour, that of minimizing the expenditure needed to achieve a specified utility level at give prices:

$$\text{minimize } M = p_1 x_1 + p_2 x_2$$
$$\text{subject to } U(x_1, x_2) = U^0$$

The Lagrangian for this model is $L = p_1 x_1 + p_2 x_2 + \lambda(U^0 - U(x_1, x_2))$. Differentiating with respect to x_1, x_2 and λ as before produces the following NFOC and SSOC:

$$L_1 = p_1 - \lambda U_1(x_1, x_2) = 0 \quad (5a)$$

$$L_2 = p_2 - \lambda U_2(x_1, x_2) = 0 \quad (5b)$$

$$L_\lambda = U^0 - U(x_1, x_2) = 0 \quad (5c)$$

$$H = \begin{vmatrix} -\lambda U_{11} & -\lambda U_{12} & -U_1 \\ -\lambda U_{21} & -\lambda U_{22} & -U_2 \\ -U_1 & -U_2 & 0 \end{vmatrix} < 0 \quad (6)$$

On the assumption that the SSOC holds, eqs. (5) can be solved simultaneously for the Hicksian demand functions

$$x_1 = x_1^U(p_1, p_2, U^0) \tag{7a}$$

$$x_2 = x_2^U(p_1, p_2, U^0) \tag{7b}$$

and the Lagrange multiplier

$$\lambda = \lambda^U(p_1, p_2, U^0) \tag{7c}$$

Eliminating λ from eqs. (5a) and (5b) produces the same ‘tangency’ relation as eliminating λ from (1a) and (1b):

$$\frac{U_1}{U_2} = \frac{p_1}{p_2} \tag{8}$$

In both models, the consumer chooses the point on an indifference curve where the budget or expenditure line has the same slope as the indifference curve. Alternatively, the consumer chooses a mix of goods such that

$$\frac{U_1}{p_1} = \frac{U_2}{p_2} \tag{9}$$

That is, the individual consumes each good until the marginal benefit (utility) per dollar is the same across all commodities. *At the margin*, all goods consumed are perfect substitutes. Given an increment of income, the consumer would be indifferent as to how to spend it, since he or she has already equalized the marginal utility of a dollar across all goods. However, the comparative statics of these two models are not the same. For the Hicksian demand functions, when a price changes utility is held constant. That is, the consumer is

constrained to slide along the same indifference curve. Thus, the responses to changes in prices are, by definition, the pure substitution effects. In the Marshallian case, as a price changes utility changes also, in the opposite direction.

Substituting the Hicksian functions (7a) and (7b) into the objective function, we obtain the expenditure function $M = M^*(p_1, p_2, U^0)$. This is the indirect objective function in this model; it gives the minimum expenditure needed to achieve utility level U^0 at prices p_1 and p_2. Since $M^*(p_1, p_2, U^0)$ is a minimum expenditure, by definition, $M^* \leq M$, so that the function $F = p_1x_1 + p_2x_2 - M^*(p_1, p_2, U^0)$ has a (constrained) minimum (of zero) with respect to not only x_1 and x_2, but also p_1, p_2, and U^0. The Lagrangian for this 'primal-dual' problem is

$$L = p_1x_1 + p_2x_2 - M^*(p_1, p_2, U^0) + \lambda(U^0 - U(x_1, x_2))$$

The first-order equations with respect to x_1, x_2 and λ are just eqs. (5); with respect to p_1, p_2 and U^0 we have

$$L_{p_1} = F_{p_1} = x_1 - M^*_{p_1} = 0 \quad (10a)$$

$$L_{p_2} = F_{p_2} = x_2 - M^*_{p_2} = 0 \quad (10b)$$

$$L_{U^0} = M^*_{U^0} - \lambda = 0 \quad (10c)$$

Eqs. (10a) and (10b) are sometimes referred to as Shephard's lemma (Shephard, 1970): the Hicksian demand functions are the first partials of the expenditure function. Moreover, since p_1 and p_2 do not enter the constraint, F has an *unconstrained* minimum in p_1 and p_2. The second-order conditions include

$$F_{p_ip_i} = -M^*_{p_ip_i} \geq 0 \; i = 1, 2 \quad (11)$$

so that the expenditure function is concave in prices. However, since $M^*_{p_i} \equiv x_i^U(p_1, p_2, U^0)$,

$$\frac{\partial x_i^U}{\partial p_i} = M^*_{p_i p_i} \leq 0 \tag{12}$$

The pure substitution effects, which are the ordinary slopes of the Hicksian demand curves, are negative. No such sign is implied for the Marshallian demand functions, since in the associated primal–dual problem, the prices enter the constraint, eliminating any implications about slope of the demand functions based on the curvature properties of the indirect utility function. Additionally, for the Hicksian demands, we find the reciprocity condition

$$M^*_{p_1 p_2} = \partial x_1^U / \partial p_2 = \partial x_2^U / \partial p_1 = M^*_{p_2 p_1} \tag{13}$$

Homogeneity

The Marshallian demands are the solutions to eq. (8) and the budget constraint (1c). Suppose there is a proportionate increase in prices and money income. That is, $p_1 \to tp_1$, $p_2 \to tp_2$ and $M \to tM$ where $t > 0$. But eqs. (8) and (1c) are unchanged by this proportionate increase in parameters; hence their solutions must be identical. Thus, $x_i^M(tp_1, tp_2, tM) \equiv x_i^M(p_1, p_2, M)$; the Marshallian demand functions are homogeneous of degree zero in prices and money income. Consumers respond to changes in relative prices, not absolute prices. Similarly, the Hicksian demands are the solutions to eqs. (8) and the constant utility constraint (5c). If $p_1 \to tp_1$ and $p_2 \to tp_2$, these equations are unchanged, and therefore $x_i^U(tp_1, tp_2, U^0) \equiv x_i^U(p_1, p_2, U^0)$. With the use of these properties, the indirect utility function must also be homogeneous of degree zero in prices and income, while the expenditure function is homogeneous of degree one in all prices.

The Slutsky equation

Evgeny Slutsky (1916) published perhaps the seminal work in the economic theory of the consumer, in which he showed that a consumer's response to a change in price could be partitioned into two parts: a pure substitution effect which was always negative (that is, in the opposite direction to the price) and an income effect, whose sign was indeterminate. When a price, say p_1, decreases, the budget line pivots outward, intersecting the x_1 axis at a greater amount. Slutsky decreased the consumer's income by shifting the new flatter budget line back until it went through the original equilibrium. By such a ‘compensation’, Slutsky isolated the substitution effect. At various schools, particularly the University of Chicago, the Marshallian and Hicksian demand curves were referred to respectively as uncompensated and compensated demand curves. By the 1930s and 1940s, with the publications of John Hicks's *Value and Capital* (1939) and Paul Samuelson's *Foundations of Economic Analysis* (1947), the ‘pure’ substitution effect had become defined as the response to a price change when the budget line was shifted (parallel to itself) back to *the original indifference curve*, producing a response in consumption holding utility constant. (It was shown by J. Mosak and A. Wald, 1942, that, if p_1 changed by an amount Δp_1, the differences between the Slutsky and Hicks demands were of second-order smallness, that is, functions of powers of Δp_1 order two and higher. See also Silberberg and Suen, 2000, p. 304.)

By the 1940s, the profession had largely settled on the Hicksian interpretation of the pure substitution effect (though Friedman, 1949, stressed the operational advantage of Slutsky's measure). However, it was not until the 1970s that the demand functions (7a) and (7b) derived from constrained expenditure minimization came to be recognized as the analytical basis for the pure substitution effects. At that point, economists realized that the Slutsky equation showed the relationship between the Hicksian and Marshallian demand functions.

We now demonstrate this using the concept of ‘conditional demands,’ first developed by Robert Pollak (1969).

The Hicksian demand function is obtained from the Marshallian function by adjusting money income, when a price changes, to the minimum amount necessary to keep the consumer on the same utility level. Stated mathematically, this is the identity (for x_1, say)

$$x_1^U(p_1, p_2, U^0) \equiv x_1^M(p_1, p_2, M^*(p_1, p_2, U^0)) \tag{14}$$

Differentiating this identity with respect to p_1,

$$\frac{\partial x_1^U}{\partial p_1} \equiv \frac{\partial x_1^M}{\partial p_1} + \frac{\partial x_1^M}{\partial M}\frac{\partial M^*}{\partial p_1} \tag{15}$$

Applying the envelope theorem (10a) and rearranging, we get the classic Slutsky equation

$$\frac{\partial x_1^M}{\partial p_1} \equiv \frac{\partial x_1^U}{\partial p_1} - x_1\frac{\partial x_1^M}{\partial M} \tag{16}$$

The slope of the Marshallian demand equals the slope of the Hicksian demand (which is always negative in its own price) and an indeterminate income effect. If, however, the good is non-inferior, that is, the income effect is non-negative, then the Marshallian demand curve is necessarily downward sloping.

Consumer's surplus

Most of the interest in the distinction between Hicksian and Marshallian demand functions was generated by the analysis of consumer's surplus. Marshall developed the concept as follows. (I use hamburgers and dollars in place of Marshall's quainter example of tea and shillings.) Suppose, at a price of $10, a consumer will buy only one hamburger a month. At a price of $9, he will buy two; at $8, he will buy three, at $7, four, and so on. Since these prices measure the marginal values of hamburgers to this consumer at these consumption levels, we interpret these numbers as the maximum the consumer would pay for an additional hamburger. In that case, the amount the consumer would pay to consume four hamburgers rather than none would be $\$10 + \$9 + \$8 + \$7 = \$34$. Marshall thus interpreted the area under the demand curve as the all-or-nothing value of that quantity of a good, or the total utility of those units, measured in units of money income. Additionally, at a price of $7, say, the consumer would spend $28 on the good, leaving the area to the right of the demand curve above the price, $6, as the consumer's surplus. This is the additional amount the consumer would have been willing to pay to consume the four units at a price of $7. The question is: when can we interpret the area to the left of a demand curve in this fashion? Consider a decrease in the price of x_1 from p_1^0 to p_1^1. Mathematically,

$$CS = -\int_{p_1^0}^{p_1^1} x_1 \, dp_1$$

(17)

For both the Hicksian and Marshallian demand curves, we can calculate this area, which has the units of dollars – not utility – being price times quantity. However, the important question is: when we calculate this amount, what question, if any, does it answer? In the case of the Hicksian demands, the answer is clear. If we use the envelope relation (10a),

$$CS = -\int_{p_1^0}^{p_1^1} x_1^U \, dp_1 = -\int_{p_1^0}^{p_1^1} \frac{\partial M^*}{\partial p_1} dp_1 = M^*(p_1^0, p_2, U^0) - M^*(p_1^1, p_2, U^0)$$

(18)

Because the Hicksian demands are the first partials of the expenditure function, the integral is simply the savings in expenditure the consumer enjoys when the price is lowered (or, likewise, the extra expenditure the consumer must make if the price increases). Thus the area to the left of a Hicksian demand curve is the amount consumers would be willing to pay, or have to be paid, to face the new price. Moreover, suppose two prices change. That is, suppose the price of x_1 changes from \$8 to \$4, and we calculate a CS_1 (\$18 if we use the above linear demand curve). The demand curve for x_2 will have now shifted. Suppose we now lower the price of x_2 from \$7 to \$3, generating $CS_2 = \$15$, say, producing $CS = CS_1 + CS_2 = \$33$. Suppose we did the experiment in the reverse order – lowering the price of x_2 first and then x_1. Would we get the same answer for total *CS*? Indeed we would:

$$CS = -\int_{p^0}^{p^1} \sum x_i^U \, dp_i = -\int_{p_1^0}^{p_1^1} \sum \frac{\partial M^*}{\partial p_i} dp_i = -\int_{p^0}^{p^1} dM^* = M^*(p_1^0, p_2^0, U^0) - M^*(p_1^1, p_2^1, U^0)$$

(19)

Because of the reciprocity condition (13), this integral is exact; the path of price changes does not affect the value of the integral.

In the case of the Marshallian demands, no such interpretations are possible (Silberberg, 1972). The Marshallian demands x_i^M are not the first partials of any function, so the area to the left of the demand curve given by (17) has no easy interpretation. Moreover, since for the Marshallian demands $\partial x_1^M / \partial p_2 \neq \partial x_2^M / \partial p_1$ (unless the utility function is homothetic) the integral corresponding to (19) for the Marshallian demand functions is path dependent. That is, depending on which price changes first, a different answer emerges, even if all the initial and final prices are identical in the two experiments. There is simply no unique measure of a change in utility in terms of income, except for some famous special cases. (See Silberberg and Suen, 2000).

Consider Figure 1, for some good *x*. At the initial price *OA*, the consumer purchases *AB*. If the price decreases to *OF*, she moves along her Marshallian demand curve x^M, and consumes *FD*. If, however, we were to keep her on the same initial indifference curve U^0 as *p* decreased, she would move along the Hicksian demand curve to point *E*. Point *E* would be to the left of *D* if the good is normal (non-inferior), since we are eliminating this (positive) income effect. If however, the consumer started at the lower price *OF* and the price were raised to *OA*, *and* we now held her on the higher level of utility U^1 she achieved at point *D*, she would move up along the Hicksian demand curve associated with U^1, leading her to point *C*. Suppose the area to the left of the Hicksian demand curve *BE* is \$20, the area to the left of the Marshallian demand curve *BD* is \$25, and the area to the left of the Hicksian demand curve *CD* is \$30. What questions, if any, do these numbers answer? The amount the consumer would pay to face price *OF* instead of *OA* is \$20. If the price were initially *OF*, the amount she would have to be paid to voluntarily face *OA* instead is \$30. It seems odd, but it is true nonetheless that, for non-inferior goods, the amount one must be paid to face a higher price is greater than the amount we would pay to get the lower price. Lastly, there is simply no operational question for which \$25 is the answer. However, Robert Willig (1976) investigated the actual empirical differences one would be likely to encounter; not too surprisingly, they turn out to be small.

Figure 1

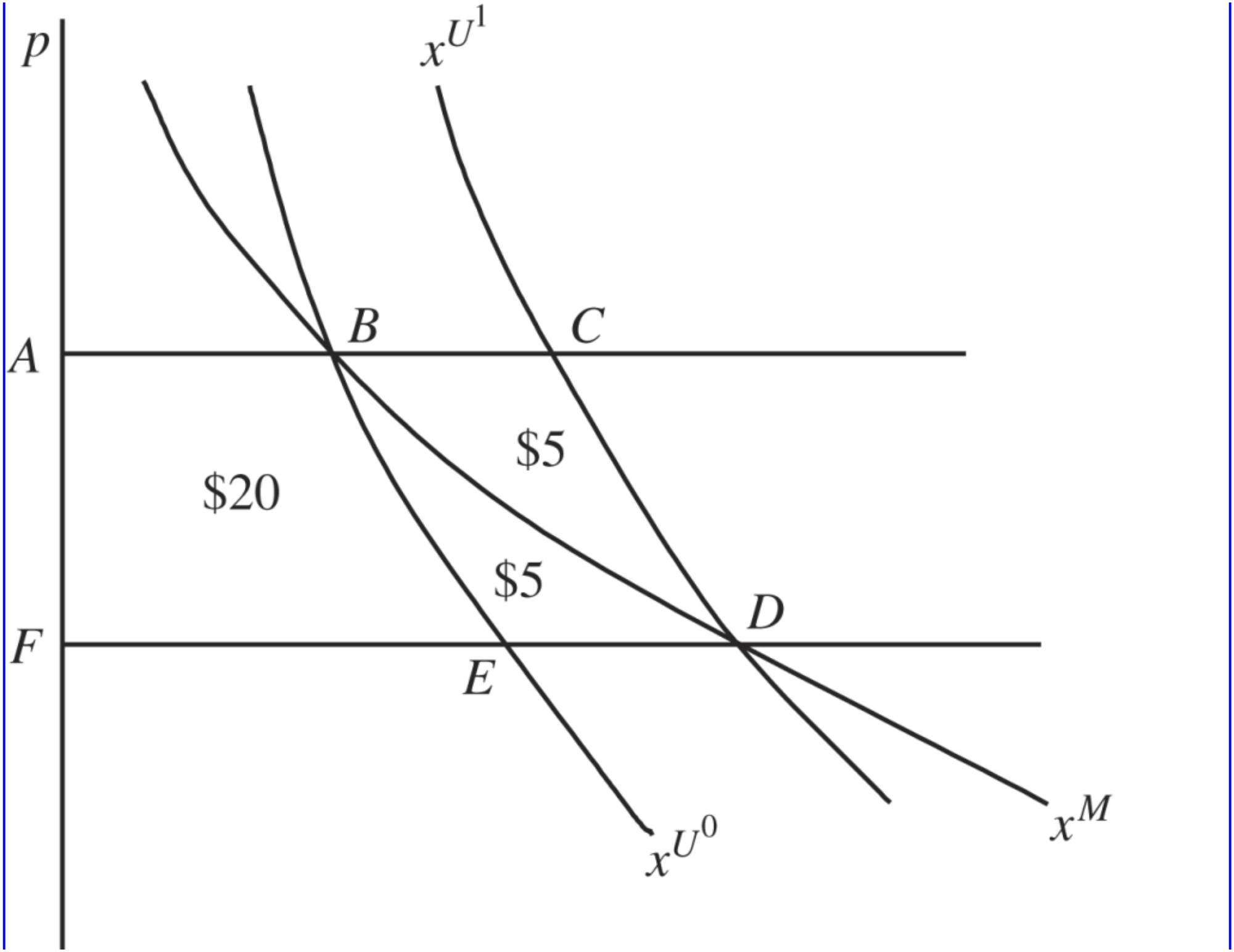
p
x^{U^1}
B
C
A
$5
$20
$5
D
F
E
x^{U^0}
x^M

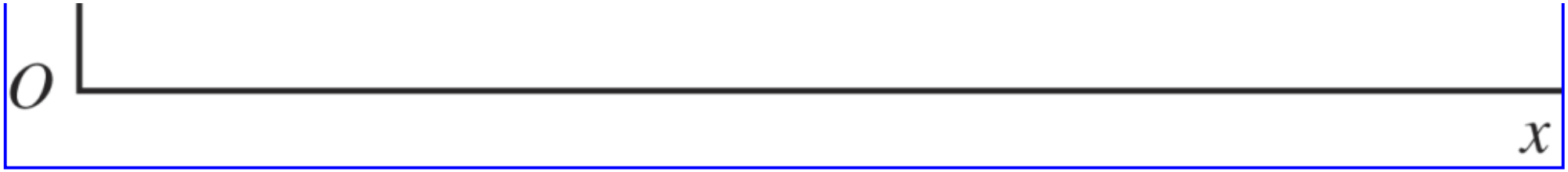

See Also

- envelope theorem
- Le Chatelier principle
- marginal utility of money

Bibliography

Edgeworth, F. 1894. Demand curves. In *Dictionary of Political Economy*, ed. R. Palgrave. London: Macmillan.

Friedman, M. 1949. The Marshallian demand curve. *Journal of Political Economy* 57, 464–74.

Hicks, J. 1939. *Value and Capital*. London: Oxford University Press.

Marshall, A. 1920. *Principles of Economics*. 8th edn. London: Macmillan.

Mosak, J. 1942. On the interpretation of the fundamental equation of value theory. In *Studies in Mathematical Economics and Econometrics*, ed. O. Lange, F. McIntyre and T. Entema. Chicago: University of Chicago Press, pp. 69–74, esp. 73–74, n. 5, which contains a rigorous proof of this statement by A. Wald.

Pollak, R. 1969. Conditional demand functions and consumption theory. *Quarterly Journal of Economics* 83, 60–78.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Shephard, R. 1970. *The Theory of Cost and Production Functions*. Princeton: Princeton University Press.

Silberberg, E. 1972. Duality and the many consumer surpluses. *American Economic Review* 62, 942–56.

Silberberg, E. and Suen, W. 2000. *The Structure of Economics*. 3rd edn. New York: Irwin/McGraw-Hill.

Slutsky, E. 1916. On the theory of the budget of the consumer. In *Readings in Price Theory*, ed. G. Stigler and K. Boulding. Homewood, IL: Richard D. Irwin, Inc., 1952.

Willig, R. 1976. Consumer's surplus without apology. *American Economic Review* 66, 589–97.

How to cite this article

Silberberg, Eugene. "Hicksian and Marshallian demands." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave

Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000167> doi:10.1057/9780230226203.0731

The New Palgrave Dictionary of Economics Online

hierarchical Bayes models

Siddhartha Chib and Edward Greenberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The standard Bayesian model is defined in terms of an outcome model and the prior density of the parameters. The latter depends on parameters called hyperparameters. A hierarchical Bayes model results when one or more of the hyperparameters are assumed to be random and modelled probabilistically. We discuss canonical versions of these models for the case when both the parameters and the hyperparameters are modelled in groups or blocks, provide relevant examples, and discuss how inference by Markov chain Monte Carlo methods makes even the fitting of complex hierarchical models practical and simple. The problem of model comparisons is also addressed.

Keywords

Bayes' th; component densities; exchangeability; hierarchical Bayes models; hyperparameters; marginal likelihood; Markov chain Monte Carlo methods

Article

Suppose that *y* is a univariate random variable or multivariate random vector and θ is a *d*-dimensional parameter vector that lies in Θ, a subset of $\Re^d$. The standard Bayesian model is then defined in terms of the density of *y* given θ (the outcome model) and the prior density of θ (the prior model). Specifically, the Bayesian model is specified as

$$y|\theta \sim p(y|\theta) \text{ (outcome model: stage 1)} \tag{1}$$

$$\theta|\gamma \sim \pi(\theta|\gamma) \text{ (prior model: stage 2)} \tag{2}$$

where γ is the vector of parameters in the prior density. These are called hyperparameters. We can assume that γ is g-dimensional and lies in Γ, a subset $\Re^g$. The labelling of the outcome model as stage 1 and the prior model as stage 2 is arbitrary, and the numbering can be reversed. The outcome model may be called the top or bottom level of the model because this difference in nomenclature has no significance.
Suppose that the researcher is not able to specify one or more of the hyperparameters in γ. In that case, the unknown hyperparameters can be assumed to be random and modelled probabilistically. This modelling of the hyperparameters leads to what is called a Bayesian hierarchical model (Berger, 1985; Lehmann and Casella, 1998). The simplest version of a Bayesian hierarchical model is defined in terms of the ingredients

$$y|\theta \sim p(y|\theta) \text{ (outcome model: stage 1)} \tag{3}$$

$$\theta|\gamma \sim \pi(\theta|\gamma) \text{ (prior model: stage 2)} \tag{4}$$

$$\gamma|\lambda \sim \psi(\gamma|\lambda) \text{ (hyperparameter model: stage 3)}, \tag{5}$$

where $\psi(\gamma|\lambda)$ is the prior density of γ. The hyperparameters λ in the stage 3 model are assumed known. In effect, a hierarchical model is a way of modelling the outcomes and the parameters through a sequence of easily interpretable steps.

In practice, it is often helpful to divide θ into natural groups or blocks $(\theta_1, \theta_2, \ldots, \theta_p)$, where, for instance, θ_1 consists of the regression coefficients, θ_2 the scale parameters and θ_p the covariance parameters. Each of these separate blocks may then be modelled independently in terms of prior densities $\pi(\theta_j|\gamma)$. In turn, γ may also be grouped into blocks $(\gamma_1, \ldots, \gamma_q)$ and, in the third stage, modelled independently through the densities $\psi(\gamma_j|\lambda)$. The resulting three-stage hierarchical model then has the form

$$y|\theta \sim p(y|\theta) \text{ (outcome model: stage 1)} \tag{6}$$

$$\theta|\gamma \sim \prod_{j=1}^{p} \pi(\theta_j|\gamma) \text{ (prior model: stage 2)} \tag{7}$$

$$\gamma|\lambda \sim \prod_{j=1}^{q} \psi(\gamma_j|\lambda) \text{ (hyperparameter model: stage 3)}. \tag{8}$$

This specification may be considered as the canonical hierarchical Bayes model.

Example 1: (Gaussian linear regression model). *Suppose that $y=(y_1,\ldots,y_n)$ is a vector of observations and θ consists of the two blocks (β, σ^2), where β is a k-vector of regression parameters. Now let*

$$y|\theta \sim N_n(y|X\beta, \sigma^2 I_n)$$
$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0) IG\left(\sigma^2 \middle| \frac{\nu_0}{2}, \frac{\delta_0}{2}\right),$$

where

$$N_k(\beta|\beta_0, B_0) = (2\pi)^{-k/2}|B_0|^{-1/2}\exp\left\{-\frac{1}{2}(\beta-\beta_0)B_0^{-1}(\beta-\beta_0)\right\}$$

is the k-variate normal density, X is the n×k matrix of covariates and

$$IG\left(\sigma^2|\frac{v_0}{2}, \frac{\delta_0}{2}\right) = \frac{(\delta_0/2)^{(v_0/2)}}{\Gamma(v_0/2)}\left(\frac{1}{\sigma^2}\right)^{(v_0/2)+1}\exp\left(-\frac{\delta_0}{2\sigma^2}\right), \sigma^2 > 0$$

is the inverse-gamma density. In this case, the hyperparameters γ *consist of the four blocks of parameters* $(\beta_0, B_0, v_0, \delta_0)$*. The top level of the model is the model of the outcome and the bottom level the model of* θ*. If it is not possible to fix the value of* β_0*, for example, one may specify a prior,* $\beta_0|\lambda \sim N_k(\beta_0|\beta_{00}, B_{00})$*, where the hyperparameters of the third stage* $\lambda = (\beta_{00}, B_{00})$ *are pre-specified. Further discussion along these lines is provided by Lindley and Smith (1972).*

Since the difficulty of specifying hyperparameters in the second stage model of the model arises in almost all applications, hierarchical Bayes modelling is of special interest and importance in Bayesian analysis. To further fix the ideas, the following example, which we develop further below, is instructive and should be studied carefully.

Example 2: (Gaussian clustered data model). *Clustered data arise when* n_i *observations are available for each subject i (*$i \leq n$*) in the sample. For example, in the panel or longitudinal set-up, there are observations across time for each subject. Let the observations on the ith subject be denoted by* $Y_i = (Y_{i1}, \ldots, Y_{in_i})$*. Assume that the observations are continuous. Binary or ordinal responses can be dealt with in much the same way by adopting the framework of Albert and Chib (1993). The data for all n subjects are collected in the vector y=(*y_1*,…,*y_n*). It is common in this context to allow for unique cluster-specific effects. Let* $W_i = (w_{i1}, \ldots, w_{in_i})'$ *be a* $n_i \times q$ *matrix of observations on q covariates* w_{ij} *whose effect on y is assumed to be cluster-specific. Also suppose that* X_{1i} *is an additional* $n_i \times k_1$ *matrix of observations on* k_1 *covariates whose effect on y is assumed to be non-cluster-specific (fixed effect). Then under the assumption that the observations across clusters are independent, a model for the outcomes is*

$$Y|\theta \sim \prod_{i=1}^{n} N_{n_i}(Y_i|X_{i1}\beta_1 + W_i\beta_{2i}, \sigma^2 I_{n_i}),$$

where the β_{2i} *are the cluster-specific effects. If the numbers of clusters is large, as is usual in practice, it is useful to assume that the effects* β_{2i} *have some structure. One possibility is to assume that the* β_{2i} *are drawn from a common distribution*

$$\beta_{2i}|Y \sim N_q(\beta_2, D)$$

independently across i. This is called the exchangeability assumption since the joint distribution of the β_{2i} *is invariant to permutation of the indices. Another possibility is the assumption that the* β_{2i} *are determined by a set of r cluster-specific covariates* a_i*:*

$$\beta_{2i}|Y \sim N_q(A_i\beta_2, D)$$

where

$$A_i = \begin{pmatrix} a_i' & 0' & \cdots & \cdots & 0' \\ 0' & a_i' & \cdots & \cdots & 0' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0' & 0' & \cdots & \cdots & a_i' \end{pmatrix},$$

$\beta_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2q})$ *is a* $k_2 = r \times q$*-dimensional vector and D, as in the first example, is a* $q \times q$ *matrix. Writing the second stage model in equivalent form as* $\beta_{2i} = A_i\beta_2 + b_i$, *where* $b_i | D \sim N_q(0, D)$, *and substituting this into the outcome model, it follows that the outcome model can be expressed as*

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}(y_i | X_i\beta + W_i b_i, \sigma^2 I_{n_i}),$$

where $\theta = (\beta, \sigma^2, b_1, \ldots, b_n)$, $X_i = (X_{1i} \vdots W_i A_i)$ *is a* $n_i \times k$ *matrix* ($k = k_1 + k_2$) *and* $\beta = (\beta_1, \beta_2)$. *The second stage of the model could now be specified as*

$$\theta | Y \sim N_k(\beta | \beta_0, B_0) IG(\sigma^2 | v_0 / 2, \delta_0 / 2) \prod_{i=1}^{n} N_q(b_i | 0, D).$$

Next suppose that there is enough prior information to fix $(\beta_0, B_0, v_0, \delta_0)$, *but that D (equivalently* D^{-1}*) cannot be fixed directly. Then* $Y = D^{-1}$. *A convenient assumption is*

$$Y | \lambda \sim \text{Wishart}_q(D^{-1} | \rho_0, R_0),$$

where

$$\text{Wishart}_q(D^{-1} | \rho_0, R_0) = c \frac{|D^{-1}|^{(\rho_0 - q - 1)/2}}{|R_0|^{\rho_0/2}} \exp\left\{-\frac{1}{2}\text{trace}(R^{-1}D^{-1})\right\}, |D^{-1}| > 0,$$

is the q-variate Wishart density,

$$c = \left(2^{\rho_0 q/2} \pi^{q(q-1)/4} \prod_{i=1}^{T} \Gamma\left(\frac{\rho_0 + 1 - i}{2}\right)\right)^{-1}$$

is its normalizing constant, and the stage 3 hyperparameters $\lambda = (\rho_0, R_0)$ *are known. Under these assumptions the full model is given by*

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}(y_i | X_i\beta + W_i b_i, \sigma^2 I_{n_i}) \tag{9}$$

$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0) IG(\sigma^2|\nu_0/2, \delta_0/2) \prod_{i=1}^{n} N_q(b_i|0, D)$$
(10)

$$\gamma|\lambda \sim \text{Wishart}_q(D^{-1}|\rho_0, R_0).$$
(11)

Putting a prior distribution on the hyperparameters γ in this way has several advantages. For one, it produces a prior distribution on θ that is less dogmatic than a prior based on specified hyperparameters since the resulting prior distribution of θ is averaged over the possible values of γ as dictated by the density $\psi(\gamma|\lambda)$:

$$\pi(\theta|\lambda) = \int \pi(\theta|\gamma)\psi(\gamma|\lambda)\,d\gamma.$$

If the hyperparameter γ is a scalar discrete quantity with support on the set $\{\gamma_1, \ldots, \gamma_G\}$, where *G* is potentially infinite, then the mixing density $\psi(\gamma|\lambda)$ is a probability mass function of the type $\sum_{j=1}^{G} p_j \delta_{\gamma_j}$, where δ_{γ_j} is the indicator function of γ_j, $0 \le p_j \le 1$ and $\sum_{j=1}^{G} p_j = 1$. The resulting conditional density $\pi(\theta|\lambda)$ is then a mixture of densities of the form

$$\pi(\theta|\lambda) = \sum_{j=1}^{G} p_j \pi(\theta|\gamma_j).$$

In this context, $\pi(\theta|\gamma_j)$ are called the component densities and p_j are the component weights. Such mixtures of component densities provide a simple mechanism for modelling θ in a flexible way.
Of course, one could have started at the outset with the prior $\pi(\theta|\lambda)$ by combining stages 2 and 3, leading to the collapsed model

$$y|\theta \sim p(y|\theta)$$
(12)

$$\theta|\lambda \sim \int \pi(\theta|\gamma)\psi(\gamma|\lambda)\,d\gamma,$$
(13)

which has the same structure as the standard 2-stage Bayesian model. This is not done, however, because the density of $\theta|\lambda$, even if tractable, is generally less easy to manage.

Example 3: (Gaussian linear regression model and Student-t prior). *Suppose that* $y = (y_1, \ldots, y_n)$ *is a vector of observations and* $\theta = (\beta, \sigma^2)$, *where* β *is a scalar regression parameter. Assume that*

$$y|\theta \sim N_n(y|X\beta, \sigma^2 I_n)$$
$$\theta|Y \sim N(\beta|\beta_0, B_0) IG\left(\sigma^2 | \frac{v_0}{2}, \frac{\delta_0}{2}\right)$$
$$B_0^{-1} \sim G\left(B_0^{-1} | \frac{v}{2}, \frac{v}{2}\right)$$

where G(·|·,·) is the gamma density and the quantities (β_0, v_0, δ_0) *and* ν *are known. Then the density of* β *marginalized over* B_0^{-1} *is Student-t,* $T(\beta|\beta_0, 1, v)$*, with location* β_0*, dispersion 1 and* ν *degrees of freedom. This Student-t prior density is not conjugate with the outcome model and therefore cumbersome to deal with.*

Bayesian hierarchical models can have additional stages. For instance, a further stage can be added by placing a prior density on λ, which leads to the model

$$y|\theta \sim p(y|\theta) \text{ (outcome model: stage 1)} \tag{14}$$

$$\theta|Y \sim \prod_{j=1}^{p} \pi(\theta_j|Y) \text{ (prior model: stage 2)} \tag{15}$$

$$Y|\lambda \sim \prod_{j=1}^{q} \psi(Y_j|\lambda) \text{ (hyperparameter model 1: stage 3)} \tag{16}$$

$$\lambda \sim \delta(\lambda) \text{ (hyperparameter model 2: stage 4)}, \tag{17}$$

where δ is the density of λ. Models with more than four stages are rare.

Posterior distributions

In a Bayesian analysis one is interested in deriving and summarizing the posterior distribution of θ given *y*. One obvious question concerns the form of this posterior distribution. Another question concerns the posterior distribution of the hyperparameters γ. Consider the canonical three-stage hierarchical model in (6)–(8). By Bayes's theorem,

$$\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{m(y)},$$

where $\pi(\theta) = \int \pi(\theta|\gamma)\pi(\gamma|\lambda)\,d\gamma$ and $m(y) = \int p(y|\theta)\pi(\theta)\,d\theta$, called the marginal likelihood, is the normalizing constant. Similarly, the posterior distribution of γ is

$$\pi(\gamma|y) = \frac{p(y|\gamma)\pi(\gamma|\lambda)}{m(y)},$$

where $p(y|\gamma) = \int p(y|\theta)\pi(\theta|\gamma)\,d\theta$. Before we discuss the tractability of these distributions we state a general result about how much information the data *y* supply about θ and γ beyond what is introduced by the prior densities $\pi(\theta)$ and $\pi(\gamma|\lambda)$. To measure this information we can use the Kullback–Leibler (KL) divergence measure, which, for any two densities *f* and *g*, is defined as

$$K(f, g) = E^f \log\frac{f}{g},$$

where E^f is the expectation with respect to the density *f*. The following result was proved by Goel and Degroot (1981). The result and proof can also be found in Lehmann and Casella (1998).

Theorem 1: For the three-stage hierarchical model,

$$K[\pi(\gamma|y), \pi(\gamma)] < K[\pi(\theta|y), \pi(\theta)].$$

This result states that the KL divergence between $\pi(\theta|y)$ and $\pi(\theta)$ is greater than between $\pi(\gamma|y)$ and $\pi(\gamma)$. In other words, the data supply more information about θ than they do about γ. Equivalently, the prior and the posterior of γ are closer than the prior and the posterior of θ. This implies that less learning is possible about the hyperparameters γ than about the parameters θ.

Much less can be said about the form of the posterior densities. In general, the posterior densities $\pi(\theta|y)$ and $\pi(\gamma|y)$ are not tractable. But if we consider the density of θ_j given (y, γ) and $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$, we have

$$\pi(\theta_j|y, \gamma, \theta_{-j}) \propto p(y|\theta)\pi(\theta_j|\gamma),$$

which is in closed form provided the prior density $\pi(\theta_j|\gamma)$ is conjugate with $p(y|\theta)$. The density $\pi(\theta_j|y, \gamma, \theta_{-j})$ is called the full conditional density of θ_j. Of course, the marginal density,

$$\pi(\theta_j|y) = \int \pi(\theta_j|y, \gamma, \theta_{-j})\pi(\gamma, \theta_{-j}|y)\,d\gamma\,d\theta_{-j},$$

where the mixing distribution is the marginal posterior distribution of (γ, θ_{-j}), is almost never available in closed form.

The same sort of difficulty arises in finding $\pi(\gamma|y)$. The problem is that the prior $\pi(\gamma|\lambda)$ generally does not combine with $p(y|\gamma)$ to produce a recognizable density. Nonetheless, just as in the case of θ_j, the calculations are easier if one considers the full conditional density of γ_j. To see this, note that

$$\pi(\gamma_j|y, \theta, \gamma_{-j}) \propto p(y|\theta)\pi(\theta|\gamma)\pi(\theta_j|\lambda) \propto \pi(\theta|\gamma)\pi(\theta_j|\lambda),$$

where the second line follows from the fact that the outcome model in stage 1 is free of γ. Thus, provided $\pi(\theta_j|\lambda)$ is conjugate with $\pi(\theta|\gamma)$, the full conditional density of γ_j can be derived in closed form.

Example 4: Consider again the clustered data model given in (9)–(11). The full conditional density of b_i is obtained as

$$\pi(b_i|y, \theta_{-b_i}, \Upsilon) \propto p(y|\theta)\pi(b_i|\Upsilon) \propto N_{n_i}(y_i|X_i\beta + W_i b_i, \sigma^2 I_{n_i}) N_q(b_i|0, D),$$

which, by standard Bayesian manipulations, is seen to be a $N_q(b_i|\hat{b}_i, B_i)$ *density, where*

$$\hat{b}_i = B_i(\sigma^{-2}W_i'(y_i - X_i\beta) \text{ and } B_i = (D^{-1} + \sigma^{-2}W_i' W_i)^{-1}.$$

Turning now to the full conditional density of D^{-1}, we obtain

$$\pi(D^{-1}|y, \theta) = \pi(D^{-1}|\{b_i\}) \propto \pi(\{b_i\}|D^{-1})\pi(D^{-1}|\lambda) \propto \prod_{i=1}^{n} N_q(b_i|0, D)\,\text{Wishart}_q(D^{-1}|\nu_0, R_0) = \text{Wishart}_q\left(D^{-1}|\rho_0 + n, \left(R_0^{-1} + \sum_{i=1}^{n} b_i b_i'\right)^{-1}\right),$$

where in the first line we have used the fact that the full conditional density of D^{-1} depends neither on y nor on β*; in the second line, Bayes's theorem; in the third line, substitutions for the needed densities; and in the fourth line, by observation that the product of the normal and Wishart prior densities is an updated Wishart distribution with the stated parameters*.

Computational issues

Difficulties in the computation of the marginal posterior densities of θ_j and γ_j were previously an impediment to the development and application of hierarchical Bayesian models. These difficulties have largely been resolved through the use of Markov chain Monte Carlo (MCMC) methods. These methods typically proceed by simulating the full conditional distributions, $\pi(\theta_j|y, \Upsilon, \theta_{-j})$ and $\pi(\Upsilon_j|y, \theta, \Upsilon_{-j})$. Under general conditions, the recursive simulation of these distributions produces a Markov chain whose limiting invariant distribution is the posterior density of interest, $\pi(\theta, \Upsilon|y)$.

Although it is not possible in this discussion to provide the theory behind MCMC methods, as outlined in Tierney (1994), and Chib and Greenberg (1995), or the range of hierarchical Bayes models that have been thus processed, it is useful to illustrate the computations with the help of the simplest MCMC method, the so-called Gibbs sampling algorithm. This algorithm was introduced by Geman and Geman (1984) in the context of image processing, but the papers of Tanner and Wong (1987) and Gelfand and Smith (1990) brought it into the limelight.

Suppose that the various blocks $\{\theta_j\}$ and $\{\Upsilon_j\}$ are chosen to ensure that the associated set of full conditional densities $\{\pi(\theta_j|y, \theta_{-j}, \Upsilon)\}$ and $\{\pi(\Upsilon_j|y, \theta, \Upsilon_{-j})\}$ are all tractable. Then one cycle of the Gibbs sampling algorithm is completed by simulating $\{\theta_j\}$ and $\{\Upsilon_j\}$ from each full conditional distribution, recursively updating the conditioning variables while moving through the set of distributions. The Gibbs sampler in which each block is revised in fixed order is defined as follows.

Algorithm:: **Gibbs sampling**

1. 1. Specify an initial value $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ and $\Upsilon^{(0)} = (\Upsilon_1^{(0)}, \ldots, \Upsilon_q^{(0)})$.
2. 2. Repeat for j=1, 2, …, n_0+M:
   - Generate $\theta_1^{(j)}$ from $\pi(\theta_1|y, \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, \Upsilon^{(j-1)})$
   - Generate $\theta_2^{(j)}$ from $\pi(\theta_2|y, \theta_1^{(j)}, \theta_3^{(j-1)} \ldots, \theta_p^{(j-1)}, \Upsilon^{(j-1)})$
   - ⋮
   - Generate $\theta_p^{(j)}$ from $\pi(\theta_p|y, \theta_1^{(j)}, \theta_2^{(j)} \ldots, \theta_{p-1}^{(j)}, \Upsilon^{(j-1)})$
   - Generate $\Upsilon_1^{(j)}$ from $\pi(\Upsilon_1|y, \theta^{(j)}, \Upsilon_2^{(j-1)}, \ldots, \Upsilon_q^{(j-1)})$
   - Generate $\Upsilon_2^{(j)}$ from $\pi(\Upsilon_2|y, \theta^{(j)}, \Upsilon_1^{(j)}, \Upsilon_3^{(j-1)} \ldots, \Upsilon_q^{(j-1)})$
   - ⋮

- Generate $Y_q^{(j)}$ from $\pi(Y_q|y, \theta^{(j)}, Y_1^{(j)}, \ldots, Y_{q-1}^{(j)})$.

3. 3. Return the values $\left\{\theta^{(n_0+1)}, Y^{(n_0+1)}, \theta^{(n_0+2)}, Y^{(n_0+2)}, \ldots, \theta^{(n_0+M)}, Y^{(n_0+M)}\right\}$.

Thus, in this algorithm, block θ_j is generated from the full conditional distribution

$$\pi(\theta_j|y, \theta_1^{(j)}, \ldots, \theta_{j-1}^{(j)}, \theta_{j+1}^{(j-1)}, \ldots, \theta_p^{(j-1)}, Y^{(j-1)}),$$

where the conditioning elements for the *j*th block reflect the fact that the previous (*j*–1) blocks of θ have already been updated, but the rest have not been. Note that the output from the first n_0 cycles (the burn-in phase) is ignored to allow the effect of the initial values to wear off.

One additional point about MCMC methods is that the blocks must be carefully chosen. Sampling over unnecessary blocks can worsen the quality of the output produced by the algorithm, where quality is measured by how quickly the serial correlations of the sampled draws decline to zero. Chains whose serial correlations decline quickly are preferred because they are closer to the ideal of independent sampling.

Example 5: Consider again the hierarchical Bayesian model for clustered data given in (9)–(11). The joint distribution of the data and the unknowns is given by

$$p(y, \theta, D^{-1}) = \pi(\beta, \sigma^2, \{b_i\}, D^{-1})\, p(y|\theta) = \pi(\beta)\pi(\sigma^2)\pi(D^{-1})\prod_{i=1}^{n} p(y_i|\theta)\pi(b_i|D). \tag{18}$$

Wakefield et al. (1994) propose a Gibbs MCMC approach for joint distribution that is based on full blocking (that is, sampling each block of parameters from its full conditional distribution). Chib and Carlin (1999) suggest a number of reduced blocking schemes. One of the simplest proceeds by first sampling β marginalized over $\{b_i\}$ and then sampling $\{b_i\}$ conditioned on β. This reduced blocking is possible because b_i in (18) can be marginalized out leaving a normal distribution that can be combined with the assumed normal prior on β. In particular,

$$p(y_i|\beta, \sigma^2, D) = \int p(y_i|\theta)\pi(b_i|D)\,\mathrm{d}b_i \propto |V_i|^{-1/2}\exp\left\{(-1/2)(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)\right\},$$

where $V_i = \sigma^2 I_{n_i} + W_i D W_i'$. The reduced conditional posterior of β is therefore

$$\pi(\beta|y, \sigma^2, D) \propto \pi(\beta)\prod_{i=1}^{n}|V_i|^{-1/2}\exp\left\{-\frac{1}{2}(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)\right\} \propto \exp\left\{-\frac{1}{2}(\beta - \hat{\beta})'B^{-1}(\beta - \hat{\beta})\right\},$$

where

$$\hat{\beta} = B\left(B_0^{-1}\beta_0 + \sum_{i=1}^{n} X_i' V_i^{-1} y_i\right) \text{ and } B = \left(B_0^{-1} + \sum_{i=1}^{n} X_i' V_i^{-1} X_i\right)^{-1}.$$

The rest of the MCMC algorithm follows the steps of Wakefield et al. (1994). In full, we sequentially sample the following distributions many times:

$$\beta \sim N_K(\hat{\beta}, B)$$

$$b_i \sim N_q\left(D_i(\sigma^{-2}W_i'(y_i - X_i\beta), D_i = (D^{-1} + \sigma^{-2}W_i'W_i)^{-1}\right), \qquad i \le n$$

$$\sigma^2 \sim IG\left(\frac{v_0 + \Sigma n_i}{2}, \frac{\delta_0 + \Sigma_{i=1}^{n} \| y_i - X_i\beta - W_i b_i \|^2}{2}\right)$$

$$D^{-1} \sim \text{Wishart}_q\left\{\rho_0 + n, \left(R_0^{-1} + \sum_{i=1}^{n} b_i b_i'\right)^{-1}\right\},$$

where the second and fourth of these distributions were derived in Example 4.

Model choice

Another inferential concern in practice is the comparison of several hierarchical Bayesian models in order to judge the extent to which the various models are supported by the data. In the context of a hierarchical model for clustered data, for instance, one may be interested in determining the support for an additional cluster-specific effect or of an additional fixed effect. Questions of this type can be answered via *Bayes factors*, or ratios of *marginal likelihoods*. The marginal likelihood of a particular model $\mathcal{M}$ is the normalizing constant of the posterior density,

$$m(y|\mathcal{M}) = \int p(y|\mathcal{M}, \theta)\pi(\theta|\mathcal{M}, \gamma)\pi(\gamma|\mathcal{M}, \lambda)\,\mathrm{d}\theta\,\mathrm{d}\gamma, \tag{19}$$

the integral of the first stage outcome density function with respect to the prior density of θ and the prior density of the hyperparameters γ. If there are two models $\mathcal{M}_k$ and $\mathcal{M}_l$, the Bayes factor is the ratio

$$\mathrm{B}_{kl} = \frac{m(y|\mathcal{M}_k)}{m(y|\mathcal{M}_l)}. \tag{20}$$

Because MCMC methods deliver draws from the posterior density and the marginal likelihood is the integral with respect to the prior $\pi(\theta|\mathcal{M}, \gamma)\pi(\gamma|\mathcal{M}, \lambda)$, MCMC output cannot be used directly to average $p(y|\mathcal{M}, \theta)$. Nonetheless, computation is feasible by the method of Chib (1995), a widely used method that we now briefly describe.

Chib (1995) begins by noting that $m(y|\mathcal{M}, \lambda)$ can be expressed as

$$m(y|\mathcal{M}) = \frac{p(y|\mathcal{M}, \theta^*)\pi(\theta^*|\mathcal{M}, \gamma^*)\pi(\gamma^*|\mathcal{M}, \lambda)}{\pi(\theta^*, \gamma^*|\mathcal{M}, y)},$$
(21)

for a given (θ^*, γ^*), usually taken to be a high density point such as the posterior mean. Thus, if we have an estimate $\hat{\pi}(\theta^*, \gamma^*|\mathcal{M}, y)$ of the posterior ordinate, the marginal likelihood on the log scale can be estimated as

$$\log m(y|\mathcal{M}) = \log p(y|\mathcal{M}, \theta^*) + \log \pi(\theta^*|\mathcal{M}, \gamma^*) + \log \pi(\gamma^*|\mathcal{M}, \lambda) - \log \hat{\pi}(\theta^*, \gamma^*|\mathcal{M}, y).$$
(22)

It turns out that it is possible to get an efficient estimate of the posterior ordinate. The basic idea is to write the posterior ordinate as

$$\pi(\theta^*, \gamma^*|\mathcal{M}, y) = \pi(\theta_1^*|\mathcal{M}, y) \times \cdots \times \pi(\theta_p^*|\mathcal{M}, y, \theta_1^*, \ldots, \theta_{p-1}^*)$$

$$\times\, \pi(\gamma_1^*|\mathcal{M}, y, \theta^*) \times \cdots \times \pi(\gamma_q^*|\mathcal{M}, y, \theta^*, \gamma_1^*, \ldots, \gamma_{p-1}^*)$$
(23)

and then to estimate each of these ordinates from the output of appropriate MCMC runs. To see what is involved, consider the ordinate $\pi(\theta_j^*|\mathcal{M}, y, \theta_1^*, \ldots, \theta_{j-1}^*)$ that appears in this decomposition. By definition,

$$\pi(\theta_j^*|\mathcal{M}, y, \theta_1^*, \ldots, \theta_{j-1}^*) = \int \pi(\theta_j^*|y, \theta_1^*, \ldots, \theta_{j-1}^*, \theta_{j+1}, \ldots, \theta_p, \gamma)\, d\pi(\theta_{j+1}, \ldots, \theta_p, \gamma|y, \theta_1^*, \ldots, \theta_{p-1}^*)$$

is the full conditional density integrated with respect to the distribution $\pi(\theta_{j+1}, \ldots, \theta_p, \gamma|y, \theta_1^*, \ldots, \theta_{p-1}^*)$. To calculate this integral by Monte Carlo one can run an MCMC algorithm in which the blocks $(\theta_1, \ldots, \theta_{p-1})$ are fixed at their starred values and sampling is over the remaining free blocks, namely $(\theta_j, \theta_{j+1}, \ldots, \theta_p, \gamma)$. This is called a *reduced* MCMC run. Let the sampled draws from this reduced run be denoted by $(\theta_{j+1}^{(r)}, \ldots, \theta_p^{(r)}, \gamma^{(r)})$, r=1, …, M. Then, provided the full conditional of θ_j is in closed form, we have the estimate

$$\hat{\pi}(\theta_j^*|\mathcal{M}, y, \theta_1^*, \ldots, \theta_{j-1}^*) = M^{-1}\sum_{r=1}^{M} \pi(\theta_j^*|y, \theta_1^*, \ldots, \theta_{j-1}^*, \theta_{j+1}^{(r)}, \ldots, \theta_p^{(r)}, \gamma^{(r)}).$$

Each ordinate is estimated in this way from the output of the appropriate reduced runs. Notice that as more blocks are fixed, fewer distributions appear in the reduced runs.

Example 6: Consider again the hierarchical Bayesian model for clustered data. In this case, we can decompose $\pi(\theta^*, \gamma^*|\mathcal{M}, y)$ *as*

$$\pi(D^{-1*}, \sigma^{2*}, \beta^*|y) = \pi(D^{-1*}|y)\pi(\sigma^{2*}|y, D^*)\pi(\beta^*|y, D^*, \sigma^{2*}),$$

so that all computations are marginalized over $\{b_i\}$. *The first term can be estimated by averaging the Wishart density given in Example 5 over draws on* $\{b_i\}$ *from the full MCMC run. To estimate the second ordinate, which is conditioned on* D^*, *we run a reduced MCMC simulation with the full conditional densities*

$$\pi(\beta|y, D^*, \sigma^2),\ \pi(\sigma^2|y, \beta, D^*, \{b_i\}),\ \pi(\{b_i\}|y, \beta, D^*, \sigma^2),$$

where each conditional utilizes the fixed value of D. The second ordinate is now estimated by averaging the inverse gamma full conditional density of σ^2 *at* σ^{2*} *over the draws on* $(\beta, \{b_i\})$ *from this reduced run. The third ordinate is multivariate normal as given in Example 5 and available directly*.

If the full conditional densities are not in closed form, the marginal likelihood can be computed by the modified Chib method as discussed in Chib and Jeliazkov (2001).

See Also

- Bayesian econometrics
- Bayesian statistics
- econometrics
- fixed effects and random effects
- longitudinal data analysis
- Markov chain Monte Carlo methods
- model selection
- simulation-based estimation
- statistical inference

Bibliography

Albert, J.H. and Chib, S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88, 669–79.

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 1313–21.

Chib, S. and Carlin, B.P. 1999. On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing* 9, 17–26.

Chib, S. and Greenberg, E. 1995. Understanding the metropolis-Hastings algorithm. *American Statistician* 49, 327–35.

Chib, S. and Jeliazkov, I. 2001. Marginal likelihood from the metropolis-Hastings output. *Journal of the American Statistical Association* 96, 270–81.

Gelfand, A.E. and Smith, A.F.M. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–41.

Goel, P.K. and Degroot, M.H. 1981. Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association* 76, 140–7.

Lehmann, E. and Casella, G. 1998. *Theory of Point Estimation*. New York: Springer.

Lindley, D.V. and Smith, A.F.M. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B* 34, 1–41.

Tanner, M.A. and Wong, W.H. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82, 528–50.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 21, 1701–62.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A. and Gelfand, A.E. 1994. Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Applied Statistics* 43, 201–21.

How to cite this article

Chib, Siddhartha and Edward Greenberg. "hierarchical Bayes models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000168> doi:10.1057/9780230226203.0732

The New Palgrave Dictionary of Economics Online

hierarchy

Luis Garicano

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Hierarchies lighten the burden of the enormous informational requirements of the price system under uncertainty by acquiring more knowledge and information than any individual can. They are thus are useful for information handling. They can also allow agents to engage in collective actions by decreasing the risk of opportunistic behaviour. But trade-offs are involved because hierarchies impose costs, including communication among agents. This article reviews the literature on this trade-off and its implications for labour markets.

Keywords

bounded rationality; earnings distribution; firm size; hierarchy; incentive conflict; information economics; monitoring theories; principal and agent; queuing theory; uncertainty

Article

Hierarchy deals with individuals' bounded rationality by allowing for more information to be used in decision-making than individual agents could possibly use and by allowing the most skilled agents to leverage their knowledge with the help of others. Hierarchy can also allow agents to engage in collective actions by decreasing the risk of opportunistic behaviour. These benefits of hierarchy do not come without costs. Hierarchies may be slow to react, may introduce noise into the communication process, and generally may require costly communication among agents. This entry reviews the literature on these trade-offs in multiple-layer, multi-agent hierarchies; that is, it leaves aside the simplest, one principal one agent-type model.

Processing information

Arrow (1974) first observed that the enormous informational requirements of the price system under

uncertainty (complete markets require one state-contingent price per commodity per state of the world) place a bound on its performance. A key role of hierarchy is to lighten this burden: hierarchies can acquire more information than any individual can, and thus are useful for information handling.

A large literature has explored this role of hierarchies in enhancing information processing. An example of this class of models is Radner (1993). Decision-making requires observing a linear combination of certain variables, and agents incur some cost of performing additive operations and some cost of communicating their results. All information must be processed. Under these conditions, organizations are asymmetric, to ensure that all agents are occupied. Bolton and Dewatripont (1994) extend this model to the case where cohorts of data arrive all the time; in this case, the optimal structures are balanced trees, and look more like the ones we arguably observe in reality. Radner and Van Zandt (1992), Van Zandt (1999) and Van Zandt and Radner (2001) take a further step by studying the problem of processing in real time, when the information relevant to a given decision is continuously arriving. The key objects of interest are the sign and size of the scale diseconomies resulting from hierarchy. That is, these authors aim to answer the question of the extent to which diseconomies of scale linked to human bounded rationality are the reason we see many firms, rather than one. The answer is not unambiguous. For example, Radner and Van Zandt (1992) find that returns to scale can vary from increasing to sharply decreasing, depending on the correlation of the data and on the cost of incorrect decisions. Vayanos (2003) extends substantially these models beyond associative operations, and considers situations with two realistic characteristics: the decisions of different agents interact; and the aggregation process entails information losses.

A separate branch of the literature, following Crémer (1980), has studied hierarchical resource allocation programmes under limited managerial processing power. Geanakoplos and Milgrom (1991) study a hierarchy in which managers can invest in information collection, but each manager can collect a limited amount of information. By decomposing hierarchically the allocation problem (so that a low-level manager allocates resources among shops, while lower-level managers allocate resources among groups of sources, and so on) the total amount of information used can be increased. Each manager is told by his superior how many resources he gets, and communicates that information to each subordinate. Managers aim to minimize the expected total costs of their units (there are no externalities). Under these conditions, the number of managers used is increasing in value of information and U-shaped in managerial ability (few managers are used if they are unproductive or if they are so productive that a few can achieve all savings). A more uncertain environment increases the number of managers needed and their average skill, and causes a decrease in their span of control.

Organizing knowledge

In Garicano (2000) a hierarchy, rather than a means to aggregate information, is a means to acquire and conserve experts’ knowledge. He considers a set of agents who face a large number of problems. They may or may not invest in learning their solution; they produce only if they do. Some problems are more common than others, and there is an *ex ante* known probability distribution of problem. Agents can ask other agents for help in solving their problems, but, crucially, they do not know who knows what. Garicano shows that an optimal organization has agents specializing in either production or problem solving; that production workers deal with routine problems and problem solvers specialize in the

exceptions; and that shape is pyramidal, with fewer agents in each successive layer. The organizing principle is management by exception. The key organizational trade-off is between acquiring knowledge and asking; that is, an extra hierarchical tier increases communication costs but also increases the utilization of expertise and results in lower knowledge acquisition costs. Given this trade-off, an increase in the cost of communication leads agents to learn more and ask less, and managers to learn less and deal with a smaller proportion of problems. Conversely, when the cost of acquiring knowledge rises, the role of the hierarchy increases as managers deal with a larger fraction of problems.

Beggs (2001) also investigates the phenomenon of ‘management by exception’, although with exogenous knowledge. He uses queuing theory to explore the optimal allocation of workers with exogenously given skills to the different layers of a hierarchy.

This type of organization of work is common in many contexts; for example, in law firms (Garicano and Hubbard, 2007) or in medicine. In this professional context, the role of the ‘juniors’ (associates, residents) is to handle the easier problems to conserve the valuable time of the seniors (attending physicians, partners) for the harder problems. Similarly, in a team engaged in technical support (Orlikowski, 1996), experts must answer customer calls, and production is organized so that juniors handle front calls, and transfer the calls they cannot handle to more senior experts.

Hierarchical allocation of talent to positions: the distribution of earnings

Another line of research has explored the relation between the distributions of income and the distributions of firm size and hierarchy. This literature has proposed that the reason why the distribution of income is more skewed than the underlying distribution of skills lies in how resources are allocated to individuals. Higher-ability managers raise the productivity of the resources they are assigned more than lower-ability managers. As a result, in equilibrium, more able managers are allocated more resources, and this leads the marginal value of their ability to increase faster than if they were working on their own. Lucas (1978) and Rosen (1982) generate full equilibrium models that yield both an equilibrium firm size and distribution of earnings. In both these papers, the manager increases the productivity under his control, which, depending on the model, may be the number of workers (Lucas, 1978), or efficiency units of labour, that is, total units of skill managed (Rosen, 1982). In these models, managerial human capital raises the marginal product of the workers or capital they are assigned, but managers’ span of control is generally limited implicitly or explicitly by managers’ time. Equilibrium assignment patterns involve scale of operations effects, which follow from the complementarity between managerial human capital and productive resources. The main equilibrium result from this class of models is that these production functions involve scale-of-operations effects: more skilled managers are assigned more resources to manage in equilibrium. As a result, the distribution of earnings is more skewed than the distribution of skills. Garicano and Rossi-Hansberg (2005) build on this line of research, but study a model of hierarchy with heterogeneous agents that extends Garicano (2000), and which involves matching between managers and workers – that is, managers do not care only about the efficiency units they manage (which would imply that a top lawyer at a law firm should be indifferent between managing two good associates or a large number of mediocre ones), but instead care about both the quality and quantity of workers. The model generates a continuum of hierarchies and an equilibrium allocation of workers to positions, as well as the income distributions. It allows for the simultaneous

exploration of changes in organization and in wage structure, and has been applied to issues such as the formation of cross-country teams (Antràs, Garicano and Rossi-Hansberg, 2006), or changes in organization and the wage structure as a result of the information technology revolution (Garicano and Rossi-Hansberg, 2005).

Monitoring and authority

An alternative class of theories study managers as agents able to fire underperforming agents or otherwise exercise their authority. Monitoring theories stem from Alchian and Demsetz (1972), who posit that hierarchies are a response to incentive problems associated with team production. In this view, lower-level individuals are directly involved in production, and upper-level individuals are specialized monitors. The view was elaborated formally by Calvo and Weillisz (1978) and Qian (1994). Their basic assumption is that supervision is necessary for ensuring performance. They study an efficiency wage setting like Becker and Stigler's (1974), where agents can work full-time and earn w or shirk and be detected and fired with probability p, in which case they earn their reservation utility. Here, the principal can induce work by increasing the monitoring intensity p through hierarchical supervision. The hierarchy then trades off the gains due to these lower wages against the cost of the supervisors.
Aghion and Tirole (1997) formally introduce the idea of decision-making agents into the study of hierarchy with incentive conflicts. Delegation by a superior functions as a commitment not to intervene, and as such delegating authority increases incentives for agents to invest. Baker, Gibbons and Murphy (1999) extend such analysis to a context where delegation is in fact a relational contract in which the centre chooses not to exercise its power. Rajan and Zingales (2001) study how the shape and size of the hierarchy responds to the problem of providing incentives for employees to protect the resources of the entrepreneur and discouraging them from stealing them. Finally, Hart and Moore (2005) consider hierarchies as chains of authority that determine priority in decisions over asset allocation, and derive conditions where optimal hierarchies have generalist coordinators on top. Their theory helps explain why generalists – individuals who know about the interactions between classes of assets – should be senior to specialists.
Overall the models reviewed in this article have the potential to address an important missing link in economic theory: the absence of managers, and of occupations, from both the theory of the firm and the theory of the determination of wages.

See Also

- implicit contracts
- incomplete contracts
- information aggregation and prices

Bibliography

Aghion, P. and Tirole, J. 1997. Formal and real authority in organizations. *Journal of Political Economy* 105, 1–29.

Alchian, A. and Demsetz, H. 1972. Production, information costs, and economic organization. *American Economic Review* 62, 777–95.

Antràs, P., Garicano, L. and Rossi-Hansberg, E. 2006. Offshoring in a knowledge economy. *Quarterly Journal of Economics* 121 (forthcoming).

Arrow, K. 1974. *The Limits of Organization*. New York: W.W. Norton.

Baker, G., Gibbons, R. and Murphy, K. J. 1999. Informal authority in organizations. *Journal of Law, Economics & Organization* 15, 56–73.

Becker, G. and Stigler, G. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3, 1–18.

Beggs, A. Queues and hierarchies. 2001. *Review of Economic Studies* 68, 297–322.

Bolton, P. and Dewatripont, M. 1994. The firm as a communication network. *Quarterly Journal of Economics* 109, 809–39.

Calvo, G. and Wellisz, S. 1978. Supervision, loss of control, and the optimum size of the firm. *Journal of Political Economy* 86, 943–52.

Crémer, J. 1980. A partial theory of the optimal organization of a bureaucracy. *Bell Journal of Economics* 11, 683–93.

Garicano, L. 2000. Hierarchies and the organization of knowledge in production. *Journal of Political Economy* 108, 874–904.

Garicano, L. and Hubbard, T. 2007. Managerial leverage is limited by the extent of the market: hierarchies, specialization and the utilization of lawyers' human capital. *Journal of Law and Economics* (forthcoming).

Garicano, L. and Rossi-Hansberg, E. 2005. Organization and inequality in a knowledge economy. Working Paper No. 11458. Cambridge, MA: NBER.

Geanakoplos, J. and Milgrom, P. 1991. A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and International Economies* 5, 205–25.

Hart, O. and Moore, J. 2005. On the design of hierarchies: coordination vs. specialization. *Journal of Political Economy* 113, 675–702.

Lucas, R. 1978. On the size distribution of business firms. *Bell Journal of Economics* 9, 508–23.

Orlikowski, W. 1996. Improvising organizational transformation over time: a situated change perspective. *Information Systems Research* 7, 63–92.

Qian, Y. 1994. Incentives and loss of control in an optimal hierarchy. *Review of Economic Studies* 61, 527–44.

Radner, R. 1993. The organization of decentralized information processes. *Econometrica* 61, 1109–46.

Radner, R. and Van Zandt, T. 1992. Information processing in teams and returns to scale. *Annales d'Economie et de la Statistique* 25/26, 265–98.

Rajan, R. and Zingales, L. 2001. The firm as a dedicated hierarchy: a theory of the origins and growth of firms. *Quarterly Journal of Economics* 116, 805–51.

Rosen, S. 1982. Authority, control, and the distribution of earnings. *Bell Journal of Economics* 13, 311–23.

Van Zandt, T. 1999. Real-time decentralized information processing as a model of organizations with boundedly rational agents. *Review of Economic Studies* 66, 633–58.

Van Zandt, T. and Radner, R. 2001. Real-time decentralized information processing and returns to scale. *Economic Theory* 17, 497–544.

Vayanos, D. 2003. The decentralization of information processing in the presence of interactions. *Review of Economic Studies* 70, 667–95.

How to cite this article

Garicano, Luis. "hierarchy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000058> doi:10.1057/9780230226203.0733

The New Palgrave Dictionary of Economics Online

Hildebrand, Bruno (1812 – 1878)

Hermann Reich

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; communism; Engels, F.; German Historical School; Hildebrand, B.; List, F.; Müller, A.; self-interest; Smith, A.; socialism; statistics

Article

Hildebrand was born in Naumburg (Thuringia), the son of a clerk to the court. He studied in Leipzig and Breslau. In 1841 he was promoted full professor of Staatswissenschaften (of government, which included political economy) at the University of Marburg.

Hildebrand had always been an activist in the liberal and patriotic movement. He faced political persecution before the 1848 revolution, during which he was elected deputy of the Frankfurt National Assembly. In the subsequent period of restoration he was forced to emigrate to Switzerland, where he became not only a professor but also the director of a railway company, and founded the first Swiss statistical office (at Berne). In 1861 he was appointed professor at the University of Jena. He was founder (in 1862) and editor of the *Jahrbücher für Nationalökonomie and Statistik* and contributed to the establishment of the statistical office of the United Thuringian States (in 1864).

Hildebrand is considered as one of the founders of the German Historical School. He was opposed to the deductive method of the classicals and denied the existence of ‘natural laws’ in economic life (1863). His most important work was *Die Nationalökonomie der Gegenwart und Zukunft* (1848), where he discussed the theories of Friedrich List, Adam Müller, and especially those of Adam Smith. With his sharp criticism of self-interest and egoism as the central determinant of Smith's economic system – and the emphasis on ethical principles and the historically changing patterns of economic development – Hildebrand launched the attacks on Smith and the classical economists that were subsequently continued by many German historical economists.

The largest part of his main work was devoted to a discussion of socialism and communism, which he sharply rejected. Hildebrand focused his attention on the then little known Friedrich Engels and his recently published *Conditions of the Working Class in England* ([1848], 1922, pp. 125–90). He particularly criticized Engel's euphemistic description of pre-industrial conditions and contrasted it with

empirical data that showed quite a different picture.

While being aware of current social problems Hildebrand perceived capitalist development most optimistically and envisioned as its last stage of development – the so-called ‘credit economy’ – a society where an advanced banking system would provide credit to a worker according to his morals and character and where thereby the monopoly of the capitalist class on capital would be broken (Hildebrand [1864], 1922, pp. 354–5). This theory of stages has to be regarded as Hildebrand's capitalist utopia, his liberal answer to socialism and communism.

Hildebrand's importance and his influence on the German Historical School has generally been underestimated; after all Hildebrand was – as Max Weber remarked – the only one really to work with the historical method. He undertook statistical studies – he regarded statistics as an important tool for detailed historical and empirical research (1865) – and wrote historical monographs (1866). He thus anticipated much of the research programme of the ‘younger historical school’ and the Verein für Socialpolitik, which he joined – as the only economist of the ‘older historical school’ – as a charter member in 1873.

Hildebrand stood for a kind of progressive liberalism that intended to reshape Germany along the lines of England, which he admired.

See Also

- Historical School, German

Selected works

1848. *Die Nationalökonomie der Gegenwart und Zukunft und andere gesammelte Schriften*. Ed. H. Gehrig, Jena: Gustav Fischer, 1922. It contains the articles Die gegenwärtige Aufgabe der Wissenschaft der Nationalökonomie (1863), *Die wissenschaftliche Aufgabe der Statistik* (1865), and *Natural-, Geld- und Kreditwirtschaft* (1864).

1866. Zur Geschichte der deutschen Wollenindustrie. *Jahrbücher für Nationalökonomie und Statistik* 6, S186–S254; 7, S81–S153.

How to cite this article

Reich, Hermann. "Hildebrand, Bruno (1812–1878)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000064> doi:10.1057/9780230226203.0734

The New Palgrave Dictionary of Economics Online

Hilferding, Rudolf (1877 – 1941)

Roy Green
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

banking industry; Bernstein, E.; Böhm-Bawerk, E. von; finance capital; Hilferding, R.; Kautsky, K.; labour theory of value; Lenin, V. I.; Luxemburg, R.; Marx, K. H.; organized capitalism; social democracy; socialism

Article

Hilferding blended Marxist economics and Social Democratic politics in a career cut tragically short by the rise of fascism in Germany. He studied medicine at the University of Vienna, but soon showed more interest in organizing the student socialist society. After graduating in 1901, he helped Max Adler to found the *Marx-Studien* (1904–23), a series which was to become the theoretical flagship of ‘Austro-Marxism’. The first volume contained a vigorous defence of the labour theory of value by Hilferding himself against Böhm-Bawerk's marginalist critique, *Zum Abschluss des Marxschen Systems* (1896). It earned him his intellectual spurs in the German-speaking socialist movement.

At the same time, Hilferding was already contributing to debate within the German Social Democratic Party (SPD) through its journal, *Die Neue Zeit*. There, on the controversial ‘mass strike’ issue, he steered a course for the party leadership between Eduard Bernstein's ‘revisionist’ abandonment of the socialist goal and Rosa Luxemburg's revolutionary commitment to it (1903/4; 1904/5). He was rewarded with an appointment in 1906 as economics lecturer at the party school in Berlin, and then as foreign editor of the party newspaper, *Vorwärts*. From 1907, he also wrote regularly for the newly established journal of the Austrian Social Democrats, *Der Kampf*.

Hilferding published his major work, *Das Finanzkapital*, in 1910; it was immediately hailed by such diverse figures as Kautsky (1911), Lenin (1916) and Bukharin (1917), as a path-breaking development of Marxist economic analysis. Essentially, Hilferding argued that the concentration and centralization of capital had led to the domination of industry and commerce by the large banks, which were transformed into ‘finance capital’ (1910, p. 225). The socialization of production effected by finance capital required a correspondingly increased economic role for the state. Society could therefore plan production by using the state to control the banking system:

> The socializing function of finance capital facilitates enormously the task of overcoming capitalism. Once finance capital has brought the most important branches of production under its control, it is enough for society, through its conscious executive organ – the state conquered by the working class – to seize finance capital in order to gain immediate control of these branches of production … . Even today, taking possession of six large Berlin banks would mean taking possession of the most important spheres of large-scale industry … (1910, pp. 367–8)

This chain of reasoning, however, tended to exaggerate not only the leverage of the banks over industry, but also the role of the state in the organization of production. While it convinced Hilferding that socialism could be introduced by a determined majority in parliament, it demonstrated to Lenin that socialism would not be possible unless the state was ‘overthrown’ by a determined minority outside parliament. Their common point of reference was the centrality of the state – rather than society – in the ‘latest phase of capitalist development’. It forced socialists to make a choice between parliamentarism and insurrection, the very nature of which contributed to the defeat of the labour movement in Germany and the rise of party dictatorship in Russia (Neumann, 1942, pp. 13–38). Although theory cannot be held responsible for the course of history, it may influence political judgements which tip the balance at decisive moments. Hilferding's generation lived through many such moments.

When war broke out in 1914, Hilferding associated himself with the SPD minority which voted against war credits and which later formed the Independent Social Democrats (USPD). He spent most of the war on the Italian front, having been drafted into the Austrian army as a doctor, and returned to Berlin as editor of the USPD journal, *Freiheit*. Hilferding successfully opposed USPD affiliation to the Third International; his speech against Zinoviev at the Halle conference of 1920 – published under the title, ‘Revolutionäre Politik oder Machtillusionen?’ – was a decisive turning point. Once the embryonic Communist Party (KPD) forced a split on the issue, however, he saw no alternative to reunification with the remnants of the SPD.

During the 1920s, Hilferding turned his attention almost entirely to the political and economic problems facing the new German republic. He was a leading member of the Reich Economic Council, twice minister of finance and an active participant in the discussions on ‘workers’ councils’ and the government's ‘socialization’ programme. Hilferding's first stint as minister of finance lasted only seven weeks in the Stresemann government of 1923. Although he had no opportunity to implement his proposals, he devised a plan for currency reform involving the introduction of a *Rentenmark* backed by gold as part of an anti-inflation package. By the time Hilferding returned to the same post in the Müller government of 1928/9, economic conditions had worsened; his predicament was appreciated by Schumpeter who wrote, ‘we now have a socialist minister who faces the exceptionally difficult task of curing or improving a situation bequeathed by non-socialist financial policies’ (quoted in Gottschlacht, 1962, p. 24). A less sympathetic observer, however, portrayed Hilferding at this time as ‘the theorist of coalition politics in the period of capitalist stabilisation’ (see Gottschlacht, 1962, p. 204), blinded by theory to the imminent fascist danger.

Pursuing the logic of *Das Finanzkapital*, Hilferding had developed a theory of ‘organized capitalism’, a term he first used in 1915 in *Der Kampf*, and then explained more fully in 1924 in *Die Gesellschaft*. He

summarized the approach at the SPD's Kiel conference in 1927: ‘Organized capitalism means replacing free competition by the social principle of planned production. The task of the present Social Democratic generation is to invoke state aid in translating this economy, organized and directed by the capitalists, into an economy directed by the democratic state’ (see Neumann, 1942, p. 23). Ironically, this was the very position of an earlier Social Democratic leadership which Marx had singled out for criticism. Commenting on the demand for a ‘free state’ in the 1875 Gotha programme, Marx wrote:

> It is by no means the goal of workers who have discarded the mentality of humble subjects to make the state ‘free’. In the German Reich the ‘state’ has almost as much ‘freedom’ as in Russia. Freedom consists in converting the state from an organ superimposed on society into one thoroughly subordinate to it; and even today state forms are more or less free depending on the degree to which they restrict the ‘freedom of the state’. (Marx, 1891, p. 354)

While Hilferding understood that in capitalist society power lay with capital and was exercised by the representatives of capital in the management structure of the great corporations, he failed to see that democratic control over the productive forces would require a change in the relationship of power *within* the corporation itself. Organized labour could use the state to accelerate this process of social transformation and to create the centralized institutional machinery necessary for the ‘associated producers’ to plan directly the whole economy; but the notion that the state itself could perform this task rested upon an illusion. In attempting to replace the domination of capitalist employers with the domination of a ‘democratic state’, Hilferding and the party leadership achieved only one practical result: ‘Unwittingly, they strengthened the monopolistic trends in German industry’ (Neumann, 1942, p. 21). The state domination which followed was far from democratic.

Hilferding, a Jew, was forced into exile after 1933, first in Switzerland via Denmark and then in France. In an unfinished manuscript, *Das historische Problem*, he set about revising his whole conception of the state. The problem was now said to consist ‘in the change in the relation of the state to society, brought about by the *subordination of the economy* to the coercive power of the state …’ (quoted by Bottomore, Introduction to Hilferding, 1981, p. 16, emphasis in original). Hilferding briefly presented his new approach in the New York *Socialist Courier* in 1940; there, like Marx, he drew a rueful comparison between Germany and Russia. The state had not ‘withered away’ under Soviet communism:

> History, that ‘best of all Marxists’, has taught us another lesson. It has taught us that, in spite of Engels’ expectations, the ‘administration of things’ may become an unlimited ‘domination over men’, and thus lead not only to the emancipation of the state from the economy but even to the subjection of the economy by the holders of state power. (1981, p. 376 n.)

It was too late for Hilferding's brave reassessment to influence the course of events. In 1941, he died in the hands of the Gestapo.

Selected works

Books

1904. *Böhm-Bawerk's Criticism of Marx*. Ed. P. Sweezy, London: Merlin Press, 1975.

1910. *Finance Capital: A Study of the Latest Phase of Capitalist Development*. London: Routledge & Kegan Paul, 1981.

Articles

1902/3. Der Funktionswechsel des Schutzzolles. Tendenz der modernen Handelspolitik. *Die Neue Zeit* 21, 2.

1903/4. Zur Frage des Generalstreik. *Die Neue Zeit* 22, 1.

1904/5. Parliamentarismus und Massenstreik. *Die Neue Zeit* 23, 2.

1915. Historische Notwendigkeit und notwendige Politik. *Der Kampf* 8.

1915. Arbeitsgemeinschaft der Klassen? *Der Kampf* 8.

1924. Probleme der Zeit. *Die Gesellschaft* 1, 1.

1924. Realistischer Pazifismus. *Die Gesellschaft* I, 2.

1933. Zwischen den Entscheidungen. *Die Gesellschaft* 10.

1933/4. Revolutionärer Sozialismus. *Zeitschrift für Sozialismus* 1.

1934/5. Macht ohne Diplomatie – Diplomatie ohne Macht. *Zeitschrift fur Sozialismus* 2.

1940. State capitalism or totalitarian state economy. *Socialist Courier*, New York; reprinted in *Modern Review* 1 (1947).

Published speeches

1919. Zur Sozialisierungsfrage. 10th Congress of the German trade unions, Nuremberg, 30 June – 5 July. Berlin.

1920. Revolutionäre Politik oder Machtillusionen? Speech against Zinoviev at the annual conference of

the USPD in Halle, Berlin.

1920. Die Sozialisierung und die Machtverhältnisse der Klassen. 1st Congress of Works Councils, 5 October. Berlin.

1927. Die Aufgaben der Sozialdemokratie in der Republik. Annual conference of the SPD in Kiel. Berlin.

1931. Gesellschaftsmacht oder Privatmacht über die Wirtschaft. 4th AFA (Allgemeiner freier Angestelltenbund) trade union congress in Leipzig. Berlin.

Bibliography

Böhm-Bawerk, E. von. 1896. *Karl Marx and the Close of his System*. Ed. P. Sweezy, London: Merlin Press, 1975.

Bukharin, N. 1914. *Imperialism and World Economy*. London: Merlin Press, 1972.

Gottschlacht, W. 1962. *Struktur veränderungen der Gesellschaft und politisches Handeln in der Lehre von Rudolf Hilferding*. Berlin: Duncker & Humblot.

Kautsky, K. 1911. Finanzkapital und Krisen. *Die Neue Zeit* 29, 764–72, 797–803, 838–46, 874–83.

Lenin, V.I. 1916. *Imperialism, the Highest Stage of Capitalism*. Moscow: Foreign Languages Publishing House, 1947.

Marx, K. 1891. Critique of the Gotha Programme. In *The First International and After*, ed. D. Fernbach. Harmondsworth: Penguin, 1974.

Neumann, F. 1942. *Behemoth*. London: Victor Gollancz.

How to cite this article

Green, Roy. "Hilferding, Rudolf (1877–1941)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000065> doi:10.1057/9780230226203.0735

The New Palgrave Dictionary of Economics Online

Hill, Polly (1914 – 2005)

C.A. Gregory

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

development economics; economic anthropology; Hill, P.; inequality; Lewis, W. A.

Article

Polly Hill was born on 10 June 1914 into a remarkable Cambridge family that includes Nobel Prize winning physiologist A.V. Hill (her father) and J.M. Keynes (her mother's brother) among its many distinguished members. She graduated from Cambridge in 1936 with a degree in economics.

Her first job upon leaving university was with the Royal Economic Society as an editorial assistant, a position she held for two years (1936–8). Her next appointment was a one year (1938–9) research position with the New Fabian Research Bureau (which almost immediately re-amalgamated with the Fabian Society) where she wrote her first book, *The Unemployment Services* (1940). This book was concerned to expose the inefficiency and inhumanity of the system of unemployment relief and to make constructive proposals. Polly Hill's commitment to social justice never waned: economic inequality is the central theme of all her books.

At the outbreak of the war she was obliged, as an unmarried young woman, to become a temporary civil servant. She worked first, briefly, in the Treasury, then for a long time in the Board of Trade and finally in the Colonial Office. She resigned in 1951. After a period of unemployment she became a journalist for the weekly *West Africa*. She married in 1953 and moved to Ghana with her husband where, at the age of 40, she began her academic career. The academic posts she held there involved no teaching and she was able to become, as she put it, ‘a pupil of the migrant cocoa farmers of southern Ghana’. She began her fieldwork as an economist and collected data using the questionnaire method, producing her second book, *The Gold Coast Cocoa Farmer: A Preliminary Survey* (1956) with characteristic speed and efficiency. The prevailing orthodoxy had it that sedentary food farmers in southern Ghana had suddenly taken up cocoa farming at the end of the 19th century with such a degree of success that cocoa exports had risen from nil to over 50,000 tons by 1914 – the largest quantity for any country. Polly Hill had uncritically accepted this orthodoxy and her subsequent realization that most farmers appeared to be migrants who had bought their land was to have a profound effect upon her intellectual methods. She

abandoned the questionnaire method of data collection in favour of one that sought to develop generalizations on the basis of: (1) detailed fieldwork in one village; (2) fieldwork done by others elsewhere; (3) archival sources. She also began a lifelong struggle with development economists and other purveyors of orthodoxies based on casual empirical observation and ‘common sense’. She drifted towards anthropology and history where the qualities of her empirical findings were recognized for what they were: revolutionary. She spent three and a half years collecting detailed evidence to substantiate her claim that the cocoa farmers were migrants and made many fascinating discoveries in the process. For example, she found that the matrilineal farmers adopted an entirely different mode of migration from patrilineal farmers: the former bought family lands with the aid of their kin, and were prepared to grant usufructural rights to their male and female kinsfolk; the latter clubbed together in so-called ‘companies’, groups of non-kin, the land being divided into strips from a base line, according to the contribution each had made, with subsequent division on inheritance always being longitudinal. Upon hearing of this Professor Meyer Fortes, then Professor of Social Anthropology at Cambridge, encouraged her to apply for a Smuts Visiting Fellowship. This enabled her to write *The Migrant Cocoa-Farmers of Southern Ghana: A Study in Rural Capitalism* (1963) which is now widely regarded as a classic. (She was awarded a Ph.D. in social anthropology from Cambridge under new special regulations in 1966 on the basis of it.) Mainstream writers on development have by and large ignored the book even though it contains telling criticisms of aspects of W.A. Lewis's work.

Following more fieldwork in Ghana, Nigeria and India she produced a further stream of books (1970a; 1970b; 1972; 1977; 1982; 1985; 1986) and many articles of outstanding quality which established her reputation as the world's foremost economic anthropologist. She was appointed a Fellow of Clare Hall in Cambridge in 1965 and subsequently to the prestigious Smuts Readership in Commonwealth Studies (1973–9). Her publications documented in painstaking detail the complexity of agrarian relations in the tropical regions of the world in which she had worked. The books as a whole constitute an encyclopaedia of knowledge on the socio-economic conditions of poverty and economic inequality and her work ranged in scope from ‘agrestic servitude’ to ‘zamindars’. Her oeuvre was much more than a compilation of facts, though. Her own data and that of others are presented in a theoretical context which broadened as her own field experience widened. She was unrelenting in her empirically based critiques of development economists and her 1986 book *Development Economics on Trial: The Anthropological Case for a Prosecution* was a concerted attempt to make them see the error of their ways.

Selected works

1940. *The Unemployment Services*. London: Routledge.

1956. *The Gold Coast Cocoa Farmer: A Preliminary Survey*. Oxford: Oxford University Press.

1963. *The Migrant Cocoa-Farmers of Southern Ghana: A Study in Rural Capitalism*. Cambridge: Cambridge University Press.

1970a. *The Occupations of Migrants in Ghana*. Ann Arbor: University of Michigan Press.

1970b. *Studies in Rural Capitalism in West Africa*. Cambridge: Cambridge University Press.

1972. *Rural Hausa: A Village and a Setting*. Cambridge: Cambridge University Press.

1977. *Population, Prosperity and Poverty: Rural Kano 1900 and 1970*. Cambridge: Cambridge University Press.

1982. *Dry Grain Farming Families: Hausaland (Nigeria) and Karnataka (India) Compared*. Cambridge: Cambridge University Press.

1985. *Indigenous Trade and Market Places in Ghana, 1962–64*. Jos Oral History and Literature Texts. Nigeria: University of Jos.

1986. *Development Economics on Trial: The Anthropological Case for a Prosecution*. Cambridge: Cambridge University Press.

How to cite this article

Gregory, C. A. "Hill, Polly (1914–2005)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000066> doi:10.1057/9780230226203.0736

The New Palgrave Dictionary of Economics Online

Hirshleifer, Jack (1926 – 2005)

Elizabeth Landaw and David K. Levine

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Jack Hirshleifer was one of the leaders of the ‘information and uncertainty’ revolution in economics. His work on the role of time and uncertainty in asset markets and the value of information plays a fundamental role in modern economic thought. Hirshleifer was also a leader in the ‘imperial’ school of economics, taking the lead in expanding economic thought to areas such as evolution and conflict, which traditionally were studied by other social science disciplines.

Keywords

altruism; arms races; best-shot game; conflict resolution; cost–benefit analysis; evolutionary economics; free-rider problem; Hirshleifer, J.; indeterminacy of equilibrium; information economics; intellectual property; intertemporal choice; Modigliani–Miller theorem; patent races; public investment; risk; speculation; status and economics; term structure of interest rates; uncertainty; value of information

Article

Jack Hirshleifer was born on 26 August 1926 in Brooklyn, New York. He graduated at the top of his class of 855 students from Erasmus Hall High School in New York, then enrolled at Harvard in 1942, studying government and other social sciences. He was quickly drawn, however, to economics, which provided him with a ‘useful set of tools and methods’. In 1943 Hirshleifer's career as a budding economist went on hold when he enlisted for active service duty in the US Naval Reserve, serving on an aircraft carrier in the Pacific until 1945. This experience inspired in him a long-lasting and deep interest in military arms races. After the war, he resumed his studies at Harvard, receiving a Ph.D. in economics in 1950. Hirshleifer's research career started at the RAND Corporation in Santa Monica. In 1955 he became an assistant professor at the Graduate School of Business at the University of Chicago, and then returned to Los Angeles in 1960 as an Associate Professor of Economics at UCLA, becoming full professor two years later. In 1975 he became what is now called a ‘Distinguished University Professor’,

thus becoming a member of the most elite group of the University of California faculty.

Hirshleifer was an economic theorist with broad-ranging interests. He not only wrote extensively in areas of general economic interest such as capital theory or economics of uncertainty and information, but also wrote and often laid out the foundations for areas outside the traditional scope of economics, including conflict theory and evolutionary modelling.

Some of Hirshleifer's early work focused on the intertemporal theory of interest and investment. Today, this research helps us better understand such topics as intertemporal choice, decisions under uncertainty, the choice of discount rate for public investments, or liquidity and the term structure. His early interest in capital theory led not only to scores of influential articles but also to pioneering and detailed examination of the concepts of interest rate, investment and capital, which are integrated into his book *Investment, Interest and Capital* (1979) and later in the volume of collected articles, *Time, Uncertainty, and Information* (1989). The earlier book and associated articles became a framework for modern finance theory and for understanding investment decisions under uncertainty.

Hirshleifer also made a lasting contribution to the theory of speculation. He showed that differences in taste are not enough to explain speculation; rather, speculation must arise from differences in beliefs. He was the first to analyse speculation in a full general-equilibrium model, with different structures of market completeness carefully considered. Although not generally recognized as such, the 1975 *Quarterly Journal of Economics* paper is also the first paper to point out the indeterminacy of equilibrium when markets are incomplete.

Early in his career, Hirshleifer was instrumental in the information economics revolution and is considered today to be one of its founding fathers. He made the abstract ideas of contingent claims concrete through his examples and applications. In the process, he helped develop fundamental tools, such as the covariance of risks, the analysis of gambling and insurance, the Modigliani–Miller theorem, and the analysis of public investment. Most notably, his 1971 *American Economic Review* paper, ‘The Private and Social Value of Information and the Reward to Inventive Activity’, became highly influential and one of the most cited papers in the economics of information. The paper demonstrates that competitive markets need not reflect the social value of information. Hirshleifer's example of an inventor who can invest based on the knowledge of the impact of his invention shows that there can be an oversupply of inventive activity. This ‘race to be first’ has its reflection in the current literature on patent races, starting with Fudenberg et al. (1983) and continuing through such work as Gallini and Scotchmer (2001). It is the key to understanding a fundamental problem in intellectual property law, which the profession is only now coming to grips with. Hirshleifer also identifies what the profession now refers to as the ‘Hirshleifer effect’: new and more reliable information can have a negative social value if the early information on risks makes these risks uninsurable.

In addition to his founding contributions in information economics, Hirshleifer had a lifelong interest in conflict, beginning with his earliest work on war damages. Late in his career this area was the focus of his contributions, and he was a leader in extending economic methods to problems more traditionally studied in political science. Just as Hirshleifer was first drawn to economics for its methods and tools, he argued that the traditional assumptions of microeconomic theory are too narrow. One such idea, he maintained, the idea of cooperation or ‘mutually beneficial exchange via markets’, is only one form of many different forms of human interactions. An alternative way would be simply to take what you want away from other parties. This is still economics, since scarcity and competition and optimization and equilibrium are all involved. Conflicts, and indeed all struggles for power and influence, are important

economic activities, as important as exchange. He explored an economic approach to conflicts not only in the context of war but also crime, litigation, strikes and political campaigns.

His work on conflict shows how ‘Peace is more likely to the extent that the decisiveness of conflict is low, or … if the stakes are small or the technology favors the defense. More surprisingly, perhaps, increased productive complementarity between the parties does not systematically favor peace…the poorer side is generally motivated to invest more heavily in fighting effort. So conflict can become an income-equalizing process’ (1991, p. 133). It is what Hirshleifer calls the ‘paradox of power’: poor or weaker contestants defeat large ones. Subsequent work shows how a narrow range of possible settlements increases the potential for conflict and how increasing returns followed by diminishing returns explains the monopoly on military force within the state, while also explaining the multiplicity of states. A number of his papers analysing conflict as opposed to cooperation are collected in *Economic Behavior in Adversity* (1987). Hirshleifer wrote broadly on expanding the domain of economic discourse to include the ‘rational’ evolutionary analysis of altruism and spite. He believed that the standard economic postulate of fixed preferences is wrong and instead argued that evolution plays a pivotal role in shaping not only people's physical make-up but also tastes. In one of his most influential papers, ‘The Expanding Domain of Economics’ (1985), Hirshleifer reviews how the economic logic of optimization, trade-off and of equilibrium can and should be applied to a wide variety of ‘non-economic’ problems. He writes that economics constitute ‘the universal grammar of social sciences’ (1985, p. 53) but that there is the wide area of ‘noneconomics’ that economists have to become aware of and get over their ‘tunnel vision about the nature of man and social interactions’.

The paper examines different kinds of altruistic preferences, including what would now be called by experimentalists the ‘warm-glow’ effect. As an application, Hirshleifer discusses Becker's ‘rotten kid’ theorem, showing how a selfish parent can gain from altruism. Still other theories of preferences, including models of status, such as the rat-race are examined. Hirshleifer opened up new areas; by now, much of this ‘non-economic’ economics is widely studied by economists, and models of altruism and status proliferate.

Key to Hirshleifer's contribution is the underlying point of view of ‘as-if’ rationality – altruism must provide some benefit to the altruist. This was the starting point of much of the modern evolutionary economics literature – for example, the work of Kandori, Mailath and Rob (1993) and Young (1993).

From this perspective, Hirshleifer examined models such as the psychological model of ‘anger, gratitude, response’ and argued that this seemingly irrational behaviour does indeed benefit the individual. Yet Hirshleifer's view of evolution was an eminently practical one: it was firmly grounded in his desire to understand why voluntary exchange arises in some situations, but conflict in others.

Although not primarily an experimentalist, Hirshleifer, together with Glenn W. Harrison, conducted a fundamental experiment on the incentives to free ride (1989). As Hirshleifer surely imagined, increasing incentives to free ride lead to more free riding. The experiment introduced the ‘best-shot’ game, a public goods contribution game in which only the largest contribution to the public good matters. In this type of game it is socially and individually optimal for only one player to contribute, and, unlike many other types of public goods games, this theoretical prediction is exactly what happens in the laboratory.

Hirshleifer's interest in risk and investment extended to public investment and cost–benefit analysis. Although the fact is not widely known, he co-authored an important study of alternative routes for bringing water from northern to southern California, as well as a follow-up years later after one of the projects was chosen and built. He was fond of saying that much of his scepticism of government arose

from the fact that of three routes one was clearly worse than the other two – and that was the one that was actually built.

Jack Hirshleifer's love of social sciences, particularly economics, was one of his endearing traits. He liked nothing better than contemplating new puzzles and exchanging ideas with his colleagues. Although officially he changed his status to Professor Emeritus in 1991, he never ceased working, writing, reviewing, and lecturing. Colleagues would find him working every day in his office, door open, sitting behind his cluttered desk with an inviting smile. He continued to work until the very end of his life, and was proud that he was able to proofread – he sent back the galleys of the seventh edition of his very popular textbook, *Price Theory and Applications*. He hosted Thursday lunches at the UCLA Faculty Club, which became a gathering place famous for spirited discussions. A kind and approachable man dedicated to his work, Jack's rule was economics and not gossip. Those who knew him remember him for his personal warmth and sense of humour.

Although the two areas in economics that have especially felt the impact of Hirshleifer's work are information economics and conflict resolution, Hirshleifer shed light on many other fields including capital theory, finance, bioeconomics and experimental economics. With his insatiable intellectual curiosity, he was never short of good ideas, illustrating them through carefully worked out and accessible examples. He would plant many seeds and often leave to others to develop sophisticated theories. Yet with his standing concern for the value of rigorous scholarship, Hirshleifer was one of the pioneers who transformed economics into the scholarly science that it is today.

See Also

- evolutionary economics
- intertemporal choice
- risk

Selected works

1958. On the theory of optimal investment decision. *Journal of Political Economy* 66, 329–52.

1960. (With J. C. DeHaven.) *Water Supply: Economics, Technology and Policy*. Chicago: University of Chicago Press.

1961. Risk, The discount rate, and investment decisions. *American Economic Review* 51, 112–20.

1965. Investment decision under uncertainty: choice-theoretic approaches. *Quarterly Journal of Economics* 79, 509–36.

1966. Investment decision under uncertainty: applications of the state-preference approach. *Quarterly Journal of Economics* 80, 252–77.

1967. (With J. W. Milliman.) Urban water supply: a second look. *American Economic Review* 57, 169–78.

1970. *Investment, Interest, and Capital*. Englewood Cliffs, NJ: Prentice-Hall.

1971. The private and social value of information and the reward to inventive activity. *American Economic Review* 61, 561–74.

1975. Speculation and equilibrium: information, risk, and markets. *Quarterly Journal of Economics* 89, 519–42.

1978. Competition, cooperation, and conflict in economics and biology. *American Economic Review* 68, 238–43.

1979. *Investment, Interest and Capital*. Saddle River: Prentice Hall.

1985. The expanding domain of economics. *American Economic Review* 75, 53–68.

1987. *Economic Behavior in Adversity*. Chicago: University of Chicago Press.

1987. Conflict and settlement. In *The New Palgrave: A Dictionary of Economics*, vol.1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

1987. Investment decision criteria. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

1989. (With G. Harrison.) An experimental evaluation of weakest link/best shot models of public goods. *Journal of Political Economy* 97, 201–25.

1989. *Time, Uncertainty, and Information*. Oxford: Basil Blackwell.

1991. The technology of conflict as an economic activity. *American Economic Review* 81, 130–4.

1992. (With J. Riley.) *Analytics of Uncertainty and Information*. Cambridge: Cambridge University Press.

2005. (With A. Glazer and D. Hirshleifer.) *Price Theory and Applications: Decisions, Markets, and Information*. 7th edn. Cambridge: Cambridge University Press.

1989. *Time, Uncertainty, and Information*. Oxford: Basil Blackwell.

Bibliography

Fudenberg, D.R., Gilbert, J., Stiglitz, J. and Tirole, J. 1983. Preemption, leapfrogging and competition in patent races. *European Economic Review* 22, 3–31.

Gallini, N. and Scotchmer, S. 2001. *Intellectual Property: When Is It the Best Incentive System?* Berkeley: University of California Press.

Kandori, M., Mailath, G. and Rob, R. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61, 29–56.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

How to cite this article

Landaw, Elizabeth and David K. Levine. "Hirshleifer, Jack (1926–2005)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000181> doi:10.1057/9780230226203.0737

The New Palgrave Dictionary of Economics Online

historical demography

Carl Mosk

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Historical demography deals with population dynamics prior to and during early phases of industrialization. Using family reconstruction historical methodology, demographers have found partial answers to Malthusian questions revolving around mortality and fertility rates in religious records yielding estimates for marriage, life expectancy and reproduction within marriage. Employing cause of death estimates and Hutterite index measures for the proportion of women married and the level of their reproduction within marriage, historical demographers have developed tentative answers to demographic transition queries. Historical demography has contributed much to our understanding of historical population dynamics.

Keywords

Black Death; child labour; climate; demographic transition; European Fertility Project; family reconstitution method; fertility; fertility transition; food prices; Henry, L.; historical demography; homeostatic equilibrium; industrialization; industrial revolution; iron law of wages; Malthus, T. R.; Malthusian economy; marriage; mortality; population dynamics; population growth

Article

Prior to European industrialization population grew in fits and starts, because the effects of the introduction of new crops like the potato or the reclaiming of uncultivated grasslands and forested slopes for irrigated rice paddy were short-lived, typically ushering in periods of stagnation. Why did pre-industrial populations increase in such a manner, slowly groping upward from one plateau to the next, perhaps even tumbling backward to ever lower plateaus before resuming forward progress? Was it fertility or mortality or an interaction of the two that constrained the growth process?

The impact of family reconstitution

Our understanding of the dynamics of pre-industrial populations has been immeasurably increased by research in historical demography, fuelled by the development of family reconstitution for analysing records of births, deaths and marriages lodged in religious quarters – Catholic, Anglican and Lutheran parishes, Buddhist temples – in clan genealogies and in military records. Developed in the 1950s and 1960s by French demographers, most notably by Louis Henry, the family reconstitution methodology exploits the fact that individuals are separately listed in vital registers that can be linked together to yield life histories moving from birth to marriage and to death. Henry's ingenuity lay in rigorously defining the period over which a family is under observation for the purposes of deducing its mortality and fertility history.

It should be emphasized that records of religious bodies, of clans and of military organizations are not the only sources that can be tapped by historical demographers. Other sources include censuses (Quebec initiated systematic census-taking in 1665); fiscal documents, for instance taxpayer lists (Japanese population counts for the rice tax paying population of the country are available from the early 17th century); property inventories and wills; archeological remains including preserved garbage dumps; cemetery data, both skeletons and gravestones; and eyewitness accounts recorded in literary documents. Hollingsworth (1969) offers a thorough review of the various methods, pinpointing strengths and deficiencies.

Still, it was the pioneering of a carefully elaborated family reconstitution methodology by French scholars working from records of parishes from the time of Louis XIV and Louis XV that opened the floodgates for systematic analysis of fertility and mortality in pre-industrial Europe and pre-industrial Asia. Particularly important was application of the methodology to England, where several thousand parish registers beginning prior to 1600 exist, and to Japan, where Akira Hayami and others have trained Henry's methodology upon Buddhist religious records (shūmon-aratame-chō) of births, deaths and marriages in analysing the population dynamics of villages during the Tokugawa (1600–1868) period. Hayami (1997) provides a useful history, replete with concrete examples, of the impact that historical demography has had on the understanding of pre-industrial population dynamics in Japan.

What is clear from the analysis of Buddhist registers for Japanese villages is that fertility within marriage was kept quite low in many parts of the country from the early 18th century onward, the intervals between births being drawn out through a combination of infanticide and taboos against having too many small offspring in the household at any one time. Whether Japanese peasants were concerned about excess competition for the family headship (only one child could take over the headship from the patriarch of the household), responding to a falling off in the demand for child labour on densely populated paddy rice fields, or whether they were attempting to maximize survivorship rates for each child allowed to live remains a matter for scholarly debate. What historical demography has shown is that the debate must be about why fertility was fairly low, not why mortality was fairly high.

Low- and high-pressure homeostatic equilibriums

Systematic analysis of the English parish data has yielded one of the crowning achievements of post-Second World War historical social science: the securing of over 3.5 million totals for baptisms, burials and marriages drawn from 404 carefully selected Anglican parish records by a research team at Cambridge University headed up by E. A. Wrigley and R. Schofield. Developing a novel technique for

projecting back population totals from the census of 1871 and from national level estimates of births and deaths generated from the 404 parish figures, Wrigley and Schofield (1981) were able to estimate population totals, and fertility (including the gross reproduction rate that gives the number of female births a woman is expected to have across her reproductive life) and mortality rates (including life expectancy at age zero) for England between 1550 and 1871. The Wrigley–Schofield 1981 volume was path-breaking not only in offering a remarkable data-set and a remarkable set of estimates for pre-industrial fertility and pre-industrial mortality. It was also path-breaking in contesting the standard Malthusian interpretations of pre-Industrial Revolution British population dynamics.

In the standard argument the force explaining fluctuations in population size and growth rates was mortality. The Black Death reduced the ranks of the populace in the 14th century. More generally, plagues occurring between 1350 and 1660 acted as negative exogenous shocks absorbed by the British population, peasants and aristocrats alike being decimated by these waves of disease. In the Malthusian model this mechanism for regulating numbers is the positive mortality check, and populations so regulated are described as operating in a high-pressure homeostatic equilibrium, feedback running from population increase to increased food prices to enhanced mortality, thereby reducing population.

Wrigley and Schofield (1981) suggested that pre-industrial England operated as a low-pressure rather than a high-pressure equilibrium system, fluctuations in fertility driving fluctuations in population size over the long run. Indeed, the authors went so far as to suggest that there was a 50-year lag at work, surges in real wages generating surges in marriage and in births over a 50-year period. In offering a theory based upon the idea that the real wage drives population growth through its impact upon births, Wrigley and Schofield (1981) put forward a novel interpretation of the iron law of wages. This proposition states that increases in real wages due to accumulation of capital or technological improvements are ultimately choked off by population increase initiated by the improvement in wages.

The low-pressure homeostatic story accounting for the iron law of wages was not satisfactory to R. Lee, who devoted much effort to analysing the response of real wages to exogenous fluctuations in population size. For instance, Lee (1980) estimated an elasticity of minus one-and-a-half for the impact of population increase on real wages, a ten per cent increase in human numbers diminishing real earnings by 15 per cent. In Lee (1987), he pointed out that the 50-year lag is only one story that is consistent with the long-run movements in fertility and real wages advanced by Wrigley and Schofield. In any event, the 50-year lag of Wrigley and Schofield and Lee's estimates for the impact of population increase on real wages are both based upon long-run movements in population, fertility and mortality.

Equally interesting are the short-run dynamics for pre-industrial populations, fluctuations in climate – when the spring thaw permitting planting of new crops in the fields takes place, when the onset of cold fall temperatures dictates harvesting – driving movements in food prices, resulting in fluctuations in marriages, pregnancies, births and deaths. Analysing a large number of historical cases, Lee (1987) concluded that the vital rates do respond to upward and downward movements in food prices, pre-industrial societies being regulated in a homeostatic fashion that was responsive to exogenous changes in climate.

To examine more systematically the impact of fluctuations in food prices upon demographic behaviour in pre-industrial Europe and Asia, the Eurasian Project in Population and Family History has pioneered the use of longitudinal databases of household and individual records, eschewing the computation and analysis of aggregate demographic statistics generated from massive family reconstitution exercises like

that carried out by Wrigley and Schofield (1981). A good illustration of the type of analysis stemming from this approach is Bengtsson, Campbell, Lee et al. (2004). Generating results for Scania in southern Sweden, for eastern Belgium, for three villages in northern Italy, for a village in northern Japan, and for Liaodong in north-eastern China, the Eurasian Project suggests that demographic responses to short-run stress (that is, spikes in food prices) were fundamentally different in the West and in the East. In the East power, especially gender-based power, played a crucial role in shaping household demographic behaviour in the face of food scarcity, females getting less access to nutrition than males in the typical scenario. In the West, socio-economic status, especially ownership of land, mattered a great deal. When climatic variation forced up the price of foodstuffs, the landless suffered in Europe. In Asia it was young females who bore the brunt of the crisis.

Onset of the demographic transition

In addition to shedding light on Malthusian questions – on the relative importance of the positive mortality check and the preventive fertility check – historical demography has shed light on the question of when the fertility and mortality transitions began. To what extent did the onset of industrialization influence fertility and mortality? Is there evidence of fertility decline in early industrializing – or even completely pre-industrial – settings? The general overlap of industrialization and the demographic transition is evident. Heavily industrialized countries enjoy low fertility and low mortality. What is not evident is that there is a direct relationship between the onset of industrialization and the onset of mortality and fertility declines.

Nor is the short-run relationship between mortality and industrialization obvious. In the 19th century before the germ theory of disease had led to advances in sanitation (for example, chlorination of water) and the treatment of food and drink (for example, pasteurization of milk), densely populated cities were unhealthy places. Germs spread as waves of immigrants flocked into metropolitan centres rife with a diverse menu of infections, the immigrants coming from rural isolates too tiny to support the host of infectious diseases with which they were now assailed.

Only in the late 19th century and after did cities become healthy as knowledge of water purification, the importance of proper sewer systems, and flush toilets spread in the West. With the 20th century development of sulpha drugs followed by the chance discovery of penicillin and the mass manufacture of antibiotic drugs, the scale economies in distribution enjoyed by cities came to the fore. Preventing infection through public health and treating infectious cases came at a lower unit cost in dense, congested, jurisdictions that had once been mortality sink holes.

In the remainder of this article our focus will be on the onset of the fertility transition and its connections with industrialization.

The most important project that laid out the empirical groundwork for analysing questions about the onset of the fertility transition is the European Fertility Project that carried out at the Office of Population Research at Princeton University during the 1960s and 1970s under the direction of A. Coale. Coale and his colleagues wanted to construct measures of fertility and its components – reproduction within marriage, proportion married, the incidence of reproduction outside of marriage (illegitimate fertility) – that could be generated from a relatively small amount of data, data that they could secure for every province throughout 19th-century western Europe and Europe.

The European Fertility Project hit upon the ingenious procedure of comparing the actual fertility

experiences of the populations they were studying with the fertility experience of the Hutterites who thereby entered the historical demography literature as a much utilized standard. Why use Hutterite reproduction as a standard? Hutterite women in the period between the world wars married at very young ages and had as many children as possible. The Hutterite sect took very seriously the Biblical injunction to ‘be fruitful and multiply’. Moreover, the Hutterites who settled in the great plains of the United States and the prairies of Canada lived on large farms and had a strong demand for child labour. A typical Hutterite woman had a total fertility rate (the sum of the age specific birth rates, an approximation to the total number of children she would give birth to over her reproductive life) of more than 12. Using the Hutterite standard allows us to estimate the degree to which a population falls short of its maximal reproductive potential.

The Hutterite indices generated by the European Fertility Project measure the relative level of marital fertility, illegitimate fertility, proportion married and overall fertility for any jurisdiction that has counts of births classified by legitimacy status and counts of population classified by gender and marital status in the five-year age groups. The idea is to use figures on women and married women in the five-year age groups in a given population of interest to the researcher to compute the level of fertility and marital fertility that would occur if these women reproduced at the rate of Hutterite women in the cohorts of the 1920s and 1930s. The age specific rates (for five-year age groups) at which Hutterite wives reproduced are known and these are used in conjunction with the actual data on population and births to compute the Hutterite indices.

In assessing why populations fall below maximal reproductive potential it is important to separate out the impact of low proportions married from the impact of sharply diminished reproduction within marriage. The Hutterite index for marital fertility (I_g) for a given population is the ratio of the legitimate births occurring in that population to the number that would occur if the women reproduced at the rate of the Hutterites. The Hutterite index for proportion married (I_m) is the ratio of married women weighted by the Hutterite fertility schedule – take the number of married women in each age group and multiply this number by the corresponding level of Hutterite fertility for the age group, thereby giving heaviest weight to the most reproductive ages – divided by the total number of women weighted by the Hutterite schedule. The Hutterite index for illegitimate fertility (I_h) is the ratio of the number of illegitimate births to those that would occur had the unmarried women reproduced as the Hutterite women had reproduced. The overall Hutterite index of fertility (I_f) is the ratio of total births occurring in a population to those that would have occurred had the women been as fruitful as the Hutterite women. The last measure offers an overall summary for fertility.

Not surprisingly, the Hutterite indices for illegitimate fertility – in 19th-century Europe and Asia – tend to be low, typically falling below a value of 0.10. By contrast the Hutterite index for marital fertility in most 19th-century western European provinces tended to be fairly high, around 0.80 in many cases.

One of the convenient properties of the Hutterite indices is their multiplicative property. If the index of illegitimate fertility is zero (typically it is close to zero), then the Hutterite index for overall fertility is the product of the Hutterite indices for marital fertility and proportion married, namely $I_f = I_g * I_m$.

To see why constructing these indices yields useful information about the nature of pre-transition fertility and the dating of the fertility transition, consider the following. In 19th-century western Europe prior to the sustained decline in marital fertility (the European Fertility Project defines the onset of the fertility transition as a drop in I_g of ten per cent initiating irreversible decline, no subsequent return to the

pre-decline level occurring), a typical value for the Hutterite index for proportion married was around 0.5, the corresponding Hutterite index for marital fertility being between 0.8 and 0.9. Multiplying the two gives a value of between 0.4 and 0.45, meaning that in western Europe women reproduced far less than did the Hutterites, not because of what they did within marriage, but rather because they were not marrying very young or, in some cases, at all. By contrast in pre-decline Japan, China and Korea, the levels of I_m were usually between 0.8 and 0.9, women marrying very early and almost universally. However, levels of reproduction within marriage I_g were quite low in pre-transition Asia, around 0.5 in many cases. Again, taking the product, we get a range for I_g between 0.4 and 0.45.

So, in both pre-transition Asia and pre-transition western Europe, overall levels of reproduction were modest, but for different reasons in the two regions. In Europe the key was late marriage and low proportions marrying. This was something Malthus approved of, believing that the path of demographic virtue lay in late marriage and abstinence outside of marriage. In Asia the key was relatively low levels of reproduction within marriage, something Malthus was less enthusiastic about. Indeed, he probably would have labelled it vice.

To return to the question of what the European Fertility Project's findings tell us about the relationship between industrialization and the fertility transition, some of the most striking findings of the project need stating. First, France was the region in western Europe enjoying the earliest decline in marital fertility, its irreversible fall beginning in the early 19th century, occurring prior to sustained industrialization there. Second, the irreversible decline in English marital fertility did not occur until the 1870s, a full century after the Industrial Revolution began there. Third, language and culture seem to have been important in shaping the spread of marital fertility decline. For instance in Belgium, language difference separates early-decline provinces from late-decline provinces. For these reasons the European Fertility Project concluded that stopping behaviour within marriage – having a specific number of offspring, then ceasing having more children altogether – was an innovation. As Coale and Watkins (1986) demonstrate, the consensus opinion in the European Fertility Project was that the innovation of regulating reproduction diffused through contact between individual households, this diffusion channelled through and within distinctive cultural groups.

In short, there is no simple story for western Europe involving the short-run relationship between the onset of industrialization and the onset of marital fertility decline. It is apparent that both are important to modernization. But the interaction of the two is certainly complex.

When we turn to Asia, the complexity of the relationship is even more evident. For instance, in Japan, China, Korea and Asiatic Russia marital fertility appears to have risen before it began its irreversible decline in the 20th century. Mosk (1983) offers one hypothesis about the rise in fertility in Japan that is consistent with the idea that there is a long-run linkage between industrialization and low marital fertility. His explanation rests on the idea that in the short run a rising standard of living may actually induce a rise in marital fertility provided marital fertility has been suppressed through infanticide and sexual taboos aimed at lengthening the intervals between live births. In particular, he argues that rural areas that were experiencing land reclamation due to the diffusion of rice seed varieties from the south-west to the north-east spawned new family managed farms, increasing the demand for child labour and easing pressure on parents concerned with finding marriage and/or farming opportunities for their offspring. Additionally, improved food consumption affected the length of intervals between live births considered optimal, promoting a rise in I_g between the 1880s and the 1920s. Better-fed households felt

less constrained to space their births far apart lest they fall short of the nutritional resources required to guarantee survival for all of their youngsters. To these arguments one can add the fact that the opening of the country to international trade in the late 19th century created a strong export market for silk, which was produced by family labour especially in the north-east and the Japanese Alps.

In sum, the literature dealing with the overlap of industrialization and the onset of the demographic transition suggests that the interaction of the two secular transformations crucial to defining modernity is complex and intriguing. As with the issues involving the Malthusian economy, much is known. The general contours of the issues involved are clear enough. But, as with so many other things, the devil is in the details. At the detailed level, it is clear what we do not know is as important as what we know. In this sense historical demography has opened up as many questions for future research as it has provided answers to questions thrown up by previous generations of scholars.

See Also

- demographic transition
- Industrial Revolution
- Malthus, Thomas Robert
- Malthusian economy

Bibliography

Bengtsson, T., Campbell, C., Lee, J. et al. 2004. *Life under Pressure: Mortality and Living Standards in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press.

Coale, A. and Watkins, S., eds. 1986. *The Decline of Fertility in Europe: The Revised Proceedings of a Conference on the European Fertility Project*. Princeton: Princeton University Press.

Hayami, A. 1997. *The Historical Demography of Pre-modern Japan*. Tokyo: University of Tokyo Press.

Hollingsworth, T. 1969. *Historical Demography*. Ithaca, NY: Cornell University Press.

Lee, J. and Campbell, C. 1997. *Fate and Fortune in Rural China: Social Organization and Population Behavior in Liaoning 1774–1873*. Cambridge, MA: Harvard University Press.

Lee, R. 1980. An historical perspective on economic aspects of the population explosion: the case of preindustrial England. In *Population and Economic Change in Developing Countries*, ed. R. Easterlin. Chicago: University of Chicago Press.

Lee, R. 1985. Population homeostasis and English demographic history. *Journal of Interdisciplinary History* 25, 635–60.

Lee, R. 1987. Population dynamics of humans and other animals. *Demography* 24, 443–65.

Mosk, C. 1983. *Patriarchy and Fertility: Japan and Sweden, 1880–1960*. New York: Academic Press.

Wrigley, E. and Schofield, R. 1981. *The Population History of England 1541–1871*. Cambridge, Massachusetts: Harvard University Press.

How to cite this article

Mosk, Carl. "historical demography." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000143> doi:10.1057/9780230226203.0738

The New Palgrave Dictionary of Economics Online

historical economics, British

J. Maloney
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Ashley, W.; British historical economics; Cliffe Leslie, T. E.; Cunningham, W.; economic man; evolutionary economics; Foxwell, H. S.; German Historical School; Hewins, W. A. S.; historicism; Hobson, J. A.; holism; individualism versus holism; Ingram, J. K.; Jones, R.; marginalism; Marshall, A.; protection; Rogers, J. E. T.; Toynbee, A.

Article

A group of economists whose heyday was from 1875 to 1890 and whose major figures were John Kells Ingram (1823–1907), James E. Thorold Rogers (1822–1890), T.E. Cliffe Leslie (1827–1882), William Cunningham (1849–1919), Arnold Toynbee (1852–1883), William Ashley (1860–1927) and W.A.S. Hewins (1865–1931). H.S. Foxwell (1849–1936) was sympathetic to their approach but outside the group's mainstream. All were united by an inductive approach to economics, a determination to stress that no economic theory or policy could be appropriate to all times and places, and a conviction that classical and neoclassical economics alike were already too abstract to give state or citizen much practical help, and were getting worse.

The movement's most important forerunner was Richard Jones (1790–1855), whose criticisms of Ricardian economics – both for its hyper-deductive character and its pretensions to universality – enjoyed intelligent public attention without much persuasive power. Jones offered neither a historically relative political economy to put in Ricardianism's place nor even any substantial contribution to economic history. But, in any case, the time was not right for Jones's ideas to take hold. By the 1870s a number of factors had combined to prepare the ground for a far more influential historical critique of orthodox economics. There was the influence of John Stuart Mill, who in his later years both practised and lent his philosophical authority to a more inductive approach to political economy. Yet when Mill's influence was removed by his death in 1873, silencing the most authoritative voice in economics, the collapse of classical orthodoxy was further accelerated. And of its two main potential heirs, marginalism and historicism, it was the historicists who were more in tune with the general intellectual climate of the time.

As Darwinian ideas were absorbed into social science, the call went up for an evolutionary (and hence relativistic) science of political economy. (No one was to call for it more loudly than Marshall.) The Comtean critique of overspecialization within social science was still near its zenith, and applied with especial force to the increasingly narrow world of neoclassical economics. ‘Straight’ history was increasingly emphasizing its economic aspects in the work of F.W. Maitland, F. Seebohm and P. Vinogradoff. And, for those who were prepared to listen, Karl Marx was reiterating the potential scope and grandeur of economic dynamics.

The representatives of the English historical school drew on such influences with varying degrees of emphasis. Ingram used his presidency of Section F of the British Association (the social science section) to mount an explicitly Comtean attack on political economy's ‘narrowness’ in 1878. Ashley painstakingly catalogued the aspects of Marxism with which he was and was not in agreement. The one conditioning factor which, oddly enough, was of limited influence was the work of the German Historical School of economists. English historicists might invoke the authority of their German contemporaries; Ashley and Hewins had important contacts with the later German Historical School; but it is hard to point to any German historicist as a major formative influence on any English counterpart.

What, then, was the detailed message of the Historical School? (In answering this question we shall be able to throw light on how far it should be regarded as a distinct ‘school’ at all.) First, as has already been mentioned, they were reacting against the narrow scope of orthodox economics. Thus, Ingram's address of 1878, while accepting the arguments in favour of doing ‘one thing at a time’, warned that the social sciences were still branches of one subject ‘and the relations of the branches may be precisely the most important thing to be kept in view respecting them’. Ingram saw the narrow intellectual vision of orthodox economists as both cause and consequence of their neglect of moral issues, and further argued that once it was accepted that ‘the idea of forming a true theory of the economic frame and working of society apart from its other sides is illusory’ it necessarily followed that ‘the economic structure of society and its mode of development cannot be deductively foreseen but must be ascertained by direct historical investigation’ (Ingram, 1878).

But should one's methodological stance in fact depend on one's assessment of the appropriate intellectual boundaries of economics? J.A. Hobson was later to argue that the two issues had nothing whatever to do with one another. However, historicists to a man – albeit with different degrees of emphasis – followed Ingram's lead in using their calls for a broader-based discipline to buttress their onslaught on unbalanced deductivism. The link was ‘economic man’, seen by historicists as an unreal psychological stereotype wholly unable to support the pyramids of deductive logic burdened upon him by Ricardians and Jevonians alike. Whether it was wealth or utility that he was supposed to maximize, he turned out very much the same, ‘an abstraction confounding a great variety of different and heterogeneous motives which have been mistaken for a single homogeneous force’ (Cliffe Leslie, 1879). Other Ricardian propositions which, in Leslie's view, contradicted actual experience included the quantity theory of money and the contention that competition operated so as to equalize rates of profit across the economy. Leslie's suggestion that the whole edifice of Ricardian economics be levelled to the ground, prior to economists making a fresh and cautious start, marked the high point of historicist iconoclasm. There were a number of different stopping-places (most of them inhabited by Ashley at one time or another) along the road from orthodoxy to this extreme point. Yet the historicists hang together as a school because of their common emphasis on factual and statistical thoroughness, on the relativity of economic doctrines, and on entering unfamiliar territory with an open mind and doing painstaking research before

allowing the first tentative inductive generalizations to filter through. The most orthodox of the school, Thorold Rogers, made the most impressive statistical contribution with his *History of Agriculture and Prices in England* (1866) which, among other objectives, sought to marshal the figures needed to refute Ricardian rent and wage theory. Ashley's verdict, however, that Rogers' practice of merely illustrating his preconceived opinions with historical material was alien to a genuine historical method has been endorsed by modern commentators.

It would be wrong to conclude from the above that the Historical School was hostile to deduction as such. 'Deduction', said Ingram, 'is a legitimate process when it sets out not from *a priori* assumptions, but from proved generalisations'. The historicist position, in effect, was that one had to ascertain by factual investigation exactly how amenable to deductive analysis different economic phenomena actually were. That the calculating maximizing spirit (where it existed) was amenable to Ricardian treatment was conceded on all sides. This point had been heavily stressed by Walter Bagehot (in his centenary essay on *The Wealth of Nations*) in the hope of rendering orthodox economics more plausible by demarcating its boundaries as those of the modern commercial world. Ashley's inaugural lecture at Harvard in 1893 endorsed this point; Cunningham's *Modern Civilisation in Some of its Economic Aspects* (1896) asserted that deductive analysis was coming into its own because 'business of a modern type is being extended over a larger and larger area'. That this last tendency was – on balance – welcomed by Ashley and regretted by Cunningham may help explain the difference in their attitudes to Marshallian economics. Ashley (who was to become professor of commerce at Birmingham in 1901) shared Marshall's enthusiasm for most of what the modern businessman represented. In Cunningham, by contrast, distaste for the modern world and nostalgia for the Middle Ages predominated. But personal temperament counted for just as much in explaining the contrast between Ashley's relatively placatory attitude to Marshall and Cunningham's violently hostile one.

Marshall's inaugural lecture at Cambridge in 1885 had met, head-on, the historicist assertion that the forces of custom and habit in economic life were strong enough to make orthodox economics, with its basic postulate of maximization, widely redundant. Marshall predicted that 'economic science' would soon be even more successful than it was already in 'break[ing] up and explain[ing] economic customs'; asserted that statements that this or that economic arrangement was due to custom were little more than confessions of ignorance of true causes; and entrusted economic analysis with the illumination of such ignorance – the demonstration, for example, that 'rents seldom diverge much for a long time from their Ricardian level in the East' (Marshall, 1885). Cunningham, while regarding the whole lecture as a personal and public affront, fastened especially onto this last point, telling the British Association (1889) that 'Professor Marshall, instead of accepting the description of mediaeval or Indian economic forms as they actually occur, sets himself to show that the accounts of them can be so arranged and stated as to afford illustrations of Ricardo's law of rent.' Marshall's *Principles of Economics*, published the following year, opened with a long historical introduction which Ashley saw as a conciliatory gesture and Cunningham as a further provocation. (Today it reads as neither.) In 'The Perversion of Economic History' (*Economic Journal*, September 1892), Cunningham joyously rebuked what he saw as Marshall's hasty and amateurish style of historiography. It would all have read more convincingly if Cunningham had refrained from grotesquely out-of-context quotation, even at one point inserting a rogue word into Marshall's text to make it sound marginally more implausible.

Marshall's reply to Cunningham's criticisms (it took Cunningham three years and seven polemics to induce it) was seen in most quarters as the final statement in the dispute (if only because the *Economic*

Journal refused Cunningham the space for a counter-riposte.) Ashley, in his Harvard inaugural the following year, praised the historical chapters in the *Principles* and claimed that ‘to most of us the recent exchange of hostilities between two distinguished English economists has seemed almost an anachronism’.

The methodological debate, then, subsided after the early 1890s. But the protectionist controversy which began when Joseph Chamberlain disavowed free trade in 1903 saw survivors of the old historicists grouping reconstituted for a new battle. The episode is best approached via a general look at historicist attitudes to policy questions.

It is no coincidence that the entire Historical School, regardless of whether as individuals they were of the ‘left’ or the ‘right’, favoured an acceleration of the existing trend towards increased state intervention in the economy. Irish social reform, the recognition and legal protection of the trades unions, and the conditions of industrial and agricultural workers were all seen as urgent areas of responsibility for the state. The general view was well summarized by Foxwell (1885):

> We have been suffering for a century from an acute outbreak of individualism unchecked by the old restraints and invested with almost a religious sanction by a certain soul-less school of writers. The narrowest selfishness has been recommended as public virtue.

Ingram praised the German Historical School for upholding the power of the state as ‘the organ of the nation for all ends which cannot be adequately effected by voluntary individual effort’. Cunningham's *Politics and Economics* (1885) introduced his readers to ‘National Husbandry’, Cunningham's scheme for an economic policy holistic in its inspiration and nationalistic in its objectives: ‘the duty we owe to posterity [is] to make the future of our nation as great and noble as lies within our power.’

The link between holism (refusal to isolate the individual as a unit of analysis) and historical relativism was an irreproachably logical one: only if an individual can be isolated from his social context can a theory involving him be isolated from time and place. And Cunningham for one kept his readers’ eyes firmly on the fact that policy recommendations were as historically relative as economic principles, even suggesting at one point that the fact that a measure had worked well in very different circumstances was a consideration *against* proposing it here and now. Such pragmatism characterized much of the protectionist campaign. If free-trading economists were to be charged with inflexible dogmatism, intellectual arrogance and subservience to abstractions, it was essential that no such taint could be thought to cling to the protectionist cause. Ashley, indeed, never went beyond recommending temporary and selective tariffs for purposes of retaliation, and stressed that ‘with England as she has been for some centuries the notion that imports are paid for by money which might otherwise be spent at home is the crudest of popular fallacies’. Cunningham – eventually – did arrive at a more thoroughly protectionist stance than this, but it took him until 1910 to do so. And by 1910 the steam was running out of the protectionist campaign anyway, at least as far as the Historical School was concerned. Ashley's administrative responsibilities at Birmingham and Hewins's parliamentary ones virtually terminated their contributions to serious economic debate; Cunningham turned his attention to the relations between Christianity, political practice and social science. The Historical School's achievements were complete by 1914.

How significant were they? Today their part in the foundation of economic history as a subject in its

own right is more obvious than their contribution to economics. Their lack of facility with marginal analysis – no historicist tried to master the neoclassical ‘paradigm’ and it must be doubted whether most of them would have been able to handle it even if they had tried – relegated them to outsiders’ roles once the dominance of neoclassicism was secured. Could they have prevented this dominance? The answer depends on whether one thinks that the inductive, historically based economics which they demanded but ostentatiously failed to supply could ever have been a feasible project. As it was, their lack of solid achievement inevitably weakened their position even as critics. Yet they forced both Marshall and his disciples to change both their thoughts and their presentation of these thoughts in a number of ways. Economic concepts were more carefully defined, and the bounds of their applicability more precisely demarcated. Policy recommendation became more cautious and less likely to be accompanied by exaggerated statements of the contributions of pure theory. The modern economist, said L.L. Price (1906),

> evinces a readiness to recognise without reserve those qualifications of subtle delicate theory which a comparison with rough, unyielding facts must necessarily require. This reasonable attitude is largely due to the abiding influence of the vigorous controversy in which Cliffe Leslie bore a leading part.

See Also

- Historical School, German

Bibliography

Ashley, W.J. 1903. *The Tariff Problem*. London: P.S. King & Son.

Coats, A.W. 1954. The historicist reaction in English political economy, 1870–90. *Economica* 21(May), 143–53.

Cliffe Leslie, T.E. 1879. *Essays in Political and Moral Philosophy*. Dublin: Hodges, Figgis & Co.

Cunningham, W. 1885. *Politics and Economics*. London: Kegan Paul, Trench & Co.

Cunningham, W. 1896. Modern Civilisation in Some of its Economic Aspects. London: Methuen.

Foxwell, H.S. 1885. What is political economy? *The Eagle*, No. 79.

Ingram, J.K. 1878. *The Present Position and Prospects of Political Economy*. Dublin and London: Longmans. Repr. in *Essays in Economic Method*, ed. R.L. Smith, London: Duckworth, 1962.

Koot, G. 1980. English historical economics and the emergence of economic history in England. *History*

of Political Economy 12, 174–205.

Marshall, A. 1885. *The Present Position of Economics*. London: Macmillan.

Rogers, J.E.T. 1866–1902. *A History of Agriculture and Prices in England*. Oxford: Clarendon Press.

Semmel, B. 1960. *Imperialism and Social Reform*. London: George Allen & Unwin.

How to cite this article

Maloney, J. "historical economics, British." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000189> doi:10.1057/9780230226203.0739

The New Palgrave Dictionary of Economics Online

Historical School, German

Heath Pearson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The German Historical School was an influential heterodoxy in 19th-century political economy. It diverged from the classical school crucially in its scepticism that universal laws of social behaviour could be established. Its members were also more interventionist, tending to favour protection, regulation colonization and the welfare state, though by no means unanimously on every point. In line with their relativity, they accepted that their policy recommendations, too, were historically contingent. Their influence among economists was greater in developing countries than in western Europe, but it has everywhere had a lasting impact on allied branches of social science such as sociology.

Keywords

American Academy of Political and Social Science; American Economic Association; Brentano, L.; Bücher, K.; Chayanov, A.; Commons, J.; Durkheim, E.; German historical school; Germany, economics in (20th century); Gide, C.; Hildebrand, B.; India, economics in; Italy, economics in; Japan, economics in; Jones, R.; Keynes, J.M.; Knapp, G.; Knies, K.; Lamprecht, K.; List, F.; Lubbock, J.; Maine, H.; Malinowski, B.; Marshall, A.; Methodenstreit; Mitchell, W.; Polanyi, K.; relativity; Roscher, W.; Schmoller, G.; Schumpeter, J.; Sombart, W.; Spiethoff, A.; Veblen, T.; Verein für Socialpolitik (Association for Social Policy); von Inama-Sternegg, K.; Wagner, A.; Weber, M.

Article

The German Historical School has a fair claim to be the most thoroughgoing and influential heterodoxy in 19th-century political economy. Scholars conventionally date its origins to 1843, with the publication of Wilhelm Roscher's *Outline of Lectures on Political Economy, according to the Historical Method*. Again by convention, the school is divided chronologically into three generations. The ‘older’ generation included Roscher (1817–94), Bruno Hildebrand (1812–78), and Karl Knies (1821–98). It was succeeded by a ‘younger’ generation, led by imperial Germany's most prominent economist, Gustav

Schmoller (1838–1917), and including Lujo Brentano (1844–1931), G.F. Knapp (1842–1926), K.T. von Inama-Sternegg (1843–1908), and Karl Bücher (1847–1930). A ‘youngest’ generation included Werner Sombart (1863–1941) and Arthur Spiethoff (1873–1957). All were professors of political economy or of *Staatswissenschaft* (‘state science’), and were widely known outside academic circles. A full account of the German Historical School would include many lesser-known figures, as well as several famous scholars who have often been associated with its agenda: Friedrich List (1789–1846), Adolph Wagner (1835–1917), Karl Lamprecht (1856–1915), and Max Weber (1864–1920), among others. There is no consensus date for the school's demise, but most would agree that by 1918 it was losing momentum, and that by 1945 it was a spent force.

In so far as they are remembered for belonging to the school, history has assigned to these economists the role of dramatic foil vis-à-vis classical political economy. Where the classicals were cosmopolitan children of the Enlightenment, the historical economists are remembered as romantics, idealists, nationalists; where the former were motivated to understand the nature and prospects of the commercial society taking shape around them, the latter were oriented to the economic past and its evolution towards the present; where the former were Newtonian in their aspirations for a master theory of the market order, the latter were satisfied to explore the peculiarities of specific situations; where the former offered a robust defence of private enterprise, the latter were just as robust in their vindication of state intervention; and where the classicals proved endlessly adaptable as circumstances varied, the German Historical School was sterile, a creature of one time and place. There is something to be said for each of these contrasts, but each of them can be – and has been – overdrawn, and the trend of recent scholarship has been to mitigate them.

Method

The Historical School partook of the ethic of professional historiography, seeking not to ransack the past but rather to understand it on its own terms. As history was an integral part of the *Staatswissenschaft* curriculum through which most German economists passed, it could hardly have been otherwise. But it is also fair to say that in this curriculum history was yoked, not to say subordinated, to the established discipline of *Statistik*, which had long meant the comparative study of social phenomena for purposes of effective statecraft. This more instrumentalist approach to historical inquiry is clearly in evidence in the works of the Historical School; it accords too with the participation of many members (notably Knies, Hildebrand, Inama-Sternegg, Bücher, Knapp, Brentano, Wagner, and Spiethoff) in the development and use of statistics in its more strictly modern sense, and the interest of many others (notably Roscher, Schmoller, Inama-Sternegg, and especially Bücher) in contemporary ethnography. In this sense they bear more than a passing resemblance to an Enlightenment polymath named Adam Smith, whom Roscher (1843, p. 150) named among the forefathers of historical economics.

But how would those copious data be used? A widespread view, born especially from the famous *Methodenstreit* between Schmoller and Carl Menger in the early 1880s, is that the historical economists took their scientific brief to include description, collation, and not much else; valid theoretical knowledge would emerge, if at all, only in the fullness of time and of its own accord. It is indeed true that in the heat of the dispute Schmoller made some ill-advised statements to this effect, and it is true also that he and his colleagues consistently denounced what they saw as the deductive excesses of Ricardian theory. However, in general they were far from denying the validity of deductive inference in

principle (see Schmoller 1901–04, pp. 108–11); in practice they did engage in generalization and in theoretical speculation, sometimes so sweepingly as to make a classical economist blush. Prominent among the grander visions was their penchant for evolutionary ‘stage theories’ of economic development. It is these stage theories which have attracted the taint of holism, teleology, and crypto-Hegelian idealism. Once again, these charges are less than outrageous (see Weber, 1902–05) but significantly overstated. Turgot, Smith and Marx made similar efforts to reduce such broad historical processes to patterns of individual behaviour, which in turn are explicable in terms of the overall context. Even as it evolves, Roscher wrote in the introduction to his influential textbook, the economy remains ‘a natural product of the faculties and drives which make the human being human’ (1854, s.14). Perhaps the best single term to distinguish their brand of science is ‘relativity’. Unlike Newton and his admirers, who envisioned law-like relationships that were invariant as to time or place, the historical economists thought this a hopeless task for the human sciences. The theories they aspired to had fewer constants, more variables (psyche, environment, institutions, and so on) and in some versions a large error term – but they were still recognizable as theories. According to Roscher's formulation, the classicals had postulated rules, to which recent critics had pointed out myriad exceptions. ‘Now it would be above all necessary’, he went on, ‘to broaden the rules themselves to the point where those exceptions are incorporated’ (quoted in Eisermann, 1956, p. 150). Or, as Schmoller himself put it near the end of his career, German economists of his persuasion had achieved progress in economic theory the way Smith had, ‘by placing man and society at its center; but they did not thereby exclude the methods of natural science, or general concepts, or regularities. They did not claim that all the phenomena of economic life are individual and unique’ (1911, p. 434).

Policy

As regards policy recommendations, it is clear that the modal opinions of German historical economists were distinctly more interventionist than the Anglo-French norm. The renowned Verein für Socialpolitik (Association for Social Policy, arguably the world's first economic think tank) was founded in 1872 primarily by members of the school, with a mission that stood as a plain rebuke to the principle of laissez-faire. It is for this reason that the epithet *Kathedersozialisten* (‘socialists of the lectern’) was rather indiscriminately applied to them in their own day, and for this reason, too, that historians have come to view government activism as just as intrinsic to historical economics as any methodological precept. Paradigmatic in this view is Schmoller's belief in a ‘social monarchy’ and its capacity to reconcile the goals of private property, national development, and distributive justice through a programme of protection, regulation, and colonization. Once again, however, this generalization must be handled gingerly. It understates the diversity of political opinion within historical economics; specifically, it ill serves those economists who called for participatory government (Brentano, Bücher), who tended towards state socialism (Wagner, young Sombart), who opposed Bismarck's tariffs (Brentano, Bücher, Weber), and who doubted the capacity of regulation to improve upon market outcomes in general (Roscher, Hildebrand, Weber). It also ignores their essentially relativistic outlook: like List before them, who had promoted protection as the policy for *his* time but not for *all* time, the German historical economists – not excluding Schmoller himself – recognized the historical contingency of their specific recommendations.

Influence

It is also the case that the German Historical School was less parochial than has been suggested. It is true that the influence of German economists was conspicuously weak in francophone Europe through most of the 19th century, despite an early translation of Roscher's *Principles*, and despite the efforts of the Belgian economist Emile de Laveleye. This situation began to improve after about 1880, however, with the creation of the first chairs of political economy in the law faculties of the French universities. Since the new professors were perforce trained in law, and since French jurisprudence had already begun to fall under the sway of German historicism, they were better disposed to the historical economists than their predecessors (Gide, 1908). Charles Gide's critical appreciation of the German economists was characteristic of this younger generation, as was their reception by Paul Cauwès, François Simiand and Emile Levasseur. Historical economics fared better in the United Kingdom, thanks largely to the indigenous examples of Richard Jones, Henry Maine, John Lubbock, and others. T.E. Cliffe Leslie praised their endeavours, as did W.J. Ashley, William Cunningham, J.S. Nicholson, and W.A.S. Hewins. Alfred Marshall, whose name is not typically associated with historical economics, in fact affirmed that the school's work ‘has thrown light on economic theory, has broadened it, has verified, and has corrected it’ (1890, p. 74).

Despite these successes, in western Europe the German historical economists remained exotic specimens of a minor genus. Elsewhere they fared better, with their influence waxing in rough proportion with the developmental ambitions of the society in question. The historical school's rise to prominence in Gilded Age America (c.1876–1914), due largely to Germany's pre-eminence as a site for higher education in economics, has been well documented (Dorfman, 1955; Herbst, 1965; Rodgers, 1998). The American Economic Association (AEA) and the American Academy of Political and Social Science were both originally modelled on the Verein für Socialpolitik; all told, 20 of the first 26 presidents of the AEA had studied in Germany. While the leading American ‘institutionalists’ of the early 20th century did not have first-hand experience in Germany, they can be seen as carrying on the school's agenda: Thorstein Veblen its methodological dissent, W.C. Mitchell its statistical inquiries, J.R. Commons its social reformism.

The Italian case offers a fairly close parallel. Young Italian economists were drawn to advanced study in Germany, and while the guardians of orthodoxy could inveigh against the trend of *germanismo economico*, they could not staunch the school's influence among economists such as Luigi Cossa, Vito Cusumano, Giacomo Luzzatti, and Achille Loria (Schiera, 1989). Elsewhere the German historical school achieved something close to intellectual hegemony by the turn of the 20th century, for example in Russia (Balabkins, 1988; Kingston-Mann, 1999; Barnett, 2004) and in Finland (Heinonen, 2002). Interestingly, its greatest influence relative to other schools was in Asia. In British India, the German dissent from orthodoxy was praised in the work of M.G. Ranade, G.K. Gokhale, R.C. Dutt, and G.S. Iyer, and left its imprint in the field of ‘Indian Economics’ that they founded. In Meiji Japan, meanwhile, the school established a decisive beachhead at the Imperial University in Tokyo in the early 1880s, thanks to direct German influence and especially to that of German-inspired American professors (Pyle, 1974; Sugiyama and Muzuta, 1988). A Society for Social Policy was founded there in 1897 on the model of the Verein für Socialpolitik, membership in which soon became an essential qualification for professional economists in that country.

Finally we turn to the question of the German Historical School's influence, or lack thereof, into the later

20th century and beyond. The school had a lasting impact on allied branches of social science. In economic sociology, Emile Durkheim's early works were fairly deeply engaged with historical economics (Steiner, 2003); Joseph Schumpeter held Schmoller up as a pioneer in the field (Schumpeter, 1926); and Max Weber was so deeply rooted in the school that he himself has occasionally been called a member. In the field of economic anthropology, Bronislaw Malinowski, Karl Polanyi and A.V. Chayanov had all been exposed to this literature in their youth (Kahn, 1990). It is only within economics itself that the German Historical School's star waned quickly after 1930. The reasons are no doubt complex and entwined with the drama of German political history; but surely the ‘formalist revolution’ in economic theory at large – where clarity and elegance gained great popularity, occasionally at the expense of verisimilitude and relevance – played a significant role. The matter is crystallized in J.M. Keynes's obituary for Marshall in 1924, where he characterized the school's work as ‘learned but half-muddled’. One pictures Marshall nodding in reluctant agreement, and then asking aloud what more could be asked of true social science. For Keynes's successors, however, that indictment could hardly have been more damning.

See Also

- American Economic Association
- Bücher, Karl Wilhelm
- economic anthropology
- economic sociology
- Hildebrand, Bruno
- historical economics, British
- institutionalism, old
- Knies, Karl Gustav Adolf
- List, Friedrich
- Methodenstreit
- Roscher, Wilhelm Georg Friedrich
- Schmoller, Gustav von
- Sombart, Werner
- Spiethoff, Arthur August Kaspar
- Weber, Max

Bibliography

Balabkins, N. 1988. Schmoller in Tsarist Russia. *Journal of Institutional and Theoretical Economics* 144, 581–90.

Barnett, V. 2004. Historical political economy in Russia, 1870–1913. *European Journal of the History of Economic Thought* 11, 231–53.

Dorfman, J. 1955. The role of the German Historical School in American economic thought. *American Economic Review* 45, 17–28.

Eisermann, G. 1956. *Die Grundlagen des Historismus in der deutschen Nationalökonomie*. Tübingen: Enke.

Gide, C. 1908. L'Ecole économique française dans ses rapports avec l'Ecole anglaise et l'Ecole allemande. In *Die Entwicklung der deutschen Volkswirtschaftslehre im neunzehnten Jahrhundert*, vol. 1, ed. S. Altmann et al. Leipzig: Duncker and Humblot.

Heinonen, V. 2002. The Influence of the German Historical School in Finnish economic thought around the turn of the century. In *Economic Thought and Policy in Less Developed Europe: The Nineteenth Century*, ed. M. Psalidopoulos and M. Mata. London: Routledge.

Herbst, J. 1965. *The German Historical School in American Scholarship: A Study in the Transfer of Culture*. Ithaca: Cornell University Press.

Kahn, J. 1990. Towards a history of the critique of economism: the nineteenth-century German origins of the ethnographer's dilemma. *Man* 25, 230–49.

Kingston-Mann, E. 1999. *In Search of the True West: Culture, Economics, and Problems of Russian Development*. Princeton: Princeton University Press.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Pyle, K. 1974. Advantages of followership: German economics and Japanese bureaucrats, 1890–1925. *Journal of Japanese Studies* 1, 127–64.

Rodgers, D. 1998. *Atlantic Crossings: Social Politics in a Progressive Age*. Cambridge, MA: Belknap.

Roscher, W. 1843. *Grundriβ zu Vorlesungen uber die Staatswirthschaft, nach geschichtlicher Methode*. [Outline of Lectures on Political Economy, according to the Historical Method.] Göttingen: Dieter.

Roscher, W. 1854. *Grundlagen der Nationalökonomie*. 24th edn. *System der Volkswirtschaft*, vol. 1. Stuttgart: Cotta, 1906.

Schiera, P., ed. 1989. *Gustav Schmoller e il suo tempo: la nascita delle scienze sociali in Germania e in Italia*. Milan: Il Mulino.

Schmoller, G. 1901–04. *Grundriβ der allgemeinen Volkswirtschaftslehre*. 2 vols. Leipzig: Duncker and Humblot.

Schmoller, G. 1911. Volkswirtschaft, Volkswirtschaftslehre und -methode. In *Handwörterbuch der Staatswissenschaften*, vol. 8, ed. J. Conrad et al. 3rd edn. Jena: Fischer.

Schumpeter, J. 1926. Gustav v. Schmoller und die Probleme von heute. *Schmollers Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft* 50, 337–88.

Steiner, P. 2003. Durkheim's sociology, Simiand's positive political economy and the German Historical School. *European Journal of the History of Economic Thought* 10, 249–78.

Sugiyama, C. and Muzuta, H., eds. 1988. *Enlightenment and Beyond: Political Economy Comes to Japan*. Tokyo: University of Tokyo Press.

Weber, M. 1902–05. *Roscher and Knies: The Logical Problems of Historical Economics*, trans. G. Oaks. New York: Free Press, 1975.

How to cite this article

Pearson, Heath. "Historical School, German." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000176> doi:10.1057/9780230226203.0740

The New Palgrave Dictionary of Economics Online

history of economic thought

Craufurd D. Goodwin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Attention was paid to the history of economic thought (HET) by pioneers of economics such as Dupont de Nemours and Adam Smith. Classical economists like J.R. McCulloch in the 19th century used HET to establish a canon of economic literature, and their successor marginalists such as William Stanley Jevons to demonstrate progress in the subject. From the First World War until the 1960s, leading economists, from Jacob Viner to Wesley Mitchell, employed HET to cast light on current research. In the 1970s HET became a separate sub-discipline with its own periodicals and meetings. The number of scholars who worked in HET did not decline, even though the major research and postgraduate training centres lost interest.

Keywords

American Economic Association; Austrians economics; Blanqui, J.-A.; Böhm-Bawerk, E. von; Cannan, E.; classical economics; Du Pont de Nemours, P. S.; economic history; Historical School; history of economic thought; History of Economics Society; Ingram, J. K.; James, E. J.; Jevons, W. S.; Knight, F. H.; marginal revolution; Marshall, A.; McCulloch, J. R.; Mitchell, W. C.; Physiocrats; Schumpeter, J. A.; Smith, A.; Twiss, T.; Viner, J.

Article

The history of economic thought (hereafter HET) is explored today for the most part within a sub-discipline of economics. (The literature on this topic is rather limited. Blaug, 1991 is an anthology of relevant articles. Two useful bibliographical works are Howey, 1982 and Stark, 1994. The history of economic thought in Britain is examined in Backhouse, 2004. Selected histories of economic thought are reprinted in Backhouse, 2000.) It shares a category in EconLit, the indexing service of the American Economic Association, with methodology, where it is called ‘Schools of Economic Thought’. Scholars in the sub-discipline conduct various kinds of studies: interpretive biographies, narrative accounts of the

growth of ideas and their impact on society, rational reconstructions of the emergence of theory, the behaviour of scientific and intellectual communities, and more. Some 476 members of the American Economic Association declared ‘methodology and the history of economic thought’ as a field of interest in 2006. There are more than 1,000 scholars seriously interested in HET worldwide. The three main journals in the field (*History of Political Economy*, *Journal of the History of Economic Thought* and *European Journal of the History of Economic Thought*) have a combined circulation of about 2,000. Approximately 200 scholars attend each of the annual meetings of the continental societies for the study of HET, and the Japanese society has over 800 members.

The location and style of HET today are in contrast to those of the histories of most other scientific disciplines, which are found usually not within the discipline under study but within one of the sub-disciplines of history known as ‘history of science’ or ‘intellectual history’. Only the more humanistic disciplines like literature and art history and, within the social sciences, political science tend still to study their history within their known communities. Unlike those studying most other scientific disciplines, historians of economics have generally been trained as economists rather than as historians; this training gives them the perspective on their subject of insiders, but also, sometimes, the historical skills of amateurs. Scholars of HET are likely to teach in economics, not in history.

From approximately the First World War until the 1960s HET was lodged comfortably in the ‘core’ of economics. One or two courses were required of students at both the undergraduate and graduate levels, taught alongside micro and macro theory and statistics. Economics faculty began their courses on almost any subject with an introduction to the evolution of relevant theory. Indeed, HET was thought of as simply an historical extension of theory, and practitioners as simply a special kind of theorist with a long time horizon. Scholars of HET met other economists at conferences of the national and international economics societies. They did not think of themselves as a separate sect within the discipline, and saw no reason to have their own meetings or associations. They published in the mainstream economics journals and in the publications of several friendly adjacent disciplines such as history, philosophy, sociology and political science.

However, in the 1950s and 1960s this landscape changed. HET was banished from the core of economics to the margins of the discipline, ostensibly to make room for more technical economic theory and burgeoning econometrics. From being a requirement in the curriculum, HET became an option for graduate and undergraduate students – if there was someone to teach it, and increasingly there was not. The mainline professional societies and journals showed less and less hospitality to HET. Even the sister sub-discipline, economic history, then in the grip of the cliometric revolution and under scrutiny itself for relevance, seemed more and more uneasy about close relations with a subject that was ‘literary’. More and more of the major postgraduate training programmes abandoned HET formally when those who taught the subject retired and were not replaced.

The response to this crisis among those in HET in the 1960s was to regroup and create a new infrastructure in which to operate, and a sub-discipline of HET effectively came into existence. The first journal dedicated exclusively to the field, *History of Political Economy*, began in 1969, and the History of Economics Society (HES) for specialists in the subject was established in 1974. Both of these new institutions, although based in the United States, were intended to serve a worldwide community. Joint sessions of the HES with the American Economic Association and other bodies of economists continued, but the HES annual meetings became the most popular gatherings where specialists might gather and interact. A paradoxical situation, then, exists in HET in the first decade of the 21st century.

While the memberships in societies and numbers of books and articles published annually is at least stable, coverage of the subject in the premier graduate training and research centres and in the mainstream periodicals of economics has steadily decreased almost to nothing. In the United States those few graduate students who specialize in the field do so usually through a jerry-built tutorial programme with a faculty mentor, and a dissertation dictated by the job market made up of only one essay in HET and two in more saleable fields. External funding for HET, unless it is camouflaged as policy studies or theory, is almost non-existent. So what then explains the impressive place gained by HET in the economics discipline at the middle of the 20th century and its precipitate fall, in prestige and respect, within the larger discipline at least, by the end of the century? The answer lies in the subject's own history, beginning in the 18th century. Five distinct historical periods can be discerned.

Period I. The Enlightenment: HET as rhetoric

HET began at about the same time as the discipline that it studies. The 18th-century Physiocrats clearly held in low regard many of the early thinkers on questions with which they were engaged; for example, they often denigrated the thinking of Colbert, the French Minister of Finance. But they used HET less as a weapon against those with whom they disagreed than to proclaim their own remarkable accomplishments. The Physiocrats assigned Pierre Samuel Du Pont de Nemours the task of historian. His short monograph, *De L'Origine et des progrès d'une science nouvelle* (1768), may be considered the earliest treatise in HET. Dupont claimed that Quesnay and his colleagues had for the first time discovered a body of doctrine that ‘following the nature of man, exposed the laws *necessary* for a government to make for man in all climates and in all countries’ (1768, p. 35). His book was mainly a celebration of this achievement.

Adam Smith was not as cautious in his criticism as were the Physiocrats. He was exceptionally well read, knew the economic literature of his day intimately, and was not shy about offering judgements. He cited some writers on economic topics in support of his views, from Aristotle onwards, and condemned others. But he did not in any sense produce a serious and balanced history of economic thought. He had favourites, such as his friend David Hume, and pointed out some whose ideas were intriguing, like Matthew Decker and Bernard Mandeville. But he did not present the work of his predecessors as constituting a unified body of thought or leading inexorably to his own. Smith praised the work of the Physiocrats, and especially that of ‘the very ingenious and profound author … Mr. Quesnai’. But he also condemned out of hand earlier thinkers who held fundamentally different views. About as charitable as Smith could be towards those who had expressed policy conclusions at variance with his own was that their ‘arguments were partly solid and partly sophistical’ (Smith, 1776, p. 433). Neither the Physiocrats’ self-congratulation nor Smith's imaginary debates with his predecessors were important contributions to a history of economic thought.

Period II. Classical political economy: HET for cartography and doctrinal cleansing

Thomas Robert Malthus and David Ricardo were neither very interested in nor respectful of their intellectual ancestors; they made occasional references to earlier work (Smith's Wealth of Nations was particularly important to them) but they made no systematic attempt to frame it as a whole. Not so their

immediate successors, the second generation of what came to be known as the classical economists: James Mill, Nassau Senior, Robert Torrens, James Ramsay McCulloch and others. These later classicals came increasingly to believe that, despite the continuing disputes over important points of theory, something approaching ultimate truth had been achieved in the work of the founders. Senior suggested in the 1820s even that the core of political economy could be expressed in a few simple propositions derived from the founding fathers' work. From these propositions could be inferred both principles of high policy by which governments should abide, and principles to guide individual human action (Senior, 1827, pp. 35–6). Yet among the loose community of businessmen, journalists, public servants and others who pursued classical political economy during the first three-quarters of the 19th century there was relatively little agreement about what should be included in the canon. There were various elementary primers for those entering the field but no definitive textbooks (with the possible exception of J.S. Mill's Principles in 1848), professorial oracles, or dominant professional periodicals to which one might turn for definitive judgements. Indeed, virtually anyone could make a claim for inclusion of his ideas in classical political economy simply by publishing in one of the many generalist reviews.

It was to correct this condition of seeming doctrinal anarchy and inconsistency, and to impose some discipline upon an unruly conversation, that the classical economists turned to HET. Historical investigation could, perhaps, help map the new discipline and discern who and what were respectable contributions to political economy and who and what were not. Each of the doctrinal cartographers and cleansers had his own ideas of what orthodoxy should be imposed (Villeneuve-Bargemont, 1841, even named consistency with Christian theology as a criterion for inclusion). Some were Smithians, some Ricardians and some paid allegiance to an amalgam of doctrines. But their common purpose in going to the past was to sort out just what should guide the present. An example of a work to this end is the book *View of the Progress of Political Economy in Europe since the Sixteenth Century* (1847), which contained a course of lectures delivered by Travers Twiss, Professor of Political Economy in the University of Oxford. Twiss aimed to demonstrate that genuine works of political economy, as the subject had evolved since Adam Smith, employed the scientific method, which he described as testing theory by history so as to produce results that could benefit society: 'leading doctrines are the conclusions of an enlarged experience, and are not, as many persons suppose, mere deductions from arbitrary premises skillfully assumed' (1847, p. v). Twiss described the ill effects that could follow from the 'unsound theory' of such writers as Colbert and John Law. Twiss explained clearly how he proposed to use HET as a device to purge political economy of any false doctrines by which it had become corrupted. 'I have attempted in the course of the above inquiry to assign to the chief writers their due shares respectively in furthering the progress of sound opinions, but I have purposely omitted the names of many authors of eminence, who have struggled to retard that progress, although they may have indirectly furthered it by the controversy which they have provoked' (1847, p. viii).

On several occasions John Ramsay McCulloch, like Twiss, gave an account of the progress of political economy as a morality tale. He pictured truth ultimately conquering error despite the strong forces massed against it. In his pioneering textbook *Principles of Political Economy* (first published in 1825) McCulloch included a chapter on 'the rise and progress of the science'. He explained that dissension amongst early economists had tended to discredit the subject among scientists generally, and political economists needed to present a united front: 'The differences which have subsisted among the most eminent of its professors have proved exceedingly unfavourable to its progress, and have generated a disposition to distrust its best-established conclusions' (1825, p. 14). One of McCulloch's primary

objectives was to sort out truth from falsehood, so that political economy could gain the reputation and influence that it deserved: ‘the errors with which this science was formerly infected are now fast disappearing; and a very few observations will suffice to shew, that it really admits of as much certainty in its conclusions as any science founded on *fact and experiment* can possibly do’ (1825, p. 15). McCulloch's view was that there had to be broad agreement in any subject for it to be considered a science, and therefore the history must be presented as leading towards consensus.

Histories of economic thought in the classical period often took on a distinctly nationalist tone. The cartographic function was perceived not only as filling in the map of the new discipline but also as making sure that some of the territory at least bore the home country's colours. Not all the map should be British red. The publication of these histories seems almost like the intellectual equivalent of the scramble for colonies that was in progress among the European nations at this time. Adolphe Blanqui, in what was as much an economic history of Europe as a history of economic thought, gave two chapters to Smith and Malthus wedged in between segments on the Physiocrats, Rousseau, the French Revolution and J.B. Say. He was relieved that the doctrines of the British ‘industrial school’ were no longer accepted without question thanks to the work of Sismondi and other French critics (Blanqui, 1837, p. 262). A similar work in Italian was by Luigi Cossa (1876).

Period III. Neoclassical and historical economics: HET as literature review

Beginning with the marginal revolution of the 1870s HET took on a new role derived from what had become fashionable in the physical sciences and mathematics: the literature review. If economists were to be seen as true scientists, insisted economists such as William Stanley Jevons, they must walk and talk like them. They must not use the history of their subject to demonstrate a stable orthodoxy, as McCulloch and others had sought to do. The past of a science contained not an accumulation of what was true but of what had been found to be false and had been displaced by current doctrine. In praising the accomplishments of the Austrian marginalists, Böhm-Bawerk used an evolutionary metaphor to describe HET as the study of illness in scientific infancy and childhood. The Austrians, he wrote, ‘are of the opinion that the errors of the classical economists were only, so to speak the ordinary diseases of the childhood of science … Their greatest fault was they were forerunners; our greatest advantage is that we came after’ (Böhm-Bawerk, 1973, p. 362). The essence of science was progress and change. The purpose of the literature review to be included with any major work in science should be twofold: to pay due respects to worthy ancestors, and more particularly to use the past to demonstrate how certain prior works led inexorably to the present, superior, one. The literature review in a work of theory while acknowledging worthy predecessors also established claims to priority in the novel ideas set forth. HET had come into the service of Whig history. Above all, the emphasis had to be on change rather than on stability. William Stanley Jevons insisted that attention to the past should be seen as liberating and not as stifling deference to orthodoxy. He observed how ‘in the other sciences the weight of authority has not been allowed to restrict the free examination of new opinions and theories; and it has often been ultimately proved that authority is on the wrong side’ (1871, pp. v–vi). In the books of the marginal revolutionaries the literature review was placed usually in the preface or in an appendix. Jevons used both. The marginalists were as ready as Twiss or McCulloch to dismiss some predecessors out of hand; but their dismissal was focused especially upon those who differed in particular methods or results from

the work currently being presented. All predecessors had necessarily been supplanted. Those who walked the right road, though too slowly, deserved to be remembered. Those who took the wrong road deserved to be condemned. Here is what Jevons wrote of McCulloch's heroes: ‘When at length a true system of Economics comes to be established, it will be seen that that able but wrong-headed man, David Ricardo, shunted the car of Economic Science on to a wrong line – a line, however, on which it was further urged towards confusion by his equally able and wrong-headed admirer, John Stuart Mill’ (Jevons, 1871, pp. li–lii). Jevons could congratulate Von Thunen, Dupuit, and Cournot; but for others, like John Stuart Mill, who were not on the right road to the marginal revolution, he had only contempt.

Each of the pioneer marginalists had his own way of incorporating a review of the literature into his text. In Menger the historical commentaries were long footnotes that so annoyed the translators of his *Principles of Economics* (1871) into English in 1950 that they appear there as a series of appendices. Marshall began with an introductory historical section on ‘the growth of economic science’ in the first edition of his *Principles of Economics* (1890) but shifted this material in the fifth edition (1907) to an appendix. Irving Fisher, lacking a single broad-based treatise of his own to which he could append an historical review of the literature, attached one to the translation of Augustin Cournot's *Researches into the Mathematical Principles of the Theory of Wealth* (1897). These reviews of the literature by the marginalists often have a strikingly unsystematic and personalized appearance with offhand comments that seem out of place in a carefully reasoned text. For example, the following comment by Marshall in a generally laudatory mention of Ricardo and his work seems to reflect more his own casual prejudices than a serious study of history. Marshall wrote:

> his [Ricardos's] aversion to inductions and his delight in abstract reasonings are due, not to his English education, but, as Bagehot points out, to his Semitic origin. Nearly every branch of the Semitic race has had some special genius for dealing with abstractions, and several of them have had a bias towards the abstract calculations connected with the trade of money dealing, and its modern developments; and Ricardo's power of threading his way without slip through intricate paths to new and unexpected results has never been surpassed. But it is difficult even for an Englishman to follow his track; and his foreign critics have, as a rule, failed to detect the real drift and purpose of his work’. (Marshall, 1920, p. 629 n.)

Edwin Cannan's *A History of the Theories of Production and Distribution in English Political Economy 1776–1848* (1917) was a generalized and highly critical literature review of what had become settled doctrine a half century before the ‘new’ economics of Alfred Marshall. It set out to demonstrate that only with marginal tools had economics become a science. Cannan's book was not like those of Twiss and McCulloch, which had sought to sift the wheat from the chaff in the confident belief that a pile of genuine truth would thereby be revealed. Cannan's message was that everything before marginal economics was hardly worth a glance because none of it was science.

The marginalists were not the only ones in the late 19th century to use HET to bolster the legitimacy of their approach. The Historical School also concluded that a literature review demonstrated the strength of their position. The essence of their claim was that the usefulness of economic theory was relative to

the circumstances in which the theory was applied. Different circumstances required different theory, and the history and appraisal of past theory had to keep in mind the tasks for which the earlier theory had been designed. The American historical economist E.J. James suggested that:

> the axioms and theorems which apply to one form of society may have little or no applications in another form, and any attempt to make such application may result in the most absurd conclusions … Nor will a theory which is adequate to the demands of an industrial state like England or America suit such a country [*sic*] as India or Africa. (Ingram, 1888, p. vii)

The historians wrote specifically in opposition to ‘The assertion of J.B. Say’ a doctrinal cleanser ‘that the history of Political Economy is of little value, being for the most part a record of absurd and justly exploded opinions’ (Ingram, 1888, p. 2). This they found to be an unjustified dismissal of early economic thought.

The correct way to view the history of ideas, they were convinced, was as the record of how theory was useful at particular times and places and not either as a gradual but final movement towards some kind of ultimate truth, or as a steady accretion of scientific understanding. At the same time it must be conceded that the consequence of this posture by the historians was not very different from that of the marginalists; the details of HET, they implied, were largely of antiquarian interest. The difference between them was that the historical economists looked with more sympathy upon their predecessors, even those with whom they disagreed in their modern application.

The most detailed history in English taking the historical approach was by the Irish economic historian, John Kells Ingram. Ingram's findings were in part similar to and in part a contrast to those of Cannan. He agreed with Cannan on the failings of the classical economists, and he insisted on the need to discover new theory. But his road map was different from that of Cannan. He found that the marginal successors were far too much like the classical economists they followed. He wanted a turn to modern science, but a different kind of science: an empirical science unconstrained by a body of high theory. He said: ‘the science must be cleared of all the theologico-metaphysical elements or tendencies which still encumber and deform it. Teleology and optimism on the one hand, and the jargon of “natural liberty” and “indefeasible rights” on the other, must be finally abandoned’ (1888, p. 241). Instead, economics must become an experimental science ‘forming only one department of the larger science of Sociology’ (1988, p. 242). Only in this way could economists change ‘the attitude of true men of science towards this branch of study, which they regard with ill-disguised contempt, and to whose professors they either refuse or very reluctantly concede a place in their brotherhood’ (1988, p. 240).

Other contemporary interpretations of HET in the same tradition as Ingram, that economic ideas were necessarily embedded in economic history, were posited by Price (1891) and Ashley (1894).

Period IV. The golden age: HET as heuristic device

Beginning around the First World War and continuing for almost half a century, HET went through a remarkable transformation. After serving in the 19th century as little more than a minor weapon in the arsenals of combatants in one professional conflict or another, and appropriately consigned to prefaces

and appendices by major figures and taken up extensively by no more than a few minor ones, HET came now to be pursued with energy and great seriousness by many of the leading figures in economics. Many of these converts produced significant book-length studies; others wrote articles. Some who did not devote years or an entire career to the subject still engaged in it soberly for the production of one or two studies before moving on. This new approach was not the 'throwaway HET' that had come before. Nor was it simply hagiography by members of a proud new community of professional economists. The authors in the golden age were committed to understanding problems through use of HET as an analytical device. They saw HET as heuristically significant. The golden agers did not think of HET as a separate new sub-discipline, as ultimately it was to become, but as an overlay of all economics, a distinct approach to all economic problems that should be explored as fully as other theoretical and empirical approaches. Moreover, the new interest was not confined to those holding any one ideological, methodological or doctrinal position. The following is an incomplete but illustrative list of some of those prominent economists who engaged in HET during this golden age apparently in search of answers to pressing questions: among the Austrian marginalists, J.A. Schumpeter, Gottfried Haberler, Karl Pribram, Erich Schneider, and Fritz Machlup; among English and American marginalists, John Hicks, Lionel Robbins, Frank Knight, George Stigler, and Jacob Viner; among the American Institutionalists, Wesley Mitchell, John R. Commons, Clarence Ayres, and John Kenneth Galbraith; among those intrigued with Marx, Eric Roll, Martin Bronfenbrenner, John Elliott, and Maurice Dobb; and among the new macroeconomists Piero Sraffa, G.L.S. Shackle, Gunnar Myrdal, and John Maynard Keynes himself. It was during this time that serious interpretive HET, rather than simply obituary notices, literature surveys, and review articles, entered the main publications of the profession, in writings by major figures such as those listed above, and lesser lights. HET was not only welcomed by the 'top' journals during the golden age, it became routinely the subject of presidential addresses and other ceremonial pronouncements. Most of the senior economists who took up HET also gave graduate courses in the field, and they encouraged some of their best graduate students to write dissertations in the area and to contemplate specializing in the field professionally.

Why this sudden turnabout? Why this unexpected fascination with history at the highest levels in the discipline? The most likely explanation lies in the circumstances of the time, which were certainly very different from those of the century before. Above all, a loss of confidence struck economics after the First World War. Before the war, economists of the mainstream such as Alfred Marshall, John Bates Clark, Léon Walras and Carl Menger concluded that they worked in an advancing science of a conventional sort and that they had the answers to most observable problems. The First World War, and the depression that followed, shattered all illusions that economic problems were that simple. No longer was it clear that relatively unconstrained rational men living in democracies and free market economies could count on enjoying peace and prosperity. The evidence seemed to prove the contrary and to suggest that all social constructs perfected during the Victorian age, including the global economy based on European empires, had to be re-examined from bottom up. Economics could not yet think of itself, as Keynes suggested it might be able to do some day, as analogous to dentistry seeking progress through technical improvements in familiar procedures. Where there had once been certainty now there was mainly doubt. And all of a sudden it seemed for many economists that HET might point the way toward undiscovered answers to some at least of the challenges newly arisen. HET was recognized as a vital tool in research. It could help economists find their bearings at several levels as they sought to be useful. Another factor behind the new interest in HET may have been the kind of scholar attracted to the

economics discipline at this time. The questions that were coming to the fore were not of a type that could be addressed effectively by narrow technicians, and the questions attracted persons who insisted on supplementing conventional economic analysis with philosophical, sociological, psychological and historical enquiry. So what were these questions that prominent economists came to believe might be tractable through HET? They were methodological, including, how to reconcile and integrate the approaches of the different national traditions of marginalist economics, for example British and American partial equilibrium with the general equilibrium of the Walrasians? Were mathematical economics and econometrics essential to progress within the discipline, and how should they be used? More generally, was it possible to retain under one disciplinary tent economists who were so different in their approaches and objectives as the varieties of neoclassical marginalists, Institutionalists, economic historians, Marxists, Keynesians and others? Was such heterogeneity virtue or vice? The questions were also theoretical; might early and forgotten theory cast light on such topics of sudden new concern as imperfect markets or business cycles? And some questions were directly policy oriented. What was the proper place for economics, and economists, in the policy process? Should the economists, rejecting the advice of most marginalist pioneers, sally forth from their ivory towers and connect directly with policymakers, perhaps even entering government as the German historians had done? If so, how? Should there be a ministry of economic affairs? Advisory councils to political leaders? Think tanks entirely outside of government? What about central planning? The Russian Revolution of 1917 raised this question for urgent public reconsideration even though it seemed to be settled for most professional economists by that date.

On all these questions, in contrast to the sense of self-confidence that characterized the first decade of the 20th century, when the most serious issues of economic policy were how to perfect the fine-tuning of the welfare economics of A.C. Pigou, the post-war mood demanded creative and fresh thinking. A notorious manifestation of this thinking across the disciplines was the hugely successful set of short biographies, *Eminent Victorians* (1918), by Lytton Strachey in which four prominent 19th-century institutions were held up for re-examination and reform: the military, the Church, the public schools and Victorian woman. Might this kind of historical enquiry reveal where the economy and economics had gone wrong, and show how they might be put back on the right track? Certainly Keynes believed so when he wrote *The Economic Consequences of the Peace* (1919), patterned substantially after *Eminent Victorians*.

Large structural questions without and within the economics discipline also were raised by the First World War and its aftermath. Was economics truly a science? This question became critical again during and after the Second World War, when public support for science, more than for other forms of enquiry, was contemplated and then implemented. These were years when the sub-disciplines of economics were just getting organized, and questions of boundaries and inclusions or exclusions had to be addressed. To some prominent scholars HET seemed a promising place to seek guidance. Jacob Viner's *Studies in the Theory of International Trade* (1937), Joseph Spengler's *French Predecessors of Malthus* (1942), Arthur Marget's *The Theory of Prices* (1938–42), Gottfried Haberler's *Prosperity and Depression* (1937), George Stigler's *Production and Distribution Theories* (1941) and Arthur Cole's *The Historical Development of Economic and Business Literature* (1957) were all milestones in HET and in the formation of the sub-disciplines of, respectively, international economics, economic demography, macro-economics, industrial organization and management science. Not all of those who pursued HET

in the golden age were the Renaissance men of the discipline. A few specialists did focus on single figures from the past, for example, Werner Stark on Bentham, Piero Sraffa on Ricardo, William Jaffé on Walras, and Joseph Dorfman on Veblen.

Not many of the giants of the golden age explained in detail the reasons for their new commitment to HET. Often the most we have to go on is an offhand remark or two. Jacob Viner said that his objectives were ‘to resurrect forgotten or overlooked material worthy of resurrection, to trace the origin and development of the doctrines which were later to become familiar, and to examine the claims to acceptance of familiar doctrine’ (Viner, 1937, p. xiii). For Joseph Schumpeter the study of HET was an integral part of discovering a vision of economic evolution, which contained the key to understanding the economy (Schumpeter, 1954). Frank Knight remarked that ‘A major lesson to be learned from the history of ideas is to realize the ‘glacial’ tardiness of men, including the best minds, in seeing what it later seems should have been obvious at the first look’ (1973, p. 46). Wesley Mitchell explored the question at some length at the start of his classes in HET at Columbia University and his reflections are revealing. In the transcription of his lectures, edited by Joseph Dorfman, Mitchell says that HET is necessary not so much to understand modern economics as to advance the subject through graduate education and research:

> All that I contend for is that so long as the social sciences continue to make progress each generation of economists will find problems in the history of their science which earlier generations have not thought out, and that these problems will probably attract workers who feel their fascination; that is, I think there is a difference between the social sciences and the natural sciences, which makes the past history of their subjects more interesting and more pertinent to the workers in the social field than to workers in the natural-science field.
>
> Our interest in the history of economics changes with the development of economics itself. The history of economics needs to be re-written by every generation of economists for the same reason that history at large needs to be re-written. (Mitchell, 1967, p. 2)

Mitchell's point was an important one. He suggested that HET was valuable especially for graduate students and young scholars who had the responsibility ultimately to move economic science forward. Without historical sensibility graduate students would be at a serious disadvantage on the research frontier. Mitchell said that the HET he was teaching was fundamentally different from that which had come before:

> Working in this spirit we find ourselves concerned more with the larger aspects of economic history than our predecessors. What we can get light upon and what we therefore think most about is not the letter of the laws laid down, the traces of a man's thinking to be found in his predecessors, the logical inconsistencies which minute criticism may develop among his formulations – it is not these things which interest us so much as the type of problems the man attacks, his way of formulating them, what materials he had to work with, the general method he employed, the things he took for granted without inquiry, the grounds for the confidence he felt in his results, what use he

> put these results to, their acceptance or rejection by his contemporaries and the reaction of his scientific work upon social processes. (1967, pp. 6–7)

Mitchell suggested that HET should help economists gain ‘knowledge of ourselves and free us from over-narrow specialization’ (1967, p. 7). It would also ‘give us clearer insight into the conditions which promote or retard the progress of knowledge in the social sciences. Perhaps some at least among these conditions will prove to be amenable to control’ (1967, p. 7). HET might also give students the background with which to select among rival theoretical claims. ‘Some of them become neo-Marshallians, some neo-Marxists, some neo-Austrians, some mathematical theorists, some institutionalists. If anyone is going to make any such choice he ought to make it with open eyes; i.e. he ought to understand what other types of theory are; what they offer. If he knows, perhaps he wont become an ardent follower of any school’ (1967, p. 10). Finally, Mitchell noted that the sheer joy of historical inquiry should attract students to it. ‘The fascination of the work itself, the possibility of gaining keener insights and more certainty as we follow up our leads, may have more to do with the future progress of such work than the indirect gains it promises for economic theory’ (1967, p. 8).
This golden age of HET came to an end in the 1950s and 1960s. The cause of its death is as much a puzzle as its birth. One explanation could be that most of the leading figures retired or left the field. But that is a description of what happened more than an explanation. Why did these leaders not have successors? Why was not the next generation of leaders in economics fascinated in the same way by the history of their subject? The best explanation seems to be that by the 1960s economics had once again regained its self-confidence and there was a reversion to the set of attitudes that prevailed before the First World War. Most of the issues that appeared after the war (depression, doctrinal conflict, war itself) seemed either to be answered or to have gone away by the 1950s. There was no longer a need to look backwards, it seemed, only ahead. One of the most powerful forces leading to a high level of self-confidence in economics was its own performance during the Second World War compared with that during the First. Macroeconomic understanding proved helpful in maintaining full employment with price stability, while optimizing models taken directly from applied microeconomics and sometimes including the new tool of game theory were found to be useful in processes as different as aiming a machine gun and planning air raids.

Period V. Building a new sub-discipline of HET

Most close observers of HET in the 1950s and 1960s might have predicted that its life within the economics discipline was over and that it was on its way to join the histories of other academic subjects in the deep recesses of history departments. At best it might leave a few champions within the larger discipline, such as Edwin Cannan proclaiming the faults of the old and the promise of the new. But this did not happen. HET lived on in economics, albeit without the powerful leaders of the golden age, without a place in most of the prominent research departments, and indeed without many opportunities for graduate training. So, without these assets how did the field survive? Several factors seem to have been in play.
The most important factor may have been the momentum carried over from the golden age. While most of those who had turned to history as a heuristic tool were gone by the 1960s, a few remained, and during this decade they joined in preparing a response to the new charge of irrelevancy. In the lead were

George Stigler, Lionel Robbins, Terence Hutchison, Joseph Spengler, Joseph Dorfman and Martin Bronfenbrenner. Also sympathetic but less directly involved were Kenneth Arrow, Kenneth Boulding, James Buchanan, John Chipman, Earl J. Hamilton, Paul Samuelson and James Tobin. But more important than this rearguard action by the last golden agers was a cadre of young and middle-aged scholars trained in HET and committed now to retaining it within the economics discipline. These children of the golden age were well placed in teaching jobs and their careers had often been encouraged by their mentors. From being an overlay of the economics discipline during the golden age, HET moved during the 1960s and 1970s to become an independent sub-discipline, led by, among others: in Britain, R. D.C. Black, Mark Blaug, Tony Brewer, A.W. Coats, David Collard, Ronald Meek, Denis O'Brien, Andrew Skinner, and Donald Winch; in the USA and Canada, William Allen, William Barber, Hans Brems, Robert Ekelund, Frank Fetter, William Grampp, Samuel Hollander, Todd Lowry, Larry Moss, Mark Perlman, Warren Samuels, Robert Smith, Vincent Tarascio, Carl Uhr, Anthony Waterman and Donald Walker; in Israel, Haim Barkai and Ephraim Kleiman; John Pullen, Michael White, and Peter Groenewegen in Australia. Outside the English-speaking world leadership was taken by, among others, Pier Luigi Porta, Maria Cristina Marcuzzo, and Pierangelo Garegnani in Italy; Erich Streissler in Austria, Heinz Kurz, Harald Hagemann, and Bertram Schefold in Germany; Arnold Heertje in the Netherlands; Yuichi Shionoya and Takashi Negishi in Japan; and Lars Jonung and Bo Sandelin in Sweden. In addition to building and supporting the infrastructure of specialized periodicals and societies, such as *HOPE*, *JHET*, *EJHET*, and others, these scholars helped to mobilize and sustain a variety of other resources that have strengthened the field: translations and republications of canonical writings, collected works and letters of major authors, variorum editions, and ephemera, as in the Kress-Goldsmith micro-film project of works published before 1800. Collections of manuscripts of prominent economists, saved sometimes at the last minute from the garbage dump, made possible for the first time the close study of the interactions among economists and how they constructed their articles and books. The most substantial of these is the Economists' Papers Project at Duke University in the United States. In the United Kingdom the guide to archives prepared by Paul Sturgess documented where materials were located in that country. Access to manuscripts made possible meticulously documented biographies of great economists, for example, of Marshall by Peter Gronewegen (1995), of Hayek by Bruce Caldwell (2004) and of Keynes by Donald Moggridge (1992). Increasingly HET was defined as ending as recently as yesterday, and so oral history too became an essential tool of the historian.

An important movement that began in the 1960s was to explore ways in which HET could be incorporated more successfully into the curriculum of graduate students, economics majors, and even non-specialist liberal arts undergraduates. The teaching of HET in the golden age had been confined very largely to graduate and honours students using original sources and a few commentaries from the secondary literature. The textbooks that were available were by then very old – for example, those by Gray (1931), Gide and Rist (1909), and Haney (1911) – and not very appealing. The first rigorous new-style textbook, mainly for graduate students, was Mark Blaug's *Economic Theory in Retrospect* (1962). It concentrated on expressing old ideas in modern guise. Other similar texts that joined it over the years were by Hans Brems (1986) and Jurg Niehans (1990). A plethora of textbooks for undergraduate courses were published with styles, degrees of rigour, and ideologies for most tastes (for example, those by Landreth, 1976; Ekelund and Hebert, 1975; Rima, 1967; and Spiegel, 1971). One of the pioneering works in this genre was William Barber's *History of Economic Thought* (1967). An important

publication landmark was Robert Heilbroner's *The Worldly Philosophers* (1953) which, with sales reputedly above a million copies, attracted generations of undergraduates to a more extended investigation of HET. Although leaders of the economics discipline in the years after the golden age expelled HET from the graduate curriculum (not even Blaug's new textbook could stem that tide), it was important for the employment prospects of those trained in HET that the appeal of the subject as an elective course for undergraduates remained.

Progress in research in HET since the 1970s has helped to sustain the positive response to the challenge of the 1950s and 1960s. The creation of a new sub-discipline was strengthened by the flush of interest in the philosophy of science in the 1970s. There were stimulating attempts to use new interpretive tools derived from the writings of Thomas Kuhn (1962), Imre Lakatos (1970), and others to understand the history of economics. And Deidre McCloskey's examination of the rhetoric of economics (1998) reverberates still in HET. Other substantial research projects that were a stimulus to the new sub-discipline of HET, both as inspiration and as source of consternation, include Samuel Hollander's reconsideration and reinterpretation of classical economics (1973; 1979; 1985; 1996), Philip Mirowski's exploration of the linkages between the history of economics and progress in other disciplines (1989), Roy Weintraub's account of the mathematization of economics (2002b), and studies of developments in modern economics by Mary Morgan (1990), Esther-Mirjam Sent (1998), Judy Klein (1997), and others.

The emergence of a new generation of leaders of HET in the decades after the golden age, leaders who were able to gain secure positions in colleges and universities, has been a reassuring development. These include Jurgen Backhaus, Roger Backhouse, Bradley Bateman, Peter Boettke, Mauro Boianovsky, Bruce Caldwell, Jose Luis Cardoso, Avi Cohen, David Colander, William Coleman, John Davis, Robert Dimand, Neil De Marchi, Ross Emmett, Jerry Evensky, Evelyn Forget, Dan Hammond, Wade Hands, Robert Hebert, Kevin Hoover, Sue Howson, John King, Judy Klein, Robert Leonard, John Lodewijx, Harro Maas, Steven Medema, Perry Mehrling, Don Moggridge, Mary Morgan, Malcolm Rutherford, Margaret Schabas, Neil Skaggs, Karen Vaughn and Jim Wible. Often these scholars have combined their interest in HET with commitment to another sub-field of economics, sometimes by keeping their interests in HET quiet until they achieved tenure. These grandchildren of the golden age, as it were, have kept the momentum for the perpetuation of the new sub-discipline alive into the 21st century.

Certain developments outside HET as well as within helped to strengthen the field in the latter decades of the 20th century. A number of distinguished economists moved to history rather late in their careers. Usually they addressed questions still alive in their original sub-disciplines, but they have employed the historian's tools and perspectives. Examples of these mid-career migrants to HET include Walter Eltis, Geoff Harcourt, Don Patinkin, David Laidler and John Whitaker.

A second kind of migrant has been more problematic for HET. When the homogenization of economics reached a crescendo in the 1980s and 1990s, some of those who felt alienated or squeezed out of the discipline for methodological or ideological reasons found comfort and welcome in HET. Some who resisted the increasing technical complexity of the new theory also sought refuge in this cross-over. These refugees, while providing welcome additions to the ranks of HET and offering different perspectives on a variety of issues, have tended to mark the entire sub-discipline as made up of malcontents.

A third kind of migrant to HET came from specialized communities within economics that had become too small or marginalized to continue on their own. They sought and received hospitality within HET whether their interests were primarily historical or not. They include some Marxists, neo-Austrians, Post

Keynesians, Institutionalists, Sraffians, and others.
HET has been enriched in recent years by visits, short or long, from members of other disciplines who came not as refugees but attracted by specific research questions. They came from social, intellectual, political and economic history, as well as from sociology, philosophy and political science. Prominent visitors have included Peter Clark, Robert Skidelsky, Heath Pearson, S.M. Amadae and Yuval P. Yonay.

The prospect ahead

The future of HET is uncertain (Weintraub, 2002b). On the one hand, the strong infrastructure of societies and publications is encouraging, as are the numbers of scholars who identify with the field. It is gratifying, moreover, that the field has demonstrated persuasively its capacity to survive adversity and to face challenges constructively. But though these are reasons for optimism for the future there are reasons also for unease. And this leads to a final question. What uses will be found for HET in the future, and can any of these be discerned from study of the past? The original use for HET in the rhetoric of policy debates persists, but mainly on the surface. Libertarians wear Adam Smith ties and opponents of an active government in the economy dismiss their opponents collectively as Keynesians, but in both cases the combatants understand little beyond the labels. HET as doctrinal cleansing is still performed, but mainly in review articles and chapters, such as those in the *Journal of Economic Literature* and the various Handbook series, prepared not by specialists in HET but by high priests of the various sub-disciplines. The more focused and celebratory literature reviews, such as those that gained popularity after the marginal revolution, can be found still in Nobel Prize acceptance speeches and presidential addresses, but neither serious history nor professional historians of economic thought are much involved. In this spirit are the innumerable biographical and hagiographic dictionaries of ‘great economists’ categorized in various ways, as women, dissenters, or something else. The use for HET which was its greatest strength during the golden age, in the training of graduate students and in the search for answers to large questions on the research frontier, has largely disappeared, and there seems no immediate prospect of it being resurrected. Among the more recent uses for HET as a home for refugees of various kinds and as a component in the undergraduate curriculum to relieve the tedium of increasingly technical abstraction, only the latter seems secure and likely to grow in strength.
The overriding question remains: can a sub-discipline survive for long when it is little valued by the discipline of which it is a part and where there is no graduate training available through which to sustain and renew the leadership? One bright spot may be the liberal arts college, where breadth as well as depth is still rewarded and which is likely to express forcefully in the labour market its preferences for kinds of faculty training. Or it may take another loss of confidence within the economics discipline overall, such as that experienced early in the 20th century, to cause economists to find once again something of relevance in their past!

Bibliography

Ashley, W.J. 1894. *An Introduction to English Economic History and Theory*. 3rd edn. London: Longmans Green.

Backhouse, R.E. 2000. *Early Histories of Economic Thought 1824–1914*. London: Routledge.

Backhouse, R.E. 2004. History of economics, economics and economic history in Britain, 1824–2000. *European Journal of the History of Economic Thought* 11, 107–24.

Barber, W. 1967. *A History of Economic Thought*. New York: Penguin.

Blanqui, A. 1837. *Histoire de L'Economie Politique*. Paris: Guillamin.

Blaug, M. 1962. *Economic Theory in Retrospect*. Homewood, IL: Irwin.

Blaug, M. 1991. *The Historiography of Economics*. Aldershot: Edward Elgar.

Böhm-Bawerk, E.V. 1973. The Austrian economists. In *Dictionary of the History of Ideas*, ed. P. Wiener. New York: Charles Scribner's Sons.

Brems, H. 1986. *Pioneering Economic Theory, 1630–1980*. Baltimore: Johns Hopkins University Press.

Caldwell, B. 2004. *Hayek's Challenge*. Chicago: University of Chicago Press.

Cannan, E. 1917. *A History of the Theories of Production and Distribution in English Political Economy 1776–1848*. Bristol: Thoemmes, 1994.

Cole, A.H. 1957. *The Historical Development of Economic and Business Literature*. Boston: Harvard Graduate School of Business Administration.

Cossa, L. 1876. *Guida allo studio dell'economica politica*. Milano: Ulrico Hoepli. (Eng. trans. 1880.)

Cournot, A. 1897. *Researches into the Mathematical Principles of the Theory of Wealth*. Homewood, IL: Irwin, 1963.

Eckland, R.B. and Hebert, R.F.. 1975. *A History of Economic Theory and Method*. New York: McGraw-Hill.

Du Pont De Nemours. 1768. *De L'Origine et des progrès d'une science nouvelle*. Paris: Geuthner, 1910.

Gide, P.H.C. and Rist, C. 1909. *Histoire des doctrines économiques*. Paris: L. Larose and L. Tenin.

Gray, A. 1931. *The Development of Economic Doctrine*. London: Longmans Green.

Gronewegen, P.D. 1995. *A Soaring Eagle*. London: Edward Elgar.

Haberler, G. 1937. *Prosperity and Depression*. Geneva: League of Nations.

Haney, L. 1911. *History of Economic Thought*. New York: Macmillan.

Heilbroner, R. 1953. *The Worldly Philosophers*. New York: Simon and Schuster.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hollander, S. 1973. *The Economics of Adam Smith*. Toronto: University of Toronto Press.

Hollander, S. 1979. *The Economics of David Ricardo*. Toronto: University of Toronto Press.

Hollander, S. 1985. *The Economics of John Stuart Mill*. Toronto: University of Toronto Press.

Hollander, S. 1996. *The Economics of Thomas Robert Malthus*. Toronto: University of Toronto Press.

Ingram, J.K. 1888. *A History of Political Economy*, with a preface by E.J. James. New York: Macmillan.

Jevons, W.S. 1871. *The Theory of Political Economy*. New York: Augustus M. Kelley, 1965.

Keynes, J.M. 1919. *Economic Consequences of the Peace*. New York: Harcourt Brace, 1920.

Klein, J. 1997. *Statistical Visions in Time*. Cambridge: Cambridge University Press.

Knight, F. 1956. *On the History and Method of Economics*. Chicago: University of Chicago Press.

Knight, F. 1973. Economic History. In *Dictionary of the History of Ideas*, ed. P. Wiener. New York: Charles Scribner's Sons.

Kuhn, T.S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakatos, I. and Musgrave, A. 1970. *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Landreth, H. 1976. *History of Economic Thought*. Boston: Houghton Mifflin.

Marget, A.W. 1938–1942. *The Theory of Prices*. New York: Prentice-Hall.

Marshall, A. 1920. *Principles of Economics*. 8th edn. London: Macmillan, 1964.

McCloskey, D. 1998. *The Rhetoric of Economics*, 2nd edn. Madison: University of Wisconsin Press.

McCulloch, J.R. 1825. *The Principles of Political Economy*. London: William, 1830.

McCulloch, J.R. 1845. *The Literature of Political Economy*. London: London School of Economics and Political Science, 1938.

Menger, C. 1950. *Principles of Economics*, trans. and ed. J. Dingwall and B.F. Hoselitz. Glencoe, IL: Free Press.

Mill, J.S. 1848. *Principles of Political Economy*, ed. W.J. Ashley. London: Longmans, 1909.

Mirowski, P. 1989. *More Heat than Light*. Cambridge: Cambridge University Press.

Mitchell, W.C. 1967. *Types of Economic Theory*. New York: Augustus M. Kelley.

Moggridge, D. 1992. *Maynard Keynes: An Economist's Biography*. New York: Routledge.

Morgan, M. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Niehans, J. 1990. *A History of Economic Theory*. Baltimore: Johns Hopkins University Press.

Price, L.L. 1891. *A Short History of Political Economy in England, from Adam Smith to Arnold Toynbee*. London: Methuen.

Rima, I. 1967. *Development of Economic Analysis*. Homewood, IL: R.D. Irwin.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Senior, N. 1827. *An Introductory Lecture on Political Economy*. London: J. Mawman.

Sent, E.-M. 1998. *The Evolving Rationality of Rational Expectations*. Cambridge: Cambridge University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Clarendon, 1976.

Spengler, J.J. 1942. *French Predecessors of Malthus*. Durham, NC: Duke University Press.

Spiegel, H. 1971. *The Growth of Economic Thought*. Durham, NC: Duke University Press.

Stark, W. 1994. *History and Historians of Political Economy*. New Brunswick: Transaction.

Stigler, G.J. 1941. *Production and Distribution Theories, 1870 to 1895*. New York: Macmillan.

Strachey, L. 1918. *Eminent Victorians*. London: Chatto and Windus.

Twiss, T. 1847. *View of the Progress of Political Economy in Europe since the Sixteenth Century*. London: Longman.

Villeneuve-Bargemont, J.-P.-A. 1841. *Histoire de l'économie politique*. Paris: Guillaumin.

Viner, J. 1937. *Studies in the Theory of International Trade*. London: George Allen and Unwin, 1955.

Weintraub, E.R. (ed.) 2002a. *The Future of the History of Economics*. Durham, NC: Duke University Press.

Weintraub, E.R. 2002b. *How Economics Became a Mathematical Science*. Durham, NC: Duke University Press.

How to cite this article

Goodwin, Craufurd D. "history of economic thought." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000174> doi:10.1057/9780230226203.0741

The New Palgrave Dictionary of Economics Online

Hobson, John Atkinson (1858 – 1940)

Entry in the first edition by Peter Clarke revised by Roger E. Backhouse and P.J. Cain
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

John Atkinson Hobson, a self-styled economic heretic, had a long and prolific career as an economist and political activist. His heresies included underconsumptionism and a critique of orthodox welfare economics based on ideas from John Ruskin, the former being elaborated into a theory of imperialism that influenced Lenin. He was belatedly recognized as a forerunner by Keynes in his General Theory, but this does not do justice to the range of Hobson's work.

Keywords

accelerator; Clark, J. B.; effective demand; Fabian economics; forced gains; Hobson, J. A.; imperialism; involuntary unemployment; Keynes, J. M.; Lenin, V. I.; living wage; marginal productivity theory; marginalism; Mill, J. S.; Pigou, A. C.; productive and unproductive surplus; progressive and regressive taxation; protection; quantity theory of money; rent; surplus value; underconsumptionism

Article

John Atkinson Hobson was born in Derby in 1858 and died at home in Hampstead in 1940. He was educated at Derby School and Lincoln College, Oxford, where he read Greats from 1876 to 1880, but only gained a Third. He taught classics at Faversham and Exeter in 1880–81, before moving to London, where he supplemented his private income (from the Derby newspaper which his father had owned) with intermittent earnings from journalism, lecturing and his books (Clarke, 1978). A prolific writer, he propagated his economic views through more than 50 books and 700 articles, many of them in a series of organs of radical liberal and socialist leanings. Hobson thus left an oeuvre which is not easy to assess and in which formal inconsistencies are not difficult to find: but he conveys, nonetheless, a general vision of the scope and nature of economics that is both distinctive and coherent. His reputation has been coloured by his supposed role as a predecessor not only of Lenin and his theory of imperialism but also of Keynes and his concept of effective demand. Neither connection is wholly factitious but both have

been open to unhistorical distortions of Hobson's own concerns.

Hobson has long been best known as an underconsumptionist. His first book (Mummery and Hobson, 1889) was written in collaboration with A.F. Mummery, a businessman, who seems to have been the senior partner. The book set out to expose fallacies in classical political economy as expounded by J.S. Mill. Its central proposition was that trade depression was caused by a deficiency in effective demand since it was the level of consumption in the immediate future that limited profitable production. It followed that there was a limit to the amount of useful savings which a community could make. Each individual could save with advantage to himself, but the overall result might be a position of underconsumption, for which over-saving was another name. Hobson was to seize on this self-defeating process as an example of what he called the protean fallacy of individualism – an idea that pervades his work in a far more general way than the particular concept of underconsumption. The polemical thrust of this early book was thus against the tendency of economists to extol thrift in so far as this neglected the crucial importance of maintaining sufficient demand. Hobson and Mummery provided an account (complete with a numerical example) of the accelerator, a concept commonly believed to have originated in the 20th century (1889, pp. 85–6; cf. Backhouse, 1990). Though the book attracted hostile comment from established economists, it did not, as Hobson alleged, blight his career. He carried on teaching economics as a university extension lecturer, the job for which he was well suited temperamentally (Kadish, 1990). Later, he was proud to proclaim himself an ‘economic heretic’ (Hobson, 1938).

This early statement of the underconsumptionist case was reiterated in two further books (Hobson, 1894; 1896) the second of which made use of the newly coined term ‘unemployment’, defining it in terms of involuntary leisure suffered by the working classes. He broadened rather than narrowed his dissent from neoclassical analysis through his distrust of marginalism, which he rejected on the ground that it rested upon an unreal individualism, marking a further breach with Marshallian orthodoxy (Hobson, 1901b; 1926a). A later book (Hobson, 1913), which was savagely reviewed by J.M. Keynes, sought to expose the errors of the quantity theory of money, recently popularized by Irving Fisher: this shows the extent to which Hobson was still thinking as a classical economist brought up on Mill, failing to fully take account of the innovations of his contemporaries such as Marshall and Fisher (Backhouse, 1990).

Hobson was to supplement his account of underconsumption with a theory of distribution (Hobson, 1900) which drew heavily upon the Fabian theory of rent. This theory built on a marginal productivity theory of distribution that had first been published in 1891 in the *Quarterly Journal of Economics*, alongside John Bates Clark's article on the same subject. Hobson distinguished the costs of subsistence for any factor of production from its rent element, and argued that in principle surplus value might accrue to land, labour or capital. He further introduced the idea of ‘forced gains’ as an assertion of superior bargaining power in this process, with the result that ‘unearned income’ accrued to certain individuals and classes. He also assumed that the proportion of income which was in this sense economically functionless varied directly with the absolute level of income received. It followed that progressive taxation would not in practice impair any necessary incentive to production.

This analysis was later elaborated (Hobson, 1909b) to distinguish a ‘productive surplus’ that covered the costs of growth from an ‘unproductive surplus’, distributed according to no functional principle. Morally this was the property of the community which had created it. If redistributive taxation could restore it to its rightful possessors, over-saving by the rich would be curtailed and underconsumption by the poor rectified. This functional view of the proper working of the economic system, with effort matched to

reward by rooting out parasitism, reappears constantly as a paradigm in Hobson's writings. He dignified it with the name ‘the organic law’ and often suggested an evolutionary provenance for it. But he also claimed the authority of John Ruskin, of whom he wrote an admiring study (Hobson, 1898), for seeing consumption, not production, as the qualitative end of economic activity. He sought to unite these ideas in one of the most frequently reprinted of his books (Hobson, 1894) by adopting the formula: ‘From each according to his powers, to each according to his needs.’

Hobson's view, taken from Ruskin, that attention should be focused on the human cost of economic activity was the basis for Work and Wealth: A Human Valuation (Hobson, 1914), which offered a systematic response to Pigou's welfare economics, the first systematic exposition of which had been published two years earlier. As in his writings on underconsumption, and distribution, he adopted terminology that emphasized, and possibly exaggerated, his differences with orthodoxy. Resting on clear value judgements about the worth of different activities, such an approach fell out of favour in the 1930s, and even before that failed to dislodge the Cambridge approach, especially in Britain. However, his work was much better received in the United States, where he had significant personal connections and where some institutionalists considered him the leading representative of English welfare economics.

In the early 1890s, Hobson was inclined to believe that protection and economic imperialism could mitigate underconsumption. As his political radicalism intensified, however, he dismissed protection as a device for safeguarding the incomes of the wealthy, thereby aggravating the problem of over-saving. In the wake of the scramble for China and the outbreak of the South African War (1899–1902) Hobson also developed a novel theory of economic imperialism. He identified speculative investment in undeveloped territories as a cause of imperialism and claimed that it arose from over-saving by a parasitic class at home. In this sense underconsumption was the economic taproot of imperialism (Hobson, 1902). What he vigorously rejected was the proposition that there was sufficient profit to the country as a whole from trade and investment in Africa to counterbalance the costs of aggression. In contrast to Lenin, therefore, Hobson denied that imperialism was a structural necessity of the metropolitan economy. It could and should be checked at home by a policy of redistributive taxation, which would have the reciprocal effect of cutting the taproot (ending over-saving) and stimulating domestic demand (ending underconsumption).

The economic implication was that Britain could easily make up any loss on foreign trade by generating wealth at home – an argument that could be used by protectionists. Nonetheless, it was the Liberal and Labour Parties, with their commitment to free trade, to which Hobson looked for reformist amelioration. He was confident that imperialism could be beaten by democratic means precisely because it did not serve the interests of the majority but only of a privileged section of the nation. In his most famous book, therefore, Hobson devotes more than twice as much space to the politics than to the economics of imperialism (Hobson, 1902). He needed to do so because the puzzle was how a policy that was bad business for the nation as a whole had come to be adopted. The answer was that finance was the ‘governor’ of an engine whose motor power came from the forces of nationalism and social psychology that fuelled the politics of self-assertion (Hobson, 1901a). His analysis of imperialism changed over time and was often strongly coloured by passing political events. In at least one book (Hobson, 1911) he commended cosmopolitan finance as a force for peace and saw imperialism as a step on the road to world economic development. During the First World War, he made a partial return to his earlier views and between the wars his position was often an uneasy compromise between the stances adopted in 1902 and 1911. The fact that he chose to republish *Imperialism: A Study* in 1938 virtually unaltered obscured

the complexity of his response to empire (Cain, 2002).

It will be apparent that Hobson was no single-minded underconsumptionist. In the early 1900s his energies were directed towards permeating the Liberal Party with a broad-based conception of economics that would justify it in rejecting the classical nostrums of laissez-faire in favour of interventionist policies designed to further social justice (Hobson, 1909b). The publication of Hobson's *The Industrial System*, which consolidated much of his previous work, opportunely coincided with Lloyd George's People's Budget of 1909 and offered a defence of the policy of redistributive taxation via the concept of the surplus. This aspect overshadowed the restatement of Hobson's underconsumptionist position; though he now went further than before in analysing the dynamic process by which over-saving reduced all real incomes in the economy until automatic checks came into play (Hobson, 1909b, ch. 18). One might call this Hobson's most accomplished exercise in macroeconomics.

It was in the context of the depression after the First World War that Hobson once more returned to this theme (Hobson, 1922; 1930), and it was in this period that his economic views enjoyed greatest publicity. He was now loosely identified with the Labour Party and found a natural application for his ideas in mounting an economic case for a ‘living wage’ (Hobson, 1926b). His central contentions on over-saving continued to be refined (King, 1994) and, amid widespread unemployment, they found a more sympathetic response, even among professional economists who had previously accepted a full-employment assumption. In particular, by 1930 Hobson was on cordial terms with J.M. Keynes, who had in earlier years scorned his work. But Keynes was still anxious to keep his distance, as he made clear (Keynes, 1930, pp. 160–1). The reason was that when Keynes wrote of over-saving he meant under-investment; whereas for Hobson saving and investment were two names for the same thing, and by over-saving he had always meant under-spending. It followed also that Keynes had more interest in policies of public works as a means of promoting investment, whereas Hobson concentrated on the case for redistribution as a means of stimulating consumption. It was not until Keynes had virtually finished the General Theory that he fully realized that he had done Hobson and Mummery an injustice; and so he paid them a handsome, if belated, tribute (Keynes, 1936, pp. 364–71).

See Also

- Clark, John Bates
- Keynes, John Maynard
- Lenin, Vladimir Ilyich [Ulyanov]
- Pigou, Arthur Cecil
- underconsumptionism

Selected works

1889. (With A.F. Mummery.) *The Physiology of Industry*. London: Murray.

1894. *The Evolution of Modern Capitalism*. London: Walter Scott.

1896. *The Problem of the Unemployed*. London: Methuen.

1898. *John Ruskin, Social Reformer*. London: Nisbet.

1900. *The Economics of Distribution*. New York: Macmillan.

1901a. *The Psychology of Jingoism*. London: Nisbet.

1901b. *The Social Problem*. London: Nisbet.

1902. *Imperialism: A Study*. London: Nisbet.

1909a. *The Crisis of Liberalism*, ed. P.F. Clarke, Brighton: Harvester Press, 1974.

1909b. *The Industrial System*. London: Longman.

1911. *An Economic Interpretation of Investment*. London: Financial Review of Reviews.

1913. *Gold, Prices and Wages*. London: Methuen.

1914. *Work and Wealth*. London: Macmillan.

1919a. *Taxation in the New State*. London: Methuen.

1922. *The Economics of Unemployment*. London: Macmillan.

1926a. *Free Thought in the Social Sciences*. London: Allen & Unwin.

1926b. (With H.N. Brailsford, A. Creech Jones and E.F. Wise.) *The Living Wage*. London: Independent Labour Party.

1930. *Rationalisation and Unemployment*. London: Allen & Unwin.

1938. *Confessions of an Economic Heretic*. With an introduction by M. Freeden, Brighton: Harvester Press, 1976.

Bibliography

Backhouse, R.E. 1990. J.A. Hobson as a macroeconomic theorist. In Freeden (1990).

Cain, P.J. 2002. *Hobson and Imperialism: Radicalism, New Liberalism and Finance, 1887–1938*. Oxford: Oxford University Press.

Clarke, P.F. 1978. *Liberals and Social Democrats*. Cambridge: Cambridge University Press.

Freeden, M. 1990. *Reappraising J. A. Hobson*. London: Unwin Hyman.

Kadish, A. 1990. Rewriting the confessions: Hobson and the extension movement. In Freeden (1990).

Keynes, J.M. 1930. *A Treatise on Money*, vol. 1: *The Pure Theory of Money*. London: Macmillan for the Royal Economic Society, 1971.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan for the Royal Economic Society, 1973.

King, J.E. 1994. J. A. Hobson's macroeconomics: the last ten years. In Pheby (1994).

Pheby, J. 1994. *J. A. Hobson after Fifty Years*. Basingstoke: Macmillan.

How to cite this article

Clarke, Peter, Roger E. Backhouse and P.J. Cain. "Hobson, John Atkinson (1858–1940)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000075> doi:10.1057/9780230226203.0742

The New Palgrave Dictionary of Economics Online

hold-up problem

Yeon-Koo Che and József Sákovics
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Hold up arises when part of the return on an agent's relationship-specific investment is *ex post* expropriable by his trading partner. The hold-up problem has played an important role as a foundation of modern contract and organization theory, as the associated inefficiencies have justified many prominent organizational and contractual practices. We formally describe the main inefficiency hypothesis and sketch out the remedies suggested, as well as the more recent re-examination of the relevance of these theories.

Keywords

bargaining; Coase theorem; contract failure; fixed-price contracts; hold-up; incomplete contracts; ownership structures; relationship-specific investment; risk neutrality; underinvestment hypothesis; vertical integration

Article

Investments are often geared towards a particular trading relationship, in which case the returns on them within the relationship exceed those outside it. Once such an investment is sunk, the investor has to share the *gross* returns with her trading partner. This problem, known as hold-up, is inherent in many bilateral exchanges. For instance, workers and firms often invest in firm-specific assets prior to negotiating for wages. Manufacturers and suppliers often customize their equipment and production processes to the special needs of their partners, knowing well that future (re)negotiation will confer part of the benefit from customization to their partners. Clearly, the risk of the investor being held up discourages him or her from making socially desirable investments.

We first describe a simple model of hold up and illustrate the main *underinvestment* hypothesis (see Grout, 1984, and Tirole, 1986, for the first formal proof). A buyer and a seller, denoted B and S, can trade quantity $q \in [0, \bar{q}]$, where $\bar{q} > 0$. The transaction can benefit from the seller's (irreversible) investment. The investment decision is binary, $I \in \{0, 1\}$, with $I = 1$ meaning ‘invest’ and $I = 0$ meaning ‘not invest’. The investment *I* costs the seller $k \cdot I$, where $k > 0$. Given investment *I*, the buyer's gross surplus from

consuming q is $v_I(q)$ and the seller's cost of delivering q is $c_I(q)$, where both v_I and c_I are strictly increasing with $v_I(0) = c_I(0) = 0$. Let $\varphi_I = \max_{q \in Q}[v_I(q) - c_I(q)]$ denote the efficient social surplus given S's investment, and let q_I^* be the associated socially efficient level of trade. The net social surplus is then $W(I) := \varphi_I - kI$. Suppose that

$$\varphi_1 - k > \varphi_0, \tag{1}$$

so it is socially desirable for S to invest.

A crucial assumption is that S's investment decision, although observable to the parties, is not verifiable, and therefore it cannot be contracted upon. For the moment, assume as well that the nature of trade is sufficiently ‘inchoate’ so that the parties can contract on q only after S's investment decision has been made. We model the negotiation of this contract *à la* Nash, yielding the efficient trading decision q_I and splitting the gross surplus ϕ_I equally between the parties. The seller thus appropriates only a fraction (a half, in this case) of her investment return, while she bears the entire cost of investment, k, so her net payoff will be $U_S(I) := \frac{1}{2}\varphi_I - kI$, following her investment. Suppose

$$\frac{1}{2}\varphi_1 - k < \frac{1}{2}\varphi_0. \tag{2}$$

Then, even though the investment is socially desirable, S will not invest. Hence underinvestment arises.

Organizational remedies

One interpretation of the inefficiency is the failure of the Coase Theorem. The parties cannot achieve the efficient outcome since the non-contractibility of S's investment decision prevents them from meaningfully negotiating over that decision *ex ante*. From this perspective, the hold-up problem entails a transaction cost of market/bargaining mechanisms, and, as Coase (1937) suggested, the transaction cost may be avoided or reduced via other organizational structures. Indeed, Klein, Crawford and Alchian (1978) and Williamson (1979) suggested *vertical integration* as an organizational response.

Just how the hold-up problem disappears or at least diminishes through integration is not clear, however, and requires a theory of how a particular ownership structure affects the parties’ exposure to hold up. This is precisely what Grossman and Hart (1986) and Hart and Moore (1990), hence forth GHM accomplish (see also Hart, 1995, for an excellent synopsis). According to them, the ownership of an asset gives the owner the right to determine the use of the asset that is contractually not specifiable. The parties will still

negotiate the terms of trade (presumably to achieve an efficient outcome), but this *residual right* – and thus ownership – matters, since it determines the status quo payoffs of the parties in the negotiation.

To illustrate how the status quo payoffs may affect the incentives, consider our model above and suppose that either B or S can own all assets necessary for the vertical operations. The former type of integration is called B-*integration* and the latter type is called S-*integration*. Fix *i*-integration and fix S's investment decision $I \in \{0, 1\}$. If they fail to agree on the trade decision, party *i* can unilaterally realize the (status quo) payoff of $\psi_i^i(I)$ and party $j \neq i$ can realize the payoff of $\psi_j^i(I)$. It is reasonable to assume that, for $i = 1, 2$

Assumption: **GHM**: (i) $\psi_i^i(I) + \psi_j^i(I) \leq \varphi_I$, $I \in \{0, 1\}$; (ii) $\psi_S^i(1) - \psi_S^i(0) < \varphi_1 - \varphi_0$; (iii) $\psi_i^i(1) > \psi_i^i(0)$ and $\psi_j^i(1) = \psi_j^i(0)$.

Assumption GHM-(i) means that the status quo is welfare dominated by efficient trade; (ii) means that S's investment is specific to the relationship; and (iii) means that the investment improves the owner's status quo payoff but not the non-owner's.

Given the assumption that the parties split the surplus over and above the status quo payoffs, S's payoff will be

$$U_S^i(I) = \psi_S^i(I) + \frac{1}{2}(\varphi_I - \psi_B^i(I) - \psi_S^i(I)) - kI = \frac{1}{2}\varphi_I + \frac{1}{2}(\psi_S^i(I) - \psi_B^i(I)) - kI.$$

Hence, S's gain from investing under *i*-integration is

$$U_S^i(1) - U_S^i(0) = \frac{1}{2}(\varphi_1 - \varphi_0) + \frac{1}{2}\Delta^i - k, \tag{3}$$

where

$$\Delta^i := \psi_S^i(1) - \psi_S^i(0) - [\psi_B^i(1) - \psi_B^i(0)].$$

Given assumption GHM-(ii) and -(iii), $\varphi_1 - \varphi_0 > \Delta^S > 0 > \Delta^B$. Hence,

$$U_S^S(1) - U_S^S(0) > U_S(1) - U_S(0) > U_S^B(1) - U_S^B(0).$$

This shows that the S-*integration* is the optimal ownership structure, dominating symmetric (non-integrated) structure, which in turn dominates B-*integration* structure. In particular, if $U_S^S(1) - U_S^S(0) > 0 > U_S(1) - U_S(0)$, then the investment is sustainable if and only if the seller has the asset ownership. This result reveals the main tenet of GHM that asset ownership can serve to reduce the owner's exposure to hold up.

Remark 1. *The effects of alternative ownership structures may depend on the particular bargaining solution assumed. For example, the outside option bargaining or a Bertrand bidding solution may change the relative rankings of the alternative structures and may eliminate inefficiencies altogether. If the buyer's outside option is binding either from the buyer's owning more assets (that is, B-integration) or from the seller being subject to competition from another seller, then the seller is forced to make the buyer indifferent to that option, which causes the seller to internalize the social return of her investment. For this reason, B-integration may perform better than S-integration (Chiu, 1998*; *De Meza and Lockwood, 1998), or competition/non-integration may solve the hold-up problem (Bolton and Whinston, 1993*; *Che and Hausch, 1996*; *Cole, Mailath and Postlewaite, 2001*; *Felli and Roberts, 2001*; *MacLeod and Malcomson, 1993).*

Contractual solutions

In the above model, the trade decision is contractible only after the investment decision has been made. While this assumption resonates with many real business situations, it is difficult to reconcile with the fact that the parties can accurately calculate the payoff consequences of their behaviour (Maskin and Tirole, 1999). It is also crucial: if the parties *can* contract on *q* prior to the investment decision, the underinvestment problem may be solved, without requiring the organizational remedies discussed above. To illustrate, suppose the parties sign a contract requiring them to trade $\hat{q}$ for the total price of $\hat{t}$. Unless renegotiated, this contract will give S a payoff of $\hat{t} - c_I(\hat{q}) - kI$ if she chooses $I \in \{0, 1\}$. If $\hat{q} \neq q_I^*$, though, both parties will be better off by renegotiating to implement q_I^*. Given the assumption that this renegotiation splits the surplus equally, S's *ex ante* payoff will be

$$\hat{U}_S(I; \hat{q}) := \hat{t} - c_I(\hat{q}) + \frac{1}{2}[\varphi_I - (v_I(\hat{q}) - c_I(\hat{q}))] - kI.$$

Hence, her net benefit from investing under this contract is

$$\hat{U}_S(1; \hat{q}) - \hat{U}_S(0; \hat{q}) = \frac{1}{2}(\varphi_1 - \varphi_0) - \frac{1}{2}(v_1(\hat{q}) - v_0(\hat{q})) - \frac{1}{2}(c_1(\hat{q}) - c_0(\hat{q})) - k.$$

(4)

Whether a contract like this can create a sufficient incentive for S to invest depends on the nature of the investment made. Suppose first that the investment is *selfish*, so that it only decreases S's cost but does not affect B's valuation (that is, $v_1(\cdot) = v_0(\cdot)$). In this case, the trade contract can indeed protect S's incentive for investment. Observe that

$$c_0(q_1^*) - c_1(q_1^*) = v_1(q_1^*) - c_1(q_1^*) - [v_0(q_1^*) - c_0(q_1^*)] \geq \varphi_1 - \varphi_0.$$

By the same logic, $c_0(q_0^*) - c_1(q_0^*) \leq \varphi_1 - \varphi_0$. Since $c_i(\cdot)$ is continuous, there exists $\hat{q}^*$ between q_0^* and q_1^* such that $c_0(\hat{q}^*) - c_1(\hat{q}^*) = \varphi_1 - \varphi_0$. Consequently, $\hat{U}_S(1; \hat{q}^*) - \hat{U}_S(0; \hat{q}^*) = W(1) - W(0)$, so S will indeed invest whenever it is efficient to do so. Edlin and Reichelstein (1996) show that a fixed-price contract can provide efficient incentives for a selfish investment by either side and, with an additional condition, for selfish investments by both, in a more general environment with continuous investment. This result implies that, as long as the investments are selfish, the organizational remedies mentioned above will not be necessary.

Remark 2. *Aghion, Dewatripont and Rey (1994)and Chung (1991) have noted that efficiency can be achieved for investments by both sides via a contract that manipulates the status quo payoff of one party in the same way as above and gives the full bargaining power to the other party at the renegotiation stage, thus making that party a residual claimant of the social surplus in the marginal sense. The idea of contractual manipulation of bargaining powers also appears in Hart and Moore (1988) and Nöldeke and Schmidt (1995).*

Contract failure

Contracts may not restore efficiency if the investments are not selfish. Suppose the investment is *cooperative*: $c_1(\cdot) = c_0(\cdot)$. So, S's investment increases B's valuation only, worsening the former's bargaining position. Such a cooperative nature of investments underlies many instances of the hold-up problem (for example, quality-enhancing R&D investment by a supplier and customization efforts by partners). In this case, any commitment to trade exacerbates rather than alleviates the investor's vulnerability to hold up. Formally, given $c_1(\cdot) = c_0(\cdot)$, S's *ex ante* payoff will be

$$\hat{U}_S(1; \hat{q}) - \hat{U}_S(0; \hat{q}) = \frac{1}{2}(\varphi_1 - \varphi_0) - \frac{1}{2}(v_1(\hat{q}) - v_0(\hat{q})) - k \leq \frac{1}{2}(\varphi_1 - \varphi_0) - k = U_S(1) - U_S(0) < 0.$$

for any $\hat{q}$. In other words, no such trade contract creates more incentives for S than the null contract. In fact, Che and Hausch (1999) demonstrated that all feasible contracts are worthless if investments are cooperative.

A similar result can be obtained if the investment is selfish, but it is difficult to predict the ‘type’ of trade that will benefit from the investment (Hart and Moore, 1999; Segal, 1999). Specifically, suppose that there are *n* potential goods the parties may wish to trade but that only one of them becomes a ‘special’ type and *only* the special type will benefit from an investment. Assume that each of the *n* goods has an equal chance of becoming that special type *ex post*, so the parties can predict the special type only with probability 1/*n*. Adapted to our model, the surplus from trading the special type is ϕ_I given investment $I \in \{0, 1\}$, and the surplus from trading a ‘generic’ type is ϕ_0, regardless of the investment decision. Assume for simplicity that $q_I^* = 1$, for $I = 0, 1$. As the contract is renegotiable, under a contract requiring the parties to trade any good, S's *ex ante* payoff from choosing $I \in \{0, 1\}$, becomes

$$\tilde{U}_S(I) := \frac{1}{n}(\hat{t} - c_I(1)) + \frac{n-1}{n}(\hat{t} - c_0(1)) + \frac{1}{2}\left[\varphi_I - \frac{1}{n}\varphi_I - \frac{n-1}{n}\varphi_0\right] - kI.$$

In other words, S's investment influences her status quo payoff only when the good they contracted to trade turns out to be the special type, an event that arises with probability 1/*n*. This feature weakens the ability of a contract to provide incentives, as can be seen from S's gain from investing:

$$\tilde{U}_S(1) - \tilde{U}_S(0) = \frac{1}{n}(c_0(1) - c_1(1)) + \frac{1}{2}\left[\varphi_1 - \varphi_0 - \frac{1}{n}(\varphi_1 - \varphi_0)\right] - k = \frac{1}{2}\left(1 + \frac{1}{n}\right)(\varphi_1 - \varphi_0) - k.$$

Further, as the environment becomes ‘complex’ in the sense that $n \to \infty$, S's incentive reduces to that under the null contract, thus rendering contracts virtually worthless.

Several implications can be drawn from these two results. First, the contract failure result implies that the true challenge of the hold-up problem may lie with the nature of specific investments – either the ‘cooperative’ nature or the ‘unpredictability of investment benefit’. Second, the general failure of contracting to protect against hold up lends credence and relevance to the GHM analysis of the ownership structures or organizational theory in general based on the hold-up problem as a source of inefficiency. Third, for the above results it is crucial for the parties to be unable to commit not to renegotiate their contract. Were such commitment available, they could devise a contract that would induce them to reveal truthfully S's investment decision, say, by having both parties report simultaneously about the decision and penalizing both of them for any inconsistency via zero trade and zero transfer. Then, S can easily be

induced to invest by a sufficient amount of bonus given to her only conditional on both parties reporting ‘S has invested’. If a contract is renegotiable, such a costless revelation of information is impossible to achieve: Inconsistent reports do not reveal the identity of the liar, and both parties cannot be simultaneously punished, since they will renegotiate back to the Pareto frontier.

Remark 3. *Several elements are crucial for the contract failure result. First, it requires the existence of an opportunity to renegotiate following any contract-specified action. If there is some non-renegotiable action, then an efficient outcome may be achievable. Rogerson (1984) shows that liquidated damages achieve the efficient outcome if a contract can be breached non-renegotiably. Likewise, if in the last period of renegotiation the buyer can irrevocably determine the terms of trade, then buyer-option contracts can overcome the hold-up problem (see Lyon and Rasmusen, 2004). Contract failure re-emerges, however, in the case of cooperative investment if the parties discount delayed exercise of the option (Wickelgren, 2007). Second, risk neutrality is important for contract failure. If the parties were risk averse, then a lottery could be used to punish both parties even in the presence of renegotiation, and could achieve the first-best (Maskin and Tirole, 1999). Third, it is important for the contract to be bilateral. If a third party can be involved, efficiency can be achieved even when the contract is subject to renegotiation or collusion (Baliga and Sjostrom, 2005). Last, Watson (2006) gives a general treatment of how renegotiation opportunities arising at different stages interact with the technology of trade, and recognizes the relevance of modeling technological details of trade, i.e., whether the trade is individual or public*.

Dynamics

The basic hold-up model assumes that there is a single opportunity to invest, followed by the distribution of the surplus. Not too surprisingly, if the interaction is repeated, inefficiencies can be greatly reduced, in accordance with the Folk Theorem for repeated games (see, for example, Klein and Leffler, 1981). More surprisingly, allowing for dynamic investment patterns can have a dramatic effect even in a one-shot interaction, as shown by Che and Sákovics (2004a). When the agents can continue to invest even after the negotiation of the terms of trade has started, the anticipated investment dynamics can influence the way the parties negotiate and improve the incentives for investment.

To see how this works, modify our running example by allowing S to invest in the following period if she has not invested in the past and no agreement has been reached yet. If the parties discount their future very little, S's ‘invest’ can be sustained in a subgame-perfect equilibrium. In this equilibrium, hold up still arises on the equilibrium path in that S receives only the fraction of the gross surplus commensurate with his bargaining power. Yet this does not stop S from investing. Suppose S does not invest today but is expected to invest tomorrow in case no agreement is reached today. Then, there will be more surplus to divide tomorrow than there is today. Since the cost of tomorrow's investment will be borne solely by the investor, the prospect of the investor raising his investment tomorrow causes his partner to demand more to settle today. The investment dynamics thus results in a worse bargaining position for the party upon not investing, and creates a stronger incentive for investing than would be possible if such investment dynamics – that is, the option to invest in the future – were not allowed. As a result, investment can be supported in equilibrium.

In sum, dynamics in the trading relationship and/or investment technology lessens either the risk of hold up or the degree of inefficiencies caused by it. This questions the relevance of the hold-up problem as a rationale for organization and/or contractual remedies. At the same time, the presence of dynamics alters

the nature of the incentive problems and calls for different types of contractual or organizational prescriptions against hold up than those proposed based on the static models, as seen by Baker, Gibbons, and Murphy (2002), Che and Sákovics (2004b) and Halonen (2002).

See Also

- Coase theorem
- contract theory
- incomplete contracts
- procurement

Bibliography

Aghion, P., Dewatripont, M. and Rey, P. 1994. Renegotiation design with unverifiable information. *Econometrica* 62, 257–82.

Baker, G., Gibbons, R. and Murphy, K. 2002. Relational contracts and the theory of the firm. *Quarterly Journal of Economics* 117, 39–83.

Baliga, S. and Sjostrom, T. 2005. Contracting with third parties. Mimeo, Northwestern University.

Bolton, P. and Whinston, M. 1993. Incomplete contracts, vertical integration and supply assurance. *Review of Economic Studies* 60, 121–48.

Che, Y.-K. and Hausch, D. 1996. Cooperative investments and the value of contracting: Coase vs. Williamson. Mimeo., University of Wisconsin-Madison.

Che, Y.-K. and Hausch, D. 1999. Cooperative investments and the value of contracting. *American Economic Review* 89, 125–47.

Che, Y.-K. and Sákovics, J. 2004a. A dynamic theory of holdup. *Econometrica* 72, 1063–103.

Che, Y.-K. and Sákovics, J. 2004b. Contractual remedies to the hold up problem: a dynamic perspective. Social Science Research Institute Working Paper N. 2004-03, University of Wisconsin, and Economics Discussion Paper No. 100, University of Edinburgh.

Chiu, Y.S. 1998. Noncooperative bargaining, hostages, and optimal asset ownership. *American Economic Review* 88, 882–901.

Chung, T.-Y. 1991. Incomplete contracts, specific investment and risk sharing. *Review of Economic Studies* 58, 1031–42.

Coase, R. 1937. The nature of the firm. *Economica* 4, 386–405.

Cole, H., Mailath, G. and Postlewaite, A. 2001. Efficient non-contractible investments in large economies. *Journal of Economic Theory* 101, 333–73.

De Meza, D. and Lockwood, B. 1998. Does asset ownership always motivate managers? Outside options and the property rights theory of the firm. *Quarterly Journal of Economics* 113, 361–86.

Edlin, A. and Reichelstein, S. 1996. Holdups, standard breach remedies and optimal investment. *American Economic Review* 86, 478–501.

Felli, L. and Roberts, K. 2001. Does competition solve the hold-up problem? STICERD Theoretical Economics Discussion Paper No. TE/01/414, London School of Economics.

Grossman, S. and Hart, O. 1986. The costs and benefits of ownership: a theory of lateral and vertical integration. *Journal of Political Economy* 94, 691–719.

Grout, P. 1984. Investment and wages in the absence of binding contracts: a Nash bargaining approach. *Econometrica* 52, 449–60.

Halonen, M. 2002. Reputation and the allocation of ownership. *Economic Journal* 112, 539–58.

Hart, O. 1995. *Firms, Contracts, and Financial Structure*. Oxford: Clarendon Press.

Hart, O. and Moore, J. 1988. Incomplete contracts and renegotiation. *Econometrica* 56, 755–85.

Hart, O. and Moore, J. 1990. Property rights and the nature of the firm. *Journal of Political Economy* 98, 1119–58.

Hart, O. and Moore, J. 1999. Foundations of incomplete contracts. *Review of Economic Studies* 66, 115–38.

Klein, B., Crawford, R. and Alchian, A. 1978. Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics* 21, 297–326.

Klein, B. and Leffler, K. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89, 615–41.

Lyon, T. and Rasmusen, E. 2004. Buyer-option contracts restored: renegotiation, inefficient threats, and the holdup problem. *Journal of Law, Economics, and Organization* 20, 148–69.

Macaulay, S. 1963. Non-contractual relations in business: a preliminary study. *American Sociological*

Review 28, 55–70.

MacLeod, W. and Malcomson, J. 1993. Investments, holdup, and the form of market contracts. *American Economic Review* 83, 811–37.

Maskin, E. and Tirole, J. 1999. Unforeseen contingencies and incomplete contracts. *Review of Economic Studies* 66, 83–114.

Nöldeke, G. and Schmidt, K. 1995. Option contracts and renegotiation: a solution to the hold-up problem. *RAND Journal of Economics* 26, 163–79.

Rogerson, W. 1984. Efficient reliance and damage measure for breach of contract. *RAND Journal of Economics* 15, 39–53.

Segal, I. 1999. Complexity and renegotiation: a foundation for incomplete contracts. *Review of Economic Studies* 66, 57–82.

Tirole, J. 1986. Procurement and renegotiation. *Journal of Political Economy* 94, 235–59.

Watson, J. 2006. Contracts, mechanism design and technological detail. *Econometrica* (in press).

Wickelgren, A. 2007. The limitations of buyer-option contracts in solving the hold-up problem. *Journal of Law, Economics, and Organization* 23(1) (forthcoming).

Williamson, O. 1979. Transactions-cost economics: the governance of contractual relations. *Journal of Law and Economics* 22, 233–62.

How to cite this article

Che, Yeon-Koo and József Sákovics. "hold-up problem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000171> doi:10.1057/9780230226203.0743

The New Palgrave Dictionary of Economics Online

home production

Yongsung Chang and Andreas Hornstein

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Studying the incentives and constraints in the non-market sector – that is, home production – enhances our understanding of economic behaviour in the market. In particular, it helps us to understand (*a*) small variations of labour supply over the life cycle, (*b*) the low correlation between employment and wages over the business cycle, and (*c*) large income differences across countries.

Keywords

capital accumulation; Cobb–Douglas functions; elasticity of substitution; home production; hours worked; income differences; intertemporal substitution elasticity; IS–LM; labour supply; labour supply elasticities; leisure; Michigan Time Use Survey; national income accounts; neoclassical growth theory; non-market activity; real business cycles; stochastic growth model; technical progress; total factor productivity

Article

Studies such as the Michigan Time Use Survey (Hill, 1984; Juster and Stafford, 1991) indicate that a typical married couple allocates only about one-third of its discretionary time working for paid compensation in the market. The allocation of time for non-market activities, such as home production or leisure, may be as important for economic welfare as is the time spent working. Starting with Becker (1965) and Mincer (1962), the value of non-market activity has been explicitly incorporated into economic analysis in terms of forgone earnings. Since household decisions on the allocation of time to market and non-market activities are undertaken jointly, studying the incentives and constraints in the non-market sector – home production – enhances our understanding of economic behaviour in the market sector. We discuss three examples where the inclusion of home production has improved our understanding of macroeconomic issues: (*a*) low estimates of the labour supply elasticity from panel data; (*b*) low correlation between return to working and hours worked over the business cycle; and (*c*)

large differences in measured output across countries.

In a standard neoclassical growth model with home production, a household derives utility not only from the consumption of market goods but also from the consumption of non-market goods. Non-market goods are produced in a home production sector using work effort and capital. The household's utility also depends on the consumption of leisure, which is the household's time endowment minus work effort supplied to the market and the home production sector. One usually assumes that the economy's technology is such that investment goods that can be used to augment the capital stock in the market and non-market sectors are produced only in the market sector of the economy. Important factors in the determination of the dynamics of a neoclassical growth model with home production are the substitution elasticity between the consumption of market and non-market goods, the substitution elasticity between capital and labour in market and home production, the relative capital intensity of production in the market and the home production sectors, and the correlation of total factor productivity in the two sectors. Examples of the neoclassical growth model augmented with home production are Benhabib, Rogerson and Wright (1991) and Greenwood and Hercowitz (1991).

Business cycle analysis

The allocation of hours worked – employment – is at the heart of business cycle analysis. Table 1 shows the standard deviations and correlation of the cyclical components of total hours worked and returns to working for the US economy, 1964–2003.

Business cycle statistics of the US labour market, 1964–2003

σ_n / σ_w	$\sigma_n / \sigma_{y/n}$	$cor(n, w)$	$cor(n, y/n)$
1.51	1.72	.38	.01

Note: All variables are logged and de-trended with the use of the Hodrick–Prescott filter. Hours worked (n) represents the total hours employed in the non-agricultural business sector. Wages (w) are the real hourly earnings of production and non-supervisory workers. Labour productivity (y/n) is output divided by hours worked. The period covered is from 1964:I to 2003:II.

Sources: DRI-WEFA Basic Economics Database; Global Insight.

Two features are of great interest to macroeconomists. First, hours worked is substantially more volatile than the return to working. Second, hours worked is not highly correlated with the return to working. Employment in other countries also exhibits similar features (for example, Backus, Kehoe and Kydland, 1992). These facts present a serious challenge to modern business cycle theory that builds on the idea of intertemporal substitution of work effort. Intertemporal substitution assumes that people work relatively more hours in some years than in others because the return from working in the market is unusually high in those years (for example, Lucas and Rapping, 1969). According to Table 1, on the one hand it appears as if employment would have to be very elastic in its response to changes in the return to work, but on the other hand the returns to work appear to be only weakly correlated with the supply of work time.

Estimates of labour supply elasticity

Business cycle theory that builds on the stochastic growth model – for example, Kydland and Prescott (1982) – indeed requires a large intertemporal elasticity of substitution in order to account for the relatively large fluctuations of hours worked. Yet a substantial empirical literature based on micro data finds that households' willingness to substitute hours is quite low – less than 0.5 (for example, MaCurdy, 1981; Altonji, 1986). Home production provides a potential resolution of this problem.

Most micro estimates of the intertemporal substitution elasticity rely on the variation of hours worked and wages over the life cycle of households. Rupert, Rogerson and Wright (2000) show that these estimates may underestimate the true willingness to substitute hours across time if one does not take into account the fact that households simultaneously decide on the supply of hours for market and non-market activities. Essentially, conventional estimates of labour supply elasticities suffer from an omitted variable bias: home work is positively correlated with market work and should be included in the estimation. For simplicity we assume that households' preferences are log-linear in a consumption aggregator of market, c_{mt}, and home-produced consumption, c_{ht}, and work time, be it in the market, n_{mt}, or at home, n_{ht}:

$$u(c_{mt}, c_{ht}, n_{mt}, n_{ht}) = \log c(c_{mt}, c_{ht}) - B\frac{(n_{mt} + n_{ht})^{1+1/\gamma}}{1 + 1/\gamma}.$$

Then the optimal labour supply of a household that is t years old can be written as

$$\log w_t = (1/\gamma)\log(n_{mt} + n_{ht}) + A_t,$$

where w_t denotes the market wage rate, and A_t represents other terms that may depend on age. The parameter γ denotes the willingness to substitute total hours over time – intertemporal substitution elasticity. For conventional estimates of the labour supply elasticity, which ignore home production, time spent for home production activities represents an unobserved supply shifter for market labour.

A typical worker faces a hump-shaped wage profile in his life: wage rates rise, reach a peak at age 45–55, and decline from then on. It is not unreasonable to assume that the consumption of non-market goods, and therefore hours worked in home production, is correlated with the market wage profile over the life cycle. For example, high earning years tend to be around the years in which one buys a house or has children, both of which call for more time spent in home production. The fact that home work and market work are positively correlated over the life cycle, but home work is omitted from the estimation equation, implies that the estimated inverse labour supply elasticity $1/\hat{\gamma}$ will be biased upward.

Wage – employment correlations

One of the primary empirical patterns that have puzzled many business cycle theorists is the lack of a systematic relationship between employment and wages. On the one hand, Keynesian IS–LM models assume that real wages and hours worked lie on a stable, downward-sloped marginal product of labour schedule, and predict a strong negative correlation between real wages and hours worked (for example, Dunlop, 1938). On the other hand, real business-cycle models, such as that of Kydland and Prescott (1982), where productivity shocks shift the labour demand schedule along a relatively stable positively sloped market labour supply curve, tend to predict a strong positive correlation between wages and employment. Incorporating home production into the neoclassical growth model helps account for the low correlation between market work and wages as well as the large variation of employment. Technical progress not only augments the marginal product of labour in the market sector but also affects the marginal product of labour in the home production sector. Consider, for example, technical progress that is embodied in consumer durables, such as vacuum cleaners and washers. This kind of technological progress often reduces the required work effort in the home sector for household chores, and thereby shifts the supply curve of market work outward along a negatively sloped market demand for labour curve. Thus, while technical progress in the market sector causes a positive correlation between market hours and wages, technical progress in the non-market sector can cause a negative correlation between market hours and wages. If technical progress in the market is positively correlated with that in the non-market sector, then market hours may fluctuate substantially without any accompanying changes in real wages.

In general, the allocation of hours between the market and home depends on (*a*) the covariance structure of productivity in the market and home, (*b*) the substitution elasticity between market goods and home-produced goods, and (*c*) the substitution elasticity between capital and labour in the home production function – in particular, if the purchase of home capital (for example, a home theatre system) requires or saves hours in home production. Recently, rich structures between the market and home production have been introduced to study the various features of business cycles – for example, McGrattan, Rogerson and Wright (1997), Hornstein and Praschnik (1997), Fisher (1997), Einarsson and Marquis (1997), Ingram, Kocherlakota and Savin (1997), Perli (1998), Chang (2000), Gomme, Kydland and Rupert (2001).

Cross-country income differences

There are enormous income differences across countries, and such disparity has persisted over time. According to Heston, Summers and Aten (2002), the ratio of the average per capita GDP (based on purchasing power parity price) of the richest fifth of all countries to that of the poorest fifth of all countries was about 12 in 1960 and had doubled to almost 25 by 2000. In the standard neoclassical growth model, distortions to capital accumulation contribute to income differences. For a reasonably calibrated neoclassical growth model, the distortions that are required to account for the observed income differences are, however, unreasonably large. Parente, Rogerson and Wright (2000) show that the required distortions are substantially reduced once we distinguish between an economy's market sector whose output is measured in the national income accounts and a home-production sector whose output is not measured. With home production, distortions to capital accumulation not only reduce the capital stock but also can reallocate economic activity from the market sector to the non-market sector.

Moreover, the measured income differences overstate the true differences in welfare, and the unmeasured consumption from home production may explain how individuals in some countries can survive on the very low levels of reported income.

Consider the neoclassical growth model with log preferences in consumption, c_m, and leisure, l. Output, y_m, is produced using capital, k_m, and labour, n_m, as inputs to a constant returns to scale Cobb–Douglas production function, $y_m = k_m^{\alpha_m}(z_m n_m)^{1-\alpha_m}$. Output can be used for consumption and investment, x_m, to increase the capital stock: $k_{m,t+1} = (1-\delta)k_{mt} + x_{mt}/\pi$, where δ is the depreciation rate. With capital accumulation distortions, investment increases the capital stock less than one for one: $\pi \geq 1$ (for example, Parente and Prescott, 1994). It is easily conceivable that there are substantial inefficiencies in capital accumulation in less developed economies (for example, inefficient governments, ill-protected property rights). Given commonly assumed preferences and technology, the investment rate and work effort on the balanced growth path will be independent of the magnitude of capital distortions, but the capital stock and output will decline with the capital distortion. Two countries that look alike in terms of the investment rates may nevertheless have very different output levels. Conditional on a reasonable parameterization of the economy, we would, however, have to assume capital distortions, $\pi \geq 100$, in order to account for observed output differences of a factor of at least 10 (for example, Parente, Rogerson and Wright, 2000).

A straightforward extension of the neoclassical growth model that includes home production assumes that preferences are defined over a consumption aggregator that includes market consumption and non-market consumption, c_h, from the home-production sector. The home-production sector also uses capital, k_h, and work effort, n_h, as inputs to a Cobb–Douglas production function. The household's time endowment can now be used in the market and the non-market sectors, and market production can be used for investment in the market and the non-market sectors. If home production is less capital-intensive than market production, and market and non-market goods are sufficiently close substitutes, a higher capital distortion not only reduces total capital accumulation but also leads to a reallocation of the available capital and work effort from the market sector to the non-market sector. Parente, Rogerson and Wright (2000) argue that, for reasonable substitution elasticities between market and home-production consumption and capital shares in the home-production sector, capital distortions as low as $\pi = 15$ can account for income differences in the market sector of a factor of ten.

See Also

- business cycle measurement
- economic growth, empirical regularities in
- labour supply
- real business cycles
- time use

Any opinions expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

Bibliography

Altonji, J. 1986. Intertemporal substitution in labor supply: evidence from micro data. *Journal of Political Economy* 94, S176–S215.

Backus, D., Kehoe, P. and Kydland, F. 1992. International real business cycles. *Journal of Political Economy* 100, 745–75.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–517.

Benhabib, J., Rogerson, R. and Wright, R. 1991. Homework in macroeconomics: household production and aggregate fluctuations. *Journal of Political Economy* 99, 1166–87.

Chang, Y. 2000. Comovement, excess volatility, and home production. *Journal of Monetary Economics* 46, 385–96.

Dunlop, J. 1938. The movement of real and money wage rates. *Economic Journal* 48, 413–34.

Einarsson, T. and Marquis, M. 1997. Home production with endogenous growth. *Journal of Monetary Economics* 39, 551–69.

Fisher, J. 1997. Relative prices, complementarities and comovement among components of aggregate expenditures. *Journal of Monetary Economics* 39, 449–74.

Gomme, P., Kydland, F. and Rupert, P. 2001. Home production meets time-to-build. *Journal of Political Economy* 109, 1115–31.

Greenwood, J. and Hercowitz, Z. 1991. The allocation of capital and time over the business cycles. *Journal of Political Economy* 99, 1188–214.

Heston, A., Summers, R. and Aten, B. 2002. *Penn World Table Version 6.1*. Philadelphia: Center for International Comparisons, University of Pennsylvania.

Hill, M. 1984. Pattern of time use. In *Time, Goods and Well–Being*, ed. F. Juster and F. Stafford. Ann Arbor: University of Michigan Press.

Hornstein, A. and Praschnik, J. 1997. Intermediate inputs and sectoral comovement in the business cycle. *Journal of Monetary Economics* 40, 573–95.

Ingram, B., Kocherlakota, N. and Savin, N. 1997. Using theory for measurement: an analysis of the

cyclical behavior of home production. *Journal of Monetary Economics* 40, 435–56.

Juster, F. and Stafford, F. 1991. The allocation of time: empirical findings, behavior models, and problems of measurement. *Journal of Economic Literature* 29, 471–522.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lucas, R., Jr. and Rapping, L. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77, 721–54.

MaCurdy, T. 1981. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 88, 1059–85.

McGrattan, E., Rogerson, R. and Wright, R. 1997. An equilibrium model of the business cycle with household production and fiscal policy. *International Economic Review* 38, 267–90.

Mincer, J. 1962. On-the-job training: costs, returns and its implications. *Journal of Political Economy* 70, 50–79.

Parente, S. and Prescott, E. 1994. Barriers to technology adoption and development. *Journal of Political Economy* 102, 298–321.

Parente, S., Rogerson, R. and Wright, R. 2000. Homework in development economics: household production and the wealth of nations. *Journal of Political Economy* 108, 680–7.

Perli, R. 1998. Indeterminacy, home production, and the business cycle: a calibrated analysis. *Journal of Monetary Economics* 41, 105–25.

Rupert, P., Rogerson, R. and Wright, R. 2000. Homework in labor economics: household production and intertemporal substitution. *Journal of Monetary Economics* 46, 557–79.

How to cite this article

Chang, Yongsung and Andreas Hornstein. "home production." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000185> doi:10.1057/9780230226203.0744

The New Palgrave Dictionary of Economics Online

homogeneous and homothetic functions

J.-P. Crouzeix

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Homothetic orderings

Given a cone E in the Euclidean space $\mathbb{R}^n$ and an ordering $\preccurlyeq$ on E (i.e. a reflexive and transitive binary relation on E), the ordering is said to be homothetic if for all pairs $x, y, \in E$

$$x \preceq y \Rightarrow \lambda x \preceq \lambda y \text{ for all } \lambda > 0.$$

For each $x \in E$, denote by $L(x)$ the indifference surface

$$L(x) = \{y \in E: y \preceq x \text{ and } x \preceq y\}.$$

Hence, geometrically, if the ordering is homothetic, then for all $x \in E$ and $\lambda > 0$

$$L(\lambda x) = \{\lambda y: y \in L(x)\}.$$

Homothetic functions

Recall that a real function f on a set E defines a complete (or total) ordering on E via the relation

$$x \leq y \text{ if and only if } f(x) \leq f(y).$$

By definition, *f* is said to be homothetic if the ordering is homothetic (implying that the domain *E* of *f* is a cone). Thus utility functions which represent a homothetic ordering are homothetic.
Assume, now, that *f* is a homothetic and differentiable function on an open cone *E* of $\mathbb{R}^n$. Assume also that $\nabla f(x) \neq 0$ for all $x \in E$. Hence for all $\lambda > 0$ and all $x \in E$ there exists $k > 0$ such that

$$\frac{\partial f}{\partial x_i}(\lambda x) = k\frac{\partial f}{\partial x_i}(x), \text{ for } i = 1, 2, \ldots, n.$$

In economic terms, this property means that the marginal rate of substitution remains constant along any ray from the origin. In fact, under some suitable assumptions, this property characterizes homothety of functions.

Positively homogeneous functions

A real function *f* defined on a cone *E* of $\mathbb{R}^n$ is said to be positively homogeneous of order *p* if for all $x \in E$

$$f(\lambda x) = \lambda^p f(x) \text{ for all } \lambda > 0.$$

If *p*=1, the function is said to be positively homogeneous or linearly homogeneous. If *p*=0, then the definition becomes

$$f(\lambda x) = f(x) \text{ for all } \lambda > 0 \text{ and } x \in E.$$

Clearly, positively homogeneous functions of any order are homoethetic. Conversely, under some suitable assumptions on *E* and *f* (for instance *E* is the positive orthant in $\mathbb{R}^n$ and *f* is increasing on *E*) then, if *f* is homothetic there exist a positively homogeneous function *g* of order 1 on *E* and an increasing function *k* on $\mathbb{R}$ such that

$$f(x) = k[g(x)] \text{ for all } x \in E.$$

(This property is sometimes used as an alternative definition of homothety for functions.) As a consequence, under reasonable economic assumptions, a homothetic preference ordering can be represented by a linearly homogeneous utility function.
Production functions are often assumed to be positively homogeneous of order *p*. For example, the so-called Cobb–Douglas function

$$f(x_1, x_2, \ldots, x_n) = K x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n} x_i > 0,$$

where K, α_1, $\alpha_2, \ldots$, α_n are positive constants, is homogeneous of order $p = \alpha_1 + \alpha_2 + \ldots + \alpha_n$.
In consumer theory, demand functions are positively homogeneous of order zero in prices and wealth.

Positively homogeneous convex (or concave) functions

Since convexity is a fundamental concept in economics, special attention should be paid to positively homogeneous functions which are convex or concave.
Let E be a convex cone and f a real function on E. Then a necessary and sufficient condition for f to be convex (concave) and positively homogeneous of order 1 on E is that for all $x \in E$ and $\lambda \geq 0$

$$f(\lambda x) = \lambda f(x)$$

and for all pairs $x, y \in E$

$$f(x + y) \leq (\geq) f(x) + f(y).$$

The producer's cost function illustrates a concave positively homogeneous function: assuming that only one output is produced using *n* inputs, the cost function is given by

$$c(y, p) = \underset{x}{\text{Min}} [p^t x : F(x) \geq y]$$

where p_i, i=1, 2,…, n, is the unit price of input i and $F(x)$, the production function, is the maximal amount of output which can be produced with the input vector $x=(x_1, x_2,\ldots, x_n)$. Then, for a fixed price vector p, $c(y, p)$ is the minimal cost of producing y units of the output. For y fixed, $c(y, p)$ is concave and positively homogeneous of order 1 in p. Similarly, in consumer theory, if F now denotes the consumer's utility function, the $c(y, p)$ represents the minimal price for the consumer to obtain the utility level y when p is the vector of utility prices.

A fundamental property is as follows. Let f be a real continuous function on a closed convex cone of $\mathbb{R}^n$. Then f is convex and positively homogeneous of order 1 if and only if there exists a closed convex set S of $\mathbb{R}^n$ such that

$$f(x) = \mathrm{Sup}[y'x \,/\, y \in S].$$

This set S is unique and the function is called the support function of S (by symmetry, the same result holds when replacing convex by concave and Sup by Inf). Duality in consumer's (as well as in producer's) theory is based on this property.

We conclude with three examples of functions widely used in mathematics. A *semi-norm* on $\mathbb{R}^n$ is a convex positively homogeneous function f of order one on $\mathbb{R}^n$ such that $f(x)=f(-x)$ for all x (then $f(x) \geq 0$ for all x). A *norm* is a semi-norm for which x=0 whenever $f(x)$=0. Finally, given a convex set C which contains the origin, the *gauge* of C is the function f defined by

$$f(x) = \mathrm{Inf}\,[\lambda \geq 0 \,/\, x \in \lambda C].$$

A gauge function is convex and positively homogeneous of order one. Moreover, if the origin belongs to the interior of C and C is balanced (i.e. $x \in C$ implies that $x \in -C$), then the gauge is a norm.

Positively homogeneous quasi-concave (quasi-convex) functions

Let $\succsim$ be a preference ordering on a set E. In view of economic considerations, a common and reasonable assumption is the convexity of the ordering (i.e. for all $x \in E$, the set $\{y \in E/y \succsim x\}$ is convex). Then the utility functions which represent the ordering are quasi-concave but in general, a concave representation does not exist. However, in the case where the ordering is homothetic, it does. Indeed, a quasiconcave linearly homogeneous function which takes only positive (negative) values on the interior of its domain is concave [Newman] (by symmetry the same result holds for quasi-convex functions). It follows that a representable preference ordering which is homothetic and convex admits a

representation by a concave linearly homogeneous utility function.

See Also

- aggregate demand theory
- Cobb–Douglas functions
- Euler's theorem
- quasi-concavity
- separability

Bibliography

Barten, A.P. and Bohrn, V. 1981. Consumer theory. In *Handbook of Mathematical Economics*, Vol. 2, ed. K.J. Arrow and M.D. Intriligator. New York: North-Holland Publishing Company, 381–429.

Diewert, W.E. 1981. Duality approaches to microeconomic theory. In *Handbook of Mathematical Economics*, Vol. 2, ed. K.J. Arrow and M.D. Intriligator, New York: North-Holland Publishing Company, 353–99.

Green, J. and Heller, W.P. 1981. Mathematical analysis and convexity with application to economics. In *Handbook of Mathematical Economics*, Vol. 1, ed. K.J. Arrow and M.D. Intriligator, New York: North-Holland Publishing Company, 1–52.

Katzner, D.W. 1970. *Static Demand Theory*. New York: Macmillan.

Newman, P. 1969. Some properties of concave functions. *Journal of Economic Theory* 1, 291–314.

How to cite this article

Crouzeix, J.- P. "homogeneous and homothetic functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000081> doi:10.1057/9780230226203.0745

The New Palgrave Dictionary of Economics Online

horizontal and vertical equity

Jean-Yves Duclos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes the concepts of vertical and horizontal equity and provides some normative and positive justifications for them. It then outlines a few of the measures that have been proposed to assess whether government policies, and tax and transfer systems in particular, are vertically and horizontally equitable. It also points to useful references in the literature.

Keywords

concentration curve; discrimination; equality of opportunity; equality of resources; Hobbes, T.; horizontal equity; liability progression; Locke, J.; Lorenz curve; Nozick, R.; positioning; procedural equity; progressive and regressive taxation; Rawls, J.; redistribution of income; relative deprivation; residual progression; Sen, A.; social justice; utilitarianism; veil of ignorance; vertical equity; well-being

Article

Two broad principles govern the redistributive analysis of government policies. The first one, *vertical equity*, helps assess the distributive equity of a policy's impact on individuals with differing initial levels of welfare. The second, *horizontal equity*, serves to evaluate the policy's impact across individuals who are similar in all relevant ethical aspects – including their initial level of welfare.
In terms of taxation, the principle of vertical equity (VE) requires that the net fiscal burden increase with individuals' capacity to pay (measured by pre-tax income, say). A strong form of this principle is usually accepted: it postulates that the capacity to pay increases more rapidly than income, and that the net tax burden should thus also rise faster than income, and should therefore be *progressive*. It can be shown that the application of this principle serves to decrease relative inequality in income, net of the tax burden. The principle of horizontal equity, in turn, stipulates that similar individuals should receive a similar tax treatment from the government. Application of this second principle also controls for the emergence of vertical disparities among initially similar individuals. Though the two principles are

generally applied to the monetary dimension of the impact of government policies, they can also prove pertinent to the analysis of other dimensions thereof.

Vertical equity

Concern about inequality in resource allocation has a long history in moral and political philosophy, and features prominently in all major religions. It is mostly based on a belief in the fundamental dignity that is equally shared by all human beings as well as on a natural social aversion to material and human deprivation. VE in government policies is one of the tools most often advocated to bring about greater equality in resource allocation. VE in resource distribution has also long been considered a condition for social cohesion and stability. Two thousand and four hundred years ago, Plato indeed expressed the following concern about equality:

> We maintain that if a state is to avoid the greatest plague of all – I mean civil war, though civil disintegration would be a better term –extreme poverty and wealth must not be allowed to arise in any section of the citizen-body, because both lead to both these disasters. That is why the legislator must now announce the acceptable limits of wealth and poverty. The lower limit of poverty must be the value of the holding. The legislator will use the holding as his unit of measure and allow a man to possess twice, thrice, and up to four times its value. (*The Laws*, Book V, quoted in Cowell, 1995, pp. 21–2)

A utilitarian justification for a concern for VE is that surveys on the subjects of happiness and health suggest that the consumption of unnecessary goods essentially represents a consumption for ‘positioning’ vis-à-vis others. Such consumption improves the individual's position relative to others but, in and of itself, yields little or no increase in the individual's welfare and decreases others’ relative sense of well-being, causing anxiety, stress, and hostility. Individuals also appear to have difficulty dealing with feelings of relative deprivation and exclusion, which can be detrimental to the good functioning of markets and institutions. The purpose of VE is then to reduce inequality in the distribution of welfare so as to mitigate the effects of inequality's negative externalities.
An influential ethical foundation for the principle of VE has also appeared since the 1970s in the writings of a number of philosophers, the most well-known probably being John Rawls and Amartya Sen (for example, Rawls, 1971; Sen, 1985). Rawls in particular has argued that in the absence of preferences and socio-economic interests (that is, behind a *veil of ignorance*; see, for instance, Harsanyi, 1955), individuals would agree that social justice implies maximizing the set of opportunities and well-being of the least well-off group, namely, equalizing opportunities ‘upwards’ so that the greatest possible well-being be available to all.

Horizontal equity

As already mentioned, the principle of horizontal equity (HE) stipulates that ethically similar individuals must be treated similarly by the government. ‘Ethically similar’ also implies having a similar level of well-being, since as seen above it can be ethically justified for governments to distinguish between poor

and rich. Two initially similar individuals must therefore find themselves at approximately the same welfare level after the effect of a government policy has been accounted for, regardless of the individuals' initial preferences or socio-economic characteristics. This is the *classical* formulation of the principle of horizontal equity. An important corollary is that government interventions should not reverse the ranking of individuals in the distribution of welfare, unless it can be shown that the initial ranking was unjust – this is the alternative and popular *reranking* formulation of the HE principle.

The rationale for HE is primarily borne of a concern for procedural equity. Unlike for VE, it is not the result that is judged, but the process. For example, it can be argued that a reranking of two individuals by the government (in which one of the two receives assistance) can reduce the income distance (and vertical inequality) between the two, but this reranking must be considered horizontally inequitable if the initial ranking was not demonstrably unjust.

The HE principle is not only universally simple to appreciate, but it generally also garners more support from philosophers than the VE principle (though see Kaplow, 1989, for a critique). The most important ethical justification for HE is the avoidance of all forms of arbitrary discrimination in the government's treatment of citizens. Individuals of similar ethical worth should be treated and valued equally by the government. Notice that we are here dealing with individuals who are ethically similar, though not necessarily identical in all respects. Limiting the principle of HE to individuals who are identical in all points would strip it of virtually all practical relevance and would arguably leave governments too much latitude to practise arbitrary discrimination between individuals.

Drawing on the 17th-century social contract theories of Thomas Hobbes and John Locke, the foundations of this procedural justice were promoted *inter alia* by Nozick (1974), for whom the usual theories of justice place too much emphasis on outcomes in the redistribution of welfare, utility, or capacities. However, the bases for HE also follow from theories of vertical equity, since the unequal treatment of equals can only increase the distance between them. Robert Musgrave, an influential contributor to the development of the HE principle (see in particular Musgrave, 1959), summarizes this as follows:

> The requirement of HE remains essentially unchanged under the various formulations of distributive justice, ranging from Lockean entitlement over utilitarianism and fairness solutions. That of VE, on the contrary, undergoes drastic changes under the various approaches. While HE is met by the various VE outcomes, this does not mean that HE is derived from VE. If anything, it suggests that HE is a stronger primary rule. (Musgrave, 1990, p. 116)

There are also various utilitarian foundations for the principle of HE. Government policies that discriminate between ethically comparable individuals give rise to resentment and insecurity amongst them and can also lead to social and political unrest. Exclusion and discrimination can have an impact on both individual welfare and on feelings of social cohesion; this is particularly for policies that discriminate among those that are alike since individuals often specifically compare their treatment with that of others who enjoy a similar standard of living or characteristics.

There are two major sources of horizontal inequity (HI). The first is that the impact of public policy often varies purposefully with individual characteristics and preferences, and the second is that public

policy is typically non-deterministic by design and/or in application. Instances of HI occur in practice because of the difficulties faced by policies to account appropriately for household heterogeneity, and because of informational problems, administrative errors, incomplete take-up, tax evasion, randomness in the effect of programs and policies, and outright or implicit discriminatory behaviour by the government.

Measurement

Local measures of VE and progressivity

Let X and N represent respectively pre-tax income and post-tax incomes, and let $T(X)$ be taxes, with $N = X - T(X)$ – and suppose for a moment that the tax system is deterministic (or non-stochastic) and differentiable. Denote the average rate of taxation at pre-tax income X by $t(X) = T(X) / X$, and the derivative of $t(X)$ and $T(X)$ at $X = x$ by $t'(x)$ and $T'(x)$. A tax $T(X)$ is said to be

- locally progressive at $X = x$ if the average rate of taxation increases with X, that is, if $t'(x) > 0$;
- locally proportional at $X = x$ if the average rate of taxation stays constant with X, that is, if $t'(x) = 0$;
- and locally regressive at $X = x$ if the average rate of taxation decreases with X, that is, if $t'(x) < 0$.

The elasticity of taxes with respect to X, also called *liability progression*, is then given by:

$$LP(X) = \frac{X}{T(X)} T'(X) = \frac{T'(X)}{t(X)}. \tag{1}$$

$LP(X)$ is the *local* ratio of the marginal tax rate over the average tax rate at X. A second local measure of progression, $RP(X)$, called *residual progression*, is the elasticity of net income with respect to pre-tax income:

$$RP(X) = \frac{\partial (X - T(X))}{\partial X} \cdot \frac{X}{N} = \frac{1 - T'(X)}{1 - t(X)}. \tag{2}$$

A tax system is everywhere progressive if $RP(X) < 1$ everywhere.

Lorenz and concentration curves

For several reasons, we can expect the tax system to be stochastically linked to *X*, and can thus express taxes *T* as $T = T(X) + \nu$, where *T*(*X*) and ν are respectively a deterministic and stochastic tax determinant. The Lorenz curve $L_X(p)$ for *X* is the proportion of the total *X* that is held by those whose percentile in the distribution of *X* is *p* or lower. A frequent tool for measuring the VE of the tax *T* is the concentration curve, defined as $C_T(p) = \int_0^p \bar{T}(q)dq / \mu_T$, where $\bar{T}(q)$ is the expected tax paid by those at percentile *q* in the distribution of *X*, and where μ_T is the average of *T* in the entire population. $C_T(p)$ thus shows the proportion of total taxes paid by the *p* bottom proportion of the population. The concentration curve $C_N(p)$ for net incomes is analogously defined as the proportion of total *N* that is enjoyed by those whose percentile in the distribution of *X* is *p* or lower. Finally, let *t* be the average tax as a proportion of average pre-tax income: $t = \mu_T / \mu_X$. On the assumption of no reranking from the pre-tax to the post-tax distribution, the following conditions are then equivalent:

1. 1. $t'(X) > 0$ for all *X*;
2. 2. $LP(X) > 1$ for all *X*;
3. 3. $RP(X) < 1$ for all *X*;
4. 4. $L_X(p) > C_T(p)$ for all $p \in]0, 1[$ and for any distribution of pre-tax income;
5. 5. $L_N(p) > L_X(p)$ for all $p \in]0, 1[$ and for any distribution of pre-tax income.

Progressive taxation thus makes the distribution of *N* unambiguously more equal than the distribution of *X*, in the sense that it pushes up the Lorenz curve for incomes *whatever* the distribution of pre-tax incomes. *Tax progressivity* and *vertical equity* can in that sense be used interchangeably. Analogous results can be obtained for the more general case in which *T* can be negative (in the context of a *tax and benefit* system, say; see Duclos and Araar, 2006, for more details). In the presence of reranking (when $T'(X) > 1$ or when the tax system is stochastic), result 5 does not hold anymore.

Global measures of VE and progressivity

There are two major approaches to measuring *global* progressivity: the tax-redistribution (*TR*) approach, and the income-redistribution (*IR*) approach.

1. 1. A tax *T* is *TR*-progressive if $C_T(p) < L_X(p)$ for all $p \in]0, 1[$.
2. 2. A tax *T* is *IR*-progressive if $C_N(p) > L_X(p)$ for all $p \in]0, 1[$.

For two taxes, T_1 and T_2, if $LP_1(X) > LP_2(X)$ at all values of *X*, then the tax 1 is necessarily more *TR*-progressive than the tax 2; if $RP_1(X) < RP_2(X)$ at all values of *X*, then the tax 1 is necessarily more *IR*-progressive than the tax 2. In the absence of reranking, a more *IR*-progressive tax system is one which

decreases inequality by more and is therefore more vertically equitable.

Horizontal equity

The literature on the measurement of HE has evolved very significantly since around 1980. There have been two sub-periods, the first of which focused on the measurement of reranking using concentration and Lorenz curves and indices based thereon. One central result is that $C_N(p)$ will never be lower than the Lorenz curve $L_N(p)$, and will be strictly greater than $L_N(p)$ for at least one value of p if there is reranking in the redistribution of incomes. A tax T will thus cause reranking (and hence horizontal inequity) if and only if $C_N(p) > L_N(p)$ for at least one value of p. The difference between the Lorenz curve of post- and pre-tax incomes can then be expressed as:

$$L_N(p) - L_X(p) = \underbrace{C_N(p) - L_X(p)}_{\text{VE: progressivity}} - \underbrace{(C_N(p) - L_N(p))}_{\text{HI: Reranking}}. \tag{3}$$

This shows why a progressive tax system that causes reranking can push the Lorenz curve down and therefore increase inequality.
A recent promising approach to measuring classical HE has been to estimate the impact of the variability of taxes conditional on some initial value of pre-tax income. Capturing the impact of this variability can be done using many of the popular social welfare and inequality indices; see *inter alia* Aronson, Johnson, and Lambert (1994); Aronson and Lambert (1994); Lambert and Ramos (1997); Duclos and Lambert (2000); and Auerbach and Hassett (2002). This has typically led to total redistribution being expressible as the difference between VE and HI components.

Further reading

Classical texts on the concept and the measurement of VE and tax progressivity include Musgrave and Thin (1948); Slitor (1948); Blum and Kahen, Jr. (1963); Vickrey (1972); Fellman (1976); Jakobsson (1976); Kakwani (1977a; 1977b); Suits (1977); Reynolds and Smolensky (1977); Atkinson (1979); Plotnick (1981; 1982), King (1983); and Pfahler (1987). Recent literature surveys on the meaning and the measurement of HE can be found in Jenkins and Lambert (1999); Lambert (2001); and Duclos and Araar (2006).

See Also

- redistribution of income and wealth
- tax incidence

- taxation and poverty

Bibliography

Aronson, R., Johnson, P. and Lambert, P. 1994. Redistributive effect and unequal income tax treatment. *Economic Journal* 104, 262–70.

Aronson, R. and Lambert, P. 1994. Decomposing the Gini Coefficient to reveal the vertical, horizontal, and reranking effects of income taxation. *National Tax Journal* 47, 273–94.

Atkinson, A. 1979. Horizontal equity and the distribution of the tax burden. In *The Economics of Taxation*, ed. H. Aaron and M. Boskin. Washington, DC: Brookings Institution.

Auerbach, A. and Hassett, K. 2002. A new measure of horizontal equity. *American Economic Review* 92, 1116–25.

Blum, W. and Kahen, H., Jr. 1963. *The Uneasy Case for Progressive Taxation*. Chicago: University of Chicago Press.

Cowell, F. 1995. *Measuring Inequality*. London: Prentice Hall, Harvester Wheatsheaf.

Duclos, J.-Y. and Araar, A. 2006. *Poverty and Equity: Measurement, Policy and Estimation with DAD*. Boston: Springer.

Duclos, J.-Y. and Lambert, P. 2000. A normative approach to measuring classical horizontal inequity. *Canadian Journal of Economics* 33, 87–113.

Fellman, J. 1976. The effect of transformations on Lorenz curves. *Econometrica* 44, 823–4.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Jakobsson, U. 1976. On the measurement of the degree of progression. *Journal of Public Economics* 5, 161–68.

Jenkins, S. and Lambert, P. 1999. Horizontal inequity measurement: a basic reassessment. In *Handbook of Income Inequality Measurement. With a Foreword by Amartya Sen*, ed. J. Silber. Boston: Dordrecht and London: Kluwer.

Kakwani, N. 1977a. Applications of Lorenz curves in economic analysis. *Econometrica* 45, 719–28.

Kakwani, N. 1977b. Measurement of tax progressivity: an international comparison. *Economic Journal* 87, 71–80.

Kaplow, L. 1989. Horizontal equity: measures in search of a principle. *National Tax Journal* 42, 139–54.

King, M. 1983. An index of inequality: with applications to horizontal equity and social mobility. *Econometrica* 51, 99–116.

Lambert, P. 2001. *The Distribution and Redistribution of Income*, 3rd edn. Manchester and New York: Manchester University Press; distributed by Palgrave, New York.

Lambert, P. and Ramos, X. 1997. Horizontal inequity and vertical redistribution. *International Tax and Public Finance* 4, 25–37.

Musgrave, R. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

Musgrave, R. 1990. Horizontal equity, once more. *National Tax Journal* 43, 113–22.

Musgrave, R. and Thin, T. 1948. Income tax progression 1929–48. *Journal of Political Economy* 56, 498–514.

Nozick, R. 1974. *Anarchy, State and Utopia*. Oxford: Basil Blackwell.

Pfahler, W. 1987. Redistributive effects of tax progressivity: evaluating a general class of aggregate measures. *Public Finance/Finances Publiques* 42, 1–31.

Plotnick, R. 1981. A measure of horizontal inequity. *Review of Economics and Statistics* 62, 283–88.

Plotnick, R. 1982. The concept and measurement of horizontal inequity. *Journal of Public Economics* 17, 373–91.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Reynolds, M. and Smolensky, E. 1977. *Public Expenditure, Taxes and the Distribution of Income: The United States, 1950, 1961, 1970*. New York: Academic Press.

Sen, A. 1985. *Commodities and Capabilities*. Amsterdam: North-Holland.

Slitor, R. 1948. The measurement of progressivity and built-in flexibility. *Quarterly Journal of Economics* 62, 309–13.

Suits, D. 1977. Measurement of tax progressivity. *American Economic Review* 67, 747–52.

Vickrey, W. 1972. *Agenda for Progressive Taxation*, 1st edn. New York: Ronald Press.

How to cite this article

Duclos, Jean-Yves. "horizontal and vertical equity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000177> doi:10.1057/9780230226203.0746

The New Palgrave Dictionary of Economics Online

Hotelling, Harold (1895 – 1973)

Kenneth J. Arrow

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Harold Hotelling was devoted mainly to mathematical statistics but had a deep influence on economics. His famous 1929 paper on stability in competition introduced the notions of locational equilibrium in duopoly, with implications for political competition. His application of the calculus of variations to the allocation of a fixed stock over time formed the basis of subsequent work on the subject. In his 1938 presidential address to the Econometric Society he argued that marginal-cost pricing was necessary for Pareto optimality even for decreasing-cost industries, and showed that suitable line integrals were a generalization of consumers' and producers' surplus for many commodities.

Keywords

Accademia Nazionale dei Lincei; American Economic Association; Arrow, K. J.; calculus of variations; competition; confidence intervals; consumer surplus; depreciation; dummy variables; Dupuit, A.-J.-L.; Econometric Society; econometrics; exhaustible resources; game theory; Hotelling, H.; Institute of Mathematical Statistics; local equilibrium; marginal cost pricing; market socialism; mathematical economics; mathematical statistics; National Academy of Sciences; political competition; producer surplus; Royal Statistical Society; simultaneous equations models; statistics and economics; subgame perfection; welfare economics

Article

Harold Hotelling, a creative thinker in both mathematical statistics and economics, was born in Fulda, Minnesota, on 29 September 1895 and died in Chapel Hill, North Carolina, on 26 December 1973. His influence on the development of economic theory was deep, though it occupied a relatively small part of a highly productive scientific life devoted primarily to mathematical statistics; only ten of some 87 published papers were devoted to economics, but of these six are landmarks which continue to this day to lead to further developments. His major research, on mathematical statistics, had, further, a generally

stimulating effect on the use of statistical methods in different specific fields of application, including econometrics.

His early interests were in journalism; he received his BA in that field from the University of Washington in 1919. Later in classes, he would illustrate the use of dummy variables in regression analysis by a study (apparently never published) of the effect of the opinions of different Seattle newspapers on the outcome of elections and referenda. The mathematician and biographer of mathematicians, Eric T. Bell, discerned talent in Hotelling and encouraged him to switch his field. He received an MA in mathematics at Washington in 1921 and a PhD in the same field from Princeton in 1924; he worked under the topologist, Oswald Veblen (Thorstein Veblen's nephew), and two of his early papers dealt with manifolds of states of motion.

The year of completing his PhD, he joined the staff of the Food Research Institute at Stanford University with the title of Junior Associate. In 1925 he published his first three papers, one on manifolds, one on a derivation of the F-distribution, and one on the theory of depreciation. Here, apparently for the first time, he stated the now generally accepted definition of depreciation as the decrease in the discounted value of future returns. This paper was a turning-point both in capital theory proper and in the reorientation of accounting towards more economically meaningful magnitudes.

In subsequent years at Stanford he became Research Associate of the Food Research Institute and Associate Professor of Mathematics, teaching courses in mathematical statistics and probability (including an examination of Keynes's *Treatise on Probability*) along with others in differential geometry and topology. In 1927, he showed that trend projections of population were statistically inappropriate and introduced the estimation of differential equations subject to error; he returned to the statistical interpretation of trends in a notable joint paper (1929a) with Holbrook Working, largely under the inspiration of the needs of economic analysis.

The same year he published the famous paper on stability in competition (1929b), in which he introduced the notions of locational equilibrium in duopoly. This paper is still anthologized and familiar to every theoretical economist. As part of the paper, he noted that the model could be given a political interpretation, that competing parties will tend to have very similar programmes. Although it took a long time for subsequent models to arise, these few pages have become the source for a large and fruitful literature.

The paper was in fact a study in game theory. In the first stage of the game, the two players each chose a location on a line. In the second, they each chose a price. Hotelling sought what would now be called a subgame perfect equilibrium point. However, there was a subtle error in his analysis of the second stage, as first shown by d'Aspremont, Gabszewicz, and Thisse (1979). Hotelling indeed found a local equilibrium, but the payoff functions are not concave; if the locations are sufficiently close to each other, the Hotelling solution is not a global equilibrium. Unfortunately, this is the interesting case, since Hotelling concluded that the locations chosen in the first stage would be arbitrarily close in equilibrium. In fact, the optimal strategies must be mixed (Dasgupta and Maskin, 1986, pp. 30–32).

His paper on the economics of exhaustible resources (1931a) applied the calculus of variations to the problem of allocation of a fixed stock over time. All of the recent literature, inspired by the growing sense of scarcity (natural and artificial), is essentially based on Hotelling's paper. Interestingly enough, according to his later accounts, the *Economic Journal* rejected the paper because its mathematics was too difficult (although it had published Ramsey's papers earlier); it was finally published in the *Journal of Political Economy*.

In 1931, he was appointed Professor of Economics at Columbia University, where he was to remain until 1946. There he began the organization of a systematic curriculum in theoretical statistics, which eventually attained the dignity of a separate listing in the catalogue, though not the desired end of a department or degree-granting entity. Toward the end of the 1930s, he attracted a legendary set of students who represented the bulk of the next generation of theoretical statisticians. His care for and encouragement of his students were extraordinary: the encouragement of the self-doubtful, the quick recognition of talent, the tactfully made research suggestion at crucial moments created a rare human and scholarly community. He was as proud of his students as he was modest about his own work.

He also gave a course in mathematical economics. The general environment was not too fortunate. The predominant interests of the Columbia Department of Economics were actively anti-theoretical, to the point where no systematic course in neoclassical price theory was even offered, let alone prescribed for the general student. Nevertheless, several current leaders in economic theory had the benefit of his teaching. But his influence was spread more through his papers, particularly those (1932, 1935) on the full development of the second-order implications for optimization by firms and households (contemporaneous with Hicks and Allen) and above all by his classic presidential address (1938) before the Econometric Society on welfare economics. Here we have the first clear understanding of the basic propositions (Hotelling, as always, was meticulous in acknowledging earlier work back to Dupuit), as well as the introduction of extensions from the two-dimensional plane of the typical graphical presentation to the calculation of benefits with many related commodities. He argued that marginal-cost pricing was necessary for Pareto optimality even for decreasing-cost industries, used the concept of potential Pareto improvement, and showed that suitable line integrals were a generalization of consumers’ and producers’ surplus for many commodities. Here also we have the clearest expression in print of Hotelling's strong social interests which motivated his technical economics. His position was undogmatic but in general it was one of market socialism. He had no respect for acceptance of the *status quo* as such, and the legitimacy of altering property rights to benefit the deprived was axiomatic with him; but at the same time he was keenly aware of the limitations on resources and the importance in any human society of the avoidance of waste.

One of Hotelling's contributions which has had very extensive practical use is not contained in a paper. In 1947, the Director of the National Park Service asked a number of economists how to evaluate the benefits to visitors to national parks. Since the fee is small, the net benefit is undoubtedly considerable. Hotelling observed in a letter (Hotelling, 1947) that individuals incur considerable travel costs in coming to a park. Those individuals with the largest distance travelled can be assumed to receive zero net benefits, so that their gross benefits equal their travel costs. Nearer individuals receive a surplus that can easily be calculated.

Important as was his contribution to economics, most of his effort and his influence were felt in the field of mathematical statistics, particularly in the development of multivariate analysis. In a fundamental paper (1931b), he generalized Student's test to the simultaneous test of hypotheses about the means of many variables with a joint normal distribution. In the course of this paper, he gave a correct statement of what were later termed ‘confidence intervals’. In two subsequent papers (1933, 1936) he developed the analysis of many statistical variables into their principal components and developed a general approach to the analysis of relations between two sets of variates. The statistical methodologies of these papers and in particular the last contributed significantly to the later development of methods for estimating simultaneous equations in economics.

In 1946, he finally had the long-desired opportunity of creating a department of mathematical statistics, at the University of North Carolina, where he remained until retirement. He continued his active interest in economics there.

Space forbids more than the brief mention of his important work in the foundation of two learned societies, the Econometric Society and the Institute of Mathematical Statistics, both of which he served as President at a formative stage. He received many formal honours during his lifetime, including honorary degrees from Chicago and Rochester; he was the first Distinguished Fellow of the American Economic Association when that honour was created, as well as member of the National Academy of Sciences and the Accademia Nazionale dei Lincei, Honorary Fellow of the Royal Statistical Society and Fellow of the Royal Statistical Society.

Selected works

1925. A general mathematical theory of depreciation. *Journal of the American Statistical Association* 20, 340–53.

1927. Differential equations subject to error. *Journal of the American Statistical Association* 22, 283–314.

1929a. (With H. Working). Applications of the theory of error to the interpretation of trends. *Journal of the American Statistical Association* 24, 73–85.

1929b. Stability in competition. *Economic Journal* 39, 41–57.

1931a. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

1931b. The generalization of Student's ratio. *Annals of Mathematical Statistics* 21, 360–78.

1932. Edgeworth's taxation paradox and the nature of supply and demand functions. *Journal of Political Economy* 40, 577–616.

1933. Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology* 24, 417–41, 498–520.

1935. Demand functions with limited budgets. *Econometrica* 3, 66–78.

1936. Relation between two sets of variates. *Biometrika* 28, 321–77.

1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6, 242–69.

1947. Letter of 18 June to Newton B. Drury. Included in R.A. Prewitt, *The Economics of Public Recreation: An Economic Study of the Monetary Evaluation of Recreation in the National Parks*. Mimeo. Washington, DC: Land and Recreational Planning Division, National Park Service, 1949.

Bibliography

Dasgupta, P. and Maskin, E. 1986. The existence of equilibrium in discontinuous economic games, II: Applications. *Review of Economic Studies* 53, 27–41.

d'Aspremont, C., Gabszewicz, J.-J. and Thisse, J. 1979. On Hotelling's ‘Stability in Competition’. *Econometrica* 47, 1145–50.

How to cite this article

Arrow, Kenneth J. "Hotelling, Harold (1895–1973)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000089> doi:10.1057/9780230226203.0747

The New Palgrave Dictionary of Economics Online

hours worked (long-run trends)

Jeremy Greenwood and Guillaume Vandenbroucke
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

From 1830 to 2000 hours worked fell on two accounts: a drop in the market workweek and a decline in housework. The end result was that leisure rose. What caused this? The answer is technological progress. First, rising living standards implied that people could work less. Second, the introduction of new forms of leisure goods enhanced the value of time off. Third, time-saving household products reduced the need for housework. The time released allowed women to switch from home into market production. These points are illustrated with the use of historical evidence, economic theory, and numerical examples.

Keywords

Edgeworth–Pareto complements and substitutes; elasticity of substitution; hours worked; household production; housework; income effects; leisure; non-market goods; real wage rates; recreation; substitutes and complements; taxation of labour income; technological progress; wealth effect; women's work and wages

Article

Between 1830 and 2000, the average number of hours worked per worker declined, both in the marketplace and at home. Technological progress is the engine of such transformation. Three mechanisms are stressed:

- the rise in real wages and its corresponding wealth effect;
- the enhanced value of time off from work, due to the advent of time-using leisure goods; and
- the reduced need for housework, due to the introduction of time-saving appliances.

These mechanisms are incorporated into a model of household production. The notion of Edgeworth–Pareto complementarity/substitutability is key to the analysis. Numerical examples link theory and data.

Facts

Hours worked dropped precipitously over the course of the 19th and 20th centuries, both in the marketplace and at home. In 1830 the average workweek for an American worker in the marketplace was 70 hours. This had plunged to just 41 hours by 2002. At the same time there was a ninefold gain in real wages. Figure 1 shows the shrinkage of the market workweek and the leap forward in real wages. Likewise, the amount of time spent on housework dropped. A famous study of Middletown, Indiana, documented that in 1924 87 per cent of housewives spent more than four hours per day on housework (see Figure 2). None spent less than one hour. By 1999 only 14 per cent toiled more than four hours per day in the home, while 33 per cent spent less than one hour.

Figure 1

The fall in the US market workweek and the gain in real wages, 1830–2002. *Sources*: Average weekly hours data for 1830–80: Whaples (1990, Table 2.1). 1890–1970: *Historical Statistics of the United States: Colonial Times to 1970* (Series D765 and D803). 1970–2002: *Statistical Abstract of the United States*. Wage data: Williamson (1995, Table A1.1)

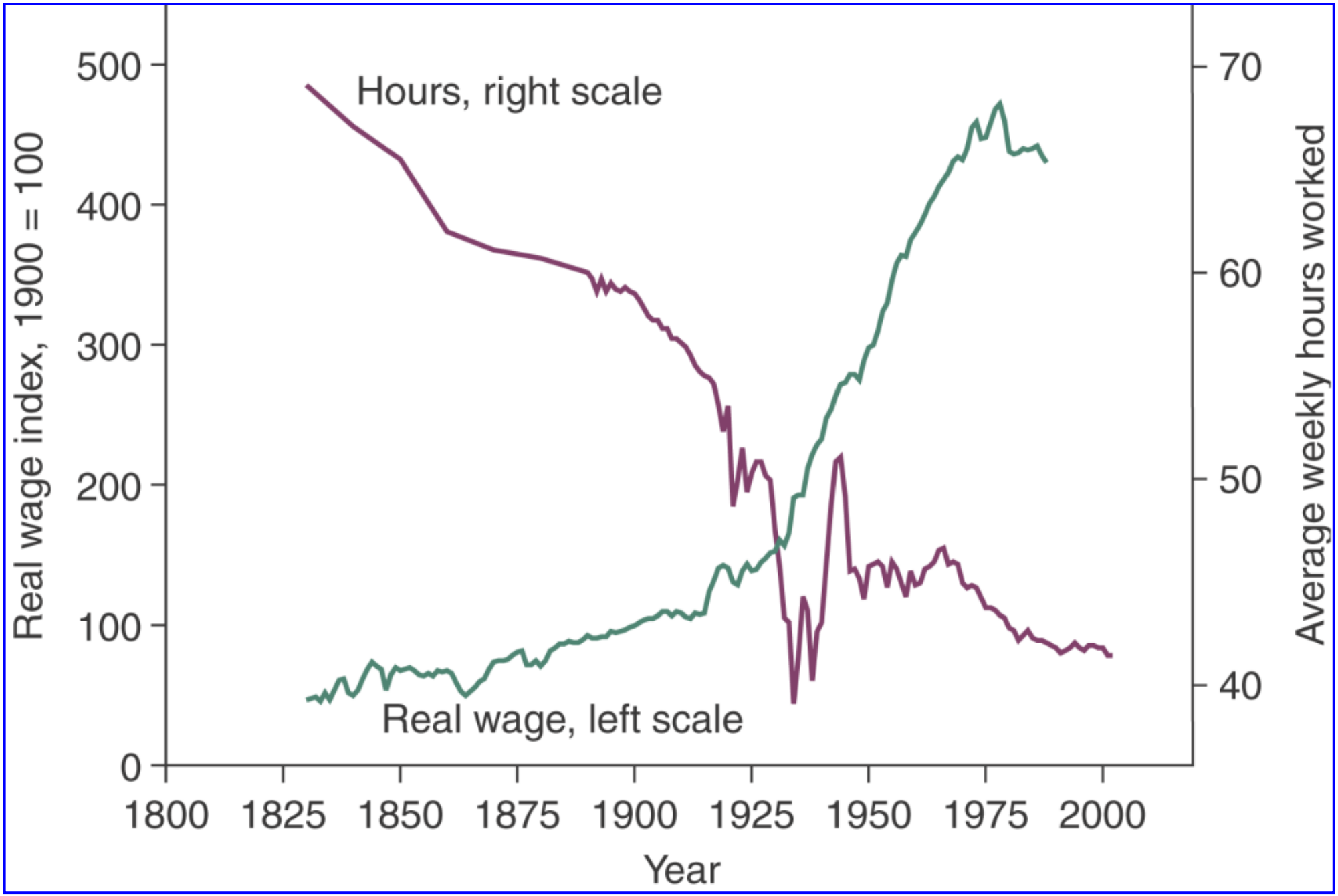

Figure 2
The ascent of US female labour-force participation and the reduction in housework, 20th century. *Sources*: Time spent on housework in Middletown: Caplow, Hicks and Wattenberg (2001, p. 37). Female labour-force participation: *Statistical Abstract of the United States*

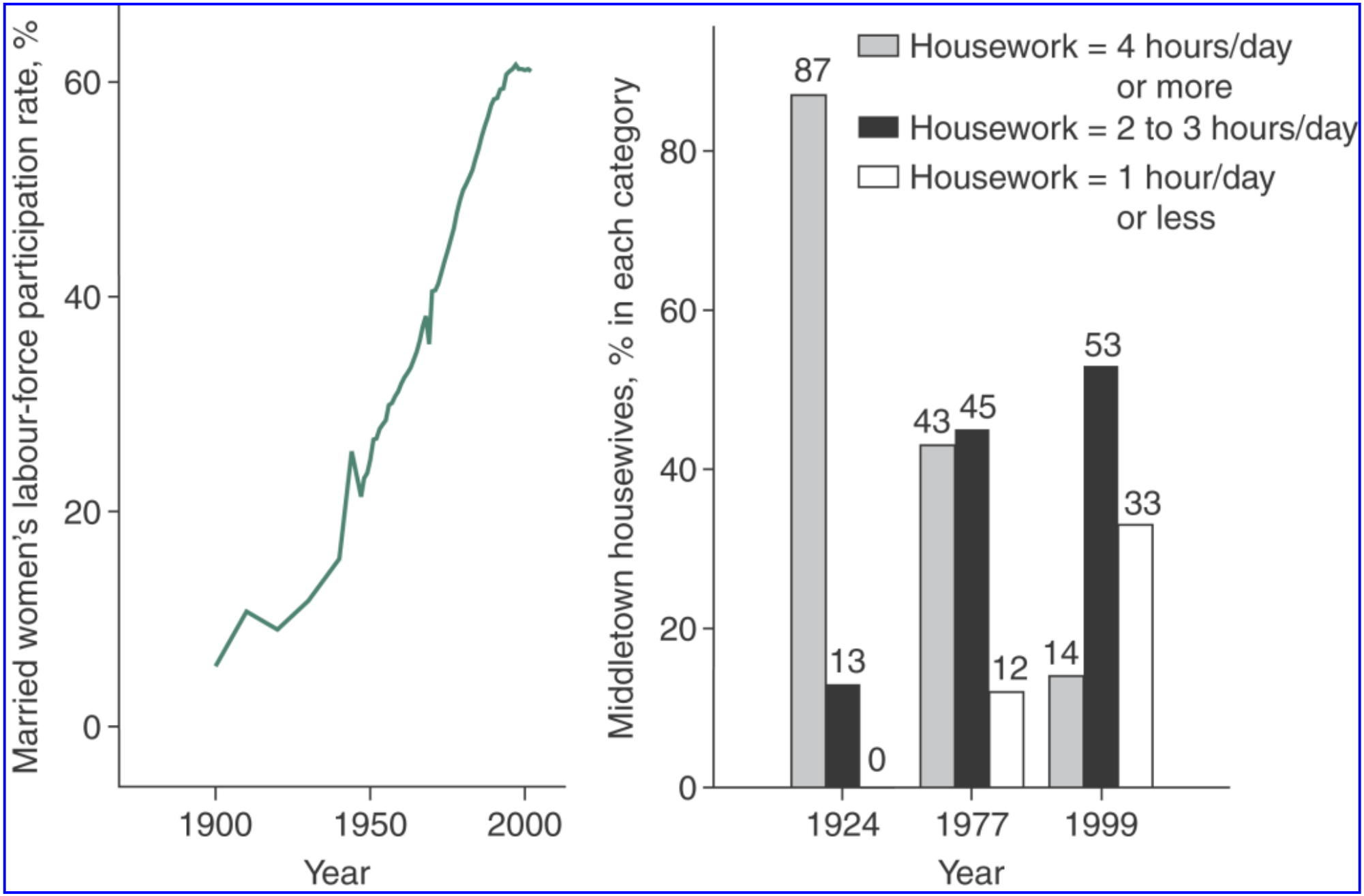

This decline in hours worked, both in the market and at home, was met by a rise in leisure. One implication of the increase in leisure is the uptrend in the share of personal consumption expenditure spent on recreation. This rose from three per cent in 1900 to 8.5 per cent in 2001, as Figure 3 illustrates. Additionally, the amount of time that a person needs to work in order to buy the goods used for leisure has fallen by at least 2.2 per cent a year – real wages grew at an annual rate of 1.65 per cent over the 1901–88 period. This price decline neglects the fact that many new forms of leisure goods have become available over time, or that old forms have improved. As the workweek – or the time spent on work both in the market and at home – dropped, more and more women entered the marketplace to work. This may seem a little paradoxical. Only four per cent of married women worked in 1890 as compared with 49 per cent in 1980 – again, see Figure 2.

Figure 3

The increase in recreation's share of expenditure and the decline in the time price of leisure in the US, 20th century. *Sources*: Recreation's share of expenditure for the years 1900–29: Lebergott (1996, Table A.1). 1929–2000: *Statistical Abstract of the United States*. Time price of leisure goods: Kopecky (2005)

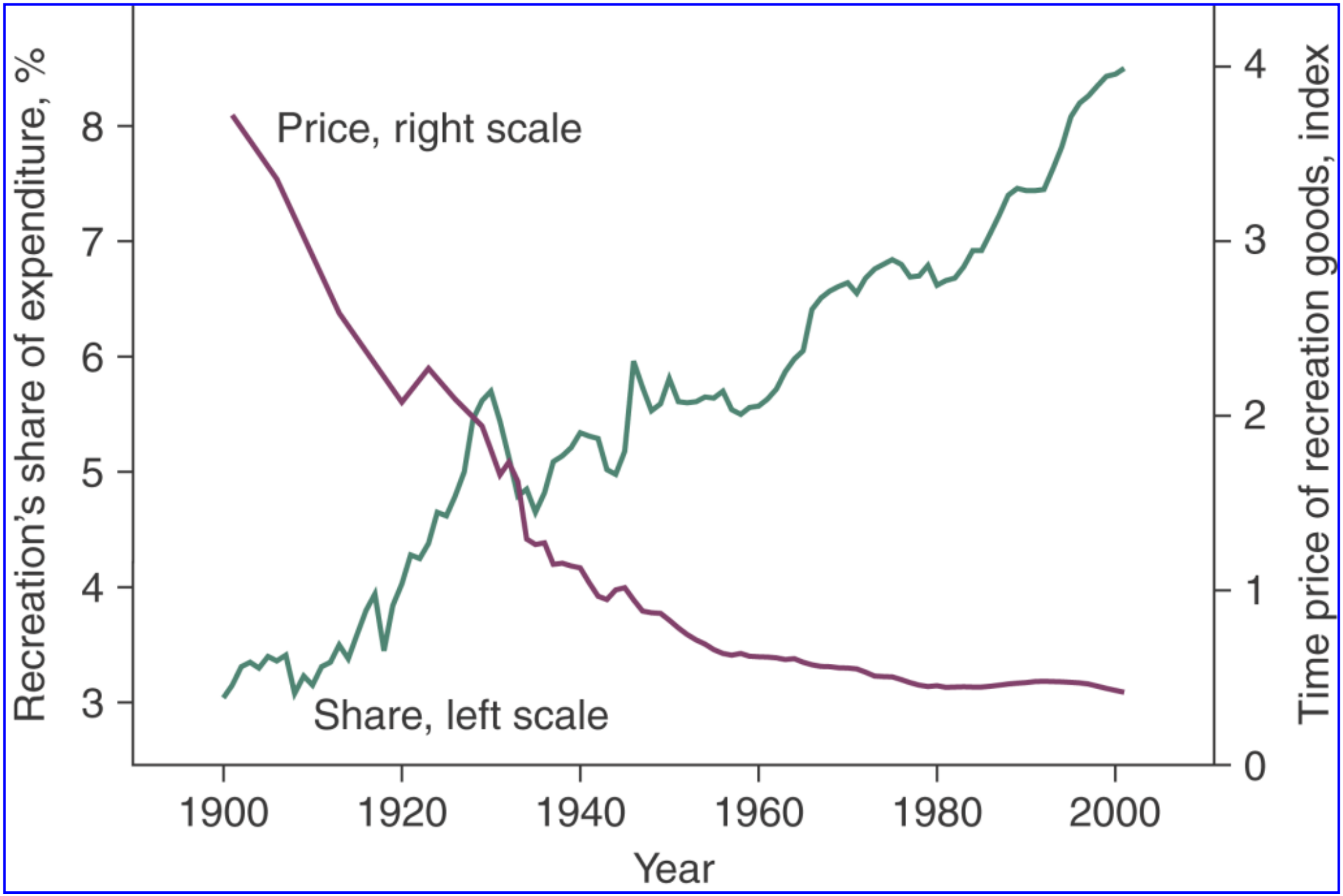

What can explain these facts? The answer is nothing mysterious: technological progress. Three channels of effect are stressed here. First, technological progress increases wages. On the one hand, an increase in real wages should motivate more work effort since the price of consumption goods in terms of forgone leisure has fallen. On the other hand, for a given level of work effort a rise in wages implies that individuals are wealthier. People may desire to use some of this increase in living standards to enjoy more leisure. Second, the value

of not working rises with the advent of new leisure goods. Leisure goods by their very nature are *time using*. Think about the impact of the following products: radio, 1919; Monopoly, 1934; television, 1947; videocassette recorder, 1979; Nintendo and Trivial Pursuit, 1984. Third, other types of new household goods reduce the need for housework. These household goods are *time saving*. Examples are: electric stove, 1900; iron, 1908; frozen food, 1930; clothes dryer, 1937; Tupperware, 1947; dishwasher, 1959; disposable diaper (Pampers), 1961; microwave oven, 1971; food processor, 1975. Some goods can be both time using and time saving, depending on the context: the telephone, 1876; IBM PC, 1984. A model is now developed to analyse the channels through which technological progress can affect hours worked in the market and time spent at home.

Analysis

Setup

Let tastes be represented by

$$U(c) + V(n), \text{ with } U_1,\ V_1 > 0 \text{ and } U_{11}, V_{11} < 0.$$

Here the utility functions U and V are taken to have the standard properties, while c and n represent the consumption of a market good and a non-market good. Now, suppose that the non-market good is produced in line with the constant-returns-to-scale production function

$$n = H(l, d) = dH\left(\frac{l}{d}, 1\right), \text{ with } H_1,\ H_2 > 0 \text{ and } H_{11},\ H_{22} < 0,$$

where H has standard properties, d represents purchased household inputs, and l is time spent in household production. The idea that non-market goods are produced by inputs of time and goods, just as market ones are, was introduced in classic work on household production theory by Becker (1965) and Reid (1934). Assume for simplicity that there is some indivisibility associated with d. The household must use the quantity $d = \delta$. (This assumption is innocuous. Greenwood, Seshadri and Yorukoglu, 2005, Section 6, and Vandenbroucke, 2005, illustrate how it can easily be relaxed.) This fixed quantity of the household input sells at price q, which is measured in terms of time. Last, an individual has one unit of time that he can divide between working in the market and using at home. The market wage rate is w.
Now, define the function

$$X(l, d) = V\left(dH\left(\frac{l}{d}, 1\right)\right).$$

Household time, l, and purchased household inputs, d, are Edgeworth–Pareto complements in utility when $X_{12} > 0$ and substitutes when $X_{12} < 0$ (cf. Pareto, 1906, eqs. (63) and (64)). When l and d are Edgeworth–Pareto complements in utility, an increase in d raises the marginal utility from l, or X_1, and likewise more l increases the marginal utility from d, or X_2.
The individual's optimization problem is

$$W(w, q) = \max_l \left\{U(w(1 - l) - qw) + X(l, \delta)\right\}.$$

The upshot of this maximization problem is summarized by the first- and second-order conditions written below.

$$wU_1(w(1-l)-qw) = X_1(l,\delta) = V_1\left(\delta H\left(\frac{l}{\delta},1\right)\right)H_1\left(\frac{l}{\delta},1\right), \tag{1}$$

and

$$\Sigma \equiv w^2U_{11} + X_{11} < 0.$$

The left-hand side of (1) represents the marginal cost of an extra unit of time spent at home. An extra unit of time spent at home results in a loss of wages in the amount *w*. This is worth $wU_1(w(1-l)-qw)$ in terms of forgone utility. The right-hand side gives the marginal benefit derived from spending an extra unit of time at home, $X_1(l, \delta)$. The solution for *l* is portrayed in Figure 4.

Figure 4

The determination of time spent at home, *l*

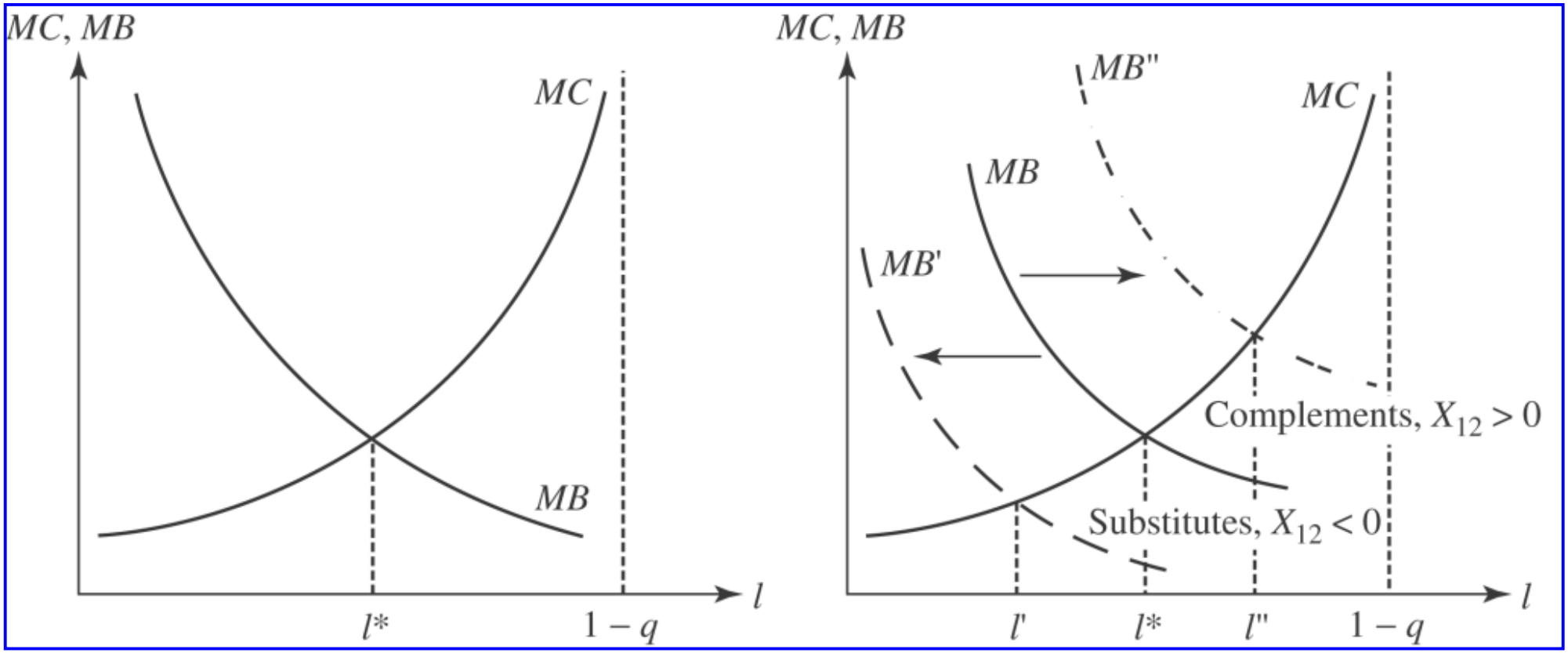

Effect of technological progress in household goods

Now, suppose that there is technological progress in household goods. In particular, let this be manifested by an increase in the amount of home inputs, δ, that can be purchased for q forgone units of time. How will this affect the amount of time spent at home? It is easy to calculate that

$$\frac{dl}{d\delta} = -\frac{X_{12}}{\Sigma} \gtreqless 0 \text{ as } X_{12} \gtreqless 0.$$

Therefore, time spent on household activities will rise or fall depending on whether time and goods are complements or substitutes in household utility. When time and purchased inputs are complements in utility, an extra unit of d raises the worth of staying at home. So, time spent at home should rise. Leisure goods, such as television, fall into this category. Such goods have contributed to the decline in work (either in the marketplace or at home) by both men and women. A detailed account of how this mechanism can contribute to the long-run drop in hours worked is provided by Vandenbroucke (2005). This case is shown in Figure 4 by a rightward shift in the marginal benefit curve from MB to MB'', causing time spent at home to rise from l^* to l''. The opposite is true when d and l are substitutes. This is portrayed in the figure by the leftward movement in the marginal benefit curve from MB to MB'. Time-saving household appliances, such as the microwave oven, are an example of this case. Such products have reduced the need for housework and have contributed to the increase in market work by women. Greenwood, Seshadri and Yorukoglu (2005) show how the increase in female labour-force participation can be explained along these lines. Therefore, technological advance in household products is consistent with the long-run decline in the market workweek (leisure goods) and the rise in female labour-force participation (time-saving appliances and goods).

When are two goods Edgeworth–Pareto complements or substitutes? From (1) the marginal benefit of time spent at home, $X_1(l, \delta)$, is the product of two terms, the marginal utility from non-market goods, $V_1(\delta H(l/\delta, 1))$, and the marginal product of household time, $H_1(l/\delta, 1)$. The marginal utility of housework is decreasing in δ, while the marginal product of household time is increasing in it. Thus, the net effect of an increase in δ will depend upon whether the former falls faster with an increase in δ than the latter rises. Specifically,

$$X_{12} = -V_{11}H_1^2(l/\delta) - V_1H_{11}l/\delta^2 + V_{11}HH_1,$$

so that

$$X_{12} \lesseqgtr 0 \text{ as } \frac{-(l/\delta)H_{11}}{H_1} \lesseqgtr \frac{-nV_{11}}{V_1}\frac{\delta(H - H_1l/\delta)}{n}.$$

In other words, whether or not $X_{12} \lesseqgtr 0$ depends on whether the elasticity of the marginal product of labour with respect to the time–goods ratio, $-(l/\delta)H_{11}/H_1$, is smaller or larger than the elasticity of marginal utility with respect to the home good, $-nV_{11}/V_1$, weighted by the share of purchased inputs in output, $\delta(H - H_1l/\delta)/n$. Thus, l and δ are likely to be substitutes in utility when: (*a*) the responsiveness of the marginal product of l/δ is small with respect to a change in δ; (*b*) the marginal utility of home goods declines quickly with more consumption; (*c*) when purchased inputs are important in production.

Example 1: **(The impact of leisure goods on hours worked)** Let $U(c) = \varphi \ln(c)$ and $V(n) = (1-\varphi)\ln(n)$. Represent the household technology by the constant-elasticity-of-substitution production function $H(l, \delta) = (\delta^\rho + l^\rho)^{1/\rho}$. The household's budget constraint is $c = w(1 - l - q)$. Given this set-up, the first-order condition (1) can be rewritten as

$$\frac{\varphi}{1-\varphi} = \frac{1-l-q}{\delta^\rho + l^\rho}l^{\rho-1}. \quad (2)$$

Observe that a change in wages, *w*, does not affect hours worked in the market, $1 - l$. The length of the workweek in the 1890s was about 42 per cent above that of the 1990s. In 1995 the typical worker spent about one-third of his available time working in the market. So, set $1 - l_{1995} = 1/3$ and $1 - l_{1895} = 1.42 \times 1/3$. Let $\delta_{1895} = 0.1$. The share of leisure goods in expenditures, *s*, is given by $s = q/(1-l)$. Costa (1997) reports that this share was two per cent in the 1890s and six per cent in the 1990s. Thus, the time-price *q* is given by $q_t = (1 - l_t)s_t$, for $t = 1895$ and 1995. Finally, pick $\rho = -0.6$, which implies an elasticity of substitution between leisure time and leisure goods of 0.63. Proceed now in two steps. First, use (2) to back out the value of ϕ that is consistent with $l = l_{1895}$, $q = q_{1895}$, and $\delta = \delta_{1895}$. This results in $\varphi = 0.19$. Second, use this equation to find the value of δ_{1995} that is in agreement with $l = l_{1995}$, $q = q_{1995}$, and $\varphi = 0.19$. This leads to $\delta_{1995} = 0.69$. Voilà, an example has now been constructed where the change in market hours matches exactly the corresponding figure in the US data. Additionally, the share of expenditure spent on leisure is in line with the data. In physical units, households in 1995 had 6.90 times more leisure goods than did households in 1895. This number depends upon the elasticity of substitution between leisure time and leisure goods. The higher the degree of complementarity (or the smaller is ρ), the less is the required increase in δ.

Remark: An example can be constructed in very similar fashion to show that labour-saving household inputs (or the case of Edgeworth–Pareto substitutes) can account for the rise in female labour-force participation. The interested reader is referred to Greenwood and Seshadri (2005, Example 5, p. 1256).

Effect of an increase in wages

How will rising wages impact hours worked? It's easy to calculate that

$$\frac{dl}{dw} = \frac{U_1 + w(1-l-q)U_{11}}{\Sigma} \gtrless 0 \text{ as } U_1 \lessgtr -w(1-l-q)U_{11}.$$

On the one hand, a boost in wages increases the opportunity cost of staying at home. This should reduce the time spent at home, *l*, and is represented by the substitution effect term, $U_1 / \Sigma < 0$. On the other hand, higher wages make the individual wealthier. The individual should use some of this extra wealth to increase his time spent at home. This income effect is shown by the term, $w(1-l-q)U_{11} / \Sigma > 0$. Thus, time spent at home can rise or fall with wages depending on whether the income effect dominates the substitution effect. In general, then, anything can happen, as the following two specialized cases for *U* make clear.

1. 1. Let $U(c) = \ln(c)$, the macroeconomist's favourite utility function. Here, $U_1 = 1/c$ and $w(1-l-q)U_{11} = -1/c$. Therefore, the substitution and income effects from a change in wages exactly cancel each other out. Long-run changes in wages have no impact on hours worked in the market, $1-l$.
2. 2. Suppose $U(c) = \ln(c-\epsilon)$, where $\epsilon > 0$ is some subsistence level of consumption. Now, $U_1 = 1/(c-\epsilon)$ and $w(1-l-q)U_{11} = -c/(c-\epsilon)^2$. Therefore, $dl/dw = -\epsilon/[(c-\epsilon)^2\Sigma] > 0$. Consequently, rising wages lead to a fall in market hours, $1-l$. The intuition is simple. At low levels of wages an individual must work hard to meet his subsistence level of consumption, ϵ. Achieving the subsistence level of consumption becomes easier as wages rise and this allows the individual to ease up on his work effort. Thus, this form for the utility function is in accord with a long-run decline in hours worked. Additionally, it is consistent with the observation reported in Vandenbroucke (2005) that unskilled workers laboured longer hours in 1900 than did skilled ones, while today they work about the same.

Can an increase in wages explain the decline in the workweek? The answer is ‘yes’, as the following example makes clear.
Example 2: **(The impact of rising wages on hours worked)** Let $U(c) = \ln(c-\epsilon)$ and $V(n) = \alpha n$. Represent the household technology by $H(l, d) = l$. Equation (1) appears as

$$1 - l = \frac{1}{\alpha} + \frac{\epsilon}{w}, \tag{3}$$

which gives a very simple solution for hours worked, $1-l$. Let the time period for this example be 1830 to 1990. The real wage rate in 1990 (actually in 1988) was 9.15 times the wage rate of 1830 (Williamson, 1995). So, set $w_{1830} = 1$ and $w_{1990} = 9.15$. Following the discussion in Example 1, fix hours worked in 1830 and 1990, or $1 - l_{1830}$ and $1 - l_{1990}$, using the equations $1 - l_{1830} = 1.65 \times 1/3$ and $1 - l_{1990} = 1/3$. Employing these restrictions in conjunction with (3) leads to a system of two equations in the two unknown parameters α and ϵ. Specifically, one obtains

$$1 - l_{1830} = \frac{1}{\alpha} + \frac{\epsilon}{w_{1830}},$$

and

$$1 - l_{1990} = \frac{1}{\alpha} + \frac{\epsilon}{w_{1990}}.$$

Solving yields $\alpha = 3.26$ and $c = 0.24$. The subsistence level of consumption, c, amounts to 44 per cent of consumption in 1830, and eight per cent in 1990.
The 20th century saw the advent of labour income taxation. So perhaps the previous example should have focused on the rise of after-tax wages. This is easy to amend.
Example 3: **(The effect of higher labour income taxation on hours worked)** Take the setup from Example 3 with one modification, to wit the introduction of labour income taxation. In particular, suppose that wages are taxed at rate τ. A fraction θ of the revenue the government receives is rebated back to the worker via lump-sum transfer payments, *t*. The rest goes into worthless government spending on goods and services, *g* – or equivalently one could assume that it enters into the consumer's utility function in a separable manner. Hence, the worker's budget constraint reads $c = (1-\tau)w(1-l) + t$, while the government's appears as $g + t = \tau w(1-l)$. The first-order condition for this setting is

$$\frac{(1-\tau)w}{c-c} = \alpha.$$

Combining the worker's and government's budget constraints yields $c = [1-\tau(1-\theta)]w(1-l)$. Using this fact in the above first-order condition results in

$$1 - l = \frac{1-\tau}{\alpha[1-\tau(1-\theta)]} + \frac{c}{w[1-\tau(1-\theta)]}. \quad (4)$$

Observe that when $c = 0$ and $\theta = 0$ (no rebate) an increase in the tax rate will have no impact on hours worked, because the substitution and income effects exactly cancel each other out. When $c = 0$ and $\theta = 1$ (full rebate) higher taxes will dissuade hours worked since only the substitution effect is operational. Alternatively, if $c > 0$ and $\theta = 0$ (no rebate), then it transpires that a rise in taxes will cause hours worked to move up. Here the negative income effect from the increase in government spending, which will result in more hours being worked, outweighs the substitution effect. Therefore, in general the effect of labour income taxation on hours worked is ambiguous. The result will depend on how the government uses the revenue it raises, and the functional forms and parameter values used for tastes and technology.
Take labour income taxes to be zero in 1830. Assume a rate of 30 per cent in 1990, in line with numbers reported by Mulligan (2002). Fix $\theta = 0.33$, its value for 1990 as measured by the National Income Product Accounts. By following the procedure in Example 3, it can be deduced that the observed fall in hours worked is occurs when $\alpha = 2.86$ and $c = 0.20$. Furthermore, it can be inferred that the rise in wages accounts for 93 per cent of the fall, while the increase in taxes explains the remaining seven per cent. (For those interested, the decomposition is done as follows: Represent the right-hand side of (4) by $L(w, \tau)$. Then,

$$(1-l') - (1-l) = [L(w', \tau') - L(w, \tau') + L(w', \tau) - L(w, \tau)] / 2 + [L(w', \tau') - L(w', \tau) + L(w, \tau') - L(w, \tau)] / 2.$$

The first term in brackets is a measure of the change in hours worked, $(1-l') - (1-l)$, due to the shift in wages from *w* to w', while the second term gives the change due to a movement in taxes from τ to τ'.)
All of the above examples are intended solely as illustrations of some secular forces that potentially influence hours worked. A quantitative assessment of the impact that taxes have on hours worked will depend upon the particulars of the model used. A serious study is conducted in Prescott (2004).
The real world seems to have experienced two conflicting trends: a decline in market work and a rise in female-labour participation. A more general model could be consistent with both of these facts. To see this, imagine a framework with two types of labour, male and female. There is a division of labour in the home. Men work primarily in the market. Females do housework and, time permitting, market work. Households purchase both time-saving and time-using household inputs. Female labour-force participation would rise as labour-

saving goods economize on the amount of housework that has to be done. Simultaneously, the market workweek would decline, due either to the introduction of leisure goods or to an income effect associated with a rise in wages. The value of leisure would rise for both men and women. Interestingly, Aguiar and Hurst (2006) document a dramatic increase in leisure for both men and women over the period 1965–2003. They construct various measures of leisure. They all showed a gain over the period under study. The narrowest definition rose by 6.4 hours a week for men and 3.8 hours for women, after adjustment for demographic changes in the population. This measure included time spent on activities such as entertainment, recreation, and relaxing. The authors' preferred measure increased by 7.9 hours a week for men and 6.0 hours for women. This broader definition also included activities such as eating, sleeping, personal care, and childcare. Another manifestation of the rise in the value of leisure is the increase in the fraction of life spent retired. Kopecky (2005) relays that a 20-year-old man in 1850 could expect to spend about six per cent of his life retired, while one in 1990 should enjoy about 30 per cent of his life in retirement. She shows how the trend towards enjoying more retirement can be analysed in much the same way as the decline in the workweek.

See Also

- household production and public goods
- labour supply
- leisure
- technical change
- time use

Bibliography

Aguiar, M. and Hurst, E. 2006. Measuring trends in leisure: the allocation of time over five decades. Working Paper No. 06-2, Federal Reserve Bank of Boston.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–17.

Caplow, T., Hicks, L. and Wattenberg, B. 2001. *The First Measured Century: An Illustrated Guide to Trends in America, 1900–2000*. Washington, DC: AEI Press.

Costa, D. 1997. Less of a luxury: the rise of recreation since 1888. Working Paper No. 6054. Cambridge, MA: NBER.

Greenwood, J., Seshadri, A. and Yorukoglu, M. 2005. Engines of liberation. *Review of Economic Studies* 72, 109–33.

Greenwood, J. and Seshadri, A. 2005. Technological progress and economic transformation. In *Handbook of Economic Growth*, vol. 1B, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Kopecky, K. 2005. The trend in retirement. Economie d'avant garde Research Report No. 12, Department of Economics, University of Rochester.

Lebergott, S. 1996. *Consumer Expenditures: New Measures and Old Motives*. Princeton, NJ: Princeton University Press.

Mulligan, C. 2002. A century of labor–leisure distortions. Working Paper No. 8774. Cambridge, MA: NBER.

Pareto, V. 1906. *Manual of Political Economy*. New York: Augustus M. Kelly. 1971.

Prescott, E. 2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28, 2–13.

Reid, M. 1934. *Economics of Household Production*. New York: John Wiley.

US Bureau of the Census. 1975. *Historical Statistics of the United States: Colonial Times to 1970*. Washington, DC: US Department of Commerce.

US Bureau of the Census. 2003. *Statistical Abstract of the United States*, Mini Historical Statistics (suppl.). Washington, DC: US Department of Commerce.

Vandenbroucke, G. 2005. A model of the trends in hours. Economie d'avant garde Research Report No. 11, Department of Economics, University of Rochester.

Whaples, R. 1990. The shortening of the American work week: an economic and historical analysis of its context, causes, and consequences. Ph.D. dissertation, University of Pennsylvania.

Williamson, J. 1995. The evolution of global labor markets since 1830: background evidence and hypotheses. *Explorations in Economic History* 32, 141–96.

How to cite this article

Greenwood, Jeremy and Guillaume Vandenbroucke. "hours worked (long-run trends)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000238> doi:10.1057/9780230226203.0748

The New Palgrave Dictionary of Economics Online

household portfolios

Michael Haliassos

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This entry presents recent work on portfolio behaviour of households and its possible departures from optimal behaviour. Topics include the role of household characteristics in influencing participation in stockholding and portfolio shares conditional on participation; portfolio implications of housing and housing debts; and portfolio coexistence of consumer debt, liquid assets and illiquid assets, with emphasis on credit card debt.

Keywords

age effects; asset allocation; asset ignorance; asset location; asset trading; bankruptcy; bequest motives; borrowing constraints; business equity; cohort effects; computational methods; conditional portfolio share; consumer debt; consumption risk; correlation between income risk and stock returns; credit card debt; credit cards; debt; debt refinancing; delinquency; diversification; earnings shocks; elasticity of substitution; Epstein–Zin preferences; equities; equity premium; financial wealth; fixed entry costs; fixed-rate mortgages; participation costs; home equity loans; homeownership; household finance; household portfolios; housing; housing collateral; hyperbolic discounting; interest-rate wedge; marginal investors; mortgages; mutual funds; real estate; refinancing; retirement; retirement accounts; risk aversion; social interactions; stockholding; stockholding participation rate; stockholding puzzle; stockholding risk; stocks; strategic default motive; time effects; trust; wealth distribution

Article

Household portfolios comprise the array of assets – financial (such as liquid accounts, stocks, bonds, and shares in mutual funds) and real (such as primary residence, investment real estate, and private businesses) – as well as liabilities held by a household, such as mortgages and consumer debt. This article focuses on three areas of active research – stockholding, housing, and credit cards – with respective household participation rates for the United States of the order of 50 per cent, two-thirds, and

two- thirds. European participation rates vary. Stockholding participation approaches 60 per cent in Sweden and 40 per cent in the UK, but it is less than 20 per cent in France, Germany, and Italy. Homeownership rates are closer to that of the United States, but in some countries, such as Germany, the majority does not own a home. The features of credit cards vary across European countries. In some countries, households have only debit cards linked to accounts with overdraft facilities.

The study of household portfolios, or ‘household finance’, is a partner to corporate finance and asset pricing, and it bridges economics and finance by extending analyses of saving to incorporate portfolio choice. It has grown considerably since the early 1990s, along with the complexity of household portfolios, in the face of ‘supply side’ developments encouraging risky asset holding. Privatization of public utilities in Europe was often accompanied by broad campaigns to educate households on the nature and benefits of stockholding. The demographic transition encouraged introduction of tax-deferred retirement accounts, promoted through educational campaigns, first in the United States and subsequently in Europe. The internet facilitated provision of information, opening of accounts, and trading internationally.

The development of household-level databases has in turn facilitated empirical research by allowing study of overall portfolios and their links to demographics and attitudes. Modern computational methods have enhanced understanding of behaviour towards non-diversifiable, background risk regarding income or health expenditures. Observed portfolio behaviour often differs from predictions of standard models, creating puzzles variously attributed to inadequate models or ‘investment mistakes’.

Stockholding

Understanding household stockholding is important, as it embodies key aspects of behaviour towards risk. In most countries, the majority of households holds no stocks, even indirectly through mutual funds, retirement, or managed accounts (Guiso, Haliassos and Jappelli, 2001; 2003). Exceptions were Sweden and the United States in 2001 (57 per cent and 52 per cent, respectively), but the United States fell back to 48 per cent in 2004. Non-participation despite an expected return premium (‘equity premium’) is inconsistent with standard expected utility maximization and constitutes the ‘stockholding puzzle’ (Mankiw and Zeldes, 1991; Haliassos and Bertaut, 1995). For a non-stockholder, stocks dominate bonds in expected return and do not contribute to consumption risk as they have zero covariance with consumption.

Various explanations have been proposed for limited participation in stock markets, given its widespread nature. Restrictions preventing borrowing at the riskless rate and short sales of stock yield zero stockholding, but only for poor households with no assets (Haliassos and Michaelides, 2003). Positive correlation between labour income risk and stock returns, coupled with short sales constraints, could justify zero stockholding among households intending to short stocks to hedge income risk, but is exhibited in practice by households likely to hold stocks – for example, the more educated and entrepreneurs.

The most widely accepted cause of limited participation is fixed entry or participation costs, actual or perceived, that discourage small potential investors. Costs can be wide-ranging, from brokerage costs to costs of one's time devoted to monitoring the stock market. In their presence, factors contributing to higher costs or lower desired stockholding, such as risk aversion or low resources, become relevant for

non-participation. An interest-rate wedge between borrowing and saving rates coupled with an empirically based assumption that borrowing rates are roughly equal to the expected return on equity also generates limited stock demand. Although Davis, Kubler and Willen (2006) offered this as an alternative to fixed costs for explaining non-participation, it could usefully serve also as a complement. Empirical estimates by Paiella (2001) and Vissing Jorgensen (2002), and numerically computed costs in Haliassos and Michaelides (2003) imply that relatively small fixed costs could justify observed patterns of non-participation.

The empirical participation literature provides various findings consistent with the presence of fixed costs (see contributions in Guiso, Haliassos and Jappelli, 2001; Rosen and Wu, 2004). More educated, financially alert, healthy households that belong to ethnic or education groups traditionally targeted by the financial sector are likely to face lower entry costs and to be more likely to participate, consistent with empirical findings. Similarly, households with greater expressed willingness to bear risk and those who do not perceive binding borrowing constraints are more likely to plan sizeable stock holdings and thus to overcome any given entry costs.

Empirical studies also point to other, often ignored, factors, which seem relevant for non-participation by those unlikely to be small investors, such as the rich. Limited social interactions and associated opportunities to exchange stockholding experiences, or lower expressed willingness to trust others, contribute to non-participation (Hong, Kubik and Stein, 2004; Guiso, Sapienza and Zingales, 2005). This can justify non-stockholding by some rich, in addition to possible substitution of private businesses for stocks (Heaton and Lucas, 2000). Non-participation also arises naturally if there is widespread ignorance of certain assets. Guiso and Jappelli (2005) found that only one-third of Italian households have simultaneous knowledge of stocks, mutual funds, and managed accounts. Moreover, although most of the literature has largely ignored tax considerations, tax laws have been shown to affect asset allocation, asset location, and trading (Bergstresser and Poterba, 2004).

Given that stockholding participation has increased, it is important to understand its economy-wide implications, as well as its future prospects in the face of changing stock market conditions. The limited existing theoretical literature already points to ambiguous effects of increased participation on wealth distribution (Peress, 2004; Guvenen, 2006). Since certain characteristics were empirically found to encourage participation, the composition of the stockholder pool is likely to change as participation spreads. If increased participation means progressive entry of ‘marginal’ investors with more limited resources and investment ability, it can contribute to lower stockholding levels, overtrading that lowers realized returns, and possibly greater wealth inequality. Households with lower education and resources have been shown to be more prone to ‘investment mistakes’ in terms of (non)participation, (under) diversification, and lack of debt refinancing (Campbell, 2006). Bilias, Georgarakos and Haliassos (2005; 2006) find evidence that the 1990s upswing attracted to the US stockholder pool households with characteristics, attitudes, and practices conducive to small stockholding levels, but this was reversed by entry and exit following the downswing. Overtrading characterizes households with brokerage accounts, but not the general population.

Households that do clear the participation hurdle need to decide what portfolio share to hold in stocks. Theory generates strong predictions on how this conditional portfolio share should be affected by household characteristics, but these are often not confirmed by the data. For example, under expected utility, constant relative risk aversion, and income risk, the share is predicted to fall with age or with the

ratio of cash on hand to permanent income (Cocco, Gomes and Maenhout, 2005). Either factor causes households to rely more on assets rather than on human wealth for financing consumption, and this reduces willingness to invest heavily in stocks. Yet the wealthy have conditional portfolio shares of risky assets about double those for the remaining population (Carroll, 2001). Although it is impossible to identify separately age, time, and cohort effects using cross-sectional data (Ameriks and Zeldes, 2005), regressions setting cohort effects to zero fail to find consistent dependence of conditional portfolio shares on age or resources (Guiso, Haliassos and Jappelli, 2003). Data from retirement accounts show great inertia in changing portfolio shares (Ameriks and Zeldes, 2005), while studies using discount brokerage accounts find overtrading (Barber and Odean, 2000). Representative data imply inertia in the population at large (Brunnemeier and Nagel, 2005; Bilias, Georgarakos and Haliassos, 2006). Gomes and Michaelides (2005) exploited the additional flexibility of departures from expected utility maximization in the form of Epstein-Zin preferences to approximate observed portfolio shares more closely. Under expected utility maximization and preferences exhibiting constant relative risk aversion, the risk aversion, prudence, and intertemporal elasticity of substitution parameters are linked. Lowering risk aversion (which increases the risky portfolio share) lowers prudence (thus precautionary wealth) and raises elasticity of substitution (thus saving for retirement). Epstein-Zin preferences allow simultaneous lowering of risk aversion and elasticity parameters. Households with low risk aversion, prudence, and elasticity parameters smooth earnings shocks with small assets, and almost never invest in equities in the presence of fixed costs. Those who clear the participation hurdle have higher parameters and moderate portfolio shares in stocks.

Housing

Although stocks are an interesting part of a household's portfolio, housing is the largest part, and it is both important and challenging to understand how homeownership interacts with the rest of the portfolio. Due to housing investment, younger and poorer investors have limited wealth to invest in stocks (Cocco, 2005). Payment commitments on mortgages may also discourage risky asset holding. Renters accumulating down payments for a house may be unwilling to jeopardize their accumulations by assuming financial risk. On the other hand, homeowners have access to home equity loans and other collateralized loans not available to renters, and ability to borrow may encourage financial risk taking. Understanding housing as a portfolio element cannot be accomplished without studying the structure of mortgages and their risk implications, on which there is surprisingly little research. Campbell and Cocco (2003) show that adjustable-rate mortgages are attractive to households that face no binding borrowing constraints but large inflation risk relative to real interest rate risk, and to potentially borrowing-constrained households with low risk aversion. They are unattractive to constrained, highly risk-averse households. Sluggishness to refinance despite significant rate drops has been found, especially among households with less wealth or financial sophistication (Campbell, 2006).

Credit card debt

Having discussed some household assets, financial and real, let us now turn to household debt and its

coexistence with assets, which received considerable attention as participation rates and median levels of indebtedness grew. Credit card debt behaviour is topical at the time of writing (2006), given increases in bankruptcy and delinquency rates that cannot be attributed to changes in debtor characteristics or supply factors (Gross and Souleles, 2002a). Gross and Souleles (2002b) documented two US credit card debt puzzles: (*a*) coexistence of high-interest card debt with substantial asset accumulation for retirement, suggesting a combination of short-run impatience with considerable patience for longer run objectives; and (*b*) coexistence of credit card debt with sizeable low-interest liquid assets that could have been used to pay it off.

The nature of these puzzles and the wide perception that credit cards make it difficult to control spending have led researchers mainly to behavioural explanations. Laibson, Repetto and Tobacman (2003) showed that a single rate of time preference has problems generating the former coexistence, and proposed hyperbolic discounting. The current self borrows because of short-run impatience. Accumulating illiquid assets is a way to control the future self, who will be impatient as retirement approaches.

The second puzzle seems to run against usual notions of arbitrage. Bankruptcy law allows households to rescue some assets, and this creates strategic default motives that discourage paying off debt, but strategic defaulters could avoid interest costs by buying exempt assets right before filing (Gross and Souleles, 2002b).

Bertaut and Haliassos (2002) and Haliassos and Reiter (2005) propose an ‘accountant-shopper’ model that generates both types of coexistence. The accountant self (or household member) revolves debt (partly) to constrain the amount charged by the impatient credit-card shopper, but this is not inconsistent with accumulating assets for retirement or other purposes. Caplin and Leahy (2004) model an absent-minded consumer who does not keep track of his spending. Credit cards may lead to overspending because they provide less information on spending flows than cash transactions.

Household portfolios entail numerous research challenges. They include further understanding of: interactions between real and financial assets and debts; sources of international differences in portfolio structure, especially around retirement; which part of unexplained portfolio behaviour is due to investment mistakes rather than model shortcomings; how labour market behaviour influences portfolios; the role of intra-household bargaining and risk sharing; the role of inattention and financial advice in the face of agency; and other incentive problems.

See Also

- consumption-based asset pricing models (theory)
- credit card industry
- financial market anomalies
- household surveys
- inheritance and bequests
- intertemporal choice
- non-expected utility theory
- precautionary saving and precautionary wealth

- recursive preferences
- risk aversion

Bibliography

Ameriks, J. and Zeldes, S. 2005. How do household portfolio shares vary with age? Mimeo, Columbia University.

Barber, B. and Odean, T. 2000. Trading is hazardous to your wealth: the common stock investment performance of individual investors. *Journal of Finance* 55, 773–806.

Bergstresser, D. and Poterba, J. 2004. Asset allocation and asset location: household evidence from the Survey of Consumer Finances. *Journal of Public Economics* 88, 1893–916.

Bertaut, C. and Haliassos, M. 2002. Debt revolvers for self-control. Mimeo. Federal Reserve Board.

Bilias, Y., Georgarakos, D. and Haliassos, M. 2005. Equity culture and the distribution of wealth. Working Paper No. 2005/20. CFS, University of Frankfurt.

Bilias, Y., Georgarakos, D. and Haliassos, M. 2006. Portfolio inertia and stock market fluctuations. Working Paper No. 2006/14. CFS, University of Frankfurt.

Brunnermeier, M. and Nagel, S. 2005. Do wealth fluctuations generate time-varying risk aversion? Micro-evidence on individuals' asset allocation. Mimeo, Princeton University.

Campbell, J.Y. 2006. Household finance. *Journal of Finance* 61, 1553–604.

Campbell, J. and Cocco, J. 2003. Household risk management and optimal mortgage choice. *Quarterly Journal of Economics* 118, 1449–94.

Caplin, A. and Leahy, J. 2004. The absent-minded consumer. Working Paper No. 10216. Cambridge, MA: NBER.

Carroll, C. 2001. Portfolios of the rich. In *Household Portfolios*, ed. L. Guiso, M. Haliassos and T. Jappelli. Cambridge, MA: MIT Press.

Cocco, J. 2005. Portfolio choice in the presence of housing. *Review of Financial Studies* 18, 535–67.

Cocco, J., Gomes, F. and Maenhout, P. 2005. Consumption and portfolio choice over the life-cycle. *Review of Financial Studies* 18, 491–533.

Davis, S., Kubler, F. and Willen, P. 2006. Borrowing costs and the demand for equity over the life cycle. *Review of Economics and Statistics* 88, 348–62.

Gross, D. and Souleles, N. 2002a. An empirical analysis of personal bankruptcy and delinquency. *Review of Financial Studies* 15, 319–47.

Gross, D. and Souleles, N. 2002b. Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. *Quarterly Journal of Economics* 117, 149–85.

Gomes, F. and Michaelides, A. 2005. Optimal life-cycle asset allocation: understanding the empirical evidence. *Journal of Finance* 60, 869–904.

Guiso, L., Haliassos, M. and Jappelli, T. 2001. *Household Portfolios*. Cambridge, MA: MIT Press.

Guiso, L., Haliassos, M. and Jappelli, T. 2003. Stockholding in Europe: where do we stand and where do we go? *Economic Policy* 36, 117–64.

Guiso, L. and Jappelli, T. 2005. Awareness and stock market participation. *Review of Finance* 9, 537–67.

Guiso, L., Sapienza, P. and Zingales, L. 2005. Trusting the stock market. Working Paper No. 11648. Cambridge, MA: NBER.

Guvenen, F. 2006. Reconciling conflicting evidence on the elasticity of intertemporal substitution: a macroeconomic perspective. *Journal of Monetary Economics* (forthcoming).

Haliassos, M. and Bertaut, C. 1995. Why do so few hold stocks? *Economic Journal* 105, 1110–29.

Haliassos, M. and Michaelides, A. 2003. Portfolio choice and liquidity constraints. *International Economic Review* 44, 143–78.

Haliassos, M. and Reiter, M. 2005. Credit card debt puzzles. Working Paper No. 2005/26. CFS, University of Frankfurt.

Heaton, J. and Lucas, D. 2000. Asset pricing and portfolio choice: the importance of entrepreneurial risk. *Journal of Finance* 55, 1163–98.

Hong, H., Kubik, J. and Stein, J. 2004. Social interaction and stock market participation. *Journal of Finance* 59, 137–63.

Laibson, D., Repetto, A. and Tobacman, J. 2003. A debt puzzle. In *Knowledge, Information, and Expectations in Modern Economics: In Honor of Edmund S. Phelps*, ed. P. Aghion et al. Princeton:

Princeton University Press.

Mankiw, N. and Zeldes, S. 1991. The consumption of stockholders and non-stockholders. *Journal of Financial Economics* 29, 97–112.

Paiella, M. 2001. Limited financial market participation: a transaction cost-based explanation. Working Paper No. 01/06. London: IFS.

Peress, J. 2004. Wealth, information acquisition and portfolio choice. *Review of Financial Studies* 17, 879–914.

Rosen, H. and Wu, S. 2004. Portfolio choice and health status. *Journal of Financial Economics* 72, 457–84.

Vissing Jorgensen, A. 2002. Towards an explanation of household portfolio choice heterogeneity: nonfinancial income and participation cost structures. Working Paper No. 8884. Cambridge, MA: NBER.

How to cite this article

Haliassos, Michael. "household portfolios." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000173> doi:10.1057/9780230226203.0749

The New Palgrave Dictionary of Economics Online

household production and public goods

Reuben Gronau

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Home production constitutes even in modern economies about one-third of GNP. The article discusses Becker's theory of home production and its critiques. It develops a general model where welfare is a function of market and home goods, market work, work-at-home and leisure, focusing on problems of its identification arising from the fact that home output is not traded in the market. These problems are aggravated in the multi-person household framework, since intra-household allocation is unobserved. These difficulties have serious ramifications for the measurement of adult equivalent scales, productivity at home and home output.

Keywords

Barten method; Barten, A.; Becker's household production model; children; equivalence scales; family; fertility; home goods vs. market goods; household production and public goods; intra-household distribution; Kuznets, S.; leisure; marginal productivity; psychic income; real business cycles; schooling; shadow prices; time use; value of time; women's work and wages

Article

The concept of household production (or home production) is not new to economics. It is often used synonymously with ‘cottage industries’ – production taking place at home – and is generally associated with less-developed economies. Mincer (1962) emphasized the importance of the substitution between work at home and work in the market in developed economies for the understanding of married women's labour supply decisions. He was also the first to point out the importance of time scarcity for the analysis of fertility decisions, the demand for maids, and the choice of transport modes (Mincer, 1963). It was, however, Becker's seminal paper (1965) that made the concept of household production an integral part of economic theory.

Becker's theory of household production and its critiques

Becker introduced two novel elements into classical consumption theory. Whereas the classical consumer maximizes welfare subject to the budget constraint, and the object of welfare is the goods consumed, in Becker's analysis the object of welfare is the household's activities (‘commodities’, in Becker's terminology), where each activity is a combination of market goods and time inputs. The household maximizes welfare subject to two constraints – the budget constraint and the time constraint (the fact that the different time uses, at home and in the market, cannot exceed total time available). In this model the household's decisions can be divided into two stages: (*a*) the production stage (how to ‘produce’ each activity?) and (*b*) the consumption stage (what is the optimal activity bundle that will maximize welfare?). The ‘household production’ decisions are determined by the household technology and the relative factor prices, and the consumption decisions are determined by the activity ‘shadow’ prices and by the total resource constraint.

The new theory diverges from classical consumption theory in several important respects. Whereas in classical theory all households face the same prices, in the new framework different households place different values on their time, choose a different input mix, and consequently face different activity prices. Different consumption bundles consumed by households with identical incomes do not attest necessarily to differences in preferences, but may be traced to differences in home technology or in the implicit value of time. Specifically, when time can be moved freely from home uses to work in the market, and when work in the market does not involve any direct disutility, the implicit value of time will equal the (marginal) wage rate. Consumers who earn higher wages are expected to produce each activity using a more goods-intensive input mix – conserving on time. The more time-intensive the activity is (for example, sleep or watching television) the more expensive it becomes, and the less favourable it becomes for high-wage earners, who are expected to choose a more goods-intensive set of activities. In the Becker framework the theory of consumption is integrated with labour supply analysis.

The model of household production was instrumental in the development of demand analysis for fertility (Willis, 1973; Becker and Lewis, 1973), health (Grossman, 1972), transport (Gronau, 1970) and other applications. The popularity of the model can be traced to the insights gained by combining consumption and production theory to explain household behaviour. One of the few dissenting voices was that of Pollak and Wachter (1975). In Becker's original model the shadow prices of the activities are independent of the amount of the activities consumed. The authors point out that this assumption is satisfied only under very restrictive conditions. For this to hold, the marginal inputs of time and goods cannot vary with ‘output’, and the shadow price of time has to be constant. The first assumption requires that the production process be subject to constant returns to scale, and the second assumption requires that the time inputs do not generate any direct utility per se (that is, in Becker's model one enjoys the commodity ‘children’ but not the childcare going into their ‘production’). The first assumption rules out the existence of increasing returns to scale, often mentioned as one of the economic motives for the establishment of multi-person households, while the second is at odds with the standard distinction (emphasized by Mincer) between leisure and work at home. In this formulation a meal is a meal, regardless whether one worked on it for two hours and ate it in five minutes, or worked on it for five minutes and ate it in two hours.

A three-way allocation of time: market work, work at home and leisure

The distinction between the two types of home time was resurrected by Gronau (1977), who proposed a model where the consumer allocates his time between three time uses: leisure, work in the market, and work at home, the last of these serving as an input in the production of ‘home goods’. In the most general formulation welfare is defined over the three time uses and the two types of goods (home and market goods) $U = U(X_m, X_h, T_m, T_h, L)$, where X_m denotes market goods, X_h home goods, T_m market time, T_h work at home time, and L leisure. The home production function is $X_h = F(T_h)$. The constraints confronting the person are the budget constraint $X_m = wT_m + V$, where w is the real wage rate, and V non-labour sources of income; and the time constraint $T_m + T_h + L = 1$. The first order conditions for an interior solution (that is, $T_m > 0$, $T_h > 0$) are

$$(U_L - U_{Tm}) / U_{Xm} = w \text{ and } (U_L - U_{Th}) / U_{Xh} = F',$$

where F' denotes the marginal productivity of work at home. Combining the two equations, one obtains the familiar factor demand equation

$$(U_{Xh} / U_{Xm}) F' = [(U_L - U_{Th}) / (U_L - U_{Tm})] w,$$

stating that the value of marginal productivity of work at home equals the ‘shadow’ price of time at home. (U_{Xh} / U_{Xm}) denotes the ‘shadow’ price of home goods, and the ‘shadow’ price of time is corrected for the differential in direct utilities of work in the market compared with work at home $[(U_L - U_{Th}) / (U_L - U_{Tm})]$.

Unfortunately, three out of the four terms in this equation are unobserved (the ‘shadow’ price of home goods, the marginal productivity of work at home, and the price of time correction factor), limiting the applicability of this equation for empirical research. Thus, changes in the observed variable, the wage rate, can be used to trace the parameters of any of the unobserved terms, but only if the parameters of the other two unobserved terms are arbitrarily restricted.

Gronau (1977) assumed that home and market goods are perfect substitutes $(U_{Xh} = U_{Xm})$, as are home and market work, yielding the dual condition for an interior optimum $(U_L - U_T) / U_X = F' = w$. The existence of two separate margins allows the tracing of the slope of the production function and the contours of the indifference curve between work time and goods. In this scheme the choice between leisure and goods is governed by preferences, and the allocation of work time between home and market is determined by technology.

Other studies tried to isolate other components of the equation, imposing a different set of restrictions. For example, Kerkhofs and Kooreman (2003), following Graham and Green (1984), estimated the

psychic income from work at home by assuming perfect substitution between home and market goods and restricting the marginal productivity F' to be a linear function of work at home. Rupert, Rogerson and Wright (1995) focused on the elasticity of substitution between home and market goods; but in order to obtain credible estimates of this parameter they had to assume that home and market work are perfect substitutes and impose specific values on the home production elasticity.

Home production and intra-household distribution

Becker's original model strictly applies only to a one-person household. Several attempts have been made to adapt it to a multi-person environment (and specifically to the husband–wife case). The multi-person household models add to the household decisions a third dimension – the intra-household distribution. Given the difficulties encountered in separating consumption from production, adding a third set of unobservables does not contribute to the tractability of the models.

The models agree that each spouse's leisure should appear separately in the welfare function, and that the spouses’ work at home is mutually substitutable in the home production function. There is, however, disagreement over whether home goods are private or public goods. The specific formulation of the welfare function varies depending on whether the researcher belongs to the ‘unitary’ or the ‘collective’ camp.

The empirical analysis reflects the difficulty in separating consumption (that is, the shadow price of home goods, the psychic income from work at home), household production technology (that is, the marginal productivity of work at home) and intra-household distribution effects. The most important ‘output’ of home production in most households is their children. Children (in particular when they are young) are associated universally with increased work at home and childcare. It is, however, impossible to tell how much of the increased time input is due to the increased shadow price of home goods, and how much should be attributed to the increased psychic income derived from work at home.

Similar difficulties affect the analysis of the factors affecting home productivity. A central theme in this analysis is the estimation of the returns to scale in home production or, alternatively, how important is the public-goods component of home goods (for example, home repair, house cleaning, laundry, cooking, shopping).

The analysis of returns to scale is one of the oldest chapters in empirical economics, dating back, under the heading of ‘Adult equivalence scales’, to the studies of Engel at the end of the 19th century. Equivalence scales are index numbers intended to allow comparisons of welfare (or real incomes) across households of different size and composition. The discussion of the methods of estimation of these index numbers on the basis of observed consumption patterns generated an extensive literature. The literature is unanimous in concluding that there exist substantial returns to scale in consumption. According to a survey paper by Van Praag and Warnaar (1997), discussing 76 studies, it is found that on average a two-person household ‘needs’ per person only 80 per cent of the resources ‘needed’ by a one-person household, and a three-person household ‘needs’ per person only two-thirds of a single-person household. Unfortunately, these estimates suffer from the shortcomings common to all multi-person models of household production: they cannot separate unobserved household technology from unobserved intra-household distribution rules. This difficulty is, perhaps, best demonstrated by one of the more sophisticated methods of estimation of the equivalence scales – the Barten method.

Barten (1964), recognizing that demographic changes may have different effects on different goods, allowed for goods-specific scales. In his formulation welfare is a function of the deflated quantity of goods (X_i/M_i), where the value of the goods-specific deflator M_i reflects the returns to scale in its consumption. Barten's formulation looks very similar to that of Becker's household production model, where time inputs are omitted, and where it is assumed that the marginal productivity of goods in the production of activities (‘commodities’) is constant and equal to $1/M_i$. If we follow this analogy, retrieving M_i should yield an estimate of the parameters of household technology.

Various suggestions have been made on how to estimate the deflators M_i by comparing the consumption patterns of households of different size. However, when total consumption is given, differences in consumption patterns between a single-person household and a multi-person household reflect both the difference in the consumption patterns of the household's members and the resources each of them commands (if all members allocate their resources identically between all goods, the single-person household and a multi-person household will have the same consumption patterns, and the comparison will generate only ‘noise’). Hence, preferences (or technology) and distribution are inseparably entangled, and there is no way to separate returns to scale from the distribution rule (Gronau, 1988).

Home production productivity

Productivity is positively correlated with physical capital investments, and there is unanimity that married women's increased productivity at home, due to increased investment in home equipment, has been an important factor explaining their increase in labour force participation since the 1950s. Investments in human capital (schooling, health and on-the-job training) have been shown to increase a person's productivity in the market. Do these investments have side benefits at home? Michael (1973), who studied the consumption patterns of households with different schooling, concluded that schooling significantly increases productivity in the use of goods in home production. Gronau (1973) studied the impact of schooling on the productivity of time use. He focused on the schooling effect on married women's reservation wage, where the unobserved reservation wage is imputed from their labour-force participation decisions. He found that college education raises the value women place on their time at home by 20 per cent compared with high-school graduates (about half the effect schooling has on their productivity in the market). Finally, Gronau and Hamermesh (2001) argued that schooling makes people more productive at home by allowing them to squeeze more ‘leisure activities’ into a smaller amount of ‘free’ time.

Home production and the macroeconomy

Inspired by Becker's original analysis, the orientation of most of the studies of household production was microeconomic, emphasizing the behaviour of individual households. This orientation changed following Becker's 1987 AEA presidential address (1988), demonstrating the implications of family economics and household production for growth and the macroeconomy. The challenge was met by Benhabib, Rogerson and Wright (1991) and Greenwood and Hercowitz (1991). The two teams tried to explain some irregularities in the traditional model of the real business cycle (RBC) in terms of shifts

from the market economy to home production. Benhabib, Rogerson and Wright tried to obtain a better explanation for the fluctuation in labour inputs over the business cycle, whereas Greenwood and Hercowitz focused on capital formation. Both topics have become the theme of several sequels, while other authors use home production to explain a wide range of additional macro phenomena: endogenous growth, development, fiscal policy and the business cycle, and the welfare cost of inflation. The common technique employed in the new generation of models is calibration, the models sharing the common assumption that market and home goods are close, though not perfect, substitutes. (An early survey of the topic and the literature is contained in Cooley, 1995, chs. 1, 5, and 6.)

The measurement of household output

While the new breed of macroeconomic studies is purely theoretical, the emphasis of an older macroeconomic branch, closely related to the national income accounting family, is purely on measurement. The exclusion of the output of the home sector has long been recognized as a major omission in national accounting (Kuznets, 1944), an omission that can seriously bias international comparisons of standards of living and estimates of growth rates. Several attempts have been made to correct this lapse.

The value of output in the home sector, as that of other non-market sectors (such as the government and non-profit organizations) is measured in terms of the value of inputs used in the production process. There is, however, an inherent difference between the home sector and the other non-market sectors, namely, that the time inputs used in home production do not carry an explicit price tag. Two methods have been suggested to circumvent this difficulty: (*a*) the market opportunity cost method, and (*b*) the market alternative method. According to the first method, the time inputs are evaluated according to the price they can command in the market. The second approach tries to evaluate home services at their market prices. Both methods are vulnerable to serious conceptual objections.

The objection to the ‘opportunity cost’ method stems from the fact that the same service (say, childcare) is evaluated at different prices if the provider gets a different wage in the market. The objection to the ‘market alternative’ method is that the household could have bought the home services at these prices but has rejected this option. These difficulties can be traced again to the inherent problem of identification of the work-at-home demand equation.

If one could assume that work at home yields no psychic income, then the ‘opportunity cost’ method should be employed, since in this case differences in the market wage attest to differences in the evaluation of home goods (for example, because of differences in the conceived quality of service). On the other hand, if women with different market wages perform the same service at home merely because of differences in psychic income, then the ‘market alternative’ approach should be preferred.

Even if the conceptual difficulties could be resolved, some technical difficulties remain. The ‘opportunity cost’ method has to cope with the problem that a substantial fraction of home output is produced by ‘full-time’ workers in the home production ‘industry’ (that is, house-persons) who receive no market wage (Gronau, 1973). Moreover, these workers should be regarded as self-employed, and the evaluation of their output should incorporate the returns to their entrepreneurial capacity (Gronau, 1980). The ‘market alternative’ approach advocates are undecided whether to use the cost of a maid as the market alternative or whether each home service should be priced separately.

Given the often-heated debate between the proponents of the two methods, the imputation outcomes show a surprising degree of similarity. Hawrylyshyn (1976), who compared nine international studies of both types, found that the average estimate of the value of home production is 35 per cent of GNP, with the estimates ranging from 32 to 39 per cent.

See Also

- Becker, Gary S.
- childcare
- collective models of the household
- equivalence scales
- intrahousehold welfare
- Kuznets, Simon
- leisure
- real business cycles
- time use
- value of time

Bibliography

Barten, A. 1964. Family composition, prices and expenditure patterns. In *Econometric Analysis for National Economic Planning*, ed. P. Hart, G. Mills and J. Whitaker. London: Butterworth.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–517.

Becker, G. 1988. Family economics and macro behavior. *American Economic Review* 78, 1–13.

Becker, G. and Lewis, H. 1973. On the interaction between quantity and quality of children. *Journal of Political Economy* 81, S279–S288.

Benhabib, J., Rogerson, R. and Wright, R. 1991. Homework in macroeconomics: household production and aggregate fluctuations. *Journal of Political Economy* 99, 1166–87.

Cooley, T., ed. 1995. *Frontiers of Business Cycle Research*. Princeton: Princeton University Press.

Graham, J. and Green, C. 1984. Estimating the parameters of the household production function with joint products. *Review of Economics and Statistics* 66, 277–82.

Greenwood, J. and Hercowitz, Z. 1991. The allocation of capital and time over the business cycle. *Journal of Political Economy* 99, 1188–214.

Gronau, R. 1970. The effect of traveling time on the demand for passenger transportation. *Journal of Political Economy* 78, 377–94.

Gronau, R. 1973. The effect of children on the housewife's value of time. *Journal of Political Economy* 81, Part II, S168–S199.

Gronau, R. 1977. Leisure, home production, and work-the theory of the allocation of time revisited. *Journal of Political Economy* 85, 1099–123.

Gronau, R. 1980. Home production – a forgotten industry. *Review of Economics and Statistics* 62, 408–16.

Gronau, R. 1988. Consumption technology and the intrafamily distribution of resources: adult equivalence scales reexamined. *Journal of Political Economy* 96, 1183–205.

Gronau, R. and Hamermesh, D. 2001. The demand for variety: a household production perspective. Working Paper No. 8509. Cambridge, MA: NBER.

Grossman, M. 1972. The demand for health: a theoretical and empirical investigation. Occasional Paper No. 119. New York: NBER.

Hawrylyshyn, O. 1976. The value of household services: a survey of empirical estimates. *Review of Income and Wealth* 22, 101–31.

Kerkhofs, M. and Kooreman, P. 2003. Identification and estimation of a class of household production models. *Journal of Applied Econometrics* 18, 337–69.

Kuznets, S. 1944. *National Income and its Composition*, vol. 2. New York: NBER.

Michael, R. 1973. Education in non-market production. *Journal of Political Economy* 81, Part I, 306–27.

Mincer, J. 1962. Labor force participation of married women: a study of labor supply. In *Aspects of Labor Economics*, ed. H. Lewis. Princeton: Princeton University Press.

Mincer, J. 1963. Market prices, opportunity costs, and income effects. In *Measurement in Economics*, ed. C. Christ. Stanford: Stanford University Press.

Pollak, R. and Wachter, M. 1975. The relevance of the household production function and its implications for the allocation of time. *Journal of Political Economy* 83, 255–78.

Rupert, P., Rogerson, R. and Wright, R. 1995. Estimating substitution elasticities in household production models. *Economic Theory* 6, 179–93.

Van Praag, B. and Warnaar, M. 1997. The cost of children and the use of demographic variables in consumer demand. In *Handbook of Population and Family Economics*, Vol.1A, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Willis, R. 1973. A new approach to the theory of fertility behavior. *Journal of Political Economy* 81, S14–S64.

How to cite this article

Gronau, Reuben. "household production and public goods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000164> doi:10.1057/9780230226203.0750

The New Palgrave Dictionary of Economics Online

household surveys

Duncan Thomas
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Household surveys play a pivotal role in empirical economics. Cross-section and longitudinal surveys are regularly conducted worldwide. A description of survey design and sampling methods provides the foundation for discussing survey errors. These include errors associated with sampling, survey coverage and non-response (which includes attrition from panel surveys), and errors of observation or measurement. In recent years, surveys have tended to become more complex and broader in scope with many reaching beyond measuring economic choices, constraints and outcomes. This trend will likely continue and exciting technological innovations in survey methods and implementation promise to revolutionize the field.

Keywords

bootstrap; clustering; cohort survey; consumer expenditure; coverage error; cross-section surveys; demographic surveys; Engel's Law; fertility surveys; health surveys; household production; household surveys; human capital; jackknife; longitudinal (panel) surveys; non-observational and observational errors; probability sampling; sampling error; synthetic panels

Article

Household surveys provide one of the pillars upon which some of the most important innovations in economics during the last half of the 20th century have been built. Enumeration of households dates back at least to the collection of budget data in the late 18th century. Eden (1797) compiled information on the diet, dress, fuel, and habitation spending as well as earnings of households from 86 households in England, while Davies (1795) reported detailed budgets of 127 households engaged in agriculture. Both studies sought to describe the lot of the poorest in England and so the budgets are not representative of the English population at the time. Ducpétiaux (1855) published the budgets of 199 Belgian households. Those data provided the empirical foundation for Engel's Law (Engel, 1857) which posits an inverse

relationship between income and the share of the budget spent on food.

Statistical foundations

The development of practical methods of probability sampling and a theory to support estimation and inference based on those samples had a major impact on the design and implementation of household surveys. Work by Neyman (1934) on stratified designs and work on randomization in agricultural experiments by Fisher (1935) were especially influential, and their work, in combination with contributions by inter alia Bowley (1926), Deming (1950), Kaier (1895) and Yates (1935) provided a theoretical foundation for survey design.

The importance of scientific surveys was underscored by some spectacular failures. For example, in 1936 the *Literary Digest* mailed out ten million questionnaires in a poll about the election of the next US president. About two million respondents mailed back their questionnaires, and the *Digest* predicted a victory for the Republican candidate, Alfred Landon. The election was won by a landslide – not by Landon but by his opponent, Franklin Roosevelt. There were also very influential survey successes. For example, Mahalanobis (1940) highlighted the advantages of surveys in terms of cost and timeliness of results. Using a sample survey of jute producers in Bengal, he estimated the area under jute within three per cent of the official estimate based on a complete census. The cost of the sample survey was only about eight per cent of the cost of the census. His sample survey cost eight per cent of the census.

These advances laid the foundation for an explosion in the quantity and quality of household surveys during the second half of the 20th century. Many of the surveys have been designed and implemented by national statistical agencies. At a substantive level, there are at least three important classes of household surveys, each of which has specific goals.

First, household budget surveys collect detailed information on the spending patterns of households. They are used to calculate price indices and poverty lines and to estimate the incidence of poverty. These include the Indian National Sample Survey, the Family Expenditure Survey (FES) in the UK and the Consumer Expenditure Survey (CEX) in the United States. Nowadays, virtually every country in the world conducts household budget surveys periodically. In some cases, respondents are asked to maintain a diary of spending over a pre-specified time period. In other surveys respondents are interviewed and asked to recall spending on items, often with varying recall periods depending on the item. The diary method typically covers a relatively short time period, which complicates modelling low frequency purchases and interpreting reported spending as indicative of longer-run resource availability. The interview method is potentially affected by recall error. This includes forgetting (which increases with the recall period) and telescoping, which may be positive (if spending before the recall time frame is telescoped into the recall period) or negative (if spending during the recall time frame is telescoped out of the period). Whether the interview or diary method yields less measurement error remains an open question.

Second, labour force and income surveys are collected routinely to monitor inter alia labour force participation, unemployment and earnings. Labour force surveys tend to be administered frequently and samples are large enough to detect small changes in the labour market. In the United States, for example, the Current Population Survey (CPS) is a monthly survey of over 50,000 households that has been conducted for over 50 years. Some surveys focus on income and wealth. The Survey of Consumer

Finances measures the financial health of the US population and includes a special over-sample of the most wealthy households.

The third class of surveys measure non-economic domains of well-being. Fertility surveys provide information on marriage and living arrangements, reproductive health including pregnancies and births, and use of health services. These are important for documenting the dramatic changes in family formation, composition and size that has occurred over the 20th century. Health surveys monitor the health of the population. In some cases, such as the National Health and Nutrition Examination Survey (NHANES), an extensive physical examination is performed by trained medical personnel in conjunction with a detailed questionnaire about health status and health-related behaviours. Several surveys integrate demographic with health information including the Demographic and Health Surveys (DHS), which grew out of the World Fertility Surveys and have been collected in over 75 countries. Surveys of attitudes, like the General Social Survey, are routinely collected across the globe.

In practice, the distinction between these classes of surveys is not clear-cut since many of the economic surveys record demographic, health or attitudinal information, and vice versa. To be sure, these topic-specific surveys are extremely important for monitoring the prevalence of indicators of interest to researchers and policymakers. However, the surveys are often inadequate for testing hypotheses about behaviours of individuals and their families.

In the late 1960s, surveys were designed to address this limitation, explicitly drawing on the theoretical models of household behaviour suggested by Gary Becker, T.W. Schultz, and their collaborators and students. One class of surveys explicitly recognized the dual role of households in agricultural economies as both producers and consumers of food. See, for example, Evenson (1978) for a discussion of a series of innovative household surveys conducted by nutritionists and economists in Laguna Province, Philippines. These surveys collect detailed information on farm inputs and output, non-farm activities, consumption, health and demographic behaviour.

Another class of surveys relied on the economic model of household production to guide the collection of information on individual choices and constraints people face. For example, the RAND Malaysian Family Life Survey (MFLS) was designed to capture multiple domains of the lives of each individual respondent, their family and community to better understand the determinants of fertility and investment in children during early life (Butz and DaVanzo, 1975). As a result of the scope of the questions, MFLS has been used to address a far broader array of questions in economics and demography than those for which it was originally conceived. The International Crops Research Institute for Semi-Arid Tropics (ICRISAT) village-level studies (VLS) followed a similar approach. The best known of these was conducted in six villages in three regions of semi-arid India and collected very detailed data on a very broad array of topics from 240 farm households surveyed annually for ten years (Walker and Ryan, 1990).

The Living Standard Measurement Surveys (LSMS) conducted by the World Bank drew heavily on the experiences of the Laguna, MFLS and ICRISAT studies among others. Conceived as broad-purpose surveys to monitor poverty and material well-being in developing countries and also contribute to the design of social policy, the surveys collect a wide array of indicators of well-being and behaviours of households along with extensive community data. Initiated in the mid- 1980s, a hallmark of the LSMS program is a framework that is broadly consistent across many countries. Having been implemented in many low-income and transition countries around the world, LSMS and DHS stand out as leaders in the development of comparable survey data collected from a wide spectrum of social and economic contexts.

Survey design

A typical household survey selects a sample of households from a frame which is the population of interest for the research. In many cases, the frame is a census and the sample is representative of a geographic area, although this need not be the case. The simplest sampling strategy randomly selects households from the frame. In practice, most household surveys follow a two-stage (or multi-stage) sampling design in which clusters are selected and then households are selected from those clusters. There are several advantages associated with geographically-defined clusters. Administration costs are lower for surveys that involve face-to-face interviews. Clusters may facilitate incorporating neighbourhood- or community-level data in the survey or, alternatively, models might highlight variation within communities and control community-level heterogeneity with a fixed effect, for example.

Clustering also carries disadvantages since two sampled units within a cluster tend to be more similar than two randomly selected units. The loss of independence across sampled units results in lower precision and thus larger standard errors of estimates. The magnitude of this effect for a particular indicator is often summarized by the design effect which is the ratio of the variance, with the cluster design taken into account, to the variance if households were randomly selected. An alternative summary statistic is provided by the intra-cluster correlation coefficient. The greater the covariance within clusters relative to differences across clusters, and the larger the number of households within a cluster, the greater is the design effect and the greater is the loss of precision due to clustering. It is standard practice to estimate standard errors by taking account of the clustering following the method of Huber (1967) or a re-sampling approach such as the jackknife or bootstrap (Efron, 1982). In short, clustering buys more information per unit cost but less information per sampled unit.

Many surveys are designed to oversample specific sub-populations, in which case estimates are typically adjusted for the probability of a household being selected into the sample. An important principle underlying population-based sampling is that because the probability of selection of every eligible unit is known and greater than zero, with appropriate weights, it is possible to reconstruct the population, although in some instances the complexity of survey designs becomes overwhelming.

Survey errors

There are at least two classes of error in any survey. ‘Non-observational’ errors occur when part of the target population is not measured. ‘Observational’ errors are the result of incorrect measurement. Sampling error, the most familiar survey error, is a form of non-observational error. It reflects the fact that any sample is a subset of the underlying population and so an estimate based on the sample will not be identical to the population value.

Coverage error, another source of non-observational error, arises when the sampling frame excludes part of the target population. Many sample frames are based on a list of household dwellings; those samples exclude homeless people and so are not representative of the entire population. If a household listing is based on an old census, more mobile people are at risk of being under-represented. A sampling frame based on telephone numbers (or e-mail addresses) will exclude those who do not have a telephone (or e-

mail address) and oversample those with multiple numbers (or addresses).

A third source of non-observation error arises from non-response, of which there are two categories. First, survey non-response occurs when a target respondent cannot be located. It will also arise if the respondent refuses to participate in the survey (or fails to answer the telephone, respond to an e-mail or return a mailed-out survey). Second, item non-response occurs when a respondent fails to answer one or more questions in the survey either because he or she refuses to answer or does not know the answer to the question(s). The incidence of the latter is reduced by probing, and unfolding brackets have proved to be particularly useful for economic quantities (Hurd et al., 1998).

Broadly speaking, non-response rates tend to rise with the value of time of the respondent, and there has been a secular trend of increased non-response in many developed countries. Survey non-response in developing country household surveys is typically substantially lower than in higher income countries. If, conditional on observed characteristics, coverage and non-response error are random, appropriate weights can be computed so that survey statistics are representative of the underlying population. Complications arise when these errors are selected on unobserved characteristics. Several procedures have been suggested to deal with non-response error including hot deck or matching procedures (Rosenbaum and Rubin, 1983) and modelling the selection process with a control function (Heckman, 1978).

The most familiar source of observational error is respondent failure to answer a question correctly. This may be intentional (in order to misrepresent reality) or unintentional. Interviewers may make errors in the administration of the survey, and there may be interviewer-specific effects in the ways questions are asked. Survey instruments are also prone to error. In general, the extent of observational error likely depends on interactions among the sources of error and also on the mode of the survey. Respondents in telephone surveys tend to provide shorter answers than those in face-to-face interviews, and web-based surveys are more likely to be ended prematurely.

While the distinction between observational and non-observational error is conceptually useful, in practice the distinction is often blurred. For example, survey non-response is typically related to interviewer characteristics. Both item non-response and respondent error have been shown to be related to questionnaire design and interviewer characteristics. Groves (1989) provides an excellent discussion of these and related issues.

Typology of surveys

Cross-section surveys provide a snap-shot of a target population at a point in time. They are the bread and butter of research based on household surveys. Many cross-section surveys are repeated regularly, with independent samples drawn from the same target population, so that it is possible to track the evolution over time of indicators such as unemployment, poverty or inequality as well as map changes for population sub-groups. Synthetic panels of individuals created using repeated cross-section follow the same population subgroup over time, such as a birth cohort. They are straightforward to interpret if there are no entrants into or exits from the target population via, for example, immigration, emigration or death. Synthetic panels of households are more complicated. Household composition changes due, for example, to marriage or divorce result in changes over time in the unit being followed. It is difficult to distinguish composition changes from true change. Similar issues arise with synthetic panels of

communities.

Longitudinal or panel surveys follow the same respondent over time, which provides opportunities for exploration not feasible with cross-section surveys. First, tracing the dynamic evolution of choices and outcomes over the individual's life provides insights into, for example, early life experiences and later life outcomes, resilience and recovery from adversity as well as the characteristics of those who cycle in and out of some state (such as poverty, unemployment, public assistance or poor health).

Second, panel data provide expanded options for treating unobserved heterogeneity in models like

$$y_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}$$

where μ_i is an unobserved individual-specific characteristic. If μ_i is correlated with x_{it}, OLS estimates of β are biased. With repeated observations on the same individual in a panel, μ_i can be estimated (or the model cast in first-differences) to consistently estimate β. The 'fixed effect' μ_i absorbs all time-invariant individual characteristics that enter the model in a linear and additive way.

The advantage of a longitudinal survey is that the same sampling unit is followed over time. This is also its Achilles heel. Attrition from longitudinal surveys is a particular form of non-response error. The nature and magnitude of attrition varies with the study design. For example, in face-to-face interviews in the home, individuals who move are followed to their new location and interviewed there. Those who move the furthest are often the hardest and most expensive to find. Attrition tends to be selected on traits associated with migration – younger, better-educated adults being the most likely to move. The selectivity of the sample is exacerbated in panel surveys that do not follow people who leave the location in which they were interviewed at baseline. Attrition in telephone and web-based surveys have less to do with tracking people to new geographical locations and more to do with retaining the cooperation of respondents – an issue that also confronts face-to-face interviews. In multi-wave panel surveys, it is important to attempt to re-contact respondents who have been skipped in prior waves so that attrition does not cumulate. There are many examples of well-designed panel surveys that have kept attrition low across multiple rounds.

Statistical adjustments for attrition are the same as other forms of non-response error. Re-weighting will be effective when attrition is selected on observed characteristics. When selection is on unobserved characteristics, a control function approach is more likely to be successful. In analytical models, the importance of adjusting for attrition will vary with the research question. The stronger the association is between attrition and observed or unobserved characteristics in the model, the more important the adjustment is for attrition.

An alternative approach to treating attrition is to replace a respondent who attrits from the survey with a new, similar respondent – frequently people living in the same housing structure as the respondents in the previous wave (or the person who is assigned the telephone number, e-mail address, and so forth). There are several problems with this approach. First, it assures the study population appears stable since no primary sampling units will lose population; the reality may be quite different. Second, housing structures can change, be torn down or difficult to relocate, resulting in a different type of attrition. Third, even if populations are stable in aggregate and housing structures do not change, it is assumed

that the replacement and original respondents are ‘exchangeable’ or effectively identical. It is not clear that this will be true as in the case of a respondent who died. Fourth, the key advantage of a longitudinal survey – following the same person through the life course – is lost.

It follows that a panel survey of households has little conceptual appeal. Although a household survey is often the baseline for a panel of individuals, households change over time and it is individuals who will be followed – possibly all the original household members. These respondents will often be interviewed along with the people in their new household and so the panel is a series of household surveys embedded in which is a longitudinal survey of individuals. A small number of longitudinal surveys have sought to follow family members over time.

The Panel Survey of Income Dynamics (PSID) is a long-running panel and one of the most widely used surveys in economics. Initiated in 1968, with a nationally representative sample of 5,000 households, interviews spanning 40 years with household members, and children born to them, has provided unique insights into the dynamics of income, human capital, health and living arrangements over the life course, across cohorts and across generations (Duncan, Hofferth and Stafford, 2004).

A cohort survey is a special type of longitudinal survey which follows a specific cohort of respondents, often a birth cohort. The advantages of the design are that, because of shared environments, cohort members are less heterogeneous than the entire population and there are power benefits to comparing people making similar life course transitions at the same time. A disadvantage is that age and period effects cannot be disentangled. To address this, cohort studies often draw new cohorts. The British Cohort Studies, for example, have mounted four large-scale population-representative birth cohort studies since the 1930s. The Health and Retirement Survey (HRS) is an innovative cohort study that focuses on the health and economic well-being of older Americans. The HRS has been replicated in several countries across the globe.

Statistical innovation presaged the explosion in household surveys since the 1950s. Technological innovation is likely to provide the foundation for the next revolution in survey design. For example, electronic communication devices, geographical information systems and innovations in health measurement along with sophisticated analytical tools have already begun to profoundly affect the scope and quality of household surveys.

Bibliography

Bowley, A.L. 1926. Measurement of precision attained in sampling. *Bulletin de l'Institut International de Statistique* 22(Suppl. to Livre 1), 6–62.

Butz, W.P. and DaVanzo, J. 1975. *The Malaysian Family Life Survey: Summary Report*. Santa Monica: The Rand Corporation.

Davies, D. 1795. *The State of Labourers in Husbandry Stated and Considered*. Bath.

Deming, W.E. 1950. *Some Theory of Sampling*. New York: Wiley.

Ducpétiaux, E. 1855. *Budgets économiques des classes ouvrières en Belgique*. Bruxelles.

Duncan, G., Hofferth, S. and Stafford, F. 2004. Evolution and change in family in come and wealth: the Panel Survey of Income Dynamics 1968–2000 and beyond. In *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*, ed. J. House, T. Juster and R. Kahn. Ann Arbor: University of Michigan Press.

Eden, F.M. 1797. *The State of the Poor*. London: J. Davis.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: SIAM.

Engel, E. 1857. Die Productions- und Consumptions verhältnisse des Königreichs Sachsen. *Zeitschrift des Statistischen Büreaus des Königlich Sächsischen Ministerium des Innern* 8(9), 22 November. Repr. in *Bulletin de I'Institut International de la Statistique* 9 (1985), 1–54.

Evenson, R.E. 1978. Time allocation in rural Philippine households. *American Journal of Agricultural Economics* 60, 322–30.

Fisher, R.A. 1935. *The Design of Experiments*. London: Oliver & Boyd.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Heckman, J.J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46, 931–59.

Huber, P.J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4, 221–33.

Hurd, M.D., McFadden, D., Chand, H., Gan, L., Merrill, A. and Roberts, M. 1998. Consumption and saving balances of the elderly: experimental evidence on survey response bias. In *Frontiers in the Economics of Aging*, ed. D. Wise. Chicago: University of Chicago Press.

Kaier, A.N. 1895. Observations et expériences concernant des dénombrements représentatifs. *Bulletin of the International Statistical Institute* 9 (Livre 2), 176–83.

Mahalanobis, P.C. 1940. Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society, A* 109, 329–78.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97, 558–625.

Rosenbaum P.R. and Rubin, D.B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Walker, T.S. and Ryan, J.G. 1990. *Village and Household Economies in India's Semi-Arid Tropics*. Baltimore, MD: Johns Hopkins University Press.

Yates, F. 1935. Complex experiments. *Journal of the Royal Statistical Society*, Suppl. 2, 181–247.

How to cite this article

Thomas, Duncan. "household surveys." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000160> doi:10.1057/9780230226203.0751

The New Palgrave Dictionary of Economics Online

housing policy in the United States

John M. Quigley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The most significant and most expensive housing policy in the United States is the treatment of owner-occupied housing for tax purposes. This treatment of housing under the tax code is analogous to that in many other countries (for example, Sweden), but certainly not in all developed countries (for example, Canada). Federal subsidies to US renter households are much smaller. Policy has evolved from programmes in which the government built, owned, and managed dwellings to programmes emphasizing housing demand through vouchers and rent certificates awarded to eligible households.

Keywords

Great Depression; homeownership; housing; housing costs; housing expenditures; housing externalities; housing finance; housing policy in the United States; housing subsidies; housing tax credit; income tax; inflation; interest rates; Internal Revenue Service (IRS); land-use regulation; rent control; social capital; tax expenditures; taxation of capital income

Article

Public concern over housing arises from three sources. First, housing is the single largest expenditure item in the budgets of families and individuals in most modern economies. The average household in western Europe and the United States devotes more than one quarter of its income to housing expenditures. Thus, increased efficiency in the provision of housing services or reduced occupancy costs can have a large impact on non-housing consumption and household well-being. Second, consumers' housing and location choices condition many other aspects of the quality of urban life. For example, the transport, schooling, and neighbourhood opportunities of urban households are themselves greatly affected by the housing opportunities available to them. Third, it is widely presumed that there are significant externalities in housing consumption. These external effects range all the way from the consequences of the social and physical isolation of those living in low-income residential

neighbourhoods to the presumed benefits of the ‘social capital’ and the increased political participation of households who own their homes.

In the United States, important policies providing subsidies to housing consumers are made by the central (‘federal’) government. Other policies governing housing – the regulation of house-building, service provision, and occupancy – are determined by local governments. At the national level, subsidies provided to selected housing consumers and producers are implemented by two government agencies: the Internal Revenue Service (IRS) and the Department of Housing and Urban Development (HUD). The policies administered by the IRS are clearly more important quantitatively, and they have large welfare effects.

The federal tax code

The IRS administers two housing subsidy programmes: the tax expenditures to owner-occupants for housing consumption specified in the personal income tax code, and the tax expenditures for builders of rental housing under the Low Income Housing Tax Credit programme specified in the Tax Reform Act of 1986. This latter programme is small, having originated in the Tax Reform Act of 1986. The former programme is large, and has existed in its current form since the personal income tax was established in 1915. Indeed, the benefits to homeowners under these tax policies are among the most generous in the developed world. (But the form of these subsidies is certainly not unique to the United States. See Englund, 2003, for a comparative discussion.)

Consider an individual who chooses between an investment in owner-occupied housing and an equivalent investment in some other asset – common stocks, say. The investment in owner-occupied housing offers three distinct tax advantages. First, under the US Internal Revenue Code, the returns on the investment in owner-occupied housing are untaxed (these returns are in the form of the housing services consumed in any year). In contrast, the dividends yielded by common stock are reported as income and are taxed in the year accrued. Second, capital gains arising from the housing investment can be deferred indefinitely. Moreover, a large capital gains exclusion is available to those over the age of 55. In contrast, capital gains in the stock market are taxed in the year they are realized. Third, some of the expenses associated with homeownership, notably property taxes and mortgage interest payments, can be itemized as deductions in computing federal tax liability under the personal income tax. No other interest payments are deductible as personal expenses under the Internal Revenue Code. This favourable treatment also extends to personal income taxation under the laws of all of the 50 states.

The net effect of these provisions of the US tax law is to reduce the price of homeownership, relative to renting, by a sizeable amount. Moreover, as a result of these policies, the relative price of homeownership varies by income level and the level of inflation.

It is useful to think of the price of homeownership as the cost of using the stock of residential capital. The rent R for using a unit of capital V is merely

$$R = iV, \tag{1}$$

where *i* is the real interest rate. *i* is simply the price of using a unit of capital *V* for a year. Housing is subject to local property tax at effective rate *t*. Annual expenditures of 100*d* per cent are required to maintain the property and to offset depreciation. The owner can expect real capital gains at a rate *g*. Let π be the rate of inflation. For housing, the user cost relationship is thus

$$R_1 = [(i + \pi) - t - d - (g + \pi)]\, V, \tag{2}$$

where the term in square brackets is the user cost of residential capital. Note that, in the absence of tax considerations, the user cost is insensitive to the level of inflation π. Now suppose nominal capital gains are untaxed and that mortgage interest payments and property taxes are deductible from gross income. Suppose net income is taxed at the rate of *T* per cent. Under these circumstances the user cost relationship is

$$R_2 = ([i + \pi]\,[1 - T] + t\,[1 - T] + d - [g + \pi])\, V, \tag{3}$$

or

$$R_2 = R_1 - T\,(i + \pi + t)\, V. \tag{4}$$

The system of taxes leads to a reduction in the net price of housing capital by the amount of the second term. Note that the after-tax cost of homeownership declines with the value of the house, the real interest rate, the property tax rate, and the marginal income tax rate.

If federal tax rates increase with income or if higher-income households live in jurisdictions with higher property tax rates, the cost of homeownership declines with income. More important, as long as housing is a normal good with a positive income elasticity, the net cost of homeownership declines with income. Furthermore, a given level of inflation in the economy reduces the user cost more for higher-income than for lower-income homeowners.

More generally, the analysis shows that the costs of homeownership are sensitive to macroeconomic stabilization policies and to the structure of income tax rates. The marginal tax rates of the highest-income US households fell from 70 per cent to 30 per cent and then rose to 40 per cent during the 1980s and 1990s, before falling again in 2001. At the same time, the inflation rate plummeted from 15 per cent

to less than three per cent. These changes have meant that the implicit policy toward housing and homeownership varied substantially.

For example, at reasonable values of the variables in eq. 4 (say, $i = g = 3\%$, $t = d = 2\%$, $T = 30\%$), then as inflation declines from six per cent to 1 per cent, the after-tax user cost of residential capital roughly doubles. Similarly, at reasonable values of the variables (for example, $\pi = 3\%$ and, as before, $i = g = 3\%$, $t = d = 2\%$), then, as income tax rates decrease from 40 per cent to 20 per cent, the after-tax cost of owner occupancy increases by more than one-third. These are substantial price changes induced entirely by taxation and macroeconomic considerations which may be completely unrelated to any objective of housing policy.

These reductions in the user cost of housing capital may be expected to increase housing consumption; reductions in the price of owning relative to renting may be expected to increase homeownership. But econometric research suggests that the demand for housing is moderately price-inelastic. It also appears, at least for the United States, that the elasticity of homeownership with respect of the relative price of homeownership is quite small. Thus, the effects of these large subsidies on housing outcomes are quite small.

In contrast, the magnitude of the implicit subsidy arising from the personal income tax code is large and extremely regressive. The subsidy is available only to owners, who are typically more affluent than renters, and only to those who find it advantageous to itemize their deductions in computing their tax liabilities. (Under US tax law, households may claim a ‘standard’ deduction for expenses or they may list deductions separately. The propensity to itemize deductions separately increases with income.) Finally, as noted above, for those owners who do itemize deductions, the magnitude of the subsidy increases with income.

The second programme administered by the IRS, the low-income housing tax credit, was established in 1986 and expanded in 2001. Under this programme, tax credits are remitted to each state in proportion to population. These credits are awarded by states to developers who propose new construction of housing reserved for low-income tenants who pay 30 per cent of their incomes in rent. The credits, in turn, are sold to firms and high-income individuals, and the proceeds are invested in the designated projects. The IRS monitors the compliance of these projects with the tax law requiring occupancy by low-income tenants for a 15-year period after construction.

The revenues forgone by the federal treasury as a result of these programmes are routinely estimated by the Joint Committee on Taxation of the Congress. The revenue costs of these subsidies are large. In 2005, for example, it is estimated that tax expenditures for owner-occupied housing totalled about $147 billion – $69 billion for the mortgage interest deduction, $33 billion for the capital gains exclusion on home sales, $28.6 billion for the exclusion of imputed rent, and $16.6 billion for the property tax deduction. It is estimated that more than half of the benefits of the tax expenditures for homeowners accrue to the top 15 per cent of the income distribution.

In contrast, in 2005 the tax expenditures arising from the low-income housing tax credit were about $4.8 billion (in present value terms). Presumably much of this benefit accrues to low-income renters.

A more relevant benchmark for the costs of these tax expenditures may be a comparison with the housing programmes managed by HUD, whose principal beneficiaries are low-income households. Direct expenditures under these programmes are currently $41 billion, or about 28 per cent of the tax expenditures on behalf of owner occupants.

Subsidies for renters

Federal housing policies for renters administered by HUD provide subsidies to about a third of low-income households. These programmes have evolved from those providing housing owned and managed by government to those providing direct cash assistance for deserving renters. The Public Housing Program was established in 1937 to subsidize local governments in building housing for those temporarily unemployed and also in providing construction jobs for unemployed urban labour during the Great Depression. Until the end of the 1970s, the programme subsidized virtually all of the capital costs of designated public housing dwellings and none of the operating costs. Since rent rolls were fixed at 25–30 per cent of tenant income, project managers who chose to serve households with the lowest incomes faced severe budgetary problems. Changes in the subsidy formulas helped local managers avoid this Hobson's choice, but the legacy of the original subsidy formula, the overcapitalization of projects to economize on maintenance expenses, is still manifest in the long-lived capital produced by the Public Housing Program.

The private sector was first induced to build, manage and provide rental dwellings for low-income tenants in the 1960s, through generous depreciation allowances provided to limited dividend corporations (under programmes such as Section 235 of the Housing Act of 1968). But it was not until 1974 that the subsidy provided to deserving tenants was divorced from the cost of supplying newly constructed housing.

The innovation in Section 8 of the Housing Act of 1974 was a programme of project-based housing assistance based upon long-term contracts in which the federal government guaranteed that participating landlords would receive the average rent in the local housing market (rather than the cost of building new housing). Low-income households pay 30 per cent of their incomes to a participating landlord and the difference, up to the ‘fair market rent’ in the housing market, is supplied under federal contract.

The radical departure to subsidize directly the demanders of low-income housing rather than the builders and suppliers of that housing was thoroughly tested by the Housing Allowance Experiments of the 1970s and 1980s, the most expensive social experiment in history, and the results were incorporated over time into the current Housing Choice Voucher Program which allocates vouchers or certificates to local authorities for distribution to low-income households. Under this programme, a qualifying household receives a voucher which pays the difference between 30 per cent of tenant income and the ‘fair market rent’. This programme is administered by Local Housing Authorities, who screen applicants and certify eligibility. Under current practice, households with incomes below 80 per cent of the area median income are eligible for vouchers, but three-quarters of the vouchers are reserved for very low-income households, those whose incomes are below 30 per cent of the area median income. In principle, the voucher is completely portable. It can be used anywhere by a recipient to enter into a rental contract within 90 days of issue.

Vouchers offer several clear advantages over the alternative supply oriented housing subsidy programmes. First, they are considerably cheaper per household served than programmes linking subsidies to construction costs, including the Public Housing Program, but also the Low Income Housing Tax Credit Program. Second, they remove questions about the location of dwellings occupied by low-income subsidized households from the local political process. Third, they preserve the anonymity of the low-income recipients of these subsidies. Fourth, they foster the spatial decentralization of the low-income population, reducing the concentration of disadvantaged households

in particular neighbourhoods. Fifth, they better facilitate the operation of the labour market by encouraging recipients to live closer to actual or potential worksites.

Although new commitments by HUD for subsidies to low-income renters are concentrated in the voucher programme, the legacy of past programmes will remain for a considerable period. For example, in the last year for which complete data are available (1998), 1.3 million units of government-owned public housing were used to provide housing subsidies, as were 1.0 million units of Section 8 project-based housing and 750,000 units of housing produced by other supply-oriented programmes. In contrast, 1.4 million households were subsidized by tenant-based voucher programmes.

Local housing regulations impose a potentially serious impediment to the efficiency of vouchers as a vehicle for housing subsidies. With local property taxes as the basis for local service provision, it is often in the fiscal interests of individual governments to limit the construction of new housing and to restrict the construction of high-density housing. The land-use regulations of individual jurisdictions are not well coordinated regionally in the United States, and the resulting regulatory pattern may make the housing supply relatively inelastic. This may lead to higher housing prices in response to increases in demand throughout the market, and it may mean that housing may be less available to voucher recipients in some metropolitan areas.

Despite these real concerns, the most important factor keeping the rent-to-income ratio of the poor high is the limited availability of housing subsidies. In 2001, it was estimated that almost 14.5 million renter households paid more than 30 per cent of their incomes on rent, and more than 7 million paid more than half of their incomes on rent. In contrast, only about 5 million renter households received subsidies from all federal government housing programmes.

See Also

- housing supply
- local public finance
- markets
- urban economics
- urban housing demand
- rent control
- residential real estate and finance

Bibliography

Englund, P. 2003. Taxing residential housing capital. *Urban Studies* 40, 937–52.

Gabriel, S. 1996. Urban housing policy in the 1990s. *Housing Policy Debate* 7, 673–93.

Quigley, J. 2000. A decent home: housing policy in perspective. *Brookings–Wharton Papers on Urban Affairs* 1(1), 53–100.

Quigley, J. and Raphael, S. 2004. Is housing unaffordable? Why isn't it more affordable? *Journal of Economic Perspectives* 18(1), 191–214.

How to cite this article

Quigley, John M. "housing policy in the United States." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000098> doi:10.1057/9780230226203.0752

The New Palgrave Dictionary of Economics Online

housing supply

Raven E. Saks

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the key factors that influence the elasticity of housing supply in the United States. When housing demand increases, the response of the housing stock is determined by physical construction costs (materials, labour and land) and government regulation. During the past several decades, a widespread adoption of restrictive land-use policies has substantially reduced the elasticity of housing supply in many parts of the United States. As the housing stock has become more inelastic, housing supply conditions have become progressively more important for understanding the dynamics of house prices and the form of urban growth and decline.

Keywords

construction costs; house prices; housing supply; income distribution; land-use regulation; public housing; real estate; urban economics; urban growth; zoning

Article

The supply of housing has exerted a growing influence on the dynamics of US housing markets since the 1970s. An increase in aggregate housing demand is ultimately met by an expansion of the housing stock somewhere in the United States, but the response of the local housing supply to a change in demand varies substantially across geographic locations. Some metropolitan areas, like Charlotte, NC, have grown rapidly with only moderate increases in house prices, suggesting that the supply of housing is elastic in these locations. By contrast, in locations like New York City, large increases in house prices and low levels of construction activity indicate a considerably more inelastic supply. Places experiencing persistent declines in housing demand, like Detroit, illustrate yet another aspect of the housing supply. The durability of housing prevents sharp contractions of the housing stock when housing demand falls, limiting population outflows from these locations and contributing to the persistence of urban decline. The heterogeneity of supply responses across local housing markets has become a topic of great interest among urban economists, particularly as the supply of housing has

become more inelastic in a growing number of areas in the United States.

Increases in supply

The response of the housing stock to an increase in demand is governed by the need for three elements: a physical structure, land, and government approval to put the structure on the land. The costs associated with each of these elements determine the extent to which increases in demand are accompanied by an expansion of the housing stock or by higher house prices. A combination of rising prices and declines in construction activity in many parts of the United States suggests that there has been a secular decline in the elasticity of housing supply since the 1970s. Low barriers to entry and exit and the absence of significant returns to scale combine to make the home-building industry fairly competitive, so that changes in the elasticity of housing supply mainly reflect the costs of the three component elements.

Structure construction costs

The technology of homebuilding has not changed dramatically since the first half of the 20th century, so the costs of building a housing structure are largely determined by the input prices of construction materials and labour. Although these costs account for the majority of new construction outlays, their importance has declined over time, and they have accounted for no more than 65 per cent of the total market value of residential real estate since the mid-1980s (Davis and Heathcote, 2005). Typically, labour makes up about two-thirds of these physical costs, and geographic variation in construction worker wages is the primary source of differences in construction costs across locations. The response of the housing supply to changes in physical structure construction costs is relatively elastic (Somerville, 1999b; Gyourko and Saiz, 2006), but increases in these costs cannot account for the entire decline in residential construction activity that has occurred during the past several decades (Glaeser, Gyourko and Saks, 2005a).

Land availability

The housing supply is also a function of the amount of land available for new residential construction. Topography, the existence of bodies of water, and the geologic composition of the land can all contribute to the difficulty of building new houses, reducing the elasticity of housing supply. In a sample of 45 large cities, Rose (1989) estimates that about 30 per cent of the variation in land prices across locations can be explained by natural restrictions on the supply of land. The availability of land is clearly important in explaining why some cities grow more quickly than others, but it is unlikely to be able to account for an inelastic housing supply in areas like Austin, TX. Moreover, places with a limited supply of land could expand the stock of housing by building taller structures. Even in places with little vacant land like Manhattan, many residential buildings are shorter than can be explained by the cost of an additional story (Glaeser, Gyourko and Saks, 2005b).

Government regulation

The third factor influencing the elasticity of housing supply involves the permission to build. Even when the costs of materials, labour and land are low enough to generate an incentive to expand the housing stock, government restrictions often prevent developers from building as many residential units as they would like. Local governments have regulated the placement of residential structures ever since the 1920s, when zoning laws began to separate residential land uses from commercial and industrial development. While these regulations altered the geographic distribution of residential structures within cities, initially they did not have a notable impact on the aggregate supply of urban housing (Fischel, 2004). It was not until the 1970s that municipalities began to enact growth controls and other exclusionary zoning practices designed to limit the absolute number of residential units in their jurisdiction. The popularity of these types of regulations has grown over the past several decades, and local governments now employ a wide range of regulatory practices including height and lot size restrictions, development moratoria, historic preservation rules and urban growth boundaries.

In contrast to these restrictive regulations, some government policies attempt to increase the supply of housing by providing tax incentives or subsidies to build units that will be affordable to low-income households. However, these policies do not have a notable impact on the aggregate stock of housing, as they mostly substitute for unsubsidized housing units (Malpezzi and Vandell, 2002). Federally owned housing appears to be less substitutable for private units, but there has been virtually no new construction of public housing units since the early 1980s (Green and Malpezzi, 2003).

Because land-use regulations are enacted by local governments and are frequently customized to meet the needs of individual neighbourhoods, these laws vary substantially across locations in both form and severity. This heterogeneity makes the degree of regulation difficult to classify in a manner that lends itself well to systematic empirical analysis. Despite this complexity, most empirical research has found a strong correlation of land-use regulation with higher house prices and less residential construction (Malpezzi, 1996; Mayer and Somerville, 2000; Saks, 2005). Thus, these regulations appear to reduce the elasticity of housing supply in the areas in which they are enacted.

As the number of municipalities with restrictive residential land-use policies has expanded, researchers have become progressively more interested in trying to understand the political economy of these regulations. Recent decades contrast sharply with the regulatory environment during the 1950s and 1960s, when builders were generally able to influence the decisions of local zoning boards (Molotch, 1976). Since that time, homeowners have become more successful at restricting residential construction in their neighbourhoods. The incentive of homeowners to constrain development has been linked to several motivations including the reduction of congestion costs, the preservation of local amenities (Hilber and Robert-Nicoud, 2006), insurance against shocks to household wealth (Ortalo-Magne and Prat, 2007), the reduction of free-riding on the provision of public goods (Fischel, 2001), and the growing likelihood that homeowners work in a different jurisdiction from their place of residence (Fischel, 2001). In addition to changes in homeowners' incentive to limit new construction, the rise of regulation may also be a function of their improved ability to influence the political process (Glaeser, Gyourko and Saks, 2005a).

While theories explaining the existence of supply restrictions have multiplied, empirical evidence on the determinants of zoning remains thin. Richer towns with more educated populations exhibit a higher propensity to restrict residential development (Evenson and Wheaton, 2003; Glaeser, Gyourko and Saks,

2005a), and cities are more likely to enact land-use regulations when the policies of neighbouring municipalities are also restrictive (Brueckner, 1998). However, these studies are based on cross-sectional evidence, making it difficult to distinguish causal mechanisms from location-specific characteristics and geographic differences in housing demand.

Decreases in supply

The durability of housing structures means that the elasticity of housing supply is asymmetric in response to increases versus decreases in demand. Because housing depreciates slowly, the housing stock does not contract immediately in response to a decline in housing demand. Instead, places experiencing persistent declines in housing demand have low house prices relative to construction costs. The availability of cheap housing encourages households to remain in declining cities rather than moving to a location with growing labour demand. Thus, urban decline is slow and highly persistent (Glaeser and Gyourko, 2005). The durability of housing may also influence urban growth through its impact on local land-use planning decisions (Turnbull, 2006).

Broader consequences of the elasticity of housing supply

The effects of the housing supply extend far beyond changes in the relative distribution of house prices and city sizes across the United States. For example, by restricting the number of households in a location, the housing supply can limit the supply of workers, altering the dynamics of local wage and employment growth (Case, 1991; Saks, 2005). Aggregate economic activity may also be reduced as workers are prevented from living in the location where they would be most productive. The housing supply also affects the distribution of income across and within cities. By altering relative house-price differentials, supply restrictions will cause high-income households to sort into metropolitan areas with highly valued amenities (Gyourko, Mayer and Sinai, 2006). Moreover, the composition of the population within metropolitan areas will also depend on the elasticity of housing supply, as demographic groups with a higher propensity to move relocate in response to rising house prices. While this article has focused on the United States, the underlying forces that shape housing supply conditions are similar around the world. Housing investment as a share of GDP in the United States has been around the median of other OECD countries since the late 1990s. In some countries, construction activity is lower than in the United States due to a greater scarcity of land and more restrictive land-use regulations. By contrast, some other developed countries have higher rates of housing investment due to a more active government role in subsidizing residential construction (Ball, 2003). Given the widespread reductions in the elasticity of housing supply in many parts of the United States during the past few decades, further investigations into the determinants and implications of housing supply conditions promise to be an important direction of future research in both urban economics and macroeconomics.

See Also

- housing policy in the United States

- low-income housing policy
- residential real estate and finance
- urban economics
- urban growth
- urban housing demand

Bibliography

Ball, M. 2003. Markets and the structure of the housebuilding industry: an international perspective. *Urban Studies* 40, 897–916.

Brueckner, J.K. 1998. Testing for strategic interaction among local governments: the case of growth controls. *Journal of Urban Economics* 44, 438–67.

Case, K.E. 1991. The real estate cycle and the economy: consequences of the Massachusetts boom of 1984–87. *New England Economic Review*, (September), 37–46.

Davis, M. and Heathcote, J. 2005. The price and quantity of residential land in the United States. Discussion Paper No. 5333, Center for Economic Policy Research.

Evenson, B. and Wheaton, W.C. 2003. Local variation in land use restrictions. *Brookings-Wharton Papers on Urban Affairs* 2003, 221–50.

Fischel, W. 2001. *The Homevoter Hypothesis: How Home Values Influence Local Government Taxation, School Finance, and Land Use Policies*. Cambridge, MA: Harvard University Press.

Fischel, W. 2004. An economic history of zoning and a cure for its exclusionary effects. *Urban Studies* 41, 317–40.

Glaeser, E.L. and Gyourko, J. 2005. Urban decline and durable housing. *Journal of Political Economy* 113, 345–75.

Glaeser, E.L. Gyourko, J. and Saks, R.E. 2005a. Why have house prices gone up? *American Economic Review Papers and Proceedings* 95, 329–33.

Glaeser, E.L., Gyourko, J. and Saks, R.E. 2005b. Why is Manhattan so expensive? Regulation and the rise in house prices. *Journal of Law and Economics* 48, 331–69.

Green, R.K. and Malpezzi, S. 2003. *A Primer on U.S. Housing Markets and Housing Policy*. Washington, DC: Urban Institute Press.

Gyourko, J., Mayer, C. and Sinai, T. 2006. Superstar cities. Working Paper No. 12355. Cambridge, MA: NBER.

Gyourko, J. and Saiz, A. 2006. Construction costs and the supply of housing structure. *Journal of Regional Science* 46, 661–80.

Hilber, C. and Robert-Nicoud, F. 2006. Owners of developed land versus owners of undeveloped land: why land use is more constrained in the bay area than in Pittsburgh. Discussion Paper No. 870, Centre for Economic Policy Research.

Malpezzi, S. 1996. Housing prices, externalities, and regulation in U.S. metropolitan areas. *Journal of Housing Research* 7, 209–41.

Malpezzi, S. and Vandell, K. 2002. Does the low-income housing tax credit increase the supply of housing? *Journal of Housing Economics* 11, 360–80.

Mayer, C.J. and Somerville, C.T. 2000. Land use regulation and new construction. *Regional Science and Urban Economics* 30, 639–62.

Molotch, H. 1976. The city as a growth machine. *American Journal of Sociology* 82, 309–30.

Ortalo-Magne, F. and Prat, A. 2007. The political economy of housing supply: homeowners, workers and voters. Discussion Paper No. TE/2007/514, Suntory-Toyota International Centers for Economics and Related Disciplines.

Rose, L. 1989. Urban land supply: natural and contrived restrictions. *Journal of Urban Economics* 25, 325–45.

Saks, R.E. 2005. Job creation and housing construction: constraints on metropolitan area employment growth. Finance and Economics Discussion Series 49, Board of Governors of the Federal Reserve System (U.S.)

Somerville, C.T. 1999a. The industrial organization of housing supply: market activity, land supply and the size of homebuilder firms. *Real Estate Economics* 27, 669–94.

Somerville, C.T. 1999b. Residential construction costs and the supply of new housing: endogeneity and bias in construction cost indexes. *Journal of Real Estate Finance and Economics* 18, 43–62.

Turnbull, G.K. 2006. The investment incentive effects of land use regulations. *Journal of Real Estate Finance* 31, 357–95.

How to cite this article

Saks, Raven E. "housing supply." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000161> doi:10.1057/9780230226203.0753

The New Palgrave Dictionary of Economics Online

human capital, fertility and growth

Oded Galor

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The worldwide demographic transition of the past 140 years has been identified as one of the prime forces in the transition from stagnation to growth. The unprecedented increase in population growth during the early stages of industrialization was ultimately reversed. The rise in the demand for human capital in the second phase of industrialization brought about a significant reduction in fertility rates and population growth in various regions of the world, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita.

Keywords

Agricultural Revolution; child labour; child quality; consumption smoothing; credit constraints; demographic transition; economic growth; educational attainment; fertility; gender gap; growth take-offs; human capital; Industrial Revolution; industrialization; infant mortality; life expectancy; mortality; old-age security hypothesis; population growth; public education; technological progress; wage differentials

Article

The transition from stagnation to growth has been the subject of intensive research in recent years. The rise in the demand for human capital and the associated decline in population growth have been identified as the prime forces in the movement from an epoch of stagnation to a state of sustained economic growth. They have brought about a significant formation of human capital along with a reduction in fertility rates and population growth, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita.

Historical evidence

The evolution of economies throughout human history has been characterized by Malthusian stagnation. Technological progress and population growth were minuscule by modern standards, and the average growth rate of income per capita was even slower, due to the offsetting effect of population growth on the expansion of resources per capita. In the past two centuries, on the other hand, the pace of technological progress increased significantly, alongside the process of industrialization. Various regions of the world departed from the Malthusian trap and initially experienced a considerable rise in the growth rates of income per capita and population. In contrast to episodes of technological progress in the pre-Industrial Revolution era, which failed to generate sustained economic growth, the increasing role of human capital in the production process in the second phase of the Industrial Revolution ultimately prompted a demographic transition, liberating the gains in productivity from the counterbalancing effects of population growth. The decline in population growth and the associated advancement in technological progress and human capital formation paved the way for the emergence of the modern state of sustained economic growth.

The evolution of population growth in the world economy has been non-monotonic. The growth of world population was sluggish during the Malthusian epoch, creeping at an average annual rate of about 0.1 per cent over the years 0–1820 (Maddison, 2001). The Western European take-off along with that of the Western Offshoots (that is, the United States, Canada, Australia and New Zealand) brought about a sharp increase in population growth in these regions. The world annual average rate of population growth increased gradually reaching 0.8 per cent in the years 1870–1913. The take-off of less developed regions and the significant increase in their income per capita generated a further increase in the world rate of population growth, despite the decline in population growth in Western Europe and the Western Offshoots, reaching a high level of 1.92 per cent per year in the period 1950–73. Ultimately, the onset of the demographic transition in less developed economies in the second half of the 20th century, reduced population growth to an average rate of 1.63 per cent per year in the period 1973–98.

The timing of the demographic transition differed significantly across regions. A reduction in population growth occurred in Western Europe, the Western Offshoots, and Eastern Europe towards the end of the 19th century and in the beginning of the 20th century, whereas Latin America and Asia experienced a decline in the rate of population growth only in the last decades of the 20th century.

The demographic transition in Western Europe occurred towards the turn of the 19th century. A sharp reduction in fertility took place simultaneously in several countries in the 1870s, and resulted in a more than 30 per cent decline in fertility rates within a 50-year period. Over the period 1875–1920, crude birth rates declined by 44 per cent in England, 37 per cent in Germany, and 32 per cent in Sweden and Finland. A decline in mortality rates preceded the decline in fertility rates in most of Western Europe. It began in England nearly 140 years prior to the decline in fertility, and in Sweden and Finland the corresponding figure was 100 years. The decline in fertility outpaced the decline in mortality rates and brought about a decline in the number of children who survived to their reproduction age.

A similar pattern characterizes mortality and fertility decline in less developed regions. The total fertility rate over the period 1960–99 plummeted from 6 to 2.7 in Latin America, from 6.14 to 3.14 in Asia, and declined moderately from 6.55 to 5 in Africa, along with a sharp decline in infant mortality rates.

Theories of the demographic transition

The decline in infant and child mortality

The decline in infant and child mortality rates has been a dominating explanation for the onset of the decline in fertility in many developed countries, with the notable exceptions of France and the United States. Nevertheless, this viewpoint appears inconsistent with historical evidence. While it is highly plausible that mortality rates were among the factors that affected the level of fertility throughout human history, historical evidence does not lend credence to the argument that the decline in mortality rates accounts for the reversal of the positive historical trend between income and fertility.

The mortality decline in Western Europe started nearly a century before the decline in fertility and was associated initially with increasing fertility rates in some countries and non-decreasing fertility rates in others. In particular, the decline in mortality started in England in the 1730s, and until 1820 was accompanied by a steady increase in fertility rates. The significant rise in income per capita in the post-Malthusian regime apparently increased the desirable number of surviving offspring and thus, despite the decline in mortality rates, fertility increased significantly so as to reach this higher desirable level. The decline in fertility during the demographic transition occurred in a period in which this pattern of increased income per capita (and its potential effect on fertility) was intensified, while the pattern of declining mortality (and its adverse effect on fertility) maintained the trend that existed in the 140 years preceding the demographic transition. The reversal in fertility patterns in England and in other Western European countries in the 1870s suggests therefore that the demographic transition was not prompted by a decline in infant and child mortality.

Furthermore, most relevant from an economic point of view is the cause of the reduction in net fertility (that is, the number of children reaching adulthood). The decline in the number of surviving offspring that was observed during the demographic transition is unlikely to have been a result of mortality decline. Mortality decline would have led to a reduction in the number of surviving offspring if the following implausible conditions had been met: (*a*) there existed a precautionary demand for children, that is, individuals were risk averse with respect to the number of surviving offspring; (*b*) risk aversion with respect to consumption was smaller than risk aversion with respect to fertility (evolutionary theory would suggest the opposite); (*c*) sequential fertility (that is, replacement of non-surviving children) was modest.

The rise in the level of income per capita

The rise in income per capita prior to the demographic transition has led some researchers to argue that the demographic transition was triggered by the asymmetric effects of the rise in income per capita on household income and on the opportunity cost of bringing up children. Becker (1981) argues that the rise in income induced a fertility decline because the positive income effect on fertility was dominated by the negative substitution effect that was brought about by the rising opportunity cost of children. Similarly, he argues that the income elasticity with respect to child quality is greater than that with respect to child quantity, and hence a rise in income led to a decline in fertility along with a rise in the investment in each child.

This theory suggests that the timing of the demographic transition across countries in similar stages of development would reflect differences in income per capita. However, remarkably, the decline in fertility occurred in the same decade across Western European countries despite their differing significantly in their income per capita. In 1870, on the eve of the demographic transition, England was

the richest country in the world, with a GDP per capita of 3,191 dollars (measured in 1990 international dollars: Maddison, 2001). In contrast, Germany, which experienced the decline in fertility in the same years as England, had in 1870 a GDP per capita of only 1,821 dollars (that is, 57 per cent of that of England). Sweden's GDP per capita of 1,664 dollars in 1870 was 48 per cent of that of England, and Finland's GDP per capita of 1,140 dollars in 1870 was only 36 per cent of that of England, but their demographic transitions occurred in the same decade. The simultaneity of the demographic transition across Western European countries that differed significantly in their income per capita suggests that the high level of income reached by Western Europeans countries in the post-Malthusian regime had a very limited role in the demographic transition.

The rise in the demand for human capital

The gradual rise in the demand for human capital in the second phase of the Industrial Revolution (and in the process of industrialization of less developed economies) and its close association with the timing of the demographic transitions has led researchers to argue that the increasing role of human capital in the production process induced households to increase investment in the human capital of their offspring, ultimately leading to the onset of the demographic transition.

Galor and Weil (1999; 2000), argue that the acceleration in the rate of technological progress gradually increased the demand for human capital in the second phase of the Industrial Revolution, inducing parents to invest in the human capital of their offspring. The increase in the rate of technological progress and the associated increase in the demand for human capital brought about two effects on population growth. On the one hand, improved technology eased households’ budget constraints and provided more resources for the quality as well as the quantity of children. On the other hand, it induced a reallocation of these increased resources towards child quality. In the early stages of the transition from the Malthusian regime, the effect of technological progress on parental income dominated, and the population growth rate as well as the average quality increased. Ultimately, further increases in the rate of technological progress, stimulated by human capital accumulation, induced a reduction in fertility rates, generating a demographic transition in which the rate of population growth declined along with an increase in the average level of education. Thus, consistent with historical evidence, the theory suggests that prior to the demographic transition, population growth increased along with investment in human capital, whereas the demographic transition brought about a decline in population growth along with a further increase in human capital formation.

Galor and Weil's theory suggests that a universal acceleration in technological progress raised the demand for human capital in the second phase of the Industrial Revolution and generated a simultaneous increase in educational attainment and demographic transition across Western European countries that differed significantly in their levels of income per capita. Consistent with the theory, the growth rates (as opposed to the levels) of income per capita among these Western European countries were rather similar during their demographic transition, ranging from 1.9 per cent per year over the period 1870–1913 in the UK, 2.12 per cent in Norway, 2.17 per cent in Sweden, to 2.87 per cent in Germany. Moreover, the demographic transition in England was associated with a significant increase in the investment in child quality as reflected by years of schooling. Moreover, international trade and its differential effects on the demand for human capital had an asymmetric effect of the timing of the demographic transition (Galor

and Mountford, 2006).

Evidence about the evolution of the return to human capital over this period is scarce and controversial, but it does not indicate that the skill premium increased markedly in Europe over the course of the 19th century, nor is it an indication of the absence of a significant increase in the demand for human capital. Technological progress in the second phase of the Industrial Revolution brought about an increase in the demand for human capital, and indeed, in the absence of a supply response, one would have expected an increase in the return to human capital. However, the significant increase in schooling in the 19th century, and in particular the introduction of publicly provided education, which lowered the cost of education, generated a significant increase in the supply of educated workers. Some of this supply response was a direct reaction to the increase in the demand for human capital, and thus may only operate to partially offset the increase in the return to human capital. However, the removal of the adverse effect of credit constraints on the acquisition of human capital (for example, Galor and Zeira, 1993 and Galor and Moav, 2006), as reflected by the introduction of publicly provided education, generated an additional force that increased the supply of educated labour and operated towards a reduction in the return to human capital.

The decline in child labour

The effect of the rise in the demand for human capital on the reduction in the desirable number of surviving offspring was magnified via its adverse effect on child labour. It gradually increased the wage differential between parental labour and child labour, inducing parents to reduce the number of their children and to further invest in their quality (Hazan and Berdugo, 2002). Moreover, the rise in the importance of human capital in the production process induced industrialists to support education reforms (Galor and Moav, 2006) and thus laws that abolished child labour (Doepke, 2004; Doepke and Zilibotti, 2005), and thus fertility.

The rise in life expectancy

The impact of the increase in the demand for human capital on the decline in the desirable number of surviving offspring was reinforced by improvements in health and life expectancy. Despite the gradual rise in life expectancy prior to the demographic transition, investment in human capital was insignificant as long as a technological demand for human capital had not emerged. The technologically based rise in the demand for human capital during the second phase of the Industrial Revolution and the rise in the expected length of productive life increased the potential rate of return to investments in children's human capital, reinforcing the inducement for investment in education and the associated reduction in fertility rates (Galor and Weil, 1999; Moav, 2005; Soares, 2005).

Natural selection and the evolution of preference for offspring's quality

The impact of the increase in the demand for human capital on the decline in the desirable number of surviving offspring may have been magnified by cultural or genetic evolution in the attitude of individuals towards child quality. Galor and Moav (2002) propose that during the epoch of Malthusian

stagnation that characterized most of human existence, individuals with a higher valuation for offspring quality (in the context of the quantity-quality survival strategies) gained an evolutionary advantage and their representation in the population gradually increased. The Agricultural Revolution facilitated the division of labour and fostered trade relationships across individuals and communities, enhancing the complexity of human interaction and raising the return to human capital. Moreover, the evolution of the human brain in the transition to *Homo sapiens* and the complementarity between brain capacity and the reward for human capital has increased the evolutionary optimal investment in the quality of offspring. The distribution of valuation for quality lagged behind the evolutionary optimal level and individuals with traits of higher valuation for their offspring's quality generated higher income and, in the Malthusian epoch, a higher number of offspring. Thus, the trait of higher valuation for quality gained the evolutionary advantage. This evolutionary process was reinforced by its interaction with economic forces. As the fraction of individuals with high valuation for quality increased, technological progress intensified, raising the rate of return to human capital. The increase in the rate of return to human capital along with the increase in the bias towards quality in the population reinforced the substitution towards child quality, setting the stage for a more rapid decline in fertility along with a significant increase in investment in human capital and a transition to sustained economic growth.

The decline in the gender gap

The rise in the demand for human capital and its impact on the decline in the gender gap in the last two centuries could have reinforced a demographic transition and human capital formation. Galor and Weil (1996; 1999) argue that technological progress and capital accumulation complemented mental-intensive tasks and substituted for physical-intensive tasks in industrial production. In light of the comparative physiological advantage of men in physical-intensive tasks and women in mental-intensive tasks, the demand for women's labour input gradually increased in the industrial sector, decreasing monotonically the wage differential between men and women. In early stages of industrialization, the wages of both men and women increased, but the rise in women's wages was not sufficient to induce a significant increase in the female labour force. Fertility, therefore, increased due to the income effect that was generated by the rise in men's absolute wages. Ultimately, however, the rise in women's relative wages was sufficient to induce a significant increase in labour force participation. It increased the cost of bringing up children proportionally more than household income, generating a decline in fertility and a shift from stagnation to growth.

The old-age security hypothesis

The old-age security hypothesis (Caldwell, 1976) has been proposed as an additional mechanism for the onset of the demographic transition. It suggests that in the absence of capital markets that permit intertemporal lending and borrowing, children are assets that permit parents to smooth consumption over their lifetime. The process of development and the establishment of capital markets reduce this motivation for bringing up children, contributing to the demographic transition. The significance of the decline in the role of children as assets in the onset of the demographic transition is questionable. The rise in fertility rates prior to the demographic transition, in a period of improvements in the credit

markets, raises doubts about the significance of the mechanism. Furthermore, cross-section evidence (Clark and Hamilton, 2006) from the pre-demographic transition era indicates that wealthier individuals, who presumably had better access to credit markets, had a larger number of surviving offspring.

Concluding remarks

The rise in the demand for human capital in the second phase of industrialization and its effect on decline in population growth have been among the prime forces in the transition of economies from an epoch of stagnation to a state of sustained economic growth. They brought about a significant formation of human capital along with a reduction in fertility rates and population growth, enabling economies to advance technologically and to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita.

See Also

- demographic transition
- economic growth in the very long run
- growth take-offs

Bibliography

Becker, G.S. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Caldwell, W.J. 1976. Toward a restatement of demographic transition theory. *Population and Development Review* 2, 321–66.

Clark, G. and Hamilton, G. 2006. Survival of the richest. *Journal of Economic History* 66, 707–36.

Doepke, M. 2004. Accounting for fertility decline during the transition to growth. *Journal of Economic Growth* 9, 347–83.

Doepke, M. and Zilibotti, F. 2005. The macroeconomics of child labor regulation. *American Economic Review* 95, 1492–524.

Galor, O. and Moav, O. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117, 1133–92.

Galor, O. and Moav, O. 2006. Das human-kapital: a theory of the demise of the class structure. *Review of Economic Studies* 73, 85–117.

Galor, O. and Mountford, A. 2006. Trade and the great divergence: the family connection. *American*

Economic Review 96, 299–303.

Galor, O. and Weil, D.N. 1996. The gender gap, fertility, and growth. *American Economic Review* 86, 374–87.

Galor, O. and Weil, D.N. 1999. From Malthusian stagnation to modern growth. *American Economic Review* 89, 150–4.

Galor, O. and Weil, D.N. 2000. Population, technology and growth: from the Malthusian regime to the demographic transition and beyond. *American Economic Review* 110, 806–28.

Galor, O. and Zeira, J. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60, 35–52.

Hazan, M. and Berdugo, B. 2002. Child labor, fertility and economic growth. *Economic Journal* 112, 810–28.

Lagelof, N. 2003. Gender equality and long-run growth. *Journal of Economic Growth* 8, 403–26.

Maddison, A. 2001. *The World Economy: A Millennia Perspective*. Paris: OECD.

Moav, O. 2005. Cheap children and the persistence of poverty. *Economic Journal* 115, 88–110.

Soares, R.R. 2005. Mortality reductions, educational attainment, and fertility choice. *American Economic Review* 95, 580–601.

How to cite this article

Galor, Oded. "human capital, fertility and growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000166> doi:10.1057/9780230226203.0755

The New Palgrave Dictionary of Economics Online

human capital

Sherwin Rosen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Human capital refers to the productive capacities of human beings as income producing agents in the economy. The concept is an ancient one, but the use of the term in professional discourse has gained currency only in the past twenty-five years. During that period much progress has been made in extending the principles of capital theory to human agents of production. Capital is a stock which has value as a source of current and future flows of output and income. Human capital is the stock of skills and productive knowledge embodied in people. The yield or return on human capital investments lies in enhancing a person's skills and earning power, and in increasing the efficiency of economic decision-making both within and without the market economy. This account sketches the main ideas, and the bibliography is necessarily restrictive. For additional detail and alternative interpretations, the reader should consult the surveys by Blaug, Rosen, Sahota and Willis, which also present complete bibliographies.

Differences in form between human and non-human capital are of less import for analysis than are differences in the nature of property rights between them. Ownership of human capital in a free society is restricted to the person in whom it is embodied. By and large a person cannot, even voluntarily, sell a legally binding claim on future earning power. For this reason the exchange of human capital services is best analysed as a rental market transaction. Quantitative analysis is restricted to the income and output flows that result from human capital investments: wage payments and earnings flows are viewed as the equivalent of rentals of human capital value, because a person cannot sell asset claims in himself. Even the long-term commitments found in enduring employment relationships are best viewed as a sequence of short-term, renewable rental contracts. By contrast, the legal system places many fewer restrictions on the sale and voluntary transfer of title to non-human capital. In fact, substantial activity on non-human capital asset markets is a hallmark of an enterprise system of organization.

Flexibility must be maintained, however, in these distinctions, which are not always hard and fast. The institution of slavery was the primary example of a transferable property right in human capital. To be sure, the involuntary elements of slavery are essential, but even voluntary systems have not been unknown. Similarly, indentured servitude was an example of a legally enforceable long-term contractual claim on the human capital services of others. And in many societies today there are severe legal restrictions on transfer of title to non-human capital: the chief example is collective and state ownership

of non-human capital in planned economies.

Background

Classical economics maintained a tripartite distinction among the factors of production, Land, Labour and Capital; whereas modern economics is much less rigid in these divisions. Viewed from the perspective of supply, factors of production, whatever their form, can be increased and improved at some cost. To the extent that these improvements involve weighing future benefits against current costs, the principles of capital theory are applicable.

William Petty, the early actuary and national income accountant, is generally credited with the first serious application of the concept of human capital, when in 1676 he compared the loss of armaments, machinery and other instruments of warfare with the loss of human life. Elements of such comparisons survive to the present day. However, Adam Smith set the subject on its main course. *The Wealth of Nations* identified the improvement of workers' skills as a fundamental source of economic progress and increasing economic welfare. It also contained the first demonstration of how investments in human capital and labour market skills affect personal incomes and the structure of wages. Alfred Marshall stressed the long-term nature of human capital investments, and the role of the family in undertaking them. He also pointed out that non-monetary considerations would play a unique role in these decisions because of the dual nature of workers as factors of production and as consumers of their work environments. The distinguished actuary and scientist Alfred Lotka provided the first quantitative application of human capital in collaboration with Dublin, calculating the present value of a person's earnings to serve as guidelines for the rational purchase of life insurance. J.R. Walsh made the first cost imputation of human capital value. Frank Knight focused upon the role of improvements in society's stock of productive knowledge in overcoming the law of diminishing returns in a growing economy. These early contributions stand as landmarks. However, the impetus for rapid progress in this area came from the quantitative revolution in economics after World War II, when extensive data sources revealed certain systematic regularities. The first of these stems from economists' interest in understanding the nature and sources of economic growth and development in the 1950s and 1960s. Detailed calculations by national income accountants showed that conventional aggregate output measures grow at a more rapid pace than aggregate measures of factor inputs. A fundamental conservation law in economics would be violated unless the unexplained 'residual' was identified with (unexplained) technical change. Research associated with T.W. Schultz and Edward Denison attributed much of the measured residual to improvements in factor inputs. Schultz adopted an all-inclusive concept of human capital. At its heart lay secular improvements in workers' skills based on education, training and literacy; but he also pointed to sources of progress in improved health and longevity, the reduction in child mortality and greater resources devoted to children in the home, and the capacity of a more educated population to make more intelligent and efficient economic calculations. John Kendrick systematically pursued the empirical implications of these ideas and demonstrated that the rate of return on these inclusive human capital investments is of comparable magnitude to yields on non-human capital. This line of research as a whole proves that an investment framework is of substantial practical value in accounting for many of the sources of secular economic growth.

Another parallel strand of development arose from professional interest in the nature and determinants of the personal distribution of income and earnings. This problem was propelled, in addition, by

substantial public interest in the problem of poverty and prospects for redistributing resources to the poor. Empirical bases for this inquiry were, and continue to be, supported by extensive personal survey instruments (such as Census and allied records) that have become widely available in the post-war period. Much of this work has focused on the role of education and training as important determinants of personal wealth and income. Herman Miller's updating and elaboration of Dublin and Lotka's calculation found a strong and systematic relationship between education and personal economic success, a finding that has been replicated many times in virtually every country where data are available to make the calculations.

The fundamental conceptual framework of analysis for virtually all subsequent work in this area was provided by Gary Becker, who not only organized the emerging empirical observations but also provided a systematic method for seeking new results and implications of the theory. Practically every idea in his book has been pursued at length in the research of the past two decades. Following Schultz's lead, Becker organized his theoretical development around the rate of return on investment, as calculated by comparing the earnings streams in discounted present value on alternative courses of actions. Rational agents pursue investments up to the point where the marginal rate of return equals the opportunity cost of funds. Hence, conditional on the sources of financing investments through the market and family resources, there is a tendency for rates of return to be equated at the margin. This theory of *supply* of human capital implies empirically refutable restrictions on intertemporal and interpersonal differences in the patterns of earnings and other aspects of productivity. In focusing on the development of a person's skills and earning capacity over the life cycle, human capital theory has evolved as a theory of ‘permanent income’ and wealth.

Becker also made a distinction between human capital that is specific to its current employment in a firm, and that which has more general value over a broader set of employments. The concept of firm-specific capital is closely allied with organizational capital, a person's contribution to a specific organization, the value of which is lost and must be reproduced by costly investment when the employment relationship is terminated. General human capital represents skills that are not specifically tied to a single firm and whose employment can be transferred from one firm to another without significant loss of value. This distinction has proved valuable for analysing the determinants of turnover and firm-worker attachments and its ramifications are still being pursued. For example, the concept of firm-specific capital underlies the transactions cost basis for recent research on labour market and other contracts.

The rate of return

The connection between the rate of return on investment in human capital and observable earnings is illustrated by Smith's discussion of the relative earnings of physicians and other professional workers. A person who contemplates entering one of these fields must look forward to a long period of training and costly personal investment before any income is forthcoming. Furthermore, the long training period cuts into the period of actual practice and reduces the period of positive earnings. Consequently earnings must *compensate* for the cost and effort required to practice the trade: if they did not, fewer people would find it attractive to enter.

The compensatory nature of earnings on prior investments, equivalent to a rate of return, is the fundamental insight of human capital theory. First, it points to the opportunities foregone by an action as

a fundamental cost of undertaking it. Thus the direct tuition and other costs of education are only one component of the true cost. The fact that the person defers entering the market and gives up a current source of earnings is also properly counted as a cost. Second, the focus on the intertemporal and life-cycle nature of these decisions leads to a much different concept of income and inequality than simply examining current earnings. Human capital theory suggests that the distributions of *lifetime earnings* and human capital *wealth* are the keys to analysing the distribution of economic welfare, because earnings are the result of prior investments.

Two methods are widely used to calculate the return on human capital investments. Consider one alternative, call it the null alternative, which yields an earnings flow of $x_0(t)$. Consider another alternative, call it the investment alternative, which yields an earnings flow of $x_1(t)$. For example, in the leading case $x_0(t)$ is the expected flow of earnings in year t if one terminates education after high school graduation and $x_1(t)$ is the earnings that can be expected if one continues on to college. The time index t commences as of high school graduation, so $x_1(t)$ will typically show a phase (during the period of college attendance) of much smaller values than does $x_0(t)$. However, in later life $x_1(t)$ is generally larger than $x_0(t)$. This is precisely the investment content of the decision to continue school: there is a current cost in terms of income foregone, but a deferred benefit in terms of greater earnings prospects in the future. Write the difference $z(t)=x_1(t)-x_0(t)$. Then $z(t)$ shows a systematic pattern of negative values when t is small and positive values when t is large; $z(t)$ is increasing from negative to positive in between. Observed earnings in the two choices allows calculation of the internal rate of return, defined as the rate of interest which equates the present discounted value of the two earnings streams. If i is the internal rate, then $\sum z(t)/(1+t)'=0$.

Of course, it is not possible to observe earnings in the path not taken. A person either stops school or continues on to the next level. In practice, the calculation is made by using observed average earnings of college graduates at different ages as an estimate of $x_1(t)$ and using the observed average earnings of high school graduates as an estimate of $x_0(t)$. The typical calculation produces an estimate of i in the neighbourhood of ten per cent, comparable to the rate of return on investment in physical capital. Hanoch presents the most complete treatment of this problem. Remarkably, rates of return on education in the vicinity of ten per cent are found in a wide variety of countries and economic institutions.

Another method of calculation, first presented by Jacob Mincer, brings out the economic aspects of these estimates more clearly. Suppose a person contemplates a level income in amount $y(s)$ over the life work-life cycle if s years of schooling are undertaken. If schooling is productive we must have that $y'(s)=dy/ds$ is positive, that is, anticipated earnings must be increasing in years of schooling. The present discounted value of wealth associated with some choice s, from the point of view of the present time, is simply

$$W(s) = y(s)\int_s^n e^{-rt}dt,$$

where the index of integration runs from s, the time the person completes school and enters the market,

to n, the time the person retires. Since n is large, we may take the approximation

$$W(s) = y(s)\int_s^{\infty} e^{-rt}dt = y(s)\,e^{-rs}/r.$$

Assume that the schooling decision is made to maximize human capital wealth $W(s)$. Then differentiating with respect to s, the first order condition is $[y'(s)-ry(s)]\,e^{-rs}=0$, or $y'(s)/y(s)=r\cdot y'/y$ is nothing other than the marginal internal rate of return on investment in schooling, so schooling is chosen such that its marginal internal rate equals the rate of interest. This rule, similar to the economic problem of when to cut a tree or uncork the wine, is one that maximizes lifetime consumption prospects for the person.

Now extend this argument to many people. In an economy with many similar individuals making schooling choices, all would choose the same value of s, satisfying $d\log y(s)/d\log s=r$. Since there would be no differences in schooling choices among them occupations and jobs that required either more or less education would go unfilled, and the labour market would not clear. Yet, if we observe that in the market equilibrium different people choose different amounts of schooling, with some actually choosing more education and some actually choosing less, then the market earnings on jobs with different schooling requirements must adjust so that the marginal condition is an identity for all possible values of s. That is, people must be indifferent as to how much education they choose. Viewing the marginal condition as a differential equation in y and s and integrating yields the *restriction* $y(s)=y_0e^{rs}$, where y_0 is the earnings of a person without any schooling. Substituting this back into the definition of $W(s)$, we have

$$W(s) = y_0e^{rs}\int_s^{\infty} e^{-rt}dt = y_0/r$$

is *independent of s*. Writing $W(s)=W$ to reflect this fact, we have $y(s)=(rW)\,e^{rs}$, and $\log y(s)=\log(rW)+rs$. Think of this last expression as a regression equation. Then after adjusting the income data for age and experience, a regression of the log of income on years of school yields an estimate of the marginal internal rate of return to education (r) as the regression coefficient on schooling. The constant term in the regression estimates ‘earning capacity’ $\log(rW)$.

The economic logic underlying this development clearly shows the compensatory nature of the returns to schooling and its relationship to the theory of supply. The equilibrium earnings–schooling function is an equalizing difference on the foregone opportunity and other costs of attending school. If people are alike, earnings must rise with schooling to cover the direct and interest costs. Otherwise no one would be inclined to undertake these investments. Notice that in this example, income differences are equalized on cost at every point and that the human wealth (W) is the same for all. Thus there is inequality of

earnings, but complete equality of human capital wealth or life cycle earnings. Restricting attention to inequality in the observed distribution of earnings would give a highly misleading indication of inequality in the true distribution of economic welfare in this case.

This simple decision problem provides a convenient and powerful conceptual framework around which much of the research in this area has been organized. The value of this framework was first demonstrated by Becker, who expanded it to include interpersonal differences in abilities and talents and in family circumstances. Interpersonal differences in the rate of interest *r*, are identified with financial constraints on human capital investments associated with family background and related factors. A person confronting a higher rate of interest would be unable to finance human capital investments on favourable terms and would therefore rationally choose to invest less than a person who was able to borrow at lower rates. Similarly, there may be interpersonal differences in talents among people. Some may be more skilled in learning, which makes schooling effectively cheaper for them, or they may have natural talents which either complement or substitute for schooling in producing earning capacity. Considerations such as these lead to an identification problem in the schooling–earnings relationship observed *across* different individuals (see Rosen, 1977, for elaboration; also Willis). To begin, let us isolate the effects of family background and financial constraints by restricting attention to a subset of individuals with the same natural talents and abilities. Then differences in school choices within this group would be provoked by corresponding differences in family backgrounds and financial constraints. The reason for this goes back to the institutional feature of human capital assets noted above, that a person cannot sell an asset claim to future earning power. Thus human capital does not serve as collateral for investments in anywhere near the same way as title to physical capital does for non-human investment. A house, for example, serves as collateral for a mortgage. If the purchaser defaults on the mortgage then the creditor gains title to the house, which can then be sold to settle the debt. Non-transferable titles to human capital make this kind of arrangement impossible for personal investments. Relaxing these kinds of constraints is, of course, the fundamental economic logic behind the public provision of education in most countries throughout the world. But since direct tuition and related costs are only a part of the true costs of schooling, the importance of foregone earnings costs suggests that financial constraints would still remain a factor in educational decision-making. As Marshall noted, the social and economic status of the family play an important role in educational choices.

From the point of view of econometric estimation, observing a subset of the population where abilities are roughly constant, but where financial constraints dictate different schooling choices allows identification of the schooling–earnings relationship for that ability level. This in turn enables the analyst to calculate the social rate of return on investment, and to determine empirically the effect on personal and aggregate wealth of social policies that relax the financial constraints. Earnings of otherwise similar people who were less constrained serve as excellent estimates of the true earnings prospects for more constrained individuals.

Extensive empirical investigation of the connection between schooling, earnings, and family background shows a very strong and systematic relationship between parents’ socioeconomic status and background and the school quality and completion levels of their children (e.g. Griliches, 1970, 1977). This is prima facie evidence of financial constraints on educational choices, though it does not rule out other routes by which family background affects a person's economic success, such as complementary investments in the home in child care and quality. These studies also indicate a direct connection between family background and earnings given the schooling choices of children. The causal link between these direct

effects of family background and earnings remain to be established. It could reflect common but unobserved variance components across generations within families, such as unobserved ability; and also unmeasured factors, such as school quality and the quality of parental inputs, that are correlated with family background. Whatever their source, these direct linkages are numerically small compared with the effect of schooling itself on earnings. Most of the effect of family background on economic success works through its effects on the educational decisions of children and through that to economic success as measured by income and earnings. The direct effect on income, while persistent and significant, is quantitatively small.

Some applications

Perhaps the main policy area where these ideas on financial constraints are important is in public provision of training and ‘manpower’ development programmes for the poor. The logic of these policies rests on the proposition that a person's income in a market economy reflects the quantity of resources that the person controls and the value of these resources. People who are permanently poor have less skills and also less valuable skills then the non-poor. So an attractive policy to help eliminate poverty is to give them more and better resources through education and training. The rate of return has been widely used for programme evaluation. For if the social return to investment in subsidized training is less than the rate of return on other forms of social investment, then programmes emphasizing direct monetary and other transfers to the poor are better bets for society overall than devoting resources to skill enhancement. There now exists a voluminous literature on manpower programme evaluation along these lines, largely stemming out of the social programmes that were instituted in the 1960s and 1970s in the United States. The evidence is mixed. While many examples of successful programmes can be found, the prevailing assessment among experts is that the average programme has not been clearly successful (Ashenfelter, 1978). This empirically based conclusion suggests that the underlying causes of poverty are more complicated than simple family constraints on resources which thwart human capital investments. Lack of motivation, discrimination, ability, low quality prior education and insufficient investments in children in the home, as well as constraints on financing are among many of the possibilities that present themselves as causal factors in reducing personal investments in human capital.
The changing role of women in the workplace and in the home has refocused current professional interest on the role of families in determining economic success of children. While these intergenerational connections between the wealth and economic status of parents and their children have long been recognized as a key element in the question of poverty and the size distribution of income, these aspects have only been linked to human capital theory in very recent years. Again, the impetus for this interest lies in the empirical findings summarized above, and also in some that have come from unexpected quarters, namely the economic success of immigrants and their children.
Recent work by Barry Chiswick (1978) has established a systematic empirical pattern for many immigrant groups into the United States. Chiswick finds that members of the first generation of immigrants earn less than comparable native born citizens in the first two decades of their life in the US. At that point their incomes reach parity with native born citizens and beyond it actually surpass the incomes of the native population. More remarkably, the sons of these immigrants – the members of the second generation – earn incomes which exceed those of the sons of native born workers. However, by the third generation there is parity, and the effects of foreign-born status wash out. While certain aspects

of Chiswick's findings remain controversial and are being studied at length, they support the ‘melting pot’ view of economic life in the US. There is obviously substantial interest and importance in examining similar phenomena in other countries.

The chief theoretical work in the intergenerational transmission of wealth and economic status through families is contained in the research of Becker and Tomes. This work directly addresses intergenerational linkages through preferences and attitudes of parents toward their children, through natural hereditary transfers of ability and through discretionary transfers of resources through the generations. This work is the most complete theoretical description of the intergeneration distribution of wealth available so far. Inheritability of abilities is known from statistical theory to imply a regression-toward-the-mean phenomenon. Thus the fortunes of one generation are not only linked by direct transfers of non-human wealth and human capital investments, but also by inherited traits. These two forces interact in the intergenerational transmission mechanism. The economic fortunes of generations are more closely linked the greater the degree of inheritability of ability and the greater the propensity of parents to invest in their children's human capital. The effects of good fortune in one generation spills over to the next through the transfer mechanism. Interestingly, it may spill over to several subsequent generations. Thus regression toward the mean may occur only after several generations rather than after only one. When borrowing constraints are imposed on this structure even more persistence is implied because low income families do not have sufficient resources to invest in their children, whose incomes as parents are smaller than they would otherwise be. These issues are important for understanding social and economic mobility, and only recently have data become available to study them empirically. In the end this may be one of the most important developments in human capital theory.

Ability bias

The other major area where considerable research progress has been made is the role of ability in determining economic success. In terms of the decision model above, interpersonal differences in ability shift the earnings–schooling relationship. More able persons earn more at a given level of schooling than the less able, so the observed income-schooling relationship does not necessarily represent the returns available to a given person. Thus consider a group of individuals who have the same financial resources (the same value of r in the term discussion above). If ability is complementary with schooling then the rate of return to schooling will be larger for the more able and they will choose to invest more. A person observed choosing less education rationally does so because the personal return is relatively small under these circumstances. Comparing the earnings of persons who choose less education with those of persons choosing more education leads to a biased assessment of the returns due to differences in their abilities. This ‘ability bias’ issue has been examined in much detail.

The basic issue was originally posed by Becker, using the discounted earning stream comparisons presented above. If $x_0(t)$ is the earnings stream of people who stop school after high school completion and $x_1(t)$ is the earning stream of those who continue to college, then $x_1(t)$ is likely to be a biased estimate of the earnings prospects of high school graduates had they continued on to college. In so far as their average ability is lower than college graduates, their earnings had they chosen to continue on to college are likely to be smaller than $x_1(t)$. Similarly, the higher average abilities of college going persons makes it probable that $x_0(t)$ is a downward biased measure of what they would have earned had they

stopped their education after high school graduation. Thus comparing $x_1(t)$ with $x_0(t)$ yields an upward biased estimate of the rate of return to education for either group.

In order to correct this bias it is necessary to purge the earnings data of the direct effects of ability. Several methods have been proposed, and most find that the effect of ability biases in rate of return calculations is positive but relatively small (Griliches). The fundamental reason for this is due to a finding of Welch, that while the direct effect of measured ability on earnings is positive (given schooling), its numerical effect is quite small. Even a person whose measured ability is one standard deviation above the mean receives, on average, an income that is only a few percentage points above average.

Most of the research in this area has concentrated on indexes of ability associated with IQ and other measures meant to predict school performance. However, predictors of school performance and grades are not necessarily good predictors of economic success. The most sophisticated studies employ factor analytic statistical models, in which measured abilities embodied in IQ scores and the like serve only as indicators of underlying and unobserved ‘true’ abilities. These studies show that ‘raw’ rate of return estimates unadjusted for ability differences overstate ‘true’ rate of return calculations by only a few percentage points. The rate of return to school remains substantial, and of comparable magnitude to that on other forms of investment even after ability adjustments have been made.

Most of this ability-bias research assumes that ability can be captured statistically as a single factor (in the statistical sense). However, some recent work is based on a multiple-factor view of ability in which there are different dimensions and components (Willis and Rosen, 1978). This multi-factor framework is familiar from the theory of comparative advantage in economics. A unidimensional specification of ability only allows for absolute advantage, where a person who is more able in one thing is necessarily more able in everything else. By contrast, a comparative advantage specification allows for both absolute and relative advantages. A person may be very talented in all things (absolute advantage), but may also be relatively more talented in some things than others. Furthermore, absolute advantage may not be so important. A great musician is not necessarily adept at non-musical activities such as accounting; and the typical accountant may well have no more than the average musical ability in the entire population. An extension of the model above shows that people would naturally select themselves into those occupations and educational categories that exploit their comparative advantage. Thus those who choose to specialize their human capital investments in musical activities would be likely to have more natural talent for it than the population at large. Similarly, those who learn the plumbing trade would be likely to have more mechanical ability than those who make some other choice. These types of selection problems gain research interest because educational and occupational choices are closely linked. While much important work remains to be done in this area, available evidence is at least consistent with the existence of comparative advantage and occupational selection. If so, the overall ability bias in simple rate of return calculations is likely to be relatively small.

The question of ability bias and selection comes up in a quite different manner in the literature on educational screening and signalling (Spence, 1973). In its most extreme form, the signalling literature maintains the hypothesis that education has no direct effect on improving a person's skills, but rather serves as an informational device for identifying more and less talented people. This model rests on a unidimensional view of ability and also on the suppositions that direct observation of a person's ability and productivity is very costly and that a person knows much more about his own abilities than other persons do. In these circumstances, education serves as a signal of ability if the more able can purchase

the educational signal on more favourable terms than the less able. For then education and ability are highly correlated, and the higher income earned by those with more schooling is supported in equilibrium by their higher ability-productivity.

Several points must be made in this connection. The first is, that taken on its own terms, the signalling and human capital models have very similar implications for the rational choice of schooling. In fact they appear to be econometrically indistinguishable on the basis of income and schooling data alone. The chief difference is a normative one, that schooling has a little social value when it serves as a signal, and has much social value when it produces real human capital. Second, the data reveal considerable 'noise' in the schooling–earnings relationship. An investigator does very well when a third of the total variance in earnings can be 'explained' in the analysis of variance sense by observable personal factors such as education, experience, ability measures, family background and other factors. The schooling–earnings relationship is very strong in the sense of population averages, but the error in prediction is very large for any given person. Large personal prediction errors dull the value of education as a signal. This fact also suggests that education is a personally risky investment. Third, when the signalling model is expanded, it does not necessarily imply that educational signals are socially unproductive. Education may have significant social value in identifying naturally talented people if there is social value in classification and sorting. For example, there may be significant interactions among workers in an organization. If so, then the organization must be structured to choose the optimal *distribution* of talent within it; for example, it may be socially beneficial for the most talented people to work together. In so far as the educational system serves to classify people for these purposes, it is producing a form of human capital (information in this case) which has both private and social value. Finally, the value of education in assisting persons to find their niche in the overall scheme of the economy, precisely because they do not know so much about themselves, has never been quantified.

Signalling and information

A definitive empirical study capable of distinguishing signalling and human capital views of investment in education is yet to be produced in spite of many attempts to do so. Most work in this area has floundered on the fact that the two views imply very similar equilibrium implications about the observed relationship between earnings and schooling, so that if any real progress is to be made, future investigations will have to look elsewhere. A promising area is to examine the direct effects of education on productivity (and not on income alone). Much research has been done on educational production functions, which have an obvious bearing on these linkages and how a different form of education might affect them. For example, some evidence suggests that preschool training can overcome the adverse effects of a poor home environment in educational success. Hanushek (1977) reviews the literature on educational production.

Surprisingly few studies have attempted to examine the schooling–productivity linkage directly, probably because data on personal productivity measures are hard to find, but those few that have managed to do so have found some very impressive results. Griliches reviews the issues at the aggregate level. However, the sharpest results have arisen in agriculture, a sector which has shown an enormous and sustained growth in productivity for at least five decades. The rate of return to education among farmers is substantial. Since most of these persons are self-employed and sell their produce in impersonal, competitive markets, it is difficult to make an *a priori* case that signalling plays any

significant role in their educational decisions. Moreover, detailed study shows how these returns come about. More educated farmers control larger resources in the form of larger farms. It is possible that there is a common connection with family background and wealth. However, available evidence suggests that these farmers are also much more efficient in their techniques of production, and that their education is used primarily to keep them informed of recent technological changes in agricultural production, which they adopt with greater frequency and with quicker response. The case that education makes farmers more efficient processors of new information is very well made in the work of Welch (1976, 1979). Schultz indicates that similar findings would apply to much of agricultural production throughout the world, and broadens the argument to make it more generally applicable to all walks of life.

Non-monetary considerations

Another potential source of bias in rate of return calculations arises from the limitations of earnings data. Using expected discounted earnings as the choice criterion is a first order approximation to a more complete formulation. Discounted expected *utility* is the ideal choice index, because an employment relationship is a tie-in between the productive services rendered by human capital skills on the one hand, and the consumption of non-pecuniary aspects of the work environment on the other. The imputed monetary equivalent value of these job-consumption items should be added to earnings in a complete calculation. The same is true of the skills that are utilized outside of the market sector, such as in home production (see Michael, 1982).

That individuals may differ in their tastes for employment of alternative forms of human capital leads to the existence of rents in human capital valuations. Furthermore, the evidence suggests that on-the-job consumption values increase with education and skill. Jobs which require more schooling are likely to be more desirable on *both* monetary and non-monetary grounds (this evidence is reviewed in Rosen, 1986). Economic theory suggests that some portion of earning capacity would be ‘spent’ on more desirable and more amenable jobs. To the extent that the value of work amenities increase with schooling, observed earnings are a downward biased estimate of total earnings for the more educated, and measured rates of return are downward biased.

These issues are most sharply drawn in the treatment of hours worked in rate of return calculations. For example, if observed earnings alone are used in the calculations, groups such as physicians are found to exhibit large rates of return on their medical education, whereas groups such as teachers are found to earn much lower returns. But physicians work very long hours, perhaps as much as 40 per cent more than the typical worker, whereas teachers work far fewer hours than most other workers; they do not work in the summer, for instance. It is necessary to make judgements about the imputed value of leisure to deal adequately with these differences. If leisure is valued at the wage rate, the proper calculation refers to ‘full’ income at a common hours-worked standard. Similar considerations apply to growth accounting calculations: The secular increase in embodied skills and human capital has been accompanied by a secular decrease in working hours among the employed population. The imputed value of the quantity and quality of increased ‘leisure’ should be counted in a measure of welfare. Also, using only market transactions as a basis for calculation conceals the significant value of human capital in home production among those groups, especially women, whose activities have shifted between the

non-market and market sectors.

Occupational choice

The discussion so far has concentrated on the role of formal schooling in human capital production. A small but important literature has used these ideas to analyse occupational choice, especially among the professions. The first, and still significant work in this area is due to Friedman and Kuznets, who set the general framework in terms of wealth maximization and rate of return calculations on entry into law, medicine and dentistry. Subsequent literature, of which the work of Freeman is especially notable, has applied modern time-series statistical methods to these problems, concentrating especially on the role of income prospects in attracting or repelling new entrants into a profession.

The human capital perspective suggests that longer term income prospects should play an important role in occupational decisions of the young and that short-term and transitory fluctuations should be of lesser consequence because they have small impact on expected lifetime wealth. Nevertheless, a central finding in this literature is that current market conditions have large effects on occupational choice, and that supply to a specific occupation is relatively elastic with respect to current wages. The effects of long-term prospects have been much more difficult to isolate empirically, depending as they do on specific formulations of expectations and the connections between future earnings expectations and current and past realizations. In so far as a person is ‘locked in’ to a profession after choosing it, economic theory suggests that long-term expectations should be the primary determinant of choice. The finding that current prospects are highly significant in these choices suggests considerable mobility and recalibration of choices after training. For example, many lawyers use their skills outside the formal practice of law and in complementary ways in the business sector more generally. However, the nature and extent of ex-post mobility possibilities remains to be thoroughly examined.

Learning from experience

From the theoretical point of view, formal schooling decisions are only half the story in human capital accumulation and skill development. Investment does not cease after schooling: there is another sense in which it just begins. Formal schooling sets the stage for accumulation of specific skills and learning in concrete work situations, through on-the-job training. The human capital literature interprets the term ‘on-the-job training’ very broadly. Only a small part of the overall concept is included in formal training programmes, apprenticeships and the like. The greater part is associated with learning from experience. This broad and inclusive interpretation is supported by persistent empirical observations on the evolution of earnings over the life-cycle. The age structure of earnings shows remarkably systematic patterns. Earnings rise rapidly in the first several years of working life, but the rate of growth falls toward mid-career and tends to turn negative toward retirement. In panel data, wage rates rise throughout the life cycle, with the greatest rate of increase in the early years. An attractive interpretation of these observations is that the increase in earnings with work experience is due to increasing productivity and human capital accumulation over the entire life cycle.

A fruitful empirical approach for studying these patterns has been developed by Jacob Mincer (1974). The conception of the problem extends the education model above. A person is viewed as making

human capital investment choices at each point in the life cycle. Workers who choose to invest more pay for their choice by accepting lower earnings when young and earn returns on their prior investments in the form of larger earnings when they are older. This is essentially a choice between a level experience–earnings pattern (if investments are small) and a ‘tilted’ one, starting at a lower point and rising to a higher one if investments are large. Mincer develops the concept of ‘overtaking’ to impute the total return to human capital. The basic idea extends the Smithian principle of compensation to on-the-job training investments. Suppose a person has a large variety of possible investment opportunities after completing school. If no further investments are made, the experience earnings profile is relatively flat. The slope of the earnings-experience profile is increasing and the intercept of the profile is decreasing with the magnitude of investment. Hence the investment level defines an entire family of age earnings profiles, which are spun out around a roughly common crossing point, labelled the ‘overtaking’ point, if in market equilibrium wealth is approximately independent of investment.

The model has a very sharp empirical prediction that in a cohort of individuals with the same schooling level and different post-school investments, the interpersonal variance of earnings should be decreasing with experience up to the overtaking point and increasing thereafter. These systematic variance patterns have been found by many investigators in a variety of data sources. The assumptions that on-the-job investments are completely equalizing and that human wealth is the same for all investment paths makes it possible to decompose total investments into formal education and on-the-job components. Mincer reports that the on-the-job components are substantial, of the order of a third or more of the total.

The complete education–experience human capital model has important implications for the analysis of poverty and income distributions. In a nutshell, human capital theory suggests that life-time earnings is the appropriate construct for understanding inequality. To the extent that age-earnings patterns are the result of rational investments in human capital, it is misleading to use unadjusted cross-section annual earnings data for inequality analysis. For those young persons who are intensively engaged in investment activities and whose current income is therefore small at present may be classified erroneously as poor even though they are not poor in the lifetime sense. These life cycle issues have not been given sufficient attention in the extensive literature on the social welfare consequences of inequality, in spite of the fact that Paglin (1975) conclusively shows that they have large consequences for the measurement of inequality. Taking the life cycle view yields Gini coefficient estimates of real inequality that are smaller than when only current incomes are used in the calculations.

More detailed econometric work on the dynamic structure of individual earnings based on panel data helps resolve questions of the extent to which poverty status is permanent or transitory over the life cycle. The most sophisticated study so far (Lillard and Willis, 1978) decomposes earnings into several components. One is measureable characteristics of persons, such as education and experience, which reflect human capital and other considerations. Another is a ‘person effect’ capturing unmeasured components of ability, health, and related factors which permanently affect a person's earning power relative to his cohort. Finally, the third component reflects more transient variations, reflecting such factors as luck and other random events which may persist for a time but which eventually die out. Each component explains about one-third of the total variance of earnings. Since the measurable factors are, by human capital theory, largely equalizing on prior investments and the transitory effects have only small effects on life cycle wealth, this leaves about one-third of the total variance of life cycle earnings as attributable to permanent differences among persons or to ‘pure’ inequality. Certainly this is quite a different picture than emerges from examining the cross-section distribution of current earnings.

Other approaches to understanding age–earnings profiles in the human capital framework have used a more formal capital–theoretic structure. Here human capital is associated with the latent stock of embodied skills and investment with skill acquisition and learning. A person must give up current income to learn more and increase the stock of skills available for rental at a later date. The optimal investment programme maximizes the present value of lifetime earnings. This basic set-up of the problem was first formulated in an important paper by Ben-Porath (1967), who structured the investment control as choice of the division of a person's time between working and investing. An extension by Rosen (1972) structures it as choice among a spectrum of jobs which offer different learning environments and opportunities. The wage on a job that offers more learning possibilities is lower and the programme is implemented by a ‘stepping stone’ progression of positions.

This capital theoretic formulation of the problem has virtues in demonstrating the conceptual commonalities between capital and growth theory and human capital theory. However, its generality comes at the cost of providing less robust predictions. Thus it seems fair to say that extensive work attempting to implement these rigorous ideas empirically has not met with overwhelming success in extracting information from observed age–experience trajectories. It appears that other important forces also affect these patterns. Several possibilities have been suggested. One relates to investments in information and search for enduring long job attachments. Job turnover is much larger among young workers than older ones. While this is a form of human capital accumulation and much recent work has been devoted to these issues, it has so far proven difficult to link this class of problems with the ideas reviewed here. Nor has human capital theory yet adequately come to terms with the fact that job patterns typically exhibit discrete jumps and ‘promotions’, where the character of human capital services rendered changes at each step. Competition for higher ranking positions is properly considered within the human capital framework, but little analysis is available so far.

Any review of human capital would be remiss in not calling attention to parallel developments and important applications in economic historians’ interpretation of slavery. The work of Fogel and Engerman (1974) stands out as the primary example of the approach. Here the empirical work focuses on direct human capital valuations rather than on earnings. The principles of capital valuation are used to examine such issues as the long-term economic viability of slavery as an economic institution in the absence of intervention. In addition, some important and fascinating agency problems must be confronted because of an inherent conflict in the master-slave relationship. The conflict arises because the owner naturally desires more effort than the slave prefers to put forth. Various institutions, involving *both* punishments and rewards, were structured to help resolve these conflicts. Mention also should be made of research on indentured servitude by economic historians (Galenson, 1981), which is analysed as a response to a capital market imperfection. A person voluntarily indentured himself for a period of years as payment for a loan to provide transportation and connections in the New World. Repayment was guaranteed by a legally binding claim on the person's services for the period of the contract.

Demographic effects

Over the years there has been increasing recognition of the relationship between human capital and economic demography. This is inherent in the role of families as both producers and financiers of human capital investments. Two important recent developments strongly rest on these connections.

The first one is related to large demographic changes in the age structure of the population in the post-war period (the ‘baby boom’) in the United States. Rates of return on education had remained remarkably constant for a thirty-year period. This in spite of the fact that there had been an enormous increase in education over that period. However, Freeman identified a decline in the rate of return commencing in the late 1960s. The evidence currently available suggests that the rate fell by several percentage points for a 10–12 year period throughout the 1970s, but had gradually returned to its prior level. The leading explanation for this has been provided by Welch (1979) and relates to increased competition for jobs within cohorts as a function of their size.

A stable age distribution of the working population provides a naturally stable progression of work and job opportunities over a person's working life. Not only the level, but also the nature and productive role of human capital changes over the life cycle. Young workers perform different tasks and have different responsibilities than do older workers. Therefore competition and supply of human capital of various types in the labour market is strongly age related. Thus as the large birth cohorts of the 1950s began to enter the market in the late 1960s and 1970s, the increased supply of educated young workers lowered their wage rates and reduced the rate of return. These effects are diffused as the large cohort ages and works its way through the age distribution, and as the structure of work is altered to accommodate their large numbers. The weight of extensive research in this area has shown that returns and wage rates are affected by cohort size. The consequences of this research for the future development of human capital theory will be important, because it requires considering heterogeneous human capital investments and the evolution and development of different types of skills over working life. It may ultimately require analysing how work itself is organized and structured.

Human capital and discrimination

A final important recent development proceeds on somewhat more conventional theoretical grounds. It addresses the role of human capital in observed wage differences between men and women, and is ultimately related to questions of labour market discrimination. The work in this area is firmly based on empirical calculations. The main fact to be explained is that women earn less than men, even after adjusting for differences in occupational status and hours worked. Labour market discrimination against women is one possible interpretation. However, there may be more subtle forces at work. Mincer and Polachek (1974) build an alternative interpretation on the observation that earnings–experience profiles of women are flatter and exhibit much less life-cycle growth than that of men, and tied it to the well known fact that women traditionally have exhibited less stronger labour force attachments than men due to the sexual division of labour in the home and the bearing and raising of children.

The value of an investment increases with its rate of utilization. Compare two persons: one who expects to utilize an acquired skill very intensively and one who expects to utilize it less intensively. Suppose further that the costs of acquiring the skill are approximately independent of its subsequent utilization. Then the rate of return on investment is larger for the intensive user and that person will tend to invest more. The application to male–female wage differential is apparent upon connecting intensity of utilization with labour force attachments and hours worked. In so far as married women play dual roles in the market and in the household, there is a tendency to invest less in labour market skills and more in non-market skills. The opposite is true of men, given prevailing marriage institutions. These differential incentives can account for differences in age earnings patterns between men and women as well as the

larger average wages of men. Research on female labour supply supports the point by showing overwhelming evidence that labour force activities of married women are severely constrained by the presence of children in the home. Mincer and Polachek provided direct empirical support by demonstrating that earnings of never-married women closely approximate those of men. Considerable research is in progress on these ideas (see, for example, *Journal of Labor Economics*, 1985). At a minimum, the human capital perspective shows that these issues are more complicated than appears on the surface. Yet there are some unresolved puzzles. In spite of the vast increase in female labour force participation in the past two decades, the relative wages of men and women have not changed very much in the United States, though they have come closer to parity in a number of other countries. Part of this may be due to differences in the importance of the government sector as employers of women, as well as differences in compliance with equal pay legislation. A definitive answer is not yet on the horizon.

This essay started by noting the twin origins of developments of the theory of human capital in understanding the sources of economic growth on the one hand and the distribution of economic rewards on the other. Much progress has been made on both counts. However, these two branches have not yet been clearly joined. Future progress will have to come to terms with the issue of how private incentives to acquire human capital affect the available social stock of productive knowledge and how changes in social knowledge become embodied in the skills of subsequent generations.

See Also

- cognitive ability
- family economics
- gender differences (experimental evidence)
- gender roles and division of labour
- household production and public goods
- measurement
- signalling and screening
- value of time

Bibliography

Ashenfelter, O. 1978. Estimating the effect of training programs on earnings. *Review of Economics and Statistics* 60(1), February, 47–57.

Becker, G. 1964. *Human Capital*, 2nd edn. New York: Columbia University Press, 1975.

Becker, G. and Tomes, N. 1978. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87(6), December, 1153–89.

Ben-Porath, Y. 1967. The production of human capital and the life cycle of earnings. *Journal of*

Political Economy 75(4), Pt 1, August, 352–65.

Blaug, M. 1976. The empirical status of human capital theory: a slightly jaundiced survey. *Journal of Economic Literature* 14(3), September, 827–55.

Chiswick, B. 1978. The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy* 86(5), October, 897–921.

Denison, E. 1962. *The Sources of Economic Growth in the United States and the Alternatives Before Us*. New York: Committee for Economic Development.

Dublin, L. and Lotka, A. 1930. *The Monetary Value of a Man*. New York: Ronald Press.

Fogel, R. and Engerman, S. 1974. *Time on the Cross*. New York: Little, Brown.

Freeman, R. 1971. *The Market for College-Trained Manpower*. Cambridge, Mass.: Harvard University Press.

Freeman, R. 1976. *The Overeducated American*. New York: Academic Press.

Friedman, M. and Kuznets, S. 1954. *Income from Independent Professional Practice*. Princeton: Princeton University Press.

Galenson, D. 1981. *White Servitude in Colonial America*. Cambridge: Cambridge University Press.

Griliches, Z. 1970. Notes on the role of education in production functions and growth accounting. In *Education, Income and Human Capital*, ed. L. Hansen. New York: National Bureau of Economic Research.

Griliches, Z. 1977. Estimating the returns to schooling: some economic problems. *Econometrica* 45(1), January, 1–22.

Hanushek, E. 1977. *A Reader's Guide to Educational Production Functions*. Institution for Social Policy Studies, Yale University.

Kendrick, J. 1976. *The Formation and Stocks of Total Capital*. New York: National Bureau of Economic Research.

Knight, F. 1944. Diminishing returns from investment. *Journal of Political Economy* 52(March), 26–47.

Lillard, L. and Willis, R.J. 1978. Dynamic aspects of earnings mobility. *Econometrica* 46(5), September,

985–1012.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan, 1930.

Michael, R. 1982. Measuring non-monetary benefits of education: a survey. In *Financing Education: Overcoming Inefficiency and Inequity*, ed. W. McMahon and T. Geske. Urbana: University of Illinois Press.

Miller, H. 1960. Annual and lifetime income in relation to education, 1929–1959. *American Economic Review* 50(December), 962–86.

Mincer, J. 1958. Investment in human capital and personal income distribution. *Journal of Political Economy* 66(August), 281–302.

Mincer, J. 1974. *Schooling, Experience and Earnings*. New York: Columbia University Press.

Mincer, J. and Polachek, S. 1974. Family investment in human capital: earnings of women. *Journal of Political Economy* 82(2), Pt II, March–April, S76–S108.

Paglin, M. 1975. The measurement and trend of inequality: a basic revision. *American Economic Review* 65(4), September, 589–609.

Petty, W. 1676. *Political Arithmetic*. In *The Economic Writings of Sir William Petty*, Vol. 1, ed. C. Hull. Cambridge: Cambridge University Press, 1899.

Rosen, S. 1977. Human capital: a survey of empirical research. In *Research in Labor Economics*, Vol. 1, ed. R. Ehrenberg. Greenwich, Conn.: JAI Press.

Rosen, S. 1985. The theory of equalizing differences. In *Handbook of Labour Economics*, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

Rosen, S. 1986. The theory of equalizing differences. In *Handbook of Labour Economics*, ed. O. Ashenfelter and R. Layard.

Sahota, G. 1978. Theories of personal income distribution: a survey. *Journal of Economic Literature* 16 (1), March, 1–55.

Schultz, T. 1961. Investment in human capital. *American Economic Review* 51(March), 1–17.

Schultz, T. 1975. The value of the ability to deal with disequilibria. *Journal of Economic Literature* 13 (3), September, 827–46.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, Modern Library Edition. New York: Random House, 1947.

Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87(3), August, 355–74.

Walsh, J. 1935. Capital concept applied to man. *Quarterly Journal of Economics* 49(February), 255–85.

Welch, F. 1970. Education in production. *Journal of Political Economy* 78(1), January–February, 35–59.

Welch, F. 1976. Ability tests and measures of differences between black and white Americans. Rand Corporation.

Welch, F. 1979. Effects of cohort size on earnings: the baby boom babies' financial bust. *Journal of Political Economy* 87(5), Pt II October, S65–97.

Willis, R. 1986. Wage determinants: a survey and reinterpretation of human capital earnings functions. In *Handbook of Labour Economics*, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

Willis, R. and Rosen, S. 1978. Education and self-selection. *Journal of Political Economy* 87(5), Pt II, October, S65–S97.

How to cite this article

Rosen, Sherwin. "human capital." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000100> doi:10.1057/9780230226203.0754

The New Palgrave Dictionary of Economics Online

Hume, David (1711 – 1776)

Eugene Rotwein
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capital supply; class; economic growth; economic individualism; economic psychology; excise taxes; free trade; gains from trade; happiness; historical economics; human nature; Hume, D.; institutional economics; international trade; invisible hand; laws of human behaviour; liberty; luxury; multiplier; natural history; optimum taxation; probability; psychological hedonism; public debt; quantity theory of money; rate of interest; Smith, A.; specie-flow mechanism; tax incidence; utilitarianism; utility of poverty doctrine

Article

David Hume's economic essays (which originally appeared in 1752 in a volume entitled *Political Discourses*) comprise a small portion of his writings. The scope of Hume's thought was vast. He wrote extensively in philosophy (the area in which his reputation primarily lies), explored several of the social sciences and the humanities, and was deeply interested in history. His multi-volume *History of England* (1754–61) was a path-breaking work in the field. Nonetheless, in the literature Hume's economic writings have typically been treated as an entirely self-contained aspect of his work. This is not surprising, since in his economic essays he does not allude to his other writings, and subsequent disciplinary specialization has not encouraged consideration of any interrelationships between the two. For their part, philosophers have often treated Hume's philosophical writings in isolation from his other work.

For Hume, however, there was no such sharp disjunction. In the Advertisement prefixed to his first and major philosophical work, *A Treatise of Human Nature* (1739), he states that he expects his philosophy to serve as the ‘capital or centre’ of all the ‘moral’ (that is, psychological and social) sciences and that he hopes to expand the *Treatise* to accommodate a study of these areas. Owing perhaps to the poor reception accorded his *Treatise*, Hume did not carry out his original intention. His treatment of the moral sciences was left mainly to his essays. But there are many links between Hume's philosophical thought and his essays, and this is true with respect to his economic essays. Indeed, in light of the importance of

these links, Hume may be regarded as the outstanding philosopher economist of the 18th century.
Viewed in most general form, what is the nature of the relationship between Hume's economic and philosophical thought? Hume regarded the foundation of his entire philosophical system – its ‘capital or centre’ – as a body of ‘principles of human nature’, or elements and relations concerning human understanding and human passions that he believed to be irreducible and universal. These principles, which constitute the analytical phase of Hume's system of thought, are treated in Books I and II of the *Treatise*. In the second and synthetic phase Hume then relates various aspects of ‘human nature’ to environmental forces in seeking to frame laws of human behaviour, or generalizations indicating how man may be expected to behave under different specific conditions. These generalizations comprise the substance of the ‘moral sciences’ with which, as indicated, Hume dealt principally in the essays. An explicit and deep interest in psychology is thus a salient characteristic of Hume's treatment of the ‘moral sciences’ in general, and this is conspicuously evident in his economic analysis.
What were Hume's views concerning the prospects of developing reliable generalizations in the ‘moral sciences?’ That Hume should have distinct views on this issue is scarcely surprising in light of the depth of his interest, as a philosopher, in the epistemological basis of science. As he had argued, the contrary of any generalization concerning relations between matters of fact is always conceivable and hence always possible. Consequently, the only way of developing an understanding of these relations, he contended, is through empirical observation; and this can only yield probabilities, never certainty. With respect to his own principles of human nature, Hume believed that his propositions carried the highest order of probability because of the abundance of evidence on which they rested.
On the other hand, recognizing the complexity of the interrelationships between man's ‘nature’ and his environment, he stressed the difficulty in framing valid laws of human behaviour. He calls attention to the effect on human behaviour of imperceptible influences, emphasizes the extent to which it could be altered by changing conditions and notes the impracticality of conducting controlled experiments in the realm of psychological phenomena. He thus warns that in the social sciences ‘all general maxims … ought to be established with the greatest caution’ and states that ‘I am apt … to entertain a suspicion that the world is still too young to fix many general truths in [the area of the social sciences] which will remain true to the latest posterity’ (Hume, 1875, vol. 3, pp. 156–7). Of all the social fields, however, he believed that a field such as economics lent itself especially well to scientific study, and here he was cautiously optimistic concerning the possibility of developing reliable generalizations through direct observation of man in the course of his day-to-day affairs. As he argued, behaviour here was governed by mass passions, which were ‘gross’ or ‘stubborn’, or were not as affected by imperceptible influences as passions governing the behaviour of small numbers of individuals. Uniformities in behaviour therefore could here be more readily discerned (1875, vol. 3, p. 176). It should be noted that, in accord with this view, Hume introduces his economic essays by contrasting the potential for scientific analysis in economics with the very limited prospects for such analysis in a field such as foreign diplomacy, where events are controlled by the behaviour of a small number of individuals (1955, pp. 3–4).
To return to the substance of Hume's economic thought, in addition to emphasizing psychological considerations Hume's analysis displays a deep interest in historical sequence. Hume's interest in history developed at a very early age, even before he undertook his *Treatise*. As it appears in his essays, however, his treatment of history differs from conventional historiography (with its concern with unique particulars) which predominates in his *History of England*. For, writing as a ‘moral scientist’, Hume sought to reduce historical sequence to generalizations which explain how transformations in human

behaviour result from the impact of changing historical circumstance on ‘human nature’. This type of study (which bore a relationship to the ‘conjectural history’ and the French ‘*histoire raisonée*’ of the period) Hume termed ‘natural history’ – the term ‘natural’ here denoting the recurrent or probable, or the substance of laws of human behaviour. There are clusters of what Hume regards as historical laws of human behaviour in several of the essays. One essay bears the title ‘The Natural History of Religion’. And in the economic essays the approach of ‘natural history’ is of fundamental importance.

This can be seen when Hume's economic essays are viewed on three different levels of analysis. The first is economic psychology, where Hume deals with economic motivation, or what he terms the ‘causes of labour’. This is the most basic level of his economic analysis in the sense that here one finds the links between his economic thought and his treatment of ‘human nature’ in the *Treatise*. On this level the analysis takes the form of a natural history of ‘the rise and progress of commerce’. In a word, Hume introduces the question of economic motivation in seeking to explain how changing environmental influences stimulated the economic growth of his general period through their impact on various human passions. Here Hume observes that there are four ‘causes of labour’ – the desire for consumption, the desire for action, the desire for liveliness and the desire for gain.

The first of these, which is commonly stressed by economists, simply denotes all the wants that may be gratified by consumption. The desire for action refers to a desire for challenging activity as such. However, its full effectuation, as Hume stressed, requires activity whose end or objective has independent value. Like hunting and gaming, economic pursuits (and especially the activities of the merchant and, more generally, the ‘industrious professions’) are seen as meeting these conditions. By the desire for liveliness Hume meant the desire for the experience of active passion as such (which he contrasts with a state of no passion, or in effect a state of waking sleep). This is not a completely independent cause of labour but is an important ingredient common to both consumption and interesting activity. The last cause of labour is the desire for monetary gain, which is a desire to accumulate the tokens of success in the economic ‘game’.

Hume argues that all these motives play a role in a nation's economic growth – the initial stimulus to which he finds in the expansion of international trade. As compared with the treatments of economic motivation by economists (which commonly accord exclusive or over shadowing emphasis to the desire for consumption), a striking characteristic of Hume's treatment lies in its multidimensionality. This multidimensionality is also found in Hume's criticism of the doctrine of psychological hedonism. Here he argues that, in addition to seeking pleasure, man is driven by a variety of ‘instincts’ which lead him to do things for their own sake, and therefore will not automatically lead him to act in his own best interests. Hume's position thus precludes any simple identification of wealth with welfare.

The second level of Hume's economic analysis is his political economy, or his treatment of market relations. It is this which makes up the bulk of his economic essays. Here Hume considers several of the major economic issues of his own period, including monetary theory, interest theory, the question of free versus regulated trade, the shifting and incidence of taxes, and fiscal policy. In this context the natural history of ‘the rise and progress of commerce’ plays a dominant role. For repeatedly in his critical treatment of the economic doctrines of his period Hume seeks to show that their major deficiency lies in a failure to give proper attention to the importance of economic growth and to the underlying psychological and other factors associated with this growth process.

Let us consider first Hume's quantity theory specie flow doctrine, which he presents (in the essay ‘Of the Balance of Trade’) in criticism of the mercantilist view that without restraints on international trade a

nation would suffer losses in its money supply. Hume's position, which has been recognized as an early anticipation of the classical view, is that, owing to the effects of specie flows on price levels in trading nations, the amount of specie in each automatically tends towards an equilibrium at which its exports and imports are in balance. Any attempt through restraints on trade to increase the amount of specie beyond this equilibrium level, as Hume argues, is destined to fail (on the assumption that the money circulates domestically) because the specie movement from abroad will raise the nation's prices relative to those abroad, reduce exports and increase imports, and generate a return outflow of specie.

The relationship of this analysis to Hume's historical perspective is evident in the purpose with which he introduces this doctrine. For in employing the quantity theory of money he is here arguing that the extent to which a specie inflow into a nation affects its prices depends on its total output. Consequently, as he is seeking to show, it is the level of a nation's economic development, or its productive capacity as determined by its population and the spirit of industry of its people, that controls the amount of specie a nation can attract and retain. As he states, ‘I should as soon dread that all our springs and rivers should be exhausted as that money should abandon a kingdom where there are people and industry’ (1955, p. 61).

To consider another of Hume's anticipations of the classical position – his interest theory presented in his essay ‘Of Interest’ – here he attacks the mercantilist view that the rate of interest is determined by the money supply. On quantity theory grounds he argues that an increased money supply will simply raise all prices and, necessitating an offsetting increased demand for loans to finance expenditures, will leave interest rates unaffected. It is therefore the supply of real capital that determines interest rates. The bulk of Hume's discussion, however, is concerned with the factors affecting the supply of real capital itself; and here he turns to a historical analysis in which he considers the effect of economic growth on the class structure of society and, through this, on economic incentives. In this context every ‘cause of labour’ considered in the natural history of ‘the rise and progress of commerce’ is brought into his treatment. In a feudal society, he points out, the supply of capital is low because there are only two classes – the peasants and the landed aristocracy. The peasants cannot save since they are poor. On the other hand, the landed aristocracy tend to be heavy borrowers. For, as they are idle and lack the sense of liveliness that interesting activity affords, they seek liveliness wholly through extravagant consumption expenditures. Capital is therefore scarce and interest rates are high. Economic development, however, spawns the growth of the merchant class and the industrious professions. These groups derive a sense of liveliness from economic activity. Consumption expenditure drops for this reason and also because the pursuit of profit nourishes a desire to accumulate gain as a token of success in the economic game. As the new industrious classes earn a substantial share of the growing national income, their disposition to save thus results in a significant increase in the capital supply and a decline in interest rates.

As noted, Hume employs the quantity theory of money in criticizing the mercantilist position. But Hume's monetary theory also exhibits a similarity to the mercantilist view. However, his treatment here too springs from an attempt to call attention to the importance of economic growth. Thus (in his essay ‘Of Money’) Hume – assuming a condition of less than full employment – grants that an increase in the quantity of money (as against a greater absolute quantity of money as such) need not simply raise prices but can stimulate economic activity. Here, in tracing the impact of the increased money supply as it courses through the economy, he presents a lucid description of the multiplier process. He denies, however, that the stimulating effect on industry – when resulting from a short-run increase in the money supply – can prove anything more than ephemeral. No justification for this view is given. But it serves to

underscore the conclusion of his analysis. For he goes on to argue that, if the increase in the money supply is gradual and continues over a long period of time, its stimulating effects on output will prove enduring because it will nourish the ‘spirit of industry’ and therefore economic growth itself. Similarly, although Hume argued that an increase in the money supply does not affect interest rates, near the conclusion of his essay ‘Of Interest’ he points out that a long-run increase in the supply of money, by stimulating economic growth and inducing a change in spending and saving patterns, can increase the supply of capital and lower the interest rate.

Another noteworthy area of Hume's analysis is his treatment of the issue of free versus regulated markets. Since the relevant comments are not found in his economic essays but rather lie scattered through his *History of England*, the full extent to which Hume anticipated Adam Smith's ‘invisible hand’ argument has not been generally recognized. These comments make clear that Hume understood the role of a free price mechanism is governing the allocation of resources (1955, pp. 1xxviii–1xxx).

In applying the argument for free markets to the case of international trade, Hume emphasizes that free trade makes it possible for nations to enjoy the gains from an exchange of the products of their different resource endowments. However, in his most thorough treatment of the issue of international free trade (in his essay ‘Of the Jealousy of Trade’) it is not this static approach to the question that predominates. Rather, once again, it is economic growth considerations that receive primary emphasis. For here, where Hume seeks to meet the mercantilist argument that foreign economic development adversely affects home industry and employment, he takes the position that expansion abroad, on the contrary, commonly promotes economic development at home. By increasing foreign income, he argues, economic growth abroad not only leads to an expansion of foreign demand for domestic output but, through an emulation of foreign technological innovations, promotes the advance of technology at home. Hume goes on to argue that even when foreign expansion competes with domestic output, there is no need for concern provided the nation's ‘spirit of industry’ – which is itself nourished by foreign trade – is preserved. For as long as a nation remains industrious it need not fear that other nations will encroach on the market for its staple and, even in the unlikely event that this does occur, an industrious nation can readily divert its resources to other uses. Moreover, in stimulating the spirit of industry, foreign trade also promotes the diversification of a nation's resource use, and so reduces the impact of any shrinkage of demand that may occur from time to time in particular markets.

There are indications that Hume was more fully aware of the possible costs of free trade than one would gather from the main argument in the essay ‘Of the Jealousy of Trade’. Elsewhere he treats the interests of poor and rich countries as incompatible, and in one place he also justifies the use of a tariff in specific cases (1955, pp. 34–5, 76, 199–205). In the essay ‘Of the Jealousy of Trade’ itself he recognizes, in a modification of his main argument, that there are circumstances in which a nation facing a loss of markets to foreign countries may find resource diversion difficult (1955, p. 81). The character of this essay as a whole (which appeared six years after the other economic essays) suggests, however, that after much reflection and groping Hume had concluded that free trade would have a markedly favourable effect on long-term economic growth for all nations, and that, with this end in view, any associated costs – which would be of a shorter-term nature – would be well worth sustaining.

A further illustration of the role of natural history in Hume's political economy is found in his treatment of the shifting and incidence of taxes (in his essay ‘Of Taxes’), where he considers the view that an expansion of taxes creates an expanded ability to pay the levies by increasing ‘proportionably the industry of the people’. This view was commonly held by the mercantilists and, in what came to be

known as ‘the utility of poverty’ doctrine, was employed to justify the imposition of excises on goods consumed by the poor. Hume's position here is twofold. He points out that history shows that natural burdens, such as relatively infertile soil, often stimulate industry, and he argues that artificial burdens such as taxes may have the same effect. This position springs from Hume's view concerning the importance of a desire for interesting action as a ‘cause of labour’ since he here emphasizes that in order to prove interesting the activity must be difficult and challenging. On the other hand, he emphasizes that, since economic activity is also motivated by a desire for consumption, increasing difficulty beyond a certain level in achieving consumption ends will lead to despair. From the viewpoint of its stimulating effect on industry there is thus an optimum tax level, and Hume takes the view that taxes on the poor throughout Europe have already so substantially exceeded that optimum that they are threatening to ‘crush all art and industry’. Considered as a whole, Hume's position represents an amalgam of both the mercantilist and the later classical view. He rejects the mercantilist ‘utility of poverty’ doctrine with its unqualified endorsement of higher taxes on goods consumed by the poor, but also would reject the view (which is based on the subsistence or accustomed standard of living theory of wages found in the writings of Smith and Ricardo) that any tax on labour would inevitably result in a reduction in its supply.

Hume's treatment of fiscal policy – the last major aspect of his political economy – does not reveal significant relationships to his natural history of the rise and progress of commerce. Owing to space limitations, his analysis – contained in the long essay ‘Of Public Credit’ – cannot here be considered in detail. It should be observed, however, that this essay, which deals specifically with the question of large and continually mounting public debt, constitutes in all essential respects a ‘natural history of the rise and collapse of public credit’. Particularly noteworthy in this analysis are the extensive relationships Hume draws between economic and other social developments, especially of a political and sociological character. Of all aspects of his political economy, this essay most fully exhibits Hume's awareness, as a moral scientist, of significant interrelations between different realms of social experience.

The third and last level of Hume's economic thought in his economic philosophy, which is his appraisal, on ultimate moral grounds, of the desirability of a commercial and industrial society. In light of his general concern, as a philosopher, with moral questions, it is hardly surprising to find that the question of the moral aspects of commercial and industrial growth was of basic importance for Hume. Appearing in the second of the economic essays – ‘Of Refinement in the Arts’ – he considers this question before turning to an analysis of market problems. Although the essay is brief, its scope is broad; for Hume discusses the impact of the development of an advanced economy both on the individual and on society as a whole.

The standard for moral judgement Hume employs is drawn from the utilitarian ethic – a position which he himself had expounded and defended in his philosophical analysis. And here the role played by his natural history of the rise and progress of commerce is fundamental. As observed, in this natural history Hume dealt with various ‘causes of labour’. In his economic philosophy three of these motives – the desires for consumption, for interesting activity and for liveliness – are now treated as ends which are regarded as major ingredients of the happiness of the individual. Here he argues that, by providing new consumption experiences, enlarging the scope for the enjoyment of economic activity as a form of interesting action and (through both the latter) enhancing a sense of liveliness, economic growth advances the fulfilment of all these ends. Economic growth, he contends, contributes to the fulfilment of a fourth end of importance to human welfare – a sense of peace and tranquillity or a state of no passion – which he argues is enjoyable only in ‘recruiting the spirits’ after intensive indulgence in lively

experiences. It is noteworthy that Hume's treatment of these ingredients of human happiness bears a direct relationship to the principal conceptions of the good life as Hume construes these in an earlier series of essays entitled ‘The Epicurean’, ‘The Stoic’ and ‘The Platonist’. Further, the pluralism reflected in his multidimensional prescription for human happiness springs from the position taken in a fourth essay on the good life entitled ‘The Sceptic’ (1955, pp. xcv–xcix).

Turning to a treatment of the effect of economic development on major aspects of social relations, Hume now expands the ‘natural history’ to encompass non-economic considerations. He argues that economic growth contributes to the growth of knowledge in the liberal as well as the mechanical arts, nurtures a sense of humanity and fellow-feeling, enhances a nation's spiritual as well as its economic ability to defend itself and, through its impact on the growth of knowledge and fellow-feeling, advances an understanding of the art of government and political harmony. A final political consideration, to which Hume gives special attention, is the charge (drawn from the experience of Rome) that luxury is corrupting and debasing and therefore is inimical to liberty. Hume argues that history shows that precisely the opposite is true. For the growth of commerce brings the expansion of the merchant class – the ‘middling rank of men’ who above all are interested in uniform laws protecting their property; and it is this development, he emphasizes, which has led to the growth of parliamentary government and the associated respect for individual liberty. Hume thus perceived the link between the growth of economic individualism and political liberty that has drawn so much attention since his time. Although Hume recognized that the development of commerce and industry could produce evils of its own, he argued that these were outweighed by its benefits. Owing apparently to an overzealous desire to counter the common religious objections to luxury, Hume overextends himself and leaves some of his arguments in support of economic growth open to criticism (1955, pp. cii–civ). His treatment nonetheless stands as an unusually broad and penetrating appraisal of a wealth-orientated individualistic society. In light of this it deserves recognition as an early classic.

Throughout our discussion, attention has been given to Hume's interest in the psychological and historical aspects of economic activity. A similar interest – pursued in varying degree – is found among other writings of Hume's own period. However, owing to his own searching analysis as a philosopher and historian, Hume's treatment was of a particularly high order; equally extraordinary was the extent to which he employed the method of ‘natural history’ in the treatment of a wide range of issues of economic theory and policy.

Comparing Hume with Adam Smith (his close friend), one is struck by the brevity of Hume's economic writings. Hume wrote a series of relatively short ‘discourses’ on selected topics. Smith's *Wealth of Nations* (1776) is a general economic treatise. In contrast to Smith, Hume moreover gives little systematic attention to price and distribution theory, which was to become the major concern of classical and neoclassical economics. In point of the general analysis of psychological and historical influences on economic activity, however, Hume's work is more comprehensive, more highly organized and more penetrating than Smith's. When dealing with the subjective aspects of human behaviour, Smith not infrequently regards them as universals (for example, his assertion that there is an innate disposition among men to ‘truck and barter’), where Hume treats them as historical variables and himself seeks to explain the nature of the specific historical influences at work (1955, pp. cvii–cx). In this Hume did not foreshadow the mainstream of subsequent economic thought; it was Adam Smith's tendency in his economic theory to abstract from history that was to become the dominant characteristic of later economic analysis. In point of general perspective (though often not its conceptual framework) Hume's

economic thought bears a relation to other subsequent lines of development – to the historical and institutional schools of economics, to the more current revived analytical interest in economic growth along with its associated cultural aspects, to the concern with psychological factors in dealing both with macroeconomics and the economics of non-competitive markets, and to the normative appraisals of economic systems in their fuller social settings.

In the standard histories of economic thought Hume has been accorded relatively little attention. He is often ignored altogether or treated cursorily as a predecessor of Adam Smith. Various studies of the technical aspects of economic analysis have called attention to several of Hume's contributions. These aspects of Hume's analysis are noteworthy in their own right. Their significance deepens and broadens when they are related to Hume's work as a philosopher and historian and are seen to take form within the context of ‘natural history’.

See Also

- bullionist controversies (empirical evidence)
- gold standard
- specie-flow mechanism

Selected works

1752. *Political Discourses*. Edinburgh: A. Kincaid & A. Donaldson.

1875. *The Philosophical Works of David Hume*, ed. with notes by T.H. Green and T.H. Grose, 4 vols. London: Longmans, Green & Co.

1955. *Writings on Economics*, ed. E. Rotwein. London: Nelson.

How to cite this article

Rotwein, Eugene. "Hume, David (1711–1776)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000104> doi:10.1057/9780230226203.0756

The New Palgrave Dictionary of Economics Online

hunters, gatherers, cities and evolution

Paul Seabright

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Human beings evolved in hunter-gatherer bands, and tended to flee from or to fight with strangers. They have subsequently learned to live in cities among a multitude of such strangers, at levels of violence far lower than those that characterized prehistory. The key to this development was the adoption of agriculture, which obliged humans to become sedentary to and to develop institutions to manage their encounters with strangers. We describe the evolution of the psychological preconditions for the agricultural revolution, and its consequences for social life.

Keywords

agricultural revolution; city; cooperation and its evolution; division of labour; equality; evolution; group selection; hierarchy; hunter-gatherers; learning; other-regarding preferences; population growth; reciprocity; repeated games; signalling; slavery; trust; urbanization

Article

Modern human beings (*Homo sapiens*) evolved in Africa but now occupy all the continents of the globe. The environmental conditions in which they live are mostly quite different from those in the woodland savanna in which they evolved, since they occupy habitats that vary enormously in terms of temperature, humidity, terrain and vegetation, available foodstuffs and building materials, and dominant predators. More surprisingly, the social conditions in which they live are dramatically different too. The latter change has happened much more recently and much more suddenly. For almost all of their existence, including during the time that they were fanning out from Africa to other continents, human beings have lived in bands numbering from a few dozen to a maximum of a few hundred individuals, and have survived through hunting and gathering. These individuals would have known each other fairly well, and many of them (in particular the men) would have had reasonably close genetic ties. At the beginning of the 21st century, however, around half of the world's population lives in urban areas, a proportion likely

to rise to 60 per cent by 2030, and around 40 per cent of these live in agglomerations of more than one million inhabitants (United Nations, 1999).

Many interesting questions are raised by this development, of which the most puzzling concern how human beings have managed to sustain a complex web of cooperation, such as that which underpins the sophisticated modern division of labour, between individuals who do not have ties of kinship. The goods that are consumed by the modern urban household are manufactured in many different stages by different people who have no relation to one another and may not even live in the same country. The theory of kin selection (Hamilton, 1964) explains the evolution of cooperation among genetically related individuals, which is widespread in the animal kingdom (most famously among the social insects). But it predicts fierce rivalry among unrelated individuals, especially among males under conditions of strong sexual selection. With a few unimportant exceptions, significant cooperation among unrelated individuals has never evolved in any species other than man. Particularly puzzling is cooperation among strangers, which is the foundation of modern urban life. There is evidence that, while hunter-gatherer bands were mostly close knit and highly cooperative, encounters between strangers for much of human evolution have been accompanied by serious, often lethal violence. Human psychology has therefore been powerfully shaped by the fact that for much of our past we were one another's most dangerous predators (Sterelny 2003). But most human beings now encounter strangers in their thousands every day without giving the matter a second thought, and even in the world's more dysfunctional cities they run a risk of violence that is far lower than it was during the whole of human prehistory. Deaths by violence average a little over one per cent of all deaths in the world as a whole, whereas in prehistoric times they are estimated to have averaged between 10 and 40 per cent of all fatalities (WHO, 2002; Keeley, 1996).

How has this remarkable transformation of human social existence come about?

A look at the dating of these changes makes clear that the answer does not lie in a change in the genetic basis of human social psychology, but rather in the flexible adaptation of an existing psychology to a new social environment. DNA and archaeological evidence both suggest that the basic genetic architecture of the mind of modern man was in place many tens and possibly some hundreds of millennia ago. Yet hunting and gathering in relatively small bands remained universal until around 10,000 years ago, and the first large-scale agricultural civilizations did not emerge until around 5,000 years later. Modern human beings are navigating in their social lives with instruments that evolved to guide them in a quite different world.

A further puzzle about human social life is that the primate species from which we evolved – including the great apes, our cousins – live in bands governed by strong status hierarchies. Modern human social life is no less governed by strong inequalities of rank, status and access to economic resources as well as to intangible goods such as esteem. However, hunter-gatherer communities in the late Paleolithic, at a stage intermediate between hunting and gathering, appear to have been fairly egalitarian in the distribution of both resources and esteem, at least between individuals of the same sex (it is likely that relations between the sexes were more unfavourable to females than among our closest primate relatives, if only because human females were more dependent on males for both food and protection than is the case among chimpanzees and bonobos). How did human beings achieve such a degree of equality, and why did they lose it again?

The answer to these puzzles turns crucially on our understanding of the first agricultural revolution, which spread at a remarkable pace around the world, and which obliged human beings to become largely sedentary, encouraging them in the process to move into villages and towns for protection. It also

enabled the production and storage of a surplus over subsistence that could be devoted to other economic ends, spurring the division of labour and the growth of complex and hierarchical civilizations. Three main questions stand out: first, which of our mental capacities that had evolved before the agricultural revolution was to prove most important in shaping how human beings responded to that dramatic development? Second, what caused the agricultural revolution, and why did it spread so fast? And third, what were its consequences for human social life?

The mental capacities that mark *homo sapiens* out from our ancestors and cousins in the hominid line have been the subject of much debate, many aspects of which remain unresolved. It seems likely, though, that they included most or all of the following elements: a capacity for symbolic thought, the ability to contemplate and refer to absent or invisible objects and events, an enhanced concern for the future and one's own place in that future, a ‘theory of mind’ that enabled greatly enhanced prediction of the behaviour of other people, a sophisticated ability to detect cheating on the part of those others, and a greatly enhanced capacity to imitate their behaviour in a flexible and creative way (Cosmides and Tooby, 1992; Deacon, 1997; Tomasello 1999). Steven Mithen (1996) has even argued that only our own species has the capacity for consciousness in a proper sense, and has offered an intriguing theory of its evolution. For our purposes the crucial point is that these capacities would have been the very ones that, as described in the literature on cooperation in repeated games, enable human beings to cooperate even in the presence of conflicting interests.

The capacities to represent and care for the future, to predict how the behaviour of others may respond to our own, to respond appropriately to their trustworthiness or dishonesty, and to learn from the successes and mistakes of others – all these would no doubt have been highly useful for undertaking cooperatively the increasingly complex challenges of hunter-gatherer existence. Once in place they could then contribute to Renaissance statecraft, higher mathematics and running a 21st-century corporation, among other things. In addition, the ability to represent absent or invisible objects and events would have greatly enhanced the strategy space for would-be cooperators. If I know that by stealing a rival's food I risk not only his own retaliation (which might be restricted by his limited information or physical strength, or by the fact that since he is a stranger he is likely never to see me again), but also that of Mr Plod, Inspector Maigret and Judge Jeffreys, I shall have much more to lose. Other primates have sophisticated strategies of peacemaking (De Waal, 1989) but only limited means with which to enforce them. Our own species’ mental capacities enabled the invention of much more ingenious institutions of enforcement than had ever been available to hunter gatherers.

There is more controversy, however, over whether large-scale cooperation required evolutionary developments in the affective as well as the cognitive components of human psychology. It is argued by many that cooperation requires human beings to display ‘other-regarding preferences’, which depart from those that would maximize an individual's enlightened self-interest (interpreted as inclusive fitness according to the Hamiltonian model of kin selection). Specifically, such other-regarding preferences must include ‘strong reciprocity’ – a preference for repaying cooperation with cooperation, and cheating with revenge, even when this is not what the calculus of self-interest would require (see Henrich et al., 2004; Seabright, 2004). These authors claim that purely self-interested behaviour, even of the sophisticated kind described in the repeated game literature, would not have permitted complex cooperation because of the problems of limited observability and consequent mistakes in the implementation of retaliation strategies. Conversely, a small amount of strong reciprocity, even among a

subset of the relevant population, can go a long way in reinforcing cooperative behaviour (but see Binmore, 2005; Gintis et al., 2006). How such preferences could have evolved remains an open question, with some favouring a version of group selection (Gintis, 2000), and others preferring a form of signalling, in which the presence of other-regarding preferences made individuals more attractive as partners in cooperative activities (Frank, 1988).

However this controversy is resolved, it seems indisputable that communities of human hunter-gatherers were governed by strong cooperative norms, held in place by some combination of kin altruism, mutual monitoring under repeated interaction, and other-regarding preferences. Boehm (1999) has argued that these communities were more egalitarian (among males) than any before or since. This was not because humans had lost the strong sense of rivalry, including status rivalry, displayed by other primate species, but because the strong competitiveness of individual motivation was held in check by social mechanisms that retaliated against overweening displays of power or arrogance by any successful individual. Under the circumstances of hunting and gathering, great disparities of wealth or status were neither possible (since mobility precluded storage) nor desirable (since hunting required too much flexibility to be undertaken by the unwilling or the enslaved).

But strong community solidarity coexisted with violent inter-community rivalries. Although it used to be believed that hunter-gather communities were inherently peaceful, this is now known to be a myth (Ember, 1978; LeBlanc, 2003). Though trading links existed between different communities (including for the exchange of marriageable women), encounters between strangers or historic rivals were frequently violent, much as they are known to be often violent when groups of foraging chimpanzees of unequal strength encounter each other by chance in the wild (Ghiglieri, 1999).

How did all this change? Beginning around ten thousand years ago, agriculture was independently invented in at least seven different places (Anatolia, Mexico, the Andes of South America, northern China, southern China, the eastern United States, and in sub-Saharan Africa at least once and possibly up to four times; see Richerson, Boyd and Bettinger, 2001). The techniques of agriculture spread rapidly around the world (Bellwood, 2005), not simply by emulation but by the migration of the farmers themselves (Cavalli-Sforza, Menozzi and Piazza, 1994). It was all the more surprising that agriculture should catch on so fast because studies of the bones and teeth of some of the earliest agricultural communities of the Near East show that farmers had worse health, due to poorer nutrition, than the hunter-gatherers who preceded them (Cohen and Armelagos, 1984). Increases in agricultural productivity in later millennia more than made up for this eventually, but even so the puzzle remains: what prompted agriculture to be adopted so quickly and often within a comparatively short space of time, if it did not achieve the one thing that a new agricultural technique surely ought to achieve – to leave people better fed than they were before?

Explanations for the paradox have included the depletion of game, which lowered the productivity of hunting (see hunting and gathering economies), and the fact that agriculture once adopted led to population growth and crowding, thereby reducing food availability and increasing disease (Bar-Yosef and Belfer-Cohen, 1989; Robson, 2005). Consistent with these views, and adding force to the view that agriculture might have been irresistible even if disadvantageous to the health and nutrition of its adopters, is the idea that sedentarism significantly increased the effort and the resources human societies had to devote to defence (Seabright, 2006).

Those who are sedentary are also vulnerable. When enemies attack, farmers have much more to lose than hunter-gatherers, who can melt into the forest without losing houses, chattels and stores of food. So farmers need to spend time, energy and resources defending themselves – building walls, manning watchtowers, guarding herds, patrolling fields. This means less time and energy, fewer resources, devoted to making food. The greater productivity of the hours they spend growing and raising food could even be outweighed by the greater time they must spend defending themselves and the food they have grown – meaning that they produce less food in all.

But why should the first farmers have adopted agriculture at all? And why should this new technology have spread with such rapidity? Stunted farmers would hardly have been a good advertisement to their hunter-gatherer neighbours of the qualities of their new wonder diet. What is needed is an account that explains how agricultural adoption could have been individually rational even if perhaps collectively self-defeating, at least in the short run.

Agriculture dramatically raised the advantages to mankind of banding together for self-defence. Once constrained by a sedentary lifestyle and unable any longer to play hide-and-seek with its enemies, a large group is much more secure than its members could be in multiple smaller groups. But once the first farmers began to invest systematically in defence, they became a threat to their neighbours, including communities who were on the margins of adopting agriculture themselves. There is no such thing as a purely defensive technology. Even walls around a town can make it easier for attacking parties to travel out to raid nearby communities in the knowledge they have a secure retreat. The club that prehistoric man used to ward off attackers was the same club he used to attack others. Once a community has invested in even a modest army, whether of mercenaries or of its own citizens, the temptation to encourage that army to earn its keep by preying on weaker neighbours can become overwhelming. So, even if the first farming communities were not necessarily any better off than they would have been if no one had adopted agriculture, once the process had started many communities had an interest in joining in. These interactions could lead each to act ineluctably against the collective interests of all. It is a logic well known from the theory of contests (Becker, 1983; Hirshleifer, 1989).

However, the necessity of self-defence was also in time the mother of an astonishing array of technological and institutional mechanisms for keeping the peace and encouraging social cooperation, albeit much more effectively within communities than between them. Many of these mechanisms were subject to significant economies of scale, which encouraged the growth of cities even before their more subtle effects on economic development had been remarked (by Adam Smith among others). They led also to the accumulation of wealth, status and power by a minority of individuals within society who had access to land or capital, or to control of the means of inflicting physical violence. Added to the fact that agriculture could be carried out by slaves under constant surveillance, as hunting and gathering could not, this led to a dramatic increase in inequality. Almost no societies did not enslave others at some time in their history, with slavery becoming more likely the wealthier the society concerned, at least until it became wealthy enough to afford to take a stand against slavery on principle (Nieboer, 1900; Fogel and Engerman, 1974).

The institutions that now keep the peace in an urban environment are extraordinarily subtle, as the work of Jane Jacobs (1961) has notably emphasized. The police and courts are but the apex of an informal structure of eyes and ears that depends on the willing participation of citizens in a neighbourhood. Formal authority alone can never establish order, as Raymond Chandler recognized when in 1950 he wrote of a ‘world in which gangsters can rule nations and can almost rule cities’. The historian Peter

Hall (1998) has also noted that the characteristics that turn some cities into crucibles of artistic creativity and economic innovation depend on subtle networks of interaction that are impossible to plan in detail. They are an organic outgrowth of human beings' acquired capacity to build trust with strangers in a daily multitude of individually insignificant but collectively remarkable encounters.

In these and other ways the consequences of the developments in human psychology that permitted the adoption of agriculture were momentous for human life. A long-standing literature in political theory, going back to Ibn Khaldun (1377) and excellently discussed by Ernest Gellner (1994), considers the need to raise a surplus for defence as constituting the foundation of the division of labour and as giving rise to some of the most intractable problems of political organization. It can be said, therefore, to be at the root of both the most remarkable intellectual and economic achievements of human society and of its most deplorable cruelties and excesses.

See Also

- hunting and gathering economies

Bibliography

Bar-Yosef, O. and Belfer-Cohen, A. 1989. The origins of sedentism and farming communities in the Levant. *Journal of World Prehistory* 3, 447–97.

Becker, G. 1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 98, 371–400.

Bellwood, P. 2005. *First Farmers*. Oxford: Blackwell.

Binmore, K. 2005. *Natural Justice*. Oxford: Oxford University Press.

Boehm, C. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

Cavalli-Sforza, L., Menozzi, P. and Piazza, A. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Cohen, M.N. and Armelagos, G.J. 1984. *Paleopathology at the Origins of Agriculture*. New York: Academic Press.

Cosmides, L. and Tooby, J. 1992. Cognitive adaptations for social exchange. In *The adapted mind*, ed. J. Barkow, L. Cosmides and J. Tooby. New York: Oxford University Press.

Deacon, T. 1997. *The Symbolic Species: The Co-Evolution of Language and the Human Brain*. London:

Allen Lane.

De Waal, F. 1989. *Peacemaking Among Primates*. Cambridge, MA: Harvard University Press.

Ember, C. 1978. Myths about hunter-gatherers. *Ethnology* 17, 439–48.

Fogel, R. and Engerman, S. 1974. *Time on the Cross: The Economics of American Negro Slavery*. New York: Little Brown.

Frank, R. 1988. *Passions Within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton.

Gellner, E. 1994. *Conditions of Liberty: Civil Society and its Rivals*. London: Hamish Hamilton.

Ghiglieri, M. 1999. *The Dark Side of Man: Tracing the Origins of Male Violence*. Cambridge, MA: Perseus Publishing.

Gintis, H. 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 213, 103–19.

Gintis, H. et al. 2006. Symposium on Ken Binmore's natural justice. *Politics, Philosophy and Economics* 5(1).

Hall, P. 1998. *Cities in Civilisation*. London: Weidenfeld & Nicolson.

Hamilton, W. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7, 1–52.

Henrich, J. et al. 2004. *Foundations of Human Sociality*. Oxford: Oxford University Press.

Hirshleifer, J. 1989. Conflict and rent-seeking success functions: ratio vs. difference models of relative success. *Public Choice* 63, 101–12.

Ibn Khaldun, A.Z. 1377. *The Muqadimmah*, trans. F. Rosenthal. Princeton, NJ: Princeton University Press. 1969.

Jacobs, J. 1961. *The Death and Life of Great American Cities*. New York: Vintage Books, 1992.

Keeley, L. 1996. *War before Civilization: The Myth of the Peaceful Savage*. Oxford: Oxford University Press.

LeBlanc, S. 2003. *Constant Battles*. New York: St Martin's Press.

Mithen, S. 1996. *The Prehistory of the Mind*. London: Thames & Hudson.

Nieboer, H.J. 1900. *Slavery as an Industrial System*. New York: Burt Franklin, 1971.

Richerson, P., Boyd, R. and Bettinger, R. 2001. Was agriculture impossible during the pleistocene but mandatory during the holocene? A climate change hypothesis. *American Antiquity* 66, 387–411.

Robson, A. 2005. A bioeconomic view of the neolithic and recent demographic transitions. Mimeo, Simon Fraser University.

Seabright, P. 2004. *The Company of Strangers: A Natural History of Economic Life*. Princeton: Princeton University Press.

Seabright, P. 2006. Warfare and the multiple adoption of agriculture after the last ice age. Discussion paper, Centre for Economic Policy Research.

Sterelny, K. 2003. *Thought in a Hostile World: The Evolution of Human Cognition*. Oxford: Blackwell.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

United Nations. 1999. *World Urbanization Prospects: The 1999 Revision*. New York: Population Division, United Nations.

WHO (World Health Organization). 2002. *The World Health Report 2002*. Statistical Annex Table 2. Online. Available at http://www.who.int/whr/2002/en/annex_table2.xls, accessed 28 November 2006.

How to cite this article

Seabright, Paul. "hunters, gatherers, cities and evolution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000163> doi:10.1057/9780230226203.0757

The New Palgrave Dictionary of Economics Online

hunting and gathering economies

Vernon L. Smith
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agricultural revolution; commodity money; human capital; hunting and gathering economies; leisure; multilateral contracting

Article

Men and women (*Homo erectus*) who were culturally and biologically distinguishable from other hominoids have lived on the planet Earth for about 1.6 million years (Pilbeam, 1984). It is likely that the biological changes since that time form a microevolutionary continuum: archaic *H. sapiens*, including the Neanderthal, appeared 125,000 years ago and anatomically modern *H. sapiens* appeared about 45,000 years ago. The record suggests that *H. erectus* fabricated and used tools, and his use of fire may have begun by 700,000 years ago. The changes identified in the prehistoric period appear only to distinguish less advanced from more advanced stone age technology. Consequently, the dominating message seems to be that over almost the whole of man's epoch on earth he lived successfully as an exceptionally well-adapted hunter. It is only recently, in the last 8,000–10,000 years (less than one per cent of his time on Earth), that man abandoned the nomadic life of the hunter to begin growing crops, husbanding domesticated animals, and living in villages. It is difficult to exaggerate the importance of this agricultural or first economic revolution (North and Thomas, 1977) in understanding who we are, and what we have become. Once man opted for the farmer–herder way of life it was but a short step to mankind's much more sophisticated development of specialization and exchange, greatly enlarged production surpluses, the emergence of the state, and finally the Industrial Revolution. Our direct knowledge of early man is confined to the record of the durables he left behind. Yet when combined with anthropological evidence from the study of recent hunter–gatherer economies the evidence can be interpreted as demonstrating that all the ingredients associated with the modern wealth of nations – investment in human capital, specialization and exchange, the development of property right or contracting institutions, even environmental ‘damage’ – had their development in the course of that vast prehistorical, pre-agricultural, period.
What accounts for this sudden abandonment of the nomadic hunting life? We do not know for we have

no direct observations on the transformation from hunting to agriculture. This transformation is perhaps the pre-eminent scientific mystery, since all of that which we have called civilization, all the great achievements of industry, science, art and literature stem from that momentous event within the last few minutes of man's day on Earth. Yet there are common factors that dominated the evolution of man from his earliest form to modern *H. sapiens*, and his primary intellectual and social development, which suggest an underground continuity between the pre-agricultural, Paleolithic hunting period, and the agricultural and subsequent periods.

Man the hunter – gatherer

There are many widely held beliefs concerning the characteristic features of the hunter–gatherer way of life that stretch back several hundred years in academic writings, and persist as part of the folklore of contemporary man's misperception of his own prehistoric past; until recently these beliefs dominated even the anthropological view of hunter–gatherer ‘subsistence’. These beliefs tend to obscure the striking continuity in man's ability to respond to changes in his environment by substituting new inputs (labour, capital and knowledge) for old, and develop new products to replace the old when effort prices were altered by the environment.

Ever since Hobbes there has prevailed the perception that life in the state of nature was ‘solitary, poor, nasty, brutish and short’. A more accurate representation (if not strictly correct in all aboriginal societies) would argue that the hunter culture was the original affluent society (Lee and DeVore, 1968).

Extensive earlier data on extant hunter–gatherers show that with rare exceptions (such as the Netsilik Eskimos) their food base was at minimum reliable, at best very abundant. The African Kung Bushman inhabited the semi-arid north-west region of the Kalahari Desert, an inhospitable environment, characterized by drought every second or third year. These conditions had served more to isolate the Kung from their agricultural neighbours than to condemn them to a brutish existence. Adults typically worked 12–19 hours per week in getting food. As with all such societies, for the most part the women gathered, the men hunted. The caloric-protein returns exceeded several measures of nutritional adequacy. Gathering was the more reliable and productive activity with women producing over twice as many edible calories per hour as men. Both men and women bought leisure with this work schedule – resting, visiting, entertaining and (for the men) trance dancing. About 40 per cent of the population were children, unmarried young adults (15–25 years of age) or elderly (over 60 years of age), who did not contribute to the food supply and were not pressured to contribute.

A comparable macroeconomic picture applied to the Hazda in Tanzania. Large and small animals were numerous and all – with the exception of the elephant – were hunted and eaten by Hazda. Hunting was the speciality of men and boys, conducted as an individual pursuit that relied primarily on poisoned arrows. The Hazda spent on average no more than two hours a day hunting. The principal leisure activity of the men was gambling, which consumed more time than hunting.

Other hunting (or fishing) peoples of Africa, Australia, the Pacific Northwest, Alaska, Malaya and Canada have shown comparably effective adaptation to this form of livelihood. Malnutrition, starvation and chronic diseases were rare or infrequent, although accidental death was high in certain cases such as the Eskimo.

The argument that life in the Paleolithic must have been intolerably harsh is simply not borne out by the many ethnographic studies of extant hunting societies in the past century. With few exceptions such

societies have fared well, and did not leap to embrace the agricultural or pastoral pursuits of their neighbours. Whether life in the Paleolithic mirrored this modern experience cannot be known with any assurance, but certainly there is no support for the proposition that hunting, per se, means an intolerably harsh existence. In fact, the Paleolithic hunting economy had demonstrably high survival value in a world far more plentifully endowed with game than has existed since the great megafaunal extinctions of the late Pleistocene, and therefore a world which might indeed have been marked by numerous original affluent economies.

Although it is natural to suppose that man's uniqueness derived from his intellectual superiority, what is more likely is that man's physical superiority was also important in giving him a superpredator's advantage over other species. His endowment of physical human capital would probably have been of significance even in the absence of his investment in tools and the human capital required to produce and use tools. As noted by J.B.S. Haldane, only man can swim a mile, walk twenty, and then climb a tree. Add to this observation the four-minute mile, unsurpassed long-distance endurance running, the ability to carry loads in excess of body weight, high altitude performance, American Indian capacity literally to run down a horse or deer by pacing the animal, the incredible accomplishments of acrobats and gymnasts, and finally the finger agility and coordination required to milk a cow, and you are left with the physical portrait of an astonishingly superior species. It appears that man's basic foundation of physical superiority was laid by his upright stance, to which of course the addition of knowledge made him truly formidable, even in the presence of the various giant proboscidea (mastodon, mammoth, elephant) which early man did not hesitate to hunt and to kill on three continents.

The idea that primitive man was too puny and too few in number to have had a significant influence on his environment underestimates man's uniqueness as a tool using, fire using, highly mobile species who, with minor exceptions (Madagascar, New Zealand and Antarctica), had populated the world by 8000•bc. The archaeological record suggests that man was a big game hunter par excellence. He hunted mammoth, mastodon, horse, bison, camel, sloth, reindeer, shrub oxen, red deer, aurochs (wild cattle), and other large mammals, for perhaps a minimum of 30,000–40,000 years, ceasing only with the great megafaunal extinctions throughout much of the world some 8,000–12,000 years ago. Paul Martin (1967) has argued the case for the overkill hypothesis that man was a significant causative factor in these extinctions. Essentially, the argument is that the alternatives to overkill, principally the climate hypothesis, fail to account for the worldwide pattern of these extinctions which appear to have begun in Africa and perhaps southeast Asia 40,000–50,000 years ago, spread north through Eurasia 11,000–13,000 years ago, jumped to Australia perhaps 13,000 years ago, and entered North America in the last 11,000 years, followed by South America 10,000 years before the present. The most recent extinctions are in New Zealand (numerous species of flightless moa birds) 900 years ago and in Madagascar 800 years ago, shortly after the remarkably late migration of man to those islands.

Man's use of fire as a tool in the management and control of natural resources must be counted as having a profound effect on his ecological environment. Numerous authors who have studied patterns of land burning by primitive peoples have concluded that most of the greatest grasslands of the world represent fire-vegetation that is man-made (see Heizer, 1955, for a summary). Where tree growth is strongly favoured by climatic conditions, regular burning will select for certain species of tree such as the pine stands of southern New York and to the West, which have been attributed to Indian burning. Contemporary man's attempts to prevent fires, which today are almost entirely caused by lightning, has probably produced far more ecological damage than the controlled use of fire that has characterized

aboriginal cultures. Recurrent fire prevents the accumulation of brush which then fuels the holocaust wildfire that destroys all forest vegetation.

A third source of ecological change produced by primitive peoples was their transportation of seed, in their migrations as hunter–gatherers, which introduced numerous botanical exotics into new regions. Archaeologists have frequently observed the association of various plants with ancient campsites and dwellings. For example, the wide distribution of wild squash, gathered for its seed, appears to be associated with man. The introduction of exotics can and has produced significant environmental changes in modern times, but the phenomenon has ancient origins and may have been considerably more disruptive as the first men moved from one ‘pristine natural’ region to another.

Success as a hunter–gatherer requires human capital usually associated only with agricultural and industrial man: learning, knowledge transfer, tool development and social organization. Comprehensive studies of the aboriginal use of fire for game and plant management show clearly that primitive men demonstrated extensive knowledge of the reproductive cycles of shrubs and herbaceous plants, and used fire to encourage the growth and flowering of the plants used in gathering, and to discourage the growth of undesirable plants (Lewis, 1973). This required one to know when, where, how and with what frequency to apply the important tool of controlled burning for managing the resources that allow gathering to make an efficient, productive and sustainable contribution to living. Primitive men knew that the growing season can be advanced by spring burns designed to warm the earth, that in dry weather fires should be set at the top of hills to prevent wild fires, but in damp air they should be set in depressions to avoid being extinguished, that the burning of underbrush aided the growth of the oak whose acorns were eaten and attracted moose who avoid underbrush, and that deer and other animals congregate to feed on the proliferation of tender new plants that sprout following a fall burn.

To live by hunting is to be committed to an intellectually and physically demanding activity that requires technology, skill, social organization, some division of labour, knowledge of animal behaviour, the habit of close observation, inventiveness, problem solving, risk bearing, and high motivation, since the rewards are great and the penalties severe. Such exceptional demands could have been highly selective in man's long evolution, and disciplined the development of the intellectual and genetic equipment that facilitated his subsequent rapid creation of modern civilization. This natural selection could have been intensified by the widespread practice among aboriginals of rewarding superior hunters with many wives. It was as a hunter that man learned to learn. In particular he understood that young boys must be imbued with the habit of goal-oriented observations, and with knowledge of animal behaviour and anatomy. To know that many ungulates travel in an arc meant that tracking success could be improved by transversing the chord. Knowledge of animal behaviour was a substitute for weapon development. Even the weapons of the later pre-agricultural period (spears, bow and arrow, harpoon) required the hunter to approach the prey within ten yards for a best shot. This might require hours crouched on the ground waiting for a shift in the wind, for just the right change in the animal's position, or for the mammoth to get deeper into the bog in a watering hole. The weapons changed with shifts to new prey. Thus the Clovis fluted point, widely distributed throughout North America, was used to kill mammoth and mastodon 11,000–12,000 years ago. The Folsom point was then developed and used to kill the large, now extinct *Bison antiquus*, which then gave way to the Scottsbluff point associated with the killing of the slightly smaller, now extinct *Bison occidentalis* (Haynes, 1964; Wheat, 1967). These observations suggest high specialization which required new forms of human and physical capital to meet the specialized demands of new prey.

The organizational requirements of the hunt are illustrated at the Olsen–Chubbuck site in Colorado, where the excavated remains of bones and projectile points of the Scottsbluff design show that about 8,500 years ago some 200 *Bison occidentalis* were stampeded into an arroyo 5–7 feet deep. Armed hunters in the arroyo on each side of the stampede then slaughtered the injured or escaping animals with their weapons (Wheat, 1967).

Primitive man has often been modelled as ‘cultural’ not ‘economic’ man, but the power and importance of the opportunity cost principle in conditioning the choice of all peoples was perceptively stated by the Kung Bushman, who, when asked why he had not turned to agriculture, replied, ‘Why should we plant, when there are so many mongongo nuts in the world?’ (Lee and DeVore, 1968, p. 33). This Bushman, I would hypothesize, stated the answer to the scientific question: why did man the hunter tend to abandon that which appeared to serve him so well for 1.6 million years and to which he seems to have adapted ever more successfully, as indicated by the growing complexity of his tools and weapons as he evolved from *H. erectus* to anatomically modern *H. sapiens*? Man would not have given up the hunter–gatherer life had there not been a change in the terms of trade between man and nature that made the hunting way of life more costly relative to agriculture. This *hypothesis* does not leave ‘culture’ out of the equation. Thus to describe hunter–gatherers as directly seeking the cultural goal of prestige does not contradict the hypothesis that man, like nature, ever economizes. Attaching prestige to the hunt may simply be an astute means of advertising, teaching and propagating the discovery that hunting and its attendant technology is the best means of livelihood, with the result that each new generation does not have to rediscover this knowledge. Myths of the great hunter, of great rewards, of great penalties for lost technique, of killing the goose that lays golden eggs are part of the oral tradition by which the economy preserves this human capital.

The hypothesis that the agricultural revolution was due to a major decrease in the productivity of labour in hunting–gathering relative to agriculture (Smith, 1975; North and Thomas, 1977) is consistent with the observations that this cultural shift (*a*) occurred at different times in different parts of the world, with small aboriginal hunting enclaves still in existence, and (*b*) did not occur once and for all in every such tribe. With respect to (*a*), the great wave of terrestrial animal extinctions occurred over a period of several thousand years, and therefore the relative increase in the cost of hunting struck different regions at different times. Also different peoples in different environments with different opportunity costs would be expected to provide different mechanisms of adaptation, with some persisting as gatherers and small game hunters, and others turning to or perhaps persisting as fisherman (for example, the Aleutian Eskimos and the Pacific Northwest Indians) in regions unsuitable for agriculture. With respect to (*b*) the reintroduction of the horse in North America by the Spanish (in the hardy form of *Equus caballus* just 8,000 years after other members of the genus became extinct in the Americas) had a major modifying impact on the economy of the plains Indians. In the northern plains the ‘fighting’ Cheyenne, as they were later to be termed by the Europeans, and the Arapahoe quickly abandoned their villages along with their pottery arts and horticulture to become nomadic Bison hunters (see the references in Smith, 1975). Apparently, agricultural productivity was dominated by the enormous increase in the bison harvest made possible by a technological change that combined the horse with the bow and arrow. To the south, where the growing season was longer and the climate more favourable, the Pawnee preserved their maize agriculture when they turned to Bison hunting, creating a mixed agricultural–hunting economy. The south-western Apache, reported by Coronado in 1541 to be subsisting as bison hunters, simply adapted the horse to their pre-existing hunter culture. The vast bison-hide tepee encampments witnessed by the

first Europeans to cross the plains were already the product of a technologically transformed native American, many of whom had only recently abandoned their agricultural economies.

Pleistocene extinctions and the rise of agriculture

Here then is a model of the epoch of man: he arrives 1.6 million years ago as a hunter among hunters, but distinguishable in terms of his human capital endowment and his ability to invest in the development of human and physical capital. His tools become more complex and knowledge of the use of fire, perhaps his most significant tool, is added to his stock of human capital. There is a gradual improvement in weapons technology – clubs, stones, stone axes, spears, stone projectile points, the atlatl (which applies the leverage principle) and, in the late pre-agricultural period, the bow (which combines the leverage principle with temporary storage of energy for increased mechanical advantage). The combination of his physical superiority, tools and fire make him a superpredator without equal. At some unknown point this success brings relative affluence, and the important commodity ‘leisure’, which might have contributed to the development of language and other forms of investment in human and physical capital.

Although *H. erectus* and archaic *H. sapiens* were advanced hunters who apparently spread from Africa to Eurasia and Asia, it remained for modern *H. sapiens* to establish himself as a big game hunter par excellence, who populated most of the world by 8000 bc. Associated with this radiation is recorded a wave of extinction that was largely confined to the large terrestrial herbivores and their dependent carnivores and scavengers. (Other extinction episodes in the Earth's history had affected plants and marine life, as well as animals.) There appear to be no continents or islands where these accelerated late Pleistocene extinctions precede man's invasion (Martin, 1967). Whether men caused these extinctions cannot be known with any certainty, but Martin's overkill hypothesis is clearly consistent with a common property resource model of the economics of megaherbivore hunting (Smith, 1975). Thus the large gregarious animals that suffered extinction provided low search cost and high kill value. The lack of appropriation (branding or domestication) provided disincentives for conservation and sustained yield harvesting. There are numerous stampede kill sites (pitfalls and cliffs) in Russia, Europe and North America that indicate wastage killing in excess of immediate butchering requirements. Considering the complex of suitabilities necessary for the remains of such a site to have been preserved, it is likely that only the tip of such phenomena has been observed. Finally, the slow growth, long lives and long maturation of the megafauna made them more vulnerable than other animals to extinction by hunting pressure.

But our model of economizing man need not sustain such a controversial hypothesis as overkill. It is sufficient that the easy, valuable prey disappeared, precipitating a decline in the productivity of hunting. Substitution is to be expected, given a change in relative effort ‘prices’. Hence, it is in this late pre-agricultural period that the archaeological record shows the appearance of bows and arrows, seed grinding stones, boiling vessels, boats, more advanced houses, even ‘villages’ (probably clan group abodes), animal-drawn sledges and the dog (almost certainly derived from domesticating the wolf). These developments strongly suggest the substitution of new tools and techniques for the old, which allowed new products to substitute for the loss of big game that could be harvested by stampeding and/or dispatch with thrusting or throwing weapons. Now the bow and arrow becomes adaptive, and gathering becomes more crucial to maintaining overall food productivity. Whereas formerly, gathering

emphasized seeds and plants that could be eaten on the run, now some of the seeds gathered were inedible without grinding, soaking, boiling. All this paraphernalia implies more sedentary, less nomadic, hunting and gathering.

Hence the incentive to invest in facilities such as utensils, sledges and houses. The boat allows fishing, sealing and whaling. The wolf, also characterized by its capacity to apply organization to the hunt, is now enlisted with man in the hunting of the game still available. Perhaps more important, the wolf may have been the model for domesticating other animals since the dog was a companion and pet that enabled children to learn about domesticated animal behaviour. With a more sedentary life, and the accumulation of personal property and real estate, would come more complex property right and contracting arrangements. The study of pre-colonial aboriginal societies in Northwest America and Melanesia reveals the existence of elaborate *multilateral contracting* arrangements in the form of ‘ceremonial exchanges’ such as the potlatch, kula, moka and abutu (Dalton, 1977). The use of valuables or commodity money (bracelets, pearl shells, cowries, young women) in these primitive societies was more complex than that of cash used in nation states with well-defined legal bases for exchange. These valuables not only bought other valuables in ordinary internal or external market exchange, they bought kinship ties with the exchange of women, military assistance when attacked, the right of refuge if invasion required the abandonment of homes, and emergency aid in times of poor harvest, hunting or fishing. In short they bought political stability, and a property right environment that made ordinary exchange and specialization possible. Property was owned by corporate descent lineages and included land, fishing sites, cemetery plots and livestock, but, interestingly, also public goods like crests, names, dances, rituals and trade routes, that could be assigned to many groups or individuals. These practices, which characterize stateless hunter–gatherer aboriginals, demonstrate that the phenomenon of multilateral contracting (Williamson, 1983), so common to the market economy in nation states, has ancient origins which antedate the state and the agricultural revolution.

Man's long existence as a hunter had brought knowledge of animals; extinction brought a change in relative costs; gathering brought knowledge of seeds and eggs; life became more sedentary, with property, contracting and exchange becoming more important. Under these more stable conditions it was a short step for mankind to plant for harvest, and/or to husband some of the more docile game that had been hunted previously. With agriculture and herding came a more sophisticated development of the earlier hunter–gatherer institutions of contract, property, exchange and specialization; and ultimately the continuing industrial-communication revolution. But long before these sweeping changes can be seen the dim outline of continuity in the development of man's capacity to adapt by creating cheaper products and techniques to substitute for dearer ones.

See Also

- economic anthropology
- hunters, gatherers, cities and evolution
- population and agricultural growth

Bibliography

Dalton, G. 1977. Aboriginal economies in stateless societies: interaction spheres. In *Exchange Systems in Pre-History*, ed. J. Erickson and T. Earle. New York: Academic Press.

Haynes, C.V. 1964. Fluted projectile points: their age and dispersion. *Science* 145, 1408–13.

Heizer, R. 1955. Primitive man as an ecological factor. *Krober Anthropological Society Papers* 13, 1–31.

Lee, R.B. and DeVore, I. 1968. *Man the Hunter*. Chicago: Aldine.

Lewis, H. 1973. *Patterns of Indian Burning in California: Ecology and Ethnohistory*. Anthropological Papers No. 1. Romana, CA: Ballena Press.

Martin, P. 1967. Prehistoric overkill. In *Pleistocene Extinctions*, ed. P.S. Martin and H.E. Wright, Jr., New Haven: Yale University Press.

North, D.C. and Thomas, R.P. 1977. The first economic revolution. *Economic History Review* 30, 229–41.

Pilbeam, D. 1984. The descent of hominoids and hominids. *Scientific American* 250(3), 60–9.

Smith, V.L. 1975. The primitive hunter culture, pleistocene extinction, and the rise of agriculture. *Journal of Political Economy* 83, 727–55.

Wheat, J.B. 1967. A Paleo-Indian bison kill. *Scientific American* 216(1), 44–52.

Williamson, O. 1983. Credible commitments: using hostages to support exchange. *American Economic Review* 73, 519–40.

How to cite this article

Smith, Vernon L. "hunting and gathering economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000106> doi:10.1057/9780230226203.0758

The New Palgrave Dictionary of Economics Online

Huskisson, William (1770 – 1830)

Roy Green
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

currency; Huskisson, William; political economy; real bills doctrine

Article

Huskisson is better remembered for the manner of his death than for his not inconsiderable achievements as a statesman and economist. While it is true that he enjoyed ‘little success in public life compared with that which his rare abilities should have commanded’ (*Dictionary of National Biography*), there were few major debates which were not enhanced by his contribution. Huskisson first entered Parliament in 1796 and remained a member, with only one short break, for over 30 years. He served in the cabinet from 1823, and held a number of key government posts, including Secretary of the Treasury, President of the Board of Trade and Secretary of State for War and the Colonies. He figured prominently in the Bullion controversy and the subsequent discussion on the resumption of cash payments; and he initiated the process of tariff reform which was to culminate in the repeal of the Corn Laws.

His abilities may be gauged by the tributes paid by his contemporaries. It was said that ‘there is no man in Parliament, or perhaps out of it, so well versed in finance, commerce, trade or colonial matters’ (Charles Greville, in Melville, 1931, p. viii); and that ‘the knowledge of theory and practice were never possessed by any one in so high a degree’ (Kirkman Finlay, in Huskisson, 1831, I, p. 161; also Alexander Baring and Henry Brougham, ibid., pp. 120–1). Indeed, according to some observers, Huskisson might easily have become Chancellor of the Exchequer, but for his almost disingenuous loyalty to George Canning and the offence which he regularly caused to traditional Tory interests. These ‘failings’ earned him a remarkably fulsome tribute from J.S. Mill: ‘With the exception of Turgot, the history of the world does not perhaps afford another example of a minister steadfastly adhering to general principles in defiance of the clamours of the timid and interested of all parties …’ (*Westminster Review*, 1826, *cit*. Tucker introduction to Huskisson, 1830, p. xv).

Even his closest supporters, however, could not pretend that Huskisson was an eloquent speaker; to his everlasting shame, he was born and brought up outside London and the Home Counties. As a consequence, no doubt, he was ‘a wretched speaker with no command of words, with awkward motions,

and a most vulgar, uneducated accent' (Sir Egerton Brydges, *cit*. *Dictionary of National Biography*). Huskisson's interest in political economy began in Paris, where, as a young man, he moved in French liberal circles, and is said to have met Franklin and Jefferson. There, in 1790, he presented a paper on the currency to the monarchist 'Club of 1789'; once the French Government started issuing assignats, however, he resigned from the club and, shortly afterwards, returned to Britain. In 1810, Huskisson had an opportunity to make his mark on British financial policy; he did so in conjunction with Henry Thornton and Francis Horner in the Bullion Report, and then on his own in a pamphlet defending the report against its 'anti-bullionist' critics. This pamphlet, *The Question Concerning the Depreciation of our Currency* (1810), ran to several editions and drew praise not only from Ricardo, as might be expected, but also from the more critical Thomas Tooke (1838/57, IV, p. 98); its main target was the 'real bills doctrine' pleaded by the Bank of England directors as an adequate principle of limitation even when the currency was inconvertible. In the Parliamentary debates on the Bullion Report, Huskisson likened the views of the Bank directors to those of John Law, and made a strong case for the resumption of cash payments (Fetter, 1965, p. 43). After the passage of resumption legislation in 1819, however, Huskisson confessed to private doubts: 'The wheel of depreciation producing high prices, etc., was turning one way whereby many interests suffered and were ruined; to attempt to turn the wheel back, without some equitable adjustment … has always appeared to me madness' (Letter to J.C. Herries, 20 December 1829, *cit*. Melville, 1932, p. 312). The sharp decline in prices which followed resumption particularly affected agricultural products.

A Committee on Agriculture was formed in 1821 whose report – drafted mainly by Huskisson and Ricardo – accepted many of the arguments against the Act of 1819 but came down in favour of its retention. Thomas Attwood, after giving evidence to the Committee, wrote: 'The stupid landowners … are all as dull as beetles, whilst Huskisson and Ricardo are as sharp as *needles* and as active as bees' *(cit*., Ricardo, 1951/73, VIII, p. 370). A year later, Huskisson headed off Western's motion to reopen the issue with an amendment in the same terms as Montague's resolution of 1696, 'That this House will not alter the Standard of Gold or Silver, in fineness, weight, or denomination'.

During the 1820s, Huskisson became an effective spokesman for the manufacturing interest, defending 'with singular success and ability, the general principles of commercial freedom' (Tooke, 1838/57, V, p. 414). He took part in debates on the silk trade, agricultural protection, tax reform, shipping and the repeal of the Combination Acts; and he was almost alone in foreseeing the crisis of 1825, expressing concern as early as March 1822, 'that this universal Jobbery in Foreign Stock will turn out the most tremendous Bubble ever known' (Hudson Gurney, *cit*. Fetter, 1965, pp. 111–12). Having disregarded his warnings, the Bank of England directors sought to blame Huskisson for promoting the crisis:

> Such is the detestation in which he is held in the City that Ld L[iverpool] & Mr. Canning did not think it prudent to summon him to London till all the Cabinet were sent for &, in the discussions with the Bank, he is kept out of sight. He repays them with equal hatred …. (Mrs Arbuthnot, 17 December 1825, *cit*. Fetter, 1965, p. 117)

In June 1827 Huskisson, responding to a memorandum circulated by James Pennington, wrote of the need to 'prevent … those alterations of excitement and depression which have been attended with such alarming consequences to this country'. He went on:

> This, for a long time, has appeared to me one of the most important matters which can engage the attention of the Legislature and the Councils of this country. The subject is certainly intricate and complicated; but the too great facility of expansion at one time, and the too rapid contraction of paper credit (I speak of it in the largest sense) at another, is unquestionably an evil of the greatest magnitude. (*cit*. Fetter, 1965, p. 131; also Viner, 1937, p. 224)

Huskisson asked Pennington for suggestions as to how these fluctuations could be minimized, and Pennington submitted a second memorandum which was to form the basis of the ‘currency principle’. Huskisson resigned from the government in 1828 over a seemingly trivial but symbolic issue – the allocation of a parliamentary seat to a sparsely populated rural hundred, instead of a manufacturing town. He died soon afterwards in unusual, not to say bizarre, circumstances. On 15 September 1830, he attended the opening ceremony of the Manchester and Liverpool Railway:

> At that moment several engines were seen approaching along the rails between which Huskisson was standing. Everybody made for the carriages on the other line. Huskisson, by nature uncouth and hesitating in his motions, had a peculiar aptitude for accident … . On this occasion he lost his balance in clambering into the carriage and fell back upon the rails in front of the Dart, the advancing engine. It ran over his leg … He lingered in great agony for nine hours, but gave his last directions calmly and with care, expiring at 9 P.M. (*Dictionary of National Biography*)

That would be the end of the story but for a fine piece of detective work by G.S.L. Tucker and his assistant, Helen Bridge, who in 1976 established beyond reasonable doubt that the author of an anonymously published 1830 tract, *Essays on Political Economy*, was none other than William Huskisson. In addition to the circumstantial evidence of style and argument, the publisher's Commission Ledger was signed by a certain ‘George Robertson’, a name unknown to political economy at that time. It was then demonstrated by Detective Sergeant D.G. Stuckey of the Document Examination Unit, New South Wales Police, that the signature belonged not to ‘George Robertson’ at all but to Huskisson's half-brother, Thomas, with whom he was on close terms (Fay, 1951, pp. 300–1). Thomas Huskisson was a captain in the Royal Navy; and there is evidence that in return for career advancement (William Huskisson was treasurer of the Navy from 1823 to 1827), he would perform errands of this kind (ibid.). Although the *Essays* had a poorer reception than if they had appeared under Huskisson's own name, he presumably felt that he could not take the risk of further embarrassing the government with his forthright views. The *Essays* are basically Smithian in approach, and, in most respects, were already superseded by Ricardo's *Principles*. They do, however, propose some important financial reforms (Huskisson, 1830, pp. 149–51 and 152–3), repudiate the landowners' monopoly (*ibid.*, *p*. 255) and, most notably, anticipate J.S. Mill's concept of a ‘general glut’ (*ibid.*, pp. 448–52 and 454–5). Overall, they epitomize Huskisson's economic philosophy and were even cited approvingly by Marx (1867, p. 495n.); this philosophy was reflected clearly and consistently in a life of ceaseless activity: ‘Whatever ridicule might be attempted to be thrown on the science of political economy’, he said, ‘that science could not be discredited. It was the

result of general principles warranted by observation, and constituted the guide in the regulation of political measures' (Huskisson, 1831, II, p. 128).

Selected works

1830. *Essays on Political Economy*. Canberra: Australia National University, 1976.

1831. *The Speeches of the Right Honourable William Huskisson*. London: John Murray.

Bibliography

Fay, C.R. 1951. *Huskisson and his Age*. London: Longmans.

Fetter, F.W. 1965. *Development of British Monetary Orthodoxy, 1797–1875*. Reprinted, Fairfield: Kelley, 1978.

Marx, K. 1867. *Capital*, Vol. 1. Moscow: Progress Publishers, n.d.

Melville, L. 1931. *The Huskisson Papers*. London: Constable.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*, Ed. P. Sraffa. Cambridge: Cambridge University Press.

Tooke, T. 1838–57. *A History of Prices, and of the State of Circulation, from 1792 to 1856*. London: P. S. King, 1928.

Viner, J. 1937. *Studies in the Theory of International Trade*. London: George Allen & Unwin.

How to cite this article

Green, Roy. "Huskisson, William (1770–1830)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000107> doi:10.1057/9780230226203.1859

The New Palgrave Dictionary of Economics Online

Hutcheson, Francis (1694 – 1746)

Andrew Skinner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Holder of the Chair of Moral Philosophy at Glasgow University, Hutcheson, counted Adam Smith among his pupils. His moral philosophy resembled Smith's in emphasizing the role of sentiment, though Smith rejected his notion of an internal moral sense. Hutcheson's economic analysis embraced the division of labour, property, and money. His theory of value, which stressed the role of subjective judgement as a determinant of value in exchange, was influenced by Pufendorf, but Hutcheson went beyond Pufendorf (and foreshadowed Smith) in arguing that goods exchange at a rate that is in part determined by the quantity of labour embodied in them.

Keywords

barter; benevolence; Carmichael, G.; division of labour; Ferguson, A.; Hume, D.; Hutcheson, F.; Locke, J.; money; moral sense; natural rights; property; Pufendorf, S.; self-interest; Smith, A.; supply and demand; value in exchange; value theory; *Wealth of Nations*

Article

Biographical

Hutcheson was born on 8 August 1694. His father, John, was a Presbyterian minister in Armagh, Ireland, and Francis spent his early years at nearby Ballyrea. In 1702 Francis and his elder brother, Hans, went to live with their grandfather, Alexander Hutcheson, at Drumalig in order to further their schooling. At the age of 14 Francis moved to a small denominational academy at Killyleagh, County Down.

In 1711 Hutcheson matriculated at Glasgow University, where he was particularly influenced by Robert Simson (mathematics), Gerschom Carmichael (moral philosophy), Alexander Dunlop (Greek) and John Simpson (the 'heretical divine'). Hutcheson graduated in 1713 and embarked upon a course of study in

theology under Simpson's guidance.
Hutcheson was back in Ireland in 1719 when he was licensed as a probationary minister but moved to Dublin where he established an academy of which he remained head until 1730. His reputation established, Hutcheson was elected to the Chair of Moral Philosophy in Glasgow, succeeding Carmichael. It was as a lecturer that he made his mark, brilliant and stylish, using English rather than Latin. Hutcheson's career as author and teacher amply confirms Adam Smith's famous reference to the 'abilities and virtues of the never-to-be-forgotten' master.
Hutcheson lectured five days a week on natural religion, morals, jurisprudence, and government – an order which was to be followed by Adam Smith on his appointment to the Chair of Moral Philosophy in 1752. On three days he lectured on classical theories of morality, thus contributing (with Dunlop) to a revival of classical learning in Glasgow, which formed an important channel for stoic philosophy; a philosophy which was to have an important influence on Adam Smith. Hutcheson died on 8 August 1746 (his birthday) and was buried in St Mary's churchyard in Dublin.

Social order

Although this article is concerned primarily with Hutcheson's economic analysis it will be convenient to say a little regarding his ethical work.
Adam Smith identified two key questions which the moral philosopher must confront. First, wherein does virtue consist, and, secondly:

> how and by what means does it come to pass, that the mind prefers one tenor, of conduct to another, denominates the one right and the other wrong; considers the one as the object of approbation, honour and reward, and the other of blame, censure and punishment. (TMS, VII, i.2)

Hutcheson addressed both questions, identifying virtue with benevolence while explaining the processes of judgement in terms of a particular sense, the 'moral sense'. Smith was to reject Hutcheson's answer to the first question on the ground that while important, the emphasis on benevolence neglected the role of self-command and the 'inferior' virtue of prudence. In the same way, while welcoming his master's emphasis on sentiment rather than reason in explaining the means by which the mind forms judgements concerning what is fit and proper to be done or to be avoided, Smith rejected the notion of a special (internal) sense, the moral sense.
The common element evident in the work of Hutcheson, Hume and Smith is the emphasis on sentiment. But they also share another preoccupation, namely the attempt to explain the origins of social order; a crucially important element in the treatment, inter alia, of economic phenomena. The basic task was to explain how it was that a creature endowed with both self- and other-regarding propensities was fitted for the social state.
When we turn to Hutcheson it is to discover marked similarities with the work of his successor, especially in the context of his belief that 'We may see in our species, from the vary cradle, a constant propensity to action and motion' (*System*, I, p. 21). But in some respects the position is subtler than that stated by Smith. To begin with, Hutcheson argued that man has powers of perception which 'introduce into the mind all the materials of knowledge' and which are associated with 'acts of the

understanding' (*System*, I, p. 7). Acts of the understanding assist in the isolation of objects to be attained (for example, sources of pleasure) or to be avoided, and culminate in acts of will.

Acts of will, which may be calm or turbulent, were divided in turn into the selfish or the benevolent. Benevolent acts of will which may be described as calm, tend towards the 'universal happiness of others' while the turbulent include 'pity, condolence, congratulation, gratitude'.

Acts of will which are selfish but calm include 'an invariable constant impulse towards one's own perfection and happiness of the highest kind' (*System*, I, p. 9) and do not rule out 'deliberate purposes of injury' (*System*, I, p. 73). The turbulent and selfish embrace 'hunger, thirst, lust, passions for sensual pleasure, wealth, power or fame' (*System*, I, pp. 11-12).

In Hutcheson's case, the problem is that of attaining degree of balance between the turbulent and the calm, the selfish and the benevolent:

> the general tenor of human life is an incoherent mixture of many social, kind, innocent actions, and of many selfish, angry, sensual ones; as one or other of our natural dispositions happens to be raised, and to be prevalent over others. (*System*, I, p. 37)

While Smith was correct in identifying Hutcheson with that school of thought which found virtue to consist in benevolence, there is equally no doubt that he (Hutcheson) gave a prominent place to self-love:

> Our reason can indeed discover certain bounds, within which we may not only act from self-love consistently with the good of the whole; but every mortal's acting thus within these bounds for his own good, is absolutely necessary for the good of the whole; and the want of self-love would be universally pernicious … But when self-love breaks over the bounds above mentioned, and leads us into actions detrimental to others, and to the whole; or makes us insensible of the generous kind affections; then it appears vicious, and is disapproved. (1725, III.v)

As in the case of Smith, what is critically important is man's desire to be approved of:

> an high pleasure is felt upon our gaining the approbation and esteem of others for our good actions, and upon their expressing their sentiments of gratitude; and on the other hand, we are cut to the very heart by censure, condemnation, and reproach. (*System*, I, p. 25)

On Hutcheson's argument an important source of control is represented by a capacity for judgement, including moral judgement, which is linked to man's deployment of internal senses such as the 'sympathetic' which differ from external senses such as sight, sound, or taste, and 'by which, when we apprehend the state of others, our hearts naturally have a fellow-feeling with them' (*System*, I, p. 19). It was Hutcheson's contention that men were inclined to, and fitted for, society: 'their curiosity, communicativeness, desire of action, their sense of honour, their compassion, benevolence, gaiety and the moral faculty, could have little or no exercise in solitude' (*System*, I, p. 34).

This discussion was to lead to Hutcheson's treatment of natural rights and of the state of nature in a manner which is reminiscent of Locke. He also advances the Lockian claim that the state of nature is a state not of war but of inconvenience which can only be resolved by the establishment of government in terms of a complex double contract.

This has been described as the 'Real Whig position' (Winch, 1978, p. 46; Robbins, 1968) and may explain the considerable influence of Hutcheson's political ideas in the American colonies (Norton, 1976). Hutcheson's 'warm love of liberty' was attested by Principal Leechman in his introduction to the *System* (I, pp. xxxv–xxxvi); a sentiment which was echoed by Hugh Blair (Winch, 1978, pp. 47–8) in a contemporary review of the book.

While agreeing that an essential precondition of social stability is some system of 'magistracy' (TMS, VII.iv.36), Adam Smith (like Hume) was to emerge as a critic of the contract theory. In addition, he criticized Hutcheson for seeming to imply that self-love was 'a principle which could never be virtuous in any degree or in any direction' (TMS, VII.ii.3.12). But for the economist it is important to note that Hutcheson distinguished often more clearly than did Smith between approval and moral approbation. As Hutcheson put it:

> A penetrating genius, capacity for business, patience of application and labour … are naturally admirable and relished by all observers, but with quite a different feeling from moral approbation. (*System*, I, p. 28)

Whatever the differences of emphasis and of analysis which are disclosed in the writings of Hutcheson and Smith, the arguments reviewed in this section are or should be important to the economist for three reasons. First, it appears that social order as a basic precondition for economic activity depends in part upon a capacity for moral judgement. Secondly, it is alleged that the psychological drives which explain economic activity must be seen in a context wider than the economic. Finally, the argument suggests that all forms of activity are subject to the scrutiny of our fellows.

Economic analysis

There are five major topics covered in Hutcheson's *System*, which is generally assumed to follow closely the content of his lecture course as a whole. The economic analysis is not given in the form of a single coherent discourse, but rather woven in the broader treatment of jurisprudence. Perhaps for this reason Hutcheson's work did not attract a great deal of attention from early historians of economic thought. But the situation was transformed as a result of Edwin Cannan's discovery of Smith's *Lectures on Jurisprudence*. Cannan recalled that:

> On April 21, 1895, Mr Charles C Maconochie, Advocate, whom I then met for the first time, happened to be present when, in course of conversation with the literary editor of the Oxford Magazine, I had occasion to make some comment about Adam Smith. Mr Maconochie immediately said that he possessed a manuscript report of Adam Smith's lectures on jurisprudence, which he regarded as of considerable interest. (1896, p. xv)

While Cannan's reaction may be imagined, the lectures had the effect of confirming Hutcheson's influence upon his pupil on a broad front, but especially in the area or economic analysis (as distinct from policy). For what Cannan discovered was that the *order* of a large part of Smith's course and its content corresponded closely with what Hutcheson was believed to have taught. It is this correspondence which served to renew interest in Hutcheson's economics with remarkable speed. Quite apart from Cannan's introduction to the *Lectures*, the same theme is elaborated in his introduction to the *Wealth of Nations* (1904). The link had also been noted, following the publication of the *Lectures*, in the Palgrave *Dictionary of Political Economy* (1896) and received its most elaborate statement in W.R. Scott's *Francis Hutcheson* (1900). The most modern treatment of this kind is to be found in W.L. Taylor's influential work *Francis Hutcheson and David Hume as Predecessors of Adam Smith* (1965).

But Cannan noted something else, namely that it may be that the 'germ of the *Wealth of Nations*' is to be found in Hutcheson's treatment of value (1896, p. xxvi). It is this topic which forms the central feature of the remainder of the present argument although it will be convenient to begin with Hutcheson's views on the division of labour where his influence on Smith may be particularly obvious.

But before we pass on to these subjects, it should be noted that Hutcheson's work on economic topics has its own history. It is evident that he admired the work of his immediate predecessor in the Chair of Moral Philosophy – Gershom Carmichael (1672–1729), and especially his translation of, and commentary on, Samuel Pufendorf. In Hutcheson's address to the 'students in Universities' (Taylor, 1965, p. 25) the *Introduction to Moral Philosophy* (1742) is described thus:

> The learned will at once discern how much of this compound is taken from the writing of others, from Cicero and Aristotle, and to name no other moderns, from Pufendorf's smaller work, *De Officio Hominis et Civis Juxta Legem Naturalem* which that worthy and ingenious man the late Professor Gerschom Carmichael of Glasgow, by far the best commentator on that book has so supplied and corrected that the notes are of much more value than the text.

Carmichael's influence as a student of ethics and of jurisprudence has been frequently celebrated, notably by Sir William Hamilton who stated that he may be regarded 'on good grounds, as the true founder of the Scottish school of philosophy' (Taylor, 1965, p. 253). But it is to W.L. Taylor that we are indebted for the reminder that Carmichael (and Pufendorf) may have shaped Hutcheson's economic ideas. Taylor concluded that:

> The interesting point for the development of economic thought in all this is the very close parallelism between Pufendorf's *De Officio* and Hutcheson's *Introduction to Moral Philosophy*. Each man covered almost exactly the same field … The inescapable conclusion is that Francis Hutcheson took over almost in whole, from Carmichael, the economic ideas of Pufendorf. (1965, pp. 28–2)

The division of labour

A key issue for both Hutcheson and Pufendorf arose from the comparison of the social as distinct from

the solitary state; or, as Pufendorf put it,

> it would seem to have been more wretched than that of any wild beast, if we take into account with what weakness man goes forth into this world, to perish at once, but for the help of others; and how rude a life each would lead, if he had nothing more than what he owed to his own strength and ingenuity. On the contrary, it is altogether due to the aid of other men, that out of such feebleness, we have been able to grow up, that we now enjoy untold comforts, and that we improve mind and body for our own advantage and that of others. And in this sense of natural state is opposed to a life improved by the industry of men. (*De Officio*, 1682, II, pp. 8–9)

This broad line of argument was developed in the *System* (II, p. 4) where Hutcheson offered two specific economic applications. First, he noted that the '*joint* labours of twenty men will cultivate forests, or drain marshes, for farms to each one, and provide houses for habitation, and enclosures for their stocks, much sooner than the *separate* labours of the same number' (*System*, II, p. 289).
Secondly, Hutcheson drew attention to the importance of the *division* of labour:

> Nay 'tis well known that the produce of the labours of any given number, twenty, for instance, in providing the necessaries or conveniences of life, shall be much greater by assigning to one, a certain sort of work of one kind, in which we will soon acquire skill and dexterity, and to another assigning work of a different kind, than if each one of the twenty were obliged to employ himself, by turns in all the different sorts of labour requisite for his subsistence, without sufficient dexterity in any. In the former method each procures a great quantity of goods of one kind, and can exchange a part of it for such goods obtained by the labours of others as he shall stand in need of. One grows expert in tillage, another in pasture and breeding cattle, a third in masonry, a fourth in the chace, a fifth in iron-works, a sixth in the arts of the loom, and so on throughout the rest. Thus all are supplied by means of barter with the works of complete artists. In the other method scarce any one could be dextrous and skilful in any one sort of labour. (*System*, II, pp. 288–9)

Property

The discussion of the division of labour implied that members of society are interdependent in respect of the satisfaction of their wants. It also led to two further analytical developments: security of property and the problem of value in exchange (see especially Brown, 1987).
Much of the discussion in Book 2, Chapter 6 of the *System* is concerned with 'the right of property'. But Hutcheson also noted that:

> If we extend our views further and consider what the common interest of society may require, we shall find the right of property further confirmed. Universal industry is plainly necessary for the support of mankind. Tho' men are naturally active, yet their activity

> would rather turn toward the lighter and pleasanter exercises, than the slow, constant, and intense labours requisite to procure the necessaries and conveniences of life, unless strong motives are presented to engage them to these severer labours. Whatever institution therefore shall be found necessary to promote universal diligence and patience, and make labour agreeable or eligible to mankind, must also tend to the public good; and institutions or practices which discourage industry must be pernicious to mankind. Now nothing can so effectually excite men to constant patience and diligence in all sorts of useful industry, as the hopes of future wealth, ease, and pleasure to themselves, their offspring, and all who are dear to them, and of some honour too to themselves on account of their ingenuity, and activity, and liberality. All these hopes are presented to men by securing to every one the fruits of his own labours, that he may enjoy them, and dispose of them as he pleases.
>
> Nay the most extensive affections could scarce engage a wise man to industry, if no property ensued upon it. (*System*, II, pp. 320–1)

Hutcheson attached a great deal of importance to freedom of choice and in fact concluded this phase of the argument by rejecting any suggestion that ‘magistrates’ may be involved, passages that may well have attracted the attention of the youthful Smith. (*System*, II, pp. 322–3)

The theory of value

It is Hutcheson’s treatment of value that shows most clearly the influence of Pufendorf and of Carmichael where the latter observed that:

> In general we may say that the value of goods depends upon these two elements, their *scarcity*, and the difficulty of acquiring them. Furthermore, *scarcity* is to be regarded as combining two elements, the number of those demanding, and the *usefulness* thought to adhere in the good or service, and which can add to the utility of human life. (Quoted in Taylor, 1965, p. 65)

Pufendorf’s analysis received its most elaborate statement in the *De Jure*, in the long chapter ‘On Price’ (Book 5, Chapter 1). The most succinct statement, on which Carmichael commented, is to be found in Book 1, Chapter 14, of *De Officio*.
Hutcheson opened his analysis of the problem by pointing out that the ‘natural ground of all value or price is some sort of use which goods afford in life’, adding that ‘by the use causing a demand we mean not only a natural subserviency to our support, or to some natural pleasure, but any tendency to give any satisfaction by prevailing custom or fancy, as a matter of ornament or distinction’ (*System*, II, pp. 53–4). He continued:

> But when some aptitude to human use is presupposed, we shall find that the prices of goods depend on these two jointly, the *demand* on account of some use or other which many desire, and the *difficulty* of acquiring, or cultivating for human use. When goods are

> equal in these respects men are willing to interchange them with each other; nor can any artifice or policy make the values of goods depend on any thing else. When there is no *demand*, there is no price, where the *difficulty* of acquiring never so great: and where there is no *difficulty* or labour requisite to acquire, the most universal *demand* will not cause a price; as we see in fresh water in these climates. Where the demand for two sorts of goods is equal, the prices are as the difficulty. Where the difficulty is equal, the prices are as the demand. (*System*, II, p. 54)

Hutcheson then added two points which are reminiscent of Pufendorf in commenting on issues that affect supply price and the rate of exchange. First, he argued:

> In like manner by difficulty of acquiring, we do not only mean great labour or toil, but all other circumstances which prevent a great plenty of the goods or performances demanded. Thus the price is increased by the rarity or scarcity of the materials in nature, or such accidents as prevent plentiful crops or certain fruits of the earth; and the great ingenuity and nice taste requisite in the artists to finish well some works of art, as men of such genius are rare. The value is also raised, by the dignity of station in which, according to the custom of the country, the men must live or provide us with certain goods, or works of art. Fewer can be supported in such stations than in the meaner; and the dignity and expense of their stations must be supported by the higher prices of their goods or services. Some other singular considerations may exceedingly heighten the values of goods to some men, which will not affect their estimation with others. These above mentioned are the chief which obtain in commerce. (*System*, II, pp. 54–5)

As regards the *rate of exchange*, Hutcheson commented:

> In commerce it must often happen that one may need such goods of mine as yield a great and lasting use in life, and have cost a long course of labour to acquire an cultivate, while yet he has none of those goods I want in exchange, or not sufficient quantities; or what goods of his I want, may be such as yield but a small use, and are procurable by little labour. In such cases it cannot be expected that I should exchange with him. I must search for others who have the goods I want, and such quantities of them as are equivalent in use to my goods, *and require as much labour to produce them*; and the goods on both sides must be brought to some estimation or value. (*System*, II, p. 53)

But although these positions do not differ significantly from those of Pufendorf, Hutcheson does seem to have taken notice of two additional points. First, he seems to suggest, as the above quotation indicates, that goods will exchange at a rate that will be in part determined by the quantity of labour embodied in them (a point later taken up by Smith). Secondly, he noted in a passage that may have been ‘foreshadowed’ by Pufendorf, that some commodities: ‘of great use have no price, either because they are naturally destined for community, or cannot come into commerce but as appendages of something else, the price of which may be increased by them, though they cannot be separately

estimated' (Hutcheson, 1742b; quoted in Taylor, 1965, p. 66).

Money

The discussion of value in exchange led Hutcheson on quite logically to consider the medium of exchange, namely money, and here too he followed an old tradition which had already been commented upon by Pufendorf. In Book I, Chapter 14 of *De Officio* he noted the inconvenience of exchange by barter:

> But after men departed from their primitive simplicity and various kinds of gain were introduced, it was readily understood that common value alone was not sufficient for the transactions of men's affairs and their increased dealings.

Once more, Hutcheson followed suit in explaining the problems of barter and the need to establish a standard or 'common measure' when settling the 'values or goods for commerce'.

> The qualities requisite to the most perfect standard are these: it must be something generally desired so that men are generally willing to take it in exchange. The very making of any goods the standard will of itself give them this quality. It must be *portable*; which will often be the case if it is rare, so that small quantities are of great value. It must be *divisible* without loss into small parts, so as to be suited to the values of all sorts of goods: and it must be *durable*, not easily wearing by use, or perishing in its nature. One or other of these prerequisites in the standard, shews the inconvenience of many of our commonest goods for that purpose. The man who wants a small quantity of my corn will not give me a work-beast for it, and his beast does not admit division. I want perhaps a pair of shoes, but my ox is of far greater value, and the other may not need him. I must travel to distant lands, my grain cannot be carried along for my support, without insufferable expense, and my wine would perish in the carriage. 'Tis plain therefore that when men found any use for the rarer metals, silver and gold, in ornaments and utensils, and thus a demand was raised for them, they would soon also see that they were the fittest standards of commerce, on all the accounts above-mentioned. (*System*, II, pp. 55–6)

The familiar arguments concerning the need for coinage and the dangers of debasement follow (*System*, II, ch. 12), while there is also a hint of the need to find an invariable measure of value at least over long periods of time.

> We say indeed commonly, that the rates of labour and goods have risen since these metals grew plenty; and that the rates of labour and goods were low when the metals were scarce; conceiving the value of the metals as invariable, because the legal names of the pieces, the pounds, shillings, or pence, continue to them always the same till a law alters them. But a days digging or ploughing was as uneasy to a man a thousand years ago as it is now, tho' he could not then get so much silver for it: and a barrel of wheat, or beef, was then of the

> same use to support the human body, as it is now when it is exchanged for four times as much silver. Properly, the value of labour, grain, and cattle, are always pretty much the same, as they afford the same uses of life, where no new inventions of tillage, or pasturage, cause a greater quantity in proportion to the demand. 'Tis the metal chiefly that has undergone the great change of value, since these metals have been in greater plenty, the value of the coin is altered tho' it keeps the old names. (*System*, II, p. 58)

The analytical section of the work is concluded in the following chapter where Hutcheson demonstrated the need for *interest*, since if it were prohibited 'none would lend' (*System*, II, p. 72). He argued that the rate would be determined 'by the state of trade and the quantity of coin, recognizing that 'as men can be supported by smaller gains upon proportion upon their large stocks, the profit made upon any given sum employed is smaller, and the interest the trader can afford must be less' (*System*, II, p. 72). Hutcheson was well aware of the relationship between interest and other forms of return, such as rent, and also introduced an allowance for risk. In sum, an interesting and often sophisticated analysis, taken as whole, which is likely to have made an impression of the youthful Smith.

Conclusion

This article has pursued a number of themes. First, it endeavours to establish a link between Hutcheson and Pufendorf. Secondly, the argument has elaborated on the parallel between Hutcheson's order of argument and that developed by Adam Smith as suggested by W.R. Scott (1900; 1937), Cannan (1896; 1904) and W.L. Taylor (1965). While these parallels are important, it is noteworthy that Smith's treatment of economic topics is worked out as a single discourse, while Hutcheson's treatment is woven into the broader fabric of his analysis of jurisprudence. Finally, the argument has sought to give prominence to the role of subjective judgement as regards the determinants of value in exchange. Edwin Cannan, as we have seen, considered that Hutcheson's emphasis on the utility of goods to be acquired and on the effort (disutility) involved in creating the goods to be exchanged, with the attendant emphasis on demand and supply considerations, provided the 'kernel' of the *Wealth of Nations*. Taylor, on the other hand, suggested that Smith's concern with material welfare served to obscure the line of argument set out by Hutcheson. Robertson and Taylor concluded that:

> It is evident that the *magnum opus* was cast in a mould of a powerful unifying conception. Now within this framework it is evident that the measurement, in real terms, of the wealth of nations, and in particular of its progress would seem to call for some unvarying standard of value which would enable valid comparisons to be made through time … for this reason, if for no other, it does not appear inexplicable that Adam Smith no longer paid so much attention to the lines of argument taken over from Hutcheson, which had served well enough in the *Lectures*. (1957, pp. 194–5)

What Robertson and Taylor did not note was that Smith's preoccupation with a real measure of value may also have owed much to Hutcheson (Skinner, 1996, 148–50).

Selected works

All citations of *System* in the text refer to *A System of Moral Philosophy* (1755). Hutcheson's *Collected Works* are published by Georg Olms, Hildesheim, 1969.

1725. *Inquiry into the Original of our Ideas of Beauty and Virtue*. London, 2nd edn, 1726.

1725–26. Reflections upon laughter and Remarks on the Fable of the Bees. *Dublin Journal* No. 11 (5 June 1725); No. 12 (12 June); No.13 (19 June). Also in Hibernicus, *Letters* (1729).

1728a. *Essay on the Nature and Causes of the Passions, with illustrations upon the Moral Sense*. London and Dublin.

1728b. Letters between the late Mr G. Burnet and Mr Hutcheson. *London Journal*.

1742a. *The Meditations of Marcus Aurelius. Newly translated from the Greek, with Notes and an Account of His Life*. Glasgow.

1742b. *A Short Introduction to Moral Philosophy in Three Books, containing the Elements of Ethics and the Law of Nature*. Glasgow.

1755. *A System of Moral Philosophy in Three Books. Published from the original Ms. by his son. Francis Hutcheson MD to which is prefaced Some Account of the Life, Writings and Character of the Author. By the Reverend William Leechman, D.D., Professor of Divinity in the same University*. Glasgow.

Bibliography

Brown, V. 1987. Value and property in the history of economic thought: an analysis of the emergence of scarcity. *Oeconomia* 7, 85–112.

Campbell, T.D. 1982. Francis Hutcheson. In *The Origins and Nature of the Scottish Enlightenment*, ed. R.H.Campbell and A.S. Skinner. Edinburgh: John Donald.

Cannan, E. 1896. *Adam Smith's Lectures on Justice, Police, Revenue and Arms*. Oxford: Clarendon Press.

Cannan, E. (ed.) 1904. *The Wealth of Nations*. London: Methuen.

Hutcheson, T. 1988. *Before Adam Smith: The Emergence of Political Economy 1622–1776*. Oxford: Blackwell.

Kaye, F.B. (ed.) 1924. *The Fable of the Bees*. Oxford: Clarendon Press.

McCosh, J. 1875. *The Scottish Philosophy from Hutcheson to Hamilton*. Princeton: Princeton University Press.

Meek, R.L. 1976. New light on Adam Smith's Glasgow Lectures on Jurisprudence. *History of Political Economy* 8, 439–77.

Naldi, N. 1993. Gershom Carmichael on demand and difficulty of acquiring. *Scottish Journal of Political Economy* 40, 456–70.

Norton, D.F. 1976. Francis Hutcheson in America. In *Studies on Voltaire and the Eighteenth Century* 154, 1547–68.

Pesciarelli, E. 1986. On Adam Smith's Glasgow Lectures on Jurisprudence. *Scottish Journal of Political Economy* 33, 74–85.

Pufendorf, S. von. 1682. *De Officio et Civis Juxta Legem Naturalem Libri Duo*. Trans. F.G.Moore, 2 vols, Oxford: Oxford University Press, 1927.

Pufendorf, S. von. 1688. *De Jure Naturae et Gentium Libri Octo*. Trans. C.H. W.A.Oldfather, Oxford: Clarendon Press, 1934.

Robbins, C. 1954. When is it that colonies may turn independent: an analysis of the environment and politics of Francis Hutcheson. *William and Mary Quarterly* 11, 214–51.

Robbins, C. 1968. *The Eighteenth Century Commonwealth Man*. New York: Atheneum.

Robertson, H.M. and Taylor, W.L. 1957. Adam Smith's approach to the theory of value. *Economic Journal* 67, 181–98.

Scott, W.R. 1900. *Francis Hutcheson: His Life, Teaching and Position in the History of Philosophy*. Cambridge: Cambridge University Press.

Scott, W.R. 1932. Francis Hutcheson. In *Encyclopaedia of the Social Sciences*, vol. 7, ed. E.R.A. Seligman. New York: Macmillan.

Skinner, A.S. 1996. *A System of Social Sciences: Papers relating to Adam Smith*. Oxford: Clarendon Press.

Skinner, A.S. 2006. Francis Hutcheson. In *A History of Scottish Economic Thought*, ed. Alexander and S. Dow. London: Routledge.

Smith, A. 1759. *The Theory of Moral Sentiments* (TMS), ed. D.D. Raphael A.L. Macfie. Oxford, Clarendon Press, 1976.

Smith, A. 1776. *The Wealth of Nations*, ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Clarendon Press, 1976.

Taylor, W.L. 1965. *Francis Hutcheson and David Hume as Predecessors of Adam Smith*. Durham, NC: University of North Carolina Press.

Winch, D. 1978. *Adam Smith's Politics: An Essay in Historiographic Revision*. Cambridge: Cambridge University Press.

How to cite this article

Skinner, Andrew. "Hutcheson, Francis (1694–1746)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000108> doi:10.1057/9780230226203.0759

The New Palgrave Dictionary of Economics Online

Hutchison, Terence Wilmot (1912 – 2007)

Roger E. Backhouse
From The New Palgrave Dictionary of Economics Online, ,
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic policy; economic theory; history of economics; Hutchison, Terence Wilmot; Machlup, Fritz; Ricardo, David; Robbins, Lionel; Robinson, Joan; statistics; truth in economics

Article

Terence Hutchison, a specialist in economic methodology and the history of economic thought, defended the idea that, if economics was to make progress, economic propositions needed to be testable and confronted with evidence. This, together with scepticism about theory based on the assumption of perfect knowledge, informed not only his methodological writing but also his work on the history of economics.

Career

Hutchison was born in Bournemouth on 13 August 1912, and attended Tonbridge School. He went to Cambridge in 1931, to read classics, but switched to economics in which he had Joan Robinson as his tutor, obtaining a first in 1934. Though much of his subsequent work can be seen as a rebellion against his tutor's economics and her politics, he acknowledged her role in training him to think. In his final year he picked up some of Wittgenstein's ideas from two of his friends, to whom Wittgenstein was dictating the lectures that comprised his *Blue Book*. Hutchison attended the now-famous lectures in which John Maynard Keynes worked his way towards the *General Theory*, and later rued the loss of his lecture notes in his wartime travels.
After a year spent going to lectures at the London School of Economics and reading widely, in 1935 he obtained a job as *Lektor* in Bonn, where his main duty was to give lectures which could be on any subject, so long as they were in good English. He remained there for around three years, learning German and developing the interest in German economic and methodological writing, the latter having been stimulated by his undergraduate exposure to Wittgenstein, that ran through all his work. While there he married. As his wife was German, they decided not to move to England, but to Baghdad, where

he taught at a teacher training college. With the coming of a pro-Nazi government which wanted to reduce British influence there, he managed to get his family out, via Basra, to Bombay. A while later, he was allowed out to join them, and he joined up. He served on the Northwest Frontier and later in Egypt where he worked as an intelligence officer. He spent the last years of the war in Delhi, at one point working with All India Radio.

Hutchison's British university career began, in 1946, with a year at Hull, after which he moved to the London School of Economics. There, working alongside Lionel Robbins, who shared and stimulated his interest in continental European writing, he taught courses on the history of economic thought since 1870 and on the history of economic controversies. In 1956 he was appointed Mitsui Professor of Economics at the University of Birmingham, the position he held until his retirement in 1978. He taught the history of economic thought until 1980, when university regulations forced him to stop. In retirement, his research continued unabated till only a few years before his death.

Away from his academic pursuits, he had a passion for cricket. He played the game in Egypt during the war, and in the 1950s became a good club cricketer. He first visited Lords (Middlesex versus the Australians) with his mother in 1921, and during the final match between England and Australia in 2005, he appeared on television to give an account of the corresponding game in 1926 (perhaps he was by then the only person alive who had seen all four days of that match).

Economic methodology and the history of economic thought

Hutchison's reputation was established with his first book, *The Significance and Basic Postulates of Economic Theory* (1938). This was a response to the recently published *Essay on the Nature and Significance of Economic Science* (1932/1935) in which Lionel Robbins had defended economic theory as a body of propositions deduced from the assumption of scarcity. Hutchison argued that most economic theory comprised tautologies that said nothing about the real world. Economists should instead seek to develop testable propositions and confront them with evidence. The book's significance lay partly in its being the first attempt systematically to apply to economics philosophical ideas being developed in the 1930s, the most prominent of which went under the label of logical positivism. Hutchison was particularly critical of any theorizing based on the assumption of perfect knowledge. The book received unexpected attention when it was the subject of a 32-page review article, ‘“What is truth” in economics?’ in the *Journal of Political Economy* for 1940 by the eminent Chicago economist Frank Knight, to which Hutchison replied from wartime Baghdad (Knight, 1940; Hutchison, 1941).

Though Hutchison continued to emphasize testability and the limitations of theorizing based on perfect knowledge, one strand of his methodological work involved engaging with ideas coming from the philosophy of science. In the 1950s he became involved in an exchange with Fritz Machlup, after being described as an ‘ultra-empiricist’ (Machlup, 1955, 1956; Hutchison, 1956). The framework within which this debate, over the extent to which propositions needed to be testable, took place reflected the concerns of the so-called ‘received view, then dominant in the philosophy of science. In the 1970s, Hutchison brought detailed knowledge of the history of economics to bear on the question of whether economics had exhibited revolutionary changes corresponding to those that Thomas Kuhn and Imre Lakatos claimed to have identified in the history of science (Hutchison, 1976, 1978, chapter 3)

This knowledge of the history of economics was first demonstrated in *A Review of Economic Doctrines*,

1870–1929 (1953b), a book that arose out of the course Hutchison taught at LSE, which provided a systematic coverage of the subject from the date of the so-called marginal revolution to the onset of the Great Depression. It was unjustly overshadowed by the appearance of Joseph Schumpeter's posthumous magnum opus a year later. Methodological themes were never far from the surface. Interestingly, the book concluded with a discussion of the growth of economic statistics, on which what he thought ‘the most spectacular progress in economic knowledge was necessarily being founded’ (p. 427). This view that the development of economic statistics was the main example of progress in economics was one that he maintained throughout his career (see, for example, Hutchison 1977, chapter 2; 1992; 1994, chapter 8). He became increasingly critical of theoretical work that was not grounded in empirical work, criticizing the ‘crisis of abstraction’ of the 1970s (Hutchison, 1977) and later the ‘formalist revolution’ (Hutchison, 1992, 2000) and the literature that developed from around the 1980s, dismissing a focus on prediction as outdated positivism.

The other strand in Hutchison's methodological work was analysis of policy. *“Positive” Economics and Policy Objectives* (1964), though a methodological book that sought to bring clarity to policy discussions through applying the positive-normative discussion, had a strong historical dimension, analysing economists' statements over several centuries. Most prominent, however, was *Economics and Economic Policy in Britain, 1946–1966* (1968). This examined what economists had said on economic policy, in some instances contrasting this with what they later claimed to have said. He followed this up with an essay, ‘Economic knowledge and ignorance in action’, which showed that, despite claims to the contrary, economists simply did not agree on the questions of whether sterling should have been devalued in the 1960s, or whether Britain should have entered the European Community (Hutchison 1977, chapter 5). He clearly delighted in pointing out how reviewers considered it an outrage to hold economists to account for claims they had made in newspaper articles or correspondence columns and the suggestion that this was, somehow, merely journalism. His own view was that to understand the policy process it was necessary to take account of economists' views, wherever they were published.

Though concerned throughout with methodological questions and with what had shaped modern economics, his interests extended much further back. *Before Adam Smith* (1985) was the first English-language work to analyse systematically the entire century of economic writing before Adam Smith's *Wealth of Nations*. As his use of the phrase ‘contentious essays’ in one of his book titles suggests, he never shirked controversy, often challenging widely accepted beliefs about major figures in economics. As with his work on economists' policy advice, he repeatedly pointed out inconsistencies in the statements of economists who upheld dogmatic views. A particular target was the Marxian ideology of his former teacher, Joan Robinson, and Maurice Dobb, and the way it coloured their interpretation of the past. He believed that readers of their historical interpretations should be informed about their views on Stalin's Soviet Union and Mao's cultural revolution (Hutchison, 1981, chapter 3). He argued that early ‘marginalists’ were not unqualified supporters of laissez-faire, concerned to defend capitalism against Marxist critics, but supporters of extensive pragmatic government intervention in economic activity. Similarly, he pointed out that in the early 1930s the differences between A. C. Pigou and Keynes were slight: Pigou advocated fiscal cures for unemployment and Keynes attributed part of the problem to the rigidity of money incomes (Hutchison, 1978: 179).

Hutchison's most controversial target was David Ricardo, who he saw as the source of the excessively abstract theorizing that plagued modern economics (1952, 1953a, 1978, 1994). When reviewing Piero Sraffa's edition of David Ricardo's collected works, he feigned surprise that its sponsor had been the

Royal Economic Society, not the Moscow State Publishing house (Hutchison, 1952: 421). He questioned not only the Marxist interpretation of Ricardo but, even more controversially, made the heretical suggestion that Ricardo was less original and less central to the history of economics than was commonly assumed. Decades later (1994, chapter 5), he ridiculed the idea that this believer in the sanctity of private property was, despite his influence on Marx, a man of the left. Ricardo was, he claimed, ‘something of an innocent abroad, whose inconsistent ideas … fell into the hands of people too keen on exploiting them for their own ideological purposes, and who had to pretend that these inconsistencies were not there’ (Hutchison, 1994: 99).

However, his criticisms were not just directed against those on the left. He also raised questions about Friedrich Hayek and the Austrians (Hutchison, 1981, 1994). The common theme running through his writing was the need for clear thinking informed by knowledge of what economists had actually said.

Acknowledgements

This article draws on an obituary written by the author for The Times, published on December 5, 2007. See Coats (1983a,b) for appraisals of Hutchison's work, and the interviews recorded in Tribe (1997) and Hart (2002).

See Also

- falsificationism
- history of economic thought
- methodology of economics
- philosophy and economics

Selected works

1938. *The Significance and Basic Postulates of Economic Theory*. London: Macmillan.

1941. The significance and basic postulates of economic theory: a reply to Professor Knight. *Journal of Political Economy* 49(5), 732–50.

1952. Some questions about Ricardo. *Economica* 19, 415–32.

1953a. Ricardo's correspondence. *Economica* 20, 263–73.

1953b. *A Review of Economic Doctrines, 1870–1929*. Oxford: Clarendon Press.

1956. Professor Machlup on verification in economics. *Southern Economic Journal* 22(4), 476–83.

1964. *“Positive” Economics and Policy Objectives*. London: George Allen and Unwin.

1968. *Economics and Economic Policy in Britain, 1946–1966*. London: George Allen and Unwin.

1976. On the history and philosophy of science and economics. In *Method and Appraisal in Economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press, 181–206.

1977. *Knowledge and Ignorance in Economics*. Oxford: Basil Blackwell.

1978. *On Revolutions and Progress in Economic Knowledge*. Cambridge: Cambridge University Press.

1981. *The Politics and Philosophy of Economics: Marxians, Keynesians and Austrians*. Oxford: Basil Blackwell.

1985. *Before Adam Smith: The Emergence of Political Economy, 1662–1776*. Oxford: Basil Blackwell.

1992. *Changing Aims in Economics*. Oxford: Basil Blackwell.

1994. *The Uses and Abuses of Economics: Contentious Essays on History and Method*. London: Routledge.

2000. *On the Methodology of Economics and the Formalist Revolution*. Cheltenham: Edward Elgar.

Bibliography

Coats, A.W. 1983a. Half a century of methodological controversy in economics: as reflected in the writings of T.W. Hutchison. In *Methodological Controversy in Economics: Historical Essays in Honor of T.W. Hutchison*, pp. 1–42, ed. A.W. Coats. Greenwich, CT: JAI Press.

Coats, A.W. 1983b. T.W. Hutchison as a historian of economics. In *The Craft of the Historian of Economic Thought. Research in the History of Economic Thought and Methodology: A Research Annual*, vol. 1, pp. 187–208, ed. W.J. Samuels. Greenwich, CT: JAI Press.

Hart, J. 2002. A conversation with Terence Hutchison. *Journal of Economic Methodology* 9(3), 359–377.

Knight, F.H. 1940. ‘What is truth?’ in economics. *Journal of Political Economy* 48(1), 1–32.

Machlup, F. 1955. The problem of verification in economics. *Southern Economic Journal* 22(1), 1–21.

Machlup, F. 1956. Rejoinder to a reluctant ultra-empiricist. *Southern Economic Journal* 22(4), 483–93.

Robbins, L.C. 1932. *Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Second edition, 1935.

Tribe, K. 1997. Terence Hutchison. *Economic Careers: Economics and Economists in Britain, 1930–1970*. London: Routledge, chapter 8.

How to cite this article

Backhouse, Roger E. "Hutchison, Terence Wilmot (1912–2007)." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000190> doi:10.1057/9780230226203.1873

The New Palgrave Dictionary of Economics Online

hyperinflation

Juan Pablo Nicolini

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A hyperinflation occurs when price indexes of broadly defined baskets of goods increase at extremely high rates. As such, hyperinflations are rare. However, the few known cases share many things in common. First, they can occur only in paper currency systems that are not pegged by the central bank to any good. Second, they occur when the quantity of paper currency also grows at extremely high rates. Finally, the force behind the process is always a fiscal imbalance that is financed by issuing currency.

Keywords

Bretton Woods system; budget deficits; commodity money; convertibility; fiat money; Fisher, I.; German hyperinflation; gold standard; hyperinflation; inflation; Laffer curve; paper money; price control; price stability; quantity theory of money; seigniorage; stabilization policy; wage control

Article

Price stability shares with a healthy knee a particular feature: both are precious, but you do not realize how much until you miss them. When you run, your knees perform amazing functions, without you even being aware of them. That is price stability. Under some circumstances, one of your knees may be under some stress and you may be forced to use medication to be able to run well. While you run, you are aware of your knee. That is inflation. Eventually, your knee hurts so much you can only walk. That is high inflation. Finally, in the worst case, your knee is broken and you must lie in bed. That is hyperinflation.

Money – that is, a commodity that is widely used as a medium of exchange – has been in use in the world since commerce became a social activity. However, to the extent that money was a particular commodity or was paper money but pegged to a commodity like silver or gold, there was no risk of long-run inflation.

From the point of view of the theory, this premise comes from the quantity equation that was first

formalized by Irving Fisher (1934). He argued that the general level of prices was a constant proportion to the ratio of the supply of currency and some index of the total quantity of goods that are traded in a year. Thus, there cannot be long-run inflation without long-run growth of the net supply of the commodity that serves as money or that backs the paper money in circulation, where by ‘net supply’ I mean the rate of growth of money in excess of the growth rate of the index of total goods.

The first known example of inflation occurred during the 16th century in Europe, precisely because of the increase in the supply of gold and silver that came from South America after the Spanish conquest. It is interesting to recall, however, that this first inflation was roughly 100 per cent during the whole century or, equivalently, 0.7 per cent a year. According to the theory, this means that the net supply of gold and silver doubled in 100 years. (The ability of Fisher's quantity framework to explain low inflation events during relatively short periods of time like a few years has been rightly called into question. However, for the kind of episodes that I discuss here, which involve very high inflation rates, this conceptual framework is perfectly suitable. See Marcet and Nicolini, 2005, and all references therein.)

By 1900 paper money was the norm, but all economies were functioning under some form of commodity standard in the sense that money was backed by some commodity, typically gold. Governments would suspend convertibility in some circumstances, like wars, but would eventually restore it. Thus, the ability to increase the net supply of paper money depended on the ability of the issuer to accumulate the commodity that backed it. As a consequence, the economic history of the world does not have records of persistent increases in the general level of prices up to the 20th century, except for the cases of the exceptional gold and silver inflows after the Spanish conquest of America mentioned above.

In a seminal paper, Cagan (1956) defined monthly inflation rates that exceed 50 per cent a month as hyperinflations. To generate a hyperinflation according to this definition, Columbus would have had to double Europe's net supply of gold and silver in a little less than two months!

The 20th century witnessed, among other things, a key change in the functioning of our monetary systems. Today, almost without exception, all modern economies function under fiat money arrangements in the sense that paper money circulates, is widely accepted and used in transactions, and is not backed to any particular commodity. Thus, the size of its net supply depends only on the will of the issuer.

All episodes of hyperinflation we observed during the 20th century, no matter how we define them, and with absolutely no exception, occurred during periods of unbacked paper money. All of them, no matter how we define them and with absolutely no exception, occurred during periods in which the net supply of paper money increased at enormous rates. And all of them occurred in times of substantial fiscal imbalances, represented by excessive government expenditures, inadequate government revenues or a huge government debt burden – or a combination of these.

The first burst of hyperinflations occurred in the 1920s in countries that lost the First World War, most notably Germany and Hungary. Sargent (1992) provides a very neat description of the causes and remedies for each of the cases. It is remarkable that the only cases registered in the first half of the century were highly concentrated in time and space: all occurred between 1922 and 1923 and in central Europe. A common story can be told about those episodes: political instability, large fiscal imbalances due, in part, to war and huge increases in the money supply.

It is also interesting to note that the first half of the century was still characterized mainly by convertible monetary systems. The four hyperinflationary experiences described by Sargent occurred during

temporary suspensions of the gold standard. By the mid-1970s, however, after the fall of the Bretton Woods arrangement, the world moved to a fiat money system, in which no commodity serves as backing.
The second half of the century also witnessed hyperinflationary episodes. Somewhat surprisingly, the second wave of hyperinflationary episodes was concentrated in the period 1985–94. And they were concentrated in two regions; it would appear, though, that the temporal coincidence was just random. The countries involved were Argentina, Brazil, Bolivia and Peru in Latin America, and Yugoslavia and Poland in central Europe. Again, a common story could be told: in the first years of the 1980s, the four Latin American countries experienced major financial crises, including default in international debt markets. As a consequence, the ability of the governments to smooth temporary fiscal shocks via credit markets was severely restricted. The four countries had experienced in the previous decade substantial political instability, including military dictatorships and weak democratic governments. On the other hand, both Poland and Yugoslavia were undergoing substantial political and economic transformation after the fall of the USSR. In all cases, there were major fiscal imbalances: government deficits were chronic and volatile. As consequence, money printing became the only source of revenues and major bursts in inflation rates occurred.
It is interesting to note that other Latin American countries (Colombia, Uruguay, Mexico) also suffered financial and debt crises, but did not experience inflation rates of this magnitude, and other central European countries underwent major political and economic transformation and did not have hyperinflations.
Indeed, what we have learned (see Bruno et al., 1988; 19911991) is that major political and economic crises are a necessary condition for hyperinflations to occur. But crisis will lead to hyperinflation if, and only if, the crisis manifests itself in serious fiscal imbalances that are financed by the central bank issuing unbacked paper money. There is a wide consensus in the literature about this.
Although we know very precisely the conditions under which hyperinflations are almost unavoidable, it is difficult to tell exactly when the burst will start and how large it will be.
The subtlety of hyperinflationary dynamics has been explored in a sequence of papers (Eckstein and Leiderman, 1992; Zarazaga, 1993; Marcet and Nicolini, 2003) that can be seen as complementary. All these models share the property, supported by evidence, that hyperinflations can occur only in economies with large and persistent fiscal deficits that are purely financed by printing money, or seigniorage. In all the models, the problem arises because the required seigniorage is close to the maximum revenue that can be raised, given the demand for real money, that is, the maximum of the Laffer curve. Eckstein and Leiderman (1992) argue that if the elasticity of money demand with respect to the inflation rate approaches one form above, when average seigniorage is very high, very small shocks to it can generate drastic changes in the required inflation rate.
Zarazaga (1993) introduces a decentralized government with a common pool of resources and private information on the shock to the spending opportunities of each member of the government. Hyperinflations occur when there are too many positive expenditure shocks, there is too much demand for resources, and the required seigniorage is too high. When this happens, a price war-type strategy follows in which all agencies become excessively demanding and the central bank ends up issuing enormous amounts of currency. Finally, Marcet and Nicolini (2003) introduce very small departures from rationality and show that the dynamics of the most simple seigniorage model change in a way that fits the evidence surprisingly well.

From the point of view of inflation stabilization policies, the debate has taken three routes. The first claims that the key for a successful stabilization policy is to correct the fundamentals, this is, to make a drastic and permanent change in fiscal policy so as to eliminate the need to print money. This kind of policy is called ‘orthodox’. The second puts the emphasis on ‘heterodox’ policies, that is, a combination of nominal anchors like fixing the nominal exchange rate – eventually moving towards a gold or strong currency standard – and price and wage controls. Finally, a third approach points to the need to combine the other two policies. From the point of view of experience and the theory, it is clear that no attempt to stabilize the economy without orthodox policies has any chance of success in the medium term. And it appears from experience that in most successful cases (although there has been some debate on whether this was true in all of them), some type of nominal anchor, typically the exchange rate, was also important. While not all theoretical models put much weight on the nominal anchor (Marcet and Nicolini, 2003, is the most notable exception), in all of the models these policies are either harmless or good for the success of the stabilization effort.

A final word regarding Cagan's (1956) definition: as with any definition, it is arbitrary. Had we taken a lower inflation rate per month, like 25 per cent, the number of experiences would have been greater, and many more countries would have been involved in our discussion. However, the general lessons one learns are essentially the same. Quantity theory predictions work extremely well, and the most appropriate policies to deal with these experiences are the same.

See Also

- German hyperinflation
- inflation

Bibliography

Bruno, M., Di Tella, G., Dornbusch, R. and Fisher, S. 1988. *Inflation Stabilization: The Experience of Argentina, Brazil, Israel and Mexico*. Cambridge, MA: MIT Press.

Bruno, M., Fisher, S., Helpman, E. and Liviatan, N. 1991. *Lessons of Economic Stabililzation and its Aftermath*. Cambridge, MA: MIT Press.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Eckstein, Z. and Leiderman, L. 1992. Seigniorage and the welfare cost of inflation. *Journal of Monetary Economics* 29, 389–410.

Fisher, I. 1934. *Stable Money: A History of the Movement*. New York: Adelphi.

Marcet, A. and Nicolini, J. 2003. Recurrent hyperinflations and learning. *American Economic Review*

93, 1476–98.

Marcet, A. and Nicolini, J. 2005. Money and prices in models of bounded rationality in high-inflation economies. *Review of Economic Dynamics* 8, 452–79.

Sargent, T. 1992. The ends of four big inflations. In *Rational Expectations and Inflation*, 2nd edn. New York: Harper and Row.

Zarazaga, C. 1993. Hyperinflation and moral hazard in the appropriation of seigniorage. Working paper, Federal Reserve Bank of Philadelphia.

How to cite this article

Nicolini, Juan Pablo. "hyperinflation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000112> doi:10.1057/9780230226203.0760

The New Palgrave Dictionary of Economics Online

ideal output

J. De V. Graaff

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Pigou's notion of ‘the ideal output’ as ‘the output in any industry which maximizes the national dividend, and, apart from the differences in the marginal utility of money to different people, also maximises satisfaction’ has long been eclipsed by the ‘general optimum of production and exchange’, in which the welfare of each member of the community is maximized in turn, subject to certain constraints – even though the more modern theory, despite its advantages, does not necessarily reach any substantially different conclusions. But ideal output theory is by now no more than an episode in the history of economic thought.

Keywords

Barone, E.; competitive equilibrium; external economies; general optimum of production and exchange; ideal output; imperfect competition; marginal equivalence; marginal social product; monopoly; Pareto, V.; Pigou, A. C.

Article

Pigou, writing in *The Economics of Welfare*, calls ‘the output in any industry which maximizes the national dividend, and, apart from the differences in the marginal utility of money to different people, also maximises satisfaction, the ideal output’. He goes on to argue that ‘this output is attained – the possibility of multiple maximum positions being ignored – when the value of the marginal social net product of each sort of resource invested in the industry under review is equal to the value of the marginal social net product of resources in general’. And, finally, it ‘will be that output which makes the demand price of the output equal to the money value of the resources engaged in producing a marginal unit of output’ (1932, pp. 802, 803).

The line of argument that comes through so clearly in these quotations can be traced back to Pigou's earlier *Wealth and Welfare* (1912) and indeed to Marshall; but since the 1930s it has been overtaken by

the development of a more powerful strand of analysis that stems from Pareto (1897) and Barone (1908) and has culminated in the theory of the general optimum of production and exchange. In it one maximizes in turn the welfare of each member of the community, subject to the constraint of the social production function and to holding on each occasion the welfare of each other member constant. The resulting first-order conditions include the marginal equivalences enumerated in the theory of ideal output (Graaff, 1957). Any modern discussion of the theory must therefore be set against the background of the one that has incorporated and replaced it.

The more modern theory has the virtues of elegance, simplicity and generality. It embraces exchange as well as production. It deals with commodities and firms (or event plants) instead of industries. It does not need the doctrine of maximum satisfaction, or any assumption about interpersonal comparisons of utility. But at the end of the day it does not reach any substantial conclusion that the theory of ideal output, correctly employed, would not itself have reached.

The problem, especially in the early development of the theory, was that it was not all that easy to apply it correctly. It was not originally recognized that (at least in a closed economy) the correct way to reckon the value of a marginal social net product is at *constant* prices. The same remark applies to the calculation of marginal social cost. If higher prices have to be offered to factors of production to attract them to an industry undergoing expansion, the element of the cost of the expansion caused by the higher prices represents a transfer payment to the factors (in the form of a rent or quasi-rent), not a cost to society. The cost to society is the value of the output sacrificed when the factors are withdrawn from their previous use. That value was reckoned at the original prices of the factors. Those prices must therefore be used in reckoning their cost to society in their new use.

Clarification of this issue was the result of a famous debate of the 1920s – much of it reprinted in *Readings in Price Theory* (Stigler and Boulding, 1953) – on the desirability of taxing industries subject to diminishing returns, and paying bounties to those subject to increasing returns, a result to which the theory of ideal output at one stage seemed to point. As competitive conditions were meant to be prevailing, the industries enjoying increasing returns had to be assumed to comprise firms whose unit costs were falling because of *external* economies; and as external economies were themselves recognized as possible reasons for a divergence between private and social net products, the opportunities for getting muddled were legion. It is to the credit of the participants – among them D.H. Robertson, G.F. Shove, F.H. Knight and J. Viner – that these dangers were largely avoided.

Much of the motivation for the theory of ideal output seems to have been a desire to see when competitive output was ideal, and when interference in a competitive economy would be justified. Today we ask, rather more formally (cf. Debreu, 1959), when a competitive equilibrium would also be a general optimum. The answer, very briefly, is when the technology is convex, there are no external effects in production or consumption, no public goods and no foreign trade.

Apart from the fact that the existence of public goods was glossed over, ideal output theory would not have given a very different answer. The importance of the foreign trade exception was recognized. (The marginal social cost of importing goods subject to rising supply price is higher than the marginal private cost. The rents that accrue to *foreigners* are not mere transfers within the domestic community, but a part of social cost). Divergences between private and social costs due to external economies and diseconomies in production, and between private and social benefits due to external economies and diseconomies in consumption, were fully discussed. The counterpart of the modern insistence on a convex technology was the painstaking treatment of increasing returns. The conditions under which

competitive output would approach the ideal were pretty clearly defined.

Pigou also discussed the deviation from the ideal of the outputs of discriminating monopolists. (Not surprisingly, they fell short.) R.F. Kahn (1935) extended the analysis to imperfect competition. He argued that (taking diseconomies as negative economies) all industries could be arranged in descending order on a scale according to the extent of the external economies they generated and the degree of monopoly (measured by the gap between price and marginal cost) they enjoyed and that at a certain point on the scale there would be an average industry. Above this point all should expand to produce ideal outputs; below it all should contract. Adjustment could be achieved by a set of taxes and bounties. When all industries had expanded or contracted to conform to the average degree of monopoly and the average capacity to create external economies, their marginal social products would diverge from their marginal private products to the same extent and ideal output would be attained.

Note that this treatment avoids the error of making ‘piecemeal’ recommendations of the sort so often found in partial analysis. All industries must move to the average. It may not help if one or two do. That may just increase the gap between those that conform and those that do not. (In technical terms, the first-order conditions for a maximum must be satisfied simultaneously.)

In this sense Kahn's treatment is very general. In another it is not general enough. Proportionality of marginal products is not sufficient. For a full optimum, equality is essential (Lerner, 1944, ch. 9). This may require an adjustment in the number of hours worked, and an expansion or contraction in the level of output as a whole.

The view that suitable corrective taxes and bounties can and should be used to bring marginal private products into line with marginal social products, when they diverge, was once very popular. On the whole it has weathered less well than ideal output theory itself, although the latter is by now no more than an episode in the history of economic thought.

See Also

- Pareto efficiency

Bibliography

Barone, E. 1908. II Ministerio della Produzione nello stato collectivista. *Giornale degli Economisti*. Trans. in *Collectivist Economic Planning*, ed. F.A. Hayek, London: Routledge, 1935.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Graaff, J.de.V. 1957. *Theoretical Welfare Economics*. Cambridge: Cambridge University Press.

Kahn, R.F. 1935. Some notes on ideal output. *Economic Journal* 45, (March), 1–35.

Lerner, A.P. 1944. *The Economics of Control*. New York: Macmillan.

Pareto, V. 1897. *Cours d'économie politique*. Lausanne: Rouge.

Pigou, A.C. 1912. *Wealth and Welfare*. London: Macmillan.

Pigou, A.C. 1932. *The Economics of Welfare*. 4th edn, London: Macmillan.

Stigler, G.J. and Boulding, K.E. (ed.) 1953. *Readings in Price Theory*. London: Allen & Unwin.

How to cite this article

Graaff, J. De V. "ideal output." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000002> doi:10.1057/9780230226203.0761

The New Palgrave Dictionary of Economics Online

identification

Jean-Marie Dufour Cheng Hsiao

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The problem of identification is defined in terms of the possibility of characterizing parameters of interest from observable data. This problem occurs in many fields, such as automatic control, biomedical engineering, psychology, systems science, the design of experiments, and econometrics. This article focuses on identification in econometric models, which typically involve random variables. Identification in general parametric statistical models is defined, and its meaning in a number of specific econometric models is considered: regression (collinearity), simultaneous equations, dynamic models, and nonlinear models. Identification in nonparametric models, weak identification, and the statistical implications of identification failure are also discussed.

Keywords

Bayes' th; collinearity; endogeneity and exogeneity; identification; instrumental variable; linear models; multivariate regression models; nonparametric estimation; nonparametric models; probability; random variables; returns to schooling; separability; serial correlation; simultaneous equations models; treatment effect; weak identification; weak instruments

Article

In economic analysis, we often assume that there exists an underlying structure which has generated the observations of real-world data. However, statistical inference can relate only to characteristics of the distribution of the observed variables. Statistical models which are used to explain the behaviour of observed data typically involve parameters, and statistical inference aims at making statements about these parameters. For that purpose, it is important that different values of a parameter of interest can be characterized in terms of the data distribution. Otherwise, the problem of drawing inferences about this parameter is plagued by a fundamental indeterminacy and can be viewed as 'ill-posed'.
To illustrate, consider X as being normally distributed with mean $E(X) = \mu_1 - \mu_2$. Then $\mu_1 - \mu_2$ can be

estimated using observed *X*. But the parameters μ_1 and μ_2 are not uniquely estimable. In fact, one can think of an infinite number of pairs $(\mu_i, \mu_j), \quad i, \quad j = 1, 2, \ldots (i \neq j)$ such that $\mu_i - \mu_j = \mu_1 - \mu_2$. In order to determine μ_1 and μ_2 uniquely, we need additional prior information, such as $\mu_2 = 3\mu_1$ or some other assumption. Note, however, that inference about the variance of *X* remains feasible without extra assumptions.

More generally, *identification failures* – or situations that are close to it – complicate considerably the statistical analysis of models, so that tracking such failures and formulating restrictions to avoid them is an important problem of econometric modelling.

The problem of whether it is possible to draw inferences from the probability distribution of the observed variables to an underlying theoretical structure is the concern of econometric literature on identification. The first economists to raise this issue were Working (1925; 1927) and Wright (1915; 1928). The general formulations of the identification problems were made by Frisch (1934), Marschak (1942), Haavelmo (1944), Hurwicz (1950), Koopmans and Reiersøl (1950), Koopmans, Rubin and Leipnik (1950), Wald (1950), and many others. An extensive treatment of the theory of identification in simultaneous equation systems was provided by Fisher (1976). Surveys of the subject can be found in Hsiao (1983), Prakasa Rao (1992), Bekker and Wansbeek (2001), Manski (2003), and Matzkin (2007); see also Morgan (1990) and Stock and Trebbi (2003) on the early development of the subject.

In this article, we first define the notion of identification in general parametric models (Sections 1 and 2) and discuss its meaning in a number of specific statistical models used in econometrics, such as regression models (collinearity), simultaneous equations, dynamic models, and nonlinear models (Section 3). Identification in nonparametric models (Sections 4 and 5), weak identification (Section 6), and the statistical implications of identification failure (Section 7) are also considered.

1 Definition of parametric identification

It is generally assumed in econometrics that economic variables whose formation an economic theory is designed to explain have the characteristics of random variables. Let *y* be a set of such observations. A structure *S* is a complete specification of the probability distribution function of *y*. The set of all a priori possible structures, *T*, is called a model. In most applications, *y* is assumed to be generated by a parametric probability distribution function $F(y, \theta)$, where the probability distribution function *F* is assumed known, but the *q*×1 parameter vector θ is unknown. Hence, a structure is described by a parametric point θ, and a model is a set of points $A \subseteq \mathbb{R}^q$.

Definition 1: Two structures, $S^0 = F(y, \theta^0)$ *and* $S^* = F(y, \theta^*)$ *are said to be* observationally equivalent *if* $F(y, \theta^0) = F(y, \theta^*)$ *for* (*'almost'*) *all possible y. A model is identifiable if A contains no two distinct structures which are observationally equivalent. A function of* θ, $g(\theta)$, *is identifiable if all observationally equivalent structures have the same value for* $g(\theta)$.

Sometimes a weaker concept of identifiability is useful.

Definition 2: A structure with parameter value θ^0 *is said to be locally identified if there exists an open neighborhood of* θ^0, *W, such that no other* θ *in W is observationally equivalent to* θ^0.

2 General results for identification in parametric models

Lack of identification reflects the fact that a random variable has the same distribution for some if not all values of the parameter. R.A. Fisher's information matrix provides a sensitivity measure of the distribution of a random variable due to small changes in the value of the parameter point (Rao, 1962). It can therefore be shown that, subject to regularity conditions, θ^0 is locally identified if and only if the information matrix evaluated at θ^0 is nonsingular (Rothenberg, 1971).
It is clear that unidentified parameters cannot be consistently estimated. There are also pathological cases where identified models fail to possess consistent estimators (for example, Gabrielson, 1978). However, in most practical cases, we may treat identifiability and the existence of a consistent estimator as equivalent; for precise conditions, see Le Cam (1956) and Deistler and Seifert (1978).

3 Some specific parametric models

The choice of model structure is one of the basic ingredients in the formulation of the identification problem. In this section we briefly discuss some identification conditions for different types of models in order to demonstrate the kind of prior restrictions required.

3.1 Linear regression with collinearity

One of the most common models where an identification problem does occur is the linear regression model:

$$\mathbf{y} = X\beta + \mathbf{u} \tag{1}$$

where **y** is an $n\times1$vector of dependent observable variables, X is an $n\times k$ fixed matrix of observable variables, β a $k\times1$ unknown coefficient vector, and **u** is an $n\times1$ vector of disturbances whose components are (say) independent and identically distributed according to a normal distribution $N(0, \sigma^2)$ with unknown positive variance σ^2.
In this model, the value of β must be determined from the expected value of $\mathbf{y}$: $E(\mathbf{y}) = X\beta$. If the latter equation has a solution for β (that is, if the model is correct), the solution is unique if and only the regressor matrix X has rank k. If X has rank zero (which entails X=0), all values of β are equivalent (β is completely *unidentifiable*). If $1 \le \mathrm{rank}(X) < k$, then not all the components can be determined, but some linear combinations of the components of β (say $c'\beta$) can be determined (that is, they are *identifiable*). A necessary and sufficient condition for $c'\beta$ to be estimable (identifiable) is that $c = (X'X)d$ for some vector d. Linear combinations that do not satisfy this condition are not identifiable. The typical way out of such collinearity problems consists in imposing restrictions on β

(identifying restrictions) which set the values of the unidentifiable linear combinations (or components) of β.

Correspondingly, when *X* does not have full rank, the equation $(X'X)\hat{\beta} = X'\mathbf{y}$, which defines the least squares estimator $\hat{\beta}$, does not have a unique solution. But all solutions of the least squares problem can be determined by considering $\hat{\beta} = (X'X)^{-}X'\mathbf{y}$ where $(X'X)^{-}$ is any generalized inverse of $(X'X)$. Different generalized inverses then correspond to different identifying restrictions on β. For further discussion, see Rao (1973, ch. 4).

3.2 Linear simultaneous equations models

Consider a theory which predicts a relationship among the variables as

$$B\mathbf{y}_t + \Gamma\mathbf{x}_t = \mathbf{u}_t, \; t = 1, \ldots, n, \tag{2}$$

where $\mathbf{y}_t$ and $\mathbf{u}_t$ are $G\times1$ vectors of observed and unobserved random variables, respectively, $\mathbf{x}_t$ is a $K\times1$ vector of observed non-stochastic variables, *B* and Γ are $G\times G$ and $G\times K$ matrices of coefficients, with *B* nonsingular. We assume that the $\mathbf{u}_t$ are independently normally distributed with mean 0 and variance-covariance matrix Σ. Equations (2) are called structural equations. Solving for the endogenous variables, **y**, as a function of the exogenous variables, **x**, and the disturbance u, we obtain:

$$\mathbf{y}_t = -B^{-1}\Gamma\mathbf{x}_t + B^{-1}\mathbf{u}_t = \Pi\mathbf{x}_t + \mathbf{v}_t, \tag{3}$$

where $\Pi = -B^{-1}\Gamma, \; E\mathbf{v}_t = 0, \; E\mathbf{v}_t\mathbf{v}_t' = V = B^{-1}\Sigma(B^{-1})'$. Equations (3) are called the *reduced form* equations derived from (2) and give the conditional likelihood of $\mathbf{y}_t$ for given $\mathbf{x}_t$ that summaries the information provided by the observed ($\mathbf{y}_t$, $\mathbf{x}_t$). The variables in $\mathbf{x}_t$ are often also called ‘instruments’. From (3), we see that the simultaneous equations model can be viewed as a special case of a multivariate regression model (MLR), such that the regression coefficient matrix Π satisfies the equation:

$$B\Pi = -\Gamma. \tag{4}$$

Provided the matrix $X=[x_1,\ldots,x_n]'$ has full rank K (no collinearity), the regression coefficient matrix Π is uniquely determined by the distribution of $Y=[y_1,\ldots,y_n]'$ (it is *identifiable*). The problem is then whether B and Γ can be uniquely derived from eq. (4). Premultiplying (2) by a $G\times G$ nonsingular matrix D, we get a second structural equation:

$$B^*\mathbf{y}_t + \Gamma^*\mathbf{x}_t = \mathbf{u}_t^*, \tag{5}$$

where $B^*=DB$, $\Gamma^*=D\Gamma$, and $\mathbf{u}_t^* = D\mathbf{u}$. It is readily seen that the reduced form of (5) is also (3). So eq. (4) cannot be uniquely solved for B and Γ, given Π. Therefore, the two structures are observationally equivalent and the model is *non-identifiable*.

To make the model identifiable, additional prior restrictions have to be imposed on the matrices B, Γ and/or Σ. Consider the problem of estimating the parameters of the first equation in (2), out of a system of G equations. If the parameters cannot be estimated, the first equation is called *unidentified* or *underidentified*. If given the prior information, there is a unique way of estimating the unknown parameters, the equation is called *just identified*. If the prior information allows the parameters to be estimated in two or more linearly independent ways, it is called *overidentified*. A necessary condition for the first equation to be identified is that the number of restrictions on this equation be no less than $G-1$ (order condition). A necessary and sufficient condition is that a specified submatrix of B, Γ and Σ be of rank $G-1$(rank condition) (see Fisher, 1976; Hausman and Taylor, 1983). For instance, suppose the restrictions on the first equation are in the form that certain variables do not appear. Then this rank condition says that the first equation is identified if and only if the submatrix obtained by taking the columns of B and Γ with prescribed zeros in the first row is of rank $G-1$ (Koopmans and Reiersøl, 1950).

3.3 Dynamic models

When both lagged endogenous variables and serial correlation in the disturbance term appear, we need to impose additional conditions to identify a model. For instance, consider the following two equation system (Koopmans, Rubin and Leipnik, 1950):

$$y_{1t} + \beta_{11}y_{1,t-1} + \beta_{12}y_{2,t-1} = u_{1t}, \beta_{12}y_{1t} + y_{2t} = u_{2t}. \tag{6}$$

If (u_{1t}, u_{2t}) are serially uncorrelated, (6) is identified. If serial correlation in (u_{1t}, u_{2t}) is allowed, then

$$y_{1t} + \beta_{11}^{*} y_{1,t-1} + \beta_{12}^{*} y_{2,t-1} = u_{1t}^{*}, \beta_{12} y_{1t} + y_{2t} = u_{2t}, \quad (7)$$

is observationally equivalent to (6), where $\beta_{11}^{*} = \beta_{11} + d\beta_{21}$, $\beta_{12}^{*} = \beta_{12} + d$, and $u_{1t}^{*} = u_{1t} + du_{2t}$. Hannan (1971) derives generalized rank conditions for the identification of this type of model by first assuming that the maximum orders of lagged endogenous and exogenous variables are known, then imposing restrictions to eliminate redundancy in the specification and to exclude transformations of the equations that involve shifts in time. Hatanaka (1975), on the other hand, assumes that the prior information takes only the form of excluding certain variables from an equation, and derives a rank condition which allows common roots to appear in each equation.

3.4 Nonlinear models

For linear models, we have either global identification or else an infinite number of observationally equivalent structures. For models that are linear in parameters, but nonlinear in variables, there is a broad class of models whose members can commonly achieve identification (Brown, 1983; McManus, 1992). For models linear in the variables but nonlinear in the parameters, the state of the mathematical art is such that we only talk about local properties. That is, we cannot tell the true structure from any other substitute; however, we may be able to distinguish it from other structures which are close to it. A sufficient condition for local identification is that the Jacobian matrix formed by taking the first partial derivatives of

$$\omega_i = \Psi_i(\theta), \quad i = 1, \ldots, n, 0 = \Phi_j(\theta), \quad j = 1, \ldots, R, \quad (8)$$

with respect to θ be of full column rank, where the ω_i are n population moments of y and the Φ_j are the R a priori restrictions on θ (Fisher, 1976).

When the Jacobian matrix of (8) has less than full column rank, the model may still be locally identifiable via conditions implied by the higher-order derivatives. However, the estimator of a model suffering from first-order lack of identification will in finite samples behave in a way which is difficult to distinguish from the behaviour of an unidentified model (Sargan, 1983).

3.5 Bayesian analysis

In Bayesian analysis all quantities, including the parameters, are random variables. Thus, a model is said to be identified in probability if the posterior distribution for θ is proper. When the prior distribution for θ is proper, so is the posterior, regardless of the likelihood function of *y*. In this sense unidentifiability causes no real difficulty in the Bayesian approach. However, basic to the Bayesian argument is that all probability statements are conditional, that is, they consist essentially in revising the probability of a fixed event in the light of various conditioning events, the revision being accomplished by Bayes' theorem. Therefore, in order for an experiment to be informative with regard to unknown parameters (that is, for the posterior to be different from the prior), the parameter must be identified or estimable in the classical sense and identification remains as a property of the likelihood function (Kadane, 1975).

Drèze (1975) has commented that exact restrictions are unlikely to hold with probability 1 and has suggested using probabilistic prior information. In order to incorporate a stochastic prior, he has derived necessary rank conditions for the identification of a linear simultaneous equation model.

4 Definition of identification in nonparametric models

When the restrictions of an economic model specify all functions and distributions up to the value of a finite dimensional vector, the model is said to be parametric. When some functions or distributions are left parametrically unspecified, the model is said to be semiparametric. The model is nonparametric if none of the functions and distributions are specified parametrically. The previous discussion is based on parametric specification. We now turn to the issue of whether economic restrictions such as concavity, continuity and monotonicity of functions, equilibrium conditions, the implications of optimization, and so on, may be used to guarantee the identification of some nonparametric models and the consistency of some nonparametric estimators (see Matzkin 1994).

Formally, an econometric model is specified by a vector of observable dependent and independent variables, a vector of unobservable variables, and a set of known functional relationships among the variables. When such functional relationships are unspecified, the nonparametric identification studies what functions or features of function can be recovered from the joint distribution of the observable variables.

The set of restrictions on the unknown functions and distributions in an econometric model defines the set of functions and distributions to which these belong. Let the model *T* denote the set of all a priori possible unknown functions and distributions. Let *m* denote a vector of the unknown functions and distributions in *T* and $P(m)$ denote the joint distribution of the observable variables under *m*. Then the identification of *m* can be defined as follows.

Definition 3: The vector of functions m is identified in T if for any other vector, $m^* \in T$ *such that* $m \neq m^*$, $P(m) \neq P(m^*)$.

Let $C(m)$ denote some feature of *m*, such as the sign of some coordinate of *m*.

Definition 4: The feature $C(m)$ *of m is identified if* $C(m)=C(m^*)$ *for all m,* $m^* \in T$ *such that* $P(m)=P(m^*)$.

5 Examples of nonparametric identification

Contrary to the parametric model, there is no general result for nonparametric identification. We shall

therefore give some examples of how restrictions can be used to identify nonparametric functions.

5.1 Generalized regression models

Economists often consider a model of the form

$$y = g(\mathbf{x}) + u. \tag{9}$$

When $E(u|\mathbf{x})=0$ and $g(\cdot)$ is a continuous function $g: \mathbf{x} \to \mathbb{R}$, then $g(\cdot)$ can be recovered from the joint distribution of $(y, \mathbf{x})$ because $E(y|\mathbf{x})=g(\mathbf{x})$.
In some cases, the object of interest is not a conditional mean function $g(\cdot)$, but some 'deeper' function, such as a utility function generating the distribution of demand for commodities by a consumer. For example, $\mathbf{x}$ in (9) can be a price vector for K commodities and the income of a consumer. Mas-Colell (1977) has shown that we can recover the underlying utility function from the distribution of demand if we restrict $g(\cdot)$ to be monotone increasing, continuous, concave and strictly quasi-concave functions.

5.2 Simultaneous equations models

Suppose $(\mathbf{y}, \mathbf{x})$ satisfies the structural equations

$$\mathbf{r}(\mathbf{x}, \mathbf{y}) = \mathbf{u}, \tag{10}$$

where $\mathbf{y}$ and $\mathbf{u}$ denote $G\times1$ vectors of observable endogenous and unobservable variables, respectively, $\mathbf{x}$ is a $K\times1$ vector of observable exogenous variables, $\mathbf{r}$ denotes the G unknown functions, and let $p(\mathbf{r})$ and $p(\mathbf{r}^*)$ represent the joint distributions of the observables under $\mathbf{r}$ and $\mathbf{r}^*$ respectively. Assume also that: (i) $\forall(\mathbf{x}, \mathbf{y})$, $\partial\mathbf{r}/\partial\mathbf{y}$ has full rank, (ii) there exists a function $\pi(\cdot)$ such that $y=\pi(\mathbf{x}, \mathbf{u})$ (for conditions ensuring this, see Benkard and Berry, 2006), and (iii) $\mathbf{u}$ is distributed independently of $\mathbf{x}$. Then a necessary and sufficient condition guaranteeing that $p(\mathbf{r}^*)=p(\mathbf{r})$ is that

$$\operatorname{rank}\begin{pmatrix} \dfrac{\partial \mathbf{r}_i^*}{\partial(\mathbf{x},\mathbf{y})} \\ \dfrac{\partial \mathbf{r}}{\partial(\mathbf{x},\mathbf{y})} \end{pmatrix} < G + 1,$$

(11)

for all (**x**, **y**) and i=1,…,G, and all, where $\mathbf{r}_i^*$ denotes the i-th coordinate function of $\mathbf{r}^* \in T$ (see Roehrig, 1988; Matzkin, 2007).

5.3 Latent variable models and the measurement of treatment effects

For each person i, let (y_{0i}^*, y_{1i}^*) denote the potential outcomes in the untreated and treated states, respectively. Then the treatment effect for individual i is

$$\Delta_i = y_{1i}^* - y_{0i}^*$$

and the average treatment effect (ATE) is defined as

$$E(\Delta_i) = E(y_{1i}^* - y_{0i}^*); \tag{12}$$

see Heckman and Vytlacil (2001).
Let the treatment status be denoted by the dummy variable d_i where d_i=1 denotes the receipt of treatment and d_i=0 denotes nonreceipt. The observed data are often in the form

$$y_i = d_i y_{1i}^* + (1 - d_i) y_{0i}^*. \tag{13}$$

Suppose $y_{1i}^* = \mu_1(\mathbf{x}_i, u_{1i})$, $y_{0i}^* = \mu_0(\mathbf{x}_i, u_{0i})$, and $d_i^* = \mu_D(\mathbf{z}_i) - u_{di}$, where d_i=1 if $d_i^* \geq 0$ and 0 otherwise, $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors of observable exogenous variables and (u_{1i}, u_{0i}, u_{di}) are unobserved random variables. The average treatment effect and the complete structural econometric model can be identified with parametric specifications of ($\mu_1(\cdot)$, $\mu_0(\cdot)$, $\mu_D(\cdot)$) and the joint distributions of (u_{1i}, u_{0i},

u_{di}) even though we do not simultaneously observe y_{1i}^* and y_{0i}^*. In the case that neither ($\mu_1(\cdot)$, $\mu_0(\cdot)$, $\mu_D(\cdot)$) nor the joint distribution of (u_1, u_0, u_d) are specified, certain treatment effects may still be nonparametrically identified under weaker assumptions. For instance, under the assumption that d_i is orthogonal to (y_{1i}^*, y_{0i}^*) conditional on a set of confounders (x, z) (conditional independence or ignorable selection), the ATE is identifiable and estimable by comparing the difference of the average outcomes from the treatment group and from the untreated (control) group (Heckman and Robb, 1985; Rosenbaum and Rubin, 1985). If the focus is on the average treatment effect for someone who would not participate if $p(\mathbf{z}) \leq p(\mathbf{z}_0)$ and would participate if $p(\mathbf{z})>p(\mathbf{z}_0)$ (the local average treatment effect (LATE)), where $p(\mathbf{z})=\text{Prob}(d=1|\mathbf{z})$ (propensity score), Imbens and Angrist (1994) show that under the assumptions of separability of the effects of observable factors and unobservable factors and independence between observed factors and unobserved factors, they can be estimated by the sample analogue of

$$\Delta^{LATE}(\mathbf{x}, p(\mathbf{z}), p(\mathbf{z}_0)) \equiv \frac{E(y|\mathbf{x}, p(\mathbf{z})) - E(y|\mathbf{x}, p(\mathbf{z}_0))}{p(\mathbf{z}) - p(\mathbf{z}_0)} \tag{14}$$

where, without loss of generality, we assume $p(\mathbf{z})>p(\mathbf{z}_0)$. The limit of LATE provides the local instrumental variable (LIV) estimand (Heckman and Vytlacil, 1999):

$$\Delta^{LIV}(\mathbf{x}, p(\mathbf{z})) \equiv \frac{\partial E(y|\mathbf{x}, p(\mathbf{z}))}{\partial p(\mathbf{z})}. \tag{15}$$

Heckman and Vytlacil (2001) give conditions that suitably weighted versions of LIV identify the ATE.

6 Weak instruments and weak identification

The most common way of trying to achieve identification consists in imposing exclusion restrictions on the variables of a structural equation. In model (2), suppose that $\mathbf{y}_t$ and $\mathbf{x}_t$ are partitioned as $\mathbf{y}_t = (y_{1t}, \mathbf{y}_{2t}', \mathbf{y}_{3t}')'$ and $\mathbf{x}_t = (\mathbf{x}_{1t}', \mathbf{x}_{2t}')'$ where y_{1t} is a scalar, $\mathbf{y}_{it}$ has dimension G_i(i=2, 3) and $\mathbf{x}_{it}$ has dimension K_i(i=1, 2). If $\mathbf{y}_{3t}$ and $\mathbf{x}_{2t}$ are excluded from the first equation and the coefficient of y_{1t} is normalized to one, this yields an equation of the form:

$$y_{1t} - \mathbf{y}_{2t}'\beta_1 = \mathbf{x}_{1t}'\gamma_1 + u_{1t},\ t = 1, \ldots, n. \tag{16}$$

Let us also rewrite the reduced equation for $\mathbf{y}_{2t}$ in terms of $\mathbf{x}_{1t}$ and $\mathbf{x}_{2t}$:

$$\mathbf{y}_{2t} = \Pi_{21}\mathbf{x}_{1t} + \Pi_{22}\mathbf{x}_{2t} + \mathbf{v}_{2t}. \tag{17}$$

Then, substituting (17) into (16), we see that the reduced form for y_{1t} is:

$$y_{1t} = \Pi_{11}\mathbf{x}_{1t} + \Pi_{12}\mathbf{x}_{2t} + \mathbf{v}_{1t}, \tag{18}$$

where $\mathbf{v}_{1t} = u_{1t} + \mathbf{v}_{2t}'\beta_1$, $\Pi_{11} = \gamma_1' + \beta_1'\Pi_{21}$ and

$$\Pi_{12}' = \Pi_{22}'\beta_1. \tag{19}$$

Since γ_1 is free, Π_{11} is not restricted, but eq. (19) determines the identifiability of β_1, hence also of γ_1. Provided eq. (19) has a solution (that is, if eq. (16) is consistent with the data), the solution is unique if and only if the rank of the $G_2 \times K_2$ matrix Π_{22} is equal to G_2, the dimension of β_1:

$$\mathrm{rank}(\Pi_{22}) = G_2. \tag{20}$$

If rank(Π_{22})<G_2, the vector β_1 is not identifiable. However, it is completely unidentifiable only if rank (Π_{22})=0, or equivalently if Π_{22}=0. If 1<rank(Π_{22})<G_2, some linear combinations $c'\beta_1$ are

identifiable, but not all of them. Failure of the identification condition means that the regressors (or the 'instruments') $\mathbf{x}_{2t}$ do not move enough to separate the effects of the different variables in $\mathbf{y}_{2t}$. Condition (20) underscores two important things: first, exclusion and normalization restrictions – which are easy to check – are not sufficient to ensure identification; second, identification depends on the way the exogenous variables $\mathbf{x}_{2t}$ *excluded* from the structural equation of interest (16) are related to endogenous variables $\mathbf{y}_{2t}$ included in the equation. The latter feature is determined by the matrix Π_{22} whose rows should be linearly independent. Since Π_{22} is not observable, this may be difficult to determine in practice.

A situation that can lead to identification difficulties is the one where the identification condition (20) indeed holds, but, in some sense, Π_{22} is 'close' not to have sufficient rank. In such situations, we say that we have *weak instruments*. In view of the fact that the distributions of most statistics move continuously as functions of Π_{22}, the practical consequences of being close to identification failure are essentially the same. Assessing the closeness to non-identification may be done in various ways, for example by considering the eigenvalues of the matrices which measure the 'size' of Π_{22}, such as $\Pi_{22}\Pi_{22}'$, $\Pi_{22}X_2'M(X_1)X_2\Pi_{22}'$ or a *concentration matrix* $\Sigma_{22}^{-1/2}\Pi_{22}X_2'M(X_1)X_2\Pi_{22}'\Sigma_{22}^{-1/2}$ where $X_1=[\mathbf{x}_{11},\ldots,\mathbf{x}_{1n}]'$, $X_2=[\mathbf{x}_{21},\ldots,\mathbf{x}_{2n}]'$, Σ_{22} is the covariance matrix of $\mathbf{v}_{2t}$, $\Sigma_{22}^{-1/2}$ is its square root, and $M(X_1) = I_n - X_1(X_1'X_1)^{-1}X_1'$. More generally, any situation where a parameter may be difficult to determine because we are close to a case where a parameter ceases to be identifiable may be called *weak identification*. Weak identification was highlighted as a problem of practical interest by Nelson and Startz (1990), Bound, Jaeger and Baker (1995), Dufour (1997), and Staiger and Stock (1997); for reviews, see Stock, Wright and Yogo (2002) and Dufour (2003).

7 Statistical consequences of identification failure

Identification failure has several detrimental consequences for statistical analysis:

1. 1. Parameter estimates, tests and confidence sets computed for unidentified parameters have no clear inpt; this situation may be especially misleading if the statistical instruments used do not reveal the presence of the problem.
2. 2. Consistent estimation is not possible unless additional information is supplied.
3. 3. Many standard distributional results used for inference on such models are not anymore valid, even with a large sample size (see Phillips, 1983; 1989; Rothenberg, 1984).
4. 4. Numerical problems also easily appear, due for example to the need to invert (quasi) singular matrices.

Weak identification problems lead to similar difficulties, but may be more treacherous in the sense that standard asymptotic distributional may remain valid, but they constitute very bad approximations to what happens in finite samples:

1. 1. Standard consistent estimators of structural parameters can be heavily biased and follow distributions whose form is far from the limiting Gaussian distribution, such as bimodal distributions, even with fairly large samples (Nelson and Startz, 1990; Hillier, 1990; Buse, 1992).
2. 2. Standard tests and confidence sets, such as Wald-type procedures based on estimated standard errors, become highly unreliable or completely invalid (Dufour, 1997).

A striking illustration of these problems appears in the reconsideration by Bound, Jaeger and Baker (1995) of a study on returns to education by Angrist and Krueger (1991). Using 329,000 observations, these authors found that replacing the instruments used by Angrist and Krueger (1991) with randomly generated (totally irrelevant) instruments produced very similar point estimates and standard errors. This result indicates that the original instruments were weak. Recent work in this area is reviewed in Stock, Wright and Yogo (2002) and Dufour (2003).

8 Concluding remarks

The study of identifiability is undertaken in order to explore the limitations of statistical inference (when working with economic data) or to specify what sort of a priori information is needed to make a model estimable. It is a fundamental problem concomitant with the existence of a structure. Logically it precedes all problems of estimation or of testing hypotheses.

An important point that arises in the study of identification is that without a priori restrictions imposed by economic theory it would be almost impossible to estimate economic relationships. In fact, Liu (1960) and Sims (1980) have argued that economic relations are not identifiable because the world is so interdependent as to have almost all variables appearing in every equation, thus violating the necessary condition for identification. However, almost all the models we discuss in econometrics are only approximate. We use convenient formulations which behave in a general way that corresponds to our economic theories and intuitions, and which cannot be rejected by the available data. In this sense, identification is a property of the model but not necessarily of the real world. It is also important to be careful about situations where identification almost does not hold (weak identification), since these are in practice as damaging for statistical analysis as identification failure itself.

The problem of identification arises in a number of different fields such as automatic control, biomedical engineering, psychology, systems science, and so on, where the underlying physical structure may be deterministic (for example, see Aström and Eykhoff, 1971). It is also aptly linked to the design of experiments (for example, Kempthorne, 1947; Bailey, Gilchrist and Patterson, 1977). Here, we restrict our discussion to economic applications of statistical identifiability involving random variables.

See Also

- econometrics
- endogeneity and exogeneity
- simultaneous equations models
- treatment effect

Bibliography

Angrist, J.D. and Krueger, A.B. 1991. Does compulsory school attendance affect schooling and earning? *Quarterly Journal of Economics* 106, 979–1014.

Aström, K.J. and Eykhoff, P. 1971. System identification – a survey. *Automatica* 7, 123–62.

Bailey, R.A., Gilchrist, F.H.L. and Patterson, H.D. 1977. Identification of effects and confounding patterns in factorial designs. *Biometrika* 64, 347–54.

Bekker, P. and Wansbeek, T. 2001. Identification in parametric models. In *Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Benkard, C.L. and Berry, S. 2006. On the nonparametric identification of nonlinear simultaneous equations models: comment on Brown (1983) and Roehrig (1988). *Econometrica* 74, 1429–40.

Bound, J., Jaeger, D.A. and Baker, R.M. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–50.

Brown, B.W. 1983. The identification problem in systems nonlinear in the variables. *Econometrica* 51, 175–96.

Buse, A. 1992. The bias of instrumental variables estimators. *Econometrica* 60, 173–80.

Deistler, M. and Seifert, H.-G. 1978. Identifiability and consistent estimability in econometric models. *Econometrica* 46, 969–80.

Drèze, J. 1975. Bayesian theory of identification in simultaneous equations models. In *Studies in Bayesian Econometrics and Statistics*, ed. S.E. Fienberg and A. Zellner. Amsterdam: North-Holland.

Dufour, J.-M. 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. *Econometrica* 65, 1365–89.

Dufour, J.-M. 2003. Identification, weak instruments and statistical inference in econometrics. *Canadian Journal of Economics* 36, 767–808.

Fienberg, S.E. and Zellner, A., eds. 1975. *Studies in Bayesian Econometrics and Statistics*. Amsterdam: North-Holland.

Fisher, F.M. 1976. *The Identification Problem in Econometrics*. Huntington, NY: Krieger.

Frisch, R. 1934. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: Universitetes Okonomiske Institutt.

Gabrielson, A. 1978. Consistency and identifiability. *Journal of Econometrics* 8, 261–3.

Griliches, Z. and Intriligator, M.D., eds. 1983. *Handbook of Econometrics*, vol. 1, Amsterdam: North-Holland.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supp.), 1–115.

Hannan, E.J. 1971. The identification problem for multiple equation systems with moving average errors. *Econometrica* 39, 751–66.

Hatanaka, M. 1975. On the global identification of the dynamic simultaneous equations model with stationary disturbances. *International Economic Review* 16, 545–54.

Hausman, J.A. and Taylor, W.E. 1983. Identification, estimation and testing in simultaneous equations models with disturbance covariance restriction. *Econometrica* 51, 1527–49.

Heckman, J. and Robb, R. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman and B. Singer. New York: Cambridge University Press.

Heckman, J.J. and Vytlacil, E. 1999. Local instrumental variables and latent variables models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* 96, 4730–4.

Heckman, J.J. and Vytlacil, E. 2001. Local instrumental variables. In *Nonlinear Statistical Modeling Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune and J.L. Powell. Cambridge: Cambridge University Press.

Hillier, G.H. 1990. On the normalization of structural equations: properties of direction estimators. *Econometrica* 58, 1181–94.

Hsiao, C. 1983. Identification. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Hurwicz, L. 1950. Generalization of the concept of identification. In *Statistical Inference in Dynamic*

Economic Models, ed. T.C. Koopmans. New York: Wiley.

Imbens, G. and Angrist, J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 467–76.

Kadane, J.B. 1975. The role of identification in Bayesian theory. In *Studies in Bayesian Econometrics and Statistics*, ed. S.E. Fienberg and A. Zellner. Amsterdam: North-Holland.

Kempthorne, O. 1947. A simple approach to confounding and factorial replication in factorial experiments. *Biometrika* 34, 255–72.

Koopmans, T.C. 1950. *Statistical Inference in Dynamic Economic Models*. New York: Wiley.

Koopmans, T.C. and Reiersøl, O. 1950. The identification of structural characteristics. *Annals of Mathematical Statistics* 21, 165–81.

Koopmans, T.C., Rubin, H. and Leipnik, R.B. 1950. Measuring the equation systems of dynamic economics. In *Statistical Inference in Dynamic Economic Models*, ed. T.C. Koopmans. New York: Wiley.

Le Cam, L. 1956. On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press.

Liu, T.C. 1960. Underidentification, structural estimation, and forecasting. *Econometrica* 28, 855–65.

Manski, C. 2003. *Partial Identification of Probability Distributions*. New York: Springer.

Marschak, J. 1942. Economic interdependence and statistical analysis. In *Studies in Mathematical Economics and Econometrics*, ed. O. Lange, F. McIntyre and T.O. Yntema. Chicago: University of Chicago Press.

Mas-Collel, A. 1977. On the recoverability of consumers preferences from market demand behavior. *Econometrica* 45, 1409–30.

Matzkin, R. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Matzkin, R. 2007. Nonparametric Identification. In *Handbook of Econometrics*, vol. 6, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

McManus, D.A. 1992. How common is identification in parametric models? *Journal of Econometrics* 53, 5–23.

Morgan, M.S. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Nelson, C.R. and Startz, R. 1990. The distribution of the instrumental variable estimator and its *t*-ratio when the instrument is a poor one. *Journal of Business* 63, 125–40.

Phillips, P.C.B. 1983. Exact small sample theory in the simultaneous equations model. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Phillips, P.C.B. 1989. Partially identified econometric models. *Econometric Theory* 5, 181–240.

Prakasa Rao, B.L.S. 1992. *Identifiability in Stochastic Models: Characterization of Probability Distributions*. New York: Academic Press.

Rao, C.R. 1962. Problems of selection with restriction. *Journal of the Royal Statistical Society*, Series B 24, 401–5.

Rao, C.R. 1973. *Linear Statistical Inference and its Applications*, 2nd edn. New York: Wiley.

Roehrig, C.S. 1988. Conditions for identification in nonparametric and parametric models. *Econometrica* 56, 433–77.

Rosenbaum, P. and Rubin, D. 1985. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–24.

Rothenberg, T.J. 1971. Identification in parametric models. *Econometrica* 39, 577–91.

Rothenberg, T.J. 1984. Approximating the distributions of econometric estimators and test statistics. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Sargan, J.D. 1983. Identification and lack of identification. *Econometrica* 51, 1605–33.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Staiger, D. and Stock, J.H. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.

Stock, J.H. and Trebbi, F. 2003. Who invented IV regression? *Journal of Economic Perspectives* 17(3), 177–94.

Stock, J.H., Wright, J.H. and Yogo, M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20, 518–29.

Wald, A. 1950. Note on the identification of economic relations. In *Statistical Inference in Dynamic Economic Models*, ed. T.C. Koopmans. New York: Wiley.

Working, E.J. 1927. What do statistical demand curves show? *Quarterly Journal of Economics* 41, 212–35.

Working, H. 1925. The statistical determination of demand curves. *Quarterly Journal of Economics* 39, 503–43.

Wright, P.G. 1915. Moore's economic cycles. *Quarterly Journal of Economics* 29, 631–41.

Wright, P.G. 1928. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan.

How to cite this article

Dufour, Jean-Marie and Cheng Hsiao. "identification." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000004> doi:10.1057/9780230226203.0762

The New Palgrave Dictionary of Economics Online

identity

George Akerlof and Rachel Kranton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

gender roles; identity; signalling; social norms; well-being

Article

A person's *identity* is broadly defined as a person's self-image or sense of self. The concept of identity has wide use in most social sciences outside economics, especially sociology, anthropology and psychology. Many social scientists hold that preserving or enhancing identity is a prime motivation for individual and group behaviour. At the time of this writing, economists are beginning to explore the implications of identity for economic outcomes.

To do so, researchers primarily include identity as an aspect of utility. In this view, a person's actions and consumption of goods and services not only affect their material well-being, but also their psychological well-being. Researchers then ask how the inclusion of identity in utility can affect economic outcomes, such as charitable contributions (Bénabou and Tirole, 2006), information acquisition (K•szegi, 2006), schooling rates (Akerlof and Kranton, 2002), and the design of workplace incentives (Akerlof and Kranton, 2005).

We can divide the economic research on identity into two strands. The first considers an individual's self-image, as in Bénabou and Tirole (2005) and K•szegi (2006). The second considers an individual's self-image as it relates to societal norms and ideals (Akerlof and Kranton, 2000; 2002; 2005).

The first strand of research explores the simple proposition that people like to feel good about themselves. There are then trade-offs between standard economic costs and benefits, and the costs and benefits for one's own self-image. K•szegi (2006) uses such a utility function to explain why people may not undertake profitable investment projects, as the downside payoffs also reduce a person's sense of his own abilities. Bénabou and Tirole (2005) use identity to explain why monetary compensation can reduce the levels of pro-social activities (such as volunteer work and blood donations), as found in several studies and experimental work. They posit a utility function where an individual's action yields a monetary payoff and an ‘intrinsic’ payoff. Individuals can have different valuations/preferences for the

monetary payoff and the intrinsic payoff. Individuals like to think of themselves as placing, and like to think others think they place, high values on intrinsic payoff. That is, they want to think of themselves as enjoying the pro-social action for its own sake. But preferences are not observable, perhaps even to oneself. As individuals choose different actions, they and others make inferences about preferences. Hence, actions serve as ‘self-signal’ and a signal to others. The main results concern the trade-off between monetary payoffs and the signalling value of an action. When the monetary compensation for an action increases, the signal conveys less information about a person's underlying value for intrinsic payoffs. Hence, introducing monetary rewards can lower the levels of pro-social activity.

The second strand of research considers identity and norms. Sen (1985) and Elster (1989) were among the earliest proponents of the importance for economics of utility-based norms. Akerlof and Kranton (2000; 2002; 2005) relate a person's self-image to societal norms and ideals for different people in society. Whether or not a person feels good about herself depends on how that person *should* act, according to her place in society. Thus, to take the most obvious example, men are supposed to act differently from women, and identity utility will depend on the match between a person's actions and these gender norms. This notion of identity reflects a large body of research on ‘social identity’ in psychology, reviewed in Haslam (2001). Philosophy has also been another important influence on the connection between identity and norms, especially for Elster (1989) and Sen (1985).

Akerlof and Kranton (2000) posits the following utility function for an individual *j*:

$$U_j(a_j, a_{-j}, I_j)$$

where a_j are *j*'s actions, a_{-j} are others’ actions, and I_j is *j*'s ‘identity utility’ which is itself a function:

$$I_j(a_j, a_{-j}; c_j, \varepsilon_j, N)$$

where c_j denotes *j*'s *social category*, *N* denotes the norms of behaviour and ideal attributes for different social categories, and ε_j denotes *j*'s own attributes. The inclusion of others’ actions allows for identity externalities. In the simplest case, an individual *j* chooses actions a_j to maximize utility *U*, taking as given c_j, ε_j, and *N* and the actions of others. In some applications, individuals may also choose the category assignment c_j, as social categories may be more or less ascriptive. Individual actions may also affect the norms, *N*, the set of social categories, *C*, as well as the status of different categories reflected in I_j•(·). With respect to gender, for example, the women's movement strived to reduce status differences between men and women and change prescribed behaviour. Gender categories themselves have become varied and complex over time. There may be no universal agreement about social categories and prescriptions. Indeed, they are the subject of much debate and controversy and the source of new

externalities.

This utility function highlights a different motivation for behaviour from a standard model, and shows how social identity can affect economic outcomes. For example, in the workplace different workers may feel more or less part of an organization (*insiders* versus *outsiders*), and work incentives will depend on norms for these different categories of workers. This utility function has implications for supervisory and management policy, as in Akerlof and Kranton (2005). A firm could choose a strict supervisory policy where a supervisor reports to upper management on workers' behaviour. This policy yields greater information, but can lead to workers adopting an *outsider* identity, with lower work norms. A looser supervisory policy yields less information to management, but workers develop a work group identity with possibly higher work norms. We use our utility function to explore the implications of identity in other realms, including race and poverty (Akerlof and Kranton, 2000), gender in the labour market (Akerlof and Kranton, 2000), and schools, student identity and education (Akerlof and Kranton, 2002).

See Also

- culture and economics
- social norms

Bibliography

Akerlof, G. and Kranton, R. 2000. Economics and identity. *Quarterly Journal of Economics* 115, 715–33.

Akerlof, G. and Kranton, R. 2002. Identity and schooling: some lessons for the economics of education. *Journal of Economic Literature* 40, 1167–201.

Akerlof, G. and Kranton, R. 2005. Identity and the economics of organizations. *Journal of Economic Perspectives* 19(1), 9–32.

Bénabou, R. and Tirole, J. 2006. Incentives and prosocial behavior. *American Economic Review* 96(5), 1652–78.

Elster, Jon. 1989. Social norms and economic theory. *Journal of Economic Perspectives* 3(4), 99–117.

Haslam, S. 2001. *Psychology in Organizations: The Social Identity Approach*. London, UK and Thousand Oaks, CA: Sage.

K•szegi, B. 2006. Ego utility, overconfidence, and task choice. *Journal of the European Economic Association* 4, 673–707.

Sen, A. 1985. Goals, commitment, and identity. *Journal of Law, Economics, and Organization* 1, 341–55.

How to cite this article

Akerlof, George and Rachel Kranton. "identity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000260> doi:10.1057/9780230226203.0763

The New Palgrave Dictionary of Economics Online

immigration and the city

Anna Hardman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

At the end of the 20th century, international migrants, legal and undocumented, were a highly visible and economically significant feature of major cities in high- and middle-income countries, including the United States. As numbers of immigrants rose, many were concentrated spatially in a small number of cities (‘ports of entry’) and within those cities in ethnically homogeneous neighbourhoods, enclaves or ghettos. An extensive literature documents the impact of immigrants on host cities, examines their patterns of assimilation and explores their interactions with native-born populations and previous immigrants.

Keywords

economies of agglomeration; ghettoes; housing markets; information sharing; internal migration; international migration; labour market discrimination; migrants in cities; neighbourhoods; network externalities; residential assimilation; residential segregation; selection bias and self-selection; spatial correlation; urban agglomeration; urban economics

Article

At the end of the 20th century, international migrants, legal and undocumented, were again a highly visible and economically significant feature of major cities in high- and middle-income countries, including the United States. As numbers of immigrants rose, many were concentrated spatially in a small number of cities (‘ports of entry’) and within those cities in ethnically homogeneous neighbourhoods (enclaves or ghettos). That was not a new phenomenon: in the ‘first great migration’ to the United States in the late 19th and early 20th centuries immigrants were highly concentrated and a highly visible feature of the largest cities. In 1870, the foreign-born constituted 35.6 per cent of the population of US cities over 100,000 and almost 50 per cent of the population of San Francisco and Chicago, though only 14.4 per cent of the national population. By 1940 the immigrant share had declined to 16.2 per cent of

the population of cities over 100,000 and 8.8 per cent of the total population (Gibson and Lennon, 1999, Tables 18 and 23).

In 15 Organisation for Economic Co-operation and Development (OECD) countries at the beginning of the 21st century the foreign-born made up between 8.3 and 32.6 per cent of the national population (Dumont and Lemaitre, 2005, Table 1). In the United States in 2000, the foreign-born constituted 26.9 per cent of the population in the central cities of metropolitan areas with a population of five million or more and 16.2 per cent in the suburbs. In the cities of New York and Los Angeles the foreign-born made up 35.9 and 40.9 per cent of the population respectively. In the ten metropolitan areas with the largest immigrant populations the foreign-born were between 35 and 54.9 per cent of the population. The foreign-born were correspondingly rare outside large cities: less than four per cent of the population in metropolitan areas with a population of 500,000 or less and even rarer outside metropolitan areas (US Census of Population, 2000). Similarly in the UK in 2001 8.3 per cent of the total population was born overseas. In the same year, the foreign-born were about 25 per cent of the London metropolitan area's total population, and were concentrated in a few neighbourhoods. For example, in Southall, Wembley, Hyde Park and Kensington, over 45 per cent of the population was foreign-born (National Statistics, 2005; BBC News, 2007).

As a result, whereas until the 1970s race and ethnicity were typically absent from analyses of urban economies (in Europe) or modelled as a black–white dichotomy (in the United States), by the mid-1980s economists had begun to explore the impact of immigrants from a wide range of source countries on cities beyond their effect on the wages and employment of natives. Urban economists have explored residential assimilation, looking at location choices, crowding and housing tenure and asked whether the location and housing consumption of immigrants relative to natives has differed because of selection, country of origin or changing make-up of successive cohorts of immigrants. The literature is dominated by studies of the United States both because of its rapidly growing immigrant population and because of micro (individual or household-level) and spatially disaggregated data on immigrant status, race and ancestry.

Immigrants are attracted to ports of entry or places with a stock of previous immigrants, because migration is path-dependent, because immigrants in enclaves benefit from *network externalities* and because immigrant enclaves offer *economies of agglomeration*. As a result, immigrants and particularly unskilled immigrants are less mobile within host societies than the native-born. The behaviour of the native-born in the host economy also drives spatial outcomes, both because of *discrimination* or avoidance of immigrants in labour and housing markets and because natives' location decisions across cities within a host country are more sensitive to wages than those of immigrants. There is evidence that the concentration of immigrants in ports of entry has led some US natives to leave gateway cities or to move to alternative destinations (Filer, 1992).

Early empirical work on immigration and wages estimated the impact of immigration on the labour market using a cross-sectional 'spatial correlations' approach that compared wages over time in metropolitan areas with different proportions of immigrant stocks and flows. The spatial correlations approach generally found weak links at best between the immigrant share and the wages and on employment of natives, both in the USA (Borjas, 1994) and more recently in the UK (Hatton and Tani, 2005). If natives' location decisions are more sensitive to labour conditions than immigrants', then the observed wage impact of immigration is attenuated because it is dispersed across the whole economy

rather than concentrated in the port of entry cities.

The impact of immigration on internal migration has been pursued by geographers and demographers (Wright, Ellis and Reibel, 1997; Kritz and Gurak, 2001) with some expressing fear that the United States faced demographic or spatial ‘balkanization’ or the concentration of immigrants in a few cities shunned by natives (Frey, 1995; 1996). However, in the United States immigrants’ location patterns are changing. In the 1990s growing numbers of immigrants moved to urban and suburban areas remote from the traditional ports of entry. The immigrant population grew more rapidly in non-traditional destinations. For example, the US 2000 Census found that in ten metropolitan areas (with a median population of over 160,000) over two-thirds of the foreign-born population had entered the USA in the previous decade. The new immigrants were moving to metropolitan areas where only a median 4.55 per cent of the population was foreign-born by 2000. Some were in states without a recent tradition of immigration (Iowa, Indiana, North and South Dakota and Nebraska); others were in or close to states with a significant immigrant presence already (Arizona, Georgia, North Carolina and Tennessee).

A notable feature since 1985 is the increasing dispersion of Mexican immigrants, who for a long time were highly concentrated in Los Angeles and elsewhere in Southern California and Texas (Alba et al., 1999). That migration is also credited with changing the industry mix in destination regions (Card and Lewis, 2007). Immigrants who move again within the USA have higher skills than other new immigrants. Moreover, migration beyond immigrant gateways and enclaves is associated with faster assimilation, although this is in part probably attributable to reverse causation since secondary migrants are self-selected (Zhang, 2004; 2006).

In the absence of detailed information on the immediate spatial areas where immigrants live, most of the analysis of residential location, however, focuses on individuals. Immigrants often live in households with partners or family members who are natives or second- or third-generation immigrants. Household-level analysis of confidential Current Population Survey data for Los Angeles shows much greater dispersion of immigrants living in mixed households (Ellis and Wright, 2005).

The urban economics literature on immigrants in cities has been concerned with ‘residential assimilation’: the progress of new immigrants towards parity with natives in housing tenure, consumption of housing and intra-city location, in or outside of ethnic enclaves (see, for example, Painter, Gabriel and Myers, 2001). While US immigrant homeownership rates are consistently lower than natives’, they rise with age and years. Increases in the gap between natives and immigrants in homeownership rates between 1980 and 2000 are explained by differences in location decisions and by changes in the national origin mix of the immigrant population that are associated with lower skills and wages for the most recent immigrant waves (Borjas, 2002).

In contrast to labour markets, where impacts of immigration on wages have been elusive, there is evidence that the arrival of immigrants raises metropolitan area housing prices and rents (Saiz, 2003). Saiz and Wachter (2006) also find immigration associated with relatively slower house price appreciation in immigrant enclaves. The latter is attributed both to native avoidance and to low-income immigrants’ preference for the cheapest housing.

Another facet of housing consumption and hence residential assimilation is residential ‘crowding’ (large numbers of occupants per dwelling or per room). Crowding increased in the USA in the 1980s and the 1990s, after decreases in every decade from 1940 to 1980; the increases were almost all in areas with large concentrations of immigrants. Cohort studies have found that immigrants initially choose higher

densities, which decline with time in the United States for most ethnic groups with the exception of Hispanic immigrants (Myers and Lee, 1996; Simmons, 2002).

Economists have begun to explore the role of ethnic enclaves, neighbourhoods with a high concentration of immigrants, usually from the same source country or region. They are characterized by forces that parallel those that drive the formation of cities and concentrations of firms: shared inputs (enclaves offer stores which provide ethnic foods, clothing, goods used both for consumption and for production by local firms); information-sharing as immigrants in enclaves benefit from news of job opportunities and learn skills essential in the job market and for everyday life in the host country; lastly, new and particularly unskilled immigrants as well as entrepreneurs in the enclave benefit from labour-market pooling in the enclave labour market.

Empirical studies of enclave economies provide evidence that immigrants value location near others from the same source country or region and that there are measurable economic benefits. Gonzalez (1998) estimated the implicit ‘price of culture’ using 1990 Census data for California and Texas, and found both lower earnings and higher rents for Mexican immigrants in enclaves with larger concentrations of Mexicans. Other studies find that immigrants within enclaves earn less than those outside, but a problem with such studies is that immigrants in enclaves are self-selected. A notable recent finding comes from Edin, Frederiksson and Åslund (2003) who exploit a natural experiment in Sweden in which asylum-seekers and refugees were randomly assigned to different cities. They find evidence that selection bias leads to significant underestimates of the value of living in enclaves, with an earnings gain in the order of four to five per cent for migrants living in enclaves, compared with the earnings losses observed before correcting for selection.

See Also

- ghettoes
- housing supply
- urban agglomeration

Bibliography

Alba, R., Logan, J.R., Stults, B.J., Marzan, G. and Zhang, W. 1999. Immigrant groups in the suburbs: A reexamination of suburbanization and spatial assimilation. *American Sociological Review* 64, 446–60.

BBC News. 2007. Born abroad: an immigration map of Britain. Online. Available at http://news.bbc.co.uk/2/shared/spl/hi/uk/05/born_abroad/around_britain/html/overview.stm, accessed 30 January 2007.

Borjas, G.J. 1994. The economics of immigration. *Journal of Economic Literature* 32, 1667–717.

Borjas, G.J. 2002. Homeownership in the immigrant population. *Journal of Urban Economics* 52, 448–76.

Card, D. and Lewis, E. 2007. The diffusion of Mexican immigrants during the 1990s: explanations and impacts. In *Mexican Immigration to the United States*, ed. G. Borjas. Chicago: University of Chicago Press and NBER.

Dumont, J.-C. and Lemaitre, G. 2005. Counting immigrants and expatriates in OECD countries: a new perspective. Social, Employment and Migration Working Papers No. 25. Paris: OECD.

Edin, P.-A., Fredriksson, P. and Åslund, O. 2003. Ethnic enclaves and the economic success of immigrants – evidence from a natural experiment. *Quarterly Journal of Economics* 118, 329–57.

Ellis, M. and Wright, R. 2005. Assimilation and differences between the settlement patterns of individual immigrants and immigrant households. *Proceedings of the National Academy of Sciences* 102, 15325–30.

Filer, R.K. 1992. The effect of immigrant arrivals on migratory patterns of native workers. In *Immigration and the Work Force: Economic Consequences for the United States and Source Areas*, ed. G.J. Borjas and R.B. Freeman. Chicago: University of Chicago Press.

Frey, W.H. 1995. Immigration and internal migration flight from US metropolitan areas: toward a new demographic balkanisation. *Urban Studies* 32, 733–57.

Frey, W.H. 1996. Immigration, domestic migration, and demographic balkanization in America: new evidence for the 1990s. *Population and Development Review* 22, 741–63.

Gibson, C.J. and Lennon, E. 1999. Historical census statistics on the foreign-born population of the United States: 1850–1990. Working Paper No. 29. Washington, DC: Population Division, US Bureau of the Census.

Gibson, C.J. and Jung, K. 2006. Historical census statistics on the foreign-born population of the United States: 1850–2000. Population Division Working Paper No. 81. Washington, DC: US Bureau of the Census.

Gonzalez, A. 1998. Mexican enclaves and the price of culture. *Journal of Urban Economics* 43, 273–91.

Hatton, T.J. and Tani, M. 2005. Immigration and inter-regional mobility in the UK 1982–2000. *Economic Journal* 115, 342–58.

Kritz, M.M. and Gurak, D.T. 2001. The impact of immigration on the internal migration of natives and immigrants. *Demography* 38, 133–45.

Myers, D. 1999. Immigration: fundamental force in the American city. *Housing Facts and Findings* 1 (4), 3–5. Fannie Mae Foundation.

Myers, D. and Lee, S.W. 1996. Immigration cohorts and residential overcrowding in southern California. *Demography* 33, 51–65.

National Statistics 2005. People and migration: Foreign-born. Online. Available at http://www.statistics.gov.uk/cci/nugget.asp?id=1312, accessed 1 February 2007.

Painter, G., Gabriel, S. and Myers, D. 2001. Race, immigrant status and housing tenure choice. *Journal of Urban Economics* 49, 150–67.

Saiz, A. 2003. Room in the kitchen for the melting pot: immigration and rental prices. *Review of Economics and Statistics* 85, 502–21.

Saiz, A. and Wachter, S. 2006. Immigration and the neighborhood. Working Paper 06–22, Research Department, Federal Reserve Bank of Philadelphia.

Simmons, P.A. 2002. Patterns and trends in overcrowded housing: early results from Census 2000. Fannie Mae Foundation Census Note 09. Fannie Mae Foundation. Online. Available at fanniemaefoundation.org/programs/pdf/census/census_note9.pdf, accessed 1 February 2007.

US Census of Population. 2000. United States – Metropolitan Areas. SF 3 Table GCT-P10. Online. Available at http://factfinder.census.gov, accessed 2 April 2007.

Wright, R.A, Ellis, M. and Reibel, M. 1997. The linkage between immigration and internal migration in large metropolitan areas in the United States. *Economic Geography* 73, 234–54.

Zhang, W. 2004. Secondary migration of recent immigrants in the United States: a study of Mexicans and Chinese. Ph.D. thesis, State University of New York at Albany.

Zhang, W.C. 2006. Internal migration of Mexicans in the United States, 1990 and 2000. Brown University Department of Sociology. Presented at the Annual Meeting of the Population Association of America, Los Angeles, and 30 March–1 April.

How to cite this article

Hardman, Anna. "immigration and the city." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000254> doi:10.1057/9780230226203.0764

The New Palgrave Dictionary of Economics Online

immiserizing growth

Jagdish N. Bhagwati
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

autarky; directly unproductive profit-seeking (DUP) activities; distortions; free trade; immiserizing growth; shadow pricing; tariffs; tariff seeking; terms of trade

Article

The theory of immiserizing growth has been developed by theorists of international trade, though it has recently been the focal point of research also by mathematical economists. It is central to understanding several important paradoxes in economic theory and has significant policy implications.

That growth in a country could immiserize it is a paradox that was first noted by trade theorists such as Bhagwati (1958) and Johnson (1955) in the context of the post-war discussions of dollar shortage. They established conditions under which, in a two-country, two-traded-goods framework of conventional theory, the growth-induced deterioration in the terms of trade would outweigh the primary gain from growth. It was shown that this paradox, unlike the paradox of donor-enriching and recipient-immiserizing transfers, was compatible with Walras-stability.

The phrase ‘immiserizing growth’ was invented by Bhagwati (1958) and has now been widely accepted (including by literary editors who have long ceased to insist on changing it to the correct English versions such as ‘immiserating’), the theory itself being generally attributed (for example, Johnson, 1967) to this 1958 article. Interestingly, as often in economics, Bhagwati happened to chance upon an early contribution by Edgeworth (1894), where Edgeworth developed an example of what he called ‘indamnifying’ growth; and the controversy surrounding this result at the time and its relationship to the Bhagwati–Johnson analyses of the 1950s was reviewed in Bhagwati and Johnson (1960).

Later, Johnson (1967) demonstrated another paradox of immiserizing growth. If a small country had a distortionary tariff in place, and then exogenously it experienced growth, the result again could be to immiserize the country. Later, Bertrand and Flatters (1971) and Martin (1977) established formally the conditions under which this new paradox of immiserizing growth could arise.

Bhagwati (1968) got to the bottom of these paradoxes and produced the central insight that explains why

these, and other immiserizing-growth paradoxes, can readily arise. He showed that, if an economy was suboptimally organized, the primary gain from growth, measured hypothetically as if the economy had an optimal policy in place before and after the growth, could be outweighed by accentuation of the loss from the distortion-induced suboptimality when growth occurred. In the original Bhagwati (1958) example, since the terms of trade could deteriorate, the economy had monopoly power in trade but was following free trade policy which is evidently suboptimal. In the Johnson (1967) example, the tariff was being used by a small country with given terms of trade and was therefore also a suboptimal policy. In both cases the suboptimal policy produced losses which were accentuated by the growth and then managed to outweigh the primary gains from growth that would have occurred if optimal policies were in place. The result was a powerful generalization that placed the theory of immiserizing growth squarely into the central theory of distortions and policy intervention (Srinivasan, 1987) that lies at the core of the modern theory of trade and welfare. Evidently, immiserizing-growth paradoxes could arise only if there was a distortion present.

This central result has immediate implications. If an economy has a suboptimal money supply, growth could be immiserizing. If trade policy is highly distorted, growth could be immiserizing. The well-known results of trade theory, which show that free trade need not be welfare-improving relative to autarky (for example, Haberler, 1950) under distortions are also seen as instances of immiserizing-growth theory; free trade augments the availability set relative to autarky, implying ‘as-if’ growth, and if distortions are present, then there is no surprise to the immiseration that free trade brings. Again, if a country uses tariffs to induce foreign investment (the so-called tariff-jumping investment that developing countries often used in the post-war period), such investment could immiserize the host country: this being a simple extension of the Johnson (1967) demonstration, argued to be relevant to analysis of developing countries in Bhagwati (1978), and analysed extensively in Bhagwati (1973), Brecher and Alejandro (1977), Hamada (1974), Minabe (1974), Uzawa (1969) and Brecher and Findlay (1983). Yet another important insight from the immiserizing-growth theory is that, in the new and growing theory of DUP (directly-unproductive profit-seeking) activities, which incorporates several quasi-political activities essentially into the corpus of economic theory, a DUP activity that wastes resources directly need not cause ultimate loss of welfare. This is because the waste may occur from a suboptimal situation, thus resulting in welfare-improvement paradoxically. This is the obverse of immiserizing growth: in one case, growth immiserizes; in the other, throwing away or wasting resources enriches. This is at the heart of the contention in Bhagwati (1980) that an exogenous tariff at *t* per cent may be welfare-superior to an endogenous tariff, procured by tariff-seeking lobbies that have diverted uses to such DUP activity, also at *t* per cent. Several such implications of the theory of immiserizing growth are discussed in Bhagwati and Srinivasan (1983, ch. 25).

Two further developments need to be cited. First, the dual of immiserizing growth, when such growth is due to factor accumulation, clearly yields negative shadow factor prices. This aspect is relevant to certain formulations in cost–benefit analysis; see, in particular, Findlay and Wellisz (1976), Diamond and Mirrlees (1976), Srinivasan and Bhagwati (1978), Bhagwati, Srinivasan and Wan (1978) and Mussa (1979).

Next, mathematical economists such as Aumann and Peleg (1974), and then Mas-Colell (1976) and Mantel (1984) among others, have rediscovered the original immiserizing-growth paradox, illustrating

how economists working apart or in different traditions may rediscover one another's findings, often decades apart. A synthesis of the two literatures has been provided in Bhagwati, Brecher and Hatta (1984). A complete and formal reconciliation of the conditions established in Bhagwati (1958) and in Mas-Colell (1976) and Mantel (1984) for the original immiserizing-growth paradox is provided by Hatta (1984).

See Also

- directly unproductive profit-seeking (DUP) activities
- terms of trade

Bibliography

Aumann, R.J. and Peleg, B. 1974. A note on Gale's example. *Journal of Mathematical Economics* 1, 209–11.

Bertrand, T. and Flatters, F. 1971. Tariffs, capital accumulation and immiserizing growth. *Journal of International Economics* 1, 453–60.

Bhagwati, J. 1958. Immiserizing growth: a geometrical note. *Review of Economic Studies* 25, 201–5. Reprinted in *International Trade: Selected Readings*, ed. J. Bhagwati, Cambridge, MA: MIT Press, 1981.

Bhagwati, J. 1968. Distortions and immiserizing growth: a generalization. *Review of Economic Studies* 35, 481–5. Reprinted in J. Bhagwati, *The Theory of Commercial Policy*, vol. 1, Cambridge, MA: MIT Press, 1983.

Bhagwati, J. 1973. The theory of immiserizing growth: further applications. In *International Trade and Money*, ed. M. Connolly and A. Swoboda. Toronto: University of Toronto Press.

Bhagwati, J. 1978. *Foreign Trade Regimes and Economic Development: The Anatomy and Consequences of Exchange Control*. Cambridge, MA: Ballinger.

Bhagwati, J. 1980. Lobbying and welfare. *Journal of Public Economics* 14, 355–63.

Bhagwati, J., Brecher, R. and Hatta, T. 1984. The paradoxes of immiserizing growth and donor-enriching ‘recipient-immiserizing’ transfers: a tale of two literatures. *Weltwirtschaftliches Archiv* 120(4), 228–43.

Bhagwati, J. and Johnson, H.G. 1960. Notes on some controversies in the theory of international trade. *Economic Journal* 60, 74–93.

Bhagwati, J. and Srinivasan, T.N. 1983. *Lectures on International Trade*. Cambridge, MA: MIT Press.

Bhagwati, J. and Srinivasan, T.N. and Wan, H., Jr. 1978. Value subtracted, negative shadow prices of factors in project evaluation, and immiserizing growth: three paradoxes in the presence of trade distortions. *Economic Journal* 88, 121–5.

Brecher, R. and Diaz-Alejandro, C. 1977. Tariffs, foreign capital and immiserizing growth. *Journal of International Economics* 7, 317–22. Reprinted in *International Trade: Selected Readings*, ed. J. Bhagwati, Cambridge, MA: MIT Press, 1981.

Brecher, R. and Findlay, R. 1983. Tariffs, foreign capital and national welfare with sector-specific factors. *Journal of International Economics* 14, 277–88.

Diamond, P. and Mirrlees, J. 1976. Private constant returns and public shadow prices. *Review of Economic Studies* 43, 41–8.

Edgeworth, F.Y. 1894. The theory of International values. *Economic Journal* 4, 35–50, 424–43, 606–38.

Findlay, R. and Wellisz, S. 1976. Project evaluation, shadow prices and trade policy. *Journal of Political Economy* 84, 543–52.

Haberler, G. 1950. Some problems in the pure theory of international trade. *Economic Journal* 60, 223–40.

Hamada, K. 1974. An economic analysis of the duty-free zone. *Journal of International Economics* 4, 225–41.

Hatta, T. 1984. Immiserizing growth in a many-economy setting. *Journal of International Economics* 17, 335–45.

Johnson, H.G. 1955. Economic expansion and international trade. *Manchester School of Economic and Social Studies* 23(2), 95–112.

Johnson, H.G. 1967. The possibility of income losses from increased efficiency or factor accumulation in the presence of tariffs. *Economic Journal* 77, 151–4. Reprinted in *International Trade: Selected Readings*, ed. J. Bhagwati, Cambridge, MA: MIT Press, 1981.

Mantel, R. 1984. Substitutability and the welfare effects of endowment increases. *Journal of International Economics* 17, 325–34.

Martin, R. 1977. Immiserizing growth for a tariff-distorted, small economy. *Journal of International*

Economics 3, 323–6.

Mas-Colell, A. 1976. En torno a una propiedad poco atractiva del equilibrio competitivo. *Moneda y Credito* 136, 11–27.

Minabe, N. 1974. Capital and technology movements and economic welfare. *American Economic Review* 64, 1088–100.

Mussa, M. 1979. The two-sector model in terms of its dual: a geometric exposition. *Journal of International Economics* 9, 513–26. Reprinted in *International Trade: Selected Readings*, ed. J. Bhagwati, Cambridge, MA: MIT Press, 1981.

Srinivasan, T.N. and Bhagwati, J. 1978. Shadow prices for project selection in the presence of distortions: effective rates of protection and domestic resource costs. *Journal of Political Economy* 86, 97–116. Reprinted in *International Trade: Selected Readings*, ed. J. Bhagwati, Cambridge, MA: MIT Press, 1981.

Uzawa, H. 1969. Shinon jiyutato kokumin keizai [Liberalization of foreign investments and the national economy]. *Ekonomisuto* 23, (December), 106–22 (in Japanese).

How to cite this article

Bhagwati, Jagdish N. "immiserizing growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000010> doi:10.1057/9780230226203.0765

The New Palgrave Dictionary of Economics Online

implicit contracts

Costas Azariadis

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

An implicit contract is a theoretical construct meant to describe complex agreements, written and tacit, between employers and employees, which govern the exchange of labour services when various types of job-specific investments inhibit labour mobility and opportunities to shed risk are limited by imperfectly developed markets for contingent claims. This construct differs from the more familiar one of a neoclassical labour exchange in emphasizing a trading process, frequently over a long period of time, between two *specific* economic units (say a worker and a firm, union and management, and so on) rather than the impersonal, and often instantaneous, market process in which wages decentralize and coordinate the actions of labour suppliers and labour demanders.

Adam Smith's exposition of occupational wage differentials (1776, book I, ch. 10) recognized very early the idiosyncratic nature of the labour market and, in particular, that employment risk affected wages in various occupations. Since then economists have accumulated many facts, raw or stylized, which are best understood if one abandons the traditional view that the shadow price of labour is simply the wage rate. Prominent among explananda are the widespread use of temporary layoffs as a means of regulating the volume of employment (Feldstein, 1975); the continuity of jobs by many primary wage earners (Hall, 1982); the collective bargaining tradition of leaving the volume of employment at the discretion of management while predetermining money wage rates two or three years in advance.

To these, one must add certain ‘impressions’ or softer facts about the labour market which arise from the central role labour services possess in macroeconomic models. There is, indeed, among macroeconomists a shared impression (Hall, 1980) that, over a typical business cycle, average real compensation per hour fluctuates considerably less than does the marginal revenue-product of labour or, for that matter, the total volume of employment.

One consequence is that wage and price rigidity are among the key assumptions of Keynesian macroeconomics, both in the Hicksian IS-LM framework and in the concept of quantity-constrained equilibrium originally developed by Clower (1965) and formalized by Bénassy (1975) and Drèze (1975). Another is the overwhelming importance of words like ‘jobs’ and ‘unemployment’, both in our colloquial vocabulary and in the specialized lexicon of economics. In particular, ‘involuntary unemployment’ is for many academic economists the *sine qua non* of modern macroeconomics.

The technically minded reader will find many of these issues surveyed in a number of specialized papers of which the most recent are Hart (1983) and Rosen (1985).

Wages and employment

The earliest literature on implicit contracts exploits an insight of Frank Knight (1921), who argued that inherently ‘confident and venturesome’ entrepreneurs will offer to relieve their employees of some market risks in return for the right to make allocative decisions. The formal development of this idea began with three independently written papers by Baily (1974), Gordon (1974), and Azariadis (1975), motivated by the seeming puzzle of layoffs. In an unusual coincidence, all three authors took the employment relation not simply as a sequential spot exchange of labour services for money, but as a more complicated long-term attachment; labour services are traded in part for an insurance contract that protects workers from random, publicly observed fluctuations in their marginal revenue-product. The idea was that workers could purchase insurance only from their employers, not from third parties.
Risk-averse workers deal with risk-neutral entrepreneurs who head firms consisting of three departments: a production department that purchases labour services and credits each worker with his marginal revenue-product (MRPL); an insurance department that sells actuarially fair policies and, depending on the state of nature, credits the worker with a net insurance indemnity (NII) or debits him with a net insurance premium; and an accounting department that pays each employed worker a wage, w, with the property that w=MRPL+NII in every state of nature.
Favourable states of nature are associated with high values of MRPL; in these the net indemnity is negative and wage falls short of the MRPL. Adverse states of nature correspond to low values of MRPL, to positive net insurance indemnities, and to wages in excess of MRPL. An implicit contract is then a complete description, made before the state of nature becomes known, of the labour services to be rendered unto the firm in each state of nature, and of the corresponding payments to be delivered to the worker. The contract is implementable if we assume the state of nature is as easily verifiable as events are in a normal insurance contract.
An immediate consequence of this framework is that wages are disengaged from the marginal revenue-product of labour. In fact, if the amount of labour performed by employed workers per unit time is fixed institutionally, then each worker's consumption is proportional to the wage rate; an actuarially fair insurance policy should make this consumption independent of the MRPL by stabilizing the purchasing power of wages over states of nature. Therefore, the real wage rate is rigid.
In traditional macroeconomic models of course, wage rigidity by itself is sufficient to cause unemployment: if wages do not adjust for some reason, then neither does the demand for labour. The argument does not carry over to implicit contracts because of the very separation between wages and the marginal revenue product of labour. A complete theory of unemployment must explain why layoffs are preferred to work-sharing in adverse states of nature, and why laid-off workers are worse off than their employed colleagues.
This is not a simple task if one thinks of implicit contracts as ordinary, explicit, timeless insurance contracts between risk-averse workers and risk-neutral entrepreneurs. All contracts of this type would share a basic property of optimum insurance schemes; namely, keeping the worker's marginal utility of consumption independent of all random, publicly observed events – including such events as ‘employment’ or ‘unemployment’.

To explain layoff unemployment, we need to distort or complicate the insurance contract in some significant way. A distortion that was noted early in the implicit contract literature is the dole. In an extremely adverse state of nature, the flow of insurance indemnities to workers can become a substantial drain on profit; one way to staunch losses is to place the burden of insurance on an outside party, the dole.

The practice of layoffs is simply the administrative counterpart of this insurance-shifting manoeuvre; workers consent in advance that some of them may be separated from their jobs in order to become eligible for unemployment insurance (UI) payments from an outside public agency. Furthermore, no worker will contract his labour unless the expected value (utility) of the total package, taken over all possible states of nature, exceeds the value of being on the dole in every state. This means, in turn that employed workers receive a wage in excess of UI payments and are therefore to be envied by their laid-off colleagues – a situation that many economists would call ‘involuntary unemployment’.

The fact that laid-off workers would gladly exchange places with their employed colleagues is not in itself sufficient to establish a misallocation of resources. After all, accident victims may very well envy more fortunate individuals without any implication that the insurance industry works poorly. Layoffs, by themselves, could be no more than the luck of the draw unless we can demonstrate that they constitute, in some sense, socially inefficient underemployment. This is clearly impossible within the Walras–Arrow–Debreu model; and it is for this reason that the early literature on contracts turned to institutions like the dole in order to explain layoff unemployment.

Private information

One fundamental departure from the Walrasian paradigm that received much attention in the early 1980s was a weakening of the information assumptions: information becomes ‘private’ or ‘asymmetric’, which simply means that not everyone is equally informed about the relevant state of nature. This is a perfectly sensible observation, for what justifies the trading of implicit contracts in the first place is that third parties simply are not as well informed about someone's income or employment status as is his employer; the employer, in turn, may be less informed about an employee's non-labour income and job opportunities than is the worker himself.

The thread was picked up by a number of authors who studied the properties of wages and employment for two main cases: in the first, entrepreneurs possess superior information about labour demand (Hall and Lilien, 1979; Grossman and Hart, 1981; Azariadis, 1983; Farmer, 1984); in the second case, workers possess superior information about labour supply, as in Cooper (1983). Suppose, for instance, that wages and employment do not depend on the unobservable true state of nature but on what the better informed contractant (say, the employer) *announces* that state to be. The question now becomes how to design contracts that reward entrepreneurs who tell the truth and punish those who lie.

One desirable property of contracts is that the truth should be the value-maximizing strategy for firms: truth-telling ought to be consistent with equality between the marginal cost and the marginal revenue-product of labour. Furthermore, entrepreneurs who misrepresent actual conditions should be punished, say, for knowingly under-reporting demand.

Under-reporting demand does turn out to be a problem in contracts that permit employers to slash both workforce and the wage bill when demand is slack, and do it in such a manner as to reduce cost more

than revenue. To avoid this temptation, a properly designed contract specifies a highly variable pattern of employment over states of nature; that is, one in which employment is below what is socially optimal and the marginal product of labour is correspondingly above the marginal rate of substitution between consumption and leisure. It is in this sense that asymmetric information is said to result in socially inefficient underemployment or unemployment.

What relation is there between the layoffs we all know and the inefficient underemployment of a model economy that suffers from asymmetric information? To go from the latter to the former, one must understand first why layoffs are a more common means of reducing employment than is work-sharing. Second, a general equilibrium picture of underemployment would require an explanation of why underemployed (or unemployed) individuals are not hired by other employers. Third, and most important, the unemployment found in this private-information story is a response to private, firm-specific risk; most economists, however, consider the unemployment observed in market economies to be a reaction to social risks, especially to business cycles set in motion by aggregate demand disturbances. Unless one intends to make the far-fetched claim that the general public is unaware of, or cannot observe, whatever disturbances set off business cycles (such as changes in government consumption, money supply or consumer confidence), does it not appear that information-based unemployment simply describes the behaviour of an isolated firm?

The answer is not obvious. Note, however, that in order to have an inefficient volume of equilibrium employment, it is sufficient that *some but not all* information be private. In fact, it is not difficult to imagine general equilibrium extensions of the work we are discussing that would include both public and private information. Such extensions will be useful, especially if they manage to establish a firm link between inefficient underemployment and extreme values of some publicly observed aggregate disturbance.

Empirical implications

Whether information is publicly shared or in the private domain, wages in implicit contracts do not merely reflect the marginal product of labour or the workers' marginal rate of substitution between consumption and leisure, as they might in more conventional theories. The empirical implications of this insight are just being worked out, and they seem to be quite considerable. At the most aggregative level, one can make sense of the oft-verified fact (Neftci, 1978) that hourly wages in manufacturing show little cyclical variability and are best described as a random walk.

In fact, it seems preferable to have empirical investigations of this sort at a less aggregated level. Aggregate studies are victims of selection bias: they fail to capture changes in the composition of output or of the labour force, which are themselves sufficient to induce substantial cyclical movement in economy-wide wages even if the business cycle does not affect the real wage of any skill grade in any industry.

Consider, for instance, a fictitious economy with homogeneous labour in which almost all industries experience little cyclical fluctuation except one, the quad industry, which is thoroughly buffeted by the business cycle. If labour mobility is good across industries, quad workers will suffer more layoffs and enjoy a wage higher than elsewhere whenever they are employed. The economy-wide average wage will vary procyclically.

Another phenomenon accounted for naturally by implicit contracts is the behaviour of occupational

wage differentials (that is of the unskilled-to-skilled wage ratio). These have shown a definite countercyclical tendency, widening in contractions and narrowing in booms, both in the United States and in the UK.

To see why, suppose that we drop the postulate of labour homogeneity in the economy just described and admit two skill grades. For simplicity, assume that the cycle is of such amplitude that there is no unemployment outside the quad industry, while unemployment in the quad industry falls solely on common labourers. These workers are thus the only group in the economy to suffer layoffs; in return they receive a wage above that of common workers outside the quad industry and below that of skilled workers – in the quad industry or out. As the cycle unfolds, then, the economy-wide wage average for craftsmen remains unaltered, the one for labourers changes procyclically, and occupational wage differentials follow a countercyclical pattern.

Intertemporal labour supply models of the type pioneered by Lucas and Rapping (1969) are another area that may in the future make fruitful use of implicit contracts. Econometric work on intertemporal labour substitution identifies the preferences of a ‘typical’ working household from time-series data on wages and salaries. The outcome is invariably an estimate of the wage-elasticity of labour supply that is so low as to be inconsistent with time-series data on employment (Kydland and Prescott, 1982). In other words, someone who believes that the wage rate represents an important conditioning factor for labour supply and demand will find that wage rates do not vary sufficiently over the business cycle to account for observed fluctuations in employment.

Employment in an implicit contract, however, reflects the underlying value of labour's marginal revenue-product, whereas wages are smoothed averages of the MRPL over time or states of nature. Small fluctuations in contract wages are in principle consistent with substantial variations in contract employment; whether these are mutually consistent *in practice* remains to be seen from empirical work.

Macroeconomic aspects

From empirical labour economics we turn to the macroeconomic issues that provided the original impetus for the development of implicit contracts. Unemployment, says this theory, is the result of differential information: a credible signal from employers to employees that product demand is slackening, or one from employees to employers that job opportunities are really better elsewhere. Newer ideas that seem to be building on this basic piece of intuition are outlined later in this article. But whatever progress we have made towards understanding fluctuations in employment has not dispelled the dense fog that still shrouds the issue for wage rigidity. All we have to go on is the early result of Martin Baily that insurance makes the wage rate less variable than it otherwise might be. This stickiness, however, is a property of the *real* rather than the nominal wage rate, and it is the latter that is assumed to be rigid in Keynesian macroeconomics.

Rigidity, of course, does not necessarily imply complete time-invariance, nor does it require money wages to change less frequently than other prices; it is simply an information-processing failure. The standard procedure in collective bargains, for instance, is to predetermine money wages several years in advance; more often than not those wages are invariant to any information that may accumulate over the duration of the contract. Only in exceptional circumstances are money wages in the United States allowed to reflect *any* contemporaneous developments in the cost of living (indexation) or in the profitability of the employer (bankruptcy).

The mystery of wage rigidity is then the failure of contracts to set money wages as *functions* of publicly available information that is obviously relevant to the welfare of all parties. Why does the wage-setting process choose to ignore this information? One answer is transaction costs and/or bounded rationality: contracts are cheaper to evaluate and implement when they are defined by a few simple numbers rather than by complicated rules that condition employment or wages on contingent events. Another possibility is to exploit the great multiplicity of equilibria that is typical of economies with missing securities markets (Azariadis and Cooper, 1985). One of these equilibria features predetermined prices and wages, while employment and other quantity variables adjust fully to short-term disturbances. Wage rigidity here is like a Nash equilibrium: it is the best response of a firm in a labour market in which the wages paid by all other firms fail to reflect new information instantaneously.

Implementation

An implicit contract is formally defined as a collection of schedules describing how the terms of employment for one person or group of persons change in response to unexpected changes in the economic environment. What brings contractants together? How detailed are their agreements? And what mechanisms are there to enforce such agreements once they are reached? After an initial stage of fairly rapid development, research is returning to these elementary questions as if trying to clarify the axiomatic basis of the underlying theory.

What brings potential contractants together is the opportunity jointly to reap substantial returns on investments peculiar to their relationship. The idea is apparent in Becker's theory of specific human capital (1964) and in Williamson's hypothesis (1979) of physical assets that are specific to a given supplier–customer pair. To reap any returns, contractants must wed themselves to one partner, forsaking all others, for some period of time. Maintaining such a special relationship involves the transactions costs of creating an idiosyncratic asset, as well as an implicit contract; that is, a number of rules that define how the partners have decided to share the returns in various possible future circumstances.

There are, of course, circumstances that are not explicitly covered, either because they are not observable at reasonable cost or because contractants think of them as unlikely or unworthy of note. Irrespective of the possible events that are covered and of the prior rules that govern the distribution of returns to shared investments, all contractants are required to bear risk and to subordinate their short-term interest to longer-term considerations.

Workers, for instance, suffer layoffs in recessions while firms hoard labour in order to preserve a long-term relationship. What mechanisms keep contractants together in adverse circumstances?

One mechanism – studied extensively by Radner (1981), Townsend (1982) and others – is reputation: if somebody deviates from the terms of the contract, the deviation becomes widely known, and the deviant finds it difficult to locate trading partners in the future. That works well if the time horizon is fairly long or the future is fairly important relative to the present; reputations are likely to be important for firms, less so for workers.

Another method of enforcement is by a third party: a monitor, arbitrator or court of law. In order for a third party to enforce a contract, it has to be able to observe all the prices and all the quantities specified in it – the employment status, hours worked and wage rate of every worker. That is an unreasonably large informational burden to place on someone who is outside the special relationship called a contract.

Outsiders can be expected to observe at low cost only certain aggregates or averages, but not very much in the way of idiosyncratic detail.

How does one design and enforce contracts when outsiders are poorly informed about the trades among contractants? According to Hölmstrom (1983) and Bull (1986), self-interest will enforce contracts that third parties are not sufficiently informed to implement.

In particular, workers will put in the required amount of effort on the job, not because effort can be ascertained easily by an outside arbitrator but rather because they know that their wages and speed of promotion depend on performance. And employers will be careful not to break even the most implicit of their commitments if doing so will compromise their ability to attract workers in the future. As of this writing, the design of self-enforcing contracts seems to be the central theoretical problem in the field of implicit contracts.

See Also

- labour economics
- layoffs

Bibliography

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83, 1183–202.

Azariadis, C. 1983. Employment with asymmetric information. *Quarterly Journal of Economics* 98, Supplement, 157–72.

Azariadis, C. and Cooper, R. 1985. Nominal wage-price rigidity as a rational expectations equilibrium. *American Economic Review, Papers and Proceedings* 75, 31–5.

Baily, M. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41, 37–50.

Becker, G. 1964. *Human Capital*. New York: Columbia University Press.

Bénassy, J.-P. 1975. Neo-Keynesian disequilibrium in a monetary economy. *Review of Economic Studies* 42, 502–23.

Bull, C. 1986. The existence of self-enforcing implicit contracts. *Quarterly Journal of Economics*.

Clower, R. 1965. The Keynesian counter-revolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cooper, R. 1983. A note on overemployment/underemployment in labor contracts under asymmetric information. *Economics Letters* 12, 81–7.

Drèze, J. 1975. Existence of an equilibrium under price rigidity and quantity rationing. *International Economic Review* 16, 301–20.

Farmer, R. 1984. A new theory of aggregate supply. *American Economic Review* 74, 920–30.

Feldstein, M. 1975. The importance of temporary layoffs: an empirical analysis. *Brookings Papers on Economic Activity* 3, 725–44.

Gordon, D.F. 1974. A neoclassical theory of Keynesian unemployment. *Economic Inquiry* 12, 431–49.

Grossman, S. and Hart, O. 1981. Implicit contracts, moral hazard and unemployment. *American Economic Review, Papers and Proceedings* 71, 301–7.

Hall, R. 1980. Employment fluctuations and wage rigidity. *Brookings Papers on Economic Activity* 1, 91–124.

Hall, R. 1982. The importance of lifetime jobs in the US economy. *American Economic Review, Papers and Proceedings* 72, 716–24.

Hall, R. and Lilien, D. 1979. Efficient wage bargains under uncertain supply and demand. *American Economic Review* 69, 868–79.

Hart, O. 1983. Optimal labour contracts under asymmetric information: an introduction. *Review of Economic Studies* 50, 3–35.

Hölmstrom, B. 1983. Equilibrium long-term labor contracts. *Quarterly Journal of Economics* 98, Supplement, 23–54.

Knight, F. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lucas, R. and Rapping, L. 1969. Real wages, employment and inflation. *Journal of Political Economy* 77, 721–54.

Neftci, S. 1978. A time-series analysis of the real wages–employment relationship. *Journal of Political Economy* 86, 281–91.

Radner, R. 1981. Monitoring cooperative agreements in a repeated principal–agent relationship. *Econometrica* 49, 1127–48.

Rosen, S. 1985. Implicit contracts: a survey. *Journal of Economic Literature* 23, 1144–75.

Smith, A. 1776. *The Wealth of Nations*. London: Pelican Books, 1970.

Townsend, R. 1982. Optimal multiperiod contracts and the gain from enduring relationships under private information. *Journal of Political Economy* 90, 1166–86.

Williamson, O. 1979. Transaction-cost economics: the governance of contractual relations. *Journal of Law and Economics* 22, 233–61.

How to cite this article

Azariadis, Costas. "implicit contracts." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000017> doi:10.1057/9780230226203.0766

The New Palgrave Dictionary of Economics Online

impulse response function

Helmut Lütkepohl

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Impulse response functions are useful for studying the interactions between variables in a vector autoregressive model. They represent the reactions of the variables to shocks hitting the system. It is often not clear, however, which shocks are relevant for studying specific economic problems. Therefore structural information has to be used to specify meaningful shocks. Structural vector autoregressive models and the estimation of impulse responses are discussed and extensions to models with cointegrated variables or nonlinear features are considered.

Keywords

Bayesian methods; bootstrap; cointegrated variables; cointegration; conditional moment profiles; dynamic multipliers; forecast error impulse responses; generalized impulse responses; impulse response functions; integrated variables; least squares; linear models; maximum likelihood; nonlinear time series models; orthogonalized impulse responses; simultaneous equations models; structural impulse responses; structural vector autoregressions; vector autoregressions; Wold causal ordering; Wold moving average

Article

Sims (1980) questioned the way classical simultaneous equations models were specified and identified. He argued in particular that the exogeneity assumptions for some of the variables are often problematic. As an alternative he advocated the use of vector autoregressive (VAR) models for macroeconometric analysis. These models have the form

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

where $y_t = (y_{1t}, \ldots, y_{Kt})'$ (the prime denotes the transpose) is a vector of K observed variables of interest, the A_i's are $(K \times K)$ parameter matrices, p is the lag order and u_t is an error process which is assumed to be white noise with zero mean, that is, $E(u_t) = 0$, the covariance matrix, $E(u_t u_t') = \Sigma_u$, is time invariant and the u_t's are serially uncorrelated or independent. There are usually also deterministic terms such as constants, seasonal dummies or polynomial trends. These terms are neglected here because they are not of interest in what follows. The relations between the variables in a VAR model are difficult to see directly from the parameter matrices. Therefore, *impulse response functions* have been proposed as tools for interpreting VAR models.

A VAR model can be written more compactly as $A(L)y_t = u_t$, where the lag or back-shift operator L is defined such that $Ly_t = y_{t-1}$ and $A(L) = I_K - A_1 L - \cdots - A_p L^p$ is a matrix polynomial in the lag operator. If the polynomial in z defined by det $A(z)$ has all its roots outside the complex unit circle, the process is stationary and has a Wold moving average (MA) representation

$$y_t = A(L)^{-1} u_t = u_t + \sum_{i=1}^{\infty} \Phi_i u_{t-i}. \tag{1}$$

In this framework impulse response analysis may be based on the counterfactual experiment of tracing the marginal effect of a shock to one variable through the system by setting one component of u_t to one and all other components to zero and evaluating the responses of the y_t's to such an impulse as time goes by. These impulse responses are just the elements of the Φ_i matrices. Because the u_t's are the one-step ahead forecast errors of the system, the resulting functions are sometimes referred to as *forecast error impulse responses* (for example, Lütkepohl, 2005, section 2.3.2).

Such a counterfactual experiment may not properly reflect the actual responses of an economic system of interest because the components of u_t are instantaneously correlated, that is, Σ_u may not be a diagonal matrix. In that case, forecast error impulses are just not the kinds of impulses that occur in practice, because an impulse in one variable is likely to be accompanied by an impulse in another variable and should not be considered in isolation. Therefore, *orthogonalized impulse responses* are often considered in this context. They are obtained from (1) by choosing some matrix B such that $BB' = \Sigma_u$ or such that $B^{-1}\Sigma_u B'^{-1}$ is a diagonal matrix and defining $\varepsilon_t = B^{-1}u_t$. Substituting in (1) gives

$$y_t = B\varepsilon_t + \sum_{i=1}^{\infty} \Theta_i \varepsilon_{t-i}, \tag{2}$$

where $\Theta_i = \Phi_i B$, $i = 1, 2, \ldots$. The ε_t's have a diagonal or even a unit covariance matrix and are hence contemporaneously uncorrelated (orthogonal). Thus, ε_t shocks may give a more realistic picture of the reactions of the system. The problem is, however, that the matrix *B* is not unique and many different orthogonal shocks exist. Thus, identifying restrictions based on non-sample information are necessary to find the unique impulses of interest which represent the actual responses of the system to shocks that occur in practice. These considerations have led to what is known as *structural VAR (SVAR) models* and *structural impulse responses*.

SVAR models

Various types of restrictions have been considered for identifying the structural innovations or, equivalently, for finding a unique or at least locally unique *B* matrix. For example, using a triangular *B* matrix obtained from a Choleski decomposition of Σ_u is quite popular (for example, Sims, 1980; Christiano, Eichenbaum and Evans, 1996). Choosing a lower-triangular matrix amounts to setting up a recursive system with a so-called *Wold causal ordering* of the variables. One possible interpretation is that an impulse in the first variable can have an instantaneous impact on all other variables as well, whereas an impulse in the second variable can also have an instantaneous effect on the third to last variables but not on the first one, and so on. Because such a causal ordering is sometimes difficult to defend, other types of restrictions have also been proposed. Examples are:

1. 1. Instantaneous effects of some shocks on certain variables may be ruled out. In other words, zero restrictions are placed on *B* just as in the Choleski decomposition approach. The zero restrictions do not have to result in a triangular *B* matrix, however.
2. 2. Identification is achieved by imposing restrictions on the instantaneous relations of the variables. In this case a structural form model of the type $A_0 y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t$ may be considered and typically linear restrictions are imposed on A_0. Usually the elements on the main diagonal of A_0 will be normalized to unity. The restrictions on A_0 imply restrictions for $B = A_0^{-1}$. For example, if A_0 is triangular, then so is *B*.
3. 3. It is also possible to set up a model in the form $A_0 y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + B\varepsilon_t$ and impose restrictions on both A_0 and *B* to identify structural shocks. Combining restrictions on *B* with those on the instantaneous effects on the observed variables results in the so-called *AB*-model of Amisano and Giannini (1997).
4. 4. There may be prior information on the long-run effects of some shocks. In this case restrictions may be placed on $B + \Sigma_{i=1}^{\infty} \Theta_i = A(1)^{-1} B$ (for example, Blanchard and Quah, 1989). For instance, demand shocks may be assumed to have no accumulated long-run effects on some variable (in their case output). In fact, distinguishing between shocks with permanent and transitory effects is perhaps done more naturally in models which allow for integrated variables.

They will be discussed later.

5. 5. Sign restrictions may be imposed on the impulse responses (for example, Canova and De Nicoló, 2003; Uhlig, 2005), that is, one may want to require that certain shocks have positive or negative effects on certain variables. For example, a restrictive monetary shock should reduce the inflation rate.

Integrated and cointegrated variables

If the VAR operator has unit roots, that is, det $A(z) = 0$ for $z = 1$, then the variables have stochastic trends. Variables with such trends are called integrated. They can be made stationary by differencing. Moreover, they are called cointegrated if stationary linear combinations exist. If the VAR model contains integrated and cointegrated variables, impulse response analysis can still be performed as for stationary processes. For the latter processes the Φ_i's go to zero for $i \to \infty$ and, hence, the marginal response to an impulse to a stationary process is transitory, that is, the effect goes to zero as time goes by. In contrast, some impulses have permanent effects in cointegrated systems. In fact, in a *K*-dimensional system with $r < K$ cointegration relations, at least $K - r$ of the *K* shocks have permanent effects and at most *r* shocks have transitory effects (King et al, 1991; Lütkepohl, 2005, ch. 9). These facts open up the possibility to find identifying restrictions for the structural innovations by taking into account the cointegration properties of the system.

Estimation of impulse responses

Estimation of reduced form and structural form parameters of VAR processes is usually done by least squares, maximum likelihood or Bayesian methods. Estimates of the impulse responses are then obtained from the VAR parameter estimates. Suppose the VAR coefficients are contained in a vector α and denote its estimator by $\widehat{\boldsymbol{\alpha}}$. Any specific impulse response coefficient θ is a (nonlinear) function of α and may be estimated as $\hat{\theta} = \theta(\widehat{\boldsymbol{\alpha}})$. If $\widehat{\boldsymbol{\alpha}}$ is asymptotically normal, that is, $\sqrt{T}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\widehat{\boldsymbol{\alpha}}})$, then, under general conditions, $\hat{\theta}$ is also asymptotically normally distributed, $\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_{\hat{\theta}}^2)$. The variance of the asymptotic distribution is $\sigma_{\hat{\theta}}^2 = \frac{\partial \theta}{\partial \boldsymbol{\alpha}}' \Sigma_{\widehat{\boldsymbol{\alpha}}} \frac{\partial \theta}{\partial \boldsymbol{\alpha}}$. Here $\partial \theta / \partial \boldsymbol{\alpha}$ denotes the vector of first order partial derivatives of θ with respect to the elements of α (see Lütkepohl, 1990, for the precise expressions). This result can be used for setting up asymptotic confidence intervals for impulse responses in the usual way.

Asymptotic normality of $\hat{\theta}$ requires that $\sigma_{\hat{\theta}}^2$ is non-zero, which follows if $\Sigma_{\widehat{\boldsymbol{\alpha}}}$ is non-singular and $\partial \theta / \partial \boldsymbol{\alpha} \neq 0$. In general the covariance matrix $\Sigma_{\widehat{\boldsymbol{\alpha}}}$ will not be non-singular for cointegrated systems, for example. Moreover, the impulse responses generally consist of sums of products of the VAR coefficients and, therefore, the partial derivatives will also be sums of products of such coefficients. Consequently, the partial derivatives will also usually be zero in parts of the parameter space. Thus,

$\sigma_\theta^2 = 0$ may hold and, hence, $\hat{\theta}$ may actually converge at a faster rate than $\sqrt{T}$ in parts of the parameter space (cf. Benkwitz, Lütkepohl and Neumann, 2000).

Even under ideal conditions where the asymptotic theory holds, it may not provide a good guide for small sample inference. Therefore, bootstrap methods are often used to construct confidence intervals for impulse responses (for example, Kilian, 1998; Benkwitz, Lütkepohl and Wolters, 2001). If one uses these methods, deriving explicit forms of the analytical expressions for the asymptotic variances of the impulse response coefficients can be avoided. Unfortunately, bootstrap methods generally do not overcome the problems due to zero variances in the asymptotic distributions of the impulse responses. In fact, they may provide confidence intervals which do not have the desired coverage level even asymptotically (Benkwitz, Lütkepohl and Neumann, 2000).

Confidence bands for impulse response functions can also be constructed with Bayesian methods (for example, Koop, 1992). Prior information on the VAR parameters or the impulse responses can in that case be considered. It is not uncommon to report confidence intervals for individual impulse response coefficients and connecting them to get a confidence band around an impulse response function. This approach has been criticized by Sims and Zha (1999), who propose likelihood-characterizing error bands instead.

Extensions

There are a number of extensions to the models and impulse response functions considered so far. For example, all observed variables are treated as endogenous. A main criticism regarding problematic exogeneity assumptions in classical simultaneous equations models is thereby accounted for. On the other hand, this approach often results in heavily parameterized models and imprecise estimates. Therefore, it is occasionally desirable to classify some of the variables as exogenous or consider partial models where we condition on some of the variables which remain unmodelled. In this case one may be interested in tracing the effects of changes in the exogenous or unmodelled variables on the endogenous variables. The resulting impulse response functions are often referred to as dynamic multipliers in the literature on simultaneous equations (see Lütkepohl, 2005, for an introductory treatment). The inference problems related to these quantities are similar to those discussed earlier for VAR impulse responses.

It was also acknowledged in the related literature that finite order VAR models are at best good approximations to the actual data generation processes of multiple time series. Therefore, inference for impulse responses was also considered under the assumption that finite order VAR processes are fitted to data generated by infinite order processes (for example, Lütkepohl, 1988; Lütkepohl and Saikkonen, 1997).

Impulse responses associated with linear VAR models have the property of being time invariant and their shape is invariant to the size and direction of the impulses. These features make it easy to represent the reactions of the variables to impulses hitting the system in a small set of graphs. Such responses are often regarded as unrealistic in practice, where, for instance, a positive shock may have a different effect from a negative shock or the effect of a shock may depend on the state of the system at the time when it is hit. Hence, the linear VAR models are too restrictive for some analyses. These problems can be resolved by considering nonlinear models. Although nonlinear models have their attractive features for

describing economic systems or phenomena, their greater flexibility makes them more difficult to interpret properly. In fact, it is not obvious how to define impulse responses of nonlinear models in a meaningful manner. Gallant, Rossi and Tauchen (1993) proposed so-called *conditional moment profiles* which may give useful information on important features of nonlinear multiple time series models. For example, one may consider quantities of the general form $E[g(y_{t+h})|y_t + \xi, \Omega_{t-1}] - E[g(y_{t+h})|y_t, \Omega_{t-1}], \quad h = 1, 2, \ldots$, where $g(\cdot)$ denotes some function of interest, ξ represents the impulses hitting the system at time *t*, and $\Omega_{t-1} = (y_{t-1}, y_{t-2}, \ldots)$ denotes the history of the variables at time *t*. In other words, the conditional expectation of some quantity of interest, given the history of y_t in period *t*, is compared to the conditional expectation that is obtained if a shock ξ occurs at time *t*. For example, defining $g(y_{t+h}) = [y_{t+h} - E(y_{t+h}|\Omega_{t+h-1})][y_{t+h} - E(y_{t+h}|\Omega_{t+h-1})]'$ results in conditional volatility profiles, which may be compared to a baseline profile obtained for a specific history of the process and a zero impulse. Clearly, in general the conditional moment profiles depend on the history Ω_{t-1} as well as the impulse ξ. Similar quantities were also considered by Koop, Pesaran and Potter (1996), who called them *generalized impulse responses* (see also Pesaran and Shin, 1998).

Although these quantities may be interesting to look at, they depend on *t*, *h*, and ξ. Hence, there is a separate impulse response function for each given *t* and ξ. In empirical work it will therefore be necessary to summarize the wealth of information in the conditional moment profiles in a meaningful way – for instance, by considering summary statistics. In practice, an additional obstacle is that the actual data generation process is unknown and estimated models are available at best. In that case, the conditional moment profiles or generalized impulse responses will be estimates, and it would be useful to have measures for their sampling variability. It is not clear how this additional information may be computed and presented in the best way in practice.

See Also

- cointegration
- long run and short run
- measurement error models
- multiplier analysis
- structural vector autoregressions
- vector autoregressions

Bibliography

Amisano, G. and Giannini, C. 1997. *Topics in Structural VAR Econometrics*, 2nd edn. Berlin: Springer.

Benkwitz, A., Lütkepohl, H. and Neumann, M. 2000. Problems related to bootstrapping impulse responses of autoregressive processes, *Econometric Reviews* 19, 69–103.

Benkwitz, A., Lütkepohl, H. and Wolters, J. 2001. Comparison of bootstrap confidence intervals for impulse responses of German monetary systems. *Macroeconomic Dynamics* 5, 81–100.

Blanchard, O. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.

Canova, F. and De Nicoló, G. 2003. On the sources of business cycles in the G-7. *Journal of International Economics* 59, 77–100.

Christiano, L., Eichenbaum, M. and Evans, C. 1996. The effects of monetary policy shocks: evidence from the flow of funds. *Review of Economics and Statistics* 78, 16–34.

Gallant, A., Rossi, P. and Tauchen, G. 1993. Nonlinear dynamic structures. *Econometrica* 61, 871–907.

Kilian, L. 1998. Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics* 80, 218–30.

King, R., Plosser, C., Stock, J. and Watson, M. 1991. Stochastic trends and economic fluctuations. *American Economic Review* 81, 819–40.

Koop, G. 1992. Aggregate shocks and macroeconomic fluctuations: a Bayesian approach. *Journal of Applied Econometrics* 7, 395–411.

Koop, G., Pesaran, M. and Potter, S. 1996. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics* 74, 119–47.

Lütkepohl, H. 1988. Asymptotic distribution of the moving average coefficients of an estimated vector autoregressive process. *Econometric Theory* 4, 77–85.

Lütkepohl, H. 1990. Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics* 72, 116–25.

Lütkepohl, H. 2005. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Lütkepohl, H. and Saikkonen, P. 1997. Impulse response analysis in infinite order cointegrated vector autoregressive processes. *Journal of Econometrics* 81, 127–57.

Pesaran, M. and Shin, Y. 1998. Generalized impulse response analysis in linear multivariate models. *Economics Letters* 58, 17–29.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Sims, C. and Zha, T. 1999. Error bands for impulse responses. *Econometrica* 67, 1113–55.

Uhlig, H. 2005. What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics* 52, 381–419.

How to cite this article

Lütkepohl, Helmut. "impulse response function." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000283> doi:10.1057/9780230226203.0767

The New Palgrave Dictionary of Economics Online

imputation

Murray N. Rothbard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Aristotle; Austrian School; Böhm-Bawerk, E. von; factor prices; imputation; marginal productivity theory; Menger, C.; socialist calculation debate; subjective theory of value; Wieser, F. F. von; Mises, L. E. von; Hayek, F. A. von

Article

‘Imputation’ is a term introduced into economics as *Zurechnung* by the Austrian School economist Friedrich Freiherr von Wieser (Wieser, 1889). The term was a legal one, and the analogy was based on the legal method by which the jurist *imputes* guilt or liability to one or another criminal or tortfeasor. Imputation was a central concern of the Austrian School, since its analysis centred on the nature of the means–ends relationship (Mises, 1949) and on the process by which the subjective valuations and value-preferences of individual consumers ‘impute’ value to the goods being produced. As Carl Menger, founder of the Austrian School, pointed out, the valuations by consumers of their satisfactions, or ends, impute values to the consumer goods, the means, that are expected to satisfy those wants (Menger, 1871). And since producers’ goods are only means to the production and sale of consumer goods, the values of the factors of production will in turn be determined by and be equal to the expected values of the consumer goods to the consumers. In short, values are ‘imputed’ back to the prices of the factors of production; the rents of Champagne land are high because the consumers value the champagne highly, and not the other way round. ‘Costs’ of resources are reflections of the value of products forgone. While this process was clear in principle, there were considerable difficulties in working out the specifics. Essentially, Menger and his student Böhm-Bawerk stuck close to the realities of the market process, and focused on value imputation as a process of estimating how much of a product would be lost if the producer were deprived of one unit of a factor. Wieser, on the other hand, presumed that the marginal value of each factor could be found with great precision; in doing so, he assumed illegitimately that subjective values can be added and multiplied to arrive at the total value of a quantity of goods. But by its nature subjective value is an expression of ordinal preferences and therefore can neither be added

nor measured.

The modern theory of marginal productivity has essentially solved these problems and shown how values of products can be imputed back to productive factors. One exception is the current assumption that the existence of variable proportions solves the problem of pricing factors and leaves no theoretical room for arbitrary bargaining between factor owners. But the more important solution depends on whether factors are purely specific to one line of production or are relatively non-specific, that is, can be employed in the production of more than one good. If two factors are each purely specific to a given product, then, even if their proportions are variable, there is still no principle by which the market can determine their relative prices except by arbitrary bargaining (Mises, 1949, p. 336). In the real world, of course, the existence of such purely specific factors, and hence the scope for such bargaining, will be extremely limited.

The other important point is that values cannot be added or divided, and that the imputation process takes place, not automatically or precisely in an abstract realm of ‘values’, but only concretely and by trial and error, in the realistic market process of changing prices. In other words, although consumers can evaluate consumer goods and determine their prices directly by valuation, the prices of productive factors are only determined indirectly through market prices and entrepreneurial trial and error. There is no direct, abstract or pure process of imputing values.

This problem became strikingly relevant during the well-known debate over the Mises–Hayek demonstration that socialist governments cannot calculate economically. Joseph Schumpeter brusquely dismissed this contention with the statement that economic calculation under socialism follows ‘from the elementary proposition that consumers in evaluating (“demanding”) consumers’ goods *ipso facto* also evaluate the means of production which enter into the production of these goods’ (Schumpeter, 1942, p. 175). Hayek's perceptive reply points out that the ‘*ipso facto*’ assumes complete knowledge of values, demands, scarcities, and so on, to be ‘given’ to everyone, thereby ignoring the reality of the universal lack of complete knowledge, as well as the necessary function of the market economy, and the market price system, in conveying knowledge to all its participants (Hayek, 1945).

The analysis of imputation began in a neglected work of Aristotle, the *Topics*. Here, Aristotle analysed the ends–means relationship, and pointed out that the means, or ‘instruments of production’, necessarily derive their value from the ends, the final products useful to man, ‘the instruments of action’. The more desirable the final good, the more valuable will be the means to arrive at the product. Aristotle introduced the theme of marginality by stating that, if the addition of a good *A* to an already desirable good *C* yields a more desirable result than the addition of good *B*, then *A* will be more highly valued than *B*. Indeed, he also added a pre-Böhm-Bawerkian note by stressing the differential value of the loss rather than the addition of a good. Good *A* will be more valuable than *B* if the loss of *A* is considered to be worse than the loss of *B*. While critics have noted that Aristotle only slightly applied his analysis to the economic realm, his imputation theory was still an important contribution to the general theory of action of which economic theory is a highly developed part (Spengler, 1955).

See Also

- Austrian economics
- marginal productivity theory

Bibliography

Aristotle. *Topica*. Trans. W.A. Pickard-Cambridge and included. in *The Works of Aristotle*. vol. 1, ed. W. D. Ross, Oxford: Clarendon Press, 1928.

Hayek, F.A. 1926. Some remarks on the problem of imputation. In *Money, Capital, and Fluctuations: early Essays*, ed. F.A. Hayek. Chicago: University of Chicago Press, 1984.

Hayek, F.A. 1945. The use of knowledge in society. In *Individualism and Economic Order*, ed. F.A. Hayek. Chicago: University of Chicago Press, 1948.

Kauder, E. 1965. *A History of Marginal Utility Theory*. Princeton: Princeton University Press.

Menger, C. 1871. *Principles of Economics*. Glencoe, IL: Free Press, 1950.

Mises, L.von. 1949. *Human Action: A Treatise on Economics*. 3rd edn , Chicago: Regnery, 1966.

Schumpeter, J.A. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper.

Spengler, J. 1955. Aristotle on economic imputation and related matters. *Southern Economic Journal* 21 (April), 371–89.

Stigler, G. 1941. *Production and Distribution Theories: The Formative Period*. New York: Macmillan.

Wieser, F.von. 1889. *Natural Value*. New York: Kelley & Millman, 1956.

How to cite this article

Rothbard, Murray N. "imputation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000022> doi:10.1057/9780230226203.0768

The New Palgrave Dictionary of Economics Online

incentive compatibility

John O. Ledyard

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Incentive compatibility – a characteristic of mechanisms whereby each agent knows that his best strategy is to follow the rules, no matter what the other agents will do – is desirable because it promotes the achievement of group goals. But it is elusive because pervasive opportunities exist for misbehaviour, such as by misrepresenting preferences. This article reviews attempts to solve or at least to manage the incentive compatibility problem. Incentive compatibility provides a basic constraint on the possibilities for normative analysis, and so serves as the fundamental interface between what is desirable and what is possible in a theory of organizations.

Keywords

allocation mechanisms; auctions; Bayes' equilibrium; Borda, J.-C. de; capital budgeting; central planning; Cobb–Douglas functions; false preferences; free rider problem; games of incomplete information; incentive compatibility; Ledyard, J. O.; majority rule; market failure; mechanism design; monotonicity; Nash equilibrium; no-trade option; offer curves; Pareto efficiency; principal and agent; public enterprise management; public goods; regulation of monopoly; revelation principle; self-selection; social welfare functions; synthetic markets; transfer pricing; tâtonnement processes; von Neumann–Morgenstern utility function

Article

Allocation mechanisms, organizations, voting procedures, regulatory bodies, and many other institutions are designed to accomplish certain ends such as the Pareto-efficient allocation of resources or the equitable resolution of disputes. In many situations it is relatively easy to conceive of feasible processes; processes which will accomplish the goals if all participants follow the rules and are capable of handling the informational requirements. Examples of such mechanisms include marginal cost pricing, designed to attain efficiency, and equal division, designed to attain equity. Of course once a feasible mechanism is

found, the important question then becomes whether such a mechanism is also informationally feasible and compatible with ‘natural’ incentives of the participants. Incentive compatibility is the concept introduced by Hurwicz (1972, p. 320) to characterize those mechanisms for which participants in the process would not find it advantageous to violate the rules of the process.

The historical roots of the idea of incentive compatibility are many and deep. As was pointed out in one of a number of recent surveys,

> the concept of incentive compatibility may be traced to the ‘invisible hand’ of Adam Smith who claimed that in following individual self-interest the interests of society might be served. Related issues were a central concern in the ‘Socialist Controversy’ which arose over the viability of a decentralized socialist society. It was argued by some that such societies would have to rely on individuals to follow the rules of the system. Some believed this reliance was naive; others did not. (Groves and Ledyard, 1986, p. 1).

Further, the same issues have arisen in the design of voting procedures. Concepts and problems related to incentives were already identified and documented in the 18th century in discussions of proposals by Borda to provide alternatives to majority rule committee decisions. (See strategy-proof allocation mechanisms for further information on voting procedures.)

Incentive compatibility is both desirable and elusive. The desirability of incentive compatibility can be easily illustrated by considering public goods, goods such that one consumer's consumption of them does not detract from another consumer's simultaneous consumption of that good. The existence of these collective consumption commodities creates a classic situation of *market failure*; the inability of markets to arrive at a Pareto-optimal allocation. It was commonly believed, prior to Groves and Ledyard (1977), that in economies with public goods it would be impossible to devise a decentralized process that would allocate resources efficiently since agents would have an incentive to ‘free ride’ on others’ provision of those goods in order to reduce their own share of providing them. Of course Lindahl (1919) had proposed a feasible process which mimicked markets by creating a separate price for each individual's consumption of the public good. This designed process was, however, rejected as unrealistic by those who recognized that these ‘synthetic markets’ would be shallow (essentially monopsonistic) and therefore buyers would have no incentive to treat prices as fixed and invariant to their demands. The classic quotation is ‘… it is in the selfish interest of each person to give *false* signals, to pretend to have less interest in a given collective consumption activity than he really has…’ (Samuelson, 1954, pp. 388–9). Allocating public goods efficiently through Lindahl pricing would be feasible and successful if consumers followed the rules; but, it would not be successful since the mechanism is not incentive compatible. If buyers do not follow the rules, efficient resource allocation will not be achieved and the goals of the design will be subverted because of the motivations of the participants. Any institution or rule, designed to accomplish group goals, must be incentive compatible if it is to perform as desired.

The elusiveness of incentive compatibility can be most easily illustrated by considering a situation with only private goods. Economists generally model behaviour in private goods markets by assuming that buyers and sellers ‘follow the rules’ and take prices as given. It is now known, however, that as long as the number of agents is finite then any one of them can still gain by misbehaving and, furthermore, can do so in a way which can not be detected by anyone else. The explanation is provided in two steps. First,

if there are a finite number of traders, and none have a perfectly elastic offer curve (which will be true if preferences are non-linear) then one trader can gain by being able to control prices. For example, a buyer would want to set price where his marginal benefit equalled his marginal outlay and thereby gain monopsonistic benefits. Of course, if the others know that buyer's demand curve (either directly or through inferences based on revealed preference) then they would know that the buyer was not ‘taking prices as given’ and could respond with a suitable punishment against him. This brings us to our second step. Even though others can monitor and prohibit price setting behaviour, our benefit-seeking monopsonist has another strategy which can circumvent this supervision. He calculates a (false) demand curve which, when added to the others’ offer curves, produces an equilibrium price equal to that which he would have set if he had direct control. He then calculates a set of preferences which yields that demand curve and participates in the process *as if he had these (false) preferences*. Usually this involves simply acting as if one has a slightly lower demand curve than one really does. Since preferences are not able to be observed by others, he can follow this behaviour which looks like it is price-taking, and therefore ‘legal’, and can do individually better. The unfortunate implication of such concealed misbehaviour is that the mechanism performs other than as intended. In this case, resources are artificially limited and too little is traded to attain efficiency.

In 1972 Hurwicz established the validity of the above intuition. His theorem can be precisely stated after the introduction of some notation and a framework for further discussion.

The impossibility theorem

The key concepts include economic environments, allocation mechanisms, incentive compatibility, the no-trade option, and Pareto-efficiency. We take up each in turn.

An *economic environment*, those features of an economy which are to be taken as given throughout the analysis, includes a description of the agents, the feasible allocations they have available and their preferences for those allocations. While many variations are possible, I concentrate here on a simple model. Agents (consumers, producers, politicians, etc.) are indexed by $i = 1, \ldots, n$. X is the set of feasible allocations where $x = (x^1, \ldots, x^n)$ is a typical element of X. (An exchange environment is one in which X is the set of all $x = (x^1, \ldots, x^n)$ such that $x^i \geq 0$ and $\Sigma x^i = \Sigma w^i$, where w^i is i's initial endowment of commodities.) Each agent has a selfish utility function $u^i(x^i)$. The environment is $e = [I, X, u^1, \ldots, u^n]$. A crucial fact is that initially *information is dispersed* since i, and only i, knows u^i. We identify the specific knowledge i initially has as i's *characteristic*, e^i. In our model, $e^i = u^i$.

Although there are many variations in models of allocation mechanisms, I begin with the one introduced by Hurwicz (1960). An *allocation mechanism* requests information from the agents and then computes a feasible allocation. It requests information in the form of messages m^i from agent i through a *response* function $f^i(m^1, \ldots, m^n)$. Agent i is told to report $f^i(m, e^i)$ if others have reported m and i's characteristic is e^i. An equilibrium of these response rules, for the environment e, is a joint message m such that $m^i = f^i(m, e^i)$ for all i. Let $\mu(e, f)$ be the set of equilibrium messages for the response functions f in the environment e. The allocation mechanism computes a feasible allocation x by using an *outcome* function $g(m)$ on equilibrium messages. The net result of all of this in the environment e is the

allocation $g[\mu(e, f)] = x$ *if all i follow the rules, f*. Thus, for example, the *competitive mechanism* requests agents to send their demands as a function of prices which are in turn computed on the basis of the aggregate demands reported by the consumers. In equilibrium, each agent is simply allocated their stated demand. (An alternative mechanism, yielding exactly the same allocation in one iteration, would request the demand function and then compute the equilibrium price and allocation for the reported demand functions.) It is well known, for exchange economies with only private goods, that if agents report their true demands then the allocations computed by the competitive mechanism will be Pareto-optimal.

It is obviously important to be able to identify those mechanisms, those rules of communication, that have the property that they are self-enforcing. We do that by focusing on a class of mechanisms in which each agent gains nothing, and perhaps even loses, by misbehaving. While a multitude of misbehaviours could be considered it is sufficient for our purposes to consider a slightly restricted range. In particular we can concentrate on undetectable behaviour, behaviour which no outside agent can distinguish from that prescribed by the mechanism. We model this limitation on behaviour by requiring the agent to restrict his misrepresentations to those which are consistent with some characteristic he might have. An allocation mechanism is said to be *incentive compatible* for all environments in the class E if there is no agent i and no environment e in E and no characteristic e^{*i} such that (e/e^{*i}) is in E (where (e/e^{*i}) is the environment derived from e by replacing e^i with e^{*i}) and such that

$$u^i\{g[\mu(e, f)], e^i\} < u^i\{g[\mu(e/e^{*i}, f)], e^i\}$$

where $u^i(x^*, e^i)$ is i's utility function in the environment e. That is, no agent can manipulate the mechanism by pretending to have a characteristic different from the true one and do better than acting according to the truth. The agent has an incentive to follow the rules and the rules are compatible with his motivations.

Incentive compatibility is at the foundation of the modern *theory of implementation*. In that theory, one tries to identify conditions under which a particular social choice rule or performance standard, $P: E \to X$, can be recreated by an allocation mechanism under the hypothesis that individuals will follow their self-interest when they participate in the implementation process. In our language, the rule P is implementable if and only if there is an incentive compatible mechanism (f, g) such that $g[\mu(e, f)] = P(e)$ for all e in E. The theory of implementation seeks to answer the question 'which P are implementable?' We will see some of the answers below for P which select from the set of Pareto-efficient allocations. Those interested in more general goals and performance standards should consult Dasgupta, Hammond and Maskin (1979) or Postlewaite and Schmeidler (1986).

An allocation mechanism is said to have the *no trade-option* if there is an allocation θ at which each participant may remain. In exchange environments the initial endowment is usually such an allocation. Mechanisms with a no-trade option are non-coercive in a limited sense. If an allocation mechanism possesses the no-trade option then the allocation it computes for an environment e, if agents follow the rules, must leave everyone at least as well off, using the utility functions for e, as they are at θ. That is,

for all i and all e in E

$$u^i\{g[\mu(e, f)], e^i\} > u^i(\theta, e^i).$$

An allocation mechanism is said to be *Pareto-efficient in E* if the allocations selected by the mechanism, when agents follow the rules, are Pareto-optimal in e. That is, for each e in E, there is no allocation x^* in X such that, for all i,

$$u^i(x^*, e^i) \geq u^i\{g[\mu(e, f)], e^i\}$$

with strict inequality for some i.

With this language and notation, Hurwicz's theorem on the elusive nature of incentive compatibility in private markets, subsequently expanded by Ledyard and Roberts (1974) to include public goods environments, can now be easily stated. *Theorem:* In classical (public or private) economic environments with a finite number of agents, there is no incentive compatible allocation mechanism which possesses the no-trade option and is Pareto-efficient. (Classical environments include pure exchange environments with Cobb–Douglas utility functions.)

A more general version of this theorem, in the context of social choice theory, has been proven by Gibbard (1973) and Satterthwaite (1975) with the concept of a ‘non-dictatorial social choice function’ replacing that of a ‘mechanism with the no-trade option’. (See strategy-proof allocation mechanisms.)

There are a variety of possible reactions to this theorem. One is simply to give up the search for solutions to market failure since the theorem seems to imply that one should not waste any effort trying to create institutions to allocate resources efficiently. A second is to notice that, at least in private markets, if there are a very large number of individuals in each market then efficiency is ‘almost’ attainable (see Roberts and Postlewaite, 1976). A third is to recognize that the behaviour of individuals will generally be different from that implicitly assumed in the definition of incentive compatibility. A fourth is to accept the inevitable, lower one's sights, and look for the ‘most efficient’ mechanism among those which are incentive compatible and satisfy a voluntary participation constraint. We consider the last two options in more detail.

Other behaviour: Nash equilibrium

If a mechanism is incentive compatible, then each agent knows that his best strategy is to follow the rules according to his true characteristic, *no matter what the other agents will do*. Such a strategic structure is referred to as a dominant strategy game and has the property that no agent need know or

predict anything about the others' behaviour. In mechanisms which are not incentive compatible, each agent must predict what others are going to do in order to decide what is best. In this situation agents' behaviour will not be as assumed in the definition of incentive compatibility. What it will be continues to be an active research topic and many models have been proposed. Since most of these are covered in Groves and Ledyard (1986), I will concentrate on the two which seem most sensible. Both rely on game-theoretic analyses of the strategic possibilities. The first concentrates on the outcome rule, g, and postulates that agents will not choose messages to follow the specifications of the response functions but to do the best they can against the messages sent by others. Implicitly this assumes that there is some type of iterative process (embodied in the response rules) which allows revision of one's message in light of the responses of others. We can formalize this presumed strategic behaviour in a new concept of incentive compatibility. An allocation mechanism $(f,\bullet g)$ is called *Nash incentive compatible* for all environments in E if there is no environment e, no agent i, and no message m^{*i} which i can send such that

$$u^i\left(g\left[\mu(e, f) / m^{*i}, e^i\right]\right) > u^i\left(g\left[\mu(e, f), e^i\right]\right)$$

where $\mu(e, f)$ is the 'equilibrium' message of the response rules f in the environment e, $g(m)$ is the outcome rule, and $[m/m^{*i}]$ is the vector m where m^{*i} replaces m^i. In effect this requires the equilibrium messages of the response rules to be Nash equilibria in the game in which messages are strategies and payoffs are given by $u[g(m)]$. It was shown in a sequence of papers written in the late 1970s, including those by Groves and Ledyard (1977), Hurwicz (1979), Schmeidler (1980), and Walker (1981), that Nash incentive compatibility is not elusive. The effective output of that work was to establish the following.
Theorem: In classical (public or private) economic environments with a finite number of agents, there are many Nash incentive compatible mechanisms which possess the no-trade option and are Pareto-efficient.
With a change in the predicted behaviour of the participants in the mechanism, in recognition of the fact that in the absence of dominant strategies agents must follow some other self-interested strategies, the pessimism of the Hurwicz theorem is replaced by the optimistic prediction of a plethora of possibilities. (See Dasgupta, Hammond and Maskin, 1979, Postlewaite and Schmeidler (1986) and Groves and Ledyard (1986) for comprehensive surveys of these results including many for more general social choice environments.) Although it remains an unsettled empirical question whether participants will indeed behave this way, there is a growing body of experimental evidence that seems to me to support the behavioural hypotheses underpinning Nash incentive compatibility, especially in iterative tâtonnement processes.

Other behaviour: Bayes' equilibrium

The second approach to modelling strategic behaviour of agents in mechanisms, when dominant strategies are not available, is based on Bayesian decision theory. These models, called *games of incomplete information* (see Myerson, 1985), concentrate on the beliefs of the players about the situation

in which they find themselves. In the simplest form, it is postulated that there is a common knowledge (everyone knows that everyone knows that…) probability function, π (e), which describes everyone's prior beliefs. Each agent is then assumed to choose that message which is best against the expected behaviour of the other agents. The expected behaviour of the other agents is also constrained to be ‘rational’ in the sense that it should be best against the behaviour of others. This presumed strategic behaviour is embodied in a third type of incentive compatibility. (It could be argued that the concept of incentive compatibility remains the same, based on non-cooperative behaviour in the game induced by the mechanism, while only the presumed information structure and sequence of moves required to implement the allocation mechanism are changed. Such a view is not inconsistent with that which follows.) An allocation mechanism (f,•g) is called *Bayes incentive compatible* for all environments in E given π on E if there is no environment e^*, no agent i, and no message m^{*i} which i can send such that

$$\int u^i\left\{g[\mu(e, f) / m^{*i}], e^{*i}\right\} d\pi(e|e^{*i}) > \int u^i\left\{g[\mu(e, f), e^{*i}] d\pi(e, |e^{*i})\right\}$$

where, as before, μ is the equilibrium message vector and g is the outcome rule. Further, $\pi(e/e^{*i})$ *is the conditional probability measure on e* given e^{*i}, and u^i is a von Neumann–Morgenstern utility function. In effect, this requires the equilibrium messages of the response rules to be Bayes equilibrium outcomes of the incomplete information game with messages as strategies, payoffs $u[g(m)]$ and common knowledge prior π.

There are two types of results which deal with the possibilities for Bayes incentive compatible design of allocation mechanisms, neither of which is particularly encouraging. The first type deals with the possibilities for incentive compatible design which is independent of the beliefs. The typical theorem is illustrated by the following result proven by Ledyard (1978). *Theorem*: In classical economic environments with a finite number of agents, there is no Bayes incentive compatible mechanism which possesses the no-trade option and is Pareto-efficient *for all* π *on E*. Understanding this result is easy when one realizes that any mechanism (f, g) is Bayes incentive compatible for all π for all e in E if and only if it is (Hurwicz) incentive compatible for all e in E. Thus the Hurwicz impossibility theorem again applies.

The second type of result is directed towards the possibilities for a specific prior π; that is, towards what can be done if the mechanism can depend on the common knowledge beliefs. The most general characterizations of the possibilities for Bayes incentive compatible design can be found in Palfrey and Srivastava (1987) and Postlewaite and Schmeidler (1986). They have shown that two conditions, called monotonicity and self-selection, are necessary and sufficient for a social choice correspondence to be implementable in the sense that there is a Bayes incentive compatible mechanism that reproduces that correspondence. The details of these conditions are not important. What is important is that many correspondences do not satisfy them. In particular, there appear to be many priors π and many sets of environments E for which there is no mechanism which is Bayes incentive compatible, provides a no-trade option and is Pareto-efficient. Thus, impossibility still usually occurs even if one allows the mechanism to depend on the prior.

One recent avenue of research which promises some optimistic counterweight to these negative results can be found in Palfrey and Srivastava (1987). In much the same way that the natural move from Hurwicz incentive compatibility to Nash incentive compatibility created opportunities for incentive compatible design, these authors have shown that a move back towards dominant strategies may also open up possibilities. Refinements arise by varying the equilibrium concept in a way that reduces the number of (Bayes or Nash) equilibria for a given e or π. Moore and Repullo use subgame perfect Nash equilibria. Palfrey and Srivastava eliminate weakly dominated strategies from the set of Nash equilibria. They have discovered that, in pure exchange environments, virtually all performance correspondences are implementable if behaviour satisfies these refinements. In particular, any selection from the Pareto-correspondence is implementable for these refinements, and so there are many refined-Nash incentive compatible mechanisms which are Pareto-efficient and allow a no-trade option. It is believed that these results will transfer naturally to refinements of Bayes equilibria, but the research remains to be done.

Incentive compatibility as a constraint

Another of the reactions to the Hurwicz impossibility result is to accept the inevitable, to view incentive compatibility as a constraint, and to design mechanisms to attain the best level of efficiency one can. If full efficiency is possible, it will occur as the solution. If not, then one will at least find the second-best allocation mechanism. Examples of this rapidly expanding research literature include work on optimal auctions (Harris and Raviv, 1981; Matthews, 1983; Myerson, 1981), the design of optimal contracts for the principle-agent problem, and the theory of optimal regulation (Baron and Myerson, 1982). As originally posed by Hurwicz (1972, pp. 299–301), the idea is to adopt a social welfare function $W(x,\bullet e)$, a measure of the social welfare attained from the allocation x if the environment is e and then to choose the mechanism (f, g) to maximize the (expected) value of W subject to the 'incentive compatibility constraints', the constraint that the rules (f, g) be consistent with the motivations of the participants. One chooses (f, g) to

$$\text{maximize} \int W\{g[\mu(e, f)], e\} d\pi(e)$$

subject to, for every i, every e, and every e^{*i},

$$\int u^i\left\{g[\mu(e / e^{*i}, f)], e^i\right\} d\pi(e|e^i) \leq \int u^i(g[\mu(e, f), e^i]) d\pi(e|e^i).$$

As formalized here the incentive compatibility constraints embody the concept of Bayes incentive compatibility. Of course, other behavioural models could be substitued as appropriate.

Sometimes a voluntary participation constraint, related to the no-trade option of Hurwicz, is added to the optimal design problem. One form of this constraint requires that (*f*, *g*) also satisfy, for every *i* and every *e*,

$$\int u^i\left\{g[\mu(e)],\ e^i\right\}d\pi(e|e^i) \geq \int u^i(\theta[e],\ e^i)d\pi(e|e^i).$$

In practice this optimization can be a difficult problem since there are a large number of possible mechanisms (*f*, *g*). However, an insight due to Gibbard (1973) can be employed to reduce the range of alternatives and simplify the analysis. Now called the *revelation principle*, the observation he made was that, to find the maximum, it is sufficient to consider only mechanisms, called direct revelation mechanisms, in which agents are asked to report their own characteristics. The reason is easy to see. Suppose that (f^*,•g^*) solves the maximum problem. Let (F^*,•G^*) be a new (direct revelation) mechanism defined by $F^{*i}(m,\ e^i) = e^i$ and $G^*(m) = g[\mu(m,\ f)]$. Each *i* is told to report his characteristic and then G^* computes the allocation by computing that which would have been chosen if the original mechanism (*f*,•G^*) had been used honestly in the reported environment. (F^*,•G^*) yields the same allocation as (f^*,•g^*), *if each agent reports the truth*. But the incentive compatibility constraints, which (f^*,•g^*) satisfied, ensure that each agent will want to report truthfully. Thus, whatever can be done, by any arbitrary mechanism subject to the Bayes incentive compatibility constraints, can be done with direct revelation mechanisms subject to the constraint that each agent wants to report their true characteristic. One need only choose a function $G: E \rightarrow X$ to

$$\text{maximize} \int W\{G(e),\ e\}d\pi(e)$$

subject to, for every *i*, *e* and e^i,

$$\int u^i\left\{G(e\,/\,e^{*i}),\ e^i\right\}d\pi(e|e^i) \leq \int u^i[G(e),\ e^i]d\pi(e|e^i),$$

and

$$\int u^i[G(e), e^i]d\pi(e|e) \geq \int u^i(\theta[e], e^i)d\pi(e|e^i).$$

There are at least two problems with this approach to organizational design. The first is that the choice of mechanism depends crucially on the prior beliefs, π. This is a direct result of the use of Bayes incentive compatibility in the constraints. Since the debate is still open let me simply summarize some of the arguments. One is that if the mechanism chosen for a given situation does not depend on common knowledge beliefs then we would not be using all the information at our disposal to pursue the desired goals and would do less than is possible. Further, since the beliefs are common knowledge we can all agree as to their validity (misrepresentation is not an issue) and therefore to their legitimate inclusion in the calculations. An argument is made against this on the practical grounds that one need only consider actual situations, such as the introduction of new technology by a regulated utility or the acquisition of a major new weapons system by the government, to understand the difficulties involved in arriving at agreements about the particulars of common knowledge. Another argument against is based on the feeling that mechanisms should be robust. A ‘good’ mechanism should be able to be described in terms of its mechanics and, while it probably should have the capacity to incorporate the common knowledge relevant to the current situation, it should be capable of being used in many situations. How to capture these criteria in the constraints or the objective function of the designer remains an open research question.

The second problem with the optimal auction approach to organizational design is the reliance on the revelation principle. Restricting attention to direct revelation mechanisms, in which an agent reports his entire characteristic, is an efficient way to prove theorems, but it provides little guidance for those interested in actual organization design. For example it completely ignores the informational requirements of the process and any limitations, if any, in the information processing capabilities of the agents or the mechanism. Writing down one's preferences for all possible consumption patterns is probably harder than writing down one's entire demand surface which is certainly harder than simply reacting to a single price vector and reporting only the quantities demanded at that price. A failure to recognize the information processing constraints in the optimization problem is undoubtedly one of the reasons there has been limited success in using the theory of optimal auctions to explain the existence of pervasive institutions, such as the first-price sealed-bid auction used in competitive contracting or the posted price institution used in retailing.

Summary

Incentive compatibility captures the fundamental positivist notion of self-interested behaviour that underlies almost all economic theory and application. It has proven to be an organizing principle of great scope and power. Combined with the modern theory of mechanism design, it provides a framework in which to analyse such diverse topics as auctions, central planning, regulation of monopoly, transfer pricing, capital budgeting, and public enterprise management. Incentive compatibility provides a basic constraint on the possibilities for normative analysis. As such it serves as the fundamental interface between what is desirable and what is possible in a theory of organizations.

See Also

- efficient allocation
- externalities
- Lindahl equilibrium
- public goods

Bibliography

Baron, D. and Myerson, R. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50, 911–30.

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules: some general results on incentive compatibility. *Review of Economic Studies* 46, 185–216.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–602.

Groves, T. and Ledyard, J. 1977. Optimal allocation of public goods: a solution to the ‘free rider’ problem. *Econometrica* 45, 783–809.

Groves, T. and Ledyard, J. 1986. Incentive compatibility ten years later. In *Information, Incentives, and Economic Mechanisms*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Harris, M. and Raviv, A. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49, 1477–99.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical Methods in the Social Sciences*, ed. K. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press, 27–46.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and Organization: A Volume in Honor of Jacob Marschak*, ed. R. Radner and C. B. McGuire. Amsterdam: North-Holland, 297–336.

Hurwicz, L. 1979. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *Review of Economic Studies* 46, 217–25.

Ledyard, J. 1978. Incomplete information and incentive compatibility. *Journal of Economic Theory* 18, 171–89.

Ledyard, J. and Roberts, J. 1974. On the incentive problem with public goods. Discussion Paper No. 116, Center for Mathematical Studies in Economics and Management Science, Northwestern University.

Lindahl, E. 1919. *Die Gerechtigkeit der Besteuerung*. Lund. Partial translation in*Classics in the Theory of Public finance*, ed. R. A. Musgrave and A. T. Peacock. London: Macmillan, 1958.

Matthews, S. 1983. Selling to risk averse buyers with unobservable tastes. *Journal of Economic Theory* 30, 370–400.

Moore, J. and Repullo, R. 1986. Subgame perfect implementation. Working paper, London School of Economics.

Myerson, R.B. 1981. Optimal auction design. *Mathematics of Operations Research* 6, 58–73.

Myerson, R.B. 1985. Bayesian equilibrium and incentive compatibility: an introduction. In *Social Goals and Social Organization: Essays in Memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler and H. Sonnenschein. Cambridge: Cambridge University Press.

Palfrey, T. and Srivastava, S. 1986. Implementation in exchange economies using refinements of Nash equilibrium. Graduate School of Industrial Administration, Carnegie-Mellon University, July 1986.

Palfrey, T. and Srivastava, S. 1987. On Bayesian implementable allocations. *Review of Economic Studies* 54, 193–208.

Postlewaite, A. and Schmeidler, D. 1986. Implementation in differential information economics. *Journal of Economic Theory* 39, 14–33.

Roberts, J. and Postlewaite, A. 1976. The incentives for price-taking behavior in large economies. *Econometrica* 44, 115–28.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36, 387–9.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Schmeidler, D. 1980. Walrasian analysis via strategic outcome functions. *Econometrica* 48, 1585–93.

Walker, M. 1981. A simple incentive compatible scheme for attaining Lindahl allocations. *Econometrica* 49, 65–71.

How to cite this article

Ledyard, John O. "incentive compatibility." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000027> doi:10.1057/9780230226203.0769

The New Palgrave Dictionary of Economics Online

income mobility

Gary S. Fields

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Income mobility means different things to different people. This article explains the six different mobility concepts used in the literature, reviews the various indices used in the mobility literature to measure these concepts, summarizes the difference the use of different mobility concepts and measures makes in practice, presents the axiomatic approach to income mobility, and discusses a number of other issues that arise in the mobility literature.

Keywords

absolute mobility; exchange mobility; income mobility; inequality (measurement); intergenerational income mobility; movement studies; relative mobility; structural mobility; time-independence studies; translation invariance

Article

What is income mobility? Extensive surveys of the income and earnings mobility literatures may be found in Atkinson, Bourguignon, and Morrisson (1992), Maasoumi (1998), Solon (1999), and Fields and Ok (1999a). (‘Income’ refers to income from all sources while ‘earnings’ refers to income earned in the labour market.) Mobility analysts agree on one defining feature: ‘income mobility’ is about how much income each recipient receives at two or more points in time. In this way, income mobility studies are distinguished from studies of the inequality and poverty aspects of income distribution, both of which are based (typically) on anonymous cross sections or (less frequently) marginal distributions of the joint distributions.

The following notation is used throughout this article. Let $x = (x^1, \ldots, x^n)$ denote a vector of ‘incomes’ in an initial year. This vector is ‘personalized’ in the sense that the same recipient units are followed over time. It is conventional to array the recipients in the base year from lowest income to highest. Whether this convention is followed or not, it is essential to keep the same order for subsequent years (or

generations). Denote the ordered vector in a subsequent year by $y = (y^1, \ldots, y^n)$. The micro-mobility data, also termed in the literature the pattern of ‘distributional change’, is summarized by the transformation $x \to y$ in the two-period case or more generally the transformation $x \to y \to z \to \cdots$ in the T-period case. The extent of mobility associated with the transformation $x \to y$ will be denoted by $m(x, y)$.

Beyond agreeing that income mobility studies are about transformations of the type $x \to y$ or $x \to y \to z \to \cdots$, the literature is marked by considerable disagreement. This is because the term ‘income mobility’ connotes precise but *different* ideas to different researchers. It is for this reason that mobility analysts often have trouble communicating with each other, with other social scientists, or with the general public. Furthermore, these differences in notions of what income mobility is remain even after agreement is reached on a number of other aspects of the mobility under consideration. These other aspects, discussed in the following paragraphs, are whether the context is intergenerational or intragenerational, what the indicator of social or economic status is, and whether the analysis is at the macro-mobility or micro-mobility level.

One issue is whether the aspect of mobility of interest is intergenerational or intragenerational. In the *intergenerational* context, the recipient unit is the family, specifically a parent and a child. In the *intragenerational* context, the recipient unit is the individual or family at two different dates. The issues discussed in this article apply equally to both.

Second, agreement must be reached on an indicator of social or economic status and the choice of recipient unit. For brevity, I shall talk about mobility of ‘income’ among ‘individuals’.

Third, the mobility questions asked and our knowledge about mobility phenomena may be grouped into two categories, macro and micro. *Macro*-mobility studies start with the question, ‘How much economic mobility is there?’ Answers are of the type ‘*a* per cent of the people stay in the same income quintile’, ‘*b* per cent of the people moved up at least $1,000 while *c* per cent of the people moved down at least $1,000’, ‘the mean absolute value of income change was $*d*,’ and ‘in a panel of length T, the mean number of years in poverty is t^*.’ The macro-mobility studies often go beyond this question to ask, ‘Is economic mobility higher here than there and what accounts for the difference?’ Answers would be of the type, ‘economic mobility has been rising over time’, ‘*A* has more upward mobility than *B* because economic growth was higher in *A* than in *B*’, and ‘incomes are more stable in *C* than in *D* because *C* has a better social safety net’. *Micro*-mobility studies, on the other hand, start with the question, ‘What are the correlates and determinants of the income or positional changes of individual income recipients?’ The answers to these questions would be of the type, ‘unconditionally, income changes are higher for the better-educated’ and ‘other things equal, higher initial income is associated with lower subsequent income growth’.

These three issues – intergenerational versus intragenerational, changes in the distribution of what among whom, and macro-mobility versus micro-mobility – help determine which kind of mobility analysis is being undertaken. Yet major differences remain. It is to these that we now turn.

Mobility concepts and measures

At least 20 mobility measures have been used in the literature. Many empirical mobility studies divide base- and final-year incomes into quantiles (for example, quintiles or deciles) and calculate immobility

ratios, mean upward movements, and the like (Fields, 2001). Other studies estimate correlation coefficients between base-year and final-year incomes (Atkinson, Bourguignon and Morrisson, 1992). In the intergenerational mobility literature, it is common to calculate intergenerational elasticities, that is, the coefficient obtained when the logarithm of the child's income is regressed on the logarithm of the parent's (Solon, 1999).

In each case, we may ask, what are the various measures measuring? The essential answer is this: *different indices measure different underlying entities*. Whenever one of these underlying entities is measured, other information contained in the joint distribution of initial and final incomes is lost.

What are the different underlying entities that the various income mobility measures measure? The first distinction to be drawn is between measures of time independence and measures of movement. The question asked by *time-independence* studies is, how dependent is current income on past income? One commonly used measure of time independence is the beta coefficient commonly calculated in the intergenerational mobility literature by regressing the log-income of the child on the log-income of the parent.

Movement studies ask a different question, namely: in comparisons of incomes of the same individuals between one year and another, or of parents and children between one generation and another, how much income movement has taken place? The various movement indices in the literature may usefully be classified into five categories or concepts (‘concepts’ because they are different underlying entities, not alternative measures of the same underlying entity).

Positional movement (or ‘quantile movement’) is about the movement of individuals among various positions (quintiles, deciles, centiles, or ranks) in the income distribution. An individual experiences positional movement if and only if he or she changes quintiles, deciles, centiles, or ranks. Positional movement in a population is greater the more such positional changes there are and/or the larger these positional changes are. King (1983) derived a broad class of positional movement indices axiomatically, one member of which is

$$M_K(x, y) = 1 - \exp\left[-\frac{\gamma}{n}\sum_{i=1}^{n}\frac{|z_i - y_i|}{\mu(y)}\right],$$

where γ is the observer's degree of immobility aversion, z_i is the income level agent i would have obtained if his or her rank order did not change during the process $x \to y$, and $\mu(y)$ is the mean income in distribution y.

Like positional movement, *share movement* is relative but it is relative in a different way. Share movement takes place if and only if an individual's income rises or falls relative to the mean. Thus, an individual can experience upward or downward share movement even if his or her income in dollars is unchanged and/or if he or she does not change position within the income distribution. Share movement in the population reflects the frequency and magnitude of these individual share changes. One attractive index of share movement in a population is the mean absolute value of share changes

$$M_S(x, y) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i}{\mu_y} - \frac{x_i}{\mu_x}\right|,$$

where $\mu(x)$ and $\mu(y)$ are the means of distributions x and y respectively.
Another concept is *non-directional income movement* (also called '*flux*'), which gauges the extent of fluctuation in individuals' incomes. To illustrate, suppose that in a two-person economy one person's income goes up by $10,000 while another's goes down by $10,000. Those who see an average income change of $10,000 are non-directional income movement adherents. Two indices of non-directional income movement have been suggested by Fields and Ok (1996; 1999b):

$$M_{F-O_1}(x, y) = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

and

$$M_{F-O_2}(x, y) = \frac{1}{n}\sum_{i=1}^{n}|\log y_i - \log x_i|.$$

Suppose, however, that, when one person's income goes up by $10,000 and another's goes down by $10,000, the observer cares not only about the amounts of the income changes but also about their direction. *Directional income movement* may be judged using a linear or a concave valuation function. One valuation function which embodies concavity is the mean change in log-incomes (Fields and Ok, 1999):

$$M_{F-O_3}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log x_i).$$

As a fifth and final notion of income movement, consider how the income changes experienced by individuals cause the inequality of longer-term incomes to differ from the inequality of base-year incomes. *Mobility as an equalizer of longer-term incomes* would judge that a pattern of income change $(1, 3) \rightarrow (1, 5)$ would *disequalize* longer-term income relative to the base, while a pattern of income

change $(1, 3) \rightarrow (5, 1)$ would *equalize* longer-term income relative to the base. This concept is well-established in the literature (Schumpeter, 1955; Shorrocks, 1978b; Atkinson, Bourguignon, and Morrisson, 1992; Slemrod, 1992; Krugman, 1992; Jarvis and Jenkins, 1998), but only recently has a class of measures of this concept been proposed (Fields, 2005). One family within this class is

$$\xi \equiv 1 - (I(a) / I(x)),$$

where *x* is the vector of base-year incomes, *y* is the vector of final-year incomes, *a* is the vector of average incomes, the i'th element of which is $a^i \equiv \frac{x^i + y^i}{2}$, and *I*(.) is a cross-sectional inequality measure such as the Gini coefficient or the Theil index.

We thus have six mobility concepts and a large number of measures. Because these concepts are fundamentally different from one another, it is important for analysts to choose the concepts that are of greatest interest to them and then measure those concepts. Let us now turn to a brief empirical review of studies that have used two or more of these concepts.

Different mobility concepts in practice

The previous section distinguished between time independence, positional movement, share movement, non-directional income movement, directional income movement, and mobility as an equalizer of longer-term incomes. How do these six concepts and the measures of them compare in empirical work? Specifically, which country has more mobility than another? Has mobility been rising or falling over time within a country? Are some groups in the population more or less mobile than others?

The answers to these questions have been shown empirically to depend on which mobility concept is used. In comparing *OECD countries*, some countries were found to be more mobile than others with the use of measures of some concepts and less mobile than others with the use of measures of other concepts (OECD, 1996; 1997). When we looked over time, in the *United States* measures of four concepts (time independence, positional movement, share movement, and income flux) all peaked in 1980–5 but measures of two other concepts did not: directional income movement exhibits a saw-tooth pattern, while mobility as an equalizer of longer-term incomes exhibits a peak followed by a valley (Fields, Leary and Ok, 2002; Fields, 2005). In *France*, mobility differences among demographic groups have been explored (Buchinsky et al., 2004). The answers to the questions ‘Who has more mobility: women or men? Better-educated or less-educated workers?’ were shown to differ depending on which mobility concept was used. By gender, women in France have *more* time independence and positional movement than men, *less* share movement than men, *about the same* non-directional and directional movement in logs, and *about the same* amount of mobility as an equalizer of longer-term incomes. By education, those with the highest educational attainments have *less* time independence and positional movement, and if anything *more* share movement, flux, and directional income movement in logs. In *Argentina*, too, measures of the six different concepts produced qualitatively different results (Sánchez Puerta, 2005).

Looking at changes over time, some mobility indices increased, some decreased, and some showed no clear trend. Comparing population subgroups (genders, educational levels, age ranges, regions, initial quintiles, and initial sector), some groups were found to have higher earnings mobility for some concepts and lower earnings mobility for others; no group was found to have higher mobility than others for every mobility concept. Finally, in both *Venezuela* and *Mexico*, the time trend of mobility was found to vary according to the notion of mobility measured (Freije, 2001; Duval Hernández, 2005).

The conclusion is that at both levels, macro and micro, it makes an important qualitative difference which mobility concept is being gauged. When a layperson asks an economist which of two situations is the more mobile, the answer 'It depends' is not very satisfying. An answer of the type 'Current incomes are more dependent on past incomes in the United Kingdom than in the United States (that is, the UK is *less* mobile in this respect than the USA), but the United Kingdom has more quintile movement than the United States (and therefore is *more* mobile than the USA in this sense)' is more informative, even if less clear-cut than the questioner may have been hoping for.

The axiomatic approach to income mobility

We have seen that there are different income mobility concepts and that the indices measuring these concepts behave differently from one another. How is the analyst to decide which notion(s) best capture (s) the essence of 'income mobility' for him or her? One approach is to proceed axiomatically, that is, to say that 'for me, mobility is such and such' and then to see which concepts, if any, embody these axioms. Two broad approaches to axiomatization may be found in the literature. In one approach, mobility is conceptualized in social welfare terms (Atkinson, 1980; King, 1983; Chakravarty, Dutta and Weymark, 1985; Dardanoni, 1993; Gottschalk and Spolaore, 2002; Ruiz-Castillo, 2004). In the other, a descriptive approach is used, wherein analysts specify the properties they wish income mobility concepts and measures to possess, and then proceed to deduce which indices, if any, have these properties (Cowell, 1985; Fields and Ok, 1996; 1999b; D'Agostino and Dardanoni, 2005). The work of Shorrocks (1978a; 1978b) makes use of both of these approaches. This difference between the ethical and the descriptive axiomatizations in the mobility literature parallels the two strands of the inequality literature (Foster and Sen, 1997): for Atkinson (1970), inequality is the amount of social welfare lost because incomes are distributed the way they are rather than being distributed perfectly equally, whereas for Sen (1973, p. 2), inequality is objective in the sense that 'one can distinguish between (*a*) "seeing" more or less inequality, and (*b*) "valuing" it more or less in ethical terms'. Note that under both the ethical and the descriptive approaches the amount of mobility recorded has or may have welfare significance. For example, many observers would say that an economy with more directional income movement has performed better than an economy with less directional income movement.

The literature offers a wide variety of axioms, some of which were designed with particular mobility concepts in mind, others of which have been explored to help sharpen what is meant by 'mobility'. Shorrocks (1993) presents 12 axioms for mobility and shows that they are mutually incompatible. In view of their incompatibility, there is a need for judgements as to which ones an analyst wants a measure to embody.

Fields and Ok (1999a) and Fields (2001) have suggested that analysts choose among the axioms by considering their views on simple examples. For example, consider the following three situations:

$$\begin{aligned} \text{I:} &\quad (1, 3) \rightarrow (1, 3) \\ \text{II:} &\quad (1, 3) \rightarrow (2, 6) \\ \text{III:} &\quad (2, 6) \rightarrow (4, 12) \end{aligned}$$

and the corresponding degree of mobility $m(x, y)$. (As above, $\rightarrow$ denotes a change in the ordered (personalized) vector of incomes.) The axiom of strong relativity, if accepted, would maintain that $m(\lambda x, \alpha y) = m(x, y)$ for all $\lambda, \alpha > 0$ and all $x, y \in \Re^n_+$. If strong relativity is accepted, it requires that Situations I, II, and III all have the same mobility. In Situation I, the only sensible amount of mobility for there to be is zero, and therefore strong relativity requires that Situations II and III also have zero mobility. An analyst who sees non-zero income mobility in Situations II and III is therefore not a strong relativity adherent.

Similarly, (weak) relativity specifies that $m(\lambda x, \lambda y) = m(x, y)$ for all $\lambda > 0$ and all $x, y \in \Re^n_+$. This axiom requires that Situations II and III have the same mobility, though not necessarily the same mobility as Situation I. Therefore, an analyst who sees more mobility in Situation III than in Situation II is not a (weak) relativity adherent either.

The literature offers characterizations of some of the mobility measures that have been used – for example, Fields and Ok's (1996; 1999b) measures of non-directional and directional income movement and Chakravarty, Dutta and Weymark's (1985) index of mobility as welfare change. More commonly, though, the axioms are used to state a number of desirable properties and then display a measure or a family of measures consistent with these properties.

In summary, a fruitful way for the analyst to choose which mobility concept(s) is (are) most salient for oneself is to consider the axiomatic judgements underlying each of the concepts. To date, some but not all of the income mobility concepts have been so characterized.

Other issues

The income mobility literature has a number of other issues that remain more or less contentious, not because the different views have not been worked out but because different analysts hold genuinely different positions on a number of important matters.

Is all distributional change ‘ mobility ’ or only some of it?

Lurking in the background of some writings on income mobility is a fundamental difference of opinion about what income mobility is. For the majority of analysts, the notion of ‘income mobility’ has both absolute and relative components. For example, if all incomes double, most would judge there to be more mobility than if all incomes remain unchanged. For some analysts, though, the notion of ‘income mobility’ is relative only; therefore, the change in the mean needs to be taken out, and ‘mobility’ applies

only to what is left.

Thinking of ‘mobility’ in this way can lead to some controversial judgements. For example, Chakravarty, Dutta and Weymark (hereafter CDW) (1985) propose the following mobility index:

$$M_{CDW} \equiv (E(y_{agg}) / E(b)) - 1,$$

where $E(.)$ is an equality measure, y_{agg} is a vector of aggregate incomes over the observation period, and b is the benchmark vector of incomes under the assumption of complete relative immobility following the first period. In the case in which $E(.)$ is a relative equality measure, the term $E(b)$ is replaced by $E(x)$, where x is the vector of first-period incomes. In the view of these authors (CDW, 1985, p. 8): ‘Socially desirable mobility is associated with income structures having positive index values while socially undesirable mobility is associated with income structures having negative index values.’ Thus, given their index, CDW judge that mobility contributes positively to social welfare if and only if y_{agg} is distributed more equally than x. Thus, if all incomes rise but the percentage gains are larger at the top end of the income distribution than they are at the bottom, mobility would be judged by CDW to have been socially *undesirable*, in direct contradiction to the quasi-Paretian welfare judgement that an increase in some incomes with no decline in others *raises* social welfare. This difference of views – whether ‘income mobility’ includes the growth aspect of distributional change or whether ‘mobility’ is what remains after growth has been taken out – underlies much of the mobility literature, but rarely is it made explicit.

What is ‘relative mobility’?

As already noted, the term ‘relative mobility’ is used ambiguously, sometimes to refer to mobility notions characterized by strong relativity $m(\lambda x, \alpha y) = m(x, y)$ for all $\lambda, \alpha > 0$ and all $x, y \in \Re^n_+$ and sometimes to refer to those characterized by weak relativity $m(\lambda x, \lambda y) = m(x, y)$ for all $\lambda > 0$ and all $x, y \in \Re^n_+$. Note that for both of these relativity notions the basis for determining whether a given individual is experiencing upward or downward relative mobility is that individual's change in *income* relative to the *income* changes of others.

However, the term ‘relative mobility’ is used in yet another sense, namely, to refer to *positional* movements. On this view, an individual experiences relative mobility if and only if he or she changes position (quintile, decile, centile, or rank) from base year to final year. For example, Jenkins and Van Kerm (2003) break down trends in income inequality into a ‘pro-poor income growth’ component and an ‘income mobility’ component. The ‘income mobility’ component involves re-rankings and only re-rankings. Thus, for them as for some others, mobility *is* positional movement and nothing more.

Finally, D'Agostino and Dardanoni (2005) have yet a different definition of relative mobility. For them, relative mobility involves a change in an individual's relative standing with respect to all others, whereas absolute status is something that can be derived by looking at data regarding the individual taken in

isolation.
This last point raises the issue of what is meant by ‘absolute mobility,’ to which we now turn.

What is ‘absolute mobility’?

The term ‘absolute mobility’ is used in at least three different ways in the income mobility literature. One way is to express a concern with gains and losses of *income* rather than *income shares* or *positions*. In this sense, the concept of directional income movement and the various measures of that concept are about absolute mobility. Second, ‘absolute mobility’ is sometimes used to mean that the analyst is concerned with the *absolute value* of income changes, as would be the case in studies of non-directional income movement, or flux. Third, the term is used in the sense of *translation invariance*, in the sense that, if all initial and final incomes are increased by the same amount, the new situation has the same absolute mobility as the original one, that is, $m(x + \alpha, y + \alpha) = m(x, y)$.
As is the case elsewhere in economics, when a term has more than one meaning within the same literature, it is probably best to drop the term altogether. Henceforth, researchers would do better to speak of dollar-based, absolute-value-based, or translation-invariant income mobility measures in preference to ‘absolute mobility’.

Is ‘income mobility’ decomposable, and if so, how?

Consider the total income mobility recorded in a population. Under what circumstances can the total be broken down into component parts?
Of the six income mobility concepts considered above, one involves the time-independence aspect of mobility and the other five involve the movement aspect of mobility. The time-independence aspect of mobility is not decomposable. However, there have been decompositions of various movement measures. One type of decomposition is subgroup decomposability, that is, if the population is divided into J subgroups, the total income mobility in the population as a whole equals a (possibly) weighted average of the mobility in each of the subgroups:

$$m(x, y) = \sum_{j=1}^{J} w_j m_j(x, y) .$$

A number of income mobility measures are subgroup decomposable; examples are Fields and Ok's (1996; 1999b) non-directional income movement measures

$$m_1(x, y) \equiv \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

$$\text{and}\, m_2(x, y) \equiv \frac{1}{n}\sum_{i=1}^{n}|\log y_i - \log x_i|$$

and their directional income movement measure

$$m_3(x, y) \equiv \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log x_i)\ .$$

A second kind of decomposition is into substantively meaningful components. There is a long tradition in the sociology literature (for example, Bartholomew, 1982) of breaking down the movement of individuals among occupations or social classes into two component parts: (*a*) changes that can be attributed to the increased availability of positions in the better occupations and social classes (‘structural mobility’) and (*b*) changes that can be attributed to increased movement of individuals among occupations and social classes for a given distribution of positions among these classes (‘exchange mobility’). Bridging the economics and sociology literatures, Markandya (1982; 1984) proposes two alternative decompositions of income mobility along these lines. The first defines exchange mobility as the proportion of the change in welfare that could have been obtained if the income distribution had stayed constant through time, in which case structural mobility is defined as the residual welfare change. The second defines structural mobility as the change in welfare that would have taken place if the two-period or two-generation transition matrix had exhibited complete immobility, in which case exchange mobility is defined as the residual. Along similar lines, Ruiz-Castillo (2004) shows how the CDW (1985) index of welfare due to mobility could be decomposed into either (*a*) a precisely defined structural component and a residual representing exchange mobility or (*b*) a precisely defined exchange component and a residual representing structural mobility. In all these cases, the residual component makes the decomposition exact but in a rather unexciting way.

The results just cited do *not* mean that an exact additive decomposition of income mobility is impossible. Fields and Ok (1996) show that their mobility index $m_1(x, y) \equiv \frac{1}{n}\Sigma_{i=1}^{n}|y_i - x_i|$ is decomposable into the sum of appropriately defined structural and exchange components. In the case of a growing economy, the decomposition equation is $m_1(x, y) = (\Sigma_{i=1}^{n} y_i - \Sigma_{i=1}^{n} x_i) + 2\Sigma_{\{i: y_i < x_i\}}(x_i - y_i)$. An analogous decomposition holds for a contracting economy. Along similar lines, Fields and Ok (1999b) show that their directional movement measure $m_3(x, y) \equiv \frac{1}{n}\Sigma_{i=1}^{n}(\log y_i - \log x_i)$ is decomposable into social utility growth and social utility transfer components. In all of these cases, the weakness of Markandya's and Ruíz-Castillo's

residual approaches is averted.

What other empirical issues arise?

Empirical researchers should bear in mind two additional issues. One is that, as an empirical matter, the longer the observation period, the greater is the amount of mobility registered (Atkinson, Bourguignon and Morrisson, 1992). Therefore, care should be taken not to compare, for example, two-year mobility in one context with, for example, five-year mobility in another.

Second, measurement error is a serious issue. There is an ample literature on mismeasurement of earnings *levels* but, as yet, only a very limited literature on mismeasurement of earnings *changes* (Deaton, 1997; Bound, Brown and Mathiowetz, 2001). A task for the future is to estimate empirically the effect of measurement error on estimates of both macro-mobility and micro-mobility.

Conclusions

The income mobility literature is fundamentally unsettled. This is because the very term ‘income mobility’ connotes different things to different people. This article has reviewed a number of dimensions in which differences arise: which of six notions most accurately captures the fundamental idea of ‘income mobility’, which indices best measure each of the concepts, which axioms best characterize the essence of ‘income mobility’, how income mobility has been evolving over time in different countries, which demographic groups have more mobility than others in different settings, and which theoretical refinements to the notion of ‘income mobility’ hold the greatest promise.

Given the unsettled state of the field, before researchers ‘do a mobility study’, it is important that we specify which concept or concepts of mobility we are considering, which measures of these concepts we are using, and which questions we are answering. More than once, when I have given seminars, a member of the audience has raised his or her hand and said, ‘But that's not what mobility *is*’. Let us do all that we can to clarify what we are talking about so that we do not talk past one another any more than we have to.

See Also

- inequality (measurement)
- intergenerational income mobility
- longitudinal data analysis

Bibliography

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 244–63.

Atkinson, A. 1980. The measurement of economic mobility. In *Inkomensverdeling en Openbare Financién*, ed. P. Eijgelshoen and L. van Gemerden. Utrecht: Het Spectrum.

Atkinson, A.B., Bourguignon, F. and Morrisson, C. 1992. *Empirical Studies of Earnings Mobility*. London: Harwood Academic Publishers.

Bartholomew, D. 1982. *Stochastic Models for Social Processes*. London: Wiley.

Bound, J., Brown, C. and Mathiowetz, N. 2001. Measurement error in survey data. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Buchinsky, M., Fields, G., Fougère, D. and Kramarz, F. 2004. Francs or ranks? Earnings Mobility in France, 1967–1999. Mimeo. Paris: INSEE.

Chakravarty, S., Dutta, B. and Weymark, J. 1985. Ethical indices of income mobility. *Social Choice and Welfare* 2, 1–21.

Chronic Poverty Research Centre. 2004. *The Chronic Poverty Report 2004–05*. Manchester: Chronic Poverty Research Centre.

Cowell, F. 1985. Measures of distributional change: an axiomatic approach. *Review of Economic Studies* 52, 135–51.

D'Agostino, M. and Dardanoni, V. 2005. The measurement of mobility: a class of distance indices. Paper presented to the Society for the Study of Economic Inequality, Palma de Mallorca, Spain, July.

Dardanoni, V. 1993. Measuring social mobility. *Journal of Economic Theory* 61, 372–94.

Deaton, A. 1997. *The Analysis of Household Surveys*. Baltimore, MD: Johns Hopkins University Press.

Duval Hernández, R. 2005. Dynamics of labor market earnings and sector of employment in urban Mexico. Doctoral dissertation, Cornell University.

Fields, G. 2001. *Distribution and Development: A New Look at the Developing World*. Cambridge, MA: MIT Press and Russell Sage Foundation.

Fields, G. 2005. Does income mobility equalize longer-term incomes? New measures of an old concept. Mimeo, Cornell University.

Fields, G., Leary, J. and Ok, E. 2002. Stochastic dominance in mobility analysis. *Economics Letters* 75, 333–9.

Fields, G. and Ok, E. 1996. The meaning and measurement of income mobility. *Journal of Economic*

Theory 71, 349–77.

Fields, G. and Ok, E. 1999a. The measurement of income mobility: an introduction to the literature. In *Handbook on Income Inequality Measurement*, ed., J. Silber. Boston: Kluwer.

Fields, G. and Ok, E. 1999b. Measuring movement of incomes. *Economica* 66, 455–72.

Foster, J. and Sen, A. 1997. *On Economic Inequality*, expanded edition. Oxford: Oxford University Press.

Freije, S. 2001. Household income dynamics in Venezuela. Unpublished doctoral. dissertation, Cornell University.

Gottschalk, P. and Spolaore, E. 2002. On the evaluation of economic mobility. *Review of Economic Studies* 69, 191–208.

Jarvis, S. and Jenkins, S. 1998. How much income mobility is there in Britain? *Economic Journal* 108, 1–16.

Jenkins, S. and Van Kerm, P. 2003. Trends in income inequality, pro-poor income growth, and income mobility. Discussion Paper No. 904. Bonn: IZA.

King, M. 1983. An index of inequality, with applications to horizontal equity and social mobility. *Econometrica* 51, 99–115.

Krugman, P. 1992. The rich, the right, and the facts. *American Prospect* 11, 19–31.

Maasoumi, E. 1998. On mobility. In *Handbook of Applied Economic Statistics*, ed. D. Giles and A. Ullah. New York: Marcel Dekker.

Markandya, A. 1982. Intergenerational exchange mobility and economic welfare. *European Economic Review* 17, 301–24.

Markandya, A. 1984. The welfare measurement of changes in economic mobility. *Economica* 51, 457–71.

OECD (Organisation for Economic Co-Operation and Development). 1996. *Employment Outlook 1996*. Paris: OECD.

OECD. 1997. *Employment Outlook 1997*. Paris: OECD.

Ruiz-Castillo, J. 2004. The measurement of structural and exchange mobility. *Journal of Economic Inequality* 2, 219–28.

Sánchez Puerta, M. 2005. Earnings mobility in urban Argentina. Unpublished doctoral dissertation, Cornell University.

Schumpeter, J. 1955. *Imperialism and Social Classes*. New York: Meridian Books.

Sen, A. 1973. *On Economic Inequality*. New York: Norton.

Shorrocks, A. 1978a. Income inequality and income mobility. *Journal of Economic Theory* 19, 376–93.

Shorrocks, A. 1978b. The measurement of mobility. *Econometrica* 46, 1013–24.

Shorrocks, A. 1993. On the Hart measure of income mobility. In *Industrial Concentration and Economic Inequality*, ed. M. Casson and J. Creedy. London: Edward Elgar.

Slemrod, J. 1992. Taxation and inequality: a time-exposure perspective. In *Tax Policy and the Economy*, vol. 6, ed. J. Poterba. Cambridge, MA: MIT Press for the NBER.

Solon, G. 1999. Intergenerational mobility in the labor market. In *Handbook of Labor Economics*, vol. 3, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

How to cite this article

Fields, Gary S. "income mobility." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000271> doi:10.1057/9780230226203.0770

The New Palgrave Dictionary of Economics Online

income taxation and optimal policies

Louis Kaplow

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Various economic literatures address the question whether first-best prescriptions for government policy require modification because redistributive income taxation distorts labour supply and cannot achieve the distributive ideal. Perhaps second-best rules for public goods provision, corrective taxation, public sector pricing, and other government activity should reflect concerns about distribution and labour supply distortion. Recent work demonstrates, however, that in basic cases first-best principles remain applicable. Demonstrations make use of income tax adjustments that preserve not only budget balance but also the pre-reform distribution of utility.

Keywords

commodity taxation; corrective taxes; cost–benefit analysis; distortion; distribution; environmental taxation; externalities; income taxation; lump-sum tax; marginal cost pricing; optimal taxation; Pigouvian tax; public goods; public sector pricing; Ramsey taxation; redistribution; regulation; Samuelson, P.; second best; taxation

Article

Optimal policy analysis is complicated by problems of the second best. Two of the most important problems – non-ideal distribution and labour supply distortion – are intimately connected with limitations of income taxation. In a first-best world, individualized lump-sum taxes can be used to achieve any desired distribution without causing distortion. Accordingly, the optimal design of other government policies is dictated by familiar first-best rules: the Samuelson cost-benefit test for public goods, the Pigouvian prescription for externalities to equate the full marginal social costs and benefits, marginal cost pricing for publicly provided goods and services and for regulated utilities, and so forth. In practice, however, informational limitations require the use of distortionary instruments, notably labour income taxation, so even at the optimum (Mirrlees, 1971) the distributive ideal is not achieved.

Due to the second-best nature of the optimal income taxation problem, it is natural to consider whether first-best prescriptions for other government policies should be modified in order to assist the redistributive function. In addition, such other policies – most obviously but not exclusively those that raise or expend revenue – may affect labour supply, which also may require modification of standard policy rules. Particularly since the explosion of interest in optimal taxation in the 1970s, extensive literatures have developed to address these issues in each particular context. Much work focuses on distortion, some on distribution, and a portion considers both simultaneously. A range of adjustments to first-best formulas have been proposed, revisions that in general depend on the initially prevailing income tax and on the modification thereof that is assumed to accompany the underlying policy reform. Another strand of research offers a new view of the second-best problem in each of these areas and allows a substantial synthesis across these seemingly different contexts. To analyse these issues, this literature employs a construction under which the income tax modification hypothesized to accompany any policy change is one that, in combination with the altered policy, holds the distribution of utility constant. In a simple standard model, it turns out that first-best policy principles are applicable without refinements: there is no need for distributive adjustments since distribution is unaffected; and, as it happens, holding distribution constant also leaves labour supply unchanged, rendering unnecessary any adjustments on account of labour supply distortion.

The analysis of income taxation and optimal government policy is best introduced in the most fundamental setting, in which the only question is whether a labour income tax should be supplemented by differential commodity taxes. As will be elaborated in the first section below, the answer is negative in simple cases regardless of whether the initial income tax is optimal, a result that in an important sense displaces principles of Ramsey taxation (and, as will subsequently be noted, other applications of Ramsey principles as well). The next section explains how a range of government policies – including public goods provision, regulation of externalities, and public sector pricing – are all formally analogous to differential commodity taxation. Hence, the results (and qualifications) can readily be extended, which allows for the understanding of second-best problems in these disparate fields to be unified substantially. Two final sections relate the analysis to classical and contemporary work and explore further implications of this approach for second-best policy analysis.

Commodity taxation

The problem of optimal commodity taxation with labour income taxation can be stated as follows. Individuals choose commodity vectors $\boldsymbol{x}$ and labour effort l to maximize the utility function $u(v(\boldsymbol{x}), l\bullet)$, where v is a subutility function. This form of the utility function entails what is referred to as weak separability of labour: for a given level of after-income-tax income, individuals will allocate their disposable income among commodities in the same manner regardless of the level of labour effort required to earn that level of income.

An individual's budget constraint requires that expenditures, $\rho\,\boldsymbol{x}(wl\bullet)$, not exceed before-tax income, wl, minus income taxes, $T(wl\bullet)$, which can be negative, thereby allowing for net transfers; ρ is the consumer price vector, w is an individual's wage, and $\boldsymbol{x}(wl\bullet)$ denotes the consumption vector chosen by an individual who earns wl. Individuals' wages w have density $f(w)$, and the government is assumed to know this density but not each individual's wage, which renders individualized lump-sum taxes infeasible. The consumer price vector ρ is understood as the sum of a producer price vector (taken to be

constant and equal to production costs) and a vector of commodity taxes (which, if negative, are subsidies).

The government's maximization problem is to select commodity taxes (equivalently, ρ) and an income tax schedule $T(wl\bullet)$ to maximize a standard concave social welfare function, subject to meeting a given revenue requirement and to incentive compatibility constraints deriving from individuals' maximization problems. If commodity taxes are taken to be zero, we have the optimal nonlinear income tax problem of Mirrlees (1971).

Atkinson and Stiglitz (1976) demonstrated that, when the income tax is set optimally, commodity taxes should be undifferentiated (i.e., uniform) in this basic setting. The derivation to follow is taken from Kaplow (2006), who does not require that the income tax be optimal and provides a more intuitively accessible approach.

For any commodity tax reform, which changes the consumer price vector from ρ to ρ^*, suppose that the income tax schedule is initially adjusted from $T(wl\bullet)$ to $T^\circ(wl\bullet)$ such that $V(\rho^*, T^\circ, wl\bullet)=V(\rho, T, wl\bullet)$ for all wl, where V is an indirect subutility function indicating the maximized value of $v(\boldsymbol{x})$, subject to the budget constraint, where ρ, T, and wl are taken as given. That is, one adjusts the income tax schedule to the $T^\circ(wl\bullet)$ that restores the original level of subutility achieved at each level of disposable income; hence, $T^\circ(wl\bullet)-T(wl\bullet)$ is the schedule of utility-compensating changes in disposable income.

This income tax schedule adjustment has a number of properties. First, if individuals do not change their level of labour supply, they achieve the same utility, for u depends only on v (which is held fixed, given $l\bullet$) and l.

Second, faced with this income tax adjustment, individuals will not in fact change their level of labour supply: each individual's (each type w's) total utility u for any choice of l after this combined reform of commodity taxes and the income tax precisely equals the total utility for that choice of l before the reform; therefore, whatever l previously maximized utility must continue to do so.

Third, the hypothetical reform will in general affect government revenue. Specifically, it can be shown that there will be a surplus if and only if the reform increases efficiency in the narrow sense – by reducing aggregate distortion among commodities – a condition that will prevail, for example, if all commodity taxes (and subsidies) are moved proportionally toward zero, including the case of complete abolition of differential commodity taxation. The reason is that reducing consumption distortion, ceteris paribus, raises individuals' utilities; because the income tax adjustment is set to hold utility constant, it must therefore reduce individuals' disposable income to offset what would otherwise be a utility increase. Accordingly, net tax collections must rise.

Finally, to complete the analysis, budget balance can be restored by further adjusting T to rebate the surplus pro rata: $T^*(wl\bullet)=T^\circ(wl\bullet)-c$, where c is some positive constant. The result is a Pareto improvement, for utility was unchanged until this final stage of the reform. To summarize, if any commodity tax reform is accompanied by an income tax adjustment that, when combined with the underlying reform, holds utility constant (until the rebate stage), there is no effect on distribution, labour supply is unchanged, and there is a surplus, allowing a Pareto improvement, if and only if the underlying commodity tax reform is efficient in a narrow, conventional sense.

It is useful to consider the intuition behind this result. It is familiar from the general theory of second-best analysis (Lipsey and Lancaster, 1956) that first-best conditions do not generally govern once some distortion is introduced. However, in the present setting the only unavoidable distortion is of the labour–

leisure choice, and differential commodity taxation does not help to alleviate it. Thus, differential taxes involve the cost of distorting consumption without any offsetting benefit. The reason that differential commodity taxes cannot help offset the labour–leisure distortion is the assumption of weak separability. Just as different levels of labour supply do not change preferences among commodities, so different consumption allocations do not change the disutility of labour.

This result on the inefficiency of differential commodity taxation provides an important benchmark for understanding and analysis. The conclusion is subject to many qualifications, each of which is best appreciated by reference to this basic starting point. First, as follows immediately from the preceding remarks, weak separability may be violated. This is the point, first elaborated by Corlett and Hague (1953), that it tends to be efficient to tax leisure complements (perhaps beach attendance or reading) and subsidize complements to labour (possibly central city transit or amenities). Second, preferences were taken to be homogeneous, but if preferences depend on unobservable ability it would be optimal to tax commodities preferred by the more able (independent of income per se), perhaps high-brow art, and to subsidize those preferred by the less able. Additional qualifications have been offered, including, importantly, concerns with administration and tax avoidance that may affect income taxation, especially in developing countries.

The foregoing analysis is usefully contrasted with that of Ramsey (1927) taxation, which involves a substantial, widely known literature that itself provides the foundation for much economic analysis of myriad other policy applications (including all those examined in the following section). Most familiar is the rule that commodity taxes should be inversely proportional to the elasticity of demand, with refinements for demand interdependencies. Also well known are modifications due to distributive concerns, which favour taxing luxuries and subsidizing necessities, commands that often conflict with the inverse elasticity rule and thus require trade-offs (Feldstein, 1972; Diamond, 1975). As initially emphasized in Atkinson and Stiglitz (1976), however, neither prescription is apt if there is an income tax. In the original Ramsey model in which all individuals are identical and thus there are no distributive concerns, the optimal tax is a uniform lump-sum extraction (a limiting case of an income tax), which, it should be noted, neither requires information about individuals' types nor is distributively objectionable in this setting. When differences in earning ability are admitted, the optimal tax is a nonlinear income tax, and in typical cases the lump-sum component involves a uniform lump-sum subsidy. Nevertheless, optimal commodity taxation still is not guided either by the familiar inverse-elasticity rule or by the general preference for harsher treatment of luxuries than of necessities; as noted, in the basic case, optimal differentiation is nil regardless of the demand elasticity or how demand changes with income.

Paradoxically, the literatures that build upon Ramsey's path-breaking contribution are motivated by second-best concerns, yet it turns out that a more complete second-best analysis – notably, incorporating the income tax, the primary distributive tool and also a central cause of unavoidable distortion that calls for second-best inquiry – returns us to a simple, first-best rule in the benchmark case. Here, that prescription is against differential commodity taxes on account of the resulting distortion of consumption. As will now be explained, this pattern of analysis is replicated with regard to a broad range of government policies.

Government policies generally

The foregoing framework can be employed to address the optimal provision of public goods, the optimal control of externalities, and other government actions, as developed by Kaplow (1996; 2004; 2008). The reason is that departures from first-best rules in these contexts are formally analogous to differential commodity taxation and hence are inefficient in the basic case (a conclusion that also is subject to similar qualifications).

To see this, suppose now that individuals have the utility function $u(v(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{g}), l\bullet)$. Here, $\boldsymbol{e}$ is a vector of externalities (suppose, for example, that each element of $\boldsymbol{e}$ is the population's total consumption of the corresponding commodity in the vector $\boldsymbol{x}$), and $\boldsymbol{g}$ is a vector of public goods. This functional form maintains the assumption that labour is weakly separable from other sources of individuals' utility.

We can again consider reforms, here of commodity taxes (and subsidies) ρ, but now with the thought of internalizing externalities, or of $\boldsymbol{g}$. Again, we can construct $T°(wl\bullet)$ such that individuals' subutility v is kept constant if they choose to supply the same level of labour. As before, this reform is distribution neutral and in fact induces all individuals to supply the same labour effort. (A review of the foregoing analysis will confirm that nothing depended on the fact that the reform was only of commodity taxes or that there were no externalities or public goods involved.)

The question, then, is whether the intermediate adjustment of the income tax schedule, from $T(wl\bullet)$ to $T°(wl\bullet)$, will produce a surplus or a deficit. With externalities, if, for example, one sets all commodity taxes equal to the marginal external effect of consumption on individuals' utilities – the traditional Pigouvian prescription (Pigou, 1920) – there will be a surplus: individuals may be better or worse off because of being subject to a different vector of commodity prices, and they may be better or worse off on account of changes in the levels of externalities; however, it can be demonstrated that the net effect on revenue is positive, essentially because of traditional efficiency considerations. (Note that the income tax adjustment from $T(wl\bullet)$ to $T°(wl\bullet)$ taxes away all sources of surplus and compensates for any disutility; hence, the sign of the net revenue effect is given by the sign of the total of all changes in individuals' surplus from the underlying reform.) Observe that this result is very similar in spirit to that on commodity taxation without externalities. There, the optimum involves setting consumer prices equal to true marginal resource costs of commodities; with externalities, the same principle holds, but true resource costs now include not only production costs but also effects on others' utilities.

For public goods, the total revenue effect has two components. The first (which is negative) is the production cost of the public goods, and the second is (by the method of construction of $T°(wl\bullet)$) the integral of individuals' surplus from changes in the levels of the public goods. Hence, there is a surplus (deficit) if and only if the reform passes (fails) the Samuelson (1954) cost–benefit test, which asks whether the integral of individuals' benefits exceeds the cost of producing the public goods. The essence of the argument is again similar to that for the basic case with commodity taxation. For example, supplying less of a public good than dictated by the Samuelson test corresponds to imposing a differential tax on a private good. To push the analogy further, consider a hypothetically decentralized regime in which consumer prices for private goods correspond to Lindahl prices for public goods, and commodity taxes on public goods are defined as the difference between the price charged to a consumer in the imaginary regime and that consumer's marginal rate of substitution. The source of the allocative inefficiency is again a failure of the prices faced by consumers to equal true marginal resource costs.

In the present setting, therefore, moving to the first best – now regarding internalization of externalities or provision of public goods rather than setting commodity taxes in a simpler world – makes possible a Pareto improvement. Concerns about distribution and labour supply effects caused by the income tax

can be ignored because they are moot.

Similar logic can be employed to address other areas of government policy, most obviously regulations that mimic corrective taxation but also seemingly unrelated fields like public sector pricing and utility regulation. Thus, marginal cost pricing will be optimal in spite of distributive concerns or the distortionary cost of raising funds to meet deficits because, if the income tax is adjusted in the manner described, distribution will be unaffected and there will be a net surplus if the reform is (narrowly) efficient in the basic case.

Historical development of second-best policy rules

First-best principles have a long and familiar lineage. The command to internalize externalities is inspired by Pigou's (1920) classical treatment, and the cost-benefit test for public goods is due to Samuelson's (1954) elegant formulation. It is notable that Samuelson (1954) explicitly said that he was considering a first-best setting in which individualized lump-sum taxes permitted any social welfare optimum to be implemented.

Second-best qualifications start with another of Pigou's (1928) books, in which he observed that, on account of the resource cost of raising revenue, public goods probably should have to meet a higher standard. Refinements appeared in Atkinson and Stern (1974), Diamond and Mirrlees (1971), and Stiglitz and Dasgupta (1971), with subsequent research crystallized by Ballard and Fullerton (1992). Analogous work on environmental taxation – addressed to the possibility of a ‘double dividend’ (a tax might both internalize an externality and raise revenue distortion-free) and qualifications implying a more negative view of corrective policies – became intense in the 1990s (see Bovenberg and Goulder, 2002; Goulder, 2002). Largely separate literatures proposed second-best adjustments to account for distributive effects (Weisbrod, 1968; Drèze and Stern, 1987). See also Bös (1985) on public sector pricing.

Much of this work builds on Ramsey's (1927) model of taxation and extensions thereof. Often, such analyses employ the original representative-individual model in which distribution is immaterial; yet, at the same time, the possibility of income taxation is ignored (specifically, the possible use of a uniform grant that, as noted above, makes commodity taxation unnecessary) or the income tax adjustments that are stipulated turn out not to be distribution-neutral. Literature focusing on distribution also often ignores the availability of the income tax.

The lessons presented in the prior sections arise from another line of work that developed intermittently and largely independently of the foregoing literatures. Hylland and Zeckhauser (1979) used a distribution-neutral income tax adjustment with a special case of individuals’ utility functions to show that distributive weights are inappropriate in cost–benefit analysis. Shavell (1981) offers a similar demonstration for legal rules. Christiansen (1981) and Boadway and Keen (1993) show that, with an optimal income tax, the basic cost–benefit test for public goods is appropriate. Kaplow (1996; 2004; 2006; 2008) considers both distribution and labour supply distortion, does not require the income tax to be optimal, and examines a broad range of government policies.

Implications

Ever since Lipsey and Lancaster (1956), economists have sought to develop principles to provide guidance in a second-best world; indeed, in the area of taxation, the search had already begun. The inability to achieve an ideal distribution without distortion is one of the most important unavoidable deviations from the first best. Thus, not surprisingly, substantial research addresses second-best concerns regarding income taxation and commodity taxation as well as all manner of government policies that may have distributive effects or influence government revenue.

Perhaps surprisingly, a number of first-best principles prove to be rather robust in basic, benchmark cases. Important caveats were noted, but, importantly, they are largely orthogonal to the original second-best concerns that motivate most research in these fields.

One further qualification deserves attention. The present analysis assumes that the income tax will be adjusted in a distribution-neutral manner. This is hardly an unnatural assumption. For example, if the initial income tax does not optimally trade off distribution and distortion, the divergence may arise from political forces that dictate some other degree of redistribution. If so, particular reforms might be expected to leave that distributive balance unaltered.

Nevertheless, consider the possibility of non-distribution-neutral adjustments of the income tax. As suggested in Kaplow (1996; 2004; 2008), a simple two-step decomposition is illuminating in this case:

1. 1. Assume that, initially, the underlying policy is implemented in the previously hypothesized distribution-neutral fashion.
2. 2. Assume also that, a moment later, a further income tax adjustment transforms the policy in step 1 into the actually imagined policy.

Analysis of step 1 can proceed as before. Step 2, observe, is a purely redistributive reform. Accordingly, the analysis is in the province of optimal income taxation and involves the familiar distribution-distortion trade-off. Significantly, the analysis of step 2 is generic – that is, it is the same regardless of whether step 1 involves changing commodity taxes, one or another regulation, the level of some public good, or indeed nothing at all (a purely redistributive overall reform). For economists, this allows substantial specialization. Step 2 analysis must be undertaken anyway and, as noted, tends to be independent of step 1. Step 1 analysis can be undertaken by experts on gasoline taxes, health care, electric utilities, and so forth, who need not concern themselves with redistribution. Policymakers can combine analyses as appropriate.

Specialization has an additional virtue in this context: it facilitates communication, both among researchers and to policymakers. For example, a study of a highway project that does not focus on step 1 will need to include analysis of (*a*) direct effects of the highway project (such as on pollution or congestion), (*b*) what other, budget-accommodating tax adjustment will in fact be made in the long run (an exercise in political economy), (*c*) an analysis of the effects of the resulting change in the extent of redistribution, and (*d*•) a social welfare assessment, requiring choice of a social welfare function. Relatedly, when studies of a highway project reach different conclusions, the discrepancies may arise from any combination of these four components, making it difficult to compare and synthesize research. A particular concern arises with much work in these literatures, both abstract and highly applied, because step 1 is often combined with an incomplete analysis of step 2. For example, work might identify a redistributive benefit from a policy; yet, if there is not a complete analysis of redistributive

taxation, the likely associated increase in labour supply distortion may be overlooked. Contrariwise, much work identifies increases in distortion, failing to recognize that the increases are due to effects on labour supply that accompany an implicit increase in redistribution, the benefit of which is omitted. Because of the original second-best problem, involving redistribution through distortionary taxation, redistribution is not an unambiguous good because (usually) it comes at a cost, and distortion – particularly of labour supply – is not an unmitigated evil because (frequently) it is symptomatic of an underlying benefit. Analysis that incorporates one side of the balance while excluding the other may be the worst approach of all.

To summarize, Ramsey principles are widely acknowledged and broadly employed as a foundation for second-best policy analysis. However, at least in developed economies in which an income tax is feasible, the model's most familiar implications for differential commodity taxation are inapt and, by extension, so are its applications to public goods provision, regulation of externalities, public sector pricing, and other policy areas. In the basic case, the problem of optimal redistribution – involving the trade-off of distribution and labour supply distortion – is separable from these other realms. Accordingly, traditional first-best principles that focus on efficiency in the area under consideration provide a useful benchmark. Complications abound, but for the most part they do not replicate the adjustments called for by the original Ramsey model or typical applications thereof. Instead, they are best understood by direct reference to the problem of redistributive income taxation.

See Also

- compensation principle
- environmental economics
- Mirrlees, James
- optimal taxation
- Pigouvian taxes
- public goods
- Ramsey model
- redistribution of income and wealth
- taxation of income

I am grateful to Steven Shavell for comments and the John M. Olin Center for Law, Economics, and Business at Harvard University for financial support.

Bibliography

Atkinson, A. and Stern, N. 1974. Pigou, taxation, and public goods. *Review of Economic Studies* 41, 119–28.

Atkinson, A. and Stiglitz, J. 1976. The design of tax structure: direct versus indirect taxation. *Journal of Public Economics* 6, 55–75.

Ballard, C. and Fullerton, D. 1992. Distortionary taxes and the provision of public goods. *Journal of Economic Perspectives* 6, 117–31.

Boadway, R. and Keen, M. 1993. Public goods, self-selection and optimal income taxation. *International Economic Review* 34, 463–78.

Bös, D. 1985. Public sector pricing. In *Handbook of Public Economics*, vol. 1., ed. A. Auerbach and M. Feldstein. Amsterdam: Elsevier.

Bovenberg, A. and Goulder, L. 2002. Environmental taxation and regulation. In *Handbook of Public Economics*, vol. 3, ed. A. Auerbach and M. Feldstein. Amsterdam: Elsevier.

Christiansen, V. 1981. Evaluation of public projects under optimal taxation. *Review of Economic Studies* 48, 447–57.

Corlett, W. and Hague, D. 1953. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21, 21–30.

Diamond, P. 1975. A many-person Ramsey tax rule. *Journal of Public Economics* 4, 283–99.

Diamond, P. and Mirrlees, J. 1971. Optimal taxation and public production. II: Tax rules. *American Economic Review* 61, 261–78.

Drèze, J. and Stern, N. 1987. The theory of cost–benefit analysis. In *Handbook of Public Economics*, vol. 2., ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Feldstein, M. 1972. Equity and efficiency in public sector pricing: the optimal two-part tariff. *Quarterly Journal of Economics* 86, 175–87.

Goulder, L. 2002. *Environmental Policy Making in Economies with Prior Tax Distortions*. Cheltenham, UK: Edward Elgar.

Hylland, A. and Zeckhauser, R. 1979. Distributional objectives should affect taxes but not program choice or design. *Scandinavian Journal of Economics* 81, 264–84.

Kaplow, L. 1996. The optimal supply of public goods and the distortionary cost of taxation. *National Tax Journal* 49, 513–33.

Kaplow, L. 2004. On the (ir)relevance of distribution and labor supply distortion to government policy. *Journal of Economic Perspectives* 18, 59–75.

Kaplow, L. 2006. On the undesirability of commodity taxation even when income taxation is not optimal. *Journal of Public Economics* 90, 1235–50.

Kaplow, L. 2008. *The Theory of Taxation and Public Economics*. Princeton: Princeton University Press.

Lipsey, R. and Lancaster, K. 1956. The general theory of second best. *Review of Economic Studies* 24, 11–32.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 68, 175–208.

Pigou, A. 1920. *The Economics of Welfare*. London: Macmillan.

Pigou, A. 1928. *A Study in Public Finance*. London: Macmillan.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36, 387–89.

Shavell, S. 1981. A note on efficiency vs. distributional equity in legal rulemaking: Should distributional equity matter given optimal income taxation? *American Economic Review* 71, 414–18.

Stiglitz, J. and Dasgupta, P. 1971. Differential taxation, public goods, and economic efficiency. *Review of Economic Studies* 38, 151–74.

Weisbrod, B. 1968. Income redistribution effects and cost–benefit analysis. In *Problems in Public Expenditure Analysis*, ed. S. Chase. Washington, DC: Brookings Institution.

How to cite this article

Kaplow, Louis. "income taxation and optimal policies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000245> doi:10.1057/9780230226203.0771

The New Palgrave Dictionary of Economics Online

incomplete contracts

Oliver Hart
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The past decade has witnessed a growing interest in contract theories of various kinds. This development is partly a reaction to our rather thorough understanding of the standard theory of perfect competition under complete markets, but more importantly to the resulting realization that this paradigm is insufficient to accommodate a number of important economic phenomena. Studying in more detail the process of contracting – particularly its hazards and imperfections – is a natural way to enrich and amend the idealized competitive model in an attempt to fit the evidence better.

In one sense, contracts provide the foundation for a large part of economic analysis. Any trade – as a quid pro quo – must be mediated by some form of contract, whether it be explicit or implicit. In the case of spot trades, however, where the two sides of the transaction occur almost simultaneously, the contractual element is usually down-played, presumably because it is regarded as trivial (although we will argue below that this need not be the case). In recent years, economists have become much more interested in long-term relationships where a considerable amount of time may elapse between the quid and the quo. In these circumstances, a contract becomes an essential part of the trading relationship.

Research on contracts has progressed along several different lines. Two prominent areas of work are principal-agent theory and implicit labour contract theory. In these literatures, the focus is on risk-sharing or income-smoothing as the motivation for a contract; that is, on the gains the parties receive from transferring income from one state of the world or one period to another. For example, in implicit contract theory, it is supposed that workers are constrained in their ability to get insurance or to borrow on the open market and that employers therefore offer these services as part of an employment contract.

While ‘income-smoothing’ is undoubtedly important, there are arguably more fundamental factors underlying the existence of long-term contracts. A basic reason for long-term relationships is the existence of investments which are to some extent party specific; that is, once made, they have a higher value inside the relationship than outside. Given this ‘lock-in’ effect, each party will have some monopoly power *ex-post*, although there may be plenty of competition *ex-ante*, before investments are sunk. Since the parties cannot rely on the market once their relationship is underway, a long-term contract is an important way for them to regulate, and divide up the gains from, their trade. This will be the case even if the parties are risk neutral and have access to perfect capital markets, that is, even if the income-smoothing role is completely inessential. Moreover, in the case, say, of supply contracts involving large firms, risk neutrality and perfect capital markets may be reasonable approximations in view of the many outside insurance and borrowing/lending opportunities available to such parties.

In spite of their importance, contracts whose *raison d’être* is the regulation of specific relationships have been the subject of little analysis. A notable early reference is Becker's (1964) analysis of worker training. More recently, Williamson (1985) and Klein et al. (1978) have emphasized the difficulty of writing contracts which induce efficient relationship-specific investments as an important factor in explaining vertical integration.

In this entry I will try to summarize what is known theoretically about contracts of this type. I will focus particularly on the problems which arise when the parties write a contract which is incomplete in some respects. Given the rudimentary state of our knowledge of the area, the entry is inevitably quite speculative in nature. The reader who is interested in an elaboration of some of the ideas presented here, and how they fit into the rest of contract theory, might want to consult Hart and Holmstrom (1987).

1 The benefits of writing long-term contracts given relationship-specific investments

The role of a long-term contract when there are relationship-specific investments can be seen from the following example (based on Grout, 1984). Let B, S be, respectively, the buyer and seller of (one unit of) an input. Suppose that in order to realize the benefits of the input, B must make an investment, *a*, which is specific to S; for example, B might have to build a plant next to S. Assume that there are just two periods; the investment is made at date 0, while the input is supplied and the benefits are received at date 1. S's supply cost at date 1 is

c, while B's benefit function is $b(a)$ (all costs and benefits are measured in date 1 dollars).

If no long-term contract is written at date 0, the parties will determine the terms of trade from scratch at date 1. If we assume that neither party has alternative trading partners at date 1, there is, given B's sunk investment cost a, a surplus of $b(a)–c$ to be divided up. A simple assumption to make is that the parties split this 50:50 (this is the Nash bargaining solution). That is, the input price p will satisfy $b(a)–p=p–c$. This means that the buyer's overall payoff, net of his investment cost, is

$$b(a) - p - a = \frac{b(a) - c}{2} - a. \quad (1)$$

The buyer, anticipating this payoff, will choose a to maximize (1), i.e. to maximize $1/2\ b(a)–a$.

This is to be contrasted with the efficient outcome, where a is chosen to maximize total surplus, $b(a)–c–a$. Maximizing (1) will lead to underinvestment; in fact, in extreme cases, a will equal zero and trade will not occur at all. The inefficiency arises because the buyer does not receive the full return from his investment – some of this return is appropriated by the seller in the date 1 bargaining. Note that an upfront payment from S to B at date 0 (to compensate for the share of the surplus S will later receive) will not help here, since it will only change B's objective function by a constant (it is like a lump-sum transfer). That is, it redistributes income without affecting real decisions.

Efficiency can be achieved if a long-term contract is written at date 0 specifying the input price p^* in advance. Then B will maximize $b(a)–p^*–a$, yielding the efficient investment level, a^* An alternative method is to specify that the buyer must choose $a=a^*$ (if not he pays large damages to S) – the choice of p can then be left until date I, with an upfront payment by S being used to compensate B for his investment. The second method presupposes that investment decisions are publicly observable, and so in practice may be more complicated than the first (see below).

We see then that a long-term contract can be useful in encouraging relationship-specific investments. The word ‘investment’ should be interpreted broadly here; the same factors will apply whenever one party is forced to pass up an opportunity as a result of a relationship with another party (e.g., A's ‘investment’ in the relationship with B may be not to lock into C). That is, the crucial element is a sunk cost (direct or opportunity) of some sort (an effort decision is one example of a sunk cost). Note that the income-transfer motive for a long-term contract is completely absent here; there is no uncertainty and everything is in present value terms.

Given the advantages of long-term contracts in specific relationships, the question that obviously arises is why we do not see more of them, and why those we do see seem often to be limited in scope. To this question we now turn.

2 The costs of writing long-term contracts

Contract theory is sometimes dismissed because ‘we don't see the long-term contingent contracts that the theory predicts’. In fact, there is no shortage of complex long-term contracts in the world. Joskow (1985), for example, in his recent study of transactions between electricity generating plants and mine-mouth coal suppliers finds that some contracts between the parties extend for fifty years, and a large majority for over ten years. The contractual terms include quality provisions, formulae linking coal prices to costs and prices of substitutes, indexation clauses, and so on. The contracts are both complicated and sophisticated. Similar findings are contained in Goldberg and Erickson's (1982) study of petroleum coke.

At a much more basic level, a typical contract for personal insurance, with its many conditions and exemption clauses, is not exactly a simple document. Nor for that matter is a typical house rental agreement. On the other hand, labour contracts are often surprisingly rudimentary, at least in certain respects (for example, there is little indexation of wages to retail prices or to firm employment or sales; layoff pay is limited, etc.).

Given that complex long-term contracts are found in some situations but not others, it is natural to explain any observed contract as an outcome of an optimization process in which the relative benefits and costs of additional length and complexity are traded off at the margin. In the last section, we indicated some of the benefits of a long-term contract. (The example considered was sufficiently straightforward that the ideal long-term contract was a simple noncontingent one; however, with the inclusion of such factors as uncertainty about payoffs and variable quality of the input, the optimal contract would be a (possibly much more complex) contingent one.) But what about the costs? These are much harder to pin down since they fall under the general heading of ‘transaction costs’, a notoriously vague and slippery category. Of these, the following seem to be important: (1) the cost to each party of anticipating the various eventualities that may occur during the life of the relationship: (2) the cost of deciding, and reaching an agreement about, how to deal with such eventualities; (3) the cost of writing the contract in a sufficiently clear and unambiguous way that the terms of the contract can be enforced; and (4) the legal cost of enforcement.

One point to note is that *all* these costs are present also in the case of *short-term* contracts, although presumably they are usually smaller. In particular, since the short-term future is

more predictable, the first cost is likely to be much reduced, and so possibly is the third. However, it certainly is not the case that there is a sharp division between short-term contracts and long-term contracts, with, as is sometimes supposed, the former being costless and the latter being infinitely costly.

It is also worth emphasizing that, when we talk about the cost of a long-term contract, we are presumably referring to the cost of a 'good' long-term contract. There is rarely significant cost or difficulty in writing *some* long-term contract. For example, the parties to an input supply contract could agree on a fixed price and level of supply for the next fifty years. They do not presumably because such a rigid arrangement would be very inefficient. (In some cases the courts will not enforce such an agreement, taking the point of view that the parties could not really have intended it to apply unchanged for such a long time. A clause to the effect that the parties really do mean what they say should be enough to overcome this difficulty, however. In other cases, it may be impossible to write a binding long-term contract because the identities of some of the parties involved may change. For example, one party may be a government that is in office for a fixed period, and it may be impossible for it to bind its successors. This latter idea underlies the work of Kydland and Prescott (1977) and Freixas et al. (1985).)

Due to the presence of transaction costs, the contracts people write will be *incomplete* in important respects. The parties will quite rationally leave out many contingencies, taking the point of view that it is better to 'wait and see what happens' than to try to cover a large number of individually unlikely eventualities. Less rationally, the parties will leave out other contingencies that they simply do not anticipate. Instead of writing very long-term contracts the parties will write limited term contracts, with the intention of renegotiating these when they come to an end. (A paper which explores the implications of this is Crawford, 1986.) Contracts will often contain clauses which are vague or ambiguous, sometimes fatally so.

Anyone familiar with the legal literature on contracts will be aware that almost every contractual dispute that comes before the courts concerns a matter of incompleteness. In fact, incompleteness is probably at least as important empirically as asymmetric information as an explanation for departures from 'ideal' Arrow–Debreu contingent contracts. In spite of this, relatively little work has been done on this topic, the reason presumably being that an analysis of transaction costs is so complicated. One problem is that the first two transaction costs referred to above are intimately connected to the idea of bounded rationality (as in Simon, 1982), a successful formalization of which does not yet exist. As a result, perhaps, the few attempts that have been made to analyse incompleteness have concentrated on the third cost, the cost of writing the contract.

One approach, due to Dye (1985), can be described as follows. Suppose that the amount of input, q, traded between a buyer and seller should be a function of the product price, p, faced by the buyer: $q=f(p)$. Writing down this function is likely to be costly. Dye measures the costs in terms of how many different values q takes on as p varies; in particular, if • $\{q|\ q=f(p)$ for some $p\}=n$, the cost of the contract is $(n-1)c$, where $c>0$. This means that a noncontingent statement '$q=5$ for all p' has zero cost, the statement '$q=5$ for $p\leq 8$, $q=10$ for $p>8$' has cost c, and so on.

The costs Dye is trying to capture are real enough, but the measure used has some drawbacks. It implies for example, that the statement '$q=p^{1/2}$ for all p' has infinite cost if p has infinite domain, and does not distinguish between the cost of a simple function like this and the cost of a much more complicated function. As another example, a simple indexation clause to the effect that the real wage should be constant (i.e. the money wage= λp for some λ) would never be observed since, according to Dye's measure, it too has infinite cost. In addition, the approach does not tell us how to assess the cost of indirect ways of making q contingent; for example, the contract could specify that the buyer, having observed p, can choose any amount of input q he likes, subject to paying the seller σ for each unit.

There is another way of getting at the cost of including contingent statements. This is to suppose that what is costly is describing the state of the world ω rather than writing a statement per se. That is, suppose that ω cannot be represented simply by a product price, but is very complex and of high dimension – e.g., it includes the state of demand, what other firms in the industry are doing, the state of technology, etc. Many of these components may be quite nebulous. To describe the state *ex-ante* in sufficient detail that an outsider, e.g. The courts, can verify whether a particular state $\omega = \widehat{\omega}$ has occurred, and so enforce the contract, may be prohibitively costly. Under these conditions, the contract will have to omit some (in extreme cases, all) references to the underlying state.

Similar to this is the case where what is costly is describing the *characteristics* of what is traded or the *actions* (e.g. investments) the parties must take. For example, suppose that there is only one state of the world, but that q now represents the quality of the item traded rather than the quantity. An ideal contract would give a precise description of q. However, quality may be multidimensional and very difficult to describe unambiguously (and vague statements to the effect that quality should be 'good' may be almost meaningless). The result may be that the contract will have to be silent on many aspects of quality and/or actions.

Models of this sort of incompleteness have been investigated by Grossman and Hart (1987) and Hart and Moore (1985) for the case where the state of the world cannot be described and by Bull (1985) and Grossman and Hart (1986, 1987) for the case where quality and/or actions cannot be specified. These models do not rely on any asymmetry of information between the parties. Both parties may recognize that the state of the world is such that the buyer's benefit is high or the seller's cost is low, or that the quality of an item is good or bad or that an investment decision is appropriate or not. The difficulty is conveying this information to others. *That is, it is the asymmetry of information between the parties on the one hand, and outsiders, such as the courts, on the other, which is the root of the problem.*

To use the jargon, incompleteness arises because states of the world, quality and actions are *observable* (to the contractual parties) but not *verifiable* (to outsiders).

We describe an example of an incomplete contract along these lines in the next section.

3 Incomplete contracts: an example

We will give an example of an incomplete contract for the case where it is prohibitively costly to specify the quality characteristics of the item to be exchanged or the parties' investment decisions. Similar problems arise when the state of the world cannot be described. The example is a variant of the models in Grossman and Hart (1986, 1987), Hart and Moore (1985).

Consider a buyer B who wishes to purchase a unit of input from a seller S. B and S each make a (simultaneous) specific investment at date 0 and trade occurs at date 1. Let I_B, I_S denote, respectively, the investments of B and S, and to simplify assume that each can take on only two values, H or L (high or low). These investments are observable to B and S, but are not verifiable (they are complex and multidimensional, or represent effort decisions) and hence are noncontractible. We assume that at date 1 the seller can supply either 'satisfactory' input or 'unsatisfactory' input. 'Unsatisfactory' input has zero benefit for the buyer and zero cost for the seller (so it is like not supplying at all). 'Satisfactory' input yields benefits and costs which depend on *ex-ante* investments. These are indicated in Figure 1.

Figure 1

	$I_B = H$	$I_B = L$
$I_s = H$	(10, 6)	(9, 7)
$I_s = L$	(9, 7)	(6, 10)

The first component refers to the buyer's benefit, v, and the second to the seller's cost, c. So when $I_S=H$, $I_B=H$, $v=10$ and $c=6$ (if input is 'satisfactory'). From these gross benefits and costs must be subtracted investment costs, which we assume to be 1.9 if investment is high and zero if it is low (for each party). (All benefits and costs are in date 1 dollars.) Note that there is no uncertainty and so attitudes to risk are irrelevant.

Our assumption is that the characteristics of the input (e.g. whether it is 'satisfactory') are observable to both parties, but are too complicated to be specified in a contract. The fact that they are observable means that the buyer can be given the option to reject the input at date 1 if he does not like it. This will be important in what follows.

An important feature of the example is that the seller's investment affects not only the seller's costs but also the buyer's benefit and the buyer's investment affects not only the buyer's benefit but also the seller's costs. The idea here is that a better investment by the seller increases the quality of 'satisfactory' input; and a better investment by the buyer reduces the cost of producing 'satisfactory' input, that is input that can be used by the buyer.

For instance, one can imagine that B is an electricity generating plant and *S* a coal mine that the plant is sited next to. I_B might refer to the type of coal-burning boiler that the plant installs and I_S to the way the coal supplier develops the mine. By investing in a better boiler, the power plant may be able to burn lower quality coal, thus reducing the seller's costs, while still increasing its gross (of investment) profit. On the other hand, by developing a good seam, the coal supplier may raise the quality of coal supplied while reducing its variable cost.

The first-best has $I_B=I_S=H$, with total surplus equal to (10–6)–3.8=0.2 (if $I_B=H$ and $I_S=L$, or vice versa, surplus=0.1 and if $I_B=I_S=L$, no trade occurs and surplus is zero). This could be

achieved if *either* investment *or* quality were contractible as follows. If investment is contractible, an optimal contract would specify that the buyer must set $I_B=H$ and the seller $I_S=H$ and give the buyer the right to accept the input at date 1 at price p_1 or reject it at price p_0. If $10>p_1-p_0>6$, the seller will be induced to supply satisfactory input (the gain, p_1-p_0, from having the input accepted exceeds the seller's supply cost) and the buyer to accept it (the buyer's benefit exceeds the increment price p_1-p_0. If, on the other hand, quality is contractible, the contract could specify that the seller must supply input with the precise characteristics which make it satisfactory when $I_B=I_S=H$. Each party would then have the socially correct investment incentives since, with specific performance, neither party's investment affects the other's payoff (there is no externality).

We now show that the first-best cannot be achieved if investment and quality are both noncontractible. A second-best contract can make price a function of any variable that is verifiable. Investment and quality are not verifiable (nor is v or c), but we shall suppose that whether the item is accepted or rejected by the buyer is, so the contract can specify an acceptance price, p_1, and a rejection price, p_0. In fact, p_0, p_1 can also be made functions of (verifiable) messages that the buyer and seller send each other, reflecting the investment decisions that both have made (as in Hart and Moore, 1985). The following argument is unaffected by such messages and so, for simplicity, we ignore them (the interested reader is referred to Hart and Holmstrom, 1987).

Can we sustain the first-best by an appropriate choice of p_0, p_1? The seller always has the option of choosing $I_S=L$ and producing an item of unsatisfactory quality, which yields him a net payoff of p_0. In order to induce him not to do this, we must have

$$p_1 - 6 - 1.9 \geq p_0, \text{ i.e. } p_1 - p_0 \geq 7.9. \tag{2}$$

Similarly the buyer's net payoff must be no less than $-p_0$ since he always has the option of choosing $I_B=L$ and rejecting the input. That is,

$$10 - p_1 - 1.9 \geq -p_0, \text{ i.e. } p_1 - p_0 \geq 8.1. \tag{3}$$

So $(p-p_0)$ must lie between 7.9 and 8.1.

Now the seller has an additional option. If he expects the buyer to set $I_B=H$, he can choose $I_S=L$ and, given that $8.1\geq p_1-p_0\geq 7.9$, still be confident that trade of ‘satisfactory’ input will occur under the original contract at date 1 (the buyer will accept satisfactory input since $v=9>p_1-p_0$, while the seller will supply it since $p_1-p_0>7=c$). But if the seller deviates, his payoff rises from $p_1-6-1.9$ to p_1-7. (The example is symmetric and so a similar deviation is also profitable for the buyer.) Hence the $I_B=I_S=H$ equilibrium will be disrupted.

We see, then, that the first-best cannot be sustained if investment and quality are both noncontractible. The reason is that it will be in the interest of the seller (or the buyer) to reduce investment since, although this reduces social benefit by lowering the buyer's (or seller's) benefit, it increases the seller's (or buyer's) own profit. The optimal second-best contract will instead have $I_B=H$, $I_S=L$ (or vice versa), which will be sustained by a pair of prices p_0, p_1 such that $9>p_1-p_0>7$. Total surplus will be 0.1 instead of the first-best level of 0.2. (Note the importance of the assumption that both the buyer and seller can choose $I=H$ or L. If only the buyer (or the seller) can choose $I=L$, the first-best can be achieved by choosing p_1-p_0 between 6 and 7 (or 9 and 10): any deviation by the buyer (or the seller) will then be unprofitable since it will lead to no trade.)

The conclusion is that inefficiencies can arise in incomplete contracts even though the parties have common information (both observe investments and both observe quality). The particular inefficiency that occurs in the model analysed is in *ex-ante* investments. *Ex-post* trade is always efficient relative to these investments since p_1, p_0 can and will be chosen such that $v>p_1-p_0>c$, i.e. the seller wants to supply and the buyer to receive satisfactory input. The example can be regarded as formalizing the intuition of Williamson (1985) and Klein et al. (1978) that relationship-specific investments will be distorted due to the impossibility of writing complete contingent contracts – note that this result is achieved without imposing arbitrary restrictions on the form of the permissible contract (e.g. we have not ruled out the existence of long-term contracts from the start). (There is one exception to this statement – we have excluded the participation of a third party to the contract; for a discussion and justification of this, see Hart and Holmstrom, 1987.)

The example may be used to illustrate a theory of ownership presented in Grossman and Hart (1986, 1987). It is sometimes suggested that when transaction costs prevent the writing

of a complete contract, there may be a reason for firm integration (see Williamson, 1985). Consider the payoffs of Figure 1 and suppose that B takes over S. The control that B thereby gains over S's assets may allow B to affect S's costs in various ways, and this may reduce the possibility of opportunistic behaviour by S. To take a very simple (and contrived) example, suppose that if S chooses $I_S=L$, B can take some action, α with respect to S's assets at date 1 so as to make S's cost of supplying either satisfactory or unsatisfactory input equal to 9 (in the coal-electricity example, α might refer to the part of the mine's seam the coal is taken out of; note that we now drop the assumption that the cost of supplying unsatisfactory input is zero). Imagine furthermore that this action increases B's benefit, so that B will indeed take it at date 1 if S chooses L. Then with this extra degree of freedom, the first-best can be achieved. In particular, if $p_1=p_0+6.1$, $I_S=I_L=H$ is a Nash equilibrium since, by the above reasoning, any deviation by the seller will be punished, while if the buyer deviates, the seller will supply unsatisfactory input given that $p_1<p_0+7$.

Note that if action α could be specified in the initial contract, there would be no need for integration: the initial contract would simply say that B has the right to choose α at date 1. Ownership becomes important, however, if (i) α is too complicated to be specified in the date 0 contract and therefore qualifies as a residual right of control; and (ii) residual rights of control over an asset are in the hands of whomever owns that asset. The point is that under incompleteness the allocation of residual decision rights matters since the contract cannot specify precisely what each party's obligations are in every state of the world. To the extent that ownership of an asset guarantees residual rights of control over that asset, vertical and lateral integration can be seen as ways of ensuring particular – and presumably efficient – allocations of residual decision rights. (While in the above example, integration increases efficiency, this is in no way a general conclusion. In Grossman and Hart (1986, 1987), examples are presented where integration reduces efficiency.)

Before concluding this section, we should emphasize that for reasons of tractability we have confined our attention to incompleteness due to a very particular sort of transaction cost. In practice, some of the other transactions costs we have alluded to are likely to be at least as important, if not more so. For example, in the type of model we have analysed, although the parties cannot describe the state of the world or quality characteristics, they are still supposed to be able to write a contract which is unambiguous and which anticipates all eventualities. This is very unrealistic. In practice, a contract might, say, have B agreeing to rent S's concert hall for a particular price. But suppose S's hall then burns down. The contract will usually be silent about what is meant to happen under these conditions (there is no hall to rent, but should S pay B damages and if so how much?), and so, in the event of a dispute, the courts will have to fill in the ‘missing provision’. (A situation where it becomes impossible or extremely costly to supply a contracted for good is known as one of ‘impossibility’ or ‘frustration’ in the legal literature.) An analysis of this sort of incompleteness, although extremely hard, is a very important topic for future research. It is likely to yield a much richer and more realistic view of the way contracts are written and throw light on how courts should assess damages (this latter issue has begun to be analysed in the law and economics literature; see, e.g., Shavell, 1980).

4 Self-enforcing contracts

The previous discussion has been concerned with explicit binding contracts that are enforced by outsiders, such as the courts. Even the most casual empiricism tells us that many agreements are not of this type. Although the courts may be there as a last resort (the shadow of the law may therefore be important), these agreements are enforced on a day to day basis by custom, good faith, reputation, etc. Even in the case of a serious dispute, the parties may take great pains to resolve matters themselves rather than go to court. This leads to the notion of a self-enforcing or implicit contract (the importance of informal arrangements like this in business has been stressed by Macaulay (1963) and Ben-Porath (1980) among others).

People often by-pass the legal process presumably because of the transaction costs of using it. The costs of writing a ‘good’ long-term contract discussed in Section 2 are relevant here. So also is the skill with which the courts resolve contractual disputes. If contracts are incomplete and contain missing provisions as well as vague and ambiguous statements, appropriate enforcement may require abilities and knowledge (what was in the parties’ minds?) that many judges and juries do not possess. This means that going to court may be a considerable gamble – and an expensive one at that. (This is an example of the fourth transaction cost noted in Section 2.)

Although the notion of implicit or self-enforcing contracts is often invoked, a formal study of such agreements has begun only recently (see, e.g. Bull, 1985), with a considerable stimulus coming from the theory of repeated games. This literature has stressed the role of *reputation* in ‘completing’ a contract. That is, the idea is that a party may behave ‘reasonably’ even if he is not obliged to do so in order to develop a reputation as a decent and reliable trader. In some instances such reputational effects will operate only within the group of contractual parties – this is sometimes called *internal* enforcement of the contract – while in others the effects will be more pervasive. The latter will be the case when some outsiders to the contract, for example other firms in the industry or potential workers for a firm, observe unreasonable behaviour by one party, and as a result are more reluctant to deal with it in the future. In this case the enforcement is said to be *external* or *market-based*. Note that there may be a tension between this external enforcement and the reasons for the absence of a legally binding contract in the first place – the more people can observe the behaviour, the more likely it is to be verifiable.

The distinction between an incomplete contract and a standard asymmetric information contract should be emphasized here. It is the former that allows reputation to operate since the parties have the same information and can observe whether reasonable behaviour is being maintained. In the latter case, it is unclear how reputation can overcome the asymmetry of information between the parties that is the reason for the departure from an Arrow–Debreu contract.

The role of reputation in sustaining a contract can be illustrated using the following model (based on Bull (1985) and Kreps (1984); this is an even simpler model of incomplete

contracts than that of the last section). Assume that a buyer, B, and a seller, S, wish to trade an item at date 1 which has value v to the buyer and cost c to the seller, where $v>c$. There are no *ex-ante* investments and the good is homogeneous, so quality is not an issue. Suppose, however, that it is not verifiable whether trade actually occurs. Then a legally binding contract which specifies that the seller must deliver the item and the buyer must pay p, where $v>p>c$, cannot be enforced. The reason is that, assuming (as we shall) that simultaneous delivery and payment are infeasible, if the seller has to deliver first, the buyer can always deny that delivery occurred and refuse payment, while if the buyer has to pay first, the seller can always claim later that he did deliver even though he did not. As a result, if the parties must rely on the courts, a gainful trading opportunity will be missed.

The idea that not even the level of trade is verifiable is extreme, and Bull (1985) in fact makes the more defensible assumption that it is the quality of the good that cannot be verified (in Bull's model, S is a worker and quality refers to his performance). Bull supposes that quality is observable to the buyer only with a lag, so that take it or leave it offers of the type considered in the last section are not feasible. As a result the seller always has an incentive to produce minimum quality (which corresponds in the above model to zero output). Making quantity nonverifiable is a cruder but simpler way of capturing the same idea (this is the approach taken in Kreps, 1984).

Note that in the above model incompleteness of the contract arises entirely from transaction cost (3), the difficulty of writing and enforcing the contract.

To introduce reputational effects one supposes that this trading relationship is repeated. Bull (1985) and Kreps (1984) follow the supergame literature and assume infinite repetition in order to avoid unravelling problems. This approach, as is well known, suffers from a number of difficulties. First, the assumption of infinite (or in some versions, *potentially* infinite) life is hard to swallow. Secondly, ‘reasonable’ behaviour, i.e. trade, is sustained by the threat that if one party behaves unreasonably so will the other party from then on. While this threat is ‘credible’ (more precisely, subgame perfect), it is unclear why the parties could not decide to continue to trade after a deviation, i.e. to ‘let bygones be bygones’ (see Farrell, 1984.)

It would seem that a preferable approach is to assume that the relationship has finite length, but introduce asymmetric information, as in Kreps and Wilson (1982) and Milgrom and Roberts (1982). The following is based on some very preliminary work that Bengt Holmstrom and I have undertaken along these lines.

Suppose that there are two types of buyers in the population, honest and dishonest. Honest buyers will always honour any agreement or promise that they have made while dishonest ones will do so only if this is profitable. A buyer knows his own type, but others do not. It is common knowledge that the fraction of honest buyers in the population is π, $0<\pi<1$. In contrast, all sellers are known to be dishonest. All agents are risk neutral.

Assume for simplicity that a single buyer and seller are matched at date 0 with neither having any alternative trading partners at this date or in the future. Consider first the one-period case. Then a date 0 agreement can be represented as follows. The interpretation is that the buyer promises to pay the seller p_1 before date 1 (stage I); in return, the seller promises to supply the item at date 1 (stage II); and in return for this, the buyer promises to make a further payment of p_2 (stage III).

We should mention one further assumption. Honest buyers, although they never breach an agreement first, are supposed to feel under no obligation to fulfil the terms of an agreement that has already been broken by a seller (interestingly, although this is a theory of buyer psychology, it has parallels in the common law). Note that if a buyer ever breaks an agreement first, he reveals himself to be dishonest, with the consequence that no further self-enforcing agreement with the seller is possible and hence trade ceases.

What is an optimal agreement? Consider Figure 2. The seller knows that he will receive p_2 only with probability π since a dishonest buyer will default at the last stage. Since the seller is himself dishonest, he will supply at Stage II only if it is profitable for him to do so, i.e. only if

$$\pi p_2 - c \geq 0. \tag{4}$$

Assume for simplicity that the seller has all the bargaining power at date 0 (nothing that follows depends on this). Then the seller will wish to maximize his overall payoff

$$p_1 + \pi p_2 - c, \tag{5}$$

subject to (4) which makes it credible that he will supply at stage II and also the constraint that he does not discourage an honest buyer from participating in the agreement at date 0. Since with (4) satisfied, buyers know that they will receive the item for sure, this last condition is

$$v - p_1 - p_2 \geq 0. \tag{6}$$

Note that a dishonest buyer's payoff $v–p_1$ is always higher than an honest buyer's payoff given in (6), so there is no way to screen out dishonest buyers. In the language of asymmetric information models, the equilibrium is a pooling one.

Figure 2

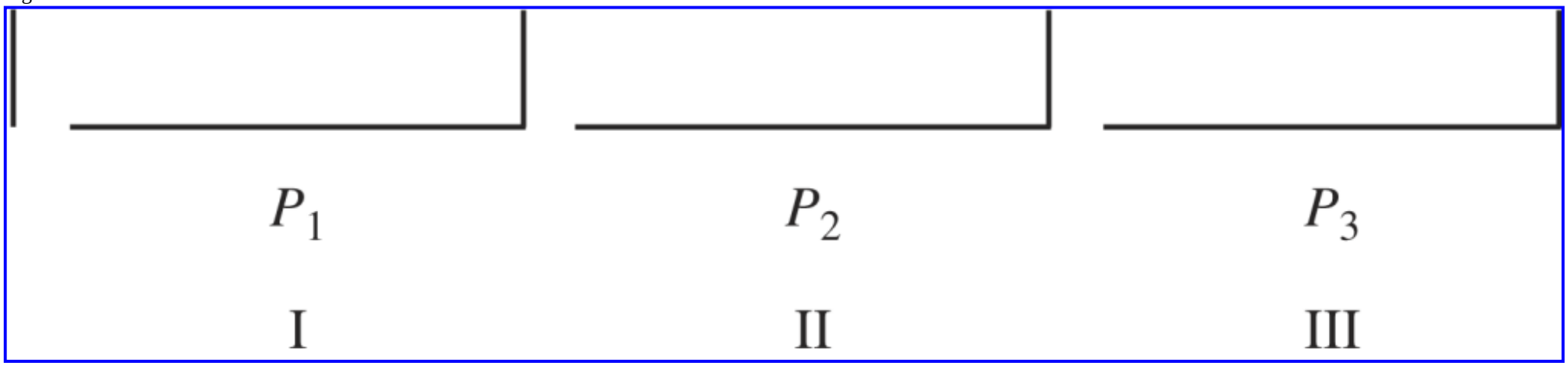

Since the seller's payoff is increasing in p_1, (6) will hold with equality (the buyer gets no surplus). (More generally, changes in p_1 simply redistribute surplus between the two parties without changing either's incentive to breach.) If we substitute for p_1 in (5), the seller's payoff becomes $v–p_2(1–\pi)–c$, which, when maximized subject to (4), yields the solution $p_2=c/\pi$. The maximized net payoff is

$$v - c/\pi, \tag{7}$$

which is less than the first-best level, $v–c$.

We see then that the conditions for trade are more stringent in the absence of a binding contract. If $c/(\pi)>v>c$, there are gains from trade which would not be realized in a one-period relationship.

Suppose now that the relationship is repeated. Consider a two-period version of the above and assume no discounting. Now the diagram shown in Figure 3 applies. That is, the agreement says that the buyer pays, the seller supplies the first time, the buyer pays more, the seller supplies a second time, and the buyer makes a final payment. Rather than solving for the optimal arrangement, we shall simply show that the seller can do better than in the one period case. Let $p_3=c/\pi$, $p_2=c$ and $p_1=2v–c–(c)/\pi$. Then (i) the seller will supply at Stage IV (if matters have got that far), knowing that he will receive p_3 with probability π (ii) both honest and dishonest buyers will pay p_2 at Stage III, the latter because, at a cost of c, they thereby ensure supply worth $v>c$ at Stage IV; (iii) the seller will supply at stage II because this gives him a net payoff of $p_2+\pi p_3–2c\geq 0$, while if he does not the arrangement is over and his payoff is zero; (iv) an honest buyer is prepared to participate since his surplus is non-negative (actually zero).

Figure 3

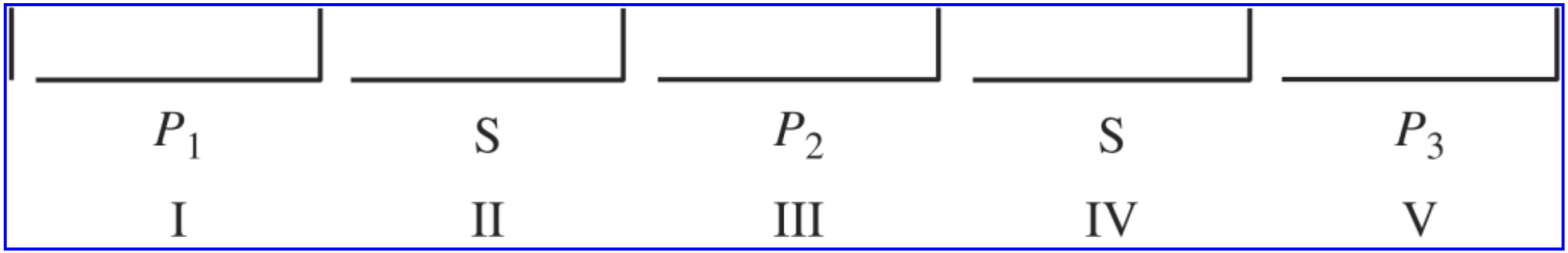

The seller's overall expected net payoff is

$$p_1 + \pi_2 + \pi p_3 - 2c = 2v - c - c/\pi, \tag{8}$$

which *exceeds* twice the one-period payoff. Hence trade is more likely to take place in a two-period relationship than in a one-period one. In fact it can be shown that the above is an *optimal* two-period agreement.

Repetition improves things by allowing the honest buyer to pay less second time round (Stage III) than third time round (Stage V). That is, the arrangement *back-loads* payments. This is acceptable to the seller because he knows that even a dishonest buyer will not default at Stage III since he has a large stake in the arrangement continuing. To put it another way, the dishonest buyer does not want to reveal his dishonesty at too early a stage.

The same arrangement can be used when there are more than two periods: the buyer promises to pay c at every stage except the last, when he pays (c/π). In fact the per period surplus of the seller from such an arrangement converges to the first-best level $(v–c)$ as the number of periods tends to ∞ (assuming no discounting, of course).

Although the above analysis is extremely provisional and sketchy, we can draw some tentative conclusions about the role of reputation and indicate some directions for further research. First, the notion of a psychic cost of breaking an agreement seems to be a useful – as well as a not unrealistic – basis for a theory of self-enforcing contracts. It is obviously desirable to drop the assumption that some agents are completely honest and others completely dishonest, and assume instead that the typical trader has a finite psychic cost of breaking an agreement, where this cost is distributed in the population in a known way. In other words, everybody ‘has their price’, but this price varies. Preliminary work along these lines suggests that the above results generalize; in particular, repetition makes it easier to sustain a self-enforcing agreement.

Of course, asymmetries of information about psychic costs are not the only possible basis for a theory of reputation. For example, the buyer and seller could have private information about v and c, and might choose their trading strategies to influence perceptions about the values of these variables (the role of uncertainty about v and c in determining reputation has been investigated by Thomas and Worrall, 1984). A theory of self-enforcing contracts should ideally generate results which are not that sensitive to where the asymmetry of information is placed. The work of Fudenberg and Maskin (1986) in a related context, however, suggests that this may be a difficult goal to achieve.

There are a number of other natural directions in which to take the model. One is to introduce trade with other parties. For example, the seller may trade with a succession of buyers rather than a single one. The extent to which repetition increases per period surplus in this case depends on whether new buyers observe the past broken promises of the seller. (This determines the degree to which external enforcement operates; more generally, ‘a new buyer may observe that default occurred in the past, but be unsure about who was responsible for it.) If new buyers do not observe past broken promises, repetition achieves nothing, which gives a very strong prediction of the possible benefits of a long-term relationship between a fixed buyer and seller. Even if past broken promises are observed perfectly, it appears that, *ceteris paribus*, a single long-term agreement may be superior to a succession of short-term ones. The reason is that in the latter case the constraint is imposed that each party must receive non-negative surplus over *their* term of the relationship whereas in the former case there is only the single constraint that surplus must be non-negative over the whole term (see Bull, 1985; Kreps, 1984).

Probably the most important extension is to introduce incompleteness due to other sorts of transaction costs, e.g. the ‘bounded rationality’ costs (1) and (2) discussed in Section 2. The problem is that the same factors which make it difficult to anticipate and plan for eventualities in a formal contract apply also to informal arrangements. That is, an informal arrangement is also likely to contain many ‘missing provisions’. But then the question arises, what constitutes ‘reasonable’ or ‘desirable’ behaviour (in terms of building a reputation) with regard to states or actions that were not discussed *ex-ante*? Custom, among other things, is likely to be important under these conditions: behaviour will be ‘reasonable’ or ‘desirable’ to the extent that it is generally regarded as such (for a good discussion of this, see Kreps, 1984). This raises many new and interesting (as well as extremely difficult) questions.

5 Summary and conclusions

The vast majority of the theoretical work on contracts to date has been concerned with what might be called ‘complete’ contracts. In this context, a complete contract means one that specifies each party's obligations in every conceivable eventuality, rather than a contract that is fully contingent in the Arrow–Debreu sense. In particular, according to this terminology, the typical asymmetric information contract found in the principal-agent or implicit contract literatures (see Hart and Holmstrom, 1987) is complete.

In reality it is usually impossible to lay down each party's obligations completely and unambiguously in advance, and so most actual contracts are seriously incomplete. In this entry, we have tried to indicate some of the implications of such incompleteness. Among other things, we have seen that incompleteness can lead to departures from the first-best even when there are no asymmetries of information among the contracting parties (and, moreover, the parties are risk neutral).

More important perhaps than this is the fact that incompleteness raises new and difficult questions about how the behaviour of the contracting parties is determined. To the extent that incomplete contracts do not specify the parties’ actions fully, i.e. they contain ‘gaps’, additional theories are required to tell us how these gaps are filled in. Among other things, outside influences such as custom or reputation may become important under these conditions. In addition, outsiders, such as the courts (or arbitrators), may have a role to play in filling in missing provisions of the contract and resolving ambiguities rather than in simply enforcing an existing agreement. Incompleteness can also throw light on the importance of the allocation of decision rights or rights of control. If it is too costly to state precisely how a particular asset is to be used in every state of the world, it may be efficient simply to give one party ‘control’ of the asset, in the sense that he is entitled to do what he likes with it, subject perhaps to some explicit (contractible) limitations.

While the importance of incompleteness is very well recognized by lawyers, as well as by those working in law and economics, it is only beginning to be appreciated by economic theorists. It is to be hoped that work in the next few years will lead to significant advances in our formal understanding of this phenomenon. Unfortunately, progress is unlikely to be easy since many aspects of incompleteness are intimately connected to the notion of bounded rationality, a satisfactory formalization of which does not yet exist.

As a final illustration of the importance of incompleteness, consider the following question. Why do parties frequently write a limited term contract, with the intention of renegotiating this when it comes to an end, rather than writing a single contract that extends over the whole length of their relationship? In a complete contract framework such behaviour cannot be advantageous since the parties could just as well calculate what will happen when the contract expires and include this as part of the original contract. It is to be hoped that future work on incomplete contracts will allow this very basic question to be answered.

See Also

- adverse selection
- contract theory
- exchange
- implicit contracts
- moral hazard
- rationality, bounded

Bibliography

Becker, G. 1964. *Human Capital*. New York: Columbia University Press.

Ben-Porath, Y. 1980. The F-connection: families, friends, and firms and the organization of exchange. *Population and Development Review* 6(March), 1–30.

Bull, C. 1985. The existence of self-enforcing implicit contracts. C.V. Starr Center, New York University.

Crawford, V. 1986. Long-term relationships governed by short-term contracts. Princeton University.

Dye, R. 1985. Costly contract contingencies. *International Economic Review* 26(1), 233–50.

Freixas, X., Guesnerie, R. and Tirole, J. 1985. Planning under incomplete information and the ratchet effect. *Review of Economic Studies* 52(2), 169, 173–92.

Fudenberg, D. and Maskin, E. 1986. The Folk Theorem in repeated games with discounting and with incomplete information. *Econometrica* 54(3), 533–54.

Goldberg, V. and Erickson, J. 1982. Long-term contracts for petroleum coke. Department of Economics Working Paper Series No. 206, University of California, Davis, September.

Grossman, S. and Hart, O. 1986. The costs and benefits of ownership: a theory of vertical and lateral integration. *Journal of Political Economy* 94(4), August, 691–719.

Grossman, S. and Hart, O. 1987. Vertical integration and the distribution of property rights. In *Economic Policy in Theory and Practice*, Sapir Conference Volume, London: Macmillan Press.

Grout, P. 1984. Investment and wages in the absence of binding contracts: a Nash bargaining approach. *Econometrica* 52(2), March, 449–60.

Hart, O. and Holmstrom, B. 1987. The theory of contracts. In *Advances in Economic Theory, Fifth World Congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

Hart, O. and Moore, J. 1985. Incomplete contracts and renegotiation. London School of Economics, Working Paper.

Joskow, P. 1985. Vertical integration and long-term contracts. *Journal of Law, Economics and Organization* 1, Spring.

Klein, B., Crawford, R. and Alchian, A. 1978. Vertical integration, appropriable rents and the competitive contracting process. *Journal of Law and Economics* 21, 297–326.

Kreps, D. 1984. Corporate culture and economic theory. Mimeo, Stanford University, May.

Kreps, D. and Wilson, R. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27, 253–79.

Kydland, F. and Prescott, E. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85(3), 473–92.

Macaulay, S. 1963. Non-contractual relations in business: a preliminary study. *American Sociological Review* 28(February), 55–67.

Milgrom, P. and Roberts, D.J. 1982. Predation, reputation and entry deterrence. *Journal of Economic Theory* 27, 280–312.

Shavell, S. 1980. Damage measures for breach of contract. *Bell Journal of Economics* 11(2), Autumn, 466–90.

Simon, H. 1982. *Models of Bounded Rationality*. Cambridge, Mass.: MIT Press.

Thomas, J. and Worrall, T. 1984. Self-enforcing wage contracts. Mimeo, University of Cambridge.

Williamson, O. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

How to cite this article

Hart, Oliver. "incomplete contracts." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000041> doi:10.1057/9780230226203.0772

The New Palgrave Dictionary of Economics Online

incomplete markets

Charles Wilson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Incomplete markets’ describes a market structure in which there are effective constraints on which bundles of goods may be exchanged with each other. When incompleteness arises from markets that are sequentially segmented, some of the basic properties of general equilibrium are affected. First, equilibrium may not exist even under the usual regularity assumptions. Second, allocations may not be Pareto optimal, even after the limitations imposed by the market structure are taken into account. Third, if securities are denominated in nominal values, the equilibrium allocation is generally not locally unique.

Keywords

Arrow–Debreu markets; complete markets; demand functions; existence of equilibrium; incomplete markets; indeterminacy of equilibrium; local uniqueness of equilibrium; marginal rate of substitution; spot markets; uncertainty

Article

Incomplete markets arise when agents are unable to exchange every good either directly or indirectly with every other agent. In the case of a single market with no limitations on the exchange of goods, relatively mild assumptions guarantee the existence, Pareto optimality, and local uniqueness of a competitive equilibrium. However, once we impose restrictions on the trade of goods and introduce sequential markets so that not all trade take place in a single market, any one of these properties may fail to be satisfied. A large literature has evolved that examines the conditions under which different sets of securities generate a complete set of markets and the properties of the equilibrium allocations when they do not. In this article I illustrate a few of the main ideas in this literature.
A good starting point is the work of Arrow (1973), who demonstrated that static competitive analysis can be extended to deal with the case of uncertainty, but only by expanding the set of markets to include

a separate price for each commodity in each state of the world. With a complete set of markets in state contingent commodities, it follows immediately that, under the usual conditions on preferences, the competitive allocation exists, is *ex-ante* Pareto efficient and locally unique. Although this approach solves the problem of extending general equilibrium analysis to deal with uncertainty, it strains the credibility of the model by requiring an unrealistic number of goods to be simultaneously exchanged. It is important, therefore, to examine the extent to which the same allocation can be attained with a different market structure requiring a smaller number of instruments.

Consider an economy with S possible realizations of an uncertain state of the world with N goods in each state. We will refer to the state contingent good i in state s as good is. To allow for more general market structures, we suppose that all trading takes place in securities, which are claims on the vector of state contingent goods. A *simple* security promises delivery of one unit of only one state contingent good. Observe that any security may be represented as a linear combination of simple securities. The *span* of any set of securities is the set of state-contingent goods that can be obtained by some linear combination of those securities. A *market* is a set of securities and a price vector at which they may be exchanged. An Arrow–Debreu market is a market consisting of the complete set of simple securities, and an Arrow–Debreu allocation is the competitive equilibrium for an Arrow–Debreu market. A *spot* market for state s is a market in which only the simple securities for s goods are traded. We will always assume that the spot market for each state s is ‘complete’ in the sense that the set of feasible trades spans the set of all simple securities for state s.

To reduce the number of securities traded in any market, Arrow considers a two-stage market structure. In the first stage, before the state is realized, all agents have access to a ‘securities’ market. In this market, there is one security f^s for each state s, which represents a claim of one unit of each good in that state. In the second stage, after the state is realized and the claims of the first-stage securities are realized, the corresponding spot market opens and the final allocation is determined by the spot market equilibrium. Arrow demonstrates that, when agents have perfect foresight of the future spot market prices, any Arrow–Debreu allocation can be attained as a competitive equilibrium for this two-stage market structure. Since spot markets operate only when the actual state is realized, the total number of securities that are required to obtain the Arrow–Debreu allocation is reduced from NS to N+S.

To demonstrate the logic of Arrow's result, let p(s) denote the vector of spot prices in state s and let qs denote the first-stage price of security f s. Then, defining pis=qspi(s) for each good is, we obtain an NS vector (p_{is}) that defines the relative prices for all state contingent goods. For example, to exchange good is for good js', an agent exchanges security $f^{s'}$ for security f^s in the first-stage securities market and then uses the spot markets to obtain the desired net exchange. Alternatively, given a vector (p_{is}) of state-contingent prices, we may obtain the equivalent prices for the two-stage market by defining each $q_s = \Sigma_i p_{is}$ and each $p_i(s) = p_{is}/q_s$. Then, since each agent effectively faces the same budget constraint in both market structures, it follows that both market structures generate the same equilibrium allocation of goods.

Notice that the only role of the first-stage securities market is to transfer purchasing power across states. For instance, the set of simple securities of good 1 would work just as well. The essential requirement is that the set of securities spans the set of all possible transfers of purchasing power across states. Furthermore, so long as there is an ‘insurance’ security for each state s that delivers only state s contingent goods, the spanning condition is necessarily satisfied (at least if the vectors of all spot prices

are strictly positive). Other sets of securities may also satisfy the spanning condition. However, the consideration of a more general set of securities also introduces some complications that may impact on the existence of equilibrium and the welfare analysis of the market structure.

The problem is not just that the set of available securities might not span the entire commodity space. In fact, as long as all trade takes place in a single market so that any feasible security can be traded with another, the span of the market is fixed. So if we simply redefine the commodity space as the market span and restrict preferences accordingly, the existence, Pareto optimality, and local uniqueness of equilibria (for any basis of securities) follow from the standard arguments. With multi-stage markets, however, the security markets are essentially segmented. Consequently, a change in relative spot market prices may translate into changes in the space of feasible transfers of purchasing power that may be obtained by exchanging any given set of securities.

To illustrate, suppose there are only two goods in each state and the first-stage securities market consists of just two 'forward' securities, which respectively represent the claim of one unit of good X or one unit of good Y regardless of the state. Now fix the spot market prices in each state. Then since there are only two securities, it follows immediately that the dimension of the space of income transfers (measured, say, in terms of good X in each state) that can be obtained using the first-stage securities is at most two. Furthermore, if there are more than two states, the space of transfers that are spanned by the securities market depends on the relative prices in the different spot markets. For instance, suppose that the relative spot prices are the identical in states 1 and 2. Then any transfer of income to state 1 must be accompanied by the same transfer of income to state 2. However, if $p_x(1)/p_y(1)<1<p_x(2)/p_y(2)$, income can be transferred from state 1 to state 2 by exchanging one of forward security X for one unit of forward security Y. For the general case with N goods and S states, Townsend (1978) shows that when all first-stage securities are forward securities, the income transfers of these securities span R^S, the space of income transfers, if and only if there are at least S securities and the set of spot market price vectors are linearly independent.

The existence of equilibrium

When the dimension of the span of the transfers of a set of securities depend on the prices in the spot markets, the usual regularity assumptions on preferences no longer guarantee the existence of an equilibrium. Consider the following example based on Hart (1975). There are two agents, a and b, and two states. In each state there are two goods, labelled X and Y which must be consumed in non-negative amounts by each agent. The preferences and endowments of the agents are given in Table 1, where x_{as} and y_{as} are the respective amounts of good X and Y consumed by agent α in state s.

Agent	Endowments		Utility
	(X_1, Y_1)	(X_2, Y_2)	
a	(2, 2)	(1,1)	$3x_{a1}+y_{a1}+3x_{a2}+y_{a2}$
b	(1, 1)	(2, 2)	$x_{b1}+3y_{b1}+x_{b2}+3y_{b2}$

Agent a is endowed with two units of each good in state 1 and one unit of each good in state 2. His marginal rate of substitution between X and Y in either state is 3, and his marginal rate of substitution

between goods across states is 1. Agent *b* is endowed with one unit of each good in state 1 and two units of each good in state 2. His preferences are the same as those of agent *a* except that the role of *X* and *Y* is reversed. It is easy to check that in the unique Arrow–Debreu equilibrium, the price of all state contingent goods must be equal.

Consider next the case in which the first-stage market consists of the two forward securities. We will show that a competitive equilibrium does not exist. As above, let $p_x(s)$ and $p_y(s)$ denote the equilibrium prices of goods *X* and *Y* in the state *s* spot market, and let q_x and q_y denote the equilibrium prices of the two forward securities. Now suppose some agent α exchanges q_x units of security *Y* for q_y units of security *X*. Then his income in state *s* changes by the amount $p_{xs}q_y - p_{ys}q_x$. Therefore, in equilibrium either (*a*) q_x/q_y lies between $p_x(1)/p_y(1)$ and px(2)/py(2), or (b) qx/qy=px1/py1=px2/py2. Otherwise, one security dominates the other in the sense that an exchange of securities raises or lowers purchasing power in both states.

We show first that case (*b*) in which the relative spot prices are equal is not consistent with equilibrium. In this case, an exchange of securities leaves the income in both spot markets unchanged. Consequently, the equilibrium allocation and prices in the spot market must be the same as if no securities market existed. But the solution to either spot market then yields the allocation in which agent *a* obtains all three units of good *X* and agent *b* all three units of good *Y*. However, to clear the spot markets, the spot prices in the two states must differ, with $p_{x1}/p_{y1}=2$ and $p_{x2}/p_{y2}=1/2$. We conclude that the relative spot prices cannot be equal in equilibrium.

Now suppose the relative spot prices are different. Then, using both the securities market and the spot markets, an agent may exchange good *X* in state 1 for good *X* in state 2 at the relative price $(p_x(1)/p_y(1))$ $([p_x(2)q_y - p_y(2)q_x]/[p_y(1)q_x - p_x(1)q_y])$. Since markets are now effectively complete, the equilibrium prices in the market structure with forward securities must generate an Arrow–Debreu allocation. But we have already observed that the prices of state contingent goods must all be equal in an Arrow–Debreu equilibrium. It then follows that the relative spot prices in the two states must also be equal, which contradicts our conclusion above. We conclude that there is no competitive equilibrium for the forward security market structure.

In this example, an equilibrium fails to exist for the market structure with forward securities because the dimension of the resulting space of feasible net trades in state contingent goods abruptly shrinks at certain prices. As the relative prices of future securities and spot prices converge to the same ratio, the volume of trade in future securities that is required for a given transfer of purchasing power across states goes to infinity. Consequently, the demand functions for securities may be unbounded even in regions where all relative prices are bounded away from zero. To avoid this problem, Radner (1972) imposes an exogenous lower bound on short sales of securities and shows that this is sufficient to guarantee the existence of equilibrium under standard assumptions. Another approach is to assume that the set of securities is sufficiently rich to guarantee that the dimensionality of net trades does not vary with the price as in Geanokoplos and Polemarchakis (1986). Under these conditions, the demand functions remain bounded and continuous, so there is no need for an exogenous lower bound on excess demand. Kreps (1979) also notes that the set of transfers for any set of securities has full rank for almost all spot prices and therefore that the existence problem is not generic. A general theorem for the generic existence of equilibrium is established by Duffie and Shafer (1985).

Pareto efficiency

As observed above, whenever the two-stage market structure generates a complete set of markets, the equilibrium allocation is an Arrow–Debreu allocation and is therefore Pareto optimal. However, if the first-stage market does not span the space of income transfers, then markets are not complete and the equilibrium allocation is generally not Pareto optimal. In this case, it may be of more use to restrict attention to a more limited set of allocations that reflect the restrictions imposed by the market structure. With the segmentation of markets, however, it is not immediately obvious how we should redefine the set of feasible allocations. For instance, if we permit a central planner to reallocate securities in each spot market, then any technologically feasible allocation can be obtained. To capture the restrictions implied by the market structure, therefore, we must impose some restrictions on how the spot market securities may be allocated.

One possibility is to permit the central planner to arbitrarily allocate securities in the first-stage market, but leave the allocation of securities in the spot markets to be determined by market clearing prices. This approach leads to the following definition suggested by Hart (1975). Let F denote the set of securities in the first-stage market. An allocation of state contingent goods is *constrained Pareto efficient* if (*a*) it is attained as an equilibrium in the spot markets for some feasible distribution of securities in F, and (*b*) there is no Pareto superior allocation of state contingent goods attained as an equilibrium in the spot markets for some other feasible distribution of securities in F.

We will show that when the number of securities in F is less than S, an equilibrium need not be even constrained Pareto efficient. The reason is that a redistribution of the ownership of securities generally leads to a change in the spot market prices and hence to a change in the vector of income transfers associated with each security. As we observed above, when the set of securities in F does not span R^S, the space of transfer vectors that are spanned by the securities in F generally depends on the spot market prices. Consequently, the transfer of real income generated by the redistribution of securities following the adjustment of prices in the spot markets typically lies outside the span of the transfers generated by the set of securities at the competitive equilibrium prices. By redistributing existing securities, therefore, it may be possible to increase the welfare of every agent in the economy.

To illustrate, consider an economy with three agents, *a*, *b* and *c*, and two states of the world, 1 and 2. In each state *s* there are two goods, labelled X and Y. Suppose the preferences and endowments of the agents are given by Table 2, where $x_{\alpha s}$ and $y_{\alpha s}$ are the respective consumption of goods X and Y by agent α in state *s*.

Agent	Endowments		Utility
	(X_1, Y_1)	(X_2, Y_2)	
a	(0, 2)	(2, 0)	$x_{a1} + \varepsilon \min\{x_{a2}, y_{a2}\}$
b	(2, 0)	(0, 2)	$\varepsilon \min\{x_{b1}, y_{b1}\} + x_{b2}$
c	(1, 1)	(1, 1)	$y_{c1} + y_{c2}$

In this economy, agent *a* is endowed with two units of good Y in state 1 and two units of good X in state 2. He consumes only good X in state 1 and always consumes an equal amount of both goods in state 2.

For each pair of units of the two goods he consumes in state 2 he is willing to give up ε units of his consumption of good *X* in state 1. The endowment and preferences of agent *b* are the same except that the role of the two states is reversed. Agent *c* is endowed with one unit of good *X* in both states but consumes only good *Y*. His marginal rate of substitution between consumption in the two states is one. Suppose there is a single security that promises to deliver one unit of good *X* in each state. Since there is nothing for which to exchange this security, the equilibrium income and spot prices in each state will be determined solely by the endowments of the agents in that state. It is easy to check that the relative price of the two goods is one in both states. Agent *a* consumes two units of good *X* in state 1 and one unit of each good in state 2. Agent *b* consumes one unit of each good in state 1 and one unit of good *X* in state 2. Agent *c* consumes two units of good *Y* in both states.

Although the security will never be traded in the market, it can still be used by the government to redistribute purchasing power in the two states and thereby change the spot prices. Suppose, for instance, that agents *a* and *b* must each supply agent *c* with two units of the security. Then the effect is the same as if the endowments were changed as to the endowments listed in Table 3.

Agent Endowments

	(X_1, Y_1)	(X_2, Y_2)
a	(–2, 2)	(0, 0)
b	(0, 0)	(–2, 2)
c	(5, 1)	(5, 1)

For this economy the equilibrium price of good *Y* in terms of good *X* in each state is 5/2. Agent *a* consumes the three units of good *X* in state 1 and nothing in state 2. Agent *b* consumes nothing in state 1 and all three units of good *X* in state 2. Agent *c* consumes the three units of good *Y* in both states.

Now compare the welfare of the two agents in the two economies. Without the transfer payments, agents *a* and *b* attain an expected utility of $2+\varepsilon$ while agent c attains an expected utility of 4. With the transfer payments, agent *a* and *b* both attain an expected utility of 3 while agent c attains a utility of 6. Consequently, for $0<\varepsilon<1$, the equilibrium with transfer payments Pareto dominates the equilibrium without transfer payments. By transferring purchasing power to agent *c* in both states, the economy has made the price of the goods demanded by agents *a* and *b* cheaper in those states where they value their increased welfare the most.

The possibility that securities can be reallocated to attain a Pareto superior allocation when markets are incomplete was first illustrated by Hart (1975). He provided an example in which removing securities and hence decreasing the possibilities for trade actually resulted in a Pareto superior allocation. The intuition is similar to that provided in the example above. If markets are not compete, the introduction of a new security may change the spot market prices in such a way that utilities of all agents decrease unless they can make trades that are not available with the existing set of securities.

Geanokoplos and Polemarchakis (1986) consider a model with two periods and enlarge the commodity space to include consumption before the state of nature is realized. With a complete set of spot markets in the second period and a combined spot and securities market in the first period, they establish that the competitive equilibrium is almost never constrained Pareto optimal whenever the number of securities in *F* is less than *S* and there are at least two goods in each state. Geanokoplos et al. (1990) establish a similar result for a general equilibrium model of the stock market.

Nominal securities and the indeterminacy of equilibrium

Cass (1985) investigates the implications for equilibrium when some of the securities are 'nominal'. These are securities in which the returns in any state are denominated in some unit of account. When all securities are nominal an equilibrium always exists. However, if the dimension of the span of these securities is less than S, the equilibrium is generally not locally unique. In fact, the dimension of indeterminacy is generally equal to S–1.

This result derives from the fact that the real income actually transferred to any state by a nominal security depends on the price level in that state. Suppose the prices in each spot market s are normalized so that they sum to q_s. Then for each vector, $q=(q_1, \ldots q_S)$, any given nominal security f that promises delivery of f_s units of income in each state s corresponds to a unique 'real' security which pays f_s/q_s units of each good in each state s. Let $g_{f(q)}$ denote the real security $(f_1/q_1, \ldots, f_S/q_S)$, and let $F(q)$ denote the set of all such securities generated by the initial set of nominal securities. Then, for any security in $F(q)$, the relative prices in each state do not affect the amount of real income that is transferred by any given exchange of securities. Consequently, there will generally be a locally unique equilibrium associated with each vector q.

Suppose that the set of nominal securities does not span R^S. Then, any non-proportional change in q (generically) changes the span of $F(q)$. Consequently, when we replace the market of nominal securities with a market of real securities $F(q)$, each (normalized) vector q generally produces a distinct equilibrium allocation. Observe, however, that each of these allocations can be realized as an equilibrium with the same set of nominal securities. Therefore, since the dimension of normalized vectors q is S–1, it follows that the dimension of equilibrium allocations associated with any incomplete set of nominal securities is generically S–1.

Notice that this argument only works when the set of nominal securities does not span RS. When the span is complete, the possibilities for distributing real income using the artificial real securities no longer depend on q. Consequently, any equilibrium must yield an Arrow–Debreu allocation.

See Also

- Arrow–Debreu model of general equilibrium
- multiple equilibria in macroeconomics
- uncertainty and general equilibrium

Bibliography

Arrow, K. 1973. *The role of securities in the optimal allocation of risk-bearing. In Essays in the Theory of Risk-Bearing*. Chicago: Markham.

Cass, D. 1985. On the 'number' of equilibrium allocations with incomplete financial markets. Working

Paper No. 85–16, CARESS, University of Pennsylvania.

Duffie, D. and Shafer, W. 1985. Equilibrium with incomplete markets, I: a basic model of generic existence. *Journal of Mathematical Economics* 14, 285–99.

Geanokoplos, J. and Polemarchakis, H. 1986. Existence, regularity, and constrained, suboptimality of competitive allocations when markets are incomplete. In *Essays in Honor of Kenneth Arrow*, vol. 3, ed. W. Heller, R. Starr and D. Starrett. Cambridge: Cambridge University Press.

Geanokoplos, J. and Polemarchakis, H. 1990. Observability and optimality. *Journal of Mathematical Economics* 19, 153–66.

Geanokoplos, J., Magill, M., Quinzii, M. and Dreze, J. 1990. Generic inefficiency of stock market equilibrium when markets are incomplete. *Journal of Mathematical Economics* 19, 113–42.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11, 418–43.

Kreps, D. 1979. Three essays on capital markets. Technical Report No. 298. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Radner, R. 1972. Existence of equilibrium of plans, prices, and price expectations. *Econometrica* 40, 289–303.

Townsend, R. 1978. On the optimality of forward markets. *American Economic Review* 68, 54–66.

How to cite this article

Wilson, Charles. "incomplete markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000045> doi:10.1057/9780230226203.0773

The New Palgrave Dictionary of Economics Online

indentured servitude

Farley Grubb
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

auctions; colonialism; incomplete contracts; indentured servitude; international migration; labour contracts; redemption; slavery

Article

Indentured servants were workers – mostly unmarried young adult males – who voluntarily entered alienable forward-labour contracts typically lasting between three and five years in exchange for passage to an overseas destination.

Indentured servitude was important to European overseas expansion and labour migration from the 17th into the 20th century. It was initially prominent among English, Scots, and Irish workers moving to colonies in British America. French and German servants joined this trade in the 18th century, going primarily to Canada and Pennsylvania, respectively (Emmer, 1986). The servant trade had disappeared among British, Irish, and French migrants by the Napoleonic era and among Germans by 1820 (Grubb, 1994). Approximately half of the transatlantic migrants in this period were indentured. Servants dominated the colonial labour force early on but, by 1700, African slaves south of Pennsylvania and colonial-born free workers north of Virginia eclipsed them in importance (Galenson, 1981; Grubb and Stitt, 1994).

After 1830, the repression of the African slave trade and the abolition of slavery in many European colonies led to the revival of the servant trade, especially to tropical sugar-plantation colonies. Between 1834 and 1918 around 1,500,000 indentured servants from India, 250,000 from China, 80,000 from Japan, 50,000 from Portuguese Atlantic islands, and 100,000 from Melanesia were sent to British, French, Dutch, Spanish, German, and US colonies in the Caribbean, Indian Ocean, South and West Africa, Malaya, Australia, Peru, Hawaii, Fiji, and Samoa (Emmer, 1986; Northrup, 1995).

Servant contracts in the transatlantic trade were typically preprinted single-page forms with blank spaces where negotiated terms were handwritten in. Contracts specified the destination, length of servitude, transferability rights, and ‘freedom dues’ to be paid at the contract's completion – typically two suits of

clothing. In the post-1830 trade freedom dues typically were return-passage tickets. The work to be performed and the maintenance to be received by servants during their contracts were incompletely specified, with contracts typically stating only that servants were to perform customary labour and masters were to provide food, apparel, and lodging (Grubb, 2000).

Because passage was provided first, servants had an incentive after arrival to run away or not work hard. Running away was criminalized and runaways were harshly penalized with whippings and forced contract extensions. The disincentive to work was remedied through the contract's incompleteness. With the servants' daily provisions incompletely specified contractually, masters could adjust daily provisions to elicit the optimal daily diligence from servants. Freedom dues compensated servants for their masters' incentive to withhold semi-durable provisions (clothing) from servants near the end of the contract (Grubb, 2000).

In the transatlantic trade, markets were largely unregulated and competitive. Servants bargained with shippers over the length of servitude and fixed contract terms before sailing. At debarkation shippers sold these contracts to the highest bidders, thereby recouping their shipping expenses. Competition led to servants signing the shortest contracts necessary to secure passage and to shippers earning zero economic profits on servant cargo. Passage costs were relatively constant across servants but labour productivity was not. Less productive servants had to sign longer contracts for the same passage cost. Contract lengths were inversely related to, whereas auction prices in America were unrelated to, servant productivity known at embarkation (Galenson, 1981; Grubb, 1985). Servants were also charged about 15 per cent more than free passengers (who paid cash in advance) to compensate shippers for forgoing other investment opportunities and to cover expected servant defaults through mortality, morbidity, and escape.

In the mid-18th century a new variant – redemption – came into use primarily among German immigrants. Under redemption passengers entered fixed-debt passage contracts before sailing that required them to enter servitude at debarkation, if necessary, to clear the debt. Redemption shifted the voyage risk and forecast error in the market from shipper to migrant. With passage debts, but not contract lengths, fixed before sailing, shippers no longer had to forecast at embarkation the amount of labour needed in a servant contract for it to sell at debarkation for enough to cover shipping costs. Instead, at debarkation migrants had to offer however much labour was needed to clear the passage debt contractually guaranteed to the shipper before sailing (Galenson, 1981). Migrants accepted this risk because it gave them greater flexibility over selecting their American masters, negotiating contingency clauses into their contracts, and using a single labour contract to pay both the passage debt and any pre-voyage debts transferred to the shipper.

In the post-1830 trade, markets were more highly regulated. For example, in the Melanesian trade to Queensland, Australia, the British government fixed the length of labour contracts at three years and servant wages at six pounds sterling per year, did not allow unrestricted recruiting, and did not allow servants to be auctioned upon arrival. Shippers were licensed to recruit only the number of servants requested by planters and were paid a set fee per recruit. Officials assigned arriving servants to planters according to the number requested. This perversely induced shippers to recruit low-quality labour.

The transatlantic servant trade ended because the supply of servants collapsed, not because American demand declined. Prospective servants found better jobs elsewhere, such as military service during the Napoleonic Wars, or better ways to pay for passage, such as borrowing from already emigrated family members (Grubb, 1994). Many post-1830 servant trades were ended by government action or the

changing fortunes of the global sugar industry (Emmer, 1986; Northrup, 1995).

See Also

- auctions (empirics)
- compensating differentials
- convict labour
- human capital, fertility and growth
- international migration
- labour market institutions

Bibliography

Emmer, P., ed. 1986. *Colonialism and Migration; Indentured Labour Before and After Slavery*. Dordrecht: Martinus Nijhoff.

Galenson, D. 1981. *White Servitude in Colonial America*. New York: Cambridge.

Grubb, F. 1985. The market for indentured immigrants: evidence on the efficiency of forwardcontracting in Philadelphia, 1745–1773. *Journal of Economic History* 45, 855–68.

Grubb, F. 1994. The end of European immigrant servitude in the United States: an economic analysis of market collapse, 1772–1835. *Journal of Economic History* 54, 794–824.

Grubb, F. 2000. The statutory regulation of colonial servitude: an incomplete-contract approach. *Explorations in Economic History* 37, 42–75.

Grubb, F. and Stitt, T. 1994. The Liverpool emigrant servant trade and the transition to slave labor in the Chesapeake, 1697–1707: market adjustments to war. *Explorations in Economic History* 31, 376–405.

Northrup, D. 1995. *Indentured Labor in the Age of Imperialism, 1834–1922*. New York: Cambridge.

How to cite this article

Grubb, Farley. "indentured servitude." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000255> doi:10.1057/9780230226203.0774

The New Palgrave Dictionary of Economics Online

index numbers

W. Erwin Diewert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Index numbers are used to aggregate detailed information on prices and quantities into scalar measures of price and quantity levels or their growth. The article reviews four main approaches to bilateral index number theory where two price and quantity vectors are to be aggregated: fixed basket and average of fixed baskets, stochastic, test or axiomatic and economic approaches. The article also considers multilateral index number theory where it is necessary to construct price and quantity aggregates for more than two value aggregates. A final section notes some of the recent literature on related aspects of index number theory.

Keywords

Allen quantity index; Bowley, A.L.; Carli price index; chain indexes; consumer price index; Edgeworth, F. Y.; Fisher ideal index; Fisher, I.; fixed base indexes; Frisch, R. A. K.; ideal indexes; index number theory; index numbers; Jevons price index; Jevons, W. S.; Konüs price index; Konüs–Pollak quantity index; Laspeyres price index; Laspeyres–Konüs quantity index; logarithmic price ratios; Lowe index; Malmquist quantity index; Marshall, A.; Marshall–Edgeworth index; Paasche price index; Pierson, N. G.; producer price index; productivity indexes; Scrope, G. P.; Sidgwick, H.; Sidgwick–Bowley index; superlative indexes; Theil, H.; Törnqvist–Theil price index; Walsh index; Walsh, C. M.; Young index

Article

1 Introduction

Each individual consumes the services of thousands of commodities over a year and most producers utilize and produce thousands of individual products and services. Index numbers are used to reduce and summarize this overwhelming abundance of microeconomic information. Hence index numbers impinge on virtually every empirical investigation in economics.

The index number problem may be stated as follows. Suppose we have price data $p^t \equiv (p_1^t, \ldots, p_N^t)$ and quantity data $q^t \equiv (q_1^t, \ldots, q_N^t)$ on N commodities that pertain to the same economic unit at time period t (or to comparable economic units) for $t = 0, 1, 2, \ldots, T$. The *index number problem* is to find T+1 numbers P^t and T+1 numbers Q^t such that

$$p^t Q^t = p^t \cdot q^t \equiv \sum_{n=1}^{N} p_n^t q_n^t \text{ for } t = 0, 1, \ldots, T. \quad (1)$$

P^t is the *price index* for period t (or unit t) and Q^t is the corresponding *quantity index*. P^t is supposed to be representative of all of the prices p_n^t, $n = 1, \ldots, N$ in some sense, while Q^t is to be similarly representative of the quantities q_n^t, $n = 1, \ldots, N$. In what precise sense P^t and Q^t represent the individual prices and quantities is not immediately evident, and it is this ambiguity that leads to different approaches to index number theory. Note that we require that the product of the price and quantity indexes, P^tQ^t, equals the actual period (or unit) t expenditures on the N commodities, $p^t \cdot q^t$. Thus if the P^t are determined, then the Q^t may be implicitly determined using eq. (1), or vice versa.

The number P^t is interpreted as an aggregate period t price level while the number Q^t is interpreted as an aggregate period t quantity level. The *levels* approach to index number theory works as follows. The aggregate price level P^t is assumed to be a function of the components in the period t price vector, p^t while the aggregate period t quantity level Q^t is assumed to be a function of the period t quantity vector components, q^t; that is, it is assumed that

$$P^t = c(p^t) \text{ and } Q^t = f(q^t); \quad t = 0, 1, \ldots, T. \quad (2)$$

The functions c and f are to be determined somehow. Note that we are requiring that the functional forms for the price aggregation function c and for the quantity aggregation function f be independent of time. This is a reasonable requirement since there is no reason to change the method of aggregation as time changes.

Substituting (2) into (1) and dropping the superscripts t means that c and f must satisfy the following functional equation for all strictly positive price and quantity vectors:

$$c(p) f(q) = p \cdot q \equiv \sum_{n=1}^{N} p_n q_n \text{ for all } p \gg 0_N \text{ and for all } q \gg 0_N. \quad (3)$$

Note that $p \gg 0_N$ means that each component of p is positive, $p \geq 0_N$ means each component is non-negative and $p > 0_N$ means each component is non-negative and at least one component is positive. We now could ask what properties the price aggregation function c and the quantity aggregation function f should have. We could assume that c and f satisfied various ‘reasonable’ properties and hope that these properties would determine the functional form for c and f. However, it turns out that we have only to make the following very weak *positivity* assumptions on f and c in order to obtain an impossibility result:

$$c(p) > 0 \text{ for all } p \gg 0_N; \quad f(q) > 0 \text{ for all } q \gg 0_N. \quad (4)$$

Eichhorn (1978, p. 144) proved the following result: if the number of commodities N is greater than 1, then there do not exist any functions c and f that satisfy (3) and (4). Thus this *levels approach* to index number theory comes to an abrupt halt. As we shall see later, when the economic approach to index number theory is studied, this is not quite the end of the story: in (3) and (4), we allowed p and q to vary independently from each other, and this is what leads to the impossibility result. If instead we allow p to vary independently but assume that q is determined as the result of an optimizing model, then eq. (3) can be satisfied.

If we change the question that we are trying to answer slightly, then there are practical solutions to the index number problem. The change is that instead of trying to decompose the value of the aggregate into price and quantity components for a single period, we instead attempt to decompose a *value ratio* pertaining to two periods, say periods 0 and 1, into a *price change component P* times a *quantity change component Q*. Thus we now look for two functions of $4N$ variables, $P(p^0,p^1,q^0,q^1)$ and $Q(p^0,p^1,q^0,q^1)$ so that:

$$p^1 \cdot q^1 / p^0 \cdot q^0 = P(p^0, p^1, q^0, q^1)\, Q(p^0, p^1, q^0, q^1). \tag{5}$$

Note that if some approach to index number theory determines the 'best' functional form for the price index $P(p^0,p^1,q^0,q^1)$, then the *product test* (5) can be used to determine the functional form for the corresponding quantity index, $Q(p^0,p^1,q^0,q^1)$.

If we take the *test or axiomatic approach* to index number theory, then we want eq. (5) to hold for all positive price and quantity vectors pertaining to the two periods under consideration, p^0,p^1,q^0,q^1. If we take the *economic approach*, then only the price vectors p^0 and p^1 are regarded as independent variables while the quantity vectors, q^0 and q^1, are regarded as dependent variables. In Section 4 below, we will pursue the test approach and in Sections 5 to 7, we will take the economic approach. In Sections 2 to 7, we take a *bilateral approach to index number theory*; that is, in making price and quantity comparisons between any two time periods, the relevant indexes use *only* price and quantity information that pertains to the two periods under consideration. It is also possible to take a *multilateral approach*; that is, we look for functions, P^t and Q^t, that are functions of *all* of the price and quantity vectors, $p^0,p^1,\ldots,p^T,q^0,q^1,\ldots,q^T$. Thus we look for $2(T+1)$ functions, $P^t(p^0,p^1,\ldots,p^T,q^0,q^1,\ldots,q^T)$ and $Q^t(p^0,p^1,\ldots,p^T,q^0,q^1,\ldots,q^T)$, $t=0,1,\ldots,T$, so that

$$p^t \cdot q^t = P^t(p^0, p^1, \ldots, p^T, q^0, q^1, \ldots, q^T)\, Q^t(p^0, p^1, \ldots, p^T, q^0, q^1, \ldots, q^T) \text{ for } t = 0, 1, \ldots, T. \tag{6}$$

We briefly pursue the multilateral approach to index number theory in Section 9.

The four main approaches to bilateral index number theory will be covered in this review: (i) the *fixed basket approach* (Section 2), (ii) the *stochastic approach* (Section 3), (iii) the *test approach* (Section 4) and (iv) the *economic approach*, which relies on the assumption of maximizing or minimizing behaviour (Sections 5–7).

Section 8 discusses fixed base versus chained index numbers, and Section 10 concludes by mentioning some recent areas of active research in the index number literature.

2 Fixed basket approaches

The English economist Joseph Lowe (1823) developed the theory of the consumer price index in some detail. His approach to measuring the price change between periods 0 and 1 was to specify an approximate representative commodity basket quantity vector, $q \equiv (q_1, \ldots, q_N)$, which was to be updated every five years, and then calculate the level of prices in period 1 relative to period 0 as

$$P_{Lo}(p^0, p^1, q) \equiv p^1 \cdot q \,/\, p^0 \cdot q \tag{7}$$

where p^0 and p^1 are the commodity price vectors that the consumer (or group of consumers) face in periods 0 and 1 respectively. The fixed basket approach to measuring price change is intuitively very simple: we simply specify the commodity 'list' q and calculate the price index as the ratio of the costs of buying this same list of goods in periods 1 and 0.

As time passed, economists and price statisticians demanded more precision with respect to the specification of the basket vector q. There are two natural choices for the reference basket: the period 0 commodity vector q^0 or the period 1 commodity vector q^1. These two choices lead to the Laspeyres (1871) price index P_L defined by (8) and the Paasche (1874) price index P_P defined by (9):

$$P_L(p^0, p^1, q^0, q^1) \equiv p^1 \cdot q^0 \,/\, p^0 \cdot q^0; \tag{8}$$

$$P_P(p^0, p^1, q^0, q^1) \equiv p^1 \cdot q^1 \,/\, p^0 \cdot q^1. \tag{9}$$

The above formulae can be rewritten in an alternative manner that is very useful for statistical agencies. Define the period t expenditure share on commodity n as follows:

$$s_n^t \equiv p_n^t q_n^t \,/\, p^t \cdot q^t \text{ for } n = 1, \ldots, N \text{ and } t = 0, 1. \tag{10}$$

Following Fisher (1911), the Laspeyres index (8) can be rewritten as follows:

$$P_L(p^0, p^1, q^0, q^1) = \sum_{n=1}^{N} p_n^1 q_n^0 \,/\, p^0 \cdot q^0 = \sum_{n=1}^{N} (p_n^1 \,/\, p_n^0)\, p_n^0 q_n^0 \,/\, p^0 \cdot q^0 = \sum_{n=1}^{N} (p_n^1 \,/\, p_n^0)\, s_n^0 \quad \text{using definitions (10).} \tag{11}$$

Thus the Laspeyres price index P_L can be written as a base period expenditure share weighted average of the N price ratios (or price relatives using index number terminology), p_n^1 / p_n^0. The Laspeyres formula (until the very recent past when in 2003 the US Bureau of Labor Statistics introduced its chained consumer price index) has been widely used as the intellectual basis for country consumer price indexes (CPIs) around the world. To implement the formula, the country statistical agency collects information on expenditure shares s_n^0 for the index domain of definition for the base period 0 and then collects information on *prices* alone on an ongoing basis. Thus a Laspeyres-type CPI can be produced on a timely basis without one having to know current period quantity information. In fact, the situation is more complicated than this: in actual CPI programmes, prices are collected on a monthly or quarterly frequency and with base month 0 say, but the quantity vector q^0 is typically *not* the quantity vector that pertains to the price base month 0; rather, it is actually equal to a *base year quantity vector*, q^b say, which is typically prior to the base month 0. Thus the typical CPI, although loosely based on the Laspeyres index, is actually a form of Lowe index; see (7) above. Instead of using the Lowe formula for their CPI, some statistical agencies use the following Young (1812) index:

$$P_Y(p^0, p^1, s^b) \equiv \sum_{n=1}^{N} (p_n^1 / p_n^0) s_n^b \tag{12}$$

where the s_n^b are base year expenditure shares on the N commodities in the index. For additional material on Lowe and Young indexes and their use in CPI and producer price index (PPI) programmes, see the ILO (2004) and the IMF (2004).

The Paasche index can also be written in expenditure share and price ratio form as follows:

$$P_P(p^0, p^1, q^0, q^1) = 1 / \left[\sum_{n=1}^{N} p_n^0 q_n^1 / p^1 \cdot q^1\right] = 1 / \left[\sum_{n=1}^{N} (p_n^0 / p_n^1) p_n^1 q_n^1 / p^1 \cdot q^1\right] = 1 / \left[\sum_{n=1}^{N} (p_n^1 / p_n^0)^{-1} s_n^1\right] \text{ using definitions (10)} = \left[\sum_{n=1}^{N} (p_n^1 / p_n^0)^{-1} s_n^1\right]^{-1}. \tag{13}$$

Thus the Paasche price index P_P can be written as a period 1 (or current period) expenditure share weighted *harmonic* average of the N price ratios.

The problem with the Paasche and Laspeyres index number formulae is that they are equally plausible but, in general, they will give different answers. This suggests that, if we require a single estimate for the price change between the two periods, then we need to take some sort of evenly weighted average of the two indexes as our final estimate of price change between periods 0 and 1. Examples of such symmetric averages are the arithmetic mean, which leads to the Sidgwick (1883, p. 68) Bowley (1901, p. 227) index, $(1/2)P_L+(1/2)P_P$, and the geometric mean, which leads to the Fisher (1922) ideal index, P_F, which was actually first suggested by Bowley (1899, p. 641), defined as

$$P_F(p^0, p^1, q^0, q^1) \equiv [P_L(p^0, p^1, q^0, q^1) P_P(p^0, p^1, q^0, q^1)]^{1/2}. \tag{14}$$

At this point, the fixed basket approach to index number theory is transformed into the *test approach* to index number theory; that is, in order to determine which of these fixed basket indexes or which averages of them might be best, we need *criteria* or *tests* or *properties* that we would like our indexes to satisfy. We will pursue this topic in

more detail in Section 4, but we give the reader an introduction to this topic in the present section because some of these tests or properties are useful to evaluate other approaches to index number theory.

Let a and b be two positive numbers. Diewert (1993b, p. 361) defined a *symmetric mean* of a and b as a function $m(a, b)$ that has the following properties: (i) $m(a, a)=a$ for all $a>0$ (mean property); (ii) $m(a, b) = m(b, a)$ for all $a>0$, $b>0$ (symmetry property); (iii) $m(a, b)$ is a continuous function for $a>0$, $b>0$ (continuity property) and (iv) $m(a, b)$ is a strictly increasing function in each of its variables (increasingness property). Eichhorn and Voeller (1976, p. 10) showed that, if $m(a, b)$ satisfies the above properties, then it also satisfies the following property: (v) $\min\{a, b\} \leq m\{a, b\} \leq \max\{a, b\}$ (min–max property); that is, the mean of a and b, $m(a, b)$, lies between the maximum and minimum of the numbers a and b. Since we have restricted the domain of definition of a and b to be positive numbers, it can be seen that an implication of the last property is that m also satisfies the following property: (vi) $m(a, b)>0$ for all $a>0$, $b>0$ (positivity property). If in addition, m satisfies the following property, then we say that m is a *homogeneous symmetric mean*: (vii) $m(\lambda a, \lambda b)= \lambda m(a, b)$ for all $\lambda>0$, $a>0$, $b>0$.

What is the best symmetric average of P_L and P_P to use as a point estimate for the theoretical cost of living index? It is very desirable for a price index formula that depends on the price and quantity vectors pertaining to the two periods under consideration to satisfy the *time reversal test*. We say that the index number formula $P(p^0,p^1,q^0,q^1)$ satisfies this test if

$$P(p^1, p^0, q^1, q^0) = 1 / P(p^0, p^1, q^0, q^1); \tag{15}$$

that is, if we interchange the period 0 and period 1 price and quantity data and evaluate the index, then this new index $P(p^1,p^0,q^1,q^0)$ is equal to the reciprocal of the original index $P(p^0,p^1,q^0,q^1)$. For the history of this test (and other tests), see Diewert (1992a, p. 218; 1993a).

Diewert (1997, p. 138) proved the following result: the Fisher ideal price index defined by (14) above is the *only* index that is a homogeneous symmetric average of the Laspeyres and Paasche price indexes, P_L and P_P, that also satisfies the time reversal test (15) above.

Thus the symmetric basket approach to index number theory leads to the Fisher ideal index as the best formula. It is interesting to note that this symmetric basket approach to index number theory dates back to Bowley, one of the early pioneers of index number theory, as the following quotations indicate:

> If [the Paasche index] and [the Laspeyres index] lie close together there is no further difficulty; if they differ by much they may be regarded as inferior and superior limits of the index number, which may be estimated as their arithmetic mean … as a first approximation. (Bowley, 1901, p. 227)

> When estimating the factor necessary for the correction of a change found in money wages to obtain the change in real wages, statisticians have not been content to follow Method II only [to calculate a Laspeyres price index], but have worked the problem backwards [to calculate a Paasche price index] as well as forwards.… They have then taken the arithmetic, geometric or harmonic mean of the two numbers so found. (Bowley, 1919, p. 348)

Instead of taking a symmetric average of the Paasche and Laspeyres indexes, an alternative average basket approach takes a symmetric average of the baskets that prevail in the two periods under consideration. For example, the average basket could be the arithmetic or geometric mean of the two baskets, leading the Marshall (1887) Edgeworth (1925) index P_{ME} or the Walsh (1901, p. 398; 1921a, pp. 97–101) index P_W:

$$P_{ME}(p^0, p^1, q^0, q^1) \equiv \sum_{n=1}^{N} p_n^1(1/2)(q_n^0 + q_n^1) / \sum_{m=1}^{N} p_j^0(1/2)(q_m^0 + q_m^1); \tag{16}$$

$$P_W(p^0, p^1, q^0, q^1) \equiv \sum_{n=1}^{N} p_n^1 (q_n^0 q_n^1)^{1/2} \Big/ \sum_{m=1}^{N} p_m^0 (q_m^0 q_m^1)^{1/2}. \tag{17}$$

Diewert (2002b, pp. 569–71) showed that the Walsh index P_W emerged as being best in this average basket framework; see also ILO (2004, chs 15 and 16).
We turn now to the second major approach to bilateral index number theory.

3 The stochastic approach to index number theory

> In drawing our averages the independent fluctuations will more or less destroy each other; the one required variation of gold will remain undiminished. (Jevons, 1884, p. 26)

The stochastic approach to the determination of the price index can be traced back to the work of Jevons (1865; 1884) and Edgeworth (1888; 1923; 1925) over 100 years ago. For additional discussion on the early history of this approach, see Diewert (1993a, pp. 37–8; 1995b).

The basic idea behind the stochastic approach is that each price relative, p_n^1 / p_n^0 for n=1,2,…,N can be regarded as an estimate of a common inflation rate α between periods 0 and 1; that is, it is assumed that

$$p_n^1 / p_n^0 = \alpha + \varepsilon_n; \quad n = 1, 2, \ldots, N \tag{18}$$

where α is the common inflation rate and the ε_n are random variables with mean 0 and variance σ^2. The least squares estimator for α is the Carli (1764) price index P_C defined as

$$P_C(p^0, p^1) \equiv \sum_{n=1}^{N} (1/N)(p_n^1 / p_n^0). \tag{19}$$

Unfortunately, P_C does not satisfy the time reversal test, namely, $P_C(p^1, p^0) \neq 1 / P_C(p^0, p^1)$. In fact, Fisher (1922, p. 66) noted that $P_C(p^0, p^1) P_C(p^1, p^0) \geq 1$ unless the period 1 price vector p^1 is proportional to the period 0 price vector p^0; that is, Fisher showed that the Carli (and the Young) index has a definite upward bias. He

urged statistical agencies not to use these formulae.

Now assume that the logarithm of each price relative, $\ln(p_n^1 / p_n^0)$, is an unbiased estimate of the logarithm of the inflation rate between periods 0 and 1, β say. Thus we have:

$$\ln(p_n^1 / p_n^0) = \beta + \varepsilon_n; \quad n = 1, 2, \ldots, N \tag{20}$$

where $\beta \equiv \ln \alpha$ and the ε_n are independently distributed random variables with mean 0 and variance σ^2. The least squares estimator for β is the logarithm of the geometric mean of the price relatives. Hence the corresponding estimate for the common inflation rate α is the Jevons (1865) price index P_J defined as:

$$P_J(p^0, p^1) \equiv \prod_{n=1}^{N} (p_n^1 / p_n^0)^{1/N}. \tag{21}$$

The Jevons price index P_J satisfies the time reversal test and hence is much more satisfactory than the Carli index P_C.

Bowley (1928) attacked the use of both (19) and (21) on two grounds. First, from an empirical point of view, he showed that price ratios were not symmetrically distributed about a common mean and their logarithms also failed to be symmetrically distributed. Second, from a theoretical point of view, he argued that it was unlikely that prices or price ratios were independently distributed. Keynes (1930) developed Bowley's second objection in more detail; he argued that changes in the money supply would not affect all prices at the same time. Moreover, real disturbances in the economy could cause one set of prices to differ in a systematic way from other prices, depending on various elasticities of substitution and complementarity. In other words, prices are not randomly distributed, but are systematically related to each other through the general equilibrium of the economy. Keynes (1930, pp. 76–7) had other criticisms of this *unweighted stochastic approach* to index number theory, including the point that that there is no such thing as *the* inflation rate; there are only price changes that pertain to well-specified sets of commodities or transactions; that is, the domain of definition of the price index must be carefully specified. Keynes also followed Walsh in insisting that price movements must be weighted by their economic importance, that is, by quantities or expenditures:

> It might seem at first sight as if simply every price quotation were a single item, and since every commodity (any kind of commodity) has one price-quotation attached to it, it would seem as if price-variations of every kind of commodity were the single item in question. This is the way the question struck the first inquirers into price-variations, wherefore they used simple averaging with even weighting. But a price-quotation is the quotation of the price of a generic name for many articles; and one such generic name covers a few articles, and another covers many.… A single price-quotation, therefore, may be the quotation of the price of a hundred, a thousand, or a million dollar's worth, of the articles that make up the commodity named. Its weight in the averaging, therefore, ought to be according to these money-unit's worth. (Walsh, 1921a, pp. 82–3)

Theil (1967, pp. 136–7) proposed a solution to the lack of weighting in (21). He argued as follows. Suppose we draw price relatives at random in such a way that each dollar of expenditure in the base period has an equal chance of being selected. Then the probability that we will draw the nth price relative is equal to $s_n^0 \equiv p_n^0 q_n^0 / p^0 \cdot q^0$, the period 0 expenditure share for commodity *n*. Then the overall mean (period 0 weighted) logarithmic price change is $\sum_{n=1}^{N} s_n^0 \ln(p_n^1 / p_n^0)$. Now repeat the above mental experiment and draw price relatives at random in such a way that each dollar of expenditure in period 1 has an equal probability of being selected. This leads to the overall

mean (period 1 weighted) logarithmic price change of $\Sigma_{n=1}^{N} s_n^1 \ln(p_n^1 / p_n^0)$. Each of these measures of overall logarithmic price change seems equally valid so we could argue for taking a symmetric average of the two measures in order to obtain a final single measure of overall logarithmic price change. Theil (1967, p. 138) argued that a nice symmetric index number formula can be obtained if we make the probability of selection for the nth price relative equal to the arithmetic average of the period 0 and 1 expenditure shares for commodity *n*. Using these probabilities of selection, Theil's final measure of overall logarithmic price change was

$$\ln P_T(p^0, p^1, q^0, q^1) \equiv \sum_{n=1}^{N} (1/2)(s_n^0 + s_n^1) \ln(p_n^1 / p_n^0). \tag{22}$$

We can give the following *descriptive statistics* interpretation of the right hand side of (22). Define the *n*th logarithmic price ratio r_n by:

$$r^n \equiv \ln(p_n^1 / p_n^0) \text{ for } n = 1, \ldots, N. \tag{23}$$

Now define the discrete random variable, *R* say, as the random variable which can take on the values r_n with probabilities $\rho_n \equiv (1/2)[s_n^0 + s_n^1]$ for *n*=1,…,*N*. Note that, since each set of expenditure shares, s_n^0 and s_n^1, sums to one, the probabilities ρ_n will also sum to one. It can be seen that the expected value of the discrete random variable *R* is

$$E[R] \equiv \sum_{n=1}^{N} \rho_n r_n = \sum_{n=1}^{N} (1/2)(s_n^0 + s_n^1) \ln(p_n^1 / p_n^0) = \ln P_T(p^0, p^1, q^0, q^1) \tag{24}$$

using (22) and (23). Thus the logarithm of the index P_T can be interpreted as *the expected value of the distribution of the logarithmic price ratios* in the domain of definition under consideration, where the *N* discrete price ratios in this domain of definition are weighted according to Theil's probability weights, $\rho_n \equiv (1/2)[s_n^0 + s_n^1]$ for *n*=1,…,*N*. If we take antilogs of both sides of (24), we obtain the Törnqvist (1936), Törnqvist and Törnqvist (1937) Theil price index, P_T. This index number formula has a number of good properties. Thus the second major approach to bilateral index number theory has led to the Törnqvist–Theil price index P_T as being best from this perspective. Additional material on stochastic approaches to index number theory and references to the literature can be found in Selvanathan and Rao (1994), Diewert (1995b), Wynne (1997), ILO (2004), IMF (2004) and Clements, Izan and Selvanathan (2006).

Formulae (8), (9), (14) and (22) (the Laspeyres, Paasche, Fisher and Törnqvist–Theil formulae) are the most widely used formulae for a bilateral price index. But Walsh (1901) and Fisher (1922) presented hundreds of functional forms for bilateral price indexes – on what basis are we to choose one as being better than the other? Perhaps the next approach to index number theory will narrow the choices.

4 The test approach to index number theory

In this section, we will take the perspective outlined in Section 1 above; that is, along with the price index $P(p^0,p^1,q^0,q^1)$, there is a companion quantity index $Q(p^0,p^1,q^0,q^1)$ such that the product of these two indexes equals the value ratio between the two periods. Thus, throughout this section, we assume that P and Q satisfy the product test (5) above.

If we assume that the product test holds means that as soon as the functional form for the price index P is determined, then (5) can be used to determine the functional form for the quantity index Q. However, as Fisher (1911, pp. 400–6) and Vogt (1980) observed, a further advantage of assuming that the product test holds is that we can assume that the quantity index Q satisfies a 'reasonable' property and then use (5) to translate this test on the quantity index into a corresponding test on the price index P.

If N=1, so that there is only one price and quantity to be aggregated, then a natural candidate for P is p_1^1 / p_1^0, the single price ratio, and a natural candidate for Q is q_1^1 / q_1^0, the single quantity ratio. When the number of commodities or items to be aggregated is greater than 1, then what index number theorists have done over the years is to propose properties or tests that the price index P should satisfy. These properties are generally multidimensional analogues to the one good price index formula, p_1^1 / p_1^0. Below, following Diewert (1992a), we list 20 tests that characterize the Fisher ideal price index.

We shall assume that every component of each price and quantity vector is positive; that is, $p^t \gg 0_N$ and $q^t \gg 0_N$ for t=0,1. If we want to set $q^0=q^1$, we call the common quantity vector q; if we want to set $p^0=p^1$, we call the common price vector p.

Our first two tests, due to Eichhorn and Voeller (1976, p. 23) and Fisher (1922, pp. 207–15), are not very controversial and so we will not discuss them.

- T1: *Positivity*: $P(p^0,p^1,q^0,q^1)>0$.
- T2: *Continuity*: $P(p^0,p^1,q^0,q^1)$ is a continuous function of its arguments.

Our next two tests, due to Laspeyres (1871, p. 308), Walsh (1901, p. 308) and Eichhorn and Voeller (1976, p. 24), are somewhat more controversial.

- T3: *Identity or constant prices test*: $P(p,p,q^0,q^1)=1$.

That is, if the price of every good is identical during the two periods, then the price index should equal unity, no matter what the quantity vectors are. The controversial part of this test is that the two quantity vectors are allowed to be different in the above test.

- T4: *Fixed basket or constant quantities test*: $P(p^0, p^1, q, q) = \Sigma_{i=1}^N p_i^1 q_i / \Sigma_{i=1}^N p_i^0 q_i$.

That is, if quantities are constant during the two periods so that $q^0=q^1\equiv q$, then the price index should equal the expenditure on the constant basket in period 1, $\Sigma_{i=1}^N p_i^1 q_i$, divided by the expenditure on the basket in period 0, $\Sigma_{i=1}^N p_i^0 q_i$. The origins of this test go back at least 200 years to the Massachusetts legislature which used a constant basket of goods to index the pay of Massachusetts soldiers fighting in the American Revolution: see Willard Fisher (1913). Other researchers who have suggested the test over the years include Lowe (1823, Appendix, p. 95), Scrope (1833, p. 406), Jevons (1865), Sidgwick (1883, pp. 67–8), Edgeworth (1887, p. 215), Marshall (1887, p. 363), Pierson (1895, p. 332), Walsh (1901, p. 540; 1921b, p. 544), and Bowley (1901, p. 227). Vogt and Barta (1997, p. 49) also observed that this test is a special case of Fisher's (1911, p. 411) proportionality test for quantity indexes which Fisher (1911, p. 405) translated into a test for the price index using the product test (5).

The following four tests restrict the behaviour of the price index P as the scale of any one of the four vectors p^0,p^1,q^0,q^1 changes. The following test was proposed by Walsh (1901, p. 385), Eichhorn and Voeller (1976, p. 24) and Vogt (1980, p. 68).

- T5: *Proportionality in Current Prices*: $P(p^0, \lambda p^1,q^0,q^1)= \lambda P(p^0,p^1,q^0,q^1)$ for $\lambda >0$.

That is, if all period 1 prices are multiplied by the positive number λ, then the new price index is λ times the old price index. Put another way, the price index function $P(p^0,p^1,q^0,q^1)$ is (positively) homogeneous of degree one in the components of the period 1 price vector p^1. Most index number theorists regard this property as a very fundamental one that the index number formula should satisfy.

Walsh (1901) and Fisher (1911, p. 418; 1922, p. 420) proposed the related proportionality test $P(p, \lambda p,q^0,q^1)= \lambda$. This last test is a combination of T3 and T5; in fact Walsh (1901, p. 385) noted that this last test implies the identity test, T3.

In the next test, due to Eichhorn and Voeller (1976, p. 28), instead of multiplying all period 1 prices by the same number, we multiply all period 0 prices by the number λ.

- T6: *Inverse proportionality in base period prices*: $P(\lambda p^0,p^1,q^0,q^1)= \lambda^{-1}P(p^0,p^1,q^0,q^1)$ for $\lambda>0$.

That is, if all period 0 prices are multiplied by the positive number λ, then the new price index is $1/\lambda$ times the old price index. Put another way, the price index function $P(p^0,p^1,q^0,q^1)$ is (positively) homogeneous of degree minus one in the components of the period 0 price vector p^0.

The following two homogeneity tests can also be regarded as invariance tests.

- T7: *Invariance to proportional changes in current quantities*: $P(p^0,p^1,q^0, \lambda q^1)=P(p^0,p^1,q^0,q^1)$ for all $\lambda>0$.

That is, if current period quantities are all multiplied by the number λ, then the price index remains unchanged. Put another way, the price index function $P(p^0,p^1,q^0,q^1)$ is (positively) homogeneous of degree zero in the components of the period 1 quantity vector q^1. Vogt (1980, p. 70) was the first to propose this test and his derivation of the test is of some interest. Suppose the quantity index Q satisfies the quantity analogue to the price test T5, that is, suppose Q satisfies $Q(p^0,p^1,q^0, \lambda q^1)= \lambda Q(p^0,p^1,q^0,q^1)$ for $\lambda>0$. Then using the product test (5), we see that P must satisfy T7.

- T8: *Invariance to proportional changes in base quantities*: $P(p^0,p^1, \lambda q^0,q^1)=P(p^0,p^1,q^0,q^1)$ for all $\lambda>0$.

That is, if base period quantities are all multiplied by the number λ, then the price index remains unchanged. Put another way, the price index function $P(p^0,p^1,q^0,q^1)$ is (positively) homogeneous of degree zero in the components of the period 0 quantity vector q^0. If the quantity index Q satisfies the following counterpart to T8: $Q(p^0,p^1, \lambda q^0, q^1)= \lambda^{-1}Q(p^0,p^1,q^0,q^1)$ for all $\lambda>0$, then, using (5), the corresponding price index P must satisfy T8. This argument provides some additional justification for assuming the validity of T8 for the price index function P. This test was proposed by Diewert (1992a, p. 216).

T7 and T8 together impose the property that the price index P does not depend on the *absolute* magnitudes of the quantity vectors q^0 and q^1.

The next five tests are invariance or symmetry tests. Fisher (1922, pp. 62–3, 458–60) and Walsh (1921b, p. 542) seem to have been the first researchers to appreciate the significance of these kinds of tests. Fisher (1922, pp. 62–3) spoke of fairness but it is clear that he had symmetry properties in mind. It is perhaps unfortunate that he did not realize that there were more symmetry and invariance properties than the ones he proposed; if he had realized this, it is likely that he would have been able to provide an axiomatic characterization for his ideal price index, as will be done shortly. Our first invariance test is that the price index should remain unchanged if the *ordering* of the commodities is changed:

- T9: *Commodity reversal test* (or invariance to changes in the ordering of commodities):

$$P(p^{0*}, p^{1*}, q^{0*}, q^{1*}) = P(p^0, p^1, q^0, q^1)$$

where p^{t*} denotes a permutation of the components of the vector p^t and q^{t*} denotes the same permutation of the components of q^t for t=0,1. This test is due to Fisher (1922),

and it is one of his three famous reversal tests. The other two are the time reversal test and the factor reversal test which will be considered below.

- T10: *Invariance to changes in the units of measurement* (commensurability test):

$$P(\alpha_1 p_1^0, \ldots, \alpha_N p_N^0; \alpha_1 p_1^1, \ldots, \alpha_N p_N^1; \alpha_1^{-1} q_1^0, \ldots, \alpha_N^{-1} q_N^0; \alpha_1^{-1} q_1^1, \ldots, \alpha_N^{-1} q_N^1) = P(p_1^0, \ldots, p_N^0; p_1^1, \ldots, p_N^1; q_1^0, \ldots, q_N^0; q_1^1, \ldots, q_N^1) \text{ for all } \alpha_1 > 0, \ldots, \alpha_N > 0.$$

That is, the price index does not change if the units of measurement for each commodity are changed. The concept of this test was due to Jevons (1884, p. 23) and the Dutch economist Pierson (1896, p. 131), who criticized several index number formula for not satisfying this fundamental test. Fisher (1911, p. 411) first called this test *the change of units test* and later, Fisher (1922, p. 420) called it the *commensurability test*.

- T11: *Time reversal test*: $P(p^0,p^1,q^0,q^1)=1/P(p^1,p^0,q^1,q^0)$.

That is, if the data for periods 0 and 1 are interchanged, then the resulting price index should equal the reciprocal of the original price index. We have already encountered this test: see (15) above. Obviously, in the one good case when the price index is simply the single price ratio, this test is satisfied (as are all of the other tests listed in this section). When the number of goods is greater than one, many commonly used price indexes fail this test; for example, the Laspeyres and Paasche price indexes, P_L and P_P defined earlier by (8) and (9) above, both *fail* this fundamental test. The concept of the test was due to Pierson (1896, p. 128), who was so upset by the fact that many of the commonly used index number formulae did not satisfy this test that he proposed that the entire concept of an index number should be abandoned. More formal statements of the test were made by Walsh (1901, p. 368; 1921b, p. 541) and Fisher (1911, p. 534; 1922, p. 64).
Our next two tests are more controversial, since they are not necessarily consistent with the economic approach to index number theory. However, these tests are quite consistent with the weighted stochastic approach to index number theory discussed in Section 3 above.

- T12: *Quantity reversal test* (quantity weights symmetry test): $P(p^0,p^1,q^0,q^1)=P(p^0,p^1,q^1,q^0)$.

That is, if the quantity vectors for the two periods are interchanged, then the price index remains invariant. This property means that if quantities are used to weight the prices in the index number formula, then the period 0 quantities q^0 and the period 1 quantities q^1 must enter the formula in a symmetric or even-handed manner. Funke and Voeller (1978, p. 3) introduced this test; they called it the *weight property*.
The next test proposed by Diewert (1992a, p. 218) is the analogue to T12 applied to quantity indexes:

- T13: *Price reversal test* (price weights symmetry test):

$$\left\{\sum_{i=1}^{N} p_i^1 q_i^1 \Big/ \sum_{i=1}^{N} p_i^0 q_i^0\right\} \Big/ P(p^0, p^1, q^0, q^1) = \left\{\sum_{i=1}^{N} p_i^0 q_i^1\right\} \Big/ P(p^1, p^0, q^0, q^1).$$

Thus, if we use (5) to define the quantity index Q in terms of the price index P, then it can be seen that T13 is equivalent to the following property for the associated quantity index Q:

$$Q(p^0, p^1, q^0, q^1) = Q(p^1, p^0, q^0, q^1). \tag{25}$$

That is, if the price vectors for the two periods are interchanged, then the quantity index remains invariant. Thus if prices for the same good in the two periods are used to weight quantities in the construction of the quantity index, then property T13 implies that these prices enter the quantity index in a symmetric manner.
The next three tests are mean value tests. The following test was proposed by Eichhorn and Voeller (1976, p. 10):

- T14: *Mean value test for prices*:

$$\min_i(p_i^1 / p_i^0 : i = 1, \ldots, N) \leq P(p^0, p^1, q^0, q^1) \leq \max_i(p_i^1 / p_i^0 : i = 1, \ldots, N).$$

That is, the price index lies between the minimum price ratio and the maximum price ratio. Since the price index is supposed to be some sort of an average of the N price ratios, p_i^1 / p_i^0, it seems essential that the price index P satisfy this test.
The next test proposed by Diewert (1992a, p. 219) is the analogue to T14 applied to quantity indexes:

- T15: *Mean value test for quantities*:

$$\min_i(q_i^1 / q_i^0 : i = 1, \ldots, n) \leq \left\{V^1 / V^0\right\} / P(p^0, p^1, q^0, q^1) \leq \max_i(q_i^1 / q_i^0 : i = 1, \ldots, n)$$

where V^t is the period t value aggregate $V^t \equiv \Sigma_{n=1}^N p_n^t q_n^t$ for t=0,1. Using (5) to define the quantity index Q in terms of the price index P, we see that T15 is equivalent to the following property for the associated quantity index Q:

$$\min_i(q_i^1 / q_i^0 : i = 1, \ldots, N) \leq Q(p^0, p^1, q^0, q^1) \leq \max_i(q_i^1 / q_i^0 : i = 1, \ldots, N). \tag{26}$$

That is, the implicit quantity index Q defined by P lies between the minimum and maximum rates of growth q_i^1 / q_i^0 of the individual quantities.
In Section 2, it was argued that it was very reasonable to take an average of the Laspeyres and Paasche price indexes as a single best measure of overall price change. This point of view can be turned into a test:

- T16: *Paasche and Laspeyres bounding test*: The price index P lies between the Laspeyres and Paasche indexes, P_L and P_P, defined by (8) and (9) above.

Bowley (1901, p. 227) and Fisher (1922, p. 403) both endorsed this property for a price index.

Our final four tests are monotonicity tests; that is, how should the price index $P(p^0,p^1,q^0,q^1)$ change as any component of the two price vectors p^0 and p^1 increases or as any component of the two quantity vectors q^0 and q^1 increases.

- T17: *Monotonicity in current prices*: $P(p^0,p^1,q^0,q^1)<P(p^0,p^2,q^0,q^1)$ if $p^1<p^2$.

That is, if some period 1 price increases, then the price index must increase, so that $P(p^0,p^1,q^0,q^1)$ is increasing in the components of p^1. This property was proposed by Eichhorn and Voeller (1976, p. 23) and it is a very reasonable property for a price index to satisfy.

- T18: *Monotonicity in base prices*: $P(p^0,p^1,q^0,q^1)>P(p^2,p^1,q^0,q^1)$ if $p^0<p^2$.

That is, if any period 0 price increases, then the price index must decrease, so that $P(p^0,p^1,q^0,q^1)$ is decreasing in the components of p^0 . This very reasonable property was also proposed by Eichhorn and Voeller (1976, p. 23).

- T19: *Monotonicity in current quantities*: *if $q^1<q^2$, then*

$$\left\{\sum_{i=1}^{N} p_i^1 q_i^1 \Big/ \sum_{i=1}^{N} p_i^0 q_i^0\right\} \Big/ P(p^0,\ p^1,\ q^0,\ q^1) < \left\{\sum_{i=1}^{N} p_i^1 q_i^2 \Big/ \sum_{i=1}^{N} p_i^0 q_i^0\right\} \Big/ P(p^0,\ p^1,\ q^0,\ q^2).$$

T20: *Monotonicity in base quantities: if $q^0<q^2$, then*

$$\left\{\sum_{i=1}^{N} p_i^1 q_i^1 \Big/ \sum_{i=1}^{N} p_i^0 q_i^0\right\} \Big/ P(p^0,\ p^1,\ q^0,\ q^1) > \left\{\sum_{i=1}^{N} p_i^1 q_i^1 \Big/ \sum_{i=1}^{N} p_i^0 q_i^2\right\} \Big/ P(p^0,\ p^1,\ q^2,\ q^1).$$

If we define the implicit quantity index Q that corresponds to P using (1), we find that T19 translates into the following inequality involving Q:

$$Q(p^0,\ p^1,\ q^0,\ q^1) < Q(p^0,\ p^1,\ q^0,\ q^2) \text{ if } q^1 < q^2. \tag{27}$$

That is, if any period 1 quantity increases, then the implicit quantity index Q that corresponds to the price index P must increase. Similarly, we find that T20 translates into:

$$Q(p^0, p^1, q^0, q^1) > Q(p^0, p^1, q^2, q^1) \text{ if } q^0 < q^2. \tag{28}$$

That is, if any period 0 quantity increases, then the implicit quantity index Q must decrease. Tests T19 and T20 are due to Vogt (1980, p. 70).

Diewert (1992a, p. 221) showed that the only index number formula $P(p^0,p^1,q^0,q^1)$ which satisfies tests T1–T20 is the Fisher ideal price index P_F defined earlier by (14), as the geometric mean of the Laspeyres and Paasche price indexes.

P_F satisfies yet another test, T21, which was Fisher's (1921, p. 534; 1922, pp. 72–81) third reversal test (the other two being T9 and T11):

- T21: *Factor reversal test* (functional form symmetry test):

$$P(p^0, p^1, q^0, q^1)P(q^0, q^1, p^0, p^1) = \sum_{i=1}^{N} p_i^1 q_i^1 \Big/ \sum_{i=1}^{N} p_i^0 q_i^0.$$

A justification for this test is the following one: if $P(p^0,p^1,q^0,q^1)$ is a good functional form for the price index, then if we reverse the roles of prices and quantities, $P(q^0,q^1,p^0,p^1)$ ought to be a good functional form for a quantity index (which seems to be a correct argument) and thus the product of the price index $P(p^0,p^1,q^0,q^1)$ and the quantity index $Q(p^0,p^1,q^0,q^1)=P(q^0,q^1,p^0,p^1)$ ought to equal the value ratio, V^1/V^0. The second part of this argument does not seem to be valid and thus many researchers over the years have objected to the factor reversal test. However, if one is willing to embrace T21 as a basic test, Funke and Voeller (1978, p. 180) showed that the only index number function $P(p^0,p^1,q^0,q^1)$ which satisfies T1 (positivity), T11 (time reversal test), T12 (quantity reversal test) and T21 (factor reversal test) is the Fisher ideal index P_F defined by (14).

Other characterizations of the Fisher price index can be found in Funke and Voeller (1978) and Balk (1985; p. 1995).

The Fisher price index P_F satisfies all 20 of the tests listed above. Which tests do other commonly used price indexes satisfy? Recall the Laspeyres index P_L defined by (8), the Paasche index P_P defined by (9) and the Törnqvist–Theil index P_T defined by (22). Straightforward computations show that the Paasche and Laspeyres price indexes fail only the three reversal tests, T11, T12 and T13. Since the quantity and price reversal tests, T12 and T13, are somewhat controversial and hence can be discounted, the test performance of P_L and P_P seems at first sight to be quite good. However, the failure of the time reversal test, T11, is a severe limitation associated with the use of these indexes.

The Törnqvist–Theil price index P_T fails nine tests: T4 (the fixed basket test), the quantity and price reversal tests T12 and T13, T15 (the mean value test for quantities), T16 (the Paasche and Laspeyres bounding test) and the four monotonicity tests T17 to T20. Thus the Törnqvist–Theil index is subject to a rather high failure rate from the perspective of this particular axiomatic approach to index number theory.

However, it could be argued that the list of tests or axioms that was used to establish the superiority of the Fisher ideal index might have been chosen to favour this index. Thus Diewert (2004), following the example of Walsh (1901, pp.104–05) and Vartia (1976), developed a set of axioms for price indexes of the form $P(p^0,p^1,v^0,v^1)$ where v^0 and v^1 are vectors of expenditures on the N commodities in the index and these vectors replace the quantity vectors q^0 and q^1 as weighting vectors for the prices. In this new axiomatic framework, the Törnqvist–Theil index P_T emerged as the best.

The consistency and independence of various bilateral index number tests was studied in some detail by Eichhorn and Voeller (1976). Our conclusion at this point echoes that of Frisch (1936): the test approach to index number theory, while extremely useful, does not lead to a single unique index number formula. However, two test

approaches that take alternative approaches to the methods for weighting prices do lead to the Fisher and Törnqvist–Theil indexes as the best in their respective axiomatic frameworks.

For additional material on the test approach to bilateral index number theory, see Balk (1995), Reinsdorf and Dorfman (1999), Balk and Diewert (2001), Vogt and Barta (1997) and Reinsdorf (2007).

In the following three sections, we consider various economic approaches to index number theory. In the economic approach to price index theory, quantity vectors are no longer regarded as being exogenous variables; rather, they are regarded as solutions to various economic optimization problems.

5 The economic approach to price indexes

Before a definition of a microeconomic price index is presented, it is necessary to make a few preliminary definitions.

Let $F(q)$ be a function of N variables, $q \equiv (q_1, \ldots, q_N)$. In the consumer context, F represents a consumer's preferences; i.e. if $F(q^2)>F(q^1)$, then the consumer prefers the commodity vector q^2 over q^1. In this context, F is called a *utility function*. In the producer context, $F(q)$ might represent the output that could be produced using the input vector q. In this context, F is called a *production function*. In order to cover both contexts, we follow the example of Diewert (1976) and call F an *aggregator function*.

Suppose the consumer or producer faces prices $p \equiv (p_1, \ldots, p_N)$ for the N commodities. Then the economic agent will generally find it is useful to minimize the cost of achieving at least a given utility or output level u; we define the *cost function* or *expenditure function* C as the solution to this minimization problem:

$$C(u, p) \equiv \min_q \{ p \cdot q : F(q) \geq u \} \tag{29}$$

where $p \cdot q \equiv \sum_{n-1}^{N} p_n q_n$ is the inner product of the price vector p and quantity vector q.

Note that the cost function depends on $1+N$ variables; the utility or output level u and the N commodity prices in the vector p. Moreover, the functional form for the aggregator function F completely determines the functional form for C.

We say that an aggregator function is *neoclassical* if F is: (i) continuous, (ii) positive; i.e. $F(q)>0$ if $q \gg 0_N$ and (iii) linearly homogeneous; that is, $F(\lambda q)=\lambda F(q)$ if $\lambda>0$.

If F is neoclassical, then the corresponding cost function $C(u, p)$ equals u times the unit cost function, $c(p) \equiv C(1, p)$, where $c(p)$ is the minimum cost of producing one unit of utility or output; that is,

$$C(u, p) = uC(1, p) = uc(p). \tag{30}$$

Shephard (1953) formally defined an aggregator function F to be *homothetic* if there exists an increasing continuous function of one variable g such that $g[F(q)]$ is neoclassical. However, the concept of homotheticity was well known to Frisch (1936) who termed it expenditure proportionality. If F is homothetic, then its cost function C has the following decomposition:

$$C(u, p) \equiv \min_q \{ p \cdot q : F(q) \geq u \} = \min_q \{ p \cdot q : g[F(q)] \geq g(u) \} = g(u) c(p) \tag{31}$$

where $c(p)$ is the unit cost function that corresponds to $g[F(q)]$.

Let $p^0 \gg 0_N$ and $p^1 \gg 0_N$ be positive price vectors pertaining to periods or observations 0 and 1. Let $q>0_N$ be a non-negative, non-zero reference quantity vector. Then the Konüs (1924) *price index* or *cost of living index* is defined as:

$$P_K(p^0, p^1, q) \equiv C[F(q), p^1] / C[F(q), p^0]. \tag{32}$$

In the consumer (producer) context, P_K may be interpreted as follows. Pick a reference utility (output) level $u \equiv F(q)$. Then $P_K(p^0,p^1,q)$ is the minimum cost of achieving the utility (output) level u when the economic agent faces prices p^1 relative to the minimum cost of achieving the same u when the agent faces prices p^0. If N=1 so that there is only one consumer good (or input), then it is easy to show that $P_K(p_1^0, p_1^1, q_1) = p_1^1 q_1 / p_1^0 q_1 = p_1^1 / p_1^0$.

Using the fact that a cost function is linearly homogeneous in its price arguments, it can be shown that P_K has the following homogeneity property: $P_K(p^0, \lambda p^1,q)= \lambda P_K(p^0, p^1,q)$ for $\lambda>0$ which is analogous to the proportionality test T5 in the previous section. P_K also satisfies $P_K(p^1,p^0,q)=1/P_K(p^0,p^1,q)$ which is analogous to the time reversal test, T11.

Note that the functional form for P_K is completely determined by the functional form for the aggregator function F, which determines the functional form for the cost function C.

In general, P_K depends not only on the two price vectors p^0 and p^1, but also on the reference vector q. Malmquist (1953), Pollak (1983) and Samuelson and Swamy (1974) have shown that P_K is independent of q and is equal to a ratio of unit cost functions, $c(p^1) / c(p^0)$, if and only if the aggregator function F is homothetic.

If we knew the consumer's preferences or the producer's technology, then we would know F and we could construct the cost function C and the Konüs price index P_K. However, we generally do not know F or C and thus it is useful to develop *bounds* that depend on observable price and quantity data but do not depend on the specific functional form for F or C.

Samuelson (1947) and Pollak (1983) established the following bounds on P_K. Let $p^0 \gg 0_N$, and $p^1 \gg 0_N$. Then for every reference quantity vector $q>0_N$, we have

$$\min_n \left\{ p_n^1 / p_n^0 \right\} \le P_K(p^0, p^1, q) \le \max_n \left\{ p_n^1 / p_n^0 \right\}; \tag{33}$$

that is, P_K lies between the smallest and largest price ratios. Unfortunately, these bounds are usually too wide to be of much practical use.

To obtain closer bounds, we now assume that the observed quantity vectors for the two periods, $q^i \equiv (q_1^i, \ldots, q_N^i)$, i=0,1, are solutions to the producer's or consumer's cost minimization problems; that is, we assume:

$$p^i \cdot q^i = C[F(q^i), p^i], \quad p^i \gg 0_N, \quad q^i > 0_N, \quad i = 0, 1. \tag{34}$$

Given the above assumptions, we now have two natural choices for the reference quantity vector q that occurs in the definition of $P_K(p^0,p^1,q)$: q^0 or q^1. The *Laspeyres–Konüs price index* is defined as $P_K(p^0,p^1,q^0)$ and the *Paasche–Konüs price index* is defined as $P_K(p^0,p^1,q^1)$.
Under the assumption of cost minimizing behaviour (34), Konüs (1924) established the following bounds:

$$P_K(p^0, p^1, q^0) \leq p^1 \cdot q^0 / p^0 \cdot q^0 \equiv P_L(p^0, p^1, q^0, q^1); \tag{35}$$

$$P_K(p^0, p^1, q^1) \geq p^1 \cdot q^1 / p^0 \cdot q^1 \equiv P_P(p^0, p^1, q^0, q^1), \tag{36}$$

where P_L and P_P are the Laspeyres and Paasche price indexes defined earlier by (8) and (9). If in addition, the aggregator function is homothetic, then Frisch (1936) showed that for any reference vector $q>0_N$,

$$P_P \equiv p^1 \cdot q^1 / p^0 \cdot q^1 \leq P_K(p^0, p^1, q) \leq p^1 \cdot q^0 / p^0 \cdot q^0 \equiv P_L. \tag{37}$$

In the consumer context, it is unlikely that preferences will be homothetic; hence the bounds (37) cannot be justified in general. However, Konüs (1924) showed that bounds similar to (37) would hold even in the general non-homothetic case, provided that we choose a reference vector $q \equiv \lambda q^0 + (1-\lambda) q^1$ which is a λ, $(1-\lambda)$ weighted average of the two observed quantity points. Specifically, Konüs showed that there exists a λ between 0 and 1 such that if $P_P \leq P_L$, then

$$P_P \leq P_K[p^0, p^1, \lambda q^0 + (1-\lambda) q^1] \leq P_L \tag{38}$$

or if $P_P>P_L$, then

$$P_L \leq P_K[p^0, p^1, \lambda q^0 + (1-\lambda) q^1] \leq P_P. \tag{39}$$

The bounds on the microeconomic price index P_K given by (37) in the homothetic case and (38)–(39) in the non-homothetic case are the best bounds that we can obtain without making further assumptions on F. In the time series context, the bounds given by (38) or (39) are usually quite satisfactory: the Paasche and Laspeyres price indexes for consecutive time periods will usually differ by less than one per cent (and hence taking the Fisher geometric average will generally suffice for most practical purposes). However, in the cross-section context where the observations represent, for example, production data for two producers in the same industry but in different regions, the bounds are often not very useful since P_L and P_P can differ by 50 per cent or more in the cross-sectional context: see Ruggles (1967) and Hill (2006a).

For generalizations of the above single household theory to many households, see Pollak (1980, p. 276; 1981, p. 328), Diewert (1983a; 2001) and in ILO (2004, ch. 18).

In Section 7, we will make additional assumptions on the aggregator function F or its cost function dual C that will enable us to determine P_K exactly. Before we do this, in the next section we will define various quantity indexes that have their origins in microeconomic theory.

6 Economic approaches to quantity indexes

In the one commodity case, a natural definition for a quantity index is q_1^1 / q_1^0, the ratio of the single quantity in period 1 to the corresponding quantity in period 0. This ratio is also equal to the expenditure ratio, $p_1^1 q_1^1 / p_1^0 q_1^0$, divided by the price ratio, p_1^1 / p_1^0. This suggests that in the N commodity case a reasonable definition for a quantity index would be the expenditure ratio divided by the Konüs price index, P_K. This type of index was suggested by Pollak (1983). Thus the *Konüs–Pollak quantity index*, Q_K, is defined by:

$$Q_K(p^0, p^1, q^0, q^1, q) \equiv p^1 \cdot q^1 / p^0 \cdot q^0 \; P_K(p^0, p^1, q) = \left\{C[F(q^1), p^1] / C[F(q), p^1]\right\} / \left\{C[F(q^0), p^0] / C[F(q), p^0]\right\} \tag{40}$$

where the second line follows from the definition of P_K, (32), and the assumption of cost minimizing behaviour in the two periods, (34).

The definition of Q_K depends on the reference vector q which appears in the definition of P_K. The general definition of Q_K simplifies considerably if we choose the reference q to be q^0 or q^1. Thus define the *Laspeyres–Konüs* quantity index as

$$Q_K(p^0, p^1, q^0, q^1, q^0) \equiv C[F(q^1), p^1] / C[F(q^0), p^1] \tag{41}$$

and the *Paasche–Konüs* quantity index as

$$Q_K(p^0, p^1, q^0, q^1, q^1) \equiv C[F(q^1), p^0] / C[F(q^0), p^0].$$

(42)

The indexes defined by (41) and (42) are special cases of another class of quantity indexes. For any reference price vector $p \gg 0_N$, define the *Allen* (1949) *quantity index* by

$$Q_A(q^0, q^1, p) \equiv C[F(q^1), p] / C[F(q^0), p].$$
(43)

If p is chosen to be p^0, (43) becomes (42) and if $p=p^1$, then (43) becomes (41).

Using the properties of cost functions, it can be shown that if $F(q^1) \geq F(q^0)$, then $Q_A(q^0, q^1, p) \geq 1$, while if $F(q^1) \leq F(q^0)$, then $Q_A(q^0, q^1, p) \leq 1$. Thus the Allen quantity index correctly indicates whether the commodity vector q^1 is larger or smaller than q^0. It can also be seen that Q_A satisfies a counterpart to the time reversal test; that is, $Q_A(q^1,q^0,p)=1/Q_A(q^0,q^1,p)$.

Just as the price index P_K depended on the unobservable aggregator function, so also do the quantity indexes Q_K and Q_A. Thus it is useful to develop bounds for the quantity indexes that do not depend on the particular functional form for F.

Samuelson (1947) and Allen (1949) established the following bounds for (41) and (42):

$$Q_A(q^0, q^1, p^0) = Q_K(p^0, p^1, q^0, q^1, q^1) \leq p^0 \cdot q^1 / p^0 \cdot q^0 \equiv Q_L;$$
(44)

$$Q_A(q^0, q^1, p^0) = Q_K(p^0, p^1, q^0, q^1, q^0) \geq p^1 \cdot q^1 / p^1 \cdot q^0 \equiv Q_P.$$
(45)

Note that the observable *Laspeyres* and *Paasche quantity indexes*, Q_L and Q_P, appear on the right hand sides of (44) and (45).

Diewert (1981), utilizing some results of Pollak (1983) and Samuelson and Swamy (1974), established the following results: if the underlying aggregator function F is neoclassical and (32) holds, then for all $p \gg 0_N$ and $q \gg 0_N$,

$$Q_P \leq Q_A(q^0, q^1, p) = Q_K(p^0, p^1, q^0, q^1, q) = F(q^1) / F(q^0) \leq Q_L.$$
(46)

Thus if the aggregator function F is neoclassical, then the Allen quantity index for all reference vectors p equals the Konüs quantity index for all reference quantity vectors q, which in turn equals the ratio of aggregates, $F(q^1)/F(q^0)$. Moreover, Q_A and Q_K are bounded from below by the Paasche quantity index Q_P, and bounded from above by the Laspeyres quantity index Q_L in the neoclassical case.

In the general non-homothetic case, Diewert (1981) showed that there exists a λ between 0 and 1 such that $Q_K(p^0,p^1,q^0,q^1, \lambda q^0+(1-\lambda)q^1]$ lies between Q_P and Q_L and there exists a λ^* between 0 and 1 such that $Q_A(q^0,q^1, \lambda^* p^0+(1-\lambda^*)p^1)$ also lies between Q_P and Q_L. Thus the observable Paasche and Laspeyres quantity indexes bound both the Konüs quantity index and the Allen quantity index, provided that we choose appropriate reference vectors between q^0 and q^1 and p^0 and p^1 respectively.

Using the linear homogeneity property of the cost function in its price arguments, we can show that the Konüs price index has the desirable homogeneity property, $P_K(p^0, \lambda p^0,q)=\lambda$ for all $\lambda>0$; that is, if period 1 prices are proportional to period 0 prices, then P_K equals this common proportionality factor. It would be desirable for an analogous homogeneity property to hold for quantity indexes. Unfortunately, it is not in general true that $Q_K(q^0, \lambda q^0,p^0,p^1,q)=\lambda$ or that $Q_A(q^0, \lambda q^0,p)=\lambda$. Thus we turn to a third economic approach to defining a quantity index which has the desirable quantity proportionality property.

Let q^1 and q^2 be the observable quantity vectors in the two situations as usual, let $F(q)$ be an increasing, continuous aggregator function, and let $q \gg 0$ be a reference quantity vector. Then the *Malmquist* (1953) *quantity index* Q_M is defined as:

$$Q_M(q^0, q^1, q) \equiv D[F(q), q^1] \,/\, D[F(q), q^0] \tag{47}$$

where $D(u, q^t) \equiv \max_k \left\{k: F(q^t / k) \geq u, \ \ k > 0\right\}$ is the *deflation* or *distance function* which corresponds to F. Thus $D[F(q),q^1]$ is the biggest number which will just deflate the quantity vector q^1 onto the boundary of the utility (or production) possibilities set $\{z: F(z) \geq F(q)\}$ indexed by the reference quantity vector q while $D[F(q),q^0]$ is the biggest number which will just deflate the quantity vector q^0 onto the set $\{z: F(z) \geq F(q)\}$ and Q_M is the ratio of these two deflation factors. Note that there is no optimization problem involving prices in the definition of the Malmquist quantity index, but the definition of the distance function involves certain deflation problems that can be interpreted as technical efficiency optimization problems.

Q_M depends on the unobservable aggregator function F and as usual, we are interested in bounds for Q_M.

Diewert (1981) showed that Q_M satisfied bounds analogous to (33); that is,

$$\min_n \left\{q_n^1 / q_n^0\right\} \leq Q_M(q^0, q^1, q) \leq \max_n \left\{q_n^1 / q_n^0\right\}. \tag{48}$$

As noted above, the assumption of cost minimizing behaviour is not required in order to define the Malmquist quantity index or to establish the bounds (46). However, in order to establish the following bounds due to Malmquist (1953) for Q_M, we do need the assumption of cost-minimizing behaviour (32) for the two periods under consideration, and we require the reference vector q to be q^0 or q^1:

$$Q_M(q^0, q^1, q^0) \leq p^0 \cdot q^1 / p^0 \cdot q^0 \equiv Q_L;$$

(49)

$$Q_M(q^0, q^1, q^1) \geq p^1 \cdot q^1 / p^1 \cdot q^0 \equiv Q_P.$$
(50)

Diewert (1981) showed that, under the hypothesis of cost-minimizing behaviour, there exists a λ between 0 and 1 such that $Q_M(q^0,q^1, \lambda q^0+(1-\lambda)q^1)$ lies between Q_P and Q_L. Thus the Paasche and Laspeyres quantity indexes provide bounds for a Malmquist quantity index for some reference indifference or product surface indexed by a quantity vector which is a λ, $(1-\lambda)$ weighted average of the two observable quantity vectors, q^0 and q^1.

Pollak (1983) showed that, if F is neoclassical, then we can extend the string of equalities in (46) to include the Malmquist quantity index $Q_M(q^0,q^1,q)$, for any reference quantity vector q. Thus, in the case of a linearly homogeneous aggregator function, all three theoretical quantity indexes coincide and this common theoretical index is bounded from below by the Paasche quantity index Q_P and bounded from above by the Laspeyres quantity index Q_L.

In the general case of a non-homothetic aggregator function, our best theoretical quantity index, the Malmquist index, is also bounded by the Paasche and Laspeyres indexes, provided that we choose a suitable reference quantity vector. In order to improve upon the bounding approach, Caves, Christensen and Diewert (1982b) show that, if one is willing to assume optimizing behaviour and make certain functional form assumptions about the underlying technology, then it is possible to obtain exact expressions for the Malmquist quantity index.

We noted in the price index context that the Paasche and Laspeyres price indexes were usually quite close in the time series context. A similar remark also applies to the Paasche and Laspeyres quantity indexes. Thus taking an average of the Paasche and Laspeyres indexes, such as the Fisher price and quantity indexes, will generally approximate underlying microeconomic price and quantity indexes sufficiently accurately for most practical purposes. However, this observation does not apply to the cross-sectional context, where the Paasche and Laspeyres indexes can differ widely. In the following section, we offer another microeconomic justification for using the Fisher indexes that also applies in the context of making inter-regional and cross-country comparisons.

7 Exact and superlative indexes

Assume that the producer or consumer is maximizing a neoclassical aggregator function f subject to a budget constraint during the two periods. Under these conditions, it can be shown that the economic agent is also minimizing cost subject to a utility or output constraint. Moreover, the cost function C that corresponds to f can be written as $C[f(q), p] = f(q)c(p)$ where c is the unit cost function (see (28) above).

Suppose a bilateral price index $P(p^0,p^1,q^0,q^1)$ and the corresponding quantity index $Q(p^0,p^1,q^0,q^1)$ that satisfy (5) are given. The quantity index Q is defined to be *exact* for a neoclassical aggregator function f with unit cost dual c if for every $p^0 \gg 0_N$, $p^1 \gg 0_N$ and $q^i \gg 0_N$ which is a solution to the aggregator maximization problem $\max_q \{f(q): p^i \cdot q \leq p^i \cdot q^i\} = f(q^i) > 0$ for i=0,1, we have

$$Q(p^0, p^1, q^0, q^1) = f(q^1) / f(q^0).$$
(51)

Under the same hypothesis, the price index P is *exact* for f and c if we have

$$P(p^0, p^1, q^0, q^1) = c(p^1) / c(p^0). \tag{52}$$

In (51) and (52), the price and quantity vectors are not regarded as being independent. The p^i can be independent, but the q^i are solutions to the corresponding aggregator maximization problem involving p^i, for i=0,1. Note that, if Q is exact for a neoclassical f, then Q can be interpreted as a Konüs, Allen or Malmquist quantity index and the corresponding P defined implicitly by (5) can be interpreted as a Konüs price index.

The concept of exactness is due to Konüs and Byushgens (1926). Below, we shall give some examples of exact index number formulae. Additional examples may be found in Afriat (1972), Pollak (1983), Samuelson and Swamy (1974) and Diewert (1976; 1992b).

Konüs and Byushgens (1926) showed that Irving Fisher's ideal price index P_F defined by (14) and the corresponding quantity index Q_F defined implicitly by (5) are exact for the homogeneous quadratic aggregator function f defined by

$$f(q_1, \ldots, q_N) \equiv \left(\sum_{n-1}^{N} \sum_{m-1}^{N} a_{nm} q_n q_m\right)^{1/2} \equiv (q \cdot Aq)^{1/2} \tag{53}$$

where $A \equiv [a_{nm}]$ is a symmetric $N \times N$ matrix of constants. Thus, under the assumption of maximizing behaviour, we can show that $f(q^1) / f(q^0) = Q_F$ and $c(p^1) / c(p^0) = P_F$ where f is defined by (51) and c is the unit cost function that corresponds to f. The important point to note is that f depends on $N(N+1)/2$ unknown a_{nm} parameters but we do not need to know these parameters in order to be able to calculate $f(q^1) / f(q^0)$ and $c(p^1) / c(p^0)$.

Diewert (1976) showed that the Törnqvist–Theil price index P_T defined by (22) is exact for the unit cost function $c(p)$ defined by:

$$\ln c(p) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n \ln p_n + (1/2) \sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_{mn} \ln p_m \ln p_n \tag{54}$$

where the parameters α_n and α_{mn} satisfy the following restrictions:

$$\sum_{n=1}^{N} \alpha_n = 1, \ \sum_{n=1}^{N} \alpha_{mn} = 0 \text{ for } m = 1, \ldots, N \text{ and } \alpha_{mn} = \alpha_{nm} \text{ for all } m, n.$$

(55)

Thus we may calculate $c(p^1)/c(p^0) = P_T$ and $f(q^1)/f(q^0) = p^1 \cdot q^1 / p^0 \cdot q^0 P_T \equiv Q_T$ where c is the unit cost function defined by (54), f is the aggregator function which corresponds to this c, and Q_T is the implicit Törnqvist–Theil quantity index. Note that we do not have to know the parameters α_n and α_{mn} in order to evaluate $c(p^1)/c(p^0)$ and $f(q^1)/f(q^0)$.

The unit cost function defined by (54) is the *translog* unit cost function defined by Christensen, Jorgenson and Lau (1971). Since P_T is exact for this translog functional form, P_T is sometimes called the *translog price index*.

Define the following family of quantity indexes Q_r that depend on a number, $r \neq 0$:

$$Q_r(p^0, p^1, q^0, q^1) \equiv \left[\sum_{n=1}^{N} s_n^0 (q_n^1 / q_n^0)^{r/2}\right]^{1/r} \left[\sum_{m=1}^{N} s_m^1 (q_m^1 / q_m^0)^{-r/2}\right]^{-1/r}$$

(56)

where $s_n^i \equiv p_n^i q_n^i / p^i \cdot q^i$ is the period i expenditure share for good n. For each $r \neq 0$, define the corresponding implicit price index by:

$$P_r^*(p^0, p^1, q^0, q^1) \equiv p^1 \cdot q^1 / p^0 \cdot q^0 \; Q_r(p^0, p^1, q^0, q^1).$$

(57)

A quick algebraic calculation will show that when r=2, $P_2^* = P_F$, the Fisher price index defined by (14) and when r equals 1, P_1^* equals:

$$P_1^* = \sum_{n=1}^{N} p_n^1 (q_n^0 q_n^1)^{1/2} / \sum_{m=1}^{N} p_m^0 (q_m^0 q_m^1)^{1/2} = P_W$$

(58)

where P_W is the Walsh price index defined earlier by (17).

Diewert (1976) showed that Q_r and P_r^* are exact for the *quadratic mean of order r aggregator function* f_r defined as follows:

$$f_r(q_1, \ldots, q_N) \equiv \left(\sum_{m=1}^{N} \sum_{n=1}^{N} a_{mn} q^{r/2} q_n^{r/2} \right)^{1/r} \tag{59}$$

where $A \equiv [a_{mn}]$ is a symmetric matrix of constants. Thus the Walsh and Fisher price indexes, P_W and P_F, are exact for $f_1(q)$ and $f_2(q)$ respectively, defined by (59) when r=1 and 2.

Diewert (1974) defined a linearly homogeneous function f of N variables to be *flexible* if it could provide a second-order approximation to an arbitrary twice continuously differentiable linearly homogeneous function. It can be shown that f defined by (53), c defined by (54) and (55) and f_r defined by (59) for each $r \neq 0$ are all examples of flexible functional forms.

Let the price and quantity indexes P and Q satisfy the product test equality, (5). Then Diewert (1976) defined P and Q to be *superlative indexes* if either P is exact for a flexible unit cost function c or Q is exact for a flexible aggregator function f. Thus P_F, P_W, P_T and P_r^* are all superlative price indexes. Thus from the viewpoint of the economic approach to index number theory, all of these indexes can be judged to be equally good.

At this point, it is useful to review the various approaches to bilateral index number theory discussed in the previous sections. In Section 2, it was found that the best average basket approaches led to the Fisher or Walsh price indexes. In Section 3, the index from the viewpoint of the stochastic approach was the Törnqvist–Theil index. In Section 4, the test approach led to the Fisher or the Törnqvist–Theil indexes as being best. Finally, in this section, the economic approach led to the Fisher, Walsh and Fisher or the Törnqvist–Theil indexes as being equally good. *Thus all four major approaches to index number theory led to the same three indexes as being best*. But which one of these three formulae, P_F, P_W and P_T, should we choose? Fortunately, it does not matter very much which of these formulae we choose to use in applications; they will all give the same answer to a reasonably high degree of approximation. Diewert (1978, p. 889) showed that all known superlative index number formulae approximate each other to the second order when each index is evaluated at an equal price and quantity point. This means the P_F, P_W, P_T and each P_r^* have the same first and second order partial derivatives with respect to all 4N arguments when the derivatives are evaluated at a point where $p^0=p^1$ and $q^0=q^1$. A similar string of equalities also holds for the corresponding implicit quantity indexes defined using the product test (5). In fact, these derivative equalities are still true provided that $p^1 = \lambda p^0$ and $q^1 = \mu q^0$ for any numbers $\lambda > 0$ and $\mu > 0$. However, although Diewert's approximation result is mathematically true, Hill (2006) has shown that superlative indexes of the form P_r^* for r very large in magnitude do not necessarily empirically approximate the standard superlative indexes P_F, P_W and P_T very closely. But these standard superlative indexes typically approximate each other to something less than 0.2 per cent in the time series context and to about two per cent in the cross-section context; see Fisher (1922), Ruggles (1967), Diewert (1978, pp. 894–5) and Hill (2006) for empirical evidence on this point.

Diewert (1978) also showed that the Paasche and Laspeyres indexes approximate the superlative indexes to the first order at an equal price and quantity point. In the time series context, for adjacent periods, the Paasche and Laspeyres price indexes typically differ by less than 0.5 per cent; hence these indexes may provide acceptable approximations to a superlative index.

After consideration of the case of two observations at length, the many-observation case is considered in the following two sections.

8 The fixed base versus the chain principle

In this section, the merits of using the chain system for constructing price indexes in the time series context versus using the fixed base system are discussed.

The chain system, introduced independently into the economics literature by Lehr (1885, pp. 45–6) and Marshall (1887, p. 373), measures the change in prices going from one period to a subsequent period using a bilateral index number formula involving the prices and quantities pertaining to the two adjacent periods. These one period rates of change (the links in the chain) are then cumulated to yield the relative levels of prices over the entire period under consideration. Thus, if the bilateral price index is P, the chain system generates the following pattern of price levels for the first three periods:

$$1,\ P(p^0,\ p^1,\ q^0,\ q^1),\ \ P(p^0,\ p^1,\ q^0,\ q^1)\ P(p^1,\ p^2,\ q^1,\ q^2). \tag{60}$$

On the other hand, the fixed base system of price levels using the same bilateral index number formula P simply computes the level of prices in period t relative to the base period 0 as $P(p^0,p^t,q^0,q^t)$. Thus the fixed base pattern of price levels for periods 0,1 and 2 is:

$$1,\ P(p^0,\ p^1,\ q^0,\ q^1),\ \ P(p^0,\ p^2,\ q^0,\ q^2). \tag{61}$$

Due to the difficulties involved in obtaining current period information on quantities (or equivalently, on expenditures), as was indicated in Section 2, many statistical agencies loosely base their consumer price index on the use of the Laspeyres formula and the fixed base system. Therefore, it is of some interest to look at some of the possible problems associated with the use of fixed base Laspeyres indexes.

The main problem with the use of the fixed base Laspeyres index is that the period 0 fixed basket of commodities that is being priced out in period t can often be quite different from the period t basket. Thus, if there are systematic *trends* in at least some of the prices and quantities in the index basket, the fixed base Laspeyres price index $P_L(p^0,p^t,q^0,q^t)$ can be quite different from the corresponding fixed base Paasche price index, $P_P(p^0,p^t,q^0,q^t)$. This means that both indexes are likely to be an inadequate representation of the movement in average prices over the time period under consideration.

As Hill (1988) noted, the fixed base Laspeyres quantity index cannot be used for ever: eventually, the base period quantities q^0 are so far removed from the current period quantities q^t that the base must be changed. Chaining is merely the limiting case where the base is changed each period.

The main advantage of the chain system is that under normal conditions, chaining will reduce the spread between the Paasche and Laspeyres indexes; see Diewert (1978, p. 895) and Hill (1988; 1993, pp. 387–8). These two indexes each provide an asymmetric perspective on the amount of price change that has occurred between the two periods under consideration, and it could be expected that a single point estimate of the aggregate price change should lie between these two estimates. Thus the use of either a chained Paasche or Laspeyres index will usually lead to a smaller difference between the two and hence to estimates that are closer to the ‘truth’.

Hill (1993, p. 388), drawing on the earlier research of Szulc (1983) and Hill (1988, pp. 136–7), noted that it is not appropriate to use the chain system when prices oscillate or ‘bounce’, to use Szulc's (1983, p. 548) term. This phenomenon can occur in the context of regular seasonal fluctuations or in the context of price wars. However, in the context of roughly monotonically changing prices and quantities, Hill (1993, p. 389) recommended the use of chained symmetrically weighted indexes. The Fisher, Walsh and Törnqvist–Theil indexes are examples of symmetrically weighted indexes.

It is possible to be more precise about the conditions under which one should chain or not chain. Following arguments due to Walsh (1901, p. 206; 1921a, pp. 84–5) and Fisher (1911, pp. 204 and 423–4), one should chain if the prices and quantities pertaining to adjacent periods are *more similar* than the prices and quantities of more distant periods, since this strategy will lead to a narrowing of the spread between the Paasche and Laspeyres indexes at each link. Of course, one needs a measure of how similar the prices and quantities pertaining to two periods are. The similarity measures could be *relative* ones or *absolute* ones. In the case of absolute comparisons, two vectors of the same dimension are similar if they are identical and dissimilar otherwise. In the case of relative comparisons, two vectors are similar if they are proportional and dissimilar if they are non-proportional. Once a similarity measure has been defined, the prices and quantities of each period can be compared with each other using this measure, and a ‘tree’ or path that links all the observations can be constructed where the most similar observations are compared with each other using a bilateral index number formula. Fisher (1922, pp. 271–6) informally suggested this strategy. However, the more recent literature on this approach is due to Robert Hill. Initially, Hill (1999a; 1999b; 2001) defined the price structures between the two countries to be more dissimilar the bigger is the spread between P_L and P_P, that is, the bigger is $\max\{P_L/P_P,\ P_P/P_L\}$. The problem with this measure of dissimilarity in the price structures of the two countries is that it could be the case that $P_L=P_P$ (so that the Hill measure would register a maximal degree of similarity) but p^0 could be very different from p^t. Thus there is a need for a more systematic study of similarity (or dissimilarity) measures in order to

pick the best one that could be used as an input into Hill's (1999a; 1999b; 2001; 2004; 2006b; 2007) spanning tree algorithm for linking observations; see Diewert (2007a).

The method of linking observations explained in the previous paragraph based on the similarity of the price and quantity structures of any two observations may not be practical in a statistical agency context since the addition of a new period may lead to a reordering of the previous links. However, the above ‘scientific’ method for linking observations may be useful in deciding whether chaining is preferable or whether fixed base indexes should be used while making month-to-month comparisons within a year.

Some index number theorists have objected to the chain principle on the grounds that it has no counterpart in the spatial context:

They [chain indexes] only apply to intertemporal comparisons, and in contrast to direct indices they are not applicable to cases in which no natural order or sequence exists. Thus the idea of a chain index for example has no counterpart in interregional or international price comparisons, because countries cannot be sequenced in a ‘logical’ or ‘natural’ way (there is no *k*+1 nor *k*–1 country to be compared with country *k*). (von der Lippe, 2001, p. 12)

This is of course correct but the approach of Robert Hill leads to a ‘natural’ set of spatial links. Applying the same approach to the time series context will lead to a set of links between periods which may not be month-to-month but it will in many cases justify year-over-year linking of the data pertaining to the same month.

It is of some interest to determine if there are index number formulae that give the same answer when either the fixed base or chain system is used. If we compare the sequence of chain indexes defined by (60) above with the corresponding fixed base indexes defined by (61), it can be seen that we will obtain the same answer in all three periods if the index number formula *P* satisfies the following functional equation for all price and quantity vectors:

$$P(p^0, p^2, q^0, q^2) = P(p^0, p^1, q^0, q^1)\ P(p^1, p^2, q^1, q^2). \tag{62}$$

If a bilateral index number formula *P* satisfies (62), then *P* satisfies the *circularity test*; see Westergaard (1890, pp. 218–19) and Fisher (1922, p. 413).

If it is assumed that the index number formula *P* satisfies certain properties or tests in addition to the circularity test above, then Funke, Hacker and Voeller (1979) showed that *P* must have the following functional form due originally to Konüs and Byushgens (1926, pp. 163–6):

$$\ln P_{KB}(p^0, p^1, q^0, q^1) \equiv \sum_{i=1}^{N} \alpha_i \ln\ (p_i^1 / p_i^0) \tag{63}$$

where the *N* constants α_i satisfy the following restrictions:

$$\sum_{i=1}^{N} \alpha_i = 1 \text{ and } \alpha_i > 0 \text{ for } i = 1, \ldots, N. \tag{64}$$

Thus, under very weak regularity conditions, the only price index satisfying the circularity test is a weighted geometric average of all the individual price ratios, the weights being constant through time. This result vindicates Irving Fisher's (1922, p. 274) intuition when he asserted that ‘the only formulae which conform perfectly to the circular

test are index numbers which have *constant weights*…'.

The problem with the indexes defined by Konüs and Byushgens is that the individual price ratios, p_n^1 / p_n^0, have weights that are *independent* of the economic importance of commodity *n* in the two periods under consideration. Put another way, these price weights are independent of the quantities of commodity *n* consumed or the expenditures on commodity *n* during the two periods. Hence, these indexes are not really suitable for use by statistical agencies at higher levels of aggregation when expenditure share information is available.

The above results indicate that it is not useful to ask that the price index *P* satisfy the circularity test *exactly*. However, it is of some interest to find index number formulae that satisfy the circularity test to some degree of *approximation* since the use of such an index number formula will lead to measures of aggregate price change that are more or less the same no matter whether we use the chain or fixed base systems. Irving Fisher (1922, p. 284) found that deviations from circularity using his data-set and the Fisher ideal price index P_F were quite small. This relatively high degree of correspondence between fixed base and chain indexes has been found to hold for other symmetrically weighted formulae like the Walsh index P_W defined earlier. It is possible to give a theoretical explanation for the approximate satisfaction of the circularity test in the time series context for symmetrically weighted index number formulae, such as P_F and P_W. Another symmetrically weighted formula is the Törnqvist–Theil index P_T. Alterman, Diewert and Feenstra (1999, p. 61) showed that if the logarithmic price ratios $\ln\ (p_n^t / p_n^{t-1})$ trend linearly with time *t* and the expenditure shares s_n^t also trend linearly with time, then the Törnqvist index P_T will satisfy the circularity test exactly. Since many economic time series on prices and quantities satisfy these assumptions approximately, then the Törnqvist index P_T will satisfy the circularity test approximately. As was noted earlier, the Törnqvist index generally closely approximates the symmetrically weighted Fisher and Walsh indexes, so that for many economic time series (with smooth trends) all three of these symmetrically weighted indexes will satisfy the circularity test to a high enough degree of approximation so that it will not matter whether we use the fixed base or chain principle.

Walsh (1901, p. 401; 1921a, p. 98; 1921b, p. 540) introduced the following useful variant of the circularity test:

$$1 = P(p^0, p^1, q^0, q^1)\ P(p^1, p^2, q^1, q^2)\ldots P(p^{T-1}, p^T, q^{T-1}, q^T)\ P(p^T, p^0, q^T, q^0). \tag{65}$$

The motivation for this test is the following. Use the bilateral index formula $P(p^0,p^1,q^0,q^1)$ to calculate the change in prices going from period 0 to 1, use the same formula evaluated at the data corresponding to periods 1 and 2, $P(p^1,p^2,q^1,q^2)$, to calculate the change in prices going from period 1 to 2, … , use $P(p^{T-1},p^T,q^{T-1},q^T)$ to calculate the change in prices going from period T–1 to T, introduce an artificial period T+1 that has exactly the price and quantity of the initial period 0 and use $P(p^T,p^0,q^T,q^0)$ to calculate the change in prices going from period T to 0. Finally, multiply all these indexes together, and since we end up where we started the product of all of these indexes should ideally be 1. Diewert (1993a, p. 40) called this test a *multiperiod identity test*. Note that, if T=2 (so that the number of periods is 3 in total), then Walsh's test reduces to Fisher's (1921, p. 534; 1922, p. 64) *time reversal test*.

Walsh (1901, pp. 423–33) showed how his circularity test could be used in order to evaluate how 'good' any bilateral index number formula was. What he did was invent artificial price and quantity data for five periods, and he added a sixth period that had the data of the first period. He then evaluated the right-hand side of (65) for various bilateral formula, $P(p^0,p^1,q^0,q^1)$, and determined how far from unity the results were. His best formulae had products that were close to 1. Fisher (1922, p. 284) later used this methodology as well.

This same framework is often used to evaluate the efficacy of chained indexes versus their direct counterparts. Thus if the right hand side of (65) turns out to be different from unity, the chained indexes are said to suffer from 'chain drift'. If a formula suffers from chain drift, it is sometimes recommended that fixed base indexes be used in place of chained ones. However, this advice, if accepted, would *always* lead to the adoption of fixed base indexes, provided that the bilateral index formula satisfies the identity test, $P(p^0,p^0,q^0,q^0)$=1. Thus it is not recommended that Walsh's circularity test be used to decide whether fixed base or chained indexes should be calculated. However, it is fair to use Walsh's circularity test as he originally used it, namely, as an approximate method for deciding how good a particular index number formula is. In order to decide whether to chain or use fixed base indexes, one should decide on the basis of how similar the observations being compared are, and choose the method which will best link up the most similar observations.

Robert Hill's method for linking observations can be regarded as a multilateral index number method, one which is based on a suitable bilateral formula, a measure of the similarity of any two price and quantity vectors and an algorithm for linking the observations via a path that links the most similar observations. In the following section, we review some other multilateral methods.

9 Multilateral indexes

Assume that there are I positive price vectors $p^i \equiv (p_1^i, \ldots, p_N^i)$ and I quantity vectors $q^i \equiv (q_1^i, \ldots, q_N^i)$ with $p^i \cdot q^i > 0$ for $i=1,\ldots,I$. We wish to find $2I$ positive numbers P^i (price indexes) and Q^i (quantity indexes) such that $P^i Q^i = p^i \cdot q^i$ for $i=1,\ldots,I$. The I data points (p^i, q^i) will typically be observations on production or consumption units that are separated spatially but yet are still comparable. For the sake of definiteness, we shall refer to the I data points as countries. Each commodity n is supposed to be the same across all countries. This can always be done by a suitable extension of the list of commodities.

Our first approach to the construction of a system of multilateral price and quantity indexes is based on the use of a bilateral quantity index Q. In this method, the first step is to pick the best bilateral index number formula, for example, the Fisher quantity index Q_F defined by (14) and (5) or the implicit Törnqvist–Theil quantity index Q_T defined by (22) and (5). Secondly, pick a numeraire country, say country 1, and then calculate the aggregate quantity for each country i relative to country 1 by evaluating the quantity index $Q(p^1,p^i,q^1,q^i)$. In order to put these relative quantity measures on a symmetric footing, we convert each relative to country 1 quantity measure into a share of world quantity by dividing through by $\Sigma_{k=1}^{I} Q(p^1, p^k, q^1, q^k)$. For a general numeraire country j, define the *share of world quantity for country i, using country j as the numeraire country*, by:

$$\sigma_i^j(p, q) \equiv Q(p^j, p^i, q^j, q^i) \,/\, \sum_{k=1}^{I} Q(p^j, p^k, q^j, q^k); \quad i = 1, \ldots, I, \tag{66}$$

where $p \equiv (p^1, \ldots, p^I)$ is the N by I matrix of price data and $q \equiv (q^1, \ldots, q^I)$ is the N by I matrix of quantity data. Once the numeraire country j has been chosen and the country i shares σ_i^j calculated, we may set $Q^i \equiv \sigma_i^j$ and $p^i \equiv p^i \cdot q^i / Q^i$ for $i=1,\ldots,I$. Thus we have provided a solution to the multilateral index number problem (1). Of course, one is free to renormalize the resulting P^i and Q^i if desired: all Q^i can be multiplied by a number provided all P^i are divided by this same number. Kravis (1984) called this method the *star system*, since the numeraire country plays a starring role: all countries are compared with it and it alone.

Of course, the problem with the star system for making multilateral comparisons is its lack of invariance to the choice of the numeraire or star country. Different choices for the base country will in general give rise to different indexes P^i and Q^i. This problem can be traced to the lack of circularity of the bilateral formula Q: if Q satisfies the time reversal test and the circular test for quantity indexes, then $\sigma_i^j = \sigma_i^k$ for all i, j and k; that is, the shares σ_i^j defined by (66) do not depend on the choice of the numeraire country j. However, given that the chosen best bilateral formula does not satisfy the circularity test (as is the case with Q_F and Q_T), how can we generate multilateral indexes that treat each country symmetrically?

Fisher (1922, p. 305) recognized that the simplest way of achieving symmetry was to average base specific index numbers over all possible bases. Thus define country i's share of world output $S_i(p, q)$ by

$$S_i(p, q) \equiv \sum_{j=1}^{I} \sigma_i^j(p, q) \,/\, I, \quad i = 1, \ldots, I \tag{67}$$

where the σ_i^j are defined by (66). We can now define country *i* quantities and prices by

$$Q^i \equiv S_i(p, q); \quad P^i \equiv p^i \cdot q^i / Q^i, \quad i = 1, \ldots, I. \tag{68}$$

Fisher (1922, p. 305) called this method of constructing multilateral indexes the *blend method* while Diewert (1986) called it the *democratic weights method*, since each share of world output using each country as the base is given an equal weight in the formation of the average.

Of course, there is no need to use an arithmetic average of the σ_i^j as in (67); one can use a geometric average:

$$\sigma_i(p, q) \equiv \left[\prod_{j=1}^{I} \sigma_i^j(p, q)\right]^{1/I}, \quad i = 1, \ldots, I. \tag{69}$$

Using (69), the resulting shares no longer sum to one in general, so country *i*'s share of world output is now defined as:

$$S_i(p, q) \equiv \sigma_i(p, q) / \sum_{k=1}^{I} \sigma_k(p, q), \quad i = 1, \ldots, I. \tag{70}$$

If the Fisher index Q_F is used in the definition of the σ_i^j, then

$$S_i(p, q) / S_j(p, q) = \left[\prod_{k=1}^{I} Q_F(p^k, p^i, q^k, q^i) / \prod_{m=1}^{I} Q_F(p^m, p^j, q^m, q^j)\right]^{1/I} \tag{71}$$

and in this case the multilateral method defined by (71) reduces to a method recommended by Gini (1924; 1931), Eltetö and Köves (1964) and Szulc (1964), the *GEKS method*. Instead of using the Fisher formula in (71), Caves, Christensen and Diewert (1982a) advocated the use of the (direct) Törnqvist–Theil quantity index while Diewert (1986) suggested the use of the implicit translog quantity index Q_T defined by (5) when P is P_T defined by (22), since Q_T is well defined even in the case where some quantities q_n^i are negative. We call the indexes generated by (69) and (70) for a general bilateral index Q *generalized GEKS* indexes.

When forming averages of the σ_i^j as in (67) or (69), there is no necessity to use equal weights: one can define country j's value share of world output as $\beta_j \equiv p^j \cdot q^j / \Sigma_{k=1}^{l} p^k \cdot q^k$ (this requires all prices to be measured in units of a common currency) and then we may define a plutocratic share weighted average of the σ_i^j:

$$S_i(p, q) \equiv \sum_{j=1}^{l} \beta_j(p, q)\ \sigma_i^j(p, q). \tag{72}$$

Diewert (1986) called this method of constructing multilateral indexes the *plutocratic weights method*.

Another multilateral method that is based on a bilateral index Q may be described as follows. Define

$$\sigma_i(p, q) \equiv \sum_{j=1}^{l} \left[Q(p^j, p^i, q^j, q^i)^{-1} \right]^{-1}; \quad i = 1, \ldots, l. \tag{73}$$

If there is only one commodity so that N=1 and the bilateral index Q satisfies quantity counterparts to tests T3 and T5, then $\alpha_i = \left[\Sigma_{j=1}^{l} (q^i / q^j)^{-1} \right]^{-1} = \left[\Sigma_{j=1}^{l} q^j / q^i \right]^{-1} = q^i / \Sigma_{j=1}^{l} q^j$ which is country i's share of world product. In the general case where N>1, the 'shares' α_i do not necessarily sum up to unity, so it is necessary to normalize them:

$$S_i(p, q) \equiv \alpha_i(p, q) / \sum_{k=1}^{l} \alpha_k(p, q); \quad i = 1, \ldots, l. \tag{74}$$

Diewert (1986; 1988; 1999b) called this the *own share method* for making multilateral comparisons.

The above methods for achieving consistency and symmetry rely on averaging over various bilateral index number comparisons. Fisher (1922, p. 307) realized that symmetry could be achieved by making comparisons with an average; he called this broadening the base. Thus the *average basket method* (see Walsh, 1901, p. 431; Gini, 1931, p. 8; Fisher, 1922, p. 307; Ruggles, 1967; and Diewert, 1999b, pp. 24–5) may be described as follows. The price level of country I relative to country j is set equal to

$p^j \cdot \left(\Sigma_{k=1}^{l} q^k / l\right) / p^j \cdot \left(\Sigma_{k=1}^{l} q^k / l\right)$. Now define $Q^{ji} \equiv [p^i \cdot q^i / p^j \cdot q^j] / [p^i \cdot (\Sigma_k q^k) / p^j \cdot (\Sigma_k q^k)]$ to be the implicit output of country *i* relative to *j*. Choose a *j* as a numeraire country and calculate country *i*'s share of world output as:

$$S_i(p, q) \equiv Q^{ji} / \sum_{k=1}^{l} Q^{jk} = \left(p^i \cdot q^i / p^i \cdot \sum_k q^k\right) / \sum_{m=1}^{l} \left(p^m \cdot q^m / p^m \cdot \sum_k q^k\right); i = 1, \ldots, l. \quad (75)$$

Note that the final expression for S_i does not depend on the choice of the numeraire country *j*. As usual, once the share functions, S_i, have been defined, the aggregate Q^i and P^i may be defined by (68).

A variation on the basket method due to Geary (1958) and Khamis (1972) is defined by (76)–(78) below:

$$\pi_n \equiv \sum_{i=1}^{l} p_n^i q_n^i / P^i \sum_{k=1}^{l} q_n^k, \quad n = 1, \ldots, N; \quad (76)$$

$$P^i \equiv \sum_{n=1}^{N} p_n^i q_n^i / \sum_{m=1}^{N} \pi_m q_m^i, \quad i = 1, \ldots, l; \quad (77)$$

$$Q^i \equiv p^i \cdot q^i / P^i, \quad i = 1, \ldots, l. \quad (78)$$

π_n is interpreted as an average international price for good *n*. From (77), it can be seen that P^i, the price level or purchasing power parity for country *i*, is a Paasche-like price index for country *i* except that the base prices are chosen to be the international prices π_n. The π_n and $(P^i)^{-1}$ can be solved for as a system of simultaneous linear equations (up to a scalar normalization) or the $(P^i)^{-1}$ may be determined as the components of the eigenvector that corresponds to the maximal positive eigenvalue of a certain matrix. The P^i can be normalized so that the quantities Q^i defined by (78) sum up to unity. This GK *method* for making multilateral comparisons has been widely

used in empirical applications; for example, see Kravis et al. (1975).

We have defined seven methods for making multilateral comparisons: the star method (66), the democratic (67) and plutocratic (72) weights methods, the GEKS method (71), the own share method (74), the average basket method (75) and the GK method (78). Many additional methods have been suggested; for example, see Hill (1997), Diewert (1986; 1988; 1999b), Rao (1990) and Balk (1996). How can we discriminate among them? One helpful approach would be to define a system of *multilateral tests* and then evaluate how the above methods satisfy these tests. Space does not permit the development of this approach in this short survey; for applications of this approach, see Diewert (1988; 1999b) and Balk (1996). A clear consensus on the best multilateral method has not yet emerged.

We conclude this section by looking at a stochastic or descriptive statistics approach to making multilateral comparisons: namely, Summer's (1973) country product dummy (CPD) method for making multilateral comparisons. If there are *I* countries in the comparison and *N* products, the relationship of the prices between the various countries using the CPD model is given (approximately) by the following model:

$$p_n^c \approx \alpha_c \beta_n; \quad c = 1, \ldots, I; \quad n = 1, \ldots, N; \tag{79}$$

$$\alpha_1 = 1 \tag{80}$$

where p_n^c is the price (in domestic currency) of commodity *n* in country *c*. Quantities for each commodity in each country are assumed to be measured in the same units. Equation (80) above is an identifying normalization; that is, we measure the price level of each country relative to the price level in country 1. Note that there are *IN* prices in the model and there are *I*–1+*N* parameters to ‘explain’ these prices. Note also that the basic hypothesis that is implied by (79) is that commodity prices are approximately proportional between the two countries. Taking logarithms of both sides of (79) and adding error terms leads to the following CPD regression model:

$$\ln p_n^c = \ln \alpha_c + \ln \beta_n + \varepsilon_n^c; \quad c = 1, \ldots, I; \quad n = 1, \ldots, N. \tag{81}$$

The main advantage of the CPD method for comparing prices across countries over traditional index number methods is that we can obtain *standard errors* for the country price levels $\alpha_2, \alpha_3, \ldots, \alpha_I$. This advantage of the stochastic approach to index number theory was stressed by Summers (1973) and more recently by Selvanathan and Rao (1994).

The recent literature on the CPD method notes that it is a special case of a hedonic regression model and this recent literature makes connections between weighted hedonic regressions and traditional index number formulae; see Triplett and McDonald (1977), Diewert (2003; 2005b; 2005c; 2007b), de Haan (2004a; 2004b), Silver (2003) and Silver and Heravi (2005).

10 Other aspects of index number theory

There are many important recent developments in index number theory that we cannot cover in any depth in this brief survey. Some of these developments are:

- Sampling problems and the construction of indexes at the first stage of aggregation: see Dalén (1992), Diewert (1995a), ILO (2004) and IMF (2004).
- *The treatment of seasonality*: see Turvey (1979), Balk (1980) (2005), Diewert (1983c) (1998b) (1999a), Hill (1996), Alterman, Diewert and Feenstra (1999), ILO (2004) and Armknecht and Diewert (2004).
- *The analysis of sources of bias in consumer price indexes*. This topic was greatly stimulated by the Boskin Commission Report; see Boskin et al. (1996). For additional contributions to this subject, see Diewert (1987; 1998a; Reinsdorf (1993), Schultze and Mackie (2002), Lebow and Rudd (2003), Balk and Diewert (2004) and ILO (2004).
- *Productivity indexes*. As more and more countries start programmes to measure sectoral and economy wide productivity, this topic has become more important. The original methodology for measuring productivity using index number techniques is due to Jorgenson and Griliches (1967; 1972) and it was first adopted by the U.S. Bureau of Labor Statistics (1983) and subsequently by Canada, Australia and more recently by New Zealand and Switzerland. Diewert (1976; 1983b) Caves, Christensen and Diewert (1982b), Diewert and Morrison (1986), Kohli (1990), Morrison and Diewert (1990), Balk (1998; 2003), Schreyer (2001), Diewert and Fox (2004), Diewert and Nakamura (2003) and Diewert and Lawrence (2006) all made contributions connecting productivity measurement with index number theory.
- *Contribution analysis*. Suppose an aggregate price or quantity index shows a certain change over a certain period. Many analysts want to be able to compute the contribution of price or quantity change of specific components of the overall index and the problem of precisely defining such contributions has given rise to a fairly substantial recent literature. Contributors to this literature include Diewert (1983b; 2002a), Diewert and Morrison (1986), van IJzeren (1957; 1983; 1987), Kohli (1990; 2003; 2004; 2007), Morrison and Diewert (1990), Fox and Kohli (1998) and Reinsdorf, Diewert and Ehemann (2002).
- *Quality change*. The analysis thus far has assumed that the list of commodities in the aggregate is fixed and is unchanging and thus it is not able to deal with the problem of quality change. For extensive discussions of this problem, see Triplett (2004) and the chapters on quality change in ILO (2004) and IMF (2004).
- *Index number theory in terms of differences rather than ratios*. Hicks (1941–42) noticed the similarities between measuring welfare change (difference measures) and index numbers of quantity change (ratio measures). The early literature on the difference approach dates back to Bennet (1920) and Montgomery (1929; 1937). More recent contributions to this subject may be found in Diewert (1992b; 2005a).

Since the mid-1980s interest in index number theory and economic measurement problems in general has increased. Perhaps influenced by Hill (1993), who in turn was influenced by Diewert (1976) (1978), national statistical agencies are moving towards using chained superlative indexes as their target indexes: see Moulton and Seskin (1999) and Cage, Greenlees and Jackman (2003) for US developments. International agencies have also endorsed the use of superlative indexes as target indexes: see the manuals produced by the ILO (2004) and the IMF (2004). These manuals are a useful development since they help disseminate best practices and they help to harmonize statistics across countries, leading to a higher degree of accuracy and comparability. One hopes that these positive developments will continue.

Bibliography

Afriat, S.N. 1972. The theory of international comparisons of real income and prices. In *International Comparisons of Prices and Outputs*, ed. D.J. Daley. New York: Columbia University Press.

Allen, R.G.D. 1949. The economic theory of index numbers. *Economica* NS 16, 197–203.

Alterman, W.F., Diewert, W.E. and Feenstra, R.C. 1999. *International Trade Price Indexes and Seasonal Commodities*. Washington, DC: Bureau of Labor Statistics.

Armknecht, P.A. and Diewert, W.E. 2004. Treatment of seasonal products. In *Producer Price Index Manual: Theory and Practice*, Washington, DC: International Monetary Fund.

Balk, B.M. 1980. A method for constructing price indices for seasonal commodities. *Journal of the Royal Statistical Society* Series A 143, 68–75.

Balk, B.M. 1985. A simple characterization of Fisher's price index. *Statistische Hefte* 26, 59–63.

Balk, B.M. 1995. Axiomatic price index theory: a survey. *International Statistical Review* 63, 69–93.

Balk, B.M. 1996. A comparison of ten methods for multilateral international price and volume comparisons. *Journal of Official Statistics* 12, 199–222.

Balk, B.M. 1998. *Industrial Price, Quantity and Productivity Indices*. Boston: Kluwer Academic Publishers.

Balk, B.M. 2003. The residual: on monitoring and benchmarking firms, industries, and economies with respect to productivity. *Journal of Productivity Analysis* 20, 5–47.

Balk, B.M. 2005. Annual and quarterly productivity measures. Paper presented at the Economic Measurement Group Workshop, Coogee, Australia, 12–13 December.

Balk, B.M. and Diewert, W.E. 2001. A characterization of the Törnqvist price index. *Economics Letters* 73, 279–81.

Balk, B.M. and Diewert, W.E. 2004. The Lowe consumer price index and its substitution bias. Discussion Paper No. 04–07. Department of Economics, University of British Columbia.

Bennet, T.L. 1920. The theory of measurement of changes in cost of living. *Journal of the Royal Statistics Society* 83, 455–62.

Boskin, M.J. (chair), Dullberger, E.R., Gordon, R.J., Griliches, Z. and Jorgenson, D.W. 1996. *Final Report of the Commission to Study the Consumer Price Index*. US Senate, Committee on Finance, Washington, DC: US Government Printing Office.

Bowley, A.L. 1899. Wages, nominal and real. In *Dictionary of Political Economy*, vol. 3, ed. R.H.L. Palgrave. London: Macmillan.

Bowley, A.L. 1901. *Elements of Statistics*. Westminster: P.S. King and Son.

Bowley, A.L. 1919. The measurement of changes in the cost of living. *Journal of the Royal Statistical Society* 82, 343–72.

Bowley, A.L. 1928. Notes on index numbers. *Economic Journal* 38, 216–37.

Cage, R., Greenlees, J.S. and Jackman, P. 2003. Introducing the chained CPI. Paper presented at the Seventh Meeting of the International Working Group on Price Indices, (Ottawa Group), Paris.

Carli, G.-R. 1764. Del valore e della proporzione de' metalli monetati. In *Scrittori classici italiani di economia politica*, vol. 13. Milano: G.G. Destefanis, 1804.

Caves, D.W., Christensen, L.R. and Diewert, W.E. 1982a. Multilateral comparisons of output, input and productivity using superlative index numbers. *Economic Journal* 92, 73–86.

Caves, D.W., Christensen, L.R. and Diewert, W.E. 1982b. The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica* 50, 1393–414.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1971. Conjugate duality and the transcendental logarithmic production function. *Econometrica* 39, 255–6.

Clements, K.W., Izan, H.Y. and Selvanathan, E.A. 2006. Stochastic index numbers: a review. *International Statistical Review* 74, 235–70.

Dalén, J. 1992. Computing elementary aggregates in the Swedish consumer price index. *Journal of Official Statistics* 8, 129–47.

de Haan, J. 2004a. Direct and indirect time dummy approaches to hedonic price measurement. *Journal of Economic and Social Measurement* 29, 427–43.

de Haan, J. 2004b. Hedonic regression: the time dummy index as a special case of the imputation Törnqvist index. Paper presented at the 8th Ottawa Group Meeting, Helsinki.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of Quantitative Economics*, vol. 2, ed. M.D. Intriligator and D.A. Kendrick. Amsterdam: North-Holland.

Diewert, W.E. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4, 114–45.

Diewert, W.E. 1978. Superlative index numbers and consistency in aggregation. *Econometrica* 46, 883–900.

Diewert, W.E. 1981. The economic theory of index numbers: a survey. In *Essays in the Theory and Measurement of Consumer Behaviour in Honour of Sir Richard Stone*, ed. A. Deaton. London: Cambridge University Press.

Diewert, W.E. 1983a. The theory of the cost of living index and the measurement of welfare change. In *Price Level Measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Diewert, W.E. 1983b. The theory of the output price index and the measurement of real output change. In *Price Level Measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Diewert, W.E. 1983c. The treatment of seasonality in a cost of living index. In *Price Level Measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Diewert, W.E. 1986. Microeconomic approaches to the theory of international comparisons. Technical Working Paper No. 53. Cambridge, MA: NBER.

Diewert, W.E. 1987. Index numbers. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Diewert, W.E. 1988. Test approaches to international comparisons. In *Measurement in Economics*, ed. W. Eichhorn. Heidelberg: Physica-Verlag.

Diewert, W.E. 1992a. Fisher ideal output, input and productivity indexes revisited. *Journal of Productivity Analysis* 3, 211–48.

Diewert, W.E. 1992b. Exact and superlative welfare change indicators. *Economic Inquiry* 30, 565–82.

Diewert, W.E. 1993a. The early history of price index research. In *Essays in Index Number Theory*, ed. W.E. Diewert and A.O. Nakamura. Amsterdam: North-Holland.

Diewert, W.E. 1993b. Symmetric means and choice under uncertainty. In *Essays in Index Number Theory*, ed. W.E. Diewert and A.O. Nakamura. Amsterdam: North-

Holland.

Diewert, W.E. 1995a. Axiomatic and economic approaches to elementary price indexes. Discussion Paper No. 95-01. Department of Economics, University of British Columbia.

Diewert, W.E. 1995b. On the stochastic approach to index numbers. Discussion Paper No. 95-31. Department of Economics, University of British Columbia.

Diewert, W.E. 1997. Commentary on Mathew D. Shapiro and David W. Wilcox: alternative strategies for aggregating price in the CPI. *Federal Reserve Bank of St. Louis Review* 79, 127–37.

Diewert, W.E. 1998a. Index number issues in the consumer price index. *Journal of Economic Perspectives* 12(1), 47–58.

Diewert, W.E. 1998b. High inflation, seasonal commodities and annual index numbers. *Macroeconomic Dynamics* 2, 456–71.

Diewert, W.E. 1999a. Index number approaches to seasonal adjustment. *Macroeconomic Dynamics* 3, 48–68.

Diewert, W.E. 1999b. Axiomatic and economic approaches to multilateral comparisons. In *International and Interarea Comparisons of Income, Output and Prices*, ed. A. Heston and R.E. Lipsey. Chicago: University of Chicago Press.

Diewert, W.E. 2001. The consumer price index and index number purpose. *Journal of Economic and Social Measurement* 27, 167–248.

Diewert, W.E. 2002a. The quadratic approximation lemma and decompositions of superlative indexes. *Journal of Economic and Social Measurement* 28, 63–88.

Diewert, W.E. 2002b. Harmonized indexes of consumer prices: their conceptual foundations. *Swiss Journal of Economics and Statistics* 138, 547–637.

Diewert, W.E. 2003. Hedonic regressions: a review of some unresolved issues. Paper presented at the 7th Meeting of the Ottawa Group, Paris.

Diewert, W.E. 2004. A new axiomatic approach to index number theory. Discussion Paper No. 04-05. Department of Economics, University of British Columbia.

Diewert, W.E. 2005a. Index number theory using differences instead of ratios. *American Journal of Economics and Sociology* 64, 311–60.

Diewert, W.E. 2005b. Weighted country product dummy variable regressions and index number formulae. *Review of Income and Wealth* 51, 561–71.

Diewert, W.E. 2005c. Adjacent period dummy variable hedonic regressions and bilateral index number theory. Discussion Paper No. 05-11. Department of Economics, University of British Columbia.

Diewert, W.E. 2007a. Similarity indexes and criteria for spatial linking. In *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, ed. D. S. Prasada Rao. Cheltenham: Edward Elgar.

Diewert, W.E. 2007b. On the stochastic approach to linking the regions in the ICP. In *Price and Productivity Measurement*, ed. W.E. Diewert et al. Vancouver: Trafford Press.

Diewert, W.E. and Fox, K.J. 2004. On the estimation of returns to scale, technical progress and monopolistic markups. Discussion Paper No. 04-09. Department of Economics, University of British Columbia.

Diewert, W.E. and Lawrence, D. 2006. *Measuring the Contributions of Productivity and Terms of Trade to Australia's Economic Welfare*. Report by Meyrick and Associates to the Australian Government. Canberra: Productivity Commission.

Diewert, W.E. and Morrison, C.J. 1986. Adjusting output and productivity indexes for changes in the terms of trade. *Economic Journal* 96, 659–79.

Diewert, W.E. and Nakamura, A.O. 2003. Index number concepts, measures and decompositions of productivity growth. *Journal of Productivity Analysis* 19, 127–59.

Edgeworth, F.Y. 1888. Some new methods of measuring variation in general prices. *Journal of the Royal Statistical Society* 51, 346–68.

Edgeworth, F.Y. 1923. The doctrine of index numbers according to Mr. Correa Walsh. *Economic Journal* 33, 343–51.

Edgeworth, F.Y. 1925. *Papers Relating to Political Economy*, vol. 1. New York: Burt Franklin.

Eichhorn, W. 1978. *Functional Equations in Economics*. London: Addison-Wesley.

Eichhorn, W. and Voeller, J. 1976. *Theory of the Price Index*. Berlin: Springer-Verlag.

Eltetö, O. and Köves, P. 1964. On a problem of index number computation relating to international comparison. *Statisztikai Szemle* 42, 507–18.

Fisher, I. 1911. *The Purchasing Power of Money*. London: Macmillan.

Fisher, I. 1921. The best form of index number. *Journal of the American Statistical Association* 17, 535–7.

Fisher, I. 1922. *The Making of Index Numbers*. Boston: Houghton Mifflin.

Fisher, W.C. 1913. The tabular standard in Massachusetts history. *Quarterly Journal of Economics* 27, 417–51.

Fox, K.J. and Kohli, U. 1998. GDP growth, terms of trade effects and total factor productivity. *Journal of International Trade and Economic Development* 7, 87–110.

Frisch, R. 1936. Annual survey of economic theory: the problem of index numbers. *Econometrica* 4, 1–39.

Funke, H. and Voeller, J. 1978. A note on the characterization of Fisher's ideal index. In *Theory and Applications of Economic Indices*, ed. W. Eichhorn et al. Würzburg: Physica-Verlag.

Funke, H., Hacker, G. and Voeller, J. 1979. Fisher's circular test reconsidered. *Schweizerische Zeitschrift für Volkswirtshaft und Statistik* 115, 677–87.

Geary, R.G. 1958. A note on comparisons of exchange rates and purchasing power between countries. *Journal of the Royal Statistical Society* Series A 121, 97–9.

Gini, C. 1924. Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues. *Metron* 4, 3–162.

Gini, C. 1931. On the circular test of index numbers. *Metron* 9, 3–24.

Hicks, J.R. 1941–42. Consumers' surplus and index numbers. *Review of Economic Studies* 9, 126–37.

Hill, R.J. 1997. A taxonomy of multilateral methods for making international comparisons of prices and quantities. *Review of Income and Wealth* 43, 49–69.

Hill, R.J. 1999a. Comparing price levels across countries using minimum spanning trees. *Review of Economics and Statistics* 81, 135–42.

Hill, R.J. 1999b. International comparisons using spanning trees. In *International and Interarea Comparisons of Income, Output and Prices*, vol. 61, ed. A. Heston and R.E. Lipsey. Studies in Income and Wealth, NBER, Chicago: University of Chicago Press.

Hill, R.J. 2001. Measuring inflation and growth using spanning trees. *International Economic Review* 42, 167–85.

Hill, R.J. 2004. Constructing price indexes across space and time: the case of the European Union. *American Economic Review* 94, 1379–410.

Hill, R.J. 2006a. Superlative index numbers: not all of them are super. *Journal of Econometrics* 130, 25–43.

Hill, R.J. 2006b. When does chaining reduce the Paasche–Laspeyres spread? An application to scanner data. *Review of Income and Wealth* 52, 309–29.

Hill, R.J. 2007. Comparing per capita income levels across countries using spanning trees: robustness, prior restrictions, hybrids and hierarchies. In *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, ed. D.S. Prasada Rao. Cheltenham: Edward Elgar.

Hill, T.P. 1988. Recent developments in index number theory and practice. *OECD Economic Studies* 10, 123–48.

Hill, T.P. 1993. Price and volume measures. In *System of National Accounts 1993*. Luxembourg, Washington, DC, Paris, New York, and Washington, DC: Eurostat, IMF, OECD, UN and World Bank.

Hill, T.P. 1996. *Inflation Accounting*. Paris: OECD.

ILO/IMF/OECD/UNECE/Eurostat/World Bank. 2004. *Consumer Price Index Manual: Theory and Practice*, ed. P. Hill. Geneva: International Labour Office.

IMF/ILO/OECD/UNECE/Eurostat/ World Bank. 2004. *Producer Price Index Manual: Theory and Practice*, ed. P. Armknecht. Washington, DC: International Monetary Fund.

Jevons, W.S. 1865. The variation of prices and the value of the currency since 1782. *Journal of the Statistical Society of London* 28, 294–320. Repr. in *Investigations in Currency and Finance*, London: Macmillan and Co, 1884.

Jevons, W.S. 1884. A serious fall in the value of gold ascertained and its social effects set forth (1863). In *Investigations in Currency and Finance*, London: Macmillan and Co.

Jorgenson, D.W. and Griliches, Z. 1967. The explanation of productivity change. *Review of Economic Studies* 34, 249–83.

Jorgenson, D.W. and Griliches, Z. 1972. Issues in growth accounting: a reply to Edward F. Denison. *Survey of Current Business* 52, 65–94.

Keynes, J.M. 1930. *Treatise on Money*, vol. 1. London: Macmillan.

Khamis, S.H. 1972. A new system of index numbers for national and international purposes. *Journal of the Royal Statistical Society* Series A 135, 96–121.

Kohli, U. 1990. Growth accounting in the open economy: parametric and nonparametric estimates. *Journal of Economic and Social Measurement* 16, 125–36.

Kohli, U. 2003. Growth accounting in the open economy: international comparisons. *International Review of Economics and Finance* 12, 417–35.

Kohli, U. 2004. Real GDP, real domestic income and terms of trade changes. *Journal of International Economics* 62, 83–106.

Kohli, U. 2007. Terms of trade, real exchange rates, and trading gains. In *Price and Productivity Measurement*, ed. W.E. Diewert et al. Vancouver: Trafford Press.

Konüs, A.A. 1924. The problem of the true index of the cost of living. Trans. in *Econometrica* 7(1939), 10–29.

Konüs, A.A. and Byushgens, S.S. 1926. K probleme pokupatelnoi cili deneg. *Voprosi konyunkturi* 2, 151–72.

Kravis, I.B. 1984. Comparative studies of national incomes and prices. *Journal of Economic Literature* 22, 1–39.

Kravis, I.B., Kenessey, Z., Heston, A. and Summers, R. 1975. *A System of International Comparisons of Gross Product and Purchasing Power*. Baltimore: Johns Hopkins University Press.

Laspeyres, E. 1871. Die berechnung einer mittleren waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik* 16, 296–314.

Lebow, D.E. and Rudd, J.B. 2003. Measurement error in the consumer price index: where do we stand? *Journal of Economic Literature* 41, 159–201.

Lehr, J. 1885. *Beitrage zur Statistik der Preise*. Frankfurt: J.D. Sauerlander.

Lowe, J. 1823. *The Present State of England in Regard to Agriculture, Trade and Finance*, 2nd edn. London: Longman, Hurst, Rees, Orme and Brown.

Malmquist, S. 1953. Index numbers and indifference surfaces. *Trabajos de Estadistica* 4, 209–42.

Marshall, A. 1887. Remedies for fluctuations of general prices. *Contemporary Review* 51, 355–75.

Montgomery, J.K. 1929. *Is there a Theoretically Correct Price Index of a Group of Commodities*? Rome: Roma L'Universale Tipogr. Poliglotta (privately printed).

Montgomery, J.K. 1937. *The Mathematical Problem of the Price Index*. Orchard House, Westminster: P.S. King & Son.

Morrison, C.J. and Diewert, W.E. 1990. Productivity growth and changes in the terms of trade in Japan and the United States. In *Productivity Growth in Japan and the United States*, ed. C.R. Hulten. Chicago: University of Chicago Press.

Moulton, B.R. and Seskin, E.P. 1999. A preview of the 1999 comprehensive revision of the national income and product accounts. *Survey of Current Business* 79, 6–17.

Paasche, H. 1874. Über die preisentwicklung der letzten Jahre nach den hamburger borsennotirungen. *Jahrbücher für Nationalökonomie und Statistik* 12, 168–78.

Pierson, N.G. 1895. Index numbers and appreciation of gold. *Economic Journal* 5, 329–35.

Pierson, N.G. 1896. Further considerations on index-numbers. *Economic Journal* 6, 127–31.

Pollak, R.A. 1980. Group cost-of-living indexes. *American Economic Review* 70, 273–8.

Pollak, R.A. 1981. The social cost-of-living index. *Journal of Public Economics* 15, 311–36.

Pollak, R.A. 1983. The theory of the cost-of-living index. In *Price Level Measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Rao, D.S.P. 1990. A system of log-change index numbers for multilateral comparisons. In *Comparisons of Prices and Real Products in Latin America*, ed. J. Salazar-Carillo and D.S. Prasada Rao. New York: Elsevier.

Reinsdorf, M. 1993. The effect of outlet price differentials on the U.S. consumer price index. In *Price Measurement and their Uses*, ed. M.F. Foss, M.E. Manser and A.H. Young. Chicago: University of Chicago Press.

Reinsdorf, M. 2007. Axiomatic price index theory. In *Measurement in Economics: A Handbook*, ed. M.J. Boumans. Amsterdam: Elsevier.

Reinsdorf, M.B., Diewert, W.E. and Ehemann, C. 2002. Additive decompositions for the fisher, Törnqvist and geometric mean indexes. *Journal of Economic and Social Measurement* 28, 51–61.

Reinsdorf, M. and Dorfman, A. 1999. The monotonicity axiom and the Sato-Vartia index. *Journal of Econometrics* 90, 45–61.

Ruggles, R. 1967. Price indexes and international price comparisons. In *Ten Economic Studies in the Tradition of Irving Fisher*, ed. W. Fellner. New York: John Wiley.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. and Swamy, S. 1974. Invariant economic index numbers and canonical duality: survey and synthesis. *American Economic Review* 64, 566–93.

Schreyer, P. 2001. *OECD Productivity Manual: A Guide to the Measurement of Industry-Level and Aggregate Productivity Growth*. Paris: OECD.

Schultze, C.L. and Mackie, C. 2002. *At What Price? Conceptualizing and Measuring Cost-of Living and Price Indices*. Washington, DC: National Academy Press.

Scrope, G.P. 1833. *Principles of Political Economy*. London: Longman, Rees, Orme, Brown, Green and Longman.

Selvanathan, E.A. and Prasada Rao, D.S. 1994. *Index Numbers: A Stochastic Approach*. Ann Arbor: University of Michigan Press.

Shephard, R.W. 1953. *Cost and Production Functions*. Princeton: Princeton University Press.

Sidgwick, H. 1883. *The Principles of Political Economy*. London: Macmillan.

Silver, M. 2003. The use of weights in hedonic regressions: the measurement of quality adjusted price changes. Room document for the 7th Meeting of the Ottawa Group, Paris.

Silver, M. and Heravi, S. 2005. A failure in the measurement of inflation: results from a hedonic and matched experiment using scanner data. *Journal of Business and Economic Statistics* 23, 269–81.

Summers, R. 1973. International comparisons with incomplete data. *Review of Income and Wealth* 29, 1–16.

Szulc, B. 1964. Indices for multiregional comparisons. *Przeglad Statystyczny* 3, 239–54.

Szulc, B.J. 1983. Linking price index numbers. In *Price Level Measurement*, ed. W.E. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Theil, H. 1967. *Economics and Information Theory*. Amsterdam: North-Holland.

Törnqvist, L. 1936. The bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin* 10, 1–8.

Törnqvist, L. and Törnqvist, E. 1937. Vilket är förhällandet mellan finska markens och svenska kronans köpkraft? *Ekonomiska Samfundets Tidskrift* 39, 1–39. Repr. in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: Research Institute of the Finnish Economy, 1981.

Triplett, J.E. 2004. *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*. Working Paper 2004/9. Paris: Directorate for Science, Technology and Industry, OECD.

Triplett, J.E. and McDonald, R.J. 1977. Assessing the quality error in output measures: the case of refrigerators. *Review of Income and Wealth* 23, 137–56.

Turvey, R. 1979. The treatment of seasonal items in consumer price indices. *Bulletin of Labour Statistics*, Fourth Quarter, 13–33. Geneva: International Labour Office.

US Bureau of Labor Statistics. 1983. Trends in multifactor productivity, 1948–81. *Bulletin* 2178. Washington, DC: U.S. Government Printing Office.

van IJzeren, J. 1983. *Index Numbers for Binary and Multilateral Comparisons*. Statistical Studies No. 34. The Hague: Central Bureau of Statistics.

van IJzeren, J. 1987. *Bias in International Index Numbers: A Mathematical Elucidation*. Dissertation for the Hungarian Academy of Sciences. The Hague: Koninklijke Bibliotheek.

van IJzeren, J. 1957. *Three Methods of Comparing the Purchasing Power of Currencies*, Statistical Studies No. 7. The Hague: Central Bureau of Statistics.

Vartia, Y.O. 1976. *Relative Changes and Index Numbers*. Helsinki: Research Institute of the Finnish Economy.

Vogt, A. 1980. Der Zeit und der Faktorumkehrtest als ‘Finders of Tests’. *Statistische Hefte* 21, 66–71.

Vogt, A. and Barta, J. 1997. *The Making of Tests for Index Numbers*. Heidelberg: Physica-Verlag.

von der Lippe, P. 2001. *Chain Indices: A Study in Price Index Theory*, Spectrum of Federal Statistics vol. 16. Wiesbaden: Statistisches Bundesamt.

Walsh, C.M. 1901. *The Measurement of General Exchange Value*. New York: Macmillan and Co.

Walsh, C.M. 1921a. *The Problem of Estimation*. London: P.S. King & Son.

Walsh, C.M. 1921b. Discussion. *Journal of the American Statistical Association* 17, 537–44.

Westergaard, H. 1890. *Die Grundzüge der Theorie der Statistik*. Jena: Fischer.

Wynne, M.A. 1997. Commentary on measuring short run inflation for central bankers. *Federal Reserve Bank of St. Louis Review* 79, 161–7.

Young, A. 1812. *An Inquiry into the Progressive Value of Money in England as Marked by the Price of Agricultural Products*. London.

How to cite this article

Diewert, W. E. "index numbers." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000053> doi:10.1057/9780230226203.0775

The New Palgrave Dictionary of Economics Online

India, economics in

Deepak Lal
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article outlines the debates amongst Indian economists on planning, transforming agriculture, poverty and income distribution, and political economy and institutions. It shows that much of this work pioneered many analyses which have come to define the sub-discipline of ‘development economics’.

Keywords

agricultural subsidies; agriculture and economic development; attached labour; Bauer, P.; Bhagwati, J.; caste system; central planning; comparative advantage; conjunctural poverty; Delhi School; destitution; development economics; dirigisme; dual economies; Dutch disease; exchange controls; famine; farm size; Fel'dman, G. A.; food security; free trade; Friedman, M.; Green Revolution; homo economicus; import substitution; India, economics in; industrialization; Johnson, H. G.; land reform; Leisure; Lewis, W. A.; Mahalanobis, P. C.; Malthus's theory of population; Meade, J. E.; pastoralism; peasants; population growth; poverty; poverty alleviation; price controls; privatization; protection; public works; Ramaswami, V. K.; redistribution of income; rent seeking; rural employment; Sen, A.; shadow prices; sharecropping; Shenoy, B. R.; Stalin, J.; structural poverty; surplus labour; terms of trade; usury

Article

Economics in India has been mainly concerned with finding means to alleviate its ancient and pervasive poverty. In this article I will concentrate on the debates amongst Indian economists, highlighting the contributions they have made in the process to the new discipline of ‘development economics’.
The Indian economic debate began in the early 20th century when after nearly a century of British colonial rule there were few signs of poverty alleviation, with only a modest rise in per capita income over the period (Sivasubramonian, 2000). A nationalist and Marxist literature evolved, which laid the blame for this economic stagnation on alien rule and the implementation – since the 1850s – of the twin classical liberal principles (dominant in the metropolitan centre) of laissez-faire and ‘free trade’. Alien

rule was epitomized by the fiscal drain of resources from India to Britain (Naoroji, 1901; Dutt, 1904). Free trade was held responsible for India's failure to industrialize and the destruction of its extensive pre-colonial handloom textile industry.

By the 1930s, the Great Depression and Stalinist Russia's success in rapidly industrializing a large, poor and mainly agrarian economy coloured the thinking of Indian economists and political leaders like Nehru. A series of economic plans were drawn up by various groups and individuals, including the National Planning Committee of the Indian National Congress (Visveswarya, 1934; Nehru, 1946; Banerjee et al., 1944; Thakurdas et al. 1944; Agarwal, 1960), that anticipated most post-war debates and ideas on development objectives, strategy and policy in academia and international organizations. The plans saw poverty alleviation as the basic development objective, outlined a ‘basic needs’ strategy and covered ‘redistribution with growth’, the development of agriculture versus industry, heavy industry-based industrialization and import substitution, the respective roles of large- and small-scale industries and of the state versus the market (see Srinivasan, 2001).

The rise and fall of the planning syndrome

With the setting up of the Planning Commission in the 1950s India embarked on a public sector dominated by heavy industry and an import-substituting industrialization strategy as the answer to alleviate its ancient poverty. Professor P.C. Mahalanobis (1953; 1955), a distinguished statistician and the father of Indian planning, provided its rationale in a formal model, taken largely from the model that the Soviet economist Fel'dman had developed for Stalin's industrialization strategy. This showed that, with a binding foreign exchange constraint (which, on the basis of the export pessimism generated by the experience of the Great Depression, was assumed to confront India) independent of a savings constraint to limit the growth rate of the economy, a higher sustainable development path could be attained by using limited foreign exchange to import (and so support the industrial structure vertically) machines to make machines, until India was producing everything she needed, except for the raw materials that could not be obtained domestically (see Bhagwati and Chakravarty, 1969; Lal, 1972a).

The Perspective Planning Division of the Planning Commission, headed by its intellectually curious and energetic head, Pitamber Pant, and the branch of Mahalanobis’ Indian Statistical Institute (ISI) attached to it, then became the centre of intense intellectual debate. In the 1960s it employed a growing number of Indian economists trained in Western universities (Bhagwati, Bardhan, Minhas, Parikh, Srinivasan, Tendulkar among others), and in association with a programme set up by Rosenstein Rodan at Massachusetts Institute of Technology (MIT) became host to a galaxy of foreign economists (Swan, Reddaway, Lewis, Little and Harberger). The Delhi School of Economics, under the leadership of K.N. Raj, engaged Chakravarty and Sen, and at the Finance Ministry I.G. Patel invigorated the newly established Indian Economic Service by engaging V.K. Ramaswami and Manmohan Singh as economic advisors. Meanwhile, the USAID mission was headed by J.P. Lewis, and the number of foreign economists visiting and participating in the economic debates of the time expanded to include Milton Friedman and Peter Bauer.

The Mahalanobis model was to form the analytical basis for India's second Five Year Plan. The Planning Commission had convened a panel of economists to discuss its framework, and most of them endorsed the broad objectives and strategy of the plan. The only dissenting voice was that of B.R. Shenoy, who

questioned, amongst other issues, the massive deficit financing on which the plan depended. In this he was supported by two of the visiting foreign economists, Peter Bauer and Milton Friedman. Whilst Komiya (1959) and Bronfrenbrenner (1960) provided explicit critiques of the Mahalanobis model. But most of these criticisms were disregarded by the prevailing intellectual consensus in favour of dirigiste, state-led planning, though the technocratic basis of the planning models on which it was based was increasingly questioned by Indian economists (see Rudra, 1975).

With the emergence of what J.P. Lewis (1963) accurately described as a ‘quiet crisis’ in India, engendered by the foreign exchange crisis caused by the fiscal expansion the dissenters had predicted (which had led to draconian foreign trade-cum-exchange and price controls), new voices arose in the 1960s providing the intellectual basis for the subsequent neoclassical resurgence in development economics. Developing ideas presaged in the writings of James Meade and Harry Johnson, two Indian economists, Jagdish Bhagwati (who was at the ISI) and V.K. Ramaswami, economic advisor at the Ministry of Finance, produced a path-breaking paper that began the process of separating the case for free trade from that for laissez-faire (Bhagwati and Ramaswami, 1963). In a series of papers with T.N. Srinivasan (also at the ISI), they established the modern theory of trade and welfare which shows that most of the arguments for protection are second best as they depend upon ‘domestic distortions’ in the working of the price mechanism, which are best dealt with by direct domestic taxes and subsidies rather than the indirect method of protection.

Two major books, by Bhagwati and Desai (1970) and Bhagwati and Srinivasan (1975), written as part of two large-scale multi-country comparative studies of trade and industrialization directed by I.M.D. Little, T. Scitovsky and M. Fg. Scott for the Organization for Economic Cooperation and Development (OECD), and by J. Bhagwati and A. Krueger for the National Bureau of Economic Research (NBER), provided a detailed empirical analysis of the relevance of this newly developed theory, besides documenting the immense inefficiency and corruption that the dirigiste planning system had engendered. This marked the beginning of the end of the planning syndrome that had held Indian economists in thrall for nearly a century.

Furthering this disenchantment was the disappointing performance of Indian industry where the net effect of the control system was shown to be a capital-intensive bias and low or negative growth of total factor productivity in post-Independence industrial performance (I.J. Ahluwalia, 1985).

Moreover, Manmohan Singh (1964), in a detailed study of Indian exports, had shown that the export pessimism underlying the assumption of a foreign-exchange constraint in the Mahalanobis model was unjustified, as it was not lack of external demand but the consequences of India's domestic economic policies that had led to the disappointing Indian export performance.

Nor was the panacea offered by the Gandhians – which was promulgated with reservations for various small-scale industries (particularly cotton textiles) on the grounds that they promoted employment growth – found to be valid. P.N. Dhar and H.F. Lydall (1961) in an empirical study of these industries showed that these small-scale industries were technically inefficient than their larger modern brethren because they used both more labour and capital per unit of output produced.

The planners’ belief that the public sector, given monopoly production rights in the ‘commanding heights’ of the economy, would be dynamic and through rising profits augment domestic savings was discredited. Numerous official empirical studies documented the growing inefficiency of the public sector and its growing drain on the nation's savings. As part of the debate on their reform which came to

the fore in the 1970s, two major manuals of project evaluation were developed to improve the efficiency of the public sector. One was produced for the UN's Industrial Development Organization by P. Dasgupta, A.K. Sen and S. Marglin the other for the OECD by I.M.D. Little and J.A. Mirrlees. With the implicit adoption of the latter by a newly set up Project Appraisal Division in the Planning Commission, Lal (1980) produced the first comprehensive set of ‘shadow prices’ based on the ‘world price rule’ for use in the evaluation of public projects in India. But the social cost–benefit analysis they were meant to support soon descended into social cosmetic analysis, as politicians continued to choose and run public projects for rent-seeking reasons rather than social profitability. It was not until the fiscal-cum-foreign exchange crisis of 1991 that planning, and the system of controls on industry and foreign trade it had engendered, finally came to a *de facto* if not *de jure* end. The market increasingly came to replace the plan, and a programme of privatization was slowly and fitfully begun.

Transforming agriculture

An implicit assumption of the Mahalanobis framework was that agriculture could be left alone, merely being a source of ‘surplus labour’ and of the limited savings and foreign exchange for the heavy industrialization strategy. By the mid-1960s this neglect had led to a severe food crisis. The transformation of agriculture, which until then had been seen largely as a means of promoting equity through land reforms, then became a matter of debate.

Nationalist and Marxist literature in India, basing itself on the perceived outcomes of the laissez-faire period of colonial rule, had maintained that the commercialization of agriculture through the creation, definition and enforcement of saleable and mortgageable land rights, and the integration of the internal economy through the railways had led to an increased concentration of land, the proletarianization of the peasantry and the growth of landless labour and a shift to cash crops from foodgrains, which in turn had led to famine. Subsequent research (summarized in Kumar and Desai, 1983, and Lal, 1988), has questioned the empirical bases of these beliefs, whilst Sen (1981a) has argued that the periodic famines that have blighted the subcontinent over the millennia were not due to a shortage of food but to ‘exchange entitlement failures’. Whenever the monsoon failed there was a drastic fall in the demand for landless labour and thence wages, leading to a reduction in ‘exchange entitlement’ in terms of food, which in extremity would lead to a famine. The British had already realized this at the end of the 19th century, when they set up a famine code whereby, when the rains failed, local District Commissioners were empowered to fund food-for-work public works to provide the necessary exchange entitlements. As a result, apart from the 1944 famine in Bengal, which was caused by disruptive wartime conditions, India did not see serious famines in the 20th century.

One of the implicit assumptions underlying the neglect of agriculture in the early plans was that peasants were not subject to economic incentives. Detailed empirical studies by Dharm Narain (1965) and Raj Krishna (1963) of peasant response to the changing relative prices of crops shows that they behaved like *homo economicus* by shifting cropping patterns to crops with higher expected relative prices.

A second tenet (following the famous Arthur Lewis model of a dual economy) was the existence of vast pools of ‘surplus labour’ in agriculture which could be removed for industrialization without affecting agricultural output. Mehra (1966) provided empirical content by using farm management studies to estimate the surplus labour time available in various states in India. But these and other studies

estimating surplus labour did not take account of the wage at which people are willing to work, or the leisure–income choice facing rural workers. They assumed that they would continue to work for an unchanged wage up to a normal number of working hours per day. But, as Sen (1966) showed, even in an overpopulated country, ‘surplus labour’ – in the sense of a perfectly elastic supply of labour at a constant wage – would imply that leisure was an inferior good. Empirical studies estimating wage elasticities for rural labour in India soon showed that this assumption was invalid (Bardhan, 1979; 1984a; Binswanger and Rosenzweig, 1984; Lal, 1989).

The means to transform Indian agriculture have not changed since the 1893 report by J. Volcker (1893), consultant chemist to the Royal Agricultural Society. His remedies were: irrigation, fertilizers, better seeds and improvements in land tenure. This has been the conventional wisdom on raising Indian agricultural productivity ever since.

An empirical finding from the Indian farm management studies that there was an inverse relationship between the size of farm and productivity per hectare (Sen, 1975, Appendix C) was used to argue for land reforms that would break up large farms and create small, family-labour based and family-owned peasant farms, which would promote both equity and efficiency (Rudra and Sen, 1980). However, Bhalla and Roy (1988) showed that, once appropriate adjustments were made for differences in land quality, the inverse relationship between farm size and productivity disappears. This undermined the case for land reform in India.

Lal (1988; 2005; 2006) argued that the Malthusian view that population pressure would lead to a stagnation of rural and industrial wages was invalid, as the alternative Boserupian perspective (Boserup, 1965) provided a better description of the changing fortunes of Indian agriculture. Boserup argued that population pressure both induces and facilitates the adoption of more intensive forms of agriculture. She identifies the differing input-per-hectare requirements of different agrarian systems by the frequency with which a particular piece of land is cropped. Thus settled agriculture is more labour- and capital-intensive than nomadic pastoralism, which is in turn more intensive in these inputs than hunting and gathering or the slash-and-burn agriculture practised until recently in parts of Africa and the tribal regions of India. Contrary to Malthusian presumptions, population growth leads to the adoption of more advanced techniques that raise yield per acre. Because these new techniques require increased labour effort, they will not be adopted until rising population reduces the per capita food output that can be produced with existing techniques and forces a change. Lal marshals empirical evidence to show that Indian agriculture's long trajectory fits this Boserupian framework, with the population expansion beginning from the early 1900s leading in the post-Independence period to an intensification of agriculture, and with the availability of the new high-yielding varieties (HYV) of seeds, to the Green Revolution in the late 1960s and 1970s.

Many of those adhering to the Marxist canon believed and hoped that the bulk of the income gains arising from the massive increases in output brought about by the Green Revolution would accrue to landowners, and that rural real wages would stagnate, leading to the revolution turning red. But the evidence showed that with the massive shift in the labour-demand curve that resulted from the new technology there was a marked rise in rural real wages (Ahluwalia, 1978; Lal, 1976; 1989).

As the new HYV technology required an assured water supply along with high dosages of fertilizers, Volcker's other major means of transforming Indian agriculture, namely irrigation, came to the fore. Surface irrigation was expanded during the Raj (the period of British rule in India), particularly in the

drier regions where the marginal social returns from irrigation were likely to be the highest. But these schemes were devised by engineers and their direct and indirect economic effects were not estimated, leading in many cases to long-term losses through salination, water-logging and the creation of malarial swamps (see Whitcomb, 1971). In the 1970s two studies of irrigation – of a major surface water scheme, the Bhakra dam, by Minhas, Parikh and Srinivasan (1972) and of groundwater (well) irrigation in the Deccan plateau by Lal (1972b) – provided economic analyses of irrigation and their optimal design.
One of the deleterious effects of the system of protection set up during the Permit Raj was the heavy implicit tax on agriculture. From 1965 efforts were made to correct this by price supports to farmers, which led to an improvement in the terms of trade. But this changed again in the 1980s with growing but inefficient input subsidies becoming the main form of supporting agriculture. With the post-1991 liberalization of trade largely affecting industrial products, part of the bias against agriculture was removed. The debate then moved to removing the remaining agricultural protection (particularly for cereals), with proponents (Gulati, 1998) arguing for domestic prices of agricultural products to be aligned with world prices to allow agriculture to develop in line with its revealed comparative advantage, and opponents (Patnaik, 1996) arguing against, on grounds of food security.

Poverty and income distribution

A continuing debate concerns the effects on income distribution and poverty of rapid capitalist growth. Indian economists have been in the forefront in both setting out the conceptual basis as well as the measurement of poverty (see Sen, 1976; Sen, 1981a; Sen, 1981b; Dandekar and Rath, 1971; Bardhan and Srinivasan, 1974; Srinivasan, 1983). The internationally adopted headcount ratio (HCR) of the poor below a nutritionally based poverty line of 15 rupees per capita (at 1960–1 prices) was based on this efflorescence of research in the 1970s (but see Sukhatme, 1978; Srinivasan and Bardhan, 1988). The continuing debate has centred on whether rapid (capitalist) growth would alleviate poverty without adverse effects on income distribution, or whether more direct methods of redistribution would be needed to alleviate poverty and prevent any worsening of income distribution. A summary of the evidence from these numerous studies based on two large national surveys undertaken by the official National Sample Survey and those undertaken by the unofficial National Council of Applied Economic Research (NCAER) is provided in Lal, Mohan and Natarajan (2001). There seems to be no clear trend in the Gini coefficient during the 50 years since Independence in 1947, whilst the fluctuating HCR for poverty shows no marked change until the acceleration of the growth rate after the economic liberalization of the 1990s, since when there has been a fall of varying magnitudes, depending upon which study one trusts.
The nationalist-cum-Marxist School unsurprisingly has argued that ‘trickle down’ would not alleviate poverty. Given the abysmally poor growth record during the planning period, which was characterized as the Hindu rate of growth (of about 1.5 per cent a year in per capita income from the 1950s to early 1980s) it would have been surprising if there had been any marked alleviation of India's mass structural poverty. Nevertheless, influential voices on the Left articulated a critique of the capitalist growth process. This critique, purportedly supported by Indian data, was soon shown to be false. Thus it was argued that the alleviation of poverty and equitable growth within the ‘existing institutional framework’ would not occur because of an increased concentration of land (Raj, 1976; refuted by Sanyal, 1977a;

1977b); the increasing proletarianization of the countryside (Raj, 1976; refuted by Visaria, 1977); increasing rural indebtedness and usury (disputed by Ghatak, 1976); a continual improvement in the agricultural terms of trade which damaged industrial development (Bagchi, 1970; Chakravarty, 1974; Sau, 1981; Vaidyanathan, 1977; and Mitra, 1977), which were critiqued by Desai, (1981); and the inimical effects of foreign investment (Sau, 1981) which is countered in Lal et al. (1975). These are now seen as shibboleths, particularly after the death of the countries of ‘really existing socialism’ and the economic liberalizations of the 1990s. The intemperate debate this provoked between the left-wing radicals and neoclassical liberalizers showed up the ideological nature of this debate, with Rudra (1991) stating: ‘I put my ideological cards on the table. I hate capitalism’, and Srinivasan (1992) rightly responding: ‘In Rudra's value system competition, without which the market economy cannot efficiently function, is an instrument with a negative value connotation. In this he would be in the good company of monopolists and oligopolists and state capitalists of the world who would also dearly love to eliminate competition!’

While growth is being increasingly accepted as necessary for the sustainable alleviation of mass structural poverty (see Tendulkar, 1998), Lal and Myint (1996) argue that two other forms of poverty, destitution and conjunctural poverty, require income transfers, though not necessarily public ones. Though Dasgupta (1993) claims to be about destitution, it is more about mass structural poverty and income distribution (Srinivasan, 1994). The only study of destitution (Lipton, 1983) based on village studies found no obvious correlates to identify an extremely heterogenous group. Thus Dasgupta's reasonable assertion that widows become destitute was belied by the evidence in Drèze and Srinivasan (1995).

Public policy has thus sought to deal with the third triad of poverty, conjunctural poverty, which is largely associated with climatic variations through a continuation of the Raj's famine code to prevent famine and by rural employment guarantee schemes to offset seasonal unemployment by offering jobs on public works at a wage only the needy will accept, which because of self-targeting have been shown to be efficacious (Ravallion, 1991).

The major advocate of the direct route for poverty alleviation (where the three categories distinguished above are amalgamated) remains Sen (1981), whose earlier empirical evidence on the superiority of this route in low-growth economies (Sri Lanka) and regions (Kerala in India) was questioned by Bhalla and Glewwe (1986). The debates in Drèze and Sen (1989) concentrate on the public provision of food for the malnourished and the merit goods of health and education. But empirical studies of the nearly 50-year-old public programmes to deal with these aspects do not provide much hope for success (Parikh, 1993; World Bank, 2000; PROBE, 1999). Similarly, the dismal state of publicly owned and operated infrastructure (Ahluwalia 1998; Ahluwalia and Little, 1998) has led to a search for decentralized private solutions to provide these ‘public goods’ with public funding (Mitra, 2006; Bardhan and Mookherjee, 2006).

Political economy and institutions

With the growing corruption engendered by the Permit Raj, there have been attempts to measure what

Krueger (1974) has designated as the ‘rent-seeking society’. Her attempts at measuring the rents created by the Permit Raj in India has been supplemented by other studies (see Acharya, 1985; Mohammad and Whalley, 1984), whilst her rent-seeking model has been expanded by Bhagwati and Srinivasan to encompass a whole host of what they term ‘directly unproductive activities’ (Bhagwati and Srinivasan, 1980).

A large political economy literature has arisen to explain the economic outcomes in India's democratic polity. Much of this has a Marxist lineage (Raj, 1973; Jha, 1980; Bardhan, 1984b). Lal (1984; 1988; 2005) on the other hand has developed a model of ‘the predatory state’ which maximizes net revenue and has argued that the successive empires in north India were predatory states that fell when they attempted to extract more than the natural ‘rent’ the economic system could provide. Lal (1987) and Lal and Myint (1996) also provide a theory which seeks to explain the role of crises in generating economic reforms in previously repressed economies. This is borne out by the liberalization undertaken in the face of a serious fiscal, foreign exchange and inflationary crisis in 1991 caused by the cumulative effects of the dirigisme of the Permit Raj.

There have also been attempts to explain various institutions that have shaped economic outcomes: the caste system (Lal, 1988; 2005) as a means of tying scarce labour down to abundant land, and a theory of interrelated factor markets which seeks to explain seemingly inefficient institutions like sharecropping, attached labour, and usurious interest rates as second-best adaptations to problems of risk and the uncertainty to which tropical agriculture is subject (Bardhan, 1980; Bardhan and Rudra, 1978; Srinivasan, Bell and Udry, 1997; Basu, 1983).

The macroeconomy

Post-Independence India followed an orthodox monetary policy based on the system of fiscal and monetary accounting left by the Raj. In the 1980s, however, in order to push up the growth rate it began to undertake risky macroeconomic policies, and, with the crisis of 1991, macroeconomic issues came to the fore. The best account of India's macroeconomy since Independence was provided by Joshi and Little (1994), whilst Bhagwati and Srinivasan (1993) and Virmani (2001) provide analyses of the genesis of the crises and the lineaments of the partial and still incomplete economic liberalization that occurred in the wake of the crisis.

With the opening of the economy and (by the standards of the planning era) large inflows of foreign capital, India faced the prospect of Dutch disease – with a rise in the real exchange rate reducing the profitability of tradable relative to non-traded goods. The authorities responded by sterilizing these inflows and building up large foreign-exchange reserves, thus stalling an appreciation of the nominal exchange rate, to maintain the competitiveness of Indian exports (which, after their post-Independence stagnation, in the 1990s began to take off with the gradual integration of India into the world economy). Because of the continuing large fiscal deficits, particularly of the states in the Indian federation (Lal, Bhide and Vasudevan, 2001), the government was also reluctant to open the capital account for fear of these deficits spilling over and causing another foreign debt crisis. A lively debate began in the early part of the 21st century on the correct monetary and exchange-rate policy for India to follow in the light of the continuing build-up in foreign exchange reserves. Lal, Bery and Pant (2003) argued for

liberalizing the capital account and floating the rupee. Joshi and Sanyal (2004) demurred, arguing for capital account controls and a managed exchange rate, largely on grounds of exchange-rate protection. The debate is still ongoing as of 2007, and the government has reconstituted an official committee which in the late 1990s had cautioned on opening the capital account.

The economic debates in India have thus moved on to what are no longer distinctively Indian issues, and local contributions are now less likely to be ground-breaking or to deal uniquely with issues in the current debates on development in the subcontinent.

See Also

- agriculture and economic development
- development economics
- planning
- poverty
- poverty alleviation programmes

Bibliography

Acharya, S. 1985. *Report on the Underground Economy*. New Delhi: National Institute of Public Finance and Policy.

Agarwal, S.N. 1960. *Principles of Gandhian Planning*. Bombay: Kitab Mahal.

Ahluwalia, I.J. 1985. *Industrial Growth in India: Stagnation since the mid-1960s*. New Delhi: Oxford University Press.

Ahluwalia, I.J. and Little, I.M.D., eds. 1998. *India's Economic Reforms and Development: Essays for Manmohan Singh*. New Delhi: Oxford University Press.

Ahluwalia, M.S. 1978. Rural poverty and agricultural performance in India. *Journal of Development Studies* 14, 298–323.

Ahluwalia, M.S. 1998. Infrastructure development in India's reforms. In Ahluwalia and Little (1998).

Bagchi, A.K. 1970. Long term constraints on India's industrial growth 1951–1968. In *Economic Development in South Asia*, ed. E.A.G. Robinson and M. Kidron. London: Macmillan.

Banerjee, B.N., Parikh, G.D. and Tarkunde, V.M. 1944. *People's Plan for Economic Development of India*. Bombay: Indian Federation of Labour.

Bardhan, P. 1979. Labor supply functions in a poor agrarian economy. *American Economic Review* 69, 73–83.

Bardhan, P. 1980. Interlocking factor markets and Agrarian development: a review of the issues. *Oxford Economic Papers* 32, 82–98.

Bardhan, P. 1984a. *Land, Labour and Rural Poverty*. New Delhi: Oxford University Press.

Bardhan, P. 1984b. *The Political Economy of Development in India*. Oxford: Blackwell.

Bardhan, P. and Mookherjee, D. 2006. Decentralization and accountability in infrastructure delivery in developing countries. *Economic Journal* 116, 107–33.

Bardhan, P. and Rudra, A. 1978. Interlinkage of land, labour, and credit relations: an analysis of village survey data in East India. *Economic and Political Weekly* 13, 367–84.

Bardhan, P. and Srinivasan, T.N., eds. 1974. *Poverty and Income Distribution in India*. Calcutta: Statistical Publishing Society.

Basu, K. 1983. The emergence of isolation and interlinkage in rural markets. *Oxford Economic Papers* 35, 262–80.

Bhagwati, J. and Chakravarty, S. 1969. Contributions to Indian economic analysis: a survey. *American Economic Review* 59, 2–73.

Bhagwati, J. and Desai, P. 1970. *India: Planning for Industrialization*. London: Oxford University Press.

Bhagwati, J. and Ramaswami, V.K. 1963. Domestic distortions, tariffs, and the theory of optimum subsidy. *Journal of Political Economy* 71, 44–50.

Bhagwati, J. and Srinivasan, T.N. 1975. *Foreign Trade Regimes and Economic Development: India*. New York: Columbia University Press.

Bhagwati, J. and Srinivasan, T.N. 1980. Revenue seeking: a generalization of the theory of tariffs. *Journal of Political Economy* 88, 1069–87.

Bhagwati, J. and Srinivasan, T.N. 1993. *India's Economic Reforms*. New Delhi: Ministry of Finance, Department of Economic Affairs.

Bhalla, S. and Glewwe, P. 1986. Growth and equity in developing countries: a reinterpretation of the Sri Lanka experience. *World Bank Economic Review* 1, 35–63.

Bhalla, S. and Roy, P. 1988. Misspecification in farm productivity analysis: the role of land quality. *Oxford Economic Papers* 40, 55–73.

Binswanger, H. and Rosenzweig, M., eds. 1984. *Contractual Arrangements, Employment and Wages in Rural Labor Markets in Asia*. New Haven, CT: Yale University Press.

Boserup, E. 1965. *The Conditions of Agricultural Growth*. London: Allen & Unwin.

Bronfrenbrenner, M. 1960. A simplified Mahalanobis development model. *Economic Development and Cultural Change* 9, 45–51.

Chakravarty, S. 1974. Reflections on the growth process in the Indian economy. In *Some Problems of Indian Economic Policy*, 2nd edn. ed. C.D. Wadhwa. New Delhi: Tata-McGraw Hill.

Dandekar, V.M. and Rath, N. 1971. *Poverty in India*. Poona: Indian School of Political Economy.

Dasgupta, P. 1993. *An Inquiry into Well-Being and Destitution*. Oxford: Clarendon Press.

Desai, A.V. 1981. Factors underlying the slow growth of Indian industry. *Economic and Political Weekly* 16(10–12).

Dhar, P.N. and Lydall, H.F. 1961. *The Role of Small Enterprises in Indian Economic Development*. Bombay: Asia Publishing House.

Drèze, J. and Sen, A.K. 1989. *Hunger and Public Action*. Oxford: Clarendon Press.

Drèze, J. and Srinivasan, P.V. 1995. Widowhood and poverty in India: some inferences from survey data. Discussion Paper No. 62. Development Economics Research Program, London School of Economics.

Dutt, R.C. 1904. *The Economic History of India*, 2 vols. New Delhi: Publications Division, Ministry of Information and Broadcasting.

Ghatak, S. 1976. *Rural Money Markets in India*. New Delhi: Macmillan.

Gulati, A. 1998. Indian agriculture in an open economy: will it prosper? In Ahluwalia and Little (1998).

Jha, P.S. 1980. *The Political Economy of Stagnation*. New Delhi: Oxford University Press.

Joshi, V. and Little, I.M.D. 1994. *India: Macroeconomics and Political Economy 1964–1991*. Washington, DC: World Bank.

Joshi, V. and Sanyal, S. 2004. Foreign inflows and macroeconomic policy in India. In *NCAER/ Brookings: India Policy Forum 2004*, Washington, DC: Brookings Institution.

Komiya, R. 1959. A note on Professor Mahalanobis' model of Indian economic planning. *Review of Economics and Statistics* 41, 29–35.

Krishna, R. 1963. Farm supply response in India: Pakistan. *Economic Journal* 73, 477–87.

Krueger, A.O. 1974. The political economy of the rent-seeking society. *American Economic Review* 64, 291–303.

Kumar, D. and Desai, M., eds. 1983. *The Cambridge Economic History of India*, vol. 2: *c. 1750–1970*. Cambridge: Cambridge University Press.

Lal, D. 1972a. The foreign exchange bottleneck revisited. *Economic Development and Cultural Change* 20, 722–30.

Lal, D. 1972b. *Wells and Welfare: An Exploratory Cost-Benefit Study of the Economics of Small Scale Irrigation in Maharashtra*. Paris: OECD.

Lal, D. 1976. Agricultural growth, real wages and the rural poor in India. *Economic and Political Weekly* 11(26), A47–A61.

Lal, D. 1980. *Prices for Planning: Towards the Reform of Indian Planning*. London: Heinemann Educational Books.

Lal, D. 1984. *The Political Economy of the Predatory State*, Discussion Paper DRD 105, Development Research Department, Washington, DC: World Bank.

Lal, D. 1987. The political economy of economic liberalization. *World Bank Economic Review* 1, 273–99.

Lal, D. 1988. *The Hindu Equilibrium, vol.1: Cultural Stability and Economic Stagnation, India c. 1500 bc–ad 1980*. Oxford: Clarendon Press.

Lal, D. 1989. *The Hindu Equilibrium, vol. 2: Aspects of Indian Labour*. Oxford: Clarendon Press.

Lal, D. 2005. *The Hindu Equilibrium, India c. 1500 bc–2000 ad*, abridged and revised edn. Oxford: Oxford University Press.

Lal, D. 2006. India: population change and its consequences. *Population and Development Review* 32 (Supplement), 145–82.

Lal, D., Bery, S. and Pant, D.K. 2003. The real exchange rate, fiscal deficits and capital flows, India: 1981–2000. *Economic and Political Weekly* 38, 4965–76.

Lal, D., Bhide, S. and Vasudevan, D. 2001. Financial exuberance: savings deposits, fiscal deficits and interests rates in India. *Economic and Political Weekly* 37, 4196–203.

Lal, D., Cave, M., Hare, P. and Thompson, J. 1975. *Appraising Foreign Investment in Developing Countries*. London: Heinemann Educational Books.

Lal, D., Mohan, R. and Natarajan, I. 2001. Economic reforms and poverty alleviation: a tale of two surveys. *Economic and Political Weekly* 36, 1017–28.

Lal, D. and Myint, H. 1996. *The Political Economy of Poverty, Equity and Growth: A Comparative Study*. Oxford: Clarendon Press.

Lewis, J.P. 1963. *Quiet Crisis in India*. Bombay: Asia Publishing House.

Lipton, M. 1983. Labour and poverty. World Bank Staff Working Papers No. 616, Washington DC: World Bank.

Mahalanobis, P.C. 1953. Some observations on the process of growth in national income. *Sankhya* 12, 307–12.

Mahalanobis, P.C. 1955. The approach of operational research to planning in India. *Sankhya* 16, 3–130.

Mehra, S. 1966. Surplus labour in Indian agriculture. *Indian Economic Review* 1(1).

Minhas, B.S., Parikh, K. and Srinivasan, T.N. 1972. *Scheduling the Operations of the Bhakra System*. Calcutta: Statistical Publishing Society.

Mitra, A. 1977. *Terms of Trade and Class Relations*. London: Frank Cass.

Mitra, B.S. 2006. Grassroots capitalism thrives in India. In *2006 Index of Economic Freedom*, ed. M. Miles, K.R. Holmes and M.A. O'Grady. Washington, DC: Heritage Foundation.

Mohammad, J. and Whalley, J. 1984. Rent seeking in India: its costs and policy significance. *Kyklos* 37, 387–413.

Naoroji, D. 1901. *Poverty and Unbritish Rule in India*. London: Swan Sonnenschein and Co.

Narain, D. 1965. *The Impact of Price Movements on Areas of Selected Crops in India 1900–39*. Cambridge: Cambridge University Press.

Nehru, J. 1946. *The Discovery of India*. New York: John Day Co.

Parikh, K.S. 1993. *Who gets how much from the PDS – How Effectively does it Reach the Poor?* Mumbai: Indira Gandhi Institute of Development Research.

Patnaik, U. 1996. Export oriented agriculture and food security in developing countries and India. *Economic and Political Weekly* 31, 2429–49.

PROBE Team. 1999. *Public Report on Basic Education in India*. New Delhi: Oxford University Press.

Raj, K.N. 1973. The politics and economics of intermediate regimes. *Economic and Political Weekly* 8, 1189–98.

Raj, K.N. 1976. Trends in rural unemployment in India. *Economic and Political Weekly* 11, 1781–862.

Ravallion, M. 1991. Reaching the poor through public employment: arguments, evidence and lessons from South Asia. *World Bank Research Observer* 6(2), 153–75.

Rudra, A. 1975. *Indian Plan Models*. New Delhi: Allied Publishers.

Rudra, A. 1991. Privatization and deregulation. *Economic and Political Weekly* 27, 2933–6.

Rudra, A. and Sen, A.K. 1980. Farm size and labour use: analysis and policy. *Economic and Political Weekly* 15, 391–4.

Sanyal, S.K. 1977a. Trends in rural unemployment: a comment. *Economic and Political Weekly* 12, 145–8.

Sanyal, S.K. 1977b. Trends in some characteristics of landholdings: an analysis for a few states I and II. *Sarvekshana* 1(1–2).

Sau, R. 1981. *India's Economic Development: Aspects of Class Relations*. New Delhi: People's Publishing House.

Sen, A.K. 1966. Peasants and Dualism, with and without surplus labour. *Journal of Political Economy* 74, 425–50.

Sen, A.K. 1975. *Employment, Technology and Development*. Oxford: Oxford University Press.

Sen, A.K. 1976. Poverty: an ordinal approach to measurement. *Econometrica* 44, 219–31.

Sen, A.K. 1981a. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Clarendon Press.

Sen, A.K. 1981b. Public action and the quality of life in developing countries. *Oxford Bulletin of Economics and Statistics* 43, 287–318.

Singh, M. 1964. *India's Export Trends and the Prospects for Self Contained Growth*. Oxford: Clarendon Press.

Sivasubramonian, S. 2000. *The National Income of India in the Twentieth Century*. New Delhi: Oxford University Press.

Srinivasan, T.N. 1983. Hunger: defining it, estimating its global incidence and alleviating it. In *Role of Markets in the World Food Economy*, ed. D. Gale Johnson and E.G. Schuh. Boulder, CO: Westview Press.

Srinivasan, T.N. 1992. Privatization and deregulation. *Economic and Political Weekly* 27, 843–8.

Srinivasan, T.N. 1994. Destitution: a discourse. *Journal of Economic Literature* 32, 1842–55.

Srinivasan, T.N. 2001. *Economic Policy and State Intervention*. New Delhi: Oxford University Press.

Srinivasan, T.N. and Bardhan, P., eds. 1988. *Rural Poverty in South Asia*. New York: Columbia University Press.

Srinivasan, T.N., Bell, C. and Udry, C. 1997. Rationing, spillover, and interlinking in credit markets: the case of rural Punjab. *Oxford Economic Papers* 49, 557–85.

Sukhatme, P.V. 1978. Assessment of adequacy of diets of different economic levels. *Economic and Political Weekly* 13(Special number), 1373–84.

Tendulkar, S. 1998. Indian economic policy reforms and poverty: an assessment. In Ahluwalia and Little (1998).

Thakurdas, P.J., Birla, G., Dalal, A., Ram, S., Lalabhai, K., Shroff, A. and Matthai, J. 1944. *A Plan of Economic Development of India*. London: Penguin Books.

Vaidyanathan, A. 1977. Performance and prospects of crop production in India. *Economic and Political Weekly* 2, 1355–68.

Virmani, A. 2001. *India's 1990–91 Crisis: Reforms, Myths and Paradoxes*. New Delhi: Planning Commission, Government of India.

Visaria, P. 1977. Trends in rural unemployment in India: a comment. *Economic and Political Weekly* 12, 145–8.

Visveswarya, M. 1934. *Planned Economy for India*. Bangalore: Bangalore Press.

Volcker, J.A. 1893. *Report on the Development of Indian Agriculture*. Bombay.

Whitcombe, E. 1971. *Agrarian Conditions in Northern India*. Berkeley and Los Angeles: University of California Press.

World Bank. 2000. *India: Reducing Poverty, Accelerating Development*. New Delhi: Oxford University Press.

How to cite this article

Lal, Deepak. "India, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000267> doi:10.1057/9780230226203.0776

The New Palgrave Dictionary of Economics Online

indicative planning

Klaus Nielsen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Indicative planning aims to coordinate private and public investment and output plans through forecasts or targets. Compliance is voluntary. The underlying logic is that the plan can supply economically valuable information which, as a public good, the market mechanism cannot disseminate efficiently. It may be perceived as a substitute for non-existing forward markets. However, indicative planning takes into account only endogenous market uncertainty, not exogenous uncertainty (technology, foreign trade and so on). Indicative planning has been most consistently and continuously implemented in France and Japan but has been used in many other countries, although decreasingly so since the 1970s.

Keywords

Austrian economics; bounded rationality; forecasting; forward markets; general equilibrium; imperfect information; indicative planning; planning; rational expectations; uncertainty

Article

Indicative planning is a means of improving the performance of an economy through the elaboration of a set of consistent numerical forecasts or targets for the economic future. The aim is to coordinate private and public sector investment and output plans through the provision of economically valuable information. As distinct from directive central planning, as practised in the Soviet Union from the late 1920s, it is planning without compulsion. Compliance is purely voluntary. It is based on the idea that, if the plan is appropriately constructed, it will indicate an optimal path for the economy, which would then be spontaneously followed by the economic actors, without the need for compulsion. Decision-making is formally fully decentralized, but some versions of indicative planning include consultation with major private actors and the concertation of private investment plans. Furthermore, compliance is encouraged and facilitated by persuasion and cognitive framing and is sometimes supported by incentives. In addition, state-controlled investment funds may be guided into favoured projects in accordance with the plan. Furthermore, public sector commitment to implement planned public investment and output targets

may constitute an element of certainty that facilitates the intended voluntary compliance.

The best-known examples of indicative planning are the plans elaborated by the French Commissariat Général du Plan and the Japanese Planning Agency since the Second World War. After the Second World War several European countries, such as the Netherlands, developed some sort of indicative planning, often linked to the building of multi-sector econometric models of the economy. Indicative planning was widely practised in developing countries during the post-war period until the 1980s (Belassa, 1990). After the collapse of Communism, indicative planning was briefly adopted in Poland, and is still being used in some of the former republics of the Soviet Union. In 1965 an indicative National Plan was implemented in the United Kingdom, but was abandoned after a year as an effect of a balance of payments crisis. Today (in 2007), the European Union is involved in soft coordination activities that have some resemblance to indicative planning.

The presence of imperfect information is a market failure, and indicative planning can be seen as an attempt to bridge the information gap. The underlying logic is that the plan can supply economically valuable information which, as a public good, the market mechanism does not disseminate efficiently. Indicative planning makes it possible to overcome the problems that arise from the economic actors' ignorance of the intentions of the other actors. The collective market research involved in indicative planning should, in principle, make it possible to anticipate potential overcapacity and shortage and to avoid states of disequilibrium with unfulfilled expectations. If every economic actor informs the planners about their prospective demand and supply intentions for the forthcoming plan period, this information could be aggregated into an indicative plan and appropriate adaptations could be made by the economic actors.

The indicative plan may be perceived as a substitute for non-existing forward markets, or as a calculated general equilibrium representing an optimal allocation of resources that it would be in everybody's interest to implement on condition that the plan was correctly worked out. J.L. Meade (1971) demonstrates that the optimality features of the welfare-maximizing general equilibrium model can be obtained even if a full set of forward markets does not exist, provided that the economic agents make honest non-binding declarations about intended actions for any future date. Based on this information, equilibrium prices and quantities could be calculated and the forecasts of the indicative plan would necessarily be realized, since they correspond to optimal behaviour by market agents.

However, the assumption that agents declare their true intentions contradicts the assumption of rational behaviour if individual agents are large enough to influence prices that provide them with an incentive not to reveal their true preferences. Furthermore, indicative planning is capable of taking into account only endogenous market uncertainty, and works only in a closed economy. Environmental, or exogenous, uncertainty (including changes in technology and foreign trade) is ignored. In theory, the indicative plan may operate with as many future paths as there are possible scenarios for the exogenous environment. However, this procedure for transformation of uncertainty to risk is hardly of any practical relevance, and it does not recognize the existence of genuine uncertainty that makes it impossible to elaborate appropriate scenarios, even in theory.

Economic internationalization and technological change have the effect that the overwhelming source of uncertainty has become exogenous, which has made the forecasting exercises of indicative planning increasingly difficult and ultimately useless. As a result, indicative planning has been widely abandoned or its ambitions have been significantly curtailed. France is the major example of a continuous commitment to indicative planning. Until 2006, planning documents covering successive five-year

planning periods were elaborated by the Commissariat Général du Plan. However, from the early 1970s and onwards, the plans became less ambitious and less influential. Targets and concertation were abandoned. The plans became internal governmental strategic documents that were, from 1993, no longer presented to Parliament. From 2006, indicative planning was formally abandoned, and the Commissariat Général du Plan was succeeded by a new Centre d'Analyse Stratégique.

It is fair to ask whether indicative planning, following its almost universal decline, is now devoid of contemporary relevance, if it ever had some, and has become a phenomenon of merely historical interest. Is indicative planning irrelevant, even in its less comprehensive and more pragmatic version that stresses the virtues of its contribution to develop shared expectations, or ‘a common view of the future?’ If the economic agents are seen as capable of developing rational expectations there is surely no role for indicative planning. In this view, attempts to influence expectations are ineffective and wasteful. From the point of view of Austrian economics, collective forecasting is even worse; it is not only ineffective but harmful. Indicative planning can be misleading, which may lead to too many eggs being put into one wrong basket. The plurality of information in a world of decentralized decision-making with no public attempts to influence expectations is seen as preferable by far.

However, from a more pragmatic point of view, it is exactly the role of indicative planning in forming common expectations concerning macroeconomic development trends that may contribute, not to the achievement of the nirvana of an optimal growth path, but rather to an improved state of disequilibrium (Holmes, 1987). If optimal equilibrium is seen to be of little practical relevance as a result of widespread genuine uncertainty and the bounded rationality of economic agents, pragmatic means to improve the situation are important, although these may not in any way be seen as leading to a utopian state of optimal allocation of resources. At least three factors make indicative planning in the form of macroeconomic forecasts highly valuable in this context: (*a*) the public good character of the collected information, (*b*) the economies of scale of information processing, and (*c*) the fact that the government is no doubt a particularly well-informed actor in relation to macroeconomic developments.

See Also

- forecasting
- market failure
- planning
- public goods
- uncertainty

Bibliography

Belassa, B. 1990. Indicative planning in developing countries. Policy Research Working Paper No. 439. Washington, DC: World Bank.

Holmes, P. 1987. Indicative planning. In *The New Palgrave Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Meade, J.E. 1971. *The Theory of Indicative Planning*. Manchester: Manchester University Press.

How to cite this article

Nielsen, Klaus. "indicative planning." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000060> doi:10.1057/9780230226203.0777

The New Palgrave Dictionary of Economics Online

indirect inference

Anthony A. Smith, Jr.

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Indirect inference is a simulation-based method for estimating the parameters of economic models. Its hallmark is the use of an auxiliary model to capture aspects of the data upon which to base the estimation. The parameters of the auxiliary model can be estimated using either the observed data or data simulated from the economic model. Indirect inference chooses the parameters of the economic model so that these two estimates of the parameters of the auxiliary model are as close as possible. The auxiliary model need not be correctly specified; when it is, indirect inference is equivalent to maximum likelihood.

Keywords

auxiliary models; Bayesian inference; criterion functions; discrete-choice models; dynamic stochastic general equilibrium (DSGE) models; estimation; Lagrange multipliers; likelihood; likelihood ratios; linear probability models; maximum likelihood; models; probability density functions; reduced-form models; semi-nonparametric (SNP) models; simulated moments estimation; simultaneous equations; vector autoregressions; Wald test

Article

Indirect inference is a simulation-based method for estimating, or making inferences about, the parameters of economic models. It is most useful in estimating models for which the likelihood function (or any other criterion function that might form the basis of estimation) is analytically intractable or too difficult to evaluate. Such models abound in modern economic analysis and include nonlinear dynamic models, models with latent (or unobserved) variables, and models with missing or incomplete data. Like other simulation-based methods, indirect inference requires only that it be possible to simulate data from the economic model for different values of its parameters. Unlike other simulation-based methods, indirect inference uses an approximate, or auxiliary, model to form a criterion function. The auxiliary

model does not need to be an accurate description of the data generating process. Instead, the auxiliary model serves as a window through which to view both the actual, observed data and the simulated data generated by the economic model: it selects aspects of the data upon which to focus the analysis.
The goal of indirect inference is to choose the parameters of the economic model so that the observed data and the simulated data look the same from the vantage point of the chosen window (or auxiliary model). In practice, the auxiliary model is itself characterized by a set of parameters. These parameters can themselves be estimated using either the observed data or the simulated data. Indirect inference chooses the parameters of the underlying economic model so that these two sets of estimates of the parameters of the auxiliary model are as close as possible.

A formal definition

To put these ideas in concrete form, suppose that the economic model takes the form:

$$y_t = G(y_{t-1}, x_t, u_t; \beta), \; t = 1, 2, \ldots, T, \tag{1}$$

where $\{x_t\}_{t=1}^{T}$ is a sequence of observed exogenous variables, $\{y_t\}_{t=1}^{T}$ is a sequence of observed endogenous variables, and $\{u_t\}_{t=1}^{T}$ is a sequence of unobserved random errors. Assume that the initial value y_0 is known and that the random errors are independent and identically distributed (i.i.d.) with a known probability distribution F. Equation (1) determines, in effect, a probability density function for y_t conditional on y_{t-1} and x_t. Indirect inference does not require analytical tractability of this density, relying instead on numerical simulation of the economic model. This is not the most general model that indirect inference can accommodate – indirect inference can be used to estimate virtually any model from which it is possible to simulate data – but it is a useful starting point for understanding the principles underlying indirect inference. The econometrician seeks to use the observed data to estimate the k-dimensional parameter vector β.
The auxiliary model, in turn, is defined by a conditional probability density function, $f(y_t|y_{t-1}, x_t, \theta)$, which depends on a p-dimensional parameter vector θ. In a typical application of indirect inference, this density has a convenient analytical expression. The number of parameters in the auxiliary model must be at least as large as the number of parameters in the economic model (that is, $p \geq k$).
The auxiliary model is, in general, incorrectly specified: that is, the density f need not describe accurately the conditional distribution of y_t determined by eq. (1). Nonetheless, the parameters of the auxiliary model can be estimated using the observed data by maximizing the log of the likelihood function defined by f:

$$\hat{\theta} = \arg\max_{\theta} \sum_{t=1}^{T} \log\ f(y_t | y_{t-1}, x_t, \theta).$$

The estimated parameter vector $\hat{\theta}$ serves as a set of 'statistics' that capture, or summarize, certain features of the observed data; indirect inference chooses the parameters of the economic model to reproduce this set of statistics as closely as possible.

The parameters of the auxiliary model can also be estimated using simulated data generated by the economic model. First, using a random number generator, draw a sequence of random errors $\{\tilde{\nu}_t^m\}_{t=1}^{T}$ from the distribution *F*. Typically, indirect inference uses *M* such sequences, so the superscript *m* indicates the number of the simulation. These sequences are drawn only once and then held fixed throughout the estimation procedure. Second, pick a parameter vector β and then iterate on eq. (1), using the observed exogenous variables and the simulated random errors, to generate a simulated sequence of endogenous variables: $\{\tilde{y}_t^m(\beta)\}_{t=1}^{T}$, where the dependence of this simulated sequence on β is made explicit. Third and finally, maximize the average of the log of the likelihood across the *M* simulations to obtain:

$$\hat{\theta}(\beta) = \arg\max_{\theta} \sum_{m=1}^{M} \sum_{t=1}^{T} \log\ f(\tilde{y}_t^m(\beta) | \tilde{y}_{t-1}^m(\beta), x_t, \theta).$$

The central idea of indirect inference is to choose β so that $\hat{\theta}(\beta)$ and $\hat{\theta}$ are as close as possible. When the economic model is exactly identified (that is, when $p = k$), it is, in general, possible to choose β so that the economic model reproduces exactly the estimated parameters of the auxiliary model. Typically, though, the economic model is over-identified (that is, $p > k$): in this case, it is necessary to choose a metric for measuring the distance between $\hat{\theta}$ and $\hat{\theta}(\beta)$; indirect inference then picks β to minimize this distance.

As the observed sample size *T* grows large (with *M* held fixed), the estimated parameter vector in the simulated data, $\hat{\theta}(\beta)$, converges to a so-called 'pseudo-true value' that depends on β; call it $h(\beta)$. The function h is sometimes called the binding function: it maps the parameters of the economic model into the parameters of the auxiliary model. Similarly, the estimated parameter vector in the observed data, $\hat{\theta}$, converges to a pseudo-true value θ_0. In the limit as *T* grows large, then, indirect inference chooses β to satisfy the equation $\theta_0 = h(\beta)$. Under the assumption that the observed data is generated by the economic model for a particular value, β_0, of its parameter vector, the value of β that satisfies this equation is precisely β_0. This heuristic argument explains why indirect inference generates consistent estimates of the parameters of the economic model.

Three examples

Example 1: a simple system of simultaneous equations

The first example is drawn from the classical literature on simultaneous equations to which indirect inference is, in many ways, a close cousin. Consider a simple macroeconomic model, adapted from Johnston (1984), with two simultaneous equations: $C_t = \beta Y_t + u_t$ and $Y_t = C_t + X_t$. In this model, consumption expenditure in period *t*, C_t, and output (or income) in period *t*, Y_t, are endogenous, whereas non-consumption expenditure in period *t*, X_t, is exogenous. Assume that the random error u_t is i.i.d. and normally distributed with mean zero and a known variance; the only unknown parameter, then, is β. There are many ways to estimate β without using indirect inference, but this example is useful for illustrating how indirect inference works. To wit, suppose that the auxiliary model specifies that C_t is normally distributed with conditional mean θX_t and a fixed variance. In this simple example, the binding function can be computed without using simulation: a little algebra reveals that $\theta = \beta / (1 - \beta) \equiv h(\beta)$. To estimate β, first use ordinary least squares (which is equivalent to maximum likelihood in this example) to obtain a consistent estimate, $\hat{\theta}$, of θ. Then evaluate the inverse of *h* at $\hat{\theta}$ to obtain a consistent estimate of β: $\hat{\beta} = \hat{\theta} / (1 + \hat{\theta})$. This is precisely the indirect inference estimator of β. This estimator uses an indirect approach: it first estimates an auxiliary (or, in the language of simultaneous equations, a reduced-form) model whose parameters are complicated functions of the parameters of the underlying economic model and then works backwards to recover estimates of these parameters.

Example 2: a general equilibrium model of the macroeconomy

In this example, the economic model is a dynamic, stochastic, general equilibrium (DSGE) model of the macroeconomy (for a prototype, see Hansen, 1985). Given choices for the parameters describing the economic environment, this class of models determines the evolution of aggregate macroeconomic time series such as output, consumption, and the capital stock. The law of motion for these variables implied by the economic model is, in general, nonlinear. In addition, some of the key variables in this law of motion (for example, the capital stock) are poorly measured or even unobserved. For these reasons, in these models it is often difficult to obtain a closed-form expression for the likelihood function.
To surmount these obstacles, indirect inference can be used to obtain estimates of the parameters of the economic model. A natural choice for the auxiliary model is a vector autoregression (VAR) for the variables of interest. As an example, let y_t be a vector containing the values of output and consumption in period *t* (expressed as deviations from steady-state values) and let the VAR for y_t have one lag: $y_{t+1} = Ay_t + \varepsilon_{t+1}$, where the ε_ts are normally distributed, i.i.d. random variables with mean 0 and covariance matrix Σ.
In this example, the binding function maps the parameters of the economic model into the parameters *A* and Σ of the VAR. To obtain a simulated approximation to the binding function, pick a set of parameters for the economic model, compute the law of motion implied by this set of parameters,

simulate data using this law of motion, and then use OLS to fit a VAR to the simulated data. Indirect inference chooses the parameters of the economic model so that the VAR parameters implied by the model are as close as possible to the VAR parameters estimated using observed macroeconomic time series. Smith (1993) illustrates the use of indirect inference to estimate DSGE models.

Example 3: a discrete-choice model

In this example, the economic model describes the behaviour of a decision-maker who must choose one of several discrete alternatives. These models typically specify a random utility for each alternative; the decision-maker is assumed to pick the alternative with the highest utility. The random utilities are latent: the econometrician does not observe them, but does observe the decision-maker's choice. Except in special cases, evaluating the likelihood of the observed discrete choices requires the evaluation of high-dimensional integrals which do not have closed-form expressions.

To use indirect inference to estimate discrete-choice models, one possible choice for the auxiliary model is a linear probability model. In this case, the binding function maps the parameters describing the probability distribution of the latent random utilities into the parameters of the linear probability model. Indirect inference chooses the parameters of the economic model so that the estimated parameters of the linear probability model using the observed data are as close as possible to those obtained using the simulated data. Implementing indirect inference in discrete-choice models poses a potentially difficult computational problem because it requires the optimization of a non-smooth objective function. Keane and Smith (2003), who illustrate the use of indirect inference to estimate discrete-choice models, also suggest a way to smooth the objective surface.

Three metrics

To implement indirect inference when the economic model is over-identified, it is necessary to choose a metric for measuring the distance between the auxiliary model parameters estimated using the observed data and the simulated data, respectively. There are three possibilities corresponding to the three classical hypothesis tests: Wald, likelihood ratio (LR), and Lagrange multiplier (LM).

In the Wald approach, the indirect inference estimator of the parameters of the economic model minimizes a quadratic form in the difference between the two vectors of estimated parameters:

$$\hat{\beta}^{Wald} = \arg\min_{\beta} (\hat{\theta} - \hat{\theta}(\beta))' W (\hat{\theta} - \hat{\theta}(\beta)),$$

where W is a positive definite ‘weighting’ matrix.

The LR approach to indirect inference forms a metric using the (approximate) likelihood function defined by the auxiliary model. In particular,

$$\hat{\beta}^{LR} = \arg\min_{\beta}\left(\sum_{t=1}^{T}\log\ f(y_t|y_{t-1}, x_t, \hat{\theta}) - \sum_{t=1}^{T}\log\ f(y_t|y_{t-1}, x_t, \hat{\theta}(\beta))\ .\right).$$

By the definition of $\hat{\theta}$, the objective function on the right-hand side is non-negative, and its value approaches zero as $\hat{\theta}(\beta)$ approaches $\hat{\theta}$. The LR approach to indirect inference chooses β so as to make this value as close to zero as possible. Because the first term on the right-hand side does not depend on β, the LR approach can also be viewed as maximizing the approximate likelihood subject to the restrictions, summarized (for large *T*) by the binding function *h*, that the economic model imposes on the parameters of the auxiliary model.

Finally, the LM approach to indirect inference forms a metric using the derivative (or score) of the log of the likelihood function defined by the auxiliary model. In particular,

$$\hat{\beta}^{LM} = \arg\min_{\beta} S(\beta)'VS(\beta),$$

where

$$S(\beta) = \sum_{m=1}^{M}\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\log\ f(\tilde{y}_t^m(\beta), x_t, \hat{\theta})$$

and *V* is a positive definite matrix. By definition, $\hat{\theta}$ sets the score in the observed data to zero. The goal of the LM approach, then, is to choose β so that the (average) score in the simulated data, evaluated at $\hat{\theta}$, is as close to zero as possible.

For any number, *M*, of simulated data-sets, all three approaches deliver consistent and asymptotically normal estimates of β as *T* grows large. The use of simulation inflates asymptotic standard errors by the factor $(1 + M^{-1})^{1/2}$; for $M \geq 10$, this factor is negligible. When the economic model is exactly identified, all three approaches to indirect inference yield numerically identical estimates; in this case, they all choose β to solve $\hat{\theta}(\beta) = \hat{\theta}$.

When the economic model is over-identified, the minimized values of the three metrics are, in general, greater than zero. These minimized values can be used to test the hypothesis that the economic model is correctly specified: sufficiently large minimized values constitute evidence against the economic model.

If the weighting matrices *W* and *V* are chosen appropriately, then the Wald and LM approaches are asymptotically equivalent in the sense that they have the same asymptotic covariance matrix; by

contrast, the LR approach, in general, has a larger asymptotic covariance matrix. If, however, the auxiliary model is correctly specified, then all three approaches are asymptotically equivalent not only to each other but also to maximum likelihood (for large *M*). Because maximum likelihood is asymptotically efficient (that is, its asymptotic covariance matrix is as small as possible), the LM approach is sometimes called the ‘efficient method of moments’ when the auxiliary model is close to being correctly specified; in such a case, this name could also be applied to the Wald approach.
When estimating the parameters of the auxiliary model is difficult or time-consuming, the LM approach has an important computational advantage over the other two approaches. In particular, it does not require that the auxiliary model be estimated repeatedly for different values of the parameters of the economic model. To estimate continuous-time models of asset prices, for example, Gallant and Tauchen (2005) advocate using a semi-nonparametric (SNP) model as the auxiliary model. As the number of its parameters increases, an SNP model provides an arbitrarily accurate approximation to the data generating process, thereby permitting indirect inference to approach the asymptotic efficiency of maximum likelihood. For this class of auxiliary models, which are nonlinear and often have a large number of parameters, the LM approach is a computationally attractive way to implement indirect inference.

Concluding remarks

Indirect inference is a simulation-based method for estimating the parameters of economic models. Like other simulation-based methods, such as simulated moments estimation (see, for example, Duffie and Singleton, 1993), it requires little analytical tractability, relying instead on numerical simulation of the economic model. Unlike other methods, the ‘moments’ that guide the estimation of the parameters of the economic model are themselves the parameters of an auxiliary model. If the auxiliary model comes close to providing a correct statistical description of the economic model, then indirect inference comes close to matching the asymptotic efficiency of maximum likelihood. In many applications, however, the auxiliary model is chosen, not to provide a good statistical description of the economic model, but instead to select important features of the data upon which to focus the analysis.
There is a large literature on indirect inference, much of which is beyond the scope of this article. Gouriéroux and Monfort (1996) provide a useful survey of indirect inference. Indirect inference was first introduced by Smith (1990; 1993) and later extended in important ways by Gouriéroux, Monfort, and Renault (1993) and Gallant and Tauchen (1996). Although indirect inference is a classical estimation method, Gallant and McCulloch (2004) show how ideas from indirect inference can be used to conduct Bayesian inference in models with intractable likelihood functions. There have been many interesting applications of indirect inference to the estimation of economic models, mainly in finance, macroeconomics, and labour economics. Because of its flexibility, indirect inference can be a useful way to estimate models in all areas of economics.

See Also

- maximum likelihood

Bibliography

Duffie, D. and Singleton, K.J. 1993. Simulated moments estimation of Markov models of asset prices. *Econometrica* 61, 929–52.

Gallant, A.R. and McCulloch, R. 2004. On the determination of general scientific models. Working paper, Fuqua School of Business, Duke University.

Gallant, A.R. and Tauchen, G. 1996. Which moments to match? *Econometric Theory* 12, 657–81.

Gallant, A.R. and Tauchen, G. 2005. Simulated score methods and indirect inference for continuous-time models. In *Handbook of Financial Econometrics*, ed. Y. Ait-Sahalia and L. Hansen. Amsterdam: North-Holland.

Gouriéroux, C. and Monfort, A. 1996. *Simulation-Based Econometric Methods*. New York: Oxford University Press.

Gouriéroux, C., Monfort, A. and Renault, E. 1993. Indirect inference. *Journal of Applied Econometrics* 8, S85–S118.

Hansen, G.D. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16, 402–17.

Johnston, J. 1984. *Econometric Methods*, 3rd edn. New York: McGraw-Hill Book Company.

Keane, M. and Smith, A.A., Jr. 2003. Generalized indirect inference for discrete choice models. Working paper, Yale University.

Smith, A.A., Jr. 1990. Three essays on the solution and estimation of dynamic macroeconomic models. Doctoral dissertation, Duke University.

Smith, A.A., Jr. 1993. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8, S63–S84.

How to cite this article

Smith, Anthony A., Jr. "indirect inference." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000259> doi:10.1057/9780230226203.0778

The New Palgrave Dictionary of Economics Online

indirect utility function

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

After many independent discoveries that were widely separated in time and space, the indirect utility function has in the last 35 years gradually become a standard part of demand theory. Its first discovery was made as early as 1886 by Antonelli in Italy, who also derived what has come to be known as Roy's Identity (see Chipman's introduction to the translation of Antonelli (1886) in Chipman et al., 1971). Later contributions came from Konyus (1924, 1926) and Byushgens in Russia, from Hotelling (1932) and Court (1941, pp. 284–97) in the United States, from Roy (1942, 1947) and Ville (1946) in France, and from Wold (1943–4) and Malmquist (1953) in Sweden; a good brief history may be found in Diewert (1982, pp. 547–50).
But it was not until the early 1950s and the contributions of Houthakker (1951–2, 1960) that the indirect utility function became an integral part of the theory of consumer's behaviour. Indeed, the very names in standard use appear to be due to him, ‘indirect utility function’ in (1951–2, p. 157) and ‘Roy's Identity’ in (1960, p. 250).

1 Definition and simple properties

Suppose that the consumer has completely preordered preferences defined over the commodity space $R^n{}^+$ of non-negative bundles $x=(x_1, x_2,\ldots, x_n)$, that those preferences are representable by a real-valued utility function u, that he (or she) faces competitively determined positive money prices $(p_1, p_2,\ldots, p_n)=p$ for the n goods, and has exogenously determined monetary wealth $\omega>0$. It is standard in demand theory to assume that the consumer chooses a bundle x^* by solving the optimization problem:

$$\text{Max}(p, \omega) \text{ Find } x \in R^{n+} \text{ to max } u(x) \text{ subject to } (p, x) \leq \omega \tag{1}$$

where the notation $\langle\cdot,\cdot\rangle$ means the inner product of the two vectors concerned.
Assume that Max(p, ω) has a unique solution x^*, for which it suffices that preferences be monotonically increasing and strictly convex. Then the number

$$v^* = u(x^*) \tag{2}$$

is the *value* of Max(p, ω). This joint determination of solution and value once p and ω are known implies the existence of two functions of the price-wealth pair (p, ω), called respectively the *ordinary* (or Marshallian) *demand function* f: $R^{n++}\times R^{++}\rightarrow R^{n++}$, defined by

$$x^* = f(p, \omega) \tag{3}$$

and the *indirect utility function* v: $R^{n++}\times R^{++}\rightarrow R$, defined by

$$v^* = v(p, \omega) \tag{4}$$

Define the attainable (or budget) set $A(p, \omega)$ by

$$A(p, \omega) = \left\{x \in R^{n+} : \langle p, x\rangle \leq \omega\right\}$$

From (1) it follows that for any $\lambda>0$, $A(\lambda p, \lambda\omega)=A(p, \omega)$, so that both f and v are positively homogeneous of degree zero in (p, ω). Next, if $(p^1-p^2)\in R^{n+}$ and $p^1 \neq p^2$ then $A(p^1, \omega)\subset A(p^2, \omega)$, from which $v(p^1, \omega)\leq v(p^2, \omega)$; for similar reasons, $v(p, \cdot)$ is nondecreasing. It can be shown further that if u is continuous then so is v (see e.g. Varian, 1984, pp. 121, 326–7).
A useful result is that $v(\cdot, \omega)$ is quasi-convex. To prove this let $p^t=tp^1+(1-t)p^2$, where $t\in[0, 1]$. Then for

any $x \in A(p^t, \omega)$,

$$t\langle p^1, x\rangle + (1-t)\langle p^2, x\rangle < \omega. \tag{5}$$

If t=0, $x \in A(p^2, \omega)$, while if t=1, $x \in A(p^1, \omega)$. Otherwise, suppose that x is in neither $A(p^1, \omega)$ nor $A(p^2, \omega)$. Then $t\langle p^1, x\rangle > t\omega$ and $(1-t)\langle p^2, x\rangle > (1-t)\omega$, which on addition yield a contradiction to (5). So x is in either $A(p^1, \omega)$ or $A(p^2, \omega)$. Hence $v(p^t, \omega)$, which is the sup of $u(\cdot)$ on $A(p^1, \omega)$, can be no larger than $\max[v(p^1, \omega), v(p^2, \omega)]$, which are themselves the sups of $u(\cdot)$ on $A(p^1, \omega)$ and $A(p^2, \omega)$, respectively. But the condition $v(p^t, \omega) \leq \max[v(p^1, \omega), v(p^2, \omega)]$ is the original definition of the quasi-convexity of the function $v(\cdot, \omega)$ (see Fenchel, 1953, p. 117).

2 Relations between the ordinary demand functions and the indirect utility function

For simplicity, the following assumptions are made: (a) x^* is a strictly positive vector. (b) Each function involved is as differentiable as required. (c) At any $x \in R^{n+}$ there is at least one commodity in which u is strictly increasing (this implies local non-satiation of preferences).
Suppose that at x^* the constraint (1) is 'slack', i.e. $\omega - \langle p, x^*\rangle = \delta > 0$. Let k be any good with property (c) at x^*, and define a new bundle x^1 by putting $x_i^i = x_i^*$ for $i \neq k$, and $x_k^1 = x_k^* + (\delta / p_k)$. Then by construction $\langle p, x^1\rangle = \omega$, while from (c) $u(x^1) > u(x^*)$, contradicting the hypothesis that x^* solves $\max(p, \omega)$. So

$$\langle p, x^*\rangle = \omega \tag{6}$$

Next, define L: $R^{n+} \times R^{n++} \rightarrow R$ by

$$L(x^1, p^1) = v(p^1, \langle p^1, x^1\rangle) - u(x^1) \tag{7}$$

where x^1 and p^1 are arbitrary. From (2) and (4), for any x^1 the value $v(p^1, \langle p^1\ x^1\rangle)$ is the *maximized* level of utility when prices are p^1 and wealth $\langle p^1, x^1\rangle > 0$. Hence, $L(\cdot, p^1)$ is positive semi-definite, i.e. $x^1 \in R^{n+}$

implies $L(x^1, p^1) \geq 0$ for any p^1. Putting $p^1=p$, the actual prices, if x^* solves Max(p, ω) it follows from (2), (4) and (6) that

$$L(x^*, p) = v(p, (p, x^*)) - u(x^*) = 0 \quad (8)$$

Hence x^* attains the infimum of $L(\cdot, p)$. So from (6), (8) and the Chain Rule,

$$\forall\ i = 1, 2, \ldots, n\ v\omega(p, \omega)\, p_i = u_i(x^*) \quad (9)$$

From (c), $u_i(x^*)>0$ for at least one i. Since $p_i>0$ this implies the simple but important result

$$v\omega(p, \omega) > 0 \quad (10)$$

i.e. the marginal utility of wealth is positive.
From (2), (3) and (4) the equation

$$v(p, \omega) = u(f(p, \omega))$$

is an identity in (p, ω). So differentiating each of the individual demand functions f_i with respect to (wrt) each p_j and ω yields,

$$\forall\ j = 1, 2, \ldots, n\ v_j(p, \omega) = \Sigma u_i(x^*) f_{ij}(p, \omega)$$

$$v\omega(p, \omega) = \Sigma u_i(x^*) f_i\omega(p, \omega)$$
(11)

From (6) and (3),

$$\{p, f(p, \omega)\} = \omega$$

This is another identity in (p, ω), and differentiating it wrt each p_j and ω results in

$$\forall \; j = 1, 2, \ldots, n f_j(p, \omega) + \Sigma p_i f_{ij}(p, \omega) = 0$$
(12)

$$\Sigma p_i f_i\omega(p, \omega) = 1$$

From (11) and (9),

$$\forall \; j = 1, 2, \ldots, n v_j(p, \omega) = v\omega(p, \omega)\Sigma p_i f_{ij}(p, \omega)$$

and from this and (12),

$$\forall \; j = 1, 2, \ldots, n v_j(p, \omega) = - f_j(p, \omega) v\omega(p\omega)$$
(13)

Equation (13) is the main result connecting *v* with *f*. From (3), (a), (10) and (13) there follow

$$\forall\ j = 1, 2, \ldots, n\ v_j(p, \omega) < 0 \tag{14}$$

and Roy's Identity (1942, p. 24; 1947, p. 217),

$$\forall\ j = 1, 2, \ldots, n\ x_j^* = -v_j(p, \omega)/v\omega(p, \omega) \tag{15}$$

As deservedly famous as is (15), its equivalent version (13) reveals the structures involved more clearly, since it focuses sharply on the relations between the functions *v* and *f* rather than the particular quantities x_j^*. In each of Roy's contributions the identity is first given in the form $v_j(p, \omega)/x_j^* = -v\omega(p, \omega)$, and is used primarily to prove (14); later, in Roy (1947, p. 220), the identity takes the more usual form (15). Since (13) is an identity, differentiating it wrt any p_i yields $\forall\ i, j=1, 2,\ldots, n$

$$-v_{ji}(p, \omega) = f_{ji}(p, \omega)v\omega(p, \omega) + f_j(p, \omega)v\omega_i(p, \omega)$$

Applying Young's Theorem to these equations, by symmetry,

$$\forall\ i, j = 1, 2, \ldots, n\ f_{ij}(p, \omega)v\omega(p, \omega) + f_i(p, \omega)v_{\omega j}(p, \omega) = f_{ij}(p, \omega)v_{\omega i}(p, \omega) \tag{16}$$

Now make the quite restrictive assumption that for each p_i, $v_{\omega i}(p, \omega)=0$; this requires in effect that each good have unitary elasticity of demand (see Samuelson, 1942, pp. 80–81). Then from (10) and (16).

$$\forall\ i, j = 1, 2, \ldots, n\ f_{ij}(p, \omega) = f_{ij}(p, \omega)$$

which are Slutsky-like equations that apply not to compensated but to ordinary demand functions.

3 Relations with the cost function and the compensated demand functions

Suppose now that a *target* level of utility is specified and the following new optimization problem posed:

$$\text{Min}(p, \tau): \text{Find } x \in R^{n+} \text{ to min}(p, x) \text{ subject to } u(x) \geq \tau \tag{17}$$

Assume that a unique solution x^{**} to this problem exists, yielding a value $\langle p, x^{**}\rangle$. This implies the existence of two functions of the price-target pair (p, τ), called the *compensated* (*or Hicksian*) demand function h: $R^{n++}\times R\rightarrow R^{n+}$, defined by

$$x^{**} = h(p, \tau) \tag{18}$$

and the *cost* (*or expenditure*) function γ : $R^{n++}\times R\rightarrow R^{+}$, given by

$$(p, x^{**}) = \gamma(p, \tau) \tag{19}$$

Retain assumptions (a)–(c), replacing x^* by x^{**}. Define M: $R^{n+}\times R^{n++}\rightarrow R$ by putting

$$M(x^1, p^1) = \gamma(p^1, u(x^1)) - (p^1, x^1) \tag{20}$$

where x^1 and p^1 are arbitrary, as before. It follows that $M(\cdot, p^1)$ is negative semi-definite. Putting $p^1=p$, the actual prices, it follows that if x^{**} solves Max (p, ω) then

$$M(x^{**}, p) = \gamma(p, u(x^{**})) - \langle p, x^{**} \rangle = 0$$
(21)

so that x^{**} maximizes $M(\cdot, p)$. Then a development *exactly* like that of the last section leads to a simple but basic result on the interrelations betwen h and γ, namely:

$$\forall\ j = 1, 2, \ldots, n\ \gamma_j(p, \tau) = h_j(p, \tau)$$
(22)

where h_j is the compensated demand function for the jth good. From (22) and (a),

$$\forall\ j = 1, 2, \ldots, n\ \gamma_j(p, \tau) > 0$$
(23)

From (18), (22) can be rewritten in the more customary version that has come to be called *Shephard's Lemma* (Shephard, 1953), although it dates back at least to Hotelling (1932).

$$\forall\ j = 1, 2, \ldots, n\ x_j^{**} = \gamma_j(p, \tau)$$
(24)

Thus (22) (or the Lemma) plays a role in the analysis of this problem which is symmetrical to that played by (13) (or Roy's Identity) in the analysis of Max(p, ω).
However, there are two important structural asymmetries between the problems max(p, ω) and min(p, τ). First, suppose that for some reason (such as incompleteness of preferences) the utility function u does not exist, so that v does not exist either. Clearly, since Max(p, ω) requires a scalar measure of utility it cannot be defined in this new situation. However, by replacing the target level τ of utility by a target bundle x^{τ}, one can still define a perfectly sensible minimum problem Min(p, x^{τ}).
The second asymmetry is that while $v(\cdot, \omega)$ is only quasiconvex, $\gamma(\cdot, \tau)$ is actually concave, and this without any assumptions on preferences. Since (full) concavity imposes sharper restrictions on any function than does quasi-convexity, the analysis of Min (p, τ) (or of Min (p, x^{τ})) yields easier proofs

of basic results than does that of Max (p, ω). For example, from (22) $h_{jj}(p, \tau) = \gamma_{jj}(p, \tau)$, and since $\gamma(\cdot, \tau)$ is concave $\gamma_{jj}(p, \tau) \leq 0$, proving that the substitution effect is non-positive.

4 Duality

It is not productive to oppose the virtues of minimum problems to those of maximum problems. Indeed, the most efficient path of the derivation of such propositions as the 'Fundamental Equation or Value Theory' (Hicks, 1939, p. 309) is by a judicious mixture of the two, i.e. by first solving max(p, ω) to obtain $\tau^* = v(p, \omega)$ and $x^* = f(p, \omega)$, and then showing that x^* also solves min(p, τ^*). One interesting result that one can reach by this route relates all four functions v, f, γ and h in one equation:

$$\forall\ i, j = 1, 2, \ldots, n \gamma_i(p, \tau^*) f_j \omega(p, \omega) = -v_i(p, \omega) h_{jT}(p, \tau^*) \tag{25}$$

Since from Shephard's Lemma the left-hand side of (25) is the Hicksian income effect of a change in p_i on the demand for good j, so is the right-hand side (RHS). Notice that although each of the components of the RHS is affected by choice of the utility index u, their product is not.

Revert now to the assumptions of section 1. The problems Max(p, ω) and Min(p, τ^*) are often referred to in the literature as dual to each other. For reasons given in detail in the entry on cost minimization and utility maximization, this usage seems inappropriate. However, as pointed out by Konyus and Byushgens (1926, p. 159) and Houthakker (1951–2, pp. 157–8), there *is* an interesting duality between the functions u and v. To show this, first rewrite the given prices and income (p, ω) as (p^*, ω^*), where $\omega^* > 0$ will be kept constant throughout. Next, define new income-normalized prices $q \in R^{n++}$ for any p by

$$q = (\omega^*) - 1 p$$

Then use the homogeneity of f and v in (p, ω) to put them in the normalized forms F: $R^{n++} \rightarrow R^{n+}$ and w: $R^{n++} \rightarrow R$, defined by

$$F(q^*) \equiv f(p^*, \omega^*) \tag{26}$$

and

$$w(q^*) \equiv v(p^*, \omega^*) \tag{27}$$

Let

$$A(q^*) = \left\{x \in R^{n+} : (q^*, x) \leq 1\right\} = A(p, \omega).$$

Then Max(p^*, ω^*) can also be written in a new form:

$$\text{Max}(q^*): \text{Find } x \in A(q^*) \text{ to max } u(x).$$

The data of Max(q^*) are q^* and u. In the same way, the chosen bundle x^* and w are the data for a problem dual to Max(q^*). Let $B(x^*)=\{q \in R^{n++}; \langle x^*,q\rangle \leq 1\}$. Then the dual problem, situated in the space of normalized prices q, is

$$\text{Min}(x^*): \text{Find } q \in B(x^*) \text{ to min } w(q).$$

A unique solution q^{**} to Min(x^*) (for which the strict quasi-convexity of w would suffice) implies the existence of two functions φ: $R^{n++} \rightarrow R^{n+}$ and U: $R^{n++} \rightarrow R$, defined analogously to (3) and to (2)-cum-(4) by

$$q^{**} = \varphi(x^*) \tag{28}$$

$$U(x^*) = w(q^{**}). \tag{29}$$

By the construction of Min(x^*), $x^* \in A(q)$ for any q. So $w(q)$ must be at least as large as the utility level at x^*. But since x^* is bought at q^*, that utility level is $w(q^*)$. Thus

$$\forall\ q \in B(x^*)\, w(q) \geq w(q^*) \tag{30}$$

Since min(x^*) is assumed to have a unique solution, (30) says that it must be q^*. If follows from this, (26) and (28) that ϕ is actually the *inverse* demand function F^{-1}. Moreover, $U(x^*)=w(q^*)$. So from this, (27), (4) and (2),

$$U(x^*) = u(x^*) \tag{31}$$

However, it cannot be concluded from (31) that $U \equiv u$ unless every bundle x in the domain of u is bought at some price-income pair (p, ω) and so can be an optimizing bundle such as x^*. This property requires that u be strictly quasiconcave. Granted that, (31) shows that the direct utility function u is recoverable from the indirect utility function w, just as w is obtainable from u.

See Also

- demand theory
- index numbers
- Roy, René François Joseph

Bibliography

Antonelli, G.B. 1886. *Sulla teoria matematica della economia politica*. Pisa: Tipografia del Folchetto.

Chipman, J.S., Hurwicz, L. Richter, M.K. and Sonnenschein, H.F., eds. 1971. *Preferences, Utility and Demand*. New York: Harcourt, Brace, Jovanovich.

Court, L.M. 1941. Entrepreneurial and consumer demand theories for commodity spectra. *Econometrica* 9, 135–62, 241–97.

Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of Mathematical Economics*, Vol. II, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland, ch. 12, 535–99.

Fenchel, W. 1953. Convex cones, sets and functions. Department of Mathematics, Princeton University, mimeo.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40, 577–616.

Houthakker, H.S. 1951–2. Compensated changes in quantities and qualities consumed. *Review of Economic Studies* 19, 155–64.

Houthakker, H.S. 1960. Additive preferences. *Econometrica* 28, 244–57.

Konyus, A.A. 1924. The problem of the true index of the cost of living. *Economic Bulletin of the Institute of Economic Conjuncture* 9–10, 64–71. Trans. in *Econometrica* 7, 1939, 10–29.

Konyus, A.A. and Byushgens, S.S. 1926. K probleme pokupatelnoi cili deneg. *Voprosi Konyunkluri* 2 (1), 151–72.

Malmquist, S. 1953. Index numbers and indifference surfaces. *Trabajos de estadistica* 4, 209–41.

Roy, R. 1942. *De l'utilité: contribution à la théorie des choix*. Paris: Hermann.

Roy, R. 1947. La distribution du revenu entre les divers biens. *Econometrica* 15, 205–25.

Samuelson, P.A. 1942. Constancy of the marginal utility of income. In *Studies in Mathematical Economics and Econometrics: In Memory of Henry Schultz*, ed. O. Lange, F. McIntyre and T.O. Yntema, Chicago: University of Chicago Press, 75–91.

Shephard, R.W. 1953. *Cost and Production functions*. Princeton: Princeton University Press.

Varian, H.R. 1984. *Microeconomic Analysis*. 2nd edn, New York: Norton.

Ville, J. 1946. Sur les conditions d'existence d'une ophelimité totale et d'un indice du niveau des prix. *Annales de l'Université de Lyon* 9, 32–9. Trans. in *Review of Economic Studies* 19, 1951, 123–8.

Wold, H.O.A. 1943–4. A synthesis of pure demand analysis. *Skandinavisk Aktuarietidskrift* 26, 85–144, 220–75; 27, 69–120.

How to cite this article

Newman, Peter. "indirect utility function." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000066> doi:10.1057/9780230226203.0779

The New Palgrave Dictionary of Economics Online

individual learning in games

Teck H. Ho
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews individual models of learning in games. We show that the experience-weighted attraction (EWA) learning nests different forms of reinforcement and belief learning, and that belief learning is mathematically equivalent to generalized reinforcement, where even unchosen strategies are reinforced. Many studies consisting of thousands of observations suggest that the EWA model predicts behaviour out-of-sample better than its special cases. We also describe a generalization of EWA learning to investigate anticipation by some players that others are learning. This generalized framework links equilibrium and learning models, and improves predictive performance when players are experienced and sophisticated.

Keywords

belief learning; curse of knowledge; equilibrium; experience-weighted attraction (EWA) learning; extensive-form games; fictitious play; forgone payoffs; individual learning in games; individual models of learning; maximum likelihood; mixed-strategy equilibrium; noise; overconfidence; population models of learning; quantal response equilibrium; reinforcement learning; signalling; social calibration; sophisticated players

Article

1 Introduction

Economic experiments on strategic games typically generate data that, in early rounds, violate standard equilibrium predictions. However, subjects normally change their behaviour over time in response to experience. The study of learning in games is about how this behavioural change works empirically. This empirical investigation also has a theoretical payoff: if subjects' behaviour converges to an equilibrium, the underlying learning model becomes a theory of equilibration. In games with multiple equilibria, this same model can also serve as a theory of equilibrium selection, a long-standing challenge for theorists.

There are two general approaches to studying learning: population models and individual models.

Population models make predictions about how the aggregate behaviour in a population will change as a result of aggregate experience. For example, in replicator dynamics, a population's propensity to play a certain strategy will depend on its 'fitness' (payoff) relative to the mixture of strategies played previously (Friedman, 1991; Weibull, 1995). Models like this submerge differences in individual learning paths.

Individual learning models allow each person to choose differently, depending on the experiences each person has. For example, in Cournot dynamics, subjects form a belief that other players will always repeat their most recent choice and best-respond accordingly. Since players are matched with different opponents, their best responses vary across the population. Aggregate behaviour in the population can be obtained by summing individual paths of learning.

This article reviews three major approaches to individual learning in games: experience-weighted attraction (EWA) learning, reinforcement learning, and belief learning (including Cournot and fictitious play). These models of learning strive to explain, for every choice in an experiment, how that choice arose from players' previous behaviour and experience. These models assume strategies have numerical evaluations, which are called 'attractions'. Learning rules are defined by how attractions are updated in response to experience.

Attractions are then mapped into predicted choice probabilities for strategies using some well-known statistical rule (such as logit).

The three major approaches to learning assume players that are adaptive (that is, they respond only to their own previous experience and ignore others' payoff information) and that their behaviour is not sensitive to the way in which players are matched. Empirical evidence suggests otherwise. There are subjects who can anticipate how others learn and choose actions to influence others' path of learning in order to benefit themselves. So we describe a generalization of these adaptive learning models to allow for this kind of sophisticated behaviour. This generalized model assumes that there is a mixture of adaptive learners and sophisticated players. An adaptive learner adjusts his behaviour according to one of the above learning rules. A sophisticated player does not learn and rationally best-responds to his forecast of others' learning behaviour. This model therefore allows 'one-stop shopping' for investigating the various statistical comparisons of learning and equilibrium models.

2 EWA learning

Denote player *i*'s *j*th strategy by s_i^j and the other player(s)' strategy by s_{-i}^k. The strategy actually chosen in period *t* is $s_i(t)$. Player *i*'s payoff for choosing s_i^j in period *t* is $\pi_i(s_i^j, s_{-i}^k(t))$. Each strategy has a numerical evaluation at time *t*, called an attraction $A_i^j(t)$. The model also has an experience weight, $N(t)$. The variables $N(t)$ and $A_i^j(t)$ begin with prior values and are updated each period. The rule for updating attraction sets $A_i^j(t)$ to be the sum of a depreciated, experience-weighted previous attraction $A_i^j(t-1)$ plus the (weighted) payoff from period *t*, normalized by the updated experience weight:

$$A_i^j(t) = \frac{\varphi \cdot N(t-1) \cdot A_i^j(t-1) + [\delta + (1-\delta) \cdot I(s_i^j, s_i(t))] \cdot \pi_i(s_i^j, s_{-i}(t))}{N(t)} \tag{2.1}$$

where indicator variable $I(x, y)$ is 1 if $x = y$ and 0 otherwise. The experience weight is updated by:

$$N(t) = \rho \cdot N(t-1) + 1. \tag{2.2}$$

Let $\kappa = \frac{\varphi - \rho}{\varphi}$. Then $\rho = \varphi \cdot (1-\kappa)$ and $N(t)$ approaches the steady-state value of $\frac{1}{1-\varphi\cdot(1-\kappa)}$. If $N(0)$ begins below this value, it steadily rises, capturing an increase in the weight placed on previous attractions and a (relative) decrease in the impact of recent observations, so that learning slows down.

Attractions are mapped into choice probabilities using a logit rule (other functional forms fit about equally well; Camerer and Ho, 1999):

$$P_i^j(t+1) = \frac{e^{\lambda \cdot A_i^j(t)}}{\Sigma_k e^{\lambda \cdot A_i^k(t)}}, \tag{2.3}$$

where λ is the payoff sensitivity parameter. The key parameters are δ, ϕ and κ (which are generally assumed to be in the [0,1] interval).

The most important parameter, δ, is the weight on forgone payoffs relative to realized payoffs. It can be interpreted as a kind of 'imagination' of forgone payoffs, or responsiveness

to forgone payoffs (when δ is larger players move more strongly toward *ex post* best responses). We call it ‘consideration’ of forgone payoffs. The weight on forgone payoff δ is also an intuitive way to formalize the ‘learning direction’ theory of Selten and Stoecker (1986). Their theory consists of an appealing property of learning: subjects move in the direction of *ex post* best-response. Broad applicability of the theory has been hindered by defining ‘direction’ only in terms of numerical properties of ordered strategies (for example, choosing ‘higher prices’ if the *ex post* best response is a higher price than the chosen price). The parameter δ defines the ‘direction’ of learning set-theoretically by shifting probability towards the set of strategies with higher payoffs than the chosen ones.

The parameter ϕ is naturally interpreted as depreciation of past attractions, $A_i^j(t-1)$. In a game-theoretic context, ϕ will be affected by the degree to which players realize other players are adapting, so that old observations on what others did become less and less useful. So we can interpret ϕ as an index of (perceived) ‘change’ in the environment.
The parameter κ determines the growth rate of attractions, which in turn affects how sharply players converge. When κ =0, the attractions are weighted averages of lagged attractions and payoff reinforcements (with weights $\varphi \cdot N(t-1) / (\varphi \cdot N(t-1) + 1$ and $1 / (\varphi \cdot N(t-1) + 1)$. When $\kappa = 1$ and $N(t) = 1$, the attractions are cumulations of previous reinforcements rather than averages (that is, $A_i^j(t) = \varphi \cdot A_i^j(t-1) + [\delta + (1-\delta) \cdot I(s_i^j, s_i(t))] \cdot \pi_i(s_i^j, s_{-i}(t))$). In the logit model, the *differences* in strategy attractions determine their choice probabilities. When κ is high the attractions can grow furthest apart over time, making choice probabilities closer to zero and one. We therefore interpret κ as an index of ‘commitment’.

3 Reinforcement learning

In cumulative reinforcement learning (Harley, 1981; Roth and Erev, 1995), strategies have levels of attraction which are incremented by only received payoffs. The initial reinforcement level of strategy j of player i, s_i^j, is $R_i^j(0)$. Reinforcements are updated as follows:

$$R_i^j(t) = \begin{cases} \varphi \cdot R_i^j(t-1) + \pi_i(s_i^j, s_{-i}(t)) & \text{if } s_i^j = s_i(t), \\ \varphi \cdot R_i^j(t-1) & \text{if } s_i^j \neq s_i(t). \end{cases}$$
(3.1)

Using the indicator function, the two equations can be reduced to one:

$$R_i^j(t) = \varphi \cdot R_i^j(t-1) + I(s_i^j, s_i(t)) \cdot \pi_i(s_i^j, s_{-i}(t)).$$
(3.2)

This updating formula is a special case of the EWA rule, when δ =0, $N(0)$=1, and κ =1.
In average reinforcement learning, updated attractions are *averages* of previous attractions and received payoffs (for example, Mookerjhee and Sopher, 1994; 1997; Erev and Roth, 1998). For example

$$R_i^j(t) = \varphi \cdot R_i^j(t-1) + (1-\varphi) \cdot I(s_i^j, s_i(t)) \cdot \pi_i(s_i^j, s_{-i}(t)).$$
(3.3)

A little algebra shows that this updating formula is also a special case of the EWA rule, when $\delta=0$, $N(0)=\frac{1}{1-\varphi}$, and $\kappa=0$. Since the two reinforcement models are special cases of EWA learning, their predictive adequacy can be tested empirically by setting the appropriate EWA parameters to their restricted values and seeing how much fit is compromised (adjusting, of course, for degrees of freedom).

4 Belief learning

In belief-based models, adaptive players base their responses on beliefs formed by observing their opponents' past plays. While there are many ways of forming beliefs, we consider a fairly general 'weighted fictitious play' model, which includes fictitious play (Brown, 1951; Fudenberg and Levine, 1998) and Cournot best-response (Cournot, 1960) as special cases. It corresponds to Bayesian learning if players have a Dirichlet prior belief.

In weighted fictitious play, players begin with prior beliefs about what the other players will do, which are expressed as ratios of strategy choice counts to the total experience. Denote total experience by $N(t)=\Sigma_k N^k_{-i}(t)$. Express the belief that others will play strategy *k* as $B^k_{-i}(t)=\frac{N^k_{-i}(t)}{N(t)}$, with $N^k_{-i}(t)\geq 0$ and $N(t)>0$.

Beliefs are updated by depreciating the previous counts by ϕ, and adding one for the strategy combination actually chosen by the other players. That is,

$$B^k_{-i}(t)=\frac{\varphi\cdot N^k_{-i}(t-1)+I(s^k_{-i},s_{-i}(t))}{\Sigma_h[\varphi\cdot N^h_{-i}(t-1)+I(s^h_{-i},s_{-i}(t))]}. \tag{4.1}$$

This form of belief updating weights the belief from one period ago ϕ times as much as the most recent observation, so ϕ can be interpreted as how quickly previous experience is discarded. When $\phi=0$ players weight only the most recent observation (Cournot dynamics); when $\phi=1$ all previous observations count equally (fictitious play).

Given these beliefs, we can compute expected payoffs in each period *t*,

$$E^j_i(t)=\sum_k B^k_{-i}(t)\,\pi(s^j_i,s^k_{-i}). \tag{4.2}$$

The crucial step is to express period *t* expected payoffs as a function of period *t–1* expected payoffs. This yields:

$$E^j_i(t)=\frac{\varphi\cdot N(t-1)\cdot E^j_i(t-1)+\pi(s^j_i,s_{-i}(t))}{\varphi\cdot N(t-1)+1}. \tag{4.3}$$

By expressing expected payoffs as a function of lagged expected payoffs, we make the belief terms disappear. This is because the beliefs are only used to compute expected payoffs, and when beliefs are formed according to weighted fictitious play, the expected payoffs which result can also be generated by generalized reinforcement according to previous

payoffs. More precisely, if the initial attractions in the EWA model are expected payoffs given some initial beliefs (that is, $A_i^j(0) = E_i^j(0)$), $\kappa = 0$ (or $\phi = \rho$), and foregone payoffs are weighted as strongly as received payoffs ($\delta = 1$), then EWA attractions are *exactly* the same as expected payoffs. Put differently, belief learning is 'mathematically equivalent' or 'observationally equivalent' to EWA learning with $\delta = 1$, $\kappa = 0$ and $A_i^j(0) = E_i^j(0)$.

This demonstrates a close kinship between reinforcement and belief approaches. Belief learning is nothing more than generalized attraction learning in which strategies are reinforced equally strongly by actual payoffs and foregone payoffs and attractions are weighted averages of past attractions and reinforcements. Hopkins (2002) compares the convergence properties of reinforcement and fictitious play and finds that they are quite similar in nature and that they will in many cases have the same asymptotic behaviour.

5 A graphical representation

Since reinforcement and belief learning are special cases of EWA learning, it is possible to represent all three learning models in a three-dimensional EWA cube (see Figure 1). The vertex $\delta = 1$ and $\kappa = 0$ corresponds, to weighted fictitious play models. The corners $\phi = 0$ and $\phi = 1$ correspond to Cournot best-response dynamics and fictitious play, respectively. Reinforcement models in which only chosen strategies are reinforced according to their payoffs correspond to vertices in which $\phi = 0$, and $\kappa = 1$ (cumulative reinforcement) or $\kappa = 0$ (averaged reinforcement). Interior configurations of parameter values incorporate both the intuition behind reinforcement learning, that realized payoffs weigh most heavily ($\delta < 1$), and the intuition implicit in belief learning, that foregone payoffs matter too ($\delta > 0$).

Figure 1

EWA's model parametric space

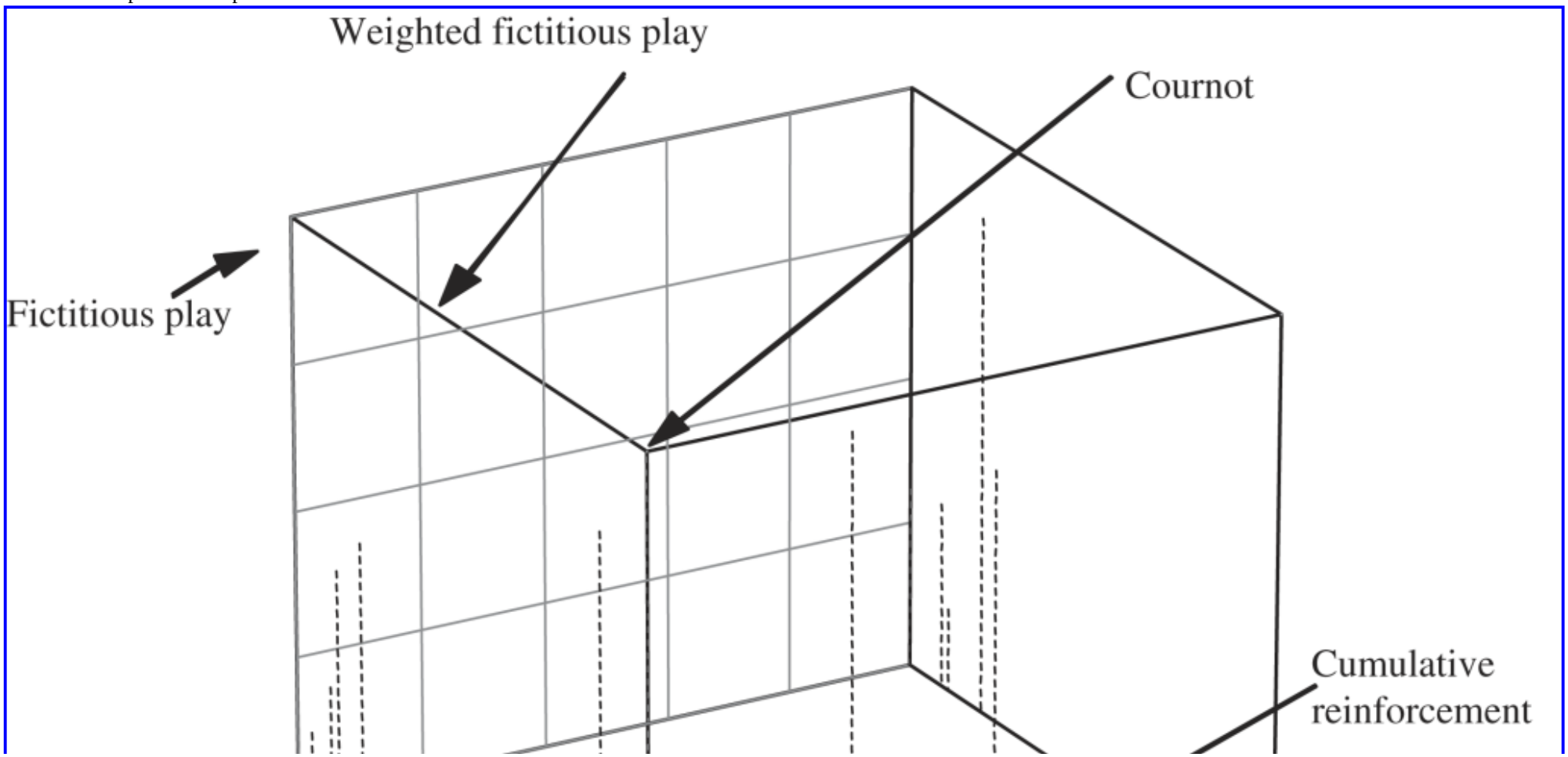

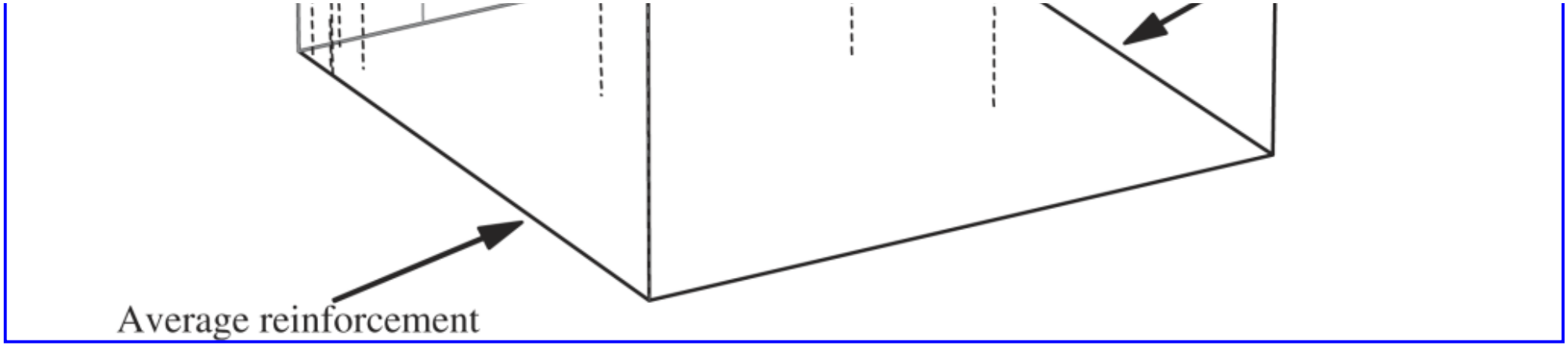

The cube shows that contrary to popular belief for many decades, reinforcement and belief learning are simply two extreme configurations on opposite edges of a three-dimensional cube, rather than fundamentally unrelated models. Figure 1 also shows estimates of the three parameters in 20 different studies (Camerer, Ho and Chong, 2002). Each point is a triple of estimates. These parameter estimates were typically obtained by the maximum likelihood method. Initial attractions could be either estimated using data or set to plausible values using the cognitive hierarchy model of one-shot games; see Camerer, Ho and Chong (2004) for details. Most points are sprinkled throughout the cube, rather than at the extreme vertices mentioned in the previous paragraph, although some (generally from games with mixed-strategy equilibria) are near the averaged reinforcement corner $\delta=0$ and $\kappa=\phi=1$. Ho, Camerer and Chong (2007) provide an explanation for how δ and ϕ vary across games by endogenizing them as functions of game experience. Parameter estimates are generally significantly inside the interior of the cube rather than near the vertices. Thus, we may conclude that subjects' behaviour is often neither belief nor reinforcement learning.

6 Linking learning and equilibrium models

The adaptive learning models presented above do not permit players to anticipate learning by others. Omitting anticipation logically implies that players do not use information about the payoffs of other players, and that whether players are matched together repeatedly or are randomly re-matched should not matter. Both of the latter implications are unintuitive, and experiments with experienced subjects have provided evidence to show otherwise.

In Camerer, Ho and Chong (2002) and Chong, Camerer and Ho (2006), we proposed a simple way to include 'sophisticated' anticipation by some players that others are learning, using two additional parameters. We assume a fraction α of players are sophisticated. Sophisticated players think that a fraction $(1-\alpha')$ of players are adaptive and the remaining fraction α' of players are sophisticated like themselves. They use the EWA model (which nests reinforcement and belief learning as special cases) to forecast what the adaptive players will do, and choose strategies with high expected payoffs given their forecast.

All the adaptive models discussed above (EWA, reinforcement, belief learning) are special cases of this generalized model with $\alpha=0$. The assumption that sophisticated players think some others are sophisticated creates a small whirlpool of recursive thinking which implies that quantal response equilibrium (QRE; McKelvey and Palfrey, 1995) and Nash equilibrium are special cases of this generalized model. Our specification also shows that equilibrium concepts combine two features which are empirically and psychologically separable: 'social calibration' (accurate guesses about the fraction of players who are sophisticated, $\alpha=\alpha'$); and full sophistication ($\alpha=1$). Psychologists have identified systematic departures from social calibration called 'false uniqueness' or overconfidence ($\alpha>\alpha'$) and 'false consensus' or curse of knowledge ($\alpha>\alpha'$).

Formally, adaptive learners follow the EWA updating equations given above (that is, (2.1) and (2.2)). Sophisticated players have attractions $B_i^j(t)$ and choice probabilities $Q_i^j(t+1)$ specified as follows:

$$B_i^j(t)=\sum_k[(1-\alpha')\cdot P_{-i}^k(t+1)+\alpha' Q_{-i}^k(t+1)]\cdot\pi_i(s_i^j,s_{-i}^k), \tag{6.1}$$

$$Q_i^j(t+1) = \frac{e^{\lambda \cdot B_i^j(t)}}{\sum_k e^{\lambda \cdot B_i^k(t)}}. \tag{6.2}$$

The generalized model has been applied to experimental data from ten-period *p*-beauty contest games (specific details of data collection are given in Ho, Camerer and Weigelt, 1998). In these games, seven subjects choose numbers in [0,100] simultaneously. The subject whose number is closest to *p* times the average (where *p*=.7 or .9) wins a fixed prize. Subjects playing for the first time are called ‘inexperienced’; those playing another ten-period game (with a different *p*) are called ‘experienced’.

The estimation results show that for inexperienced subjects, adding sophistication to adaptive EWA improves log likelihood (LL) substantially both in- and out-of-sample. The estimated fraction of sophisticated players is $\hat{\alpha} = .24$ and their estimated perception $\hat{\alpha}' = 0$. Experienced subjects show a much larger improved fit from sophistication, and a larger estimated proportion, $\hat{\alpha} = .75$. Their perceptions are again too low, $\hat{\alpha}' = .41$, showing a degree of overconfidence. The increase in sophistication due to experience reflects a kind of ‘learning about learning’, which is similar to rule learning (that is, subjects switch their learning rule over time (Stahl, 2000; Ho, Camerer and Chong, 2007). Overall, these results suggest that subjects are not socially calibrated, that not all subjects are sophisticated, and that the proportion of sophistication grows with experience.

7 Conclusions and future research

We describe three major approaches of adaptive learning models. We show that EWA learning is a generalization of reinforcement and belief learning and that the latter two nested models are intimately related. Specifically, they differ mainly in the way they treat forgone payoffs; reinforcement learning ignores them and belief learning treats them the same as actual payoffs. Estimation results from dozens of studies show that the emergence of behaviour is neither reinforcement nor belief learning in most games. The EWA cube provides a simple way for detecting how these simpler models fail and why.

We also describe a generalization of these adaptive models to study anticipation by some players that others are learning. This generalized model nests equilibrium and the adaptive learning models as special cases and is a powerful framework for analysing both equilibrium and learning simultaneously. We show that it can improve the predictive performance of the adaptive learning models when players are experienced and able to anticipate how others learn.

There are three promising areas of future research, all of which aim to make the above learning models more amenable to field applications.

1. *Transfer of learning across similar games*. In practice, it is unreasonable to expect people play the identical game again and again. Since people are more likely to face with similar but non-identical strategic situations, it is important to determine whether they are able to transfer learning from one situation to another. Cooper and Kagel (2004) provide evidence that subjects who have learned to play strategically in one signalling game can transfer most of this knowledge to related games. This transfer of learning occurs because the proportion of sophisticated players grows with experience (just like what we observed in *p*-beauty contest games discussed above). This positive evidence is encouraging but more work is necessary to determine whether this finding indeed generalizes to other games.

2. *Learning in extensive-form games*. Most of the learning literature focuses on strategic or normal-form games (for an exception see Anderson and Camerer, 2000). This is done in part to simplify the learning context to situations where each action unambiguously corresponds to a final outcome. In extensive-form games or many field settings, where a final outcome is typically a result of a series of actions taken sequentially over time, there is a natural question how an action step taken at a particular time contributes to the final outcome. This ‘credit assignment’ problem is important because different agents might be responsible for different action steps, and some steps might be more crucial than others at determining the final outcome. A good learning model should assign credit appropriately to each action step.

3. *Learning in noisy experiments*. There is a general belief that, given a sufficiently high stake and that people play repeatedly with a clear feedback, their behaviour will converge to equilibrium in the long run. However many real-world environments provide noisy feedback. So it is important to study how noise in feedback affects rates of learning and the likelihood of convergence to equilibrium.

See Also

- experimental economics
- learning and evolution in games: belief learning
- maximum likelihood.

Bibliography

Anderson, C. and Camerer, C. 2000. Experience-weighted attraction learning in sender-receiver signaling games. *Economic Theory* 16, 689–718.

Brown, G. 1951. Iterative solution of games by fictitious play. In *Activity Analysis of Production and Allocation* New York: Wiley.

Camerer, C.F. and Ho, T.-H. 1999. Experience-weighted attraction learning in normal-form games. *Econometrica* 67, 827–74.

Camerer, C., Ho, T.-H. and Chong, J.-K. 2002. Sophisticated learning and strategic teaching. *Journal of Economic Theory* 104, 137–18.

Camerer, C.F., Ho, T.-H. and Chong, J.-K. 2004. A cognitive hierarchy model of one-shot games. *Quarterly Journal of Economics* 119, 861–98.

Chong, J.-K., Camerer, C. and Ho, T.-H. 2006. A learning-based model of repeated games with incomplete information. *Games and Economic Behavior* 55, 340–71.

Cooper, D. and Kagel, J. 2004. Learning and transfer in signaling games. Working paper, Ohio State University.

Cournot, A. 1960. *Recherches sur les principes mathématiques de la théorie des richesses*. Trans. N. Bacon as *Researches in the Mathematical Principles of the Theory of Wealth*. London: Haffner.

Erev, I. and Roth, A. 1998. Modelling predicting how people play games: reinforcement learning in experimental games with unique, mixed-strategy equilibria. *American Economic Review* 88, 848–81.

Friedman, D. 1991. Evolutionary games in economics. *Econometrica* 59, 637–66.

Fudenberg, D. and Levine, D. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Harley, C. 1981. Learning the evolutionary stable strategies. *Journal of Theoretical Biology* 89, 611–33.

Ho, T.-H., Camerer, C. and Weigelt, K. 1998. Iterated dominance and iterated best-response in *p*-beauty contests. *American Economic Review* 88, 947–69.

Ho, T.-H., Camerer, C. and Chong, J.-K. 2007. Self-tuning experience-weighted attraction learning in games. *Journal of Economic Theory* 133, 177–98.

Hopkins, E. 2002. Two competing models of how people learn in games. *Econometrica* 70, 2141–66.

McKelvey, R. and Palfrey, T. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10, 6–38.

Mookerjhee, D. and Sopher, B. 1994. Learning behavior in an experimental matching pennies game. *Games and Economic Behavior* 7, 62–91.

Mookerjee, D. and Sopher, B. 1997. Learning and decision costs in experimental constant-sum games. *Games and Economic Behavior* 19, 97–132.

Roth, A.E. and Erev, I. 1995. Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* 8, 164–212.

Selten, R. and Stoecker, R. 1986. End behavior in sequences of finite Prisoner's Dilemma supergames: a learning theory approach. *Journal of Economic Behavior and Organization* 7, 47–70.

Stahl, D. 2000. Rule learning in symmetric normal-form games. *Games and Economic Behavior: Theory and Evidence* 32, 105–38.

Weibull, J. 1995. *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

How to cite this article

Ho, Teck H. "individual learning in games." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000055> doi:10.1057/9780230226203.0780

The New Palgrave Dictionary of Economics Online

individual retirement accounts

Jonathan Skinner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Individual Retirement Accounts in the United States are tax-preferred saving vehicles designed to encourage saving for retirement. Many countries have adopted similar saving mechanisms such as Individual Saving Accounts in Britain, Special Saving Incentive Accounts in Ireland, and Tax-Preferred Deposit Accounts in Belgium. Enrolment rates are substantially higher among high income taxpayers, while the saving effects are often found to be quite modest. However, given the often low revenue cost of these tax preferred accounts, they may be reasonably cost-effective in terms of new saving per lost unit of revenue.

Keywords

capital gains; Individual Retirement Accounts; national saving; precautionary saving; retirement; pensions

Article

The Individual Retirement Account (IRA) in the United States was first introduced in 1974, but languished in relative obscurity until the Economic Recovery Tax Act of 1981 expanded eligibility to all US taxpayers. Contributions jumped from $4.8 billion in 1981 to $28.3 billion in 1982, before peaking at $37.8 billion in 1986 (Holden et al., 2005). The traditional IRA provided a tax break when contributions were made to qualified accounts, but taxed the entire withdrawal (principle plus interest) upon withdrawal. Restrictions included a ten per cent penalty for withdrawing money before age $59\frac{1}{2}$, and the requirement that the taxpayer implement a systematic withdrawal plan by age $70\frac{1}{2}$. In 2007, the limit for tax-deductible contributions was $4,000, or $5,000 for taxpayers over age 50.
The IRA came under fire during the mid-1980s because of revenue costs and concerns that it was being used as a tax shelter for high-income taxpayers. The Tax Reform Act of 1986 tax instituted income

limits and, as a result, contributions dropped off rapidly, from 37.8 billion in 1986 to 14.1 billion in 1987 (Holden et al., 2005). That contributions even fell by 30 per cent among those still eligible to contribute suggests that confusion about eligibility (Hrung, 2001), or a decline in advertising, may have affected taxpayer participation adversely. While the introduction of the eponymous Roth IRA, under which the taxpayer contributed after-tax dollars which were allowed to accumulate (and be withdrawn) tax-free, was popular among contributors, IRAs remain a relatively unimportant source of new saving, accounting for less than 0.2 per cent of GDP.

Nonetheless, the stock of IRA assets had grown to $3 trillion by 2007, when it comprised 20 per cent of US retirement saving (Holden et al., 2005). The reason IRAs comprise such a large fraction of wealth is that workers changing jobs or retiring are allowed to ‘roll over’ defined contribution (401(k)) balances into IRAs without any tax penalty. Thus IRA growth has been fuelled by these rollovers, which in 2000–1 comprised more than $200 billion annually, or about ten times the contributions by savers (Holden et al., 2005).

A number of IRA-like saving vehicles have been introduced in other countries, often with more generous eligibility and contribution rules. For example, the current (2007) contribution limit for Registered Retirement Saving Plans (RRSPs) in Canada is C$19,000. As well, taxpayers may carry forward past unused contributions, so the effective limit is generally much larger. In the United Kingdom, Individual Saving Accounts (ISAs) were introduced in 1999, replacing Personal Equity Plans (PEPs) and Tax-Exempt Special Saving Accounts (TESSAs) (Attanasio, Banks and Wakefield, 2004). The contribution limit for the ISA in 2007 was £7,000 and resembled a Roth IRA in that contributions were made after taxes were paid but withdrawals and accumulated build-up were tax-free. Many other developed countries offer similar tax incentives, such as tax-preferred saving accounts for children and grandchildren in Denmark, Special Saving Incentive Accounts in Ireland, and Tax-Preferred Deposit Accounts in Belgium (see Maffini, 2007). Other tax-preferred saving schemes, most notably employer-based defined-contribution pension plans such as 401(k)s in the United States, are discussed elsewhere (see pensions).

Economic incentives

As noted above, there are two basic flavours of IRAs, traditional IRAs with an ‘up-front’ deduction and Roth-style IRAs, whereby taxpayers invest tax-free dollars and withdraw the accumulated amount tax-free. The economic effects of IRAs are simplest to see in the case of a standard bond that pays a constant rate of return r^* for n years until retirement when the entire IRA is withdrawn. The marginal tax rate on an extra dollar of interest income is τ_m, at which point the tax rate shifts to τ_o while retired. If the investor invests one dollar in the conventional bond, her after-tax return at retirement will be $(1+r^*(1-\tau_m))^n$, while an investment in a classic IRA will yield

$$\frac{(1-\tau_o)(1+r^*)^n}{1-\tau_m}$$

and a Roth IRA will return $(1+r^*)^n$. It is straightforward to demonstrate that both the classic and the Roth IRA strictly dominate the conventional bond investment, and that the classic IRA dominates the Roth IRA under the assumption that $\tau_o < \tau_m$. This may not always be a sound assumption, particularly if retirees are too worried about higher future tax rates to pay for financially strained social insurance programmes. Gokhale, Kotlikoff and Neumann (2001) have observed that in some cases taxable income while retired may be subject to a higher marginal tax rate because of peculiarities in the US tax code, diminishing the advantage of traditional IRAs relative to Roth IRAs.

The decision becomes more complex when considering whether to hold equity investments inside an IRA. When a substantial fraction of the asset appreciation occurs through capital gains, one must trade off the tax advantages of the conventional IRA with the necessity to withdraw the (appreciated) assets from the IRA account starting in age $70\frac{1}{2}$ (Note that the Roth IRA does not require a withdrawal plan). As well, keeping the stock outside of an IRA retains its availability for precautionary purposes, and makes it eligible for preferential treatment of capital gains and dividends, and the possibility of stepping up the tax basis at death.

There are two key reasons why countries may decide to create IRA-style accounts. The first is to stimulate national saving, while the second is to improve the financial security of retirees, particularly those without access to employer-based pensions. The two are not necessarily overlapping. A programme that successfully stimulates saving among millionaires and billionaires may have a large impact on aggregate national saving, but do little or nothing to enhance the financial security of these households already well-prepared for the risks of retirement. Similarly, a programme that encourages low- or lower-middle-class savers by supplementing financial resources by (say) $10,000 would have a small impact on national saving but could exert a much larger proportional impact on available financial resources. That IRA inflows (excluding rollovers) in the United States comprise less than 0.2 per cent of GDP suggests a small upper limit for its impact on aggregate saving. (The size of these plans in other countries, relative to GDP, also appears modest; see Maffini, 2007.)

Whether as a mechanism to increase aggregate saving or to encourage retirement security for specific households, the impact of IRAs on net saving is theoretically ambiguous. If IRA wealth and non-IRA taxable wealth were perfect substitutes, clever taxpayers could simply shuffle money from their taxable wealth accounts into IRAs, and enjoy the future or current tax rebate. If the tax incentive is further financed through deficit spending, and taxpayers spent part of the tax break, net national saving could *decline* following the introduction of an IRA programme. How individual accounts affect individual and national saving is therefore an empirical question.

Empirical evidence

There has been considerable debate regarding the impact of IRAs on net saving. The first set of studies was by Venti and Wise (1986; 1990) who estimated that IRA and non-IRA savings were imperfect substitutes, thus suggesting that IRAs led to roughly 60 cents of new saving per dollar of IRA contributions, with most of the remaining 40 cents representing the tax subsidy. Similarly, Engelhardt (1996) found large saving effects by comparing saving rates in Canada before and after the cessation of

the tax-subsidized Registered Home Ownership Savings Plan.

Gale and Scholz (1994) specified a more general model allowing for differences in tastes for saving between IRA and non-IRA contributions, and using the nonlinearity of the budget constraint – due to the IRA limits – to help identify the true saving effects of IRAs. They arrived at a quite different conclusion, namely, that IRAs in fact reduced net saving because contributors ‘shuffled’ savings from taxable accounts. Ultimately, their estimates were found to be very sensitive to the inclusion or exclusion of a few observations (Poterba, Venti and Wise, 1996), underscoring the difficultly of testing for causality using observational data. Even in a dynamic setting (as in Feenberg and Skinner, 1989 one cannot rule out the possibility that former spendthrifts who suddenly start pouring money into IRAs would have done so even without IRAs available for their use.

Attanasio, Banks and Wakefield (2004) used the natural experiment of the 1999 shift in the United Kingdom from PEPs and TESSAs to the less restrictive (and hence more popular) ISAs to test the resulting impact on saving. The resulting (albeit very noisy) patterns of changes in saving rates were not supportive of a positive impact on national saving. Another study used the difference between contribution rates of taxpayers making their first year's contribution to an IRA and those of later contributors (Attanasio and De Leire, 2002; see also Joines and Manegold, 1995). They found that new contributors exhibit shuffling behaviour from existing assets into IRAs. Less clear is whether later contributors (the majority of IRA inflows) were increasing net national saving (Hubbard and Skinner, 1996; Attanasio, Banks and Wakefield, 2004).

The strongest evidence of how IRAs affect saving comes from a randomized trial conducted in the St. Louis metropolitan area by H&R Block, a large tax preparation firm (Duflo et al., 2006). In this study, tax filers at H&R Block were provided with different incentives to open an ‘express’ IRA funded with either tax refunds or other sources. Duflo et al. found enrolment rates of 3 per cent for the control group, 10 per cent for the treatment group with a 20 per cent match and 17 per cent for those with the 50 per cent match. Conditional on enrolment, contributions (excluding the match) amounted to $860, $1,280, and $1,310, respectively. The researchers were not able to measure offsetting effects for non-IRA wealth, but large and significant effects in the treatment group were observed even in households with low median income or without saving accounts, thus minimizing the potential for ‘shuffling’ from other assets within these groups.

However, these results cannot be generalized to the saving effects of conventional IRAs in the United States or in other countries. As the authors noted, the treatment effect depended strongly on the specific tax professional; some tax professionals just couldn't ‘sell’ the IRAs no matter how attractive the match. Furthermore, the IRA was offered at an auspicious time when the refund had not yet been issued. The IRA match may have been a necessary, but apparently it was not a sufficient, condition to persuade all contributors (even those in high income brackets) to sign up.

Conclusions

It is unfortunate that we still know so little about the saving effects of IRAs and similar saving incentives. While we don't know the incremental effect of IRAs, we might expect the strongest saving effects to arise among lower-income households where the opportunity to shuffle assets is most constrained (for example, Engen and Gale, 2000; Benjamin, 2005 in the case of 401(ks)). And the

evidence we do observe – the sharp drops in contributions among those still eligible following the 1986 cutbacks, and the importance of individual tax professional effects, for example – suggests that behavioural or marketing factors are critically important in ‘selling’ IRAs to the households where the tax advantages are perhaps no so apparent and the distributional effects of the subsidy are not so inequitable (see Bernheim, 1997).

What about the saving effects of IRAs among higher-income households? Because IRA accounts typically require the enrolee to write a check, one may never expect it to exhibit the same saving effects of a 401(k) plan that automatically withdraws money before the paycheck is cashed. But, as Hubbard and Skinner (1996) argue, the saving effects need not be large in order to justify the government provision of saving accounts. Recall that the net revenue loss to the government for the traditional IRA is the up-front deduction, less the present value of the discounted future tax payments. This difference may be quite modest when strong stock market gains build up equity inside traditional IRAs, leading to higher future revenue collections as the IRAs are gradually drawn down (Dusseault and Skinner, 2000; also see Gravelle, 2000).

In sum, IRAs provide tax-preferred wealth accumulation to those without employer pensions or who seek to accumulate something extra for retirement. Saving effects are likely to be largest when IRAs are designed to appeal to low- or middle-income households where opportunities for shuffling are minimized. Finally, governments may find policy changes irresistible as they realize how much future tax revenue lies within traditional IRA assets, or how much potential tax liability lies within rapidly growing Roth or ISA assets.

See Also

- capital gains taxation
- pensions
- retirement
- taxation of income
- taxation of wealth

Bibliography

Attanasio, O., Banks, J. and Wakefield, M. 2004. Effectiveness of tax incentives to boost (retirement) saving: theoretical motivation and empirical evidence. *OECD Economic Studies* 39(2), 145–72.

Attanasio, O. and De Leire, T. 2002. IRAs and household savings revisited: some new evidence. *Economic Journal* 112, 504–38.

Benjamin, D. 2005. Does 401(k) eligibility increase saving? Evidence from propensity score subclassification. *Journal of Public Economics* 87, 1259–90.

Bernheim, B.D. 1997. Rethinking savings incentives. In *Fiscal Policy: Lessons from Economic*

Research, ed. A. Auerbach. Cambridge: MIT Press.

Duflo, E., Gale, W.G., Liebman, J.B., Orszag, P.R. and Saez, E. 2006. Saving incentives for low- and middle-income families: evidence from a field experiment with H&R block. *Quarterly Journal of Economics* 121, 1311–46.

Dusseault, B. and Skinner, J. 2000. Did individual retirement accounts actually raise revenue? *Tax Notes* 86, 851–6.

Engelhardt, G. 1996. Tax subsidies and household saving: evidence from Canada. *Quarterly Journal of Economics* 111, 1237–68.

Engen, E.M. and Gale, W.G. 2000. The effects of 401(k) plans on households wealth: differences across earnings groups. Working Paper No. 8032. Cambridge, MA: NBER.

Engen, E.M., Gale, W.G. and Scholz, J.K. 1996. The illusory effects of saving incentives on saving. *Journal of Economic Perspectives* 10(4), 113–38.

Feenberg, D. and Skinner, J. 1989. Sources of IRA saving. In *Tax Policy and the Economy 3*, ed. L.H. Summer. Cambridge, MA: MIT Press.

Gale, W.G. and Scholz, J.K. 1994. IRAs and household saving. *American Economic Review* 84, 1233–60.

Gokhale, J., Kotlikoff, L.J. and Neumann, T. 2001. Does participating in a 401(k) raise your lifetime taxes? Working Paper No. 8341. Cambridge, MA: NBER.

Gravelle, J. 2000. IRAs as revenue-raisers: a critique. *Tax Notes* 86(9).

Holden, S., Ireland, K., Leonard-Chambers, V. and Bogdan, M. 2005. The individual retirement account at age 30: a retrospective. *Investment Company Institute Perspective* 11(1), 1–23.

Hrung, W. 2001. Information and IRAs: the influence of tax preparers. *Journal of Public Economics* 80, 467–84.

Hubbard, G.R. and Skinner, J.S. 1996. Assessing the effectiveness of saving incentives. *Journal of Economic Perspectives* 10(4), 73–90.

Joines, D.H. and Manegold, J.G. 1995. IRA and saving: evidence from a panel of taxpayers. Mimeo, University of Southern California.

Maffini, G. 2007. *Encouraging Saving Through Tax-Preferred Accounts*. Tax Policy Studies No. 15. Paris: OECD.

Poterba, J., Venti, S. and Wise, D. 1996. How retirement saving programs increase saving. *Journal of Economic Perspectives* 10(4), 91–112.

Venti, S. and Wise, D. 1986. Tax-deferred accounts, constrained choice, and estimation of individual saving. *Review of Economic Studies* 53, 579–601.

Venti, S. and Wise, D. 1990. Have IRAs increased US saving? Evidence from consumer expenditure surveys. *Quarterly Journal of Economics* 105, 661–89.

How to cite this article

Skinner, Jonathan. "individual retirement accounts." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000067> doi:10.1057/9780230226203.0781

The New Palgrave Dictionary of Economics Online

individualism versus holism

Harold Kincaid

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Issues about individualism and holism in economics surface because economics is committed to understanding both institutions and large-scale economic processes, in terms of constrained maximizing of individuals. Three key questions are at issue. Can a theory of individual economic behaviour capture everything we want to explain about the economy in principle? To what extent do our accounts of individual economic behaviour trump or constrain other economic explanations that are not directly about individuals? Are non-individual economic entities real, and what is their relation to individual behaviour? These questions are answered in light of developments in economics and in philosophy of science.

Keywords

causation; evolutionary game theory; explanation; game theory; holism; Hoover, K.; methodological individualism; new institutional economics; rational expectations; reductionism; representative agents; Schumpeter, J. A.; Smith, A.

Article

The idea that economic outcomes result from and thus are to be explained by the maximizing choices of individual human beings has been essential to economics at least since Adam Smith. Yet this maxim of methodological individualism has consistently existed alongside and in tension with the important role that institutions play in economic outcomes and the desire of economists to explain large-scale phenomena such as the rate of inflation and unemployment. Debates over individualism and holism have generally been vaguely formulated and argued at an abstract level with a questionable relationship to the actual practice of economics. The purpose of this article is to clarify the theses and arguments at work and replace rhetoric with identifiable empirical issues with real ties to economic practice. Some recent developments in economics – for example, the new economics of information asymmetry and

institutions, and the obstacles to the refinement programme in classical rational choice game theory and to rational expectations programmes in macroeconomics – argue against the most extreme forms of individualism while leaving a plausible place for various more modest individualist constraints.
There are at least three questions at the centre of economic debates over holism and individualism:

1. 1. Can a theory of individual economic behaviour capture everything we want to explain about the economy in principle?
2. 2. To what extent do our accounts of individual economic behaviour trump or constrain other economic explanations that are not directly about individuals?
3. 3. Are non-individual economic entities real, and what is their relation to individual behaviour?

The first question can be thought of as a question about theory reduction. Can a well-formulated theory of individual behaviour replace all economic explanations that are not directly about individuals, at least those we think are relatively well confirmed? The second question is usually put as a thesis about mechanism: every economic explanation has to be given individualist mechanisms. The final question is about ontology: what entities populate the economic realm and how are they related?
Individualists tend to answer the first two questions affirmatively and either deny that social entities exist at all or assert that individuals are in some sense prior to them. Extreme holists take the opposite stance, answering negatively to the first two questions and arguing that social entities are real and in some sense prior to individuals.
These three questions are no doubt related, and it is often asserted that an answer to one of these questions tells us the answer to the others. Yet, for the most part, discussions in the literature do not clearly identify which theses are at issue nor exactly what their relationship is to each other.
To what extent is economics committed to a version of individualism? Schumpeter (1954), in his classic history of economic analysis, apparently coined the term ‘methodological individualism’ and argued that it, along with a fundamental focus on prices and general equilibrium analysis, was the common core of economics since Smith. There is little evidence that Schumpeter was right about the classical economists. They were interested in the distribution of the total economic product to social classes and the factors influencing its growth. Their accounts frequently involved taking institutional structure as given rather than explained; Smith explicitly acknowledged that invisible hand processes work against a background of social institutions, customs, and the like (Gordon, 1991).
However, the neoclassical revolution certainly ushered in an explicit commitment to individualism. Many past and recent elements of modern economics are directly motivated by the individualist theses mentioned above, among them:

1. (*a*) the general equilibrium programme of explaining all economic phenomena on the basis of individual preferences and initial endowments;
2. (*b*) the rational choice game-theory programme of explaining norms, institutions, the behaviour of the firm, and so on, completely in terms of the behaviour of maximizing individuals; and
3. (*c*) the rational-expectations programme which seeks to model macroeconomic phenomena in terms of the expectations states of individual maximizing agents with given preferences, technology, and so on.

These doctrines make the three kinds of claims listed above: (*a*) the ontological claim that all economic phenomena consist of the actions of individuals, (*b*) the reductive claim that a theory based on such primitives can explain and thus reduce all economic phenomena to individual behaviour, and (*c*) the claim that individual-based mechanisms are a requirement of good explanation. Thus the individualism–holism controversy in economics crucially involves, at a minimum, evaluating the extent to which these programmes succeed in realizing their individualism, putting aside other worries such as the assumption of equilibrium outcomes and so on.

To start with the ontological issues first, there is no good evidence that I know of that any major economist from the classicals on affirmed the extreme holist ontological thesis that society acts or exists entirely independently of the behaviour of individuals. Moreover, the question whether aggregate economic entities are real seems to me to be the least interesting and controversial issue in the individualism–holism debate in economics. No one denies that firms, for example, are collections of individuals. If each of the individuals in the firm exists, then the sum of them exists as well. The real issue I would argue is how far we can go in *explaining* the aggregate in terms of the individuals composing it – that is a question of reduction.

Hoover (2001) has argued for the reality of macroeconomic aggregates on the grounds that something is real if it stands in causal relations and that macroeconomic aggregates do stand in such relations. However, the advocate of rational expectations is not denying that the GDP or the rate of inflation exists but instead is asserting that their causal efficacy can be explained in terms of the actions of individuals and that, given rational expectations, we cannot expect there to be stable causal relations among aggregates invariant to policy changes. Again the issues seem to be more about explanation, not whether aggregates are real.

The second ontological claim – that the characteristics of economic aggregates are dependent upon the facts about individuals – has more content. Borrowing from philosophical discussions of physicalism (Hellman and Thompson, 1975), reductionism in general and in the philosophy of mind in particular (Fodor, 1974), we can take this claim to be asserting that all non-individual facts supervene on or are determined by the facts about individuals. The basic idea is that one set of facts A supervenes on or is determined by another set B just in case once the B facts are set, so are the A facts. In other words, there is no difference in the A facts without a corresponding difference in the B facts. As we will see below, this asserts only a one-way conditional from the Bs to the As and not the stronger biconditional of A if and only if B that is typical of reduction. (So the individualist thesis would be that the economic facts about individuals fix the facts about other aggregate or collective economic entities.)

Talk of ‘facts’ is vague. We can be more precise by asking if the truths or assertions of a particular theory fix those of another – in this case, whether a particular economic theory referring only to individuals determines or fixes the truths of an economic theory that includes terms referring to collective economic phenomena. Put this way, the debate over this individualist thesis is really many different debates, depending on what particular models are at issue, and individualism might be plausible in some cases and not others. So, for example, we can ask whether downward-sloping demand curves of consumer choice theory ensure that aggregate market demand curves are likewise downward-sloping. The evidence seems to be that they are only under very restrictive conditions that are unlikely to hold (Deaton and Muellbauer, 1980). More generally, the Sonnenschein–Mantel–Debreu theorem shows that individual excess demand functions do not ensure a unique equilibrium. Similar difficulties face

some attempts to derive macroeconomic implications from choice theoretic models of individual behaviour (Martel, 1996). Yet it seems clear that there must be *some* model of individual behaviour that determines such aggregate relations.

Individualism acquires it clearest statement as a claim about theory reduction. There is an extensive literature on the requirements for theory reduction in general as exemplified by the reduction of the gas laws to statistical mechanics. The gas laws refer essentially to temperature while no such notion is a fundamental category in statistical mechanics. However, temperature has an analogue in the mean kinetic energy of the molecules in a gas – we can define the former in terms of the latter. A definition requires at a minimum some kind of biconditional relationship between the terms involved. When we can produce such a definition, then to reduce we should be able to reproduce the explanation given by one theory in the vocabulary of another, for example, by showing that the gas laws follow from laws of statistical mechanics once temperature is equated with mean kinetic energy.

There are at least three possible ways that one theory might turn out to be irreducible (Kincaid, 1996; 1997).

Multiple realizations: if we wanted to reduce ordinary claims about chairs, for example, to particle physics, then we would need to find a one-to-one correspondence between chair categories and quantum mechanical descriptions, because there are indefinitely many ways to bring chairs about in physical terms and no natural way to capture them in terms of physics. Chairs are in that sense multiply realized and thus there is no link that allows physical explanations to replace common sense ones. The root idea here is thus that categories at one level of description may pick out kinds that look disparate in another vocabulary. It is important to note that the multiple realizations problem undermines the common conclusion that aggregate economic phenomena *must* be reducible because they are made from individual behaviour. Reduction is a claim about what specific theories can in principle explain. From the fact that As are composed of Bs it does not follow that a specific theory applying to the Bs has the explanatory resources to eliminate explanations in the categories that describe the As – we make various claims about chairs that cannot be cashed out in quantum mechanics even though chairs are made entirely of atoms.

One–many relations: reductive definitions can fail in the other direction in that in the reducing theory the descriptions used are not sufficient to fix the descriptions in the theory to be reduced. This is failure of the one-to-one mapping as well, but in the other direction.

Presuppositions: the reducing theory may find itself implicitly in need of categories from the theory to be reduced in its own accounts. To take an example from reductionism debates outside economics, attempts to explain antibodies in purely biochemical structural terms arguably fail, because in the end none of the physical descriptions suffices unless an immune response also occurs (Kincaid, 1997). But appealing to immune response seems not be giving a physical explanation but a biological one.

It is an empirical issue whether in fact these sorts of obstacles are real for reduction in economics. When and where they are real need not be uniform across every economic sub-domain and economic model – reduction might be feasible for some and not for others. However, there is some good evidence to think that reduction is often unfeasible for the following kinds of reasons:

1. Much explanation in economics that might seem individualist in spirit is really nothing of the sort. One case in point is the widespread use of representative agents who are not flesh and blood individuals and who cannot be legitimized as reasonable aggregations of individual behaviour. Another is the widespread practice of taking household and firms as basic entities. These are social, aggregative entities

that, when treated as black boxes, belie a commitment to individualism.

2. There are various reasons to think that multiple realizations of economic categories in terms of individual behaviour are likely. (*a*) Many claims in economics are at least implicitly motivated by selectionist arguments, for example, that firms must be profit maximizers if they are not to be weeded out by competitive selection. However, such selective mechanisms do not ‘care’ about how profitability comes about, only that it does. This means that there may be multiple ways of organizing individual behaviour that meet the criterion. This possibility is reinforced by results in the theory of the firm, where there are many plausible models of how profit-maximizing behaviour might be brought about by the organization of individual incentives. (*b*) Much applied economics consists of estimating aggregate supply and demand curves in specific markets. This work proceeds quite well without estimating individual utility functions. This suggests that these aggregate phenomena are indifferent to the exact details of individual behaviour. Becker's (1976) argument that downward-sloping demand curves would be expected from random choices given budget constraints provides a theoretical account of why this is likely to be the case. (*c*) Solid results from physics and elsewhere suggest that complex causality in aggregate phenomena often show structural relations or ‘universalities’ (Batterman, 2001) that are indifferent to a wide range of underlying detail and that in fact are described by categories that are ‘scale relative’ in that they have no counterpart in smaller scales of resolution (Ladyman and Ross, 2007). Hoover (2001) makes an argument like this about macroeconomic variables: the rate of inflation or the GDP has no obvious meaning at very fine scales of measurement. Similarly, equilibrium analysis in terms of strong attractors and the like in evolutionary game theory also presents a parallel situation. The properties of an equilibrium can be understood while there is a wide range of actual dynamic paths to that equilibrium, the details of which are inessential to the equilibrium explanation.

3. Economic explanations involving individuals often rest on – they take as given and unexplained – information about institutions, structures, norms, and so on that are not cashed out in individualist terms. In short, they presuppose rather than eliminate social processes. As we noted above, Schumpeter's claim that economics since its inception has been methodological individualist in orientation is implausible in any strong form, because Smith, for example, is quite clear that institutions and customs matter in fundamental ways. Not surprisingly much work in economics after the neoclassical revolution has carried the banner of methodological individualism in its rhetoric but its practice is much closer to Smith. One clear illustration of this comes from recent developments in rational choice game theory explanations in economics. The failure of the refinement programme to plausibly eliminate all multiple equilibria means that the focal points that are often used to explain which equilibrium is selected will bring in unexplained norms. Bayesian agents in games reach equilibrium when they have sufficiently similar priors, assuming rather than explaining the social processes that produce consensus (Janssen, 1993). Most fundamentally, game theory explanations have to take as given the possible payoffs, the utility functions of individuals, the information available to them, and the initial distribution of resources. This assumes rather than explains much institutional structure. Property rights have to be defined and so on. Much of the new institutional economics is about how these institutional differences can have strong influences on outcomes. The conclusion to draw is that explanations in terms of individuals have to be supplemented with accounts of collective social and economic phenomena, making reduction – full explanation in individual terms – unlikely.

Another example where the individualist rhetoric can outrun the actual practice comes from explanations

of the distribution of income in standard neoclassical models. The goal is to explain what individuals get in terms of the traits of individuals, for example, investment in human capital and so on. Institutional structure is in the background here as well. The preferences of workers and initial distributions of wealth are taken as given. It is generally assumed that there is a direct link between productivity and earnings, which the considerable work on the theory of the firm shows holds only under specific institutional contexts that are often not satisfied. Most fundamental, those models generally take the distribution of jobs or positions as given. In effect, what is being explained concerns what determines on which rung on the ladder individuals stand. Left unexplained is the number of rungs and the distances between them (see Sattinger, 1993).

Of course, nothing precludes the individualist from seeking further explanations of all these unexplained collective phenomena in purely individual terms. But the breath of these problems and the lack of individualist explanations at this point suggest that the current evidence for the reductionist version of individualism in economics is slim.

I turn finally to the version of individualism claiming that individualist *mechanisms* are necessary. Here I think is the most important individualist insight, though it is important to distinguish various versions of this claim, for some are considerably more plausible than others. The notion of a mechanism is nebulous. The root idea, going back to Maxwell and before in physics, seems to be that of a continuous causal process – one that is not gappy as it were. Taken that way, a mechanism might be either horizontal or vertical. To find the causes between A and B is to find a horizontal mechanism and explaining how the parts of A contribute to its causal influence on B is identifying the vertical sense of mechanism. A related important distinction concerns how a mechanism is described and in what detail. An ‘antibody’ and a ‘compound of such and such a structure’ may commit us to different things.

With these distinctions in hand, here are some general things that can be said about mechanisms in science in general. We can sometimes know that A causes B without knowing either the horizontal or the vertical mechanism. To cite a common sense example, I can know that the flying baseball caused the broken window without knowing the quantum descriptions of the baseball's constitution or the exact details of how the ball surface interacted with the glass. Furthermore, the notion of knowing the ‘full’ mechanism is not well defined, since we can generally give more fine-grained descriptions of the constituting parts of or of the time periods between causes; no account of a causal process is a complete explanation in the sense that there are no unanswered questioned that might be answered. Finally, the place of specific mechanisms in our accounts of the world seems to depend on three things: how solid our knowledge is at the level of description we are using to pick out the mechanism, how solid our knowledge is about that process for which we are seeking mechanisms, and to what extent the two make presuppositions about the other. An account of large-scale brain structures, for example, that required neuronal processes at speeds beyond the known synaptic firing times would be suspect. An explanation that was well confirmed at the scale of brain structures through experiment and physical tracing and that relied on no very specific view of neuronal details should not be strongly constrained by molecular mechanism, particularly if our understanding of the molecular details was much less solid than our understanding at the level of brain structure.

From this general perspective, some claims that individualist mechanisms are essential in economics are plausible and some are not. Among the implausible is that no economic explanation ever succeeds until there is an account in terms of individual maximizing behaviour and general equilibrium. There is just too much good work in economics that provides apparently well-confirmed explanations without

meeting this requirement. As noted above, much applied economics is about aggregate supply and demand that has no general equilibrium foundations. Other compelling explanations in industrial organization describe firm behaviour in competitive environments of various kinds with no pretence of providing a foundation in individual (as opposed to firm) maximizing behaviour. Good econometric work in macroeconomics can use structure breaks between macroeconomic variables to show causation without any account of underlying individual behaviour (see Hoover, 2001).

Of course, these accounts could certainly be made stronger by providing some account of how they relate to individual behaviour. Given everything we know from experimental and behavioural economics, however, the theory would not be a simple picture of individuals maximizing utility functions. In any case, the fact that non-individualist explanations can be made stronger does not thereby mean they are bad explanations – if does not follow from the fact that I cannot answer all questions about a domain that I can answer none.

Alternatively, mechanisms in terms of individual behaviour can be plausible requirements indeed in the right circumstances. Critics of Keynesian orthodoxy had reason to be critical in that Keynesian models required individuals to be systematically fooled. Critics of rational expectations models, however, could with equal justification turn the tables and reject those models because of their lack of individualist mechanisms in that they require individuals to make the best econometric forecast given the available data. This thus illustrates the theme of this article, which is that, once we move beyond the individualist rhetoric typical of the economics profession, individualism and holism have many different claims that vary enormously in plausibility – the devil is in the details.

See Also

- aggregation (theory)
- explanation
- methodological individualism

Bibliography

Batterman, R. 2001. *Devil in the Details*. Oxford: Oxford University Press.

Becker, G. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Deaton, A. and Muellbauer, J. 1980. *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.

Fodor, J. 1974. Special sciences (or the disunity of science as a working hypothesis). *Synthèse* 28, 97–115.

Gordon, S. 1991. *The History and Philosophy of Social Science*. London: Routledge.

Hellman, G. and Thompson, F. 1975. Physicalism: ontology, determination, and reduction. *Journal of Philosophy* 72, 551–64.

Hoover, K. 2001. *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Janssen, M. 1993. *Microfoundations: A Critical Inquiry*. London: Routledge.

Kincaid, H. 1997. *Individualism and the Unity of Science*. Lanham, MD: Rowman and Littlefield.

Kincaid, H. 1996. *Philosophical Foundations of the Social Sciences*. Cambridge: Cambridge University Press.

Ladyman, J. and Ross, D. 2007. *Every Thing Must Go*. Oxford: Oxford University Press.

Martel, R. 1996. Heterogeneity, aggregation, and a meaningful macroeconomics. In *Beyond Microfoundations*, ed. D. Colander. Cambridge: Cambridge University Press.

Sattinger, M. 1993. Assignment models of the distribution of income. *Journal of Economic Literature* 31, 831–80.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

How to cite this article

Kincaid, Harold. "individualism versus holism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000277> doi:10.1057/9780230226203.0782

The New Palgrave Dictionary of Economics Online

indivisibilities

William J. Baumol
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A commodity is indivisible if it has a minimum size below which it is unavailable, at least without significant qualitative change. Indivisible inputs yield economies of scale and scope. But even where indivisibilities impose large fixed costs, if they are not sunk, potential competition can impose behaviour upon incumbents that is consistent with economic efficiency. Perhaps the most significant way in which indivisibilities can impede efficiency in pricing is the existence of indivisible input–output vectors that are efficient but which are not profit maximizing at any positive scalar prices. Integer programming is naturally suited to optimality analysis involving indivisibilities.

Keywords

barriers to entry; circular flow; economies of scale; economies of scope; Euler's theorem; fixed costs; indivisibilities; integer programming; marginal cost pricing; natural monopoly; non-convexity; Ramsey pricing; sunk costs

Article

A commodity is indivisible if it has a minimum size below which it is unavailable, at least without significant qualitative change. Most commodities are indivisible but this is often unimportant. Half a chair has little use, but this makes little difference for analysis of market demand because so many are sold that there is little inaccuracy in treating an increase in sales of chairs from 10 million to 10,000,001 as a change in a continuous variable. In other cases, minimum size is so large relative to usage that it requires special analytic approaches and has substantial behavioural consequences: a Boeing 747 passenger aircraft is a large outlay for any airline; to carry *any* freight from New York to Chicago a railroad must lay at least two rails, each about 1,000 miles long.

Fixed cost and sunk cost

The *fixed cost* of a firm is defined as the minimum outlay it must incur to carry out any activity. If we write (assuming input prices fixed) the long-run cost function as $C(y) = k + f(y)$, where $k = \text{constant}$, $f(0) = 0$, and y=the vector of output quantities, then k is the fixed cost. As in the railroad example, the need for indivisible equipment is the normal source of fixed costs.
Fixed costs are important in economics as a source of economies of scale, of impediments to the workings of the price mechanism, of breakdown in the convexity conditions usually relied upon in optimization calculations and in the uniqueness of solutions.
Fixed costs are often confused with *sunk costs*, which are also related to indivisibilities. A sunk cost may or may not be larger than the minimum outlay a firm needs to operate but, once incurred, it cannot be withdrawn for some substantial period without significant loss. An automobile producer may build a plant much larger than the minimum needed to turn out one car, and once the capital is sunk it may only be possible to retrieve it gradually as vehicles are sold. Thus, sunk costs (like the car plant) need not be fixed and fixed costs (like an aircraft) need not be sunk.

Economies of scale and scope

Indivisible inputs by their nature yield economies of scale and scope. An indivisibility requires a producer of even a small output volume to acquire relatively large capacity, part of which must be unused. The firm can then increase its outputs without increasing costs proportionately (economies of scale). Formally, strict economies of scale are defined to be present at output vector *y* if $C(ay) / a > C(y)$, where $0 < a < 1$, that is, if average cost is declining along the ray *ay*. With fixed costs this becomes $[k + f(ay)] / a > k + f(y), k > 0$. Then, assuming that $f(y)$ is bounded from both above and below, say, $0 \leq f(y) \leq M < \infty$, the scale economies criterion must clearly be satisfied as *a* approaches zero. Thus, the presence of fixed costs always introduces scale economies (so defined), at least in any neighbourhood of the origin.

If the indivisible item is not too specialized, the firm can add commodities to its product line without the combined costs equalling the sum of those of several more specialized enterprises which together produce the same output vector as our firm. The latter attribute is referred to as *economies of scope*. Formally, using the three product case $y=(y_1, y_2, y_3)$ for simplicity, strict economies of scope are defined by $C(y) < C(y_1, 0, 0) + C(0, y_2, 0) + C(0, 0, y_3)$.

Together, economies of scale and scope are what underlie the phenomenon of natural monopoly. An industry is said to be a *natural monopoly* at output *y* if one single firm can produce *y* more cheaply than can be done by *any* combination of two or more firms. Formally, if y^i is the output vector of firm *i*, then the industry is a natural monopoly at *y* if $C(y) < \Sigma C(y^i)$ for each and every set of y^i such that $\Sigma\ y^i=y$.

Scale economies lead to natural monopoly because in their absence it may be possible to save resources by dividing the industry's output among several firms, each providing similar proportions of the industry's output vector. Specifically, the absence of (weak) scale economies at *y* means that, for some values of *a*, $C(ay) / a < C(y),\ \ 0 < a < 1$. Suppose there exists such a value of *a* at which $b=1/a$ is an integer. Then the industry can reduce cost by dividing output among *b* firms each producing $y^i=ay$, at total cost

$$\sum C(y^i) = bC(ay) = C(ay) / a < C(y),$$

thus violating the criterion of natural monopoly. Economies of scope are relevant because in their absence it may be possible to save resources by dividing up the industry's products among specialized enterprises. Specifically, for example in the two-product case, absence of weak economies of scope means $C(y_1, 0) + C(0, y_2) < C(y) = C(y_1, y_2)$, also violating the natural monopoly requirement.

It can also be shown that scale economies together with an attribute closely related to economies of scope are sufficient (but not necessary) for an industry to be a natural monopoly (see Baumol, Panzar and Willig, 1982, pp. 178, 187–8).

Indivisibilities, sunk costs and barriers to entry

The literature offers various definitions of ‘barriers to entry’, some mutually inconsistent. If one defines them as impediments to the invisible hand mechanism, then sunk costs are entry barriers while fixed costs are not.

The need to sink capital into an enterprise constitutes a risk which obviously can deter a potential entrant and thus can protect incumbents from potential competition. So, in an industry with relatively large sunk costs, monopoly profits and inefficiencies become possible.

On the other hand, even where indivisibilities impose large fixed costs, if they are not sunk, potential competition can impose behaviour upon incumbents that is consistent with economic efficiency. Where the fixed capital is highly mobile and there is an active market on which it can readily be sold (as with, for example, ocean cargo vessels) then the fixed capital constitutes no special risk and is no impediment to entry. Even if the indivisibilities make the industry a natural monopoly it will be unable to earn excess profits, operate inefficiently or behave like a protected monopolist in other ways, because this will attract entry that – with no sunk costs – incurs little risk and punishes the misbehaving monopolist.

Indivisibilities as impediment to efficient pricing

Perhaps the most significant of the ways in which indivisibilities can impede efficiency in pricing is the existence of indivisible input–output vectors that are efficient but which are not profit maximizing at any positive scalar prices. This is best shown diagrammatically. In Figure 1 (Frank, 1969, pp. 5, 42–3), $y_1 \leq 0$ and $y_2 \geq 0$ are the input and ouput quantities respectively. With both of them indivisible, the dots, or *lattice points*, represent the only feasible input–output combinations. Point $A=(-2, 1)$ is efficient since no feasible lattice point lies to its northeast. However, *A* lies inside the convex hull of the (non-convex) feasible region whose northeast boundary is ray *OR*. Hence, any line given by $p_1y_1+p_2y_2$=profit,

through point *A* must lie below at least one lattice point on *OR* (here, either *B* or 0). Thus, at any non-negative prices efficient point *A* must be less profitable than 0 or *B* – no simple prices can lead profit maximizing firms to produce *A*. Only a set of 'nonlinear prices' (for example, two-part tariffs), which lead to a curved isoprofit locus such as *PP*, can induce production of *A*.

Figure 1

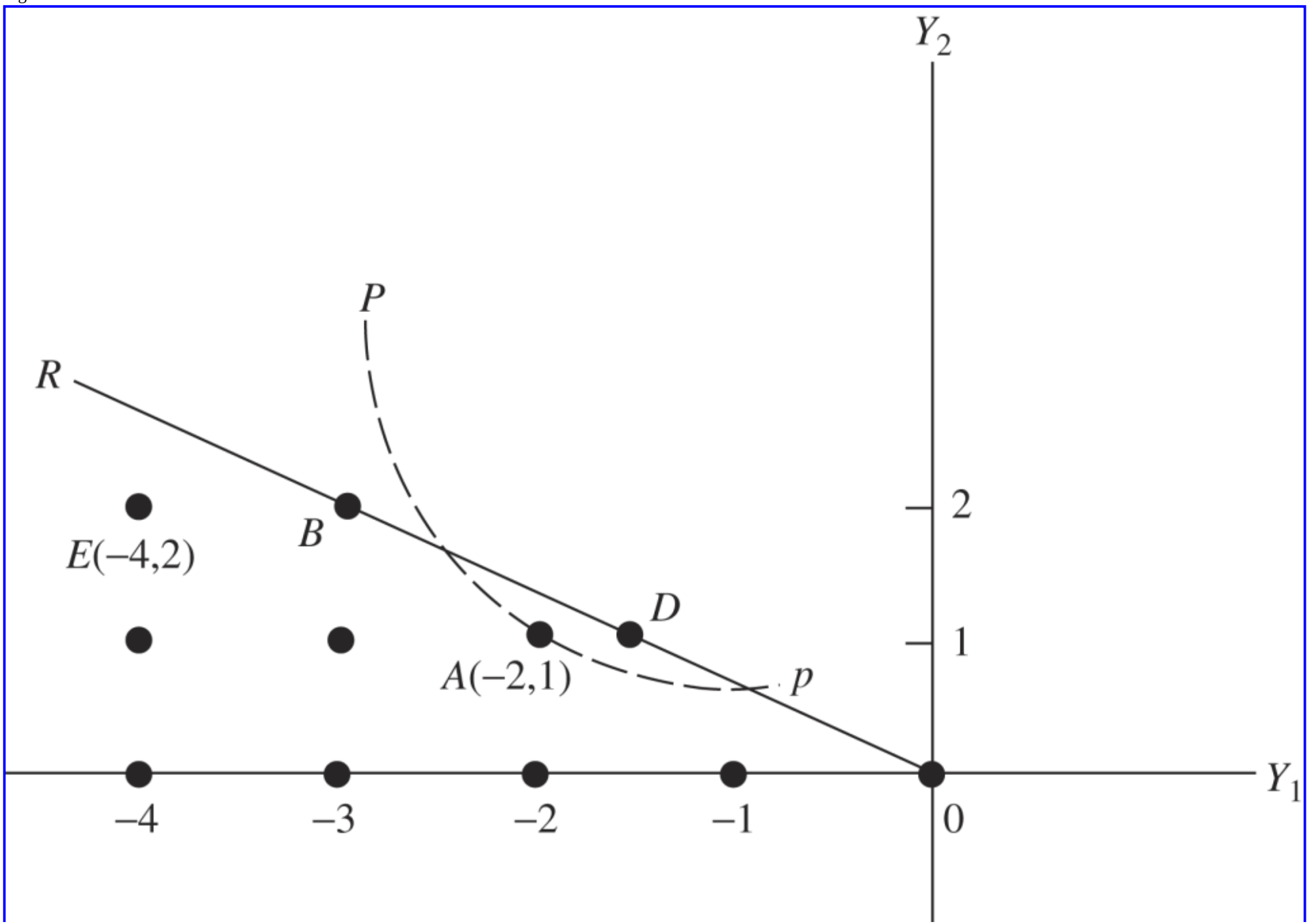

The diagram demonstrates how indivisibilities lead to non-convexity. For example, a line segment connecting points *A* and *B* in Figure 1 clearly is not composed entirely of lattice points, that is, it is not entirely contained in the feasible set of lattice points, and so that set is not convex.

The graph also shows in another way how indivisibilities introduce scale economies. Consider *D*, a non-lattice point on *OR* to the right of *A*. Let *c* be the smallest integer for which *c* (distance *AD*)$\geq$1 (in the graph *c*=2). Then *cA* will be a feasible lattice point (point *E*), but there will be a point (*B*) between *E* and *E* – *c*(*AD*) which is a feasible lattice point, with the same output and a smaller input quantity than those at *E*. Since *E* is an integer multiple of efficient point *A*, one can multiply output by $c > 1$ while multiplying input by a smaller amount, that is, there must be scale economies.

Indivisibilities impede the price system in yet another way. By creating scale economies they make marginal cost pricing unprofitable. Specifically, let *y* be an output vector at which there are scale economies so that $C(ay)=a^bC(y)$ in the neighbourhood of *y*, with $b < 1$. Then, the function is locally (approximately) homogeneous of degree *b* and by Euler's theorem $\Sigma\, y_i \partial C / \partial y_i = bC < C$. Hence, if prices are set equal to marginal costs the supplier must lose money. In that case, financial feasibility requires the substitution of Ramsey prices (see Ramsey pricing) for marginal cost prices to achieve a second-best optimal resource allocation. This is true not only for the individual firm – the entire economy may have no parametric price option that is superior to Ramsey prices. For all outputs must be sold to suppliers of inputs and the receipts from output sales are paid out as wages, profits, and so on, to the input suppliers. This imposes (in the absence of lump sum payments with parametric prices of inputs and outputs) the economy's circular flow requirement $\Sigma\, p_i y_i=0$, again taking input quantities to be negative. Now, a set of Pareto optimal prices p_i^* will, in general, not satisfy this constraint. The second-best prices, p_i, which are constrained to satisfy this requirement, are by definition the Ramsey prices and the differences $t = p_i - p_i^*$ between Ramsey prices and optimal prices may be interpreted as the optimal vector of taxes needed for compliance with the economy's circular flow constraint.

That is the form in which Frank Ramsey's original treatment is expressed. As we have just seen from the Euler's theorem argument, where costs are differentiable with respect to outputs, the first-best prices of the outputs, which are their marginal costs, will not satisfy the circular flow constraint when there are scale economies. This shows that in general, where indivisibilities create scale economies, optimality in pricing cannot avoid the complications of Ramsey theory.

There is a third way in which indivisibilities complicate the optimization process. As is well known, where the feasible set is not convex, as must be true when there are indivisibilities, a multiplicity of local maxima is likely to be present and an iterative solution process that always follows a direction in which profit (or the value of the social objective function) is increasing may well lead towards a local optimum rather than one which is global.

Integer programming and the analysis of indivisibilities

Integer programming is the mathematical technique that is naturally suited to optimality analysis involving indivisibilities. An integer programme is a mathematical programme in which only integer values are admissible for some or all of the variables. The constraint requiring *x*=number of locomotives to be an integer is what keeps the solution from including the absurd recommendation that 1.783 locomotives be produced.

Integer programming also permits the solution of more subtle indivisibility problems, such as those involving scale economies or either/or choices, which have resisted other analytical techniques. As an example, consider a firm required to produce *y* units of output using either a machine of type 1 or a machine of type 2, where *x* is the vector of other inputs, and x_1 and x_2 are the respective numbers of the two types of machines purchased, $\Pi(y, x, x_1, x_2)$ is the profit function and $y \leq f(x, x_1, x_2)$ is the production constraint. Then the firm must

$$\text{maximize } \Pi\ (y, x, x_1, x_2)$$

subject to the constraints

$$y \leq f(x, x_1, x_2)\, y, x, x_1, x_2 \geq 0\, x_1 + x_2 \leq 1\, x_1, x_2 \text{ integer}.$$

The last two constraints guarantee that x_1 will take either the value zero or unity and that (at least) one of them will be zero, as an either/or decision requires.

Economies of scale and scope raise related issues. Such cases tend to yield corner rather than interior solutions. If there are n firms, each with different attributes, which are candidate producers of industry output vector y, it is likely to be most economical for just one of them to produce all of y. But which one of the n firms should do the job? That is obviously an extended either/or issue whose formal statement is perfectly analogous to that just described.

Indivisibilities give rise to other complex combinatorial problems. The choice among m machines may, for example be constrained by the fact that a machine of type A will work only if a machine of type B is also purchased. This is dealt with via the constraints $x_a \leq x_b$, x_a, x_b integer. In such problems the indivisibility feature is fundamental and cannot be avoided by non-integer approximation. In sum, indivisibilities raise basic issues for theory and for methods of analysis which bear little resemblance to those pertinent to cases of divisibility.

See Also

- contestable markets
- Ramsey pricing

Bibliography

Baumol, W.J., Panzar, J.C. and Willig, R.D. 1982. *Contestable Markets and the Theory of Industry Structure*. San Diego: Harcourt Brace, Jovanovich.

Dupuit, J. 1844. De la mésure de l'utilité des travaux publiques. *Annales des Ponts et Chausées*, 2nd Series, vol. 8. Repr. in *International Economic Papers* No. 2 (1952), London: Macmillan.

Frank, C.R., Jr. 1969. *Production Theory and Indivisible Commodities*. Princeton: Princeton University Press.

Gomory, R.E. 1965. On the relation between integer and non-integer solutions to linear programs. *Proceedings of the National Academy of Sciences* 53, 260–65.

Gomory, R.E. and Baumol, W.J. 1960. Integer programming and pricing. *Econometrica* 28, 521–50.

Koopmans, T.C. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Lewis, W. 1949. *Overhead Costs*. London: George Allen & Unwin.

How to cite this article

Baumol, William J. "indivisibilities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000069> doi:10.1057/9780230226203.0783

The New Palgrave Dictionary of Economics Online

industrial relations

Thomas A. Kochan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Industrial relations is an interdisciplinary field of study that encompasses all aspects of work and employment relations. Originating in institutional economics and Fabian socialism, it has evolved to address employment problems and issues ranging from wage determination and collective bargaining to human resource management and labour market dynamics and policies. Globalization has increased the importance of international and comparative analysis of employment practices and outcomes. Shifts in employment from manufacturing to services has rendered the term ‘industrial relations’ obsolete and led scholars to use the term ‘work and employment relations’ to describe their work and this field of study.

Keywords

American Economic Association; arbitration; class conflict; collective bargaining; Commons, J. R.; corporate governance; corporations; Ely, R. T.; Fabian socialism; globalization; Great Depression; health insurance; human capital; human relations; human resource management; incentive compensation; industrial psychology; industrial relations; innovation; institutional economics; internal labour markets; Labor and Employment Relations Association (USA); labour economics; layoff; Marx, K. H.; minimum wages; networks; New Deal; pensions; personnel economics; power; scientific management; social insurance; social networks in labour markets; strikes; team production theory of the firm; technical change; trade unions; training; unemployment insurance; wage determination; Webb, S. and B.; women's work and wages; workers’ compensation

Article

Industrial relations is an interdisciplinary field devoted to the study of all aspects of work and employment relations. It emerged historically out of the works of the Fabian Socialists Sydney and Beatrice Webb (1894; 1897) in Great Britain and institutional economists such as John R. Commons

(1909; 1934) in the United States. Both sets of scholars and their students were searching for ways to understand *and* influence employment relations in ways that distinguished their normative, theoretical and methodological approaches from Marx (1849) on the one hand and classical or neoclassical economics (Marshall, 1920) on the other.

Over the years the field has evolved and broadened considerably to incorporate concepts and methods from other social sciences such as psychology, sociology, and political science and from disciplines outside the social sciences such as history and law. In recent years the term ‘industrial relations’ has become somewhat dated, given the growth of the service sector and the decline of traditional manufacturing industries, leading a number of research units working in this scholarly tradition to redefine the field as the study of ‘work and employment relations’. But the underlying normative, theoretical and methodological features of the field carry on the distinctive features of industrial relations.

Origins and initial intellectual debates

Karl Marx provided the intellectual rationale and stimulus to the field of industrial relations. His most enduring contribution was to assert that labour was more than just a commodity or factor of production subject to deterministic laws of supply and demand. Instead, the free will and power that reside in human beings make labour more than an inanimate object. This basic insight serves as an enduring normative premise in industrial relations and motivates much of the work in the field to this day. That is, while affected by market forces similar to other factors, labour deserves and requires special treatment in theory and public policy because workers can take individual or collective actions to influence market outcomes, and work and employment relationships affect important human values and have important social as well as economic consequences. For these reasons, industrial relations research, public policies and practices need to be as concerned about equity as efficiency at work (Barbash, 1984; Meltz, 1989).

Moreover, freedom of association at work is recognized as a fundamental human right in democratic societies and, therefore, the ability of workers to have a voice in determining their employment conditions serves as an equally important industrial relations outcome (Budd, 2004).

While Marx provided the starting point for the field of industrial relations, much of the scholarship in the field has taken issue with other aspects of Marxian analysis. This is especially true of the Marxian view of the source of labour conflict in employment relations. Marx saw conflict at work as inevitable and all-encompassing, arising out of class differences rooted in the capitalist system of production. Conflict could be eliminated only by the revolutionary overthrow of that system. This became a major point of differentiation between Marxist and labour process schools of industrial relations on the one hand (Hyman, 1975) and on the other hand the more mainstream pluralist model which has grown to dominate European and Anglo-Saxon research traditions (Clegg, 1970; Fox, 1971; Kochan, 1980).

Sidney and Beatrice Webb (the Webbs) were among the first to challenge Marx with their model of Fabian socialism. They shared with Marx a concern for the plight of the growing working class. Beatrice Webb was a student of the factory conditions prevailing in 19th century as Britain ushered in the first Industrial Revolution. Her empirical observations of factory conditions convinced her that the average worker suffered from an inherent imbalance of power in dealing with his or her employer. Trade unions were therefore needed to provide increased social support and bargaining power. Over time, however,

unions were expected to evolve into institutions that promoted orderly government regulation that worked to the common benefit of all workers and for the overall community. Thus through evolution, not the revolution predicted by Marx, societies would evolve to better balance the needs of workers, employers and the communities in which they were embedded.

At about the same time as the Webbs were doing their work in Britain a quiet revolt was taking place in the field of economics in the United States. Leading economists were frustrated with the highly deductive and mathematical features of late 19th-century economic research. As a result, in 1886 Robert Ely at the University of Wisconsin led other colleagues to form a new American Economic Association in an effort to bring a more inductive, empirical, and institutional brand of economics to bear on the critical problems of the day. Labour economics, more than any other sub-field within the economics profession, took up the institutional approach. Led by Ely's protégé at Wisconsin, John R. Commons, a new field was born. It focused on the study of labour and working conditions using empirically based, inductive methodological methods and focused as much on the collective institutions and organizations of workers and employers governing work and employment relations as on the actions of individuals in response to market forces.

Like Marx, these institutionalists believed that labour was more than a commodity. But unlike Marx, Commons and those that followed in building the field of industrial relations saw the conflicts of interests between employers and employees as part of natural, legitimate and ongoing differences in economic interests, not as a function of the capitalist system. Employers have the responsibility of promoting efficient use of scarce resources, including labour. Employees have the right and need to pursue their self-interests, individually or collectively, to improve their security, wages, working conditions, and other features of their work lives they value. These conflicting interests are not, however, absolute. Employers and employees also have some common interests that tie them together in ongoing interdependent relationships. Both want to generate value from their relationships so that there is more value to share. Safety and security may be other shared values. Commitment to the mission of the organization and service to their clients, customers, patients and so forth may be other shared values and objectives. Thus employment relations involve an inevitable mix of separate, perhaps conflicting, and common or shared goals. The task of industrial relations theory, research, teaching and policy therefore focuses on both finding an equitable resolution of differences or conflicts and ways to support value, creating solutions where interests overlap or are held in common (Walton and McKersie, 1965).

The early institutionalists were strong proponents of empirical research and active involvement in policymaking and institution building. They studied labour market dynamics and labour management relations through field work more than through deductive model building. Their collective body of research and personal involvement generated most of the ideas and policy proposals embedded in the labour legislation of the New Deal. Unemployment insurance, workers’ compensation, child and women's labour protections, minimum wages and social security all were ideas developed and studied at state and local levels of the economy between 1900 and 1930. Commons is now widely recognized as the intellectual father both of the New Deal labour legislation and the study of industrial relations in America (Kaufman, 1993).

Debates with alternative disciplines

Kuhn (1970) argues that a new paradigm for the study of a phenomenon must be judged ultimately by

whether it is better able than its alternatives to solve problems. So it is appropriate to examine industrial relations against this criterion at critical stages in its development.

Scientific management: scientific management and industrial engineering dominated the study and practice of management and the design of work systems in the United States in the first two decades of the 20th century. The objective was to use engineering principles to find the optimal, most efficient methods for carrying out tasks, organizing them into a clear hierarchy and controlling labour through appropriate economic incentives and supervision to conform to the specified work process. In following these scientific engineering principles one would eliminate any potential conflicts of interests at work (Taylor, 1895). This view of work and employment relations saw no rationale for worker voice, representation, or policies that would balance power between workers and managers. Its primary theoretical prediction was that efficient organization and supervision of work, when supported by the right individual incentive compensation system, would generate maximum efficiency. Because efficient work would be rewarded, it would in turn generate worker satisfaction. This virtuous cycle would keep conflict from emerging in employment relationships. Thus, scientific management theory and efforts to implement it in practice stood in sharp contrast to industrial relations theories and normative assumptions.

Industrial psychology: at the same time industrial psychology was emerging as a field of study that paralleled and complemented the engineering approach. The study of personnel management largely grew out of industrial psychology. In contrast with the institutional economists, individuals, not collective groups or organizations, were the central unit of analysis and the firm was viewed more as a closed system, on the assumption that management controlled workplace decisions. Institutionalists reflected their economics' training by treating work and organizational practices as influenced by both organizational and external market and technological forces.

Human relations: in the 1920s the field of human relations was born out of the Hawthorne experiments (social-psychological experiments conducted at the Hawthorne Works' plant of Western Electric) in group behaviour and gave rise to another competing paradigm for the study of work and employment relations. The human relations school focused on work groups as the key unit of analysis and the social dynamics that shaped worker attitudes and behaviour. Human relations theorists reversed the theoretical argument of scientific management by proposing that worker satisfaction drove efficiency at work rather than the other way around (Rothlisberger and Dickson, 1939). This school of thought provided intellectual foundation for the emergence of welfare capitalism in the 1920s. Large firms sought to provide a set of benefits and positive working conditions in order to achieve efficiency and in the process of doing so eliminate the incentives of workers to join trade unions (Jacoby, 1991).

These were the alternative paradigms competing for influence with industrial relations over the first 30 years of the 20th century. The Great Depression of the 1930s raised industrial relations ideas, policies and research to a more prominent and perhaps dominant place in the intellectual and policy debates about work and employment relations. With the rise in industrial conflict and massive unemployment came the recognition of the need to establish a floor on labour standards and a means for workers to bargain as relative equals with their employers to improve on these minimum conditions. Thus, it was the dramatic deterioration in economic conditions, the threats unregulated conflict posed to democracy and social stability, and the shift in the political environment that allowed the ideas and research evidence of the institutional economists to emerge as the intellectual basis for much of the New Deal legislation passed in the 1930s.

The New Deal era, the Second World War and the War Labor Board

From 1932 to 1945 industrial relations scholars and practitioners had an unprecedented impact on national policy and private practices of employment relations in the United States. The War Labor Board (WLB) (1941–5) that was charged with controlling wages and mediating collective bargaining negotiations played a key role in legitimating and starting the long-term diffusion of many modern personnel and labour relations practices and benefits including grievance arbitration, cost of living wage increases, paid time off for holidays and sick leave, paid health insurance and private pensions.

The first two decades of the post-war era were dominated by institutional economists and scholars from sociology, political science, law, labour history, and psychology who united around a common desire to better understand and regulate labour management relations. In 1946 strike levels in the USA reached their historic peak. Concern over the escalating labour–management conflict led a number of state legislatures to create new multidisciplinary schools or centres of industrial relations in leading universities such as Cornell, Wisconsin, Illinois, Michigan State, Rutgers and the University of California at both Berkeley and UCLA. In 1947 a new scholarly professional association, the Industrial Relations Research Association, was created. This association continues today under the name of the Labor and Employment Relations Association.

Two sets of questions featured prominently in industrial relations research in decades following the Second World War: (*a*) how does collective bargaining work and (*b*) what are the effects of unions and collective bargaining on management, the workforce and the economy? A debate arose over whether political (that is, pressures from union members and the need for union leaders to match settlements achieved in closely aligned industries or occupations) (Ross, 1948) or economic forces (Dunlop, 1944) were the primary drivers of wage determination. While never fully resolved, the evidence suggested that both play roles – political forces are influential within a range but are limited by market conditions. A reformulation of the debate by one institutional economist suggested that bargaining power includes a mixture of political, economic and ‘pure power’ forces and that these should be incorporated into a more complete theory of wage determination under collective bargaining (Levinson, 1968).

The growing presence and pressure of unions and collective bargaining from the 1930s through the 1950s exerted what one set of researchers called a shock effect on management. Personnel practices had become more professionalized and applied in more uniform fashion and management had to search for ways to improve productivity to recoup the higher wage costs resulting from collective bargaining (Slichter, Livernash and Healy, 1960). Much of industrial relations research over this time period examined the dynamics of labour management relations and the causes of strikes and/or industrial peace (Golden and Parker, 1955). Most of this work was carried out using qualitative case studies or historical studies of specific unions or of industrial relations in particular industries.

Dunlop (1958) criticized post-war industrial relations research for being characterized by too many facts chasing too little theory. He sought to correct this problem by proposing a general systems theory of industrial relations. He argued that the central task for industrial relations theory was to explain variations in the rules governing employment relations. These rules were set in interactions among three key actors – labour, management and government – and conditioned by external market, technological and societal forces. The system was bound together by what Dunlop argued was a shared ideology

valuing democracy, respect for market forces and worker rights. Although Dunlop's framework never reached the level of being accepted as a general theory of industrial relations, it became the starting point for much of what constituted industrial relations research in the decades following publication of this important work.

Public sector unions and collective bargaining

Government employees were not covered under the National Labor Relations Act (NRLA) of 1935 and, with a few exceptions such as postal employees, remained largely non-union until the 1960s. In 1958 Wisconsin enacted the first of what would grow to be a surge of state legislation protecting state and local employees' right to unionize and engage in bargaining. By 1976 38 states had enacted similar statutes, employing various forms of mediation, fact-finding, and arbitration to resolve contract disputes in lieu of the right to strike. Only a handful of states provided public employees the right to strike and even in these cases police and firefighters were not given the right to strike. Federal employees were granted similar rights to negotiate over non-wage and benefit issues, first through an Executive Order enacted in 1962 and then through legislation enacted in 1978.

As a result, unionism among public employees grew from its minimal level prior to 1960 to reach its present level (in 2007) of approximately 37 per cent of all government employees. These developments produced a significant body of new research on public sector collective bargaining throughout the 1960s and 1970s. Most of this work focused on the performance of mediation, fact-finding and arbitration as deterrents to strikes. The consensus findings of these studies is that arbitration has been successful in deterring strikes of public employees (Olson, 1988). Other studies have focused on the effects of public sector unions and collective bargaining on wages and government budgets. The general findings of these studies are that unions can increase wages. Prior to 1980, estimates suggested the union effect was around five per cent. After 1980 it rose to 20 per cent for local government employees and ten per cent for federal employees (Gunderson, 2007).

Internal labour markets

The study of labour market behaviour represents another longstanding strand of research in industrial relations, dating back to Commons's (1909) classic historical study of changes in labour and product markets of shoemakers. Throughout the 1940s and 1950s studies of the dynamics of external labour markets followed the institutional tradition by examining the development of industry and regional wage structures (Lester, 1952; Rees and Shultz, 1970).

Interest turned to the study of internal labour markets in the 1970s and thereafter by both economists (Doeringer and Piore, 1972; Osterman, 1984) and sociologists (Baron and Bielby, 1980; Pfeiffer and Baron, 1988). Internal labour markets refer to firm-level rules governing hiring and termination, arrangement of jobs into job ladders, compensation structures that link jobs, and access to and mobility of personnel within and across job ladders. The primary questions of interest in these studies is what substantive rules govern the organization of jobs and job ladders and what factors give rise to the development, continuity and decline of internal labour market rules and practices. There is more consensus over the factors giving rise to internal labour markets than to the degree to which are the

causes of their decline. Internal labour markets arise as a function of pressure from unions, governments and tight labour markets (Jacoby, 1985). Over time these rules gain sufficient acceptance to become norms that sustain them even in the face of changing conditions in the external labour market (Osterman, 1984). A major topic of debate in emerged in the 1990s over whether, and if so why, internal labour markets are declining in importance as firms appear to be more willing to lay off workers, adjust compensation to external market signals and hire more workers from outside the firm rather than train and promote current employees (Cappelli, 1999; Jacoby, 1999). There is no conclusive outcome to this debate. Micro firm-level studies tend to find more significant changes in firm-level rules and employment practices and outcomes while macro labour market studies tend to observe modest reductions in employee tenure (for men but not for women). The relative consensus is that norms governing layoff decisions and internal wage structures have led leading firms to be less reluctant to lay off hourly and managerial employees and more willing to allow their internal wage structures to become more disparate or unequal.

Resurgence of the basic disciplines

In the 1960s and 1970s the disciplines and methodologies from which industrial relations researchers drew became more quantitative as econometric and psychometric tools advanced, micro data-sets on labour market behaviour became more readily available, and computer power became more readily accessible. The vast majority of newly trained labour economists moved away from institutional analysis in favour of drawing propositions from neoclassical economics that could be tested with econometric methods. Studies of individual labour market behaviour grew and studies of collective behaviour, where data were less available, declined. Research on discrimination, mobility, labour supply, returns to education, and human capital flourished while the study of unions and collective bargaining declined. The exception to the shift away from unions and collective bargaining was the use of econometrics to estimate the impact of unions on relative wages of individuals (Lewis, 1963). The consensus estimates of these studies were that private sector unions raised wages of their members relative to comparable non-members between 10 and 15 per cent. These estimates rose to 15 to 22 per cent in the 1970s (Kochan and Helfman, 1981). Unions also were shown to have positive effects on other outcomes such as health and pension coverage, wage inequality, productivity, worker retention and satisfaction with wages (Bennett and Kaufman, 2007). Unions have negative effects on firm profits and satisfaction with non-wage outcomes (such as satisfaction with job content) (Kochan and Helfman, 1981; Freeman and Medoff, 1984).

The development of human capital theory (Becker, 1975) further encouraged the movement of labour economics back into the mainstream of the economics discipline and away from its institutional orientation. Becker's work stimulated others (Lazear, 1998) to apply economic analysis to personnel decisions and practice. The study of alternative forms of incentive compensation and their effects on motivation and performance lies at the heart of personnel economics. A paradox appears to exist: the empirical evidence documents the economic value of incentive compensation to the firm, while use of individual incentives has not grown and in some countries appears to be in decline. Explaining this paradox requires consideration of the social context and other institutional forces that seek to reduce competition among workers and enhance social cohesion at work. Personnel economics’ models of

incentive compensation, therefore, need to be supplemented with sociological theories of group norms and other institutional factors that shape wage determination in contemporary organizations.

The same movement back to their mother disciplines could be observed by the 1970s in the work of psychologists and sociologists studying work and employment issues. Models of motivation, job satisfaction, work performance, turnover, and other aspects of individual attitudes and behaviours, based most often on survey, laboratory, or other data-sets assembled by these researchers, became the dominant topics and methodologies. This has given rise to the more applied field of human resource management research. Human resource management combines analysis of firm-level personnel functions (selection, compensation, performance appraisal, and so forth) with analysis of the links between human resource strategies and individual or organizational performance (Dyer, 1984; Schuler and Jackson, 1987). Most of this work adopted the normative premises of the human relations and scientific management schools rather than those of industrial relations. Thus they focused on how to manage employees through the use of modern personnel and human resource practices and strategies to overcome any sources of conflict in the employment relationship and to foster firm performance.

1980s: a time of transformation

The 1980s proved to be a watershed decade for both the study and practice of industrial relations. A central debate arose over whether reductions in real and nominal wage and other changes observed in collective bargaining were simply temporary adjustments to the deep recession of 1981–3 or signalled a more permanent structural shift in the wage determination process and in industrial relations more generally. Few today doubt that the wage determination and industrial relations practices shifted in fundamental ways in the 1980s by reducing the power of the strike threat, and weakening unions in general. Strike rates (measured in percentage of contract negotiations that involve a strike or percentage of annual work hours lost to strikes) have declined precipitously to the point they are no longer reported by government agencies.

The confluence of the deep recession, increased international competition and a shift to a conservative government in the United States under President Ronald Reagan unleashed a set of changes that created a set of anomalies for much of post-war industrial relations theory and empirical research. Management became more openly hostile and aggressive in avoiding new union organizing, moving operations from union to non-union workplaces. Management replaced unions as the driving force in shaping the process and outcomes of collective bargaining. Nominal wage reductions were negotiated in many employment contracts. New approaches to work organization and employee participation challenged traditional job structures and labour management relations. These developments led to an expanded model of industrial relations that emphasized how the choices made by management in particular (but labour and government as well) in structuring relations at the workplace, in collective bargaining or personnel policies and in high-level business/competitive strategies shape employment relationships and outcomes (Kochan, Katz, and McKersie, 1986).

Analysis of how these choices played out and affected outcomes featured significantly in industrial relations research throughout the 1980s and 1990s. Researchers began to assess the effects of different combinations of employment practices on firm performance, reflecting the systems' perspective of industrial relations and the emerging emphasis on complementary practices in personnel economics

(Milgrom and Roberts, 1992). By the end of the 20th century the evidence suggested that flexible work systems and employee involvement in production and workplace decisions served as positive complements to investments in technology and training, produced significant improvements in productivity and service quality (Ichniowski et al., 1996; Appelbaum et al., 2000). The theory and evidence suggested a high-wage, high-productivity equilibrium was possible in sectors as diverse as manufacturing, airlines, health care and financial services. Yet these ‘high performance’ work systems did not diffuse naturally across the economy, in part because of the costs of transitioning from more traditional practices, and in part because they competed with a low-wage, low-cost equilibrium. These two competing models of human resource practice and industrial relations compete with each other across most industries and occupations in the USA and other countries. A central theoretical and policy question in the field today focuses on whether a high-wage, high-productivity equilibrium can be sustained in the face of low-wage, low-cost competition in domestic *and* international labour and product markets, and, if so, how to best encourage adoption of these strategies.

Policy debates

Concerns over public policy rose in parallel to these theoretical and empirical developments. The central proposition driving policy debates was that the changes in the workforce, nature of work, and the economy had outpaced adaptations in public policies, institutions, and practices in employment relations and that this gap was imposing costs on workers and the economy (Osterman et al., 2001). Efforts to build consensus on changes needed in labour and employment policies consistently failed from the late 1970s to the 1990s (Kochan, 1995). The result is that the field of industrial relations has come full circle to where it began in the early years of the 20th century when Commons and his students documented the mismatch between policies and institutions and workplace relations as the economy transitioned from its agrarian base to a manufacturing base. Today the mismatch is playing out on a global rather than a domestic scale, and therefore the theoretical, institution building, and public policy challenges are broader and perhaps more complex than ever before. As yet, however, there is little public or political support for comprehensive reforms of labour and employment policies. Consistent with the history of policy changes in the United States, it will likely take a significant crisis, combined with a major shift in political power, to achieve a change in policy.

Rebirth of sociological studies of work and labour markets

While sociologists have studied various aspects of work, employment and careers throughout the 20th century (Hughes, 1958; Barley and Kunda, 2001), since the 1980s there has been a significant growth in interest in these topics among sociologists, who now label their work as the study of economic sociology. Economic sociologists implicitly (and sometimes explicitly) seek to counter purely economic models of labour-market behaviour by demonstrating that individual and organizational decisions reflect the social and institutional structures in which they are embedded. Much of this work examines how networks of workers and/or organizations influence labour market behaviour. An early study in this tradition (Granovetter, 1974) documented how networks affect access to job opportunities. Later studies have shown networks to be important in influencing migration (Portes and Sensenbrenner, 1993;

Saxenian et al., 2002), promotions and upward mobility (Burt, 1992), and cooperative relationships among firms in industrial regions (Putnam, 1973; Piore and Sabel, 1984; Locke, 1995). These studies extend the institutional tradition of industrial relations by drawing more heavily on classical sociological theories of Weber (1962), Durkheim (1893) and Selznick (1984).

From industrial relations to work and employment relations

By the beginning of the 21st century many scholars began to recognize that the term ‘industrial relations’ had become increasingly problematic as a label of the study of people at work. The majority of the workforce is employed in services, not manufacturing. Thus many researchers and university programmes have gradually changed the labels used to describe their field of enquiry and/or teaching from industrial relations to work and employment relations, human resource management, work and organizational relations, and a variety of other terms. At the same time, more scholars from traditional disciplines of sociology, political science, economics and social psychology have taken up the study of work and employment issues, which has led to an expansion of the field and to a new round of competition among these different disciplines for influence in shaping the future study and practice of employment relations.

The research questions that are most central to this field today reflect two interrelated realities: (*a*) globalization of economic activity, and (*b*) the importance of knowledge and innovation in structuring work and shaping economic outcomes. Globalization and changes in technology have increased the mobility of capital, work, and workers thereby weakening the influence of national laws, institutions, and norms in shaping employment relationships and outcomes. Once again, today as in the Commons era of the early 20th century, wages and labour costs are under intense competition, only this time more labour markets are international in scope.

The increased ease of locating work and expansion of trade across national borders affects a wide range of work and employment issues and outcomes. Globalization has been associated with, among other things, changes in the distribution of wages and profits, growth in income inequality, and greater and more widely distributed job insecurity. Within firms, globalization of production and supply chains diffuses responsibility for employment decisions and policies, blurring the traditional distinction between employers and employees. All these effects are being subjected to intense analysis, debate, measurement of the direction and magnitude of their effects, and debate over how to adapt policies and institutions to cope with them. These international and organizational developments also make it more difficult to regulate employment relations with national laws and firm-centred rules and policies.

These developments have also generated a debate over the appropriate goals of the modern corporation and its role in society and as an employer. Since the early 1980s the view that firms exist primarily or even solely to maximize shareholder value has dominated academic and public discourse. This view is now being challenged. Blair and Stout (1999) offer a critique of the view that firms exist solely or primarily to maximize shareholder wealth and instead propose a team production theory of the firm. In their view the appropriate underlying view is that the firm should maximize the total value of wealth produced for all the constituents that supply resources and add value to the organization. Human capital plays a central role in this theory since workers contribute and put at risk their human capital by joining and staying with a given firm. The longer workers stay with a given firm, the higher the costs of losing

their job. Thus, like those who invest and put at risk their financial capital, workers are residual risk bearers should the firm fail.

The outcome of this debate could have important consequences for the design of institutions of worker voice in employment relationships. Since 1935 American labour law has taken as a guiding premise that employees should be allowed to bargain over wages, hours, and working conditions and that management should remain free to make strategic business decisions on its own. If by investing their human capital employees become a residual risk bearer similar to financial investors, then there is no logical basis for excluding workers from a voice in strategic decisions and corporate governance. Thus the study of industrial relations has expanded to engage issues of corporate strategy and governance and theories of the firm.

The field has also expanded in response to changes in the relationships between work and family–personal life. Work and family life were tightly linked in the pre-industrial agrarian economy because they were co-located (families lived and worked on the farm) and men, women, and children all contributed to the production process. With the growth of the industrial economy came a clearer division of labour and physical separation in work and family life. The male breadwinner emerged as the prototypical worker, with the assumption that he had a wife at home attending to family responsibilities. With the growth in the labour-force participation of women from the 1960s onward and the slowdown in the growth of real wages, working hours have both been spread more evenly between men and women, and particularly between mothers and fathers. This once again increases the interdependence of work and family life and calls for changes in workplace and human resource practices to provide flexibility in hours and career options for women and men. Thus work and family issues have become an important topic of research and policy analysis within the field of work and employment relations (Bailyn, 2006; Kossak, 2006; Drago, 2007).

International studies

The study of work and employment relations across the world parallels most of the trends observed in the USA. Throughout much of the 20th century, studies of labour movements and labour conflict dominated both country-specific research and international comparisons of industrial relations systems. In the 1960s a debate arose over whether technological changes and increasing economic interdependencies would lead to a convergence in employment systems and practices or whether differences observed across countries would endure because of the influence of national culture and other institutional forces (Kerr et al., 1960). This debate continues today, although researchers have shifted to more micro level (industry, occupational and regional) comparisons to sort out forces leading to convergence and divergence in employment relationships (Katz and Darbashire, 2000; Bamber, Lansbury and Wailes, 2004). Moreover, researchers active in the field of international industrial relations (Kaufman, 2004) are actively analysing and debating most of the issues and developments discussed in this article in countries across the globe.

Historical parallel

In the USA and Britain, the field of work and employment issues of industrial relations have come full

circle to their origins. As in the first two decades of the 20th century, contemporary researchers are driven by a broad proposition that the nature of the economy, workforce, the nature of work and its relationship to other institutions such as family life have all changed dramatically while public policies and institutions remain tailored to a fading industrial-based economy. The gap between policies and institutions and the contemporary realities of work and family life lie at the heart of the tensions and pressures building up in workplaces in America and, increasingly, across the world. The central task of work and employment researchers today, as for their industrial relations forefathers, is to conduct research and policy analysis that prepares for the day that the political forces align to make it possible to begin the updating and modernization process.

See Also

- collective bargaining
- Commons, John Rogers
- human capital
- institutional economics
- Slichter, Sumner Huber
- social contract
- strikes
- Taylorism
- unemployment

Bibliography

Appelbaum, E., Bailey, T., Berg, P. and Kalleberg, A. 2000. *Manufacturing Advantage*. Ithaca, NY: Cornell University/ILR Press.

Bailyn, L. 2006. *Breaking the Mold*, 2nd edn. Ithaca, NY: Cornell/ILR Press.

Bamber, G., Lansbury, R. and Wailes, N. 2004. *International and Comparative Employment Relations*. London: Sage.

Barbash, J. 1984. *The Elements of Industrial Relations*. Madison: University of Wisconsin Press.

Barley, S. and Kunda, G. 2001. Bringing work back in. *Organization Science* 12, 76–95.

Baron, J. and Bielby, W. 1980. Bringing the firms back in: stratification, segmentation, and the organization of work. *American Sociological Review* 45, 737–65.

Becker, G. 1975. *Human Capital: A Theoretical and Empirical Analysis*. New York: Columbia

University Press.

Bennett, J. and Kaufman, B. 2007. *What Do Unions Do: A Twenty-Year Perspective*. New Brunswick, NJ: Transaction Press.

Blair, M. and Stout, L. 1999. A team production theory of corporate law. *Virginia Law Review* 85, 247–328.

Budd, J. 2004. *Employment Relations with a Human Face*. Ithaca, NY: Cornell University/ILR Press.

Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Cappelli, P. 1999. Career jobs are dead. *California Management Review* 42, 146–67.

Clegg, H. 1970. *The System of Industrial Relations in Great Britain*. London: Blackwell.

Commons, J. 1909. American shoemakers, 1648–1895. *Quarterly Journal of Economics* 24, 39–98.

Commons, J. 1934. *Institutional Economics*. New York: Macmillan.

Doeringer, P. and Piore, M. 1972. *Internal Labor Markets and Manpower Analysis*. Lexington, MA: D. C. Heath.

Drago, R. 2007. *Striking a Balance*. Boston, MA: Dollars & Sense.

Dunlop, J. 1944. *Wage Determination under Trade Unions*. New York: Macmillan.

Dunlop, J.T. 1958. *Industrial Relations Systems*. New York: McGraw Hill.

Durkheim, E. 1893. *The Division of Labor in Society*. New York: Free Press, 1984.

Dyer, L. 1984. Studying human resource strategy: an approach and agenda. *Industrial Relations* 23, 155–69.

Fox, A. 1971. *A Sociology of Work and Industry*. London: Collier MacMillan.

Freeman, R. and Medoff, J. 1984. *What Do Unions Do?* New York: Basic Books.

Golden, C. and Parker, V. 1955. *Causes of Industrial Peace*. New York: Harper and Row.

Granovetter, M. 1974. *Getting a Job: A Study of Contacts and Careers*. Cambridge, MA: Harvard University Press.

Gunderson, M. 2007. Two faces of union voice in the public sector. In *What Do Unions Do: A Twenty-Year Perspective*, ed. J. Bennet and B. Kaufman. New Brunswick, NJ: Transaction Press.

Hughes, E. 1958. *Men and their Work*. Glencoe, IL: Free Press.

Hyman, R. 1975. *Industrial Relations: A Marxist Introduction*. London: Macmillan.

Ichniowski, C., Kochan, T., Levine, D., Olson, C. and Strauss, G. 1996. What works at work: overview and assessment? *Industrial Relations* 35, 299–333.

Jacoby, S. 1985. *Employing Bureaucracy*. New York: Columbia University Press.

Jacoby, S. 1991. *Masters to Managers*. New York: Columbia University Press.

Jacoby, S. 1999. Are career jobs headed for extinction? *California Management Review* 42, 123–45.

Katz, H. and Darbashire, O. 2000. *Converging Divergences*. Ithaca, NY: Cornell/ILR Press.

Kaufman, B. 1993. *The Origins and Evolution of the Field of Industrial Relations*. Ithaca, NY: Cornell University/ILR Press.

Kaufman, B. 2004. *The Global Evolution of Industrial Relations*. Geneva: International Labour Organization.

Kerr, C., Dunlop, J., Harbison, F. and Myers, C. 1960. *Industrialism and Industrial Man*. Cambridge, MA: Harvard University Press.

Kochan, T. 1980. *Industrial Relations: From Theory to Policy and Practice*. Homewood, IL: Irwin.

Kochan, T. 1995. Using the Dunlop Report for mutual gain. *Industrial Relations* 34, 350–66.

Kochan, T. and Helfman, D. 1981. The effects of collective bargaining on economic and behavioral job outcomes. In *Research in Labor Economics*, ed. R. Ehrenberg. Greenwich, CT: JAI Press.

Kochan, T., Katz, H. and McKersie, R. 1986. *The Transformation of American Industrial Relations*. New York: McGraw Hill.

Kossak, E. 2006. Work and family in America: growing tensions between employment policy and a

transformed workforce. In *America at Work*, ed. E. Lawler and J. O'Toole. New York: Palgrave Macmillan.

Kuhn, T. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lazear, E. 1998. *Personnel Economics for Managers*. New York: Wiley.

Lester, R. 1952. A range theory of wage differentials. *Industrial and Labor Relations Review* 5, 433–50.

Lewis, H. 1963. *Unions and Relative Wages in the United States*. Chicago: University of Chicago Press.

Levinson, H. 1968. Wage determination under collective bargaining. In *Collective Bargaining*, ed. A. Flanders. London: Palgrave Books.

Locke, R. 1995. *Remaking the Italian Economy*. Ithaca, NY: Cornell University Press.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Marx, K. 1849. *Wages, Labor Capital. Repr*. In *The Marx-Engels Reader*, ed. R. Tucker. New York: Norton, 1972.

Meltz, N. 1989. Industrial relations: balancing efficiency and equity. In *Theories and Concepts of Comparative Industrial Relations*, ed. J. Barbash and K. Barbash. Columbia, SC: University of South Carolina Press.

Milgrom, P. and Roberts, J. 1992. *Economics, Organizations, and Management*. Upper Saddle River, NJ: Prentice Hall.

Olson, C. 1988. Dispute resolution in the public sector. In *Public Sector Bargaining*, 2nd edn. ed. B. Aaron, J. Najita and J. Stern. Washington, DC: Bureau of National Affairs.

Osterman, P. 1984. *Internal Labor Markets*. Cambridge, MA: MIT Press.

Osterman, P., Kochan, T., Locke, R. and Piore, M. 2001. *Working in America: A Blueprint for the New Labor Market*. Cambridge, MA: MIT Press.

Pfeiffer, J. and Baron, J. 1988. Taking workers back out: recent trends in the structuring of employment. In *Research on Organizational Behavior*, vol. 10, ed. B. Staw and L. Cummings. Greenwich, CT: JAI Press.

Piore, M. and Sabel, C. 1984. *The Second Industrial Divide*. New York: Basic Books.

Portes, A. and Sensenbrenner, J. 1993. Embeddedness and immigration: notes on the social determinants of economic action. *American Journal of Sociology* 98, 1320–50.

Putnam, R. 1973. *The Beliefs of Politicians*. New Haven: Yale University Press.

Rees, A. and Shultz, G. 1970. *Workers and Wages in an Urban Labor Market*. Chicago: University of Chicago Press.

Roethlisberger, F. and Dickson, W. 1939. *Management and the Worker*. Cambridge, MA: Harvard University Press.

Ross, A. 1948. *Trade Union Wage Policy*. Berkeley: University of California Press.

Saxenian, A., Motoyama, Y., Quan, X. and Wittenborn, D. 2002. *Local and Global Networks of Immigrant Professionals in Silicon Valley*. San Francisco: Public Policy Institute of California.

Schuler, R. and Jackson, S. 1987. Linking competitive strategies with human resource management practices. *Academy of Management Executive* 1, 207–19.

Selznick, P. 1984. *Leadership in Administration*. Berkeley: University of California Press.

Slichter, S., Livernash, R. and Healy, J. 1960. *The Impact of Collective Bargaining on Management*. Washington, DC: Brookings Institution.

Taylor, F. 1895. A piece rate system, being a step toward partial solution of the labor problem. *Transactions* 16, 856–83.

Walton, R. and McKersie, R. 1965. *A Behavioral Theory of Labor Negotiations*. New York: McGraw Hill.

Webb, S. and Webb, B. 1894. *The History of Trade Unions*. London: Longmans, Green & Co.

Webb, S. and Webb, B. 1897. *Industrial Democracy*. London: Longmans, Green & Co.

Weber, M. 1962. *Basic Concepts in Sociology*. New York: Philosophical Library.

Wright, P.M. and McMahan, G.C. 1992. Theoretical perspectives for human resource management. *Journal of Management* 18, 295–320.

How to cite this article

Kochan, Thomas A. "industrial relations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000077> doi:10.1057/9780230226203.0784

The New Palgrave Dictionary of Economics Online

Industrial Revolution

Gregory Clark
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘Industrial Revolution’ has come to mean two very different things: first, the transformation the British economy experienced between 1760 and 1850, to become the first modern industrialized, fast-growing economy; second, the general switch between the pre-industrial world of slow technological advance, high fertility and little human capital to the modern world of rapid efficiency gains, low fertility and large investments in human capital. Modern economists’ theories of this second worldwide transition have proved difficult to reconcile with the details of Britain's transition.

Keywords

child mortality; demographic transition; education; Enlightenment; family planning; fertility; fertility–income relationship; human capital; Industrial Revolution; Innovation; Malthusian economy; population growth; real wage rates; skill premium; technical change; total factor productivity

Article

The Industrial Revolution is an ambiguous term, freighted with multiple meanings, interpreted differently by different writers. First, it describes the extraordinary transformation the British economy experienced between 1760 and 1850. In these years Britain moved from being a largely self-sufficient, self-sustaining, and still principally agrarian society, to being an economy where a substantial fraction of food, raw materials and energy was imported, or mined from the earth as coal, and where the great majority of the population was engaged in industry and commerce. But second, and more importantly, it has come to mean the general move in the world economy in about 1800 from the pre-industrial economy, which experienced extremely low rates of efficiency growth, to the modern economy, where efficiency growth is rapid and persistent. That shift from low rates of efficiency advance to rapid rates had nothing inherently to do with industry or industrialization. Efficiency advance in agriculture has been as rapid as in the rest of the economy since 1800. So for the more general use of the term ‘Industrial Revolution’ the ‘industrial’ component is a misnomer, but a misnomer that we have to live with.

The Industrial Revolution of the historians

The ‘Industrial Revolution’ more traditionally describes a specific period in British history, most commonly taken as 1760 to 1850. In 1760 Britain was a prosperous but still heavily agrarian economy, with half the labour force employed in agriculture. Foreign trade was insubstantial. Britain was largely self-sufficient in staple foods. The main imports were Mediterranean or tropical products such as sugar and spices, wines, raisins, coffee and tea. The main export was woollen cloth produced by domestic weavers or handloom workshops. London was already a huge city with over 750,000 inhabitants, but the other towns in England circa 1760 were mostly small. The next biggest city was Bristol with only 50,000 people. Travel and communication were slow and costly. The road system was poorly maintained, and there were few canals.
By 1850 the share of the population employed in agriculture in Britain had dropped to less than a quarter. Staple foods and raw materials such as timber had become major imports. Exports were dominated by factory-produced textiles, but included a whole range of manufactured goods and even substantial amounts of coal. The urban population had grown enormously. Manchester, for example, had grown from about 20,000 in 1770 to over 300,000 by 1851. London had nearly 2.4 million by 1851, more than 13 per cent of English

people, and was the largest city in the world. The road system had greatly improved, and alongside the roads there were now about 2,000 miles of canals and improved river navigations, as well as more than 5,000 miles of the new railways.

Rapid population growth accompanied the change in occupational structure, location and trade patterns. The English population grew from seven million in the 1770s to 19 million by the 1850s. Periods of population growth earlier in English history, as in the 13th and the 16th centuries, were associated with declining living standards. The Industrial Revolution represented a sharp break with this past. For the first time living standards improved even as the population swelled. Figure 1 shows the real wage of building workers vis-à-vis the English population from 1250 to 1850. The unusual character of experience in the Industrial Revolution era is clear.

Figure 1

Real building workers' day wages vis-à-vis population by decade, 1280–1849. *Note*: The line summarizing the trade-off between population and real wages for the pre-industrial era is fitted using the data from 1280–9 to 1590–9. *Source*: Clark (2005b, Figure 5).

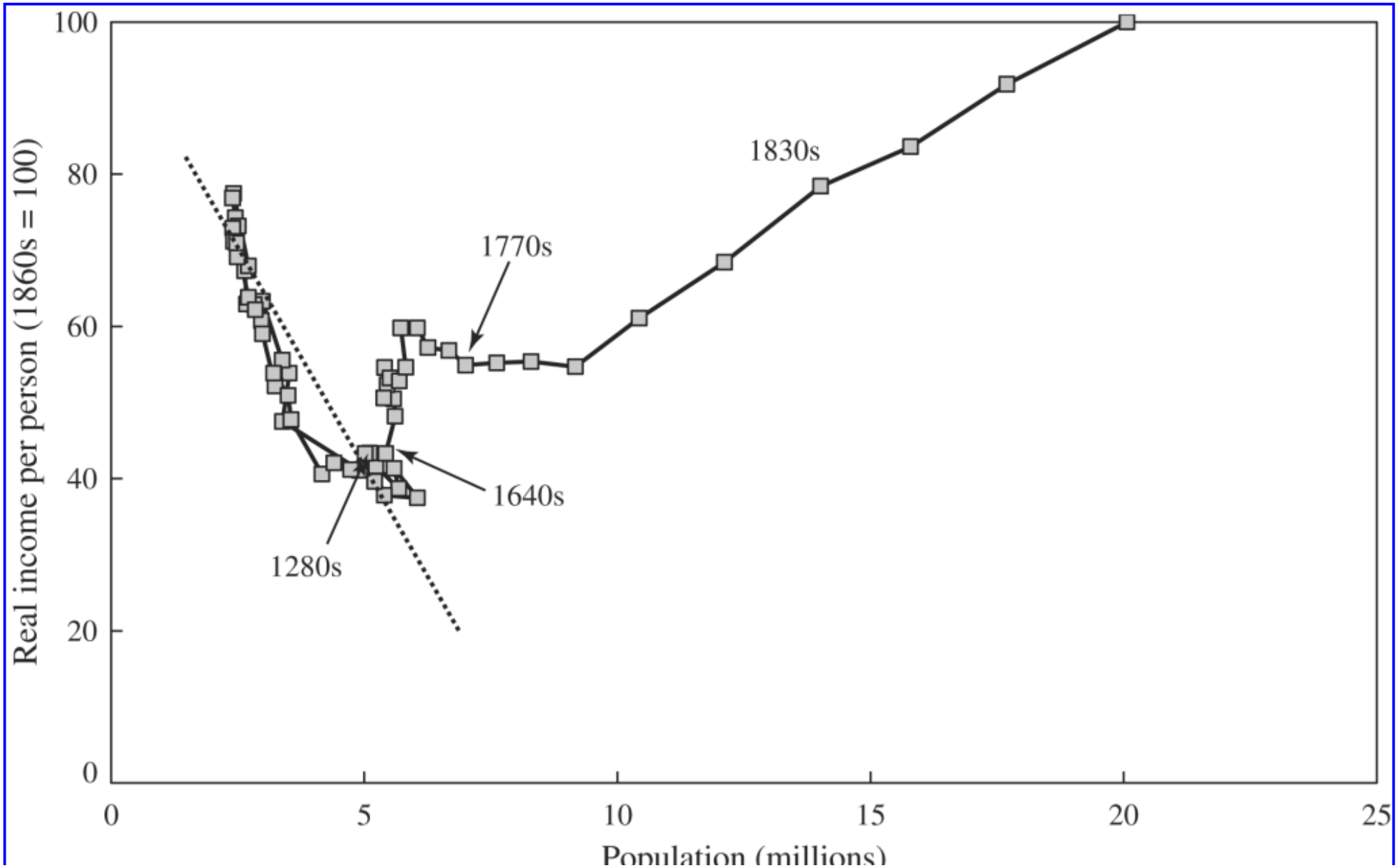

Between 1760 and 1850 England experienced what was cumulatively profound economic change, though the actual rate of change for most measures of the economy such as gross output per person or the fraction of the population employed in agriculture was by modern standards very slow. Indeed, the changes were so slow that many economists writing in this period – such as Adam Smith, Thomas Malthus and David Ricardo – had little comprehension of the fundamental break from the past that was occurring.

The recent consensus has been that the immediate cause of the Industrial Revolution was the dramatic increase in efficiency in a minority of the economy: yarn and cloth production, iron and steel making, and rail transport. Most of the economy, including surprisingly the coal industry, saw little technological advance (Clark and Jacks, 2007). Textiles alone explain perhaps 60 per cent of all measured technological advance from 1760 to 1850. The concentration of technological advance in textiles, aligned with the move of production there into factories, explains why the general move around 1800 towards economies with faster technological advance came to be labelled the ‘Industrial Revolution’.

In textiles we see a whole series of innovations, especially from the 1760s onwards, which transformed the industry. These innovations had no direct connection with the scientific advances of the previous 150 years and were indeed mainly made by artisans and craftsmen with no formal scientific training. Nor were the new production processes in these industries particularly capital-using. Water and steam powered textile mills were modest in their capital requirements compared with later innovations like the railways, but also compared with existing industries like agriculture. The demands of these mills were mainly for unskilled labour. Tending the new spinning and weaving machines did not require literacy, and involved skills fully mastered within a year of employment. Thus the Industrial Revolution in the first instance did not involve great investments in either physical or human capital.

The question of why England first experienced the Industrial Revolution, and why only in the 1760s, has occupied the energies of an enormous number of historians and economists. There has been an intense debate on the features of the British economy in 1760 that precipitated the break from the past. Generations of economic historians have thrown themselves at the problem, like waves of infantry in the First World War going over the top of the trenches. Their explanations, however, have generally fared no better than the average First World War soldier when tested against the history of England in these years.

Putative explanations of the Industrial Revolution can be separated into those based on the supply of or the demand for innovations, as portrayed in Figure 2. Some emphasize greater returns to innovation as inducing more innovation, others a greater supply of innovators.

Figure 2

Demand and supply interpretations of the Industrial Revolution

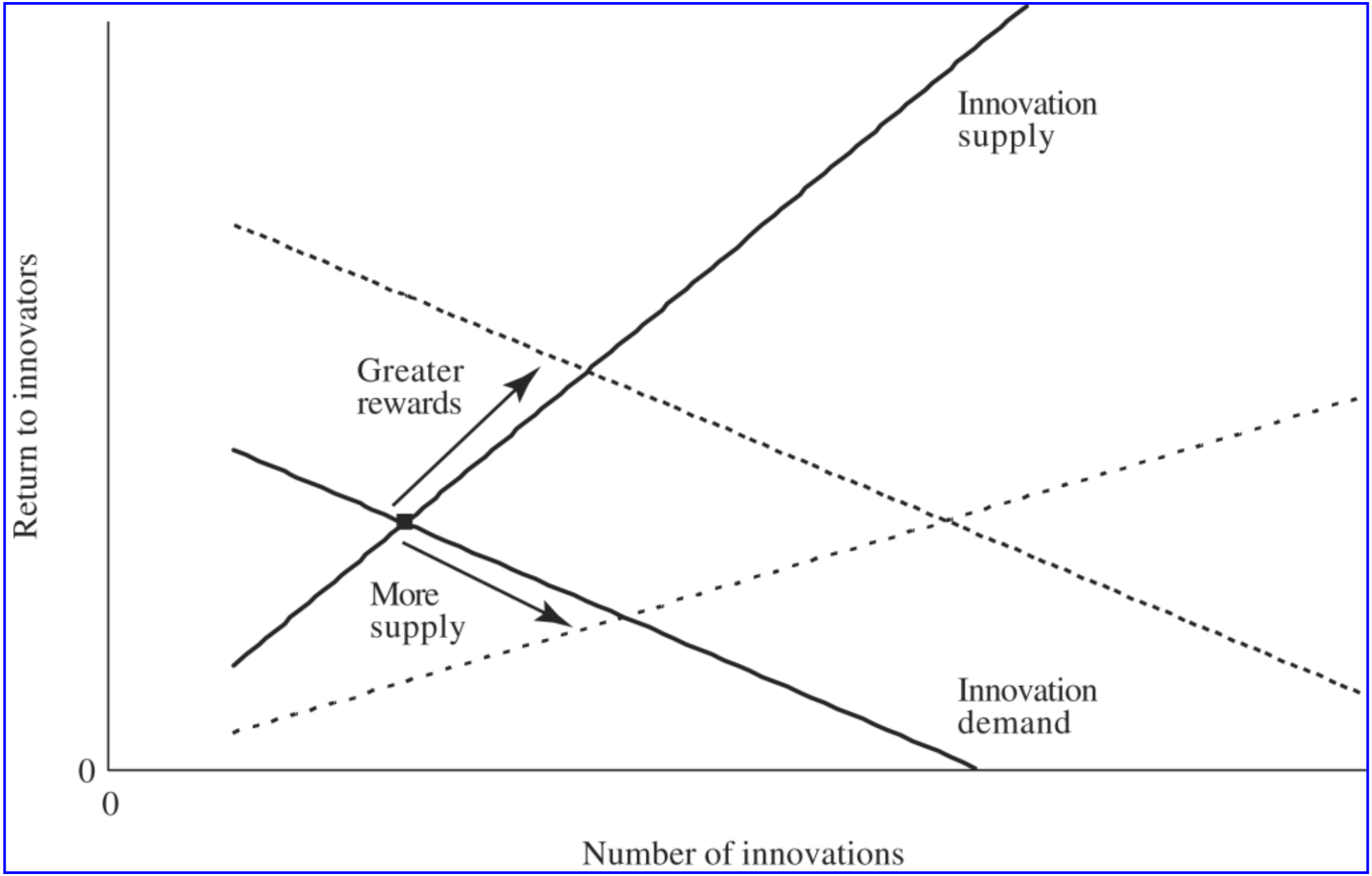

Much attention has been given, for example, to the institutional changes that preceded the English Industrial Revolution, and raised the benefits to innovation. Douglass North and Barry Weingast proclaimed the Glorious Revolution of 1688–9, which established the institutional framework of the modern British state with a figurehead monarch and control by an elected parliament, as the key precondition for economic growth (North and Weingast, 1989). The development of a government restrained from seizing the profits of investors increased the expected returns to investment in general in the economy.

There are numerous problems with this identification. The gap between the institutional changes and the onset of the Industrial Revolution is a generous 80 years or so. In those 80 years there were was no speed up in the rate of efficiency advance in the economy, as Figure 3 shows. The efficiency of the economy, known also as the total factor productivity (TFP), is the amount of output delivered per unit of input of capital, labour and land. From 1689 to 1760 the English economy had efficiency growth rates no faster than those of the 'bad' days of the old regime in 1600–89, when England experienced considerable political turmoil.

Figure 3
Efficiency level of the English economy, 1600–1860. *Notes*: The figure shows the estimated efficiency of the English economy by year (dotted line) and as an 11-year moving average (solid line). *Source*: Clark (2007, Figure 12.6).

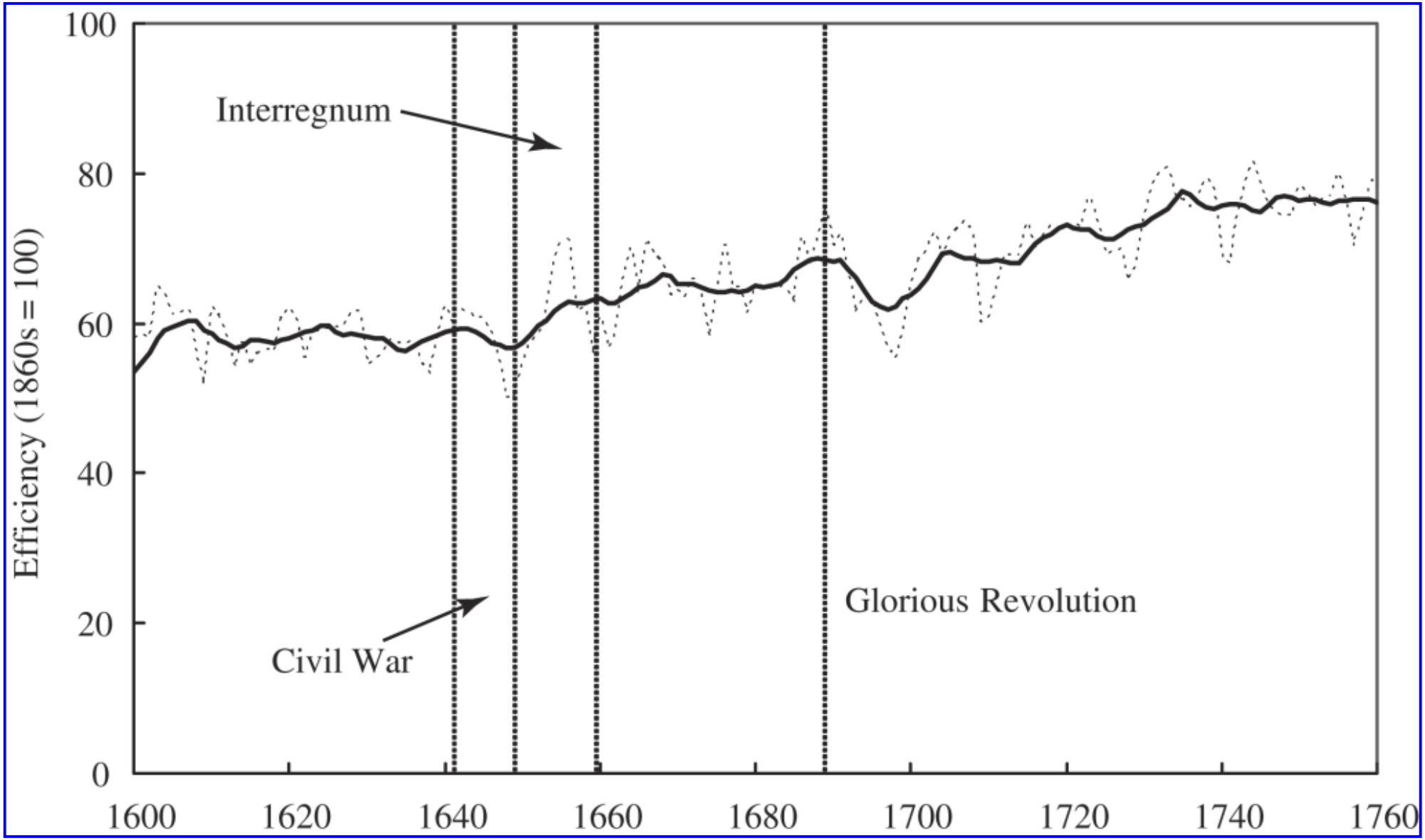

Also, contemporary economic actors seem to have attached no importance to the political changes of 1688–9. Gross rates of return on capital in the private economy, for example, did not decline, as would be expected if the new regime had ushered in more secure property rights (Clark, 1996). Finally, societies such as that of England had most of the institutional prerequisites of modern growth – stable politics, free markets, factor mobility, and low taxation – hundreds of years before any growth appeared (Clark, 2007, ch. 8).

Kenneth Pomeranz has argued that the Industrial Revolution was triggered in England in the 1760s, and not in other sophisticated societies such as China, because of the accidents of coal and colonies (Pomeranz, 2000). The chance location of coal fields in England, and the ability of North America to supply massive imports of raw materials liberated England from the energy and raw material constraints that had limited growth before in the self-sustaining organic pre-industrial economy. But the concentration of growth in cotton textiles, an industry that was present also in Japan and China by 1800, where water power could supply all the energy required, suggests that the elements Pomeranz concentrates on were

actually peripheral to the Industrial Revolution (Clark and Jacks, 2007).

Other economists, such as Joel Mokyr, have argued alternately that the root cause of the Industrial Revolution was an increased supply of innovation, promoted by the Enlightenment, the intellectual movement which swept Europe in the 18th century (Mokyr, 2005). Mokyr shows that while the Enlightenment was an important intellectual movement in many European countries such as France it was a particularly prominent part of intellectual life in England. And if we look at many other measures – literacy, numeracy, publications – England was becoming a more intellectually sophisticated society in the years leading up to the Industrial Revolution at all levels of the society. But Mokyr offers no account of why this intellectual movement should have taken hold in England in particular, and only in the 18th century.

The Industrial Revolution of the economists

From a broader perspective, the Industrial Revolution that brought us from the static pre-industrial economy to the modern dynamic economy is characterized by a three key features. Most important is the appearance of persistent total factor productivity growth. Such growth occurs when output rises faster than the measured inputs. Thus if *y* is output per worker hour, *k* capital services per worker hour, and *z* land services per worker hour, and *A* the level of efficiency (TFP) of the economy, A grows at the rate

$$g_A = g_y - a \cdot g_k - c \cdot g_z$$

where *g* denotes a growth rate, and *a* and *c* are the shares of capital and land in total factor costs. Since 1850 in the most successful economies TFP has grown at one per cent or more per year. Before 1800, over extended periods, even for successful economies TFP grew at rates of 0.01–0.1 per cent per year.

We can estimate TFP growth before 1800 using population. On average before 1800 output per worker-hour, *y*, did not rise (see the Malthusian economy). In this case we can simplify the equation above. In such a static economy, labour hours *L* will be proportionate to population *N*. Since the land area is fixed

$$g_z = -g_L = -g_N.$$

Similarly income per capita was constant over the long run. On the assumption that the rate of return on capital did not change, capital per person would have been constant, so that,

$$g_k = 0.$$

Substituting both these relations into the basic equation above implies that for the pre-industrial world the growth rate of efficiency over the long run was just

$$g_A = c \cdot g_N.$$

Thus long-run technological advance at a world scale before 1800 is proportionate to long-run population growth, as Kremer (1993) pointed out. Since plagues or disorder can result in wages departing from the long-run equilibrium, this calculation serves only for the long run. Table 1 shows the details. For the world as a whole there is no long period before 1700 when the rate of technological advance even exceeds 0.1 per cent per year.

Growth rate of world population and TFP before 1800

Year	Population (millions)	Population growth rate (%)	Technology growth rate (%)
130,000 bc	0.1	–	–
10,000 bc	7	0.004	0.001
1 ad	300	0.038	0.009
1000 ad	310	0.003	0.001
1250 ad	400	0.102	0.025
1500 ad	490	0.081	0.020
1750 ad	770	0.181	0.045

Source: Clark (2007, Table 7.1).

The second general feature of the broader Industrial Revolution has been declining fertility, measured as births per woman. English women, for example, average five births each all the way from the 1540s to the 1890s. Figure 4 shows the gross reproduction rate (GRR), the number of daughters born per woman living to the age of 50, by decade in England from the 1540s to 1990s. The ‘demographic transition’ to modern fertility rates in Europe and North America, except for France, began only in the 1880s. By 2000 English women gave birth on average to fewer than two children.

Figure 4

English fertility history, 1540–2000. *Notes*: GRR=gross reproduction rate. NRR=net reproduction rate. The data for the years after 1837 is for the whole population. Before 1837 it is from a sample of parishes. *Sources*: Clark (2005a, Figure 2).

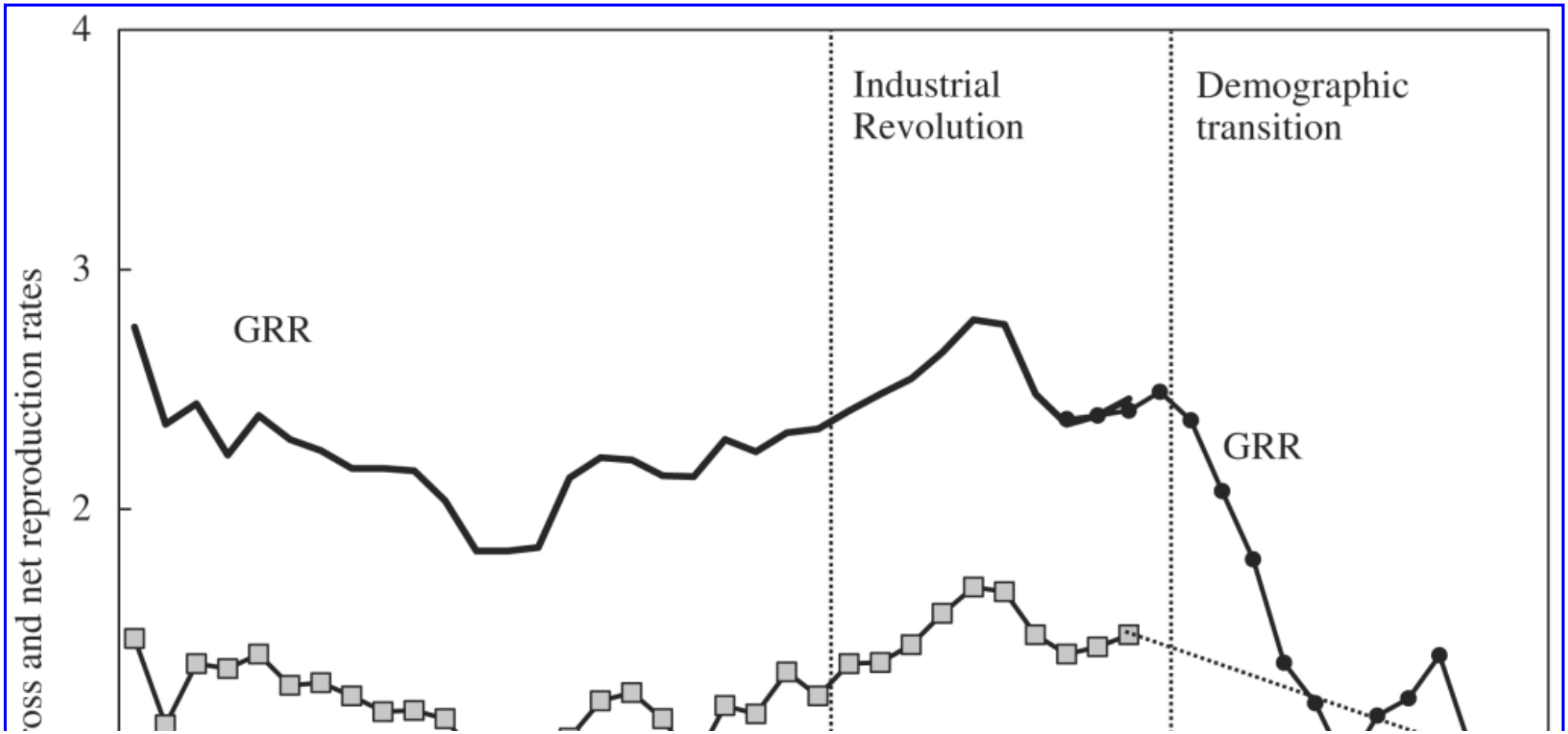

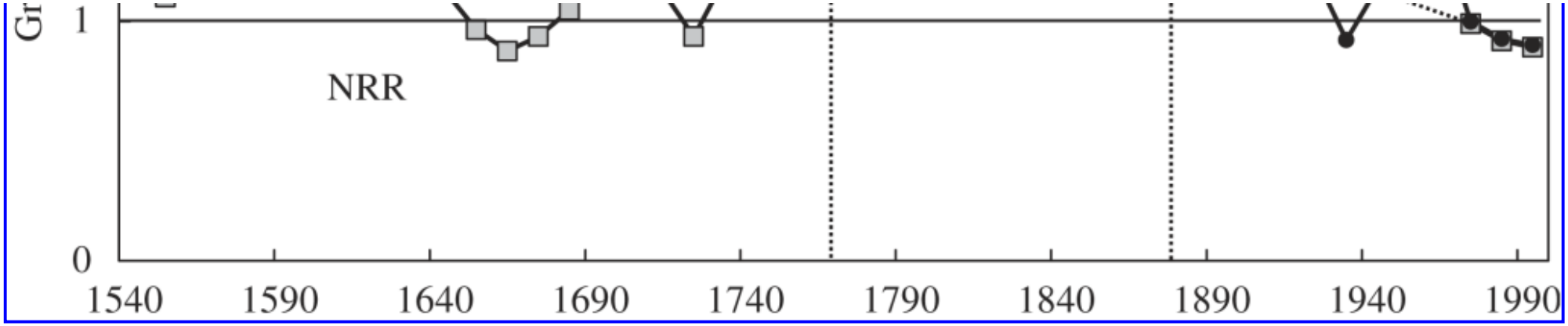

Since pre-industrial child mortality rates were high, however, the net reproduction rate (NRR), the number of daughters the average woman gave birth to over her lifetime, fell much less in the modern world than in the pre-industrial era. Figure 4 shows also the NRR for England. England in 1540–1800 had an unusually high NRR for pre-industrial society, where this number would normally be just slightly above 1. Note that the GRR and NRR both rose in England in the course of the classic Industrial Revolution.

The decline in gross fertility after the 1880s was crucial in allowing enhanced efficiency in the economy to translate into higher incomes. Had this not happened, so that population growth would have been much more rapid, then the share of payments to land as a factor, *c*, would not have declined so rapidly and might even have increased. Then in the first equation above the increase of population per acre would have been faster, and its weight greater, leading to a greater drag on income growth.

The third key feature of the transition to the modern world has been an increase in human capital per person, investments in education and training. In most pre-industrial societies the mass of the population was illiterate and innumerate. Along with the Industrial Revolution came a transition to a society where the implied value of human capital is nearly as great as for physical capital.

English education levels increased over the Industrial Revolution years. Figure 5 shows a measure of basic literacy, the fraction of men and women signing their names on witness statements or marriage registers. However, if one compares Figure 5 with Figure 4 there appears to be no connection between changes in literacy rates and changes in fertility: the fertility transition in England occurred after the attainment of mass literacy.

Figure 5

Literacy in England, 1580–1920. *Source*: Clark (2005a, Figure 3).

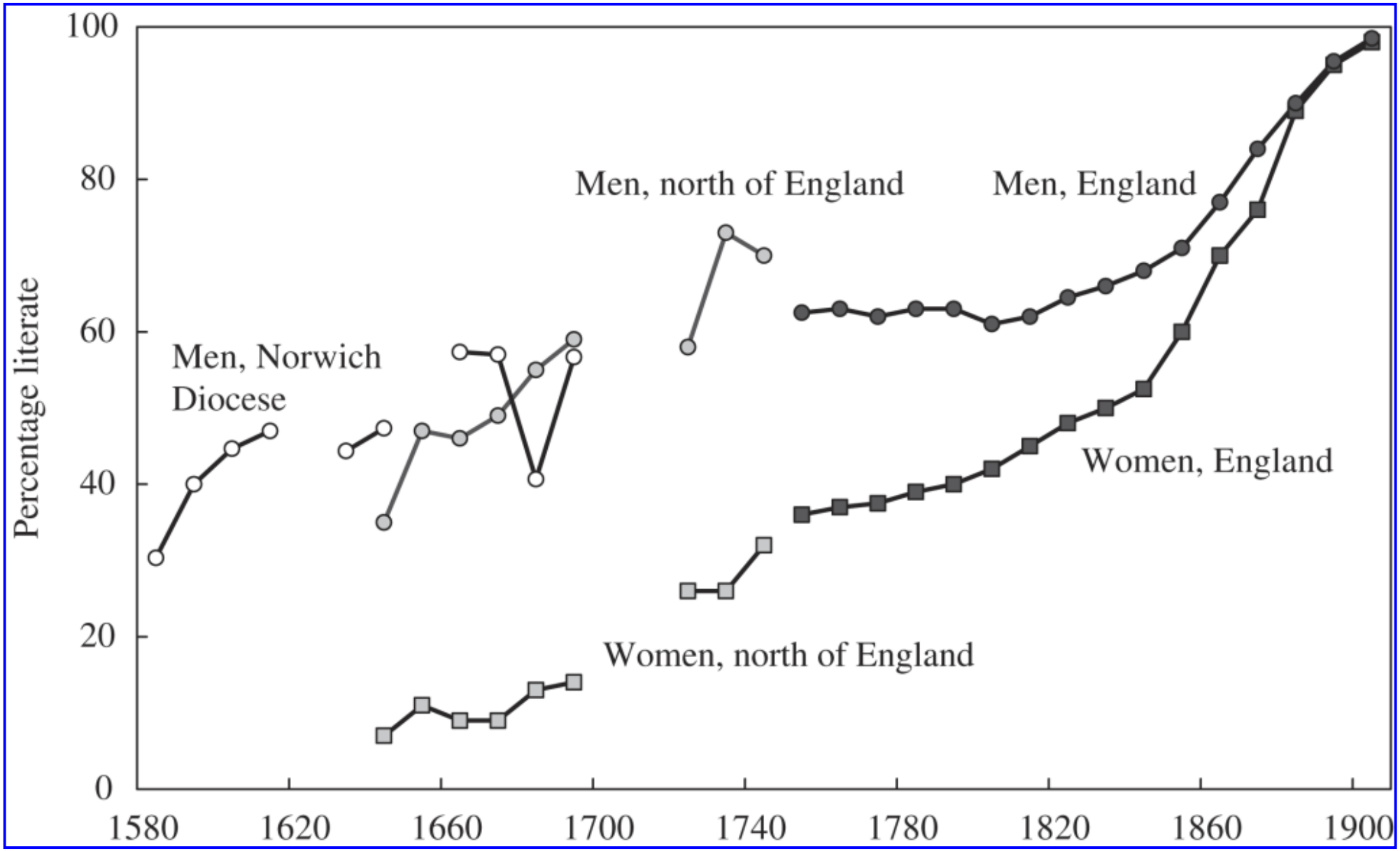

The coincidence of these three great changes in societies – technological advance, declining birth rates and increased education – has led economists in recent years to attempt theories of the broader Industrial Revolution that unify these elements (Becker, Murphy and Tamura, 1990; Galor, 2005; Galor and Moav, 2002; Galor and Weil, 2000; Lucas, 2002). These theories, however, face formidable obstacles in reconciling themselves to the facts of the Industrial Revolution in England.

One method of unification would posit the technological advances as primary, and have the income gains from these spur both lower fertility and more investment in human capital. In the years of the demographic transition in both the USA and in Europe between 1880 and 1920, higher-income families were the first to reduce fertility (Clark, 2007, Table 14.5; Jones and Tertilt, 2006, pp. 23–7). Indeed, Larry Jones and Michele Tertilt conclude that, for female birth cohorts in the USA between 1828 and 1958, income explains most of the decline in gross fertility. Figure 6, for example, shows the hourly real wage of building workers in England from 1200 to 2000. After the 1860s real wages begin to rise rapidly, and after the 1860s fertility declined substantially. In the modern world there is a strong negative fertility–income relationship across countries.

Figure 6

Real day wages of English building workers, 1200–2000. *Source*: Clark (2005b, Figure 1).

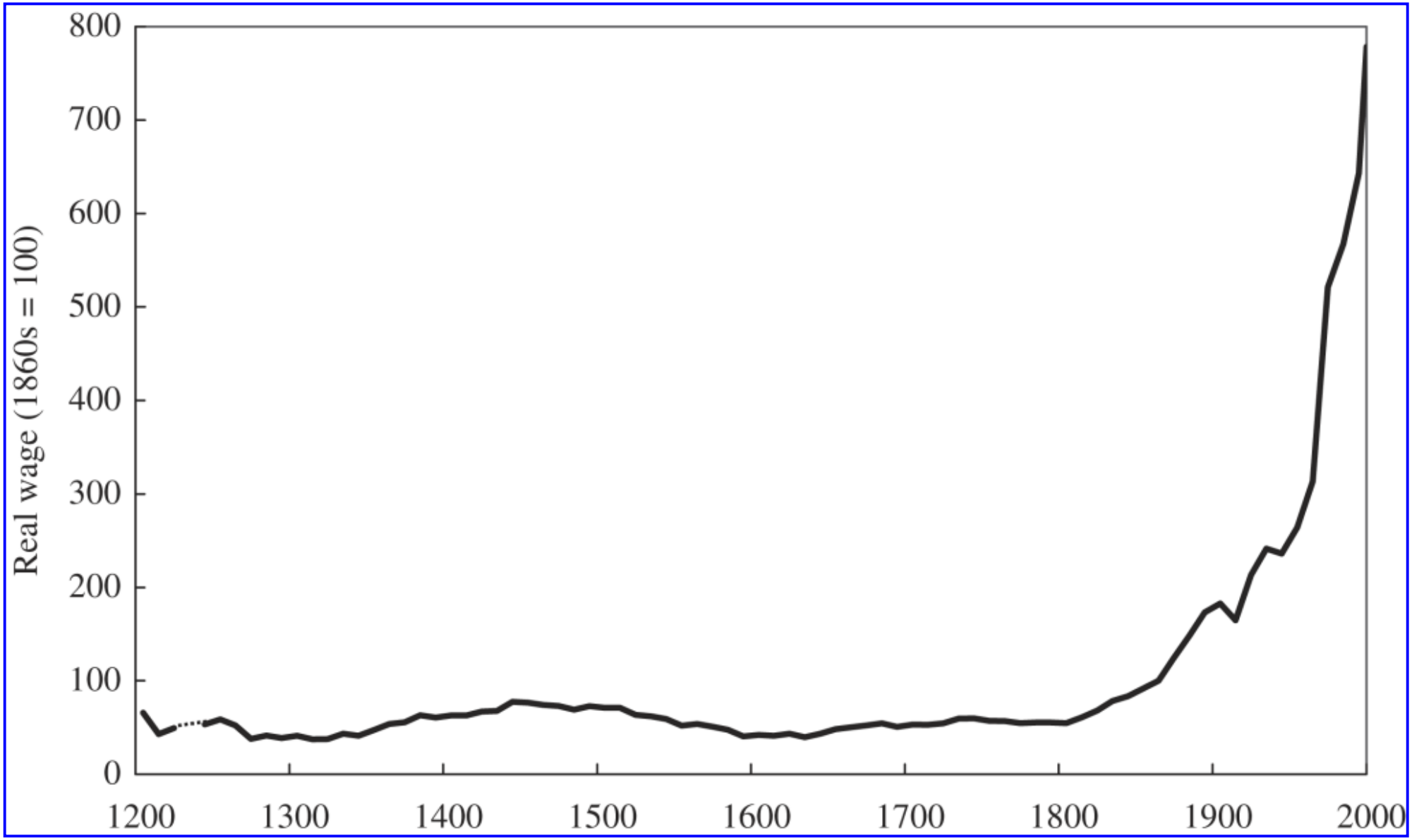

The problem with explaining the fertility transition through income is that all plausible models of population regulation before 1800 depend on a *positive* association between fertility and income. Empirical information on pre-industrial fertility and income is rare. But in pre-industrial England we get an insight into the connection through evidence from the wills of male testators (Clark and Hamilton, 2006). Connecting information on assets at death to parish records reveals the average numbers of births per testator for each bequest class. Figure 7 shows that a man leaving less than £25 at death would typically father fewer than four children, while one with assets of more than £1,000, six children. Thus in pre-industrial England there was a positive association between income and both gross and net fertility over a wide range of incomes. This stands in sharp contrast to the association in the modern world.

Figure 7

Births by assets of testator, 1585–1636. *Source*: Clark (2007, Figure 4.3).

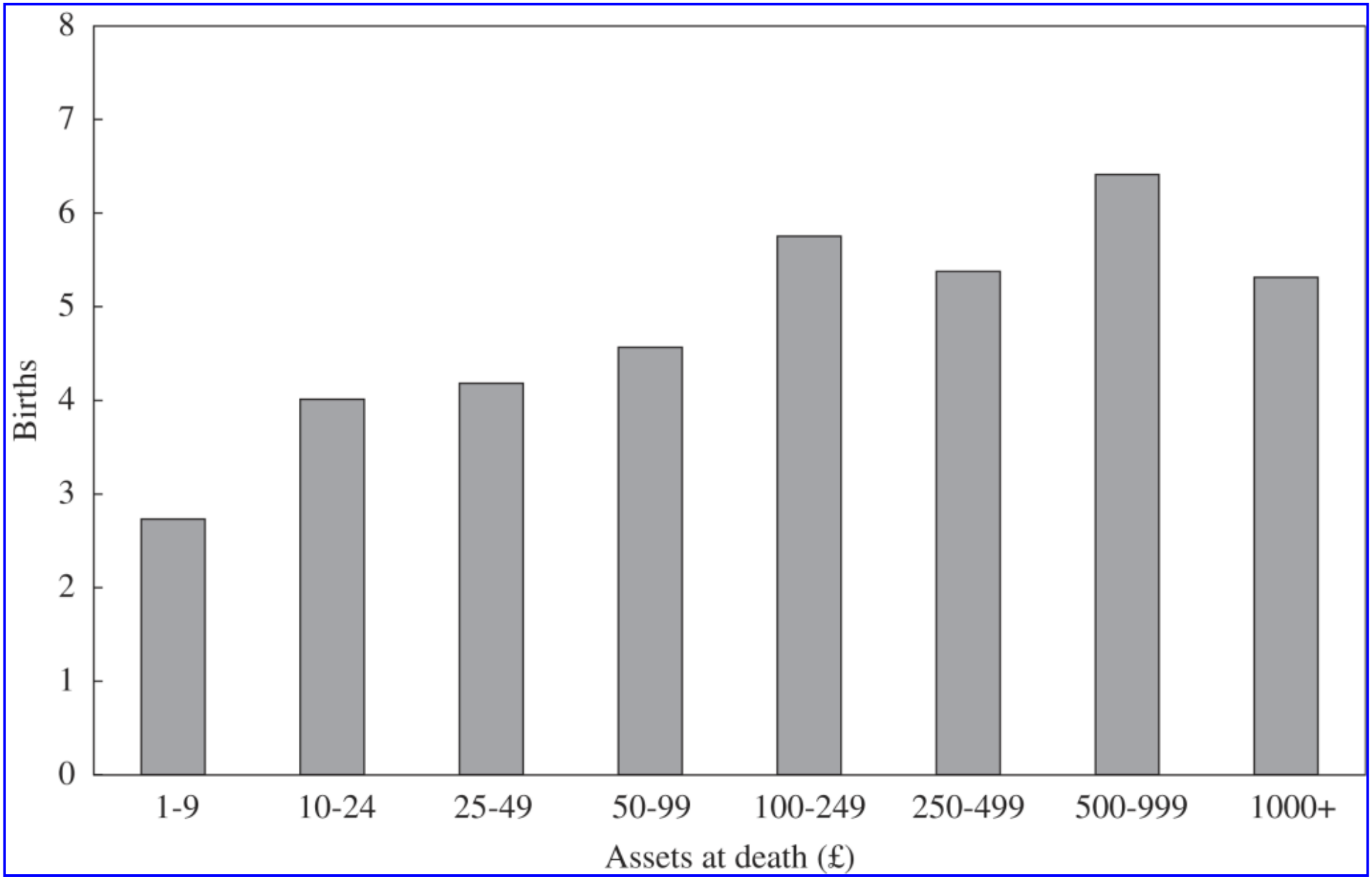

This positive association between fertility and income became negative in the period of demographic transition. But in current high-income, low-fertility societies there seems to be only the most modest negative association between income and fertility. A recent study of female fertility found on average little association between household income and fertility, measured as the numbers of children present in the households of married women aged 30–42, for 1980 and 2000, for the six Organisation for Economic Cooperation and Development (OECD) countries (Dickmann, 2003, Table 2). The income–fertility relationship within societies has changed dramatically over time.

All this makes constructing a link between fertility and income challenging. Why does fertility increase with income in the pre-industrial world? Authors who have addressed this have concentrated on explaining the association for incomes close to subsistence level. Galor and Weil (2000) and Galor and Moav (2002) assume a minimum consumption level that parents must achieve before producing children. Lucas (2002) assumes children require a minimum consumption transfer. We see in Figure 7, however, that the richest families in pre-

industrial England, people who would have high incomes even by the standards of 1900 showed high gross fertility rates.

The third problem with using income to explain declining family size is that, as Figure 6 shows, we cannot explain rising human capital in the years prior to Industrial Revolution through income gains. Human capital gains preceded the income gains of the Industrial Revolution. Finally, as noted above, we still lack any institutional or other explanation for the transition towards higher rates of efficiency advance after 1800.

Another mechanism that might explain both the rise in human capital and the decline in fertility and the Industrial Revolution would be an increase in the premium paid for human capital in the Industrial Revolution era. In most settled pre-industrial economies the bulk of labour demand was for agricultural work, where levels of human capital were low. In such an economy, it is argued, parents would favour quantity over ‘quality’ in children.

However, for this explanation is to be compatible with individual incentives, the return from investments in human capital before the Industrial Revolution has to be low. In England, and in a variety of other pre-industrial economies, rewards to human capital were higher than in the modern economy. We have, for example, the skill premium in the building industry: the ratio of the wages of craftsmen to building labourers. Figure 8 shows the wages of craftsmen relative to labourers in England by decade from 1200. The period 1600–1900, when literacy rates increased markedly, featured a near constant skill premium. When fertility rates fell after 1800 it was in a labour market where the premium for skills was also declining markedly. Thus gross fertility is highest where the premium for skills in the labour market is greatest. A demand interpretation of fertility decline, on its own, will not work either in England or as a general explanation of the fertility transition.

Figure 8

The skill premium for building workers, 1200–2000. *Source*: Clark (2005b, Figure 2).

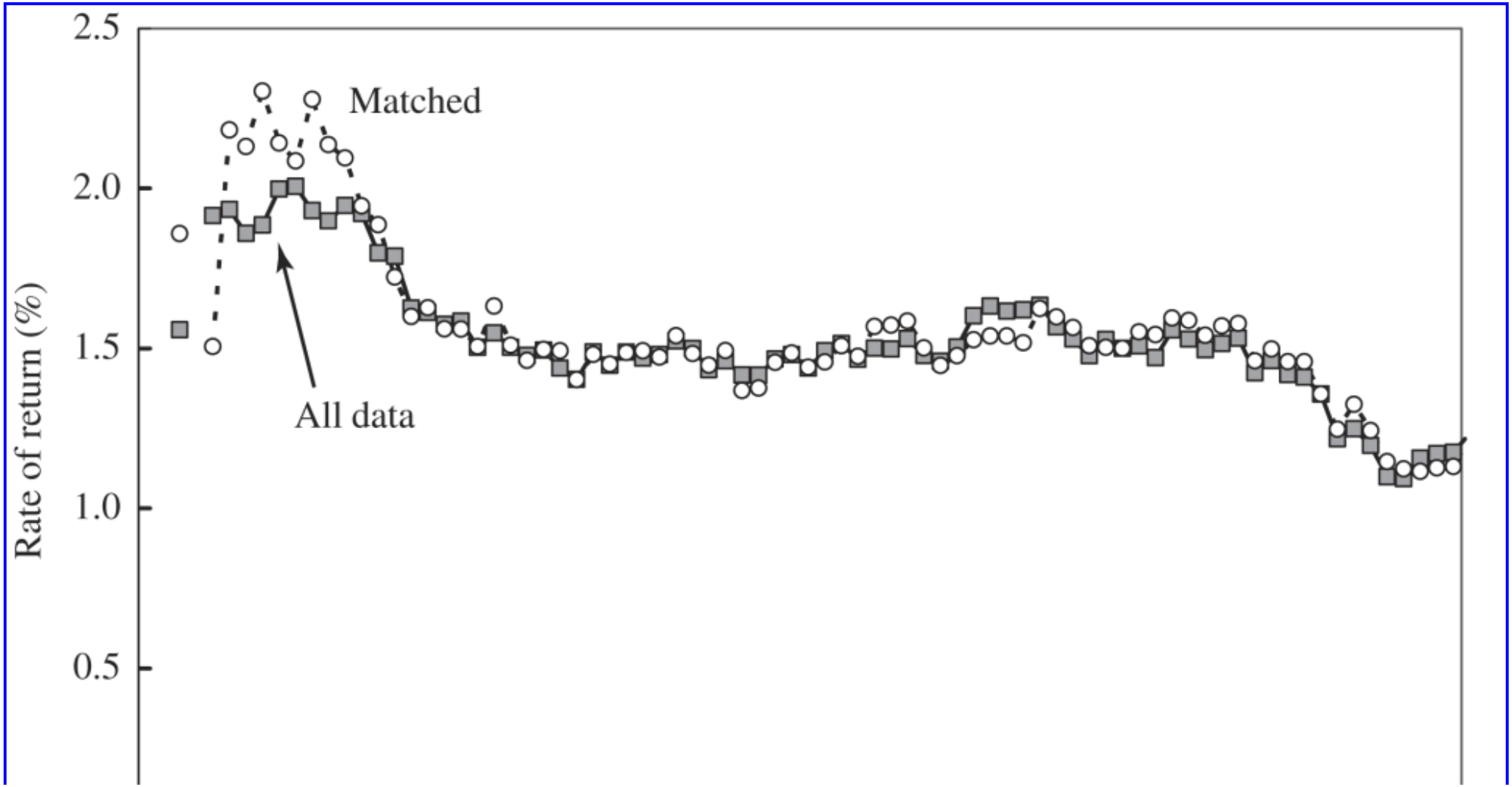

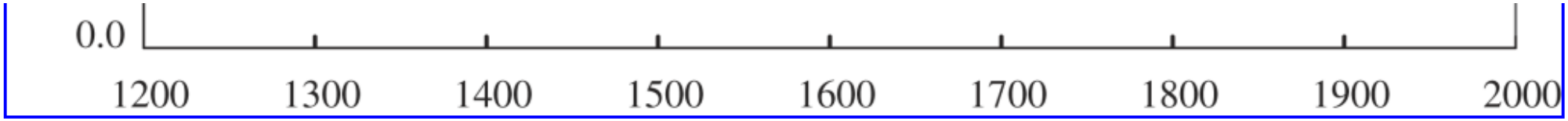

Since the expansion of human capital first occurred when the return to human capital was constant, the gains of human capital in the Industrial Revolution era had to involve significant supply shifts. Galor and Moav (2002) posit that the supply shift was created by Darwinian competition in the pre-industrial economy between families with different tastes for child ‘quality’. In the Malthusian world each family can have a NRR only slightly above 1. But ‘high-quality’ families do better. High-quality types produce offspring who, because of their greater human capital and hence higher incomes, have more children. Thus, when incomes are close to subsistence, but only then, they out-produce the ‘low-quality’ types. There should be an inverse U-shape of fertility with income. Figure 7, however, is inconsistent with this proposed mechanism. Even the richest in pre-industrial England show the highest gross and net fertility rates.

Clark (2007), however, argues that more general Darwinian selection mechanisms in the pre-industrial era could explain the move to more human capital, and the greater supply of innovations in the Industrial Revolution. Just as people were shaping economies, the economy of the pre-industrial era was shaping people, at the least culturally, perhaps also genetically. The Neolithic revolution created agrarian societies that were just as capital intensive as the modern world. At least in England, the emergence of such an institutionally stable, capital-intensive economic system created a society that rewarded middle-class values with reproductive success, generation after generation. This selection process was accompanied by changes in characteristics of the pre-industrial economy that owe much to the population displaying more middle-class preferences. Interest rates fell, murder rates declined, work hours increased, the taste for violence declined, and numeracy and literacy spread even to the lower reaches of society. These selection mechanisms thus provide an economic underpinning to the intellectual developments such as the Enlightenment of the 18th century that Mokyr identifies as a key background to the Industrial Revolution in England.

But such an explanation for the onset of the Industrial Revolution, which emphasizes the greater fertility of the rich in the pre-industrial era, leaves declining fertility after 1880 as a conundrum. If the economic system prior to the Industrial Revolution selected those with a tendency to use higher incomes to achieve greater net fertility, why did all this change in the 1880s? There are several possible explanations.

One is that the desired number of children per married couple is actually independent of income, and was always for just two or three surviving children. But to ensure a completed family size of even two children in the high-mortality environment of the Malthusian era required six or more births. For example, in pre-industrial England where 60 per cent of children died before adulthood, to ensure a 90 per cent chance of getting a surviving son would require giving birth to seven children. Nearly 40 per cent of the poorest married men leaving wills in 17th century England had no surviving son. Even among the richest married men nearly one-fifth left no son. The average rich man left four children because some families had large numbers of surviving children. Hence the absence of any sign of fertility control by richer families in pre-industrial England may stem largely from the uncertainties of child survival in the Malthusian era. This may have led to an unwillingness on the part of all families to limit births. As the fraction of children surviving increased in the late 19th century, even risk-averse families could afford to begin limiting births.

In the late 19th century child mortality in England had fallen substantially from the levels of the 18th century, and the rate of that decline was strongly correlated with income. For families living in homes with ten or more rooms only 13 per cent of children failed to reach the age of 15, while for those in one room still 47 per cent of children failed to reach that age (Clark, 2007, p. 00). Thus the lower gross fertility of high-income groups at the end of the 19th century translates into a more muted decline in net fertility. And these groups faced a substantially reduced variance in family size outcomes compared with low-income groups.

Another possible element in the decline of fertility since the Industrial Revolution is the increased social status of women. Men may well have had greater desire for children in pre-industrial society than women. Women, not men, bore the very real health risks of pregnancy, and did most of the work involved in bringing up the children. But typically men had a much more powerful position within the family. Thus women may always have desired smaller numbers of surviving children than men, but have been able to effect those desires only in the late 19th century.

Women's relative status and voice was clearly increasing in the late 19th century in England, when literacy rates for women had advanced to near equality with those of men. Women had gained access to universities by 1869, enhanced property rights within marriage by 1882, votes in local elections in 1894, and finally a vote in national elections in 1918. The gain in the relative status and voice of women proceeded most rapidly among higher-income groups.

These assumptions could explain why net fertility falls after the late 19th century – even though in cross section in the 16th century – and in 2000 there is either a positive connection between income and net fertility or no connection. They could also explain why the demographic transition appeared first in the higher socio-economic status groups, so that net fertility is negatively related to income in the transition period.

See Also

- historical demography
- Malthusian economy

Bibliography

Becker, G., Murphy, K. and Tamura, R. 1990. Human capital, fertility and economic growth. *Journal of Political Economy* 98, S12–S37.

Clark, G. 1996. The political foundations of modern economic growth: England, 1540–1800. *Journal of Interdisciplinary History* 26, 563–88.

Clark, G. 2005a. Human capital, fertility, and the industrial revolution. *Journal of the European Economic Association* 3, 505–15.

Clark, G. 2005b. The condition of the working-class in England, 1209–2004. *Journal of Political Economy* 113, 1307–40.

Clark, G. 2007. *A Farewell to Alms: A Brief Economic History of the World*. Princeton: Princeton University Press.

Clark, G. and Hamilton, G. 2006. Survival of the richest: the Malthusian mechanism in pre-industrial England. *Journal of Economic History* 66, 707–36.

Clark, G. and Jacks, D. 2007. Coal and the Industrial Revolution, 1700–1869. *European Review of Economic History* 11, 39–72.

Crafts, N.F.R. 1985. *British Economic Growth during the Industrial Revolution*. New York: Oxford University Press.

Crafts, N.F.R. and Harley, C.K. 1992. Output growth and the Industrial Revolution: a restatement of the Crafts–Harley view. *Economic History Review* 45, 703–30.

Dickmann, N. 2003. Fertility and family income on the move: an international comparison over 20 years. Maxwell School of Citizenship and Public Affairs, Syracuse University, Working Paper No. 360.

Galor, O. 2005. From stagnation to growth: unified growth theory. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Galor, O. and Moav, O. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117, 1133–91.

Galor, O. and Weil, D.N. 2000. Population, technology and growth: from Malthusian stagnation to the demographic transition and beyond. *American Economic Review* 90, 806–28.

Jones, L.E. and Tertilt, M. 2006. An economic history of fertility in the U.S. 1826–1960. Working Paper 12796. Cambridge, MA: NBER.

Kremer, M. 1993. Population growth and technological change: one million bc to 1990. *Quarterly Journal of Economics* 108, 681–716.

Lucas, R.E. 2002. The industrial revolution: past and future. In *Lectures on Economic Growth*, ed. R.E. Lucas. Cambridge, MA: Harvard University Press.

McCloskey, D.N. 1981. The industrial revolution 1780–1860: a survey. In *The Economic History of Britain since 1700*, vol. 1, ed. R. Floud and D. McCloskey. Cambridge: Cambridge University Press.

Mokyr, J. 1999. Editor's introduction: the new economic history and the Industrial Revolution. In *The British Industrial Revolution: An Economic Perspective*, 2nd edn., ed. J. Mokyr. Boulder, CO: Westview Press.

Mokyr, J. 2005. The intellectual origins of modern economic growth. *Journal of Economic History* 65, 285–351.

North, D.C. and Weingast, B. 1989. Constitutions and commitment: evolution of institutions governing public choice in seventeenth century England. *Journal of Economic History* 49, 803–32.

Pomeranz, K. 2000. *The Great Divergence: China, Europe, and the Making of the Modern World Economy*. Princeton: Princeton University Press.

How to cite this article

Clark, Gregory. "Industrial Revolution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000078> doi:10.1057/9780230226203.0785

The New Palgrave Dictionary of Economics Online

inequality (global)

Steve Dowrick
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The seemingly inexorable rise in global inequality from the early 19th century may have reached a plateau at the end of the 20th century, although there are disputes about the methodology underlying that conclusion. Increasing global inequality in the 20th century was driven largely by increasing income gaps between nations. Inequality within countries fell sharply at the beginning of the 20th century, rising slightly towards the end. The strong economic growth of the Chinese economy is tending to reduce global inequality as China moves up towards the middle of the income ladder.

Keywords

Balassa–Samuelson effect; catch-up; Geary–Khamis purchasing power parity; Gini index; inequality (global); inequality (measurement); inter-country inequality; intra-country inequality; Kuznets, S.; mean logarithmic deviation; national accounting; purchasing power parity; standard deviation of logarithmic income; Theil index; uneven development

Article

The seemingly inexorable rise in global inequality in the 19th and 20th centuries may have reached a plateau in the 1980s.

The causes and consequences of changing global inequality are a hotly contested area of economic research and debate. The intensity of the debate is in part due to the moral outrage felt by many at revelations such as those from the International Comparison Program (2007), henceforth ICP. The ICP 1996 data on average real expenditures per person reveal expenditure exceeding 1,000 dollars on the luxuries of alcoholic beverages, recreation and restaurant meals in each of the world's 20 richest countries, an amount that exceeds the total national income in each of the world's 12 poorest countries and exceeds total expenditure on food in each of the world's 70 poorest countries.

Income distribution estimates reveal that in the year 2000 more than one in ten of the world's population eked out a living around or below the World Bank's intermediate poverty line of two dollars per person per day, whilst the richest five per cent enjoyed incomes at or above 100 dollars per person per day. According to the World Bank (2006a), out of every 100 child born today, less than one child in the USA is expected to die before the age of five, but for children born in Mali, 24 children will not survive.

The extent of current global inequality far exceeds the inequalities of previous eras, apparently giving the lie to theories that the forces of global integration reduce inequality through factor-price equalizing trade, boosting demand for low-wage labour in the poorest countries, and through capital mobility, whereby global investment flows to the poorest and least capital-intensive countries, boosting labour productivity and real wages – although these observations must be tempered by the evidence that some aggregate measures of global inequality peaked towards the end of the 20th century and by the evidence of the highly successful catch-up growth of many East Asian economies in the second half of that century.

There are many problems in conceptualizing and measuring inequality: are we concerned with measured incomes, with consumption or with well-being? Is inequality measured across nations, across households or across individuals? What is the appropriate index of inequality to use? For the most part I will focus on inequalities in measured income based on national accounting conventions or on survey data. Rather than debate the merits of different indices of inequality, I report a range of commonly used measures – noting that many studies find that different indices tend to move in the same direction over time even if their levels differ. Towards the end of this article I consider some of the methodological problems.

Inequality over the centuries

Looking back to the year 1500, Angus Maddison (2003) has dared to publish estimates of average income levels – or, more precisely, real GDP per capita measured at 1990 international prices, which I refer to as ‘income’ for short. His estimates suggest that over the first three centuries global income rose very slowly – from 566 dollars per person in 1500 to 667 dollars in 1820. Over this period, national income levels did not differ by very much, most of the nations being less than 50 per cent above or below the world average. As world income growth began to accelerate through the 19th and 20th centuries, led first by the United Kingdom and then by the United States, income gaps began to widen. By the end of the 20th century the world's richest major nation, the United States, was more than 100 times richer than the world's poorest nation.

These broad trends in growth and inequality are illustrated in Figure 1, which displays average income levels across eight populous countries and regions at approximately 50-year intervals from 1500 to 2000. Averaging incomes across regions does of course understate the true extent of inter-country inequality, particularly in the case of Africa where the 2000 average of nearly 1,500 dollars disguises a maximum income of over 10,000 dollars in Mauritius and a minimum of just 218 dollars in Zaire.

Figure 1

Long-run development: real GDP per capita, 1500–2005. *Note*: Data for 1550, 1650 and 1760 have been interpolated. *Source*: Maddison (2003) extended to 2005 using World Bank (2006b).

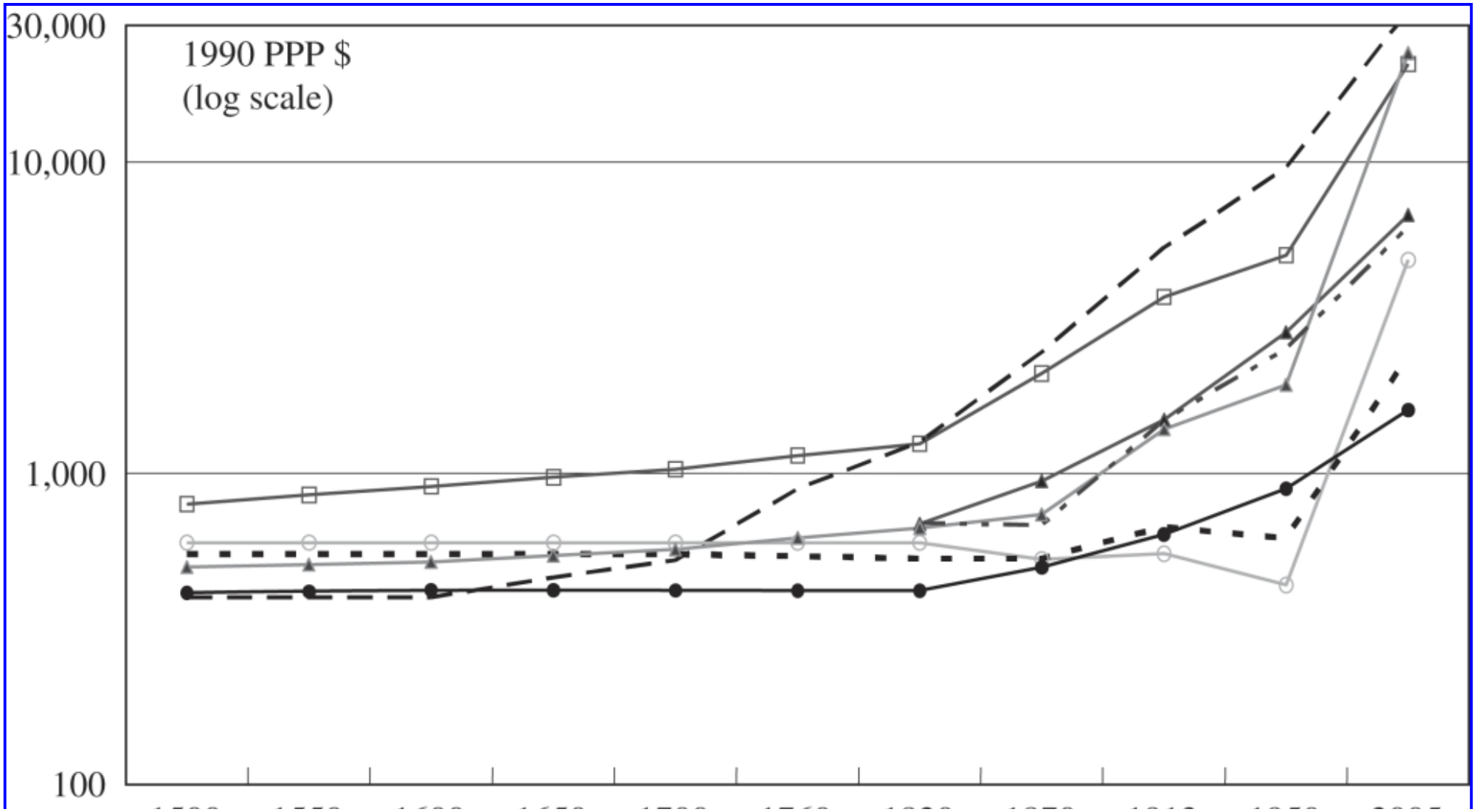

It is also the case that averaging incomes within countries disguises the true extent of inequality across individuals or households (or inequality by gender or ethnic groups). The paucity of historical data on income distribution within countries makes disaggregation below the national level an extremely difficult task for eras before the late 20th century. This task has, however, been attempted by François Bourguignon and Christian Morrisson (2002), who estimate global inequality across a group of 33 countries/country-groups reaching back to 1820 using historical income distribution data and extrapolating across countries judged to be similar. Their results are displayed as the four solid lines in Figure 2.

Figure 2

Inequality within and between 33 countries, 1820–1992. *Source*: Bourguignon and Morrison (2002).

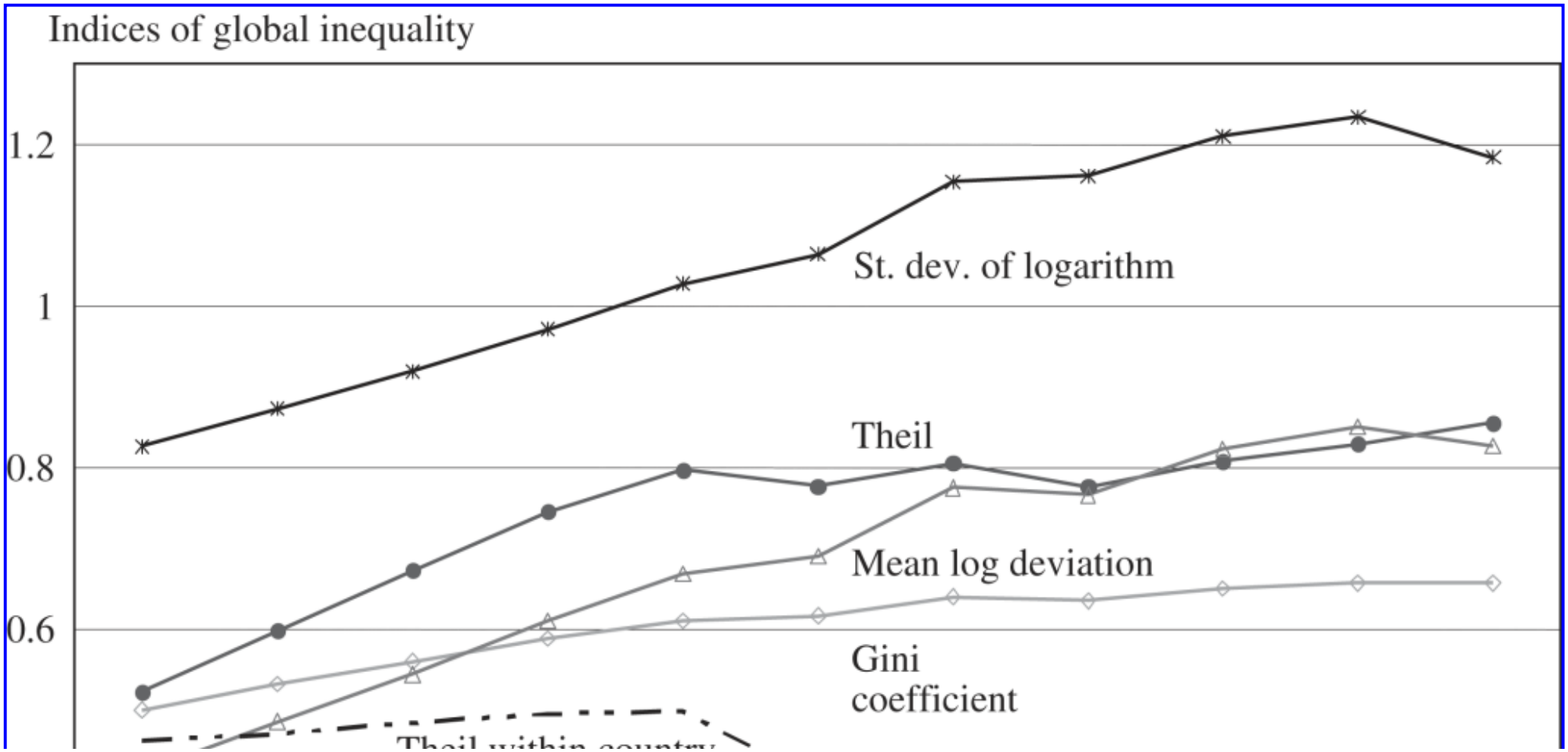

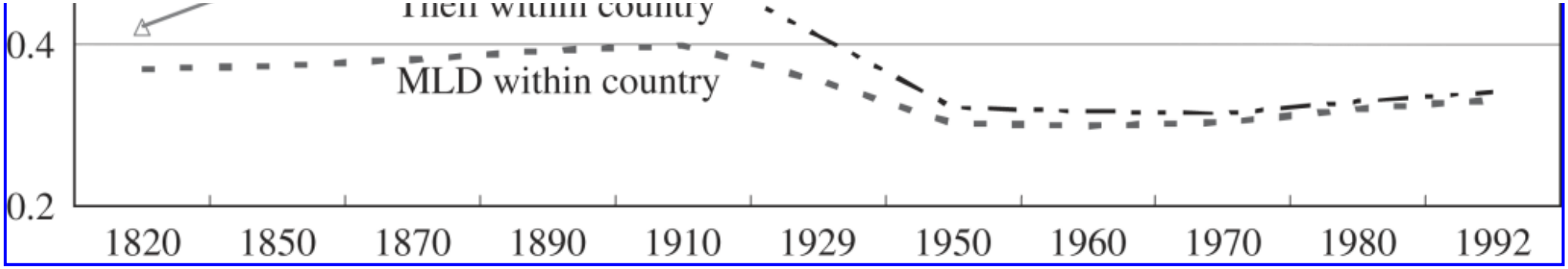

It is apparent that global income inequality rose strongly in the 19th century on all four of their measures: the Gini index, the Theil index, the mean logarithmic deviation (MLD) and the standard deviation of logarithmic income. Bourguignon and Morrisson's estimates indicate a slowing down in the rate of increase in inequality in the 20th century, although each measure displays slightly different trends. The Gini flattens out after 1970, both of the logarithmic measures peak in 1980, whilst the Theil measure is flat between 1910 and 1970 but rises up to 1992.

Both the Theil and the MLD can be decomposed exactly into the contributions of inequality within countries and inequality between countries. The within-country contributions to global inequality are shown as the dashed lines in Figure 2. It is apparent that within-country inequality was high and stable in the 19th century but fell substantially in the first half of the 20th century. On both measures, the contribution of within-country inequality to total inequality fell from nearly 90 per cent in 1820 to 40 per cent over the second half of the 20th century.

Global income inequality in the late 20th century

Data availability is far less of a problem for the second half of the 20th century than for previous eras (though problems of data definition and reliability persist) due to the publication of time series data on real GDP across most of the world's economies by Maddison (2003), by Robert Summers and Alan Heston (1991) and by Heston, Summers and Bettina Aten (2002) – the latter two studies producing successive versions of the Penn World Table. All these authors extrapolate over time and across countries from the benchmark price surveys, which are carried out periodically by the International Comparison Program.

Klaus Deininger and Lyn Squire (1996) have compiled sporadic time series on income distribution within countries – typically by decile or quintile groups. The gaps in their annual and country coverage have been filled by James K. Galbraith and Hyunsub Kum (2003), who extrapolate using data on wage inequality. Branko Milanovic (2002; 2005) has independently compiled a large number of national surveys of the distribution of income or expenditure at household level. I draw on a number of studies that have analysed global inequality using these sources of data.

A majority of these studies concludes that global income inequality peaked in the 1970s or 1980s and has subsequently declined slightly. The majority position has been challenged by Milanovic (2002; 2005) who uses household income surveys and World Bank estimates of current purchasing power of currencies to show that global inequality rose in the 1990s, in contradiction to Xavier Sala-i-Martin (2006) who demonstrates falling global inequality over the same period. Sala-i-Martin's methodology differs from that of Milanovic in that he uses the Deininger and Squire data on within-country inequality and converts currencies using the constant price estimates of purchasing power parity from the Penn World Table.

The majority position is also contested by Steve Dowrick and Mohammed Akmal (2005), who show that the Penn World Table's method of measuring real GDP at constant prices is subject to time-varying substitution bias, which understates the true level of inequality across countries. The evidence on this debate from Bourguignon and Morrisson (2002) is equivocal since two of their measures of global inequality, the standard deviation and the mean deviation of logarithmic income, fall between 1980 and 2000 whilst their other two measures, the Gini and Theil indices, are flat or rising after 1980.

This debate on recent trends is important in that it identifies key methodological problems and it emphasizes the fact that any attempt to measure global inequality is subject to a considerable margin of error. The debate is heated because the majority view can be interpreted as support for the equalizing tendencies of global capitalism, giving some comfort to those embarrassed by the evidence of relentless growth in inequality.

Nevertheless, the ‘big pictures’ of both Maddison (2003) and Bourguignon and Morrisson (2002) – see Figures 1 and 2 – prevail. After 150 years of unparalleled growth and rising inequality, global inequality appears to have stabilized towards the end of the 20th century.

Decomposing global income inequality

Within the context of this big picture, I will examine the principal components that contribute to the overall extent of global inequality: inequality across countries; weighting countries by population; and inequalities within countries.

Examining inequality in national average incomes (or GDP per capita) has been part of the focus of research into economic growth and convergence. The consensus in that literature has been aptly summarized in the title of a paper by Lant Pritchett (1997), ‘Divergence, big time’. Some of the growth research has concentrated on evidence of conditional convergence, whereby there is a tendency for poorer countries to grow faster than richer countries provided that some growth determinants are held constant. Conditional convergence is not, however, a sufficient condition for inequality to fall over time, since random shocks will tend to increase dispersion of income levels, and many of the common conditioning factors, such as investment rates or levels of human capital, are distributed in such a way as to limit the growth rates of the poorer countries. So there is no logical contradiction between evidence of conditional convergence and evidence of increasing inequality between countries.

Trends in inter-country inequality are illustrated in Figure 3, where I plot four measures of inequality across 112 countries which together account for nearly 90 per cent of the global population. The time series are represented by the four solid lines. All four measures trend upwards between 1961 and 1996.

Figure 3

Inequality across 112 countries, 1961–1996: population weighted and unweighted. *Source*: Penn World Table 6.1 (Heston, Summers and Aten, 2002).

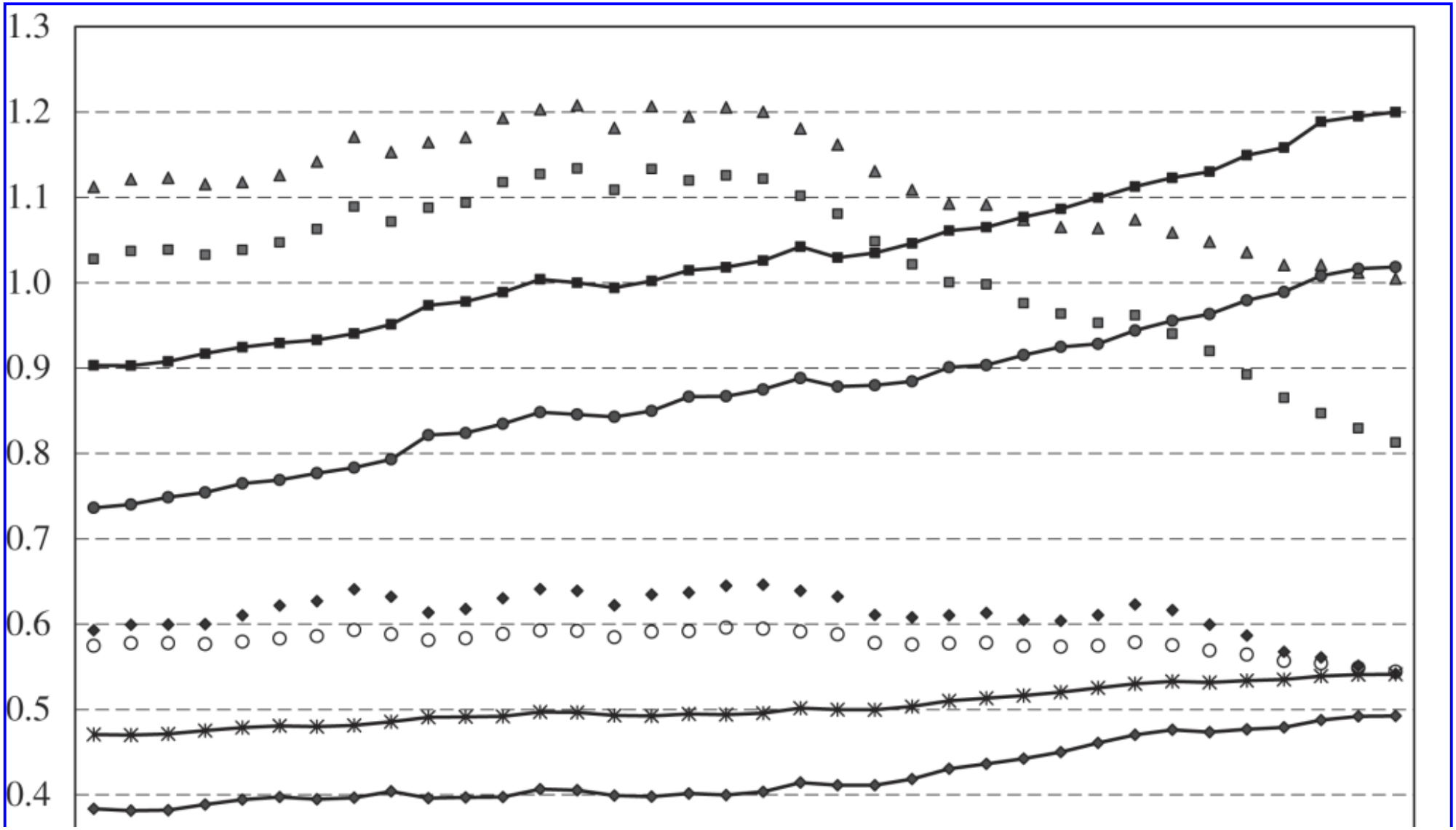

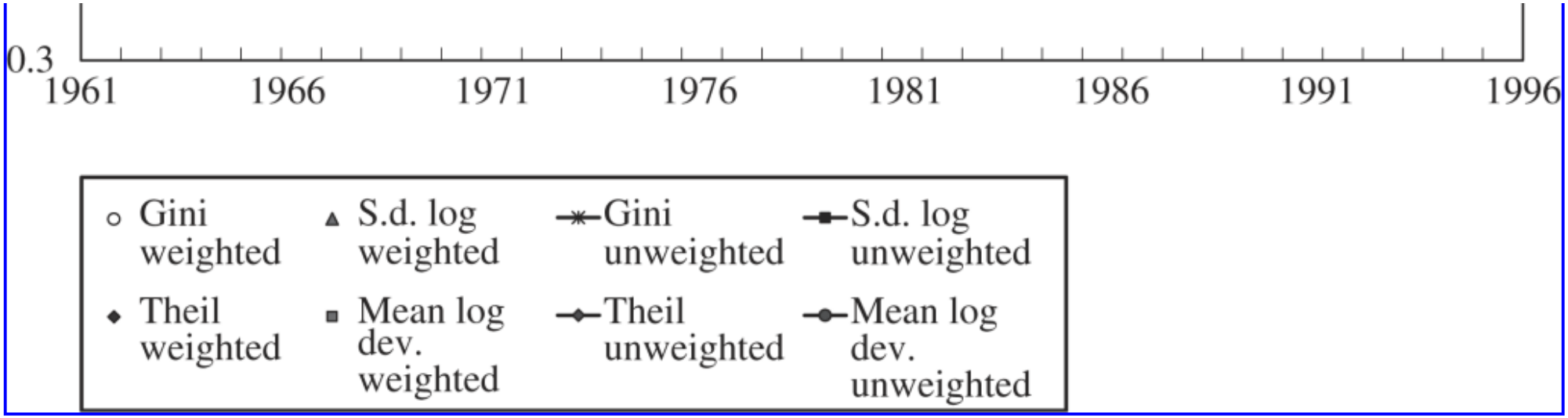

The proximate causes of this rise in inequality between countries in the period include the relatively rapid growth of the already rich United States (averaging 2.2 per cent per year in growth of real GDP per capita), the even faster growth of the relatively rich economies of western Europe (averaging 2.7 per cent per year) which have benefited from technological catch-up with the USA, and the tragedy of African economies which, on average, recorded less than one per cent growth per year. Fifteen African economies experienced falling income levels. With the rich nations becoming relatively richer and the poorest nations becoming relatively poorer, it is no surprise that all four measures of inter-country inequality record increases.

These comparisons, in the tradition of the literature on economic growth, give equal weight to each country. When examining inequality, however, we are often interested in inequality across households or individuals, so it makes sense to weight each country's average income by the population of that country. As many researchers have pointed out, this procedure changes the picture drastically – as illustrated by the dashed lines in Figure 3, which are non-monotonic.

All four of the population-weighted measures of inequality between countries reach a peak in the late 1970s. This peak corresponds to the time when the growth rate of the Chinese economy took off. Through the 1960s and 1970s the Chinese economy grew at a moderate rate, moving average income from 21 per cent of the world average in 1960 to 26 per cent by 1978, still below African income levels. Over the next two decades, the growth rate accelerated, moving Chinese average income in 1996 up to 69 per cent of the world average. This movement of one-fifth of the world's population away from the bottom and towards the middle of the country income distribution is the principal cause of the substantial fall in population-weighted inequality across countries. Another contributory factor was the rise in the growth rate of the Indian economy, which moved from income at 21 per cent of the world average in 1980 to 32 per cent by 2000. (Relative income levels are derived from Maddison, 2003.)

The final dimension to global inequality is inequality within countries. There has been widespread concern within the rich industrialized economies that the rapid expansion in the 1980s and 1990s of trade with low-wage economies such as China would cause increasing inequality as less skilled workers faced wage cuts or unemployment in the face of competitive imports. At the same time, real wages were rising for workers in developing economies who found jobs in the expanding export sectors. Indeed, it has been the case that many of the richest economies have experienced rising income inequality, with Gini coefficients averaging a rise of 3.5 points between 1970 and 1995 in the richer half of the sample of countries. Income inequality also increased in many of the poorer countries, averaging a rise of 2.2 points. (Data on inequality within countries are from Galbraith and Kum, 2003, supplemented by estimates for China in 1970 and 1995 from the UNU-WIIDER data-set, sourced to Dowling and Soo, 1983, and to Khan and Riskin, 1998, respectively.)

Kuznets (1955) has famously observed that, over the course of economic development in the 19th century and the first half of the 20th century, income inequality first rose as labour moved from agriculture into industrial sectors with higher wages and then declined as industrial employment stabilized and wages were equalized. The ensuing implication of a hump-shaped cross-sectional relationship between inequality and income levels is not, however, supported by the cross-sectional evidence from 1970 and 1995, which is illustrated in Figures 4 and 5. Each figure plots the Gini coefficient on the vertical axis against the income level. The best-fit quadratic regression line has been added to each figure. For each year, it is evident that there is a fairly strong tendency for income inequality to fall as average income levels rise. This graphical analysis confirms the results of the econometric study conducted by Schultz (1998).

Figure 4

Intra-country inequality, 1970. *Source*: Galbraith and Kum (2003) and UNU–WIDER (2005).

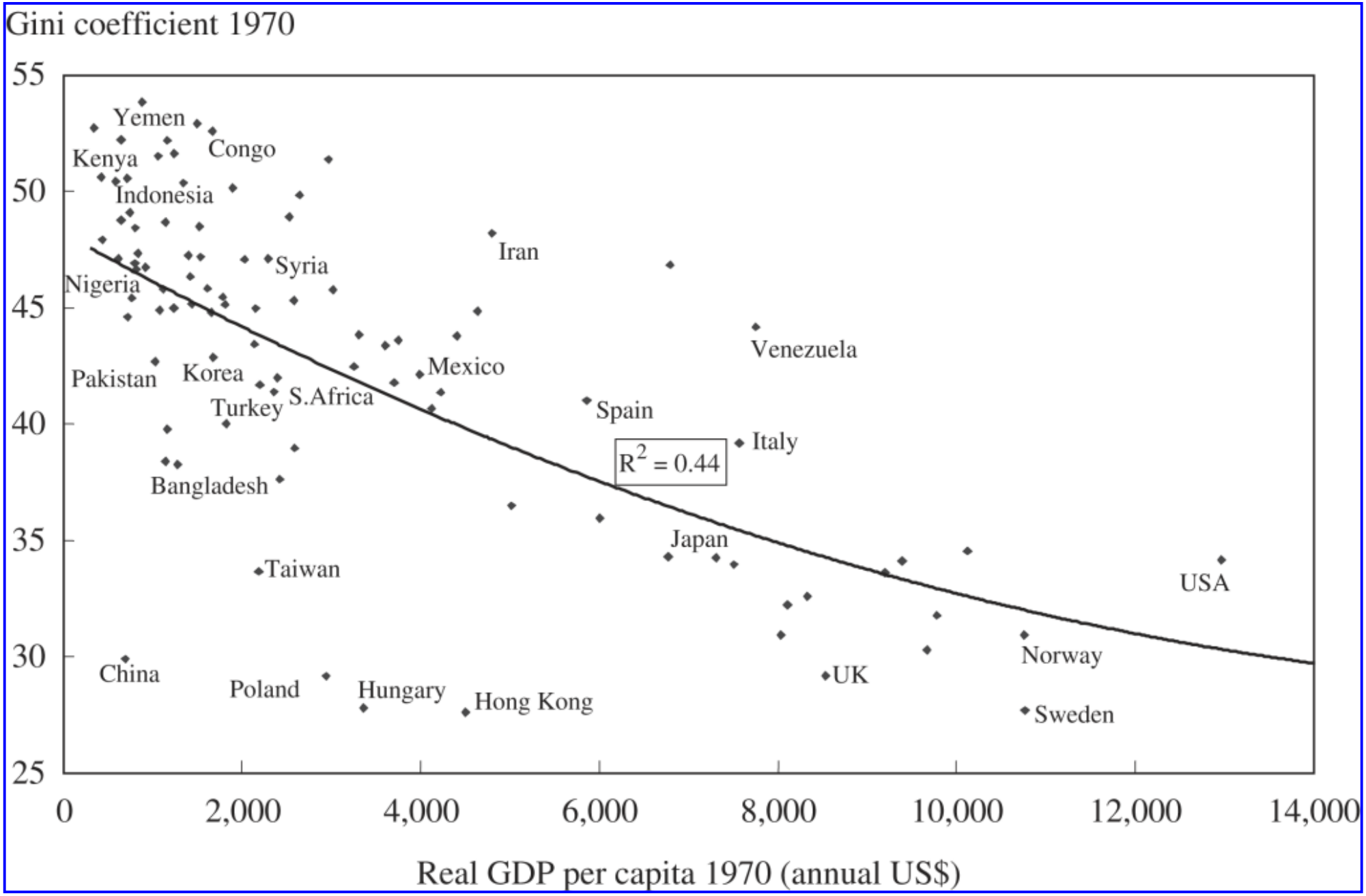

Figure 5
Intra-country inequality, 1995. *Source*: Galbraith and Kum (2003) and UNU–WIDER (2005).

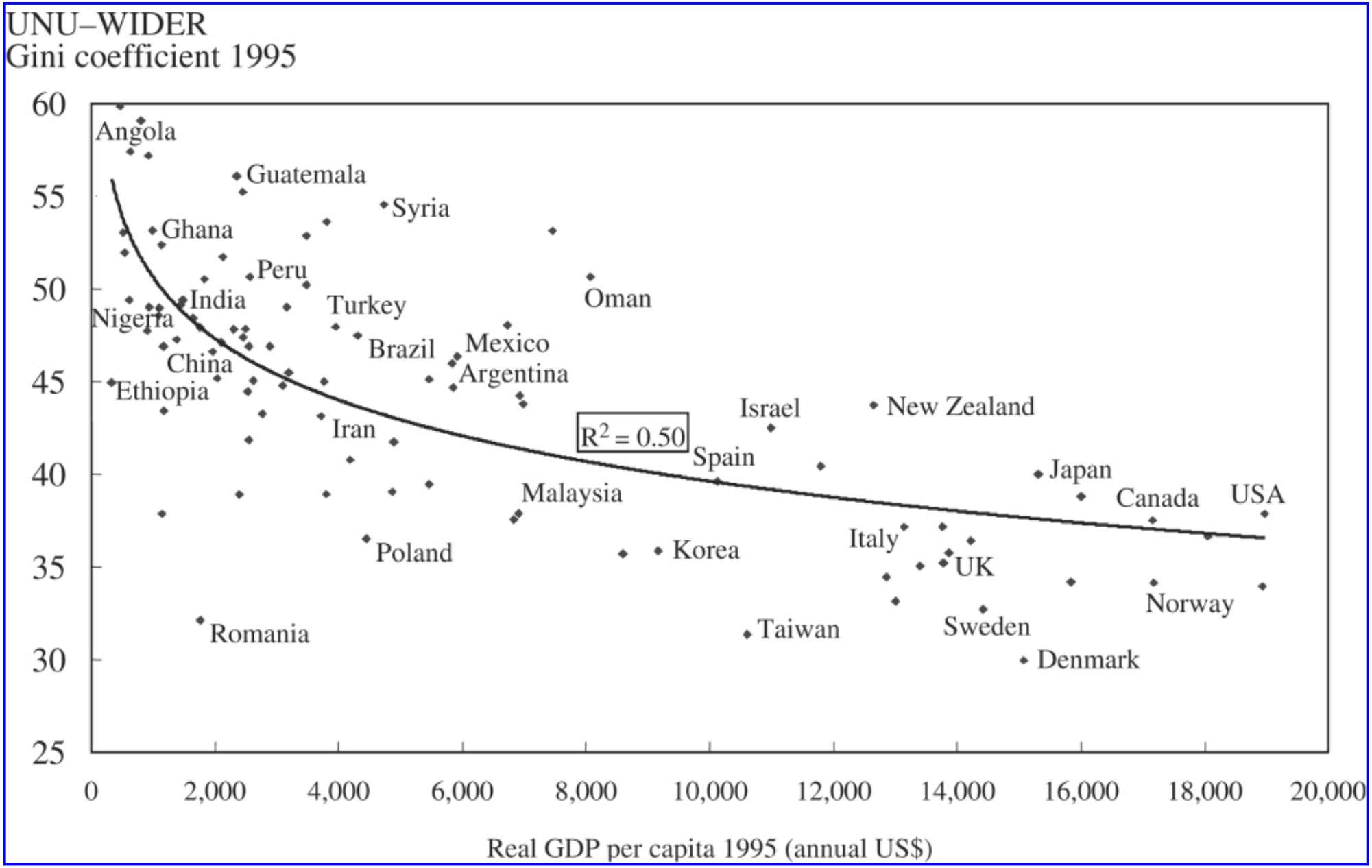

Comparison of Figures 4 and 5 confirms the tendency for inequality to have risen within countries over the 25-year period. It is of particular interest to note the sharp rise in estimated inequality for China, from a Gini around 30 in 1970 – commensurate with the low levels of inequality observed in the Communist countries of eastern Europe – to a Gini of 45 in 1995 – commensurate with the more generally observed levels of inequality amongst other countries at the same level of development. This sharp rise in inequality is in keeping with accounts of rising inequality between the provinces in China, reflecting uneven development between rural regions and the rapidly industrializing coastal cities. Over the same period, the Indian Gini coefficient of inequality was fairly stable at 46.9 in 1970 and 47.2 in 1995.

It might be expected that the general rise in inequality within countries after 1970, particularly within China, would offset any tendency for population-weighted inequality between countries to decline in the 1980s and 1990s. There is some evidence of this offsetting in the Bourguignon and Morrisson (2002, Table 2) data on global inequality, which is illustrated in Figure 2. Two of their measures, the Theil index and the MLD, allow an exact decomposition into within-country and between-country inequality. The within component of the

Theil index rises from 0.315 in 1970 to 0.342 in 1992, whilst the within component of the MLD rises from 0.304 to 0.332. Their Theil measure of global inequality does indeed continue to rise after 1980, although the MLD falls slightly.

Similar results on within-country inequality are reported by Sala-i-Martin (2006), who finds that the within component of the Theil index rises from 0.255 to 0.284 between 1970 and 2000, and the within component of the MLD rises from 0.246 to 0.319. However, his methodology differs from Bourguignon and Morrisson (2002) in that he studies a much larger number of countries and uses nonparametric estimates of within-country distribution. His overall conclusion is that global inequality fell towards the end of the 20th century despite the rise in inequality within countries.

It is noteworthy that 20th century movements in within-country inequality tend to be dominated by movements in population-weighted inequality across countries. This is not surprising since the most glaring inequalities are found in comparisons across countries. Typical values for the quintile ratio (the income of the richest fifth relative to the income of the poorest fifth) are around seven or eight when we look within countries, but across countries the quintile ratio is over 20. This is very different from the situation at the beginning of the 19th century when the dominating influence on global inequality was the extent of inequality within countries.

Methods of comparing income levels across countries

Most studies that examine population-weighted inequality between countries conclude that inequality peaked in the 1970s and declined in the 1980s and 1990s. These studies depend on estimates of GDP per capita evaluated at purchasing power parities (PPP) using data from either Maddison (2003) or the Penn World Table. Maddison's data is used by Bourguignon and Morrisson (2002) and by Sutcliffe (2004). The Penn World Table data are used by Schultz (1998), Firebaugh (1999), Melchior, Telle and Wiig (2000) and Sala-i-Martin (2006), among others.

Several of these studies have contrasted their results with those obtained by Korzeniewicz and Moran (1997) and the United Nations Development Report (UNDP, 2006), who use market rates of exchange rather than PPP exchange rates to compare incomes across countries. The use of market exchange rates leads to the conclusion that income inequality across countries was rising rather than falling over the final decades of the 20th century. There is widespread agreement that exchange rate comparisons are not appropriate if income inequality measures are being calculated in an attempt to evaluate inequality in human welfare. They suffer from two major defects. Market exchange rates are volatile, implying unrealistically sharp short-term movements in real incomes. They also systematically exaggerate real income differentials due to the Balassa–Samuelson effect whereby market exchange rates take no account of the relative cheapness of non-traded goods and services in low-wage low-income countries. Market rates of exchange systematically undervalue incomes in poor countries.

There are, however, some purposes for which the exchange rate measures of inequality may be more appropriate than PPPs. If we are concerned with the ability of poor countries to catch up with the technologies of the rich, and if this depends on their ability to purchase high-tech equipment from the major exporters of capital equipment, then it is the exchange rate which is the appropriate measure of their capacity to develop. The same may well be true when we consider the bargaining power of the poorer nations at international forums such as the World Trade Organization.

To the extent that we are interested in income comparison as an approximation to welfare comparison, it is clearly preferable to compare incomes across countries at purchasing power parity. There is, however, a complication: which measure of purchasing power parity should we use? The PPPs used by both Maddison and the Penn World Table rely on the Geary–Khamis method, which calculates a weighted average of relative prices across all of the countries surveyed by the International Comparison Project in a benchmark year and values the GDP bundles of all countries in all years at that fixed set of prices. The weighting procedure uses country expenditure shares in world GDP, generating ‘world prices’ which are close to the price relativities prevailing in the rich countries of the Organisation for Economic Co-operation and Development (OECD) but very different from the relative prices prevailing in the world's poorer economies. The Geary–Khamis procedure induces substitution bias, valuing the abundant and cheap local services in low-wage economies at the much higher relative price of the rich economies. The effect is the opposite of the bias in the exchange rate comparisons. The Geary–Khamis PPPs systematically overvalue incomes in poorer countries, resulting in measures of global income inequality which are biased downwards. Dowrick and Akmal (2005) demonstrate that the use of Geary–Khamis PPPs can also distort the trend, since the magnitude of the bias changes over time, and show that an unbiased measure of global inequality does not fall between 1980 and 1993.

Further problems with the standard methods of comparing incomes across countries are pointed out by Milanovic (2005). He argues that it is illogical or at least inconsistent to use household survey data to estimate income distribution within countries, but to use national accounting measures rather than the survey measures when computing differences in average income levels across countries. Using average survey income, converted at PPP, he finds that global inequality rose between 1988 and 1993 before falling slightly by 1998. Milanovic notes that average survey income is always less than national accounts measures of average income, or GDP per capita, because it omits public expenditures. Lacking data on the distribution of public expenditures, he argues that survey income is the preferable measure.

Concluding remarks

Global income inequality rose to historically unprecedented proportions, morally repugnant to many, through most of the 19th and 20th centuries – at the same time as average income levels rose to previously unimaginable heights. Since the 1970s, the level of inequality appears to have halted or, by some measures, has begun to fall slightly.
Prospects for the future evolution of global inequality depend crucially on two questions. First, will China continue to follow the trail of development blazed by Japan and Korea several decades earlier? If one-fifth of the world's population does indeed follow this path, then we can expect measures of global inequality to fall as Chinese income level approach the world average; but inequality will then increase as Chinese income levels catch up with those of the global rich. Second, can the desperately poor nations of Africa find a way, with or without the assistance of the rest of the world, to follow the successful development path on which China and India embarked in the 1980s and the 1990s? If African development fails to take off and if population growth continues to exceed that of the other continents, then global inequality may well resume its rising trend in the course of the 21st century.

See Also

- Gini ratio
- inequality (measurement)
- Kuznets, Simon

Bibliography

Bourguignon, F. and Morrisson, C. 2002. Inequality among world citizens: 1820–1992. *American Economic Review* 92, 727–44.

Deininger, K. and Squire, L. 1996. A new data set measuring income inequality. *World Bank Economic Review* 10, 565–91.

Dowling, J.M. and Soo, D. 1983. Income distribution and economic growth in developing Asian countries. Staff Paper No. 15. Manila: Asian Development Bank.

Dowrick, S. and Akmal, M. 2005. Contradictory trends in global income inequality: a tale of two biases. *Review of Income and Wealth* 51, 201–29.

Firebaugh, G. 1999. Empirics of world income inequality. *American Journal of Sociology* 104, 1597–630.

Galbraith, J.K. and Kum, H. 2003. Estimating the inequality of household incomes: filling gaps and fixing problems in Deininger and Squire. Working paper, Inequality Project, University of Texas.

Heston, A., Summers, R. and Aten, B. 2002. Penn World Table Version 6.1. Center for International Comparisons, University of Pennsylvania.

International Comparison Program 2007. 1996 price survey data: http://pwt.econ.upenn.edu/Downloads/benchmark/benchmark.html.

Khan, A.R. and Riskin, C. 1998. Income and inequality in China: composition, distribution and growth of household income. *China Quarterly* 154, 221–53.

Korzeniewicz, R.P. and Moran, T.P. 1997. World economic trends in the distribution of income, 1965–92. *American Journal of Sociology* 102, 1000–39.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45, 1–28.

Maddison, A. 2003. *The World Economy: Historical Statistics*. Paris: OECD.

Melchior, A., Telle, K. and Wiig, H. 2000. *Globalization and Inequality: World Income Distribution and Living Standards, 1960–1998*. Oslo: Norwegian Institute of Foreign Affairs.

Milanovic, B. 2002. True world income distribution, 1988 and 1993: first calculations based on household surveys alone. *Economic Journal* 112, 51–92.

Milanovic, B. 2005. *Worlds Apart: Measuring International and Global Inequality*. Princeton and Oxford: Princeton University Press.

Pritchett, L. 1997. Divergence, big time. *Journal of Economic Perspectives* 11(3), 3–17.

Sala-i-Martin, X. 2006. The world distribution of income: falling poverty and … convergence, period. *Quarterly Journal of Economics* 121, 351–97.

Schultz, T.P. 1998. Inequality in the distribution of personal income in the world: how it is changing and why. *Journal of Population Economics* 11, 307–44.

Summers, R. and Heston, A. 1991. The Penn World Table (Mark 5): an expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* 106, 327–68.

Sutcliffe, B. 2004. World inequality and globalization. *Oxford Review of Economic Policy* 20, 15–37.

UNDP (United Nations Development Programme). 2006. *Human Development Report 2006*. New York: Oxford University Press.

UNU–WIDER. 2005. World income inequality database 2.0a. Online. Available at http://www.wider.unu.edu/wiid/wiid.htm, accessed 8 March 2007.

World Bank. 2006a. *World Development Report*. New York: Oxford University Press.

World Bank. 2006b. *World Development Indicators 2006*. Online. Available at http://go.worldbank.org/3SGLDH5V10, accessed 12 April 2007.

How to cite this article

Dowrick, Steve. "inequality (global)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000270> doi:10.1057/9780230226203.0787

The New Palgrave Dictionary of Economics Online

inequality (international evidence)

Andrea Brandolini, Timothy M. Smeeding
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The methodological assumptions underlying international comparisons of levels and trends in inequality are discussed, starting with the choice of the evaluative space. Empirical evidence shows that at the end of the 1990s, the United States had the highest level of disposable income inequality among high-income economies, while northern and central European countries had the lowest levels. Only in Russia and Mexico, two middle-income economies, was disposable income more unequally distributed. No common trend in inequality is observed since the 1970s across rich nations. Public redistribution through taxes and benefits influence both levels and changes in inequality.

Keywords

Atkinson index; capability approach; consumer price index; disposable income; expenditure; Gini index; human capital; income; income inequality; inequality (measurement); inequality, international evidence of; Kuznets, S.; Lorenz curve; Luxembourg Income Study; market income; Pareto's law; purchasing power parity; redistribution of income; relative inequality; standard of living; Theil index

Article

The comparison of inequality across countries and over time has a long tradition in economics. In 1897 Pareto used data from tax returns for a heterogeneous group of nations, spanning a period of almost four centuries, to conclude that income inequality was remarkably constant over time and space. An intense debate followed, such that the editors of *Econometrica* devoted the second ‘Annual Survey of Statistical Data’ to Pareto's law (Bresciani-Turroni, 1939), which served to bring to an end the idea of a ‘natural’ constancy of the distribution of income.

The study of international differences in income distribution gathered new momentum after the Second World War. In the 1950s, United Nations agencies pioneered the assembly of international data-sets on income inequality (for example, United Nations, 1951) and Kuznets (1955) stated his celebrated hypothesis of an inverted-U relationship between inequality and growth. Since those early days, international agencies and individual scholars have increasingly been engaged in collecting information on income distribution and comparing levels and trends of inequality across nations (Gottschalk and Smeeding, 1997; Atkinson and Brandolini, 2001). Cross-country comparisons of income inequality have become common in analysis that informs policymaking: measures of income distribution are featured among the indicators of social cohesion agreed by the European Union to monitor the performance of member countries (Atkinson et al., 2002), and one of the first charts of the 2006 *World Development Report* ranks nations by the Gini index of income (or expenditure) to show that ‘Africa and Latin America have the world's highest levels of inequality’ (World Bank, 2005, Fig. 2.9, p. 39; the underlying data are reported in Table 1).

World Bank's estimates of inequality levels: income and expenditure. Gini indices

Country	Year	Gini index	Income group
High-income economies			
Expenditure			

•Taiwan	2000	0.24	HIC
•Italy	2000	0.31	HIC
•Israel	2001	0.35	HIC
•Greece	1998	0.36	HIC
Income			
•Finland	2000	0.25	HIC
•Japan	1993	0.25	HIC
•Sweden	2000	0.25	HIC
•Belgium	2000	0.26	HIC
•Denmark	1997	0.27	HIC
•Norway	2000	0.27	HIC
•Austria	1997	0.28	HIC
•Germany	2000	0.28	HIC
•Luxembourg	2000	0.29	HIC
•Netherlands	1999	0.29	HIC
•France	1994	0.31	HIC
•Ireland	2000	0.31	HIC
•Switzerland	1992	0.31	HIC
•Australia	1994	0.32	HIC
•Republic of Korea	1998	0.32	HIC
•Canada	2000	0.33	HIC
•United Kingdom	1999	0.34	HIC
•Spain	2000	0.35	HIC
•New Zealand	1997	0.37	HIC
•United States	2000	0.38	HIC
•Portugal	1997	0.39	HIC
•Singapore	1998	0.43	HIC
Middle East and North Africa			
Expenditure			
•Yemen	1998	0.33	LIC
•Egypt	2000	0.34	LMC
•Algeria	1995	0.35	LMC
•Morocco	1998	0.38	LMC
•Jordan	2002	0.39	LMC
•Tunisia	2000	0.40	LMC
•Iran	1998	0.43	LMC
South Asia			

Expenditure			
•Pakistan	2001	0.27	LIC
•Bangladesh	2000	0.31	LIC
•India	1999/2000	0.33	LIC
•Nepal	1996	0.36	LIC
•Sri Lanka	2002	0.38	LMC
East Asia and Pacific			
Expenditure			
•Mongolia	1998	0.30	LIC
•Indonesia	2000	0.34	LMC
•Lao PDR	1997/1998	0.35	LIC
•Vietnam	2002	0.35	LIC
•Cambodia	1997	0.40	LIC
•Thailand	2002	0.40	LMC
•China	2001	0.45	LMC
•Philippines	2000	0.46	LMC
Income			
•Malaysia	1997	0.49	UMC
Europe and Central Asia			
Expenditure			
•Hungary	2002	0.24	UMC
•Bosnia & Herzegovina	2001	0.25	LMC
•Armenia	2003	0.26	LMC
•Uzbekistan	2000	0.27	LIC
•Bulgaria	2003	0.28	LMC
•Romania	2002	0.28	LMC
•Serbia & Montenegro	2003	0.28	LMC
•Slovenia	1998	0.28	HIC
•Croatia	2001	0.29	UMC
•Kyrgyzstan	2002	0.29	LIC
•Lithuania	2000	0.29	UMC
•Belarus	2000	0.30	LMC
•Kazakhstan	2003	0.30	LMC
•Albania	2002	0.31	LMC
•Poland	2002	0.31	UMC
•Estonia	1998	0.32	UMC
•Russian Federation	2002	0.32	UMC

•Tajikistan	2003	0.32	LIC
•Latvia	1998	0.34	UMC
•Azerbaijan	2001	0.36	LMC
•Macedonia	2003	0.36	LMC
•Moldova	2001	0.36	LIC
•Turkey	2002	0.37	UMC
•Georgia	2002	0.38	LMC
•Turkmenistan	1998	0.41	LMC
Income			
•Czech Republic	1996	0.25	UMC
•Slovak Republic	1996	0.26	UMC
•Ukraine	1999	0.29	LMC
Latin America and the Caribbean			
Expenditure			
•Trinidad & Tobago	1992	0.39	UMC
•Nicaragua	2001	0.40	LIC
•Jamaica	2001	0.42	LMC
•St. Lucia	1995	0.44	UMC
•Peru	2000	0.48	LMC
•Panama	2000	0.55	UMC
Income			
•Venezuela	2000	0.42	UMC
•Uruguay (urban)	2000	0.43	UMC
•Guyana	1998	0.45	LMC
•Costa Rica	2000	0.46	UMC
•Dominican Republic	1997	0.47	LMC
•Mexico	2002	0.49	UMC
•El Salvador	2002	0.50	LMC
•Argentina (urban)	2001	0.51	UMC
•Chile	2000	0.51	UMC
•Honduras	1999	0.52	LMC
•Colombia	1999	0.54	LMC
•Ecuador	1998	0.54	LMC
•Paraguay	2001	0.55	LMC
•Bolivia	2002	0.58	LMC
•Guatemala	2000	0.58	LMC
•Brazil	2001	0.59	LMC

•Haiti	2001	0.68	LIC

Notes: Economies are classified by the World Bank according to 2004 per capita gross national income in the following income groups: low-income economies (LIC), $825 or less; lower-middle-income economies (LMC), $826–$3,255; upper-middle income economies (UMC), $3,256–$10,065; and high-income economies (HIC), $10,066 or more. *Source*: World Bank (2005, Table A2, pp. 280–1).

Focal variable

As Sen suggests, the relative advantages and disadvantages that people have, compared with each other, can be judged in terms of many different variables, e.g. their respective incomes, wealths, utilities, resources, liberties, rights, quality of life, and so on. The plurality of variables on which we can possibly focus (the *focal variables*) to evaluate interpersonal inequality makes it necessary to face, at a very elementary level, a hard decision regarding the perspective to be adopted. (Sen, 1992, p. 20)

Pareto saw the distribution of income as a reflection of the natural distribution of abilities among persons, while Kuznets regarded its evolution as one of the characteristics of the process of economic growth; but they both agreed that the focal variable should be income. However, other dimensions of economic inequality are relevant in international comparisons. Earnings dispersion and differences in employment rates capture inequality in the labour market. Wealth may be seen as an indicator of the capacity to face adverse events or of the power to control the resources of the society. The standard of living is much influenced by non-monetary aspects, such as a person's health status or human capital – as stressed by the ‘capability approach’ advocated by Sen (1992).

In this article, the focal variable is taken to be income, the most common indicator of (current) economic resources in rich countries. Expenditure is an alternative variable often used, especially in less developed countries. The World Bank (2005, Table A2, pp. 280–1) reports income-based Gini indices for 22 of the 27 high-income economies for which the statistics are available vis-à-vis 20 of the 60 middle-income economies and only one of the 39 low-income economies. Mixing income-based and consumption-based statistics confounds international comparisons, as income tends to be more unequally distributed than expenditure – and to an extent that varies considerably from country to country (for example, World Bank, 2005, Box 2.5, p. 38).

Wealth (net worth) is much more concentrated than income. Moreover, international comparisons of net worth are very problematic (Wolff, 1996; Davies and Shorrocks, 2000) as the assembling of cross-nationally comparable databases on household net worth is still in its infancy (Sierminska, Brandolini and Smeeding, 2006).

Methodology

International comparisons of income inequality crucially depend on the underlying measurement assumptions. This has been known at least since Kravis (1962) and Kuznets (1963) and has received growing attention from the mid-1970s (for example, Atkinson, 1974; Sawyer, 1976; Lydall, 1979). However, it was not until the assembling of the cross-nationally comparable database of the Luxembourg Income Study (LIS) that the impact of these assumptions was fully understood (Smeeding, 2004). Differences in methodology arise in the definition of income, the choice of the recipient unit, the quality of underlying sources, the treatment of individual data (O'Higgins, Rainwater and Smeeding, 1990; Atkinson, Rainwater and Smeeding, 1995; Gottschalk and Smeeding, 1997, 2000; Atkinson and Brandolini, 2001).

Income definitions differ in comprehensiveness, as certain income sources like capital gains, imputed rents on owner-occupied dwellings, or home production may or may not be included. There are also widespread differences in the treatment of taxes (and social security contributions), as income may be taken before taxes, before taxes but after allowing for tax deductions, or after taxes. The definition of income may be augmented to include the imputed value of public in-kind benefits for education, health care and housing or to deduct indirect taxes. Moreover, income may be measured over a variety of time periods: the reference is often the year, but in some cases it is some ‘current’ period (for example, the most recent pay period for earnings in household surveys for the United Kingdom) and then the annual amount must be estimated.

The *reference unit* may be the household, the related or extended family, the tax unit, or the individual income earner. Information obtained from income tax records typically relates to the tax unit only, while sample surveys generally provide data for all members of a household. The total income may be adjusted for the size and the composition of the reference unit by dividing by an *equivalence scale*. Indeed, not adjusting income implies that the welfare achievable in a household with a certain income is independent of the number of its occupants. At the other extreme, taking income per capita amounts to an assumption that no economies of scale arise from cohabitation and that people do not differ in their needs.

The *welfare unit* may be the person (person-weighted) or the household (household-weighted): in the former case the welfare indicator represented by (equivalent) income is counted as many times as there are persons in the household, while in the latter it is counted only once. This welfare weighting is a separate issue from that of the equivalence scale: for instance, the European Commission (2002) typically reports statistics for the distribution of *equivalent disposable incomes* among *persons*, while the U.S. Census Bureau (2005) presents figures for the distribution of *unadjusted money incomes before taxes* among *households*.

Diversity in definitions is not the only factor that affects the comparability of income inequality statistics. There are also differences in the *nature of the data source*, the most

important distinction being between sample surveys and administrative archives. Data may cover the whole population or only the household population, excluding people living permanently in institutions like boarding houses, nursing homes for the elderly, prisons, or military bases. Administrative data reflect the purposes for which they were collected. Even when sources have the same nature, they may considerably vary in quality, through differences in the response rate, the under-reporting of certain income components, or the coverage of the bottom and the top of the distribution. Lastly, significant differences can originate in the way data are processed. For example, the Gini index may be computed from micro-data or from observations grouped by income classes. When the ranking of observations is based on a variable different from that of concern, say before-tax income instead of after-tax income, measures of inequality are understated.

All these factors influence international comparisons of income inequality, as shown for instance by Buhmann et al. (1988) with regard to equivalence scales and by Smeeding et al. (1993) with regard to the inclusion of in-kind public benefits. These differences need to be kept in mind when making international comparisons. While perfect comparability is not achievable, it is important to raise the ratio of signal to noise by minimizing data and methodological differences across nations (Gottschalk and Smeeding, 2000).

Relative inequality levels

Figure 1 compares the distribution of equivalent disposable income among persons in 32 nations for various years around the turn of the 21st century, or for the most recent year available in the LIS database. Disposable income is defined as the sum of wages, salaries and earnings from self-employment, cash receipts from property, private pension schemes, alimony and child support, public transfer payments (retirement pensions, family allowances, unemployment compensation, and welfare benefits) less income taxes and social security contributions. Observations are top- and bottom-coded in order to reduce the influence of anomalous income values. Total household income, the sum over all household members, is divided by a simple equivalence coefficient (the square root of the household size) and then attributed to each person in the household.

Figure 1

The distribution of disposable income in 32 high- and middle-income economies. *Notes*: P10 and P90 are the ratios to the median of the tenth and 90th percentiles, respectively. Observations are bottom-coded at 1% of the mean of equivalent disposable income and top-coded at ten times the median of unadjusted disposable income. Incomes are adjusted for household size by the square-root equivalence scale. See note to Table 1 for the definition of high- and middle-income economy. *Sources*: Authors' calculations from the Luxembourg Income Study database, as of 10 March 2007 (figures coincide with those reported in http://www.lisproject.org/keyfigures/ineqtable.htm) and the European Community Household Panel database, Waves 1–8, December 2003 for Portugal; statistics for Japan were computed according to the same methodology as all other figures by Tsuneo Ishikawa for Gottschalk and Smeeding (2000).

	P10 (Low income)	Length of bars represents the gap between high and low income individuals	P90 (High income)	P90/P10 (Decile ratio)	Gini index
High-income economies					
Denmark 2000	57		155	2.8	0.225
Norway 2000	57		159	2.8	0.251
Finland 2000	57		164	2.9	0.247
Sweden 2000	57		168	3.0	0.252
Netherlands 1999	56		167	3.0	0.248
Slovenia 1999	53		167	3.2	0.249
Austria 2000	55		173	3.2	0.260
Luxembourg 2000	57		184	3.2	0.260
Belgium 2000	53		174	3.3	0.277
Switzerland 2000	55		182	3.3	0.280
Germany 2000	54		180	3.4	0.275

France 2000	55	188	3.4	0.278
Taiwan 2000	52	196	3.8	0.296
Canada 2000	48	188	3.9	0.302
Japan 1992	46	192	4.2	0.315
Australia 2001	47	199	4.2	0.317
Italy 2000	45	199	4.5	0.333
Ireland 2000	41	189	4.6	0.323
United Kingdom 1999	47	215	4.6	0.343
Greece 2000	43	207	4.8	0.338
Spain 2000	44	209	4.8	0.340
Israel 2001	43	216	5.0	0.346
Portugal 2000	45	226	5.0	0.363
United States 2000	37	212	5.7	0.370
Middle-income economies				
Slovak Republic 1996	56	162	2.9	0.241
Czech Republic 1996	59	179	3.0	0.259
Romania 1997	53	180	3.4	0.277
Hungary 1999	54	194	3.6	0.295
Poland 1999	52	188	3.6	0.293
Estonia 2000	46	234	5.1	0.361
Russia 2000	33	276	8.4	0.434
Mexico 2000	32	331	10.4	0.491

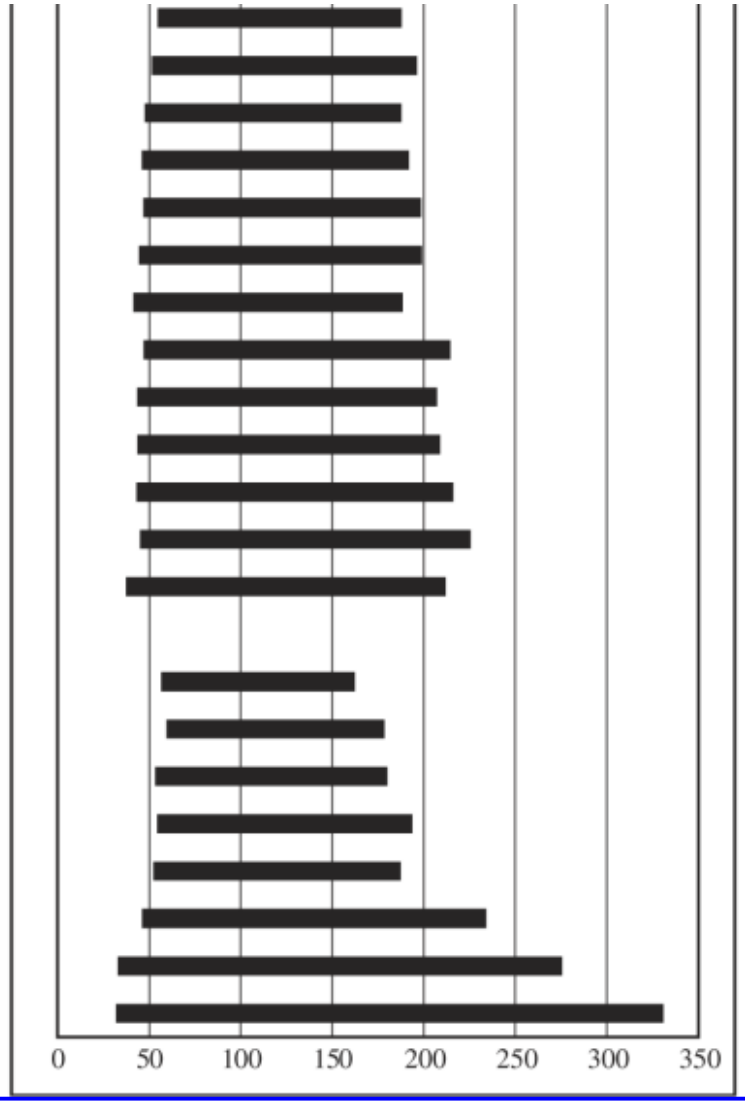

Figure 1 reports, for each country, the ratio to the median of the income of a person at the tenth percentile (P10 or ‘low income’) and a person at the 90th percentile (P90 or ‘high income’). P10 and P90 provide some indication of how far below or above the middle of the distribution the poor and the rich are on the continuum of income. The ratio between P90 and P10, the ‘decile ratio’, is a measure of the gap between the rich and the poor. While these statistics refer to specific points of the distribution, the Gini index measures inequality across the entire distribution. For non-negative values, it varies between zero (perfect equality) and one (maximum inequality).

There is a wide range of income inequality among the nations of Figure 1. The United States is an outlier among rich nations, and only Russia and Mexico, two middle-income economies, have higher levels of inequality. A low-income American at the 10th percentile in 2000 had an income that was only 37 per cent of the median income. By contrast, in most countries of central, northern and eastern Europe the income of the poor exceeded 50 per cent of the income of middle-income person; in the other English-speaking nations and in the southern European countries, plus Israel, it was above 40 per cent. Only in Russia and Mexico did the poor fare relatively worse than in the United States. In Greece, Portugal, Spain, Israel as well as the United States and the United Kingdom the rich persons earn more than twice the national median incomes. In poorer countries the 90th percentile can also be very high in relative terms, for example in Mexico, Russia, and Estonia.

The countries in Figure 1 fall into distinctive clusters. Inequality, as measured by the decile ratio, is least in Nordic countries, the Netherlands and the Czech and Slovak Republics with values of 3 or less. The two other Benelux countries (Belgium and Luxembourg), Central Europe (France, Switzerland, Germany, Austria, Slovenia) and three other Eastern European countries (Hungary, Poland, Romania) come next at 3.2–3.6. These precede four English-speaking nations (Canada, Australia, Ireland and the United Kingdom), which

have decile ratios comprised between 3.9 and 4.6, and the southern European countries (Italy, Spain, Greece and Portugal) and Israel, whose ratios fall between 4.5 and 5.0. Only the United States, Estonia, Mexico and Russia have values in excess of 5. With decile ratios around 4, the two Asian countries, Taiwan and Japan, are in an intermediate position. Inequality differs much more across middle-income than high-income economies. While Estonia, Russia and Mexico show a very unequal distribution of income, the other five countries, all from eastern Europe, exhibit moderate or low levels of inequality. The shape of the income distribution was noticeably different even in the mid-1980s across these formerly planned economies, with Czechoslovakia showing the least inequality and the Soviet Union the highest (Atkinson and Micklewright, 1992).

In Figure 1 countries are arranged, within the two categories of high-income and middle-income, by the decile ratio, from lowest to highest. This country rank order does not need to coincide with that based on the other statistics reported: P10, P90 and the Gini index. For instance, Sweden shows the second highest P10 but the seventh lowest Gini index. This follows from the fact that the Swedish at the 90th percentile is less closer to the middle than the equivalent person in Denmark, Finland or the Slovak Republic. (These differences should not be overstressed as they are small and likely to be within the bounds of sampling error.) The rankings of countries in international comparisons depends on which part of the distribution is analysed, for example, the bottom with P10 or the top with P90, or in the way single observations are weighted by a summary measure of inequality like the Gini index, or the Theil and Atkinson indices. Different summary measures may produce different results reflecting differences at the top and bottom of the distribution. More robust, but partial, rankings are obtained by comparing the entire distributions by ‘Lorenz dominance’, whereby inequality is assessed to be unequivocally higher in country *A* than in country *B* if the Lorenz curve of country *A* lies everywhere below that of country *B*, but no unambiguous conclusion is achieved if the two curves intersect. Although countries may switch their relative positions, indices are still in general highly correlated: for instance, the correlation between decile ratio and Gini index in Figure 1 is 0.97. The basic patterns of international inequality are clear regardless of the measure of inequality employed.

Redistribution

Every nation's tax and benefit system reduces market income inequality, but not all are equally effective in doing so. The efficiency with which nations accomplish this redistribution may vary over time as well as space. A common measure of the level of redistribution is represented by the difference between the Gini index for market incomes, that is, before public transfers are added and taxes and social security contributions are deducted, and the Gini index for disposable incomes. This difference provides only a first estimate of the actual impact of public redistribution, as it ignores how market income inequality would be different if there were no taxes and benefits. Table 2 shows the extent of redistribution in 16 countries using LIS data.

Gini indices of market income and disposable income in 16 countries (per cent)

Country	Year	Gini index for market income	Gini index for disposable income	Absolute reduction	Percentage reduction
		[1]	[2]	[3]=[1]–[2]	[4]=[3]/[1]
High-income economies					
•Denmark	2000	42	23	20	47
•Finland	2000	38	25	14	36
•Netherlands	1999	39	25	14	36
•Norway	2000	41	25	16	39
•Sweden	2000	46	25	21	45
•Germany	2000	48	28	21	43
•Switzerland	2000	36	28	8	22
•Taiwan	2000	33	30	3	9
•Canada	2000	42	30	12	28
•Australia	2001	48	32	17	34
•United Kingdom	1999	51	34	17	33
•Israel	2001	52	35	17	33
•United States	2000	48	37	11	23

Middle-income economies					
•Czech Republic	1996	44	26	18	41
•Romania	1997	38	28	10	27
•Poland	1999	50	29	21	41

Notes: Observations for disposable income are bottom-coded at 1% of the mean of equivalent disposable income and top-coded at ten times the median of unadjusted disposable income. Changes in disposable incomes due to bottom- and top-coding are entirely attributed to market incomes. Both market and disposable incomes are adjusted for household size by the square-root equivalence scale. *Source*: Authors' calculations from the Luxembourg Income Study database, as of 10 March 2007.

In all nations disposable incomes are more equally distributed than market incomes, suggesting that the tax and benefit system narrows the overall distribution. On average, inequality falls by about a third, from a Gini index of 44 to one of 29 per cent. Cross-country variation in original inequality is wider than after redistribution: the Gini index ranges from 33 to 52 per cent for market incomes, and from 23 to 37 per cent for disposable incomes. The United States has the highest inequality of disposable incomes, although the dispersion of market incomes is on the high side but not far from most other countries; it is as high as in Germany and Australia and below the values recorded for the United Kingdom, Poland and Israel. The fact is that the percentage reduction in before-tax-and-benefit inequality in the United States is a mere 23 per cent. If we exclude Taiwan, where redistribution has a tiny impact, only Switzerland shows a reduction as low as the United States, but the Swiss start from a much more equal distribution and end with a Gini index below the average.
These percentage reductions are very consistent with the patterns of aggregate public spending. High-spending northern and central European nations have the highest degree of inequality reduction, from 36 to 47 per cent; the English-speaking Anglo-Saxon (excluding the United States) nations and Israel are next with 28 to 33 per cent reductions; the United States and Switzerland are, as just seen, at the bottom of the scale. The degree of redistribution in southern Europe is lower than in Ireland and the United Kingdom, especially if public pensions are not included among transfers, according to the EUROMOD estimates based on micro-simulations rather than the records of the original micro-data sources (Immervoll et al., 2005). The nations that redistribute the most are not necessarily those with the greatest degree of market income inequality: before-tax-and-benefit incomes in Finland and the Netherlands are far more equally distributed than in the United States.

Absolute inequality levels

The comparisons in Figure 1 relate to *relative* inequality. The income of the poor at the tenth percentile is compared with the income of the person at the middle of the distribution in the same country. When average standards of living differ across nations, results may look quite different if comparisons are made in terms of *real* income, that is, the amount of goods that a certain income can purchase.
The statistics in Figure 2 on real incomes in 2000 international dollars are derived by adjusting the original incomes by the national consumer price indices (CPI) and converting them by means of the purchasing power parities (PPP) for gross domestic product (GDP). The real P10 and P90 are then recomputed as a fraction of the US median real income. These comparisons are very rough indicators of differences in 'real living standards'. First, the conversion to real income across countries and time is sensitive to the PPP and consumer price indices used. Second, the PPPs are computed for national accounts which are intrinsically different from survey data (Deaton, 2005). For instance the ratios of total survey incomes to GDP aggregates vary considerably across these countries. Thus countries with surveys that capture less of national income appear to have much lower mean living standards than countries whose surveys or administrative records capture a larger share of that income. Third, it is questionable that the same conversion factor should be applied across the entire distribution. Lastly, real income does not account for goods and services such as education and health care that are provided at different prices and under different financing schemes in different nations. As low-income citizens in some countries need to spend more out of pocket for these goods than do low-income citizens in other countries, their living standard is relatively lower than that measured by PPP-adjusted income.
Figure 2
The distribution of real disposable income in 32 high- and middle-income economies. *Notes*: Real P10 and P90 are the percentage ratios to the US median of the tenth and 90th percentiles, respectively; real median is expressed as a percentage ratio of the US median. Observations are bottom-coded at 1% of the mean of equivalent disposable income and top-coded at ten times the median of unadjusted disposable income. Incomes are adjusted for household size by the square-root equivalence scale. *Sources*: Authors' calculations from the Luxembourg Income Study database, as of 10 March 2007, and the European Community Household Panel database, Waves 1–8, December 2003 for Portugal; statistics for Japan were computed according to the same methodology as all other figures by Tsuneo Ishikawa for Gottschalk and Smeeding (2000). Consumer price indices and purchasing power parity conversion factors from local currency units to international dollars are from International Monetary Fund (2006).

	Real P10 (Low income)	Length of bars represents the gap between high and low income individuals	Real P90 (High income)	P90/P10 (Decile ratio)	Real median

High-income economies				
Denmark 2000	45	125	2.8	80
Norway 2000	48	134	2.8	84
Finland 2000	36	103	2.9	63
Sweden 2000	34	99	3.0	59
Netherlands 1999	39	116	3.0	69
Slovenia 1999	23	73	3.2	44
Austria 2000	41	129	3.2	75
Luxembourg 2000	65	209	3.2	114
Belgium 2000	38	125	3.3	72
Switzerland 2000	45	150	3.3	82
Germany 2000	39	130	3.4	72
France 2000	34	117	3.4	62
Taiwan 2000	39	150	3.8	76
Canada 2000	39	152	3.9	81
Japan 1992			4.2	
Australia 2001	30	129	4.2	65
Italy 2000	25	111	4.5	56
Ireland 2000	30	136	4.6	72
United Kingdom 1999	31	141	4.6	66
Greece 2000	20	96	4.8	46
Spain 2000	25	121	4.8	58
Israel 2001	23	115	5.0	53
Portugal 2000	18	91	5.0	40
United States 2000	37	212	5.7	100
Middle-income economies				
Slovak Republic 1996	14	40	2.9	25
Czech Republic 1996	18	54	3.0	30
Romania 1997	8	27	3.4	15
Hungary 1999	12	44	3.6	23
Poland 1999	12	45	3.6	24

Estonia 2000	9	47	5.1	20
Russia 2000	3	26	8.4	10
Mexico 2000	4	46	10.4	14

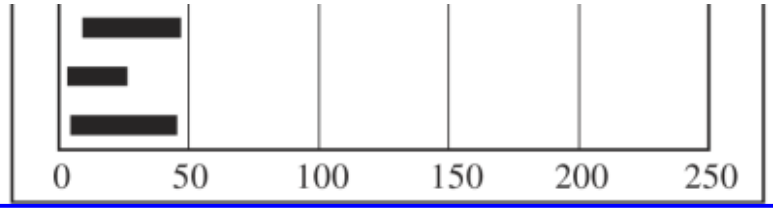

The living standard of the median German or Belgian appears to be 72 per cent of that of the median American; but the living standard of poor Germans and Belgians is just above that of their American counterparts, 38–39 per cent against 37 per cent of the US median. Low-income people in Denmark, Norway, Switzerland and, especially, Luxembourg are much better off than elsewhere. In all southern European countries but also, to a lesser extent, in Australia, Ireland and the United Kingdom, the living standards of low-income households were lower than in the United States. Of course, they are a great deal lower in all middle-income economies. At the other extreme, the rich Americans far surpass the rich in any other nation observed, save for the Luxembourgers.

Long-run trends in high-income economies

Movements of inequality over time follow irregular trajectories rather than smooth profiles, with more substantial changes often concentrated in few episodes (Atkinson, 1997). Some causes are common to many countries, such as the spreading of skilled-biased technologies, the greater world economic integration, or the aging of population in more recent decades; some others are more specific to national experiences, typically changes in tax-and-benefit systems but also modifications in institutions such as wage setting policies. The evolution of inequality reflects the joint working of these factors, which sometimes balance out and sometimes reinforce each other, making it an arduous task to disentangle common trends from idiosyncratic variations. Moreover, changes in data collection and statistical methodology interrupt the continuity of time series. And so the interpretation of long-run movements needs to allow for the patchwork nature of the evidence.

The temporal patterns show some similarity in the United States and the United Kingdom, where inequality was considerably less in the 1940s than before the Second World War. It then moderately declined until the mid-1970s, when this trend abruptly reversed. But we have no consistent overall time series running this far back for other nations (see Gottschalk and Smeeding, 2000, Figs. 6a and 6b, for the longer-term US and UK trends). The best we can do on a reasonably comparative basis is shown in Figure 3, covering four decades from 1965 to 2005. (Estimates reflect national practices and are not to be compared across countries.) Indeed, the 1980s saw a substantial rise of inequality, more pronounced in Britain than in the United States, though the starting level was lower. In the 1990s the two nations parted: income distribution kept widening in the United States, while it broadly stabilized in the United Kingdom. Both Finland and Sweden experienced a fall in inequality until the early 1980s and then a modest rise afterwards, which has strengthened around the turn of the century. A tendency towards higher inequality followed by a period of stability seems to characterize the 1980s and the 1990s in the Netherlands and Norway as well as Australia and New Zealand. Canadian income inequality exhibited some variation but no clear trend from 1965 to the mid-1990s, when it started to slowly rise. In the Federal Republic of Germany a sharp fall between 1962 and 1973 was followed by a period of stability and a modest rise over the 1990s. Income distribution narrowed in Italy from the 1970s to the 1980s; after a sharp widening at the beginning of the 1990s, there was virtually no change until 2004. In France alone, inequality steadily decreased between 1970 and the mid-1990s, and remained stable afterwards.

Figure 3

Inequality trends in selected high-income economies (Gini index, per cent), 1965–2005. *Source*: Authors' elaboration on national sources.

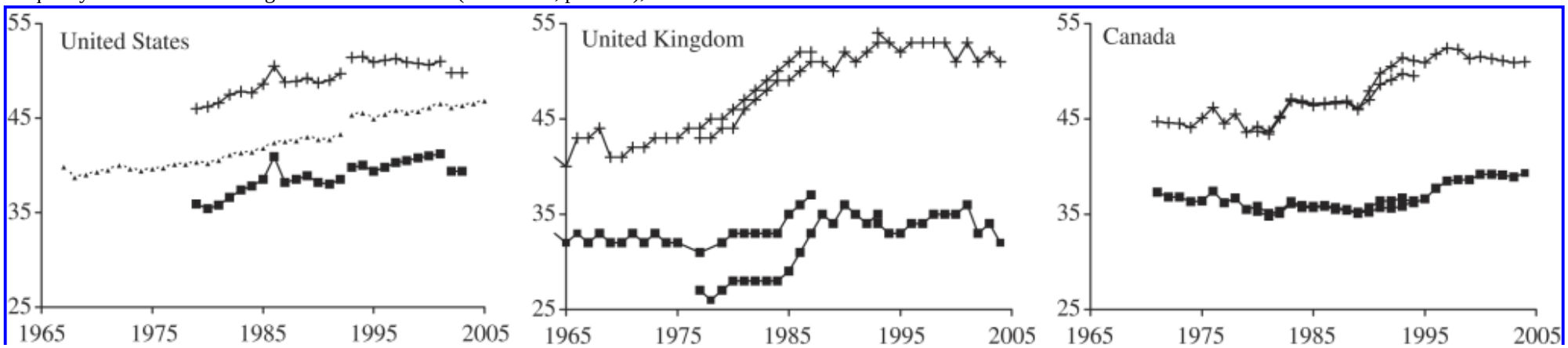

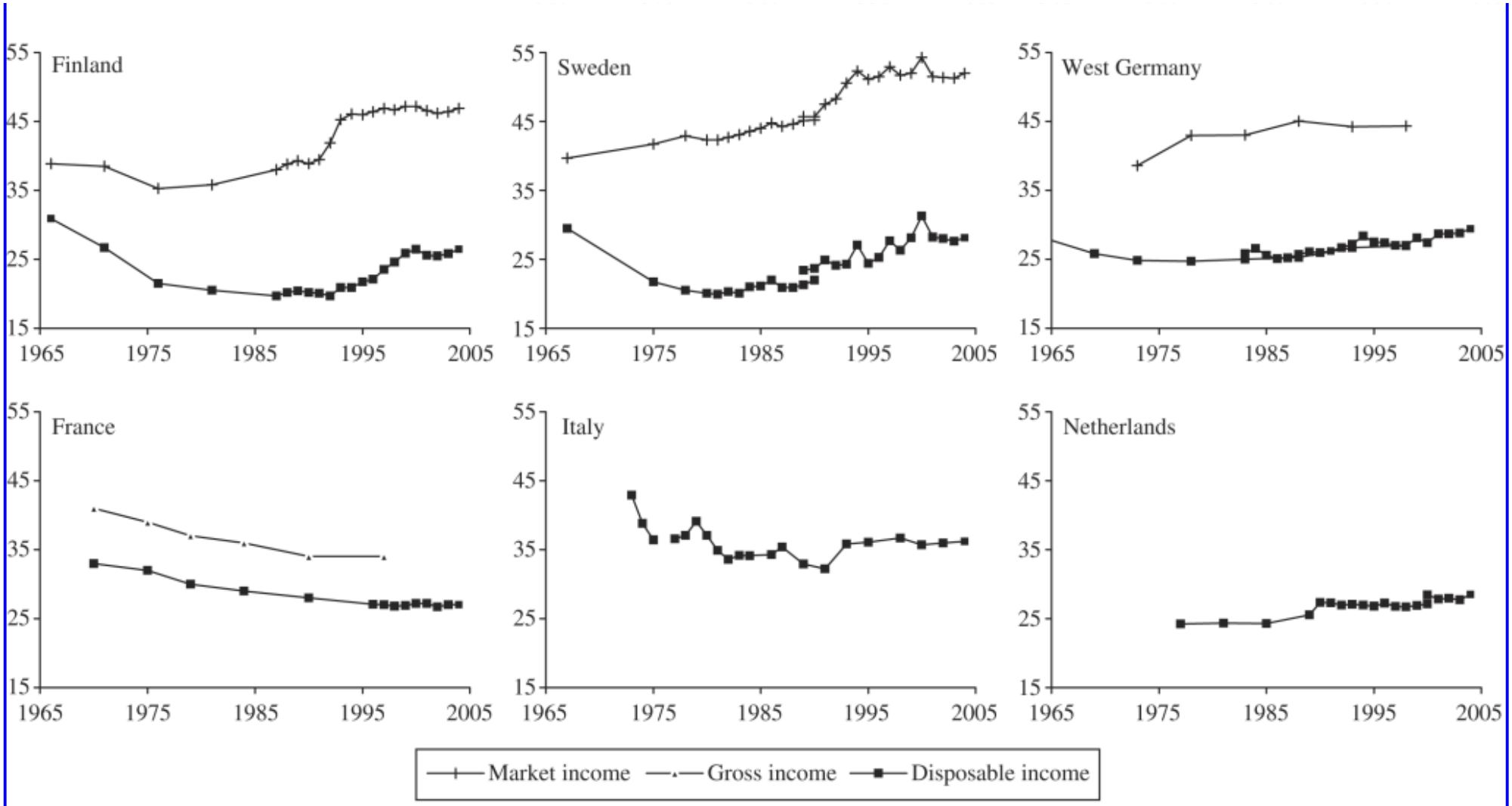

Source: Authors' elaboration on national sources.

In summary, national experiences vary and there is no one overarching common story. However, there was a general tendency for the disposable income distribution to narrow until the mid-1970s. Some increase in inequality was experienced by most nations in the 1980s to the 1990s, but its timing and magnitude differed widely across countries. In particular there was and is no regression to the mean pattern of change in the United States, which began with the most inequality in the late 1970s and has increasingly pulled away from the other nations through the early years of the 21st century.

These observations mainly relate to disposable incomes. In the six countries for which data are available (Canada, the Federal Republic of Germany, Finland, Sweden, the United Kingdom and the United States), movements in market income inequality appear to be more synchronous, with a rise in the 1980s followed by stability thereafter. Changing public redistribution appears to be an important determinant of the time pattern of the inequality of disposable incomes. If we take, as before, the absolute difference between Gini indices, the redistributive impact of taxes and transfers initially increased and then stabilized or dropped in all countries except for the United States, where it remained quite stable over time. The United Kingdom stands out for having the most dramatic switch of regime, as in the early 1980s it apparently shifted from a situation not too different from the two Nordic countries to a model closer to that of the two North American countries. It is not possible to infer from this simple measure whether changes in redistribution are the automatic response of a progressive tax-and-benefit system to changes in the distribution of market incomes, or are instead the product of explicit policy choices (Atkinson, 2004). Nevertheless, they confirm that a widening of the market income distribution need not result in a drastic increase in the inequality of disposable incomes.

See Also

- Gini ratio
- household surveys
- income mobility
- inequality (measurement)
- Lorenz curve
- Pareto distribution
- Pareto, Vilfredo
- redistribution of income and wealth
- survey data, analysis of

Bibliography

Atkinson, A.B. 1974. *The Economics of Inequality*. Oxford: Clarendon Press.

Atkinson, A.B. 1997. Bringing income distribution in from the cold. *Economic Journal* 107, 297–321.

Atkinson, A.B. 2004. Increased income inequality in OECD countries and the redistributive impact of the government budget. In *Inequality, Growth, and Poverty in an Era of Liberalization and Globalization*, ed. G.A. Cornia. Oxford: Oxford University Press.

Atkinson, A.B. and Brandolini, A. 2001. Promises and pitfalls in the use of secondary data-sets: income inequality in OECD countries as a case study. *Journal of Economic Literature* 39, 771–800.

Atkinson, A.B. and Micklewright, J. 1992. *Economic Transformation in Eastern Europe and the Distribution of Income*. Cambridge: Cambridge University Press.

Atkinson, A.B., Rainwater, L. and Smeeding, T.M. 1995. *Income Distribution in OECD Countries: The Evidence from the Luxembourg Income Study (LIS)*. Paris: OECD.

Atkinson, T., Cantillon, B., Marlier, E. and Nolan, B. 2002. *Social Indicators: The EU and Social Inclusion*. Oxford: Oxford University Press.

Bresciani-Turroni, C. 1939. Annual survey of statistical data: Pareto's law and the index of inequality of incomes. *Econometrica* 7, 107–33.

Buhmann, B., Rainwater, L., Schmaus, G. and Smeeding, T.M. 1988. Equivalence scales, well-being, inequality, and poverty: sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database. *Review of Income and Wealth* 34, 115–42.

Davies, J.B. and Shorrocks, A.F. 2000. The distribution of wealth. In *Handbook of Income Distribution*, ed. A.B. Atkinson and F. Bourguignon. Amsterdam: North-Holland.

Deaton, A. 2005. Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics* 87, 1–19.

European Commission. 2002. *European Social Statistics – Income, Poverty and Social Exclusion: 2nd Report*. Data 1994–1997. Luxembourg: Office for Official Publications of the European Communities.

Gottschalk, P. and Smeeding, T.M. 1997. Cross-national comparisons of earnings and income inequality. *Journal of Economic Literature* 35, 633–87.

Gottschalk, P. and Smeeding, T.M. 2000. Empirical evidence on income inequality in industrialized countries. In *Handbook of Income Distribution*, ed. A.B. Atkinson and F. Bourguignon. Amsterdam: North-Holland.

Immervoll, H., Levy, H., Lietz, C., Mantovani, D., O'Donoghue, C., Sutherland, H. and Verbist, G. 2005. Household incomes and redistribution in the European Union: quantifying the equalising properties of taxes and benefits. Working Paper No. EM9/05, EUROMOD.

International Monetary Fund. 2006. *World Economic Outlook, September 2006*. Washington, DC: International Monetary Fund. Online. Available at http://www.imf.org/external/pubs/ft/weo/2006/02/pdf/weo0906.pdf, accessed 1 April 2007.

Kravis, I.B. 1962. *The Structure of Income. Some Quantitative Essays*. Philadelphia: University of Pennsylvania Press.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45, 1–28.

Kuznets, S. 1963. Quantitative aspects of the economic growth of nations: 8. Distribution of income by size. *Economic Development and Cultural Change* 11, 1–80.

Lydall, H.F. 1979. Some problems in making international comparisons of inequality. In *Income Inequality: Trends and International Comparisons*, ed. J.R. Moroney. Lexington, MA: Lexington Books.

O'Higgins, M., Rainwater, L. and Smeeding, T.M. 1990. *Poverty, Inequality, and Income Distribution in Comparative Perspective: The Luxembourg Income Study (LIS)*. Washington, DC: Urban Institute Press, Wheatsheaf Books.

Pareto, V. 1897. *Cours d'économie politique*. Lausanne and Paris: Rouge & Pichon. Repr. in *Oeuvres Complètes*, ed. G.-H. Bousquet and G. Busino. Genève: Droz, 1964.

Sawyer, M. 1976. *Income distribution in OECD countries. OECD Economic Outlook. Occasional Studies*. Paris: OECD.

Sen, A.K. 1992. *Inequality Reexamined*. Oxford: Clarendon Press.

Sierminska, E., Brandolini, A. and Smeeding, T.M. 2006. The Luxembourg Wealth Study – a cross-country comparable database for household wealth research. *Journal of Economic Inequality* 4, 375–83.

Smeeding, T.M. 2004. Twenty years of research on income inequality, poverty, and redistribution in the developed world: introduction and overview. *Socio-Economic Review* 2, 149–63.

Smeeding, T.M., Saunders, P., Coder, J., Jenkins, S., Fritzell, J., Hagenaars, A.J.M., Hauser, R. and Wolfson, M. 1993. Poverty, inequality, and family living standards impacts across seven nations: the effect of noncash subsidies for health, education, and housing. *Review of Income and Wealth* 39, 229–56.

United Nations. 1951. *National income and its distribution in under-developed countries. Statistical Papers, Series E, No. 3*. New York: United Nations.

U.S. Census Bureau. 2005. *Income, Poverty, and Health Insurance Coverage in the United States: 2004*. Current population reports, P60–229.Washington, DC: U.S. Government Printing Office.

Wolff, E.N. 1996. International comparisons of wealth inequality. *Review of Income and Wealth* 42, 433–51.

World Bank. 2005. *World Development Report 2006. Equity and Development*. New York and Oxford: Oxford University Press.

How to cite this article

Brandolini, Andrea and Timothy M. Smeeding. "inequality (international evidence)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000273> doi:10.1057/9780230226203.0788

The New Palgrave Dictionary of Economics Online

inequality (measurement)

Frank A. Cowell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides an overview of the key issues in inequality measurement and shows how theoretical concepts are related to practical judgements. The principal axioms of distributional analysis are used to show the social-welfare underpinnings of standard ranking principles and to derive families of inequality indices. Recent developments that focus on income differences and reference income levels are examined.

Keywords

Atkinson inequality index; coefficient of variation; generalized Lorenz curve; Gini coefficient; index numbers; inequality (measurement); Lorenz curve; Pen's parade; poverty; poverty line; principle of transfers; ranking; relative deprivation; risk aversion; social-welfare function; Theil, H.; well-being

Article

1 Introduction

Inequality measurement is principally concerned with the comparison of personal income distributions in quantitative terms. In its modern form it is a branch of welfare economics although it clearly derives some of its intellectual heritage from statistics. It is distinct from the measurement of poverty and relative deprivation, although there are close analytical links to these topics. The motivation for taking the subject of inequality seriously is both analytical and practical: the principal concepts reviewed in this article are of concern to theoretical economists and are also used by policymakers. The subject touches on questions addressed by philosophers and by social scientists.

The type of issue under consideration can be illustrated by a simple example as depicted in Tables 1 and 2. These tables do not pretend to be the most general or the most suitable representation of the facts, but they are from an easily accessible source and give a convenient snapshot of what happened to the distribution of income in the United States over a span of about 30 years. From Table 1 it is clear that the bottom decile income experienced a 12.2 per cent growth over the period (in real terms) while the median grew by half as much again (18.3 per cent) and the top decile grew by almost four times as much (44.8 per cent). Table 2 describes what happened to the average incomes of particular *groups*. The average income of households in the bottom fifth of the distribution grew by just 10.1 per cent over the 30 years while the average income of households in the top fifth grew by 58.6 per cent. We return to the use of the concepts of quantiles and shares after introducing some of the technical equipment needed for analysing income distributions.

Quantile incomes and growth, United States 1974–2004

q	*q*-quantile		Growth
	1974	2004	
10%	$9,741	$10,927	12.2%
20%	$16,285	$18,500	13.6%

50%	$37,519	$44,389	18.3%
80%	$64,781	$88,029	35.9%
90%	$83,532	$120,924	44.8%
95%	$102,534	$157,185	53.3%

Note: Columns 2 and 3 give the upper limit of the bottom 10%, 20%,… of the population. Incomes are in 2004 dollars; the income-receiving unit is the household.*Source*: DeNavas-Walt, Proctor and Lee (2005, Appendix Table A3).

Growth in average incomes for the five quintile groups and overall. United States, 1974–2004

Group	Average income		Growth
	1974	2004	
1st	$9,324	$10,264	10.1%
2nd	$23,176	$26,241	13.2%
3rd	$37,353	$44,455	19.0%
4th	$53,944	$70,085	29.9%
Top	$95,576	$151,593	58.6%
Overall	*$43,875*	*$60,528*	*38.0%*

Note: Columns 2, 3 give the average incomes of the bottom fifth, second fifth,•…•. Incomes are in 2004 dollars; the income-receiving unit is the household.*Source*: as for Table 1.

The thumbnail sketch suggests a substantial increase in inequality in the United States over the last quarter of the 20th century. But how much did inequality increase? In what ways can the impressionistic method of inequality comparisons suggested in the example be made precise and interpreted within the context of standard economic analysis? The purpose of this article is to provide a succinct overview of the role played by economic theory and other abstract principles in this class of problem and how to make sense of inequality comparisons such as those suggested in the example.

The sketch example in Tables 1 and 2 also illustrates some of the essential practicalities that have to be taken into account when implementing the principles of inequality measurement. Should we be focusing on households or individuals? What is the appropriate definition of income?

To follow the analysis there are few prerequisites: an understanding of utility and preference analysis is helpful but not essential to grasping the basic points that will be discussed.

2 Basics

2.1 Components of the problem

The framework adopted here is not the most general approach, but one that is suitable for setting out the key ideas. We begin by considering the basic building blocks and then show how to assemble the constituent parts.

2.1.1 Income and income distribution

At the heart of the problem there is some scalar entity to be called ‘income’, but in practice this entity could be wealth, expenditure or some other economic quantity, the distribution of which is of particular interest. Income is distributed among a number of ‘income receivers’, which we will refer to as ‘persons’ (although the income receiver in practice may be a family or household). Suppose that there is a known number of income receivers n and that person i has income x_i. The *income distribution* is then simply the vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_n). \tag{1}$$

The set of all possible income distributions X is a subset of $\mathbb{R}^n$. The nature of X is going to depend in practice upon the precise definition of 'income': is it logically possible to have a zero value of x_i, for example? Or a negative value? As a working assumption we will take it that X consists of all vectors (1) such that $x_i \geq \underline{x}$ and leave open the specification of the lower bound $\underline{x}$ for particular instances of the inequality-measurement problem. Representations of the income distribution other than (1) will appear later in the discussion.

2.1.2 Indices

The topic of inequality measurement presumes that there is an inequality measure. An obvious interpretation of this is that there is some index I that, given a particular income distribution $\mathbf{x}$, yields a real number that is taken to be the amount of inequality exhibited by the distribution. In some ways the index I works like other well-known summary statistics of distributions, such as the mean

$$\mu(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

and the variance

$$\mathrm{var}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} [x_i - \mu(\mathbf{x})]^2. \tag{3}$$

Indeed, the variance itself is sometimes used as an inequality index, although it is more common to use a transformed version of it known as the *coefficient of variation*:

$$I_{\mathrm{CV}}(\mathbf{x}) := \frac{\sqrt{\mathrm{var}(\mathbf{x})}}{\mu(\mathbf{x})}. \tag{4}$$

One of the most commonly used indices in practice is the *Gini coefficient* defined as

$$I_{\mathrm{Gini}}(\mathbf{x}) := \frac{1}{2n^2\mu(\mathbf{x})}\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|. \tag{5}$$

There are many more. However, rather than running through an exhaustive list of candidate indices it is more useful to examine the principles that have usually been applied to

construct indices; this we do by considering a priori what constitutes a ‘suitable’ inequality measure, the issue addressed in Section 2.2.

2.1.3 Ranking and dominance

An apparently more flexible interpretation of the idea of inequality measurement is the idea of an inequality *ranking*. This is a partial ordering that picks up the general flavour of the kind of comparisons that we suggested in the introduction; the partial ordering is typically captured by a simple representational tool. Consider three of these.

The first of these tools is *Pen's parade* (named after the famous parable introduced by Pen, 1974, ch. 3), which is simply the inverse of the empirical distribution function. To depict it let $x_{[i]}$ denote the *i*th smallest component in the vector (1) – the *i*th smallest income. Then take the collection of points

$$\left(\frac{i}{n}, x_{[i]}\right), \quad i = 1, 2, \ldots, n. \tag{6}$$

From this simple definition we can also introduce the idea of dominance. Take two distributions $\mathbf{x}'$ and $\mathbf{x}''$ in X where $\mathbf{x}' = \left(x_1', x_2', \ldots, x_n'\right)$ and $\mathbf{x}'' = \left(x_1'', x_2'', \ldots, x_n''\right)$. If it is true that $x_{[i]}' > x_{[i]}''$ for all $i = 1, 2, \ldots, n$ then we say that $\mathbf{x}'$ strictly Parade-dominates $\mathbf{x}''$.

The resulting graph plots income quantiles against population proportions: $x_{[i]}$ is the quantile corresponding to the bottom q per cent of the population where $q = 100\frac{i}{n}$. To illustrate the concept we use the information in Table 1 to produce a graph that looks like Figure 1. In Pen's parable we imagine the whole population (seen as individuals rather than households) arranged in order on the [0,1] interval where each person's height has been altered in proportion to his/her income; the average-height income recipient in 1974 is located at position 0.57 in Figure 1 (in other words, at a point 57 per cent along the horizontal axis the height of the Parade is exactly mean income) but in 2004 the average-height income recipient is located at position 0.61. Although the distribution of 2004 Parade-dominates the distribution in 1974, it is clear from Table 1 that overall the Parade shifted upwards in a lopsided fashion over the 30 years with the incomes of the very rich (95 per cent quantile) growing more than four times faster than those of the poor (10 per cent quantile); this shift suggests increased inequality over the period. However, by itself the Parade does not tell us much about inequality directly, although concepts closely related to it are widely used to characterize inequality comparisons. It is common to use *quantile ratios* for distributional comparisons: for example the popular ‘90–10 ratio’ is given by $x_{[k]}/x_{[j]}$ where j and k are, respectively, the smallest integers satisfying $j/n \geq 10\%$ and $k/n \geq 90\%$: in the example above this ratio increased from 8.6 to 11.1. Furthermore, there is an important welfare-economic interpretation of the Parade that is discussed in Section 3.3 below.

Figure 1

Parade diagram corresponding to Table 1. *Source*: As for Table 1.

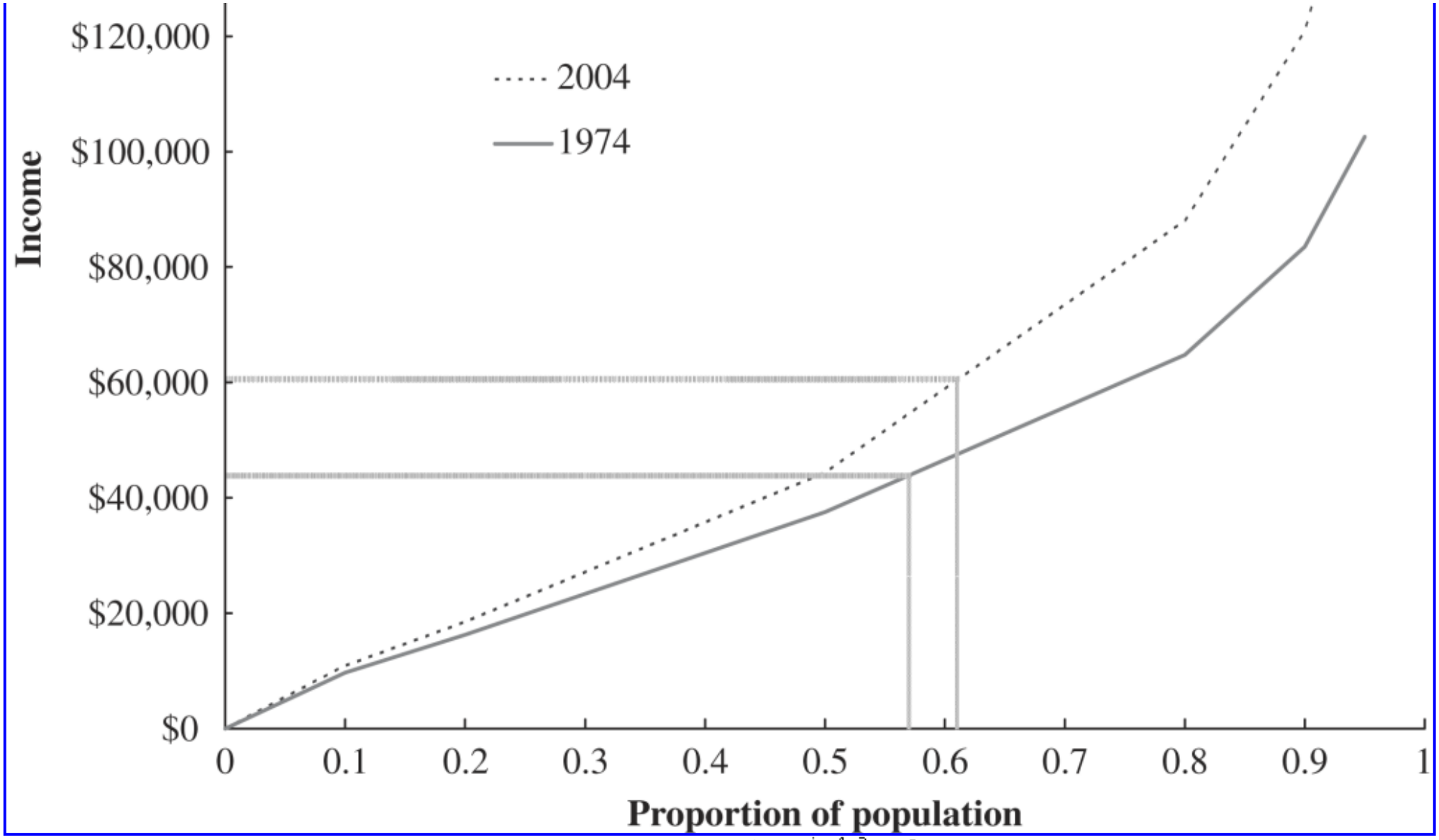

For the second and third concepts we use the $x_{[i]}$ to derive the normalized income cumulations; for any $i = 1, 2, \ldots, n$ these are

$$c_i := \frac{1}{n}\sum_{j=1}^{i} x_{[j]}. \tag{7}$$

Then the *generalized Lorenz curve* (GLC) is given by the graph of

$$\left(\frac{i}{n}, c_i\right), i = 1, 2, \ldots, n. \tag{8}$$

Again we have a natural definition of dominance: for two distributions $\mathbf{x}'$ and $\mathbf{x}''$ in X if it is true that $c_i' > c_i''$ for all $i = 1, 2, \ldots, n$ then we say that $\mathbf{x}'$ strictly GLC-dominates $\mathbf{x}''$. For the example we used earlier the GLC is illustrated in Figure 2, derived from Table 2. (Note that the definitions of Parade- and GLC-dominance can be extended to cases where the two distributions do not have the same number of incomes – this step makes use of the 'population principle' defined in Section 2.2. In some cases it is useful to consider the weak – non-strict – versions of the dominance criteria introduced here.)

Figure 2

Generalized Lorenz curve. *Source*: As for Table 1.

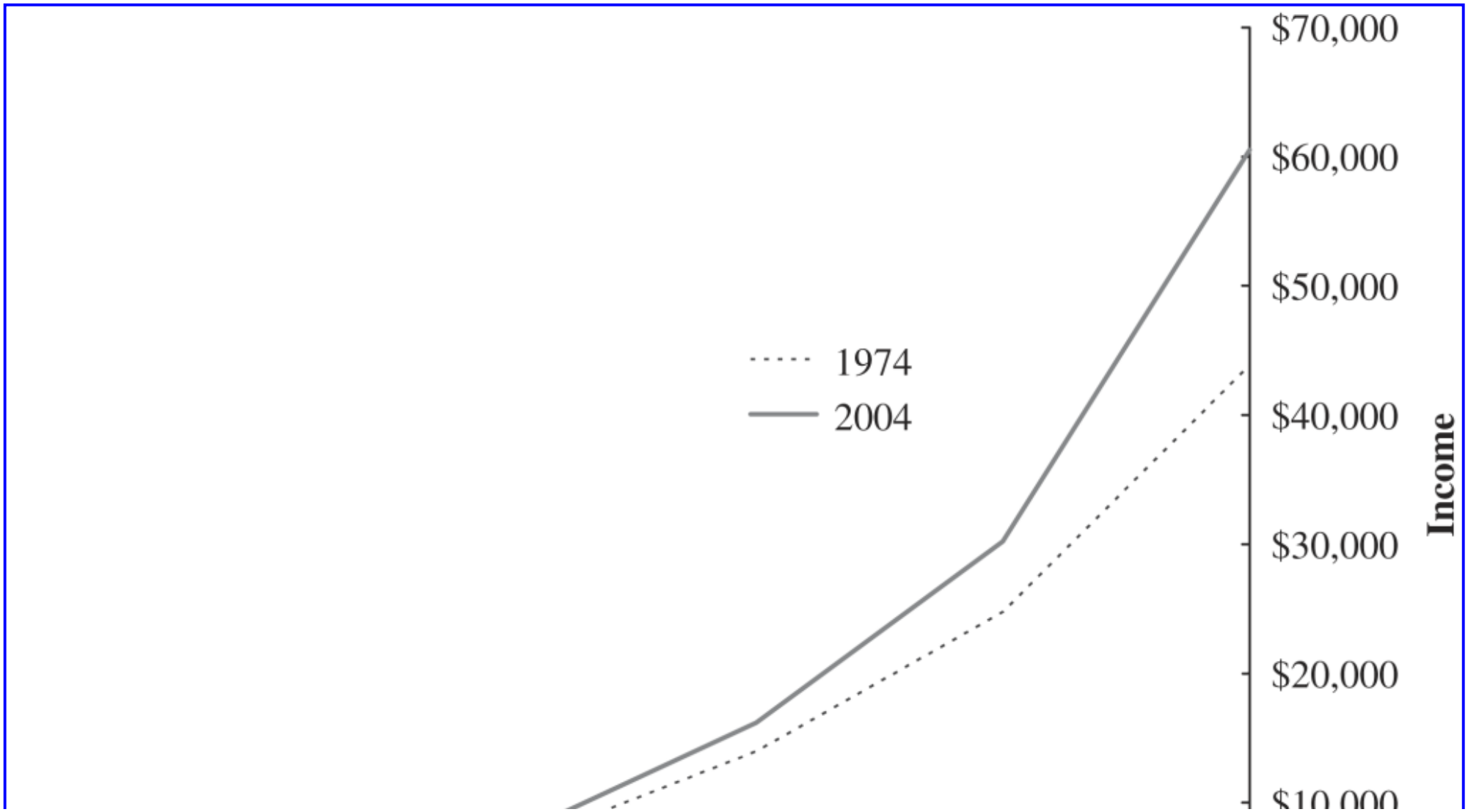

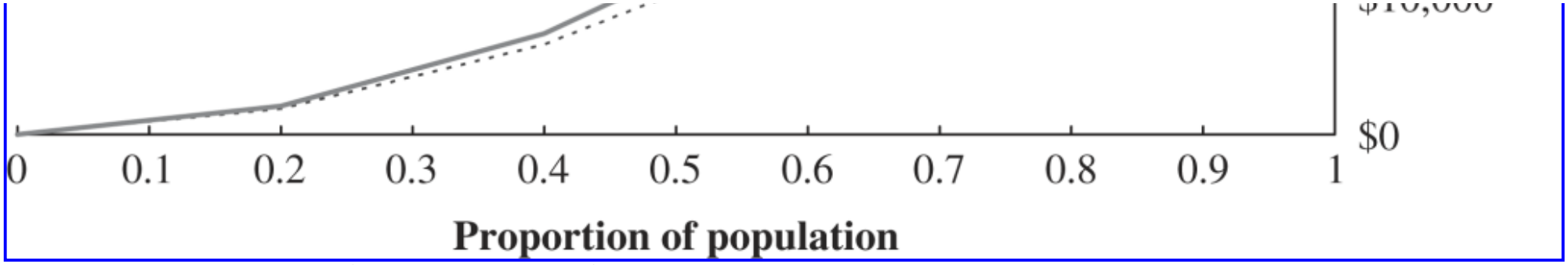

The GLC plots the normalized income of the bottom 100*q* per cent of the population against *q* and, although the 2004 distribution GLC-dominates 1974, it is clear that over the period the growth of these group averages was not evenly distributed – the higher was *q*, the higher was the growth over 1974 to 2004. (This is easily inferred from Table 2: for example, the average income of the top 20 per cent grew almost six times as fast as the average income of the bottom 20 per cent.) Once again, although the GLC does not give information about inequality comparisons directly, there is an important welfare-economic interpretation (in Section 3.3). In addition, a small modification of the GLC yields one of the central concepts of distributional analysis. Dividing c_i in (7) by the mean $\mu(\mathbf{x})$ gives the income share of the bottom $100\frac{i}{n}$ per cent of the population. The graph of the (population-proportion, income-share) pairs

$$\left(\frac{i}{n}, \frac{c_i}{\mu(\mathbf{x})}\right), \quad i = 1, 2, \ldots, n \tag{9}$$

gives the *Lorenz curve*. Also, for two distributions $\mathbf{x}'$ and $\mathbf{x}''$, if it is true that, $c_i' / \mu(\mathbf{x}') > c_i'' / \mu(\mathbf{x}'')$ for all $i = 1, 2, \ldots, n-1$ then we say that $\mathbf{x}'$ strictly Lorenz-dominates $\mathbf{x}''$. In the case of the example using US data this is illustrated in Figure 3: the Lorenz curve plots the income share of the bottom 100*q* per cent of the population against *q* and the diagonal line depicts a hypothetical distribution of perfect equality. (Take the area trapped between the Lorenz curve and the equality diagonal. Using (7) and (9) we can show that the ratio of this area to the area of the whole triangle is given by the weighted sum $\sum_{i=1}^{n} \kappa_i x[i]$ where the weights are $\kappa_i = [2i - 1 - n] / [n\mu(\mathbf{x})]$. This is exactly the Gini coefficient (5).) It is clear that for each *q* the share was smaller in 2004 than it was in 1974 – the 1974 distribution Lorenz-dominates that for 2004. This simple intuitive notion of greater inequality conforms exactly with a fundamental principle to be explained below.

Figure 3

Lorenz curve. *Source*: As for Table 1.

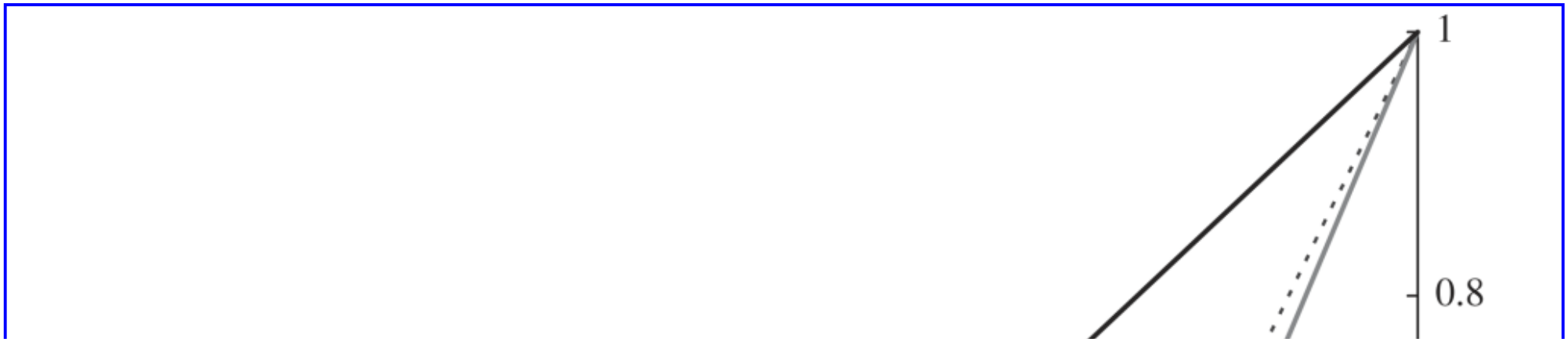

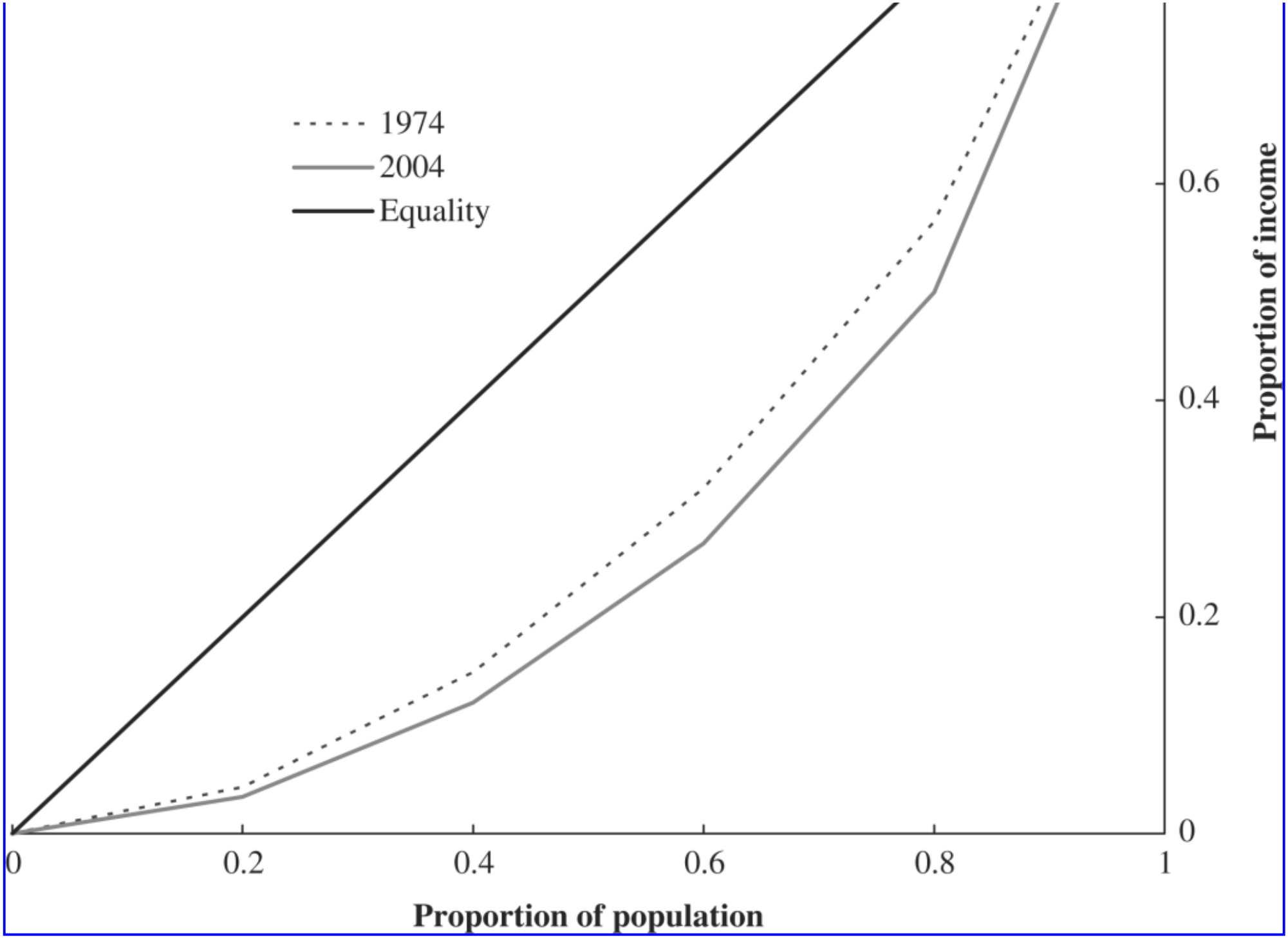
1974
2004
Equality
0.6
0.4
0.2
0
Proportion of income
0
0.2
0.4
0.6
0.8
1
Proportion of population

2.2 Axioms

An inequality index *I* is in some ways like a utility function in consumer theory: it is a representation of an inequality ordering on the members of *X* and is usually taken to be continuous and ordinal – although there is often a 'natural' cardinal representation of a particular index, a formal argument for one representation rather than another is not usually provided (why not use the square or the log of the Gini coefficient?). Ordinality is sufficient for making comparing income distributions, the primary task of inequality analysis. Axioms are essentially formal statements of the principles of assessment that are used to give meaning to the ordering represented by *I*. The treatment here does not claim to generality; rather, it focuses on those principles that are central to modern approaches to inequality. Rather than presenting the axioms as formal statements, however, it is more useful here to introduce the underlying key principles discursively.
Assume that everywhere in the following discussion the vector **x** in (1) is any arbitrary member of the set *X*.

- First, it seems reasonable that the labelling of the components of **x** be irrelevant: it does not matter which income receiver gets which income. This means that *I* has the *symmetry* property:

$$I(x_1, x_2, \ldots, x_n) = I(x_2, x_1, \ldots, x_n) = I(x_3, x_1, \ldots, x_n) = \ldots \tag{10}$$

- We will always assume that this holds and we may therefore adopt the convention that incomes have been labelled such that $x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n$.
- Second, we need some coherent way of characterizing inequality in different-sized populations. Perhaps the most obvious assumption is that simple replications of an income vector (1) leave inequality unchanged. This is the *population principle*:

$$I(x_1, x_2, \ldots, x_n) = I(x_1, x_1, x_2, x_2, \ldots, x_n, x_n) = I(x_1, x_1, x_1, x_2, x_2, x_2, \ldots, x_n, x_n, x_n) = \ldots \tag{11}$$

  Taken in conjunction with symmetry this allows one to represent distributions purely in terms of a distribution function.
- A key assumption that is commonly invoked focuses on the effect on inequality of a hypothetical small income transfer. Suppose $x_i<x_j$ and consider some positive number δ such that $x_i - \delta \geq \underline{x}$, then the *principle of transfers* (Dalton, 1920) requires that:

$$I(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n) < I(x_1, \ldots, x_i - \delta, \ldots, x_j + \delta, \ldots, x_n) \tag{12}$$

  - – a poorer-to-richer income transfer will always increase inequality.
- • As a counterpart to the assumption relating to different sizes of population (eq. 11) it is useful to have an assumption relating to different amounts of total income. The standard assumption is that of *scale independence*. This requires that, for any scalar $\lambda > 0$:

$$I(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) = I(x_1, x_2, \ldots, x_n) \tag{13}$$

- – double all incomes or halve all incomes and inequality is left unaltered.
  An alternative assumption that is sometimes used is *translation independence*. Take any real number δ such that $x_1 + \delta \geq \underline{x}$; then

$$I(x_1 + \delta, x_2 + \delta, \ldots, x_n + \delta) = I(x_1, x_2, \ldots, x_n) \tag{14}$$

- – add or subtract one dollar from every income and inequality is left unaltered.

Clearly this brief list raises some important questions. Why use these particular axioms? Some of them appear to be quite strong; for example, although scale independence seems attractive if the 'incomes' x_i here are measured in dollars and we consider just dividing through by some rate of exchange so as to work with incomes in some other monetary units, it may seem less attractive if we want to consider the impact on inequality of redistribution policies at different stages of economic growth: a rearrangement of income shares that constitutes a reduction in inequality in a low-income society might not be considered as a reduction in inequality if the whole population is prosperous. Furthermore, the axioms captured by eqs (10)–(13), for example, are satisfied by both (4) and (5) as well as other important classes of inequality measures; on the other hand, the axioms captured by eqs (10)–(12) and (14) are satisfied by (3) and another rich class of inequality measures. Following on from this question, what more is required to get a specific index or well-defined family of indices that is both theoretically appropriate and practical to implement?
To answer this we need to be precise about what it means to say that one distribution is more unequal than another and the intellectual basis used for making such comparisons. The meaning of inequality can be further clarified through one of several routes: this article will analyse three of these in turn, namely, social welfare, decomposition, income differences.

3 Social welfare and inequality

The welfare-economic approach to the subject starts from the position that inequality is about 'illfare' – the opposite of welfare. If we adopt this approach then the definition of inequality follows almost immediately. The idea is similar to the conventional measurement of economic waste and the basis for a simple model can be laid with only a little more theorizing.
The *social-welfare function* (SWF) is a real-valued function W defined on the space of distributions X. The social welfare associated with a particular income distribution (1), given by

$$W(x_1, x_2, \ldots, x_n), \tag{15}$$

is to be interpreted as follows: suppose we are given a specific SWF $W(\cdot)$ and that for two separate income distributions $\mathbf{x}'$ and $\mathbf{x}''$ we have $W(\mathbf{x}') > W(\mathbf{x}'')$; then social welfare associated with the distribution $\mathbf{x}'$ is higher than the social welfare associated with the distribution $\mathbf{x}''$. In principle W is an ordinal function so that the scale of measurement of welfare levels can be subjected to arbitrary monotonic-increasing transformations.
This basic specification raises a number of important questions:

- Why express social welfare as a function of income? Income defined how?
- What particular form should W take?
- What is the relation between the functions I and W?

The answer to the first question helps to pin down the relationship between inequality measurement as conventionally practised and standard welfare economics – see Section 1. The answers to the last two questions will determine the form of a class of inequality measures and permit us to establish some important welfare-economic results: these are addressed in

Sections 2 and 3.

3.1 Welfare and income

We need to rectify a point that was fudged in the discussion of the US example: how to do the trick of passing from a distribution of dollar income among households to a standard welfare analysis that is typically concerned with the levels of economic well-being of individuals. The standard approach is as follows. We require a method of appropriately capturing the relationship between the living standard that is attainable by an individual and the income that he/she is presumed to have access to within the household. This is conventionally done by defining a function $\nu(\cdot)$ that has as its argument a list of non-income attributes **a** that might include household size, age and sex of household members and health status; $\nu(\mathbf{a})$ determines the number of *equivalent adults* in the household with attributes **a** such that

$$x = \frac{y}{\nu(\mathbf{a})}, \tag{16}$$

where y is nominal income and x is *equivalized income* that is taken to be comparable across different household types. Note that the equivalization function ν is typically specified as independent of income although this simplification is not essential; of course, the way in which the function ν is determined – from ethical considerations or econometric studies – is an important issue in its own right, but one that lies outside the present discussion. The function ν transforms a distribution of dollar incomes among n households

$$\mathbf{y} = (y_1, y_2, \ldots, y_n) \tag{17}$$

into a distribution of equivalized incomes by households given by (1). In order to complete the welfare interpretation we need to recognize that social-welfare considerations are usually represented in terms of individuals rather than households, and so, for example, households consisting of couples should receive more weight in social-welfare evaluations than households consisting of single individuals. Therefore, if the income-receiving units consist of households of differing size, we might want to represent this by introducing a corresponding set of population weights w_i for the observations, so that the distribution becomes an ordered list of pairs:

$$((w_1, x_1), (w_2, x_2), \ldots, (w_n, x_n)) \tag{18}$$

where w_i is the number of persons in household i divided by the number of persons in the whole population. There is little analytical complication in using (18) rather than (1) as a representation of the distribution of equivalized incomes by individuals. Typically it is just a matter of a minor redefinition of formulas for inequality measures and the like: for example, the coefficient of variation (4) would now be written

$$\sqrt{\sum_{i=1}^{n} w_i \left[\frac{x_i}{\mu} - 1\right]^2} \tag{19}$$

where μ is the appropriately redefined mean $\sum_{i=1}^{n} w_i x_i$. (More generally: all measures that can be written in the form $\Phi(\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i), \mu)$ just need to be rewritten in the form $\Phi(\sum_{i=1}^{n} w_i\varphi(x_i), \mu)$. A similar modification applies to the Gini coefficient.)
However, having introduced this important theoretical qualification we will now neglect it – for expositional purposes it is convenient to assume (*a*) that the population consists of isolated individuals that are identical in every relevant respect other than income and (*b*) that income appropriately represents individual welfare. So, from here on, *i* indexes individuals or households and the distinction between *x* and *y* is dropped.

3.2 Social welfare and inequality measures

The idea of the SWF was introduced without discussing specific properties of the function *W*. Some properties must be imposed on *W* if we require there to be a specific relationship between social welfare and inequality and we impose specific assumptions on the function *I*. However, in addition it is particularly important to be explicit about how *W* should respond to an increase in one or more incomes. This is the usual principle that is applied:

- Suppose we consider any income distribution $(x_1, x_2, \ldots, x_n)$ and some positive number δ. Then *monotonicity* requires that:

$$W(x_1, x_2, \ldots, x_i + \delta, \ldots, x_n) > W(x_1, x_2, \ldots, x_i, \ldots, x_n) \tag{20}$$

On the assumption that monotonicity holds and that *W* is a continuous function, the SWF can itself be used to derive a family of inequality measures. There are several ways of doing this, but a standard approach is to represent social welfare using a money metric: we can always do this in view of the ordinal nature of *W* and the requirement that it be monotonic and continuous. The *equally distributed equivalent* (EDE) income is a real number ξ such that for any $(x_1, x_2, \ldots, x_n)$ in *X*:

$$W(\xi, \xi, \ldots, \xi) = W(x_1, x_2, \ldots, x_n). \tag{21}$$

(Note that monotonocity is unnecessarily strong for this step: for example one could define ξ in cases where one required only that *W* is increasing if *all* incomes are increased by δ, not just if *some* income is increased by δ. However, the assumption of monotonicity is useful for other results that follow.)
Clearly the relationship (21) can be used to derive EDE as a function of the income distribution, $\xi(\mathbf{x})$ and the function $\xi(\cdot)$ is a valid way of representing social welfare.
Suppose we require that the principle of transfers apply to *W*; this by analogy with (12) means that a mean-preserving poorer-to-richer income transfer will *decrease* social welfare.
Then it is always true that $\xi(\mathbf{x}) \leq \mu(\mathbf{x})$ and the normalized gap between ξ and μ provides a natural basis for an inequality index

$$1 - \frac{\xi(\mathbf{x})}{\mu(\mathbf{x})}. \tag{22}$$

It is clear that this index is bounded between zero and 1 and that if there were perfect equality then we would have $\xi(\mathbf{x}) = \mu(\mathbf{x})$ and inequality in (22) would be zero. Furthermore, if the scale-independence property (13) is also satisfied, then EDE income takes the form of a generalized mean:

$$\xi(\mathbf{x}) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}, \ \varepsilon > 0 \tag{23}$$

and (22) gives the class of *Atkinson indices*:

$$I_{\mathrm{A}}^{\varepsilon}(\mathbf{x}) := 1 - \left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{\mu(\mathbf{x})}\right]^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}. \tag{24}$$

(The limiting forms of (23) and (24) as $\varepsilon \to 1$ are, respectively, $\xi(\mathbf{x}) = \exp(\frac{1}{n}\Sigma_{i=1}^{n}\log(x_i))$ and $I_{\mathrm{A}}^{1}(\mathbf{x}) = 1 - \exp(\frac{1}{n}\Sigma_{i=1}^{n}\log(x_i)) / \mu(\mathbf{x})$.)

The number ε – the degree of (relative) inequality aversion – is a parameter that characterizes individual members of the class of inequality measures. For any given unequal income distribution, the larger is ε the larger is the Atkinson inequality index – there is an example of this in Table 3 above. There is a close analogy with a class of risk indices in the case of constant relative risk aversion. This is unsurprising since this approach was explicitly founded on the formal similarity between distributional comparisons in terms of inequality and of risk (Atkinson, 1970).

Inequality indices for the example in Table 1

	1974	2004
$I_{\mathrm{A}}^{0.25}$	0.067	0.097
$I_{\mathrm{A}}^{0.5}$	0.134	0.190
$I_{\mathrm{A}}^{0.75}$	0.207	0.286
I_{A}^{1}	0.297	0.418
I_{Gini}	0.395	0.466
I_{GE}^{0}	0.352	0.542
I_{GE}^{1}	0.267	0.406

If, instead of the scale-independence property, we required *I* to satisfy translation independence (14), then we would obtain a different class of indices

$$I_{\mathrm{K}}^{\beta}(\mathbf{x}) := \frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\beta[x_i - \mu(\mathbf{x})]}\right)$$

(25)

where $\beta > 0$ is a sensitivity parameter indexing members of the class (Kolm, 1976). The connection of (25) with constant absolute risk aversion is evident. (One could also use 'regularity' assumptions other than scale- or translation-independence – see Bossert and Pfingsten, 1990.)

3.3 Ranking distributions

As noted earlier, there are important results available about welfare and inequality comparisons that do not require the usage of specific indices. They follow from standard first- and second-order dominance results that are familiar from finance and other disciplines. Take the special class of *additive* welfare functions where *W* in (15) can be written in the form $\sum_{i=1}^{n} u(x_i)$ for some function *u*. If *W* is additive and satisfies the monotonicity axiom, then *u* must be a strictly increasing function; if, furthermore, *W* satisfies the principle of transfers then *u* must be strictly concave. Then the following powerful results are available for any two distributions $\mathbf{x}'$ and $\mathbf{x}'' \in X$:

- *First-order*: $\mathbf{x}'$ strictly Parade-dominates $\mathbf{x}''$ if and only if $W(\mathbf{x}') > W(\mathbf{x}'')$ for any additive *W* that satisfies the principle of monotonicity.
- *Second-order*: $\mathbf{x}'$ strictly GLC-dominates $\mathbf{x}''$ if and only if $W(\mathbf{x}') > W(\mathbf{x}'')$ for any additive *W* that satisfies monotonicity and the principle of transfers (Shorrocks, 1983).

(From the statement of these results it is clear that the Parade-dominance and GLC-dominance criteria are formally equivalent to first-order and second-order stochastic dominance in the analysis of probability distributions – see stochastic dominance.)
A version of the second-order result applies to the conventional Lorenz curve and it accords with the intuitive argument presented in the introduction. Take the class of SWFs that satisfy the principle of transfers (they do not have to be additive). Then, for two distributions $\mathbf{x}'$ and $\mathbf{x}''$ that have the same mean, the statement '$W(\mathbf{x}') > W(\mathbf{x}'')$ for any *W* in this class' is true if and only if $\mathbf{x}'$ strictly Lorenz-dominates $\mathbf{x}''$. Furthermore, under these circumstances for any inequality index *I* that satisfies the principle of transfers it must be the case that $I(\mathbf{x}') < I(\mathbf{x}'')$. The implication of this is that all inequality measures that satisfy the principle of transfers 'go the same way' if one distribution Lorenz-dominates the other. This is illustrated in Table 3 (which again uses the distribution of household income by households). Rows 1 to 4 give the results for the Atkinson indices: notice that in each case measured inequality is closer to 1 (the maximum) the higher is the degree of inequality aversion. The indices in the last two rows of Table 3 are discussed in the next section.

4 Decomposition

The axioms discussed in Section 2.2 induced some structure on inequality measures. By introducing the idea of *decomposing* inequality we can impose more structure and thereby obtain a useful class of indices. There are two principal types of decomposition: by subgroups of the population (regions, age groups,…) and by components of income (labour income, income from capital,…). Here we focus just on the population-subgroup issue.
Imagine that the population of *n* persons can be partitioned into a collection of *m* groups so that any individual falls into just one of these *m* groups. Each group *j* could be considered as a sub-population of size n_j in its own right (where $\sum_{j=1}^{m} n_j = n$) and one could compute inequality within this subpopulation as

$$I_j = I(\mathbf{x}_j) \tag{26}$$

where $\mathbf{x}_j$ is the income distribution consisting of just the members of subgroup *j*. The essence of the decomposition problem is to represent inequality overall as a function of inequality in each group $j = 1, \ldots, m$

$$I(\mathbf{x}) = F(\iota_1, \iota_2, \ldots, \iota_m; \pi_1, \ldots, \pi_m, s_1, \ldots, s_m) \tag{27}$$

where F is an aggregation function and the terms after the ';' show that aggregation may depend on the groups' shares of the population $\pi_j := n_j / n$ and the groups' shares of total income $s_j := n_j \mu(\mathbf{x}_j) / n\mu(\mathbf{x})$. A consistency requirement on (27) is that, if the income distribution within subgroup j changes so as to increase ι_j in (26), all other things remaining the same, then inequality overall should increase. Insisting on this requirement on F for all logically possible partitions induces a type of separability on the function $I(\cdot)$ so that the index must be of the general form mentioned just after eq. (19) above. If we also require that scale-independence hold, then the inequality index must take the specific form

$$I^{\alpha}_{\mathrm{GE}}(\mathbf{x}) = \frac{1}{\alpha^2 - \alpha}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{\mu(\mathbf{x})}\right]^{\alpha} - 1\right] \tag{28}$$

or some monotonic transform of it, where α is a real number. The 'GE' used in the labelling of (28) stands for the *generalized entropy* class, which is a generalization of the two indices introduced by Theil (1967). (Theil's two indices are those corresponding to the special forms in the cases $\alpha = 0, 1: I^{0}_{\mathrm{GE}}(\mathbf{x}) := -\frac{1}{n}\Sigma_{i=1}^{n}\log(x_i / \mu(\mathbf{x}))$ and $I^{1}_{\mathrm{GE}}(\mathbf{x}) = \frac{1}{n}\Sigma_{i=1}^{n}[x_i / \mu(\mathbf{x})]\log(x_i / \mu(\mathbf{x}))$. The values of these indices for the US example are given in the last two rows of Table 3.) The α in (28) is a parameter that characterizes different members of the GE class: a high positive value of α yields an index that is very sensitive to income transfers at the top of the distribution; specifying a negative value will produce an index that is sensitive to income transfers among the poor. (There is a functional relationship between the class (24) and the class (28). For any $\alpha < 1$ we have $I^{\varepsilon}_{\mathbf{A}}(\mathbf{x}) = 1 - [1 + \alpha[\alpha - 1]I^{\alpha}_{\mathrm{GE}}(\mathbf{x})]^{1/\alpha}$ where $\varepsilon = 1 - \alpha$.)

5 Income differences

The third way forward from the basic argument outlined in Section 2.2 focuses on fundamental income differences. This is one of the key ways in which one can motivate usage of the very well-known inequality indices mentioned in Section 2.1.2. The variance and the coefficient of variation (4) can be thought of as a representation of the averaged squared difference between each income x_i and the mean. A compelling argument for the Gini coefficient is that it is the (normalized) expected value of the absolute difference between any two randomly selected incomes in the population.

However, there are other types of income difference that are of special relevance to inequality measurement. Just as some poverty indices can be characterized as a kind of average distance of individual incomes from a reference income level – the poverty line (many poverty indices can be written in the form $\frac{1}{n}\Sigma_{i=1}^{n} p(z - x_i)$ where z is the poverty line and $p(\cdot)$ is a non-decreasing function that is zero for all $x_i \geq z$) – so also some inequality measures use the idea of a reference level income. In the case of inequality the reference income level has been suggested as either that of the best-off person in society, or of the average income of all those who are better off than any given person i (Temkin, 1993). In each of these cases application of standard axioms about the structure of inequality orderings leads to a class of inequality indices that bears a functional similarity to poverty indices and to indices of relative deprivation (Cowell and Ebert, 2004).

6 Implementation

The practical issues associated with the exposition of the example in Tables 1 and 2 highlight some of the problems in implementing inequality measures and associated tools – the definition of income, income receiver, and so on. Given the way in which income data are usually obtained, issues of sampling and measurement error usually need to be treated

carefully. Furthermore, the special nature of income and wealth distributions and the sensitivity of inequality indices to very high or very low incomes usually require that particular attention be paid to the problem of outliers. Finally, it should be noted that it is still sometimes the case that the data required for estimating inequality indices are made available only in grouped form rather than as microdata so that special techniques may be required for interpolation within income intervals and for modelling the tails of the distribution.

7 Further reading

For the welfare-economic issues, see Atkinson (1983) and Sen and Foster (1997). For literature surveys see Cowell (2000; 2007) and Lambert (2001).

See Also

- poverty
- stochastic dominance

My thanks go to Yoram Amiel, Tony Atkinson, Sanghamitra Bandyopadhyay, Kristof Bosmans, Udo Ebert, Giovanni Ko, Peter Lambert and Abigail McKnight, who made helpful comments on an earlier draft.

Bibliography

Atkinson, A.B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 244–63.

Atkinson, A.B. 1983. *The Economics of Inequality*, 2nd edn. Oxford: Clarendon Press.

Bossert, W. and Pfingsten, A. 1990. Intermediate inequality: concepts, indices and welfare implications. *Mathematical Social Science* 19, 117–34.

Cowell, F.A. 2000. Measurement of inequality. In *Handbook of Income Distribution*, ed. A.B. Atkinson and F. Bourguignon. Amsterdam: North Holland.

Cowell, F.A. 2007. *Measuring Inequality*, 3rd edn. Hemel Hempstead: Oxford University Press.

Cowell, F.A. and Ebert, U. 2004. Complaints and inequality. *Social Choice and Welfare* 23, 71–89.

Dalton, H. 1920. Measurement of the inequality of incomes. *Economic Journal* 30, 348–61.

DeNavas-Walt, C., Proctor, B.D. and Lee, C.H. 2005. U.S. Census Bureau, Current Population Reports, P60–229, *Income, Poverty, and Health Insurance Coverage in the United States: 2004*. Washington, DC: U.S. Government Printing Office. Online. Available at http://www.census.gov/prod/2005pubs/p60-229.pdf, accessed 24 November 2006.

Kolm, S.-C. 1976. Unequal inequalities I. *Journal of Economic Theory* 12, 416–42.

Lambert, P.J. 2001. *The Distribution and Redistribution of Income*, 3rd edn. Manchester: Manchester University Press.

Pen, J. 1974. *Income Distribution*, 2nd edn. London: Allen Lane, Penguin Press.

Sen, A.K. and Foster, J.E. 1997. *On Economic Inequality*, 2nd edn. Oxford: Clarendon Press.

Shorrocks, A.F. 1983. Ranking income distributions. *Economica* 50, 3–17.

Temkin, L.S. 1993. *Inequality*. Oxford: Oxford University Press.

Theil, H. 1967. *Economics and Information Theory*. Amsterdam: North Holland.

How to cite this article

Cowell, Frank A. "inequality (measurement)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000269> doi:10.1057/9780230226203.0789

The New Palgrave Dictionary of Economics Online

inequality between nations

François Bourguignon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Inequalities between nations are large but determining their magnitude and whether they have increased or decreased over time is not a simple matter. The conclusions that one reaches depend on the concept of inequality that one uses and on the point of view that one adopts.

Keywords

child mortality; educational attainment; Gini ratio; global inequality; globalization; household surveys; income mobility; inequality between nations; inter-country inequality; international inequality; nutrition; power; purchasing power parity; value judgements; voice; well-being

Article

It is useful to distinguish three different concepts of world inequality (see Milanovic (2005), Bourguignon, Levin and Rosenblatt, 2004). The *inter-country* distribution measures the level of inequality across representative citizens of each country in the world. This is a distribution of unweighted gross national income (GNI) per capita. The *international* distribution uses country GNI per capita *weighted* by their population size: it measures the inequality in the distribution of the world's citizens if each citizen were assigned the average income of the country in which he or she resides, adjusted for purchasing power parity, instead of his or her own income. Finally, there is the *global* distribution of individual incomes. This third concept lines up all citizens of the world (not countries) and calculates the distribution of their actual incomes, adjusted for purchasing power parity. *Global* inequality can be decomposed into inequality attributable to inequalities *within country* – that is, among persons within each country – and the differences of mean income *between countries*, that is *international* inequality.
The first section of this article describes the evolution of world inequality according to these various definitions and shows that they lead to somewhat contradictory conclusions, with inter-country inequality rising more or less continuously since the 1950s, international inequality declining, and global inequality increasing until the late 1980s and falling somewhat afterwards. The reason for these differences is easily understood, and has to do with population weights and the role played by giant developing countries like India or China and, to a lesser extent, with the evolution of inequality within countries. The second section tries to reconcile these various views on the evolution of world inequality by considering the mobility of world citizens within the income scale – this is similar to watching a movie rather than photographs of the world distribution of income at various points of time. The final section extends the income framework by providing some information on the evolution of world inequalities in a few non-income dimensions.

Evolution of world inequality according to alternative definitions

The evolution of inter-country and international inequality between 1950 and 2000 is shown in Figure 1, which shows that world income inequality, as measured by the Gini index, has been a story of increasing inter-country inequality and declining international inequality. (The Gini index is probably the most widely used measure of inequality. In theory it varies between 0 – perfect equality – and 1 – perfect inequality. Practically, it ranges from .20 to .25 in most egalitarian countries like the Nordic countries, and .6 or slightly more for the most inegalitarian countries in the world (for instance, Brazil or South Africa). Other measures are used below because of their decomposability property. But the evolution of the

various definitions of world inequality is the same whatever the inequality measure being used.) A 20-year plateau was reached for inter-country inequality starting in the early 1960s, but the unequalizing trend resumed with the crisis of the world economy in the early 1980s. Due to differential demographic growth, no plateau is observed in the decline of the international inequality, but it can be seen that, since the 1987 or so, this decline is essentially fuelled by the fast growth of the two giant developing countries, namely, China and – to a lesser extent – India.

Figure 1

Inter-country and international distribution of income, 1950–2000. *Source*: Milanovic (2005).

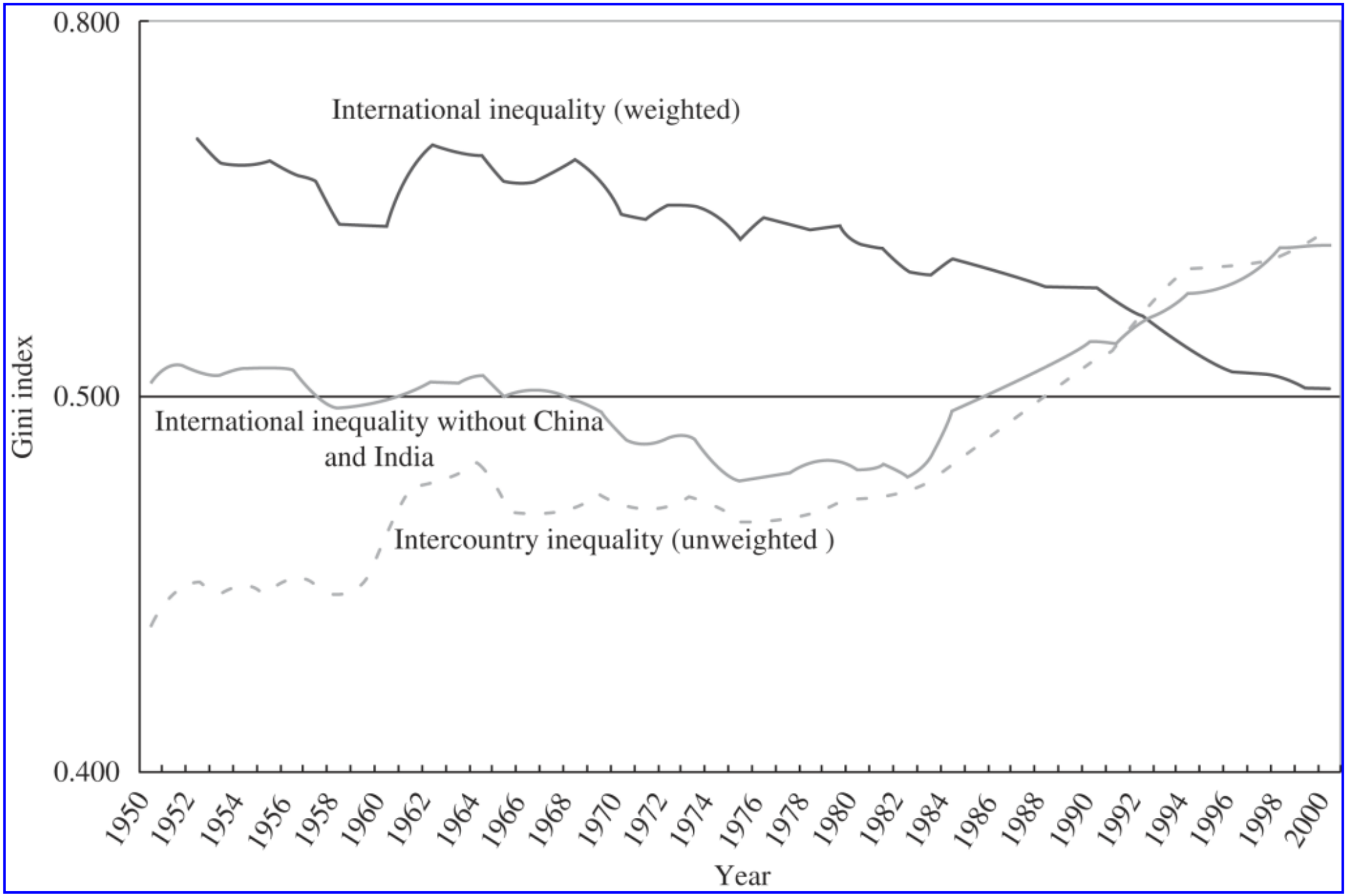

As China and India catch up to the world average, their equalizing effect on the international distribution of income will diminish. If they continue to develop at similar rates than in the past two decades, the effect of their growth will soon be unequalizing (Sala-i-Martin, 2002c). Both inter-country and international inequality will then most likely increase unless countries at the bottom of the two distributions – sub-Saharan African economies in particular – begin to experience healthy growth. This suggests that, in the future, whether the world income distribution is equalizing or unequalizing will increasingly be a function of economic growth in Africa (and some other low-income countries), especially if population growth rates in Africa remain above world average.

To estimate *global* inequality, some knowledge of the distribution of inequality within each country is necessary. Producing estimates for long periods requires strong assumptions. Bourguignon and Morrisson (2002) measure global inequality over 1820–1992 using historical estimates of countries' GDP per capita and rough figures on the distribution of income by deciles within countries to estimate the global distribution of income. They use income distribution information for a limited number of countries and assume that geographically and culturally similar countries or groups of countries have the same distributions. They then reconstitute the global distribution by assuming that individuals in each decile of a country's distribution have the same income. With this method they estimate that the Gini coefficient for the global distribution increased from approximately .5 in 1820 to .66 in 1992, making today's world more probably more unequal than any single country. The mean logarithmic deviation shows the same dramatic evolution of global inequality in Figure 2. But, because, this measure is decomposable into *within* and *between-country* inequality, it permits a better understanding of the forces behind that evolution. In particular, it turns out that *international* inequality was negligible at the turn of the 19th century (accounting for roughly 12 per cent of global inequality) but increased very rapidly until the Second World War. It then stabilized and started to decline after 1980. (The difference with Figure 1 where international inequality declines more or less continuously after 1960 is due to the definition of countries – 33 'country groups' in Bourguignon and Morrisson as against 120 countries in Milanovic – and to the fact that Bourguignon and Morrisson considered discrete years rather than the whole annual series – for instance, 1960 is a 'low' point in the series shown in Figure 1.) Within country inequality, however, reached its peak around 1910 and declined dramatically (mainly due to equalizing forces in the now developed countries) between the two world wars, and started creeping back up only after the 1970s. The combined effect of these changes is an increase in the share of international inequality from roughly ten per cent in 1820 to more than 60 per cent by 1992.

Figure 2

Decomposition of global inequality into within-country and between-country inequality over a long period, 1820–1992. *Sources*: Bourguignon and Morrisson (2002); World Bank (2005).

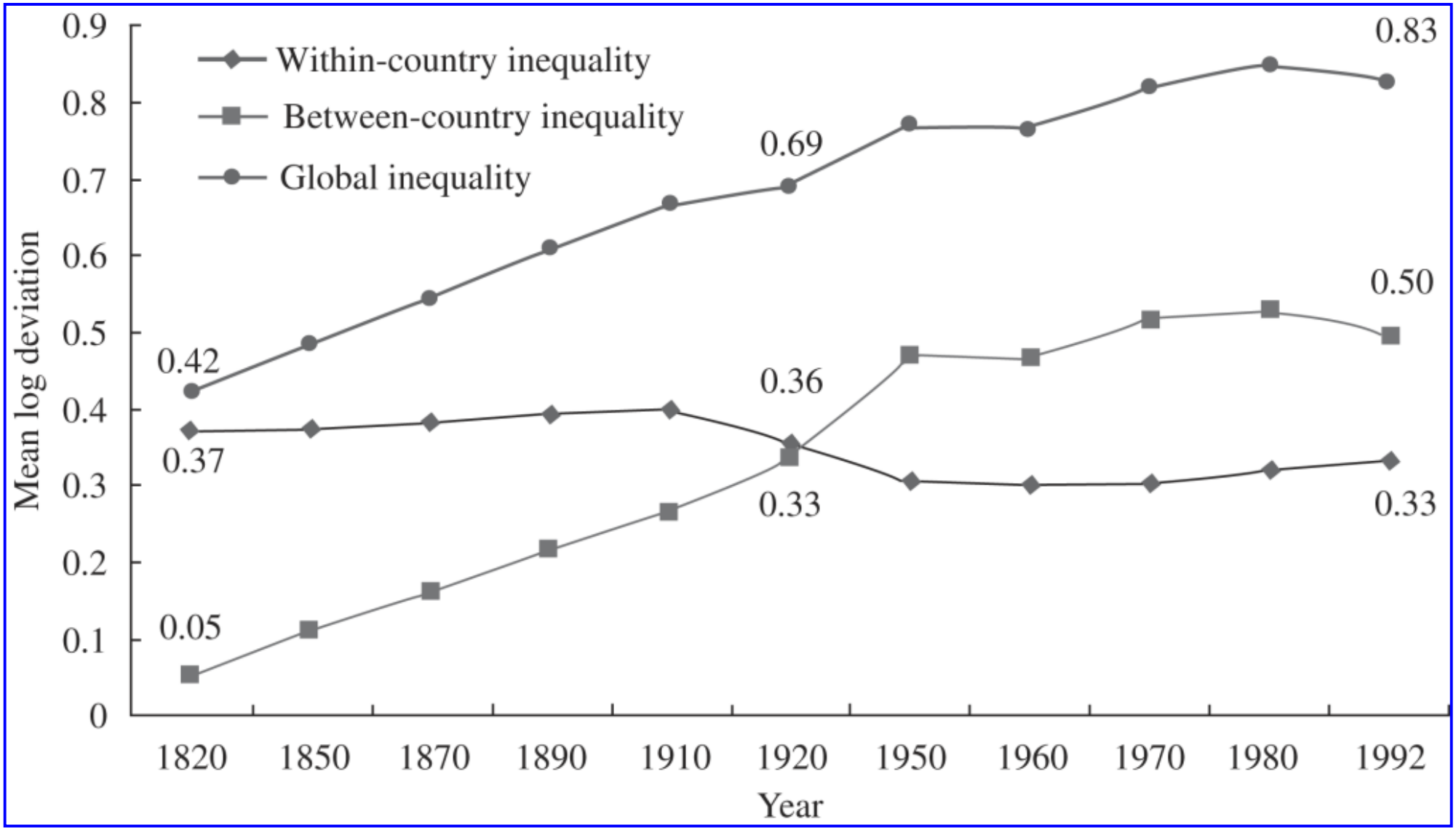

For the recent past, trends in global inequality can be estimated using information from household surveys – which, for developing countries, have been available at regular time intervals only since the 1980s. Inequality within a country cannot increase drastically and indefinitely over the long term. Therefore, even though strong assumptions have to be made, Figure 2, in which the evolution of the international inequality dominates, probably approximates rather well actual long-term trends. The same cannot be said for shorter periods. At the same time, estimates of global inequality become more uncertain over short time intervals because of problems of measurement and comparability between surveys for different countries or at different points of time.

What happened to global inequality in the recent past has been the subject of fierce debate in the context of globalization. In terms of method, Sala-i-Martin (2002a; 2002b) uses an approach that is similar to Bourguignon and Morrisson (2002), combining GDP per capita figures and a constant rough continuous approximation of the distribution of income. Milanovic (2005) uses another method, estimating the parameters of country distributions on the basis of some predetermined functional form and grouped data from all available (comparable) household surveys over the 1980s and 1990s (which cover more than 90 per cent of the world population and world GNI). The sample of countries also differs non-marginally between the two studies.

In a three year comparison, Milanovic (2005) finds that global inequality increased slightly between 1988 and 1993 and then declined between 1993 and 1998. Sala-i-Martin (2002b;

2002c) and others (see, for example, Bhalla, 2002; Firebaugh and Goesling, 2004) argue that it has declined. By and large, however, global inequality did not change much in 1988–2000. In all existing studies, variations of inequality do not exceed a few percentage points, whatever the inequality measure being used. This is in strong contrast with what was observed historically until the Second World War.

The latest and probably the most comprehensive estimate in terms of income distribution data being used is the World Bank (2005) – see Figure 3. It confirms the evolution found by Milanovic. In agreement with Figure 1, it also indicates that the share of global inequality that can be attributed to inequality between countries or *international* inequality declined steadily from 77 per cent around 1988 to 72 per cent around 1993 and to 67 per cent by around 2000. As global inequality stayed roughly the same during this time period, within-group inequality increased at a somewhat steady pace. These results are consistent with international inequality decreasing due to fast income growth in China and India – and with the evidence that inequality in China and in many other countries, including OECD countries, has been increasing over this period (see Ravallion and Chen, 2007, for China; Atkinson and Brandolini, 2004; Cornia, 2004; Katz and Autor, 1999; Gottschalk and Smeeding, 1997). Interestingly enough, this evolution is the opposite of what was observed earlier in history. Does this evolution bring some support to the view derived from standard theoretical models of trade that globalization should tend to substitute inequality across countries by inequality within countries (see Bourguignon and Guesnerie, 1999)?

Figure 3

Decomposition of global inequality into within-country inequality and between-country inequality using household survey data, 1980–2002. *Source*: World Bank (2005).

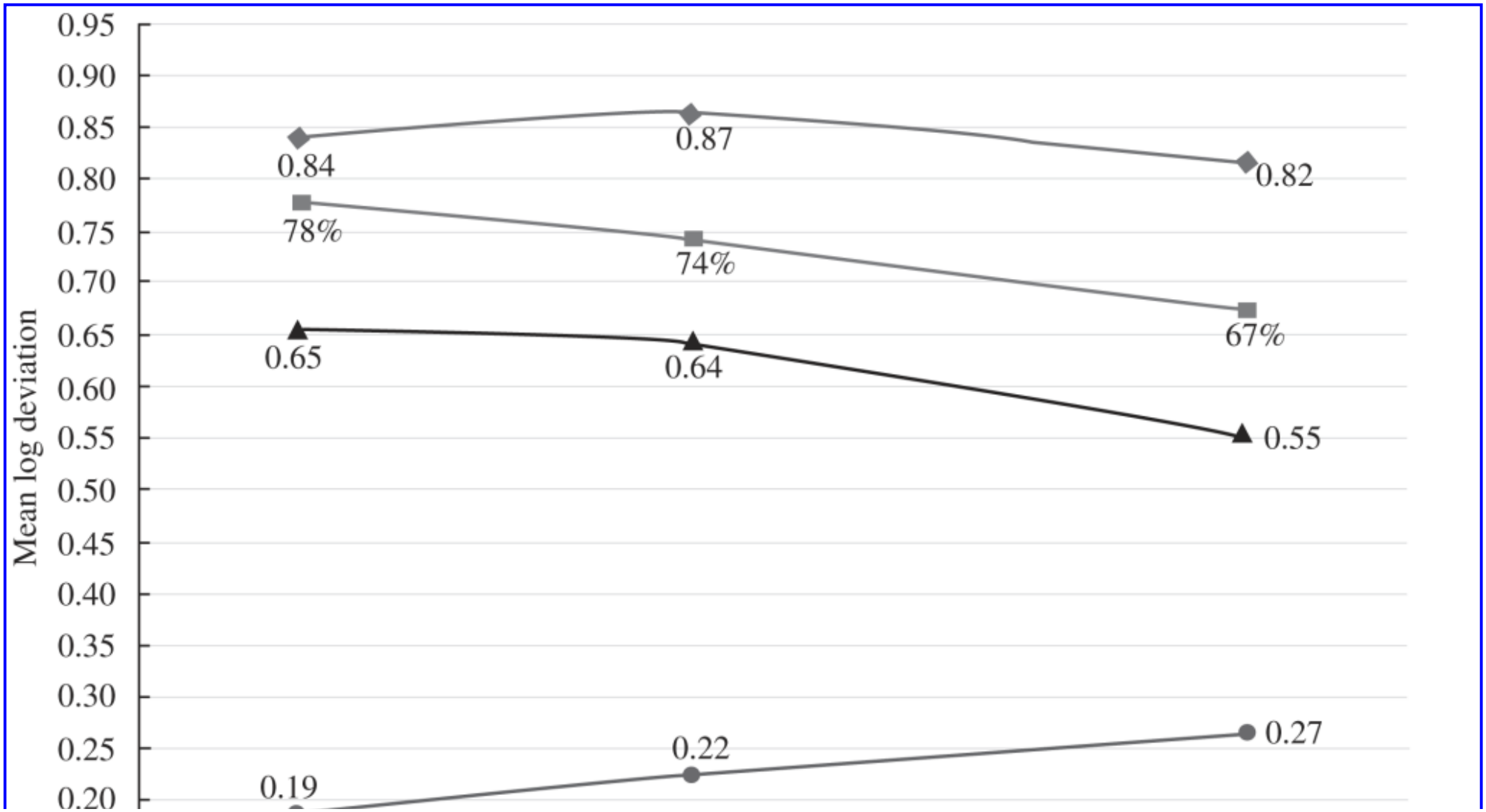

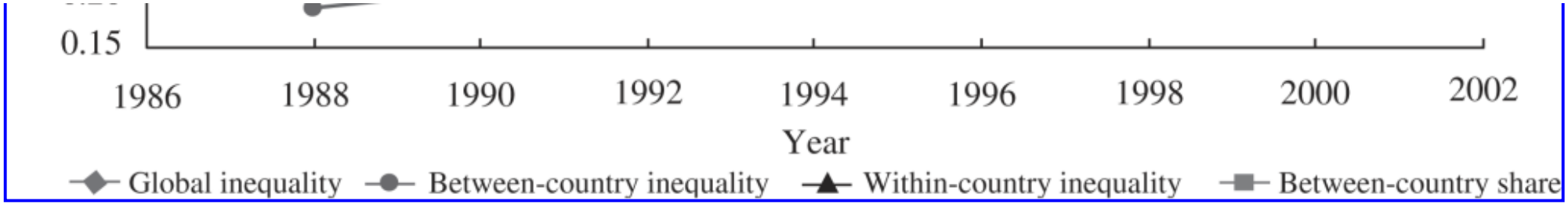

What should we conclude from this review of evidence on the evolution of the world inequality of income? There is no doubt that the world has been increasingly unequal from the beginning of the industrial revolution until the end of the Second World War. But has it become more or less unequal afterwards? Should we rely on the *inter-country* definition and conclude for more inequality, or should we rely on the *international* – and *global* – definitions and conclude for less inequality? Should we give the same weight to China and the Liechtenstein, or should we have the world distribution of income depend on the relative performances of a few giant countries? It is argued in what follows that there is no right or wrong answer to these questions. Alternative definitions of world inequality correspond to different perspectives about the same evidence. The issue is to know whether this evidence is sufficient to form a final judgement or whether more is needed. The next section suggests that the common approach behind the figures reviewed so far may be misleading. Considering the income dynamics of countries, and country citizens, rather than viewing the 'anonymous' distribution at two points of time is more informative and somehow allows this apparent contradiction to be resolved.

Mobility on the world income scale

The approach of inequality in the preceding section focuses only on final outcomes and disregards initial starting positions. A better approach would be to track mobility. It departs from the conventional 'anonymous' view behind inequality measurement and can explain divergent opinions on changes in inequality between nations. If mobility itself forms part of the welfare criterion behind distributional judgements, then one is led to a conclusion about the change in the world distribution of income that is more nuanced and is consistent with both increasing and decreasing inequality. The main point is simply that the evolution of the distribution of income in the world since the 1980s has not been Pareto improving. Some countries among the poorest and their inhabitants have lost income whereas the majority of world citizens have gained. Putting more emphasis on the former would then lead one to conclude that world welfare has fallen, that is, inequality has increased, whereas giving more weight to the latter leads to the opposite conclusion.

A simple way of tracking changes in the international distribution of income is to create 'mobility matrices' of world citizens moving over time from one income range to another (Bourguignon, Levin and Rosenblatt, 2004; Milanovic, 2005). Such a mobility matrix is shown in Table 1. In the calculations behind this table, within-country inequality is assumed away and all citizens within a country are assumed to receive the same income. But there would be little difference if within country inequality were taken into account. In all cases, the world inhabitants that occupied the bottom range of the world income scale (less than 710 US dollars annually after correction for purchasing power parity (PPP), approximately the limit of the first quintile of the international distribution) in 1980 lived mostly in China and a few sub-Saharan African countries. In 2002, however, the Chinese had moved to an upper-income range whereas some sub-Saharan Africans who were initially in the second and third income range had fallen back to the first range. In other words, some poor people with income initially above 710 dollars in 1980 had fallen below that threshold by 2002.

Mobility matrix in absolute country per capita annual income levels (US dollars), 1980–2002

	Income in 2002				
Income in 1980	< 710	711–1,100	1,101–2,890	2,890–10,000	10,001>
< 710	1.28%	1.64%	0.00%	97.08%	0.00%
711–1,100	8.23%	3.89%	87.88%	0.00%	0.00%
1101–2,890	8.09%	0.56%	59.08%	32.28%	0.00%
2,890–10,000	0.00%	0.00%	0.98%	90.84%	8.17%
10,001>	0.00%	0.00%	0.00%	3.99%	96.01%

Source: Bourguignon, Levin and Rosenblatt (2004).

Even though only eight per cent of each of the second and third income ranges fell into the bottom range over these two decades, this evolution clearly shows that no Pareto improvement has taken place in the world income distribution between 1980 and 2002. When considered in an anonymous way, as in the previous section, it may be the case that the

average income of people in the poorest deciles of the international distribution of income has increased and may have even become closer to the mean world income. Yet this hides the fact that the composition of these deciles was very much modified. Chinese went out and were replaced by people from poor countries, initially richer than China, whose income fell after 1980.

Overall, whether the world distribution of income is judged as improving or worsening depends essentially on whether one considers that the increase in income of the Chinese and other poor people who climbed the income scale between 1980 and 2002 over- or under-compensates for the drop in the income of those people whose income fell. Looking at *international* inequality with population weights is equivalent to taking the former view, whereas focusing on *inter-country* inequality leads to the second conclusion. If the initial income position matters in assessing the social welfare of a population observed at a given point of time, then the social cost of falling incomes is not necessarily compensated for by the social gain of increasing incomes even if these changes take place towards and from the same income range. Deciding whether such a change increases or decreases social welfare requires some value judgements. The difference between those who feel that the world distribution of income has become worse and those who feel the opposite may simply be due to such differences in their value judgements.

Non-income dimensions of world inequalities

Most of the existing empirical evidence on inequality between nations concerns income, but recent studies have examined inequality in other dimensions of well-being, mainly health and education (see for example, Araujo, Ferreira and Schady, 2004; Deaton, 2004; Goesling and Firebaugh, 2004; Sala-i-Martin, 2002c; Schady, 2005). These studies indicate convergence in health and education indicators but divergence (or at least lack of convergence) in income. International inequalities in educational attainment and child mortality have been steadily declining, though improvements in life expectancy at birth have been set back since the early 1990s due to the devastating effects of HIV/AIDS and the difficult circumstances facing the former USSR and other transitions economies. Unlike global inequalities in income, global inequalities in educational attainments are attributable mostly to inequalities within countries.

What explains that there is convergence for health and education indicators and divergence for incomes? Deaton (2004, p. 109) points out that, while gains in income were undoubtedly important for improving nutrition and for funding better water and sanitation schemes, some countries made progress in reducing child mortality even in the absence of economic growth. These improvements came from the globalization of knowledge, facilitated by local political, economic and educational conditions. A possible explanation for the disconnect between the convergence in education and the divergence in incomes is that education is not the only determinant of income and that the rise in per-worker schooling explains only a small part of the growth in output per worker.

Finally, it is worth insisting that there are major inequalities in voice and power between nations in participating to international decisions. These are discussed in detail in World Bank (2005). As Deaton (2004) puts it, poor countries lack the financial and human resources that would allow them to be equal participants in the international bodies in which decisions are taken that affect them and, beyond that, in setting the rules under which the international system operates.

Equity at the global level means that people should face the same opportunities for living the life they want regardless of where they are born. Income inequality among nations is only a sign that we are far from such a goal. Convergence in some non-income dimensions is an encouraging sign. Global action is possible in a number of areas to promote world equity, from improvements in international law and human rights, to promoting fairness in global markets, allowing free trade and free migration of labour, to more aid to the poorest, to a more equitable management of the environment and the global commons.

See Also

- inequality (global)
- inequality (international evidence)

Bibliography

Araujo, C., Ferreira, F. and Schady, N. 2004. *Is the World Becoming More Unequal? Changes in the World Distribution of Schooling*. Washington, DC: World Bank.

Atkinson, A.B. and Brandolini, A. 2004. Global world inequality: absolute, relative or intermediate? Paper presented at the 28th General Conference of the International Association for Research on Income and Wealth. Cork, Ireland, 22 August.

Bhalla, S. 2002. *Imagine there is No Country: Poverty, Inequality and Growth in the Era of Globalization*. Washington, DC: Institute of international Economics.

Bourguignon, F. and Guesnerie, R. 1999. L'économie mondialisée: inégalités entre nations hier, inégalité au sein des nations demain? Paper presented at an interdisciplinary seminar, Ecole des Haute Etudes en Sciences Sociales, Paris, 10 February.

Bourguignon, F., Levin, V. and Rosenblatt, D. 2004. Declining economic inequality and economic divergence: reviewing the evidence through different lenses. *Économie Internationale* 100, 13–25.

Bourguignon, F. and Morrison, C. 2002. Inequality among world citizens: 1820–1992. *American Economic Review* 92, 727–44.

Cornia, G.A. 2004. Inequality, growth and poverty: an overview of changes over the last two decades. In *Inequality, Growth and Poverty in an Era of Liberalization and Globalization*, ed. G.A. Cornia. Oxford: Oxford University Press.

Deaton, A. 2004. Health in an age of globalization. *Brookings Trade Forum 2004*. Washington, DC: Brookings Institution.

Firebaugh, G. and Goesling, B. 2004. Accounting for the recent decline in global income inequality. *American Journal of Sociology* 110, 283–312.

Gottschalk, P. and Smeeding, T. 1997. Cross-national comparisons of earnings and income inequality. *Journal of Economic Literature* 35, 633–87.

Goesling, B. and Firebaugh, G. 2004. The trend in international health inequality. *Population and Development Review* 30, 131–46.

Katz, L.F. and Autor, D.H. 1999. Changes in the wage structure and earnings inequality. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter, R. Layard and D. Card. Amsterdam: North-Holland.

Kenny, C. 2005. Why are we worried about income? Nearly everything that matters is converging. *World Development* 33, 1–19.

Milanovic, B. 2005. *Worlds Apart: Measuring International and Global Inequality*. Princeton, NJ: Princeton University Press.

Ravallion, M. 2004. Competing concepts of inequality in the globalization debate. *Brookings Trade Forum 2004*. Washington, DC.: Brookings Institution.

Ravallion, M. and Chen, S. 2007. China's (uneven) progress against poverty. *Journal of Development Economics* 82, 1–42.

Sala-i-Martin, X. 2002a. The disturbing 'rise' of world income inequality. Working Paper No. 8904. Cambridge, MA: NBER.

Sala-i-Martin, X. 2002b. The world distribution of income. Working Paper No. 8905. Cambridge, MA: NBER.

Sala-i-Martin, X. 2002c. Unhealthy people are poor people … and vice versa. Keynote address at the European Conference on Health Economics of the International Health Economics Organization, Paris, 7 July. Online. Available at http://www.columbia.edu/~xs23/papers/parisconference.pdf, accessed 24 June 2007.

Schady, N. 2005. *Changes in the Global Distribution of Life Expectancy and Education*. Washington, DC: World Bank.

World Bank. 2005. *World Development Report 2006. Equity and Development*. New York: Oxford University Press for the World Bank.

How to cite this article

Bourguignon, François. "inequality between nations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 01 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000081> doi:10.1057/9780230226203.0786

The New Palgrave Dictionary of Economics Online

infant-industry protection

Douglas A. Irwin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The infant industry argument for protection holds that new industries in developing countries should be promoted through trade or industrial policy measures to allow them to mature and compete successfully against established foreign rivals. This article assesses the theory and evidence behind this qualification to the case for free trade.

Keywords

Brazil; capital market failure; cost–benefit analysis; counterfactual simulation; Denmark; dynamic learning effects; growth and learning-by-doing; Hamilton, A.; imperfect information; infant industry protection; informational barriers to entry; learning-by-doing; learning-by-doing spillovers; List, F.; Mill, J. S.; research and development; Smith, A.; trade policy, political economy of; transfer of technology; underproduction

Article

The infant industry argument for trade protection is one of the oldest and most widely debated qualifications to the case for free trade. The argument holds that certain new industries should be protected from foreign competition in the expectation that they will eventually mature and successfully compete against more experienced foreign rivals. The case for infant industry protection involves temporary and selective, not permanent and across-the-board, government assistance and is often discussed in the context of trade policies that might promote economic development.

The idea of infant industries can be traced back as far as the 17th century (Irwin, 1996). In Book IV of the *Wealth of Nations* (1776), Adam Smith was sceptical that trade restrictions would create new wealth, arguing that they would just divert scarce resources into less productive endeavours. Other writers, such as Alexander Hamilton (1791) and Friedrich List (1841), believed that policies to promote manufacturing industries would be beneficial in encouraging economic diversification and growth in

developing countries. The classical economist John Stuart Mill (1848, p. 922) lent his authority to the case by endorsing it in this way:

> The only case in which, on mere principles of political economy, protecting duties can be defensible, is when they are imposed temporarily (especially in a young and rising nation) in hopes of naturalizing a foreign industry, in itself perfectly suitable to the circumstances of the country. The superiority of one country over another in a branch of production, often arises only from having begun it sooner. There may be no inherent advantage on one part, or disadvantage on the other, but only a present superiority of acquired skill and experience. A country which has this skill and experience yet to acquire, may in other respects be better adapted to the production than those which were earlier in the field . . . . But it cannot be expected that individuals should, at their own risk, or rather to their certain loss, introduce a new manufacture, and bear the burthen of carrying it on until the producers have been educated up to the level of those with whom the processes are traditional. A protecting duty, continued for a reasonable time, will [*changed to 'might' in later editions*] sometimes be the least inconvenient mode in which the nation can tax itself for the support of such an experiment. But ['*it is essential that' added in later editions*] the protection should be confined to cases in which there is good ground of assurance that the industry which it fosters will after a time be able to dispense with it; nor should the domestic producers ever be allowed to expect that it will be continued to them beyond the time necessary for a fair trial of what they are capable of accomplishing.

In the 19th century, the debate over infant industry protection centred on whether such protection would (*a*) create new wealth and capital, or merely divert it from other more profitable activities, (*b*) stimulate domestic producers to acquire new technology and skills, or just stifle the incentive for such efforts, and (*c*) generate long-term net benefits, or simply foster costly industries that would require ongoing government support. Unfortunately, economic analysis proved to be of little assistance in evaluating these claims, as one could envision the successful maturation of an infant industry but also see the possibility of protection breeding inefficiencies; a priori, neither outcome could be dismissed.

In the modern literature, the infant industry argument hinges on dynamic learning effects, which allow an industry that is not currently competitive to become so after a temporary period of protection. As such, the conditions for infant industry protection include the following: (*a*) irreversible technological external economies that cannot be captured by the protected industry, (*b*) a limited period of protection, and (*c*) sufficient long-run economic benefits (lower production costs that generate producer surplus) that will more than compensate for the costs associated with protection, with a rate of return at least equal to that on other investments (Kemp, 1960).

If condition (*a*) is not fulfilled, the private market should deliver an efficient outcome unless there is some other market imperfection (poorly functioning capital markets, imperfect information) so that risks to starting the industry are overestimated by private agents and there is underproduction from the social point of view. Condition (*b*) states that protection must be time-limited, and not persist indefinitely. Condition (*c*) requires an intertemporal cost–benefit analysis, wherein the initial costs of protecting the industry will be more than offset by long-run benefits.

The modern literature on infant industries also focuses on identifying the specific market failure or

distortion that makes government intervention necessary as well as the ranking of alternative policy instruments in terms of their ability to correct the market failure or distortion. If an industry is characterized by learning-by-doing spillovers, wherein production costs for any firm fall as a result of production experience by anyone in the industry (that is, the learning benefits are external to the firm), then this dynamic economy of scale may lead to a divergence between private and social costs of production. The knowledge generated by research and development expenditures can also create external benefits which could lead to underinvestment by the private market. And if there are capital market failures, such that firms cannot acquire credit (Flam and Staiger, 1991; Bond, 1993), or informational barriers to entry (Grossman and Horn, 1988), there may be a case for selective interventions.

Once the specific obstacle facing the infant industry is identified (that is, related to production experience, technology transfer, or imperfect capital markets), then the policy recommendation can be specifically targeted to address the problem. In general, trade protection will not be the first-best policy intervention to correct the distortion that hinders the development of an infant industry. Baldwin's (1969) classic critique of the infant industry argument stresses that import protection fails to provide the right incentive for an infant firm to make additional investments in acquiring technological knowledge and does not necessarily solve a firm's appropriability problem of securing the benefits of investments in knowledge or production experience. But, by reducing foreign competition and raising the domestic price, it does make the status quo more profitable.

If the economic conditions giving rise to infant industries are difficult to assess, the implementation of a welfare-improving policy also poses difficulties for governments. The government must differentiate among various industries (ignoring the lobbying of firms for government assistance), pick those to support with preferential policies, select the proper policies to ensure that firms have the incentive to respond the right way, and be able to resist pressure from firms to maintain protection indefinitely (Tornell, 1991).

Despite many theoretical articles on infant industry protection, Krueger and Tuncer (1982, p. 1142) wrote that ‘there has been virtually no systematic examination of the empirical relevance of the infant industry argument’ through *ex post* evaluations or other studies. They found little correlation between various measures of the effective rate of protection and industry productivity growth in Turkey, but did not perform counterfactual simulation or cost–benefit analysis (see Harrison, 1994).

To date, there are still relatively few evaluations of infant industry policies. Luzio and Greenstein (1995) studied performance of the Brazilian microcomputer industry under protection that started in the early 1990s. They found that the rates of technological advance in Brazil were rapid but lower than that of potential international competition. As a result, the technical frontier in Brazil lagged that best performance practices in international markets by three to five years, and forgone consumer surplus due to protection approached 20 per cent of domestic expenditure on microcomputers.

Hansen, Jensen and Madsen (2003) examined the welfare effects of subsidies in Denmark for the production of electricity from wind power. They found strong learning-by-doing productivity growth in the Danish windmill industry, and the industry achieved a dominant position in the world market. By their calculation, these subsidies passed a cost–benefit test: the costs consist of the efficiency loss from diverting electricity production from using fossil fuels to utilizing wind power, but the benefits include reductions in the environmental damage of using fossil fuels and the emergence of a new export sector.

They concluded that the subsidies pass a cost–benefit test because the value of the windmill firms at the stock exchange far exceeds the accumulated distorted losses in electricity production.

Economists have also looked at historical cases of infant industry protection, as it is commonly contended that high-income countries such as the United States, Japan and Germany rose to industrial prominence by protecting infant industries in the late 19th and early 20th centuries (Chang, 2003). Two studies examined the US iron and steel industry in the late 19th century. Head (1994) sought to account for the individual roles of learning-by-doing, changing resource endowments, and tariff protection in the emergence of the steel rail industry. In a counterfactual simulation of what would have happened under free trade, he concluded that learning effects were very strong and that, even though the steel rail tariff hurt rail users in both short and long runs, the tariff's overall effect on welfare was positive but fairly small. Irwin (1998) studied the US tinplate industry which, after earlier failures, flourished after receiving tariff protection in 1890. His counterfactual simulation indicated that, without the additional duties, domestic tinplate production would have arisen about a decade later as US iron and steel input prices converged to those in Britain. Although the tariff accelerated the industry's development, welfare calculations suggest that protection did not pass a cost–benefit test.

In general, however, economists have been sceptical about the relevance of the infant industry argument for current developing countries, and for the ability of governments to implement the policy wisely. For example, reviewing the empirical literature on manufacturing establishments in developing countries, Tybout (2000) found that unexploited economies of scale in developing countries are insignificant and that protection tends to reduces average efficiency levels by allowing lower-productivity, higher-cost firms to survive in the market. He concluded that ‘although the econometric evidence on technology diffusion in [developing countries] is limited, it does suggest that protecting “learning” industries is unlikely to foster productivity growth’ (2000, p. 39).

See Also

- growth and international trade
- international trade theory
- strategic trade policy
- trade policy, political economy of
- trade, technology diffusion and growth

Bibliography

Baldwin, R. 1969. The case against infant-industry tariff protection. *Journal of Political Economy* 77, 295–305.

Bell, M., Ross-Larson, B. and Westphal, L. 1984. Assessing the performance of infant industries. *Journal of Development Economics* 16, 101–28.

Bond, E. 1993. Capital market imperfections and the infant industry argument for protection. In *Theory,*

Policy and Dynamics in International Trade: Essays in Honor of Ronald W. Jones, ed. W. Ethier, E. Helpman and J. Neary. New York: Cambridge University Press.

Chang, H.-J. 2003. Kicking away the ladder: infant industry promotion in historical perspective. *Oxford Development Studies* 31, 21–32.

Flam, H. and Staiger, R. 1991. Adverse selection in credit markets and infant industry protection. *International Trade and Trade Policy*, ed. E. Helpman and A. Razin. Cambridge, MA: MIT Press.

Grossman, G. and Horn, H. 1988. Infant-industry protection reconsidered: the case of informational barriers to entry. *Quarterly Journal of Economics* 103, 767–88.

Hamilton, A. 1791. Report on manufactures. In *The Papers of Alexander Hamilton*, ed. H. Syrett. New York: Columbia University Press, 1961.

Hansen, J., Jensen, D. and Madsen, E. 2003. The establishment of the Danish windmill industry – was it worthwhile? *Review of World Economics* 139, 324–47.

Harrison, A. 1994. An empirical test of the infant industry argument: comment. *American Economic Review* 84, 1090–5.

Head, K. 1994. Infant industry protection in the steel rail industry. *Journal of International Economics* 37, 141–65.

Irwin, D. 1996. *Against the Tide: An Intellectual History of Free Trade*. Princeton: Princeton University Press.

Irwin, D. 1998. Did late-nineteenth-century US tariffs promote infant industries? Evidence from the tinplate industry. *Journal of Economic History* 60, 335–60.

Kemp, M. 1960. The Mill–Bastable infant-industry dogma. *Journal of Political Economy* 68, 65–7.

Krueger, A. and Tuncer, B. 1982. An empirical test of the infant industry argument. *American Economic Review* 72, 1142–52.

List, F. 1841. *National System of Political Economy*. Philadelphia: J. Lippincott.

Luzio, E. and Greenstein, S. 1995. Measuring the performance of a protected infant industry: the case of Brazilian microcomputers. *Review of Economics and Statistics* 77, 622–33.

Mill, J. S. 1848. *Principles of Political Economy*. London: Longmans, 1909.

Smith, A. 1776. *The Wealth of Nations*. Oxford: Clarendon Press, 1976.

Tornell, A. 1991. Time inconsistency of protectionist programs. *Quarterly Journal of Economics* 106, 963–74.

Tybout, J. 2000. Manufacturing firms in developing countries: how well do they do, and why? *Journal of Economic Literature* 38, 11–44.

How to cite this article

Irwin, Douglas A. "infant-industry protection." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000281> doi:10.1057/9780230226203.0790

The New Palgrave Dictionary of Economics Online

inflation dynamics

Timothy Cogley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There have been a number of changes in monetary policy rules in the United States and UK since the early 1960s. The Lucas critique says that this should induce changes in the equilibrium law of motion. This article summarizes reduced-form evidence on the evolving law of motion for inflation in the USA and the UK. Since the 1970s, inflation has become lower on average, less volatile and less persistent. There is also less uncertainty about the central bank's long-run target for inflation.

Keywords

Bank of England; Bretton Woods system; central-bank independence; commitment; European Monetary System (EMU); Federal Reserve System; fixed exchange rates; Greenspan, A.; inflation; inflation dynamics; inflation gap; inflation targeting; inflation volatility; Lucas critique; monetary policy rules; New Keynesian macroeconomics; output gap; Phillips curve; prediction-error variance; sticky price models; supply shocks; Taylor rule; Thatcher, M.; trend inflation; vector autoregressions; Volcker, P.

Article

As with other macroeconomic data series, economists have long been interested in the dynamics of inflation. This series is of particular interest because of its relationship to alternative theories of aggregate fluctuations and associated effects of alternative monetary policies. Indeed, for understanding the behaviour of inflation, it is important to take into account the alternative monetary policies that were in force during the period being studied. Shifts in monetary policy rules alter the fundamentals that drive inflation and therefore also alter its dynamic properties. Accordingly, one must distinguish inflation variation arising within a stable monetary regime from movements that follow from shifts in policy rules.
For the United States, Taylor's (1993) rule is often used to describe Federal Reserve behaviour. Although his rule was originally intended as a normative proposal, Taylor and others soon discovered that it closely approximated Federal Reserve behaviour during the Volcker–Greenspan era (1979–2006). Shortly thereafter, a number of economists began to explore whether the Taylor rule also described Fed behaviour prior to that (for example, Clarida, Gali and Gertler, 2000; Lubik and Schorfheide, 2004). They found that it did, although with different coefficients. Prior to Volcker's term as chairman, the Fed reacted more strongly to fluctuations in output and was less sensitive to movements in inflation. In fact, although the Fed increased the nominal funds rate when inflation rose, it increased the funds rate by less than one for one, so that the real funds rate actually declined, thus amplifying the initial movement in inflation. According to Clarida, Gali and Gertler and Lubik and Schorfheide, this was an important factor behind the high and volatile inflation of the 1970s.
Important changes have also occurred in UK monetary policy. After the Second World War, the Bank of England at first operated under the Bretton Woods system of fixed exchange rates; the breakdown and float in the early 1970s was followed by attempts to re-establish fixed exchange rates in the 1980s, the decision to opt out of EMU after the foreign-exchange crisis of 1992, the adoption of inflation targeting in 1992, and finally central-bank independence from the Treasury in 1997. The last two steps in particular altered the way the Bank of England conducts monetary policy, with the Bank now placing a higher priority on controlling inflation.
The Lucas critique says that a change in a government policy rule should alter the equilibrium law of motion for endogenous variables. In this article I summarize research on changes in the dynamic properties of inflation since the early 1960s. I focus on the USA and the UK because they are the two economies studied most extensively in the literature.

Evolving inflation dynamics in the USA and the UK

I use a vector autoregression (VAR) to summarize the dynamic properties of inflation. For a historical period during which government and private decision rules are unchanged, one can estimate a time-invariant VAR. Here I am concerned about a period with changing monetary policy rules, however, so VAR parameters must be allowed to vary, in accordance with the Lucas critique. Consequently, much of the literature on evolving inflation dynamics studies Bayesian VARs with drifting conditional mean and variance parameters; for example, Benati and Mumtaz (2006), Canova and Gambetti (2006), Cogley and Sargent (2001; 2005), Cogley, Morozov, and Sargent (2005), and Primiceri (2005a).

Cogley and Sargent (2005) estimate VARs of the form

$$y_t = X_t' \theta_t + \varepsilon_t, \tag{1}$$

where y_t is a vector consisting of inflation, a short-term nominal interest rate, and a real activity variable such as unemployment or GDP. The right-hand variables X_t consist of a constant plus lags of y_t, and the conditional mean parameters θ_t evolve as a driftless random walk subject to reflecting barriers. The driftless random walk assumption makes θ_t vary as

$$\theta_t = \theta_{t-1} + v_t, \tag{2}$$

where v_t is $NID(0, Q)$. The reflecting barriers prevent θ_t from wandering into the region of the parameter space where the system has explosive autoregressive roots. This representation puts a unit root in inflation, because the reflecting barriers do not restrict drift in the VAR intercepts, but it prohibits more than one unit root in inflation.

The VAR innovations ε_t are assumed to be conditionally normal with mean zero and drifting variance

$$R_t = B^{-1} H_t B^{-1'}, \tag{3}$$

where H_t is diagonal and B is lower triangular:

$$H_t = \begin{pmatrix} h_{1t} & 0 & 0 \\ 0 & h_{2t} & 0 \\ 0 & 0 & h_{3t} \end{pmatrix}, \tag{4}$$

$$B = \begin{pmatrix} 1 & 0 & 0 \\ \beta_{21} & 1 & 0 \\ \beta_{31} & \beta_{32} & 1 \end{pmatrix}. \tag{5}$$

The diagonal elements of H_t are univariate stochastic volatilities that evolve as driftless, geometric random walks:

$$\ln h_{it} = \ln h_{it-1} + \sigma_i \eta_{it}. \tag{6}$$

The random-walk assumption is designed to fit permanent shifts in innovation variances such as those associated with the 'Great Moderation' in the USA (McConnell and Perez Quiros, 2000). This formulation allows time-varying correlations among the VAR innovations, and it guarantees that R_t is positive definite. Primiceri (2005a) extends the model to allow for drifting covariances as well, $R_t = B_t^{-1} H_t B_t^{-1'}$.
The results reported below illustrate various aspects of the Bayesian posterior distribution for this model. Readers who are interested in the technical details should consult the original sources.

Trend inflation

Figure 1 depicts trend inflation in the USA and the UK. Following Beveridge and Nelson (1981), I define trend inflation in terms of long-horizon forecasts. At date *t*, trend inflation is the level at which inflation is expected to settle after the transient variation dies out,

$$\bar{\pi}_t = \lim_{j \to \infty} E_t \pi_{t+j}. \tag{7}$$

To approximate $\bar{\pi}_t$, write the VAR in companion form as

$$z_t = \mu_t + A_t z_{t-1} + \varepsilon_{zt}, \tag{8}$$

where μ_t contains the VAR intercepts and A_t the autoregressive parameters. Then trend inflation $\bar{\pi}_t$ can be approximated by the long-horizon VAR forecast,

$$\bar{\pi}_t \doteq \mu_t (1 - A_t)^{-1}. \tag{9}$$

The figure portrays the posterior median of $\bar{\pi}_t$ at each date along with the interquartile range. (All the figures are based on the author's calculations.)

Figure 1

Trend inflation in the USA and the UK. *Sources*: Federal Reserve Economic Database (USA), Bank of England (UK).

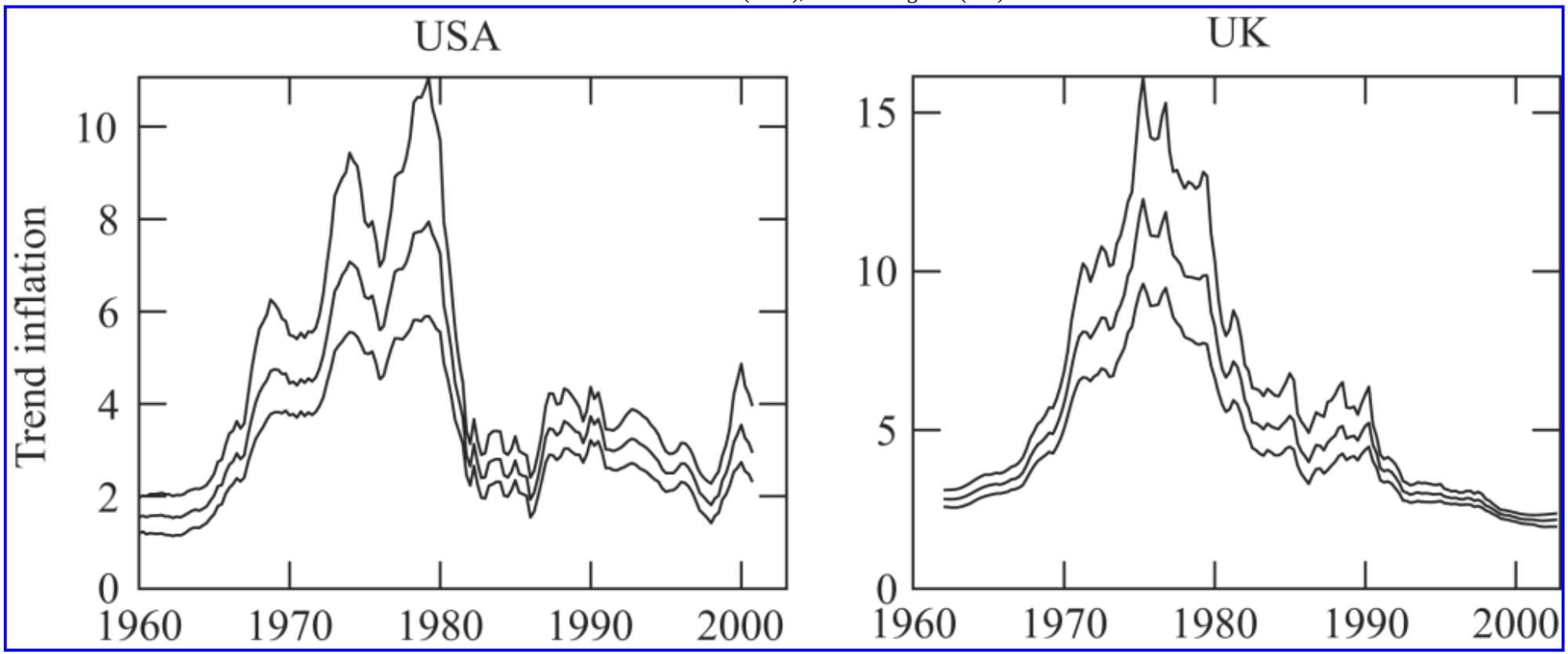

The median estimate of trend inflation was a bit below two per cent in the USA in the early 1960s, and it was just shy of three per cent in the UK. It increased sharply in both countries in the late 1960s and early 1970s and peaked in the mid- to late 1970s. In the UK, a peak of 12 per cent was reached in 1975, and $\bar{\pi}_t$ remained in double digits until 1980. In the USA, trend inflation ratcheted to 4.5 percent in 1970, then to seven per cent in 1974, and finally to almost eight per cent in 1979. In the early 1980s, Paul Volcker's disinflation in the USA and Margaret Thatcher's monetarist experiment in the UK brought $\bar{\pi}_t$ quickly back down to more tolerable levels. In the USA, trend inflation has fluctuated around 2–3.5 per cent since 1985. In the UK, $\bar{\pi}_t$ settled in the neighbourhood of 2.5 per cent when the Bank of England adopted an explicit inflation target in 1992, and then declined gradually to around two per cent after 1997.

Measures of uncertainty about trend inflation also rise and then fall. The inter-quartile range for $\bar{\pi}_t$ is narrow at the beginning of the sample, widens substantially in the middle, and then narrows again at the end. Thus, the 1970s was not only a decade when trend inflation was high, but also a time of substantial uncertainty about its exact value. For example, in the USA, when the median estimate of $\bar{\pi}_t$ peaked at eight per cent, there was a fifty-fifty chance that it was somewhere outside the interval 5.75–12 per cent. Similarly, when the median estimate of UK trend inflation peaked at 12 per cent, there was a fifty-fifty chance that it could exceed 16 per cent or fall short of 12 per cent.

The inter-quartile range quantifies level uncertainty, that is, uncertainty about where $\bar{\pi}_t$ is at a particular date. One can also quantify how rapidly trend inflation drifts. Figure 2 portrays the median absolute deviation for $\Delta\bar{\pi}_t$. (I report this statistic instead of the standard deviation because of outliers.) Not only is there a lot of uncertainty about the level of $\bar{\pi}_t$

in the 1970s, but $\bar{\pi}_t$ was also drifting more rapidly then. The rate of drift fell considerably in both countries in the early 1980s, and in the UK it also declined sharply after 1992. Stock and Watson (2005) were the first to report a result like this in the context of an unobserved-components model of US inflation. The result also holds for our drifting-parameter VARs.

Figure 2

Median absolute deviation for $\Delta\bar{\pi}_t$. *Sources*: Federal Reserve Economic Database (USA), Bank of England (UK).

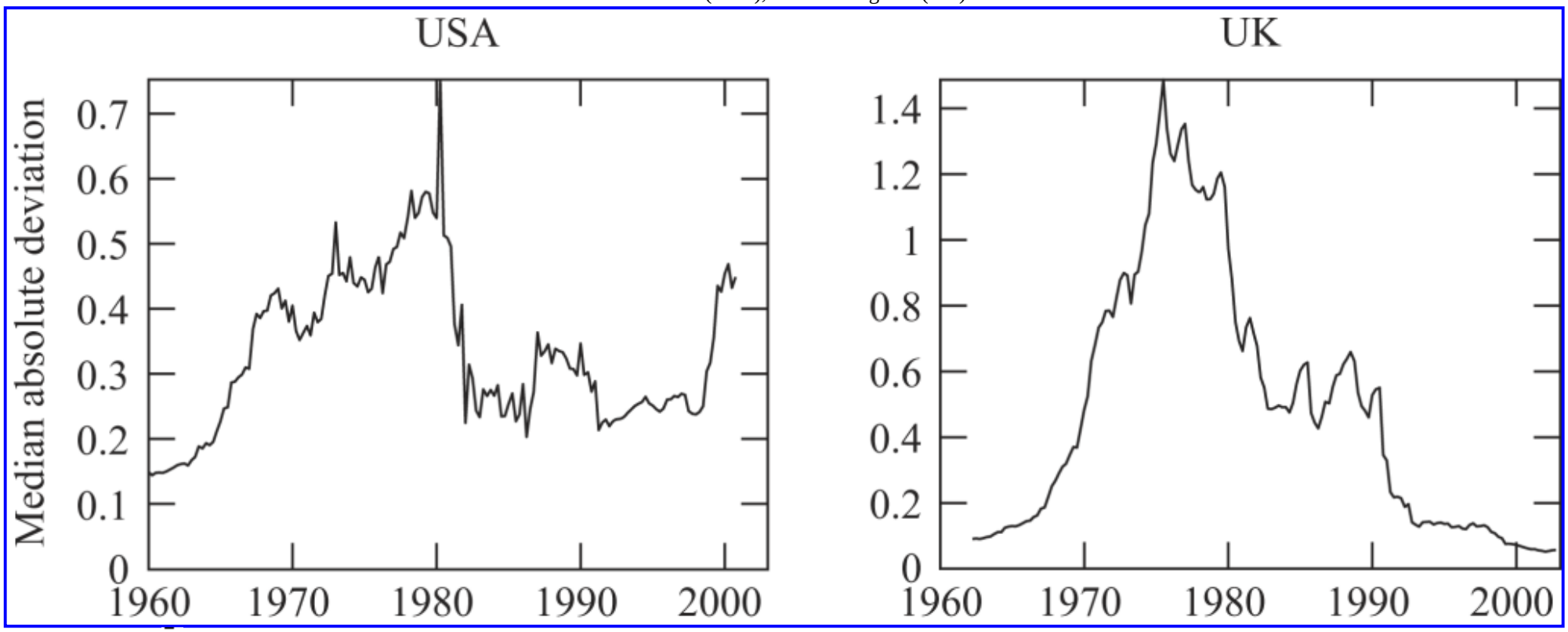

Uncertainty about $\bar{\pi}_t$ presumably reflects uncertainty about the central bank's long-run target for inflation, or doubt about its commitment to the target, or both. For the UK, it is interesting to note how the inter-quartile range narrowed and the rate of drift declined after the adoption of inflation targeting in 1992. In contrast, the inter-quartile range for the USA was about as wide in the 1990s as in the 1980s, and its width was also comparable to that of the UK for the 1980s. Similarly, the rate of drift in US trend inflation was considerably higher at the end of the sample than in the UK. The difference, of course, is that the USA has not adopted a formal inflation target. Taken at face value, these figures illustrate how an explicit inflation target can anchor long-run inflation expectations.

Inflation gap variability

Next I turn to changes in inflation volatility. To a first-order approximation, trend inflation is a random walk, which means that inflation itself has infinite unconditional variance. Here I focus instead on the volatility of de-trended inflation, $\pi_t - \bar{\pi}_t$. A central bank that behaves as if minimizing an undiscounted quadratic loss function will adjust its policy instrument so that inflation eventually converges to its target. Thus, I interpret $\bar{\pi}_t$ as a measure of target inflation and $\pi_t - \bar{\pi}_t$ as the deviation from the target or ‘inflation gap’. At each date in the sample, I approximate its instantaneous standard deviation as

$$\sigma_t(\pi_t - \bar{\pi}_t) = \left[\sum_{j=0}^{\infty} e_\pi' A_t^j R_t A_t^{j\prime} e_\pi\right]^{1/2},$$

(10)

where e'_{π} is a selection vector that picks out π_t from the vector z_t. The top row of Figure 3 portrays the evolution of the posterior median and interquartile range for σ_t.

Figure 3

Standard deviation of inflation gaps and inflation innovations. *Sources*: Federal Reserve Economic Database (USA), Bank of England (UK).

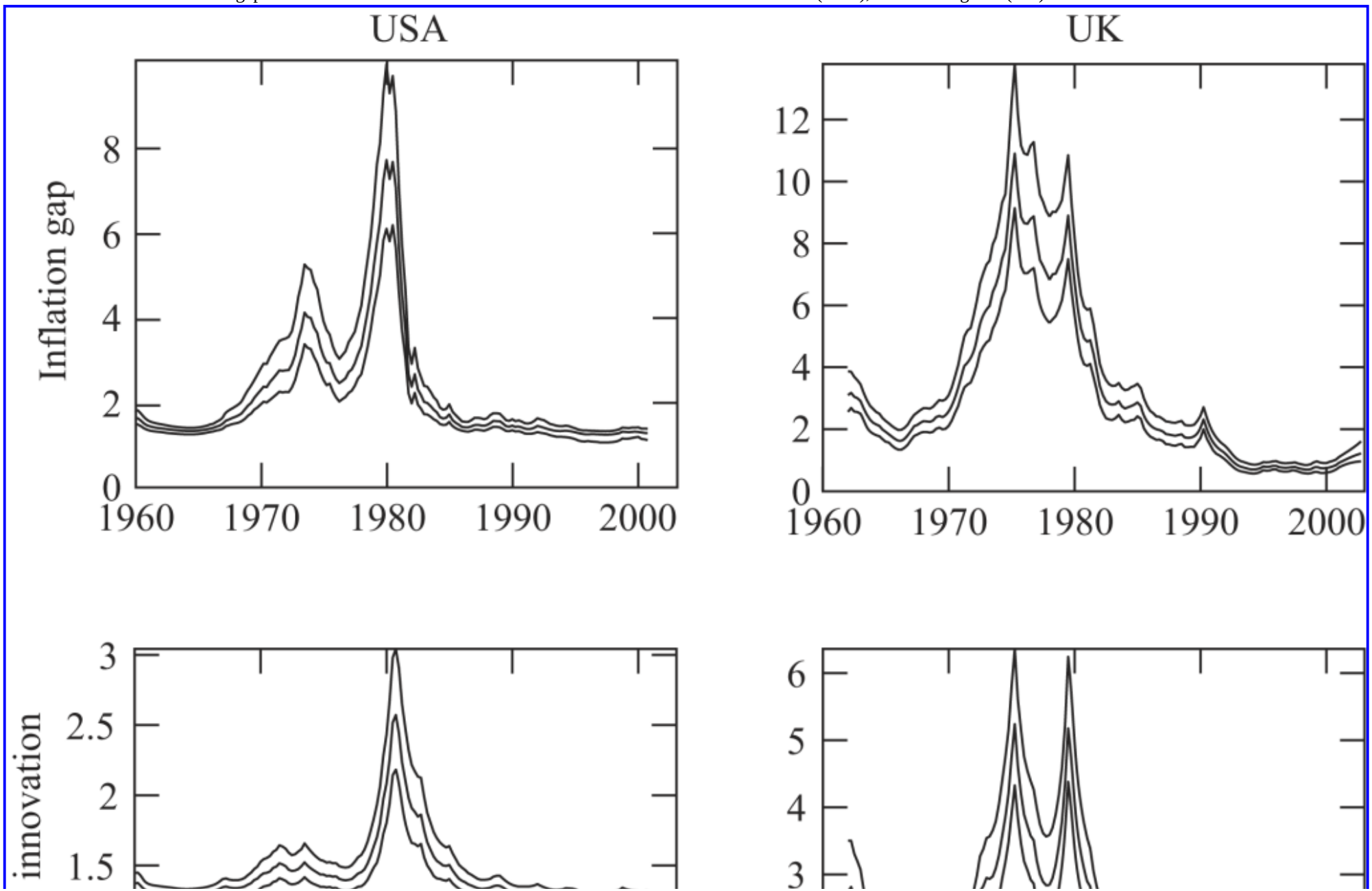

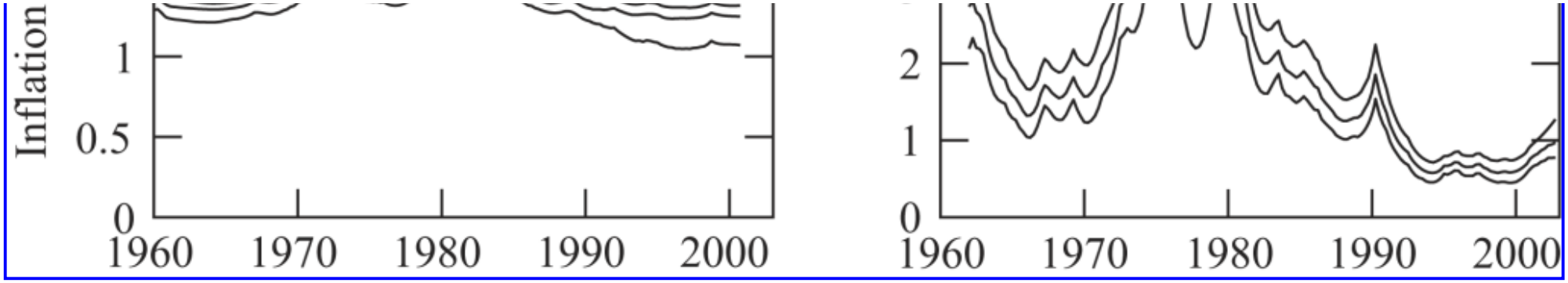

Along with Figure 1, Figure 3 depicts a positive correlation between inflation gap volatility and trend inflation. Inflation volatility was low at the beginning and ends of the samples when trend inflation was low, and it was high during the great inflation of the 1970s. Thus, not only has average inflation declined since the 1970s, but so has inflation gap volatility. Moreover, this did not come at the expense of an increase in output or unemployment volatility, which follow much the same trajectory as for inflation. Sticky price models often predict a trade-off between the variability of inflation and real variables, but that trade-off is not apparent here. Instead we see a simultaneous decline in both after 1980.

Whether the greater stability of inflation and output is the result of better policy or better luck (that is, smaller shocks) is the subject of much current research. The absence of a volatility trade-off suggests that better luck is a promising candidate, however, for smaller shocks would deliver a simultaneous decline in both in standard models.

Inflation can be volatile either because shocks are volatile or because they are persistent. Thus, we can drill down by examining innovation variances and measures of persistence. In the bottom row of Figure 3, I report the standard deviation of one-step ahead VAR prediction errors for inflation. For the UK, the pattern is the same as in the other figures: prediction errors were large in magnitude during the 1970s and smaller before and after. For the USA, there was only a slight increase during much of the 1970s, but a sharp spike during the brief window in 1979–80 when the Federal Reserve was targeting monetary aggregates.

To the extent that monetary policy affects inflation with long and variable lags, these pictures also hint that good luck in the form of smaller shocks is part of the story. But the movements in innovation variances do not necessarily disprove the bad-policy story, for better policy can take the form of smaller policy shocks. It is also conceivable that better policy could damp the impact of non-policy shocks. Perhaps more importantly, if policy in the 1970s was so bad that sunspots affected equilibrium outcomes, then better policy could eliminate one of the shocks altogether (that is, the sunspot), and that would reduce the VAR prediction error variance for inflation and other variables.

Stock and Watson (2005) also report a decline in the prediction-error variance after the great inflation, but they point out another sense in which inflation has simultaneously become harder to forecast. Consider the R^2 statistic for the VAR forecast of inflation,

$$1 - R_t^2 = \frac{\sigma_t^2(\pi_t - E_{t-1}\pi_t)}{\sigma_t^2(\pi_t - \bar{\pi}_t)}. \tag{11}$$

The numerator is the VAR innovation variance shown in the bottom row of Figure 3, and the denominator is the total variance depicted in the top row. Since both terms of the ratio decline after the great inflation, it is not obvious whether the R^2 statistic has increased or decreased. One-step ahead prediction errors are smaller after 1980, but so is the total amount of transient variation that one hopes to predict.

As shown in Figure 4, for our time-varying VARs the denominator actually falls by more, so that the R^2 for inflation declines. Furthermore, this decline is statistically significant. For example, for the USA the posterior probability of a decline in R^2 between 1980 and 2000 is 0.998. Stock and Watson report a similar finding for an unobserved-components model of inflation. Thus, inflation has become both more and less predictable: inflation forecast errors are smaller in absolute value, but they account for a larger proportion of inflation-gap variability.

Figure 4

Predictability of the inflation gap. *Sources*: Federal Reserve Economic Database (USA), Bank of England (UK).

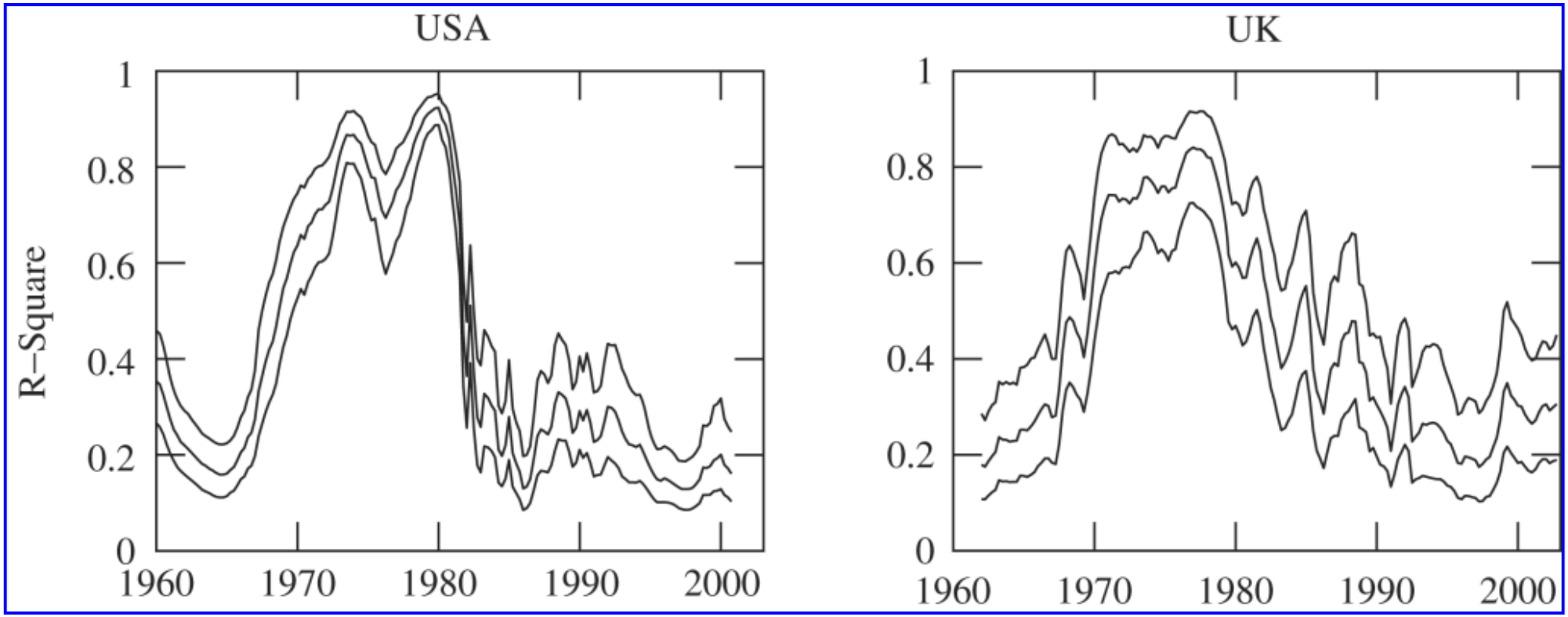

This means the inflation gap has become less autocorrelated and also less cross-correlated with lags of other macroeconomic variables. In other words, it is closer to a martingale-difference variate. In Figure 5, I summarize changes in inflation gap persistence by graphing its normalized spectrum,

$$g_{\pi\pi}(\omega, t) = \frac{2\pi f_{\pi\pi}(\omega, t)}{\int_{-\pi}^{\pi} f_{\pi\pi}(\omega, t)\, d\omega}. \tag{12}$$

In the numerator, $f_{\pi\pi}(\omega, t)$ represents the instantaneous power spectrum,

$$f_{\pi\pi}(\omega, t) = e_{\pi}'(I - A_t e^{-i\omega})^{-1} \frac{R_t}{2\pi} (I - A_t e^{i\omega})^{-1\prime} e_{\pi}. \tag{13}$$

The denominator is the instantaneous variance of $\pi_t - \bar{\pi}_t$. Thus, $g_{\pi\pi}(\omega, t)$ measures autocorrelation rather than autocovariance. I also multiply by 2π so that the units are easy to interpret. In these units, a white noise variate has $g_{\pi\pi}(\omega) = 1$ at all frequencies. Relative to that benchmark, excess power at low frequencies represents positive autocorrelation, and excess power at high frequencies signifies negative autocorrelation. If the ordinate at frequency zero is less than 1, then the price level is partially mean reverting (Cochrane, 1988).

Figure 5

Normalized spectrum for the inflation gap. *Sources*: Federal Reserve Economic Database (USA), Bank of England (UK).

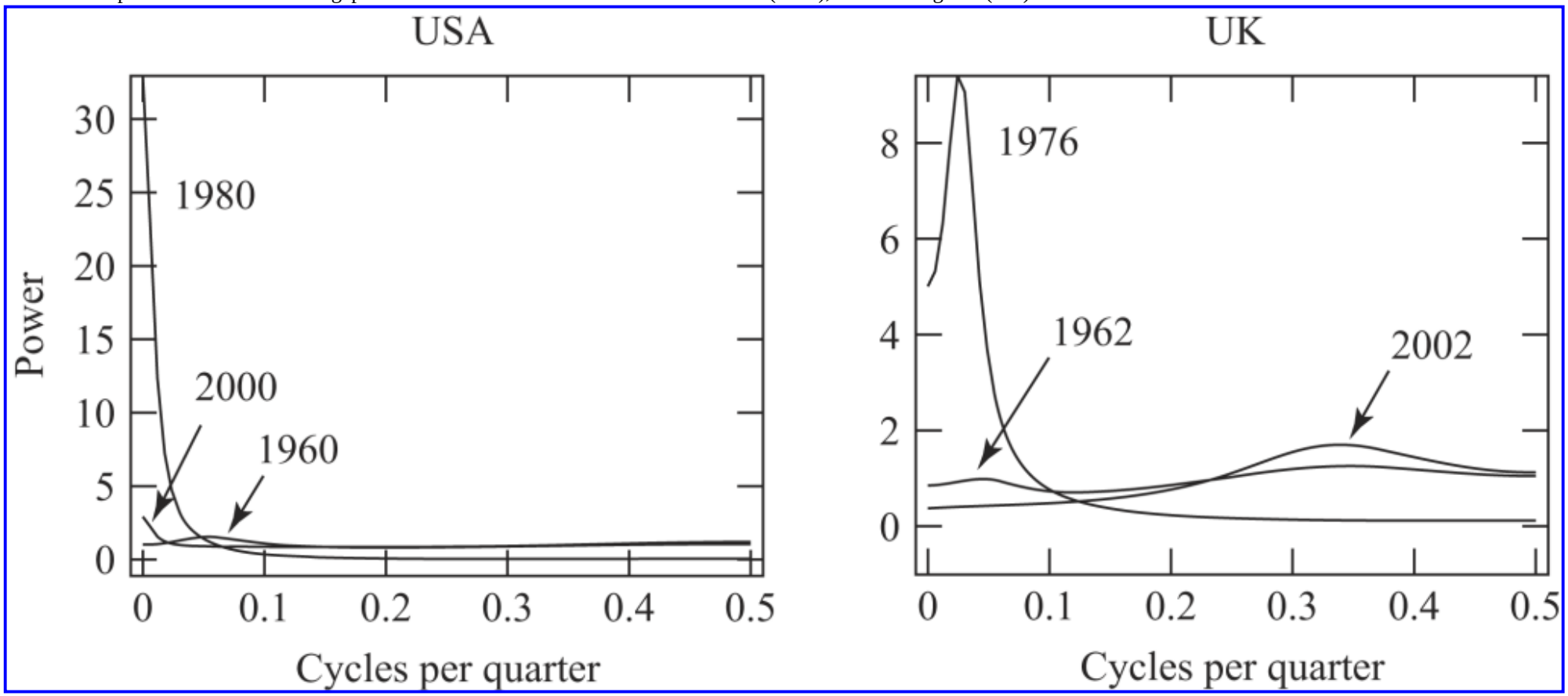

In the early 1960s, the spectrum was relatively flat in both countries, and the inflation gap was not far from being white noise. The gap became more persistent by the mid- to late 1970s, however, with power concentrated at frequencies of eight years per cycle or longer. This signifies the presence of substantial transient fluctuations in inflation. Evidently, the monetary authorities were permitting inflation fluctuations to go unchecked for years at a time, only gradually bringing π_t back toward $\bar{\pi}_t$. These policies were reversed after the early 1980s, and by the end of the sample the spectrum had again become relatively flat. Thus, we also see a positive correlation between trend inflation and persistence of the inflation gap. Trend inflation was low and the gap was weakly persistent at the beginning and the end of the sample, and they were high and strongly persistent, respectively, in the middle.

For the USA, Cogley and Sargent (2005b) and Primiceri (2005b) explain this association in terms of changing Fed beliefs about the sacrifice ratio. In Cogley and Sargent's model, the central bank wants to reduce inflation in the 1970s, but it wants to move very slowly. Their hypothetical central bank prefers gradualism because it puts some weight on Keynesian Phillips-curve models which at that time predicted intolerable sacrifice ratios – much higher than the predictions of the same models in the 1980s or 1990s. Thus, when inflation was highest, optimal policy called for an extremely gradual adjustment towards the target, making the inflation gap highly persistent.

Cogley and Sargent's story gains credibility when one reviews the analyses of leading policy economists from the late 1970s. For example, Arthur Okun (1978, p. 284) wrote that ‘recession will slow inflation, but only at the absurd cost in production of roughly $200 billion per point’. At that time, $200 billion amounted to roughly ten per cent of GDP, and, if we extrapolate Okun's estimate to zero inflation, the total cost amounts to three-quarters of a year's GDP. Like the central bank in Cogley and Sargent's model, Okun recommended gradualism in the 1970s because he thought the cost of aggressive actions would be exorbitant.

This explanation dovetails nicely with the work of Orphanides (2001; 2003), who demonstrates that the Fed overestimated the magnitude of the output gap in the 1970s because it was slow to detect the productivity slowdown. Because the estimated output gap was too big, they also initially exaggerated the amount of disinflation that would ensue. When that disinflation failed to materialize, they became pessimistic about the amount of slack needed to slow inflation, concluding that the sacrifice ratio was bigger than previously thought. Output gap misperceptions are not an element of Cogley and Sargent's model (their hypothetical central bank is better at filtering than the Fed was), but it may be an important part of

the bigger picture.

In retrospect, the high estimates of sacrifice ratios in the 1970s may seem excessive because current estimates are quite a bit lower. Indeed, that is probably one reason why central banks now react more strongly to inflation. In any event, what matters for understanding monetary policy in the 1970s was what economists believed then, not what we believe now with the benefit of hindsight.

Finally, it is interesting to contrast the shape of the spectrum for the USA and the UK at the end of the sample. For the UK, the spectrum has a trough at frequency zero and a gentle positive slope, a shape that signifies partial mean reversion in the price level. For the USA, there is still a peak above 1 at frequency zero and a downward sloping spectrum, hence no mean reversion in the price level.

The contrast is noteworthy because it connects with questions about optimal monetary policy. In a textbook version of a dynamic New Keynesian model, the first-order condition for optimal policy is

$$\pi_t = -\frac{\lambda}{\kappa}(x_t - x_{t-1}), \tag{14}$$

where x_t is the output gap and λ and κ are parameters (Woodford, 2003, ch. 7). Because x_t is a stationary random variable, the right-hand side is over-differenced, implying that optimal policy induces mean reversion in the price level. Woodford explains that a partially mean-reverting price level is a feature of optimal policy in many versions of the New Keynesian model, because a credible commitment on the part of the central bank to roll back future price increases restrains a firm's incentive to increase its price today. To make that promise credible, the central bank must follow through by taking actions to reverse realized movements in the price level. The end-of-sample UK inflation spectrum implies a partial rollback of the price level, but the US inflation spectrum does not.

Conclusion

During the great inflation of the 1970s, inflation outcomes worsened in many dimensions. Inflation was higher on average, more volatile and more persistent. There was more uncertainty about the central bank's long-run target for inflation and also more uncertainty about where inflation would be one quarter ahead. All of that has been reversed, possibly because of improved monetary policy rules, possibly because we have not experienced the severe adverse supply shocks that central bankers had to contend with in the 1970s. Sorting out the reasons behind the improvement in inflation outcomes is the subject of much ongoing research.

See Also

- inflation
- inflation targeting
- monetary policy, history of

Bibliography

Benati, L. and Mumtaz, H. 2006. The great stability in the UK: good policy or good luck? Working paper, Bank of England.

Beveridge, S. and Nelson, C. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* 7, 151–74.

Canova, F. and Gambetti, L. 2006. Structural changes in the U.S. economy: bad luck or bad policy? Discussion Paper No. 5457. London: CEPR.

Clarida, R., Gali, J. and Gertler, M. 2000. Monetary policy rules and macroeconomic stability: evidence and some theory. *Quarterly Journal of Economics* 115, 147–80.

Cochrane, J. 1988. How big is the random walk in GNP? *Journal of Political Economy* 96, 893–920.

Cogley, T., Morozov, S. and Sargent, T. 2005. Bayesian fan charts for U.K. inflation: forecasting and sources of uncertainty in an evolving monetary system. *Journal of Economic Dynamics and Control* 29, 1893–925.

Cogley, T. and Sargent, T. 2001. Evolving post-World War II U.S. inflation dynamics. In *NBER Macroeconomics Annual 16*, ed. B. Bernanke and K. Rogo. Cambridge: MIT Press.

Cogley, T. and Sargent, T. 2005a. Drifts and volatilities: monetary policies and outcomes in the post World War II U.S. *Review of Economic Dynamics* 8, 262–302.

Cogley, T. and Sargent, T. 2005b. The conquest of U.S. inflation: learning and robustness to model uncertainty. *Review of Economic Dynamics* 8, 528–63.

Lubik, T. and Shorfheide, F. 2004. Testing for indeterminacy: an application to U.S. monetary policy. *American Economic Review* 94, 190–217.

McConnell, M. and Perez Quiros, G. 2000. Output fluctuations in the United States: what has changed since the early 1980s? *American Economic Review* 90, 1464–76.

Okun, A. 1978. Discussion. In *Curing Chronic Inflation*, ed. A. Okun and G. Perry. Washington, DC: Brookings Institution.

Orphanides, A. 2001. Monetary policy rules based on real-time data. *American Economic Review* 91, 964–85.

Orphanides, A. 2003. The quest for prosperity without inflation. *Journal of Monetary Economics* 50, 633–63.

Primiceri, G. 2005a. Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies* 72, 821–52.

Primiceri, G. 2005b. Why inflation rose and fell: policymakers’ beliefs and U.S. postwar stabilization policy. Working Paper No. 11147. Cambridge, MA: NBER.

Stock, J. and Watson, M. 2005. Has inflation become harder to forecast? Mimeo, Harvard University and Princeton University.

Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39, 195–214.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

Cogley, Timothy. "inflation dynamics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000272> doi:10.1057/9780230226203.0792

The New Palgrave Dictionary of Economics Online

inflation expectations

Bennett T. McCallum

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Agents' expectations about future values of inflation play an important role in macroeconomic analysis. From a steady-state perspective, higher expected inflation rates induce agents to hold smaller real money balances and, in most models, to hold different amounts of capital. In dynamic analysis, inflationary expectations affect agents' decisions regarding saving and price adjustments, and affect monetary policy behaviour in ways that have become increasingly important. Over time, analysts' treatment of expectations evolved from distributed-lag, adaptive models to rational expectations, a change that had major analytical implications. Analysis of learning behavior has become more prominent, supplementing or occasionally replacing rational expectations.

Keywords

Bretton Woods System; commodity money; distributed lags; E-stability; fiat money; Fisher, I.; hyperinflation; infinite horizons; inflation; inflation targeting; inflationary expectations; learning in macroeconomics; leisure; long-term interest rates; lump-sum taxes; monetary policy; monetary policy rules; neutrality of money; New Keynesian macroeconomics; New Neoclassical synthesis; Phillips curve; rational expectations; real business cycles; real vs nominal interest rates; recursive least squares learning; stabilization policy; steady-state theorizing; sticky prices; Thornton, H.; time preference; transversality condition; Wicksell, K.

Article

The concept of inflationary expectations came to the fore in the work of Chicago School economists during the 1950s, with notable contributions including those of Cagan (1956), Bailey (1958), and Friedman (1960; 1969) on hyperinflation experiences, the cost of inflation, and the optimal steady-state inflation rate. The upsurge of inflation experienced in many countries following the 1971–3 demise of the Bretton Woods System led to additional interest, which was increased again by the spread of rational

expectations analysis during the 1970s. Yet another boost in prominence came from the widespread influence of monetary policy strategies based on the notion of inflation targeting, beginning around 1990 and continuing unabated as of 2006. Major developments in technical analysis relating to monetary policy have lent additional interest to inflationary expectations in various ways that are touched upon below.

The following exposition begins by considering ways that inflationary expectations are important in terms of comparative steady-state analysis, pertaining to ‘long-run’ phenomena, before turning to expectations’ role in dynamic analysis that corresponds to cyclical fluctuations. Next, the manner in which expectations are formed is discussed, with emphasis on the rational expectations hypothesis and also on recent attempts to depart in a disciplined manner from the strict rationality requirement. Finally, a brief historical note is included.

Steady-state effects

From a steady-state perspective it is natural to presume that actual and expected rates of inflation (and other variables) coincide, so it is common to discuss the welfare cost of inflation, super-neutrality, and so on, in terms of actual rather than expected inflation. Most of the allocational effects are, however, attributable in principle to expected rather than realized inflation. Even in an economy in which the real rate of interest is invariant to expected inflation, the nominal interest rate – and therefore the quantity of real money balances held – will be influenced by these expectations. In particular, a relatively high expected inflation rate will induce individuals to hold (*ceteris paribus*) relatively small shares of their wealth in the form of money (which, as the medium of exchange, pays its holders interest at a lower rate – often assumed to be zero – than other assets). Consequently, since reduced real money balances entail reduced quantities of the transaction-facilitating services that are provided by the medium of exchange, agents are required to devote relatively more time and/or resources to the activity of ‘shopping’, that is, conducting transactions. In addition, the volume of transactions conducted may fall. A reduced level of utility is then the consequence for each individual agent, *ceteris paribus*, of an increased rate of expected inflation. In two classic contributions, Friedman (1960; 1969) argued that, on the assumption that there are virtually no resource costs associated with the creation and management of fiat money, overall efficiency requires a rate of expected inflation that drives the opportunity cost of holding money to zero and thereby satiates agents with the transaction-facilitating services of money. An exposition that extends the argument to models with finite-lived agents and considers the modification needed when lump-sum tax/transfers are not feasible is provided by McCallum (1990).

In many well-articulated models, expected inflation also has steady-state effects on other real macroeconomic variables – that is, money is not ‘superneutral’ (Barro and Fischer, 1976). In models with finite-lived agents, for example, the steady-state real rate of interest will be affected by inflationary expectations and, consequently, so will the per capita stock of capital and rate of consumption. But, even if individuals are modelled as having infinite time horizons and a fixed rate of time preference – features which (together with exogenous growth rates) fully determine the steady-state real rate of interest – capital and consumption per capita will under most specifications depend (though probably weakly) on the expected inflation rate. (The well-known model of Sidrauski, 1967, provides an exception, but only because it ignores individuals’ desire for leisure.) In sum, the magnitude of inflationary expectations

may have significant allocative consequences, even if one neglects the practically important effects of tax schedules that are set in nominal terms. Such allocative effects are in principle operative also at business-cycle frequencies, but the large magnitude of the capital stock/investment ratio in developed countries leads to the presumption that these effects are of quantitative significance only over longer spans of time.

Dynamic effects

There are three main ways in which expectations of future inflation affect period-by-period equilibria in dynamic analyses with typical macroeconomic models. First, intertemporal decisions depend significantly on *real* rates of interest, which are nominal rates adjusted for expected inflation. Second, expected inflation rates are important determinants of price-setting behaviour in most models in which there is some form of nominal price stickiness, reflecting a failure of prices to adjust immediately to values that would prevail under full flexibility. Third, monetary policy decisions may be based in substantial part on expected inflation rates, as with the strategy of ‘inflation forecast targeting’ that has been prominent in recent years (for example, Bernanke et. al., 1999; Svensson and Woodford, 2005). With respect to the second of these, there has been much disagreement over the best way to represent departures from full price flexibility. (Indeed, an important school of macroeconomic thought adheres to the real-business-cycle view that it is best to assume full flexibility.) In recent years (for example, 1998–2006) variants of the Calvo (1983) price-adjustment scheme have been most prominent, but over the years specifications due to Lucas (1972a; 1973), Fischer (1977), Taylor (1980), Mankiw and Reis (2002), and others have also attracted significant support. Some of the models advanced in the 1970s imply that any real stimulus resulting from inflation will be smaller, the greater is the extent to which this inflation was previously expected. Indeed, a prominent and important line of thought originated by Friedman (1966; 1968) and Phelps (1967) contends that inflation will provide a stimulus to output and employment (via the so-called Phillips-curve relation) *only* to the extent that it is unexpected. As the validity of that viewpoint - that there is no long-lasting trade-off between unemployment and inflation – is highly relevant for stabilization policy, many attempts were made to conduct statistical tests. The appropriate design of such tests will, of course, depend significantly on the way in which expectations are formed, a matter discussed below.

Rational expectations

As mentioned above, for steady states it is natural to presume that expected inflation rates will match those actually realized, and virtually all contemporary steady-state theorizing proceeds under that assumption. Analysis of quarter-to-quarter or year-to-year movements requires, however, some more ambitious formulation concerning expectational behaviour. From the time of Cagan's (1956) study of hyperinflations until the mid-1970s, the most widely used hypothesis was that of adaptive expectations – which makes each period's change in the relevant variable proportional to the most recent expectational error – with other autoregressive representations also used to some extent. During the 1970s it became clear, however, that adaptive and other fixed autoregressive specifications permit the occurrence of repeated, systematic expectational *errors*. But, since such errors are costly to the individual agents who

make them, standard neoclassical reasoning suggests that it would be analytically fruitful to assume that agents typically eliminate any systematic source of expectational error, subject to available information. This hypothesis of rational expectations was introduced by Muth (1961) and developed in a macroeconomic context by Lucas (1972a; 1973) and Sargent (1973). It met with some initial resistance, perhaps because of a mistaken impression that it implies homogeneity of information and expectations across agents and/or that activist macroeconomic stabilization policy must necessarily be ineffective. Scepticism about agents' cognitive abilities also played a role, probably. But by the end of the 1970s the rational expectations hypothesis – implying that an agent's expectational errors are uncorrelated over time with all elements of his information set – had become dominant in both theoretical and applied macroeconomics.

The early development of techniques for the econometric implementation of rational expectations involved attempts to test the Friedman–Phelps no-trade-off hypothesis mentioned above. Various estimates of the crucial slope parameter attached to the expected-inflation variable in a Phillips-type relationship had been obtained, during the late 1960s and early 1970s, with econometric procedures relying upon the assumption of adaptive expectations (or fixed autoregressive expectations with lag weights summing to 1.0). Typical estimates of the slope parameter obtained in these studies were in the vicinity of 0.4–0.6, well below the value of unity implied by the Friedman–Phelps theory (for example, Solow, 1969). It was shown analytically by Sargent (1971) and Lucas (1972b), however, that the test strategy utilized would not identify the parameter at issue if expectations are in fact formed rationally. Instead, the estimate would tend to equal this parameter value times the sum of lag coefficients in a univariate forecasting equation for the inflation rate, a sum that need not equal the value of 1.0 presumed by the procedure in question. Estimates using similar (quarterly, US) data-sets but taking account of this insight were then found to yield values close to unity (McCallum, 1976). The resulting interpretation – that the true parameter value is approximately unity and that expectations are at least approximately rational – subsequently received indirect support from additional estimates presuming fixed autoregressive expectations, as the values obtained rose over time during the 1970s (Gordon, 1976). As the univariate autoregressive representation of actual inflation was also changing during this period, with the sum of lag coefficients rising from around 0.5 to nearly 1.0, these findings accorded well with the Sargent–Lucas interpretation of the evidence.

As important implication of the Sargent–Lucas analysis is that, if expectations are in fact rational, one cannot generally measure the 'long-run' effect (that is, comparative steady-state effect) of one variable on another by the sum of coefficients in a distributed-lag relationship. For example, since expected inflation affects interest rates to a different extent from unexpected inflation, the sum of coefficients in a distributed-lag regression of interest on inflation will depend on the stochastic properties of inflation (the variable being forecast) as well as the slope coefficient measuring the effect of expected inflation on interest. To test hypotheses about the latter effect, it is necessary to take some account of the type of process generating the variable being forecast. That this principle continues to obtain when frequency-domain statistical techniques are employed was emphasized by Whiteman (1984) and McCallum (1984), but King and Watson (1992) and Fisher and Seater (1993) developed procedures that can be used for model-free tests if the monetary policy rule in force (over the sample period) is such that it generates unit-root behaviour of the log of the money stock.

During the second half of the 1990s, expectations came to play an increasingly important role in

monetary policy analysis as ‘New Keynesian’ or ‘New Neoclassical Synthesis’ models, firmly based on optimizing analysis while incorporating sluggish price adjustments, became the norm for researchers in academia and central banks alike (Goodfriend and King, 1997; Rotemberg and Woodford, 1997; Clarida, Galí and Gertler, 1999; Woodford, 2003). In these models with forward-looking expectations in the price-adjustment equations, optimal policy requires ‘history-dependent’ rules (Woodford, 2003) that take account of expectations in a manner not recognized in traditional optimal-control analysis. Various developments, including consideration of issues implied by the zero lower bound on nominal interest rates, made policy analysis increasingly a matter of ‘managing expectations’ (Eggertsson and Woodford, 2003). From the perspective of actual policy practice rather than theory, Goodfriend and King (2005) present documentation, based on Federal Open Market Committee transcripts, indicating that as early as November 1979 the committee was using long-term interest rates as an indicator of inflationary expectations, which were being used to help guide the disinflation of 1979–84. This episode has, more recently, come to be widely regarded as a major turning point in the remarkable worldwide reduction in inflationary difficulties that took place between the late 1970s and the early 1990s.

Issues regarding expectations

The principle mentioned above, involving distributed-lag coefficient sums, can remain true even if expectations are not strictly rational. In particular, it will apply if expectations are formed in a manner that reflects full but delayed responsiveness to the properties of the generating system. Expectational behaviour of that type, which might be termed ‘asymptotic rationality,’ can be expressed analytically by the condition that the unconditional mean of the expectational error process equals zero, a weaker requirement than that the error must be uncorrelated with all information variables available at the time of expectation formation. This less stringent type of partial rationality has not been prominent to date, but may become important eventually. It is not analytically similar, it should be said, to hypotheses involving *learning* – that is, changing perceptions over time regarding the structure of the system. During recent years, analysis of learning behaviour has developed into an important and influential ingredient in reasoning about expectations, in two different respects. First, it is a much-noted fact that in most monetary macro-models there is a multiplicity (that is, two or more) of rational expectations (RE) solutions that are dynamically stable, that is, non-explosive. (Dynamically unstable solutions can usually be ruled out by recognition of a transversality condition that is relevant in the explicit or implicit optimization problem solved by the model's agents.) A widely held point of view is that in cases of indeterminacy – that is, two or more stable RE solutions – there is a presumption of substantial non-optimality for that reason alone, so that (for example) a monetary policy rule that permits such indeterminacy should be strenuously avoided (see, for example, Benhabib and Farmer, 1999; Woodford, 2003.) An alternative possibility, expressed most explicitly by McCallum (2003), is that in such cases only a single stable RE solution may be economically relevant. From that perspective there is evidently good reason to believe that a necessary (not sufficient!) condition, for a particular RE solution to be plausible, is that it be *learnable* by some process that enables individual agents in an economy to obtain empirical information about the parameters that govern the behaviour of the economy. (It seems implausible that individuals could obtain such information by processes that do not involve inference from data generated by the economy.) Extensive study of such processes has been one feature of a large

body of work summarized in the influential treatise of Evans and Honkapohja (2001). The leading contender for a learning process for this first purpose is recursive least squares learning, with the issue being whether such a learning process converges to a RE solution as time passes and an unlimited quantity of data is accumulated. There exist alternative learning algorithms, of course, but the one in question is in several respects specified so as to be highly conducive to learnability, so that if a RE solution is not learnable by this procedure then it should not be regarded as a plausible candidate for an economically relevant equilibrium. This point of view may, in some cases, eliminate concerns regarding solution multiplicity. In this first type of learning analysis, the concept of E-stability – due to DeCanio (1979), Evans (1986), Marcet and Sargent (1988), and Evans and Honkapohja (1992) – is used extensively, as it provides a convenient technique for determining the learnability status of particular RE solutions. A notable application to monetary policy rules is Bullard and Mitra (2002).

A second, quite different, and more ambitious application of learning algorithms is to represent distinct hypotheses, as alternatives to RE, concerning the formation of expectations. This line of work has been pursued extensively by Evans and Honkapohja (2001) and numerous other researchers including Sargent (1993) and Orphanides and Williams (2005). An attractive feature of this approach is that it views expectational behaviour as departing somewhat from full, strict expectational rationality, but nevertheless retaining much of the intellectual discipline imposed by the RE requirement that behaviour regarding expectation formation be governed by the same optimizing benchmark that characterizes neoclassical economics more generally. From the substantive perspective, the use of learning processes (such as constant-gain modifications of least-squares learning that have the effect of down-weighting older data) in place of RE in calibrated macroeconomic models often gives rise to additional serial correlation in endogenous variables, thereby generating model properties that are viewed by many analysts as being more nearly consistent with actual macro time-series data.

Historical considerations

The foregoing discussion is somewhat historical in nature yet includes no references to literature predating the Second World War. Is there some explanation for this absence? The most celebrated discussion of inflationary expectations in the ‘pre-war’ literature is, probably, that of Irving Fisher. In *Appreciation and Interest* (1896), Fisher emphasizes the real versus nominal interest rate distinction that is often associated with his name, and in *The Theory of Interest* (1930) he estimates a distributed-lag regression relating interest to current and past inflation rates (interpreting the long lags as due to ‘delayed adjustment’). In addition, several other economists (such as Marshall, 1890) devoted some attention to the effects of expected inflation, the contribution of Henry Thornton (1802) being perhaps the most prescient (on this topic, see Humphrey, 1983). All in all, however, it seems that the subject attracted little attention in the pre-war literature. Even in Knut Wicksell's famous analysis of the ‘cumulative process’ of inflation, there is only brief passing mention (1898, pp. 96, 148) of the possibility that the inflation will be anticipated. Discussion of the effects of sustained inflationary expectations on capital formation seems to be entirely absent. This neglect may perhaps be satisfactorily explained by first noting that it is sustained inflation that is relevant and then recalling that during this earlier era the world's major economies normally adhered to some commodity-money standard, thereby sharply reducing the scope for substantial inflation to arise or to be sustained.

See Also

- adaptive expectations
- expectations
- inflation
- inflation dynamics
- learning in macroeconomics
- Lucas critique
- new classical macroeconomics
- Phillips curve
- rational expectations

Bibliography

Bailey, M.J. 1956. The welfare cost of inflationary finance. *Journal of Political Economy* 64(2), 93–110.

Barro, R.J. and Fischer, S. 1976. Recent developments in monetary theory. *Journal of Monetary Economics* 2, 133–67.

Benhabib, J. and Farmer, R.E.A. 1999. Indeterminacy and sunspots in macroeconomics. In *Handbook of Macroeconomics*, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Bernanke, B.S., Laubach, T., Mishkin, F.S. and Posen, A.S. 1999. *Inflation Targeting: Lessons from the International Experience*. Princeton: Princeton University Press.

Bullard, J. and Mitra, K. 2002. Learning monetary policy rules. *Journal of Monetary Economics* 49, 1105–29.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Clarida, R., Galí, J. and Gertler, M. 1999. The science of monetary policy: a New Keynesian perspective. *Journal of Economic Literature* 37, 1661–707.

DeCanio, S. 1979. Rational expectations and learning from experience. *Quarterly Journal of Economics* 94, 47–57.

Eggertsson, G.B. and Woodford, M. 2003. The zero bound on interest rates and optimal monetary policy. *Brookings Papers on Economic Activity* 2003(1), 139–233.

Evans, G.W. 1986. Selection criteria for models with non-uniqueness. *Journal of Monetary Economics* 18, 147–57.

Evans, G.W. and Honkapohja, S. 1992. On the robustness of bubbles in linear RE models. *International Economic Review* 33, 1–14.

Evans, G.W. and Honkapohja, S. 2001. *Learning and Expectations in Macroeconomics*. Princeton: Princeton University Press.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85 (February), 191–205.

Fisher, I. 1896. *Appreciation and Interest*. New York: American Economic Association.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Fisher, M.E. and Seater, J.J. 1993. Long-run neutrality and superneutrality in an ARIMA framework. *American Economic Review* 83, 402–15.

Friedman, M. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, M. 1966. Comments. In *Guidelines, Informal Controls, and the Market Place*, ed. G.P. Shultz and R.Z. Aliber. Chicago: University of Chicago Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review Papers and Proceedings* 58, 1–18.

Friedman, M. 1969. *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

Goodfriend, M. and King, R.G. 1997. The new neoclassical synthesis and the role of monetary policy. In *NBER Macroeconomics Annual 1997*, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Goodfriend, M. and King, R.G. 2005. The incredible Volcker disinflation. *Journal of Monetary Economics* 52, 981–1015.

Gordon, R.J. 1976. Recent developments in the theory of inflation and unemployment. *Journal of*

Monetary Economics 2, 185–219.

Humphrey, T.M. 1983. The early history of the real/nominal interest rate relationship, *Federal Reserve Bank of Richmond Economic Review* 69(3), 2–10.

King, R.G. and Watson, M.W. 1992. Testing long-run neutrality. Working Paper No. 4156. Cambridge, MA: NBER.

Lucas, R.E., Jr. 1972a. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R.E., Jr. 1972b. Econometric testing of the natural rate hypothesis. In *The Econometrics of Price Determination Conference*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.

Lucas, R.E., Jr. 1973. Some international evidence on output inflation tradeoffs. *American Economic Review* 63, 326–34.

Mankiw, N.G. and Reis, R. 2002. Sticky information versus sticky prices: a proposal to replace the new Keynesian Phillips curve. *Quarterly Journal of Economics* 117, 1295–328.

Marcet, A. and Sargent, T.J. 1989. Convergence of least-squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory* 48, 337–68.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

McCallum, B.T. 1976. Rational expectations and the natural rate hypothesis: some consistent estimates. *Econometrica* 44, 43–52.

McCallum, B.T. 1984. Low-frequency estimates of long-run relationships in macroeconomics. *Journal of Monetary Economics* 14, 3–14.

McCallum, B.T. 1990. Inflation: Theory and Evidence. In *Handbook of Monetary Economics*, ed. B.M. Friedman and F.H. Hahn. Amsterdam: North-Holland.

McCallum, B.T. 2003. Multiple-solution indeterminacies in monetary policy analysis. *Journal of Monetary Economics* 50, 1153–75.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Orphanides, A. and Williams, J.C. 2005. Imperfect knowledge, inflation expectations, and monetary policy. In *The Inflation Targeting Debate*, ed. B.S. Bernanke and M. Woodford. Chicago: University of

Chicago Press.

Phelps, E.S. 1967. Phillips curves, expectations of inflation, and optimal unemployment over time. *Economica* 34, 254–81.

Rotemberg, J.J. and Woodford, M. 1997. An optimization-based econometric framework for the evaluation of monetary policy. In *NBER Macroeconomics Annual 1997*, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Sargent, T.J. 1971. A note on the accelerationist controversy. *Journal of Money, Credit and Banking* 3, 721–5.

Sargent, T.J. 1973. Rational expectations, the real rate of interest, and the natural rate of unemployment. *Brookings Papers on Economic Activity* 1973(2), 429–72.

Sargent, T.J. 1993. *Bounded Rationality in Macroeconomics*. Oxford: Oxford University Press.

Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review* 57, 534–44.

Solow, R.M. 1969. *Price Expectations and the Behavior of the Price Level*. Manchester: Manchester University Press.

Svensson, L.E.O. and Woodford, M. 2005. Implementing optimal policy through inflation-forecast targeting. In *The Inflation-Targeting Debate*, ed. B.S. Bernanke and M. Woodford. Chicago: University of Chicago Press.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*, ed. F.A. von Hayek, Fairfield: Kelley, 1978.

Whiteman, C.H. 1984. Lucas on the quantity theory: hypothesis testing without theory. *American Economic Review* 74, 742–9.

Wicksell, K. 1898. *Interest and Prices*. Trans. R.F. Kahn, London: Macmillan, 1936.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

McCallum, Bennett T. "inflation expectations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000098> doi:10.1057/9780230226203.0793

The New Palgrave Dictionary of Economics Online

inflation measurement

David E. Lebow and Jeremy B. Rudd

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Inflation measurement is the process whereby changes in the prices of individual goods and services are combined to yield a measure of general price change. This article discusses the conceptual framework for thinking about inflation measurement and considers practical issues associated with determining an inflation measure's scope; with measuring individual prices; and with combining these individual prices into a measure of aggregate inflation. We also discuss the concept of ‘core inflation’ and summarize the implications of inflation measurement for economic theory and policy.

Keywords

aggregate price level definition; asset price measurement; Cobb–Douglas functions; conditional cost-of-living index; constant elasticity of substitution; consumer price index; core inflation; cost-of-goods index; cost-of-living index; dynamic factor models; ‘exclusion’ measures of inflation; Fisher index; GDP price index; headline inflation; hedonic price functions; index numbers; indexation for inflation; inflation measurement; Laspeyres index; lifetime utility; limited-influence measures of inflation; matched model price index; measurement error; ‘neo-Edgeworthian’ index; Paasche index; quality-adjustment problem; reduced-form Phillips curve; rental equivalence; shadow prices; ‘stochastic’ approach to inflation measurement; Törnqvist index; underlying inflation

Article

Inflation measurement is the process whereby changes in the prices of individual goods and services are combined to yield a measure of general price change. In formal terms, we may specify the time-*t* rate of aggregate inflation P_t as

$$P_t = F(p_t^1, p_t^2, \ldots, p_t^I), \quad (1)$$

where $F(.)$ is a function that aggregates a set of I individual time-t price changes p_t^i. Writing the problem in this manner highlights three basic issues associated with inflation measurement. First, we must decide what collection of price changes we wish to include (or, more generally, what should be the measure's *scope*); second, we must ensure that the *individual* price changes are correctly measured; and finally, we must choose a method for combining those changes into a measure of aggregate inflation. While the problem of inflation measurement can be broadly described in these terms, dealing with the numerous complications that emerge in practice requires some explicit conceptual framework. Probably the simplest way to construct a measure of overall inflation involves defining the aggregate price level in terms of the cost of a fixed basket of goods and services. Such a measure – sometimes labeled a *cost-of-goods index* (COGI) – has several practical advantages; in particular, for a broad enough basket of goods, the change in a COGI comes very close to what most people intuitively mean by an inflation rate, and a COGI-type measure can easily be defined for any sub-component of expenditure or production (such as consumption, investment, or the output of intermediate goods). However, this simple measure of inflation faces an important practical difficulty. In a dynamic economy, the composition and nature of output will evolve as existing goods are consumed or produced in different quantities, as the characteristics of existing products change, or as entirely new goods are introduced; these changes make the COGI's fixed bundle of goods become less representative over time. The COGI approach provides no guidance as to how to address this problem, suggesting that a more comprehensive conceptual framework is needed.

If we confine our attention to consumption prices, then a natural guiding principle is provided by the concept of a *cost-of-living index* (COLI), which measures the expenditure needed for an optimizing consumer to maintain a specified level of utility as prices change. The strength of the COLI framework derives from its grounding in the theory of consumer behaviour, which can provide clear-cut suggestions (at least in principle) as to how to deal with such problems as changes in expenditure patterns or the introduction of new goods. That said, this feature of the COLI approach can also be a weakness to the extent that consumer theory provides an incorrect characterization of actual behaviour (NRC, 2002, pp. 53–8) or is insufficiently well developed to handle a particular practical situation. In addition, because the COLI concept pertains only to consumption, it provides little or no guidance about the construction of broader measures of inflation that include prices for other components of output (these might be of interest, for instance, to a monetary policymaker). The COLI framework therefore provides a natural guide to the construction of a consumer price index (CPI), which attempts to measure the prices of goods and services consumed by households; it will not, however, be able to inform the construction of a price index for overall GDP, which is defined to include the prices of *all* domestically produced final output – whether purchased by consumers, businesses, governments, or the rest of the world. (While a literature does exist on the measurement of price change from a *producer* perspective – see Diewert, 1983, for an overview – it has generally received much less attention than the corresponding consumer-based approach.)

Despite these potential shortcomings, a COLI-based approach is commonly employed as a framework for informing inflation measurement – in the United States, for example, the CPI uses the COLI concept both as its explicit measurement objective and as a reference for making practical decisions about index construction (U.S. Bureau of Labor Statistics, 2005). In much of what follows, therefore, we follow common practice in using the concept of a cost-of-living index to guide our discussion of the three basic issues – scope, individual price measurement, and aggregation – that are associated with the measurement of inflation. In addition, we discuss the concept of *core inflation*, which can be motivated and interpreted in terms of an alternative approach to inflation measurement. Finally, we conclude by considering the implications of these measurement issues for economic research and policy.

What items should be included in an inflation measure?

The scope or domain of a cost-of-goods index – whether it is defined for consumption goods or more broadly – is defined to include all items that are purchased and sold in market transactions, and, hence, that have well-defined prices. (In reality, of course, any inflation measure will include only a subset of goods consumed or produced in the economy, so sampling in order to provide a representative characterization of aggregate price change represents an important practical concern.) By contrast, the scope of a cost-of-living index is much broader than that of a corresponding COGI for consumption goods in as much as a COLI needs to account for *anything* that affects utility, including changes over time in ‘background’ or ‘environmental’ factors such as weather, pollution, crime, or the provision of public goods.

For a COLI-based measure of consumption price inflation, therefore, the relevant set of price changes $p_t^1, p_t^2, \ldots, p_t^I$ should in principle include changes in both market prices and the ‘shadow prices’ of environmental factors (with the latter defined in the sense of Pollak, 1989). In practice, however, it is almost impossible to correctly measure the effect on utility of these sorts of changes (even if we could do so, inclusion of such factors strays beyond what most people understand by the term ‘inflation’). These considerations lead to the concept of a ‘conditional’ COLI, which (to follow Pollak, 1989, again) is defined as the smallest change in expenditure that is required in order to maintain a reference utility level following a change in prices, *with the state of the environment fixed*. Although intuitive, the concept of a conditional COLI has its own conceptual difficulties. In particular, since preferences over market goods will likely depend on the environment (for example, demand for medical care depends on the incidence of disease), the rate of inflation implied by a conditional COLI will depend on the particular state of the environment that we condition on.

While the concept of a conditional COLI provides useful guidance regarding the relevant domain of a measure of consumption price inflation, it cannot unambiguously solve all questions about scope. For example, many households receive an implicit flow of services from owner-occupied housing. On the assumption that the ‘price’ of these services could be measured, it is unclear whether they should enter a COLI given that they are not generated by a market transaction or explicit expenditure (and are not closely related to a conventional notion of a price); of course, the initial home purchase does meet these criteria. A similar problem extends to a number of other goods that are consumed by households but not directly purchased by them (one example is banking services furnished without explicit charge, which are included in most national accounts’ definitions of consumption). And, again, once one moves outside

of the realm of private consumption, the conditional COLI framework provides no practical guidance regarding the construction of inflation measures for other components of production or spending (such as investment) or for broader measures of inflation (such as the GDP price index).

Finally, a particularly difficult and controversial issue concerns the proper role of asset prices in inflation measures. If we extend the theory of a cost-of-living index to an intertemporal or multi-period context (see Pollak, 1975), then expected changes in the price of future consumption streams can affect current inflation through their impact on *lifetime* utility. We can therefore consider a cost-of-living index that is defined to include current and future prices of consumption goods; furthermore, to the extent that information about future consumption prices is contained in current asset prices, an argument can be made for including these prices in a COLI-based inflation measure (Alchian and Klein, 1973). In practice, however, the volatility of asset prices – as well as the related fact that observed movements in asset prices can stem from sources unrelated to expected future consumption-price changes – typically precludes their inclusion in conventional inflation measures. (The current purchase prices of durable goods, which *are* often included, provide a partial exception.)

How should individual price changes be measured?

A number of practical problems complicate the measurement of individual price changes. First, in a modern economy the characteristics of existing goods can change over time; likewise, new goods and services will constantly be entering – and old goods leaving – production and consumption. Left unaddressed, these problems will render it impossible to track the price changes for an identical set of goods and will cause the set of goods being priced to become increasingly less representative of actual consumption and production. This will obviously affect COGI-based measures of inflation, and it will also affect COLI-based measures to the extent that changes in the characteristics or variety of available goods have an effect on the utility that is realized from their consumption.

Several techniques exist for dealing with non-trivial changes in the characteristics (loosely speaking, the ‘quality’) of existing products; all of these involve some procedure for dividing the observed price change into a component that reflects changes in the good's characteristics and a component that reflects ‘pure’ price change, where only this latter component is appropriate for inclusion in an inflation measure. (Moulton and Moses, 1997 and NRC, 2002, provide a detailed description and assessment of these various methods of quality adjustment in the context of the US CPI; see also ILO et al., 2004.) For example, when the original and modified products exist in the same period, any difference in their prices can be attributed to differences in the goods’ characteristics. Alternatively, in the more common case where a good exists in one form in period *t* and in another in period $t+1$, the ‘pure’ price change over the intervening period can be imputed from the observed average price change for a similar group of goods. (A ‘matched model’ index, which only includes price changes for goods that remain in the sample without change – and so implicitly assigns that average price change to other items – is a common example.) Finally, additional information may be brought to bear on the problem: under certain assumptions, for example, data on the cost to manufacturers of modifying the characteristics of a product can be used to compute the effect of these modifications on the good's price.

When detailed information about a product's characteristics is available, so-called ‘hedonic’ methods may be used. The hedonic approach relates the observed price of a good to its characteristics; any

change in characteristics can then be explicitly controlled for and removed from the good's total price change. Specifically, when the individual effects of a good's characteristics on its price are stable over time, a measure of pure price change can be obtained by permitting the level of the price-characteristics relation to shift in each period. In the more realistic case where different hedonic functions exist for each period, a measure of pure price change between periods *t* and $t+1$ can be defined as $h^{t+1}(z_t)/h^t(z_t)$ or $h^{t+1}(z_{t+1})/h^t(z_{t+1})$, where $h^i(z)$ denotes the hedonic function in period *i* relating the good's price to its set of characteristics *z*. (Here, the first expression yields a ‘Laspeyres-like’ price measure as the hedonic function is evaluated with the set of characteristics from the variety that is purchased in the base period; similarly, the second expression yields a ‘Paasche-like’ measure.)

An important advantage of the hedonic approach to dealing with quality change is that it can be explicitly grounded in cost-of-living theory. Under relatively weak conditions, $h^{t+1}(z_t)-h^t(z_t)$ provides an upper bound for the compensating variation associated with a given price change; likewise, $h^{t+1}(z_{t+1})-h^t(z_{t+1})$ gives a lower bound for the equivalent variation (NRC, 2002, pp.153–4). It is unknown, however, whether these bounds are particularly tight. In addition, statistical agencies typically find real-time production of measures like these too difficult, and instead produce quality-adjusted price changes by scaling the observed price change for a good by the ratio $h^{t-j}(z_t)/h^{t-j}(z_{t+1})$, where the $t-j$ superscript makes apparent the dependence of the estimated hedonic function on an earlier period's data. Such a procedure cannot, in general, be justified in terms of a COLI-based approach (Pakes, 2002).

The ‘new goods’ problem can be thought of as a more difficult variant of the quality-adjustment problem in which the new good contains features or characteristics that have never existed before (in a sense, the dimension of the ‘characteristics space’ has increased): examples include the introduction of the video cassette recorder or cellular telephone. In this case, one needs a method for imputing the price of a newly introduced good in the period prior to its first appearance in the economy; as was suggested by Hicks (1940, pp. 114–15), one logical imputation involves setting this pre-introduction price equal to the price at which the demand for the good is just equal to zero. While such an approach can be explicitly motivated in terms of a COLI-based framework, its implementation requires a degree of information about consumer preferences that is unlikely to be realized in practice (see Hausman, 1997, for a representative example). It is therefore common for statistical agencies to attempt to mitigate the new goods problem through the more rapid addition of new items into the set of price changes being tracked over time; while intuitive, this approach may not always ameliorate the effects of new-goods introduction (Pakes, 2002).

Another problem that arises in measuring individual price changes relates to the fact that even identical goods can sell for different prices across different sellers. These differentials could reflect true price differences – a particular outlet might simply be able to charge a lower price – but they could also reflect characteristics of the outlet itself, such as customer service or convenience. In the latter case, two otherwise identical goods should be treated as different products if they are sold at different outlets; similarly, when the outlet used to price a particular good changes, some adjustment – akin to the sorts of quality adjustments discussed above – must be made to the good's price.

One final issue relating to the measurement of individual price changes is that a good's purchase price need not be related to its effect on current-period utility if it provides consumption services in more than

one period (as is the case for a durable good) or if it can be stored for later consumption. For a durable good, the conceptually relevant measure of the change in the good's price in a given period is the change in its user cost. In practice, the user cost turns out to be difficult to estimate and often implies erratic price movements. In the presence of a well-functioning rental market, the cost of hiring a good can serve as a proxy for its user cost; this ‘rental equivalence’ procedure is used in the US CPI for owner-occupied housing. However, the absence of rental markets for most durable goods limits the usefulness of this technique, and in practice the purchase prices of many durable goods are directly included in most inflation measures.

How should the individual price changes be combined?

The combination of individual price changes into an aggregate measure of inflation falls into the domain of the theory of *index numbers*, a full discussion of which is beyond the scope of this survey. We therefore focus on some of the practical issues that arise in choosing and implementing an aggregation formula.

A natural way to construct a cost-of-goods index involves weighting the individual price changes for the components of the fixed market basket by their shares in overall expenditures. When the initial period of the index is the same as the period used to specify the expenditure weights, the resulting measure corresponds to a Laspeyres index. As is well known, however, a Laspeyres index overstates changes in the cost of living when consumer substitution occurs in response to changes in relative prices; hence, alternative formulas that do capture substitution behaviour can provide a more accurate approximation to a COLI. Examples include the Törnqvist and Fisher ideal indexes (both members of the ‘superlative’ class of index numbers defined by Diewert, 1976), which employ aggregation weights derived from quantities purchased in both the initial and final periods of the comparison. Although the theory is not as well developed as that for consumer expenditures, similar justifications for commonly employed superlative aggregation formulas may exist for broader measures of output prices as well (for example, see Diewert, 1983, for a production-based interpretation of a Törnqvist index).

In addition, statistical agencies often make use of ‘chaining’ (Fisher, 1911, ch. 10; Forsyth and Fowler, 1981) when constructing long time series of inflation rates; with this procedure, the price changes implied by a sequence of indexes defined over various sub-periods are ‘chained’ or cumulated together. In the COGI context, chaining carries an intuitive or pragmatic appeal in as much as it ensures that the basket being priced will remain reasonably representative of actual consumption patterns over time. However, chaining by itself cannot correctly capture consumer substitution. (Feenstra and Shapiro, 2003, and Szulc, 1983, consider other problems that can arise with chained indexes.)

In many circumstances, price indexes must be constructed in the absence of timely data on expenditures. A superlative aggregation formula cannot be used in real time in these cases (indeed, the fact that the Laspeyres index requires only expenditures from an earlier base period accounts for much of its appeal). A compromise procedure, which requires only base-period expenditure data, involves using a weighted constant elasticity of substitution (CES) aggregator (this includes the weighted geometric mean – which measures the cost of living when utility takes a Cobb–Douglas form – as a special case). Based on historical evidence, one could form a judgement about the likely degree of substitutability across items and then use an appropriately calibrated CES formula (Shapiro and Wilcox, 1997). Such a procedure is

now employed by the US CPI to aggregate *individual* prices (that is, prices *within* item-area strata), with a geometric means formula used for the majority of cases and a Laspeyres formula reserved for strata where substitution is deemed unlikely a priori.

Accurately capturing substitution behaviour is not the only relevant issue for choosing an aggregation formula. Statistical agencies typically measure a sample of prices (where the number of price quotes for a given sub-index may be quite small), and commonly used formulas can differ in their susceptibility to small-sample biases. Indeed, Bradley (2001) has argued that small-sample bias in Laspeyres indexes – not a failure to capture substitution across categories of goods – accounts for most of the observed difference between the published (Laspeyres) version of the US CPI and a superlative (Törnqvist) variant.

At least two other issues arise in choosing how to combine individual price changes into a measure of overall inflation. First, the weights selected for use in aggregation can reflect explicit or implicit judgements as to which agents are to be represented in the index. By employing aggregate expenditure weights, the typical consumer price measure in effect gives a larger weight to the inflation rates faced by richer households – a so-called ‘plutocratic’ weighting scheme. (Alternatively, we could compute the simple average of each household-specific inflation rate; this ‘democratic’ weighting scheme might be more representative of a ‘typical’ household's experience.) For some purposes, one might also explicitly choose to measure the inflation rate faced by a particular segment of the population, such as wage earners, the poor, or the elderly.

Second, correct measurement of the quantities used in aggregation is critical. To the extent that these are subject to measurement error (as might occur if they are estimated from survey data), and to the extent that mismeasured weights are systematically associated with items that display above- or below-average price changes, the resulting aggregate inflation rate will be mismeasured. (Lebow and Rudd, 2003, present evidence of this in the US CPI.)

The concept of core inflation

Core inflation was originally defined as ‘the trend rate of increase’ of either ‘the price of aggregate supply’ or ‘the cost of the factors of production’ (Eckstein, 1981). More commonly, however, core inflation is understood in a statistical sense as corresponding either to ‘underlying inflation’ (the portion of overall inflation that is free from transitory influences) or to a measure of the common trend in all prices. In line with its various definitions, core inflation can be measured in a variety of ways.

The most prevalent core inflation measures are ‘exclusion’ measures that omit certain items, such as food and energy, from the calculation of overall inflation. The popularity of excluding food and energy derives in part from the experience of the 1970s and early 1980s, which saw sizable supply-driven price hikes for these items. Many prices other than food and energy may move erratically as well, however (indeed, some countries publish exclusion-based measures of core consumer price inflation that omit housing, the effects of changes in indirect taxes, or other items). Thus, a variant on the exclusion approach involves adjusting the weight of items in inverse proportion to their variability (sometimes termed a ‘neo-Edgeworthian’ index), so that items with erratic prices are downweighted rather than omitted entirely.

A second category of core inflation measures includes limited-influence measures such as medians or

trimmed means (Bryan and Cecchetti, 1994). These measures exclude a certain proportion of the largest and smallest price changes each period (in the extreme case of the median, all items but one are excluded each period). In contrast to standard exclusion measures, however, the omitted items will vary period by period. Limited-influence measures sometimes do well in statistical exercises aimed at finding measures that are well correlated with long moving averages of headline inflation, or measures that can serve as good univariate predictors of headline inflation. However, for these limited-influence measures to capture underlying inflation well, true relative price changes must be smaller than transitory fluctuations (which will not always be the case). In addition, construction of these measures is often sensitive to the degree of disaggregation employed and to the length of time over which the individual price changes are measured.

A third set of approaches uses econometric techniques to estimate core inflation (variously defined). For example, in an econometric reduced-form Phillips curve (as was employed in Eckstein's original study), lagged inflation terms can proxy for the persistent component of inflation once one controls for supply shocks and aggregate demand (the univariate analogue would involve taking simple or weighted averages of past inflation as the core inflation measure). Another approach is to use a dynamic factor model to extract a common component or ‘signal’ from a set of disaggregated inflation rates (Bryan and Cecchetti, 1993). Other econometric approaches have been proposed as well (often invoking economic theory to provide their rationale) – for example, core inflation may be defined as the component of inflation that is uncorrelated with long-run economic activity (Quah and Vahey, 1995), or best correlated with money growth. Of course, these theory-based underpinnings might be controversial; more generally, econometric approaches might be difficult to understand or communicate.

The neo-Edgeworthian, limited-influence, and dynamic factor approaches to measuring core inflation exemplify an alternative ‘statistical’ or ‘stochastic’ approach to inflation measurement that has garnered increased interest in recent years (Wynne, 1997). Wynne contends that the economic basis for these inflation concepts is ‘some concept of “monetary” inflation that…is not necessarily the same thing as changes in the cost of living’. If so, these alternative approaches will in principle imply different decisions about scope and aggregation relative to those implied by a COLI-based framework. In particular, to the extent that these measures seek to capture the portion of aggregate price movement that is attributable to changes in the supply of money, their relevant scope could be the price of *any* transaction that involves an exchange of money (including prices for financial assets and the purchase prices – not the user costs – for durable goods). In addition, the aggregation weights employed by these stochastic approaches are typically informed by purely statistical considerations, and so need not bear any resemblance to the weights implied by cost-of-living theory.

Implications for research and policy

Inflation measurement matters for at least three reasons. First, and most obviously, economic decisions often depend directly – even automatically – on published inflation measures. In the public sphere, many government programmes are indexed to inflation measures such as changes in a consumer price index: in the United States, for example, Social Security benefits, income tax schedules, and coupon payments on inflation-indexed government debt are all directly tied to changes in the CPI. Private contracts, including wage arrangements, are also indexed to changes in the CPI (although such indexation

provisions are less common today than they were when inflation was higher and more uncertain).
The use of inflation measures in indexation arrangements, in principle, should help inform the details of inflation measurement. If indexation of a payment is intended to maintain its real purchasing power for a recipient, then this goal is best served by using an inflation measure tailored to that recipient. Thus, indexation of pension payments would utilize an inflation measure that reflects the consumption patterns of pensioners; income-support payments would use a measure reflecting the consumption of the poor; and so on. Such specialized price indexes can differ from an aggregate price index in both the choice of priced items and in the weights assigned to them.
Inflation measurement is also important because inflation affects economic welfare and therefore serves as a goal of public policy in its own right – in particular, a central objective of monetary policymakers is the maintenance of low and stable inflation. Problems measuring the average level of inflation will therefore affect a central bank's choice of inflation target (whether explicit or implicit). For example, many argue that the Federal Reserve should seek to stabilize *measured* inflation at some level higher than zero, in part because the US CPI tends to overstate changes in the cost of living (for example, Bernanke et al., 1999). More problematically, if measurement errors in inflation vary over time in unknown ways, central banks could respond inappropriately to movements in observed inflation rates. Finally, because real quantities are typically estimated by deflating nominal values with a price index, inflation measurement directly affects the construction of other economic statistics (including real GDP and productivity). Thus, our ability to correctly assess the effects of technological progress, the sources of economic growth and changes in living standards over time hinges in an obvious way on the accurate measurement of individual and aggregate price movements. Furthermore, if the extent of measurement error in inflation varies over time and across items or places, then growth comparisons could be affected; examples include measuring changes in living standards over long periods (Gordon, 2005) and comparing growth and productivity performance in the United States and Europe (Ahmad et al., 2003).

See Also

- hedonic prices
- index numbers

Bibliography

Ahmad, N., Laquiller, F., Marianna, P., Pilat, D., Schreyer, P. and Wölfl, A. 2003. Comparing labour productivity growth in the OECD area: the role of measurement. OECD Science, Technology and Industry Statistics Working Papers, 2003/14. OECD Publishing. doi: 10.1787/126534183836.

Alchian, A. and Klein, B. 1973. On a correct measure of inflation. *Journal of Money, Credit, and Banking* 5, 173–91.

Bernanke, B., Laubach, T., Mishkin, F. and Posen, A. 1999. *Inflation Targeting: Lessons from the International Experience*. Princeton: Princeton University Press.

Bradley, R. 2001. Finite sample effects in the estimation of substitution bias in the consumer price index. *Journal of Official Statistics* 17, 369–90.

Bryan, M. and Cecchetti, S. 1993. The consumer price index as a measure of inflation. *Federal Reserve Bank of Cleveland Economic Review* 29, 15–24.

Bryan, M. and Cecchetti, S. 1994. Measuring core inflation. In *Monetary Policy*, ed. N. Mankiw. Chicago: University of Chicago Press.

Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4, 115–45.

Diewert, W. 1983. The theory of the output price index and the measurement of real output change. In *Price Level Measurement*, ed. W. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

Eckstein, O. 1981. *Core Inflation*. Englewood Cliffs, NJ: Prentice-Hall.

Feenstra, R. and Shapiro, M.D. 2003. High-frequency substitution and the measurement of price indexes. In *Scanner Data and Price Indexes*, ed. M. Shapiro and R. Feenstra. Chicago: University of Chicago Press.

Fisher, I. 1911. *The Purchasing Power of Money*. New York: Macmillan.

Forsyth, F. and Fowler, R. 1981. The theory and practice of chain price index numbers. *Journal of the Royal Statistical Society* (Series A) 144, 224–46.

Gordon, R. 2005. Apparel prices 1914–93 and the Hulten/Brueghel paradox. Working Paper No. 11548. Cambridge, MA: NBER.

Hausman, J. 1997. Valuation of new goods under perfect and imperfect competition. In *The Economics of New Goods*, ed. T. Bresnahan and R. Gordon. Chicago: University of Chicago Press.

Hicks, J. 1940. The valuation of the social income. *Economica* 7, 105–24.

ILO/IMF/OECD/UNECE/Eurostat/World Bank. 2004. *Consumer Price Index Manual: Theory and Practice*. Geneva: International Labour Office.

Lebow, D. and Rudd, J. 2003. Measurement error in the consumer price index: where do we stand? *Journal of Economic Literature* 41, 159–201.

Moulton, B. and Moses, K. 1997. Addressing the quality-change issue in the consumer price index. *Brookings Papers on Economic Activity* 1997(1), 305–66.

NRC (National Research Council). 2002. *At What Price? Conceptualizing and Measuring Cost-of-Living and Price Indexes*. Washington, DC: National Academy Press.

Pakes, A. 2002. A reconsideration of hedonic price indices with an application to PCs. Working Paper No. 8715. Cambridge, MA: NBER.

Pollak, R. 1975. The intertemporal cost of living index. *Annals of Economic and Social Measurement* 4, 179–95.

Pollak, R. 1989. The treatment of the environment in the cost-of-living index. In *The Theory of the Cost of Living Index*, ed. R. Pollak. New York: Oxford University Press.

Quah, D. and Vahey, S. 1995. Measuring core inflation. *Economic Journal* 105, 1130–44.

Shapiro, M. and Wilcox, D. 1997. Alternative strategies for aggregating prices in the CPI. *Federal Reserve Bank of St. Louis Review* 79(3), 113–25.

Szulc, B. 1983. Linking price index numbers. In *Price Level Measurement*, ed. W. Diewert and C. Montmarquette. Ottawa: Statistics Canada.

U.S. Bureau of Labor Statistics. 2005. The consumer price index. In *BLS Handbook of Methods*. Washington, DC: GPO.

Wynne, M. 1997. Commentary. *Federal Reserve Bank of St. Louis Review* 79(3), 161–7.

How to cite this article

Lebow, David E. and Jeremy B. Rudd. "inflation measurement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000294> doi:10.1057/9780230226203.0794

The New Palgrave Dictionary of Economics Online

inflation targeting

Lars E.O. Svensson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Inflation targeting is a monetary-policy strategy that was introduced in New Zealand in 1990, has been very successful in terms of stabilizing both inflation and the real economy, and as of 2007 had been adopted by more than 20 industrialized and non-industrialized countries. It is characterized by an announced numerical inflation target, an implementation of monetary policy that gives a major role to an inflation forecast and has been called ‘inflation-forecast targeting’, and a high degree of transparency and accountability.

Keywords

accountability; central banking; central banks; consumer price index; core consumer price index; credibility; fixed exchange rates; flexible exchange rates; flexible inflation targeting; forecast targeting; forecasting; inflation expectations; inflation targeting; liquidity trap; monetary policy; monetary policy instrument; monetary policy rules; monetary transmission mechanism; money supply; money-growth targeting; optimal instrument-rate paths; output gap; potential output; price stability; price-level targeting; stabilization policy; transparency; unit root

Article

Inflation targeting is a monetary-policy strategy that was introduced in New Zealand in 1990, has been very successful, and as of 2007 had been adopted by more than 20 industrialized and non-industrialized countries. It is characterized by (*a*) an announced numerical inflation target, (*b*) an implementation of monetary policy that gives a major role to an inflation forecast and has been called ‘inflation-forecast targeting’, and (*c*) a high degree of transparency and accountability.
The *numerical inflation target* is typically around two per cent at an annual rate for the Consumer Price Index (CPI) or a core CPI, in the form of a range, such as one to three per cent in New Zealand; or a point target with a range, such as a two per cent point target with a range/tolerance interval of plus/

minus one percentage points in Canada and Sweden; or a point target without any explicit range, such as two per cent in the UK and 2.5 per cent in Norway. The difference between these forms does not seem to matter in practice: a central bank with a target range seems to aim for the middle of the range, and the edges of the range are normally interpreted as ‘soft edges’ in the sense that they do not trigger discrete policy changes, and being just outside the range is not considered much different from being just inside.
In practice, inflation targeting is never ‘strict’ inflation targeting but always ‘flexible’ inflation targeting, in the sense that all inflation-targeting central banks (‘central bank’ is used as the generic name for monetary authority) not only aim at stabilizing inflation around the inflation target but also put some weight on stabilizing the real economy, for instance, implicitly or explicitly stabilizing a measure of resource utilization such as the output gap between actual output and ‘potential’ output. Thus, the ‘target variables’ of the central bank include not only inflation but other variables as well, such as the output gap. The objectives under flexible inflation targeting seem well approximated by a quadratic loss function consisting of the sum of the square of inflation deviations from target and a weight times the square of the output gap, and possibly also a weight times the square of instrument-rate changes (the last part corresponding to a preference for interest-rate smoothing). (The instrument rate is the short nominal interest rate that the central bank sets to implement monetary policy.) However, for new inflation-targeting regimes, where the establishment of ‘credibility’ is a priority, stabilizing the real economy probably has less weight than when credibility has been established (more on credibility below).
Because there is a lag between monetary-policy actions (such an instrument-rate change) and its impact on the central bank's target variables, monetary policy is more effective if it is guided by forecasts. The implementation of inflation targeting therefore gives a main role to forecasts of inflation and other target variables. It can be described as *forecast targeting*, that is, setting the instrument rate (more precisely, deciding on an instrument-rate path) such that the forecasts of the target variables conditional on that instrument-rate path ‘look good’, where ‘look good’, for instance, means that the inflation forecast approaches the inflation target and the output-gap forecast approaches zero at an appropriate pace.
Inflation targeting is characterized by a high degree of *transparency*. Typically, an inflation-targeting central bank publishes a regular monetary-policy report which includes the bank's forecast of inflation and other variables, a summary of its analysis behind the forecasts, and the motivation for its policy decisions. Some inflation-targeting central banks also provide some information on, or even forecasts of, their likely future policy decisions.
This high degree of transparency is exceptional in view of the history of central banking. Traditionally, central-bank objectives, deliberations, and even policy decisions have been subject to considerable secrecy. It is difficult to find any reasons for that secrecy beyond central bankers’ desire not to be subject to public scrutiny (including scrutiny and possible pressure from governments or legislative bodies). The current emphasis on transparency is based on the insight that monetary policy to a very large extent is ‘management of expectations’. Monetary policy has an impact on the economy mostly through the private-sector expectations that current monetary-policy actions and announcements give rise to. The level of the instrument rate for the next few weeks matter very little to most economic agents. What matters is the expectations of future instrument settings, which expectations affect longer interest rates that do matter for economic decisions and activity.
Furthermore, private-sector expectations of inflation for the next one or two years affect current pricing decisions and inflation for the next few quarters. Therefore, the anchoring of private-sector inflation expectations on the inflation target is a crucial precondition for the stability of actual inflation. The

proximity of private-sector inflation expectations to the inflation target is often referred to as the ‘credibility’ of the inflation-targeting regime. Inflation-targeting central banks sometimes appear to be obsessed by such credibility, there are good reasons for this obsession. If a central bank succeeds in achieving credibility, a good part of the battle to control inflation is already won. A high degree of transparency and high-quality and convincing monetary-policy reports are often considered essential to establishing and maintaining credibility. Furthermore, a high degree of credibility gives the central bank more freedom to be ‘flexible’ and also stabilize the real economy.

Whereas many central banks in the past seem to have actively avoided *accountability*, for instance by not having explicit objectives and by being very secretive, inflation targeting is normally associated with a high degree of accountability. A high degree of accountability is now considered generic to inflation targeting and an important component in strengthening the incentives faced by inflation-targeting central banks to achieve their objectives. The explicit objectives and the transparency of monetary-policy reporting contribute to increased public scrutiny of monetary policy. In several countries inflation-targeting central banks are subject to more explicit accountability. In New Zealand, the Governor of the Reserve Bank of New Zealand is subject to a Policy Target Agreement, an explicit agreement between the Governor and the government on the Governor's responsibilities. In the UK, the Chancellor of the Exchequer's remit to the Bank of England instructs the Bank to write a public letter explaining any deviation from the target larger than one percentage point and what actions the Bank is taking in response to the deviation. In several countries, central-bank officials are subject to public hearings in the Parliament where monetary policy is scrutinized; and in several countries, monetary policy is regularly or occasionally subject to extensive reviews by independent experts (for instance, New Zealand, the UK, Norway, and Sweden).

So far, since its inception in the early 1990s, inflation targeting has been a considerable *success*, as measured by the stability of inflation and the stability of the real economy. There is no evidence that inflation targeting has been detrimental to growth, productivity, employment, or other measures of economic performance. The success is both absolute and relative to alternative monetary-policy strategies, such as exchange-rate targeting or money-growth targeting. No country has so far abandoned inflation targeting after adopting it, or even expressed any regrets. For both industrial and non-industrial countries, inflation targeting has proved to be a most flexible and resilient monetary-policy regime, and has succeeded in surviving a number of large shocks and disturbances. As of 2007, a long list of non-industrial countries were asking the International Monetary Fund for assistance in introducing inflation targeting. Although inflation targeting has been an unqualified success in all the small- and medium-sized industrial countries that have introduced it, the United States, the eurozone and Japan have not yet adopted all the explicit characteristics of inflation targeting, but they are all moving in that direction.

Reservations about inflation targeting have mainly suggested that it might give too much weight on inflation stabilization to the detriment of the stability of the real economy or other possible monetary-policy objectives; the fact that real-world inflation targeting is flexible rather than strict and the empirical success of inflation targeting in the countries where it has been implemented seem to confound those reservations.

A possible alternative to inflation targeting is *money-growth targeting*, whereby the central bank has an explicit target for the growth of the money supply. Money-growth targeting has been tried in several countries but been abandoned, since practical experience has consistently shown that the relation between money growth and inflation is too unstable and unreliable for money-growth targeting to

provide successful inflation stabilization. Although Germany's Bundesbank paid lip service to money-growth targeting for many years, it often deliberately missed its money-growth target in order to achieve its inflation target, and is therefore arguably better described as an implicit inflation targeter. Many small and medium-sized countries have tried exchange-rate targeting in the form of a *fixed exchange rate*, that is, fixing the exchange rate relative to a centre country with an independent monetary policy. For several reasons, including increased international capital flows and difficulties defending misaligned fixed exchange rates against speculative attacks, fixed exchange rates have become less viable and less successful in stabilizing inflation. This has led many countries to instead pursue inflation targeting with flexible exchange rates.

A current much-debated issue concerning the further development of inflation targeting is the appropriate *assumption about the instrument-rate path* that underlies the forecasts of inflation and other target variables and the *information provided about future policy actions*. Traditionally, inflation-targeting central banks have assumed a constant interest rate underlying its inflation forecasts, with the implication that a constant-interest-rate inflation forecasts that overshoots (undershoots) the inflation target at some horizon such as two years indicates that the instrument rate needs to increased (decreased). Increasingly, central banks have become aware of a number of serious problems with the assumption of constant interest rates. These problems include that the assumption may often be unrealistic and therefore imply biased forecasts, imply either explosive or indeterminate behaviour of standard models of the transmission mechanism of monetary policy, and on closer scrutiny be shown to combine inconsistent inputs in the forecasting process (such as some inputs such as asset prices that are conditional on market expectations of future interest rates rather than constant interest rates) and therefore produce inconsistent and difficult-to-interpret forecasts. Some central banks have moved to an instrument-rate assumption equal to market expectations at some recent date of future interest rates, as they can be extracted from the yield curve. This reduces the number of problems mentioned above but does not eliminate them. For instance, the central bank may have a view about the appropriate future interest-rate path that differs from the market's view. A few central banks (notably in New Zealand, Norway, and Sweden – the last probably within the next few months) have moved to deciding on and announcing an optimal instrument-rate path; this approach solves all the above problems, is the most consistent way of implementing inflation targeting, and provides the best information for the private sector. The practice of deciding on and announcing optimal instrument-rate paths is now likely to be gradually adopted by other central banks in other countries, in spite of being considered more or less impossible, or even dangerous, only a few years ago.

Another issue is whether flexible inflation targeting should eventually be transformed into flexible *price-level targeting*. Inflation targeting as practised implies that past deviations of inflation from target are not undone. This introduces a unit root in the price level and makes the price level non-stationary. That is, the conditional variance of the future price level increases without bound with the horizon. In spite of this, inflation targeting with a low inflation rate is referred to as ‘price stability’. An alternative monetary-policy regime would be ‘price-level targeting’, where the objective is to stabilize the price level around a price-level target. That price-level target need not be constant but could follow a deterministic path corresponding to a steady inflation of two per cent, for instance. Stability of the price level around such a price-level target would imply that the price level becomes trend stationary, that is, the conditional variance of the price level becomes constant and independent of the horizon. One benefit of this compared with inflation targeting is that long-run uncertainty about the price level is smaller.

Another benefit is that, if the price level falls below a credible price-level target, inflation expectations would rise and reduce the real interest rate even if the nominal interest rate is unchanged. The reduced real interest rate would stimulate the economy and bring the price level back to the target. Thus, price-level targeting may imply some automatic stabilization. This may be highly desirable, especially in situations when the zero lower bound on nominal interest rates is binding, the nominal interest rate cannot be further reduced, and the economy is in a liquidity trap, as has been the case for several years until recently in Japan. Whether price-level targeting would have any negative effects on the real economy remains a topic for current debate and research.

See Also

- central bank independence
- inflation
- inflation dynamics
- inflation measurement

Bibliography

Roger, S. and Stone, M. 2005. On target? The international experience with achieving inflation targets. Working Paper No. 05/163. Washington, DC: International Monetary Fund.

Svensson, L.E.O. 2002. Monetary policy and real stabilization. In *Rethinking Stabilization Policy: A Symposium Sponsored by the Federal Reserve Bank of Kansas City*. Kansas City, MO: Federal Reserve Bank for Kansas City.

Svensson, L.E.O. 2007. Optimal inflation targeting: further developments of inflation targeting. In *Monetary Policy Under Inflation Targeting*, ed. F. Mishkin and K. Schmidt-Hebbel. Santiago, Chile: Banco Central de Chile.

Woodford, M. 2005. Central-bank communication and policy effectiveness. In *The Greenspan Era: Lessons for the Future – A Symposium Sponsored by the Federal Reserve Bank of Kansas City*. Kansas City, MO: Federal Reserve Bank for Kansas City.

How to cite this article

Svensson, Lars E. O. "inflation targeting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000095> doi:10.1057/9780230226203.0795

The New Palgrave Dictionary of Economics Online

inflation

Michael Parkin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article essay reviews the theoretical and empirical literature on the causes and consequences of inflation – of a continuously rising price level and falling value of money. It describes the research agendas using the analytical distinction between anticipated inflation – an idealized situation in which prices are rising at a rate at which all economic agents expect them to rise – and unanticipated inflation. The literature on the effects of inflation on economic growth and unemployment, inflation in open economies, positive theories of central bank behavior, inflation and fiscal policy, and policies towards inflation including interest rate and inflation targeting receives particular attention.

Keywords

aggregate demand shocks; budget deficits; business cycles; capital–labour ratio; cash-in-advance constraint; central bank behaviour; commodity money; Consumer Price Index; dynamic general equilibrium analysis; exchange rate determination; exchange rate regimes; fiscal theory of the price level; hyperinflation; incomplete information; inflation; inflation measurement; inflation targeting; interest rate rules; international policy coordination; labour market contracts; labour unions; monetarism; monetary and fiscal policy; monetary base; monetary policy rules; money; money supply; monopolistic competition; Mundell–Tobin effect; natural rate of unemployment; neutrality of money; new classical theory; new Keynesian macroeconomics; output gap; overlapping generations framework; overnight interest rate on inter-bank loans; Phillips curve; precommitment; prices and incomes policies; rational expectations; real business cycles; reputation; sacrifice ratios; stagflation; sticky prices; sticky wages; stylized facts; substitutes and complements; superneutrality of money; targets and instruments; technological shocks; time inconsistency; time preference; velocity of circulation

Article

‘Inflation is a process of continuously rising prices, or equivalently, of a continuously falling value of money’ (Laidler and Parkin, 1975, p. 741). Because there are several ways of measuring prices, there are also several different measures of inflation. The most commonly used measures in the modern world are the percentage rate of change in a country's Consumer Price Index or in its Gross Domestic Product deflator. Measures of inflation in earlier periods are based on fragmentary samples of prices, such as those of corn and other staple commodities, or of labour.

Inflation has been a feature of human history for as long as money has been used as a means of payment, and as Milton Friedman (1970, p. 24) famously wrote, ‘inflation is always and everywhere a monetary phenomenon, in the sense that it cannot occur without a more rapid increase in the quantity of money than in output’.

Anna J. Schwartz (1973) provides a compact account of the history of inflation from antiquity to modern times. One of the earliest documented inflations in the ancient world occurred following Alexander the Great's conquest of the Persian Kingdom (330 bc); the Roman Empire experienced rapid inflation under Diocletian at the end of the third century ad. We have no knowledge of inflation for the thousand years that followed the fall of the Roman Empire. But we do have data from the Middle Ages onwards. The inflation episodes during the Middle Ages were modest, and during those years there was a tendency for periods of rising prices to be interspersed by periods of falling prices. This pattern of intermittent inflation and deflation persisted all the way through to the Great Depression of the 1930s. Since the Great Depression, there has been a general tendency for prices to rise every year (with trivial exceptions). In the 1970s and early 1980s, serious inflations – of more than ten per cent a year – gripped most of the industrial world. But this ‘double-digit’ inflation era was short-lived, and by the mid-1980s inflation rates had returned to the more modest levels experienced in the late 1960s. In the early 2000s, there was little sign of high inflation returning in the major economies. Individual inflations of spectacular dimensions occurred in inter-war Europe, during the fall of Nationalist China (1948–9), and in modern times in some Latin American nations, Israel, and Zimbabwe. Some of these were episodes were hyperinflations – inflation rates that exceeded 50 per cent per month.

It is the fact that inflation has been so variable over time and across countries that gives rise to the question: what are the causes and the consequences of inflation? It is the enormously rich variation in inflationary experience that also provides the data which makes progress in answering those questions possible.

The literature on inflation is large, and several comprehensive, if dated, surveys of it are available (see Bronfenbrenner and Holzman, 1963; Johnson, 1963; Laidler and Parkin, 1975). No up-to-date survey of the literature on inflation was available as of 2006.

Attempts to understand inflation have been aided by the insight that anticipated inflation has different effects from unanticipated inflation. It is convenient to use that distinction in organizing this article. But it must be borne in mind that the distinction between anticipated and unanticipated inflation is analytical. It is not a distinction that has an immediate or direct correspondence with actual historical inflations.

Anticipated inflation

Anticipated inflation is an idealized situation in which prices are rising at a rate at which all economic agents expect them to rise. No one is caught by surprise. What are the effects of a fully anticipated

inflation?

There is little disagreement on the answer to this question concerning the effects on nominal variables – on such things as nominal interest rates, wages and foreign-exchange rates. Other things equal, the higher the expected rate of inflation, the higher the *level* of nominal interest rates, the higher is the rate at which wages rise, and the faster the rate of currency depreciation. Furthermore, these effects are one for one. An *x* per cent higher anticipated inflation raises nominal interest rates by *x* per cent, makes wage rates rise *x* per cent faster, and makes the currency depreciate *x* per cent faster.

There is less than complete agreement about the effects of anticipated inflation on *real* economic variables. Abstracting from transitory adjustment paths, all economic theories predict monetary neutrality: a one-shot change in the quantity of money leads to a proportionate change in the levels of all prices (and wages) and has no real effects. But not all economic theories predict monetary superneutrality – that real variables are neutral with respect to changes in the growth rate of the quantity of money.

There are three alternative views in the literature concerning money's superneutrality. One view is that money is superneutral – a change in the anticipated inflation rate has no effects on output (or economic welfare). A second view is that in increase in the anticipated inflation rate increases output (and economic welfare). Yet a third view is that a higher anticipated inflation rate lowers output (and economic welfare).

The superneutrality result has been most elegantly and clearly stated by Sidrauski (1967). The result also is present in some modern theories of money that pay detailed attention to the physical environment in which monetary exchange arises (see, for example, Townsend, 1980). The essential feature of models that generate superneutrality is that the real rate of interest is imposed by the structure of preferences (intertemporally additive with a constant rate of time preference). In equilibrium, the marginal product of capital is equal to this fixed rate of time preference so that, regardless of what happens to money, the capital stock and output rate are unaffected.

The natural rate hypothesis is a variant of the superneutrality proposition. This hypothesis, advanced by Friedman (1968) and Phelps (1968), states that money is superneutral in the particular sense that there is a unique natural unemployment rate that is independent of the anticipated rate of inflation. Any trade-off between inflation and unemployment is temporary and best thought of as a trade-off between unanticipated inflation and unemployment.

The second view that a higher anticipated rate of inflation increases output and improves economic welfare arises in two classes of models. The first is the so-called Mundell–Tobin effect (Mundell, 1963; 1965; Tobin, 1965). A higher anticipated inflation rate results in an increase in the opportunity cost of holding real money balances. According to the Mundell–Tobin view, this higher opportunity cost of holding money leads to a portfolio reallocation away from money and towards physical capital. The higher holdings of physical capital result in a higher stock of capital and therefore in a higher capital–labour ratio, which in turn leads to a higher level of output. A rise in the anticipated rate of inflation would put the economy on an adjustment path towards the new higher capital stock that would be associated with a transitory rise in the growth rate and a permanent rise in the level of output. A restatement of the Mundell–Tobin position couched in a modern rational expectations terms has been provided by Fischer (1979). The second type of model is one in which an asymmetry in price and wage adjustment – a downward rigidity – creates a long-run trade-off between inflation and the level of

economic activity – a downward-sloping long-run Phillips curve.

The third view that a higher anticipated rate of inflation lowers output and economic welfare also arises in two classes of models. First in an overlapping-generations framework (Samuelson, 1958; Wallace, 1980) a rise in the anticipated rate of inflation leads agents to economize on their holdings of money which, in turn, leads them to save less and transact on a lower scale with the succeeding generation. Second, Clower's (1967) suggested technological basis for money – the cash-in-advance constraint – generates super-non-neutrality. Using Clower's assumption, Stockman (1981) shows that, because a higher anticipated inflation rate raises the opportunity cost of holding money, this, in effect, raises the opportunity cost of undertaking all transactions and, therefore, in equilibrium lowers the scale of transactions undertaken. In Stockman's model, this results in a lower investment rate and lower capital stock. Thus a higher expected inflation rate leads to a lower level of output. A rise in the anticipated inflation rate will place the economy on an adjustment path that would result in a lower transitory growth rate and a lower permanent level of income.

Some of the above results can be thought of in terms of the substitute/complement relation between money and capital. If money and capital are substitutes in portfolios, then the Mundell–Tobin result arises. If money and capital are complements, as they implicitly are in the overlapping generations and cash-in-advance models, then higher anticipated inflation leads to lower output.

There is an abundance of empirical evidence on the alternative hypotheses about the effects of fully anticipated inflation. But the evidence is not entirely unambiguous. Because the very concept of anticipated inflation is analytical and not historical, in examining inflationary experience assumptions must be made concerning the extent to which inflations have been anticipated.

Comprehensive and systematic attempts that have addressed the question in the context of economic growth are those by Kormendi and Meguire (1985), Barro (1997), and Sala-i-Martin, Doppelhoffer and Miller (2004).

Using post-war data for 47 countries, Kormendi and Meguire analyse the effects of a change in the anticipated rate of inflation on output growth in a multivariate regression framework. Anticipated inflation was measured as simply the mean growth rate of inflation over the sample period (which went from the late 1940s to 1977). The finding of that study solidly rejects the Tobin–Mundell hypothesis and, in some formulations, fails to reject the opposite view.

Using data for about 100 countries between 1960 and 1990, Robert Barro finds that inflation has a negative effect on growth. The effect is small but significant and implies that maintained for a number of years, in inflation rate that exceeds ten per cent per year has a large cumulative effect on output. Barro is careful in his analysis of the endogeneity of inflation and growth to establish that causation runs from inflation to growth.

Barro's finding is challenged by Sala-i-Martin, Doppelhofer and Miller. Using data from 1960 to 1996 for 88 countries and 67 variables considered candidates for influencing the rate of economic growth, and using a Bayesian averaging of classical estimates approach, they find that neither average inflation rate nor the square of the inflation rate has a significant effect on the growth rate.

The work just summarized takes a reduced form and linear approach, and these features limit its utility. Future work on the effects of inflation on growth should be directed toward looking at structural accounts of the linkages and seeking highly nonlinear and perhaps nonparametric relationships between these two variables.

Investigations of the neutrality of unemployment (and output) with respect to anticipated inflation has been the subject of innumerable studies, and Laidler and Parkin (1975) review the state of this literature up to the mid-1970s. The conclusions that emerged from this work were mixed, and most of the results generated on data-sets that ended around 1970 showed the existence of a trade-off. But as the data for the 1970s (with its high inflation rate) were added, the picture changed and Laidler and Parkin concluded that it was not possible to reject the view that the unemployment rate is neutral with respect to anticipated inflation.

This conclusion is challenged in three different ways. First, the classic Sargent (1976) shows that reduced-form equations estimated for a given sampling interval over a given sampling period cannot distinguish among alternative theories, even though the theories have radically different policy implications. The implication of this result for Phillips curve trade-offs is that useful inferences can be made but only by estimating reduced forms over different sub-periods or countries across which policy rules differed systematically. As of 1976, Sargent thought that not much of this type of work had been done, so that little was known.

Second, further empirical work seemed to be consistent with the view that a permanent trade-off exists. King and Watson (1994) study the US Phillips correlations and Phillips trade-offs in a bivariate time-series analysis. They use the unit root (I(1)) inflation process to get around the Sargent (1976) problem (see Fisher and Seater, 1993, and King and Watson, 1997, for details), and estimate structural models to interpret the data and compute the long-run trade-offs and sacrifice ratios (cost of lowering inflation) associated with each model. Except for the extreme case of a real business cycle model, they find long-run trade-offs between inflation and unemployment.

The same conclusion is reached by Akerlof, Dickens and Perry (1996), but for a different reason. They report evidence of permanent downward wage stickiness, which implies a long-run trade-off. This evidence comes from four sources: ethnographic surveys, Bureau of Labor Studies data on the distribution of wage changes in manufacturing establishments, union settlements (in both the United States and Canada), and the authors' own survey of individuals in the Washington DC area. The authors' were aware that Panel Study of Income Dynamics (PSID) data showed evidence of extensive downward wage flexibility, but argue that individual reporting errors are large, and when corrected for using data from the Current Population Survey, downward rigidity is present. The presence of downward wage rigidity would constitute a serious challenge to the natural rate hypothesis – the neutrality of the unemployment rate with respect to the anticipated inflation rate. And not surprisingly, much work has been done to check the conclusion reached by Akerlof, Dickens and Perry. Parkin (2000) summarizes this work, which concludes that the money wage rate is not downwardly rigid and that the appearance of downward rigidity results from three sources of bias; measurement error, rounding error and long-term contracts. Controlling and correcting for these sources of bias points towards wage flexibility. Clearly more work is needed to settle this issue.

The third challenge to monetary neutrality comes from a series of papers by Barro (1977; 1978) and Mishkin (1982a; 1982b). Decomposing money growth into anticipated and unanticipated components, Barro reports that only unanticipated money growth influences unemployment and real GDP and (as predicted) both anticipated and unanticipated money growth influences the price level. Mishkin shows that Barro's estimation procedure, while providing consistent parameter estimates, delivers incorrect standard errors. When Mishkin replicates Barro's exercises with valid tests, he rejects the restrictions

implied by neutrality. (He does not reject the restrictions implied by rationality.)
The literature just reviewed deals with the consequences of anticipated inflation and not its causes. Questions concerning causality are more naturally addressed in the context of an investigation of unanticipated inflation.

Unanticipated inflation

It is not possible to analyse unanticipated inflation in isolation, independently of other aspects of aggregate economic performance. Fluctuations (at the business cycle frequency) in the general level of economic activity and in inflation, though far from perfectly correlated, share some common features. There is, for example, a general positive correlation between inflation and real income (or equivalently, a negative correlation between inflation and unemployment). There is also a positive correlation between money and income as well as between the velocity of circulation of money and income.
The ‘stylized facts’ about the business cycle (shared by all economies) raise difficult questions about cause and effect. Of the four variables – the price level, real output, the money supply and the velocity of circulation – which, if any, is the prime mover? Do fluctuations in the growth rate of the money supply cause fluctuations in the other variables? Do autonomous movements in the price level, perhaps stemming from wage-push pressure, initiate the fluctuations in money, velocity and output? Does the business cycle have its origin in real factors that initiate fluctuations in output, which in turn lead to induced fluctuations in money supply growth, inflation and velocity?
At one level questions such as these are statistical and are capable of being investigated using econometric methods that detect causality, such as those proposed by Granger (1969). Studies based on such methods have not, however, delivered decisive results.
Most investigations of the possible causes of inflation have sought to understand the phenomenon by identifying the sources of inflation and studying the transmission mechanism whereby those sources are translated into variations in the rate of inflation and in other economic aggregates. This approach is one which seeks to understand both inflation and the business cycle as an integrated phenomenon.
There are three broad classes of theories that have been proposed for understanding the unanticipated and cyclical aspects of inflation. The first of these stems from the work of Keynes (1936) and emphasizes both price stickiness and the potential for autonomous movements in prices. On this view, the normal state of affairs would be one in which wages and prices are relatively sticky, responding only gradually to aggregate demand shocks. Shocks to aggregate demand arise from a variety of sources. One possibility is that autonomous fluctuations in investment produce fluctuations in aggregate demand. Other possible sources of aggregate demand fluctuations are fluctuations in wealth and interest rates which in turn are induced by fluctuations in the growth rate of the money supply. Fluctuations in wealth and interest rates can induce fluctuations in investment and consumption. All of these potential sources of variation in aggregate demand lead to cycles in both output and the price level. Initially, a change in demand will have bigger output effects than price-level effects, but eventually prices and wages will adjust to reflect fully the change in aggregate demand. The resulting co-movements in output and prices will be positively, though not strongly, correlated.
From time to time this normal state of affairs is disturbed by autonomous price shocks. The most commonly hypothesized source of price shocks is wage-push. It is suggested that, at times of substantial industrial or social unrest, movements in the level of money wages will act as a type of social safety

mechanism. The idea that wage-push results from sociological phenomena was particularly popular amongst economists in the UK in the early 1970s (see, in particular, Balogh, 1970; Jones, 1972; Wiles, 1973; Hicks, 1974. By the time the first oil shock occurred (late 1973), ‘wage-push’ gave way to ‘oil-push’ as the most commonly identified source of autonomous movement in inflation.

When autonomous movements in the price level occur, the phenomenon that came to be known as ‘stagflation’ quickly follows. The autonomous price rise raises the inflation rate and lowers output (raising unemployment). If the higher unemployment and lower output induces an increase in the growth rate of the money supply, then even further price-level rises occur.

This traditional version of the Keynesian theory of inflation and the business cycle, together with some of the sociological embellishments that have been briefly reviewed above, is very thoroughly explained and elaborated in Laidler and Parkin (1975).

More recent and sophisticated versions of the Keynesian theory of cycles and inflation may be found in papers by Fischer (1977), Phelps and Taylor (1977) and Taylor (1979; 1980). The essence of these ‘New Keynesian’ theories is the existence of long-term contractual arrangements in labour markets. Such arrangements result in wages, the major element of costs, being predetermined. This stickiness of wages and costs results in a stickiness of prices, even if the expectations of prices that form the basis for the long-term labour market contracts are formed rationally.

A second approach to understanding cyclical fluctuations is one based on incomplete contemporaneous information about aggregate demand. This approach, sometimes called the ‘New Classical Theory’, was first suggested in the early 1970s by Lucas (1972; 1973). The approach is broadly consistent with the Keynesian mechanism of aggregate demand determination but proposes an alternative theory of aggregate supply. Individual economic agents are assumed to operate in informationally isolated ‘islands’ and to be incapable of distinguishing relative from absolute price level changes. The resulting confusion causes them to respond to absolute price changes as if they were relative price changes. This response results in positive co-movements in output and the price level.

In both the Keynesian and New Classical approaches, the key driving variable generating the cycle – fluctuations in both real output and the inflation rate – is a fluctuating growth rate in the money supply. This is not to deny that other things might, from time to time, shock the economy. Rather, it is a proposition about the major ongoing source of cyclical variation. Within both of the theories, positive co-movements of velocity are explained by appealing to the idea that to some degree the cycle itself is forecastable. To the extent that it is, higher rates of inflation at the cyclical peak will in part be anticipated and, therefore, reacted to. It is always efficient to reduce money holdings when the opportunity cost of holding money increases. Higher expected inflation rates, leading to higher nominal interest rates, induce such economizing and are, therefore, the major source of procyclical fluctuations in velocity.

A third approach to understanding aggregate fluctuations denies the primacy of variations in the money supply growth rate, or in any other sources of aggregate demand fluctuation in generating the cycle. This approach, known as ‘real business cycle theory’, has yet to gain a major following but has, in recent years, begun to spawn a growing and important literature (see, in particular, King and Plosser, 1984; Kydland and Prescott, 1982; Long and Plosser, 1983; Nelson and Plosser, 1982). Though differing in details, the essential proposition of the new real business cycle theories is that aggregate fluctuations emanate from technological shocks to the aggregate production function or, in some versions, from

sector-specific shocks and from the interactions between sectors of the economy – although a large literature has now incorporated Calvo (1983) price stickiness or monopolistic competition.

Technological shocks that generate fluctuations in full-employment output would, other things equal, generate negative co-movements in prices, and, presumably, to the extent that such movements were forecastable, countercyclical movements in velocity. Since such co-movements do not occur, it seems as if the real cycle theories are in substantial trouble. King and Plosser (1984) address this problem directly by proposing that technological shocks which affect real output induce responses in money and credit that accommodate – indeed over-accommodate – the real fluctuations. Thus, when there is a positive shock to aggregate supply, this induces an even bigger rise in the total volume of money and credit and, therefore, induces procyclical co-movements in money, prices and output. To the extent that these are forecastable, economizing on real balances generates procyclical velocity.

There is not, at the present time, any definitive and systematic evidence capable of disposing convincingly of any of these three alternative approaches; nor is there any overwhelming evidence suggesting that any of them is clearly in the lead.

Inflation in open economies

The alternative approaches to understanding inflation that have been reviewed so far have (implicitly) examined inflation in a closed economy. Most practical concerns about inflation arise in individual countries which are open economies. The international trade and international capital market transactions undertaken by such countries have an important bearing on their inflation performance. Also, the foreign-exchange rate regime – fixed or flexible – has an important influence upon a country's inflation performance. It was during the period of rapidly accelerating inflation in the 1970s that open economy theories and the international transmission mechanism gained in prominence (see Parkin and Zis, 1976a; 1976b).

The main feature of the analysis of inflation in an open economy is the emphasis on the limited potency of domestic monetary policy under fixed exchange rates. In a country, or more interestingly in a world, operating on fixed exchange rates, individual countries’ monetary policies have no effect on the country's rate of inflation. Instead, monetary policy influences the country's balance of payments. In such a world, inflation is a world phenomenon, not a national phenomenon. It is the growth rate of the world money supply that determines the world average rate of inflation. Theorizing along this line had, in fact, made good progress even as early as the middle of the 18th century at the hands of David Hume (1752). It was rediscovered and popularized in the 1960s and early 1970s by Mundell (1971) and Johnson (1973).

The rediscovery of David Hume's analysis provided interesting insights into the resurgence of world inflation at the end of the 1960s. An attempt on the part of the United States to finance its Great Society programme and the Vietnam War with limited tax increases and with an increase in the growth rate of the money supply – with an increase in the inflation tax – became the engine of an inflation that engulfed the entire fixed exchange-rate world.

Understanding the international generation and transmission of inflation in a flexible exchange rate world, such as that which had emerged by the mid-1970s, is still far from settled. At the centre of the problem of understanding inflation is the problem of understanding the determination of foreign

exchange rates. Large and rapid movements in foreign-exchange rates are seen as having a potentially powerful and rapid effect on domestic price levels. The forces that determine exchange rates are still, however, far from well understood. Viewing the foreign exchange rate as following a random walk is as precise as any structural theories of the exchange rate that have so far been proposed and tested.
Despite the absence of a convincing theory of inflation in an open economy, the effects of policy coordination (or its absence) have been studied. A central question addressed by Oudiz and Sachs (1984) and Obstfeld and Rogoff (2002) is whether unilateral national monetary policy rules are inferior to international monetary coordination. The answer is that they are not.

Positive theories of central bank behaviour

Recent developments in understanding inflation have been dominated by the rational expectations revolution and the related and more far-reaching revolution that has uses rigorous dynamic general equilibrium analysis. Some of the implications of that revolution have been discussed above and have been to strengthen and refine the theories of inflation that emphasize fluctuations in the growth rate of the money supply as the principal source of fluctuations in inflation and other economic aggregates.
The rational expectations hypothesis holds that expectations are formed by making predictions of future inflation on the basis of the mechanisms that generate actual inflation. If inflation is indeed caused by rapid monetary expansion, then forecasting future inflation is the same thing as forecasting future monetary policy. But monetary policy itself emerges from an ill-understood political process. In most countries the task of formulating monetary policy has been delegated to a central bank. Yet, in determining monetary policy, central banks are often influenced by the economic and political environment in which they operate and must also take account of the consequences of their actions for the behaviour of the economy as a whole.
In order to understand the inflationary process, with people forming expectations rationally, it becomes necessary to understand the policymaking mechanisms and the forces that generate varying monetary growth rates. The first serious analysis of this problem was that by Kydland and Prescott (1977) and the problem has been investigated more recently by Barro and Gordon (1983a; 1983b) and Cukierman (1992). In the models proposed by these writers, a central bank's goal is to achieve an optimal combination of inflation and unemployment. Lower inflation and lower unemployment are seen by the central bank as desirable objectives. The bank is constrained, however, by a short-run trade-off between inflation and unemployment – a trade-off arising from the considerations described above. A surprise rise in inflation would produce a cut in unemployment while a surprise drop in inflation would produce a rise in unemployment. The precise way in which the short-run trade-off between inflation and unemployment constrains the central bank depends on the expectations of private agents concerning the bank's behaviour. A central bank that can credibly precommit to a particular rule about inflation – perhaps a zero-inflation rule – would be a bank that could engender rational expectations of zero inflation. It would be optimal for such a bank to in fact precommit to a zero rate of inflation and then deliver that rate.
The ability to precommit and with credibility seems to require some mechanism for binding the central bank that does not have a readily identifiable counterpart in the real world. Central banks are, in fact, free to pursue whatever policies they wish at their discretion. Since this fact is known to all private

economic agents, it will be rational for them to take it into account when forming expectations about central bank behaviour. The equilibrium that results in this case will be such as to ensure that the actual inflation rate chosen by the bank is one that removes any temptation for the bank to depart from that rate and further exploit the short-run trade-off. Put differently, the inflation rate chosen will be the best available at the natural rate of unemployment. Only in such a situation would the central bank have no further temptation to attempt to exploit the short-run trade-off. Thus without the ability to precommit to a fixed (and presumably zero) rate of inflation, a central bank will end up delivering a higher rate of inflation than that which is socially desirable.

One feature of the positive theories of inflation developed by Kydland–Prescott and Barro–Gordon that some people find disquieting is the time inconsistency. (In game theory language, the equilibrium concept is Nash rather than sub-game perfection.) Attempts to develop positive analyses that do not have this feature have been based on reputation. One such approach, in Barro and Gordon (1983a), uses the so-called ‘trigger strategy’ model of reputation suggested by James Friedman (1971). A model is proposed in which the central bank would be punished if it delivered too high a rate of inflation and in which it takes time to restore the bank's reputation. In equilibrium, the bank never does inflate at a rate that requires the punishment to be inflicted.

An alternative approach by Barro (1986) uses the reputation analysis developed by Kreps and Wilson (1982). In this model there are two potential ‘types’ of central banker, one that likes inflation and one that dislikes it. The inflationary central banker has an incentive to masquerade as a non-inflationary type in order to induce low inflation expectations. By inducing low inflation expectations, the inflationary central bank will, at some point, be able to exploit those low expectations and produce a surprise inflation; it will do this by following initially a strategy of inflating at exactly the same rate as would be chosen by a non-inflationary central bank. At some later point it will pursue a mixed strategy – a strategy analogous to choosing an inflation rate by drawing numbers from an urn. Once this mixed strategy has resulted in a high rate of inflation, the inflationary central banker is revealed, and expectations about inflation as well as actual inflation will rise.

Another feature of the Kydland–Prescott and Barro–Gordon models that is objectionable is that the central bank targets an unemployment rate below the natural rate. If it were to target the natural rate, there is no tension between its inflation and real goals. Cukierman overcomes this objection by replacing the symmetric loss function of the standard model with an asymmetric loss function: the central bank weighs positive deviations from the natural unemployment rate more heavily than deviations below the natural rate.

Backus and Driffill (1985) and Cukierman (see Cukierman, 1992, ch. 3), have suggested another modification to the standard model: the possible interactions between labour unions (working as a unified wage-setting institution) and the central banks. In this case, inflation (and money supply growth) is determined as the outcome of a game between the central bank and the economy-wide labour union.

Empirical tests of the alternative positive theories of central bank behaviour have been conducted by Ruge-Murcia (2003) and by Cukierman and Gerlach (2003). Ruge-Murcia uses US time-series data and rejects the Barro–Gordon formulation but does not reject the Cukierman asymmetric loss function formulation. Cukierman and Gerlach use data for 22 OECD countries and reach a similar conclusion.

Other recent developments in understanding central bank behaviour arise from the normative analysis of monetary policy to achieve an inflation target, and it is convenient to discuss this topic in the context of

inflation policy below. But, before that, it is convenient to consider the links between monetary policy and fiscal policy.

Inflation and fiscal policy

A further consequence of the rational expectations and dynamic general equilibrium revolutions has been to force attention back to the connection between fiscal and monetary policy. The simple accounting fact that government expenditure must be financed, either by taxation, by borrowing or by money creation, implies that any analysis of the determination of money growth must at the same time make consistent propositions about fiscal policy and deficit financing. Of course, variations in the growth rate of interest-bearing debt can provide a good deal of insulation of money growth from the deficit. Nevertheless, large and persistent deficits may give rise to rational expectations of future money growth, even in the face of currently firm monetary policies. Sargent and Wallace (1981) have shown that, if the fiscal authority is the prime mover and follows taxation and spending policies that are independent of monetary policy, then, essentially, inflation and, ultimately, money growth are fiscal phenomena. Whether these findings are of practical importance is a matter of some controversy. Sargent (1982), studying the ends of four big inflations, has argued that adjustments in fiscal policy have been crucial to ending inflation. By implication, the emergence of a large and apparently uncontrolled deficit would be seen as the origin of serious inflation. Work by Dornbusch and Fisher (1986) offers a different interpretation, however, placing major importance on the behaviour of the foreign exchange rate.
The link between fiscal policy and inflation is most complete in Woodford's (1995) fiscal theory of the price level. Because the quantity of money demanded depends on the opportunity cost of holding money, which in turn depends on the rational expectation of the inflation rate, there is a large number (infinite) of equilibrium price level paths. The standard (mostly unstated) approach rules out all the purely speculative equilibria and selects the unique equilibrium based on the monetary fundamentals. In which the government's choice of how to finance its debt determines the inflation rate. The fiscal theory of the price level rejects this approach and rules out equilibria by the government's selection of its debt financing regime. As an example, Kocherlakota and Phelan (1999) show that, with a policy of constant taxes and constant money, the fiscal theory predicts that a one-time cut in the quantity of money generates a speculative hyperinflation (in contrast to the standard model prediction of a one-time fall in the price level).

Policy towards inflation

Analyses of policies towards inflation have changed over the years. Advocacy of gradually slowing down the growth rate of the money supply and advocacy of controls on wages and prices were the most commonly heard policy suggestions for controlling inflation in the 1960s and early 1970s. Those who saw autonomous wage and price movements as the principal source of inflation saw prices and incomes policies as the major weapon to control it. Those who saw money growth as the source of inflation embraced monetary gradualism as the most obvious cure. A prodigious amount of work attempting to evaluate alternative policies was undertaken, much of which is surveyed by Laidler and Parkin (1975). As a consequence of the rational expectations and dynamic general equilibrium revolutions, the focus of

the policy debate has shifted markedly from that of seeking to manipulate variables such as key wage settlements (the prices-and-incomes policy solution) or the growth rate of the quantity of money (the monetarist solution). Instead, attention has turned to thinking about the way in which different institutional arrangements interact to produce different inflation rates. And the emphasis has shifted from policy as an action to policy as a process or set of rules.

One line of research has examined the consequences of alternative monetary systems, including the adoption of alternative forms of commodity money (see, in particular, ‘Conference on Alternative Monetary Standards’, 1983). Another research direction has been the investigation of targeting nominal income growth as a means of conquering and avoiding inflation (Tobin, 1983; Taylor, 1985).

But the idea that has attracted most attention both in the research community and among central banks is the use of a monetary policy rule that seeks to achieve either an inflation rate target or a price level target. The study of inflation or price level targeting has both a positive and a normative dimension and sometimes the two are not explicitly distinguished.

Svensson (1999) has provided a nice distinction between what he calls ‘instrument rules’ and ‘targeting rules’ for monetary policy. In the context of inflation targeting (and that is the context of most of the recent literature on monetary policy) an instrument rule specifies how the policy instrument responds to the current state of the economy. The current state can include current forecasts of future variables. A targeting rule, in contrast, states that the policy instrument shall be set at the level that makes for forecast inflation rate equal the inflation target.

The policy instrument that features in instrument rules is either the overnight interest rate on inter-bank loans or the monetary base. Woodford (2003) provides the authoritative account and discussion of the interest rate instrument rule and shows that such a rule can, in principle, deliver low and stable inflation provided that it incorporates the ‘Taylor principle’, which states that the interest rate must change in the same direction as a change in the inflation rate but by more than the change in the inflation rate (Taylor, 1993; 1999).

McCallum (1988) has explored the use of a monetary base rule and compared the robustness of interest rate and monetary base rules.

It is a curious fact about the models that explore the use of an interest rate rule that money plays either no role or no essential role in the inflation process. The models in which money plays no role are typically specified as reduced forms in which inflation is generated by expected inflation and the output gap; the output gap responds to the real interest rate, which equals the nominal interest rate set by the central bank minus the inflation rate; and expectations are rational. Other models are specified at a deeper structural level with consumers maximizing intertemporal utility of consumption and leisure and monopolistically competitive firms setting prices according to a Calvo (1983) formula.

In some models, money enters through a ‘shopping time’ function (King and Wolman, 1996). But whether present in the model or not, money plays no essential role in the inflation process. This fact is emphasized in Woodford (2003 by his exploration of the cashless economy and is seen as a virtue because it might provide insights on inflation in a future economy when technological change has driven money, as we know it, out of existence.

It is also a curious fact that inflation targeting amounts to targeting a variable whose value cannot be influenced by a central bank's current actions until well beyond the bank's forecast horizon. It is the long and variable lags in the response of inflation (and output) to monetary policy that led Friedman to his

original advocacy of a money stock growth rate target.
The evolution of inflation over the coming years will provide valuable evidence on both the inflation process, the currently out-of-favour monetarist ideas, and the wisdom of the current policy regimes.

Conclusion

Macroeconomics in general, and the theory of inflation in particular, is in a fluid state. The foregoing has attempted to review that state and provide a picture of the path that we have taken in getting to it. We have broad agreement on the facts to be explained and broad agreement on the behaviour of nominal variables (for given real variables) in an inflationary economy in which the path of inflation is anticipated. We also have broad agreement that fully anticipated inflations, though in many theoretical models capable of generating non-neutralities, are nevertheless to a good approximation neutral. Beyond that there is little in the way of firm knowledge. We have a variety of models of macroeconomics and inflation, and many clear theoretical results. We do not have much, however, in the way of solidly based rejections of any of the available models. Uncertainty surrounds both the issue of the impulse (or impulses) that generate inflation and other fluctuations and the issue of the propagation mechanisms that translate those impulses into movements in output and the price level.

See Also

- central bank independence
- cost-push inflation
- demand-pull inflation
- hyperinflation
- inflation dynamics
- inflation expectations
- inflation measurement
- inflation targeting
- neutrality of money

Bibliography

Akerlof, G.A., Dickens, W.T. and Perry, G.L. 1996. The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1996(1), 1–76.

Backus, D. and Driffill, J. 1985. Rational expectations and policy credibility following a change in regime. *Review of Economic Studies* 3, 211–21.

Balogh, T. 1970. *Labour and Inflation*. London: Fabian Society.

Barro, R.J. 1977. Unanticipated money growth and unemployment in the United States. *American*

Economic Review 67, 101–15.

Barro, R.J. 1978. Unanticipated money, output, and the price level in the United States. *Journal of Political Economy* 86, 549–80.

Barro, R.J. 1986. Reputation in a model of monetary policy with incomplete information. *Journal of Monetary Economics* 17, 3–20.

Barro, R.J. 1997. *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge, MA: MIT Press.

Barro, R.J. and Gordon, D.B. 1983a. Rules, discretion and reputation: a model of monetary policy. *Journal of Monetary Economics* 12, 101–21.

Barro, R.J. and Gordon, D.B. 1983b. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91, 589–610.

Bernanke, B.S. and Woodford, M. 2004. *The Inflation-Targeting Debate*. Chicago: University of Chicago Press.

Bronfenbrenner, M. and Holzman, F.D. 1963. A survey of inflation theory. *American Economic Review* 53, 593–661.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6, 1–8.

Conference on alternative monetary standards. 1983. *Journal of Monetary Economics* 12(1).

Cukierman, A. 1992. *Central Bank Strategy, Credibility, and Independence*. Cambridge, MA: MIT Press.

Cukierman, A. and Gerlach, S. 2003. The inflation bias revisited: theory and some international evidence. *The Manchester School* 71, 541–65.

Dornbusch, R. and Fischer, S. 1986. Stopping hyperinflations past and present. *Weltwirtschaftliches Archiv* 122, 1–47.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85, 191–205.

Fischer, S. 1979. Anticipations and the non-neutrality of money. *Journal of Political Economy* 87, 228–52.

Fisher, M. and Seater, J. 1993. Long run neutrality and superneutrality in an ARIMA framework. *American Economic Review* 83, 402–15.

Friedman, J.W. 1971. A non-cooperative equilibrium for supergames. *Review of Economic Studies* 38, 1–12.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Friedman, M. 1970. *The Counter-Revolution in Monetary Theory*. London: Institute of Economic Affairs.

Granger, C.W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.

Hicks, J.R. 1974. *The Crisis in Keynesian Economics*. Oxford: Blackwell.

Hume, D. 1752. Of money; Of interest; Of the balance of trade. *Political Discourses*; reprinted in *Three Essays: Moral, Political and Literary*. London: Oxford University Press, 1963.

Johnson, H.G. 1963. A survey of theories of inflation. *Indian Economic Review* 6(4), 29–69.

Johnson, H.G. 1973. Secular inflation and the international monetary system. *Journal of Money, Credit, and Banking* 5(1, Part II), 509–20.

Jones, A. 1972. *The New Inflation: The Politics of Prices and Incomes*. London: Penguin Books and André Deutsch.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

King, R.G. and Plosser, C.I. 1984. Money, credit, and prices in a real business cycle model. *American Economic Review* 74, 363–80.

King, R.G. and Wolman, A.L. 1996. Inflation targeting in a St. Louis model of the twenty-first century. *Federal Reserve Bank of St. Louis Review* 78(3), 83–107.

King, R.J. and Watson, M.W. 1994. The post-war U.S. Phillips curve: a revisionist econometric history. *Carnegie-Rochester Conference Series on Public Policy* 41, 157–219.

King, R.J. and Watson, M.W. 1997. Testing long-run neutrality. *Federal Reserve Bank of Richmond Quarterly Review* 83, 69–101.

Kocherlakota, N. and Phelan, C. 1999. Explaining the fiscal theory of the price level. *Federal Reserve Bank of Minneapolis Quarterly Review* 23(4), 14–23.

Kormendi, R.C. and Meguire, P.G. 1985. Macroeconomic determinants of growth: cross-country evidence. *Journal of Monetary Economics* 16, 141–63.

Kreps, D. and Wilson, R. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27, 253–79.

Kydland, F.E. and Prescott, E.C. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Laidler, D. and Parkin, M. 1975. Inflation: a survey. *Economic Journal* 85, 741–809.

Long, J.B. and Plosser, C.I. 1983. Real business cycles. *Journal of Political Economy* 91, 39–69.

Lucas, R.E., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R.E., Jr. 1973. Some international evidence on output-inflation tradeoffs. *American Economic Review* 63, 326–34.

McCallum, B.T. 1988. Robustness properties of a rule for monetary policy. *Carnegie-Rochester Conference Series on Public Policy* 29, 173–203.

Mishkin, F.S. 1982a. Does anticipated monetary policy matter? An econometric investigation. *Journal of Political Economy* 90, 22–51.

Mishkin, F.S. 1982b. Does anticipated aggregate demand policy matter? Further econometric results. *American Economic Review* 72, 788–802.

Mundell, R.A. 1963. Inflation and real interest. *Journal of Political Economy* 71, 280–83.

Mundell, R.A. 1965. Growth, stability and inflationary finance. *Journal of Political Economy* 73, 97–109.

Mundell, R.A. 1971. *Monetary Theory: Inflation, Interest and Growth in the World Economy*. Pacific Palisades, CA: Goodyear Publishing Co.

Nelson, C.R. and Plosser, C.I. 1982. Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10, 139–62.

Obstfeld, M. and Rogoff, K. 2002. Global implications of self-oriented national monetary rules. *Quarterly Journal of Economics* 117, 503–35.

Oudiz, G. and Sachs, J. 1984. Macroeconomic policy coordination among the industrial economies. *Brookings Papers on Economic Activity* 1984(1), 1–76.

Parkin, M. 2000. What have we learned about price stability? *Price Stability and the Long-Run Target for Monetary Policy*. Ottawa, Bank of Canada.

Parkin, M. and Zis, G. 1976a. *Inflation in Open Economies*. Manchester: Manchester University Press.

Parkin, M. and Zis, G. 1976b. *Inflation in the World Economy*. Manchester: Manchester University Press.

Phelps, E.S. 1968. Money, wage dynamics, and labor market equilibrium. *Journal of Political Economy* 76, 678–711.

Phelps, E.S. and Taylor, J.B. 1977. Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85, 163–90.

Ruge-Murcia, F.J. 2003. Does the Barro–Gordon model explain the behavior of US inflation? A reexamination of the empirical evidence. *Journal of Monetary Economics* 50, 1375–90.

Sala-i-Martin, X., Doppelhoffer, G. and Miller, R.I. 2004. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94, 813–35.

Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66, 467–82.

Sargent, T.J. 1976. The observational equivalence of natural and unnatural rate theories of macroeconomics. *Journal of Political Economy* 84, 631–40.

Sargent, T.J. 1982. The ends of four big inflations. In *Inflation: Causes and Effects*, ed. R.E. Hall. Chicago: University of Chicago Press.

Sargent, T.J. and Wallace, N. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis, Quarterly Review* 5(3), 1–17.

Schwartz, A.J. 1973. Secular price change in historical perspective. *Journal of Money, Credit, and Banking* 5(1, Part II), 243–69.

Sidrauski, M. 1967. Inflation and economic growth. *Journal of Political Economy* 75, 796–810.

Stockman, A.C. 1981. Anticipated inflation and the capital stock in a cash-in-advance economy. *Journal of Monetary Economics* 8, 387–93.

Svensson, L. 1999. Inflation targeting as a monetary policy rule. *Journal of Monetary Economics* 43, 607–54.

Taylor, J.B. 1979. Staggered wage setting in a macro model. *American Economic Review, Papers and Proceedings* 69(2), 108–13.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

Taylor, J.B. 1985. What would nominal GNP targeting do to the business cycle? *Carnegie-Rochester Conference Series on Public Policy* 22, 61–84.

Taylor, J.B. 1993. Discretion versus policy rules in practice, *Carnegie-Rochester Conference Series on Public Policy* 39, 195–214.

Taylor, J.B., ed. 1999. *Monetary Policy Rules*. Chicago: University of Chicago Press.

Tobin, J. 1965. Money and economic growth. *Econometrica* 33, 671–84.

Tobin, J. 1983. Monetary policy: rules, targets and shocks. *Journal of Money, Credit, and Banking* 15, 506–18.

Townsend, R. 1980. Models of money with spatially separated agents. In *Models of Monetary Economies*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Wallace, N. 1980. The overlapping generations model of fiat money. In *Models of Monetary Economies*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Wiles, P. 1973. Cost inflation and the state of economic theory. *Economic Journal* 83, 377–98.

Woodford, M. 1995. Price level determinacy without control of a monetary aggregate. *Carnegie-*

Rochester Conference Series on Public Policy 43, 1–46.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Parkin, Michael. "inflation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000088> doi:10.1057/9780230226203.0791

The New Palgrave Dictionary of Economics Online

informal economy

Keith Hart
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The informal economy or sector has become the preferred term for unregulated economic activities, in both rich and poor countries. Based on Weber's theory of rationalization, it was coined during the early 1970s in response to proliferating self-employment and casual labour in Third World cities. Now its range of reference is very wide, embracing everything from high-level political corruption to home improvement. The phenomenon is real enough and of some antiquity, but its definition remains elusive. Operating beyond the rules of bureaucracy, the informal economy may be understood dialectically as division, content, negation or residue.

Keywords

bureaucracy; capital accumulation; economic development in the long run; exploitation; informal economy; labour surplus economies; poverty; protection; rational enterprise; rule of law; rural–urban migration; self-employment; Weber, M.

Article

The term ‘informal economy’ became current in the 1970s as a label for economic activities that take place outside the framework of bureaucratic public and private sector establishments. It arose in response to the proliferation of self-employment and casual labour in Third World cities; but later the expression came to be used with reference to societies like Britain, where it competed with epithets of deindustrialization – the ‘hidden’, ‘underground’, ‘black’ economy, and so on.
The social phenomenon is real enough and of some antiquity. London's East End in the mid-19th century is a stark example of informal economic organization which rivals in scale any of today's tropical slum areas (Davis, 2006). Nevertheless, the empirical referents of the ‘informal economy’ remain elusive, ranging as they do between the extremes of corrupt public finance in Congo and do-it-yourself in a London suburb. The intellectual history of the concept is clearer. It was provoked by the

failure of prevalent economic models to address a large part of the world that they claimed to offer prescriptions for. Sociologists, anthropologists, geographers and historians have grasped the opportunity to embarrass economists by pointing out this deficiency. More remarkably, many economists, including employees of bureaucracies such as the World Bank and the International Labour Organization (ILO), have identified the ‘informal sector’ as something they must deal with. Whereas once the effects of ‘informality’ were thought to be palliative, they are now often seen as a threat to legitimate businesses.

Some notable attempts have been made to document the economy of the streets. Henry Mayhew's investigations for the *Morning Chronicle* in the 1850s, published as *London Labour and the London Poor* (1861–2), are a classic source, as are Oscar Lewis's several accounts of the ‘culture of poverty’ (for example, *La Vida*, 1964). Very little of all this impinged on the world of development economists. The dualistic models of economic development that prevailed in the 1960s took their lead from W. Arthur Lewis's (1954) theory of development with unlimited supplies of labour, whereby underemployed rural workers migrated to find wage employment in a higher productivity urban economy.

In *Peddlers and Princes* (1963), Clifford Geertz identified two economic ideal types in a Javanese town. The majority were occupied in a street economy that he labelled ‘bazaar-type’. Opposed to this was the ‘firm-type’ economy consisting largely of Western corporations that benefited from the protection of state law. These had *form* in Weber's (1981) sense of ‘rational enterprise’ based on calculation and the avoidance of risk. National bureaucracy lent these firms a measure of protection from competition, thereby allowing the systematic accumulation of capital. The ‘bazaar’, on the other hand, was individualistic and competitive, so that accumulation was well-nigh impossible. Geertz considered what it would take for a group of reform Muslim entrepreneurs to join the modern ‘firm’ economy. They were rational and calculating enough; but they were denied the institutional protection of state bureaucracy, which was the preserve of the existing corporations.

A decade later and in the context of growing unease over Third World urban unemployment, Keith Hart (1973, based on a conference paper of 1971) argued that the masses who were surplus to the requirements for wage labour in African cities were not ‘unemployed’ but rather were positively employed, even if often for erratic and low returns. He proposed that these activities be contrasted with the ‘formal’ economy of government and organized capitalism as ‘informal income opportunities’. Moreover, he suggested that the aggregate inter-sectoral relationship between the two sources of employment might be of some significance for models of economic development in the long run. In particular, the informal economy might be a passive adjunct of growth originating elsewhere or its dynamism might be a crucial ingredient of economic transformation in some cases.

The dualism (formal–informal) and some of the thinking behind it received immediate publicity through its adoption in an influential ILO (1972) report on incomes and employment in Kenya, which elevated the ‘informal sector’ to the status of a major source for national development by the bootstraps, as it were. This was enough to encourage legions of researchers to adopt the term. Before long a substantial critique of the ‘informal sector’ concept had emerged. Marxists claimed that its proponents mystified the essentially regressive and exploitative nature of this economic zone, which they preferred to call ‘petty commodity production’. The study of Third World urban poverty rapidly became a new segment of the academic division of labour; as a key term in its discourse, the informal economy attracted an unusual volume of debate (Bromley, 1978). Later, sociologists applied the term to industrial societies (Pahl, 1984).

Hernando De Soto argued that Peru was a mercantilist state whose over-regulated and impenetrable national bureaucracy excluded the vast majority from effective participation in development. The latter were an entrepreneurial peasantry flocking in ever-larger numbers to the main cities. They were forced to operate informally, that is, outside the law, in sectors such as housing, trade and transport. Later, he portrayed poor countries like Peru as being trapped in a world economy dominated by the first industrial nations (De Soto, 2000). Red tape is mainly an effect of a global regime that forces marginal states to adopt inappropriate institutional practices. The result is the same: migrants pile up in the cities and are forced to work outside the law. Countries like the USA, which dominates this global financial bureaucracy, made the transition to modern capitalism by giving informal practices and decentralized violence full rein in their own development. Similar flexibility has to be shown today if the poor urban masses are to have a chance of joining global development on less unequal terms.

The idea of an ‘informal economy’ is entailed by the institutional effort to organize society along formal lines (Hart, 2006). ‘Form’ is *the rule*, an idea of what ought to be universal in social life; and for most of the 20th century the dominant forms have been those of bureaucracy, particularly of national bureaucracy, since society has become identified to a large extent with nation states. This identity may now be weakening in the face of the neoliberal world economy and a digital revolution in communications. Popularity as a jargon word has not helped the informal economy acquire a measure of analytical precision. For many it is a convenient name for an unambiguous empirical phenomenon – what you find in the slums of Manila. Others refer to size (large-scale–small-scale), productivity (high–low), visibility (enumerated–unenumerated), pattern of rewards (wages–self-employment), market conditions (monopoly–competitive) and much else. Hart (1973), like Geertz, explicitly derived his analysis from Weber's theory of rationalization. Much that goes on in developing countries today is only marginally the product of state regulation: it is thus ‘informal’ relative to the forms of publicly organized economic life. This is a qualitative distinction.

‘Form’ is the rule, the invariant in the variable. Idealist philosophers from Plato onwards thought the general idea of something was more real than the thing itself. The ‘formal sector’ is likewise an idea, a collection of people, things and activities that share an idea; but we should not mistake the idea for the reality that it partially identifies. What makes something ‘formal’ is its conformity with such an idea or rule. Thus formal dress in some societies means that the men will come dressed like penguins, but the women are free to wear something extravagant that suits them personally – they come as variegated butterflies. Formality endows a class of people with universal qualities, with being the same and equal. What makes dress ‘informal’ is therefore the absence of such a shared code. But informality is relative to the eye of the beholder. Any observer of an informally dressed crowd will notice that the clothing styles are not random. We might ask what these informal forms are and how to account for them. The world's ruling elite can be identified as ‘the men in suits’, because they choose to wear a style invented in the 1920s as an informal alternative to formal evening dress.

What the public and private sectors share is conformity to the rule of law at the national and increasingly international levels. How then might non-conformist economic activities, ‘the informal economy’, relate to this formal order? They may be related in any of four ways: as *division*, as *content*, as *negation* and as *residue*. The first two imply a positive relationship of interdependence, the third is antagonistic and the last relatively autonomous. The moral economy of capitalist societies is based on an attempt to keep separate impersonal and personal spheres of social life. The establishment of a formal public sphere entailed another based on domestic privacy. Most people, traditionally men more than women, divide

themselves every day between production and consumption, paid and unpaid work, submission to impersonal rules in the office and the free play of personality at home. Money is the means whereby the two sides are brought together, so that their interaction is an endless process of separation and integration that I call *division*.

For any rule to be translated into human action, something else must be brought into play, such as personal judgement. So informality is built into bureaucratic forms as unspecified *content*. Take a chain of commodities from their production by a transnational corporation to their final consumption in a Third World city. At several points invisible actors appear filling the gaps that the bureaucracy cannot handle directly, from the factories to the docks to the supermarkets and street traders who supply the cigarettes to smokers. Informal processes are indispensable to the trade, as variable content to the universal form. Of course, some of these activities may break the law, through a breach of health and safety regulations, tax evasion, smuggling, the use of child labour, selling without a licence, and so on.

The third way that informal activities relate to formal organization is thus as its *negation*. Rule breaking takes place both within bureaucracy and outside it; and so the informal is often illegal. The informalization of the world economy is to a large extent criminal and this includes white-collar crime.

The fourth category is not so obviously related to the formal order as the rest. Some ‘informal’ activities exist parallel to it, as *residue*. They are just separate from the bureaucracy. It would be stretching the logic of the formal–informal pair to include peasant economy, traditional institutions and much else besides within the rubric of the ‘informal’. Yet the social forms endemic to these often shape informal economic practices.

It is inconsistent to claim that the urban poor have an informal economy but their rich masters do not; or that the Third World has an informal sector but not the industrialized West. As long as there is formal economic analysis and the *partial* institutionalization of economies around the globe along capitalist lines, there will be a need for some such remedial concept as the informal economy. Its application to concrete conditions is stimulated by palpable discrepancies between prevalent models and observed realities. Such a discrepancy provoked the emergence of the concept in the 1970s, when Third World economies bore the brunt of the depression that marked the end of the West's post-war miracle. Later the accelerating decline of the British economy encouraged some social scientists to adopt the term there. The common strand is the growing inability of modern states to control the wider economic environment that sustains them. Hence the need for a dualistic model, such as that offered by the ‘informal economy’ concept.

See Also

- development economics
- economic anthropology
- labour surplus economies

Bibliography

Bromley, R., ed. 1978. The urban informal sector: critical perspectives. *World Development* 6(9–10).

Davis, M. 2006. *Planet of Slums*. New York: Verso.

De Soto, H. 2000. *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. London: Bantam.

Geertz, C. 1963. *Peddlers and Princes*. Chicago: University of Chicago Press.

Hart, K. 1973. Informal income opportunities and urban employment in Ghana. *Journal of Modern African Studies* 11, 61–89.

Hart, K. 2006. Bureaucratic form and the informal economy. In *Linking the Formal and Informal Economy: Concepts and Policies*, ed. B. Guha-Khasnobis, R. Kanbur and E. Ostrom. Oxford: Oxford University Press.

ILO (International Labour Organization). 1972. *Incomes, Employment and Equality in Kenya*. Geneva: ILO.

Lewis, O. 1964. *La Vida: A Puerto Rican Family in the Culture of Poverty – San Juan and New York*. New York: Random House.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economics and Social Studies* 22(2), 139–91.

Mayhew, H. 1861–2. *London Labour and the London Poor*, 4 vols. London.

Pahl, R. 1984. *Divisions of Labour*. Oxford: Basil Blackwell.

Weber, M. 1981. *General Economic History*. New Brunswick NJ: Transaction Publishers.

How to cite this article

Hart, Keith. "informal economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000102> doi:10.1057/9780230226203.0796

The New Palgrave Dictionary of Economics Online

information aggregation and prices

James Jordan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists commonly interpret market-clearing prices as the signals that competitive markets transmit to economic agents to facilitate the efficient allocation of resources. Informational decentralization theory formalizes this interpretation by characterizing the market mechanism as the unique decentralized mechanism that achieves efficient allocation with the minimal required communication. Rational expectations equilibrium theory formalizes a different aspect of the interpretation, showing that markets transmit to each trader all of the decision-relevant information in the market.

Keywords

Cobb–Douglas functions; efficient markets hypothesis; First Fundamental Welfare Theorem; full communication equilibrium; general equilibrium under uncertainty; Hayek, F.; information aggregation and prices; informational decentralization; rational expectations equilibrium

Article

Market-clearing prices aggregate decision-relevant information that is initially dispersed throughout the economic environment. Hayek (1945) asserts that competitive markets economize on the communication needed to achieve efficient allocations by embedding in prices all that any individual needs to know about the rest of the economy. During the 1970s and 80s, Hayek's famous assertion was interpreted and formalized by two distinct literatures in economic theory. Hurwicz's (1960) model of decentralized allocation mechanisms stimulated a literature on *informational decentralization theory* that led to the characterization of the market as the unique informationally decentralized allocation mechanism that minimizes the communication needed to achieve Pareto-efficient allocations. This result verifies Hayek's assertion, although the minimal message communicated by the market mechanism necessarily includes the market clearing trades as well as prices.
Hayek's assertion has a second connotation in financial asset markets and other markets involving trade

under uncertainty, where the dispersed information may be of direct interest to all traders. In this setting Hayek's assertion can be interpreted as a version of the *strong form* of the *efficient market hypothesis* (for example, Fama, 1970), which states that market prices constitute a sufficient statistic for all of the decision-relevant information in the market. Adding *rational expectations* (Muth, 1961) to models of general equilibrium under uncertainty makes it possible to formalize Hayek's assertion as follows. If all information in the market is directly communicated to all traders, the resulting equilibrium is also a rational expectations equilibrium. That is, each trader would find the market-clearing prices statistically sufficient for all of the decision-relevant private information of others. This version of Hayek's assertion is also verified, again provided that the market clearing trades are added to the prices in forming the sufficient statistic. There are also interesting cases in which the prices alone form a sufficient statistic.

Informational decentralization

Mount and Reiter (1974) formalized the general model of informationally decentralized allocation as follows. There are n agents, indexed $1 \leq i \leq n$. The private information, or environment, of agent i is an element e^i of a set E^i. The set of economic environments is the Cartesian product $E = E^1 \times \cdots \times E^n$, with generic element $e = (e^1, \ldots, e^n)$. Let Y denote a set of outcomes or allocations. The desired performance is modelled by a *performance correspondence* $g: E \rightarrow\rightarrow Y$, which associates with each environment e a set of desired allocations $g(e)$, where the double arrow is used to denote that g is set-valued. The communication among agents is embodied in messages $m \in M$, where M is the *message space*. Each message is associated with an allocation via an *outcome* function $h: M \rightarrow Y$. Finally, each environment is associated with a set of messages via a *message correspondence* $\mu: E \rightarrow\rightarrow M$. An allocation mechanism (μ, M, h) *realizes* the performance correspondence g if for all $e \in E$,

$$\mu(e) \neq \varnothing; \tag{1.1}$$

and

$$h(m) \in g(e) \text{ for all } m \in \mu(e). \tag{1.2}$$

An allocation mechanism (μ, M, h) is *informationally decentralized* if there are individual message correspondences $\mu^i: E^i \rightarrow\rightarrow M$ such that for each $e \in E$,

$$\mu(e) = \cap_{i=1}^{n} \mu^i(e^i). \tag{1.3}$$

This rather abstract definition can be interpreted by imagining that a computer chooses a message *m* at random and displays it to all agents. If for some agent *i*, $m \notin \mu^i(e^i)$, then agent *i* vetoes this message and another is chosen at random, until a message *m* is found that is accepted by all agents, that is, $m \in \cap_i \mu^i(e^i)$, at which point the allocation *h*(*m*) is chosen. The message *m* embodies the communication needed to achieve the allocation $h(m) \in g(e)$, because each agent *i* makes the veto \accept decision based only on e^i and *m*. An informationally decentralized allocation mechanism (μ ,*M*, *h*) that realizes a given *g* is *informationally efficient* if there is no other such mechanism that uses a message space that is smaller, in an appropriate sense, than *M*. If *E* and *Y* are finite, then the cardinality of *M* is the appropriate measure of size. In this case, the minimum required size agrees with the *communication complexity* of *g*, as that measure is defined in the computer science literature (for example, Karchmer, 1989). In models with continua of environments and allocations, it is more common to require that *M* be a manifold and interpret its dimension as its size. In this case, informational efficiency has the interpretation of using messages with the minimum number of real variables.

The application of this model to competitive markets is direct. Let *E* denote the set of pure exchange environments with *l* commodities. Assume that each trader's consumption set is the non-negative orthant R^l_+, so that trader *i*'s private information is $e^i=(u^i,\omega^i)$, where $u^i: R^l_+ \to R$ is trader *i*'s utility function and ω^i is an initial endowment bundle in R^l_+. Let *Y* be the set of all net trades that balance in the aggregate, $Y = \{y = (y^1, \ldots, y^n) \in R^{nl}: \Sigma_i y^i = 0\}$. The desired allocations are simply the Pareto efficient allocations that are also non-coercive, in the sense that traders are not forced below the utility level of their initial endowments. Formally, define $g: E \to \to Y$ as $g(e) = \{y \in Y: (\omega^1 + y^1, \ldots, \omega^n + y^n)$ is a Pareto-efficient allocation satisfying $u^i(\omega^i + y^i) \geq u^i(\omega^i)$ for all $i\}$. Non-coerciveness, which is sometimes called ‘individual rationality’, excludes the possibility of achieving Pareto efficiency by giving everything to one trader.

The *competitive allocation mechanism* is defined as follows. Let *P* denote the interior of the unit simplex in R^l_+, and define the *competitive message space* $M_c = \{(p, y) \in P \times Y: py^i = 0$ for all $i\}$. The outcome function $h_c: M_c \to Y$ is the projection $h_c(p, y) = y$, and the individual message correspondence $\mu^i_c: E^i \to \to M_c$ is defined as $\mu^i_c(u^i, \omega^i) = \{(p, y) \in M_c: \omega^i + y^i \in R^l_+$ and $u^i(\omega^i + y^i) \geq u^i(x^i)$ for all $x^i \in R^l_+$ satisfying $px^i \leq p\omega^i\}$. In effect, $\mu^i_c(e^i)$ is trader *i*'s offer curve, and the *competitive message correspondence* μ_c, defined as $\mu_c(e) = \cap_i \mu^i_c(e^i)$, is the intersection of the offer curves.

The competitive allocation mechanism (μ_c,M_c,h_c) is informationally decentralized by construction. If the sets E^i are restricted by conventional assumptions (for example, utility functions are continuous,

quasi-concave and strictly increasing, and endowments are strictly positive) then $\mu_c(e)$ is equal to the non-empty set of competitive equilibria for e, and the First Fundamental Welfare Theorem implies that the competitive allocation mechanism realizes g.

The informational efficiency of the competitive allocation mechanism was established by Hurwicz (1977) and Mount and Reiter (1974). The dimension of the competitive message space, M_c, is $n(l-1)$, so the informational efficiency of (μ_c, M_c, h_c) means that any other allocation mechanism (μ, M, h) which is informationally decentralized and realizes g must have $\dim M \geq n(l-1)$. This requires imposing sufficient mathematical regularity on (μ, M, h) so that dim M is well-defined and μ cannot behave as a Peano curve, encoding multi-dimensional information into the unit interval, for example. In particular, it is sufficient to require that M be (homeomorphic to) an open subset of a Euclidean space and that, on the set of exchange environments in which all traders have Cobb–Douglas utility functions, the correspondence μ admits a continuous selection on some open subset (on Cobb–Douglas environments, μ_c is itself single-valued and continuous). More general conditions are given by Mount and Reiter (1974), where the non-coerciveness requirement on g is also relaxed to require merely that for Cobb–Douglas environments, $g(e)$ includes only interior Pareto-efficient allocations. This excludes the mechanism that gives everything to trader 1 by using a message space of dimension $(n-1)l$ to enable traders 2,…, n to communicate their endowments.

The informational efficiency of competitive markets leaves open the possibility that other allocation mechanisms are also informationally efficient. This possibility is excluded by Jordan (1982b), albeit under stronger mathematical regularity conditions. In particular, the message correspondence μ is required to be single-valued and continuous on Cobb–Douglas environments, as opposed to merely having a local continuous selection. The non-coerciveness assumption is also much less dispensable for this result.

The informational decentralization literature verifies Hayek's assertion by characterizing the market mechanism as the unique informationally efficient mechanism that achieves non-coercive Pareto-efficient allocations. However, the competitive message is more than just the $l-1$ relative prices that are the focus of Hayek's insight. The realization of non-coercive Pareto-efficient allocations requires $n(l-1)$-dimensional messages because of the need to communicate the equilibrium trades as well as the prices.

Rational expectations equilibrium

A simple version of the rational expectations equilibrium model can be described as follows. Before trade, each trader i observes her endowment, $\omega^i \in R^l_+$, and a private signal, z^i, which is jointly distributed with the future state s, which is common to all traders, that determines her utility function $u^i(\cdot, s): R^l_+ \to R$. Assume for simplicity that there is only a finite number of possible private signal values, each of which has positive probability. In a rational expectations equilibrium, each trader i maximizes her expected utility conditional on her private signal z^i and any endogenous market variables she observes. To formulate the information aggregation condition, suppose that all private signals were publicly observable. Then every trader would observe all of the information in the market, so there

would be no need to infer information from market variables. The appropriate equilibrium concept would be a *full communication equilibrium* (FCE), defined as associating with each profile of signals $z = (z^i)_i$, a price vector $p(z)$ and net trades $y(z) = (y^i(z))_i$ satisfying, for each z,

$$\Sigma_i y^i(z) = 0, \tag{2.1}$$

and for each trader i,

$$\omega^i + y^i(z) \text{ maximizes } E\left\{u^i(x^i, \cdot)|z\right\} \text{subject to } p(z)x^i \leq p(z)\omega^i, \tag{2.2}$$

where the expectation is taken over s with respect to the conditional distribution over s given z. Thus, an FCE allocation is an allocation that would result if every trader possessed all of the information in the market. The information aggregation question is thus whether an FCE can be supported if traders are given only their private information and, for example, the equilibrium price vector. More precisely, does an FCE $(p(\cdot), (y^i(\cdot))_i)$ also satisfy

$$\omega^i + y^i(z) \text{ maximizes } E\left\{u^i(x^i, \cdot)|z^i, p(z)\right\} \text{subject to } p(z)x^i \leq p(z)\omega^i, \tag{2.3}$$

for each z and each trader i? Functions $p(\cdot), (y^i(\cdot))_i$ that satisfy (2.1) and (2.3) constitute a *rational expectations equilibrium* (REE). Thus the information aggregation question is whether an FCE is also an REE.

Kreps (1977) provides a simple example showing not only that the answer is ‘no’, but that an REE can easily fail to exist. In the Kreps example, there are two traders, two commodities and two equiprobable states of the world, $s \in \{a, b\}$. The traders’ endowments are $\omega^1 = \omega^2 = (3, 3)$ and their state-dependent utility functions are given by

$$u^1(x, a) = 2\ln x_1 + \ln x_2, u^1(x, b) = \ln x_1 + 2\ln x_2,$$

$$u^2(x, a) = \ln x_1 + 2\ln x_2, u^2(x, b) = 2\ln x_1 + \ln x_2.$$

Trader 1's signal is the state itself, $z^1 = s$, so trader 1 is fully informed; and trader 2's signal is a constant, $z^2 = \bar{z}^2$, so trader 2 is uninformed. Suppose by way of contradiction that $p(\cdot)$ is an REE price function. There are two possible cases. If $p(a, \bar{z}^2) \neq p(b, \bar{z}^2)$, then the price reveals the state to trader 2, and both traders are fully informed. However, it is easily seen that the (fully informed) environments $(\omega^i, u^i(\cdot, a))_i$ and $(\omega^i, u^i(\cdot, b))_i$ have the same unique equilibrium price $p = (\frac{1}{2}, \frac{1}{2})$. There remains only the case that $p(a, \bar{z}^2) = p(b, \bar{z}^2)$, in which trader 1 is fully informed but trader 2 remains uninformed. However, the exchange environments $((\omega^1, u^1(\cdot, a)), (\omega^2, \frac{1}{2}u^2(\cdot, a) + \frac{1}{2}u^2(\cdot, b)))$ and $((\omega^1, u^1(\cdot, b)), (\omega^2, \frac{1}{2}u^2(\cdot, a) + \frac{1}{2}u^2(\cdot, b)))$ have unique and distinct equilibrium prices, thus eliminating this case as well. Thus market prices cannot always aggregate all the information in the market, and the use of market prices as information signals as well as rates of exchange can prevent even the existence of equilibrium.

If traders condition their expectations on the entire competitive message, however, full information aggregation occurs. In fact, for each trader *i*, conditioning on $p(z)$ and $y^i(z)$ is enough. More precisely, every FCE also satisfies (2.3)

$$\omega^i + y^i(z) \text{ maximizes } E\{u^i(x^i, \cdot) | p(z), y^i(z)\} \text{subject to } p(z)x^i \leq p(z)\omega^i, \tag{2.4}$$

for each *z* and every trader *i*. It follows from the FCE property (2.2) that for any observed *p* and y^i, $\omega^i + y^i$ maximizes $E\{u^i(x^i, \cdot) | z\}$ subject to $px^i \leq p\omega^i$ for every *z* in the observed event $\{z: (p(z), y^i(z)) = (p, y^i)\}$. Thus the conditional expected utility function in (2.4) is a convex combination of expected utility functions, each of which is maximized by $\omega^i + y^i(z)$ subject to $p(z)x^i \leq p(z)\omega^i$. Therefore, the convex combination has the same maximum at the same constraint. Moreover, $(p(z), y^i(z))$ is the minimal market data needed for full information aggregation. Jordan

(1982a) shows that, if traders condition expectations on non-constant functions of the competitive message $(p(z), (v^i(z))_i)$, then each trader's data must be sufficient to reveal $(p(z), v^i(z))$ to each trader *i*, not only to ensure full information aggregation but even to avoid examples of nonexistence of equilibrium.

The simultaneous determination of expectations, prices and trades in a rational expectations equilibrium begs the question of how the private information *z* becomes embedded in the prices and trades, from which each trader can infer the private information of others. A dynamic interpretation is given by Jordan (1982c). Suppose that, initially, each trader *i* conditions expectations on the private signal z^i alone. This leads to an initial equilibrium price vector $p_1(z)$, but suppose that, before the equilibrium trades are executed, traders update their expectations using the information revealed by $p_1(z)$. This leads to a second equilibrium price vector $p_2(z)$, and so on, until a price vector $p_T(z)$ is reached that reveals no new information that changes any trader's demand. This process may fail to reveal all decision-relevant information to every trader. For example, if the only trader with a non-constant signal has state-independent preferences, no information will be revealed and the process will terminate at the first step. However, for each trader *i*, the final price and net trade $(p_T(z), v_T^i(z))$ is a sufficient statistic for all of the decision-relevant information trader *i* has learned from z^i and the temporary equilibrium prices. In this sense, the final prices and net trades summarize all of the private information revealed by prices along the temporary equilibrium path. The sequence of temporary equilibria is virtual in the sense that the temporary equilibrium trades are never executed. If they were, expectations of interim capital gains and losses could lead to nonexistence of temporary equilibrium, which is shown by an example in Jordan (1982c).

The Kreps (1977) example described above shows that prices alone cannot always support full information aggregation. However, the example is non-generic in the sense that a slight perturbation of the state-dependent utility functions can make the full communication equilibrium prices different in the two states, resulting in full revelation. Radner (1979) develops a financial asset market model in which the FCE price function p(·) is generically 1–1. In Radner's model, the set of future states and current signal values are both finite. Each future state corresponds to a vector of values for the assets that are currently traded. Each signal *z* is associated with a conditional probability vector over the future states, and thus the future asset values. Radner shows that the set of signal-conditional probability arrays that give rise to FCE price functions that are not 1 – 1 is a closed nowhere dense set of Lebesgue measure zero. This line of research was greatly extended in a series of papers by Allen (see Allen and Jordan, 1998, for a more detailed survey). Let *Z* denote the range of possible signal values *z*, and suppose that the relation between the signal *z* and conditional expected utility functions is sufficiently regular that the set of full communication environments $(E\{u^i(\cdot)|z\}, \omega^i)_i$ has dimension no larger than the dimension of *Z*. If *Z* is finite, as in Radner's model, both sets have dimension zero. Allen (1981) shows that, if $\dim Z < \frac{1}{2}\dim P$, then an FCE price function p(·) is generically 1 – 1. Allen (1982b) shows that, if $\dim Z < \dim P$, then an FCE price function is generically 1 – 1 except on a subset of *Z* having Lebesgue measure zero. This implies that, if the probability distribution over signals has a density function on *Z*, then an FCE price function is 1 – 1 on a set of signals having probability one, so that prices are again

fully informative. The dimensional inequality is crucial. Jordan and Radner (1982) provide a robust example of the nonexistence of price-conditional rational expectations equilibrium with $\dim Z = \dim P = 1$.

Most of the results described above do not substantially restrict the way in which traders' preferences depend on the unknown future state of the world. Financial asset market models, in contrast, typically involve special kinds of state-dependent preferences that give rise to some interesting cases in which market prices are fully revealing. The earliest full revelation result was obtained by Jerry Green (1973) in an Arrow-securities markets model. In this case, the securities traded are wealth claims contingent on each future state, and traders have private signals about the probability distribution over the future states. Green (1973) shows that the derivative of market excess demand with respect to the state probabilities has a dominant diagonal property that ensures that the function from the full communication probabilities to the FCE price vector is 1 – 1. Grossman (1981) generalizes Green's model to obtain the full revelation of decision-relevant information even when the FCE prices are not 1 – 1. However, Green (1977) shows that, if 'noise' is included in the environment in the form of random endowments, rational expectations equilibrium can fail to exist.

The Green–Grossman full revelation result depends on the completeness of the securities markets. In the absence of complete markets, full revelation through prices can be obtained under restrictions on the nature of the uncertainty or on traders' utility-of-wealth functions. Grossman (1978) considers a model with a single riskless asset and several risky assets. The future values of the risky assets have a joint normal distribution. Traders have private signals about the mean of this distribution, but the covariance matrix is fixed. Grossman (1978) shows that, if traders' utility-of-wealth functions exhibit non-increasing absolute risk-aversion, then the FCE price vector is a 1 – 1 function of the full communication mean, and thus reveals all decision-relevant information. The same asset markets are studied by Jordan (1983), but arbitrary small perturbations are allowed in traders' endowments and the joint probability distribution over private signals and future risky asset values. In this case, if the number of private signal variables exceeds the number of risky assets ($\dim Z > \dim P$), full revelation by prices is assured only for three special classes of utility-of-wealth functions: linear, exponential, and constant relative risk aversion with the same constant for all traders.

The full revelation of private information by the market seems inconsistent with the acquisition of costly private information. For this reason, Grossman and Stiglitz (1980) introduced a financial asset market model, generalized by Hellwig (1980), which has a price-conditional rational expectations equilibrium that is only partially revealing. This model assumes that traders have exponential utility, and that future risky asset values are normally distributed. Full revelation is prevented by adding noise to the model in the form of randomness in the aggregate supply of the risky asset. Unfortunately, the existence of rational expectations depends on the special parametric assumptions of the model. Allen (1982a; 1985a; 1985b) and Anderson and Sonnenschein (1982) develop general models of partially revealing approximate rational expectations equilibria, but the rational expectations equilibrium literature has not produced a general model of partially revealing equilibrium.

See Also

- efficient markets hypothesis

Bibliography

Allen, B. 1981. Generic existence of completely revealing equilibria for economies with uncertainty when prices convey information. *Econometrica* 49, 1173–99.

Allen, B. 1982a. Approximate equilibria in microeconomic rational expectations models. *Journal of Economic Theory* 26, 244–60.

Allen, B. 1982b. Strict rational expectations equilibria with diffuseness. *Journal of Economic Theory* 27, 20–46.

Allen, B. 1985a. The existence of rational expectations equilibria in a large economy with noisy price observations. *Journal of Mathematical Economics* 14, 67–103.

Allen, B. 1985b. The existence of fully rational expectations approximate equilibria with noisy price observations, *Journal of Economic Theory* 37, 213–53.

Allen, B. and Jordan, J. 1998. The existence of rational expectations equilibrium: a retrospective. In *Organizations with Incomplete Information*, ed. M. Majumdar. Cambridge: Cambridge University Press.

Anderson, R. and Sonnenschein, H. 1982. On the existence of rational expectations equilibrium. *Journal of Economic Theory* 26, 261–78.

Fama, E. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25, 383–417.

Green, J. 1973. Information, efficiency and equilibrium. Discussion Paper No. 284. Cambridge, MA: Harvard Institute of Economic Research, Harvard University.

Green, J. 1977. The non-existence of informational equilibria. *Review of Economic Studies* 44, 451–63.

Grossman, S. 1978. Further results on the informational efficiency of competitive stock markets. *Journal of Economic Theory* 18, 81–101.

Grossman, S. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48, 541–59.

Grossman, S. and Stiglitz, J. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 35, 393–408.

Hayek, F. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

Hellwig, M. 1980. On the aggregation of information in competitive markets. *Journal of Economic Theory* 22, 477–98.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical Methods in the Social Sciences*, ed. K. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

Hurwicz, L. 1977. On the dimensional requirements of informationally decentralized Pareto-satisfactory processes. In *Studies in Resource Allocation Process*, ed. K. Arrow and L. Hurwicz. Cambridge: Cambridge University Press.

Jordan, J. 1982a. Admissible market data structures: a complete characterization. *Journal of Economic Theory* 28, 19–31.

Jordan, J. 1982b. The competitive allocation process is informationally efficient uniquely. *Journal of Economic Theory* 28, 1–18.

Jordan, J. 1982c. A dynamic model of expectations equilibrium. *Journal of Economic Theory* 28, 235–54.

Jordan, J. 1983. On the efficient markets hypothesis. *Econometrica* 51, 1325–44.

Jordan, J. and Radner, R. 1982. Rational expectations in microeconomic models: an overview. *Journal of Economic Theory* 26, 201–23.

Karchmer, M. 1989. *Communication Complexity: A New Approach to Circuit Depth*. Cambridge, MA: MIT Press.

Kreps, D. 1977. A note on ‘fulfilled expectations’ equilibria. *Journal of Economic Theory* 14, 32–43.

Mount, K. and Reiter, S. 1974. The informational size of message spaces. *Journal of Economic Theory* 8, 161–92.

Muth, J. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Radner, R. 1979. Rational expectations equilibrium: generic existence and the information revealed by prices. *Econometrica* 47, 665–78.

How to cite this article

Jordan, James. "information aggregation and prices." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000247> doi:10.1057/9780230226203.0797

The New Palgrave Dictionary of Economics Online

information cascade experiments

Lisa Anderson and Charles A. Holt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bayes' rule; imitation; imperfect information; information cascade experiments; private information

Article

Cascade experiments test the theory that conformity can result from individuals receiving private imperfect information and making public decisions in a sequence (see information cascades).

Cascade theories provide a rational explanation for imitation even when people receive different private information. If a person gathers additional information by observing others' decisions, then a sequence of decisions that matches one alternative might be strong enough to outweigh that person's contrary private information. When the initial decisions in a sequence are correct, cascades can lead to better overall decision-making than private information alone. However, information cascades are problematic when the initial decision-makers in a queue receive incorrect information and convey it to others through their public (incorrect) decisions.

Anderson and Holt (1997) designed the first laboratory cascade experiment to test the theory described in Bikhchandani, Hirshleifer and Welch (1992). Participants were shown two cups labelled A and B. Cup A contained two light marbles and one dark marble. Cup B contained two dark marbles and one light marble. A six-sided die was used to determine whether Cup A or Cup B was selected at the start of each decision-making round. The cups were equally likely to be selected by the die throw. Once a cup was selected, each person saw one private draw from the cup, with the marble being returned to the cup after each draw. Each participant made a public prediction about which cup (A or B) was being used for the draws in a randomly determined sequence that changed from round to round. Sessions included six decision-makers who were paid two dollars for a correct prediction and nothing otherwise for each of 15 rounds.

In any given round, if the first two public predictions matched (AA or BB) it was rational (based on Bayes' rule) for all subsequent decision-makers to follow, regardless of which marble they saw drawn from the cup (see Bayesian statistics). Starting with prior probabilities of 1/2 for each cup, if the first decision-maker predicted cup A, others could rationally infer that he saw a light marble, since there were

more light marbles than dark marbles in Cup A. With this new information, the probability of Cup A should have been updated to 2/3. If the second decision-maker predicted Cup A, others could infer that he also saw a light marble, and the probability of Cup A being used for the draws should have been updated to 4/5. Even if the third person observed a dark marble, it was still more likely that Cup A was being used for the draws, and a cascade should start with the third decision-maker. Alternatively, if the first two decision-makers cancelled each other out (AB or BA) and the next two matched, then a cascade could start with the fifth person in the sequence.

Cascades were possible, based on the private draws and the decision-making sequence, in about half the Anderson and Holt (1997) experiments and actually formed in about 70 per cent of these cases. Almost all the people who did not join rational cascades were following private information that conflicted with the cascade. This type of deviation is explained by cascade models with small amounts of noisy behaviour, as described in Anderson and Holt (1997) and Goeree et al. (2007), who showed that incorrect cascades are not likely to persist in experiments with long sequences of decisions.

From a policy perspective, cascades are a concern because they hide information, since the private information of cascade followers is not revealed by their decisions. Kübluer and Weizsäcker (2004) studied whether or not people recognized the lack of information in conforming decisions by making participants pay a fee to see a private signal. In one version of their experiment, it was rational for only the first person in the sequence to purchase information, but the authors found that many people made irrational purchases. Some of this behaviour can be explained by a model with error, since it is rational to buy information if one cannot completely trust the quality of public decisions.

In addition to the studies discussed above, laboratory experiments have been used to test other variations of the seminal cascade theory including applications to voting (Hung and Plott, 2001), investment (Alsopp and Hey, 2000), markets (Drehmann, Oechssler and Roider, 2005; and Cipriani and Guarino, 2005) and advice-giving (Çelen, Kariv and Schotter, 2005).

See Also

- Bayesian statistics
- information cascades

Bibliography

Alsopp, L. and Hey, J.D. 2000. Two experiments to test a model of herd behavior. *Experimental Economics* 3, 121–36.

Anderson, L.R. and Holt, C.A. 1997. Information cascades in the laboratory. *American Economic Review* 87, 847–62.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1992. A theory of fads, fashion, custom and cultural change as information cascades. *Journal of Political Economy* 100, 992–1026.

Çelen, B., Kariv, S. and Schotter, A. 2005. Words speak louder than actions and improve welfare: an experimental test of advice and social learning. Working paper, Center for Experimental Social Science, New York University.

Cipriani, M. and Guarino, A. 2005. Herd behavior in a laboratory financial market. *American Economic Review* 95, 1227–443.

Drehmann, M., Oechssler, J. and Roider, A. 2005. Herding and contrarian behavior in financial markets: an internet experiment. *American Economic Review* 95, 1203–426.

Goeree, J.K., Palfrey, T.R., Rogers, B.W and McKelvey, R.D. 2007. Self-correcting information cascades. *Review of Economic Studies* (forthcoming).

Hung, A.A. and Plott, C.R. 2001. Information cascades: replication and an extension to majority rule and conformity-rewarding institutions. *American Economic Review* 91, 1508–20.

Kübler, D. and Weizsäcker, G. 2004. Limited depth of reasoning and failure of cascade formation in the laboratory. *Review of Economic Studies* 71, 425–41.

How to cite this article

Anderson, Lisa and Charles A. Holt. "information cascade experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000594> doi:10.1057/9780230226203.0798

The New Palgrave Dictionary of Economics Online

information cascades

Sushil Bikhchandani, David Hirshleifer and Ivo Welch
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An information cascade occurs when individuals, having observed the actions and possibly payoffs of those ahead of them, take the same action regardless of their own information signals. Informational cascades may realize only a fraction of the potential gains from aggregating the diverse information of many individuals, which helps explain some otherwise puzzling aspects of human and animal behaviour. For example, why do individuals tend to converge on similar behaviour? Why is mass behaviour prone to error and fads? The theory of observational learning, and particularly of information cascades, has much to offer economics and other social sciences.

Keywords

convergent behaviour; herding; imitation; information aggregation; information cascades and observational learning; observational learning; product life cycle; signalling; uncertainty

Article

An *information cascade* is a situation in which an individual makes a decision based on observation of others without regard to his own private information.

Social observers have long recognized that human beings have a deep-rooted proclivity to imitate. According to Machiavelli (1514, p. 152), ‘Men nearly always follow the tracks made by others and proceed in their affairs by imitation.’ Even animals imitate in choices of mate and territories. A common view among social scientists equates the conformity of individuals in large groups with irrationality – ‘fads’, ‘mass psychology’, or the ‘madness of crowds’.

However, there has also been recent recognition of the benefits of social influence. For example, zoologists have argued that, despite its possible disadvantages, imitation is an evolutionary adaptation that has promoted survival over thousands of generations by allowing individuals to take advantage of the hard-won knowledge of others (Gibson and Hoglund, 1992).

Nevertheless, as this article discusses, even when individuals are entirely rational, observational influence helps surprisingly little, leading to social outcomes that are inefficient and superficially may seem irrational. Irrationality undoubtedly affects social behaviour. Recent developments in the theory of observational learning, however, give reason to be sceptical about casual attributions of perverse social outcomes to irrational passions.

Why do people tend to ‘herd’ on similar actions? Why is mass behaviour prone to error and fads? The theory of observational learning helps explain some otherwise puzzling phenomena about human behaviour, and offers a vantage point for treating issues in economics and business strategy.

We call influence resulting from rational processing of information gained by observing others *observational learning* or *social learning*. Observational learning is only one of several possible causes of convergent behaviour. The simplest reason is that individuals can have identical beliefs and decision problems. Alternative reasons for conformity include positive payoff externalities, which lead to conventions such as driving on the right-hand side of the road; preference interactions, as with everyone desiring to wear the more ‘fashionable’ clothing as determined by what others are wearing; and sanctions against deviants, as with a dictator punishing opposition.

Among these theories, however, only observational learning explains why mass behaviour is error-prone, idiosyncratic, and often fragile in the sense that small shocks might lead to large shifts in behaviour. To understand how these effects arise, consider a sequence of rational individuals who take identical decisions under uncertainty. Each individual makes use of all relevant information – his own private signal and any inferences drawn from observing the choices of preceding individuals. As soon as the information gleaned from publicly

observable choices of others is even slightly more informative than the individual's private signal, he imitates his immediate predecessor without regard to his private information. Therefore, this individual's choice is uninformative about his signal, and at that point an information cascade starts. His immediate successor finds herself in an identical position; she imitates him (her immediate predecessor) and ignores her private signal. Based on the information conveyed by the actions of the first few individuals – the ones not in a cascade – every succeeding individual takes the same action. This action may be an incorrect one, so even small shocks such as the possible arrival of a different type of individual or a little new information can overturn it. Thus, observational learning explains not only conformity but also rapid and short-lived fluctuations such as fads, fashions, booms and crashes.

The social outcome is highly error-prone because there is an information externality. If an individual selects an action that depends on his information signal, his action provides useful information to later decision-makers. However, it is in the self-interest of an individual in a cascade to ignore his signal; therefore, later individuals do not get the benefit of learning his private signal. Thus, the failure of individuals to take into account the welfare of later decision-makers leads to inefficient information aggregation.

This entry focuses on the situation where individuals with diverse private information learn by observing the actions of others or the consequences of these actions. (Previous surveys of this literature include Bikhchandani, Hirshleifer and Welch, 1998, and Chamley, 2004.)

Observable actions versus observable signals

Consider a setting in which individuals choose an action in a chronological order. Each individual starts with some private information, obtains some information from predecessors, and then decides on a particular action. We consider two scenarios. In the *observable actions* scenario, individuals can observe the actions but not the signals (that is, private information) of their predecessors. As demonstrated below, cascades will arise in this model. We compare this with a benchmark *observable signals* scenario in which individuals can observe both the actions and the signals of predecessors. (See Welch, 1992; Bikhchandani, Hirshleifer and Welch, 1992; Banerjee, 1992.)

The main ideas are seen in the following simple example. Several risk-neutral individuals decide in sequence whether to *adopt* or *reject* a possible action. The payoff to adopting, *V*, is either 1 or –1 with equal probability; the payoff to rejecting is 0. In the absence of further information, the two alternatives are equally desirable. The order in which individuals decide is given and known to all.

Each individual's signal is either High (H) or Low (L). It is H with probability $p > 1/2$ if $V = 1$, and with probability $1 - p$ if $V = -1$. Bayes' rule implies that, after observing one H, an individual's posterior probability that $V = 1$ is *p*; if instead one L is observed the probability that $V = 1$ is $1 - p$. All private signals are identically distributed and independent conditional on *V*. Naturally, an individual's posterior belief about *V* also depends on information derived from predecessors. All this is common knowledge among the individuals.

In the observable signals scenario, each individual observes predecessors' information signals. As the pool of public information keeps increasing, later individuals will settle on the correct choice (adopt if $V = 1$, reject if $V = -1$) and thus behave alike.

Because actions reflect information, it is tempting to infer that, if only the actions of predecessors are observable, the public information set will also gradually improve until the true value is revealed almost perfectly. But that is not the case. In the observable actions case, individuals often converge fixedly on the same wrong action – that is, the choice that yields a lower payoff, *ex post*. Furthermore, behaviour is *idiosyncratic* in that the choices of a few early individuals determine the choices of all successors.

To return to our example, the first individual, Asterix, adopts if his signal is H and rejects if it is L. All successors can infer Asterix's signal perfectly from his decision. If Asterix adopted, then Beatrix, the second individual, should also adopt if her private signal is H; as Beatrix sees it, there have now been two H signals, the one she inferred from Asterix's actions and the one she observed privately. However, if Beatrix's private signal is L, it exactly offsets Asterix's signal H. She is indifferent between adopting and rejecting. We assume, for expositional simplicity, that, as Beatrix is indifferent between the two alternatives, she tosses a coin to decide. (By similar reasoning, if Asterix rejected, then Beatrix should reject if she observes L, and toss a coin if her signal is H.)

The third individual, Cade, faces one of three possible situations: both predecessors adopted (AA), both rejected (RR), or one adopted and the other rejected (AR or RA). In case AA, Cade also adopts. He knows that Asterix observed H and that more likely than not Beatrix observed H too (although she may have seen L and flipped a coin). Thus, even if Cade sees a signal L, he adopts. Consequently, Cade's decision to adopt provides no information to his successors about the desirability of adopting. Cade is therefore in an *information cascade*; his optimal action does not depend on his private information. The uninformativeness of Cade's action means that no further information accumulates. Everyone after Cade faces the same decision and also adopts based only on the observed actions of Asterix and Beatrix. By similar reasoning, RR leads to a cascade of rejection starting with Cade.

In the remaining case where Asterix adopted and Beatrix rejected (or vice versa), Cade knows that Asterix observed H and Beatrix observed L (or vice versa). Thus, Cade's belief based on the actions of the first two individuals is that the $V = 1$ and $V = -1$ are equally likely. He finds himself in a situation identical to that of Asterix, so Cade's decision is based only on his private signal. Then, the decision problem of the fourth individual, Daisy, is the same as Beatrix's. Asterix's and Beatrix's actions have offset and thus carry no information to Eeyore. And if Cade and Daisy both take the same action – say, adopt – then an adoption cascade starts with Eeyore.

An individual's optimal decision rule is as follows. Let *d* be the difference between the number of predecessors who adopted and the number who rejected. If $d > 1$, then adopt regardless of private signal. If $d = 1$, then adopt if private signal is H and toss a coin if signal is L. If $d = 0$, then follow private signal. The decisions for $d = -1$ and $d < -1$ are symmetric. The difference between adoptions over rejections evolves randomly, and very quickly hits either the upper barrier of +2 and triggers an adoption cascade, or the lower barrier of –2 to trigger a rejection cascade. With virtual certainty, all but the first few individuals end up doing the same thing.

Order of information, noise, and information externalities

The reason the outcome with observable actions is so different from the observable signals benchmark is that, once a cascade starts, public information stops accumulating. An early preponderance towards adoption or rejection causes all subsequent individuals to ignore their private signals, which thus never join the public pool of knowledge. Nor does the public pool of knowledge have to be very informative to cause individuals to disregard their private signals. As soon as the public pool becomes slightly more informative than the signal of a single individual, individuals defer to the actions of predecessors and a cascade begins.

Furthermore, the type of cascade depends not just on how many H and L signals arrive, but on the order in which they arrive. For example, if signals arrive in the order HHLL…, then all individuals adopt, because Cade begins an adoption cascade. If, instead, the same set of signals arrive in the order LLHH…, all individuals reject, as Cade begins a rejection cascade. Thus, in the observable actions scenario, whether individuals on the whole adopt or reject is *path dependent*.

A cascade is likely even when private signals are noisy. Specifically, in the above example, let the probability that the signal is correct be $p = 0.51$. The probability that an adoption or rejection cascade forms after the first two individuals is close to 75 per cent! (The signal sequences HH – that is, Asterix observes H and Beatrix observes H – and LL cause adoption and rejection cascades respectively, starting with Cade. Similarly, HL and LH each lead to adoption and rejection cascades with probability 0.5 each, if the action chosen by Beatrix after a coin flip is the same as Asterix's. The sum of the probabilities of these events is about 0.75.) After eight players the probability is only 0.004 that the individuals are not in a cascade. (This is the probability $|d| < 2$ for each of individuals 3 through 8.)

Although a cascade starts eventually with probability one, the probability of being in a correct cascade (that is, an adoption cascade when $V = 1$ and a rejection cascade when $V = -1$) is only 0.5133. (The calculation can be found in Bikhchandani, Hirshleifer and Welch, 1992.) If individuals do not observe their predecessors' choices (or information), then they would choose an action based only on the private signal; the probability that an individual's choice is correct is 0.51. Thus, the increase in accuracy from observing the actions of predecessors is small. Contrast this with the observable signals scenario, where after many individuals the publicly observed information signals of predecessors are virtually conclusive as to the right action.

More generally, even when individuals have more accurate signals (*p* is much greater than 0.5), the information contained in a cascade is substantially short of efficient information aggregation. Consider the benchmark observable signals scenario. Individuals far enough out would know the true state almost perfectly. The correctness of these individuals' actions increases from *p* to 1 due to information revelation. Figure 1 graphs, as a function of the signal accuracy *p*, the fraction of potential accuracy improvement realized in the observable actions scenario. (The fraction of potential accuracy improvement realized is $(Pr[\text{correct cascade}] - p)(1 - p)$. From (3) in Bikhchandani, Hirshleifer and Welch, 1992, $Pr[\text{correct cascade}] = p(p+1)/2(1-p+p^2)$. This fraction increases from 0 for very noisy signals to 0.50 for very informative signals. Thus, in the basic model, at most half of the potential gains are realized.

Figure 1

Gains from action observability. *Note*: Fraction of potential accuracy improvement realized $\left(\left[\frac{p(p+1)}{2(1-p+p^2)} - p\right] / (1-p)\right)$ as a function of signal accuracy (*p* is the probability that the signal is high given that the true value is high).

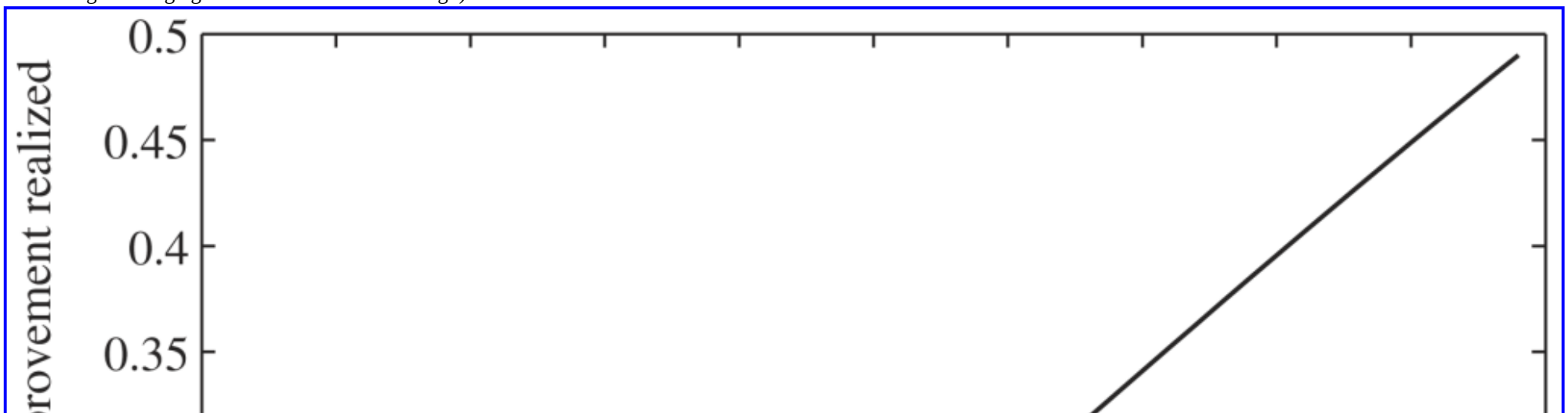

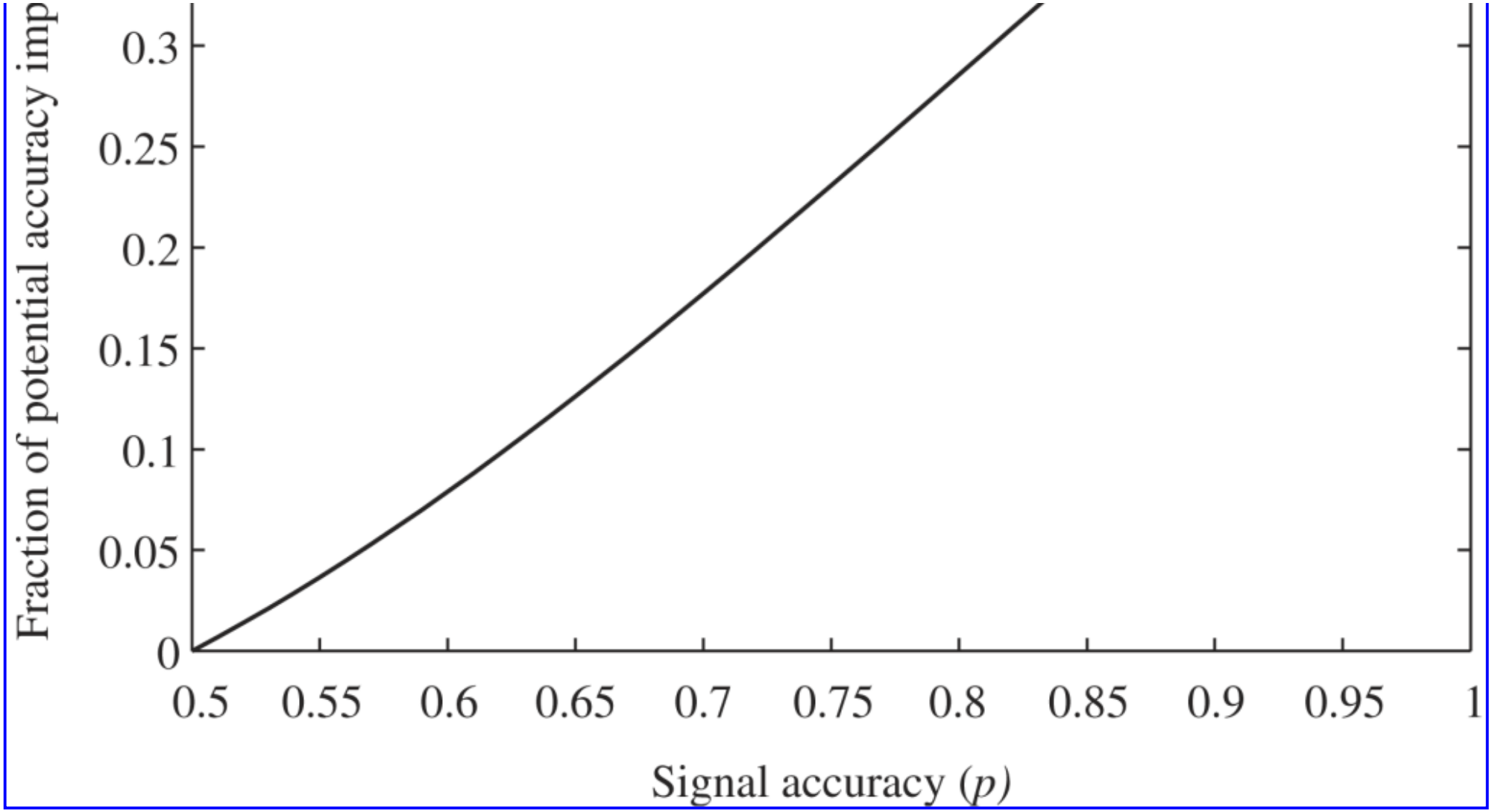

An individual's private information is useful to others. However, in choosing the optimal action, the individual ignores this benefit: with the onset of a cascade in the observable actions scenario, individuals rationally take uninformative imitative actions. This information externality reduces information aggregation. To see this, consider an alternative benchmark scenario in which (*a*) each individual maximizes a discounted sum of payoffs to all individuals and (*b*) no individual can directly reveal his private information; others learn of his information only through this individual's choice of action. The onset of cascades in this scenario is delayed (compared with the observational actions scenario); information aggregation is efficient subject to the constraint that private information is revealed only through actions.

Fragility

Of course, in reality we do not expect a cascade to last for ever. The arrival of better-informed individuals or the release of new public information can easily dislodge a cascade. Indeed, participants in a cascade know that the cascade is based on information that is only slightly more accurate than the private information of an individual. Thus, a key prediction of the theory is that behaviour in cascades is *fragile* with respect to small shocks. (In some models in which conformity is enforced by the threat of sanctions upon defectors, rare shifts occur when the system crosses a critical value that shifts the outcome from one equilibrium to another; Kuran, 1989.)

How robust are the conclusions that cascades are born quickly and idiosyncratically, and shatter easily? When some assumptions in the example are relaxed, is the aggregation of

information still inefficient or delayed?

Robustness of the basic model

The conclusions of the basic model remain robust along a number of dimensions. We discuss here alternative assumptions about the action space and the signal, which affect the conclusions to some extent.

The action space

In the basic model, players make inferences about others' signals from observed choices. When there are many possible actions, the action choice can convey more information. If the set of actions is continuous and unbounded, then actions fully reveal players' information and cascades do not arise (Lee, 1993). (If the action space is a continuous but bounded interval, then when an individual optimally chooses one of the end points of the interval, the value of his signal is not revealed by his action. In consequence, incorrect cascades can form at the end points of the interval.) For example, if a set of firms cannot invest less than zero, they may incorrectly cascade on zero investment.
However, if players are even slightly unsure of the payoff functions of other players, then there is a discontinuous shift to a slower learning process in which information aggregation is inefficient (Vives, 1993). In many real-world settings, the action space is bounded or partly discrete: investment projects that have a minimum efficient scale, elections amongst a discrete set of alternatives, a car purchase of a Ford or a Toyota, a takeover decision of whether to bid or not bid for a target firm, and a decision to hire or fire a worker.

The signal space

As in the simple two signal example presented above, in settings with a large but discrete set of signal values cascades occur with probability close to one and are sometimes incorrect. In some continuous signal settings cascades do not form (Smith and Sorensen, 2000), but an informational externality remains and information aggregation is inefficient. Furthermore, with substantial probability individuals soon follow the behaviour of recent predecessors, and with some probability that action is incorrect. Indeed, with any finite number of individuals, a continuous signal setting is observationally similar to a discrete signal setting that approximates the continuous model. In other words, in a continuous signals setting herds tend to form in which an individual follows the behaviour of his predecessor with high probability, even though this action is not necessarily correct. Thus, the welfare inefficiencies of the discrete cascades model are also present in continuous settings (Chamley, 2004, ch. 4).

Observability of payoffs or signals

Several papers consider the inefficiency of social learning when there is some degree of observability of payoffs (Caplin and Leahy, 1994). Furthermore, even if individuals can observe the payoffs of predecessors, inefficient cascades can form and with positively probability last for ever, because a cascade can lock into an inferior choice before sufficient trials have been performed on the other alternative to persuade later individuals that this alternative is superior (Cao and Hirshleifer, 2002). Indeed, if individuals can observe a subset of past signals, such as the past *k* signals, inefficient cascades can form.

Other assumptions of basic model

When individuals have the freedom to delay their action choice, in equilibrium there is delay, followed by a sudden onset of cascades when an individual commits to an action (Chamley and Gale, 1994; Zhang, 1997). The existence, idiosyncrasy and fragility of cascades are robust to relaxing other assumptions as well, including allowing for differing information precision, costly information acquisition, and heterogeneous observable tastes (see Bikhchandani, Hirshleifer and Welch, 1998, and the references therein). Inefficient cascades still form when individuals have reputational as well as informational motives to herd (Ottaviani and Sorensen, 2000). When individuals are imperfectly rational, inefficient cascades still form, but overconfident individuals provide social value when their impetuous choices shatter incorrect cascades (Bernardo and Welch, 2001).

Applications

There has been extensive testing of information cascades models in the laboratory. Experiments provide some support for information cascades and observational learning (Anderson

and Holt, 1997).

Demand for goods and securities

The information cascades theory implies not just that consumer purchase decisions will be influenced by others, as occurs, for example, in automobile purchases in Finland (Grinblatt, Ikäheimo and Keloharju, 2004), but that the source of this influence is informational. In consequence, the cascades approach implies that the incorrect cascades arise in settings in which individuals observe summary statistics of others' behaviour, such as whether one product is outselling another. Golder and Tellis (2004) provide evidence that information cascades play a role in the dynamics of product life cycles. The cascades theory also implies that individuals who are viewed by others as being better informed will be fashion leaders, in the sense that their decisions can trigger immediate cascades. This can explain the effectiveness of a star basketball player's endorsement of a brand of sneakers, but not of his or her endorsement of a brand of beer.

Even without fashion leaders, there are ways for individuals to have disproportionate effects on the onset of information cascades. In a salient 1995 episode, management gurus Michael Treacy and Fred Wiersema secretly purchased 50,000 copies of their business strategy book in order to inflate the sales measures used to construct the *New York Times* best-seller list. Despite mediocre reviews, their book not only made the best-seller list but subsequently sold well enough to continue as a best-seller without further demand intervention by the authors.

The ubiquitous and legitimate marketing method of offering a low initial price may be a successful scheme for introducing an experience good: early adoptions induced by the low price help start a positive cascade. This idea was first analysed by Welch (1992) to explain why initial public offerings of equity are on average severely underpriced by issuing firms. Indeed, a seller may be tempted to cut price secretly for early buyers, so that later buyers will attribute the popularity of the product to high quality rather than low price.

Medicine

Most doctors cannot stay fully abreast of relevant medical advances in their specialties, suggesting that they may select among new treatments based primarily on observation of choices made by other doctors. The cascades approach implies that medical treatments will be characterized by localized conformity and occasional reversals triggered by limited information, and that doctors perceived as having special expertise will have disproportionate influence. It has indeed been claimed that a blind reliance by physicians upon their colleagues' medical decisions commonly leads to surgical fads and even to treatment-caused illnesses (Robin, 1984). Many dubious practices seem to have been adopted initially based on weak information (elective hysterectomy, ileal bypass and tonsillectomy), and then later abandoned. A few decades ago, differences in tonsillectomy frequencies in different countries and regions were extreme.

Politics

People learn about others' political beliefs by observing how they vote and from opinion and exit polls. Several studies of political momentum show that early respondents carry disproportionate weight (see Bartels, 1988). A possible non-informational explanation is that individuals have a direct preference to conform, but we would expect such an effect to be stronger when an individual is personally exposed to acquaintances with strong views than when the individual observes a polling statistic. Furthermore, polling numbers influence not just preference between candidates, but 'thermometer score' ratings of the perceived quality of candidates. Iowa voters gave an obscure candidate named Jimmy Carter a conspicuous early success in the 1976 US presidential campaign. Many southern states have coordinated their primaries early in the election cycle on the same date ('Super Tuesday') in order to increase their influence on the presidential election. The expanding turnout of protestors in Leipzig in 1989, which triggered the fall of communism in East Germany, has been modelled as an information cascade (Lohmann, 1994). More broadly, a recent literature on the social diffusion of ideas emphasizes that individual signals are sometimes not reflected in public discourse, leading to poor information aggregation in public policy decisions (Kuran and Sunstein, 1999).

Finance

The decision of individual investors to participate in the stock market and the buying and selling decisions of mutual fund managers are influenced by their peers' decisions (Hong, Kubik and Stein, 2005), and there is some indication that herding by mutual funds influences prices (Wermers, 1999). The rise in popularity of investment clubs and of day-trading in the 1990s was probably due in part to a self-feeding effect in which individuals learned from the media or word of mouth that many others were day trading. Several theoretical models of securities market trading (Avery and Zemsky, 1998) and market crashes (Lee, 1998) have been developed which embody either cascade or cascade-like features. Hirshleifer and Teoh (2003) review the theory and evidence of social learning and cascades in finance.

Zoology

Zoologists have documented observational learning, and proposed that information cascades are exhibited in a variety of animal behaviours, including ‘false alarm’ flights from possible predators, selection of night roosts by birds, and mate-choice copying in various animal species (Giraldeau, Valone and Templeton, 2002).

See Also

- product life cycle
- psychology of social networks
- social interactions (empirics)
- social interactions (theory)
- social norms

Bibliography

Anderson, L. and Holt, C. 1997. Information cascades in the laboratory. *American Economic Review* 87, 847–62.

Avery, C. and Zemsky, P. 1998. Multi-dimensional uncertainty and herd behavior in financial markets. *American Economic Review* 88, 724–48.

Banerjee, A. 1992. A simple model of herd behavior. *Quarterly Journal of Economics* 107, 797–818.

Bartels, L. 1988. *Presidential Primaries and the Dynamics of Public Choice*. Princeton, NJ: Princeton University Press.

Bernardo, A. and Welch, I. 2001. On the evolution of overconfidence and entrepreneurs. *Journal of Economics and Management Strategy* 10, 301–30.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1992. A theory of fads, fashion, custom and cultural change as informational cascades. *Journal of Political Economy* 100, 992–1026.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1998. Learning from the behavior of others: conformity, fads, and informational cascades. *Journal of Economic Perspectives* 12(3), 151–70.

Cao, H. and Hirshleifer, D. 2002. Taking the road less traveled: does conversation eradicate pernicious cascades? Dice Center Working Paper No. 2002–8. Ohio State University.

Caplin, A. and Leahy, J. 1994. Business as usual, market crashes, and wisdom after the fact. *American Economic Review* 84, 548–65.

Chamley, C. 2004. *Rational Herds: Economic Models of Social Learning*. Cambridge: Cambridge University Press.

Chamley, C. and Gale, D. 1994. Information revelation and strategic delay in irreversible decisions. *Econometrica* 62, 1065–85.

Choi, J. 1997. Herd behavior, the penguin effect, and suppression of information diffusion: an analysis of informational externalities and payoff interdependency. *RAND Journal of Economics* 28, 407–25.

Gale, D. 1996. What have we learned from social learning? *European Economic Review* 40, 617–28.

Gibson, R. and Hoglund, J. 1992. Copying and sexual selection. *Trends in Ecological Evolution* 7, 229–32.

Giraldeau, L.-A., Valone, T. and Templeton, J. 2002. Potential disadvantages of using socially acquired information. *Philosophical Transactions of the Royal Society: Biological Sciences* 357, 1559–66.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111, 507–48.

Golder, P. and Tellis, G. 2004. Growing, growing, gone: cascades, diffusion, and turning points in the product life cycle. *Marketing Science* 23, 207–18.

Grinblatt, M., Ikäheimo, S. and Keloharju, M. 2004. Interpersonal effects in consumption: evidence from the automobile purchases of neighbors. Working Paper No. 10226. Cambridge, MA: NBER.

Hirshleifer, D. and Teoh, S. 2003. Herding and cascading in capital markets: a review and synthesis. *European Financial Management* 9, 25–66.

Hong, H., Kubik, J. and Stein, J. 2005. Thy neighbor's portfolio: word-of-mouth effects in the holdings and trades of money managers. *Journal of Finance* 60, 2801–24.

Kuran, T. 1989. Sparks and prairie fires: a theory of unanticipated political revolution. *Public Choice* 61, 41–74.

Kuran, T. and Sunstein, C. 1999. Availability cascades and risk regulation. *Stanford Law Review* 51, 683–768.

Lee, I. 1993. On the convergence of informational cascades. *Journal of Economic Theory* 61, 396–411.

Lee, I. 1998. Market crashes and informational avalanches. *Review of Economic Studies* 65, 741–59.

Lohmann, S. 1994. The dynamics of informational cascades: the Monday demonstrations in Leipzig, East Germany, 1989–91. *World Politics* 47, 42–101.

Machiavelli, N. 1514. *The Prince*, ed. Q. Skinner and R. Price. Cambridge: Cambridge University Press, 1988.

Ottaviani, M. and Sorensen, P. 2000. Herd behavior and investment: comment. *American Economic Review* 90, 695–704.

Robin, E. 1984. *Matters of Life and Death: Risks vs. Benefits of Medical Care*. New York: Freeman and Co.

Smith, L. and Sorensen, P. 2000. Pathological outcomes of observational learning. *Econometrica* 68, 371–98.

Taylor, R. 1988. *Medicine Out of Control: The Anatomy of a Malignant Technology*. Melbourne: Sun Books.

Welch, I. 1992. Sequential sales, learning and cascades. *The Journal of Finance* 47, 695–732.

Wermers, R. 1999. Mutual fund herding and the impact on stock prices. *Journal of Finance* 54, 581–622.

Vives, X. 1993. How fast do rational agents learn? *Review of Economic Studies* 60, 329–47.

Zhang, J. 1997. Strategic delay and the onset of investment cascades. *RAND Journal of Economics* 28, 188–205.

How to cite this article

Bikhchandani, Sushil, David Hirshleifer and Ivo Welch. "information cascades." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000103> doi:10.1057/9780230226203.0799

The New Palgrave Dictionary of Economics Online

information sharing among firms

Xavier Vives
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Firms may have efficiency or strategic incentives to share information about current and past behaviour or intended future conduct. This article examines those incentives and the welfare consequences from the perspective of static oligopoly and monopolistic competition models. It concludes with a review of the available evidence.

Keywords

antitrust; Bertrand competition; collusion; common values; Cournot competition; disclosure; disclosure rules; hard information; independent values; information sharing among firms; misrepresentation; monopolistic competition; private values; product differentiation; soft information; trade associations

Article

Information sharing (IS) among firms has been a contentious topic in antitrust and has received substantial attention from researchers. Firms may share information about current and past behaviour of, for example, customers, orders and prices, as well as cost and demand conditions. This type of information exchange typically involves hard or verifiable information. Firms may also exchange information about intended future conduct – for example, planned prices, production, new products or capacity expansion. This typically involves soft information. Firms may have incentives to share information for efficiency or strategic reasons. The latter include influencing the behaviour of rivals or sustaining collusion. We will discuss here the results of static models, leaving out dynamic models of collusion and information signalling (see for those models Vives, 1999, sects. 8.4, 8.5 and 9.1.5; Kühn and Vives, 1995, sect. 8).

Firms may exchange cost or demand information in order to better adapt their output and pricing decisions to uncertainty. From the firm's point of view, the main effects of IS are the increased precision of information to be used by itself and rivals, and the corresponding impact on firms’ strategies. In

general, increased precision has a positive effect on a firm's expected profits, while the effect of increased precision of rivals and the induced strategy correlation depends on the nature of competition and shocks.

Information exchange is typically modelled as a two-stage game in which firms first unilaterally decide whether to reveal their signals, and then, after receiving those signals and possibly revealing them, compete *à la* Cournot or Bertrand. It is assumed that firms report their signals truthfully if they decide to share information. The workhorse model has quadratic payoffs and normal distributions (or distributions yielding linear conditional expectations) for signals and uncertain parameters such as demand intercepts and marginal costs. The assumptions yield linear equilibria at the second stage and explicitly computable payoffs. (See Vives, 1999, sect. 8.3.1, and Kühn and Vives, 1995, sects 2–5.) A sample of the literature is Novshek and Sonnenschein (1982), Clarke (1983), Vives (1984), Fried (1984), Gal-Or (1985; 1986), Li (1985), Sakai (1985), Shapiro (1986), Kirby (1988), Sakai and Yamato (1989), Raith (1996), and the extensions in Malueg and Tsutsui (1996; 1998). In the subgame-perfect equilibria of the two-stage game (excepting Bertrand competition with cost uncertainty) unilaterally revealing information is a dominant strategy with independent values, private values (that is, where each firm receives a signal with no noise about its payoff-relevant parameter), or common values with strategic complements. With common value and strategic substitutes, not revealing is a dominant strategy.

If firms are able to enter into industry-wide agreements, the determining factor is whether the information pooling situation increases or reduces expected profits. With the exception of Bertrand competition under cost uncertainty, expected profits with IS are always larger than without, under independent values, private values, and common value and strategic complements. With (for example, Cournot with substitutes), IS yields higher (lower) expected profits for a high (low) degree of product differentiation or steeply (slowly) rising marginal costs. Note that since IS often raises profits under one-shot interaction, IS cannot be taken as prima facie evidence of collusion.

IS agreements are usually mediated by trade associations that typically disclose an aggregate statistic of firms’ private signals. Monopolistic competition, where no firm has a significant impact on aggregate market outcomes, is suitable for examining the role of such associations’ disclosure rules. A firm first decides whether or not to join the association and reveal its private information. Under non-exclusionary disclosure, information is made available to everyone in the market; under exclusionary disclosure, it is provided to members only. Obviously, with a non-exclusionary disclosure rule, IS will not ensue if the sharing is costly (by not joining, a firm, being negligible in terms of aggregate market impact, can free ride and obtain market information costlessly, with no effect on market aggregates). With an exclusionary disclosure rule, IS may occur if the membership fee is not too high (see Vives, 1990).

The impact of IS on consumer surplus and total surplus depends on the type of competition and uncertainty, and on the number of firms. Three effects operate: output adjustment to information, output uniformity across varieties (given consumer preference for variety), and selection among firms of different efficiencies. IS may allow firms to better adjust to demand and/or costs shocks (output adjustment effect). This will tend to improve welfare except if the firm is a price setter and demand is uncertain. In this case, more information will give the firm greater scope to extract consumer surplus – an insight already valid for a monopolist. In monopolistic competition, where variety must be taken into account, IS tends to make the outputs of varieties more similar with common value uncertainty and less so with private value uncertainty, thus increasing (decreasing) expected total surplus under demand

uncertainty and Cournot (Bertrand) competition (Vives, 1990).

Analysis of the oligopoly case is complex, but several generalizations hold. Under demand uncertainty and Cournot competition, IS increases expected total surplus (ETS); under demand uncertainty and Bertrand competition, it decreases consumer surplus (as well as ETS, under monopolistic competition). With common values, IS always increases ETS, except under price competition, when goods are poor substitutes and/or there are many firms. (See Kühn and Vives, 1995, sect. 5.2, and Vives, 1999, sect. 8.3.3.) There are potentially large efficiency benefits from information exchange. For example, the production rationalization effect of cost information exchange under Cournot can be very large and is of a larger order of magnitude than the market power effect (Vives, 2002).

What happens when there is no trade association to provide a mechanism to share information truthfully? Assume private cost information that is exchangeable only at an interim stage, once each firm learns its own cost but does not know its rivals'. In this case, if information is not verifiable and there are no other signalling possibilities, information revelation is impossible, since all firms would like to be perceived as being low-cost. With verifiable information, full revelation ensues if disclosure is costless and it is known whether firms have information (Okuno-Fujiwara, Postlewaite and Suzumura, 1990; Van Zandt and Vives, 2006). The lowest-cost firm will reveal its type and then all other types will unravel. Information could also be revealed through costly signalling in the form of wasteful advertising (for example, Ziv, 1993), or via dynamic competition in which production levels are observable (Mailath, 1989) or with sales reports (Jin, 1994). In the latter case, sharing sales reports eliminates the incentive to misrepresent and changes the consequences of IS. If it is possible to verify information but not whether the firm is informed, then the unravelling result need not hold, and firms can selectively disclose acquired information (Jansen, 2005).

Evidence on the effect of IS among firms is scant. Genesove and Mullin (1999) study information exchange in the Sugar Institute and find no misreporting, but some information withholding, suggesting that information can be verified. Doyle and Snyder (1999) study production plans announcements in the trade press in the automobile industry and find that a firm's announcement affects competitors' responses. Announcements of increased production are met by upward adjustments in production, which they interpret as consistent with announcements signalling a common demand parameter. Christensen and Caves (1997) study capacity announcements in the pulp and paper industry and find that unexpected announcements by rivals promote project abandonment in sub-industries with low concentration levels (and the opposite in concentrated sub-industries); they compare these results with IS models of cost information. Armantier and Richard (2003) examine exchange of cost information in the multi-market context of the airline industry. The authors account for entry decisions in a Cournot setting with complementary goods across markets, and simulate a hypothetical agreement to share cost information by American Airlines and United Airlines at Chicago O'Hare airport. They find that IS would improve airline profitability and moderately harm consumers (although, theoretically, cost IS need not necessarily hurt consumers in such a situation). The experimental results in Cason (1994) suggest that pricing behaviour is influenced by IS decisions. Ackert, Church and Sankar (2000) find that in a Cournot game with cost uncertainty, where it cannot be verified whether a firm has received information, when a firm receives information about industry-wide cost unfavourable information is disclosed but favourable information is withheld. Contrary to theory, when information is about a cost-specific shock, disclosure

is not affected by the favourableness of information.

See Also

- cartels

Bibliography

Ackert, L., Church, B. and Sankar, M. 2000. Voluntary disclosure under imperfect competition: experimental evidence. *International Journal of Industrial Organization* 18, 81–105.

Armantier, O. and Richard, O. 2003. Exchanges of cost information in the airline industry. *RAND Journal of Economics* 34, 461–77.

Cason, T. 1994. The impact of information sharing opportunities on market outcomes: an experimental study. *Southern Economic Journal* 61, 18–39.

Christensen, L. and Caves, R. 1997. Cheap talk and investment rivalry in the pulp and paper industry. *Journal of Industrial Economics* 45, 47–73.

Clarke, R. 1983. Collusion and the incentives for information sharing. *Bell Journal of Economics* 14, 383–94.

Doyle, P. and Snyder, C. 1999. Information sharing and competition in the motor vehicle industry. *Journal of Political Economy* 107, 1326–64.

Fried, D. 1984. Incentives for information production and disclosure in a duopolistic environment. *Quarterly Journal of Economics* 99, 367–81.

Gal-Or, E. 1985. Information sharing in oligopoly. *Econometrica* 53, 329–43.

Gal-Or, E. 1986. Information transmission-Cournot and Bertrand equilibria. *Review of Economic Studies* 53, 85–92.

Genesove, D. and Mullin, W. 1999. The Sugar Institute learns to organize information exchange. In *Learning by Doing in Markets, Firms and Countries*, ed. N. Lamoreaux, D. Raff and P. Temin. Chicago: University of Chicago Press for the NBER.

Jansen, J. 2005. Information acquisition and strategic disclosure in oligopoly. Mimeo Social Science Research Center Berlin (WZB) and Humboldt University, Berlin.

Jin, Y. 1994. Information sharing through sales report. *Journal of Industrial Economics* 42, 323–33.

Kirby, A. 1988. Trade associations as information exchange mechanisms. *RAND Journal of Economics* 19, 138–46.

Kühn, K. and Vives, X. 1995. *Information Exchanges among Firms and their Impact on Competition*. Luxembourg: Office for Official Publications of the European Communities.

Li, L. 1985. Cournot oligopoly with information sharing. *RAND Journal of Economics* 16, 521–36.

Mailath, G. 1989. Simultaneous signalling in an oligopoly model. *Quarterly Journal of Economics* 104, 417–27.

Malueg, D. and Tsutsui, S. 1996. Duopoly information exchange: the case of unknown slope. *International Journal of Industrial Organization* 14, 119–36.

Malueg, D. and Tsutsui, S. 1998. Distributional assumptions in the theory of oligopoly information exchange. *International Journal of Industrial Organization* 16, 785–97.

Novshek, W. and Sonnenschein, H. 1983. Fulfilled expectations Cournot duopoly with information acquisition and release. *Bell Journal of Economics* 13, 214–18.

Okuno-Fujiwara, M., Postlewaite, A. and Suzumura, K. 1990. Strategic information revelation. *Review of Economic Studies* 57, 25–47.

Raith, M. 1996. A general model of information sharing in oligopoly. *Journal of Economic Theory* 71, 260–88.

Sakai, Y. 1985. The value of information in a simple duopoly model. *Journal of Economic Theory* 36, 36–54.

Sakai, Y. and Yamato, T. 1989. Oligopoly, information and welfare. *Journal of Economics (Zeitschrift für Nationalökonomie)* 49, 3–24.

Shapiro, C. 1986. Exchange of cost information in oligopoly. *Review of Economic Studies* 53, 433–46.

Van Zandt, T. and Vives, X. 2006. Monotone equilibria in Bayesian games of strategic complementarities. *Journal of Economic Theory* (forthcoming).

Vives, X. 1984. Duopoly information equilibrium: Cournot and Bertrand. *Journal of Economic Theory* 34, 71–94.

Vives, X. 1990. Trade association, disclosure rules, incentives to share information and welfare. *RAND Journal of Economics* 22, 446–53.

Vives, X. 1999. *Oligopoly Pricing: Old Ideas and New Tools*. Boston: MIT Press.

Vives, X. 2002. Private information, strategic behavior, and efficiency in Cournot markets. *RAND Journal of Economics* 33, 361–76.

Ziv, A. 1993. Information sharing in oligopoly: the truth-telling problem. *RAND Journal of Economics* 24, 455–65.

How to cite this article

Vives, Xavier. "information sharing among firms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000250> doi:10.1057/9780230226203.0800

The New Palgrave Dictionary of Economics Online

information technology and the world economy

Dale W. Jorgenson and Khuong Vu

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article analyses the impact of investment in information technology (IT) on the recent resurgence of world economic growth. We describe the growth of the world economy, seven regions, and 14 major economies during the period 1989–2004. We allocate the growth of world output between input growth and productivity and find, surprisingly, that input growth greatly predominates. Moreover, differences in per capita output levels are explained by differences in per capita input rather than variations in productivity. The contributions of IT investment have increased in all regions, but especially in industrialized economies and Developing Asia.

Keywords

Asian miracle; growth accounting; human capital; information technology and the world economy; input growth; productivity growth

Article

1 Introduction

This article analyses the impact of investment in information technology (IT) equipment and software on the recent resurgence in world economic growth. The crucial role of IT investment in the growth of the US economy has been thoroughly documented and widely discussed. (See Dale Jorgenson and Kevin Stiroh, 2000, and Stephen Oliner and Daniel Sichel, 2000. The growth accounting methodology employed in this literature is discussed by Jorgenson, Mun Ho and Stiroh, 2005, and summarized by Jorgenson, 2005.) Jorgenson (2001) has shown that the remarkable behaviour of IT prices is the key to understanding the resurgence of American economic growth. This behaviour can be traced to developments in semiconductor technology that are widely understood by technologists and economists. Jorgenson (2003) has shown that the growth of IT investment jumped to double-digit levels after 1995 in

all the G7 economies – Canada, France, Germany, Italy, Japan, and the United Kingdom, as well as the United States. (Nadim Ahmad, Paul Schreyer and Anita Wolfl, 2004, have analysed the impact of IT investment in OECD countries. Bart van Ark et al., 2003; 2005), and Francesco Daveri, 2002, have presented comparisons among European economies.) These economies account for nearly half of world output and a much larger share of world IT investment. The surge of IT investment after 1995 resulted from a sharp acceleration in the rate of decline of prices of IT equipment and software. Jorgenson (2001) has traced this to a drastic shortening of the product cycle for semiconductors from three years to two years, beginning in 1995.

In Section 2 we describe the growth of the world economy, seven economic regions, and 14 major economies given in Table 1 during the period 1989–2003. (We include 110 economies with more than one million in population and a complete set of national accounts for the period 1989–2003 from Penn World Table, 2002, and World Bank Development Indicators Online, 2004. These economies account for more that 96 per cent of world output.) The world economy is divided among the G7 and Non-G7 industrialized economies, Developing Asia, Latin America, Eastern Europe and the former Soviet Union, North Africa and the Middle East, and sub-Saharan Africa. The 14 major economies include the G7 economies listed above and the developing and transition economies of Brazil, China, India, Indonesia, Mexico, Russia, and South Korea.

The world economy: shares in size and growth by group, region, and major economies. The measures for groups and the world are averages weighted by GDP (in PPP$) share

Group/region	Period 1989–1995			Period 1995–2003		
GDP growth	Average share		GDP growth	Average share		
GDP	Growth	GDP	Growth			
World (110 economies)	**2.50**	**100.00**	**100.00**	**3.45**	**100.00**	**100.00**
G7 (7 economies)	2.18	47.44	41.33	2.56	45.26	33.62
Developing Asia (16)	7.35	20.76	61.13	5.62	26.05	42.56
Non-G7 (15)	2.03	8.38	6.77	3.01	8.13	7.10
Latin America (19)	3.06	8.35	10.20	2.11	8.07	4.94
Eastern Europe (14)	–7.05	9.32	–26.76	2.87	6.57	5.47
Sub-Saharan Africa (28)	1.21	2.13	1.03	2.88	2.01	1.68
N. Africa and Middle East (11)	4.36	3.61	6.29	4.08	3.91	4.64
Economy	Period 1989–1995					Period 1995–2003

GDP growth	Avg. GDP share		Growth share		GDP growth	Avg. GDP share		Growth share		
Group	World	Group	World	Group	World	Group	World			
Seven world major economies (G7)										
Canada	1.39	4.90	2.32	3.12	1.29	2.51	4.78	2.17	4.69	1.58
France	1.30	7.10	3.37	4.23	1.75	1.92	6.76	3.06	5.05	1.70
Germany	2.34	10.80	5.12	11.58	4.79	0.86	10.20	4.63	3.41	1.15
Italy	1.52	7.42	3.52	5.17	2.14	1.48	6.99	3.17	4.05	1.36
Japan	2.56	16.23	7.70	19.03	7.88	1.39	15.73	7.13	8.54	2.88
United Kingdom	1.62	7.45	3.54	5.53	2.29	2.55	7.37	3.34	7.32	2.46
United States	2.43	46.11	21.87	51.34	21.26	3.56	48.16	21.76	66.92	22.46
All G7	**2.18**	**100.0**	**47.4**	**100.00**	**41.4**	**2.56**	**100.0**	**45.3**	**100.0**	**33.6**
Seven major developing and transition economies (GD7)										
Brazil	1.97	12.10	3.16	6.89	2.48	1.94	10.16	2.93	3.80	1.65
China	9.94	29.26	7.64	84.23	30.36	7.13	37.79	10.91	51.99	22.55
India	5.03	18.98	4.95	27.65	9.97	6.15	20.69	5.97	24.54	10.65
Indonesia	6.82	7.12	1.86	14.07	5.07	2.41	6.98	2.02	3.25	1.41
Mexico	2.19	7.48	1.95	4.74	1.71	3.56	6.74	1.95	4.63	2.01
Russian Federation	–8.44	19.92	5.20	–48.71	–17.56	3.18	12.17	3.52	7.46	3.24
South Korea	7.48	5.14	1.34	11.13	4.01	4.09	5.47	1.58	4.32	1.87
All GD7	**3.45**	**100.0**	**26.1**	**100.0**	**36.0**	**5.18**	**100.0**	**28.9**	**100.0**	**43.4**

We have sub-divided the period in 1995 in order to focus on the response of IT investment to the accelerated decline in IT prices. As shown in Table 1, world economic growth has undergone a powerful revival since 1995. The per capita growth rate jumped nearly a full percentage point from 2.50 per cent during 1989–95 to 3.45 per cent in 1995–2003. We can underline the significance of this difference by pointing out that per capita growth of 3.45 per cent doubles world output per capita in a little over two decades, while slower growth of 2.50 per cent doubles per capita output in slightly less than three decades.

In Section 3 we allocate the growth of world output between input growth and productivity. Our most astonishing finding is that input growth greatly predominated! Productivity growth contributed only one-fifth of the total during 1989–95, while input growth accounted for almost four-fifths. Similarly, input growth contributed more than 70 per cent of growth after 1995, while productivity accounted for less than 30 per cent. The only important departure from this worldwide trend is the Asian miracle before 1995, when the rate of economic growth in Developing Asia far outstripped the rest of the world and productivity growth predominated.

In Section 3 we distribute the growth of input per capita between investments in tangible assets, especially IT equipment and software, and investments in human capital. The world economy, all seven

regions, and the 14 major economies, except Indonesia and Mexico, experienced a surge in investment in IT after 1995. The soaring level of US IT investment after 1995 was paralleled by jumps in IT investment throughout the industrialized world. The contributions of IT investment in Developing Asia, Latin America, Eastern Europe, North Africa and the Middle East, and sub-Saharan Africa more than doubled after 1995, beginning from much lower levels. By far the most dramatic increase took place in Developing Asia.

In Section 4 we present levels of output per capita, input per capita and productivity for the world economy, the seven economic regions, and the 14 major economies. We find that differences in per capita output levels are primarily explained by differences in per capita input, rather than variations in productivity. Taking US output per capita in 2000 as 100.0, world output per capita was a relatively modest 23.9 in 2003. If we use similar scales for input and productivity, world input per capita in 2003 was a substantial 42.4 and world productivity a robust 56.3. Section 6 concludes the paper.

2 World economic growth, 1989 – 2003

In order to set the stage for analysing the impact of IT investment on the growth of the world economy, we first consider the shares of world product and growth for each of the seven regions and the 14 major economies presented in Table 1. Following Jorgenson (2001), we have chosen GDP as a measure of output. We employ the Penn World Table, generated by Alan Heston, Robert Summers and Bettina Aten (2002), as the primary data source on GDP and purchasing power parities for economies outside the G7 and the European Union, as it existed prior to enlargement in May 2004. (Maddison, 2001, provides estimates of national product and population for 134 countries for varying periods from 1820 to 1998 in his magisterial volume, *The World Economy: A Millennial Perspective*.)

We have revised and updated the US data presented by Jorgenson (2001) through 2003. Comparable data for Canada have been constructed by Statistics Canada (see John Baldwin and Tarek Harchaoui, 2003). Data for France, Germany, Italy, and the UK and the economies of the European Union before enlargement have been developed for the European Commission by Bart van Ark et al. (2003). Finally, data for Japan have been assembled by Jorgenson and Kazuyuki Motohashi (2005) for the Research Institute on Economy, Trade, and Industry. We have linked these data by means of the OECD's purchasing power parities for 1999 (OECD, 2002).

The G7 economies accounted for slightly under half of world product from 1989 to 2003. The per capita growth rates of these economies – 2.18 per cent before 1995 and 2.56 per cent afterward – were considerably below world growth rates. The growth acceleration of 0.60 per cent for the G7 economies lagged behind the jump in world economic growth. The G7 shares in world growth were 41.3 per cent during 1989–95 and 33.6 per cent in 1995–2003, well below the G7 shares in world product of 47.4 per cent and 45.3 per cent, respectively.

During 1995–2003 the United States accounted for 21.8 per cent of world product and 48.2 per cent of G7 output. After 1995 Japan fell from its ranking as the world's second largest economy to third largest after China. Germany dropped from fourth place before 1995, following the United States, China and Japan, to fifth place during 1995–2003, ranking behind India as well. Japan remained the second largest of the G7 economies, while Germany retained its position as the leading European economy. France,

Italy and the UK were similar in size, but less than half the size of Japan. Canada was the smallest of the G7 economies.

The US growth rate jumped from 2.43 per cent during 1989–95 to 3.56 per cent in 1995–2003. The period 1995–2003 included the shallow US recession of 2001 and the ensuing recovery, as well as the IT-generated investment boom of the last half of the 1990s. The United States accounted for more than half of G7 growth before 1995 and more than two-thirds afterward. The US share in world growth fell below its share in world product before 1995, but rose above the US product share after 1995. By contrast Japan's share in world economic growth before 1995 exceeded its share in world product, but fell short of the product share after 1995. The remaining G7 economies had lower shares of world growth than world product before and after 1995.

The 16 economies of Developing Asia generated slightly more than a fifth of world output before 1995 and more than a quarter afterward. The burgeoning economies of China and India accounted for more than 60 per cent of Asian output in both periods. (Our data for China are taken from the Penn World Table, 2002. Alwyn Young, 2003, presents persuasive evidence that the official estimates given, for example, by the World Development Indicators, 2004, exaggerate the growth of output and productivity in China.) The economies of Developing Asia grew at 7.35 per cent before 1995 and 5.62 per cent afterward. These economies were responsible for an astounding 61 per cent of world growth during 1989–95! Slightly less than half of this took place in China, while a little less than a third occurred in India. Developing Asia's share in world growth declined to 43 per cent during 1995–2003, remaining well above the region's share of 26.1 per cent of world product. China accounted for more than half of this growth and India about a quarter.

The 15 Non-G7 industrialized economies generated more than eight per cent of world output during 1989–2003. These economies were responsible for lower shares in world growth than world product before and after 1995. Prior to the fall of the Berlin Wall and the collapse of the Soviet Union, the 14 economies of Eastern Europe and the former Soviet Union were larger in size than the Non-G7, generating 9.3 per cent of world product. All of the economies of Eastern Europe experienced a decline in output during 1989–95. Collectively, these economies subtracted 26.8 per cent from world growth during 1989–95, dragging their share of world product down to 6.6 per cent. During 1989–1995 Russia's economy was comparable in size to Germany's, but from 1995 to 2003 the Russian economy was only slightly larger than the UK economy.

During 1989–95 the ten per cent share of the Latin American economies in world growth exceeded their eight-and-a-half percent share in world product. After 1995 these economies had a substantially smaller six per cent share in world growth, while retaining close to an eight-and-a-half share in world product with Brazil and Mexico responsible for more than 60 per cent of this. Brazil's share in world growth was below its three per cent share in world product before and after 1995, while Mexico's growth was lower than its product share before 1995 and higher afterward.

The 11 economies of North Africa and the Middle East, taken together, were comparable in size to France, Italy, or the UK, while the 30 economies of sub-Saharan Africa, as a group, ranked with Canada. The economies of North Africa and the Middle East had a share in world growth of 6.3 per cent during 1989–95, well above their 3.6 per cent share in world product. After 1995 their share in world growth fell to 4.6 per cent, still above the share in world product of 3.9 per cent. Growth in the economies of sub-Saharan Africa lagged behind their shares in world product during both periods.

3 Sources of world economic growth

We next allocate the sources of world economic growth during 1989–2003 between the contributions of capital and labour inputs and the growth of productivity. We find that productivity, frequently touted as the primary engine of economic growth, accounted for only 20–30 per cent of world growth. Nearly half of this growth can be attributed to the accumulation and deployment of capital and another a quarter to a third to the more effective use of labour. Our second objective is to explore the determinants of the growth of capital and labour inputs, emphasizing the role of investment in information technology equipment and software and the importance of investment in human capital.

We have derived estimates of capital input and property income from national accounting data for the G7 economies. We have constructed estimates of hours worked and labour compensation from labour force surveys for each of these economies. We measure the contribution of labour inputs, classified by age, sex, educational attainment, and employment status, by weighting the growth rate of each type of labour input by its share in the value of output. Finally, we employ purchasing power parities for capital and labour inputs constructed by Jorgenson (2003). (Purchasing power parities for inputs follow the methodology described in detail by Jorgenson and Eric Yip, 2000.)

We have extended these estimates of capital and labour inputs to the 103 Non-G7 countries using data sources and methods described in Section 5. (We employ data on educational attainment from Barro and Lee, 2001, and governance indicators constructed by Kaufmann, Kraay and Mastruzzi, 2004, for the World Bank; for further details, see Section 5.)

We have distinguished investments in information technology equipment and software from investments in other assets for all 110 economies in our study. We have derived estimates of IT investment from national accounting data for the G7 economies and those of the European Union before enlargement. We measure the contribution of IT investment to economic growth by weighting the growth rate of IT capital inputs by the shares of these inputs in the value of output. Similarly, the contribution of Non-IT investment is a share-weighted growth rate of Non-IT capital inputs. The contribution of capital input is the sum of these two components.

We have revised and updated the US data presented by Jorgenson (2001) on investment in information technology and equipment. (US data on investment in IT equipment and software, provided by the Bureau of Economic Analysis, BEA, are the most comprehensive and detailed. The BEA data are described by Grimm, Moulton and Wasshausen, 2005.) Data on IT investment for Canada have been have been constructed by Statistics Canada (Baldwin and Harchaoui, 2003). Data for the countries of the European Union have been developed for the European Commission by van Ark et al. (2003). Finally, data for Japan have been assembled by Jorgenson and Motohashi (2005). We have relied on the WITSA *Digital Planet Report* (2002; 2004) as the starting point for estimates of IT investment for the remaining economies. (WITSA stands for the World Information Technology and Services Alliance. Other important sources of data include the International Telecommunication Union, ITU, telecommunications indicators, the UNDP *Human Development* reports, and the Business Software Alliance, 2003. Additional details are given in Section 5.)

We have divided labour input growth between the growth of hours worked and labour quality, where quality is defined as the ratio of labour input to hours worked. This reflects changes in the composition

of labour input, for example, through increases in the education and experience of the labour force. The contribution of labour input is the rate of growth of this input, weighted by the share of labour in the value of output. Finally, productivity growth is the difference between the rate of growth of output and the contributions of capital and labour inputs.

The contribution of capital input to world economic growth before 1995 was 1.18 per cent, a little more than 47 per cent of the growth rate of 2.50 per cent. Labour input contributed 0.79 per cent or slightly less than 32 per cent, while productivity growth of 0.53 per cent or just over 21 per cent. After 1995 the contribution of capital input climbed to 1.56 per cent, around 45 per cent of output growth, while the contribution of labour input rose to 0.89 per cent, around 26 per cent. Productivity increased to 0.99 per cent or nearly 29 per cent of growth. We arrive at the astonishing conclusion that the contributions of capital and labour inputs greatly predominated over productivity as sources of world economic growth before and after 1995!

We have divided the contribution of capital input to world economic growth between IT equipment and software and Non-IT capital input. The contribution of IT almost doubled after 1995, less than a quarter to more than a third of the contribution of capital input. However, Non-IT was more important before and after 1995. We have divided the contribution of labour input between hours worked and labour quality. Hours rose from 0.39 per cent before 1995 to 0.62 per cent after 1995, while labour quality declined from 0.40 per cent to 0.27 per cent. Labour quality and hours worked were almost equal in importance before 1995, but hours worked became the major source of labour input growth after 1995.

The acceleration in the world growth rate after 1995 was 0.95 per cent, almost a full percentage point. The contribution of capital input explained 0.38 per cent of this increase, while the productivity accounted for 0.46 per cent. Labour input contributed a relatively modest 0.10 per cent. The jump in IT investment of 0.26 per cent was most important source of the increase in capital input. This can be traced to the stepped-up rate of decline of IT prices after 1995 analysed by Jorgenson (2001). The substantial increase of 0.23 per cent in the contribution of hours worked was the most important component of labour input growth.

Table 2 presents the contribution of capital input to economic growth for the G7 economies, divided between IT and Non-IT. Capital input was the most important source of growth before and after 1995. The contribution of capital input before 1995 was 1.28 or almost three-fifths of the G7 growth rate of 2.18 per cent, while the contribution of 1.43 per cent after 1995 was 55 per cent of the higher growth rate of 2.56 per cent. Labour input growth contributed 0.49 per cent before 1995 and 0.46 per cent afterward, about 22 per cent and 18 per cent of growth, respectively. Productivity accounted for 0.42 per cent before 1995 and 0.67 per cent after 1995 or less than a fifth and slightly more than a quarter of G7 growth, respectively.

Sources of output growth: 1995–2003 vs. 1989–1995. The measures for groups and the world are averages weighted by GDP (in PPP$) share

Economy	Period 1989–1995				Period 1995–2003		
GDP Growth	Sources of growth (% points per annum)			GDP Growth	Sources of growth (% points per annum)		
	Capital	Labour	TFP		Capital	Labour	TFP

ICT	Non-ICT	Hours	Quality	ICT	Non-ICT	Hours	Quality					
World (110 economies)	**2.50**	**0.27**	**0.91**	**0.39**	**0.40**	**0.53**	**3.45**	**0.53**	**1.03**	**0.62**	**0.27**	**0.99**
G7 (7 economies)	2.18	0.38	0.90	0.07	0.42	0.42	2.56	0.69	0.74	0.28	0.18	0.67
Developing Asia (16)	7.35	0.15	1.73	1.19	0.42	3.86	5.62	0.43	2.27	0.81	0.38	1.72
Non-G7 (15)	2.03	0.32	0.68	0.21	0.21	0.61	3.01	0.49	0.77	1.06	0.20	0.49
Latin America (19)	3.06	0.16	0.58	1.20	0.37	0.75	2.11	0.39	0.61	1.10	0.34	–0.32
Eastern Europe (14)	–7.05	0.10	–0.15	–0.86	0.36	–6.50	2.87	0.23	–0.81	0.01	0.39	3.06
Sub-Saharan Africa (28)	1.21	0.3	0.24	1.66	0.56	–1.39	2.88	0.29	0.68	1.18	0.42	0.32
N. Africa and Middle East (11)	4.36	0.15	0.72	1.43	0.56	1.50	4.08	0.40	0.88	2.02	0.49	0.30
Seven world major economies (G7)												
Canada	1.39	0.49	0.27	0.07	0.55	0.01	2.51	0.65	0.61	0.68	0.16	0.42
France	1.30	0.19	0.93	–0.17	0.61	–0.26	1.92	0.36	0.75	0.22	0.07	0.52
Germany	2.34	0.26	1.05	–0.42	0.33	1.12	0.86	0.40	0.50	–0.24	0.09	0.11
Italy	1.52	0.26	0.86	–0.35	0.38	0.37	1.48	0.46	0.96	0.61	0.27	–0.82
Japan	2.56	0.31	1.16	–0.39	0.54	0.94	1.39	0.56	0.26	–0.32	0.22	0.67
United Kingdom	1.62	0.27	1.69	–0.73	0.49	–0.10	2.55	0.65	0.19	0.38	0.26	1.07
United States	2.43	0.49	0.71	0.57	0.36	0.31	3.56	0.88	1.01	0.50	0.17	0.99
All G7	**2.18**	**0.38**	**0.90**	**0.07**	**0.42**	**0.42**	**2.56**	**0.69**	**0.74**	**0.28**	**0.18**	**0.67**
Seven major developing and transition economies (GD7)												
Brazil	1.97	0.09	0.29	0.99	0.39	0.20	1.94	0.46	0.24	0.67	0.37	0.21
China	9.94	0.17	2.12	0.87	0.45	6.33	7.13	0.63	3.17	0.45	0.39	2.49
India	5.03	0.09	1.18	1.27	0.43	2.06	6.15	0.26	1.77	1.22	0.41	2.49
Indonesia	6.82	0.10	1.62	1.64	0.43	3.04	2.41	0.09	1.47	0.91	0.41	–0.47
Mexico	2.19	0.24	0.95	1.48	0.38	–0.87	3.56	0.23	1.11	1.76	0.31	0.14
Russian Federation	–8.44	0.07	–0.07	–1.02	0.37	–7.79	3.18	0.10	–1.30	0.21	0.44	3.73

South Korea	7.48	0.29	2.31	1.45	0.31	3.13	4.09	0.46	1.67	0.86	0.26	0.85
All GD7	**3.45**	**0.13**	**1.17**	**0.72**	**0.41**	**1.03**	**5.18**	**0.40**	**1.70**	**0.74**	**0.39**	**1.96**

The powerful surge of IT investment in the United States after 1995 is mirrored in jumps in the growth rates of IT capital through the G7. The contribution of IT capital input for the G7 increased from 0.38 during the period 1989–95 to 0.69 per cent during 1995–2003, rising from 30 per cent of the contribution of capital input to more than 48 per cent. The contribution of Non-IT capital input predominated in both periods, but receded slightly from 0.90 per cent before 1995 to 0.74 per cent afterward. This reflected more rapid substitution of IT capital input for Non-IT capital input in response to swiftly declining prices of IT equipment and software after 1995.

The modest acceleration of 0.38 per cent in G7 output growth after 1995 was powered by investment in IT equipment and software, accounting for 0.31 per cent, while the contribution of Non-IT investment slipped by 0.16 per cent. Before 1995 the contribution of labour quality of 0.42 per cent accounted for more than 80 per cent of the contribution of G7 labour input, while the contribution of hours worked of 0.28 per cent explained more than 60 per cent after 1995. The rising contribution of hours worked was offset by the declining contribution of labour quality, while productivity growth rose by 0.25 per cent.

In Developing Asia the contribution of capital input increased from 1.88 per cent before 1995 to 2.70 per cent after 1995, while the contribution of labour input fell from 1.61 per cent to 1.19 per cent. These opposing trends had a slightly positive impact on growth. The significant slowdown in the Asian growth rate from 7.35 per cent to 5.62 per cent can be traced entirely to a sharp decline in productivity growth from 3.86 to 1.72 per cent. Productivity explained slightly over half of Asian growth before 1995, but only 30 per cent after 1995.

The first half of the 1990s was a continuation of the Asian Miracle, analysed by Krugman (1994), Lau (1999), and Young (1995). This period was dominated by the spectacular rise of China and India and the continuing emergence of the Gang of Four – Hong Kong, Singapore, South Korea, and Taiwan. However, all the Asian economies had growth rates considerably in excess of the world average of 2.50 per cent. The second half of the 1990s was dominated by the Asian financial crisis but, surprisingly, conforms much more closely to the ‘Krugman thesis’ attributing Asian growth to input growth rather than productivity.

The Krugman thesis was originally propounded to distinguish the Asian Miracle from growth in industrialized countries. According to this thesis, Asian growth was differentiated by high growth rates and a great predominance of inputs over productivity as the sources of high growth. In fact, productivity growth exceeded the growth of input during the Asian Miracle of the early 1990s! Moreover, growth in the world economy and the G7 economies was dominated by growth of capital and labour inputs before and after 1995. Productivity growth played a subordinate role and fell considerably short of the contributions of capital and labour inputs to world and G7 growth.

Developing Asia experienced a potent surge in investment in IT equipment and software after 1995. The contribution of IT investment more than doubled from 0.15 percent to 0.43 per cent, explaining less than eight per cent of the contribution of capital input before 1995, but almost 16 per cent afterward. The rush in IT investment was particularly powerful in China, rising from 0.17 per cent before 1995 to 0.63 per cent afterward. India fell substantially behind China, but outperformed the region as a whole, increasing the contribution of IT investment from 0.09 to 0.26 per cent.

Indonesia was the only major economy to experience a decline in the contribution of both IT and Non-IT

investment after 1995. South Korea's IT investment increased from 0.29 before 1995 to 0.46 per cent afterward, while Non-IT investment dropped as a consequence of the Asian financial crisis. The contribution of Non-IT investment in Asia greatly predominated in both periods and also accounted for most of the increase in the contribution of capital input after 1995. The contributions of hours worked and labour quality declined after 1995 with hours worked dominating in both periods.

Economic growth in the 15 Non-G7 industrialized economies accelerated much more sharply than G7 growth after 1995. The contribution of labour input slightly predominated over capital input before and after 1995. The contribution of labour input was 0.81 per cent before 1995, accounting for about 40 per cent of Non-G7 growth, and 1.26 after 1995, explaining 39 per cent of growth. The corresponding contributions of capital input were 0.75 per cent and 1.12 per cent, explaining 37 and 34 per cent of Non-G7 growth, respectively. Non-G7 productivity also rose from 0.47 before 1995 to 0.89 percent afterward; however, productivity accounted for only 23 and 27 per cent of growth in these two periods. The impact of investment in IT equipment and software in the Non-G7 economies doubled after 1995, rising from 0.22 per cent to 0.44 per cent or from 29 per cent of the contribution of Non-G7 capital input to 39 per cent. This provided a substantial impetus to the acceleration in Non-G7 growth of 1.25 per cent. Non-IT investment explained another 0.14 per cent of the growth acceleration. However, the increased contribution of hours worked of 0.49 per cent and improved productivity growth of 0.42 per cent predominated.

The collapse of economic growth in Eastern Europe and the former Soviet Union before 1995 can be attributed almost entirely to a steep decline in productivity during the transition from socialism. This was followed by a modest revival in both growth and productivity after 1995, bringing many of the transition economies close to levels of output per capita that prevailed in 1989. The contribution of capital input declined both before and after 1995, even as the contribution of IT investment jumped from 0.09 to 0.26 per cent. Hour worked also declined in both periods, but labour quality improved substantially.

Latin America's growth decelerated slightly after 1995, falling from 2.95 to 2.52 per cent. The contribution of labour input was 1.92 per cent before 1995 and 1.89 per cent afterward, accounting for the lion's share of regional growth in both periods. The contribution of capital input rose after 1995 from 0.72 per cent to 0.99 per cent, but remained relatively weak. Mexico's IT investment declined slightly after 1995, while Non-IT investment increased. Nonetheless the contribution of IT investment in Latin America more than doubled, jumping from 0.15 per cent before 1995 to 0.34 per cent afterward or from 21 per cent of the contribution of capital input to 34 per cent. Productivity was essentially flat from 1989 to 2001, rising by 0.31 per cent before 1995 and falling by 0.36 per cent after 1995.

Productivity in sub-Saharan Africa collapsed during 1989–95 but recovered slightly, running at minus 1.63 per cent before 1995 and 0.36 per cent afterward. The contribution of labour input predominated in both periods, but fell from 2.77 per cent to 1.89 per cent, while the contribution of capital input rose from 0.52 per cent to 0.99 per cent. Productivity in North Africa and the Middle East, like that in Latin America, was essentially stationary from 1989 to 2001, falling from a positive rate of 0.50 per cent before 1995 to a negative rate of minus 0.46 per cent afterward.

4 World output, input and productivity

The final step in our analysis of the world growth resurgence is to describe and characterize the levels of

output, input, and productivity for the world economy, the seven economic regions, and the 14 major economies in Table 3. We present levels of output per capita for 1989, before the transition from socialism, 1995, the start of the worldwide IT investment boom, and 2003, the end of the period covered by our study. We also present input per capita and productivity for the years 1989, 1995 and 2003, where productivity is defined as the ratio of output to input.

Levels of output and input per capita and productivity (US=100 in 2000). The levels for groups and the world are averages weighted by population share

Region/country	Output per capita			Input per capita			Productivity		
1989	1995	2003	1989	1995	2003	1989	1995	2003	
World	**18.9**	**20.0**	**23.9**	**38.5**	**38.5**	**42.4**	**49.0**	**52.0**	**56.3**
G7	66.9	72.8	85.5	72.8	77.4	86.4	91.9	94.1	99.0
Developing Asia	6.0	8.5	12.1	19.1	21.5	26.2	31.7	39.7	46.1
Non-G7	51.5	56.0	68.0	61.9	64.9	75.9	83.2	86.4	89.5
Latin America	18.6	20.0	21.0	27.1	28.2	30.5	68.4	71.0	68.7
Eastern Europe	34.3	22.5	29.3	43.2	41.4	42.6	79.4	54.4	68.8
Sub-Saharan Africa	5.3	4.8	5.0	15.7	15.7	16.7	33.5	30.6	30.0
N. Africa and Middle East	12.5	14.2	17.0	22.3	23.2	27.3	55.9	61.1	62.3
Seven world major economies (G7)									
Canada	79.4	80.2	91.0	75.0	75.7	83.2	105.9	105.9	109.5
France	54.5	57.4	64.7	53.7	57.4	62.1	101.5	100.0	104.2
Germany	59.0	65.5	69.4	71.6	74.3	78.0	82.4	88.2	89.0
Italy	57.7	62.5	69.9	55.9	59.2	70.7	103.2	105.6	98.9
Japan	56.3	64.4	70.8	72.5	78.3	81.7	77.7	82.2	86.7
United Kingdom	56.9	61.8	73.7	61.7	67.5	73.9	92.2	91.6	99.8
United States	80.6	86.3	106.4	84.4	89.1	101.4	95.5	96.9	104.9
All G7	**66.9**	**72.8**	**85.5**	**72.8**	**77.4**	**86.4**	**91.9**	**94.1**	**99.0**
Seven major developing and transition economies (GD7)									
Brazil	19.9	20.5	21.5	29.3	29.8	30.8	67.9	68.7	69.8
China	4.8	8.1	13.4	17.9	20.7	28.0	26.9	39.3	48.0
India	5.0	6.0	8.6	15.9	17.0	19.9	31.2	35.3	43.1
Indonesia	8.3	11.3	12.2	23.7	26.8	29.9	35.3	42.3	40.7
Mexico	22.2	22.3	26.6	28.0	29.7	34.9	79.3	75.3	76.1
Russian Federation	41.8	25.1	33.5	50.0	48.0	47.4	83.6	52.4	70.6
South Korea	24.3	35.8	46.5	37.1	45.4	55.0	65.4	78.9	84.5
All GD7	**9.0**	**10.2**	**14.0**	**24.4**	**24.0**	**28.3**	**36.8**	**42.4**	**49.6**

The G7 economies led the seven economic regions in output per capita, input per capita, and

productivity throughout the period 1989–2003. Output per capita in the G7 was, nonetheless, well below US levels. If we take US output per capita in 2000 as 100.0, G7 output per capita was 66.9 in 1989, 72.8 in 1995 and 85.5 in 2003. For comparison: US output per capita was 80.6, 86.3, and 106.4 in these years. The output gap between the United States and the other G7 economies has widened considerably, especially after 1995. Canada was very close to the United States in output per capita in 1989, but dropped substantially behind by 1995. The United States–Canada gap widened further during the last half of the 1990s. Germany, Japan, Italy, and the UK had similar levels of output per capita throughout 1989–2003, but remained considerably behind North America. France lagged the rest of the G7 in output per capita in 1989 and failed to make up lost ground.

The United States was the leader among the G7 economies in input per capita throughout the period 1989–2003. If we take the United States as 100.0 in 2000, G7 input per capita was 72.8 in 1989, 77.4 in 1995, and 86.4 in 2003, while US input per capita was 84.4, 89.1, and 101.4, respectively. Canada, Germany and Japan were closest to US levels of input per capita with Canada ranking second in 1989 and 2003 and Japan ranking second in 1995. France lagged behind the rest of the G7 in input per capita throughout the period with Italy and the UK only modestly higher.

Productivity in the G7 has remained close to US levels, rising from 91.7 in 1989 to 93.9 in 1995 and 96.7 in 2001, with the United States equal to 100.0 in 2000. Canada was the productivity leader throughout 1989–2003 with Italy and France close behind. The United States occupied fourth place in 1989 and 1995, but rose to second in 2003. Japan made substantial gains in productivity, but lagged behind the other members of the G7 in productivity, while Germany surpassed only Japan.

Differences among the G7 economies in output per capita can be largely explained by differences in input per capita rather than gaps in productivity. The range in output was from 64.7 for France to 106.4 for the United States, while the range in input was from 62.1 for France to 101.4 for the United States. Productivity varied more narrowly from 86.7 for Japan to 109.5 for Canada with French productivity of 104.2 closely comparable to the United States.

In the economies of Developing Asia output per capita rose spectacularly from 6.0 in 1989 to 8.5 1995 and 12.1 in 2003 with the United States equal to 100.0 in 2000. Levels of output per capita in Asia's largest economies, China and India, remained at 13.4 and 8.6, respectively, in 2003. These vast shortfalls in output per capita relative to the industrialized economies are due mainly to differences in input per capita rather than variations in productivity. Developing Asia's levels of input per capita were 19.1 in 1989, 21.5 in 1995, and 26.2 in 2003, while Asian productivity levels were 31.7, 39.7, and 46.1, respectively.

China made extraordinary gains in output per capita, growing from 4.8 in 1989 to 8.1 in 1995 and 13.4 in 2003 with the United States equal to 100.0 in 2000. India had essentially the same output per capita as China in 1989, but grew less impressively to levels of only 6.0 in 1995 and 8.6 in 2003. China's input per capita – 17.9 in 1989, 20.7 in 1995, and 28.0 in 2001 – exceeded India's throughout the period. India's 31.2 productivity level in 1989 considerably surpassed China's 26.9. China's productivity swelled to 39.3 in 1995, outstripping India's 35.3. China expanded its lead with a productivity level of 48.0 in 2003 by comparison with India's 43.1.

Indonesia and South Korea grew impressively from 1989 to 1995, but fell victim to the Asian financial crisis during the period 1995–2003. Indonesia maintained its lead over India in output per capita, but dropped behind China in 2003. Indonesia led both China and India in input per capita during 1989–2003. Indonesia's productivity level led both China and India in 1995, but fell behind both economies by

2003. South Korea made substantial gains in productivity, achieving a level close to Japan in 2003, while falling considerably short of Japan's impressive input per capita.

The 15 Non-G7 industrialized economies, taken together, had levels of output per capita comparable to Germany, Italy, Japan, and the UK during 1989–2003. Input per capita for the 15 Non-G7 economies was also very close to these four G7 economies. However, productivity for the group was comparable to that of Germany, the second lowest in the G7.

Before the beginning of the transition from socialism in 1989, output per capita in Eastern Europe and the former Soviet Union was 34.3, well above the world economy level of 18.9, with the United States equal to 100.0 in 2000. The economic collapse that accompanied the transition reduced output per capita to 22.5 by 1995, only modestly higher than the world economy level of 20.0. A mild recovery between 1995 and 2003 brought the region back to 29.3, below the level of 1989, but well above the world economy average of 23.9. Input in the region was stagnant at 43.2 in 1989, 41.4 in 1995, and 42.6 in 2003. Productivity collapsed along with output per capita, declining from 79.4 in 1989 to 54.4 in 1995, before climbing back to 68.8 in 2003.

The downturn in output per capita and productivity was especially severe in the economies of the former Soviet Union. Russia's level of output per capita fell from 41.8 in 1989 to 25.1 in 1995 before recovering feebly to 33.5 in 2003. Russian input per capita remained essentially unchanged throughout the period 1989–2003, while productivity mirrored the decline and subsequent recovery in output, falling from a West European level of 83.6 in 1989 to 52.4 in 1995 before recovering to 70.6 in 2003. We conclude that the transition from socialism failed to restore Eastern Europe and the former Soviet Union to pre-transition levels of output and input per capita by 2003, while productivity remained weaker than before the transition.

For the Latin American region output per capita rose from 18.6 to 21.0 during 1989–2003, input per capita rose from 27.1 to 33.0, but productivity was essentially unchanged at about two-thirds of the US level in 2000. The stall in productivity from 1989 to 2003 was pervasive, contrasting sharply with the rise in productivity in the G7 economies, the Non-G7 industrialized economies, and Developing Asia. Nonetheless, Latin America's lagging output per capita was due chiefly to insufficient input per capita, rather than a shortfall in productivity.

Brazil's economic performance has been anaemic at best and has acted as a drag on the growth of Latin America and the world economy. Despite productivity levels comparable to the rest of Latin America, Brazil was unable to generate substantial growth in input per capita. Although Mexico lost ground in productivity between 1989 and 2003, rising input per capita produced gains in output per capita after 1995, despite a slight decline in the contribution of IT investment.

Output and input per capita in sub-Saharan Africa was the lowest in the world throughout the period 1989–2003, but the level of productivity was slightly higher than Developing Asia in 1989. All the economies of North Africa and the Middle East fell short of world average levels of output and input per capita. Output per capita grew slowly but steadily for the region as a whole during 1989–2003, powered by impressive gains in input per capita, but with stagnant productivity.

5 Methods and data sources

To measure capital and labour inputs and the sources of economic growth, we employ the production possibility frontier model of production and the index number methodology for input measurement

presented by Jorgenson (2001). For the G7 economies we have updated and revised the data constructed by Jorgenson (2003). For the remaining 103 economies, we rely on two primary sources of data: the Penn World Table (2002) and *World Bank Development Indicators Online* (2004) provide national accounting data for 1950–2003 for all economies in the world except Taiwan. WITSA's *Digital Planet Report* (2002; 2004) gives data on expenditures on IT equipment and software for 70 economies, including the G7. (Other important sources of data include the International Telecommunication Union, ITU, telecommunications indicators, and the UNDP *Human Development* reports.)

US data on investment in IT equipment and software, provided by the Bureau of Economic Analysis (BEA), are the most comprehensive. (The BEA data are described by Grimm, Moulton and Wasshausen, 2004). We use these data as a benchmark in estimating IT investment data for other economies. For the economies included in the *Digital Planet Report* we estimate IT investment from IT expenditures. The *Digital Planet Report* provides expenditure data for computer hardware, software, and telecommunication equipment on an annual basis, beginning in 1992.

Expenditure data from the *Digital Planet Report* are given in current US dollars. However, data are not provided separately for investment and intermediate input and for business, household, and government sectors. We find that the ratio of BEA investment to WITSA expenditure data for the United States is fairly constant for the periods 1981–90 and 1991–2001 for each type of IT equipment and software. Further, data on the global market for telecommunication equipment for 1991–2001, reported by the ITU, confirms that the ratio of investment to total expenditure for the United States is representative of the global market.

We take the ratios of IT investment to IT expenditure for the United States as an estimate of the share of investment to expenditure from the *Digital Planet Report*. We use the penetration rate of IT in each economy to extrapolate the investment levels. This extrapolation is based on the assumption that the increase in real IT investment is proportional to the increase in IT penetration.

Investment in each type of IT equipment and software is calculated as follows:

$$I_{c,A,t} = \eta_{c,A,t} * E_{c,A,t}$$

where $I_{c,A,t}$, $\eta_{c,A,t}$, and $E_{c,A,t}$ are investment, the estimated investment-to-expenditure ratio, and the *Digital Planet Report* expenditures, respectively, for asset *A* in year *t* for country *c*.

The IT expenditures for years prior to 1992 are projected by means of the following model:

$$\ln(Ec_{it-1}) = \beta_0 + \beta_1 \ln(Ec_{it}) + \beta_2 \ln(y_{it-1})$$

where $Ec_{i\,t}$ represents expenditure on IT asset *c* and the subscripts *i* and *t* indicate country *i* in year *t*, and $y_{i\,t}$ is GDP per capita. The model specifies that, for a country *i*, spending on IT asset *c* in year *t*–1 can be

projected from GDP per capita in that year and spending on asset *c* in period *t*.

Given the estimated IT investment flows, we use the perpetual inventory method to estimate IT capital stock. We assume that the geometric depreciation rate is 31.5 per cent and the service life is seven years for computer hardware, 31.5 per cent and five years for software, and 11 per cent and 11 years for telecommunication equipment. Investment in current US dollars for each asset is deflated by the US price index to obtain investment in constant US dollars.

To estimate IT investment for the 66 economies not covered by the *Digital Planet Reports*, we extrapolate the levels of IT capital stock per capita we have estimated for the 70 economies included in these *Reports*. We assume that IT capital stock per capita for the 40 additional economies is proportional to the level of IT penetration. The details are as follows:

For computers we divide the 70 economies included in the *Digital Planet Reports* into ten equal groups, based on the level of personal computer (PC) penetration in 2003. We estimate the current value s^i_{HW} of computer stock per capita in 2003 for an economy *i* as:

$$s^i_{HW} = \bar{s}^I_{HW} * (p^i_{HW} / \bar{p}^I_{HW}),$$

where $\bar{s}^I_{HW}$ is the average value of computer capital per capita in 2001 of Group I for countries included in the *Digital Planet Report*, p^i_{HW} and $\bar{p}^I_{HW}$ are the PC penetration rates of economy *i* and the average PC penetration of Group I, respectively.

For the economies with data on PC penetration for 1995, we use the growth rates of PC penetration over 1989–2003 to project the current value of computer capital stock per capita backwards. We estimate computer capital stock for each year by multiplying capital stock per capita by population. For economies lacking the data of PC penetration in 1995 and 1989, we estimate computer capital stock by assuming that the growth rates in the two periods, 1995–2003 and 1989–95, are the same as those for the group to which it belongs.

For software capital stock, we divide the 110 countries into ten categories by level of PC penetration in 2003. We subdivide each of these categories into three categories by degree of software piracy, generating 30 groups. (The information on software piracy is based on study conducted by the Business Software Alliance, 2003.) We assume that the software capital stock-to-hardware capital stock ratio is constant in each year for each of the 30 groups:

$$s^i_{SW} = \bar{s}^I_{SW} * (s^i_{HW} / \bar{s}^I_{HW})$$

where $\bar{s}^I_{SW}$ is the average software capital stock per capita of Subgroup I in 2003. Since the value of

computer stock per capita has been estimated for 1995 and 1989, this enables us to estimate the software capital stock per capita for these two years.

Finally, we define the penetration rate for telecommunications equipment as the sum of main-line and mobile telephone penetration rates. These data are available for all 110 economies in all three years – 1989, 1995 and 2003. We have divided these into ten groups by the level of telecommunications equipment penetration for each year. The current value of telecommunications capital stock per capita is estimated as:

$$s_{TLC}^{it} = \bar{s}_{TLC}^{It} * (p_{TLC}^{it} / \bar{p}_{TLC}^{It})$$

where $\bar{s}_{TLC}^{It}$ is the average current of telecommunications equipment capital stock per capita in year *t* of Group I for economies included in the *Digital Planet Reports* and p_{TLC}^{it} and $\bar{p}_{TLC}^{It}$ are the telecommunications equipment penetration rates of economy *i* and the average penetration rate of Group I in year *t*.

We employ Gross Fixed Capital Formation for each of the 103 economies provided by the Penn World Table, measured in current US dollars, as the flow of investment. We use the Penn World Table investment deflators to convert these flows into constant US dollars. The constant dollar value of capital stock is estimated by the perpetual inventory method for each of the 103 economies for 1989 and the following years. We assume a depreciation rate of seven per cent and a service life of 30 years.

The current value of the gross capital stock at a year is the product of its constant dollar value and the investment deflator for that year. We estimate the current value of Non-ICT capital stock of an economy for each year by subtracting the current value of IT stock from the current value of capital stock in that year. Given the estimates of the capital stock for each type of asset, we calculate capital input for this stock, using the methodology presented by Jorgenson (2001).

Finally, labour input is the product of hours worked and labour quality:

$$L_t = H_t * q_t$$

where L_t, H_t, and q_t, respectively, are the labour input, the hours worked, and labour quality. A labour quality index requires data on education and hours worked for each of category of workers.

We extrapolate the labour quality indexes for the G7 economies by means of the following model:

$$q_{i,t} = \beta_0 + \beta_1 \text{Education}_{i,t} + \beta_2 \text{Institution 1}_i + \beta_3 \text{Institution 2}_i + \beta_4 \text{Income 1989}_i + \beta_5 T$$

where subscripts *i* and *t* indicate economy *i* in year *t*. Education is the educational attainment of the population aged 25 or over from the data-set constructed by Barro and Lee (2001). Institution 1='Rule of Law' and Institution 2='Regulatory Quality' are constructed by Kaufmann, Kraay and Mastruzzi (2004) for the World Bank; Income 1990 is GDP per capita for 1990 from the Penn World Table and *T* is a time dummy.

Labour quality is largely explained by educational attainment, institutional quality and living conditions. The model fits well (R^2=0.973) and all the explanatory variables are statistically significant. We assume that hours worked per worker is constant at 2000 hours per year, so that growth rates of hours worked are the same as employment.

6 Summary and conclusions

World economic growth, led by the industrialized economies and Developing Asia, experienced a strong resurgence after 1995. Developing Asia accounted for an astonishing 60 per cent of world economic growth before 1995 and 40 per cent afterward, with China alone responsible for half of this, but output per capita remained well below the world average. Sub-Saharan Africa and North Africa and the Middle East languished far below the world average. Eastern Europe and the former Soviet Union lost enormous ground during the transition from socialism and have yet to recover completely.

The growth trends most apparent in the United States have counterparts throughout the world. Investment in tangible assets, including IT equipment and software, was the most important source of growth. However, Non-IT investment predominated. The contribution of labour input was next in magnitude with labour quality dominant before 1995 and hours worked afterward. Finally, productivity was the least important of the three sources of growth, except during the Asian Miracle before 1995.

The leading role of IT investment in the acceleration of growth in the G7 economies is especially pronounced in the United States, where IT is coming to dominate the contribution of capital input. The contribution of labour input predominated in the Non-G7 industrialized economies, as well as Latin America, Eastern Europe, sub-Saharan Africa, and North Africa and the Middle East. Productivity growth was the important source of growth in Developing Asia before 1995, but assumed a subordinate role after 1995. Productivity has been stagnant or declining in Latin America, Eastern Europe, sub-Saharan Africa, and North Africa and the Middle East.

All seven regions of the world economy experienced a surge in investment in IT equipment and software after 1995. The impact of IT investment on economic growth has been most striking in the G7 economies. The rush in IT investment was especially conspicuous in the United States, but jumps in the contribution of IT capital input in Canada, Japan, and the UK were only slightly lower. France, Germany and Italy also experienced a surge in IT investment, but lagged considerably behind the leaders. While IT investment followed similar patterns in the G7 economies, Non-IT investment varied considerably and explains important differences among growth rates.

Although the surge in investment in IT equipment and software is a global phenomenon, the variation in the contribution of this investment has increased considerably since 1995. Following the G7, the next most important increase was in Developing Asia, led by China. The Non-G7 industrialized economies followed Developing Asia. The role of IT investment more than doubled after 1995 in Latin America, Eastern Europe, and North Africa and the Middle East, and sub-Saharan Africa.

See Also

- production functions
- technical change
- technology

The Economic and Social Research Institute provided financial support for work on the G7 economies from its program on international collaboration through the Nomura Research Institute. Alessandra Colecchia, Mun S. Ho, Kazuyuki Motohashi, Koji Nomura, Jon Samuels, Kevin J. Stiroh, Marcel Timmer, and Bart van Ark provided valuable data. The Bureau of Economic Analysis and the Bureau of Labor Statistics assisted with data for the United States and Statistics Canada contributed the data for Canada. We are grateful to all of them but retain sole responsibility for any remaining deficiencies.

Bibliography

Ahmad, N., Schreyer, P. and Wolfl, A. 2004. ICT investment in OECD countries and its economic impact. In OECD, *The Economic Impact of ICT: Measurement, Evidence, and Implications*, Paris: OECD.

Baldwin, J.R. and Harchaoui, T.M. 2003. *Productivity Growth in Canada – 2002*. Ottawa: Statistics Canada.

Barro, R.J. and Lee, J.-W. 2001. International data on educational attainment: updates and implications. *Oxford Economic Papers* 53, 541–563.

Business Software Alliance. 2003. *Global Software Piracy Study: Trends in Piracy, 1994–2002*. Washington, DC: Business Software Alliance USA, June. Online. Available at http://www.caast.com/resources/2003_global_study.pdf, accessed 22 April 2007.

Daveri, F. 2002. The new economy in Europe: 1992–2001. *Oxford Review of Economic Policy* 18, 345–62.

Grimm, B., Moulton, B. and Wasshausen, D. 2005. Information processing equipment and software in the national accounts. In *Measuring Capital in the New Economy*, ed. C. Corrado, J. Haltiwanger and D. Sichel. Chicago: University of Chicago Press.

Heston, A., Summers, R. and Aten, B. 2002. *Penn World Table Version 6.1*. Philadelphia: Center for International Comparisons at the University of Pennsylvania (CICUP), October. Online. Available at http://pwt.econ.upenn.edu/aboutpwt.html, accessed 1 February 2007.

International Telecommunications Union. 2004. *World Telecommunications Indicators*. Geneva: International Telecommunications Union, October. Online. Available at http://www.itu.int/ITUD/ict/publications/world/world.html, accessed 1 February 2007.

Jorgenson, D.W. 2001. Information technology and the U.S. economy. *American Economic Review* 91, 1–32.

Jorgenson, D.W. 2003. Information technology and the G7 economies. *World Economics* 4(4), 139–70.

Jorgenson, D.W. 2005. Accounting for growth in the information age. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Jorgenson, D.W., Ho, M.S. and Stiroh, K.J. 2005. *Information Technology and the American Growth Resurgence*. Cambridge, MA: MIT Press.

Jorgenson, D.W. and Motohashi, K. 2005. Information technology and the Japanese economy. *Journal of the Japanese and International Economies* 19, 460–81.

Jorgenson, D.W. and Stiroh, K.J. 2000. Raising the speed limit: U.S. economic growth in the information age. *Brookings Papers on Economic Activity* 2000(1), 125–211.

Jorgenson, D.W. and Yip, E. 2000. Whatever happened to productivity growth? In *New Developments in Productivity Analysis*, ed. C.R. Hulten, E.R. Dean and M.J. Harper. Chicago: University of Chicago Press.

Kaufmann, D., Kray, A. and Mastruzzi, M. 2004. *Governance Matters III: Governance Indicators for 1996–2002*. Washington, DC: World Bank.

Krugman, P. 1994. The myth of Asia's miracle. *Foreign Affairs* 73(6), 62–78.

Lau, L.J. 1999. The sources of East Asian economic growth. In *The Political Economy of Comparative Development in the 21st Century*, ed. G. Ranis, S.-C. Hu and Y.-P. Chu. Northampton, MA: Edward Elgar.

Maddison, A. 2001. *The World Economy: A Millennial Perspective*. Paris: OECD.

Oliner, S.D. and Sichel, D.J. 2000. The resurgence of growth in the late 1990s: is information technology the story? *Journal of Economic Perspectives* 14(4), 3–22.

OECD (Organisation for Economic Co-operation and Development). 2002. *Purchasing Power Parities and Real Expenditures, 1999 Benchmark Year*. Paris: OECD.

Piatkowski, M. and van Ark, B. 2005. ICT and productivity growth in transition economies: two-phase convergence and structural reforms. TIGER Working Paper Series No. 72, Warsaw, January.

UNDP. 2004. *Human Development Report 2004*. New York: United Nations.

van Ark, B., Melka, J., Mulder, N., Timmer, M. and Ypma, G. 2003, updated 2005. *ICT Investment and Growth Accounts for the European Union, 1980–2000*. Brussels: European Commission, March.

World Bank. 2004. *World Development Indicators 2004*. Washington, DC: World Bank.

WITSA (World Information Technology and Services Alliance). 2002, 2004. *Digital Planet Report*. Washington, DC: WITSA.

Young, A. 1995. The tyranny of numbers: confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 106, 641–80.

Young, A. 2003. Gold into base metals: productivity growth in the People's Republic of China during the reform period. *Journal of Political Economy* 111, 1220–61.

How to cite this article

Jorgenson, Dale W. and Khuong Vu. "information technology and the world economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000275> doi:10.1057/9780230226203.0801

The New Palgrave Dictionary of Economics Online

Ingram, John Kells (1823 – 1907)

R.D. Collison Black

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Comte, A.; deductive method; historical method; Ingram, J. K.; positivism

Article

Ingram's whole professional career was spent at Trinity College, Dublin, of which he became a Fellow in 1846. He subsequently held a remarkable variety of offices there – Professor of Oratory (1852) and English Literature (1855), Regius Professor of Greek (1866), Librarian (1879) and Vice-Provost (1898) – but was never a professional teacher of political economy.

Nevertheless, Ingram played a notable part in the debates of the 1870s on the future of political economy and became one of the leading advocates in English of the use of the historical method in that science. Ingram's views were initially stated in his presidential address to Section F of the British Association in 1878. Here he attacked the ‘vicious abstraction’ and attachment to the deductive method of the classical economists, blaming this for the low repute into which political economy had fallen. He advocated the replacement of the deductive by the historical method and that ‘the study of the economic phenomena of society … be systematically combined with that of other aspects of social existence’. In adopting this approach Ingram was influenced partly by his contemporary T.E. Cliffe Leslie (1826–1882) but chiefly by the positivist philosophy of Auguste Comte, of whom he was an active and lifelong disciple. Of his later economic writings, the best known was, and still is, his *History of Political Economy*, which was for a long time the fullest account in English of the work of the historical school in Germany, France and Belgium. All Ingram's economic work displayed the holistic and normative outlook which he derived from Comte, but did not go far towards fulfilling the programme of historical and comparative studies to which his earlier critique of classical economics pointed.

Selected works

1878. The present position and prospects of Political Economy. *Report of the British Association for the Advancement of Science*, 641–58. Reprinted in *Essays in Economic Method*, ed. R.L. Smyth. London:

Duckworth, 1962.

1880. Work and the workman. Address to the Trade Union Congress, Dublin, September. Dublin: E. Ponsonby.

1888. *A History of Political Economy*. Edinburgh: A. & C. Black.

1895. *A History of Slavery and Serfdom*. London: A. & C. Black.

1901. *Human Nature and Morals according to Auguste Comte*. London: A. & C. Black.

How to cite this article

Black, R. D. Collison. "Ingram, John Kells (1823–1907)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000108> doi:10.1057/9780230226203.0802

The New Palgrave Dictionary of Economics Online

inheritance and bequests

Kathleen McGarry

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The importance of bequests, their role in capital accumulation, and the motivation behind these transfers has long been the subject of debate among economists. Various models of intergenerational transfers yield different predictions about the responsiveness of bequests to changes in incomes of the donors and recipients and thus to the impact public policy. Yet, despite the intuitive appeal of these models, none has proved to be consistent with empirical patterns. This article discusses the alternative theories of transfer behaviour, examines the empirical work testing their predictions, and discusses the role of estate and gift taxes in affecting bequest behaviour.

Keywords

accidental-bequest motive; altruism; annuities; bequest motive; bequests; charitable contributions; consumption smoothing; estate taxation; exchange motive; gift taxation; health insurance; inheritance taxation; inheritances; *inter vivos* transfers; intergenerational transfers; life insurance; marginal utility of consumption; National Longitudinal Surveys (NLS); pensions; savings behaviour; Social Security (USA); succession laws; tax avoidance; transfer taxation; wills

Article

Fascination with inheritances and bequests began long before economists formalized models of transfer behaviour. Literature and history are rife with examples of the role of inheritances (for example, Shakespeare's *King Lear*). Societies have laws governing bequest behaviour and governments have long employed bequest, gift and/or inheritance taxes (jointly termed ‘transfer taxes’) as a means of raising revenue. Economists, in turn, have examined the motivation behind bequests and their importance in driving economic behaviours. These transfers have been theorized to play a central role in the accumulation of wealth, the degree of inequality present in a society, and the interactions among generations. This article touches briefly upon several economic dimensions of inheritances and bequests.

Although the focus in here is on inheritances, the distinction between bequests, made at the time of death, and *inter vivos* transfers, made during life, is somewhat arbitrary. Intended bequests may be made prior to death as a means of reducing estate or inheritance taxes, avoiding other legal requirements pertaining to the settling of an estate (such as probate), or alleviating liquidity constraints and smoothing the consumption of an intended heir. Conversely, resources transferred during life could well have been saved and transferred at death in the form of a bequest. Indeed, much of the literature attempting to assess the importance of bequests in contributing to the capital stock has included the magnitude of *inter vivos* transfers along with bequests in any calculations. Similarly, economic models of the motivation for bequests are generally applicable to *inter vivos* transfers as well. Finally, in many cases, transfer taxes apply to both *inter vivos* transfers and bequests.

In this discussion I focus primarily on bequests but, where appropriate, I draw on the research examining *inter vivos* transfers as well. I use the generic term ‘transfers’ to refer to either bequests or *inter vivos* transfers. Also, for ease of exposition I occasionally refer to donors as parents and recipients as children. Obviously, bequests are frequently made to non-child heirs, but the use of this terminology makes the discussion less abstract and also accurately reflects the situation for the majority of bequests.

Much of the literature examining bequests has sought to assess the relative importance of inherited wealth and life-cycle savings as components of the existing wealth stock. Estimates of the relative importance of bequests have varied widely. Numerous researchers have put the fraction of wealth due to transfers at 15–20 per cent (see Modigliani, 1988, for a discussion), but some studies argue that the figure is much higher, concluding that transfers account for a large share of wealth holdings (for example, Kotlikoff and Summers, 1981). Although the existing estimates bracket an extremely large range, even the lower figures indicate that these transfers are an important economic phenomenon and crucial to understanding patterns of savings and life-cycle behaviour. Furthermore, inheritances and *inter vivos* transfers can potentially have substantial impacts on the well-being of the recipient, his economic behaviour, and on broader measures of the distribution of income and measures of inequality.

Intentional versus accidental bequests

The importance of bequests and their impact on macroeconomic measures such as saving rates and individual well-being depends to a great extent on the motivation driving the transfer. One school of thought argues that bequests are accidental, the result of an uncertain length of life. Individuals save to finance consumption during their retirement years and whatever wealth remains when they die is bequeathed to their heirs (Davies, 1981). Because they do not know how many years of consumption they must finance and do not want to exhaust their resources prior to death, individuals will typically die with some amount of wealth. Hurd (1987) tests the data for consistency with an accidental-bequest motive. He argues that, if bequests were intentional, one would find that individuals who had a strong bequest motive would dissave at a slower rate than those with a weak bequest motive. As a proxy for the strength of a bequest motive, Hurd uses the presence of children. He finds no difference in rates of spend-down of assets for those with and those without children, and thus concludes that there is no operative bequest motive: observed bequests are the result of an uncertain date of death.

The uncertainty in this accidental-bequest scenario need not arise solely from uncertainty about the length of life, but could stem from a variety of sources: an individual might conserve assets to guard

against expenses arising from a negative health shock, the need for long-term care, or uncertainty about returns on investment.

Additional corroborating evidence for the notion of accidental bequest comes from the failure of many individuals to specify a particular distribution of their estates. Although laws in the United States, as in much of the world, allow an individual to distribute his estate in any way he wishes through the use of a will, the use of wills to divide resources is far from universal. An individual who dies without a will is said to have died intestate. In these cases the assets of the deceased are distributed according to the laws specific to the area in which she resided. In the United States, laws differ by state but earmark a large fraction of the estate for a surviving spouse, followed by children, grandchildren and parents, with shares equally divided within kinship category. Although the reliance on succession laws to distribute assets suggests that the individual may not have thought about bequests and therefore does not have a bequest motive, it can also be argued that, if the succession laws mirror (or come close to) the distribution that the deceased would have chosen herself, a reasonable person might forgo the trouble and expense of writing a will and allow the state to divide her assets. Because wills, when they do exist, typically divide estates equally among children, as do succession laws, the failure to execute a will may indeed reflect a satisfaction with the default distribution.

The chief criticism of the notion of accidental bequests is that an individual who is concerned about uncertain future expenses could instead purchase insurance protecting against these expenditures. Insurance against outliving one's assets is available in the form of annuities which guarantee a stream of income for life and eliminate the possibility of dying with unspent wealth; instruments such as health insurance and long-term care insurance can protect against other types of unplanned expenditures. The accidental-bequest motive thus requires that these insurance markets function imperfectly.

Indeed, the potential for annuities to eliminate the possibility of an accidental bequest has been used to test for a bequest motive. If all wealth is annuitized, one has nothing to leave as a bequest. If complete annuitization is not optimal and a bequest is desired, an individual can convert a portion of his annuity income into a bequest by purchasing a life-insurance policy. Thus, life insurance can be used to offset the effects of an annuity. ‘Over-annuitization’ may not be uncommon; the prevalence of mandatory old age pensions (either public or private) suggests that many workers may retain a substantial portion of their retirement resources in annuities, perhaps more than they would choose. Bernheim (1991) examines the relationship between annuitization, in the form of US Social Security benefits, and the holdings of life insurance and private pensions. He finds that, conditional on lifetime resources, those with a greater Social Security benefits hold more life insurance and somewhat smaller private pensions. This result suggests that these tools are used to de-annuitize wealth and support the notion of an operative bequest motive. Indeed, the extensive life-insurance holdings observed in the population provide prima facie evidence that individuals are sufficiently concerned about the well-being of their heirs that they are willing to reduce own consumption.

Another argument against the accidental-bequest motive comes from the growing literature on *inter vivos* transfers. The large number of *inter vivos* transfers observed in the data are unquestionably intended and suggest that bequests might likewise be intentional.

Motivation for bequests

If individuals intentionally leave bequests, the next question is: why? What motivates an individual to

forgo consumption in order to leave assets to his heirs? Several behavioural models have been offered to address this question, but the results of empirical tests remain inconclusive.

Perhaps the most obvious explanation for the existence of bequests is that donors are altruistic; they care about the well-being of their heirs. The standard specification of the altruism model (Barro, 1974; Becker 1974) includes the heir's utility as an argument in the utility function of the donor. Formally, the utility of the donor (say, parent), U_p, is written as

$$U_p = U(C_p, U_k(C_k))$$

where C_k is the consumption of the heir (say, child). (Note that this formalization is based on very specific assumptions and thus has implications that may differ from what one might more generally regard as altruistic behaviour in other contexts; Pollak, 2003.) Consumption for *p* and *k* depend on the resources of each party prior to the bequest and on the size of the bequest. In this specific formulation the donor will make transfers until the marginal utilities are equalized across arguments of the utility function. Because the marginal utility of consumption is assumed to be decreasing, bequests will increase with the income of the donor but decline with increases in the income of the (potential) recipient. If there is more than one child, the parent will endeavour to equalize the marginal utility of consumption across children. Again, because marginal utility is decreasing in consumption, less well-off children will receive larger bequests. Thus, within a family, bequests will be compensatory and will serve to mitigate inequality.

Alternatively, transfers may be part of an exchange regime wherein the donor reimburses the recipient for specific services or behaviours. A parent compensating a child for providing home health care or simply for paying attention to the parent (Bernheim, Schleifer and Summers, 1985) would be an example of possible exchange-related transfers. In this case the donor's utility function has as its arguments her own consumption and the goods or services ‘purchased’ from the child. Formally,

$$U_p = U(C_p, S_k)$$

where S_k is a measure of services provided to the donor. The price of the services depends on the price of the recipient's time, with services purchased from high-income individuals being more costly than those purchased from low-income individuals. As the price of the good or service increases, the quantity purchased declines. In this case, then, the parent will be less likely to purchases services from a high-income child and the *probability* of a transfer will decline with the income of the child. However, conditional on purchasing services, the relationship between the transfer and the income of the recipient is indeterminate: the total amount of the transfer, price multiplied by quantity, can either rise or fall with the income of the heir, depending on the relevant elasticities.

Several other models have been discussed in the literature but have received less attention. A ‘paternalistic’ model argues that parents care not just about the utility of their children but about their actual consumption bundles. In this case a parent might bequeath money to a child through a trust specifying that it be used for certain purposes, such as schooling, or available only at certain ages when the parent believes the child's preferences will more closely mirror her own.

A ‘warm glow’ model posits that donors receive utility from the act of giving itself and not from the impact the gift has on the utility of the recipient (Becker, 1974; Behrman, Pollak and Taubman, 1982; Andreoni, 1989). Such a model might be relevant in the decision to make charitable gifts, wherein the donor is unlikely to observe the increase in utility accruing to the beneficiary as a result of the donation, yet she derives satisfaction from making the gift.

A good deal of research has attempted to discern which of the models best represents observed behaviour. The models are typically written in a static one-period framework and in such a case testing the altruism model is straightforward. Simple tests of the relationship between the probability and amount of the transfer on the one hand and the income of the potential recipient on the other should reveal a negative relationship: that is, transfers should be compensatory. However, there is a stricter test of the altruism model based on the magnitude of the response to variations in the incomes of the donor and the recipient. Specifically, the model requires that, conditional on transfers being made, an increase of one dollar in the income of the donor, accompanied by a decrease of one dollar in the income of the recipient, must be met by an increase of one dollar in the amount of the transfer (Cox, 1987). This test imposes a strict ‘adding up’ constraint on the estimated coefficients on the donor's and the recipient's income variables in a regression equation for the amount of a transfer (conditional on a positive amount). In contrast to this strict test of the altruism model, nearly any relationship between income and transfers is possible in an exchange regime. This ambiguity makes it difficult to discredit the exchange model. Not only can the relationship between the income of the recipient and the amount of the transfer go in either direction, but the components of the exchange need not be made coincidently, making it difficult to observe both sides of the transaction in data.

Observed patterns

Although *inter vivos* transfers and bequests appear to be substitutes to some extent, the two forms of giving exhibit strikingly different patterns. *Inter vivos* transfers have nearly uniformly been found to be compensatory, with more going to the less well-off children. This negative relationship between the income of the recipient and the probability and amount of a transfer is consistent with the altruism model, but is also consistent with an exchange regime wherein the donor purchases more services from lower-income heirs. Where the strict test for altruism based on the relationship of the income derivatives (that is, the magnitude of the responsiveness of transfers to changes in the incomes of the donor and recipient) has been applied, however, it has failed decisively, with estimated responsiveness closer to zero than to the value of 1 predicted by the model (Altonji, Hayashi and Kotlikoff, 1997).

Perhaps the seminal article testing for the existence of an exchange motive is Bernheim, Schleifer and Summers (1985). In that paper the authors hypothesize that parents hold bequeathable wealth and use the possibility of disinheritance to elicit desired behaviour from their children. The study finds a positive correlation between parental bequeathable wealth and the amount of attention children pay to their

parents. Recent work has questioned the empirical results (Perozek, 1998) but the notion of a parent reimbursing a child for the provision of care or other behaviour has some appeal as does the idea of an altruistic parent using bequests to compensate a less well-off child.

Although economists often shy away from directly questioning individuals about their motives, one method of attempting to discern the motivation behind the division of bequests is to ask parents about their intentions. The National Longitudinal Surveys (NLS) included such questions, explicitly asking those respondents who reported that their wills provided for unequal division of their estates why they were allocating their assets in such a way. Light and McGarry (2003) examine this question and find that motives based on altruistic concerns and those based on some sort of exchange were of nearly equal importance.

Despite the predictions of the altruism and exchange models and the compensatory transfers observed for *inter vivos* giving, examinations of both actual bequests and existing wills find that equal division among children is the norm. Some of the first work in this arena found evidence that bequests were compensatory (Tomes, 1981), but other work appeared to contradict this conclusion – for example, Menchik, (1980). More recent studies have found overwhelming evidence that estates are typically equally divided. Wilhelm (1996) uses a sample of US estate tax returns and finds that two-thirds of decedents with two or more children divided their estate exactly equally among the children and three-quarters used a division in which inheritances differed by no more than two per cent from the within-family average. Although Wilhelm's study is necessarily limited to decedents whose estates filed a tax return and who were therefore in the upper tail of the wealth distribution, similar results have been found for the general population. McGarry (1999) examines reports about existing wills for those who are still living and finds that more than 80 per cent of respondents report that their will divides their estate ‘approximately equally’ among their children.

This equal division is difficult to reconcile with either the altruism or the exchange model, both of which predict a correlation between the income of the recipient and the magnitude of the bequest. This empirical regularity has thus led several authors to propose alternative models of behaviour. Wilhelm (1996) and Bernheim and Severinov (2003) posit that unequal division is costly to parents in that they foresee that such a division could lead to unhappiness on the part of the children/intended heirs. If the difference between the utility obtained through an equal division and that obtained through an unequal allocation is greater than the utility cost (in terms of unhappy heirs) of unequal division, the parent will simply divide her estate equally among her children. McGarry (1999) provides an alternative model wherein the parent's uncertainty about the future incomes of her children lead her to resort to equal division, except in cases where large differences in the future incomes of children are expected.

Transfer taxes

Bequest and inheritance taxes have an extremely long history, dating back thousands of years, and can arouse strong feelings. Historically these taxes have been imposed as a revenue-raising mechanism, often in times of war and as a means of diluting the concentration of wealth (see Johnson and Eller, 2001, for a discussion of the history of estate taxes). In the United States the modern estate tax was implemented in 1916 to help finance the war effort (Joulfaian, 1998). Although the fraction of estates

owing a tax has varied over time, it has typically been small, hovering around two per cent. The future of the estate tax in the United States is uncertain. Under current law the tax is being gradually phased out, to be completely eliminated in 2010 but reinstated in 2011.

The form transfer taxes take varies across countries. In the United States taxes are levied on bequests and gifts, with the tax rate applied, broadly speaking, to the total value of the transfer regardless of how it is divided, although various aspects of the tax code lead to important differences in the cost of the two types of transfers (Jolfaian, 1998). Transfers to spouses and charitable organizations are exempt from tax. Not all governments have used the same approach as that employed in the United States. Many countries instead have enacted inheritance taxes wherein the tax owed depends on the amount received by an individual heir and often on the legal relationship between the decedent and the heir. These different tax bases, and the particular rules governing the evaluation of the transfers, produce varying incentives for the distribution of estates and gifts. However, the specific behavioural responses also depend on the motivation behind transfers. Although uncertainty exists about this motivation and thus about some of the predicted effects, numerous studies have shown that transfers (both *inter vivos* transfers and bequests) are responsive to tax rates (for example, Bernheim, Lemke and Scholz, 2004; Joulfaian, 2005), and there exist sizable segments of the financial and legal industries devoted to estate planning (that is, reducing estate and gift tax liabilities; see Cooper, 1979, for a fascinating look into methods for tax avoidance).

Despite these findings, and the public sentiment against the tax, several empirical studies have shown that some of the simplest tax avoidance schemes often go unexploited. For instance, in the United States *inter vivos* gifts of less than a given amount in any specific year are exempt from gift tax and can thus be used to ‘spend down’ a potentially taxable estate. Despite this opportunity, at least half of those whose estates appear likely to incur estate tax do not make such transfers (Poterba, 1998). Numerous hypotheses have been proposed to explain the failure to make ‘early bequests’, including the fear that the resources will be needed at some future date, the utility obtained from holding wealth, or the mistrust of children and their ability to manage the funds. None of these explanations appears sufficient to explain this behaviour fully.

Charitable giving

Bequests are made not just to individuals but often to charitable institutions. In the United States the tax-exempt status granted to charitable bequests reduces the price of donations to these sorts of organization relative to the price of giving to other non-spousal heirs. Numerous studies have found that the lower tax price substantially increases charitable donations. A recent study by the United States Congressional Budget Office estimates that total charitable bequests would decline by 6 to 12 per cent in the absence of the estate tax, an amount similar to the range of estimates produced by various studies over the years (U. S. Congressional Budget Office, 2004).

Behaviour of heirs

Much of the research assessing the importance of bequests in affecting economic behaviours has focused on the behaviour of the donor, the motivation for the transfer, the response to estate and gift taxes, and

the effect of desired bequests on savings behaviour. Less frequently examined is the economic response of the heirs.

From the point of view of the heir, inheritances increase financial resources and would therefore be expected to increase the consumption of normal goods, including leisure. The potential reduction in the labour supply of heirs is often cited as a motivation for a transfer tax (for example, Carnegie, 1962). Despite the theoretical implications, the several studies examining this issue have found relatively small negative effects on earnings of workers (for example, Joulfaian and Wilhelm, 1994). The small responses are not surprising in that the distribution of bequests is extremely skewed, so that for most heirs the amounts received are small relative to their lifetime incomes. Furthermore, as recent work in labour economics has demonstrated, it may be difficult to adjust hours of work on the margin. Indeed, evidence of a negative labour market effect is somewhat larger with respect to whether one participates in the labour force at all (Holtz-Eakin, Joulfaian and Rosen, 1993), suggesting that the length of the working life may be a dimension along which adjustments are more easily made.

Finally, if bequests are fully anticipated, their effect on desired hours of work ought to have been already incorporated into behaviour, and there should be no discernible response at the time the heir receives the inheritance. Thus, only unanticipated bequests or bequests received by previously liquidity-constrained heirs would be expected to spark a change in behaviour. As evidence of the potential importance of liquidity constraints, Holtz-Eakin, Joulfaian and Rosen (1993) find that inheritances can spur entrepreneurial activity.

Conclusion

Bequests play a central role in numerous economic models and as such have long attracted the attention of economists. The strength of the desire to leave bequests and the motivation behind these transfers have direct implications for such fundamental behaviours as life-cycle savings and consumption. From a public policy point of view, bequests affect the accumulation of the capital stock, the distribution of income, and the ties across generations. They also provide a source of tax revenue. Furthermore, estimates suggest that an enormous amount of wealth could be bequeathed in the coming decades, making the issues quite timely.

In attempting to understand the motivation behind bequests economists have offered several theoretical models, all of which have some intuitive appeal. Although no agreement has been reached on the most plausible theory or their relative importance, the recent availability of richer data-sets and the use of administrative records provide some hope that patterns of intergenerational transfers will be better understood. Gaining insight into the motivation behind transfer behaviour will help us to assess the potential impacts of tax policies and public transfer programmes, and to understand more completely the impact of population ageing.

See Also

- altruism, history of the concept
- estate and inheritance taxes
- intergenerational income mobility

Bibliography

Altonji, J., Hayashi, F. and Kotlikoff, L. 1997. Parental altruism and inter vivos transfers: theory and evidence. *Journal of Political Economy* 105, 1121–66.

Andreoni, J. 1989. Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy* 97, 1447–58.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1063–94.

Becker, G. 1974. A theory of social interactions. *Journal of Political Economy* 82, 1063–94.

Behrman, J., Pollak, R. and Taubman, P. 1982. Parental preferences and provision for progeny. *Journal of Political Economy* 90, 52–73.

Bernheim, B. 1991. How strong are bequest motives? Evidence based on estimates of the demand for life insurance and annuities. *Journal of Political Economy* 99, 899–927.

Bernheim, B., Lemke, R. and Scholz, J. 2004. Do estate and gift taxes affect the timing of private transfers? *Journal of Public Economics* 88, 2617–34.

Bernheim, B. and Severinov, S. 2003. Bequests as signals: an explanation for the equal division puzzle. *Journal of Political Economy* 111, 733–64.

Bernheim, B., Schleifer, A. and Summers, L. 1985. The strategic bequest motive. *Journal of Political Economy* 93, 1045–76.

Carnegie, A. 1962. The advantages of poverty. In *The Gospel of Wealth and Other Timely Essays*, ed. E. Kirkland. Cambridge, MA: Harvard University Press.

Cooper, G. 1979. *Voluntary Tax? New Perspectives on Sophisticated Estate Tax Avoidance*. Washington, DC: Brookings Institution Press.

Cox, D. 1987. Motives for private income transfers. *Journal of Political Economy* 95, 508–46.

Davies, J. 1981. Uncertain lifetime, consumption, and dissaving in retirement. *Journal of Political Economy* 89, 561–77.

Gale, W. and Perozek, M. 2001. Do estate taxes reduce savings? In *Rethinking Estate and Gift Taxation*, ed. W. Gale, J. Hines and J. Slemrod. Washington, DC: Brookings Institution Press.

Holtz-Eakin, D., Joulfaian, D. and Rosen, H. 1993. The Carnegie conjecture: some empirical evidence. *Quarterly Journal of Economics* 108, 413–35.

Hurd, M. 1987. Savings of the elderly and desired bequests. *American Economic Review* 77, 298–312.

Johnson, B. and Eller, M. 2001. *Federal Taxation of Inheritance and Wealth Transfers*. Washington, DC: Statistics of Income Division, Internal Revenue Service. Online. Available at http://www.irs.gov/pub/irs-soi/inhwlttr.pdf, accessed 7 March 2006.

Joulfaian, D. 1998. The federal estate and gift tax: description, profile of taxpayers, and economic consequences. Office of Tax Analysis Paper 80. Washington, DC: US Department of Treasury.

Joulfaian, D. 2005. Choosing between gifts and bequests: how taxes affect the timing of wealth transfers. *Journal of Public Economics* 89, 2069–91.

Joulfaian, D. and Wilhelm, M. 1994. Inheritance and labor supply. *Journal of Human Resources* 29, 1205–34.

Kopczuk, W. and Slemrod, J. 2001. The impact of the estate tax on the wealth accumulation and avoidance behavior of donors. In *Rethinking Estate and Gift Taxation*, ed. W. Gale, J. Hines and J. Slemrod. Washington, DC: Brookings Institution Press.

Kotlikoff, L. and Summers, L. 1981. The role of intergenerational transfers in aggregate capital accumulation. *Journal of Political Economy* 89, 706–32.

Light, A. and McGarry, K. 2003. Why parents play favorite: explanations for unequal bequests. *American Economic Review* 94, 1669–81.

McGarry, K. 1999. Inter vivos transfers and intended bequests. *Journal of Public Economics* 73, 321–51.

Menchik, P. 1980. Primogeniture, equal sharing, and the U.S. distribution of wealth. *Quarterly Journal of Economics* 94, 299–316.

Modigliani, F. 1988. The role of intergenerational transfers and life cycle saving in the accumulation of wealth. *Journal of Economic Perspectives* 2(2), 15–40.

Perozek, M. 1998. A reexamination of the strategic bequest motive. *Journal of Political Economy* 106, 423–45.

Pollak, R. 2003. Gary Becker's contributions to family and household economics. *Review of Economics*

of the Household 1, 111–41.

Poterba, J. 1998. Estate and gift taxes and incentives for inter vivos giving in the U.S. *Journal of Public Economics* 79, 237–64.

Tomes, N. 1981. The family, inheritance, and the intergenerational transmission of inequality. *Journal of Political Economy* 89, 928–58.

U.S. Congressional Budget Office. 2004. *The Estate Tax and Charitable Giving*. Washington, DC: Congressional Budget Office.

Wilhelm, M. 1996. Bequest behavior and the effect of heirs' earnings: testing the altrusitic model of bequests. *American Economic Review* 86, 874–92.

How to cite this article

McGarry, Kathleen. "inheritance and bequests." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000251> doi:10.1057/9780230226203.0803

The New Palgrave Dictionary of Economics Online

Innis, Harold Adams (1894 – 1952)

Ian M. Drummond

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Canada, economics in; economic history; Innis, H. A.; McLuhan, M.; natural resources

Article

Canadian economist, historian and university administrator. Born in rural Ontario, Innis was educated at McMaster University, Toronto (BA, MA), and at the University of Chicago (Ph.D.). Having served in the First World War, he joined the faculty of the Department of Political Economy, University of Toronto, in 1920. From 1937 until his death he was head of the department, and from 1947 he was also Dean of the Graduate School; at his death he was President of the American Economic Association.
A prolific and thoughtful scholar, Innis began by scrutinizing Canadian economic history, both in shorter writings and in such major works as *The Fur Trade in Canada* (1930) and *The Cod Fisheries* (1940), where he concentrated his attention on such great ‘staple products’ as codfish, fur, wheat and timber. In these works, which have been read with interest in other lands whose economic structures appear to be similar, such as Australia, Innis developed a vision of Canadian economic history that centred on the successive development of natural-resource-based industries. The physical characteristics of these industries’ products, Innis believed, had shaped not only the economic but the political and cultural history of Canada. Few would now accept Innis's interpretation of Canadian history as the mere reflection of the ‘staple products’. Yet for 40 years that interpretation shaped the teaching and writing of economy history in English-speaking Canada, and it affected political historiography as well. Innis's undergraduate education was aimed at the Baptist ministry, and perhaps it was a misfortune that he turned to economics; if he had followed some more speculative vocation, the particular powers of his intellect might have developed more widely and less eccentrically, although Canadian economic history would have been deprived of its most creative practitioner. The broadening of Innis's interests beyond economic history can be detected in his early writings on what he called ‘the penetrative power of the price system’–the ability of market mechanisms to reshape social relationships. Uncertain in his grasp of modern economics, Innis ignored the Keynesian Revolution, and he was profoundly sceptical about the

potential contribution to rational national policymaking which might come not only from economists but from other scholars; better, he thought, for university folk to concentrate upon the safeguarding of the Western cultural tradition. In his later years Innis wrote almost exclusively about very large questions – the interconnections, over very long periods, among imperial structures and means of communication. These works – *The Bias of Communications*, *Changing Concepts of Time*, *Empire and Communications*, *Minerva's Owl* – have had little impact on economists or economic historians, although they have influenced some students of the humanities – most notably the Canadian literary scholar Marshall McLuhan. Also, during the 1970s and 1980s Innis's writings attracted attention from Canadian nationalists, more or less regardless of discipline; furthermore, in these decades efforts were made to find and explicate new profundities in his writings, or to reinterpret this pessimistic and conservative thinker as an unconscious proto-Marxist. Few economists and fewer historians have found these efforts persuasive.

See Also

- linkages

Selected works

1930. *The Fur Trade in Canada*. New Haven: Yale University Press; revised edn, Toronto: University of Toronto Press, 1956.

1940. *The Cod Fisheries*. New Haven: Yale University Press; revised edn, Toronto: University of Toronto Press, 1954.

1950. *Empire and Communications*. London: Oxford University Press; revised edn, Toronto: University of Toronto Press, 1972.

1951. *The Bias of Communications*. Toronto: University of Toronto Press.

1956. *Essays in Canadian Economic History*. Toronto: University of Toronto Press.

Bibliography

Neill, R. 1972. *A New Theory of Value: The Canadian Economics of H.A. Innis*. Toronto: University of Toronto Press.

How to cite this article

Drummond, Ian M. "Innis, Harold Adams (1894–1952)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000113> doi:10.1057/9780230226203.0804

The New Palgrave Dictionary of Economics Online

input – output analysis

Wassily Leontief

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Input–output analysis is a practical extension of the classical theory of general interdependence which views the whole economy of a region, a country and even of the entire world as a single system and sets out to describe and to interpret its operation in terms of directly observable basic structural relationships. Wassily Leontief, a Russian-born American economist, started the construction of the first input-output tables of the American economy when he joined the faculty at Harvard University in 1932. These tables, for the years 1919 and 1929, were published together with the formulation of a corresponding mathematical model and numerical computation based on it in 1936 and 1937. Thus from the very outset the new methodology – for the development of which Leontief was awarded 40 years later a Nobel prize – emphasized the importance of close mutual alignment of systematic fact finding and theoretical formulation.

In the late 1920s Leontief spent three years at the Institute for the World Economy at the University of Kiel (Germany) on derivation of statistical supply and demand curves. That early experience with curve fitting taught him not to rely on indirect statistical inference as a substitute for painstaking direct factual inquiry.

With its emphasis on disaggregation permitting detailed quantitative description of the structural properties of all component parts of a given economic system, the input–output analysis moved in a direction directly opposite to that of the highly aggregative approach that began, approximately at the same time, to dominate fundamental economic research under the powerful influence of the Keynesian paradigm presented in Keynes's *General Theory*. Hand-in-hand with a disaggregated data base went an equally disaggregated theoretical model, the empirical implementation of which involved numerical computations exceeding in their complexity and scale anything that had been carried out up to that time along these lines in economics or any other social science.

The limited capabilities of the Wilbur linear analog computer used in the first large scale computation forced Leontief to scale down his problem by neglecting some of the detail contained in the disaggregated data base. Subsequent rounds of computation were carried out at first on Howard Aiken's, Mark I and Mark II computers, and later on the early electronic machines. Thirty years later the race between the economists and statisticians compiling more and more detailed factual information, and engineers constructing more and more powerful machines, was won hands down by the latter.

A standard input–output table contains square arrays of figures arranged in chess-board fashion. Each row and the corresponding column bears the name of one particular sector, say, steel industry, automobile industry, electric power utilities, advertising services, and so on. Each individual entry represents the amount (which can, of course, be zero) of the commodity or service produced by the sector – identified by the name of the row in which it appears – that has been delivered to the sector named at the head of the column in which that entry is placed. The small schematic input–output table presented below (Table 1) describes intersectoral transactions between the three sectors of the elementary economy described by it.

	Agriculture	Manufacturing	Households	Total
Agriculture	25	20	55	100 bushels
Manufacturing	14	6	30	50 yards of cloth
Households	80	180	—	260 man-years

Examining these figures, one finds that to produce one bushel of wheat, agriculture requires 0.25 bushels of wheat (seed), 0.14 tons of steel and 0.80 man years of labour. A similar set of technical coefficients – 0.40 units of agricultural and 0.12 of manufactured products – describe the input requirements for production of one yard of cloth. Listed column by column these sets of technical input coefficients represent the structural matrix at the producing part of the given economy. While the figures in Table 2 were derived from the input–output table (Table 1), estimates of the magnitudes of the technical coefficients could be, and in some instances actually are, obtained directly from technical, engineering data sources.

	Sector 1	Sector 2
Sector 1	0.25	0.40
Sector 2	0.14	0.12
Household	0.80	3.60

The structural matrix of an economy provides a basis for determination of total sectoral output as well as magnitude of inter-sectoral transactions that would enable the producing sectors to deliver to households and to other so-called final users a specified ‘bill of goods’. Considering the vector of final demand, consisting of 55 bushels of wheat and 30 yards of cloth, as given, the following set of balanced equations can be used to determine the total amounts of wheat (x_1), of cloth (x_2), as well as the total amount of labour (L) needed to balance under these particular technological conditions the outputs and inputs of both producing sectors,

$$(1-0.25)x_1 - 0.14x_2 = y_1 - 0.40x_1 + (1-0.12)x_2 = y_2 \quad (1)$$

The general solution of these two equations:

$$1.457y_1 + 0.662y_2 = x_1 0.232y_1 + 1.242y_2 = x_2$$
(2)

permits us to compute the total levels of output of wheat, x_1 and cloth, x_2 required directly and indirectly to satisfy any given vector (y_1, y_2) of 'final demand'.

An increase in the final deliveries of agricultural products, y_1 by one unit would for instance require a rise of total agricultural output, x_1, by 1.1457 units, 0.1457 of which will have to be used to satisfy the additional input requirements of the agricultural and manufacturing sectors.

Formulated in short-hand matrix notation, the balance equations (1), describing the relationship between the column vector of final demand, y, and the column vector, x, of total outputs of all producing sectors can be written as:

$$(I - A)x = y$$
(3)

where A represents the upper, square part of the structural matrix (Fig. 2) describing the material input requirements of all producing sectors, x is the column vector of total outputs and y, the column vector of final deliveries of both goods. The general solution of that linear equation is,

$$x = (I - A)^{-1}y$$
(4)

where $(I - A)^{\bullet-\bullet 1}$ represents the so-called inverse of matrix $(I - A)$.

Total labour requirement can be computed in a separate step,

$$L = l'x = l'(I - A)^{-1}y$$
(5)

where l' is a row vector of technical labour coefficient representing the technologically determined amounts of labour that each industry employs per unit of its total output.

The same set, A, of structural coefficients that controls the physical flows, determines also the

relationship between the prices of goods and services produced by different industries and the ‘value added’ payments (expressed in the monetary units) made by each industry per unit of its output. These include wages, profits, taxes, etc. In short, all payments other than those made for goods and services purchased from other producing sectors.

This set of value added–price equations, (often referred to as a ‘dual’ to set (3) of physical input–output relationships) can be formulated as follows,

$$(I - A')P = V \tag{6}$$

and its solution for the unknown prices as,

$$P = (I - A')^{-1}V \tag{7}$$

where *P* is the column vector of prices of all sectoral outputs and *V* is the given column vector of values added (per unit of their respective outputs), in different sectors.

In the schematic input–output table (Fig. 1) considered above all amounts entered along a particular row are measured in the same appropriately selected physical unit, for instance, wheat – in bushels; cloth – in yards; labour – in man years. No column totals are entered, since adding amounts measured in incomparable physical units would make no sense. In most published input–output tables, all transactions are measured however in value terms – usually in ‘base year’ prices. Since these are assumed to satisfy the price-value added equations described above – each column total, including the value added per unit of total output, must naturally be equal to the total output figures entered at the end of the corresponding row.

Value figures entered along a particular row can, however, also be interpreted as representing physical amounts of the good in question, provided the physical unit in which they are measured is implicitly defined as the quantity of that good purchasable for, say, one dollar.

In the case of a table, some rows of which are presented in conventional physical amounts, say kwh of electric power, or tons of copper, while some other rows are presented in monetary units, appropriate ‘equilibrium prices’ can be computed through solution of the corresponding ‘dual’ equation (7). To do so it would suffice to re-define the physical unit of the products of each sector as the amount purchasable for, say, one dollar, or some other monetary unit, at the price actually used in determination of the value figures entered on the base year table. These prices might of course be different from the equilibrium prices.

From the outset the development of input–output analysis was marked by a succession of empirical applications. In Leontief's early volume, *The Structure of American Economy, 1919–1929* (1941), this

was the computation of the effects of changes in the input structure of different industries on levels of output and prices of their products, and in particular on the ‘standard of living’ of households.

With the onset of Second World War, attention was centred on the transition from peacetime to a war economy: in particular, on the effects of changes in the level and composition of final demand on the intersectoral distribution of output and employment. The first official US input–output table – for the year 1939, compiled for the US Bureau of Labor Statistics – provided a basis for preparation of a detailed multisectoral projection of postwar production and employment levels. Correctly predicting serious steel shortages, instead of large surpluses anticipated by leading economic and industry experts, this report gained wider interest in the new approach, not only in government circles, but among large industrial corporations as well. The Western Electric Company (the manufacturing arm of A.T.&T.) having successfully employed input-output analysis to anticipate impending shortages of lead, one of its principal raw materials, even produced an educational film describing the methodology used.

In one of the early applications of the same modelling technique as that which later on became known as operations research, the small input–output team organized – under the name Project Scoop – by the US Air Force constructed a detailed structural matrix of its far-flung material procurement and training operations. It was not a square, but rather a rectangular matrix showing for some sectors not one but several input vectors corresponding to two or more alternative technologies that could be used to produce a particular weapon or to provide a particular type of pilot training. Confronted with the problem of optimal choice between alternative ‘cooking recipes’, Dr George Dantzig, a young mathematician on the Project's staff, invented the still very widely used Simplex method of linear programming, which consists of a series of inversion of structural input–output matrices with sequential substitution at alternative vectors of technical coefficients.

Not unlike research conducted in modern natural sciences, input–output analysis was from the outset most successfully conducted by closely coordinated teams rather than individual investigators. The first of such academic research groups was the Harvard Economic Research Project directed by Leontief over a period of nearly thirty years. Another centre was organized by Richard Stone in the Department of Applied Economics at the University of Cambridge. He was responsible for formal incorporation of input–output tables in the United Nations system of national accounts designed by him.

Many of the young foreign economists who came to the United States to complete or postgraduate studies spent from a few months up to several years at the HERP, and after returning home introduced input–output analysis not only as a subject of academic instruction and research but also as a new field of governmental statistics.

In Norway, Canada, Japan and in many other countries governmental planning agencies and central statistical offices compile national input–output tables and carry out practical applications of input–output analysis, but also engage in fundamental methodological research. In Soviet Russia this was the first non-marxist, mathematical approach to economics adapted, on the recommendation of Oscar Lange, after World War II as a subject of academic instruction and as a tool of economic planning.

The first International Conference on Input–Output Analysis organized by Professor Tinbergen was held in Dreibergen, Holland in 1950; the eighth has been held in Japan in 1986. Proceedings of these and of other similar scientific meetings published in book form provide a good account of the current state of the art in the general field of input–output analysis and its various applications.

One of the fundamental theoretical questions that came up in connection with the early input–output computations concerned the conditions under which none of the elements of the inverse $(I - A)^{-1}$ can

be negative. The answer to it was provided by Herbert Simon – the future Nobel prizewinner – and David Hawkins, a philosopher, in the form of the following theorem:

The necessary and sufficient conditions for some of the elements of $(I - A)^{-1}$ to be positive, and all to be non-negative, are:

$$|1-\alpha_{11}| > 0, \begin{vmatrix} (1-\alpha_{11}) & -\alpha_{12} \\ -\alpha_{21} & (1-\alpha_{22}) \end{vmatrix} > 0, \ldots \begin{vmatrix} (1-\alpha_{11}) & -\alpha_{12}\ldots & -\alpha_{1n} \\ -\alpha_{21} & (1-\alpha_{22})\ldots & -\alpha_{2n} \\ -\alpha_{n1} & -\alpha_{n2} & \ldots(1-\alpha_{nn}) \end{vmatrix} > 0 \quad (8)$$

If these conditions are satisfied for any particular numbering of sectors it will necessarily be satisfied for any other numbering sequence too. The economic interpretation of this theorem is that for a system, in which each sector functions by absorbing directly or indirectly outputs of some other sectors, to be able not only to sustain itself but also to make some positive deliveries to final demand, each one of the smaller and smaller sub-systems contained within it has to be capable of sustaining itself and yielding a surplus deliverable to outside users as well.

An example of a system unable to sustain itself in this sense could be an economy so badly damaged by some natural catastrophe or war that only external assistance, taking the form of an import surplus, could prevent it from complete collapse. Exports are entered in a standard input-output table and in the corresponding set of balance equations, as positive and exports as negative components of the final bill of goods. The negative elements of the inverse $(I - A)^{-1}$ multiplied into such negative components of the vector y of final demand would yield in this case positive total outputs x.

In an attempt to reconcile at least to some extent the so-called fixed coefficient assumption of linear input–output models with the neoclassical production functions allowing for input substitution, Kenneth Arrow, Tjalling Koopmans and Paul Samuelson provided independently from each other three different proofs of the ‘non-substitution theorem’. They considered a multisectoral economy in which each productive sector operates on the basis of a neoclassical production function and all sectors use the same single primary factors of production, say labour. The input combinations used by different sectors are chosen so as to minimize the total amount of labour that has to be employed by that economy in order to enable it to deliver to final users an exogenously specified bill of goods. The non-substitution theorem states that the combination of the relative amounts of different inputs chosen in each sector will be independent of the composition of the final bill of goods. That means that even if the structure of final demand changes all producing sectors will behave as if they were operating on the basis of fixed coefficients of production.

Restrictive assumptions – particularly those postulating invariability of production functions that control the operations of all sectors – deprive the non-substitution theorem of much of its practical significance. However, it calls attention to the difference between the ways in which the terms technology, and technological change, are used in neoclassical and in input-output theory. In input–output modelling the technology used in any particular sector is described as a given column vector of coefficients, and a

change in any element of that vector is called technological change. In neoclassical modelling the state of the technology employed by a particular sector is described by a much more general – and because of that much more complex – kind of functional relationship that in input–output analysis would have to be viewed as a set of many (strictly speaking, infinitely many) different technologies, each described by a different column vector of input coefficients. While providing a convenient basis for deductive reasoning, the neoclassical terminology makes the task of actual observation of the technological structure of a particular economy and empirical description of processes of technological change extremely, not to say prohibitively, difficult.

Since direct observation of a set of isoquants is hardly ever possible, empirical implementation of standard neoclassical models involves nearly exclusive reliance on more and more sophisticated methods of indirect statistical inference.

Neither of the two definitions of technology and technological change can be said to be more correct than the other. The employment of the simpler definition however permitted input–output analysis to advance in the direction of systematic detailed factual inquiry, while reliance on a definition, much less serviceable for purposes of empirical description but much richer in its theoretical implications, propelled neoclassical economics towards construction of elaborate theoretical models erected on a narrow, fragile data base or even on quite arbitrary, purely theoretical assumptions.

In static input–output models, additions to the stocks of building, machinery, and other kinds of productive stocks are treated as a component part of the final demand vector, entered in the right-hand side of the balance equation (6). In the following formulation of a simple dynamic model these terms are transferred to its left-hand sides and described explicitly as serving technologically determined capacity expansion required for a rise in the level of output.

$$(I - A)X_t - B(X_{t+1} - X_t) = Y_t \tag{9}$$

B is a square matrix of technical capital coefficients, each column of which consists of stock-flow ratios, describing the stocks of products of different industries which the sector in question must have on hand per unit of its capacity output.

If the time unit in terms of which the process is observed and described is relatively long – say, covering a five or even ten year period – the stocks might be engaged in production in the same time period during which they have been produced. In this case, the second term on the left-hand side would be $B(X_t - X_{t-1})$. Current inputs required for maintenance of the existing capital stock have of course to be accounted for by the appropriate elements of the A matrix.

While bringing to the fore the crucial role that a complete set of capital coefficients has to play – in addition to a complete set of current input coefficients – in the detailed description of the structural framework of a given economy, such a set of difference equations is too rigid a tool to be used to describe and project the actual process of economic development and change.

More effective, because more flexible, is an approach which takes the form of a step-by-step construction of complete input–output tables of the economy for successive periods of time, each based

on the knowledge of its state in the previous period, of anticipated changes in the final bill of goods and expected technological changes.

In more general terms, the input–output relationship between goods produced and consumed over a sequence of successive years can be formally described exactly in the same terms as relationships between different sectors are presented in an ordinary ‘static’ input-output table for a single year. The solution of a time-phased system of linear equations describing the intertemporal balances of inputs and outputs of goods and services produced and consumed over a long stretch of successive periods of time can be interpreted as inversion of a large triangular matrix; triangular because outputs of one year can become inputs in later years, but not vice versa. The results of this operation describing the direct and indirect relationships between all appropriately timed inputs and outputs has been called the ‘dynamic inverse’. Since the sets of flow and capital coefficients controlling the input–output balances in successive stretches of such an historical process do not have to remain the same, both that dynamic matrix and its inverse can accurately represent all kinds of structural change, including elimination of old and introduction of entirely new goods.

Introduction of capital coefficients permits subdivision of the value-added term, *V*, on the right-hand side of the dual system (8) into its two parts – the returns on capital and wage income:

$$(I - A')P = \lambda B' cP + 1w \tag{10}$$

or, solving for *P*:

$$(I - A')P - \lambda B' P = 1w$$

λ represents the rate of return on invested capital and *w*, the wage rate. These equations can be used for calculating the ‘trade-off curve’ between real wages (that is, money wage rate divided by a price index) and the rate of return on capital for any given state of technology. Comparison of such curves, each reflecting a different combination of alternative technologies available in different sectors, provides a base for numerical assessment of the influence of the distribution of income between the return on capital and wages upon technological choice.

Practical concerns led quite early to construction of regional input–output tables. The municipal government of the city of Stockholm was the first to compile a detailed metropolitan table. The complex fact-finding task of putting together a detailed input–output map of a particular region seemed to have been inspired sometimes by the desire to assert distinct identity. In Canada, French-speaking economists were the first to construct a regional table, that of Quebec. In Belgium one was compiled for the autonomy-seeking Flemish provinces. In addition to pressing needs of developmental planning, similar considerations seem to have prompted early compilation of input–output tables of many less developed

countries.

The next step was construction of multiregional input–output tables and models in which intraregional transactions were linked with each other by interregional flows of goods and services. While comparison of labour, capital and natural resource ‘contents’ was the object of some of the earliest input-output studies of domestic and internationally traded goods, neither the theoretical formulation nor the available data base are yet sufficiently advanced to permit input–output modelling of international economic transactional trade to be solidly based on direct empirical implementation of the comparative cost theory. In most multiregional input–output models the structure of international transactions is controlled by sets of empirically determined export and import coefficients. A large multiregional input–output model of the world economy constructed under the auspices of the United Nations was published in 1977. Originally intended to provide a basis for a set of alternative projections of the future growth of eight groups of developed and seven groups of less developed countries, this large, highly disaggregated model was used in a series of other studies such as the analysis of economic effects of international arms trade, detailed long-run projections of the production and consumption of non-ferrous metals in the United States and construction of alternative multiregional scenarios of future exploration of agricultural and energy resources.

As the range of its practical applications widened, the scope of input–output modelling had to be broadened, along with the contents of the requisite data bases.

Analysis of the petroleum refining industry in the early Fifties required modelling of multiproduct processes. Thirty years later a similar approach was employed to describe within the framework of a national input–output table the generation and elimination of various polluting substances. Modelling devices adapted in description of the allocation of the output of transportation and trade sectors have later on been adapted in modelling the activities of all service industries. Separation of the description of the physical from the price and costing aspects of government operations proved to be useful in construction and theoretical interpretation of input–output tables of simple, not yet fully monetized economies of the less developed economies. Richard Stone offered the conceptual framework of input-output analysis for the formal description of demographic processes.

To the extent to which it can provide a bridge between aggregative analysis and detailed description of production and consumption of specific goods and services, input–output analysis has been incorporated into most of the well-known forecasting econometric models.

The general nature of the approach has made the development of input–output analysis a cumulative process. Each refinement in theoretical structure and each addition to or improvement in the accuracy of factual information incorporated in its data base potentially improved the performance of the general model in application to all special problems.

See Also

- Hawkins–Simon conditions
- Leontief paradox

Bibliography

Brody, A. 1970. *Proportions, Prices and Planning: A Mathematical Restatement of the Labor Theory of Value*. Amsterdam: North-Holland.

Brody, A. and Carter, A.P., eds. 1970. *Applications of Input-Output Analysis*. Proceedings of the Fourth International Conference on Input-Output Techniques, Geneva, 8–12 January, 1968, Vol. 2, Amsterdam: North-Holland.

Brody, A. and Carter, A.P., eds. 1972. *Input-Output Techniques*. Proceedings of the Fifth International Conference on Input-Output Techniques, Geneva, January, 1971. Amsterdam: North-Holland.

Bulmer-Thomas, V. 1982. *Input-Output Analysis in Developing Countries: Sources, Methods and Applications*. New York: John Wiley & Sons.

Carter, A.P. 1970. *Structural Change in the American Economy*. Cambridge, Mass.: Harvard University Press.

Leontief, W. 1941. *The Structure of American Economy, 1919–1939: An Empirical Application of Equilibrium Analysis*. 2nd edn, enlarged, White Plains, NY: International Arts and Sciences Press, 1951.

Leontief, W. et al. 1953. *Studies in the Structure of the American Economy: Theoretical and Empirical Explorations in Input-Output Analysis*. White Plains, NY: International Arts and Sciences Press.

Leontief, W. 1966. *Input-Output Economics*. 2nd edn, New York: Oxford University Press, 1986.

Leontief, W., Carter, A.P. and Petri, P.A. 1977. *The Future of the World Economy*. A United Nations Study. New York: Oxford University Press.

Leontief, W. and Duchin, F. 1985. *The Future Impact of Automation on Workers*. New York: Oxford University Press.

Meyer, U. 1980. *Dynamische Input-Output-Modelle*. Königstein: Athenaum Ökonomie Verlag.

Miller, R.E. and Blair, P.D. 1985. *Input-Output Analysis: Foundations and Extensions*. Englewood Cliffs, NJ: Prentice-Hall.

Polenske, R. and Skolka, Ji•i V., eds. 1976. *Advances in Input-Output Analysis*. Proceedings of the Sixth International Conference on Input-Output Techniques, Vienna, 22–26 April, 1974, Cambridge, Mass.: Ballinger.

Schumann, J. 1968. *Input-Output-Analyse*. Berlin and Heidelberg: Springer-Verlag.

Smyshlyaev, A., ed. 1983. *Proceedings of the Fourth IIASA Task Force Meeting On Input-Output Modeling 29 September – 1 October 1983*. Laxenburg, Austria: International Institute for Applied Systems Analysis.

How to cite this article

Leontief, Wassily. "input–output analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000118> doi:10.1057/9780230226203.0805

The New Palgrave Dictionary of Economics Online

inside and outside money

Ricardo Lagos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A distinction is drawn between outside money, which is either of a fiat nature or backed by some asset that is not in zero net supply within the private sector, and inside money, which is an asset backed by any form of private credit that circulates as a medium of exchange.

Keywords

banking theory; bonds; commitment; fiat money; finance theory; inside and outside money; money; open market operations; Patinkin, D.; private credit

Article

Money is an asset that serves as a medium of exchange.
Outside money is money that is either of a fiat nature (unbacked) or backed by some asset that is not in zero net supply within the private sector of the economy. Thus, outside money is a net asset for the private sector. The qualifier ‘outside’ is short for ‘(coming from) outside the private sector’.
Inside money is an asset representing, or backed by, any form of private credit that circulates as a medium of exchange. Since it is one private agent's liability and at the same time some other agent's asset, inside money is in zero net supply within the private sector. The qualifier ‘inside’ is short for ‘(backed by debt from) inside the private sector’.

Background

In 1960, John G. Gurley and Edward S. Shaw published *Money in a Theory of Finance*, in which they attempted to develop a theory of finance that encompasses the theory of money and a theory of financial institutions that includes banking theory.
Consider a simple economy similar to the one considered by Gurley and Shaw. The economy has fiat

money – an intrinsically useless asset with no backing whatsoever – that is generally accepted as a means of payment. A monetary authority or ‘government’ has the monopoly over issuing this asset. The economy is closed and consists of three sectors: households, firms and government. Firms issue debt in the form of homogeneous, perfectly safe nominal bonds. (For example, think of these bonds as being promises to pay one dollar at some future date.)

Table 1 shows hypothetical sectoral balance sheets for this economy. In this example, households hold only financial wealth (that is, no real wealth such as houses), in particular money, equity in firms, and the bonds issued by the firms. Here households have no liabilities, so their net worth (NW) is just the sum of the value of their assets. The assets owned by firms consist of cash and physical capital. A part of these assets has been financed with debt (bonds), and another part by issuing equity. The former represent the firms’ liabilities toward the bond holders, and the latter represent the firms’ liabilities towards share holders. The firms’ net worth (net of equity) is zero. The government has no real assets, but at some point in the past it issued financial assets – money – to pay for expenditures, and from an accounting point of view these outstanding government-issued pieces of paper constitute liabilities. (If the money was backed by a real asset, for example gold, and also fully convertible, then the value of the gold would show up on the government's Assets column. In this case, the money issued is literally a liability representing the government's commitment to redeem the money for gold. In the case of fiat money, there need not be a counterpart on the Assets column of the government's balance sheet.)

Households		Firms		Government	
Assets	Liabilities	Assets	Liabilities	Assets	Liabilities
Money 50		Money 100	Bonds 25		Money 150
Bonds 25		Capital 200	Equity 275		
Equity 275					
NW 350		NW 0	NW –150		

Table 2 shows what happens if we consolidate the balance sheets of the private sector. The bonds are debts from private agents (in this example the firms) to other private agents (in this example the households), so they have cancelled out. The only assets left in the balance sheet of the public sector are physical capital and the money issued by the government. Money can be thought of as a ‘claim’ held by consumers and firms against the government. From the standpoint of the private sector, it is a net external, or outside, claim: it is *outside money*.

Combined private sector			Government	
Assets		Liabilities	Assets	Liabilities
Money	150			Money 150
Capital	200			
NW	350	NW –150		

Gurley and Shaw (1960) were interested in considering the effects of ‘open market operations’ whereby the government issues money to purchase private bonds. Suppose, for example, that they purchase $15 worth of private bonds. The resulting balance sheets are those in Table 3, which should be compared with those in Table 1. The government now has $15 worth of assets (the private bonds it purchased), and

its liabilities have increased by $15 because of the money issued to pay for these bonds. Households still hold $350 worth of assets, but the composition of their portfolio has changed: they now hold $65 in money and $10 in bonds, as opposed to the $50 in money and $25 in bonds of Table 1. The additional $15 in money holdings comes from the new issue of money, backed by private bonds. These $15 are government debt, but they are issued in payment for government purchases of private securities. They are a claim of consumers and firms against the world outside the private sector, but they are counterbalanced by private debt to the world outside, that is, to the government. These additional cash balances are based on internal debt, so Gurley and Shaw referred to these $15 as *inside money*.

Households		Firms		Government	
Assets	Liabilities	Assets	Liabilities	Assets	Liabilities
Money 65		Money 100	Bonds 25	Bonds 15	Money 165
Bonds 10		Capital 200	Equity 275		
Equity 275					
NW 350	NW 0		NW –150		

To use the terminology of Gurley and Shaw, the $165 stock of money in the economy of Table 3 consists of $150 of outside money and $15 of inside money. Both types of money are really the same physical object, for example, green pieces of paper: The qualifiers *inside* and *outside* refer to the asset counterpart of the money. Inside money is backed by private domestic debt. Outside money is of a fiat nature (or backed by some other asset that is not in zero net supply within the private sector, such as gold). Note that, if we consolidate the balance sheets of the private sector in Table 3, the net worth of the private sector is still $350, just as in Table 2. Also, note that inside money is ‘endogenous’ in that if, for example, firms pay off their whole debt, *ceteris paribus* the money supply would shrink by $15. Most likely, Gurley and Shaw were led to stress the distinction between inside and outside money because they viewed money and private debt as assets that played distinct roles in exchange, so that an economy with the balance sheets of Table 1, where households hold $50 in cash and $25 in private bonds, would function differently from an economy with the balance sheets of Table 3, where households hold $65 in cash and $10 in private bonds. (See Gurley and Shaw, 1960, pp. 82–8, the section titled ‘Monetary Policy in a Modified Second Model’.) The theoretical analysis throughout the book is predominantly verbal, so it is not clear which are the precise trade-offs that agents consider when making a portfolio decision between money and bonds. The fact that households treat them as different assets is explicit in the Mathematical Appendix, where Alain C. Enthoven assumes distinct reduced-form demand functions for the two financial assets. Note that, since bonds are nominal and riskless in this set-up, it is not obvious why households would not treat them as perfect substitutes for money.)

The contemporary literature on monetary theory in general, and the subfield that deals with inside and outside money in particular, does not take it as given that money and bonds play different roles. Instead, it seeks to understand whether they indeed do, and whether they ought to. The recent emphasis has been on trying to gain a deeper understanding of the precise roles that fiat money and private debt play and ought to play, both as media of exchange and as vehicles to channel resources across economic agents, towards their most efficient use. This change of emphasis has led to a slightly different definition of inside money. The more modern use of the concept does not rely on the type of open market operations

of Gurley and Shaw. Inside money need not be defined narrowly as circulating fiat money backed by private debt; the private debt *itself* is regarded as inside money if it circulates as means of payment among the private agents. The more modern definition given at the beginning of this article encompasses both the case where private debt circulates directly and Gurley and Shaw's original example. To illustrate, consider again the economy of Table 1. According to the modern use of the term, there is not enough information in that table to decide how much inside money there is in the economy; there are $25 of inside assets, that is, assets that are in zero net supply within the private sector, but whether these assets constitute inside money depends on whether they circulate as means of payment. If they do not – for example if lenders merely hold the bonds until maturity to redeem them – then these bonds are not inside money.

Contemporary perspectives

Gurley and Shaw (1960) simply asserted that agents would want to hold government-issued fiat money (this weakness was stressed by Patinkin, 1961), and for their purposes the distinction between inside and outside money was relevant because they implicitly regarded them as imperfect substitutes. The modern literature on monetary theory seeks to identify the fundamental features of the basic economic environment that can make fiat money, or, more generally, any asset that serves as a medium of exchange, valuable and socially beneficial. Modern theory also focuses on the differences and similarities between inside and outside money. When is outside money valued? Under which circumstances does inside money arise? Are inside and outside money substitutes or complements? Under which circumstances can they coexist? Are they *both* needed to achieve efficient outcomes?
Inside money is private debt that also circulates as a tangible medium of exchange. Thus, an economy with inside money must perform a delicate balancing act. On the one hand, it must have enough commitment or enforcement for credit to be feasible, but at the same time credit must not function too well, for otherwise a tangible medium of exchange would be inessential. For example, Kocherlakota (1998) shows that a tangible medium of exchange is not essential if agents can commit to future actions or if their trading histories are public. Starting from this observation, Cavalcanti and Wallace (1999a) consider an environment where trading histories are public for a subset of agents but private for the rest, and show that a social optimum requires note issue by those agents with public trading histories. In addition, those notes are in turn used in trade among the agents whose trading histories are private. Thus, in their environment an optimum requires inside money.
Kiyotaki and Moore (2002a) instead consider an environment where everyone is anonymous, and emphasize the importance of the agents’ ability to make bilateral and multilateral commitments. The degree of (bilateral) commitment a borrower can make to an initial lender when selling a paper claim places a bound on the entire stock of private debt. The degree of (multilateral) commitment a borrower can make to repay any bearer determines the extent to which the borrower's debt can circulate in equilibrium. Kiyotaki and Moore find that only outside money circulates in economies with very low degrees of bilateral commitment. For higher, but still low, degrees of bilateral commitment, outside and inside money circulate alongside each other in equilibrium. For yet higher degrees, only inside money circulates, and, when the agents’ ability to make bilateral commitments is large enough, the economy can manage without any money, inside or outside.

I thank Narayana Kocherlakota and Warren Weber for comments. I also thank the C.V. Starr Center for Applied Economics at New York University for financial support. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

Bibliography

Cavalcanti, R. de O., Erosa, A. and Temzelides, T. 1999. Private money and reserve management in a random-matching model. *Journal of Political Economy* 107, 929–45.

Cavalcanti, R. de, O. and Wallace, N. 1999a. A model of private bank-note issue. *Review of Economic Dynamics* 2, 104–36.

Cavalcanti, R. de O. and Wallace, N. 1999b. Inside and outside money as alternative media of exchange. *Journal of Money, Credit, and Banking* 31, 443–57.

Gurley, J. and Shaw, E. 1960. *Money in a Theory of Finance*. Washington, DC: Brookings Institution.

Kiyotaki, N. and Moore, J. 2002a. Inside money and liquidity. Mimeo, London School of Economics.

Kiyotaki, N. and Moore, J. 2002b. Evil is the root of all money. *American Economic Review* 92(2), 62–6.

Kiyotaki, N. and Moore, J. 2005. Financial deepening. *Journal of the European Economic Association* 3, 701–13.

Kocherlakota, N. 1998. Money is memory. *Journal of Economic Theory* 81, 232–51.

Patinkin, D. 1961. Financial intermediaries and the logical structure of monetary theory: a review article. *American Economic Review* 51(1), 95–116.

Wallace, N. 2000. Knowledge of individual histories and optimal payment arrangements. *Federal Reserve Bank of Minneapolis Quarterly Review* 24(3), 11–21.

How to cite this article

Lagos, Ricardo. "inside and outside money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000119> doi:10.1057/9780230226203.0806

The New Palgrave Dictionary of Economics Online

insider trading

Andrew Metrick
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Insider trading has two definitions: securities trading by a corporate insider, and securities trading while in the possession of material non-public information about the security. This article reviews the two main strands of economic literature on insider trading. First, scholars on the intersection of law and economics analyse the social-welfare implications of insider-trading regulation. Second, financial economists use empirical evidence on insider trading to analyse the efficiency of stock markets.

Keywords

adverse selection; capital asset pricing model; cost of capital; efficient markets hypothesis; informational efficiency; insider trading; law, economic analysis of; market liquidity; non-public information; regulation costs; risk-adjusted returns; Securities and Exchange Act 1934 (USA); Securities and Exchange Commission (USA); securities regulation; short selling; stock markets; tipping chains

Article

Federal securities law defines a ‘corporate insider’ to be an officer, director, or major shareholder of a corporation. The first definition of ‘insider trading’ refers to any purchase or sale of public-corporation stock by an insider of that corporation. The second definition does not require the trader to be a corporate insider, but does require that the trader possess material non-public information. Within this article, all uses of the generic term ‘insider trading’ encompass both of these definitions. When necessary, the two definitions are referred distinctly as ‘trading by insiders’ (any transaction that is made by a corporate insider) and ‘trading on inside information’ (a transaction that requires material non-public information but need not be made by a corporate insider).

In the United States, prior to 1934 insider trading was regulated by state-level corporate law. In the first few decades of the 20th century, states used a variety of criteria to adjudicate cases, with a substantial minority of states holding that corporate directors had a duty to disclose material information before

buying (but not selling) stock. Federal regulation of insider trading did not begin until the Securities and Exchange Act (SEA) of 1934. Rule 10b of the SEA made it unlawful for any person ‘to use or employ, in connection with the purchase or sale of any security registered on a national securities exchange or any security not so registered, any manipulative device or contrivance in contravention of such rules and regulations as the Commission may prescribe …’ (Bainbridge, 2001). This sweeping language does not directly mention either corporate insiders or material non-public information, but later judicial interpretations expanded its scope to these cases. Corporate insiders do appear in Section 16a of the SEA, which requires that open-market trades by insiders be reported to the Securities and Exchange Commission (SEC) within ten days after the end of month in which they took place. These reports, filed on the SEC's ‘Form 4’, are the source of data for almost all of the empirical studies of trading by insiders. It was not until 1961 that the SEC took its first administrative action on an insider-trading case (Cady, Roberts, & Co.), and it would be another seven years before the first federal insider-trading case was decided by the courts (*SEC* v. *Texas Gulf Sulphur Co*. (1968)). In the decades since these seminal cases, the courts have reaffirmed and expanded the SEC's role in the regulation of trading on inside information. Despite the long judicial record, there is still considerable confusion and a continuing evolution about the scope of regulation, with debates about the type of information that is considered to be ‘non-public’ or ‘material’, and about the necessity of the trader having some fiduciary relationship to the company. A discussion of these issues is beyond the scope of this survey; readers are referred to Bainbridge (2001) for a summary.

The United States was the first country to use securities regulation to prohibit trading on inside information. Other countries were slow to adopt similar regulations: Bhattacharya and Daouk (2002) report that, as of 1990, of 103 countries with stock markets, only 34 had any prohibitions on insider trading, and only 9 had enforced their prohibitions with a prosecution. The same paper, however, reports that, by 1998, 87 countries had prohibitions and 38 had made at least one prosecution.

The remainder of this article reviews the two main strands of economic literature on insider trading. First, scholars on the intersection of law and economics analyse the social-welfare implications of insider-trading regulation. Second, financial economists use empirical evidence of trading by insiders to analyse the efficiency of stock markets. Each of these two topics has developed an extensive literature since the 1960s.

Social welfare

Prior to the 1960s, scholars gave little thought to the social-welfare implications of insider-trading regulation. With the Cady case of 1961, the first federal regulation in the United States stimulated a large literature on the topic, beginning with Manne (1966). The economic debate revolves around six main issues: market liquidity, informational efficiency, market manipulation, efficient managerial compensation, the costs of regulation, and the necessity of federal law.

Market liquidity. The pro-regulation side argues that, if trading on inside information were pervasive, then non-insiders would be discouraged from trading, thus reducing market liquidity and all of the other good things that come from having well-functioning capital markets. The logic here is straightforward: if non-insiders perceive that counterparties are likely to possess inside information, then they face an adverse-selection problem, and will demand a discount (if buying) or premium (if selling). The resultant

spread between bid and ask prices would then act effectively as a tax on every transaction, which lowers the amount of trade.

In response to this argument, the anti-regulation side argues that the total amount of insider trading is very small, and is thus unlikely to create much of an adverse-selection problem for most stocks. Under the current regulatory regime in the United States, these adverse-selection costs do indeed seem to be low. Jeng, Metrick and Zeckhauser (2003) examine all reported trading by insiders in the United States from 1975 to 1996. After estimating the profits earned by insiders on these trades, the authors estimate that non-insiders have expected trading losses of about ten cents per $10,000 trade for non-insider sales and less than one cent per $10,000 trade for non-insider purchases. These results require two caveats. First, the study considers only the trades that were reported to the SEC. If the most profitable trades by insiders go unreported, then non-insiders may face larger expected losses. Second, these costs reflect the regulatory regime in place during the relevant period in the United States. If insider-trading restrictions were significantly loosened, then the frequency and profitability of insider trades might be quite different.

Informational efficiency. A second argument in favour of regulation is that, in the absence of regulation, insiders might be induced to hoard information until such time as it could be exploited in the most profitable way. For example, suppose that the managers of company XYZ have just learned of a major problem at one of their production facilities, which they expect to reduce company value by ten per cent. At the same time, managers also learn that a major research breakthrough has been made on another project, which would have an offsetting effect on firm value. Under these assumptions, if both pieces of information were immediately released, there would be no stock-price reaction. However, if insider trading were always permitted, managers would have an incentive to delay one of these announcements. For example, managers could release the bad news first, decreasing the stock price, and then buy stock in advance of releasing the good news.

The anti-regulation side provides a direct counterargument, claiming that insider trading is likely to speed up the flow of information to the market. As a counter to the example presented above, imagine that managers learn only the bad news about the production facility, with no good news about research. In this case, one can imagine these managers trying to contain this information for as long as possible, perhaps in the hope that the problem can be fixed before it is made public. In a regime without any insider trading, this strategy might be possible. With no restrictions on insider trading, however, managers would have a strong incentive to sell shares. In an extreme case, they could even sell shares they do not own (‘short selling’), thus providing a virtually unlimited amount of selling, and driving the price to its ‘correct’ level. Opponents of regulation argue that this kind of scenario is common, and that insider trading would allow stock prices to adjust more quickly to new information. Unfortunately, there is no empirical evidence to give us more insight into this debate, nor is it easy to imagine a plausible data-set that could provide such evidence.

Market manipulation. Once again, consider the situation of company XYZ, with problems at its production facility and the potential of research breakthroughs. For managers who live through these events, it is only a short leap to imagine the possibilities of market manipulation. For example, a well-placed rumour – coming from an insider – could move the stock price and allow for profitable trading. Opponents of regulation could counter that market manipulation can be illegal, even if trading on inside information is not. The game can grow more complex, however, if managers engage in real activities that allow for higher volatility and increased trading opportunities. For example, a CEO can increase

expenditure on research and development well beyond optimal levels, safe in the knowledge that this combination of projects will increase the real underlying volatility of corporate value. In this case, the manager is manipulating the economic activities of the firm in an economically wasteful manner.

Efficient managerial compensation. Opponents of regulation argue that profitable insider trading is mostly a transfer of wealth from shareholders to managers, and thus can be treated the same as any other form of managerial compensation. For example, if shareholders believe that the CEO of their company can earn about $5 million per year from trading on inside information, then the company can reduce the CEO's other compensation by that same amount. In this scenario, the shareholders are not injured at all by the insider trading. Of course, this argument rests on the absence of the other costs discussed above: market liquidity, informational efficiency, and market manipulation.

The costs of regulation. Opponents of regulation argue that effective enforcement would be prohibitively costly. Insiders have many vehicles to exploit their superior information. In addition to stock trading in their own account, they can tip other traders, sometimes using complex ‘tipping chains’ that are difficult to detect. Furthermore, insiders may be able to exploit superior information by not trading at all. For example, if a manager of XYZ was planning to buy stock, but then learns bad news about a production problem, he could then decide not to buy. Since the manager has taken no action, there is no conceivable way that this exploitation of inside information could be detected. Of course, these opportunities for insider ‘non-trading’ are limited, since they presuppose a standing (but reversible) decision to trade. The importance of this argument is ultimately an empirical question. In the absence of more complete information, it is impossible to know the frequency of different kinds of trading opportunities and the costs of detecting each type. Proponents of regulation can also argue that, even when detection probabilities are low, sufficiently high penalties can still provide effective deterrence.

Necessity of federal law. Prior to the Cady case of 1961, insider trading in the United States was governed by state law. Opponents of regulation argue that these state laws are sufficient, and the regulation of insider trading under federal securities laws is illogical and inefficient. There is much legal scholarship to support this view (Bainbridge, 2001), as the legal theories of insider trading are still struggling for a solid foundation, having adopted and discarded several models in the decades since Cady. The economic justification for leaving insider-trading regulation to the states rests on the identification of insider trading as a private issue between a company and its shareholders, with no externalities to security markets. If it is indeed a private issue, then opponents of regulation are correct that insider trading is the purview of other corporate law, which is left to individual states. If externalities exist – for example due to effects on market liquidity or informational efficiency – then federal regulation can be justified.

Overall, these six issues comprise the main topics of debate between the pro-regulation and anti-regulation sides. As seen by this survey, the empirical evidence on each of these issues is limited. For the debate as a whole, the best evidence comes from the aforementioned paper of Bhattacharya and Daouk (2002). After surveying the 103 countries with stock markets to assess the existence and enforcement of insider-trading laws, the authors used a variety of methods to estimate the cost of capital in each country. They find significant evidence that the cost of capital falls after the first enforcement of insider-trading laws. In contrast, the establishment of laws (prior to the first enforcement) has no effect on the cost of capital. Thus, for some combination of reasons – liquidity, informational efficiency, and so on – it is cheaper for firms to raise capital in markets that enforce prohibitions against trading on inside information.

Market efficiency

While law-and-economics scholars focused on social welfare, financial economists saw a good opportunity to use insider-trading data to test market efficiency. This literature began in earnest with the definitions of the efficient markets hypothesis (EMH) (Roberts, 1967), which comes in three versions: weak, semi-strong, and strong. Weak-form efficiency means that current asset prices incorporate all information contained in past prices; semi-strong efficiency means that current asset prices incorporate all public information; strong-form efficiency means that current asset prices incorporate all relevant information, both public and non-public.

Data on trading by insiders can be used to test both the strong and semi-strong versions of the EMH. If the strong form of the EMH holds, then insiders should not be able to make excess profits on their trades, since any information possessed by insiders would already be incorporated in market prices. One can test this implication of the EMH by analysing the risk-adjusted returns earned by insiders, where the main complication is the definition of ‘risk-adjusted returns’. The capital asset pricing model (CAPM) was the first model of risk-adjusted returns to be widely adopted by economists. Finnerty (1976) uses the CAPM to evaluate the equally weighted returns to all insider trades in NYSE stocks from 1969 to 1972. He finds that insider buys overperform and insider sales underperform their CAPM benchmarks, thus providing the first direct evidence against the strong form of the EMH.

In the decades that followed Finnerty's study, researchers developed several other methods of computing risk-adjusted returns. Jeng, Metrick and Zeckhauser (2003) test the strong form of EMH using these more modern methods on 25 years of disclosed insider trading: they conclude that insiders earn positive risk-adjusted returns on their purchases but not on their sales. Since both the Finnerty and Jeng, Metrick and Zeckhauser studies focus on transactions reported to the SEC, they may both be underestimating insider profits if the most profitable transactions are unreported. While comprehensive data on unreported transactions is, by definition, unavailable, a unique study by Meulbroek (1992) does provide some evidence. Using proprietary data from SEC investigations of insider trading, then-SEC employee Meulbroek concluded that these transactions earned substantial risk-adjusted profits. Overall, the Finnerty, Jeng, Metrick and Zeckhauser, and Meulbroek studies provide significant evidence against the strong form of the EMH.

While the strong form of the EMH is of interest to regulators and academics, the semi-strong version commands far greater attention from investors; if the semi-strong version is false, then there exist profitable trading strategies based on public information. Economists have focused on insider-trading data as one possible source of such information. The first study of this data is Smith (1941), who finds no trading advantage for insiders, a result that discouraged other researchers until the work of Lorie and Niederhoffer (1968). These authors point out the severe problems of the SEC data, with trade dates often off by several weeks. These data problems invalidated the Smith study and opened the door to a new generation of analyses.

To handle these problems, Lorie and Niederhoffer devised a strategy that has dominated the insider-trading literature to this day: analyse the risk-adjusted returns to firms in relation to the ‘intensity’ of insiders’ purchases and sales over well-defined periods. For example, a stock may be labelled an ‘insider buy’ for a month if at least three insiders bought the stock and no insiders sold it. In the decades that

followed, many authors adopted this methodology, with the most important examples being Jaffe (1974) and Seyhun (1986). These many studies use a variety of intensive-trading criteria for many different sample periods, and are nearly unanimous in concluding that stocks that are intensely bought tend to outperform relevant benchmarks over a subsequent period, and that those that are intensely sold tend to underperform. They provide mixed evidence on whether other investors can profit, after transactions costs, by using this information. Seyhun (1998) summarizes this evidence and concludes that several different trading rules lead to profits. Overall, this literature provides strong evidence against the semi-strong version of the EMH. As in all tests of the EMH, this conclusion is specific to the time period studied and the models used to estimate risk-adjusted returns. Defenders of the EMH can always propose that the effect will go away once investors learn about it, or that researchers will discover some additional risk factor to explain the results.

Conclusion

After 40 years of intense study, research in insider-trading has made substantial progress. Scholars of law and economics have identified the main arguments for and against the regulation of insider trading, and the limited empirical evidence on these arguments has sharpened the debate for future researchers. Further progress is most likely using data-sets from the many countries that have recently begun to regulate insider trading. For financial economists, the evidence on market efficiency is more straightforward. There is significant evidence that insiders profit on their own trades, and that outsiders can profit by gleaning information from the trades of insiders. Under the assumption that there is no missing risk factor that can explain these results, this evidence argues against both the strong and the semi-strong versions of the efficient markets hypothesis.

See Also

- capital asset pricing model
- efficient markets hypothesis
- law, economic analysis of

Bibliography

Bainbridge, S. 2001. Insider trading. In *Encyclopedia of Law and Economics, Volume III. The Regulation of Contracts*, ed. B. Bouckaert and G. De Geest. Cheltenham: Edward Elgar.

Bhattacharya, U. and Daouk, H. 2002. The world price of insider trading. *Journal of Finance* 57, 75–108.

Finnerty, J.E. 1976. Insiders and market efficiency. *Journal of Finance* 31, 1141–8.

Jaffe, J.F. 1974. Special information and insider trading. *Journal of Business* 47, 410–28.

Jeng, L.A., Metrick, A. and Zeckhauser, R.J. 2003. Estimating the returns to insider trading: a performance-evaluation perspective. *Review of Economics and Statistics* 85, 453–71.

Lorie, J.H. and Niederhoffer, V. 1968. Predictive and statistical properties of insider trading. *Journal of Law and Economics* 11, 35–53.

Manne, H. 1966. *Insider Trading and the Stock Market*. New York: Free Press.

Meulbroek, L.K. 1992. An empirical analysis of illegal insider trading. *Journal of Finance* 47, 1661–99.

Roberts, H. 1967. Statistical versus clinical prediction of the stock market. Unpublished manuscript, Centre for research in Security Prices, University of Chicago.

Seyhun, H.N. 1986. Insiders' profits, costs of trading and market efficiency. *Journal of Financial Economics* 16, 189–212.

Seyhun, H.N. 1998. *Investment Intelligence from Insider Trading*. Cambridge, MA: MIT Press.

Smith, F.P. 1941. *Management Trading, Stock-market Prices, and Profits*. New Haven, CT: Yale University Press.

How to cite this article

Metrick, Andrew. "insider trading." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000122> doi:10.1057/9780230226203.0807

The New Palgrave Dictionary of Economics Online

institutional economics

Warren J. Samuels

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Institutional economics, also concerned with resource allocation and the level and distribution of aggregate income, is primarily concerned with the organization and control of the economy, that is, its power structure, which governs whose interests count. Institutionalists have a broader or deeper set of explanatory variables, including the fundamental economic role of government, the socialization of the individual, and the consequences of a business system. Thus, prices are a function of demand and supply, these a function of markets and rights, manifest in the actions of firms and governments, and the latter a matter of business control of government.

Keywords

Adams, J.; agricultural economics; antitrust enforcement; Association for Evolutionary Economics; Ayres, C.; Barbash, J.; Boulding, K.; Bromley, D.; capitalism; Clark, J. M.; Commons, J.; competition; conspicuous consumption; corporatism; DeGregori, T.; Dewey, J.; Dugger, W.; economic development; equilibrium; European Association of Evolutionary Political Economy; evolutionism; Fusfeld, D.; Galbraith, J. Kenneth; German Historical School; Gordon, W.; Gruchy, A.; Hamilton, D.; heterodox economics; Hobson, J.; Hodgson, G.; holism; industrial organization; institutional economics; labour economics; law and economics; macroeconomics; Means, G.; methodological individualism; Mitchell, W.; Myrdal, G.; natural resource economics; Neale, W.; neoclassical economics; optimality; Parsons, K.; Perlman, S.; Peterson, W.; Post Keynesian economics; power; preference formation; resource allocation; Rutherford, M.; Schmid, A.; Schumpeter, J.; social norms; socialism; Solo, R.; Stanfield, R.; status; Taussig, F.; technology; Tool, M.; Trebing, H.; value-free economics; Veblen, T.; Waller, W.; wealth; Weber, M.; Wieser, F. von; Witte, E.

Article

Apart from Marxism, which has also the character of a social movement, institutional economics has become the principal school of heterodox thought in economics. Originating in and still concentrated largely, but by no means exclusively, within the United States, institutionalism has served the dual functions of providing critiques of mainstream neoclassical (and Marxian) economics and producing an alternative conception of the economy, and of doing economic research and analysis. In so doing, it has represented in part a continuation of the German and English historical traditions, including Max Weber, as well as other writers such as John Hobson.

Early position

The place of institutional economics thus described applies principally to the post-Second World War period. During the interwar period, the picture was substantially different. For most economists, institutionalist ideas and theories were very much a part of economics. Many economists, typified by Frank William Taussig, John Maurice Clark, Friedrich von Wieser and Joseph A. Schumpeter, did not make a fundamental distinction in their own work between institutional and neoclassical economics or, if they did differentiate the two, nonetheless pursued both modes of doing economics. They could work on aspects of the problem of organization and control and on the institutional foundations of markets pretty much simultaneously with work on the theory of competition and the working of pure abstract markets, with each enriching the other. Some economists were less eclectic in their orientation. They continued the antagonism of Thorstein Veblen, on the one hand, or developed the antagonism of those suspicious of institutionalism as another form of interventionism and as largely unreceptive to the development of mathematical formalism in economic theory, on the other hand. The work of Malcolm Rutherford and others has shown a discipline largely undifferentiated in terms of institutionalism versus neoclassicism during the interwar period.

The precise relationship of heterodox institutional economics to orthodox neoclassical economics in the post-Second World War period is complicated by several considerations: the awkward sociological status of heterodoxy within the discipline; the ambivalence within institutionalism as to the relationship, some institutionalists feeling that the two schools are complementary and others that the two are mutually exclusive; and the presence within institutionalism of two different and to some extent conflicting traditions, one emanating from Thorstein Veblen and continuing through Clarence Ayres, the other starting with John R. Commons. The Veblen–Ayres tradition focuses on the progressive role of technology and the inhibitive role of institutions; the Commons tradition is less enamoured of the imperatives of technology and approaches institutions, as modes of collective action, more neutrally; both groups accept that actual economic performance is a function, inter alia, of both technology and institutions. Notwithstanding their differences, there is a common core of institutional analysis of perhaps no greater variety of formulation than within neoclassicism or Marxism.

Relation to mainstream economics

Mainstream economists maintain that the central economic problems are the allocation of resources, the distribution of income, and the determination of the levels of income, output and prices. In contrast, institutional economists assert the primacy of the problem of the organization and control of the economic system, that is, its structure of power. Thus, whereas orthodox economists have a strong

tendency to identify the economy solely with the market, institutional economists argue that the market is itself an institution, comprised of a host of subsidiary institutions, and interactive with other institutional complexes in society. In short, the economy is more than the market mechanism: it includes the institutions which form, structure, and operate through, or channel the operation of, the market. The fundamental institutionalist position is that it is not the market but the organizational structure of the larger economy which effectively allocates resources.

To the extent, then, that institutional and neoclassical economists study the same questions (for example, resource allocation) the institutionalists generally encompass a broader or deeper set of explanatory variables: instead of having price and resource allocation be a function of demand and supply in a purely conceptual market, these latter are in turn related to the structure of power (wealth, institutions) which help form them. Power structure in turn is related to legal rights, thence to the use of government in forming legal rights of economic significance and thereby influencing the allocation of resources, level of income, and distribution of wealth.

Institutionalists are generally less concerned with price and resource allocation per se and more with the problem of the organization and control of the economy: that is, with performance seen as specific to power (rights) structure, as well as to technology. Institutionalists are interested, for example, in the formation and role of institutions, and the interrelations between economic and legal systems and between power and belief systems.

If institutionalists insist that the economy comprises more than the market mechanism, they also object to the equilibrium and presumptive optimality modes of analysis of neoclassical economics. The search for the deterministic technical conditions of stable equilibrium, it is felt, obscures the fundamental power and choice aspects of the economy. The search for optimality, or for optimal solutions, it is also felt, is either formally empty or can be given substance only by the introduction, typically implicitly, of antecedent normative assumptions as to whose interests count, whereas in the real world such questions have to be worked out both within institutions and through contests over institutional adjustment and reformation.

Principal ideas

The central features of institutional thought are its holism and evolutionism. Thus the further principal themes of institutional economics include the following:

1. 1. A theory of social change, and an activist orientation towards social institutions, through focusing on both the substantive impact of institutions on economic performance and the processes of institutional change, treating institutions not as something to be taken as given but as man-made and changeable, both deliberatively and non-deliberatively.
2. 2. A theory of social control and collective choice, or a theory of institutions, a focus on the formation and operation of institutions as both cause and consequence of the power structure and societized behaviour of individuals and subgroups, and as the mode through which economies are organized and controlled. Instead of focusing on the mechanics of choice from within opportunity sets, a focus on the formation of opportunity sets; instead of a focus on unfettered market freedom, a focus on the total, complex pattern of freedom and control, that is, on the formation and operation of the system of control through which both actual opportunity sets and multi-

dimensional freedom are formed.

3. 3. A theory of the economic role of government, as a principal social process through which both itself and other institutions of economic significance are in part formed and revised. Instead of treating government, law, and the system of rights as either given and/or exogenous, these are treated as both dependent and independent, and always critical, not merely aberrational, economic variables.
4. 4. A theory of technology, as defining and determining the relative scarcity of all resources, as a principal force in the evolution of economic structure (including the operation of institutions) and performance, and as the basis of the logic of industrialization marking the mentality as well as the practices of modern economies.
5. 5. The fundamental principle that the real determinant of resource allocation is not the market but the organizational – institutional, power – structure of society.
6. 6. An emphasis on facets of the value conception which transcend price, on the values represented in and given effect by the habits and customs of social life, on the pragmatic, instrumental values ensconced in the transcendental notion of the life process of man and society, and on the constructive values latent within and given effect by the working rules of law which are both the foundation and the product of the power structure of society. Included are attempts to understand the process by which values are changed, in contrast to the orthodox assumption of given values; that is, to consider within economics such questions as where the values come from, how they are tested, and how they are changed.

In amplification of these themes one finds, for example, Veblen's emphasis on status emulation as a principal force in the formation of economic behaviour, including (through conspicuous consumption and the making of invidious comparisons) the formation of consumer demands; Commons's analysis of the evolution of the fundamental legal foundations of the modern economy; John Dewey's theory of instrumental logic and social value; John Maurice Clark's analysis of the social control of business; Wesley Mitchell's emphasis on the economy as a pecuniary phenomenon; Commons's and Selig Perlman's analyses of labour unions as a mode of representing worker interests and of generating institutional change; Edwin E. Witte's, and Commons's, efforts at creating new institutions for the embodiment and protection of rising interests and for the creative resolution of social conflict and the development of a body of analysis of institutional genesis and adjustment; and, inter alia, Veblen's and Ayres's analyses of the formation of the human belief system, including that of economists, under the impact of the contest between traditional and new ways of doing things.

Apropos of the last point, institutionalists have freely pointed to the selectivity and typically implicit nature of the operative assumptions of neoclassical analysis. They insist that, by its taking institutional or power structure as given or, more typically, by its selective specification of institutions and power structure, there is a strong tendency towards selective apologetics in orthodox economics, especially in that work which is directed to the identification of ‘optimal’ solutions. The institutionalist solution to such problems is that of Gunnar Myrdal: to avoid the pretence of value-free economics by making all, or substantially all and certainly the operative, value premises explicit and by generating appraisals thereof. Accordingly, institutional economists have tended to avoid recourse to methodological individualism and to abstain from puzzle-solving research in the context of models devoid of institutional embodiment and stressing equilibrium, optimality, and purely competitive markets. They have rather attended to

theoretical and empirical analyses of real-world problems, such as the operation of particular institutions, business–government relations, and the conditions of economic development. In so far as they have dealt with economic variables at fundamental conceptual levels, such as government and rights, they have at least tried to do so in both analytically credible and non-presumptive ways.

Internal conflict

Conflict within institutionalism has largely been on two issues. One involves the putative dichotomy of technology and institutions. The other is between those who call for government planning to modify if not replace private enterprise and those who favour private enterprise but call for strong antitrust enforcement to ensure a competitive market economy.

John Kenneth Galbraith

The best-known contemporary version of the institutionalist conception of the economy has been that of John Kenneth Galbraith. Following the course laid down by Veblen, and grafting it on to a version of Keynesian economics, Galbraith explored the corporate nature and planning modes of the business system and the impact of what he considers to be technological imperatives, the social formation of individual preferences underlying demand functions, the power and continuous interaction of the state and the corporate core of the economy, the factors and forces which influence the formation of opinion and policy in the public sector, and the inevitability of resolving conflicts of interest on the basis of some conception of public purpose.

Widespread practice

In such fields as labour economics, industrial organization, economic development, law and economics, agricultural and natural resource economics, and macroeconomics, institutionalists, through their primary attention to power structure and belief system, in the context of their overriding concerns with social change and social control, have produced understandings of economic reality quite different from those of neoclassical economists. These contributions have come through the recent work, in addition to Galbraith, of John Adams, Jack Barbash, Kenneth E. Boulding, Dan Bromley, Thomas DeGregori, William Dugger, Daniel R. Fusfeld, Wendell C. Gordon, Allan G. Gruchy, David B. Hamilton, Gardiner C. Means, Walter C. Neale, Kenneth Parsons, Wallace Peterson, A. Allan Schmid, Robert Solo, Ron Stanfield, Paul Strassmann, Marc Tool, Harry M. Trebing and William Waller, among others. Some of this work appears in the *Journal of Economic Issues*, published by the Association for Evolutionary Economics. Also in the United States, institutional economists have joined with Post Keynesian economists and with varieties of political economists to explore empirically and theoretically topics central to those fields. In Europe, Geoffrey Hodgson and others have pursued the development and application of evolution theory to the array of institutionalist topics. Some have studied the formation, use and impact of technology and others, for example, the organizational theory applicable to the corporation. The European Association of Evolutionary Political Economy has become the major forum for European institutionalists, and even for many Americans.

Altogether this work has constituted an alternative analysis of the economic system, especially of capitalism but also of socialism, and a critique of both existing economic systems and orthodox schools of economics.

See Also

- Ayres, Clarence Edwin
- Clark, John Maurice
- Commons, John Rogers
- evolutionary economics
- Galbraith, John Kenneth
- heterodox economics
- institutionalism, old
- Mitchell, Wesley Clair
- Post Keynesian economics
- Veblen, Thorstein Bunde
- Witte, Edwin Emil

Bibliography

Canterbery, E. et al. 1984. Galbraith symposium. *Journal of Post-Keynesian Economics* 7(1).

Dorfman, J. et al. 1963. *Institutional Economics*. Berkeley: University of California Press.

Gruchy, A. 1947. *Modern Economic Thought*. New York: Prentice-Hall.

Parker, R. 2005. *John Kenneth Galbraith: His Life, His Politics, His Economics*. New York: Farrar, Straus and Giroux.

Sharpe, M. 1974. *John Kenneth Galbraith and the Lower Economics*. 2nd edn. White Plains, NY: International Arts and Sciences Press.

Thompson, C. (ed.) 1967. *Institutional Adjustment*. Austin: University of Texas Press.

Tool, M. 1979. *The Discretionary Economy*. Santa Monica: Goodyear.

How to cite this article

Samuels, Warren J. "institutional economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000125> doi:10.1057/9780230226203.0808

The New Palgrave Dictionary of Economics Online

institutional trap

Victor Polterovich
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

One of the main obstacles for successful economic development is the formation of institutional traps, inefficient yet stable norms of behaviour. Domination of barter exchange, arrears, corruption and black market activities are examples of institutional traps that have hampered reforms in transition economies. Institutional traps are supported by mechanisms of coordination, learning, linkage and cultural inertia. The acceleration of economic growth, systemic crisis, the evolution of some cultural characteristics and the development of civil society may result in breaking out of institutional traps. Examples from the history of the United States and Russia are considered.

Keywords

arrears; barter; civic culture; civil society; coordination failures; corruption; cultural inertia; hysteresis; institutional trap; linkage effect; lock-in; multiple equilibria; path dependence; rent seeking; reputation; systemic crises; transaction costs; transformation costs; transitional rent; trust

Article

Institutional trap is a stable but yet inefficient equilibrium in a system where agents choose a norm of behaviour (an institution) among several options. It is usually implied that multiplicity of equilibria prevails in the system, and that an institutional trap is Pareto dominated.
The concept of institutional trap is closely related to the notion of lock-in used by Arthur (1988) and North (1990); these authors showed that inefficient technical or institutional development can be self-supporting. In fact institutional traps have been studied in many papers (see for example Ickes and Ryterman, 1992; Tirole, 1996; Bicchieri and Rovelli, 1995; Jonson, Kaufman and Shleifer, 1997; Uribe, 1997). In Polterovich (2000; 2004) a general scheme for the formation of an institutional trap was described. The theory developed was successful in explaining a number of important features of wide-scale institutional transformation in Russia and other post-communist countries where the evolution of

institutional traps was clearly observable. In particular, it was shown that such different phenomena as barter, mutual arrears, tax evasion, and corruption were intensified and supported during the reforms due to similar mechanisms. Also studied were possible strategies for a country to get out of an institutional trap.

Norm-fixing mechanisms and institutional trap formation

A norm is a rule that large groups of people can or must obey. In any area of life and at each moment in time, a multitude of alternative norms is available, and every agent has to make his or her choice. For example, an official may choose either corruption or honest service.

Each agent who interacts with partners within the framework of a certain behavioural norm has to bear the corresponding transaction costs. For example, the possibility of being caught while taking a bribe would cause a transaction cost component for an official who has chosen corruption as the norm.

The costs of transition from one norm to another are called transformation costs. These may be incurred by an individual, a firm or the state. If a firm decides to switch from black market to legal operations, it has to search for new partners. Search expenditure is a part of the transformation cost.

For a behavioural norm to be stable, individuals should feel that it is unprofitable or disadvantageous for them to deviate from it. This means that the present value of the difference between the transaction cost of a prevailing norm and any alternative norms has to be less than the related transformation cost. The main type of stabilizing mechanism is based on the coordination effect, according to which the more consistently a norm is observed in a society the greater are the costs incurred by each individual deviating from it. For example, the coordination effect takes place if a personal probability to be punished for a rule-breaking activity decreases with the number of people involved in the activity. In this respect, institutional traps belong to a broader class of coordination failures (Howitt, 2003; see also poverty traps).

With time, the transaction costs of a norm's observance decreases due to learning effect since the agents learn to operate more efficiently. If the payment of taxes is considered a norm within a society, the taxpaying technology improves. If, on the contrary, tax evasion is a norm, the relevant techniques develop. A decrease of the transaction costs fixes the norm.

Another mechanism, referred to as the linkage effect, is also important. With time, an established norm finds itself linked with a multitude of other rules, and becomes part of a system of other norms. Therefore, non-observance of this norm triggers a chain of other transformations and, consequently, leads to high transformation costs. By increasing transformation costs, the linkage effect, too, contributes to a norm's fixation.

There is yet another norm-fixing mechanism, cultural inertia, which denotes agents’ reluctance to review those behavioural stereotypes that have already proven viable. Inertia effects may be supported by a formal or informal system of punishments and awards for past behaviour. For example, a person with a good reputation tries to maintain that reputation by following respectable norms of conduct.

As with any other norm, an institutional trap's stability means that a system absorbing a small external impact will remain in the institutional trap, having perhaps slightly changed its parameters, and will return to the former equilibrium state once the source of destabilizing pressure is removed. An individual or a small group of people loses if it deviates from an institutional trap. However, the simultaneous adoption by all agents of an alternative norm may be Pareto improving. Thus the lack of

coordination is the main cause of the institutional trap stability.

The emergence of institutional traps is an important source of risk associated with any reform process. The universal norm-fixing mechanisms described above, the coordination, learning and linkage effects, as well as cultural inertia, are responsible for institutional trap formation.

Consider a system with multiplicity of equilibria, and let an efficient norm prevail. Under a strong perturbation, the equilibrium may lose its stability or disappear so that the system moves to an alternative stable equilibrium, a potential institutional trap. After the disturbing factor is removed the system remains in the new equilibrium, which is now inefficient. This is the so-called hysteresis effect, which is a form of a system's dependence on its former path of development (path dependence).

A number of unexpected phenomena observed during the wide-scale reforms of the 1990s, including the rise and persistence of arrears, corruption, black market activity, and barter exchange, may be considered as institutional traps. Using the Russian experience, one can describe barter and corruption traps formation in greater detail.

Example 1: barter

In modern economies, barter is associated with higher transaction costs than monetary transactions. When the inflation rate increases, paper money loses its value. Economic agents try to diminish their losses and seek to accelerate the rates of money circulation, which means an increase of their transaction costs. The transaction costs of monetary exchanges may grow very rapidly, if the finance system fails to cope with the rocketing number of transactions.

In economies with advanced banking systems the share of barter is rather modest, even when inflation is high. But after price liberalization in 1992, Russia proved to be ripe for barter. With the banking system still unformed, money transfers within Moscow could take up to two weeks, and beyond the capital, over a month. It sometimes made more sense to carry bags of cash from city to city by plane than to transfer money from one bank account to another. Many firms soon found that barter transaction costs were lower than those for monetary exchange. Moreover, the transformation costs of a shift to barter looked acceptable, given the pre-reform direct links between supplier and consumer that had been typical in the centrally planned economy. The search for prospective partners and the process of trade negotiations were facilitated by the spread of sophisticated means of communication. The larger the number of firms choosing barter, the lower the barter transaction costs for a fixed barter volume since it was easier to find partners and put together barter chains (a coordination effect). In those conditions, as the share of barter exchanges increased, even more companies became involved.

Thus the environment conducive to barter had been created by changes in fundamental parameters, such as the rate of inflation and the risk of arrears, which radically increased the ratio of monetary exchange transaction costs to barter exchange transaction costs. The coordination effect triggered a rapid formation of a barter economy. Later, the transaction costs of barter exchanges continued to decrease due to the learning effect: companies learned to design elaborate chains of barter exchanges. The newly established norm gave birth to a new institute of barter exchange intermediaries and proved to be an efficient instrument of tax evasion (linkage effect).

By 1997, inflation in Russia had decreased dramatically, and monetary exchange technology had notably improved. Barter practices, however, were not dropped altogether. Barter-driven behaviour was supported by the coordination effect; it has been fixed through learning, linkage and cultural inertia. Any

agent deciding to break out of the barter system would be exposed to inevitable transformation costs. He or she would be forced to sever long-established connections, to look for new partners, and to be ready to come face to face with the tax-collecting authorities. The barter intermediaries, who would lose their main sources of income if barter practices were eliminated, formed a potential group of pressure for perpetuation of the relevant norm. This is the hysteresis effect mentioned above.

Example 2: corruption

Every potential bribe-taker makes decisions comparing his or her gains from bribes and from honest behaviour. In Russia, income inequality jumped sharply during transition because of uneven transitional rent expropriation. The state was not able to properly adjust the salaries of bureaucrats, so the salaries were insignificant in comparison to bribes from the newly rich. This caused an increase in corruption activity. Inefficient government policy, inadequate legislation, unclear norms for new market behaviour and weak mechanisms of government control contributed to a rise in corruption.

The larger the scale of corruption, the smaller were the chances for a bribe-taker to be caught. Corruption technologies were developed with time, corruption hierarchies arose, and corruption activities were closely linked with other shadow economy mechanisms. Corruption turned out to be habitual for both the bureaucrats and the population. The coordination, learning, and linkage mechanisms as well as cultural inertia made the corruption system even more stable.

One can find institutional traps in the history of many developed countries. The United States of 19th century presents a good illustration of the corruption trap (Knott and Miller, 1987, pp. 15–31). The time between 1815 and 1840 was a period of intensive transformations of political institutions in the United States. Property ownership requirements were abandoned to allow the lower classes to vote. These democratic reforms had unanticipated consequences, however. The political party machine became an effective instrument for some party bosses to get rich. Such men allocated public service positions (including those of postmaster, customs official, and policemen) among their supporters without taking into account competence or skills. Office workers were forced to pay a proportion of their wage to the political party through whom they had obtained their jobs. The police were a political tool rather than a law enforcement agency. Businessmen paid bribes for franchises. Low-level policemen took payments for ‘permitting’ local vice operations, and the money was distributed among the police hierarchy and the political bosses. Many people understood that the situation had to be changed, but nobody wanted to make a move. This was a corruption trap.

Once it has fallen into an institutional trap, the system chooses a non-efficient path of development, and, with time, returning to efficient development may be very difficult even if possible.

Escaping from an institutional trap

However, there are reasons to believe that some institutional traps are stable in the medium run only and that an economy can gradually develop mechanisms conducive to its escaping from institutional traps. The theory outlined above gives us a framework for the systematic consideration and classification of different mechanisms that may facilitate this transformation.

One has to reach at least one of the following goals: (*a*) to increase the transaction costs of the prevalent inefficient norm; (*b*) to decrease the transaction costs of an alternative efficient norm; (*c*) to bring down

the transformation costs of the transition to an efficient norm. The coordination, linkage or/and inertia mechanisms have to be influenced for these purposes.

Below we consider microeconomic measures and macroeconomic policies that may be taken by a government, as well as spontaneous tendencies that are helpful for an economy to escape institutional traps.

Microeconomic measures and macroeconomic policies

The simplest way of increasing the transaction costs of an inefficient norm is the introduction of a high penalty for deviating behaviour: for example, a strong punishment for corruption or a special tax on barter exchange. However, high penalties are very costly. There are at least three sources of penalty costs. First, enforcement of stronger penalties requires larger resources to be spent. Large fees may result in strong resistance on the part of the penalized persons. Second, a penalty directed to decrease the intensity of an inefficient norm may increase the intensity of its even more inefficient substitutes. Fee increasing may shift the system to another institutional trap instead of shifting it to an efficient equilibrium. For example, strong punishment for arrears could create additional incentives for firms to escape into the underground economy. Third, one should take into account the possibility of wrong decisions. The stronger the punishment of an innocent person, the larger the social losses.

The development of reputation mechanisms is another way of increasing the transaction costs of corruption, arrears, or tax evasion (Tirole, 1996). These mechanisms also decrease transaction costs of efficient norms, creating incentives to observe them. At the start of the Russian transition, old reputation mechanisms were totally destroyed. New mechanisms arose gradually, due to strengthening of the state and formation of new business networks.

Amnesty is an instrument of weakening inertia effects in the cases of tax evasion, arrears and corruption. Many governments use this measure. The outcome is mixed, however. To be successful the amnesty has to be an unexpected event, conducted at an appropriate moment when fundamental causes for a trap are exhausted, and it has to be complemented by other measures weakening linkage and coordination effects. The rotation of officials may be an effective measure for destroying unproductive coordination (see a theory of rotation in Ickes and Samuelson, 1987).

Macroeconomic policy also influences the evolution of institutional traps. In choosing tax, social, or industrial policies, one has to take into account that they can create incentives or disincentives for participation in black market operations or corruption.

Spontaneous exit

There are some spontaneous tendencies which, being unintended, may nevertheless facilitate exit from institutional traps.

A number of institutional traps (corruption and tax evasion traps, for example) are connected with rent-seeking behaviour. Each economic agent may invest his or her money and time into production or into rent-seeking activity. The choice depends on the relative efficiency of these two options. If rent-seeking dominates, then many agents choose this option, and an institutional trap may arise.

At a time of major institutional transformation, some economic agents are able to derive additional

income – transitional rent – exclusively from their fortunate positions. Price liberalization gives the advantage to suppliers of goods in high demand. Foreign trade liberalization allows importers and exporters to profit from differences between domestic and world prices. The emergence of new stock exchanges and securities markets creates ample arbitrage opportunities for financial intermediaries.

If the state does not take special measures to extract transitional rent, rent-seeking becomes much more profitable than production. An increasing number of economic agents find themselves to be involved in rent-seeking activity, and increasing volumes of resources are diverted from productive activities. The rate of production growth falls, and this makes production even less attractive for investors. Coordination, learning, linkage, and inertia mechanisms start to work and form institutional traps.

If, however, the rate of economic growth substantially increases due to improvements of technology or term of trade, then some agents may decide to increase their investment into production. This supports growth and creates new incentives for the next cohort of agents to switch their efforts from rent-seeking to production. As a result, an institutional trap may disappear. Growth diminishes the transaction costs of ‘good behaviour’ and facilitates improvement of institutions. This conclusion was corroborated by econometric calculations (Chong and Calderon, 2000) as well as theoretical research (Balatsky, 2002).

Evolution of civic culture

One way out of an institutional trap is disadvantageous for each isolated economic agent but advantageous for society as a whole. The root of the problem is lack of coordination. The ability of agents to coordinate their efforts depends on the prevailing civic culture and the development of civil society.

Most studies of economic growth consider civic culture as a fixed and non-changing factor. However, some important parameters of civic culture may change drastically during a period of 10–20 years; therefore long-term considerations have to take them into account. For example, the proportion of people who revealed political interest in Germany was 27 per cent in 1952 and 50 per cent in 1977; the proportion of affirmative answers on the question ‘Can most people be trusted?’ increased from 9 per cent in 1948 to 39 per cent in 1976 (Conradt, 1989). Political interest and social trust are important preconditions for social activity and the strengthening of civil society. Note that the proportion of respondents who belonged to a voluntary organization grew in Germany from 44 per cent in 1959 to 50 per cent in 1967, and 59 per cent in 1975.

Lack of trust has direct economic consequence: it increases transaction costs and decreases investment (Zak and Knack, 2001). If social activity is intensified and the degree of social trust increases, coordination becomes less costly; and there are more chances to escape from institutional traps.

The history of the US corruption trap, mentioned above, demonstrates the importance of the development of civil society (Knott and Miller, 1987, pp. 33–53). By the turn of the 19th century, a powerful progressive movement had emerged. The movement combined the efforts of several groups of citizens including middle-class taxpayers, small businessmen, farmers, and professionals of various sorts. Their main goal was an administrative reform that would separate politics from administration. They required administration according to rules, the selection of civil officers according to merit and qualification, the standardization and simplification of procedures, the centralization of administrative authority under a single executive in accordance with the principles of hierarchy. Progressives created a number of organizations such as the New York Municipal Research Bureau, New York Citizen's Union,

and the Milwaukee Free Press, and occupied leading positions in both Republican and Democratic Parties. The US Republican President Theodore Roosevelt and Democratic President Woodrow Wilson conducted reforms in accordance with progressive ideas and constructed a new system of governance based on independent commissions. The elimination of the corruption trap was a result of these reforms.

Systemic crises

Sometimes systemic crises can be helpful in helping an economy escape from an institutional trap. (The idea that a systemic crisis may be advantageous has been put forward and studied in a number of papers: see Drazen and Grilli, 1993.)

A crisis drastically changes system parameters and even destroys supporting mechanisms so that an economy may find itself outside the attraction area of the inefficient norm. The evolution of the barter trap in Russia serves as a remarkable illustration of this statement.

The barter trap was broken in 1998 due to systemic financial crisis. In consequence of the rouble devaluation the dollar has strengthened against the rouble by about two times in real term. Imports dropped drastically – in 1999 to 56 per cent of the 1997 level. Exports decreased because of the rise in oil prices. Real wage rates also dropped. However, the overall demand for domestic goods increased, labour costs diminished and the economy started to grow. The crisis totally destroyed the government bond market, which diverted money flows from production purposes. Enterprises started to earn money and used it for investments. Their real balances increased. All these changes contributed into a strong decrease in monetary exchange transaction costs. The share of barter in industrial sales fell dramatically. In 2002 it was about ten per cent. The barter trap disappeared, including the complicated system of barter intermediaries. The crisis achieved what the government had not been able to do.

Conclusion

Institutional traps are serious obstacles to economic development. Many countries have found themselves in institutional traps. Some were able to escape, others have been searching for an exit for a long time.

The main cause of institutional traps is lack of coordination. The market is a powerful coordination mechanism; however, if the market fails, the government may try to prevent an institutional trap or facilitate getting out of it by developing reputation mechanisms, implementing an amnesty, improving administration and choosing appropriate macroeconomic policies. In many cases, however, neither market nor government measures are effective in the short run. Civil society institutions have to be developed to reach the necessary coordination. This is a point that may be helpful in integrating cultural and civil society studies into the theory of economic development.

See Also

- arrears
- barter
- path dependence

- poverty traps
- state capture and corruption in transition economies

Bibliography

Arthur, W.B. 1988. Self-reinforcing mechanisms in economics. In *The Economy as an Evolving Complex System*, ed. P.W. Anderson, K. Arrow and D. Pines. Santa Fe: Addison-Wesley.

Balatsky. 2002. Functional properties of institutional traps. Economics and mathematical methods (Funkzional'nye svoistva instituzional'nyh lovushek). *Ekonomika i matematicheskie metody* 38, 54–72. (In Russian).

Bicchieri, C. and Rovelli, C. 1995. Evolution and revolution: the dynamics of corruption. *Rationality and Society* 7, 201–24.

Chong, A. and Calderon, C. 2000. Causality and feedback between institutional measures and economic growth. *Economics and Politics* 12, 201–24.

Conradt, D.P. 1989. Changing German political culture. In *Civic Culture Revisited*, ed. G.A. Almond and S. Verba. Newbury Park: Sage Publications.

Drazen, A. and Grilli, V. 1993. The benefit of crises for economic reforms. *American Economic Review* 83, 598–607.

Howitt, P. 2003. Coordination failures. In *An Encyclopedia of Macroeconomics*, ed. B. Snowdon and H. R. Vane. Cheltenham: Edward Elgar.

Ickes, B.W. and Ryterman, R. 1992. The interenterprise arrears crisis in Russia. *Post-Soviet Affairs* 8, 331–61.

Ickes, B.W. and Samuelson, L. 1987. Job transfers and incentives in complex economic organizations: thwarting the Ratchet effect. *Rand Journal of Economics* 18, 275–86.

Jonson, S., Kaufman, D. and Shleifer, A. 1997. The unofficial economy in transition. *Brookings Papers on Economic Activity* 2, 159–239.

North, D. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

Polterovich, V. 2000. Institutional traps. In *The New Russia: Transition Gone Awry*, ed. L.R. Klein and M. Pomer. Stanford: Stanford University Press.

Polterovich, V. 2005. Institutional traps: is there a way out? *Social Sciences, A Quarterly Journal of the Russian Academy of Sciences* 36, 30–40.

Tirole, J.A. 1996. A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies* 63, 1–22.

Uribe, M. 1997. Hysteresis in a simple model of currency substitution. *Journal of Monetary Economics* 40, 185–202.

Zak, P.J. and Knack, S. 2001. Trust and growth. *Economic Journal* 3l, 295–321.

How to cite this article

Polterovich, Victor. "institutional trap." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000262> doi:10.1057/9780230226203.0809

The New Palgrave Dictionary of Economics Online

institutionalism, old

Malcolm Rutherford
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

What is now called ‘old’ institutional economics was a central part of a pluralistic American economics during the inter-war period. It is a tradition that still exists today but as a marginal heterodoxy to a dominant neoclassical mainstream. By the early 1920s it had established itself as an appealing programme with a major presence at leading universities and research institutes. Institutionalist work over the inter-war period included significant contributions to economic measurement and analysis. A number of factors led to the decline of institutional economics after the Second World War, but institutionalism has continued in a modified form, and still attracts adherents today.

Keywords

Adams, H.C.; administered pricing; Ayres, C.E.; business cycles; Clark, J.M.; collective bargaining; Commons, J.R.; conciliation and mediation; conspicuous consumption; Cooley, C.H.; Dewey, J.; Ely, R. T.; Hamilton, W.H.; Hobson, J.A.; household economics; household production; institutional economics; institutionalism, old; marginal utility theory; market failure; Mitchell, W.C.; monopoly; National Bureau of Economic Research; New Deal; overhead costs; perfect competition; Planning; public utilities; public utility regulation; separation of ownership and control; social norms; underconsumptionism; Veblen, T.

Article

What is now often referred to as the ‘old’ institutional economics was a central part of American economics during the inter-war period, and is a tradition of economics that still exists today.
The explicit identification of something called the ‘institutional approach’ to economics, or ‘institutional economics’, goes back to 1918 and to Walton Hamilton's American Economic Association (AEA) conference paper, ‘The Institutional Approach to Economic Theory’ (Hamilton, 1919).
Hamilton's paper was a call for the profession at large to adopt the ‘institutional approach’. For

Hamilton, anything that ‘aspired to the name of economic theory’ had to be (i) capable of giving unity to economic investigations of many different areas; (ii) relevant to the problem of social control; (iii) relate to institutions as both the ‘changeable elements of economic life and the agencies through which they are to be directed’; (iv) concerned with ‘process’ in the form of institutional change and development; and (v) based on an acceptable theory of human behaviour, in harmony with the ‘conclusions of modern social psychology’. According to Hamilton, only an approach to economics that focused on the institutions that make up the ‘economic order’ could meet these tests. He identified H.C. Adams, Charles Horton Cooley (his own teachers at Michigan), Thorstein Veblen and Wesley Mitchell as the leaders of this movement. At the same session of the AEA conference, J.M. Clark (Clark, 1919), argued for an economics both ‘relevant to the issues of its time’ and based on an ‘ideal of scientific impartiality’. Walter Stewart (Hamilton's friend and colleague) chaired the session, and argued that economics needed to be ‘organized around the central problem of control’, should utilize the ‘most competent thought in the related sciences of psychology and sociology’, and combine ‘the statistical method and the institutional approach’ (Stewart, 1919, p. 319).

The exact timing of this effort to promote ‘institutional economics’ as a distinctive approach probably had much to do with the end of the First World War. The war had impressed upon many the great importance of improved economic data and policy analysis, and of the potential role of government in the economy. The period of reconstruction seemed to offer significant opportunities for bringing changes to the conduct of economic research, education, and policy. The 1918 session of the AEA conference was followed by further efforts to promote institutional economics. Another AEA session critical of traditional theory was organized in 1920. This featured J.M. Clark's paper ‘Soundings in Non-Euclidian Economics’ (Clark, 1921), which criticized orthodox theoretical propositions. In 1924 Mitchell argued in his presidential address to the AEA that quantitative methods would transform economics by displacing traditional theory and leading to a much greater stress on institutions (Mitchell, 1925). Lionel Edie called this address ‘a genuine manifesto of quantitative and institutional economics’, one that stated ‘the faith of a very large part of the younger generation of economists’ (Edie, 1927, p. 417). In the same year Rexford Tugwell edited *The Trend of Economics*, a book again seen as something of an institutionalist manifesto and which included papers from Mitchell and Clark as well as from younger people of institutionalist persuasion such as Tugwell himself, F.C. Mills, Sumner Slichter, Morris Copeland, and Robert Hale (Tugwell, 1924).

During the inter-war period institutionalism developed a significant following, with a concentrated presence at a number of major schools and research institutes. In addition to Veblen, Hamilton, Clark, Mitchell, and Commons, who were the most visible proponents of institutionalism, there were many others associated with the movement (Rutherford, 2000a; 2000b). The two major centres for institutionalism over the whole inter-war period were Columbia and Wisconsin, at that time among the leading doctoral departments of economics in the country. Wisconsin's department included Commons (until he retired in 1933), E.E. Witte, Harold Groves, Martin Glaeser, Selig Perlman and several others (Rutherford, 2006). Columbia was an even bigger centre for institutionalism with Mitchell, Clark, Rexford Tugwell, F.C. Mills, A.R. Burns, Joseph Dorfman, Leo Wolman, Carter Goodrich, James Bonbright, and Robert Hale all in the Economics Department or Business School at various times, and Gardiner Means, Adolf A. Berle, and many other people of related views in other departments (Rutherford, 2004). Chicago had an institutionalist contingent at least until Clark left for Columbia in

1926, and Walton Hamilton was at the centre of groups first at Amherst (1915–23) and later at the Robert Brookings Graduate School (1923–8). Other institutionalist groups existed at Texas, where Clarence Ayres joined Robert Montgomery in 1930, and in a number of other schools and colleges (Rutherford, 2003; 2007).

Among research institutes, the Institute of Economics, which became part of the Brookings Institution, was heavily institutionalist in character (the research staff included Isador Lubin, and Edwin Nourse among others). The National Bureau of Economic Research (NBER) was closely associated with Mitchell's quantitative approach and his programme of business cycle research and employed many of his Columbia colleagues and students. The quantitative and policy orientation of the work done by these organizations attracted funding from foundations such as Carnegie and Rockefeller (Rutherford, 2005a).

The sources and appeal of institutional economics

The elements that went to make up the core of the institutional approach as defined by Hamilton, were all present in American economics before 1918. Institutionalism as it formed in the inter-war period was an approach to economics that derived from several sources. While the single most significant source of inspiration for institutionalism was the work of Thorstein Veblen, it is important to understand that institutionalism was a blending of ideas taken from Veblen with those from others (Rutherford, 2001), and was never simply Veblenism.

At the most basic level the most important element in the institutionalist approach is the conception of the economic system as a set of evolving social institutions. In this, institutions are seen as much more than constraints on individual action. Social norms, conventions, laws, and common practices embody generally accepted ways of thinking and behaving, and they work to mould the preferences and values of individuals brought up under their sway. A good part of this orientation came from Veblen, but also from sociologists such as Charles Horton Cooley, and from a previous generation of German-influenced scholars (such as R.T. Ely and H.C. Adams). At this time, in line with the German model, sociology was commonly taught within economics departments.

On a more specific level, Veblen's framework, which stressed the role of new technology in bringing about institutional change (by changing the underlying ways of living and thinking) and the predominantly ‘pecuniary’ character of the existing set of American institutions, was widely influential among institutionalists. Within this framework Veblen developed his analyses of ‘conspicuous consumption’; the effect of corporate finance on the ownership and control of firms; business and financial strategies for profit-making, salesmanship and advertising; the emergence of a specialist managerial class; business fluctuations; and many other topics (Veblen, 1899; 1904).

For Veblen, the existing legal and social institutions of America were outmoded and inadequate for the task of the social control of modern large-scale industry. Veblen perceived a *systemic* failure of ‘business’ institutions to channel private economic activity in ways consistent with the public interest. He attacked the manipulative, restrictive, and unproductive tactics used by business to generate income (including consolidations, control via holding companies and interlocking directorates, financial manipulation, insider dealing, sharp practices, and unscrupulous salesmanship), the ‘waste’ generated by monopoly restriction, unemployment, conspicuous consumption, and competitive advertising, and he held out little hope of change short of a complete rejection of ‘business’ principles.

Cooley also analysed pecuniary institutions but in more measured tones, and it must be emphasized that many institutionalists, including Hamilton, Clark, Commons, and Hale placed a much greater emphasis on the evolution of legal institutions than did Veblen. Both Hamilton and Hale moved into law schools and had close connections with legal scholars of the realist school. The major sources of this emphasis on legal institutions were Ely (who taught Commons) and Adams (who taught Hamilton). This greater emphasis on law and on legal evolution helped to shift the character of institutionalism away from Veblen's radicalism and connect it to a pragmatic philosophy, based primarily on the work of John Dewey, which looked to legislative and legal reform concerning such issues such as business regulation, labour law, collective bargaining, health and safety regulations, and consumer protection. Thus, in the hands of institutionalists such as Hamilton, Clark, Mitchell, and Commons, the problem became one of supplementing the market with other forms of ‘social control’ of business.

Another important element was the linking of institutional economics with ‘modern psychology’. Veblen had provided a particularly penetrating criticism of the hedonistic psychology implicit in marginal utility theory (Veblen, 1898) and pointed to an alternative based on instinct/habit psychology. What was important for institutionalists, however, was less Veblen's specific formulation but the impetus he gave to the idea that economics needed to be reconstructed on the basis of a theory of human behaviour in harmony with the conclusions of modern psychology (see Mitchell, 1910a; 1910b).

Finally, and of central importance to the attraction of institutionalism, was the claim that it represented the ideal of empirical science. An important influence here was Mitchell's combination of Veblenian ideas concerning the significance of the institutions of the ‘money economy’ with the quantitative and statistical approach he had absorbed as a student at Chicago. Mitchell's *Business Cycles* (1913) was enthusiastically received and widely regarded at the time as a paradigm for a scientific economics. Mitchell thought of business cycles as a phenomenon arising out of the patterns of behaviour generated by the institutions of a developed money economy (Mitchell, 1927), and he explicitly connected quantitative work and the institutional approach, arguing that it is institutions that create the regularities in the behaviour of the mass of people that quantitative work analyses (Mitchell, 1924; 1925). Mitchell's quantitative bent was shared by many other institutionalists, but the scientific method, for institutionalists, was not confined to the statistical or quantitative, and included all work that was genuinely ‘investigative’ in character. It is important to comprehend that at this time it was institutionalists, not neoclassicals, who were claiming to be following the methods of natural science (Rutherford, 1999), and seemed to be at one with the general movement in American social science towards greater empiricism and ‘realism’.

At its inception, then, institutionalism could be seen as a very promising programme – modern, scientific, pointing to a critical investigation and analysis of the existing economic system and its performance, in tune with the latest in psychological, social scientific, and legal research, established at leading universities and research institutes, and involved in important issues of economic policy and reform (see also Yonay, 1998).

The contributions of interwar institutionalism

Mark Blaug has stated that institutionalism ‘was never more than a tenuous inclination to dissent from orthodox economics’ (Blaug, 1978, p. 712), and this view still finds wide currency. In fact,

institutionalsm in the inter-war period was a major part of a pluralistic mainstream economics (Morgan and Rutherford, 1998). That institutionalists did have a positive programme of research in mind should be clear from the above. Not all elements of this programme were pursued successfully, but there can be no doubt that institutionalists did make important positive contributions to economics, and this is particularly true of the period when institutionalism was at its peak. Just a few of these contributions will be highlighted.

Institutionalist took the task of improving economic measurement seriously. The NBER not only produced many empirical studies relating to business cycles, labour, and price movements, but also played a vital role in the development of national income accounting, through the work of Mitchell's student, Simon Kuznets. In conjunction with the Federal Reserve, the NBER also did much to develop monetary and financial data. Moreover, during the New Deal, institutionalists were heavily involved in the effort to improve the statistical work of government agencies (Rutherford, 2002).

As noted above, one of the claims of institutionalists was that a ‘scientific’ economics would have to be consistent with ‘modern’ psychology. A typical argument was that economics ‘is a science of human behaviour’ and any conception of human behaviour that the economist may adopt ‘is a matter of psychology’ (Clark, 1918, p. 4). Clark made one of the most interesting efforts to develop the psychological basis of institutional economics. Building on the work of William James and Cooley, he argued that the ‘effort of decision’ is an important cost, and one that prevents maximization. Clark was considering both the costs of information gathering and of calculation, and his argument is a clear precursor of more recent conceptions of bounded rationality leading to the use of habits or routines.

Interesting work on the economics of consumption and the household, was pursued by Hazel Kyrk and Theresa McMahon. McMahon made use of Veblen's conception of emulation in consumption, while Kyrk was critical of marginal utility theory as a basis for a theory of consumption and emphasized the social nature of the formation of consumption values. Consumption patterns relate to habitual ‘standards of living’, and Kyrk undertook to measure and critically analyse existing standards of living, and to create policy to help achieve higher standards of living. In her later work she discussed the household in both its producing and consuming roles, the division of labour between the sexes, employment and earnings of women, adequacy of family incomes, and issues of risks of disability, unemployment, provision for the future and social security, and the protection and education of the consumer (Kyrk, 1923; 1933; McMahon, 1925).

There was much work dealing with the inadequacy of the standard models of perfect competition and pure monopoly. The soft coal industry received particular attention. In that industry investigators such as Hamilton found little that corresponded to the ideal of a competitive industry. Competition within the industry had resulted not in efficient low-cost production but in persistent excess capacity, inefficiency, irregular operation, poor working conditions and low earnings (Hamilton and Wright, 1925). This represented a common institutionalist theme – that, particularly under conditions of high overheads and rapid technological advance, competition could lead to ‘disorder’ and inefficiency rather than to order and efficiency. Institutionalists also studied such things as common pool problems in the oil industry, production cycles in agriculture, including the cobweb model and its implications for the orthodox view of ‘self-regulating’ markets, and the vast array of restrictive practices to be found in many industries (Hamilton and Associates, 1938).

A related theme was that technological change had altered the structure of costs faced by firms and had

altered their behaviour. This argument derived from Clark's *Overhead Costs* (1923). For Clark, the growth of overhead costs as a result of capital-intensive methods of production had resulted in price discrimination, an extension of monopoly and an increase in price inflexibility over the cycle. A little later Gardiner Means (1935) developed his theory of administered pricing, which sparked a vast literature on relative price inflexibility.

On issues of corporate finance and ownership, Bonbright and Means co-authored *The Holding Company*, and Berle and Means *The Modern Corporation and Private Property*, both in 1932. These works much extended Veblen's earlier discussions of corporate consolidation and the separation of ownership and control. Berle and Means's work raised important issues of agency, and whether managers would maximize profits.

On labour market issues, institutionalists concerned themselves with studying unions and the history of the labour movement, developing in the process both classifications of unions and explanations for the particular pattern of trade union development in America (Perlman, 1928). Wage determination was also a problem that attracted the attention of institutionalists. Walton Hamilton's 1923 book *The Control of Wages* (with Stacy May) was praised by Clark for providing not an ‘abstract formulation of the characteristic outcome’ but a ‘directory of the forces to be studied’ in any particular case (Clark, 1927, pp. 276–7). Discussions of trade unions and wage bargaining were provided by other institutional labour economists such as Commons (1924) and Sumner Slichter (1931). In this work much attention was given to issues of collective bargaining and systems of conciliation and mediation.

Public utilities, including issues relating to the valuation of utility property and the proper basis for rate regulation, were major areas of institutionalist research. Both Clark and Commons devoted considerable attention to the concept of intangible property, goodwill, and valuation issues (Commons, 1924; Clark, 1926). Bonbright dealt with the difference between commercial and social valuation in connection with public utilities. Bonbright, Hale, and Martin Glaeser all wrote extensively on issues of public utility regulation, with Hale probably having the greatest impact with his campaign of criticism of the ‘fair value’ concept as a basis for rate regulation (Hale, 1921; Bonbright, 1961, p. 164).

In his *Social Control of Business* (1926) Clark argued that business cannot be regarded as a purely private affair. This idea of private business being broadly ‘affected with a public interest’ was absolutely central to the institutionalist argument for regulation of business. Clark expresses the idea in his claim that ‘every business is “affected with a public interest” of one sort or another’ (Clark, 1926, p. 185), and the argument also appears in as a central theme in Tugwell's early work on regulation (Tugwell, 1921; 1922), and in Walton Hamilton's and Robert Hale's extensive writings on law and economics (Rutherford, 2005b; Fried, 1998).

More general interconnections between law and economics and the operation of markets were addressed by Hale, Commons, and Hamilton. Commons's approach was the most developed and was built on his notions of the pervasiveness of distributional conflicts, of legislatures and courts as attempting to resolve conflicts (at least between those interest groups with representation), and of the evolution of the law as the outcome of these ongoing processes of conflict resolution. He developed his concept of the ‘transaction’ as the basic unit of analysis (later adopted by Oliver Williamson). In turn, the terms of transactions were determined by legal rights and by economic (bargaining) power. Market transactions always involved some degree of ‘coercion’, in the sense of some degree of restriction upon alternatives

(Commons, 1924; 1932; Hale, 1923). He also provided a theory of the behaviour of legislatures based on ‘log-rolling’, and a theory of judicial decision-making based on the concept of ‘reasonableness’, a concept that included, but was not limited to, a concern with efficiency (Commons, 1932; 1934).

The institutionalist programme dealing with business cycles, in the period before the depression, was centred on Wesley Mitchell's work and that he promoted through the NBER. As noted above, Mitchell explicitly placed his work on business cycles within an institutional context by associating cycles with the functioning of the system of pecuniary institutions. Mitchell's 1913 volume *Business Cycles*, with its discussion of the four-phase cycle driven by an interaction of factors such as the behaviour of profit seeking firms, the behaviour of banks, and the leads and lags in the adjustment of prices and wages, became the standard institutionalist reference. At the NBER, Mitchell focused heavily on promoting work that would add to the understanding of business cycles, generating a stream of research studies far too long to list here, but contributing to the development of national income measures, business cycle indicators, and much more. In addition, Clark developed his concept of the accelerator out of his study of Mitchell's 1913 work, and the accelerator mechanism soon became a standard part of cycle theory (Clark, 1917). Mitchell's work was not the only approach to business cycles to be found within institutionalism. Many institutionalists, including Hamilton, had an interest in the work of J.A. Hobson, and Hobson's underconsumptionism became popular among institutionalists in the 1930s (Rutherford, 1994).

On issues of market failure, broadly conceived, Clark (1926) discussed a large number of types of market failure in his *Social Control of Business*. These included monopoly, maintaining the ethical level of competition, protecting individuals where they are unable to properly judge alternatives, problems of agency, relief for people displaced by rapid economic and technological change, relief of poverty (including social security and minimum wages), regulation of advertising and the provision of information and standards, increasing equality of opportunity, externalities (‘unpaid costs of industry’), public goods (‘inappropriable services’), the wastes of ‘arms race’ types of competition (such as competitive advertising), unemployment, the interests of posterity or future generations, and any other discrepancy between private and social accounting. Slichter (1924) provided a list of problems almost as long, including the pro-cyclical behaviour of banks, overexploitation of natural resources, discrimination in employment, advertising and salesmanship, lack of market information, pollution and other external effects, uncertainty and unemployment, economic waste and inefficiency, and economic conflict. All these problems were seen as justifying some additional ‘social control’ of business activity.

Finally, and intimately related to the above, institutionalists made important contributions to policy in their roles in the development of unemployment insurance, workmen's compensation, social security, labour legislation, public utility regulation, agricultural price support programmes, and in the promotion of government ‘planning’ to create high and stable levels of output. Commons had pioneered public utility regulation, unemployment insurance, and workmen's compensation in Wisconsin, and the Wisconsin model was widely influential. Many institutionalists were active members of the American Association of Labor Legislation (AALL), and the AALL promoted many reforms to labour legislation. Medical insurance programmes were also pursued by the AALL, and also by the Committee on the Cost of Medical Care, which involved both Hamilton and Mitchell.

Institutionalists had significant influence within the New Deal. Many of Commons's students played leading roles in the development of the federal social security programme. Berle and Tugwell were two

of Roosevelt's original ‘Brains Trust’, and Tugwell, Means, and Mordecai Ezekiel were the leading advocates of the ‘structuralist’ or planning approach that had influence in the early part of the New Deal (Barber, 1996). Hamilton and several others were deeply involved in the labour legislation and consumer protection aspects of the New Deal. Hamilton later worked with Thurman Arnold in developing their case by case approach to anti-trust (Rutherford, 2005b).

Institutional economics after 1945

Institutionalism attained a significant position in American economics in the interwar period, both in academia and in government, but then declined in position and prestige after the Second World War. At this point institutionalism fell out of the mainstream of American economics to become a heterodox tradition on the margins of the discipline. There are quite a number of overlapping reasons for this, some of which reach back into the 1920s and 1930s, but the focus here will be limited to just a few of the more important issues.

Institutionalism clearly did not live up to its own early promise, particularly in its failure to pin down exactly what foundations in ‘modern psychology’ it was supposed to have. After the mid-1920s, psychologists abandoned the instinct/habit approach in favour of a behaviourism that became increasingly narrow and difficult to see as an adequate foundation for economics. In this climate, the enthusiasm for new psychological approaches that had played such a role in the institutionalist movement's beginnings could not be sustained. Institutionalism probably played a part in ridding economics of explicitly hedonistic language, but it did not develop the alternative basis to convince the profession as a whole to abandon its traditional views of rationality (Lewin, 1996).

It must also be said that institutionalists failed to develop their theories of social norms, technological change, legislative and judicial decision-making, transactions, and forms of business enterprise (apart from issues of ownership and control) much beyond the stage reached by Veblen and Commons. The reasons for this lack of development relate partly to the focus of interwar institutionalists on immediate and pressing policy problems, like business cycles, labour law, and social security. In addition, from the late 1920s on, sociology separated itself from economics and became established in separate departments, taking much of the subject matter of social norms and institutions with it.

It is also the case that, from the 1930s onwards, many new developments in theory and methods occurred within economics: developments that tended to displace institutionalist ideas and methods. Hicks's revision of demand theory seemed to free economics from the shifting basis of psychology, while the work of Joan Robinson and Edward Chamberlin provided treatments of imperfect competition more amenable to neoclassical approaches. The discussion of externalities in terms of market failure was also much clarified. Neoclassicism developed a language capable of encompassing at least some of the issues of concern to institutionalists; issues that had formerly fallen outside the neoclassical theoretical compass.

Moreover, institutionalist approaches to business cycles were replaced by Keynesian ideas. In many respects, Keynesian economics took over the role of the exciting ‘new’ economics that institutionalism had played in the early 1920s. In addition, neoclassical and Keynesian economics gained an empirical component with the rise of econometrics. Institutionalists could no longer claim greater ‘scientific’ standing because of their empiricism; indeed, they were accused by Koopmans (1947) of ‘measurement

without theory'; a much exaggerated view, but one often repeated and widely accepted.

In these ways more 'orthodox' economic theory took over those aspects of institutionalism amenable to 'model analysis' (Copeland, 1951) while other aspects were absorbed into what became applied field areas, such as industrial organization, labour economics, and industrial relations. At least until the 1960s these field areas had only loose ties to the theoretical core of the discipline, and maintained a substantial institutional component.

Finally, a significant part of the institutionalist agenda of social reform had come to pass, both removing some of the original causes of the institutionalist movement, and prompting a reaction in the form of critiques of the expanded role for government that institutionalists had done so much to put forward.

Under these circumstances, it is not difficult to see why institutionalism slipped from being a central part of American economics to a more marginalized position. This change did not happen overnight, but was hastened by the significant amount of new hiring on the part of American universities immediately after the Second World War. These new faculty were predominantly Keynesians or neoclassicals equipped with the latest in mathematical and econometric tools. The retirement of the last of the older generation of institutionalists in the 1950s completed the process.

American institutionalism did not disappear, but it certainly changed. Insitutionalists formed the small 'Wardman Group' in 1959, an organization that later became the Association for Evolutionary Economics, still the primary organization of 'old' institutionalists in America, and the publisher of *The Journal of Economic Issues*. Institutionalism disassociated itself from the positivism that had gained popularity elsewhere (a positivism that, ironically, Mitchell and the NBER had played an important part in creating), and turned away from the methods and the core areas of the discipline that had been taken over by neoclassical and Keynesian economics. Institutionalists continued to work in applied areas, and to argue for more active government regulation and 'planning' of the economy (Gruchy, 1974), but there was also something of a movement back to the broader institutional themes found in Veblen and Commons.

This tendency was especially promoted by Clarence Ayres, in his *Theory of Economic Progress* (1944). Ayes attempted to renew the Veblenian emphasis on technology as the driving force behind institutional change, and developed the Veblenian distinction between business and industry into a general dichotomy between the ceremonial and instrumental aspects of culture. Ayres's charismatic personality attracted a number of students to the institutionalist ranks, and they spread his version of institutionalism to many south-western universities. The University of Texas, too, retained its institutionalist character longer than most, and in the 1960s was still the home of a substantial institutionalist group. Other institutionalist groups existed at Maryland and at Michigan State. J.K. Galbraith produced widely read and distinctly Veblenian analyses in his *Affluent Society* (1958) and *New Industrial State* (1971), while the Commons tradition in law and economics has been kept alive by Daniel Bromley, Allan Schmid, and Warren Samuels (Samuels, 1971; Schmid, 1978; Bromley, 1989).

Perhaps the most important recent development within the 'old' institutionalist tradition has been the growing interest in the work of Veblen and Commons among a new generation of European economists attracted to institutional and evolutionary ideas. One outstanding example of this is to be found in the work of Geoffrey Hodgson, who has argued forcefully for the development of an institutional economics along lines he sees as having been originally pioneered by Veblen in his evolutionary and Darwinian approach to institutions and institutional change (Hodgson, 1988; 2004).

See Also

- Ayres, Clarence Edwin
- Clark, John Maurice
- Commons, John Rogers
- Mitchell, Wesley Clair
- United States, economics in (1885–1945)
- Veblen, Thorstein Bunde

Bibliography

Ayres, C.E. 1944. *The Theory of Economic Progress*, 2nd edn. New York: Schocken, 1962.

Barber, W.J. 1996. *Designs within Disorder: Franklin D. Roosevelt, the Economists, and the Shaping of Economic Policy, 1933–1945*. New York: Cambridge University Press.

Blaug, M. 1978. *Economic Theory in Retrospect*, 3rd edn. London: Cambridge University Press.

Bonbright, J.C. 1961. *Principles of Public Utility Rates*. New York: Columbia University Press.

Bromley, D.W. 1989. *Economic Interests and Institutions: The Conceptual Foundations of Public Policy*. Oxford: Basil Blackwell.

Clark, J.M. 1917. Business acceleration and the law of demand: a technical factor in business cycles. *Journal of Political Economy* 25, 217–35.

Clark, J.M. 1918. Economics and modern psychology, I and II. *Journal of Political Economy* 26, 1–30 136–66.

Clark, J.M. 1919. Economic theory in an era of social readjustment. *American Economic Review* 9, 280–90.

Clark, J.M. 1921. Soundings in non-Euclidean economics. *American Economic Review* 11, 132–43.

Clark, J.M. 1923. *Studies in the Economics of Overhead Costs*. Chicago: University of Chicago Press.

Clark, J.M. 1926. *Social Control of Business*. Chicago: University of Chicago Press.

Clark, J.M. 1927. Recent developments in economics. In *Recent Developments in the Social Sciences*, ed. E.C. Hayes. Philadelphia: Lippencott.

Commons, J.R. 1924. *The Legal Foundations of Capitalism*. Madison: University of Wisconsin Press, 1968.

Commons, J.R. 1932. The problem of correlating law, economics and ethics. *Wisconsin Law Review* 8, 3–26.

Commons, J.R. 1934. *Institutional Economics: Its Place in Political Economy*. New York: Macmillan.

Copeland, M.A. 1951. Institutional economics and model analysis. *American Economic Review* 41, 56–65.

Edie, L.D. 1927. Some positive contributions of the institutional concept. *Quarterly Journal of Economics* 41, 405–40.

Fried, B.H. 1998. *The Progressive Assault on Laissez Faire: Robert Hale and the First Law and Economics Movement*. Cambridge, MA: Harvard University Press.

Galbraith, J.K. 1958. *The Affluent Society*. Boston: Houghton Mifflin.

Galbraith, J.K. 1971. *The New Industrial State*, 2nd edn. Boston: Houghton Mifflin.

Gruchy, A.G. 1974. Government intervention and the social control of business: the neoinstitutionalist position. *Journal of Economic Issues* 8, 235–49.

Hale, R.L. 1921. The “physical value” fallacy in rate cases. *Yale Law Journal* 30, 710–31.

Hale, R.L. 1923. Coercion and distribution in a supposedly non-coercive state. *Political Science Quarterly* 38, 470–94.

Hamilton, W.H. 1919. The institutional approach to economic theory. *American Economic Review* 9, 309–18.

Hamilton, W.H. and May, S. 1923. *The Control of Wages*. New York: Augustus M. Kelley, 1968.

Hamilton, W.H. and Wright, H.R. 1925. *The Case of Bituminous Coal*. New York: Macmillan.

Hamilton, W.H. and Associates 1938. *Price and Price Policies*. New York: McGraw Hill.

Hodgson, G.M. 1988. *Institutions and Economics: A Manifesto for a Modern Institutional Economics*. Cambridge: Polity Press.

Hodgson, G.M. 2004. *The Evolution of Institutional Economics: Agency, Structure and Darwinism in American Institutionalism*. London: Routledge.

Koopmans, T.C. 1947. Measurement without theory. *Review of Economic Statistics* 29, 161–72.

Kyrk, H. 1923. *A Theory of Consumption*. Boston: Houghton Mifflin.

Kyrk, H. 1933. *Economic Problems of the Family*. New York: Harper.

Lewin, S. 1996. Economics and psychology: lessons for our own day from the early twentieth century. *Journal of Economic Literature* 35, 1293–323.

McMahon, T. 1925. *Social and Economic Standards of Living*. Boston: D.C. Heath.

Means, G.C. 1935. *Industrial Prices and their Relative Inflexibility*. Senate Document 13, 74th Congress, 1st Session. Washington, DC: US Government Printing Office.

Mitchell, W.C. 1910a. The rationality of economic activity, I. *Journal of Political Economy* 18, 97–113.

Mitchell, W.C. 1910b. The rationality of economic activity, II. *Journal of Political Economy* 18, 197–216.

Mitchell, W.C. 1913. *Business Cycles*. Berkeley: University of California Press.

Mitchell, W.C. 1924. The prospects of economics. In *The Trend of Economics*, ed. R.G. Tugwell. Port Washington, NY: Kennikat Press, 1971.

Mitchell, W.C. 1925. Quantitative analysis in economic theory. *American Economic Review* 15, 1–12.

Mitchell, W.C. 1927. *Business Cycles: The Problem and its Setting*. New York: NBER.

Morgan, M.S. and Rutherford, M. 1998. From Interwar Pluralism to Postwar Neoclassicism. Annual Supplement to vol. 30 of History of Political Economy, Durham, NC: Duke University Press.

Perlman, S. 1928. *A Theory of the Labor Movement*. New York: Macmillan.

Rutherford, M. 1994. J.A. Hobson and American institutionalism: underconsumption and technological change. In *J.A. Hobson after Fifty Years*, ed. J. Pheby. London: Macmillan.

Rutherford, M. 1999. Institutionalism as ‘scientific’ economics. In *From Classical Economics to the*

Theory of the Firm: Essays in Honour of D.P. O'Brien, ed. R. Backhouse and J. Creedy. Aldershot: Edward Elgar.

Rutherford, M. 2000a. Institutionalism between the wars. *Journal of Economic Issues* 34, 291–303.

Rutherford, M. 2000b. Understanding institutional economics: 1918–1929. *Journal of the History of Economic Thought* 22, 277–308.

Rutherford, M. 2001. Institutional economics: then and now. *Journal of Economic Perspectives* 15, 173–94.

Rutherford, M. 2002. Morris A. Copeland: a case study in the history of institutional economics. *Journal of the History of Economic Thought* 24, 261–90.

Rutherford, M. 2003. Walton Hamilton, Amherst, and the Brookings graduate school: institutional economics and education. *History of Political Economy* 35, 611–53.

Rutherford, M. 2004. Institutional economics at Columbia University. *History of Political Economy* 36, 31–78.

Rutherford, M. 2005a. Who's afraid of Arthur Burns? The NBER and the foundations. *Journal of the History of Economic Thought* 27, 109–39.

Rutherford, M. 2005b. Walton H. Hamilton and the public control of business. In *The Role of Government in the History of Political Economy*, ed. S. Medema and P. Boettke. Supplement to volume 37, *History of Political Economy*. Durham, NC: Duke University Press.

Rutherford, M. 2006. Wisconsin institutionalism: John R. Commons and his students. *Labor History* 47, 161–88.

Rutherford, M. 2007. Chicago economics and institutionalism. In *Elgar Companion to Chicago Economics*, ed. R. Emmett. Aldershot: Edward Elgar (forthcoming).

Samuels, W.J. 1971. The interrelations between legal and economic processes. *Journal of Law and Economics* 14, 435–50.

Schmid, A.A. 1978. *Property, Power, and Public Choice*. New York: Praeger.

Slichter, S.H. 1924. The organization and control of economic activity. In *The Trend of Economics*, ed. R.G. Tugwell. Port Washington, NY: Kennikat Press, 1971.

Slichter, S.H. 1931. *Modern Economic Society*. New York: H. Holt.

Stewart, W.W. 1919. Economic theory: discussion. *American Economic Review* 9, 319–20.

Tugwell, R.G. 1921. The economic basis for business regulation. *American Economic Review* 11, 643–58.

Tugwell, R.G. 1922. *The Economic Basis of Public Interest*. New York: Augustus M. Kelley, 1968.

Tugwell, R.G. 1924. *The Trend of Economics*. Port Washington, NY: Kennikat Press, 1971.

Veblen, T. 1898. Why is economics not an evolutionary science? In *The Place of Science in Modern Civilisation*. New York: Russell & Russell, 1961.

Veblen, T. 1899. *The Theory of the Leisure Class*. London: George Allen & Unwin, 1924.

Veblen, T. 1904. *The Theory of Business Enterprise*. Clifton, NJ: Augustus M. Kelley, 1975.

Yonay, Y.P. 1998. *The Struggle over the Soul of Economics: Institutionalist and Neoclassical Economists in America between the Wars*. Princeton, NJ: Princeton University Press.

How to cite this article

Rutherford, Malcolm. "institutionalism, old." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000268> doi:10.1057/9780230226203.0810

The New Palgrave Dictionary of Economics Online

instrumental variables

Original article by Charles E. Bates, revised by Moshe Buchinsky and Steven N. Durlauf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Instrumental variables methods are an essential tool in modern econometric practice. The method itself is of ancient lineage and historically is closely connected with the econometrics of simultaneous equations. This article describes the statistical foundations of instrumental variables methods with a focus on their classical development.

Keywords

central limit theorems; errors in variables; Euler equations; generalized method of moments estimation; instrumental variables; law of large numbers; natural experiments; returns to schooling; serial correlation; simultaneous equations models; treatment effect; two-stage least squares estimator

Article

In one of its simplest formulations the problem of estimating the parameters of a system of simultaneous equations with unknown random errors reduces to finding a way of estimating the parameters of a single linear equation of the form $Y=X\beta_0+\varepsilon$, where β_0 is unknown, Y and X are vectors of data on relevant economic variables and ε is the vector of unknown random errors. The most common method of estimating β_0 is the method of least squares: $\hat{\beta}_{OLS} \equiv \arg\min \varepsilon(\beta)'\varepsilon(\beta)$, where $\varepsilon(\beta)\equiv Y-X\beta$. Under fairly general assumptions $\hat{\beta}_{OLS}$ is an unbiased estimator of β_0 provided $E(\varepsilon_t|X)=0$ for all t, where ε_t is the tth-coordinate of ε.
Unfortunately for the empirical economist, it is often the case that the basic orthogonality condition between the errors and the explanatory variables is not satisfied by economic models, due to correlation between the errors and the explanatory variables. Particularly relevant examples of this situation include (1) any case where the data contain errors introduced by the process of collection (errors in variables problem); (2) the inclusion of a dependent variable of one equation in a system of simultaneous

equations as an explanatory variable in another equation in the system (simultaneous equations bias); and (3) the inclusion of a lagged dependent variable as an explanatory variable in the presence of serial correlation. For all of these cases,

$$E(\hat{\beta}_{OLS}) = E[(X'X)^{-1}X'Y] = \hat{\beta}_0 + E\left[(X'X)^{-1}\sum_{i=1}^{n} X'E(\varepsilon_i|X)\right] \neq \beta_0$$

in general, and the bias introduced cannot be determined because the errors ε are unknown. Furthermore, in every case the bias fails to go to zero as the sample size increases. Clearly the method of least squares is unsatisfactory for many situations of relevance to economists.

In 1925 the US Department of Agricultural published a study by the zoologist Sewall Wright where the parameters of a system of 6 equations in 13 unknown variables were estimated using a method he referred to as ‘path analysis’. In essence his approach exploited zero correlations between variables within his system of equations to construct a sufficient number of equations to estimate the unknown parameters. The idea which underlies this approach is that, if two variables are uncorrelated, then the average of the product of repeated observations of these variables will approach zero as the number of observations is increased without bound except for a negligible number of times. Thus if we know that a variable of the system Z_i is uncorrelated with the errors ε, we can exploit the fact that $n^{-1}\Sigma_{t=1}^{n} Z_{ti}\varepsilon_t \equiv n^{-1}\Sigma_{t=1}^{n} Z_{ti}(Y_t - X_t\beta_0)$ approaches zero to construct a useful relationship between parameters of the system by setting such averages equal to zero. Provided a sufficient number of such relationships can be constructed which are independent, this provides a method for estimating the parameters of a system of simultaneous equations which should become more accurate as the number of observations increases.

Since the 1940s, when Reiersøl (1941; 1945) and Geary (1949) presented the formal development of this procedure, the variables Z which are instrumental in the estimation of the parameters β_0 have been called ‘instrumental variables’. Associated with each instrumental variable Z_i is an equation formed as described in the previous paragraph, called a normal equation, which can be used to form the estimates of the unknown parameters. Frequently there are more instrumental variables than parameters to be estimated. As the equations are formed from relationships between random variables, generally no solution will exist to a system of estimating equations formed in this manner using all possible instrumental variables. As each estimating equation contains relevant information about the parameters to be estimated, it is undesirable just to ignore some of them. Thus we can define a fundamental problem in the application of this method: how can we make effective use of all the information available from the instrumental variables? This problem will occupy the rest of this article.

Let $\varepsilon_t(\theta) \equiv F_t(X_t, \bullet Y_t, \bullet\, \theta)$ be a $p\times1$ vector-valued function defined on a domain of possible parameter values $\Theta \subseteq \mathbb{R}^k$ which represents a system of p simultaneous equations with dependent variables Y_t, a $p\times1$ random vector, and an $m\times s$ random matrix of explanatory variables X_t for all t=1, 2, …, n. Standard formulations of $F_t(X_t, \bullet Y_t, \bullet\, \theta)$ are the linear model $\varepsilon_t(\theta) \equiv Y_t - X_t\theta$ and the nonlinear model $\varepsilon(\theta) \equiv Y_t$

$-f_t(X_t,\bullet\,\theta)$. Let $W_t(\theta)$ be a $p\times r$ random valued matrix defined on Θ for all t=1, 2, …, n. Assume that there exists a unique value θ_0 in Θ such that $E(\varepsilon_1^0|W_1^0)=0$ for all t=1, 2, …, n, where $\varepsilon_t^0 \equiv F_t(X_t, Y_t, \theta_0)$ and $W_t^0 \equiv W_t(\theta_0)$. Finally, let $Z_t(\theta)$ be a $p\times 1$ random matrix such that $E(Z_t(\theta)|W_t(\theta))=Z_t(\theta)$ for all θ in Θ. Any such variables $Z_t(\theta_0)$ may serve as instrumental variables for the estimation of the unknown parameters θ_0 since

$$E(Z_t^{0\prime}\varepsilon_t^0) = E(E(Z_t^{0\prime}\varepsilon_t|W_t^0)) = E(Z_t^0 E(\varepsilon_t^0|W_t^0)) = 0$$

for all t=1, 2, …, n, as long as the functions F_t and W_t and the data generating process satisfy sufficiently strong regularity assumptions to ensure that the uniform law of large numbers is satisfied, that is

$$n^{-1}\sum_{t=1}^{n} Z_t(\theta)'\varepsilon_t(\theta) \overset{P}{\square} n^{-1}\sum_{t=1}^{n} E\left[Z_t(\theta)'\varepsilon(\theta)\right]$$

uniformly in θ on Θ.

Identification of the unknown parameters θ_0 requires that there be at least as many instrumental variables as there are parameters to be estimated, that is, $l\geq k$. On the other hand, if there are more instrumental variables than parameters to be estimated, there will be no solution to $n^{-1}\Sigma_{t=1}^{n} Z_t(\theta)'\varepsilon_t(\theta)=0$ in general for finite n as indicated above. One possible solution to this problem is simply to use k of the instrumental variables in the estimation of θ_0. The omitted instrumental variables may then be used to construct statistical tests of the l–k overidentifying restrictions of the unknown parameter vector. A drawback of this approach is that not all of the information available to us is used in the estimation of the unknown parameters and hence, the estimates will not be as precise as they should be. An alternative approach which effectively uses all of the available instrumental variables is to be preferred.

Even though in general the moment function $n^{-1}\Sigma_{t=1}^{n} Z_t'(\theta)\varepsilon_t(\theta)\neq 0$ for any value of θ, its limiting function $n^{-1}\Sigma_{t=1}^{n} E[Z_t'(\theta)'\varepsilon_t(\theta)]$ does vanish when $\theta=\theta_0$. This suggests estimating θ_0 with that value of Θ which makes $n^{-1}\Sigma_{t=1}^{n} Z_t'(\theta)\varepsilon_t(\theta)$ as close to zero as possible. The criterion of closeness is of some interest to the econometrician. It affects the size of the confidence ellipsoids of the estimator about θ_0 and hence the precision of the estimate. The nonlinear instrumental variables estimator (NLIV), $\hat{\theta}_{n,\mathrm{NLIV}} = \arg\min_{\theta\in\Theta}$

$$\left[\sum_{t=1}^{n} Z_t'(\theta)\varepsilon_t(\theta)\right]' \cdot \left[\mathrm{Var}\sum_{t=1}^{n} Z_t'(\theta_0)\varepsilon_t(\theta_0)\right]^{-1} \cdot \left[\sum_{t=1}^{n} Z_t'(\theta)\varepsilon_t(\theta)\right]$$

is the optimal instrumental variables estimator in this respect (Bates and White, 1986a).

The NLIV estimator simplifies to well-known econometric estimators in a variety of alternative specifications of the underlying probability model which generated the variables. When the data generating process is independent and identically distributed, $\hat{\theta}_{n,\mathrm{NLIV}}$ is the nonlinear three-stage least squares estimator of Jorgenson and Laffont (1974). The additional restriction of consideration to a single equation (p=1) results in the nonlinear two-stage least squares estimator of Amemiya (1974).

Furthermore, if the model $\varepsilon(\theta)$ is linear in θ, $\hat{\theta}_{n,\mathrm{NLIV}}$ then simplifies to the three-stage least squares estimator of Zellner and Theil (1962) for a system of simultaneous equations and to the two-stage least squares estimator of Theil (1953), Basmann (1957) and Sargan (1958) for the estimation of the parameters of a single equation. On the other hand, if we allow for heterogeneity by restricting the data generating process only to be independent, $\hat{\theta}_{n,\mathrm{NLIV}}$ simplifies to White's (1982) two-stage instrumental variables estimator of the parameters of a single linear equation.

As indicated above, it is desirable from consideration of asymptotic precision to include as many instrumental variables as are available for the estimation of the unknown parameters θ_0. This raises the question of the existence of a set of instrumental variables $\{Z^*\}\in\Gamma$ that renders the inclusion of any further instrumental variables redundant, where Γ is the set of all sequences of instrumental variables such that $\hat{\theta}_{n,\mathrm{NLIV}}$ is a consistent estimator of θ_0 with an asymptotic covariance matrix. Bates and White (1986b) provide conditions which imply that such instrumental variables exist, though it may not be possible to obtain them in practice. Suppose there exists a sequence of k instrumental variables $\{\mathcal{Z}\}$ such that for all $\{Z\}$ in Γ

$$E[\mathcal{Z}(\theta_0)'\nabla_\theta\varepsilon(\theta_0)] = E[\mathcal{Z}(\theta_0)'\varepsilon(\theta_0)\varepsilon(\theta_0)'Z(\theta_0)].$$

Then $\mathcal{Z}(\theta_0)$ is optimal in Γ in the sense of asymptotic precision. Suppose it is also the case that Σ is an $np\times np$ matrix with representative element $\sigma_{th\tau g}\equiv E(\varepsilon_{th}(\theta_0)\cdot\varepsilon_{\tau g}(\theta_0)\,|W_{th}(\theta_0), W_{\tau g}(\theta_0))$, is nonsingular a.s. and that

$$E[E(\nabla_\theta\varepsilon_{th}(\theta_0)|W_{th}(\theta_0))|W_{\tau g}] = E(\nabla_\theta\varepsilon_{th}(\theta_0)|W_{th}(\theta_0))$$

for all t, $\tau=1, 2, \ldots, n$ and $h, g=1, 2, \ldots, p$ such that $\sigma^{th\tau g}\equiv 0$, where $\sigma^{th\tau g}$ is a representative element of Σ^{-1}. Let Z^* be an $np\times k$ matrix with rows

$$Z^*_{th} \equiv \sum_{\tau=1}^{n}\sum_{g=1}^{p} \sigma^{th\tau g} E[\nabla_\theta \varepsilon_{\tau g}(\theta_0) | W_{\tau g}(\theta_0)].$$

If $\{Z^*\}$ is in Γ then $\{Z^*\}$ is optimal in Γ.

In many situations it will not be possible to make use of such instrumental variables in practice. However, for some important situations optimal instrumental variables are available. Suppose that $\varepsilon(\theta)\equiv Y-X\theta$ and the explanatory variables X are independent of the errors $\varepsilon(\theta_0)$. If the errors are independent and identically distributed for all $t=1, 2, \ldots, n$ and $h=1, 2, \ldots, p$, then $Z^*=X$. Thus the optimal instrumental variables estimator is given by

$$\hat{\theta}_{n,\mathrm{NLIV}} \equiv \arg\min_{\theta\in\Theta} \varepsilon(\theta)' X[\sigma^2 E(X'X)]^{-1} X'\varepsilon(\theta),$$

where $\sigma^2\equiv \mathrm{var}[\varepsilon_{th}(\theta_0)]$ is a real, nonstochastic scalar for all t and h. If it is also the case that $n^{-1}E(X'X)-n^{-1}X'X\to 0$ as $n\to\infty$, $\hat{\theta}_{n,\mathrm{NLIV}}$ is asymptotically equivalent to $\arg\min_{\theta\in\Theta}\varepsilon'(\theta)X(X'X)^{-1}X'\varepsilon(\theta)$, that is ordinary least squares is the optimal instrumental variables estimator. If there is contemporaneous correlation only, that is, $\mathrm{var}(\varepsilon_t(\theta_0))=\Omega$, a $p\times p$ nonstochastic matrix, then Zellner's (1962) seemingly unrelated regression estimator (SURE), is the optimal instrumental variables estimator. If we further relax these assumptions so that $\mathrm{var}(\varepsilon(\theta_0))$ is an arbitrary positive definite $np\times np$ matrix, the generalized least squares (Aitken, 1935) is the optimal instrumental variables estimator.

Since the development of the two-stage least squares estimator in the mid-1950s, the method of instrumental variables has come to play a prominent role in the estimation of economic relationships. In turn, in modern econometric practice, the use of instrumental variables methods has been very much influenced by the evolution of economic theory as well as the evolution of the relationship between theory and empirical practice. Within macroeconomics, the standard approach to structural economic analysis, Hansen's (1982) generalized method of moments estimation (GMM) procedure, which is a generalization of nonlinear instrumental variables estimation, typically relies on economic theory to identify valid instruments. The most prominent application of GMM methods is to Euler equations

estimation, in which valid instruments are defined by a theoretical specification of how one set of variables may be understood to be the expected value of another set.

While economic theory can identify instruments that are, in principle, valid, it typically does not provide insights into whether such instruments are strongly or weakly correlated with the variables that are to be instrumented. Concern over the problem of weak instruments was stimulated by the demonstration in Bound, Jaeger and Baker (1995) that the failure to account for weak instruments called into question findings on the return to schooling in the work of Angrist and Krueger (1991). Important analyses of the effects of weak instruments on estimation and inference include Dufour and Taamouti (2005), Staiger and Stock (1997) and Stock and Wright (2000). This new literature is well surveyed in Dufour (2003) and Andrews and Stock (2006). This work demonstrates how lack of attention to the possibility of weak instruments can lead to very misleading inferences for a broad range of contexts. Furthermore, it provides new ways for conducting inference with respect to the parameters of interest.

When economic theory identifies valid instruments, it may be the case that the set of potential instruments is unbounded. This is evident in Euler equation contexts, where the validity of the instrument vector Z_{t-k} usually implies the validity of $Z_{t-k-l}, l > 0$. This has led to research on the properties of estimators where the number of available instruments grows with the number of observations. Donald and Newey (2001) and Hahn (2002) are examples of analyses of this type.

Other developments in the use of instrumental variables have occurred because of a desire to avoid the use of structural models in developing substantive economic claims. In particular, the literature on treatment effects may be understood as an effort to show how the causal effects of alternative policies may be uncovered outside the context of a structural model. A part of this literature has focused on the analysis of data from natural experiments and quasi-natural experiments, but, as argued by Heckman (1996), such analyses are in fact forms of instrumental variable estimation. Perhaps unsurprisingly, the effort to develop causal claims based on statistical rather than economic assumptions has led to controversy since the two types of assumptions are in fact interrelated. See Angrist, Imbens and Rubin (1996) as well as the discussion of this paper for different perspectives. Heckman (1997) provides a wide-ranging critique of the use of instrumental variables in contemporary research; Roy model describes constructive ways to proceed.

Finally, it is important to remember that instrumental variable estimates are normally evaluated on the basis of asymptotic properties, that is, the law of large numbers and the central limit theorems. Since in general it is not possible to know how much data are required to arrive at acceptable estimates, conclusions derived from instrumental variables estimates should be tempered with a healthy dose of scepticism.

See Also

- generalized method of moments estimation
- matching estimators
- natural experiments and quasi-natural experiments
- Roy model
- treatment effect

- two-stage least squares and the *k*-class estimator

Bibliography

Aitken, A.C. 1935. On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* 55, 42–8.

Amemiya, T. 1974. The nonlinear two-stage least-squares estimator. *Journal of Econometrics* 2, 105–10.

Andrews, D. and Stock, J. 2006. Inference with weak instruments. In *Advances in Econometrics: Proceedings of the Ninth World Congress of the Econometric Society*, ed. R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press.

Angrist, J.D., Imbens, G.W. and Rubin, D.B. 1996. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 91, 444–72.

Angrist, J. and Krueger, A. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979–1014.

Basmann, R.L. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25, 77–83.

Bates, C.E. and White, H. 1986a. Efficient estimation of parametric models. Working Paper No. 166, Department of Political Economy, Johns Hopkins University.

Bates, C.E. and White, H. 1986b. An asymptotic theory of estimation and inference for dynamic models. Working paper, Department of Political Economy, Johns Hopkins University.

Bound, J.D., Jaeger, D.A. and Baker, R. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–50.

Donald, S. and Newey, W. 2001. Choosing the number of instruments. *Econometrica* 69, 1365–87.

Dufour, J.-M. 2003. Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics* 36, 767–808.

Dufour, J.-M. and Taamouti, M. 2005. Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica* 73, 1351–65.

Geary, R.C. 1949. Determination of linear relations between systematic parts of variables with errors in

observation, the variances of which are unknown. *Econometrica* 17, 30–58.

Goldberger, A.S. 1972. Structural equation methods in the social sciences. *Econometrica* 40, 979–1001.

Hahn, J. 2002. Optimal inference with many instruments. *Econometric Theory* 18, 140–68.

Hausman, J.A. 1983. Specification and estimation of simultaneous equation models. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Heckman, J. 1996. Randomization as an instrumental variable. *Review of Economics and Statistics* 78, 336–41.

Heckman, J. 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 32, 441–62.

Jorgenson, D.W. and Laffont, J. 1974. Efficient estimation of nonlinear simultaneous equations with additive disturbances. *Annals of Economic and Social Measurement* 3, 615–40.

Reiersøl, O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9, 1–24.

Reiersøl, O. 1945. Confluence analysis by means of instrumental sets of variables. *Arkiv for Mathematik, Astronomi och Fysik* 32A, 1–119.

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.

Staiger, D. and Stock, J. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.

Stock, J. and Wright, J. 2000. GMM with weak instruments. *Econometrica* 68, 1055–96.

Theil, H. 1953. *Estimation and Simultaneous Correlation in Complete Equation Systems*. The Hague: Centraal Planbureau.

White, H. 1982. Instrumental variables regression with independent observations. *Econometrica* 50, 483–500.

White, H. 1984. *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

White, H. 1985. Instrumental variables analogs of generalized least squares estimators. *Journal of Advances in Statistical Computing and Statistical Analysis* 1, 173–227.

Wright, S. 1925. *Corn and Hog Correlations*. Washington, DC: US Department of Agriculture, Bulletin 1300.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57, 348–68.

Zellner, A. and Theil, H. 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica* 30, 54–78.

How to cite this article

Bates, Charles E., Moshe Buchinsky and Steven N. Durlauf. "instrumental variables." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000127>
doi:10.1057/9780230226203.0812

The New Palgrave Dictionary of Economics Online

instrumentalism and operationalism

Lawrence A. Boland

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Instrumentalism and Operationalism are the methodological doctrines associated respectively with Milton Friedman and Paul Samuelson. Each has a long philosophical history. Instrumentalism was the 18th-century doctrine created to deal with the Newton mechanics; Operationalism was the early 20th-century doctrine created to deal with Einstein' general relativity. With Instrumentalism one can say that theories do not have to be true, just useful – as Friedman argued in 1953. With Operationalism one is required to express theories only in terms of observable and measurable variables. Samuelson's early work was designed to demonstrate how theory can be made operational and thus potentially refutable.

Keywords

assumptions; Austrian economics; behaviouralism; Berkeley, Bishop G.; Bridgman, P.; consumer's demand curve; Einstein, A.; Friedman, M.; Instrumentalism; mathematics and economics; methodology of economics; neoclassical economics; Newton, I.; Operationalism; operationally meaningful; perfect competition; positive economics; positivism; psychology; revealed preference theory; Samuelson, P. A.; Slutsky equation; tautologies; utility functions; verificationism; weak axiom of revealed preference

Article

For any reader familiar with economic literature, the conjunction of the two methodological doctrines of Instrumentalism and Operationalism immediately calls to mind the methodological pronouncements of Milton Friedman (1953) and Paul Samuelson (1947; 1965; 1983). Interestingly, when making their methodological pronouncements, neither Friedman nor Samuelson mentions a philosopher to support his methodological viewpoint or to indicate an inspiration. Nevertheless, each of these methodologies is intended to solve a philosophical problem. Each has a philosophical history.

Friedman's Instrumentalism

Historically, Instrumentalism was a response to the success of Newton's laws of physics that were claimed to explain facts of nature and in particular explain the movements of the planets. Given the success of Isaac Newton's physics, the problem was thought to be that ordinary people would look to science instead of the church for true explanations of nature. Thus, the fear was that any recognized authority of science might undermine faith. Bishop George Berkeley in the early 18th century promoted the idea that we should allow science to be seen *only* as instruments or tools to solve practical problems (for example, instruments to calculate movements of the planets and thus make astronomical predictions). In this way, scientific explanations should not be considered true, only useful. If people would accept the doctrine of Instrumentalism then science and religion could comfortably coexist.
Friedman's Instrumentalism has nothing to do with religion. Instead it is a response to the demands of the critics of neoclassical economics and in particular criticism of the assumptions of perfect competition. In the 1930s and 1940s, the climate of philosophical opinion concerning any claim to scientific knowledge was that it must be realistic in the sense that any scientific claim can be verified as true. With this in mind, critics of neoclassical economics claimed that the assumptions of neoclassical theory would have to be shown to be true if they are claimed to be the basis of true explanations of economic behaviour – or, perhaps more importantly, if they are used to form economic policy such as that involving labour and employment (for example, Lester, 1946; 1947). The key question was: must the assumptions of economic theory be true in order to be useful? In his 1953 article ‘The Methodology of Positive Economics’, Friedman argues that useful theories do not have to be based on true assumptions, or can even, in some cases be based on assumptions that are known to be false.
Few economists today would think any theory should have to be absolutely true for any serious consideration of that theory, whether or not it is to be used for the formation of an economic policy (for example, Aumann, 1985). That is, few today think we should be concerned with the question of the absolute truth status of any scientific theory. Instead, all that is hoped is that the theory is the best available as determined by the current scientific conventions. Instrumentalism is an answer to a different question: ‘What is the *role* of scientific theories?’ As noted above, Bishop Berkeley had long ago answered this question. Scientific theories should not be considered true or false, simply because, so long as they are useful tools of analysis or prediction, their truth status does not matter. Instrumentalism will never be seen as a satisfactory methodology to economists who think the truth status of economic theory matters (for example, Lawson, 1997; 2003).
As it turns out, Instrumentalism can be a very useful tool for the defense against any competing methodological doctrine that claims either that scientific theories are true or that scientific theories must be proven true before they can be used to deal with real practical problems. Proponents of Instrumentalism can always respond to critics by merely claiming that the use of Instrumentalist methodology has proven to be very useful.

Samuelson's Operationalism

Operationalism is usually attributed to physicist Percy Bridgman's 1927 book, although several writers may have taken similar positions before (see Mirowski, 1998; Hands, 2004). Bridgman's contribution was apparently a response to the growing interest in Albert Einstein's general relativity-based theory of

physics that was being seen as the successor to the classical physics of Newton. Einstein's physics was a challenge for most people of the day to understand. Bridgman was seen to be offering a common-sense view of physical theory and methodology that fitted better with what most people understood.

The basic idea of Operationalism is that explanations should be based only on concepts and variables that can be defined by the operations used to measure them. Interestingly, Einstein in 1905 started out trying to explain his theory of special relativity by showing how what we mean by ‘simultaneous’ is not as obvious as we might think should we try to operationalize it, that is, try to define it in terms of the operations used to measure it. However, he ultimately rejected such an operationalist approach as it made the development of his theory of general relativity paradoxical or impossible (Schilpp, 1949).

Apparently, in the 1930s Operationalism was seen as a plausible means of implementing positivism. Supposedly, if every concept used in one's theory can be operationally defined and is thereby observable, then any empirical verification of a scientific theory would be beyond dispute. Moreover, in the 1930s it was commonly believed that scientific theories were meaningful while philosophical or religious theories were not. And the basic notion to support this was that one could verify scientific theories but not philosophical or religious theories. Operationalism was seen by some to be an avenue to make the common verificationist methodological perspective plausible.

About the same time as Bridgman was arguing for Operationalism, there was a movement towards behaviouralism in psychology. The motivation there was that we should avoid constructs such as consciousness or mind and use only observable behaviour. Being observable behaviour means that one could in principle define operations such that one would be able to measure human behaviour. And this is exactly what can be seen in Samuelson's advocacy of Operationalism in economics. What Samuelson wanted to purge from economic theory and in particular from the theory of the consumer was psychology (1938). That is, in Marshallian neoclassical theory, the consumer is thought to be maximizing utility whenever making a decision about what to buy; however, we cannot observe, let alone measure, the level of utility achieved. So how do we know that it is maximized? Do we have to turn our analysis of consumer decisions over to the psychologists? Samuelson thought not. Perhaps his motivation was only the recognition that, if one were to promote mathematical formulations of economics, everything would need to be quantifiable. But the advocacy of Operationalism goes beyond this by requiring the use of only observable variables to derive the fundamental laws of economics, such as the so-called Slutsky equation and the consumer's demand curve. Samuelson claimed that one did not have to assume the existence of a utility function for the consumer but only that the consumer makes well-defined and consistent observable choices. These choices and whether they are consistent is completely and directly observable and are so without any reference to psychology or utility.

Samuelson went on in his Ph.D. thesis (published in 1947; see also 1998) to say that it is possible to construct all of the important ideas in economics in such a manner that they can be shown to be in principle falsifiable. He called empirically falsifiable statements ‘operationally meaningful statements’. It should be noted, however, that by 1947 he was no longer taking the extreme view taken in 1938, which required that all assumptions of a theory be directly observable, but instead said that a theory need only be shown to have implications that are falsifiable with observable data. Any theory or model that has such implications would henceforth be deemed to be ‘operationally meaningful’.

This invocation of Operationalism is somewhat suspect. At the time Samuelson was promoting the use of mathematics in economics, critics, particularly some of the Austrian School, were claiming that all mathematical propositions are tautologies (see Hutchison 1935; 1938). Samuelson, by requiring any

theory or model to be ‘operationally meaningful’ (that is, falsifiable), was really just avoiding the critics, for there is no conceivable evidence that can refute a tautology. So, when he showed that falsifiable statements can be derived from his mathematical model of an economic theory, he proved that his mathematical model is not a tautology. Thus, Operationalism offered a means of dealing with the critics of the use of mathematics in economics.

The methodological failures of Friedman and Samuelson

By 1950 Samuelson had to admit that his original operationalist programme to purge psychological concepts such as utility from consumer theory was a failure. Consistency of today's choice with past choices presumes no change in tastes. When he revisited his version of consumer theory in 1948, he introduced preferences into the discussion. His key assumption was to be called the weak axiom of revealed preference. However, once it was recognized that the weak axiom was not sufficient for the purposes of consumer theory, the introduction of a strong version revealed that his revealed preference analysis was logically equivalent to old fashioned utility-based analysis (see Wong, 1978; 2006). This undermined any further interest in promoting Operationalism in economics.

Friedman's version of Instrumentalism begs many questions. Who decides what is meant by ‘useful’ or which empirical facts need to be predicted with one's economic model? What one economist might think an obviously successful policy that justified the use of obviously false assumptions might not be accepted as successful by critics of the realism of those assumptions. So some may think that Instrumentalism can be used to defend its use in a self-referential form by economists, others can just as easily say that the primary and perhaps sole use of Instrumentalism is to avoid criticism of theories and hence of policies recommended on the basis of those theories. Instrumentalism still lives in economics, although in a somewhat muted form (see Boland, 2003, chs. 4 and 5). The dominance of formal mathematical modelling has often been supported by appeals to Instrumentalism (for example, Aumann 1985, pp. 31–2). Rarely, however, will a promoter of mathematics in economics refer to Friedman's essay, but the methodological position taken is the same.

See Also

- assumptions controversy
- conventionalism
- Friedman, Milton
- Samuelson, Paul Anthony

Bibliography

Aumann, R. 1985. What is game theory trying to accomplish? In *Frontiers of Economics*, ed. K. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

Boland, L. 1979. A critique of Friedman's critics. *Journal of Economic Literature* 17, 503–22.

Boland, L. 2003. *The Foundations of Economic Method: A Popperian Perspective*. London: Routledge.

Bridgman, P. 1927. *The Logic of Modern Physics*. New York: Macmillan.

Friedman, M. 1953. Methodology of positive economics, *in Essays in Positive Economics*. Chicago: University of Chicago Press.

Hands, D.W. 2004. On operationalisms and economics. *Journal of Economic Issues* 38, 953–68.

Hutchison, T. 1935. A note on tautologies and the nature of economic theory. *Review of Economic Studies* 2, 159–61.

Hutchison, T. 1938. *The Significance and Basic Postulates of Economic Theory*. London: Macmillan.

Lawson, T. 1997. *Economics and Reality*. London: Routledge.

Lawson, T. 2003. *Reorienting Economics*. London: Routledge.

Lester, R. 1946. Shortcomings of marginal analysis for wage-employment problems. *American Economic Review* 36, 63–82.

Lester, R. 1947. Marginalism, minimum wages, and labor markets. *American Economic Review* 37, 135–48.

Mirowski, P. 1998. Operationalism. In *Handbook of Economic Methodology*, ed. D. W. Hands, J. Davis and U. Mäki. Cheltenham: Edward Elgar.

Samuelson, P. 1938. A note on the pure theory of consumer behaviour. *Economica* 5 (n.s.), 61–71.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge: Harvard University Press.

Samuelson, P. 1948. Consumption theory in terms of revealed preference. *Economica* n.s. 15, 243–53.

Samuelson, P. 1950. The problem of integrability in utility theory. *Economica* n.s. 17, 355–85.

Samuelson, P. 1965. *Foundations of Economic Analysis*. 2nd edn. New York: Atheneum.

Samuelson, P. 1983. *Foundations of Economic Analysis*. 3rd edn. Cambridge, MA: Harvard University Press.

Samuelson, P. 1998. How *Foundations* came to be. *Journal of Economic Literature* 36, 137–86.

Schilpp, P.A., (ed.) 1949. *Albert Einstein: Philosopher-Scientist*. Evanston, IL: Library of Living Philosophers.

Wong, S. 1973. The ‘F-twist’ and the methodology of Paul Samuelson. *American Economic Review* 63, 312–25.

Wong, S. 1978. *The Foundations of Paul Samuelson's Revealed Preference Theory*. London: Routledge & Kegan Paul.

Wong, S. 2006. *The Foundations of Paul Samuelson's Revealed Preference Theory*. 2nd edn. London: Routledge.

How to cite this article

Boland, Lawrence A. "instrumentalism and operationalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000265> doi:10.1057/9780230226203.0811

The New Palgrave Dictionary of Economics Online

insurance mathematics

James C. Hickman and Edward W. Frees

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Insurance mathematics is concerned with the valuation of obligations arising from insurance contracts. At contract initiation, valuation is known as premium determination or ratemaking, whereas, for a contract already in force, valuation is known as reserve determination. Updating these values as information is revealed involves important techniques known as experience adjustment. Models of insurance mathematics are based on probability theory and financial economics. These models are calibrated with insurance experience and present values from returns on investments in asset markets.

Keywords

benefit premium; calibration; central limit theorems; collective risk theory; compound interest; continuous interest rate; defined benefits; equivalence principle; expected values; health insurance; insurance mathematics; liability; life insurance; mortality; pensions; portfolio theory; present value; probability density function; recursion relationships; risk management; risk theory; selection bias; workers’ compensation insurance

Article

Mathematics and insurance have developed along parallel paths during the past 350 years. It is difficult to identify an economic activity more closely tied to mathematics than insurance. Since the genesis of probability ideas in the mid-17th century, there have been times when mathematical developments were ahead of insurance practice. At other times, commercial necessity required improvisations that did not rest on solid mathematical foundations. In general the science and the application moved together.

Reserves and premiums: long-term coverages

Two related valuation problems are to establish a price, or premium, and to estimate the liability created

by a contract. The basic tools for solving these problems are expected values and compound interest combined with an economic concept, the equivalence principle. The equivalence principle requires, at the time the coverage is activated, that the expected present value of premiums equals to expected present value of benefits. Following the issuance of the coverage, the principle can be extended to define the liability of the insurer as the expected present value of future benefits less the expected present value of future premiums.

Long-term insurance contracts have the possibility of extending for many years. The time of benefit payment, and for some contracts the amount of payments, often depend on the length of the survival time of the insured. Specifically, let T denote the random variable time until death. One example of a long-term coverage is life insurance with a single payment of benefits at time T. Another example is a life annuity with many payments of benefits paid during survival, up to time T. The life insurance model would apply to financing the replacement of equipment from light bulbs to generators. The mathematics of annuities would apply to funding equipment maintenance costs.

To illustrate, consider a life insurance policy paying a benefit b at death to be funded by a premium π, paid at a continuous annual rate until death. For the time until death random variable, let $s(t)=\Pr(T>t)$ be the survival function. Then the equivalence principle determines the premium π by the equation

$$-b\int_0^\infty e^{-\delta t}s'(t)\,dt = \pi\int_0^\infty e^{-\delta t}s(t)\,dt. \tag{1}$$

Here, $-s'(t)$ is the probability density function of time until death and δ is the continuous interest rate, also called the force of interest. It will be assumed constant for simplicity. It is defined through the relation $1+i=e^{\delta}$, where i is the annual effective rate of interest. The premium rate π is known as a 'benefit premium'; it is computed assuming that firms are risk neutral and that there are no transactions costs. In commercial practice, the benefit premium π will be increased to a contract premium G, $G>\pi$. The contract premium will contain provisions for expenses, profits and risk. The equivalence principle can be extended to include these elements.

The liability of the insurer, denoted by ${}_sV$, given survival to s, $s\geq 0$, would be given by the equivalence principle as

$${}_sV + \pi\int_s^\infty e^{-\delta(t-s)}s(t-s|s<T)\,dt = -b\int_s^\infty e^{-\delta(t-s)}s'(t-s|s<T)\,dt. \tag{2}$$

In words, the liability is the expected present value, also called the actuarial present value, of future benefits less the actuarial present value of future premiums. In this equation, the conditional

survivorship function is $s(t - s|s < T) = s(t) / s(s)$.
In eq. (1) for the premium rate and eq. (2) for the reserves, making benefits and premiums a function of time survived, b(*t*) and $\pi(t)$, creates no conflict with the equivalence principle. There are practical reasons for requiring $sV \geq 0$. This prevents a voluntarily withdrawing insured from leaving the insurer with a negative liability, a non-collectable asset.
As another special case, we now consider a life annuity with benefits at an annual rate *b* starting at retirement time *r*, funded by a continuously paid premium rate π paid during survival to *r*. Such a contact would be a building block of a pension plan. The equivalence principle yields

$$\pi \int_0^r e^{-\delta t} s(t)\, dt = b \int_r^\infty e^{-\delta t} s(t)\, dt, \tag{3}$$

allowing us to compute the premium rate π based on survivorship and interest information.
To illustrate how other contracts can be accommodated, we consider the life annuity case that also includes a so-called ‘return of premiums’ feature. With this feature, there is an additional benefit consisting of the accumulated premiums (with interest) that are paid at death before time *r*. The benefit side of formula (3) is increased by

$$-\pi \int_0^r \left(\int_0^t e^{\delta x} dx \right) e^{-\delta t} s'(t)\, dt = -\pi \left(s(r) \int_0^r e^{-\delta t} dt + \int_0^r e^{-\delta t} s(t)\, dt \right), \tag{4}$$

where the right-hand side is from an integration by parts. With this additional benefit, from formula (3) we have

$$\pi s(r) \int_0^r e^{-\delta t} dt = b \int_r^\infty e^{-\delta t} s(t)\, dt,$$

a result that might have been derived by general reasoning from the equivalence principle.
We return to the life annuity premium displayed in formula (3). The equivalence principle yields a reserve liability at time *s*, $0 \leq s$, of

$$_sV + \pi\int_s^r e^{-\delta(t-s)}S(t-s|s<T)\,dt = b\int_s^\infty e^{-\delta(t-s)}S(t-s|s<T)\,dt 0<s<r$$

$$_sV = b\int_s^\infty e^{-\delta(t-s)}S(t-s|s<T)\,dt r<s.$$

The key role played by the survival function and the assumed interest rate in these typical formulas is clear.

Reserves and premiums: short-term coverages

Short-term coverages include most individual property/casualty, health and group insurance policies. They are characterized by the reduced role of present values. In addition, the benefit amount is typically a random variable. Its value will depend in health insurance on the services provided, and in property insurance on the extent of the property damage. Premiums and reserves will continue to be determined by the equivalence principle. In the time period between the occurrence of a loss event and its settlement, available information about the loss event will determine reserve amounts.
The expected value of benefit payments for short-term coverages is given by

$$\pi = E\left(\sum_{i=1}^{N} X_i\right) = E\left(E\left(\sum_{i=1}^{N} X_i\right)|N=n\right) = \mu E(N),$$

where N denotes the random number of losses during the insurance period, X_i is the loss amount arising from loss i and $E\,X_i = \mu$.
If the distribution of N is Poisson and N and the loss amounts are independent, then

$$S = X_1 + \ldots + X_N \tag{5}$$

has a compound Poisson distribution. Clearly many distributions, such as the binomial or negative

binomial, could be used for the distribution of *N*.
The reserve liability for short-term coverages uses information about loss events and the loss reserve is

$$E(X_1 + \ldots + X_N | N = n) = nE(X) = n\mu$$

and *n* is the number of losses incurred.
Risk theory is the study of the distribution of total losses and the management of their inconvenient consequences. The earliest contributions to risk theory build on the model for long-term coverages. We start with loss variables

$$L_i = b_i e^{-\delta T_i} - \pi_i \int_0^{T_i} e^{-\delta s} ds,$$

and study the distribution of $S = L_1 + \ldots + L_n$. Here, T_i is a random variable representing the future lifetime of an individual. This study is known as individual risk theory because the variable *S* is based on *n* individual loss variables. If the loss variables are assumed to be mutually independent, then

$$Z = \frac{S - E(S)}{\sqrt{\mathrm{Var}(S)}}$$

will have, as a result of an extension of the central limit theorem, an approximate normal distribution, with mean zero and variance one. In contrast, the direct study of the distribution of *S* as in formula (5) is called collective risk theory. Approximating the distribution of *S* has been an active topic in actuarial research since early in the 20th century.

Experience adjustment: long-term coverages

Valuation of long-term coverages requires assumptions about the realizations about interest rates and mortality in the distant future. In this dynamic world it is almost certain that the results expected by an insurance system will not be obtained. For many contracts, it has become customary for insurers to make assumptions that many financial analysts would view as conservative for pricing at contract initiation. As better than anticipated experience is realized, excess funds are realized that can be directed to the insured in a mutual insurance organization or to owners of the insurance company. For the insured, these are additional (non-contractual) benefits; depending on the regulatory environment, these additional

benefits come in the form of dividends or bonuses.
Reconciling anticipated to actual experience is done periodically, not just at the conclusion of the contract. Because of this periodic reconciliation, recursion relationships are important tools for measuring and adjusting for deviations from expected results.
Specifically, let ${}_{s-1}F$ be the fund, possibly the insurance reserve, at the end of policy year *s*-1. Define ${}_sP$ to be the premium paid at the beginning of policy year *s*, E(${}_sB$) the expected benefits paid at the end of policy year *s* and ${}_sF$ the expected fund at the end of policy year *s*. We simplify and assume that $E({}_sB) = bq_s$. A basic recursive relationship is

$$({}_{s-1}F + {}_sP)(1+i) - bq_s = {}_sF\ p_s, \tag{5a}$$

where *i* is the expected annual interest rate and $p_s = 1 - q_s$. Formula (5) can be written as

$$({}_{s-1}F + {}_sP)(1+i) - q_s(b - {}_sF) = {}_sF. \tag{6}$$

If the actual experience yields i' and q_s', then formula (6) can be written as

$$({}_{s-1}F + {}_sP)\left(1+i'\right) - q_s'(b - {}_sF) = {}_sF + D, \tag{7}$$

where *D* is a deviation of actual from expected results. If $D > 0$, the amount might be paid to the insured in a mutual insurance organization or to owners of the insurance company.
Subtracting formula (6) from (7), yields

$$D = \left(q_s - q_s'\right)(b - {}_sF) + \left(i' - i\right)({}_{s-1}F + {}_sP). \tag{8}$$

The first term on the right-hand side of formula (8) is called the mortality contribution and the second term the interest contribution. Formulae used in practice also contain a term for the difference between expense loading and actual expenses.
To study life annuities, the general formula (6) can be modified during the benefit payment period to yield

$$_{s-1}F(1+i) - p_s b = p_s\,{}_sF. \tag{9}$$

Replacing the expected parameters with experience parameters, we have

$$_{s-1}F\left(1+i'\right) - p_s' b = p_s'\left({}_sF + D\right). \tag{10}$$

Subtracting formula (9) from formula (10) yields

$$_{s-1}F\left(i' - i\right) + \left(p_s - p_s'\right)\left(b + {}_sF\right) = p_s' D.$$

If $D > 0$, this expression could be the basis of a dividend to surviving annuitants.
These recursion relationships are also the basis for flexible coverages where premiums and benefits can be changed by the insured within contractual limits.

Experience adjustment: short-term coverages

In the first decade of the 20th century industrial accidents were a leading cause of death, a source of much litigation and a major social concern. The advent of workers' compensation insurance replaced litigation with a system based on defined benefits. Employers, in most cases, were required by statute to provide workers' compensation benefits. Because of great variation in the hazards faced in different industries and the lack of loss statistics, initial premiums were set by judgement. The goal was to develop a self-correcting rate estimation process that would also provide incentives to employers to improve industrial safety.
The solution came from the formula

$$(New\ rate) = Z(n) \times (observed\ average\ losses) + [1 - Z(n)] \times (Initial\ Rate),$$

where the credibility factor is $Z(n)$, n a measure of exposure and $0 \leq Z(n) \leq 1$. To provide intuition, consider the case where $Z(n) = 1$, known as the 'full credibility' case. Here, the employer's next period premium would consist entirely of observed average losses from the prior period. If the employer had introduced practices to improve industrial safety then this would be reflected in a lower premium. In contrast, consider the case where $Z(n) = 0$. Here, the premium would consist of an initial rate that presumably would reflect industry results but not the employer's actual experience. The case $Z(n) = 0$ is the standard for individual coverages. Many employers would fall in the intermediate case, $0 < Z(n) < 1$, known as 'partial credibility'. Premiums for employers in this category would reflect their own industrial safety records as well as benefit from the pooling of risks within an industry.

For the credibility factor, one typically requires $Z'(n) > 0$ and $Z''(n) < 0$. Thus, other things equal, employers with larger exposure (n) enjoy larger credibility but the rate of increase decreases with exposure. A typical credibility function is of the form $Z(n) = n / (n + k),\ k > 0$. The establishment of k with a satisfactory intellectual foundation has come from Bayesian statistics after its introduction into practice. The credibility idea for experience adjustments is now used in many short-term coverages.
Another type of insurance plan available for groups is known as 'stop-loss' or 'excess of loss' coverage. Large group insurance plans, usually based on employee groups, have distinctly different risk characteristics from individual policies. The sponsor, usually a large organization, is typically willing and able to absorb some variation in benefit payments. Only large and unexpected payments are financially inconvenient to the sponsor. The insurance company is paid to adjudicate and pay benefit claims, and to absorb large and inconvenient benefit payments. Typically, the sponsor maintains an internal account of losses known as an 'experience account'. This account records premiums as income and losses and expenses as expenditures.
We let X be the losses in an experience period and d be the stop loss amount (or d for 'deductible'). The experience account is charged for losses up to d. If $X > d$, then X-d is not charged to the experience account of the sponsor. A risk premium for this experience adjustment is charged on the basis of

$$\int_d^{\infty} (x - d)\, f(x)\, dx.$$

Model calibration: experience studies

The models introduced suggest that extensive work must be done in estimating survival functions in implementing long-term insurance models. For short-term models, the distribution of the number of losses N, per policy period, and the distribution of X, the loss amount, must be estimated. These efforts are in most applications special cases of statistical estimation.

These estimation projects are generally observational studies. The data come from insurance experience and the subjects have purchased insurance or gained insurance as an employee benefit. The use of general population statistics for insurance purposes has hazards because of potential biases. To illustrate, when studying annuitant mortality, it is well-known that mortality is substantially lower than the general population mortality. This is a selection bias issue; seldom do those in substandard health purchase a life annuity.

Rapid increases in the cost of health services and jury awards in some areas have increased the need to estimate time trends for the distribution of *X*, loss costs. Because of longer settlement time in some coverages, this estimation has become a major project in loss reserve determination. The rate of increase in health care costs in recent years has been such that estimates of the distribution of *X*, benefit amount random variable, using information from previous years would result in a distribution significantly to the left of the distribution for the current year. The rate of growth of health care costs is the most important single pricing decision for health insurance.

Model calibration: classification

The distributions that enter insurance models are all conditional distributions. Clearly, the distribution of *X*, loss amount, depends on the time, location and other facts surrounding the insurance loss incident. The distribution of *T*, time until death, in life insurance depends on a set of classification variables. The purpose of observing these classification variables is to increase the likelihood that the assumed distribution of *T* will be approximately realized.

The selection of these classification variables may be constrained by law and expense. For example, a determination of the degree of aggression of an applicant for automobile insurance might have a significant impact on the distribution of *N*, but the expense of collecting the information might be greater than its value in reducing variability.

Model calibration: financial economics

The critical role played by the force of interest δ in premiums and reserves for long-term coverage is clear. The use of an assumed force of interest for an extended period of time will lead, according to common experience, to serious deviations between actual and expected results. Options for moderating these deviations are numerous.

- A statistical model for the force of interest, estimated from past data, could be constructed and the equivalence principle extended to take expectations over the joint distribution of δ_T and *T*. The joint distribution might also be used to fix an interest rate risk loading into premiums to minimize the inconvenient consequence of variations in δ.
- Arrange the timing of investment cash flows to approximately match the expected cash flows from the insurance operations.
- Pass variations in interest earnings directly to the policy owner as indicated in the section on Experience adjustment, long-term coverages. The insured's account *F* would absorb variation in investment earnings.

- Use a program of financial derivative contracts to stabilize, for a price, variations in investment income.

Financial economics has not only enriched insurance mathematics by providing risk management tools for the investment risk in conventional insurance contracts, it has also created the possibility of absorbing many traditional insurance risks into special securities traded in worldwide investment markets. The idea is to use the capital in investment markets, and not just the capital held by insurance companies, to manage risk.

The idea of special securities with contractual payments that approximately match payments from an insurance system has already been developed for several coverages – for example, catastrophe bonds with modified payments following a catastrophe, fitting the definition of the security. A second example is a survivorship bond with regular coupon payments proportional to the number of survivors in a defined group. Such bonds could spread the risk if mortality improvement exceeds the capacity of the sponsor of a pension system.

The market for such special securities is determined, in part, from ideas in financial economics. Portfolio theory would predict that investors would seek securities that have cash flows that are not positively correlated with the regular business cycle. Tying security payments to natural disasters, such as earthquakes and hurricanes, might achieve the sought for independence.

See Also

- health insurance, economics of
- liability for accidents
- life tables
- mortality
- pensions
- present value
- social insurance

Bibliography

Booth, P., Chadburn, R., Cooper, D., Haberman, S. and James, D. 1999. *Modern Actuarial Theory and Practice*. London: Chapman and Hall/CRC.

Bowers, N., Gerber, H., Hickman, J., Jones, D. and Nesbitt, C. 1997. *Actuarial Mathematics*. Schaumburg, IL: Society of Actuaries.

Klugman, S., Panjer, H. and Willmot, G. 2004. *Loss Models: From Data to Decisions*, 2nd edn. New York: Wiley.

Panjer, H. (ed.) 1998. *Financial Economics: With Applications to Investments, Insurance and Pensions*.

Schaumburg, IL: Actuarial Foundation.

How to cite this article

Hickman, James C. and Edward W. Frees. "insurance mathematics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000284> doi:10.1057/9780230226203.0813

The New Palgrave Dictionary of Economics Online

intangible capital

Daniel E. Sichel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Intangible capital has played an increasingly important role in economic growth, although firm-level financial and national income accounting practices provide little information about intangibles and do not count many purchases of intangible capital as investment.

Keywords

economic growth; financial accounting; financial market valuations; growth accounting; information technology; intangible capital; labour productivity; national income accounting; National Income and Product Accounts (USA); total factor productivity

Article

Economists have long understood that advances in knowledge and technology play a crucial role in economic growth. An important recent contribution to this literature is research on the magnitude and role of intangible capital. As defined by Corrado, Hulten and Sichel (2005; 2006), intangible investment is expenditures by businesses that are intended to boost output in the future but that are not traditional, tangible physical capital; examples include outlays for computer software, research and development, training, brand equity, and improvements in organizational structure and efficiency.
Recent interest in intangible capital was generated by a sense in some quarters that official statistics may not be capturing the full dynamism of the US economy as well as by the resurgence of US productivity growth in the mid-1990s. That resurgence led many researchers, including Oliner and Sichel (2000; 2002), Jorgenson and Stiroh (2000), and Jorgenson, Ho and Stiroh (2002), to focus on the contribution of information technology (IT) to economic growth. And that focus on IT, as well as the run-up in equity valuations that occurred at about the same time, turned researchers' attention to intangible capital. Many analysts observed that firms using IT effectively did more than simply install it; they made sizable collateral investments to revamp their operations in order to exploit the new technologies. For example,

Walmart developed a more efficient supply chain, Dell linked demand and production more tightly, Amazon pioneered a new distribution channel, and Google and eBay developed entirely new businesses. In each case, the collateral investments consisted largely of expenditures on intangible inputs. Many observers believe that these intangible investments, as well as intangible investments that may not be tied to IT, are playing an increasingly important role in the economy.

Despite the apparent importance of investments in intangible capital, relatively little is known about these investments. At the firm level, financial accounting provides little information about such expenditures and the return earned by them. Moreover, these outlays are considered a current-period expense, not an investment creating an asset on the firm's balance sheet. Because of this lack of information, Lev (2004) argues that managers may make poor investment decisions and financial markets may incorrectly value firms and therefore may inefficiently allocate capital. At the level of the National Income and Product Accounts (NIPAs) used to measure gross domestic product (GDP) in the United States, historical practice has classified such expenditures as intermediate inputs, and thus they are not counted as investment in GDP. (The inclusion of business software as an investment in the NIPAs is a notable exception to this practice.) Moreover, the GDP accounts, like firm-level financial accounts, provide very little information about most intangible expenditures.

Research has begun to fill this gap with three broad approaches to measuring intangible capital. The first uses financial market valuations to gauge the value of intangible capital, inferring a measure of intangible capital from the gap between the market and book value of firms. As summarized in Hall (2005), such an estimate was quite large around 2000, about equal to the stock of tangible capital. At the firm level, Brynjolfsson and Hitt (2005) regress market value on capital and labour inputs as well as various proxies for intangible capital. Their work highlights the link between intangible investments and investments in computers, and suggests that intangible investments may exceed tangible investments in computers by as much as a factor of ten. Considerable controversy has surrounded estimates of intangible capital that are derived from financial market valuations.

The second broad category of research relies on other performance measures (such as productivity or earnings) to gauge the magnitude of intangible capital; for examples, see McGrattan and Prescott (2005), Cummins (2005), and Lev and Radhakrishnan (2005). Lev (2004) summarizes a methodology for estimating the value of intangibles at the level of individual firms, starting from earnings. This literature also finds a large role for intangibles.

The third broad category of research uses expenditure data to develop measures of intangible capital. Nakamura (1999; 2001; 2003) was the first to develop expenditure measures. Corrado, Hulten and Sichel (2005) expanded on Nakamura's work and more tightly integrated estimates of intangible investment with the NIPAs. Marrano and Haskel (2006) applied the methodology of Corrado, Hulten, and Sichel (2006) to the United Kingdom, and obtained similar results.

Corrado, Hulten, and Sichel (2006) classify business spending on intangibles into three broad groups: computerized information, innovative property and economic competencies. Computerized information consists mainly of computer software. Innovative property includes scientific R&D and non-scientific R&D such as product development expenditures in financial services and in the entertainment industry. Economic competencies include brand equity (advertising) and firm-specific resources such as training and organizational capital. Corrado, Hulten, and Sichel use a variety of data sources to develop time series of nominal expenditures for each category. These figures suggest that nominal intangible business

investment from 2000 to 2003 averaged $1.2 trillion per year, about $1 trillion of which was not counted as investment in the NIPAs.

This research highlights the magnitude and importance of intangibles but does not quantify their contribution to economic growth. This question is taken up in Corrado, Hulten, and Sichel (2006), which extends their earlier paper, and embeds intangibles in a conventional growth accounting framework. Specifically, Corrado, Hulten, and Sichel develop time series of the real stock of intangible capital for the United States, using their earlier estimates of investment in intangibles. According to their numbers, the nominal stock of intangible capital was about $3.6 trillion in 2003, about $3.1 trillion of which is not included in official measures. These figures imply that official measures may be understating the stock of business capital by roughly 20 per cent.

Corrado, Hulten, and Sichel (2006) embed their estimates of intangible capital into a standard growth accounting decomposition and present estimates for the period from 1973 to 2003 for the United States. They compare a decomposition based on data that exclude intangibles to one based on data that include intangible assets. Several important results emerge from this analysis. First, the inclusion of intangibles as investment boosts the estimated growth rate of labour productivity in the non-farm business sector by 10 to 20 per cent relative to a baseline case that completely ignores intangibles. Second, the contribution of intangibles to economic growth has increased dramatically since 1995, and including intangibles has a considerable effect on the composition of the mid-1990s pickup in labour productivity growth. Third, once intangibles are included, greater use of capital (including both tangible and intangible capital) becomes a more important source of growth. This contrasts with the traditional result (when intangibles are largely excluded), where total factor productivity – the residual after accounting for the contributions from labour and capital – plays a larger role. Finally, the majority of the contribution of intangibles comes from categories of intangibles that have received relatively little attention in the past, such as non-scientific R&D and firm-specific resources. Scientific R&D – perhaps the most studied and most ‘traditional’ category of intangibles – accounts for only about one-tenth of the contribution of intangibles to labour productivity growth.

Taken together, the research indicates that business investment in intangible capital is quite sizable and has played an important role in the US economy. Moreover, these results indicate that both firm-level and national income accounting practice miss some important features of economic activity. Nevertheless, the quantitative estimates discussed here are clearly provisional, and this area appears to be a fruitful one for further research.

See Also

- growth accounting
- intellectual property

Bibliography

Brynjolfsson, E. and Hitt, L. 2005. Remarks. In Corrado, Haltiwanger and Sichel (2005).

Corrado, C.A., Haltiwanger, J. and Sichel, D., eds. 2005. *Measuring Capital in the New Economy*.

Chicago: University of Chicago Press.

Corrado, C.A., Hulten, C.R. and Sichel, D.E. 2005. Measuring capital and technology: an expanded framework. In Corrado, Haltiwanger and Sichel (2005).

Corrado, C.A., Hulten, C.R. and Sichel, D.E. 2006. Intangible capital and economic growth. Finance and Economics Discussion Series Paper No. 2006–24. Washington, DC: Federal Reserve Board.

Cummins, J.G. 2005. A new approach to the valuation of intangible capital. In Corrado, Haltiwanger and Sichel (2005).

Hall, R.E. 2005. Remarks. In Corrado, Haltiwanger and Sichel (2005).

Jorgenson, D.W. and Stiroh, K.J. 2000. Raising the speed limit: U.S. economic growth in the information age. *Brookings Papers on Economic Activity* 2000(1), 125–211.

Jorgenson, D.W., Ho, M.S. and Stiroh, K.J. 2002. Projecting productivity growth: lessons from the U.S. growth resurgence. *Federal Reserve Bank of Atlanta, Economic Review* 87(3), 1–14.

Lev, B. 2004. Sharpening the intangibles edge. *Harvard Business Review* 82(6), 109–16.

Lev, B. and Radhakrishnan, S. 2005. The valuation of organizational capital. In Corrado, Haltiwanger and Sichel (2005).

McGrattan, E.R. and Prescott, E.C. 2005. Taxes, regulations, and the value of U.S. and U.K. corporations. *Review of Economic Studies* 72, 767–96.

Marrano, M.G. and Haskel, J. 2006. How much does the UK invest in intangible assets? Working Paper No. 578, Department of Economics, Queen Mary College, University of London.

Nakamura, L. 1999. Intangibles: what put the *new* in the new economy? Federal Reserve Bank of Philadelphia, *Business Review* (July–August), 3–16.

Nakamura, L. 2001. What is the US gross investment in intangibles? (At least) one trillion dollars a year! Working Paper No. 01–15, Federal Reserve Bank of Philadelphia.

Nakamura, L. 2003. The rise in gross private investment in intangible assets since 1978. Mimeo, Federal Reserve Bank of Philadelphia.

Oliner, S.D. and Sichel, D.E. 2000. The resurgence of growth in the late 1990s: is information technology the story? *Journal of Economic Perspectives* 14(4), 3–22.

Oliner, S.D. and Sichel, D.E. 2002. Information technology and productivity: where are we now and where are we going? *Federal Reserve Bank of Atlanta Economic Review* 87(3), 15–44.

How to cite this article

Sichel, Daniel E. "intangible capital." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000299> doi:10.1057/9780230226203.0814

The New Palgrave Dictionary of Economics Online

integrability of demand

Donald W. Katzner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Integrability of demand arguments start with consumer demand functions having properties that would be implied by constrained utility maximization were they generated from that source. Using a process of mathematical integration, the arguments then proceed to demonstrate the existence of utility functions from which those demand functions could be derived.

Keywords

demand function; integrability of demand; marginal rate of substitution; revealed preference; Slutsky substitution functions; utility function

Article

The lines of reasoning linking individual (ordinal) utility functions (or preference orderings) to individual demand functions run in both directions. Progressions from the former to the latter often begin with assumptions about the characteristics of a consumer's utility function and the requirement that he or she always chooses so as to maximize utility subject to a budget constraint, and then go on to derive the demand functions and the properties of those demand functions that logically ensue from such premises. Depending on context, certain of the properties of the demand functions so derived are expressed in differential terms (that is, symmetry and negative definiteness of matrices of Slutsky substitution functions where the latter are defined) or in discrete revealed preference form (for example, weak and strong axioms of revealed preference). The reverse course takes the individual's demand functions and their properties as given and determines the existence of a utility function from which, upon constrained maximization, the original demand functions could have been generated. In this second case, when the starting point includes the differential rather than revealed preference properties of demand, the argument often involves (in part) the integration of a system of one or more differential equations. Hence the name ‘integrability of demand’ affixed to it.
There are several ways to structure an integrability of demand argument. Perhaps the most

straightforward approach (the only one considered in detail here) is simply to backtrack over the path that yields demand functions from utility functions via the theorem of Lagrange on maximization subject to constraint. That path may be summarized as follows. Begin with a utility function $u = u(x)$ defined over the commodity space $\{x: x \geq 0\}$, where $x=(x_1,\ldots,x_I)$ is a vector of quantities of commodities x_i and $x \geq 0$ means $x_i \geq 0$ for every $i=1,\ldots, I$. Let $u_i(x)$ be the partial derivative of the utility function u with respect to its ith argument. For each vector $(p, m)>0$, where $p=(p_1,\ldots, p_I)$, p_i is the price of good i, and m is a scalar denoting the consumer's income, vectors $x>0$ that maximize $u(x)$ subject to the budget constraint $\Sigma_{i=1}^{I} p_i x_i = m$ are, according to Lagrange's theorem, characterized by

$$\frac{p_i}{p_I} = \frac{u_i(x)}{u_I(x)}, \qquad i = 1, \ldots, I-1, \tag{1}$$

$$\frac{m}{p_I} = x_I + \sum_{i=1}^{I-1} \frac{u_i(x)}{u_I(x)} x_i. \tag{2}$$

Equations (1) state that, at a constrained maximum, the marginal rates of substitution or the negatives of the partial derivatives of indifference functions equal the price ratios, and equation (2) is a form of the budget constraint. Equations (1) and (2) together may be thought of as a system of inverse functions which are solved to secure demands x_i as functions h^i, of prices and income:

$$x_i = h^i\left(\frac{p_1}{p_I}, \ldots, \frac{p_{I-1}}{p_I}, \frac{m}{p_I}\right), \quad i = 1, \ldots, I. \tag{3}$$

Evidently, (3) may be written in the equivalent form

$$x_i = H^i(p, m), \quad i = 1, \ldots, I,$$

where $H^i(p, m) = h^i(p_1/p_I, \ldots, p_{I-1}/p_I, m/p_I)$ and H^i is homogeneous of degree zero. Of course, sufficient properties have to be imposed on u so as to ensure the existence of a usually unique constrained maximizing x for each $(p, m)>0$, and these properties, in turn, imply the well-known characteristics of the h^i or the H^i.

Consider now an integrability of demand argument that reverses the above steps. Start with the demand functions (3) having all of the properties that would be implied by constrained utility maximization were they determined as previously described. The aim is to show the existence of a utility function generator of these demand functions. Clearly, for this latter utility function to generate the h^i, it must exhibit properties such as those stated above that yield unique constrained maxima. Backtracking from (3), solve for price–price and income–price ratios as functions, g^i, of x:

$$\frac{p_1}{p_I} = g^i(x), \quad i = 1, \ldots I-1, \tag{4}$$

$$\frac{m}{p_I} = g^I(x). \tag{5}$$

If the h^i are to be derivable from constrained utility maximization, then the g^i of (4) should indicate the negatives of the partial derivatives of appropriate indifference functions and g^I should be related to the budget constraint; that is, equations (4) should correspond to equations (1), and equation (5) to equation (2). But to say that the g^i are the negatives of the partial derivatives of indifference functions means that

$$\frac{\partial x_I}{\partial x_i} = -g^i(x), \quad i = 1, \ldots, I-1. \tag{6}$$

Thus, at every $x>0$, the ‘slopes’ of the indifference surface through x in the direction of each of the coordinate axes are given by the g^i, for $i=1,\ldots,I-1$. Integrating the differential equation system (6) ‘fits’

all of these slopes for each surface together to form an indifference map from which a utility function is deduced. (It is possible to integrate alternative, though related, systems of differential equations which yield the utility function directly.) It should be noted, however, that the mathematics employed in this integration process usually shows only that such an indifference map, and hence a utility function, exists and typically does so without providing the means to specify the exact forms it will take. Lastly, the appropriate general characteristics of this utility function and the fact that its constrained maximization produces the original demand functions (3) are established.

Naturally, the properties of the demand functions h^i are crucial for such an integrability of demand argument to hold up. Among other things, these properties must permit the inversion of the h^i into the g^i and must ensure that the integration step can be carried out. Invertibility means that the h^i specify a 1–1 correspondence between values of the vectors $x=(x_1,\ldots, x_I)$ and $(p_1/p_I, \ldots, p_{I-1}/p_I, m/p_I)$. For the integration of (6) it is necessary that the g^i be continuous and, when $I>2$, that a certain 'integrability' condition be satisfied. This guarantees that at least one indifference surface passes through every $x>0$. To make certain that no more than one indifference surface passes through each x, a Lipschitz condition has to be in force. It turns out that the g^i are continuous as long as the h^i are continuous, that the integrability condition is equivalent to the existence and symmetry, for all $(p, m)>0$, of the matrix of Slutsky substitution functions, and that the Lipschitz condition is implied if certain partial derivatives of the g^i are bounded. All of these properties of demand functions except the last two are derivable from the constrained maximization of utility functions that are twice continuously differentiable, increasing, strictly quasi-concave, and whose indifference surfaces do not touch the boundaries of the commodity space. (Although such utility-function characteristics imply symmetry of the matrix of Slutsky substitution functions, they do not guarantee that those functions, and hence the matrix, will be defined everywhere.) Even so, the properties of demand functions obtained from such utility functions are still 'roughly' sufficient to support the integrability of demand argument outlined above.

Problems arise when the properties of demand functions are derived from utility functions with modified characteristics. For example, the previously mentioned 1–1 correspondence may not appear in the h^i and hence invertibility from the h^i to the g^i may break down. In such a situation it is possible to restructure the integrability of demand argument to avoid the invertibility issue at the level of the h^i altogether. Since it turns out that the demand functions $H^i(p, m)$ may also be viewed as partial derivatives with respect to p_i of the expenditure or income compensation function (obtained in the progression from utility to demand by minimizing expenditure for a given level of utility)

$$m = E(p, \mu),$$

where μ varies over all utility levels and p ranges over all vectors $p>0$, this is accomplished by integrating the system

$$\frac{\partial m}{\partial p_i} = H^i(p, m), \quad i = 1, \ldots, l, \tag{7}$$

and converting the resulting expenditure function into a utility function. Once again the appropriate characteristics of the derived utility function have to be established, constrained maximization of it has to produce the given H^i, and enough properties of the H^i need to be present to sustain the argument. Antonelli (1886) is usually credited with introducing economists to the integrability of demand argument. He began with the functions g^i and obtained a utility generator by integrating a system of differential equations related to (6). Many years later in a mathematical appendix, Samuelson (1950) inverted the h^i and then secured an indifference map by integrating another differential equation related to (6). In between, Antonelli's work seems to have been almost forgotten. Fisher (1892) independently ‘rediscovered’ the integrability problem in his doctoral dissertation, and various aspects of it were taken up subsequently by Pareto (1906a; 1906b), Volterra (1906), Allen (1932), Georgescu-Roegen (1936), Wold (1943; 1944), and others. It is interesting that Volterra's contribution was to point out that, in Pareto's initial (1906a) discussion of integrability for the case of more than two goods (that is, in the first edition of his *Manuale*), the integrability condition had been conspicuously omitted. More detailed history is given by Samuelson (1950) and Chipman et al. (1971, intro. to Part II). Hurwicz and Uzawa (1971) were the first to structure an integrability of demand argument based on the integration of (7).

Bibliography

Allen, R. 1932. The foundations of a mathematical theory of exchange. *Economica* 12, 197–226.

Antonelli, G. 1886. On the mathematical theory of political economy. In *Preferences, Utility, and Demand*, ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich, 1971.

Chipman, J., Hurwicz, L., Richter, M. and Sonnenschein, H., eds. 1971. *Preferences, Utility, and Demand*. New York: Harcourt Brace.

Fisher, I. 1892. Mathematical investigations in the theory of value and prices. *Transactions of the Connecticut Academy of Arts and Sciences* 9(July), 1–124. Reprinted, New York: Augustus M. Kelley, 1965.

Georgescu-Roegen, N. 1936. The pure theory of consumer's behavior. *Quarterly Journal of Economics* 50, 545–93.

Hurwicz, L. and Uzawa, H. 1971. On the integrability of demand functions. In *Preferences, Utility, and Demand*, ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich.

Katzner, D. 1970. *Static Demand Theory*. New York: Macmillan.

Pareto, V. 1906a. *Manuale di economia politica*. Milan: Societa Editrice Libraria.

Pareto, V. 1906b. Ophelimity in nonclosed cycles. In *Preferences, Utility, and Demand*, ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich, 1971.

Samuelson, P. 1950. The problem of integrability in utility theory. *Economica* 17, 355–85.

Volterra, V. 1906. Mathematical economics and Professor Pareto's new manual. In *Preferences, Utility, and Demand*, ed. J. Chipman et al. New York: Harcourt Brace, Jovanovich, 1971.

Wold, H. 1943. A synthesis of pure demand analysis I, II, *Skandinavisk Aktaurietidskrift* 26, 85–118, 221–63.

Wold, H. 1944. A synthesis of pure demand analysis III, *Skandinavisk Aktaurietidskrift* 27, 69–120.

How to cite this article

Katzner, Donald W. "integrability of demand." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000237> doi:10.1057/9780230226203.0815

The New Palgrave Dictionary of Economics Online

intellectual property, history of

Zorina Khan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The evolution of patents and copyrights followed different paths over time and across countries. Initially, intellectual property rules were endogenously determined according to social and economic priorities in each society. International patent laws subsequently were heavily influenced by early American policies that favoured the rights of original inventors. By contrast, US copyrights were among the weakest in the world; international copyright laws converged towards European doctrines that were based on non-economic rationales for inherent authors' rights. The intellectual property system in the 21st century therefore constitutes an anomaly, since previously no country simultaneously adhered to strong patent rights and strong copyrights.

Keywords

barriers to entry; copyright; fair use; innovation; intellectual property; intellectual property rights; patents; technology; total factor productivity

Article

Intellectual property rights primarily have their origins in 15th-century monopoly privileges granted in Europe. Specific features of these rights of exclusion varied enormously and some constituted broad national claims that existed in perpetuity. By the 18th century, such differentiated privileges had evolved into standardized legal rights whose boundaries were delimited by statute. Most notably, the British Statute of Monopolies (1624) and the Statute of Anne (1710) established the longest continuous intellectual property system in existence. In Europe the philosophy and enforcement of intellectual property laws, the structure of patent and copyright systems, and the resulting patterns of invention (broadly defined to include technological and cultural creations) were all consistent with the oligarchic structure of these societies.

European patents were viewed as 'pernicious monopolies', which had to be narrowly interpreted, monitored, and restricted. This perspective was reinforced by the grant of patents to anyone who paid the exceedingly high fees, regardless of whether they were true inventors. The Crown reserved the right to expropriate any innovations that it wished, and kept others secret. Few provisions were made to ensure ready access to information. The legal system was biased against patents in general, and incremental improvements in particular. High transactions and monetary costs, as well as the prevailing prejudices towards non-elites, combined to create barriers to entry that discouraged the poor or disadvantaged from making contributions to technological innovation. Markets in patent rights and in patented inventions were thin and risky. As a result, trade secrecy probably played a more prominent part in protecting new discoveries, diffusion was certainly inhibited, the distribution of inventors and inventions was skewed, and potential inventors faced a great deal of uncertainty. The elites who were privileged by these biases had little inducement to adopt institutional reforms that might generate social benefits at their expense. Administrators and patent agents lobbied against amendments and many had to be compensated for their lost rents before the system could be revised. Thus, despite their inefficiencies, patent rules and standards in both France and England remained essentially unchanged for stretches of over 100 years. In Britain, patent grants favoured a narrow range of capital-intensive industries and unbalanced growth paths. Clearly, despite these drawbacks, European economies still experienced industrialization and expansion; nevertheless, total factor productivity gains were quite modest and Britain was unable to sustain its initial advantage. Indeed, the record for Britain and other countries suggests that patent systems and their specific rules and standards had a significant effect. As Figure 1 shows, when Britain reformed its laws in line with the United States in 1852 and 1883, patenting rates immediately increased. Similarly, Swiss patent reforms in the 1880s and Taiwanese revisions in the 1980s changed the rate and direction of their inventive activity (Khan, 2005; Lo, 2005).

Figure 1
Patents per capita issued in Britain and the United States, 1790–1860. *Source*: Khan (2005).

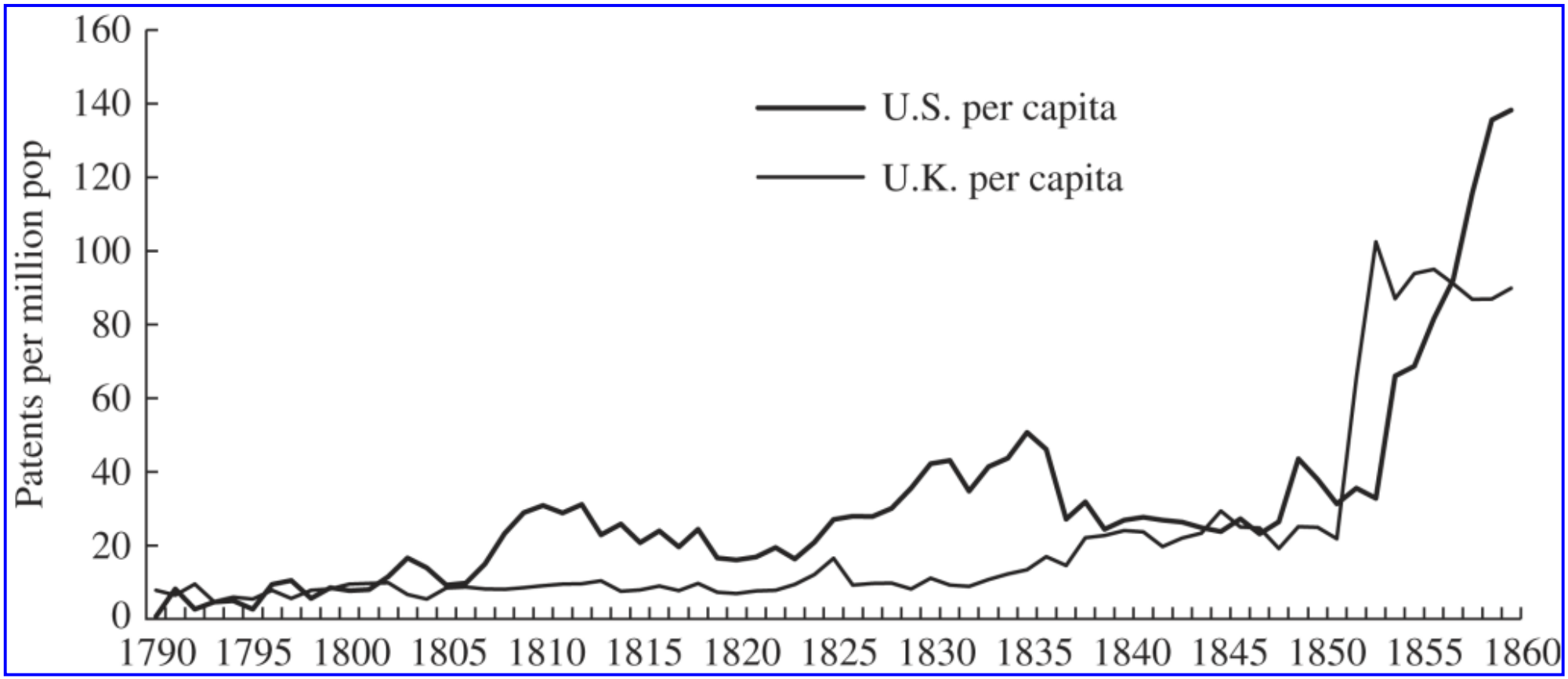

In the United States policymakers were well aware of the European experience. They carefully weighed the grant of intellectual property rights against alternative strategies such as state subsidies and prizes. Legislators did not shrink from novel approaches, which they estimated would increase social welfare, regardless of how great the popular outcry. In accordance with the US Constitution, the utilitarian objective of the intellectual property system was to promote the public welfare. Patent and copyright laws were clearly distinguished in separate statutes in 1790, and developed along diametrically different lines based on a rational assessment of their costs and benefits.

The leading industrial nations acknowledged that patent rights might increase the rate of invention, but it was less conventional to propose that the background or the identity of inventors was irrelevant to their productivity. The US patent system exemplified one of the country's most democratic institutions, offering secure property rights to true inventors, regardless of colour, marital status, gender, or economic standing. Patent data, when linked to biographical information, show that the expansion of markets and profit opportunities stimulated increases in inventive activity by attracting wider participation from relatively ordinary individuals. The roster of patentees included not only scientists and engineers, but also senators, schoolteachers, housewives, and even economists. The characteristics and patterns of patenting for American ‘great inventors’ were strikingly similar to those of ordinary patentees, unlike Europe where inventors were much more likely to be drawn from the elites.

Such patterns were due in part to the conscious design of US patent institutions to ensure open access. These included transparent rules and administration, explicit measures for the diffusion of information, low fees, protection of the rights of the first and true inventor, a centralized examination system, and a legal system that balanced the rights of patentees with social welfare. American judges understood that secure private property rights and market competition comprised effective counters to oligarchical tendencies. Unlike the situation England, where the Crown reserved the right to expropriate inventions, in the United States even federal government claims could not trump the patentee's property right. The examination system ensured that all inventors were able to secure the services of professional examiners at minimal cost. Patents helped transform inventive ideas into tradeable assets, and this securitization of invention enhanced market efficiency.

The second industrial revolution from 1870 to 1920 was a transitional period that hinted at future changes in the nature and organization of technology. This era is usually characterized as the age of professional, science-based invention conducted by teams in research laboratories. Indeed, formal college education, human capital accumulation, and

financial capital mobilization through corporate ties became more important, but relatively uneducated rural inventors were no less likely to produce valuable inventions. By the 1920s the rate of assignments sharply increased (Figure 2), even as patenting per capita declined (Figure 3), in part because inventive activity was increasingly conducted within corporations that appropriated returns through alternative strategies. Figure 3 indicates that per capita patenting by US residents in the 21st century remains lower than during the second industrial revolution, whereas the so-called ‘patent explosion’ after the 1980s was largely the result of increases in grants to foreign inventors.

Figure 2

US patents assigned at time of issue as a proportion of all patents granted, 1840–1930. *Source*: Khan (2005).

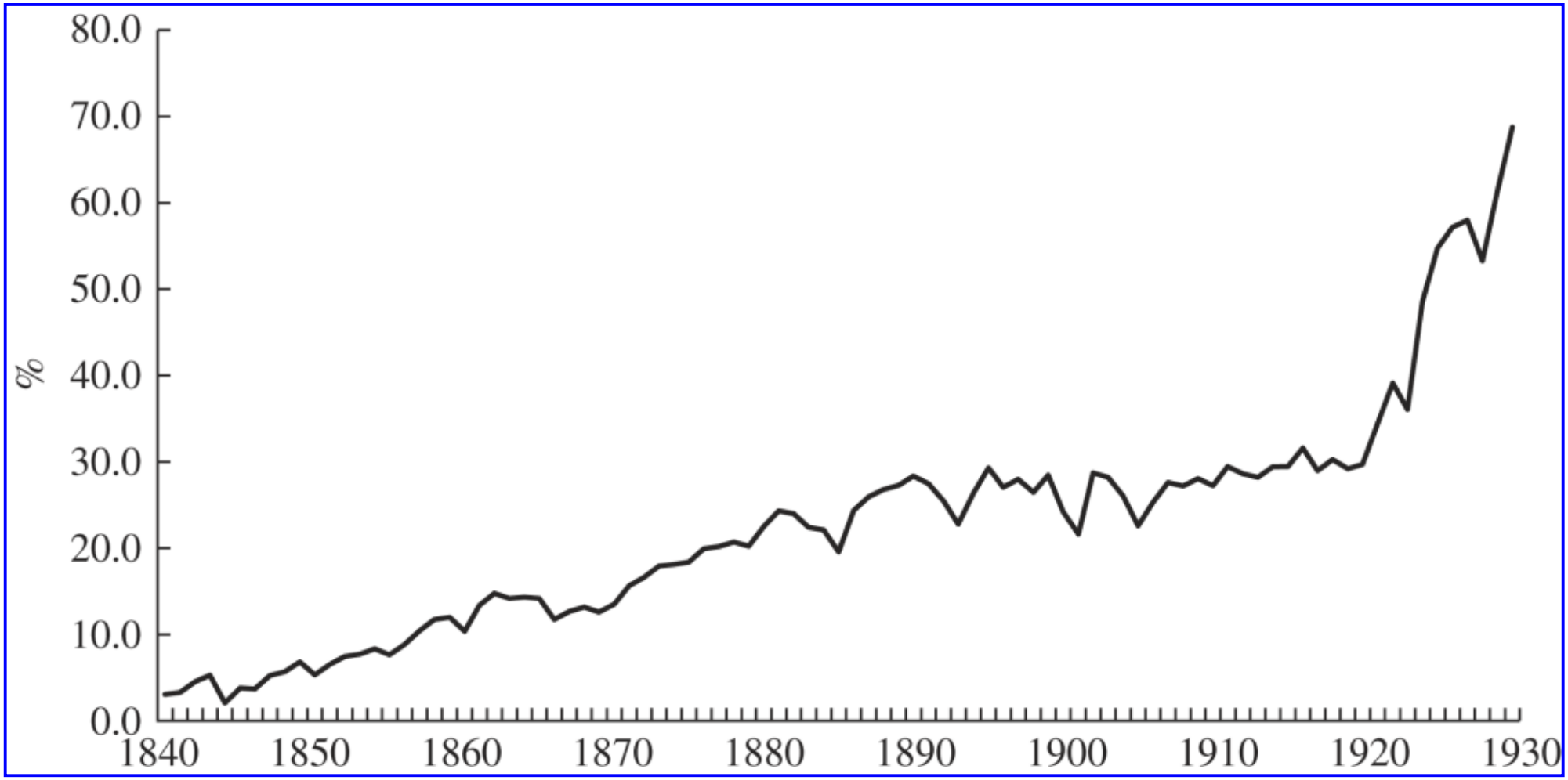

Figure 3

Per capita patenting: total US patents and patents issued to domestic residents, 1850–2000. *Source*: US Patent Office.

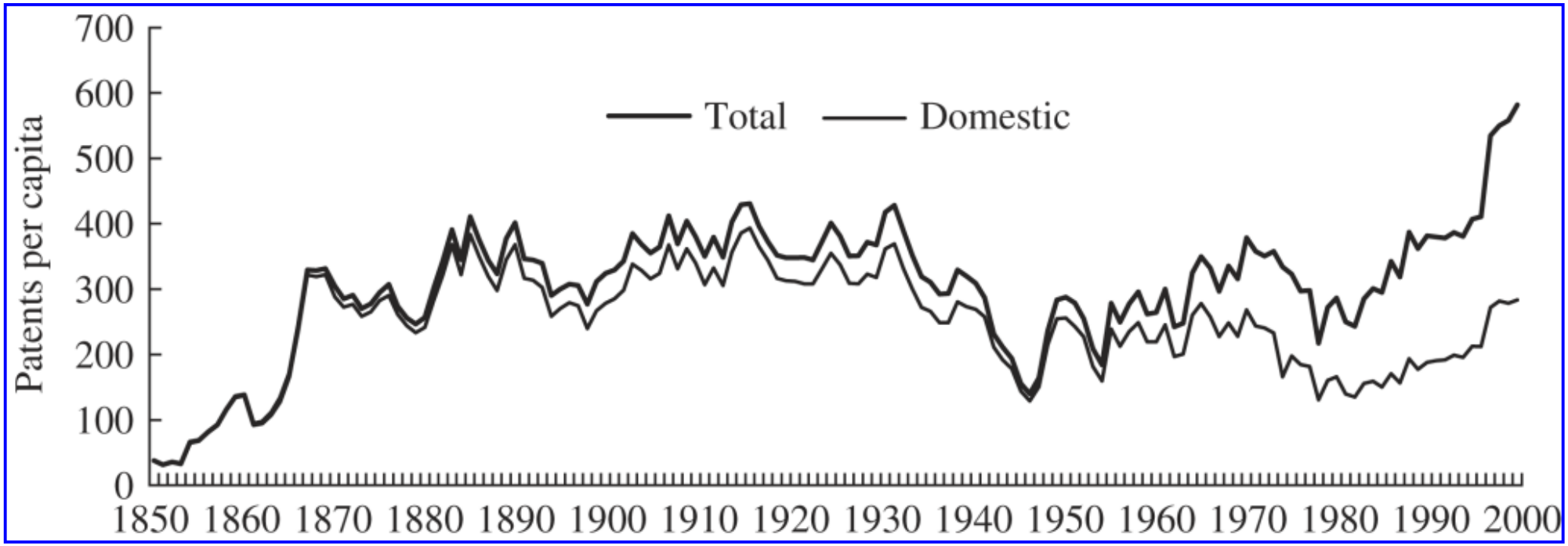

The US patent system was soon acknowledged as the most advanced in the world, and other countries drew causal connections between American achievements and its strong protection of patent property. Follower countries such as Germany and Japan patterned their own patent regime after the American model, but they introduced measures that addressed the particular needs of their own societies. These included the likelihood that patents would predominantly be granted to foreigners, the wish to raise revenues, and the need to foster domestic ingenuity. Their patent policies incorporated exemptions to protect social welfare in crucial industries such as food and pharmaceuticals, and restricted monopolistic tendencies through compulsory licensing and working requirements. Still, despite resistance from follower nations, patent harmonization over the 19th and 20th centuries converged towards the American model.

Copyrights

However much they praised and emulated US patent policies, other countries failed to understand the rationale for its copyright policies. The intellectual property clause of the US Constitution was the common source of both patent and copyright doctrines, and the same individuals were responsible for their formulation and implementation. American patent and copyright policies differed precisely because the objective of both systems was to promote the general welfare. This objective required a judicious balancing of private and public interests, the weighing of costs and benefits, and estimations of incentives and outcomes. Interests, costs and incentives differed across technical inventions and cultural goods, and also altered over time. Intellectual property adapted endogenously to meet these changing circumstances in a way that contrasted directly with the institutional sclerosis in Europe.
The rationale for US copyrights was not based on European notions of inherent rights of personhood but, rather, on purely pragmatic and utilitarian grounds. Instead of a bona fide property right, American copyrights often mimicked more limited legal mechanisms such as contract, trade restraint or even liability rules. Americans viewed copyright trade-offs with greater concern. First, the economic processes that produced cultural goods differed from technological innovations: many copyrighted items might be produced even in the absence of financial incentives because their producers could benefit from ancillary returns such as enhanced reputations or greater demand for complementary goods. Second, the risk of unwarranted monopolies was higher, because cultural goods incorporated ideas that belonged to the public domain in ways that made it difficult to distinguish between the contributions of the author and those of society in general. Third, the enforcement of copyright had more serious implications for a democratic society. Restrictions on free diffusion could result in significant social costs in terms of knowledge, education and free speech, in ways that promised to bolster the narrow redistributive claims of elites and interest groups. Although policymakers protected property rights, their primary objective was not to benefit authors or publishing companies per se, so the advantages of a privileged few were circumscribed in order to protect the public domain.
It is, therefore, unsurprising that throughout US history patents were treated differently from copyrights. The first copyright statute granted protection to both ‘authors and proprietors’ for the instrumental purpose of learning, whereas only the first and true inventor could claim patent rights. Similarly, for much of the 19th century, work-for-hire doctrines led to

weak employee rights in the case of copyrights, but not in the case of patents. Copyrights were administered in a registration system and were overturned if authors did not strictly comply with the rules; since 1836 patents were granted through an examination system and could not be revoked except for fraud. Patent policies were hostile to compulsory licences and unauthorized use of patent rights. By contrast, US copyright laws enshrine the world's most pervasive ‘fair use’ doctrines, which allow free unauthorized access for socially justifiable purposes (such as academic research and education) if such access did not significantly reduce the author's returns.

Although they excelled at pragmatic contrivances, 19th-century Americans were advisedly less sanguine about their efforts in the realm of music, art, literature, and drama, and so the USA was initially a net debtor in flows of material culture from Europe. The first copyright statute recognized this when it authorized international copyright piracy that persisted for a century. Proposals to reform the law were repeatedly brought before Congress and rejected because the net effects for Americans would be ‘on the wrong side of the ledger’. It was only in 1891, when the balance of trade in cultural goods was more favourable to the United States, that an international copyright law was finally passed. Even then, the bill almost failed, and its passage required protectionist exemptions in favour of American workers and printing enterprises that remained in place until 1986. This policy was a dramatic departure from the evolution of international copyright laws in European countries. Early on in the 19th century France accorded national treatment to all countries, and led the movement for international harmonization towards strong copyright laws, which culminated in the 1886 Berne Convention. While it took a leadership role in patent conventions, the United Stated did not enter the Berne Convention until 1988, and it still has not completely complied with its provisions.

Today intellectual property rights are at the forefront of economic policy issues for developed and developing countries alike. Questions from four centuries ago are still current, ranging from the philosophical underpinnings of intellectual property to proposals for the abolition of all such rights. A 19th-century economist could assess contemporary policies that substituted tariffs and taxes for revenues to copyright owners, and would have been equally familiar with analyses about whether uniformity in intellectual property rights across countries benefited global welfare. However, throughout their history, patent and copyright regimes have accommodated ‘new eras’ that were no less significant and contentious for their time than the ‘digital dilemmas’ of the 21st century.

Economic history indicates that intellectual property institutions best stimulated early economic growth when they enabled flexible endogenous responses to socio-economic circumstances. However, the movement to harmonize patent and copyright laws encouraged a ‘race to the top’: it arose from two separate sources that culminated in stipulations for a system of uniformly strong patents and strong copyrights regardless of the level of economic development. Such a system did not exist anywhere in the world before the late-20th century, when countries enjoyed greater freedom to choose appropriate institutions. The more limited menu of choices today – especially for developing countries but even in the United States – constitutes an economic and historical anomaly.

See Also

- intellectual property
- patents

Bibliography

Bugbee, B. 1967. *The Genesis of American Patent and Copyright Law*. Washington, DC: Public Affairs Press.

Dutton, H. 1984. *The Patent System and Inventive Activity during the Industrial Revolution, 1750–1852*. Manchester: Manchester University Press.

Khan, B.Z. 2005. *The Democratization of Invention: Patents and Copyrights in American Economic Development*. New York and Cambridge: NBER and Cambridge University Press.

Khan, B.Z. and Sokoloff, K.L. 2004. Institutions and democratic invention in 19th-century America. *American Economic Review* 94, 395–401.

Khan, B.Z. and Sokoloff, K.L. 2001. The early development of intellectual property institutions in the United States. *Journal of Economic Perspectives* 15(3), 233–46.

Lamoreaux, N.R. and Sokoloff, K.L. 2001. Market trade in patents and the rise of a class of specialized inventors in the nineteenth-century United States. *American Economic Review* 91, 39–44.

Lo, Shih-tse. 2005. Strengthening intellectual property rights: the experience of the 1986 Taiwanese patent reforms. Ph.D. thesis, University of California.

MacLeod, C. 1988. *Inventing the Industrial Revolution, the English Patent System, 1660–1800*. Cambridge: Cambridge University Press.

Sokoloff, K.L. 1988. Inventive activity in early industrial America: evidence from patent records, 1790–1846. *Journal of Economic History* 48, 813–50.

How to cite this article

Khan, Zorina. "intellectual property, history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000256> doi:10.1057/9780230226203.0817

The New Palgrave Dictionary of Economics Online

intellectual property

Michele Boldrin and David K. Levine

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Intellectual property refers to patents, copyrights, trademarks and other forms of ownership of ideas. It results in monopoly power that has significant consequences for discouraging as well as encouraging innovation and growth. The discouragement effect is especially important when ideas are used as building blocks for other ideas. The economics literature has examined the need for intellectual property; optimal systems of intellectual property; the optimal duration of intellectual property; how innovation takes place in the absence of intellectual property; and the rent-seeking behaviour induced by intellectual property.

Keywords

Arrow, K.; copyright; dynamic vs static efficiency; first-mover advantage; innovation, competitive; monopoly vs incentive to innovate; no-compete contract clauses; non-disclosure agreements; patent law; patent races; patents; Plant, A.; rent seeking; Schumpeter, J.; shrink-wrap agreements; Stigler, G.; trademarks; transaction costs

Article

Intellectual property refers to patents, copyrights, trademarks and other forms of ownership of ideas. While property and ownership are not controversial topics among economists, patents and copyrights have long been. By contrast, trademarks – serving merely to identify individuals and businesses – are not controversial. The economic analysis of patents and copyrights applies also to a variety of private contractual arrangements that are used to enforce ‘intellectual property’ such as non-disclosure agreements, no-compete contract clauses, and software shrink-wrap agreements.
The controversy surrounding patents and copyright has both theoretical and policy relevance. The theoretical relevance arises because models of economic growth, trade, and industrial regulation all put innovative activity at their core. Two fundamental views of innovation have been advanced. In the first,

which can be traced back to Arrow (1962) and has recently been developed by Romer (1986; 1990), it is abstract ideas that matter. These are produced subject to fixed costs that cannot be recouped under competition because, once discovered, ideas are non-rivalrous and infinitely reproducible at constant marginal cost, which is often treated as equal to zero. In the second, traces of which are found in Schumpeter (1911), Plant (1934) and Stigler (1956), and is formalized by Boldrin and Levine (2003), it is the concrete embodiment of copies of ideas that has economic value. Embodied ideas are initially characterized by indivisibility, but their reproduction is limited by capacity constraints; the latter generate competitive rents that may cover the indivisibility cost, hence ideas can be produced and traded under competition.

From both perspectives, ‘intellectual property’ is a grant of monopoly power over the right to make copies of ideas, and not simply the extension of ‘normal’ property rights to the realm of ideas. The first view argues in favour of intellectual property not because it is property and serves to protect the value of individual investment, but rather because monopoly over ideas can be a good thing. This argument has recently been linked, by such authors as Aghion and Howitt (1992) or Grossman and Helpman (1991), to a Schumpeterian theme (Schumpeter, 1942). It posits a trade-off between ‘static efficiency’, which requires competition, and ‘dynamic efficiency’, which can be achieved only through technological progress driven, in turn, by the desire to acquire a monopoly. In this view it is monopoly power that drives the innovative process.

On the policy side, intellectual property has become controversial largely because of three developments. The first is the high price and restrictive policies of pharmaceutical companies, for example with regard to AIDS drugs. Second is the damaging impact of intellectual property on the growth perspectives of the less developed countries, especially when they are denied free trade with the more developed ones unless they adopt strict standards for intellectual property. Third is the impact of the internet on the ‘piracy’ of music, books and movies.

On one side of the policy debate stand those who benefit from existing monopolies and are eager to protect their way of doing business, arguing typically that their ‘property’ should be protected from ‘theft’. On the other side is a broad array of people who resent having the free use of their copies of ideas restricted by creators.

The central issue is whether the monopoly power achieved through copyrights and patents is truly necessary, in the words of the US Constitution, ‘To promote the progress of science and useful arts’ or whether it in fact hinders progress and innovation.

Optimal systems of intellectual property

A key question is what an optimal system of intellectual property looks like. Most research has focused on patents, and much attention has been devoted to the issue of the breadth versus the duration of patents – that is, whether long but narrow patents are preferable to short but broad ones. The seminal paper on the subject is that of Gilbert and Shapiro (1990), which models breadth as a price ceiling limiting the patent holder's ability to price at the full monopoly price. In the Gilbert–Shapiro setup, the conclusion is that optimal patents should be as long and narrow as possible. Subsequent authors have contested this conclusion – Gallini (1992), in particular, argues that the model of ‘breadth’ as a price ceiling does not reflect what ‘breadth’ is likely to mean in practice, and that a more reasonable model of breadth leads to

the opposite conclusion. Subsequent work by Gallini and Scotchmer (2001) concludes that optimal patent protection should probably be broad but short.

Recent research, for example by Hopenhayn and Mitchell (2001), has examined the possibility of providing not a single breadth-cum-duration for all patents, but rather allowing patent applicants to select from a menu of alternatives: some might choose broad but short, others narrow but long. They show that a properly calibrated system of this type can be superior to a one-size-fits-all system.

Many of these models recognize that innovators will earn something even without patent protection, and it is generally true in these models that if enough rents are earned without patents the optimal system is to have no patents at all. From the perspective of designing a system, this answer – no patents – is not interesting, and for understandable reasons this case tends to be underplayed. However, while how not to design a patent system may be less intellectually challenging than how to design one, the case in which adequate rents are earned without patents may well be empirically more relevant.

Optimal duration of intellectual property

Quite apart from the details of system design, the question arises of how much protection an IP system should provide. Why, for example, should monopoly be limited rather than unrestricted? Answering this question requires trading off the monopoly power created by intellectual property against the incentive to innovate. This issue was first studied in the context of copyrights by Ian and Waldman (1984). They examine a model in which ‘innovation’ has the dimension of higher product quality. However, they assume away the harmful effects of monopoly power by assuming that demand is completely elastic up to an upper bound. Not surprisingly, in this setting stronger intellectual property is unambiguously good.

Beginning with Liebowitz (1985), Stan Liebowitz has also extensively studied copyrights; focusing largely on a single creation, he argues that the indirect appropriability of competitive rents is generally an inadequate incentive to create.

More recently, Grossman and Lai (2004) and Boldrin and Levine (2005a) have examined a general equilibrium setting in which ideas of different quality are produced. Both papers show that optimal protection is generally limited rather than unlimited; and, because markets are growing over time, they consider the consequences of expanded markets for optimal protection. Boldrin and Levine show that optimal protection always declines when the market is large enough. They also give an elasticity condition under which protection should always decline with market size and, based on examination of existing data, conclude that this condition is likely to be satisfied in practice. This empirical analysis builds heavily on an empirical literature, stemming from Pakes (1986), that tries to estimate the distribution of patent values by examining such things as patent renewal rates.

Patent races

Some time ago, theoretical work focused on patent races, in which firms over-invest in R&D in an effort to obtain a valuable patent before a rival. Fudenberg et al. (1983) and Harris and Vickers (1985) are two of the earliest papers on these lines. Contrary to the traditional problem of too little innovation, this line of research suggests that the desire to acquire the monopoly power that patents confer may encourage too much expenditure in wasteful R&D. These models seem to have fallen out of favour in recent years,

perhaps because there is little empirical evidence that patent races are quantitatively important in determining the pace of actual technological innovation. The fact that legal battles over patents have become a persistent feature of contemporary business strategy may well restore currency to a modified version of these models.

Ideas as building blocks

One critical element of innovation – and artistic creation as well – is that new ideas are generally built upon existing ideas. That this is true for patentable innovations is fairly obvious (in the realm of copyright see Lessig, 2004, and Vaidhyanathan, 2003). Scotchmer (1991) points out that strong patent protection can have the dual effect of increasing the return to innovation and at the same time increasing the cost of acquiring the rights needed to innovate. This point is developed further in Boldrin and Levine, who show that under certain conditions a patent system may serve only to discourage innovation (2003), and that, when the innovator is better informed about the value of a new idea than the holders of rights to previous ideas, a patent system serves strictly to discourage innovation (2005b). Intuitively, all the additional profit from the new innovation is absorbed by the existing rights holders; if there are many of them, there is a public goods problem, with each ‘little monopolist’ setting a price that is too high because much of the cost of decreased likelihood of innovation is borne by the other ‘little monopolists’. This type of holdup problem is not dissimilar to the problem pointed out by Chari and Jones (2000) in the context of externalities more broadly; interestingly, in this case externalities are created by the existence of intellectual property, and would be altogether absent without it.

The practical impact of intellectual property in a setting where the use of existing ideas is important is well documented by Bessen and Hunt (2003), who examine the software industry in the United States during the era of personal computers; they find that intellectual property has been antithetical to innovation in this industry.

The role of transactions costs that arise when it is necessary to acquire many rights in order to innovate is underlined by David Friedman's (1994) striking hypothetical example of what would happen if every word in the English language was copyrighted, so that any writer had to pay for each use of every word.

Competitive innovation

Since there is a well-documented downside to intellectual property, it is important to understand how markets might function in its absence. Arnold Plant and George Stigler, among others, provide important examples of innovation and creation taking place without the benefit of monopoly. Plant (1934, p. 173) writes that, although in the 19th century English authors could not copyright their works in the United States,

> … American publishers found it profitable to make arrangements with English authors … English authors sometimes received more from the sale of their books by American publishers, where they had no copyright, than from their royalties in [England].

Similarly, Stigler (1956, p. 274) argues that monopoly is completely unnecessary to provide incentives

for innovation.

> There can be rewards – and great ones – to the successful competitive innovator. For example, the mail-order business … The innovators … were Aaron Montgomery Ward, who opened the first general merchandise establishment in 1872, and Richard Sears … Sears soon lifted his company to a dominant position by his magnificent merchandising talents, and he obtained a modest fortune, and his partner Rosenwald an immodest one. At no time were there any conventional monopolistic practices, and at all times there were rivals within the industry and other industries making near-perfect substitutes …

In more recent times, Liebowitz (1985), Boldrin and Levine (2003), Quah (2002), Legros (2005), and Hellwig and Irmen (2001) have all examined the competitive rents that accrue to innovators due to 'limited capacity' – the fact that in a competitive market the owners of a fixed factor (first copy of an idea) are the recipients of all downstream rents originating from it, and that an infinite number of copies cannot be made instantaneously. The conclusion is that innovation will take place even without intellectual property – as it often has in the past (see for example the cases mentioned by Moser, 2002). While some of this work shows that there may be too little innovation under competition due to the indivisible nature of the initial copy of ideas, it also suggests that the appropriate remedy is unlikely to be a government-granted monopoly.

In modern times, evidence that patents are unnecessary to provide the adequate incentive to innovate can be found in the widespread cross-licensing agreements found in chip manufacturing. The evidence is discussed by Shapiro (2001), who argues that the sharing of information between chip firms is much more important to them than any short-term advantage gained through a patent, and the primary function of patenting in this industry is to block entry by potential rivals.

First-mover advantage

Regardless of the presence of competitive rents, an innovator is likely to have a substantial advantage by being first to the market. This is largely what Plant and Stigler had in mind. The important impact of first-mover advantage in the market for new types of financial securities prior to the advent of patents in that industry has been ably documented by Tofuno (1989), and the theory explained carefully, together with further evidence, by Herrera and Schroth (2002; 2003).

Besides the temporary monopoly that results from being first, there are less obvious advantages. Hirshleifer (1971) first, and Anton and Yao (1994) subsequently, show how advance knowledge of an innovation can give an edge in asset markets. This can be illustrated through the example of the 'Segway' scooter – much publicized when it was introduced as a revolution in transportation. Suppose for the moment that these claims were true: how could the inventor have profited from this information without – as he did – surrounding himself with a thicket of patents? The Hirshleifer scheme would see him selling short automobile stocks, which would drop through the floor as soon as he announced his discovery. The Anton–Yao scheme would have the inventor sell the idea to, say, Ford in exchange for a share of the profits. Since he would share the profits, he would then have no incentive to try to sell the idea to other automobile companies, and so Ford would be happy to pay for the resulting monopoly. If it

simply took the idea without paying, Ford would lose the monopoly when the inventor told the other companies how to build Segways. Along similar lines, Baccara and Razin (2004) have developed a bargaining model for the case when the inventor must share the idea with others to implement it, and the latter can 'run away' after the idea is revealed. Even in the absence of any intellectual property, as the idea is revealed to more and more people and market power dilutes, the 'threat of competition' is enough to make the collaborator comply and to guarantee the innovator a substantial (larger than one-third) share of the surplus.

Rent seeking

The most significant downside of government grants of monopoly is the rent seeking they trigger. For example, although they have a generally favorable view of patents, historical research by Lamoreaux and Sokoloff (2001) shows that tightening of patent law resulted in a large upswing in innovation – presumably because it eliminated the nuisance of 'submarine' and other patents designed to appropriate value from the true innovators.

Outside the direct line of those whose existing way of business is threatened by innovation, enormous concern has been expressed at the consequences of rent seeking for the limitations it imposes on personal liberty and the threat it poses to economic progress. For example, the efforts of large media giants to 'protect' their 'intellectual property' through government-mandated hardware installed in computers poses a significant threat to innovation in the much larger IT industry.

Throughout history governments with little ability to monitor transactions and collect tax revenue have often fallen back on grants of monopoly to private individuals. Current patent and copyright systems seem to be remnants from this era, and many economists wonder if it is not time to replace them with more efficient modern systems of graduated incentives such as tax subsidies.

See Also

- patents

Bibliography

Aghion, P. and Howitt, P. 1992. A model of growth through creative destruction. *Econometrica* 60, 323–51.

Anton, J. and Yao, D. 1994. Expropriation and inventions: appropriable rents in the absence of property rights. *American Economic Review* 84, 190–209.

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity*, ed. R. Nelson. Princeton, NJ: Princeton University Press.

Baccara, M. and Razin, R. 2004. From thought to practice: appropriation and endogenous market

structure with imperfect intellectual property rights. Discussion Paper No. 4419. London: CEPR.

Bessen, J. and Hunt, R. 2003. An empirical look at software patents. Working paper No. 2003–17. Philadelphia: Federal Reserve Bank of Philadelphia.

Boldrin, M. and Levine, D. 2003. Perfectly competitive innovation. Mimeo. University of Minnesota and UCLA. Online. Available at http://www.econ.umn.edu/~mboldrin/Papers/pci39.pdf, accessed 14 July 2005.

Boldrin, M. and Levine, D. 2005a. IP and market size. Mimeo. University of Minnesota and UCLA. Online. Available at http://www.dklevine.com/papers/scale22.pdf, accessed 14 July 2005.

Boldrin, M. and Levine, D. 2005b. The economics of ideas and intellectual property. *Proceedings of the National Academy of Sciences* 102, 1252–6.

Chari, V. and Jones, L. 2000. A reconsideration of the problem of social cost: free riders and monopolists. *Economic Theory* 16, 1–22.

Friedman, D. 1994. Standards as intellectual property: an economic approach. *University of Dayton Law Review* 19, 1109–29.

Fudenberg, D., Gilbert, R., Stiglitz, J. and Tirole, J. 1983. Preemption, leapfrogging and competition in patent races. *European Economic Review* 22, 3–31.

Gallini, N. 1992. Patent policy and costly imitation. *RAND Journal of Economics* 23, 52–63.

Gallini, N. and Scotchmer, S. 2001. Intellectual property: when is it the best incentive system? Economics Working Papers E01–303. Berkeley: University of California.

Gilbert, R. and Shapiro, C. 1990. Optimal patent length and breadth. *RAND Journal of Economics* 21, 106–12.

Grossman, G. and Helpman, E. 1991. Quality ladders in the theory of growth. *Review of Economic Studies* 58, 43–61.

Grossman, G. and L.-C.Lai, E. 2004. International protection of intellectual property. *American Economic Review* 94, 1635–53.

Harris, C. and Vickers, J. 1985. Patent races and the persistence of monopoly. *Journal of Industrial Economics* 33, 461–81.

Hellwig, M. and Irmen, A. 2001. Endogenous technical change in a competitive economy. *Journal of Economic Theory* 101, 1–39.

Herrera, H. and Schroth, E. 2002. Profitable innovation without patent protection: the case of derivatives. Mimeo. Mexico, D.F.: Department of Economics, ITAM. Online. Available at http://ciep.itam.mx/~helios/deriv.pdf, accessed 14 July 2005.

Herrera, H. and Schroth, E. 2003. Developer's expertise and the dynamics of financial innovation: theory and evidence. Mimeo. Mexico, D.F.: Department of Economics, ITAM. Online. Available at http://ciep.itam.mx/~helios/express.pdf, accessed 14 July 2005.

Hirshleifer, J. 1971. The private and social value of information and the reward to inventive activity. *American Economic Review* 61, 561–74.

Hopenhayn, H. and Mitchell, M. 2001. Innovation variety and patent breadth. *RAND Journal of Economics* 32, 152–66.

Ian, E. and Waldman, M. 1984. The effects of increased copyright protection: an analytic approach. *Journal of Political Economy* 92, 236–46.

Lamoreaux, N. and Sokoloff, K. 2001. Market trade in patents and the rise of a class of specialized inventors in the nineteenth-century United States. *American Economic Review. Papers and Proceedings* 91, 39–44.

Legros, P. 2005. Art and internet: blessing the curse? Mimeo. Bruxelles: ECARES, Université Libre de Bruxelles.

Lessig, L. 2004. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York: Penguin Press.

Liebowitz, S.J. 1985. Copying and indirect appropriability: photocopying of journals. *Journal of Political Economy* 93, 945–57.

Moser, P. 2002. How do patent laws influence innovation? Evidence from nineteenth century world fairs. Mimeo. Sloan School of Management, MIT.

Pakes, A.S. 1986. Patents as options: some estimates of the value of holding European patent stocks. *Econometrica* 54, 755–84.

Plant, A. 1934. The economic aspect of copyright in books. *Economica* 1, 167–95.

Quah, D. 2002. 24/7 Competitive innovation. Mimeo. London School of Economics. Online. Available at http://econ.lse.ac.uk/staff/dquah/p/0204-247.pdf, accessed 4 July 2005.

Romer, P. 1986. Increasing returns and long run growth. *Journal of Political Economy* 94, 1002–3.

Romer, P. 1990. Endogenous Technological Change. *Journal of Political Economy* 98, S71–102.

Shapiro, C. 2001. Navigating the patent thicket: cross licenses, patent pools and standard setting. In *Innovation Policy and the Economy*, vol. 1, ed. A. Jaffe, J. Lerner and S. Stern. Cambridge, MA: MIT Press for the NBER.

Schumpeter, J. 1911. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press, 1934.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. New York: Harper and Brothers.

Scotchmer, S. 1991. Standing on the shoulders of giants: cumulative research and the patent law. *Journal of Economic Perspectives* 5, 29–41.

Stigler, G. 1956. Industrial organization and economic progress. In *The State of the Social Sciences*, ed. L. White. Chicago: University of Chicago Press.

Tofuno, P. 1989. First mover advantages in financial innovation. *Journal of Financial Economics* 3, 350–70.

Vaidhyanathan, S. 2003. *Copyrights and Copywrongs: The Rise of Intellectual Property and How It Threatens Creativity*. New York: New York University Press.

How to cite this article

Boldrin, Michele and David K. Levine. "intellectual property." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000246> doi:10.1057/9780230226203.0816

The New Palgrave Dictionary of Economics Online

interacting agents in finance

Cars Hommes

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Interacting agents in finance represent a behavioural, agent-based approach in which financial markets are viewed as complex adaptive systems consisting of many boundedly rational agents interacting through simple heterogeneous investment strategies, constantly adapting their behaviour in response to new information and strategy performance, and through social interactions. An interacting agent system acts as a noise filter, transforming and amplifying purely random news about economic fundamentals into an aggregate market outcome exhibiting important stylized facts such as unpredictable asset prices and returns, excess volatility, temporary bubbles and sudden crashes, large and persistent trading volume, clustered volatility and long memory.

Keywords

agent-based modelling; arbitrage; asset prices; asymmetric information; behavioural finance; bounded rationality; bubbles; clustered volatility; discrete choice; complexity; evolutionary selection mechanisms; excess volatility; expectations; finance; herding; cost of information; interacting agents in finance; long memory; multiple equilibria; noise traders; no-trade theorems; random choice; rational expectations; representative agent; social interaction; stylized facts

Article

Interacting agents in finance represent a new behavioural, agent-based approach in which financial markets are viewed as complex adaptive systems consisting of many boundedly rational, heterogeneous agents interacting through simple investment strategies, constantly learning from each other as new information becomes available and adapting their behaviour accordingly over time. Simple interactions at the individual, micro level cause sophisticated structure and emergent phenomena at the aggregate, macro level. Recent surveys of this approach are Hommes (2006) and LeBaron (2006).
The traditional approach in finance is based on a representative, rational agent who makes optimal

investment decisions and has rational expectations about future developments. Friedman (1953) made an early, strong argument in favour of rationality, arguing that ‘irrational’ agents would lose money whereas rational agents would earn higher profits. This is essentially an evolutionary argument saying that irrational agents will be driven out of the market by rational agents. In a perfectly rational world, information is transmitted instantaneously, asset prices reflect economic fundamentals and asset allocations are efficient. In the traditional view, agents interact only through the price system.

In contrast, Keynes earlier stressed that prices of speculative assets are not solely driven by market fundamentals, but that ‘market psychology’ also plays an important role. Another early critique on perfect rationality is due to Simon (1957), who emphasized that agents are limited in their computing abilities and face information gathering costs. Therefore individual behaviour is more accurately described by simple, suboptimal ‘rules of thumb’. Along similar lines, Tversky and Kahneman (1974) in psychology argued that individual decision behaviour under uncertainty can be better described by simple *heuristics* and *biases*. Since the 1990s the traditional view of financial markets has been challenged through developments in bounded rationality (for example, Sargent, 1993), behavioural finance (for example, Barberis and Thaler, 2003) and computational, agent-based modelling (for example, Tesfatsion and Judd, 2006).

Fundamentalists versus chartists

Most interacting agents models in finance include two important classes of investors: *fundamentalists* and *chartists*. Fundamentalists base their investment decisions upon market fundamentals, such as interest rates, growth of the economy, company's earnings, and so on. Fundamentalists expect the asset price to move towards its fundamental value and buy (sell) assets that are undervalued (overvalued). In contrast, chartists or technical analysts look for simple patterns, for example, trends in past prices, and base their investment decisions upon extrapolation of these patterns. For a long time, technical analysis has been viewed as ‘irrational’ and, according to the Friedman argument, chartists would be driven out of the market by rational investors. Frankel and Froot (1986) were among the first to emphasize the role of fundamentalists and chartists in real financial markets. Evidence from survey data on exchange rate expectations (for example, Frankel and Froot, 1987; Allen and Taylor, 1990) shows that at short time horizons (say, up to three months) financial forecasters tend to use *destabilizing*, trend-following forecasting rules, whereas at longer horizons (say 3–12 months or longer) they tend to use *stabilizing*, mean-reverting, fundamental forecasts. Frankel and Froot (1986) argue that the interaction of chartists and fundamentalists amplified the strong rise and subsequent fall of the dollar exchange rate in the mid-1980s.

Another simple interacting agent system with chartists and fundamentalists driven by *herding* behaviour is due to Kirman (1991; 1993). This model was motivated by the puzzling behaviour of ants observed by entomologists. A colony of ants facing two identical food sources distributes asymmetrically, say 80–20 per cent, over the two sources. Moreover, at some point in time the distribution suddenly reverses to 20–80 per cent. Kirman (1993) proposed a simple stochastic model explaining ants’ behaviour and applied it to a financial market setting (Kirman, 1991). Agents can choose between two investment strategies – a fundamentalist or a chartist strategy– to invest in a risky asset. Two agents meet at random and with

some interaction-conversion probability one agent will adopt the view of the other. There is also a small self-conversion probability that the agent will change her view no matter what the other agent believes. It turns out that, when the interaction-conversion probability is relatively high compared with the self-conversion probability, the distribution of agents is bimodal. The behaviour of the agents is very persistent and the market tends to be dominated by one group for a long time, but then the majority of agents suddenly switches to the other view, and so on.

But what about the Friedman argument? Will not ‘irrational’ technical trading rules be driven out of the market by rational investment strategies? DeLong et al. (1990) presented one of the first models showing that this need not be the case. Their model contains two types of traders, *noise traders*, with erroneous stochastic beliefs, and rational traders who are perfectly rational and take into account the presence of noise traders. Noise traders create extra risk and risk-averse rational traders are not willing to fully arbitrage away the mispricing. Noise traders bear more risk and can earn higher realized returns than rational traders, and therefore noise traders can survive in the long run. Lux (1995) presents a herding model with fundamentalists and chartists, whose behaviour is driven by imitation and past realized returns, leading to temporary bubbles and sudden crashes. Furthermore, Brock, Lakonishok and LeBaron (1992) showed empirically, using 90 years of daily Dow Jones index data, that technical trading rules can generate significant above-normal returns.

Markets as complex adaptive systems

Since the end of the 1980s, multidisciplinary research as done at the Santa Fe Institute (SFI) (for example, Anderson, Arrow and Pines, 1988) has stimulated a lot of work on interacting agents in economics and finance. Models of interacting particle systems in physics served as examples of how local interaction at the micro level may explain structure, for example a phase transition, at the macro level. This has motivated economists to study *the economy as an evolving complex system*.

Arthur et al. (1997) consider the so-called SFI artificial stock market consisting of an ocean of different types of agents choosing among many simple investment strategies. Agents’ investment decisions are affected by their expectations or beliefs about future asset prices. Beliefs affect realized prices, which in turn determine new beliefs, and so on. Prices and beliefs about prices thus *co-evolve* over time, and agents continuously adapt their behaviour as new observations become available, replacing less successful strategies by more successful ones. Are simple forecasting strategies irrational and will rational traders outperform technical traders in such an artificial market? In general, no. The reason is that a speculative asset market is an *expectations feedback system*. Imagine a situation where an asset price is overvalued and the majority of traders remains optimistic expecting the rising trend to continue. Aggregate demand will increase and as a result the asset price will rise even further. Optimistic expectations thus become *self-fulfilling* and chartists will earn higher realized returns than fundamental traders who sold or shortened the asset because they expected a decline in its price. As long as optimistic traders dominate the market and reinforce the price rise, fundamentalists will lose money. Even when the fundamentalists may be right in the long run, there are ‘limits to arbitrage’, for example due to short selling constraints, preventing them from holding their positions long enough against a prevailing optimistic view, as stressed by Shleifer and Vishny (1997).

Emergent phenomena and stylized facts

The interacting agents approach has been strongly motivated by a number of important stylized facts observed in many financial time series (for example, Brock, 1997): (*a*) unpredictable asset prices and returns; (*b*) large, persistent trading volume; (*c*) excess volatility and persistent deviations from fundamental value, and (*d*) clustered volatility and long memory. According to (*a*) asset prices are difficult to predict. New information is absorbed quickly in asset prices and there is ‘no easy free lunch’, that is, arbitrage opportunities are difficult to find and exploit. The traditional rational, representative agent framework can explain (*a*), but has difficulty in explaining the other stylized facts (*b*)–(*d*). In particular, in a world with only rational, risk-averse investors with asymmetric information there can be no trade, because no trader can benefit from superior information since other rational traders will anticipate that this agent must have superior information and therefore will not agree to trade (for example, Fudenberg and Tirole, 1991). These no-trade theorems are in sharp contrast to the huge daily trading volume observed in real financial markets, which suggests that there must be other types of heterogeneity such as *differences in opinion* about future movements. Stylized fact (*c*) means that fluctuations in asset prices are much larger than fluctuations in underlying market fundamentals. This point has been emphasized by, for example, by Shiller (1981). When markets are excessively volatile, prices can deviate from their fundamental values for a long time. Stylized fact (*d*) means that price fluctuations are characterized by irregular switching between quiet, low volatility phases, with small price fluctuations and turbulent phases of high volatility and large swings in asset prices. Interacting agent models have been able to explain these stylized facts simultaneously (for example, LeBaron, Arthur and Palmer, 1999; Lux and Marchesi, 1999).

Evolutionary selection of strategies

Blume (1993) and Brock (1993) present a general probabilistic framework for strategy selection motivated by results from interacting particle systems in physics (see also Föllmer, 1974). The probability of agents using strategy *h* changes over time according to a random utility fitness measure of the general form

$$U_{ht} = \pi_{ht} + S_{ht} + \varepsilon_{ht}. \tag{1}$$

Here π_{ht} represents *private utility*, for example given by (a weighted average of) realized profit, realized utility or forecasting performance. S_{ht} represents *social utility* measuring herding behaviour or social interactions (see Brock and Durlauf, 2001a; 2001b). For example, agents may behave as conformists, that is, they are more likely to follow strategies that are more popular among the population (global interaction) or among their neighbours (local interaction). Agents observe the performance of

each strategy with some idiosyncratic errors, represented by ε_{ht}.

A frequently used model for the probabilities or fractions of the different strategy types is the *discrete choice* or *multinomial logit model*

$$n_{ht} = e^{\beta U_{h,t-1}} / Z_{t-1}, \quad (2)$$

where $Z_{t-1} = \Sigma_j e^{\beta U_{j,t-1}}$ is a normalization factor so that the fractions add up to one. When the errors ε_{ht} in (1) are independently and identically distributed according to a double exponential distribution, the probability of choosing strategy *h* is exactly given by (2). The crucial feature of (2) is that, the higher the fitness of trading strategy *h*, the more agents will select strategy *h*, and therefore it is essentially an *evolutionary selection* mechanism. Agents are *boundedly rational* and tend to follow strategies that have performed well in the (recent) past. The parameter β is called the *intensity of choice* and is inversely related to the variance of the noise ε_{ht}. It measures how sensitive agents are to selecting the optimal strategy. The extreme case $\beta = 0$ corresponds to noise with infinite variance, so that differences in fitness cannot be observed and all fractions will be equal to 1/*H*, where *H* is the number of strategies. The other extreme $\beta = +\infty$ corresponds to the case without noise, so that the deterministic part of the fitness is observed perfectly, and in each period *all* agents choose the optimal forecast. An increase in the intensity of choice β represents an increase in the degree of rationality concerning strategy selection. Brock and Hommes (1997; 1998) propose a simple, analytically tractable heterogeneous agent model to show how non-rational strategies can survive evolutionary selection. Brock and Hommes (1997) consider a market with an *endogenous evolutionary selection* of expectations rules described by the multi-nomial logit model (2), with fitness given by past realized profits. Agents choose between a set of different forecasting rules and tend to switch to forecasting strategies that have performed well in the recent past. When agents face information gathering costs, because sophisticated rational strategies are more costly to obtain, simple rule of thumb strategies can survive in this market. In Brock and Hommes (1998) this evolutionary selection of strategies is applied to a standard asset pricing model similar to but much simpler than the SFI artificial stock market. Agents choose between fundamentalists' and chartists' investment strategies. When the sensitivity to differences in past performance of the strategies is high (that is, the parameter β is high), evolutionary selection of strategies destabilizes the system and leads to complicated, possibly chaotic asset price fluctuations around the benchmark rational expectations fundamental price. The fluctuations are characterized by an irregular switching between a quiet phase with asset prices close to the fundamentals and a more turbulent phase with asset prices following (temporary) trends or bubbles. In contrast with Friedman's argument, chartists can survive in this evolutionary competition and may on average earn (short-run) profits equal to or even higher than (short-run) profits of fundamentalists.

A common finding in these models is that more rationality, that is, a larger intensity of choice, leads to *instability*. The intuition is that random choice leads to stability, because agents will be evenly

distributed over the strategy space without systematic biases. In contrast, correlated choice may cause instability when, for example, many traders switch to a profitable trend-following strategy. Another common finding is that, when the social interaction effect is strong, multiple equilibria exist and it depends sensitively on the initial state to which of the many equilibria the market system will settle down (for example, Brock and Durlauf, 2001a; 2001b).

Summary and future perspectives

Although the approach in finance is relatively new, interacting agent models have been able to explain important stylized facts simultaneously. An interacting agents system acts as a noise filter, transforming and amplifying purely random news about economic fundamentals into an aggregate market outcome exhibiting excess volatility, temporary bubbles and sudden crashes, large and persistent trading volume, clustered volatility and long memory. It should be emphasized that at the aggregate level these asset price fluctuations are highly irregular and unpredictable, there exists no easy free lunch, and arbitrage will be very difficult and risky in such a market.
Much more theoretical work is needed in this area, for example, to find the ‘simplest tractable model’ explaining all important stylized facts. Speculative bubbles have been observed in laboratory experiments of Smith, Suchanek and Williams (1988) and more recently in Hommes et al. (2005), showing that coordination on trend-following rules can destabilize a laboratory experimental asset market. Another important topic for future research is estimation of interacting agent models on financial data. Boswijk, Hommes and Manzan (2007) is one of the first attempts to estimation of an evolutionary model with fundamentalists versus trend-following chartists using yearly S&P 500 data, suggesting that trend-following behaviour amplified the strong rise in stock prices at the end of the 1990s. More laboratory experiments and estimation of interacting agents models are needed to test the robustness and empirical relevance of the interacting agents approach.

See Also

- economy as a complex system
- ergodicity and nonergodicity in economics
- finance
- finance (new developments)
- mathematics of networks
- network formation
- rationality, bounded
- residential real estate and finance
- social interactions (empirics)
- social interactions (theory)

Bibliography

Allen, H. and Taylor, M. 1990. Charts, noise and fundamentals in the London foreign exchange market. *Economic Journal* 100, 49–59.

Anderson, P., Arrow, K. and Pines, D., eds. 1988. *The Economy as an Evolving Complex System* Reading, MA: Addison-Wesley.

Arthur, W., Holland, J., LeBaron, B., Palmer, R. and Taylor, P. 1997. Asset pricing under endogenous expectations in an artificial stock market. In *The Economy as an Evolving Complex System II*, ed. W. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison-Wesley.

Barberis, N. and Thaler, R. 2003. A survey of behavioral finance. In *Handbook of the Economics of Finance*, ed. G. Constantinidis, M. Harris and R. Stulz. Amsterdam: North-Holland.

Blume, L. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5, 387–424.

Boswijk, H., Hommes, C. and Manzan, S. 2007. Behavioral heterogeneity in stock prices. *Journal of Economic Dynamics and Control* (forthcoming).

Brock, W. 1993. Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance. *Estudios Económicos* 8, 3–55.

Brock, W. 1997. Asset price behavior in complex environments. In *The Economy as an Evolving Complex System II*, ed. W. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison-Wesley.

Brock, W. and Durlauf, S. 2001a. Discrete choice with social interactions. *Review of Economic Studies* 68, 235–60.

Brock, W. and Durlauf, S. 2001b. Interactions-based models. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Brock, W. and Hommes, C. 1997. A rational route to randomness. *Econometrica* 65, 1059–95.

Brock, W. and Hommes, C. 1998. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control* 22, 1235–74.

Brock, W., Lakonishok, J. and LeBaron, B. 1992. Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance* 47, 1731–64.

DeLong, J., Shleifer, A., Summers, L. and Waldmann, R. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98, 703–38.

Föllmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1, 51–62.

Frankel, J. and Froot, K. 1986. Understanding the US dollar in the Eighties: the expectations of chartists and fundamentalists, *Economic Record*, special issue, 24–38.

Frankel, J. and Froot, K. 1987. Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review* 77, 133–53.

Friedman, M. 1953. The case of flexible exchange rates. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Fudenberg, D. and Tirole, J. 1991. *Game Theory*. Cambridge. MA: MIT Press.

Hommes, C. 2006. Heterogeneous agent models in economics and finance. In *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*, ed. L. Tesfatsion and K. Judd. Amsterdam: North-Holland. pp. 1109–86.

Hommes, C., Sonnemans, J., Tuinstra, J. and van de Velden, H. 2005. Coordination of expectations in asset pricing experiments. *Review of Financial Studies* 18, 955–80.

Kirman, A. 1991. Epidemics of opinion and speculative bubbles in financial markets. In *Money and Financial Markets*, ed. M. Taylor. London: Macmillan.

Kirman, A. 1993. Ants, rationality and recruitment. *Quarterly Journal of Economics* 108, 137–56.

LeBaron, B. 2006. Agent-based computational finance. In *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*, ed. L. Tesfatsion and K. Judd. Amsterdam: North-Holland. pp. 1187–1233.

LeBaron, B., Arthur, W. and Palmer, R. 1999. Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control* 23, 1487–516.

Lux, T. 1995. Herd behavior, bubbles and crashes. *Economic Journal* 105, 881–96.

Lux, T. and Marchesi, M. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397, 498–500.

Sargent, T. 1993. *Bounded Rationality in Macroeconomics*. Oxford: Clarendon Press.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends?

American Economic Review 71, 421–36.

Shleifer, A. and Vishny, R. 1997. The limits of arbitrage. *Journal of Finance* 52, 35–55.

Simon, H. 1957. *Models of Man*. New York: Wiley.

Smith, V., Suchanek, G. and Williams, A. 1988. Bubbles, crashes and endogenous expectations in experimental spot asset markets. *Econometrica* 56, 1119–51.

Tesfatsion, L. and Judd, K., eds. 2006. *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*. Amsterdam: North-Holland.

Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–31.

How to cite this article

Hommes, Cars. "interacting agents in finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000249> doi:10.1057/9780230226203.0818

The New Palgrave Dictionary of Economics Online

interdependent preferences

Peter C. Fishburn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Interdependent preferences arise in economic theory in the study of both individual decisions and group decisions. We imagine that a decision is required among alternatives in a set X and that the decision will depend on preferences between the elements in X. If the preferences represent different points of view about the relative desirability of the alternatives, of if they are based on multiple criteria that impinge on the decision, then we encounter the possibility of interdependent preferences.

There are two predominant approaches to interdependent preferences, the synthetic and the analytic. The synthetic approach begins with a set of preference relations on X and attempts to aggregate them into a holistic representative preference relation on X. This is done in social choice theory, where each original relation refers to the preferences of an individual in a social group. The aggregate relation is then referred to as a social preference relation. The synthetic approach also appears in studies of individual preferences, as when an individual rank-orders the alternatives for each of a number of criteria and then seeks a holistic ranking that combines the criteria rankings in a reasonable way.

In contrast, the analytic approach begins with a holistic preference relation on X and seeks to analyse its internal structure. This may involve a decomposition into components of preference, or it may concern trade-offs between factors that describe interactive contributions to overall preferences.

The synthetic approach often considers a list $(\square_1, \square_2, \ldots, \square_n)$ of preference relations on X, where $x \bullet_i y$ could mean that person i prefers x to y, or that an individual prefers x to y on the basis of criterion i. The problem may then be to specify a holistic relation $\square = f(\square_1, \square_2, \ldots, \square_n)$ for each possible n-tuple of individual relations.

The analytic approach often begins with X as a subset of the product $X_1 \times X_2 \times \cdots \times X_n$ of n other sets. It considers a holistic *is preferred to* relation $\bullet$ on X and asks how $\bullet$ depends on the X_i considered separately or in combination. Under suitably strong *independence* assumptions it may be possible to *define* $\bullet_i$ for each i in a natural way from $\bullet$ on X, and perhaps to establish a functional dependence of $\bullet$ on the $\bullet_i$. However, interdependencies among the factors will often preclude such a simple resolution.

Historical remarks

During the rise of marginal utility analysis in the latter part of the 19th century (Stigler, 1950), the utility of each commodity bundle in a set $X=X_1\times X_2\times\cdots\times X_n$ was thought of as an intuitively measurable quantity. Founders such as Jevons, Menger and Walras regarded x as preferred to y precisely when $u(x)$, the utility of x, is greater than $u(y)$. Their analytic approach ignored interdependencies since they used the independent additive utility form $u(x) = u_1(x_1) + \cdots + u_n(x_n)$.

Later writers such as Edgeworth, Fisher, Pareto and Slutsky discarded the additive decomposition for the general interdependent form $u(x_1, x_2,\ldots, x_n)$. Their ordinalist view of utilities as a mere reflection of a preference ordering remains dominant, and they considered interactive effects among goods, such as complementarities and substitutabilities. A fine example of interdependent analysis appears in Fisher (1892).

Fisher was also one of the first people to mention explicitly the interpersonal effect on individual utility (Stigler, 1950, p. 324). This occurs when one's utility and consequent demand depend on other people's consumption and could generally be expressed by $u_i(x_1,\ldots,\bullet x_n)$ as consumer i's utility when x_j denotes the commodity bundle of consumer j. Pigou (1903) considered the interpersonal effect in modest detail, and Duesenberry (1949) explored it in greater depth, but it has never been a prominent concern in economic theory.

Early examples of the synthetic approach in social choice theory come from Borda and Condorcet in the late 1700s. They asked: Given a list of voter preference rankings on a set X of $m\bullet\geq\bullet 3$ nominees, what is the best way of selecting a winner? Borda's answer was to assign $m, m-1,\ldots,1$ points to each first, second,…, last place nominee in the rankings and to elect the nominee with the largest point total. Condorcet advocated the election of a nominee who is preferred by a simple majority of voters to each other nominee in pairwise comparisons. Black (1958) contains an excellent review of their work and the proposals of later writers. The debate over good election methods continues today (Brams and Fishburn, 1983).

The turning point for social choice theory was Arrow's (1951) discovery that a few appealing conditions for aggregating individual preference orders on three or more candidates into social preference orders were jointly incompatible. The avalanche of research set off by Arrow's discovery is represented in part by Sen (1970, 1977), Fishburn (1973), Pattanaik (1971), and Kelly (1978).

In the area of risky decision theory, we envision a risky alternative as a probability distribution x on potential outcomes in a set C and observe that such decisions involve multiple factors since they entail both chances and outcomes. Bernoulli (1738) argued that a reasonable person will choose a risky alternative from a set X of distributions that maximizes his expected utility $\sum x(c)u(c)$. He proposed that u be assessed without reference to chance since he held an intuitive measurability view of utility. Consequently, his approach is wholly synthetic.

Little changed in the foundations of risky decisions during the next two centuries. Then, in a complete turnabout, von Neumann and Morgenstern (1944) introduced the analytic approach by beginning with a preference relation • on X. Axioms for • on X were shown to imply the existence of a real valued function u on C such that, for all x and y in X, $x \bullet y$ precisely when x has greater expected utility than y, and u is to be assessed on the basis of comparisons between distributions. With a few exceptions, most notably Allais (1953), subsequent research has adopted the von Neumann-Morgenstern approach.

In the rest of this essay we comment further on multiattribute preferences under ‘certainty’, interdependent preferences in risky decisions, and social choice theory.

Multiattribute preferences

We assume throughout this section that • is a strict preference relation on $X=X_1\times X_2\times\cdots\times X_n$. A given X_i could represent amounts of commodity *i*, consumption bundles available to person *i*, levels of income and/or consumption in period *i*, or values that elements in *X* might have for criterion *i*. Also let *u* on *X* and u_i denote real valued functions.

A non-empty proper subset *N* of $\{1, 2, \ldots, n\}$ is defined to be •-*independent* if, for all x_N and y_N in the product of the X_i over *N* and for all $z_{(N)}$ in the product of the X_i over *i not* in *N*,

$$(xN, z(N)) \square (yN, z(N)) \square (xN, w(N)) \square (yN, w(N)).$$

Most research for • on *X* involves •-independence for some *N*, but this need not exclude elementary notions of preference interdependencies. Two models that presume all *N* to be •-independent are the additive model (see Krantz, Luce, Suppes and Tversky, 1971)

$$x \square y \Leftrightarrow u_1(x_1) + \cdots + u_n(x_n) > u_1(y_1) + \cdots + u_n(y_n),$$

and the lexicographic model (Fishburn, 1974a) that places a value hierarchy on the factors.

Relationships between factors in the additive model and the more general model $x \square y \Leftrightarrow u(x) > u(y)$ with *u* continuous, are often characterized by indifference maps or iso-utility contours. Interdependence arises in the lexicographic model from the fact that a small change in one factor overwhelms all changes in factors that are lower in the hierarchy.

Situations in which only some of the *N* are •-independent are reviewed by Keeney and Raiffa (1976, ch. 3) and Krantz, Luce, Suppes and Tversky (1971, ch. 7). Among other things, these models allow complete reversals in preferences over one factor at different fixed levels of the other factors. This, of course, is a very strong form of interdependence under which all *N* may fail to be •-independent.

Other general models for interdependent preferences are discussed by Fishburn (1972) for finite sets, and by Dyer and Sarin (1979) when *u* is viewed in the intuitive measurability way.

Models that explicitly incorporate the interpersonal effect in economic analysis have been investigated by Pollak (1976) and Wind (1976), among others. Pollak explores the influence of several versions of interdependence among individuals on short-run and long-run consumption within a group. Using models of demand that are locally linear in others’ past consumption, he concludes that the distribution of income need not be a determinant of long-run per capita consumption patterns. Wind's work is

representative of empirical approaches to the influence of others on an individual's choice behaviour.

Risky decisions

Interdependent preferences in risky decisions fall into two categories. The first concerns special forms for $u(c) = u(c_1, c_2, \ldots, c_n)$ in the context of von Neumann-Morgenstern expected utility theory when the outcome set C is a subset of a product set $C_1 \times C_2 \times \cdots \times C_n$. The second focuses on changes in the basic model that occur when the independence axiom that gives rise to the expected utility form $\sum x(c)u(c)$ is relaxed or dropped.

Decompositions of $u(c_1, c_2, \ldots, c_n)$ in the expected utility model have been axiomatized by various people. Reviews and extensions of much of this work appear in Keeney and Raiffa (1976) and Farquhar (1978). The simplest independent decompositions are the additive form and a multiplicative form. The first of these requires x and y to be indifferent whenever the marginal distributions of x and y on X_i are the same for every i. The multiplicative form arises when, for each non-empty proper subset N of $\{1, \ldots, n\}$, the preference order over marginal distributions on the product of the C_i for i in N, conditioned on fixed values of the other factors, does not depend on those fixed values.

An example of a more involved interdependent decomposition is the two-factor model (Fishburn and Farquhar, 1982) $u(c_1, c_2) = f_1(c_1)g_1(c_2) + \cdots + f_m(c_1)g_m(c_2) + h(c_1),$ which clearly allows a variety of interactive effects.

In the basic formulation for expected utility, assume that X is closed under convex combinations $\lambda x + (1-\lambda)y$ with $0<\lambda<1$ and x and y in X. The *independence axiom* for expected utility asserts that, for all x, y and z in X and all $0<\lambda<1$.

$$x \square y \Rightarrow \lambda x + (1-\lambda)z \square \lambda y + (1-\lambda)z.$$

Systematic violations of this axiom uncovered in experiments by Allais (1953), Kahneman and Tversky (1979) and MacCrimmon and Larsson (1979) among others, have led to new theories of risky decisions (Kahneman and Tversky, 1979; Machina, 1982; Chew, 1983; Fishburn, 1982) that do not assume independence. Machina (1982) proposes a model that approximates expected utility locally but not globally. Fishburn (1982) weakens the usual transitivity and independence assumptions to obtain a non-separable model $x \square y \Leftrightarrow \varphi(x, y) > 0$ that allows preference cycles.

Related interdependent generalizations of Savage's subjective expected utility model for decisions under uncertainty are developed by Loomes and Sugden (1982) and Schmeidler (1984).

Social choice

Many problems in social choice theory are related to Condorcet's phenomenon of cyclical majorities. This phenomenon occurs when voters have transitive preferences yet every nominee is defeated by

another nominee under simple majority comparisons. The simplest example has three nominees and three voters with $x \succ_1 y \succ_1 z$, $z \succ_2 y$ and $y \succ_3 z \succ_3 x$; *x* beats *y*, *y* beats *z*, and *z* beats *x*. Borda's point-summation procedure can fail to satisfy Condorcet's majority-choice principle, and it is notoriously sensitive to strategic voting. Moreover, all summation procedures based on decreasing weights for positions in voters' rankings are sensitive to nominees who have absolutely no chance of winning, but whose presence can affect the outcome.

Various problems and paradoxes for multicandidate elections that arise from combinatorial aspects of synthetic methods are discussed by Fishburn (1974b), Niemi and Riker (1976), Saari (1982) and Fishburn and Brams (1983). Analyses of strategic voting, which suggest that no sensible election method is immune from manipulation by falsification of preferences, are reviewed in Kelly (1978) and Pattanaik (1978).

Arrow's (1951) theorem offers a striking generalization of Condorcet's cyclical majorities phenomenon. Suppose *X* contains three or more nominees, each of *n* voters can have any preference ranking on *X*, and an aggregate ranking $\succ = f(\succ_1, \succ_2, \ldots, \succ_n)$ is desired for each list $(\succ_1, \succ_2, \ldots, \succ_n)$ of individual rankings. The question addressed by Arrow is whether there is any way of doing this that satisfies the following three conditions for all *x* and *y* in *X*:

1. (1) Pareto optimality: if $x \succ_i y$ for all *i*, then $x \succ y$;
2. (2) Binary independence: the aggregate preference between *x* and *y* depends solely on the voters' preferences between *x* and *y*;
3. (3) Non-dictatorship: there is no *i* such that $x \succ y$ whenever $x \succ_i y$. Arrow's theorem says that it is impossible to satisfy all three conditions.

Several dozen related impossibility theorems have subsequently been developed by others. Many of these are noted in Kelly (1978) and Pattanaik (1978). As well as multi-profile theorems, like Arrow's, that use different lists of preference rankings to demonstrate impossibility, there are single-profile theorems (Roberts, 1980) that use only one list with sufficient variety in the rankings to establish impossibility.

Impossibility theorems, voting paradoxes, and results on strategic manipulation highlight the difficulty of designing good election procedures. Recent research to alleviate such problems (Dasgupta, Hammond and Maskin, 1979; Laffont and Moulin, 1982) focuses on the design of preference-revelation mechanisms (generalized ballots) and aggregation procedures that encourage people to vote in such a way that the outcome will agree with some theoretically best decision based on the true but unknown preferences of the voters. Other work, such as that on approval voting (Brams and Fishburn, 1983), continues to search for simple synthetic methods that minimize the problems that beset these methods.

See Also

- Arrow's theorem
- externalities

Bibliography

Allais, M. 1953. Le comportement de 1'homme rationnel devant de risque; critique des postulats et axiomes de 1’école Américaine. *Econometrica* 21, 503–46.

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley. 2nd edn, 1963.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientarium Imperialis Petropolitanae* 5, 175–92. Trans. by L. Sommer as ‘Exposition of a new theory on the measurement of risk’, *Econometrica* 22 (1954), 23–36.

Black, D. 1958. *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.

Brams, S.J. and Fishburn, P.C. 1983. *Approval Voting*. Boston: Birkhäuser.

Chew, S.H. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* 51, 1065–92.

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules: some general results on incentive compatibility. *Review of Economic Studies* 46, 185–216.

Duesenberry, J.S. 1949. *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, Mass.: Harvard University Press.

Dyer, J.S. and Sarin, R.K. 1979. Measurable multiattribute value functions. *Operations Research* 27, 810–22.

Farquhar, P.H. 1978. Interdependent criteria in utility analysis. In *Multiple Criteria Problem Solving*, ed. S. Zionts, Berlin: Springer-Verlag, 131–80.

Fishburn, P.C. 1972. Interdependent preferences on finite sets. *Journal of Mathematical Psychology* 9, 225–36.

Fishburn, P.C. 1973. *The Theory of Social Choice*. Princeton: Princeton University Press.

Fishburn, P.C. 1974a. Lexicographic orders, utilities, and decision rules: a survey. *Management Science* 20, 1442–71.

Fishburn, P.C. 1974b. Paradoxes of voting. *American Political Science Review* 68, 537–46.

Fishburn, P.C. 1982. Nontransitive measurable utility. *Journal of Mathematical Psychology* 26, 31–67.

Fishburn, P.C. and Brams, S.J. 1983. Paradoxes of preferential voting. *Mathematics Magazine* 56, 207–14.

Fishburn, P.C. and Farquhar, P.H. 1982. Finite-degree utility independence. *Mathematics of Operations Research* 7, 348–53.

Fisher, I. 1892. Mathematical investigations in the theory of values and prices. *Transactions of the Connecticut Academy of Arts and Sciences* 9, 1–124. Reprinted, New York: Augustus M. Kelley, 1965.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Keeney, R.L. and Raiffa, H. 1976. *Decisions with Multiple Objectives: preferences and value tradeoffs*. New York: Wiley.

Kelly, J.S. 1978. *Arrow Impossibility Theorems*. New York: Academic Press.

Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A. 1971. *Foundations of Measurement*. Volume I: *Additive and Polynomial Representations*. New York: Academic Press.

Laffont, J.-J. and Moulin, H., eds. 1982. Special issue on implementation. *Journal of Mathematical Economics* 10(1).

Loomes, G. and Sugden, R. 1982. Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal* 92, 805–24.

MacCrimmon, K.R. and Larsson, S. 1979. Utility theory: axioms versus ‘paradoxes’. In *Expected Utility Hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen. Dordrecht: Reidel, 333–409.

Machina, M.J. 1982. ‘Expected utility’ analysis without the independence axiom. *Econometrica* 50, 277–323.

Niemi, R.G. and Riker, W.H. 1976. The choice of voting systems. *Scientific American* 234, 21–7.

Pattanaik, P.K. 1971. *Voting and Collective Choice*. Cambridge: Cambridge University Press.

Pattanaik, P.K. 1978. *Strategy and Group Choice*. Amsterdam: North-Holland.

Pigou, A.C. 1903. Some remarks on utility. *Economic Journal* 13, 58–68.

Pollak, R.A. 1976. Interdependent preferences. *American Economic Review* 66, 309–20.

Roberts, K.W.S. 1980. Social choice theory: the single-profile and multi-profile approaches. *Review of Economic Studies* 47, 441–50.

Saari, D.G. 1982. Inconsistencies of weighted summation voting systems. *Mathematics of Operations Research* 7, 479–90.

Schmeidler, D. 1984. Subjective probability and expected utility without additivity. Preprint #84, Institute for Mathematics and its Application, University of Minnesota.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Sen, A.K. 1977. Social choice theory: a re-examination. *Econometrica* 45, 53–89.

Stigler, G.J. 1950. The development of utility theory, Part I and II. *Journal of Political Economy* 58, 307–27, 373–96.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press. 2nd edn, 1947; 3rd edn, 1953.

Wind, Y. 1976. Preference of relevant others and individual choice models. *Journal of Consumer Research* 3, 50–7.

How to cite this article

Fishburn, Peter C. "interdependent preferences." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000142> doi:10.1057/9780230226203.0819

The New Palgrave Dictionary of Economics Online

intergenerational income mobility

Gary Solon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

One important aspect of income inequality is the extent to which position in the income distribution is passed from parents to children. Theoretical models suggest that both intergenerational persistence and equilibrium income inequality increase with the responsiveness of earnings to human capital investment and with the heritability of income-generating traits, and decrease with the progressivity of public investment in children's human capital. A rapidly growing empirical literature is documenting the extent of intergenerational income mobility in many countries and is beginning to explore why intergenerational transmission is as high (and low) as it is.

Keywords

assortative mating; Becker, G.; Cobb–Douglas function; human capital investment; income mobility; inequality of income; intergenerational income mobility; longitudinal surveys; panel surveys

Article

‘Intergenerational income mobility’ refers to the degree to which position in the income distribution persists or changes from one generation to the next.
For example, a society in which individuals’ adult income is altogether independent of their parents’ income is a highly mobile society. A society in which one's percentile in the income distribution is always identical to one's parents’ percentile is completely immobile. Neither extreme is ideal, and neither corresponds to the intergenerational patterns typically observed in actual societies. Which intergenerational association between the extremes is desirable depends on the processes generating it. In any case, reasonable and well-informed observers are likely to disagree about the optimal level of intergenerational mobility because of differences in their values, such as feelings about equity–efficiency trade-offs. See Jencks and Tach (2005) for a discussion of some of the normative issues.
Because understanding intergenerational mobility is important for understanding the nature of income

inequality, intergenerational mobility has received a great deal of attention from both theoretical and empirical researchers. Since the late 1980s, empirical research describing the extent of intergenerational mobility has made considerable progress. Much work remains to improve our understanding of the causal processes underlying the observed extent of intergenerational mobility.

Theory

The classic theoretical analysis by Becker and Tomes (1979) encompasses a multitude of reasons why relative income status is correlated across generations. In a recent variant of that analysis, Solon (2004) adopts the functional form assumptions consistent with the log-linear intergenerational mobility regression commonly estimated by empirical researchers. In this model, an individual parent divides her income between her own consumption and investment in an individual child's human capital so as to maximize a Cobb–Douglas utility function in which the two goods are the parent's consumption and the child's adult income. The mapping from the parent's investment in her child's human capital to the child's subsequent income as an adult operates through two functions.

First, a semi-logarithmic human capital production function relates the child's level of human capital to the logarithm of the sum of the parent's investment and public investment (for example, publicly supported education and health care for children) plus a variable representing the human capital endowments children receive regardless of the investment choices of their families and the government. These more mechanically determined endowments to children are, in Becker and Tomes's (1979, p. 1158) words, ‘determined by the reputation and “connections” of their families, the contribution to the ability, race, and other characteristics of children from the genetic constitutions of their families, and the learning, skills, goals, and other “family commodities” acquired through belonging to a particular family culture’. The transmission of these endowments is assumed to follow a first-order autoregressive process across generations. Thus, intergenerational transmission occurs both because higher-income parents have greater wherewithal to invest in the human capital of their children and because of the genetic and cultural heritability of human capital. Second, the mapping to the child's income is completed by a semi-logarithmic earnings function that relates the child's log earnings to her level of human capital.

This simple model leaves out some important aspects of intergenerational transmission. For example, it assumes that the parent cannot borrow against the child's prospective earnings and does not bequeath financial assets to the child. See Becker and Tomes (1986) and Mulligan (1997) for analyses that relax this assumption. Also, the model's single-parent/single-child structure ignores the role of assortative mating, which is discussed by Lam and Schoeni (1994) and Chadwick and Solon (2002). Nevertheless, the model is rich enough to illustrate some key aspects of the intergenerational transmission process. These are embodied in the following result concerning the intergenerational income elasticity β, which is the coefficient in the regression of the child's log income on the parent's log income:

$$\beta \cong \frac{(1-\gamma)\theta+\lambda}{1+(1-\gamma)\theta\lambda} \tag{1}$$

where θ is the elasticity of earnings with respect to human capital investment, λ is the autoregressive parameter representing the genetic and cultural heritability of income-generating traits, and γ is an index of the progressivity of public investment in children's human capital.
This result implies that the intergenerational income elasticity increases with the responsiveness of earnings to human capital investment and with the heritability of income-generating traits, and decreases with the progressivity of public investment in children's human capital. Cross-country differences in intergenerational mobility could arise from differences in any of these factors. So could changes over time in a particular country's intergenerational mobility. Finally, it can be shown that the same factors that increase the intergenerational income elasticity also increase the equilibrium level of cross-sectional income inequality within a generation. Thus, we should not be surprised if societies with particularly high income inequality also exhibit high intergenerational persistence of income status.

Empirical evidence

If lifetime income data were available for both the parents' and children's generations in a nationally representative sample, estimation of the intergenerational income elasticity β could be performed simply by applying least squares to the regression of the children's log lifetime income on the parents' log lifetime income. In most countries, however, the ideal data are not available. As of the 1980s, data constraints had forced most of the then-small empirical literature to rely on short-term income measures, such as annual income in only one year, for peculiarly homogeneous samples. As summarized in Becker and Tomes (1986, p. S25), the resulting estimates suggested that 'a 10% increase in father's earnings (or income) raises son's earnings by less than 2%'. As discussed in detail in Solon (1989; 1992), however, these estimates were biased substantially downward. The 'right-side' measurement error from using short-term parental income measures to proxy for parents' lifetime income can serve as a good classroom example for the econometrics textbook analysis of the attenuation bias resulting from 'noisy' measurement of an explanatory variable. And when the estimates were based on relatively homogeneous parent samples, this bias was aggravated by the diminished 'signal variance' in the explanatory variable. By the 1990s, empirical researchers in the United States had the benefit of better data. By that time, two longitudinal surveys initiated in the late 1960s, the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS) of labour market experience, had generated new data with an intergenerational span. Because these surveys used national probability samples, they were less subject to the problems from homogeneous samples. And because the longitudinal surveys repeatedly collected income information at each re-interview, they enabled exploration of the impact of using longer-term income measures. Many of the new studies, surveyed in Solon (1999), treated the errors-in-variables issue by averaging the parental income measure over several years. A typical finding was that, in a regression of son's log earnings on a multi-year measure of father's log earnings, the estimated slope coefficient was about 0.4 – that is, double the 0.2 value that previously had been described as an upper bound for the intergenerational elasticity. A few studies treated the errors-in-variables problem by performing instrumental variables estimation with parental characteristics like education or occupation used to instrument for measured parental income. That approach usually produced somewhat higher intergenerational elasticity estimates, but, as explained in Solon (1992), the consistency of such

instrumental variables estimation depends on the ‘excludability’ of the instruments from the model for children's income.

Even the new estimates based on multi-year parental income data probably were too low. As emphasized in Solon (1992) and Mazumder (2005), averaging parental income over several years reduces but does not eliminate attenuation bias. Non-random attrition from the longitudinal surveys probably generated a weaker version of the sample homogeneity that had plagued earlier data-sets. And, as discussed in Reville (1995) and Haider and Solon (2006), many of the newer estimates have been biased by ‘left-side’ measurement error. At the time researchers began to use intergenerational data from the PSID and NLS, the offspring were only about 30 years old or even younger. For workers in their twenties, the log of current income as a proxy for log lifetime income is subject to ‘mean-reverting’ measurement error, instead of the classical measurement error typically analysed in econometrics textbooks. The mean reversion occurs because the workers who eventually will have high lifetime earnings typically experience steeper earnings growth. As a result, the early career gap in current earnings between workers with high and low lifetime earnings tends to understate their lifetime gap. This sort of mean-reverting measurement error in a dependent variable is still another source of attenuation bias. Once all these downward biases in the estimation of the intergenerational elasticity are considered, it becomes plausible that the intergenerational elasticity in the United States may well be as large as 0.5 or 0.6.

In recent years, researchers have estimated intergenerational elasticities for many other countries, sometimes with much larger samples than are available from the US surveys. As summarized in Solon (2002), the elasticity estimates for the United States and United Kingdom are towards the high end among developed countries, with considerably smaller estimates appearing for Canada, Sweden, Finland and Norway. Some new estimates for developing countries in Latin America (Dunn, 2004; Ferreira and Veloso, 2004; Grawe, 2004) are even higher than the US and UK estimates. By and large, these cross-country comparisons accord with the theoretical prediction of greater intergenerational income persistence in countries with greater income inequality, higher returns to human capital, and less progressive public investment in children's human capital. A related question is whether the changes in income inequality experienced by many countries since the 1970s have been accompanied by changing intergenerational elasticities. In most of the time-trends research conducted so far, the time spans and sample sizes have been too limited to permit strong conclusions.

Cross-country comparisons have only begun to illuminate *why* intergenerational income associations are as large (and as small) as they are. To what extent does intergenerational transmission occur because higher-income parents invest more in their children's human capital? What are the roles of genetic and cultural heritability? One intriguing line of research seeks clues from comparisons of relatives with varying degrees of genetic and environmental relatedness. Sibling studies of this type (Taubman, 1976; Bjorklund, Jantti and Solon, 2005) have compared correlations in socio-economic status among monozygotic twins, dizygotic twins, non-twin full siblings, half-siblings, and biologically unrelated adoptive siblings, and also have compared biological siblings reared together and apart. (A related literature – Solon, Page and Duncan, 2000; Page and Solon, 2003a, 2003b; Oreopoulos, 2003; Raaum, Salvanes and Sorensen, 2006 – has compared sibling correlations and correlations among unrelated children that grew up in the same neighbourhood. The typical finding that the sibling correlations are considerably larger than the neighbour correlations suggests that family influences loom larger than

neighbourhood influences in accounting for the effects of origins on socioeconomic outcomes.) Intergenerational studies (Bjorklund, Lindahl and Plug, 2006; Plug, 2004; Sacerdote, 2004) have compared parent-child outcome associations in biological and adoptive families. Some empirical patterns consistent with an important role for genetic transmission are as follows: outcome correlations are particularly high among monozygotic twins; correlations for dizygotic twins and non-twin full siblings exceed those for half-siblings and adoptive siblings; correlations for biological siblings are positive even when the siblings are reared apart; intergenerational associations are higher for biologically related parents and children; and adoptive children's outcomes are positively associated with those of their biological parents (even after the adoptive parents’ outcomes are controlled for). Empirical patterns consistent with important environmental factors are as follows: outcome correlations are positive among biologically unrelated adoptive siblings; correlations among biological siblings tend to be higher when the siblings are reared together; and adoptive children's outcomes are positively associated with those of their adoptive parents (even after the biological parents’ outcomes are controlled for).

See Also

- income mobility

Bibliography

Becker, G. and Tomes, N. 1979. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87, 1153–89.

Becker, G. and Tomes, N. 1986. Human capital and the rise and fall of families. *Journal of Labor Economics* 4, S1–S39.

Bjorklund, A., Jantti, M. and Solon, G. 2005. Influences of nature and nurture on earnings variation: a report on a study of various sibling types in Sweden. In *Unequal Chances: Family Background and Economic Success*, ed. S. Bowles, H. Gintis and M. Osborne Groves. New York: Russell Sage Foundation.

Bjorklund, A., Lindahl, M. and Plug, E. 2006. The origins of intergenerational associations: lessons from Swedish adoption data. *Quarterly Journal of Economics* 121, 999–1028.

Chadwick, L. and Solon, G. 2002. Intergenerational income mobility among daughters. *American Economic Review* 92, 335–44.

Dunn, C.E. 2004. The intergenerational transmission of earnings: evidence from Brazil. Doctoral dissertation. University of Michigan.

Ferreira, S. and Veloso, F. 2004. Intergenerational mobility of wages in Brazil. Mimeo. Rio de Janeiro: IBMEC Business School.

Grawe, N. 2004. Intergenerational mobility for whom? The experience of high- and low-earning sons in international perspective. In *Generational Income Mobility in North America and Europe*, ed. M. Corak. Cambridge: Cambridge University Press.

Haider, S. and Solon, G. 2006. Life-cycle variation in the association between current and lifetime earnings. *American Economic Review* 96, 1308–20.

Jencks, C. and Tach, L. 2005. Would equal opportunity mean more mobility? In *Mobility and Inequality: Frontiers of Research from Sociology and Economics*, ed. S. Morgan, D. Grusky and G. Fields. Stanford: Stanford University Press.

Lam, D. and Schoeni, R. 1994. Family ties and labor markets in the United States and Brazil. *Journal of Human Resources* 29, 1235–58.

Mazumder, B. 2005. Fortunate sons: new estimates of intergenerational mobility in the U.S. using Social Security earnings data. *Review of Economics and Statistics* 87, 235–55.

Mulligan, C. 1997. *Parental Priorities and Economic Inequality*. Chicago: University of Chicago Press.

Oreopoulos, P. 2003. The long-run consequences of living in a poor neighborhood. *Quarterly Journal of Economics* 118, 1533–75.

Page, M. and Solon, G. 2003a. Correlations between brothers and neighboring boys in their adult earnings: the importance of being urban. *Journal of Labor Economics* 21, 831–55.

Page, M. and Solon, G. 2003b. Correlations between sisters and neighbouring girls in their subsequent income as adults. *Journal of Applied Econometrics* 18, 545–62.

Plug, E. 2004. Estimating the effect of mother's schooling on children's schooling using a sample of adoptees. *American Economic Review* 94, 358–68.

Raaum, O., Salvanes, K. and Sorensen, E. 2006. The neighbourhood is not what it used to be. *Economic Journal* 116, 200–22.

Reville, R. 1995. Intertemporal and life cycle variation in measured intergenerational earnings mobility. Mimeo. Santa Monica, CA: RAND.

Sacerdote, B. 2004. What happens when we randomly assign children to families? Working Paper No.

10894. Cambridge, MA: NBER.

Solon, G. 1989. Biases in the estimation of intergenerational earnings correlations. *Review of Economics and Statistics* 71, 172–4.

Solon, G. 1992. Intergenerational income mobility in the United States. *American Economic Review* 82, 393–408.

Solon, G. 1999. Intergenerational mobility in the labor market. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Solon, G. 2002. Cross-country differences in intergenerational earnings mobility. *Journal of Economic Perspectives* 16(3), 59–66.

Solon, G. 2004. A model of intergenerational mobility over time and place. In *Generational Income Mobility in North America and Europe*, ed. M. Corak. Cambridge: Cambridge University Press.

Solon, G., Page, M. and Duncan, G. 2000. Correlations between neighboring children in their subsequent educational attainment. *Review of Economics and Statistics* 82, 383–92.

Taubman, P. 1976. The determinants of earnings: genetics, family, and other environments; a study of white male twins. *American Economic Review* 66, 858–70.

How to cite this article

Solon, Gary. "intergenerational income mobility." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000252> doi:10.1057/9780230226203.0820

The New Palgrave Dictionary of Economics Online

intergenerational transmission

Lance Lochner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Intergenerational transmission refers to the transfer of individual abilities, traits, behaviours and outcomes from parents to their children. This article analyses the key theoretical and empirical issues in studies of intergenerational transmission of educational attainment, welfare receipt and fertility. Mechanisms that lead to intergenerational transmission of these outcomes are discussed in detail. The role of government policy in affecting intergenerational transmission is also considered.

Keywords

ability; altruism; borrowing; credit markets; education; educational attainment; educational subsidies; fertility; human capital; inequality; instrumental variables; intergenerational correlation; intergenerational income mobility; intergenerational transmission; patience; preferences; risk aversion; skills; social mobility; twin studies; unobserved heterogeneity; welfare receipt

Article

Intergenerational transmission refers to the transfer of individual abilities, traits, behaviours and outcomes from parents to their children. Economists have largely focused on the intergenerational transmission of educational attainment, earnings and income, wealth, fertility decisions and welfare receipt. When intergenerational transmission is strong, children turn out much like their parents, and social mobility is low.

Raw intergenerational correlations in education, earnings, teenage childbearing and welfare receipt in the United States are sizable. Correlations between parents’ and children's educational attainment and earnings are both around 0.4. Daughters of teenage or welfare mothers are nearly twice as likely to have a child when they are teenagers compared to daughters of older or non-welfare mothers. Mothers who grew up in a welfare family are four to six times more likely to receive welfare themselves than other mothers.

What gives rise to the intergenerational transmission of these outcomes? Parents may genetically pass on abilities, endowments, or preferences to their children that predispose them to choose actions similar to those they themselves chose. This can generate an intergenerational correlation in outcomes even if there is no actual causal effect of a parent's behaviour or outcome on the child. However, parents' actions themselves may encourage their children to take similar actions. For example, parents' schooling choices may directly impact on their children's decisions to stay in school. Intergenerational transmission incorporates both causal and non-causal channels.

Identifying the mechanisms that lead to the intergenerational transmission of education, earnings, fertility, or welfare receipt is central to understanding the role played by economic conditions or government policies in shaping those relationships. For example, if differences in earnings or welfare receipt primarily reflect differences in genetically endowed abilities, then policies to expand educational opportunities may have little effect on the intergenerational transmission of earnings and welfare. On the other hand, if ability primarily influences earnings by altering individual access to, or the financial rewards from, schooling, then college subsidies for low-income families should weaken the link between parents' and their children's earnings and welfare receipt.

This article offers detailed analyses of the key theoretical and empirical issues in studies of intergenerational transmission of educational attainment, welfare receipt and fertility. See also intergenerational income mobility for a discussion of earnings and income transmission.

Educational attainment

The economics literature has emphasized the role of skill and human capital development in analysing intergenerational transmission. To begin, consider an overlapping generations economy that generalizes the model of Becker and Tomes (1986) in which parents choose between investing in their children's human capital, their own current consumption, and borrowing or saving in the form of debts or bequests left for their children. Parents care about their own current consumption, but they also care about the consumption of their children and all future generations. While schooling is costly for parents and children, it raises human capital (or skill) levels, which increases subsequent earnings. Suppose that the production of a child's human capital, H_c, depends positively on parental human capital levels, H_p, the child's 'natural' ability, A_c, and the total years of child schooling, s_c, such that $H_c=h(H_p, A_c, s_c)$. Further, assume that both child ability and parental human capital raise the marginal productivity of schooling (that is, $\partial^2 h/\partial s_c \partial H_p \geq 0$ and $\partial^2 h/\partial s_c \partial A_c \geq 0$). These assumptions imply that: (*a*) for any given level of child investment or schooling, an increase in parental education or child ability produces more child human capital and (*b*) more able children from more educated parents will tend to invest more in their skills through schooling. Finally, assume that the abilities of children and parents are positively (but not perfectly) correlated. That is, bright parents tend to have bright children while dull parents tend to have dull children, but there is, on average, regression to the mean.

If parents are free to leave any amount of bequests/debts to their children, optimal child schooling for each generation will be chosen to maximize discounted earnings less investment costs. In this case, a child's schooling, $s_c=\sigma(H_p, A_c, \pi)$ is an increasing function of his parents' human capital and his own ability, while it is decreasing in the price of schooling, π. Importantly, the optimal schooling level will not depend on parental earnings or wealth, although it may be correlated with both since they depend on

parental abilities and human capital. In this simple model, a positive correlation in schooling between parents and children arises for two primary reasons: (1) parental human capital directly raises the productivity of child schooling and (2) abilities are positively correlated across generations and ability raises the productivity of schooling. Economists interested in identifying the 'causal effect' of parental schooling on child schooling attempt to estimate effect (1). This reflects the amount child schooling would increase if policy interventions were to raise parental schooling (and all else were held constant). Effect (2) depends on the intergenerational transmission of ability. As this is driven by genetics, it reflects the main role played by nature. If a child's human capital depended only on his own ability and schooling (so $\partial h/\partial H_p=0$ as assumed in Becker and Tomes, 1986), only effect (2) would matter, and the intergenerational transmission of educational attainment would be driven by the intergenerational transmission of ability. Even in this case, nurture plays a role in that schooling and other family investments are choices made by families. When schools change their prices (or quality), schooling decisions and the intergenerational transmission of educational attainment are affected.

Imperfect credit markets with limited borrowing opportunities also weaken the link between ability and schooling for poor families. When poor parents cannot borrow against their own future earnings or leave debts for their children, they may be forced to compromise on both their own consumption and schooling for their children (see, for example, Becker and Tomes, 1986; Caucutt and Lochner, 2006). Among constrained families, schooling choices depend on family income, I_p, so that $s_c=\sigma'(H_p, A_c, \pi, I_p)$ where $\partial\sigma'/\partial I_p \geq 0$. Poorly educated (and, consequently, low-income) parents lucky enough to have bright children may not be able to afford the efficient amount of schooling for them. (This need not be true when parental human capital has a very strong effect on the marginal product of schooling; in this case, poorly educated parents may not want to invest much in their children, even when they are bright.) This implies a strong intergenerational transmission of schooling among the least educated who cannot escape their misfortune. Since more educated and wealthier parents can afford efficient investments in their children, their behaviour is driven by the forces described earlier (that is, $s_c=\sigma(H_p, A_c, \pi)$). That the most disadvantaged underinvest in their children (while the most advantaged do not) when borrowing opportunities are limited implies that policies designed to subsidize the schooling of poor children will help to reduce economic inequality while improving aggregate efficiency.

Most researchers agree that the primary reason many college-age children from poor families do not attend college is that they are ill-prepared and not because they are unable to borrow for college. This raises the question as to whether these youths are ill-prepared because their parents have been unable to borrow the resources needed to prepare them for college in the first place. Direct evidence is scant, but indirect evidence suggests that poor parents sometimes fail to make early educational investments in their children that have substantial long-run payoffs. Cunha et al. (2007), therefore, argue that policies promoting early investments (for example, pre-school) in children do not face the same equity-efficiency trade-off that late investments (for example, college or post-school training) do.

The intergenerational transmission of preferences (for example, altruism, patience, or risk aversion) and other causal channels (for example, schooling may stimulate intellectual curiosity that is passed on to children) may also play important roles in the intergenerational transmission of education. While Mulligan (1997) explores the implications of endogeneous altruism, most economists have not incorporated these channels into their theoretical models.

The empirical literature typically considers a linearized version of the schooling decision described

earlier:

$$S_{ci} = \alpha S_{pi} + \beta A_{ci} + \gamma I_{pi} + X_i\delta + \varepsilon_i, \quad (1)$$

where X_i reflects variables that may affect the costs or benefits of schooling (for example, parenting skills, neighbourhood characteristics, school quality, or tuition prices) for child *i*. With ideal data, estimates from this equation inform us about the schooling choice function. Estimates of α tell us the direct effect of an increase in parental schooling, net of any effects parental schooling has on family income (or neighbourhood and school characteristics included in *X*). To obtain the total effect of parental education ($\alpha_T = \alpha + \gamma\, \partial I_p/\partial S_p + \delta\, \partial X/\partial S_p$), one must incorporate its effects through family income and the *X* variables. These effects are typically referred to as causal effects, since they measure how much a change in parents' education causes children's education to change. Most empirical studies suggest that the difference between α and α_T is small. See Haveman and Wolfe (1995) or Behrman (1997) for surveys of standard multivariate regression estimates of eq. (1).

Since data do not typically contain reliable measures of child ability, neighbourhood and school peer quality, or parental skills in bringing up children, most regression-based estimates of eq. (1) are probably upward biased for α. Researchers have begun to exploit three alternative econometric techniques that aim to reduce or eliminate biases arising from these types of unobserved factors: comparisons of children born of twin mothers or fathers, studies of adopted children, and instrumental variable approaches.

Some researchers have estimated how schooling differences between cousins whose parents are identical twins depend on the educational differences between their twin parents. This approach assumes that schooling differences among twin mothers or fathers are random rather than the result of different abilities or environments – an assumption often questioned. If the effects of unmeasured ability and parenting skill differences are additively separable from the effects of parental schooling, within-twin-parent estimators remove the effects of genetic differences in parental ability (from the twin parent side of the family) as well as any variation in the twins' parenting skills owing to the similarity of their upbringing – two potential sources of bias. Twin-parent-based estimates generally imply an important role for unobserved ability and parenting skills in determining child schooling levels. Using recent US data on the children of twins, Behrman and Rosenzweig (2002) find that within-twin-parent estimates of the effect of father's schooling are positive and statistically significant, while the estimated effect for mother's schooling is not. That is, differences in schooling between cousins with fathers who are twins are positively correlated with the difference between their fathers' schooling. For cousins with twin mothers, differences in child and differences in mothers' schooling are uncorrelated. (Controlling for differences in spouses' schooling or earnings has little effect on these conclusions.) In explaining the finding that a mothers' schooling does not affect child schooling, the authors argue that more educated mothers spend more time working and may, therefore, spend less time bringing up their children. However, this was not true in the 1970s in the United States (Leibowitz, 1974) nor is it true today in

rural India (Behrman et al., 1999), where women work little outside the home. This shows that the economic environment plays an important role in determining intergenerational relationships.

A different approach estimates the effects of parents' schooling on adopted children. When the effects of nature and nurture are additively separable and adoptees are randomly assigned to adoptive parents, the estimated effects of the adoptive parents' education on adoptees' schooling eliminates any bias due to the genetic transmission of ability. Under these circumstances, the estimated effects from adoptees provide a measure of the role played by nurture. However, they need not reflect the causal effect of parental education if some unobserved parenting skills are correlated with (but not caused by) parents' educational attainment. Bjorklund, Lindahl and Plug (2006) use a unique data-set from Sweden that contains educational attainment for adopted children and *both their biological and their adopting parents*. This enables them to regress adoptee schooling on the schooling of both biological parents, both adoptive parents, and even the interaction of biological and adoptive parents' schooling. While their results suggest important effects of the biological and adoptive father's and biological mother's education on their children, evidence of the adoptive mother's role is mixed. Interestingly, they estimate a positive and significant interaction between the biological and adoptive mother's schooling, suggesting an important nature–nurture complementarity. This interaction raises questions about methods that rely on the assumption that genetic and environmental effects are additively separable (for example, twin-parent studies or other adoptee studies that do not use data on both biological and adoptive parents).

Finally, some recent studies use changes in compulsory schooling laws in the United States and Europe as instrumental variables for changes in parental schooling. The legal changes largely affect the educational outcomes of parents at the low end of the distribution; thus, the studies' findings measure the impacts of increasing schooling among less-educated parents. Furthermore, the laws alter the population distribution of schooling, which may impact marriage markets. As such, they do not necessarily measure the effects of changing a single parent's schooling level. A Norwegian study (Black, Devereux and Salvanes, 2005) estimates little causal effect of parental schooling (except for the mother–son relationship) when using an increase in compulsory schooling as an instrument, but the effects are not very precisely estimated. By contrast, a US study (Oreopoulos, Page and Stevens, 2006) finds that a mother's and father's education has a significant effect on the probability that a young child is a year behind at school.

To summarize, most researchers conclude that parental education has a causal effect on child education, albeit substantially smaller than raw correlations suggest. While a few recent studies that compare children with twin parents or that focus on adopted children suggest that changes in a mother's education may have very small effects, instrumental variables studies do not confirm this pattern. Adoptee studies suggest that the educational outcomes of biological parents are important even when the child is brought up by others. Thus, the genetic transmission of abilities and preferences plays an important role in intergenerational transmission. Bjorklund, Lindahl and Plug (2006) estimate an important interaction between nature and nurture that is often neglected in empirical analyses. Finally, even studies that estimate causal effects do not separately identify the mechanisms by which parents' schooling affects child schooling. We are still left wondering whether schooling changes the preferences or information of parents, or whether it changes the marginal productivity of investing in one's children.

Teenage and non-marital fertility and welfare receipt

Studies of intergenerational fertility transmission have typically focused on non-marital and teenage births, as these are often associated with a wide range of negative outcomes for mothers and their children. Studies of intergenerational welfare receipt invariably discuss intergenerational patterns for education, earnings, and fertility as well. Economic theories of fertility (for example, Becker, 1991) generally say little about intergenerational patterns in childbearing and marital decisions. Formal economic models of intergenerational welfare transmission are also notably absent. Despite a lack of formal theory, social scientists have identified a number of factors that may affect the intergenerational transmission of fertility and welfare outcomes, including intergenerational correlations in cognitive ability, age of puberty, education and earnings. Economists are most interested in causal channels, however. Studies of teenage and non-marital fertility often refer to parental role-model effects and the impacts of early/non-marital childbearing on subsequent family structure and economic resources. Studies of intergenerational welfare patterns stress that parental welfare receipt may affect children's views about accepting public transfers, inform children about the welfare system, limit connections in and information about the labour market, and augment family resources.

Empirical researchers primarily aim to estimate the causal effects of parental teenage or out-of-wedlock childbearing and welfare receipt on daughters’ choices; however, it is difficult to separate causal effects from other factors that contribute to intergenerational correlations. Analyses typically employ multivariate regression techniques to control for measured family and environmental conditions, but concerns about unobserved heterogeneity plague most studies. Unmarried welfare mothers almost certainly differ from married mothers who are not on welfare, even when current family income and other observable characteristics are the same.

Kahn and Anderson (1992) estimate very different roles of teen motherhood on the fertility decisions of black and white children. They find that teen motherhood largely affects white daughters’ marital teen childbearing whereas black daughters’ non-marital teen childbearing is most affected. Differences in family background drive much of the intergenerational correlation of teen motherhood for whites but not blacks. Biological links related to the age of puberty play no role in teen fertility for either race. Two more studies (Haveman, Wolfe and Pence, 2001; Wolfe, Wilson and Haveman, 2001) separate the effects of the mother's age and her marital status at childbirth on the probability that a daughter has an out-of-wedlock birth as a teenager. The first study finds that mother's age is the more important factor, while the second concludes that marital status is more important. There is no consensus in the literature as to the relative importance of mother's age or marital status at the time of birth on her daughter's subsequent fertility decisions.

Most empirical studies of intergenerational welfare receipt control for parental income levels (or welfare eligibility), and attempt to estimate how parental welfare acceptance itself affects daughters’ future welfare receipt. Some studies use instrumental variables (typically, local unemployment rates or state welfare benefit levels) to further account for unobserved heterogeneity in family tastes or productivity levels (for example, Levine and Zimmerman, 1996; Pepper, 2000). Gottschalk (1996) exploits the timing of parental welfare receipt (while the daughter lives at home and afterwards) in an attempt to control for unobserved permanent family characteristics. These studies generally conclude that parental welfare receipt increases the daughter's subsequent welfare receipt and childbearing, but much (or even most) of the raw intergenerational correlation is attributed to the correlation in both income and unobserved heterogeneity. Recent studies suggest that there is a small positive causal effect of family income on

children's educational outcomes (for example, see Dahl and Lochner, 2006); however, most intergenerational welfare studies find that income-enhancing effects from parental welfare payments do not reduce the probability of daughter's welfare receipt enough to offset other direct effects on daughters’ tastes or information.

See Also

- education production functions
- family economics
- human capital
- intergenerational income mobility

Bibliography

Becker, G.S. 1991. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Becker, G.S. and Tomes, N. 1986. Human capital and the rise and fall of families. *Journal of Labor Economics* 4, S1–39.

Behrman, J. 1997. Mother's schooling and child education: a survey. Working Paper No. 97–025, PIER.

Behrman, J., Foster, A., Rosenzweig, M. and Vashishtha, P. 1999. Women's schooling, home teaching, and economic growth. *Journal of Political Economy* 107, 682–714.

Behrman, J. and Rosenzweig, M. 2002. Does increasing women's schooling raise the schooling of the next generation? *American Economic Review* 92, 323–34.

Bjorkland, A., Lindahl, M. and Plug, E. 2006. The origins of intergenerational associations: lessons from Swedish adoption data. *Quarterly Journal of Economics* 121, 999–1028.

Black, S., Devereux, P. and Salvanes, K. 2005. Why the apple doesn't fall far: understanding intergenerational transmission of human capital. *American Economic Review* 95, 437–42.

Caucutt, E. and Lochner, L. 2006. Early and late human capital investments, borrowing constraints, and the family. Working paper, University of Western Ontario.

Cunha, F., Heckman, J.J., Lochner, L. and Masterov, D. 2007. Interpreting the evidence on life cycle skill formation. In *Handbook of the Economics of Education*, vol. 1, ed. E. Hanushek and F. Welch. Amsterdam: North-Holland.

Dahl, G. and Lochner, L. 2006. The impact of family income on child achievement. Working Paper No. 11279. Cambridge, MA: NBER.

Gottschalk, P. 1996. Is the correlation in welfare participation across generations spurious? *Journal of Public Economics* 63, 1–25.

Haveman, R. and Wolfe, B. 1995. The determinants of children's attainments: a review of methods and findings. *Journal of Economic Literature* 33, 1829–78.

Haveman, R., Wolfe, B. and Pence, K. 2001. Intergenerational effects of nonmarital and early childbearing. In *Out of Wedlock: Causes and Consequences of Nonmarital Fertility*, ed. L. Wu and B. Wolfe. New York: Russell Sage.

Kahn, J. and Anderson, K. 1992. Intergenerational patterns of teenage fertility. *Demography* 29, 39–57.

Leibowitz, A. 1974. Education and home production. *American Economic Review* 64, 243–50.

Levine, P. and Zimmerman, D. 1996. The intergenerational correlation in AFDC participation: welfare trap or poverty trap? Discussion Paper No. 1100–96, Institute for Research on Poverty.

Mulligan, C. 1997. *Parental Priorities and Economic Inequality*. Chicago: University of Chicago Press.

Oreopoulos, P., Page, M. and Stevens, A.H. 2006. The intergenerational effects of compulsory schooling. *Journal of Labor Economics* 24, 729–60.

Pepper, J. 2000. The intergenerational transmission of welfare receipt: a nonparametric bound analysis. *Review of Economics and Statistics* 82, 472–88.

Wolfe, B., Wilson, K. and Haveman, R. 2001. The role of economic incentives in teenage nonmarital childbearing choices. *Journal of Public Economics* 81, 473–511.

How to cite this article

Lochner, Lance. "intergenerational transmission." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_F000317> doi:10.1057/9780230226203.0821

The New Palgrave Dictionary of Economics Online

intergovernmental grants

Nora E. Gordon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Intergovernmental grants are payments from one level of government to another, such as from the federal government to a state government, or from a city to a school district. Theoretically, such grants allow more local choice in public goods provision than purely centralized provision would, while still enabling some redistribution across local jurisdictions. Empirical research on these grants has focused on the extent to which these grants ultimately affect spending by receiving jurisdictions, both on the intended programme area and overall, and on other unintended consequences of the grants.

Keywords

block grants; bureaucratic capture; crowding out; fiscal federalism; flypaper effect; intergovernmental grants; interjurisdictional spillovers; matching grants; public spending; targeted public spending; Tiebout hypothesis

Article

Intergovernmental grants are payments from one level of government to another, such as from the federal government to a state government, or from a city to a school district.
Intergovernmental grants are widely used in the United States across a range of policy functions and are an important tool for redistribution in a federalist context. Under the Tiebout hypothesis, providing public goods locally rather than centrally improves match quality between individual preferences and local provision levels and generates competition in efficiency of public goods provision across communities, limiting bureaucratic capture. In a purely local system, however, any spillovers to public spending across local jurisdictions generate inefficient levels of public spending, and the ability to redistribute is limited to within local borders. Intergovernmental grants provide a mechanism to retain some benefits of local provision, while allowing for more optimal levels of public spending in the presence of interjurisdictional spillovers and increasing the capacity for redistribution.

The economic literature on intergovernmental grants investigates both their fiscal and their non-fiscal effects. Research on the fiscal impact of intergovernmental grants focuses on the extent to which they supplement local revenue formerly dedicated to the programme area, rather than supplanting it. Because intergovernmental grants are used in such a variety of policy functions, they have the capacity – especially if they do not crowd out local revenue – to affect a wide range of non-fiscal outcomes. Before discussing the research on the effects of intergovernmental grants, I briefly discuss the main types of intergovernmental grant structures.

Block grants and matching grants

The most important distinction between block grants and matching grants is that matching grants change the relative prices facing the receiving jurisdiction, making the publicly provided good or service in question relatively cheaper, while block grants provide income but do not change prices. Both types of grant typically are directed to particular agencies or programmes.

Block grants transfer funds from one jurisdiction to another, and are theoretically equivalent to the receiving jurisdiction facing a positive income shock from any source. A conditional block grant requires that the receiving jurisdiction spend at least the grant amount on the governmental activity targeted by the grantor jurisdiction. The extent to which the condition is binding depends on the preferences of the receiving jurisdiction. Despite this constraint, the fungibility of grant income makes it difficult to force receiving jurisdictions to increase spending by the full grant amount. Grantor jurisdictions often attempt to address this issue through ‘maintenance of effort’ requirements, by which receiving jurisdictions are required to continue funding the programme to which the grant is dedicated at some set percentage of previous years’ levels in order to receive the grant.

When a grantor jurisdiction offers matching grants, it sets a rate at which it will match contributions from the grantee jurisdiction. These rates may vary depending on the level of contributions. Matching grants differ from block grants in fundamentally changing incentives for spending on education by making education spending ‘cheaper’ than other spending.

Data on intergovernmental grants in the United States

The Census of Governments, conducted every five years in years ending in 2 and 7, collects data from states, counties, cities and other municipalities, independent schools districts, and special districts on all revenues and expenditures, including intergovernmental grants. For intergovernmental grants, the Census of Governments details the source of revenue or destination of payments (federal, state, or local) and the policy function to which it is dedicated (for example, health, education, or fire).

Evidence of fiscal impacts: the flypaper effect

Economic theory predicts that a jurisdiction receiving an intergovernmental lump-sum grant targeted to a particular function of government will view the grant as income and spend it as such, with a fraction going to the targeted function, and the remainder going to other projects or to private consumption through reductions in tax rates. Many empirical studies, however, have observed that the marginal

propensity to spend an intergovernmental grant on public expenditures is higher than the marginal propensity to spend other income on public expenditures. Arthur Okun called this phenomenon the flypaper effect, because the money ‘sticks where it hits’ (see Gramlich, 1977). There are three main categories of explanation for these observed effects: (*a*) they are real and reflect the preferences of bureaucrats but not of voters; (*b*) they are real and reflect voters’ preferences, but voter preferences may reflect some behavioural anomalies, such as loss aversion and lack of fungibility; (*c*) they are not real, but are generated by econometric misspecification. Hines and Thaler (1995) describe more specific cases within these categories in detail.

Given the current and historical prevalence of such grants, whether they ultimately supplement, or ‘stick to’, local spending is, unsurprisingly, the subject of a lengthy empirical literature, of which Hines and Thaler provide an excellent review. Studies included typically find that intergovernmental grants increase expenditures on the targeted programme by 25 to 100 per cent of the grant amount, with most estimates clustered at the high end of the range. This is much more than the receiving government's estimated propensity to spend on public programmes out of regular income (here Hines and Thaler estimate that only five to ten per cent of new non-grant income would be spent on public programmes), corresponding to a strong flypaper effect. One of the most convincing studies in their review is that of Ladd (1992), which shows that plausibly exogenous increases in state tax bases (stemming from the fact that some states link their tax base definition to the federal one, and exploiting changes in the federal income tax base following the Tax Reform Act of 1986) generate increases of about 40 per cent in state revenue. Many other studies simply correlate intergovernmental grants with spending, often in a cross-sectional context, without regard to potential bias from the fact that the same factors which make some jurisdictions receive more intergovernmental payments in a particular policy area may also make them have higher demand for public spending in that area.

Several recent additions to this literature have focused more explicitly on isolating exogenous variation in grant levels, and in doing so have yielded much less ‘sticky’ results. Knight (2002) accounts for political endogeneity in the amount of federal highway aid received by states by exploiting variation in legislative bargaining power due to seniority of state representatives in the US House. His technique reveals significant crowd-out of states’ own support of their highway programmes.

A number of recent papers focus on the heterogeneity of flypaper effects. Gordon (2004) shows that governments receiving intergovernmental grants may need time to adjust other revenue sources in response. Federal Title I grants to school districts for compensatory education, based largely on child poverty counts, appeared to stick completely to school spending in the first year following a shock to grant amount after the release of new census poverty data. Three years after the shock, however, there appears to be no effect on spending. Baicker and Staiger (2005) highlight the importance of institutional factors in determining how much receiving jurisdictions are capable of crowding out. In examining state responses to federal Medicaid Disproportionate Share Hospital (DSH) grants, they find that states which allow different levels of government to transfer funds directly between one another crowded out about half the federal grants. In states without this institutional capacity, the DSH funds were much stickier. Strumpf (1998) shows that the share of local spending on administrative overhead (a proxy for bureaucratic power) predicts the extent to which intergovernmental payments stick to local budgets, supporting a bureaucratic capture explanation of the flypaper effect.

Evidence of non-fiscal impacts

Intergovernmental grants have a wide range of effects, intended and unintended, on non-fiscal outcomes. The intended effects of intergovernmental grants may be due to the productive use of the grant. For example, Baicker and Staiger (2005) go on to show that federal DSH grants have significant impacts on mortality, despite the substantial crowd-out observed. Their findings suggest that the effects on mortality are due to the sticky part of the grant, which improves quality of hospital care. More often, studies evaluate the effect of the total intergovernmental grant amount rather than the effective or sticky grant amount on the outcome targeted by the grant. Such studies may conclude that public spending in that area is not effective, when in fact other revenue was crowded out so that total public spending in that area did not rise.

Jurisdictions making intergovernmental grants may do so to create incentives for the receiving governments that differ from simply spending the payment as designated. For example, Title I of the Elementary and Secondary Education Act of 1965 strengthened incentives for school districts to desegregate in compliance with the Civil Rights Act of 1964, and school districts responded accordingly (Cascio et al., 2005), though Title I funded compensatory education activities rather than desegregation-related costs. The current incarnation of this programme, the No Child Left Behind Act of 2001, similarly uses the threat of losing compensatory education funds as an incentive for schools to meet criteria for academic achievement growth benchmarks.

Finally, intergovernmental grants may create incentives that generate consequences unintended by the granting jurisdiction. For example, Cullen (2003) attributes 40 per cent of the significant rise in the special education classification of Texas public (government) school students from 1991 to 1996 to increased payments from the state to districts on a per-classified-student basis.

See Also

- fiscal federalism
- Tiebout hypothesis

Bibliography

Baicker, K. and Staiger, D. 2005. Fiscal shenanigans, targeted federal health care funds, and patient mortality. *Quarterly Journal of Economics* 120, 345–86.

Bailey, S.J. and Connolly, S. 1998. The flypaper effect: identifying areas for future research. *Public Choice* 95, 335–61.

Cascio, E., Gordon, N., Lewis, E. and Reber, S. 2005. Financial incentives and the desegregation of Southern schools. Working paper.

Cullen, J.B. 2003. The impact of fiscal incentives on student disability rates. *Journal of Public*

Economics 87, 1557–89.

Gramlich, E.M. 1977. Intergovernmental grants: a review of the empirical literature. In *The Political Economy of Fiscal Federalism*, ed. W.E. Oates. Lexington MA: Heath Publishers.

Gordon, N. 2004. Do federal funds boost school spending? Evidence from Title I. *Journal of Public Economics* 88, 1771–92.

Gruber, J. 2005. *Public Finance and Public Policy*. New York: Worth Publishers.

Hines, J.R. and Thaler, R.H. 1995. Anomalies: the flypaper effect. *Journal of Economic Perspectives* 9 (4), 217–26.

Knight, B. 2002. Endogenous federal grants and crowd-out of state government spending: theory and evidence from the federal highway aid program. *American Economic Review* 92, 71–92.

Ladd, H. 1992. State responses to the TRA86 revenue windfalls: a new test of the flypaper effect. *Journal of Policy Analysis and Management* 12, 82–103.

Oates, W.E. 1999. An essay on fiscal federalism. *Journal of Economic Literature* 37, 1120–49.

Strumpf, K.S. 1998. A predictive index for the flypaper effect. *Journal of Public Economics* 69, 389–412.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

How to cite this article

Gordon, Nora E. "intergovernmental grants." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000295> doi:10.1057/9780230226203.0822

The New Palgrave Dictionary of Economics Online

internal economies

G. Becattini

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Marshall introduced the idea of ‘internal’ economies, which accompany the growth of the ‘individual representative firm’, as opposed to the ‘external’ economies accompanying the growth of ‘a national or a local industry’. In principle the pursuit of internal economies would lead to a world composed of firms each one producing a great share of a very small range of commodities. But while Marshall shared the classical view of an increasing average size of the business unit, he put it into a dynamical and historical context. Tendencies and countertendencies may result in different outcomes in terms of market structures.

Keywords

concentration; external economies; internal economies; market structure; Marshall, A.; mechanization

Article

The expression ‘internal economies’ is considered here solely in terms of Alfred Marshall's own formulation, which is quite different from current terminology referring to internal economies of scale. Modern terminology refers to a reduction in the average cost of production of a well-specified commodity in relation to increases in the quantity produced, assuming, for every given quantity produced, the most appropriate utilization of the optimum productive plant. Marshall's concept of internal economies is analytically looser than this, but richer in empirical content and, possibly, in philosophical insight.

The twin terms, ‘internal’ and ‘external’ economies (and diseconomies) were first used by Marshall ‘for indicating the fundamental distinction between the “internal” economies and wastes which come with an increase in the size of the individual representative firm; and those “external” economies and wastes which come with an increase in the aggregate volume of a national or a local industry’ (Marshall, 1890, vol. 2, p. 347).

The main aim of the distinction was to help the applied economist in his attempts to disentangle the intricacies of contemporary socio-economic reality, rather than to provide an integral part of a formal theory of the relative values of the commodities; this is shown clearly enough by the ambiguous references to ‘national or local industry’ and by the use of such a fuzzy concept as the ‘individual representative firm’. We must add that many times Marshall conveys the impression of confining external economies in the straitjacket of a single-product, homogeneous industry. Moreover, we must bear in mind, as Loasby aptly pointed out, that Marshall ‘made no clear distinction between the theory of value and the theory of growth’ (Loasby, 1978, p. 1, n. 1).

This vein of Marshallian thought derives from three sources: his vast and detailed knowledge of the literature on contemporary British and American industry; his own ruminations on the Smith–Babbage arguments on the division of labour and the internal organization of the firm, and, finally, his own early studies of mental science.

There is a passage in the *Principles* that contains the kernel of the Marshallian ideas on the internal growth of the firm. ‘Practice makes perfect’, starts Marshall, taking up the well-known Smithian theme:

> physiology, [he continues], in some measure explains this fact. For it gives reasons for believing that the change is due to the gradual growth of new habits of more or less reflex or automatic action. Perfectly reflex actions … are performed by the responsibility of the local nerve centre without any reference to the supreme central authority of the thinking power, … But all deliberate movements require the attention of the chief central authority: it receives information from the nerve centre or local authorities and perhaps in some cases direct from the sentient nerves, and sends back detailed and complex instructions to the local authorities, or in some cases direct to muscular nerves, and so co-ordinates their action as to bring about the required results. (1890, vol. 1, pp. 250–1)

This quotation helps us put together the scattered pieces of the Marshallian theory of the growth of the firm under competitive conditions.

Under the spell of all the usual drives of the human mind (money-making propensity, ‘instinct of the chase, desire for fame’, and so on), a business unit, working in a competitive context, is subject to a continual pressure to rationalize its most typical recurring operations and the tools used. So we have, simultaneously, both the development of ‘skills’ (a ‘sort of capital of nerve force’), allowing the saving of time and of physical and, above all, nervous energies, and a rationalization of the process and the tools used. Alert to the danger of sliding to an abstract conception of the industrial process, Marshall makes room for historical and geographical peculiarities of the ‘skilling’ and ‘rationalizing’ processes. But there comes a point ‘when the action has thus been reduced to routine [that] it has nearly arrived at the stage at which it can be taken over by machinery’ (1890, vol. 1, p. 254) At this point it is very probable that someone will invest the money and the inventive power required for the realization of the appropriate appliance.

When a machine is introduced into a manufacturing firm, its product becomes more uniformly specified and a cumulative process of mechanization and standardization can start. Marshall speaks of ‘a great architectonic principle’ according to which

> a well-driven machine tool could become the parent of new machine work more exact

> than itself … and so on … By successive steps larger and more delicate work is thrown upon the apparatus … at last it becomes … a thinking acting on hints given from within…. When all is in order, the machine is nearly self-sufficient. (1890, vol. 1, pp. 206–7)

The gradual introduction of specialized machinery results in more time and more nervous energies being made free at the hierarchical summit of the firm, in such a way that the entrepreneur can devote more of his time and energies to the ‘broadest and most fundamental problems of his trade’ (1890, vol. 1, p. 284), that is, to the collection and evaluation of information about general market trends and technological and organizational innovations.

The growing of a business unit above the other units of an industry gives to it the opportunity of taking advantage of a better allocation of skills, of getting hold of ‘big brains’, of introducing innovations out of reach of the others, of obtaining better terms in buying, selling and borrowing. And consequently, in the words of Marshall: ‘lowers the price at which he can afford to sell’ (1890, vol. 1, p. 315).

The basic constraint to the development of the individual firm lies in the conflict between the urge of the entrepreneur to decipher the environmental conditions of growth and the organizational requirements of the productive process. From this second viewpoint the best results can be attained by concentrating the entrepreneur's efforts on a narrow range of tasks. The simpler the work of direction, the larger the volume of output which can be efficiently controlled by a single mind, the greater the scope for the introduction of machines and uniform continuous processes. It would seem that the combined effect of these constraints would be a world composed of firms each one producing a great share of a very small range of commodities.

But this outcome would be apparently self-destroying for a world of competitive (albeit imperfectly) firms, like the Marshallian one. Marshall's answer to this challenge is both complex and stimulating. First of all, to make use of all the possible internal economies, a certain amount of individual volition is needed. The entrepreneur ‘works hard and lives sparely … subordinates trust him and he trusts them … every improved process is quickly adopted …’. If this behaviour ‘could endure for a hundred years, he and one or two others like him would divide between them the whole of that branch of industry in which he is engaged’. But life is short and those who follow are not always fit to take over the task. The firms of many industries, at least before ‘the great recent development of vast joint-stock companies, which often stagnate but do not readily die’, like the trees of the forest ‘gradually lose vitality and one after another … give place to others’.

We must also remember that ‘many of the lines of division between the trades which are nominally distinct are becoming narrower and less difficult to be passed … A watch factory with those who worked in it could be converted without any overwhelming loss into a sewing-machine factory’ (1890, vol. 1, pp. 258–9). This continual trespassing of ‘industrial’ borderlines systematically frustrates the inner tendencies towards concentration and monopolization.

It must also be taken into account that the continual formation of economies, external to the single firm but internal, either to an industry or to some group of industries, in that they apply even to the smallest firms, systematically erodes part of the advantage of the bigger businesses. A particularly relevant example of this is provided by the case of a localized population (the Marshallian ‘industrial district’) of medium-small sized firms, which, grouping together and specializing in various stages of the production process, achieve many of the large-scale economies typical of the giant firms.

Marshall shares the classical view of an increasing average size of the business unit, but he is very careful to put it into a dynamical and historical context. Tendencies and countertendencies may result in different outcomes in terms of market structures. What is necessary for the process to be self-perpetuating is that the system should reproduce the complex of motivations which, given the structural characteristics of the industrial field, nourish the basic tendency of man towards liberation from purely mechanical tasks. In the words of R.A. Jenner: ‘external and internal economies thus form counterbalanced forces of competition around which the disturbing thrusts of evolutionary change are held in control’ (Jenner, 1964, p. 311).

See Also

- external economies
- returns to scale

Bibliography

Jenner, R.A. 1964. The dynamic factor in Marshall's economic system. *Western Economic Journal* 3(1), 21–38.

Loasby, B.J. 1978. Whatever happened to Marshall's theory of value? *Scottish Journal of Political Economy* 1, 1–12.

Marshall, A. 1890. *Principles of Economics*. 9th (Variorum) edition, 2 vols, ed. C.W. Guillebaud. London: Macmillan, 1961.

Marshall, A. 1919. *Industry and Trade*. London: Macmillan.

How to cite this article

Becattini, G. "internal economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000157> doi:10.1057/9780230226203.0823

The New Palgrave Dictionary of Economics Online

internal migration

James R. Walker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Migration is a shared topic within social sciences attracting interest from members of all sub-disciplines. This attention reflects both the importance of the flows and the complexity of the behaviour. This article presents a short overview of the basic theoretical perspectives on individual migration decision making, and it considers empirical challenges to bringing these models to the data.

Keywords

family decision making; human capital; internal migration; labour market search; learning; social networks

Article

Migration is a shared topic within social sciences attracting interest from members of all sub-disciplines. This attention reflects both the importance of the flows and the complexity of the behaviour. This article presents a short introduction to economic analyses of internal migration.

Theory

Since the seminal work of Sjaastad (1962), economists have recognized that migration is a form of human capital. In the simplest model of wealth maximization the fixed costs of moving are balanced against the net present value of earnings streams available in the alternative location. This framework explains why, as was first noted by Ravenstein (1885), migration is an activity primarily of the young. The young are most likely to move, according to the human capital perspective, for three related but distinct reasons. First, they should move to take advantage of economic opportunities as soon as they are independent economic actors. Second, the young have a longer horizon over which to amortize the fixed cost of migration; hence, relatively small gains in earnings may tip the scales in favour of moving. And

third, the young have fewer location-specific investments that serve to tie them to the current location (such as children).

Since its publication, Sjaastad's framework has been extended in a variety of ways (see Greenwood, 1997, for a useful summary). Wealth maximization, as a motivation for migration, has given way to utility maximization, with uncertainty, information and local amenities given special attention. Perhaps the richest set of behavioural models appears in the development literature, where models of family behaviour incorporating notions of risk sharing, intergenerational transfers, and household bargaining have been developed. (See Lucas, 1997, for a comprehensive summary, and Stark, 1991, for several case studies which tailor the model to a particular context or issue.)

Nearly all the research on migration adopts a static framework, usually within a binary mover–stay decision framework. A classic example is Mincer's paper ‘Family Migration Decisions’ (1978), which is an early contribution to the now popular area of decision making in multi-person households. Mincer assumes wealth maximization and that spouses have separate preferences and different opportunities across locations. His basic insight is that the location of an individual's maximum may not coincide with the location of joint maximum. Indeed, the location of the joint maximum may not coincide with the location of the individual maximum of either spouse. This gives rise to the concepts of ‘tied movers’ and ‘tied stayers’ and sharp predictions on who should remain married and who should separate. One of his interesting predictions is that the incidence of migration should increase soon after a divorce or separation as the now independent individuals move from their ‘tied’ locations. Mincer also predicts that these forces become stronger as women's labour force participation and earnings increase.

The limitations of a static framework are also evident in Mincer's paper. (It is noteworthy that to date no one has extended Mincer's work in a meaningful way.) Mincer presumes marriage and does not investigate who marries whom (forward-looking agents may consider the possible consequences of different spatial opportunities before consummating the match). And, restricted to a single period, the analysis cannot investigate the timing or temporal sequence of separation and migration. Indeed, to study temporal linkages of migration and other important life-cycle choices (such as marriage or retirement) requires a dynamic framework. And, as illustrated above, a static framework begs the question as to the nature of initial equilibrium. Allowing households to make multiple migration decisions substantially increases the model's complexity. Now the model must determine *where* and *when* to move. Moreover, prior moves influence subsequent opportunities, giving these models their own natural dynamics.

Empirical implementation

One of the first empirical regularities gleaned from individual migration histories is their diversity (DaVanzo and Morrison, 1981, is an early contribution). The richness of the life histories appears in the diverse terminology describing the types of moves. Concepts such as ‘repeat’ (an individual's second or higher order move), ‘onward’ (a move to a new location – all first moves are ‘onward’) and ‘return’ (movement back to a previous location, most commonly the individual's childhood location or self-identified ‘home’) appear. At an aggregate level, notions of ‘circular’ and ‘chain’ migration are commonly used.

Data from the US National Longitudinal Survey of Youth, 1979 Cohort (NLSY79) (US Department of

Labor, 2006), can be used to give an estimate of the magnitude of these flows. Data through 1994 when cohort members were in their mid-30s show that roughly 80 per cent of the cross-sectional sample had never moved out of their childhood state of residence, considered as their ‘home’ state. Of the 20 per cent of movers, more than half move again, and of the repeat movers, approximately 55 per cent ever return to their home state. Interestingly, few differences appear by gender, but the proportion of movers is U-shaped in completed education – individuals with a high school degree move the least, whereas those with some college or less than a high-school education are more likely to move. (Long, 1988; 1991; 1992, and the numerous reports of William Frey at the University of Michigan and the Brookings Institute are among the best sources of descriptive evidence on internal migration flows in the United States.)

To study the influence of labour market opportunities on migration, we would like to define locations corresponding to distinct local labour markets. Within the United States, if we define local labour markets crudely as equivalent to counties, the model admits a choice set of approximately 3,100 elements. (Models of residential choice and occupation are sometimes said to be isomorphic. From an empirical standpoint they are not. Models of occupation choice typically have relatively few alternatives, say five or ten, and educational or experience requirements or other characteristics offer a natural ordering to the occupational alternatives. See Neal, 1999, for a recent contribution.) Consequently, there is a fundamental trade-off between the economic definition of locations and statistical measures available. For this reason many studies of internal migration use the decennial Census. Yet the decennial Census has its limitations, most importantly that virtually all individual and household characteristics are measured as of the date of the census. Census data offer detailed descriptive summaries of migration flows over narrow geographic regions, but, with no measures of pre-migration characteristics, are of problematic use for unravelling cause and effect.

Extending the analysis to panel data and multiple decision periods makes greater demands on the data and the analysis. Opportunities must be measured for each period of time, and some decision must be made on the persistence of economic opportunities. As in models of job search, the analyst must decide whether ‘recall’ is available: do migrants have the ability to remember and possibly return to a previous wage offer? If so, the size of the state space within the dynamic program formulation increases exponentially with the agent's memory length. And an important empirical challenge for dynamic analyses is sample attrition, as not being able to locate a respondent who has moved is one of the reasons for not securing an interview. (Survey organizations quickly developed expertise in locating respondents in the early years of the large-scale surveys such as the Panel Studies of Income Dynamics and the National Longitudinal Surveys. Most commonly, respondents are not interviewed because they refuse, not because they could not be located by the survey organization. See Olsen and Reagan, 2000, for detailed information on the experience for the NLSY79.)

Nevertheless, Bellman's principle can be usefully applied to represent the decision problem of the individual (or household). Kennan and Walker (2005) adopt a dynamic programming approach for analysing the migration histories within the National Longitudinal Survey of Youth, 1979 Cohort. We find that earnings are an important (economic and statistical) determinant of migration flows, and their inclusion significantly improves the model's fit to the migration flows within the NLSY79. Respondents in the NLSY79 are more likely to leave a poor local labour market but do not necessarily move to ‘the best’ labour market as predicted by the model. Our findings are consistent with the interpretation that

economic factors are an important determinant of migration, but not the *only* factor.

Research frontiers

Research on internal migration flows remains on the frontier. Models and analysis of life-cycle migration are still in their infancy, with plenty of room for growth. An important challenge to decision theorists is developing frameworks for understanding return migration – why it is optimal to leave and return home. (Learning or more generally the resolution of uncertainty will be part of the explanation. For an early investigation based on learning, see Pessino, 1991.)

Investigating the timing and the relationship between migration and other life-cycle choices such as marriage and retirement is another likely active research area. Certainly the migration behaviour by baby boomers will be of increasing interest to federal and local policymakers.

There is broad consensus that economic and family factors are the primary determinants of internal migration flows. Yet no analysis satisfactorily combines both factors. The barriers to doing so are more empirical than conceptual. The set of family members who may potentially influence migration choice is large, with no consensus as to which relationships must be surveyed. Except for the central role of parents and children, there is little additional information to guide one's choice. Obtaining information on the spatial distribution of family members and their avenues of influence (for example, income pooling or information sharing) is time consuming and thus costly. The influence of broader social networks need also be considered. The Great Black Migration within the United States during the first half of the 20th century illustrates the importance of social factors for migration streams (Lemann's 1991 classic *The Promised Land: The Great Black Migration and How it Changed America* offers an engaging account of the family, social and economic factors that stimulated these flows). Migration research may offer another avenue to explore the influence of social interactions, and perhaps provide stronger ties among social science disciplines.

See Also

- collective models of the household
- family decision making
- human capital
- labour supply

Bibliography

DaVanzo, J.S. and Morrison, P.A. 1981. Return and other sequences of migration in the United States. *Demography* 18, 85–101.

Frey, W.H. 2006. US demographic publications: research reports by William H. Frey. Online. Available at http://www.frey-demographer.org/reports.html, accessed 21 November 2006.

Greenwood, M. 1997. Internal migration in developed countries. In *Handbook of Population and Family Economics*, vol. 1b, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Kennan, J. and Walker, J.R. 2005. The effect of expected income on individual migration decisions. Unpublished manuscript, University of Wisconsin-Madison.

Lemann, N. 1991. *The Promised Land: The Great Black Migration and How it Changed America*. New York: Vintage Books.

Long, L.H. 1988. *Migration and Residential Mobility*. New York: Russell Sage.

Long, L.H. 1991. Residential mobility differences among developed countries. *International Regional Science Review* 14, 133–47.

Long, L.H. 1992. Changing residence: comparative perspectives on its relationship to age, sex, and marital status. *Population Studies* 46, 141–58.

Lucas, R.E.B. 1997. Internal migration in developing countries. In *Handbook of Population and Family Economics*, vol. 1b, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Mincer, J. 1978. Family migration decisions. *Journal of Political Economy* 86, 749–73.

Neal, D. 1999. The complexity of job mobility of young men. *Journal of Labor Economics* 17, 237–61.

Olsen, R. and Reagan, P. 2000. You can go home again: evidence from longitudinal data. *Demography* 37, 339–50.

Pessino, C. 1991. Sequential migration theory and evidence from Peru. *Journal of Development Economics* 36, 55–87.

Ravenstein, E.G. 1885. The laws of migration. *Journal of the Statistical Society of London* 48, 167–235.

Sjaastad, L. 1962. The costs and returns of human migration. *Journal of Political Economy* 70, 80–9.

Stark, O. 1991. *The Migration of Labor*. London: Basil Blackwell.

US Department of Labor. 2006. National Longitudinal Survey of Youth 1979. Online. Available at http://www.bls.gov/nls/nlsy79.htm, accessed 30 November 2006.

How to cite this article

Walker, James R. "internal migration." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000358> doi:10.1057/9780230226203.0824

The New Palgrave Dictionary of Economics Online

internal rate of return

Harald Hagemann
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The internal rate of return of an investment project is that discount rate or rate of interest Y which makes the stream of net returns x_t associated with the project equal to a present value of zero. It is the solution for i in the following equation in which indicates the physical lifetime of the investment project.

$$C(0, \theta) = \sum_{t=0}^{\theta} x_t(1 + i)^{-t} = 0 .$$

The internal rate of return is compared with the market rate of interest in order to determine whether a proposed project should be undertaken or not.
Among the criteria to be used in determining the profitability of an investment project two others are frequently considered. Whereas the payout-period criterion is a crude rule of thumb which for much of the time ignores pattern of receipts, the net present value criterion is the most relevant ‘rule’ for optimal investment behaviour. If the present value (using the market rate of interest as the rate of discount) of a project's expected earnings is greater than its cost (including discounted future operating and maintenance costs), that is, if the net present value is positive, the investment project is potentially worth undertaking.
Whereas the net-present-value rule and the internal-rate-of-return rule lead to identical results in the two-period case and in the perpetuity case (which in essence is only a variant of the former), the two criteria may lead to different results in the multiperiod case. Figure 1 illustrates such a case in which the choice between two alternative investment options will lead to identical results for $i>i^*$ whereas the two criteria lead to different results for market rates of interest smaller than the cross-over rate i^* where the present value of i is higher while II has the higher internal rate of return. The failure of the internal rate of return criterion is the consequence of the implicit assumption that all intermediate receipts, positive or negative, are treated as if they could be compounded at the ‘*internal*’ rate of return itself whereas the only appropriate *external* discounting rate is the market rate of interest (reinvestment problem).
Figure 1

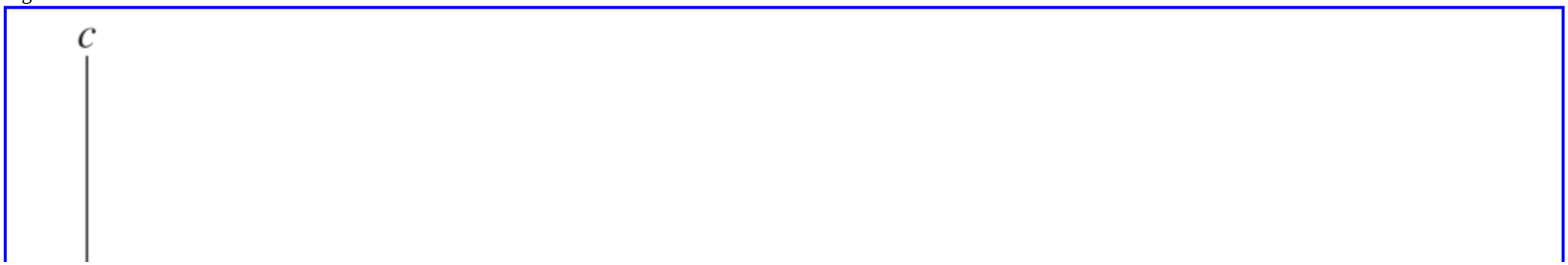

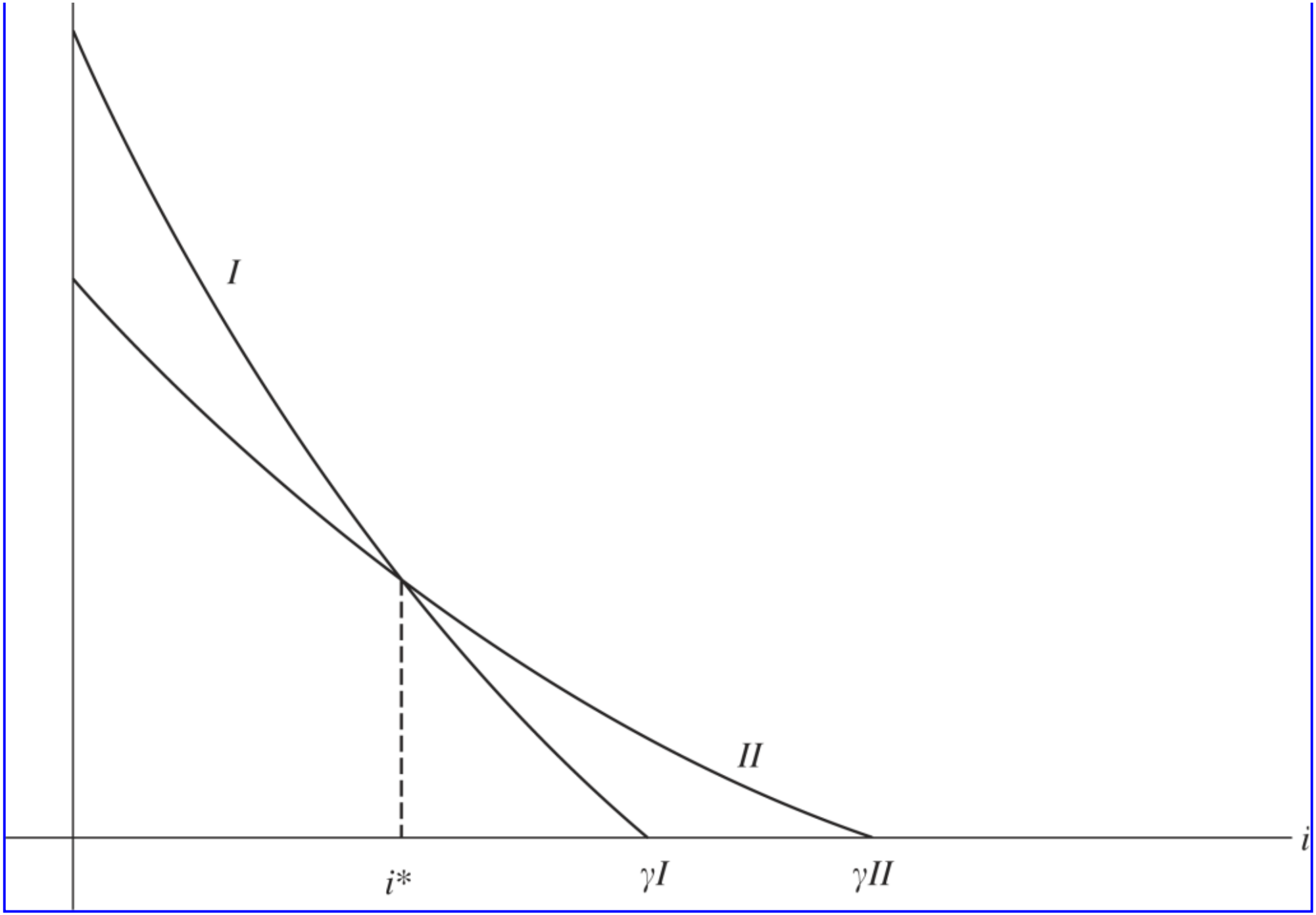

When the investment projects are independent and with a perfect capital market (in which the lending and borrowing rates of interest are identical) the net present value is, in general, the only universally correct criterion of appraising investment projects (see Hirshleifer, 1958 and 1970, ch. 3). For the multiperiod case the internal-rate-of-return rule is not generally correct. Furthermore, there may be *multiple rates of return* that will equate the present value of a project to zero. A necessary condition for non-uniqueness of the internal rate of return is that there be more than one change of sign in the stream of receipts over the lifetime of a project.

The controversy about the multiplicity of the internal rate of return in the late 1950s led to the development of the *truncation theorem*. This theorem turns out to be important for the general problem of choosing the optimal investment period (for a historical survey of truncation theorems see Matsuda and Okishio, 1977). In 1969 Arrow and Levhari presented a new version of the truncation theorem which contrasted sharply with the other economists' method of choosing a truncation period so as to maximize the internal rate of return. They rightly pointed out that this criterion would not be adequate for the choice of the truncation period. Instead they advocated the maximization of the present value of the investment project as the proper criterion. It was demonstrated that the possibility of truncating investment projects at any age different from their physical lifetimes and at no extra costs leads to the following results:

1. (1) The maximized present value of the project is a monotonically decreasing function of the rate of interest. A corollary of this is that the internal rate of return is always unique.
2. (2) A rise (fall) in the rate of interest will always lower (raise) the present value of the remaining future net returns at all stages of the production process. Consequently the optimal economic lifetime, too, is a monotonically decreasing function of the rate of interest.

Flemming and Wright (1971) dropped the assumption of a constant rate of discount per unit of time and tried to generalize the theorem to the case of different interest rates over time, a case where the deficiency of the internal-rate-of-return rule is most obvious. However, the 'generalization' does not take us very far because the calculation would require perfect foresight of future rates. The authors emphasize that a 'slight relaxation' of this assumption is allowed because 'a change in expectation which causes' all rates 'to be revised in the same "direction" will alter the present values of all costlessly terminable projects•…•in a common direction' (Flemming and Wright, 1971, p. 262). But even this proposition holds, in general, only when the change takes place uniformly, so that there is no change in the weights of the time pattern of the stream of net returns.

More interesting is the discussion of the impact of a consequence stream, that is, costs and benefits following from truncation. Whereas a positive *scrap value* can easily be incorporated the range of validity of the truncation theorem is severely limited in the case of *shut down costs*. Shut down costs can occur before and after truncation. Sen (1975) has shown that in the general case of a consequence stream following from truncation only minimal sufficiency conditions can be formulated: non-negative consequence sums (NCS) and non-negative consequence remainders (NCR), that is, the present value of the consequence stream for each t before and after the actual point θ of truncation has to be non-negative. Neither NCS nor NCR requires the present value of the consequence stream at θ to be non-negative, that is, a negative present value of the remaining process does not endanger the monotonicity result. But the conditions are very restrictive, because NCR is violated if the last item or the discounted value of the tail of the consequence stream is negative. This may be the case because of for example, redundancy payments, environmental protection or shut down costs of a nuclear power station.

The truncation theorem was originally developed in a *partial framework*. Nevertheless, Hicks (1973) and Nuti (1973) considered it applicable in a *general framework*. However, Eatwell's (1975) criticism of these authors has clarified that important propositions of the theorem do not carry over to the general framework (see also Hagemann and Pfister, 1978). At the partial level all prices in the economy are taken as given, that is, the individual's stream of net returns is considered not to be affected by changes in the discount rate. This assumption is impermissible when considering investment processes for society as a whole. At the general level the rate of profit is represented by the internal rate of return of the process as a whole for a given real wage. A variation of the discount factor, that is, the profit rate implies an opposite variation of the real wage rate. Because the present value of the whole process is both maximum and zero in competitive equilibrium the slope of the wage–profit curve is negative throughout. This is the only result one can draw under the conditions of the truncation theorem in the general setting. Neither the inverse relationship between the present value of the rest of the process and the rate of profit nor that between the optimal economic process length and the rate of profit invariably hold.

Furthermore, the analysis raises serious doubts as to the existence of an inverse monotonic relationship between interest and investment. The implication for Keynes's concept of the 'marginal efficiency of capital' is close at hand. As is well known, Keynes considered his concept 'identical' with Fisher's definition of the 'rate of return over cost' and stressed that there is no material difference 'between my schedule of the marginal efficiency of capital or investment demand-schedule and the demand curve for capital contemplated by some of the classical writers' (Keynes, 1936, pp. 140 and 178). To be sure, there are passages which indicate that Fisher was aware of the fact that prices and therefore not only the present values of the streams of net receipts but the net receipts themselves vary with variations in the rate of interest (see especially the 'more intricate than important' complication discussed in Fisher, 1930, pp. 170–71). However, the fixed-price assumption he commonly referred to implies a partial framework where the relationship between interest rates and prices is eliminated. It is therefore impossible to construct a demand curve for investment on the basis of a *ceteris paribus* clause for prices simply by variations of the rate of interest. An inverse macroeconomic relation between interest and investment cannot be derived from monotonicity results reached in a microeconomic framework. The difficulties encountered by Fisher and Keynes are discussed by Alchian and Garegnani from different points of view. Alchian (1955, p. 942) stresses that 'a schedule of investment demand at different market rates of interest requires that one compute the internal rates of return in terms of the prices that would prevail at each potential market rate of interest'. Garegnani (1978–9) brings into focus the problems involved in Keynes's concept of the schedule of the marginal efficiency of capital.

The return of the same truncation period and reswitching of techniques are closely linked phenomena occurring in a general framework. Some authors have tried to draw another analogy between the reswitching problem and the well-known possibility of the existence of multiple rates of return. Apparently the intention was to play down the importance of

reswitching. This is reflected by the proposition that ‘there is no new thing under the sun’ (Bruno, Burmeister and Sheshinski, 1966, p. 553). However, multiple internal rates of return are a phenomenon related to the partial framework from which a generalization to the general level is not admissible. Truncation ensures the uniqueness of the internal rate of return but cannot rule out reswitching. Therefore, an analogy between the two phenomena does not exist.

See Also

- investment decision criteria

Bibliography

Alchian, A.A. 1955. The rate of interest, Fisher's rate of return over costs and Keynes’ internal rate of return. *American Economic Review* 45, December, 938–43.

Arrow, K.J. and Levhari, D. 1969. Uniqueness of the internal rate of return with variable life of investment. *Economic Journal* 79, September, 560–66.

Bruno, M., Burmeister, E. and Sheshinski, E. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80, November, 526–53.

Eatwell, J. 1975. A note on the truncation theorem. *Kyklos* 28(4), 870–75.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Flemming, J.S. and Wright, J.F. 1971. Uniqueness of the internal rate of return: a generalisation. *Economic Journal* 81, June, 256–63.

Garegnani, P. 1978–9. Notes on consumption, investment and effective demand, I and II, *Cambridge Journal of Economics*, Pt. I, 2(4), December 1978, 335–53; Pt II, 3(1), March 1979, 63–82.

Hagemann, H. and Pfister, J. 1978. Zur Relevanz des Truncation-Theorems in partialanalytischer und totalanalytischer Sicht. *Jahrbücher für Nationalökonomie und Statistik* 193(4), August, 359–79.

Hicks, J. 1973. *Capital and Time. A Neo-Austrian Theory*. Oxford: Clarendon Press.

Hirshleifer, J. 1958. On the theory of optimal investment decision. *Journal of Political Economy* 66, August, 329–52.

Hirshleifer, J. 1970. *Investment, Interest, and Capital*. Englewood Cliffs, NJ: Prentice-Hall.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Matsuda, K. and Okishio, N. 1977. Theorems of investment truncation. *Annals of the School of Business Administration*, Kobe University, 73–90.

Nuti, D.M. 1973. On the truncation of production flows. *Kyklos* 26(3), 485–96.

Sen, A. 1975. Minimal conditions for monotonicity of capital value, *Journal of Economic Theory* 11(3), December, 340–55.

How to cite this article

Hagemann, Harald. "internal rate of return." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000159> doi:10.1057/9780230226203.0825

The New Palgrave Dictionary of Economics Online

international capital flows

Shang-Jin Wei

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Cross-border capital flows may be regarded as either too small (known as the Lucas paradox) or too big (against the Samuelson theorem of factor price equalization). The resolution to the conflicting views may require thinking out of the neoclassical box. In theory, international capital flows can promote economic growth, but the data do not reveal a strong, robust, and causal effect, particularly for developing countries. The theoretical results and the empirical patterns can be reconciled through either a composition effect or a threshold effect. Some emerging evidence suggests that the two effects are related.

Keywords

composition hypothesis on capital flows; corruption; economic growth; factor price equalization theorem; financial globalization; foreign aid; foreign debt; foreign direct investment; Heckscher–Ohlin–Samuelson model; institutional quality; international capital asset pricing model (ICAPM); international capital flows; Lucas paradox; portfolio equity flows; threshold hypothesis on capital flows; total factor productivity; trade costs

Article

Cross-border capital flows worldwide have risen substantially since the mid-1970s, from US$1.2 trillion in 1980 to $5.8 trillion in 2004. The pace of the growth (at an average annual rate of 6.6 per cent) surpasses by a big margin those of the world GDP (at 1.7 per cent per annum) and the world exports (at 3.1 per cent per annum). Developed economies are the most important source countries, accounting for 92 per cent of the aggregate outward capital flows in 2004. They are also the most important recipients, accounting for 91 per cent of the aggregate inward capital flows in 2004. A small number of developing countries – commonly known as emerging market economies – receive the lion's share (nearly 70 per cent) of the remaining international capital flows in 2004. More than 130 other developing economies

are more or less bypassed by the surge in the capital flows. (For these calculations, developed countries consist of the following 25 countries: Australia, Austria, Belgium, Canada, Cyprus, Denmark, Euro Area, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Japan, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, the United Kingdom, and the United States. Emerging market economies consist of the following 22 economies: Argentina, Brazil, Chile, China, Colombia, Egypt, Hong Kong SAR, Indonesia, India, Israel, Korea, Morocco, Mexico, Malaysia, Pakistan, Peru, Philippines, Singapore, Thailand, Turkey, Venezuela, and South Africa. Aggregate capital flows for any set of countries are calculated by summing up the values for individual countries in the set.)

The first part of this article, which draws from joint work (Ju and Wei, 2006), provides an analytical perspective on the volume of international capital flows, which can be regarded as either too low (known as the Lucas paradox) or too high (when compared with the logic of factor price equalization). The second part, which draws from a different set of recent work (Prasad et al., 2003; Wei, 2006; Kose et al., 2006), examines some apparent mismatch between theory and empirics on the economic consequences of international capital flows, and discusses ways to reconcile them.

The volume of international capital flows: paradoxes and possible solutions

The extent of cross-border capital movement can be measured by flows at a given point in time or by stocks accumulated over time. Capital inflows are net purchases of domestic assets by foreign residents, whereas capital outflows are net purchases of foreign assets by domestic residents. These data are well described in the International Monetary Fund's Balance of Payments statistics. For stock data, the IMF reports information for a few countries in recent years. Lane and Milesi-Ferreti (2001; 2005) expand the country and year coverage by combining this information with cumulative flows adjusted for valuation effects.

A country's exposure to international capital flows can be measured either by its government's policies (restrictions or incentives vis-à-vis capital flows) or by the actual amount of capital movement (scaled by the size of the recipient economy). The latter, the *de facto* measure, does not need to agree with the former, the *de jure* measure. For example, some countries may have many legal restrictions on capital movement (and hence a low exposure to capital flows by the *de jure* measure), but massive capital flight (and hence a high exposure by the *de facto* measure). A practical *de facto* measure of a country's exposure to cross-border capital movement is the sum of the country's total foreign assets and total foreign liabilities, divided by the country's GDP. For some economic questions, such as the effect of international capital flows on economic growth, the *de facto* measure may be more meaningful than the *de jure* measure.

Is the volume of capital flows observed in the data consistent with economic theory? Using a one-sector model, Lucas (1990) argues that it is a paradox that more capital does not flow from rich to poor countries. His reasoning goes as follows. Let $y = f(L, K)$ be a constant-returns-to-scale production function, where y is the output produced using labour L and capital K. Let p be the price of the good, and w and r be the returns to labour and capital, respectively. Firm's profit maximization problem gives

$$r = p\partial f(L, K) / \partial K = p\partial f(1, K / L) / \partial K \quad (1)$$

If the product price is equalized across countries under free trade, the law of diminishing marginal product implies that *r* is higher in the country with a lower capital–labour ratio. As an illustration, Lucas calculates that the return to capital in India should be 58 times as high as that in the United States based on their factor endowment. Facing a return differential of this magnitude, one should observe a lot more capital flowing from rich to poor countries. That too little is observed in the data has come to be known as the ‘Lucas paradox’.

Lucas (1990) discusses three possible explanations (within a one-sector framework): (*a*) a worker in a rich country could be several times more productive than her counterpart in a poor country; (*b*) human capital may be a missing factor and is likely much higher in a rich country; and (*c*) political risk and hence the required risk premium may be substantially higher in a poor country. Reinhart and Rogoff (2004) illustrate the last point for a set of countries with frequent default on their external debt.

Lucas's logic can be turned on its head in a multi-sector model. More precisely, in a standard Heckscher–Ohlin–Samuelson model with two goods, two factors, and two countries, firms earn zero profit. So one must have:

$$p_1 = c_1(w, r) \text{ and } p_2 = c_2(w, r) \quad (2)$$

where *c*(.) is the unit cost function and the numerical subscripts represent sectors. This implies that the factor prices are uniquely determined by product prices, and are independent of factor endowments. Since free trade in goods equalizes the product prices across countries, factor returns must also be equalized even in the absence of cross-border capital and labour movement. This was first pointed out by Samuelson (1948) and has become known as the ‘factor price equalization theorem’. Two countries with different capital–labour ratios would simply produce different mixes of outputs, but the marginal returns to physical capital are the same everywhere. In other words, zero capital flow is needed in equilibrium. This is true with or without cross-country differences in effective labour, human capital or political risk. The actual capital flow appears excessive on this logic.

One might think that the theorem of factor price equalization is too naive, requiring restrictive assumptions that surely do not hold in a more realistic setting with many countries, goods and factors. However, Ju and Wei (2006) show that, in a generalized neoclassical framework, relatively weak conditions are sufficient for factor prices to be equalized across countries (without factor movement). In particular, while the United States and India may not appear to satisfy the conditions for the factor prices to be equalized between them in a two-country model, it is nonetheless possible for factor prices to be equalized through a chain of country pairs (for example, the United States and Spain, Spain and Greece, Greece and Thailand, and Thailand and India). This means that it may be more difficult than it first

appears to escape from the logic of factor price equalization within a neoclassical framework, and that free trade in goods can completely substitute for capital mobility.

Obstfeld and Rogoff (2001) proposed that the existence of trade costs could explain the low but positive international capital flow. Trade costs do break factor price equalization even in a two-sector, two-factor model. However, as tariffs and transport costs decline over time, factor prices (including returns to capital) should converge across countries. This should lead to a decline in international capital flow (by the logic of factor price equalization), which is contradicted by the data.

Cross-country differences in total factor productivity (TFP) is another influential explanation of the Lucas paradox. While the discussion is usually couched in a one-sector model, it *could* work even in a multi-sector model. In particular, in a two-sector model, if the TFPs in both sectors are many times higher in the United States than in India, then return to capital in the United States could be only slightly lower than in India, justifying the observed small amount of capital flow. What drives the TFP differential across countries can be the quality of institutions, including the protection of property rights and the control of bureaucratic corruption. However, the TFP story can also go in the opposite direction in principle, exacerbating rather than resolving the Lucas paradox. In particular, if the United States has a greater TFP advantage in the labour-intensive sector than in the other sector, then this could further depress the return to capital from what already results from a high capital–labour ratio. This suggests that one has to be precise about the nature of the TFP differences in order to deliver predictions on the sign and the size of international capital flows.

Moving outside the neoclassical box, Ju and Wei (2006) introduce financial contracts and heterogeneous firms into an otherwise standard two-sector, two-factor framework. A key implication of the model is the separation between return to physical capital and return to financial investment. In particular, India could have a high return to physical capital due to its relatively low capital–labour ratio, but a low return to financial investment due to its relatively inefficient financial system. In addition, heterogeneous firms give rise to diminishing marginal returns at the sector level even though every firm has a constant returns technology. As a result, factor price equalization (before factor movement) does not hold in this model. In equilibrium, it is possible for financial capital to leave India for the United States, and for physical investment to flow in the reverse direction, resulting in a moderate amount of net flow. In this model, the return to capital (before capital flows) is still higher in India (with a lower capital-to-labour ratio) than the United States, but the differential in return is much smaller than in a one sector model. Thus, Ju and Wei's non-neoclassical two-sector, two-factor model partially restores the result of a typical one-sector model (that is, return to capital is determined in part by factor endowment) but does not generate the Lucas paradox.

Effects of international capital flows on economic growth

The gap between theories and empirics

International capital flows have the potential to bring a variety of benefits to recipient countries. In theory, financial globalization could raise a country's economic growth rate through a number of direct and indirect channels.

The direct channels include (*a*) augmenting domestic savings, (*b*) reducing the cost of capital through

better allocation of risks (Henry, 2000; and Stulz, 1999), (*c*) transferring technology and managerial know-how (Grossman and Helpman, 1991), and (*d*) stimulating development of the domestic financial sector (Levine, 1996 and 2005). The indirect channels include (a) promoting specialization (Brainard and Cooper, 1968; and Imbs and Wacziang, 2003), and (b) committing to better economic policies (Gourinchas and Jeanne, 2004; Tytell and Wei, 2004).

Yet a massive body of empirical papers has often found mixed results, suggesting that the benefits are not straightforward. Kose et al. (2006) survey 20 scholarly articles written between 1994 and 2005 that have empirically estimated the effect of exposure to international capital flows on economic growth. A majority of these papers (16 out of 20) find no, or at best mixed, effects. This echoes the conclusion in earlier survey articles by Eichengreen 2001 and Prasad et al. (2003) that it is not easy to find a strong and robust causal effect from financial globalization to economic growth, especially for developing countries.

Indeed, one alleged source of collateral damage of financial globalization is an increased propensity for developing countries to experience currency crises or other types of financial turmoil. For example, while the pace of cross-border capital flows picked up in the 1980s, there have also been more financial crises since around 1990, including the crises in Mexico in 1994, the Asian financial crisis during 1997–9, the Russian meltdown in 1999, and the Argentinean and Uruguayan crises of 2001–2. Most such crises tend to set countries back in their growth aspirations for a number of years.

Reconciling theories with empirical patterns

Financial crises do not prove that financial integration is a bad thing. Indeed, almost all developed countries are financially integrated, and very few developing countries, once embarked on a path of integration, would go back to financial isolation. So why do countries aspire to become financial integrated and yet experience so many bumps and potholes along the way? The literature has proposed *independently* two views: a composition hypothesis and a threshold hypothesis.

The composition hypothesis maintains that not all capital flows are equal. International direct investment, and perhaps international portfolio flows, appear to be robustly associated with a positive effect on economic growth (Borensztein, de Gregorio and Lee, 1998; Bekaert, Harvey and Lundblad, 2004). In contrast, there is no strong evidence that private foreign debt including international lending has robustly promoted economic growth. Indeed, one sometimes finds evidence that international lending is negatively associated with economic growth. Official aid flows do not robustly support growth either (Rajan and Subramanian, 2005).

Composition of capital flows has also been related to a country's propensity to experience a currency crisis. In their study of all episodes of currency crises in emerging markets during 1971–92, Frankel and Rose (1996) report that, while virtually no variable has a strong predictive power for subsequent currency crashes, the composition of capital inflows is one of the very few variables that are robustly related to the probability of a currency crisis. In particular, the share of foreign direct investment (FDI) in a country's total capital inflow is negatively associated with the probability of a currency crisis. This is confirmed in several subsequent studies including Frankel and Wei (2005). Other dimensions of composition are the maturity structure of external debt (the greater the share of short-term debt, the more

likely a crisis), and the currency denomination of external debt (the greater the share of foreign currency debt, the more likely a crisis) (Frankel and Rose, 1996; Radelet and Sachs, 1998; Rodrik and Velasco, 1999).

The threshold hypothesis states that certain minimum conditions have to be met before a country can be expected to benefit from financial globalization. Otherwise, the country could experience more crises and lower growth. The threshold effect comes in various versions. Only countries with reasonably good public institutions (for example, adequate control of corruption) and a minimum level of human capital seem to be able to translate exposure to financial globalization into stimulus to investment and growth on a sustained basis (see the surveys by Prasad et al., 2003; Kose et al., 2006). It is not difficult to imagine why countries with weak institutions may not benefit from financial globalization. In a highly corrupt country, for example, more capital inflows are likely to result in more consumption by a few elite families or in bigger Swiss bank accounts rather than more productive investment. So more capital flows may not result in higher growth rates. If capital inflows help to promote excessively risky projects backed by governments, then more inward capital flows could translate into an increased probability of a financial crisis.

Is the composition effect a consequence of the threshold effect?

Rather than viewing the threshold effect and the composition effect as two *rival* hypotheses, Wei (2000a; 2000b; 2001) suggests a concrete connection between the two: countries with better public institutions are likely to attract more international direct investment than international bank loans. Wei derives evidence from data on bilateral FDI reported by OECD source countries, and bilateral international lending reported by Bank for International Settlements (BIS) member countries. In the earlier work, Wei measures quality of public institutions by perception of corruption reported in surveys of firms such as those conducted by the World Economic Forum for its *Global Competitiveness Report* or by the World Bank for its *World Development Report*.

Recent evidence on investment by international mutual funds suggests that better institutions measured by a high degree of government and corporate transparency help to attract more international equity investment than that predicted by the international capital asset pricing model (ICAPM) (Gelos and Wei, 2005). So the composition effect and the threshold effect are perhaps just the two sides of the same coin.

Not everyone has found the same result. Hausmann and Fernandez-Arias (2000) report no relationship between share of FDI in total capital inflows and good institutions. In a panel of advanced and developing countries, Albuquerque (2003) finds the share of FDI in total inflows to be negatively related to good credit rating. It is important to note that Albuguerque's measure is about financial development rather than quality of public institutions generally, whereas Haumann and Fermandez-Arias mix measures of financial development and property rights institutions. As Ju and Wei (2006) point out, financial development and quality of public institutions have different effects, in theory, on the composition of capital flows. Furthermore, none of these studies employs instrumental variables to correct for possible measurement errors and endogeneity of the corruption or other institutional measures.

In any case, more recent papers with an instrumental variable approach and arguably better data again affirm the earlier conclusion that there may be an intimate relationship between the institutional

threshold effect and the composition effect. Using data from the IMF on balance of payments, Alfaro, Kalemli-Ozcan and Volosovych (2005) find that good institutional quality is a key determinant of total capital inflows. Papaioannou (2005) reports that foreign asset holdings by BIS banks, including their portfolio assets and direct investments, tend to be higher in destinations with better institutions.

Using recently available data from the IMF on member countries' international investment position (IMF, 2002), Faria and Mauro (2004) present evidence that countries with strong institutions are likely to attract more equity-like capital flows (FDI and portfolio equity flows) than other types of capital. Their measure of institutional quality is the average of six indicators – voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law, and control of corruption – as computed and reported by Kaufmann, Kraay and Mastruzzi (2003). An important feature of the study is that the authors address explicitly the possibility that the composite institutional index may be measured with errors and/or be endogenous. They employ as instrumental variables log settler mortality during the early colonial period as proposed by Acemoglu, Johnson and Robinson (2001) and ethno-linguistic fragmentation first used by Mauro (1995). The instrumental variable approach reaffirms their basic conclusion.

Wei (2006) furnishes evidence that the effects of the quality of public institutions and the level of financial development can indeed be different. In particular, weak public institutions strongly discourage FDI, and possibly foreign debt, as shares of a country's total foreign liabilities, but appear to encourage borrowing from foreign banks. In comparison, low financial-sector development discourages inward portfolio equity flows but encourages inward FDI. The finding that poor financial development could encourage FDI may sound surprising. A possible story is set out in Ju and Wei (2006). Essentially, in countries with poor financial systems but also low capital–labour ratios, the return to financial capital is low. Hence domestic households would want to take savings out of the country, and international portfolio investors do not wish to come in. As the same time, as long as the risk of expropriation is not too high, the depression of domestic investment due to poor domestic financial development could raise the return to FDI.

To gain confidence that these patterns reflect causal relations, Wei (2006) employs instrumental variables for the institutional measures based on the economic histories of the countries in the sample, in particular the log mortality rate of the European settlers in former colonies à la Acemoglu, Johnson and Robinson (2001), and the origin of legal systems à la La Porta et al. (1998). The instrumental variable approach bolsters the argument that weak institutions are a *cause* of the unfavourable composition of capital inflows.

To summarize, the cumulative evidence points to the strong possibility that weak public institutions tilt the composition of capital flows into a country away from FDI and portfolio equity flows and towards debt, including bank loans, making the country more vulnerable to a currency crisis and less able to translate a given amount of capital inflow into stimulus for economic growth. While the composition and the threshold effects are not identical, they are very likely related.

For an institutionally challenged country, more research is needed to determine whether it should wait for its institutions to be sufficiently improved before opening up to global capital flows, or use exposure to the international capital market as a disciplinary device to improve its institutions.

See Also

- financial liberalization
- foreign direct investment
- globalization

The author is grateful to Andre Faria, Jiandong Ju, Ayhan Kose, Paolo Mauro, Gian Maria Milesi-Ferretti, Romain Ranciere, Kenneth Rogoff, and Irina Tytell for helpful discussion, and Yuanyuan Chen and Patricia Medina for capable research and editorial assistance. The views expressed in this article are the author's own, and do not reflect those of the IMF or any other organization he is associated with.

Bibliography

Acemoglu, D., Johnson, S. and Robinson, J. 2001. The colonial origins of comparative development: an empirical investigation. *American Economic Review* 91, 1369–401.

Albuquerque, R. 2003. The composition of international capital flows: risk sharing through foreign direct investment. *Journal of International Economics* 1, 353–83.

Alfaro, L., Kalemli-Ozcan, S. and Volosovych, V. 2005. Why doesn't capital flow from rich to poor countries? An empirical investigation. Working Paper No. 11901. Cambridge, MA: NBER.

Bekaert, G., Harvey, C. and Lundblad, C. 2005. Does financial liberalization spur growth? *Journal of Financial Economics* 77, 3–56.

Borensztein, E., de Gregorio, J. and Lee, J.-W. 1998. How does foreign direct investment affect economic growth? *Journal of International Economics* 45, 115–35.

Brainard, W. and Cooper, R. 1968. Uncertainty and Diversification of International Trade. *Food Research Institute Studies in Agricultural Economics, Trade, and Development* 8, 257–85.

Caballero, R., Farhi, E. and Gourinchas, P.-O. 2005. An equilibrium model of global imbalances and low interest rate. Working paper, MIT and UC Berkeley.

Eichengreen, B. 2001. Capital account liberalization: what do cross-country studies tell us? *World Bank Economic Review* 15, 341–65.

Faria, A. and Mauro, P. 2004. Institutions and the external capital structure of countries. Working Paper No. 04/236. Washington, DC: IMF.

Frankel, J. and Rose, A. 1996. Currency crashes in emerging markets: an empirical treatment. *Journal of*

International Economics 41, 351–66.

Frankel, J. and Wei, S.-J. 2005. Managing macroeconomic crises: policy lessons. In *Managing Economic Volatility and Crises: A Practitioner's Guide*, ed. J. Aizenman and B. Pinto. New York: Cambridge University Press.

Gelos, G. and Wei, S.-J. 2005. Transparency and positions of international mutual funds. *Journal of Finance* 60, 2987–3020.

Gourinchas, P.-O. and Jeanne, O. 2004. On the benefits of capital account liberalization for emerging economies. Working paper, University of California, Berkeley, and the International Monetary Fund.

Grossman, G. and Helpman, E. 1991. Trade, knowledge spillovers, and growth. *European Economic Review* 35, 517–26.

Hausmann, R. and Fernandez-Arias, E. 2000. Foreign direct investment: good cholesterol? Working Paper No. 417. Research Department, Inter-America Development Bank.

Henry, P. 2000. Stock market liberalization, economic reform, and emerging market equity prices. *Journal of Finance* 55, 529–64.

Imbs, J. and Wacziang, R. 2003. Stages of diversification. *American Economic Review* 93, 63–86.

IMF (International Monetary Fund). 2002. *International Investment Position: A Guide to Data Sources*. Online. Available at www.imf.org/external/np/sta/iip/guide/index.htm, accessed 11 May 2006.

Ju, J. and Wei, S.-J. 2006. A solution to two paradoxes of international capital flows. NBER working paper, forthcoming. Cambridge: National Bureau of Economic Research.

Kaufmann, D., Kraay, A. and Mastruzzi, M. 2003. Governance matters III: governance indicators for 1996–2002. Policy Research Working Paper No. 3106. Washington DC: The World Bank.

Kose, M., Prasad, E., Rogoff, K. and Wei, S.-J. 2006. Financial globalization: a reappraisal and synthesis. Prepared for the *Journal of Economic Literature*.

Lane, P. and Milesi-Ferreti, G. 2001. The external wealth of nations: measures of foreign assets and liabilities for industrial and developing nations. *Journal of International Economics* 55, 263–94.

Lane, P. and Milesi-Ferreti, G. 2005. The external wealth of nations mark II: revised and extended estimates of foreign assets and liabilities, 1970–2004. Working paper No. 06/69. Washington, DC: IMF.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R. 1998. Law and finance. *Journal of Political Economy* 106, 1113–55.

Levine, R. 1996. Foreign banks, financial development, and economic growth. In *International Financial Markets: Harmonization* Versus *Competition*, ed. C. Barfield. Washington, DC: AEI Press.

Levine, R. 2005. Finance and growth: theory and evidence. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Lucas, R. 1990. Why doesn't capital flow from rich to poor countries? *American Economic Review* 80, 92–6.

Mauro, P. 1995. Corruption and growth. *Quarterly Journal of Economics* 110, 681–712.

Obstfeld, M. and Rogoff, K. 2001. The six major puzzles in international macroeconomics: is there a common cause? *NBER Macroeconomics Annual* 15, 339–90.

Papaioannou, E. 2005. What drives international bank flows? Politics, institutions and other determinants. Working Paper No. 437. Frankfurt: European Central Bank.

Prasad, E., Rogoff, K., Wei, S.-J. and Kose, M. 2003. *Effects of Financial Globalization on Developing Countries: Some Empirical Evidence*. Occasional Paper No. 220. Washington, DC: IMF.

Radelet, S. and Sachs, J. 1998. The east Asian financial crisis: diagnosis, remedies, and prospects. *Brookings Papers on Economic Activities* 1998(1), 1–74.

Rajan, R. and Subramanian, A. 2005. What undermines aid's impact on growth? Working Paper No. 05/126. Washington, DC: IMF.

Reinhart, C. and Rogoff, K. 2004. Serial Default and the ‘paradox’ of rich to poor capital flows. *American Economic Review* 94(2), 52–8.

Rodrik, D. and Velasco, A. 1999. Short-term capital flows. Working Paper No. 7364. Cambridge, MA: NBER.

Samuelson, P. 1948. International trade and equalization of factor prices. *Economic Journal* 58, 163–84.

Smarzynska, B. and Wei, S.-J. 2000. Corruption and the composition of foreign direct investment: firm-level evidence. Working Paper No. 7969. Cambridge, MA: NBER.

Stulz, R. 1999. Globalization of equity markets and the cost of capital. Working Paper No. 7021.

Cambridge, MA: NBER.

Tytell, I. and Wei, S.-J. 2004. Does financial globalization induce better macroeconomic policies? Working Paper No. 04/84. Washington, DC: IMF.

Wei, S.-J. 2000a. How taxing is corruption on international investors? *Review of Economics and Statistics* 82, 1–11.

Wei, S.-J. 2000b. Local corruption and global capital flows. *Brookings Papers on Economic Activities* 2000(2), 303–54.

Wei, S.-J. 2001. Domestic crony capitalism and international fickle capital: is there a connection? *International Finance* 4, 15–45.

Wei, S.-J. 2006. Connecting two views on financial globalization: can we make further progress? Working paper.

How to cite this article

Wei, Shang-Jin. "international capital flows." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000163> doi:10.1057/9780230226203.0826

The New Palgrave Dictionary of Economics Online

international economics, history of

M. June Flanders

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Starting with mercantilist theories, the article deals with laissez-faire rejections of mercantilism, the Ricardian justification of free trade and its extension to multiple countries and commodities. Heckscher–Ohlin trade theory, factor-price equalization and the ‘Leontieff paradox’ debate follow. Intra-industry trade is related to increasing returns, imperfect competition, and product differentiation. Trade and growth, economic geography, and tariffs and trade restrictions are summarized. Regarding macroeconomics, Hume and the monetary approach to the balance of payments are compared with income adjustment theories. International monetary regimes, exchange rate regimes, capital transfers, internal–external balance, and new international macroeconomics are discussed.

Keywords

absorption approach to the balance of payments; balance of trade; balance of payments; Bank of England; beggar-my-neighbour; brand identification; Cantillon, R.; classical economics; comparative advantage; cone of diversification; convertibility; Corn Laws; devaluation; diffusion of technology; dirigisme; economic geography; elasticities approach to the balance of payments; Eurozone; exchange rate regimes; factor price equalization; fixed exchange rates; Fleming, J. M.; flexible exchange rates; foreign direct investment; full employment; gains from trade; Gervaise, I.; gold standard; growth and international trade; Haberler, G.; Hayek, F.; Heckscher–Ohlin trade theory; Heckscher–Ohlin–Samuelson model; Hume, D.; imperfect competition; increasing returns; infant-industry protection; innovation; intermediate products; international capital flows; international economics, history of; international monetary theory; international trade theory; internal–external balance; intra-industry trade; labour theory of value; Leontief paradox; Lerner, A. P.; Locke, J.; long-term economic growth; Meade, J. E.; mercantilism; microfoundations; monetary approach to the balance of payments; money multiplier; money supply; multinational corporations; Mundell, R.; new classical macroeconomics; new international macroeconomics; offer curve or reciprocal demand curve; Ohlin, B.; outsourcing; product

differentiation; production possibility frontier; quantity theory of money; quotas and tariffs; Ricardo, D.; Smith, A.; specialization; specie-flow mechanism; sterilization; sticky wages; subsistence; Taussig, F.; technological progress; terms of trade; Tobin tax; Torrens, R.; trade policy, political economy of; transfer problem; vent for surplus; Walrasian general equilibrium

Article

The real theory of international trade

Contemporary international trade theory has its roots in classical economics, which developed in opposition to the widely accepted views, known (since Adam Smith's introduction of the term into British discourse) as mercantilism, held until the mid-18th century by both policymakers and many analysts. This was a loose body of doctrine advocating extensive government control and interference in economic activity. In the context of international trade it refers to the imposition of tariffs, quotas, and prohibitions, designed to maximize the balance of payments surplus and the net inflow of precious metals (specie). The justification for such policies took many forms. Since what we know as mercantilism lasted for about two centuries and arose in widely differing social and political contexts, there was room for substantial differences of opinion. In some cases it was apparently a fear of goods, in others an identification of specie with real wealth, and in still others, as Keynes (1937) suggested, a desire to stimulate employment. (Detailed critical surveys and analyses are to be found in Heckscher, 1935; Viner, 1937; Blaug, 1985.)

Though he was not the first to oppose dirigisme and to see the advantages of trade and specialization, Smith, in his *Wealth of Nations* (1776), provided the starting point for classical theories of trade. He argued both that gold and silver are not real wealth and that generating a balance of trade surplus is not the only way to acquire them. His discussion of interference with trade in goods and services is the same as his treatment of other governmental interferences: he dealt with the interactions of markets for goods and factors, showing the effects of tariffs and subsidies on each. He recognized situations in which restrictions on trade might be justified: these included defence needs, retaliation and infant-industry arguments.

Smith did not develop a theory of comparative advantage, though he came close to it when he noted that Britain was more productive in manufactures relative to Poland than it was in agriculture (Smith 1776, pp. 6–7). But his main argument in favour of trade is indeed that which is now labelled the ‘vent for surplus’. This is enunciated at several points, but a clear statement occurs when he says that, because of international trade, ‘the narrowness of the home market does not hinder the division of labour in any particular branch of art or manufacture from being carried to the highest perfection. ‘By opening a more extensive market for whatever part of the produce of their labour may exceed the home consumption’ (1776, p. 415). Mill (1848) cites Smith as having viewed exports as an outlet for surplus, and Bastable (1897) takes Smith very literally on the surplus and attacks him vigorously for it, arguing that it implies the existence of unemployed resources. Schumpeter (1954, p. 374) accuses Smith of having ‘believed that under free trade all goods would be produced where their absolute costs in terms of labor are lowest’. But Schumpeter notes that Viner indicates that Smith, and others before him, had

formulated the more general proposition that, under free trade, commodities would be imported whenever they could be obtained most cheaply in this way. This includes the case where exports ‘cost less to produce than it would cost to produce the corresponding imports at home and thus implies the theorem of comparative costs’. A contemporary evaluation of the several interpretations of Smith's trade theory is provided by Blecker (1997).

It was Ricardo, who, in his *Principles of Political Economy and Taxation* (1817), first articulated explicitly and emphasized and publicized the theory of comparative advantage (though Torrens, in 1815, had come very close to it), noting that absolutely lower costs in the production of all goods was not a sufficient reason for producing them all, and that it was generally to the advantage of a country to specialize in the production of that which it did best. Here, too, the objective was to influence policy. Ricardo, like Smith, developed his views on trade within the framework of his model of how an economy did or should operate. British manufacturers had no need of protection from imports, and his main concern was with duties on grain imports, the Corn Laws, which protected British agriculture. Ricardo attacked these because they raised food prices and hence real wages at the expense of profits, thus discouraging investment and growth. The equilibrium result would be zero saving and investment and a long-run stationary economy with wages at subsistence level. With free trade, national income would be maximized by keeping real wages in industry low and by concentrating output in the relatively low-cost sectors, regardless of absolute advantage or disadvantage. It was here that he inserted a brief monetary statement, that there was a natural distribution of specie, which depended on the economic size of each country and the parameters of its monetary variables.

Ricardo's trade theory, like classical economics in general, is based on a model of supply: price is the long-run supply price. He implies that complete specialization is the norm, but he neither makes this explicit nor specifies how the gains from trade are divided between the trading countries. He considered the possibility of non-traded goods and the idea that a country could produce and export more than one good if country size or demand patterns were highly uneven.

Mill (1848, pp. 583–606) brought demand and, implicitly, its elasticity into the theory, thereby explaining how the terms of trade were established between the limits set by comparative advantage. He extended the analysis to include transport costs, more commodities (which limited the range of the terms of trade) and more countries. He also noted that, since the terms of trade must be the same for all countries, the gains from trade will be greater for those for which the opportunity cost of the exportable good is lower. These refinements were spelled out precisely in the 20th century by Graham (1923), who noted that the terms of trade and the probability of specialization depended on the relative size of the (two) trading countries and the relative importance of the (two) traded goods in total consumption. Furthermore, when more countries and/or more goods were brought into the picture, the final terms of trade were narrowed down even further, as was the exact number of goods traded by any given country.

Marshall (1879; 1923) generalized the theory into a two-country multi-commodity analysis (using the device of ‘bales’ of goods) and derived offer curves to depict graphically the general equilibrium in production and consumption that Mill had analysed only verbally. Edgeworth (1925) produced a similar analysis. The derivation of the offer (reciprocal demand) curves was not spelled out; all domestic markets were assumed to be equilibrium at each point on an offer curve. Marshall (1879) has a detailed analysis of equilibrium and stability conditions, but not, explicitly, elasticity.

A sizeable literature developed, well into the 20th century, both testing Ricardian comparative cost

theories and justifying economists' generally free-trade position. In a fairly laboured list of possibilities, Samuelson (1939) showed that some trade is always better than no trade. This holds for various shapes of production possibilities curves, but its welfare implications depend on the welfare device of the ability of gainers to compensate the losers, or, as he put it, 'by Utopian co-operation everyone can be made better off as a result of trade' (1939, p. 204).

In the 1930s and 1940s, there was extensive discussion of arguments for and against tariffs, quotas, and other impediments, and the real gain – or loss – from imposing them. This included the debate on the relationship of trade structure to growth, which continues to the present.

A major paradigm shift came with the formal use of what are often called 'neoclassical' assumptions about technology and preferences in the analysis of comparative costs. In the entry in this dictionary on Haberler, Gottfried, it is claimed that, although Barone in 1908 (but not subsequently) had a (non-concave) production-possibility frontier and a community indifference curve, it was Haberler's independent discovery in 1930, and the use to which he put it, that transformed the theory of international trade. Haberler thereby broke with the labour theory of value, the production possibility frontier becoming standard in all economic theorizing and teaching. Lerner went on to draw a 'compound indifference curve' in 1932, and in 1934 developed the demand side fully. Both Lerner and Haberler (1936) dealt with the possibility of increasing returns, Haberler granting that this could justify tariffs.

In fact, Haberler was preceded by Bickerdike, and by Heckscher in 1918 and Ohlin in 1924, but the latter two published in Swedish, reducing the visibility of their work. They addressed the issue of factor prices under free trade. Ohlin published an enlarged version of his 1924 Ph.D. thesis in English in 1933; Heckscher's seminal paper was translated (partially) into English only in 1949.

Heckscher's general equilibrium analysis, in 1918, was wholly verbal. He examined the reasons for, and results of, large-scale Swedish migration to America. In studying this, he showed that under certain circumstances trade in goods could substitute for movement of factors in equalizing factor prices. If factor endowments were not too different, factor prices would inevitably be equal throughout the world. Ohlin, his student, drew from this and, using Cassell's version of Walrasian general equilibrium analysis, developed a more formal approach (1924). Lerner demonstrated factor price equalization with arithmetic and geometry in a seminar paper in 1933, the Swedish work being unknown.

Ohlin himself rejected the conclusion of the equalization of factor prices. In his formal analysis he simply (inexplicably) assumed that each country was completely specialized; in verbal discussion, he always saw exceptions to any generalization. Among the obstacles to factor price equalization were insufficient geographical and occupational mobility of factors between industries within a country, increasing returns to scale, excessive imbalance in factor supplies, taxes, transport costs, and imperfect competition. All these had to be taken into account when explaining the actual patterns of trade.

Furthermore, the verbal analysis is explicitly dynamic. One example is his discussion of the effects, almost year by year, of an increase in world demand for a country's exports. There is an increased demand for labour in that industry, so labour moves into it; individuals move, gradually, into the skill group required for those goods; labour moves, again gradually, geographically. All these take time: some years later, if demand remains as it was with no further increases, the supply of productive factors will have adjusted to the new demand levels. Ohlin's earlier work (his thesis) is replete with numerical examples of international and interregional differences in price levels, in expenditures for food and

housing, and the like.

The factor price equalization aspect of the model was largely ignored until Stolper and Samuelson unearthed it in 1941. (When White reviewed Ohlin's book in 1934 and compared his model with the first German edition of Haberler's *International Trade*, he did not mention the factor price equalization issue at all.) The model was formalized subsequently by Samuelson (1948; 1949; 1951–2; 1953–4), and later elaborated, extended, and modified by Jones (2000, passim) and others. Originally these formalizations required very strict assumptions: two commodities, two countries, constant returns to scale, both goods being produced in both countries. The last requirement of the Heckscher–Ohlin–Samuelson model, as it became known, was very stringent. It meant that factor proportions had to be within what was called the cone of diversification, where both goods could be produced when prices were equalized across countries. Later models picked up some of the complications that Ohlin had enumerated earlier: Jones developed variations, and welfare implications, for the cases of specific factors (immobile between industries), increased numbers of goods and factors, trade in intermediate goods, and tariffs. A penetrating discussion of fundamental methodological issues in the development of the theory and the subsequent debate on the Leontief paradox is provided by De Marchi (1976).

The Heckscher–Ohlin paradigm remained untested and empirical work in trade proceeded along Ricardian lines until Leontief applied his input–output model to a test of United States trade and concluded that US exports were labour-intensive relative to imports. The implied paradox gave rise to an avalanche of theoretical and empirical work, ranging from multifactor models to specific explanations in terms of the period and the structure of world trade in the early post-Second World War period which was used in the study. One explanation offered was the possibility of reversals of relative factor intensities as relative factor prices changed, a proposition which led to a spate of ongoing debates in capital theory (Arrow et al., 1961.)

Baggott (1970) presents an exhaustive list and discussion of the various proffered explanations, including her own, that the United States in the relevant period was exporting capital directly through its balance of trade surplus, and also notes the possibility that some of the capital-intensive commodity imports were produced by branches of American firms, with American capital. Caves and Jones (1977) give some of the highlights in the debate. Others sought to resolve the paradox in a wide variety of ways: for example, by claiming that there had been a conceptual misunderstanding, that labour needed to be augmented by considering human capital, and that during Leontief's sample period capital intensity was highly correlated with high natural-resource use. For an exhaustive treatment of the paradox and of the numerous attempts to resolve it, see Chipman (1965–66, 33 51–70).

Economic analysis and investigation move with events. Both output and trade became increasingly characterized by differentiated products, brand identifications, oligopolistic behaviour and the scale economies which these generated, strategic investment and marketing decisions, and expanding trade in intermediate products. Historically, the first phenomenon observed in the post-Second World War period was the large and growing intra-industry trade (see Grubel and Lloyd, 1975). Helpman and Krugman (1985) treat such trade as a result of product differentiation, generally associated with monopolistic competition and increasing returns, and model it as coexisting with inter-industry trade based on factor endowments. A survey of trade theory based on imperfect competition due primarily to increasing returns can be found in Krugman (1987). The behaviour of multinational corporations and conglomerates, often producing and marketing a wide variety of products, demanded examination and

explanation (see Caves, 1982). Dixit and Norman (1980) include in their formal, general-equilibrium rewriting of trade theory the case of oligopolistic markets. Game-theoretic study of innovation strategies, outsourcing, and the political economy of trade restrictions policy followed. Refinements of these trends continue and are at the forefront of international trade theoretical and empirical work today (such as that of Grossman and Helpman, 1991; 2002).

There has also developed a large literature on the mutual interaction of growth and international trade, which is a subject in and of itself, starting with Smith's vent for surplus theory, through issues of putative exploitation of developing countries, into contemporary empirical/historical studies of long-term growth, and the effects of differential increases in the several factors of production. The differential rates of technical progress, the diffusion (or lack of it) of technology, and the concomitant international differences in growth rates are some of the topics being analysed and explored. The existence of increasing returns, both to firms and to industries, first discussed by Haberler, has led to renewed interest in economic geography, a field being examined today both by economists and geographers and, perhaps, waiting for some real interdisciplinary exploration. A critical review of the field, and a plea for integration of new explorations in economic geography with industrial geography is offered by Martin and Sunley (1996).

Macro-monetary theory

Many of the mercantilists, with their concern for the accumulation of specie, evinced no sense of the impossibility of having a permanent balance of trade surplus. Some saw money as working capital that would drive an increasing volume of trade; others saw little problem in absorbing specie, perceiving trade as constrained by a shortage of coin. In the 16th and 17th centuries, there was increasing awareness of a link between money and the price level, culminating in Locke's formulation of the quantity theory (1696). Despite a partial anticipation of the result by Cantillon (1755; written c. 1730), it is Hume (1752) who is commonly credited with the price-specie-flow mechanism and the implied endogeneity of the money supply in an open economy. However, both were preceded by Isaac Gervaise, whose pamphlet of 1720 was almost totally ignored. Gervaise had an adjustment mechanism, essentially the monetary approach to the balance of payments (Gervaise, 1720), which Ricardo 100 years later was to enunciate as the ‘natural distribution of specie’. Beyond that, he had a model of financial adjustment through a money multiplier and real effects in the form of inter-industry shifts in production.

Hume argued that attempts to acquire precious metals, or prevent their export, would result in price level changes, affecting the balance of trade and reversing the specie flows. The same line of reasoning (though under flexible exchange rates) informed the work of Ricardo (1811) and Henry Thornton (1802). Other highly sophisticated monetary writings of the time are discussed in Hollander (1910–11). The main point, that the money supply is endogenous, resurfaced almost 150 years later in the monetary approach to the balance of payments (see Frenkel and Johnson, 1976).

In subsequent British writings, through much of the 19th century, the distinction between international and domestic monetary theory barely existed. The British economy was open; the London money market was the world's financial centre. The brilliant stream of monetary debate in 19th-century England (see Fetter, 1965; Bagehot, 1873), carried out largely by bankers and business people, concentrated on issues of monetary policy for an open economy, but with little attention to the effect of policy on the real sector

or on long-term financial markets. The analysis focuses in general on the short run: international disturbances were both exogenous and temporary in that world, and the issue was really how to ride the storm – the underlying structure was always, implicitly, in equilibrium. This was a reasonably accurate picture of the England of the day, where long-term capital outflows were effected by changes in trade flows in the equilibrating direction, and much of any needed real adjustment was performed in the periphery (see Ford, 1962). This literature had to do primarily with the reciprocal relationship between specie holdings and specie flows on the one hand and the domestic money supply on the other. Much of this involved differences in the definition of money, the role of the Bank of England, the vulnerability of the bank's gold stock, by law subject to drain by holders of Bank of England notes (paper money), and the appropriate measures which this very public, privately owned, institution could and should take to protect itself. (Evidence of the time, and subsequent histories, point to the uniqueness of the Bank of England; yet in the paradigm of the adjustment mechanism it is always treated as the prototype of central banks.)

Convertibility of Bank of England notes into gold was suspended during the Napoleonic Wars; when it was resumed, in 1819, the bank entered a period of more than half a century of recurrent crises and near crises, but managed to maintain the convertibility of its notes into gold. Discussion and debate in this environment was almost brought to an end when Walter Bagehot published his influential *Lombard Street* in 1873. Based on experience, and possibly as a result of his hectoring, the Bank of England learned in the subsequent decades how to handle its huge constituency of world and domestic finance on the basis of very small reserves of gold, developing, gradually, a number of highly sophisticated ‘tricks’ in the money markets to protect itself and forestall runs. Nothing about this management, meticulously documented by Sayers in his two separate studies, had to do with interactions between the real and financial sectors, certainly not in the context of long-run relationships. There was little if any discussion of the balance of trade, except when the original disturbance was a trade imbalance (usually temporary), such as a crop failure. And nothing remotely suggested automaticity of adjustment; the effects on the domestic money supply were to be avoided or at least mitigated. This relative neglect continued in discussion of British monetary policy and international adjustment until 1925, when it appeared in the debate on the resumption of gold convertibility. The previous examination, by the Cunliffe Committee in 1918, was noteworthy for its brevity. There was really nothing to discuss: specie outflow led to the Bank of England's changing bank rate, which reversed the outflow. Changes in expenditures play a secondary role in the process of adjustment to gold flows (see Flanders, 1989).

Despite these objections, the automatic price-specie flow mechanism, stemming from Hume, lived on, until it was challenged in the 20th century by several analyses of what became known as the transfer problem. The origin of this is traced to Thornton, Ricardo, and others in the early 19th century who distinguished between money transfers and deficits caused by harvest failures. Bastable (1889) emphasized the impact of a monetary transfer, such as a tribute, on demand and the possibility of effecting it with no change in the terms of trade. But this insight faded and reappeared only in the 20th century, when Taussig and his Harvard students, including Jacob Viner, John Williams and Harry Dexter White, examined the adjustments to the huge capital flows of the 19th and early 20th centuries. They found that income and expenditure changes played a much more critical role than had been thought. Adjustment was too smooth and too fast for it to have worked through Humean kinds of changes in price levels and thence in trade balances.

It was at this time, during the academic year 1922–23, that Ohlin, visiting Harvard, developed the approach first expressed in his 1929 debate with Keynes over German reparations, and then in great detail in his book (1933). He spelled out explicitly an expenditure-driven adjustment to international capital movements. (The irony is that Keynes could not see this.) Capital flows would lead to changes in total spending and hence directly in net exports. Price and wage changes might ensue but were not essential to the adjustment process.

The price-specie-flow story was challenged later, from a different flank, by Brown (1940), who demonstrated that the canonical view of the historical gold standard was mistaken and that in fact it had been a sterling standard, managed by the Bank of England. The system worked reasonably smoothly because trade and long-term capital movements were consistent with long-run equilibrium in the balance of payments. When this ceased to be true, England went off gold. Following the chaos of the inter-war period and the controls of the war and post-war years, the establishment of the International Monetary Fund constituted a recognition that the textbook adjustment mechanism of a metallic standard could not be relied upon. But what emerged in fact was a dollar standard rather than the intended multilateral system.

The textbooks continued to describe the international monetary system in terms of the price-specie flow mechanism and to treat capital movements as either factor flows (foreign investment) or short-term financial adjustments. In 1937, Hayek had taken the position that there had never been a true test of the price-specie flow mechanism in a multilateral world system in which domestic money supplies were endogenous and adjustments were automatic. In a neglected lecture at the London School of Economics he argued that the fixed exchange rate system, or gold standard, should not be abandoned on grounds of its failure, since it had never been operated correctly. There had never been a time, he said, when domestic money supplies were made to vary in response to specie flows as they would have had there been no sterilization or offsetting policies, that is, had the specie-flow mechanism functioned in the manner of the traditional paradigm.

Hayek's complaint, in the mid-1930s, was made in response to growing sentiment in favour of fluctuating or, more accurately, administratively pegged exchange rates. Given the willingness to consider changing the peg, the rate became a policy tool and a literature developed around what was called ‘internal–external policy’. Beginning with Joan Robinson's attack on ‘beggar-my-neighbour’ devaluations (1937), there ensued a discussion of the effects of exchange rate changes on income and expenditure, as well as on the balance of payments; starting with the elasticities approach (elasticities of demand for and supply of exports, which proved to depend on general equilibrium in the domestic goods markets), moving on to the absorption approach to the balance of payments (introducing monetary effects as devaluation altered price levels) and culminating in Meade's massive multi-equation model of an open economy (1951). Meade rang the changes of the effects of various domestic monetary and fiscal policies directed at the level of employment and balance of payments equilibrium, under different conditions of price flexibility, capital mobility, wages policy, and various types of initial disturbance in regimes of (*a*) pegged and (*b*) flexible exchange rates (for a more detailed account, see Flanders, 1989). Not dissimilar in aim and scope is a neglected attempt by Stuvel (1950) to formalize the effects of exchange rate changes; both he and Meade, by the way, confine themselves to comparative statics. Discussions of optimal exchange rate regimes were based on the assumption that the large pre-war capital movements, characterized by political and economic speculative flights, were expected not to

continue or, in any case, not be permitted. (As early as 1936 Williams outlined possible forms of international monetary and exchange arrangements.) In this light the recommendations for exchange rate flexibility of Friedman (1953), Meade (1955) and others were simple arguments in favour of permitting goods markets to clear by price variation.

The issue of the effects of internal financial policy on the foreign balance and hence on its success in achieving its domestic goals was explored with much simpler models, by Fleming (1962) and by Mundell (1960). These were initially designed to address the problem that domestic full employment policies, if successful, would worsen the trade balance. Some of their two-country models necessarily dealt with the impact on foreign countries as well. (Williams had raised this issue as early as 1934.) They produced models that dealt with the possibility of balance in external payments and attainment of a targeted level of domestic expenditure provided international financial capital flows were sufficiently elastic with respect to interest rate differentials. Metzler (1960) dealt with similar issues but concentrated primarily, in the spirit of Wicksell and Keynes, on the implications for domestic money markets and interest rates as the channel for influence on real absorption. His is a full employment model, so there is no government stabilization activity.

The question of whether monetary and fiscal authorities can maintain desired levels of inflation, real output and the real exchange rate continues to exercise the profession to the present day. Now, given the trend back into administered, if not fixed, exchange rates, the enormous stocks and flows of international financial assets, and the large current account imbalances that these permit, mirrored by the gaps between domestic savings and investment, the issue of the ‘adjustment mechanism’ takes the form of questions as to the sustainability of these imbalances and the consequences of diminishing or eliminating them. There is general agreement that the imbalances prevailing currently are not sustainable; the manner and consequence of their elimination is less obvious. See Clarida (2006) for an excellent summary of a National Bureau of Economic Research conference.

The overwhelming size and volatility of international financial flows (which, not by chance, coincided with the abandonment of the worldwide fixed exchange-rate system in 1973) have informed the reactivated discussions of optimum exchange rate regimes. Should rates be pegged (temporarily or in perpetuity, as in a model of dollarization) or allowed to float, freely or with some intervention? And if pegged, then to what? This leads to discussion of currency baskets, pegging to a weighted average of currencies; the question then is, what determines the weights (Flanders and Helpman, 1979)? If the baskets are weighted by trade shares, international capital flows can prove highly disruptive. We are led, in turn, to the issue of whether capital movements can be controlled, and, if so, should they be, and which countries should be encouraged or permitted to attempt such controls. One line of discussion on this subject revolves around the proposed ‘Tobin Tax’ (Tobin, 1978), designed to put a little ‘sand in the wheels’ of international monetary flows, which have become huge relative to commodity flows, and which can be highly erratic in response to short-run volatile shifts in expectations.

In Mundell's work the internal-external balance issue led naturally into a discussion of the requirements, in terms of both labour and capital mobility, for a group of countries to constitute an optimum currency area. (Abba Lerner had hinted at something like that in 1944.) At the time, the question was a theoretical curiosum. Twenty years later it led to substantive questions about the viability of the Eurozone as an optimum currency area, including analogies to studies of the United States as such (see Rockoff, 2003), and whether a single currency and central bank can be sustained without a single fiscal authority which

effects intra-area transfers.

At the same time, the recent neo-monetarism or new neoclassical trend in macroeconomics has been paralleled by a ‘new international macroeconomics’. An exhaustive and perceptive survey is provided in Lane (2001). Starting from the work of Obstfeld and Rogoff (1995), there have been numerous explorations of the impact of monetary shocks (and some consideration of technological shocks) on trade, prices, welfare, real exchange rates, real terms of trade in models of two countries, many countries, and a single small country. Some have sticky wages, all have administered prices, of various types. Different assumptions are made as to consumption elasticities, technology, non-traded goods, bias toward home goods, the inclusion of capital, financial structure and completeness of financial markets, *inter alia*. Some attempts at calibration of the models have been made, with varying success. While the spelling out of the microfoundations of international macroeconomics is intellectually satisfying, the results, as Lane, himself a contributor, avers, are ‘highly sensitive to the precise denomination of price stickiness, the specification of preferences and financial market structure. For this reason, any policy recommendations emanating from this literature must be highly qualified’ (Lane, 2001, p. 262).

Bibliography

Arrow, K.J., Chenery, H.B., Minhas, B.S. and Solow, R.M. 1961. Capital–labor substitution and economic efficiency. *Review of Economics and Statistics* 43, 225–50.

Bagehot, W. 1873. *Lombard Street: A Description of the Money Market*. London: Henry S. King and Company.

Baggott, N. 1970. The modern theory of international trade and the Leontief Paradox. Ph.D. dissertation, Purdue University.

Bastable, C.F. 1889. On some applications of the theory of international trade. *Quarterly Journal of Economics* 4, 1–17.

Bastable, C.F. 1897. *The Theory of International Trade, with Some of its Applications to Economic Policy*, 4th edn. London: Macmillan, 1903.

Blaug, M. 1985. *Economic Theory in Retrospect*, 4th edn. Cambridge University Press.

Blecker, R.A. 1997. The ‘unnatural and retrograde order’: Adam Smith's theories of trade and development reconsidered. *Economica* 64, 527–37.

Brown, W.A., Jr. 1940. *The International Gold Standard Reinterpreted* 1914–1934, 2 vols. New York: NBER.

Cantillon, R. 1755. *Essai sur la nature du commerce en général* (written c. 1730). Ed. and trans. H.

Higgs, London: Macmillan, 1931.

Caves, R.E. 1982. *Multinational Enterprise and Economic Analysis*. Cambridge: Cambridge University Press.

Caves, R.E. and Jones, R.W. 1977. *World Trade and Payments*, 2nd edn. Boston: Little Brown.

Chipman, J.S. 1965–66. A survey of the theory of international trade. *Econometrica* 33, July, 477–519; 33, October, 685–760; 34, January, 18–76.

Clarida, R. 2006. G7 current account imbalances: sustainability and adjustment. Working Paper No. 12194. Cambridge, MA: NBER.

Clarida, R., ed. 2007. *G7 Current Account Imbalances: Sustainability and Adjustment*. Chicago: University of Chicago Press.

Cunliffe Committee. 1918. *The Interim Report of the Committee on Currency and Exchanges After the War*. London: HMSO.

De Marchi, N. 1976. Anomaly and the development of economics: the case of the Leontief paradox. In *Method and Appraisal in Economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

Dixit, A. and Norman, V. 1980. *The Theory of International Trade: A Dual, General Equilibrium Approach*. Cambridge: Cambridge University Press.

Edgeworth, F.Y. 1925. *Papers Relating to Political Economy*, 3 vols. London: Macmillan for the Royal Economic Society.

Ellis, H.S. and Metzler, L.A., eds. 1949. *Readings in the Theory of International Trade*. American Economic Association. Philadelphia: Blakiston.

Fetter, F.W. 1965. *Development of British Monetary Orthodoxy 1797–1875*. Cambridge, MA: Harvard University Press.

Flanders, M.J. 1974. Some problems of stabilization policy under floating exchange rates. In *Trade, Stability and Macroeconomics, Essays in Honor of Lloyd A. Metzler*, ed. G. Horwich and P. Samuelson. New York: Academic Press.

Flanders, M.J. 1989. *International Monetary Economics 1870–1960*. Cambridge: Cambridge University Press.

Flanders, M.J. and Helpman, E. 1979. An optimal exchange rate peg in a world of general floating. *Review of Economic Studies* 46, 533–42.

Fleming, J. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9, 369–79.

Ford, A.G. 1962. *The Gold Standard 1880–1914: Britain and Argentina*. Oxford: Clarendon Press.

Frenkel, J. and Johnson, H. 1976. *The Monetary Approach to the Balance of Payments*. London: Allen & Unwin.

Friedman, M. 1953. A case for flexible exchange rates. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Gervaise, I. 1720. *The System or Theory of the Trade of the World*. London: A. Woodfall. Baltimore: Johns Hopkins University Press, 1954.

Graham, F.D. 1923. The theory of international values re-examined. *Quarterly Journal of Economics* 38, 54–86. Repr. in Ellis and Metzler (1949).

Grossman, G.M. and Helpman, E. 1991. *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.

Grossman, G.M. and Helpman, E. 2002. *Interest Groups and Trade Policy*. Princeton: Princeton University Press.

Grubel, H.G. and Lloyd, P.J. 1975. *Intra-industry Trade: The Theory and Measurement of International Trade in Differentiated Products*. New York: Wiley.

Haberler, G. 1930. Die Theorie der komparativen Kosten und ihre Aüswertung für die Begründung des Freihandels. *Weltwirtschaftliches Archiv* 32, 349–70. Trans. as ‘The theory of comparative costs and its use in the defense of free trade’ in *Selected Essays of Gottfried Haberler*, ed. and trans. A.Y.C. Koo. Cambridge, MA: MIT Press, 1985.

Haberler, G. von. 1936. *The Theory of International Trade*, English translation of German original, 1933. London: Hodge.

Hayek, F.A. 1937. *Monetary Nationalism and International Stability*. New York: Augustus M. Kelley, 1964.

Heckscher, E. 1935. *Mercantilism*. 2nd edn, ed. E.F. Soderlund. London: George Allen & Unwin, 1955.

Heckscher, E.F. 1918, and Ohlin, B. 1924. *Heckscher–Ohlin Trade Theory*, trans. and ed. with introduction by H. Flam and M.J. Flanders. Cambridge, MA: MIT, 1991.

Helpman, E. and Krugman, P. 1985. *Market Structure and Foreign Trade*. Cambridge, MA: MIT Press.

Hollander, J.H. 1910–11. The development of the theory of money from Adam Smith to David Ricardo. *Quarterly Journal of Economics* 25, 429–70.

Hume, D. 1752. Of the balance of trade, from *Political Discourses*. In *David Hume: Writings on Economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.

Jones, R.W. 2000. *Globalization and the Theory of Input Trade*. Cambridge, MA: MIT Press.

Keynes, J.M. 1929. The German transfer problem. *Economic Journal* 39, 1–7. Repr. in Ellis and Metzler (1949). Subsequent ‘Rejoinder’ and ‘Reply to Ohlin and Rueff’ in *Economic Journal* 39, June and September, 1929, 179–82, 404–8.

Keynes, J.M. 1937. *General Theory of Employment, Interest, and Money*. London: Macmillan.

Krugman, P. 1987. Is free trade passé? *Journal of Economic Perspectives* 1(2), 131–44.

Lane, P. 2001. The new open economy macroeconomics: a survey. *Journal of International Economics* 54, 235–66.

Leontief, W.W. 1953. Domestic production and foreign trade: the American capital position re-examined. *Proceedings of the American Philosophical Society* 97, 332–49. Repr. in H.G. Johnson and R. E. Caves, *Readings in International Trade*. Homewood, IL: R.D. Irwin, 1968.

Lerner, A.P. 1932. The diagrammatical representation of cost conditions in international trade. *Economica* 12, 346–56. Repr. in Lerner (1953) and Lerner (1983).

Lerner, A.P. 1933. Factor prices and international trade. Seminar paper, LSE. Repr. in Lerner (1953) and Lerner (1983).

Lerner, A.P. 1934. The diagrammatical representation of demand conditions in international trade. *Economica* n.s. 1, 319–34. Repr. in Lerner (1953) and Lerner (1983).

Lerner, A.P. 1944. *Economics of Control*. New York: Macmillan.

Lerner, A.P. 1953. *Essays in Economic Analysis*. London: Macmillan.

Lerner, A.P. 1983. *Selected Economic Writings of Abba P. Lerner*, ed. D.C. Colander. New York: New York University Press.

Locke, J. 1696. *Several Papers Relating to Money, Interest and Trade, Etcetera*. New York: Augustus M. Kelley, 1968.

Marshall, A. 1879. *The Theory of Foreign Trade and other Portions of Economic Science Bearing on the Principle of Laissez Faire*. ed. J.K. Whitaker. New York: The Free Press, 1975.

Marshall, A. 1923. *Money, Credit and Commerce*. London: Macmillan.

Martin, R. and Sunley, P. 1996. Paul Krugman's geographical economics and its implication for regional development theory: a critical assessment. *Economic Geography* 72, 259–92.

Meade, J.E. 1955. The case for variable exchange rates. *Three Banks Review* 27 (September), 3–27.

Meade, J.E. 1951. *The Balance of Payments and the Balance of Payments: Mathematical Supplement*. Oxford: Oxford University Press.

Metzler, L.A. 1960. The process of international adjustment under conditions of full employment: a Keynesian view. In *Collected Papers*, ed. L.A. Metzler. Cambridge, MA: Harvard University Press, 1973; and *Readings in International Economics*, ed. R.E. Caves and H.G. Johnson, American Economic Association. London: Allen & Unwin, 1968.

Mill, J.S. 1848. *Principles of Political Economy*. Oxford: Oxford University Press, 1994.

Mundell, R.A. 1960. The monetary dynamics of international adjustment under fixed and flexible exchange rates. *Quarterly Journal of Economics* 74, 227–57.

Mundell, R.A. 1961. A theory of optimum currency areas. *American Economic Review* 51, 657–65.

Obstfeld, M. and Rogoff, K. 1995. Exchange rate dynamics *redux*. *Journal of Political Economy* 103, 624–60.

Ohlin, B. 1924. *Handelns teori*. Repr. In *Heckscher–Ohlin Trade Theory*, ed. and trans. H. Flam and M. J. Flanders. Cambridge, MA: MIT Press, 1991.

Ohlin, B. 1929. The reparation problem: a discussion, and a rejoinder to Jacques Rueff. *Economic Journal* 39. Repr. in Ellis and Metzler (1949).

Ohlin, B. 1933. *International and Interregional Trade*. Cambridge, MA: Harvard University Press, 1967.

Ricardo, D. 1811. The high price of bullion, a proof of the depreciation of bank notes. In *Pamphlets and Papers 1809–1811*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1817. *Principles of Political Economy and Taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Robinson, J. 1937. Beggar-my-neighbour remedies for unemployment. In *Essays on the Theory of Employment*. New York: Macmillan, 1937, 1947; Repr. in Ellis and Metzler (1949) from the 1947 edition.

Rockoff, H. 2003. How long did it take the United States to become an optimal currency area? In *Monetary Unions: Theory, History, Public Choice: International Studies in Money and Banking* 18, ed. F.H. Capie and G.E. Wood. London and New York: Routledge.

Samuelson, P.A. 1939. The gains from international trade. *Canadian Journal of Economics and Political Science* 5, 195–205. Repr. in Samuelson (1966).

Samuelson, P.A. 1948. International trade and the equalisation of factor prices. *Economic Journal* 58, 163–84. Repr. in Samuelson (1966).

Samuelson, P.A. 1949. International factor-price equalisation once again. *Economic Journal* 59, 181–97. Repr. in Samuelson (1966).

Samuelson, P.A. 1951–2. A comment on factor-price equalisation. *Review of Economic Studies* 19, 121–2. Repr. in Samuelson (1966).

Samuelson, P.A. 1953–4. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21, 1–20. Repr. in Samuelson (1966).

Samuelson, P.A. 1966. *The Collected Scientific Papers of Paul A. Samuelson*, ed. J. Stiglitz. Cambridge, MA: MIT Press.

Sayers, R.S. 1936. *Bank of England Operations 1890–1914*. Westport, CT: Greenwood Press, 1970.

Sayers, R.S. 1976. *The Bank of England, 1891–1944*, 3 vols. Cambridge: Cambridge University Press.

Schumpeter, J.A. 1954. *History of Economic Analysis*. George Allen & Unwin.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. ed. E. Cannan. New York: Modern Library, 1937.

Stolper, W.F. and Samuelson, P.A. 1941. Protection and real wages. Repr. in Ellis and Metzler (1949).

Stuvel, G. 1951. *The Exchange Stability Problem*. Leyden, New York: Augustus M. Kelley.

Taussig, F.W. 1928. *International Trade*. New York: Macmillan.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*, ed. F.A. Hayek. New York: Augustus M. Kelley.

Tobin, J. 1978. A proposal for international monetary reform. *Eastern Economic Journal* 4, 153–9. Repr. in J. Tobin, *Essays in Economics: Theory and Policy*, Cambridge, MA: MIT Press, 1982.

Torrens, R. 1820. *An Essay on the External Corn Trade 1820*. London: Longman, Rees, Orme, Brown & Green.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper and Bros.

White, H.D. 1934. Haberler's Der internationale Handel, Ohlin's interregional and international trade. *Quarterly Journal of Economics* 48, 727–41.

Williams, J.H. 1929. The theory of international trade reconsidered. *Economic Journal* 39, 195–209. Repr. in Williams (1944).

Williams, J.H. 1934. The world's monetary dilemma: internal versus external stability. *Proceedings of the Academy of Political Science* 16(1), 62–8. Repr. in Williams (1944).

Williams, J.H. 1937. The adequacy of existing currency mechanisms under varying circumstances. *American Economic Review* 27(1), 151–68. Repr. in Williams (1944) with the title ‘International monetary organization and policy’.

Williams, J.H. 1944. *Postwar Monetary Plans and Other Essays*. New York: Albert A. Knopf.

How to cite this article

Flanders, M. June. "international economics, history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000266> doi:10.1057/9780230226203.0827

close print preview

The New Palgrave Dictionary of Economics Online

international finance

Maurice Obstfeld
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Fundamental to international finance is the idea of ‘external balance’, whereby a country's external indebtedness does not threaten its ability to meet its international obligations. The requirements of external balance have varied with the nature of the linkages among economies across historical episodes. This article both reviews the major developments in the economic analysis of external balance and traces how nations have sought to achieve it from the era of the gold standard in 19th century through the Bretton Woods system to the era of floating exchange rates that began in 1973.

Keywords

balance of payments; balanced trade; Bank of England; barter trade theory; beggar-thy-neighbour; Bretton Woods system; central banks; consumption smoothing; current account; default; Dornbusch, R.; external balance; fiscal policy; Fisher, I.; fixed exchange rates; floating exchange rates; Friedman, M.; gold standard; gold-exchange standard; Great Depression; Hume, D.; income–specie–flow mechanism; inflation; inflationary expectations; interest rates; internal balance; international capital flows; international finance; international financial adjustment; International Monetary Fund; international reserves; intertemporal maximization hypothesis; investment; Keynes, J. M.; liquidity preference; Lucas, R.; Meade, J.; mercantilism; money supply; moral hazard; Mundell, R.; Nash–Cournot equilibrium; national income identity; Nurkse, R.; Obstfeld, M.; Ohlin, B.; Pigou, A.; price stability; price–specie–flow mechanism; protectionism; public debt; quantity theory of money; real-balance effect; Ricardo, D.; Samuelson, P; Seigniorage; sovereign debt; Sterilization; sticky prices; sticky wages; terms of trade; Third World debt; transfer problem; Triffin, R.; unemployment; Viner, J.; Walras's law; willingness-to-pay hypothesis

Article

International finance is concerned with the determination of real income and the allocation of consumption over time in economies linked to world markets. Fundamental to international finance is the somewhat elusive idea of ‘external balance’, which in practice entails a path of external indebtedness that does not threaten a country's ability to meet its international obligations. Because the nature of the linkages among economies has varied across historical episodes, the requirements of external balance have varied as well. International finance studies the policies and market forces which may lead to external balance under various conditions. The history of the subject illustrates how the nature of world market linkages has itself been changed by national efforts to cope with external constraints.
The national income identity is the necessary groundwork for any discussion of external balance. The national income of an open economy equals domestic product plus net factor payments from abroad plus net international transfer payments; the current account equals net exports of goods and services (including all net factor payments) plus net transfers. If national expenditure is defined as the sum of consumption and investment (by both the public and private sectors), the national income identity asserts that national income less national expenditure equals the current account. When in surplus, the current account therefore measures the growth of the economy's external assets; when in deficit, it measures the growth of external debt.

The classical paradigm

The classical Ricardo–Mill barter trade theory shows how the terms of trade and international production pattern are determined in a stationary world economy with balanced trade. The classical analysis of the transition to balanced trade may be viewed as an account of the convergence process to the long-run barter equilibrium. As Ricardo noted in the *Principles* (1817):

> Gold and silver having been chosen for the general medium of circulation, they are, by the competition of commerce, distributed in such proportions amongst the different countries of the world as to accommodate themselves to the natural traffic which would take place if no such metals existed, and the trade between countries were purely a trade of barter.

Historically, however, the classical paradigm of external adjustment preceded Ricardo. Major elements of the theory had been expounded quite clearly by the early 18th century, but the most coherent and effective exposition was given by Hume in 1752.

Hume assumed a world economy that settles trade imbalances exclusively through imports or exports of precious metals that also serve as money. Building on the quantity theory of money, he constructed a full dynamic model of the balance of payments and the terms of trade. The famous price–specie–flow mechanism was put forth as an automatic market process that always works to restore balanced trade. Hume's goal was to refute mercantilist and protectionist arguments by showing that market forces would ensure in the long run a ‘natural’ distribution of specie among countries.

Hume invited his readers to imagine that four-fifths of Great Britain's money supply were ‘annihilated in one night’. British prices would naturally fall, he argued, cheapening British exportables relative to foreign goods and creating a trade surplus. As a result of this surplus Britain would accumulate foreign wealth in the form of specie, seeing its money supply, and hence its prices, rise. Abroad, the drain of specie would lower prices. Britain's trade surplus would dwindle and eventually disappear once its terms of trade had improved sufficiently, and at this point, the natural distribution of specie would prevail. A hypothetical fivefold increase in Britain's money supply would set off the reverse process, involving an initial improvement in Britain's terms of trade and a trade balance deficit. Over time, specie would flow abroad as the terms of trade deteriorated and external equilibrium was restored.

There is little exaggeration in saying that issues raised by Hume's analysis dominated writing in international finance up until the inter-World War years. In a period that culminated in the classical gold standard, it was natural to take as the benchmark of external balance an absence of international specie movements. Hume had placed relative price movements at the centre of his account of how external balance would be attained, but subsequent writers asked whether direct income or wealth effects might also be operative, and whether external adjustment could take place in some cases without price changes. Such questions arose in the 1929 Keynes–Ohlin debate over the German transfer problem, but as Viner (1937) showed, the questions had been raised much earlier.

A simple model of a Humean world makes apparent some of the assumptions underlying the price–specie–flow mechanism. Such a model also serves as a springboard for understanding later developments in the analysis of external adjustment. (A more detailed exposition of a similar model is given by Dornbusch, 1973, whose analytical approach is, however, somewhat different from that taken here.)

Assume a world of two countries, each specialized in the production of a single commodity that is consumed in both countries. With given supplies of capital and labour within each country and perfect wage flexibility, home-country output is fixed at the full-employment level x and foreign-country output is fixed at y. Let q denote the price of y-goods in terms of x-goods (the terms of trade), z domestic expenditure measured in x-goods, and z^* foreign expenditure, also measured in x-goods. Then the domestic demands for the two goods are $c_x(q, z)$ and $c_y(q, z)$, while the foreign demands are $c_x^*(q, z^*)$ and $c_y^*(q, z^*)$.

Expenditure is determined by monetary conditions. The money supplies M and M^* are for simplicity taken to consist entirely of gold, and P and P^* denote the money prices of home and foreign goods, respectively. The exchange rate between domestic and foreign currency can be set at unity with no loss of generality, so the terms of trade, q, equal P^*/P. In each country there is a desired long-run (or ‘natural’) money supply: this is proportional to nominal output, and saving behaviour is governed by discrepancies between natural and actual money supplies. Because a country's net saving here equals its current account, which by assumption is settled in specie, saving behaviour determines the evolution of national money supplies. These evolve according to the laws

$$dM/dt = \theta(\chi Px - M) \quad dM^*/dt = \theta^*(\chi^* P^* y - M^*),$$

where χ (χ^*) is the reciprocal of the home (foreign) country's long-run monetary velocity and θ (θ^*) is the home (foreign) marginal propensity to dissave out of monetary wealth. Expenditure levels are therefore

$$z(1-\theta\chi)x+\theta M/P,\ z^{*}=(1-\theta^{*}\chi^{*})qy+\theta^{*}M^{*}/P,$$

where $\theta\ \chi,\ \theta^{*}\chi^{*}<1$.
The model is closed by two equilibrium conditions. With a given world stock of monetary gold, M^{w}, home saving must equal foreign dissaving, that is, world expenditure must equal world output. In addition, the market for domestic goods must clear. By Walras's Law, these two equilibrium conditions imply equilibrium in the market for foreign goods.
The condition of zero desired world saving is $(dM/dt)+(dM^{*}/dt)=0$, or

$$P=\frac{\theta M+\theta^{*}(M^{W}-M)}{\theta\chi x+\theta^{*}\chi^{*}qy} \tag{1}$$

Equation (1) shows that, for given terms of trade and money supplies, the world price level adjusts to maintain consistency between the countries' saving plans. In equilibrium, this condition makes P a function of q and M, $P=P(q,M)$, with

$$\frac{q}{P}\frac{\partial P}{\partial q}=\frac{-\theta^{*}\chi^{*}qy}{\theta\chi x+\theta^{*}\chi^{*}qy}>-1,\ \frac{M}{P}\frac{\partial P}{\partial M}=\frac{(\theta-\theta^{*})M}{\theta\chi Px+\theta^{*}\chi^{*}P^{*}y}\lesseqgtr 0.$$

The market for x-goods clears when

$$c_{x}[q,(1-\theta\chi)x+\theta M/P]+c_{x}^{*}[q,(1-\theta^{*}\chi^{*})qy+\theta^{*}M^{*}/P]=x. \tag{2}$$

Substitution of $P=P(q,M)$ and $M^{*}=M^{W}-M$ into (2) gives the curve describing combinations of M and q at which both goods markets clear and aggregate world saving is zero. The curve is labelled XX in Figure 1 and is shown with a negative slope. The assumptions giving rise to this negative slope are crucial for analysing the Humean adjustment process. An increase in M (which necessarily implies an equal fall in M^{*}) causes an excess demand for x-goods equal to

$$\frac{\theta\theta^{*}(M+M^{*})}{P(\theta M+\theta^{*}M^{*})}[\partial c_{x}/\partial z-\partial c_{x}^{*}/\partial z^{*}]$$

near the system's long-run equilibrium (where $dM/dt = dM^*/dt = 0$). The term $\partial c_x/\partial z - \partial c_x^*/\partial z^*$ is the difference between the two countries' marginal propensities to spend on home-country goods; if the home-country marginal propensity is larger – the 'orthodox' presumption in transfer analysis (Samuelson, 1971) – a redistribution of nominal balances in favour of the home country creates an incipient excess demand for its output. This excess demand is eliminated by a fall in *q* if the home-goods market is Walras stable, so XX slopes downward under standard assumptions concerning marginal spending propensities and Walrasian stability.

Figure 1

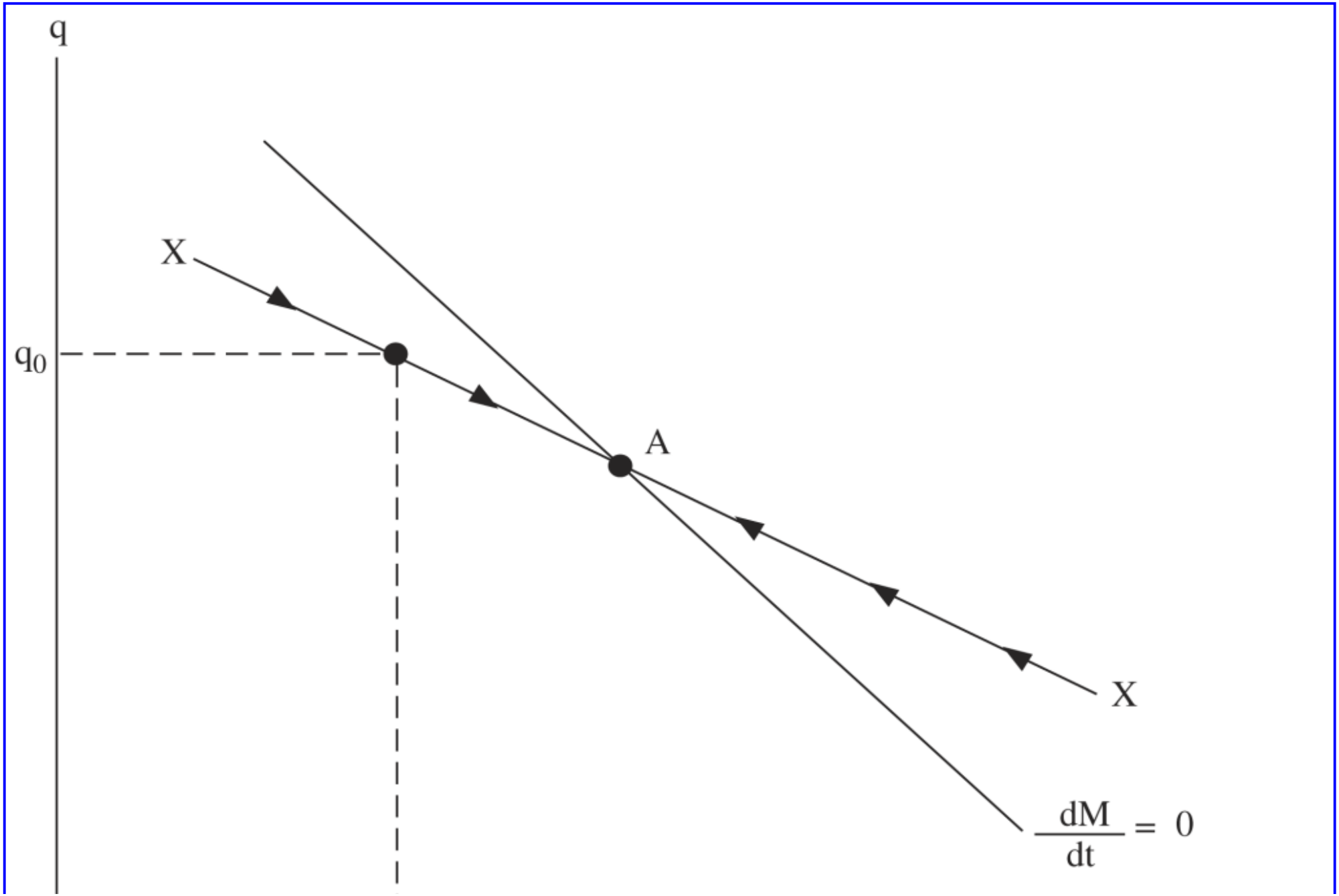

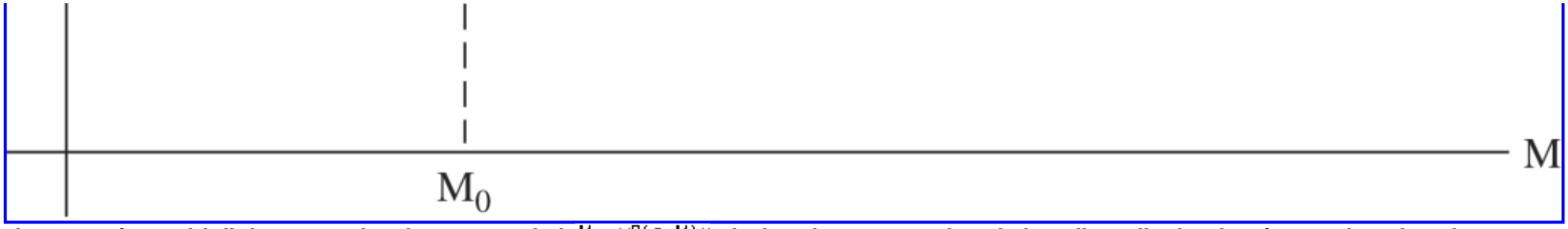

The curve in figure 1 labelled $dM/dt = 0$ describes points at which $M = xP(q, M)x$. This locus has a negative slope algebraically smaller than that of XX. With goods markets continuously in equilibrium, the world economy travels along XX to its long-run equilibrium at point A, where international prices and the distribution of specie give rise to balanced trade.

In most respects the model confirms Hume's account of the external adjustment process. A (small) fall in M to M_0, for example, leads to terms of trade q_0, which are worse for the home country. The terms-of-trade change is a direct result of the transfer of purchasing power to foreigners, which produces an excess supply of home goods at the initial prices. The home balance-of-payments surplus that simultaneously emerges causes a gradual redistribution of money in favour of the home country, so the home terms of trade improve during the transition to external balance.

If $\theta = \theta^*$, equilibrium P is a function of q alone, with a negative elasticity greater than –1. The rise in q caused by a fall in M is thus accompanied by a less-than-proportional fall in P and a rise in P^* that are reversed as the economy returns to point A. These results are in accord with Hume's predictions, but they need not hold if the expenditure responses to real balances differ sufficiently in the two countries. If $\theta > \theta^*$, a transfer of money abroad raises world saving for given terms of trade, so P^* may *fall* along with P and then rise during the subsequent adjustment. Likewise, if $\theta < \theta^*$, a money transfer abroad may reduce world saving sufficiently that P must *rise*, along with P^*, to restore goods-market equilibrium in the short run. In this case, the initial response to the disturbance is followed by price deflation in both countries.

This stylized version of Hume's paradigm may be used to analyse the transfer problem. Suppose that ownership of a portion of the foreign country's endowment is given to the home country. Does the home trade deficit necessarily increase by the amount of the transfer, or is the transfer undereffected, requiring a flow of specie to the home country to balance international accounts? A second focus of debate in the literature is the possibility that the transfer imposes a ‘secondary burden’ on the paying country by adding an equilibrium terms-of-trade deterioration to the primary income burden. Keynes and Ohlin clashed on this point in 1929, with Keynes arguing that the secondary burden is inevitable.

To simplify, suppose that $\theta = \theta^*$ and $\gamma = \gamma^*$. Since long-run money demand rises with income, the $dM/dt = 0$ locus shifts to the right, implying that the transfer at first is undereffected and that the world's gold stock is redistributed toward the home country. Under the standard assumption regarding marginal spending propensities, the transfer also creates an excess demand for x-goods at the initial terms of trade, so XX shifts downward. A secondary burden is thus imposed on the paying country, and this burden worsens over time as balanced trade is reestablished.

The interwar period

The years between the World Wars saw a partial and ultimately unsuccessful return to the gold standard, followed by extensive experimentation with floating exchange rates and direct controls on international payments as means of attaining external balance. Nurkse's (1944) account of the period is probably the most influential one. Writers on international finance continued to conceptualize external balance in terms of reserve movements. The spread of the gold-exchange standard, under which central banks held as foreign reserves currencies tied to gold as well as gold itself, broadened the class of assets through which balance-of-payments deficits were financed.

International capital movements were discussed increasingly in the theoretical literature, but they were viewed for the most part as an adjunct to the classical balance-of-payments adjustment mechanism. The theoretical discussions merely formalized a mechanism that had long been exploited by the Bank of England to regulate gold flows. A country that suddenly developed a trade deficit would face declining international reserves, a declining money supply, and rising interest rates. Rising interest rates would, however, attract foreign capital inflows and thus dampen the resulting deficit in the balance of payments. On this view, interest-sensitive capital flows had a potentially stabilizing role to play in discouraging protracted reserve flows. Given the turbulent conditions of the period, contemporary writers fully recognized that capital flows motivated by fears of devaluation or political instability could just as well destabilize an already bad external payments problem.

Such ‘short-term’ or interest-sensitive capital movements were generally discussed separately from ‘long-term’ international capital movements which directly financed investment or government expenditures. Theoretical discussions of long-term capital movements focused mainly on the transfer mechanism, the balance-of-payments and terms-of-trade adjustments that would accompany an inter-country transfer of capital. Conspicuously absent from the literature were attempts to develop a normative intertemporal theory of international capital transfer. Such a theory naturally would have extended the prevailing external balance concept to comprise changes in nations’ *overall* indebtedness rather than

just changes in the central bank's foreign assets. It had been known, at least since Ricardo's *Principles*, that producers and consumers could gain if long-term foreign investment equalized profits internationally. The insight did not dominate thinking about the nature of external balance.

This gap in the literature is surprising in view of the developments in international capital markets over the previous century. Huge flows of long-term capital, primarily from Britain, had financed railroad construction and other investment in the Western Hemisphere. France and Germany also made significant foreign loans. In the early 1930s, widespread foreign debt default among the Latin American countries highlighted the need to analyse formality the sustainability of external debt paths. In the world assumed by Hume, specie flows had been the only means of settling current-account imbalances, and a concept of external balance based on balance-of-payments equilibrium had been defensible. Such a concept of external balance was outmoded, however, in a world where other types of asset trade could finance the current account.

The necessary change of perspective did not occur for several decades. Instead, the events and ideas of the interwar period led international financial theory to turn away sharply from the concern with the dynamics of international adjustment underlying the classical model. Emphasis shifted inward, to the interaction between the balance of payments and domestic economic conditions.

The Bretton Woods period

The interwar experience had a profound influence on both the institutional framework of postwar international finance and the theoretical orientation of researchers. The international agreement reached at Bretton Woods in 1944 set up a world trading community linked by fixed dollar exchange rates, with a United States commitment to peg the dollar price of gold at $35 per ounce providing an anchor for the world price level. The agreement's provisions aimed to promote free trade in goods, but private capital movements were viewed as potentially disruptive and the widespread capital controls then in force were not discouraged. A prevailing view that flexible exchange rates had failed during the interwar period motivated the adoption of a fixed-rate system. Provision was made, however, for infrequent exchange-rate adjustment, after due consultation, in circumstances of ‘fundamental disequilibrium’ in the balance of payments.

Central to the design of the Bretton Woods system was a desire to avoid unemployment and ensure price-level stability. In the interwar years, many governments had resorted to competitive currency depreciations and trade restrictions aimed at reducing domestic unemployment. These ‘beggar-thy-neighbour’ moves made all countries worse off. Having recently experienced the hardships of the worldwide Great Depression, the Bretton Woods signatories recognized the goal of ‘internal balance’ – full employment with price stability – as a key aim of government policy. An International Monetary Fund was set up to reconcile the goals of internal and external balance. It was hoped that the availability of Fund credit would make it unnecessary for members to tolerate high unemployment in pursuing external balance, or to interfere with trade flows in pursuing internal balance.

In an environment of fixed exchange rates and extremely limited capital mobility, the overriding external consideration for governments was the available stock of foreign, particularly dollar, reserves. The operative external target was therefore the acquisition of as many dollars as possible through balance-of-payments surpluses. As the reserve centre, the United States enjoyed the privilege of being able to finance its own balance-of-payments deficits by borrowing dollars from foreign central banks. In reality, however, the United States was not totally free of a reserve constraint. Foreign central banks could, and did, use their dollars to buy gold from the US authorities at the official price. The problem of gold losses became important as the postwar period of ‘dollar shortage’ ended in the late 1950s. In 1960, Triffin put the American external dilemma in its most sombre light: Once foreign official dollar holdings exceed the official value of the US gold stock, it would become impossible to satisfy all foreign claims to US gold without a rise in the dollar price of metal. The resulting confidence problem, Triffin predicted, would undermine the stability of the Bretton Woods system.

As it developed immediately after World War II, international financial theory reflected the new institutional arrangements, along with the economic assumptions underlying Keynes's (1936) diagnosis of the unemployment of the 1930s. The new paradigm, set forth very effectively by Metzler (1948, pp. 212–13), assumed sticky price levels and wages along with fixed exchange rates, thus precluding the relative-price adjustments at the heart of the classical paradigm while opening the door to employment fluctuations:

> The important feature of the classical mechanism … is the central role which it attributes to the monetary system. The classical theory contains an explicit acceptance of the Quantity Theory of Money as well as an implied assumption that output and employment are unaffected by international monetary disturbances. In other words, the classical doctrine assumes that an increase or decrease in the quantity of money leads to an increase or decrease in the aggregate money demand for goods and services, and that a change in money demand affects prices and costs rather than output and employment … . The essence of the new theory is that an external event which increases a country's exports will also increase imports *even without price changes*, since the change in exports affects the level of output and hence the demand for all goods. In other words, movements of output and employment play much the same role in the new doctrine that price movements played in the old.

An increase in external demand for a country's exports, for example, would raise the country's trade surplus in the first instance, but once the multiplier effect of the disturbance had raised income and hence import spending, the initial impact on the trade balance would be reduced. Metzler noted, however, that even if one assumed that investment spending responds positively to a rise in real income, it was unlikely that multiplier effects alone would ensure complete trade-balance adjustment in the short run.

The Keynesian account of external adjustment therefore contained an important gap. Private capital movements were largely ruled out in the Keynesian models, so incomplete trade-

balance adjustment implied incomplete balance-of-payments adjustment and growing or shrinking central-bank foreign reserves. The models pushed monetary factors to the background, implicitly or explicitly assuming that central-bank sterilization operations were offsetting any monetary effects of the balance of payments. Only a few of the early postwar theorists, notably Meade (1951), assigned an important role to monetary factors.

Even if the sterilization assumption were granted, however, consideration of the system's inherent dynamics made clear the infeasibility of a *permanent* sterilization policy. Countries with persistent deficits would ultimately exhaust their available international reserves, including IMF credit; and even surplus countries might be unable to sterilize indefinitely if domestic financial markets were thin. How, then, could trade-balance equilibrium even be restored after a permanent external shock? Fiscal policy could be effective in situations where the needs of internal and external balance were both served by the same measure. In dilemma situations where fiscal measures could move the economy toward external balance only at the cost of increasing its distance from internal balance, the ‘fundamental disequilibrium’ clause of the IMF Articles of Agreement could be invoked and the currency devalued. But no *automatic* market mechanism pushing the economy toward balance-of-payments equilibrium was featured in the early postwar writing.

In a series of remarkable papers published in the early 1960s, Mundell revived the explicit dynamic analysis of international adjustment. His models placed the monetary sector in the foreground, adopting a Keynesian liquidity-preference view of interest-rate determination. A prescient paper by Metzler (1960), written at about the same time, took a similar approach.

Mundell's paper on ‘The International Disequilibrium System’ (1961) criticized the Keynesian model's failure to account for the dynamic effects of payments imbalances. Even in a Keynesian world, Mundell argued, an income–specie–flow mechanism, analogous to Hume's price–specie–flow mechanism, ensures long-run balance-of-payments equilibrium. A ‘fivefold increase’ in a country's money supply, for example, depresses domestic interest rates, stimulates investment spending, and creates a deficit in the balance of payments. As the central bank loses reserves, however, the interest rate gradually rises and reduces investment, the process coming to an end (for a small country) only when the domestic money supply, the interest rate, investment, and output have returned to their original levels. The introduction of dynamic adjustment made it clear that sterilization could have only limited success as a policy response to permanent balance-of-payments disturbances. One source of dynamic effects, however, was not explicitly analysed in Mundell's work of the period. The omitted effect was the real-balance effect on expenditure, central to the classical account but possibly relevant (as Pigou had shown) under Keynesian conditions as well.

In line with the increasing international capital mobility that followed the European move toward currency convertibility in 1958, Mundell gave the capital account a prominent role in his models. The presence of capital mobility suggested a solution to the policy dilemmas that could arise under fixed exchange rates when the goals of internal and external balance appeared to conflict. Mundell showed that by gearing monetary policy to external balance and fiscal policy to internal balance, governments could simultaneously attain both goals. The key to the argument is the observation that monetary and fiscal expansion both raise output but have different effects on the capital account, monetary expansion causing capital outflows (by driving down the home interest rate) and fiscal expansion causing capital inflows (by raising the interest rate). With two independent instruments, both internal and external policy targets can be attained simultaneously.

While a major step forward, the Mundellian argument for a policy mix suffered from two drawbacks. First, the theoretical specification of the capital account as a function of international interest-rate levels was weak: it seemed unlikely that capital would flow at a uniform level forever even if the interest differential remained fixed. Missing was a discussion of stock equilibrium in international asset markets. The second problem with the policy mix was its definition of external balance. Would any policymaker view with satisfaction a permanently high interest rate that brought about balance-of-payments equilibrium by crowding out domestic investment and encouraging a build-up of external debt? Key considerations omitted from Mundell's model were the stock of net foreign claims and the associated flows of interest payments. Mundell himself (1968, p. 207) recognized that in many contexts, the definition of external balance as balance-of-payments equilibrium might be inadequate:

> Just as the composition of output is important (the division of output between investment and consumption affects additional growth targets), so an appropriate composition of the balance of payments is a legitimate target of policy.

Indeed, in spite of the continuing obligation to peg dollar exchange rates, the standard definition of external balance was becoming increasingly outmoded by the late 1960s. The balance of payments remained a legitimate concern, of course, in part because a large or persistent imbalance might look like ‘fundamental disequilibrium’ to the market and spark a speculative attack on the currency involved. But the increasing integration of national financial markets – a development epitomized by the growth of Eurocurrency trading – weakened the bite of the balance-of-payments constraint. In a hypothetical world of *perfect* capital mobility, a central bank short on reserves can essentially borrow them from abroad at no net cost simply by contracting domestic credit. Such an action, by causing an incipient rise in the home interest rate, leads to an instantaneous private capital inflow and an official reserve gain equal to the fall in domestic credit. The home interest rate, the money supply, output, and the national external debt are unchanged in the final equilibrium: the central bank holds more foreign assets and fewer domestic assets, while the home private sector, having made the mirror-image adjustment, holds fewer foreign assets and more domestic assets.

The case of perfect capital mobility is an extreme one that does not fit the facts of the late Bretton Woods period. None the less, the opportunities for central banks to borrow dollar reserves in the international capital market had grown since the early 1960s. The situation facing the United States was quite different. As the primary international reserve issuer, its responsibility was to peg the dollar price of gold, a responsibility that would have required the gearing of US monetary policy to that external commitment. In spite of such expedients

as the two-tier gold market established by central banks in 1968, the US did not succeed in preserving the dollar's link to gold. Triffin had been right. After a series of violent speculative attacks, the US severed the dollar's gold link in August 1971 and in December 1971 devalued the dollar against major foreign currencies. The patchwork system of fixed exchange rates proved unstable, and in the first months of 1973 the postwar period of floating exchange rates began.

Floating exchange rates

The industrialized countries adopted floating dollar exchange rates as an interim measure, but in fact a significant body of economists had come to advocate floating rates by 1973. Friedman's (1953) powerful case for flexible rates was the opening shot in a campaign to revise the then-prevailing view, expounded by Nurkse (1944), that the floating-rate experiments of the interwar years were disastrous. By the time Johnson wrote his well-known polemic of 1969, Friedman's views had gained many adherents.

The fundamental argument for floating rates was that they would free governments of the balance-of-payments constraint and allow them to use monetary policy to attain domestic economic goals. Equilibrium in the balance of payments would be automatic if central banks simply refrained from intervening in the foreign exchange market. At the same time, floating rates would permit central banks to target their nominal money supplies without being frustrated by offsetting interest-sensitive foreign reserve flows. Widespread restrictions on trade and capital movements, motivated in part by a desire to impede reserve flows under the fixed-rate regime, could be dismantled.

Subsequent experience was to provide only partial vindication to the advocates of floating. In the decade after 1973, barriers to capital movement were reduced to insignificant levels in many of the industrial countries. This development helped spark unprecedented growth in international financial intermediation. Under the new exchange-rate regime, however, policymakers became more acutely aware that the traditional definition of internal balance as full employment cum price stability really involved two, quite distinct, goals. Under a floating exchange rate, monetary expansion aimed at domestic unemployment translates immediately into currency depreciation, higher import prices, and heightened inflationary expectations. Conversely, a rapidly adjusting exchange rate provides a powerful channel through which inflationary expectations can have a direct and immediate effect on inflation in an open economy. Any short-run tradeoff between inflation and unemployment would therefore be less favourable under a floating rate. Floating rates certainly allow countries to choose their own trend inflation rates. But it soon became evident that if disturbances to the economy originated predominantly outside the money market, the inflationary cost of using monetary policy to target employment could be quite high.

Sharp exchange-rate movements might also have adverse distributional effects in the economy, and these, together with a desire for price-level stability, led central banks to intervene, at times heavily, in the foreign exchange market. Correspondingly, the predicted drop in central banks’ demand for international reserves did not materialize (although the composition of reserves did change over time as the Deutschmark and yen became important reserve currencies and the pound sterling retreated). Central banks’ use of foreign reserves to manage exchange rates did not necessarily imply an operative balance-of-payments constraint, however, since in many countries the same exchange-rate effects could have been achieved at an unchanged reserve level through domestic credit measures.

Under conditions of limited capital mobility, such as those existing in the early 1950s when Friedman wrote, the automatic balancing of international reserve movements by a floating exchange rate amounted essentially to the automatic balancing of the current account. With means other than reserve flows available to settle current-account imbalances, however, there is no theoretical necessity for a floating rate to balance the current account in the short run. A current-account deficit, say, can be financed entirely through domestic borrowing abroad with no decline in the central bank's foreign assets. Experience was to show that floating exchange rates themselves could not prevent the emergence of large and persistent current-account imbalances. These imbalances were problematic not only because they usually entailed costs of shifting productive resources between the economy's tradable and nontradable sectors, but also because they implied changes in foreign debt and thus in sustainable future consumption levels.

Attention therefore shifted to the mechanism of current-account adjustment under floating exchange rates and capital mobility, with researchers asking, as Hume had, if market forces would automatically push economies toward current-account balance. The new generation of dynamic open-economy models produced in the mid-1970s built on a number of antecedents in the literature. One of these was the neoclassical monetary approach to the balance of payments, which stressed the real balance effect and the transition to long-run payments equilibrium (see, for example, Frenkel and Johnson, 1976). The second important antecedent was the closed-economy literature on money and growth, which had clarified the stock-flow distinction in multi-asset models with wealth accumulation. As suggested by the rational-expectations revolution in macroeconomics, many model builders endowed agents with forward-looking exchange-rate expectations that played a key role in clearing the asset markets.

The intrinsic dynamic mechanism in these models is fuelled by wealth, broadly defined to include not only real monetary balances, but also foreign assets and possibly capital, physical as well as human. (See Obstfeld and Stockman, 1985, for a survey.) In line with the long-run nature of the inquiry, the ‘classical’ conditions of price flexibility and full employment were generally assumed, giving a productions structure similar to the Humean model set out above. Where the models differed essentially from Hume was in the wider spectrum of marketable assets, and in the resulting portfolio problem of private agents. Each given configuration of world asset stocks determines a short-run equilibrium defined by the requirements of market clearing in asset as well as goods markets. The resulting equilibrium wealth levels and real interest rates determine consumption levels at home and abroad, but there is no necessary requirement of current-account balance in the short run: goods-market equilibrium implies only that one country's planned current-account surplus equals the other's planned current-account deficit. The international adjustment process can now be visualized. All else equal, the deficit country is running down its wealth by borrowing from abroad, so its consumption is falling and foreign consumption is rising. Under the orthodox transfer criterion, this redistribution of wealth between the countries

causes the deficit country's terms of trade to deteriorate over time; if anticipated, the evolution of the terms of trade has further repercussions on world real interest rates and expenditure levels. The process comes to an end once the deficit country's consumption has fallen into line with its income, which is lower than initially because of the increased interest burden of the external debt. (A very similar adjustment process would take place with mobile capital and a fixed exchange rate, but reserve movements rather than exchange-rate movements would contribute to asset-market balance during the transition to long-run equilibrium.)

This simple picture of the adjustment process becomes more complicated once domestic capital accumulation is allowed. A current-account deficit may now finance an investment boom in which the deficit country's terms of trade improve over time. Eventually, however, the international wealth-flow mechanism restores a balanced current account. Further complications arise when the classical assumptions are dropped and Keynesian price stickiness in output markets is assumed. In such models, the approach to the long-run, full-employment equilibrium can be oscillatory.

For a single economy with Keynesian features, there is an analogue to the Mundellian idea of using monetary and fiscal policy simultaneously to attain internal and external targets. Figure 2, which is developed more fully in Obstfeld (1985, pp. 408–10), illustrates this approach. The downward sloping internal-balance schedule shows combinations of monetary and fiscal ease consistent with full employment. On the assumption that monetary ease improves the current account by depreciating the currency, the external-balance schedule, which shows policy settings consistent with some current-account target, slopes upward. The intersection of the two schedules shows how policies should be set to achieve both of the government's goals in the short run.

Figure 2

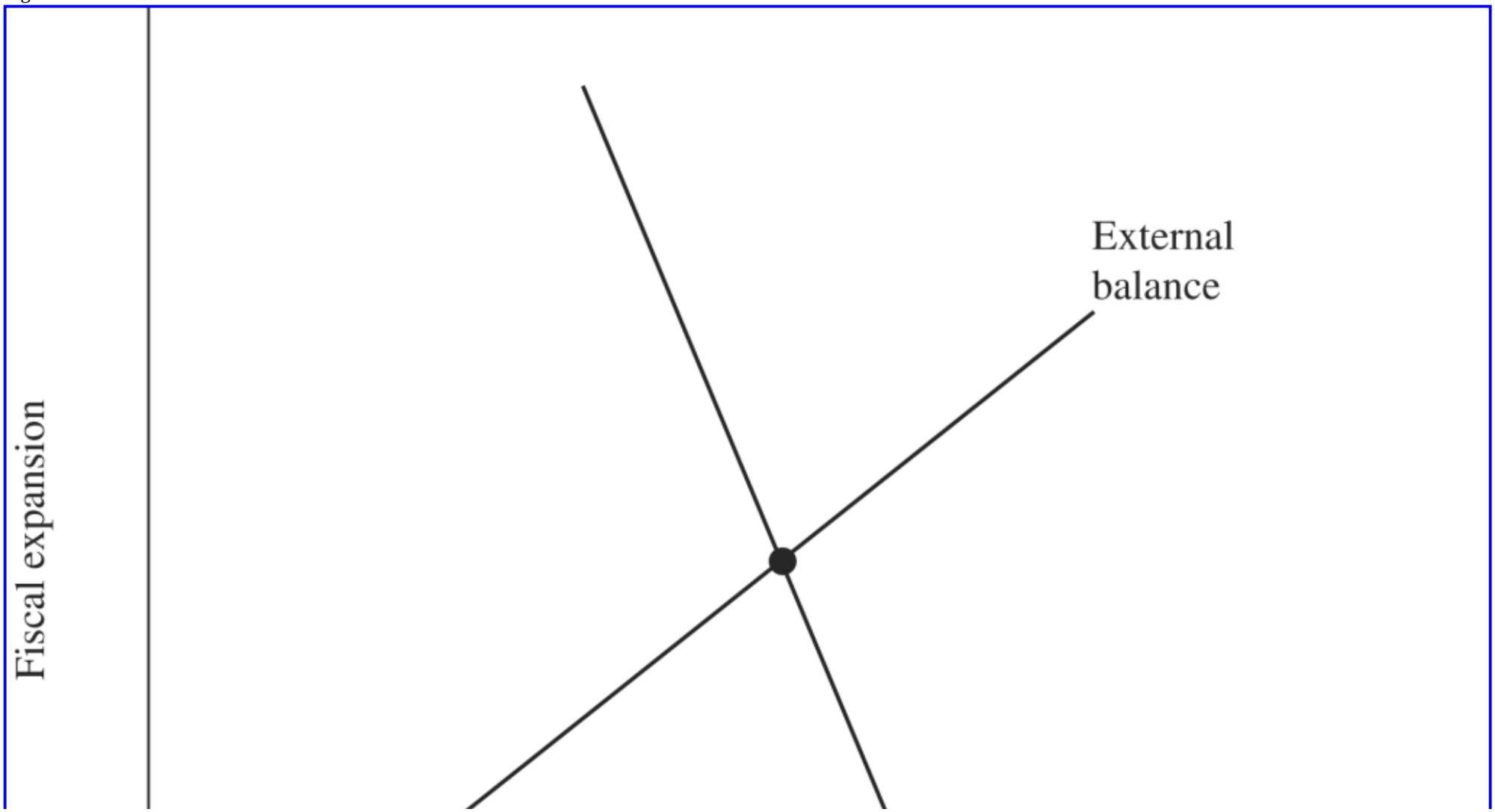

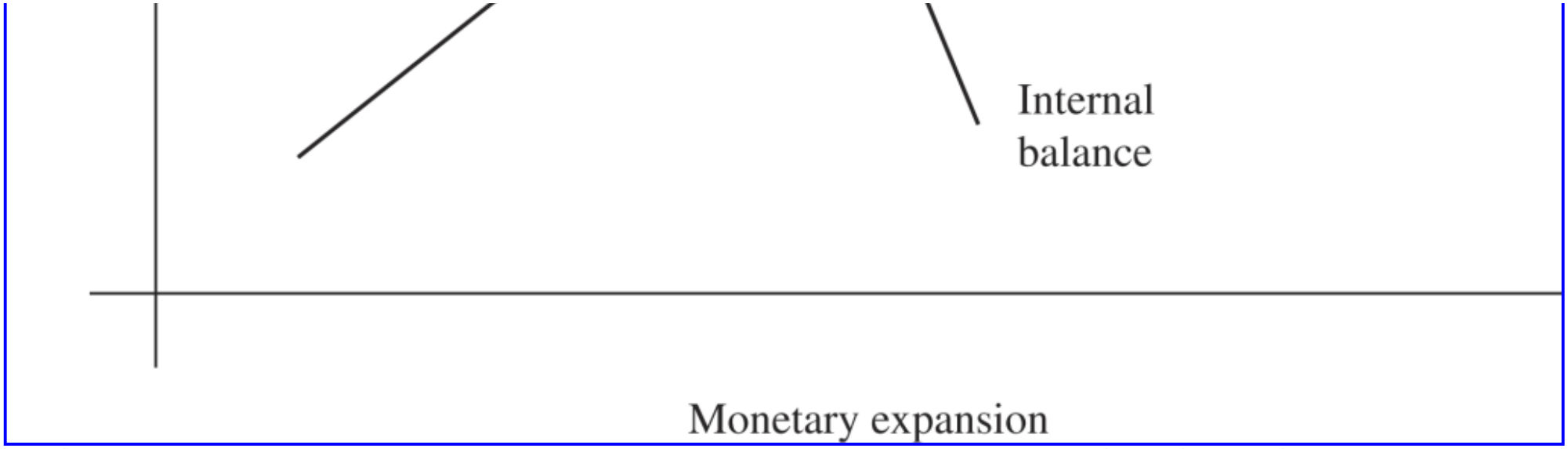

Even if one leaves aside the complex game-theoretic problems surrounding interactions between expectations and policy, the usefulness of the above framework as a normative guide is limited by its failure to incorporate some key dynamic elements. If the government can hit its targets only by running a budget deficit, its fiscal stance must eventually be reversed if the government debt is to be serviced. In addition, the policy equilibrium shown in Figure 2 may imply a domestic investment rate that is socially sub-optimal. Finally, the framework itself gives no guidance as to the appropriate external-balance criterion. The balanced current account reached in the hypothetical long-run equilibrium of a stationary world economy may be far off the mark in the short run in which policy decisions must be made. Recently, the theory of international finance has made partial progress in addressing these issues.

The intertemporal analysis of external balance

In the 1980s, it became increasingly common to analyse the dynamic behaviour of open economies in terms of the intertemporal maximization hypothesis applied by Fisher (1930) to the theory of saving and investment. As usual, this trend was the result of both new theoretical approaches in macroeconomics generally and of economic events that existing open-economy models seemed ill-equipped to analyse.

Lucas's (1976) influential critique of econometric policy evaluation was important in motivating the intertemporal approach. Lucas argued that the standard econometric models of the time would generally not be invariant to policy changes. Because the parameters estimated were not the ‘deep’ parameters describing preferences or technology, but instead reflected both deep structure and the policy environment prevailing over the estimation period, the models could not be used to analyse *changes* in the policy environment. Lucas's analysis suggested that more reliable policy conclusions might be drawn from open-economy models if demand and supply functions were derived from the optimal decision rules of maximizing households and firms.

Further impetus to develop an intertemporal approach came from events in the world capital market, particularly the international pattern of current accounts following the sharp oil-price increases of 1973–4 and 1979–80. The divergent patterns of a current-account adjustment by industrialized and developing countries raised the inherently intertemporal problem of characterizing the optimal response to external shocks. Neither classical nor Keynesian transfer analysis offered any reliable guidance on this question. Similarly, the explosion in bank lending to developing countries after the first oil shock sparked fears that some countries’ external debt burdens would become unsustainable. The need to assess developing-country debt levels again led naturally to the notion of an intertemporally optimal current-account deficit.

Any intertemporal analysis of external balance must begin by specifying the economy's technological and market opportunities for shifting consumption over time. These opportunities are described by the economy's intertemporal budget constraint, which specifies the terms on which the economy can borrow or lend abroad, as well as the domestic investment technology. Separate analysis of the public and private sector's budget constraints illuminates the link between the public finances and external imbalances, as measured by the balance of payments or by the current account. The economy-wide budget constraint results from consolidation of the public- and private-sector constraints.

Assume for simplicity that a single good is consumed and produced on each date, and consider the position of a small open economy that can borrow or lend internationally at the real interest rate ρ. For each date t, the government of the economy chooses a level of real government consumption, $g(t)$, and a (possibly negative) level of real transfers to the private sector, $\tau(t)$. The government finances its outlays by issuing debt, by printing money, and by drawing on the interest paid by the central bank's foreign reserves. (For present purposes,

the central bank's budget is best viewed as a component of the government's budget.) Let $b^G(t)$ denote real government bond holdings (other than central-bank foreign reserves), $D(t)$ the money value of central-bank domestic credit, $P(t)$ the money price level, and $r(t)$ real foreign reserves. If the government pays the interest rate ρ on the public debt $[-b^G(t)]$, then the path of government bond holdings satisfies the equation:

$$db^G(t)/dt = \rho[b^G(t) + r(t)] + [1/P(t)]dD(t)/dt - g(t) - \tau(t). \quad (3)$$

Changes in the economy's money supply, $M^s(t)$, result from changes in the central bank's foreign or domestic assets. If the world price level P^* is constant (so that proportional changes in $P(t)$ equal proportional changes in the exchange rate), then the central-bank balance-sheet identity implies $dM^S(t)/dt = P(t)[dr(t)/dt] + dD(t)/dt$. Let $m(t)$ denote the private sector's desired real money balances and $\pi(t)$ the home inflation rate. On the assumption that the money market is continuously in equilibrium, $m(t) = M^S(t)/P(t)$ and equation (3) becomes

$$d[b^G(t) + r(t)]/dt = \rho[b^G(t) + r(t)] + \pi(t)m(t) + [dm(t)/dt] - g(t) - \tau(t). \quad (4)$$

Integrate (4) forward from $t = 0$ and impose the condition $\lim_{t\to\infty}\exp(-\rho t)[b^G(t) + r(t)] \geq 0$, which restricts the government to borrowing paths such that the public debt is assymptotically paid off. The result is the intertemporal budget constraint of the government,

$$\int_0^\infty [g(t) + \tau(t)]\exp(-\rho t)dt \leq \int_0^\infty [\pi(t)m(t) + dm(t)/dt]\exp(-\rho t)dt + b^G(0) + r(0).$$

The inequality states that the present value of net government outlays must be less than the present value of the seigniorage from money creation plus the government's initial asset position.The latter quantity, in turn, equals central-bank foreign reserves less the public debt. For a world of perfect capital mobility, the constraint makes clear that it is the government's overall asset position that is relevant for assessing solvency. The level of foreign reserves $r(0)$ has little significance in itself. As noted earlier, the central bank can increase its reserves by selling other government assets (thus reducing $b^G(0)$ by an amount equal to the rise in reserves). The transaction requires no change in the path of planned government outlays, $g(t) + \tau(t)$.

Consider next the private sector. Let $b(t)$ denote net private real bond holdings and $k(t)$ real capital holdings. (By assumption capital's real price equals unity.) Foreigners do not hold domestic money or capital, although the analysis could easily be modified to account for these possibilities. Given an inelastic labour supply normalized at unity and a neoclassical production function $x[k(t), t]$, private-sector assets obey the equation

$$d[b(t) + k(t) + m(t)]/dt = x[k(t), t] + \rho b(t) + \tau(t) - c(t) - \pi(t)m(t). \quad (5)$$

Define investment $i(t)$ as $dk(t)/dt$. The sum of (4) and (5) is

$$d[b(t) + b^G(t) + r(t)]/dt = x[k(t), t] + \rho[b(t) + b^G(t) + r(t)] - c(t) - i(t) - g(t).$$

The sum $b(t) + b^G(t) + r(t)$ will be denoted by $f(t)$: $f(t)$ equals the economy's overall net claims on the rest of the world. Integrated forward and combined with the condition $\lim_{t\to\infty}\exp(-\rho t) f(t) \geq 0$, the above equation implies the economy's overall intertemporal budget constraint,

$$\int_0^\infty \{c(t) + i(t) + g(t) - x[k(t), t]\}\exp(-\rho t)dt \leq f(0). \quad (6)$$

(The same constraint is relevant when the private sector is prohibited from transacting in the world capital market, but the paths of consumption, investment, and output would generally change if such a prohibition were imposed.)

Inequality (6) states that the present value of the economy's expenditures cannot exceed the present value of output plus initial net external assets. Alternatively, (6) constrains the present value of the economy's trade balance deficits to its initial foreign asset stock. The initial foreign asset stock thus limits the economy's ability to maintain absorption levels in excess of output.

An implication of the analysis is that the most appropriate indicator of flow disequilibrium in external transactions is the change in the economy's overall external assets – the current account. A surplus in the balance of payments may indicate low domestic credit expansion or growing domestic money demand; but when the government has unlimited access to the world capital market, a growing stock of foreign reserves is, in itself, neither a necessary nor a sufficient condition for a sound external position.

The important consequences of current-account flows do not imply that external balance and current-account balance are the same. In analogy with the idea of a high-employment government budget surplus, external balance could be defined roughly as a current account that maintains the highest possible steady consumption level consistent with the economy's expected intertemporal budget constraint. (A more exact definition would require a more explicit treatment of the preferences of households and the government.) Temporary unfavourable movements in output, world interest rates, or the terms of trade are appropriately offset by temporary current-account deficits, while temporary surpluses are an appropriate response to temporary favourable shocks. External balance in the face of a permanent shock, however, generally requires a rapid adjustment to current-account balance. Similarly increases in the productivity of investment can justify a current-account deficit that is fully consistent with external balance in a long-run sense. In terms of equation (6), a technological innovation implying a gradual upward shift of the production function $x[k(t), t]$ generates higher levels of consumption and investment, and thus an initial current-account deficit. The ability to borrow abroad prevents the sharp rise in the interest rate that would occur initially in a closed economy; a higher investment level than under intertemporal autarky is supported by the foreign capital inflow. As productivity growth returns to normal, investment falls and current-account balance is restored with consumption and output at permanently higher levels.

These points can be made graphically in terms of a two-period Fisherian model (see Figure 3). The axes measure amounts of the two goods available, present and future consumption, and the indifference curves show preferences over those goods. Investment opportunities are described by the production-possibilities frontier, which indicates the amount of future consumption obtained from a given input of present consumption. With the opportunity to borrow abroad at an interest rate ρ, the economy chooses to invest at point A and consume at point B, both of these points lying on the economy's budget line, which has slope $-(1+\rho)$. Given preferences and technology, it is optimal for this economy to run a first-period current-account deficit equal to the horizontal distance between B and A; in period two, the country runs a surplus to repay its earlier borrowing. External balance thus entails an initial current-account deficit for the country shown, but surpluses for countries whose autarky interest rates are less than the equilibrium world rate ρ. The model is a parable of the development process.

Figure 3

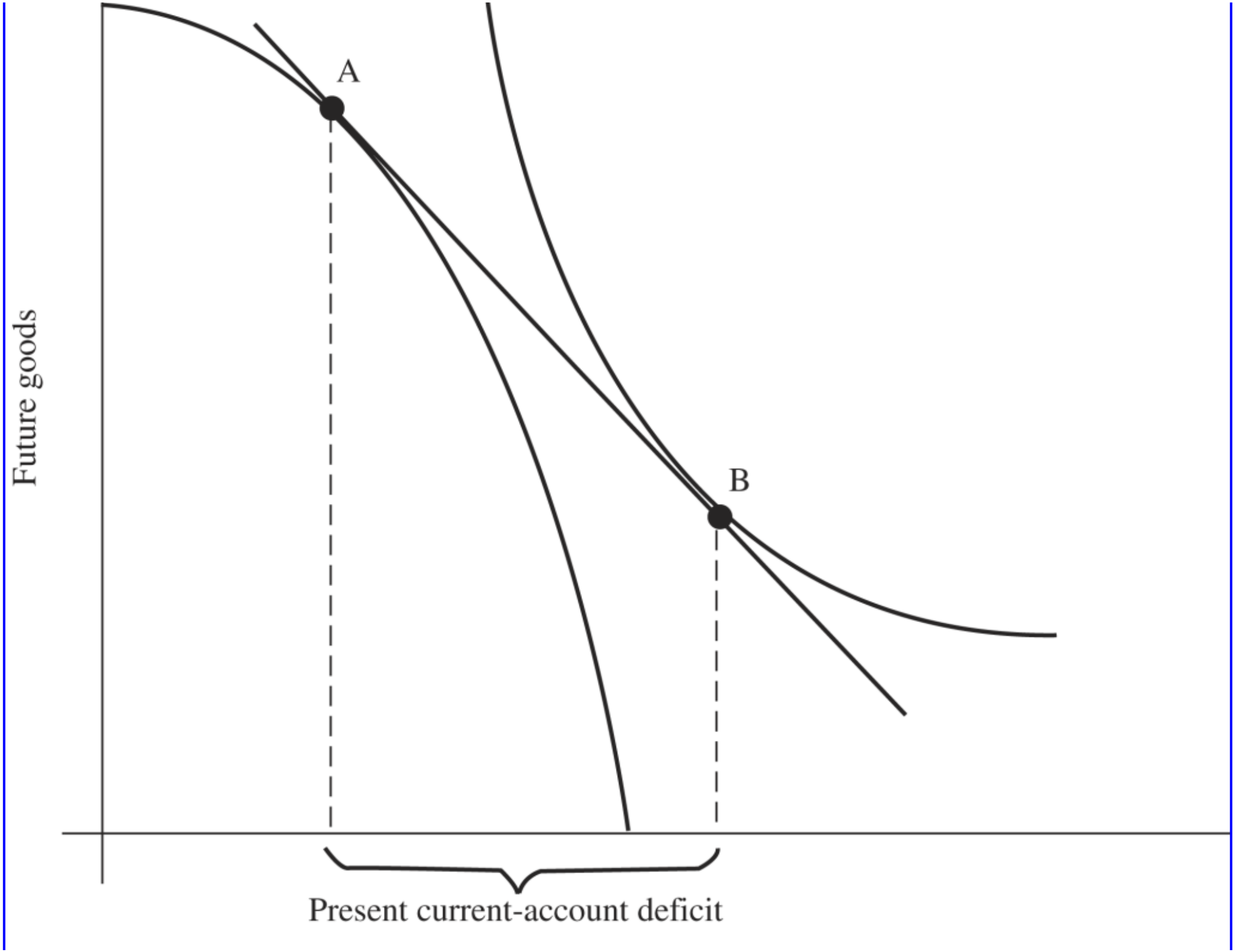
A
B
Future goods
Present current-account deficit

Present goods

When distortions in the economy cause the actual current account to diverge from its optimal level, governments may find it appropriate to adopt policies, such as taxes or subsidies on capital movement, that move the economy closer to the ideal external balance. Policies that operate directly on the distortions in question (if these can be identified) will, as usual, be best. Interesting problems arise when the countries being analysed are large enough that their governments can affect world real interest rates (and other world prices) through their actions. In this situation, the normative guidelines offered by the above approach are not directly applicable to policy analysis, and governments instead condition their actions on the conjectured responses of other governments. A Nash–Cournot equilibrium, in which each government maximizes over policy settings taking as *given* the policies of other governments, will in general be Pareto-inefficient from a global viewpoint. When governments recognize their policy interdependence, welfare in each country can be improved though policy cooperation. The practical difficulty lies in the negotiation process through which all parties agree to choose a particular point on the world contract curve.

Sovereign borrowing and credit constraints

The intertemporal analysis of external balance sketched above assumes a world in which individuals, or at least governments, can borrow unlimited amounts in the world capital market, subject only to their intertemporal budget constraints. Individual and sovereign borrowers alike, however, often appear to face binding credit constraints as a result of nonrepayment risk. After the early 1980s, the extreme difficulty for many industrializing countries of tapping world credit markets focused attention on how countries' borrowing possibilities are affected by the possibility of sovereign debt default. The problem is a central one because most developing-country debts are either contracted directly by government agencies or are government-guaranteed.

Eaton and Gersovitz (1981) presented an early explicit analysis of the sovereign repudiation problem in an international setting. Claims on sovereign debtors are usually not legally enforceable, so the analysis of sovereign default cannot be conducted in terms of bankruptcy laws that govern cases of individual default. Eaton and Gersovitz hypothesized that a sovereign debtor defaults whenever the present discounted benefit of doing so exceeds the present discounted cost. Potential lenders, understanding the debtor's decision rule, will never lend so much that a sure incentive to default is created. Accordingly, sovereign borrowers may find themselves credit-rationed, unable to borrow as much as would normally be optimal at the interest rate quoted by lenders.

There are several potential costs of sovereign default. A defaulting country's external assets, such as foreign reserves or goods in transit, can be seized. The country could, in addition, find itself unable to borrow in the future in response to unexpected changes in its income or technology. Continued participation in the world trade and payments system might become infeasible altogether.

This 'willingness to pay' hypothesis has radical implications for the analysis of external balance. The borrowing country shown in Figure 3, for example, would repudiate its foreign debt if that action were costless, thus avoiding the resource transfer it would otherwise have to make in the second period. As a result, period-one borrowing would take place at a country-specific interest rate reflecting the probability of default, with the extent of borrowing limited by the market's estimate of default costs. At interest rates so high that default was certain, no lending at all would occur.

The analysis of external balance becomes much more complex in such a setting. Not only is the allowable current-account deficit more severely circumscribed; in addition, the policymaker must consider how various policy actions will affect the costs of default and hence the availability of foreign credit. Trade liberalization measures that move the economy away from an autarkic production allocation increase the cost of default by making the economy more vulnerable to disruption of its foreign trade. Such measures will therefore ease international credit constraints at the same time as they improve the static allocation of national resources. Conversely, trade restrictions aimed at improving the current account may well reduce a country's creditworthiness.

The traditional balance-of-payments target has a rationale if the government believes that foreign credit lines may disappear unexpectedly. There is then a case for holding precautionary reserves to finance current-account deficits that may become necessary at times when credit happens to be tight or non-existent. The same purpose would be served, however, if foreign assets held by government agencies other than the central bank were run down at such times.

Internal and external balance may be irreconcilable for countries that seek to continue external debt service in the face of severe limitations on foreign borrowing. After the early 1980s, many developing countries were able to obtain private external finance only through 'forced' bank lending orchestrated by the IMF and central banks. Measures to reduce current-account deficits in line with the external funds available (and in line with IMF stabilization targets) pushed many economies into deep recession. As of this writing, it is unclear how long it will remain politically feasible for debtor governments to downplay internal-balance goals in order to continue avoiding default. There are increasingly frequent calls for some form of debt relief. Such proposals amount to the *ex post* indexation of debt contracts to adverse contingencies that were not entirely under the debtors' control.

The debt crisis of the 1980s has raised deep and consequential questions about the types of assets traded between developed and developing countries. Before the debt crisis, the typical loan contract between banks and developing-country borrowers was indexed only to the London Inter-Bank Offered Rate, and not to other factors that might alter the borrower's ability to repay. Trade between developed and developing countries in a wider spectrum of state-contingent assets would improve the international allocation of risk, and thus help to avoid future debt crises. A greater share for equity in settling current-account imbalances is one possible step in this direction. Such reforms would not eliminate the sovereign-default problem entirely, nor would they eliminate the moral-hazard problem emphasized by critics of debt-relief proposals. The possibility of a widespread and

synchronized default could be sharply reduced, however, under innovative external financing arrangments.

The structure of international financial intermediation also has implications for the mutual adjustment process of industrialized countries. Current-account imbalances are only one avenue through which countries can maintain long-run consumption levels in the face of real income fluctuations or changes in investment productivity. Similar consumption-smoothing can be obtained with smaller current-account imbalances if there is a greater degree of international portfolio diversification. Lucas (1982), for example, models a world of two exchange economies with perfect international risk sharing in which consumption levels can be perfectly correlated internationally even though current-account imbalances never take place. The problem of external balance therefore never arises in Lucas's idealized setting. In reality, the extent of international portfolio diversification seem to be much smaller than plausible financial models of an integrated world capital market would predict. Why this should be so is a major empirical puzzle, and a problem for policy as well.

See Also

- purchasing power parity
- specie-flow mechanism

Recent general perspectives on international finance may be found in International Monetary Fund, new open economy macroeconomics and World Bank, *as well as entries on various specific aspects of international economics*.

Bibliography

Dornbusch, R. 1973. Currency depreciation, hoarding, and relative prices. *Journal of Political Economy* 81, 893–915.

Eaton, J. and Gersovitz, M. 1981. Debt with potential repudiation: theoretical and empirical aspects. *Review of Economic Studies* 48, 289–309.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Frenkel, J.A. and Johnson, H.G., eds. 1976. *The Monetary Approach to the Balance of Payments*. Toronto: University of Toronto Press.

Friedman, M. 1953. The case for flexible exchange rates. In M. Friedman, *Essays in Positive Economics*, Chicago: University of Chicago Press.

Hume, D. 1752. Of the balance of trade. In D. Hume, *Essays Moral, Political and Literary*. London: Longmans Green, 1898.

Johnson, H.G. 1969. The case for flexible exchange rates, 1969. *Federal Reserve Bank of St. Louis Monthly Review* 51(6), 12–24.

Keynes, J.M. 1936. *The General Theory of Employment Interest and Money*. London: Macmillan.

Lucas, R.E., Jr. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A.H. Meltzer. Amsterdam: North-Holland.

Lucas, R.E., Jr. 1982. Interest rates and currency prices in a two-country world. *Journal of Monetary Economics* 10, 335–59.

Meade, J.E. 1951. *The Balance of Payments*. London: Oxford University Press.

Metzler, L.A. 1948. The theory of international trade. In *A Survey of Contemporary Economics*, ed. H.S. Ellis. Philadelphia: Blakiston.

Metzler, L.A. 1960. The process of international adjustment under conditions of full employment: a Keynesian view. In *Readings in International Economics*, ed. R.E. Caves and H. G. Johnson. Homewood: Irwin, 1968.

Mundell, R.A. 1961. The international disequilibrium system. In R.A. Mundell, *International Economics*. New York: Macmillan, 1968.

Mundell, R.A. 1968. The nature of policy choices. In R.A. Mundell, *International Economics*. New York: Macmillan.

Nurkse, R. 1944. *International Currency Experience: Lessons of the Inter-war Period*. Geneva: League of Nations.

Obstfeld, M. 1985. Floating exchange rates: experience and prospects. *Brookings Papers on Economic Activity* 1985(2), 369–464.

Obstfeld, M. and Stockman, A.C. 1985. Exchange-rate dynamics. In *Handbook of International Economics*, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.

Ricardo, D. 1817. *The Principles of Political Economy and Taxation*. London: J.M. Dent & Sons, 1911.

Samuelson, P.A. 1971. On the trail of conventional beliefs about the transfer problem. In *Trade, Balance of Payments, and Growth*, ed. J.N. Bhagwati et al. Amsterdam: North-Holland.

Triffin, R. 1960. *Gold and the Dollar Crisis*. New Haven: Yale University Press.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper & Brothers.

How to cite this article

Obstfeld, Maurice. "international finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000169> doi:10.1057/9780230226203.0828

The New Palgrave Dictionary of Economics Online

international financial institutions (IFIs)

John Toye

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The International Monetary Fund was established to manage international payments post-Second World War. The gold exchange standard re-established current account convertibility in the industrialized nations and oversaw rapid growth of international trade. After that standard collapsed in 1971 the IMF ran stabilization programmes for developing countries, with mixed success. The World Bank was set up to provide medium-term loans at concessional interest rates for (post-war) reconstruction and to develop capital-poor areas. In 1979 it initiated programme lending with conditions to promote economic adjustment. Conditionality has been under-enforced but increasingly loans go to countries that show commitment to liberal economic reforms.

Keywords

Baker Plan; Bretton Woods agreement; capital account controls; conditionality; convertibility; fixed exchange rates; floating exchange rates; gold exchange standard; international capital flows; International Development Agency (IDA); international financial institutions; International Monetary Fund; Keynes, J.M.; Mexican debt crisis 1982; Poverty Reduction Strategy Papers; Special Drawing Rights (SDRs); World Bank

Article

The two major international financial institutions, the International Monetary Fund (IMF) and the World Bank, were designed on an Anglo-American plan, negotiated by John Maynard Keynes and Harry Dexter White. After gaining wider approval at the Bretton Woods conference in July 1944, they were set up in order to introduce new elements of multilateral regulation into the working of the international economy (Skidelsky, 2000)

The International Monetary Fund

The IMF, established in 1947 to manage international payments in the economic chaos that followed the Second World War, had four Charter objectives: to restore a system of multilateral payments for current transactions between its members; to minimize disequilbrium in the members' international balances of payments; to promote exchange stability; and 'to facilitate the expansion and balanced growth of international trade, and to contribute thereby to the promotion and maintenance of high levels of employment and real income'. In promoting all of these objectives, the Fund originally acted as the umpire of a set of rules of international monetary behaviour.

Originally, the IMF managed a system of fixed, but adjustable, exchange rates against the US dollar, which itself was pegged to gold. In order to keep exchange rate fluctuations within set limits, each member country – and the membership then (59) was much smaller than it was in 2005 (over 180) – paid into the Fund a capital sum, determined according to its importance in world trade, and was given a borrowing 'quota' related to its capital. Voting power in the organization is related to the size of this capital. In balance of payments difficulties, members were permitted to borrow from the Fund and repay over the following two or three years. Thus the Fund acted as a bank, but the scale of the 'banking' operation was initially small. Between 1947 and 1955, 14 out of 59 members made drawings, at an annual rate of $46•m. This equalled 0.06 per cent of world imports. In 1990–8, when 78 out of 182 members made drawings, the rate was $13.4bn, or 0.29 per cent of world imports.

The gold exchange standard succeeded in re-establishing current account convertibility in the industrialized nations, while permitting countries to maintain capital account controls. The IMF had less success in shortening and reducing the severity of balance of payments disequilibria (Killick, 1985).

Nevertheless, under this system, international trade did grow rapidly, and employment and real income also grew faster than subsequently. Fear of liquidity shortage led the IMF in 1967 to create Special Drawing Rights (SDRs) – the First Amendment of the Fund's Articles of Agreement – but they came too late to save the anchor of the system, the $35 an ounce fixed parity of the official price of gold. The ratio of US gold reserves to its liquid liabilities had fallen from 2.73 in 1950 to 0.41 by 1968. Once the private market gold price rose above the official price, dollar-gold convertibility was suspended de facto, and officially abandoned in 1971. The collapse of the gold exchange system was thus due to an inherent design flaw, and not to any particular failures of the IMF.

The Fund soon ceased to be a banker to OECD countries, and began to cast around for a new role in the developing countries. However, this changed the Fund from an institution of collective action for industrial countries into their instrument for disciplining others. The Second Amendment to the IMF Articles in 1978 allowed all forms of national exchange-rate mechanism, except pegging to gold. Many larger economies chose to float their currency. Many smaller economies chose to peg their exchange rate to other currencies or baskets of currencies. Systemic international economic coordination was replaced by G7 meetings that tried to 'talk down' or 'talk up' particular key currencies.

Under the gold exchange standard, developing countries had been of little interest to the IMF. Many had never been properly integrated into the system, although in Peru and Paraguay the Fund did pioneer policy-conditioned lending. From the early 1960s, under UN pressure, the Fund developed additional 'banking' facilities relevant to the needs of developing countries – for example, the Compensatory Financing Facility in 1963 and the Extended Fund Facility (EFF) in 1974, which provided medium-term finance, beyond the limits of normal lending, to support agreed stabilization programmes requiring structural adjustment.

The Mexican debt crisis of 1982 was a turning point in the history of the Fund. Following the Baker Plan of 1985, the US Administration recruited the Fund, along with the World Bank, to be its managers at one remove of the prolonged debt crisis that for some years threatened the survival of major Western banks. The capital available to both institutions was increased. Building on the EFF, new longer-term lending facilities were created to channel credit to indebted developing countries. In 1986 the Structural Adjustment Facility was set up, and in 1987 the Extended Structural Adjustment Facility (ESAF), to provide loans to low-income countries suffering protracted balance of payments problems at 0.5 per cent interest over five and a half to ten years. Policy conditionality is strong under ESAF loans, and is specified in the Poverty Reduction Strategy Papers of the borrowing country.

However, IMF stabilization programmes frequently broke down before completion. Between 1979 and 1993, 53 per cent of 305 Fund programmes were uncompleted, often because of inadequate financing (Killick, 1995, pp. 58–65). If sustained, they improved the current account and the overall balance of payments, and slowed inflation, but at the cost of a short-term reduction in growth. The Fund came under particularly fierce criticism for its handling of the Asian financial crisis of 1997–8 (Stiglitz, 2002). It has since introduced reforms to improve its own transparency and member countries' data reporting standards.

The World Bank Group

The International Bank for Reconstruction and Development (IBRD) was established in 1946 in order to provide medium-term loans at less than commercial interest rates to governments for (post-war) reconstruction and for the development of capital-poor areas. Since then other parts of what is now called the World Bank Group have been added – the International Finance Corporation (IFC), set up in 1956 for lending to the private sector, the International Centre for the Settlement of Investment Disputes (1966) and the Multilateral Investment Guarantee Agency (MIGA) (1988). However, the most significant addition was the International Development Agency (IDA) in 1960, to provide long-term, highly concessional loans to the poorest countries.

Having largely missed out on post-war reconstruction lending, the Bank focused on project lending for economic development. Its procedure was to borrow on the developed country capital markets and re-lend (plus a small margin) for specific investment projects in developing countries. In the early years, this was a slow process, originally concerned with large physical infrastructure schemes, such as dams and electricity generation. After IDA gave the Group a development agency function, the composition of Bank investments began to change, gradually including agricultural and urban redevelopment projects. The criterion of project success was the *ex post* rate of return on each project. In 1973, a semi-independent Operations Evaluation Department was established to calculate this. The Bank's participation almost certainly produced a better quality of project than would have occurred in its absence. However, if fungibility exists, the economic effect of the investment cannot be measured by its *ex post* rate of return. Although fungibility need not concern a development bank, whose chief aim is to recover its loans, it should worry a multilateral aid agency funded by public capital, whose main objective is to promote the sound development of the borrower's economy. To ensure that, projects need to be appraised as part of a comprehensive development plan.

In the 1970s, when the economies of developing countries were disturbed by substantial economic shocks, the World Bank decided that the success of their individual loan projects, as measured by their

ex post rates of return, was being affected negatively by their broader economic environment (rising oil price, high inflation, fixed nominal exchange rates, import restrictions, and so on). In 1979, the Bank initiated programme lending, previously regarded as an unsound banking practice. The new types of loans, structural (SAL) and sectoral (SECAL) adjustment lending, provided rapidly disbursing foreign exchange on condition that the borrowing government undertook economic policy changes, either economy-wide or sectorally.

Programme lending with policy conditions attached provided the instrument that the Bank could bring to the task of co-managing the 1980s debt crisis with the Fund. A Fund–Bank ‘concordat’ in 1989 established effective (though not formal) cross-conditionality of Fund and Bank loans. Bank adjustment lending became conditional on a pre-existing Fund programme, and a statement of economic policy for the borrowing country had to be agreed by both institutions – entitled the Poverty Reduction Strategy Paper.

The evaluation of the effects of programme lending is more difficult and controversial. Governments of developing countries have been reluctant to comply with some of the conditions for policy change laid down in the loan agreements. This is often described as a result of their ‘lack of ownership’ of the economic reform process. The Bank itself faces incentives that make it unlikely that it will react to non-compliance consistently with a discontinuation of funding (Mosley, Harrigan and Toye, 1995). Thus the evidence suggests that the Bank's loan conditionality is a weak instrument for inducing policy change (see Ferreira and Keely, 2000). At the start of the 21st century, the Bank was moving towards a lending strategy of selectivity, in which future loans are directed increasingly to countries that have already demonstrated their zeal for neoliberal economic reform.

The Bank has been criticized on the grounds that private flows to developing countries can do the job instead (Krueger, 1998). In 1970, IBRD net lending was about ten per cent of net private flows. In 1996, this share had fallen to 0.7 per cent. In 25 years, private flows had increased 40-fold, while IBRD flows had increased threefold in nominal terms. The original justification of IBRD loans in terms of imperfect private capital markets seems weak in the light of these figures, although private finance is very concentrated geographically and the Asian crisis showed how short term and volatile private money can be.

Apart from lending, the Bank undertakes many other activities. It conducts what is probably the largest single publication programme on development issues in the world. This includes its own research across the field of development problems, published in two house journals, flagship reports like the annual *World Development Report*, a host of monographs and a multitude of Working Papers. The Bank has also become a major provider of statistical data, including regular published series and data from household and firm surveys. It regards itself as a ‘knowledge agency’.

See Also

- foreign aid
- Keynes, John Maynard
- third world debt

Bibliography

Akyuz, Y. 2002. *Reforming the Global Financial Architecture: Issues and Proposals*. Geneva: UNCTAD.

Boughton, J. 2001. *Silent Revolution: The International Monetary Fund, 1979–89*. Washington, DC: IMF.

Buira, A., ed. 2005. *The IMF and the World Bank at Sixty*. London: Anthem Press.

Ferreira, F. and Keely, L. 2000. The World Bank and structural adjustment: lessons from the 1980s. In *The World Bank: Structure and Policies*, ed. C. Gilbert and D. Vines. Cambridge: Cambridge University Press.

Kapur, D., Lewis, J. and Webb, R. 1997. *The World Bank: Its First Half Century. Volume 1: History*. Washington, DC: Brookings Institution.

Killick, T. 1985. *The Quest for Economic Stabilization: The IMF and the Third World*. Aldershot: Gower.

Killick, T. 1995. *IMF Programmes in Developing Countries: Design and Impact*. London: Routledge.

Krueger, A. 1998. Whither the World Bank and the IMF? *Journal of Economic Literature* 36, 1983–2020.

Mallaby, S. 2005. Saving the World Bank. *Foreign Affairs* 84(3), 75–85.

Mosley, P., Harrigan, J. and Toye, J. 1995. *Aid and Power: The World Bank and Policy-based Lending*, vol. 1, 2nd edn. London: Routledge.

Ritzen, J. 2005. *A Chance for the World Bank*. London: Anthem Press.

Skidelsky, R. 2000. *John Maynard Keynes: Fighting for Britain 1937–1946*. London: Macmillan.

Stiglitz, J. 2002. *Globalization and Its Discontents*. London: Allen Lane.

Toye, J. and Toye, R. 2004. *The UN and Global Political Economy: Trade, Finance and Development*. Bloomington: Indiana University Press.

Woods, N. 2001. Making the IMF and the World Bank more accountable. *International Affairs* 77, 83–100.

How to cite this article

Toye, John. "international financial institutions (IFIs)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000244> doi:10.1057/9780230226203.0829

The New Palgrave Dictionary of Economics Online

international migration

George J. Borjas
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The resurgence of large-scale immigration in many countries has stimulated a great deal of research on many aspects of the economics of immigration. A key insight of economic theory is that the impact of immigration depends on how the skills of immigrants compare with those of natives in the host country. This article examines the ideas and models that are typically used to analyse flows of persons across countries, and illustrates how this framework increased our understanding of the determinants of the direction, size, and skill composition of immigrant flows, and of the consequences of those flows on economic outcomes.

Keywords

elasticity of complementarity; elasticity of substitution; immigrant self-selection; immigrant skills; immigration and the welfare state; immigration surplus; international migration; labour flows; labour markets; migration costs; National Academy of Sciences (US); redistribution of income; Roy model; social insurance

Article

There was a significant resurgence in international migration in the last three decades of the 20th century. By the end of the century, about 175 million persons – or almost three per cent of the world's population – resided in a country where they were not born. Nearly 9 per cent of the population in Germany, 11 per cent in France or Sweden, 12 per cent in the United States, 19 per cent in Canada, 23 per cent in New Zealand, and 25 per cent in Switzerland is foreign-born (United Nations, 2002). These sizable labour flows altered economic opportunities for workers in both sending and receiving countries, and generated a great deal of debate over the economic impact of immigration and over the types of immigration policies that host countries should pursue.

Labour flows across labour markets – whether within or across countries – play a central role in any discussion of labour market equilibrium. These labour flows help markets attain a more efficient allocation of resources. This article surveys the economic analysis of immigration. In particular, it investigates the determinants of the immigration decision and the impact of that decision on economic conditions in the receiving country.

The discussion emphasizes the ideas and models that are used to analyse flows of persons across countries, and examines the implications of these models for empirical research and for our understanding of the labour market effects of immigration. A key insight of economic theory is that the economic impact of immigration depends on how the skills of immigrants compare with those of natives in the host country. As a result, much of the research effort in the immigration literature has been devoted to: (*a*) understanding the factors that determine the relative skills of the immigrant flow; (*b*) measuring the relative skills of immigrants in the host country; and (*c*) evaluating how the skill composition of the immigrant influx affects economic outcomes.

Because the discussion focuses on the impact of immigration on economic conditions in the host country, the analysis ignores a number of equally important issues, in terms of both their theoretical implications and their empirical significance. Immigration, after all, alters economic opportunities not only in the host country, but in the source country as well. Few studies, however, investigate what happens to economic opportunities in a source country when a selected subsample of its population moves elsewhere. Similarly, the discussion focuses on the economic impact of immigrants, and ignores the long-run impact of the children and grandchildren of immigrants on the host country.

The impact of immigration on a host country's labour market

Consider initially the simplest theoretical framework that can be used to understand how immigration alters the economic rewards accruing to various factors of production in a host country. Suppose the linear homogeneous aggregate production function in the country is given by $Q = f(K, L)$, where Q is output, K is capital, and L is labour. The workforce contains N native and M immigrant workers, and all workers are perfect substitutes in production $(L = N + M)$. Natives own the entire capital stock in the host country and initially the supply of capital is perfectly inelastic. The supplies of both natives and immigrants are also perfectly inelastic. Finally, let the price of the output be the numeraire.
In a competitive equilibrium, each factor price equals the respective value of marginal product. The rental rate of capital in the pre-immigration equilibrium is $r_0 = f_K(K, N)$ and the price of labour is $w_0 = f_L(K, N)$. In the pre-immigration regime, national income accruing to natives, Q_N, is:

$$Q_N = r_0 K + w_0 N. \quad (1)$$

Figure 1 illustrates this initial equilibrium. Because the supply of capital is fixed, the area under the marginal product of labour curve (f_L) gives the value of the economy's total output. The national income accruing to natives Q_N is given by the trapezoid ABN0.

Figure 1
The immigration surplus

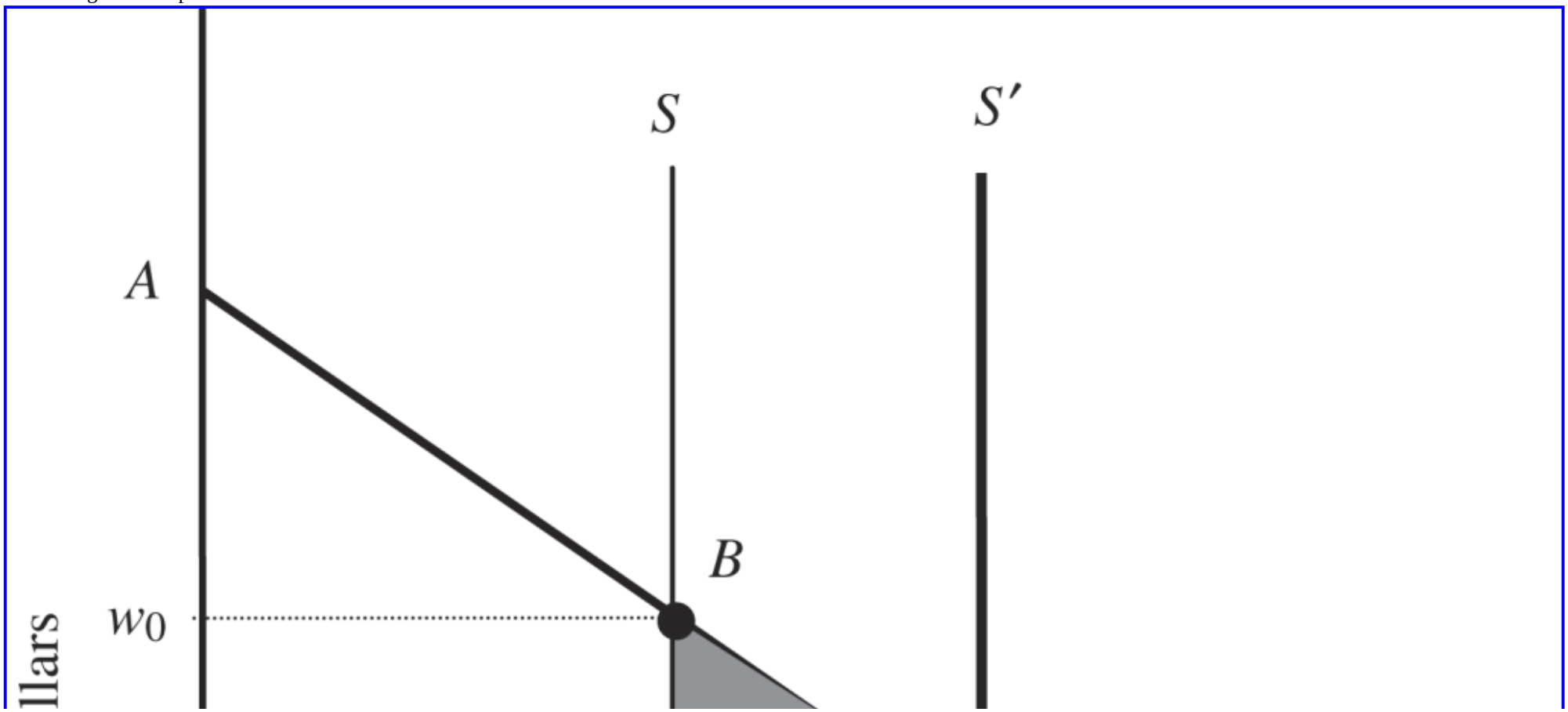

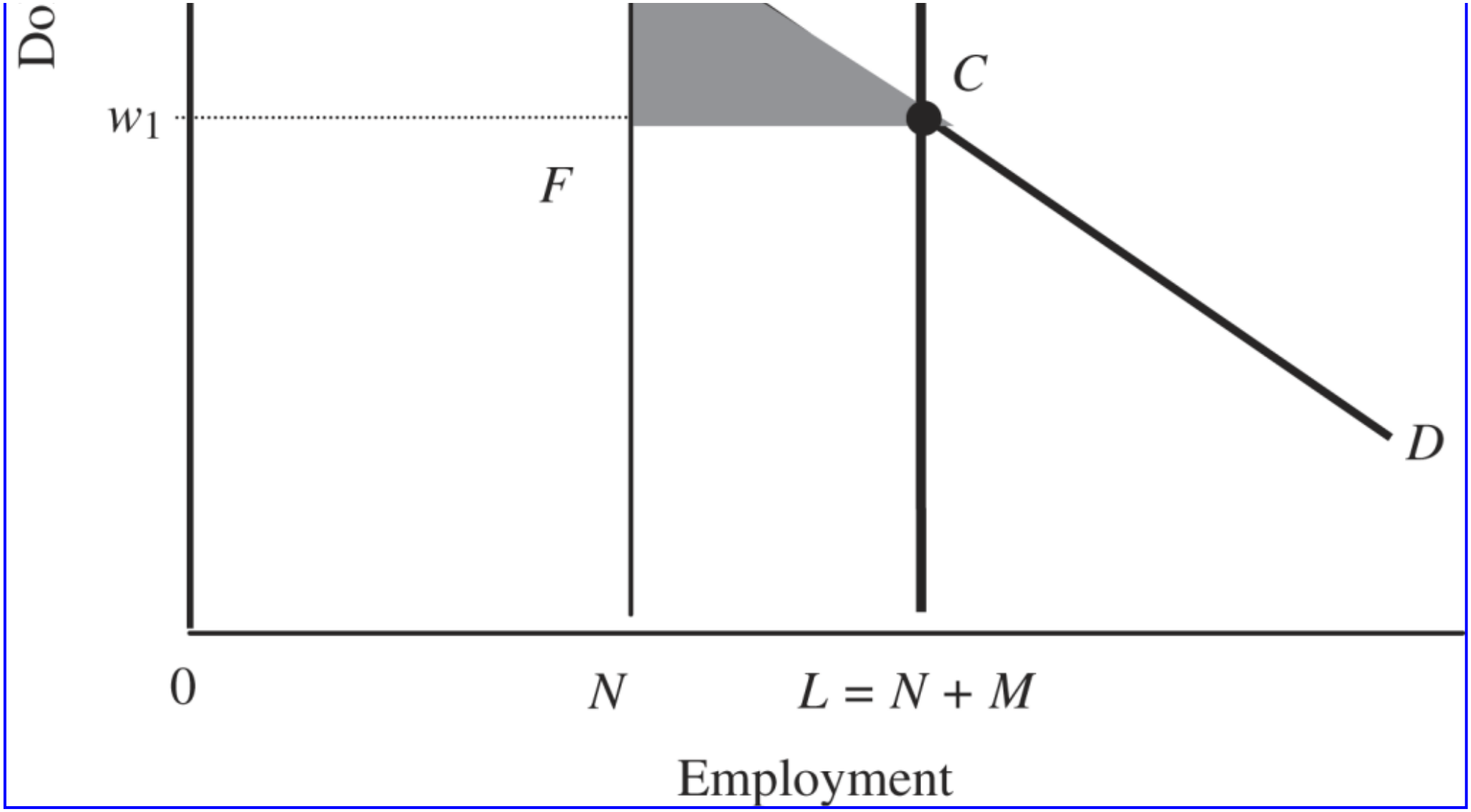

The entry of M immigrants shifts the supply curve and lowers the market wage to w_1. The area in the trapezoid ACL0 now gives national income. Immigrants receive part of the increase in national income as labour earnings (w_1M). The area in the triangle BCF gives the increase in national income that accrues to natives, or the 'immigration surplus'. If we use the approximation that $(w_0 - w_1) \approx (\partial w / \partial L) \times M$, the immigration surplus as a fraction of national income equals:

$$\frac{\Delta Q_N}{Q} = -\frac{1}{2}\alpha_L \varepsilon_{LL} m^2, \quad (2)$$

where α_L is labour's share of national income $(\alpha_L = wL / Q)$; ε_{LL} is the elasticity of factor price for labor ($\varepsilon_{LL} = d\log w / d\log L$, with marginal cost held constant); and m is the fraction of the workforce that is foreign born $(m = M / L)$.

Equation (2) can be used to calculate how much a host country gains from immigration. In the United States, for example, the share of labour income is about 70 per cent, and the fraction of immigrants in the workforce was 13 per cent in 2000. Hamermesh's survey (1993, pp. 26–9) of the empirical evidence on labour demand suggests that the elasticity of factor price for labor may be around –.3. The US immigration surplus, therefore, is on the order of 0.2 per cent of GDP.

Immigration also redistributes income from labour to capital. As Figure 1 shows, native workers lose the area in the rectangle w_0BFw_1, and this quantity plus the immigration surplus accrues to capitalists. The net changes in the incomes of native workers and capitalists are approximately given by:

$$\left.\frac{\text{Change in native labour earnings}}{Q}\right|_{dK=0} = \alpha_L \varepsilon_{LL} m(1-m), \tag{3}$$

$$\left.\frac{\text{Change in income of capitalists}}{Q}\right|_{dK=0} = -\alpha_L \varepsilon_{LL} m\left(1-\frac{m}{2}\right). \tag{4}$$

Consider again the back-of-an-envelope calculation for the United States. If the elasticity of factor price is –.3, native-born workers lose about 2.4 per cent of GDP, while native-owned capital gains about 2.6 per cent of GDP. A small immigration surplus may disguise a sizable income transfer from workers to the users of immigrant labour.

The derivation of the surplus in (2) assumed that the host country's capital stock is fixed. Alternatively, suppose that the supply of capital is perfectly elastic at the world price $(dr = 0)$, so that in the long run the capital stock adjusts completely to the increased labour supply. Differentiating the marginal productivity condition $r = f_K(K, L)$ implies that the immigration-induced change in the capital stock is:

$$\left.\frac{dK}{dM}\right|_{dr=0} = -\frac{f_{KL}}{f_{KK}} > 0. \tag{5}$$

The derivative in (5) is positive because $f_{KL} > 0$ when the production function is linear homogeneous.

The elasticity of complementarity for any input pair *i* and *j* is defined by $c_{ij} = f_{ij}f / f_i f_j$. (The elasticity of complementarity is the dual of the elasticity of substitution. Hamermesh, 1993, ch. 2, presents a detailed discussion of the properties of the elasticity of complementarity.) The immigration-induced wage change is then given by:

$$\left.\frac{d\log w}{d\log M}\right|_{dr=0} = \frac{\alpha_L}{c_{KK}}(c_{KK}c_{LL} - c_{LK}^2)\, m. \tag{6}$$

The linear homogeneity of the production function implies that $(c_{KK}c_{LL} - c_{LK}^2) = 0$, so that the host country's wage is independent of immigration. The immigration-induced capital flow re-establishes the pre-immigration capital–labour ratio in the host country. Immigration does not alter the price of labour or the returns to capital, and natives neither gain nor lose from immigration. The long-run immigration surplus is zero.

The conclusion that immigration does not alter labour market conditions in the long run depends critically on the assumption of a homogeneous labour force. Suppose there are two types of workers in the host country's labour market, skilled (L_S) and unskilled (L_U). The linear homogeneous aggregate production function is:

$$Q = f(K, L_S L_U) = f[K, bN + \beta M, (1-b)N + (1-\beta)M], \tag{7}$$

where b and β denote the fraction of skilled workers among natives and immigrants, respectively. The price of each factor of production, r for capital and w_i (i=S, U) for labour, is determined by the respective marginal productivity condition. The assumption that r is fixed implies that the immigration-induced adjustment in the capital stock equals (see Borjas, 1999, pp. 1703–5):

$$\left.\frac{dK}{dM}\right|_{dr=0} = -\frac{[f_{KS}\beta + f_{KU}(1-\beta)]}{f_{KK}}. \tag{8}$$

We can determine the impact of immigration on the wage of skilled and unskilled workers by differentiating the respective marginal productivity conditions and by imposing the restriction in equation (8). The wage effects of immigration are:

$$\left.\frac{d\log w_S}{d\log M}\right|_{dr=0} = \frac{\alpha_S}{c_{KK}}[c_{SS}c_{KK} - c_{SK}^2] \times \frac{(\beta - b)}{p_S p_U}(1-m)m, \tag{9}$$

$$\left.\frac{d\log w_U}{d\log M}\right|_{dr=0} = \frac{-\alpha_U}{c_{KK}}[c_{UU}c_{KK} - c_{UK}^2] \times \frac{(\beta - b)}{p_S p_U}(1-m)m, \tag{10}$$

where α_i is the share of national income accruing to factor i; and p_S and p_U are the shares of the workforce that are skilled and unskilled, respectively.

The assumption that the isoquants between any pair of inputs in the production function $f(K, L_S, L_U)$ have the typical convex shape implies that $c_{11}c_{22} - c_{12}^2 > 0$. Equations (9) and (10) then reveal that the impact of immigration on the wage structure depends on how the skill distribution of immigrants compares with that of natives. If the two skill distributions

are equal $(\beta = b)$, immigration has no impact on the wage structure of the host country. If immigrants are relatively unskilled $(\beta < b)$, the unskilled wage declines and the skilled wage rises. If immigrants are relatively skilled $(\beta > b)$, the skilled wage declines and the unskilled wage rises. In the long run, therefore, immigration lowers the wage of substitutes and raises the wage of complements.

It can be shown that the immigration surplus as a fraction of national income is given by:

$$\left.\frac{\Delta Q_N}{Q}\right|_{dr=0} = \frac{-\alpha_S^2}{2c_{KK}}[c_{SS}c_{KK} - c_{SK}^2] \times \frac{(\beta - b)^2}{p_S^2 p_U^2}(1-m)^2 m^2. \tag{11}$$

The immigration surplus is zero if $\beta = b$, and positive if $\beta \neq b$. If immigrants had the same skill distribution as natives, the immigration-induced change in the capital stock implies that the wages of skilled and unskilled workers are unaffected by immigration. The gains arise only if immigrants differ from natives.

Some studies simulate this model to provide back-of-an-envelope calculations of the immigration surplus when there is heterogeneous labour (Borjas, 1995; Johnson, 1997). In the US context, the immigration surplus calculated in this more general setting is roughly of the same order of magnitude (less than 0.2 per cent of GDP) as that estimated from the simplest framework illustrated in Figure 1. The available evidence, therefore, suggests that the net measurable gains from immigration to the United States tend to be small.

Finally, it is worth emphasizing that any credible estimate of the economic benefits from immigration must rely on a theoretical framework that fully captures the various effects which inevitably arise as the impact of immigration ripples through the economy. Inevitably, different models of the economy will lead to different estimates of the economic benefits. Recent theoretical work by trade economists, for example, suggests that if one takes the Ricardian perspective that the United States provides superior economic opportunities for all factors of production – so that both capital and labour would get higher returns by migrating to the United States – immigration would actually *lower* the GDP accruing to natives substantially, by around 1.0 per cent of GDP (Davis and Weinstein, 2002). Therefore, the important point to draw from the existing evidence is that plausible models of the US economy indicate that, at best, the net gains from immigration for the native-born population are very small.

Estimating the labour market impact of immigration

As shown above, economic theory suggests that immigration into a particular labour market affects the wage structure by raising the wage of complementary workers and lowering the wage of substitutes. Almost all of the first-generation empirical studies in the literature define the labour market along a geographic dimension, such as metropolitan areas in the United States. Beginning with Grossman (1982), the typical study regresses a measure of native economic outcomes in the locality (or the change in that outcome) on the relative quantity of immigrants in that locality (or the change in the relative number). (Representative studies include Altonji and Card, 1991; Card, 1990; 2001; Pischke and Velling, 1997.) The regression coefficient is then interpreted as the impact of immigration on the native wage structure.

This approach has two well-known problems. First, immigrants may not be randomly distributed across labour markets. If immigrants endogenously cluster in areas that have done well over some time periods, this would produce a positive spurious correlation between immigration and area outcomes either in the cross-section or in the time series. Second, natives may respond to the entry of immigrants in a local labour market by moving their labour or capital to other localities until native wages and returns to capital are again equalized across areas. For example, a large immigrant flow arriving in California might well result in fewer workers moving to California, as well as a reallocation of capital from other states into California. Interregional comparisons of the wage of native workers might show little or no difference because the effects of immigration are diffused throughout the national economy.

In view of these potential problems it is not too surprising that the region-based empirical literature has produced a confusing array of results (see the survey in Friedberg and Hunt, 1995). Nevertheless, there is a tendency for the estimated cross-region correlations to cluster around zero, creating the conventional wisdom that immigrants have little impact on the labour market opportunities of native workers. It would seem, therefore, that a fundamental implication of the competitive model of the labour market – that supply shocks alter the wage structure – is soundly rejected by the data.

Because local labour markets adjust to immigration, recent research emphasizes that the labour market impact of immigration may be measurable only at the national level. Borjas (2003) used this insight to derive an estimable framework that can be used to measure the national labour market effects of immigration by linking the evolution of the wage structure in the host country to changes in immigration. As an illustration, suppose that the national workforce in the host country is composed of skill groups defined in terms of both educational attainment and work experience. The aggregate production function at time *t* is:

$$Q_t = \left[\lambda_{Kt} K_t^{\nu} + \lambda_{Lt} L_t^{\nu}\right]^{1/\nu}, \tag{12}$$

where Q is output, K is capital, L denotes the aggregate labour input; and $\nu = 1 - 1/\sigma_{KL}$, with σ_{KL} being the elasticity of substitution between capital and labour $(-\infty < \nu \leq 1)$. The vector λ gives technology parameters that shift the production frontier, with $\lambda_{Kt} + \lambda_{Lt} = 1$. The aggregate L_t incorporates the contributions of workers who differ in both education and experience. Let:

$$L_t = \left[\sum_i \theta_{it} L_{it}^{\rho}\right]^{1/\rho} \tag{13}$$

where L_{it} gives the number of workers with education i at time t, and $\rho = 1 - 1/\sigma_E$, with σ_E being the elasticity of substitution across these education aggregates $(-\infty < \rho \leq 1)$. The θ_{it} give time-variant technology parameters that shift the relative productivity of education groups, with $\Sigma_i \theta_{it} = 1$. Finally, the supply of workers in each education group is itself given by an aggregation of the contribution of similarly educated workers with different experience. In particular,

$$L_{it} = \left[\sum_j \alpha_{ij} L_{ijt}^{\eta}\right]^{1/\eta}, \tag{14}$$

where L_{ijt} gives the number of workers in education group i and experience group j at time t (given by the sum of N_{ijt} native and M_{ijt} immigrant workers); and $\eta = 1 - 1/\sigma_X$, with σ_X being the elasticity of substitution across experience classes within an education group $(-\infty < \eta \leq 1)$. Equation (14) assumes that the technology coefficients α_{ij} are constant over time, with $\Sigma_j \alpha_{ij} = 1$.

The three elasticities of substitution that summarize all the economically relevant information in the production technology can be easily estimated using data on factor prices and quantities. The empirical application of this framework to US Census data from 1960 through 2000 in Borjas (2003) indicated that $\sigma_X = 3.5$, $\sigma_E = 1.3$, and $\sigma_{KL} = 1.0$. These elasticity estimates, combined with estimates of the size of the immigrant influx for each skill group, can be used to calculate the impact of immigration on the wage structure in a host country. Define the factor price elasticity giving the impact on the wage of factor y of an increase in the supply of factor z as:

$$\varepsilon_{yz} = \frac{d \log w_y}{d \log L_z}, \tag{15}$$

It is easy to show that the factor price elasticities depend on the income shares accruing to the various factors and on the three elasticities of substitution in the three-level constant elasticity of substitution (CES) framework. The marginal productivity condition for the typical worker in education group *s* and experience group *x* can be written as $w_{sx} = D(K, L_{ij})$, where L_{ij} is a vector indicating the number of workers in each of the education–experience cells. Suppose that the capital stock is constant. The short-run impact of immigration on the log wage of group (*s*, *x*) is:

$$\Delta \log w_{sx} = \sum_i \sum_j \varepsilon_{sx,ij} m_{ij}, \quad (16)$$

where m_{ij} gives the percentage change in labour supply due to immigration in skill cell (*i*, *j*). The available evidence suggests that the 1980–2000 immigrant influx into the United States, which represented an 11 per cent increase in labour supply, lowered the wage of the typical native worker by 3.7 per cent in the short run. As indicated in the earlier discussion, the adverse wage effects of immigration on the average worker are muted as the capital stock adjusts to the supply shock.

The self-selection of immigrants

As we have seen, the economic impact of immigration depends crucially on the differences in the skill distributions of immigrants and natives. Not surprisingly, a great deal of research effort has focused on the question of how immigrant skills compare with those of native workers. Perhaps the central finding of this literature is that immigrants are not a randomly selected sample of the population of the source countries.

It is instructive to consider a two-country model (Borjas, 1987). Residents of the source country (country 0) consider migrating to the host country (country 1). Assume the migration decision to be irreversible. Residents of the source country face the earnings distribution:

$$\log w_0 = \mu_0 + v_0, \quad (17)$$

where w_0 gives the wage in the source country; μ_0 gives the mean earnings in the source country; and the random variable v_0 measures deviations from mean earnings and is normally distributed with mean zero and variance σ_0^2. For simplicity, equation (17) omits the subscript that indexes a particular individual.

Suppose the potential earnings in the host country of emigrants from country 0 can be represented by:

$$\log w_1 = \mu_1 + v_1, \quad (18)$$

where μ_1 gives the mean earnings in the host country *for this particular population*, and the random variable v_1 is normally distributed with mean zero and variance σ_1^2. The correlation coefficient between v_0 and v_1 equals ρ_{01}.

The mean μ_1 does not necessarily equal the mean earnings of native workers in the host country. After all, the average worker in the source country might be more or less skilled than the average worker in the host country. It is convenient to assume that the average person in both countries is equally skilled (or, equivalently, that any differences in average

skills have been controlled for), so that μ_1 also gives the mean earnings of natives in the host country. This assumption helps isolate the impact of the selection process on the skill composition of the immigrant influx.

Equations (17) and (18) describe the earnings opportunities available to persons born in the source country. Assume that the migration decision is determined by a comparison of earnings opportunities across countries, net of migration costs. Define the index function:

$$I = \log\left(\frac{w_1}{w_0 + C}\right) \approx (\mu_1 - \mu_0 - \pi) + (v_1 - v_0), \tag{19}$$

where C gives migration costs, and π gives a 'time-equivalent' measure of these costs $(\pi = C / w_0)$. A person emigrates if $I > 0$, and remains in the source country otherwise. Migration is costly. Because the costs vary greatly among persons, and include direct costs, forgone earnings, and psychic costs, the sign of the correlation between costs and wages is unknowable. The distribution of the random variable π in the source country's population is:

$$\pi = \mu_\pi + v_\pi, \tag{20}$$

where μ_π is the mean level of migration costs in the population, and v_π is a normally distributed random variable with mean zero and variance σ_π^2. The correlation coefficients between v_π and (v_0, v_1) are given by $(\rho_{\pi 0}, \rho_{\pi 1})$. The probability that a person migrates to the host country can be written as:

$$P(z) = Pr[v > -(\mu_1 - \mu_0 - \mu_\pi)] = 1 - \Phi(z), \tag{21}$$

where $v = v_1 - v_0 - v_\pi$, $z = -(\mu_1 - \mu_0 - \mu_\pi) / \sigma_v$, and Φ is the standard normal distribution function.

It is easy to show that the emigration rate falls when the mean income in the source country rises, when the mean income in the host country falls, and when time-equivalent migration costs rise. Most studies in the literature on the internal migration of persons within a particular country focus on testing these theoretical predictions. The empirical evidence in these studies is generally supportive of the theory.

Although it is important to determine the size and direction of migration flows, it is equally important to determine *which* persons find it most worthwhile to migrate to the host country. This question lies at the heart of the Roy model (Roy, 1951). Consider the conditional means $E(\log w_0 | \mu_0, I > 0)$ and $E(\log w_1 | \mu_1, I > 0)$. These means give the average earnings in both the source and host countries for persons who migrate. Note that the conditional means hold μ_0 and μ_1 constant. The calculation effectively assumes that the migration flow is sufficiently small so that there are no feedback effects on the performance of immigrants (or natives) in the host country or on the performance of the 'stayers' in the source country. Because the random variables v_0, v_1, and v_π are jointly normally distributed, these conditional means are given by:

$$E(\log w_0 | \mu_0, I > 0) = \mu_0 + \left[\frac{\sigma_0 \sigma_1}{\sigma_v}\left(\rho_{01} - \frac{\sigma_0}{\sigma_1}\right) - \rho_{\pi 0}\frac{\sigma_\pi}{\sigma_1}\right]\lambda, \tag{22}$$

$$E(\log w_1 | \mu_1, I > 0) = \mu_1 + \left[\frac{\sigma_0 \sigma_1}{\sigma_v} \left(\frac{\sigma_1}{\sigma_0} - \rho_{01} \right) - \rho_{\pi 1} \frac{\sigma_\pi}{\sigma_0} \right] \lambda, \tag{23}$$

where $\lambda = \varphi(z) / (1 - \Phi(z))$, and ϕ is the density of the standard normal. It is easier to interpret the results in equations (22) and (23) by assuming that $\sigma_\pi = 0$, so that time-equivalent migration costs are constant. Let $Q_0 = E(v_0 | \mu_0, I > 0)$ and $Q_1 = E(v_1 | \mu_1, I > 0)$. The Roy model identifies three cases that summarize the skill differentials between immigrants and natives:

$$\begin{aligned} &Q_0 > 0 \text{ and } Q_1 > 0, \quad \text{if} \quad \rho_{01} > \frac{\sigma_0}{\sigma_1} \text{ and } \frac{\sigma_1}{\sigma_0} > 1, \\ &Q_0 < 0 \text{ and } Q_1 < 0, \quad \text{if} \quad \rho_{01} > \frac{\sigma_1}{\sigma_0} \text{ and } \frac{\sigma_0}{\sigma_1} > 1, \\ &Q_0 < 0 \text{ and } Q_1 > 0, \quad \text{if} \quad \rho_{01} < \min\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1} \right). \end{aligned} \tag{24}$$

Positive selection occurs when immigrants have above-average earnings in both the source and host countries ($Q_0 > 0$ and $Q_1 > 0$), and negative selection when immigrants have below-average earnings in both countries ($Q_0 < 0$ and $Q_1 < 0$). Equation (24) shows that either type of selection requires that skills be positively correlated across countries. The standard deviations σ_0 and σ_1 measure the 'price' of skills: the greater the rewards to skills, the larger the inequality in wages. Immigrants are then positively selected when the source country – *relative to the host country* – 'taxes' highly skilled workers and 'insures' less skilled workers from poor labour market outcomes, and immigrants are negatively selected when the host country taxes highly skilled workers and subsidizes less skilled workers.

There also exists the possibility that the host country draws persons who have below-average earnings in the source country but do well in the host country ($Q_0 < 0$ and $Q_1 > 0$). This sorting occurs when the correlation coefficient ρ_{01} is small or negative. This correlation may be negative when a source country experiences a structural political shift, such as a Communist takeover. In its initial stages, this political system often redistributes incomes by confiscating the assets of relatively successful persons. Immigrants from such systems will be in the lower tail of the post-revolution income distribution, but will perform well in a host country's market economy.

Equation (24) shows that neither differences in mean incomes across countries nor the level of migration costs determine the type of selection that characterizes immigrants. Mean incomes and migration costs affect the size of the flow (and the extent to which the skills of the average immigrant differ from the mean skills of the population), but they do not determine whether the immigrants are drawn mainly from the upper or lower tail of the skill distribution.

The discussion assumed that migration costs are constant in the population. Variable migration costs do not alter any of the selection rules if (*a*) time-equivalent migration costs are uncorrelated with skills, or (*b*) the variation in migration costs is 'small' relative to the variation in earnings. Otherwise, variable migration costs can change the nature of selection. Suppose that π is negatively correlated with earnings, perhaps because less skilled persons find it more difficult to find jobs in the host country. This negative correlation increases the likelihood that the immigrant flow is positively selected.

Some of the implications of the Roy model have been tested empirically by estimating the correlation between the earnings of immigrants in a host country, typically the United States, and measures of the rate of return to skills in source countries. The evidence provides mixed support for the Roy model's prediction that immigrants originating in countries with higher rates of return to skills have lower earnings in the United States. Borjas (1987) reports that measures of income inequality in the source country, which are a very rough

proxy for the rate of return to skills, are weakly negatively correlated with the earnings of immigrant men, while Cobb-Clark (1993) reports a similar finding for immigrant women. In contrast, Chiquiar and Hanson's study (2005) of Mexican emigration finds that the least skilled persons in the Mexican population tend to be under-represented in the Mexican-born workforce in the United States. Because the Mexican wage distribution has a larger variance than that of the United States, these low-skill workers would presumably be the persons most motivated to emigrate. This finding suggests that non-constant migration costs in the population of some source countries may play an important role in determining the selection of immigrants.

Measuring trends in immigrant skills

Beginning with the work of Chiswick (1978), a large literature has developed that attempts to measure the skill differential between immigrants and natives at the time of entry and how this differential changes over time as immigrants adapt to the host country's labour market. A key result of this literature is that there exists a cross-sectional positive correlation between the earnings of immigrants and the number of years that have elapsed since immigration. As will be seen below, there has been a great deal of debate over the interpretation of this correlation.

The empirical analysis of the relative economic performance of immigrants was initially based on the cross-section regression model:

$$\log w_\ell = X_\ell \beta_0 + \beta_1 I_\ell + \beta_2 y_\ell + \varepsilon_\ell, \quad (25)$$

where $w_\bullet$ is the wage rate of person • in the host country; $X_\bullet$ is a vector of socioeconomic characteristics (which includes age); $I_\bullet$ is a dummy variable set to unity if person • is foreign-born; and $y_\bullet$ gives the number of years that the immigrant has resided in the United States and is set to zero if • is a native. (The models used in empirical studies typically include higher-order polynomials in age and years-since-migration. These nonlinearities, however, do not play a role in the identification issue discussed below.) Because the vector X includes the worker's age, the coefficient β_2 measures the differential value that the host country's labour market attaches to time spent in the host country versus time spent in the source country.

Cross-section studies of immigrant earnings in several host countries have typically found that β_1 is negative and β_2 is positive. (Although most of the empirical evidence focuses on the US experience, the literature finds a similar correlation in Canada – Baker and Benjamin, 1994; Australia – Beggs and Chapman, 1991; and Germany – Dustmann, 1993.) Chiswick's (1978) analysis of the 1970 US Census data indicates that immigrants earn about 17 per cent less than ‘comparable’ natives at the time of entry, and this gap narrows by slightly over one percentage point per year. As a result, immigrant earnings overtake those of their native counterparts after about 15 years in the United States. The steeper age-earnings profiles of immigrants was interpreted as meaning that immigrants accumulated more human capital than natives as the assimilation process took hold, closing the wage gap between the two groups. The overtaking phenomenon was then explained by assuming that immigrants were positively selected. As we have seen, this assumption about the selection process is not necessarily implied by income-maximizing behaviour on the part of immigrants.

Borjas (1985) suggested an alternative interpretation of the cross-section evidence. Instead of interpreting the positive β_2 as a measure of assimilation, he argued that the cross-section data might be revealing a decline in relative skills across successive immigrant cohorts. In the United States, the post-war era witnessed major changes in immigration policy and in the size and national origin mix of the immigrant flow. If these changes generated a less-skilled immigrant flow, the cross-section correlation indicating that more recent immigrants earn less may say little about the process of wage convergence, but may instead reflect innate differences in ability or skills across cohorts.

To illustrate the identification problem, consider a hypothetical situation where there are three separate immigrant waves, and these waves have distinct productivities. One wave arrived in 1960, the second arrived in 1980, and the last arrived in 2000. Suppose also that all immigrants enter the United States at age 20.

Assume that the earliest cohort has the highest productivity level of any group in the population, including US-born workers. If we could observe their earnings in every year after they arrive in the United States, their age-earnings profile would be given by the line *PP* in Figure 2. For the sake of argument, let's assume that the last wave of immigrants (that is, the 2000 arrivals) is the least productive of any group in the population, including natives. If we could observe their earnings throughout their working lives, their age-earnings profile would be given by the line *RR* in the figure. Finally, suppose that the immigrants who arrived in 1980 have the same skills as natives. If we could observe their earnings at every age in their working lives, the age-earnings profiles of this cohort and of natives would overlap and be given by the line *QQ*. Note that the age-earnings profiles of each of the immigrant cohorts is parallel to the age-earnings profile of the native population. There is *no* wage convergence between immigrants and natives in this hypothetical example.

Figure 2

Cohort effects and the immigrant age-earnings profile

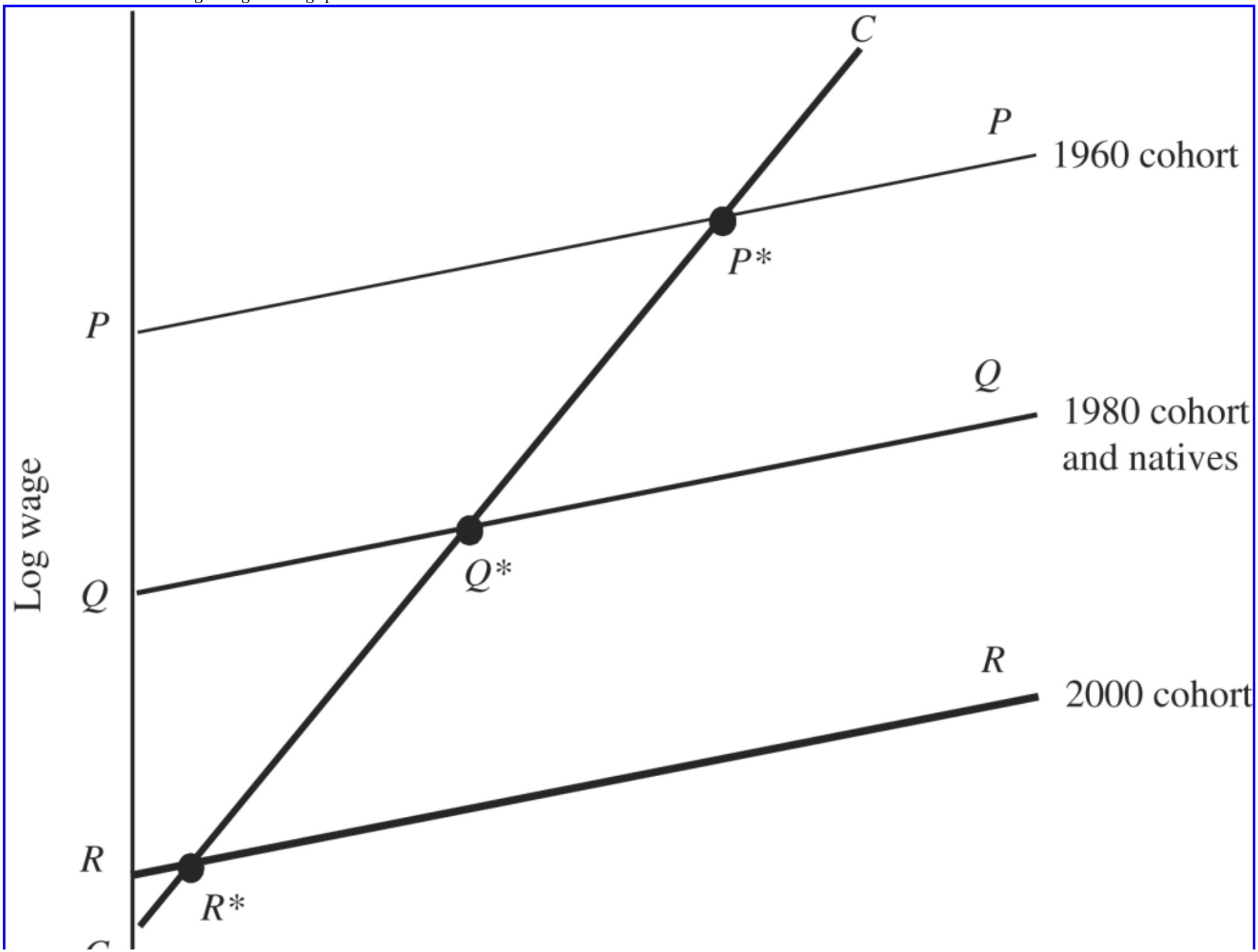

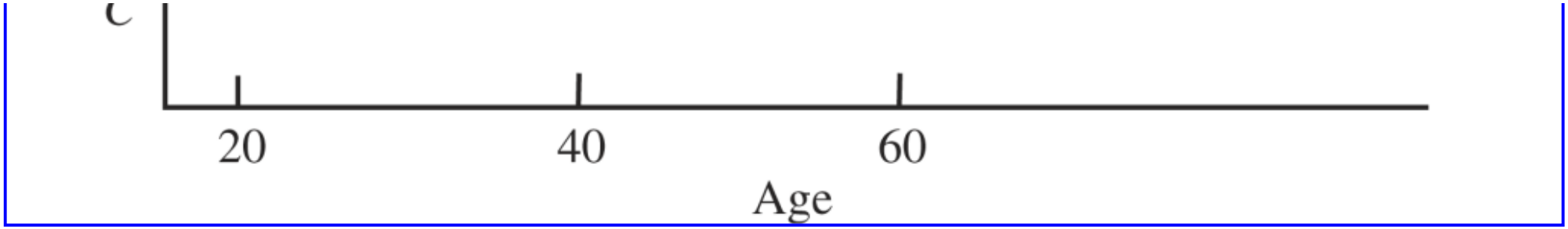

Suppose we now have access to data drawn from the 2000 decennial census. This cross-section data-set, which provides a snapshot of the US workforce as of 1 April 2000, provides information on each worker's wage rate, age, whether native- or foreign-born, and on the year the worker arrived in the United States. As a result, we can observe the wage of immigrants who have just arrived as part of the 2000 cohort when they are 20 years old (see point *R** in Figure 2). We can also observe the wage of immigrants who arrived in 1980 when they are 40 years old (point *Q**), and we observe the wage of immigrants who arrived in 1960 when they are 60 years old (point *P**). A cross-section data-set, therefore, allows us to observe only one point on each of the immigrant age-earnings profiles.

If we connect points *P**, *Q**, and *R**, we trace out the immigrant age-earnings profile that is generated by the cross-sectional data, or line *CC* in Figure 2. This cross-section line has two important properties. First, it is substantially steeper than the native age-earnings profile. The tracing out of the age-earnings profile of immigrants using cross-section data makes it seem as if there is wage convergence between immigrants and natives, when in fact there is none. Second, the cross-section line *CC* crosses the native line at age 40. This gives the appearance that immigrant earnings overtake those of natives after they have been in the United States for 20 years. In fact, no immigrant group experienced such an overtaking.

The identification of aging and cohort effects raises difficult methodological problems in many demographic contexts. Identification requires the availability of longitudinal data where a particular worker is tracked over time, or, equivalently, the availability of a number of repeated cross-sections so that specific cohorts can be tracked across survey years. Suppose that a total of Ω cross-section surveys are available, with cross-section $\tau(\tau = 1, \ldots, \Omega)$ being obtained in calendar year T_τ. Pool the data for immigrants and natives across the cross-sections, and consider the regression model:

$$\text{Immigrant equation: } \log w_{\ell\tau} = X_{\ell\tau}\varphi_{i\tau} + \alpha\, y_{\ell\tau} + \beta\, C_{\ell\tau} + \sum_{\tau=1}^{\Omega} \gamma_{i\tau}\pi_{\ell\tau} + \varepsilon_{\ell\tau}, \tag{26}$$

$$\text{Native equation: } \log w_{\ell\tau} = X_{\ell\tau}\varphi_{n\tau} + \sum_{\tau=1}^{\Omega} \gamma_{n\tau}\pi_{\ell\tau} + \varepsilon_{\ell\tau}, \tag{27}$$

where $w_{\bullet\tau}$ gives the wage of person • in cross-section τ, *X* gives a vector of socio-economic characteristics (including age); $C_{\bullet\tau}$ gives the calendar year in which the immigrant arrived in the host country; $y_{\bullet\tau}$ gives the number of years that the immigrant has resided in the host country $(y_{\ell\tau} = T_\tau - C_{\ell\tau})$; and $\pi_{\bullet\tau}$ is a dummy variable indicating if person • was drawn from cross-section τ.

The identification problem arises from the identity:

$$y_{\ell\tau} \equiv \sum_{\tau=1}^{\Omega} \pi_\tau (T_\tau - C_{\ell\tau}). \quad (28)$$

Equation (28) introduces perfect collinearity among the variables $y_{\bullet\tau}$, $C_{\bullet\tau}$ and $\pi_{\bullet\tau}$ in the immigrant earnings function. As a result, the key parameters of interest – α, β, and the vector γ_i – are not identified. Some type of restriction must be imposed to separately identify the aging effect, the cohort effect, and the period effects. Borjas (1985) proposed the restriction that the period effects are the same for immigrants and natives:

$$Y_{i\tau} = Y_{n\tau}, \quad \forall \tau. \quad (29)$$

Put differently, trends in aggregate economic conditions change immigrant and native wages by the same percentage amount. A useful way of thinking about this restriction is that the period effects for immigrants are calculated from *outside* the immigrant wage determination system.

The measurement of cohort and assimilation effects has received a great deal of attention, particularly in the US, context where the data indicate that cross-section age-earnings profiles overestimate the rate of convergence between immigrant and native earnings due to the presence of cohort effects like those illustrated in Figure 2.

Immigration and the welfare state

In addition to the labour market consequences, immigration has fiscal impacts on host countries because there may be significant costs associated with providing social services to the immigrants, and these costs will depend both on the skill composition of the immigrant population and on the generosity of the host country's welfare state. In fact, the immigration debate in many receiving countries has often focused on the possibility that immigrants may become public charges. Since 1882, for example, the United States has banned the entry of ‘any persons unable to take care of himself or herself without becoming a public charge’. Similarly, the US immigration statutes declare that ‘any alien who, within five years after the date of entry, has become a public charge … is deportable’.

There has been a great deal of concern over the possibility that the relatively generous welfare programmes offered by industrialized Western economies have become a magnet for immigrants. It is possible, for example, that generous welfare programmes attract immigrants who otherwise would not have migrated, or that the safety net discourages immigrants who ‘fail’ in the host country from returning to their origin. These magnetic effects raise questions about both the political legitimacy and economic viability of the welfare state. Who is entitled to the safety net that the host country's taxpayers pay for? And can the richer host countries afford to extend that safety net to the population of poorer countries? Surprisingly, few studies attempt to determine whether such magnetic effects are empirically important.

Much of the empirical debate over the link between immigration and welfare in recent years has instead been dominated by the bottom line: do immigrants pay their way in the welfare state? There exist a large number of accounting exercises, each purporting to calculate the amount of taxes paid by immigrants and the amount of social expenditures that can be attributed to immigrants. The estimates provided by many of these studies are often unconvincing, with the conclusion typically dictated by the accounting assumptions employed in the exercise. For example, how does one allocate expenditures in various public goods between immigrants and natives? In 1996, the US National Academy of Sciences attempted to settle the issue by examining in detail this contentious issue (Smith and Edmonston, 1997, ch. 6).

The National Academy report measured the ‘short-run’ impact of immigration on the fiscal ledger sheet of states and local governments, that is, the fiscal impact during a particular fiscal year. For two major immigrant-receiving states, California and New Jersey, the National Academy conducted an item-by-item accounting of expenditures incurred and taxes collected, and calculated how immigration affected each of these entries.

California attracts a disproportionately large number of the welfare recipients in the immigrant population, and provides a wide array of expensive services, ranging from generous welfare assistance to a world-class system of public universities and a sophisticated and well-maintained system of roads and freeways. It turns out that immigration increased the state and local taxes paid by the typical native household in California in 1995 by almost $1,174 annually. The cost-benefit calculation for New Jersey is less dramatic. Because New Jersey provides fewer state and local services, and because New Jersey attracts a different type of immigrant (more skilled and less prone to use government services), immigration increased the annual tax bill of New Jersey's typical native household by only $229.

If one were to extrapolate these estimates nationwide, the National Academy concluded that immigration increased the taxes of the typical native household in the United States by around $200 annually in the mid-1990s. There are approximately 90 million native households in the United States, so that the national fiscal burden is around $18 billion per year. Recall that the annual immigration surplus in the United States is estimated to be around 0.2 per cent of GDP, or roughly around $20 billion in 2000. In the short run, therefore, the available evidence suggests that the net gain (to the native population) from immigration is essentially zero.

It is important to note that this type of accounting exercise is myopic because it does not consider the long-run impact of immigration on government expenditures. For example, it has been argued that immigration may provide an important mechanism to alleviate the fiscal crisis that most industrialized countries will face as their populations age and the dependency ratio rises, putting much greater pressures on social insurance and the fiscal solvency of the welfare state. However, careful simulations of the fiscal consequences of this demographic transition – and of the costs and benefits from immigration in industrialized economies – suggest that immigration can play only a limited role in alleviating the fiscal stress.

Using an overlapping generations framework, these studies examine how the payroll tax rate must adjust to cover the expenses that will be inevitably incurred over the 21st century to provide social benefits to a relatively larger aging population (Fehr, Jokish and Kotlikoff, 2004; Storesletten, 2000). One can then simulate the impact of different immigration scenarios on the required payroll tax rate. These simulations typically suggest that the social insurance tax rate in industrialized economies will not fall drastically even if immigration were greatly expanded (such as doubling the size of the flow) over the next century.

The reason for the relative unimportance of immigration in this fiscal exercise is that immigrants themselves generate an increase in social expenditures, and this increase may reduce much of the perceived benefit from simply having a larger population over which to amortize the required expenses. In addition, social insurance programmes in many industrialized host countries tend to be progressive, so that the immigrant population, which is relatively low-skill, will generally contribute less to their funding and receive higher benefits. In short, immigration is not the panacea that can resolve the fiscal problems associated with an aging population in these societies.

See Also

- economic demography
- globalization
- Roy model

Bibliography

Altonji, J. and Card, D. 1991. The effects of immigration on the labor market outcomes of less-skilled natives. In *Immigration, Trade, and the Labor Market*, ed. J. Abowd and R. Freeman. Chicago: University of Chicago Press.

Baker, M. and Benjamin, D. 1994. The performance of immigrants in the Canadian labor market. *Journal of Labor Economics* 12, 369–405.

Beggs, J. and Chapman, B. 1991. Male immigrant wage and unemployment experience in Australia. In *Immigration, Trade, and the Labor Market*, ed. J. Abowd and R. Freeman. Chicago: University of Chicago Press.

Borjas, G. 1985. Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of Labor Economics* 3, 463–89.

Borjas, G. 1987. Self-selection and the earnings of immigrants. *American Economic Review* 77, 531–53.

Borjas, G. 1995. The economic benefits from immigration. *Journal of Economic Perspectives* 9(2), 3–22.

Borjas, G. 1999. The economic analysis of immigration. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: Elsevier.

Borjas, G. 2003. The labor demand curve is downward sloping: reexamining the impact of immigration on the labor market. *Quarterly Journal of Economics* 118, 1335–74.

Card, D. 1990. The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review* 43, 245–57.

Card, D. 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics* 19, 22–64.

Chiquiar, D. and Hanson, G. 2005. International migration, self-selection, and the distribution of wages: evidence from Mexico and the United States. *Journal of Political Economy* 113, 239–81.

Chiswick, B. 1978. The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy* 86, 897–921.

Cobb-Clark, D. 1993. Immigrant selectivity and wages: the evidence for women. *American Economic Review* 83, 986–93.

Davis, D. and Weinstein, D. 2002. Technological superiority and the losses from migration. Working Paper No. 8971. Cambridge, MA: NBER.

Dustmann, C. 1993. Earnings adjustment of temporary immigrants. *Journal of Population Economics* 6, 153–68.

Fehr, H., Jokisch, S. and Kotlikoff, L. 2004. The role of immigration in dealing with the developed worlds demographic transition. *FinanzArchiv* 69, 296–324.

Friedberg, R. and Hunt, J. 1995. The impact of immigration on host county wages, employment and growth. *Journal of Economic Perspectives* 9(2), 23–44.

Grossman, J. 1982. The substitutability of natives and immigrants in production. *Review of Economics and Statistics* 54, 596–603.

Hamermesh, D. 1993. *Labor Demand*. Princeton, NJ: Princeton University Press.

Johnson, G. 1997. Estimation of the impact of immigration on the distribution of income among minorities and others. In *Help or Hindrance? The Economic Implications of Immigration for African-Americans*, ed. D. Hamermesh and F. Bean. New York: Russell Sage Press.

Pischke, J.-S. and Velling, J. 1997. Employment effects of immigration to Germany: an analysis based on local labor markets. *Review of Economics and Statistics* 79, 594–604.

Roy, A. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.

Smith, J. and Edmonston, B., eds. 1997. *The New Americans: Economic, Demographic, and Fiscal Effects of Immigration*. Washington, DC: National Academy Press.

Storesletten, K. 2000. Sustaining fiscal policy through immigration. *Journal of Political Economy* 108, 300–23.

United Nations. 2002. *International Migration, 2002*. New York: Population Division, Department of Economics and Social Affairs, United Nations.

How to cite this article

Borjas, George J. "international migration." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000182> doi:10.1057/9780230226203.0830

The New Palgrave Dictionary of Economics Online

International Monetary Fund

Brett House, David Vines and W. Max Corden

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The International Monetary Fund (IMF) was set up in 1944 and charged with supervising the post-war Bretton Woods system of pegged but adjustable exchange rates as a means of promoting international monetary cooperation. Since the Bretton Woods system broke down in 1971, the IMF's role has become more complicated. It has exercised surveillance over its members' policies, worked to ensure the stability of the international financial system, and assisted the world's poorest economies. This article reviews the history and achievements of the IMF as well as the challenges it faces in redefining its role at the beginning of the 21st century.

Keywords

balance of payments disequilibria; balanced growth; banking crises; beggar-thy-neighbour; Bretton Woods system; capital account liberalization; capital controls; catch-up; clearing union; concessionality; conditionality; convertibility; currency crises; European Monetary Union; Exchange Rate Mechanism (EMU); fixed exchange rates; floating exchange rates; foreign aid; General Agreement on Tariffs and Trade; gold exchange standard; gold standard; incomes policy; inflation; inflation targeting; international capital flows; International Monetary Fund; intertemporal trade; Keynes, J.M.; Keynesianism; Meade, J.E.; monetarism; moral hazard; NAFTA; nominal exchange rates; pegged exchange rates; poverty alleviation; price control; real exchange rates; sovereign debt; special drawing rights (IMF); Taylor doctrine; trade liberalization; World Bank

Article

The International Monetary Fund (henceforth 'the IMF' or 'the Fund') was conceived at a conference at the Mount Washington Hotel in Bretton Woods, New Hampshire, in July 1944 and its Articles of Agreement entered into force in December 1945. The World Bank (henceforth 'the Bank') was set up at the same time. The IMF was established to promote international monetary cooperation and the

elimination of exchange restrictions on current account transactions; to facilitate trade, economic growth and high levels of employment; to foster exchange rate stability; and to provide temporary financial assistance to countries so as to ease balance of payments adjustment. More specifically, it was given the role of supervising a system of pegged but adjustable exchange rates, which became known as the Bretton Woods system. In the first two sections of this entry we explain how the Bretton Woods system worked, and why it broke down in 1971. In the following sections we consider the roles which the Fund now plays, which differ from its original activities. They are: surveillance, ensuring stability for the international financial system and for individual economies within this system, and assisting the world's poorest economies. As part of each of these three activities, the Fund also provides policy advice and technical assistance. This is a much less clear collection of responsibilities, and, as a result, the future direction of the Fund is somewhat uncertain. The aim of this article is to review the achievements of the Fund, and also the challenges that lie ahead. A related overview of some of the issues discussed here can be found in Gilbert and Vines (2004).

1 The Bretton Woods system

1.1 Intentions

As the Second World War drew to a close, the United Kingdom, the United States and their allies, inspired in part by the *General Theory* of John Maynard Keynes (Keynes, 1936), established a policy framework in which countries would be able to promote high levels of employment and output, by means of demand management policies, focused mainly on fiscal measures. This would – it was hoped – avert slumps in growth and would thereby prevent the re-emergence of the kind of global depression that had occurred in the 1930s. (See Williamson, 1983a; Moggridge, 1986.)

From early on, Keynes had seen that such policies would need global support. This is because they would have to be reconciled with the need for each country to be sufficiently competitive; that is, each country would need to be able to export enough to pay for the imports that would be purchased at full employment. In 1942, Keynes put forward plans for a new post-war international monetary system designed to make this possible, which he called a ‘Clearing Union’. (See Keynes, 1971–88, vol. 25, pp. 41–67; van Dormael, 1978; Gardner, 1956.) His plan drew on the theoretical arguments in his *General Theory*, and also on the harsh practical example provided by the United Kingdom's return to the gold standard in 1925 (Eichengreen, 1992). He argued that, for many countries, sufficient competitiveness would not be assured if the world returned to a gold standard after the war. Such a standard would require that any country with balance of payments difficulties, of the kind which Britain was likely to have, would need to rely on downward adjustment of its wages and prices in order to make its goods sufficiently attractive in world markets. Keynes judged that, in the political climate of the post-war world, such wage and price adjustments might not be possible. Nevertheless, because of the exchange rate instability of the early 1920s and the 1930s, he also showed no enthusiasm for floating exchange rates. The need for something different was discussed in much detail over the next two years with Harry Dexter White and others from the United States (Keynes, 1971–88, vol. 25, pp. 338 ff.), including during a visit that Keynes made to Washington in 1943.

The analytical content of these immensely difficult negotiations is explained in Meade, James Edward,

and is discussed in more detail in Vines (2003), which draws on the wonderful historical account by Skidelsky (2000). Skidelsky makes clear that Keynes was propelled in these discussions by the knowledge that the generous provision by the United States of wartime funding to the United Kingdom (‘Lend Lease’) had put the United States in a position in which it would be able to dismember the British Empire after the war. Keynes, who had been accustomed to Britain managing the global economy, wanted to create a new global order in which prospects for Britain remained acceptable, even although global economic hegemony would pass to the United States. He feared that difficulties in the balance-of-payments adjustment process might impose, on deficit countries like Britain, an obligation to deflate demand below full employment, something which might not be matched by symmetrical over-expansion by surplus countries, and might thereby create pressures towards global deflation. This is why he wanted his Clearing Union to be able to create global liquidity. (Like a bank, it would ‘clear’ the overdrafts which countries could obtain from it.) He differed in this view from Harry Dexter White, who feared an outcome in which liquidity would be so freely available that there would be a great post-war worldwide inflation.

What emerged at Bretton Woods was a global system of pegged but adjustable exchange rates, to be overseen by an International Monetary Fund. The currency system was to have three major features. First, each country would establish a *par value* for its currency in terms of gold or dollars. Second, all exchange controls would be removed for current-account transactions and all currencies would be freely convertible into dollars, although controls on international capital flows would remain in place. Third, dollars would be freely convertible into gold. Thus, the system was to be a ‘gold exchange standard’; it would differ from a gold standard in being a club rather than a unilateral pegging arrangement, and in allowing for occasional exchange rate changes.

The IMF would do two things in this system. First, exchange-rate pegs would only be adjusted if the approval of the IMF's Executive Board had been obtained. That approval would not be given unless there were deemed to be a ‘fundamental disequilibrium’. This term was imprecisely defined, but it meant a situation in which an exchange rate was not at a level that would ensure that exports could equal imports at full employment. This kind of test was designed, with the 1930s in mind, to prevent countries pursuing a ‘beggar-thy-neighbour’ devaluation of their currencies so as to steer towards full employment by ‘stealing’ jobs from other countries rather than by expanding expenditure at home. A country with longer-term difficulties would be declared to be in ‘fundamental disequilibrium’ and would be expected to devalue its currency by an appropriate amount after consulting with the Fund and getting the required approval. Similarly, a country with an excessively large and sustained balance of payments surplus would be expected to revalue its currency.

Second, the Fund would be set up like a credit union, into which members would place deposits; a country in temporary balance of payments difficulty rather than ‘fundamental disequilibrium’ would be able to draw on a short-term basis from the Fund to help it address the problem. It was thought that these loans would be repaid quite rapidly (that is, within three to five years), since more fundamental difficulties would be addressed by exchange rate adjustments. Each country in this credit union was to be given a ‘quota’, based on a nonlinear equation that took account of a country's national income, its international trade, and its official reserves; services, other external current account transactions, and a measure of volatility were further added to the quota formula in the 1960s. The quotas would define each country's capital contribution, its borrowing entitlement, and, in aggregate, the Fund's lending capacity. The US quota was initially about 20 per cent of the total (less than would have been implied by

a strict calculation based on the variables noted above), and originally the United Kingdom had, by design, the second largest quota. This was *not* like Keynes's Clearing Union, and Keynes was dismayed at how little the Fund would be able to lend (see Vines, 2003). There have been a number of substantial increases in total quotas under regular quinquennial reviews, but they have not grown in such a way as to keep pace with the expansion of the world economy and international financial flows. In addition, as the relative size and importance of countries have changed, there has been a need to adjust both quota shares and the factors used in the calculation of these quotas. Both of these types of adjustment have been politically difficult; a (small and interim) adjustment for four emerging-market countries (China, Korea, Mexico, and Turkey) happened in September 2006.

The quota system partly determined the relative voting entitlements of countries on the Executive Board of the Fund. It seemed obvious, for a credit union to which money had been contributed, to make voting power depend partly on the amount contributed, and on the amount which could be borrowed at a time of difficulty, rather than using a one-member, one-vote system of governance like that adopted at the United Nations. However, there were also a number of ‘basic votes’ allotted equally to all members, whose effect was to mitigate a little the voting power of large countries.

The Fund's Articles and their subsequent amendments established that a member is allowed to borrow up to a certain proportion of its quota as of right, without policy conditions. This amount was referred to as the ‘reserve tranche’; it was equal to 25 per cent of quota and corresponded to the amount that a member had paid into the Fund in hard foreign currencies. Beyond the reserve tranche, a country had an option to borrow up to four ‘credit tranches’, each of which represented 25 per cent of quota. Access to the first credit tranche was relatively easy; borrowing under the subsequent or ‘upper’ credit tranches was normally made available through what were (and still are) described rather quaintly as ‘stand-by arrangements’.

1.2 Consequences

The international monetary system followed only imperfectly the intentions underpinning the Bretton Woods system, and only until 1971. (See de Vries, 1976.) Current-account convertibility, for most European currencies, was not achieved until 1958 (the year after a large US current account deficit). There was a reluctance to alter exchange rates even in the presence of ‘fundamental disequilibrium’. And the Fund was unable to stop France from implementing a multiple currency system in 1948. One major currency, the Canadian dollar, floated from 1950 to 1962 and the Fund acquiesced in this. The Fund ratified British devaluations in 1949 and 1967 at short notice (though it was closely involved in discussions in the second case). It had little influence on US policies – and has had little influence ever since. It played virtually no role in the later US decision to end gold convertibility in August 1971, a decision which brought the Bretton Woods system crashing down. And it had limited influence on the policies of the principal surplus countries in the 1960s. On the other hand, the Fund did have a role in the exchange rate realignments of other currencies that took place in 1949, 1967 and 1971 as a result of the sterling and dollar devaluations, seeking to ensure ‘orderly adjustment’. The most important point is that the IMF had an influence mainly through the conditions it could impose on those countries (such as the United Kingdom in 1976) which needed its funds.

When the Fund began providing stand-by arrangements in 1952 they were typically of short duration and did not feature any conditions. This may seem surprising now, given the close association in the

popular imagination between conditional lending and the IMF. Policy conditions were first added to Fund-supported programmes in 1954, partially in light of the increase in the size of borrowing under stand-by arrangements, as compared with first-credit tranche financing. Quantitative targets or ‘performance criteria’ followed in 1957, in order to provide a clear baseline for policymaking under IMF-supported programmes, and an objective yardstick by which the effects of these policies – and the possible need for further adjustments – might be assessed. They were calibrated using the Fund's financial programming framework, developed by Polak (1957), and came to be a nearly universal feature of Fund-supported programmes by the mid-1960s. (See IMF, 1987; 2004a; Mussa and Savastano, 1999.) This combination of policy or ‘structural’ commitments and quantitative performance criteria came to characterize the ‘conditionality’ attached to IMF lending from the 1960s to the present. This was justified – then as now – not so much as a way of collateralizing IMF lending, and guaranteeing a turnover of the IMF's funds, but rather as a means of ensuring the viability of Fund-supported programmes and the quick adjustment of countries in crisis back to a balanced growth path. The period from 1945 to 1971 was one of extraordinary dynamism (a ‘golden age’): it was a time in which Europe and Japan were first rebuilt after the war and then proceeded to catch up with the United States. The Bretton Woods system appears to have played a part in ensuring that this happened. In this system, the Fund was helped by the World Bank, whose role was to lend money for longer periods than the Fund, first for reconstruction after the war, and then, later on, to help finance development. (Keynes once helpfully remarked that in order to comprehend the Bretton Woods institutions one has to understand that the Fund is a bank, and the Bank is a fund.) The purpose of this World Bank lending was to enable these countries to borrow abroad (in a world in which there was little international mobility of private capital), to run balance of trade deficits, to invest, and to grow – with the expectation that the borrowing would then be repaid out of the increased export proceeds that investment and growth made possible. In addition, a conference in Geneva in 1947 established the General Agreement on Tariffs and Trade (or GATT) to supplement the Bretton Woods system by encouraging the growth of international trade. The GATT's role in promoting the liberalization of trade restrictions supplemented the Fund's role in promoting the liberalization of exchange restrictions on current account transactions. In due course, a series of GATT ‘rounds’ brought about tariff reductions, which helped to create markets for exports as countries expanded. With high employment, with balance-of-payments deficits dealt with as described above, and with many countries growing by exporting, there were clear incentives for most countries to support trade liberalization. That, in turn, made exports and imports more sensitive to exchange-rate levels and so made balance of payments adjustment easier to achieve by exchange-rate adjustments. Yet, these linkages between different aspects of the overall post-war policy framework are difficult to pin down empirically. This explains why economic historians still differ in their view as to how important the Bretton Woods system actually was in sustaining the golden age of growth observed in the 1950s and 1960s. (See Matthews, Feinstein and Odling-Smee, 1982; Matthews and Bowen, 1988; Temin, 2002; papers in Eichengreen, 1995; and Eichengreen, 2007.)

2 Breakdown and reconfiguration

Up to the 1960s the growth of gold reserves had been slow, and the need for additional international liquidity was increasingly met by the use of the US dollar as a ‘reserve currency’. This led to calls for

the IMF to create a more multilateral way to augment official reserves. The IMF's Articles of Agreement were eventually amended in 1969 to allow the Fund to create ‘special drawing rights’ (SDRs) that would act as the Fund's unit of account and which could be used as a source of credit for member countries. (See Corden, 1983a; Boughton, 2001.)

In the 1960s, imbalances also began to emerge: by the latter part of the decade, the United States had a large balance of payments deficit. A belief emerged that the dollar price of gold might rise as economic growth in Europe and Japan weakened the US dollar's role as anchor of the Bretton Woods system. In 1968, central banks ceased their efforts to control the dollar price of gold in private markets, which meant that the prevailing fixed price of gold applied only to central bank dealings. The market price of gold rose: in August 1971, following a massive speculative attack on the dollar, the United States ended the gold convertibility of dollars held by central banks and, as a result, the entire gold exchange standard broke down. A reluctant movement from a pegged exchange-rate system to a system with floating exchange rates followed. This outcome can best be explained by three sets of factors. (See Corden, 1993.)

First, many countries were unwilling to adjust the exchange rates for their currencies in the face of fundamental disequilibria. It was particularly problematic that the core country, the United States, behaved in this way. Because US productivity growth lagged behind that of the countries which were catching up with it, the trade position of the United States was at risk by the late 1960s. In addition, the United States fought the Vietnam War and launched its ‘Great Society’ programmes at the same time, without adequately raising taxes. The result was a large balance of payments deficit for the United States, the correction of which required both real exchange rate depreciation and restraint of domestic expenditure. Neither of these actions was forthcoming.

Second, the growth of international capital flows – which was in part a result of the international stability associated with the golden age – helped to undermine the system. As first demonstrated by the 1967 sterling crisis, it was no longer possible for the IMF and national governments to set exchange rates without reference to the forward-looking perceptions of private markets about what sustainable exchange rates might be. With increasingly mobile capital, once a suspicion was generated that that there would be (or might need to be) a devaluation of a country's currency to preserve external balance, speculation could make it difficult or impossible for central banks to defend an existing rate. By 1971, the balance of payments deficit of the United States had caused a large build-up of mobile dollar holdings in offshore or ‘Euro-dollar’ accounts. These funds were used to finance the speculative attack on the dollar in 1971.

Third, the Keynesian macroeconomic policy framework established after the Second World War contained no clear responsibility for preventing inflation. Although there were periods of (generally unsuccessful) price controls or ‘incomes policy’, the seeds of incipient inflation were sown by this omission. Eventually, tensions generated by the oil price shock of 1973, and by the period of undisciplined inflation which followed it, led to more than the collapse of the Bretton Woods system. The entire structure of Keynesian, interventionist, high-employment policies, which had been at the centre of the post-war policy architecture, came tumbling down, both in the United States and in Europe. For the ten years after 1971, macroeconomic policy was in a state of worldwide disarray.

The great inflation of the 1970s led to significant movements in the real exchange rates between countries, which killed nearly all of the (many) attempts made at the time to reconstruct an international monetary system with pegged exchange rates. (See Williamson, 1977.) There was only one lasting,

partial, attempt to reconfigure such a system, in Europe, which led to the European Monetary Union. For a period of time it appeared that the Keynesian approach to macroeconomic policy might be replaced by monetarist policies of a non-interventionist kind. But this alternative proved unsuccessful. Instead, with great difficulty, activist macroeconomic policies were reconstructed by the 1990s within inflation-targeting regimes, in which an inflation target was pursued through interest rate changes. This new system quickly came to be allied with a system of floating exchange rates in which there was a high degree of international capital mobility. In this new set-up, a floating exchange rate would help to stabilize demand, and movements in the exchange rate would become an important part of the process of inflation control. If a country suffered from a shock which raised prices, then its monetary policymakers would set higher interest rates, and the nominal exchange rate of the country would appreciate. This would reduce net exports and import costs, and so inflation.

As a result of this reconfiguration of policy assignments, a second revision of the Fund's Articles of Agreement was made in 1976 and came into effect in 1978. At Bretton Woods, the Fund had been set up to manage a pegged exchange rate system. But it came to be realized that a country cannot have, at the same time, an independent monetary policy, capital markets which are open to the rest of the world, and a pegged exchange rate. (These three things, taken together, have become known as an ‘impossible trinity’. The reason that these things cannot occur together is to be found in the Mundell–Fleming macroeconomic model, which was developed by Fleming and Mundell, at the IMF, in the early 1960s.) As a result, the Fund's revised Articles ratified a new form of international monetary system in which a country did not have to establish a par value for its exchange rate, but could instead have exchange rate arrangements of its own choice.

Since 1978, the Fund has gradually been drawn into new roles, in support of this revised, and more flexible, system. As described in the introduction, its work now has three aspects. First, the Fund's Articles, as revised in 1976, require it to exercise surveillance and influence over macroeconomic policies, and to monitor and guard against the development of unsustainable conditions that could lead to financial crisis. The Fund still lends to countries in balance of payments difficulty, and its second activity has been to do this for emerging-market economies and for ‘transition economies’ moving from central planning to market-based systems. More than this, the Fund helps such countries to deal with, and to prevent, the financial crises that have afflicted a number of them. Third, the Fund has lent money to the poorest developing countries, which generally do not have capital-market access. In these cases, Fund lending has often been indistinguishable from other long-term concessional development assistance, and the Fund's main distinctive contribution has been to work with central banks and finance ministries in crafting credible macroeconomic frameworks that can elicit further support from aid donors. We consider each of these three activities in turn.

3 The IMF and policy surveillance

Countries that are creditworthy, and which have access to highly mobile international capital under floating exchange rate regimes, no longer need to borrow from the Fund in the way they did when the Fund was first established. Such countries can adjust to balance of payments disequilibria through exchange rate movements, supported by foreign borrowing from sources other than the Fund. (See Corden, 1983b, and Dam, 1982). At the time of writing, no advanced country had agreed a borrowing arrangement with the Fund since the substantial stand-by arrangements with the United Kingdom and

with Italy in 1976. Fund lending is only required at a time when a country ceases to be perceived as clearly creditworthy, something which, as of mid-2007, had not happened in industrial countries since 1976. This was true even at the time of the crisis of the Exchange Rate Mechanism of the European Monetary System in 1992. The Fund did not at that time provide financing to assist Sweden, Italy, the United Kingdom, or France in a defence of their currencies. When crisis struck, these countries (eventually) allowed their currencies to float downwards, rather than using lending from the IMF to defend further their exchange rates.

Nevertheless, a world with a high degree of international capital mobility is not without difficulties. In such a system, the spending decisions of nations can move away from permanently sustainable positions for very long periods of time, an outcome with an external current account deficit (or surplus) offset by an external capital account surplus (or deficit). The ‘global imbalances’ that can result have, as of mid-2007, been substantial at three points of time since the 1960s. In the late 1960s, as we have seen, the US ran a large current account deficit; current account surpluses of a number of European economies and of Japan, which, as noted above, were engaged in a process of export-led growth and ‘catch-up’, were the ‘other side of the coin’. Nearly 20 years later, in the early to mid-1980s, President Reagan increased defence expenditures and cut taxes. Tight monetary policy was used to restrain demand in the United States, which caused the dollar to appreciate, and the result was a large current account deficit. Japanese current account surpluses were on the other side of this coin. Twenty years later, in 2007, the United States was again running a large fiscal deficit and an (unprecedentedly) large current account deficit; and again Japan was running the corresponding current account surpluses, along with China, other emerging-market economies in East Asia and elsewhere, and a number of oil-producing countries.

These global imbalances reflect decisions by countries to de-link income and spending over time. Of course, such ‘intertemporal trade’ can be welfare-improving. But such imbalances might instead reflect an urge by a deficit country to spend beyond its means. This was clearly the case for the United States in the late 1960s and the mid-1980s, and might also be the case from 2000 (and especially from 2005). Conversely, these imbalances might also partly reflect a desire by some countries to maintain their currencies at artificially devalued levels against the US dollar, in order to grow quickly through a process of export-led catch-up. This is something which, at one time, would have been called ‘beggar-thy-neighbour’ behaviour of the kind which the IMF was established to prevent. As noted above, one can argue that this may have been what was done by western Europe and Japan in the late 1960s. Some commentators have argued that a number of emerging-market economies in East Asia, and elsewhere, were behaving the same way in the early 21st century (Dooley, Landau and Garber, 2003; Roubini and Setser, 2005). These commentators, in recognition of the parallel, suggested that we were living under a ‘Bretton Woods II’ regime.

But global imbalances eventually unwind. They must do so if countries are eventually to repay what they owe. In 1971, global imbalances led to crisis, and to the collapse of the Bretton Woods financial system. By contrast, the imbalances of the mid-1980s were resolved in an orderly way. (See Eichengreen, 2004; Eichengreen and Park, 2006; Corden, 2007; Joshi, Lane, and Vines, 2006; Williamson, 2006.) Such orderly adjustment requires the deficit country to cut expenditure, and its currency to depreciate significantly (unless it grows its way out of difficulty). It also requires, in addition, that expenditure in surplus countries expands so that global expenditure is maintained, or, if this does not happen, that global interest rates fall so that global expenditure is stimulated by other means. If all of this happens, as it did in the late 1980s, then the benefits of intertemporal separation

between spending and income may not be diminished by the costs of an adjustment crisis.

There are four main ways in which the existence of the Fund helps global imbalances to unwind in an orderly manner.

First, ever since the second amendment of the Fund's Articles described above, the Fund has been required to exercise ‘firm surveillance’ over the exchange rate and macroeconomic policies of its members. As a result, the Fund regularly sends to each country an ‘Article IV mission’ whose purpose is to review the country's macroeconomic policies. This is done annually for most countries, and at interludes of up to 24 months in countries with active Fund-supported programmes. (For such countries the Article IV cycle is elongated since policies are reviewed frequently in the context of semi-annual or quarterly programme reviews.) All aspects of macroeconomic policy are considered on these occasions. Following the emerging-markets crises of the 1990s and early 2000s, the Article IV consultation process has been supplemented by detailed review of countries’ financial sectors under the World Bank and IMF's joint Financial Sector Assessment Program (FSAP).

Second, the Fund provides a vast amount of published information and analysis, both about the world economy and financial system in general and about particular countries. The Fund's biannual *World Economic Outlook* provides a forecast for the world economy, and analyses multilateral and regional issues; this report is supplemented by *Regional Economic Outlook*s. These products are based in part on Article IV consultations and would not be possible without that process. The Fund also publishes a biannual *Global Financial Stability Report* which monitors markets, and several statistical publications that compile economic and financial data supplied by member countries, including *International Financial Statistics*.

Third, the Fund plays an important role in keeping the governments of all members in touch with developments in other countries and globally. The Article IV missions to the largest economies (and the related research, published in *Selected Economic Issues* papers that are companions to the Fund's Article IV staff reports) are particularly important in helping to keep governments informed of policies and developments that are likely to affect the world economy as a whole. Additionally, the Annual Meetings of the Boards of Governors of the IMF and the World Bank enable an informed exchange of ideas between countries, as do the Spring Meetings. The Fund thus provides a valuable global information network.

Finally, the Fund has also created a valuable global human network. Fund staff are of high quality, something which is necessary since they have to deal with senior officials in many countries. The offices of Executive Directors of the Fund in Washington act as valuable means of communication between the member nations of the Fund. And in many national capitals a large number of public servants and elected officials have served on the Fund staff earlier in their careers, or have been located in Washington as Executive Directors at the Fund or as members of staff in Executive Directors’ offices. This experience has made many decision-makers more internationally minded than they might otherwise have been.

Nevertheless, some have argued that the Fund's ‘firm surveillance’ is not firm enough. Arriazu, Crow and Thygesen (1999) discuss the impact of Fund surveillance, country by county, in the Article IV consultation process. They note that, although these consultations have been ‘taken seriously’, it does not appear that these reviews by the Fund have had more than an occasional impact on national policy decisions in some countries. A more recent assessment of Article IV consultations by Meyer et al. (2004) reaches similar conclusions. When an Article IV mission goes to a country that does not borrow

from the Fund (and which therefore does not require the Fund's *imprimatur* in order to obtain loans from other official creditors or from banks), the mission is usually relegated to a mainly advisory role, for which ‘surveillance’ may be too grand a label. But this *de facto* situation is not inevitable, since the *de jure* position of the Fund is that it should assess and appraise as well as advise. Goldstein (2006) asserts that there are gaps in the current practice of bilateral surveillance and argues, in particular, that the Fund's dealings with China in the early 21st century have not been satisfactory in addressing and effecting remedies for exchange rate misalignments. He further observes that the Fund's Managing Director has only rarely used the power granted to him by the 1977 and 1979 Board decisions on ad hoc and ‘supplemental’ consultations with members to address cases where a country's exchange rate policies appear inconsistent with the exchange rate principles of the Fund's Articles. (See Boughton, 2001.)

It is important to note that these critics do *not* seek policy changes from countries, in the interests of the greater good, that such countries would find unattractive if left to make policy choices on their own. That is, it is not suggested that the Fund could enforce a ‘cooperative’ outcome in macroeconomic policymaking when countries would prefer a different selfish, or ‘Nash,’ outcome. (This difference between Nash and cooperative outcomes was much discussed in the 1980s literature on policy coordination, summarized by McKibbin, 1997). Instead, it is argued that the Fund could enable cooperative outcomes, so that any adjustments in countries’ policies that need to happen in the face of global imbalances might happen in the right sequence rather than in a disorganized manner. The capacity to enforce even this modest form of coordination might occasionally be important in the adjustment processes. (See Kumar, 2006; Wolf, 2005; 2006; Joshi, Lane and Vines, 2006.)

There was action of this kind under the Plaza Accord of September 1985, although it was not coordinated by the Fund. At this time, the finance ministers of the world's five largest national economies agreed that the value of the dollar needed to go down. They also arrived at some (rather general) agreements on the monetary and fiscal policies that would be needed in order for this fall in the dollar to be achievable, and announced coordinated intervention in foreign-exchange markets to help bring it about.

To act effectively in this way requires the Fund to come to terms with the difficult tension between its strengths as a universalist institution and the need, on occasion, to bring together a more limited group of players. But it is an objective of the Fund's current Medium-Term Strategy that it should provide such a forum (IMF, 2005b). The Fund's Multilateral Consultation on global imbalances began by consulting with the United States, the European Union, Japan, China and Saudi Arabia, and it reported on its findings in April 2007. This work ran in parallel with similar discussions at summit meetings of Heads of Government of the Group of Eight Countries (or G8), and at meetings of the finance ministers and central bank governors of these countries. The G8 consists of the United States, Russia, Japan, Germany, Britain, France, Italy, and Canada. This is a powerful collection of countries, but it is not clear that these G8 meetings have had the right participants to deal with the global imbalances of the early 2000s. China and India have not been members of this group (though they have been observers), nor have many of the major oil-producing economies; by contrast, Canada and Italy, while committed to the G8 process, have been perhaps too small to contribute substantially to coordinated efforts to unwind global imbalances. The Fund may therefore have more to offer than such G8 gatherings, since the Fund can act as a locus of coordination amongst subsets of its membership, convening small groups of countries to deal with

particular problems.

Nevertheless there are three reasons why further progress may be slow on this front.

First, in the words of the IMF's Independent Evaluation Office (IEO) (IMF, 2006a, p. 2), ‘As a result of its … [country-by-country] orientation, multilateral surveillance has not sufficiently explored options to deal with policy spillovers in a global context’. Pursuing this theme, Mervyn King, Governor of the Bank of England, made it clear (King, 2006b) that more effective multilateral surveillance would require: (i) that countries made clearer commitments about their objectives for macroeconomic policies (that is, fiscal, monetary and financial); (ii) that the Fund's Article IV and the *World Economic Outlook* processes focused more transparently on cases when these policy commitments, and the countries’ policy actions, are not globally consistent; and (iii) that this process also transparently demonstrated the negative spillover effects that come from such lack of consistency and proposed actions to reduce such negative spillovers. But, given the limits to the precision of what we know about the international economy at any given time, doing this would be difficult. And it should be noted that the Fund's management issued a rejoinder to the 2006 IEO report which explained this difficulty.

Second, there may well be governance limitations on such firm surveillance. As of 2007, Article IV consultations were not finalized by the Fund Staff sent on the Article IV mission, but by the Fund's Executive Board, whose views were conveyed to the authorities of the country concerned after discussion at the Board. It is possible that this has compromised the space for missions to assess and appraise frankly. If the process of IMF surveillance were made more independent of the IMF's Executive Board, then this might allow clearer messages to be delivered to the Fund's member countries. As against this, the messages might then lose political weight because they would no longer be seen as the views of the global community represented in the Executive Board.

Third, and fundamentally, the Fund is not an agent of a sovereign state in the way that central banks (except the European Central Bank) are, however ‘independent’ these central banks may be. As a result, the Fund has no actual instruments of its own with which its recommendations on global cooperation can be implemented. It must always rely on being able to persuade its members to act.

4 The IMF and crises in emerging markets since 1980

In the mid-to-late 1970s, after the rise in the price of oil in 1973, funds flooded from oil producers on to the international capital market and flowed to middle-income countries. The early to mid-1990s saw a further massive surge of private capital flows into emerging market economies, and this was repeated in the mid-2000s. The economic benefits of such international mobility are obvious: if capital flows from relatively rich to relatively poor countries, and if the rate of return is high in poor countries, the potential gains are high for both borrower and lender. But such funds are not always used well, the volatility of these flows can be very high, and they can create dangerous mismatches in the maturities and currencies of assets and liabilities. Indeed, these flows contributed to three major waves of financial crises, in Latin America, East Asia and Russia, something which called into question the stability of the entire international financial system. Across these regions of the world, the IMF has been required to help prevent such crises through surveillance. It has also been required to assist in the orderly workout of crises, through lending and through ongoing engagement in the development of macroeconomic policies in the countries which it assists. We explain how the Fund's activities have evolved in these emerging-market economies, and how its role has broadened. We do this by examining the three generations of

emerging-markets crises that occurred from the early 1980s onward.

4.1 The Latin American debt crisis: a ‘ first-generation ’ crisis

Oil money, facilitated by loans from international banks, financed a spending boom in Latin America and elsewhere during the 1970s. This led to a rapid increase in foreign debts (Little et al., 1993) in countries which were not in a position subsequently to adjust and service these debts. In due course, significant balance of payments problems emerged when, in 1980–82, real interest rates rose, driven by tight monetary policy in the United States and by a world recession which worsened the terms of trade for many emerging-market economies. These countries rediscovered the truth of what Keynes had maintained 40 years earlier: adjustment to external difficulties requires both good budgetary control and an appropriately competitive real exchange rate (Corden, 1990; Little, 1993). This turned out to be something which many policymakers in Latin America, and elsewhere, were unable to engineer, and monetized fiscal deficits led to reserve losses, uncontrolled devaluations of currencies and inflation, and difficulties in meeting foreign-currency-denominated debt obligations. Currency and debt crises were triggered more or less mechanically as macroeconomic fundamentals drove reserves down to critical levels, resulting in what has become known as a ‘first-generation’ crisis.

Although Latin America is most closely associated with the debt crisis of the early 1980s, other countries, including Morocco, were also involved. The crisis placed the IMF at the centre of the world stage in a way which made it more prominent than it had ever been under the Bretton Woods system. The Fund played four roles. First, it offered financial support with stand-by arrangements and other lending facilities. Second, the Fund came to define the broad envelope of resources that a country could be expected to devote to meeting its residual obligations under a debt rescheduling. In turn, the Fund, together with the United States and other bilateral creditors in the Paris Club, pressed creditor banks to reschedule debts and to engage in ‘concerted lending’ programmes, threatening to provide no support for indebted countries if banks did not cooperate, and, hence, making defaults more likely. Third, the Fund's advice and conditionality, together with that of the World Bank, had significant effects on indebted governments’ policies: they were encouraged to undertake growth-oriented structural reforms to escape from their debt problems. Fourth, the Fund's reports and conditionality provided the ‘seal of good housekeeping’ on the basis of which banks and bilateral creditors could justify rescheduling existing debt and providing new funds.

This use of the Fund, and the broader strategy surrounding it, is usually associated with James Baker, then Secretary of the US Treasury. It was a success only to the extent that it made the financial crisis manageable. The strategy avoided explicit debt reduction and insisted that indebted countries meet their obligations, although over an extended period of time. (This lengthening of the repayment profile did, of course, lead to some reduction in the net present value of debt.) Such an approach was advocated by the governments of major industrialized countries, especially the United States, that were concerned about systemic risks to their own banking systems arising from widespread write-downs of debt. The Fund was criticized in some quarters for agreeing to this strategy and for acting as an ‘enforcer’ of debt service on behalf of private banks.

A policy shift took place in 1989. Under the Brady Plan, also initiated by the US administration, the Fund and the World Bank provided encouragement and some financial support for debt reduction programmes for those countries (notably Mexico) where major policy reforms were being undertaken.

The shift from the Baker Plan to the Brady Plan represented a tilt in favour of debtor countries relative to creditor banks. But this came only after a long period in which these banks were able to rebuild their balance sheets, thereby putting them in a position to weather debt restructuring. The US Treasury induced creditors to grant write-downs to debtor countries by collateralizing the debt that emerged from these restructurings. The Fund backed up this carrot by concluding financing packages with debtor countries before the terms of debt reschedulings had been determined: a practice that came to be known as ‘lending into arrears’. This acted as a stick to weaken creditor leverage in the negotiation process, and it also greatly strengthened the role of the Fund in debt work-outs since, during the negotiations, Fund staff came to play a major role in influencing debtor countries’ macroeconomic policies.

4.2 The Mexican ‘ Tequila ’ crisis: a ‘ second-generation ’ crisis

The Latin American debt crisis of the early 1980s had been caused by *public*-sector overspending. But in 1994 something new happened. A major financial crisis, caused by the outflow of *private* capital, of the kind which had brought down the Bretton Woods system in 1971 and the European Monetary System in 1992, happened in Mexico. The Mexican crisis was different from the Latin American turmoil of the 1980s in that it was set off not just by fundamental weaknesses, such as unsustainable fiscal and current account deficits, but also by currency mismatches on the public-sector balance sheet. (See Calvo and Mendoza, 1996.) These caused a ‘second-generation crisis’ in the form of a self-fulfilling currency run.

This crisis presented new challenges for the IMF since it marked the first of a series of crises in emerging markets that originated in the capital account, rather than the current account, of the external balance of payments. The IMF was called on to assist Mexico despite the fact that its Articles of Agreement provide it with only limited jurisdiction over capital account issues.

Mexico had implemented a comprehensive reform programme in the early 1990s, which included financial liberalization and the completion of the North American Free Trade Agreement (NAFTA) in 1993. This led to a surge in investment financed mainly by foreign capital flows. The result was a large (real) overvaluation of the peso and a very large current account deficit. Initially, the government maintained prudent fiscal policy. But during 1994 many began to question the sustainability of the exchange rate, the fiscal position and current account deficit. By December 1994 there was a massive reversal of capital flows, and the peso plummeted. The consequences for Mexico were severe: inflation rose from 7 per cent in 1994 to 35 per cent in 1995; and GDP fell by 6.2 per cent in 1995 compared with a growth rate of 4.4 per cent in the preceding year.

The pain inflicted on Mexico by private investors led to a view that pegged exchange-rate regimes are unviable everywhere, not just in advanced industrial countries. (Mexico had a ‘crawling peg’ at the time.) And in Mexico there was a new emerging-market feature. Much of the Mexican government's debt was denominated in US dollars (for example, the ‘*tesobonos*’) because of the difficulty and high costs of borrowing in local currency; much of the government's revenue stream, by contrast, was peso-denominated (although oil revenue was denominated in dollars). This mismatch meant that the collapse of the peso led the government to the verge of default in early 1995.

The Fund played a critical role in stabilizing the crisis. In particular, drawing on financing from bilateral creditors, it coordinated assistance, mainly from the United States, that totalled more than five times Mexico's quota entitlements at the IMF. After a significant real devaluation of the peso and fiscal correction, exports rebounded, the economy grew, although only slowly, and Mexico earned enough

foreign exchange to repay the exceptional financing that had been provided to it during the crisis. Some subsequent analyses (see, for example, Calvo and Goldstein, 1996) were critical of the IMF's role in both surveillance and in crisis management for Mexico. But the arguments cut both ways.

On surveillance, it was claimed that IMF reports prior to the crisis placed insufficient emphasis on the vulnerabilities of public-sector and financial-sector balance sheets to the possibility of a run on the currency. Some authors argued that the Fund should have been more frank in conveying its views on macroeconomic and exchange-rate policy to its members, and that it should publish these appraisals. But there may well have been inadequate provision of information by Mexico to the Fund, as well as to the public. In particular, it appears that incomplete data may have been provided on official international reserves and liabilities (although the Mexican authorities disagreed with this claim when it was made). As a result, following the Mexican crisis, the Fund began a drive to get countries to sign on to transparency standards, such as the Fund's Special Data Dissemination Standards (which were established in 1996; see Fischer, 2004, p. 127). Additionally, the Fund began the practice of publishing Board documents, except when the authorities of a country objected. But this heightened focus on transparency left the Fund unclear on whether it should assist countries confidentially to prevent crises or spur corrective action by bringing bad news to the market. Given the sometimes self-fulfilling mechanics of second-generation currency crises, solving this dilemma is critical in defining the future role of the Fund in crisis prevention.

On crisis management, no clear conclusions emerged, either. *Ex post* it appeared that the private sector should have been prepared to lend short term to the Mexican government in the way that the IMF and the United States did. Overcoming such a market failure is surely a role of the IMF and national governments, and giving the IMF the capacity to provide such big loans seemed important to many observers. From this experience, Sachs (1995) concluded that the Fund should be given an explicit international lender-of-last-resort capacity, well beyond that formally possible under its ‘credit-union’ status, so as to enable it to be ready to respond forcefully and quickly to emerging crises, as it had done in the Mexican crisis. (See also Fischer, 1999.) With such firm IMF action, currency crises could be contained as liquidity crises rather than becoming solvency crises. Indeed, it appears that the combination of large-scale IMF financing, combined with significant adjustment by the authorities, prevented the development of a solvency crisis in Mexico. However, some authors began to warn that, if the IMF always acted as a lender of last resort in the face of crisis, then this might create moral hazard on the part of lenders to emerging markets, who might expect to be able to lend virtually risk-free with any possibility of default prevented by IMF action. (The Fund-led bailout of *tesobonos* holders strengthened these fears.) These critics suggested that efforts be made to make sovereign debt rescheduling easier and more orderly (Eichengreen and Portes, 1995), thereby containing the threat of creditor moral hazard.

4.3 The Asian financial crisis of 1997 – 98: the ‘ third generation ’ of crises

Two and a half years later these issues re-emerged in Asia, in a crisis which interrupted a long period of sustained economic growth financed by exports and foreign capital inflows. Unlike the earlier Latin American debt crisis, or even in Mexico, fiscal profligacy played no *explicit* part in the East Asian crisis. But there were two other main policy failings. (See Bluestein, 2001; Corbett and Vines, 1999a; 1999b;

Corbett, Irwin and Vines, 1999.)

First, much more than in Mexico, an under-developed financial system and over-protected financial sector in some Asian economies meant that the private sector had to rely on borrowing, rather than equity issuance, to raise investment funds. As a result, firms became highly leveraged, but banks continued to lend because they were underpinned by *implicit* government guarantees. When growth slowed, as it first did in Thailand in 1996, and then in other East Asian economies, these banks were exposed to the inability of borrowers to repay loans.

Second, a further difficulty arose, as so many times before, from the existence of fixed exchange-rate systems in some East Asian economies, but with a new twist. Banks financed much of their domestic corporate lending by borrowing in foreign exchange from abroad, often at shorter maturities than those employed when they lent onwards in domestic currency. Very little of this borrowing was hedged as a result of the implicit guarantee on the exchange rate. As noted in the previous paragraph, the financial sector was already in difficulty after the initial slow down in growth in 1996. Currencies fell in mid- to late 1997 because of foreign investors' concerns about these difficulties; as a consequence, widespread bankruptcies and potential bank failures loomed because of the unhedged foreign-currency obligations. Fear grew that fiscal systems would be unable to bear the cost of large-scale bank rescues (Irwin and Vines, 2003).

The East Asian debacle marked the advent of 'third-generation' crises in which currency crises and banking crises are intimately intertwined – situations in which vulnerabilities in the private balance sheet can quickly translate into a public debt crisis.

As in Mexico, the Fund played a large part in resolving the crises. The IMF moved quickly to lend very large sums to Thailand, Korea and Indonesia. Nevertheless, there has been widespread criticism of the Fund's behaviour before and after the crisis. (See, for example, Stiglitz, 2002.)

Two difficulties must be acknowledged in the Fund's *crisis prevention* work in East Asia. First, the Fund may have underestimated the risks associated with capital account liberalization. Second, the Fund may not have been firm enough in warning of the difficulties inherent in maintaining a fixed exchange-rate peg. Nevertheless, Thailand, for instance, was warned privately by the Fund several times in the year leading up to the 1997 currency crisis. The Fund, like some private-sector analysts, saw problems looming in Thailand, but its advice was not heeded.

Concerning the Fund's work on *crisis management*, there are three points to consider.

First, as the Fund has acknowledged in both its own reviews of the East Asian crisis and in the evaluations performed by its Independent Evaluation Office (IEO) (IMF, 2003), its programmes may have placed too much emphasis on tightening budgets in countries that were already running prudent fiscal policies. Stanley Fischer, then the Fund's First Deputy Managing Director (FDMD), argues, however, that this approach was driven by a need to boost government savings to support the current account and provision for the impending cost of bank restructurings. (See Fischer, 2004.) Furthermore, the credibility of an adjustment programme at a time of crisis may hinge on policy erring towards being too tight, in order to send a clear signal to markets. Once the scale of the economic downturn became apparent in East Asia and current account balances improved, Fischer argues that the Fund programmes shifted to addressing structural problems. (See also Corden, 1999; Boorman et al., 2000.)

Second, monetary policy was also tightened in an attempt to defend currencies. There is an inevitable trade-off between raising interest rates in order to moderate exchange rate depreciations and lowering

interest rates so as to ease the stress on both the banking system and on corporations that depended on domestic credit. Stiglitz (2002) argues that the tightening was too forceful. However, it does appear that this tightening was essential in order to stem capital flight. Nevertheless, this tightening was not followed by a concerted move to an inflation-targeting regime of a kind that might have allayed concerns of further depreciation. Hence, pressure on the region's currencies continued. And rather than stimulating recovery, these depreciations proved contractionary, at least initially, owing to their effects on external debt burdens. (See Krugman, 1999.)

Third, the Fund did not have a mandate to declare ‘standstills’ on external debt payments during the crisis. In corporate bankruptcies, standstills force creditors to share in the burden of crisis and agree to reasonable debt reschedulings. In the context of a currency crisis, a standstill mechanism would similarly ‘bail in’ foreign private-sector creditors and then make reschedulings possible to reduce debt to sustainable levels. The fact that a standstill was not imposed in Thailand, Korea or Indonesia enabled creditors to race to get their assets out of these countries. Negotiations with foreign creditors to Korea and Indonesia did ensure some rollover of existing short-term lending, with effects similar to those that might have resulted from standstills. In both cases, however, negotiations were pursued too late and without sufficient coordination to maximize their impact (though they did stave off collapse in Korea). The only comprehensive brake on external payments was that imposed in Malaysia through the implementation of capital controls rather than a standstill by the government of Prime Minister Mahathir bin Mohamad in late 1998, a move that contravened the Fund's advice. But this was done only after substantial capital outflows from Malaysia had already taken place.

Because the Fund lacked a mandate to impose standstills, it lent countries money in an attempt to allay the concerns of foreign creditors and to stem capital flight. Given the scale of the external capital-account movements in these countries, the size of IMF financing packages soared, especially after it became clear that smaller lending programmes would be unlikely to produce adequate results. In the case of Korea, the authorities of the IMF's large shareholder governments, notably the United States and Japan, also made a key decision to pursue a debt rollover plan and to exert moral suasion on creditor banks. These banks presumably realized that the alternative would have been partial default. The IMF played a useful role in facilitating communication among the different actors, in providing information, and in certifying that the policies to be pursued by the Korean authorities were appropriate. The IMF's Independent Evaluation Office writes, ‘No single national government, nor any private sector institution, could have played this role as effectively’ (IMF, 2003, p. 115).

Although the Fund's work in Korea showed that the IMF could effectively manage a debt workout, its conduct elsewhere in the East Asia crisis had the effect of shifting the balance of power in debt workouts back toward creditors. IMF programmes did *not* reduce the debt overhang in Indonesia and Thailand. Instead, governments rescued banks and corporations by shifting their debt to the public balance sheet. Taxpayers in these countries still bear the burden of this debt. Rather than ‘bailing in’ private creditors, the Fund's handling of the crisis in these countries may have provided creditors with an even bigger bailout than they might have expected under the terms established in the 1990s’ Brady Plan.

Partially out of dissatisfaction with this result, Anne Krueger, who followed Fischer as the Fund's FDMD in 2001, proposed a bankruptcy or standstill procedure for countries, the ‘Sovereign Debt Restructuring Mechanism’ (SDRM) (Krueger, 2002). The US Treasury and financial markets both opposed this proposal out of a concern it would create unrestrained debtor moral hazard. Under what came to be known as the ‘Taylor Doctrine’ (after John Taylor, then US Treasury Under Secretary for

International Affairs), the US government argued that countries should be left on their own to negotiate with their creditors. But this is only feasible when the number of external creditors is small, which for most countries has not been the case since the 1980s when external borrowing was provided mainly under loans from banks. To help remedy this problem, the US supported the introduction of ‘collective action clauses’ (CACs) in bond contracts with commercial creditors. These clauses prevent rogue creditors from holding out in restructuring negotiations in order to extract a premium from the bond issuer; they work by enforcing a restructuring if a pre-specified minimum proportion of creditors have agreed to its terms. CACs do not, however, provide a framework to guide the allocation of losses between borrowers and lenders, which is necessary in any restructuring. In the absence of a clear means of sharing these losses, it may prove impossible to renegotiate debt owed to commercial creditors. When faced with debt-servicing problems, debtor countries may then decide to borrow from official sources (including the IMF, whose debt is senior to other external liabilities and not reschedulable) in order to repay private sector creditors, as happened in Korea, Thailand and Indonesia. Since private-sector creditors are likely to believe that this will happen, the Taylor doctrine's approach, even when coupled with CACs, might promote creditor moral hazard, something which has been feared ever since the Mexican crisis. Thus, although the Taylor doctrine's approach has the virtue of minimizing debtor moral hazard, it appears to go in the opposite direction by promoting creditor moral hazard.

4.4 Default: the Russian and Argentine crises

Russia. The fall of the Berlin wall in 1989 and the dissolution of the Soviet Union in 1991 enabled the IMF at last to become a (nearly) universal institution. In three years, membership increased from 152 countries to 172, the most rapid increase since the influx of African members in the 1960s. The IMF supported programmes in most former Eastern Bloc countries and newly independent ex-Soviet Republics to help ease the transition to a market economy. The contribution the IMF made to the speed and relative smoothness of this transition is, perhaps, one of its most singular and least-heralded achievements.

Russia, however, got off to an inauspicious start under the first stand-by arrangement with the Fund in 1992. The IMF encountered intense difficulties in influencing the Russian leadership (Odling-Smee, 2004). GDP fell for several years under the IMF-supported combination of macroeconomic stabilization and industrial restructuring. Although the IMF can claim credit for helping to instil some monetary discipline by the mid-1990s, the process took time, foreign direct investment remained low, tax collection was poor, and the fiscal deficit remained large. Growth in real GDP did re-emerge by 1997. But, following the onset of the East Asian crisis, the ruble came under speculative attack in November 1997. Pressure on the ruble was compounded by foreign investors’ attempts to hedge their ruble holdings, as well as by a drop in the price of oil, which accounted for about one-third of Russia's foreign-exchange inflows.

Russia sought additional IMF financing in early 1998, but agreement on the terms of a new programme could not be reached owing, in part, to a failure by the Russian authorities to secure an increase in fiscal revenue. As a result, foreign investors began to unload Russian assets and about US$4 billion fled the country in the summer of 1998. By the time additional IMF financing was agreed in July 1998, fears of a devaluation led to such a pronounced sell-off of Russian securities that the authorities were forced to devalue the ruble and halt payments on both domestic and foreign debt.

Although the Fund is routinely criticized for providing cover for private capital flight from Russia in the first half of 1998, private investors who maintained faith that the Fund would rescue Russia sustained even greater losses when the ruble was devalued. This was perhaps the largest case to that point where the Fund stepped away from a floundering member, declared a solvency crisis, and let private creditors sustain substantial losses. It marked a different approach to the challenge of balancing creditor and debtor interests from that which the Fund had adopted in East Asia. And in some ways it set a precedent for the Fund's handling of the Argentine crisis in 2001.

Argentina. After a sustained period of hyperinflation in the 1980s, Argentina decided in 1991 to peg its currency, the peso, to the US dollar under a quasi currency-board regime at a one-to-one parity. Although the Fund cautioned that Argentina had neither the fiscal discipline nor the robust export sector needed to sustain such a system, it went along with the authorities' plans and supported their macroeconomic programme under a series of lending arrangements. By the late 1990s, Argentina was widely hailed as a model of successful economic reform as the rate of inflation fell to single digits and growth increased. In addition, the economy had successfully weathered the global turbulence caused by the East Asian crisis of 1997–8, and the Russian crisis of 1998.

But the seeds of the problems identified by the Fund back in the early 1990s were beginning to bear fruit by the end of the decade. Fiscal policy remained insufficiently tight owing to the lack of effective central government control on provincial borrowing, and this stimulated domestic demand for imports. Argentina's export sector remained too small to finance these imports, and its real exchange rate made its goods uncompetitive on regional and international markets. As a result, Argentina chose to borrow substantial amounts in US dollars to finance its imports. Brazil's decision to float the real in 1999 in response to pressure from the Russian crisis made it even harder for Argentina to compete under its quasi currency-board regime. The Argentine authorities allowed the peso to float in January 2002, and it quickly collapsed from parity with the US dollar to an exchange rate of nearly 3.9 to the dollar in June 2002. Output fell sharply, inflation reignited, the government defaulted on its debt, and the banking system was largely paralysed.

The Argentine debacle rightly cast several doubts on the Fund's conduct of both crisis prevention and crisis management in emerging markets. At the outset of the 1990s, the Fund proved incapable of resisting Argentina's arguably doomed effort to impose its quasi currency board. Subsequently, the Fund endorsed Argentina's exchange rate peg in a series of programmes through the 1990s that coincided with an accumulation of macroeconomic vulnerabilities. When the regime became unsustainable in 2001 (or earlier), the Fund maintained lending until the end of that year in an attempt to save the peg. After the crisis, the Fund resumed lending to an insolvent Argentina in 2003 at the behest of the Executive Board, even although misgivings were expressed by the Fund staff. IMF lending ceased again later in 2003 and Argentina pursued an aggressive 'take it or leave it' strategy with private creditors. The Argentinean authorities achieved a roughly 75 per cent write-down on the country's defaulted foreign bonds, while leaving nearly US$20 billion in unexchanged bonds in default (IMF, 2005a).

The Fund's experience with Argentina demonstrates at least four things. First, it can be very difficult for Fund staff to resist Executive Board pressure to support a country with IMF lending, either when inappropriate policies are being pursued (for example, the creation of the quasi currency board) or when a country is insolvent (as Argentina was by 2003). Second, the Fund has sometimes found it just as hard as its members to take a stand against an inappropriate fixed-exchange-rate regime. Third, the absence of any international standstill process or debt restructuring mechanism makes it difficult and time

consuming to reconstruct a financial system and to reach a balanced solution with creditors once a crisis has occurred. The Taylor doctrine has not worked out wholly as planned. Fourth, once damaged, the quality of the policy dialogue between the Fund and its members is difficult to restore. Since the crisis, Argentina's policies have appeared unsustainable: Argentina has contrived to keep its exchange rate at a level at which its exports seem to be excessively competitive, while relying heavily on high international primary commodity prices to sustain its balance of payments. These policies do not seem consistent with the world envisaged in the second amendment of the Fund's Articles, a world in which the Fund exercises firm surveillance over member countries' policies in its role as steward of the international financial system.

4.5 Conclusions

The capital account crises of the 1990s and 2000s represent a new chapter in the Fund's history: they mark a distinct shift from the Fund's previous bread-and-butter work of dealing with current account crises. These capital account crises created new challenges and strains on the Fund – some of which it responded to well, some less so.

On *crisis prevention* the Fund has learned much. After the Mexican crisis it promoted regulatory reform, increased transparency, and better monitoring in emerging market economies. The Fund's Articles prevent it from pronouncing on countries' particular choice of exchange-rate regimes. But in its policy advice the Fund has made clear that the trilogy of floating exchange rates, carefully sequenced liberalization of capital accounts and financial systems, and inflation targeting can work well (Blejer et al., 2001; Corden, 2002; Batini, Kuttner and Laxton, 2005); by contrast, the Fund has given clear advice about the difficulties faced by fixed exchange-rate regimes. The Fund has also attempted to reinvent itself as a lender of 'first resort' through the creation of contingent or 'pre-approved' lending facilities aimed at crisis prevention. These lending windows would provide members with an added incentive to pursue sound policies and a signalling framework under which they could commit to these policies. But the Fund's first effort in this direction – 1999's Contingent Credit Lines (CCL) – expired in 2003 after four years without use, owing to somewhat stringent qualification criteria, less than full automaticity in disbursements, and concerns amongst members that a request for a CCL might send a negative signal to capital markets. New effort was invested in the design of such an instrument, initially called the Reserve Augmentation Line (RAL), during 2006–07.

On *crisis management*, much work has been done to understand better how to construct, balance and sequence macroeconomic policy restraint at a time of crisis. The Fund has developed a detailed debt sustainability framework and complemented its traditional analysis of financial flows with a 'balance sheet approach' to analysing stock imbalances, so as to enable it to understand the financial vulnerabilities of countries. This tool was designed to help Fund staff draw a clearer distinction between liquidity crises and solvency cases. (On this see Irwin and Vines, 2005; Cohen and Portes, 2004; Portes, 2004.) But from the early 1980s onward, the three generations of crises outlined above also threw into sharp relief the problem of moral hazard arising from IMF lending. The need to balance better debtor moral hazard and creditor moral hazard became one of the key challenges facing the Fund in the design of its lending facilities and its accompanying policy responses to crises. This article has highlighted the manner in which the Fund has occasionally oscillated between favouring creditor interests and favouring debtor interests, in an attempt to balance these interests in an acceptable way.

The Fund's experience with crisis management in the 1990s revealed difficulties with Fund conditionality. By then the conditions attached to Fund loans had grown far beyond what had earlier been thought necessary to ensure adequate macroeconomic adjustment, and came to include substantial structural conditionalities. Some of these concerned macroeconomic issues of proper concern to the Fund. But there was also an explicit concern with a range of microeconomic reform issues, and, even more broadly, with poverty-reduction questions. Many observers, including Arriazu, Crow and Thygesen (1999), IFIAC (2000) and Williamson (2000), have questioned the wisdom of this policy creep, although it should be said that, in some cases (for example, poverty reduction), the spread of IMF conditionality reflected the concerns of member countries rather than an attempt by the Fund to expand its mandate. Following member country dissatisfaction with the comprehensive conditionalities included in their programmes (Indonesia's programmes in the late 1990s are particularly relevant cases), there has been much work at the IMF since 2000 on streamlining conditionality, and on pulling back from a range of concerns about structural issues that are not deemed ‘macro critical’. This led to a careful restatement during 2002 of the principles governing the IMF's design and implementation of conditionality, with a view to ensuring that the conditions attached to IMF lending focus only on policies essential to the macroeconomic viability of Fund-supported programmes. (See IMF, 2002a; Boughton and Mourmouras, 2004.)

At the time of the preparation of this article (2007) there was a lull in the frequency of crises, and a significant decline in the volume of Fund lending. The Asian, Russian and Argentinean borrowings which originated in the crises described above had all been repaid. There is a striking parallel here with the end of the 1980s, when the Fund's stock of outstanding loans to emerging markets was also quite modest. At that time, the Latin American arrangements that had originated in the crisis years 1980–83 had been repaid. But, just as then, risks remain; the international community must remain engaged in the task of ensuring that the Fund is prepared to respond to and manage crises when they occur.

Dissatisfaction with the Fund's crisis management in the 1990s and early 2000s cast a long shadow over the Fund's relations with many emerging-market economies, which may have some consequences. A number of East Asian countries, over the ten years following the East Asian crisis, accumulated in excess of a trillion US dollars of reserves. This massive reserve accumulation reflected a persistent excess of saving over investment across these economies, which may, at least in part, represent a conscious choice to amass reserves as a form of self-insurance against future crises. These countries went about a pooling of some of these reserves into a common fund, a process which began in 2000 when ASEAN, Japan, China and the Republic of Korea agreed to set up a bilateral currency swap scheme known as the Chiang Mai Initiative. There were some suggestions that this might one day form the basis of an Asian regional alternative to the IMF that would be designed to help these countries to co-insure and spread risks. But taking this step would require difficult decisions by these countries in order to make surveillance between the pool's members effective and enforceable. And such a common pool of reserves might also create its own form of moral hazard if it were to encourage countries to take excessive risks with foreign borrowing.

5 The IMF and low-income countries

Until the mid-1970s, the Fund's work in its role as coordinator and monitor of the international monetary

system was concerned mainly with monetary, exchange-rate and trade issues. To the extent that the IMF also functioned as a credit union for countries in balance of payments difficulties, its lending focused on the provision of short-term, self-liquidating loans to buttress central banks through temporary balance of payments difficulties. The Fund's cornerstone principle of equal treatment of member countries dictated that finance to low-income countries was provided largely under stand-by arrangements on the same terms as those approved for emerging markets and industrialized countries. The oil crises of the 1970s, however, made it increasingly clear that intractable structural issues in many low-income countries needed to be tackled if balance of payments difficulties were to be addressed. As a result, the 1970s saw a lengthening of the average maturity of stand-by arrangements in both emerging markets and low-income countries, accompanied by the advent of lending on concessional terms, with lower interest rates, to low-income countries. This created some tension between the Fund's essentially monetary character and its deepening role in the provision of longer-term resources in support of broad macroeconomic adjustment in developing countries.

In order to provide member countries with more breathing room to enact structural economic reforms, the Fund created a series of new lending instruments from the mid-1970s onward. The first amongst these, the Extended Financing Facility (EFF), provided greater financing and longer maturities than traditional stand-by arrangements, but its terms were not concessional. The Fund's Articles of Agreement did not provide for the use of IMF resources for concessional lending to a subset of the Fund's membership, and the EFF's market-linked interest rates were identical to those of other Fund arrangements. An EFF did, however, typically carry more stringent conditionality than a stand-by arrangement in response to concerns that the EFF's greater financing implied a need for greater adjustment.

The obstacle to financing concessional lending posed by the Fund's Articles was overcome in the 1970s by the solicitation of donor funds and the sale of a portion of the IMF's gold. Concessional IMF lending began under the 1975 Oil Facility Subsidy Account, in which contributions from 25 countries were used to reduce the interest cost of borrowing from a Fund facility set up to assist countries deemed to have been most severely affected by the sudden rise in oil prices. In the following year, the IMF created a Trust Fund for all low-income countries out of profits from the sale of a portion of the Fund's stock of gold. The Trust Fund offered long-term low-interest loans to low-income countries from 1976 until its resources were fully committed in 1981. Borrowing under the Trust Fund was similar to financing under the first credit tranche: in order to obtain financing, low-income countries had only to demonstrate a balance-of-payments need and explain the efforts they were taking to reduce it.

These new financing windows provided concessional loans to developing countries, but it was feared that the weak conditionality attached to these loans did not induce sufficient adjustment (Boughton, 2001). In the early to mid-1980s prices for many primary commodities collapsed, and several developing countries faced new external balance of payments challenges. The Fund moved to reinvigorate its concessional lending by using the repayments of Trust Fund loans to finance a new round of concessional credit under what, in 1986, came to be known as the Structural Adjustment Facility (SAF). The SAF marked a determined attempt by the Fund to integrate concessionality with conditionality. In part, this twinning of concessionality with conditionality allowed the Fund to lobby for new donor loans and grants, which expanded the SAF some threefold into the Enhanced SAF (ESAF) in 1987.

Boughton (2001) contends that the ESAF became one of the IMF's great success stories, as it allowed

the Fund to send billions of dollars to the world's poorest counties on concessional terms with longer maturities than was possible under previous IMF facilities. (See also Tarp, 1993.) The ESAF also had a catalytic effect on lending from other official creditors, and IMF collaboration with the World Bank and the regional development banks, as well as with, inter alia, the UN, UNICEF, UNDP and bilateral donors, all appeared to improve under the ESAF process (Boughton, 2001). In addition, IMF technical assistance to many developing countries on monetary, fiscal, and trade policy, as well as debt management, also expanded substantially in order to help countries achieve their programme commitments. This increase in technical assistance has been very valuable.

Despite these gains, and even although the ESAF was technically distinct from the Fund's general resources, some critics have charged that the ESAF marked an unfortunate departure from the Fund's monetary focus. Others have questioned the strict conditionality on adjustment agreed under ESAF-supported programmes, especially because some of the structural conditions have appeared to intrude on the traditional territory of the World Bank. In reply it might be said that this has happened partly because the Bank has not proved capable of devising appropriate macroeconomic conditions for its own loans. (See Gilbert and Vines, 2000.)

Despite the Fund's efforts – both to revive its concessional lending in 1986 and 1987, and to increase its accompanying technical assistance – it was clear by 1988 that many low-income countries would find it impossible to grow without debt relief. Under the auspices of the Paris Club of bilateral creditors, a series of progressively more concessional refinancing terms for bilateral debts were agreed from 1988 onward, for both emerging market, and relatively poor, indebted countries. Nevertheless, even with this bilateral debt relief, many low-income countries had trouble meeting the payment obligations on their stand-by arrangements and EFFs. But the absence of a serious lobby of private creditors (most low-income countries' external debt was owed to the Paris Club and other public creditors) may have delayed efforts to find a comprehensive solution to the debt problems of developing countries until the late-1990s.

By the 1990s, the Fund's engagement in low-income countries had become the target of a rising chorus of concern. Some civil society organizations and academics, as well as some low-income governments themselves, contended that IMF conditionality and programme design in low-income countries tended to prioritize adjustment over poverty reduction, growth, and income distribution concerns. This criticism is summarized by Easterly (2005). It arose despite the fact that the Fund has been helping to produce, in many low-income countries, a marked stabilization in macroeconomic indicators, and in some cases the beginning of sustained periods of growth. In response to critics' concerns, and in a further step in the evolution of Fund lending, IMF Managing Director Michel Camdessus advocated in the mid-1990s a fresh model of engagement with low-income countries in which there would be a renewed role for the Fund in reducing global poverty and in promoting high-quality growth in developing countries.

This new strategy featured three main elements. First, along with bilateral donors and other international financial institutions, the Fund recognized that catalysing growth in low-income countries would require more profound debt relief, including treatment of previously unrescheduled multilateral concessional debt. The 1996 Heavily Indebted Poor Countries' (HIPC) Initiative represented the concerted efforts of the international community to address the external debt overhang in poor countries; the Initiative was later enhanced in 1999 to provide deeper and faster debt reduction. The HIPC Initiative was novel, particularly in that debt relief was explicitly tied to plans to spend debt-service savings on poverty-alleviating social expenditure. From 1999, these plans were articulated in a country-based Poverty

Reduction Strategy Paper (PRSP). This approach, initiated by the Fund in conjunction with the World Bank, formed the second prong of the Fund's renewed engagement with low-income countries. The PRSP approach aimed to provide a clear country-owned link between national policy frameworks, donor support, and development outcomes. The PRSP approach also dovetailed neatly with the United Nations' Millennium Development Goals (MDGs). These goals were articulated at the UN Millennium Summit in 2000 and were centred on halving global poverty by 2015. The PRSPs were also intended to form the basis of the targets and policy conditions in programmes supported by the IMF's Poverty Reduction and Growth Facility (PRGF). This was the successor in 1999 to the ESAF and formed the third element of the Fund's new approach to low-income countries.

The results of these initiatives by the early 21st century were mixed. Reviews of the PRGF by IMF staff in 2002 (IMF, 2002b) and by the IMF's IEO in 2004 (IMF, 2004b) found that PRGF-supported programs had become more accommodating to higher public expenditure, in particular pro-poor spending. Nevertheless, a review of PRGF programme design by the IMF Executive Board in September 2005 (IMF, 2005c) found that per capita income and growth rates remained low despite some improvements in a range of macroeconomic indicators. More recently, the IEO found in its evaluation of Fund engagement in sub-Saharan Africa (IMF, 2007b) that the PRGF and PRSP approaches had not had a significant positive effect on catalysing new aid flows. This is despite the fact that commitments to increase such flows were made in 2002 under the 'Monterrey Consensus' and at the Gleneagles G8 summit in 2005. The IMF's Spring 2007 *Regional Economic Outlook* noted, however, that Sub-Saharan Africa's growth performance since 2004 had been the best in more than three decades (IMF, 2007d). In sum, the impact of the PRGF and PRSP on aid and spending in low-income countries remained inconclusive, but their growth effects appeared increasingly positive by 2007.

The advent of the HIPC Initiative, the PRSP and the PRGF together intertwined the work of the IMF and World Bank in developing countries to an unprecedented extent. The Multilateral Debt Relief Initiative (MDRI) agreed at the Gleneagles G8 Summit in 2005, and which provided a framework for the write-off of nearly all remaining HIPC-country debts to the IMF, World Bank and African Development Bank, represented a major step forward in this collaboration. While the MDRI drew a welcome line under the multilateral debt relief process, it left several questions about the next phase of IMF and World Bank support for low-income countries unanswered. Having written off so much concessional debt, the MDRI implied that future multilateral support for low-income countries should be provided only as grants, not loans. The source of financing for such grants remained unclear. And in some cases, financing, whether by grants or loans, may not be the most crucial contribution that the international financial institutions could make to development. The Fund's 2005 Policy Support Instrument (PSI), essentially a 'no money' programme, acknowledged that Fund macroeconomic advice, rather than short-term balance of payments financing, might be a valuable channel of support for developing countries. These matters have been complicated by the growth of 'South–South' flows in development assistance from new donors such as China and Brazil. These flows have raised doubts about the future necessity of concessional financing from the Bretton Woods institutions. But they have also called into question the conditionality that comes attached to IMF and World Bank money. Such financing from non-traditional donors could also complicate future debt restructurings, should they prove necessary, since most new donors have not been members of the Paris Club.

Throughout this section we have noted the latent tension between the Fund's monetary character and its long-term support for low-income countries. This tension is heightened by the intertwining of the work

of the Fund and the World Bank, which we have just reviewed. The report of the external review committee on Bank–Fund collaboration (IMF, 2007c) provided some suggestions on strengthening Bank–Fund collaboration, while reducing overlap between the two institutions.

6 The Future of the IMF: next steps

In mid-2004 the Fund's Managing Director, Rodrigo de Rato, launched a review of the role of IMF in light of the challenges posed by a changing and increasingly complex global economic system. Stemming from this review, De Rato presented the aforementioned Medium-Term Strategy for the Fund (IMF, 2005b) to the World Bank–IMF Annual Meetings in September 2005, and shortly thereafter followed up with a plan for the Strategy's implementation (IMF, 2006b). The plan focused on specific proposals to ensure that the Fund:

- provides more effective surveillance and better monitoring of policies in advanced economies, with a renewed emphasis on exchange rates;
- provides better monitoring of emerging markets economies, re-explores financing mechanisms to help prevent crises, and reconsiders issues regarding capital account liberalization;
- enhances the role of IMF in low-income countries, and sharpens its focus;
- reforms IMF governance, particularly country representation; and
- restructures the IMF's own budget, including by broadening the Fund's income base, and its management practices.

The plan also expressed an intention to expand the role of the IMF as a provider of technical assistance and training, while improving Fund communications and transparency to ensure that the Fund would play a more central role in global policy debates.
The Fund's Medium-Term Strategy is a clear response to the three dominant tasks it has assumed following the collapse of the Bretton Woods system of fixed exchange rates in 1971, tasks which we have reviewed in Sections 3–5 of this article. But if the Fund is to be able to act effectively in relation to these tasks it will need to have: (i) a better system of governance; (ii) a more secure and robust source of income so that it can cover its operating expenses; and (iii) a larger stock of resources to lend for crisis prevention and resolution. We conclude this article by briefly discussing these three issues. (See also Lane, 2006.)

6.1 Governance

The first subsection of the Fund's Articles of Agreement made clear that its founding purpose was ‘to promote international monetary co-operation through a permanent institution which provides the machinery for consultation and collaboration on international monetary problems’. At the time of the Fund's creation, most countries stood a reasonable chance of alternating between being a creditor to and borrower from the Fund over time. Since then, the ranks of creditors and borrowers have diverged as industrial countries have stopped using IMF financing, a role which has instead been filled by emerging market economies and low-income countries. A number of reformers such as Woods (2006) argue that

the Fund's capacity to facilitate solutions to international monetary problems depends on the Fund's decision-making structure being made more reflective of the interests and voices of the emerging markets and developing countries which borrow from it, and which see their public policy frameworks at least partly determined by Fund conditionality. The demand for such reform is bolstered by the fact that the relative distribution of quotas, which determine the voting power in the Fund, has become separated from the relative economic (and political) weights of many emerging markets in the global economy. In addition, the relative power of basic votes, which were intended to provide some measure of fairness to poorer countries, has been substantially eroded relative to the contribution of quotas to voting weights at the Executive Board. The ad hoc provision of increased quota shares to China, Korea, Mexico, and Turkey in 2006 under the Fund's Medium-Term Strategy was a first step toward realigning voting power in the Fund with emerging markets' growing share of the world economy; further steps will be more difficult since increased voting shares for some countries will inevitably mean painful decisions to reduce the shares of others. It may, however, be possible for countries to change the way in which the 24 chairs on the IMF's Executive Board are allocated in order to compensate partly for changes in relative voting shares.

Changing the Fund's voting structure would not in and of itself alter the way in which the Fund operates, suddenly making it better able to deliver on the objectives set out in its 2005 Medium-Term Strategy. De Gregorio et al. (1999); King (2006a, p. 12); Dodge (2006a; 2006b) and Kenen (2006) have all argued, however, that parallel changes in the Fund's governance arrangements might help the Fund in its push towards these objectives.

One proposal would put the responsibility for the delivery of improved policies more firmly in the hands of the management of the IMF. Up to 2007, the Executive Board of the Fund had involved itself in day-to-day reviews of Article IV reports, approved all lending decisions, and reviewed the design of the Fund's lending programmes. Stepping back from this activity would enable Directors to pay proportionately more attention to strategic issues. That would move the governance structure of the Fund closer to the relationship between management and advisory boards that one sees in the private sector, where non-executive directors bring dispassionate external views to broad questions of corporate operations and strategy, and clearly delegate day-to-day operations to management.

Evolution in this direction could strengthen the accountability of the Managing Director and his Deputies. In one version of this type of arrangement, all of the Managing Director, the Deputy Managing Directors, and Department Directors would report on a regular basis to the Board, but Executive Directors would be more removed from many of the day-to-day decisions of the institution. Doing this could have an effect – even if only implicit or indirect – on the Fund's ability to function better in its pursuit of more dispassionate surveillance. It might also lead to more effective crisis prevention and resolution through a careful balancing of debtor moral hazard and creditor moral hazard in Fund lending; and also to a clearer focus in the Fund's work with low-income countries.

A move to a non-resident Executive Board would draw a clearer line between the work of Directors and management. Such a move would leave the Managing Director in control of the execution of the Fund's work since the Executive Directors would give only part-time oversight and direction. Making this change would take the governance of the Fund closer to Keynes' original vision. (See King, 2006a.) Directors would be the senior public servants that steer policy in their national capitals, and not, as in 2007, their proxies resident in Washington. In contrast with 1946, the ease of modern travel makes a non-resident Board, with meetings some six to eight times a year, entirely feasible. Any move in this

direction would, however, need to ensure that the nexus of communication between capitals, which the Board currently provides, is preserved in some other way.

6.2 Income

In May 2006, the Managing Director established a committee (the ‘Crockett Committee’), chaired by a former General Manager of the Bank for International Settlements, Andrew Crockett, to study options for sustainable long-term financing of the IMF. The Committee's report, released on 31 January 2007 (IMF, 2007a), argued that the IMF's current funding model was unsustainable and that a more diversified income stream needed to be developed in order to guarantee the institution's financial future. The IMF's revenue stream had been primarily based on income derived from its lending for crisis resolution (IMF, 2007a, Annex 2, p. 2). This financing mechanism was not entirely appropriate, because, as Crockett said during the press briefing to launch the Committee's report, ‘it's a concentrated income source … It's volatile, because when the Fund is lending a lot …it generates large resources. When the Fund is not lending, it doesn't generate resources.’ In a low-lending environment, as existed in the early 21st century, the Fund's income model appeared untenable over the longer term; in the shorter term, it could also be inconsistent with sound incentives to minimize moral hazard in Fund lending.

The Committee considered some alternative sources of income for the Fund. In assessing these possibilities, the committee observed that the Fund's activities could be broken down into three types of functions that cut across the full membership of industrialized countries, emerging markets, and low-income economies: financial intermediation, the provision of global public goods (for example, data, standards and codes, and combating terrorist financing), and the provision of bilateral services, in the form of capacity building and technical assistance.

The Committee concluded that revenue from Fund lending should be sufficient to cover its ongoing costs arising from financial intermediation. The Committee also noted that this income should not be used to cross-subsidize the provision of global public goods because (i) this income was too volatile for this purpose and (ii) cross-subsidization could cause IMF lending to become too expensive compared with private financing.

In order to ensure that the Fund could continue to provide its key global public goods, the Committee noted that the Fund could, like the United Nations, assess a periodic levy on member countries. The Committee did not, however, favour this source of income, as it ‘would risk politicising the activities of the Fund’ by making its work subject to regular financing calls. Nevertheless, the Committee did note that charging fees for some services might generate a small amount of additional revenue.

The Committee's core proposal concerned the creation of an endowment for the IMF that would provide a reliable income stream without relying on annual requests to member countries. The Committee suggested a further sale of IMF gold as a possible source of endowment funds. Such sales had been mooted at various points in the past for a variety of purposes; this was done to finance the establishment of the trust funds that underwrote the 1996 HIPC Initiative. But other plans for such sales have usually failed to gain enough support in the face of opposition from the United States and from gold-producing countries. To allay these fears, the Committee report suggested a ‘balanced’ approach, in which the Fund would also invest some of its quota resources in highly rated securities so that the burden of creating an investment endowment would not fall exclusively on the sale of gold.

As this article was being drafted, discussion was continuing on the exact form an endowment for the

Fund could take. In meantime, the Fund had begun to invest some of its retained earnings from lending in investment grade securities in an effort to supplement its income.

6.3 Resources

The relative size of the Fund shrank markedly from the 1970s ownard in comparison with, inter alia, global reserves, international trade, financial flows, stocks of financial assets and world output. This decline in pecuniary stature has distorted some of the debates about the Fund's work, most notably on creditor and debtor moral hazard. Much of the debate over the implications of jumbo or ‘exceptional access’ arrangements in the 1990s (arrangements in which lending was equivalent to 300 per cent of quota or more) would be moot if regular quota increases had maintained the Fund's relative size in the global economy. Indeed, had the Fund grown through regularly scheduled quota increases, very few of the arrangements of the 1990s and 2000s would have been deemed at all exceptional. This suggests a simple yardstick for an appropriately-sized IMF: at any given time, the sum of the Fund's quotas should enable a risk-adjusted subset of its membership to borrow from the Fund on non-exceptional terms to finance their adjustment needs.

Accepting the validity of such a yardstick depends critically, however, on one's ultimate view of the role the IMF should play in the international system: trusted macroeconomic advisor, catalyst for private capital inflows and foreign assistance, or potential lender of last resort at time of crisis? To some extent the Fund played all of these roles at the turn of the 21st century, though its reduced relative size meant that the lender-of-last-resort function was credible only for its smaller members. The Fund staff, its shareholders, and those who care about the future of the multilateral system will need to decide which of these roles the IMF should continue to play.

See Also

- Bretton Woods system
- currency crises
- development economics
- emerging markets
- Keynes, John Maynard
- monetary approach to the balance of payments
- World Bank

Bibliography

Agénor, R., Miller, M. Vines, D. and Weber, A., eds. 1999. *The Asian Financial Crises: Causes, Contagion, and Consequences*. Cambridge: Cambridge University Press.

Arriazu, R., Crow, J. and Thygesen, N. 1999. *External Evaluation of IMF Surveillance*. Washington, DC: IMF. Online. Available at http://www.imf.org/external/pubs/ft/extev/surv/eval.pdf, accessed 6 June

2007.

Batini, N., Kuttner, K. and Laxton, D. 2005. Does inflation targeting work in emerging markets? In *World Economic Outlook*, September. Washington, DC: IMF. Online. Available at http://www.imf.org/external/pubs/ft/weo/2005/02/pdf/chapter4.pdf, accessed 6 June 2007.

Blejer, M., Leone, A.M., Raubanal, P. and Schwartz, G. 2001. Inflation targeting in the context of IMF-supported adjustment programs. Working Paper No. 01/31. Washington, DC: IMF.

Bluestein, P. 2001. *The Chastening: Inside the Crisis that Rocked the International Financial System and Humbled the IMF*. New York: Public Affairs.

Boorman, J., Lane, T., Schulze-Ghattas, M., Bulir, A. et al. 2000. Managing financial crises: the experience in east Asia. Working Paper No. 00/107. Washington, DC: IMF.

Boughton, J. 2001. *Silent Revolution: The International Monetary Fund 1979–1989*. Washington, DC: IMF.

Boughton, J. 2004. The IMF and the force of history: ten events and ten ideas that have shaped the institution. Working Paper, No. 04/75. Washington: IMF.

Boughton, J. and Mourmouras, A. 2004. Whose programme is it? Policy ownership and conditional lending. In *The IMF and its Critics: Reform of Global Financial Architecture*, ed. D. Vines and C. Gilbert. Cambridge: Cambridge University Press.

Bordo, M.D. 1993. The Bretton Woods international monetary system: an historical overview. In *A Retrospective on the Bretton Woods System, Lessons for International Monetary Reform*, ed. M. Bordo and B. Eichengreen. Chicago and London: University of Chicago Press.

Calvo, G. and Goldstein, M. 1996. What role for the official sector? In *Private Capital Flows to Emerging Markets After the Mexican Crisis*, ed. G. Calvo, M. Goldstein and E. Hochreiter. Washington, DC: Institute for International Economics.

Calvo, G. and Mendoza, E. 1996. Reflections on Mexico's balance of payment crisis: a chronicle of a death foretold. *Journal of International Economics* 41, 235–64.

Cohen, D. and Portes, R. 2004. Towards a lender of first resort. Discussion paper No. 4615. London: CEPR.

Corbett, J., Irwin, G. and Vines, D. 1999. From Asian miracle to Asian crisis: why vulnerability, why collapse? In *Capital Flows and the International Financial System*, ed. D. Gruen and L. Gower. Sydney:

Reserve Bank of Australia. Repr. in Irwin and Vines (2001).

Corbett, J. and Vines, D. 1999a. Asian currency and financial crises: lessons from vulnerability, crisis, and collapse. *World Economy* 22, 155–77.

Corbett, J. and Vines, D. 1999b. The Asian crisis: lessons from the collapse of financial systems, exchange rates, and macroeconomic policy. In Agénor et al. (1999).

Corden, W.M. 1983a. Is there an important role for an international reserve asset such as the SDR? In *International Money and Credit: The Policy Roles*, ed. G.M. von Furstenberg. Washington, DC: IMF.

Corden, W.M. 1983b. The logic of the international monetary non-system. In *Reflections on a Troubled World Economy*, ed. F. Machlup, G. Fels and H. Muller-Groeling. London: Macmillan.

Corden, W.M. 1990. Macroeconomic adjustment in developing countries. In *Public Policy and Economic Development*, ed. M. Scott and D. Lal. Oxford: Clarendon Press.

Corden, W.M. 1993. Why did the Bretton Woods System break down? In *A Retrospective on the Bretton Woods System*, ed. M. Bordo and B. Eichengreen. Chicago: University of Chicago Press.

Corden, W.M. 1999. *The Asian Crisis: Is There a Way Out*? Singapore: Institute of South Asian Studies.

Corden, W.M. 2002. *Too Sensational: On the Choice of Exchange Rate Regimes*. Cambridge, MA: MIT Press.

Corden, W.M. 2007. Those current account imbalances: a sceptical view. *World Economy* 30, 363–82.

Dam, K. 1982. *The Rules of the Game*. Chicago: University of Chicago Press.

De Gregorio, J., Eichengreen, B., Ito, T. and Wyplosz, C. 1999. *An Independent and Accountable IMF*. Geneva Reports on the World Economy, No. 1. London: CEPR.

de Vries, M.G. 1976. *The International Monetary Fund 1966–71: The System under Stress*. Washington, DC: IMF.

de Vries, M.G. 1985. *The International Monetary Fund 1972–78: Cooperation on Trial*. Washington, DC: IMF.

de Vries, M.G. 1987. *Balance of Payments Adjustment, 1945 to 1986: The IMF Experience*. Washington, DC: IMF.

Dodge, D. 2006a. Global imbalances: why worry? What to do? Speech made to the New York Association for Business Economics, 29 March. *BIS Review*, 4 April. Online. Available at http://www.bis.org/review/r060404a.pdf, accessed 19 June 2007.

Dodge, D. 2006b. The evolving international monetary order and the need for an evolving IMF. Speech given at the Woodrow Wilson School of Public and International Affairs, 30 March. *BIS Review*, 5 April. Online. Available at http://www.bis.org/review/r060405a.pdf, accessed 19 June 2007.

Dooley, M., Landau, F. and Garber, P. 2003. An essay on the revived Bretton Woods system. Working Paper No. 9971. Cambridge, MA: NBER.

Easterly, W. 2005. What did structural adjustment adjust? The association of policies and growth with repeated IMF and World Bank adjustment loans. *Journal of Development Economics* 76, 1–22.

Edwards, S. 1989. The international monetary fund and the developing countries: a critical evaluation. In *IMF Policy Advice, Market Volatility, Commodity Price Rule, and Other Essays*, ed. K. Brunner and A.H. Meltzer. Carnegie-Rochester Conference Series on Public Policy. Amsterdam: North-Holland.

Eichengreen, B. 1992. *Golden Fetters: The Gold Standard and the Great Depression 1919–1939*. Oxford: Oxford University Press.

Eichengreen, B., ed. 1995. *Europe's Postwar Growth*. New York: Cambridge University Press.

Eichengreen, B. 2004. The dollar and the new Bretton Woods system. Henry Thornton Lecture, delivered at the Cass School of Business, London, 15 December.

Eichengreen, B. 2007. *The European Economy Since 1945: Coordinated Capitalism and Beyond*. Princeton: Princeton University Press.

Eichengreen, B. and Park, Y.C. 2006. Global imbalances and emerging markets. Paper presented to the Asia-Europe Economic Forum Conference in Beijing on 13 July, 3, 7–68.

Eichengreen, B. and Portes, R. 1995. *Crisis, What Crisis? Orderly Workouts for Sovereign Debtors*. London: CEPR.

Finch, C.D. 1989. The IMF: the record and the prospects. Essays in International Finance No. 175. Princeton: International Finance Section, Department of Economics, Princeton University.

Fischer, S. 1999. On the need for an international lender of last resort. *Journal of Economic Perspectives* 13(4), 85–104.

Fischer, S. 2004. *IMF Essays from a Time of Crisis: The International Financial System, Stabilization, and Development*. Cambridge, MA: MIT Press.

Gardner, R. 1956. *Sterling Dollar Diplomacy*. Oxford: Oxford University Press.

Gilbert, C. and Vines, D., eds. 2000. *The World Bank: Structure and Policies*. Cambridge: Cambridge University Press.

Gilbert, C. and Vines, D. 2004. The IMF and international financial architecture: liquidity and solvency. In Vines and Gilbert (2004).

Goldstein, M. 2006. Currency manipulation and enforcing the rules of the international monetary system. In *Reforming the IMF for the 21st Century*, ed. E. Truman. Special Report No. 19. Washington, DC: Institute for International Economics. Online. Available at http://www.iie.com/publications/chapters_preview/3870/05iie3870.pdf, accessed 6 June 2007.

Gwin, C., Sachs, J., Kenen, P., Feinberg, R. and Nelson, J. 1990. *The International Monetary Fund in a Multipolar World: Pulling Together*. US–Third World Policy Perspectives No. 13, Overseas Development Council. New Brunswick: Transaction Books.

Horsefield, J.K. 1969. *The International Monetary Fund 1945–65: Twenty Years of International Monetary Cooperation*, 3 vols. Washington, DC: IMF.

IFIAC (International Financial Institution Advisory Commission). 2000. *Report of the International Financial Institution Advisory Commission*. Washington, DC: IFIAC, US Congress.

IMF (International Monetary Fund). 1987. *Theoretical Aspects of the Design of Fund-Supported Adjustment Programs*. Occasional Paper No. 55. Washington, DC: IMF.

IMF. 2002a. Guidelines on conditionality. Prepared by the Legal and Policy Development and Review Departments, 25 September. Online. Available at http://www.imf.org, accessed 30 May 2007.

IMF. 2002b. Review of the poverty reduction and growth facility (PRGF) and the poverty reduction strategy paper (PRSP) Approach, 15 March. Online. Available at http://www.imf.org, accessed 30 May 2007.

IMF. 2003. *The IMF and Recent Capital Account Crises: Indonesia, Korea, Brazil*. Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Online. Available at http://www.imf.org, accessed 6 June 2007.

IMF. 2004a. Policy formulation, analytical frameworks and policy design. Prepared by the Policy

Development and Review Department. Online. Available at http://www.imf.org, accessed 30 May 2007.

IMF. 2004b. *Evaluation of the IMF's Role in Poverty Reduction Strategy Papers and the Poverty Reduction and Growth Facility*. Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Online. Available at http://www.imf.org, accessed 6 June 2007.

IMF. 2005a. *Argentina: 2005 Article IV Consultation – Staff Report*. IMF Country Report No. 05/236. Washington, DC: IMF.

IMF. 2005b. The Managing Director's report on the fund's medium-term strategy, 15 September. Online. Available at http://www.imf.org, accessed 30 May 2007.

IMF. 2005c. Review of PRGF Program Design – Overview. 8 August. Policy Development and Review Department. Online. Available at http://www.imf.org, accessed 6 June 2007.

IMF. 2006a. *Multilateral Surveillance*. Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Online. Available at http://www.imf.org, accessed 6 June 2007.

IMF. 2006b. The Managing Director's report on implementation of the fund's medium-term strategy. 5 April. Online. Available at http://www.imf.org, accessed 30 May 2007.

IMF. 2007a. Committee to study sustainable long-term funding of the IMF: final report. Online. Available at http://www.imf.org, accessed 6 June 2007.

IMF. 2007b. *The IMF and Aid to Sub-Saharan Africa*. Report of the Independent Evaluation Office of the IMF. Washington, DC: IMF. Available at http://www.imf.org, accessed 6 June 2007.

IMF. 2007c. Final report of the external review committee on bank–fund collaboration. February. Washington, DC: IMF. Online. Available at http://www.imf.org, accessed 6 June 2007.

IMF. 2007d. *Regional Economic Outlook: Sub-Saharan Africa*. Washington, DC: IMF. Online. Available at http://www.imf.org/external/pubs/ft/reo/2007/AFR/ENG/sreo0407.htm, accessed 6 June 2007.

Irwin, G. and Vines, D., eds. 2001. *Financial Market Integration and International Capital Flows*. Cheltenham: Edward Elgar.

Irwin, G. and Vines, D. 2003. Government guarantees, investment, and vulnerability to financial crises. *Review of International Economics* 11, 860–74.

Irwin, G. and Vines, D. 2005. Policies for the resolution of international financial crises: how to avoid moral hazard. *International Journal of Money and Finance* 10, 233–50.

Joshi, V., Lane, T. and Vines, D. 2006. The US, East Asia and Europe: how to achieve an orderly resolution of global imbalances. Mimeo, Oxford University.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1971–1988. *The Collected Writings of J.M. Keynes*, ed. E. Johnson and D. Moggridge. London: Macmillan.

Kenen, P. 2006. *Comments on the address of the Managing Director of the IMF*. Washington, DC: Institute for International Economics. Online. Available at http://www.iie.com, accessed 30 May 2007.

Killick, T. 1986. *The Quest for Economic Stabilisation: The IMF and the Third World*. Aldershot: Gower.

King, M. 2005. The international monetary system. Remarks at Advancing Enterprise 2005 Conference, London, 4 February. Online. Available at http://www.bankofengland.co.uk, accessed 30 May 2007.

King, M. 2006a. Reform of the international monetary fund. Speech given to the Indian Council for Research on International Economic Relations. New Delhi, 20 February. Online. Available at http://www.bankofengland.co.uk, accessed 30 May 2007.

King, M. 2006b. Through the looking glass: reform of the international institutions. Inaugural International Distinguished Lecture, Melbourne Centre for Financial Studies, Australia, 21 December. Online. Available at http://www.bankofengland.co.uk, accessed 30 May 2007.

Krueger, A. 2002. A new approach to sovereign debt restructuring. Address given at the Indian Council for Research on International Economic Relations, Delhi, 20 December. Online. Available at http://www.imf.org, accessed 6 June 2007.

Krugman, P. 1999. Balance sheets, the transfer problem, and financial crises. *International Tax and Public Finance* 6, 459–72.

Kumar, M. 2006. Economic policy cooperation and coordination: analytical issues and historical experience. Mimeo, Fiscal Affairs Department, IMF.

Lane, T. 2006. Tensions in the role of the IMF and directions for reform. *World Economics* 6(2), 47–66.

Little, I.M.D. 1993. *Macroeconomic Analysis and the Developing Countries, 1970–1990*. Occasional

Paper No. 41. San Francisco: International Centre for Economic Growth.

Little, I.M.D., Cooper, R., Corden, W.M. and Rajapatirana, S. 1993. *Boom, Crisis and Adjustment: The Macroeconomic Experience of Developing Countries*. Oxford: Oxford University Press.

Matthews, R.C.O. and Bowen, A. 1988. Keynesian and other explanations of postwar macroeconomic trends. In *Keynes and Economic Policy*, ed. W.A. Eltis and P.J.N. Sinclair. London: NEDO.

Matthews, R.C.O., Feinstein, C.H. and Odling-Smee, J.C. 1982. *British Economic Growth 1956–73*. Oxford: Clarendon Press.

McKibbin, W. 1997. Empirical evidence on international economic policy coordination. In *Handbook of Comparative Economic Policies Volume 5: Macroeconomic Policy in Open Economies*, ed. M. Fratianni, D. Salvatore and J. Von Hagen. London: Greenwood Press.

Meyer, L., Doyle, B., Gagnon, J. and Henderson, D. 2004. International coordination of macroeconomic policies: still alive in the new millennium? In *The IMF and its Critics: Reform of Global Financial Architecture*, ed. D. Vines and C. Gilbert. Cambridge: Cambridge University Press.

Moggridge, D. 1986. Keynes and the international monetary system 1909–1946. In *International Monetary Problems and Supply Side Economics: Essays in Honour of Lorie Tarshis*, ed. J.S. Cohen and G.C. Harcourt. London: Macmillan.

Mussa, M. and Savastano, M. 1999. The IMF approach to economic stabilization. In *NBER Macroeconomics Annual 1999*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Odling-Smee, J. 2004. The IMF and Russia in the 1990s. Working Paper No. 04/155. Washington, DC: IMF.

Polak, J. 1957. Monetary analysis of income formation and payments problems. *IMF Staff Papers* 6 (November), 1–50.

Portes, R. 2004. Resolution of Sovereign debt crises: the new old framework. Discussion paper No. 4717. London: CEPR.

Roubini, N. and Setser, B. 2005. Will the Bretton Woods 2 regime unravel soon? The risk of a hard landing in 2005–2006. Unpublished manuscript. Online. Available at http://pages.stern.nyu.edu/~nroubini/papers/BW2-Unraveling-Roubini-Setser.pdf, accessed 19 June 2007.

Sachs, J. 1995. Do we need an international lender of last resort? F.D. Graham Lecture at Princeton University, 20 April.

Skidelsky, R. 2000. *John Maynard Keynes: Fighting for Britain, 1937–1946*. London: Macmillan.

Stiglitz, J. 2002. *Globalization and Its Discontents*. New York: W.W. Norton & Company.

Tarp, F. 1993. *Stabilization and Structural Adjustment: Macroeconomic Frameworks for Analysing the Crisis in Sub-Saharan Africa*. London: Routledge.

Temin, P. 2002. The golden age of European growth reconsidered. *European Review of Economic History* 6, 3–22.

van Dormael, A. 1978. *Bretton Woods: Birth of a Monetary System*. London: Macmillan.

Vines, D. 2003. John Maynard Keynes 1937–1946: the creation of international macroeconomics. *Economic Journal* 113, F338–61.

Vines, D. and Gilbert, C., eds. 2004. *The IMF and its Critics: Reform of Global Financial Architecture*. Cambridge: Cambridge University Press.

Williamson, J. 1977. *The Failure of World Monetary Reform 1971–74*. New York: New York University Press.

Williamson, J. 1983a. Keynes and the international economic order. In *Keynes and the Modern World*, ed. G.D.N. Worswick and J. Trevithic. Cambridge: Cambridge University Press.

Williamson, J., ed. 1983b. *IMF Conditionality*. Washington, DC: Institute for International Economics.

Williamson, J. 2000. The role of the IMF: a guide to the reports. Policy Brief 00–5. Washington, DC: Institute for International Economics.

Williamson, J. 2006. The potential of international policy coordination. Paper presented at a seminar on Implications for the IMF's Role in Surveillance and Policy Coordination, at a Roundtable on International Economic Cooperation for a Balanced World Economy, Chongqing, China, 12–13 March. Online. Available at http://www.iie.com, accessed 30 May 2007.

Wolf, M. 2005. Multilateral leadership can right the ship. *Financial Times*, 28 June.

Wolf, M. 2006. The world needs a tough and independent monetary fund. *Financial Times*, 22 February.

Woods, N. 2006. *The Globalisers: The IMF, the World Bank and their Borrowers*. Ithaca and London: Cornell University Press.

How to cite this article

House, Brett, David Vines and W. Max Corden. "International Monetary Fund." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000183>
doi:10.1057/9780230226203.0831

The New Palgrave Dictionary of Economics Online

international monetary institutions

Stanley W. Black

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International monetary institutions are required to support payments arrangements between countries with different currencies and exchange rate arrangements. Reserve assets and adjustment and financing mechanisms are provided to assist markets in balancing conflicting objectives including economic growth and price stability, growing international trade and payments, and convertibility of currencies at reasonably stable exchange rates. The evolution of the Bretton Woods system has proceeded through floating exchange rates, increased capital mobility, financial crises, and various reform proposals. The development of regional monetary institutions has led to creation of the European Monetary Union and some steps towards increased Asian monetary cooperation.

Keywords

Asian Development Bank; Association of South East Asian Nations (ASEAN); Bretton Woods system; capital controls; competitive devaluation; conditionality; convertibility; European Central Bank; European Currency Unit (ECU); European Monetary System (EMS); European Monetary Union (EMU); European Payments Union; exchange rate risk; fixed exchange rates; flexible exchange rates; foreign exchange controls; General Arrangements to Borrow (IMF); gold exchange standard; gold standard; inflation; international financial adjustment; international liquidity; International Monetary Fund; international monetary institutions; international reserves; Maastricht Treaty (EU); monetary policy; pegged exchange rates; political risk; reserve assets; Smithsonian Agreement (1971); special drawing rights (SDRs); Stability and Growth Pact (EU); Tchebychev's inequality; Triffin, R.

Article

Domestic money is conceived of by society as a device to facilitate transactions in the marketplace, as a temporary store of value, and as a unit of account for contracts. Given the possibilities of fraud and counterfeiting, domestic monetary authorities have been established to regulate the quality of the

domestic monetary unit in most countries. Such regulations attempt to guarantee the interchangeability of the different media, such as currency and the deposits of different banks, as well as stability in the value of the monetary unit, under conditions of prosperity.

International monetary arrangements are required under conditions of international trade when residents of different countries must make payments to each other, and yet wish to hold most of their assets in terms of domestic currency. Such arrangements are designed to guarantee *convertibility* of assets denominated in different currencies, so that payments may be made independent of country of residence, thus facilitating a free and open trading system. International monetary *institutions* such as the International Monetary Fund are designed to support international monetary arrangements by enforcing rules of behaviour, assisting countries in difficulties, and encouraging good practices.

Alternative exchange rate mechanisms

Under a *gold standard*, domestic residents and foreign residents may freely convert domestic currency into gold at a fixed rate of exchange. This type of convertibility was eliminated in the 1930s in favour of a gold *exchange* standard, which allowed only foreign monetary authorities to exchange domestic currency for gold. Gold convertibility of both types was ended as part of the Smithsonian Agreement of 1971 (see below).

Under a system of *pegged* exchange rates between different currencies, as established by the Bretton Woods system (see below), convertibility implies that domestic residents are free to obtain foreign currency at a *fixed* rate of exchange for the purchase of foreign goods and services, inclusive of normal trade credit. Likewise, foreign residents are free to sell domestic currency obtained by sale of goods and services or to use it for purchase of domestic goods and services, at the same fixed rate of exchange. This definition does not require free convertibility for capital account transactions (those arising from exchanges of financial assets only).

Under a system of floating or flexible exchange rates, convertibility still implies that both domestic and foreign residents may freely convert domestic and foreign currency at the same rate of exchange for current account transactions, but the exchange rate at which this may be done is determined on a daily basis by market transactions, rather than being guaranteed by the domestic monetary authorities of the respective countries.

In 2005, only 20 out of the 184 member countries of the International Monetary Fund (IMF) declined to accept the obligations to current account convertibility. But in a large number of countries various types of restrictions limited convertibility in some way or created differences in the exchange rates applying to exports and imports. Non-unified exchange rates lead to inefficient allocation of resources, as previously documented by Bhagwati (1978). For example, 70 countries required repatriation and surrender of proceeds of exports or invisible transactions, 57 countries had payments arrears of one kind or another, and 11 countries maintained either dual or multiple exchange rates for different types of transactions. With respect to capital account transactions, the situation is much more restrictive: 126 countries had controls on international transactions in capital market securities, and 143 countries maintained controls on direct investment flows.

Reserve assets

In order to guarantee convertibility of the domestic currency into other convertible currencies, monetary authorities hold stocks of *reserve assets*, which are liquid assets held in readily accepted international media of exchange, such as dollars, euros, and a few other currencies. In addition, IMF member countries have access to unconditional borrowing rights to obtain additional reserve assets in the form of their reserve positions in the Fund and Special Drawing Rights. These, together with reserve asset holdings, make up *international liquidity*.

Since most international payments are handled by inter-bank transactions, banks have sought to minimize transactions costs by channelling their foreign exchange transactions through one or more *vehicle* currencies, the pound sterling in earlier days, but more recently the US dollar and to some extent the euro. Because the dollar is so widely used in private exchange transactions, monetary authorities also find it convenient to operate in dollars to ensure the convertibility of their currencies.

Adjustment mechanisms

The existence of different national currencies and the need to maintain convertibility of the different currencies lead to the concept of balance of payments adjustment mechanism. At a given exchange rate, as long as the amount of foreign exchange earned through exports of goods and services and capital inflows just pays for imports and capital outflows, no external imbalance exists. If international capital markets were perfect and if investors were risk neutral so that assets denominated in different currencies were perfect substitutes for one another in private portfolios, there would in practice be a single world interest rate for short-term borrowing. Then imbalances between foreign exchange earnings and payments could simply be *financed* by borrowing in the international capital market. There would be no real distinction between the convertibility characteristics of the official liabilities of different borrowers. But, in fact, countries face very real limits on the amount of foreign currency they can borrow abroad in exchange for domestic currency because of *exchange rate* risk, which limits the willingness of risk-averse foreign lenders to acquire domestic currency assets. According to the doctrine of *original sin*, countries with a history of convertibility problems are unable to issue foreign debt in their own currency (Eichengreen and Hausmann, 2005). The ability to repay foreign currency debt is dependent on balance of payments adjustment. *Political* risk involves the possibility that exchange controls may be imposed in the future, preventing the repayment of foreign currency debt on the promised terms. Thus it is desirable for countries to have access to a variety of adjustment mechanisms to eliminate external imbalances, as well as a variety of sources of official financing in the form of international liquidity. The primary mechanisms of balance of payments adjustment are through movements in exchange rates and adjustments of income and price levels via monetary and fiscal policies. The need for adjustment can be postponed by imposition of tariffs and subsidies, quantitative restrictions on current account or capital account transactions, or controls over the allocation of foreign exchange. But tariffs, quantitative restrictions, and exchange controls generally involve inefficiencies in the allocation of resources, including in the latter case loss of convertibility of the domestic currency. Changes in monetary and fiscal policies or exchange rates have their own costs in terms of domestic policy objectives forgone.

Financing

Thus, a mixture of adjustment policies and financing mechanisms is provided in a system of international monetary arrangements. *Official* financing is provided either by drawing on holdings of official reserve assets or by borrowing from international institutions. *Private* financing can be arranged by a monetary authority borrowing from foreign banks or the international bond market. Either provides the ability to postpone adjustment. The optimum mix of adjustment and financing for an individual country depends on the costs of the various alternatives. By setting the costs of these alternatives, international monetary arrangements influence the behaviour of the world economy.

A model of adjustment versus financing

In the theory of adjustment versus financing, a country is faced with random balance of payments deficits and surpluses, which it may either finance by drawing on reserve assets or adjust by one of the adjustment mechanisms mentioned above. In one branch of the theory, due to Heller (1966) and others, the cost of adjustment is assumed to be a linear function of the size of the adjustment, so that any adjustments are postponed to the last minute, at which time full adjustment takes place. Alternatively, one may assume a nonlinear cost of adjustment, leading to a theory of partial adjustment. Kelly (1970) and Clark (1970) assume that the country's welfare function depends on the mean and variance of income, so that gradual adjustments are preferred. The analysis determines both the optimum level of reserve holdings, R^*, and the optimum rate of adjustment α to that level, according to the equation

$$\Delta R = \alpha [R^* - R_{-1}] + u, \tag{1}$$

where u is normally distributed with mean zero and variance σ^2 and R_{-1} is the stock of reserves at the end of the previous period. This equation assumes that changes in the stock of reserves arise from both the random shocks in the balance of payments and the desired rate of adjustment to the optimal level of reserves. From eq. (1) we find that the variance of reserve holdings decreases as the speed of adjustment α increases from zero to one.

Tchebychev's inequality then enables one to show that, for a given probability of not exhausting reserves and given opportunity cost r of holding reserves, the optimum reserve holding R^* decreases with increasing α. As α increases, the need for more frequent adjustments raises the variance of income. Therefore the speed of adjustment should be chosen such that the welfare loss from increased variance in income due to a small increase in α is just counterbalanced by the welfare saving due to holding slightly smaller reserves.

According to this theory, international monetary institutions will strongly affect the behaviour of national policies concerning balance of payments adjustment and acquisition of reserves. Specifically, international money institutions will determine the opportunity cost of holding reserves, the penalty attached to running out of reserves, and the availability of different types of adjustment policies. By influencing countries’ balance of payments adjustment policies, international institutions will also

influence their domestic policies, since there is a trade-off between internal and external objectives of policy.

The role of markets and institutions

An optimal design for the international monetary system depends on balancing among a group of conflicting objectives: growth of real income and employment, stable prices, efficient allocation of resources, maintenance of convertibility of currencies, improving the distribution of income, and growth of world trade. The relevant trade-offs can be understood in the context of an economic model. According to the model of adjustment and financing outlined above, reductions in the opportunity cost of holding reserves will lead to increased reserve holdings, a reduction in the speed of adjustment to imbalances, increased use of financing, and a decline in the variability of income. The slowdown in the speed of adjustment implies a change in the allocation of resources among countries. The increased use of financing may imply an increase in the rate of inflation. An optimal international system should balance these various considerations. For discussion of efforts to design such a system, see Solomon (1982) and the documents of the IMF's Committee of Twenty (IMF, 1974).

In a purely laissez-faire system, market borrowing instead of official reserves would be the source of financing to postpone adjustment. Fluctuations in market interest rates would determine the terms of trade between adjustment and financing. As is usual in market solutions, the wealthy are in a better position to negotiate terms on loans. By contrast, a more institutionalized system provides access to financing at lower rates to those with a weaker market position, with more conditions on the use of the funds. Evaluating the difference between two such systems is a complex task. For an attempt, see Jones (1983).

The evolution of international monetary institutions

Between the close of the Napoleonic Wars and 1880, the international monetary system gradually moved onto the gold standard, which was fully achieved during the period 1880–1914. Under the leadership of Great Britain, sterling operated as a vehicle currency during this period, allowing an efficient international payments mechanism to develop. The increasing substitution of bank deposits for currency allowed an ever-larger volume of payments to be supported by a gradually rising supply of gold. Despite the best efforts of the Bank of England and other central banks, periodic crises interfered with the continued convertibility of individual currencies. And the system was characterized by substantial fluctuations in employment and prices, albeit about a rising trend of employment with no trend in prices.

Following the First World War, gold convertibility was resumed on a limited basis, until the Great Depression of 1929–33 brought it to an end. A period of fluctuating exchange rates, competitive devaluations, and increasing use of trade restrictions to promote domestic employment ensued. It is generally believed that the economic difficulties of the interwar period were major factors bringing on the Second World War.

The Bretton Woods system

The United States and Great Britain took the lead in constructing the post-war international monetary institutions, with Harry Dexter White and John Maynard Keynes drawing up rival designs for the new system agreed at the Bretton Woods Conference in 1944. The Articles of Agreement of the International Monetary Fund provided for a system based on pegged, but adjustable, exchange rates and an institution which would lend reserve assets to countries that were having temporary difficulties in maintaining convertibility. Resort to floating exchange rates, competitive devaluations, and trade restrictions to promote domestic employment were explicitly to be avoided, in the light of the problems of the 1930s. Convertibility for current account transactions was promoted, while capital account convertibility was required only for those transactions necessary for financing current payments.

The lending power of the IMF was based on *quotas* of gold and domestic currency contributed by each member country. Only the gold was to be paid in initially, but, if the Fund needed convertible currency to lend out, it would obtain it from any member whose currency was considered strong enough to be *usable*. Members could borrow automatically up to the amount of the gold portion or *tranche* of the quota, but only on demonstration of balance of payments need, and thereafter they could borrow more subject to meeting conditions on economic and financial policies. For further discussion of IMF policies, see Williamson (1983), Kenen (2001), and Truman (2006).

The initial post-war problem involved the establishment of a payments system that would promote economic recovery and the growth of trade among the former combatants. The International Monetary Fund limited itself to establishing a set of agreed par values for pegged exchange rates which could promote the growth of trade, leaving the provision of loans and grants for economic recovery to the United States, the strongest economy. Under this system, which was a form of gold exchange standard, countries declared their par values in terms of the US dollar, which in turn was convertible into gold at $35 an ounce. Thus the dollar became the key currency of the system, and most foreign exchange reserves came to be held in the form of dollars. Within Europe, convertibility remained limited until 1958, and the European Payments Union was established to facilitate intra-European payments. The re-establishment of convertibility led to fears that the IMF might have inadequate resources to deal with the problems of large member countries. In 1962 the General Arrangements to Borrow were created, to enable the Fund to mobilize additional resources from its largest members, the Group of Ten.

With the recovery of the European economies in the 1950s and the achievement of convertibility in 1958, the US dollar became gradually overvalued relative to gold and other currencies. As Robert Triffin (1960) pointed out, the key currency system required the United States to continue to run balance of payments deficits in order to supply other countries with increased foreign exchange reserves. As it did so, the gold reserve of the United States became increasingly inadequate to guarantee gold convertibility of growing US official dollar liabilities at $35 an ounce.

A variety of solutions to this problem were proposed, including the creation of an artificial reserve asset to substitute for dollars, an increase in the dollar price of gold, and the adoption of floating exchange rates. In 1968 the First Amendment to the Articles of Agreement of the International Monetary Fund permitted the creation of Special Drawing Rights (SDRs), which have twice been allocated to member countries in proportion to their existing quotas in the Fund. SDRs, when utilized, permit the user to acquire convertible currencies from other members, upon the payment of interest. They represent a centralized mechanism for increasing the stock of reserves. By the early 1970s the gold convertibility of the dollar was under increasing pressure, for a variety of reasons. In August 1971 the dollar was

unilaterally set loose from gold. The Smithsonian Agreement of December 1971 attempted to save the Bretton Woods system by multilateral realignment of exchange rates, including a devaluation of the dollar against gold and a widening of the narrow bands of fluctuation permitted around the newly fixed values. Some members of the European Communities (EC) agreed to maintain narrower margins of fluctuations versus each other's currency, in an arrangement that became known as the ‘EC Snake’. Despite these efforts, the revised Bretton Woods system lasted only a little more than a year.

Floating exchange rates

In March 1973, exchange rates of most of the major industrial countries began floating. At the same time, most developing countries continued to peg their currencies to the dollar or another developed country currency, and the EC maintained the ‘Snake’. About this time, a major effort to reconstruct international monetary institutions on the basis of pegged exchange rates began under the auspices of the IMF's Committee of Twenty. This effort collapsed in 1974, in part under the impact of the quadrupling of world oil prices by the Organization of Petroleum Exporting Countries.

In Jamaica in January 1976, the Interim Committee of the Board of Governors of the International Monetary Fund agreed on a Second Amendment to the Fund's Articles of Agreement, ratifying the system of floating exchange rates. First, stability of exchange rates was to be sought through stability of underlying monetary and fiscal policies rather than through pegging. Second, floating rates should be subject to a process of ‘firm surveillance’ by the IMF. Third, it was hoped that the SDR would ‘become the principal reserve asset’, with the role of gold and the dollar being reduced. Fourth, the fixed official price of gold was abolished and one-third of the IMF's gold was disposed of. Acceptance of the status quo was all that could be accomplished. The result, according to Corden (1983), was an international laissez-faire system.

In 2005 some 88 countries made use of floating exchange rates, while 51 had pegged exchange rates of one type or another and 48 operated within currency unions with other countries.

Increased capital mobility, the Asian crisis and reform proposals

Beginning in the 1970s, international capital mobility increased significantly, as middle-income developing countries found new access to foreign borrowing and industrialized countries increasingly opened production facilities in each others’ markets. In the early 1990s, the IMF began discussions of a possible amendment that would promote capital account convertibility as an additional goal of the international monetary system, on the argument that improved allocation of capital would lead to increased economic growth. But a series of crises in emerging market economies interfered with this project, most notably the Asian financial crisis of 1997, followed by the Russian crisis of 1998 and the Argentine crisis of 2001. Each of these events was preceded by substantial capital inflows seeking higher returns, which overwhelmed under-regulated and underprepared domestic economies and financial systems. The convertibility of affected currencies was often temporarily impaired (Black, Christofides and Mourmouras, 2006). In some cases the IMF was seen as creating a permissive environment prior to the crisis, followed by harsh demands for domestic reforms subsequently, in attempts to restore confidence and bring an end to capital outflow.

A substantial body of criticism on one side argued that, by its willingness to provide large amounts of financing to countries in crisis, the Fund had created ‘moral hazard’, encouragement to over-borrowing and over-lending in expectation of a bailout (International Financial Institution Advisory Commission, 2000). On the other side, others claim that the Fund by its harsh requirements for reform was stifling economic recovery and growth (Stiglitz, 2002). Both of these viewpoints may have had some validity, but in a sense they cancel each other out (see Kenen, 2001). The Fund itself proposed creation of an international Sovereign Debt Restructuring Mechanism to assist defaulting countries in negotiations with creditors (Krueger, 2003). This was rejected in favour of a more modest approach encouraging the use of collective action clauses in bond indentures requiring minority bondholders to accept terms of repayment agreed to by a majority.

Another criticism of the IMF is that its voting shares and representation appear outdated, as compared with the changing economic importance of different groups of countries (Truman, 2006). In particular, large emerging market economies such as China, India, and Brazil are under-represented, while the European Union countries with 32 per cent of the voting power are over-represented. Obviously, changes in representation are extremely difficult to achieve, but will still be necessary to remedy a situation in which the rich creditor countries that do not utilize the Fund's resources have disproportionate voting power relative to the debtor nations that have greater need for use of its facilities.

The ‘new’ Bretton Woods and Asian monetary cooperation

Following recovery from the Asian crisis of 1997, countries such as Korea, China, Malaysia, Taiwan and India sharply increased their accumulations of international reserves, as developing Asian countries in total raised their reserves (minus gold) from SDR 414 billion to SDR 1,039 billion between the ends of 1998 and 2004. China, Hong Kong and Malaysia in particular sought to maintain exchange rates pegged to the US dollar, while the other countries managed their floating exchange rates so as to avoid undue appreciation against the US dollar, accumulating enormous reserves in the process. An influential paper by Dooley, Folkerts-Landau and Garber (2004) argued that this relationship was a new version of the old Bretton Woods system, whereby other countries pegged their exchange rates to the US dollar, enabling the United States to run large current account deficits, while the creditor nations increased their exports to the United States. Alternatively, the vastly increased reserve holdings of Asian countries could be regarded as a precautionary response to insure the availability of financing to avoid the prospect of another sharp adjustment, following the unpleasant experiences of the 1997 Asian crisis.

The combination of increased regional reserve holdings and recent bad experience with internationally supervised adjustment has led Asian countries to embark on steps towards regional monetary cooperation, culminating in the so-called Chiang Mai Initiative for regional currency swaps among the Association of South East Asian Nations (ASEAN) plus China, Japan, and Korea (see Park and Wang, 2005). ASEAN members realized that the industrial countries of the Group of Ten had previously used currency swaps among central banks to lend each other money in times of crisis and thus avoid the need for borrowing from the IMF with its conditionality. With growing availability of reserves in Asia, the ASEAN+3 concluded that they might similarly help each other out in future. Under the leadership of the Asian Development Bank, further steps are contemplated, possibly including an Asian Monetary Fund and an Asian Currency Unit.

The European Monetary Union

The enlargement and strengthening of the EC ‘Snake’ in 1978, which was in the process renamed the European Monetary System (EMS), gradually led to the creation of the European Monetary Union with a unit of account, the European Currency Unit (ECU). The objectives of the enlarged EMS were to reduce intra-European exchange rate fluctuations, to promote convergence of macroeconomic policies within Europe, and to reduce European dependence on US monetary policies. Over a period of 15 years, the EMS succeeded in these objectives, at the cost of a series of exchange rate realignment crises culminating in a major collapse of the system in 1992–3, when the narrow margins (plus or minus $2\frac{1}{4}$ per cent) were expanded (to plus or minus 15 per cent). The crisis was brought on by a combination of increasingly rigid exchange rates within the system, increased capital mobility as a component of the Single Market programme of the European Union, and stresses brought on by the unification of East and West Germany.

In response to these factors, and to further strengthen the integration of European markets and achieve a more symmetrical sharing of decision making in monetary policy, the Maastricht Treaty ratified in 1993 brought into being in 1999 the European Monetary Union, with a single currency, the *euro*, with monetary policy controlled by a European Central Bank (ECB) in Frankfurt, Germany, replacing the currencies of the 12 member countries of the eurozone. While the euro has been quickly accepted as an international currency, in both the member countries and their neighbours, the relatively conservative operations of the ECB together with the constraints on member countries’ fiscal policy embodied in the Stability and Growth Pact have proven controversial in the light of slow economic growth in the eurozone.

The euro is gradually becoming more important in international transactions and in the foreign exchange market as a rival to the US dollar. In 2006 the IMF redefined the SDR currency basket reflecting the importance of currencies in international trade and finance to be composed of 44 per cent US dollars, 34 per cent euro, 11 per cent Japanese yen and 11 per cent pound sterling, as compared with the previous weights of 45 per cent US dollars, 29 per cent euro, 15 per cent yen and 11 per cent pound sterling.

See Also

- capital controls
- gold standard
- international capital flows
- international reserves

Bibliography

Bhagwati, J. 1978. *Anatomy and Consequences of Exchange Control Regimes*. Cambridge, MA: Ballinger.

Black, S. 1985. International money and international monetary arrangements. In *Handbook of International Economics*, vol. 2, ed. R. Jones and P. Kenen. Amsterdam: North-Holland.

Black, S., Christofides, C. and Mourmouras, A. 2006. Convertibility risk: the precautionary demand for foreign currency in a crisis. *Annals of Finance* 2, 141–65.

Clark, P. 1970. Optimum international reserves and the speed of adjustment. *Journal of Political Economy* 75, 356–76.

Corden, W. 1983. The logic of the international monetary non-system. In *Reflections on a Troubled World Economy*, ed. F. Machlup et al. London: Macmillan.

Dooley, M., Folkerts-Landau, D. and Garber, P. 2004. The revived Bretton Woods system. *International Journal of Finance and Economics* 9, 307–13.

Eichengreen, B. and Hausmann, R., eds. 2005. *Other People's Money: Debt Denomination and Financial Instability in Emerging Market Economies*. Chicago and London: University of Chicago Press.

Heller, H. 1966. Optimal international reserves. *Economic Journal* 74, 333–52.

IMF (International Monetary Fund). 1974. *International Monetary Reform: Documents of the Committee of Twenty*. Washington, DC: IMF.

International Financial Institution Advisory Commission. 2000. *Report*. Washington. United States Congress.

Jones, M. 1983. International liquidity: a welfare analysis. *Quarterly Journal of Economics* 98, 1–23.

Kelly, M. 1970. The demand for international reserves. *American Economic Review* 60, 655–67.

Kenen, P. 2001. *The International Financial Architecture: What's New? What's Missing?* Washington, DC: Institute for International Economics.

Krueger, A. 2003. Sovereign debt restructuring: messy or messier? *American Economic Review* 93(2), 70–4.

Park, Y. and Wang, Y. 2005. The Chiang Mai Initiative and beyond. *World Economy* 28, 91–101.

Solomon, R. 1982. *The International Monetary System, 1945–1981*. New York: Harper & Row.

Solomon, R. 1999. *Money on the Move: The Revolution in International Finance Since 1980*. Princeton:

Princeton University Press.

Stiglitz, J. 2002. *Globalization and its Discontents*. New York and London: Norton.

Triffin, R. 1960. *Gold and the Dollar Crisis*. New Haven: Yale University Press.

Truman, E. 2006. *A Strategy for IMF Reform*. Washington, DC: Institute for International Economics.

Williamson, J., ed. 1983. *IMF Conditionality*. Washington, DC: Institute for International Economics.

How to cite this article

Black, Stanley W. "international monetary institutions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000184> doi:10.1057/9780230226203.0832

The New Palgrave Dictionary of Economics Online

international outsourcing

Deborah L. Swenson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International outsourcing involves the import of intermediate inputs or services from unaffiliated foreign suppliers. While it implies that the production of a final product involves production activities in more than one country, this trade in intermediate inputs can be explained by traditional theories of international trade where countries have comparative advantage in different stages of production. However, since outsourcing relationships involve interaction with foreign partners, the choice of organizational form for these transactions is also influenced by industrial organization factors, such as search costs or contract incompleteness. This article discusses these issues and the effects of outsourcing on the international economy.

Keywords

comparative advantage; factor price equalization; foreign direct investment; hold-up problem; incomplete contracts; intermediate inputs; international outsourcing; matching; North–South economic relations; offshoring; search costs; sunk costs; trade costs; thick markets; transport costs; vertical integration

Article

In complex production processes firms face a classic make-or-buy question: should they purchase parts, assembly or services from an outside vendor, or perform those tasks themselves? In the domestic context, the benefits and costs of vertical integration are already well understood. However, declines in international transport costs, advances in remote management technologies and improved communications technologies have brought an international dimension to this question, as they have enabled an increasing number of firms to engage in international outsourcing, purchasing parts, assembly or services from unaffiliated international suppliers.
As with trade in final products, international trade in intermediate inputs is shaped by international

differences in comparative advantage that reflect cross-country differences in factor costs, or relative productivities for different stages of production. Hence, when firms decide where to complete the production activities required for the creation of a final product – design, materials extraction, parts production, and assembly – comparative advantage influences the ideal country placement for each production stage. However, if there are international frictions, such as international transport costs or tariffs, outsourcing imports will emerge only when the international outsourcing benefits stemming from comparative advantage exceed the costs associated with these international frictions.

Broad examination of international trade data, such as that of Hummels, Ishii and Yi (2001), or Yeats (2001) indicates that international trade in intermediate inputs has grown even more rapidly than the generally large growth in international trade since 1960. This trade in intermediate inputs represents both outsourcing purchases from unrelated suppliers and ‘offshoring’, which is the import of parts or services from related overseas suppliers, such as foreign subsidiaries. Yi (2003) argues that two key economic factors explain why recent declines in international trading costs have generated such an exceptional increase in intermediates trade. First, trade in intermediates increases along an extensive margin as declines in trade costs enable products that were previously produced domestically to be more profitably produced through an internationally integrated production process. Second, the effects of declining frictional costs are magnified when intermediates trade involves multiple border crossings, since the benefits of falling tariff or transportation charges apply to each border crossing involved in the creation of the final product.

Nonetheless, while international differences in international factor costs provide an incentive for international outsourcing, differences in international factor costs are not sufficient in themselves to guarantee that outsourcing relationships will develop. Thus, there are two strands in the literature on international outsourcing that explain firm choices. The first emphasizes limits on outsourcing that relate to search costs and matching, while the second focuses on the firm's choice of organizational form when contracts are incomplete.

When the search for an appropriate outsourcing partner is costly, firms will search for a foreign partner if the expected increase in profit generated by the search exceeds the sunk costs of searching for an international partner. Thus, Grossman and Helpman (2005) demonstrate that, when the cost of foreign search is particularly high, firms may choose domestic outsourcing in the high-wage home country over foreign outsourcing in the lower-cost foreign country. In addition, if the appearance of potential outsourcing partners is endogenous, the increased demand for partners in a particular location generates a market thickness externality. Since the entry of potential partners increases the likelihood that searches will be successful, the increase in expected profits in thick markets increases the equilibrium number of searches in the market that becomes more densely populated with suppliers. As a result, search cost frictions, and the market thickness externalities they generate support an international equilibrium in which firms may be indifferent between searching for partners at home and searching for them abroad, even though wages and factor costs are not equalized across countries. In support of these ideas, Swenson (2005) finds that, while costs matter for some US outsourcing decisions, the cost sensitivity is largest for industries that are less capital intense and for industries that have thicker international markets for suppliers. The quality of country institutions, such as the strength and efficacy of a country's legal system, also influence the strength of market thickness externalities, since favourable country institutions increase demand for international outsourcing partnerships, thus increasing entry by outsourcing suppliers.

Even when comparative advantage is sufficiently strong to favour the overseas purchase of intermediate inputs, problems caused by contract incompleteness present a second reason why firms may not choose to engage in international outsourcing. In this case, firms may alternatively choose to integrate vertically with their foreign suppliers, or to set up foreign subsidiaries to conduct and support their purchase of overseas inputs and supplies. Firms are particularly likely to choose such ‘offshoring’ arrangements when problems arising from contract incompleteness are present in an industry that requires significant relationship-specific investments (Head, Ries and Spencer, 2004; Qui and Spencer, 2002). Antras (2003) argues that contract incompleteness will be more problematic in capital-intense industries: an idea that finds empirical support in his observation that the fraction of US imports that are intra-firm is higher for capital-intense industries and for trade with more capital-abundant countries.

Since the set-up of a foreign affiliate often involves substantial foreign direct investment expenditure, not all firms will choose offshoring over outsourcing for their international purchases of intermediate inputs. In this vein, Antras and Helpman (2004) show that, when firms are heterogeneous in their productivity, different sourcing strategies will coexist in equilibrium, as each firm chooses the sourcing method that maximizes its profits. Feenstra and Hanson (2005) provide further evidence of heterogeneous organizational choices in the case of Chinese processing trade, where organizational variation across Chinese industries and Chinese provinces supports their model of firm organization which is based on a property-rights description of the firm. Finally, thick market externalities may generate multiple outsourcing equilibria, as McClaren (2000) describes in a setting where independent input suppliers face a hold-up problem when they develop special components for a specific foreign purchaser of intermediate inputs. Here, an increase in market thickness, due to an increase in the number of final goods firms who search for suppliers, reduces the hold-up problem, since the bid of the next-closest purchaser increases.

Cross-country cost differences influence firms’ international outsourcing decisions. In turn, the growth of international outsourcing may lead to changes in the international equilibrium. First, if country endowments differ dramatically, ordinary trade in final goods may narrow cross-country differences in factor rewards, but fail to bring about factor price equalization. Using traditional trade models, Deardorff (2001) shows that outsourcing may facilitate factor price equalization. Deardorff also shows that outsourcing may reduce a country's welfare if changes in international prices cause a terms of trade loss which reduces a country's gains from trade, as compared with the gains it reaped from trade in final goods only. However, outsourcing will not harm, and may even help, country welfare when international prices are unaffected.

The effect of outsourcing on international factor rewards depends crucially on the nature of the production process. In a three factor world where production of intermediate inputs involves the combination of capital with skilled and unskilled labour, Feenstra and Hanson (1996) demonstrate how outsourcing may exacerbate income inequality in all countries, where income inequality is measured by the compensation of high-skilled relative to low-skilled workers. Intermediate goods are ordered according to their relative use of skilled to unskilled labour, while all intermediate inputs have equal capital cost shares. The production of the final good involves the costless assembly of the full range of intermediate inputs. In this framework, outsourcing brought about by capital flows from the Northern country to the Southern country reduces the relative cost of capital in the south, thus lowering the relative cost of producing each intermediate input in the south. This causes the skill intensity of southern

production to rise as the country begins to produce an expanded range of intermediate inputs, which were previously completed in the North. From the South's perspective, the activities are more skilled-labour intense than their previous set of activities, while the activities were the least skilled-labour intense activities of those that the North produced. Thus, the shift in intermediates production increases the compensation of high-skill relative to low-skill workers in both locations since the relative demand for skilled workers rises in both the North and the South. At a firm level, Head and Ries (2002) observe that the skill level of Japanese workers in Japanese multinationals rose especially rapidly when the Japanese firms imported an increasing portion of their products from low-income, presumably labour-abundant countries. In the end, the fact that international outsourcing provokes such strong political concern reflects the fact that outsourcing, like international trade, has the potential to influence relative factor rewards.

See Also

- foreign direct investment
- incomplete contracts
- international trade theory
- vertical integration

Bibliography

Antras, P. 2003. Firms, contracts and trade structure. *Quarterly Journal of Economics* 118, 1375–418.

Antras, P. and Helpman, E. 2004. Global sourcing. *Journal of Political Economy* 112, 1375–418.

Deardorff, A.V. 2001. Fragmentation in simple trade models. *North American Journal of Economics and Finance* 12, 121–37.

Feenstra, R.C. and Hanson, G.H. 1996. Globalization, outsourcing, and wage inequality. *American Economic Review* 86, 240–5.

Feenstra, R.C. and Hanson, G.H. 2005. Ownership and control in outsourcing to China: estimating the property-rights theory of the firm. *Quarterly Journal of Economics* 120, 729–61.

Grossman, G.M. and Helpman, E. 2005. Outsourcing in a global economy. *Review of Economic Studies* 72, 135–59.

Head, K. and Ries, J. 2002. Offshore production and skill upgrading by Japanese manufacturing firms. *Journal of International Economics* 58, 81–105.

Head, K., Ries, J. and Spencer, B.J. 2004. Vertical networks and U.S. auto parts exports: is Japan

different? *Journal of Economics and Management Strategy* 13, 37–67.

Hummels, D., Ishii, J. and Yi, K.-M. 2001. The nature and growth of vertical specialization in world trade. *Journal of International Economics* 54, 75–96.

McLaren, J. 2000. ‘Globalization’ and vertical structure. *American Economic Review* 90, 1239–54.

Qui, L.D. and Spencer, B.J. 2002. Keiretsu and relationship-specific investment: implications for market-opening trade policy. *Journal of International Economics* 58, 49–79.

Swenson, D.L. 2005. Overseas assembly and country sourcing choices. *Journal of International Economics* 66, 107–30.

Yeats, A.J. 2001. Just how big is global production sharing? In *Fragmentation: New Production Patterns in the World Economy*, ed. S.W. Arndt and H. Kierzkowski. Oxford: Oxford University Press.

Yi, K.-M. 2003. Can vertical specialization explain the growth of world trade? *Journal of Political Economy* 111, 52–102.

How to cite this article

Swenson, Deborah L. "international outsourcing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000105> doi:10.1057/9780230226203.0833

The New Palgrave Dictionary of Economics Online

international policy coordination

Paul R. Bergin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Coordination among national governments as they formulate macroeconomic policies has been proposed as a response to global integration among national markets. Policy coordination may be beneficial by preventing the externalities created by policy spillovers, as well as by promoting international risk sharing. The usefulness of coordination depends upon numerous characteristics of an economy, including the degree of openness in goods and asset markets.

Keywords

asset market integration; beggar-thy-neighbour; Bretton Woods system; Commitment; coordinated solutions; European Central Bank; European Monetary Union; exchange rate targets; fiscal expansion; globalization; goods market integration; information sharing; international capital flows; international migration; international policy coordination; Keynesianism; labour market integration; microfoundations; monetary policy externalities; Nash solutions; policy spillovers; risk sharing; sticky price; terms of trade

Article

Coordination among national governments as they formulate macroeconomic policies has been proposed as a response to global integration among national markets.
Awareness has grown over time of how national macroeconomies are interconnected in a global marketplace. Rising trade volumes indicate international integration among goods markets, large international financial flows indicate integration in asset markets, and highly visible immigration flows reflect increasing integration in national labour markets. Progressive globalization in the private economic sphere has prompted the question of whether public policy likewise should be global. Should the policies that nations use to manage their national macroeconomies be coordinated jointly with other nations? This is not a new question, and economists have voiced a variety of opinions and theories. Most

academic economists have tended to be sceptical about the need for explicit international policy coordination.

To date there is limited coordination of macroeconomic policies in practice. Under the Bretton Woods arrangement of fixed exchange rates following the Second World War II until 1973, the monetary policies of member countries were constrained by the need to maintain an exchange rate target. If a national central bank were to attempt to increase the domestic money supply or lower domestic interest rates as a means of stimulating domestic production, this would tend to lower the value of its national currency relative to others and violate the fixed exchange rate agreement. Since the dissolution of this system in the 1970s, many nations learned to appreciate the resulting freedom to use their monetary policy to pursue domestic objectives.

Nonetheless, over the decades since the end of the Bretton Woods system, economic officials of major industrial countries periodically have met to discuss exchange rate intervention and options for monetary and fiscal policies. Examples include the Plaza and Louvre accords in the 1980s. Without binding public agreements, it is not clear how much coordination takes place at such meetings, and the function served by them may simply be sharing information regarding policy intentions. In some regions of the world, a subset of countries have taken steps on their own to more formally coordinate their policies. The most dramatic form of international macroeconomic policy coordination of late has been the formation of the European Monetary Union in 1999. Eleven initial member countries ceded sovereignty over national monetary policy to a European Central Bank, where a single monetary policy must be agreed upon for the whole region.

The opinions of academic economists on the advisability of policy coordination have varied over time, largely in response to the introduction of new tools of economic analysis. Milton Friedman (1953) and others recommended against explicit coordination, suggesting that private market forces could be trusted to achieve a desirable outcome. In particular, exchange rate movements could serve a useful function of insulating countries against the macroeconomic shocks of their neighbours. In contrast, economists of the 1970s and 1980s were able to find theoretical rationales for policy coordination, using Keynesian models that featured frictions that prevented economic markets from operating efficiently on their own. Finally a renewed interest in the subject since 2000, employing models with more microeconomic foundations, has produced new theoretical reasons to question the usefulness of policy coordination.

The rest of this article considers two primary motivations for policy coordination: preventing policy spillovers and promoting pooling of international risk. The article discusses each motivation in turn along with its limitations.

Policy spillovers

One motivation for policy coordination is the possibility that the effects of policy spill over national borders to affect the macroeconomies of trading partners. For example, suppose there is a global shock that lowers global demand below some desirable level, such as a wave of pessimistic expectations that lowers investment expenditure. This might be undesirable, to the degree the excess inventories may lead to recession, with a scaling back of production and lower levels of employment. Keynesian theory indicates that one way policymakers can combat such a shortfall in demand is through expansionary fiscal policy, with a rise in government expenditure or a cut in taxes to stimulate private consumption demand. However, globalization affects this policy prescription. National policymakers may fail to

respond if they fear that some of the benefit will leak abroad: a fiscal expansion may lead to a currency appreciation, making domestic goods less competitive than foreign goods. As a result, some of the increase in demand generated by domestic government debt will be used to purchase foreign goods and employ foreign workers.

Coordination of policymakers across countries may provide a way of eliminating the problem created by this externality. If a mechanism of coordination existed to make sure that all countries symmetrically expanded government spending, each government could be reassured that it would benefit from spillovers of demand from abroad, to compensate for the negative spillover of demand leaking abroad. A coordinated global fiscal expansion, the theory says, is an effective way of combating a global shortfall in demand.

Externalities also apply to monetary policy. Monetary expansions tend to cause currency depreciations that make domestic goods more competitive compared to foreign goods. The use of such policy to shift demand from foreign goods toward home goods to raise domestic production at the expense of lower foreign production is labelled ‘beggar-thy-neighbour’. One might imagine repeated rounds of such policies, with each country progressively increasing money supply to regain competitiveness. In the end competing policies will have no net effect on the exchange rate and competitiveness, but the net rise in the money supply of each country would produce the undesirable outcome of excessive inflation. Coordination agreements may commit countries to avoid such policy outcomes; they may agree to forswear beggar-thy-neighbour policies if there is a credible commitment from other countries to do the same. The end result is a better outcome for all.

The spillover argument in favour of coordination clearly depends on the degree to which the private economies are interdependent internationally. Consider goods market integration. If exports tend to be a small fraction of a country's GDP, a currency depreciation raising exports a certain percentage will have a small effect on GDP in absolute terms. The international implications of any policy just wouldn't matter very much. Asset market integration also has been found to be important. If asset markets do not view government debt issued by different countries as equivalent, then a fiscal expansion that raises the issue of debt in one currency could cause a currency depreciation rather than an appreciation, reversing the direction of the fiscal spillovers described above.

Policy spillovers and strategic interactions of policymakers are topics introduced in research by Hamada (1974), Oudiz and Sachs (1984), and Canzoneri and Gray (1985). When a Keynesian theoretical model embodying the spillover arguments above was quantified by Oudiz and Sachs (1984), it was found that the gains from coordination were too small to justify the effort. US merchandise exports to Europe at the time amounted only to 1.6 per cent of US GNP. As a result, the gains from coordination were estimated at only about 0.5 percentage points of GDP for the United States. The lesson was that since international integration was actually quite low, there was little or no role for policy coordination. A question that is addressed later in this article is whether this conclusion continues to hold in a progressively more globalized and integrated world.

Policymaker objectives

The relevance of policy spillovers has been qualified by recent research that studies the objectives of policymakers; see Obstfeld and Rogoff (2002), Corsetti and Pesenti (2005), and Canzoneri, Cumby and

Diba (2005). This research features microeconomic foundations to describe the behaviour of consumers, workers, and producers. One benefit of deriving consumer behaviour from assuming they are trying to maximize a particular utility function is that this utility function provides a natural metric by which to evaluate the benefits of alternative policies. Further, it facilitates predictions about how policymakers will act, on the assumption that their behaviour is driven by the goal of improving the welfare of private consumers. For example, one might assume that the policymakers of each country act independently to maximize the utility of citizens in their own country. This 'Nash' solution can be contrasted with a coordinated solution, where an international coordinator chooses the policies of all the countries jointly to maximize the sum of utility of citizens across countries. Only if the outcome of the latter coordinated solution supersedes that of the independent Nash solution is there a clear motivation for international policy coordination.

Consider a simplified theoretical world of two countries populated by representative agents that consume and produce. Production involves labour supplied by these agents, combined with technology that is subject to uncertain shocks each year. Suppose these economies exhibit a market imperfection in the form of prices that must be set ahead of time and that cannot change in response to surprise fluctuations in productivity. Given this environment, imagine there is a negative productivity shock that lowers the level of output. In contrast to the argument of the previous section, it no longer is clear that a policymaker should respond by trying to restore output to its previous level by stimulating demand. This would make the welfare of the citizens even worse, because it would force them to work harder during periods where their labours are less rewarded. Instead, utility is made highest by using monetary policy to replicate the outcome of an economy that is free from the sticky-price market imperfection. In this flexible-price version of the world, citizens would choose to work and consume less during periods of low productivity, and choose to work more and acquire wealth during periods when productivity shocks are favourable.

Although it may seem counter-intuitive, a policymaker wishing to maximize the welfare of his or her citizens in such an economy often will contract the money supply when output falls due to the productivity shock. This has the effect of raising the relative price of home goods and reducing demand and hence production. The outcome of this Nash game differs from the outcome described in the previous section, and does not involve any beggar-thy-neighbour strategy. The domestic policymaker is perfectly capable of replicating the flexible price outcome by the appropriate application of domestic policy. Under certain conditions to be discussed below it turns out that the coordinated solution is identical to that for the Nash solution above. If policy in the two countries were dictated by a central coordinator trying to eliminate all externalities, the set of policies he would prescribe for each country would be identical to the policies that each country would have chosen independently. In this world the spillover argument fails to apply, and there is no benefit from international policy coordination.

International risk sharing

A second type of motivation offered for international policy coordination is the possibility that countries can benefit by mutually insuring each other against the effects of shocks. Ideally private asset markets would include trade in securities contingent on the incidence of shocks, which households could use to insure themselves. For example, in the case of a fall in productivity and output in just one country, such securities would require a transfer of wealth from abroad to this country as a way of buffering the level

of consumption despite the fall in domestic production. International trade in equities could serve this function. Suppose the residents of two countries each own half of the firms in the other country's stock market, and they thereby have claim to half of the output of each other's total production. If a productivity shock lowers the output of the home country but leaves the foreign country unaffected, when each country sends half of its respective production to the other country, this implies a net payoff from the foreign country to the home country. This transfer effectively spreads the impact of the productivity shock over the consumption levels of both countries and acts as a type of insurance. However, in the absence of a private market for such securities, there may be a role for policy coordination to replicate these insurance benefits.

For example, consider again the story above of a negative shock to productivity in one country. Another motivation for the policymaker to employ a contractionary monetary policy is to raise the value of the domestic currency, in order to raise the relative price of its exports to imports, the terms of trade. By making home goods more valuable, he or she raises the revenue from export sales abroad, transferring wealth to the affected country. The ability to manipulate the exchange rate to transfer wealth from the foreign to home country clearly could present a temptation to pursue beggar-thy-neighbour policies. But in the hands of a central policy coordinator, this becomes a means of making transfers between countries when useful for insurance purposes.

Note that a coordinated policy motivated by the objective of risk sharing might in principle conflict with the motivation for coordination laid out in previous sections. There is no reason to suppose that the degree of monetary contraction needed to transfer enough wealth to pool risk is also that degree needed to discourage production to the level consistent with flexible prices. That is, it may not be possible to use policy coordination to offset two economic distortions at the same time, the sticky-price and imperfect risk-sharing distortions. This is a point emphasized in the influential work of Obstfeld and Rogoff (2002).

Extensions to more realistic economies

While recent research has noted additional theoretical rationales for coordination, this may not change the conclusion that the gains are too small quantitatively to justify the effort. When Obstfeld and Rogoff (2002) calibrate with reasonable parameter values a model that combines imperfect risk sharing with nominal rigidities, it does find there is some positive gain from a coordinator choosing a policy as opposed to each country optimizing separately. But the additional benefit from coordination is small. As long as policymakers act wisely to replicate flexible price outcomes in their domestic economy, the benefit of coordinating with foreign countries is smaller by an order of magnitude. Several features of the theoretical economic environment are key to this result. Clearly key is the supposition that policymakers will act in a manner to maximize the welfare of their residents when given the freedom to do so. But also essential are assumptions about the behaviour of consumers, such as the willingness to substitute across home and foreign goods to maintain their level of utility, and a desire to smooth consumption levels over time.

As progressively more realistic economic environments are explored, the list is augmented of economic features that affect the decision to coordinate. One such feature is the nature of price stickiness. When exporting firms set their prices, many will set them in the currency of the buyer's market. If prices are

sticky in the local currency, any fluctuations in the nominal exchange rate will have no effect on the price that consumers face in the market. So any attempt to use monetary policy to manipulate the terms of trade as an insurance device will fail. As demonstrated in Devereux and Engel (2003), local currency pricing kills off a primary motivation for policy coordination as well as the temptation to pursue beggar-thy-neighbour policies in a sticky-price world.

On the other hand, some other realistic economic features tend to augment the benefits of coordination. These include the reliance on imported goods as intermediates in the domestic production process, in which case random fluctuations in the exchange rate can severely disrupt domestic production. Such issues are likely to be most important for small economies, especially those that specialize in assembly operations of imported components for final export. Another relevant feature is the presence of nontraded goods. If the productivity shocks hitting the nontraded sector differ from the traded sector, it can become difficult for international trade in asset markets to insure against them. Calibrating and simulating models with these more realistic features indicates that it is possible for some economies to benefit substantially from policy coordination (see Tchakarov 2004).

In sum, the size of benefits from coordination depends on a number of key characteristics of economies. These include how developed asset markets are, how responsive trade flows are to relative prices, how important it is to households to smooth their consumption levels over time, how imports are used, and how sticky prices are set. Whether policy coordination is worthwhile for a country depends largely on the individual characteristics of that country.

Openness reconsidered

While the discussion above has offered two motivations for policy coordination, namely, risk sharing and price stickiness, a revealing distinction between the two is how they are affected by openness and globalization. Consider first openness in the form of international economic integration in goods markets. Goods trade itself may have built-in mechanisms that can help insure a country against country-specific output shocks. For example, if a country is hit by a fall in its production, the relative scarcity of home goods would induce a rise in their relative price. Depending on consumer preferences, such as a type implying constant expenditure shares over home and foreign goods, this terms-of-trade effect will be able to compensate home agents for the fact they have a smaller quantity of home goods. In particular, they will be able to import more foreign goods in exchange for the smaller quantity of home exports, and thereby enjoy a comparable level of overall consumption and utility as the foreign country. This means that goods markets potentially can do the job of pooling risk internationally without the need for an international policy coordinator.

This conclusion stands in sharp contrast to earlier literature. Recall that Oudiz and Sachs (1984) concluded that the need for coordination was small precisely because the degree of goods trade was small. But here we conclude that the need for coordination is small when goods market integration is high.

Consider also the implications of integration in asset markets. In the limiting case where asset markets were complete, with assets to insure against all shocks, private agents would be able to pool the risk of asymmetric shocks internationally on their own. Again, if private markets pool risk, there is no need for policy coordination to serve this function. Clearly the world remains far from complete asset markets, but international trade in equities is definitely on the rise, and international capital flows of various types

have ballooned. One gets the impression that international integration has progressed faster in asset markets than in goods markets, so that this type of integration may be more important. Nevertheless, integration in both markets works in the same direction here. A high level of integration, be it in either asset markets or goods markets, indicates there is less need for explicit international policy coordination to pool national risks. Contrary to the predictions of some analysts, as the age of globalism progresses we might see less pressure for international policy coordination rather than more.

See Also

- international finance
- international monetary institutions
- international real business cycles
- macroeconomic effects of international trade
- monetary and fiscal policy overview

Bibliography

Canzoneri, M.B., Cumby, R.E. and Diba, B.T. 2005. The need for international policy coordination: what's old, what's new, what's yet to come? *Journal of International Economics* 66, 363–84.

Canzoneri, M.B. and Gray, J. 1985. Monetary policy games and the consequences of non-cooperative behavior. *International Economic Review* 26, 547–64.

Corsetti, G. and Pesenti, P. 2005. International dimensions of optimal monetary policy. *Journal of Monetary Economics* 52, 281–305.

Devereux, M. and Engel, C. 2003. Monetary policy in the open economy revisited: price setting and exchange rate flexibility. *Review of Economic Studies* 70, 765–83.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Hamada, K. 1974. Alternative exchange rate systems and the interdependence of monetary policies. In *National Monetary Policies and the International Financial System*, ed. R. Aliber. Chicago: University of Chicago Press.

Obstfeld, M. and Rogoff, K. 2002. Global implications of self-oriented national monetary rules. *Quarterly Journal of Economics* 117, 503–35.

Oudiz, G. and Sachs, J. 1984. Macroeconomics policy coordination among the industrial countries. *Brookings Papers on Economic Activity* 1984(1), 1–64.

Tchakarov, I. 2004. The gains from international monetary cooperation revisited. IMF Working Paper No. WP/04/1. Washington, DC: International Monetary Fund.

How to cite this article

Bergin, Paul R. "international policy coordination." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000298> doi:10.1057/9780230226203.0834

The New Palgrave Dictionary of Economics Online

international real business cycles

Mario J. Crucini
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International business cycle research seeks to summarize the statistical properties of worldwide economic fluctuations and model them as the outcome of purposeful decisions by individuals, firms and policymakers who react to changes in their economic environment and an uncertain future. The focus is on identifying the sources of fluctuations and how interactions of economic actors play out in terms of cyclical movements in variables such as gross domestic product. The term ‘real’ indicates a sub-area of the research programme that focuses on non-monetary dimensions such as changes in productivity and fiscal policy rather than in the money supply and monetary policy.

Keywords

adjustment costs; Armington aggregator; asset market structure; business cycle measurement; business cycles; capital accumulation; capital utilization; consumption function; consumption smoothing; current account; Euler equations; expectations; factor mobility; imperfect competition; incomplete markets; international finance; international real business cycles; international taxation; international trade; intertemporal substitution effect; law of large numbers; law of one price; nominal exchange rates; open economy analysis; permanent-income hypothesis; purchasing power parity; rational expectations; real business cycles; real exchange rates; risk sharing; state space models; sticky prices; terms of trade; total factor productivity; trade costs; wealth; world business cycle

Article

1 International real business cycles

Business cycles are the recurrent fluctuations of national output relative to its long-term growth trend. The qualitative features of these fluctuations are common to virtually all economies, with their quantitative properties differing somewhat across countries and time periods. Modern research seeks to summarize the statistical properties of business cycles and formally model them as the outcome of purposeful decisions by individuals and firms who react to changes in their economic environment and an uncertain future. Whereas closed-economy analysis focuses on responses to domestic shocks and policy actions, open economy analysis *adds* to this international policy interaction and spillovers of foreign shocks to the domestic economy. The term ‘real’ indicates a sub-area of the business cycle research programme that focuses on non-monetary dimensions such as changes in productivity, taxes and government spending, rather than changes in the money supply and monetary policy.

2 Measuring international business cycles

What may be surprising to the uninitiated is the controversy surrounding business cycle measurement itself. Measures most often cited in the press are the calendar dates of business cycle peaks and troughs. In the United States, these dates are identified by the Business Cycle Dating Committee at National Bureau of Economic Research. A committee affiliated with the Center for Economic Policy Studies serves the same function for Europe. The logic of the methods used by both committees dates back to the classic contribution of Burns

and Mitchell (1946), pioneers of formal business cycle measurement.

In academic work, economists favour econometric methods in which the logarithm of real gross domestic product, y_t, is decomposed into a growth trend, $y_{g,t}$, and a business cycle component, $y_{c,t}$:

$$y_t = y_{g,t} + y_{c,t}. \tag{2.1}$$

A large applied econometrics literature achieves trend and cycle decompositions by applying identifying assumptions on the innovations to the trend and cycle components of aggregate output. See, for example, Beveridge and Nelson, 1981; Cochrane, 1994; Crucini and Shintani, 2006; Stock and Watson, 2005. Here we employ the Hodrick–Prescott (1997) filter to achieve this decomposition since it is widely used in the literature. The Hodrick–Prescott filter provides a smooth estimate of the growth trend, $y_{g,t}$, and the cycle is computed as the difference between the growth trend and the original series.

Figure 1 displays the business cycle component of the logarithm of gross domestic product for eight industrialized countries: Australia, Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States. As is evident, business expansions and contractions are persistent. One also sees common features such as the emergence of a recession in the 1980s simultaneously in most countries.

Figure 1

Business cycle component of the logarithm of gross domestic product for eight industrialized countries, 1970–2005. *Source*: OECD Quarterly National Accounts, CD-ROM and author's calculations.

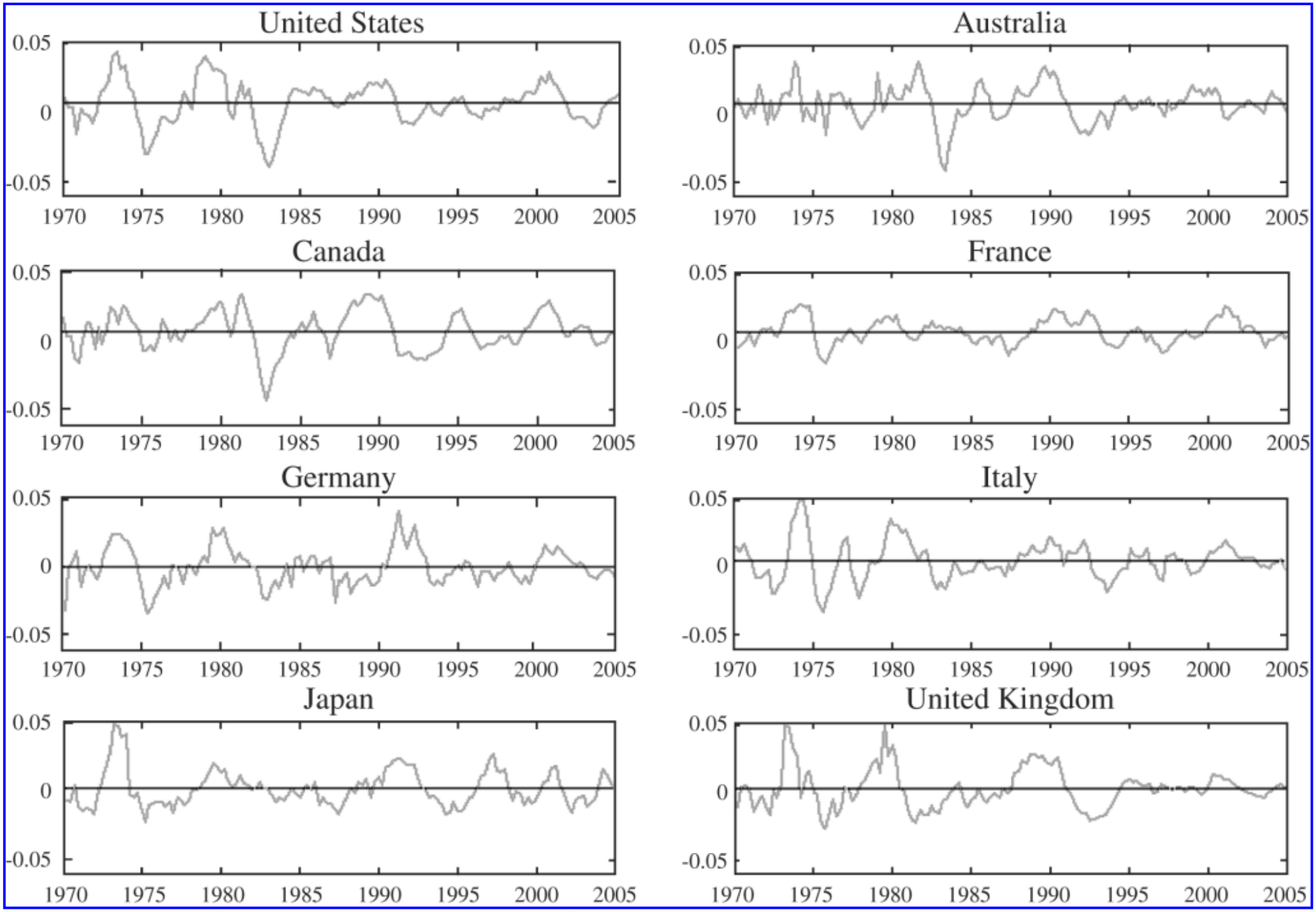

We organize our discussion of business cycle facts around two equations. The first is the national income and product accounts (NIPA) accounting identity. (The OECD data satisfy this identity when changes in inventories and a statistical discrepancy are included. We subtract these two items from output when we perform the variance decomposition of output from the expenditure side.)

$$Y_t = C_t + I_t + G_t + (X_t - M_t). \tag{2.2}$$

In words: the amount of output produced in the home country equals the sum of its uses in domestic private consumption and investment, C_t and I_t, government spending, G_t, and exports, X_t. Imports are deducted to avoid double counting since they are already counted in the other expenditure components.

The variables have been ordered in terms of the fraction of output accounted for by each component. Averaged across time periods and countries, consumption accounts for about 58 per cent of output and investment accounts for 23 per cent, while the percentages for government consumption, exports and imports are almost identical, at 18 per cent, 19 per cent, and 19 per cent, respectively. With the exception of exports and imports, the ratios differ modestly across industrialized countries when long-time averages are taken. We use eq. (2.2) below to perform an *expenditure-side* decomposition of output variability.

The second relationship is a theoretical construct. The prototype model assumes that output is produced with two inputs, capital and labour. The production function relating inputs to outputs usually takes the form:

$$Y_t = A_t K_t^{\alpha} N_t^{1-\alpha} \tag{2.3}$$

where A_t is total factor productivity, K_t is the stock of physical capital in place at time t, N_t is total hours of input at time t. The exponent $1-\alpha$ measures the share of national income paid to labour (salaries and wages) since labour is paid its value marginal product in the model.

Taking logarithms of eq. (2.3) provides the basis for the second variance decomposition:

$$y_t = a_t + \alpha k_t + (1-\alpha) n_t. \tag{2.4}$$

We compute a_t as a residual, setting $\alpha = \frac{1}{3}$ (the share of capital income in national income) and using standard measures of physical capital and aggregate hours, as the inputs on the right-hand-side of the equation. We call this our *production-side* decomposition.

Table 1 contains business cycle statistics for each country using data from the first quarter of 1970 to the first quarter of 2005. Beginning with the variance of the cycle itself, we see that the United States has the most variable business cycle, with a standard deviation of 1.58 per cent per quarter, while France, at the other end of the scale, has a standard deviation of only 0.91 per cent. Australia, Canada, Germany, Italy, Japan and the UK have remarkably similar volatility, in the range of 1.32–1.48 per cent.

Business cyclical properties of eight industrial countries, 1970Q1–2005Q1

	US	Australia	Canada	France	Germany	Italy	Japan	UK
Std. dev. of output	1.58	1.32	1.46	0.91	1.36	1.43	1.35	1.48
Panel A. Standard deviations relative to output								
Consumption	0.80	0.77	0.79	0.97	0.87	0.93	0.92	1.14
Investment	2.85	3.41	2.83	3.11	2.59	2.29	2.36	2.49
Government	0.54	1.26	0.78	0.78	0.86	0.55	0.92	0.72
Exports	2.68	3.00	2.66	3.11	3.00	2.71	3.21	1.97

Imports	3.26	4.83	3.16	3.95	2.32	3.24	4.31	2.54
Savings	4.46	4.88	3.72	4.07	3.70	3.03	2.52	4.17
Productivity	0.56	0.76	0.64	1.04	0.72	0.94	0.68	0.80
Capital	0.39	0.43	0.42	2.87	1.11	1.16	0.55	0.35
Labour	0.83	1.01	0.94	0.65	0.66	0.65	0.62	1.19
Panel B. Correlation with own-country output								
Consumption	0.85	0.42	0.82	0.71	0.67	0.75	0.79	0.79
Investment	0.95	0.78	0.60	0.83	0.81	0.79	0.93	0.66
Government	–0.18	0.07	–0.15	–0.20	0.10	–0.03	0.04	–0.19
Exports	0.42	0.11	0.67	0.72	0.62	0.24	0.05	0.48
Imports	0.81	0.45	0.73	0.80	0.74	0.70	0.62	0.68
Savings	0.88	0.86	0.89	0.81	0.82	0.82	0.84	0.74
Productivity	0.84	0.67	0.77	0.45	0.76	0.87	0.91	0.58
Capital	0.26	0.20	–0.14	0.25	0.26	–0.08	0.37	0.09
Labour	0.90	0.68	0.84	0.67	0.81	0.47	0.75	0.67
Net export ratio	–0.44	–0.32	–0.09	–0.28	0.08	–0.38	–0.41	–0.30
Correlation of savings and investment	0.63	0.44	0.67	0.58	0.44	0.80	0.47	0.83

Notes: All variables except the net export ratio are the Hodrick–Prescott cycle components. All nominal variables are deflated by the Gross Domestic Product Deflator.*Source*: OECD Quarterly National Accounts, CD-ROM.

Turning to the details, we see that investment and trade flows are much more variable than output; consumption is less variable than output while government spending is the least variable. There are some quantitative differences across countries, but the rankings are robust.

The correlation of variables with output indicates the cyclicality of a variable. If the correlation is positive, the variable is said to be pro-cyclical: on average, it rises when the economy is in an expansionary phase and falls when the economy is in a contractionary phase. All variables except government spending and the net export ratio are strongly pro-cyclical, consumption and investment particularly so. In a statistical sense, government spending seems to provide some stabilization by virtue of its low variability and near-zero correlation with the cycle. Imports are consistently more highly correlated with domestic output than are exports. This makes economic sense since import demand is influenced by domestic income while export demand depends on potentially diverse income developments across a country's trading partners.

On the production side of the equation, capital is less cyclically variable than either productivity or labour input (a notable exception is France). The ranking of the variability of labour input relative to productivity is ambiguous.

2.1 Variance decompositions

The variance decomposition of output from the expenditure side or production side is computed as:

$$std(y) = \sum_{z} s_z \cdot std(z) \cdot corr(z, y) \tag{2.5}$$

where s_z is either the expenditure share or the production share for variable z (productivity gets a weight of one), $std(z)$ is the standard deviation of component z over the cycle and $corr(z,y)$ is the correlation between component z and income. The variance decomposition is exact in levels, but approximate in logs, because the NIPA identity involves levels. The

variance decomposition is exact on the production side because of the log-linearity of the production function.

On the expenditure side consumption and investment account for about 95 per cent of the cyclical variation in aggregate demand. There is no consistent ordering of their relative importance. The reason for consumption's impact is that about two-thirds of aggregate demand is accounted for by this component. While investment is a paltry 23 per cent of aggregate demand, it is about twice as variable as consumption and therefore exerts an influence on the cycle larger than its expenditure share would suggest. Imports are often as important as consumption or investment, while the contribution of exports is not robust across countries. However, since imports and exports enter the national income and product identity with opposite signs, they tend to cancel out. Fluctuations in government spending contribute little to the cycle, for three reasons. First, government spending accounts for a relatively small amount of aggregate demand, close to the investment and trade shares and much lower than that of private consumption. Second, government spending is typically less variable than output. Third, the correlation between government spending and output is close to zero, on average. (In periods of war, such as the Second World War, the picture is very different since government spending is a much larger fraction of output and is strongly pro-cyclical.)

To turn to the *production side*, total factor productivity and changes in labour input account for virtually all of the cyclical variation in output (the cross-country average contribution of these two combined is 95 per cent). This is because each of these variables is highly variable and highly correlated with output, much more so than is true of the physical capital stock. Moreover, capital's share in income is exactly one-half that of labour's, reducing its influence relative to labour. While productivity and labour have a comparable influence, the source of the influence differs. Labour input is more variable than productivity, but gets a weight of two-thirds, less than the unit coefficient on total factor productivity (see eq. (2.4)).

It should be stressed that, while these accounting-based decompositions are useful in framing the discussion, they do not tell us what the underlying sources of business cycles are. To see this, consider the distinction between choice variables and exogenous variables. In the prototype real business cycle model, productivity is the only exogenous source of economic change, all other variables are responding optimally to this variable. The model, then, tells us that productivity variation accounts for all of business cycle variation and the various facets of how this plays out across macroeconomic aggregates reflect the choices made by individuals, firms and governments, in response to these productivity changes.

Thus, in practice, there is a subtle link between exogenous impulses and endogenous responses to them. For example, Imbs (1994) introduces variable capital utilization into the model described above. Since capital utilization is not part of what we are measuring in our physical capital stock series, we incorrectly allocate variation in capital utilization to productivity. It is natural to think that this leads us to overestimate the role of productivity. Baxter and Farr (2005) show, however, that, when one moves from a model with constant utilization of capital to one with variable utilization, the response of the economy to a productivity change of a fixed size is larger when utilization is variable than when it is fixed. This moves the bias in the other direction. The lesson here is that theory and measurement work best in concert to achieve the most accurate possible attribution of economic variance.

2.2 International dimensions of the business cycle

We turn, now, to key international facets of business cycles: (*a*) the current account balance, (*b*) international business cycle co-movement and (*c*) relative price determination.

2.2.1 The current account

An important goal of international business cycle research is to improve our understanding of the time path of the current-account balance or the trade balance. International trade focuses on the direction and composition of trade and often assumes balanced trade. International finance focuses on the current account, modelling the dynamics of savings and investment over time. Since the business cycle involves time variation, it is natural to emphasize the international finance perspective.

The current account equals the difference between savings and investment. National savings is the sum of private savings and public savings. Private savings is the difference between disposable income and private consumption while public savings is the difference between tax revenue and government expenditure.

$$CA_t = S_t - I_t S_t = \underbrace{(Y_t - T_t - C_t)}_{\text{private saving}} + \underbrace{(T_t - G_t)}_{\text{public saving}} . \tag{2.6}$$

In a closed economy, of course, the current account is identically equal to zero – each dollar of savings must be allocated to domestic investment. An open economy, freed from this constraint, rarely finds itself with a current account balance; when current savings fall short of (or exceed) current investment levels, a current account deficit (or surplus) obtains. Feldstein and Horioka (1980) vividly demonstrated that, when the data are averaged over long periods of time, savings and investment rates are highly positively correlated – countries with higher than average savings rates tend to have higher than average investment rates. Business cycle correlations of saving and investment tend to be lower than the

Feldstein–Horioka values, suggesting that large deviations in the current account are transitory. The correlation of national saving and national investment over the cycle ranges from a high of 0.80 in Italy to a low of 0.44 in both Australia and Germany (see Table 1).

2.2.2 International business cycle co-movement

International co-movement may be expressed in different ways. Kose, Otrok and Whiteman, 2003, among others, use state space models in which there are world, country and idiosyncratic factors in the income process as well as in each component of aggregate demand. This method avoids an arbitrary choice of numeraire and helps to identify what economists refer to as the 'world business cycle'. Here we use the correlation of a foreign variable with its US counterpart. As is evident in Table 2, positive movements of foreign variables with their US counterparts are the rule rather than the exception. In terms of rankings, output tends to be more correlated than the components of aggregate demand; investment and government spending have particularly low international correlations. The rankings are more ambiguous in a statistical sense and for a broader range of countries than Table 2 suggests; see Ambler, Cardia and Zimmerman (2004).

International business cycle co-movement correlation with US counterpart, 1970Q1–2005Q1

	Australia	Canada	France	Germany	Italy	Japan	UK
Output	0.46	0.71	0.36	0.42	0.32	0.43	0.64
Panel A. Demand side							
Consumption	–0.09	0.53	0.37	0.37	0.01	0.35	0.50
Investment	0.29	0.16	0.25	0.47	0.15	0.42	0.40
Government	0.22	0.29	–0.04	0.05	–0.01	0.07	0.06
Exports	0.03	0.33	0.40	0.34	0.10	0.25	0.32
Imports	0.13	0.45	0.36	0.32	0.40	0.29	0.50
Savings	0.53	0.68	0.38	0.39	0.33	0.51	0.37
Net exports	–0.18	–0.50	–0.08	0.23	–0.29	–0.16	0.07
Panel B. Supply side							
Productivity	0.42	0.53	–0.07	0.21	0.04	0.27	0.36
Capital	0.33	0.18	0.08	0.17	0.09	0.31	0.55
Labour	0.42	0.59	0.36	0.39	–0.17	0.42	0.60

Source: Author's calculations.

To turn to the *production side*, we see that US labour input has the highest correlation with its counterpart abroad, ranging from 0.60 with the UK to a low of minus 0.17 with Italy. International productivity levels also tend to be positively correlated, though not to the extent of labour input. Changes in capital formation have a low international correlation, consistent with other facets of this input documented above. The highest international business cycle correlations are between Canada and the United States, geographic neighbours with similar institutions and extensive trade relations.

2.2.3 Real exchange rates and the terms of trade

The two key international relative prices are the real exchange rate and the terms of trade. The real exchange rate is:

$$Q_t^R = \ln(E_t P_t^* / P_t) \tag{2.7}$$

where E_t is the nominal exchange rate between the home and foreign country and P_t and P_t^* are home and foreign price indices (usually the consumer price index), respectively. In words: Q_t^R is the cost of the foreign consumption basket relative to the domestic consumption basket after converting to a common currency. According to the purchasing power parity proposition, the dollar goes just as far in foreign countries as it does in the United States in terms of purchasing power. This implies that $Q_t^R = 1$ at each point in time.

In practice, however, the real exchange rate is highly variable and very persistent. High variability suggests large absolute departures from parity, while high persistence implies that, when a price gap opens up internationally, it tends to remain open for many months rather than days or weeks. In terms of the time series measurement of this property, at business cycle frequencies, it appears that the real and nominal exchange rates have approximately the same variance while the price ratio term (P_t^* / P_t) is very stable. For example, the standard deviation of the nominal exchange rate between the United States and France is about 8.52, close to the standard deviation of their bilateral real exchange rate at 7.95, while the price ratio has a standard deviation of only 1.17 (see Table 3). These numbers are typical of US bilateral real exchange rates with respect to other industrialized countries. One also finds that the real exchange rate is not highly correlated with quantity variables such as output or even net exports (not shown).

Cyclical properties of real exchange rates and the terms of trade

	US	Australia	Canada	France	Germany	Italy	Japan	UK
Panel A. Standard deviations								
Price ratio				1.17	1.42	1.67		1.74
Nominal exchange rate				8.52	8.37	8.51		8.20
Real exchange rate				7.95	8.06	7.80		
Terms of trade	2.90	5.21	2.44	3.50	2.61		5.68	2.64
Trade ratio		9.94	4.60	3.66	3.90		7.29	3.94
Panel B. Contemporaneous cross correlations								
Output and net exports	–0.30	–0.19	–0.43	–0.30	–0.05		–0.23	–0.25
Output and the terms of trade	–0.08	–0.30	–0.11	–0.14	–0.09		–0.09	0.22
Terms of trade and net exports	0.28	–0.07	0.06	–0.51	0.00		–0.50	–0.54

Sources: Terms of trade moments are from Table 1 in Backus and Crucini (2000). Sample periods are as follows: Canada, the United Kingdom and the United States, 1955Q1–1990Q3; Australia, 1960Q1–1990Q3; France, 1970Q1–1990Q3; Germany, 1968Q1–1990Q3; Italy, 1970Q1–1990Q2; Japan, 1955Q2–1990Q3. Real exchange rate moments are from Chari, Kehoe and McGrattan (2002); sample period is 1973Q1–2000Q1.

To turn to the terms of trade, it is defined as:

$$Q_t^T = P_t^m / P_t^x \tag{2.8}$$

where P_t^m and P_t^x are import and export price indices for a particular country. Since these price indices are domestic deflators, they are already expressed in the home currency terms, and the spot exchange rate is not needed to convert them to common units. Unlike the real exchange rate, economic theory does not place strong restrictions on the time series or cross-country behaviour of the terms of trade. Given the presumption that countries import different goods from those they export, we expect the terms of trade to be different from unity, and it should fluctuate, too.

Australia and Japan have the highest terms-of-trade variability, about twice that of the other countries, with the exception of France, which experiences terms-of-trade variability between these extremes. The terms of trade does not have a robust correlation with either output or net exports.

3 Modelling international business cycles

Quantitative theoretical investigations of business cycles seek to account for business cycle facts using models in which consumers are thoughtful and informed, firms employ workers and utilize capital efficiently, and policymakers use a combination of rules and discretion to achieve various economic objectives. The key dimensions of study are those unique to international economics: matching the international character of the world business cycle and the business cycle properties of the current account, the real exchange rate and the terms of trade.

3.1 The current account

The most rudimentary model of current account behaviour is one in which a small open economy faces an exogenous world interest rate and income stream. To fix ideas, think of a small country that produces mostly oil with perfect access to international capital markets. If the country is always producing at capacity, all of its income variation is due to changes in the price of oil in world markets. What does the intertemporal approach to the current account predict in this circumstance?

The theory reduces the NIPA identity to: $S_t = Y_t - C_t$, so that consumption decisions effectively determine saving decisions. Investment is absent since we are abstracting from changes in production capacity and its utilization. While this model seems simplistic, the identity is deceptive since it suggests that only current income enters into the current consumption–savings decision. In fact, the most widely used set-up has its roots in the seminal contribution of Friedman (1957), with individuals assumed to be able to draw upon the entire present discounted value of their future labour income. Whereas current income is the traditional argument in the Keynesian consumption function, wealth plays this role in modern macroeconomics. Since wealth is the sum of the market value of financial assets and all future anticipated flows of income, expectations play a central role in the modern consumption function. (There are many extensions to this basic framework that prevent individuals from drawing upon their lifetime wealth for present consumption: collateral requirements, limits on debt-to-income ratios and credit histories. Discussion of these extensions is beyond the scope of this survey.)

Much of the intuition for the impact of a changing income profile on the current account of a small open economy is available from Quah's (1990) formulation of the permanent-income hypothesis. He assumes a constant interest rate, quadratic preferences and rational expectations. He allows income to contain both permanent and transitory shocks. If we assume income follows a first-order autoregressive process: $Y_{t+1} = \rho Y_t + v_{t+1}$, where v_{t+1} is news about income (that is, under rational expectations, news about income is: $E_{t+1} Y_{t+1} - E_t Y_{t+1} = v_{t+1}$), the predicted change in consumption in response to this news is:

$$\Delta C_{t+1} = \frac{r}{1 + r - \rho} v_{t+1} \tag{3.1}$$

and the change in the current account on impact, on the assumption it was in balance initially, is:

$$\Delta CA_{t+1} = \Delta Y_{t+1} - \Delta C_t = v_{t+1} - \frac{r}{1 + r - \rho} v_{t+1}. \tag{3.2}$$

Since output deviations from trend (the business cycle) are persistent, it is safe to assume, $\rho > 0$. A plausible value for *r* is 0.05 (a five per cent real interest rate). Note that the consumption response depends positively on persistence since wealth effects are rising in the persistence of the income change. As persistence moves from zero toward unity, the effect on the current account rises from close to unity toward zero. This algebra delivers a key prediction of the intertemporal approach, that consumption smoothing leads to current account surpluses during booms unless the income change is viewed as permanent (that is, $\rho = 1$) in which case the current account is predicted to remain unchanged.

While there is evidence to suggest an interest rate channel on consumption, it does not help to resolve the counterfactual prediction of a pro-cyclical current account from the consumption side, just established. There are two reasons for this. First, if interest rates are higher during a boom in the home country, individuals would tend to tilt consumption from current to future periods (that is, postpone durable goods purchases) – the intertemporal substitution effect. This would reinforce rather than overturn our prediction that the current

account moves into surplus during a boom. If real interest rates actually fell during a boom, the intertemporal substitution effect would operate in the right direction, but the evidence on the cyclicality of the real interest is ambiguous. Second, when we move to a general equilibrium setting, incorporating home and foreign responses, the increase in the real interest rate is shared by the two countries and therefore incapable of delivering the asymmetric consumption responses necessary to move the current account balance. This leaves us with the need to look elsewhere for a channel that moves the current account in a countercylical direction.

To return to the algebra of the current account identity, it would appear that what is needed for a countercyclical current account is for domestic investment to rise more than domestic savings during a business cycle expansion:

$$\Delta CA_t = \Delta S_t - \Delta I_t = (\Delta Y_t - \Delta C_t) - \Delta I_t. \tag{3.3}$$

The identity reveals the tension between the consumption smoothing channel, whereby a transitory change in income is mostly saved, pushing the current account towards surplus and the investment channel, which pulls in the opposite direction, towards a deficit.

In a model with only one good, the consumption smoothing channel wins the contest unless the shocks are highly persistent (see, for example, Backus, Kehoe and Kydland, 1992; Baxter and Crucini, 1993; Mendoza, 1991). Persistence, by increasing the impact response of consumption due to the larger wealth effect, helps to push the current account towards balance, leaving the investment channel to produce a deficit. Extensive empirical investigations of the intertemporal approach to the current account may be found in Glick and Rogoff (1995) and Nason and Rogers (2006). (Sachs (1981) provides early evidence on the investment channel.)

Extensions of the model to multiple goods helps avoid this unpleasant arithmetic because individuals want to increase consumption of both the domestic good and the foreign good, increasing import demand and reinforcing the tendency towards a deficit from a traditional trade channel. Demonstrations of this effect under complete and incomplete risk sharing are found in Backus, Kehoe and Kydland (1994) and Arvanitis and Mikkola (1996), respectively. (JoAnne Feeney (1994) provides an insightful exposition of this issue.)

To summarize, early developments of the intertemporal approach to the current account emphasized the consumption smoothing channel and predicted that current account surpluses would occur when output was temporarily above trend. Current account surpluses are often described as *good* based on the idea that surpluses flow from good economic times. The complete model of the current account adds investment dynamics and allows for the possibility that investment-led booms produce current account deficits. These theoretical developments and their empirical implications have led to a more balanced view of the current account: that we need to understand the sources of the changes in the current account before making value judgements about them. Kollman (1998), appears to be the first quantitative simulation of US and European current account dynamics using a modern real business cycle analysis that incorporates variation in productivity, government spending and national tax rates.

3.2 The world business cycle

Conceptually, the world business cycle is simple to define: the deviation of world output from its growth trend. The practical difficulty is the measurement of world output because national output is denominated in domestic currency. Converting nominal output into a common currency using spot nominal exchange rates greatly exaggerates fluctuations in output because nominal exchange rates are much more volatile than either real production or price levels. Moreover, prices vary considerably across nations even after conversion to a common currency, making it difficult to construct an appropriate deflator to convert nominal gross international product into real gross international product. Here we follow much of the existing literature and use real gross domestic product of each country, and compute correlations across them. If real output is highly correlated across countries, we have evidence of a world business cycle. As we documented earlier, most macroeconomic aggregates are positively correlated across countries, indicative of a world business cycle. How do business cycle researchers account for this fact?

There are two channels through which positive economic co-movement may arise: endogenous propagation and exogenous propagation. Positive endogenous propagation refers to a situation in which a disturbance originating in one country has a positive impact on both home and foreign output levels. For example, rapid development in China drives up demand for crude petroleum and fuels economic expansions in countries that are specialized in oil production. Positive exogenous propagation refers to the correlation of shocks across countries. For example, the Second World War witnessed dramatic increases in national output in most industrialized countries as government spending rapidly expanded during the conflict. In practice, endogenous propagation and exogenous propagation are difficult to distinguish, presenting one of the key challenges of business cycle research.

Real business cycle researchers have devoted most of their effort to measuring total factor productivity, which has been found to be highly persistent and positively correlated across countries. Correlations over the business cycle are typically lower than correlations over long periods of measurement, suggesting more commonality in the technological trend than in the productivity cycle. (When analysis extends to small developing countries the business cycle correlations sometimes exceed the growth correlations.) Given the lower

correlation of fiscal variables with the cycle and their modest cyclical variation, it is not surprising that they have received less attention in empirical and theoretical analysis than productivity. Two key studies of the empirical behaviour of international taxes and their equilibrium implications using dynamic equilibrium theory are Mendoza, Razin and Tesar (1994), and Mendoza and Tesar (1998), respectively. Both studies suggest international taxation is more relevant for secular and long-run trends than it is for business-cycle fluctuations.

Apart from the obvious role of the correlation of the shocks that drive business cycles, three economic factors have proven critical in determining the ability of dynamic equilibrium models to generate international co-movements resembling those we see in the data. The first is the extent to which domestic and foreign goods are substitutes in demand. The second is the extent to which factors of production are internationally mobile. The third is the extent of international financial linkages.

The first generation of models by Backus, Kehoe and Kydland (1992) and Baxter and Crucini (1993) followed the analytical structure of the closed economy models by Kydland and Prescott (1982) and King, Plosser and Rebelo (1988) quite closely. Despite the similarity, however, international economists were immediately confronted with two key modelling issues. The first had to do with factor mobility across countries, which is obviously absent in the closed economy setting. The mobility of labour across countries seemed minor enough to ignore, physical capital mobility was not. Since physical capital takes real resources to reallocate, the standard approach has been to subject capital accumulation to adjustment costs (or time to build as in Backus, Kehoe and Kydland, 1992). Without some cost of physical capital mobility, capital would be predicted to move rapidly and in large amounts across national boundaries in response to persistent changes in productivity or taxes. Such factor movements generate strongly negative correlations of output from the supply side and unrealistically volatile investment over the business cycle.

The second issue model builders were confronted with was asset market structure. Much of aggregative economics is predicated on the basis that idiosyncratic shocks are irrelevant to macroeconomic fluctuations. In an economy with millions of individual agents and thousands of firms, the law of large numbers combined with not-too-objectionable restrictions of preferences and technology provided a compelling argument to abstract from idiosyncratic variation. At the aggregative international level, the number of shocks is small (in many models it equals the number of countries), and countries are large and few in number. Thus, it makes little sense to rely on the law of large numbers, so researchers adopted the assumption that agents pool nation-specific risks, avoiding the need to track the wealth distribution across countries.

Unfortunately, complete risk pooling in the one-sector model leads to a presumption that output is negatively correlated across countries while consumption is close to perfectly positively correlated. In the data, the reverse rankings of correlation tend to prevail, and the absolute level of consumption correlations is well below unity. The prediction of near-perfect consumption co-movement across countries derives from the risk-pooling assumption and the fact that agents face common prices and interest rates.

The negative correlation of income is driven by cost-minimizing production decisions where firms allocate plants and equipment to the most productive location. Thus, an increase in home productivity increases domestic output relative to foreign output directly, and this is reinforced by the flow of capital from the less productive country to the more productive country. Risk pooling also enhances the supply-side response by neutralizing the wealth redistribution effects on home and foreign labour supplies.

Debate continues as to what the appropriate asset market structure should be and how to incorporate changes in asset diversification in business cycle models. Baxter and Crucini (1995) and Kehoe and Perri (2002) show that, when risk pooling is limited, positive output co-movements are more likely to arise the more persistent are the deviations to relative international productivity. Also, consumption correlations may actually fall below output correlations if the shocks are close to permanent, a feature that is prevalent in the data and difficult to explain from a number of standard theoretical paradigms.

Researchers have had more success accounting for positive international output co-movement in models where countries depend on their trading partners for final goods or intermediate inputs they themselves do not produce. Examples of work along these lines include an extension of the multisector model with intermediate inputs of Long and Plosser (1983) to the open economy by Ambler, Cardia and Zimmerman (2002), a model of the North–South business cycle by Michael Kouparitsas (1996) which emphasizes trade of manufactures for primary inputs across these two regions, and the introduction of home production by Canova and Ubide (1998). A contribution that extends the incomplete markets model developed by Baxter and Crucini (1995) to the two-good setting is Arvanitis and Mikkola (1996).

3.3 Real exchange rates and the terms of trade

Multiple sectors take centre stage when one considers the real exchange rate and the terms of trade. Approaches to international relative price determination may be usefully placed into two categories. One category focuses on the determination of international relative prices of different goods. A second category focuses on deviations from the law of one price, meaning identical goods trade at different prices in different countries.

A classic contribution in the former category is Backus, Kehoe and Kydland (1994) (BKK), who develop a two-country, two-good model. Each country specializes in the production of one of the two goods and the two goods are combined in production, via an Armington aggregator, to create a composite final good which is, in effect, the single final good in each economy.

The Armington aggregator is a function that describes how substitutable the two goods are in achieving a particular output level of the final good. To match low trade shares with the specialization-in-production assumption, home bias is assumed in the aggregator function. This means that the home country uses more of the home good when producing the

composite good, and the foreign country behaves symmetrically.

This is an elegant model that ties in nicely with the one-sector two-country framework. The key difference between this model and the one-sector model is that specialization provides a motivation for keeping production levels more nearly equal across locations, since individuals have demands for each type of good. The model allows us to study the terms of trade, a key international relative price absent from the one-sector model, by construction. In the BKK model, the terms of trade and trade ratio are related as follows:

$$q_t = \ln\left(\frac{p_{bt}}{p_{at}}\right) = \omega + \frac{1}{\sigma}\ln(a_t / b_t). \tag{3.4}$$

In words: an increase in production of the home good, *a*, drives down its relative price. The home terms of trade turn against the country experiencing the expansion, a pro-cyclical terms of trade, as BKK define it. In the data, the correlation varies substantially across countries in magnitude and sign. The model has difficulty matching both the observed volatility of the terms of trade and the quantity ratio; as the Euler equation makes clear, there is a trade-off between terms of trade and quantity ratio variability as the elasticity is altered. If we view *a* and *b* as the final consumption levels of each good, the quantity ratio is not nearly volatile enough, given a plausible degree of elasticity, to generate the terms of trade variation we see in the data. Backus and Crucini (2000) add an oil producing region (and sector) and find that the model does better in matching the cyclicality of the terms of trade and the trade balance than the original BKK model. Kose (2002) provides an extensive quantitative analysis of the variation of international relative prices and their role in the business cycles of small open economies.

Models that consider deviations from the law of one price differ in the source of the price deviations and their duration. Sticky-price models consider the deviations to be transitory, with nominal prices responding with a lag to changes in the economic environment. These models also assume trade in an infinite number of varieties, which allows individual firms to charge a markup of price over marginal cost. Key contributions in this area are Svensson and van Wijgenbergen (1989) and Obstfeld and Rogoff (1995).

Trade cost models treat price deviations as a consequence of a real resource cost of trading, or operating businesses, in different locations. The simplest version allows prices to vary across locations by a shipping cost, usually treated as proportional to the marginal cost of the producer/supplier. The seminal contribution is Samuelson (1952), with more recent contributions by Eaton and Kortum (2002) and Sercu, Uppal and van Hulle (1995). An alternative variant is to distinguish traded and non-traded goods with traded goods not subject to trade costs and non-traded goods assumed to be subject to prohibitive trade costs, as in the original Salter (1959) and Swan (1960) models. Stockman and Tesar (1995) conduct a quantitative investigation of the business cycle predictions of this class of model.

Recent efforts have focused on quantifying the role of sticky prices, imperfect competition and trade costs in accounting for international relative price deviations and their business cycle implications. Chari, Kehoe and McGrattan (2002) conduct a quantitative evaluation of the sticky-price, imperfect-competition model and find that it can account for only a small part of the persistence and somewhat more of the volatility of the real exchange rate. (See also Betts and Devereux, 2000; Bergin and Feenstra, 2000.) Corsetti, Dedola and Leduc (2005) and Ravn and Mazzenga (2004) show the promise of models that combine imperfect competition with real trading costs.

What is missing from existing models is a clear distinction between economic activities that take place at the dock and exchange in retail markets. Transportation costs alone cannot account for all of the retail price dispersion we observe. Presumably, this is because much of what the retail market entails are local inputs of land, labour and infrastructure (some of it publicly provided). Models and empirical evidence are just now being developed to make these distinctions, such as Burstein, Neves and Rebelo (2003) and Crucini, Telmer and Zachariadis (2005), respectively.

See Also

- business cycle measurement
- macroeconomic effects of international trade
- real business cycles
- stylized facts

The author is grateful to Ben Eden for comments and to Hakan Yilmazkuday for excellent research assistance.

Bibliography

Ambler, S., Cardia, E. and Zimmerman, C. 2002. International transmission of the business cycle in a multi-sector model. *European Economic Review* 46, 273–300.

Ambler, S., Cardia, E. and Zimmermann, C. 2004. International business cycles: what are the facts? *Journal of Monetary Economics* 51, 257–76.

Arvanitis, A.V. and Mikkola, A. 1996. Asset-market structure and international trade dynamics. *American Economic Review* 86, 67–70.

Backus, D.K. and Crucini, M.J. 2000. Oil prices and the terms of trade. *Journal of International Economics* 50, 185–213.

Backus, D.K., Kehoe, P. and Kydland, F. 1992. International real business cycles. *Journal of Political Economy* 100, 745–75.

Backus, D.K., Kehoe, P. and Kydland, F. 1994. Dynamics of the trade balance and the terms of trade: the J-curve. *American Economic Review* 84, 84–103.

Baxter, M. and Crucini, M.J. 1993. Explaining saving-investment correlations. *American Economic Review* 83, 416–36.

Baxter, M. and Crucini, M.J. 1995. Business cycles and the asset structure of foreign trade. *International Economic Review* 36, 821–54.

Baxter, M. and Farr, D. 2005. Variable capital utilization and international business cycles. *Journal of International Economics* 65, 335–47.

Bergin, P.R. and Feenstra, R.C. 2000. Staggered price setting, translog preferences, and endogenous persistence. *Journal of Monetary Economics* 45, 657–80.

Betts, C. and Devereux, M. 2000. Exchange rate dynamics in a model of pricing to market. *Journal of International Economics* 50, 215–44.

Beveridge, S. and Nelson, C. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* 7, 151–74.

Burns, A. and Mitchell, W. 1946. *Measuring Business Cycles*. New York: NBER.

Burstein, A.T., Neves, J.C. and Rebelo, S. 2003. Distribution costs and real exchange rate dynamics during exchange-rate based stabilizations. *Journal of Monetary Economics* 50, 1189–214.

Canova, F. and Ubide, A.J. 1998. International business cycles, financial markets and household production. *Journal of Economic Dynamics and Control* 22, 545–72.

Chari, V.V., Kehoe, P.J. and McGrattan, E.R. 2002. Can sticky price models generate volatile and persistent real exchange rates? *Review of Economic Studies* 69, 533–63.

Cochrane, J.H. 1994. Permanent and transitory components of GNP and stock prices. *Quarterly Journal of Economics* 109, 241–65.

Corsetti, G., Dedola, L. and Leduc, S. 2005. International risk-sharing and the transmission of productivity shocks. International Finance Discussion Paper No. 826. Board of Governors of the Federal Reserve Board.

Crucini, M.J. and Shintani, M. 2006. International comovement: is theory ahead of business cycle measurement? Mimeo, Vanderbilt University.

Crucini, M.J., Telmer, C.I. and Zachariadis, M. 2005. Understanding European real exchange rates. *American Economic Review* 95, 724–38.

Eaton, J. and Kortum, S. 2002. Technology, geography, and trade. *Econometrica* 70, 1741–79.

Feeney, J. 1994. Goods and asset market interdependence in a risky world. *International Economic Review* 35, 551–63.

Feldstein, M. and Horioka, C.I. 1980. Domestic saving and international capital flows. *Economic Journal* 90, 314–29.

Friedman, M. 1957. *A Theory of the Consumption Function*, Princeton, NJ: Princeton University Press.

Glick, R. and Rogoff, K. 1995. Global versus country-specific productivity shocks and the current account. *Journal of Monetary Economics* 35, 159–92.

Hodrick, R. and Prescott, E.C. 1997. Post-war U.S. business cycles: an empirical investigation. *Journal of Money, Credit and Banking* 29, 1–16.

Imbs, J. 1994. Technology, growth and the business cycle. *Journal of Monetary Economics* 44, 65–80.

Kehoe, P.J. and Perri, F. 2002. International business cycles with endogenous incomplete markets. *Econometrica* 70, 907–28.

King, R.G., Plosser, C.I. and Rebelo, S. 1988. Production, growth, and business cycles I: the basic neoclassical model. *Journal of Monetary Economics* 21, 195–232.

Kollman, R. 1998. U.S. trade balance dynamics: the role of fiscal policy and productivity shocks and of financial market linkages. *Journal of International Money and Finance* 17, 637–69.

Kose, A. 2002. Explaining business cycles in small open economies: how much do world prices matter? *Journal of International Economics* 56, 299–327.

Kose, A., Otrok, C. and Whiteman, C. 2003. International business cycles: world, region, and country-specific factors. *American Economic Review* 93, 1216–39.

Kouparitsas, M. 1996. North–South business cycles. Working Paper No. 96-9, Federal Reserve Bank of Chicago.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Long, J. and Plosser, C.I. 1983. Real business cycles. *Journal of Political Economy* 91, 39–69.

Mendoza, E. 1991. Real business cycles in a small open economy. *American Economic Review* 81, 797–818.

Mendoza, E., Razin, A. and Tesar, L. 1994. Effective tax rates in macroeconomics: cross-country estimates of tax rates on factor incomes and consumption. *Journal of Monetary Economics* 34, 297–323.

Mendoza, E. and Tesar, L. 1998. The international ramifications of tax reforms: supply-side economics in a global economy. *American Economic Review* 88, 226–45.

Nason, J. and Rogers, J. 2006. The present-value model of the current account has been rejected: round up the usual suspects. *Journal of International Economics* 68, 159–87.

Obstfeld, M. and Rogoff, K. 1995. Exchange rate dynamics redux. *Journal of Political Economy* 102, 624–60.

Quah, D. 1990. Permanent and transitory movements in labor income: an explanation for ‘excess smoothness’ in consumption. *Journal of Political Economy* 98, 449–75.

Ravn, M. and Mazzenga, E. 2004. International business cycles: the quantitative role of transportation costs. *Journal of International Money and Finance* 23, 645–71.

Sachs, J.D. 1981. The current account and macroeconomic adjustment in the 1970s. *Brookings Papers on Economic Activity* 1981(1), 201–68.

Salter, W.E. 1959. Internal and external balance: the role of price and expenditure effects. *Economic Record* 35, 226–38.

Samuelson, P.A. 1952. Spatial price equilibrium and linear programming. *American Economic Review* 42, 283–303.

Sercu, P., Uppal, R. and van Hulle, C. 1995. The exchange rate in the presence of transaction costs: implications for tests of purchasing power parity. *Journal of Finance* 50, 1309–19.

Stock, J. and Watson, M. 2005. Understanding changes in international business cycle dynamics. *Journal of the European Economic Association* 3, 968–1006.

Stockman, A. and Tesar, L. 1995. Tastes and technology in a two-country model of the business cycle: explaining international comovements. *American Economic Review* 85, 168–85.

Svensson, L. and van Wijgenbergen, S. 1989. Excess capacity, monopolistic competition and international transmission of monetary disturbances. *Economic Journal* 99, 785–805.

Swan, T. 1960. Economic control in a dependent economy. *Economic Record* 36, 51–66.

How to cite this article

Crucini, Mario J. "international real business cycles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000261> doi:10.1057/9780230226203.0835

The New Palgrave Dictionary of Economics Online

international reserves

Joshua Aizenman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Developing countries, particularly in East Asia, account for most of the large increase in international reserves–GDP ratios in recent decades. Possible explanations include self-insurance against the output costs of sudden stops; precautionary fiscal demand by countries with inelastic fiscal outlays, sovereign risk, volatile and limited tax capacity; and a modern incarnation of mercantilism. Empirical studies reveal that the 1997–8 East Asian financial crisis triggered a sharp increase in hoarding international reserves. They suggest prominent roles for the precautionary demand and self-insurance motives and conclude that the financial integration of developing countries is associated with greater hoarding of international reserves.

Keywords

Asian miracle; buffer stock model; exchange-rate flexibility; hot money; international capital flows; international reserves; liquidity crises; option pricing theory; self-insurance

Article

International reserves are the liquid external assets under the control of the central bank. An intriguing development since the 1960s has been that, despite the proliferation of greater exchange rate flexibility, international reserves–GDP ratios increased substantially. Flood and Marion (2002) report that reserve holdings have trended upwards; at the end of 1999, reserves were about 6 per cent of global GDP, 3.5 times what they were at the end of 1960 and 50 per cent higher than in 1990. Practically all the increase in reserves-GDP holding has been by developing countries, mostly concentrated in East Asia.
These developments stirred lively debate among economists and financial observers. The earlier literature focused on using international reserves as a buffer stock, part of the management of an adjustable-peg or managed-floating exchange-rate regime. Accordingly, optimal reserves balance the macroeconomic adjustment costs incurred in the absence of reserves with the opportunity cost of

holding reserves (see Frenkel and Jovanovic, 1981). The buffer stock model predicts that average reserves depend negatively on adjustment costs, the opportunity cost of reserves, and exchange rate flexibility; and positively on GDP and on reserve volatility, driven frequently by the underlying volatility of international trade. Overall, the literature of the 1980s supported these predictions; see Frenkel (1983), Edwards (1983), and Flood and Marion (2002) for a recent review.

While useful, the buffer stock model has limited capacity to account for the recent development in hoarding international reserves – the greater flexibility of the exchange rates exhibited in recent decades should work in the direction of reducing reserve hoarding, in contrast to the trends reported above. As an indication of excess hoarding, observers noted that developing countries frequently borrow at much higher interest rates than the one paid on reserves.

The recent literature provided several interpretations for these puzzles, focusing on the observation that the deeper financial integration of developing countries has increased exposure to volatile short-term inflows of capital (dubbed ‘hot money’), subject to frequent sudden stops and reversals (see Calvo, 1998; Edwards, 2004). Looking at the 1980s and 1990s, Aizenman and Marion (2003a) pointed out that the magnitude and speed of the reversal of capital flows throughout the 1997–8 crisis surprised most observers. Most viewed East Asian countries as being less vulnerable to the perils associated with hot money than Latin American countries. After all, East Asian countries were more open to international trade, had sounder fiscal policies, and much stronger growth performance. In retrospect, the 1997–8 crisis exposed hidden vulnerabilities of East Asian countries, forcing the market to update the probability of sudden stops affecting all countries.

The above observations suggest that hoarding international reserves can be viewed as a precautionary adjustment, reflecting the desire for self-insurance against exposure to future sudden stops. Self-insurance has several interpretations. The first focuses on precautionary hoarding of international reserves needed to stabilize fiscal expenditure in developing countries (see Aizenman and Marion, 2003b). Specifically, a country characterized by volatile output, inelastic demand for fiscal outlays, high tax collection costs and sovereign risk may want to accumulate both international reserves and external debt. External debt allows the country to smooth consumption when output is volatile. International reserves that are beyond the reach of creditors would allow such a country to smooth consumption in the event that adverse shocks trigger a default on foreign debt. Political instability, by taxing the effective return on reserves, can reduce desired current reserve holdings. The tests reported by Aizenman and Marion (2003b) are consistent with this interpretation. Another version of self-insurance and precautionary demand for international reserves follows the earlier work of Ben-Bassat and Gottlieb (1992), viewing international reserves as output stabilizers (see Aizenman and Lee, 2005; see Lee, 2004, for insurance perspectives of international reserves applying the option pricing theory). Accordingly, international reserves can reduce the probability of an output drop induced by a sudden stop and/or the depth of the output collapse when the sudden stop materializes (see Kaminsky and Reinhart, 1999).

The views linking the large increase in hoarding reserves to deeper financial integration face a well-known contender in a modern incarnation of mercantilism: international reserves accumulations triggered by concerns about export competitiveness. This explanation has been advanced by Dooley, Folkerts-Landau and Garber (2003), especially in the context of China. They interpret reserves accumulation as a by-product of promoting exports, which is needed to create better jobs, thereby absorbing abundant labour in traditional sectors, mostly in agriculture. While intellectually intriguing,

this interpretation remains debatable. Some have pointed out that high export growth is not the new kid on the block – it is the story of East Asia since the 1950s. Yet the large increase in hoarding reserves happened mostly after 1997. This issue is of more than academic importance: the precautionary approach links reserves accumulation directly to exposure to sudden stops, capital flight and volatility, whereas the mercantilist approach views reserves accumulation as a residual of an industrial policy, a policy that may impose negative externalities on other trading partners.

Aizenman and Lee (2005) test the importance of precautionary and mercantilist motives in accounting for the hoarding of international reserves by developing countries. While variables associated with the mercantilist motive (like lagged export growth and deviation from purchasing power parity) are statistically significant, their economic importance in accounting for reserve hoarding is close to zero and is dwarfed by other variables. Overall, the empirical results are in line with the precautionary demand. The effects of financial crises have been localized, increasing reserve hoarding in the aftermath of crises mostly in countries located in the affected region, but not in other regions. A more liberal capital account regime is found to increase the amount of international reserves, in line with the precautionary view. These results, however, do not imply that the hoarding of reserves by countries is optimal or efficient. Making inferences regarding efficiency would require having a detailed model and much more information, including an assessment of the probability and output costs of sudden stops, and the opportunity cost of reserves. To conclude, greater exposure of developing countries to sudden stops and reversals of hot money as well as growing trade openness go a long way towards accounting for the observed increase in international reserves–GDP ratios by developing markets.

See Also

- exchange rate volatility
- international capital flows
- liquidity constraints

Bibliography

Aizenman, J. and Marion, N. 2003a. The high demand for international reserves in the Far East: what's going on? *Journal of the Japanese and International Economies* 17, 370–400.

Aizenman, J. and Marion, N. 2003b. International reserves holdings with sovereign risk and costly tax collection. *Economic Journal* 114, 569–91.

Aizenman, J. and Lee, J. 2005. International reserves: precautionary versus mercantilist views, theory and evidence. Working Paper No. 11366. Cambridge, MA: NBER.

Ben-Bassat, A. and Gottlieb, D. 1992. Optimal international reserves and sovereign risk. *Journal of International Economics* 33, 345–62.

Calvo, G. 1998. Capital flows and capital-market crises: the simple economics of sudden stops. *Journal of Applied Economics* 1, 35–54.

Dooley, M., Folkerts-Landau, D. and Garber, P. 2003. An essay on the revived Bretton Woods system. Working Paper No. 9971. Cambridge, MA: NBER.

Edwards, S. 1983. The demand for international reserves and exchange rate adjustments: the case of LDCs, 1964–1972. *Economica* 50, 269–80.

Edwards, S. 2004. Thirty years of current account imbalances, current account reversals, and sudden stops. *IMF Staff Papers* 51(Special Issue), 1–49.

Flood, R. and Marion, P. 2002. Holding international reserves in an era of high capital mobility. In *Brookings Trade Forum 2001*, ed. S. Collins and D. Rodrik. Washington, DC: Brookings Institution Press.

Frenkel, J. 1983. International liquidity and monetary control. In *International Money and Credit: The Policy Roles*, ed. G. von Furstenberg. Washington, DC: International Monetary Fund.

Frenkel, J. and Jovanovic, B. 1981. Optimal international reserves: a stochastic framework. *Economic Journal* 91, 507–14.

Kaminsky, G. and Reinhart, C. 1999. The twin crises. the causes of banking and balance-of-payments problems. *American Economic Review* 89, 473–500.

Lee, J. 2004. Insurance value of international reserves. Working Paper 04/175. Washington, DC: IMF.

How to cite this article

Aizenman, Joshua. "international reserves." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000276> doi:10.1057/9780230226203.0836

The New Palgrave Dictionary of Economics Online

international trade and heterogeneous firms

Marc J. Melitz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Empirical studies of production units within sectors have reported a massive amount of heterogeneity in various performance measures (most notably, size and productivity). This heterogeneity, within sectors, matters for theoretical and empirical models of trade. Trade, or trade liberalization more generally, induces important reallocations between heterogeneous producers in a sector: the smallest, least productive producers are forced to exit, and market shares are further reallocated between less productive producers (who do not export) towards larger, more productive exporters. These reallocations generate a new channel for productivity and welfare gains from trade.

Keywords

comparative advantage; export market entry; extensive and intensive margins of trade; factor proportions; firm-level heterogeneity; firm-level productivity growth; international trade; intra-industry trade; market share; monopolistic competition; open economy models of growth; product differentiation; skill-biased technical change; sunk costs of entry; trade costs; trade liberalization; trade models

Article

Census-wide ‘micro’ level studies of production units for a wide range of countries at all levels of development have documented substantial heterogeneity in virtually all relevant performance measures across these production units. For example, across all US manufacturing plants in 1992, a plant one standard deviation above the mean plant size is 167 per cent bigger, and a plant one standard deviation above the mean plant productivity level (value-added per worker) is 75 per cent more productive (Bernard et al., 2003). (More precisely, the standard deviation of log sales is 1.67 and that of labour productivity is 0.75.) These represent massive differences in performance outcomes, which are also reflected in differences in other key plant characteristics. Furthermore, the extent of this heterogeneity does not diminish much when looking within narrowly defined sectors. In the case of the US plants, the

75 per cent productivity difference mentioned above only drops to a 66 per cent difference when controlling for productivity differences across more than 400 different sectors.

These large differences in firm performance are also strongly correlated with the firm decision to engage in international transactions (such as exporting, importing intermediate goods from foreign suppliers, or investing in foreign subsidiaries): only a small proportion of firms report any such activities, even within narrowly defined sectors; and those firms are substantially larger and more productive than their counterparts with no international contacts in the same sector. (This pattern has been documented at both the firm and the plant level for a very wide range of countries. From here on out, I will mostly focus on differences between exporting and non-exporting firms, although similar differences have also been documented concerning multinational firms and firms that import intermediate goods from foreign suppliers.) For the United States, Bernard et al. (2006) report that manufacturing plants are more than twice as large (value of shipments) as and 14 per cent more productive (value-added per worker) than their non-exporting counterparts in the same sector. (Bernard et al., 2006, provide an extensive description of firm-level differences related to international trade based on US manufacturing data and also survey the related empirical and theoretical literatures.) Bernard et al. (2006) also report how these exporting firms exhibit other different characteristics relative to non-exporters: they are more capital- and skill-intensive, and pay higher wages.

This strong correlation between export status and firm characteristics (notably higher productivity) naturally leads to the follow-up question of causality. A very large number of studies have examined this question, usually focusing on a firm's productivity trajectory over time relative to its export market entry decision. Virtually all these studies find a strong self-selection effect: firms are relatively more productive prior to their entry into export markets. (Two early influential papers in this area are Bernard and Jensen, 1999, and Clerides, Lach and Tybout, 1998.) Several of these studies further reject the hypothesis of firm-level productivity growth following export market entry, although some studies, especially for developing countries, do report such a link. (See, for instance, Loecker, 2007; Topalova, 2004; Biesebroeck, 2005; and the survey by Girma, Greenaway and Kneller, 2004.) However, this distinction – based on the timing of the export market entry – has been blurred given the evidence from some recent studies that firms make innovation/technology use decisions based on current or anticipated export market participation, as highlighted by Bustos (2006), Verhoogen (2007), and Trefler and Lileeva (2007). In such a case, productivity and exporting decisions are both endogenous with respect to one another, and the timing of the export market entry can no longer be used to identify causality. (Yeaple, 2005, theoretically studies this joint technology adoption and export decision by firms, and explores the consequences for the return to skill – highlighting how skill-biased technological change may be induced by trade.)

Nonetheless, the results obtained clearly indicate that it is initially more successful firms that make the joint decisions concerning innovation (or ‘higher’ technology use) and export status. In other words, the least successful firms overwhelmingly tend to undertake neither activity.

Another part of the recent empirical literature using micro-level data has examined the consequences of this link between export status and productivity when the exposure to trade is changing (predominantly because trade costs are decreasing over time). In such a case, trade liberalization induces some reallocations between exporters and non-exporters competing in the same sectors (see Tybout, 2003, for a survey of this literature). One influential such study by Pavcnik (2002) finds that most of the 25 per

cent productivity increase in export-competing sectors in Chile between 1979 and 1986 is explained by reallocations between producers (generated by entry, exit, export market entry, and market share reallocations). However, since significant changes in trade regimes are also part of a larger set of substantial macroeconomic changes for the involved countries (as was the case for Chile), it nevertheless remains difficult to associate this type of reallocation-induced productivity growth to the direct effects of trade liberalization. One notable exception is Bernard, Jensen and Schott (2006), who show that reductions in trade costs for US plants substantially increase both the probability of exit and that of exporting among non-exporters. Given the productivity advantage of exporters, this induces reallocations in favour of the more productive exporting plants and hence increases average industry productivity (which is also confirmed by Bernard et al., 2006, as a result of the decrease in trade costs).

Clearly, these empirical patterns cannot be addressed by trade models based on representative firms. Such models, by construction, predict that trade affects all firms in a sector in similar ways. (Note that extensive firm-level heterogeneity per se is not necessarily problematic for a representative firm model of trade so long as firms, on average, respond in similar ways to trade. However, the evidence reviewed clearly shows that this is not the case.) In response to this empirical evidence, theoretical models of trade have been developed to incorporate firm-level productivity differences, and analyse the consequences for the effects of trade liberalization. One class of models, developed by Bernard et al. (2003) and Eaton and Kortum (2008), introduce stochastic firm productivity into the multi-country Ricardian model analysed in Eaton and Kortum (2002). In this class of models, there is a fixed number of products that can be produced by competing firms in all countries. All these firms (both in the same country but also across countries) use different technologies to produce the same good (based on a stochastic productivity draw) – hence the Ricardian framework. Consumers in any given country buy each good from the lowest-cost producer across all countries. Due to trade costs, several firms producing the same good can survive if they are located in different countries (although each firm is the sole supplier to any given destination). This model thus emphasizes the resulting competition between firms to be this exclusive supplier. Bernard et al. (2003) show how such a model can be calibrated to fit both micro-level data on US producers and macro-level data on cross-country trade and aggregate production across countries. The calibrated model can then be used to analyse many counterfactual predictions involving the consequences of trade liberalization.

Another class of models developed in Melitz (2003) and Melitz and Ottaviano (2005) eschews the analysis of the direct competition between firms to produce the same good by using a monopolistic competition framework: each firm produces its own distinctive differentiated good. These models incorporate firm heterogeneity into the one-sector models of intra-industry trade (the ‘new’ trade theory) developed in Krugman (1979; 1980). In this type of model, the product variety available to consumers in any given country varies endogenously with the characteristics of the country and the trade costs linking it to its trading partners (these affect the endogenous number of varieties produced domestically, as well as the endogenous fraction of firms from all trading partners that export to that country). Firms face sunk costs of entry, along with uncertainty concerning their future productivity (or also possibly the quality of the differentiated good that is under development). Upon entry, each firm instantaneously learns about its productivity level, modelled as a draw from a known distribution. Due to the sunk nature of the entry costs, firms with heterogeneous productivity levels remain active and produce. The least productive firms face negative profits and therefore exit. As exporting is costly, only the relatively more productive

firms (among those surviving) choose to export, while the remaining firms only serve their domestic market. Exporting is not profitable for these firms, either because it involves fixed or sunk costs, or because import demand is driven to zero at prices below the firms' delivered cost.

Both classes of models predict that trade liberalization induces the type of reallocations between firms that was previously described: the least productive firms are constrained to exit, new firms enter the export market, and market shares are reallocated towards more productive firms. These reallocations generate both aggregate productivity and welfare gains. Both classes of models also predict an important empirical regularity regarding bilateral trade flows: that differences in these trade flows reflect both differences in the amount of each good traded (the intensive margin of trade) as well as differences in the number of goods traded (the extensive margin of trade). (See Bernard, Jensen and Schott, 2005; Broda and Weinstein, 2006; Broda, Greenfield and Weinstein, 2006; Eaton, Kortum and Kramarz, 2004; and Kehoe and Ruhl, 2003, for some empirical applications.) Helpman, Melitz and Rubinstein (2007) and Chaney (2006) show how the framework of Melitz (2003) can be extended to derive a gravity specification for bilateral trade flows where trade costs affect both the extensive and intensive margins of trade. Both papers highlight the empirical importance of incorporating changes in trade at both margins.

Due to the absence of strategic interactions between firms, the monopolistic competition model of Melitz (2003) provides a convenient framework for the modelling of additional firm-level decisions in an open economy environment – where heterogeneous firms self-select into different types of activities. This framework can thus also explain why only a fraction of firms choose to become multinationals and operate foreign affiliates (horizontal foreign direct investment, FDI) as in Helpman, Melitz and Yeaple (2004) or integrate with their foreign suppliers (vertical FDI) as in Antras and Helpman (2004). (Helpman, 2006, provides a much more extensive review of the related models.) Additionally, other firm-level decisions that are also affected by the exposure to international trade can be incorporated: the choice of technology as in Acemoglu, Antras and Helpman (2007), the level of investment in innovation as in Atkeson and Burstein (2006), or the range of products produced and exported within multi-product firms as in Bernard, Redding and Schott (2006). Lastly, the structure from Melitz (2003) has also been fruitfully integrated into various other types of models that rely on the basic monopolistic competition of trade. This includes extension to two-sector models of trade with comparative advantage and factor proportion differences (Bernard, Redding and Schott, 2007), open economy models of growth (Baldwin and Robert-Nicoud, 2006), and international macro-dynamics (Ghironi and Melitz, 2005). In each case, the addition of firm-level heterogeneity allows the models to explore additional important features upon which a model with representative firms remains silent.

See Also

- international trade theory
- international trade, empirical approaches to
- new economic geography
- Ricardian trade theory
- trade, technology diffusion and growth

Bibliography

Acemoglu, D., Antras, P. and Helpman, E. 2007. Contracts and technology adoption. *American Economic Review* 97, 916–43.

Antras, P. and Helpman, E. 2004. Global sourcing. *Journal of Political Economy* 112, 552–80.

Atkeson, A. and Burstein, A. 2006. Trade costs, pricing to market, and international relative prices. Mimeo, UCLA.

Baldwin, R.E. and Robert-Nicoud, F. 2006. Trade and growth with heterogenous firms. Working Paper No 12326. Cambridge, MA: NBER.

Bernard, A.B. and Jensen, J.B. 1999. Exceptional exporter performance: cause, effect, or both? *Journal of International Economics* 47, 1–25.

Bernard, A.B., Jensen, J.B. and Schott, P.K. 2005. Importers, exporters, and multinationals: a portrait of firms in the US that trade goods. Working Paper No. 11404. Cambridge, MA: NBER.

Bernard, A.B., Jensen, J.B. and Schott, P.K. 2006. Trade costs, firms and productivity. *Journal of Monetary Economics* 53, 917–37.

Bernard, A.B., Redding, S. and Schott, P.K. 2006. Multi product firms and trade liberalization. Working Paper No. 12782. Cambridge, MA: NBER.

Bernard, A.B., Redding, S. and Schott, P.K. 2007. Comparative advantage and heterogeneous firms. *Review of Economic Studies* 74, 31–66.

Bernard, A.B., Eaton, J., Jensen, J.B. and Kortum, S. 2003. Plants and productivity in international trade. *American Economic Review* 93, 1268–90.

Bernard, A.B., Jensen, J.B., Redding, S. and Schott, P.K. 2006. Firms in international trade. Working Paper No. 13054. Cambridge, MA: NBER.

Biesebroeck, J.Van. 2005. Exporting raises productivity in sub-Saharan African manufacturing firms. *Journal of International Economics* 67, 373–91.

Broda, C. and Weinstein, D.E. 2006. Globalization and the gains from variety. *Quarterly Journal of Economics* 121, 541–85.

Broda, C., Greenfield, J. and Weinstein, D. 2006. From groundnuts to globalization: a structural estimate of trade and growth. Working Paper No. 12512. Cambridge, MA: NBER.

Bustos, P. 2006. Rising wage inequality in the Argentinean manufacturing sector: the impact of trade and foreign investment on technology and skill upgrading. Mimeo, CREI, Universitat Pompeu Fabra.

Chaney, T. 2006. Distorted gravity: heterogeneous firms, market structure and the geography of international trade. Mimeo, University of Chicago.

Clerides, S.K., Lach, S. and Tybout, J.R. 1998. Is learning by exporting important? Micro dynamic evidence from Colombia, Mexico, and Morocco. *Quarterly Journal of Economics* 113, 903–47.

Eaton, J. and Kortum, S. 2002. Technology, geography, and trade. *Econometrica* 70, 1741–79.

Eaton, J. and Kortum, S. 2008. *Technology in the Global Economy: A Framework for Quantitative Analysis*. Princeton: Princeton University Press.

Eaton, J., Kortum, S. and Kramarz, F. 2004. Dissecting trade: firms, industries, and export destinations. *American Economic Review* 94, 150–4.

Ghironi, F. and Melitz, M.J. 2005. International trade and macroeconomic dynamics with heterogeneous firms. *Quarterly Journal of Economics* 120, 865–915.

Girma, S., Greenaway, D. and Kneller, R. 2004. Does exporting increase productivity? A micro-econometric analysis of matched firms. *Review of International Economics* 12, 855–66.

Helpman, E. 2006. Trade, FDI, and the organization of firms. *Journal of Economic Literature* 44, 589–630.

Helpman, E., Melitz, M.J. and Yeaple, S.R. 2004. Export versus FDI with heterogeneous firms. *American Economic Review* 94, 300–16.

Helpman, E., Melitz, M.J. and Rubinstein, Y. 2007. Estimating trade flows: trading partners and trading volumes. Working Paper No. 12927. Cambridge, MA: NBER.

Kehoe, T.J. and Ruhl, K.J. 2003. How important is the new goods margin in international trade? Mimeo.

Krugman, P.R. 1979. Increasing returns, monopolistic competition, and international trade. *Journal of International Economics* 9, 469–79.

Krugman, P.R. 1980. Scale economies, product differentiation, and the pattern of trade. *American*

Economic Review 70, 950–9.

Loecker, J.De. 2007. Do exports generate higher productivity? Evidence from Slovenia. *Journal of International Economics* (forthcoming).

Melitz, M.J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71, 1695–725.

Melitz, M.J. and Ottaviano, G.I.P. 2005. Market size, trade, and productivity. Working Paper No. 11393. Cambridge, MA: NBER.

Pavcnik, N. 2002. Trade liberalization, exit, and productivity improvements: evidence from Chilean plants. *Review of Economic Studies* 69, 245–76.

Topalova, P. 2004. Trade liberalization and firm productivity: the case of India. Working Paper No. 04/28, International Monetary Fund.

Trefler, D. and Lileeva, A. 2007. Does improved market access raise plant-level productivity? Mimeo.

Tybout, J.R. 2003. Plant- and firm-level evidence on ‘new trade theories’. In *Handbook of International Trade*, ed. E.K. Choi and J. Harrigan. Oxford: Basil Blackwell.

Verhoogen, E. 2007. Trade, quality upgrading and wage inequality in the Mexican manufacturing sector: theory and evidence from an exchange-rate shock. Mimeo.

Yeaple, S.R. 2005. Firm heterogeneity, international trade, and wages. *Journal of International Economics* 65, 1–20.

How to cite this article

Melitz, Marc J. "international trade and heterogeneous firms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000302> doi:10.1057/9780230226203.0838

The New Palgrave Dictionary of Economics Online

international trade theory

James E. Anderson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International trade theory provides explanations for the pattern of international trade and the distribution of the gains from trade. The theory convinces most economists of the benefits of liberal trade. But many non-economists oppose liberal trade. Opponents include some who may have encountered trade theory but nevertheless fall prey to fallacious reasoning. This article attempts to convey why trade theory is so persuasive to economists and also to deal with why many non-economists are not persuaded.

Keywords

absolute advantage; adjustment costs; autarky; bilateral trade; comparative advantage; comparative labour productivity advantage; competitive advantage; consumption taxation; economic theory of gravity; economies of scale; endogenous advantage; external budget constraint; factor immobility; factor proportions theory; free trade; gains from trade; general equilibrium trade model; Heckscher–Ohlin trade theory; imperfect competition; information limitations; international trade theory; lump-sum transfers; monopolistic competition; multilateral resistance; multiple equilibria; partial equilibrium; product variety; productivity shocks; protection; Ricardian trade theory; substitution effect

Article

Why do nations trade the products they do? Is trade a good thing? The theory of international trade provides answers. The answers are both convincing and elegant; hence, the vast majority of economists agree about the desirability of liberal trade. But the argument is also subtle and often misunderstood or distorted. Thus, a large proportion of the general population tends to oppose liberal trade from confusion. This article attempts to convey why the answers convince most economists and why their liberal trade position is so often misunderstood. The article's focus is theory, but theory convinces when it succeeds in fitting the data. Thus, passing reference will be made to empirical findings, a sensibility much more thoroughly developed in the graduate textbook of Feenstra (2003).

‘Buy low, sell high’ logic leads economists to comparative advantage theory. Comparative advantage means the comparison of *relative* price differences between nations to explain the pattern of trade. For example, compare the relative price of wheat in terms of cheese at home with the same relative price in the foreign economy in a hypothetical equilibrium with no trade (autarky) or with restricted trade. The country with the lower relative price of wheat is said to have a comparative advantage in wheat while the other country has, symmetrically, a comparative advantage in cheese. Buy low, sell high logic predicts that a country will export the good in which it has a comparative advantage. (In the case of many goods, the prediction is that a country will on average export goods which are relatively cheap in the absence of trade and import goods which are relatively expensive in the absence of trade. The prediction is about correlation. Bernhofen and Brown, 2004, show that Japan's opening to trade in the 1850s reveals data consistent with the prediction.)

Notice that the focus on relative prices tends to cancel out forces (exchange rate manipulations, environmental or labour standards) which cause national differences in levels of non-traded factor (or goods) prices. Note also that by this reasoning a country *must* have a comparative advantage in some good. Prices of non-traded factors of production adjust in general equilibrium so that each country ends up in the trade equilibrium with a competitive or absolute cost advantage in the good in which it has a comparative advantage. Partial equilibrium thinking takes factor prices as given and does not impose the external budget constraint that requires exports to pay for imports. Partial equilibrium reasoning leads to misunderstandings explored below as the absolute advantage fallacy.

Comparative advantage differences between nations are explained by exogenous differences in national characteristics. Labour differs in its productivity internationally and different goods have different labour requirements, so comparative labour productivity advantage was Ricardo's predictor of trade patterns. Ricardian trade theory is useful in its simplicity and even rather loosely confirmed by empirical evidence. The factor proportions theory added relative factor endowment differences to the exogenous explanation of comparative advantage (Jones, 1987). More capital-abundant countries have higher labour productivity, but the advantage gained relative to the less abundant countries varies with the relative capital intensity of the good's technology. Combining technology and endowment differences appears to account well for actual trade patterns (Davis and Weinstein, 2002).

Trade theory also encompasses endogenous differences between countries. One focus is on economies of scale. The wider market due to trade induces a cost advantage in an industry in one of the countries. Another theory is based on monopolistic competition, whereby the wider markets due to trade increase product variety as buyers seek the special characteristics of foreign brands. Differentiated products trade flows both ways within product categories.

Trade costs also shape the pattern of trade. The economic theory of gravity explains the complex bilateral trade patterns among countries. Actual trade is much lower than gravity predicts in a frictionless world, providing evidence of trade costs much larger than those due to policy or transportation. The costs are well explained by geography and a set of national differences. The stability of the relationships over time suggests that these costs change slowly.

There are gains from trade in all these models. But the division of the gains will be uneven and there will be losers. Distribution matters in two ways, between and within nations. Internationally, with only mild qualifications, gains are shared between nations: some trade is better than none. Each nation can act through trade policy to take more of the gain, however, leading to destructive trade wars with mutual losses. Within national economies, there are gains on average but there are ordinarily losers. National

institutions act to redistribute some of the gains (US Trade Adjustment Assistance) or provide temporary relief from losses due to trade (escape clause protection), at the cost of lowering the overall gain from trade.

The topics of this outline are developed below in more detail. Section 1 examines the causes of comparative advantage. Section 2 exposes the absolute advantage fallacy. Section 3 reviews endogenous advantage. Section 4 sets out the economic theory of gravity and its implications. The concluding section examines the gains from trade.

1 Comparative advantage

Ricardo explained comparative advantage as due to differences in labour productivity. Suppose that it takes two hours of labour to produce a bushel of wheat in the home country, while it takes four hours of labour to produce a bushel of wheat in the foreign country. Also, it takes three hours of labour to produce a pound of cheese in the home country while it takes eight hours of labour to produce a pound of cheese in the foreign country.

Ricardo saw that the world trade equilibrium would result in the home country exporting cheese and the foreign country exporting wheat. This is because, in the absence of trade, a pound of cheese is worth 1.5 bushels of wheat (three hours per pound of cheese divided by two hours per bushel of wheat) in the home country while a pound of cheese is worth two bushels of wheat in the foreign country. The labour market equilibrium which accompanies such a trade equilibrium must have a foreign wage of at most one-half of the home wage (since, with a foreign wage equal to one-half the home wage, a bushel of wheat costs the same amount in each country, allowing production in both). If we consider a low-wage foreign economy, the labour market equilibrium accompanying the trade equilibrium could have a foreign wage no lower than three-eighths of the home wage (since in this case a pound of cheese costs the same amount in each country).

Notice that countries export the good in which they have the comparative labour productivity advantage, cheese for the home country and wheat for the foreign country. The numbers chosen make no difference to the logic; what is essential is that comparative labour productivities differ. One special aspect of the numbers deserves emphasis, however: the home country has an absolute labour productivity advantage in both goods yet trade occurs regardless.

Subsequent developments of trade theory generalized the production model. The essence of comparative advantage theory remains: trade is due to differences in relative prices that would obtain in the absence of trade, and an average of each country's citizens gain from such trade. The Heckscher–Ohlin analysis of the factor proportions model predicted that a country would have a comparative advantage in the good which made relatively intensive use of its relatively abundant factor. Thus, if the home country were relatively abundant in capital (which would explain why its labour was so much more productive in the preceding example), it would have a comparative advantage in the good which used capital relatively intensively (cheese in the preceding example). Conversely, the foreign country is relatively abundant in labour and has a comparative advantage in the good which uses labour relatively intensively (wheat in the example above).

Trade in goods compensates for the international immobility of factors. The factor content extension of Heckscher–Ohlin trade theory predicts that trade patterns permit each country to consume factor services as if it were in a completely integrated world, smoothing out differences in national factor endowments.

Recent empirical work has met with striking success in combining factor endowment differences with technology differences as an explanation of observed trade patterns (Davis and Weinstein, 2002).
Comparative advantage theory is much more general than the preceding discussion of special cases (Deardorff, 1984), but predictions about the pattern of trade weaken with generality. On average a country will import goods that would be relatively expensive in the absence of trade. See the Appendix for a technical statement. See Bernhofen and Brown (2004) for confirming evidence based on Japan's opening to trade in the 1850s. The assumptions of the general model are that (*a*) price-taking consumers minimize the expenditure needed to realize any level of utility (real income), and (*b*) producers behave so as to maximize the national product given the resource endowments. Assumption (*a*) implies downward-sloping demand curves in the generalized form. Assumption (*b*) leads to upward-sloping supply curves in the generalized form. Scale economies and imperfect competition, treated below in the section on endogenous advantage, can lead to the violation of assumption (*b*).

2 The absolute advantage fallacy

Businessmen naturally compare the money cost of the same good in different locations to draw inferences about the direction of trade. Absolute cost advantage appears to imply that a nation imports goods that are cheaper abroad and exports goods that are more expensive abroad. The reasoning is insidious because it makes sense in many contexts. Absolute advantage appropriately addresses the householder's question of which good should be purchased, the businessman's question of how tough his competitors are. The individual businessman can appropriately take all other prices as given when contemplating his own actions, such as entering a new export market.
To see the difference between absolute and comparative advantage reasoning clearly, return to the Ricardian example above. If wages (measured in a common currency) were equal in the two countries prior to the opening of trade, the home country would have a ‘competitive’ or absolute advantage in both goods: it could undersell the foreign country in both wheat and cheese. Foreign businessmen would naturally be worried that they would all be driven from the market. This universal bankruptcy could not be an equilibrium, however, because the foreign workers would have no income to pay for home-produced goods. The imbalance between expenditure and income would also mirror the absence of exports to pay for imports. Market equilibrium would be reached through price changes, lowering the foreign wage or raising the home wage until the foreign workers could be employed in the industry in which the foreign economy has the comparative advantage. (Unless the two currencies were pegged, the exchange rate of the foreign economy could depreciate and create the same effect.) More general models of production lead to the same conclusion: *equilibrium costs will adjust to confer absolute advantage in the good in which each country has a comparative advantage*.
With many goods, comparative advantage applies to ranges of goods rather than to a single good, and the dividing line between comparative advantage and disadvantage is endogenous. The absolute advantage is weak in the mathematical sense in the case where both countries continue to produce the good.
Another illustration of the absolute advantage fallacy arises in popular concerns about the rapid productivity growth of China compared with that of the United States. A ten per cent improvement in productivity will indeed secure a ten per cent cost advantage for the businessman over his competitor. A

ten per cent improvement in all Chinese productivity relative to the United States is unlikely to change comparative advantage (indeed, in the Ricardian example, comparative labour productivity advantage is unchanged) because Chinese wages will rise relative to US wages. Similarly, a ten per cent drop in all US productivity due to tighter environmental regulations will be unlikely to change comparative advantage because US factor returns will fall.

The widespread practice of making international comparisons of ‘competitive advantage’ is essentially misguided because it suggests the metaphor of a race. The race metaphor is extended in concerns about ‘a race to the bottom’, which supposedly expresses the dilemma of countries seeking to implement pollution or labour standards but being pressured to lower standards by their competition with foreign countries that have low standards. But nations do not ‘compete’ as firms do. A firm may well be unable to survive after implementing pollution reduction when its competitors abroad do not follow suit and no other prices change in the new equilibrium. Nations cannot similarly put themselves out of business because factor prices will change in the new equilibrium. Polluting industries may or may not survive at the new factor prices under the new regulations, but the nation's factors will be productively employed somewhere in the economy. Pollution reduction is costly with or without trade; nothing about the nature of a trading economy makes any essential difference to the nation's ability to implement desired standards. The desirability of trade is an *essentially* separate matter.

3 Endogenous advantage

Many goods are traded because they are simply unavailable from local production. Some kinds of availability are exogenous to the interaction of nations – diamonds and oil are found only in a few locations. Endogenous availability is in contrast driven by advantage arising from the economic interaction of nations. Endogenous advantage normally coexists with comparative advantage but it is simpler to consider special cases independent of comparative advantage. Theory focuses on endogenous advantage resulting from economies of scale. (In a formal but trivial sense, oil or diamond trade can be seen as comparative advantage trade – big oil deposits lead to a low relative price of oil where they are found. Moreover, comparative advantage trade is often associated with the disappearance of some industries in some countries. Neither of these associations of comparative advantage with availability is essential to the model, however.)

Trade based on scale economies features the possibility of multiple equilibria – one country will produce a good with scale economies but which nation ends up producing it can be a matter of chance. Since advantage is endogenous, it appears attractive in developing countries to attempt to reverse the historical head start of rich countries by starting up production behind protection and then later being able to compete on world markets. The record of success in such efforts is mixed.

Openness to trade will generally allow economies of scale to be more thoroughly exploited, so this is a new source of gains from trade. Moreover, wider markets may support a wider range of products, still another source of gains from trade. Each country shares in the gains from trade with scale economies under conditions that appear to be met in practice. (This claim is based on the results from numerous simulation models of trading economies that have been developed since the mid-1970s.) The theoretical possibility that a country can lose from trade based on scale economies has drawn a lot of attention from development economists in particular (Ethier, 1982b). (Losses result when a trading equilibrium has a country importing the good with scale economies while still producing it. Since domestic scale is

smaller, unit costs are higher, meaning that market forces perversely ‘choose’ to import a good with higher price than in autarky. Simulation models have not found such equilibria but they are possible.) Gains can be guaranteed if a country expands production in goods with scale economies, so it looks more attractive to use policy to promote production of such goods.

Scale economies come in two forms: external to the firm and internal to the firm. External scale economies are typified by specialized labour markets such as Silicon Valley, where the concentration of the market reduces search costs for computer engineers. External scale economies need not be location-specific, however. Increases in the scale of downstream final production can permit carrying on upstream input production with a specialized process that is cheaper at large enough scale. Such scale economies can operate at the level of the world economy and appear to be bound up with the recent phenomenon of outsourcing (Ethier, 1982a). Global scale economies tend to guarantee mutual gains from trade among countries.

Internal scale economies are associated with imperfect competition when the size of the firm looms large relative to the market size. Trade tends to intensify competition and thus to reduce the inefficiency of monopoly, another gain from trade.

The most fruitful form of imperfect competition for trade theory has been monopolistic competition. Only Ford Motor Co. produces Ford autos (monopoly) but dozens of brands compete for auto buyers. Each design has a fixed cost of design (and marketing) which must be covered by sales net of variable cost. The total market size limits the number of designs which can profitably be produced. A signal accomplishment of trade theory in the 1980s was the embedding of monopolistic competition in a general equilibrium trade model (Helpman and Krugman, 1985; Ethier, 1982a). Progress was enabled by the simplifying assumption of symmetric firms: all brands were equally desirable and all firms’ costs were the same.

Monopolistic competition provides an explanation of the two-way international trade that is found in many products such as autos, and of why two-way trade is more prevalent between similar countries. Trade between rich and poor countries, in contrast, is explained mainly by comparative advantage as autos exchange for agriculture. Relative country size matters too, the home-market effect of Krugman (1980). Here the insight has been rigorously proved only for a two-country example. Start with two equally sized countries, then increase one relative to the other. Trade costs imply that the larger country will have a more than proportionally larger share of brands. Intuitively, with access to foreign markets being costly the home market, being larger, allows scale economies to be more readily exploited, increasing the larger country's share of differentiated goods production more than its share of world income.

Monopolistic competition theory has recently focused on the heterogeneity of firms. If the symmetry of firms on the demand side is retained, differences in firms’ productivities imply differential responses to trade. The best firms export disproportionately while imports drive out the worst firms. Fixed trade costs add explanatory power; only the best firms choose to incur the cost of trade. A key element of the model is productivity shocks, firms discover their productivity after committing fixed costs. The distribution of surviving firms is related to the distribution of productivity shocks as well as economic determinants. The models of Bernard et al. (2003) and Melitz (2003) deserve special attention. The former focuses on competition within a variety while the latter focuses on competition across varieties. Both models imply new gains from trade in the form of overall productivity gains: opening trade causes the exit of weak firms and the expansion of strong ones.

4 Bilateral trade patterns

The trade theories presented above are focused on explaining the cross-commodity trade pattern of essentially two trading countries. The contemporary world of more than 100 countries (most of which are collections of distinct economic regions) has complex trade patterns.

The economic theory of gravity complements the preceding models by providing an explanation of bilateral trade (Anderson and van Wincoop, 2004). Gravity fits the data well and reveals important information. The model is based on four assumptions: expenditure on goods from all sources is equal to income from sales to all sources, markets for all goods clear, (more restrictively) each country or region produces a unique good, and all countries have the same tastes for goods.

The third assumption – products differentiated by place of origin – appears to be the most restrictive. In practice, only models of this type do at all well in fitting bilateral trade patterns. Monopolistic competition provides one explanation for why products appear to be differentiated by place of origin. Eaton and Kortum (2002) show alternatively that productivity shocks in a Ricardian model will select producers within product lines, resulting at the aggregate level in what appears to be two-way trade. In either case, gravity ends up describing trade flows.

In a frictionless world, gravity theory predicts that the bilateral trade in a commodity as a share of world production of the commodity will be equal to the product of the source country's share of world production of the commodity times the consuming country's share of expenditure on the commodity. Alternatively, the model predicts that size-adjusted trade, the bilateral flow divided by the product of source country supply and consuming country expenditure, should be constant across country pairs in a frictionless world.

Actual trade flows are far smaller than the frictionless prediction (while shipments within regions are far larger, home bias). The deviations of actual bilateral trade from the frictionless prediction allows inference about bilateral trade costs. Distance appears to be more costly than can be accounted for by transport costs. Other costs are associated with non-contiguity, language barriers, exchange rate barriers, insecurity and other plausible bilateral characteristics. Just crossing a border imposes a cost which is larger than can be explained by policy variables.

Trade flows in the model are predicted to vary with relative resistance, equal to the ratio of the direct bilateral trade cost to the product of inward and outward multilateral resistance. Multilateral resistance is an index of bilateral trade costs, inward from every source to a particular destination or outward from a particular source to every destination. Multilateral resistance is linked to country size and thus to explaining an important aspect of trade patterns. Since borders are costly, a big country tends to have lower multilateral resistance than does a small country because a smaller fraction of its shipments must cross borders. The size-adjusted internal trade of big countries will be smaller than that of small countries because big countries have higher relative resistance to their internal trade. These differences can be quite dramatic, as shown by studies of US and Canadian trade (Anderson and van Wincoop, 2003), where the United States is about ten times larger than Canada.

5 Division of the gains

Professional economists generally support liberal trade because theory and evidence persuade them that there are gains from trade in an average sense in all these models of the determinants of trade. But the division of the gains will be uneven and there can be losers. Most policy intervention with trade is explained by the policymakers' desire to alter the distribution of gains.

The gains from trade reasoning is illustrated with comparative advantage-based trade. Focus on a 'typical' household. Suppose that in autarky equilibrium, as in the Ricardian numerical example, the Home typical householder is willing to swap 1 unit of cheese for 1.5 units of wheat. That is, he would be indifferent to moving his consumption and production a small distance to offer the market 1 cheese for 1.5 wheat or 1.5 wheat for 1 cheese. Suppose that a typical foreign country household in the autarky equilibrium is willing to swap 2 wheat for 1 cheese. Now allow frictionless trade, and suppose for illustrative purposes that the new equilibrium price is equal to 1.75 wheat per unit of cheese. (Generally the price must lie between 1.5 and 2, always implying mutual gains.) Each Home household offers cheese to Foreign households in exchange for their wheat. Formerly it cost 2 wheat for 1 cheese in Foreign but now the 2 wheat will procure 1 cheese and leave 0.25 wheat left over, a gain from trade. Similarly, each Home household can obtain 1.75 wheat for 1 cheese where formerly this would procure only 1.5 wheat, a gain from trade of 0.25 wheat. Both households and hence both nations gain from trade. The numbers chosen illustrate a general principle: mutual gains result from trade when autarky relative prices differ. See the Appendix for a more formal discussion.

The mutual gains from trade claim may seem dubious because, with the numbers chosen, trade equilibrium requires that foreign wages must be lower than home wages. In effect, trade facilitates an exchange in which more than one unit of foreign labour exchanges for one unit of home labour – the home country is 'exploiting' foreign labour. Some anti-trade sentiment on the left in rich countries is based on this observation. (Marxism embeds the observation in a wider system of analysis, but it probably is no longer a basis for much sentiment on the left.) Nevertheless, foreign labour gains from trade, as does home labour. Prior to trade, a pound of cheese cost 1.5 bushels of wheat in the home country while it cost 2 bushels of wheat in the foreign country. By specializing in wheat production and exchanging it for cheese, foreign workers can obtain cheese more cheaply, at a price somewhere between 2 and 1.5 bushels of wheat. This exchange must make them better off. As for home workers, prior to trade, a pound of cheese obtained 1.5 bushels while with trade it obtains somewhere between 1.5 and 2 bushels. This must make home workers better off. Concern about the 'fairness' of the exchange in rich countries should lead to policies which might actually help the poor countries. Trade theory shows that anti-trade policies by rich countries will instead on average harm the poor countries.

Scale economies and imperfect competition models of trade suggest further gains. With scale economies, trade implies that the force of wider markets drives costs lower. With imperfect competition, trade stimulates competition and drives profit margins lower. Trade equilibrium with monopolistic competition suggests that consumers and intermediate input users gain from more variety of differentiated products (see Helpman and Krugman, 1985).

The distribution of the gains matters, both between and within nations. Nationalist trade policy can take more of the gains, leading to destructive trade wars with mutual losses. Negotiation of trade agreements and their enforcement through international institutions such as the World Trade Organization (WTO) help to restrain the destructive tendencies of unilateral action. Nations have an incentive to participate in negotiations and to join institutions such as the WTO because some trade is better than none for each nation. Theoretical qualifications to this statement must be entered in models of trade involving scale

economies and imperfect competition, but in practice simulation of such models suggests that some trade remains better than none for each nation.

Within national economies the division of gains issue is much sharper: some members of a nation ordinarily lose from trade. Ricardo's one-factor trade model submerges income distribution. Multi-factor production models feature groups who must lose from trade. Loosely speaking, these groups are associated with import competing production (see Jones, 1987).

In equilibrium the gains must ordinarily outweigh the losses within each nation, by the preceding national-gains-from-trade argument. For an economy with non-identical households, this implies that there are gains from trade on average. Under special circumstances the gains can be redistributed so that all households gain. In practice, these circumstances are rarely met completely. Even so, most economists tend to favour efficiency-enhancing policies such as liberal trade on the pragmatic grounds that efficiency-reducing policies such as protection also cause gainers and losers, so it is better to go with the larger net gains and supplement them with feasible programmes to compensate the most obvious losers.

(A benevolent and very powerful government can in principle calculate and implement the lump-sum transfers – negative for gainers, positive for losers – that are required to achieve redistribution so that all gain. In practice, information is more limited and implementation more difficult – because households modify their behaviour to reduce their tax or increase their subsidy – than with the lump-sum story. Trade and public economic theory have relaxed the conditions somewhat. Income taxation can in some circumstances achieve redistribution with efficiency, but information limitations rule them out as a practical matter; see Guesnerie, 2001. Dixit and Norman, 1986, show that a system of consumption taxes – differentially taxing each commodity – that sacrifices some of the gains from trade is powerful in achieving gains for all. For a qualification of their argument, see Kemp and Wan, 1986. Again, information limitations vitiate the applicability of this idea. Finally, a government that can discriminate powerfully between households is sure to be lobbied intensively by those able to organize politically, to the detriment of the unorganized.)

What if losers are not compensated? A person taking this question seriously must decide on liberal trade by weighting individual gains and losses. Ethical considerations give more weight to losses or gains to the poor than to the rich. The case for liberal trade is strengthened by ethical considerations because the illiberal trade policies of rich countries hurt the poor disproportionately, as documented by Gresser (2002). Poor countries have comparative advantage based on cheap low-skilled labour, hence discrimination against their exports harms the poor citizens of poor countries. At home in rich countries, protection makes food and clothing more expensive, a regressive tax on poor consumers. Among the poor, losers from protection appear almost surely to outweigh gainers.

On the way to equilibrium, it is theoretically possible that adjustment cost losses may temporarily exceed gains, justifying temporary relief measures. For example, workers displaced by import competition may be unemployed for a time. Extensive investigation of US cases suggests that such adjustment cost losses from trade are small, of short duration, and are swamped by the gains from trade. A typical investigation reports that the net cost to the economy of using protection to re-employ a worker far exceeds the wage the worker would receive in the job, usually several (up to ten) times the wage. In practice, therefore, temporary protection for workers cannot be justified on efficiency grounds, though it remains possible to justify it on equity grounds. Economists in favour of liberal trade point out that protection can be replaced with much less inefficient methods of compensation to displaced workers.

A substantial part of the opposition to liberal trade is based on confusion and ignorance. Confusing absolute advantage with a valid theory of trade sows fear that a nation must protect itself from overwhelming competition. Greatly exaggerated notions of the size of adjustment costs leads to support for protection. Ignorance of the harm done to the world's poor by protection persuades many who support redistribution of income to support protection that harms the majority of those they seek to help. The combination of confusion and ignorance among the ‘disinterested’ with well-organized special interest groups explains the power of protectionism.

Appendix

The general statement of comparative advantage is that on average a country will import goods that are relatively expensive in autarky. Let *m* denote the vector of excess demands in equilibrium, positive for imports and negative for imports. Let *p* denote the vector of relative prices in autarky in the home country and let p^* denote the vector of relative prices in autarky in the foreign country. Then the vector inner product $(p - p^*)'m \geq 0$.

The key requirement for the proposition is ‘as if’ optimization by consumers and producers, leading downward-sloping demand and upward-sloping supply in the generalized sense (the substitution effects matrix of real income-compensated excess demands, m_p^ξ, is negative semi-definite). If the actual trade equilibrium involves trade distortions, the additional requirement is that trade not be on balance subsidized. Let *t* be the vector of trade taxes, positive for import taxes and negative for export taxes (and negative for import subsidies and positive for export subsidies). The requirement is $t'm \geq 0$.

The ‘buy low, sell high’ logic implies that a surplus is captured by trade, so comparative advantage trade is closely linked to the gains from trade. ‘As if’ optimization means that consumers lower the expenditure required to support given real income by reallocating consumption in trade equilibrium as compared with autarky, while optimization by producers means that income is raised by reallocating production in trade equilibrium as compared with autarky.

Similar comparative advantage statements can be made concerning the factor content of trade; countries tend to import (embodied in goods) the factors that are relatively expensive in autarky (see Neary and Schweinberger, 1986).

See Also

- Heckscher–Ohlin trade theory

Bibliography

Anderson, J. and van Wincoop, E. 2003. Gravity with gravitas: a solution to the border puzzle. *American Economic Review* 93, 170–92.

Anderson, J. and van Wincoop, E. 2004. Trade costs. *Journal of Economic Literature* 42, 691–751.

Bernard, A., Eaton, J., Jenson, B. and Kortum, S. 2003. Plants and productivity in international trade. *American Economic Review* 93, 1268–90.

Bernhofen, D. and Brown, J. 2004. A direct test of the theory of comparative advantage: the case of Japan. *Journal of Political Economy* 112, 48–67.

Davis, D. and Weinstein, D. 2002. An account of global factor trade. *American Economic Review* 91, 1423–53.

Deardorff, A. 1984. The general validity of the law of comparative advantage. *Journal of Political Economy* 88, 941–57.

Dixit, A. and Norman, V. 1986. Gains from trade without lump-sum compensation. *Journal of International Economics* 21, 111–22.

Eaton, J. and Kortum, S. 2002. Technology, geography and trade. *Econometrica* 70, 1741–79.

Ethier, W. 1982a. National and international returns to scale in the modern theory of international trade. *American Economic Review* 73, 389–405.

Ethier, W. 1982b. Decreasing costs in international trade and Frank Graham's argument for protection. *Econometrica* 50, 1243–68.

Feenstra, R. 2003. *Advanced International Trade: Theory and Evidence*. Princeton: Princeton University Press.

Gresser, E. 2002. Toughest on the poor. *Foreign Affairs* 81(6), 9–14.

Guesnerie, R. 2001. Second best redistributive policies: the case of international trade. *Journal of Public Economic Theory* 3, 15–25.

Helpman, E. and Krugman, P. 1985. *Market Structure and Foreign Trade*. Cambridge, MA: MIT Press.

Jones, R. 1987. Heckscher–Ohlin trade theory. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Kemp, M. and Wan, H. 1986. Gains from trade with and without lump-sum compensation. *Journal of International Economics* 21, 99–110.

Krugman, P. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70, 950–9.

Melitz, M. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71, 1695–725.

Neary, J. and Schweinberger, A. 1986. Factor content functions and the theory of international trade. *Review of Economic Studies* 53, 421–32.

How to cite this article

Anderson, James E. "international trade theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000263> doi:10.1057/9780230226203.0839

The New Palgrave Dictionary of Economics Online

international trade, empirical approaches to

Stephen J. Redding

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews empirical research in international trade, which has undergone a resurgence since the mid-1980s. The article begins with traditional trade empirics, in which cross-country differences in opportunity costs of production (comparative advantage) are the basis for trade, before turning to new trade empirics, in which consumer love of variety and increasing returns to scale give rise to trade in similar goods between similar countries. More recent empirical research has emphasized heterogeneity across products within industries and across individual plants and firms, while other recent work has focused on the political economy of trade policy.

Keywords

autarky; bilateral trade; campaign contributions; Canada–US Free Trade Agreement; comparative advantage; constant elasticity of substitution (CES) preferences; elasticity of substitution; endowments paradox; factor content of trade; factor endowments; factor price equalization; factor price equalization theorem; factor service trade; General Agreement on Tariffs and Trade (GATT); gravity equation; Heckscher–Ohlin trade theory; Hecksher–Ohlin theorem; international trade; international trade and heterogenous firms; intra-industry trade; Japan, economics in; location of production; missing trade; Most Favored Nation (MFN) principle; multilateral trade liberalization; neoclassical trade theory; new trade theory; non-neutral technology differences; opportunity costs of production; outsourcing; political competition; protection; Reciprocity; regional and preferential trade agreements; Ricardo, D.; Rybczynski theorem; skill-biased technical change; Stolper–Samuelson theorem; vertical specialization; wage heterogeneity; World Trade Organization (WTO)

Article

Traditional trade empirics

The idea that comparative advantage provides an explanation for ‘inter-industry trade’ (the international exchange of one set of goods for another) dates back to Ricardo (1817), who emphasized technology differences as the source of cross-country variation in opportunity costs of production. While some early empirical studies adopted a Ricardian perspective (for example, MacDougall, 1951), much of the empirical analysis of traditional trade frameworks has been concerned with the Heckscher–Ohlin (HO) model (Heckscher, 1919; Ohlin, 1924). In contrast to its Ricardian counterpart, the HO model assumes that countries have identical technologies, and instead emphasizes variation in country factor endowments and industry factor intensities as the source of differences in opportunity costs of production.

The stylized version of the HO model assumes two factors of production (capital and labour), two countries (one capital-abundant), and two goods (one capital-intensive at all factor prices). In this stylized case, the model yields four sharp predictions: (*a*) the HO theorem – the capital-abundant country exports the capital-intensive good; (*b*) the factor price equalization theorem – with diversified production, international trade equalizes factor prices; (*c*) the Stolper–Samuelson theorem – with diversified production, an increase in the relative price of the labour-intensive good raises the relative and real return to labour and reduces the relative and real return to capital; (*d*) the Rybczynski theorem – with diversified production, an increase in the endowment of labour leads to a more than proportionate increase in the output of the labour-intensive good and reduces output of the capital-intensive good.

Early empirical examinations of the HO model were loosely motivated by these four theorems. In seeking to test the HO theorem, Leontief (1953) found that US exports were less capital-intensive than US imports, which appeared paradoxical within the confines of the stylized HO model. The key to resolving this paradox in Leamer (1980) was in rigorously deriving the correct empirical predictions directly from the theory. Indeed, a distinguishing feature of recent empirical studies of the HO model has been the derivation of empirical specifications from general equilibrium trade theory and the explicit recognition of the complexity of the model's predictions with many goods and factors of production.

With many goods and factors of production, and in the absence of trade costs, the theorems of the HO model are considerably weaker than in the 2×2×2 stylized version, and hold only as averages or correlations. We begin by examining predictions for international trade (the generalization of the HO theorem). The many-good, many-factor version of the model does not predict the pattern of trade in individual goods, but does predict the pattern of trade in individual factor services. A country that is abundant in a factor is predicted to be a net exporter of the factor, where factor abundance is defined as an endowment exceeding the country's share of world consumption times the world factor endowment. Therefore, many empirical studies of the HO model have focused on its predictions for net trade in factor services. Following Leamer's (1984) early and influential treatment, Bowen, Leamer and Sveikauskas (1987) were the first to observe that a full test of the model's predictions for factor service trade requires three sets of separate data on international trade, factor input requirements and factor endowments. Early empirical results were discouraging from the point of view of the explanatory power of the theory. Bowen, Leamer and Sveikauskas (1987) found that the HO model performed no better than a coin toss in predicting the direction of a country's net trade in factor services. In response, Trefler (1993) argued that factor-augmenting technology differences could both explain patterns of trade in factor services and account for cross-country variation in factor prices. Under this hypothesis, first

mooted in Leontief (1953), the HO model's predictions for factor service trade and of factor price equalization hold only after one controls for cross-country differences in the efficiency of factors of production. In subsequent work, Trefler (1995) identified two systematic departures between predicted and measured net trade in factor services: (*a*) ‘The Case of the Missing Trade’, where measured factor services trade is close to zero and much smaller than predicted by the HO model; and (*b*) ‘The Endowments Paradox’, where rich countries are scarce in most factors and poor countries are abundant in most factors.

One strand of recent research has argued that the HO model's predictions for factor service trade are much closer to the measured values for trade between regions within countries, where the model's assumptions of identical technologies, factor price equalization and identical and homothetic preferences are more likely to be satisfied. Davis et al. (1997) provide evidence supporting the HO model's predictions using data for trade between Japanese regions. A second strand of research has argued that factor-augmenting technology differences are not enough to explain international trade in factor services, but that a reconciliation between theory and data is ultimately possible. Davis and Weinstein (2001) provide evidence that international trade in factor services can be successfully explained if the HO model's assumptions are relaxed to introduce cross-country differences in technology that vary between industries (‘non-neutral’ technology differences), trade costs and non-factor price equalization. While predicted and measured net factor service trade has been brought into line, the model is radically transformed by relaxing these assumptions.

We now turn to the predictions of the many-good, many-factor HO model for the international location of production (the generalization of the Rybczynski theorem). With an equal number of goods and factors of production and factor price equalization, the HO model implies a linear relationship between production and factor endowments. Estimating this relationship using cross-country data, Harrigan (1995) finds statistically significant coefficients on factor endowments, but large within-sample prediction errors, suggesting that the model performs poorly in explaining the international location of production. Gandal, Hanson and Slaughter (2004) and Hanson and Slaughter (2002) examine the HO model's prediction that, in an equilibrium where factor prices are pinned down by goods prices, changes in factor endowments should be absorbed through changes in output mix. Using immigration data for Israel and US states, they find some evidence in support of the model's prediction. More recent research reinforces conclusions from the analysis of net factor services trade by suggesting that non-neutral technology differences across industries are important for explaining the international location of production. In an influential paper, Harrigan (1997) estimates an equation for the share of sector in GDP derived from the neoclassical model of trade, which relaxes the assumptions of the HO model to allow for cross-country differences in technology. Both differences in factor endowments and differences in technology that are non-neutral across industries are found to be important in explaining cross-country variation in production structure. Other research finds evidence consistent with multiple cones of diversification within the HO model, where countries or regions specialize in a distinct set of goods, and as a result have different relative factor prices (Schott, 2003; Bernard, Redding and Schott, 2005).

We now turn to the relationship between international trade and factor prices, an issue which rose to prominence with the debate about whether the rise in wage inequality in OECD countries since the 1970s is explained by international trade or skill-biased technological change. While the labour economics literature has tended to emphasize the role of skill-biased technological change, the

international trade literature has produced mixed findings, as illustrated by the collection of studies in Feenstra (2000). One approach has examined the net factor content of trade and has typically found a relatively minor role for international trade (see, for example, Krugman, 2000). Another approach has examined the relationship between relative goods and relative factor prices within the many-good, many-factor version of HO model (the generalization of the Stolper–Samuelson theorem). Here the results have been more sanguine about the contribution of international trade. Leamer (1998) shows how zero-profit conditions and the shares of factors in unit costs for a cross-section of industries can be used to estimate the changes in factor prices mandated by observed changes in goods prices. Assumptions are made about the degree to which improvements in technology are passed through into lower goods prices, and some evidence is found that trade-induced changes in goods prices during the 1970s pushed towards increasing wage inequality in the United States. Feenstra and Hanson (1999) extend the analysis to estimate the contribution of measures of technological change and outsourcing to changes in relative goods and hence through the zero-profit conditions to relative factor prices. In their baseline specification, they estimate that computers explain around 35 per cent of the rise in the relative wages of US non-production workers over the period 1979 to 1990, and outsourcing explains around 15 per cent. One important difference between international trade and other fields, such as development economics, is that general equilibrium is central to many of the field's theoretical predictions. As it result, it has proved hard to find natural experiments that provide plausible sources of exogenous variation to identify relationships of interest. Relatedly, many of the predictions of traditional trade theory with many goods and many factors relate to movements from autarky to international trade, but autarky is rarely observed. In two creative papers, Bernhofen and Brown (2004; 2005) exploit the dramatic opening of the Japanese economy in the 19th century from a state of near-complete isolation to test some of the most fundamental predictions of general equilibrium trade theory. In their first paper, they find evidence, supporting the general law of comparative advantage, that an economy's net export vector evaluated at autarky prices is negative. In their second paper, they estimate that, during the final years of Japan's isolation during 1851–3, real income would have had to increase by around eight or nine per cent in order to afford the consumption bundle that the economy could have obtained if it were engaged in international trade during that period.

New trade empirics

Although traditional trade theory emphasizes the international exchange of one set of goods for another (inter-industry trade) due to comparative advantage (dissimilar countries), much of international trade involves the two-way exchange of goods within industries (intra-industry trade) between developed nations (similar countries). This apparent disconnect between theory and data was documented in a number of early empirical studies, which examined the extent of intra-industry trade (for example, Grubel and Lloyd, 1975) and the volume of trade between similar countries (for example, Linder, 1961). This empirical evidence was a key motivation for the ‘new trade theory’ literature following Krugman (1979; 1980) that explained these features of international trade in terms of consumer love of variety and increasing returns to scale. Firms manufacture differentiated products and concentrate production in a single location, while consumers spread their expenditure across all firms’ varieties, giving rise to two-way trade even if countries are identical. Although not the only explanation for intra-industry trade

between similar countries (see Davis, 1997), the combination of consumer love of variety and increasing returns to scale provided an entirely new intellectual framework for thinking about the causes and consequences of international trade.

In the HO model, the volume of trade is increasing in the extent of dissimilarity in countries' factor endowments, whereas in new trade theory the volume of trade is increasing in the similarity of countries' sizes. Indeed, new trade theory provides rigorous theoretical foundations for the so-called 'gravity equation', in which the volume of trade between two countries is proportional to the product of their sizes and measures of extent of trade frictions. Although the gravity equation had been known for some time to provide an extremely successful empirical explanation for bilateral patterns of international trade (classic early treatments include Tinbergen, 1962, and Linnemann, 1966), it initially suffered from a lack of theoretical foundations.

New trade theory's prediction that the volume of trade should be proportional to the similarity of country sizes was examined empirically by Helpman (1987) in specifications derived directly from the theory. Using data from 14 OECD countries over the period 1956 to 1981, both bilateral trade and the share of inter-group trade in total trade were found to be strongly increasing in the similarity of country sizes. While this appeared to strongly confirm the predictions of new trade theory, Hummels and Levinsohn (1995) found that the same patterns existed for trade between non-OECD countries, for which new trade theory's assumptions of differentiated products and identical and homothetic preferences appeared less appropriate. One explanation of why the gravity equation appears to work for such diverse groups of countries is that a number of alternative theoretical frameworks, including the HO model, yield this relationship. As argued by Deardorff (1998), the gravity relationship is a basic implication of specialization combined with identical and homothetic preferences. Therefore, the problem is not a lack but rather a surfeit of theoretical foundations. Consistent with this insight, Evenett and Keller (2002) found that increasing returns and factor endowments both played a role in explaining the empirical success of the gravity equation for a diverse cross-section of developed and developing countries.

The gravity equation has been widely used in empirical work to estimate the impact on trade of a host of frictions, policies and institutions including national borders, transport costs, tariffs, common currencies and the World Trade Organization (WTO). A notable example is McCallum (1995), who finds that trade between Canadian provinces was more than 20 times larger than trade between Canadian provinces and US states, suggesting a surprisingly large impact of national borders on trade. Anderson and Van Wincoop (2002) show, however, that theoretical derivations of the gravity equation imply that bilateral trade depends not only on trade costs between regions themselves ('bilateral resistance') but also on trade costs with all locations ('multilateral resistance'). An implication is that national borders have a larger impact on inter-regional trade than on international trade the smaller a country is and the larger its trade partner. When countries are small, international trade is a large share of overall economic activity. Therefore, the national border has a large effect on multilateral resistance, and so leads to a large reduction in the cost of inter-regional trade relative to international trade. Estimating the gravity equation in a theory-consistent way, Anderson and Van Wincoop (2002) obtain much smaller, though still large, estimates of the trade impact of the Canada–US border.

In the presence of trade costs, an important difference emerges between the predictions of new trade theory and those of traditional trade theory. The combination of consumer love of variety, increasing returns to scale and trade costs in new trade theory generates a 'home market effect', whereby an

increase in expenditure leads to a more than proportionate increase in domestic production of a good. The intuition is that increasing returns to scale imply that firms have an incentive to concentrate production, while transport costs imply that they have an incentive to concentrate production close to large markets. In contrast, traditional trade models imply that an increase in expenditure leads at most to a proportionate increase in domestic production if foreign export supply is perfectly inelastic. Otherwise, if the foreign export supply curve is upward sloping, some of the increase in expenditure is satisfied through higher foreign exports and the increase in domestic production is less than proportionate. Using international and Japanese regional data, Davis and Weinstein (1999; 2003) find evidence of home market effects for a number of manufacturing industries, which together account for a substantial share of overall manufacturing activity. Additional evidence in support of home market effects emerges from international trade data in Feenstra, Markusen and Rose (2001) and Hanson and Xiang (2004).

One feature of international trade that appears at first sight hard to reconcile with new trade theory is the large number of zeros between country pairs. The constant elasticity of substitution (CES) preferences and iceberg trade costs in new trade models imply that all country pairs trade a positive quantity of each variety. However, in an analysis of bilateral trade between 161 countries over the period 1970 to 1997, Helpman, Melitz and Rubinstein (2006) find that roughly one half of the country-partner-year observations involve zero trade. A natural explanation for zero bilateral trade flows can be created within new trade theory if firm heterogeneity and fixed trade costs are introduced following Melitz (2003). Depending on the distribution of productivity within countries, firms may or may not find it profitable to incur the fixed costs of exporting to a particular market. Helpman, Melitz and Rubinstein (2006) develop a methodology for estimating the gravity equation that not only controls for multilateral resistance as suggested above, but also controls for the existence of zero bilateral flows and the non-random selection of firms into exporting according to their productivity.

Finally, one key stylized fact about international trade since the Second World War is that it has grown far more rapidly than income. Two potential explanations are reductions in trade barriers following multilateral liberalization or regional integration, and improvements in transportation and communication technologies. Yi (2003) argues that is hard to explain the magnitude of the trade growth using standard trade models, and observed declines in trade barriers unless one assumes implausibly high elasticities of substitution. However, augmenting standard models to include intermediate inputs enables the growth in trade to be explained with a smaller elasticity of substitution. In the augmented model, tariff reductions decrease the cost of shipping both intermediate inputs and final goods, and so have a magnified impact on overall trade volumes. Indeed, the geographical separation of stages of the production process is one of the distinctive features of trade at the end of the 20th century compared with an earlier era of international integration at the end of the 19th century. This geographical separation of stages of production has been variously referred to as vertical specialization, vertical disintegration, the fragmentation of production, the slicing of the value-added chain, geographical production networks and offshoring. Hummels, Ishii and Yi (2001) define vertical specialization as occurring when the following conditions are satisfied: (*a*) goods are produced in multiple sequential stages; (*b*) two or more countries provide value-added in the good's production sequence; (*c*) at least one country uses imported inputs in its stage of the production process and some of the resulting output is exported. The authors provide empirical evidence of the rapid growth in vertical specialization in the closing decades of the 20th century alongside the rapid growth in overall trade.

The empirics of product trade

The dissemination of highly disaggregated data-sets on trade in thousands of individual products (Feenstra, Romalis and Schott, 2002; Feenstra et al., 2005) has contributed towards a shift in focus in empirical trade research towards the micro level. For the United States, data are available for over 7,000 seven-digit products of the Tariff Schedule of the United States (TS7) from 1972 to 1988 and for over 10,000 ten-digit products of the Harmonized System (HS10) from 1989 onwards.

In contrast to the empirical research on the HO model discussed above, which emphasizes specialization across products or industries, Schott (2004) provides compelling evidence of specialization *within* products. With US manufacturing imports taken as a whole in 1994, and the unit value ratio (UVR) defined as the ratio of value to quantity, the maximum UVR within products across trade partners is a factor of 24 times greater than the minimum UVR. The UVRs are higher for varieties originating in capital- and skill-abundant countries than for those sourced from labour-abundant countries, consistent with HO-based specialization. Similarly, UVRs are positively associated with the capital intensity of the production techniques that exporters use to produce them. Taken together, these and other findings in the paper suggest that comparative advantage operates at a much finer level of detail than customarily considered.

Another insight that emerges from the product-level trade data is the importance of the ‘extensive margin’ of the set of goods traded. Hummels and Klenow (2005) decompose variation in countries’ aggregate exports into the contributions of the following terms: (*a*) the quantity of each good exported (the ‘intensive margin’); (*b*) the set of goods exported (the ‘extensive margin’); (*c*) the quality of goods exported. They find that the extensive margin accounts for around 60 per cent of the greater exports of larger economies, while the remaining intensive margin contribution of 40 per cent consists of higher quantities being exported at modestly higher prices. Kehoe and Ruhl (2004) establish an important role for the extensive margin in explaining the growth of trade following trade liberalizations. The set of goods that accounted for only 10 per cent of trade prior to liberalization are found to account for as much as 40 per cent of trade after liberalization. Using micro data from the U.S. Commodity Flow Survey, Hillberry and Hummels (2005) show that trade frictions such as distance reduce the aggregate value of trade primarily through the number of commodities shipped and the number of establishments shipping commodities (the extensive margin) rather than through the average value of shipments (the intensive margin). Together these findings present a number of challenges to standard trade models. For example, in marked contrast to the data, new trade theory models without firm heterogeneity and fixed costs of trading imply that all of the adjustment to trade frictions occurs through the intensive margin.

Although consumer love of variety is one of the defining features of new trade theory, Broda and Weinstein (2006) were the first to estimate the welfare gains from an increase in the number of varieties imported over time. In their analysis, the product-level trade data is used to measure varieties, defined as the versions of a product supplied by different exporters. The methodology of Feenstra (1994) is extended to estimate separate elasticities of substitution for thousands of products and to evaluate the contribution of new varieties to the US import price index. According to their baseline estimates, conventional price indices that do not correctly control for variety growth overstate the growth in US import prices by around 1.2 percentage points per annum. The estimated contribution to US welfare

from an increase in the number of varieties imported over the period 1972 to 2001 is around 2.6 per cent of national income.

The empirics of plant and firm trade

The analysis of micro data-sets on plants and firms presents additional empirical challenges to traditional and new trade theory, and prompted a wave of subsequent theoretical research. The first set of empirical challenges relates to producer heterogeneity and persistent reallocation. Whereas traditional and new trade theories typically assume a representative firm, micro data-sets reveal vast heterogeneity across plants and firms within narrowly defined industries, in terms of productivity, capital intensity, skill intensity and other characteristics (see for example the survey by Bartelsman and Doms, 2000).

Similarly, whereas traditional trade theory emphasizes net reallocations of resources between industries in response to exogenous shocks such as trade liberalization, micro data-sets reveal persistent job creation and job destruction in all industries even in the apparent absence of exogenous shocks. Additionally, job creation and job destruction are positively correlated across industries, implying that rates of gross job creation and destruction are large relative to the net reallocation emphasized in traditional trade theory. An implication of these findings is that the changes in employment across plants and firms are greater than those required to achieve the observed between-industry reallocation of resources (‘excess job reallocation’), implying substantial within-industry reallocations of resources (see in particular Davis, Haltiwanger and Schuh, 1998).

The second set of empirical challenges relates to the export behaviour of plants and firms. Traditional trade theory predicts net exports in one set of industries and net imports in another set of industries. New trade theory implies that all firms export as a result of consumer love of variety and increasing returns to scale. Yet, in micro data-sets, all manufacturing industries display a mix of exporters and non-exporters (see Bernard and Jensen, 1995). Moreover, exporters are systematically more productive, more capital-intensive and more skill-intensive than non-exporters (see again Bernard and Jensen, 1995). These findings have led to considerable debate as to whether high-performing firms become exporters or whether exporting leads to improved firm performance. The current consensus favours causality running from good firm performance to exporting (selection into export markets): see, for example, Bernard and Jensen (1999), Clerides, Lach and Tybout (1998) and Roberts and Tybout (1997).

The third set of empirical challenges relates to evidence from trade liberalizations in both developed and developing countries. Despite traditional trade theory's emphasis on between-industry reallocations of resources, one of the central findings from empirical studies of trade liberalizations is the importance of within-industry reallocations of resources across plants and firms. In an influential paper, Pavcnik (2002) finds that between-plant reallocations of resources account for around two-thirds (12.4 percentage points) of the 19 per cent increase in aggregate productivity in the Chilean manufacturing sector following the trade liberalization of the late 1970s and early 1980s. Similarly, Trefler (2004) finds an important role for reallocation in accounting for the improvement in aggregate productivity in Canadian manufacturing in the aftermath of the Canada–US free trade agreement.

Together these empirical challenges have led to the development of new theoretical frameworks incorporating firm heterogeneity into both traditional and new trade theory (see in particular Bernard, Eaton and Kortum, 2003; Melitz, 2003). The interplay between the econometric analysis of micro data-

sets on plants and firms and the theoretical analysis of firm-based responses to international trade is one of the most exciting areas of ongoing research.

The empirics of trade policy

A final area of rapid recent progress is the empirical analysis of trade policy. A number of alternative approaches to modelling the political economy of trade policy have been taken, including median-voter theories, models where the government trades off political support from industry against consumer dissatisfaction, theories of lobbying by special interest groups, and models of electoral contribution. One of the most influential lines of research follows the seminal work of Grossman and Helpman (1994). In their model, campaign contributions are designed to influence policy choices. Interest groups move first and offer politicians campaign contributions that depend on their policy stance. Politicians next maximize a political objective function which depends on both campaign contributions and social welfare. The political objective function is derived from microeconomic foundations within a model of electoral competition. The model yields a structural equation in which the level of protection depends on the political organization of the industry, the ratio of domestic output in the industry to net trade, and the elasticity of import demand or export supply. Goldberg and Maggi (1999) estimate the structural relationship implied by the Grossman and Helpman model and find broad empirical support. In particular, the pattern of protection differs markedly between politically organized and non-organized industries, though the implied weight on social welfare relative to political contributions is larger than expected. One of the distinctive features of recent empirical work in this area is again the rigorous derivation of empirical specifications from theoretical predictions. Gawande and Krishna (2003) survey both the recent empirical evidence and the results of earlier and more ad hoc empirical specifications. A related theoretical literature has sought to model the politics of international trade agreements (for example, Grossman and Helpman, 1995; Krishna, 1998; McLaren, 2002). One issue that has attracted particular attention is the extent to which regional preferential trade agreements reinforce or retard multilateral trade liberalization. Theoretical research has also examined the microeconomic foundations for observed features of international trade institutions such as the General Agreement on Tariffs and Trade (GATT) and the World Trade Organization (WTO) (see in particular Bagwell and Staiger, 1999; 2001). Two key features are reciprocity and non-discrimination (the Most Favored Nation, MFN, principle). Empirical work in this area remains in its infancy and offers an exciting prospect for the future. In an analysis of US trade policy, Limao (2006) finds evidence that preferential trading blocs have acted as stumbling blocks for multilateral liberalization.

See Also

- factor content of trade
- Heckscher–Ohlin trade theory
- international trade and heterogeneous firms
- international trade theory
- Ricardian trade theory

- trade policy, political economy of

Bibliography

Anderson, J. and van Wincoop, E. 2002. Gravity with gravitas: a solution to the border puzzle. *American Economic Review* 93, 170–92.

Bagwell, K. and Staiger, R. 1999. An economic theory of GATT. *American Economic Review* 89, 215–48.

Bagwell, K. and Staiger, R. 2001. Domestic policies, national sovereignty and international economic institutions. *Quarterly Journal of Economics* 116, 519–62.

Bartelsman, E. and Doms, M. 2000. Understanding productivity: lessons from longitudinal microdata. *Journal of Economic Literature* 38, 569–94.

Bernard, A.B., Eaton, J. and Kortum, S.S. 2003. Plants and productivity in international trade. *American Economic Review* 93, 1268–90.

Bernard, A.B. and Jensen, J.B. 1995. Exporters, jobs, and wages in US manufacturing: 1976–87. *Brookings Papers on Economic Activity: Microeconomics* 67–112.

Bernard, A.B. and Jensen, J.B. 1999. Exceptional exporter performance: cause, effect, or both? *Journal of International Economics* 47, 1–25.

Bernard, A.B., Redding, S.J. and Schott, P.K. 2005. Factor price equality and the economies of the United States. Discussion Paper No. 5111. London: CEPR.

Bernhofen, D. and Brown, J. 2004. A direct test of the theory of comparative advantage: the case of Japan. *Journal of Political Economy* 112, 48–67.

Bernhofen, D. and Brown, J. 2005. An empirical assessment of the comparative advantage gains from trade: evidence from Japan. *American Economic Review* 95, 208–25.

Bowen, H., Leamer, E. and Sveikauskas, L. 1987. Multicountry, multifactor tests of the factor abundance theory. *American Economic Review* 77, 791–809.

Broda, C. and Weinstein, D.E. 2006. Globalization and the gains from variety. *Quarterly Journal of Economics* 121, 541–86.

Clerides, S., Lach, S. and Tybout, J. 1998. Is learning by exporting important? Micro-dynamic evidence

from Columbia, Mexico and Morocco. *Quarterly Journal of Economics* 113, 903–47.

Davis, D. 1997. Critical evidence on comparative advantage? North–north trade in a multilateral world. *Journal of Political Economy* 105, 1051–60.

Davis, S., Haltiwanger, J. and Schuh, S. 1998. *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Davis, D. and Weinstein, D. 1999. Economic geography and regional production structure: an empirical investigation. *European Economic Review* 43, 379–407.

Davis, D. and Weinstein, D. 2001. An account of global factor trade. *American Economic Review* 91, 1423–53.

Davis, D. and Weinstein, D. 2003. Market access, economic geography, and comparative advantage: an empirical assessment. *Journal of International Economics* 59, 1–23.

Davis, D., Weinstein, D., Bradford, S. and Shimpo, K. 1997. Using international and Japanese regional data to determine when the factor abundance theory of trade works. *American Economic Review* 87, 421–46.

Deardorff, A.V. 1998. Determinants of bilateral trade: does gravity work in a neoclassical world? In *The Regionalization of the World Economy*, ed. J. Frankel. Chicago: NBER and Chicago University Press.

Evenett, S. and Keller, W. 2002. On theories explaining the success of the gravity equation. *Journal of Political Economy* 110, 281–316.

Feenstra, R.C. 1994. New product varieties and the measurement of international prices. *American Economic Review* 84, 157–77.

Feenstra, R. 2000. *The Impact of International Trade on Wages*. Chicago: NBER and University of Chicago Press.

Feenstra, R. and Hanson, G. 1999. The impact of outsourcing and high-technology capital on wages: estimates for the United States, 1979–80. *Quarterly Journal of Economics* 114, 907–40.

Feenstra, R., Markusen, J. and Rose, A. 2001. Understanding the home market effect and the gravity equation. *Canadian Journal of Economics* 34, 430–47.

Feenstra, R.C., Romalis, J. and Schott, P.K. 2002. U.S. imports, exports, and tariff data, 1989–2001. Working Paper No. 9387. Cambridge, MA: NBER.

Feenstra, R.C., Lipsey, R.E., Deng, H., Ma, A.C. and Mo, H. 2005. World trade flows: 1962–2000. Working Paper No. 11040. Cambridge, MA: NBER.

Gandal, N., Hanson, G. and Slaughter, M. 2004. Technology, trade, and adjustment to immigration in Israel. *European Economic Review* 48, 403–28.

Gawande, K. and Krishna, P. 2003. The political economy of trade policy: empirical approaches. In *Handbook of International Trade*, ed. E.K. Choi and J. Harrigan. Oxford: Basil Blackwell.

Goldberg, P. and Maggi, G. 1999. Protection for sale: an empirical investigation. *American Economic Review* 89, 1135–55.

Grossman, G.M. and Helpman, E. 1994. Protection for sale. *American Economic Review* 84, 833–50.

Grossman, G.M. and Helpman, E. 1995. The politics of free trade agreements. *American Economic Review* 85, 667–90.

Grubel, H. and Lloyd, P. 1975. *Intra-industry Trade: The Theory and Measurement of International Trade in Differentiated Products*. London: Macmillan.

Hanson, G. and Slaughter, M. 2002. Labor market adjustment in open economies: evidence from U.S. states. *Journal of International Economics* 57, 3–29.

Hanson, G. and Xiang, C. 2004. The home market effect and bilateral trade patterns. *American Economic Review* 94, 1109–29.

Harrigan, J. 1995. Factor endowments and the international location of production: econometric evidence for the OECD, 1970–85. *Journal of International Economics* 39, 123–41.

Harrigan, J. 1997. Technology, factor supplies, and international specialisation: estimating the neoclassical model. *American Economic Review* 87, 475–94.

Heckscher, E.F. 1919. The effect of foreign trade on the distribution of income. *Economisk Tidsckrift*. In *Heckscher–Ohlin Trade Theory*, ed. E.F. Heckscher and B. Ohlin. Cambridge, MA: MIT Press, 1991.

Helpman, E. 1987. Imperfect competition and international trade: evidence from fourteen industrial countries. *Journal of the Japanese and International Economies* 1, 62–81.

Helpman, E., Melitz, M. and Rubinstein, Y. 2006. Trading partners and trading volumes. Mimeo, Harvard University.

Hillberry, R. and Hummels, D. 2005. Trade responses to geographic frictions: a decomposition using micro-data. Working Paper No. 11339. Cambridge, MA: NBER.

Hummels, D., Ishii, J. and Yi, K.-M. 2001. The nature and growth of vertical specialization in world trade. *Journal of International Economics* 54, 75–96.

Hummels, D. and Klenow, P. 2005. The variety and quality of a nation's exports. *American Economic Review* 95, 704–23.

Hummels, D. and Levinsohn, J. 1995. Monopolistic competition and international trade: reconsidering the evidence. *Quarterly Journal of Economics* 110, 799–836.

Kehoe, T. and Ruhl, K. 2004. How important is the new goods margin in international trade? Working Paper, Federal Reserve Bank of Minneapolis.

Krishna, P. 1998. Regionalism and multilateralism: a political economy approach. *Quarterly Journal of Economics* 113, 227–51.

Krugman, P.R. 1979. Increasing returns, monopolistic competition, and international trade. *Journal of International Economics* 9, 469–79.

Krugman, P.R. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70, 950–9.

Krugman, P.R. 2000. Technology, trade and factor prices. *Journal of International Economics* 50, 51–72.

Leamer, E. 1980. The Leontief paradox, reconsidered. *Journal of Political Economy* 88, 495–503.

Leamer, E. 1984. *Sources of International Comparative Advantage*. Cambridge, MA: MIT Press.

Leamer, E. 1998. In search of Stolper–Samuelson linkages between international trade and lower wages. In *Imports, Exports, and the American Worker*, ed. S. Collins. Washington, DC: Brookings Institution Press.

Leontief, W. 1953. Domestic production and foreign trade: the American capital position re-examined. *Proceedings of the American Philosophical Society* 97, 332–49.

Limao, N. 2006. Preferential trade agreements as stumbling blocks for multilateral trade liberalization: evidence for the United States. *American Economic Review* 96, 896–914.

Linder, S.B. 1961. *An Essay on Trade and Transformation*. New York: Wiley.

Linnemann, H. 1966. *An Econometric Study of International Trade Flows*. Amsterdam: North-Holland.

McCallum, J. 1995. National borders matter: Canada–US regional trade patterns. *American Economic Review* 85, 615–23.

MacDougall, G.D.A. 1951. British and American exports: a study suggested by the theory of comparative costs. *Economic Journal* 61, 697–724.

McLaren, J. 2002. A theory of insidious regionalism. *Quarterly Journal of Economics* 117, 571–608.

Melitz, M.J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71, 1695–725.

Ohlin, B. 1924. The theory of trade. In *Heckscher–Ohlin Trade Theory*, ed. E.F. Heckscher and B. Ohlin. Cambridge, MA: MIT Press, 1991.

Pavcnik, N. 2002. Trade liberalization, exit, and productivity improvement: evidence from Chilean plants. *Review of Economic Studies* 69, 245–76.

Ricardo, D. 1817. *The Principles of Political Economy and Taxation*. London: John Murray.

Roberts, M.J. and Tybout, J. 1997. The decision to export in Colombia: an empirical model of entry with sunk costs. *American Economic Review* 87, 545–64.

Schott, P.K. 2003. One size fits all? Heckscher–Ohlin specialization in global production. *American Economic Review* 93, 686–708.

Schott, P.K. 2004. Across-product versus within-product specialization in international trade. *Quarterly Journal of Economics* 119, 647–78.

Tinbergen, J. 1962. *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: Twentieth Century Fund.

Trefler, D. 1993. International factor price differences: Leontief was right! *Journal of Political Economy* 101, 961–87.

Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85, 1029–46.

Trefler, D. 2004. The long and short of the Canada–U.S. Free Trade Agreement. *American Economic Review* 94, 870–95.

Yi, K. 2003. Can vertical specialization explain the growth of world trade? *Journal of Political Economy* 111, 52–102.

How to cite this article

Redding, Stephen J. "international trade, empirical approaches to." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000304> doi:10.1057/9780230226203.0837

The New Palgrave Dictionary of Economics Online

Internet, economics of the

Nicholas Economides
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We discuss salient economic aspects of the Internet, including the possible abolition of net neutrality by local broadband access networks as well as potential incompatibilities and degradation of connectivity in the Internet backbone.

Keywords

antitrust; cartels; common standards; innovation; interconnection; Internet, economics of the; market power; net neutrality; network effects; network externalities; price discrimination; proprietary standards; switching costs; two-sided pricing; universal connectivity

Article

The Internet is a global network of interconnected networks that connect computers. The Internet allows data transfers as well as the provision of a variety of interactive real-time and time-delayed telecommunications services. Internet communication is based on common and public protocols. Hundreds of millions of computers are presently connected to the Internet. Figure 1 shows the expansion of the number of computers connected to the Internet.

Figure 1
Internet survey host count, 1993–2006. *Source*: Internet Systems Consortium. Online. Available at http://www.isc.org, accessed 29 January 2007.

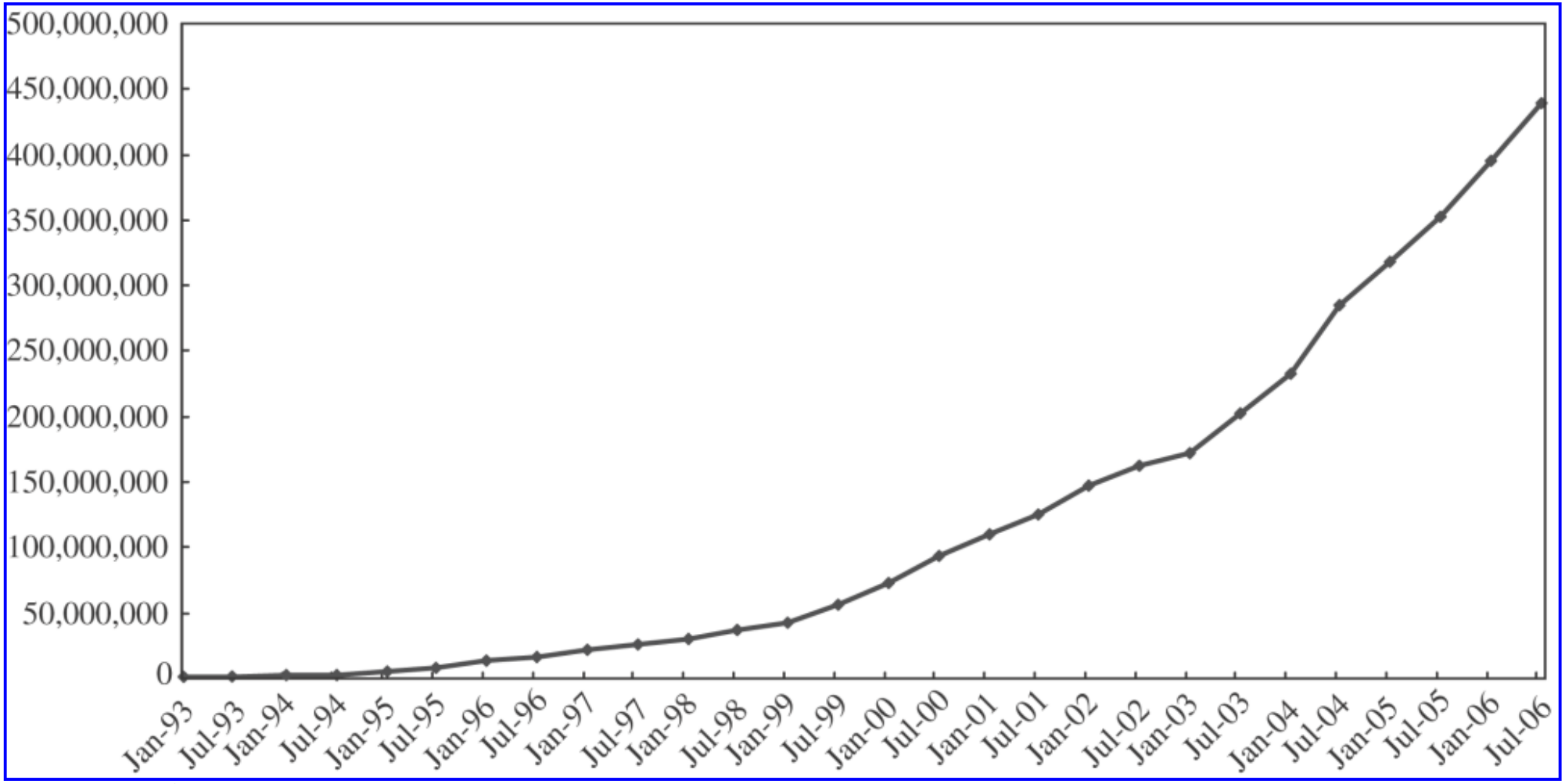

The vast majority of computers owned by individuals or businesses connect to the Internet through commercial ‘Internet service providers (ISPs). Educational institutions and government departments are also connected to the Internet but typically do not offer commercial ISP services. Users connect to the Internet either by dialing their ISP, connecting through cable modems, or residential ‘digital subscriber line’ (DSL), or through corporate networks. Typically, routers and switches owned by the ISP send the caller's packets to a local ‘point of presence’ (POP) of the Internet. Dial-up, cable modem, and DSL access POPs as well as corporate networks dedicated access circuits connect to high-speed hubs. High-speed circuits, leased from or owned by telephone companies, connect the high speed hubs forming an ‘Internet backbone network’.
The Internet is based on three basic separate levels of functions of the network:

- the hardware/electronics level of the physical network;
- the (logical) network level where basic communication and interoperability is established; and
- the applications/services level.

Thus, the Internet separates the network interoperability level from the applications/services level. Unlike earlier centralized digital electronic communications networks, such as CompuServe, AT&T Mail, Prodigy, and early America On Line (AOL), the Internet allows a large variety of applications and services to be run ‘at the edge’ of the network and not centrally.

Residential broadband access networks and net neutrality

Users pay ISPs for access to the whole Internet. Similarly, ISPs pay backbones for access to the whole Internet. ISPs pay per month for a pipe of a certain bandwidth, presumably according to their expected use. When digital content, for example, is downloaded by consumer A from provider B, both sides, that is, both A and B, pay. Consumer A pays to his ISP through his monthly subscription, and provider B pays similarly. In turn, ISPs pay to their respective backbones through their monthly subscriptions. The present regime on the Internet does not distinguish in terms of price (or in any other way) between bits or information packets depending on the services that these bits and packets are used for. This regime, called 'net neutrality', has prevailed on the Internet since its inception. Presently, a bit or information packet used for 'voice over Internet protocol' (VOIP), for search, email, for an image or for a video is priced equally as a part of the large number of packets that correspond to the subscription services of the originating and terminating ISP.

Taking advantage of a change in regulatory rules by the Federal Communications Commission that reclassified the Internet as an 'information service' rather than a 'telecommunications service', AT&T, Verizon and cable TV networks advocate price discrimination based on which application and on which provider the bits they transport come from. These local broadband access networks would like to abolish the regime of non-discrimination which has been called 'net neutrality' and substitute for it a complex price discrimination schedule where, besides the basic service for transmission of bits, there will be additional charges by the Internet access network levied to the originating party (such as Google, Yahoo or Microsoft Network, MSN) even when the application provider is not directly connected to the local access network.

The imposition of price discrimination on the provider side of the market and not on the subscriber is a version of two-sided pricing. It is uniquely possible for firms operating within a network structure. Besides traditional networks, such two-sided pricing is also possible for intermediaries in exchange networks (such as the exchanges themselves). There is presently considerable debate on the legality as well as the efficiency properties of the implementation of such complex pricing strategies by broadband Internet access networks, mainly because of the very considerable market power of such firms.

Residential retail broadband Internet access customers may well have difficulty changing ISPs. Ninety-nine per cent of US households are offered Internet access by at most two firms – a telephone company through DSL and a cable TV company through a cable 'modem' – and many households are facing a monopoly of either cable or DSL. There are also switching costs to residential customers, such as changing equipment. Finally, residential customers are much more affected by contracts that bundle broadband Internet access with other services such as telecommunications and cable television.

As discussed earlier, the Internet under net neutrality separated the network layer from the applications/services layer. This allowed firms to innovate 'at the edge of the network' without seeking approval from network operator(s). The decentralization of the Internet based on net neutrality facilitated innovation resulting in big successes such as Google, MSN, Yahoo, and Skype. Net neutrality also increased competition among the applications and services 'at the edge of the network' which did not need to own a network to compete. Additionally, the existence of network effects on the Internet implies that efficient prices to users on both sides (consumers and applications) should be lower than in a market without network effects. Instead we see an attempt to increase prices that will reduce network effects and innovation.

Abolition of net neutrality raises both horizontal and vertical antitrust issues. To start with horizontal issues, last-mile carriers (who are selling as a duopoly or monopoly to residential consumers) may reduce capacity of 'plain' broadband Internet access service and/or degrade it so that they can establish a 'premium' service for which they intend to charge content/applications providers whose content or application is used by residential subscribers. Coordinated reduction of capacity in 'plain' service is reminiscent of cartel behaviour. In general, the coordinated introduction of price discrimination schemes may reduce output, which would reduce total surplus. Therefore, introduction of coordinated price discrimination may have anti-competitive consequences.

There is also a variety of potentially anti-competitive vertical effects. For example, a carrier may favour its own content or application over the content of a competing carrier or a company that does not have its own network. VOIP provided over broadband Internet competes with traditional circuit-switched service provided by AT&T and Verizon, and could be subject to discrimination. Additionally, both AT&T and Verizon are gearing to distribute video, and could favour their video services over those of others. But the anti-competitive concerns are hardly limited to products and services currently provided by the firms with market power in the access market. The carriers can also leverage market power in broadband access to the content or applications markets through contractual relationships. For example, a carrier can contract with an Internet search engine to put it in 'premium' service while searches using other search engines face considerable delays using 'plain' service. The question that confronts the US Congress in 2007 is whether it should intervene by imposing non-discrimination restrictions or wait instead for antitrust suits to filed and resolved. The crucial role of the Internet in US economic growth argues in favour of pre-emptive restrictions.

Backbone issues

Backbone networks provide transport and routing services for information packets among high-speed hubs on the Internet. Backbone networks vary in terms of their geographic coverage. There is wide variance of ISPs in terms of their subscriber size and the networks they own. However, irrespective of its size, an ISP needs to interconnect with other ISPs so that its customers can reach all computers/nodes on the Internet. That is, interconnection is necessary to provide the universal connectivity on the Internet which is demanded by users. Internet networks interconnect in two ways: (*a*) private bilateral interconnection, and (*b*) interconnection at public network access points (NAPs). Private interconnection points and public NAPs are facilities that provide collocation space and a switching platform so that networks are able to interconnect. Interconnection services are complementary to Internet transport. In a sense, the Internet backbone networks are like freeways and the NAPs are like the freeway interchanges.

Internet networks have contracts that govern the terms under which they pay each other for connectivity. Payment takes two distinct forms: (*a*) payment in dollars for ‘transit’, and (*b*) payment in kind, that is, barter, called ‘peering’. Connectivity arrangements among ISPs encompass a seamless continuum, including ISPs that rely exclusively on transit to achieve connectivity, ISPs that use only peering to achieve connectivity, and everything in between. Although there are differences between transit and peering in the specifics of the payments method, and transit includes services to the ISP not provided by peering, these two are essentially alternative payment methods for connectivity. The transport and routing that backbone networks offer do not necessarily differ depending on whether cash (transit) or barter (peering) is used for payment.

Under transit, a network X connects to network Y with a pipeline of a certain size, and pays network Y for allowing X to reach all Internet destinations. Under transit, network X pays Y to reach not only Y and its peers, *but also any other network*, such as network Z by passing through Y, as in the diagram below.

```
X-----------Y- = - = - = - = - = -Z
```

Under peering, two interconnecting networks agree not to pay each other for carrying the traffic exchanged between them as long as the traffic originates and terminates in the two networks. In the diagram above, if X and Y have a peering agreement, they exchange traffic without paying each other so long as such traffic terminating on X originates in Y, and traffic terminating on Y originates in X. If Y were to pass to X traffic originating from a network Z that was not a customer of Y, Y would have to pay a transit fee to X (or get paid a transit fee by X), that is, it would not be covered by the peering agreement between X and Y.

Although the networks do not exchange money in a peering arrangement, the price of the traffic exchange is not zero. If two networks X and Y enter into a peering agreement, it means that they agree that the cost of transporting traffic from X to Y and vice versa that is incurred within X is roughly the same as the cost of transporting traffic incurred within Y. These two costs have to be roughly equal if the networks peer, but they are not zero.

It is a commercial decision whether interconnection takes the form of peering or transit payment. Peering is preferred when the cost incurred by X for traffic from X to Y and Y to X is roughly the same as the cost incurred by Y for the same traffic. If not, the networks will use transit. As I explain below, the decision on whether to peer depends crucially on the geographic coverage of the candidate networks.

Generally, peering does not imply that the two networks should have the same size in terms of the numbers of ISPs connected to each network, or in terms of the traffic that the two networks generate. If two networks, X and Y, are similar in terms of the types of users to whom they sell services, the amount of traffic flowing across their interconnection point(s) will be roughly the same, irrespective of the relative size of the networks. For example, suppose that network X has ten ISPs and network Y has one ISP. If all ISPs have similar features, the traffic flow from X to Y is generally equal to the traffic flow from Y to X.

What determines whether a peering arrangement is efficient for both networks is the *cost* of carrying the mutual traffic within each network. This cost will depend crucially on a number of factors, including the geographic coverage of the two networks. Even if the types of ISPs of the two networks are the same as in the previous example (and therefore the traffic flowing in each direction is the same), the cost of carrying the traffic can be quite different in network X from network Y. For example, network X (with the ten ISPs) may cover a larger geographic area and have significantly higher costs per unit of traffic than network Y. Then network X would not agree to peer with Y. These differences in costs ultimately would determine the decision to peer (barter) or receive a cash payment for transport.

Where higher costs are incurred by one of two interconnecting networks because of differences in the geographic coverage of the networks, peering would be undesirable from the perspective of the larger network. Similarly, one expects that networks that cover small geographic areas will peer only with each other. Under these assumptions, who peers with whom is a consequence of the extent of a network's geographic coverage, and may not have any particular strategic connotation. In a theoretical model, Milgrom, Mitchell and Srinagesh (2000) show how peering can emerge under some circumstances as an equilibrium in a bargaining model between backbones.

Structural conditions for Internet backbone services (ease of expansion and entry) ensure low barriers to entry and expansion, and easy conversion of other transport capacity to Internet backbone capacity. As discussed later, raw transport capacity as well as Internet transport capacity have grown dramatically. Transport capacity is almost a commodity because of its abundance. The business environment for Internet backbone services is competitive. Generally, ISPs buying transport services face flexible transit contracts of relatively short duration. This is reflected in competitive pricing. Economides (2006a) shows that AT&T and MCI had almost identical prices for transit in 1999 when AT&T's backbone business was significantly smaller than MCI's.

ISPs are not locked in by switching costs of any significant magnitude. Thus, ISPs are in good position to change providers in response to any increase in price, and it would be very difficult for a backbone profitably to increase price. Moreover, a large percentage of ISPs has formal agreements that allow them to route packets through several backbone networks and are able to control the way the traffic will be routed (multi-homing).

When an ISP reaches the Internet through multiple backbones, it has additional flexibility in routing its traffic through any particular backbone. A multi-homing ISP can easily reduce or increase the capacity with which it connects to any particular backbone in response to changes in prices of transit. Thus, multi-homing increases the firm-specific elasticity of demand of a backbone provider. Therefore, multi-homing severely limits the ability of any backbone services provider to profitably increase the price of transport. Any backbone

increasing the price of transport will face a significant decrease in the capacity bought by multi-homing ISPs.

Large Internet customers also use multiple ISPs, which is called ‘customer multi-homing’. They have chosen to avoid any limitation on their ability to switch traffic among suppliers even in the very shortest of runs. Customer multi-homing has similar effects as ISP multi-homing in increasing the firm-specific elasticity of demand of a backbone provider and limiting the ability of any backbone services provider to profitably increase the price of transport.

Like any network, the Internet exhibits network effects. Network effects are present when the value of a good or service to each consumer rises as more consumers use it, everything else being equal – see Economides (1996), Farrell and Saloner (1985), Katz and Shapiro (1985), and Liebowitz and Margolis (1994). In traditional telecommunications networks, an additional customer to the network increases the value of a network connection to all other customers, since each of them can now make an extra call. On the Internet, an additional user potentially

- adds to the information that all others can reach;
- adds to the goods available for sale on the Internet;
- adds one more customer for e-commerce sellers;
- adds to the number of people who can send and receive e-mail or otherwise interact in through the Internet.

Thus, the addition of an extra computer node increases the value of an Internet connection to each connection.

In networks of interconnected networks, there are large social benefits from the interconnection of the networks and the use of common standards. A number of networks of various ownership structures have harnessed the power of network externalities by using common standards. Examples of interconnected networks of diverse ownership that use common standards include the telecommunications network, the network of fax machines, and the Internet. Despite the different ownership structures in these three networks, the adoption of common standards has allowed each of them to reap huge network-wide benefits.

As the variety and extent of the Internet's offerings expand, and as more customers and more sites join the Internet, the value of a connection to the Internet rises. Because of the high network externalities of the Internet, consumers on the Internet demand universal connectivity, that is, to be able to connect with every website on the Internet and to be able to send electronic mail to anyone. This implies that every network must connect with the rest of the Internet in order to be a part of it. The demand for universal connectivity on the Internet is stronger than the demand of a voice telecommunications customer to reach all customers everywhere in the world. In the case of voice, it may be possible but very unlikely that a customer might buy service from a long-distance company that does not include some remote country because the customer believes that it is very unlikely that he or she would be making calls to that country. On the Internet, however, one does not know where content is located. If company A did not allow its customers to reach region B or customers of a different company C, customers of A would never be able to know or anticipate what content they would be missing. Thus, consumers’ desire for Internet universal connectivity is stronger than for voice telecommunications. Additionally, because connectivity on the Internet is two-way, a customer of company A would be losing exposure of his or her content (and the ability to send and receive e-mails) to region B and customers of company C. It would be difficult for customer A to calculate the extent of the losses accrued to him or her from such actions of company A. Thus, again, customers on the Internet require universal connectivity.

In markets with network externalities, firms may create bottleneck power by using proprietary standards. A firm controlling a standard needed by new entrants to interconnect their networks with the network of the incumbent may be in a position to exercise market power (see Economides, 2006b). Often a new technology will enter the market with competing incompatible standards. Competition among standards may have the snowball characteristic attributed to network externalities.

Economics literature has established that using network externalities to affect market structure by creating a bottleneck requires three conditions (see Economides, 1996; 1989; Farrell and Saloner, 1985; Katz and Shapiro, 1985):

- networks use proprietary standards;
- no customer needs to reach nodes of or to buy services from more than one proprietary network; and
- customers are captives of the network to which they subscribe and cannot change providers easily and cheaply.

First, without proprietary standards, a firm does not have the opportunity to create the bottleneck. Second, if proprietary standards are possible, the development of proprietary standards by one network isolates its competitors from network benefits, which then accrue to only one network. The value of each proprietary network is diminished when customers need to buy services from more than one network. Third, the more consumers are captive and cannot easily and economically change providers, the more valuable is the installed base to any proprietary network. I show below that these conditions fail in the context of the Internet backbone.

For example, if universal connectivity were not offered by a backbone network, a customer or its ISP would have to connect with more than one backbone. This would be similar to the period 1895–1930 when a number of telephone companies run disconnected networks. Eventually most of the independent networks were bought by AT&T, which had a dominant long-distance network. The refusal of AT&T to deal and interconnect with independents was effective for three key reasons: (*a*) AT&T controlled the standards and

protocols under which its network ran; (*b*) long-distance service was provided exclusively by AT&T in most of the United States; and (*c*) the cost to a customer of connecting to both AT&T and an independent was high. None of these reasons applies to the Internet. The Internet is based on public protocols. No Internet backbone has exclusive network coverage of a large portion of the United States. Finally, connecting to more than one backbone (multi-homing) is a common practice by many ISPs and does not require big costs. And ISPs can interconnect with each other through secondary peering, as explained below. Thus, the economic factors that allowed AT&T to blackmail independents into submission in the first three decades of the 20th century are reversed in today's Internet backbone, and therefore would not support a profitable refusal to interconnect by any backbone.

The Internet fails to fulfill any of the three necessary conditions stated above under which a network may be able to leverage network externalities and create a bottleneck. First, there are no proprietary standards on the Internet, so the first condition fails. The scenario of standards wars is not at all applicable to Internet transport, where full compatibility, interconnection and inter-operability prevail. For Internet transport, there are no proprietary standards. There is no control of any technical standard by service providers and none is in prospect. Internet transport standards are firmly public property (Kahn and Cerf, 1999; Bradner, 1999). As a result, any seller can create a network complying with the Internet standards – thereby expanding the network of interconnected networks – and compete in the market.

In fact, the existence and expansion of the Internet and the relative decline of proprietary networks and services, such as CompuServe, can be attributed to the conditions of inter-operability and the tremendous network externalities of the Internet. AOL, CompuServe, Prodigy, MCI and AT&T folded their proprietary electronic mail and other services into the Internet. Microsoft, thought to be the master of exploiting network effects, made the error of developing and marketing the proprietary MSN. After that product failed to sell, Microsoft re-launched the Microsoft Network as an Internet service provider, adhering fully to the public Internet standard. This is telling evidence of the power of the Internet standard and demonstrates the low likelihood that any firm can take control of the Internet backbone by imposing its own proprietary standard.

Second, customers on the Internet demand *universal connectivity*, so the second condition above fails. Users of the Internet do not know in advance what Internet site they may want to contact or to whom they might want to send e-mail. Thus, Internet users demand from their ISPs, and expect to receive, universal connectivity. This is the same expectation that users of telephones, mail and fax machines have: that they can connect to any other user of the network without concern about compatibility, location, or, in the case of telephone or fax, any concern about the manufacturer of the appliance, the type of connection (wireline or wireless) or the owners of the networks over which the connection is made. Because of the users' demand for universal connectivity, ISPs providing services to end users or to websites must make arrangements with other networks so that they can exchange traffic with *any* Internet customer.

Third, there are no 'captive' ISPs on the Internet, so the third condition fails, for a number of reasons:

- ISPs can easily and with low cost migrate all or part of their transport traffic to other network providers;
- many ISPs already purchase transport from more than one backbone to guard against network failures and for competitive reasons (ISP 'multi-homing');
- many large websites/providers use more than one ISP for their sites ('customer multi-homing'); and
- competitive pressure from their customers makes ISPs agile and likely to respond quickly to changes in conditions in the backbone market.

Competitive conditions imply that significant price increase, raising rivals' costs or degrading interconnection are unlikely to be profitable on the Internet backbone. If the large Internet backbone connectivity provider's strategy were to impose equal increases in transport costs on all customers, the response of other backbone providers and ISPs would be to reduce the traffic for which they buy transit from the large Internet backbone provider (IBP) and to instead re-route traffic and purchase more transit from each other. Thus, in response to a price increase by the large Internet backbone connectivity provider, other IBPs and ISPs would reduce the traffic for which they buy transit from the large IBP down to the minimum level necessary to reach ISPs that are *exclusively* connected to the large IBP. All other IBPs and ISPs would exchange all other traffic with each other bypassing the large IBP network.

Figure 2 shows the typical reaction of an increase in the price of a large IBP, and illustrates why the strategy of increasing price is unprofitable. Consider, for example, a situation where, prior to the price increase, four ISPs (1 to 4) purchase transit from IBP 0, which considers increasing its price. Two of these ISPs (ISP 2 and ISP 3) peer with each other. ISP 1 and ISP 4 buy transit capacity for all their traffic to IBP 0 and the other three ISPs. ISP 2 and ISP 3 buy transit capacity for all their traffic to IBP 0, ISP 1 and ISP 4.

Figure 2

Traffic flows between ISPs and a backbone

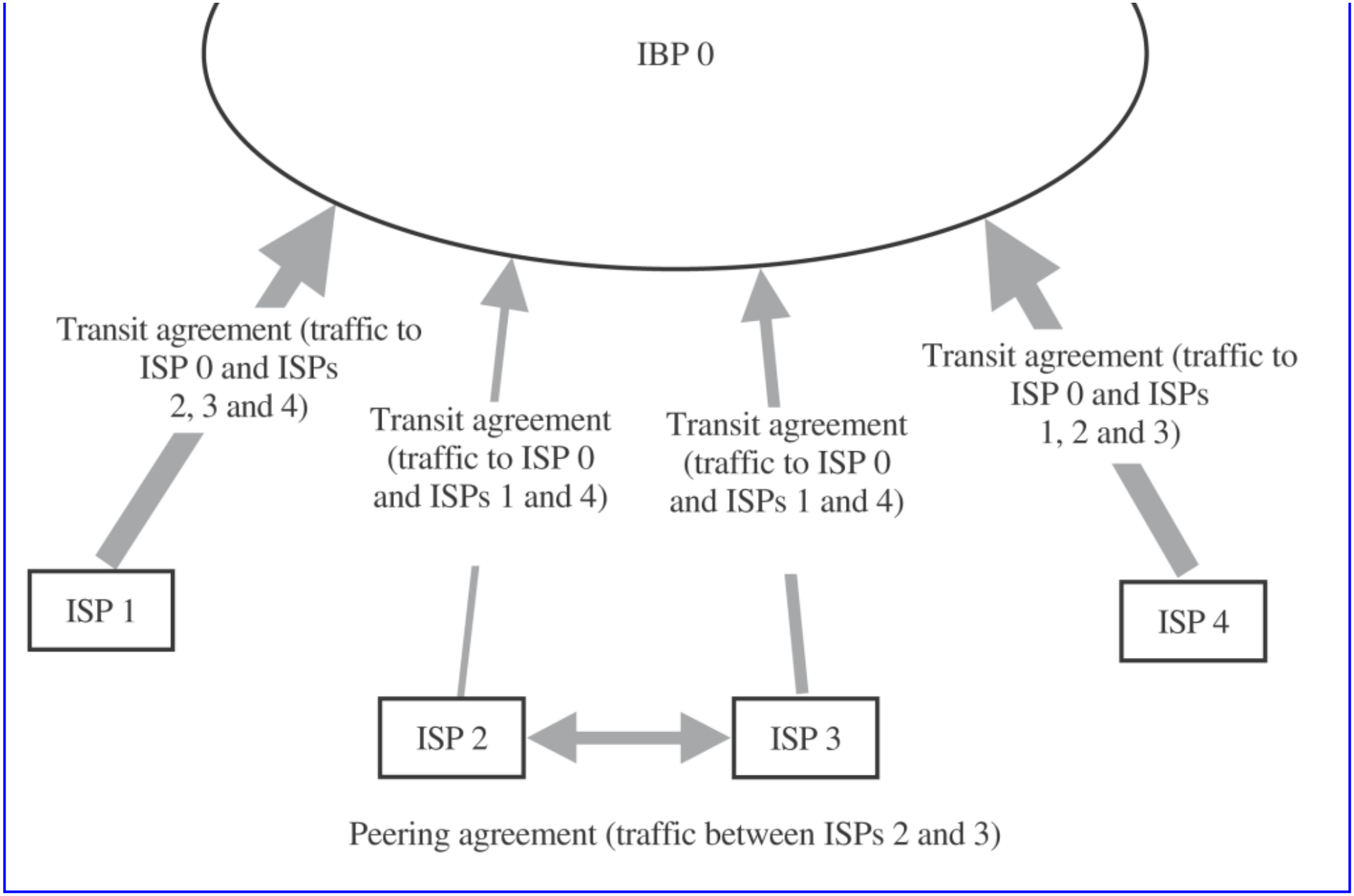

Now suppose that IBP 0 increases its transit price. In response, ISP 1 and ISP 4 decide to reduce the traffic for which they buy transit from IBP 0, and instead to re-route some of their traffic and purchase more transit from ISP 2 and ISP 3 respectively. Because of the peering relationship between ISP 2 and ISP 3, all traffic from ISP 1 handed to ISP 2 will reach ISP 3 as well as ISP 4, which is a customer of ISP 3. Similarly, by purchasing transit from ISP 3, ISP 4 can reach all the customers of ISP 1, ISP 2 and ISP 3. Thus, in response to the price increase of IBP 0, each of the ISPs 1, 2, 3 and 4 will reduce the amount of transit purchased from the IBP 0. Specifically, each of the ISPs buys from IBP 0 only capacity sufficient to handle traffic to the customers of network 0. This may lead to a considerable loss in revenues for IBP 0, rendering the price increase unprofitable. The big beneficiaries of the price increase of IBP 0 are peering ISPs 2 and 3, which now start selling transit to ISPs 1 and 4 respectively and become larger networks.

In response to a price increase by the large IBP, rivals would be able to offer their customers universal connectivity at profitable prices below the large IBP's prices. In the scenario described in the example above, market forces, responding to a price increase by a large network, re-route network traffic so that it is served by rival networks, except for the traffic to and from the ISPs connected exclusively with the large network. The rivals purchase the remaining share from the large IBP in order to provide universal connectivity. Thus, the rivals' blended cost would permit them to profitably offer all transport at prices lower than the large IBP's prices, but above cost.

A direct effect of the increase in price by the large network is that (*a*) ISPs that were originally exclusive customers of the large IBP would shift a substantial portion of their transit business to competitors, and (*b*) ISPs that were not exclusive customers of the large IBP would also shift a significant share of their transit business to competitors' networks, keeping the connection with the large IBP only for traffic for which alternative routes do not exist or for cases of temporary failure of the rivals' networks.

Similarly, degradation of interconnection to all backbones or sequentially one at a time is likely to be unprofitable. Degradation of interconnection to all backbones is clearly dominated by a price increase (since a price increase directly produces additional revenue to the firm, while interconnection degradation does not directly increase revenue), and, as we have shown above, competitive conditions severely limit price increases. Targeted degradation is also unprofitable for a large network that would initiate it for several reasons.

1. 1. ISP clients of the targeted network are likely to switch to third IBP networks that are unaffected by the degradation; it is very unlikely that any will switch to the degrading IBP network because it is itself degraded and cannot offer universal connectivity; there is no demand reward to the large IBP network.
2. 2. Degradation of interconnection hurts all the ISP customers of the targeting IBP network as well, since they lose universal connectivity; these customers of the large network would now be willing to pay less to the large network; this leads to significant revenue and profit loss.
3. 3. After losing universal connectivity, customers of the large IBP network are likely to switch to other networks that are unaffected by degradation and can provide universal connectivity; this leads to even further revenue and profit loss for the degrading network.
4. 4. Multi-homing ISPs would purchase less capacity from the large IBP network, or even terminate their relationship with the large network, which through its own actions sabotages their demand for universal connectivity; this further reduces demand and profits for the degrading network; the same argument applies to multi-homing customers of ISPs.
5. 5. As the large IBP network pursues target after target, its customers face continuous quality degradation while the target's customers face only temporary degradation; this would result in further customer and profit losses for the large IBP network.
6. 6. Prospective victims would seek alternative suppliers in advance of being targeted by the large IBP network; the scheme cannot play out the way it is proposed.
7. 7. The degradation scheme is implausible in its implementation. How large do networks need to be to become serial killers? Why have we not observed this behaviour at all?
8. 8. There is no enduring change to the number of competitors in a market caused by serial degradation in a market with negligible entry barriers; the eliminated rival is likely to be replaced by another.

In conclusion, competition on the Internet backbone is strong, with many carriers and easy entry, and thus presently there are no significant competition concerns for Internet backbone services. However, local broadband access is typically a duopoly or monopoly depending on location. As of 2007, local broadband access networks were proposing to abolish the regime of net neutrality and impose fees on content and applications providers. The legality of this proposed change is questionable, and imposition of such price discrimination may have adverse consequences for consumers' total surplus.

See Also

- computer industry

Bibliography

Bradner, S. 1999. *The Internet Standards Process*, revision 3, Network Working Group. Online. Available at ftp://ftp.isi.edu/in-notes/rfc2026.txt, accessed 24 January 2007.

Cremer, J., Rey, P. and Tirole, J. 1998. The degradation of quality and the domination of the internet. Appendix 5 in the *Submission of GTE to the European Union for the merger of MCI with WorldCom*.

Cremer, J., Rey, P. and Tirole, J. 2000. Connectivity in the commercial internet. *Journal of Industrial Economics* 48, 433–72.

Economides, N. 1989. Desirability of compatibility in the absence of network externalities. *American Economic Review* 79, 1165–81.

Economides, N. 1996. The economics of networks. *International Journal of Industrial Organization* 14, 675–99.

Economides, N. 2006a. The economics of the internet backbone. In *Handbook of Telecommunications*, vol. 2, ed. S. Majumdar, I. Vogelsang and M. Cave. Amsterdam: Elsevier Publishers.

Economides, N. 2006b. Competition policy in network industries: an introduction. In *The New Economy and Beyond: Past, Present and Future*, ed. J. Dennis. Cheltenham: Edward Elgar.

European Commission. 1998. *Statement of Objections to the MCI WorldCom Merger*.

Farrell, J. and Saloner, G. 1985. Standardization, compatibility, and innovation. *RAND Journal of Economics* 16, 70–83.

Kahn, R.E. and Cerf, V.G. 1999. What is the internet (and what makes it work). Reston, VA: Corporation for National Research Initiatives. Online. At http://www.cnri.reston.va.us/what_is_internet.html, accessed 24 January 2007.

Katz, M. and Shapiro, C. 1985. Network externalities, competition and compatibility. *American Economic Review* 75, 424–40.

Liebowitz, S.J. and Margolis, S.E. 1994. Network externality: an uncommon tragedy. *Journal of Economic Perspectives* 8(2), 133–50.

Milgrom, P., Mitchell, B.M. and Srinagesh, P. 2000. Competitive effects of internet peering policies. In *The Internet Upheaval – Raising Questions, Seeking Answers in Communications Policy*, ed. I. Vogelsang and B.M. Compaine. Cambridge, MA and London: MIT Press.

How to cite this article

Economides, Nicholas. "Internet, economics of the." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000244> doi:10.1057/9780230226203.0840

The New Palgrave Dictionary of Economics Online

interpersonal utility comparisons (new developments)

Claude d'Aspremont

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Recent developments on interpersonal utility comparisons rely on various interpretations of ‘utility’ indicators and combine in various degrees the ‘subjective’ appreciation of the social states by each individual and their ‘objective’ evaluation by the ethical observer. In a formal welfarist approach, interpersonal comparisons are specified by invariance conditions on social welfare functionals or on social welfare orderings. Interpersonal comparisons have also been introduced through scoring methods.

Keywords

bargaining solution (Nash); Bentham, J.; capabilities; cardinal utility; collective preference; difference principle (J. Rawls); egalitarian-equivalent allocation; fair allocation; happiness; Harsanyi, J. C.; impossibility theorem; independence of irrelevant alternatives; interpersonal utility comparisons; justice; leximin; Mill, J. S.; new welfare economics; *ophélimité* ; ordinal utility; Pareto, V.; preferences; primary goods; Rawls, J.; revealed preference theory; risk aversion; Sen, A.; social choice; social welfare function; social welfare ordering; utilitarianism; utility; veil of ignorance; von Neumann and Morgenstern; welfarism

Article

Distributive justice, whether in normative economics or in collective choice theory, can hardly be treated without introducing some interpersonal comparisons. But does this mean considering interpersonal *utility* comparisons? The term ‘utility’ has received so many different interpretations that the distinguishing mark of the utility approach to the evaluation of *social states* by an ethical observer is simply that it assigns to each member of the collectivity a *unidimensional individual indicator*. These indicators combine in various ways the ‘subjective’ appreciation of the social states by each individual and their ‘objective’ evaluation by the observer.
Bentham introduced an interpersonally summable notion of utility based on the objective property of

things to procure either pain or pleasure, but it is the subjective reinterpretation given by J.S. Mill, for whom utility means ‘pleasure itself’ and ‘exemption of pain’, which has prevailed (Mongin and d'Aspremont, 1998). In Harsanyi's (1977) reformulation of utilitarianism, the ethical observer is, with equal chance, any one of the individuals and obeys the rationality conditions of decision-making under risk. The criterion is expected utility computed from the individual von Neumann–Morgenstern (NM) utility functions, but ‘corrected’ for factual errors and ‘censored’ for anti-social attitudes.

Pareto has clearly distinguished between the objective notion of *utilité* (in the Bentham sense) and the subjective notion of *ophélimité*, as an ordinal measure of actual preference satisfaction. The latter has become the dominant concept in economics (the only concept in the new welfare economics), forbidding interpersonal utility comparisons and, ultimately, reducing preferences to observable individual choices (revealed preference theory). But Samuelson (1947) insisted that welfare economics cannot avoid ethical and interpersonal assumptions, and Arrow (1951) derived an ‘impossibility theorem’. For a set $N=\{1, 2, \ldots, n\}$ of individuals and a set $X=\{x, y, \ldots\}$ of (more than two) social states, there is no acceptable *social welfare function*, associating every profile of individual preference orderings with one ‘collective’ preference ordering of X, and satisfying *weak Pareto* (if all strictly prefer one state to another, so should society) and *independence of irrelevant alternatives* (if individual preferences are modified except for a subset of several alternatives, then the collective preference should not be modified on this subset). Such a social welfare function can only be *dictatorial*: one individual imposes his strict preference. Since all Arrow's assumptions concern ‘preference satisfaction’, modifying them and reinterpreting ‘utility’ involve ethical considerations.

Rawls's (1971) principles of justice are agreed upon in some original negotiation where all irrelevant personal features (including personal conceptions of the good) are ignored. Also chosen behind this ‘veil of ignorance’ is an ‘index of primary goods’ defined as an objective indicator of the fundamental resources (except for liberties and access to occupations, preliminarily and equally divided) allocated to each person to promote his own conception of the good. If such an indicator can be called ‘utility’, it is not in the sense of ‘happiness’ or ‘preference satisfaction’. Desires (however intense) and tastes (however inexpensive) are not relevant per se. A similar view is represented by Sen's (1992) notion of capabilities, that is, the set of doings or ‘functionings’ available to a person, leading to an ‘index of functionings’. What is at stake here, as in other theories concerned by opportunities (for example, Roemer, 1996), are the objectively defined conditions allowing individuals to exercise their freedom.

Extended sympathy, social welfare functionals and welfarism

To formally examine the role of interpersonal utility comparisons in social choice, it is usual to start within the framework introduced by Sen (1970), in which the basic ingredient is a utility profile given by a real-valued function U defined on elements (x,i) of the Cartesian product $X\times N$. The function U can be seen as a vector of individual indicators $U_i(x) \equiv U(x,i)$, or as the *extended utility function* of an individual, evaluating from a moral viewpoint what it is to be anyone in any social state (exercising ‘extended sympathy’ or ‘empathy’). Moreover, if for individual i one interprets the name i as designating all the characteristics of i, then one could look at the function U as a *fundamental utility function* (Harsanyi, 1977), itself a representation of ‘human nature’ (Kolm, 1972), which would then justify why every individual, when adopting the viewpoint of an ethical observer, should have the same

extended utility function (or at least the same fundamental preference). Lack of identity could lead to a dictatorial ethical observer (see Suzumura, 1996).

The fundamental utility approach is also used in econometric estimations to define ‘adult-equivalent scales’ and different forms of exact aggregation (Christensen, Jorgenson and Lau, 1975; Blackorby and Donaldson, 1991; Deaton and Muellbauer, 1980). Other measurement techniques, such as that which uses the number of ‘just-noticeable-differences’ between two alternatives (discussed in Arrow, 1951) or the ‘social indicators’ approach using questionnaires about degree of happiness (discussed in Hammond, 1991; Fleurbaey and Hammond, 2004) are differently founded.

The U function is a very flexible informational basis to start with. For every x, we can denote U_x the *utility vector* $(U(x, 1), \ldots, U(x, n))$ in $\mathbb{R}^N$. Taking all functions U in some domain $\mathcal{D}$ determines the set of all admissible utility vectors. Sen's (1970) concept of *social welfare functional* (SWFL) associates every admissible extended utility function U in $\mathcal{D}$ with one (collective) preference ordering R_U. We denote I_U and P_U the corresponding indifference and strict preference relations. Using this notation, *Pareto indifference* means ‘$U_x=U_y$ implies xI_Uy’ and *strong Pareto* requires in addition that ‘$U_x \geq U_y$ and $U_x \neq U_y$ implies $x\ P_Uy$’. Also, Arrow's independence of irrelevant alternatives can be weakened to *binary independence*, whereby for any two functions U and V with equal values on two social states x and y we have $xR_VV \Leftrightarrow xR_Uv$.

An alternative framework is to define directly a *social welfare ordering* (SWO) denoted R^* on the set of admissible utility vectors. If the set of admissible utility vectors is large enough (for example, equal to $\mathbb{R}^N$), then, under Pareto indifference and binary independence, the two frameworks coincide: $u=U_x$ and $v=U_y$ implies $uR^*v \Leftrightarrow xR_Uv$. This is called *welfarism* and is an extreme form of consequentialism. All the information required for social evaluation is contained in the final utility values. Under welfarism, strong Pareto (SP) reduces to the condition that $u \geq v$, and $u \neq v$ implies uP^*v (P^* denoting strict collective preference).

Invariance axioms

Measurement theory (Krantz et al., 1971; Roberts, 1980) associates with different measurement scales the associated meaningful statements. We are interested in meaningful statements about intrapersonal and interpersonal comparisons of utility. For instance, the Arrowian informational basis for SWFLs requires that only *intrapersonal level comparisons* are meaningful: $R_U=R_V$ whenever, for every i and all x, y, $U(x, i) \geq U(y, i)$ if and only if $V(x, i) \geq V(y, i)$. Another example (for this and others see Bossert and Weymark, 2004) is to consider meaningful *interpersonal comparisons of utility differences*: $R_U=R_V$ whenever, for all w, x, y, z and all i, j, $U(w, i) - U(x, i) \geq U(y, j) - U(z, j)$ if and only if $V(w, i) - V(x, i) \geq V(y, j) - V(z, j)$.

The more standard way (for example, Sen, 1977) to specify the measurability and comparability properties of ‘utility’ is to introduce invariance transformations $\varphi=(\varphi_1,\varphi_2, \ldots, \varphi_n)$, each φ_i being a real-valued function on $\mathbb{R}$. In the Arrowian framework, we get the invariance axiom of *ordinality and non-*

comparability (ON): if each φ_i is increasing and if for every x, $V(x,i)=\varphi_i(U(x,i))$, then $R_U=R_V$.
Corresponding to interpersonal comparisons of utility differences, we get *cardinality and unit-comparability* (CU): if each φ_i is a positive affine transformation ($\varphi_i(u_i)=a_i+bu_i$, with $b>0$) and if, for all (x,i), $V(x,i)=a_i+bU(x,i)$, then $R_U=R_V$.
But there is a third way of specifying such conditions. We have stated these two axioms as restrictions on SWFLs. Under welfarism, they can be translated into axioms on SWOs, as

$$\text{ON: for any increasing } \phi_i\text{'s}, \quad uR^*v \Leftrightarrow (\phi_1(u_1), \ldots, \phi_n(u_n))R^*(\phi_1(v_1), \ldots, \phi_n(v_n));$$

$$\text{CU: for any } a_i\text{'s}, \quad b>0, \quad uR^*v \Leftrightarrow (a_1+bu_1, \ldots, a_n+bu_n)R^*(a_1+bv_1, \ldots, a_n+bv_n).$$

Invariance axioms determine the informational basis for social evaluation. They should not be considered as purely factual. By specifying the kind of information that a social evaluation can or cannot use, these axioms are taking an ethical stance. But their strong ethical implications are better measured when combined with other axioms. To illustrate, the following axiom allows for (and only for) comparisons of utility differences that are intrapersonal. It is *cardinality and non-comparability* (CN): for any a_i's and positive b_i's, if $V(x,i)=a_i+b_iU(x,i)$, for all (x,i), then $R_U=R_V$. Such a SWFL version of this axiom does not exclude interpersonal utility comparisons. It does allow us to compare *ratios of utility differences* of the sort $(U(w,i)-U(x,i))/(U(y,i)-U(z,i))$ between different individuals, and hence to compare measures of risk aversion in case X is specified as a set of lotteries and each $U_i(x)$ as an NM utility function. But the possibility of such comparisons is erased under welfarism, under which cardinal non-comparability reduces to ordinal non-comparability and, with strong Pareto, implies dictatorship (by Arrow's theorem). Under welfarism, ON becomes equivalent to CN: for any a_i's, positive b_i's,

$$uR^*v \Leftrightarrow (a_1+b_1u_1, \ldots, a_n+b_nu_n)R^*(a_1+b_1v_1, \ldots, a_n+b_1v_n).$$

If CN is replaced by CU and dictatorship is excluded by *anonymity* (any utility vector is socially indifferent to any of its permutations), then the only possibility is the *pure utilitarian* SWO: uR^*v if and only if $\sum_{i\in N} u_i \geq \sum_{i\in N} v_i$ (d'Aspremont and Gevers, 1977).
This characterization of utilitarianism is directly related to Harsanyi's aggregation theorem since, under welfarism (and NM preferences), cardinality and unit-comparability become equivalent to NM-

independence of the collective preference ordering (Mongin and d'Aspremont, 1998).
By giving priority to liberties and access to occupations, Rawls clearly departs from welfarism, even in a formal sense. However, to allocate other primary goods, a common index $V(x_i)$ is fixed, where x_i is the vector of primary goods to be allocated to individual i. Letting, for an allocation $x=(x_1,\ldots,x_n)$, $U(x,i)=V(x_i)$, we can fall again into the welfarist formal framework. Since the index is common to all, the associated invariance axiom is *ordinality and comparability*

$$OC:\ \text{for any increasing } \widehat{\Phi},\ \ uR^*v \Leftrightarrow (\widehat{\Phi}(u_1),\ldots,\widehat{\Phi}(u_n))R^*(\widehat{\Phi}(v_1),\ldots,\widehat{\Phi}(v_n)).$$

Two other axioms are clearly required by Rawls: *anonymity* and *strong Pareto*, the latter being the reason why equal distribution of all primary goods is not the agreed-upon solution. To any u in $\mathbb{R}^N$, one can associate a (re)ordered vector $u_{i(\cdot)}$ with same components in $\{v\in\mathbb{R}^N : v_1 \le v_2 \le \ldots \le v_n\}$, the set of ordered utility vectors. *Minimal equity* requires that one should never give priority to the best-off individual over the worst-off. Then, under *separability* (in choosing between two utility vectors the indifferent individuals should not be taken into account), the solution is the ‘lexicographic maximin’ (*leximin*) SWO : for any u, v in $\mathbb{R}^N$, uP^*v if and only if, for some k, $1 \le k \le n$, $u_{i(k)}>v_{i(k)}$, and, $\forall j<k$, $1 \le j \le n$, $u_{i(j)}=v_{i(j)}$. Leximin formalizes Rawls's ‘difference principle’. Other concepts of opportunity equalizations can be so translated into welfarist terms (see Maniquet, 2004).
From a formal viewpoint, in the preceding result only utility levels are both intrapersonally and interpersonally comparable. If we add the same possibility for utility differences we have *full comparability*, that is

$$FC:\ \text{for any } a,\ b>0,\ \ uR^*v \Leftrightarrow (a+bu_1,\ldots,a+bu_n)R^*(a+bv_1,\ldots,a+bv_n).$$

With this type of invariance and the same other assumptions (Deschamps and Gevers, 1978), the SWO R^* can be either leximin or utilitarianism, but in a weak sense (that is $\sum_{i\in N} u_i > \sum_{i\in N} v_i$ implies uP^*v).
Many other invariance axioms can be introduced (for example, d'Aspremont, 1985). Let us give only two more, *ratio-scale measurability*, *without* or *with interpersonal comparisons* of utility:

$$RN:\ \text{for any positive } b_i\text{'s, and } u,v\in\mathbb{R}^N,\ \ uR^*v \Leftrightarrow (b_1u_1,\ldots,b_nu_n)R^*(b_1v_1,\ldots,b_nv_n);$$

$$\text{RC: for any positive } b, \text{ and } u, v \in \mathbb{R}^N, \ uR^*v \Leftrightarrow buR^*bv.$$

With ratio-scale measurability the origin is fixed, so that, under *RN*, utility levels are interpersonally comparable if they are of opposite sign (below or above the zero line). Moreover, under RC, all utility levels and differences are comparable. These axioms are most often applied on a positive domain (denoted $\mathbb{R}^N_{++}$). Under RN, ratios of utilities or percentage changes in utility are interpersonally comparable. If we add SP, then a continuous SWO on $\mathbb{R}^N_{++}$ can only be the Nash bargaining solution with status quo point normalized to zero: for some positive θ_i's, uR^*v if and only if $\Pi_{i=1}^n u_i^{\lambda_i} \geq \Pi_{i=1}^n v_i^{\lambda_i}$. Under *RC* and *SP* the set of continuous and anonymous SWOs is characterized by all homothetic, increasing, continuous and symmetric functions on $\mathbb{R}^N_{++}$ (for these and other results, see Bossert and Weymark, 2004).

Beyond welfarism: scoring and fair allocation rules

Even when ‘utility’ simply represents actual preference satisfaction, it can be severely adjusted by the ethical observer (or rule designer). This is the case for various voting rules or more generally for *scoring rules*. These rules violate binary independence in one way or another, so that welfarism is excluded. Voting methods, such as Borda's, are generally not acceptable for social evaluation, but some related rules are better candidates (Moulin, 1988). Another ‘scoring’ method is *relative utilitarianism* (axiomatized by Dhillon and Mertens, 1999). Each $U_i(.)$ is supposed to have both a maximum $U_i^{\max}$ and a minimum $U_i^{\min}$ and to represent a NM preference ordering on X and the observer associates to it a NM utility function $V_i(.)$ ordinally equivalent to $U_i(.)$, and then, through individual affine transformations, defines the scoring function $S_i(x) = (V_i(x) - V_i^{\min}) / (V_i^{\max} - V_i^{\min})$. The score $S_i(x)$ is the same whatever the arbitrarily chosen NM utility function $V_i(.)$ representing the preference ordering underlying the initial utility function $U_i(.)$, and measures of curvatures are preserved. The SWFL F is taken to be pure utilitarianism applied to the scores. It satisfies *ordinality and non-comparability* with respect to the utility functions $U_i(x)$ representing the individual preferences. Since the scores are defined as ratios of utility differences, they could be interpersonally compared, but, because their aggregation is utilitarian, it relies only on interpersonal comparisons of differences of scores.
Other concepts have been proposed in the literature on fair allocations, excluding welfarism in terms of the initial utility functions, but ending up applying some SWFL to some recalibrated utility. An example

is the Pareto efficient egalitarian-equivalent allocation concept (Pazner and Schmeidler, 1978). To illustrate, let $\overline{\omega}$ in $\mathbb{R}^L_+$ be the vector of total quantities of L goods and X be the set of feasible allocations $(x_1, \ldots, x_n)$: each x_i is in $\mathbb{R}^L_+$ and $\sum_{i=1}^n x_i \leq \overline{\omega}$. If $U(x,i)=U_i(x_i)$ is increasing in each argument and continuous then one can define an ordinally equivalent function $V_i(x_i)$ such that $U_i(x_i) = U_i(V_i(x_i)\overline{\omega})$. A Pareto efficient allocation $\hat{x}$ in X is *egalitarian-equivalent* if $U_i(\hat{x}_i) = U_i(V_i(\hat{x}_i)\overline{\omega})$ and $V_i(\hat{x}_i) = V_j(\hat{x}_j)$ for all i,j. As observed in Fleurbaey and Hammond (2004), if individual preferences are convex, $\hat{x}$ can be obtained by applying the leximin SWFL on the $V_i(x_i)$'s. Again, starting with a concept defined in terms of purely ordinal and non-comparable utilities (the U_i's), we end up comparing and equalizing utility levels in terms of the V_i's.

See Also

- Arrow's theorem
- bargaining
- equality of opportunity
- fair allocation
- interpersonal utility comparisons
- Pareto principle and competing principles
- revealed preference theory
- risk aversion
- social choice
- social choice (new developments)
- social welfare function
- utility
- welfare economics

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*. 2nd edn. New Haven, CT: Yale University Press, 1963.

Blackorby, C. and Donaldson, D. 1991. Adult-equivalence scales, interpersonal comparisons of well-being, and applied welfare economics. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. E. Roemer. Cambridge: Cambridge University Press.

Bossert, W. and Weymark, J.A. 2004. Utility in social choice. In *Handbook of Utility Theory*, vol. 2, ed. S. Barberà, P. Hammond and C. Seidl. Dordrecht: Kluwer.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1975. Transcendental logarithmic utility functions. *American Economic Review* 65, 367–83.

d'Aspremont, C. 1985. Axioms for social welfare orderings. In *Social Goals and Social Organizations: Essays in Memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler and H. Sonnenschein. Cambridge: Cambridge University Press.

d'Aspremont, C. and Gevers, L. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44, 199–209.

Deaton, A. and Muellbauer, J. 1980. *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.

Deschamps, R. and Gevers, L. 1978. Leximin and utilitarian rules: a joint characterization. *Journal of Economic Theory* 17, 143–63.

Dhillon, A. and Mertens, J.-F. 1999. Relative utilitarianism. *Econometrica* 67, 417–98.

Fleurbaey, M. and Hammond, P.J. 2004. Interpersonally comparable utility. In *Handbook of Utility Theory*, vol. 2, ed. S. Barberà, P.J. Hammond and C. Seidl. Dordrecht: Kluwer.

Hammond, P.J. 1991. Interpersonal comparisons of utility: why and how they are and should be made. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J.E. Roemer. Cambridge: Cambridge University Press.

Harsanyi, J.C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Kolm, S.C. 1972. *Justice et Equité*. Paris: CNRS.

Krantz, D., Luce, R.D., Suppes, P. and Tversky, A. 1971. *Foundations of Measurement*, vol. 1: *Additive and Polynomial Representations*. New York: Academic Press.

Maniquet, F. 2004. On the equivalence between welfarism and equality of opportunity. *Social Choice and Welfare* 23, 1–21.

Mongin, P. and d'Aspremont, C. 1998. Utility theory and ethics. In *Handbook of Utility Theory*, vol. 1, ed. S. Barberà, P. Hammond and C. Seidl. Dordrecht: Kluwer.

Moulin, H. 1988. *Axioms of Cooperative Decision Making*. Cambridge: Cambridge University Press.

Pazner, E.A. and Schmeidler, D. 1978. Egalitarian-equivalent allocations: a new concept of economic equity. *Quarterly Journal of Economics* 92, 671–87.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Roberts, K.W.S. 1980. Interpersonal comparability and social choice theory. *Review of Economic Studies* 47, 421–39.

Roemer, J.E. 1996. *Theories of Distributive Justice*. Cambridge, MA: Harvard, University Press.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Sen, A.K. 1977. On weights and measures: informational constraints in social welfare analysis. *Econometrica* 45, 1539–72.

Sen, A.K. 1992. *Inequality Re-Examined*. Cambridge, MA: Harvard University Press.

Suzumura, K. 1996. Interpersonal comparisons and the possibility of social choice. In *Social Choice Re-Examined*, vol. 2, ed. K.J. Arrow, A.K. Sen and K. Suzumura. London: Macmillan.

How to cite this article

d'Aspremont, Claude. "interpersonal utility comparisons (new developments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000285> doi:10.1057/9780230226203.0842

The New Palgrave Dictionary of Economics Online

interpersonal utility comparisons

John C. Harsanyi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Although we all make interpersonal utility comparisons, many economists and philosophers argue that our limited information about other people's minds renders them meaningless. If they are possible, interpersonal comparisons of utility *differences* must be distinguished from interpersonal comparisons of utility *levels*. Utilitarianism must assume the interpersonal comparability of utility differences to maximize a social welfare function, while Rawls's maximin principle requires interpersonal comparability of utility levels. Adopting an ordinalist or a cardinalist view of utility functions restricts the positions one can consistently take as to interpersonal comparability of utilities.

Keywords

Arrow, K.; interpersonal utility comparisons; maximin; Rawls, J.; Robbins, L.; utilitarianism; utility: cardinal vs. ordinal; von Neumann–Morgenstern utility function

Article

Suppose I am left with a ticket to a Mozart concert I am unable to attend and decide to give it to one of my closest friends. Which friend should I actually give it to? One thing I will surely consider in deciding this is which friend of mine would enjoy the concert *most*. More generally, when we decide as private individuals whom to help, or decide as voters or as public officials who are to receive government help, *one* natural criterion we use is who would derive the greatest benefit, that is, who would derive the *highest utility*, from this help. But to answer this last question we must make, or at least attempt to make, *interpersonal utility comparisons*.

At the common-sense level, all of us make such interpersonal comparisons. But philosophical reflection might make us uneasy about their meaning and validity. We have direct introspective access only to our *own* mental processes (such as our preferences and our feelings of satisfaction and dissatisfaction) defining our *own* utility function, but have only very indirect information about other people's mental processes. Many economists and philosophers take the view that our limited information about other

people's minds renders it impossible for us to make meaningful interpersonal comparisons of utility.

Comparisons of utility levels vs. comparisons of utility differences

In any case, if such comparisons are possible at all, then we must distinguish between interpersonal comparisons of utility levels and interpersonal comparisons of utility differences (i.e. utility increments or decrements).

It is one thing to compare the utility level $U_i(A)$ that individual i enjoys (or would enjoy) in situation A, with utility level $U_j(B)$ that another individual j enjoys (or would enjoy) in situation B (where A and B may not refer to the same situation). It is a very different thing to make interpersonal comparisons between utility differences, such as comparing the utility increment

$$\Delta U_i(A, A') = U_i(A') - U_i(A) \quad (1)$$

that individual i would enjoy in moving from situation A to situation A', with the utility increment

$$\Delta U_j(B, B') = U_j(B') - U_j(B) \quad (2)$$

that individual j would enjoy in moving from B to B'. Either kind of interpersonal comparison might be possible without the other kind being possible (Sen, 1970).

Some ethical theories would require one kind of interpersonal comparisons; others would require the other. Thus, *utilitarianism* must assume the interpersonal comparability of utility *differences* because it asks us to maximize a social utility function (social welfare function) defined as the *sum* of all individual utilities. (There are arguments for defining social utility as the *arithmetic mean*, rather than the *sum*, of individual utilities (Harsanyi, 1955). But for most purposes – other than analysing population policies – the two definitions are equivalent because if the number of individuals can be taken for a *constant*, then maximizing the sum of utilities is mathematically equivalent to maximizing their arithmetic mean.) Yet, we cannot add different people's utilities unless all of them are expressed in the same utility units; and in order to decide whether this is the case, we must engage in interpersonal comparisons of utility differences. (On the other hand, utilitarianism does not require comparisons of different people's utility levels because it does not matter whether their utilities are measured from comparable zero points or not.)

Likewise, the interpersonal utility comparisons we make in everyday life are most of the time comparisons of utility *differences*. For instance, the comparisons made in our example between the

utilities that different people would derive from a concert obviously involve comparing utility differences.

In contrast, the utility-based version of Rawls's *Theory of Justice* (1971) does require interpersonal comparisons of utility *levels*, but does not require comparisons of utility *differences*. This is so because his theory uses the *maximin principle* (he calls it the *difference principle*) in evaluating the economic performance of each society, in the sense of using the well-being of the *worst-off* individual (or the worst-off social group) as its principal criterion. But to decide which individuals (or social groups) are worse off than others he must compare different people's utility levels. (In earlier publications, Rawls seemed to define the worst-off individual as one with the lowest utility level. But in later publications, he defined him as one with the smallest amount of ‘primary goods’. For a critique of Rawls's theory, see Harsanyi, 1975.)

Ordinalism, cardinalism and interpersonal comparisons

In studying comparisons between the utilities enjoyed by *one* particular individual *i*, we again have to distinguish between comparisons of utility *levels* and comparisons of utility *differences*. The former would involve comparing the utility levels $U_i(A)$ and $U_i(B)$ that *i* assigns to two different situations *A* and *B*. The latter would involve comparing the utility increment

$$\Delta U_i(A, A') = U_i(A') - U_i(A) \tag{3}$$

that *i* would enjoy in moving from situation *A* to situation A', with the utility increment

$$\Delta U_i(B, B') = U_i(B') - U_i(B) \tag{4}$$

that he would enjoy in moving from *B* to B'.

If *i* has a well-defined utility function U_i at all, then he certainly must be able to compare the utility *levels* he assigns to various situations; and such comparisons will have a clear behavioural meaning because they will correspond to the preference and indifference relations expressed by his choice behaviour. In contrast, it is immediately less obvious whether comparing utility differences as defined under (3) and (4) has any economic meaning (but see below).

A utility function U_i permitting meaningful comparisons only between *i*'s utility levels, but not permitting such comparisons between his utility differences, is called ordinal; whereas a utility function permitting meaningful comparisons both between his utility levels and his utility differences is called

cardinal.

As is well known, most branches of economic theory use only ordinal utilities. But, as von Neumann and Morgenstern (1947) have shown, cardinal utility functions can play a very useful role in the theory of risk taking. In fact, utility-difference comparisons based on von Neumann–Morgenstern utility functions turn out to have a direct behavioural meaning. For example, suppose that U_i is such a utility function, and let Δ_i^* and Δ_i^{**} be utility differences defined by (3) and by (4). Then, the inequality $\Delta_i^* > \Delta_i^{**}$ will be algebraically equivalent to the inequality

$$\frac{1}{2}U_i(A') + \frac{1}{2}U_i(B) > \frac{1}{2}U_i(B') + \frac{1}{2}U_i(A). \tag{5}$$

This inequality in turn will have the behavioural interpretation that *i prefers* an equi-probability mixture of A' and of B to an equi-probability mixture of B' and of A. Of course, once von Neumann–Morgenstern utility functions are used in the theory of risk taking, they become available for possible use also in other branches of economic theory, including welfare economics as well as in ethical investigations. (It has been argued that von Neumann–Morgenstern utility functions have no place in ethics (or in welfare economics) because they merely express people's attitudes toward *gambling*, which has no moral significance (Arrow, 1951, p. 10; and Rawls, 1971, pp. 172 and 323). But see Harsanyi, 1984.)

Note that by taking an ordinalist or a cardinalist position, one restricts the positions one can consistently take as to interpersonal comparability of utilities:

1. (1) An *ordinalist* is logically free to *reject* both types of interpersonal comparisons. Or he may *admit* comparisons of different people's utility *levels*. But he *cannot* admit the interpersonal comparability of utility differences without becoming a cardinalist. (The reason is this. If the utility differences experienced by one individual *i* are comparable with those experienced by *another* individual *j*, this will make the utility differences experienced by *one* individual (say) *i* likewise indirectly comparable with one another, which will enable us to construct a *cardinal* utility function for each individual.)
2. (2) A *cardinalist* is likewise logically free to *reject* both types of interpersonal comparisons. Or he may *admit* both. Or else he may admit interpersonal comparisons only for utility *differences*. (Though it is hard to see why anybody might want to reject interpersonal comparisons for utility levels if he admitted them for utility differences.) But he *cannot* consistently admit interpersonal comparisons for utility *levels* while rejecting them for utility *differences*. (This can be verified as follows. If utility levels are interpersonally comparable, then we can find four situations A, A', B, and B' such that $U_i(A) = U_j(B)$ and $U_j(A') = U_j(B')$. But then we can conclude that

$$\Delta_i^* = U_i(A') - U_i(A) = \Delta_j^* = U_j(B') - U_j(B),$$

which means that at least the utility differences Δ_i^* and Δ_j^* are interpersonally comparable. But since U_i and U_j are *cardinal* utility functions, any utility difference Δ_i^{**} experienced by *i* is comparable with Δ_i^*, and any utility difference Δ_j^{**} experienced by *j* is comparable with Δ_j^*. Yet this means that *all* utility differences Δ_i^{**} experienced by *i* are comparable with *all* utility differences Δ_j^{**} experienced by *j*. Thus, cardinalism together with interpersonal comparability of utility levels *entails* that of utility differences.)

Extended utility functions

In what follows, I will use the symbols A_i, B_i,•…•to denote the economic and non-economic resources available to individual *i* in situations *A*, *B*,•…•Moreover, I will use the symbol A_j to denote an arrangement under which *j* has the same resources available to him as were available to individual *i* under arrangement A_i. These entities A_i, B_i,•…,•A_j, B_j,•…•I will call positions.
Interpersonal utility comparisons would pose no problem if all individuals had the same utility function. For in this case, any individual *j* could assume that the utility level $U_i(A_i)$ that another individual *i* would derive from a given position A_i should be the *same* as he himself would derive from a similar position. Thus, *j* could write simply

$$U_i(A_i) = U_j(A_j). \tag{6}$$

Of course, in actual fact, the utility of different people are rather *different* because people have different *tastes*, that is, they have different abilities to derive satisfactions from given resource endownments. I will use the symbols R_i, R_j,•…•to denote the vectors listing the personal psychological characteristics of each individual *i*, *j*,•…•that *explain* the differences among their utility functions U_i, U_j,•…•Presumably, these vectors summarize the effects that the genetic make-up, the education and the life experience of each incividual have on his utility function. This means that any individual *j* can attempt to assess the utility level $U_i(A_i)$ that another individual *j* would enjoy in position A_i as

$$U_i(A) = V(A_i, R_i),$$

(7)

where the function V represents the psychological laws determining the utility functions U_i, U_j,•…•of the various individuals i, j,•…•in accordance with their psychological parameters specified by the vectors R_i, R_j,•…•. Since, by assumption, all differences among the various individuals' utility functions U_i, U_j,•…•are fully explained by the vectors R_j, R_j,•…•, the function V itself will be the same for all individuals. We will call V an *extended utility function*. (See Arrow, 1978; and Harsanyi, 1977, pp. 51–60; though the basic ideas are contained already in Arrow, 1951, pp. 114–15.)

To be sure, we know very little about the psychological laws determining people's utility functions and, therefore, know very little about the true mathematical form of the extended utility function V. This means that, when we try to use equation (7), the best we can do is to use our – surely very imperfect – personal *estimate* of V, rather than V itself. As a result, in trying to make interpersonal utility comparisons, we must expect to make significant errors from time to time – in particular when we are trying to assess the utility functions of people with a very different cultural and social background from our own. But even if our judgements of interpersonal comparisons can easily be mistaken, this does not imply that they are meaningless.

Ordinalists will interpret both the functions U_i and the function V as ordinal utility functions and will interpret (7) merely as a warrant for interpersonal comparisons of utility levels (cf. Arrow, 1978). In contrast, cardinalists will interpret all these as *cardinal* utility functions and will interpret (7) as a warrant for *both* kinds of interpersonal comparison (cf. Harsanyi, 1977).

Limits to interpersonal comparisons

It seems to me that economists and philosophers influenced by *logical positivism* have greatly exaggerated the difficulties we face in making interpersonal utility comparisons with respect to the utilities and the disutilities that people derive from ordinary commodities and, more generally, from the ordinary pleasures and calamities of human life. (A very influential opponent of the possibility of meaningful interpersonal utility comparisons has been Robbins, 1932.) But when we face the problem of judging the utilities and the disutilities that other people derive from various *cultural* activities, we do seem to run into very real, and sometimes perhaps even unsurmountable, difficulties. For example, suppose I observe a group of people who claim to derive great aesthetic enjoyment from a very esoteric form of abstract art, which does not have the slightest appeal to me in spite of my best efforts to understand it. Then, there may be no way for me to decide whether the admirers of this art form *really* derive very great and genuine enjoyment from it, or merely *deceive themselves* by claiming that they do. Maybe in such cases interpersonal comparisons of utility do reach unsurmountable obstacles. But, fortunately, very few of our personal moral decisions and of our public political decisions depend on such exceptionally difficult interpersonal comparisons of utility. (References additional to those listed below will be found in Hammond, 1977 and in Suppes and Winet, 1955.)

See Also

- interdependent preferences
- interpersonal utility comparisons (new developments)
- Pigou, Arthur Cecil
- value judgements
- welfare economics

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*. 2nd edn, New York: Wiley, 1963.

Arrow, K.J. 1978. Extended sympathy and the possibility of social choice. *Philosophia* 7, 223–37.

Hammond, P.J. 1977. Dual interpersonal comparisons of utility and the welfare economics of income distribution. *Journal of Public Economics* 7, 51–71.

Harsanyi, J.C. 1955. Cardinal utility, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21. Reprinted as ch. 2 of Harsanyi (1977).

Harsanyi, J.C. 1975. Can the maximum principle serve as a basis for morality? A critique of John Rawls' theory. *American Political Science Review* 69, 594–606. Reprinted as ch. 4 of Harsanyi (1977).

Harsanyi, J.C. 1976. *Essays on Ethics, Social Behavior and Scientific Explanation*. Dordrecht: Reidel.

Harsanyi, J.C. 1977. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Harsanyi, J.C. 1984. Von Neumann–Morgenstern utilities, risk taking, and welfare. In *Arrow and the Ascent of Modern Economic Theory*, ed. G.R. Feiwel, New York: New York University Press, 545–58.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Suppes, P. and Winet, M. 1955. An axiomatization of utility based on the notion of utility differences. *Management Science* 1, 259–70.

Von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*. 2nd edn, Princeton: Princeton University Press.

How to cite this article

Harsanyi, John C. "interpersonal utility comparisons." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000196> doi:10.1057/9780230226203.0841

The New Palgrave Dictionary of Economics Online

intertemporal choice

Christopher F. Chabris, David I. Laibson and Jonathon P. Schuldt
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Decisions that have consequences in multiple time periods are intertemporal choices. Individuals typically discount delayed rewards much more than can be explained by mortality effects. The most common discount function is exponential in form, but hyperbolic and quasi-hyperbolic functions seem to explain empirical data better. Individual discount rates may be measured in a variety of ways, subject to important methodological caveats. Higher discount rates are empirically associated with a variety of substance abuse and impulsive conditions, including smoking, alcoholism, cocaine and heroin use, gambling, and risky health behaviours. By contrast, low discount rates may be associated with high cognitive ability.

Keywords

addiction; discount factor; discount rate; discounted utility; dynamic consistency; dynamic inconsistency; felicity; generalized hyperbolas; impatience; instantaneous utility; intertemporal choice; mortality; naive vs. sophisticated; neuroeconomics; normative economics; positive economics; preference reversal,; revealed preference; salience; time preference

Article

Models of intertemporal choice

Most choices require decision-makers to trade-off costs and benefits at different points in time. Decisions with consequences in multiple time periods are referred to as intertemporal choices. Decisions about savings, work effort, education, nutrition, exercise, and health care are all intertemporal choices.
The theory of discounted utility is the most widely used framework for analysing intertemporal choices. This framework has been used to *describe* actual behaviour (positive economics) and it has been used to *prescribe* socially optimal behaviour (normative economics).
Descriptive discounting models capture the property that most economic agents prefer current rewards to delayed rewards of similar magnitude. Such time preferences have been ascribed to a combination of mortality effects, impatience effects, and salience effects. However, mortality effects alone cannot explain time preferences, since mortality rates for young and middle-aged adults are at least 100 times too small to generate observed discounting patterns.
Normative intertemporal choice models divide into two approaches. The first approach accepts discounting as a valid normative construct, using revealed preference as a guiding principle. The second approach asserts that discounting is a normative mistake (except for a minor adjustment for mortality discounting). The second approach adopts zero discounting (or near-zero discounting) as the normative benchmark.
The most widely used discounting model assumes that total utility can be decomposed into a weighted sum – or weighted integral – of utility flows in each period of time (Ramsey, 1928):

$$U_t = \sum_{\tau=0}^{T-t} D(\tau) \cdot u_{t+\tau}.$$

In this representation: U_t is total utility from the perspective of the current period, t; T is the last period of life (which could be infinity for an intergenerational model); $u_{t+\tau}$ is flow utility in period $t+\tau$ ($u_{t+\tau}$ is sometimes referred to as felicity or as instantaneous utility); and $D(\tau)$ is the discount function. If delaying a reward reduces its value, then the discount function weakly *declines* as the delay, τ, *increases*:

$$D'(\tau) \leq 0.$$

Economists normalize $D(0)$ to 1. Economists assume that increasing felicity, $u_{t+\tau}$, weakly increases total utility, U_t. Combining all of these assumptions implies,

$$1 = D(0) \geq D(\tau) \geq D(\tau') \geq 0,$$

where $0 < \tau < \tau'$.
Time preferences are often summarized by the rate at which the discount function declines, $\rho(\tau)$. For differentiable discount functions, the discount rate is defined as

$$\rho(\tau) \equiv -\frac{D'(\tau)}{D(\tau)}.$$

(See Laibson, 2003, for the formulae for non-differentiable discount functions.) The higher the discount rate the greater the preference for immediate rewards over delayed rewards.
The discount factor is the inverse of the continuously compounded discount rate $\rho(\tau)$. So the discount factor is defined as

$$f(\tau) = \lim_{\Delta \to 0} \left(\frac{1}{1 + \rho(\tau)\Delta} \right)^{1/\Delta} = e^{-\rho(\tau)}.$$

The lower the discount factor the greater the preference for immediate rewards over delayed rewards.
The most commonly used discount function is the exponential discount function:

$$D(\tau) = \delta^{\tau},$$

with $0< \delta <1$. For the exponential discount function, the discount rate is independent of the horizon, τ. Specifically, the discount rate is $-\ln(\delta)$ and the discount factor is δ. Figure 1.

Figure 1
Three calibrated discount functions

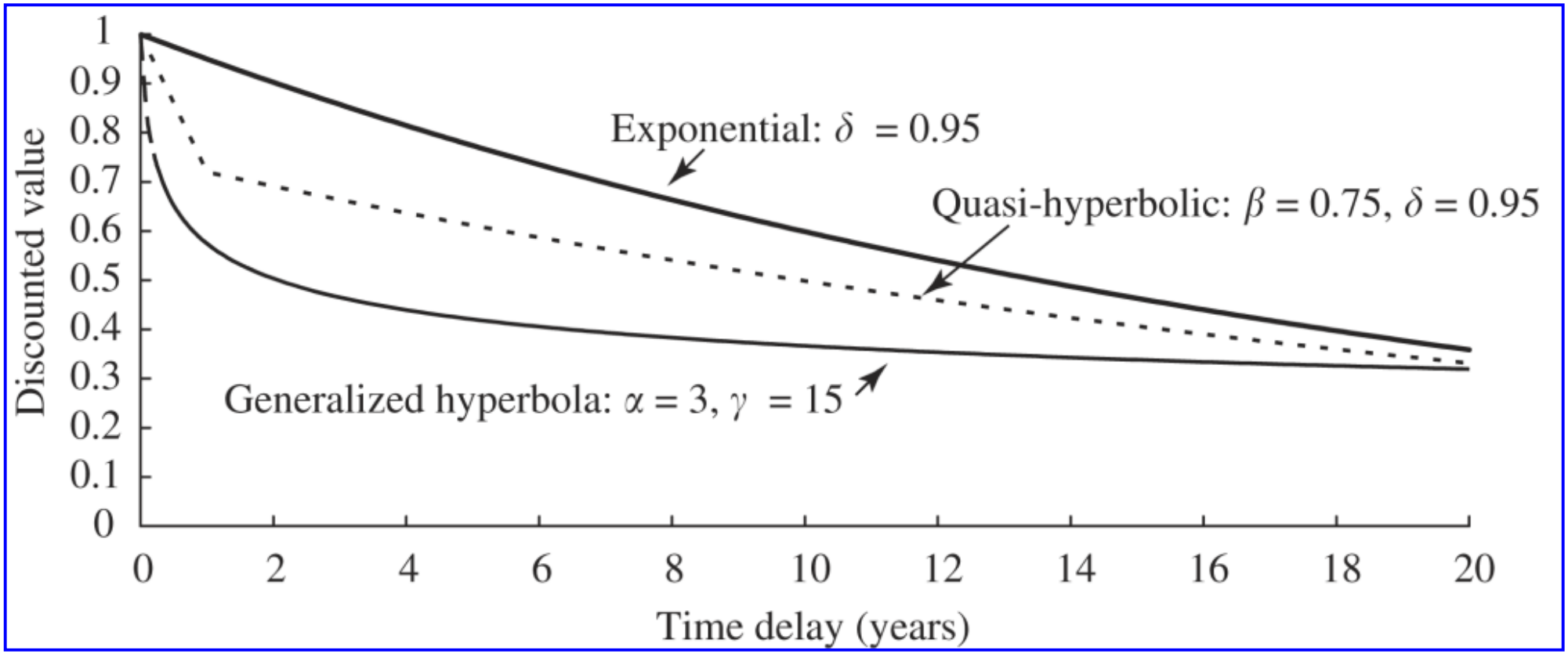

The exponential discount function also has the property of dynamic consistency: preferences held at one point in time do not change with the passage of time (unless new information arrives). For example, consider the following investment opportunity: pay a utility cost of C at date t=2 to reap a utility benefit of B at date t=3. Suppose that this project is viewed from date t=1 and judged to be worth pursuing. Hence, $-\delta C+\delta^2 B>0$. Imagine that a period of time passes, and the agent reconsiders the project from the perspective of date t=2. Now the project is still worth pursuing, since $-C+\delta B>0$. To prove that this is true, note that the new expression is equal to the old expression multiplied by $1/\delta$. Hence, the t=1 preference to complete the project is preserved at date t=2. The exponential discount function is the *only* discount function that generates dynamically consistent preferences.

Despite its many appealing properties, the exponential discount function fails to match several empirical regularities. Most importantly, a large body of research has found that measured discount functions decline at a higher rate in the short run than in the long run. In other words, people appear to be more impatient when they make short-run trade-offs – today vs. tomorrow – than when they make long-run trade-offs – day 100 vs. day 101. This property has led psychologists (Herrnstein, 1961; Ainslie, 1992; Loewenstein and Prelec, 1992) to adopt discount functions in the family of generalized hyperbolas:

$$D(\tau) = (1 + \alpha\tau)^{-\gamma/\alpha}.$$

Such discount functions have the property that the discount rate is higher in the short run than in the long run. Particular attention has been paid to the case in which $\gamma = \alpha$, implying

that $D(\tau) = (1 + \alpha\tau)^{-1}$.

Starting with Strotz (1956), economists have also studied alternatives to exponential discount functions. The majority of economic research has studied the quasi-hyperbolic discount function, which is usually defined in discrete time:

$$D(\tau) = \begin{Bmatrix} 1 & \text{if } \tau = 0 \\ \beta \cdot \delta^{\tau} & \text{if } \tau = 1, 2, 3, \ldots \end{Bmatrix}.$$

This discount function was first used by Phelps and Pollak (1968) to study intergenerational discounting. Laibson (1997) subsequently applied this discount function to intra-personal decision problems. When $0 < \beta < 1$ and $0 < \delta < 1$ the quasi-hyperbolic discount function has a high short-run discount rate and a relatively low long-run discount rate. The quasi-hyperbolic discount function nests the exponential discount function as a special case ($\beta = 1$). Quasi-hyperbolic time preferences are also referred to as ‘present-biased’ and ‘quasi-geometric’.

Like other non-exponential discount functions, the quasi-hyperbolic discount function implies that intertemporal preferences are not dynamically consistent. In other words, the passage of time may change an agent's preferences, implying that preferences are dynamically inconsistent. To illustrate this phenomenon, consider an investment project with a cost of 6 at date t=2 and a delayed benefit of 8 at date t=3. If $\beta = 1/2$ and $\delta = 1$ (see Akerlof, 1991), this investment is desirable from the perspective of date t=1. The discounted value is positive:

$$\beta(-6 + 8) = \frac{1}{2}(-6 + 8) = 1.$$

However, the project is undesirable from the perspective of date 2. Judging the project from the t=2 perspective, the discounted value is negative:

$$-6 + \beta(8) = -6 + \frac{1}{2}(8) = -2.$$

This is an example of a preference reversal. At date t=1 the agent prefers to do the project at t=2. At date t=2 the agent prefers not to do the project. If economic agents foresee such preference reversals they are said to be sophisticated and if they do not foresee such preference reversals they are said to be naive (Strotz, 1956). O’Donoghue and Rabin (2001) propose a generalized formulation in which agents are partially naive: the agents have an imperfect ability to anticipate their preference reversals.

Many different microfoundations have been proposed to explain the preference patterns captured by the hyperbolic and quasi-hyperbolic discount functions. The most prominent examples include temptation models and dual-brain neuroeconomic models (Bernheim and Rangel, 2004; Gul and Pesendorfer, 2001; McClure et al., 2004; Thaler and Shefrin, 1981). However, both the properties and mechanisms of time preferences remain in dispute.

Individual differences in measured discount rates

Numerous methods have been used to measure discount functions. The most common technique poses a series of questions, each of which asks the subject to choose between a sooner, smaller reward and a later, larger reward. Usually the sooner, smaller reward is an *immediate* reward. The sooner and later rewards are denominated in the same goods, typically amounts of money or other items of value. For example: ‘Would you rather have $69 today, or $85 in 91 days?’ The subject's discount rate is inferred by fitting one or more of the discount functions described in the previous section to the subject choices. Most studies assume that the utility function is linear in consumption. Most studies also assume no

intertemporal fungibility – the reward is assumed to be consumed the moment it is received. Many factors may confound the analysis in such studies, leading numerous researchers to express scepticism about the conclusions generated by laboratory studies. Table 1 provides a summary of such critiques.

Potential confounds that may arise in attempts to measure discount rates in laboratory studies

Factor	Description
Unreliability of future rewards	A subject may prefer an earlier reward because the subject thinks she is unlikely to actually receive the later reward. For example, the subject may perceive an experimenter as unreliable.
Transaction costs	A subject may prefer an immediate reward because it is paid in cash, whereas the delayed reward is paid in a form that generates additional transaction costs. For example, a delayed reward may need to be collected, or it may arrive in the form of a cheque that needs to be cashed.
Hypothetical rewards	A subject may not reveal her true preferences if she is asked hypothetical questions instead of being asked to make choices with real consequences. However, researchers who have directly compared real and hypothetical rewards have concluded that this difference does not arise in practice (Johnson and Bickel, 2002).
Investment versus consumption	Some subjects may interpret a choice in a discounting experiment as an *investment* decision and not a decision about the timing of consumption. For example, a subject might reason that a later, larger reward is superior to a sooner, smaller reward as long as the return for waiting is higher than the return available in financial markets.
Consumption versus receipt	Rewards, especially large ones, may not be consumed at the time they are received. For example, a $500 reward is likely to produce a stream of higher consumption, not a lump of consumption at the date of receipt. Such effects may explain why large-stake experiments are associated with less measured discounting than small-stake experiments
Curvature of utility function	A subject may prefer a sooner, smaller reward to a later, larger reward if the subject expects to receive other sources of income at that later date. In general, a reward may be worth less if it is received during a period of relative prosperity.
Framing effects	The menu of choices or the set of questions may influence the subject's choices. For example, if choices between $1.00 now and delayed amounts ranging between $1.01 and $1.50 were offered, subjects may switch preference from early to later rewards at an interior threshold – for example $1.30. However, if choices between $1.00 and delayed amounts ranging between $1.51 and $2.00 were offered, the switch might happen at a much higher threshold – for example $1.70 – implying a much higher discount rate.
Demand characteristics	Procedures for estimating discount rates may bias subject responses by implicitly guiding their choices. For example, the phrasing of an experimental question can imply that a particular choice is the right or desired answer (from the perspective of the experimenter).

Discount functions may also be inferred from field behaviour, such as consumption, savings, asset allocation, and voluntary adoption of forced-savings technologies (Angeletos et al., 2001; Shapiro, 2005; Ashraf, Karlan and Yin, 2006). However, field studies are also vulnerable to methodological critiques. There is currently no methodological gold standard for measuring discount functions.

Existing attempts to measure discount functions have reached seemingly conflicting conclusions (Frederick, Lowenstein and O'Donoghue, 2003). However, the fact that different methods and samples yield different estimates does not rule out consistent individual differences. Dozens of empirical studies have explored the relationship between individuals' estimated discount rates and a variety of behaviours and traits. A significant subset of this literature has focused on delay discounting and behaviour in clinical populations, most notably drug users, gamblers, and those with other impulsivity-linked psychiatric disorders (see Reynolds, 2006, for a review). Other work has explored the relationship between discounting and traits such as age and cognitive ability. Table 2 summarizes representative studies.

Representative empirical studies linking estimated discount rates for monetary rewards to various individual behaviours and traits

Variable	Study	*N*	Discount rate findings
Nicotine	Bickel, Odum and Madden (1999)*	66	Current smokers>never-smokers and ex-smokers
Alcohol	Bjork et al. (2004)	160	Abstinent alcohol-dependent subjects>controls
Cocaine	Coffey et al. (2003)*	25	Crack-dependent subjects>matched controls[a]
Heroin	Kirby, Petry and Bickel (1999)	116	Heroin addicts>age-matched controls
Gambling	Petry (2001b)*	86	Pathological•gamblers[b]>controls

Risky Behaviour	Odum et al. (2000)*	32	Heroin addicts agreeing to share needle in a hypothetical scenario>non-agreeing
Age	Green, Fry and Myerson (1994)*	36	Children>young adults>older adults
Psychiatric disorders	Crean, de Wit and Richards (2000)	24	'High risk' patients[c]>'low risk' patients
Cognitive ability	Benjamin, Brown and Shapiro(2006)	92	Low scorers on standardized mathematics test>high scorers

Notes: *N*=total number of participants in study.

* These studies used hypothetical rewards; others used real rewards.

[a] Results based on those choices falling within the delay range of 1 week to 25 years. Overall analyses including shorter delays (5•minutes to 5 days) also revealed the same effect, but with smaller magnitude.

[b] Gamblers with comorbid substance abuse disorders showed a greater effect than gamblers without such disorders.

[c] 'High risk' patients were those diagnosed with disorders carrying high risk for impulsive behaviour, according to *DSM-IV* criteria, such as patients with borderline personality disorder, bipolar disorder, and substance abuse disorders.

Smoking. A number of investigations have explored the relationship between cigarette smoking and discounting, together providing strong evidence that cigarette smoking is associated with higher discount rates (Baker, Johnson and Bickel, 2003; Bickel, Odum and Madden, 1999; Kirby and Petry, 2004; Mitchell, 1999; Ohmura, Takahashi and Kitamura, 2005; Reynolds et al., 2004).

Excessive alcohol consumption. While the association with alcoholism has received relatively little attention, the available data suggest that problematic drinking is associated with higher discount rates. Heavy drinkers have higher discount rates than controls (Vuchinich and Simpson, 1998), active alcoholics discount rewards more than abstinent alcoholics, who in turn discount at higher rates than controls (Petry, 2001a), and detoxified alcohol-dependents have higher discount rates than controls (Bjork et al., 2004).

Illicit drug use. Recent studies document a positive association between discount rates and drug use for a variety of illicit drugs, most notably cocaine, crack-cocaine, heroin and amphetamines (Petry, 2003; Coffey et al., 2003; Bretteville-Jensen, 1999; Kirby and Petry, 2004).

Gambling. Pathological gamblers have higher discount rates than controls, both in the laboratory (Petry, 2001b) and in a more natural setting (Dixon, Marley and Jacobs, 2003), and among a population of gambling and non-gambling substance abusers (Petry and Casarella, 1999). Moreover, Alessi and Petry (2003) report a significant, positive relationship between a gambling severity measure and the discount rate within a sample of problem gamblers. Petry (2001b) finds that gambling frequency during the previous three months correlates positively with discount rate.

Age. Patience appears to increase across the lifespan, with the young showing markedly less patience than middle-aged and older adults (Green, Fry and Myerson, 1994; Green et al. 1996; Green, Myerson and Ostazewski, 1999). Read and Read (2004) report that older adults (mean age=75) are the most patient age group when delay horizons are only one year. However, this study also finds that older adults are the *least* patient group when delay horizons are from three to ten years. This reversal probably reflects the fact that 75-year-olds face significant mortality/disability risk at horizons of three to ten years.

Cognitive ability. Kirby, Winston and Santiesteban (2005) report that discount rates are correlated negatively with grade point average in two college samples. Benjamin, Brown and Shapiro (2006) find an inverse relationship between individual discount rates and standardized (mathematics) test scores for Chilean high school students. Silva and Gross (2004) show that students scoring in the top third of their introductory psychology course have lower discount rates than those scoring in the middle and lower thirds. Frederick (2005) shows that participants scoring high on a 'cognitive reflection' problem-solving task demonstrate more patient intertemporal choices (for a variety of rewards) than those scoring low. Finally, in a sample of smokers, Jaroni et al. (2004) report that participants who did not attend college had higher discount rates than those attending at least some college.

All of these empirical regularities are consistent with the neuroeconomic hypothesis that prefrontal cortex is essential for patient (forward-looking) decision-making (McClure et al., 2004). This area of the brain is slow to mature, is critical for general cognitive ability (Chabris, 2007), and is often found to be dysfunctional in addictive and other psychiatric disorders.

More research is required to clarify the cognitive and neurobiological bases of intertemporal preferences. Future research should evaluate the usefulness of measured discount functions in predicting real-world economic decisions (Ashraf, Karlan and Yin, 2006). Finally, ongoing research should improve the available methods for measuring intertemporal preferences.

See Also

- time preference

Bibliography

Ainslie, G. 1992. *Picoeconomics*. New York: Cambridge University Press.

Akerlof, G.A. 1991. Procrastination and obedience. *American Economic Review* 81, 1–19.

Alessi, S.M. and Petry, N.M. 2003. Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behavioural Processes* 64, 345–54.

Angeletos, G.-M., Laibson, D.I., Repetto, A., Tobacman, J. and Weinberg, S. 2001. The hyperbolic consumption model: calibration, simulation, and empirical evaluation. *Journal of Economic Perspectives* 15(3), 47–68.

Ashraf, N., Karlan, D.S. and Yin, W. 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. *Quarterly Journal of Economics* 121, 673–97.

Baker, F., Johnson, M.W. and Bickel, W.K. 2003. Delay discounting in current and never-before cigarette smokers: similarities and differences across commodity, sign, and magnitude. *Journal of Abnormal Psychology* 112, 382–92.

Benjamin, D.J., Brown, S.A. and Shapiro, J.M. 2006. Who is ‘behavioral’? Cognitive ability and anomalous preferences. Unpublished manuscript, University of Michigan.

Bernheim, B.D. and Rangel, A. 2004. Addiction and cue-triggered decision processes. *American Economic Review* 94, 1558–90.

Bickel, W.K., Odum, A.L. and Madden, G.J. 1999. Impulsivity and cigarette smoking: delay discounting in current, never, and ex-smokers. *Psychopharmacology* 146, 447–54.

Bjork, J.M., Hommer, D.W., Grant, S.J. and Danube, C. 2004. Impulsivity in abstinent alcohol-dependent patients: relation to control subjects and type 1-/type 2-like traits. *Alcohol* 34, 133–50.

Bornovalova, M.A., Daughters, S.B., Hernandez, G.D., Richards, J.B. et al. 2005. Differences in impulsivity and risk-taking propensity between primary users of crack cocaine and primary users of heroin in a residential substance-use program. *Experimental and Clinical Psychopharmacology* 13, 311–8.

Bretteville-Jensen, A.L. 1999. Addiction and discounting. *Journal of Health Economics* 18, 393–407.

Chabris, C.F. 2007. Cognitive and neurobiological mechanisms of the law of general intelligence. In *Integrating the Mind*, ed. M.J. Roberts. Hove, UK: Psychology Press.

Coffey, S.F., Gudleski, G.D., Saladin, M.E. and Brady, K.T. 2003. Impulsivity and rapid discounting of delayed hypothetical rewards in cocaine-dependent individuals. *Experimental and Clinical Psychopharmacology* 11, 18–25.

Crean, J.P., de Wit, H. and Richards, J.B. 2000. Reward discounting as a measure of impulsive behavior in a psychiatric outpatient population. *Experimental and Clinical Psychopharmacology* 8, 155–62.

Dixon, M.R., Marley, J. and Jacobs, E.A. 2003. Delay discounting by pathological gamblers. *Journal of Applied Behavior Analysis* 36(4), 449–58.

Field, M., Santarcangelo, M., Sumnall, H., Goudie, A. et al. 2006. Delay discounting and the behavioural economics of cigarette purchases in smokers: the effects of nicotine

deprivation. *Psychopharmacology* 186, 255–63.

Frederick, S. 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19(4), 24–42.

Frederick, S., Loewenstein, G. and O'Donoghue, T. 2003. Time discounting and time preference: a critical review. In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, ed. G. Loewenstein, D. Read and R. Baumeister. New York: Sage.

Giordano, L.A., Bickel, W.K., Loewenstein, G., Jacobs, E.A. et al. 2002. Mild opioid deprivation increases the degree that opioid-dependent outpatients discount delayed heroin and money. *Psychopharmacology* 163, 174–82.

Green, L., Fry, A.F. and Myerson, J. 1994. Discounting of delayed rewards: a life-span comparison. *Psychological Science* 5, 33–7.

Green, L., Myerson, J. and Ostazewski, P. 1999. Discounting of delayed rewards across the life span: age differences in individual discounting functions. *Behavioural Processes* 46, 89–96.

Green, L., Myerson, J., Lichtman, D., Rosen, S. and Fry, A. 1996. Temporal discounting in choice between delayed rewards: the role of age and income. *Psychology and Ageing* 11, 79–84.

Gul, F. and Pesendorfer, W. 2001. Temptation and self-control. *Econometrica* 69, 1403–35.

Herrnstein, R.J. 1961. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior* 4, 267–72.

Hinson, J.M., Jameson, T.L. and Whitney, P. 2003. Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29, 298–306.

Holt, D.D., Green, L. and Myerson, J. 2003. Is discounting impulsive? Evidence from temporal and probability discounting in gambling and non-gambling college students. *Behavioural Processes* 64, 355–67.

Jaroni, J.L., Wright, S.M., Lerman, C. and Epstein, L.H. 2004. Relationship between education and delay discounting in smokers. *Addictive Behaviors* 29, 1171–5.

Johnson, M.W. and Bickel, W.K. 2002. Within-subject comparison of real and hypothetical money rewards in delay discounting. *Journal of the Experimental Analysis of Behavior* 77, 129–46.

Kirby, K.N. and Petry, N.M. 2004. Heroin and cocaine abusers have higher discount rates for delayed rewards than alcoholics or non-drug-using controls. *Addiction* 99, 461–71.

Kirby, K.N., Petry, N.M. and Bickel, W.K. 1999. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of Experimental Psychology: General* 128, 78–87.

Kirby, K.N., Winston, G.C. and Santiesteban, M. 2005. Impatience and grades: delay-discount rates correlate negatively with college GPA. *Learning and Individual Differences* 15, 213–22.

Laibson, D. 2003. Intertemporal decision making. In *Encyclopedia of Cognitive Science*. London: Nature Publishing Group.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112, 443–77.

Loewenstein, G. and Prelec, D. 1992. Anomalies in intertemporal choice: evidence and an interpretation. *Quarterly Journal of Economics* 107, 573–97.

Madden, G.J., Begotka, A.M., Raiff, B.R. and Kastern, L.L. 2003. Delay discounting of real and hypothetical rewards. *Experimental and Clinical Psychopharmacology* 11, 139–45.

McClure, S.M., Laibson, D.I., Loewenstein, G. and Cohen, J.D. 2004. Separate neural systems value immediate and delayed monetary rewards. *Science* 306, 503–7.

Mitchell, S.H. 1999. Measures of impulsivity in cigarette smokers and non-smokers. *Psychopharmacology* 146, 455–64.

Mitchell, S.H. 2004. Effects of short-term nicotine deprivation on decision-making: delay, uncertainty, and effort discounting. *Nicotine and Tobacco Research* 6, 819–28.

O'Donoghue, T. and Rabin, M. 2001. Choice and procrastination. *Quarterly Journal of Economics* 116, 121–60.

Odum, A.L., Madden, G.J., Badger, G.J. and Bickel, W.K. 2000. Needle sharing in opioid-dependent outpatients: psychological processes underlying risk. *Drug and Alcohol Dependence* 60, 259–66.

Ohmura, Y., Takahashi, T. and Kitamura, N. 2005. Discounting delayed and probabilistic monetary gains and losses by smokers of cigarettes. *Psychopharmacology* 182, 508–15.

Ortner, C.N.M., MacDonald, T.K. and Olmstead, M.C. 2003. Alcohol intoxication reduces impulsivity in the delay-discounting paradigm. *Alcohol and Alcoholism* 38, 151–6.

Petry, N.M. 2001a. Delay discounting of money and alcohol in actively using alcoholics, currently abstinent alcoholics, and controls. *Psychopharmacology* 154, 243–50.

Petry, N.M. 2001b. Pathological gamblers, with and without substance use disorders, discount delayed rewards at high rates. *Journal of Abnormal Psychology* 3, 482–7.

Petry, N.M. 2003. Discounting of money, health, and freedom in substance abusers and controls. *Drug and Alcohol Dependence* 71, 133–41.

Petry, N.M. and Casarella, T. 1999. Excessive discounting of delayed rewards in substance abusers with gambling problems. *Drug and Alcohol Dependence* 56, 25–32.

Phelps, E.S. and Pollak, R.A. 1968. On second-best national saving and game-equilibrium growth. *Review of Economic Studies* 35, 185–99.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–9.

Read, D. and Read, N.L. 2004. Time discounting over the lifespan. *Organizational Behavior and Human Decision Processes* 94, 22–32.

Reynolds, B. 2006. A review of delay-discounting research with humans: relations to drug use and gambling. *Behavioural Pharmacology* 17, 651–67.

Reynolds, B., Richards, J.B., Horn, K. and Karraker, K. 2004. Delay discounting and probability discounting as related to cigarette smoking status in adults. *Behavioral Processes* 65, 35–42.

Shapiro, J.M. 2005. Is there a daily discount rate? Evidence from the food stamp nutrition cycle. *Journal of Public Economics* 89, 303–25.

Silva, F.J. and Gross, T.F. 2004. The rich get richer: students' discounting of hypothetical delayed rewards and real effortful extra credit. *Psychonomic Bulletin & Review* 11, 1124–8.

Strotz, R.H. 1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23, 165–80.

Thaler, R.H. and Shefrin, H.M. 1981. An economic theory of self-control. *Journal of Political Economy* 89, 392.

Vuchinich, R.E. and Simpson, C.A. 1998. Hyperbolic temporal discounting in social drinkers and problem drinkers. *Experimental and Clinical Psychopharmacology* 6, 292–305.

How to cite this article

Chabris, Christopher F., David I. Laibson and Jonathon P. Schuldt. "intertemporal choice." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000365> doi:10.1057/9780230226203.0843

The New Palgrave Dictionary of Economics Online

intertemporal equilibrium and efficiency

E. Malinvaud

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A clear formalization of intertemporal equilibrium not only aids the fundamental conceptualization of economic activity but should also lead to comparative statics properties, which, dealing with intertemporal equilibria, have also been called ‘comparative dynamics properties’. Particular importance has been given to the question of knowing how the interest rate changes from one stationary equilibrium to another when some specific change is being brought to its exogenous determinants. The theory of the optimum allocation of resources can likewise be transposed to the intertemporal framework. Applications of these properties may give insights on the evolution of prices through time.

Keywords

Austrian economics; capital accumulation; capital theory; comparative statics; competitive temporary equilibria; decision criteria; exhaustible resources; golden rule; impatience; implicit contracts; information aggregation; information exchange; intertemporal competitive equilibrium; intertemporal decisions; intertemporal efficiency; intertemporal equilibrium; irreversibilities; joint production; Keynesian consumption function; Labour market contracts; overcapitalization; overlapping generations models; Pareto efficiency; period of production; production possibility set; productivity of investment; proportional growth; rate of interest; rationing; reswitching of technique; Samuelson–Leontief technology; shadow discount rate; stationary equilibrium; *temporary equilibrium* ; trade cycle; unbounded horizons; uncertainty

Article

People, corporations and governments take decisions for the future. What kind of consistency exists between these decisions? What role does the price system play in this respect? Is the resulting evolution efficient? How can economic organization be improved in order to permit a more satisfactory growth? Confronted with such huge questions, economists have often answered quickly. Even when attention is

limited to formal theory, which this article exclusively considers, many statements can be found which, taken as valid for a time, were later disproved. They had been obtained on special models and too easily given a broad validity. Indeed, the preliminary step should have been to find a general formal representation of economic activity through time, but this step was not given sufficient attention until the late 19th century (Böhm-Bawerk, 1888; Fisher, 1907). The central model with reference to which the whole theory can be built and developed clearly emerged only in the 1950s.

A survey on the subject must then start from first principles and note which major features of reality are still today neglected in main-stream theory. The significance of the most far-reaching results and the importance of some big question marks will then have to be assessed.

Intertemporal decisions

Households save for future consumption, employees work overtime so as to have enough to enjoy their vacation, students strive to get a diploma so as to hold good jobs later, parents want to leave bequest to their children. Firms produce to inventories in the expectation of future sales, recruit and train staff that will later improve their competitiveness, install equipment to be used for many years, build new factories.

The main theories dealing with intertemporal economic problems see such decisions as parts of plans that the relevant agents make for all their future activities. Any household, for instance, is assumed not only to decide its present supply of labour and demand for goods, but also simultaneously to choose its plan for the labour to be later supplied and the goods to be later consumed, and this up to the end of its existence.

The notion of this plan can in principle be made richer by taking *uncertainties* into account; the future decisions are then conditional on events to be later observed, but they are already specified for all conceivable combinations of events. In principle again the structure of the plan must then depend on the structure of the *information* that the agent will receive. In the main intertemporal theories these complications coming from uncertainties and information are, however, neglected, so that the concept of a plan does not appear to be unduly abstract. When the relevance of these theories is assessed, one has to wonder about the consequences of the simplification, as will be seen in the sequel.

Analysis of intertemporal behaviour can adopt the familiar approach: the constraints to which the plan is subject and the objectives that it strives to achieve must be identified; then the optimization problem is solved. The purest of all theories simply transpose the classical analysis of consumer and producer behaviour (Debreu, 1959). They assume the existence of a full system of discounted prices, with one such price for each commodity at each present or future date, a price at which agents will be able to buy or sell as much of this commodity as they may wish. They then directly reinterpret as follows the constraints and objectives that static atemporal theories made familiar.

As between the many plans that he can think of, a consumer is assumed to have a system of preferences that is often conveniently represented by a utility function, whose argument is a consumption vector with as many components as there are commodities and dates. A budget constraint requires that the discounted value of the consumption vector does not exceed a given amount, the consumer initial wealth. The chosen plan maximizes the utility function subject to the budget constraint. It then follows that the consumption of the various commodities (and the supply of labour) depend on what are the discounted prices and the initial wealth. The present saving of the consumer may be said to be equal to

the interest income earned on his initial wealth *minus* the value of his present consumption (labour income appears negatively in this value). It is immaterial in this theory to know how saving is invested. Hence, the consumption plan and the resulting saving plan are seen as involving the whole future *life cycle* of the consumer (Modigliani and Brumberg, 1954).

The plan of a producer is subject only to the constraints that technology imposes. The producer acts as a price taker. His objective is to maximize the discounted value of the plan. It follows that demand for inputs and supply of outputs are functions of the discounted prices. The balance between the value of present outputs and present inputs gives the financial surplus if positive or requirement if negative; this is subject to no direct constraint.

Such a theory of consumer and producer behaviour does not claim to apply to all problems concerning this behaviour. Clearly, analysis of the firm in particular must usually go far beyond the stylized description given above, even simply when investment behaviour is being studied (Nickell, 1978). But the theory is supposed to be appropriate for fitting into the discussion of the broad questions raised by intertemporal equilibrium and efficiency.

Even when it is so circumscribed, the intent cannot be considered as fully achieved. Significant limitations must be kept in mind, since they may forbid application of the theory to some of the problems raised by equilibrium and efficiency over time; indeed, some of these limitations have been the motivation for theoretical developments that will not be discussed at length here, but must be mentioned.

Full knowledge of the system of discounted prices for purchases of sales at all relevant future dates is of course an abstraction. Forward prices exist for only a few basic commodities and a limited horizon. Whereas the interest rates at which one can borrow or lend for more or less long durations are fairly well defined, with non-negligible transaction costs and fiscal interference, however, prices that will apply to future transactions have to be forecast by the agents. The uncertainties that their forecast necessarily contains are neglected. Among the many consequences of this major simplification, one particularly notes that it rules out fundamental problems concerning the characterization of decision criteria of business firms (Drèze, 1982).

Constraints on individual choices are also reduced to a minimum. No consideration is given to quantitative constraints, such as those following from mass unemployment on individuals looking for jobs or from business depression on firms looking for customers. When such constraints are binding, not only must the plans meet them, but also spillover effects from one period to others occur, according to laws that follow from the theory of individual behaviour under rationing (Samuelson, 1947). In particular, consumers willing but unable to borrow are constrained by their current resources, a phenomenon that gives some justification to the Keynesian consumption function relating current consumption to current income.

Neglect of financial constraints may be considered as following from other theoretical simplifications, lack of uncertainty and full knowledge of discounted prices, which rule out insolvency; but it is often particularly restrictive. The role of financial constraints on investment behaviour indeed play a major part in the development of trade cycle theories (Haberler, 1937).

Another notable feature of the theory is the simplicity of the trading relations that it assumes. Consumers and producers buy from ‘the market’ or sell to ‘the market’. A worker need not establish ties with a particular employer, nor a manufacturing firm to a particular supplier of raw material. Actually, intertemporal decisions are often subject to quite significant irreversibilities. Long-term commitments

are frequent for easily understandable reasons, some of which have to do with the specificities that characterize many production processes (for instance, most equipment, once bought, cannot be resold). Long-term contracts are also predominant on the labour market, even though many of their clauses often remain implicit. This feature motivates significant research nowadays, under the heading of ‘implicit contracts’ (Rosen, 1985).

Limited as it is, the classical theory of individual intertemporal decisions is, however, indispensable as a starting point, from which the study of the many complexities of real life can proceed. It has moreover brought to light some quite relevant results, such as the fact that, contrary to common belief, the saving of a household need not be an increasing function of interest rates or that individual choice is bound to exhibit some degree of impatience (Koopmans, 1960).

An intertemporal economy

The theory of general intertemporal equilibrium can also transpose the more familiar static theory. But clearly when so doing it does not go very far; new complications, specific to intertemporal problems, must be faced.

The simple transposition of the general competitive equilibrium assumes the existence of a terminal date, ‘the horizon’, a given set of consumers and producers whose activities end at this date, if not before. They all decide their plans at the initial date, on the basis of a full system of discounted prices, and acting as price takers. Perfect competition is assumed to imply that discounted prices are such that all markets clear; more precisely for a given date and a given commodity, aggregate supply and demand are defined by addition of corresponding individual supplies and demands contained in individual plans, which may then be considered as fully announced; at equilibrium the aggregate supply is precisely equal to aggregate demand, and this applies for any date and commodity. Hence, all individual plans are, from the initial date, mutually consistent for all future dates.

The usefulness of such an abstract equilibrium concept cannot be judged independently of its application, in particularly for the discussion of properties linking discounted prices to the agents’ individual characteristics. Before facing this discussion, it is enlightening to consider how the model can be revised; this was done in three ways.

First, the hypothesis of a full system of markets, one for each date and commodity, has been relaxed and the notion of a *temporary equilibrium* made explicit (Hicks, 1939; Arrow and Hahn, 1971; Grandmont, 1977). Markets then exist only for the exchange of commodities at the (initial) present date, as well as for the loans of one numeraire commodity from the present to the next future date. Thus, present prices and the interest rate of the first period are assumed to be determined by the law of supply and demand, individual plans being made mutually consistent for the initial data. But, in deciding their plans, individual agents have to form anticipations about future prices. Nothing guarantees that these anticipations are correct, so that individual plans will be revised with the passage of time, as actual prices are found to differ from what was expected.

Formal properties of this more realistic model will not be discussed here. Cases can be defined in which anticipations are later realized. It is then possible, but not always necessary when the future is unbounded, that the sequence of temporary equilibria coincides with the equilibrium defined from the hypothesis of a full system of markets. Thus, two sources of difficulty can arise: false anticipations and

on the other hand instability following from the myopic functioning of the market system (Hahn, 1968).
Second, coming back to the case of a full system of markets, one has relaxed the assumption of a finite horizon with a fixed set of agents. The problem of knowing which firms exist has not been considered as specific to the intertemporal models, and has not been discussed thus far in the framework of these models, given that infinitely lived firms have been assumed. But since the initial proposals of Allais (1947) and Samuelson (1958), consumers are more and more assumed to belong to overlapping generations, each generation living only for a finite time. Such a representation of the consumption sector is clearly more appropriate for long-term analysis than the assumption of a given set of consumers living for ever, but it raises new difficulties (Balasko and Shell, 1980–81).
Third, since long-term phenomena are often involved, it has been found natural and convenient to concentrate attention on specifications in which the exogenous conditions of economic activity, such as technology, tastes, size of the population, natural resources, remain the same through time or change in a simple way; for instance, population increasing at a constant rate while technology exhibits constant returns to scale and natural resources are unbounded. Within such specifications one has dealt with the particular case of a stationary equilibrium, or else with equilibria in which production and consumption all increase at the same constant rate, that is, the case of ‘proportional growth’. The analytical usefulness of this assumption of stationarity was at the centre of an important debate on the building of the theory of capital during the 1930s (Knight, 1935; Hayek, 1936). It follows from the simple form that has the price system of a stationary equilibrium: all discounted prices can be computed from the prices of the present commodities using a single interest rate that applies to all future periods of unit duration. ‘The interest rate’ is then unambiguously defined (Malinvaud, 1953).

Any general law?

A clear formalization of intertemporal equilibrium not only serves to aid progress in the fundamental conceptualization of economic activity (hence indirectly in the rigour of the discussions concerning many particular questions) but should also lead to comparative statics properties, which, dealing with intertemporal equilibria, have also been called ‘comparative dynamics properties’. Particular importance has been given to the question of knowing how the interest rate changes from one stationary equilibrium to another when some specific change is being brought to its exogenous determinants.
The study of this question concentrated on a number of conjectures, which turned out to be about as many disappointments for whose who had expected to find rigorous proofs of their general validity. It is now realized that the rate of interest is related in a very complex way to the many exogenous determinants of equilibrium and that changes of relative prices, which are associated with changes of interest, may be responsible for paradoxical effects. A brief survey of this theoretical search, that extended over many years, nevertheless reveals some basic issues.
Does a high preference of individuals for present consumption necessarily imply a high interest rate? The property was often asserted. When first publishing his *Theory of Interest* in 1907, Irving Fisher called it an impatience theory. Only later when he revised the book for the 1930 edition did he add the subtitle ‘as determined by impatience to spend income and opportunity to invest it’, which recognizes the role of the productivity of investment (Samuelson, 1967). Quite significant cases have indeed been found in the overlapping generation model for which changes of impatience leave the interest rate

unchanged (Samuelson, 1958).

Does a decrease of the rate of interest mean a lengthening of the production process? The positive answer was taken for granted, at least as long as technology was given, by many economists and was at the head of the ‘Austrian Theory’ as developed mainly by Böhm-Bawerk (1889) and Hayek (1941).

Actually, description of the production process was usually organized in such a way as to focus on the conjectured property, this being true also with such non-Austrian authors as Wicksell (1901). Final output, available for consumption at some date, was seen as resulting from a number of well-identified primary inputs made at previous dates and having ‘matured’ since then. The notion of an average period of production looked natural; an inverse relationship between this period and the rate of interest was expected. However, it turned out that, even restricting attention to the case of one primary input and one final output, one could not prove the relationship unless a special definition was given to the production period and a special phrasing to the property (Hicks, 1939; 1973). Generalization to many primary inputs, many final outputs and many interdependent production processes raises the fundamental difficulty resulting from induced variations in relative prices; it is quite unlikely that a generalized property could be proved (Sargan, 1955).

A somewhat similar property was expected with another formalization that seems to be much more appropriate for describing technology in modern industry. The property concerns the choice of techniques and the notion that different techniques should be selected at various stages of development, as relative scarcity of the two main factors, labour and capital, changes and the interest rate moves accordingly. Its formal specification actually requires a particular model. The production possibility set is seen as resulting from combination of a number of elementary processes, each one operating at constant returns to scale, with fixed input–output coefficients, and requiring a time just equal to one period. Specifying further this model and applying it to an economy with one primary factor (labour), *n* produced goods and no joint production (the ‘Samuelson–Leontief technology’), one defines a technique as a selection of *n* processes, one for the production of each good.

In this model, given any value of the interest rate, one can determine one technique that is fully appropriate for production, no matter what is the consumption basket. It then seemed natural to conjecture that techniques thus appearing as efficient at different interest rates were ordered from the less capitalistic (high interest) to the most capitalistic ones (low interest). However, this conjecture is not generally valid, even in this special model: as the interest rate progressively declines, one may have to switch at some point away from some technique but have to switch back to it at a later point: this is the case of ‘reswitching of techniques’ (Morishima, 1966).

Is the interest rate systematically smaller when, with a given technology, one shifts from a stationary equilibrium to another one using the same labour input but more productive capital? Again, this looked like a natural property to be stated. Since in a perfect equilibrium with no uncertainty the net rate of profit must be equal to the interest rate, the property was associated with the notion that capital accumulation must depress profit rates.

The property holds in a purely aggregated model with just one produced commodity, used both for consumption and as productive capital (Solow, 1956). The significance of this model for a more general situation was at the heart of hot debates in the late 1950s and early 1960s, the main opponents being located in the two academic cities named Cambridge (Robinson, 1956; Lutz and Hague, 1961). A side issue was whether one could give unambiguous definitions to such aggregate notions as the volume of

productive capital and the marginal productivity of capital. Eventually, both counterexamples and formal analysis of the problem showed that the property was not generally valid (Burmeister and Turnovsky, 1972).

The significance of these various negative theoretical results should of course not be overstated. While reflecting the basic complexity of the relationship between the full system of discounted prices and its determinants, the results do not prove that ‘pathological cases’ are often empirically relevant.

Intertemporal efficiency

In the same way as the classical theory of individual behaviour, the theory of the optimum allocation of resources can be transposed to the intertemporal framework. Pareto efficiency of a ‘programme’ made of a set of individual plans, also called ‘Pareto optimality’, is generalized in an obvious way that need not be spelled out. The two classical duality theorems directly apply as long as the horizon is bounded: the programme resulting from a competitive equilibrium of the type described above is Pareto efficient if no external effect occurs; conversely, under a convexity or atomicity assumption, to any Pareto efficient programme can be associated a set of discounted prices supporting this programme. Properties of this system of prices are similar to those of the competitive price system.

Interesting new applications of these properties may give insights on the evolution of prices through time. In particular it is easily found that, if extraction costs are negligible, the discounted efficiency price of an exploited exhaustible resource is the same for all future dates, which means that the undiscounted price increases at a rate equal to the interest rate of the numeraire (Hotelling, 1931). When forming decisions on the use of exhaustible resources, one should give as much weight to the distant future as to the present; discounting gives no comfort for such decisions.

Theoretical difficulties, however, occur when the more realistic case of an unbounded horizon is being considered. The most relevant of these difficulties concerns the Pareto efficiency of competitive equilibria; efficiency is still proved to hold if the discounted value of the productive capital that exists at date t decreases to zero when one lets t increase to infinity (Malinvaud, 1953); but examples of competitive equilibria that do not fulfil this condition and are not Pareto efficient can be found. Such examples may be characterized as cases of overcapitalization, an excessive capital stock being indefinitely maintained without this ever benefiting consumption.

When attention is limited to stationary equilibria, a negative interest rate reveals lack of efficiency, whereas a positive one implies efficiency (if no external effect exists). Similarly, the interest rate of the price system supporting an efficient proportional growth programme cannot be smaller than the rate of growth (Starrett, 1970). The borderline case of an interest rate equal to the growth rate corresponds to what was called ‘the golden rule’. More precisely, a new notion of optimality has been defined as follows for proportional growth programmes: an optimal programme is feasible and no other feasible programme leads to larger consumptions (that is, a larger consumption of some commodity at some date and no smaller consumption of any commodity at any date). This definition neglects the conditions at the initial date since an ‘optimal’ programme can require a large input of capital at this date, a larger than is required by other Pareto efficient proportional growth programmes. It was proved that a price system exists that supports such an optimal programme and contains an interest rate equal to the rate of growth (Desrousseaux, 1961; Phelps, 1961). This is another case in which discounting does not make

the distant future negligible.

When it is considered in the preceding terms, the theory of intertemporal efficiency has a somewhat unrealistic aspect; or rather it seems to be quite partial in its treatment of the various questions that intertemporal efficiency raises both for planning and for the study of actual economic evolution. Indeed, the restrictions mentioned in the first section of this article are often serious.

For the theory of planning, even restricted to the medium and long terms, for which intertemporal choices are particularly important, problems concerning the gathering and exchange of information should not be neglected. If a system of discounted prices is to be used for supporting consistency of individual decisions with national objectives, its determination must be given very serious consideration. Moreover, planning often aims at correcting handicaps, distortions or market failures preventing economic development. Its long-term achievement then depends on how well it deals with problems that are not considered here but have motivated an important literature, dealing in particular with the determination of the best shadow discount rate to be used in project evaluation (Dasgupta, Marglin and Sen, 1972).

Similarly, for assessing the performance of actual economic systems, one has still to face many questions that again often relate to problems of information. Three of them seem to deserve particular attention. First, the vision of agents exchanging in markets abstracts too much from the complexities of actual contractual arrangements, some of which deal precisely with intertemporal choices; one does not yet clearly see how these complexities react on the behaviour of the full economy, nor even how theory could approach the issue.

Second, the notion of an intertemporal competitive equilibrium should be replaced by that of a sequence of competitive temporary equilibria. It is then known that, even if anticipations are self-fulfilling along this sequence, intertemporal efficiency is not guaranteed; more precisely, the short-sightedness of equilibria seems to increase the likelihood of an overcapitalization of the type exhibited by the theory of the golden rule. This may occur because of too high saving propensities, because of risk aversion or because of oligopolistic market structures (Malinvaud, 1981). But the question of knowing whether and when this likelihood will materialize remains obscure.

Third, the dual assumption of permanent market clearing and permanently equilibrating prices rules out of consideration many issues, such as those arising from variations in the degree of unemployment or in the stimulus given by profitability. A rather common view among supporters of the market system sees these variations as negligible from a long-term perspective, economic evolution being supposed simply to oscillate around the long-term path determined by equilibrium analysis. But critics of the market system and some other economists have the opposite view: economic disequilibria would provide the main clue for an understanding of the comparative growth of nations (Schumpeter, 1934; Beckerman, 1966). Theory remains conspicuously weak with respect to solving this major debate.

See Also

- Arrow–Debreu model of general equilibrium
- general equilibrium
- Ramsey model
- sequence economies

Bibliography

Allais, M. 1947. *Economie et intérêt*. Paris: Imprimerie Nationale.

Arrow, K. and Hahn, F. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Balasko, Y. and Shell, K. 1980–81. The overlapping generations model. *Journal of Economic Theory*. I. The case of pure exchange without money 23(3) (1980), 281–306; II. The case of pure exchange with money 24(1) (1981), 112–42.

Beckerman, W. 1966. The determinants of economic growth. In *Economic Growth in Britain*, ed. P.D. Henderson. London: Weidenfeld & Nicolson.

Böhm-Bawerk, E. von. 1889. *Positive Theories des Kapitales*. Trans. as vol. 2 of *Capital and Interest*, South Holland, IL: Libertarian Press, 1959.

Burmeister, E. and Turnovsky, S.J. 1972. Capital deepening response in an economy with heterogeneous capital goods. *American Economic Review* 62, 842–53.

Dasgupta, P., Marglin, S. and Sen, A. 1972. *Guidelines for Project Evaluation*. New York: UNIDO, United Nations.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New York: Wiley.

Desrousseaux, J. 1961. Expansion stable et taux d'intérêt optimal. *Annales des Mines*, November, 31 and 46. Paris.

Drèze, J. 1982. Decision criteria for business firms. In *Current Developments in the Interface: Economics, Econometrics, Mathematics*, ed. M. Hazewinkel and A. Rinney Khan. Dordrecht: D. Reidel.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan. 2nd edn, 1930.

Grandmont, J.-M. 1977. Temporary general equilibrium theory. *Econometrica* 45, 535–72.

Haberler, G. 1937. *Prosperity and Depression*. Geneva: League of Nations. 3rd enlarged edn, 1941.

Hahn, F. 1968. On warranted growth paths. *Review of Economic Studies* 35, 175–84.

Hayek, F.A. von. 1936. The mythology of capital. *Quarterly Journal of Economics* 50(February), 199–228.

Hayek, F.A. von. 1941. *The Pure Theory of Capital*. London: Routledge & Kegan Paul.

Hicks, J. 1939. *Value and Capital*. Oxford: Clarendon.

Hicks, J. 1973. *Capital and Time*. Oxford: Clarendon.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Knight, F. 1935. The theory of investment once more: Mr. Boulding and the Austrians. *Quarterly Journal of Economics* 50(November), 36–67.

Koopmans, T.C. 1960. Stationary ordinal utility and impatience. *Econometrica* 28(April), 287–309.

Lutz, F. and Hague, D., eds. 1961. *The Theory of Capital*. London: Macmillan.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21, 233–68.

Malinvaud, E. 1962. Efficient capital accumulation: a corrigendum. *Econometrica* 30, 570–73.

Malinvaud, E. 1981. *Théorie macroéconomique*, vol. 1. Paris: Dunod.

Modigliani, F. and Brumberg, R. 1954. Utility analysis and the consumption function: an interpretation of cross-section data. In *Post-Keynesian Economics*, ed. K. Kurihara. New Brunswick, NJ: Rutgers University Press; London: George Allen & Unwin, 1955.

Morishima, M. 1966. Refutation of the nonswitching theorem. *Quarterly Journal of Economics* 80 (November), 520–25.

Nickell, S. 1978. *The Investment Decisions of Firms*. Cambridge: Cambridge University Press.

Phelps, E. 1961. The golden rule of capital accumulation. *American Economic Review* 51, 638–42.

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Rosen, S. 1985. Implicit contracts. *Journal of Economic Literature* 23, 1144–75.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P. 1958. An exact consumption-loan model of interest with or without the social contrivance

of money. *Journal of Political Economy* 66(December), 467–82.

Samuelson, P. 1967. Irving Fisher and the theory of capital. In *Ten Economic Studies in the Tradition of Irving Fisher*, ed. W.J. Fellner et al. New York: John Wiley & Sons.

Sargan, J.D. 1955. The period of production. *Econometrica* 23, 151–65.

Schumpeter, J. 1934. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70 (February), 65–94.

Starrett, D. 1970. The efficiency of competitive programmes. *Econometrica* 38, 704–11.

Wicksell, K. 1901. *Vorlesungen über Nationalökonomie*. Trans. as *Lectures on Political Economy*, London: Routledge and Kegan Paul, 1934, vol. 1.

How to cite this article

Malinvaud, E. "intertemporal equilibrium and efficiency." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000198> doi:10.1057/9780230226203.0844

The New Palgrave Dictionary of Economics Online

intrahousehold welfare

Marcel Fafchamps
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article focuses on the allocation of tasks and consumption within the household. We first discuss the role of the household in the production of various self-consumed goods and services. We then turn to the outcome of bargaining between household members, examining the empirical evidence to date. The last section makes the link between intrahousehold welfare and the matching of spouses in the marriage market.

Keywords

altruism; bargaining; Becker, G.; collective versus unitary models of the household; comparative advantage; consumption decisions; distribution of income and wealth; division of labour; economies of size; household production and public goods; human capital; identification; income-pooling test; inequality; inheritance; intrahousehold welfare; marriage and divorce; marriage markets; paternalism; preferences; pre-nuptial agreements; rotten kid theorem; sharing rule; social norms

Article

There is a voluminous economic literature on intrahousehold issues. This is hardly surprising given the many critical functions that households fulfil. They are the locus of most consumption decisions and human capital investments. By pooling resources, households generate economies of size and shelter members against unemployment and health shocks. Furthermore the formation and dissolution of households play a crucial role in the long-term distribution of income and wealth. Here we focus on intrahousehold welfare which, as Haddad and Kanbur (1990) have shown, is important to our understanding of inequality in general.

Becker was the first economist to become seriously interested in what happens within the household. Becker's contribution, which is nicely summarized in his *Treatise on the Family* (1981), emphasized three things: the organization of production within the household; the way decisions are made within the

household; and the formation of couples. All three have a bearing on intrahousehold welfare.

The household as a production unit

The household is a production and consumption unit, self-providing many services such as food preparation, child care and house chores. In developing countries, households also produce much of their own food and housing and fetch their own fuel and water. Becker (1981) pointed out that the organization of production within the household ought to follow economic principles such as the equalization of marginal returns across activities and the allocation of tasks across household members according to comparative advantage.

These simple observations have far-reaching implications because seemingly small differences between household members can have dramatic consequences. To see why, consider the allocation of wage work and household chores between husband and wife. Assume that the tasks are non-divisible and that the return to education is positive in work outside the home and zero in house chores. It follows that the husband will work outside the home if he is slightly better-educated than his wife. Anticipating this, parents may in turn decide not to invest in daughter education but rather to emphasize learning household chores among girls. This results in a self-fulfilling equilibrium in which women receive less education and are confined to household chores. To the extent that education and independent income affect bargaining within the household, such a traditional division of labour may have dramatic consequences on intrahousehold welfare.

The recent empirical literature has cast some doubt on the efficient organization of production within the household. Using data from West Africa, Udry (1996) showed that households do not equalize returns to labour and organic fertilizer across fields managed by different members. Duflo and Udry (2004) provide similar evidence, showing that household labour resources are not optimally reallocated across activities in response to weather shocks. Fafchamps and Quisumbing (2003) show that comparative advantage alone cannot explain the allocation of tasks within Pakistani households. Their evidence also suggests that most household tasks are easy to learn, contradicting Becker's conjecture that learning-by-doing locks men and women into specific work patterns. Gender differences in career choices and intrahousehold division of labour may reflect different preferences, possibly shaped by social norms, or result from differences in intrahousehold bargaining.

Intrahousehold bargaining

Most consumption takes place within households sharing a common budget. (In some societies, such as the coastal region of West Africa, spouses keep separate finances. However, whenever they both contribute to a household public good, they can be regarded as deciding consumption jointly.) Certain consumption goods are rival in the sense that consumption by one precludes consumption by another. Food is an example of a rival good. Other consumption goods – such as a house – are non-rival: they are consumed jointly by the members of the household. In the context of intrahousehold welfare, non-rival goods are usually referred to as (household) public goods.

When choosing how to allocate a limited budget to various rival and non-rival goods, the household takes into account the preferences of its members. Formally, let x_i denote a vector of rival goods

consumed by individual *i* and let *X* denote household public goods. The household's consumption choices can be represented as the solution to an optimization problem of the form:

$$\max_{\{x_1,\ldots x_N, X\}} \sum_{i=1}^{N} \omega_i U_i(x_i, X) \text{ subject to} \sum_{i=1}^{N} p x_i + qX = y \tag{1}$$

where ω_i is a welfare weight, *N* is the number of household members, *p* and *q* are prices, and *y* is income. Consumption choices depend not only on individual preferences $U_i(.)$ but also on welfare weights ω_i: individuals with large ω_i have more weight in the household's decision and hence achieve a higher individual welfare. Understanding intrahousehold welfare thus boils down to understanding the factors that affect ω_i.

In two seminal contributions, Manser and Brown (1980) and McElroy and Horney (1981) model intrahousehold bargaining as depending on threat points: when negotiating over how to allocate consumption expenditures, spouses can threaten to walk away from the couple. How much welfare they can achieve on their own determines how much bargaining power they have within marriage. Intrahousehold welfare is predicted to be determined by rules determining the devolution of assets upon divorce (including alimony, child support and welfare payments).

Lundberg and Pollak (1993) argue that the threat of divorce is too extreme to be credible in most everyday situations. Non-cooperation within marriage is a more realistic threat. In this case, intrahousehold welfare is expected to depend on the financial autonomy of spouses, such as rules determining who receives welfare payments or whether married women have independent access to credit. Lundberg, Pollak and Wales (1997) for instance, show that consumption of women's and children's clothing increased when the UK transferred a substantial child allowance from husbands to wives. McElroy (1990) provides a useful discussion of various factors thought to affect intrahousehold bargaining.

The empirical literature has explored these ideas in terms of ‘unitary’ versus ‘collective’ models of the household. A household model is said to be unitary if choices do not depend on bargaining power; otherwise it is collective. A household may be unitary for a variety of reasons, for instance because all decisions are taken by the household head, or because all household members have the same preferences over household consumption $\{x_1,\ldots x_N, X\}$. A simple way of testing the unitary model is the income-pooling test: if welfare weights do not depend on bargaining power, consumption choices should depend only on total income, not on bargaining weights. This yields a simple exclusion test that has been widely applied in the literature, often to identify variables affecting intrahousehold bargaining.

Chiappori (1988) has proposed a way of testing the efficiency of the intrahousehold bargaining process. The basic idea is that the solution to optimization problem (1) can be written as a two-step process. The household first decides how much to allocate to household public goods *X* and to the rival expenditures $y_i = px_i$ of each household member, with $pX + \sum_i y_i = y$. Then each member maximizes his or her own

utility U_i subject to $px_i = y_i$. Intrahousehold bargaining only affects how total expenditures are shared among members, that is, it affects only the share of rival expenditures that goes to each member. This observation yields testable restrictions on cross-equation parameters in a demand system. This is called the ‘sharing rule’ approach. Browning et al. (1994), for instance, apply this approach to Canadian couples without children and show that allocation of expenditures on each partner depend on their relative incomes.

Both the sharing rule and the income-pooling tests raise empirical difficulties. One difficulty arises whenever utility is transferable and all household members contribute to a household public good. In this case, Bergstrom (1997) has shown that changes in individual incomes do not affect intrahousehold welfare allocation. The reason is fungibility: reducing the income of a household member simply reduces his or her contribution to the household public good.

Another empirical difficulty is that individual preferences are not directly observable. Hence, in order to identify the effect of bargaining power on household choices, we must assume that different categories of household members have systematically different preferences over joint household consumption. The empirical literature has relied on two types of identification strategies to deal with this issue. The first strategy is to rely on stereotypes, such as ‘men prefer alcohol and cigarettes’ or ‘women care more about children’. This strategy permits identification whenever the stereotype is correct. For instance, it has been shown that, when the bargaining power of the wife increases, the household spends more on child nutrition and schooling (see Bergstrom, 1997, and the references cited therein). Based on this evidence, it has been argued that increasing the bargaining power of women is a way to improve child welfare. Such interpretation is a double-edged sword, however. It also reinforces a stereotype that could be used to argue that, since women care for children, it is acceptable for society to relegate them to a reproductive role. What we need is empirical evidence based on actual preferences, not stereotypes.

The second identification strategy is to focus on individual consumption of rival goods such as food or clothing. While this is a better strategy, it also has problems. Browning et al (1994), for instance, show that households in which the wife earns more spend more on female clothing. They interpret this result as evidence that higher income raises a woman's bargaining power. The problem is that a spouse with a higher income probably occupies a higher job position and needs better clothes to go to work. This may generate a reverse causation between income and clothing expenditures, thereby weakening inference.

Spouses probably derive utility from each other's consumption of rival goods. This point was initially made by Becker (1981), who discusses two possible cases, one in which individuals are altruistic – someone else's *utility* enters their preferences – the other in which they are paternalistic – someone else's *consumption* enters their preferences. An example of paternalistic preferences is when a parent does not want a child to smoke, although the child wishes to. In poor countries, differences in health or nutritional status between spouses have sometimes been interpreted as the result of intrahousehold bargaining (Dercon and Krishnan, 2000). Yet it would be quite foolish for even the most despotic and selfish husband to starve his wife to death as she would be of no use to him once gone. Hence, even such a husband would care about his wife's consumption.

An interesting illustration of how altruism can affect intrahousehold welfare is the so-called rotten kid theorem. In this theorem, Becker (1981) imagines a parent who, for altruistic reasons, transfers money to a child. The child can try to capture part of the household income, for instance by refusing to work or by diverting household resources. Becker shows that, as long as the parent decides the size of the transfer

after capture has taken place, capture only leads to lower household income and hence to a lower transfer. As a result, the child chooses not to capture because doing so ultimately reduces his consumption.

The marriage market

The marriage market is discussed in a separate entry in this dictionary and need not be discussed in detail here. What is important to realize for our purpose is that, if sufficient commitment mechanisms exist, intrahousehold allocation of welfare can be negotiated up front at the time of marriage. For instance, future spouses may anticipate that a wife who earns an independent income has more say in household decisions. As a result, the groom may insist that the bride will never work before agreeing to marry her. Similarly, if devolution of assets upon divorce affects bargaining power, the newlyweds may sign a pre-nuptial agreement that shapes how assets will be divided.

As first pointed out by Lundberg and Pollak (1993), this observation has deep implications regarding policy intervention. If intrahousehold welfare is entirely decided at the time of marriage, then changing the rules applying to married couples affects only those who are already married. Changes to what happens after marriage (for example, devolution of assets upon divorce) have no long-run effect because, once they have been introduced, they are anticipated in the marriage market.

Provided that this reasoning is empirically correct, the policy implication is that the best policy handle to influence intrahousehold welfare is the marriage market itself. The share of household consumption that women can (implicitly or explicitly) negotiate for themselves depends on the assets they bring to marriage. If this is true, helping women then is best achieved in the long run by improving female education and by changing inheritance rules in their favour.

See Also

- Becker, Gary S.
- collective models of the household
- household production and public goods
- marriage and divorce
- marriage markets
- rotten kid theorem

Bibliography

Becker, G. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Behrman, J. 1997. Intrahousehold distribution and the family. In *Handbook of Population and Family Economics*, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Bergstrom, T. 1997. A survey of theories of the family. In *Handbook of Population and Family*

Economics, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Browning, M., Bourguignon, F., Chiappori, P.-A. and Lechene, V. 1994. Income and outcomes: a structural model of intrahousehold allocation. *Journal of Political Economy* 102, 1067–96.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56, 63–90.

Dercon, S. and Krishnan, P. 2000. In sickness and in health: risk-sharing within households in rural Ethiopia. *Journal of Political Economy* 108, 688–727.

Duflo, E. and Udry, C. 2004. Intrahousehold resource allocation in Côte d'Ivoire: social norms, separate accounts and consumption choices. Working Paper No. 10498. Cambridge, MA: NBER.

Fafchamps, M. and Quisumbing, A. 2003. Social roles, human capital, and the intrahousehold division of labor: evidence from Pakistan. *Oxford Economic Papers* 55, 36–80.

Haddad, L. and Kanbur, R. 1990. How serious is the neglect of intra-household inequality? *Economic Journal* 100, 866–81.

Lundberg, S. and Pollak, R. 1993. Separate spheres bargaining and the marriage market. *Journal of Political Economy* 101, 988–1010.

Lundberg, S., Pollak, R. and Wales, T. 1997. Do husbands and wives pool their resources? Evidence from the United Kingdom Child Benefit. *Journal of Human Resources* 32, 463–80.

Manser, M. and Brown, M. 1980. Marriage and household decision making: a bargaining analysis. *International Economic Review* 21, 31–44.

McElroy, M. 1990. The empirical content of Nash-bargained household behavior. *Journal of Human Resources* 25, 559–83.

McElroy, M. and Horney, M. 1981. Nash-bargained household decisions: toward a generalization of the theory of demand. *International Economic Review* 22, 333–49.

Udry, C. 1996. Gender, agricultural production and the theory of the household. *Journal of Political Economy* 104, 1010–46.

How to cite this article

Fafchamps, Marcel. "intrahousehold welfare." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000243> doi:10.1057/9780230226203.0845

The New Palgrave Dictionary of Economics Online

inventory investment

James A. Kahn
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Interest in inventory investment's role in business cycle volatility goes back at least to John Maynard Keynes. This article examines some basic facts about aggregate inventory investment, emphasizing its highly volatile and pro-cyclical nature. It then outlines several approaches to modelling inventory behaviour, including a detailed discussion of the linear-quadratic model, and examines their implications for inventory investment's potential role in business cycle fluctuations. The article concludes with a discussion of the potential for progress in inventory control methods to have played a role in the decline in aggregate volatility since the mid-1980s.

Keywords

adjustment costs; business cycles; credit constraints; dynamic programming; flexible accelerator models; inventory behaviour; inventory investment; linear-quadratic models; Metzler, L. A.; national income accounts; non-convexity; s-S models; stockout-avoidance model

Article

Inventory investment is the change in the stocks of materials, works in process, and finished goods within a firm, industry, or entire economy over a specified period of time. Because in most instances the measure encompasses a variety of goods, it is usually measured in currency units, perhaps deflated (for example, in 1999 dollars). Occasionally, however, when highly disaggregated data are available, it can be measured in physical units (for example, Blanchard, 1983; Kahn, 1992).

In national income accounts, aggregate inventory investment is the difference between Gross Domestic Product (GDP) and final sales of domestic product. As a share of GDP it is tiny but highly volatile in modern industrial economies. In the post-war United States, for example, it averages 0.62 per cent of GDP, but has a standard deviation of 0.83 per cent. By comparison, fixed non-residential investment averages 10.6 per cent of GDP with a standard deviation of 1.2 per cent. (Data for these calculations come from the US National Income and Product Accounts, Table 5.)

Inventory investment is also highly pro-cyclical. For example, its correlation with real GDP growth in post-war US data is approximately 0.4, and very close to the correlation between fixed non-residential investment and real GDP growth. Also, the standard deviation of real GDP growth is substantially higher than that of final sales (4.0 versus 3.3 per cent), notwithstanding the fact that for more than half of the economy GDP and final sales are identical. Thus inventory investment ‘adds’ to the volatility of GDP growth in the accounting (though not necessarily causal) sense. Indeed, interest in inventory behaviour as a contributor to aggregate volatility goes back at least to Keynes (1936), and includes notable contributions by Metzler (1941) and Abramovitz (1950). Blinder (1981, p. 500) writes that ‘to a great extent, business cycles are inventory fluctuations’.

The pro-cyclicality of inventory investment appears inconsistent with standard microeconomic models of inventory behaviour, particularly those that stress ‘buffer stock’ or ‘production-smoothing’ motives, as noted by, among others, Blinder (1986) and West (1986). And this was not the first puzzle brought to light by research on inventory behaviour. Some ten years earlier, Feldstein and Auerbach (1976) noted the persistence of inventory-sales ratios’ deviations around their means (see also Ramey and West, 1999), particularly given the trivial adjustments needed to restore them to a (presumed) fixed target.

Researchers have also found inventory behaviour informative about the fundamental driving forces of business cycles (see West, 1990). For example, Blinder (1986), Eichenbaum

(1989), Kydland and Prescott (1982) and Christiano (1988) hypothesize supply side disturbances to account for pro-cyclical inventory investment. Others (such as Ramey, 1991; Hornstein and Fischer, 2000) consider non-convexities such as fixed costs or downward-sloping marginal cost. Kashyap, Stein and Wilcox (1994) argue for the importance of credit constraints. By contrast, Bils and Kahn (2000) argue that the counter-cyclical behaviour of inventory-sales ratios casts doubt on such supply side explanations, which imply counterfactually that inventories should be relatively tight (in relation to sales) during recessions and plentiful in expansions.

The linear-quadratic model

The workhorse of applied inventory research is the linear-quadratic cost minimization model developed by Holt et al. (1960). The firm is assumed to face a stochastic demand process independent of its inventory and production decisions. Consequently, whether it is a competitive price-taker or has monopoly power, the firm can condition on its expected sales process and minimize costs, which take the form

$$E_t\left\{\sum_{\tau=t}^{\infty}\beta^{\tau-t}\left[c_1 y_\tau + c_2 y_\tau^2 + c_3\left(h_\tau - h_\tau^*\right)^2\right]\right\}, \tag{1}$$

subject to

$$h_\tau = h_{\tau-1} + y_\tau - s_\tau, \tag{2}$$

where y denotes production, s sales, h the end-of-period inventory stock, h^* the desired or 'target' stock, and β a discount factor. Some versions of the model include additional cost terms such as a cost of changing production. The target h^* is usually assumed to be either a constant or proportional to expected sales. In addition, c_1 may be stochastic, and there may be additional additive stochastic terms (for example, materials prices).

A standard informational assumption is that production decisions at date t are based on period $t-1$ information, with the implication that h_t is not controlled directly. Letting $\theta = c_2 / c_3$, the solution to the problem (1)–(2) is:

$$E_{t-1}\{h_t\} = \lambda h_{t-1} + \lambda E_t\left\{\sum_{\tau=t}^{\infty}(\beta\lambda)^{\tau-t}\left[\theta^{-1}h_\tau^* + \beta s_{\tau+1} - s_\tau\right]\right\} \tag{3}$$

where $\lambda \in (0, 1)$ is the smaller root of $\beta\lambda^2 - \left(1 + \beta + \theta^{-1}\right)\lambda + 1$. In the limiting case with $\theta = c_2 = 0$ the solution is $E_{t-1}\{h_t\} = h_t^*$. If h^* is a constant, then the only motive for varying inventories is to smooth production. Durlauf and Maccini (1995) decisively reject this version of the model. It is worth noting that the solution (3) bears some similarity to another widely used model, the flexible accelerator of Lovell (1961) and others

$$h_t - h_{t-1} = \lambda_1\left(h_t^* - h_{t-1}\right) - \lambda_2(s_t - E_{t-1}\{s_t\}) + u_t,$$

where u_t is a disturbance term, and both λ_1 and λ_2 lie between zero and 1.

Both the linear-quadratic and flexible accelerator models, however, have a history of empirical difficulties. Blinder (1986) pointed out that, for the pure production-smoothing model (with h^* a constant), the model counterfactually implies that the variance of sales exceeds that of production. West (1986) showed that a more general variance inequality implied by the model with a target proportional to expected sales is also violated in US manufacturing data. Moreover, among studies of similar data, there is disagreement in the literature on the magnitudes, and even the signs, of key parameters. For example, Ramey (1991) finds negatively sloped marginal cost, in contrast to most other studies. West (1986) finds a relatively small cost of inventory deviations from their target. West and Wilcox (1994) find that obtaining precise estimates of the linear-quadratic model may be problematic with realistic sample sizes.

Regarding the flexible accelerator, Feldstein and Auerbach (1976) estimated small values for both λ_1 and λ_2, which is paradoxical because a small λ_1 implies large adjustment costs, but a small λ_2 implies that sales surprises are largely offset by within-period production responses. Their proposed solution is a target ratio that itself adjusts slowly over time. They do not provide a strong theoretical foundation for their ‘target adjustment’ model, however. One theme of the alternative approaches discussed in the next section is the effort to base inventory models on more rigorous microfoundations in the hope of resolving the empirical puzzles.

Other approaches

Motivated by the empirical difficulties described above, researchers have examined a number of alternative approaches to modelling inventory behaviour. One, the so-called ‘stockout-avoidance’ model, provides a rigorous microfoundation for the target stock. Building on Karlin and Carr (1962), Kahn (1987; 1992) considers a firm that faces a non-negativity constraint on its inventories, and must commit to production and pricing decisions each period before observing potential sales, or ‘demand’ x_t. Consequently, sales equal the minimum of x_t and the stock available $h_{t-1} + y_t$. If we let F denote the distribution function for x, profit maximization implies

$$p_t(1 - F(h_{t-1} + y_t)) - c_t + \beta E_t\{c_{t+1}\}F(h_{t-1} + y_t) = 0,$$

where p is price and c is marginal cost. Then, if demand uncertainty is multiplicative, for example, and p and c (and hence the markup) are constant, the firm will set $h_{t-1} + y_t$ proportional to expected demand $E_{t-1}\{x_t\}$. In addition, positive serial correlation in demand results in the variance of production exceeding the variance of sales.

Another important implication of this approach is that inventory-sales ratios depend on price–cost markups. Bils and Kahn (2000) show, in a model in which expected sales are increasing in the stock available, that the optimal inventory–sales ratio is a function of the markup and a discount rate $\beta E_t\{c_{t+1} / c_t\}$. They argue that the counter-cyclical behaviour of the inventory–sales ratio implies a counter-cyclical markup, or, equivalently, pro-cyclical marginal cost.

An alternative approach builds on the work of Scarf (1960), who modelled inventory behaviour with fixed ordering costs. Scarf provided conditions under which inventories would fluctuate between a fixed upper and lower bound, which he dubbed ‘S’ and ‘s’ respectively – hence the moniker (S,s) model. (The conditions, such as i.i.d. orders, are quite restrictive, however.) Caplin (1985) showed that this model implies that the variance of orders exceeds the variance of sales, and Hornstein and Fisher (2000) extended this approach to a general equilibrium setting. Hall and Rust (2000) provide some empirical support for ‘generalized’ (S,s) behaviour where the two limits depend on the spot price of the good in inventory. While the existence of fixed costs at the microeconomic level is well established, their importance for aggregate inventory behaviour at business cycle frequencies remains a matter of debate.

Inventories and the great moderation

Recently attention has again turned to inventory behaviour as a possible explanation for the dramatic reduction in aggregate volatility, which in the United States dates from approximately 1984 (McConnell and Perez-Quiros, 2000). Kahn, McConnell M. and Perez-Quiros (2002) show that reduced volatility is most pronounced in the durable goods sector,

and for production more than for sales. At the same time, that sector has experienced large declines in inventory–sales ratios, as shown in the accompanying Figure 1, and reduced volatility of inventory investment. They also provide a model in which improved information about demand shocks results in reduced output volatility. While there is much anecdotal evidence of efforts to improve inventory control by techniques such as ‘just-in-time’ management, there remains nonetheless considerable debate over the importance of inventories in increased aggregate stability.

Figure 1

Inventory-sales ratio, durable goods, USA, 1954–98. *Note*: Inventories and sales are in chained 2000 dollars, *Source*: US National Income and Product Accounts. Durable goods inventories are from Table 5.7.6A, and final sales are from Table 1.2.6.

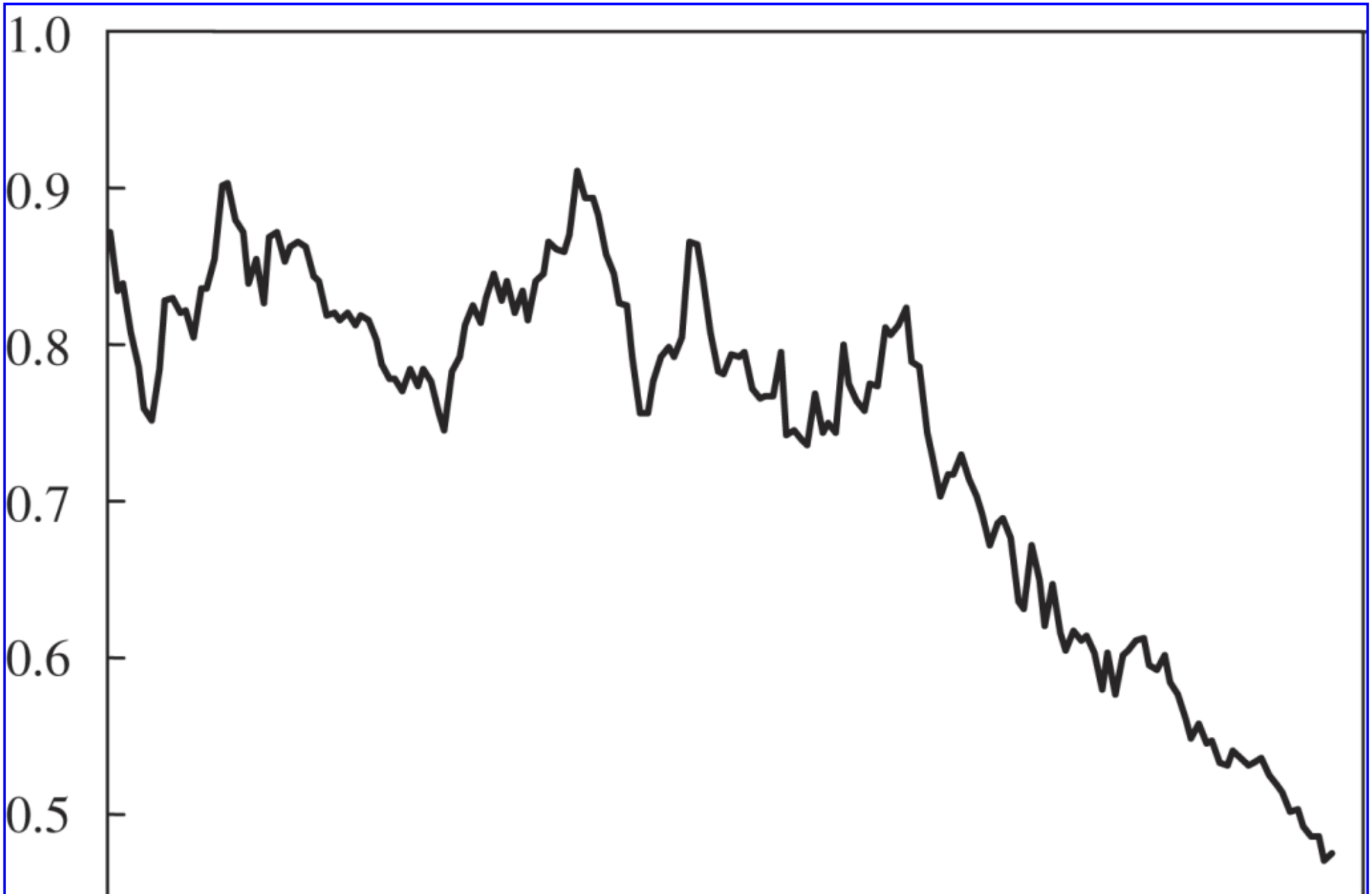

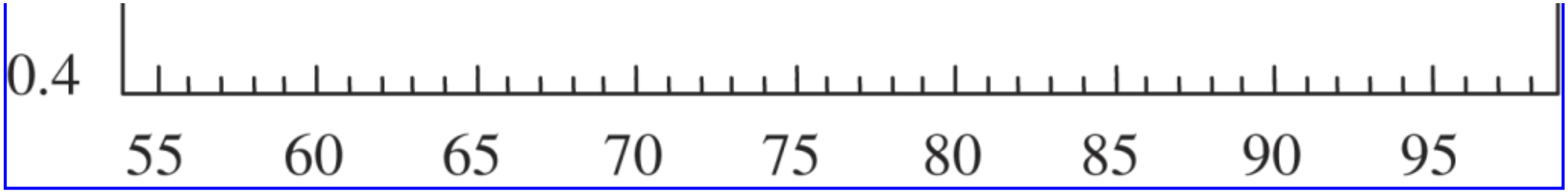

See Also

- dynamic programming
- s-S models

Bibliography

Abramovitz, M. 1950. *Inventories and Business Cycles*. New York: National Bureau of Economic Research.

Bils, M.J. and Kahn, J.A. 2000. What inventory behavior tells us about business cycles. *American Economic Review* 90, 458–81.

Blanchard, O.J. 1983. The production and inventory behavior of the U.S. automobile industry. *Journal of Political Economy* 91, 365–400.

Blinder, A.S. 1981. Retail inventory behavior and business fluctuations. *Brookings Papers on Economic Activity* 1981(2), 443–505.

Blinder, A.S. 1986. Can the production smoothing model of inventory behavior be saved? *Quarterly Journal of Economics* 101, 431–53.

Caplin, A. 1985. The variability of aggregate demand with (S,s) inventory policies. *Econometrica* 53, 1395–410.

Christiano, L. 1988. Why does inventory investment fluctuate so much? *Journal of Monetary Economics* 21, 247–80.

Durlauf, S. and Maccini, L. 1995. Measuring noise in inventory models. *Journal of Monetary Economics* 36, 65–90.

Eichenbaum, M. 1989. Some empirical evidence on the production level and production cost smoothing properties of inventory investment. *American Economic Review* 79, 853–64.

Feldstein, M.S. and Auerbach, A. 1976. Inventory behavior in durable goods manufacturing: the target adjustment model. *Brookings Papers on Economic Activity* 1976(2), 351–96.

Hall, G. and Rust, J. 2000. An empirical model of inventory investment by durable commodity intermediaries. *Carnegie-Rochester Conference Series on Public Policy* 52, 171–214.

Holt, C.C., Modigliani, F., Muth, J.F. and Simon, H. 1960. *Planning Production, Inventories, and Work Force*. Englewood Cliffs, NJ: Prentice-Hall.

Hornstein, A. and Fisher, J. 2000. (S,s) inventory policies in general equilibrium. *Review of Economic Studies* 67, 117–45.

Kahn, J.A. 1987. Inventories and the volatility of production. *American Economic Review* 77, 667–9.

Kahn, J.A. 1992. Why is production more volatile than sales? Theory and evidence on the stockout-avoidance motive for inventory-holding. *Quarterly Journal of Economics* 107,

481–510.

Kahn, J.A., McConnell, M. and Perez-Quiros, G. 2002. On the causes of increased stability in the US economy. *Economic Policy Review* 8, 183–202.

Karlin, S. and Carr, C.R. 1962. Prices and optimal inventory policy. In *Studies in Applied Probability and Management Science*, ed. K. Arrow, S. Karlin and H. Scarf. Stanford: Stanford University Press.

Kashyap, A., Stein, J. and Wilcox, D. 1994. Credit conditions and the cyclical behavior of inventories. *Quarterly Journal of Economics* 109, 565–92.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. Harcourt Brace. London: Macmillan and Co.

Kydland, F. and Prescott, E. 1982. Time-to-build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lovell, M.C. 1961. Manufacturers' inventories, sales expectations, and the acceleration principle. *Econometrica* 29, 293–314.

McConnell, M. and Perez-Quiros, G. 2000. Output fluctuations in the United States: What has changed since the early 1980s? *American Economic Review* 90, 1464–576.

Metzler, L. 1941. The nature and stability of inventory cycles. *Review of Economics and Statistics* 23, 113–29.

Ramey, V. 1991. Nonconvex costs and the behavior of inventories. *Journal of Political Economy* 99, 306–34.

Ramey, V. and West, K. 1999. Inventories. In *The Handbook of Macroeconomics*, ed. J.B. Taylor and M. Woodford. Amsterdam: Elsevier Science.

Scarf, H. 1960. The optimality of (S, s) policies in the dynamic inventory problem. In *Mathematical Methods in the Social Sciences*, ed. K. Arrow, S. Karlin and H. Scarf. Stanford: Stanford University Press.

US National Income and Product Accounts. 2007. Online. Available at http://www.bea.gov/bea/dn/nipaweb/SelectTable.asp, accessed 23 February 2007.

West, K.D. 1986. A variance bounds test of the linear quadratic inventory model. *Journal of Political Economy* 94, 374–401.

West, K.D. 1990. The sources of fluctuations in aggregate inventories and GNP. *Quarterly Journal of Economics* 105, 939–72.

West, K.D. and Wilcox, D. 1994. Estimation and inference in the linear-quadratic inventory model. *Journal of Economic Dynamics and Control* 18, 897–908.

How to cite this article

Kahn, James A. "inventory investment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000292> doi:10.1057/9780230226203.0846

The New Palgrave Dictionary of Economics Online

investment (neoclassical)

Original article by Robert M. Coen and Robert Eisner, revised by Francesca Rondina and Steven N. Durlauf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Investment is capital formation – the acquisition or creation of resources to be used in production. As such, it captures the production side of intertemporal consumption/savings decisions. This entry focuses on neoclassical approaches to the study of investment. Theoretical and empirical issues are discussed.

Keywords

acceleration principle; adjustment costs; capital–output ratio; CES production function; depreciation; distributed lags; elasticity of substitution; Euler equations; full employment; human capital; investment (Keynesian); investment (neoclassical); irreversible investment; opportunity cost of capital; public capital; saving and investment; user cost of capital

Article

Investment is capital formation – the acquisition or creation of resources to be used in production. In capitalist economies much attention is focused on business investment in physical capital, such as buildings, equipment and inventories. But investment is also undertaken by governments (see public capital), nonprofit institutions and households, and it includes the acquisition of human and intangible capital as well as physical capital. In principle, investment should also include improvement of land or the development of natural resources, and the relevant measure of production should include non-market output as well as goods and services produced for sale.
Thus, acquisition of an automobile by government or households is as much investment as acquisition of an automobile by a business firm. The car is used in all cases for the production of transport services. Similarly, government construction of roads, bridges and airports is as much investment as business acquisition of trucks and planes. Expenditures for research and development are investment whether undertaken by business, government or nonprofit universities. And, most important, education and

training, wherever undertaken, are major forms of investment in human capital.

There is a widespread mythology that investment is good and the more investment the better. But investment may be good or bad and there may be too much as well as too little.

Classical and neoclassical economists have stressed the role of investment in providing for the future. Maintaining the current level of output requires keeping up the existing means of production. Economic growth, or the increase in the rate of output, is then seen as depending considerably on the acquisition of additional means of production, that is, investment in excess of the wearing away or depreciation of existing capital. Investment may also contribute to higher output where the new capital ‘embodies’ new and improved technology. That investment will contribute to economic growth presupposes, however, that the additional capital is useful. It must have a positive net product, which is to say that the additional capital must contribute more to future production than the value of the resources used to create it.

How far one should go in allocating resources to investment depends upon our preferences for current consumption versus future consumption, or our preferences between our own consumption and that of our children and grandchildren. It also depends on the production function, that is, the terms under which additional capital can be converted into additional future output. It would hardly seem desirable to sacrifice 100 dollars of current consumption to produce 100 dollars of capital that would result in future production of only 90 dollars. The notion that this is not a relevant issue stems from the assumption that profit-seeking entrepreneurs would not freely undertake investment in which the costs are greater than the returns. It is not always perceived, however, that where governments offer subsidies or ‘tax incentives’ for businesses to undertake investment that would not otherwise seem profitable, such unproductive capital formation is exactly what may be expected.

A second major role for investment has been seen in the achievement and maintenance of full employment. This requires that aggregate investment plus aggregate consumption equal the total output that would be produced if all individuals who wish to work could find employment. Investment may then be inadequate not only in failing to provide sufficient resources for future production, it may also be inadequate if it is insufficient to bring about the full utilization of existing resources. This latter problem has received major attention as a consequence of the work and influence of John Maynard Keynes (1936).

Another way of stating the condition necessary for full employment is that aggregate investment must equal aggregate saving out of the full-employment level of income. In national income accounts, measured investment and saving are always identically equal, owing to the identity of output and income, which, apart from receipts from abroad, is earned only from production. That part of income not spent on consumption is saved. But that part of production not purchased by consumers must be acquired (or kept) by producers and hence is investment, though not necessarily intended investment. If we designate Y as income and output, C as consumption, S as saving and I as investment, we then have $S=Y-C=I$.

While realized investment is thus identically equal to saving, investment and saving may be more or less than investment demand, that is, intended investment. If investment demand is less than saving at the current level of income, producers will find that they cannot sell all that they produce. They will accumulate undesired inventories of finished goods (unintended investment in inventories), which should lead them to reduce production. Reduced production means less income and hence less consumption and saving. A shortfall of investment demand in relation to saving therefore brings on a cumulative reduction of output and income until saving and investment are brought down to equality

with the lesser investment demand. Insufficiency of investment demand has been identified with depression and recessions and tendencies towards chronic unemployment. Stimuli to investment, such as reductions in tax rates on income from capital, have thus seemed in order to bring investment demand up to the levels of saving that would be forthcoming with full employment. Conversely, excessive levels of investment demand can create inflationary pressures, calling for policies that would restrict investment.

These Keynesian perceptions as to the costs and benefits of investment are startlingly different from those of the classical models – old and new – which assume, implicitly or explicitly, that the economy is operating at full employment and full utilization of resources. In the classical models, more current production of capital must mean less current production of consumption goods and services. And more consumption now must mean less current investment and less output and consumption in the future.

In an economy with substantial unemployed resources, however, more investment need not and probably will not bring less consumption. Expenditures for additional investment will rather constitute additional incomes for their recipients, and this income will in turn largely be spent on increased consumption. Thus the production of consumption goods and services will increase rather than decline. And more consumption may bring about more investment, as producers see a need for additional capital to increase the output of consumption goods and services.

Classical and Keynesian views also differ on the principal mechanism by which intended investment and saving are equated. In the classical view, changes in the rate of interest are presumed to perform this task. Investment demand is thought to be negatively related and very sensitive to the rate of interest, which is the cost of borrowing funds to finance capital spending. If investment demand is smaller than saving at the full-employment level of incomes, the classical analysis holds that the excess of funds in the credit market will depress interest rates, thereby inducing increases in investment demand (and possibly reductions in saving as the interest earned by savers falls) until intended investment and saving are equal. Thus, no change in the level of economic activity (output) need occur as in the Keynesian analysis. The Keynesian view of the equilibrating process has interest rates playing a smaller role than changes in output, because investment demand is thought to be relatively insensitive to interest rates, being dominated instead by producers' expectations of future demand for their products. Even if the investment demand were sensitive to interest rates, expectations could be so pessimistic that, even if the rate of interest were to fall to zero, there would be insufficient investment demand.

Empirical studies have attempted to measure the influence of interest rates, taxes and expectations of future demand on investment decisions. Producers are presumed to acquire capital to increase their expected profits. The profitability of additional capital depends on its cost, on its expected productivity and on expectations on the price at which additional output can be sold. On the assumption that output is a fixed, 'well-behaved' function of capital and labour (strictly concave, with declining partial derivatives of output with the respect to capital and labour and positive cross-partial derivatives), producers will acquire capital to the point where its declining marginal product equals its cost. This will then define both the desired, or equilibrium, capital–labour ratio and capital–output ratio. With the supply of labour and the rate of output fixed and no change in the relative price of capital and labour, investment in equilibrium will be equal to depreciation, or what is necessary to maintain the existing capital stock, and net investment will be zero. Positive net investment will then stem from increases in the demand for output or reductions in the relative price of capital. Increases in output will generate investment demand to maintain the equilibrium capital–output ratio. A reduction in the cost of capital would generate investment in order to increase the capital–labour and capital–output ratios. In either case, maintaining

increased amounts of capital will generate further investment to cover increased depreciation.
In general, the desired capital stock may be written as:

$$K^* = f(p, c, Y^*), \tag{1}$$

where p is the price of output, $c = q[i - (\dot{q}/q) + d]$ is the rental price or user cost of capital, q is the supply price of capital goods, i is the opportunity cost of capital, d is the rate of economic depreciation and Y^* is desired output. If firms minimize expected costs of producing an exogenously given or expected output Y, then the wage rate, w, would be substituted for p. The rental price, or user cost of capital, c, is the cost per period of holding and maintaining one unit of capital. In the absence of taxes, it is the price of capital goods multiplied by the sum of the real interest rate and the rate of economic depreciation. The former measures the opportunity cost in terms of forgone net earnings from lending or otherwise investing money, plus the capital loss (or minus the capital gain) associated with changing prices of capital goods.
Building on this neoclassical theory of the firm developed by Haavelmo (1960), and assuming a Cobb–Douglas production function with elasticity of output with respect to capital, b, Jorgenson (1963; 1967) arrived at a demand function for capital with a particular form that has been employed in a large number of influential studies:

$$K^* = b(p/c)Y^*. \tag{2}$$

With an implicit unitary elasticity of K^* with respect to c, this formulation implies strong effects of monetary policy, via the rate of interest, and of tax policy so far as, by accelerated tax depreciation, investment subsidies or exclusion of capital gains from taxation, it affects the value of c (see below).
The more general constant-elasticity-of-substitution (CES) production function may be used to generate a demand for capital having the form:

$$K^* = h(p/c)^s(Y^*)^r, \tag{3}$$

where s, the elasticity of substitution between labour and capital, is the critical elasticity of demand for

capital with respect to the relative price of capital, and *r* is the elasticity of demand for capital with respect to output. The elasticity, *r*, will be greater than, equal to, or less than unity as the returns to scale are decreasing, constant or increasing.

If relative prices are constant, or if technology requires that capital and labour be used in fixed proportions (in which case the elasticity of substitution is zero), then with constant returns to scale, desired capital is proportional to the demand for output. This form of the demand for capital leads to the ‘acceleration principle’, according to which net investment demand, arising from a desire to change the stock of capital, depends not on the level of demand for output, but on the *change* in demand for output (Clark, 1917). To induce firms to invest (acquire more capital), demand for output must be expected to rise. Both the original formulation by Jorgenson of the demand for capital (2) and the more general formulation (3) underly a ‘flexible accelerator’, where the desired capital–output ratio is not constant but depends on prices and on the scale of output and, as seen below, investment is subject to a distributed lag process (Koyck, 1954) affected by adjustment costs and the dynamic process governing the formation of expectations of future variables (Eisner and Strotz, 1963; Helliwell and Glorieux, 1970; Lucas, 1976; Eisner, 1978).

Many early econometric studies of investment behaviour tested the accelerator in various forms, but generally they did not allow for effects of prices on the desired capital–output ratio, which is the hallmark of Jorgenson's neoclassical approach. The major competing hypothesis was that investment depends on the level of profits, on the grounds that realized profits measure expected profits, or that capital market imperfections cause firms’ capital expenditures to be constrained by the flow of internal funds (Meyer and Kuh, 1957). Reviews of these earlier investigations are found in Eisner and Strotz (1963) and Jorgenson (1971). The practice in recent studies has been to capture profit expectations by including expectations of the major determinants of profits, namely, sales, prices and wages, or to approximate them by stock market valuations of firms. The flow of internal funds may play some role in investment decisions, not as a determinant of the desired capital stock but as a factor influencing the speed of adjustment of capital (Coen, 1971).

To study the effects of tax policy on demand for capital, the rental price can be generalized to incorporate parameters of the tax system. For example, the after-tax cost of holding one unit of capital would be:

$$c = q[(1 - uv)i - (1 - uw)(\dot{q}/q) + d][1 - k - uz]/(1 - u) \tag{4}$$

where *u* is the rate of taxation of business income; *v* is the proportion of the opportunity cost of capital (such as interest, dividends and forgone earnings) that is tax deductible; *w* is the proportion of capital gains and losses effectively taxed; *k* is the effective rate of the investment tax credit or subsidy; and *z* is the present value of the tax depreciation expected from a dollar of investment (Hall and Jorgenson, 1967).

It can be seen, in this definition, that higher values of *v*, *k* and *z* (from accelerating tax depreciation)

reduce the value of c, as does a higher value of w, provided that capital goods prices are expected to rise. The value of c would also be lowered by decreasing the rate of interest or other measure of the opportunity cost of capital. A higher rate of inflation of capital goods prices has two opposing effects on c. In so far as higher inflation reduces the real after-tax opportunity cost of capital, it reduces c. However, if tax depreciation is based on the historical cost of assets rather than on replacement costs, inflation reduces the present value of tax allowances, z, and thereby raises c (Feldstein, 1982). Finally, we may note that changes in the general rate of business taxation are ambiguous in their effects on c. If v and w are unity, and if the opportunity cost of capital is unaffected by a change in the business tax rate, then a decrease in u will reduce, leave unchanged or increase c as the present value of tax allowances on a unit of investment (including the investment credit) is less than, equal to or greater than the present value of economic depreciation (Hall and Jorgenson, 1971). But then, going back to eq. (1), the effect of any of these parameters on K^* depends upon the elasticity of the latter with respect to c.

The desired capital stock does not in itself indicate the rate of investment, which is the rate of replacement of existing capital plus the rate of net additions. Both entail a combination of financial considerations and costs of adjustment, which will in turn relate to costs of acquiring information necessary to decisions, costs of planning and the supply function for capital goods, all filtered through the expectations of agents. If adjustment costs are an increasing function of the rate of investment, it will generally prove optimal not to adjust capital to the desired level immediately, but instead to distribute changes in the capital stock over time (Eisner and Strotz, 1963).

The speed of adjustment of capital to changes in its desired or equilibrium level may depend on the causes and magnitudes of the changes. An increase in the demand for output may generate investment with all due speed as expectations become firm with regard to the permanence of the increased demand. If, however, the increased demand for capital is due to a fall in its relative price (because, let us say, of a reduction in the rate of interest), thus generating a demand for more durable and hence more substantial and expensive capital, the rate of investment may be slowed by the availability of existing capacity sufficient for current production. These considerations underlie the ‘putty-clay’ model in which the capital–labour ratio can be varied on newly installed capacity but cannot be altered on existing capacity. A demand for additional housing services will bring on investment in housing as rapidly as cost considerations permit. A lower rate of interest, causing substantial investment in more durable brick houses to replace less durable houses of wood or straw, would cause the rate of investment to increase only as existing houses of wood and straw wear out and are replaced.

Investment equations should thus in principle involve separate distributed lag responses to changes in relative prices and to changes in output. They should also admit the possibility that the lag distribution is not fixed and may vary with other economic parameters and the expectations function.

A logarithmic transformation of eq. (3) yields

$$\ln K^* = \ln h + s\ln(p/c) + r\ln Y. \tag{5}$$

Putting this in first difference form, we have:

$$\Delta \ln K^{*} = s\Delta \ln(p/c) + r\Delta \ln Y. \tag{6}$$

Since the change in the logarithm of capital is the relative change in capital, we may treat the ratio of net investment to existing capital stock as approximately equal to $\Delta \cdot \ln \cdot K$, which may in turn be written as a distributed lag function of changes in the determinants of desired capital:

$$I_N / K_{-1} = s[q_1(L)\ \Delta \ln(p/c)] + r[q_2(L)\ \Delta \ln Y], \tag{7}$$

where $q_1(L)$ and $q_2(L)$ are lag operators that indeed should be functions of such variables as the rate of interest, and the cost and availability of capital. Then, finally, since investments equal net investment plus replacement, we may write

$$I = I_N + R = I_N + dK_{-1}, \tag{8}$$

where d, the replacement rate, may vary over time.

Estimates of investment functions of this type have often neglected influences of economic variables and expectations on adjustment processes and the replacement rate. Lag distributions are assumed to be of some fixed functional form, and d is assumed to be constant (for evidence that d may not be constant, see Feldstein and Foote, 1971; Eisner, 1972; Feldstein and Rothschild, 1974; Coen, 1975).

Where production and lag parameters have not been unduly constrained by a priori specifications, estimates have generally yielded values of s, the elasticity of substitution, considerably less than unity, in some cases not substantially greater than zero (see Eisner and Nadiri, 1968; Coen, 1969; Lucas, 1969; Eisner, 1978; Chirinko and Eisner, 1982). Lag distributions estimated from time series and cross-section data have usually extended over a number of years (Eisner, 1978), and they often have inverted-U shapes. Where a putty-clay formulation has been employed with separate lags on relative prices and output, the mean lags on prices are typically much longer than those on output (Bischoff, 1971).

These findings of small price elasticities of demand for capital suggest some role, but a limited one, for monetary and tax policies in directly affecting the general rate of investment through the rental cost of capital. The long lag distributions on relative prices suggest further difficulties in the use of monetary or

fiscal policy for reducing cyclical fluctuations in investment. However, policy impacts may operate not only on the desired capital stock but also on the speed of adjustment of capital.

Repeated changes in tax parameters such as *k*, the rate of investment tax credit or subsidy, may be used to bring about intertemporal substitution of investment even if the effects on its long-run average are small. Thus, when investment is low, the marginal rate of subsidy, *k*, might be raised, while if investment were deemed too great, the value of *k* could be reduced to zero or indeed made negative (an investment tax instead of subsidy). Paradoxically, a fluctuating and uncertain investment subsidy/tax may have substantial effects on investment where permanent subsidies or taxes would not. There is thus an asymmetry between effects of changes in the cost of capital and changes in the demand for output, the effects of which on investment will be proportional to the permanence with which they are perceived (Eisner, 1978).

Most investment functions, with their ad hoc, fixed lag distributions and assumptions of static expectations, fail to capture accurately the effects of economic policies on the timing of investment or to distinguish properly between the effects of temporary and permanent policy changes (Lucas, 1976). To correct these shortcomings, adjustment costs must be explicitly introduced in the firm's optimization problem, so that instead of there being a desired stock of capital towards which the firm moves in a mechanistic way, there is a desired path of capital accumulation. Along such a path, the optimal rate of investment at each point, including the present, will in general depend on expected relative prices and output over the entire planning horizon.

Obtaining solutions to the firm's dynamic optimization problem under very general specifications of technology and expectations has proven difficult. To make such an approach empirically tractable, strong assumptions are usually made, for example, that the production function is quadratic, that adjustment costs are quadratic, symmetric and separable from the rest of technology (the cost of adjusting capital, for example, does not depend on the quantities of capital and labour currently employed), and that expectations are characterized by relatively simple autoregressive processes.

The critical role in current investment of unobservable adjustment costs and of uncertain, shifting (and generally not directly observable) expectations of the future, stressed by Keynes, has sparked interest in a formulation of an investment function that directly relates demand prices and supply prices of capital. Going back to Keynes's *General Theory*, we have investment undertaken to the point where the expectation of marginal profit on investment (the ‘marginal efficiency of investment’) is equal to the rate of interest or, alternatively, the present value of expected returns from the marginal investment, using the rate of interest as the discount factor, is equal to the marginal supply price of newly produced capital goods. Building on this, Brainard and Tobin (1968) and Tobin (1969), presented a ‘*q*-theory’, which sees investment as a positive function of the ratio, *q*, of the market value of capital to its replacement cost. The former may in principle be observed in the trading prices of stock shares along with bonded indebtedness of business firms. With proper adjustment for tax considerations, when the value of *q* is greater than unity, investment will take place because the cost of additional capital will be less than the market evaluation of the present value of returns from capital. Conversely, when *q* is less than unity, business demand for capital may be better satisfied by acquisitions taking over existing firms and their facilities than by new investment. In general, the rate of investment should be greater the greater the value of *q*.

Empirical estimation of ‘*q*’ investment equations and predictions based on these estimates have not, however, proved very successful (von Furstenberg, 1977; Abel, 1980; Summers, 1981; Hayashi, 1982;

Abel and Blanchard, 1986). Suggested explanations of the difficulties include the fact that market values of firms may relate to much more than the tangible capital generally included in business investment, and the failure to distinguish marginal and average values of the cost of new capital versus the acquisition costs of existing firms (Chirinko, 1986).

Investment decisions are but one element of producers' plans for hiring or acquiring factors of production. Interrelationships between investment demand and demands for other inputs have been a subject of growing interest. Since factor demands are derived from a given production function, they share common technological parameters and may be estimated as a system of demand functions (Coen and Hickman, 1970). Such an approach calls attention to effects of investment stimuli that are often overlooked. For example, at a given level of output, the direct impact of an investment tax credit is to reduce the demand for labour, since it raises the relative cost of labour. Employment may eventually be raised, but only if the expansion in aggregate output induced by the increase in investment demand is large enough to offset the adjustment to a higher capital–labour ratio.

Additional interrelationships may arise when capital is not the only factor of production subject to adjustment costs. If labour input is also costly to change, then the rate of investment may depend not only on the desired adjustment in capital stock but also on the desired adjustment in employment (Nadiri and Rosen, 1969; Brechling, 1975; Epstein and Denny, 1983). Furthermore, since a firm must operate on its production function, factor adjustments cannot be entirely independent. If output is exogenously given and there are n inputs, $n - 1$ inputs can be independently adjusted, but the nth is determined by the production function, the level of output, and the quantities of the other inputs (Gould, 1969). It may be unreasonable to view the production function as a binding constraint, however, because it is difficult, if not impossible, to measure perfectly all inputs and their utilization rates.

With the development of dynamic optimization models of interrelated factor demands in which various types of capital and other inputs are subject to adjustment costs, and expectations are not treated as static, it is possible to estimate the magnitude of adjustment costs for capital, to see how they affect and are affected by adjustments of other inputs, and to study the impacts of changes in producers' perceptions of the *processes* generating prices, output and policy parameters. As we noted above, this approach necessitates strong restrictions on functional forms to obtain explicit decision rules for accumulation of capital and employment of other inputs (Meese, 1980). Where general forms of the production, adjustment cost and expectations functions are assumed and the model cannot be solved completely, it is still possible to estimate the first-order conditions (Euler equations) that implicitly define the evolution of the optimal inputs (Pindyck and Rotemberg, 1983; Shapiro, 1986). Such estimates do not give a complete account of the dynamics of investment behaviour for any initial conditions and stochastic environment, but they do give insights about differing short- and long-run responses to, say, an unexpected increase in the price of energy starting today versus the same increase beginning five years from now but anticipated today. An important empirical development in the study of investment is the use of more disaggregated data-sets; an important example is Cummins, Hassett and Hubbard (1994), which analyses the effects of major tax reforms on investment based on firm-level panel data. The paper is important in providing much stronger evidence on the importance of the user cost of capital than appears in aggregate studies. Chirinko (1993) is still of value as a survey.

Yet, as noted in the valuable review of Caballero (1999), a general dissatisfaction with the empirical performance of the neoclassical model led to change in investment research which emphasized the role

of irreversibilities. See irreversible investment for these new developments.

See Also

- irreversible investment
- neoclassical synthesis
- new classical macroeconomics
- public capital
- rational expectations
- Tobin's q

Bibliography

Abel, A.B. 1980. Empirical investment equations: an integrative framework. In *On the State of Macro-Economics*, ed. K. Brunner and A.H. Meltzer. Vol. 12 of Carnegie-Rochester Conference on Public Policy. Amsterdam: North-Holland.

Abel, A.B. and Blanchard, O.J. 1986. The present value of profits and cyclical movements in investments. *Econometrica* 54, 249–73.

Bischoff, C.W. 1971. The effect of alternative lag distributions. In *Tax Incentives and Capital Spending*, ed. G. Fromm. Washington, DC: Brookings Institution.

Brainard, W.C. and Tobin, J. 1968. Pitfalls in financial model-building. *American Economic Review* 58, 99–122.

Brechling, F. 1975. *Investment and Employment Decisions*. Manchester: Manchester University Press.

Caballero, R. 1999. Aggregate investment. In *Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Chirinko, R.S. 1986. Business investment and tax policy: a perspective on existing models and empirical results. *National Tax Journal* 39, 137–55.

Chirinko, R. 1993. Fixed business investment spending: a critical survey of modeling strategies, empirical results and policy implications. *Journal of Economic Literature* 31, 1875–911.

Chirinko, R.S. and Eisner, R. 1982. The effects of tax parameters in the investment equations in macroeconomic econometric models. In *Economic Activity and Finance*, ed. M.E. Blume, J. Crockett and P. Taubman. Cambridge, MA: Ballinger.

Clark, J.M. 1917. Business acceleration and the law of demand: a technical factor in economic cycles. *Journal of Political Economy* 25, 217–35 Reprinted in American Economic Association, *Readings in Business Cycle Theory*. Philadelphia: Blakiston, 1951.

Coen, R.M. 1969. Tax policy and investment behaviour: comment. *American Economic Review* 59, 370–77.

Coen, R.M. 1971. The effect of cash flow on the speed of adjustment. In *Tax Incentives and Capital Spending*, ed. G. Fromm. Washington, DC: Brookings Institution.

Coen, R.M. 1975. Investment behavior, the measurement of depreciation, and tax policy. *American Economic Review* 65, 59–74.

Coen, R.M. and Hickman, B.G. 1970. Constrained joint estimation of factor demand and production functions. *Review of Economics and Statistics* 52, 287–300.

Cummins, J.G., Hassett, K.A. and Hubbard, R.G. 1994. A reconsideration of investment behavior using tax reforms as natural experiments. *Brookings Papers on Economic Activity* 1994(2), 1–59.

Eisner, R. 1972. Components of capital expenditures: replacement and modernization versus expansion. *Review of Economics and Statistics* 54, 297–305.

Eisner, R. 1978. *Factors in Business Investment*. Cambridge, MA: Ballinger.

Eisner, R. 1985. The total incomes system of accounts. *Survey of Current Business* 65, 24–48.

Eisner, R. 1986. *How Real is the Federal Deficit?* New York: Free Press, Macmillan.

Eisner, R. and Nadiri, M.I. 1968. Investment behavior and the neo classical theory. *Review of Economics and Statistics* 50, 369–82.

Eisner, R. and Strotz, R.H. 1963. Determinants of business investment. *Commission on Money and Credit, Impacts of Monetary Policy*. Englewood Cliffs, NJ: Prentice-Hall.

Epstein, L.G. and Denny, M.G.S. 1983. The multivariate flexible accelerator model: its empirical restrictions and an application to U.S. manufacturing. *Econometrica* 51, 647–74.

Feldstein, M.S. 1982. Inflation, tax rules and investment: some econometric evidence. *Econometrica* 50, 825–62.

Feldstein, M.S. and Foote, D.K. 1971. The other half of gross investment: replacement and

modernization expenditures. *Review of Economics and Statistics* 53, 49–58.

Feldstein, M.S. and Rothschild, M. 1974. Towards an economic theory of replacement investment. *Econometrica* 42, 393–423.

Gould, J.P. 1969. The use of endogenous variables in dynamic models of investment. *Quarterly Journal of Economics* 83, 580–99.

Haavelmo, T. 1960. *A Study in the Theory of Investment*. Chicago: University of Chicago Press.

Hall, R.E. and Jorgenson, D.W. 1967. Tax policy and investment behavior. *American Economic Review* 58, 391–414.

Hall, R.E. and Jorgenson, D.W. 1971. Application of the theory of optimal capital accumulation. In *Tax Incentives and Capital Spending*, ed. G. Fromm. Washington, DC: Brookings Institution.

Hayashi, F. 1982. Tobin's marginal q and average q: a neoclassical interpretation. *Econometrica* 50, 213–24.

Helliwell, J.F. and Glorieux, G. 1970. Forward looking investment behavior. *Review of Economic Studies* 37, 499–516.

Hicks, J.R. 1950. *A Contribution to the Theory of the Trade Cycle*. Oxford: Clarendon Press.

Jorgenson, D.W. 1963. Capital theory and investment behavior. *American Economic Review, Papers and Proceedings* 53, 247–59.

Jorgenson, D.W. 1967. The theory of investment behavior. In *Determinants of Investment Behavior*, ed. R. Ferber. New York: Columbia University Press.

Jorgenson, D.W. 1971. Econometric studies of investment behavior: a survey. *Journal of Economic Literature* 9, 1111–47.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

King, M.A. and Fullerton, D.K. 1984. *The Taxation of Income from Capital*. Chicago: University of Chicago Press.

Koyck, L.M. 1954. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.

Lange, O. 1938. The rate of interest and the optimum propensity to consume. *Economica* 5, 12–32

Reprinted in American Economic Association, *Readings in Business Cycle Theory*, Philadelphia: Blakiston, 1951.

Lucas, R.E. 1969. Labor-capital substitution in U.S. manufacturing. In *The Taxation of Income from Capital*, ed. A.C. Harberger and M.J. Bailey. Washington, DC: Brookings Institution.

Lucas, R.E. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A.H. Meltzer. Vol. 1 of Carnegie-Rochester Conference in Public Policy. Amsterdam: North-Holland.

Meese, R. 1980. Dynamic factor demand schedules for labor and capital under rational expectations. *Journal of Econometrics* 14, 141–58.

Meyer, J. and Kuh, E. 1957. *The Investment Decision*. Cambridge, MA: Harvard University Press.

Nadiri, M.I. and Rosen, S. 1969. Interrelated factor demand functions. *American Economic Review* 59, 457–71.

Pindyck, R.S. and Rotemberg, J.J. 1983. Dynamic factor demands and the effects of energy price shocks. *American Economic Review* 73, 1066–79.

Shapiro, M.D. 1986. The dynamic demand for capital and labor. *Quarterly Journal of Economics* 101, 513–42.

Summers, L.H. 1981. Taxation and corporate investment: a *q*-theory approach. *Brookings Papers on Economic Activity* 1981(1), 67–140.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1, 15–29.

von Furstenberg, G.M. 1977. Corporate investment: does market valuation matter in the aggregate? *Brookings Papers on Economic Activity* 1977(2), 347–97.

How to cite this article

Coen, Robert M., et al. "investment (neoclassical)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000305> doi:10.1057/9780230226203.0848

The New Palgrave Dictionary of Economics Online

investment decision criteria

Jack hirshleifer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Investment is present sacrifice for future benefit. Individuals, firms, and governments all are regularly in the position of deciding whether or not to invest, and how to choose among the options available. An individual might have to decide whether to buy a bond, plant a seed, or undertake a course of training; a firm whether to purchase a machine or construct a building; a government whether or not to erect a dam. Under the heading of investment decision criteria, economists have addressed the problem of how to choose rationally in situations that involve a tradeoff between present and future.

The Economic Theory of Intertemporal Choice

The object of investment is taken to be to optimize one's pattern of consumption over time. The elements needed to determine an individual's investment decision are: (a) his *endowment*, in the form of a given existing income stream over time; (b) his *preference function*, which orders in desirability all possible time-patterns of consumption; and (c) his *transformation set*, which specifies the possibilities for transforming the original endowment into other time-combinations of consumption.
Figure 1 illustrates an artificially simple case of only two periods (say, this year and next) under conditions of certainty. Each point represents a combination of current consumption c_0 and future consumption c_1. The *endowment* combination Y has coordinates (y_0, y_1). *Time-preferences* are portrayed by the indifference curves U_1, U_2, U_3, …, each such curve connecting combinations yielding equal satisfaction. The curve QQ' through the endowment position Y pictures the intertemporal *productive opportunities*. By sowing seed, for example, a person can sacrifice current consumption for future consumption – represented in the diagram by a movement from Y along QQ' to the northwest. (There may also be disinvestment opportunities, i.e., the individual might be able to draw upon the future so as to augment current consumption, which would be represented by a movement from Y along QQ' to the south-east.)
Figure 1
Investment and saving in a 2-period model

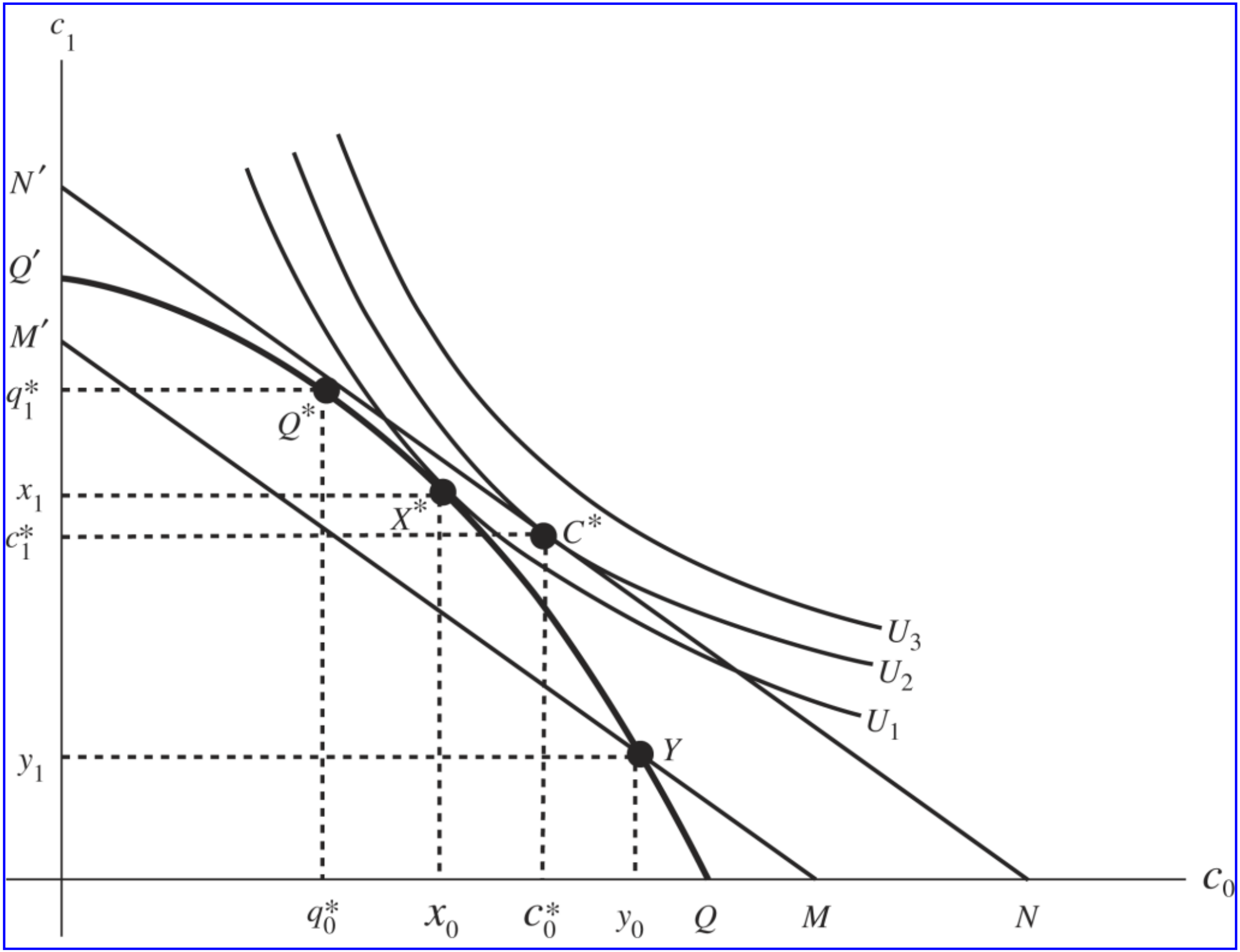
c_1
N'
Q'
M'
q_1^*
Q^*
x_1
X^*
c_1^*
C^*
U_3
U_2
U_1
y_1
Y
c_0
q_0^*
x_0
c_0^*
y_0
Q
M
N

For a Robinson Crusoe, the optimum balance of present and future consumption – which in his isolated state must necessarily be identical to his provision for present and future production – occurs at point X^* along QQ'. In the situation pictured he achieves this optimum by *investing* the quantity $y_0 - x_0$ of current consumption claims. For example, having at hand a current corn endowment of y_0, he retains x_0 for current consumption and plants the remainder as seed. Next year he will reap as return from investment the amount $x_1 - y_1$ to augment his endowed availability of future corn.

If markets for trading between present and future income claims exist, however, in contrast with the Robinson Crusoe situation the individual will be able to disconnect the amount he *invests* from the amount he *saves*. These trading opportunities are shown in Figure 1 by the family of 'market lines' whose general equation is:

$$c_0 + c_1 / (1 + r_1) = W_0 \quad (1)$$

Here r_1 is the *interest rate* that discounts one-year future claims c_1 into their equivalent value in terms of c_0 claims. Along each market line the parameter W_0 represents the associated level of *wealth*. Put another way, wealth in equation (1) measures the *present worth* of any specified (c_0, c_1) vector – the future-dated element being 'discounted' at the given market interest rate r_1. In the diagram two market lines are shown: MM' through the endowment vector $Y=(y_0, y_1)$ indicates the individual's endowed wealth $W_0^y = y_0 + y_1 / (1 + r_1)$, while NN' represents the maximum attainable level of wealth $W_0^* = q_0^* + q_1^* / (1 + r_1)$.

If an individual has both productive and market opportunities, his optimizing decision in Figure 1 can be thought of as taking place in two stages. First he locates his 'productive solution' $Q^* = (q_0^*, q_1^*)$ by moving along QQ' so as to maximize attained wealth at the tangency with market line NN'. Second, he then transacts in the funds market, by lending or borrowing (exchanging current for future claims or vice versa) along NN' to find his 'consumptive solution' $C^* = (c_0^*, c_1^*)$ at the tangency of NN' with indifference curve U_2 in the diagram. Notice that his preferences do not at all affect the productive solution, but only how he chooses to 'finance' the investments made. Specifically, in the diagram here the amount he *invests* $(y_0 - q_0^*)$ exceeds the amount he *saves* $(y_0 - c_0^*)$. By borrowing on the market, in effect he has been able to get others to undertake part of the saving necessary to finance his projected investments.

This disconnection between the individual's productive and consumptive decisions in a regime of perfect markets is known as 'Fisher's Separation Theorem'. The essential implication is that individuals with diverging time-preferences can nevertheless come together and agree upon joint productive investments. Business firms and (to some extent) governments can be regarded as institutions designed for undertaking joint investments whose scale is too large for any single individual. The underlying principle is that those investment choices maximizing wealth value or present worth of the mutual undertaking will also maximize wealth for each and every participant therein.

The Present-Value Rule

The economic theory of intertemporal choice leads immediately to what is known as the *Present-Value Rule* for investment decision. This rule can be expressed in two essentially equivalent forms:

1. (i) Among the opportunities available, adopt the set of investments that maximizes wealth W_0.
2. (ii) Adopt any single investment project if and only if its present value V_0 is positive. (Taking into account, of course, any repercussions of that project upon the returns yielded by other members of the adopted investment set.)

As an obvious corollary, if two available projects are mutually exclusive, the one with the larger present value V_0 should be chosen.

Generalizing to the multi-period context, wealth as maximand becomes:

$$W_0 = q_0 + q_1 / (1 + r_1) + q_2 / [(1 + r_2)(1 + r_1)] + \cdots + q_T / [(1 + r_T) \cdots (1 + r_2)(1 + r_1)]$$

(2)

Here the q_t are the coordinates of points along the T+1-dimensional productive opportunity surface

$$\varphi(q_0, q_1, \ldots, q_T) = 0,$$

a generalization of curve QQ' in Figure 1. T is the 'economic horizon', which may be infinite. And the r_t represent the successive short-term interest rates, each of which discounts prospective payments at any date into its wealth-equivalent at the next preceding date.
For a single project in the multi-date context, present value is defined as:

$$V_0 = z_0 + z_1 / (1 + r_1) + z_2 / [(1 + r_2)(1 + r_1)] + \cdots\cdots + z_T / [(1 + r_T)\cdots(1 + r_2)(1 + r_1)]$$
(3)

Here the z_t are the dated payments or 'cash flows' associated incrementally with the project considered. Normally the z_1 elements for earlier dates would include some with negative signs – or else the project could not be described as an investment – while those for later dates would have predominantly positive signs. In the special case where r_1=r_2=…=r_T=r – that is, where interest rates are expected to remain constant at the level r over time – the Present-Value formulas reduce to the more familiar forms:

$$W_0 = q_0 + q_1 / (1 + r) + q_2 / (1 + r)^2 + \cdots + q_T / (1 + r)^T$$
(2′)

$$V_0 = z_0 + z_1 / (1 + r) + z_2 / (1 + r)^2 + \cdots + z_T / (1 + r)^T$$
(3′)

The Present-Value solutions can also be formally generalized to allow for continuous rather than discrete time. As an illustrative simplified example, consider a project whose scale of current input or investment sacrifice i_0 is fixed while the output date is subject to choice (e.g., when to cut a growing tree). In Figure 2, horizontal distances represent time t and vertical distances value V_t at each date. Present Value V_0 is indicated by height along the vertical axis. The curve GG' represents productive growth of the asset – in the case of a tree, market value of the standing timber at any date. The 'discount curves' D, D', D'', …, are analogous to the 'market lines' of Figure 1. Each such curve represents the growth of a specific sum of present dollars by continuous compounding at a constant market rate of interest r, or alternatively the Present Value of any future payment continuously discounted at r. The optimal investment period t=t^* is then the one that maximizes Present Value V_0, subject to the constraint on the available V_t described by the curve GG', in the equation:

$$V_0 = -i_0 + V_t e^{-rt} \quad (4)$$

Geometrically, t^* is determined by the tangency of GG' with the highest discount curve (constant-wealth curve) attainable. The solution condition is then:

$$V_t' / V_t = r \quad (5)$$

Figure 2
Optimal during of investment

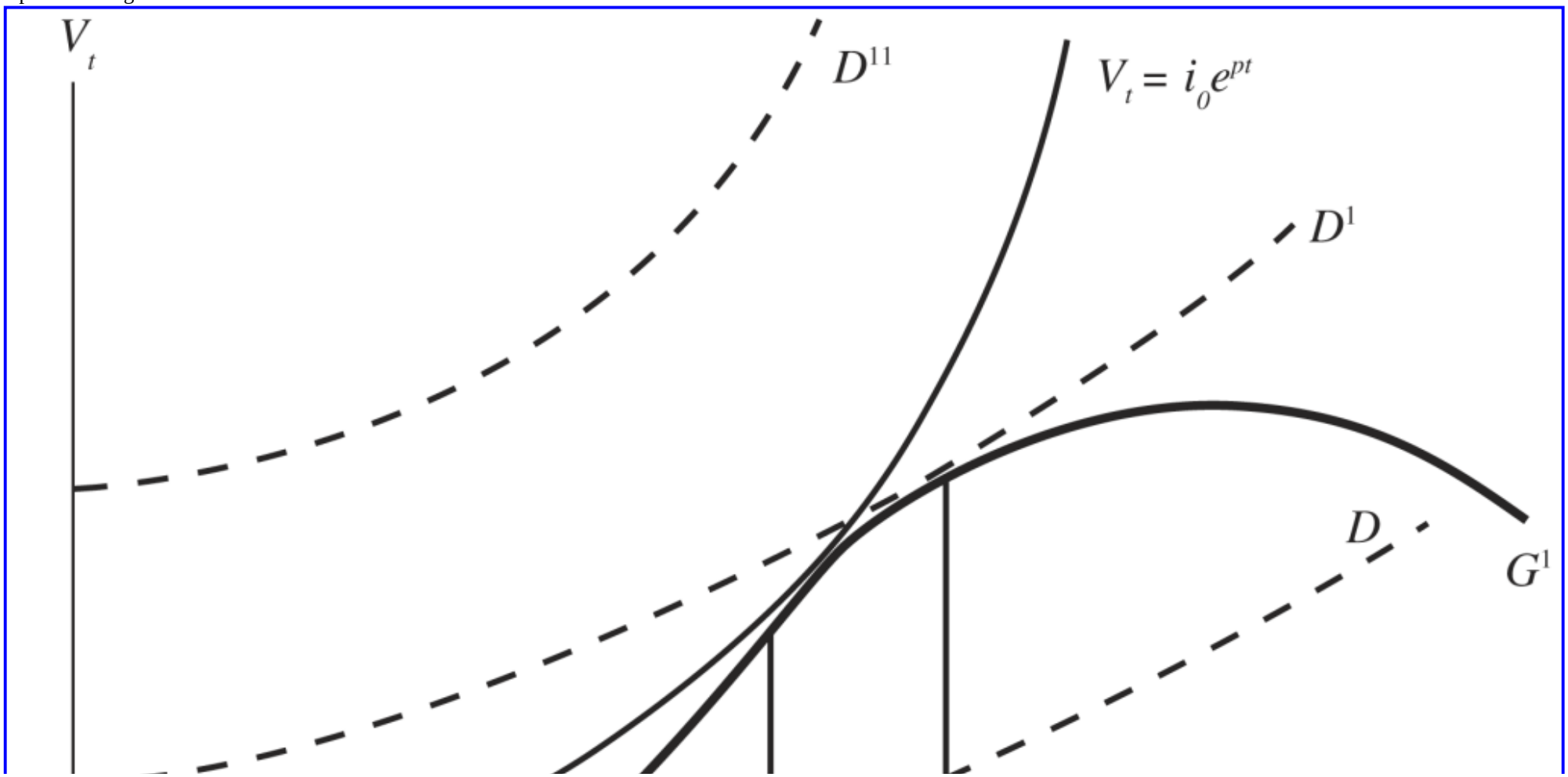

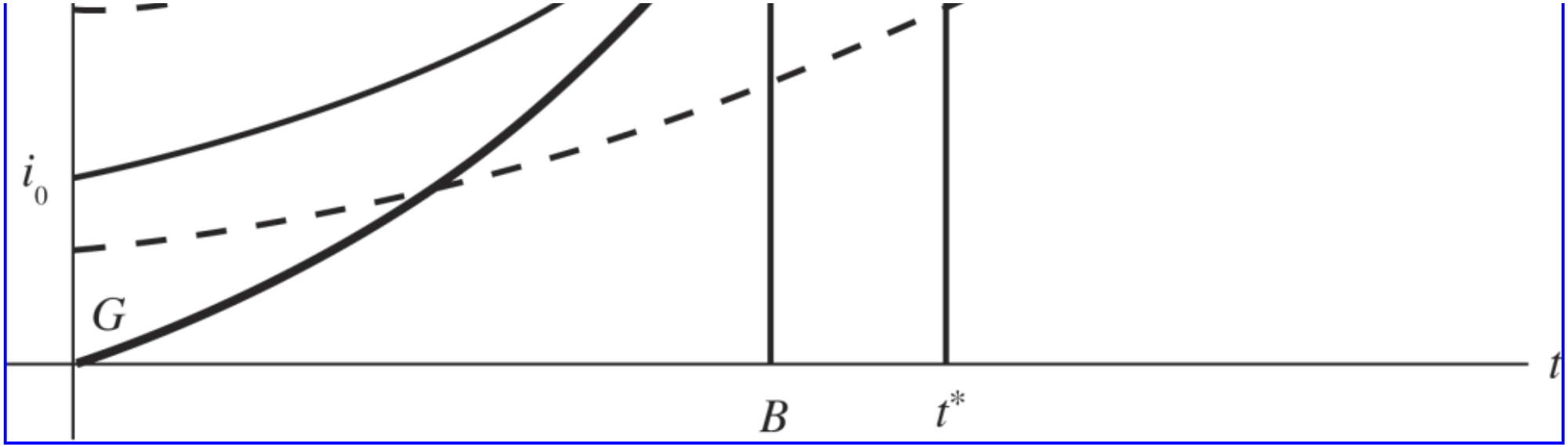

Other Investment Criteria

Certain investment criteria employed in business practice are definitely erroneous. One such is rapidity of ‘payout’ (the date when cash inflows first balance initial outlays), a formula that obviously fails to allow properly for time-discount. Controversy among theorists has centred upon a more interesting concept known variously as the ‘internal rate’ or the ‘rate of return’. The internal rate for a project (or set of projects) is defined as ρ in the discrete discounting equation:

$$0 = z_0 + z_1 / (1 + \rho) + z_2 / (1 + \rho)^2 + \cdots + z_T / (1 + \rho)^T \quad (6)$$

As before the z_t here are the successive terms, positive or negative, of the payments-receipts sequence associated incrementally with a particular project. In the special ‘deepening’ case illustrated in Figure 2, the corresponding concept under continuous compounding is defined implicitly in:

$$0 = -i_0 + V_t e^{-\rho t} \quad (7)$$

where once again the V_t at any date is described by the productive opportunity curve GG'. Under these conditions ρ represents an average compounded rate of growth.
There has been some confusion between two quite different investment decision *rules* that both employ the internal-rate measure ρ: (i) choose projects so as to maximize ρ, versus (ii) adopt projects incrementally so long as $\rho > r$.

Maximum ρ Rule

If the internal rate ρ is interpreted as the average rate of growth, it may seem plausible that the investor should maximize ρ rather than wealth W_0. (Of course, maximizing a growth *rate* would scarcely make sense unless the initial outlay or scale of investment were held constant, which would not in general hold true.) The solution of (7) that maximizes ρ is shown in Figure 2 as $t=B$, notably earlier than the Present-Value solution $t=t^*$.

In favour of B over t^* it has been argued that, if the growth opportunity were to be replicated in perpetuity, returns from choosing the earlier 'rotation period' B must ultimately dominate those associated with cutting on each cycle at t^*. That is certainly true. However, if the decision problem concerns infinite rotation rather than a one-time cutting, for a valid comparison the relevant Present-Value measure would have to be a generalized one that allows for the associated infinite sequence of discounted returns. It can be shown that this generalized Present-Value does coincide with B if the growth opportunity can be reproduced on an ever-broadening scale (e.g. on new land) – but only as funds are freed by cutting the tree or trees. This turns out to be an impossible or uninteresting case, because it implies that the productive opportunity must be of *infinite* market value if the maximized ρ exceeds the market interest rate r (and of zero value otherwise). In contrast, if the opportunity is a unique one which cannot be reproduced after cutting, as pictured in Figure 2, the simple $t=t^*$ solution remains correct. Another solution, $t=F$, found by the German forester Faustmann, is appropriate when the opportunity can be reproduced over time by cutting and replanting but cannot be broadened in scale. F would be found by maximizing the Present Value V_0 of an infinite sequence of rotations, each being a constant-scale replication of the original opportunity. Like all the correct solutions, it is equivalent to maximizing the present worth of the opportunity under the stated assumptions. (F is not shown in Figure 2 but would lie between B and t^*.)

ρ vs. r Comparison Rule

The Comparison Rule says to adopt any project whose internal rate ρ exceeds the market rate of interest r. This rule remains popular in business practice, in part because it offers a convenient division of labour: calculation of the ρ's on individual projects might be delegated to subordinates, while top decision-makers choose the cutoff rate r that corresponds to the relevant market interest rate faced by the firm. Unfortunately, however convenient such a decision of labour may be, once again this is not in general a correct method of project selection.

The difficulty with the Comparison Rule first came to be appreciated when it was discovered that a sequence of positive and negative cash flows could have more than one ρ serving as solution of equation (6) above. A project represented by the annual payments sequence –1, 5, –6, for example, has two solutions: $\rho=1$ and $\rho=2$. (It can be shown that a project with $T+1$ dated elements may have as many as T solutions.) This of course destroys the idea that the internal rate can generally be identified with a growth rate; an outlay of one dollar cannot be said to grow at both 100% and 200%. Various answers have been offered to the puzzle of which ρ to use in such cases. But the difficulty is immediately explained and resolved if we think instead in terms of Present Value. It turns out that the sequence –1, 5, –6 has positive V_0 (and is therefore worth adopting) for any constant market interest rate r *between* 100% and 200%, but at other values of r has negative Present Value (and should not be adopted). Perhaps even more illuminating is the project described by cash flows –1, 3, $-2\frac{1}{2}$. This sequence has no real solution for ρ in equation (6), the reason in Present-Value terms being that V_0 investment opportunity. After all, there is no justification for postulating (as is implicitly done by the Comparison Rule) that the anticipated sequence of market interest rates $r_1, r_2, \ldots, r_T$ must be constant over time (always equal to a common r). It turns out that the cash-flow pattern $-1, 3, -2\frac{1}{2}$ has positive Present Value (i.e., the project would be worth adopting) for many possible non-constant interest-rate sequences – for example, r_1=100% and r_2=200%.

Summing up, therefore, the Present-Value Rule for investment decision – corresponding as it does to the principle of maximizing wealth within the opportunities available – is correct itself and also serves to define the range of validity of all the other rules considered.

Generalizations and Extensions

The preceding analysis needs to be extended in at least two important ways, so as to allow for: (1) uncertainty, and (2) imperfect and incomplete markets.

Uncertainty

Investment choices, involving as they do present sacrifice for future benefit, are peculiarly sensitive to uncertainty. However, so long as we can continue to assume a regime of complete and perfect markets, the Present-Value Rule is robust enough to retain validity even in a world of uncertainty. For, the proximate goal of any individual (or group of individuals organized in a firm or other joint enterprise) will still be to undertake productive activities so as to maximize wealth. Having achieved that goal, each and every individual investor will be in a position to distribute his attained wealth as desired over all possible dated contingencies in accordance with his time-preferences, degree of risk-aversion, and

probability beliefs.

Economists use two main models for the analysis of uncertainty – state-preference and mean-versus-variability analysis. Since the latter, under certain assumptions, can be regarded as a special case of the former, for our purposes attention can be limited to the state-preference model. If markets for state-claims are complete and perfect, any pattern of varying returns over states of the world at a given date has a *certainty-equivalent* in value terms as of that date. In equations (3) and (3′), the z_t for any project can now be interpreted as certainty-equivalents (rather than as simple cash flows) defined by:

$$z_t = P_{t1}z_{t1} + P_{t2}z_{t2} + \cdots + P_{tS}z_{tS} \tag{8}$$

Here z_{ts} represents the cash flow at date t contingent upon state of the world s obtaining – there being S distinguishable such states – while P_{ts} is the price at which a unit claim to income in state s at date t can be converted into (traded for) current certainty income.

Incomplete or imperfect markets

Markets are said to be *incomplete* if some objects of choice are non-tradeable. For example, futures markets for some commodities at far-distant dates do not exist, nor is it possible to trade in claims contingent upon each and every conceivable future uncertain event. Markets are said to be *imperfect* if there are costs of trading – for example, brokerage fees, transaction taxes, or expenses in locating exchange partners. Any real-world regime of markets will necessarily be both incomplete and imperfect, but for some purposes the assumption of complete and perfect markets may be a usable idealization. Unfortunately, once we depart from this idealization the problem of investment decision criteria becomes very difficult. The reason is that the Separation Theorem fails. Only under complete and perfect markets is the concept of wealth or Present-Value unambiguously defined, so that the choice of productive investments can be entirely disconnected from individuals' personal time-preferences, risk-preferences, beliefs etc. Failure of the Separation Theorem particularly subverts the ability of investors to join together in undertaking large projects or groups of projects.

However, two different lines of analytical approach have yielded results of interest. (i) A number of techniques have been devised for locating 'utility-free' or 'efficient' investment choices. In general such techniques cannot determine an optimal project set, but they can serve to filter out options whose payoff patterns over dates and/or states are dominated by other available projects or project combinations. (ii) While investors' personal circumstances may diverge in innumerable ways, there should be some tendency for those similarly situated to group together. Thus, a firm whose investment opportunities yield far-future payoffs should tend to be owned by a 'clientele' consisting of individuals with moderate time-preferences, willing to forego current dividends in the hope of large long-term gain. It follows that unanimity as to the investment choices to be made may after all govern *within* the firm, for example as to the discount rate to employ in calculating Present Value, even in the absence of perfect and complete markets.

See Also

- internal rate of return
- present value

Bibliography

The modern theory of investment and intertemporal choice was set down in classic form by Irving Fisher as part of his great works on interest (1907, chapters 8–9; 1930, chapters 10–13). The seminal works on uncertainty theory include Arrow (1953) for the state-preference approach and Markowitz (1959) and Sharpe (1964) for the mean-versus-variability model. Choice over time and choice under conditions of uncertainty are integrated in the treatise by Hirshleifer (1970) that builds upon these foundations. All these topics have been followed up in an enormous literature, of which only a few illustrative instances can be cited here: on investment decision formulas, Samuelson (1976); on utility-free or dominant choices, Pye (1966), Hanoch and Levy (1969), and DeAngelo (1981). A survey of investment decision criteria used in current business practice appears in Schall, Sundem, and Geijsbeek (1978).

Arrow, K.J. 1953. The role of securities in the optimal allocation of risk-bearing. Reprinted in K.J. Arrow, *Essays in the Theory of Risk-Bearing*, Chicago: Markham, 1971.

DeAngelo, H. 1981. Competition and unanimity. *American Economic Review* 7(1), March, 18–27.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Hanoch, G. and Levy, H. 1969. Efficiency analysis of choices involving risk. *Review of Economic Studies* 36(3), July, 335–46.

Hirshleifer, J. 1970. *Investment, Interest, and Capital*. Englewood Cliffs, NJ: Prentice-Hall.

Markowitz, H.M. 1959. *Portfolio Selection*. New York: Wiley.

Pye, G. 1966. Present values for imperfect capital markets. *Journal of Business* 39(January), 45–51.

Samuelson, P.A. 1976. Economics of forestry in an evolving society. *Economic Inquiry* 14(4), December, 466–92.

Schall, L.D., Sundem, G.L. and Geijsbeek, W.R. 1978. Survey and analysis of capital-budgeting methods. *Journal of Finance* 33(1), March, 281–7.

Sharpe, W.F. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19(3), September, 425–42.

How to cite this article

hirshleifer, Jack. "investment decision criteria." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000215> doi:10.1057/9780230226203.0847

The New Palgrave Dictionary of Economics Online

invisible hand

Mark Blaug
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Adam Smith employed the term ‘invisible hand’ twice in his published writings, and a considerable secondary literature has explored the multiple meanings he intended to convey by the use of this metaphor. I argue that, whatever he did mean, he certainly did not mean that competition or the market mechanism promoted efficiency: instead it promoted the growth of income, even for the poor.

Keywords

Austrian economics; capital accumulation; classical competition; competition; competition as rivalry; Cournot, A.; division of labour; Ferguson, A.; first fundamental theorem of welfare economics; Hume, David; invisible hand; Mandeville, B.; mercantilism; perfect competition vs free competition; productive and unproductive labour; Scottish Enlightenment; Smith, A. on invisible hand; spontaneous order; Srewart, D.; Turgot, A.; welfare economics

Article

The ‘invisible hand’ was a metaphor used by Adam Smith to describe ‘the principle by which a beneficent social order emerged as the unintended consequence of individual human action’. This is Vaughn's succinct summary of Smith's intentions in employing the metaphor (1987, p. 997). More recently, Grampp (2000) has reviewed nine different interpretations of the famous metaphor, concluding that the three references to the invisible hand in Smith's works are not expressions of the same concept, an opinion share by many other commentators.
Smith referred to the ‘invisible hand’ twice in his published writings (there is a third reference in his unpublished ‘Essay on Philosophical Subjects’), and he did so at greatest length in Book IV, Chapter 2, of the *Wealth of Nations*. It is easier to say what he did *not* mean by the invocation to the ‘invisible hand’ than to spell out precisely what he did mean. What he definitely did not mean is the so-called first fundamental theorem of modern textbook welfare economics, although that reading has been frequently

ascribed to him (for example, Arrow and Hahn, 1971, p. 1; Mas-Colell, Whinston and Green, 1995, pp. 308, 327, 524, 545, 549). The first fundamental theorem states that, subject to certain exceptions such as externalities, economies of scale, public goods and imperfect information, every competitive equilibrium is Pareto-optimally efficient. It is indeed possible to find statements in Chapter 2 of the *Wealth of Nations* on mercantilism that appear to endorse something like the first fundamental theorem. Capitalists have a preference for domestic over foreign investment for reasons of security, Smith asserts, but the result of the free movement of capital is nevertheless of benefit to society as a whole:

> As every individual, therefore, endeavours as much as he can both to employ his capital in support of domestic industry, and so to direct that industry that its produce may be of the greatest value; every individual necessarily labours to render the annual revenue of the society as great as he can. He generally, indeed neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this as in many other cases, led by an invisible hand to promote an end which was no part of his intention. (Smith, 1776, pp. 455–6).

The natural interpretation of this passage is, at least for domestic industry, that total product is maximized by free competition. This is almost the first fundamental theorem – but not quite.
First, a presumption of maximization is not a mathematical theorem and, secondly and more significantly, free competition or free unrestricted entry into industries is a far cry from *perfect* competition without which the notion of the price-taking behaviour of numerous, small competitors, adjusting only the quantities they buy or sell, falls to the ground. Cournot invented the concept of perfect competition *de novo* in 1838 and, since the proof of the first fundamental theorem absolutely requires the concept of perfect competition, the idea that Adam Smith somehow stated a primitive version of the first theorem must be wrong; it is in fact a historical travesty.
Adam Smith clearly believed in competition, or rather ‘the simple system of natural liberty’, but his idea of competition was a behavioural one, not defined by the number of firms in the market as in Cournot. Competition, for Smith as for all the classical economists, implied rivalry by price and non-price means, rivalry among consumers bidding for a limited supply and rivalry among producers to dispose of that supply on the most advantageous terms. In other words, he had what I have called a ‘process conception of competition’, nowadays associated with Austrian economics, in contrast to the orthodox conception of economics, in which all the emphasis is directed to the nature of the final equilibrium, regardless of how that final equilibrium is attained (Blaug, 1997, p. 678; see also Coase, 1997, p. 318; Kirzner, 2000).
Although the first theorem cannot be found in the *Wealth of Nations*, what can be found is the notion that competition has desirable properties, namely, that it promotes the rate of growth of national income or what he labelled ‘the wealth of nations’, which results in the material improvement of the standard of living even of the poorest members of society. This idea is not only the mainspring of the famous opening chapter of the book on ‘The Division of Labour’ in the pin factory, but it accounts for the emphasis on capital accumulation and the crucial distinction for Smith's theory of economic growth between ‘productive and unproductive’ labour in Book II, not to mention the content of the whole of

Book III with its revealing title ‘The Different Progress of Opulence in Different Nations’, which translated into modern jargon reads ‘On Differences in the Growth Rate of Different Countries’. Much of Book III is devoted to persuading the reader that there had been material progress since Elizabeth I, a thesis which surprisingly was frequently denied at the time. In short, what was good about what he called ‘the commercial society’ was that it grew rapidly, not that it was efficient, a term and indeed a concept that never appears in the *Wealth of Nations*.

Smith's references to an ‘invisible hand’ in the *Wealth of Nations* have attracted an enormous secondary literature (for example, Hayek, 1973, ch. 2; Vaughn, 1987; Persky, 1989; Grampp, 2000; Rothschild, 2001, pp. 116 ff; Streissler, 2003, Minowitz, 2004; Vivenza, 2005), no doubt because they express three closely connected but separable ideas: (*a*) the *private* actions of individuals can have unforeseen and unintended *social* consequences; (*b*) these private self-interested actions and unintended social consequences may be harmonious in mutually promoting the interests of all members of society; and (*c*) there is an order in these harmonious outcomes as if private self-interested actions were centrally coordinated to produce a coherent overall pattern. This is a profound assembly of ideas that captures the doctrine of ‘spontaneous order’ employed by many thinkers of the Scottish Enlightenment to explain the emergence of such social institutions as language, the law, private property, the monetary system and even the market mechanism itself, not by central design or collective regulation but by individual action undertaken for quite different reasons. It arises most clearly in Adam Ferguson's *Essay on the History of Civil Society* (1767), published a decade before the *Wealth of Nations*, and even earlier in Hume's *Treatise of Human Nature* (1740). But important as the idea of a ‘spontaneous order’ may have been to Ferguson and Hume, as well as to Mandeville, Turgot and Dugald Stewart, it was not actually in the forefront of Adam Smith's thinking and, in any case, he never characterized the price system or even free competition as an ‘invisible hand’. This is a modern reading of Smith under the influence of Walras and Pareto as translated by Arrow and Debreu.

It was only in the last quarter of the 19th century (as a result of German critics of Smith) that the phrase ‘invisible hand’, which after all occurs only once in the *Wealth of Nations*, was elevated to a proposition of profound significance. Rothschild deals expertly with the subject and concludes that ‘the image of the invisible hand is best interpreted as a mild ironic joke’ (2001, p. 116). This may be going a little too far in the opposite direction to the now prevailing interpretation, but there is no doubt that Smith himself did not attach great importance to the idea of an invisible agency channelling the behaviour of self-interested individuals and instead regarded the metaphor of the invisible hand as a sardonic, if not ironic, comment on the self-deception of all of us, including moral philosophers.

Support for this view of his intentions is found in the one reference to the ‘invisible hand’ in *The Theory of Moral Sentiments*, a reference that is frequently ignored in the exegetical literature on Smith. In that passage in the *Theory of Moral Sentiments* (1759, 184–5) Smith argues that mankind has progressed in the face of pronounced and persistent inequalities and that the rich, despite their natural selfishness, end up unintentionally sharing their wealth with the poor, who for their part end up no worse than the rich themselves. Both Grampp and Minowitz, alone among all the Smithian commentators, object to this conclusion as too Panglossian. Be that as it may, this passage soon dispels the belief that Smith meant one thing and one thing only by the metaphor of ‘the invisible hand’.

The notion of a spontaneous order in the sense of a self-regulating system accounting for the existence of economic institution went underground after the Scottish Enlightenment, and references to an ‘invisible hand’ are rarely encountered in any of the classical economists, although the idea that

economics studies an underlying invisible reality beneath the surface appearance of a free market economy continued to dominate the thinking of Ricardo, J.S. Mill and particularly Karl Marx.

See Also

- competition
- Cournot, Antoine Augustin
- division of labour
- efficient allocation
- Hayek, Friedrich August von
- Hume, David
- Mandeville, Bernard
- Marx, Karl Heinrich
- Pareto, Vilfredo
- spontaneous order

Bibliography

Arrow, K. and Hahn, F. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Blaug, M. 1997. *Economic Theory in Retrospect*, 5th edn. Cambridge: Cambridge University Press.

Coase, A. 1997. The wealth of nations. *Economic Inquiry* 15, 309–25.

Grampp, W. 2000. What did Adam Smith mean by the invisible hand? *Journal of Political Economy* 108, 441–65.

Hayek, F. 1973. *Law, Legislation and Liberty. Volume 1: Rules and Order*. London: Routledge & Kegan Paul.

Kirzner, I. 2000. Competition and the market process: some doctrinal milestones. In *The Process of Competition*, ed. J. Kraft. Cheltenham: Edward Elgar.

Mas-Colell, A., Whinston, M. and Green, R. 1995. *Microeconomic Theory*. New York: Oxford University Press.

Minowitz, P. 2004. Adam Smith's invisible hands. *Econ Journal Watch* 1, 381–412.

Persky, J. 1989. Retrospectives: Adam Smith's invisible hands. *Journal of Economic Perspectives* 3(4), 195–201.

Rothschild, E. 2001. *Economic Sentiments: Adam Smith, Condorcet and the Enlightenment*. Cambridge, MA: Harvard University Press.

Smith, A. 1759. *The Theory of Moral Sentiments*, ed. D. Raphael and A. Macfie. Oxford: Clarendon Press, 1976.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R. Campbell, A. Skinner and W. Todd. Oxford: Clarendon Press, 1976.

Streissler, E. 2003. Adam Smith's ultimate invisible hand: content and rhetoric. *History of Economic Thought Newsletter* 8, 3–15.

Vaughn, K. 1987. Invisible hand. In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate and P. Newman. Basingstoke: Palgrave.

Vivenza, G. 2005. The agent, the actor, and the spectator. *History of Economic Ideas* 13, 37–56.

How to cite this article

Blaug, Mark. "invisible hand." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000220> doi:10.1057/9780230226203.0849

The New Palgrave Dictionary of Economics Online

involuntary unemployment

John B. Taylor
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand; efficiency wages; frictional unemployment; incentive wages; involuntary unemployment; labour supply; marginal rate of substitution; matching; minimum wages; natural rate of unemployment; nominal wages; non-accelerating inflation rate of unemployment (NAIRU); optimal contracts; real business cycles; real wages; search theory; staggered wage setting; sticky wages; unemployment; unemployment measurement; wage differentials; wage rigidity

Article

The most common and analytically useful definition of involuntary unemployment is based on the labour supply curve: if workers are off the labour supply curve – so that there is an excess supply of labour at the current real wage – then, by definition, there is involuntary unemployment. The amount of involuntary unemployment is equal to the amount of excess labour supply. If workers are on the labour supply curve, then, by definition, there is no involuntary unemployment. One could analogously define involuntary overemployment as a situation of insufficient supply of labour at the prevailing real wage (as may occur during wartime with wage and price controls), but the term is seldom used.
In a static, deterministic, utility maximization framework, the labour supply curve is simply the set of real wage and employment pairs for which the marginal rate of substitution of income for leisure is equal to the current real wage. Hence, involuntary unemployment can be equivalently defined using the utility function: if the real wage is greater than the marginal rate of substitution of income for leisure, then, by definition, there is involuntary unemployment. If the marginal rate of substitution of income for leisure is equal to the real wage, then there is no involuntary unemployment.

Historical examples of usage

This definition of involuntary unemployment is very close to that used by Keynes (1936). In Chapter 2 of the *General Theory*, Keynes writes '… the equality of the real wage to the marginal disutility of

employment … corresponds to the absence of “involuntary” unemployment’ (p. 15). (Keynes makes the simplification that the marginal utility of income is constant, so that the marginal disutility of employment is the same as the marginal rate of substitution of income for leisure.) Keynes excluded frictional unemployment from involuntary unemployment. However, it is important to note the Keynes also excluded unemployment ‘due to the refusal or inability of a unit of labour, as a result of legislation or social practices or of a combination for collective bargaining or of a slow response to change or of mere human obstinacy, to accept a reward corresponding to the value of the product attributable to its marginal productivity’ (Keynes, 1936, p. 6). Thus, Keynes chose to exclude union wage differentials as well as minimum wage legislation as sources of involuntary unemployment. Clearly, Keynes wanted to focus on a particular type of involuntary unemployment.

Patinkin (1965, ch. 13) also used the static labour supply definition in his well-known analysis of involuntary unemployment:

> The norm of reference to be used in defining involuntary unemployment is the supply curve for labor … as long as workers are ‘on their labor supply curve’ – that is, as long as they succeed in selling all the labor they want to at the prevailing real wage rate – a state of full employment will be said to exist in the economy. (pp. 314–15)

Although Keynes developed and emphasized the idea of involuntary unemployment much more than economists had done before, the above definition based on the labour supply curve predates Keynes writings. In fact it was used by the ‘classical’ economists. For example, in 1914 Pigou proposed measuring involuntary unemployment of a group of persons by the number of hours’ work by which employment ‘… falls short of the number of hours’ work that these persons would have been willing to provide at the current rate of wages under current conditions of employment’ (see Casson, 1983, p. 39). According to Keynes, however, classical theories (such as Pigou's) did not admit the possibility of involuntary unemployment. Unemployment of a particular group caused by union wage differentials or minimum wage legislation was admitted by the classical theory, but as mentioned above Keynes chose to classify this as voluntary.

Criticisms of the definition of involuntary unemployment

Despite the analytical simplicity of the above definition based on labour supply, the term involuntary unemployment has resulted in many critiques and controversies. One of the criticisms stems from simple conflicts between the above technical definition and everyday non-technical usage of the term involuntary. For example, Fellner (1976) wrote, ‘… distinguishing elements of voluntariness from elements of involuntariness in the unemployment problem is a hopeless endeavour …’ (p. 134) and that ‘Keynes’ definition is unhelpful and so are all variants inspired by that definition’ (p. 53). Fellner and others have been concerned that one can never determine the intentions of a given unemployed person so that the broad classification of unemployment into involuntary and voluntary is meaningless. Although the many connotations of the term involuntary may cause semantic difficulties (as may other concepts in economics such as ‘rational’ or ‘marginal’), focusing on the technical definition given above would seem to avoid these difficulties.

A second criticism arises in the practical use of the concept of involuntary unemployment for public policy. From the above definition, one criterion of good macroeconomic performance would be zero, or very small, involuntary unemployment. (Strictly speaking, this is true only if the measured real wage is equal to the marginal productivity of labour, an equality that might not hold if optimal contracts of the type described below are important in the economy.) Since government unemployment statistics are commonly taken as an indicator of economic performance, one might hope that measured unemployment could be related to the concept of involuntary unemployment. However, this is very difficult and any attempt is bound to be criticized. Government unemployment statistics typically attempt to measure the number of unemployed who are looking for work, but who have not yet found work. However, aside from the problem of determining whether someone is looking for work, or how intensively, unemployment statistics obviously include frictional unemployment and other types of unemployment that would not be included as involuntary according to the above definition. Even in a condition of relatively full employment, there exists some ‘normal’ unemployment, which government statistics need to be corrected for. Milton Friedman (1968) used the term ‘natural’ unemployment for the amount of unemployment that would exist, without excess supply, in equilibrium after wages and prices have adjusted. Another concept of normal unemployment is the non-accelerating inflation rate of unemployment (NAIRU), defined as the amount of unemployment that would exist when there is no tendency for wage or price inflation to rise or fall. Measuring the ‘natural’ rate or NAIRU in practice entails looking for an unemployment rate for which inflationary pressures are small and adjusting this rate for known changes in the demographic characteristics of the labour force. The natural rate of unemployment is not a constant, however, and these measurements have considerable error. Nevertheless, a practical alternative to involuntary unemployment as a measure of economic performance is the difference between the actual unemployment rate and the natural unemployment rate. For policy purposes, this may serve as a reasonably close approximation to involuntary unemployment, but clearly it is a different concept. In particular, note that this measure can be negative, as when the unemployment rate falls below the natural rate in boom times. Fellner (1976) suggested focusing on this measure and hence on inflation stability, rather than on involuntary unemployment, and he argued that demand management (monetary and fiscal policy) should promote the maximum amount of employment that can be achieved without inflation instability. This measure is also the criterion used in stabilization studies that characterize a macroeconomic trade-off in terms of the fluctuations of unemployment about the natural rate versus the fluctuations in inflation (see Taylor, 1980).

A third reason for criticism of the term involuntary unemployment is that the standard definition is essentially static and deterministic. In fact, the static, deterministic labour supply and demand model does not admit an explicit theory of frictional or natural unemployment. Without such a model it is difficult even to discuss whether a given level of unemployment is voluntary or optimal or not. Research on the microfoundations of unemployment (see for example Phelps et al., 1970), had as a major goal the development of a model of equilibrium unemployment – using search and matching theory. Some search models generated unemployment that was Pareto optimal (see Lucas and Prescott, 1974, for example), but others included trading externalities and generated unemployment which could be non-optimal (see Diamond, 1982, for example). While not yet definitive, at the least this research shows that for many public policy questions it is necessary to go beyond the simplest model of labour supply, and thereby beyond the simple definition of involuntary unemployment.

In the *General Theory* Keynes presented a more convoluted definition of involuntary unemployment, and this has been a fourth source of controversy. According to Keynes (1936, p. 15),

> Men are involuntarily unemployed if, in the event of a small rise in the price of wage-goods relatively to the money-wage, both the aggregate supply of labour willing to work for the current money-wage and the aggregate demand for it at that wage would be greater than the existing volume of employment.

One can clearly envisage a point off the labour supply curve from this definition. However, there is much more. Embedded in the definition of involuntary unemployment are some of Keynes's other ideas that were part of his *theory* of involuntary unemployment, but logically distinct from the *definition* of involuntary unemployment. Within the definition it is noted that workers would be willing and able to have a reduction in their real wage (and still increase their work) if it occurred through an increase in the price level, but not if it occurred through a decline in the nominal wage. This ‘stickiness’ of nominal wages, which is generated as part of the market mechanism, is of course crucial to Keynes's theory. Also embedded in the definition is the assumption that firms are in their labour demand curve, so that a lower real wage would stimulate unemployment, an idea that is much less crucial for Keynes's ideas, as Leijonhufvud (1968) has emphasized. Why did Keynes emphasize this convoluted definition of involuntary unemployment? It seems clear that he wanted to highlight the crucial difference between his theory of unemployment and what he called classical theory. This difference centred on the inability, given the way labour markets and the whole economy interact, of individual workers to reduce unemployment simply by reducing nominal wages. As indicated above, Pigou based the definition of involuntary unemployment on the labour supply curve in much the same way that Keynes did, but the classical reason for its existence – simply that real wages were too high – was much different from the theory of deficient aggregate demand put forth by Keynes. In retrospect Keynes would have added clarity to his discussion by unbundling his theory and his definition of involuntary unemployment.

Implications of recent technical research for the concept of involuntary unemployment

Five research developments since the 1960s have had great relevance for the concept of involuntary unemployment: equilibrium macroeconomics, optimal contract theory, disequilibrium macroeconomics, efficiency or incentive wage theory, and staggered-wage setting theory. However, this relevance must be inferred from the research, because the term involuntary unemployment is seldom used explicitly, and perhaps avoided by many recent researchers.

Equilibrium macroeconomics

One strand of research macroeconomics has established a strategy of trying to explain the observed fluctuations in unemployment by equilibrium models in which workers are always on their labour supply curves. Wages and prices are perfectly flexible, and all markets clear in these models. Lucas and Rapping (1969) and Kydland and Prescott (1982) represent some of the seminal work in this strand of research. Clearly if these models turn out to be successful and to dominate other models, then the idea of

involuntary unemployment would become useless for macroeconomics. Shifts of the labour supply curve – caused by intertemporal substitution of labour supply in response to temporary actual or perceived fluctuations in the real wage – are the main source of employment variability in these models. Research in this area is continuing and branching out into ‘real business cycle’ theory which ignores monetary factors in the cycle altogether. It appears, however, that a very high labour supply elasticity – by the standards of recent microeconomic empirical research (see MaCurdy, 1981) – is required for these models to be able to explain the observed fluctuations in employment.

Optimal contract theory

Studies by Azariadis (1975), Baily (1974) and others attempted to explain why involuntary unemployment would arise when there exist optimal contracts between firms and workers stipulating for fixed wage payments. However, when firms and workers have equal access to information, these studies have shown that, in the relevant sense, involuntary unemployment does not exist despite the fixed wage bill. In these optimal contract models the marginal rate of substitution of income for leisure is equal to the marginal productivity of labour – the condition for the optimality – in all possible states. Although workers are off their labour supply curve *ex post* (since the real wage is not necessarily equal to the marginal rate of substitution), this discrepancy has no welfare significance. Models in which firms have more information than workers about the nature of the shock can lead to a breakdown in the marginal conditions for optimality, but unless firms are more risk averse than workers the result is involuntary *over*-employment: the marginal productivity of labour is less than the marginal rate of substitution of income for leisure (see Green and Kahn, 1983, and Grossman and Hart, 1983). Viewed as an attempt to explain involuntary unemployment this research, therefore, has been unsuccessful. Taken literally, it shows that much of the unemployment that may have appeared as involuntary is, in fact, voluntary or at least efficient!

Disequilibrium theory

Malinvaud's (1977) careful examination of fixprice multimarket equilibria, following the tradition of Clower (1965) and Barro and Grossman (1971), has greatly helped to clarify the conceptual difference between Keynes's explanation of involuntary unemployment due to insufficient aggregate demand (where firms are constrained in product markets), and the classical unemployment associated with the real wage being too high (where firms are not constrained in product markets). This research also has had considerable policy relevance in the early 1980s because the high rates of unemployment in western Europe were diagnosed as classical rather than Keynesian by many economists.

Efficiency or Incentive Wages

Calvo (1979) and others have argued that involuntary unemployment can occur because high wages must be paid to give workers the incentive to work hard, to be productive, and not to shirk. As firms attempt to bid up their wages relative to other firms, an equilibrium is reached with all firms paying more than the wage in the absence of incentive effects and with involuntary unemployment: an excess

supply of labour with unemployed workers willing to work at the going wage. This type of unemployment is not of the deficient demand type emphasized by Keynes, and given Keynes's willingness to lump other minimum wage unemployment in with frictional unemployment, it is likely that Keynes would have classified this type of unemployment as voluntary. Incentive wages would increase the normal unemployment (natural or NAIRU) rate, but there is little empirical evidence of how quantitatively important the effect is.

Staggered-wage setting theory

In these models (see Taylor, 1980, for example), wages are set with an aim to maintain relative wages unless there is a reason for relative wages to adjust. This relative wage setting leads average nominal wages to adjust with a lag described by a predictable dynamics to changes in demand. In these models prices are set as a markup over wages, and for this reason aggregate prices are almost as sticky as nominal wages. Combined with an elementary model of aggregate demand and an aggregate demand policy that does not fully accommodate inflation, these models are designed to be compared directly with the data and in fact lead to fluctuations in unemployment which have features similar to the real world. The unemployment in these models comes close to the usual definition of involuntary unemployment, but since explaining empirical regularities is a primary objective, unemployment enters the model directly as the deviation of unemployment from the natural rate – a more readily measurable quantity than involuntary unemployment. These models show that wage rigidities need not be very long to generate the type of fluctuations in unemployment that characterize the business cycle. Like the equilibrium models discussed above, and unlike the other three research developments described above, these models are dynamic and can therefore be directly tested against time series data.
Although there has been a tendency for much recent research to avoid the term involuntary unemployment, and instead to define unemployment as appropriate to the theoretical or empirical objectives of the research itself, the term involuntary unemployment will probably continue to be used. Despite the criticism and controversy discussed above there is little harm in this usage, as long as the technical definition is emphasized. Its usage may encourage researchers to point out the connection of new results to past achievements.

See Also

- implicit contracts
- natural rate of unemployment
- search theory
- unemployment

Bibliography

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83, 1183–202.

Baily, M.N. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41, 37–50.

Barro, R.J. and Grossman, H. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

Calvo, G. 1979. Quasi-Walrasian theories of unemployment. *American Economic Review, Papers and Proceedings* 69, 102–7.

Casson, M. 1983. *Economics of Unemployment*. Oxford: Martin Robertson.

Clower, R. 1965. The Keynesian counter-revolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Diamond, P. 1982. Aggregate demand management in search equilibrium. *Journal of Political Economy* 90, 881–94.

Fellner, W. 1976. *Towards a Reconstruction of Macroeconomics: Problems of Theory and Policy*. Washington, DC: American Enterprise Institute.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Green, J. and Kahn, C. 1983. Wage employment contracts. *Quarterly Journal of Economics* 98 (Supplement), 173–87.

Grossman, S.J. and Hart, O. 1983. Implicit contracts, moral hazard, and unemployment. *American Economic Review* 71, 301–7.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kydland, F. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press.

Lucas, R.E., Jr. and Prescott, E.C. 1974. Equilibrium search and unemployment. *Journal of Economic Theory* 7, 188–209.

Lucas, R.E., Jr. and Rapping, L. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77, 721–54.

MaCurdy, T.E. 1981. An empirical model of labour supply in a life cycle setting. *Journal of Political Economy* 89, 1059–85.

Malinvaud, E. 1977. *The Theory of Unemployment Reconsidered*. Oxford: Basil Blackwell.

Patinkin, D. 1965. *Money Interest and Prices*, 2nd edn. New York: Harper & Row.

Phelps, E.S. et al., eds. 1970. *Microeconomic Foundations of Employment and Inflation Theory*. New York: W.W. Norton.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

How to cite this article

Taylor, John B. "involuntary unemployment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000222> doi:10.1057/9780230226203.0850

The New Palgrave Dictionary of Economics Online

Ireland, economics in

Tom Boylan and Renee Prendergast

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This contribution provides an overview of Irish economic thought from the 17th to the late 20 century. Broadly speaking, the pioneering contributions of the 17th and 18th centuries were concerned with the issue of improvement or economic development. The 19th century saw the formal institutionalization of political economy in Ireland and seminal contributions to value, distribution, and international trade theory in addition to work on public finance and methodology. The achievement of independence in the 20th century led to new concerns with development and policy experiments aimed at promoting lasting growth.

Keywords

Anglo–Irish Free Trade Agreement; Bastable, C.F.; Bentham, J.; Berkeley, Bishop G.; budget deficits; Busteed, J.; Butt, I.; Cairncross, A. K.; Cairnes, J. E.; Cantillon, R.; Carter, C.; catch-up; classical economists; Cliffe Leslie, T. E.; Common Agricultural Policy (EU); concentration; cost-of-production theory of value; demand schedule; development; distribution theory; Dublin School; Duncan, G.; Dutch disease; European Economic Community; European Monetary Union; factor immobility; factor pricing; fiduciary credit system; foreign direct investment; free trade; Geary, R. C.; Great Depression; Griffith, A.; Hancock, W. N.; Hearn, W. E.; Historical School; Hutcheson, F.; imperfect competition; induction; infant-industry protection; Ingram, J. K.; innovation; international trade; Ireland, economics in; Johnston, J.; Keynes, J. M.; Keynesianism; Kiernan, T. J.; land question (Ireland); Lawson, J. A.; List, F.; Longfield, M.; luxury; Lynch, P.; Malthus, T. R.; Mandeville, B.; marginal revolution; Mill, J. S.; Mill–Bastable condition; monetarism; national champions; non-competing groups; O'Brien, G.; Oldham, C. H.; Owen, R.; partnership; paper money; Petty, W.; planning; protection; rent; research and development; Ricardian socialists; Ryan, L.; self-sufficiency; statistics and economics; subjective theory of value; Swift, J.; tariffs; tax incentives; technology policy; terms of trade; Thompson, W.; utilitarianism; Whately, R.

Article

The 17th and 18th centuries

William Petty and Richard Cantillon are commonly regarded as the founders of classical political economy. Both had connections with Ireland. Petty, English by birth, came to Ireland with the Cromwellian army in 1652 and became interested in ‘political anatomy’ in the course of surveying the country in preparation for the confiscation of Irish lands. Cantillon was born in Ireland but spent his adult life in as a banker in Paris, where he wrote what many regard as the first systematic treatise on economics. Despite his nationality and his importance, Cantillon's work is not considered here. It was not written in Ireland; it was neither inspired by Irish conditions nor known to contemporaries living in Ireland.

Political Anatomy of Ireland, written in 1671–2, was Petty's first attempt to uncover the symmetry, fabric and proportion of the body politic by means of political arithmetic. Like all of Petty's writings, it contains pregnant theoretical suggestions, but our interest here is in its systematic approach to economic development. Petty sought to identify Ireland's development potential by considering the distribution and value of land and by estimating the number of ‘spare hands’ who could potentially add to local or universal (tradable) wealth. Petty identified the main causes of Irish underdevelopment as constraints on Ireland's trade with England and the plantations, insufficient coin, underdeveloped consumption patterns, perceived illegitimacy of rulers, rent-seeking and low population density. Petty's proposed remedies as set out in a *Report of the Council of Trade in Ireland*, 25 March 1676, included regularization of money, restoration of trade with the plantations and (particularly the cattle trade) with England, a bank based on landed property as security, reformation of the housing of the poor, legislative union with England and later the transportation of large numbers from Ireland to England. Despite a reference to discountenancing the use of certain foreign commodities in the report, Petty seems not to have favoured protection, arguing in *Political Anatomy* that the proceeds of exports would be more than sufficient to pay for imported products.

Partly as a result of prohibitions on the export of live cattle to England, farmers turned their attention to sheep, with the result that towards the end of the 17th century Irish wool and woollen yarn were among its most important exports. This promising development was nipped in the bud by restrictions introduced under the Wool Acts of 1698–9. This added to the fragility of an already weak economy, resulting in widespread poverty and unemployment in the early decades of the 18th century. Despite their confused and somewhat contradictory Irishness, the new generation of planter stock, including the likes of Prior, Dobbs, Browne, Molesworth, Hutcheson, Swift and Berkeley, responded with a steady stream of pamphlets advancing various proposals for improvement. These included increased agricultural investment, drainage and reclamation of bogs, improvement of inland waterways, encouragement of sea fisheries, mining and manufacturing, the setting up of a mint, consumption of locally produced goods, taxation of absentee rents, removal of restrictions on foreign trade and deportation of the undeserving poor (Kelly, 1991). The main differences were between those such as Browne and Berkeley, who were relatively positive about Ireland's development prospects, and those such as Swift, who believed that plausible sources of improvement had little realistic chance of being implemented by those with the power to do so. Swift's position was vigorously expressed in *A Modest Proposal* (1729), a powerful

satire on the pamphlet literature of his own time and one of the most telling critiques of positive economics ever to have been written anywhere.

While most authors emphasized the need to remove the constraints on trade, Berkeley argued that it would be more prudent to concentrate on those branches which were permitted, including Ireland's domestic trade (Berkeley, 1752). Development would be possible even if the country were surrounded by a wall of brass. This, however, would require the substitution of domestically produced goods for the imported luxuries consumed by the elite as well as an expansion of the wants of the poorer classes in order to make them industrious. An argument for the reform of consumption patterns had already been made in 1726 by Francis Hutcheson in his ‘Remarks upon the Fable of the Bees’ in the course of controverting Mandeville's claim that luxury and vice were inseparable from economic development. Berkeley, who was also an implacable adversary of Mandeville, was even more emphatic than Hutcheson in his opposition to luxury, and he showed himself willing to contemplate sumptuary laws to achieve this objective. While Berkeley's proposals for development on the basis of the domestic market were innovative at the time, his most radical proposals related to the adoption of paper money and the setting up of a national bank. Real wealth, Berkeley argued, consisted not in gold or silver but in the plenty of the necessaries and comforts of life and the power to command the industry of others. Money was simply a ticket or a counter for conveying or recording such power. As such, paper money and bank deposits were perfectly adequate and had some advantages over coin. The ruinous effects of the Mississippi and South Sea schemes were not due to paper money as such but to its use for speculative purposes rather than as a catalyst of industry. Private banks being subject to frauds and hazards, Berkeley proposed the setting up of a public bank, which he assumed would not suffer from these disabilities. The radicalism of Berkeley's position can be appreciated if we bear in mind that support for a fiduciary credit system as opposed to metallic money was in his time very much a minority view and remained so until recently (Murphy, 2000).

The recovery of the Irish economy which took place in the second half of the 18th century was partly due to the success of the linen industry, which had been encouraged as a replacement for wool, and partly to the gradual weakening of commercial restrictions as Britain's population grew and Ireland became an important source of food and agricultural raw materials. During a brief period of legislative independence from 1782 to 1800, the Irish Parliament took steps to encourage domestic industry with various protective measures. It also introduced a corn law to encourage corn production for the British market.

The 19th century

Following the Act of Union in 1801, Ireland was assimilated into the administrative and political jurisdiction of the United Kingdom. Many of its newly established industries went into gradual decline and corn production became a major source of employment. Population increased and with it poverty culminating in the Great Famine of 1845–50. These conditions influenced developments in political economy in Ireland and elsewhere. The need to counter the argument that high rents were a major cause of Irish poverty was a catalyst for the development of Malthus's rent theory (Prendergast, 1987). The attention devoted by John Stuart Mill to the incentive effects of different forms of land tenure was partly a response to Irish land conditions. The scale of the human tragedy of the Irish famine influenced the perception and standing of laissez-faire political economy in Ireland and elsewhere. Irish economists

became pioneers of the Historical School, which emphasized the specificity of time and place.

The formal institutionalization of political economy in Ireland began with the establishment of the Whately Chair in Trinity College in 1832. The chair was funded by Richard Whately, the Protestant Archbishop of Dublin, who came to Ireland from Oxford in 1831. The chair was part of a larger crusade by Whately to promote the dissemination of political economy with a view to encouraging more economically responsible behaviour. The Whately chair was to be filled by a number of outstanding occupants, which included Mountifort Longfield, John Elliot Cairnes and Charles Bastable. Chairs in jurisprudence and political economy were also established in the new Queen's colleges set up in at Belfast, Cork and Galway in 1845 (Boylan and Foley, 1993). Outside of the universities, the principal institutional development was the founding in 1847 of the Dublin Statistical Society, later the Statistical and Social Inquiry Society of Ireland, which had Whately as its first president (Daly, 1997). The society aimed at ‘promoting the study of Statistical and Economical Science’ and its participants included the academic, administrative and professional elite of Irish society. By the mid-19th century an extensive institutional infrastructure for the teaching and dissemination of political economy was in place (Boylan and Foley, 1992).

Irish political economists in the 19th century made original contributions to a number of theoretical areas within the discipline. In value theory, the seminal contribution of Longfield, the first holder of the Whately Chair, has received considerable attention and is recognized as providing one of the earliest attempts at formulating a subjective theory of value (Moss, 1976). A number of Longfield's immediate successors, including Isaac Butt, James Anthony Lawson and William Neilson Hancock, also subscribed, albeit in a limited way, to a subjective theory of value, which led R.D.C. Black (1945) to suggest that the early Whately professors constituted a ‘Dublin school’ of subjective value theorists who anticipated by 30 years the marginal revolution of the 1870s. Longfield's contribution contained in his *Lectures on Political Economy* (1834) is by far the most original offering reflecting his disagreement with the dominant Ricardian framework of analysis in value and distribution theory.

Longfield approached value and distribution as pricing problems. The theory of value, in which commodity prices were determined in markets by supply and demand, was at the centre of his analysis. Longfield did not neglect the influence of cost on market price through changes in supply, but his main emphasis was on demand. The concept of a demand schedule was introduced, in which market demand was conceived as a ranking of individual demands according to their intensity, where ‘the market price is measured by that demand, which being of the least intensity, yet leads to actual purchases’. Longfield invoked the concept of the individual's demand schedule as being composed of ‘several demands of different degrees of intensity’ (Longfield, 1834, pp. 113, 114). This is now interpreted as a seminal statement, foreshadowing the principle of marginal utility that was to find its more formal articulation in the marginalist writers of the 1870s.

Though not a member of the Dublin ‘school’, William Edward Hearn's *Plutology: Or the Theory of the Efforts to Satisfy Human Wants* (1863) was a significant contribution to the debate on the subjective theory of value. Hearn was appointed the first Professor of the Greek Language in Queen's College Galway in 1849. He left Galway in 1854 and became Australia's first professor of economics (Boylan and Foley, 1984b). *Plutology* contained an extended and sophisticated taxonomy of the different kinds and degrees of human wants. Hearn went on to examine how demand could influence the impact of changes in the cost of production for different kinds of commodities; he distinguished between the

demand for essential commodities or ‘necessities’ and non-essential or ‘superfluities’. In this analysis Hearn provided a valuable extension to Longfield's earlier contribution, which was well regarded by contemporaries and later writers including Jevons and Marshall.

If Longfield and the Dublin ‘school’ represented an anti-Ricardian position in value and distribution, the Ricardian tradition was powerfully represented by Cairnes, the sixth holder of the Whately Chair at Trinity from 1856 to 1861 and Professor of Jurisprudence and Political Economy at Queen's College Galway from 1859 to 1870. Cairnes was arguably the most distinguished of the 19th-century Irish academic economists, and contributed to a number of areas of economic theory and contemporary policy issues. Cairnes was a close personal friend of J. S. Mill and was strongly influenced by Mill's analysis, but he produced a more complicated version of the theory of value than Mill. In *Some Leading Principles of Political Economy Newly Expounded* (1874), Cairnes provided a cost-of-production theory of value. But it is clear not only that Cairnes's ‘normal value’ is to be identified as cost of production, but that cost should be interpreted as real cost or sacrifice. In the course of his analysis he made the innovative move of applying Mill's proposition of the determination of international values by reciprocal demand in the case of factor immobility between countries to the internal economy of a country. In the latter situation, the existence of internal factor immobility gave rise to what is arguably Cairnes's most original application of the concept of non-competing groups. Cairnes's tenure in the Whately Chair broke the intellectual continuity of the Dublin ‘school’ by virtue of his commitment to the Ricardo–Mill approach.

In the domain of distribution theory one of the most interesting contributions was made by William Thompson (1775–1833), an Owenite and supporter of the French Revolution. Thompson pursued the aim of formulating an alternative economic system based on the rights of the primary producer. He emerged as the most analytical and original thinker of the Owenite movement, which later became identified with the Ricardian socialists. Thompson was a personal friend of Bentham and it has been argued that Thompson's originality as a thinker consisted in his appropriation of the greatest happiness principle as a basis for fundamental social reform (Duddy, 2002). While radical utilitarianism provided him with a critical component of his rationale for social reform, it was the adoption of Owen's system of mutual cooperation by Thompson, as a model of social organization, that would deliver to individual primary producers the fruits of their labour, which was fundamental to the Ricardian Socialists’ doctrine.

In contrast to the Irish contributions to the Ricardian tradition of distribution theory, Longfield was forging a rather different approach in his *Lectures on Political Economy* of 1834. As Moss (1976) has argued, if the classical economists found the unifying principle for their theories of distribution in the concept of cost of production on the supply side, then Longfield could be said to have discovered his unifying principle of factor pricing in his supply and demand analysis. His identification of the role of marginal demand in the commodity market and marginal productivity in the factor market, justifies Longfield's claim as one of the leading progenitors of the neo-classical marginal theory of commodity and factor pricing.

In the area of international trade, the originality of Irish economists matched their contribution to value and distribution theory. In his *Three Lectures on Commerce and One on Absenteeism* (1835), Longfield extended the theory of comparative cost in significant directions, including the addition of both the multi-commodity and multi-factor case. He also addressed the issue of the incidence of tariffs and traced their effects on the relative price ratios between trading countries. Isaac Butt, Longfield's successor in the Whately Chair, considered the case for protection in his *Protection to Home Industry: Some Cases of Its*

Advantages Considered (1846). This work, which was influenced by conditions in Ireland, was both methodologically engaging and analytically perceptive in its assessment of the benefits and weaknesses of protection in the context of particular circumstances.

Cairnes's reputation in the area of international trade rests on his systematic integration of the concept of non-competing groups into his analysis. This allowed him to distinguish between the role played by costs of production in determining international prices where effective competition existed; but where competition was absent, as in the case of non-competing groups, the fundamental determinant of international prices was not costs of production but reciprocal demand between non-competing groups. He also provided an account of the factors determining the movements and range of a country's prices and money incomes arising from international trade, along with an original analysis of the process of international borrowing and the effects of loans on the equilibrium of international trade. This contribution that has been described as ‘perhaps of greater permanent merit than any of his doctrines’ in this area (Angell, 1926, p. 94).

If the early and middle parts of the 19th century are associated respectively with the writings of Longfield and Cairnes, the latter part of the century must be identified with the work of Charles Bastable, who occupied the Whately Chair for 50 years, from 1882 to 1932. Bastable's *Public Finance*, first published in 1892, was a pioneering treatise that integrated, for the first time since McCulloch's *Taxation and the Funding System* (1845) what had become a rapidly expanding field of enquiry. Reviewing *Public Finance* in the *Economic Journal*, L.L. Price (1892) suggested it was the most comprehensive treatment of the topic since Adam Smith's *Wealth of Nations*. Bastable also made important contributions to international trade. In *The Theory of International Trade* (1887), he introduced varying elasticities of demand, increasing and decreasing returns and an extended analysis of obstacles to competition. In *The Commerce of Nations* (1892), he provided a stringent critique of protection, while his name is associated with the celebrated ‘Mill–Bastable’ condition, which became an important part of the extended analysis of protection (Chipman, 1965).

A distinguishing characteristic of many Irish economists in the 19th century was their commitment to an inductive method of approach. This was certainly true of the early Whately Professors. Isaac Butt maintained a robust scepticism with respect to the generality of economic principles, while Lawson was highly critical of Senior's efforts to reduce political economy to an axiomatic basis. It was not that the Irish writers called into question the validity of the deductive method in political economy. Rather, their position was that empirically observed facts should provide the basis for deductive reasoning. The methodological bias towards the inductive approach has been linked to the fact that the majority of the Irish professors were lawyers by training and profession and this allied to the preoccupation with the land question influenced their concentration on detailed studies of applied issues (Black, 1947). Two of the most important representatives of the inductive approach were Thomas Edward Cliffe Leslie and John Kells Ingram. Both were major figures in the English-speaking world as pioneers of the Historical School of political economy. They were critics of the classical method of deduction and stressed the absolute necessity of an inductive approach to the study of economic issues, which in their view could never be separated from the larger social matrix of relations. The exception among the Irish contributions to economic methodology in the 19th century and to the inductivist position in particular was Cairnes who, in *The Character and Logical Method of Political Economy*, provided the most rigorous exposition of the deductive method that was produced in the course of the century.

The 20th century

At the beginning of the 20th century, Belfast was Ireland's leading industrial centre, with strengths in linen, shipbuilding, rope making and engineering. Elsewhere agriculture predominated, and manufacturing was limited mainly to the food and drink industries. Against this background, the nascent independence movement regarded the development of industry as a matter of strategic importance. Drawing on the German economist Friedrich List, Arthur Griffith, the founder of Sinn Fein, proposed a programme for balanced economic development using protection on a broad scale. Protection was not to be permanent and was to be removed when the protected industries were strong enough to meet international competition (Griffith, 2003). Industrial Development Associations throughout the country urged people to purchase Irish-made goods. Although the validity of the infant industry argument was widely acknowledged, in the main professional economists were not advocates of protection. Professor Oldham of University College Dublin argued that the relative openness and small size of the Irish economy meant that the protection of the home market could not provide a basis for development. Oldham was also concerned about the hidden costs of protection and suggested that bounties should be preferred to tariffs on grounds of their greater transparency and controllability (Oldham, 1908; 1917).

During the 1921 Treaty negotiations with Britain, which were led by Arthur Griffith, professional economists including Riordan, O'Brien and Smiddy played a valuable role in securing the right of the Irish Free State to determine its future tariff regime (Girvin, 1989). However, despite this and the fact that partition, which accompanied independence in 1922, involved the loss of Ireland's leading manufacturing centre, the new Free State government was cautious in its approach to economic policy and favoured free trade, fiscal prudence and the maintenance of the link with sterling. Bastable of Trinity and George O'Brien of University College Dublin were members of a committee set up in 1923 to consider the case for greater protection. The committee came out strongly against tariff protection for industry. Among the grounds given were that protection would raise costs for exporting industries, including the all-important agricultural sector, whose increasing efficiency and exports were seen as the main motor for growth.

The Great Depression of the late 1920s and the widespread protectionism to which it gave rise made a re-evaluation of the free-trade position necessary. In any event, a new government with a different electoral base placed strong emphasis on self-sufficiency in both agriculture and industry. A bitter Anglo–Irish dispute over land annuities added further momentum to the protectionist drive. This led the Trinity College economist Joseph Johnston to argue in his polemical *Nemesis of Nationalism* (1934) that the Anglo–Irish dispute had been provoked by Eamon de Valera, the prime minister, in order to expedite his drive towards self-sufficiency. To judge from the contents of the Statistical and Social Inquiry Society journal during the period, the most prominent academic economists were also opposed to the policy of self-sufficiency. One of the few economists to comment favourably on the policy was J. M. Keynes, who also cautioned that only a modest degree of self-sufficiency could be achieved in such a small economy without a drastic impact on the standard of living (Whitaker, 1983, p. 59).

The protectionist policies were successful in increasing industrial output and employment, but, as predicted by Keynes and as understood by Sean Lemass, the industry and commerce minister, and his top civil servant, the real challenge was to nurture industry to international competitiveness and to maintain the impetus for development once the initial easy phase of import substitution was over. In the

event, the onset of the Second World War in 1939 and the growing scarcity of imported manufactures forced a further intensification of the policy of self-sufficiency. As elsewhere, the exigencies of the war economy led to greater government involvement in the allocation of resources and entrepreneurial activity generally. This continued after the war and, as late as 1959, Professor Charles Carter, formerly of Queen's University Belfast, commended the southern government for its willingness to engage in state enterprise if private enterprise failed to work, and contrasted this with the view taken in Northern Ireland that the function of government was to create the conditions for development and offer appropriate inducements but no more than that (Carter, 1969).

In preparation for the aftermath of the war, policy debate on appropriate strategies for agriculture and employment took place in the early 1940s. A committee on agricultural policy chaired by T. A. Smiddy, De Valera's economic advisor, emphasized the importance of agricultural efficiency and the restoration of exports as a means of earning the foreign exchange that was necessary for the purchase of raw materials for industrial development. The other major policy debate of the period was occasioned by the publication in the UK of the Beveridge Report and the *White Paper on Employment*. The spectrum of Irish attitudes towards Keynesianism was reflected in a discussion of the problem of full employment held by the Statistical and Social Inquiry Society on 27 April 1945 (Lynch et al., 1945, 438–59).

Opening the debate, Patrick Lynch, an economist in the Department of Finance and later Professor at University College Dublin, argued that the time had come to accept the Keynesian analysis of the economic system (Lynch et al., 438–41). The problems of the Irish economy were acute and required increasing state intervention. Government had to concern itself with the economy as a whole and not just its own expenditure as in the past. Lynch argued that government proposals for rural electrification and building were appropriate forms of intervention by means of which employment and further development could be stimulated. On the other hand, T. K. Whittaker, then number two at the Department of Finance, argued that Ireland was less exposed to cyclical fluctuations than Britain and America (Lynch et al., 446–9). Its unemployment problem was not primarily of a cyclical nature but the result of the insufficient investment in industry and agriculture. The problem was one of underinvestment rather than fluctuations in investment. Whitaker implied that increased investment in industry or agriculture would yield bigger returns at lower cost than the social investments mentioned by Lynch. Summarizing the debate, George O'Brien felt that there was general agreement that the Beveridge analysis did not apply to Irish circumstances (Lynch et al., 456–9). He noted that the main way in which Ireland had solved its unemployment problem during the last one hundred years was through the export of its people. Dr Beddy's recent paper comparing Irish and Danish agriculture (Beddy, 1943–4) had shown that a more efficient agriculture was likely to employ fewer rather than more people. While the comparison with Denmark showed the possibilities offered by secondary industries, O'Brien himself felt that tertiary industries such as tourism had considerable potential.

While there were some attempts at policy and institutional innovation in the late 1940s and early 1950s, these involved attempts to make existing policy more effective rather than any major policy shifts. The performance of the economy was sluggish with low or sometimes negative rates of growth, and high levels of emigration. The publication of *Economic Development* (Department of Finance, 1958) prepared by the Secretary of the Department, T. K. Whitaker, is commonly regarded as a major turning point in policy. The report demonstrates a remarkable consistency of position with the positions expressed by Whitaker in the 1945 debate on the *Beveridge Report*. The emphasis was on the need for productive investment. Whitaker argued that investment for which part if not all of the cost of servicing

must be paid by the taxpayer was redistributive rather than productive, and should be replaced by productive investment. Despite the emphasis on the importance of investment, Whitaker cited A. K. Cairncross to the effect that entrepreneurial capacity was an even more important factor (1958, pp. 6–7), and in the body of the report he argued that the problem was not so much one of obtaining capital as securing ‘know-what’ as well as ‘know-how’ (1958, p. 154). These, it was suggested might have to come from external sources including foreign direct investment. The report also argued that, since the home market had been largely catered for, further development would have to depend on exports.

The 1958 Programme for Economic Expansion based on *Economic Development* involved a shift of emphasis from the promotion of domestically owned import-substituting industry to foreign-owned export-oriented industry. The implementation of the shift in policy over the following decade involved the easing of restrictions on foreign ownership and the implementation of incentives in the form of grants and tax exemptions on export generated profits. It also involved a shift from state-led enterprise to private enterprise, although Whitaker himself recognized that, since private investment was limited, productive investment would have to be engaged in by the public sector for some time to come. Another feature of the programme was its argument against existing policies of decentralization and in favour of the concentration of industries in large population centres with good internal and external communications and pools of skilled labour. A similar position on the concentration of new enterprises and infrastructure was later put forward in the Buchanan Report on regional development, which was published in 1968 and led to considerable debate (Buchanan and Partners, 1968).

The introduction to the Programme for Economic Expansion emphasized that it was a not a plan and argued that the setting of detailed targets was inappropriate in a private enterprise economy exposed to fluctuations in external trade (Chubb and Lynch, 1969). A few years later, there was much greater optimism about the value of planning. A second programme, which was developed in cooperation with the newly created Economic Research Institute (currently Economic and Social Research Institute, ESRI) and with input from Professor Louden Ryan of Trinity College, was much more specific in its targets. However, the actual performance of the economy in the period covered by the programme deviated from the planned targets, which were then abandoned. Drawing on this experience, the third programme emphasized the conditional nature of targets (Chubb and Lynch, 1969). Despite this, the policies which were evolved in the programmes for economic expansion are widely regarded as providing the underpinning for the subsequent expansion of the economy. Governments also engaged in mildly expansionary fiscal policy, which helped to avoid the deflationary impact of Whittaker's own proposals. During the 1970s an attempt was made to counter the effects of oil price shocks by means of deficit spending. As the expected recovery failed to materialize, Ireland's debt burden grew and by the 1980s had reached unsustainable levels. During this period, Irish economists were vocal in their criticisms of fiscal policy and earned something of a reputation as hard-nosed monetarists. The world recession of the early 1980s resulted in the closure of some foreign plants and made it difficult to attract new investment, so that there were net job losses in the foreign-owned sector. Irish economists criticized the Industrial Development Authority for subsidizing capital as a means of job creation. The Telesis review of industrial policy proposed a more selective approach to foreign direct investment, a shift of emphasis towards building strong indigenous national champions, the substitution of employment grants for capital grants and the setting up of a national linkage programme (NESC, 1982). The emergence of the economy from the 1980s recession, sometimes characterized as expansionary fiscal contraction, was achieved through a partnership agreement involving trade unions, employers and government.

Partnership agreements have remained in place and are now regarded by economists in Ireland and elsewhere as an important industrial relations and development innovation (Teague and Donaghey, 2004). In the decades since the 1960s, other policy issues with which economists engaged were the Anglo–Irish Free Trade Agreement, entry into European Economic Community, the Common Agricultural Policy and European Monetary Union. More recently, there has also been increasing emphasis on the factors governing the productivity and competitiveness of the economy as a whole.

When Ireland achieved independence in 1922, its cadre of professional economists was small, and the Statistical and Social Inquiry Society and its journal provided the main discussion forum for academic economists and government officials. Although some academics, including O'Brien, were literary in their approach, others such as George Duncan of Trinity College and John Busteed of University College Cork used a variety of statistical and empirical techniques. Duncan produced estimates of Irish national income to supplement T.J. Kiernan's pioneering efforts. Given that Duncan could be not regarded as anti-statistical, it is perhaps surprising that he was one of the main protagonists in a protracted debate between economists and Roy Geary. Geary was Ireland's foremost statistician but he also made important technical contributions to economics, including the Stone–Geary utility function as well as methods for updating input–output tables, for making international comparisons of real income, and for calculating the change in real income arising from changes in the terms of trade (Neary, 1997, Spencer, 1997). Geary argued that economics could become a science only through measurement and that economists' failure to appreciate the value of statistical work was due to their lack of awareness of the power of modern statistics. Geary also felt that economic theory was of very little value in the solution of practical problems and that academic economists were not sufficiently active in researching the social problems of the day. Duncan countered that the collection and manipulation of data could not by themselves advance knowledge of economic behaviour (Fanning, 1984, pp. 151–5). He also pointed out that the Irish universities were seriously underfunded and had very little resources with which to carry out research. Part of the problem was a difference in attitude. Duncan, who had Austrian sympathies, disapproved of government intervention in general and of the protectionist policies of the day in particular. Geary, on the other hand, viewed policy issues as problems that could be solved with the correct technical means.

The present generation of Irish economists are more numerous and better trained than their predecessors in the early years of the 20th century. Many are the products of graduate schools in United States and in Britain. A recent examination of the journal output of Irish economists over the period 1970 to 2001 identified a total of 659 individual authors and 1,610 contributions, of which 218 were in the 1970s, 406 in the 1980s and 1,013 in the 1990s (Barrett and Lucey, 2003). Of the total over the full period, half were in the main Irish journals: the *Economic and Social Review*, the *Journal of the Statistical and Social Inquiry Society of Ireland* and the *Irish Banking Review*. Other popular outlets for Irish economists were *Regional Studies Applied Economics* and the *Economic Journal*. A relatively small number of Irish economists contributed to the top international journals during the period surveyed by Barett and Lucey. Of these, Peter Neary made important contributions to the theory of international trade as well as consumer theory, industrial organization and macroeconomics. In international trade, he is best known for work on Dutch Disease and the implications for trade of imperfect competition and technology policy.

In his seminal work on 20th century Ireland, the historian Joseph Lee (1989) argued that Irish

economists have been impressive in the analysis of short-term movements but have contributed little to understanding the long-term development of the economy and have failed to contribute to development economics more widely. Whatever its truth in the past, this statement is no longer an accurate reflection of the state of affairs. Since the early 1990s, the Irish economy has experienced rapid growth so that its GDP per capita is now among the highest in Europe. This has led to considerable interest in understanding the nature and timing of the forces at work in Ireland's catch-up (Honohan and Walsh, 2002). Meanwhile, however, Ireland's own development challenges have changed from those of catch-up to those of innovation and growth on the frontier. Meeting these challenges will require not only the strengthening of R&D capabilities but also addressing the special challenges of innovating in a small open economy.

See Also

- agriculture and economic development
- Cairnes, John Elliott
- Cantillon, Richard
- development economics
- foreign direct investment
- Geary, Robert Charles
- Historical economics, British
- Hutcheson, Francis
- Ingram, John Kells
- Petty, William
- Swift, Jonathan

Bibliography

Angell, J.W. 1926. *The Theory of International Prices: History, Criticism and Restatement*. Cambridge, MA: Harvard University Press.

Barrett, A. and Lucey, B. 2003. An analysis of the journal article output of Irish-based economists, 1970–2001. *Economic and Social Review* 34(2), 109–43.

Bastable, C.F. 1887. *The Theory of International Trade*, 4th edn. London: Macmillan, 1903.

Bastable, C.F. 1892. *Public Finance*. Phoenix, AZ: Simon Publications, 2003.

Bastable, C.F. 1892. *The Commerce of Nations*. London: Methuen.

Beddy, J.P. 1943–4. A comparison of the principal economic features of Eire and Denmark. *Journal of*

the Statistical and Social Inquiry Society of Ireland 17, 189–220.

Berkeley, G. 1752. *The Querist*. Dublin. Reproduced in Johnston (1970).

Black, R.D.C. 1945. Trinity College Dublin, and the theory of value, 1832–1863. *Economia* (n.s.) 12, 140–8.

Black, R.D.C. 1947. Economic studies at Trinity College, Dublin – I. *Hermathena* 70, 67–80.

Boylan, T.A. and Foley, T.P. 1984a. John Elliot Cairnes, John Stuart Mill and Ireland: some problems for political economy. In *Economists and the Irish Economy: From the Eighteenth Century to the Present Day*, ed. A. E. Murphy. Dublin: Irish Academic Press.

Boylan, T.A. and Foley, T.P. 1984b. Cairnes, Hearn and Bastable: the contribution of Queen's College, Galway to economic thought. In *Galway: Town and Gown 1484–1984*, ed. D. OCearbhaill. Dublin: Gill & Macmillan.

Boylan, T.A. and Foley, T.P. 1992. *Political Economy and Colonial Ireland*. London: Routledge.

Boylan, T.A. and Foley, T.P. 1993. The teaching of economics at the Queen's Colleges in Ireland (Belfast, Cork, Galway), 1845–1900. In *The Market for Political Economy: The Advent of Economics in British University Culture, 1850–1905*, ed. A. Kadish and K. Tribe. London and New York: Routledge.

Boylan, T.A. and Foley, T.P. 2004. *John Elliot Cairnes: Collected Works*. London and New York: Routledge.

Buchanan, C. and Partners. 1968. *Regional Studies in Ireland*. Dublin: An Foras Forbartha.

Butt, I. 1846. *Protection to Home Industries: Some Cases of Its Advantages Considered*. Dublin: Hodges and Smith; London: John W. Parker.

Cairnes, J.E. 1874. *Some Leading Principles of Political Economy Newly Expounded*. London: Macmillan & Co.

Carter, C.F. 1969. Problem of economic development. In Chubb and Lynch (1969).

Chipman, J.S. 1965. A survey of the theory of international trade: Part I, the classical theory. *Econometrica* 33, 477–519.

Chubb, B. and Lynch, P. 1969. *Economic Development and Planning*. Dublin: Institute of Public Administration.

Cliffe Leslie, T.E. 1888. *Essays in Political Economy*, 2nd edn. Dublin: Hodges, Figgis & Co.

Corden, W.M and Neary, J.P. 1982. Booming sector and de-industrialization in a small open economy, *Economic Journal* 92, 825–48.

Daly, M.E. 1997. *The Spirit of Earnest Inquiry: The Statistical and Social Enquiry Society of Ireland 1847–1997*. Dublin: Statistical and Social Inquiry Society of Ireland.

Department of Finance. 1958. *Economic Development* (*The Whitaker Report*). Dublin: Stationery Office.

Duddy, T. 2002. *A History of Irish Thought*. London and New York: Routledge.

Fanning, R. 1984. Economists and governments: Ireland 1922–52. In *Economists and the Irish Economy*, ed. A.E. Murphy. Dublin: Irish Academic Press.

Girvin, B. 1989. *Between Two Worlds – Politics and Economy in Independent Ireland*. Dublin: Gill and Macmillan.

Griffith, A. 2003. *The Resurrection of Hungary: A Parallel for Ireland*, ed. P. Murray. Dublin: University College Dublin Press.

Hutcheson, F. 1726. Remarks upon the Fable of the Bees. *Dublin Journal*, 5, 12 and 19 February.

Honohan, P. and Walsh, B. 2002. Catching up with the leaders: the Irish hare. *Brookings Papers on Economic Activity* 2002(1), 1–77.

Johnston, J. 1934. *Nemesis of Nationalism*. London: P.S. King.

Johnston, J. 1970. *Bishop Berkeley's Querist in Historical Perspective*. Dundalk: Dun Dealgan Press.

Kelly, J. 1991. Jonathan Swift and the Irish economy in the 1720s. *Eighteenth-Century Ireland* 6, 7–36.

Lee, J.J. 1989. *Ireland 1912–1985 – Politics and Society*. Cambridge: Cambridge University Press.

Leahy, D. and Neary, J.P. 1997. Public policy towards R&D in oligopolistic industries. *American Economic Review* 87, 642–62.

Longfield, M. 1834. *Lectures on Political Economy*. Dublin: William Curry, Jun. and Company.

Longfield, M. 1835. *Three Lectures on Commerce and One on Absenteeism*. Dublin: William Curry,

Jun. and Company; London: Longman and Company.

Lynch, P. et al. 1945. The problem of full employment: a discussion. *Journal of the Statistical and Social Inquiry Society of Ireland* 17, 98th session, 438–59.

Moss, L.S. 1976. *Mountifort Longfield: Ireland's First Professor of Political Economy*. Ottawa, IL: Green Hill Publishers.

Murphy, A.E. 2000. Canons of monetary orthodoxy and John Law. In *Contributions to Political Economy – Essays in Honour of R.D.C. Black*, ed. A.E. Murphy and R. Prendergast. London and New York: Routledge.

Neary, J.P. 1997. R. C. Geary's contributions to economic theory. In *Roy Geary, 1896–1983: Irish Statistician*, ed. D. Conniffe. Dublin: Oak Tree Press.

Neary, J.P. and Roberts, K.W.S. 1980. Theory of household behaviour under rationing. *European Economic Review* 13, 25–42.

NESC (National Economic and Social Council). 1982. *A Review of Industrial Policy. A Report Prepared by the Telesis Consultancy Group*. Dublin.

Oldham, C.H. 1908. The economics of ‘industrial revival’ in Ireland. *Journal of the Statistical and Social Inquiry Society of Ireland* 12, 175–89.

Oldham, C.H. 1917. Industrial Ireland under free trade. *Journal of the Statistical and Social Inquiry Society of Ireland*, vol.13, part 96, 383–98.

Petty, W. 1672. Political Anatomy of Ireland. In *The Economic Writings of William Petty*, vol. 1, ed. C. H. Hull. Cambridge: Cambridge University Press, 1899.

Prendergast, R. 1987. James Anderson's political economy – his influence on Smith and Malthus. *Scottish Journal of Political Economy* 34, 388–409.

Price, L.L. 1892. [Review of] Charles Bastable *Public Finance*. *Economic Journal* 2, 671–6.

Spencer, J.E., 1997. R. C. Geary: his life and work. In *Roy Geary, 1896–1983: Irish Statistician*, ed. D. Conniffe. Dublin: Oak Tree Press.

Teague, P. and Donaghey, J. 2004. The Irish experiment in social partnership. In *The New Structure of Labor Relations: Tripartism and Decentralization*, ed. H. C. Katz and W. Lee. Ithaca, NY: ILR Division, Cornell University Press.

Viner, J. 1955. *Studies in the Theory of International Trade*. London: George Allen & Unwin.

Whitaker, T.K. 1983. *Interests*. Dublin: Institute of Public Administration.

How to cite this article

Boylan, Tom and Renee Prendergast. "Ireland, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000282> doi:10.1057/9780230226203.0851

The New Palgrave Dictionary of Economics Online

iron law of wages

Mark Blaug
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

iron law of wages; Lassalle, F.; market price of labour; natural price of labour; natural wages; population growth; real wages; Ricardo, D.; subsistence' theory of wages

Article

The 'iron (or brazen) law of wages' is a term invented by Ferdinand Lassalle (1862) to describe the inexorable tendency of real wages under capitalism to adhere to a level just sufficient to afford the bare necessities of life. This law, he claimed, was not just a socialist indictment of capitalism but was authorized by leading 'bourgeois' economists such as Malthus and Ricardo. He failed to point out, however, that in Malthus and Ricardo the so-called 'subsistence' theory of wages was predicated on a theory of population growth according to which the supply of labour responds automatically to any gap between the going 'market price' and 'natural price' of labour, the latter being defined as a real wage sufficient to reproduce a working population of given size and composition. Lassalle, however, being a socialist, followed Marx in rejecting the Malthusian theory of population; what ensured the 'iron law of wages' for Lassalle, as for Marx, was the tendency for any rise in real wages to generate unemployment, thus setting in motion forces that reversed the rise. This threw the entire weight of argument for equilibrium adjustments in the labour market on the side of employers' demand; it provided no explanation of the supply of labour and thus failed to furnish a determinate theory of wages in long-run equilibrium. Ironically, therefore, there may be an 'iron law of wages' in Malthus and Ricardo, but there is certainly no such iron law in socialist economics. The question whether Malthus and particularly Ricardo can be said to have held the iron law or subsistence theory of wages was a favourite debating question in the latter half of the 19th century (see, for example, Marshall, 1890, pp. 508–9). There is no doubt that they held the view that real wages tend to fluctuate around a natural point of 'gravity', namely, the minimum level of food and other necessities required for existence. But, in the first place, these fluctuations, depending as they did upon decisions to marry and to have children, involved a lag of at least 15–18 years, a point which Malthus (but not Ricardo) conceded explicitly. In the second place, the minimum-of-existence level of 'natural wages' was admitted to be a matter of custom and habit and

therefore subject to a secular upward drift. It was therefore perfectly possible to argue for the existence of something like a normal long-run supply price of labour – a constant real wage, everything else being the same – while at the same time granting that the ‘market price’ of labour fluctuated around an ever-rising trend. In short, rising living standards under capitalism do not violate the iron law of wages, understood as a theory about the long-run equilibrium price of labour. But that is only to say that the iron law or subsistence theory of wages amounts for all practical purposes to accepting customary wages as an institutional datum (Schumpeter, 1954, p. 665).

There has been a revisiting of the old debate about whether Ricardo held the iron law of wages, but in an entirely new form: did Ricardo hold real wages to be constant at the subsistence level in stationary equilibrium or did he allow for an initial stage of increasing real wages followed by a final stage of declining wages alongside a secular fall in the rate of profit (Hollander, 1983)? It is doubtful whether this question yields one simple, neat answer, since it is clear that Ricardo operated with a number of different models regarding the determination of the ‘natural price’ of labour. In the very opening paragraph of the chapter on wages in Ricardo's *Principles of Political Economy of Taxation*, the ‘natural price’ of labour is defined as ‘that price which is necessary to enable the labourers, one with another, to subsist and to perpetuate their race, without either increase or diminution’. This defines the natural price of labour to be the commodity wage that ensures a zero rate of population growth. But a page or two later, the natural price of labour is said to be that commodity wage which ensures a rate of growth of population equal to the rate of growth of the capital stock, so that market wages only rise above natural wages when capital accumulates faster than the growth of population. It is possible to make sense of this in terms of modern growth theory, and many have done so (see Casarosa, 1978), but it is questionable whether Ricardo himself was aware of what he was doing, the more so as he frequently resorts to the constant-subsistence-wage assumption in the later tax chapters of the *Principles*.

See Also

- wages fund

Bibliography

Casarosa, C. 1978. A new formulation of the Ricardian system. *Oxford Economic Papers* 30(1), 38–63.

Hollander, S. 1983. On the interpretation of Ricardian economics: the assumption regarding wages. *American Economic Review* 73, 314–18.

Lassalle, F.J.G. 1862. *Open Letter to the National Labor Association of Germany*. Trans. J. Ehmann and F. Bader, Cincinnati: Cincinnati Press, 1879.

Marshall, A. 1890. *Principles of Economics*. 9th (Variorum) edn, ed. C.W. Guillebaud, vol. 1, London: Macmillan, 1961.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

How to cite this article

Blaug, Mark. "iron law of wages." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000224> doi:10.1057/9780230226203.0852

The New Palgrave Dictionary of Economics Online

irreversible investment

Janice C. Eberly
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The cost of an irreversible investment cannot be recovered once it is installed. This restriction not only truncates negative investments, but also raises the threshold for positive investment. The threshold return that justifies an irreversible investment increases with uncertainty, or more precisely, with the probability mass in the lower tail of outcomes. Irreversibility constrains the ability to redeploy capital in ‘bad’ states, so the agent is particularly sensitive to these states when investing *ex ante*. This finding is analogous to valuation and exercise of financial options, and irreversible investments are valued and understood by using option pricing techniques.

Keywords

adjustment costs; irreversible investment; option pricing theory; option valuation; put–call parity; uncertainty

Article

Irreversible investment acknowledges that the value of capital may not be fully recoverable when resold.
This simple generalization has rich implications for investment. Beyond truncating disinvestment, irreversibility changes the dynamics of investment by creating a threshold level of returns for positive investments. Below this threshold, investment is zero – which immediately implies intermittent rather than continuous investment activity. Moreover, the threshold return that justifies investment exceeds the required return on a reversible investment.

Investment and options

Marschak (1949) raised the potential role of irreversibility in factor accumulation by emphasizing the convertibility or liquidity of capital. Work by Arrow (1968) and Henry (1974) considered when irreversible actions in environmental applications were justified and emphasized the idea of an option value. This idea was extended by Bernanke (1983) to the role of uncertainty in delaying investment decisions.
McDonald and Siegel's (1986) article ‘The Value of Waiting to Invest’ provides the first explicit valuation of investment allowing for irreversibility, incorporating option valuation (real options) into investment theory. McDonald and Siegel analyse a project of fixed size, so the timing of the project is the only choice to be made. They show that the value of the project includes an ‘option value of waiting’, that can be valued and interpreted using option pricing theory. The additional value of being able to choose when to invest, rather than a ‘now or never’ investment decision, can be quantitatively large, and has interesting implications for the investment decision. First, the presence of this option implies that it is optimal to delay the investment, rather than undertaking it immediately, even when immediate execution has positive value. Instead, value can be increased by waiting for additional information. Second, like most options, the value of the option to wait is increasing in uncertainty. This feature implies an effect of uncertainty on the value and timing of investments that is absent in most conventional models.
Later work by Pindyck (1988) and Bertola (1988), allows for incremental investment, so that the firm chooses both the timing and the size of its investments. They show that there is a threshold for investing with irreversibility that exceeds the return that would justify a positive reversible investment. Instead of a single investment decision, as in McDonald and

Siegel, there is an infinite sequence of investment decisions, where each satisfies the threshold condition.

An illustrative model

Most irreversible investment models work in continuous time, so that optimal investment timing can be calculated exactly. An introduction to these techniques, as well as a broader overview, is found in Dixit and Pindyck's (1994) *Irreversible Investment*. The intuition can be understood in a discrete time framework, adapted from Abel, Dixit, Eberly and Pindyck (1996), specialized to the case of irreversible investment.

Consider the decision of a single firm to undertake a capital investment at time 1. In the first period, the return to installing capital K_1 is $r(K_1)$. The total return $r(K_1)$ is strictly increasing and concave in K and satisfies the Inada conditions. The firm pays a price b per unit of capital to purchase capital. In the second period, the return to capital is uncertain and equal to $R(K, \varepsilon)$, where ε is stochastic. The derivative of $R(K, \varepsilon)$ with respect to K, $R_K(K, \varepsilon) \geq 0$, is continuous and strictly decreasing in K, continuous and strictly increasing in ε, and $R(K, \varepsilon)$ also satisfies the Inada conditions. Define a threshold value of ε by

$$R_K(K_1, \bar{\varepsilon}) = b, \tag{1}$$

as illustrated in Figure 1.

Figure 1

The second period marginal return to capital

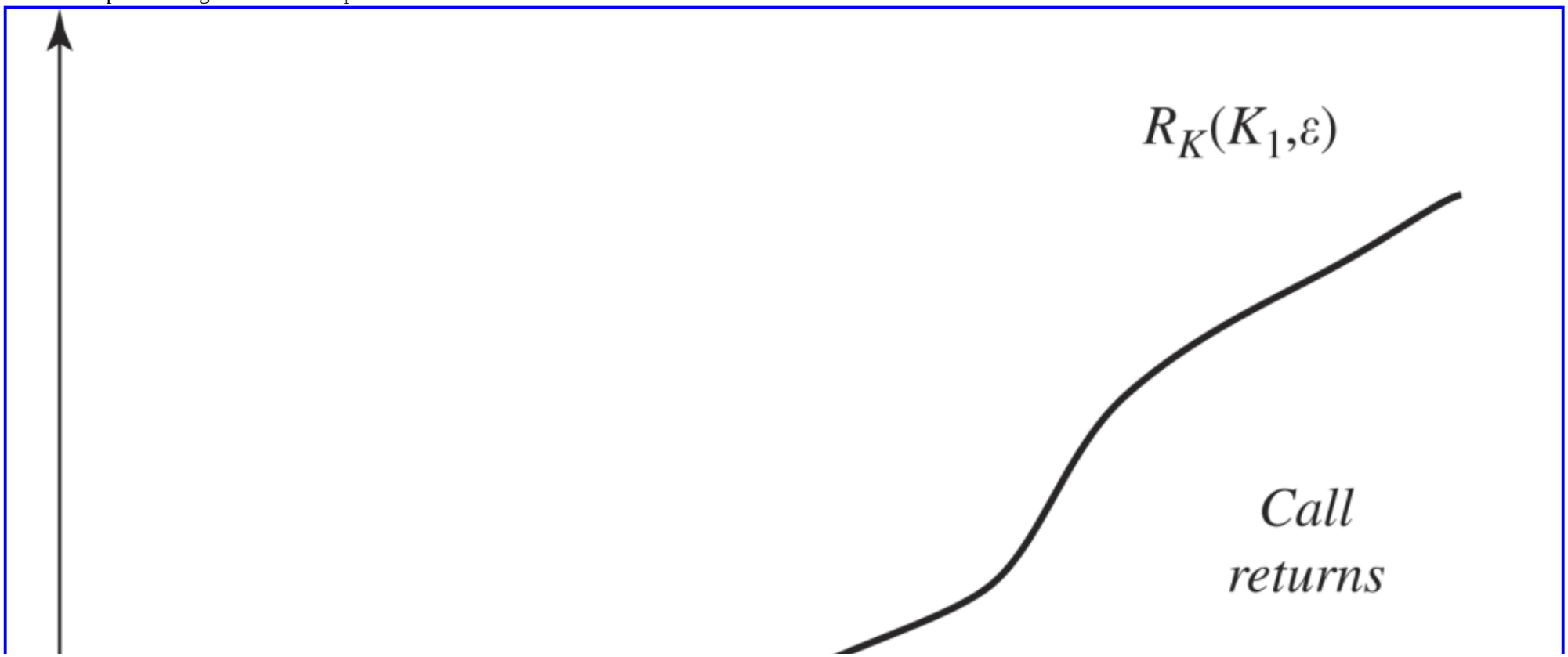

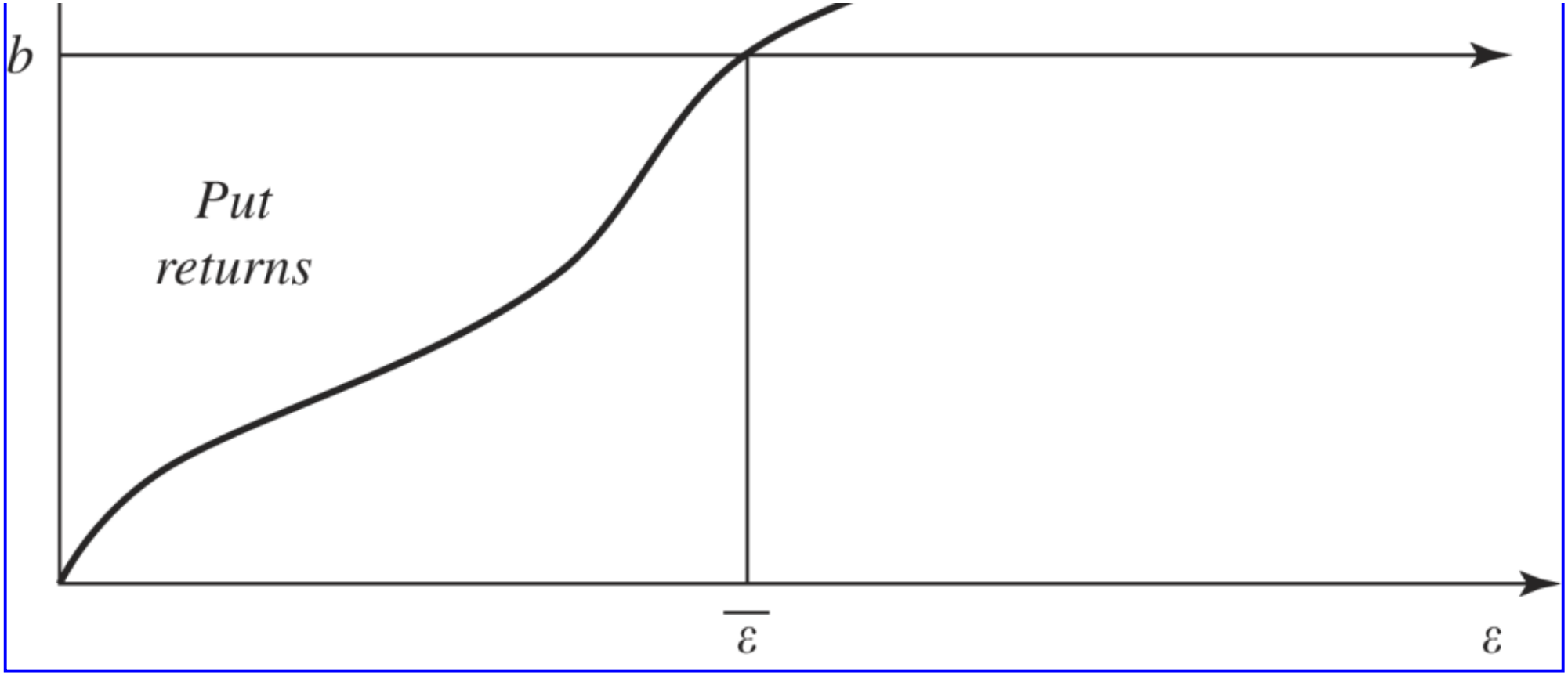

Assume that the resale price of capital is zero, or complete irreversibility. In the second period, the capital stock is optimally chosen at a level equal to $K_2(\varepsilon)$, subject to the irreversibility constraint. When $\varepsilon > \bar{\varepsilon}$, the optimal capital stock rises to satisfy the first-order condition $R_K[K_2(\varepsilon), \varepsilon]=b$. However, when $\varepsilon > \bar{\varepsilon}$, the marginal return to capital is less than its purchase price. If the firm could resell capital at its acquisition price b (costless reversibility) it would do so. However, the available resale price is zero, so the firm prefers to keep its capital stock, which has positive marginal return; in this case $K_2(\varepsilon)=K_1$. The optimal second-period marginal return to capital is graphed in Figure 1 as the lower envelope of $R_K(K_1, \varepsilon)$ and b.

Conditional on the optimal second-period capital stock, the firm chooses its capital stock at time 1 to maximize $V(K_1)-bK_1$, where $V(K_1)$ is the first period value of the firm equal to $r(K_1) + \gamma E[R(K_2, \varepsilon)]$ and $0 < \gamma < 1$ is the discount factor. The first-order condition for the optimal capital choice is

$$V'(K_1) \equiv r'(K_1) + \gamma \int_{-\infty}^{\bar{\varepsilon}} R_K(K_1, \varepsilon) dF(\varepsilon) + \gamma b[1 - F(\bar{\varepsilon})] = b, \tag{2}$$

where $F(\varepsilon)$ is the cumulative distributive function (CDF) of ε.

Notice that the term $V'(K_1)$ is the marginal value of an additional unit of capital, or marginal q. The standard investment first-order condition equating marginal q to the marginal cost of capital still holds with irreversibility. The effects of irreversibility are incorporated into the value of marginal q, so when investment is non-zero the standard q-theory first-order condition equating the marginal value and the marginal cost of investment still holds.

Embedded options

Now rewrite this first-order condition to highlight the investment options and their implications for the investment decision. Rewrite eq. (2) as

$$q(K_1) \equiv V'(K_1) \equiv n(K_1) - \gamma c(K_1) \tag{3}$$

where

$$n(K_1) \equiv r'(K_1) + \gamma \int_{-\infty}^{\infty} R_K(K_1, \varepsilon) dF(\varepsilon) > 0 \tag{4}$$

and

$$c(K_1) \equiv \int_{\varepsilon}^{\infty} [R_K(K_1, \varepsilon) - b] dF(\varepsilon) > 0. \tag{5}$$

The marginal value of an additional unit of capital is decomposed into two terms. The first term, $n(K_1)$, is equal to the present value of marginal returns to capital, evaluated at its current level, K_1. The second term subtracts the discounted value of a call option, $c(K_1)$, to add more capital, as illustrated in Figure 1, where the returns to the call option are represented by the area under $R_K(K_1, \varepsilon)$ and above the line b. The call option reduces the marginal value of capital because additional capital irrevocably reduces the marginal return to capital owing to the concavity of the revenue function. If one combines these two terms, the marginal value of capital is the discounted sum of marginal revenues on the assumption that the capital stock is fixed, less the marginal value of the option to increase the capital stock. Note that the concavity of the revenue function is crucial to this mechanism. Hence, models such as Abel and Eberly, (1997) which assume constant returns to scale, do not generate these option values.

The effects of uncertainty are not transparent in the above formulation, since both terms in eq. (3) depend on the distribution, $F(\varepsilon)$. To better discern the effect of uncertainty, rewrite eq. (2) instead as

$$q(K_1) \equiv V'(K_1) \equiv j(K_1) - \gamma p(K_1) \tag{6}$$

where

$$j(K_1) \equiv r'(K_1) + \gamma b > 0 \tag{7}$$

and

$$p(K_1) \equiv \int_{-\infty}^{\bar{\varepsilon}} [b - R_K(K_1, \varepsilon)]\,dF(\varepsilon) > 0. \tag{8}$$

The marginal value of an additional unit of capital is again decomposed into two terms. The first term, $j(K_1)$, is the discounted marginal return to costlessly reversible capital: the firm earns the marginal revenue in period one and can sell the capital for the same price b in period two. This is the Jorgensonian marginal return (Jorgenson, 1963); notice that it is independent of ε and risk free. The second component of q is the put option to sell capital at price b. When investment is irreversible, the put option is not available to the firm, since it cannot sell capital at any positive price. The value of the put option must be subtracted from the Jorgensonian valuation (where resale at price b would be permitted) to obtain the marginal value of irreversible capital. Marginal q can thus be written as a frictionless value less the value of the put option that is eliminated by the irreversibility constraint. This is illustrated in Figure 1 by subtracting the returns to the put option (the area under the line b and above the function $R_K(K_1, \varepsilon)$) from the frictionless return b.

Effects of uncertainty and put – call parity

To calculate the effect of uncertainty on marginal q from eq. (6), one need only calculate the effect of uncertainty on $p(K_1)$, since $j(K_1)$ is risk free. The effect of uncertainty on $p(K_1)$ is clear: $p(K_1)$ is an option value, and an increase in uncertainty increases the value of an option. In this case specifically, a second-order stochastic dominant shift in the distribution of ε shifts the CDF up for every value of ε. Since $R_K(K_1, \varepsilon)$ is increasing in ε, the term $[b–R_K(K_1, \varepsilon)]$ is decreasing in ε. Hence, greater uncertainty in ε shifts more weight of the CDF towards the large option payoffs in the left tail and unambiguously increases the value of the option, $p(K_1)$. Greater uncertainty unambiguously lowers the value of $q(K_1)$. Since $q(K_1)$ is decreasing in K_1, a downward shift in $q(K_1)$ reduces the optimal value of K_1 for a given value of b, as illustrated in Figure 2. This decrease is the incremental investment counterpart to McDonald and Siegel's finding that greater uncertainty increases the option value of waiting, lowering the value of investing immediately.

Figure 2

Marginal q and the optimal capital stock under low and high uncertainty

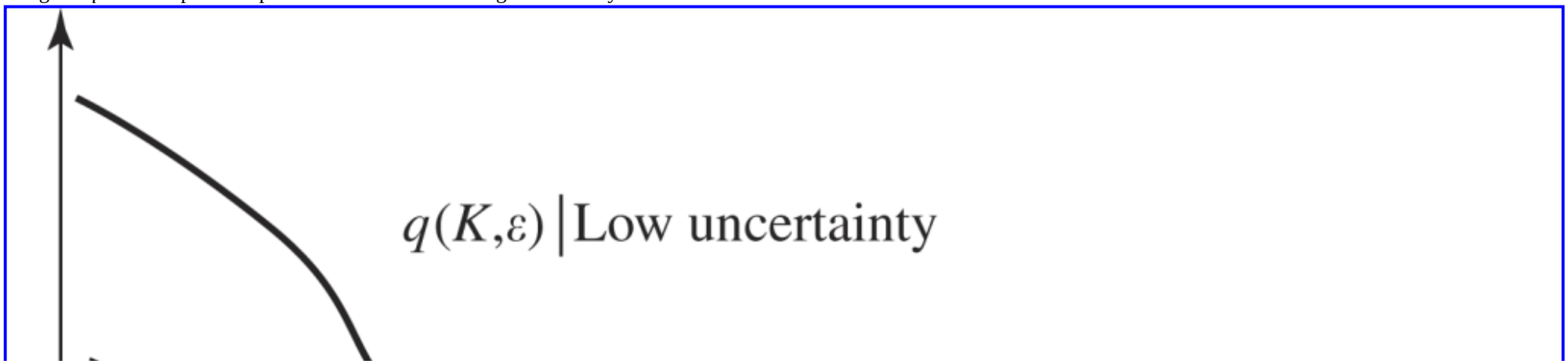

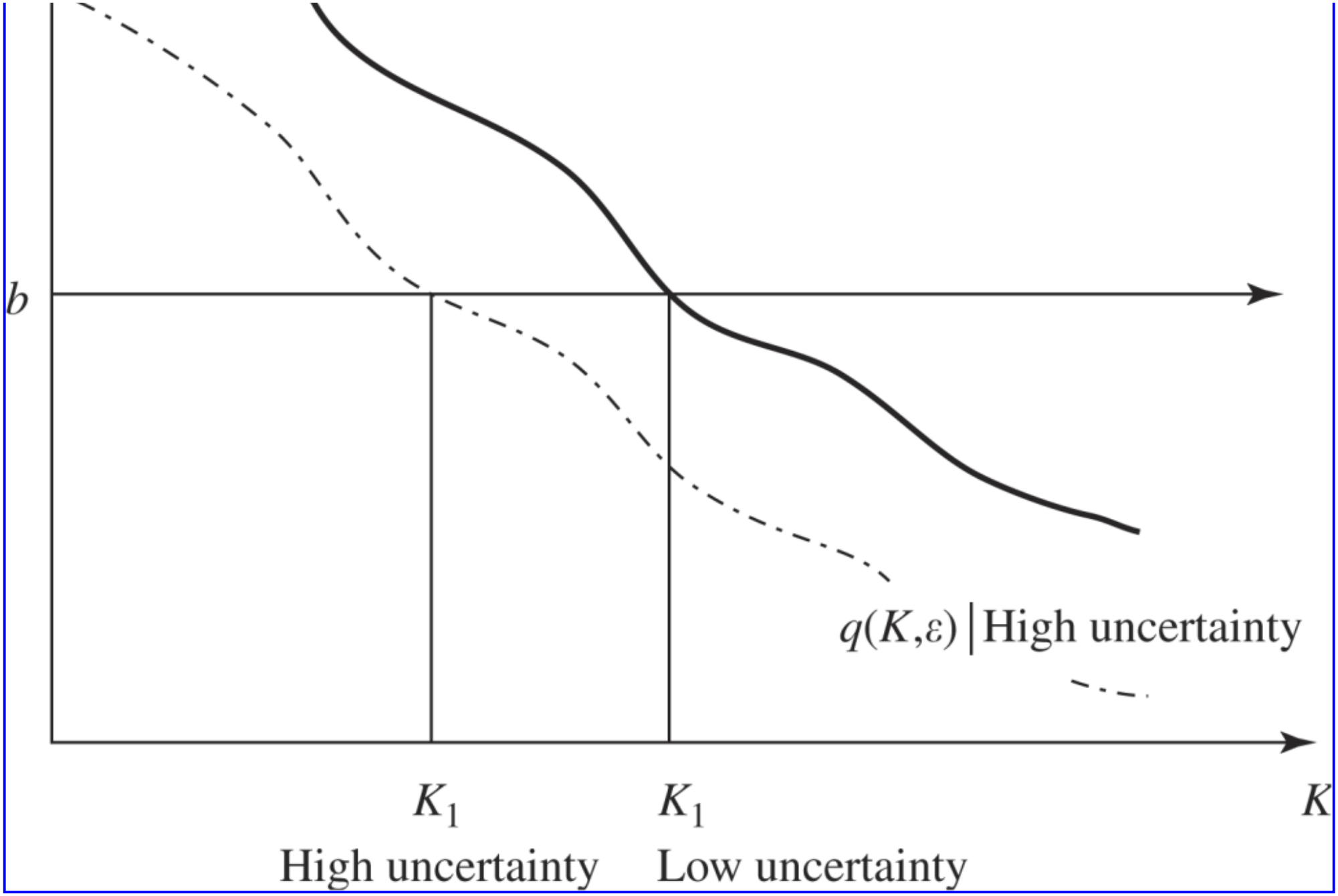

This formulation of $q(K_1)$ also demonstrates Bernanke's (1983) 'bad news principle' of irreversible investment. The distribution of ε only appears in the expression for q in eq. (6) via the put option $p(K_1)$. The put option only depends on the lower tail of the distribution of ε, below the threshold $\bar{\varepsilon}$. That is, the only part of the distribution of shocks that affects the value of $q(K_1)$ is the lower tail – or the 'bad news'. The upper tail is irrelevant, since in that region, the firm invests until the marginal product of capital equals its price. The exact realization of the shock in this region is irrelevant to the marginal return. In the lower tail, on the other hand, the firm neither invests nor disinvests, and the realization of the shock determines the marginal return to capital.

Figure 1 illustrates these arguments. The second period return to capital is the lower envelope of the price of capital, *b*, and the second-period marginal return, $R_K(K_1, \varepsilon)$ evaluated at K_1. The value of these returns depends on ε only in the lower tail of the distribution of ε (the bad news principle). The lower envelope can be expressed as either the function $R_K(K_1, \varepsilon)$ less the area labelled *call returns* in Figure 1; adding this difference to first-period marginal returns $r'(K_1)$, we obtain the expression for *q* in eq. (3). Equivalently, the second-period marginal return can be expressed the line *b* less the area labelled *put returns* in Figure 1. Adding the first-period marginal return $r'(K_1)$, we obtain the expression for marginal *q* in eq. (6). The fact that the second-period return can be written in two equivalent ways using options follows from put–call parity, a fundamental property of options prices. In fact, in this setting put–call parity is found simply by setting the two expressions for *q* in eqs (3) and (6) equal to each other. Equating these two expressions for *q* and simplifying, we find

$$\gamma p(K_1) + n(K_1) = \gamma c(K_1) + j(K_1). \tag{9}$$

This expression equates the value of a portfolio containing a put option and the underlying security, $n(K_1)$, to the value of a portfolio containing a call option and a risk-free asset. For a financial security such as a stock with price *S*, put–call parity analogously states that $P(S, \tau) + S = C(S, \tau) + X / (1 + r)^{\tau}$, where *X* is the strike price of the options and τ is the time to maturity. The terms $P(S, \tau)$ and $C(S, \tau)$ are the value of the put and call, respectively, on the underlying stock, *S*. $X / (1 + r)^{\tau}$ is the present value of a risk-free payoff (a zero coupon bond) of *X* in τ periods.

Extensions and applications

The above analysis assumes complete irreversibility. However, less stringent forms of the constraint deliver similar implications. Abel and Eberly (1996) examine costly reversibility, where capital can be disinvested and resold at a price less than the purchase price of capital. In this case, the gap between the investment and disinvestment thresholds opens quickly, even for small differences between the purchase and sale prices of capital. Moreover, this formulation has assumed kinked, linear adjustment costs, so that the degree of irreversibility is summarized by the ratio of the purchase and sale prices of capital. However, with more general cost formulations, such as Abel and Eberly (1994), capital may have a positive resale price and still be effectively irreversible when other costs of reselling capital exceed any potential benefits. In addition to a resale market discount, convex adjustment costs and fixed costs, for example, may induce irreversibility.

Research on irreversibility has branched out both empirically and theoretically. Initial applications included energy and natural resource markets (Brennan and Schwartz, 1985), with extensions to virtually all types of quasi-fixed capital, including durable goods, real estate and equipment investment. Modelling has been extended to include multiple types of quasi-fixed capital goods (Eberly and van Mieghem, 1997). Aggregating models with infrequent adjustment to incorporate equilibrium effects is challenging, and the results remain controversial. Except in very special cases (Caplin and Spulber, 1987) aggregating requires tracking a distribution of agents. However, it is precisely this feature that can match the observation that much of the volatility in empirical investment arises from the extensive margin (the number of agents adjusting) rather than the intensive margin (the average size of the adjustment). Much progress has been made in this direction (for example, Caballero and Engel, 1999), though the quantitative implications vary with modelling strategy (Veracierto, 2002).

See Also

- adjustment costs
- Marschak, Jacob
- s-S models
- Tobin, James
- Tobin's q

Bibliography

Arrow, K.J. 1968. Optimal capital policy and irreversible investment. In *Value, Capital, and Growth*, ed. J.N. Wolfe. Chicago: Aldine.

Abel, A.B., Dixit, A.K., Eberly, J.C. and Pindyck, R.S. 1996. Options, the value of capital, and investment. *Quarterly Journal of Economics* 111, 753–77.

Abel, A.B. and Eberly, J.C. 1994. A unified model of investment under uncertainty. *American Economic Review* 84, 1369–84.

Abel, A.B. and Eberly, J.C. 1996. Optimal investment with costly reversibility. *Review of Economic Studies* 63, 581–93.

Abel, A.B. and Eberly, J.C. 1997. An exact solution for the investment and value of a firm facing uncertainty, adjustment costs, and irreversibility. *Journal of Economic Dynamics and Control* 21, 831–52.

Bernanke, B.S. 1983. Irreversibility, uncertainty, and cyclical investment. *Quarterly Journal of Economics* 98, 85–106.

Bertola, G. 1988. Adjustment costs and dynamic factor demands: investment and employment under uncertainty. Ph.D. Dissertation, Cambridge, MA: Massachusetts Institute of Technology.

Brennan, M. and Schwartz, E. 1985. Evaluating natural resource investments. *Journal of Business* 58, 135–57.

Caballero, R.J. and Engel, E.M.R.A. 1999. Explaining investment dynamics in U.S. manufacturing: a generalized (S,s) approach. *Econometrica* 67, 783–826.

Caplin, A.S. and Spulber, D.F. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102, 703–26.

Dixit, A.K. and Pindyck, R.S. 1994. *Irreversible Investment*. Princeton: Princeton University Press.

Eberly, J.C. and van Mieghem, J.A. 1997. Multi-factor dynamic investment under uncertainty. *Journal of Economic Theory* 75, 345–87.

Henry, C. 1974. Option values in the economics of irreplaceable assets. *Review of Economic Studies* 41, 89–104.

Jorgenson, D. 1963. Capital theory and investment behavior. *American Economic Review* 53, 247–59.

Marschak, J. 1949. Role of liquidity under complete and incomplete information. *American Economic Review* 39, 182–95.

McDonald, R.L. and Siegel, D. 1986. The value of waiting to invest. *Quarterly Journal of Economics* 101, 707–28.

Pindyck, R.S. 1988. Irreversible investment, capacity choice, and the value of the firm. *American Economic Review* 78, 969–85.

Veracierto, M.L. 2002. Plant-level irreversible investment and equilibrium business cycles. *American Economic Review* 92, 181–97.

How to cite this article

Eberly, Janice C. "irreversible investment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000291>

doi:10.1057/9780230226203.0853

The New Palgrave Dictionary of Economics Online

Islamic economic institutions

Timur Kuran

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economic institutions of the classical Islamic world include Islamic contract law and the waqf, a form of trust. Until modern times, these two institutions were generally beneficial to economic performance. However, each had limitations that eventually blocked modern economic growth. Islamic contract law discouraged the formation of large and long-lived partnerships, thus obviating the need for business techniques and organizational forms associated with economic modernization. The waqf, designed as a rigid organization, locked capital into inefficient uses. Not until modern times has the corporation, a more flexible organizational form, entered the legal systems of the Islamic world.

Keywords

charitable contributions; choice of law; contract law; corporation; double-entry bookkeeping; industrial revolution; inheritance; interest; Islamic economic institutions; limited liability; partnerships; tax farming; waqf; zakat

Article

Prior to the 18th century, the Islamic world did not appear economically underdeveloped to outside observers. Comparative studies by economic historians confirm that it became ‘poor’ in relation to Europe during the Industrial Revolution. Until that point, economic institutions grounded in Islamic law had afforded a respectable level of wealth by standards of the day. They had also facilitated the spread of Islam across Asia, southern Europe, and the coasts of Africa.

Law of contracts

The first few centuries of Islam – *c*. 622–1000 ad – witnessed the gradual development of an elaborate law of contracts. It enabled the pooling of labour and capital through several forms of partnership,

including ones providing limited liability to passive investors. Profit shares, negotiated in advance, could be unequal or contingent. Islamic partnership contracts were enforced, with minor variations, wherever Muslims ruled. As merchants and producers moved, they carried Islamic law with them, helping to spread Islam. Huge numbers of people converted in order to gain acceptance into lucrative commercial networks managed according to Islamic law.

Islamic law limited neither the size of a partnership nor its duration. However, in practice the typical Islamic partnership consisted of two people, who pooled resources for a single economic venture expected to last just a few months (Çizakça, 1996). Lacking a life of its own, it was not what we call a firm. If a partner died during the contract period, the partnership became null and void, and the decedent's share of the assets fell to his heirs. There could be numerous claimants, for the Islamic inheritance system, by medieval standards remarkably egalitarian, assigns mandatory shares to a possibly long list of extended relatives. Accordingly, reconstituting a dissolved partnership could be very costly. Merchants and investors minimized the risk of dissolution by keeping their partnerships small and ephemeral (Kuran, 2003b).

A long-term consequence is that Islamic partnerships remained structurally simple, which obviated pressures to develop the sorts of organizational forms and business techniques that, in western Europe, gradually led to the modern economy. For instance, double-entry book-keeping did not develop, and no markets arose for trading enterprise shares. This institutional inertia made it impossible to borrow new organizational forms, except as part of a comprehensive legal reform. Advanced organizational forms, such as the joint-stock company and the corporation, reached the Islamic world in the 19th century through the imposition of secular commercial law. By that time the financing and organization of the region's external trade was largely under Western control; and, as a result of the Industrial Revolution, productivity was much higher in the West than elsewhere. The very commercial institutions that had served Muslims well through the Middle Ages were now hindering the exploitation of modern technologies.

Role of minorities

The religious minorities of the Islamic world might have escaped the limitations of Islamic commercial institutions, because they enjoyed ‘choice of law’ – the privilege to do business under legal systems of their own. Yet as individuals non-Muslims could opt unilaterally to take anyone to an Islamic court, whose decision would trump that of a non-Muslim judge or arbitrator. To achieve predictability in their economic relations, non-Muslims thus tended to base their financial and commercial contracts on Islamic law; their claims induced their own court systems to emulate Islamic legal practices. Consequently, until the 18th century the economic performance of non-Muslim peoples of the Islamic world did not diverge significantly from that of Muslims. Most non-Muslim communities started pulling ahead, however, as western Europe developed the legal infrastructure of modern capitalism. Vast numbers of Christians, Jews and other non-Muslims gained an economic advantage over Muslims by doing business under western or western-inspired laws (Kuran, 2004b; Issawi, 1982).

The waqf

Another contributor to the Islamic world's economic successes and also to its subsequent economic retardation is the waqf, Islam's distinct form of trust. From the eighth century to modern times, Muslim-governed states provided few public goods directly, beyond law and order. They left the supply of public goods largely to waqfs established in a decentralized manner. Vast resources flowed into waqfs; by the early 18th century they owned between a quarter and half of all real estate, depending on the country. The services financed through waqfs included mosques, schools, hospitals, water fountains, roads, parks, inns, bathhouses, orphanages and soup kitchens.

A waqf is an unincorporated trust established under Islamic law by an individual owner of immovable property for the perpetual provision of a service. It emerged in the early Islamic period, a time of weak property rights, partly to enable landowning high officials to shelter wealth. Converting property into waqf yielded considerable immunity against confiscation, because waqf-owned assets were considered sacred, and this made legitimacy-seeking rulers reluctant to expropriate them. In addition to social status and religious satisfaction, the founder usually obtained pecuniary benefits. He could make himself the waqf's mutawalli (trustee and manager), set his own salary, appoint relatives to paid positions, and designate his successor. This last prerogative enabled circumvention of the Islamic inheritance system. In founding a waqf, then, an individual did not simply engage in charity. In return for shouldering social responsibilities, he obtained the privilege of sheltering wealth for personal use. Local norms determined the share of a waqf's income that its mutawalli could reserve for himself and his family.

For a millennium this system for supplying public goods remained a distinguishing feature of the Islamic world. It owed this remarkable longevity to identifiable benefits that it yielded to huge groups. Property owners achieved a measure of material security. Rulers unburdened themselves of the responsibility to provide public goods. And the average person received diverse forms of philanthropy. Nevertheless, the waqf system had a flaw that became increasingly serious over time. Although some opportunities existed to reallocate resources to new uses, the waqf was designed to serve its founder's wishes for ever. As such, it could not adapt quickly to changing social needs, and it locked capital into inefficient uses. By the 19th century, a time of massive technological change, the waqf system had become conspicuously dysfunctional, and reformers took to dismantling it (Çizakça, 2000; Kuran, 2001).

Up to that time, services to the Middle East's great cities were supplied mostly by waqfs. The 19th century saw the establishment of the region's first municipalities, under secular laws. These municipalities, which attained corporate powers, could reallocate resources relatively quickly. Within a few decades, they assumed most of the functions previously relegated to the waqf sector.

Absence of the corporation

Islamic law, which borrowed from various pre-existing legal systems, had spurned the Roman concept of the corporation. Limiting legal standing to natural persons supported Islam's political mission, which was to turn Arabia's feuding tribes into an undivided religious community. Corporations might have undermined that goal by enabling tribes to form autonomous organizations. During the formative period of Islam – from the seventh through the tenth centuries – the Middle East thus experienced no incorporation wave analogous to that observed in contemporaneous western Europe. One reason is that the waqf, by providing the means for delivering perpetual services with large sunk costs, alleviated the need for corporations. Another is that the waqf system spawned constituencies with a stake in preserving its key features; yet another that merchants and producers who stood to benefit from corporate powers

could not muster the collective action necessary to reform the legal system. Not until the modern era did the concept of the corporation enter legal systems of the Islamic world.

In the late 20th century, certain predominantly Muslim countries started to revive their waqf sectors, though in modernized and secularized form. Unlike the traditional waqf, a modern waqf enjoys legal personhood, and its founder may be a group. It is managed by a mutawalli board rather than a single caretaker appointed for life. Most critical, as a self-governing organization it can remake itself. Secularists, civil rights groups and economic liberalizers are the key constituencies of coalitions formed to promote waqf founding. These groups see their mission as a vehicle for shrinking the state, strengthening local governance, and promoting democratization (Çizakça, 2000). Thus, having played an enormous role in Islamic economic history, the waqf is now turning into an agent of political and economic modernization.

Because the Qur'an does not mention the waqf, many Islamists are indifferent to ongoing efforts to reinvigorate the waqf sector. Their overriding goal is to purge interest from financial transactions, largely in the belief that the Qur'an bans interest categorically (Saleh, 1986; Lewis and Algaoud, 2001; Kuran, 2004a). In fact, Islam's prescriptions concerning interest have always been a matter of interpretation, and throughout Islamic history interest-based transactions have been common (Rodinson, 1966). Nevertheless, Islamists treat Islamic banking, intended to be free of interest, as the sine qua non of a properly Islamic economy. Yet Islamic banking is a modern creation. Pre-modern economies based on Islamic law had moneylenders but no banks (Udovitch, 1979). The first banks of the Islamic world, all foreign-owned and -managed corporations, date from the mid-19th century (Kuran, 2005).

Property rights

Until modern times economies of the Islamic world suffered from a lack of institutions to tie the hands of governments. This meant that private property rights remained weak. Although material insecurity varied across time and space, taxation was often arbitrary, and states resorted to compulsory labour. Private property rights did not achieve credibility even in the eyes of state officials – one reason why endowing waqfs was so popular. A scribe could be plucked out of obscurity to become a prosperous statesman, and then, all of a sudden, fall into disgrace and lose everything. The expropriation of large estates was especially common, all the more so in times of financial crisis. Because this practice violated the Islamic law of inheritance, typically it was based on the ground that the deceased was not the rightful owner of his estate (Findley, 1989).

In the seventh century, the first Islamic state in Arabia had instituted a tax-and-subsidy system that might have strengthened property rights. Known as zakat, it required the payment of taxes to the state in specific forms of income and wealth at predetermined rates. In providing the state the resources to fund various activities, including charity, it also capped taxation (Rahman, 1974). However, precisely because of the inflexibility of its rate structure, within a couple of generations revenue-hungry rulers abandoned it for taxes that gave them greater latitude. Thereafter zakat metamorphosed into a narrow religious duty, incumbent on people of means, to assist the poor on an annual basis (Kuran, 2003a). Modern Islamists have tried to turn zakat into a state-run social welfare system to which the wealthy make obligatory contributions. But throughout the Islamic world taxation remains an essentially secular matter. It has also become more predictable. The wealthy classes of the present have far better defences than those of

the past against government predation.

Taxation can be arbitrary without being devoid of logic. In pursuing opportunities to raise revenue, rulers sought to limit transaction costs, in particular to minimize the costs of measuring income, identifying assets, and collecting taxes. To that end they tended to collect fixed taxes directly, leaving the collection of variable taxes to local officials (Co•gel and Miceli, 2005). They also made extensive use of tax farming, which assigns collection rights to people knowledgeable about tax units.

See Also

- corporations
- development economics
- institutional trap
- public goods
- religion and economic development

Bibliography

Çizakça, M. 1996. *A Comparative Evolution of Business Partnerships: The Islamic World and Europe, with Special Reference to the Ottoman Archives*. Leiden: E.J. Brill.

Çizakça, M. 2000. *A History of Philanthropic Foundations: The Islamic World from the Seventh Century to the Present*. Istanbul: Bo•aziçi University Press.

Co•gel, M. and Miceli, T. 2005. Risk, transaction costs, and tax assignment: government finance in the Ottoman Empire. *Journal of Economic History* 65, 806–21.

Findley, C. 1989. *Ottoman Civil Officialdom: A Social History*. Princeton: Princeton University Press.

•nalc•k, H. 1994. The Ottoman State: Economy and Society, 1300–1600. In *An Economic and Social History of the Ottoman Empire, 1300–1914*, ed. H. •nalc•k and D. Quataert. Cambridge: Cambridge University Press.

Issawi, C. 1982. The transformation of the economic position of the *Millets* in the nineteenth century. In *Christians and Jews in the Ottoman Empire*, vol. 1, ed. B. Braude and B. Lewis. New York: Holmes and Meier.

Kuran, T. 2001. The provision of public goods under Islamic law: origins, impact, and limitations of the waqf system. *Law and Society Review* 35, 841–97.

Kuran, T. 2003a. Islamic redistribution through *zakat*: historical record and modern realities. In *Poverty and Charity in Middle Eastern Contexts*, ed. M. Bonner, M. Ener and A. Singer. Albany: State

University of New York Press.

Kuran, T. 2003b. The Islamic commercial crisis: institutional roots of economic underdevelopment in the Middle East. *Journal of Economic History* 63, 414–46.

Kuran, T. 2004a. *Islam and Mammon: The Economic Predicaments of Islamism*. Princeton: Princeton University Press.

Kuran, T. 2004b. The economic ascent of the Middle East's religious minorities: the role of Islamic legal pluralism. *Journal of Legal Studies* 33, 475–515.

Kuran, T. 2005. The logic of financial Westernization in the Middle East. *Journal of Economic Behavior and Organization* 56, 593–615.

Lewis, M. and Algaoud, L. 2001. *Islamic Banking*. Cheltenham: Edward Elgar.

Rahman, F. 1974. Islam and the problem of economic justice. *Pakistan Economist* 14(24 August), 14–39.

Rodinson, M. 1966. *Islam and Capitalism*, trans. B. Pearce. New York: Pantheon, 1973.

Saleh, N. 1986. *Unlawful Gain and Legitimate Profit in Islamic Law: Riba, Gharar, and Islamic Banking*. Cambridge: Cambridge University Press.

Udovitch, A. 1979. Bankers without banks: commerce, banking, and society in the Islamic world of the Middle Ages. In *The Dawn of Modern Banking*, ed. Center for Medieval and Renaissance Studies. New Haven: Yale University Press.

How to cite this article

Kuran, Timur. "Islamic economic institutions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000257> doi:10.1057/9780230226203.0854

The New Palgrave Dictionary of Economics Online

Islamic finance

Mahmoud El-Gamal

From The New Palgrave Dictionary of Economics Online, , 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Islamic laws on financial matters date back to medieval times. They were reinterpreted in the 20th century to provide guidelines for the burgeoning Islamic financial sector. Compliance with religious law is a driving force in this sector, and a variety of financial instruments have been developed that are adjudged to be acceptable for use by Muslims. These continue to evolve, and the rules could be considered to merit reinterpretation to better enable Islamic financial institutions to deal with risk factors and obey the spirit, rather than merely the letter, of medieval Islamic jurisprudence, which was regulatory in nature.

Keywords

Islamic finance; *Shari‘a* law; *mudaraba*

Article

The notion of ‘Islamic finance’ was born during the tumultuous identity-politics years of the mid-20th century. Indian, Pakistani and Arab thinkers contemplated independence from Britain, and the independence of Pakistan from India, within a context of ‘Islamic society’. Islam was assumed to inspire political, economic and financial systems that are distinctive and independent of the Western (capitalist) and Eastern (socialist) models of the epoch. The term ‘Islamic economics’ was coined by Abu al-A‘la Al-Mawdudi, whose students and followers worked to develop an ostensible Islamic social science (Kuran, 2004). Mawdudi's influence on Arab Islamists began with the writings of Sayid Qutb, the father of modern Arab political Islam, whose quasi-exegesis *Under the Qur’anic Shade* referred exclusively to Mawdudi's writings on economic matters. Mawdudi's migration from majority-Hindu Indian society to majority-Muslim Pakistan thus became a prototype for Islamist migration away from secular political and economic systems.

From Islamic economics to Islamic banks

In the first few decades of its existence, Islamic economics focused on comparative economic systems (a fashionable field at the time) as well as neoclassical and Keynesian modelling with a highly stylized *homo islamicus* (a moral and ethical individual who shuns excessive greed and consumerism) in place of mainstream economics' *homo economicus* (a selfish utility and profit maximizer) (Haneef, 1995). As a byproduct, Islamic banking emerged in the Islamic economists' literature as a financial system based exclusively on profit-and-loss sharing, which was argued to be more equitable and stable (Chapra,1996; Siddiqi, 1983). In the process, Islamic economists focused on the Islamic prohibition of *riba* or usury, which they interpreted as a prohibition of all interest-based lending, in accordance with earlier interpretations of the Judeo-Christian canon.

Classical Islamic jurisprudence had interpreted interest-based lending, the cornerstone of fractional-reserve depositary banking, as riskless – and therefore illegitimate and inequitable – return for idle capitalists. Indeed, the importance of credit and counterparty risk for any financial analysis remains conspicuously absent from the writings of the Islamic-economics faithful. The preferred financial model, they postulated, would be based on the ancient silent-partnership model known in Islamic writings as *mudaraba*, corresponding to the Jewish *heter iska* and the Christian-European *commenda* (Udovitch, 1970).

An 'Islamic bank' was envisioned as a two-tier silent partnership. Thus, deposits seeking a return (as opposed to fiduciary deposits, for which 100 per cent reserves are required) would not be guaranteed loans to the bank, but rather silent-partnership investments in the bank's portfolio. In turn, the bank's investments of those funds would not consist of loans and acquisition of debt instruments, but rather profit-and-loss sharing investments in other silent partnerships. Thus, the Islamic bank would serve its financial intermediation function (pooling of return-seeking savings and diversification of investments) through profit-and-loss sharing. This idea continues to serve as the cornerstone of Islamic banking today, despite being thoroughly debunked by prominent jurists (Tantawi, 2001; El-Gamal, 2003).

Potential loss of return-seeking deposits was assumed by Islamic-banking proponents such as the Islamic Financial Services Board (IFSB) to encourage depositor-monitoring and risk-mitigating market-based discipline. Thus, the grossly inadequate depositor-protection measures supported by the industry have focused on transparency of operations and profit-distribution mechanisms (IFSB, 2006).

The practice of Islamic banking

This risk-sharing model has continued to shape the liabilities side of Islamic banks' balance sheets, with a few exceptions in Europe and the United States, where regulators have required Islamic financial providers that function as banks to guarantee deposits. The assets side of Islamic banks and financial providers, on the other hand, has utilized multiple structured-financial models to replicate loans and fixed-return securities that limit the banks' exposure to credit risk. The transformation from the idealistic profit-and-loss sharing model of Islamic economics – which continues to be hailed as the 'Islamic ideal' by industry practitioners and commentators – to replication of modern financial products and markets in 'Islamic' garb coincided with the increased importance of classical methods of Islamic jurisprudence and a limited rhetorical role for Islamic economics.

Early models in the subcontinent during the 1950s and in Egypt during the 1960s notwithstanding, the true beginnings of Islamic banking and finance occurred in the mid-1970s. Islamic jurists including the Shiite scholar Baqir al-Sadr and many Sunni scholars in Egypt, Saudi Arabia and elsewhere collaborated with Islamist bankers to replicate loans using ancient contract forms. Baqir al-Sadr, in his classical work *The Non-Usurious Bank in Islam* (long out of print), had attempted to use similar structured products to replicate guaranteed bank deposits on the liabilities side. However, since risk-sharing depositors were clearly beneficial to the shareholders of Islamic banks, and because the latter drove innovation in Islamic banking through the retention of lawyers and religious scholars, most of the 'innovations' were restricted to the assets side of the balance sheet.

Murabaha (cost-plus sale) financing

The workhorse of Islamic banking has been the *murabaha* (cost-plus sale) contract. The logical evolution of this form of finance is indicative of the general methodology of Islamic finance to this day. In the early 1980s, Islamic banks in the Gulf were flush with petrodollars, and Western corporations were eager to borrow from them as Western-bank credit dried up following the petrodollar-driven Latin American debt crisis. Islamic banks resorted to the easiest ancient trick: introducing a property to separate lent principal from repaid principal plus interest.

In the simplest ruse, the bank could have sold some commodity to its potential borrower on credit (for principal plus interest payable later), and then bought it back for cash (principal paid immediately), thus effectively replicating the cashflows of the loan, with the commodity making a round trip from bank to customer and back. However, this ancient ruse was forbidden by name as same-item sale resale (*bay' al-'ina*). In practice, one credit sale and one spot sale of liquid commodities were still used to accomplish the desired goal by conducting the second spot sale with a third party.

Interestingly, Al-Rajhi Investment Company in Saudi Arabia, which has one of the strictest religious-scholar boards, received a question on the legitimacy of the credit sale of gold, and ruled that such sales were disallowed because gold is a monetary commodity. Promptly thereafter, the same board was asked if platinum can be sold on credit, and issued a *fatwa* that this was permitted. Thus, Islamic banks could simply trade precious metals, acquiring an amount of platinum (or other metal excluding gold and silver) equal in value to the desired loan principal. The metal was then sold on credit to the Western borrower under a *murabaha* contract, with a credit price equal to the desired principal plus interest. The customer was then able to sell the metal quickly to receive the desired borrowed principal, perhaps less a small transaction cost.

This was the juristic solution first popularized by the late banker Sami Humud in his book *Evolving Banking Transactions in Accordance with Islamic Law* (1976). The prohibition of *riba* (usury) in the Islamic canon and subsequent juristic analysis left room for such ruses. The Qur'an merely mentioned *riba* in the abstract without specifying precisely which transactions were thus forbidden. The Prophetic tradition merely listed six commodities: gold, silver, dates, wheat, barley and salt, all of which were used at some point as commodity monies in the ancient world, stipulating that those may be traded only hand-to-hand and in equal amounts measured by weight or volume. One school of jurisprudence (*Hanafi*) expanded the prohibition to all commodities measured by weight or volume, but still did not treat them as money. Therefore, while trading platinum now for platinum later, or trading gold now for silver later, would both be deemed impermissible based on the *Hanafi* interpretation, trading platinum now for

dollars later was considered permissible.

Interestingly, the *Halacha*, developed by Jewish scholars prior and in parallel to the development of Islamic *Fiqh*, forbade such embedded-interest credit sales (Reisman, 1995, p. 112). In contrast, all major schools of Islamic jurisprudence (four Sunni and four Shiite) have allowed credit sales at prices possibly exceeding the spot price. Initially, this was only a method for seller financing. Thus, the financier needed first to acquire the property before selling it on credit. In addition, to give the contract an Islamic flavour, the industry adopted the name of an ancient cost-plus sale – *murabaha*, a contract devised to protect buyers who were unfamiliar with market prices, allowing them to negotiate prices by negotiating markup over revealed cost.

The contract that emerged in the 1970s was formally known as ‘cost-plus sale to the customer who ordered the initial purchase’ (*murabaha lil-‘amir bil-shira‘*). It was initially subject to scholarly controversy, especially as bankers added provisions to eliminate all forms of risk other than customer credit. In order to eliminate property-related risks, which were ironically the basis on which jurists allowed earning a return on the transaction, they allowed banks to stipulate that the eventual buyer must guarantee to buy the property on credit once the bank acquires it. Eventually, wide consensus emerged and the contract became the workhorse of Islamic banking practices, from large multi-million-dollar loans to Western corporations to retail-bank secured lending.

Tawarruq (monetization) financing

In order to reduce transaction costs, especially for retail customers who wished to borrow cash, Islamic banks in Gulf Cooperation Council (GCC) countries revived another ancient financial trick: monetization. This transaction is very similar to the cost-plus commodity-sale finance model, with the added complication that the Islamic bank executes all three legs of the transaction: (i) buying the principal's worth of metals at the spot price, (ii) selling said metals to the customer on credit for principal plus interest, and (iii) selling the metals back to the dealer, as the customer's agent, for the spot price less a small fee. All three transactions can be concluded within minutes via fax.

This transaction avoids the forbidden two-party sale–resale trick by adding not only one commodity as a degree of separation between lent principal and repaid principal plus interest, but also a third-trading-party degree of separation (the metals dealer) so that every two parties formally trade the commodity only once. The commodity still completes a round-trip (dealer→bank→customer→dealer), spot cash in the amount of desired principal completes one trip (bank→dealer→customer), and the credit-sale-price payment of principal plus interest occurs in the future (customer→bank). This three-party variation on same-item sale resale was also known in ancient and medieval practice, and deemed forbidden or reprehensible by most schools of law. Some medieval scholars within the *Hanbali* school of jurisprudence, which is dominant in the GCC, had permitted this practice. Despite the fact that the most respected 14th-century scholars ibn Qayim and ibn Taymiya forbade the transaction (as merely an expensive and potentially more hazardous type of usury/*riba*), contemporary *Hanbali* jurists who dominate one juristic council in Saudi Arabia permitted the practice in 1998. The same council later forbade the organized practice of Islamic banks using this contract in 2003, but the practice continued to thrive.

Ijara (lease) financing, securitization and sukuk (Islamic bonds)

Despite juristic approval of credit-sale-based financing methods, the practice remained suspect in scholarly as well as general Islamist circles. In addition to objections that the practice merely replicated interest-based financing with interest characterized as profit or markup, there were problems with securitization and trading of receivables from *murabaha* and *tawarruq* facilities. Those problems arose from the fact that most jurists, with the notable exception of those in Malaysia, forbade trading of debts, except under very strict transfers at face values and resale to the debtor. This prevented the development of secondary markets that would allow banks to diversify their portfolios and sources of funds. Lease financing provided a partial solution to both problems: it was ostensibly based on real assets that continued to play a role throughout the life of the financial facility, and it was possible to trade lease receivables on secondary markets as ostensible shares in the leased assets.

Jurists were adamant that Islamic lease or *ijara* financing must be truly asset-based, and therefore must be structured as operating rather than financial leases. However, recent advances in structured finance – which helped corporations such as Enron to move debts and interest payments off balance sheets through sale-leaseback structures – had blurred the line between operating and financial leases. As a result, a prestigious juristic council declared in 2008 that more than 80 per cent of lease-based bond (*sukuk*) structures were unIslamic, since material ownership of the underlying assets was not real.

Developed initially as another mode of secured lending, lease financing proceeded by acquiring durable assets and leasing them with an option to buy – principal plus interest passing to the lessor as rent plus potential final payment. For banks in countries that forbid them from owning real estate, special purpose vehicles (SPVs) received credit that were used to acquire the assets and lease them till maturity. Shares in those SPVs were treated as shares in the leased properties, thus allowing them to trade on secondary markets. In the United States, such structures were used to originate mortgage loans that were then securitized through Fannie Mae and Freddie Mac, and marketed both domestically and in the cash-rich GCC, especially after the second wave of petrodollar flows began in 2001.

Bond structures were easily adapted from these financial forms. An entity that wished to issue a bond would create an SPV, which sold shares for the amount of financing desired. The proceeds of that sale were used to buy some asset from the originator, which asset was promptly leased back. The originator would thus collect the proceeds of the sale of its asset as principal, and pay principal plus interest in the form of rent and/or a final repurchase price, which payments were passed through to the *sukuk* or bond holders. An added advantage of this structure is that the payments were made ostensibly on shares in ownership of the real asset, thus the contract could be advertised as a form of partnership, which appealed to the earlier political-Islam inspired literature on Islamic economics.

Jurists further facilitated securitization of debts by allowing a portfolio of asset-based and purely debt-based receivables (that is, lease-based and credit-sale-based, respectively) to be traded as long as the asset-based component exceeded 51 per cent of the total face value (Usmani, 1998). This strange provision clearly imposed no significant constraints on securitization, since successive portions of pure-debt receivables could be bundled iteratively with the same asset-based ones, which could be bought back repeatedly for the purpose of bundling with pure-debt tranches. Thus, Islamic finance became an equal partner in the credit bubble the ensued in the first decade of the 21st century. In fact, the volume of *sukuk* remained sufficiently small (relative to demand by Islamic banks) to merit abnormally high prices and low yields relative to conventional debts issued by the same entities.

Islamic mutual funds

A widely publicized area of Islamic finance was the development of ‘screening’ methods to identify ‘*Shari‘a* compliant’ stocks. These screens excluded stocks of companies with significant forbidden activities (such as breweries), and also of firms with excessive debt or interest income. The debt screen chosen by the industry was particularly perplexing, as it excluded firms with debt to market capitalization ratios exceeding one-third. This rule clearly forced fund managers to buy high and sell low in highly volatile markets. Moreover, the rule diverted funds away from Muslim-owned companies, which were not allowed any degree of unsecured-loan leverage, in favour of western firms with moderate levels of leverage. The financial screens themselves had no foundation in Islamic law or reasonable economic analysis, starting as they did at 5 per cent debt to assets and evolving during the tech-stock bubble of the late 1990s into 33 per cent of debt to assets and then 33 per cent of debt to market capitalization. It is not clear whether and when these rules can be replaced with sensible ones.

Takaful (Islamic insurance) and derivatives

One of the fast-growing sectors in Islamic finance is an Islamic alternative to commercial insurance known as *takaful* (mutual support). The rhetoric of this sector is based on the idea of mutual protection against losses, but most *takaful* companies to date have not been structured as mutual insurance companies (in which policyholders and shareholders are the same individuals). Instead, *takaful* companies are generally shareholder-owned and act through silent partnership or agency to invest the policyholders’ premiums and pay legitimate claims in the form of ‘voluntary contributions’ – thus avoiding the Islamic prohibition of *gharar*, which includes trading known amounts (policy premia) for uncertain future amounts (on potential valid insurance claims). The prohibition of *gharar* was also invoked to forbid derivative securities, but forwards and options were easily synthesized from the ancient contracts of *salam* (prepaid forward sale) and *‘urbun* (downpayment call option), respectively.

Substance and form

El-Gamal (2006, 2008) has argued that the essence of the ancient religious law was regulatory. It is well known in financial economics that financial innovators eventually find means to circumvent outdated regulation, thus increasing systemic risk. Financial crises later propel political and economic authorities to impose further regulations for innovators to circumvent. In this regard, the ancient religious regulations enshrined in medieval Islamic jurisprudence, especially if interpreted naively as prohibitions of certain contracts and permissions of others, are woefully out of date, and therefore ceased to perform their regulatory function centuries ago. Indeed, that is precisely why majority-Muslim societies had abandoned those outdated contract-based frameworks before the Islamist revisionism of the mid-20th century. The ancient law, which is not uniquely Islamic, does contain many lessons for today's societies – Muslim and otherwise. However, rent-seeking behaviour by bankers, lawyers and religious scholars on the one hand, and incoherent pietism and adherence to fictional Utopian history on the other, have prevented societies from adapting this centuries-old accumulated human wisdom for any purpose

beyond short-term self-enrichment and identity-political appeasement, both of which increase rather than ameliorate systemic risks.

See Also

- Islamic economic institutions

Bibliography

Chapra, U. 1996. What is Islamic economics? IDB Prize Winner's Lecture Series, No. 9. Jeddah: Islamic Development Bank.

El-Gamal, M. 2003. *Interest* and the paradox of contemporary Islamic law and finance. *Fordham International Law Journal* 27(1), 108–49.

El-Gamal, M. 2006. *Islamic Finance: Law, Economics, and Practice*. Cambridge: Cambridge University Press.

El-Gamal, M. 2008. Incoherence of contract-based Islamic financial jurisprudence in the age of financial engineering. *Wisconsin International Law Journal* 25(4), 605–23.

Haneef, M. 1995. *Contemporary Islamic Economic Thought: A Selected Comparative Analysis*. Kuala Lumpur: Iqrak.

Humud, S. 1976. *Tatwir Al-A‘mal Al-Masrifiyya bima Yattafiqu wa Al-Shari‘ah Al-Islamiyya* (Evolving banking transactions in accordance with Islamic law). Cairo: Dar Al-Ittihad Al-‘Arabi lil-Tiba‘a.

Islamic Financial Services Board (IFSB). 2006. *Guiding Principles on Corporate Governance for Institutions Offering Only Islamic Financial Services*…. Kuala Lumpur: IFSB, Available at <http://www.ifsb.org/standard/ifsb3.pdf>, accessed 12 January 2009.

Kuran, T. 2004. *Islam and Mammon*. Princeton: Princeton University Press.

Reisman, Y. 1995. *The Laws of Ribbis*. New York: Mesorah Publications.

Siddiqi, M.N. 1983. *Banking without Interest*. Leicester: Islamic Foundation.

Tantawi, M. 2001. *Mu‘amalat Al-Bunuk wa Ahkamuha Al-Shar‘iyya*. Cairo: Nahdat Misr.

Udovitch, A. 1970. *Partnership and Profit in Medieval Islam*. Princeton, N.J.: Princeton University

Press.

Usmani, M.T. 1998. *An Introduction to Islamic Finance*. Karachi: Idaratul Ma‘arif.

How to cite this article

El-Gamal, Mahmoud. "Islamic finance." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_I000309> doi:10.1057/9780230226203.1894

The New Palgrave Dictionary of Economics Online

IS – LM in modern macro

Edward Nelson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The IS–LM framework is associated with traditional macroeconomics, but versions of IS and LM functions can be justified using dynamic general equilibrium models that assume optimizing behaviour on the part of the private sector. The baseline version of these optimizing IS–LM relationships is discussed. Relative to the traditional IS–LM specification, the IS relationship in the optimizing IS–LM framework involves an extra term, which reflects the dependence of real aggregate demand on the expected level of spending next period. This extra term is implied by the intertemporal behaviour of households.

Keywords

aggregate demand; dynamic stochastic general equilibrium (DSGE) models; infinite horizons; IS–LM in modern macro; IS–LM model; monetarism

Article

Background: traditional IS – LM

Some discussions use the term ‘IS–LM’ as a catch-all label for the approach of traditional Keynesian economics. The treatment here, however, will follow Sargent (1987, p. 53) in interpreting ‘IS–LM’ narrowly (and literally) as a pair of structural equations describing real aggregate spending as a function of the real interest rate, and real money demand as a function of scale and opportunity cost variables. From that perspective, it is not strictly accurate to talk of an ‘IS–LM model’ (since IS–LM is only a portion of a macroeconomic model) or to refer to the ‘sticky-price assumption’ of IS–LM (properly specified IS and LM equations are structural and should be independent of what is assumed about price behaviour, whose specification belongs to the supply side of a model; and IS–LM analysis in conjunction with price flexibility was considered even in Hicks, 1937). Nor should models that have

separate equations describing consumption and investment behaviour be considered models containing IS–LM equations, since deriving an IS equation necessarily involves eliminating the components of aggregate demand in favour of an expression for total demand.

It follows that, to find descendants of traditional IS–LM in modern macroeconomics, one should focus on cases where general equilibrium models produce a pair of equations clearly recognizable as corresponding to IS (real aggregate spending) and LM (real money demand)-type relationships.

IS – LM in modern macro: early literature

Early attempts to link IS–LM with dynamic optimizing macroeconomics include Aiyagari and Gertler (1985) and Fane (1985). These attempts, however, did not use infinite-horizon agents (the standard assumption in modern macroeconomics) and usually left more endogenous variables than output and the real interest rate in the equation for total spending, so this equation was not clearly recognizable as an IS relationship.

This early literature did show that it was possible to derive a conventional money demand equation from an optimizing model. This was also shown by McCallum and Goodfriend (1987) using an infinite-horizon model. A semilogarithmic version of McCallum and Goodfriend's money demand equation is:

$$rm_t = c_1 c_t + c_2 R_t \tag{1}$$

where $c_1>0$, $c_2<0$, and rm_t and c_t denote log-deviations of real money balances and real household consumption from their respective steady-state levels, with R_t being the short-term net nominal interest rate minus its steady-state value.

In light of the feasibility of deriving an LM relationship from an optimizing general equilibrium model, most discussions concentrated on whether IS-type relationships, and therefore IS–LM as a whole, are compatible with optimizing behaviour. A symposium on the subject of IS–LM and modern macroeconomics (Young and Zilberfarb, 2000), which largely predated the recent literature, was generally negative about the prospects of linking up modern macroeconomics with IS–LM.

IS – LM in modern macro: later literature

In discussing the recent literature, it is worthwhile first stepping back to Hall (1978, p. 974), who showed that an infinite-horizon dynamic general equilibrium model implied an equation for aggregate household consumption (C) of the form

$$C_t^{-(1/\sigma)} = \beta(1 + r_t) E_t C_{t+1}^{-(1/\sigma)}$$

(2)

where r_t is the short-term net real interest rate, β is the household's discount factor and $\sigma > 0$ is a utility function parameter (with a large σ value implying high intertemporal substitution in consumption). A log-linearized version of this equation is:

$$c_t = -\sigma(r_t - E(r)) + E_t c_{t+1}$$
(3)

where E(r) denotes the steady-state value of r_t. Consumption equations such as (3) continue to be present in the dynamic stochastic general equilibrium models prevalent today. What is different in the recent literature is a change in emphasis in interpreting the equation. Hall (1978) treats the real interest rate as fixed and focuses on the implied univariate behaviour of consumption. The recent literature does not treat the real interest rate as fixed, and instead builds up from the consumption condition (3) to an economy-wide description of aggregate real spending behaviour. If consumption is the only component of aggregate demand (implying the relation $c_t = y_t$), then eq. (3) implies an aggregate relationship – the optimizing IS equation – of the form:

$$y_t = b_1(R_t - E_t \pi_{t+1}) + E_t y_{t+1}$$
(4)

where $b_1 = -\sigma < 0$, π_t is inflation minus its steady-state value, and the Fisher condition $(r_t - E(r)) = R_t - E_t \pi_{t+1}$, has also been substituted in. Under the same approximation that output equals consumption, log output becomes the scale variable in the money demand function, so that eq. (1) implies an LM relationship:

$$m_t = c_1 y_t + c_2 R_t.$$
(5)

Alternative assumptions to that of strict equality between consumption and output will deliver much the same IS relationship as eq. (4). For example, one could assume constant but non-zero investment, or

random-walk exogenous investment behaviour (as in McCallum and Nelson, 1999), or proportionality between consumption and investment, and in each case derive an IS equation isomorphic to eq. (4). Whatever the precise derivation, the common element in the recent literature that starts from the Euler consumption condition is that, instead of making restrictions, as Hall did, that lead to conclusions about the unforecastability of consumption growth, it sees the condition as underpinning a structural relationship describing the level of total real spending. In this relationship total spending is a function of the real interest rate, expected future spending levels, and exogenous shocks. The negative coefficient on the real interest rate allows a parallel with the traditional interest-elastic IS relationship $Y = f(r)$. That parallel has been highlighted by the later IS–LM literature, including Koenig (1989; 1993), McCallum (1989, p. 105), Woodford (1995; 2003), Kerr and King (1996), Rotemberg and Woodford (1997), and McCallum and Nelson (1999).

Shocks

It is straightforward to justify the addition of exogenous shock terms to the optimizing IS and LM equations. Preference shocks in the household's utility function can deliver this result: for the IS equation the shock is to the marginal utility of consumption; the LM shock, on the other hand, is a combination of the shocks to the marginal utility of consumption and to the marginal utility of services generated by real money balances. In addition, the portion of government spending that is not well approximated by a (log) random walk will produce a further rationale for an IS shock.

Treatment of capital

As noted above, a restrictive assumption about investment (that is, that it is constant or random-walk exogenous) is needed to derive the optimizing IS eq. (4). Dupor (2001) criticizes such approximations on the grounds that investment is a sizable portion of aggregate demand and a major contributor, in arithmetic decompositions, to real GDP fluctuations. These facts can be accommodated, however, without making investment endogenous. One can simply assume that investment has a random-walk and a stationary component, both exogenous. The exogenous stationary component becomes a further IS shock and can be assumed to be highly variable.

General equilibrium status

Early discussions of dynamic general equilibrium models stressed the interdependence of aggregate demand and supply relationships, and, that being so, the infeasibility of labelling a subset of equations specifically ‘aggregate demand’ equations (see, for example, Sargent, 1982). By contrast, the approach that derives IS and LM relationships from a general equilibrium analysis emphasizes that a subset of equations may be labelled ‘aggregate demand’ relationships; and that other conditions describing private sector behaviour (such as firms’ pricing and hiring decisions and households’ labour supply condition) constitute the aggregate supply block. Common separability assumptions regarding the private sector's preference and cost functions justify this division of equations. The central assumption is that the terms

involving consumption, leisure, and real balances are additively separable in the private households' utility functions.

Is it an IS equation?

The optimizing IS function has been criticized as not descriptive of the original 'investment–saving' acronym, as its baseline version comes from a model with no investment fluctuation and with saving zero or constant in equilibrium. But 'IS' was always a description of how aggregate spending fluctuations related to interest-rate variations – an 'income sensitivity' or 'interest sensitivity' equation rather than really an 'investment–saving' relationship. Detailed discussion of saving issues typically would not use the assumptions (for example, those regarding infinite horizons for agents) underpinning baseline optimizing macroeconomic models.
Alternatively, the old 'investment–saving' label could be justified on the grounds that the IS equation forms part of a model describing the process by which investment and saving are equated. That description remains true of the optimizing IS equation; it happens that in the baseline model underlying this equation, the equilibrium occurs with saving and investment at constant or zero values.

Other interest rates

It is tempting to suggest that the optimizing IS eq. (4) is subject to the monetarist critique of traditional IS–LM because it excludes money from the IS equation. But in fact monetarists did not argue that money belonged in the structural IS equation. Instead, they argued that many yields mattered for aggregate demand and that these yields could not be summarized by a single interest rate (see, for example, Brunner and Meltzer, 1973). Variations in money acquired significance because this spectrum of yields also appeared in the money demand function. The monetarist critique amounts to the suggestion, first, that different financial assets are not perfect substitutes, and second, that the discrepancies between the yields might be related to the behaviour of money. Baseline IS–LM, both old and new, presumes perfect substitutability between assets, in which case the short-term real interest rate is tightly related to other real returns prevailing in the economy. McCallum and Nelson (1999) defend the perfect-substitution assumption as the appropriate benchmark for many purposes. Nevertheless, as Bernanke and Reinhart (2004) argue, for some policy issues this assumption is not appropriate and so it would be desirable to break the link between different returns on assets, and investigate the effect of monetary policy actions on various yields. Such a generalization of IS–LM would tend to put extra real yields into the IS equation and extra nominal yields into the LM function.

See Also

- Hicks, John Richard
- IS–LM

Views expressed in this paper are the author's and should not be interpreted as those of the Federal

Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

Bibliography

Aiyagari, S.R. and Gertler, M. 1985. The backing of government bonds and monetarism. *Journal of Monetary Economics* 16, 19–44.

Bernanke, B.S. and Reinhart, V. 2004. Conducting monetary policy at very low short-term interest rates. *American Economic Review* 94, 85–90.

Brunner, K. and Meltzer, A.H. 1973. Mr. Hicks and the ‘monetarists’. *Economica* 40, 44–59.

Dupor, B. 2001. Investment and interest rate policy. *Journal of Economic Theory* 98, 85–113.

Fane, G. 1985. A derivation of the IS–LM model from explicit optimizing behavior. *Journal of Macroeconomics* 7, 493–508.

Hall, R.E. 1978. Stochastic implications of the life-cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86, 971–87.

Hicks, J.R. 1937. Mr. Keynes and the ‘Classics’: a suggested interpretation. *Econometrica* 5, 147–59.

Kerr, W. and King, R.G. 1996. Limits on interest rate rules in the IS Model. *Federal Reserve Bank of Richmond Economic Quarterly* 82, 47–75.

Koenig, E.F. 1989. A simple optimizing alternative to traditional IS–LM analysis. Manuscript, *Federal Reserve Bank of Dallas*.

Koenig, E.F. 1993. Rethinking the IS in IS–LM: adapting Keynesian tools to non-Keynesian economies, part 1. *Federal Reserve Bank of Dallas Economic Review* 78, 33–49.

McCallum, B.T. 1989. *Monetary Economics*. New York: Macmillan.

McCallum, B.T. and Goodfriend, M.S. 1987. Demand for money: theoretical studies. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, P. Newman and M. Milgate. London: Macmillan.

McCallum, B.T. and Nelson, E. 1999. An optimizing IS–LM specification for monetary policy and business cycle analysis. *Journal of Money, Credit, and Banking* 31, 296–316.

Rotemberg, J.J. and Woodford, M. 1997. An optimization-based econometric model for the evaluation

of monetary policy. *NBER Macroeconomics Annual* 12, 297–346.

Sargent, T.J. 1982. Beyond demand and supply curves in macroeconomics. *American Economic Review* 72, 382–9.

Sargent, T.J. 1987. *Macroeconomic Theory*, 2nd edn. New York: Academic Press.

Woodford, M. 1995. Price-level determinacy without control of a monetary aggregate. *Carnegie-Rochester Conference Series on Public Policy* 43, 1–46.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, NJ: Princeton University Press.

Young, W. and Zilberfarb, B.Z., eds. 2000. *IS–LM and Modern Macroeconomics*. Boston, MA: Kluwer Academic Publishers.

How to cite this article

Nelson, Edward. "IS–LM in modern macro." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000301> doi:10.1057/9780230226203.0856

The New Palgrave Dictionary of Economics Online

IS – LM

Steven N. Durlauf and Donald D. Hester

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The IS–LM model is a short-run macroeconomic analytical construct for studying an economy with idle productive resources. The diagram has been especially influential because its constituent curves are loci on which the goods market (IS curve) and the money market (LM curve) are respectively in equilibrium, making it possible to infer changes in fiscal policy and monetary policy, both separate and simultaneous. The model is prominent in elementary and intermediate macroeconomic textbooks, yet it fails to accommodate the main features of modern macroeconomic theory, although modern dynamic models are sometimes interpreted as having IS–LM type features.

Keywords

asset price equilibrium; Bretton Woods System; central bank independence; crowding out; deflation; Econometric Society; expectations; Federal Reserve System; fiscal drag; fiscal policy; Fleming, J.; floating exchange rate; general equilibrium; government budget constraint; Hicks, J.; inflation; inflationary expectations; international capital markets; IS–LM model; Kahn, R.; Keynes, J. M.; liquidity trap; lump sum taxes; macroeconomic volatility; Marshall–Lerner condition; microfoundations; monetarism; monetary policy; monetary policy rules; money demand; money supply; Mundell, R.; Mundell–Tobin effect; open market operations; Pigou effect; price control; progressive taxation; real balance effect; Robinson, J.; stabilization policy; sticky prices; tâtonnement; taxation of income; Taylor rule; uncertainty; Walras's Law

Article

The IS–LM model is a short-run macroeconomic analytical construct for studying an economy with idle productive resources. In the form exposed by Hansen (1949), it is a two-dimensional diagram with the abscissa measuring real income and the ordinate the real interest rate. It has been widely and successfully employed in interpreting macroeconomic policy and is prominent in elementary and

intermediate macroeconomic textbooks. A close antecedent, the SI–LL diagram, first appeared in print in an influential article by J.R. Hicks (1937) that proposed an interpretation of Keynes's *General Theory* (1936) and effectively made Keynes's contribution accessible to large numbers of students ever since. Hicks's diagram had nominal income measured on the abscissa and was unclear about whether the interest rate was real or nominal. If prices are fixed or ‘sticky’ these differences are immaterial, but assumptions about prices and their measurement loomed large in subsequent controversies and applications of the diagram. Lange (1938) appears first to have required that variables were real magnitudes. Reflecting the times of its origin, the model describes a closed economy.

The diagram has been especially influential because its constituent curves are loci on which the goods market (IS curve) and the money market (LM curve) are respectively in equilibrium. The intersection of the two curves is a point where both markets (and, through Walras's Law, the bond market) are in equilibrium. The labour market is not required to be in equilibrium. Because fiscal policy affects the goods market through tax, transfer and expenditure changes, the effects of fiscal policy can be inferred from the change in the intersection of the IS curve with a stationary LM curve. Similarly, because changes in the money stock affect only the LM curve, the effects of monetary policy can be inferred from the change in the intersection of the LM curve with a stationary IS curve. Finally, the effects of simultaneous changes in both fiscal and monetary policies can be predicted from the change in the intersection when both curves are moved.

By way of background, the *General Theory* contains no formal mathematical model and has only one diagram. ‘Keynes believed economics was over-addicted to “specious precision” – making perfectly precise what was in reality vague and *complex*. It is significant that he refused to present the “model” of the *General Theory* in mathematical form, even though he assembled its (verbal) elements in chapter 18’ (Skidelsky, 1994, p. 540). ‘The mathematicisation of the *General Theory* started immediately it (*sic*) was published but it was left to Hicks to map the mathematics on to a two-curve diagram which became the accepted form of the *General Theory*’ (Skidelsky, 1994, p. 611).

Hicks's article emerged from the September 1936 European meetings of the Econometric Society at Oxford where a symposium on ‘Mr. Keynes’ System’ was held. Other important papers from the symposium interpreting Keynes were by R.F. Harrod (1937) and J. Meade (1937); both were published slightly earlier than Hicks's paper, but contained no trail-blazing graphical apparatus. Young (1987, p. 29) claims that all three papers had the same underlying equation system, which differed from that of the *General Theory* but may have appeared in Keynes's lectures at Cambridge as early as 1934. All three papers analysed the relation between their interpretation of what underlay the *General Theory* and the pre-existing theoretical framework. Keynes apparently did not object to the specifications of Hicks, Harrod and Meade, but stressed the importance of expectations and uncertainty in his subsequent discussions (1937) of the general theory; expectations do not appear formally in the three authors’ equation systems. If one accepts Keynes's (1936, ch. 12) discussion of how long-term expectations are formed, it may indeed be specious to describe commodity market equilibrium as if it were lying on a stationary curve, but that in no way reduces the usefulness of the model for designing and interpreting policy. The effects of monetary and fiscal policy actions are unaffected by random shocks to the two curves. However, the effects might be affected if the curves are moved by expectations about present or future policy moves, as has been suggested by Lucas (1976). Subsequently, J. Robinson (1975) and R. Kahn (1984), two of Keynes's contemporaries when the *General Theory* was being drafted, objected

strenuously to the IS–LM formulation and Hicks himself (1982) indicated dissatisfaction with it.

Basic theoretical structure

In the standard formulation of the IS–LM model, the endogenous variables to be determined are the level of aggregate output *Y* and the real interest rate *r*. Aggregate demand in the goods market is modeled via the output identity:

$$Y = C + I + G + NX$$

where C denotes consumption, *I* investment, *G* government spending and *NX* net exports. The identity is given substance by replacing *C* with a consumption function and *I* with an investment function. Typically, IS–LM analysis assumes that consumption depends positively on disposable income, which equals output *Y* minus taxes $T(Y)$ (income taxes induce dependence of the level of taxes that are collected on income) whereas investment depends on the real interest rate; one could also endogenize government spending and net exports. This leads to the IS equation:

$$\text{IS}: Y = C(Y - T(Y)) + I(r) + G + NX.$$

Money market equilibrium is defined by equating money demand and money supply. Real money demand, denoted by *L* to capture the idea that the demand for money is the demand for liquidity, is assumed to depend negatively on the nominal interest rate, which by definition equals the real interest rate plus the expected inflation rate, π. Throughout, expected inflation will be treated as exogenous; as noted below a defect of the IS–LM framework is that it does not embody expectations in an interesting way. The real money supply, $\frac{M}{P}$, is treated as exogenous, as are the price level and inflation rate. The LM equation is:

$$\text{LM}: \frac{M}{P} = L(r + \pi, Y).$$

If the demand for money does not depend on the nominal interest rate, then the LM curve uniquely determines the level of output. This special case is of historical importance in understanding the monetarist perspective on macroeconomics. In contrast, if the demand for money is infinitely elastic at some exogenous nominal interest rate, the LM curve uniquely determines the real interest rate. This case

is also of historic interest as it is the first version of a liquidity trap. It should also be noted that the LM curve can be replaced with a more sophisticated system of asset price equilibrium conditions; a significant component of James Tobin's work on monetary economics well summarized in Tobin (1969), represented an effort to enrich IS–LM analysis via a richer specification of financial markets.

Regardless of the specific assumptions on the shapes of the IS and LM schedules, the equilibrium pair (r, Y) is determined by the simultaneous solution of these two equations. Comparative static analysis may be done by changing the various exogenous variables in the IS and LM equations. Notice that this system is entirely demand driven in the sense that it does not consider resource constraints in the determination of output.

The IS–LM model has a number of well-known implications with respect to the effects of changes in government policy. Increases in government spending G increase the equilibrium levels of r and Y and decrease the equilibrium level of I. The reduction in investment induced by an increase in government spending is known as crowding out. When money demand does not depend on the nominal interest rate, there is complete crowding out in the sense that the increase in G is completely offset by a decrease in I, so that aggregate demand is unaffected; the independence of money demand from the nominal interest rate has often been treated as a hallmark of monetarism, since in this case changes in fiscal policy have no real effects. Similar results occur when one considers tax changes; a reduction in either lump sum taxes or the income tax rate increases both Y and r. With respect to monetary policy, an increase in M leads to an increase in Y, a decrease in r and an increase in I. Hence, unlike expansionary fiscal policy, expansionary monetary policy increases investment, and so causes crowding in. The exception to this result is a liquidity trap in which changes in the supply of money are accepted by the public at the initial nominal interest rate.

Beyond the evaluation of exogenous changes in government policies on aggregate outcomes, the IS–LM model has also been used to evaluate alternative government policies. Poole (1970) is particularly notable in this regard. Poole compares the stabilization properties of a monetary policy that fixes the nominal money supply with one that fixes the real interest rate. He shows that, if macroeconomic volatility derives from shocks to the IS schedule, then a fixed money stock policy stabilizes an economy more than a fixed interest rate policy; in contrast, when aggregate fluctuations are generated solely by shocks to the LM schedule, a fixed interest rate completely eliminates aggregate fluctuations. The idea that the effects of a monetary policy rule depend on the type of shocks an economy experiences has proven to be of importance in contexts far beyond the IS–LM model; one sees echoes of Poole's reasoning in discussions of the Taylor rule for interest-rate setting.

The IS–LM model has also been used to study the effects of changes in other exogenous (from the perspective of the model) variables on the macroeconomic equilibrium. An especially important issue concerns changes in the price level, because one wants to know whether price adjustments can move aggregate output towards a level consistent with full employment. In the specification so far described, a decrease in the price level raises Y and lowers r because a lower price level increases the real money supply, thereby shifting the LM schedule. This property was considered important in early expositions of IS–LM such as Modigliani (1944) that considered the possibility of a liquidity trap.

However, as argued by Pigou (1943; 1947), there is another channel through which lower prices can raise demand. Pigou argued that the level of consumption depends on the level of wealth as well as on disposable income. Since a component of wealth is nominal money, price reductions can increase demand through increases in the real value of money and, hence, wealth. This is called the Pigou (or real

balance) effect. Work on the real balance effect, in turn, affected the study of monetary policy in the IS–LM framework. The seminal paper in this regard is Metzler (1951), who argued that the real balance effect implied an important difference between the effects of a helicopter dropping of money, that is, an increase in the money supply in which additional money are simply added to individual portfolios, and an open market operation, in which an increase in the money supply is generated by the trading of money for bonds, so that one nominal asset is swapped for another, thereby keeping the aggregate nominal supply of assets constant.

Another well-known property of the model concerns the equilibrium effects of an increase in the (exogenously given) inflation rate. In the IS–LM model, an increase in π increases Y and lowers r. Intuitively, an increase in inflation reduces money demand, requiring an adjustment of the real interest rate and output to compensate. The fact that increases in inflation lead to less than one-to-one increases in the nominal interest is known as the Mundell–Tobin effect (Mundell, 1963a; Tobin, 1965). As before, this property depends on the dependence of money demand on the nominal interest rate.

From the perspective of modern macroeconomic theory, the IS–LM model has very serious deficiencies. One problem is that the model lacks well-defined microeconomic foundations. The price level is treated as exogenous; while prices may be sticky, the complete rigidity found in the IS–LM model is unappealing. Further, the consumption and investment functions are typically specified in an ad hoc fashion, rather than as the outcome of solving explicit decision problems.

Also, the IS–LM model fails to embody aggregate dynamics. The model constructs a snapshot of the macroeconomy without accounting for the fact that the snapshot is really one frame of a motion picture. While expectations variables can be introduced into the model, its static nature precludes one from considering many implications of the intertemporal government budget constraint for the real effects of changes in fiscal policy or the effects of expectations about monetary policy on the sequence of equilibrium price levels over time. This lack of dynamics was recognized early on as a defect of the model; heuristic analyses include Patinkin (1956) who tried to link IS–LM with tâtonnement adjustment of prices à la Walras. Indeed, Patinkin's book was the high point of efforts to link IS–LM with the conceptual structure of general equilibrium models. But this type of work disappeared rather quickly. Later on, efforts were made by Blinder and Solow (1973) and Tobin and Buiter (1976) to account for changes in the stocks of various assets on the IS–LM equilibrium, with particular attention paid to understanding how permanent changes in fiscal policy affect output in the presence of the requirement that the government budget balance with respect to the present discounted value of debt and taxes. These analyses found that accounting for such effects could imply that the long-run effect of a change in government spending exceeds the short-run change. One reason for this is the increased holdings of government debt induced by a fiscal expansion will increase consumption. This type of work also failed to have much effect on the use of the IS–LM framework.

To be fair, a number of recent authors have attempted to provide more rigorous microeconomic foundations to the IS–LM model; McCallum and Nelson (1999) is the most important example. See King (2000) for evaluation of such models. While progress has been made in producing better microeconomics for the IS–LM model, it seems fair to say that much of its use, especially in pedagogical and policy contexts, relies on the model we have described.

Mundell – Fleming model

A variant of the IS–LM model that has proven important in international economics is due to Mundell (1963b) and Fleming (1962); its value derives from the consideration of how the effects of monetary and fiscal policy are altered when one considers the role of the exchange rate. In this framework, a small country is assumed so that the income of the rest of the world, Y^{ROW}, is unaffected by events in the country. However, actions of the country may affect the exchange rate *e*, defined as the number of units of foreign currency one unit of the country's currency can purchase. Net exports for the country are assumed to obey

$$NX = NX\left(Y, Y^{ROW}, e\right).$$

Higher exchange rates are assumed to lead to a lower level of net exports; this is known as the Marshall–Lerner condition.

The effects of changes in monetary and fiscal policy will critically depend on the effects of a policy on the exchange rate. This, in turn, depends on the degree of integration of international capital markets. Following the classic Mundell–Fleming analysis, suppose that international capital markets are fully integrated. In this case, the real return on investments cannot differ across countries and so *r* is fixed since the economy under study is small. An increase in government spending by a small economy, in this case, will have no real effects. An increase in *G* will be entirely offset by exchange rate appreciation, that is, an increase in *e*, so that there is no net fiscal stimulus as the increase in government spending is fully offset by a reduction in net exports. In contrast, an increase in the money supply will induce an increase in output via exchange rate depreciation. These outcomes, of course, presume that the exchange rate is allowed to float.

In contrast, suppose that the central bank authority is committed to maintaining a given exchange rate, $\bar{e}$. In this case, the effects of monetary and fiscal policy are quite different. Maintenance of the exchange rate eliminates any independent role for the central bank in the sense that any action it takes to raise output will have to be undone in order to preserve the exchange rate. In contrast, a fiscal stimulus will induce a subsequent increase in the money supply in order to overcome the associated exchange rate appreciation, which reinforces the effects of the stimulus that are found in the closed economy model.

While current thinking on the interactions of the exchange rate regime with policy effects has moved far beyond the details of the Mundell–Fleming model, the ideas in the model not only proved of direct value for much subsequent research (for example, Tobin and Braga de Macedo, 1980) but has also, via its limitations, defined the agendas of alternative research directions (Obstfeld, 2001).

IS – LM in the light of macroeconomic history

Despite its theoretical limitations, the IS–LM model is illuminating in interpreting macroeconomic policy and events in the post-war period, especially in the United States, which could reasonably be viewed as a closed economy in the early years. Between the end of the Second World War and 1951, the

Federal Reserve was committed to a policy of restricting upward shifts in the yield curve in order to reduce the cost of financing the US government's large war debt. During this period, the Federal Reserve effectively ‘pegged’ the short-term nominal interest rate, so that monetary policy was expansionary whenever the real rate was negative and especially so when inflation was increasing. In the inflationary period between 1945 and 1948, the real short-term interest rate was substantially negative. Although deflation accompanied the recession of 1949, when the real rate turned briefly positive this unsustainable pegging policy together with the onset of the Korean War in 1950 led to the ‘Accord’ of 4 March 1951, which permitted the Federal Reserve to undertake discretionary monetary policy. This early inflationary monetary policy bias may not have been inappropriate, because demobilization after the Second World War led to a massive leftward shift of the IS curve due to falling government spending and essentially stagnant real net foreign and gross domestic investment until 1950. Inflation occurred with the suspension of price controls in 1946, but did not begin to accelerate until the war started in June 1950. The Korean War led to a 75 per cent increase in real government spending and a sizable increase in the real government deficit between 1950 and 1954. An expansion of accelerated depreciation allowances in 1954 was associated with a ten per cent increase in real investment in producers’ durable equipment in the subsequent three years. Both events caused a rightward shift in the IS curve. However, the shifts were rather insidiously being offset by ‘fiscal drag’ that resulted from increases in the marginal income tax rate that existed when the economy was at full employment. The tax rate rose because progressive tax schedules applied to nominal income, not income adjusted for inflation. Following the Accord, the Federal Reserve used open-market operations to fight inflation by raising nominal interest rates on several occasions with unsatisfactory results. Unemployment rates rose as interest rates rose, as might have been predicted from the resulting leftward shift in the LM curve. However, it was not widely understood that the IS curve was also shifting leftward because of the cumulative effects of fiscal drag. Seemingly, restrictive monetary policy induced three recessions in this decade with unemployment rates at troughs successively higher in each recession. Inflation temporarily abated during or shortly after each recession, but then returned, in part, because real short-term interest rates were infrequently positive until 1959.

The 1960 elections resulted in John F. Kennedy becoming president and, of greater significance for this discussion, a generation of economic advisors who understood and were intent upon applying the IS–LM model. The US Council of Economic Advisors’ *Economic Report of the President* explicitly focused on the importance of the full-employment budget surplus (1962, pp. 78–81) and, thus, fiscal drag. The Council of Economic Advisors also worried about the fact that real gross private domestic investment had been below its 1955 peak for the subsequent six years, a possible consequence of high interest rates. The administration negotiated an arrangement with the Federal Reserve whereby it would attempt to twist the yield curve through open-market operations by increasing short-term interest rates to protect the US gold stock from foreign withdrawals and lowering long-term interest rates to stimulate investment. The Treasury assisted in this effort with its debt management policies (1962, pp. 86–91). The administration increased federal government spending significantly beginning in the 1962 fiscal year and would subsequently stimulate the economy with tax cuts. The federal government deficit rose over the four years after 1960 and the IS curve shifted rightward. Because twisting of the yield curve involved two interest rates, it cannot be interpreted directly with the LM curve; however, between 1961 and 1964 the interest rate on three-month treasury bills rose 50 per cent, the rate on three–five year issues rose 6.4 per cent, and between December 1960 and December 1964 the level of Federal Reserve

credit outstanding rose by 37 per cent. In part because of continuing gold outflows, the last was the largest rate of growth of Federal Reserve credit until that date over any four-year span since the Accord. Net of gold flows Federal Reserve credit expanded by 28 per cent in this period of relatively low inflation. It is hard to argue that the LM curve didn't also shift rightward.

Tax cuts were phased in 1962 in the form of an investment tax credit that effectively reduced the required rate of return that profitable firms needed to undertake an investment project and in 1964 and 1965 in the form of 10 per cent reductions in corporate and personal income tax rates. The investment tax credit and accommodating monetary policy was associated with a 50 per cent increase in real gross private domestic investment between 1961 and 1966; the tax credit implies a rightward shift in the IS curve. The large tax rate cuts in 1964 and 1965 did not lead to an increase in the federal government deficit, partly because GDP rose considerably in response to rising investment. The unemployment rate fell from 6.7 per cent in 1961 to 3.8 per cent in 1966 and the average annual rate of inflation from the end of 1960 to the end of 1966 was less than two per cent, although it began to rise in the fourth quarter of 1965.

Because of rising inflation, a policy change occurred at the end of 1965 when the Federal Reserve signalled with an increase in its discount rate that it would begin to restrict credit. All interest rates (real and nominal) rose sharply in 1966 and the real money supply fell for four successive quarters; the LM curve was shifting leftward. The Federal Reserve was briefly successful in reducing the rate of increase in prices at the end of 1966, but then inflation rose sharply in 1967 and 1968 as large deficits resulted from the Vietnam War. One reason inflation rose was that the Federal Reserve was focusing on nominal rather than real short-term interest rates; the latter were negative on average in the last three quarters of 1967 and so monetary policy was actually expansionary. The IS curve was shifting rightward until a temporary ten per cent income tax surcharge was imposed in January 1968 on corporate income taxes and in April 1968 on personal income taxes. The federal budget deficit in the national income accounts turned into a surplus in the third quarter of 1968. Beginning in early 1968, the Federal Reserve began to raise real interest rates dramatically. With both the IS and LM curves shifting leftward, the economy began to slow and the unemployment rate began to rise in 1969, as the model predicts.

During the 1960s it was becoming less tenable to view the United States as a closed economy. Although the percentage of US exports to GNP had risen from six per cent in 1946 only to seven per cent in 1969, international events were beginning to impair the usefulness of the original Hicksian model. The quasi-fixed exchange rate system that had been established in the 1944 Bretton Woods Conference began to collapse in 1968 when the US gold stock reached a critically low level. In light of the aforementioned important contributions by Mundell and Fleming, which argued that monetary policy is ineffective and fiscal policy is powerful in a fixed-exchange regime with perfect capital markets, it is necessary to digress to explain how the model applied before the actual collapse in 1971–3 and afterwards.

At the end of the Second World War many countries restricted currency and capital flows, which allowed both fiscal and monetary policy to be effective, as in the world Hicks envisioned. These restrictions were gradually relaxed during the subsequent years. As they disappeared, the efficacy of monetary policy weakened, although imperfect capital markets allowed it to have some residual potency. Monetary policy was weakened because countries were obliged to maintain quasi-fixed exchange rates by not allowing real interest rates to vary across countries. In contrast, fiscal policy was strengthened because central banks were obligated to take actions that offset the effects of fiscal actions on real interest rates, essentially causing the LM curve to shift in the same direction that the IS curve shifted.

By 1973 the world was in a ‘dirty’ floating exchange rate system where various countries attempted to maintain some fixed bilateral exchange rates with their major trading partners. Mundell and Fleming had argued that in a pure floating exchange rate system with perfect capital mobility fiscal policy would be ineffective and monetary policy would be very strong, because any action to change a country's real interest rate relative to other countries would be reinforced by a change in its trade balance. In other words, a shift in its LM curve would be reinforced by a shift in its IS curve in the same direction because its trade balance was negatively related to its exchange rate, which was positively related to the value of its real interest rate relative to those of its trading partners. Because countries were unwilling to have their exchange rates be completely flexible, fiscal policy was considerably weakened relative to the fixed exchange rate period but still continued to have some power. Monetary policy was strengthened.

This open economy extension of the IS–LM model has proven to be illuminating about monetary and fiscal policy in the post-1971 period, again particularly in the United States. With the change in the exchange-rate regime, the trade-weighted value of the dollar fell about 20 per cent between 1971 and 1973, which, together with a recession in 1974, was sufficient to allow the United States to have a trade surplus on average through 1976. Between 1973 and 1979 the Federal Reserve allowed the real short-term (federal funds) rate to be negative on average, which led to substantial inflation and a bubble in the housing market. However, the international trade-weighted value of the dollar was essentially unchanged between the middle of 1973 and the middle of 1978, because average nominal short-term interest rates and inflation in major trading partners of the United States moved in tandem. Beginning in 1977 the US trade deficit began to increase and after mid-1978 the value of the US dollar fell unevenly until July 1980.

In July 1980 the Federal Reserve began to reduce the real money stock and, with accompanying large tax cuts in 1981–3, real and nominal interest rates rose to record levels, actions that were not offset by matching policies in foreign countries. As a result, the trade-weighted dollar appreciated from 84.65 (March 1973=100) in July 1980 to 158.43 in February 1985 and the trade deficit soared. The IS curve shifted to the right because the increase in the federal deficit was larger than the increase in the trade deficit during these years; the LM curve shifted to the left. As the IS–LM model predicts, the expansionary effects of the 1981–3 tax cuts, as measured by changes in real GDP and the unemployment rate, were much smaller than those of the similarly sized 1964–5 tax cuts, because of both the non-accommodating monetary policy and the dollar's appreciation.

In September 1985 a meeting of representatives of five major nations in New York resulted in a successful coordinated effort to reduce the trade-weighted value of the dollar, which fell 30 per cent in the succeeding two years and was followed by a sharp reduction in the US trade deficit. US short-term real interest rates fell until the middle of 1988 and the unemployment rate reached a low of 5.2 per cent in 1989. As the extended IS–LM model predicts, monetary policy was quite effective. Monetary policy effectiveness would be repeatedly evident in the following years. For example, the Federal Reserve raised real short-term interest rates between July 1988 and December 1990 to combat inflation, which resulted in a short recession in 1991. It successfully responded to rising unemployment by cutting its real overnight federal funds interest rate to near zero in 1993, which dramatically lowered the unemployment rate in 1995. By sharply raising this interest rate in 1994, the Federal Reserve managed to continue to lower the unemployment rate with negligible inflation until a stock-market bubble burst in 2001. Both fiscal and monetary policies were strongly expansionary between 2001 and 2005. The real federal funds rate had on average been negative since the end of 2001. As might have been predicted from the

Mundell–Fleming model, the trade deficit expanded enormously; its increase was roughly equal to the increase in the federal government deficit.

Finally, a troubling problem with prolonged periods of negative short-term interest rates is that they have tended to manifest themselves in rapid rates of inflation in prices of houses – both in the 1970s and in the 2000s. As Keynes and his followers warned, expectations are at the heart of the *General Theory* and they are not prominent in the textbook expositions of the IS–LM model. Many macroeconometric models can be interpreted as extensions of the textbook IS–LM model, and as time has passed many of them have increasingly attempted to incorporate expectations formation. Expectations are prominent in the Federal Reserve's recent model (Brayton et al., 1997), but their formation may not yet be accurately represented.

Conclusions

The IS–LM model occupies an awkward position in modern macroeconomics. It is still a workhorse of undergraduate teaching and still widely used by economists in developing intuition about short-run macroeconomic phenomena, including policy counterfactuals; see Colander (2004) for discussion of these roles. However, the model fails to accommodate the main features of modern macroeconomic theory, although modern dynamic models are sometimes interpreted as having IS–LM type features. We expect this dichotomy and this anomalous use to continue.

See Also

- monetarism

Bibliography

Blinder, A. and Solow, R. 1973. Does fiscal policy matter? *Journal of Public Economics* 2, 319–37.

Brayton, F., Mauskopf, E., Reifschneider, D., Tinsley, P. and Williams, J. 1997. The role of expectations in the FRB/US macroeconomic model. *Federal Reserve Bulletin* 83(4), 227–45.

Colander, D. 2004. The strange persistence of the IS–LM model. *History of Political Economy* 36 (Suppl. 1), 305–22.

Fleming, J. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9, 369–79.

Hansen, A. 1949. *Monetary Theory and Fiscal Policy*. New York: McGraw-Hill Publishing Co.

Harrod, R. 1937. Mr. Keynes and traditional theory. *Econometrica* 5, 74–86.

Hicks, J. 1937. Mr. Keynes and the ‘classics’: a suggested interpretation. *Econometrica* 5, 147–59.

Hicks, J. 1982. IS–LM: an explanation. In *Money, Interest, and Wages: Collected Essays on Economic Theory*, vol. 2. Oxford: Basil Blackwell.

Kahn, R. 1984. *The Making of Keynes's General Theory*. Cambridge: Cambridge University Press.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. New York: Macmillan.

Keynes, J.M. 1937. The general theory of employment. *Quarterly Journal of Economics* 51, 209–23.

King, R. 2000. The new IS–LM model: language, logic and limits. *Federal Reserve Bank of Richmond Economic Quarterly* 86(3), 45–103.

Lange, O. 1938. The rate of interest and the optimum propensity to consume. *Economica N.S.* 5, 12–32.

Lucas, R., Jr. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets. Carnegie Rochester Conference Series on Public Policy*, vol. 1, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

McCallum, B. and Nelson, E. 1999. An optimizing IS–LM specification for monetary policy and business cycle analysis. *Journal of Money, Credit & Banking* 31, 296–316.

Meade, J. 1937. A simplified model of Mr. Keynes’ system. *Review of Economic Studies* 4(2), 98–107.

Metzler, L. 1951. Wealth, saving, and the rate of interest. *Journal of Political Economy* 59, 93–116.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12, 45–88.

Mundell, R. 1963a. Inflation and real interest. *Journal of Political Economy* 73, 280–3.

Mundell, R. 1963b. Capital mobility and stabilization policy under fixed and flexible exchange rates. *Canadian Journal of Economics and Political Science* 29, 475–85.

Obstfeld, M. 2001. International macroeconomics: beyond the Mundell–Fleming model. *IMF Staff Papers* 47, 1–38.

Patinkin, D. 1956. *Money, Interest, and Prices*. Evanston, IL: Row Peterson.

Phelps Brown, E. 1937. Report of the Oxford meeting, September 25–29, 1936. *Econometrica* 5, 361–83.

Pigou, A. 1943. The classical stationary state. *Economic Journal* 53, 343–51.

Pigou, A. 1947. Economic progress in a stable environment. *Economica* 14, 180–8.

Poole, W. 1970. Optimal choice of optimal monetary instruments in a simple stochastic macro model. *Quarterly Journal of Economics* 84(2), 197–216.

Robinson, J. 1975. What has become of the Keynesian revolution? In *Essays on John Maynard Keynes*, ed. M. Keynes. Cambridge: Cambridge University Press.

Skidelsky, R. 1994. *John Maynard Keynes: The Economist as Savior 1920–1937*. New York: Viking Penguin.

Tobin, J. 1965. Money and economic growth. *Econometrica* 33, 671–84.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1(1), 15–29.

Tobin, J. and Braga de Macedo, J. 1980. The short-run macroeconomics of floating exchange rates: an exposition. In *Flexible Exchange Rates and the Balance of Payments: Essays in Memory of Egon Sohmen*, ed. J. Chipman and C. Kindleberger. Amsterdam: North-Holland.

Tobin, J. and Buiter, W. 1976. Long run effects of fiscal and monetary policy on aggregate demand. In *Monetarism*, ed. J. Stein. Amsterdam: North-Holland.

US Council of Economic Advisors. 1962. *Economic Report of the President together with the Annual Report of the Council of Economic Advisors*. Washington, DC: US Government Printing Office.

Young, W. 1987. *Interpreting Mr Keynes: The IS–LM Enigma*. Cambridge: Basil Blackwell.

How to cite this article

Durlauf, Steven N. and Donald D. Hester. "IS–LM." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000303> doi:10.1057/9780230226203.0855

The New Palgrave Dictionary of Economics Online

Isnard, Achylle Nicolas (1749 – 1803)

R.F. Hébert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Boisguilbert, Pierre le Pesant, Sieur de; capital theory; general equilibrium; Isnard, A. N.; mathematical economics; optimum resource allocation; Physiocracy; simultaneous equations; value; Vauban, S. le P.; Walras, L.

Article

French engineer and economist, Isnard was born at Paris on 25 February 1749; he died at Lyons on 25 February 1803. There are no details of his family history except that he had a devoted brother, J.L. Isnard, who was a lawyer and a judge, and who often interceded on his behalf. At the age of 17, Isnard entered the Ecole des Ponts et Chaussées which, even at this early date, inspired interest in political economy and exposed its students to heavy doses of mathematics and statistics. On successfully completing his studies, Isnard began his career as an apprentice engineer in the district of Besançon. While engaged in various works of construction in these environs, he took the time to write his remarkable two-volume work, *Traité des richesses*, which was published in 1781.

Isnard's *Traité* is a highly original work, despite the fact that its theoretic core is embedded in otherwise unexceptional arguments against Physiocratic doctrines. By this fact, we may infer that Isnard knew the Physiocratic literature, but we can only speculate on his acquaintance with other writers. Given his background and training, the authors he would have most likely known are Boisguilbert and Vauban (Boisguilbert's ideas were represented in 19th-century course outlines at the Ecole des Ponts et Chaussées, and Vauban's views on the professionalization of engineers were largely responsible for the establishment of the Ecole). Boisguilbert certainly had a vision of an interconnected economy and of a kind of general equilibrium, although he failed to render his conception concrete by erecting any kind of formal, theoretic structure of a mathematical nature.

Isnard, on the other hand, was the first writer to attempt a mathematical definition and a mathematical proof of an economic equilibrium. Furthermore, he gave specific form to the general equilibrium concept by constructing a set of simultaneous equations which, in general form and content, anticipated the major elements of the Walrasian system, including the general interdependence of markets and

quantities, the technical specifications of the exchange ratios, and the mathematical determination of the *numéraire*. It remained for Walras to add the engine of utility maximization and to adapt Isnard's model to his own purposes, something which, according to Jaffé (1969), he did with persistence, if not with ease. Isnard's pioneer efforts do not in any way denigrate Walras's monumental achievement, but they do lend force to the conviction that the development of economics was, and remains, a cumulative process. Isnard's *Traité* is now extremely rare. However, the mathematics of his equilibrium analysis of exchange are partially accessible in Robertson (1949), Baumol and Goldfeld (1968), Jaffé (1969) and Theocharis (1983). The significance of Isnard's performance is that he discovered early on the truth that value is not an intrinsic thing but rather is a magnitude which necessarily varies in relation to other goods, whose worth is also interdependent. Specifically, Isnard anticipated the two-good world of Walras in which, for example, the demand for eggs is the supply of wheat and the demand for wheat is the supply of eggs. This elaboration of commodity interdependencies in real terms consumes approximately the first half of Walras's *Eléments*. Mathematically, Isnard treated value as an exchange ratio, moreover, and he worked out the equilibrium process of exchange both with and without money.

It is noteworthy that Isnard extended the subjects under his analytical purview to include, besides the theory of exchange, the theories of production, capital, interest and foreign exchange. Jaffé (1969) has demonstrated that Walras's economic theory bears the imprint of Isnard in each of these areas. Underscoring merely the most striking example of the calibre of Isnard's analysis, Jaffé (1969, p. 40) emphasized his theory of capital and interest, which correctly laid down the rule for optimum resource allocation in the following terms:

> Capitals are distributed among different employments in agriculture, industry, and commerce in such a way that the ratios of their values to receipts from the sale of their products less the costs of upkeep, repair, and replacement – that is, the ratios of [invested] funds to [net] returns – are everywhere the same in all enterprises. This uniformity is achieved and equilibrium established because funds flow to and abound in places where the yield [*intérêt*] is highest and because like things have one and the same value. When things have a higher price in one place than in another, they rush there and equilibrium is re-established. Let F be the value of the funds employed in agriculture and F′ that of the funds employed in industry; let B be the payments for the value of the products of agriculture less the cost of upkeep, repairs and replacement and B′ the payments for the products of industry less the same costs, then the ratio of F to F′ must be equal to the ratio of B to B′ for the ratio of F to B to be equal to the ratio of F′ to B′ or for the rate of interest [in the sense of rate of capitalization] to be everywhere the same. This uniformity [in the rate of capitalization] is realized not only between agriculture and industry in general, but also among individual enterprises.

It is, of course, necessary that perfect knowledge obtain for this conclusion to hold, but even without always making his assumptions explicit, Isnard anticipated much of modern microeconomic theory. The scope and sweep of his analysis unquestionably entitle Isnard to a position of prominence in the history of economic thought. Yet appropriate recognition took a long time. Despite the filiation of ideas between Isnard and Walras, the ‘father’ of general equilibrium analysis mentioned Isnard's name in only

one place, and that an obscure bibliographic article (a French reprint of Jevons's famous bibliography of mathematico-economic works) published in the *Journal des Economistes* in 1878. Add to this the ambiguity, idiosyncrasy and prolixity of Isnard's treatise. His definition and use of mathematical symbols is inconsistent and the essence of his arguments difficult to extract, nested as they are in a morass of other material that is neither very original nor very interesting. Such deficiencies were bound to handicap the recognition and acceptance of Isnard's contribution. In the final analysis, however, Isnard was simply a brilliant pioneer who wrote ahead of his time, and like so many other semi-tragic heroes of economic analysis (for example, Cournot and Gossen), he failed to receive his due until long after departing the scene.

Isnard suffered in his personal life even as his ideas suffered (by neglect) in economics. Hot-tempered, yet not given to the intrigues apparently required to advance in the engineering ranks of a quasi-military public service, Isnard spent most of his career in a subordinate capacity. After he finally received a post worthy of his talents, his wife died, leaving him to raise three motherless children. At that point Isnard left government service and struggled in penury for some time. Recalled by Napoleon for the Egyptian campaign in 1798, he was inexplicably left behind. Adding insult to injury, he was forced to take an oath of allegiance to the Republic even though he was an avowed royalist. He later became a member of the Tribunate under Napoleon and took an active part in the formation of public finance and conscription policies. But upon completing his term he resumed his engineer's career at Lyons, where he died soon after, 54 years to the day from his birth. Given his apparent influence on Walras, Theocharis (1983, p. 62) probably did not exaggerate much when he labelled Isnard's *Traité* ‘one of the most important contributions in the history of the development of mathematical economics’.

Selected works

1781. *Traité des richesses*, 2 vols. London and Lausanne: F. Grasset.

1801. *Considérations théoriques sur les caisses d'amortissement de dette publique*. Paris: Duprat.

Bibliography

Baumol, W.J. and Goldfeld, S.M., eds. 1968. *Precursors in Mathematical Economics: An Anthology*. London: London School of Economics and Political Science.

Jaffé, W. 1969. A.N. Isnard, progenitor of the Walrasian general equilibrium model. *History of Political Economy* 1, 19–43.

Robertson, R.M. 1949. Mathematical economics before Cournot. *Journal of Political Economy* 57, 523–36.

Theocharis, R.D. 1983. *Early Developments in Mathematical Economics*, 2nd edn. London: Macmillan.

How to cite this article

Hébert, R. F. "Isnard, Achylle Nicolas (1749–1803)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000227> doi:10.1057/9780230226203.0857

The New Palgrave Dictionary of Economics Online

Italy, economics in

Massimo M. Augello and Marco E.L. Guidi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The history of economics in Italy reflects an interaction between scientific–educational institutions and political power, which led economists to combine a theoretical approach and political commitment. During the Enlightenment a network of circles and public academies spawned the contributions of Beccaria, Genovesi, Galiani, Ortes and Verri. The institutionalization of economics in the 19th century prepared the success of the marginalist generation led by Pantaleoni, Pareto and Barone. In the interwar period, academic economists formed a bulwark against Fascism. The post-war political climate favoured the internationalization of economics, with the importation of Keynesianism and, later, other currents of thought.

Keywords

Barone, E.; Beccaria, C.; Boccardo, G.; Bodio, L.; Boselli, P.; Botero, G.; Bresciani Turroni, C.; Caffè, F.; Carey, H. C.; Cognetti de Martiis, S.; Corbino, E.; corporatism; Cossa, L.; cost of reproduction theory of value; Croce, B.; Custodi, P.; Davanzati, B.; De Viti De Marco, A.; Del Vecchio, G.; Demaria, G.; di Fenizio, F.; division of labour; Einaudi, L.; Taviani, P.E.; Engel, E.; entrepreneurship; Equilibrium; Fanfani, A.; Fanno, M.; Ferrara, F.; Fuà, G.; Galiani, F.; Galilean revolution; Garegnani, P.; Genovesi, A.; German Historical School; Gioja, M.; Graziadei, A.; Graziani, A.; Intieri, B.; Italy, economics in; Jannaccone, P.; Keynesianism; labour theory of value; Labriola, A.; List, F.; Loria, A.; Luzzatti, L.; Machiavelli, N.; marginalism; Marrama, V.; Mazzola, U.; Messedaglia, A.; *Methodenstreit* ; Modigliani, F.; Montanari, G.; Morpurgo, E.; Nazzani, E.; neoclassical synthesis; neo-Ricardian economics; Nitti, F. S.; Ortes, G.; Pantaleoni, M.; Pareto, V.; Pasinetti, L.; Pesenti, A.; Physiocracy; Platonism; Protonotari, F.; public finance; Quetelet, A.; Rae, J.; revisionism; Ricci, U.; Romagnosi, G.D.; Saraceno, P.; Say, J.-B.; Scaruffi, G.; Scialoja, A.; Serra, A.; sharecropping; Sismondi, J. C. L. S. de; Smith, A.; Società Italiana degli Economisti; Sraffa, P.; Steve, S.; subjectivist theory of value; Supino, C.; Sylos Labini, P.; Toniolo, G.; Turati, F.; Vanoni, E.; Vasco, G.; Verri, P.

Article

This article examines the evolution of Italian economic thought from its origins to the post-Second World War years, when the professionalization and internationalization of economics reached maturity, sowing the seeds of the present vigorous state of economic studies. It offers an institutional history of political economy, distinguishing four epochs. The first epoch runs from the 16th to the 18th century. In this period, the alliances between enlightened sovereigns and groups of intellectuals produced a wealth of original contributions to economics. The second epoch corresponds to the Napoleonic age and the Restoration. This was a period in which, despite political repression in Italy, Smithian political economy penetrated many circles and was debated in journals and academies. The third epoch runs from the unification of Italy in 1860 to the rise of the Fascist regime in 1922, and is referred to as the ‘liberal age’. This period was crucial for the institutionalization of economics and for the prominent public role that was attributed to economists. Such a favourable environment was responsible for the high standard of scientific debate, which culminated in the generation of Pantaleoni and Pareto, when Italian economics, as Schumpeter wrote (1954, p. 855), ‘was second to none’. Finally, the fourth epoch regroups the Fascist era and the post-war years. During the decades of Fascism, the regime's attempts to control economic debate produced a reaction of self-defence and isolation among neoclassical economists. After the war, the new climate of liberty encouraged economists to get back to their public role. The political debate on economic planning was responsible for the acceptance of Keynesianism in the 1950s. The success of Neo-Ricardianism in the 1960s and, later, of American-based mainstream economics marked the internationalization of Italian economics.

Public happiness and geometrical method: from the origins to the Enlightenment

Although in the Middle Ages theological debate over the ‘just price’ and the legitimacy of usury flourished in many parts of Italy, and Italian authors came to be respected throughout Europe, the beginnings of modern economic science in this country date from the 16th and early 17th centuries, when the formation of regional states generated a need to regulate public finances, trade and the circulation of money. Some traces of economic analysis can be found in the treatises on politics published by the humanist Niccolò Machiavelli (1469–1527) (*Discorsi sopra la prima deca di Tito Livio*, 1513–21) and the Jesuit Giovanni Botero (1544–1617) (*Della Ragion di Stato*, 1589). But the most original analyses were contained in some short treatises of a systematic and highly formalized character, based on a hedonistic framework, a coherent theory of value and an extensive use of mathematics. The theoretical framework they provided could be employed to interpret the monetary and commercial problems of the time. Of this kind are the *Discorso sulle monete* (1582) by the aristocrat Gaspare Scaruffi, the *Lezione delle monete* (1588) by the merchant and historian Bernardo Davanzati (1529–1606), the *Breve trattato delle cause che possono far abbondare li regni d'oro e d'argento, dove non sono miniere* (1613) by Antonio Serra, whose life is shrouded in mystery, and the *Trattato mercantile della moneta* (1683) by Geminiano Montanari (1633–87), a professor of mathematics and astronomy. These works were connected to the diffusion of the Catholic currents inspired by Platonism and hostile to Aristotelian scholasticism that were at the root of the Galilean revolution. The abstract nature of these texts can be explained by the fact that they were a product of the scientific academies

created in this period under the aegis of the Italian princes – especially those in Florence – with a view to encouraging knowledge which could be useful in strengthening state power and countering the civil decay of the country, for which they blamed the political influence of the Church. The authors of these economic treatises were natural philosophers who acted as temporary consultants to government or were members of the state bureaucracy.

The long wave of Galilean and Platonic doctrines continued through the 18th century and had a recognizable impact on the theoretical structure of the economic discourse in the two circles that were at the centre of the Italian Enlightenment, combining in different ways with new ideas coming from France and Scotland and with ideas from other indigenous traditions. The first of these circles was the Accademia dei Pugni of Milan, which was run by Pietro Verri (1728–97), author of the *Meditazioni sull'economia politica* (1771) and Cesare Beccaria (1738–94), known not only for his main work, *Dei delitti e delle pene* (1764), but also for a series of theoretical articles on political economy published in the journal of the academy, *Il Caffè* (1764–6). The second group was that in Naples headed by Bartolomeo Intieri, whose discussions sparked both *Della moneta* (1751) by Ferdinando Galiani (1728–87) and *Lezioni di economia civile* (1766–7) by Antonio Genovesi (1713–69). These groups became involved in the economic reforms their monarchs were trying to bring about, as revealed by the public offices obtained by Verri, Beccaria and others. Also in other parts of the country, the academies were called upon to produce a science whose utility was measured in terms of greater dominion over nature and of the greater ability of governments to increase ‘public happiness’. But it was the 1753 foundation of the Accademia Economico-Agraria dei Georgofili in Florence that sanctioned the scientific status of economics and its political role in the strategy of reform. Right up to the end of the 18th century, this academy was to be one of the main vehicles for the spread of Physiocratic and Smithian doctrines in Italy. Its journal was an example of the many agricultural periodicals that hosted economic debates, often of a practical kind, albeit open to the new science of political economy.

Collaboration between philosophers and princes ushered in the creation of the first teaching of political economy as part of a reform of university studies whose purpose was to bring them under the umbrella of the state, combating the control that religious orders and professional bodies had traditionally exerted over them. A chair of Commercio e meccanica was founded in Naples in 1754 on the initiative (and funding) of Intieri, and conferred on Genovesi. Another professorship of political economy, established at the Scuole Palatine, Milan, in 1768, was assigned to Cesare Beccaria. Similar chairs were created in Modena, Catania and Palermo. The aim of these chairs was twofold: they should instruct government bureaucrats in the art of governing economic and financial affairs, and stimulate the application of new agricultural techniques and agrarian laws.

From a theoretical point of view, the contribution of these authors – to whom one should add at least Giammaria Ortes (1713–90) and Gianbattista Vasco (1733–96) – is quite homogeneous. Their core approach is based on a natural law framework which turns around a static rather than dynamic notion of equilibrium, and suggests that there are forces in society that tend to restore equilibrium when natural disasters, changes in tastes, or political errors create unbalances. Another basic assumption is the sensationalist view that human beings are constantly under the guidance of pleasure and pain. The underlying methodology is still abstract and ‘geometrical’, as in the works of their predecessors. The focus of analysis is on problems of exchange rather than of production (although Beccaria gave the clearest definition of the division of labour before Adam Smith). Their main contributions concern the analysis of value, based on utility and scarcity: Ortes, Beccaria and Verri attempted a mathematical

formulation of the law of demand and supply as a guide to the analysis of price adjustments. Another theme of inquiry is the theory of money, where Galiani, adopting metallist assumptions, expounded a clear distinction between short-run variations in the value of money and long-term effects, and between real and monetary effects. Finally, these authors shared a view of the sovereign as a reformer and supreme moral authority, who takes into account the feelings and needs of individuals and constructs a social order according to the dictates of reason and natural law. This order consists of an equilibrium of interests that generates ‘public happiness’ (*felicità pubblica*).

The spread of classical economics in the age of Risorgimento (1815 – 60)

Although the Napoleonic age is considered, in Europe as a whole, a period in which political economy was regarded with suspicion, in Italy the establishment of the Empire's satellite kingdoms favoured the discipline's development and its institutionalization. However, the content of teaching was radically modified: theoretical economics was reduced in order to make room for legal and statistical notions, which were considered more urgent for the training of public officials. Moreover, Napoleonic administrations concentrated on the collection of detailed statistical information about the condition of their *départements*, in order not only to promote anti-feudal reforms but also to protect French interests. A new generation of government officials was assigned to this task: among them there was the most important economist of this period, Melchiorre Gioja (1767–1829), who published his main work, the *Nuovo prospetto delle scienze economiche*, between 1815 and 1817. Gioja examined Smith's and Say's theories with a critical eye, and his original analysis of cooperation, division of labour and machinery was acknowledged by Charles Babbage as an anticipation of his own theories. Regarding economic policy, Gioja was favourable towards state intervention in order to foster the development of agriculture and manufactures.

Another Napoleonic official was Pietro Custodi (1771–1842), who from 1803 to 1805 edited the 50 volumes of a collection titled *Scrittori classici italiani d'economia politica*, which reproduced most of the Italian texts on political economy from previous centuries. Custodi aimed at stimulating the patriotic spirit of his fellow citizens by encouraging them to improve their economic and statistical knowledge. This collection produced in the next generation of intellectuals of the *Risorgimento* era an exaggerated feeling of national pride, which nevertheless encouraged the study of economics.

Economics experienced its worst period after the Restoration in 1815. The reactionary governments of the Italian regional states considered the teaching of economics to be a vehicle for liberal and democratic ideas. As a consequence, all chairs of political economy were suppressed, except in Naples and Sicily, where they were put under strict political control. Only in the 1840s, in Piedmont and Tuscany, with the establishment of constitutional governments, was the teaching of political economy restored. Antonio Scialoja (1817–77), first, and then Francesco Ferrara (1810–1900) were appointed professors at the University of Turin, while other chairs of economics were created in Pisa and Siena.

In these conditions, discourse on political economy went on largely outside universities. This does not mean that it was clandestine, since it was developed in academies and associations which enjoyed an official status. But the political control over these institutions implied that public debate on controversial issues was sometimes tolerated and sometimes heavily repressed. Already in the Napoleonic age newly founded institutions, such as the Accademia Pontaniana of Naples or the Istituto Nazionale, established in Bologna and transferred to Milan in 1810, had included departments of moral and political sciences

where political economy was discussed. Furthermore, the experiences of 18th-century agrarian academies had prompted the establishment of a network of provincial associations termed ‘agrarian’ or ‘economic societies’, which aimed at promoting the development of local economies. These associations continued their activities even during the decades following the Restoration, expanding from Piedmont to Sicily. Despite their eminently practical goals, economic societies gave an important impetus towards the spread of British and French political economy and laissez-faire ideals.

Another means by which political economy spread through the Italian learned classes was the periodical press, despite the existence of censorship. In the early decades of the 19th century the heading ‘political economy’ appeared on a growing number of articles published in new journals of ‘sciences, letters and arts’, such as the *Biblioteca italiana*, founded in Milan in 1816, the *Antologia*, created in Florence in 1821, and *Il progresso delle scienze, delle lettere e delle arti*, first published in Naples in 1832, where it acted as the main point of convergence of liberal culture. Lively exchange of ideas was also found in journals of agriculture, especially the *Giornale agrario toscano*, founded in 1827, which together with the Accademia dei Georgofili promoted an original debate on sharecropping echoing Sismondi's remarks in *Tableau de l'agriculture toscane*.

The first signs of a trend towards specialization in economic disciplines came with the birth of several journals mainly devoted to statistical and economic themes, such as the *Annali universali di statistica* and the *Giornale di statistica*. The former was first published in Milan in 1822 and had among its contributors Gioja and Giandomenico Romagnosi (1761–1835). The latter was founded in Palermo in 1836 as the organ of the Central Statistical Office. Edited by Ferrara, it achieved immediate recognition as the premier forum for debate among Sicilian laissez-faire economists. Another interesting experiment was *Il Politecnico*, launched by Carlo Cattaneo (1801–69) in 1839. The majority of the essays were composed by Cattaneo himself, and dealt with various practical issues. However, in two remarkable articles Cattaneo focused on doctrinal questions, criticizing the protectionist theories of Friedrich List, and arguing that knowledge and motivation are the most important factors of economic development. But one should also mention a number of journals created in Naples in the 1840s, which arose against the backdrop of private law schools, established as an alternative to the more conservative form of instruction offered by the universities. These journals and institutions soon became a focal point for the new school of liberal economists, of which Scialoja was the main representative.

As this description makes clear, the political economy debated in these forums was that of Smith and Say. In northern Italy a key figure was Romagnosi, a legal philosopher who – taking inspiration from Giambattista Vico's philosophy of history – formulated a peculiar version of Smithian political economy, in which the notion of ‘natural progress of opulence’ was employed to argue that economic development depended on a framework of formal and informal institutions (so-called *incivilimento*), and that government-induced industrialization would result in social disaster. Romagnosi's ‘institutionalist’ approach influenced a whole generation of economists, stimulating interesting contributions on the relationships between law and economics. In the south of Italy, the penetration of classical economics was mediated by the influence of the French *idéologues*, which caused Say's work to be received enthusiastically. The most brilliant product of this environment was Scialoja's *I principj della economia sociale esposti in ordine ideologico* (1840), translated into French in 1844, which adopted Say's subjectivist approach to value and developed the analysis of the entrepreneur in a pre-Schumpeterian sense.

But the most acute and original economist of this age was Ferrara, who in his lecture notes of 1856–8

and in the prefaces to the *Biblioteca dell'economista* – a ‘library’ containing the Italian translation of a vast number of foreign works on economics, of which he was the editor – proposed a generalization of the cost of reproduction theory of value formulated by Henry Carey and John Rae. Ferrara's version of this theory took into consideration three different cases: that of ‘physical reproduction’ by direct labour, that of physical reproduction ‘by way of exchange’ and that of ‘economic’ reproduction by substitutes. In this way, the theory highlighted the fact that value is grounded on utility and subjective opportunity costs, clearly foreshadowing marginalist analysis.

The institutionalization of economics in the liberal age (1860 – 1922)

The epoch that followed the unification of the country in 1860 was decisive for the consolidation of economic studies. Chairs of political economy were introduced in the more than 20 law faculties that existed at the time. In 1876, new university regulations added the teaching of statistics and public finance. The latter was established as a compulsory course in 1885. Likewise in the 1880s, two Higher Schools of Commerce were created in Genoa and Bari, similar to the first institution of this kind, which had been founded in Venice in 1868. This expansion multiplied the opportunities for economists to obtain university positions, and well before the end of the 19th century the social identity of the economist could be largely identified with the academic profession. But a decisive stimulus to the professionalization of economics was provided in the mid-1870s by the explosion of the Italian counterpart of the *Methodenstreit*, the dispute over methods that divided German-speaking economics. All the major economists became involved in it, and opposition between different economic and political conceptions had an important impact on the professional and academic level. These divisions induced economists to devote greater attention not only to the scientific aspects of the profession (training and specializations) but also to academic policy (increase in the number of academic chairs and control over recruitment procedures). This process resulted in a generational change within the ranks of academic staff, leading to a preponderance of the followers of *Kathedersozialismus* or ‘socialism of the chair’ in the German mould.

Another important element is represented by the increasing public role played by economists. The extension of civil liberties, coupled with the institution of a national representative system, gave them an extraordinary opportunity to spread economic knowledge and influence policymaking. Many economists became columnists for newspapers and weekly magazines, while others were active in the foundation of associations of interests, chambers of commerce, saving banks or cooperatives. Lastly, virtually all the leading economists of this age – more than 30 – became members of parliament. And although some of them were involved in parliamentary activities that bore little relation even to the broadest view of the scope of political economy, in the central debates on tariffs and trade, fiscal policies, credit, education, and in inquiries on the condition of agriculture and industry the voice of economists became a typical feature of public life. Some economists were also appointed ministers, while three of them – Paolo Boselli (1838–1932), Luigi Luzzatti (1841–1927), and Francesco Saverio Nitti (1868–1953) – became prime ministers. On the whole, these activities strengthened the scientific and social identity of economists.

The growing professionalization of economists was also reflected in the creation of new societies in which they played a central role. After the first experiments in Turin in the 1850s and early 1860s under the guidance of Ferrara, the Società di Economia Politica Italiana was established in 1868 on the

initiative of the economist Francesco Protonotari, who was the editor of the most important scientific and literary journal of the time, the *Nuova Antologia*. These associations attracted the great majority of academic economists and many representatives of the political elites. The constitution of the Società di Economia Politica significantly opened with the statement that the Society's mission was to ‘promote and disseminate economic studies’. However, very soon its activities were dominated by more practical discussions on parliamentary debates and government economic policy. Conflicts concerning the new orientation of the society's purposes led to a gradual slowdown in the pace of activities.

It was against this backcloth that the *Methodenstreit* arose, breeding the projects of two rival associations. The former, dubbed Società Adamo Smith, was set up in Florence in 1874 on the initiative of Ferrara and several laissez-faire economists and politicians belonging to the group of the so-called ‘Tuscan moderates’. The aim of the Society was that of ‘promoting, developing and defending the doctrine of economic liberties’, and of assuming the character of a scientific body, excluding from debate all that could be more properly described as political. The latter society, called Associazione per il Progresso degli Studi Economici, was created in January 1875 by a group headed by Luigi Cossa (1831–96), a powerful academic of the university of Pavia, Fedele Lampertico (1833–1906), an influential senator of the Venetian area, and by Luzzatti and Scialoja. Responding to Ferrara's splinter-group tendency, these economists had drawn up a document, known as the ‘Padua circular’, which marked the start of the counteroffensive by ‘socialists of the chair’. The society set itself the task of promoting social studies, to be accomplished partly through extension of its organizational structure to different parts of Italy.

Both associations proved to be short-lived, but the overall effect of their activities over roughly a 30-year period was that of ushering in a profound change in the institutional set-up of economic studies, reinforcing the academic and public background of the economists’ activities. This new condition was reflected in the world of publishing. To begin with, while scientific–literary journals continued in their tradition of hosting writings on economic themes, more specialized journals were established. Characteristically, in the 1870s almost all of the economic journals exhibited very close links with one or the other of the conflicting schools of economic thought. Thus, orthodox liberals used as their mouthpiece the journal *L'Economista*, created in 1874, whereas ‘socialists of the chair’ founded one year later the *Giornale degli economisti*. Rather than a genuine forum for scientific debate, however, such journals tended to become tools with which to enter the political fray. This characteristic was to a lesser extent replicated by the new journals that appeared in the 1890s, despite their more scientific and academic nature: the most important among them were the socialist periodical *Critica sociale*, directed by Filippo Turati (1857–1932), the *Rivista internazionale di scienze sociali e discipline ausiliarie*, edited by the Catholic economist Giuseppe Toniolo (1845–1918), and *Riforma sociale* (1894–1935), edited at first by Nitti, and later by Luigi Einaudi (1874–1961). Perhaps the first journal of economics in the modern sense may be considered the new series of the *Giornale degli economisti*, started in 1890 and managed by Maffeo Pantaleoni (1857–1924), Antonio De Viti de Marco (1858–1943) and Ugo Mazzola (1863–99). This journal voiced the radical laissez-faire approach of its editors, while becoming the forum of academic research and the main vehicle of penetration of marginalist theory in Italy.

Even outside the world of journals, greater attention began to be paid to the promotion of economic studies. For instance, after the first two series of the *Biblioteca dell'economista*, edited by Ferrara and published in the 1850s and 1860s, a third series was entrusted to Gerolamo Boccardo (1829–1904) in the 1870s, and a fourth and fifth series were continued by Salvatore Cognetti de Martiis (1844–1901) and

Pasquale Jannaccone (1872–1959) up to 1922. With its 71 volumes containing more than 150 classics of economics, the *Biblioteca* was extremely successful and became a unique tool for those who wanted to update their knowledge in this field. Economists also popularized their doctrines through dictionaries and encyclopaedias; the most important of them was the *Dizionario della economia politica e del commercio*, published in 1857–61 by Boccardo, who was also the editor of the *Nuova enciclopedia italiana* (1875–88).

Undoubtedly the main instrument for the spread and institutionalization of political economy was the large number of treatises, manuals and popularizations that were published during this period. The authors of these texts were both major economists and a host of lesser-known scholars, philanthropists and schoolteachers interested in the popularization of political economy. The number of works published – almost 300 from 1840 to 1920 – reveals that there was a pervasive ‘need’ for political economy, considered as a discipline that could educate the younger generations of administrators and politicians, instruct public opinion and enlighten the working classes. To judge from the number of editions, the most popular manuals were Cossa's *Primi elementi di economia politica* (1875, 17 re-editions), Emilio Nazzani's *Sunto di economia politica* (1873, 16), Boccardo's *Trattato teorico-pratico di economia politica* (1853, nine), Camillo Supino's *Principii di economia politica* (1904, nine), Achille Loria's *Corso completo di economia politica* (1909, seven), and Augusto Graziani's *Istituzioni di economia politica* (1904, six).

At the same time, the scientific quality of the work was high. Italian economists rapidly assimilated international economic debates, and in some cases they became important protagonists. The quantitative approach to statistics initiated by A. Quetelet and E. Engel was largely accepted in the mid-1860s thanks to the contributions of Angelo Messedaglia (1820–1901), whose methodological works influenced a whole generation of economists, Emilio Morpurgo (1836–1885), and Luigi Bodio (1840–1920), who organized the Central Statistical Office and was elected secretary of the International Institute of Statistics on its foundation in 1885. Some years later, Cossa, Lampertico and Luzzatti were instrumental in familiarizing Italian scholars with the methodology of the German Historical School and the social views of socialism of the chair. These economists promoted a renewal of economic studies along inductivist and quantitative lines, and adopted a critical stance vis-à-vis economic liberalism in matters of social policy. They were called the ‘Lombard–Venetian School’ since most of them taught at the universities of Pavia and Padua.

An even more vigorous and original response to outside stimuli was represented by the penetration of marginalism. Pantaleoni's *Principii di economia pura* – a work largely inspired by Jevons, Edgeworth and Marshall – dated from 1889, but the same author had already published in 1883 a work on public finance based on marginalist notions. Pantaleoni encouraged Vilfredo Pareto's (1848–1923) conversion to the new approach some years later. In 1893, the latter succeeded Walras at the chair of economics in Lausanne. In his *Cours d’économie politique* (1896–7), and more radically in *Manuale di economia politica* (1906), he revolutionized utility theory, laying the foundations of modern microeconomic analysis. A third representative of Italian marginalism was Enrico Barone (1859–1924), whose article on ‘The Ministry of Production in the Collectivist State’ (1908) was included by Hayek in his 1935 anthology on economic planning. Interesting applications to public finance were also provided by Barone himself and by De Viti and Mazzola. Their contributions lay the foundation of an original school of thought whose analysis of taxes, public expenditure, and of the political context in which fiscal structures operate, has been recognized by James Buchanan as the starting point of the development of

modern public finance theory. A distinctive feature of Italian marginalist economists was their practical and ideological commitment: they engaged themselves in political and editorial activities, staunchly defending a radical laissez-faire view. The socialists Arturo Labriola (1873–1959) and Enrico Leone (1875–1940) attempted to find a compromise between marginalism and Marxism. In the first decade of the 20th century, neoclassical economics had already become the orthodox approach.

Less vigorous, albeit no less original, was the Italian contribution to Marxist revisionism. In *La rendita fondiaria e la sua elisione naturale* (1880), Loria (1857–43) attempted to explain the functioning of a capitalist economy as a result of the structure and evolution of landed property. The historical and theoretical weaknesses of Loria's approach were then attacked at the end of the century by the Marxist philosopher Antonio Labriola (1843–1904), but his efforts to convince Benedetto Croce (1866–1952) to join his camp resulted in a relaunching of revisionism: Croce considered Marx's notion of surplus value as a simple ‘mental abstraction’ which could not explain the essence of capitalist production. On the other hand, Antonio Graziadei (1873–1953) argued that the labour theory of value was useless to explain the genesis of surplus value and the formation of market prices.

From corporatism to Keynesianism and neo-Ricardianism

After an early phase of authoritarian laissez-faire policy delegated by Mussolini to the economist and minister of finance Alberto De’ Stefani (1879–1969), a turn towards a corporatist organization of the economy was accomplished in 1926. The introduction of corporatism was the result of political decisions rather than of scientific debate, although corporatist currents of Catholic and socialist ascendancy had existed since the late 19th century. The fascist regime organized two national conferences in 1930 and 1932 to stimulate a debate on corporatist economics, but they ended up with the defeat of those intellectuals who stood for a more radical transformation of economic relationships along corporatist lines.

On the whole, only from 1925 to 1934 did corporatist economics enjoy some popularity. Its partisans proclaimed that the *homo corporativus* should replace the individualist *homo oeconomicus*, but they failed to produce significant achievements in economic theory. Orthodox economists like Einaudi and Jannaccone, initially forced into a tactical retreat, took back the lead in debate after 1934. Most academic economists – Gustavo Del Vecchio (1883–1972), Marco Fanno (1878–1965), Costantino Bresciani Turroni (1882–1963), Giovanni Demaria (1899–1998), and others – put aside their laissez-faire beliefs and attempted to interpret corporatist economy from a marginalist viewpoint. Corporatism was thus reduced to a case of economic policy, which did not modify the content of pure theory. A characteristic that distinguished these and other economists was their firm attachment to Paretian general equilibrium analysis, which they developed in a dynamic sense elaborating some suggestions derived from Pantaleoni's writings and from Pareto's sociology. At the same time, forced to defend orthodoxy against ideological attacks, these economists largely ignored or misunderstood the nature of the Keynesian revolution.

This success of orthodoxy can be mostly explained by institutional factors. The Fascist government tried to reform the organization of university studies, in 1935 transforming the teaching of economics into that of ‘corporatist political economy’. However, orthodox economists jealously defended their academic autonomy, and the younger generation they recruited was composed of disciples whose career was generally not obstructed by political intrusions. The efforts of the Fascist regime concentrated on

the creation of special schools and research institutions – such as the School of Corporatist Sciences of the University of Pisa, directed by Giuseppe Bottai (1896–1979), the Labour School of Florence, headed by Gino Arias (1879–1942), and the National Institute of Agrarian Economics, directed by Arrigo Serpieri (1877–1960).

Likewise, the major publishing houses – in particular Einaudi in Turin and Laterza in Bari – actively supported orthodox economics. The publisher that more actively sponsored corporatist economics was Sansoni in Milan, which issued a series connected to the Pisa Corporatist School, containing works on corporatism and economic planning. The lobbying activity of liberal economists also succeeded in modifying the editorial project of the *Nuova collana di economisti stranieri ed italiani* (1932–7), originally conceived as a sequel of the *Biblioteca dell'economista* and as the seal of Fascist economic culture. As a matter of fact this collection was open to recent international literature (Pigou, Sraffa, Hicks, Frisch, Hayek, Robertson and Keynes), and made no room for corporatist economics. Even the major cultural enterprise of the Fascist regime, the *Enciclopedia Italiana* edited by Giovanni Gentile, was quite impartial in the choice of authors for its economic entries.

Conversely, Mussolini's government was able to impose a considerable control over the periodical press. On the one hand, it created its own ideological mouthpieces – such as *Gerarchia* and *Critica Fascista* – and favoured the rise of economic journals – such as *Economia*, founded in 1923, and *Nuovi studi di diritto, economia e politica*, started in 1927 – that stimulated a considerable debate around the implications of corporatist economics. On the other hand, it extended its repression of journals of the liberal camp. Both *La Riforma sociale* and the *Giornale degli economisti* were discontinued for political reasons, in 1935 and 1942 respectively.

One of the costs of the Fascist years was a limited but significant ‘brain drain’: among those who were forced to emigrate were Bresciani Turroni, Umberto Ricci (1879–1946), Piero Sraffa (1898–1983), and the young Franco Modigliani (1918–2003).

The evolution of the economics profession in the post-war period was substantially influenced by the restoration of liberal-democratic institutions. First and foremost, the recovered political freedom favoured the rise of a network of centres of research and advanced studies (the Centre of Specialisation and Economic–Agrarian Research of Portici, the Svimez in Naples, the Istao in Ancona, the Research Department of the Bank of Italy in Rome) and of university departments. Scholarships were granted to young scholars who wanted to continue their studies abroad, encouraging the opening of frontiers to international debate after the relative isolation of the Fascist period. The main economic journals were restructured and new specialized periodicals emerged, adopting international standards. Another crucial event was the creation in 1951 of the Società Italiana degli Economisti, whose constitution stipulated a full economics professorship as a criterion for admission.

The new political context soon stimulated many economists to return to their traditional public vocation. Einaudi, Labriola and Nitti sat in the Constitutional Assembly (1946) with other economists of the younger generation, including Epicarmo Corbino (1890–1984), Amintore Fanfani (1908–1999), Antonio Pesenti (1910–1973), Paolo Emilio Taviani (1912–2001) and Ezio Vanoni (1903–1956). A special Commission on economic and social affairs nominated by the government was chaired by Demaria and composed of the most eminent amongst his colleagues.

It was mostly from the political side that Keynesianism made its entry into the Italian debate in the early 1950s, despite the persisting reluctance of economists to accept its theoretical underpinnings. Some Catholic economists engaged in politics, such as Fanfani and Giorgio La Pira (1904–77) declared

themselves to have been inspired by Keynes when, as members of the cabinet, they introduced a plan for subsidized housing to reduce unemployment. But a Keynesian flavour could also be discerned in the ‘Plan for labour’ propounded in 1948 by the CGIL, the communist and socialist trade union. Likewise, the ‘Scheme for the growth of employment and income in Italy in the decade 1955–1964’ presented by Vanoni was explicitly inspired by Harrod's growth model. Finally, the debate of the 1960s on economic planning, in which Ferdinando di Fenizio (1906–74), Pasquale Saraceno (1903–1991), Giorgio Fuà (1919–2000), Paolo Sylos Labini (1920–2005) and Federico Caffè (1914–1987) participated, was clearly dominated by Keynesian assumptions. This does not mean that Keynes's theory was not present in more academic debates. The second edition of di Fenizio's *Lezioni di teoria economica* (1948) reflected the neoclassical synthesis arguing that the Keynesian approach was complementary rather than alternative to classical theory. Also Caffè and Vittorio Marrama (1914–82) published a series of theoretical contributions on Keynesian economic policies. Finally, Keynesianism exerted a considerable influence on the Italian public finance tradition, especially thanks to the works of Sergio Steve (1915–2006).

The 1960s were also marked by the impact on Italian economics of works of Sraffa, who since the 1920s had migrated to Cambridge. His article on ‘The Laws of Return under Competitive Conditions’ (1926) – preceded by a paper published in 1925 in the *Giornale degli economisti* – had criticized Marshall's equilibrium analysis and paved the way for research in imperfect competition. In *Production of Commodities by Means of Commodities* (1960) Sraffa expounded an alternative to general equilibrium analysis based on a reformulation of the classical and Marxian notion of surplus. This work originated a school of thought, the neo-Ricardians, that exerted a powerful influence on international scientific debates and on Italian academic life for a couple of decades. Among its main representatives one should count Pierangelo Garegnani and Luigi Pasinetti.

The strength of neo-Ricardian economics has probably been the last distinctive feature of Italian economics, at least in its most theoretical departments. In recent decades, the growing internationalization of this discipline has caused Italian economics to move towards the American-based mainstream of economics, with its typical formalistic and quantitative features. This change has been symbolized by the move to English as *lingua franca* not only of economic discussion but also of Italian journals, conferences and Ph.D. programmes.

See Also

- Barone, Enrico
- Beccaria, Cesare Bonesana, Marchese di
- Einaudi, Luigi
- Galiani, Ferdinando
- Genovesi, Antonio
- Modigliani, Franco
- Ortes, Giammaria
- Pantaleoni, Maffeo
- Pareto, Vilfredo
- Sraffa, Piero

- Verri, Pietro

Bibliography

Asso, P.F., ed. 2001. *From Economists to Economists. The International Spread of Italian Economic Thought, 1750–1950*. Florence: Polistampa.

Augello, M.M. and Guidi, M.E.L. 1996. The emergence of economic periodical literature in Italy (1750–1900). In *Political Economy in European Periodicals, 1750–1900*, ed. M. Bianchini. Special issue of *History of Economic Ideas* 4(3), 15–62.

Augello, M.M. and Guidi, M.E.L. 2001. The associations of economists and the dissemination of political economy in Italy. In *The Spread of Political Economy and the Professionalisation of Economists. Economic Societies in Europe, America and Japan in the Nineteenth Century*, ed. M.M. Augello and M.E.L. Guidi. London: Routledge.

Augello, M.M. and Guidi, M.E.L. 2005a. The Italian economists in Parliament from 1860 to 1922: a quantitative analysis. *European Journal of the History of Economic Thought* 12, 279–319.

Augello, M.M. and Guidi, M.E.L. 2005b. Economists and political economy in Parliament from the unification of Italy to the rise of the Fascist regime (1861–1922). In *Economists in Parliament in the Liberal Age, 1848–1920*, ed. M.M. Augello and M.E.L. Guidi. Aldershot: Ashgate.

Augello, M.M. and Guidi, M.E.L., eds. 2007. *L'economia divulgata (1840–1922). Stili e percorsi italiani*, Vol. I, *Manuali e trattati*; Vol. II, *Teorie e paradigmi*; Vol. III, *La «Biblioteca dell'economista» e la circolazione internazionale dei manuali*. Milan: Franco Angeli.

Bartoli, H. 2003. *Histoire de la pensée économique en Italie*. Paris: Publications de la Sorbonne.

Barucci, P. 1972. The spread of marginalism in Italy (1871–1890). *History of Political Economy* 4, 512–31.

Barucci, P., ed. 2003. *Le frontiere dell'economia politica. Gli economisti stranieri in Italia: dai mercantilisti a Keynes*. Florence: Polistampa.

Bellanca, N. 1993. *La teoria della finanza pubblica in Italia, 1883–1946. Saggio storico sulla scuola italiana di finanza pubblica*. Florence: Olschki.

Bianchini, M. 1982. *Alle origini della scienza economica: felicità pubblica e matematica sociale negli economisti italiani del Settecento*. Parma: Studium. French translation: *Bonheur public et méthode géométrique: enquête sur les économistes italiens (1711–1803)*. Paris: Ined, 2002.

Bianchini, M. 1989. Some fundamental aspects of Italian eighteenth-century economic thought. In *Perspectives on the History of Economic Thought*, ed. D.A. Walker. Aldershot: Edward Elgar.

Bianchini, M. 1994. The Galilean tradition and the origins of economic science in Italy. In *Political Economy and National Realities*, ed. M. Albertone and A. Masoero. Turin: Fondazione Luigi Einaudi.

Buchanan, J. 1960. ‘La scienza delle finanze’: the Italian tradition in fiscal theory. In *Fiscal Theory and Political Economy. Selected Essays*. Chapel Hill, NC: University of North Carolina Press.

Di Battista, F. 1983. *L'emergenza ottocentesca dell'economia politica a Napoli*. Bari: Facoltà di Economia e Commercio.

Faucci, R. 2000. *L'economia politica in Italia dal Cinquecento ai nostri giorni*. Turin: Utet Libreria.

Faucci, R. and Perri, S. 1995. Socialism and marginalism in Italy (1880–1910). In *Socialism and Marginalism in Economics, 1870–1930*, ed. I. Steedman. London: Routledge.

Fausto, D. and De Bonis, V., eds. 2003. *The Italian Tradition of Public Finance, special issue of Il Pensiero economico italiano* 11(1).

Garofalo, G. and Graziani, A., eds. 2004. *La formazione degli economisti in Italia (1950–1975)*. Bologna: Il Mulino.

Guidi, M.E.L. 2000. Corporative economics and the Italian tradition of economic thought: a survey. *Storia del pensiero economico* 40, 31–58.

Mancini, O., Perillo, F.D. and Zagari, E., eds. 1982. *Teoria economica del corporativismo*, 2 vols. Naples: Edizioni scientifiche italiane.

Meacci, F., ed. 1998. *Italian Economists of the 20th Century*. Cheltenham: Edward Elgar.

Nuccio, O. 1984–7. *Il pensiero economico in Italia (1050–1750)*, 7 vols. Rome: Mediocredito Centrale.

Parisi, D. 1984. *Il pensiero economico classico in Italia (1750–1860). Criteri definitori ed evoluzione storica*. Milan: Vita e Pensiero.

Porta, P.L. 1993. A note on Italian economics in the early nineteenth century from Restoration to Risorgimento. *History of Economic Ideas* 1, 43–70.

Porta, P.L. 1996. Italian economics through the postwar years. In *The Post-1945 Internationalization of Economics*. Annual supplement to vol. 28 of *History of Political Economy*. Durham, NC, and London:

Duke University Press.

Quadrio Curzio, A., ed. 1996. *Alle origini del pensiero economico in Italia. 2. Economia e istituzioni. Il paradigma lombardo tra i secoli XVIII e XIX*. Bologna: Il Mulino.

Roggi, P., ed. 2001. *Le grandi 'voci' nei dizionari specializzati (e non) di economia*. Special issue of *Storia del pensiero economico* 41–42.

Romani, R. 1994. *L'economia politica del Risorgimento italiano*. Turin: Bollati Boringhieri.

Roncaglia, A., ed. 1995. *Alle origini del pensiero economico in Italia. 1. Moneta e sviluppo negli economisti napoletani dei secoli XVII–XVIII*. Bologna: Il Mulino.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen & Unwin.

How to cite this article

Augello, Massimo M. and Marco E.L. Guidi. "Italy, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000289> doi:10.1057/9780230226203.0858

The New Palgrave Dictionary of Economics Online

Jaff é , William (1898 – 1980)

Donald A. Walker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

general equilibrium; Jaffé, W.; normative economics; tâtonnement; Walras, L.

Article

Historian of economic thought whose important contributions were to the study of the work of Léon Walras, Jaffé was born in Brooklyn on 16 June 1898 and died in Toronto on 17 August 1980. He graduated from City College of New York with an AB degree in classics and English (1918), from Columbia University with an MA in history (1919), and from the University of Paris with a *Docteur en droit* in economics and political science (1924). He taught economics at Northwestern University (1928–66), and at York University in Ontario (1970–80). Jaffé translated Walras's *Eléments d'économie politique pure* into English (Walras, 1954), thereby providing a major stimulus to the study of his work; edited and exhaustively annotated Walras's scientific correspondence and related papers (Jaffé, 1965a), thereby furnishing an encyclopedic storehouse of information about his writings; and wrote many essays on Walras's economic ideas (Walker, 1983). Jaffé believed that, even in its scientific aspects, a writer's work reveals the influence of his normative views and intellectual environment, and that to understand his work fully it is therefore necessary to study his biography and the era of which he was a part (Jaffé, 1965b). He applied this thesis to the study of Walras's work, examining the aspects of his biography that had a bearing on his theories, explaining the antecedents of his scientific ideas and the philosophical sources of his normative conceptions, and interpreting and assessing his theories of demand, exchange, production, capital formation, money, tâtonnement, and general economic equilibrium.
In an extreme change of opinion, Jaffé came to believe, in the last seven years of his life, that Walras's theory of general equilibrium was intentionally a normative scheme, and that his theory of tâtonnement was intentionally a normative exercise in static analysis (Jaffé, 1980; 1981). It would be a disservice to Jaffé and a denial of his scholarship not to recognize that his soundest judgements on Walras were made during the first 43 years of his study of Walras' work, when he regarded Walras's economic theories as positive in intent and character and the theory of tâtonnement that Walras espoused during most of his career as an attempt to describe the general features of the dynamic adjustment of the market system

toward equilibrium (Jaffé, 1967).

Selected works

1942. Léon Walras' theory of capital accumulation. In *Studies in Mathematical Economics and Econometrics*, ed. O. Lange et al. Chicago: University of Chicago Press.

1965a. (ed.) *Correspondence of Léon Walras and Related Papers*, 3 vols. Amsterdam: North-Holland.

1965b. Biography and economic analysis. *Western Economic Journal* 3, 223–32.

1967. Walras' theory of *tâtonnement*: a critique of recent interpretations. *Journal of Political Economy* 75, 1–19.

1972. Léon Walras's role in the 'marginal revolution' of the 1870s. *History of Political Economy* 4, 379–405.

1980. Walras's economics as others see it. *Journal of Economic Literature* 18, 528–49.

1981. Another look at Léon Walras's theory of *tâtonnement*. *History of Political Economy* 13, 313–36.

Bibliography

Walker, D.A. 1981. William Jaffé, historian of economic thought, 1898–1980. *American Economic Review* 71, 1012–19. (This contains a complete bibliography of Jaffé's published writings.)

Walker, D.A. 1983. William Jaffé, *Officier de Liaison intellectuel*. In *Research in the History of Economic Thought and Methodology*, ed. W.J. Samuels. Greenwich, CT: JAI Press.

Walker, D.A., ed. 1983. *William Jaffé's Essays on Walras*. New York: Cambridge University Press.

Walras, L. 1954. *Elements of Pure Economics*. Homewood, IL: Richard D. Irwin, trans. and annotated by W. Jaffé.

How to cite this article

Walker, Donald A. "Jaffé, William (1898–1980)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000001> doi:10.1057/9780230226203.0859

The New Palgrave Dictionary of Economics Online

Japan, economics in

Tamotsu Nishizawa and Aiko Ikeo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economics in Japan seems to have developed in two major different ways, political economy and neoclassical economics. The traditional notion of ‘administering the nation and relieving the suffering people’ continued to exert a strong influence on political economists. The German Historical School and then Marxian economics also maintained their very strong traditional hold up to the 1960s, as in other late-developing countries. Neoclassical and Keynesian economics began to develop in the 1930s. Some theoretical economists began to make international contributions to studies of the general equilibrium approach, welfare economics, and trade theory in the 1950s.

Keywords

Akamatsu, K.; Amano, T.; Amonn, A.; Aoyama, H.; Arisawa, H.; Ashley, W. J.; Bastable, C. F.; Brentano, L.; Brouwer's fixed-point theorem; Cambridge School; Carey, H. C.; Chenery, H.; Civil Society School; cobweb theorem; comparative advantage; Cournot, A. A.; developmental state; differential–difference models; dual structure; Econometric Society; economic development; Ely, R. T.; Engel's law; Fenollosa, E.; fixed-point theorems; free trade; Fukuda, T.; Fukuzawa, Y.; Furuya, H.; general equilibrium; German Historical School; Gossen, H. H.; Hayakawa, M.; Hecksher–Ohlin trade theory; Hirai, Y.; Hobson, J. A.; Ichimura, S.; Inada, K.; industrial policy; infant-industry protection; Inoue, T.; international division of labour; international trade; Ishibashi, T.; Japan, economics in; Japanese Economic Association; Kakutani's fixed-point theorem; Kanai, N.; Kawakami, H.; Kemp, M. C.; Keynesianism; King's law; Kitamura, H.; Komiya, R.; Koza School; Kumazawa, B.; Kushida, T.; laissez-faire; law of demand; Leontief, W.; Liapunov, A. M.; List, F.; London School Institutionalists; Marshall, A.; Marxism; Masuji, Y.; mathematical economics; Mill, J. S.; Ministry of International Trade and Industry (MITI) (Japan); Miyamoto, K.; Moore, H. L.; Morishima, M.; Nakayama, I.; Negishi, T.; neoclassical economics; Nikaido, H.; Ohkawa, K.; Okawa, M.; Okita, S.; Oniki, H.; Otsuka School; Ouchi, H.; over-consumption; Oyama, M.; partial equilibrium; Persons, W. M.; Pigou, A. C.; planning;

poverty; retirement; Ricardian trade theory; Rono School; Roscher, W. G. F.; roundabout production; Sato, N.; Schumpeter, J. A.; separability; Shibata, K.; Shimomura, K.; Shinohara, M.; Smith, A.; socialism; Sono, M.; stability; Stackelberg, H. von; Sugimoto, E.; Taguchi, U.; Takano, I.; Takashima, Z.; Takata, Y.; tariffs; Tobata, S.; Tsuru, S.; two-sector models; Uchida, T.; Uchida, Y.; Ueda, T.; Uekawa, Y.; Uno, K.; Uzawa, H.; Wadagaki, K.; wage differentials; Wagner, A.; Walras, L.; Watanabe, T.; Webb, B. and S.; welfare economics; Yagi, Y.; Yamada, I.; Yamada, M.; Yamada, Y.; Yasui, T.

Article

Economics and economic thought in Japan have changed and progressed in response to various phases of Japanese historical and social development. The Meiji Restoration of 1867 and the Second World War were the two most obvious phases. Before the Meiji Restoration, there were very marked differences between Japanese and Western approaches to economic problems, though even in the Tokugawa era (1603–1867) problems common to East and West seem to have generated some similar economic answers. When Japan opened up to the West in 1867, and when the state came to play a vital role in retaining national independence and promoting rapid industrialization, it is hardly surprising that the ideas of laissez-faire had less appeal than the nation-centred developmentalism of the German Historical School, which was propagated largely from (Tokyo) Imperial University (founded by the government). However, there also existed in Japan a tradition of British liberal economics, especially at the private universities (such as Waseda) and the Higher Commercial Schools (such as Hitotsubashi) (founded by the private citizens). Further, the traditional Japanese notion of the ‘economy’, as it says in Confucianism, ‘administering the nation and relieving the suffering people’, continued to have a strong influence on Japanese political economists even after the Second World War, while the notion of economics as ‘science’ rather than an ‘art’ – the modern neoclassical view of economics – was, with a few exceptions, generally put to one side, particularly before the Second World War.

From the Meiji Restoration to the First World War: the making of modern economic thinking in Japan

With the Meiji Restoration, the flow of Western ideas into Japan turned into a flood, and the study of Western economic ideas and institutions was incorporated into Japan's new knowledge. Though Western economic liberalism awakened modern Japanese intellectuals, it is helpful to think of pre-Meiji traditions of knowledge as providing the framework that determined the types of Western ideas that were widely accepted. Japanese thinkers selected certain parts of Western knowledge as relevant to their interests and gave them a Japanese interpretation.

For the economic thinkers of the early Meiji era, the simultaneous introduction of an industrial capitalist system and its institutions and of Western theories was to create formidable intellectual problems. Two major intellectuals, Yukichi Fukuzawa (1835–1901) and Ukichi Taguchi (1855–1905), were deeply committed to the so-called ‘civilization and enlightenment’ movement and interested not just in economic thought, industry and trade, but in a wide range of subjects related to the humanities and morality. Fukuzawa aimed to promote civilization by advocating ‘wealth’ and ‘virtue’ as means of retaining national independence and of making Japan develop into a strong and wealthy nation, and so suggesting a protectionist policy. His attempt to provide a realistic response to Japan's situation meant

that his views and ideas were complex, but this very complexity fostered many good economists, industrialists and businessmen, who studied at Keio Gijuku (later Keio University), which he founded in 1858.

On the other hand Taguchi (1878), the author of Japan's version of ‘Manchesterism’, believed in a harmonious natural law and the universal applicability of free trade. He took forward the banner of laissez-faire doctrine in Meiji Japan with his journal *Tokyo Keizai Zasshi* [Tokyo Economist], which was founded in 1879 and remained active until 1923. Another major journal, *Toyo Keizai Shinpo* [Oriental Economist] was founded just after the Sino–Japanese war (1894–5), at the time of Japan's first industrialization, and edited by people such as Tameyuki Amano (1859–1938), a liberal economist at Waseda who had translated J.S. Mill's *Principles*, and by his pupils. This journal propagated the ideas and policy of new liberalism in Japan, and from 1924 was edited by Tanzan Ishibashi (1884–1973). Ishibashi was active in debates on lifting the gold embargo and later became finance minister (1946–7); he was sympathetic to the economic ideas of J.M. Keynes.

From the late 1880s to the mid-1890s (the second decade of the Meiji period), Japan's economic studies increasingly moved away from English liberal economics towards the German Historical and Social Policy School. This new historical and ethical thinking and the adoption of German financial science in Japan first came about through the 1880 English translation of *Guida allo Studio dell’ Economia Politica* (1876) by Luigi Cossa (1831–96, Italian historical economist), and the books by R.T. Ely (1854–1943, American historical economist and a founder of the American Economic Association). Economic discourse by H.C. Carey (1793–1879) and the English translation of *Das nationale System der politischen ökonomie* (1841) by Friedlich List (1789–1846) were propagated through the Japanese National Economics Association (Kokka Keizai Kai, established in 1890), and appealed to those concerned with national independence and the protection of infant industries.

The Meiji governments promoted a developmental state policy that followed the Prussian model of rapid modernization and industrialization; but this caused social problems. The (Tokyo) Imperial University (so called after the Imperial University Act of 1886) became the centre for the dissemination of German ideas in Japan, largely through the *Kokka Gakkai Zasshi* [Journal for State Science], which was founded in 1887. In 1888, Kenzo Wadagaki (1860–1919), who succeeded Ernest Fenollosa, the first professor of economics in Japan, wrote a pioneering article titled ‘Kodan Shakaito’ [The Socialist Party of the Chair]. Noboru Kanai (1865–1933) was instrumental in implanting the German Historical School in Japan and in establishing its theories and policies. Marshall and Mill were still studied, however, at private universities, and Hitotsubashi. T. Inoue (at Waseda) translated Marshall's *Elements of the Economics of Industry* (1896) into Japanese; this soon became a best seller and in 1902 went into its 11th edition.

The Japanese Association for the Study of Social Policy was set up in 1896 to investigate factory laws abroad. Faced with domestic labour problems, the Association, which was opposed to laissez-faire liberalism and to socialism, aimed to prevent class conflict and to sustain social and industrial peace by means of economic freedom and state intervention. Its thinking reflected the pre-Meiji tradition of ‘administering the nation and relieving the sufferings of the people’, and it considered that economics was interwoven with moral and political issues and embodied the duty of the government to be concerned for the social welfare of its subjects. The Association organized an annual conference and discussed not only labour, but also tariff problems, small industries, the peasantry, and other issues. Iwasaburo Takano (1871–1949), a core member of the Association who studied with Georg von Mayr, founded a strong tradition of social statistics in Japan and later directed the Ohara Institute for Social

Research (which had been founded by a cotton giant in 1918), which the Marxists expelled from Tokyo University were to make into a centre for Marxian studies before the Second World War.

In 1906, following the Russo-Japanese war, the *Kokumin Keizai Zasshi* [Journal of National Political Economy], co-edited by the staff of the Higher Commercial Schools at Hitotsubashi and Kobe, first appeared. This became Japan's first proper economics journal and a de facto organ of the Association. While Kanai and his followers at Tokyo Imperial University moved towards Adolph Wagner's style of state socialism, Tokuzo Fukuda (1874–1930) at Hitotsubashi and his followers at the Higher Commercial Schools were sympathetic to ‘reform liberalism’ and were closer to the ideas of British political economists. In the Higher Commercial Schools business economics, industrial studies (particularly in small-scale industries) and financial and monetary studies were also well developed. Such a monetary economics tradition made a good basis for the introduction of Keynesian economics into Japan. Keynes's *Treatise on Money* was translated into Japanese in 1932–4, and the ‘fever’ of the *General Theory* took hold at Hitotsubashi soon after the book's publication, giving rise to the formation of a group of Keynesian economists.

Teijiro Ueda (1879–1940) studied ‘business policy’ in England with W.J. Ashley, and in 1909, on his return to Japan, he began to lecture on business administration. Highly impressed by Ashley's ‘The Enlargement of Economics’ (1908) and his proposal for making ‘business economics’, Ueda wrote about and tried to create a business economics aimed at high efficiency rather than high profit, and a science of socially efficient management similar to that established by German business economists such as H. Nicklish. He subsequently lectured on joint stock companies, social reconstruction and the role of managers, stressing ‘the duties of managers’. Ueda published *Shakai Kaizo to Kigyo* [Social Reconstruction and Business Enterprise] (1921), *Shinjiyushugi* [New Liberalism] (1927), and others, issuing his own journal titled *Kigyo to Shakai* [Business Enterprise and the Society]. He actively pursued free trade, and was opposed to socialism, protectionism and the imperialist economic blockade in the 1930s.

In the 1920s, while Marxist studies flourished in academic circles in Japan, particularly at the imperial universities, business economics and management studies also prospered against the background of the rapid development of the corporate economy after the First World War. Ueda's business studies were followed and developed by Y. Masuji at Tokyo and Y. Hirai at Kobe, while F. Muramoto, the first Japanese MBA from Harvard, began to lecture on scientific management at Osaka Higher Commercial School in the very early 1920s. Ashley's pioneering efforts in creating the study of business economics, which were not followed up in Britain, developed at the expanding Higher Commercial Schools and the universities of commerce in Japan. In 1926, the year Ashley's *Business Economics* was published, the Japanese Society of Business Administration was founded, its original membership numbering 342.

Before the Second World War, the Higher Commercial Schools and the universities of commerce played a significant role in the development of economics and business studies. Until Marxian economics became dominant in the 1920s, economics in Japan was very much in a tradition of the German Historical and Social Policy School in broad sense. Japanese economists caught up with many developments very early on, and were innovators as well as consumers of foreign ideas, though they did not develop systematically or perceive the whole economy as a single system.

Fukuda, Kawakami and the Marxian tradition

During the years of the Taisho democracy movement, the Russian Revolution, and rice riots after the First World War, Marxism emerged and began to flourish among Japanese intellectuals, quickly replacing the Historical School. The establishment of economics faculties at the imperial universities in Tokyo and Kyoto and the inauguration of Tokyo University of Commerce took place at about the same time. Initially, Hajime Kawakami (1879–1946) at Kyoto and Fukuda were the leading figures in the study of Marxian economics, whereas Fukuda pioneered the study of welfare economics and the welfare state against Marxism. The newly created economics faculty at Tokyo produced and was dominated by a number of eminent Marxian economists, such as Hyoe Ouchi, Moritaro Yamada, Hiromi Arisawa and Kozo Uno. Many young scholars went to study in Germany.

While Fukuda and Kawakami were initially heavily influenced by the German Historical School, they began to develop original perspectives by assimilating various new trends in economics. Fukuda had been inspired by Roscher and Marshall since his student days, and in Germany studied with Brentano, with whom he co-authored *Rodo Keizairon* [Labour Economics] (Brentano and Fukuda, 1899), discussing working conditions, productivity and the working people's welfare. Fukuda's economic studies covered a wide range of subjects, the most important of which were probably welfare economics and social policy. Though he studied the orthodox welfare economics of Marshall and Pigou, it was from J.A. Hobson that Fukuda learned most about the ethical and humanist approach to welfare economics. Just like the American Institutionalists, Fukuda became openly sympathetic to the idealist, historical and ethical approach of the Oxford economists (or ‘London School Institutionalists’), rather than to the so-called neoclassical Cambridge School of utilitarian economists.

Fukuda contended for social policy (or welfare economic studies) as an alternative to Marxism, and proposed a welfare struggle, not a class struggle. Inspired by Lorenz von Stein and Anton Menger, Fukuda developed the theory of social rights, particularly the right to live (needs), and made it the foundation of social welfare policy. This was similar to the Webbs's ‘national minimum’. The art of economics would be to provide the economic basis for the minimum human life and to make cultural and moral development possible, as Fukuda learnt from his contemporaries such as A. Marshall and C.J. Fuchs. These ideas lay at the root of Fukuda's welfare economic studies, and formed the basis for the welfare state, as evaluated by people like Yuzo Yamada (1902–1996), who followed and developed Fukuda's ideas in the theory of economic planning and national income, and Ichiro Nakayama (1898–1980), who applied and extended Fukuda's ideas after the Second World War, stabilizing industrial relations in order to increase productivity and proposing the doubling of wages, which formed the basis of the income-doubling policy in the high-speed economic growth of the 1960s.

The transition to Marxism and political activism in the 1920s is well illustrated in the career of Kawakami. Initially idealistic and much concerned with problems of morality, rather like Noboru Kanai, Kawakami was deeply disturbed by the poverty that he encountered in the slums of London (see his best-selling book *Binbo Monogatari* [A Tale of Poverty], 1917). He argued that production in the capitalist system was designed not to fulfil human needs: the basic needs of the poor were ignored because they were not expressed in term of monetary demand, which led to over-consumption by the rich. Kawakami linked modern economic analysis to the moral precepts of Tokugawa philosophers such as Banzan Kumazawa, whose ‘frugality’ of the rich and proposals for the nationalization of industry and state-run welfare schemes reflected a late-Tokugawa agriculturalist Nobuhiro Sato's egalitarian nationalism. For Kawakami, the ultimate object of economics was to make human beings more fully human. Despite its wide popular appeal, *A Tale of Poverty* was criticized by younger scholars, such as his former student

Tamizo Kushida (1885–1934), whose debates were to play a vital role in developing Marxian economics among the younger generation.

Marxist ideas have had their greatest impact on the peripheral nations of the capitalist world such as Russia. Japan was a latecomer in the industrialized world and had large agrarian sectors in the pre-war period in which the pre-capitalist remnants were slowly disappearing. Marxist economic thought became entangled in questions of political strategy and generated a debate over the possibility of ‘premature’ revolution within a semi-developed capitalist society. The Koza (Lecture) School, named after the *Nihon Shihonshugi Hattatsushi Koza* [Lectures on the Historical Development of Japanese Capitalism] (1932–3), defined as its objectives the bourgeois democratic reforms that must precede a future socialist revolution. Moritaro Yamada (1897–1980) was influential in developing the distinctive Koza School approach. This was criticized by the Rono (Worker–Farmer) School, which had separated from the Communist Party and aimed to create a mass organization of workers, peasants and others that would evolve into a revolutionary movement to overthrow capitalism.

The influence of Marxism on Japanese intellectual life reached a high point in the decades after the Second World War. During the Allied occupation, many of those who had played prominent parts in Marxist debates and who had been expelled from their chairs re-emerged as dominant figures in the economics faculties of universities. Yamada suggested that the Allies’ occupational reforms had brought about Japan's long-delayed bourgeois democratic revolution, and he sat on the Land Reform Committee for creating the landed farmers and democratizing the farmland system. The Otsuka School of economic history, which emerged from the Koza School and was led by Hisao Otsuka (1907–96), and the Civil Society School led by followers of Adam Smith studies like Zenya Takashima (1904–90) and Yoshihiko Uchida (1913–89), also played an important role in the post-war modernization and democratization of Japan. Adam Smith studies constituted a strong tradition in the development of Japan's economic thought.

The years immediately after 1945 were marked by the active participation of economists in policymaking, and ‘Marxian’ economists made a disproportionate contribution. Among them were Hyoe Ouchi (1888–1980) and Hiromi Arisawa (1896–1988), who, together with Ichiro Nakayama and Seiichi Tobata, were important members of a research committee of the Foreign Ministry that published *Nihon Keizai Saiken no Kihon Mondai* [The Basic Problems of Reconstructing the Japan's Economy] (1946), stressing democratization and advocating the importance of government planning and intervention for the recovery of the Japan's economy. Arisawa, who had stressed the concept of socialization since his studies in Weimar Germany in the mid-1920s, in 1946 proposed to the Cabinet the priority of producing coal and steel, based on the Austrian idea of roundabout production and Marx's two-sector production model. He had developed his ideas of a managed economy and also generated a theory of ‘dual structure’, namely, the coexistence of large-scale and small-scale industries; the gap between them was a structural problem that was also the problem of employment and wage structure. Such a structural gap was also seen between the modern industrial sector and the pre-modern agricultural sector, which had a large underemployed labour force. The theory of dual structure was developed into a theory of wages differentials and value-added differentials across industries based on empirical and statistical studies by Miyohei Shinohara (born 1919) in the high-speed economic growth era from around 1960. Shinohara, who ‘formed his analytical framework, taking in Keynes, Hayek, Marx, properly revised, and making the reality as the last resort’ (Shinohara, 1987, vol. 4, p. 395), had studied Nakayama's economics and

Kaname Akamatsu's synthetic dialectics and soon worked with Kazushi Ohkawa (1908–93) and the Hitotsubashi group of the Institute of Economic Research on *Choki Keizai Tokei* [Long-Term Economic Statistics in Japan] (14 volumes, 1965–88), which was inspired by Yuzo Yamada's *Nihon Kokuminshotoku Suikei Shiryo* [A Comprehensive Survey of Japan's National Income Data] (1951). Along with these studies on national income statistics, advances were made in empirical, quantitative and theoretical studies on the Japanese economy.

The spread of neoclassical economics

Fukuda was the economist most responsible for introducing many of the latest economic trends into Japan. He advised his students to translate *Das Kapital* almost in parallel with Marshall's *Principles*, whose full translation came out in 1926. In the late 1920s he created a system of parallel lectures in economics, that is, Marxian and modern (neoclassical) economics, a system that was retained at most Japanese universities until the 1980s. From about 1930 international journals were another important source for the spread of neoclassical economics in Japan. Some Japanese economists were regular readers of *Zeitschrift für Nationalökonomie* (1930–), *Econometrica* (1933–), and the *Review of Economic Studies* (1933–). In 1934, the Japanese Economic Association was established by leading theoretical economists, and in 1997 the Association was re-founded, its tradition thus maintained and even expanded.

General equilibrium theory was introduced through four channels in the 1920s. First, from 1921 to 1922 Fukuda advised his student Nakayama to study Cournot, Walras and Gossen, the classics of mathematical economics. Second, Alfred Amonn (1883–1962), a Czech who had studied at the University of Vienna, explained Cassel's simplified system of general equilibrium in classes at the Imperial University of Tokyo between 1926 and 1929. Third, J.A. Schumpeter, admirer of the general equilibrium theory, had some influence on Japanese economists in the 1920s and 1930s. Two Japanese economists, Tobata and Nakayama, who later became influential in Japan, had studied under him in Bonn. Schumpeter's *Das Wesen und der Hauptinhalt der theoretischen National-ökonomie* [Essence and Main Content of Economics] (1908) was translated into Japanese: there is no English edition. Schumpeter advised two other young economic theorists, Miyoji Hayakawa (1895–1962) and Takuma Yasui (1909–95), to begin with Walras. Fourth, Cassel's system was also taught by Yasuma Takata (1883–1972) at Kyoto Imperial University after 1929. Thus general equilibrium theory circulated in Japan a little earlier than in the English-speaking world.

In 1929, Takata began to publish his *Keizaigaku Shinko* [New Lectures in Economics] (1929–32). This constituted a survey of what was happening in economics, including a discussion of both general equilibrium theory and partial equilibrium theory. Succeeding Fukuda, Nakayama lectured in neoclassical economics and statistics at Hitotsubashi, and his textbook *Junsui Keizaigaku* [Pure Economics] (1933), contributed to the popularizing of general equilibrium theory and the basic concepts of neoclassical economics in Japan. He explained the methodology of pure economics or general equilibrium theory, then the theory of economic development, following the Schumpeterian path. Nakayama further tried to take in Keynes's theory of the investment multiplier as an analytical means to connect dynamic and static aspects of economy in his *Hatten Katei no Kinko Bunseki* [Equilibrium Analysis of the Developing Process] (1939).

Some good statistical studies were made relating to rice. Yoshinosuke Yagi (1895–1944) conducted a full-scale statistical study of rice by surveying current studies. He confirmed that King's Law or the law of demand existed in the case of rice, and showed that Engel's law also held true. Yagi calculated not only the demand elasticity of rice with respect to the price, but also constructed the price and quantity indices following W.M. Persons's method. Excellent econometric studies were carried out in the 1930s as an application of Marshallian economics. Eiichi Sugimoto (1901–52), who studied under Fukuda and taught Marxian economics at Hitotsubashi, was very Marshallian and stressed partial equilibrium, time elements and elasticity, and criticized pure economics. In his *Beikoku Juyo-hosoku no Kenkyu* [Study on the Law of Demand for Rice] (1935), Sugimoto regarded per capita consumption of rice as the demand for rice following the cobweb theorem that the points on the demand curve were realized in the case of disequilibrium. Following H.L. Moore's extension of Marshallian demand analysis (1929), Sugimoto included not only the price of rice but also the prices of all other commodities and time as variables in the rice demand function. He judged that the effects of the changes in the prices of non-rice commodities on the demand for rice should cancel each other out because there were neither close substitutes nor complimentary goods for rice. He divided the rice price index by the general price index and got the rice rate to remove the effect of the changes in the other prices on the rice price. Then he estimated the demand function for every seven years using the least squares method. These early econometric works prompted the study of neoclassical theory of demand and supply.

In 1930 Kei Shibata (1902–86) at Kyoto examined Cassel's ‘mechanism of price formation’ (Shibata, 1930) and explained one of the formal problems in Cassel's simplified system of general equilibrium three years earlier than H. von Stackelberg. Shibata created numerical examples and counter-examples, and published a series of theoretical papers in English in *Kyoto University Economic Review*. It is also noteworthy that Shibata's review article (1937) of Keynes's *General Theory* (1936) included Keynes's own comments on the draft and was praised by D. Dillard's in *The Economics of John Maynard Keynes* (1948), although it was critical of Keynes's macroeconomic analysis for lacking technological changes in production and the transmission mechanism from the increase in savings to the increase in investment.

Takuma Yasui (1909–95) can be called the Japanese Samuelson. He attended Amonn's lectures at Tokyo and studied the work of Nakayama and Takata. He began to publish a series of papers on the Walrasian general equilibrium framework in 1933. In his article ‘Juyo no hosoku nitsuite’ [On the law of demand], Yasui (1940) developed a sophisticated analysis of consumer behaviour, generating the law of demand along the lines of Slutsky, Hicks and Allen. He made a step forward in obtaining the universal law of demand and tried to clarify the conditions under which the demand curve is convex or concave. Masazo Sono (1886–1969), a mathematician at Kyoto, discussed the separability of goods in his ‘Kakaku hendo ni tomonau bunrikanouzai no jukyu hendo’ [Effect of price changes on the demand and supply of separable goods] (1943). By discussing J.R. Hicks's definitions of substitutability and complementarity among commodities, Sono developed the idea of the separability of commodities in terms of utility. The English version of Sono's paper, which appeared in the *International Economic Review* in 1961, anticipated similar studies published later in English.

The study of general equilibrium theory

It is well known that a number of Japanese economists began to contribute to the study of mathematical

economics around 1950. In retrospect, this happened earlier: from around 1930 onwards Japanese mathematicians spread contemporary mathematical knowledge by producing new textbooks in Japanese. In the 1940s several Japanese economists made important contributions to stability analysis, mostly in Japanese but comparable to the studies developed in North America and Europe in the 1950s. The economists Takuma Yasui, Hideo Aoyama (1910–92) and Michio Morishima (1923–2004) and the mathematician Masazo Sono studied stability analysis and the problem of the market mechanism and economic dynamics, discussing not only the mathematical implications of the economic models but also the economic meanings of the mathematical models. By 1950, Yasui and Morishima had studied the conditions for the stability of a competitive equilibrium with the use of a system of ordinary differential equations and reached the qualitative theory of stability developed by A.M. Liapunov, who was gaining popularity outside Russia but was as yet little known to Western economists. As a result of this research, Takashi Negishi wrote his famous and influential article ‘Stability of a competitive economy: a survey article’ (1962), which provided a good survey of stability analysis.

At Kyoto Hideo Aoyama studied the dynamics of economic exchange. His article ‘Mirudaru no keizai hendo riron’ [Myrdal's theory of economic fluctuation] (1938b) started with the cumulative processes of inflation and deflation, which were articulated in Wicksell's monetary economics. He elaborated D.H. Robertson's step-by-step analysis and the period analysis by Myrdal, Lindahl and Ohlin. Aoyama also traced differential–difference models, which were set out by R. Frisch, H. Holme and M. Kalecki. Aoyama later published the English version ‘A Critical Note on D.H. Robertson's Theory of Savings and Investment’ (1940).

In his ‘Seigakuteki ippankinkoron to dogakuka no mondai’ [Static theory of general equilibrium and its dynamization] (1938a) Aoyama picked on the concept of momentary dynamic equilibrium in Frisch's ‘Statikk og dynamikk’ [Statics and dynamics] (1929) and discussed a sequence of momentary equilibrium that was established in a Walrasian exchange economy with multiple commodities. In his ‘Gendai keiki riron niokeru hanro hosoku no mondai’ [On the law of market in the contemporary theories of business cycles](1942) he examined the concept of general dynamic economic equilibrium and pointed out that Hicks's ‘temporary equilibrium’ was the same notion as Frisch's ‘momentary dynamic equilibrium’ and La Volpe's ‘general dynamic economic equilibrium’.

In the 1950s, the proof of the existence of a general competitive equilibrium utilized set theory and convex set method, and a fixed-point theory. Around 1954, Hukukane Nikaido (1923–2001) in Tokyo made a special study of the existence question independently of K.J. Arrow and G. Debreu's ‘Existence of an equilibrium for a competitive economy’ (1954). Nikaido's ‘On the classical multilateral exchange problem’ was published in *Metroeconomica* of 1956. Nikaido formulated the basic propositions of the existence of general equilibrium as a theorem relating to the excess demand correspondence in the case of multilateral exchange of many commodities. Resorting to slightly more restricted assumptions than Arrow and Debreu, Nikaido proved this with the direct use of Kakutani's fixed-point theorem.

Hirofumi Uzawa (born 1928) proved in his ‘Walras's existence theorem and Brouwer's fixed-point theorem’ (1962) that the two theorems in the title were equivalent. He noted that it had already been well established that Brouwer's fixed-point theorem implies Walras's existence theorem. He constructed an excess demand function, which satisfied the conditions describing Walras's existence theorem. By dividing a price by the summation of prices, Uzawa neatly proved that Walras's existence theorem implies Brouwer's fixed-point theorem. Though he was at Stanford University, the paper appeared in

Kikan Riron Keizaigaku [Economic Studies Quarterly], which was the formal journal of the Japanese Association of Theoretical Economics and the Japanese Econometric Society (now the *Japanese Economic Review*, published by the Japanese Economic Association).

In the 1950s, Japanese economists such as Nikaido, Uzawa, Kenichi Inada (1925–2002), Hajime Oniki (born 1933) and Takashi Negishi (born 1933) joined K.J. Arrow's project at Stanford backed by the Office of Naval Research. They played active roles in the study of the existence and stability of a general equilibrium in a competitive economy, two-sector growth models and welfare economics. The mathematical economist David Gale visited Japan, stayed at Osaka University and studied with Nikaido, Shin-ichi Ichimura (born 1925) and Morishima in the mid-1950s. The Japanese dream of intellectual cooperation with Western economists finally became a reality.

Moreover, the generous provisions of the fund for the Government and Relief in Occupied Areas (GARIOA) and, later, the Fulbright scholarship programme brought Japanese youth to the United States and to other countries for advanced study. Ichimura, Tsunehiko Watanabe (born 1926), Tadao Uchida (1923–86), and Ryutaro Komiya (born 1928) were fascinated by American empirical studies, such as the inter-industry analysis originated by Leontief and econometric modelling by Chenery. Returning to Japan, they not only taught American economics but also conducted important econometric works in making economic plans and predictions in the 1960s. Hiroshi Furuya (1920–57), who studied at Harvard from 1952 to 1954, not only strongly advised economics students to study mathematics, but also invited mathematics students such as Hirofumi Uzawa and Ken-ichi Inada to study economics. Moreover, Morishima studied at Oxford and enjoyed attending the meetings organized by J.R. Hicks in 1954 and 1955.

Trade and development

In international economics and economic policy, Japanese economists emphasized different factors from those emphasized by economists in other countries. For example, the Japanese economists who took an interest in policy issues between the 1930s and the 1970s were most interested in the relationship between economic development and international trade. While they shared interests in shifting comparative advantage, dynamic internal economies, and the protection of infant industries, we can usefully divide them in two groups by considering to their backgrounds.

First, some Japanese economists had strong connections with the German-speaking community of economists. Kaname Akamatsu (1896–1974) became known to non-Japanese audiences thanks to his flying-geese-pattern theory (Ganko Keitai Ron) after his paper in English ‘A theory of unbalanced growth in the world economy’ (1961) was published in *Weltwirtshaftliches Archiv*. Akamatsu invented the theory in the 1930s, based on his empirical studies of Japan's woollen industry and later applied it to industries related to cotton yarn, cotton cloth, spinning, weaving machinery and general machinery industries between 1870 and 1940. In the mid-1920s, having spent almost two years in Heidelberg, he garnered ideas about how to do empirical work (business barometers and case study method) during a short stay at the Harvard Business School.

Akamatsu drew three time-series curves denoting the import, the domestic production, and the export of manufactured goods in a plane with time on horizontal axis and the yen value on the vertical axis. He realized that the import curve usually increases until it reaches a peak and declines with the increase of domestic production, at which time the exports increase. This means that many import curves had a

mountain-shape with one peak. This pattern might appear to be similar to that suggested in a Hecksher–Ohlin trade model with many goods. As capital accumulates, there is shifting comparative advantage over time. Originally it was described as the ‘flying-geese-pattern’ theory of industrial development, or the ‘catching-up product cycle’ theory of development, as phrased more accurately in English. Akamatsu believed that his findings for Japan could be generalized into a theory for many countries, especially developing countries or late developers. His development theory has been applied to a picturesque description of a group of developing countries and has led to the discussion of which country is the front flyer.

Hiroshi Kitamura (1909–2002) was, like Akamatsu, critical of the Ricardian theory of free trade and the international division of labour. He left Japan for Europe in 1931, studied economics at the University of Berlin, and specialized in international economics and foreign investment at the University of Basel. In 1941 he published *Zur Theorie des internationalen Handels: Ein kritischer Beitrag* [On the Theory of International Trade: A Critical Contribution]. He developed an early macro-dynamic theory, which was considered to be more appropriate for developing countries than advanced countries, and he supplied theoretical support for protectionist policy such as that advocated by Friedrich List. He brought the theory of trade and development from the German-speaking world back to Japan in 1948 and then he was sent to the United Nations Economic Commission for Asia and the Far East from 1957 to 1969.

Other Japanese economists had close connections with English-speaking economists like Murray C. Kemp. For example, Kemp collaborated with Takashi Negishi, Michihiro Oyama (born 1938), Kouji Shimomura (1952–2007), and Masayuki Okawa (born 1953), and Inada, Uzawa, and Yasuo Uekawa (1925–94) visited the University of New South Wales. Uekawa edited the Japanese version (1981) of Kemp's *Pure Theory of International Trade and Investment* (1969). Kemp's books referred to a number of Japanese works, some of which had been published in international economics journals, others in Japanese university journals or as mimeos.

Among them, Negishi (1972) was interested in the possibility of the protection of an infant industry and debated with Max Corden. Corden and Kemp were critical of the so-called Mill–Bastable case for protecting infant industries, and in response Negishi argued for the protection of the infant industry with emphasis on dynamic internal economies. He maintained that if Bastable's test is understood in terms of increases in social welfare in some sense, then it by no means requires private profitability, and so Kemp's test is not necessary (though it is sufficient) for protection.

The social activities of Japanese economists

The first task for Japanese economists after 1945 was to hasten the recovery of the nation's ruined economy. Japan's top economists joined the Economic Stabilization Board, which was organized on the instruction of the Supreme Commander of the Allied Powers (SCAP) and later reorganized into the Economic Planning Agency. Shigeto Tsuru (1912–2006) joined the Board after he returned from Harvard, where he was trained both in Marxian and neoclassical economics. He brought not only a cosmopolitan attitude but also American economic language into the community of Japanese economists. Tsuru and Saburo Okita (1914–93) co-authored the famous first White Paper on the Japanese economy in 1947, where Tsuru's training as an economist with pragmatic inclinations was vividly revealed. Okita and other officials were trained as economists through working for the Board. Okita was sent to the Economic Commission for the Asia and the Far East (ECAFE), and he was the

chief economic analyst for the Commission in 1952–3. For many years he represented ‘the able Japanese bureaucracy’.

Japan returned to the international community on the basis of the Peace Treaty of 1952, which marked the end of 15 years of central control over the Japanese economy. Governmental agencies such as the Ministry of International Trade and Industry (MITI), the Ministry of Finance (MOF), and the Ministry of Agriculture and Forestry (MAF) began to regulate the economy. For example, the power industry was run by the government from 1939, but in 1951 nine private companies took their businesses back with the slogan of ‘democratization’ or participation. They used a neoclassical-econometric analysis to make a report for MITI on the necessary increase in installed capital, based on estimates of economic growth and the increasing demand for electricity. Neoclassical and Keynesian economists began to collaborate with MITI to rationalize the power industry more thoroughly in 1954. MITI, MOF and MAF made input–output tables of the Japanese economy independently of each other in 1951, which were completed respectively by 1956. They believed that the tables were useful in regulating the economy and in mediating between consumers and producers when Japan began to take growth-oriented policies. The MITI project was the biggest project undertaken, and was led by Shin-ichi Ichimura, who was trained at MIT. In the process of constructing the tables, the quality of the statistical data for national income and wealth was greatly improved. The agencies decided to cooperate with each other beyond bureaucratic sectionalism to make a single 1955 input–output table. From about 1960 they asked neoclassical and Keynesian economists to discuss the economic issues in making policies and to teach mainstream economics to government officials in MITI and other ministries.

The first econometric model of the Japanese economy was made by Isamu Yamada (1909–86) in 1948. After 1957, those who had been trained in the United States began to build econometric models one after another. In 1960, the Ikeda Cabinet decided on the Income Doubling Plan, that is, the economic plan of doubling per capita national income in a decade. After 1960 they asked econometricians to prepare the mid-term plan. A variety of macroeconometric models of the Japanese economy were constructed for various purposes, such as long-term economic forecasts, business cycles explained by changes in investment, and the Klein–Goldberger-type model of the Japanese economy. Tadao Uchida (1925–86), Tsunehiko Watanabe (born 1926), Masahiro Tatemoto (born 1925), Kei Mori (1932–90) and Shuntaro Shishido (born 1924) played leading roles in simulating government policies with the use of the latest computer technology.

The Japanese enjoyed a new way of life, equipped with an increasing number of durable consumer goods, which were the fruits of high-speed economic growth. Yet by 1970, when Japan had become one of the advanced countries, several negative external effects were found in the environment, such as mercury poisoning caused by drainage, air pollution and traffic jams in large cities. It was also realized that the welfare system was not sufficiently developed to provide for an enjoyable retirement. Japanese economists studied a number of problems similar to those that had interested American and European economists.

Tsuru, who had been committed to the basic tenets of Marxian political economy since the late 1920s and who had studied under Schumpeter at Harvard, wrote ‘‘Kokumin Shotoku’ gainen heno Hansei’ [Reflections on the ‘national income’ concept] in 1943, when he was first employed by what later became the Institute of Economic Research (at Hitotsubashi). This was a critique of the market-oriented concept of national income, which would not be a good indicator of welfare. This became his major concern in subsequent years and led to ‘In Place of GNP’ (1971) and formed the basis of *Kougai*

no Seijikeizaigaku [Political Economy of Environmental Disruption] (1972), by pointing out that environmental pollution was not counted negatively in the system of national accounting. He took the initiative in organizing an interdisciplinary Research Committee on Environmental Disruptions in 1963, which was widely supported by economists such as Ken-ichi Miyamoto and Hirofumi Uzawa. Tsuru was involved in a series of international academic activities, culminating in the presidency of the International Economic Association (IEA) from 1977 to 1980.

In terms of international activities, the Econometric Society holds regional meetings in East Asia. The first Far Eastern meeting was held in Tokyo independently of the Society in 1950, although its report appeared in *Econometrica* in 1951. Formal annual meetings were held in Japan from 1966 to 1970. After a long break, biennial meetings have been held every other year somewhere in East Asia since 1987. In July 1997, the fifth meeting was held in Hong Kong, which had been returned to the People's Republic of China a few days earlier. In August 1995, the World Congress of the Econometric Society was held in Tokyo.

After the turning point of 1985

The year 1985 was an important turning point for Japanese economists. American economists, including Paul R. Krugman, were the first to become interested in US–Japanese trade frictions and examined Japanese trade and industrial policies. Japanese economists were then forced to pay attention to the results of this research and the trade dispute itself. They felt obliged to make some response even though they had previously ignored criticism by American politicians and officials, and began to think that some applied economists should come out of their ivory towers, arm themselves with relevant facts and ideas, and face American professional economists in a policy debate. However, the East Asian financial crisis of 1997 made them more concerned about the inter-linkages of national economies and inclined them to conceive of a kind of transnational community or forum for regional economic stability. Although the idea of establishing an Asian Monetary Fund was rejected by the United States and China in 1997, Japanese economists aim to integrate the Chinese economy into international settings. A good economist today may have to handle Chinese, English and Japanese if he or she wants to be active in Japan.

See Also

- general equilibrium
- Historical School, German
- Morishima, Michio
- Tsuru, Shigeto
- Uno, Kozo

Bibliography

Akamatsu, K. 1961. A theory of unbalanced growth in the world economy. *Weltwirtshaftliches Archiv* 86

(2), 196–217.

Aoyama, H. 1938a. Seigakuteki ippankinkoron to dogakuka no mondai [The static theory of general equilibrium and its dynamization]. *Keizai Ronso* 47, 395–409.

Aoyama, H. 1938b. Mirudaru no keizai hendo riron [Myrdal's theory of economic fluctuation]. *Keizai Ronso* 47, 561–81.

Aoyama, H. 1940. A critical note on D.H. Robertson's theory of savings and investment. *Kyoto University Economic Review* 16(1), 49–73; 16(2), 64–81.

Aoyama, H. 1942. Gendai keiki riron niokeru hanro hosoku no mondai [On the law of the market in the contemporary theories of business cycles]. *Nihon Keizai Gakkai Nenpo* 2, 15–136.

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Ashley, W. 1908. The enlargement of economics. *Economic Journal* 18(70), 181–204.

Brentano, L. and Fukuda, T. 1899. *Rodo Keizairon* [Labour Economics]. Tokyo: Dobunkan.

Cossa, L. 1880. *Guide to the Study of Political Economy*. Translated from the 2nd Italian edn. London: Macmillan.

Dillard, D. 1948. *The Economics of John Maynard Keynes: The Theory of a Monetary Economy*. London: C. Lockwood.

Frisch, R. 1929. Statikk og dynamikk [Statics and dynamics]. *National konomisk Tidsskrift* 67, 321–79.

Gao, B. 1997. *Economic Ideology and Japanese Industrial Policy: Developmentalism from 1931 to 1965*. Cambridge: Cambridge University Press.

Ikeo, A. 1994. When economics harmonized mathematics in Japan: a history of stability analysis. *European Journal of the History of Economic Thought* 1, 577–99.

Ikeo, A., ed. 2000. *Japanese Economics and Economists since 1945*. London: Routledge.

Ikeo, A. 2006. *Economics in Japan: The Internationalization of Economics in the Twentieth Century* [in Japanese]. Nagoya: Nogoya University Press.

Kawakami, H. 1917. *Binbo Monogatari* [A Tale of Poverty]. Kyoto: Kobundo-shobo.

Kemp, M.C. 1969. *Pure Theory of International Trade and Investment*. London: Prentice-Hall.

Keynes, J.M. 1936. *The General Theory of Employment, the Interest and Money*. London: Macmillan.

Kitamura, H. 1941. *Zur Theorie des internationalen Handels: Ein kritischer Beitrag* [On the Theory of International Trade: A Critical Contribution]. Basel.

Moore, H.L. 1929. *Synthetic Economics*. New York: Macmillan.

Morris-Suzuki, T. 1989. *A History of Japanese Economic Thought*. London: Routledge.

Nakayama, I. 1933. *Junsui Keizaigaku* [Pure Economics]. Tokyo: Iwanami Shoten.

Nakayama, I. 1939. *Hatten Katei no Kinko Bunseki* [Equilibrium Analysis of the Developing Process]. Tokyo: Iwanami Shoten.

Negishi, T. 1962. Stability of a competitive economy: a survey article. *Econometrica* 30, 635–69.

Negishi, T. 1972. *General Equilibrium Theory and International Trade*. Amsterdam: North-Holland.

Nihon Keizai Saiken no Kihon Mondai [The Basic Problems of Reconstructing the Japan's Economy]. 1946. Tokyo: Foreign Office.

Nihon Shihonshugi Hattatsushi Koza [Lectures on the Historical Development of Japanese Capitalism]. 1932–33. Tokyo: Iwanami-shoten.

Nikaido, H. 1956. On the classical multilateral exchange problem. *Metroeconomica* 8, 135–45.

Ohkawa, K., Shinohara, M. and Umemura, M., eds. 1965–88. *Choki Keizai Tokei* [Long-Term Economic Statistics in Japan], 14 vols. Tokyo: Toyo Keizai Shinpo-sha.

Schumpeter, J.A. 1908. *Das Wesen und der Hauptinhalt der theoretischen National-ökonomie* [Essence and Main Content of Economics]. Leipzig: Duncker & Humblot. Translated into Japanese by K. Kimura and T. Yasui in 1936.

Shibata, K. 1930. Kasseru-shi no ‘kakaku keisei no kiko’ no ginmi [An examination of ‘the mechanism of price formation’ as explained by Mr. Cassel]. *Keizai Ronso* 30, 916–36.

Shibata, K. 1937. Some questions on Mr. Keynes’ general theory of employment, interest and money. *Kyoto University Economic Review* 12(1), 83–96.

Shinohara, M. 1987. *Chosakushu* [Collected Works], vol. 4. Tokyo: Chikuma Shobo.

Society for the History of Economic Thought. 1984. *Nihon no Keizaigaku* [Economics in Japan]. Tokyo: Toyo Keizai Shinpo-sha.

Sono, M. 1943. Kakaku hendo ni tomonau bunrikanouzai no jukyu hendo. *Kokumin Keizai Zasshi* 74, 1–51. Trans. as ‘The effect of price changes in the demand and supply of separable goods’. *International Economic Review* 2 (1961), 239–71.

Sugihara, S. and Tanaka, T. 1998. *Economic Thought and Modernization in Japan*. Cheltenham: Edward Elgar.

Sugimoto, E. 1935. *Beikoku Juyo-hosoku no Kenkyu* [A Study on the Law of Demand for Rice]. Tokyo: Japan Society for the Promotion of Science.

Taguchi, U. 1878. *Jiyukoeki Nihonkeizairon* [Free Trade. On the Japanese Economy]. Tokyo.

Takata, Y. 1929–32. *Keizaigaku Shinko* [New Lectures on Economics]. Tokyo: Iwanami Shoten.

Tsuru, S. 1943. ‘Kokumin Shotoku’ gainen heno Hansei [Reflections on the ‘National Income’ Concept]. *Hitotsubashi Ronso* 12(6), 22–41.

Tsuru, S. 1971. In place of GNP. Originally presented at the Symposium on Political Economy of Environment, Paris, 5–8 July. Repr. in S. Tsuru, *Towards a New Political Economy*. Tokyo: Kodansha, 1976.

Tsuru, S. 1972. *Kougai no Seijikeizaigaku* [Political Economy of Environmental Disruption]. Tokyo: Iwanami-shoten.

Ueda, T. 1921. *Shakai Kaizo to Kigyo* [Social Reconstruction and Business Enterprise]. Tokyo: Shimoyama-shoten.

Ueda, T. 1927. *Shinjiyushugi* [New Liberalism]. Tokyo: Dobunkan.

Uzawa, H. 1962. Walras's existence theorem and Brouwer's fixed-point theorem. *Kikan Riron Keizaigaku* [Economic Studies Quarterly] 13, 59–62.

Wadagaki, K. 1888. Kodan Shakaito [The Socialist Party of the Chair]. *Kokka Gakkai Zasshi* [Journal for State Science].

Yagi, K. 1999. *Kindai Nihon no Shakai-Keizaigaku* [The Social Economics of Modern Japan]. Tokyo:

Chikumashobo.

Yamada, Y., ed. 1951. *Nihon Kokumin Shotoku Suikei Shiryo* [A Comprehensive Survey of National Income Data]. Tokyo: Toyo Keizai Shinpo-sha.

Yasui, T. 1940. Juyo no hosoku nitsuite [On the law of demand]. *Keizaigaku Ronshu* 10, 651–66.

How to cite this article

Nishizawa, Tamotsu and Aiko Ikeo. "Japan, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000051> doi:10.1057/9780230226203.0860

The New Palgrave Dictionary of Economics Online

Jaszi, George (1915 – 1992)

E. Denison

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

government accounting; Jaszi, G.; national income accounting; national product accounting

Article

Research administrator and expert in national accounts. Born in Budapest, Jaszi attended Oberlin College, the London School of Economics (BSc, 1936) and Harvard (Ph.D., 1946). He was employed by the Bureau of Economic analysis (or its predecessors) in its National Income Division 1942–59 (Division Chief 1949–59), as Assistant Director 1959–62, and as Director 1963–85.

Jaszi helped develop the US national income and product accounts that were introduced gradually during the Second World War and fully in 1947. The accounts for the government sector were his unique contribution. One aspect of this is that all government purchases, like other purchases not for resale, are counted as final products. For four decades Jaszi influenced the United Nations standardized system in addition to guiding the United States accounts. His firm grasp of the national income and product accounts as an integrated system and of the principle that they must rest upon quantifiable concepts made him particularly skilful in explaining and vindicating the 1947 system and its subsequent improvement. Jaszi's (1958) exposition of the accounts and responses to critics at a 1955 conference and his (1971) critique of comments by 46 economists were masterful. Elsewhere (1964), Jaszi showed that hedonic and conventional methods of allowing for quality change in output measurement are conceptually equivalent.

During 1963–85 Jaszi directed all the varied statistics and analyses of the Bureau of Economic Analysis – international, national and regional. Many improvements were introduced during his tenure. He closely supervised BEA's *Survey of Current Business* and co-authored its ‘Business Situation’ section. His talks (for example, Jaszi, 1972) helped balance exaggerated claims of (*a*) damage introduced into policy formulation by errors of estimate in the NIPA and (*b*) the possibility of greatly reducing such errors.

See Also

- national accounting, history of

Selected works

1943. (With M. Gilbert.) National income and national product in 1942. *Survey of Current Business*, March, 10–26.

1947. (With others in National Income Division.) National income and product statistics of the United States 1929–46. *Survey of Current Business*, July, Supplement.

1948. (With M. Gilbert, E. Denison and C. Schwartz.) Objectives of national income measurement: a reply to Professor Kuznets. *Review of Economics and Statistics* 30(August), 179–95.

1956. Statistical foundations of gross national product. *Review of Economics and Statistics* 38, 205–14.

1958. The conceptual basis of the accounts: a reexamination. Reply and comments. *Studies in Income and Wealth*. Princeton: Princeton University Press, vol. 22, 13–127, 140–45, 209–27, 300–22, 363–71, 402, 454–7, 521–35, 579–82.

1964. Comment. *Studies in Income and Wealth*, Princeton: Princeton University Press, vol. 28, 404–9.

1971. Review: an economic accountants' ledger. *Survey of Current Business*, July, Pt II, 183–227.

1972. Taking care of soft figures: reflections on improving the accuracy of the GNP. In University of Michigan Economics Department, *The Economic Outlook for 1972*. Ann Arbor: Research Seminar in Quantitative Economics.

1981. (With C. Carson.) National income and product accounts of the United States: an overview. *Survey of Current Business*, February, 22–34.

1986. An economic accountant's audit. *American Economic Review* 76, 411–17.

Bibliography

Duncan, J. and Shelton, W. 1978. *Revolution in United States Government Statistics 1922–76*, ch. 3. US Department of Commerce, Office of Federal Statistical Policy and Standards, Washington, DC: US Government Printing Office.

How to cite this article

Denison, E. "Jaszi, George (1915–1992)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000006> doi:10.1057/9780230226203.0861

The New Palgrave Dictionary of Economics Online

J-curve

Giorgia Giovannetti

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The J curve is the description of an empirical phenomenon: the trade balance worsens immediately after a depreciation of the exchange rate, to improve in the longer term. This pattern can be ascribed to different speed of adjustment of trade prices and volumes to changes in exchange rates. Several models have been put forward, suggesting explanations for these lags that are not mutually exclusive. While the empirical evidence is not conclusive, to assess the existence of a J-curve adjustment path is relavant since the J-curve may induce dynamic instability in the exchange rates.

Keywords

balance of trade; common pool problem; consumption habits; currency depreciation; hysteresis; J-curve; Marshall–Lerner conditions; real business cycles; structural vector autoregressions; sunk costs; technology shocks; terms of trade

Article

A depreciation of the domestic exchange rate is expected to improve a country's trade balance. It has been observed, however, that in reality the trade balance often worsens immediately after depreciation. Only in the longer term, if at all, does it improve. The combination of a negative short-run effect together with a subsequent positive long-run effect of a devaluation of the exchange rate on the trade balance is referred to as the ‘J-curve’ due to the similarity of the current-account behaviour to a ‘J’.
To assess the existence of a J-curve adjustment path is relevant, since, under most circumstances, the J-curve may induce dynamic instability in exchange rates (see Beenstock, 1990; Levin, 1985). If depreciation worsens the trade balance in the short run, the exchange rate may fall further (depreciate more). Although export volumes may rise and import volumes fall, import values tend to grow faster than export values and the exchange rate instability persists. This instability may be neutralized by speculators having rational expectations. In this case, agents know the dynamics of the J-curve and allow

for it in their speculative behaviour, thereby eliminating the potentially destabilizing influence of the J-curve itself.

The J-curve is the description of an empirical phenomenon, first discussed after the 1967 devaluation of the pound sterling in NIESR (1968) and analysed in a seminal paper by Magee (1973). Theoretical models have been developed, building on some kind of frictions, such as pre-existing contracts, asymmetric use of domestic currency and foreign (international) currency, sluggishness in adding new productive capacity and sunk costs.

The lags in the adjustment can be ascribed to trade prices adjusting faster than trade volumes to changes in the exchange rate. The currency in which imports and exports are denominated, which is likely to be determined by the relative market power of traders, plays a crucial role. When both import and export contracts are expressed in domestic currency, following an unexpected devaluation, the value of existing imports rises due to the increased cost of an unchanged quantity of imports, while the value of existing exports remains constant (the price of exports in domestic currency does not change). The existence of lags on consumers' and producers' side induces stickiness and a worsening of the trade balance, until higher export and lower import volumes eventually, and on the assumption that import and export demand elasticities are sufficiently elastic to exchange rate changes so that their sum is higher than 1 (that is, the Marshall–Lerner conditions hold), generate a favourable trade balance response (that is, a J-shaped path). When both import and export contracts are expressed in foreign currency, if the value of contracts denominated in foreign currency is higher for imports than for exports, the J-curve will always ensue. On the other hand, the trade balance would improve immediately after devaluation if import contracts are denominated in domestic currency and export contracts in foreign currency.

Quantities may not adjust rapidly to exchange rate changes, since domestic demand for imported goods may be fairly inelastic due, for instance, to a reputation or brand and/or domestic supply may be a poor substitute for imported goods. Furthermore, producers may not be able to reallocate expenditure between foreign and domestic goods since most import and export orders are placed in advance (before depreciation). In the long run, however, quantities tend to adjust and, if the value of elasticity satisfies Marshall–Lerner, the trade balance improves.

Several models have suggested not mutually exclusive reasons for a short-run deterioration of the trade balance following depreciation. Knetter (1993, p. 473) maintains that: 'sellers reduce markups to buyers whose currencies have depreciated against the seller, thereby stabilising prices in the buyer's currency relative to a constant markup policy', that is, follow a 'local currency price stability' and differentiate between markets (pricing to market). Some models rely on the small open economy hypothesis and emphasize intertemporal substitution. Bacchetta and Gerlach (1994) challenge the view of a *rapid* pass-through of exchange rates to import prices, showing that the J-curve can arise even if import prices are sticky. In an intertemporal framework, an anticipated rise in future import prices after depreciation provides agents with an incentive to decrease their current expenditure and therefore revise their future purchases, eventually displaying a J-shape dynamics of trade balance.

In a continuous time open-economy framework, Mansoorian (1998) shows that a trade balance deterioration following depreciation can be due to persistence of consumption habit, on the assumption that the utility function depends not only on current consumption but also on standard of living. Similar conclusions have been recently reached by Cardi (2005) who derives a J-curve due to the 'strength' of consumption habits and capital investment 'inertia', following an unanticipated terms of trade deterioration.

Tivig (1996) solves an intertemporal maximization problem in a dynamic oligopoly contest. In a two-period model of duopolistic competition without entry, regardless of the degree of capital mobility, she provides sufficient (at least) conditions on import demand elasticity for short-run ‘perverse’ price reactions, following temporary changes in the exchange rates, so that a temporary devaluation may initially worsen and later improve the trade balance.

Tornell and Lane (1998) use political economy considerations to explain the J-curve pattern; their model with ‘voracity effects’ suggests that a positive real trade shock may exacerbate the *common pool problem*, leading to a more than proportional increase in public transfers and then to a social loss. Because of this effect, the impact of an unexpected improvement in the terms of trade may lead to deterioration in the current account.

Recent literature in dynamic general equilibrium depicts an S-curve as a dynamic response of the trade balance to technology shocks. Backus, Kehoe and Kydland (1994) find that the trade balance is negatively correlated with current and future movements in terms of trade, but positively correlated with past movements. Using a two-country version of Kydland and Prescott's (1982) closed economy model, in which each country produces imperfectly substitute goods with capital and labour, they claim that, after a once and for all positive shock to technology, domestic output increases, its relative price falls and domestic investment increases, inducing a fall in net exports. With time, the rise in investments decreases and the trade balance moves into a surplus. This dynamic gives rise to an S-curve consistent with the J-curve, since the initial deterioration and subsequent improvement of the trade balance may well deliver also an S-shaped cross-correlation function of the trade balance and the terms of trade.

Senhadji (1998) extends the Backus, Kehoe and Kydland analysis to document business cycle features of several developing countries. Within a set-up with a downward sloping export demand function and limited access to international financial markets for capital formation, he shows results completely supportive of the findings of Backus, Kehoe and Kydland.

Finally, the J-curve can be due to hysteresis (cf., for instance, Baldwin, 1988; Dixit, 1994). In the presence of sunk costs, to export is an ‘option’ and consumers value the alternative of ‘wait and see’ before reacting to exchange rate changes. The presence of a threshold induces non-standard behaviour of the trade flows when the exchange rate depreciates.

Over the years, an extensive empirical literature has emerged. Results are not conclusive. As stated by Bahmani-Oskooee and Artatrana (2004, p. 1389), ‘the general consensus is that the short run response of the trade balance to current depreciation does not follow a specific pattern’ but, if a J-curve exists, the perverse effect has a duration of one to three years (see Junz and Rhomberg, 1973; Baldwin and Krugman, 1987; Spitaller, 1980; Moffett, 1989). Koray and McMillin (1999) use a structural VAR model to show that the pattern of the trade balance after a negative monetary shock exhibits the traditional J-curve behaviour. Support for the J-curve hypothesis has been found also by Bahmani-Oskooee and Alse (1994), Marwah and Klein (1996) and Hacker and Abdulnasser Hatemi (2003). Leonard and Stockman (2002) find a weak positive evidence of the J-curve, but document strong violations in the distributional assumptions that underline previous works. More interestingly, their evidence on the J-curve is inconsistent with traditional theoretical explanation (real business cycle models included), since they present evidence that current account surpluses are usually associated with low real GDP.

Other studies have challenged the existence of a J-curve: Rose and Yellen (1989) and Rose (1990; 1991) maintain that, if import prices adjust slowly to exchange rate changes, the initial negative effect embodied in the J-curve may not occur: the value of imports does not increase and, *ceteris paribus*, the trade balance does not worsen. More recently, Demeulemeester and Rochat (1995), Hsing and Savvides (1996), Shirvani and Wibratte (1997) using different methodology, different data-set and estimating over different periods, found no evidence of a J-curve in the data.

See Also

- elasticities approach to the balance of payments
- exchange rate dynamics
- international finance

Bibliography

Artus, J.R. 1975. The 1967 devaluation of the pound sterling. *IMF Staff Papers* 22, 595–640.

Bacchetta, P. and Gerlach, S. 1994. Sticky import prices and J-curves. *Economics Letters* 44, 281–5.

Backus, D.K.,Kehoe, P. and Kydland, F. 1994. Dynamics of the trade balance and the terms of trade: the J-curve? *American Economic Review* 84, 84–103.

Bahmani-Oskooee, M. and Alse, J. 1994. Short-run versus long-run effects of devaluation: error correction modeling and cointegration. *Eastern Economic Journal* 20, 453–64.

Bahmani-Oskooee, M. and Brooks, T.J. 1999. Bilateral J-curve between U.S. and her trading partners. *Weltwirtschaftliches Archiv* 135, 156–65.

Bahmani-Oskooee, M. and Artatrana, R. 2004. The J-curve: a literature review. *Applied Economics* 36, 1377–98.

Baldwin, R. 1988. Hysteresis in import prices: the beachhead effect. *American Economic Review* 78, 773–85.

Baldwin, R. and Krugman, P. 1987. Persistence of the U.S. trade deficit. *Brookings Paper Economic Activity* 1987(1), 1–43.

Beenstock, M. 1990. Exchange rate dynamics. In *Current Issues in international Monetary Economics*, ed. D. Llewellyn and C. Milner. London: Macmillan.

Cardi, O. 2005. Another view of the J-curve. Working Paper No. 2005–29, Département des Sciences Economiques, Université catholique de Louvain.

Demeulemeester, J.L. and Rochat, D. 1995. An empirical investigation of the relationship between the exchange value of the pound and the UK trade balance. *Economia Internazionale* 48, 347–58.

Dixit, A. 1994. Hysteresis and the duration of the J curve. *Japan and the World Economy* 6, 105–15.

Hacker, R.S. and Abdulnasser Hatemi, J. 2003. How productivity and domestic output are related to exports and foreign output in the case of Sweden. *Empirical Economics* 28, 767–82.

Hsing, H.M. and Savvides, A. 1996. Does a J-curve exist for Korea and Taiwan? *Open Economies Review* 7, 126–45.

Junz, H.B. and Rhomberg, R.R. 1973. Price competitiveness in export trade among industrial countries. *American Economic Review* 63, 412–8.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–71.

Knetter, M.M. 1993. International comparisons of pricing-to-market behavior. *American Economic Review* 83, 473–86.

Koray, F. and McMillin, W.D. 1999. Monetary shocks, the exchange rate and the trade balance. *Journal of International Money and Finance* 19, 925–40.

Krugman, P. 1989. *Exchange Rate Instability*. Cambridge, MA: MIT Press.

Leonard, G. and Stockman, A.C. 2002. Current accounts and exchange rates: a new look at the evidence. *Review of International Economics* 10, 483–96.

Levin, J. 1985. Does leaning against the wind improve exchange rate performance. *Journal of International Money and Finance* 4, 135–50.

Magee, S. 1973. Currency contracts, pass-through, and devaluation. *Brookings Paper on Economic Activity* 1973(1), 303–25.

Mann, C.L. 1986. Prices, profit margins, and exchange rates. *Federal Reserve Bulletin* 72, 356–79.

Mansoorian, A. 1998. Habits and durability in consumption, and the dynamics of the current account. *Journal of International Economics* 44, 69–82.

Marwah, K. and Klein, L.R. 1996. Estimation of J-curves: United States and Canada. *Canadian Journal of Economics* 29, 523–39.

Moffett, M. 1989. The J curve revisited: an empirical examination for the United States. *Journal of International Money and Finance* 8, 425–44.

NIESR (National Institute of Economic and Social Research). 1968. The economic situation. *The Home Economy: National Institute Economic Review* 44, 4–17.

Rose, A.K. 1990. Exchange rates and trade balance: some evidence from developing countries. *Economic Letters* 34, 271–75.

Rose, A.K. 1991. The role of exchange rates in a popular model of international trade: does the Marshall–Lerner condition hold? *Journal of International Economics* 30, 301–16.

Rose, A.K. and Yellen, J.L. 1989. Is there a J-curve? *Journal of Monetary Economics* 24, 53–68.

Senhadji, A. 1998. Time-series estimation of structural import demand equations: a cross-country analysis. *IMF Staff Papers* 45, 236–68.

Shirvani, H. and Wibratte, B. 1997. The relationship between the real exchange rate and the trade balance: an empirical reassessment. *International Economic Journal* 11, 39–50.

Spitaller, E. 1980. Short run effects of exchange rate changes on terms of trade and trade balance. *IMF Staff Papers* 27, 320–48.

Tivig, T. 1996. Exchange rate pass-through in two-period duopoly. *International Journal of Industrial Organization* 14, 631–45.

Tornell, A. and Lane, P.R. 1998. Are windfalls a curse? A non-representative agent model of the current account. *Journal of International Economics* 44, 83–112.

How to cite this article

Giovannetti, Giorgia. "J-curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000007> doi:10.1057/9780230226203.0862

The New Palgrave Dictionary of Economics Online

Jenkin, Henry Charles Fleeming (1833 – 1885)

R.D. Collison Black

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

consumer surplus; Dupuit, A.-J.-E. J.; guaranteed wage; Jenkin, H. C. F.; Jevons, W. S.; supply and demand; tax incidence; trade unions; value; wages fund

Article

Jenkin was a distinguished engineer whose wide interests and clarity of mind enabled him to make notable contributions to economic analysis.

Jenkin received his early education at Edinburgh Academy, but financial exigencies forced the family to move first to France and then to Italy. Consequently, Jenkin graduated from the University of Genoa in 1850. Returning to England in 1851, he spent ten years with various engineering firms working on the design and laying of submarine cables. In 1859 he became associated with William Thomson (Lord Kelvin) and frequently collaborated with him in later years, especially in contributing to the work of the British Association's Committee on Electrical Standards. In 1866 Jenkin was appointed Professor of Engineering at University College, London, and moved to a similar chair at the University of Edinburgh in 1868. Apart from his work in civil and electrical engineering, Jenkin distinguished himself as a critic of Darwin's theory of evolution, as an advocate of improved urban sanitation, and for the development of the system of monorail electric transport called telpherage.

Between 1868 and 1872 Jenkin published three economic papers whose theoretical quality and practical value have since earned him a deserved place in the history of economic thought. Recognizing that in current debates on trade unions ‘the principles of political economy though often quoted are little understood’, Jenkin set himself in his first paper to examine their application to the labour market. In the process he revealed the emptiness of the wages-fund concept, refuted the view that trade unions could not materially benefit their members and made the first clear statement in English economic writing of the concept of supply and demand as functions of price. These ideas he further developed and generalized in his 1870 paper, in which he analysed fully the determination of market price using diagrams to present the supply and demand functions in the form of intersecting curves. Jenkin specifically noted that in the long-run cost of production chiefly determines the price of manufactured

goods, but stressed ‘how much the value of all things depends on simple mental phenonema, and not on laws having mere quantity of materials for their subject’ (1887, vol. 2, p. 93).

In a third paper Jenkin applied his techniques of supply and demand analysis to the problem of tax incidence, stating the concept of consumers’ surplus previously developed by Dupuit but apparently without knowledge of Dupuit's work.

Jenkin left two further essays on economic issues, which were published posthumously in his collected *Papers, Literary, Scientific, &c*. In ‘Is one Man's Gain another Man's Loss?’ (1884) he used a simple form of closed circuit diagram to illustrate the exchange process and its results. ‘The Time–Labour System’ contained an acute diagnosis of the differences between goods markets and labour markets with a proposal to improve the operation of the latter through what was in effect a system of guaranteed annual wages.

All Jenkin's economic writings were characterized by a striking combination of precise and lucid analysis with tolerant understanding of the facts of daily life in both the workshops and the counting-houses of the world he knew. In view of this their influence in his own time was surprisingly limited, although his ‘Graphic Representation’ (1870) does seem to have afforded the stimulus which led W.S. Jevons to publish his *Theory of Political Economy* in 1871.

Selected works

1868. Trade-unions: how far legitimate? *North British Review*, March.

1870. The graphic representation of the laws of supply and demand, and their application to labour. In *Recess Studies*, ed. Alexander Grant. Edinburgh.

1872. On the principles which regulate the incidence of taxes. *Proceedings of the Royal Society of Edinburgh 1871–2*.

1879–81. The time-labour system. Unpublished ms.

1884. Is one man's gain another man's loss? Unpublished ms.

All published in *Papers, Literary, Scientific, &c*, ed. S.C. Colvin and J.A. Ewing. London: Longmans, Green & Co., 2 vols, 1887. Reprinted as No. 9 in *London School of Economics Series of Reprints of Scarce Tracts in Economics and Political Science*, 1931.

Bibliography

Brownlie, A.D. and Lloyd Prichard, M.F. 1963. Professor Fleeming Jenkin, 1833–1885, pioneer in engineering and political economy. *Oxford Economic Papers*, NS 15, 204–16.

How to cite this article

Black, R. D. Collison. "Jenkin, Henry Charles Fleeming (1833–1885)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000009> doi:10.1057/9780230226203.0863

The New Palgrave Dictionary of Economics Online

Jevons, William Stanley (1835 – 1882)

Harro Maas
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article examines William Stanley Jevons's life and work against the background of Victorian disputes over the appropriate method of political economy. Jevons is commonly known as one of the founders of marginalist analysis in economics. As a genuine Victorian polymath, Jevons undertook research in many different fields of the sciences, meteorology, statistics and political economy in particular. This article shows how Jevons transposed his training in the natural sciences to political economy, in the process shifting from a labour to a utility theory of value and mathematizing the discipline as well.

Keywords

Bentham, J.; Boole, G.; Cairnes, J. E.; D'Avenant, C.; De Morgan, A.; equation of exchange; felicific calculus; functional analysis; hedonic theory of value; index numbers; induction; Jenkin, H. C. F.; Jevons, W. S.; King, G.; labour supply; labour theory of value; marginal revolution; mathematical methods in political economy; Mill, J. S.; political economy; productive and unproductive labour; Psychophysiology; quantity theory of money; Quetelet, A.; statistics, and economics; tendency laws; Tooke, T.; utility; value; wages fund; Whewell, W.

Article

Stanley Jevons is generally known as one of the 'fathers' of the so-called marginal revolution in economics of the last decades of the 19th century. In his *Theory of Political Economy* (1871), with its 'mechanics of utility and self-interest', he analysed decisions of economic agents by means of the calculus, in terms of deliberations over marginal increments of utility. Economic agents – whether in their role as consumers, workmen or other – came to be seen as maximizing utility functions. Jevons is thus commonly considered to have broken with the labour theory of value of the classical economists. Value came to be identified with exchange value, and Jevons identified this with what we now call marginal utilities, not with costs of production. Jevons is also remembered for his innovative contributions to the empirical (statistical) study of the economy. He much favoured the use of graphs to picture and analyse statistical data. He introduced index numbers to make causal inferences about economic phenomena such as changes in the value of gold following the gold discoveries in California and Australia. In short, there is no particle of economics, theoretical or empirical, to which Jevons did not make important contributions that even in the 21st century are considered to have altered the field of economics in revolutionary fashion.

Though this short summary of Jevons's accomplishments makes him one of the fathers of modern economics, Jevons was in many regards heavily indebted to Victorian ways of practising the natural sciences. This transpires from his commitment to a Baconian view of the natural and social sciences and the typical Victorian use of mechanical analogies to understand the world. Just as William Thomson, the later Lord Kelvin, argued that the making of a mechanical model was the ultimate 'test' of intelligibility of a natural object, so Jevons relied on mechanical models to understand the social world. For Jevons, human individuals were little machines driven by pleasures and pains; they were not rational and autonomous decision-makers. In his influential essay on the nature of political economy (first edition 1932) Lord Robbins perceived clearly that Jevons's analysis was not just rooted in Bentham's hedonics, but in psycho-physiological explanations of human conduct. Jevons's new theory of value was not a theory of rational choice. The common perception of Jevons's utility theory as a theory about individual *decision*-making is therefore not quite accurate.

This article, in following Jevons's life and work, aims to show how Jevons translated his knowledge of methods of research in the natural sciences to political economy and thus to

radically transform the scientific image of economics. The radicalism of Jevons's new approach to political economy will be better understood against the background of Victorian discussions of its proper method.

Early years

Stanley Jevons was born in Liverpool as the ninth child in a well-to-do Unitarian family. In 1882 he drowned near Hastings at the age of only 47, leaving his wife Harriet Taylor and three children. Good biographical essays are Black and Könekamp's introduction to the papers and correspondence (1972–81), Schabas's 1990 monograph, and Mosselman and White's introduction to their recent edition of Jevons's major works. The liveliest introduction to his biography is still to be gathered from his own *Letters and Journals* (1886), chosen from his *Nachlass* by his wife Harriet.

Jevons's father, Thomas Jevons, was an iron merchant with utilitarian sympathies who is said to have invented the first floating iron ship. His mother, Mary-Ann Roscoe, was the daughter of William Roscoe, a Liverpool banker and important collector of Flemish and Italian masters. Like the Booth, Hutton and Martineau families to whom they were related, the family formed part of the self-confident Unitarian circles in the heartland of the industrial revolution that shared a belief in rational argument and the advancement of science to promote the common good. It is therefore unsurprising that Jevons's early education was predominantly in the natural sciences, first at Liverpool Mechanics' High School, and, after an interlude at a grammar school that was less to his liking, at the preparatory school of University College London. In 1851 he enrolled at University College itself to study mathematics and chemistry.

Jevons's youth was not without difficulty. His mother died in 1845, his much beloved eldest brother, Roscoe, went insane in 1847; on top of this, his father's iron business went bankrupt in 1848 due to the great railway crisis the previous year. Forced to move to Manchester, the Jevons family never recovered from these financial difficulties, which lifted Stanley and his siblings from the commercial elite to what has been called the 'uneasy classes' – intellectually gifted, but without the means to leisurely pursue their interests. As we will see, the family's financial difficulties greatly influenced Jevons's early intellectual career.

At University College, Jevons enrolled in courses in experimental philosophy and chemistry, and the mathematics class of Augustus De Morgan, the first mathematics professor at University College who taught the by far most demanding course in mathematics in England at the time. De Morgan was a great propagator of French analysis, a mathematics that by 1850 was still received with considerable scepticism in the Oxbridge system because of the strong mechanical worldview that went with it. De Morgan would prove to be one of the most enduring influences on Jevons's intellectual life. Though Jevons performed well, he never considered himself a mathematician (and was not so considered in his lifetime). Jevons's forte was in chemistry and the experimental sciences.

During these first years of study Jevons won several medals, a gold medal in chemistry amongst others. In 1853, through the intervention of his cousin Harry Roscoe, the later professor of chemistry at Owens College, Manchester, Jevons was offered the opportunity to become gold assayer at the newly established Mint in Sydney. After some hesitation (and persuasion by his father) Jevons accepted the offer, because the job paid extremely well (£675 a year), and thus helped to alleviate the financial burdens on the family. Jevons sailed off to Australia in 1854 to stay there for a five-year period.

Jevons's Antipodean interlude

There has been quite some discussion about the importance of Jevons's 'Antipodean interlude' for his further intellectual career (see the relevant essays in Wood, 1988). Not only did the work at the Sydney Mint offer Jevons ample opportunities to pursue his manifold scientific interests, but the social environment of the Mint itself was highly favourable to the pursuit of science. The newly created philosophic society of New South Wales, of which Captain E.W. Ward, director of the Mint, was office bearer, provided Jevons and his colleagues ample opportunities to develop their scientific interests and to publish on them. As a typical Victorian colonial institute the Mint thus functioned as a nucleus of scientific activity that turned its 'imperial gaze' upon Australian nature and society.

The most important Australian science periodical was the *Sydney Magazine of Arts and Sciences*, to which Jevons made several contributions, most of them on meteorology. Jevons published on experiments on the formation of clouds, in which he attempted to reproduce clouds on a miniature scale in accordance with the existing classification of clouds. He made these experiments on strong mechanical assumptions and in the hope of rendering his results in mathematical form – something that proved far too difficult. His aim was to mimic the process of cloud formation in another medium (fluid rather than air) and so to uncover its underlying mechanism. It is worth noting that Jevons's experiments did not go completely unnoticed: Lord Rayleigh, the later Nobel laureate in physics, reproduced Jevons's experiments in the early 1880s at the Cavendish laboratory at Cambridge in order to study diffusion processes in fluids and gases. Jevons also contributed to Waugh's statistical almanac, in which he extensively reported on his statistical observations on the Australian weather. Jevons's cloud experiments and his work in meteorology are best covered by Raymond Schmitt (1995) and Neville Nicholls (1998).

In the 1830s Lancashire Unitarians had been instrumental in setting up statistical societies to study the 'moral and physical condition of the working classes'. In a similar vein Jevons had wandered through the poor working men's districts of London – the 'dark alleyways of Spitalfields' – to study the moral condition of the working poor during his early years of study in London. In Australia Jevons resumed these wanderings and started work on a social survey of Sydney. He published on his findings in the *Sydney Morning Herald*. Though

only fragments of the original survey remain, it is clear from his notebooks that Jevons considered his survey the beginning of a ‘science of towns’ that itself was a prelude to a general ‘science of man’. A very good recent study of Jevons's survey (Davison, 1997–8) has shown interesting parallels with the much better-known work of Henry Mayhew on the London poor and Charles Booth's famous late 19th-century social survey of London. Jevons's social survey may serve as early witness of how he transported his natural and acquired skills in decomposition and classification of natural phenomena to the social sciences, and how he translated visualizing techniques used in the natural sciences to the social domain. Another fine example, from the same Sydney period, of Jevons's use of visualizing techniques to classify social phenomena, recently explored by White (2006), is his stratigraph of the ‘industrial system of society’. Jevons classified the various occupations of Australian society in different layers corresponding to a system of human needs and to the classical distinction between productive and unproductive labour. This kind of diagram had only recently come to be used by geologists to picture the composition of the underground. It is noteworthy that the subtitle of Jevons's last great work (or rather project), the unfinished and posthumously published *Principles of Political Economy*, shows his lifelong concern with the ‘industrial mechanism of society’.

Jevons not only explored the urban wilderness of Sydney but also made excursions into New South Wales, for example to the newly discovered gold mines. Sometimes he made these trips alone, because others found them too dangerous. The barometer and thermometer always accompanied him, and his notebooks are filled with pages of meteorological observations made during these trips. Apart from his extensive and innovative use of visualizing techniques, his appetite for the experimental method in the sciences, and his sometimes daring and innovative collection and analysis of statistical facts, Jevons also pioneered in photography, which was facilitated by his knowledge of chemistry and the ease with which he could lay hands on chemical materials equally needed in photography and gold-assaying. In 2004 the Sydney Powerhouse Museum organized an exhibition on Jevons's life and work in Australia that wonderfully brought all these different influences and materials together, thus showing vividly the rich background and context of Jevons's scientific work. For a brief description of this exhibition see Barrett and Connell (2005).

Historians of economics have commonly taken Jevons's enthusiasm for a lecture on Bentham's utilitarian ethics as a decisive moment in his turn to the social sciences and to political economy in particular. Jevons's appraisal of this lecture is seen as a premonition of the hedonic theory of value that was to become the core of his *Theory of Political Economy* (1871). In addition to this, reference is often made to the interest Jevons took in the Australian railway debate (to which he made some contributions in the *Empire*, an Australian newspaper) and his interest in Lardner's *Railway Economy*, a book that was influenced by the work of Cournot on oligopolies. From the above it should be clear that Jevons's studies in social statistics and political economy were wide in scope. From a very young age he aimed at all-embracing explanations of society that were deeply rooted in a thorough engagement with statistics and fuelled by a predilection for mechanistic explanations.

This short summary of Jevons's scientific pursuits in Australia shows him to be the kind of Victorian that felt more indebted to the Belgian astronomer and revolutionary of statistics Adolphe Quetelet than to John Stuart Mill or August Comte. In a short essay on Comte for *Nature* (1875), Jevons explicitly paid tribute to Quetelet as the ‘true founder’ of the social sciences, because of his endeavours to discover, like an astronomer, regularities in the avalanche of social statistics. According to Jevons, by focusing on average values, rather than on particular data, the mechanisms could be uncovered that governed the natural and the social world. Mathematics, when properly targeted, spelled out this mechanism. When Jevons wrote from Australia to his sister Henrietta that he considered devoting his life to political economy it was not so much an application of Bentham's hedonic calculus to economics he had in mind, but rather his aim to explain the ‘industrial mechanism’ of society. The consequences of Jevons's attitude to science for political economy would become spelled out only after his return to London to take up his studies at University College, London.

The mid-century split between theory and statistics

Jevons returned to England in 1859. He once again enrolled at University College, now to study political economy, but he quickly became disappointed in the way the topic was taught by Jacob Waley, whom he considered ‘prejudiced’ against opinions and ideas that went contrary to the Mill–Ricardian orthodoxy. As in the early 1850s, Jevons most enjoyed his mathematics classes with Augustus De Morgan. Jevons completed his BA in 1861 and received his MA in 1863.

Jevons's disappointment was not just a matter of (emerging) diverging insights about one of the cornerstones of classical economics, the theory of value; it was also disappointment with its methods of research. Mid-century political economy was characterized by a sharp split between economic theory and statistics. In his writings on the definition and method of political economy, J.S. Mill (and in his footsteps John Elliot Cairnes) had ardently defended the so-called deductive or a priori method of political economy in opposition to the inductive (statistical) method. Mill had done so in his seminal essay of 1836 ‘On the Definition of Political Economy; and on the Method of Investigation Proper To It’ that was reprinted in his *Essays on Some Unsettled Questions of Political Economy* (1844) and in the famous Book VI of *Logic* (1843) that was devoted to the method of the ‘moral sciences’. Mill wrote his essay in a deliberate defence of Ricardian economics, which in the early 1830s was actually on the wane. Mill was adamant there was nothing wrong with Ricardian economics. Indeed, what was considered its most distinguishing ‘vice’ – its deductivism – was just the way the science should proceed. Mill's argument for this was quite innovative and had little to do with Ricardian theory per se. Leaning on the philosopher of mind Dugald Stewart, Mill sharply distinguished between two different fields of science, the natural and the mental (or ‘moral’, as was used somewhat equivalently at the time). Political economy was interested only in a limited set of mental motives on which observations could be made within the ‘private laboratory’ of one's own mind. Therefore, political economy was a science of *tendency laws* the consequences of which could be deduced with the same certainty as the laws of physics. Mill repeatedly emphasized that what we nowadays call ‘introspective’ observations on mental states were as good as the methods of ‘observation

and experiment' used in the natural sciences. Mill's defence of (Ricardian) political economy proved extremely successful. After the publication of his own *Principles of Political Economy* (1848) debates on the proper method of political economy certainly shifted in favour of Mill.

Following Mill, the Irish political economist John Elliot Cairnes argued in his lectures on the method of political economy (1857) that a political economist did not need the 'tedious route of induction' practised by the statistical societies. Political economy was an a priori science and the 'business' of the political economist was finished once he had traced an event back to a mental motive. More explicitly than Mill, Cairnes argued that political economy lacked the exactness of the natural sciences, because principles of the mind, by their nature, were not the kind of material fit for measurement and hence quantification. Cairnes's lectures reflect the 'curious separation between abstract theory and empirical work' in which the work of political economists and statisticians were worlds apart (Blaug, 1976), much in contrast with Jevons's own practice in Australia.

Much of Jevons's mature work in economics and statistics can be seen as a deliberate transgression of this division between (inductive) statistics and (deductive) theory. In his 1870 opening address as president of Section F of the British Association for the Advancement of Science (BAAS), Jevons explicitly argued that a 'scientific treatment' of social facts consisted in inductive and deductive processes, just as in the 'other branches of the sciences'. Jevons used his mechanical world view to cut through the distinction between theory and statistics and in the process mathematized the discipline.

Jevons's publications in the 1860s

In the early 1860s Jevons worked in frenzy on a variety of subjects. He devoted much of his time to the development of an alternative to George Boole's algebraic logic, on which he published a small tract in 1863 that received little attention. He worked on what he called his 'Statistical Atlas' project, a large-scale project clearly started in the spirit of William Playfair's *Commercial and Political Atlas* (1801), which was the first application of the graphical method to social statistics. He presented its outline and first plates to William Newmarch, then one of the leading statisticians, but Newmarch hardly paid attention to it, and Jevons ended up publishing two of his plates at his own expense. He wrote several entries for Watts's *Dictionary of Chemistry and of the Allied Branches of the Other Sciences* (1863–8) on various measuring instruments, such as the balance and the thermometer, but also on topics such as cloud formation. He wrote a short outline of a mathematical theory of political economy, read to the British Association for the Advancement of Science in 1862 and published 1866 as the *Brief Account of a Mathematical Theory of Political Economy*. It was as poorly noticed as his work in formal logic, and it is understandable that Jevons's diary at the end of 1862 showed some frustration about his worldly accomplishments. The best essay on Jevons's first brief mathematical outline of economics is by Grattan-Guinness (2002).

Jevons's first success came with his study of the fall of the value of gold, published in 1863, and republished in his *Investigations in Currency and Finance* (1884). Apart from giving an imaginative survey of the various causes of price fluctuations, it is noteworthy for the use of index numbers to assess the change in the value of gold. The question was whether the value of gold had changed as a result of the gold discoveries in California and Australia. The study can be seen as an application of a version of the quantity theory of money, but, more interestingly, Jevons compared the quantity theory equation of exchange with a mechanical balance and asked what was the more probable: that a tip of the balance had come from a variety of causes influencing individual prices on the one side of the balance, or from just one cause on the other side, that is, an increase in gold bullion. On this analogical reasoning Jevons constructed an unweighted index number, arguing that the geometric mean gave the best approximation for the 'true' fall in value (and so a rise in the general price level). Whatever its technical limitations (many of which Jevons acknowledged himself), the study was a genuine accomplishment and it was immediately recognized as such. It was well received by Cairnes, who wrote approvingly that he had come to similar conclusions though using the a priori method of research; for Cairnes, statistical data did not play any formative role in his argument.

The favourable reception of Jevons's essay on the value in the fall of gold of 1863 was even surpassed by that of *The Coal Question* in 1865. This soon made Jevons's professional frustrations vanish. Already elected a member of the London Statistical Society in November 1864, by 1865 Jevons could notice in his diary that he was considered 'by reviews of authority, a *competent* statistician'. *The Coal Question* was Jevons's definitive breakthrough. Using extensive statistical resources, the book addressed the question of England's wealth in the face of the inevitably rising costs of coal extraction. It is still excellent reading for those interested in environmental economics. It led J.S. Mill to ask questions in Parliament and Gladstone to invite the author to Downing Street (and to argue for a balanced budget). *The Coal Question* was certainly instrumental to Jevons's appointment as professor of logic and political economy at Owens College, Manchester in 1865. For someone who had just reached the age of 30, this can hardly be seen as an unfavourable professional record.

In the second half of the 1860s Jevons continued to publish in statistics and formal logic, on the development of his logical machine in particular. He presented his own formal logical system in *The Substitution of Similars* (1869) and his proposal for a mechanical representation of this system in *The Mechanical Performance of Logical Inference* (1870), both reprinted in *Pure Logic and Other Minor Works* (1890). He also published statistical investigations on seasonal variations, using ratio charts to thresh out seasonal patterns. At the end of the 1860s, Jevons was predominantly known for his work in formal logic and statistics, not as a theoretician, and it may be for that reason that political economists such as Mill and Cairnes were somewhat puzzled by the *Theory*. Given their views on the method of political economy, they substantially disagreed with Jevons's transgression of the dividing line between the natural and moral sciences, and with his concomitant use of formal methods in political economy. Jevons gave his general defence of a unified scientific method in his major work, *The Principles of Science*, published in 1874.

Formal logic and the mechanics of the mind

To understand the impact of the emergence of formal logic in Britain, pioneered by Boole, De Morgan and Jevons, it is important to invoke the distinction between the moral (or mental) and natural realm that dominated Victorian discourse from the publication of Mill's *Logic* onwards. Quantitative methods of research – mathematics and statistics – were considered fit for the natural realm, but the realm of the mind could (and should) be explored introspectively. Logic, until then, was considered a branch of the sciences of mind, not of the natural sciences, and so the idea of using mathematics for its study was considered a violation of the distinction between these realms (Richards, 2002).

Earlier in the 19th century Babbage's famous calculating engines had opened up debates in which the idea was explored that our mind was, after all, no more than a calculating machine. Challenging the distinction between mind and matter, such thoughts (and machines) also challenged one of the backbones of Victorian moral philosophy, the notion of freewill. A variety of authors who agreed on next to nothing found themselves in the same camp on this issue: William Whewell, J.S. Mill, the utilitarian philosopher James Martineau and the common sense philosophers William Hamilton and Henry Longueville Mansel all argued against the use and usefulness of algebraic methods for the study of logic, precisely because this was felt to threaten the notion of free will, and more generally the foundations of moral agency.

Augustus De Morgan clearly was of opposite opinion. In his writing on Boole's algebraic logic and in his own formal logic, he emphasized the connections between an algebraic treatment of logic and mechanical theories of the working of the mind. In his own work in formal logic, Jevons followed De Morgan's lead. Jevons enthusiastically referred to Boole and Babbage and promised a logical machine of his own ‘which shall not only solve Aristotle's dilemma's, but shall exhibit to the eyes the working of Boole's logic the most general and perfect system of logic yet proposed’. Jevons worked on such a machine in the mid-1860s and wrote several articles to describe its working. To emphasize the relation between the machine, formal logic, and the method of science, he used an image of the machine as a frontispiece to his own magnum opus on scientific method, *The Principles of Science* (1874). Jevons once described this book to his brother Herbert as a work on formal logic ‘in disguise’.

It is worthwhile quoting from De Morgan's *Formal Logic* (1847) and Jevons's *The Principles of Science* (1874) to see the similarities between their endeavours. De Morgan wrote that ‘with respect to the mind, considered as a complicated apparatus, we are not even so well off as those would be who had to examine and decide upon the mechanism of a watch, merely by observation of the function of the hands, without being allowed to see the inside’. Jevons similarly wrote that ‘we are in the position of spectators who witness the productions of a complicated apparatus, but are not allowed to examine its intimate structure’. These extracts show not only their similarities, but, more importantly, how Jevons's (and De Morgan's) reliance on mechanical analogies shifted the grounds for studying the laws of mind from an introspective to an outsider's perspective; the emergence of formal logic challenged introspection as a viable route of discovery in the realm of the mind.

So what alternatives did Jevons propose for discovery in the moral sciences? Essentially, his general answer was: statistics and mathematics; the specific answer was: psychophysiology. The first answer related to Jevons's view that all sciences were quantitative in nature and that one needed proper instruments to measure these quantities. The second related to contemporary developments in the work of the physiologists William Carpenter, Henry Maudsley and others, work that seemed to promise the unravelling of the physiological groundwork of human agency. For Jevons these developments pointed in the direction of a mathematical theory of human agency, which was to be made exact with the help of statistics. I will first discuss the relation of Jevons's logical machine to his theory of induction and then discuss the relation of his reading into psychophysiology to his new theory of value.

Rethinking induction

To see the relationship between Jevons's logical machine, statistics and induction we need a brief outline of the machine's working. The logical machine had the appearance of a small piano. Its keys were either terms in a logical proposition or operations like ‘and’, ‘or’, and ‘is’. Though the machine was of limited capacity, Jevons used it as an illustration device of the more general process of logical inference. One of his examples was ‘Iron is a metal’, or in Jevons's formalism ‘$A = AB$’, and ‘a metal is a good conductor of electricity’, formally ‘$B = BC$’, an example that referred back to Hans Christian Oersted's famous thought experiment on the relation between electricity and magnetism. Once these propositions were fed to the machine, a tip on the ‘finis’ key made the machine ‘reason upon’ them. Thus, Jevons argued, just as Babbage had created ‘in the wheels and levers of an insensible machine’ a ‘rival’ of the human mind, so his logical machine ‘really accomplishes in a purely mechanical manner … the true process of logical inference’. All conclusions that could be drawn appeared on the display at the front of the piano (in the example given, there are eight valid conclusions). The list of conclusions would grow exponentially with the number of propositions.

At the time, the fact that the machine showed *all* logical conclusions was seen as a disadvantage. Jevons turned this disadvantage into a clue about induction. He believed that observations should be considered as the conclusions poured out by the fundamental machinery of nature. The task of the scientist was to infer back from the avalanche of conclusions to the fundamental propositions underlying them – in the example given, the scientist should infer back from the eight conclusions to the two producing propositions or ‘laws’. Jevons called this process ‘indirect deductive inference’. When facing nature's complexity, such indirect deductive inferences could only be hypothetical; there was no unambiguous procedure to know with certainty that the correct fundamental propositions were touched upon.

If nature's conclusions were quantitative, then the fundamental laws were quantitative as well. Jevons gave an example that Babbage had discussed in his *Ninth Bridgewater Treatise*, Bernoulli numbers. Behind a seemingly capricious array of numbers there was a simple mathematical formula generating them all. Once one left the realm of pure mathematics and entered the real world, however, an additional complicating factor entered the scene that related to the process of scientific observation itself. According to Jevons, all scientific observations were loaded with error, so the laws of nature could never fully account for individual observations, but should always relate to average values. Errors in measurement cancelled out in the average. For this reason Jevons never considered the target of scientific explanation a mathematical form fitting all individual observations; rather, science should find the 'rational formula', that is, the mathematical form that explained the phenomena behind the data. This is exactly the procedure Jevons followed in his own statistical investigations, and it put him at a large distance from Mill. As explained best by Sandra Peart (1995), Mill's 'disturbing causes' were for Jevons just 'noxious errors'. Jevons's investigations into the 'black art of induction' were thus closely connected to his investigations in formal logic and the logical machine. They gave a pivotal role to the manipulation of statistical data in discovering the underlying laws producing these data 'in the average'.

Mathematics and the 'physiological groundwork' of economics

In the *Principles* Jevons illustrated his meaning with reference to a series of (self-) experiments on the relationship between work and fatigue, which he published in *Nature* (1870). Jevons made these experiments using proxies for physical labour (holding weights, lifting weights with pulley and block, throwing weights) to show how, contrary to the opinion of political economists like Mill and Cairnes, the investigation of the 'physical groundwork of economics' could be mathematized. Thus, these experiments form an interesting link between Jevons's views on the role of statistical data as evidence for theories, and his views on the nature of political economy.

In his lectures on the method of political economy, Cairnes had discussed two authors in particular who proposed to ground economics in the mechanics of man's physiology, Richard Jennings and Henry Dunning Macleod. Cairnes observed that 'every economist, so soon as an economic fact has been traced to a mental principle, considers the question solved', and so did not need to take recourse to a superfluous examination of mankind's physiology. These authors thus transgressed the limits Mill had set on political economy. Referring to Jennings's *Natural Elements of Political Economy* (1855) in particular, Cairnes argued that he could not see how an examination of the 'afferent trunk of nerve-fibre' would clarify, for example, the phenomenon of consumption. If political economy consisted in the study of man's physiology, Cairnes complained, 'it is evident that it will soon become a whole different study from that which the world has hitherto known it'.

Jevons read Cairnes's lectures intensely, but he did not approve of its conclusions. In contrast with Cairnes, Jevons enthusiastically embraced Jennings's suggestion to ground the laws of political economy in man's physiology. In the *Theory* Jevons wrote that Jennings 'most clearly appreciated the nature and the importance of the laws of utility' by treating the 'physical groundwork of Economy, showing its dependence on physiological laws'. While Cairnes dismissed Jennings's suggestion to 'exhibit' the 'result of the principles of human nature … by the different methods of Algebra and Fluxions', Jevons considered this 'a clear statement of the views which I have also adopted'.

In a series of excellent essays, Michael White has explored the relation of Jevons's *Theory* to his engagement with psychophysiology and his closely related interest in the emerging theory of thermodynamics (see in particular White, 1994a; 2004; also Maas, 2005). From White's investigations the following image emerges. In his early years back in London, Jevons explored the meaning of value in economics using two resources, Bentham's felicific calculus of pleasures and pains, and contemporary research in psychophysiology, in particular the work of William Carpenter, which built on Marshall Hall's theory of reflex action of the 1830s. These investigations brought him to the idea that value, or what he later called the 'final degree of utility', could be examined by means of functional analysis. Jevons's own work on his highly successful *Coal Question* (1865) importantly contributed to his own awareness of the importance of the newly emerging discourse of thermodynamics. The incorporation of the new discourse of energy already transpired from the *Theory* itself, in particular from his theory of labour supply, and is even clearer in his (unfinished) *Principles of Economics*.

In June 1860 Jevons famously wrote to his brother Herbert that he had found the decisive elements of his new theory of utility, especially 'the most important axiom' of the declining degree of what he then called the '*ratio of utility*', and the assumption that, 'on an average', this ratio of utility was 'some continuous mathematical function of the quantity of commodity'. According to Jevons, political economists had assumed this law 'under the more complex form and name of the Law of Supply and Demand'. Hence, Jevons's engagements with psychophysiology and his attempt to mathematize the theory of value were intimately connected. Psychophysiology made Jevons think of a mechanism of pleasures and pains, which expressed itself in market prices.

Jevons's first airing of his new mathematical theory of value was the short *Notice* of 1862 that was read to the BAAS and published in 1866 as the *Brief Account*. Though Jevons had been thinking about his theory throughout the 1860s, he felt prompted to write it down after William Thornton's challenge to the classical wages fund theory in 1867. Thornton's challenge led to Mill's famous recantation and to vehement debates as to the character of the 'laws of supply and demand'. Jevons's extensive exchange of letters with the engineer Fleeming Jenkin on this topic was the immediate reason to speed up publishing a written version of his mathematical theory of utility. When Jenkin published a paper in 1870 called the 'Graphical Representation of the Laws of Supply and Demand', Jevons clearly feared that priority in a mathematical theory would escape him, and in just half a year he completed the *Theory*.

This book seriously transgressed the limits set by Mill on political economy's methods and subject. In his *Essay* of 1836 Mill had relegated the '*laws* of the *consumption* of wealth' to outside the domain of political economy. Jevons, by contrast, made these 'laws of human enjoyment' the cornerstone of his new theory. To articulate these laws, Jevons used

Bentham's felicific calculus, but he grounded this calculus in man's physiological dispositions. Rather than thinking of pleasures and pains as motives on which the mind decides, Jevons transformed them into physical forces that drive a mechanical balance to equilibrium. Figure 1 sums up the main characteristics of Jevons's utility theory.

Figure 1

Note: Jevons's diagrammatic representation of the utility adjustments of an individual (‘trading body’) to the optimum at *m*, at *given* market prices of two commodities *x* and *y*. The utility curve *p′r′* for commodity *y* is inverted and superposed upon the utility curve *pr* for commodity *x*. *Source*: Jevons (1871).

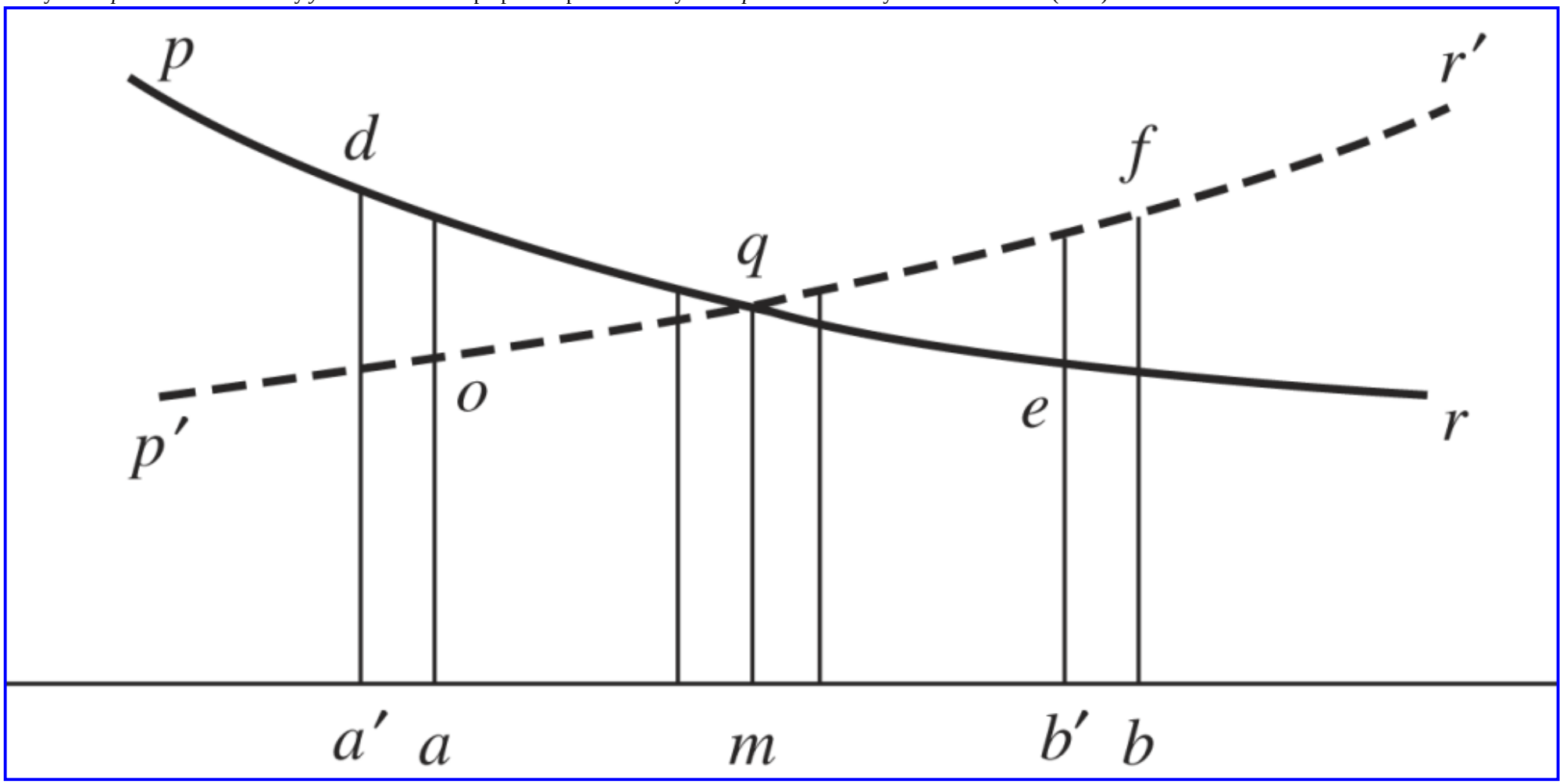

In Figure 1, two utility curves for two commodities *x* and *y* of one person (‘trading body’) are superposed and inverted upon one another. Utility is measured on the vertical axis, commodities on the horizontal. The diagram shows how this person would make a net gain in utility by extending trade from *a′* in the direction of *m*, and would lose in utility when trading beyond that point. Hence, there automatically emerges an equilibrium for this individual at *m*.

This balancing model was taken to represent the individual's balancing of pleasures and pains at the margin. As Jevons put it in the *Theory*: ‘the will is our pendulum and its oscillations are minutely registered in the price lists of the markets’. This theory enabled Jevons to state relative prices in terms of relative marginal utilities. What it did not show was *how* an equilibrium price was actually obtained through price adjustments, something that had been pointed out to him earlier in correspondence with Fleeming Jenkin. Hence, Jevons's theory did not explain price *formation*; it only showed how individuals adjusted their demands at a *given* price.

In the *Theory* Jevons suggested that numerical precision could be given to his theory by taking the so-called King–Davenant Price Quantity Table as an example, just as he used his experiments on work and fatigue to show that in principle numerical precision could be given to his theory of labour. This small table was found in the work of the 17th-century political arithmeticians Gregory King and Charles D’Avenant and allegedly contained statistical data on prices and quantities of wheat. Through the 19th century it had been widely used to argue for or against the possibility of mathematizing political economy, for example by Thomas Tooke, William Whewell, but also by Cairnes (Creedy, 1992). Jevons showed

how this table could give numerical exactness to the notion of the final degree of utility, the ‘all important element of political economy’, and so how statistical data could give precision to theory (Stigler, 1994; White, 1989).

The remainder of the book contained Jevons's theory of rent and capital that are interesting in their own right, but it was undeniably his theory of utility, worked out in his theory of exchange and labour, that brought Jevons the fame he hoped for. But not immediately.

In a letter to Cairnes of 5 December 1871 (Mill, 1972, pp. 1862–3), Mill wrote that he certainly would agree with Cairnes's negative judgement on the book:

> I have not seen Jevons's book, but as far as I can judge from such notices of it as have reached me [in Avignon], I do not expect I shall think favourably of it. He is a man of some ability, but he seems to have a mania for encumbering questions with useless complications, and with a notation implying the existence of greater precision in the data than the questions admit of. His speculations on Logic, like those of Boole and De Morgan … are infected in an extraordinary degree with this vice.

Interestingly, Mill considered Jevons's mathematical endeavours in economics on a par with his ‘speculations in logic’. Behind Mill's irritation we may guess a genuine concern with mechanistic theories of the mind, which Mill feared were a degradation of man's most ‘ennobling’ characteristics. It is only after Jevons's strong mechanistic image of man as a balance of pleasures and pains and the Victorian obsession with the problem of free will waned that his ‘mechanics of utility and self-interest’ could be considered to deal with rational choice. One will search in vain for this notion in Jevons's original work, however.

Jevons's later years

The publications dating after the *Theory* and *Principles* are generally thought to be of much less importance. Jevons turned his attention to his notorious sunspot studies, in which he attempted to establish a causal connection between solar activity and commercial crises. Though these studies were generally seen to be failures, the famous astronomer William Herschel had in fact voiced similar ideas on the relationship between agricultural output and the activity of the sun in the early 19th century. Jevons had to make increasingly far-fetched assumptions as to the causal mechanism involved, which cast doubt on the whole enterprise. Jevons also wrote a number of highly successful primers; the primer on logic went through numerous reprints (up to 1931), and his *Money and the Mechanism of Exchange* (1875) sold well too. Jevons also worked on the second edition of the *Theory*, which appeared 1879 and which contained an extensive survey of precursors in mathematical economics. In *The State in Relation to Labour* (1882) and a posthumous collection of essays on social reform (1883), Jevons turned his attention to the social and political issues of his day, issues that had been close to his mind from his formative years in Australia. At the end of the 1870s he wrote a number of vehement attacks on Mill's philosophy that, given the towering status of Mill as a political economist and philosopher, actually harmed Jevons's own intellectual status. Having moved from Owen's College to University College, London, in 1876 to take up the professorship in political economy, he resigned in 1880, partly because of problems of health but more importantly to be able to devote all his time to writing. His untimely death in 1882 left his last large project, the *Principles of Economics*, unfinished. It was published, with some additional essays, in 1905.

Jevons's investigative spirit

Jevons's statistical work in the 1850s and 1860s, his imaginative, though less well-known, work in formal logic, and in particular of course his *Theory of Political Economy* and *Principles of Science* stand out as landmark contributions to economics and to the philosophy of science. A genuine Victorian polymath, Jevons worked in many different fields of the sciences, all of which he engaged in the same investigative spirit. Though some of his work in the 1870s was quite successful, some lacked the sharpness and acuity of his earlier work. Equally well at home in theory as in the ‘black arts of inductive economics’, Jevons was, as Keynes noted in his 1936 obituary, ‘the first theoretical economist to survey his material with the prying eyes and fertile controlled imagination of the natural scientist’.

Jevons's investigative spirit, with its typical belief in the power of mathematics to capture the mechanical principles of the subject under study, irrevocably altered the image of economics, and is perhaps still with us. In many of the sciences, a satisfactory explanation nowadays requires the description of a mechanism, and economics is no exception to this. Nobel laureate Robert Lucas once described economic theory as providing an ‘explicit set of instructions for building … a mechanical imitation system’ (1980, p. 697). In retrospect we may hear the echo of Jevons's approach to economics in these words.

See Also

- analogy and metaphor
- energy economics

- equation of exchange
- functional analysis
- index numbers
- labour theory of value
- law of indifference
- marginal revolution
- mathematical methods in political economy
- statistics and economics
- utilitarianism and economic theory
- utility
- wages fund

Selected works

1865. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-mines*. London: Macmillan.

1866. Brief account of a general mathematical theory of political economy. *Journal of the Statistical Society of London* 29, 282–7.

1870. On the natural laws of muscular exertion. *Nature* 2, 158–60.

1871. *The Theory of Political Economy*. London: Macmillan.

1874. *The Principles of Science: a Treatise on Logic and Scientific Method*. New York: Dover, 1958.

1875. *Money and the Mechanism of Exchange*. London: King.

1879. *The Theory of Political Economy*, 2nd ed. London: Macmillan.

1882. *The State in Relation to Labour*. New York: Augustus M. Kelley, 1968.

1883. *Methods of Social Reform*. London: Macmillan.

1884. *Investigations in Currency and Finance*. London: Macmillan.

1886. *Letters and Journal of William Stanley Jevons*, ed. H.A. Jevons-Taylor. London: Macmillan.

1890. *Pure Logic and Other Minor Works*. New York: Burt Franklin, 1971.

1905. *The Principles of Economics and Other Papers*. London: Macmillan.

Bibliography

Aldrich, J. 1987. Jevons as statistician: the role of probability. *Manchester School of Economic and Social Studies* 55, 233–56.

Barrett, L. and Connell, M. 2005. Jevons and the logic ‘piano’. *Rutherford Journal* 1. Online. Available at http://www.rutherfordjournal.org/article010103.html, accessed 4 February 2007. Further details of the Sydney Powerhouse exhibition on Jevon's work are online, available at http://www.powerhousemuseum.com/exhibitions/jevons.asp, accessed 4 February 2007.

Black, R.D.C. 1960. Jevons and Cairnes. *Economica* 27, 214–32.

Black, R.D.C. 1972. Jevons, Bentham and De Morgan. *Economica* 39, 119–34.

Black, R.D.C, Coats, A.W. and Goodwin, C.D.W., eds. 1973. *The Marginal Revolution in Economics: Interpretation and Evaluation*. Durham, NC: Duke University Press.

Black, R.D.C. and Könekamp, R. 1972–81. *Papers and Correspondence of William Stanley Jevons*, vols 1–7. London: Macmillan.

Blaug, M. 1976. *Ricardian Economics: A Historical Study*. Westport, CT: Greenwood Press.

Cairnes, J.E. 1857. *The Character and Logical Method of Political Economy*. London: Longman Brown Green Longmans and Roberts.

Creedy, J. 1992. *Demand and Exchange in Economic Analysis: A History from Cournot to Marshall*. Aldershot: Elgar.

Davison, G. 1997–98. The unsociable sociologist: W.S. Jevons and his survey of Sydney, 1856–8. *Australian Cultural History* 16, 127–50.

De Marchi, N.B. 1972. Mill and Cairnes and the emergence of marginalism in England. *History of Political Economy* 4, 344–63.

De Morgan, A. 1847. *Formal Logic, or, The Calculus of Inference, Necessary and Probable*. London: Taylor & Walton.

Durand-Richard, M.-J. 1991. Babbage, Boole, Jevons between science and industry: the principle of analogy and the mechanization of operations. In *The Interaction between Technology and Science*, ed. B. Gremmen. Wageningen: Wageningen Agricultural University.

Grattan-Guinness, I. 2002. ‘In some parts rather rough’: a recently discovered manuscript version of William Stanley Jevons's *General Mathematical Theory of Political Economy* (1862). *History of Political Economy* 34, 685–726.

Inoue, T. 2002. *W. Stanley Jevons: Collected Reviews and Obituaries*, 2 vols. Bristol: Thoemmes.

Jenkin, F. 1996. *The Graphic Representation of the Laws of Supply and Demand and Other Essays on Political Economy*. London: Routledge/Thoemmes Press.

Jennings, R. 1855. *Natural Elements of Political Economy*. New York: Augustus M. Kelley, 1969.

Kim, J. 1995. Jevons versus Cairnes on exact economic laws. In Rima (1995).

Klein, J.L. 1995. The method of diagrams and the black arts of inductive economics. In Rima (1995).

Lucas, R.E. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12, 696–715.

Maas, H. 2005. *William Stanley Jevons and the Making of Modern Economics*. Cambridge: Cambridge University Press.

Mill, J.S. 1843. A System of Logic, Ratiocinative and Inductive: Being A Connected View Of The Principles Of Evidence And The Methods Of Scientific Investigation. In *The Collected Works of John Stuart Mill*, vols 7 and 8, general ed. J.M. Robson. Toronto: University of Toronto Press, 1973–4.

Mill, J.S. 1844. *Essays on Some Unsettled Questions of Political Economy*. 2nd edn. London: Longmans, Green, Reader, and Dyer Co., 1874. In *Essays on Economics and Society*. *The Collected Works of John Stuart Mill*, vol. 4, general ed J.M. Robson. Toronto: University of Toronto Press, 1967.

Mill, J.S. 1972. *The Later Letters of John Stuart Mill 1849–1873 Part IV. The Collected Works of John Stuart Mill*, vol. 17, general ed. J.M. Robson. Toronto: University of Toronto Press.

Mirowski, P. 1989. *More Heat than Light: Economics as Social Physics, Physics as Nature's Economics*. Cambridge: Cambridge University Press.

Mosselmans, B. 1998. William Stanley Jevons and the extent of meaning in logic and economics. *History and Philosophy of Logic* 19, 83–99.

Mosselmans, B. 2007. *William Stanley Jevons and the Cutting Edge of Economics*. London and New York: Routledge.

Mosselmans, B. and White, M.V. 2001. *Collected Economic Writings of W.S. Jevons*. London: Palgrave Macmillan.

Nicholls, N. 1998. William Stanley Jevons and the climate of Australia. *Australian Meteorological Magazine* 47, 285–93.

Peart, S.J. 1995. Disturbing causes, noxious errors, and the theory-practice distinction in the economics of J.S. Mill and W.S. Jevons. *Canadian Journal of Economics* 28, 1194–211.

Peart, S.J. 1996. *The Economics of W.S. Jevons*. London and New York: Routledge.

Peart, S.J. 2003. *W.S. Jevons: Critical Responses*, 4 vols. London and New York: Routledge.

Playfair, W. 1801. *The Commercial and Political Atlas and Statistical Breviary*, ed. H. Wainer and I. Spence. Cambridge: Cambridge University Press, 2005.

Richards, J. 2002. In a rational world all radicals would be exterminated: mathematics, logic, and secular thinking in Augustus De Morgan's England. *Science in Context* 15, 137–64.

Rima, I.H. 1995. ed. *Measurement, Quantification and Economic Analysis: Numeracy in Economics*. London: Routledge.

Robbins, L.C. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan, 1984.

Schabas, M. 1990. *A World Ruled by Number: William Stanley Jevons and the Rise of Mathematical Economics*. Princeton: Princeton University Press.

Schabas, M. 2005. *The Natural Origins of Economics*. Chicago: University of Chicago Press.

Schmitt, R.W. 1995. The salt finger experiments of Jevons (1857) and Rayleigh (1880). *Journal of Physical Oceanography* 25, 8–17.

Stigler, S.M. 1982. Jevons as statistician. *Manchester School of Economic and Social Studies* 50, 354–65.

Stigler, S.M. 1994. Jevons on the King Davenant law of demand: a simple resolution of a historical puzzle. *History of Political Economy* 26, 185–91.

Watts, H. 1868. *A Dictionary of Chemistry and the Allied Branches of Other Sciences*. London: Longmans, Green & Co.

White, M.V. 1989. Why are there no supply and demand curves in Jevons. *History of Political Economy* 21, 425–56.

White, M.V. 1994a. The moment of Richard Jennings: the production of Jevons's marginalist economic agent. In *Natural Images in Economic Thought: 'Markets Read in Tooth and Claw'*, ed. P. Mirowski. Cambridge: Cambridge University Press.

White, M.V. 1994b. 'That God-forgotten Thornton': exorcising Higgling after *On Labour*. *History of Political Economy* 26(supplement), 149–83.

White, M.V. 1994c. Bridging the Natural and the Social: Science and Character in Jevons's Political Economy. *Economic Inquiry* 32, 429–44.

White, M.V. 2004. In the lobby of the energy hotel: W.S. Jevons' formulation of the post-classical economic problem. *History of Political Economy* 36, 227–71.

White, M.V. 2006. A painful disposition to classification: W.S. Jevons' first statistical chart in political economy. Mimeo, Monash University.

Wood, J.C. 1988. *William Stanley Jevons: Critical Assessments*, 3 vols. London and New York: Routledge.

How to cite this article

Maas, Harro. "Jevons, William Stanley (1835–1882)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000012> doi:10.1057/9780230226203.0864

The New Palgrave Dictionary of Economics Online

Jewkes, John (1902 – 1988)

Z.A. Silberston

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

free markets; Jewkes, J.; planning

Article

Jewkes was educated at Barrow Grammar School and Manchester University. His first job was as Assistant Secretary of the Manchester Chamber of Commerce, 1925–26. He was then appointed lecturer in economics at Manchester University, and stayed there for three years. Following a period in the United States, he returned to Manchester as Professor of Social Economics in 1936. After holding this chair for ten years, he was appointed Stanley Jevons Professor of Political Economy at Manchester. In 1948 he became Professor of Economic Organization at Oxford, and a Fellow of Merton College, and held this chair until his retirement in 1969. His professional contacts, however, remained mainly outside Oxford. Jewkes had a distinguished wartime career. He became Director of the Economic Section of the War Cabinet Secretariat in 1941, and was appointed Director-General of Statistics and Programmes at the Ministry of Aircraft Production in 1943. This was followed by other posts, and after his return to university life he was a member of a number of royal commissions and other official committees. Jewkes's Manchester roots, together with his wartime experience, made him a powerful advocate of free-market solutions. His first notable book on this subject was *Ordeal by Planning* (1948), followed by *Public and Private Enterprise* (1965), *New Ordeal by Planning* (1968), and *A Return to Free Market Economics*? (1978). In these works he advocated the virtues of the free market, as opposed to government ownership or government planning, as a fruitful background for economic efficiency and individual initiative. He argued that government efforts to replace the market had produced one debacle after another, and also that economists claimed too much for their subject, thus reducing their potential usefulness. Before the Second World War Jewkes's work had concentrated on detailed studies of the economic and social problems of Lancashire – as, for example, in his *Wages and Labour in the Cotton Spinning Industry* (1935, with E.M. Gray). Some of his work after the war also concentrated on detailed problems, but in a national or international context. For example, he published studies, jointly with his wife Sylvia, on medicine and the National Health Service, arguing that the state-operated National

Health Service had displayed many weaknesses. A notable contribution to the literature on innovation was his *The Sources of Invention* (1958, with David Sawers and Richard Stillerman). This was one of the earliest attempts at systematic investigation in this field. It successfully established the importance of the small-scale inventor, and showed that many notable 19th- and 20th-century inventions were essentially the work of one or two individuals, working with limited resources. This may well prove to be his most lasting contribution.

Selected works

1935. (With E.M. Gray.) *Wages and Labour in the Lancashire Cotton Spinning Industry*. Manchester: Manchester University Press.

1938. (With S. Jewkes.) *The Juvenile Labour Market*. London: Victor Gollancz.

1948. *Ordeal by Planning*. London: Macmillan.

1958. (With D. Sawers and R. Stillerman.) *The Sources of Invention*. London: Macmillan.

1961. (With S. Jewkes.) *The Genesis of the British National Health Service*. Oxford: Basil Blackwell.

1965. *Public and Private Enterprise*. London: Routledge & Kegan Paul.

1968. *The New Ordeal by Planning. The Experience of the Forties and Sixties. Revised edn of Ordeal by Planning*. London: Macmillan; New York: St Martin's Press.

1978. *A Return to Free Market Economics? Critical Essays on Government Intervention*. London: Macmillan.

How to cite this article

Silberston, Z. A. "Jewkes, John (1902–1988)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000013> doi:10.1057/9780230226203.0865

The New Palgrave Dictionary of Economics Online

Jim Crow South

Gavin Wright

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The US South maintained a distinctive economic and political structure from the demise of slavery in the 1860s to the Civil Rights revolution of the 1960s. Racial wage differentials in the unskilled labour market were small. But blacks were virtually absent from higher-paying skilled jobs. Disfranchisement led to a drastic fall in relative expenditures on black schooling between 1890 and 1910. The effort to protect cheap labour reinforced regional isolation, depriving the South of dynamic stimulus from new migrants, enterprise and ideas. Conflicts between recruitment of capital and demands for racial justice were resolved only by federal intervention.

Keywords

agriculture; Arrow, K.; Becker, G.; black-white wage differentials; civil rights movement; convict labour; economic development; human capital; internal migration; Jim Crow South; Key, V.; minimum wage; racial discrimination; regional development; schooling; segregation in labour markets; slavery; trade unions

Article

‘Jim Crow’ was a blackface caricature from a minstrel show that pre-dated the US civil war. The term came to represent the regime of racial segregation that became entrenched in both law and custom by the end of the 19th century in the US South.

From the demise of slavery in the 1860s to the Civil Rights revolution of the 1960s, the southern states maintained a distinctive economic and political structure. This historical episode raises a number of issues of general interest for economics, among them the effects of segregation on efficiency, and the impact of the segregationist regime on the economic progress of the region as a whole.

Jim Crow segregation did not emerge full-blown in the aftermath of war and emancipation, but instead had its own evolution. Appearing in the midst of the school segregation debate of the 1950s, C. Vann

Woodward's classic *The Strange Career of Jim Crow* (1955) overturned the myth that the races had ‘always’ been separated in the South, noting that explicit racial codes did not appear in most states until the 1890s. Blacks participated actively in Southern politics during Reconstruction (the period of federal military control, 1865–77), and longer in many states. Only when it became clear that neither the federal government nor the Supreme Court would intervene were Southern states emboldened to disfranchise black voters, using a variety of ostensibly non-racial devices (such as poll taxes and literacy tests) whose racial intent was barely disguised. Mississippi's constitution of 1890 was the first statewide disfranchisement programme, and by 1910 the exclusion of blacks from Southern politics was nearly complete. Legally mandated segregation soon followed disfranchisement, ultimately extending not only to schools, churches, eating establishments and recreation, but to public transportation, hospitals, prisons, cemeteries and other avenues of life. Although all were aware that it was honoured more in the breach than the observance, the ‘separate but equal’ principle was upheld by the US Supreme Court in the famous 1896 case, *Plessy vs. Ferguson*.

Even more onerous than the demand for physical separation were brutal features of the Jim Crow South, such as lynching (extra-legal executions) and convict labour (leasing of prisoners to private contractors). Neither of these phenomena was exclusively racial, but their impact fell most heavily on black Southerners.

Segregation and labour markets

Extensive as the scope of legal segregation became, its limits were equally notable. Racial aspects of employment and work relations were virtually unregulated. The only industrial segregation laws of any importance – a North Carolina statute requiring separate toilets and a South Carolina law requiring segregation in cotton textiles – were adopted only in 1913 and 1915, respectively – long after prevailing racial patterns were established – and were not imitated elsewhere. Yet, despite the absence of legal enforcement, segregation was the norm in Southern industries. In his study of Virginia firms in 1900 and 1909, Higgs (1977, p. 241) found that ‘occupational workforce segregation was overwhelmingly the rule’. Interestingly, racial separation was more prevalent and more clearly delineated by industry than by location. White cotton mills and black tobacco factories coexisted in places like Durham, North Carolina, and Danville, Virginia; in Birmingham, Alabama, where two-thirds of iron and steel workers were black, the Avondale cotton mill was 98.1 per cent white.

Explaining segregation in labour markets does not pose a serious challenge for economic theory. The models of Becker (1957) and Arrow (1973), among others, show that, if whites demand a premium for working in close association with blacks, segregation dominates mixed alternatives. The issue that economists have wrestled with is not segregation per se, but wage discrimination: did segregation serve to support an ‘unjustified’ wage differential, or was it merely the market's way of avoiding the costs of mixing the races? The perhaps surprising finding of numerous studies is that, despite the prevalence of racism in the Jim Crow South, racial wage differentials in the open (unskilled) labour market were small or non-existent.

In agriculture, wage labour coexisted with sharecropping and other forms of tenancy. Although whites had a large overall advantage in farm property and incomes, whites as well as blacks could be found at all stages of the ‘agricultural ladder’. With rare exceptions, black and white farm labourers were paid the

same wage. For example, in 1887 the North Carolina Bureau of Labor Statistics posed the question of racial wage differences to landlords as well as to tenants and labourers. In 94 of the 95 counties, the landlords reported ‘no difference’ between the races, and in 77 of the 95 counties the tenants and labourers gave the same response (cited in Higgs, 1978, p. 310). Even more remarkable, evidence from pre-First World War Virginia indicates that unskilled wages were equilibrated for black and white labourers, even across highly segregated industries such as cotton textiles and tobacco manufacturing (Whatley and Wright, 1994). The large intermediating agricultural sector may have been important in maintaining this equilibration for men, because the same data suggest a 25 per cent wage gap in favour of white over black women, who did not have access to farm labour jobs.

Equally noteworthy, however, is the virtual exclusion of blacks from higher-paying skilled jobs in Southern industry. Within agriculture, blacks were often able to rise up the ladder of accumulation, from wage labour to tenancy and even to farm ownership, albeit usually on a small scale. Such advancement opportunities were rare in non-agricultural sectors. In some cases, such as railroads, barriers to black promotion were enforced by all-white craft unions (Sundstrom, 1990). Elsewhere blacks were held back even in the absence of unions and even where skills were largely acquired on the job. Promotion of blacks to supervisory positions over whites was widely seen as unthinkable. Thus, despite the efficacy of labour markets in equilibrating unskilled wages, access to skilled positions was distinctly unequal between the races.

Race and schools

One direct consequence of disfranchisement was a drastic fall in relative expenditures on black schooling. The inequity was most extreme in the black-majority counties of the lower South, where funding was simply diverted from white to black schools. For example, in Mississippi in 1907 predominantly white counties spent $3.50 per school-age child on blacks but $5.60 on whites; in predominantly black counties, $2.50 was spent on blacks but $80.00 on white children. Black schools were also characterized by lower teaching salaries, higher student-teacher ratios, shorter terms and lower educational levels of teachers (Bond, 1934).

Economists often identify poor schooling as the primary explanation for low black incomes throughout the Jim Crow era (Smith, 1984). This interpretation meshes comfortably with the perspective emphasizing that barriers to black progress operated through political channels rather than through discrimination in markets. Economic historians, however, generally interpret the politics of Jim Crow as part of a larger political-economic package.

Landowning planters actively opposed higher spending on black schools, not just because funding could be diverted towards white children but because ‘educated Negroes, in nearly all cases, become valueless as farm laborers’ (quoted in Anderson, 1988, p. 96). As one Arkansas planter put it in 1900: ‘My experience has been that when one of the younger class gets so he can read and write and cipher, he wants to go to town. It is rare to find one who can read and write and cipher in the field at work’ (quoted in Wright, 1986, p. 79). In other words, restricting black education was a way of preserving the agricultural labour force.

Even outside of agriculture, exclusion of blacks from skilled jobs exercised a feedback effect on the demand for education. When the Rosenwald Fund sought to provide funding for black high schools in

the South during the 1920s and 1930s, it learned that there were no black jobs for which a high-school education would be useful. Thus, black schools typically did not offer training in such subjects as stenography, accounting, bookkeeping, printing or typing. The fund's curriculum expert acknowledged: ‘If commercial courses were offered in the Negro school there would no doubt be tremendous pressure to get into them and the only result would be keen disappointment for nearly everyone’ (quoted in Anderson, 1988, pp. 223–4).

Because of this mutual interaction between schooling and labour markets, the interwar years saw the opening of a racial wage gap for entry-level positions, in contrast to pre-First World War patterns. Proximate reasons for divergence include stagnant world demand for cotton during the 1920s, disproportionately affecting blacks, and an upward shift of real wages in the all-white cotton textiles industry, initially because of wartime inflation, and subsequently resistant to reduction for both internal and external reasons. The long-term consequence was that racial segregation took on a different economic character, becoming more a ‘vertical’ support for wage differentials than a ‘horizontal’ separator of the races as in the earlier period. According to a 1937 survey (Perlman and Frazier, 1937), firms hiring only blacks paid starting wages one-third lower than those hiring only whites; of those hiring both, nearly 30 per cent paid blacks a lower starting wage. In contrast, no explicit racial wage differential was reported in Northern firms. The ‘separate wage rates for Negroes’ that Southern observers took to be ‘a fixed tradition’ had in reality developed and become institutionalized only in the 20th century (Whatley and Wright, 1994). Margo (1990) finds that employment segregation increased between 1900 and 1950, even after racial differences in schooling are controlled for.

Economic development in the Jim Crow South

Granted that the Jim Crow regime adversely affected African Americans, the question may be posed: what was its effect on economic development in the region? Although theory suggests that racial discrimination is inefficient, it is not straightforward to detect inhibiting effects on the growth of major Southern industries. Cotton textiles, the most racially exclusive of them all, surpassed the historic New England branch by the turn of the 20th century. Rapid growth in such diverse industries as iron and steel, fertilizer, tobacco manufactures and furniture did not seem to be hindered by the colour line in employment, and Southern value-added in manufacturing grew faster than the national average throughout the Jim Crow era. Nonetheless, per capita income in the South was roughly half the national average as of 1880, and this ratio had barely changed by 1940. Can this failure to converge on national norms be tied to Jim Crow institutions?

Growth-accounting analysis attributes much of the regional income gap to low levels of education in the South (Connolly, 2004). Underinvestment in human capital extended to Southern whites as well as blacks, a phenomenon that was also historically linked to race. Disfranchisement of blacks deprived many lower-income whites of the vote at the same time, preventing a class-based political mobilization that might have overcome planter opposition to funding for public schools (Kousser, 1974). In the classic analysis of political scientist V.O. Key (1949), regional unity on the race issue led to one-party politics, depriving the South of the popular political participation that elsewhere supported public schools and other measures favoring economic development.

Perhaps the worst effect of Jim Crow on economic development was that the effort to protect cheap

regional labour led to regional isolation, depriving the South of dynamic stimulus from new migrants, enterprise and ideas. In 1910, just two per cent of Southern residents were foreign-born, the lowest share in the nation. Connolly (2004) finds that low levels of human capital did not just lower average incomes in the South but also slowed the diffusion and generation of new technologies. In the 1930s, when regions were actively competing for newly available federal funds, the states of the South – the most solidly Democratic in the country – received the lowest levels of federal support per capita. One main reason was that Southern political and business leaders feared the effects of federal funding on wages, labour discipline and race relations.

The demise of the Jim Crow South

The Jim Crow world crumbled under federal pressure during the 1960s. But this political revolution was preceded by an earlier regime change in economic policy, occurring between the 1930s and the 1950s. At that time the South began its modern economic take-off, an acceleration of growth dated from approximately 1940. As a case study in economic modernization, the episode is highly unusual in that the acceleration coincided with massive outmigration from the region in question. Low-income, poorly educated Southerners left the countryside for cities in both North and South, while professionals and retirees began to move Southward, into fast-growing cities and sunbelt retirement areas. Migration was racially as well as economically selective. Net Southern *white* outmigration all but ended by the 1950s, while blacks continued to leave the region in large numbers through the 1960s.

With the advent of the national minimum wage (and related labour market regulations) in the 1930s, and the renewal and extension of these policies in 1950s, it was clear to business leaders of the South that an Asian-style industrialization based on cheap labor within US borders was not going to be politically acceptable. At roughly the same time, full mechanization of cotton growing became feasible, and was all but complete by 1960. Together, these developments tipped the political balance towards vigorous efforts to attract business through tax breaks, municipal bonds for plant construction, industrial development corporations, research parks and expenditures on publicity far beyond those of other regions. James C. Cobb (1982) calls it the ‘selling of the South’.

One might suppose that enlightened Southern businessmen should have led the way in breaking down racial barriers, but the evidence suggests that most were extremely reluctant to do so. In city after city, business leaders weighed in on the side of compromise, but only after political turbulence reached the point where it threatened the flow of investment capital. For their part, employers had no strong economic motives for challenging racial norms, since low-end wages were governed by federal law, and few blacks were qualified by education or experience for high-end jobs. This perverse regional equilibrium might have survived indefinitely on purely economic grounds. But ultimately the irresistible force of economic progress came into collision with the immovable object of Jim Crow.

The leverage of the movement derived from the fact that competition for outside capital required Southern leaders to present their towns and cities as safe, civilized communities, with a labour force that was well-behaved and eager for work. The most famous case in point was Little Rock, Arkansas, where a promising post-war development programme came to a standstill when Orval Faubus called out the National Guard to block court-ordered school integration in 1957. Although the city had attracted eight new plants in 1957, not a single new plant came to Little Rock during the next four years. A widely discussed *Wall Street Journal* headline for 26 May 1961 read: ‘Business in Dixie: Many Southerners

Say Racial Tension Slows Area's Economic Gains.' In her systematic review of Southern businessmen's response to the desegregation crisis, Elizabeth Jacoway writes, 'In the 1950s and 1960s, white businessmen across the South found themselves pushed – by the federal government and civil rights forces as well as by their own economic interests – and values – into becoming reluctant advocates of a new departure in southern race relations' (Jacoway and Colburn, 1982, p. 1). In a sense they had to be coerced to act in their own economic interest! Although few were willing to say so in public, many local leaders and business proprietors were privately grateful for civil rights legislation of the 1960s, at least after the fact. These measures largely put an end to disputes over public accommodations and employment segregation, while providing managers the ready-made excuse that the matter was no longer in their hands (Wright, 1999).

Since then, the South has been the most rapidly growing region in the United States. The political revolution has generated economic gains for blacks as well as whites. After the political breakthroughs of the 1960s, more than 50 years of black outmigration came to an end, and blacks have been moving into the region ever since. Net black migration into the South amounted to more than 500,000 between 1990 and 2000, whereas net black migration was negative for each of the other census regions. The attraction of the New South for blacks has economic as well as cultural, political and geographic aspects. As of 1977, the majority of the nation's black-owned businesses were in the South. Median black income grew faster in the South than elsewhere, and by the end of the 20th century equalled or surpassed median black income in the north-eastern and the mid-western regions.

See Also

- black–white labour market inequality in the United States
- regional development, geography of

Bibliography

Anderson, J. 1988. *The Education of Blacks in the South, 1860–1935*. Chapel Hill: University of North Carolina Press.

Arrow, K. 1973. The theory of discrimination. In *Discrimination in Labor Markets*, ed. O. Ashenfelter and A. Rees. Princeton, NJ: Princeton University Press.

Becker, G. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bond, H. 1934. *The Education of the Negro in the American Social Order*. New York: Prentice-Hall.

Cobb, J. 1982. *The Selling of the South: The Southern Crusade for Industrial Development, 1936–1980*. Baton Rouge: Louisiana State University Press.

Connolly, M. 2004. Human capital and growth in the Postbellum South. *Journal of Economic History*

64, 363–99.

Higgs, R. 1977. Firm-specific evidence on racial wage differentials and workforce segregation. *American Economic Review* 67, 236–45.

Higgs, R. 1978. Racial wage differentials in agriculture: evidence from North Carolina in 1887. *Agricultural History* 52, 308–11.

Jacoway, E. and Colburn, D., eds. 1982. *Southern Businessmen and Desegreation*. Baton Rouge: Louisiana State University Press.

Key, V. O. 1949. *Southern Politics in State and Nation*. New York: A. A. Knopf.

Kousser, M. 1974. *The Shaping of Southern Politics: Suffrage Restriction and the Establishment of One-Party South, 1880–1910*. New Haven: Yale University Press.

Margo, R. A. 1990. *Race and Schooling in the South, 1880–1950*. Chicago: University of Chicago Press.

Perlman, J. and Frazier, E. 1937. Entrance rates of common laborers in 20 industries, July 1937. *Monthly Labor Review* 45, 1491–510.

Smith, J. 1984. Race and human capital. *American Economic Review* 74, 685–98.

Sundstrom, W. A. 1990. Half a career: discrimination and railroad internal labor markets. *Industrial Relations* 29, 423–40.

Whatley, W. and Wright, G. 1994. Race, human capital, and labour markets. In *Labour Market Evolution*, ed. G. Grantham and M. MacKinnon. New York: Routledge.

Woodward, C.V. 1955. *The Strange Career of Jim Crow*. New York: Oxford University Press.

Wright, G. 1986. *Old South, New South: Revolutions in the Southern Economy Since the Civil War*. New York: Basic Books.

Wright, G. 1999. The civil rights revolution as economic history. *Journal of Economic History* 59, 267–89.

How to cite this article

Wright, Gavin. "Jim Crow South." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of

Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000050> doi:10.1057/9780230226203.0866

The New Palgrave Dictionary of Economics Online

Johansen, Leif (1930 – 1982)

Karl H. Borch
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bargaining; bargaining sets; central planning; coalitions; economic growth; Frisch, R. A. K.; game theory; Haavelmo, T.; Johansen, L.; labour theory of value; Leontief's input–output model; marginal utility; perfect competition; Prisoner's Dilemma; production functions; public economics

Article

Johansen was born in Eidsvoll, Norway, on 11 May 1930 and died in Oslo on 29 December 1982. He entered the University of Oslo in 1948, and received the equivalent of a Master's degree in economics (*cand.oecon*.) in 1954. He was awarded a doctors degree (dr. philos) in 1961, for a dissertation with the title ‘A Multi-Sectoral Study of Economic Growth’.

In 1951 Johansen became research assistant to Ragnar Frisch. After graduation the university awarded him a research fellowship. In 1958 he received a Rockefeller Fellowship, which he held until in 1959 he was appointed Associate Professor of Public Economics at the University of Olso. On the retirement of Frisch in 1965, Johansen became Professor of Economics at the University of Oslo, with the special duty of lecturing on macroeconomic planning.

Johansen's first important work is his doctoral dissertation, mentioned above. This book (1960) builds a bridge between the theory of economic growth, which had become fashionable in the 1950s, and Leontief's input–output model, which at the time was widely used in economic planning and forecasting. The choice of dissertation topic was undoubtedly influenced by Johansen's two mentors, Frisch and Trygve Haavelmo, who were then both working in these fields.

In the dissertation Johansen presented a theoretical model, and applied it to Norwegian data. He analysed a 23-sector model of the economy, and it seems that at first this empirical part of the work was considered the more important.

After a few years, however, it became clear that the model, often referred to as the MSG-model, had considerable merits in itself. It became the basis for long-term planning by the Norwegian Ministry of Finance, and over the years it was developed and extended. Johansen took an active part in this work. It seems that the model also influenced planning methods in several countries, and a new and enlarged

edition of the book was published in 1974.

The laws of production must play an important part in any growth model, and Johansen continued and extended the pioneering work of Frisch on production functions in a series of articles which had considerable influence. The main results in these papers are brought together and generalized in the book from 1972: *Production Functions: An Integration of Micro and Macro, Short Run and Long Run Aspects*. The subtitle indicates the high aspiration level of the book, and it did present production functions which were realistic and so general that they could be used in multisectoral planning models.

Johansen was a member of the Communist Party of Norway until he died, and he participated actively in some election campaigns. However, his political views have hardly left a trace in his professional writing. An uniformed reader will be at loss to divine which political opinions – if any – the author holds. Johansen seems to have written relatively few papers on planning in eastern Europe and on Marxist economic theory, and none after 1966. His objectives seem to be to inform and explain, rather than to convert, and often it seems that these papers are written on request – for instance his paper ‘Labour Theory of Value and Marginal Utilities’ (1963). This is an extension and clarification of some short comments he made in a discussion the year before. It shows that under certain circumstances the two theories can be reconciled. Johansen served on a number of expert committees appointed by different Norwegian governments, and was accepted as the objective scientist who would point out logical inconsistencies but never let his personal views influence the recommendations he made.

There is, however, little doubt that Johansen's political opinions had a marked effect on his career. Under the rules in force in the 1950s and 1960s it was impossible for him to obtain a visa to the USA. He therefore did not have the opportunity of spending some of his formative years at an American university. Such opportunities were regularly offered to bright young academics in western Europe and usually had a profound influence on their later work. Johansen missed this experience, and in fact never visited the USA. He remained a European, and principally a Norwegian. Most of his work was published in Europe, and about half of it was written in Norwegian.

Johansen's political views did not affect his scientific work, but his views did inevitably influence his opinions as to which economic problems were important and which should be studied. His views naturally led him to study economic planning, and this subject remained Johansen's main interest during most of his professional life. His two-volume *Lectures on Macroeconomic Planning* (1977 and 1978) is a landmark. It is essentially a textbook which gives a balanced overview of the major issues in the economics of planning, integrating the results Johansen reached over 25 years with those of the many others who contributed to the development of the subject. As often Johansen appears as a master in reconciling different views and approaches. A third volume was in preparation at the time of Johansen's death, and this might have rounded off the work, and removed the many gaps and omissions which reviewers found in the presentation.

Economic planning is closely related to, if not a part of, the subject which has become known as ‘public economics’. The subject may not be very well defined, and its contents have certainly changed over the years. The central topics, however, remain taxation, public expenditure and social welfare. When Johansen began to lecture on the subject at the University of Oslo in 1960, there was no single book which covered this heterogeneous subject. He published his own textbook in Norwegian in 1962. A revised and extended edition appeared in 1965, and was translated into English in the same year as *Public Economics*. The book did not give any clear definition, nor did it define the limits of the subject. Perhaps too tailor-made for his students at the University of Oslo, it deals very briefly with topics

covered by other courses in the curriculum. The book did, however, have an impact, and helped to establish ‘public economics’ – suitably defined – as a recognized part of economics.

Johansen was one of the founders of the *Journal of Public Economics*, which first appeared in 1972. He served as co-editor from the beginning until his death, and he contributed the opening article of the new journal. Its title, ‘On the Optimal Use of Forecasts in Economic Policy Decisions’, indicates how broadly Johansen tended to view the subject of public economics.

In his later years Johansen developed a strong interest in game theory. He seems to have been led to this subject by Arrow's proof that non-dictatorial and efficient decisions were impossible, and he wrote a penetrating paper on the subject in 1969. A model of central planning will naturally be compared – for efficiency and fairness – with a model of free competition. The assumptions leading to neoclassical equilibrium are generally considered to be unrealistic, and game theory, with its different solution concepts based on compromises between coalitions, were developed as a generalization of the standard market model. The same idea can be applied to a central planning model. Plans are rarely drawn up and executed by a consistent single-minded dictator. Usually they appear as a compromise between different interest groups (coalitions) in society, or within a bureaucracy.

Johansen's first publication on game theory seems to be a short article in Norwegian with the title ‘Plans and Games’ from 1970, contributed to a Festschrift with the general title ‘Economics and Politics’. Here he shows that, if there are several independent decision makers, with different preferences, the collective decision must necessarily be a compromise.

In the following years Johansen published a few papers in Norwegian along similar lines. His first paper on a game theory in English is ‘A Calculus Approach to the Theory of the Core of an Exchange Economy’, published in 1978. Debreu and Scarf (1963) proved that the core of a market game would, under certain conditions, shrink to the competitive equilibrium, as the number of players increased to infinity. Their proof, as well as the ones given by others, depends heavily on topological or measure theoretical arguments, which make the results inaccessible to most economists of the older generation. Johansen shows that the result can be reached by elementary methods, under the assumptions conventionally made in neoclassical economic theory. The paper does not appear to be much cited, and its main effect may have been to give Johansen a deeper understanding of the subject.

Game theory is closely related to bargaining theory, and Johansen's next paper on the subject is ‘The Bargaining Society and the Inefficiency of Bargaining’ (1979). Here he wrote: ‘I consider the game theory approach to economic problems to be the most appropriate paradigm as soon as we go beyond mere accounting and description of production technology and want to include various aspects of economic behaviour.’ The conclusion of the essentially verbal discussion in the paper is that bargaining is not an efficient way of making social decisions. At the time of writing Johansen did not seem to be aware of the concept of ‘bargaining sets’ introduced by Aumann and Maschler (1964). The different bargaining sets include the core if it is not empty, and also some subsets corresponding to the cases in which the players fail to agree on a Pareto optimal outcome.

In one of his last papers, ‘On the Status of the Nash Type of Noncooperative Equilibrium in Economic Theory’ (1982), Johansen argued that the theorem of Nash (1950) has often been misinterpreted and misused in economic literature. The theorem just states that every *n*-person game has at least one equilibrium point in mixed strategies. In this purely mathematical context equilibrium point means what in mechanics is called a ‘dead point’, where the forces are in equilibrium. There are few reasons to assume that a point with this property should have any economic optimality property. Johansen

observes, inter alia, that in the game known as the ‘Prisoner's Dilemma’ the only equilibrium point is the worst possible of all outcomes.

During his last years Johansen came to look at game theory as a general theory of economic behaviour, which contains as special cases the two extremes: completely centralized decision making and perfect competition. His work during these years showed that Johansen as always was a quick learner, and that at his death at the age of 52 he had gained mastery of the relevant parts of game theory. One can only make guesses about the general theories he might have developed if he had been given a few more years to live.

selected works

The obituary published by the Norwegian Academy of Science (*Yearbook* for 1983) lists 11 books and 138 articles written by Johansen, about half of them in Norwegian. Some of the most important in English are:

1958. The role of the banking system in a macro-economic model. *International Economic Papers* 8, 91–110.

1959. Substitution versus fixed production coefficients in the theory of economic growth. *Econometrica* 27, 157–76.

1960. *A Multi-Sectoral Study of Economic Growth*. Amsterdam: North-Holland. 2nd enlarged edn, 1974.

1963. Labour theory of value and marginal utilities. *Economics of Planning* 2, 89–103.

1965. *Public Economics*. Trans. from Norwegian. Amsterdam: North Holland.

1969a. Ragnar Frisch's contributions to economics. *Swedish Journal of Economics* 71, 302–24.

1969b. An examination of the relevance of Kenneth Arrow's General Possibility Theorem for economic planning. *Economics of Planning* 9, 5–41.

1972a. *Production Functions*. Amsterdam: North-Holland.

1972b. On the optimal use of forecasts in economic policy decisions. *Journal of Public Economics* 1, 1–24.

1977. *Lectures on Macroeconomic Planning*. Vol. 1: *General Aspects*. Amsterdam: North-Holland.

1978a. *Lectures on Macroeconomic Planning*. Vol. 2: *Centralization, Decentralization, Planning under Uncertainty*. Amsterdam: North-Holland.

1978b. A calculus approach to the theory of the core of an exchange economy. *American Economic Review* 68, 813–20.

1979. The bargaining society and the inefficiency of bargaining. *Kyklos* 32, 497–522.

1982. On the status of the Nash type of noncooperative equilibrium in economic theory. *Scandinavian Journal of Economics* 84, 421–41.

Bibliography

Aumann, R. and Maschler, M. 1964. The bargaining set for cooperative games. *Annals of Mathematical Studies* 52, 443–76.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

Nash, J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* 36, 48–9.

Solow, R. 1983. Leif Johansen 1930–1982: a memorial. *Scandinavian Journal of Economics* 85, 445–59.

How to cite this article

Borch, Karl H. "Johansen, Leif (1930–1982)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000017> doi:10.1057/9780230226203.0867

The New Palgrave Dictionary of Economics Online

Johnson, Alvin Saunders (1874 – 1971)

Warren J. Samuels
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Clark, J. B.; exploitation; institutional economics; Johnson, A. S.; marginalist economics; protection; welfare economics

Article

Alvin Johnson was born on 18 December 1874 near Homer, Nebraska, and died 7 June 1971 in Upper Nyack, New York. He received a BA (1897) and MA (1898) from the University of Nebraska and a Ph. D. from Columbia University (1902). His varied teaching career included Bryn Mawr, Columbia, Nebraska, Texas, Chicago, Stanford (twice), Cornell and the New School for Social Research of which, in 1919, he was a founder and, beginning in 1923, director. He was president of the American Economic Association in 1936 and of the American Association of Adult Education in 1939. He was active in the struggle for academic freedom and other civil rights and in providing a haven, at the New School, for refugee scholars. His students included Walton Hale Hamilton, Frank H. Knight and James Harvey Rogers.

Johnson also had an active and varied editorial career. He was assistant editor of the *Political Science Quarterly*, founder and editor of *Social Research*, associate editor of the *Encyclopedia of the Social Sciences*, economics editor of the *New International Encyclopedia*, political science editor of the American edition of *Nelson's Encyclopedia*, and on the editorial council of the *Yale Review*. He also was a founder and member of the editorial staff of *The New Republic*.

Johnson, who also published novels and short stories, wrote as an economist on a wide range of theoretical and policy problems. He was also the author of a popular and respected principles text which went through several editions. As a student of (and secretary to) John Bates Clark, Johnson adhered to his marginalist approach to economic theory but combined his neoclassicism with social and institutionalist elements. His dissertation on rent theory stressed inter-product competition and tried to develop a non-Marxian conception of exploitation (30 years prior to Joan Robinson's work). His early economic nationalism encompassed a limited pro-protectionist argument. In various writings he argued that labour-saving machinery did not necessarily raise wages; that forward shifting of the corporate

income tax requires price to be a function only of cost of production, which he deemed not prevalent; and that arguments against the minimum wage were based on static assumptions. He considered that prevailing theory offered only universalist, formal explanations to problems of price formation, whereas he found that price phenomena were also the product of a multiplicity of complex variables, and called for greater realism and empiricism. Following Clark, Johnson also anticipated Pigovian welfare economics arguing, in effect, that public ownership could be a solution to cases in which, because of non-approbriables, marginal private benefits fell short of marginal social benefits. For many years he was active in the land reclamation movement.

In general, Johnson was a cautious reformer, advocating reform within the existing social order through the expansion of non-property rights as both a corollary to the security of property itself and a mark of a progressive economy.

Selected works

1903. *Rent in Modern Economic Theory: An Essay in Distribution*. New York: Columbia University Press.

1909. *Introduction to Economics*. Boston: Heath. 2nd edn, 1922.

1952. *Pioneer's Progress: An Autobiography*. New York: Viking.

1954. *Essays in Social Economics*. New York: New School for Social Research.

How to cite this article

Samuels, Warren J. "Johnson, Alvin Saunders (1874–1971)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000018> doi:10.1057/9780230226203.0868

The New Palgrave Dictionary of Economics Online

Johnson, D. Gale (1917 – 2003)

John M. Antle

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

D. Gale Johnson, an intellectual leader in agricultural economics in the mid- to late 20th century, was an early critic of the parity price concept. His case against agricultural subsidies helped bring agricultural trade policy into the international policy arena. Johnson was a long-time observer of the Soviet Union and Chinese agricultural reforms. His analysis showed that investment in agricultural research, including biotechnology, primarily benefited the poor through lower real food prices. He argued that market and policy failures, not population growth, were the root causes of environmental problems in developing countries.

Keywords

agricultural economics; agricultural growth and population change; agricultural markets in developing countries; agricultural research; agricultural subsidies; agricultural trade policy; American Economic Association; American Farm Economic Association; biotechnology regulation; genetically modified organisms (GMOs); Johnson, D. Gale; National Research Council (US); parity pricing; population growth; price support; property rights; share contracts

Article

David Gale Johnson is widely regarded as one of the intellectual leaders in the field of agricultural economics in the mid- to late 20th century. Born and raised on an Iowa farm, in the early 1940s Johnson was an assistant professor at Iowa State University, where Theodore W. Schultz was department head. Due to a dispute over academic freedom, in 1943 Schultz resigned from Iowa State and moved with several junior faculty, including Johnson, to the economics department at the University of Chicago. Johnson became one of the founders of the Chicago School's ‘oral tradition’ and the workshop system that trained many recognized economists. In addition to his scholarly work and mentoring of students, Johnson served the University of Chicago in various capacities, as Department Chair, Dean of Social

Sciences, and Provost. He was President of both the American Farm Economic Association and the American Economic Association, served on numerous national advisory committees, and was an adviser to many governments and international agencies. He was the editor of *Economic Development and Cultural Change* from 1985 until 2003.

Over the course of his career Johnson's research addressed important topics related to the economics of agriculture in both industrialized countries and developing countries. Johnson emphasized both the welfare effects of agricultural policies on farm and rural people, and their effects on the efficiency of resource allocation within agriculture and between agriculture and other sectors of the economy. His early work focused on domestic agricultural policy design. In his influential 1947 book, *Forward Prices for Agriculture*, he provided both a critique of the parity price concept that dominated agricultural policy debate in the post-war era, and an alternative to it based on understanding the dynamics of the agricultural sector. Another focus of his early work was the role of labour resources in agriculture. His classic 1950 paper on resource allocation under share contracts anticipated much of the debate about their efficiency that was to follow in the 1970s and later. In his equally important 1950 paper on the agricultural supply function, Johnson laid the intellectual groundwork for the extensive literature on agricultural supply that would come in later decades. Importantly, this paper also debunked the claim by J.K. Galbraith and J.D. Black (1938) that the elasticity of agricultural supply is near zero, an argument that was used to rationalize the use of agricultural price supports.

From the 1950s, Johnson's attention moved increasingly to the international policy arena. Perhaps his best-known and most influential work was his 1973 book, *World Agriculture in Disarray*. In this book, Johnson used a general equilibrium model of a growing economy as the basis for his analysis of the impacts that domestic and trade policy interventions have on welfare and resource allocation. Using both theory and data, he showed that output price policies have little or no effect on the returns to the mobile resources engaged in farming (capital and labour), and that it is through the factor markets that returns of farming and other sectors of the economy are equalized. A major conclusion of Johnson's analysis is that the primary effects of subsidy programmes for agriculture is to increase the returns to and price of land, to expand agricultural output, and to induce governments to interfere with international trade. Johnson's work was highly influential in bringing the issue of agricultural trade policy into the international policy arena.

Johnson was recognized as one of the leading experts on agriculture in China, the Soviet Union, and other centrally planned economies. He was one of the first Americans to tour Russian farms in the mid-1950s and point out the inefficiencies of the communal farm system. Four decades later, he would conclude that the cost of the failed Soviet agricultural policy was a major factor in the ultimate demise of the Soviet Union. Johnson and his students were also close observers of the Chinese agricultural economy, the reforms that began in the late 1970s, and the rapid economic growth that followed those reforms.

In the 1980s and 1990s, Johnson focused an increasing amount of his research on the role of agriculture in economic growth during the 19th and 20th centuries, and its relationship to population growth and the improved well-being of the human population in both industrialized and developing countries. His approach to this topic was a direct extension of his vision of agriculture and its role in economic growth. Johnson's investigation of US agricultural incomes in the 1930s and 1940s was essentially a study of the economics of agriculture in a developing economy. Later Johnson applied insights from his earlier work to analyse economic development in an international context. His work on economic development

emphasized the contributions of improvements in agricultural productivity to economic development, the falling real price of food and consequently to improving food security. A further consequence of growing agricultural productivity, combined with the low income elasticity of demand for farm products, was migration out of agriculture.

Throughout his career, Johnson emphasized that the per capita supply of agricultural commodities has been increasing for more than a century, despite the fact that this period has experienced the highest population growth rates in human history. Johnson frequently emphasized that, since at least the 1860s, the long-term trend in the real price of agricultural commodities has been downward, and at an accelerating rate. Between 1866 and 1996, for example, the real price of wheat declined at an annual average rate of 0.89 per cent, but between 1955 and 1996 the annual rate of decline in the real price of wheat was 2.69 per cent. Combined with the fact that the poorest people of the world spend much more of their income on food than richer people, Johnson inferred that gains from agricultural growth had been widely shared and had actually benefited the poor most.

Johnson's ability to disentangle long-run trends from short-term shocks made his advice valuable to governments, but it was often at odds with conventional wisdom. For example, during the 1970s many commentators, politicians and economists thought that a new era of resource scarcity was emerging. Projections were for high and rising farm prices. Farmers were encouraged by farm price-support policy and exhortations from the Secretary of Agriculture to plant ‘fence-row to fence-row’. Johnson was one of the few voices urging caution and more appreciation of the long history of falling real farm commodity prices. Only when prices collapsed in the early 1980s and, inevitably, the budget costs of farm subsidy programmes exceeded all government projections was Johnson's message appreciated. A similar episode, concerning China's role as a grain importer, occurred almost two decades later. Again, Johnson, the source of careful economic logic and sound data analysis, pointed out the sloppy thinking behind the dramatic pronouncements like ‘Who will feed China?’ Within a few years Johnson had been again proven right as China has continued to be a significant grain exporter. Johnson summarized his views of the agricultural supply pessimists as follows; ‘…those who make their living by presenting the future of food supply in very negative terms should be called upon to show conclusively why the remarkable record of the recent past will not continue’ (Johnson, 1999, p. 23).

Johnson also addressed the issue of population growth, and chaired a National Research Council (NRC) committee whose 1986 report, *Population Growth and Economic Development*, proved to be controversial. Contrary to the conventional wisdom of the time (or of today), this report argued that population growth per se was not a major cause of low rates of economic growth or environmental problems in developing countries. The arguments in the NRC report were straightforward implications of economic reasoning. First, the report pointed out the various economic arguments, such as agglomeration economies, scale economies, and the arguments of endogenous growth theory, which suggest that higher populations and higher population densities may increase productivity. Second, the report made the point that environmental degradation is caused not by population per se but by the lack of appropriate institutions, including well-defined and legally defensible property rights. Third, the report emphasized the substitutability of many resources and the role of prices in signalling resource scarcity and in leading to requisite adjustments in resource utilization and innovation.

Given the importance that Johnson attributed to agricultural technology as a source of improvement in human well-being since the Industrial Revolution, he also argued forcefully against anti-technology sentiments. In particular, Johnson expressed concern about the potential negative impact that regulation

of biotechnology could have on the well-being of the world's poor. He outlined these concerns in one of his last publications (Johnson, 2002). He pointed out that the costs of regulating genetically modified organisms, such as bio-fortified foods, will be borne largely by the world's poor, as they are the only ones who spend a significant share of their income on food and would benefit most from an increased availability of micronutrients at a low cost. In addition, he argued that regulations on biotechnology would discourage investment in research, and he noted that biotechnology could bring significant benefits by providing natural substitutes for synthetic pesticides that are costly for poor farmers and have well-known adverse health and environmental impacts.

See Also

- agricultural economics
- agricultural markets in developing countries
- agricultural research
- population and agricultural growth

Selected works

1947. *Forward Prices for Agriculture*. Chicago: University of Chicago Press. Reissued with new foreword, New York: Arno, 1976.

1950. Resource allocation under share contracts. *Journal of Political Economy* 58, 111–23.

1950. The nature of the supply function for agricultural products. *American Economic Review* 60, 539–64.

1973. *World Agriculture in Disarray*, 2d edn. London: Macmillan, 1991.

1997. Agriculture and the wealth of nations. *American Economic Review* 87, 1–12.

1999. North America and the world grain market. In *The Economics of World Wheat Markets*, ed. J. Antle and V. Smith. Wallingford, UK: CABI.

2002. Biotechnology issues for developing economies. *Economic Development and Cultural Change* 51, 1–4.

Bibliography

Galbraith, J. and Black, J. 1938. The maintenance of agricultural production during depression: the explanation reviewed. *Journal of Political Economy* 46, 305–23.

Hoover, D. and Johnson, P. 1996. D. Gale Johnson's contribution to agricultural economics. In *The Economics of Agriculture, Vol. 2: Papers in Honor of D. Gale Johnson*, ed. J. Antle and D. Sumner. Chicago: University of Chicago Press.

National Research Council. 1986. *Population Growth and Economic Development: Policy Questions*. Washington, DC: National Academy Press.

Sumner, D., Antle, J. and Gardner, B. 2004. Contributions of D. Gale Johnson to the economics of agriculture. *Economic Development and Cultural Change* 52, 659–79.

How to cite this article

Antle, John M. "Johnson, D. Gale (1917–2003)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000049> doi:10.1057/9780230226203.0869

The New Palgrave Dictionary of Economics Online

Johnson, Harry Gordon (1923 – 1977)

Jacob A. Frenkel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; brain drain; capital theory; distortions; Econometric Society; flexible exchange rates; free trade; gains from trade; growth and international trade; immiserizing growth; inflation; international monetary economics; international trade theory; Johnson, H. G.; Keynesian revolution; monetary approach to the balance of payments; monetary statistics; monetary theory; monopoly power; optimal tariffs; Phillips curve; retaliation; second best; transfer problem

Article

Harry G. Johnson was born in Toronto, Canada on 26 May 1923 and died in Geneva, Switzerland on 9 May 1977. Throughout his professional career he was a recognized leader of the economics profession in the United States, Britain and Canada, though his influence extended worldwide. He wrote prodigiously: 526 professional scientific articles, 41 books and pamphlets and over 150 book reviews. In addition, he edited 27 books and wrote numerous pieces of journalism. His writings are characterized by creative insights and by a unique capacity to synthesize; both clarify apparently untidy and unyielding masses of seemingly unrelated and abstruse contributions. His impact on the economics profession was enhanced by his ceaseless participations in conferences around the world, and by his willingness to lecture even at the smallest campus or institute, both of which he perceived as a professional obligation. He graduated from the University of Toronto in 1943 and then spent a year at St. Francis Xavier University in Nova Scotia as Acting Professor of Economics (at the age of 20). After military service in the Canadian Infantry, he proceeded to Cambridge, England, obtaining his BA in 1946. He taught in the following year at the University of Toronto, where he also received his MA, specializing in economic history. He then spent 1947–8 at Harvard, followed by a year at Jesus College, Cambridge and then election to a Berry-Ramsey Fellowship at King's College in 1949. He was to remain a Fellow of King's, teaching also at the London School of Economics, until he left Cambridge for the University of Manchester as Professor of Economic Theory in 1956. In 1959 he joined the University of Chicago as Professor of Economics, later becoming the Charles F. Grey Distinguished Service Professor of Economics, and remained there until his death. He was soon to combine the professorship at Chicago

with a chair at the London School of Economics (1966–74) and then with the Graduate Institute for International Studies in Geneva, Switzerland (1976–7).

These shifts in location and the associated changes in intellectual environment shaped his character as a cosmopolitan economist. The years in Cambridge and in Chicago were to be the most significant. For both campuses had, in addition to Johnson himself, remarkable figures in economic science such as Dennis Robertson, Richard Kahn, Nicholas Kaldor and Joan Robinson in Cambridge, and Milton Friedman, George Stigler and Theodore W. Schultz in Chicago. The strong professional and political views and interests of many of these economists must have deepened Johnson's interest in developing theory as a tool of policymaking, and influenced the evolution of his own views and attitudes toward the various approaches to economics.

His writings span the entire range of the economics discipline: from the history of economic doctrines to the economics of the price of gold; from the theory of international commodity agreements to the theory of preferences and consumption; from an analysis of Keynesian economics to the theory of income distribution. They cover, too, the economics of reparations, the theory of productivity, growth and the balance of payments, the theory of tariffs, economic policies for Canada, Britain, the United States and developing countries, the theory of excise taxes, the economics of public goods, the economics of common markets, the economics of monetary reform, the theory of inflation, the theory of index numbers, the theory of nationalism, the state of international liquidity, the theory of advertising, the relationship between planning and free enterprise, the theory of the demand for money, the choice between fixed and floating exchange rates, the economics of basic and applied research, the economics of the brain drain, the economics of poverty and opulence, the theory of distortions, the theory of money and economic growth, the theory of effective protection, the theory of human capital, the economics of bank mergers, an analysis of efficiency of monetary management, the economics of the North–South relationship, an analysis of minimum wages, the economics of student protest, an analysis of the infant-industry argument for protection, the economics of the multinational corporation, the economics of universities, the economics of libraries, the economics of international monetary union, the economics of dumping, an analysis of the role of uncertainty, the economics of smuggling, an analysis of income policy, the economics of speculation, an analysis of mercantilism, the economics of bluffing, an analysis of equal pay for men and women, an analysis of monetarism, an analysis of buffer stocks, the economics of patents, licences and innovations, an analysis of legal and illegal migration, the economics of welfare and reversed international transfers, the monetary approach to the balance of payments, and the monetary approach to the exchange rate.

Four areas of interest and impact were clearly the most important and deserve to be highlighted: (*a*) the pure theory of international trade, (*b*) macroeconomics, (*c*) international monetary theory, and (*d*) economic policies and issues of political economy.

Johnson's work on trade theory constitutes perhaps his most important scientific contribution. His early work in this area is collected in *International Trade and Economic Growth* (1958). This book contains his important and highly original papers in the theory of trade and growth (1953a; 1954). These articles, written at the time of the dollar shortage after the war, were to address the issues from the viewpoint of differential growth of productivity among trading countries, and were to put the whole theoretical discussion into a form that dominated the work of trade theorists for years.

His writings on the general equilibrium analysis of international trade include two influential companion papers on income distribution (1959b; 1960b). In addition, among his notable contributions are those

that belong to what James Meade called the theory of trade and welfare. Four are particularly noteworthy. In chronological order, these are: his classic paper (1953b) on optimum tariffs and retaliation; the cost of protection and the scientific tariff (1960a), building on his earlier work measuring the gains from trade; optimal trade interventions in the presence of domestic distortions (1965a); and the possibility of income losses from economic growth of a small, tariff-distorted economy (1967d).

The paper on optimum tariffs and retaliation addresses the issue of whether a large country which exercises its monopoly power can be made worse off because of foreign retaliatory tariffs. Using a Cournot-type retaliation mechanism, Johnson showed that the country that initially imposes an optimal tariff can wind up better off than under free trade despite foreign tariff retaliation. From the viewpoint of Johnson's evolution as an economist, this paper is notable for two things. First, the early vintage Johnson was intrigued by analytical complexities of the kind that he found much less interesting later. Second, the policy implication of this early vintage analysis was to resurrect the classic case for the exercise of monopoly power by a large country; Johnson's later writings tended to go in the opposite direction, highlighting the great potential cost of departing from truly free trade.

The shift in Johnson's emphasis to the advantages of free trade is seen most directly in his work on the theory of optimal policy intervention in the presence of distortions and in his work on the theory of immiserizing growth. In both instances, Johnson opposed the use of tariffs, utilizing the insights of the theory of second best as applied to problems of trade and welfare.

Finally, the impact of Johnson's paper on the scientific tariff (1960b) was in two areas: (*a*) the measurement of the cost of protection and (*b*) the analytical propositions regarding optimal tariff structures. Johnson's theoretical contributions influenced empirical work on measuring the cost of protection, and on measuring the gains or losses to Britain from joining the EEC. Many of his contributions to the theory of tariffs and commercial policy are reprinted in his *Aspects of the Theory of Tariffs* (1971a).

Johnson's early contributions to macroeconomics were made during his tenure at Cambridge. In ‘Some Cambridge Controversies in Monetary Theory’ (1951b) he clarified the essence of the controversy between the Keynesian and the Robertsonian approaches to key issues like loanable funds versus liquidity preference, the savings–investment identity and the Gibson paradox, and he clearly demonstrated his talent for distilling and integrating complex issues into a coherent framework. His major contributions during that period, however, were his study of the implications of secular changes in the UK banks’ assets and liabilities consequent on the replacement of private by public debt (1951a) and his active participation in the discussion surrounding the revival of monetary policy in the UK. Johnson was critical of the quality of British monetary statistics and in a series of articles attempted to make the case that improved monetary statistics were essential for well-managed monetary policy. In ‘British Monetary Statistics’ (1959a) he published his own labouriously constructed monetary aggregates for the period 1930–57, which stimulated further research.

Johnson's move to the University of Chicago (to which he was invited as the ‘Keynesian’) marked an increased research interest in monetary theory. His major contributions in the early 1960s are ‘The *General Theory* After Twenty Five Years’ (1961), the survey article ‘Monetary Theory and Policy’ (1962b) and ‘Recent Developments in Monetary Theory’ (1963a). These three contributions have since become classics in the field of monetary economics. They established Johnson's reputation as a scholar with a rare breadth of knowledge and with broad scientific and historical perspectives. The survey article is widely acclaimed as a masterpiece in scholarship and its contribution went far beyond

surveying the ‘state of the art’. Johnson's survey suggested a list of issues that would benefit from further research. In retrospect, this list seems to have served as the research agenda in the subsequent 15 years. One of the notable issues on the list was his early scepticism on the stability of the Phillips curve in the face of changes in macroeconomic policies. His evaluations of the major developments in monetary economics (as of the early 1960s) have been influential and perceptive. These developments were the application of capital theory to monetary theory and the shift from static analysis to dynamic analysis. These contributions, along with others, are reprinted in *Money, Trade and Economic Growth* (1962c) and *Essays in Monetary Economics* (1967b) that also include his important contributions to the topic of money and economic growth.

As a result of his interest in the Keynesian revolution and his deep historical perspective, Johnson wrote his controversial article ‘The Keynesian Revolution and the Monetarist Counter-Revolution’ (1971b) which was first presented as the Richard T. Ely Lecture in 1970 and was reprinted in his *Further Essays in Monetary Economics* (1972a). This article is an exercise in the history of economic thought and scientific evolution. His interest in the various aspects of Keynes and his economic thought resulted in a series of provocative articles, some of which appeared posthumously in his joint book with his wife Elizabeth Johnson, *The Shadow of Keynes* (1978b).

Johnson's major criticism of the Keynesian model was its failure to deal with the problem of inflation at the levels of both economic theory and economic policy. He was critical of the ‘sociological’ non-economic theories of inflation, as well as of price controls and incomes policy as remedies for inflation. His analysis of inflation was approached from the perspective of an international economist who views inflation (under a fixed exchange rate regime) as a global phenomenon, a proper analysis of which requires a shift of focus from the concept of monetary developments in individual countries to the concept of the aggregate world money supply. Johnson's view of world inflation is best exemplified in his *Inflation and the Monetarist Controversy* (1972b) which was delivered as the De Vries Lecture in 1971.

Throughout his professional life, Johnson continued his research on international monetary economics. Three articles in 1950 set the stage for what later on became the typical characteristics of his style of research: courage to take positions not always popular with others, the application of relatively simple economic techniques to a new range of problems with resultant important insights, and a passion for geometry as a tool of analysis. He took an early stand against raising the price of gold in terms of all other currencies (1950a), analysed the destabilizing effect of international commodity agreements on the prices of primary products (1950b) and produced an early diagrammatic analysis of income variations and the balance of payments (1950c) – an analysis which was conducted within the then typical Keynesian framework, a framework which he later criticized.

In his writings on the theory of the transfer problem, originally developed in the context of the post-war reparations, Johnson extended earlier work by P.A. Samuelson, L.A. Metzler, F. Machlup and J.E. Meade and demonstrated the potential provided by his philosophy that individual research effort is most productive when it utilizes the work of previous theorists as a foundation for new construction. Johnson's theme was that of ‘continuity and multiplicity of effort’. In ‘The Transfer Problem and Exchange Stability’ (1956) he demonstrated that the problems of transfers and of exchange stability are formally the same and that all the possible methods of correcting balance of payments disequilibrium can be posed in terms of the analytical apparatus of the transfer problem. Almost two decades later (1974) he returned to the analysis of transfers with greater emphasis on the monetary aspects of the

problem.

Johnson's most important contribution to the understanding of international monetary economics is ‘Towards a General Theory of the Balance of Payments’ printed in his *International Trade and Economic Growth* (1958). His insight was the emphasis on the monetary nature of a balance of payments surplus or deficit. ‘[A] balance-of-payments deficit implies *either* dishoarding by residents, *or* credit creation by the monetary authorities’; the former is inherently transitory and the latter is policy induced. As for policy, Johnson coined the distinction between ‘expenditure reducing’ policies and ‘expenditure switching’ policies. The insights contained in this important article are all the more remarkable considering the intellectual environment in the mid-1950s where to a large extent the balance of payments was viewed as a ‘real’ (in contrast with ‘monetary’) phenomenon. This article may be viewed as the intellectual precursor of what would be termed 15 years later ‘the monetary approach to the balance of payments’.

Over the years, Johnson focused increasingly on policy issues with special reference to Canada (1962a; 1963b; 1965c). He supported the move to a flexible exchange rate regime (1969) but recognized, relying on the theory of optimum currency areas, that there are circumstances under which a small country (like Panama) might be better off maintaining a fixed parity.

His analysis of the international monetary system revealed his strength as a realistic political scientist. Monetary reform is not carried out in a vacuum. It is performed by representatives of independent nation states, to whom international commitments are likely to be secondary to national commitments. This view is reflected in his numerous commentaries on international monetary crises, in his doubts about the prospects of a stable European monetary union, in his appraisal of the Bretton Woods system and in his perceptive article ‘Political Economy Aspects of International Monetary Reform’ (1972d). He took a hard line on schemes designed to solve the international monetary problems by methods that channel resources to the less developed countries. He was aware that such a stance might be unpopular but his professional integrity determined his position; in his words, ‘My reason for refusing to endorse such schemes is not that I am opposed to the less developed countries receiving more development assistance but I think that no useful purpose is served by misapplying economic analysis for political ends’ (1967a, p. 8).

As world inflation accelerated in the 1960s Johnson recognized that in a world integrated through international trade in goods and assets, national rates of inflation cannot be fully analysed without a global perspective:

> I have become increasingly impressed in recent years with the conviction that the traditional division between closed-economy and open-economy monetary theory is a barrier to clear thought, and that domestic monetary phenomena for most of the countries with which economists are concerned can only be understood in an international monetary context. (1972a, p. 11)

This perception of world inflation along with the analytical insights from his earlier work ‘Towards a General Theory of the Balance of Payments’ (printed in 1958) paved the way to his work on the monetary approach to the balance of payments which he viewed as the crowning achievement of his career. The intellectual roots of the monetary approach go back to the classic writers (David Hume and David Ricardo) and its early developments can be found in the work of economists associated with the

International Monetary Fund (for example, Jacques Polak). Johnson, however, along with Robert A. Mundell and other members of the International Economics Workshop at the University of Chicago, introduced new and significant dimensions to the approach. Noting that the balance of payments is essentially a monetary phenomenon, he concluded that balance of payments policies will not produce an inflow of international reserves unless they increase the quantity of money demanded or unless domestic credit policy forces the resident population to acquire the extra money wanted through the balance of payments via an excess of receipts over payments. He saw himself as a missionary; and he was to take the lead in developing and disseminating the approach by encouraging and at times guiding the theoretical and empirical research in this field in various centres such as Chicago, London and Geneva. He co-edited with J. A. Frenkel some of the results in *The Monetary Approach to the Balance of Payments* (1976). The evolution of the international monetary system into a regime of flexible exchange rates led to further extensions of the monetary approach and resulted in a new direction of theoretical and empirical research on the economics of exchange rates. Johnson stimulated much of the early research in the area and co-edited with J. A. Frenkel *The Economics of Exchange Rates* (1978) which contains some of the resulting work.

In addition to his theoretical contributions, Johnson wrote profusely also on policy matters. His *Economic Policies Toward Less Developed Countries* (1967a) analyses proposals such as commodity schemes and preferential entry for manufactured exports of the less developed countries. Similarly, his work on the brain drain (1964; 1967c) propounded the view that the brain drain might be welfare-improving for the countries from which it occurred. This is one example of how, in his later years, his analyses increasingly questioned interventionist policies. Thus, the brain drain was beneficial rather than harmful; the multinational corporations were part of a non-zero-sum game and so on. The United Nations Conference on Trade and Development (UNCTAD), which addresses the less-developed countries' problems and demands, and which he had looked on rather benignly in the early 1960s, came under his criticism in several writings as he came to feel that professional economists had allowed themselves to be influenced by their sympathies for the poor countries to the point of being led into empathetic and non-scientific research on trade and development.

It is impossible to conclude the brief survey of Johnson's prolific research without highlighting three other important aspects of his contribution. First, he was a humane social scientist who was interested in understanding social phenomena, in contributing to the improvement of welfare, and in understanding the development of knowledge and technological advances. These qualities are particularly evident in his *On Economics and Society* (1975a) and in *Technology and Economic Interdependence* (1975b). Second, he was a gifted teacher with a deep sense of mission and responsibility. He devoted great effort to the preparation of his lectures and always undertook an extremely heavy teaching load. Some of his lucid and insightful lectures are published in *Macroeconomics and Monetary Theory* (1972c) and *The Theory of Income Distribution* (1973). Third, he was widely respected as an editor, who demonstrated both considerable judgement and a talent for recognizing and encouraging the development of new and original lines of thought. He was devoted to his sustained role as an editor of the *Journal of Political Economy*. He also served on the editorial boards of the *Review of Economic Studies*, *Economica*, the *Journal of International Economics* and *The Manchester School of Economic and Social Research*. Testifying to Johnson's impact on the economics profession is the number of articles devoted to the evaluation of his scientific contributions. Noteworthy in this respect are the special issues of the *Canadian Journal of Economics* (1978) and the *Journal of Political Economy* (1984) (which also

contain a complete bibliography of Johnson's voluminous writings), as well as the entry in the *International Encyclopedia of the Social Sciences* (Bhagwati and Frenkel, 1979), on which this present article draws.

Many honours came Johnson's way. He was invited to deliver many of the prestigious public lectures in economics: the Ely Lecture, the Wicksell Lectures, the De Vries Lecture, the Ramaswami Lecture, the Johansen Lectures and the Horowitz Lectures. He was elected to the presidency of the Canadian Political Science Association (1965–6) and the Eastern Economic Association (1976–7), was Chairman of the (British) Association of University Teachers in Economics (1968–71), and was Vice-President of the American Economic Association (1976). He was a Fellow of the Econometric Society, the British Academy, the Royal Society of Canada, the American Academy of Arts and Sciences, a Distinguished Fellow of the American Economic Association and an honorary member of the Japan Economic Research Center. He was the holder of honorary degrees from St. Francis Xavier University, University of Windsor, Queen's University, Carleton University, University of Western Ontario, Sheffield University and the University of Manchester, and he was awarded the Innis-Gérin Medal of the Royal Society of Canada, the Prix Mondial Messim Habif by the University of Geneva, and the Bernhard Harris Prize by the University of Kiel, Germany, just prior to his untimely death. The Canadian government named him an Officer of the Order of Canada in December 1976: a fitting tribute from his native country for a fully internationalist economist who had brought great distinction to his profession and his discipline.

Selected works

1950a. The case for increasing the price of gold in terms of all currencies – a contrary view. *Canadian Journal of Economics & Political Science* 16(May), 199–209.

1950b. The de-stabilising effect of international commodity agreements on the prices of primary products. *Economic Journal* 60, 626–9.

1950c. A note on diagrammatic analysis of income variations and the balance of payments. *Quarterly Journal of Economics* 64, 623–32.

1951a. Some implications of secular changes in bank assets and liabilities in Great Britain. *Economic Journal* 61, 544–61.

1951b. Some Cambridge controversies in monetary theory. *Review of Economic Studies* 19(2), 90–104.

1953a. Equilibrium growth in an international economy. *Canadian Journal of Economics & Political Science* 19, 478–500. Reprinted in (1958).

1953b. Optimum tariffs and retaliation. *Review of Economic Studies* 21(2), 142–53. Reprinted in (1958).

1954. Increasing productivity, income-price trends and the trade balance. *Economic Journal* 64, 462–85.

Reprinted in (1958).

1956. The transfer problem and exchange stability. *Journal of Political Economy* 64, 212–25. Reprinted in (1958).

1958. *International Trade and Economic Growth*. London: George Allen & Unwin.

1959a. British monetary statistics. *Economica* 26, 1–17.

1959b. International trade, income distribution and the offer curve. *Manchester School* 27(September), 241–60. Reprinted in (1971a).

1960a. The cost of protection and the scientific tariff. *Journal of Political Economy* 68, 327–45.

1960b. Income distribution, the offer curve, and the effects of the tariffs. *Manchester School* 28 (September), 215–42.

1961. The ‘General Theory’ after twenty-five years. *American Economic Review* 51(May), 1–17. Reprinted in (1962c).

1962a. *Canada in a Changing World Economy*. Alan B. Plaunt Memorial Lectures, Carleton University. Toronto: University of Toronto Press.

1962b. Monetary theory and policy. *American Economic Review* 52, 335–84. Reprinted in (1967b).

1962c. *Money, Trade and Economic Growth*. London: George Allen & Unwin.

1963a. Recent developments in monetary theory. *Indian Economic Review* 6(3), 29–69; (4), 1–28.

1963b. *The Canadian Quandary: Economic Problems and Policies*. Toronto: McGraw-Hill.

1964. The economics of brain drain. In *Sixth Annual Seminar on Canadian-American* Relations (papers). Windsor: Windsor University Press.

1965a. Optimal trade intervention in the presence of domestic distortion. In *Trade, Growth and The Balance of Payments: Essays in Honor of Gottfried Haberler*, ed. R.E. Baldwin et al. Amsterdam: North-Holland.

1965b. The theory of tariff structure, with special reference to world trade and development. *Etudes et Travaux de L’Institut Universitaire de Hautes*, Etudes Internationales de Geneva 4. Reprinted in (1971a).

1965c. *The World Economy at the Crossroads*. Oxford: Clarendon Press.

1967a. *Economic Policies Toward Less Developed Countries*. Washington, DC: Brookings Institution; London: George Allen & Unwin.

1967b. *Essays in Monetary Economics*. London: George Allen & Unwin; 2nd edn, 1969; Cambridge, MA: Harvard University Press.

1967c. Some economic aspects of brain drain. An appendix. *Notes on the Effects of Emigration of Professional People on the Welfare of Those Remaining Behind. Pakistan Development Review* 7, 379–411. Reprinted in (1975a).

1967d. The possibility of income losses from increased efficiency or factor accumulation in the presence of tariffs. *Economic Journal* 77(March), 151–4.

1969. The case for flexible exchange rates, 1969. *Federal Reserve Bank of St. Louis Review* 51(6), 12–24. Reprinted in (1972a).

1971a. *Aspects of the Theory of Tariffs*. London: George Allen & Unwin.

1971b. The Keynesian revolution and the monetarist counter-revolution. *American Economic Review* 61 (May), 1–14. Reprinted in (1972a).

1972a. *Further Essays in Monetary Economics*. London: George Allen & Unwin.

1972b. *Inflation and the Monetarist Controversy*. F. De Vries Lectures. Amsterdam and London: North-Holland.

1972c. *Macroeconomics and Monetary Theory*. Chicago: Aldine Publishing.

1972d. Political economy aspects of international monetary reform. *Journal of International Economics* 2, 401–24.

1973. *The Theory of Income Distribution*. London: Gray-Mills Publishing.

1974. The welfare economics of reversed international transfers. In *Trade, Stability and Macro-economics: Essays In Honor of Lloyd A. Metzler*, ed. G. Horwich and P. Samuelson. New York and London: Academic Press.

1975a. *On Economics and Society*. Chicago and London: University of Chicago Press.

1975b. *Technology and Economic Interdependence*. Trade Policy Research Centre. London: Macmillan.

1976. (With J.A. Frenkel, eds.) *The Monetary Approach to the Balance of Payments*. London: George Allen & Unwin; Toronto: University of Toronto Press.

1978. (With J.A. Frenkel, eds.) *The Economics of Exchange Rates*. Reading, MA: Addison-Wesley.

1978b. (With E. Johnson.) *The Shadow of Keynes*. Oxford: Basil Blackwell.

Bibliography

Bhagwati, J.N. and Frenkel, J.A. 1979. Johnson, Harry G. *International Encyclopedia of the Social Sciences*, Biographical Supplement, Vol. 18, New York: The Free Press.

Canadian Journal of Economics, Supplement, 11(4) (1978). *Journal of Political Economy* 92(4) (1984).

How to cite this article

Frenkel, Jacob A. "Johnson, Harry Gordon (1923–1977)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000019> doi:10.1057/9780230226203.0870

The New Palgrave Dictionary of Economics Online

Johnson, William Ernest (1858 – 1931)

S.L. Zabell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

exchangeability; Johnson, W. E.; statistical inference; subjective probability; utility theory

Article

English logician, philosopher, and economic theorist. The son of the headmaster of Llandaff House, a Cambridge academy, Johnson entered King's College in 1879 on a mathematical scholarship (11th Wrangler, Mathematics Tripos 1882; First Class Honours, Moral Sciences Tripos 1883). Initially a mathematics coach, then lecturer on psychology and education at the Cambridge Women's Training College, Johnson later held a succession of temporary positions at Cambridge (University Teacher in the Theory of Education, 1893 to 1898; University Lecturer in Moral Science, 1896 to 1901), until he was elected a Fellow of King's College in 1902 and appointed Sidgwick Lecturer in Moral Science in the University, where he remained until his death.

In the Cambridge of Johnson's day, economics was included among the moral sciences and, as C.D. Broad remarks, ‘it was a subject in which Johnson's mathematical, logical, and psychological interests could combine with the happiest results’ (Broad, 1931, pp. 500–1). Although he lectured on mathematical economics for many years, Johnson wrote only three papers on economics (1891; 1894; 1913), of which only the last, ‘The Pure Theory of Utility Curves’, was published during his lifetime. This latter was, however, an important paper, representing ‘a considerable advance in the development of utility theory’ (Baumol and Goldfeld, 1968, p. 96), and ‘contains several results that should secure for its author a place in any history of our science’ (Schumpeter, 1954, p. 1063n). These include an analysis of utility based on marginal utility ratios, and a proof of the consistency of expenditure and convex indifference curves.

Johnson's aversion to publication has been variously ascribed to his ‘ill health, diffidence, and a very high standard of achievement’ (Broad, 1931, p. 505), and a ‘rooted antipathy to publish anything until he was sure of everything’ (Braithwaite, 1931). Indeed, between the publication of his treatise on *Trigonometry* in 1888, and his three volume work on *Logic* in the 1920s, he published only three papers

on logic (1892; 1900; 1918) in addition to his paper on utility. Despite such a limited output Johnson retained his fellowship at King's (the continuance of which was periodically reviewed), due to the high regard in which he was held by his colleagues.

Johnson nevertheless exerted considerable influence on his colleagues and students at Cambridge through his lectures and personal interaction. One example, among many, is John Neville Keynes, the father of John Maynard Keynes and an eminent logician in his own right. When the senior Keynes was at work on the successive editions of his *Studies and Exercises in Formal Logic*, Johnson would come to lunch regularly to discuss the work; one result was that among the examples at the ends of chapters ‘the hardest, neatest, and most ingenious problems are marked “J”, which means that they were devised by Johnson’ (Broad, 1931, p. 504). Among his students were John Maynard Keynes, Frank Ramsey, Ludwig Wittgenstein, C.D. Broad and Dorothy Wrinch (an early collaborator of Harold Jeffreys).

Nevertheless, it was only after the publication of his three-volume *Logic* (1921; 1922; 1924) – written only after the encouragement and assistance of his students, in particular Naomi Bentwich – that Johnson gained recognition outside Cambridge: honorary degrees from Manchester (1922) and Aberdeen (1926), and membership of the British Academy (1923). The third volume of the *Logic* concludes with a remarkable appendix on ‘education’, in which Johnson introduced his ‘combination’ and ‘permutation’ postulates. The latter of these was none other than the concept of exchangeability, soon to be independently rediscovered by Haag and de Finetti, and employed by the latter as a key element in his theory of subjective probability and statistical inference (Dale, 1985).

Johnson was one of a remarkable group of English intellectuals – most notably Jevons, Edgeworth, Keynes and Ramsey – who combined in varying proportions interests in economic theory and the philosophical foundations of logic, probability, statistics, and scientific inference. For further biographical details, see the obituary notices by C.D. Broad (1931); R.B. Braithwaite (1931); the unsigned A.D. (1932); and the entry on Johnson by Braithwaite (1949) in the *Dictionary of National Biography 1931–1940*. R.F. Harrod (1951) contains scattered references to Johnson.

Johnson's three papers on economics are reprinted, with brief commentary, in William J. Baumol and Stephen N. Goldfeld (1968). The 1891 and 1894 papers were printed for private circulation, and are virtually unobtainable elsewhere. For a critical discussion of the 1913 paper, see F.Y. Edgeworth (1915).

Due in part to what George Stigler has termed Johnson's ‘concise and peculiar’ style, and in part to the appearance of Slutsky's classic paper two years after the appearance of Johnson's *Economic Journal* paper, there has never been widespread recognition of Johnson's achievement in utility theory, and references to his work in the economic literature are few, brief, and scattered; see, for example Joseph A. Schumpeter (1954).

Selected works

1888. *Trigonometry*. London: Macmillan.

1891. Exchange and distribution. *Cambridge Economic Club*, Lent Term, 1–6. Reprinted with brief commentary in Baumol and Goldfeld (1968).

1892. The logical calculus. *Mind* 17, 3–30, 235–50, 340–57.

1894. (With C.P. Sanger.) On certain questions connected with demand. *Cambridge Economic Club*, Easter Term, 1–8. Reprinted, with brief commentary, in Baumol and Goldfeld (1968).

1900. Sur la théorie des équations logiques. In *Bibliothèque du Congrès International de Philosophie*, vol. 3. Paris: Librairie Armand Colin, 1901.

1913. The pure theory of utility curves. *Economic Journal* 23, 483–513. Reprinted, with brief commentary, in Baumol and Goldfeld (1968), 97–124.

1918. The analysis of thinking. *Mind* 27, 1–21, 133–51.

1921. *Logic*. Pt I. Cambridge: Cambridge University Press.

1922. *Logic*. Pt II: *Demonstrative Inference: Deductive and Inductive*. Cambridge: Cambridge University Press.

1924. *Logic*. Pt III: *The Logical Foundations of Science*. Cambridge: Cambridge University Press.

1932. Probability. *Mind* 41, 1–16 (The relations of proposal to supposal), 281–96 (Axioms), 409–23 (The deductive and inductive problems).

In addition to the above, Johnson wrote several critical reviews and a note for *Mind* during the years between 1886 and 1890, and contributed several entries to the original *Palgrave*.

Bibliography

A.D. 1932. W.E. Johnson (1858–1931): an impression. *Mind* 41, 136–7.

Baumol, W.J. and Goldfeld, S.N., eds. 1968. *Precursors in Mathematical Economics: An Anthology*. London: London School of Economics and Political Science.

Braithwaite, R.B. 1931. W.E. Johnson. *Cambridge Review* 52 (30 January), 220.

Braithwaite, R.B. 1949. W.E. Johnson. In *Dictionary of National Biography 1931–1940*. Oxford: Oxford University Press.

Broad, C.D. 1922. *Logic, Part I* by W.E. Johnson. *Mind* 31, 496–510.

Broad, C.D. 1924. *Logic, Parts II and III* by W.E. Johnson. *Mind* 33, 242–61, 369–84.

Broad, C.D. 1931. W.E. Johnson. *Proceedings of the British Academy* 17, 491–514.

Dale, A.I. 1985. A study of some early investigations into exchangeability. *Historia Mathematica* 12, 323–36.

Edgeworth, F.Y. 1915. Recent contributions to mathematical economics. *Economic Journal* 25, 36–63, 189–203 at 41–62.

Good, I.J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Harrod, R.F. 1951. *The Life of John Maynard Keynes*. London: Macmillan.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

How to cite this article

Zabell, S. L. "Johnson, William Ernest (1858–1931)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000020> doi:10.1057/9780230226203.0871

The New Palgrave Dictionary of Economics Online

Jones, Richard (1790 – 1855)

C.A. Gregory
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

English Historical School; inductive method; Jones, R.; Malthus's theory of population; McCulloch, J. R.; peasant economy; Ricardian theory of rent

Article

Jones was born at Tunbridge Wells, Kent. After finishing his studies at Cambridge in 1816 he took holy orders and served as a curate at various places in England for the next decade and a half. During this time he developed an interest in political economy which culminated in his *Essay on the Distribution of Wealth: Vol. I. Rent* (1831a). Soon after publication he was appointed Professor of Political Economy at the newly established King's College, London. In 1835, following the death of Malthus, he was appointed Professor in the East India College at Haileybury and remained there until his death in 1855. He took an active part in the commutation of tithes and served as a commissioner of tithes from 1836 to 1851.

Jones never wrote the proposed second volume of his book and published very little else during his lifetime. The lectures he gave at King's College and East India College, together with other sundry essays and notes, were published soon after his death as *Literary Remains* (edited by W. Whewell, 1859). A persistent theme in Jones's work is a critique of the ahistorical, deductivist methods of the Ricardian school of political economy. He argued for a method he called ‘inductivist’ and was primarily concerned to overturn the Ricardian theory of rent with an historically based theory that distinguished between farmers’ rents and various categories of peasant rents. He also developed a number of theoretical propositions concerning population and technology that contradicted the Malthusian orthodoxy.

Jones's iconoclastic theories were not well received by his contemporaries. McCulloch, in an extended review in the *Edinburgh Review* (1831), dismissed Jones's book as ‘superficial’, ‘lacking in originality’ and ‘signally abortive’ in its attempt to overthrow the Ricardian theory of rent. This opinion was generally held in the 19th century. However, Jones has acquired something of a reputation in the 20th century. Marx's favourable review of Jones in *Theories of Surplus Value* (1905–10, ch. 24) has been a

major contributing factor in this rehabilitation. Marx argued that Jones's theories were a substantial advance on Ricardo because, among other things, Jones had a sense of the historical differences in modes of production and was thus able to conceptualize rent as a form of surplus labour. Jones's theory of peasant rents has also attracted much attention. When his book was reprinted for the first time in 1914, for example, only the first half of his book on peasant rents was republished (see Jones, 1831b). Historians of the role of British economic thought in India have shown that his theory of peasant rent had an important impact on policy debates in India in the latter part of the 19th century (Ambirajan, 1978, p. 175) and have assessed his theoretical contribution vis-à-vis the Ricardian school very favourably (Barber, 1975, ch. 12). Jones's approach to understanding the unfamiliar circumstances of rural India continues to have its advocates even today (Hill, 1982, pp. 14–15).

Miller, in two reviews of Jones's contribution to the history of economic thought, has attempted to assess the reputation to which Jones's orginality entitles him as distinct from the reputation that he has acquired. He finds that ‘Jones did not really have a distinct inductive approach to offer’ (1971, p. 206) and that his theory of rent ‘largely deserved McCulloch's harsh judgement that it lacked originality’ (1977, p. 360).

Originality is a difficult quality to assess because the theoretical perspective of the observer obviously affects any judgement made. Nevertheless it is clear that Jones's rehabilitation owes more to his advocacy of a method than to his theories. But this method is not ‘inductivist’. Jones, as Miller (1971) correctly points out, employs both inductive and deductive reasoning. This is not evidence of a contradiction in Jones's thought, as Miller would argue. Jones's use of the term ‘inductivist’ is a simple misnomer. What is distinctive about Jones's method is the comparative and historical perspective he adopts. This method is now the basis of many non-neoclassical approaches to the economy. Not only does Jones deserve to be regarded as the founder of the English Historical School (Edgeworth, 1899), he also deserves to be regarded as the founder of the English Comparative Economy School because of his contribution to the theory of peasant economy.

Selected works

1831a. *An Essay on the Distribution of Wealth and on the Sources of Taxation. Vol. I: Rent*. New York: Augustus M. Kelley, 1964.

1831b. *Peasant Rents*. New York: Macmillan, 1914.

1859. *Literary Remains*, ed. W. Whewell New York: Kelley, 1964.

Bibliography

Ambirajan, S. 1978. *Classical Political Economy and British Policy in India*. Cambridge: Cambridge University Press.

Barber, W.J. 1975. *British Economic Thought and India 1600–1858: A Study in the History of*

Development Economics. Oxford: Clarendon Press.

Edgeworth, F.Y. 1899. Jones, Richard. In *Dictionary of Political Economy*, ed. R.H.I. Palgrave. vol. 2. London: Macmillan.

Hill, P. 1982. *Dry Grain Farming Families*. Cambridge: Cambridge University Press.

Marx, K. 1905–10. *Theories of Surplus Value*. Pt 3. Moscow: Progress Publishers, 1971.

McCulloch, J.R. 1831. Review of *An Essay on the Distribution of Wealth* by R. Jones. *Edinburgh Review* 54, 84–99.

Miller, W.L. 1971. Richard Jones: a case study in methodology. *History of Political Economy* 3, 198–207.

Miller, W.L. 1977. Richard Jones's contribution to the theory of rent. *History of Political Economy* 9, 346–68.

How to cite this article

Gregory, C. A. "Jones, Richard (1790–1855)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000026> doi:10.1057/9780230226203.0872

The New Palgrave Dictionary of Economics Online

Joplin, Thomas (C. 1790 – 1847)

D.P. O'Brien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank of England; Banking School; Currency School; dual-circulation hypothesis; Joplin, T.; metallic circulation theory; monetary economics, history of

Article

Joplin was born, probably in Newcastle upon Tyne, England, about 1790, and died in Silesia in 1847. He is important both as a banking pioneer and as a monetary theorist.

Joplin's interest in both banking and money was spurred by the banking failures in Newcastle after 1815. The failure of these partnership banks led Joplin, inspired by the joint stock banks over the border in Scotland, into a campaign for the countrywide establishment of such banks in England, and for the loosening of the Bank of England's monopolistic grip upon this form of banking. Working with enormous energy, and brooking no opposition, he established two major joint stock banks; the Provincial Bank of Ireland and the National Provincial Bank of England. But the financial establishment in London, who would have found Joplin's Newcastle accent impenetrable and his rough manners repellent, froze him out, and he received little financial recompense for his achievements, despite the fact that he laid the foundations of the modern British banking system.

His most striking achievements, from an intellectual point of view, lay in the field of monetary economics. Not only did he comment actively and perceptively on monetary policy – and he has a clear claim to be the single most important influence in the development of the lender-of-last-resort doctrine (O'Brien, 2003) – but he developed a macroeconomic model of quite extraordinary sophistication (O'Brien, 1993). This involved a treatment of the circular flow of income, an income multiplier, a model of the transmission of monetary changes, an analysis of aggregate supply, and an explanation for depression and unemployment. In the course of all this he employed a dual-circulation hypothesis; and this led to an analysis of the operation of the monetary system which was fundamentally subversive of 19th-century monetary orthodoxies.

On the one hand, Joplin was quite clear – unlike the members of the Banking School – that causality ran from monetary disturbance, and the balance of payments, to the level of money income. On the other, he

was equally clear, unlike the Currency School, that controlling the note issue of the Bank of England was not the key to price and balance of payments stability.

The Bank of England circulation, he argued, supplied the financial circulation of the country, but this had only a very limited effect on prices. The price level was largely determined by the circulation of the country banks; but, because they held their lending rate rigid and varied the note issue with demand, they failed to vary the note issue in conformity with inflows and outflows of gold resulting from variations in the overall balance of payments on current (the main concern) and capital account.

Yet not only would such a response by the country banks be prudent, given that the note issue was convertible into gold, but it was required if variations in the note issue were to be corrective of external disequilibrium. Joplin was one of the earliest to put forward the theory of ‘metallic circulation’ – the idea that a mixed currency of gold coins and notes should vary in amount exactly as an identically circumstanced fully metallic currency would, in an open economy. Such fluctuation was designed not only to correct the balance of payments, through monetary contraction lowering the level of money income when gold was flowing out, and vice versa, but to act counter-cyclically, thus limiting economic fluctuations (O’Brien, 1995).

Joplin argued that the behaviour of the country banks ensured that metallic fluctuation was not achieved. The solution lay in the introduction of a currency system tied closely to gold which would prevent the perverse behaviour of the country banks.

Joplin's view of the operation of the monetary system involved hypotheses about the relationships between basic macroeconomic building blocks, which differed fundamentally from those of the ruling orthodoxies. But methodologically Joplin was far ahead of his time, writing explicitly about the need to formulate hypotheses and test them. The contrast with the apriorism of Ricardo could hardly be greater. Application of modern econometric techniques to the data collected by Joplin, supplemented by other data, not all of which were available in his lifetime, provides remarkable support for his view of the operation of the monetary system (O’Brien, 1993, ch.13; 1997). In particular, it seems clear that changes in the issues of country bank notes affected the price level, while those of the Bank of England, supposedly at the heart of the money supply, did not; that bullion flows across the exchanges did not respond to variations in the Bank of England note issue but were influenced by the country bank issues; that changes in the country bank note issues were the main source of monetary instability; and that the Bank of England note issue did not act as the high-powered money base of the system.

Joplin was an important economist, one who also offered important insights into the theory of international trade. But he was treated as an outsider, in both banking and intellectual circles. Neither the Banking School nor the Currency School seems to have deigned to take any public notice of him. Inevitably with Joplin, he did not make any attempt to ingratiate himself with others, and was free with accusations of plagiarism, directed not merely at Francis Horner (Joplin suggested the creation of a word ‘hornering’ to describe such activity) but even at Ricardo. Yet all this was extremely unfortunate; there seems little doubt that, had Joplin had more influence, and his ideas been considered more seriously, the catastrophic liquidity crises of 1847 and 1857 in Britain would have been avoided.

See Also

- Bank of England

- banking industry
- gold standard
- monetary economics, history of
- Overstone, Lord [Samuel Jones Loyd]

Selected works

1823. *An Essay on the General Principles and Present Practice of Banking in England and Scotland*, 4th ed. Reprinted with Outlines of a System of Political Economy. London: Baldwin, Cradock and Joy, reprinted New York: Kelley, 1970.

1828a. *Views on the Corn Bill of 1827, and Other Measures of Government*. London: Ridgway.

1828b. *Views on the Currency: in which the Connexion between Corn and Currency is Shown*. London: Ridgway.

1830. *An Examination of the Principles of an Improved System of Banking and the Means of Carrying it into Effect*. London: Ruffy.

1831. *The Plan of a National Establishment for Country Banking and the Principles by which it is Recommended*. London: Ruffy.

1832. *An Analysis and History of the Currency Question*. London: Ridgway.

1839. *On Our Monetary System &c. &c. with an Explanation of the Causes by which the Pressures on the Money Market are Produced, and a Plan for their Remedy, which can be Carried into Immediate Effect, without any Derangement, and with the Approbation of the Banks, both Private and Public, by which the Currency is Issued*. London: Ridgway.

1844. *Currency Reform: Improvement not Depreciation*. London: Richardson.

Bibliography

O'Brien, D. 1991. The macroeconomics of Thomas Joplin. In *Economics, Culture and Education. Essays in Honour of Mark Blaug*, ed. G. Shaw. Aldershot: Edward Elgar.

O'Brien, D. 1993. *Thomas Joplin and Classical Macroeconomics*. Aldershot: Edward Elgar.

O'Brien, D. 1995. Long run equilibrium and cyclical disturbances: the currency and banking controversy over monetary control. In *The Quantity Theory of Money From Locke to Keynes and Friedman*, ed. M. Blaug et al. Aldershot: Edward Elgar.

O’Brien, D. 1997. Monetary base control and the Bank Charter Act of 1844. *History of Political Economy* 29, 593–633.

O’Brien, D. 2003. The lender-of-last-resort concept in Britain. *History of Political Economy* 35, 1–19.

How to cite this article

O’Brien, D. P. "Joplin, Thomas (C. 1790–1847)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000027> doi:10.1057/9780230226203.0873

The New Palgrave Dictionary of Economics Online

Juglar, Cl é ment (1819 – 1905)

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; circular flow; Juglar cycles; Juglar, C.; Mitchell, W. C.; real business cycles; Schumpeter, J. A.; Wicksell, J. G. K.

Article

Like his rather more illustrious compatriot Francois Quesnay, Juglar is an example of a physician turned economist. The circular flow of economic life – which it is often said Quesnay saw in terms of an analogy to the circulatory system – in Juglar's work seems have as its counterpart the view of the economic process as one of quasi-rhythmical variations between good and bad trade. This simple idea has been of profound importance in the study of alterations in the conditions of economic prosperity ever since. Both Wesley Clair Mitchell and Joseph Schumpeter in their classic studies of business cycles (in 1927 and 1939 respectively) credit Juglar's contribution as having been seminal in the field. For Mitchell, it was Juglar's recognition of the *cyclical* character of economic crises that established him as a pioneer (1927, p. 452); for Schumpeter it was Juglar's perception of how theory, statistics and history ought to contribute to the study of industrial fluctuations (1939, pp. 162–3). There is something to each of these claims, but it should not be forgotten that other authors had also done much in both of these areas – one may mention Samuel Jones Loyd, John Wade and Amasa Walker. As theorists of industrial fluctuations, of course, Sismondi, Rodbertus and Marx would also need to be mentioned.

Juglar practised as a physician until 1848. His first work in the social sciences was on the cyclical pattern of birth, death, and marriage rates in France, and it appeared in the *Journal des Economistes* in October–December 1851 and January–June 1852. He moved on to examine the discount policy of the Bank of France and published his findings in the *Annuaire de léconomie politique* for 1856 and in the *Journal des Economistes* for April–May 1857. In 1852 he was elected into the Société d'Economie Politique and he was one of the founders of the Société de Statistique de Paris in 1860. In 1868 he published an account of the policies and practices of the French monetary authorities and their effects on the exchanges.

There is, however, little doubt that Juglar's most important work on business cycles is his *Des crises*

commerciales et de leur retour périodique en France, en Angleterre, et aux Etats-Unis, first published in 1860. Juglar's analysis of crises is essentially a monetary one – protracted periods of inflation and expansion are brought to an end when the banking system initiates a contraction in the face of unacceptable pressures on its specie reserves. This is very like the story Wicksell was later to tell, but without the sophistication of Wicksellian theory. Subsequent theories of the business cycle, which attributed the process to ‘real’ causes, were critical of this aspect of Juglar's argument. The observed periodicity of the cycle – of nine to ten years – is commonly known in the applied literature on business cycles as a Juglar cycle.

Selected works

1860. *Des crises commerciales et leur retour périodique en France, en Angleterre, et aux Etats-Unis*. Paris: Guillaumin. 2nd edn, 1889.

1868. *Du change et de la liberté d’émission*. Paris: Guillaumin.

Bibliography

Mitchell, W.C. 1927. *Business Cycles: The Problem and its Setting*. New York: NBER.

Schumpeter, J.A. 1939. *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process*, 2 vols.New York: McGraw-Hill.

How to cite this article

Milgate, Murray. "Juglar, Clément (1819–1905)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000028> doi:10.1057/9780230226203.0874

The New Palgrave Dictionary of Economics Online

just price

Odd Langholm
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Albert the Great; Alexander of Hales; Aquinas, St. T.; Aristotle; Duns Scotus, J.; just price; market price; Odonis, G.; Olivi, P.

Article

The idea of the just price is associated primarily with scholastic economics. The schoolmen suggested two ways of estimating the just price, with reference to cost and with reference to the market. The former originated in reply to some of the Church fathers, who claimed that merchants reaped an unjust profit from the toils of others. Alexander of Hales (d. 1240), Peter Olivi (d. 1298), John Duns Scotus (d. 1308) and other schoolmen together compiled a catalogue of cost elements incurred in trade: transport, storage, risk, costly training, professional expertise and diligence, as well as support of the merchant and his family. The cost estimate was confirmed by the schoolmen's interpretation of the strange formula of exchange appearing in Aristotle's *Nicomachean Ethics*. In Book V, on justice, Aristotle presents a cast of characters – a builder, a shoemaker, a farmer, a doctor. ‘As a builder is to a shoemaker, such and such a number of shoes must be to a house’ (1973, 5: ll33a22–3; author's translation). What could this mean? Albert the Great (d. 1280), the first Latin commentator, and numerous followers, suggested that it might mean equality in proportion to the labour and expenses incurred in the production of the goods offered in exchange. Albert did not merely indicate that economic exchangers deserve cost coverage but that society requires it. If a carpenter (another of Aristotle's characters) is not paid for a bed as much as it costs him to make it, he will stop making beds – a medieval hint about the law of cost. Scotus says much the same about merchants in general. If no one will be a merchant, the authorities must appoint functionaries and pay them accordingly.

The exchange formula in the *Ethics* also gave rise to the market estimate of the just price. According to Aristotle, human need is the cause of exchange. Thomas Aquinas (d. 1274) suggested, and many others agreed, that need is not only the cause of exchange; it is a measure of the value of goods in exchange as well. This could not apply to individual need, as John Buridan (d. c. 1360) points out. It would follow that a poor man should pay more for a measure of corn than a rich man because his need is greater. It

must apply to common need. In the words of Henry of Friemar (d. 1340), the need that measures goods in exchange ought not to be taken partially with regard to this or that person, but universally with regard to the whole community. This value estimate was challenged and confirmed with reference to Roman law. Some early manuscripts of the *Digest* (a compendium of Roman law compiled in the 6th century ad) contain a gloss stating that a thing is worth the amount at which it can be sold. Seemingly granting unconditional economic power to those in possession of scarce goods needed by others, this maxim was modified by the principle of commonality precisely in line with the Aristotelian formula. In a gloss to the *Digest*, the Romanist Azo (d. c. 1220) states that a thing is worth the amount at which it commonly can be sold. The canonist Laurence of Spain (d. 1248) confirmed this interpretation, which earned universal acceptance among the schoolmen. A common estimate, based on common need, can mean several things. The schoolmen tended to associate it with the market or, more precisely, with the common, competitive market price. Albert the Great explicitly defined the just price as ‘that price at which the good can be valued according to the estimation of the market at the time of the contract’ (1894, 16: 46, p. 638; author's translation).

The schoolmen envisaged no conflict between the cost and market estimates of the just price. That conflict is of a much more recent date. The two estimates were used interchangeably and are perhaps best understood as complementary and mutually supportive criteria when the market did not function properly. When it did, cost had to adapt to the market anyhow. Does the fact that these estimates were thus associated mean that the medieval schoolmen anticipated modern value theory? Certainly not, but there are suggestions worth noting in some of the *Ethics* commentaries, where the two principles are textually close. Note may be made of Gerald Odonis (d. 1349), an exceptionally perceptive and original thinker, who applied both principles to the payment of professional services rather than commodities. This is a marginal case, but it points to an important generalization. A price obtained when the market did not function properly owing to monopoly or other market irregularities was held to be unjust because it involved economic coercion. Free consent to the price on the part of both the seller and the buyer was a fundamental requirement of justice in exchange.

See Also

- Albert the Great, Saint Albertus Magnus
- Aquinas, St Thomas
- scholastic economics

Bibliography

Albert the Great, Saint Albertus Magnus. 1891. Ethica. In *Opera Omnia*, vol. 8, ed. A. Borgnet. Paris: Louis Vives.

Albert the Great, Saint Albertus Magnus. 1894. Super Quartum Sententiarum. In *Opera Omnia*, vol. 29, ed. A. Borgnet. Paris: Louis Vives.

Aquinas, St.T. 1969. Sententia Libri Ethicorum. In *Opera Omnia*, vol. 47, ed. R.A. Gauthier. Rome:

Leonine Commission.

Aristotle. 1973. Ethica Nicomachea. In *Aristoteles Latinus*, vol. 26, ed. R.A. Gauthier. Leiden.

Baldwin, J.W. 1959. The medieval theories of the just price. *Transactions of the American Philosophical Society* NS 49, 15–92.

de Roover, R. 1958. The concept of the just price. *Journal of Economic History* 18, 418–34.

Duns Scotus, J. 1894. *In Quartum Librum Sententiarum*, in *Opera Omnia*, vol. 18. Paris: Louis Vives.

Langholm, O. 1979. *Price and Value in the Aristotelian Tradition*. Oslo: Universitetsforlaget.

Langholm, O. 1992. *Economics in the Medieval Schools*. Leiden: Brill.

Langholm, O. 1998. *The Legacy of Scholasticism in Economic Thought*. Cambridge: Cambridge University Press.

How to cite this article

Langholm, Odd. "just price." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000031> doi:10.1057/9780230226203.0875

The New Palgrave Dictionary of Economics Online

Justi, Johannn Heinrich Gottlob Von (1720 – 1771)

K. Tribe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Justi, J. von; cameralism

Article

Born in modest circumstances in a Thuringian village in late December 1717, Justi is best known today as one of the architects of mid-18th-century cameralism. He studied law at Wittenberg in 1742–4, then embarked upon a career of literary activity and state service. In 1750 he was appointed tutor to the son of Haugwitz, reforming administrator of Maria Theresa, archduchess of Austria, and then later the same year to a post at the Viennese Theresianum, where he lectured on ‘commerce and public economics’ to civil servants of noble descent. The lectures were later published in 1755 under the title Staatswirthschaft, by which time Justi had made a hasty departure from Vienna and taken up a new post as Director of Police in Göttingen. This was associated with a transfer of political allegiance from Vienna to Berlin, which new allegiance forced him to leave Göttingen in 1757 when occupation by the French, allied with the Austrians, threatened. For several years he lived from his writings, before being appointed Prussian Inspector of Mines, Glass and Steel Works in 1765. Embroiled in a financial scandal of obscure origin in 1768, he was imprisoned and died in the fortress at Küstrin in 1771.

Justi's literary output and journalistic activity was extensive, if repetitive, ranging over aesthetics, philosophy, history, politics and economics. His major work is the *Staatswirthschaft* (1755; 1758), literally ‘state economy’, which details the manner in which a ruler should govern his lands to assure the ‘happiness of the state’ and a flourishing population. Cameralism had begun as a systematization of the principles followed by the administrators of the ruler's domains. In Justi these principles are identified with the management of the absolutist state, in which economic welfare is conceived identified as the path to political power. Welfare and wealth are produced by good government and the implementation of ‘good police’ – *Polizei* in the 18th-century sense of regulations covering all aspects of social action and public order. The ‘science of police’ is covered in a further textbook, *Grundsätze der Policey-Wissenschaft* (1756; 1759; 1782), which Justi claimed to be the first systematic treatment of the subject, and which was in fact republished after his death in a revised edition. Justi's influence was strong during

the later 18th century, diminishing only with the general decline of cameralism at the turn of the 19th century.

See Also

- cameralism
- German economics in the early 19th century
- Sonnenfels, Joseph von

Selected works

1755, 1758. *Staatswirthschaft*, 2 vols. Leipzig.

1756, 1759, 1782. *Grundsätze der Policey-Wissenschaft*. Göttingen.

1759. *Der Grundriss einer guten Regierung*. Frankfurt am Main.

How to cite this article

Tribe, K. "Justi, Johannn Heinrich Gottlob Von (1720–1771)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000032> doi:10.1057/9780230226203.0876

The New Palgrave Dictionary of Economics Online

justice (new perspectives)

Bertil Tungodden

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides a survey of recent normative work on justice. It shows how the concern for distributive equality has been questioned by the idea of personal responsibility and the idea that there is nothing intrinsically valuable in levelling down individual benefits. It also discusses the possibility of combining a concern for the worse off with a concern for Pareto efficiency, within both aggregative and non-aggregative frameworks, which includes a discussion of the arguments of prioritarianism, sufficientarianism, and welfarism. Finally, the article briefly reviews the modern literatures on rights-based reasoning, intergenerational justice and international justice.

Keywords

altruism; behavioural economics; choice; compensation; consequentialism; difference principle; egalitarianism; equality; equalization; evolutionary economics; fair allocation; Fairness; Harsanyi, J. C.; independence of irrelevant alternatives; indexing impasse; infinite horizons; intergenerational justice; international justice; interpersonal utility comparisons; justice; leximin; libertarianism; maximin; Nozick, R.; Pareto efficiency; Pareto indifference; Pareto principle; poverty; poverty alleviation; preferences; primary goods; prioritarianism; Rawls, J.; responsibility; self-ownership; Sen, A.; social welfare function; sufficiency; sufficientarianism; tyranny of aggregation; utilitarianism; veil of ignorance; welfarism; well-being

Article

Modern thinking on justice has been strongly motivated by the work of Rawls (1971; 1993). Rawls not only developed a prominent theory of justice that has been extensively analysed, he also expressed in a very powerful way the fundamental role justice has to play in the evaluation of social arrangements. Rawls argued that justice is the first virtue of social institutions, as truth is of systems of thought. ‘A theory however elegant and economical must be rejected or revised if it is untrue; likewise, laws and

institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust' (1971, p. 3).

The fundamental problem is that there are many divergent views of what constitutes a just society, and thus many divergent views of what are just social arrangements. Rawls introduced the notion of a reflective equilibrium, which, roughly speaking, is attained when our principles and judgements of justice coincide. The normative literature on justice can be seen as part of a process towards such a reflective equilibrium, where the aim is to attain a better understanding of both the consequences and the underlying foundation of various possible conceptions of justice.

Further understanding of different conceptions of justice is also important in the positive analysis of individual behaviour, because it is by now well-established that people in many situations are motivated by fairness considerations (Camerer, 2003). There is a substantial literature in behavioural economics that study in more detail what kind of fairness norms motivate people and to what extent these fairness norms survive in different settings (Konow, 2003), and also an important literature in evolutionary economics that aim at understanding why our concern for justice has evolved (Binmore, 2005; Skyrms, 1996; 2003).

This article is a sequel to Sen's entry on justice in the first edition of *The New Palgrave: A Dictionary of Economics* (reproduced in this edition), where Sen argues for a broader view of justice than what is captured by utilitarianism (see also Sen, 1979). Sen views utilitarianism as the amalgam of three distinct principles, namely, welfarism, sum-ranking, and consequentialism, and he shows how each of them was contested in the early modern literature on justice. In this article, I survey how these questions have been dealt with in recent normative work on justice. In particular, I focus on the role of distributive equality. Sen argued convincingly for the need to take explicit note of inequalities in the distribution of utilities or some other equalisandum, and the standard welfare economic view is presently that justice requires a trade-off between equality of utility and the sum of utility. Interestingly, however, the concern for distributive equality has been questioned from different perspectives. First, it has been argued that distributive equality neglects the role of personal responsibility, and, second, it has been argued that distributive equality legitimizes the intrinsic value of levelling down utilities. I review each of these arguments before I move on to the classical question of how to incorporate equality or a concern for the worse off in an aggregative theory of justice. Any aggregative theory of justice, however, faces what I call the tyranny of aggregation, and therefore, inspired by Rawls (1971), there have been many attempts to establish a non-aggregative framework that combines a concern for equality with a concern for Pareto efficiency. I discuss some of the most prominent non-aggregative perspectives and also some recent developments on rights-based non-consequentialistic reasoning. Finally, I review briefly the growing literature on intergenerational and international justice, which raises interesting questions on how to deal with individuals who are in asymmetric relationships to each other.

Distributive equality and personal responsibility

Modern egalitarian theories of justice seek to combine the values of equality and personal responsibility. The contemporary focus on this relationship can be traced back to Rawls (1971), but it has historical roots both in the US Declaration of Independence (1776) and the French Declaration of the Rights of Man and Citizen (1789). The American and French societies developed in rather different directions,

though, and, as noted by Nagel (2002, p. 88), ‘what Rawls has done is to combine the very strong principles of social and economic equality associated with European socialism with the equally strong principles of pluralistic toleration and personal freedom associated with American liberalism, and he has done so in a theory that traces them to a common foundation’. The ideas of Rawls have been developed further, notably by Dworkin (1981), Arneson (1989), Cohen (1989), Kolm (1996), Roemer (1993; 1996; 1998), van Parijs (1995), Bossert (1995), Fleurbaey (1995a; 1995b), Bossert and Fleurbaey (1996), and Fleurbaey and Maniquet (1996;1999), where the main achievement has been to include considerations of personal responsibility in egalitarian reasoning. The two basic conditions put forward in this literature are *the principle of equalization* and *the principle of responsibility*. The principle of equalization states that if two persons have exercised the same level of responsibility, then justice demands that they should have the same outcome in the morally relevant space. The principle of responsibility states that inequalities due to different levels of responsibility can be justified.

A fundamental question is whether the two basic principles can be combined in a coherent theory of justice. Dworkin (1981) proposes the idea of a hypothetical insurance scheme, where each person makes her choice of insurance behind a thin veil of ignorance where everyone knows his or her own preferences and is in the possession of the same amount of resources. The equilibrium outcome in this insurance market forms then the basis for the just compensation of disadvantages in the actual world. The proposal of Dworkin has been criticized by Roemer (1985), who argues that, if individuals maximize their expected utility in the insurance market, they insure against states in which they have low marginal utility. If low marginal utility happens to be the consequence of some inborn handicaps, then the hypothetical market will tax the disabled for the benefit of the others. Hence, if we do not want to hold people responsible for their handicaps, then this approach violates the principle of equalization in the actual world, even though it satisfies it if we define responsibility in relation to the choices behind the veil of ignorance. For a further discussion of this issue, see Dworkin (2002), Fleurbaey (2002) and Roemer (2002a).

Bossert (1995) and Bossert and Fleurbaey (1996) study the compatibility of the principle of responsibility and the principle of equalization within a model where pre-tax income of each person is determined by a vector of factors and where we hold people responsible for some of these factors (for example, effort) and not for others (for example, family background). They show that, if the principle of responsibility is interpreted as saying that people should be held fully responsible for the actual consequences of changes in their behaviour, then it cannot be combined with the principle of equalization. However, such an interpretation of the principle of responsibility can be questioned because in many cases it may imply that inequalities reflect differences that we do not want to hold people responsible for, including their inborn talent (Tungodden, 2005). However, there are many other possible interpretations of the principle of responsibility which can be combined with the principle of equalization (Fleurbaey and Maniquet, 2008). One possibility is captured by the egalitarian equivalent mechanism, where people face a given reward scheme for their choice of effort and then share equally the deficit or surplus that follows from this scheme.

A basic insight from this literature is that, if we want to satisfy the principle of equalization, then justice requires that people should face the same consequences from the same kind of behaviour. However, this implies that there is a general tension between the just allocation and Pareto efficiency, where the latter requires that people should face the actual consequences of their behaviour. This tension is not present if

one accepts a weaker version of the principle of equalization, which requires complete equalization only for some level of responsibility (Kolm, 1996). Such an approach is consistent with holding people fully responsible for the actual consequences of changes in their behaviour, as illustrated by the conditional egalitarian mechanism introduced by Bossert and Fleurbaey (1996).

Another interesting insight that follows from this framework is that an income tax system may be unjust in two different ways. First, it may be unjust because it does not equalize sufficiently among people exercising the same level of responsibility. Second, it may be unjust because it equalizes too much between people exercising different levels of responsibility. Within the more standard framework of welfare economics, where considerations of responsibility are not introduced, the second type of injustice is usually overlooked.

The location of the responsibility cut is essential in any application of a responsible-sensitive egalitarian theory, which is most easily seen by noticing the implications of two extreme cases. No redistribution would be justifiable if all factors are responsibility factors, while, ideally, outcomes should be equalized completely if all factors are non-responsibility factors. If there are both responsibility factors and non-responsibility factors, however, then the ideal level of redistribution also depends on the degree of inequality in the non-responsibility factors. However, it is in general not the case that the ideal level of redistribution is lower if the differences in some non-responsibility factor are eliminated or if we move to a situation where people are held responsible for more factors. This will be the case only if there are no negative correlations between various non-responsibility factors in society (Cappelen and Tungodden, 2006).

The standard way of defining the responsibility cut is to rely on the distinction between choice and circumstances, where people are held responsible for their choices but not for their circumstances (Cohen, 1989). However, this approach is controversial and raises metaphysical questions about the basis for our choices (Dennet, 2003). Alternatively, we may think of the responsibility cut in political terms, whereby people are assigned responsibility for a particular set of factors without relying on a particular metaphysical view of individual choices (Fleurbaey, 1995a). The question of where to locate the responsibility cut then mirrors the political debate on redistribution, where right-wingers argue that people should be held responsible for a large fraction of the factors influencing their lives, whereas left-wingers hold individuals responsible for a smaller set of factors.

A further problem in applying this framework is how to obtain a more precise measure of the degree of responsibility a person has exercised. To simplify, suppose that we consider a case where only labour effort and talent affect outcome, and where we do not want to hold people responsible for their talent. Roemer (1993; 1996; 1998) proposes that we partition the population into talent groups, and then consider two individuals identical in terms of responsibility if they are at the same percentile of the labour effort distribution within their class of talent. This approach can be generalized to any number of responsibility and non-responsibility factors by studying conditional distributions more generally. Roemer combines this framework with a maximin interpretation of the principle of equalization and a utilitarian interpretation of the responsibility principle. His proposal equalizes as much as possible among people who have exercised the same level of responsibility, but rewards individuals for additional labour effort only if this maximizes the total amount of utility (or some other equalisandum) within the sub-population consisting of those who receive the lowest level of utility at each percentile of labour effort level. In sum, this provides us with a complete theory of justice, not only the ideal solution,

and Roemer (2002b) illustrates how this framework may be applied in studying redistribution policies. Alternative versions of Roemer's framework are studied in Van de gaer (1993) and Ooghe, Schokkaert and Van de gaer (2006).

Distributive equality and prioritarianism

A fundamental critique of distributive equality has been launched in the debate on prioritarianism and egalitarianism (Parfit, 1995; Temkin, 1993; 2000; Scanlon, 2000), where it has been questioned whether even in situations where people have exercised the same level of responsibility we should find distributive equality intrinsically valuable.

Scanlon (2000) argues that equality very seldom seems to be what we care about and that our concern for equality in most cases can be traced back to other fundamental values. We care about a reduction in inequality because, among other things, it contributes to the alleviation of suffering, the feeling of inferiority, and the dominance of some over the lives of others. Parfit (1995) questions the intrinsic value of equality by appealing to *the levelling down objection*. A reduction in inequality can take place by harming the better off in society without improving the situation of the worse off. If equality is intrinsically valuable, then this must be good in some respect. However, to harm everyone cannot be good *in any respect*, and hence inequality cannot be intrinsically bad.

Parfit (1995) suggests that there is an alternative view, what he calls the priority view, which better captures our concern for the worse off and avoids the levelling down objection. Parfit defines prioritarianism as the view that, the worse off people are, the more important it is to benefit them. This, however, is an imprecise statement which does not clearly set apart prioritarianism from egalitarianism, and it has been questioned in the literature whether it is at all possible to distinguish these two perspectives (Broome, 2007). As pointed out by Fleurbaey (2007), a prioritarian view will always coincide with an egalitarian view that cares both for total utility (or well-being) and equality, and which measures inequality with the same index that is implicit in the prioritarian view. However, it can be argued that the two perspectives reflect different ways of justifying priority to the worse off. The prioritarian justification focuses on the absolute circumstances of the worse off, while the egalitarian justification focuses on the relative circumstances of the worse off (Tungodden, 2003).

Justice, welfarism and aggregation

A substantial literature has studied how to combine a concern for distributive equality or the worse off with other values, in particular Pareto efficiency. This raises two core questions. First, we need to establish a metric of individual advantage and, second, we need to determine how much weight to assign to distributive equality relative to other values.

Much of this work has rested on the assumption of welfarism, which states that the social ranking of alternatives must depend only on the utility levels of individuals in these alternatives (Arrow, 1951; Sen, 1970a; Bossert and Weymark, 2002). Welfarism may be assumed as a basic assumption or it may be derived from the more fundamental principles of Pareto indifference and independence of irrelevant alternatives (d'Aspremont and Gevers, 1977). There has been a huge literature criticizing welfarism. On

the one hand, it has been argued that welfarism contains an unsatisfactory representation of individual advantage. On the other hand, it has been claimed that it is impossible to apply welfarism in practice. We may label these the *pragmatic* and the *fundamental* arguments against welfarism.

The underlying idea of the pragmatic argument is that we 'must respect the constraints of simplicity and availability of information to which any practical policy conception [of justice] is subject (Rawls, 1993, p. 182). Welfarism implies that interpersonal comparisons should be based on comparisons of preference satisfaction, which in general is considered to be non-observable. Thus, the welfaristic framework does not provide a practicable public basis for considerations of justice.

The fundamental critique of welfarism is concerned with the substantive claims of this framework. Rawls (1971; 1993) argues that utility or well-being is not a relevant feature of states of affairs. Appropriate claims should refer to an idea of rational advantage that is independent of any particular comprehensive doctrine of the good, and for this purpose Rawls suggests a list of primary goods. Sen (1985; 1992a) defends the focus on well-being in social choices, but he argues against the idea of well-being implicit in welfarism. Sen introduces the framework of functionings and a capability set, where functionings are the various things that a person may value doing or being (for example, being adequately nourished, free from avoidable disease, and able to take part in the life of the community) and the capability set is the set of alternative functioning vectors available to her.

The proposals of Rawls and Sen differ, but formally they are closely related; social alternatives are characterized by a vector of valuable elements assigned to each individual. However, this raises the fundamental question of how to trade off gains and losses in the various dimensions for each individual. On possibility, as first suggested by Rawls (1971), is to establish an objective index as the basis for interpersonal comparisons in a theory of justice. The problem with this approach, as observed by Gibbard (1979), is that this in general will violate the Pareto principle. Some people will have preferences that are in disagreement with how the index implicitly makes the trade-off, and thus we face what is commonly named the indexing impasse (Sen, 1996a; Plott, 1978; Blair, 1988; Arneson, 1990). Sen suggests that the indexing impasse follows from not taking note of the citizens' preferences when constructing the index, and he argues in favour of an intersection approach which articulates only those judgements that are shared implications of all the preferences present in society. However, as shown by Fleurbaey and Trannoy (2000), Brun and Tungodden (2004), and Pattanaik and Xu (2007), this approach does not solve the problem. In any society where people have heterogenous preferences, the intersection approach runs into a conflict with the Pareto principle.

A related argument has been put forward by Kaplow and Shavell (2001; 2002). They argue that any notion of fairness or justice that implies a violation of the Pareto indifference principle will also imply a violation of the standard Pareto principle if we accept a minimal continuity condition. They apply this insight to argue against any notion of fairness or justice that does not rely on individual utilities. However, there are alternatives to welfarism that are consistent with the Pareto principle (Fleurbaey, Tungodden and Chang, 2003). In particular, there is a literature on fair allocation which exploits the fact that with a richer description of the social alternatives we may apply considerations that rely on the shape of the indifference curves of individuals when establishing a justice ranking (Fleurbaey, 2003; Fleurbaey, Suzumura and Tadenuma, 2005; Fleurbaey and Maniquet, 2006). This approach violates Arrow's independence of irrelevant alternatives, and thus shows that this condition is far from innocent in an analysis of justice.

If we now turn to the question of how much weight to assign to the worse off, then the answer may depend on our assumptions about the informational framework (Bossert and Weymark, 2002). If we assume that there is a one-dimensional measure of individual benefits (which may be utility) and no constraints on interpersonal comparability, then there is a vast number of theories of justice satisfying the Pareto principle. Hence, we need to impose other ethical conditions on the justice ranking in order to choose among the set of possible theories. One possibility is to appeal to conditions that only cover two-person situations, that is, situations where only the benefits of two persons differ in a comparison of two social alternatives, and it turns out that these conditions are extremely powerful in an analysis of this kind (d'Aspremont, 1985). By way of illustration, the utilitarian and the leximin ranking follows almost directly from assuming two-person utilitarianism and two-person leximin within such a framework. Within this informational framework, one may also show that any aggregative theory faces what we may name the tyranny of aggregation (Tungodden, 2003), whereby the interests of the worse off may be outweighed by the interests of a sufficiently large number of better off, even though the gain of each of the better off is *infinitesimal*. Although the tyranny of aggregation is well-known in the context of utilitarianism, it is important to note that this applies to any aggregative approach to social choices, including any aggregative prioritarian rule. This raises the question of whether an aggregative approach can constitute the basis of a theory of justice.

Justice and non-aggregation

A core element in Rawls's theory of justice is precisely that each person possesses an inviolability that makes aggregation impermissible. This is expressed both in the absolute priority assigned to the fulfilment of basic liberties and in the difference principle. Rawls aimed at justifying a non-aggregative approach by a constructive theory whereby people choose fairness principles behind a veil of ignorance. An extensive literature has questioned this conclusion, and following Harsanyi (1955) it is commonly argued that the veil of ignorance approach implies some version of utilitarianism (Weymark, 1991; Broome, 1991; Mongin, 2001; Mongin and d'Aspremont, 2002).

However, there are other ways of justifying a non-aggregative approach to distributive justice. One possibility is to combine a concern for equality promotion with a concern for Pareto efficiency (Barry, 1989). Tungodden and Vallentyne (2005) have investigated this approach, where the basic idea is that distributive conflicts ought to be solved by choosing the more equal distribution. They show that any consistent theory of justice satisfying this approach has to assign strict priority to the worst off in society. This result is closely related to the formal results developed by Hammond (1975) on extreme inequality aversion, and questions the claim of some philosophers that the leximin principle is not consistent with a concern for equality (McKerlie, 1994).

Another interesting non-aggregative approach has been suggested by Nagel (1979) and Scanlon (1982; 1998). They both argue that justice requires pairwise comparisons of individual claims, where the just solution is to satisfy the individual with the most urgent claim. In other words, they reject the argument that the number of persons with a particular claim should count (Taurek, 1977). They do, however, defend the view that, in order to measure the urgency of a claim, one should take into account both gains and losses and the absolute circumstances of an individual. Hence, the aim is to outline a theory of

justice that lies in the middle ground between leximin and the standard aggregative perspective. It turns out, however, that it is impossible to establish such a middle ground within a framework satisfying some basic consistency conditions (Tungodden, 2003).

Frankfurt (1987) proposes the doctrine of sufficiency, which says that justice plays no role if everyone has enough. This may be interpreted as a non-aggregative approach, whereby absolute priority is assigned to those below the sufficiency threshold. There are two fundamental and interlinked issues within this framework. First, one needs to define what it means to have enough. Second, one needs to justify why justice is not an issue among those who have enough. Anderson (1999), who appeals to the notion of democratic equality, may be seen as one way of developing Frankfurt's proposal, where people have enough if they have what is sufficient to stand as an equal in society. Crisp (2003), on the other hand, relates the idea of sufficiency to the notion of compassion, where priority should be given to the worse off only when their circumstances warrant the compassion of an impartial spectator. Underlying both these proposals is the perspective that the role of justice is limited, which of course does not exclude the possibility that there are other reasons for caring about the circumstances of people above the sufficiency threshold.

The standard view within economics is to think of justice as unlimited, that is, as relevant independent of people's circumstances. However, it is commonly recognized that by far the most pressing problem of justice in the modern world is the presence of poverty, and this has caused a substantial literature on the definition and measurement of this concept (Sen and Foster, 1997). Given any definition of poverty, absolute or relative, there is then the further question of how to fit this into a more general theory of justice. One possibility is a non-aggregative approach, where strict priority is given to the alleviation of poverty but where this is combined with a concern for distributive justice among those who do not live in poverty. Interestingly, this scheme is formally closely related to the structure of the difference principle as suggested by Rawls (1971). Rawls proposed that a relative threshold should define the worst-off group, and that we should assign strict priority to the expectations of this group (and not only the worst-off individual). However, it turns out that a relative definition of the worst-off group does not make room for an interpretation of the difference principle that differs from the standard leximin interpretation (Tungodden, 1999). Hence, an absolute threshold is needed to build an alternative non-aggregative theory of justice to leximin.

Any non-aggregative theory of justices faces what we may call the tyranny of non-aggregation, that is, it sometimes justifies that minor improvements in the lives of some people should outweigh great losses for any number of better-off people. This may seem as a knock-down argument against a non-aggregative approach. However, it is important to have in mind that by rejecting a non-aggregative approach one accepts the tyranny of aggregation, since there does not exist any reasonable theory of justice that avoids both the tyranny of aggregation and the tyranny of non-aggregation.

Libertarianism, rights, and consequentialism

We have argued that a non-aggregative theory of justice has to assign absolute priority to the worse off, and thus such a theory provides a strong protection of the rights and liberties of this group. However, this may imply the violation of the rights of others. Libertarianism, on the other hand, holds that all agents are, initially at least (for example, prior to engaging in any commitments or unjust actions), full

self-owners, and that any violation of full self-ownership is unjust. The core idea of full self-ownership is that agents own themselves in just the same way that they can fully own inanimate objects.

The modern interest in libertarianism was initiated by the work of Nozick (1974), who not only defended full self-ownership but also the view that people should be free to appropriate parts of the external world as long as no one be left worse off with the appropriation than she would be if the thing were in common use. This view of just appropriation of the external world has recently been challenged by left-libertarians, who argue that people have joint ownership of natural resources (Moulin and Roemer, 1989; Steiner, 1994; Vallentyne and Steiner, 2000; Otsuka, 2003). If we accept the premise of joint ownership, then it follows that natural resources may be justly appropriated only with the permission of, or with a significant payment to, the other members of society.

The work of Nozick (1974) was partly motivated by Sen's liberal paradox (Sen, 1970b; Gibbard, 1974), where Sen shows that there is a conflict between respecting the Pareto principle and protecting a private sphere to each individual in society. This work has initiated a large literature, which has studied alternative formulations of individual rights. In particular, it has been argued that rights should be formulated as the admissibility of actions or strategies of individuals and not as the right to impose one's preferences on the ranking of a particular set of social alternatives (Gaertner, Pattanaik, and Suzumura, 1992). However, as pointed out by Sen (1992b; 1996a), even though this provides an interesting alternative formulation of rights, it does not in itself eliminate the tension between the Pareto principle and individual rights.

Libertarianism in its various forms provides one way of justifying individual rights, but the right-based perspective is certainly not exclusive to libertarianism (see, for example, Rawls, 1971; 1993; Kolm, 1996; van Hees and Dowding, 2003). Moreover, a rights-based perspective does not necessarily have to be non-consequentialistic in the sense that it imposes side constraints that cannot be overridden by other considerations of justice. It is possible to defend a consequentialistic rights-based approach, where the best overall outcome, as judged from an impersonal standpoint which gives equal weight to the interest of everyone, is to minimize the violations of some basic rights or liberties (Scheffler, 1988). In fact, it has also been argued that side constraints and agent relativity can be accommodated by consequential reasoning if we adopt a positional view of consequences (Sen, 1982; 1993). Finally, we should note that rights and liberties also may be justified on instrumental grounds, as a way of generating good consequences.

Intergenerational justice and international justice

The intergenerational perspective introduces several interesting challenges to a theory of justice (Parfit, 1984). First, we have to consider how to deal with the non-identity problem of future people, which questions whether people that are born as a result of a particular set of policies can be harmed by these policies, given that they would not have been born at all otherwise. Second, we have to consider how to avoid the repugnant conclusion, namely, that for any given affluent population there is a better world with more people, but where everyone has an arbitrarily low level of utility or well-being. This conclusion follows from two intuitively plausible conditions, namely, that we make the world better by bringing in people who have a life worth living and that we do not make the world worse by making it

more equal (at least as long as this does not reduce the total amount of utility in society). Blackorby, Bossert and Donaldson (1997; 2005) propose critical-level utilitarianism as the best possible solution to these problems, where critical-level utilitarianism disvalues only individuals whose utility level is below some fixed, low but positive threshold.

The literature on intergenerational justice has considered how to formulate a criterion of justice when there is an infinite number of generations. The basic problem was raised by Diamond (1965), who proved that within this framework it is impossible to construct a social welfare function that satisfies the Pareto principle, a principle of intergenerational equity and continuity. Basu and Mitra (2003) strengthen this result by proving that the continuity condition is superfluous. In other words, the fundamental conflict is between the Pareto principle an intergenerational equity. However, the literature also contains positive results, which show that a criterion for intergenerational justice can be formulated with an infinite horizon if one moves away from the framework of a social welfare function (Asheim, Buchholz and Tungodden, 2001; Asheim and Tungodden, 2004; Basu and Mitra, 2003; Basu and Mitra, 2007; Bossert, Sprumont and Suzumura, 2007; Fleurbaey and Michel, 2003).

Rawls (1971) limited his theory of justice to the circumstances of a nation, and recently this has been questioned by a number of philosophers (Pogge, 1989; 1992; 1994; 2001). They find any distinction between people based on territory arbitrary, and thus argue in favour of applying Rawls's principles of justice on the global scale. Consequently, they claim that the situation of the worst-off members of the global, rather than the domestic, society ought to be the starting point for considerations of justice. This view has been rejected by Rawls (1999), who argues that there is no basic structure in the international arena that can be the primary subject of social justice, and the difference principle cannot be a demand of justice in the international realm because, among other things, the justification of the difference principle has merit only between persons who cooperate in the way that this is done within the nation state.

Concluding remarks

The normative literature on justice has expanded enormously in recent years, with an extremely fruitful exchange of ideas between the disciplines of economics and philosophy. As a result, we now have a much richer understanding of how to think about the various possible conceptions of a just society. Still, there are many unresolved questions in the literature. Let me briefly mention three of them. First, there is a need further study of how to combine egalitarian ideas of responsibility with a concern for efficiency. Basic economic theory tells us that the efficient solution is to let individuals face the actual consequences of their choices, but this is clearly to hold individuals responsible for too much in many situations. How should we deal with this tension in a just society? Second, there is a need for further study of the metric of individual advantage. There are a number of suggestions present in the literature, including primary goods, basic needs, functionings and capabilities, but still need for more research on how to combine these approaches with a respect for individual preferences and choices. Finally, there is a need for further analysis of the prioritarian proposal. The core question within prioritarianism is how much more weight to attach to the worse off, and presently we lack a clear understanding of how to move forward on this issue.

See Also

- egalitarianism
- ethics and economics
- justice
- libertarianism

Bibliography

Anderson, E. 1999. What is the point of equality? *Ethics* 109, 287–337.

Arneson, R. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56, 159–94.

Arneson, R. 1990. Primary goods reconsidered. *Noûs* 24, 429–54.

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley 2nd ed., 1963.

Asheim, G.B., Bucholz, W. and Tungodden, B. 2001. Justifying sustainability. *Journal of Environmental Economics and Management* 2, 252–68.

Asheim, G.B. and Tungodden, B. 2004. Resolving distributional conflicts between generations. *Economic Theory* 24, 221–30.

d'Aspremont, C. 1985. Axioms for social welfare orderings. In *Social Goals and Social Organization: Essays in Memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler and H. Sonnenschein. Cambridge: Cambridge University Press.

d'Aspremont, C. and Gevers, L. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44, 199–210.

Barry, B. 1989. *Theories of Justice*. Berkley: University of California Press.

Basu, K. and Mitra, T. 2003. Aggregating infinite utility streams with intergenerational equity: the impossibility of being Paretian. *Econometrica* 32, 1557–63.

Basu, K. and Mitra, T. 2007. Utilitarianism for infinite utility streams: a new welfare criterion and its axiomatic characterization. *Journal of Economic Theory* 133, 350–73.

Binmore, K. 2005. *Natural Justice*. New York: Oxford University Press.

Blackorby, C., Bossert, W. and Donaldson, D. 1997. Critical-level utilitarianism and the population–ethics dilemma. *Economics and Philosophy* 13, 197–230.

Blackorby, C., Bossert, W. and Donaldson, D. 2005. *Population Issues in Social Choice Theory, Welfare Economics and Ethics*. New York: Cambridge University Press.

Blair, D.H. 1988. The primary-goods indexation problem in Rawls's Theory of Justice. *Theory and Decision* 24, 239–52.

Bossert, W. 1995. Redistribution mechanisms based on individual characteristics. *Mathematical Social Sciences* 29, 1–17.

Bossert, W. and Fleurbaey, M. 1996. Redistribution and compensation. *Social Choice and Welfare* 13, 343–55.

Bossert, W., Sprumont, Y. and Suzumura, K. 2007. Ordering infinite utility streams. *Journal of Economic Theory* 135, 579–89.

Bossert, W. and Weymark, J.A. 2002. Utility in social choice. In *Handbook of Utility Theory*, ed. S. Barbera, P. Hammond and C. Seidl. London: Kluwer.

Broome, J. 1991. *Weighing Goods. Equality, Uncertainty, and Time*. Oxford: Basil Blackwell.

Broome, J. 2007. Equality versus priority: a useful distinction. In *Fairness and Goodness in Health*, ed. C. Murray and D. Wikler. Geneva: World Health Organization.

Brun, B.C. and Tungodden, B. 2004. Non-welfarist theories of justice: is ‘the intersection approach’ a solution to the indexing impasse? *Social Choice and Welfare* 22, 1–12.

Camerer, C.F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Cappelen, A. and Tungodden, B. 2006. Relocating the responsibility cut: should more responsibility imply less redistribution. *Politics, Philosophy and Economics* 5, 353–62.

Cohen, G.A. 1989. On the currency of egalitarian justice. *Ethics* 99, 906–44.

Crisp, R. 2003. Equality, priority, and compassion. *Ethics* 113, 745–63.

Dennet, D. 2003. *Freedom Evolves*. London: Penguin Books.

Diamond, P. 1965. The evaluation of infinite utility streams. *Econometrica* 33, 170–7.

Dworkin, R. 1981. What is equality? Parts. I and II. *Philosophy and Public Affairs* 10, 283–345.

Dworkin, R. 2002. Sovereign Virtue revisited. *Ethics* 113, 106–43.

Fleurbaey, M. 1995a. Equal opportunity or equal social outcome. *Economics and Philosophy* 11, 25–55.

Fleurbaey, M. 1995b. The requisites of equal opportunity. In *Social Choice, Welfare, and Ethics*, ed. W. A. Barnett et al. New York: Cambridge University Press.

Fleurbaey, M. 1995c. Three solutions for the compensation problem. *Journal of Economic Theory* 65, 505–21.

Fleurbaey, M. 2002. Equality of resources revisited. *Ethics* 113, 82–105.

Fleurbaey, M. 2003. On the informational basis of social choice. *Social Choice and Welfare* 21, 347–84.

Fleurbaey, M. 2007. Equality versus priority: how relevant is the distinction? In *Fairness and Goodness in Health*, ed. C. Murray and D. Wikler. Geneva: World Health Organization (forthcoming).

Fleurbaey, M. and Maniquet, F. 1996. Fair allocation avec unequal production skills: the no-envy approach to compensation. *Mathematical Social Sciences* 32, 71–93.

Fleurbaey, M. and Maniquet, F. 1999. Fair allocation with unequal production skills: the solidarity approach to compensation. *Social Choice and Welfare* 16, 569–83.

Fleurbaey, M. and Maniquet, F. 2006. Fair income tax. *Review of Economic Studies* 73, 55–83.

Fleurbaey, M. and Maniquet, F. 2008. Compensation and responsibility. In *Handbook of Social Choice and Welfare*, vol. 2, ed. K.J. Arrow, A.K. Sen and K. Suzumura. Amsterdam: North-Holland.

Fleurbaey, M. and Michel, P. 2003. Intertemporal equity and the extension of the Ramsey criterion. *Journal of Mathematical Economics* 39, 777–802.

Fleurbaey, M., Suzumura, K. and Tadenuma, K. 2005. Arrovian aggregation in economic environments: how much should we know about indifference surfaces? *Journal of Economic Theory* 124, 22–44.

Fleurbaey, M. and Trannoy, A. 2000. The impossibility of a Paretian egalitarian. *Social Choice and Welfare* 21, 243–63.

Fleurbaey, M., Tungodden, B. and Chang, H.F. 2003. Any non-welfarist method of policy assessment violates the Pareto principle: a comment. *Journal of Political Economy* 111, 1382–5.

Frankfurt, H. 1987. Equality as a moral ideal. *Ethics* 98, 21–43.

Gaertner, W., Pattanaik, P.K. and Suzumura, K. 1992. Individual rights revisited. *Economica* 59, 161–77.

Gibbard, A. 1974. A Pareto-consistent libertarian claim. *Journal of Economic Theory* 7, 388–410.

Gibbard, A. 1979. Disparate goods and Rawls' difference principle: a social choice theoretic treatment. *Theory and Decision* 11, 267–88.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Hammond, P. 1975. A note on extreme inequality aversion. *Journal of Economic Theory* 11, 119–32.

Kaplow, L. and Shavell, S. 2001. Any non-welfarist method of policy assessment violates the Pareto principle. *Journal of Political Economy* 109, 281–6.

Kaplow, L. and Shavell, S. 2002. *Fairness Versus Welfare*. Cambridge, MA: Harvard University Press.

Kolm, S.-C. 1996. *Modern Theories of Justice*. Cambridge, MA: MIT Press.

Konow, J. 2003. Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature* 41, 1188–239.

McKerlie, D. 1994. Equality and priority. *Utilitas* 6, 25–42.

Mongin, P. 2001. The impartial observer theorem of social ethics. *Economics and Philosophy* 17, 147–80.

Mongin, P. and d'Aspremont, C. 2002. Utility theory and ethics. In *Handbook of Utility Theory*, vol. 1, ed. S. Barberà, P. Hammond and C. Seidl. Dordrecht: Kluwer.

Moulin, H. and Roemer, J. 1989. Public ownership of the external world and private ownership of self. *Journal of Political Economy* 97, 347–67.

Nagel, T. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.

Nagel, T. 2002. *Concealment and Exposure & Other Essays*. Oxford: Oxford University Press.

Nozick, R. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Ooghe, E., Schokkaert, E. and Van de gaer, D. 2006. Equality of opportunity versus equality of opportunity sets. *Social Choice and Welfare*.

Otsuka, M. 2003. *Libertarianism without Inquality*. Oxford: Clarendon Press.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Parfit, D. 1995. Equality or priority. Lindley Lecture, University of Kansas. Reprinted in *The Ideal of Equality*, ed. M. Clayton and A. Williams. Basingstoke: Palgrave Macmillan, 2000.

Pattanaik, P.K. and Xu, Y. 2007. Minimal relativism, dominance, and standard of living comparisons based on functionings. *Oxford Economic Papers* 59, 354–74.

Plott, C.R. 1978. Rawls's theory of justice: an impossibility result. In *Decision Theory and Social Ethics: Issues in Social Choice*, ed. H. Gottinger and W. Leinfellner. Dordrecht: D. Reidel.

Pogge, T. 1989. *Realizing Rawls*. Ithaca, NY: Cornell University Press.

Pogge, T. 1992. Cosmopolianism and sovereignty. *Ethics* 103, 48–75.

Pogge, T. 1994. An egalitarian law of peoples. *Philosophy and Public Affairs* 23, 195–224.

Pogge, T., ed. 2001. *Global Justice*. Oxford: Blackwell.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 1993. *Political Liberalism*. New York: Columbia University Press.

Rawls, J. 1999. *The Law of Peoples*. Cambridge, MA: Harvard University Press.

Roemer, J. 1985. Equality of talent. *Economics and Philosophy* 1, 151–87.

Roemer, J. 1993. A pragmatic theory of responsibility for the egalitarian planner. *Philosophy and Public Affairs* 22, 146–66.

Roemer, J. 1996. *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.

Roemer, J. 1998. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Roemer, J. 2002a. Egalitarianism against the veil of ignorance. *Journal of Philosophy* 99, 167–84.

Roemer, J. 2002b. Equality of opportunity: a progress report. *Social Choice and Welfare* 19, 455–71.

Scanlon, T. 1982. Contractualism and utilitarianism. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams. Cambridge: Cambridge University Press.

Scanlon, T. 1998. *What We Owe Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T. 2000. The diversity of objections to inequality. In *The Ideal of Equality*, ed. M. Clayton and A. Williams. Basingstoke: Palgrave Macmillan.

Scheffler, S. 1988. *Consequentialism and Its Critics*. New York: Cambridge University Press.

Sen, A. 1970a. *Collective Choice and Social Welfare*. San Francisco: Holden Day.

Sen, A. 1970b. The impossibility of a Paretian liberal. *Journal of Political Economy* 78, 152–7.

Sen, A. 1979. Utilitarianism and welfarism. *Journal of Philosophy* 9, 463–89.

Sen, A. 1982. Rights and agency. *Philosophy and Public Affairs* 11, 3–39.

Sen, A. 1985. *Commodities and Capabilities*. Amsterdam: North-Holland.

Sen, A. 1992a. *Inequality Reexamined*. Cambridge, MA: Harvard University Press.

Sen, A. 1992b. Minimal liberty. *Economica* 59, 139–59.

Sen, A. 1993. Positional objectivity. *Philosophy and Public Affairs* 22, 126–45.

Sen, A. 1996a. Rights: formulations and consequences. *Analyse & Kritik* 18, 153–70.

Sen, A. 1996b. Individual preferences as the basis of social choice. In *Social Choice Re-examined*, ed. K. Arrow, A. Sen and K. Suzamura. Basingstoke: Macmillan.

Sen, A. and Foster, J. 1997. *On Economic Inequality*, expanded edn. Oxford: Clarendon Press.

Skyrms, B. 1996. *Evolution of the Social Contract*. New York: Cambridge University Press.

Skyrms, B. 2003. *The Stage Hunt and the Evolution of Social Structure*. Cambridge: Cambridge

University Press.

Steiner, H. 1994. *An Essay on Rights*. Oxford: Blackwell.

Taurek, J. 1977. Should the numbers count? *Philosophy and Public Affairs* 6, 293–316.

Temkin, L. 1993. *Inequality*. Oxford: Oxford University Press.

Temkin, L. 2000. Equality, priority, and the levelling down objection. In *The Ideal of Equality*, ed. M. Clayton and A. Williams. Basingstoke: Palgrave Macmillan.

Tungodden, B. 1999. The distribution problem and Rawlsian reasoning. *Social Choice and Welfare* 16, 599–614.

Tungodden. 2003. The value of equality. *Economics and Philosophy* 19, 1–44.

Tungodden, B. 2005. Responsibility and redistribution: the case of first best. *Social Choice and Welfare* 24, 33–44.

Tungodden, B. and Vallentyne, P. 2005. On the possibility of Paretian egalitarianism. *Journal of Philosophy* 102, 126–54.

Vallentyne, P. and Steiner, H., eds. 2000. *Left Libertarianism and Its Critics: The Contemporary Debate*. New York: Palgrave Publishers.

Van de gaer, D. 1993. Equality of opportunity and investment in human capital. Ph.D thesis, K.U. Leuven.

Van Hees, M. and Dowding, K. 2003. The construction of rights. *American Political Science Review* 97, 281–293.

Van Parijs, P. 1995. *Real Freedom for All*. Oxford: Oxford University Press.

Weymark, J.A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J.E. Roemer. Cambridge: Cambridge University Press.

How to cite this article

Tungodden, Bertil. "justice (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_J000052> doi:10.1057/9780230226203.0878

The New Palgrave Dictionary of Economics Online

justice

Amartya Sen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Traditionally, economists have treated justice as a component of social welfare maximization. Recently, philosophical treatments of justice have challenged the three principles underlying utilitarianism: welfarism, sum-ranking and consequentialism. Various theories of justice advance alternatives to utility (such as Rawls's notion of ‘primary goods’) as a basis for social judgements, counterpose distributional criteria (such as Sen's ‘leximin’ rule) to the aggregative approach of utilitarianism, and assert the moral priority of certain aspects of individual advantage (such as Nozick's idea of individual rights as entitlements) over consequences. This article attempts to distinguish and clarify these conceptions of justice.

Keywords

Arrow, K.; consequentialism; difference principle; entitlements; envy; equality; fairness; interpersonal utility comparisons; justice; leximin; liberty; maximin; Meade, J.; Nozick, R.; Pigou, A.; primary goods; property rights; Rawls, J.; Sen, A.; sum-ranking; utilitarianism; welfarism

Article

1 Justice and utilitarianism

The concept of justice is often invoked in economic discussions. Its relevance to economic evaluation is obvious enough. However, it is fair to say that in traditional welfare economics, when the notion of justice has been invoked, it has typically been seen only as a part of a bigger exercise, viz., that of social welfare maximization, rather than taking justice as an idea that commands attention on its own. For example, in utilitarian welfare economics (e.g. Pigou, 1952; Harsanyi, 1955) the problem of justice is not separated out from that of maximization of aggregate utility. This situation has been changing in recent years, partly as a result of developments in moral philosophy dealing explicitly with the notion of

justice as a concept of independent importance (see especially Rawls, 1971, 1980).

In the utilitarian formulation the maximand in all choice exercises is taken to be the sum-total of individual utilities. The approach can be seen as an amalgam of three distinct principles: (1) welfarism, (2) sum-ranking, and (3) consequentialism. *Welfarism* asserts that the goodness of a state of affairs is to be judged entirely by the utility information related to that state, i.e., by information about individual utilities. All other information is either irrelevant, or only indirectly relevant as a causal influence on utilities (or as a surrogate for utility measures when such measurement cannot be directly done). The second principle is *sum-ranking*, which asserts that the goodness of a collection of utilities (or welfare indicators) of different individuals, taken together, is simply the sum of these utilities (or indicators). This eliminates the possibility of being concerned with inequalities in the distribution of utilities, and the overall goodness or ‘social welfare’ is seen simply as the aggregate of individual utilities. The third principle is *consequentialism*, which requires that all choice variables, such as actions, rules, institutions, etc., must be judged in terms of the goodness of their respective consequences. The overall effect of combining these three principles is to judge all choice variables by the sum-total of utilities generated by one alternative rather than another.

2 Sum-ranking and equality

A theory of justice can take issue with each of the principles underlying utilitarianism, and in fact in the literature that has developed in recent decades, each of these principles has been seriously challenged (see the papers included in Sen and Williams, 1982). Some critiques have been particularly concerned with assessing and questioning the axiom of sum-ranking, and have considered the claims of equality in the distribution of well-being (see, for example, Phelps, 1973; Sen, 1973, 1977, 1982; Kern, 1978).

The summation formula can be defended either directly (e.g., in terms of attaching equal importance to everyone's ‘interest’: see Hare, 1981, 1982), or indirectly through invoking some model of ‘impersonality’ or ‘fairness’ (e.g., involving a hypothetical choice in a situation of primordial uncertainty, in which each person has to assume that he or she has an as if equal probability of becoming anybody else: see Vickrey, 1945; Harsanyi, 1955). Other routes to deriving sum-ranking involve independence or separability requirements of various kinds (see d’Aspremont and Gevers, 1977; Deschamps and Gevers, 1978; Maskin, 1978; Gevers, 1979; Roberts, 1980; Myerson, 1981; Blackorby, Donaldson and Weymark, 1984; d’Aspremont, 1985).

Whether the defences obtainable from these approaches are convincing enough has been a matter of some dispute. There have also been some interpretative discussions as to whether giving equal importance to everyone's ‘interest’ does, as alleged, in fact yield the formula of summing individual utilities *irrespective* of distribution, and also whether the additive formula that is obtained on the basis of hypothetical primordial choice is, in fact, a justification for adding individual utilities as they might be *substantively* interpreted in welfare economic exercises (see Pattanaik, 1971; Smart and Williams, 1973; Sen, 1982, 1985a; Blackorby, Donaldson and Weymark, 1984; Williams, 1985). It is not obvious that this debate has been in any way definitively concluded one way or the other.

3 The difference principle and leximin

Meanwhile, much attention has been paid to developing welfare-economic rules based on taking explicit note of inequalities in the distribution of utilities. A definitive departure on this came from Suppes (1966). Another major approach was developed in Rawls's (1971) *Theory of Justice*, even though Rawls himself was concerned not so much with the distribution of *utilities* but with that of the indices of primary goods (on which more later). The concern with the utility level of the worst-off individual has been formalized and reflected in various formulae suggested or derived in the rapidly growing welfare-economic literature on this theme. In particular, James Meade (1976) has provided an extensive treatment of this type of distributional issues, and it has also been penetratingly analysed by Kolm (1969), Phelps (1973, 1977), Atkinson (1975, 1983), Blackorby and Donaldson (1977), and others.

In fact, the Rawlsian ‘Difference Principle’, which judges states of affairs by the advantage of the least well-off person or group, has often been axiomatized in welfare economics and in the social-choice literature by equating advantage with utility. In this form, the ‘lexicographic maximin’ rule (proposed in Sen, 1970) has been axiomatically derived in different ways. The rule judges states of affairs by the well-being of the worst-off individual. In case of ties of the worst-off individuals’ utilities, the states are ranked according to the utility levels of the second worst-off individuals respectively. In case of ties of the second worst-off positions as well, the third worst-off individuals’ utilities are examined. And so on. There is no necessity to interpret these axioms in terms of utilities only, and in fact the analytical results derived in this part of the social-choice literature can be easily applied without the ‘welfarist’ structure of identifying individual advantage with the respective utilities. Various axiomatic derivations of lexicographic maximin – ‘leximin’ for short – can be found in Hammond (1976), Strasnick (1976), Arrow (1977), d’Aspremont and Gevers (1977), Sen (1977), Deschamps and Gevers (1978), Suzumura (1983), Blackorby, Donaldson and Weymark (1984), d’Aspremont (1985), among others. These can be seen as exercises that incorporate concern for reducing inequality, related to recognizing the claims of justice.

While the Rawlsian approach rejects the aggregation procedure of utilitarianism (i.e., ranking by sums), a major aspect of the Rawlsian theory involves the rejection of utility as the basis of social judgements (i. e., welfarism). Rawls (1971) argues for the priority of the ‘principle of liberty’, demanding that ‘each person is to have an equal right to the most extensive basic liberty compatible with similar liberty for others’. Then, going beyond the principle of liberty, claims of efficiency as well as equity are both supported by Rawls's ‘second principle’ which *inter alia* incorporates his ‘Difference Principle’ in which priority is given to furthering the powers of the worst-off group. These powers are judged by indices of ‘primary social goods’ which each person wants (Rawls, 1971, pp. 60–65).

Primary goods are ‘things that every rational man is presumed to want’, including ‘rights, liberties and opportunities, income and wealth, and the social bases of self-respect’. The Difference Principle takes the form, in fact, of maximin, or lexicographic maximin, based on interpersonal comparisons of indices of primary goods. This rule can be axiomatized in much the same way as the other ‘lexicographic maximin’ rule based on utilities, and all that is needed is a reinterpretation of the content of the axioms (with the objects of value being indices of primary goods rather than utilities).

The Rawlsian approach to justice, therefore, involves rejection both of welfarism and of sum-ranking. Furthermore, consequentialism is disputed too, since the priority of liberty might possibly go against judging all choice variables by consequences only. At least, in the more standard forms, consequentialism does involve such a conflict, even though it is arguable that the problem can be, to a

great extent, resolved by a broader understanding of consequences, which takes into account the fulfilment and violation of liberties and rights, and also the agent's special role in the actions performed (Sen, 1985a).

4 Utilities, primary goods and capabilities

The claim of primary goods to represent the demands of justice better than utilities is based on the idea that utilities do not reflect a person's advantage (in terms of well-being or powers) adequately. It is arguable that in making interpersonal comparisons of advantage, the metric of utilities (either in the form of happiness, or of desire fulfilment) may be biased against those who happen to be hopelessly deprived since the demands of unharrassed survival force people to take pleasure in small mercies and to cut their desires to shape in the light of feasibilities (see Sen, 1985a, 1985b). The status of ‘preference’ may be disputed in view of the need for critical assessment (see Broome, 1978; McPherson, 1982; Goodin, 1985, among others). Also, what types of pleasures should ‘count’ can itself be a matter for an important moral judgement. As Rawls (1971) points out, in the utilitarian formulation, we have the unplausible requirement that

> if men take a certain pleasure in discriminating against one another, in subjecting others to a lesser liberty as a means of enhancing their self-respect, then the satisfaction of these desires must be weighed in our deliberations according to their intensity, or whatever, along with other desires (pp. 30–31).

These and other types of difficulties have been dealt with by some utilitarians through moving to less straight-forward versions of utilitarianism, for example Harsanyi's (1982) exclusion of ‘all antisocial preferences, such as sadism, envy, resentment, and malice’ (p. 56); see also the refinements proposed by Hare (1981, 1982), Hammond (1982) and Mirrlees (1982).

Recently, it has been argued that primary goods themselves may be rather deceptive in judging people's advantages, since the ability to convert primary goods into useful capabilities may vary from person to person. For example, while the same level of income (included among ‘primary goods’) may give each person the same command over calories and other nutrients, the nourishment of a person depends also on other parameters such as body size, metabolic rates, sex (and if female, whether pregnant or lactating), climatic conditions, etc. This indicates that a more plausible notion of justice may demand that attention be directly paid to the distribution of basic capabilities of people (see Sen, 1982, 1985b). The approach goes back to Smith's (1776) and Marx's (1875) focus on fulfilling needs.

The achievement of capabilities will, of course, be causally related to the command over primary goods, and the capabilities, in their turn, will also influence the extent to which utilities are achieved, so that the various alternative measures will not be independent of each other. However, the basic issue is the variable that should be chosen to serve as the proper metric for judging advantages of people – the equity and the distribution of which could form the foundations of a theory of justice. On this central issue several alternative views continue to flourish in the literature.

5 Fairness and envy

A view of justice that is not altogether dissimilar from Rawls's concerned with primary goods is captured by the literature on ‘fairness’, inspired by a pioneering contribution of Foley (1967). In this approach a person's relative advantage is judged by the criterion as to whether he or she would have preferred to have had the commodity bundle enjoyed by another person. This has been seen as a criterion of ‘non-envy’. If no one ‘envies’ the bundle of anyone else, the state of affairs is described as being ‘equitable’. If a state is both equitable and Pareto efficient, it is described as being ‘fair’ (even though the term fairness is also sometimes used interchangeably with only ‘equitability’).

There has been an extensive literature on existence problems, in particular whether equitability can be combined with efficiency in all circumstances. (The answer seems to be no, especially when production is involved: Pazner and Schmeidler, 1974.) There has also been considerable exploration of the effects of varying the criterion of equitability and fairness to reflect better the common intuitions regarding the requirements of justice. Various results on these problems and related ones have been presented, among many others, by Foley (1967), Schmeidler and Vind (1972), Feldman and Kirman (1974), Varian (1974, 1975), Svensson (1980), and Suzumura (1983).

It should be remarked that the fairness criterion does not provide a complete ranking of alternative states. It identifies some requirements of justice, which makes the states fair. Varian (1974) has argued, with some force, that ‘social decision theory asks for too much out of the process in that it asks for an entire *ordering* of the various social states (allocations in this case)’, whereas ‘the original question asked only for a “good” allocation; there was no requirement to rank all allocations’ (pp. 64–5). While it is true that ‘the fairness criterion in fact limits itself to answering the original question’, the absence of further rankings may be particularly problematic if no feasible ‘fair’ allocation exists incorporating efficiency (as seems to be the case in many situations). Furthermore, while a ‘pass-fail’ criterion of justice may have attractive simplicity, it does not follow that two states, both passing this criterion, must be seen as being ‘equally just’. Various ‘finer’ aspects of justice have indeed been discussed in the literature (see particularly. Suppes, 1966; Kolm, 1969; Rawls, 1971; Meade, 1976; Atkinson, 1983).

It should also be noted that the ‘fairness’ literature deals with commodity allocations, or incomes, or some other part of the set of things that figures in Rawls's characterization of ‘primary goods’. The list is, in fact, much less extensive than that of primary goods as defined by Rawls (1971), and as such it leaves out many considerations that are regarded as important in the Rawlsian framework (e.g., the social bases of self-respect). On the other hand the criticisms – discussed earlier – of the Rawlsian focus on primary goods (based on recognizing inter-individual variations in the ability to convert primary goods into capabilities) would apply *a fortiori* to the fairness approach as well.

6 Liberty and entitlements

A different type of consideration altogether is raised by the place of liberty in a theory of justice. As was mentioned before, Rawls gives it priority. This priority has been questioned by pointing to the possibility that other things (e.g., having enough food) may sometimes be no less important than enjoying liberty without restriction by others. Rawls does, of course, attach importance to these other considerations, but in view of the priority of liberty, they may end up having too little impact on judgements regarding

justice in many circumstances, and this might not be acceptable (on this see Hart, 1973).

On the other hand, in some other theories of justice, the priority of liberty has been given even greater importance than in the Rawlsian structure. For example, in Nozick's (1974) theory of ‘entitlements’, rights are given complete priority, and since these rights are characterized quite extensively, it is not clear whether or not much remains to be supported over and above the recognition of rights. Nozick argues against any ‘patterning’ of *outcomes*, indicating that any outcome that is arrived at on the basis of people's legitimate exercise of their rights must be acceptable because of the moral force of rights as such. These rights, in Nozick's analysis, include not only personal liberty, but also ownership rights over property, including the freedom to use its fruits, to use it freely for exchange, and to donate or bequeath it to others (thereby asserting the legitimacy of inherited property).

This type of approach has been criticized partly on grounds of what has been seen as its ‘extremism’, since the constraints imposed by rights can override other important considerations, for example reducing misery and promoting the well-being of the deprived members of the society. In fact, it has been argued that a system of entitlements of the kind specified by Nozick might well co-exist with the emergence and sustaining of widespread starvation and famines, which are often the result of legally sanctioned exercises of property rights rather than of natural calamities (on this see Sen, 1981).

Although Nozick does refer to the possibility that in case of ‘catastrophic moral horrors’ rights may be compromised, it is not at all clear how his theory would accommodate such waiving of rights, in the absence of formulation of other, competing bases of moral judgements. On the other hand, there cannot be any doubt that Nozick's theory does capture some notions of justice that can be found in a less clear form in the literature. Nozick's analysis gives a well-formulated and illuminating account of an entitlement-based approach to justice.

7 Sources of difference

To conclude, theories of justice explicitly or implicitly invoked in the literature show a variety of ways in which the demands of justice can be interpreted. There are at least three different bases of variation. One source of variation concerns the *metric* in terms of which a person's *advantage* is to be judged in the context of assessing equity and justice. Various metrics have been considered in this context, including utility (as under utilitarianism and other welfarist theories of justice), primary goods index (as in the Rawlsian theory of justice), capabilities index (as in theories emphasizing what people can actually do or be, e.g., Sen, 1985b), incomes or commodity bundles (as in the literature on ‘fairness’, and on statistical measures of poverty, e.g., Foster, 1984), various notions of command over commodity bundles and resources (as in some notions of ‘equality’ developed in the literature, e.g., Archibald and Donaldson, 1973; Dworkin, 1981), and so on.

A second source of difference relates to the *aggregating* of diverse information regarding the advantages of different individuals. One approach, best represented by utilitarianism, sees nothing being needed to be ascertained other than the *sum-total* of the overall utilities of different people. Insofar as distributional considerations come into this exercise, they enter in the conversion of goods to be distributed into the appropriate metric of individual utilities. For example, inequality in the distribution of incomes may be disvalued in the approach of utilitarian justice because it may lead to a reduction in the sum-total of individual utilities, through (interpersonally comparable) ‘diminishing marginal utilities’. Other

approaches are more concerned with distributional properties related to the different individuals' relative positions (vis-à-vis each other). The Rawlsian lexicographic maximin is one example of such a distributional concern, and there are others than can be considered, such as adding concave transformations of the individual utility indices (e.g., the additive formula used by Mirrlees, 1971, for his taxation assessment), and using various 'equity' axioms (e.g., Kolm, 1969, 1972; Sen, 1973; 1982; Atkinson, 1975, 1983; Hammond, 1976, 1979; d'Aspremont and Gevers, 1977; Roberts, 1980).

The third issue concerns the claimed *priority* of some particular aspect of a person's advantage (e.g., Rawls's insistence on the priority of liberty), o*r* non-consequentialist priority of some processes over results (e.g., Nozick's, 1974, view of rights serving as unrelaxable constraints; or ideas of exploration based on counterfactual exercises of shared rights to social resources, e.g., Roemer, 1982).

Given the diversity of moral intuitions related to the complex notion of justice, which has been extensively used over centuries to arrive at normative assessment, it is not surprising that various theories of justice have been proposed in the economic and philosophical literature. The exercise of clearly understanding what the differences between distinct theories of justice consist of (and arise from) is, in some ways, the first task. This essay has been concerned with that task.

See Also

- justice (new perspectives)

Bibliography

Archibald, G.C. and Donaldson, D. 1979. Notes on economic equality. *Journal of Public Economics* 12, 205–14.

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Arrow, K.J. 1963. *Social Choice and Individual Values*. 2nd edn, New York: Wiley.

Arrow, K.J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review* 67, 219–25.

Atkinson, A.B. 1975. *The Economics of Inequality*. Oxford: Clarendon Press.

Atkinson, A.B. 1983. *Social Justice and Public Policy*. Brighton: Wheatsheaf; Cambridge, Mass.: MIT Press.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*. London: Payne. Reprinted, Oxford: Clarendon Press, 1907.

Blackorby, C. and Donaldson, D. 1977. Utility versus equity: some plausible quasi-orderings. *Journal of*

Public Economics 7(3), 365–82.

Blackorby, C., Donaldson, D. and Weymark, J.A. 1984. Social choice with interpersonal utility comparisons: a diagrammatic introduction. *International Economic Review* 25(2), 327–56.

Broome, J. 1978. Choice and value in economics. *Oxford Economic Papers* 30(3), 313–33.

d'Aspremont, C. and Gevers, L. 1977. Equity and the informational base of collective choice. *Review of Economic Studies* 44(2), 199–209.

d'Aspremont, C. 1985. Axioms for social welfare orderings. In *Social Goods and Social Organization: Essays in Memory of Elisha Pazner*, ed. L. Hurwicz, D. Schmeidler and H. Sonnenschein. Cambridge: Cambridge University Press.

Deschamps, R. and Gevers, L. 1978. Leximin and utilitarian rules: a joint characterisation. *Journal of Economic Theory* 17(2), 143–63.

Dworkin, R. 1981. What is equality? *Philosophy and Public Affairs* 10, 185–246.

Feldman, A. and Kirman, A. 1974. Fairness and envy. *American Economic Review* 64(6), 995–1005.

Foley, D. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7(1), 45–98.

Foster, J. 1984. On economic poverty: a survey of aggregate measures. *Advances in Econometrics* 3.

Gevers, L. 1979. On interpersonal comparability and social welfare orderings. *Econometrica* 47(1), 75–89.

Goodin, R.E. 1985. *Protecting the Vulnerable*. Chicago: Chicago University Press.

Gottinger, H.W. and Leinfellner, W., eds. 1973. *Decision Theory and Social Ethics: Issues in Social Choice*. Dordrecht: Reidel.

Hammond, P.J. 1976. Equity, Arrow's conditions and Rawls' Difference Principle. *Econometrica* 44(4), 793–804.

Hammond, P.J. 1979. Equity in two person situations: some consequences. *Econometrica* 47(5), 1127–36.

Hammond, P.J. 1982. Utilitarianism, uncertainty and information. In Sen and Williams (1982).

Hare, R.M. 1981. *Moral Thinking: Its Levels, Methods and Point*. Oxford: Clarendon Press.

Hare, R.M. 1982. Ethical theory and utilitarianism. In Sen and Williams (1982).

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Harsanyi, J.C. 1982. Morality and the theory of rational behaviour. In Sen and Williams (1982).

Hart, H.L.A. 1973. Rawls on liberty and its priority. *University of Chicago Law Review* 40, 534–55.

Kern, L. 1978. Comparative distributive ethics: an extension of Sen's examination of the pure distribution problem. In Gottinger and Leinfellner (1978).

Kolm, S.C. 1969. The optimum production of social justice. In *Public Economics*, ed. J. Margolis and H. Guitton. London: Macmillan.

Kolm, S.C. 1972. *Justice et équité*. Paris: Edition du Centre National de la Recherche Scientifique.

Margolis, J. and Guitton, H., eds. 1969. *Public Economics*. London: Macmillan.

Marx, K. 1875. *Critique of the Gotha Programme*. English translation, New York: International Publishers, 1938.

Maskin, E. 1978. A theorem of utilitarianism. *Review of Economic studies* 45(1), 93–6.

McPherson, M.S. 1982. Mill's moral theory and the problem of preference change. *Ethics* 92(2), 252–73.

Meade, J.E. 1976. *The Just Economy*. London: Allen & Unwin.

Mirrlees, J.A. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Mirrlees, J.A. 1982. The economic uses of utilitarianism. In Sen and Williams (1982).

Myerson, R.B. 1981. Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica* 49(4), 883–97.

Ng, Y.K. 1979. *Welfare Economics*. London: Macmillan.

Nozick, R. 1974. *Anarchy, State and Utopia*. Oxford: Blackwell.

Pattanaik, P.K. 1971. *Voting and Collective Choice*. Cambridge: Cambridge University Press.

Pattanaik, P.K. and Salles, M., eds. 1983. *Social Choice and Welfare*. Amsterdam: North-Holland.

Pazner, E.A. and Schmeidler, D. 1974. A difficulty in the concept of fairness. *Review of Economic Studies* 41(3), 441–3.

Phelps, E.S., ed. 1973. *Economic Justice*. Harmondsworth: Penguin Books.

Phelps, E.S. 1977. Recent developments in welfare economics: Justice et Equité. In *Frontiers of Quantitative Economics*, vol. 3, ed. M.D. Intriligator. Amsterdam: North-Holland.

Pigou, A.C. 1952. *The Economics of Welfare*. 4th edn, London: Macmillan.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Rawls, J. 1980. Kantian construction in moral theory. The Dewey Lectures 1980. *Journal of Philosophy* 77.

Roberts, K.W.S. 1980. Interpersonal comparability and social choice theory. *Review of Economic Studies* 47, 421–39.

Roemer, J. 1982. *A General Theory of Exploitation and Class*. Cambridge, Mass.: Harvard University Press.

Schmeidler, D. and Vind, K. 1972. Fair net trades. *Econometrica* 40(4), 637–42.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day, and Edinburgh: Oliver & Boyd; republished Amsterdam: North-Holland.

Sen, A.K. 1973. *On Economic Inequality*. Oxford: Clarendon Press; New York: Norton.

Sen, A.K. 1977. On weights and measures: informational constraints in social welfare analysis. *Econometrica* 45(7), 1539–72.

Sen, A.K. 1981. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Clarendon Press.

Sen, A.K. 1982. *Choice, Welfare and Measurement*. Oxford: Blackwell; Cambridge, Mass.: MIT Press.

Sen, A.K. 1985a. Well-being, agency and freedom: the Dewey Lectures 1984. *Journal of Philosophy* 82 (4), 169–221.

Sen, A.K. 1985b. *Commodities and Capabilities*. Amsterdam: North-Holland.

Sen, A.K. and Williams, B. 1982. *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Smart, J.J.C. and Williams, B.A.O. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Dent, 1954.

Strasnick, S. 1976. Social choice theory and the derivation of Rawls' Difference Principle. *Journal of Philosophy* 73(4), 85–99.

Suppes, P. 1966. Some formal models of grading principles. *Synthèse* 6, 284–306.

Suzumura, K. 1983. *Rational Choice, Collective Decisions and Social Welfare*. Cambridge: Cambridge University Press.

Svensson, L.G. 1980. Equity among generations. *Econometrica* 48(5), 1251–6.

Varian, H. 1974. Equity, envy and efficiency. *Journal of Economic Theory* 9(1), 63–91.

Varian, H. 1975. Distributive justice, welfare economics and the theory of fairness. *Philosophy and Public Affairs* 4, 223–47.

Vickrey, W. 1945. Measuring marginal utility by reactions to risk. *Econometrica* 13, 319–33.

Williams, B. 1973. A critique of utilitarianism. In Smart and Williams (1973).

Williams, B. 1985. *Ethics and the Limits of Philosophy*. Cambridge: Cambridge University Press.

How to cite this article

Sen, Amartya. "justice." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_J000033> doi:10.1057/9780230226203.0877

The New Palgrave Dictionary of Economics Online

Kahn, Richard Ferdinand (1905 – 1989)

Luigi L. Pasinetti
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cambridge Circus; duopoly; elasticity of substitution; ideal output; Kahn, R. F.; Keynes, J. M.; Keynesian Revolution; liquidity preference; market imperfections; multiplier; oligopoly; Radcliffe Committee view; Robinson, J. V.; Sraffa, P.

Article

Kahn was the favourite pupil and closest collaborator of John Maynard Keynes, at the time when the ‘Keynesian Revolution’ was under way (Keynes, 1936). For the whole of his academic career, he remained associated with King's College, Cambridge (Keynes's College), where he lived, as a bachelor, from his undergraduate days.

Kahn was born in London on 10 August 1905, into a Jewish family of very strict religious observance. His father, Augustus Kahn, a schoolmaster, was a first-generation Englishman (his parents being German), who went back to Germany to marry Regina Schoyer, Richard's mother. They had three daughters besides Richard, their eldest son.

Richard won a scholarship to St Paul's School, London (curiously enough, Joan Robinson was educated in the girls’ section of the same school). Then he won a scholarship to King's College, Cambridge, where he studied mathematics and physics, and graduated in physics in 1927 (being placed in the second class of the Natural Science Tripos). The scholarship gave him the right to a fourth year and he took up economics, at a time of great effervescence in Cambridge intellectual circles.

He was taught economics (in 1927–8) at King's College by Keynes and Shove, and attended university lectures delivered by Pigou, Keynes, Shove, Dennis Robertson and, in the following academic year, by Piero Sraffa, the Italian economist who had just arrived at Cambridge. He obtained his university degree in economics in June 1928 (placing himself, after only one year, in the first class of the Economic Tripos Part II), and immediately, under strong encouragement from Sraffa and Keynes, started work on a Fellowship dissertation (under the title ‘The Economics of the Short Period’), which he wrote in a surprisingly short time, obtaining a Fellowship of King's College in March 1930.

‘The Economics of the Short Period’ (Kahn, 1929), which has remained unpublished (though a

translation appeared in Italian, in 1983), is one of the two substantial works (the other being Joan Robinson's *Economics of Imperfect Competition*, 1933) that were stimulated in Cambridge by the devastating critique of Marshall launched by Piero Sraffa in the late 1920s (Sraffa, 1926). Richard Kahn and Joan Robinson worked very much in collaboration, under the strong influence of Sraffa and Keynes. For Kahn and Joan Robinson this was the beginning of an intense intellectual partnership that lasted for life.

The most interesting part of Kahn's ‘Economics of the Short Period’ is perhaps his analysis of the extent to which – in periods of depressions – market imperfections affect the way in which output gets distributed among the various firms, the essential point being that market imperfections prevent the most efficient firms from reaching an optimum utilization of their productive capacities and instead cause *all* firms (efficient and inefficient alike) to reach equilibrium at a point at which there is under-utilization of productive capacity at less than full employment. This sets obvious relations between the microeconomic behaviour of the single firms and the situations of underutilization of productive capacity for the economic system as a whole.

The only part of the dissertation that reached publication in English is the treatment (part of Chapter 7) of duopoly, which Kahn re-elaborated in the form of an elegant article (Kahn, 1937) that has since become a standard reference in the economic literature on duopoly and oligopoly. But the whole of Kahn's dissertation deserves closer scrutiny. When it becomes more readily available, it may well contribute to piecing together the great analytical puzzle of the relations between Sraffa's critique of Marshall's theory of the firm and Keynes's macroeconomic theory, or, to put it in other terms, of the microfoundations of Keynes's *General Theory*. It will also contribute to clarify the role played by Kahn in Joan Robinson's *Economics of Imperfect Competition*, whose Preface, as is well known, contains heavy acknowledgements of Kahn's help.

There can be no doubt that on a strictly intellectual level these were the most productive years in Kahn's life. It was in the summer of 1930, in the process of criticizing a paper by Keynes and Henderson on public works, that he discovered the principle of the multiplier.

The multiplier is a relation between the *increase* in exogenous aggregate expenditure and the *increase* in net national product thereby generated (and thus also in employment, if employment is proportional to net national product and the economy is in a situation of unemployment due to lack of effective aggregate demand). If c is the fraction of any increase in income that consumers tend to spend, it can be shown that any increase of £1 of exogenous expenditure (or else of such an amount that generate 1 extra job) will finally generate £$1/(1-c)$ of net national product (or else $1/(1-c)$ extra jobs). This is Kahn's multiplier. The author originally presented it in a short article with reference to employment (1931). It was then to be used by Keynes with reference to national income (and to the process of investments generating a corresponding amount of savings), as one of the major ingredients of Keynes's revolutionary work.

In 1930, Kahn started chairing and conducting the so-called ‘Cambridge Circus’, a group (or rather a closed club) of young Cambridge economists, among whom the most prominent members, besides Kahn, were Joan and Austin Robinson, Piero Sraffa and James Meade. The Cambridge Circus met regularly to discuss, criticize and propose changes to the subsequent drafts of what was later to become Keynes's *General Theory*.

The exact nature and extent of the part played by Kahn in Keynes's masterpiece will remain a matter of speculation. Schumpeter's view that Kahn's ‘share in the historic achievement cannot have fallen very

far short of co-authorship' (Schumpeter, 1954, p. 1172) may well be exaggerated. Yet, from Keynes's acknowledgement of indebtedness to him, it can surely be argued that that part must have been very large.

But this was not all. Kahn's contributions to economic theory in those years also concern two other debated subjects in the 1930s: the development of the concept of elasticity of substitution among 'factors of production', as an analytical tool in the traditional theory of income distribution (Kahn, 1933), and the laying out of the foundations of welfare economics. Kahn's notes on 'ideal output' (1935), and his article on tariffs and the terms of trade (1947) were later to be basic to de Villiers Graaff's systematic (and rather pessimistic) theoretical work on welfare economics (Graaff, 1957).

Kahn was appointed a university lecturer in 1933 and became a member of the King's economics teaching staff in 1936. Except for an interruption due to the war, he was responsible for the teaching of economics at King's College from 1936 to 1951 (first with Shove and Keynes and later, from 1949, with Kaldor). From 1939 to 1944, during the Second World War, he worked for the British government on various schemes, mainly connected with war production and war rationing. He also became head of the General Division of the Board of Trade. At the end of the war he returned to economics teaching in King's College. He was appointed a professor of economics at the University of Cambridge in 1951, and retired from that professorship in 1972.

On many occasions, and for temporary periods, Kahn worked for various international organizations: in 1955 he was a member of the Research Unit of the Economic Commission for Europe; in 1959 he was a member of a Group of Experts of the OEEC to study the problem of rising prices; in the 1960s he served as a member of four Groups of Experts of UNCTAD. In this capacity, he contributed to numerous official publications, both of British and of international organizations (see, for example, Kahn et al., 1956; 1961). Of considerable importance has been his Memorandum of Evidence submitted to the Government-appointed Committee on the Working of the Monetary System (the Radcliffe Committee) (1960). This Memorandum, jointly with his theoretical work on the extension of the concept of liquidity preference (1954a), was among the substantial pieces behind the formation of what has become known as 'the Radcliffe Committee view' on the working of the monetary system. When in the 1970s the more traditional 'monetarist' views once again became fashionable, Kahn was consistent in reacting vehemently against them and in rallying to the defence of the Keynesian approach (1976a; 1976b).

In 1965, Kahn was elevated to the House of Lords as a life peer (taking the title of Baron Kahn of Hampstead), in recognition of his services to the British government. Although a member of the Labour Party, he sat on the cross benches. For a number of years he was a reasonably active member of the House of Lords and made a number of speeches – almost exclusively on issues of government economic policy.

Kahn remained, for the whole of his life, a strong defender and faithful expositor of the original ideas contained in Keynes's major work. In a book reproducing his Mattioli Lectures and published in 1984, he gave his version – no doubt a version from the spot closest to the master – of how Keynes's historic work came into being. In one respect Kahn did go on to new ground to complete Keynes's views, and that was with reference to the inevitability of inflationary pressures in industrialized countries, once full employment is reached, unless some drastic changes are introduced into our institutions. On this subject he explored in considerable detail those institutional changes that he thought should be introduced into the process of wage negotiations (1976a; 1976b; 1977). He also took a major part in the shaping of the post-Keynesian theories of capital, growth and income distribution, as opposed to the neoclassical

theories (see, for example, Kahn, 1954b; 1959), as well as in the development of a post-Keynesian approach to planning (1958).

On the whole, Kahn was not a prolific writer. Apart from his Fellowship dissertation (still unpublished in its original version) and the publication of his Mattioli Lectures (Kahn, 1984), the only book that can be found in the library under his name is Kahn (1972), which is not in fact a proper book but the collection of his best articles, arranged together and published by two of his pupils on the occasion of his retirement from his Cambridge professorship. His most remarkable contribution to economic theory clearly remains the multiplier. But he will also be remembered as one of the crucial members of the post-Keynesian group of critics of neoclassical economics, although rarely did he come out in the battle forefront. By temperament, he constantly preferred to play the role of the meticulous scholar, never completely satisfied with any version of any work, of the relentless, sometimes even fastidious, critic, of the propounder of new alternatives and ideas to suggest to others to develop. In a few words, Kahn superbly played the role – rather congenial to him – of the *éminence grise* behind the scene. Precisely for these reasons, Kahn's association with Keynes in the late 1920s and the 1930s and his lifelong intellectual partnership with Joan Robinson will stimulate the imagination and curiosity of historians of economic thought for years to come.

Selected works

1929. The economics of the short period. Unpublished Fellowship Dissertation, Cambridge, King's College, 319. Italian translation, *L'economia del breve periodo*. Turin: Boringhieri, 1983.

1931. The relation of home investment to unemployment. *Economic Journal* 41, 173–98. Reprinted in Kahn (1972).

1933. The elasticity of substitution and the relative share of a factor. *Review of Economic Studies* 1 (October), 72–8.

1935. Some notes on ideal output. *Economic Journal* 45, 1–35.

1937. The problem of duopoly. *Economic Journal* 47, 1–20.

1947. Tariffs and the terms of trade. *Review of Economic Studies* 15(1), 14–19.

1954a. Some notes on liquidity preference. *Manchester School of Economic and Social Studies* 22 (September), 229–57. Reprinted in Kahn (1972).

1954b. (With D.G. Champernowne.) The value of invested capital: a mathematical addendum to Mrs Robinson's article. *Review of Economic Studies* 21(2), 107–11. Reprinted as an appendix to Robinson (1956).

1956. (With others.) *Economic Survey of Europe in 1955*. Geneva: United Nations.

1958. The pace of development. In *The Challenge of Development*, The Eliezer Kaplan School of Economics and Social Sciences, The Hebrew University, Jerusalem. Reprinted in Kahn (1972).

1959. Exercises in the analysis of growth. *Oxford Economic Papers* 11(June), 143–56. Reprinted in Kahn (1972).

1960. Memorandum of evidence submitted to the Radcliffe Committee (27 May 1958). Committee on the working of the Monetary System, *Principal Memoranda of Evidence*, vol. 3. London: HMSO. Reprinted in Kahn (1972).

1961. (With others.) *The Problem of Rising Prices*. Paris: OECC.

1972. *Selected Essays on Employment and Growth*. Cambridge: Cambridge University Press.

1976a. Thoughts on the behaviour of wages and monetarism. *Lloyds Bank Review* 119(January), 1–11.

1976b. Inflation – a Keynesian view. *Scottish Journal of Political Economy* 23(1), 11–16.

1977. (With M.V. Posner.) Inflation, unemployment and growth. *National Westminster Bank Quarterly Review*, November, 28–37.

1984. *The Making of Keynes' General Theory*. Cambridge: Cambridge University Press.

Bibliography

Graaff, J.de V. 1957. *Theoretical Welfare Economics*. Cambridge: Cambridge University Press.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36, 535–50.

How to cite this article

Pasinetti, Luigi L. "Kahn, Richard Ferdinand (1905–1989) ." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000002> doi:10.1057/9780230226203.0879

The New Palgrave Dictionary of Economics Online

Kahneman, Daniel (born 1934)

Matthew Rabin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A prominent figure in the new hedonic psychology, Daniel Kahneman has been influential in the emergence of behavioural economics. His research programme on heuristics and judgemental biases points to a range of important ways that economics has traditionally misunderstood human behaviour, and identifies how some common economic assumptions have been misleading. Especially in light of potential flaws in the way people manage their well-being, Kahneman and his colleagues have launched research that may help move economics towards a more realistic approach both to predicting behaviour and to welfare analysis.

Keywords

behavioural economics; biases; cognition; conjunction effect; conjunction rule; diminishing sensitivity; endowment effect; fairness; framing effect; happiness; hedonic psychology; heuristics; judgement; Kahneman, D.; law of small numbers; loss aversion; probability; representativeness heuristic; risk aversion; Thaler, R.; Tversky, A.; utility theory; well-being

Article

Daniel Kahneman was born in 1934, after his Jewish parents had emigrated from Lithuania to France. He grew up in France under German occupation, and then moved as an adolescent to what is now Israel. After studying in Israel, he entered the Ph.D. programme in psychology at the University of California in Berkeley in 1961. His intellectual journey continued from there, as he grew from a young psychologist specializing in vision research into a highly influential social and cognitive psychologist. Kahneman is a major figure in the field of psychology – his work with Amos Tversky on heuristics and biases was central to the ‘cognitive revolution’ in social psychology, and his later work on the psychology of happiness has made him a prominent figure in the new hedonic psychology. Kahneman's contribution to economics, outside his own field, is more remarkable. He is one of the two psychologists (along with his

long-standing collaborator Tversky) to most influence the field of economics. Along with Tversky, Richard Thaler, and a few others, he is one of the primary founding figures in the emerging field of ‘behavioural economics’. In recognition of his contributions, in 2002 he was awarded the Nobel Memorial Prize in Economics.

Kahneman's research presents a menu of important ways that economics has traditionally misunderstood human behaviour, and hence plays the useful role of identifying ways in which common economic assumptions have been misleading. Yet Kahneman has not just been a thorn in the side of economics; his research provides the material to improve economics. Although he has not himself specialized in conventional economic analysis, a recent spate of behavioural economic research – which attempts to improve economic analysis by incorporating greater psychological realism into economics – has been built on different strands of Kahneman's research.

Heuristics and biases

Economics has traditionally assumed that, when making decisions under uncertainty, people form subjective probabilistic assessments about the state of the world derived correctly from to the laws of probability. Kahneman's influential early research, conducted jointly with Tversky, documents departures from rationality in probabilistic judgement and decision-making under uncertainty. As Tversky and Kahneman (1974, p. 1124) frame their agenda,

> … people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors. The subjective assessment of probability resembles the subjective assessment of physical quantities such as distance or size. These judgments are all based on data of limited validity, which are processed according to heuristic rules. For example, the apparent distance of an object is determined in part by its clarity. The more sharply the object is seen, the closer it appears to be. This rule has some validity, because in any given scene the more distant objects are seen less sharply than nearer objects. However, the reliance on this rule leads to systematic errors in the estimation of distance. Specifically, distances are often overestimated when visibility is poor because the contours of objects are blurred. On the other hand, distances are often underestimated when visibility is good because the objects are seen sharply. Thus, the reliance on clarity as an indication of distance leads to common biases. Such biases are also found in the intuitive judgment of probability.

Many other approaches to the study of cognition and judgement have been explored by psychologists. Yet, by dint of its study of systematic but limited departures from a normative Bayesian model, Kahneman and Tversky's research programme on judgemental biases has shown the most promise for integration into economics.

Probably the most important biases Kahneman and Tversky identified are an array of related phenomena collected under the rubric of ‘the representativeness heuristic’. Although Bayesian updating tells us that people ought use conditional probabilities as a clue to underlying states – somebody who has symptoms

of a disease is more likely to have the disease – Kahneman and Tversky and subsequent researchers demonstrate that people tend to *overuse* 'representativeness' in assessing probabilities. One implication of this is the tendency to underuse base rates: even if a certain symptom appears always among people with a rare disease, and only occasionally among people without the disease, given the rarity of the disease Bayesian reasoning tells us that most people who have the symptom do not have the disease. Yet people tend to exaggerate the likelihood of having the disease given the symptom.

The most striking manifestation of such base-rate neglect is the common violation of the conjunction rule, a fundamental axiom of probability theory: the probability that somebody belongs to both categories A and B is less than or equal to the probability that she belongs to category B alone. Kahneman and Tversky demonstrate what they call the *conjunction effect*: when a description is representative of a person in category A but not of a person in category B, people often judge it more likely that the description matches somebody who falls into both categories A and B than into category B alone. Tversky and Kahneman (1982b, p. 92) illustrate this effect by recounting an experiment in which subjects were provided with the following description:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Subjects were then asked to rate the relative likelihood that eight different statements about Linda were true. Two statements on the list were 'Linda is a bank teller' and 'Linda is a bank teller and is active in the feminist movement'. Over 85 per cent of subjects judged it more likely that Linda was both a bank teller and a feminist than that she was a bank teller. This is because the description of Linda made her seem like a feminist, so that being a bank teller and a feminist seemed a more natural description, and thus more 'representative' of Linda, than simply being a bank teller.

Tversky and Kahneman (1971) argue that another manifestation of the representativeness heuristic is a bias they call 'The Law of Small Numbers': people exaggerate how often a small sample closely resembles the parent population or underlying probability distribution that generates the sample. We expect even small classes of students to contain very close to the typical distribution of smart ones and not-so-smart ones. Likewise, we underestimate how often a good financial analyst will be wrong a few times in a row, and how often a clueless analyst will be right a few times in a row. Such misunderstandings of variance in small samples have far-reaching implications for social and economic judgements.

Kahneman and Tversky identified and provided evidence for other heuristics and biases. Beyond the biases that Kahneman and Tversky have themselves documented, however, research on other biases (for example, the hindsight bias) that are likely to be quite important for economics has benefited from the more general research programme to which they have centrally contributed. Indeed, the early collection of papers edited by Kahneman, Slovic, and Tversky (1982c) serves as a sort of early bible of the programme, and the later collection edited by Gilovich, Griffin, and Kahneman (2002) presents much of the subsequent research indicating the insights of this research programme.

Loss aversion, prospect theory, and choice under uncertainty

One of the most important and widely cited papers in economics of recent decades is Kahneman and Tversky's (1979) ‘Prospect Theory: An Analysis of Decision Under Risk’. (An ISI Web of Science search in April 2006 indicates that this is the second most widely cited article in *Econometrica*, and the only article among the top ten most cited that was not concerned with econometric theory.)
Humans are typically more sensitive to how an outcome contrasts with reference levels than to the absolute level of the outcome itself. Kahneman and Tversky (1979, p. 277) stress that the salience of changes from reference points is a basic aspect of human nature:

> An essential feature of the present theory is that the carriers of value are changes in wealth or welfare, rather than final states. This assumption is compatible with basic principles of perception and judgment. Our perceptual apparatus is attuned to the evaluation of changes or differences rather than to the evaluation of absolute magnitudes. When we respond to attributes such as brightness, loudness, or temperature, the past and present context of experience defines an adaptation level, or reference point, and stimuli are perceived in relation to this reference point (Helson, 1964). Thus, an object at a given temperature may be experienced as hot or cold to the touch depending on the temperature to which one has adapted. The same principle applies to non-sensory attributes such as health, prestige, and wealth. The same level of wealth, for example, may imply abject poverty for one person and great riches for another – depending on their current assets.

In the context of utility theory, people often feel the effects of changes and contrasts more intensely than absolute levels. (While the role of reference levels in decision making is often inconsistent with fully rational behaviour, Tversky and Kahneman, 1991, and others have shown that many reference-level effects can be captured within the framework of utility theory.)
Kahneman and Tversky identify two pervasive ways in which reference levels influence preferences and choice. First, people are *loss averse*: in a wide variety of domains, people are more averse to losses relative to their reference level than they are attracted to same-sized gains. Second, people exhibit *diminishing sensitivity*: the marginal change in perceived well-being is greater for changes that are close to one's reference level than for changes that are further away. While these features of preferences have much broader implications, the most important and striking are in the context of choice involving monetary uncertainty. Loss aversion implies that people are significantly ‘risk averse’ for even small amounts of money. People dislike losing $10 more than they like gaining $11, and hence prefer their status quo to a 50–50 bet of losing $10 or gaining $11. While such ‘first-order’ risk aversion is widely observed, the standard concave-utility function implies that people are close to risk-neutral for small stakes. Diminishing sensitivity also has a provocative implication for risk preferences: while people are likely to be risk averse over gains, they are often risk *loving* over losses. For instance, Kahneman and Tversky (1979) find that 70 per cent of subjects report that they would prefer a 3/4 probability of losing nothing and 1/4 probability of losing $6,000 to a 1/2 probability of losing nothing and 1/4 probability each of losing $4,000 or $2,000. The preferred lottery here is a mean-preserving spread of the less-preferred lottery; hence, the responses of 70 per cent of the subjects are inconsistent with the standard concave utility-for-wealth assumption. Subsequent evidence has suggested more varied and context-specific features of risk attitudes in the loss domain, but it supports the finding that people more often have risk-seeking preferences over modest-scale losses than over modest-scale gains.

Another important phenomenon attributed by most researchers to loss aversion is the striking *endowment effect* identified by Thaler (1980; 1985), and subsequently fleshed out by Kahneman, Knetsch and Thaler (1990). Once a person comes to possess a good, she immediately values it more than before she possessed it. An experiment in the latter paper nicely illustrates this phenomenon. A decorated mug (worth about $5) was placed in front of each of one-third of a group of students. Prices at which the subjects were willing to sell their mugs were elicited in a way that made it optimal for subjects to be truthful. Other subjects were asked to give the minimal amount of money that they would prefer to receiving that mug. These two groups faced exactly the same choice, but differed in their reference level – for sellers, losing the mug was a loss, while for ‘choosers’ no loss was involved. The average selling price was about $7.00, and the average exchange value for choosers was about $3.50. The difference in these amounts reflects an instantaneous effect of owning on object on the valuation of that object. Such an endowment effect is usefully conceptualized as a case of loss aversion. Individuals treated the endowed mugs as part of their reference levels, and considered subsequently not having a mug to be a loss, whereas individuals without mugs considered not having a mug as remaining at their reference point. The inducement of a nearly instantaneous endowment effect suggests that reference points ubiquitously influence decision making – with potentially significant economic consequences.
Besides this value function, a second important element of Kahneman and Tversky's (1979) multifaceted prospect theory is the fact that people do not evaluate uncertain prospects in the linear-in-probabilities way conventionally assumed by economists. Kahneman and Tversky argue that people maximize with respect to a monotonic nonlinear function of probabilities with the following properties: they ignore *very* low probability events, but among events they don't ignore low probabilities are overweighted and moderate and high probabilities are underweighted, and the latter effect is more pronounced than the former. Tversky and Kahneman (1992, p. 297) conclude that decision weights and the value function combine to imply ‘a distinctive fourfold pattern of risk attitudes: risk aversion for gains and risk seeking for losses of high probability; risk seeking for gains and risk aversion for losses of low probability’.

Framing effects

Probably the most striking and problematic departure from rationality emphasized in Kahneman and Tversky's early research is what they call the *framing effect*: two logically equivalent statements of a problem lead decision-makers to choose different options. Examples of framing effects typically involve differing frames whose logical equivalence is neither totally transparent nor terribly obscure. Because losses resonate with people more than gains, for instance, a frame that highlights the losses associated with a choice makes that choice less attractive. Tversky and Kahneman (1986, p. S260) demonstrate framing effects in a public-health context, asking subjects the following hypothetical question:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.

> If Program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved.

Seventy-two per cent of subjects said that they preferred Program A over B. But they also asked another group of subjects the same question with the two programs framed thusly:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die.

In this second group, 78 per cent preferred Program D. Although A vs. B and C vs. D are precisely the same choices, framing the choice in terms of numbers of lives saved clearly evokes ‘risk aversion’ in gains – better to save 200 lives for sure than an uncertain number of lives averaging 200. Framing the choice in terms of number of victims dying evokes ‘risk-loving’ attitudes in losses – the chance of preventing any deaths is very attractive. (While these and other questions are hypothetical, Tversky and Kahneman found similar effects among experienced physicians judging cancer treatment, suggesting that similar patterns might play out in the real world. Moreover, similar framing effects were found in choices over lotteries with small monetary stakes.)

Perhaps the most fundamental example of a framing effect – whose centrality to understanding risk attitudes and other economic choices has only recently begun to be fully appreciated by researchers – is that people assess risky prospects in isolation, rather than by aggregating them. As an illustration of isolation errors, Tversky and Kahneman (1986, p. S255) ask subjects to ‘Imagine that you face the following pair of concurrent decisions. First examine both decisions, then indicate the options you prefer.’ To simplify, the choices were:

> Choose between:
>
> - •••A: $240 for sure
> - •••and
> - •••B: (.25+$1,000, .75 $0)
>
> Choose between:
>
> - •••C: – $750 for sure
> - •••and
> - •••D: (.75 – $1,000, .25 $0)

Eighty-four per cent of subjects chose A over B and 87 per cent chose D over C – in accordance with the principles of diminishing sensitivity. But, when subjects’ choices for both decisions were combined, 73 per cent chose the combination AD, 11 per cent chose AC, 14 per cent chose BD, and three per cent chose BC. The problem with these choices is that AD is in fact a 75 per cent chance of losing $760 and

25 per cent chance of no change, while BC is a 75 per cent chance of losing $750 and 25 per cent of no change. BC is clearly better than AD. The fact that most people made the choice AD when asked to choose separately clearly indicates that they did not integrate the decisions. (Indeed, in groups of other subjects asked to make just one of the A vs. B or C vs. D choices in isolation, 85 per cent chose A over B and 86 per cent chose D over C, virtually the same as those choosing for both choices.) Such examples were also observed for real (but smaller) monetary stakes by Tversky and Kahneman and subsequent researchers. In recent years, this notion that risk attitudes are fundamentally influenced by 'narrow framing' has become a major theme of research as researchers have begun to establish that loss aversion is not itself a sufficient explanation for modest-scale risk aversion without such narrow framing. (If people hated losses but integrated their losses and gains across different choices, they would become de facto risk neutral by cancelling out losses with gains.)

Fairness

Many economists have over the years discussed the existence and economic implications of preferences that depart from pure self-interest, as narrowly defined. But much of the credit for introducing the empirical study of the economic implications of fairness judgements into economics should go to Kahneman, Knetsch, and Thaler (1986). Their interest is positive, not normative: instead of studying normative standards of what we as policymakers (or philosophers) might consider fair allocations or appropriate social-welfare functions, they study with surveys what a typical economic actor might assess as fair or unfair behaviour. For instance, they asked subjects to assess the fairness of reducing the wages of current employees as opposed to hiring new employees at lower wages after normal turnover in response to market unemployment. They found that respondents are likely to consider lowering wages to current workers unfair, while they consider using market conditions to set new wages acceptable. Respondents also considered it unfair to raise the price of peanut butter already in stock in response to a rise in the wholesale price of peanut butter – much as people protest when gas stations immediately raise prices on petrol in stock in response to an increase in wholesale petrol prices. Kahneman, Knetsch, and Thaler identify some more general principles with their surveys: people generally find it acceptable for firms to raise prices or lower wages in response to concurrent shifts in their costs, but not in response to demand shifts or to shortages.

Decision utility, experienced utility, and happiness

The research on heuristics and biases discussed above indicates that people misjudge the probabilistic consequences of their decisions. But a spate of recent research suggests that, even when they correctly perceive the *physical* consequences of choices they make, people may systematically misperceive the *hedonic* consequences of those choices. As Kahneman (1994, pp. 20–1) argues,

> … it may be rash to assume as a general rule that people will later enjoy what they want now. The relation between preferences and hedonic consequences is better studied than postulated. These considerations suggest an explicit distinction between two notions of utility. The *experienced utility* of an outcome is the measure of the hedonic experience of

> that outcome. … The *decision utility* of an outcome, as in modern usage, is the weight assigned to that outcome in a decision.

Kahneman (2000) summarizes and provides a conceptual framework for understanding utility misprediction, and identifies several errors in this domain. He argues, for instance, that in forecasting future utility people tend to use the '*transition rule*: predictions of a person's initial reaction to a new situation, which may be quite accurate in itself, is incorrectly used as a proxy to forecast the long-term effects of that situation' (2000, p. 703). The most important error resulting from this is the tendency to under-appreciate the hedonic effects of adaptation, leading people to exaggerate changes in utility caused by small and big changes in their lives.

Kahneman and fellow researchers have also conducted a series of experiments demonstrating that another source of misprediction of future utility actually comes from a biased evaluation of past episodes: even when people might well recollect the momentary hedonic sensations from past experiences, they might be bad at 'adding up' these hedonic sensations from extended episodes. Through comparing moment-by-moment evaluation with retrospective evaluation of episodes (such as unpleasant medical procedures), they show that people are biased in sundry ways: retrospective evaluation tends to be over-influenced by such factors as extreme moments and final moments, and people are subject to 'duration neglect', with intensity of an episode looming much larger than its duration.

More generally, especially in light of potential flaws in how people manage their well-being, Kahneman and others have launched research that may help move economics towards a more realistic approach to welfare analysis. As an editor of and contributor to a recent volume (Kahneman, Diener and Schwarz, 1999), indeed, Kahneman is a leader in the exciting new focus in social science and public policy on the study of what makes people happy.

Conclusion

Daniel Kahneman, despite his accolades and influence in economics, is a psychologist and self-identifies as such. Although many of the examples and motivations in his research are quite directly inspired by economic concerns, his research rarely constitutes traditional economic analysis per se. But with the gradual rise of 'behavioural economics' as a field of research, and more recently as this research programme has moved into the mainstream, the insights established in his research has become ever more widely influential.

See Also

- behavioural public economics
- happiness, economics of

Selected works

1971. (With A. Tversky.) Belief in the law of small numbers. *Psychological Bulletin* 76, 105–10.

1972. (With A. Tversky.) Subjective probability: a judgment of representativeness. *Cognitive Psychology* 3, 430–54.

1973. (With A. Tversky.) Availability: a heuristic for judging frequency and probability. *Cognitive Psychology* 5, 207–32.

1973. (With A. Tversky.) On the psychology of prediction. *Psychological Review* 80, 237–51.

1974. (With A. Tversky.) Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–31.

1979. (With A. Tversky.) Prospect theory: an analysis of decisions under risk. *Econometrica* 47, 263–91.

1982a. (With A. Tversky.) On the study of statistical intuitions. *Cognition* 11, 123–41.

1982b. (With A. Tversky.) Judgments of and by representativeness. In *Judgment under Uncertainty: Heuristics and Biases*, ed. D. Kahneman, P. Slovic and A. Tversky. New York: Cambridge University Press.

1982c. (With P. Slovic and A. Tversky, eds.) *Judgment under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press.

1986. (With A. Tversky.) Rational choice and the framing of decisions. *Journal of Business* 59, S251–S278.

1986. (With J. Knetsch and R. Thaler.) Fairness and the assumptions of economics. *Journal of Business* 59, S285–S300.

1990. (With J. Knetsch and R. Thaler.) Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98, 1325–48.

1991. (With J. Knetsch and R. Thaler.) The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* 5(1), 193–206.

1991. (With R. Thaler.) Economic analysis and the psychology of utility: applications to compensation policy. *American Economic Review (Proceedings)* 81, 341–52.

1991. (With A. Tversky.) Loss aversion in riskless choice: a reference-dependent model. *Quarterly Journal of Economics* 106, 1039–61.

1992. (With A. Tversky.) Advances in Prospect theory: cumulative representation of uncertainty.

Journal of Risk and Uncertainty 5, 297–323.

1993. (With D. Lovallo.) Timid choices and bold forecasts: a cognitive perspective on risk taking. *Management Science* 39, 17–31.

1994. New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics* 150, 18–36.

1998. (With I. Ritov and D. Schkade.) Economists have preferences, psychologists have attitudes: an analysis of dollar responses to public issues. Mimeo, Princeton University.

1999. (With E. Diener and N. Schwarz.) *Well-being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation.

2000. Evaluation by moments: past and future. In *Choices, Values, and Frames*, ed. D. Kahneman and A. Tversky. New York: Russell Sage Foundation and Cambridge University Press.

2000. (With A. Tversky, eds.) *Choices, Values, and Frames*. New York: Russell Sage Foundation and Cambridge University Press, 2000.

2002. (With T. Gilovich and D. Griffin, eds.) *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Bibliography

Helson, H. 1964. *Adaptation-Level Theory*. New York: Harper.

Krugman, P. How I work. Online. Available at http://web.mit.edu/krugman/www/howiwork.html, accessed 9 September 2006.

Thaler, R. 1980. Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization* 1, 39–60.

Thaler, R. 1985. Mental accounting and consumer choice. *Marketing Science* 4, 199–214.

How to cite this article

Rabin, Matthew. "Kahneman, Daniel (born 1934)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000065> doi:10.1057/9780230226203.0880

The New Palgrave Dictionary of Economics Online

Kain, John Forest (1935 – 2003)

Eric A. Hanushek

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

housing markets and policy; Kain, F. K.; residential segregation; school achievement; spatial mismatch hypothesis; urban economics; urban simulation models

Article

John Kain was an empirical economist who significantly changed analysis and modeling in urban economics. Modern urban economics was extensively developed during the 1960s, and Kain's contributions were particularly important in several key areas. His most famous line of inquiry revolved around the interactions of race and urban location and the importance of housing segregation for black welfare. He was also one of the early pioneers in developing general equilibrium urban simulation models that were capable of addressing interesting and important policy questions. His analyses of urban transportation policies have been influential in both developed and developing countries (Meyer, Kain and Wohl, 1966). A fourth significant endeavour, while having some of the same underpinnings, went to the issues of educational achievement.

His influential paper on the spatial mismatch hypothesis started a large line of inquiry (Kain, 1968; 1992). The innovative idea was that housing segregation kept blacks in areas that were increasingly farther from jobs (which were rapidly decentralizing from more central locations). As commuting to work became more costly, black employment suffered. Kain's connection of urban location, housing, and labour markets was a true innovation. A second important inquiry was the investigation of how segregation affected black housing costs and home ownership (Kain and Quigley, 1972). His early urban simulation models were developed to permit investigation of how multiple housing and job locations interact with a variety of housing policies and urban dynamics (for example, Ingram, Kain, and Ginn, 1972).

These urban studies derived from his intense interest in the intersection of geography, schools and race. In a different direction, he originated the use of large-scale administrative databases on school achievement to study the elements of human capital formation (Rivkin, Hanushek and Kain, 2005). But, again, he emphasized the fundamental influence of race on opportunities and outcomes (Hanushek, Kain

and Rivkin, 2004).
Ultimately, one of his most important and lasting influences was legitimizing the study of the economics of race through showing its fundamental importance to a range of social issues. Before his systematic work, few economists considered the economic influence of race and segregation.
He received his Ph.D. from the University of California at Berkeley. Most of his career was spent in the Department of Economics at Harvard University, although he also taught at the US Air Force Academy and the University of Texas at Dallas. His last position at the University of Texas at Dallas led to his development of the extensive stacked panel databases on school performance.

See Also

- housing policy in the United States
- urban economics

Selected works

1966. (With J. R. Meyer and M. Wohl.) *The Urban Transportation Problem*. Cambridge, MA: Harvard University Press.

1968. Housing segregation, Negro employment and metropolitan decentralization. *Quarterly Journal of Economics* 82, 175–97.

1972. (With J. R. Ginn and G. K. Ingram.) *The Detroit Prototype of the NBER Simulation Model*. New York: NBER.

1972. (With J. M. Quigley.) Housing market discrimination, homeownership and savings behavior. *American Economic Review* 62, 263–77.

1992. The spatial mismatch hypothesis: three decades later. *Housing Policy Debate* 3, 371–460.

2004. (With E. A. Hanushek and S. G. Rivkin.) Disruption versus Tiebout improvement: the costs and benefits of switching schools. *Journal of Public Economics* 88, 1721–46.

2005. (With E. A. Hanushek and S. G. Rivkin.) Teachers, schools, and academic achievement. *Econometrica* 73, 417–58.

How to cite this article

Hanushek, Eric A. "Kain, John Forest (1935–2003)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_K000071> doi:10.1057/9780230226203.0881

The New Palgrave Dictionary of Economics Online

Kaldor, Nicholas (1908 – 1986)

Adrian Wood
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

buffer stocks; capital theory; cobweb theorem; compensation tests; decreasing returns; distribution theory; division of labour; economic growth; entrepreneurship; excess capacity; expenditure tax; firm, theory of; foreign trade multiplier; imperfect competition; incomes policies; increasing returns; infant-industry protection; Kaldor, N.; labour productivity; money supply; multiplier–accelerator models; nominal wages; non-price competition; oligopoly; polypoly; price formation; price leadership; real wages; relative wages; roundaboutness; steady-state models; tariffs; technical change; terms of trade; trade cycles; uneven development; Verdoorn's Law

Article

Nicholas Kaldor was born in Budapest. From 1927 to 1947 he studied and taught at the London School of Economics. Then, following two years at the Economic Commission for Europe in Geneva, he moved to Cambridge University, where he became a fellow of King's College and, in 1966, professor of economics. He was elevated to the peerage in 1974, as Baron Kaldor of Newnham. (For more biographical information, see Kaldor, 1960–89, vol. 1, pp. vii–xxxi, and Pasinetti, 1979.)

Kaldor was always passionately involved with practical problems of economic policy. As a (vigorously dissenting) member of a British Royal Commission in the early 1950s, he acquired international renown in the field of taxation. But he also constantly addressed the major domestic and international economic issues of the day – in books and journal articles, in letters to newspapers, in lectures and speeches, and through personal contacts. He was special advisor to the British Chancellor of the Exchequer in 1964–8 and 1974–6, and also gave advice to the governments of many other countries, and to various international organizations. Though a defender of the private enterprise market system, he consistently advocated government intervention to make capitalist economies more productive and equitable, and has devised many policies and instruments for this purpose.

Yet Kaldor's main intellectual interest, and the main basis of his fame as an economist, was always theory – simplified analytical description of how economies function. Driven from within by logic,

creativity and curiosity, he nonetheless became a committed advocate and exponent of the inductive method – that to be fruitful, theory must spring from (and constantly be tested against) direct observation of reality. His involvement in practical matters contributed to, as well as benefited from, his theoretical work; and he was a steady and perceptive consumer of statistics and the empirical work of other economists.

Range and evolution of Kaldor's thought

Kaldor's theoretical contributions cover an astonishing range. And on many topics, his views changed over time – sometimes evolving linearly, sometimes being rejected and replaced with contrary views. His nine volumes of *Collected Economic Essays*, with illuminating retrospective introductions, amply document the scope and chronological progress of his theorizing – important aspects of which are concealed in many of the ostensibly non-theoretical essays. Studies of his work include Nell and Semmler (1991), Targetti (1992) and Turner (1993).

During the 1930s, with Hicks, Hayek, Robbins and Scitovsky among his friends and colleagues, Kaldor made several notable contributions to mainstream neoclassical theory. He named the cobweb theorem, introduced the idea of compensation tests into welfare economics, and clarified the relationship between tariffs and the terms of trade. He also entered into the prevailing controversy on the theory of capital, defending the Austrian view against the criticisms of Knight (though he later came to accept that the concept of roundaboutness could not solve the problem of measuring capital in homogeneous units, and indeed to reject the whole idea of interpreting the rate of profit or interest as the marginal product of capital).

But much of his theoretical work during this period revolved around the firm and imperfect competition. Like others, he saw the supply curve as a weak link in competitive theory, arguing that diminishing returns to scale do not convincingly explain firm size, and hence that perfect competition is not compatible with equilibrium. He accordingly welcomed the emphasis that Joan Robinson and Chamberlin gave to demand-side limitations on firm size in imperfectly competitive markets, being particularly impressed by Chamberlin's reconciliation of imperfect competition with free entry. At the same time, however, he criticized them for assuming that firms in imperfectly competitive markets generally face conventional demand curves. More commonly, Kaldor contended, such firms must take direct account of the reactions of a relatively small group of rivals to their price and output decisions.

Though Kaldor never lost sight of the importance of microeconomic foundations, the publication of Keynes's *General Theory* greatly stimulated his interest in macroeconomics, as well as setting him on the path to becoming a committed critic of mainstream neoclassical theory. The first fruits of this change were two essays closely related to Keynes's own theorizing. One was on speculation and economic stability, focusing on the relationships between changes in demand and supply flows and changes in speculative stocks, and between price stability and income stability. Among other things, this essay argues that the long-term interest rate cannot adjust sufficiently to equate current saving and investment because it is constrained by the bond market's concept of a normal interest rate, which is simply the expected future average of short-term rates (themselves determined by the prevailing balance between the transactions demand for, and the supply of, money). The second essay – of equal analytical power, though less influential – concerned own-rates of interest.

There then followed some prominent contributions to trade cycle theory. Kaldor criticized simple

multiplier–accelerator models for having to be either unrealistically stable or unrealistically unstable. He argued that a convincing theory of purely endogenous cycles would need to be based on nonlinearities in the investment function, but also that in reality cycles are not purely endogenous. In particular, he suggested that entrepreneurial dynamism could cause cumulative upswings of investment that periodically crashed against exogenous barriers such as full employment or bottlenecks in the supply of capital goods. This approach, and perhaps also the sustained prosperity of the 1950s, gradually led him to become less interested in cycles as such, and much more interested in long-term economic growth.

The first main phase of his work on growth, which lasted until the mid-1960s, found expression in a series of formal steady-state models. Though similar to other such models in some respects, including the assumptions that full employment generally prevails and that long-term growth of output is governed by supply rather than demand, Kaldor's models have two important distinctive features. One is an original theory of distribution, whereby the share of profits in national income is determined by the share of investment, which in turn depends on the aggregate capital–output ratio and on the (independently given) aggregate growth rate. Another way of expressing this theory is to say that the aggregate profit rate on capital is determined by the aggregate growth rate – an approach further developed by Pasinetti (1962), and in Kaldor's subsequent neo-Pasinetti theorem.

The reason for this linkage between profits and investment is that the proportion of profits saved (by enterprises, not by rentiers) is much higher than the proportion of wages saved. As a result, an economy can achieve the aggregate saving rate needed to sustain any given growth rate through adjustment of the share of profits in its national income. This Kaldor regarded as a long-term generalization of a Keynesian principle – that investment determines savings, rather than the other way round. He opposed it to the neoclassical principle that ‘availability of capital’ constrains growth and governs the rate of profit.

The other distinctive feature of Kaldor's formal growth models is his technical progress function. He rejected the neoclassical concept of a production function that is shifted over time by autonomous technical progress, and indeed the whole notion that productivity gains due to capital accumulation are separable from those due to technical advance. Instead, he argued that the knowledge needed to increase productivity is acquired through a process of learning that is inseparable from the process of investment, and hence that the pace of applied technical progress depends on the rate of investment (which in turn depends on entrepreneurial expectations of profitability and risk). This principle was eventually formalized, in an ingenious vintage model developed with Mirrlees, as a relationship between the rate of change of gross investment per worker and the rate of increase in labour productivity on newly installed equipment.

During the 1960s, however, Kaldor's own empirical research and practical experience caused him to become deeply dissatisfied with formal macroeconomic growth models. Their microeconomic underpinnings seemed inadequate. They were excessively aggregated, and did not capture the different characteristics of (and critical relationships between) the various broad economic sectors. Their assumptions of full employment and of growth governed solely by supply appeared increasingly implausible. And they failed to come to grips with international economic linkages and the spatial pattern of development.

These concerns launched the second main phase of Kaldor's work on growth, in which several fundamentally new ideas displaced or modified some – but by no means all – of the principles of the first phase, and in which some of his earlier theoretical insights were reintroduced. This work was not,

however, embodied in formal models. Nor, more generally, did Kaldor ever attempt a comprehensive theoretical treatise, showing how his ideas in different areas fitted together, and how new ideas in particular areas affected his thinking in other areas.

For this reason, the rest of the present article will try to summarize the latest versions of Kaldor's theories in several areas, and to assemble them into a relatively complete Kaldorian view of how the world works. A number of questions can be raised about this view, some calling for theoretical amplification, others for more empirical research. But all theories have shortcomings; and economists choose among alternative theories ultimately on the basis of their strong points, not their weak points. It is thus on the strengths and insights of the Kaldorian view that this article will concentrate.

The process of economic growth

Kaldor's account of growth provides the context for most of his theorizing in other areas. Like Ricardo, he drew a sharp distinction between industry and primary production.

Increasing returns

In industry, especially in manufacturing, growth of output per worker arises principally from static and dynamic economies of scale, whose realization depends on (but also contributes to) expansion of markets for industrial products. Increasing returns, noted by Smith but subsequently emphasized by only a few economists such as Marx, Marshall and Allyn Young (who taught Kaldor at LSE), are a multifaceted and pervasive feature of industrial production. They often exist at plant and firm level. They are to be found also at the industry level, where larger scale permits greater internal specialization of production among different firms (many of which may in fact be small). Finally, increasing returns operate at the ‘macroeconomic’ level, partly because different industries stimulate each others’ development through demand and supply linkages, partly because all of them benefit from a common labour market large enough to justify the development of many highly specialized skills.

Increasing returns and technical progress are intimately related. This is because the construction and operation of larger-scale plants, the finer subdivision of production processes, and the emergence of more specialized skills, all require the development and application of new knowledge. Each path-breaking stage of realization of scale economies is initially painful and problematic; but effort and experience gradually eliminate the problems and realize the full potential of this technical stage, making it possible to plan and implement the next step forward. (Increasing returns are thus not simply a static function of the scale of production, but also of the cumulative amount of production over time.) Scientific advances, and better technical and general education, facilitate industrial growth, but do not drive it.

Much the same is true, in Kaldor's view, of the accumulation of physical capital. Sustained growth of labour productivity in industry requires investment, mainly because a finer division of labour can generally only be realized through increased mechanization – more capital per worker. This explains why the secular growth of modern industry has entailed continuing increases in both output per worker and the capital–labour ratio, but much less change in capital–output ratios (and why there are large differences in capital–labour ratios, but no systematic differences in capital–output ratios, between rich and poor countries). It also explains why rapid industrial growth is associated with high ratios of

investment to output. But it is not the investment that is generating the growth; rather, it is the growth that is generating the investment.

Increasing returns cause a strong statistical correlation, for example across countries, between growth of industrial labour productivity and growth of total industrial output. (This is known as Verdoorn's Law, after the economist who first noticed the correlation.) On average across industries, and over periods of a decade or more, industrial labour productivity nonetheless generally grows slower than total industrial output. This means that relatively rapid industrial labour productivity growth tends to involve relatively rapid growth of industrial employment.

Sectoral complementarities

Agriculture and mining are subject not to increasing but to decreasing returns. Over time, the productivity of most land is increased through accretion of technical knowledge; but in Kaldor's view, technical progress in primary production is more exogenous – less responsive to the need for it – than in industry, which means that there is a relatively inflexible upper limit on the rate of growth of primary production. Nor are the primary sectors subject to Verdoorn's Law – labour productivity growth is generally independent of output growth.

In a closed system, this technically determined upper limit on the rate of primary production growth is the main long-term constraint on the growth rate of industrial production, and hence on the growth of the whole economy. One reason for this is that expansion of industrial production requires increased amounts of food for industrial workers and of raw materials for processing. Relatedly, growth of primary production and incomes is a vital source of growth in demand for the products of industry. For industrial expansion cannot be self-sustaining, simply because a significant part of the incomes generated in industry is spent on non-industrial goods such as primary products.

To offset this leakage of industrial demand into other sectors, there must clearly be demand for industrial products from the incomes generated in other sectors. But Kaldor's position is stronger and more specific, namely, that growth of demand from the primary sectors in a closed economy actually *determines* the long-term growth rate of industrial production. This is because there is no enduring limit to growth within industry itself: the supply of industrial capital, labour, knowledge and skills will generally respond to whatever happens to be the rate of growth of overall demand for industrial products. It is also because expansion of demand for industrial products from within the industrial sector is in the long term passively determined by expansion of industrial production. (Kaldor regarded all this as a long-run generalization, albeit confined to industry, of the Keynesian principle that output is determined by effective demand, combined with Harrod's concept of the foreign trade multiplier and Hicks's concept of the supermultiplier.)

For primary production to constrain the growth of a closed economy, decreasing returns (and comparatively unresponsive technical advance) in agriculture and mining are essential, since otherwise primary output could be profitably increased to any required level simply by a larger allocation of capital, skilled labour and research expenditure. It is likewise important that the primary-industry terms of trade not be completely flexible, since otherwise any primary sector output constraint on industrial growth might be overcome by an increase in the prices of agricultural and mining products relative to industrial products, which could make a larger volume of primary output profitable, increase the purchasing power of primary producers over industrial goods, and switch some industrial purchasing

power from primary to industrial products. This does not happen, in Kaldor's view, because industrial wages – and hence industrial prices – are inflexible downwards in terms of their purchasing power over primary products, especially food. The primary-industry terms of trade thus cannot improve enough to prevent primary sector production from constraining the long-term pace of industrial growth.

A substantial share of output and employment in all economies of course derives from neither the primary nor the industrial sector, but from services. However, Kaldor argued that service sector expansion is not generally an active ingredient of economic growth, but rather a consequence and complement of expansion in other sectors, particularly industry. He also argued that in high-income countries the service sector acts as an industrial employment reservoir, since (for reasons discussed below) it usually contains a considerable proportion of underemployed workers, who are paid less than workers in industry and are thus willing and able to fill industrial vacancies as they arise. In low-income countries, by contrast, agriculture is the main reservoir of industrial labour. But in both sorts of countries the existence of these reservoirs is one of Kaldor's main reasons for arguing that expansion of industrial output is not normally constrained by availability of labour.

Cyclical interruptions

Economic growth is not mechanical or smooth. Even though its long-run path is governed by certain basic constraints and linkages, entrepreneurial dynamism – expectations of future growth – is what keeps the process going. Shocks and disturbances of many kinds are constantly disrupting the process in an upward or downward direction. But the resilience of entrepreneurial expectations normally tends to damp rather than to amplify these disturbances, and gives the long-term growth path a momentum sufficient to transcend temporary shocks. The momentum can be broken, however, in deep and sustained recessions, which may cause an enduring downward shift of business expectations and hence the actual growth rate for long periods to fall below the maximum imposed by technical advance in the primary sectors.

Kaldor regarded the volatility of primary product prices as an important source of economic instability, especially when expectations of normal prices for these products are weak (and hence movements in speculative stocks reinforce rather than offset price changes arising from demand or supply fluctuations). A large fall in primary product prices tends to retard industrial growth by slashing the purchasing power of primary producers over industrial products. But a large rise in primary product prices does not have symmetrical benefits for industry, since it tends to push up industrial wages and prices, and provokes governments to deflate. (For reasons of this kind, Kaldor strongly advocated international buffer stock schemes to stabilize the general level of primary product prices, ideally in the form of a commodity-backed world currency).

The resilience of entrepreneurial expectations in the face of temporary disturbances causes increasing returns to be a short-term as well as a long-term feature of industrial production. This is partly because firms deliberately expand capacity somewhat ahead of demand, partly because they base their employment decisions on medium-term prospects rather than immediate needs. As a consequence, and because increases and decreases in aggregate demand tend to be spread across efficient and inefficient firms alike, industrial labour productivity normally rises in booms and falls in slumps. Only in a severe recession, when business expectations are badly dented and financial reserves are exhausted, do the least efficient firms close down – thus checking the fall in productivity and causing the industrial sector to

display some signs of decreasing returns.

Spatial patterns and relationships

The ‘closed economy’ to which the growth model sketched above applies most directly is the world as a whole. But Kaldor also offered a theoretical explanation for the differing development paths of the various geographical subdivisions of the world economy, and a complementary account of the determinants and consequences of trade among these subdivisions. The spatial pattern of primary production and trade is more or less self-explanatory, being determined mainly by the unalterable location of natural resources. Kaldor thus focused mainly on industrial development, and especially on the reasons for its tremendous spatial unevenness – its long-term tendency to become concentrated in particular cities, regions, and countries.

Cumulative causation

The root cause of the unevenness of industrial development is increasing returns, which mean that success tends to breed further success, and that failure also tends to be self-perpetuating. For example, within a particular country or region, any locality that somehow becomes a substantial centre of industry will thereby achieve higher labour productivity than other smaller industrial centres, which, with fairly uniform wages, will mean lower unit labour costs. The firms in the larger locality will thus be able to charge lower prices or to spend more on marketing and product development, which will cause their sales to increase at the expense of their competitors in the smaller industrial centres. They will thus be able to expand production, further increasing their labour productivity and competitive advantage, and so on, with migration of workers from the declining smaller centres overcoming any labour shortages in the expanding larger centre.

Even within a single country, and even without deliberate government intervention, the eventual outcome will not necessarily be the concentration of all industry in a single place. This is because increasing returns, possibly in conjunction with diseconomies of urban agglomeration, may cause different large centres to specialize in different industrial products. There are also forces which retard the disequalizing process. One is the automatic redistributive effects of a unitary fiscal system. Another is that the smaller centres usually derive some offsetting benefits from expansion of the larger centres, including a bigger market for some of their products. But their growing competitive disadvantage is ultimately more important, and hence they tend to fall further and further behind.

Kaldor also emphasized a subtly different version of cumulative causation, based on the Verdoorn relationship between growth of industrial output and growth of industrial labour productivity, coupled with the assumption that the relative growth rates of exports from different localities depend on relative growth rates – rather than relative levels – of unit costs. This makes differences in industrial growth rates self-perpetuating. For, just as in a closed economy the long-term rate of industrial growth is governed by growth of demand from primary producers, so for a particular locality the necessary ‘external’ determinant of industrial growth is its growth rate of industrial exports. A locality whose industrial output happens to be growing relatively fast thereby has relatively rapid growth of labour productivity. In so far as wages are uniform across localities, it therefore has relatively slow growth of unit labour costs, and can thus achieve relatively rapid growth of exports. This causes the locality to

sustain relatively rapid output growth, and so on.

Kaldor regarded cumulative causation as important also in explaining the differing industrial development paths of different countries. Rapid industrial output growth, rapid labour productivity growth, and rapid export growth constitute a virtuous circle for some countries, with a corresponding vicious circle of low growth for others. But the underlying mechanism cannot be exactly the same as for different localities within a single country, because restrictions on international migration mean that wages are not uniform across countries. Moreover, empirical observation confirms what Kaldor's own theory of distribution implies, namely, that variation in the level and growth of real wages across countries is closely related to variation in the level and growth of average labour productivity. This clearly reduces the competitive advantage in international markets that countries with higher or faster-growing productivity would otherwise possess.

Indeed, Kaldor himself for many years argued that real wage rate changes brought about by regular exchange rate adjustments (with dual exchange rates for developing countries) could prevent cumulative causation at the international level. Experience with floating exchange rates after 1971, however, led him to the conclusion that neither exchange rate adjustments nor linkages between productivity and real wages are in practice sufficient to neutralize cumulative causation. There are various possible reasons for this. One is that the demand for industrial products is much more sensitive to quality than to price: increasing returns may thus involve feedback from faster output growth to faster product quality improvement, which increases international competitiveness and hence makes for faster export growth, faster output growth and so on.

Development and trade policy

Cumulative causation explains why places which acquire an initial advantage in industrial production tend to consolidate and increase this advantage at the expense of other places. Yet why do some places, rather than others, get ahead initially? Kaldor argued that this cannot usually be explained by the location of natural resources, nor by endowments of industrial capital or skills (which are for the most part generated by industrial growth itself). More important are transport facilities and general education, as well as social and institutional circumstances, which affect the willingness of individuals (or government organizations) to become entrepreneurs, and their ability to obtain bridging finance and to recruit a factory labour force. But what starts the process going is some stimulus to local industrial production, typically provided by deliberate protection against (or extraneous interruption of) imports. Kaldor therefore advocated protection or subsidies (of one sort or another) for infant industries, which can enable backward places to get the virtuous circle of industrialization started. He also argued that protection may be needed to prevent decline in an industrialized place (such as the UK) that has somehow begun to slip behind. At the same time, he emphasized that high or indiscriminate protection may be harmful. This is partly because it causes production to be spread thinly over many industries, none of which benefits sufficiently from increasing returns. It is also because protection discourages industrial exports, whose growth is in Kaldor's view essential – to pay for the increasing imports required by industrial expansion, to realize scale economies in particular industries, and to provide a dynamic external source of demand to propel the whole industrial sector.

In this and other senses, Kaldor's emphasis on increasing returns caused him to have very mixed feelings about trade. On the one hand, trade can have destructive or disequalizing effects. On the other hand,

trade is essential to the expansion of markets needed for the realization of increasing returns and hence for economic growth. The world as a whole therefore benefits from faster growth of trade, and most advanced (and some developing) countries from increased industrial specialization and exchange. The problem for policy – which Kaldor himself regarded as largely unsolved – is how to secure these collective benefits without aggravating the difficulties of industrially weak or backward places.

Markets, prices and wages

Kaldor's macroeconomic view of growth both stems from and contributes to his microeconomic view of how markets function. For primary products, his account of price formation is basically conventional – demand and supply under perfect competition – though modified to allow for the important influence of changes in stocks held by dealers and speculators.

Imperfect competition

Industrial and service enterprises, by contrast, are generally neither atomistic, passive price-takers nor static monopolies preoccupied with marginal adjustments in the face of given demand and technology. Instead the main task of most enterprises is to seek and develop both markets and technological opportunities, under competitive pressure from other enterprises. Enterprises generally do not have a well-defined optimum size, and they usually operate with excess capacity (either involuntarily or in order to be able to take advantage of unexpected sales opportunities), so that their actual size at any moment is determined by the demand for their products. Over time, by attracting more demand, many enterprises grow. Indeed, with increasing returns, industrial enterprises must grow in order to remain competitive.

Kaldor emphasized the force of competition in most markets for industrial goods and services, but also that it is imperfect, since firms have some discretion in setting their prices, and since non-price competition is important. He distinguished two main types of imperfect competition. One is polypoly, in which numerous firms – which can freely enter or leave the market – supply more or less imperfectly substitutable products. The number of competitors is too great for each firm to take direct account of the possible reactions of other firms in setting its price, which it accordingly does with a markup based on its perception of the elasticity of demand for its own product. But although prices are thus set above marginal costs, free entry ensures that the rate of profit on capital is not more than normal, by obliging firms to operate with excess capacity – underutilization of indivisible overhead inputs. (In services such as commerce, where polypoly often prevails, labour is usually treated as an overhead cost, which explains why these sectors act as reservoirs of underutilized labour, even in advanced countries.)

The other form of imperfect competition is oligopoly, where a smaller number of sellers (often the result of increasing returns) must consider each other's reactions in price setting and similar decisions. Kaldor regarded price leadership as the most common type of behaviour in this situation, with most firms tailoring their own market strategies to that adopted by one of the largest and most efficient firms. All firms tend to set their prices on the basis of markups over costs at normal capacity use, and not to vary their prices in the face of temporary fluctuations in demand. Nonetheless, the need to remain competitive with the leader greatly influences the size of the markups chosen by individual firms, and in particular means that less efficient firms must accept smaller margins of profit.

It is clearly important in this model of oligopoly to explain how the leading firm sets its own profit markup. Kaldor argued that this involves striking a balance between two opposing pressures, both arising from a desire to make the firm expand as fast as possible (which in Kaldor's view serves the interests not just of managers but also – in an uncertain world of increasing returns – of shareholders). One is the need to compete demand away from other firms, by lower prices or higher marketing and product development expenditure, which requires a low profit margin. The other is the need to finance investment in capacity expansion, which requires a high profit margin. This is because borrowing and new share issues are necessarily limited in relation to existing equity capital, and hence enterprises must rely heavily on retained profits as a source of finance.

Though Kaldor regarded this account of profit margin determination as relevant mainly to the leading firm in each market, it could apply more generally, especially when price is only one of a number of dimensions of market strategy. This is because each follower firm presumably also wants to grow, and also faces an inverse trade-off between its profit margin and expansion of its share of market demand, as well as needing to finance much of its capacity expansion internally – even though its maximum attainable growth rate will normally be less than that of the market leader. In any event, such an account of behaviour on the part of the average or representative firm (as in Wood, 1975) fits well with Kaldor's macroeconomic theory of profits, discussed earlier.

Real, relative and money wages

This theory of profits ties real wages to average productivity, except in low-income countries, where they may be governed by a subsistence minimum. Actually, Kaldor distinguished two such subsistence minima, one applicable to traditional agriculture, the other to industry, where higher food intake requirements (because of more intensive work) and other expenses of factory employment necessitate higher wages. Even at higher income levels, industrial real wages – especially in terms of primary commodities such as food, and especially where unions are powerful – tend to be inflexible downwards from whatever level they happen previously to have been raised to by productivity growth.

Although relative wage rates within the industrial sector tend to be rather rigid, and although labour market pressures may tend to equalize wages within agriculture and within services, wages are not normally uniform across these broad sectors. The initially large wage gap between agriculture and industry tends to diminish in the course of development, as more and more labour is sucked out of agriculture into other employment. But a similar wage gap between the industrial and service sectors tends to persist, fluctuating cyclically as workers released from industry in periods of recession crowd into service jobs, with a reverse flow in booms. Only perhaps at a very mature level of development would wages tend to equality across all broad sectors.

The average level of money wages, which is the main determinant of the absolute level of prices, has a life of its own, and tends to rise spontaneously in economies where collective wage bargaining is widespread, even in the face of substantial unemployment. Kaldor was inclined to believe that no simple general model can explain why money wages have risen at different rates in different places and at different times. But he identified some key elements, including efforts by unions in fast-growing industries to capture some of the profits created by productivity growth, imitative transmission of wage increases from one industry to another, and generalized resistance to real wage reductions. He has also consistently advocated the control of inflation through incomes policy (with restraint of dividends as

well as wages), despite the many practical problems involved.

Fiscal and monetary policy

Though well aware of the political obstacles to effective tax reform, Kaldor regarded taxation as the best available instrument for improving the distribution of income (as well as for altering the composition of production and expenditure). His extensive writings on tax policy covered many issues, theoretical and practical. But one consistent theme was equity as between different sources of income, and specifically the need to ensure that income from property bears a fair (but not penal) share of the tax burden. This motivated, among other things, his well-known pioneering proposal to change the basis of progressive personal taxation from income to expenditure, which better covers capital gains and other windfalls.

In macroeconomic management, Kaldor subscribed to the Keynesian view that effective demand is crucial and can be powerfully influenced through the budget. He also believed, however, that full employment and sustained growth in an open industrial economy cannot be secured simply through fiscal deficits, because these tend to be reflected in (ultimately unsustainable) foreign trade deficits. Instead, employment and growth objectives must be approached, for theoretical reasons discussed above, by operating on the foreign trade multiplier – especially on the rate of increase of exports, but also on the propensity to import. Yet this, Kaldor thought, is in practice not at all easy, since exchange rate adjustments are not very effective, even if coupled with an incomes policy to prevent offsetting money wage adjustments. Measures aimed at the basic determinants of international competitiveness, such as faster replacement of equipment and increased expenditure on training and product development, can make a significant contribution; but subsidies and protective tariffs or their equivalent may also be necessary.

Kaldor attached much less importance to monetary policy. The financial system is vital to modern capitalism, especially because it enables investment to be relatively independent of the current level of income. But the demand for money is not a stable function of income. Nor can the authorities effectively control the money supply, which in a credit money economy is largely endogeneously determined by the needs of enterprises and households. On these grounds, Kaldor always rejected the view that regulation of the money supply is an important ingredient of macroeconomic policy, even though interest rates can be directly manipulated to influence some components of domestic demand and international capital flows.

Conclusion

No second-hand account of Kaldor's economic theorizing can capture the force and vitality of the original, which greatly influenced many other economists, especially those fortunate enough to have been his students or colleagues in Cambridge. Nor can one adequately convey in a few pages the tremendous scope and depth of Kaldor's theoretical work. Finally, a survey like this is liable to give a misleadingly settled impression. Kaldor's thinking evolved constantly, and in its latest form – as at earlier stages – contained gaps, loose ends, inner tensions and unanswered empirical questions, which will provoke further progress. But it is an analytical framework of great range, power, and practical relevance, which constitutes a major contribution to our understanding of the way in which economies work.

Selected works

1955. *An Expenditure Tax*. London: Allen and Unwin.

1960–89. *Collected Economic Essays*. 9 vols, London: Duckworth.

1. *Essays on Value and Distribution*. 2nd edn, 1980.

2. *Essays on Economic Stability and Growth*, 2nd edn, 1980.

3. *Essays on Economic Policy I*.

4. *Essays on Economic Policy II*.

5. *Further Essays on Economic Theory*.

6. *Further Essays on Applied Economics*.

7. *Reports on Taxation I*.

8. *Reports on Taxation II*.

9. *Further Essays on Economic Theory and Policy*, ed. F. Targetti and A.P. Thirwall.

1982. *The Scourge of Monetarism*. Oxford: Oxford University Press.

1985. *Economics Without Equilibrium*. New York: M.E. Sharpe.

Bibliography

Nell, E. and Semmler, W., eds. 1991. *Nicholas Kaldor and Mainstream Economics: Confrontation or Convergence?* New York: St. Martin.

Pasinetti, L.L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29, 267–79.

Pasinetti, L.L. 1979. Kaldor, Nicholas. In *International Encyclopedia of the Social Sciences*, vol. 18 (Biographical Supplement). New York: Macmillan.

Targetti, F. 1992. *Nicholas Kaldor: The Economics and Politics of Capitalism as a Dynamic System*. Oxford: Oxford University Press.

Turner, M. 1993. *Nicholas Kaldor and the Real World*. New York: M.E. Sharpe.

Wood, A. 1975. *The Theory of Profits*. Cambridge: Cambridge University Press.

How to cite this article

Wood, Adrian. "Kaldor, Nicholas (1908–1986)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000003> doi:10.1057/9780230226203.0882

The New Palgrave Dictionary of Economics Online

Kalecki, Michal (1899 – 1970)

Malcolm Sawyer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article gives an outline of the life and work of Micha• Kalecki, in particular his contributions on macroeconomics in capitalist economics, including his discoveries of the role of effective demand, the significance of investment, the interplay between profits and investment and the degree of monopoly. His writings on socialism and on development are also outlined.

Keywords

agriculture and economic development; budget deficits; business cycles; capitalism; circuitist approach; distribution of income; economic development; economic growth; effective demand; endogenous money; full employment; intermediate regime; investment expenditure; Kalecki, M.; Keynes. J. M.; limit cycles; market power; market socialism; mixed differential-difference equation; monopoly; oligopoly; perfect competition; planning; political business cycles; Post Keynesian economics; principle of increasing risk; profit; rate of interest; Robinson, J. V.; socialism; state capitalism; unemployment; wage determination; workers' councils

Article

Michal Kalecki was born on 22 June 1899 in Lodz, Poland, and died in Warsaw on 17 April 1970. His academic training was in engineering, and he was self-taught in economics, influenced by writers such as Marx and Rosa Luxemburg. He obtained his first quasi-academic employment in 1929 at the Research Institute of Business Cycles and Prices in Warsaw, where his work involved the study of business cycles and the preparation of reports on specific industries. A Rockefeller Foundation Fellowship allowed him in 1936 to study abroad in Sweden and then England, where he remained for the next ten years; during the Second World War he was employed at the Oxford University Institute of Statistics. After work for the International Labour Office in Montreal, Canada, in 1945 and 1946, Kalecki was appointed at the end of 1946 as deputy director of a section of the economics department of

the United Nations secretariat in New York. In that job, a major task was the preparation of the *World Economic Reports*. When a board of directors was appointed to exercise control over that *Report*, which he and others viewed as McCarthyite American involvement in the work of the UN, he resigned in protest. Kalecki returned to Poland in 1955. He served as a consultant on economic planning with the government and then with the Planning Commission (1955 to 1964), and was heavily involved in the debates over the role of decentralization and of workers' councils, the speed of industrialization and the relative size of consumption and investment. He undertook research and teaching at the Polish Academy of Sciences between 1955 and 1961. The centre of his activities after 1961 was the Central School of Planning and Statistics.

In his analysis of capitalist economies, Kalecki discovered a range of ideas on the importance of effective demand and the role of investment similar to those discovered by Keynes, but Kalecki can claim priority of publication (Kalecki, 1933; Keynes, 1936). While there are similarities there are also differences, for example over the determinants of investment, the perception of the economy as competitive or oligopolistic (on the relationship between Kalecki and Keynes, see Sawyer, 1985, ch. 9).

A central element in Kalecki's work was the idea that the level of economic activity would be determined by the level of aggregate demand, and that investment decisions were a particularly significant element in the determination of the level of demand. Any decision to increase investment expenditure can come to fruition only if the finance is available, and the provision of additional finance comes through the banking system. Actual investment expenditure generates a corresponding amount of savings. Kalecki argued that savings were undertaken predominantly out of profits, and he often assumed as a first approximation that workers did not save, and hence investment expenditure in aggregate determined the volume of profits. As Kalecki wrote, 'capitalists as a class gain exactly as much as they invest or consume, and if – in a closed system – they ceased to construct and consume they could not make any money at all' (Kalecki, 1990–97, vol. 1, p. 79). If s_p is the propensity to save out of profits, and if there are no savings out of wages, then in a closed economy $s_p P=I$ where P is profits and I investment, with the direction of causation here running from investment to profits. The assumption that wages are spent and the view that capitalists' expenditure determines their income was reflected in the aphorism that was ascribed by Joan Robinson to Kalecki – 'the workers spend what they get, and capitalists get what they spend' (Robinson, 1966, p. 341) – though it cannot be found in the writings of Kalecki. There is also a reverse direction of causation at the level of the enterprise, whereby the profitability of the enterprise will influence its investment decisions. Profits provide internal finance for investment, and the present level of profits influences expectation on future profits.

Kalecki saw capitalism as oligopolistic and monopolistic and dismissed the notion of perfect competition as a 'dangerous myth'. His approach to pricing put forward the idea of the 'degree of monopoly' which expresses the notion that the market power which an enterprise possess will strongly influence the markup of its price over its (production) costs. The extent of market power depends on factors such as the dominance of the enterprise in its market, the barriers to entry into the industry and so forth. The degree of monopoly leads to a theory of the distribution of income and of the determination of real wages. At the level of the enterprise, the degree of monopoly sets the price–cost ratio; from this the ratio of profits to sales can be derived.

$$p = \lambda(w + m)R = \lambda(W + M)Y = \Pi + W\frac{\Pi}{Y} = \frac{(R - W - M)}{(W + \Pi)}$$

Then the share of profits is $\frac{(\lambda - 1)(1 + j)}{[1 + (\lambda - 1)(1 + j)]}$ and the share of wages $\frac{1}{[1 + (\lambda - 1)(1 + j)]}$, where *j=M/W*.

The real product wage can be calculated as $\frac{w}{p} = \frac{1}{\lambda} - \frac{m}{p}$.

Further derivation and then aggregation indicates that the share of profits in national income depends on the average degree of monopoly and on the cost of imports. Since wages are a major component of costs, the degree of monopoly impacts of the real product wage. Kalecki thus advanced a distinctive theory of the distribution of income between wages and profits, and the view that firms' pricing behaviour, rather than events in the labour market, set the real wage.

Kalecki's approach could be summarized by saying that the volume of profits depends on the level of investment, while the share of profits in national income depends on the degree of monopoly, that is, the market power possessed by firms.

The phenomenon of the business cycle was central to Kalecki's economic analysis of capitalism, and his discovery of the importance of aggregate demand for the level of economic activity was undertaken in the context of cyclical fluctuations. Kalecki viewed 'the determination of investment decisions by, broadly speaking, the level and the rate of change of economic activity' as the *pièce de résistance* of economics (Kalecki, 1968, p. 263). The central feature of Kalecki's explanation of the business cycle is the influence of investment on economic activity, and hence the determinants of investment. He distinguished between on the one hand the decision to invest and the placing of orders for investment, and, on the other hand, the actual investment taking place (for example, because it takes time to build the factory, there is a lag between investment orders and actual investment). Investment orders depend on profits, and profits are generated by actual investment (as noted above). He also postulated that investment is negatively influenced by the size of the capital stock. Combining these elements, Kalecki arrived at a mixed differential-difference equation (see, for example, Kalecki, 1990, vol. 1, pp. 82–3), for which there may be many solutions. Kalecki sought to establish that there is one solution for which the amplitude remains constant. 'This case is especially important because it corresponds roughly to the real course of the business cycle' (Kalecki, 1990, p. 90). He then argued that, with that condition satisfied, the other parameters of the model are such that a regular cycle of around ten years would be generated, which conforms with the general pattern of the time of a cycle of the order of eight to twelve years in length. The mixed differential-difference equation was the basis of Kalecki's attempt to generate a self-perpetuating cycle, which was later to be resolved through the notion of limit cycles.

The central feature of Kalecki's explanation of the business cycle is the influence of investment on economic activity, and hence the determinants of investment. Steindl (1981) identified three versions by Kalecki of the trade cycle, each with a different view of the determinants of investment, and he observed that there are differences in the ways through which profits influence investment and the impact of the size of the existing capital stock on investment (see also Sawyer, 1996).

Kalecki argued that 'the long-run trend is but a slowly changing component of a chain of short-period situations; it has no independent entity' (Kalecki, 1968, p. 263). This can be interpreted as undermining the predominant equilibrium approach to economic analysis whereby there is a long-period equilibrium

around which the economy fluctuates or towards which the economy tends and which is unaffected by the short-period movements of the economy.

The expansion of investment (and other forms of spending) has to be financed, and that comes predominantly through the creation of bank credit. In one of his earliest papers (1933), Kalecki acknowledged the link between the cycle and money creation. He asked:

> how can capitalists invest more than remains from their current profits after spending part of them for personal consumption? This is made possible by the banking system in various forms of credit inflation. Hence … without credit inflation there would be no fluctuations in investment activity. *Business fluctuations are strictly connected with credit inflation.*… A similar type of inflation is the financing of investments from bank deposits, a process usually not classified as inflation but one which perhaps has the greatest importance in the inflationary financing of investments during an upswing in the business cycle. (1990, vol. 1, pp. 148 and 149; emphasis in original)

Kalecki presented a number of ideas which now appear in the structuralist Post Keynesian analysis of endogenous money and in the circuitist approach, and he developed a substantial analysis of the workings of the monetary system (see Sawyer, 2001). Kalecki viewed the rate of interest as essentially a monetary phenomenon, and specifically not as a mechanism for bringing about the equality between savings and investment. He wrote that ‘the rate of interest cannot be determined by the demand for and supply of capital because investment automatically brings into existence an equal amount of savings. Thus, investment “finances itself” whatever the level of the rate of interest. The rate of interest is, therefore, the result of the interplay of other factors’ (Kalecki, 1997, vol. 7, p. 262).

The cost of borrowing is influenced by the ‘principle of increasing risk’ (Kalecki, 1937). Simply put, this principle is the idea that the greater the volume of borrowing a company wishes to undertake, relative to its own size and profits, the greater is the risk that the company will be unable to repay the borrowing. Any investment venture is subject to risk and uncertainty and to the vagaries of the business cycle. There is then some chance that a business will not be able to meet its loan commitments when its profits turn down. The lender would charge a risk premium on the loan, which makes the loan more expensive and increases the chances that the loan repayments cannot be met. The ‘principle of increasing risk’ then forms an upper limit on a business's ability to borrow and then to expand and grow.

The discoveries of Keynes and Kalecki in the 1930s on the principle of effective demand and the associated idea that governments could (and should) manipulate their budget stance to generate high levels of employment (rather than aim for a balanced budget) appeared to open up the way for the achievement of permanent full employment in capitalist economies. Kalecki (1943) raised many doubts on the possibilities of achieving prolonged full employment in a laissez-faire capitalist economy. Kalecki introduced an idea which was later interpreted in terms of the political business cycle. Economic activity and employment could be stimulated prior to elections to aid the chances of the governing party being re-elected. But the resulting high level of employment would not last, and at best full employment would be achieved only at the top of the cycle. There were a number of routes through which effective demand could be stimulated. Kalecki argued that the promotion of investment expenditure would be subject to important limits, namely that as investment rose, there would be a tendency for the output to capital ratio to fall (as investment adds to the capital stock) and for the rate of profit to fall. Instead,

Kalecki favoured a redistribution of income towards the working class which would stimulate spending, and the acceptance, if necessary, of a budget deficit by the government.

Socialism

Almost all Kalecki's writings on the economics of socialism were undertaken after his return to Poland in December 1954. While much of his writing was of a theoretical nature, the questions tackled and the approach adopted were much influenced by his perceptions of the Polish situation. He was directly involved in many of the debates of the mid-1950s on the development and organization of the Polish economy. His general approach can be summarized by saying that, while he sought a departure from the system of bureaucratic centralism, he thought that the main parameters of development in an economy should be centrally planned, with the market mechanism used in a subordinate role. He advocated a substantial increase in self-management by workers under a system of workers' councils, though he acknowledged that there would be tensions between them and central planning.

Soviet economic planning from the 1920s onwards and eastern European planning in the post-war period placed great weight on rapid industrialization and a heavy industry investment programme. The tendency towards overambitious plans often led to the sacrifice of consumption in favour of investment, when the overall plan could not be implemented but investment was safeguarded. Kalecki's criticisms of heavy industrialization and the sacrifice of consumption to investment brought him into conflict with the prevailing orthodoxy in Poland at the theoretical and at the practical levels.

Kalecki's approach to growth under socialism can be illustrated by reference to the basic relationship in which the growth of output is equal to the impact on productive potential of new investment minus the loss of the production through depreciation plus the change in the utilization of productive capacity. Much of Kalecki's theoretical work stemmed from this equation for the growth of output, with modifications for foreign trade, limited labour supply and technical progress. The emphasis was on the identification, and then pushing back, of the effective constraints on economic growth.

Kalecki viewed the market as involving the inefficient allocation of resources and the cause of insufficient aggregate demand. The socialist system was seen in terms of its ability to solve the problem of effective demand and to involve price–wage flexibility. Although he was critical of the decisions made under central planning, he was opposed to the market socialist alternative.

Development

Kalecki was heavily involved with teaching and research in the area of development planning from the late 1950s to the late 1960s. It is convenient to summarize his writings on development in terms of four themes. The first is that unemployment is seen to arise from a shortage of capital equipment, rather than from a deficiency of effective demand as in industrialized capitalist economies, so that constraints on employment and the pace of development arise more from the supply side than from the demand side. This led Kalecki to an identification of the binding constraints in any concrete situation: difficulties of expanding agricultural production, problems of achieving the desired rate of investment, and shortages of foreign exchange. These essentially economic constraints were generally compounded by the political resistance of powerful groups whose interests would be harmed by economic development.

The second theme is the need for the expansion of agricultural production as a part of the development

process, since development and increased incomes leads to an increased demand for food. If that increased demand for food is not satisfied, then the price of food is likely to rise, thereby reducing real wages. But the agricultural sector is likely to suffer from low productivity and outdated techniques. Since there are often powerful obstacles to the development of agriculture, such as feudal or semi-feudal relations in land tenure and the domination of the peasants by merchants and moneylenders, substantial institutional changes would be required to sustain agricultural and economic development.

The third theme is that market mechanisms, left to themselves, are unlikely to produce outcomes that Kalecki would have regarded as acceptable or desirable. He saw a strong need for planning and direct government intervention, particularly in investment and foreign trade.

The fourth theme is the distributional aspects of growth and development, and in particular a concern that the process of development should benefit the poor. This was combined with an awareness that prospective distributional consequences may block development.

In his work on developing countries, Kalecki developed the concept of an ‘intermediate regime’. Countries with intermediate regimes had generally achieved political independence after the Second World War and could not be considered as either socialist or laissez-faire capitalist economies though they sought economic development with government involvement. Kalecki argued that the governments of these intermediate regimes represented the interests of the lower-middle class, rich peasants and managers in the state sector. The poorest strata of society were still unorganized and lacked any political power. He further argued that in order to keep power these representatives of the middle class would have to achieve political and economic emancipation, carry out land reform and assure continuous economic growth. State capitalism develops at the expense of socialism in the economies of intermediate regimes because it helps the middle class to retain power by, for example, aiding faster growth and economic emancipation.

See Also

- Post Keynesian economics

Selected works

1933. *Proba teorii koniunktury*. Warsaw: Institute of Research on Business Cycles and Prices. Repr. as *Essays on the Business Cycle Theory*, in *Collected Works*, vol. 1, 1990.

1937. Principle of increasing risk. *Economica* 4, 440–7.

1943. Political aspects of full employment. *Political Quarterly* 14, 322–31.

1968. Trend and the business cycle, *Economic Journal* 78, 263–76.

1990–97. *Collected Works of Michal Kalecki*, 7 vols, ed. J. Osiatynski. Oxford: Clarendon.

Bibliography

Keynes, J.M. 1936. *General Theory of Employment, Money and Interest*. London: Macmillan.

Robinson, J. 1966. *Kalecki and Keynes in Economic Dynamics and Planning: Essays in Honour of Michal Kalecki*. Oxford: Pergamon.

Sawyer, M. 1985. *The Economics of Michal Kalecki*. Basingstoke: Macmillan.

Sawyer, M. 1996. Kalecki on the trade cycle and economic growth. In *An Alternative Macroeconomic Theory: the Kaleckian Model and Post Keynesian Economics*, ed. J. King. New York: Kluwer.

Sawyer, M. 1999. *The Legacy of Michal Kalecki*, 2 vols. Aldershot: Edward Elgar.

Sawyer, M. 2001. Kalecki on money and finance. *European Journal of the History of Economic Though* 8, 487–508.

Steindl, J. 1981. Some comments on the three versions of Kalecki's theory of the trade cycle. In *Studies in Economic Theory and Practice. Essays in Honour of Edward Lipinski*, ed. J. Los et al. Amsterdam: North-Holland. Repr. in J. Steindl, *Economic papers 1941–88*. London: Macmillan, 1990.

How to cite this article

Sawyer, Malcolm. "Kalecki, Michal (1899–1970)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000004> doi:10.1057/9780230226203.0883

The New Palgrave Dictionary of Economics Online

Kalman and particle filtering

Jesús Fernández-Villaverde
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Kalman and Particle filters are algorithms that recursively update an estimate of the state and find the innovations driving a stochastic process given a sequence of observations. The Kalman filter accomplishes this goal by linear projections, while the Particle filter does so by a sequential Monte Carlo method. With the state estimates, we can forecast and smooth the stochastic process. With the innovations, we can estimate the parameters of the model. The article discusses how to set a dynamic model in a state-space form, derives the Kalman and Particle filters, and explains how to use them for estimation.

Keywords

dynamic stochastic general equilibrium models; extended Kalman filter; Gaussian sum approximations; Kalman filter; Kalman gain; law of large numbers; maximum likelihood; Monte Carlo methods; Particle filter; sequential sampling; state space models; statistical inference

Article

The Kalman and Particle filters are algorithms that recursively update an estimate of the state and find the innovations driving a stochastic process given a sequence of observations. The Kalman filter accomplishes this goal by linear projections, while the Particle filter does so by a sequential Monte Carlo method.
Since both filters start with a state-space representation of the stochastic processes of interest, Section 1 presents the state-space form of a dynamic model. Section 2 introduces the Kalman filter and Section 3 develops the Particle filter. For extended expositions of this material, see Doucet, de Freitas, and Gordon (2001), Durbin and Koopman (2001), and Ljungqvist and Sargent (2004).

1 The state-space representation of a dynamic model

A large class of dynamic models can be represented by a state-space form:

$$X_{t+1} = \Phi(X_t, W_{t+1}; \Upsilon) \tag{1}$$

$$Y_t = g(X_t, V_t; \Upsilon). \tag{2}$$

This representation handles a stochastic process by finding three objects: a vector that describes the position of the system (a *state*, $X_t \in X \subset R^l$) and two functions, one mapping the state today into the state tomorrow (*the transition equation*, (1)) and one mapping the state into observables, Y_t (*the measurement equation*, (2)). An iterative application of the two functions on an initial state X_0 generates a fully specified stochastic process. The variables W_{t+1} and V_t are independent $i.i.d.$ shocks. A realization of T periods of observables is denoted by $Y^T \equiv \{y_t\}_{t=1}^T$ with $Y^0 = \{\varnothing\}$. Finally, γ, which belongs to the set $\Upsilon \subset R^n$, is a vector of parameters. To avoid stochastic singularity, we assume that $\dim(W_t) + \dim(V_t) \geq \dim(Y_t)$ for all t.

This framework can accommodate cases in which the dimensionality of the shocks is zero, where the shocks have involved structures, or where some or all of the states are observed. Also, at the cost of heavier notation, we could deal with more general problems. For example, the state could be a function or a correspondence, and the transition equation a functional operator. The basic ideas are, however, identical.

The transition and measurement equations may come from a statistical description of the process or from the equilibrium dynamics of an economic model. For example, dynamic stochastic general equilibrium models can be easily written in state-space form with the transition and measurement equations formed by the policy functions that characterize the optimal behaviour of the agents of the model. This observation tightly links modern dynamic macroeconomics with the filtering tools presented in this article.

It is important to note that there are alternative timing conventions for the state-space representation of a dynamic model and that, even while the timing convention is kept constant, the same model can be written in different state-space forms. All of those representations are equivalent, and the researcher should select the form that best fits her needs.

2 The Kalman filter

The Kalman filter deals with state-space representations where the transition and measurement equations are linear and where the shocks to the system are Gaussian. The procedure was developed by Kalman (1960) to transform ('filter') some original observables y_t into Wold innovations a_t and estimates of the state x_t. With the innovations, we can build the likelihood function of the dynamic model. With the estimates of the states, we can forecast and smooth the stochastic process.

We begin with the state-space system defined by the transition equation:

$$x_{t+1} = Ax_t + G\omega_{t+1},\ \omega_{t+1} \sim \mathcal{N}(0, Q)$$

and the measurement equation:

$$y_t = Cx_t + v_t,\ v_t \sim \mathcal{N}(0, R)$$

where A, G, C, Q, and R are known matrices.

There are different ways to derive and interpret the Kalman filter, including an explicitly Bayesian one. We follow a simple approach based on linear least-square projections. The reader will enhance her understanding with the more general expositions in Durbin and Koopman (2001) and Ljungqvist and Sargent (2004).

Let $x_{t|t-1} = E(x_t|Y^{t-1})$ be the best linear predictor of x_t given the history of observables until $t-1$, i.e., Y^{t-1}. Let $y_{t|t-1} = E(y_t|Y^{t-1}) = Cx_{t|t-1}$ be the best linear predictor of y_t given Y^{t-1}. Let $x_{t|t} = E(x_t|Y^t)$ be the best linear predictor of x_t given the history of observables until t, i.e., y^t. Let $\Sigma_{t|t-1} \equiv E((x_t - x_{t|t-1})(x_t - x_{t|t-1})'|Y^{t-1})$ be the predicting error variance-covariance matrix of x_t given Y^{t-1}. Finally, let $\Sigma_{t|t} \equiv E((x_t - x_{t|t})(x_t - x_{t|t})'|Y^t)$ be the predicting error variance-covariance matrix of x_t given y^t.

How does the Kalman filter work? Let's assume we have $x_{t|t-1}$ and $y_{t|t-1}$, that is, an estimate of the state and a forecast of the observable given Y^{t-1}. Then, we observe y_t. Thus, we want to revise our linear predictor of the state and obtain an estimate, $x_{t|t}$, that incorporates the new information. Note that $x_{t+1|t} = Ax_{t|t}$ and $y_{t+1|t} = Cx_{t+1|t}$, so we can go back to the first step and wait for the

y_{t+1} next period. Therefore, the key of the Kalman filter is to obtain $x_{t|t}$ from $x_{t|t-1}$ and y_t.
We do so with the formula:

$$x_{t|t} = x_{t|t-1} + K_t(y_t - y_{t|t-1}) = x_{t|t-1} + K_t(y_t - Cx_{t|t-1}),$$

that is, our new value $x_{t|t}$ is equal to $x_{t|t-1}$ plus the difference between the actual y_t and the forecasted $y_{t|t-1}$, times a matrix K_t, called the Kalman gain. Durbin and Koopman (2001) derive this formula from probabilistic foundations. Ljungqvist and Sargent (2004) find it through an application of a Gram–Schmidt orthogonalization procedure.
Then, if we choose K_t to minimize $\Sigma_{t|t}$, we get $K_t = \Sigma_{t|t-1}C'(C\Sigma_{t|t-1}C' + R)^{-1}$. This expression shows the determinants of K_t. If we made a big mistake forecasting $x_{t|t-1}$ using past information ($\Sigma_{t|t-1}$ large), we give a lot of weight to the new information (K_t large). Also, if the new information is noisy (R large), we give a lot of weight to the old prediction (K_t small).
Now, note that $\Sigma_{t|t} \equiv E((x_t - x_{t|t})(x_t - x_{t|t})'|y^t) = \Sigma_{t|t-1} - K_tC\Sigma_{t|t-1}$. Therefore, from $x_{t|t-1}$, $\Sigma_{t|t-1}$, and y_t, we compute $x_{t|t}$ and $\Sigma_{t|t}$ using K_t. Also, we derive $\Sigma_{t+1|t} = A\Sigma_{t|t}A' + GQG'$, $x_{t+1|t} = Ax_{t|t}$, and $y_{t+1|t} = Cx_{t+1|t}$.
We collect all the previous steps. We start with some estimates of the state $x_{t|t-1}$, the observables $y_{t|t-1}$, and the variance-covariance matrix $\Sigma_{t|t-1}$. Then we observe y_t and compute $x_{t+1|t}$, $y_{t+1|t}$, and $\Sigma_{t+1|t}$.
Thus, the Kalman filter can be recursively written as follows:

- $y_{t|t-1} = Cx_{t|t-1}$
- $K_t = \Sigma_{t|t-1}C'\left(C\Sigma_{t|t-1}C' + R\right)^{-1}$
- $\Sigma_{t|t} = \Sigma_{t|t-1} - K_tC\Sigma_{t|t-1}$
- $x_{t|t} = x_{t|t-1} + K_t(y_t - Cx_{t|t-1})$
- $\Sigma_{t+1|t} = A\Sigma_{t|t}A' + GQG'$
- $x_{t+1|t} = Ax_{t|t}$.

The differences between the observable and its forecast, $a_t = y_t - y_{t|t-1} = y_t - Cx_{t|t-1}$ are, by construction, Wold innovations. Moreover, since the system is linear and Gaussian, a_t is normally distributed with zero mean and variance $C\Sigma_{t|t-1}C' + R$. That is why the Kalman filter is a whitening filter: it takes as an input a correlated sequence y^T and it produces a sequence of white noise innovations a_t.
With this last result, we write the likelihood function of $y^T = \left\{y_t\right\}_{t=1}^T$ as:

$$\log L\left(y^T|A, G, C, Q, R\right) = \sum_{t=1}^T \log L\left(y_t|y^{t-1}A, G, C, Q, R\right) = -\sum_{t=1}^T\left[\frac{N}{2}\log 2\pi + \frac{1}{2}\log\left|C\Sigma_{t|t-1}C' + R\right| + \frac{1}{2}\sum_{t=1}^T a_t'\left(C\Sigma_{t|t-1}C' + R\right)^{-1}a_t\right].$$

This likelihood is one of the most important results of the Kalman filter. With it, we can undertake statistical inference in the dynamic model, both with maximum likelihood and with Bayesian approaches.
An important step in the Kalman filter is to set the initial conditions $x_{1|0}$ and $\Sigma_{1|0}$. If we consider stationary stochastic processes, the standard approach is to set $x_{1|0} = x^*$ and $\Sigma_{1|0} = \Sigma^*$ such that $x^* = Ax^*$ and

$$\Sigma^* = A\Sigma^*A' + GQG' = [I - A\otimes A]^{-1}vec(GQG').$$

Non-stationary time series require non-informative prior conditions for $x_{1|0}$. This approach, called the diffuse initialization of the filter, begins by postulating that $x_{1|0}$ is equal to:

$$x_{1|0} = \tau + \Phi\delta + G\omega_0,\ \omega_0 \sim \mathcal{N}(0, Q) \text{ and } \delta \sim \mathcal{N}(0, \kappa I_q)$$

where τ is given and Φ and G are formed by columns of the identity matrix such that $\Phi G' = 0$. This structure allows for some elements of $x_{1|0}$ to have a known joint distribution, while, by letting $\kappa \to \infty$, to formalize ignorance with respect to other elements. Clearly, $x_{1|0} = E(x_{1|0}) = \tau$. To determine the initial variance, we expand $\Sigma_{1|0} = \kappa\Phi\Phi' + GQG'$ as a power series of κ^{-1} and take $\kappa \to \infty$ to find the dominant term of the expansion. Durbin and Koopman (2001) provide details.

The Kalman filter can also be applied for smoothing, that is, to obtain $x_{t|T}$, an estimate of x_t given the whole history of observables, that is, y^T. Smoothing is of interest when the state x_t has a structural interpretation of its own. Since smoothing uses more information than filtering, the predicting error variance covariance matrix of x_t given y^T will be smaller than $\Sigma_{t|t-1}$. Finally, we note that the Kalman filtering problem is the dual of the optimal linear regulator problem.

3 The Particle filter

The Kalman filter relies on the linearity and normality assumptions. However, many models in which economists are interested are nonlinear and/or non-Gaussian. How can we undertake the forecast, smoothing, and estimation of dynamic models when any of those two assumptions are relaxed?

Sequential Monte Carlo methods, in particular the Particle filter, reproduce the work of the Kalman filter in those nonlinear and/or non-Gaussian environments. The key difference is that, instead of deriving analytic equations as the Kalman filter does, the Particle filter uses simulation methods to generate estimates of the state and the innovations. If we apply the Particle filter to a linear and Gaussian model, we will obtain the same likelihood (as the number of simulations grows) that we would if we used the Kalman filter. Since it avoids simulations, the Kalman filter is more efficient in this linear and Gaussian case.

We present here only the basic Particle filter. Doucet, de Freitas and Gordon (2001) discuss improvements upon the basic filter. Fernández-Villaverde and Rubio-Ramírez (2007) show how this Particle filter can be implemented to estimate dynamic stochastic general equilibrium models. Our goal is to evaluate the likelihood function of a sequence of realizations of the observable y^T implied by a stochastic process at a parameter value γ:

$$L\left(y^T; \gamma\right) = p\left(y^T; \gamma\right). \tag{3}$$

Our first step is to factor the likelihood function as:

$$p\left(y^T; \gamma\right) = \prod_{t=1}^{T} p\left(y_t | y^{t-1}; \gamma\right) = \prod_{t=1}^{T} \iint p\left(y_t | W^t, X_0, y^{t-1}; \gamma\right) \times p\left(W^t, X_0 | y^{t-1}; \gamma\right) dW^t dX_0, \tag{4}$$

where X_0 is the initial state of the model and the p's represent the relevant densities. In general, the likelihood function ((4)) cannot be computed analytically. The particle filter uses simulation methods to estimate it.

Before introducing the filter, we assume that, for all γ, x_0, w^t, and t, the following system of equations:

$$X_1 = \phi(x_0, w_1; \gamma)$$
$$y_m = g(X_m, V_m; \gamma) \text{for } m = 1, 2, \ldots t$$
$$X_m = \phi(X_{m-1}, w_m; \gamma) \text{for } m = 2, 3, \ldots t$$

has a unique solution, (v^t, x^t), and we can evaluate $p(v^t; \gamma)$. This assumption implies that we can evaluate the conditional densities $p(y_t | w^t, x_0, y^{t-1}; \gamma)$ for all γ, x_0, w^t, and t. Then, we have:

$$p\left(y_t | w^t, x_0, y^{t-1}; \gamma\right) = |dy(v_t; \gamma)| p(v_t; \gamma)$$

for all γ, x_0, w^t, and t, where $|dy(v_t; \gamma)|$ stands for the determinant of the Jacobian of y_t with respect to V_t evaluated at v_t.

Conditional on having N draws of $\left\{\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N\right\}_{t=1}^T$ from the sequence of densities $\left\{p(W^t, X_0 | y^{t-1}; \gamma)\right\}_{t=1}^T$, the law of large numbers implies that the likelihood function ((4)) can be approximated by:

$$p\left(y^T; \gamma\right) \simeq \prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N p\left(y_t | w^{\#t-1,i}, x_0^{\#t-1,i}, y^{t-1}; \gamma\right).$$

This observation shows that the problem of evaluating the likelihood ((4)) is equivalent to the problem of drawing from $\left\{p(W^t, X_0 | y^{t-1}; \gamma)\right\}_{t=1}^T$. Since the algorithm does not require any assumption about the distribution of the shocks except the ability to evaluate $p(y_t | w^t, x_0, y^{t-1}; \gamma)$, either analytically or by simulation, we can deal with models with a rich specification of non-Gaussian innovations. But, how do we sample from $\left\{p(W^t, X_0 | y^{t-1}; \gamma)\right\}_{t=1}^T$?

Let $\left\{x_0^{t-1,i}, w^{t-1,i}\right\}_{i=1}^N$ be a sequence of N $i.i.d.$ draws from $p(W^{t-1}, X_0 | y^{t-1}; \gamma)$. Let $\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N$ be a sequence of N $i.i.d.$ draws from $p(W^t, X_0 | y^{t-1}; \gamma)$. We call each draw $(x_0^{t,i}, w^{t,i})$ a *particle* and the sequence $\left\{x_0^{t,i}, w^{t,i}\right\}_{i=1}^N$ a *swarm of particles*. Also, define the weights:

$$q_t^i = \frac{p\left(y_t | w^{\#t-1,i}, x_0^{\#t-1,i}, y^{t-1}; \gamma\right)}{\Sigma_{i=1}^N p\left(y_t | w^{\#t-1,i}, x_0^{\#t-1,i}, y^{t-1}; \gamma\right)}.$$
(5)

The next proposition shows how to use $p(W^t, X_0 | y^{t-1}; \gamma)$, the weights $\left\{q_t^i\right\}_{i=1}^N$, and importance sampling to draw from $p(W^t, X_0 | y^t; \gamma)$:

Proposition 1: Let $\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N$ *be a draw from* $p(W^t, X_0|Y^{t-1}; \Upsilon)$. *Let the sequence* $\left\{\tilde{x}_0^i, \tilde{w}^i\right\}_{i=1}^N$ *be a draw with replacement from* $\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N$ *where* q_t^i *is the probability of* $(x_0^{\#t-1,i}, w^{\#t-1,i})$ *being drawn* $\forall i$. *Then* $\left\{\tilde{x}_0^i, \tilde{w}^i\right\}_{i=1}^N$ *is a draw from* $p(W^t, X_0|Y^t; \Upsilon)$.

Then, with a draw $\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N$ from $p(W^t, X_0|Y^{t-1}; \Upsilon)$, we get a draw $\left\{x_0^{t,i}, w^{t,i}\right\}_{i=1}^N$ from $p(W^t, X_0|Y^t; \Upsilon)$ and we generate a sequence of particles $\left\{\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N\right\}_{t=1}^T$ from the sequence $\left\{p(W^t, X_0|Y^{t-1}; \Upsilon)\right\}_{t=1}^T$. Given some initial conditions, we can recursively apply the idea of the previous proposition as summarized by the algorithm:

- **Step 0, Initialization:** Set $t \leftarrow 1$. Initialize $p(W^{t-1}, X_0|Y^{t-1}; \Upsilon) = p(X_0; \Upsilon)$.
- **Step 1, Prediction:** Sample *N* values $\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N$ from the conditional density $p(W_1^t, X_0|Y^{t-1}; \Upsilon) = p(W_t; \Upsilon)\, p(W^{t-1}, X_0|Y^{t-1}; \Upsilon)$.
- **Step 2, Filtering:** Assign to each draw $(x_0^{\#t-1,i}, w^{\#t-1,i})$ the weight q_t^i as defined in (5).
- **Step 3, Sampling:** Sample *N* times with replacement from $\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N$ with probabilities $\left\{q_t^i\right\}_{i=1}^N$. Call each draw $(x_0^{t,i}, w^{t,i})$. If $t < T$ set $t \leftarrow t + 1$ and go to Step 1. Otherwise stop.

With the algorithm's output $\left\{\left\{x_0^{\#t-1,i}, w_1^{\#t-1,i}\right\}_{i=1}^N\right\}_{t=1}^T$, we obtain the estimate of the states in each period and compute the likelihood:

$$p\left(Y^T; \Upsilon\right) \simeq \frac{1}{N}\left(\prod_{t=1}^T \frac{1}{N}\sum_{i=1}^N p\left(y_t|w^{\#t-1,i}, x_0^{\#t-1,i}, Y^{t-1}; \Upsilon\right)\right).$$

The **sampling step** is the heart of the algorithm. If we skip it and weight each draw in $\left\{x_0^{\#t-1,i}, w^{\#t-1,i}\right\}_{i=1}^N$ by $\left\{N q_t^i\right\}_{i=1}^N$, we have a sequential importance sampling. The problem with this approach is that it diverges as t grows. The reason is that, as $t \to \infty$, all the sequences become arbitrarily far away from the true sequence of states (the true sequence being a zero measure set), and the sequence that happens to be closer dominates all the remaining sequences in weight. In practice, after a few steps only one sequence has a non-zero weight. Through resampling, we eliminate this problem as we keep (and multiply) those sequences that do not diverge from the true one.

The algorithm outlined above is not the only procedure to evaluate the likelihood of nonlinear and/or non-Gaussian dynamic models. However, the alternatives, such as the extended Kalman filter, the Gaussian sum approximations, or grid-based filters, are of limited use, and many, such as the extended Kalman filter, fail asymptotically. Consequently, the Particle filter is the most efficient and robust procedure to undertake inference for nonlinear and/or non-Gaussian models, and we will witness many applications of this filter in economics in future years.

See Also

- Bayesian methods in macroeconometrics
- Markov chain Monte Carlo methods
- state space models

Bibliography

Doucet, A., de Freitas, N. and Gordon, N. 2001. *Sequential Monte Carlo Methods in Practice*. New York: Springer Verlag.

Durbin, J. and Koopman, S.J. 2001. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Fernández-Villaverde, J. and Rubio-Ramírez, J.F. 2007. Estimating macroeconomic models: a likelihood approach. *Review of Economic Studies* 74, 1059–87.

Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering* D82, 35–45.

Ljungqvist, L. and Sargent, T.J. 2004. *Recursive Macroeconomic Theory*, 2nd edn. Cambridge, MA: MIT Press.

How to cite this article

Fernández-Villaverde, Jesús. "Kalman and particle filtering." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000005>
doi:10.1057/9780230226203.0884

The New Palgrave Dictionary of Economics Online

Kantorovich, Leonid Vitalievich (1912 – 1986)

V. Makarov
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic dynamics; innovation; Kantorovich, L. V.; linear programming; mathematics; optimal planning; optimization approach to economics; socialism

Article

Kantorovich made valuable contributions to the theory of welfare and was the founder of the theory of optimal planning of socialist economics. As a professional mathematician, he also made a valuable contribution to a number of sections of modern mathematics. He is regarded (together with G. Dantzig) as the founder of linear programming, the mathematical discipline which has many applications in economics.

L.V. Kantorovich was born on 19 January 1912. He graduated from the department of mathematics of Leningrad University in 1930 at the age of 18. Four years later he became professor of mathematics at Leningrad University. In 1939, through the publishing house of Leningrad University, he published a small booklet, ‘Mathematical Methods of Organization and Planning of Production Process’.

This may be considered a historic document, containing the facts about discovery of the linear programming. The mathematical formulation of production problems of optimal planning was presented here for the first time and the effective methods of their solution and economic analysis were proposed. Thus the idea of optimality in economics was founded scientifically. This booklet and a number of subsequent articles establish Kantorovich together with F.P. Ramsey and J. von Neumann as the founders of the optimization approach to the analysis of economic problems.

His fundamental work, *The Best Uses of Economic Resources*, written in 1942 but published for the first time only in 1959, is a brilliant example of the consistent application of the optimization principle to the analysis of a wide variety of economic problems: the planning of production from the level of enterprise to the level of the national economy as a whole; a theory of price formation, which includes the principles of price formation not only for goods and services but also for the factors of production, the time factor, the space factor, natural conditions, the conditions of labour application, and so on; a theory of economic and social-economic efficiency of economic enterprises.

In fact, Kantorovich developed a powerful tool for the analysis of economic problems from the unified position of global optimum and indeed it is not necessary to find this optimum, it is enough to postulate its existence. In a number of his subsequent articles Kantorovich demonstrated the power of his method for the analysis and improvement of the mechanism of economic management of the socialist economy as a whole and its components. He proposed methods for calculating wholesale price levels for the branches of the national economy; the value of the norm of effectiveness of capital investments; the norm of depreciation allowances, and the value of transport tariffs, rent payments, and so on.

For a number of years Kantorovich showed great interest in the problems of economic dynamics. He proposed, analysed and used in practice a dynamic model of optimal planning. On the basis of this model and its different modifications Kantorovich proposed an original theory of economic evaluation of technical ventures. The essence of this theory is that the economic effect of the introduction of a scientific-technical innovation includes three components: a producer effect; a consumer effect; and an effect which is the result of the increase in general scientific-technical economic potential derived from the innovation. The third component is ignored in usual economic practice which leads to a distorted calculation of the real efficiency of innovations.

Kantorovich was also a world-famous mathematician. He made great contributions to a number of different branches of mathematics, among them the descriptive theory of functions and of sets; the constructive theory of functions; a decision method of solving a wide range of problems concerning the best approximation of functions by polynomials; calculus of variations; functional analysis, where he introduced and studied the class of semi-ordered spaces (K-spaces); approximate calculation methods; and developed several effective algorithms as well as a number of other branches of mathematics. This demonstrates his mathematical genius and the vast range of his interests and knowledge.

The author of about 300 scientific works, Kantorovich was awarded the Nobel Prize in economics in 1975.

Selected works

1939. *Matematicheskie metody organizatsii i planirovaniia proizvodstva*. Leningrad: State Publishing House. Trans. as ‘Mathematical methods of organizing and planning production’, *Management Science* 6 (1960), 363–422.

1959a. *Ekonomicheskii raschet nailuchshego isplo'zovaniia resursov*. Moscow: State Publishing House. Trans. P.F. Knightsfield, ed. G. Morton, as *The Best Uses of Economic Resources*, Oxford: Pergamon, 1965.

1959b. (With G.P. Akilov.) *Funktsional'nyi analiz v normirovannykh prostranstvakh*. Moscow: Nauka. 2nd edn, 1982. 1st edn trans. D.E. Brown, ed. Dr A.P. Robertson, as *Functional Analysis in Normed Spaces*, Oxford: Pergamon. 2nd edn trans. H. L. Silcock as *Functional Analysis*, Oxford: Pergamon, 1982.

1976. *Essays in Optimal Planning*, ed. L. Smolinski. New York: Wiley.

How to cite this article

Makarov, V. "Kantorovich, Leonid Vitalievich (1912–1986)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000007> doi:10.1057/9780230226203.0885

The New Palgrave Dictionary of Economics Online

Katona, George (1901 – 1981)

James N. Morgan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

consumer confidence; Cowles Commission; inflation; Katona, G.; psychological economics

Article

Katona developed the theory and substance of psychological economics, with particular attention to the effects of national events on the confidence, expectations, plans and ultimately behaviour of masses of individuals. From a background in Gestalt psychology, he noted that there can be major restructuring of the way people interpret their world and its future, leading to sometimes dramatic shifts in behaviour. And he had a firm belief in people's capacity to learn and to adjust their goals, so that behaviour was more than a simple response to stimuli. Like most great ideas, these were at the same time simple and profound, obvious (that attitudes would affect behaviour) but not accepted, particularly by economists who preferred to keep attitudes and expectations endogenous so they need not be measured or dealt with directly.

The theory argued that the importance of mass psychology was growing as consumers became more affluent, used credit and had to make long-term commitments to levels of investment in housing and cars, and to repayment schedules. Furthermore, he argued, the world was becoming increasingly volatile and unpredictable, so that it was necessary for people to interpret the chaos. The repeated measures of consumer confidence were useful for short-run prediction, but the long-run goal was, and is, to understand mass changes in consumer attitudes and behaviour.

Katona was born in 1901 in Budapest, and was a law student at the University of Budapest when a communist putsch under Bela Kun closed the University. Instead he studied psychology under Mueller at the University of Göttingen. His Ph.D. was on the psychology of perception. While at the University of Frankfurt, he wrote a prize-winning monograph on the psychology of comparison, with an empirical orientation. Hyperinflation drove him to work for a Frankfurt bank and he wrote a widely quoted article on the mass psychological aspects of inflation. There followed a period in Berlin studying Gestalt psychology with his friend Max Wertheimer and writing for Gustav Stolper's *Der Deutsche Volkswirt*. When Hitler closed down the paper, Katona emigrated to New York. An attack of tuberculosis ended his

career as an investment counsellor and he turned again to writing. His path-breaking book *Organizing and Memorizing* (1940), showed how organizing material in gestalts made it easier to remember and apply to other situations, and could lead to changes in expectations.

Concern with the economic effects of the war led him to write *War without Inflation: The Psychological Approach to the Problems of a War Economy* (1942). After tours at the Cowles Commission for Research in Economics at the University of Chicago, and the Division of Program Surveys of the US Department of Agriculture, he moved with that programme to the University of Michigan to help form the Survey Research Center. There he began the continuing series of surveys measuring mass changes in consumer confidence, building a body of knowledge about how people respond to events and interpret them for their own lives. A series of books starting in 1951 and continuing through 1978 summarized the research.

He died in West Germany on 18 June 1981, the day after receiving an honorary degree from the Free University of Berlin.

Selected works

1940. *Organizing and Memorizing: Studies in the Psychology of Learning and Teaching*. New York: Columbia University Press.

1942. *War without Inflation: The Psychological Approach to Problems of a War Economy*. New York: Columbia University Press.

1945. *Price Control and Business*. Washington, DC: Cowles Commission and Principia Press.

1951. *Psychological Analysis of Economic Behavior*. New York: McGraw-Hill.

1960. *The Powerful Consumer: Psychological Studies of the American Economy*. New York: McGraw-Hill.

1964. *The Mass Consumption Society*. New York: McGraw-Hill.

1965. *Private Pensions and Individual Savings*. Ann Arbor: Survey Research Center, University of Michigan.

1971. (With B. Strumpel and E. Zahn.) *Aspirations and Affluence: Comparative Studies in the United States and Western Europe*. New York: McGraw-Hill.

1975. *Psychological Economics*. New York: Elsevier.

1978. *A New Economic Era*. New York: Elsevier.

How to cite this article

Morgan, James N. "Katona, George (1901–1981)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000009> doi:10.1057/9780230226203.0886

The New Palgrave Dictionary of Economics Online

Kautsky, Karl (1854 – 1938)

Tadeusz Kowalik
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; cartels; economic planning; Engels, F.; guild socialism; imperialism; Kautsky, K.; market socialism; Marx, K. H.; Marx's analysis of capitalist production; money; peasant economy; Say's Law; socialism; stagnation; Tugan-Baranovsky, M. I.; underconsumptionism

Article

Kautsky was born in Prague on 16 October 1854 and died in Amsterdam on 17 October 1938. Marxist thinker and writer, leading theoretician of the German Social Democratic Party (SPD) and the Second International, he studied law and arts in Vienna. Fascinated by the theories of Marx and Engels (both of whom he met and befriended in London in 1881), Kautsky must be credited with the spread and development of their ideas in all his embodiments – as a prodigal and versatile columnist; as founder and editor (1883–1917) of the SPD theoretical journal *Die Neue Zeit*, which soon became the chief Marxist forum in Europe; as editor of Marx's books and unfinished manuscripts (Kautsky edited them in three volumes called *Theorien über den Mehrwert*, which appeared in 1905–10); and also as socialist thinker. Kautsky presented his ideas systematically in *Die materialistische Geschichtsauffassung* (1927), expounding a theory of social development which combined Marx's and Engels's historical materialism with Darwin's naturalism.

Kautsky's first major popular book designed to spread Marxian theories was *Karl Marx' ökonomische Lehren* (1887), which expounds the substance of the first volume of *Das Kapital*. It went into numerous editions in German and other languages, and in some countries (as in Russia) its effect on the spread of Marxism was significant.

His original contribution to Marxian theory was his *Agrarfrage* (1899), described by Lenin as the most outstanding work since the third volume of *Das Kapital* had appeared in print. In it, Kautsky analyses trends of development in agricultural production against the backdrop of Marx's theory of capitalism, of capitalist development's own specific features and, in particular, of the then much-discussed question of persistence of small peasant holdings. Kautsky studied the causes of small private farms' relative viability, a phenomenon which at that time was often cited as evidence that Marx's concentration theory

was wrong. He attributed the survival of small peasant holdings to the undernourishment and excessive toil of peasant families, to the demand for seasonal labour by large landed estates and to their interest in preserving local labour reserves. Kautsky also pointed out that, in agriculture, concentration of production does not necessarily go along with increases of crop area but may result from more intensive cultivation. Generally, though, he believed that the conquest of agriculture by capitalism was just a question of time.

Kautsky's motive for studying the agrarian question was pragmatic; he wanted to answer the question of whether or not the SPD needed an agricultural policy of its own. In particular, it was unclear whether the SPD ought to defend peasants on their own holdings against the adverse effects of capitalism. Kautsky came to believe that such a move would only hamper what was an inexorable social process, namely, the emergence of large capitalist farms relying on hired labour, and hence would hamper the ascent of socialism. Without compromising its own tenets and aspirations, Kautsky said, the SPD could demand the abolition of all vestiges of feudalism in the countryside and defend peasants as working people, as semi-proletarians. But he thought the idea of defending peasants as smallholders a reactionary utopia. He used the same logic to interpret the role of the capitalist metropolitan countries in subjugating colonies.

Kautsky wrote the *Agrarfrage*, as well as his studies concerning crises, as polemics against ‘revisionists’, who argued that the spread of cartels and trusts, along with the expansion of bank activities, eliminated the anarchy in capitalist production and hence was likely to allay or forestall crises in the future. Kautsky opposed these theories in a series of articles (1901–2) in *Die Neue Zeit* which he wrote in reaction to a German translation of Mikhail Tugan-Baranovsky's *Studien zur Theorie und Geschichte der Handelskrisen in England* (1901). Tugan-Baranovsky reinterpreted Marx's reproduction models in terms of Say's Law and attributed the causes of crises to the disproportions of capitalist development. The spread of cartels, Tugan-Baranovsky argued, eliminated those disproportions and hence also forestalled crises.

Kautsky defended the theory of underconsumption as the basis of business cycles and argued that cartels and other similar organizations of capitalists, keen as they were on maximizing profits, were unable to keep control of production and demand on a national scale, to say nothing of the world economy. He countered the optimistic picture presented by the ‘revisionists’ with his own hypothesis of capitalism's inexorable drift toward ‘a chronic depression’. That was one of the first-ever theories of stagnation. Later (1910), Kautsky was inclined to attribute the principal cause of ‘recent’ crises to the circumstance that agricultural growth was slower than and lagging behind industrial growth. He also cited this particular disproportion in his concept of imperialism as the expansion of advanced industrial countries into agrarian markets. During the First World War Kautsky formulated his well-known hypothesis portraying ultra-imperialism as an alliance of previously rival imperialist powers for a joint exploitation of world resources.

In many studies Kautsky returned to the political and economic problems of the transition to socialism and to the organization and operation of the socialist economy. At first, those problems were overshadowed by the dominant question of political revolution to seize power and of proletarian dictatorship, and Kautsky's casual remarks indicate he regarded a socialist economy simply as the negative of a market-dominated capitalist economy. But from the war onwards, especially in the 1920s, he interpreted socialism and the socialist economy as a continuation and further development of capitalist accomplishments not only in economics but also in terms of social advancement and political

progress. His writings are pervaded by a concern for freedom and democracy. Accordingly, he views the transition period as a long process of socialization of production during which those accomplishments would be preserved and economic efficiency would be maintained.

Kautsky was one of the first socialist writers to dispute the idea of a natural, that is money-free, socialist economy. Already at the turn of the century (1902) he argued that money and market were indispensable if freedom of choice in consumption and jobs was to be preserved. Two decades later, when the wave of revolution in Germany, but especially in the Soviet Union, made the construction of socialism a topical question, Kautsky considered the question in a systematic manner (1922). Apart from reaffirming the advantages of money and prices, Kautsky acknowledged the importance of money as a measure of value which permitted the quantitative assessment of production by means of accounting techniques and as a device for identifying benefits that may be gained from trade transactions. However, he failed to furnish a clear picture of how he interpreted economic choice in the allocation of resources. He was probably not quite consistent on this point. On the one hand, he wrote in the spirit of ‘market socialism’ that socialist society would be governed by the law of value. On the other hand, he overrated the benefits of economies of scale, that is, the supremacy of large-scale over small-scale production, and he was adamant in his faith in vertical and horizontal integration. If his beliefs came true, the integration was bound to lead to ubiquitous monopolistic practices on the part of socialist industrial giants.

He also believed that full socialization of production and of the bank credit system would render the latter superfluous. He accepted that interest rates might be charged by the socialized banks, but solely in order not to deprive them of their competitive edge in relations with capitalist banks and only in the transition period. His idea of economic planning also seems incompatible with ‘market socialism’. In his view, economic planning would amount to the entire community of consumers negotiating output volumes and prices with the branch producers. Since this implied that a lot of time would be needed to build an efficient system of statistical records, Kautsky believed full economic planning was a remote prospect. But what would a fulfilment of those plans actually guarantee? Kautsky failed to realize how complex a question that was, although some of his remarks, such as his comments about the important part that talented production organizers, who are as rare as talented artists, might play in socialism, sound quite up-to-date.

Opposed as he was to total state control. Kautsky was an advocate of a plurality of ownership forms in socialism. Apart from a certain scope for state ownership of production (which would not be managed by state-employed functionaries), he saw in socialism room for production cooperatives, for municipal enterprises, and for union-sponsored autonomous enterprises similar in character to those advocated by Guild Socialists. He regarded the general idea of guild socialism as excellent and inspiring, but he thought that this school focused its attention too much on producers to the detriment of consumers, and he resisted in particular attempts to present guild socialism as the only feasible production organization model for socialism.

Selected works

1887. *Karl Marx’ ökonomische Lehren*. Stuttgart.

1899. *Agrarfrage, Eine Ubersicht über die Tendenzen der modernen Landwirtschaft und die Agrarpolitik der Sozialdemokratie*. Stuttgart.

1901–2. Kriensentheorien. *Neue Zeit*, Nos. 2, 3, 4 and 5.

1902. *Die soziale Revolution*. Berlin. Trans. J.B. Askew as *The Social Revolution and On the Morrow of the Revolution*, London: Twentieth Century Press, 1903.

1910. *Vermehrung und Entwicklung in Natur und Gesellschaft*. Stuttgart.

1922. *Die proletarische Revolution und ihr Programm*. Trans. H.J. Stenning as *The Labour Revolution*, New York: Dial Press, 1925.

1927. *Die materialistische Geschichtsauffassung*. Berlin.

Bibliography

Tugan-Baranovsky, M.I. 1901. *Studien zur Theorie und Geschichte der Handelskrisen in England*. Jena: Fischer.

How to cite this article

Kowalik, Tadeusz. "Kautsky, Karl (1854–1938)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000010> doi:10.1057/9780230226203.0887

The New Palgrave Dictionary of Economics Online

kernel estimators in econometrics

Aman Ullah

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The kernel estimation method is a nonparametric procedure for analysing economic models. It is a data-based procedure which avoids the a priori parametric specification of the economic model, and it has become popular because of its wide applicability and well-developed theory. A substantial literature has developed where the local polynomial kernel estimator has been proposed to analyse various economic models, which include regression models, single-index models, dynamic time series models and panel data models. The frontier of this subject is expected to develop further in both theory and applications, especially with advances in computer technology.

Keywords

bootstrap; convergence; cross-section econometrics; curse of dimensionality; kernel estimators in econometrics; linear models; local modelling; nonlinear models; nonlinear time-series analysis; nonparametric regression; time-series econometrics

Article

For empirical research, we draw from economic theory the types of variables which can be used in the economic relationship (model) under consideration. But theory usually does not provide the functional form of the economic model. Empirical and theoretical work in econometrics is, therefore, often carried out by assuming linear or nonlinear parametric functional forms of the economic models (see Gallant, 1987, for work on nonlinear models by econometricians sparked by the work of statisticians Hartley, 1961, and Jennrich, 1969). However, these parametric models may often be mis-specified and hence they may provide biased and misleading conclusions. With this in view, econometrics moved in the direction of local modelling (local averaging), which is a data-based approach, for studying the economic relationships of unknown forms. In the regression framework this approach is also called ‘nonparametric regression’ or ‘nonparametric smoothing’. Here our focus is on nonparametric kernel

regression.

Nonparametric kernel regression methods are becoming increasingly popular for applied data analysis; they are best suited to situations involving large data-sets for which the number of variables involved is manageable. A kernel is simply a weighting function. The kernel estimation procedure was developed in the seminal published work of Rosenblatt (1956) on the density function, and later in the context of the regression function, by Nadaraya (1964) and Watson (1964). A detailed development on this subject in statistics was first presented by Prakasa Rao (1983), and then Härdle (1990) and Fan and Gijbels (1996), followed by the work of Pagan and Ullah (1999) in econometrics. There are other ways to do local modelling – for example, spline methods, series methods, differencing methods, and neural network methods (see Pagan and Ullah, 1999) – but the kernel smoothing procedure has become popular because of its vast applicability, simplicity, and well-developed theoretical underpinnings for both time-series and cross-section data. Nonparametric kernel methods essentially involve local averaging in a regression context: we can obtain a consistent estimate of the conditional mean by locally averaging those values of the dependent variable which are ‘close’ in terms of the values taken on by the regressors. The amount of local information used to construct the average is determined by a window width, also known as a ‘bandwidth’ or a ‘smoothing parameter’.

Suppose one wished to estimate the function *m* in the regression equation:

$$y_i = m(x_i) + u_i, \; i = 1, \ldots, n \tag{1}$$

where y_i is the dependent variable, x_i is a vector of q regressors, and u_i is an additive error. A parametric approach intends to fit the data to a parametric model $m(x_i) = m(x_i, \theta)$, often a linear model with $m(x_i, \theta) = \alpha + x_i\beta$, where θ is a parameter set of the model. But from the perspective of economic theory many economic models tend to be nonlinear. This makes the linear model specification inappropriate for understanding economic relationships. A way to capture the nonlinearity in data is to model the regression function locally, that is, to obtain the regression function $m(x)$ at a given point x by applying the linear regression technique to the data in a window width of size h using the linear model

$$y_i = \alpha(x) + (x_i - x)\beta(x) + u_i, \; \text{for } x_i \text{ in } x \pm \frac{h}{2} \tag{2}$$

This local linear regression method leads to the following locally weighted minimization problem:

$$\min \sum_{i=1}^{n} (y_i - \alpha(x) - (x_i - x)\beta(x))^2 K\left(\frac{x_i - x}{h}\right) \tag{3}$$

where $K(\cdot)$, a non-negative weight (kernel) function, is a decreasing function of distances of x_i from the point x, and h is a window width that determines how rapidly the weights decrease as the distance of x_i from x increases. Let $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ be the estimated local linear least squares estimators, which are the solutions of (3). Then the estimated regression function at the point $x_i = x$ is $\hat{m}(x) = \hat{\alpha}(x)$ and $\hat{\beta}(x)$ is the estimator of $\beta(x) = \partial m(x) / \partial x$ which is the local slope. If $\beta(x) = 0$ in (2) and (3), then the resulting estimator of $m(x) = \alpha(x)$ is the Nadaraya (1964) and Watson (1964) kernel regressor estimator. The local linear regression approach in (2) amounts to considering a linear Taylor series expansion of $m(x_i)$ around x in model (1). This approach can be extended to a local polynomial regression by taking a polynomial expansion of order, say p, of $m(x_i)$ around x. This provides the local polynomial least squares estimator of $m(x)$ (Stone, 1977). The local linear estimators ($p = 1$) perform better than the Nadaraya–Watson estimator ($p = 0$) with respect to bias reduction, absence of boundary effects, and the adaptation to various design situations; for $p \geq 1$ the local polynomial estimators may suffer singularity problems in applied settings.

The principle of local regression estimators can be generalized to other parametric regression settings such as local logit and probit, local proportional hazards, local quantile, robust regression, and nonlinear time-series models. For example, if we let $m(x_i, \theta)$ be a parametric model and $L_i(y_i, x_i, m(x_i, \theta))$ be the loss or the log-likelihood of the i-th observation, then we can minimize (if a loss) or maximize (if a likelihood) the objective function given by

$$L(\theta) = \sum_{i=1}^{n} L_i(y_i, x_i, m(x_i, \theta)) K\left(\frac{x_i - x}{h}\right) \tag{4}$$

The $m(x_i, \theta)$ is now locally estimated by $m(x_i, \hat{\theta}(x))$, for example, when $m(x_i, \theta) = \alpha + x_i\beta$ then $m(x_i, \theta(x)) = \alpha(x) + \beta(x)\, x_i$ and $L_i(y_i, x_i, m(x_i, \theta(x)) = (y_i - \alpha(x) - (x_i - x)\beta(x))^2$, or $L(\theta)$ is maximized with L_i written presuming normality of errors. Similarly, in a single index econometric model $L_i(y_i, x_i, m(x_i, \theta)) = \log[F(x_i\beta)^{y_i}(1 - F(x_i\beta))^{1 - y_i}]$ where $y_i = 1$ or 0 and $F(\cdot)$ is a cumulative distribution function, and in the case of a local linear *k–th* quantile regression $L_i = u_i(k - I(u_i < 0))$ where $u_i = y_i - \alpha(x) - (x_i - x)\beta(x)$, $0 < k < 1$, and $I(\cdot)$ is the usual indicator function.

The selection of window width h is by far the most important issue of nonparametric kernel estimation.

When *h* is arbitrarily small, the bias of the estimator is small but the variance is large. Conversely, when *h* is large, the estimator has a lower variance and a higher bias. Much of the literature on the methods of window width selection can really be viewed as attempts to balance this classic bias–variance trade-off. Overall, the selection rules fall into roughly three broad categories: (*a*) reference rules that would be optimal from a reference data generating process, (*b*) plug-in, penalizing, and cross-validation methods, and (*c*) bootstrap methods (see Pagan and Ullah, 1999, and Marron, 1992).

The asymptotic properties of the local polynomial estimators are well established (see Fan and Gijbels, 1996, for cross-section data and Masry, 1996, for the time-series case). The implication of these results is that the rate of convergence of the pointwise estimator of the *r–th* derivative of *m*(*x*) is the inverse of $(nh^{q+2r})^{\frac{1}{2}}, r \geq 0$, which is slower than the parametric rate of $\sqrt{n}$. In fact, as the dimensions of the regressors *q* for a given *r* increase, the rates become worse, which is the well-known ‘curse of dimensionality’ problem. However, the rate of convergence of the average of the pointwise estimators (global estimators) of these derivatives is widely known to have $\sqrt{n}$ rate of convergence. One of the most popular ways to deal with the ‘curse of dimensionality’ is to consider the nonparametric additive regression model which can be written as $y_i = \beta_0 + \Sigma_{j=1}^{q} m(x_{ij}) + u_i$. Imposing this additivity provides an estimator having a one-dimensional nonparametric rate of convergence.

In recent years the kernel regression estimation methods have progressed in various directions. These include testing for the significance of a regressor or group of regressors, consistent testing for the correct parametric functional form, estimation of the so-called ‘structural relationship’ among endogenous (dependent) variables, and the estimation of various types of semiparametric models consisting of a combination of parametric and nonparametric models (see Pagan and Ullah, 1999). Extensive work on the empirical applications of the kernel regression estimation have begun to appear in both cross-section econometrics and time-series econometrics, especially in labour economics and empirical finance. Although some related work is being done, several challenging research issues remain to be worked out. The first is the development of a unified approach towards a data-driven window width, and the development of software that permits fast computation of kernel-based estimators and test statistics for large data-sets in a desktop environment. The second is the development of kernel-based estimation of time-series models for non-stationary data. Third is the systematic development of the work on kernel estimation of panel-data models with heterogeneity parameters, especially when the time-series component of the data is large. Finally, the development of the theory of kernel estimation of various econometric models with both continuous and discrete variables is important, especially for the empirical applications of the kernel regression methods (see Racine and Li, 2004).

The nonparametric kernel regression method is a dynamic area, and there are rapid ongoing theoretical advances. With advances in computer technology, applications of the kernel regression approach continue to increase. The developments described above provide the dimensions in which the kernel-estimation procedures have been explored in econometrics and statistics. In a broad sense, the frontier of this research area has moved on, and is expected to continue with further developments in both its theory and applications.

See Also

- econometrics
- nonlinear time series analysis
- nonparametric structural models
- robust estimators in econometrics

Bibliography

Fan, J. and Gijbels, I. 1996. *Local Polynomial Modeling and its Applications*. London: Chapman and Hall.

Gallant, A. 1987. *Nonlinear Statistical Models*. New York: John Wiley & Sons.

Härdle, W. 1990. *Applied Nonparametric Regression*. New York: Cambridge University Press.

Hartley, H. 1961. The modified Gaus–Newton method for the fitting of nonlinear regression functions by least squares. *Technomterics* 3, 269–80.

Jennrich, R. 1969. Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics* 40, 633–43.

Marron, S. 1992. Bootstrap bandwidth selection. In *Exploring the Limits of Bootstrap*, ed. R. Lepage and L. Billard. New York: John Wiley.

Masry, E. 1996. Multivariate regression estimation local polynomial fitting for time series. *Stochastic Process Application* 65, 81–101.

Nadaraya, E. 1964. On estimating regression. *Theory of Probability and its Application* 9, 141–2.

Pagan, A. and Ullah, A. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Prakasa Rao, B. 1983. *Nonparametric Functional Estimation*. New York: Academic Press.

Racine, J. and Li, Q. 2004. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119, 99–130.

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* 27, 642–69.

Stone, C. 1977. Consistent nonparametric regression. *Annals of Statistics* 5, 267–84.

Watson, G. 1964. Smooth regression analysis. *Sankhya*, Series A 26, 359–72.

How to cite this article

Ullah, Aman. "kernel estimators in econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000069> doi:10.1057/9780230226203.0888

The New Palgrave Dictionary of Economics Online

Keynes, John Maynard (1883 – 1946)

Don Patinkin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

J. M. Keynes was the greatest political economist of the first half of the 20th century. This article traces the development of his thinking about economic theory and policy. It focuses largely on the inter-war trilogy, the *Tract on Monetary Reform* (1923), the *Treatise on Money* (1930), and the *General Theory of Employment, Interest and Money* (1936), in which Keynes's monetary thought evolved from the quantity-theory tradition he had inherited, changed the face of monetary theory, laid the foundation for its development into macroeconomic theory, and defined the analytical framework and research programme of this theory for decades to come.

Keywords

Beveridge, W.; Bretton Woods agreement; business cycles; Cambridge Circus; cheap money; Clark, C.; credit cycle; demand for money; Econometric Society; effective demand; Fisher, I.; fundamental equations; gold standard; Harrod, R.; Hawtrey, R.; Hicks, J.; inflation; International Monetary Fund; Kahn, R.; Kalecki, M.; Keynes, J. M.; Keynes, J. N.; Kuznets, S.; Leontief, W.; Lerner, A.; liquidity preference; liquidity premium; liquidity trap; Macmillan Committee on Industry and Finance; marginal efficiency of capital; Marshall, A.; mathematical economics; Meade, J.; Modigliani, F.; Myrdal, G.; national-income estimates; Ohlin, B.; paradox of thrift; partial equilibrium theory; Patinkin, D.; Pigou, A.; public works; purchasing power parity; quantity theory of money; Robertson, D.; Robinson, A.; Robinson, J.; Say's Law; Shackle, G.; Sraffa, P.; Stockholm School; stocks and flows; Tinbergen, J.; unemployment equilibrium; velocity of circulation; Viner, J.; wage rigidity; Walras, L.; Wicksell, K.; World Bank

Article

John Maynard Keynes was one of the great intellectual innovators of the first half of our century, and certainly its greatest political economist. He was born in Cambridge on 5 June 1883, and died at Tilton (in Sussex) on 21 April 1946. His father was John Neville Keynes, also an economist, author of *The Scope and Method of Political Economy* (1891), and later registrary of Cambridge University.

With the help of a scholarship, Keynes was educated at Eton. He then went on to King's College, Cambridge, where he took a degree in mathematics in 1905. Afterward he spent an additional year at Cambridge studying economics under the then-doyen of British economics, Alfred Marshall, as well as under the latter's student and successor-to-be as Professor of Political Economy at Cambridge, Arthur Pigou. Keynes then entered the Civil Service, where he worked for over two years in the India Office, though he never actually visited India. Out of this work grew his first book in economics, *Indian Currency and Finance* (1913), which was largely descriptive in nature, and whose main concern was not the Indian monetary system as such – and *a fortiori* not the Indian economy – but with this system as an example of the workings of a gold-exchange standard. This work also led to Keynes's first major participation in public life as a member of the Royal Commission on Indian Finance and Currency (1913–14).

In 1908 Keynes returned to Cambridge as a Lecturer in Economics (some of Keynes's notes for his lectures during this period have survived and are reproduced in *JMK* XII, pp. 689–783). During that year he continued his work on *A Treatise on Probability*, which he successfully submitted to King's College as a fellowship dissertation in 1909. This dissertation was published in a revised form in 1921 and continues to be recognized as a pioneering work in the field.

Shortly after the outbreak of World War I, Keynes took a leave of absence from Cambridge to enter the Treasury. Here his exceptional ability and capacity for work led to his rapid advancement, and by 1919 he was principal Treasury representative at the Peace Conference at Versailles. His passionate disagreement with what he considered to be the harsh

clauses of the Versailles Peace Treaty led to his resignation from the British delegation and to the writing of his vehement denunciation of the treaty in his *Economic Consequences of the Peace* (1919), which was translated into many languages and overnight made him a world celebrity. From then on Keynes was an international figure whose voice was heard on all major economic problems that arose in interwar Britain and, indeed, in the Western world as a whole.

In 1925 Keynes married the Russian ballerina Lydia Lopokova, a leading member of Diaghelev's company in the early 1920s. They had no children.

(The present essay is devoted almost entirely to the development of Keynes's thinking about economic theory and policy. For full biographical studies of Keynes, see Austin Robinson, 1947; Harrod, 1951; Milo Keynes (ed.), 1975; Moggridge, 1980; and Skidelsky 1983, 1992 and 2000.)

1. In our profession, Keynes is known primarily for his fundamental contributions to monetary economics. The *Tract on Monetary Reform* (1923; henceforth *Tract*), the *Treatise on Money* (1930; henceforth *Treatise* or *TM*), and the *General Theory of Employment, Interest and Money* (1936; henceforth *GT*): this is the inter-war trilogy that marks the development of Keynes's monetary thought from the quantity-theory tradition that he had inherited from his teachers at Cambridge; to his subsequent systematic attempt to dynamize and elaborate upon this theory and its applications; and, finally, to the revolutionary work (as Klein, 1947, so rightly termed it) which he wrote under the constant stimulus and criticism of his colleagues and students – and with which he changed the face of monetary theory, laid the foundation for its development into macroeconomic theory, and defined the analytical framework and research programme of this theory for decades to come.

(The following discussion draws freely on the material in Patinkin 1976a, 1977 and 1982, to which the reader is referred for further details; all references to Keynes's writings are to the form in which they appear in the relevant volumes (most of which were edited by Donald Moggridge) of the Royal Economic Society's edition of his *Collected Writings*, referred to henceforth as, e.g., *JMK* IX, *JMK* XIII, and so forth. Though it has its faults (see Patinkin, 1975, section I; 1980. pp. 2–3 (especially n.2 and n.6), p. 8 (n.14), and pp. 14–15 (n.22 and n.23); see also Schefold, 1980, and section 3 below), this edition – to paraphrase one of the famous passages of the *Treatise* – is verily a widow's cruse from which students of the development of Keynes's thought will continue to draw materials for years to come, without diminution in the profits to scholarship.)

Though I have referred to Keynes's three books on monetary theory as a trilogy, they differ from each other greatly not only in substance (a difference that has, of course, been a major theme of all studies of the development of Keynes's thought) but also in form and purpose. Thus the *Tract* is not really a book, but in large part a revision and elaboration of the series of article on postwar economic policy that Keynes first published in 1922 in the ‘Reconstruction Supplements’ (which he edited) of the *Manchester Guardian Commercial*, with the addition of material that is not always integrated with that from the series.

Thus chapters 1 and 3:2 of the *Tract* are based on these articles and deal with the pressing problems of inflation, deflation, and the resulting exchange rate disequilibrium that then beset Europe. Keynes analysed this disequilibrium in terms of the purchasing-power-parity theory, which he expounded in detail and tested with contemporary data from the countries involved. In the new material presented in chapter 4, he then provided a lucid analysis of the basic dilemma between the ‘alternative aims’ of stability of the internal price level and stability of the exchange rate – and strongly argued the view that he was to reaffirm in the *Treatise* of giving precedence to the aim of internal price stability. Similarly, the brief, formal presentation of monetary theory that appears in chapter 3:1 of the *Tract* – and which, as Keynes tells us (*Tract*, p. 63, n.1) ‘follows the general lines of Professor Pigou … and of Dr Marshall’ – is part of the material that Keynes added to these articles in making up the book.

In this context, Keynes presents the ‘famous quantity theory of money’ in the following terms:

> Let us assume that the public, including the business world, find it convenient to keep the equivalent of k consumption units in cash and a further k' available at their banks against cheques, and that the banks keep in cash a proportion r of their potential liabilities (k') to the public. Our equation then becomes
>
> $$n = p(k + rk')$$
>
> [where n is the quantity of money and p the price level]. So long as k, k' and r remain unchanged, we have the same result as before, namely, that n and p rise and fall together (*Tract*, p. 63).

This equation is nothing but a minor variation on the famous ‘Cambridge equation’ that Pigou had first presented in print in his classic 1917 article (p. 166), to which Keynes at this point refers.

Similarly, when he goes on to explain the determinants of k and k', Keynes states that the matter cannot be summed up better than in the words of Dr Marshall:

> ‘In every state of society there is some fraction of their income which people find it worth while to keep in the form of currency; it may be a fifth, or a tenth, or a

> twentieth. A large command of resources in the form of currency renders their business easy and smooth, and puts them at an advantage in bargaining; but on the other hand it locks up in a barren form resources that might yield an income of gratification if invested, say, in extra furniture; or a money income, if invested in extra machinery or cattle.' A man fixes the appropriate fraction 'after balancing one against another the advantages of a further ready command, and the disadvantages of putting more of his resources into a form in which they yield him no direct income or other benefit.' 'Let us suppose that the inhabitants of a country, taken one with another (and including therefore all varieties of character and of occupation), find it just worth their while to keep by them on the average ready purchasing power to the extent of a tenth part of their annual income, together with a fiftieth part of their property; then the aggregate value of the currency of the country will tend to be equal to the sum of these amounts' (*Tract*, p. 64).

The words are from Marshall's *Money, Credit and Commerce* (1923), pp. 44–5. In this source, however, Marshall indicates that in large part they go back to his testimony before the Indian Currency Committee in 1899 (reproduced in Marshall's *Official Papers* [1926], esp. pp. 267–9).

Just as this theoretical material was (by Keynes's 'revealed preference') not necessary for an understanding of the original articles in the *Manchester Guardian*, so is it not really necessary for the book: its deletion would interfere very little with an understanding of the argument of the *Tract* at other points, as indeed Keynes indicated (*Tract*, p. 61n). Conversely (and this is one of the clearest manifestations of the failure of the *Tract* to be an integrated whole) this added theoretical material in chapter 3:1 barely reflects the penetrating and elegant analysis of inflation as a tax on real money balances (including the notion of an optimum rate of inflation!) that Keynes reproduces from the aforementioned articles in chapter 2:1 of the *Tract* – and that can be read with both profit and pleasure even today.

Nor does the *Tract* incorporate the dynamic analysis of the way in which an influx of gold operates through the banking system – and thence on prices – that Keynes (basing himself on Marshall) had summarized in his long 1911 review of Irving Fisher's *Purchasing Power of Money* (1911), a review that I would essentially consider to be Keynes's first published work on monetary theory. Thus the *Tract* – as a theoretical work – is not only not integrated within itself, but even fails to reflect some major aspects of Keynes's thinking about monetary problems at the time it was published.

2. On both of these scores the *Treatise* (on which Keynes began working less than a year after the appearance of the *Tract*) is the exact opposite. It is as specifically designed for a professional audience whose major concern was with the latest developments in monetary theory as the *Tract* was designed for a general audience whose major concern was with current policy. Indeed, from the viewpoint of traditional scholarship, the *Treatise* is Keynes's most ambitious and weighty work: the two-volume work – on 'The Pure Theory of Money' and 'The Applied Theory of Money' – designed to endow him with an academic reputation that would match the public one he had already achieved. At its core (in Books III–IV of Volume I) is a formal, rigorous presentation of a theory of money that deals in detail with both the static and dynamic aspects of the problem. And in the slow, stately, and systematic manner in which an academic treatise customarily proceeds – but in which Keynes of the interwar period so rarely proceeded – it leads up to this core, first, by defining the nature of money and describing its historical origins (Book I); and then (in Book II) describing at length the various index numbers that can be used to measure the value of money, which (to use one of Keynes's favourite terms) is the quaesitum of monetary theory. And afterwards comes Volume II, which begins with a lengthy description of the respective empirical magnitudes of the critical theoretical variables described in the preceding volume – as well as the institutional features of the financial sectors which bear upon these variables (Books V–VI). Only when all this is completed does Keynes finally proceed (in Book VII) to a systematic presentation of the monetary policy, both domestic and international, that he derives from his theory.

The basic problem that Keynes set out to analyse in the *Treatise* was that of the 'credit cycle' and the fluctuations in employment and output which characterize it. His analysis was essentially a simple one: profits – by which Keynes means profits above those representing a normal return on capital – are the motive force of the economy (*TM* I, pp. 126, 163). The existence of profits causes firms to expand their respective outputs and hence their demands for the inputs of productive services – and conversely for losses. Now (in the Marshallian terms that Keynes used: *Principles*, Book III, ch. III and Book IV, ch. I), profits are the difference between the 'demand price' (i.e. market price; cf. *TM* I, pp. 186, 189) of a unit of output and its 'supply price' (i.e. cost of production). Hence the study of cyclical movements of output reduces to a study of the causes of the differential movements of prices and costs.

It is these movements that Keynes then tries to analyse rigorously by means of his 'fundamental equations'. These are derived (in Chapter 10 of the *Treatise*) after first distinguishing between 'consumption goods' and 'investment goods' and then defining the following basic variables of the analysis, where all variables refer to total or aggregate quantities. (For simplicity, and since my main concern is to compare the *Treatise* with the *General Theory*, I disregard the variables relating to foreign investment, which actually plays an important role in the *Treatise*):

E=current money income=factor earnings (including normal return on capital)=costs of production; all exclusive of abnormal profits;

O=the same, at base-period prices;

I' =that part of E earned in the investment-goods sector=current money costs of producing investment goods;

C=the same, at base-period prices;

I=the same, at current market prices, i.e. the current market value of investment goods produced;

$E–I'$ =that part of E earned in the consumption-goods sector=current money costs of producing consumption goods; and

R=the same, at base-period prices.

Keynes then proceeds to define the price variables

P=current price level of consumption goods;

P' =the same, for investment goods; and

Π=the same; for output as a whole=the weighted average of P and P' =the general price level.

Keynes implicitly (and sometimes explicitly) assumes that the base period is one of the equilibrium – defined as a situation in which per-unit price=per-unit costs in both the consumption-goods and investment-goods sectors. Hence there is no difference between evaluating current output at base-period prices and evaluating it at base-period costs of production. He then defines what are effectively (1) an index of the money wages per unit of labour, W (where labour represents factors-of-production-in-general) and (2) an index of output per worker, e (or the 'coefficient of efficiency'); and he implicitly assumes that both of these indexes change in exactly the same way in both sectors. From these definitions it then follows that the change in the cost of production with respect to the base period in both the consumption and investment sectors is

$$E / O = W / e = W_1,$$

where W_1 (which Keynes calls 'the rate of efficiency earnings') is accordingly an index of costs of production per unit of output.

From all this, Keynes then derives his two fundamental equations in the following alternative forms:

$$P = (E / O) + (Q_1 / R) = (W / e) + (Q_1 / R) = W_1 + (Q_1 / R) \quad \text{(i)}$$

$$\Pi = (E / O) + (Q / O) = (W / e) + (Q / O) = W_1 + (Q / O) \quad \text{(ii)}$$

where Q_1 and Q represent profits in the consumption sector and in the economy as a whole, respectively. Thus all that fundamental equation (i) consists of is the quite obvious statement that the change (with respect to the base period) in the price of consumption goods equals the change in the per-unit costs of production of these goods (the first term of equation (i)) plus the change in the per-unit (abnormal) profits, assumed zero in the base period (the second term); and equation (ii) makes a correspondingly obvious statement for output as a whole.

The deeper meaning that Keynes attributed to these equations stemmed from his demonstration that profits Q_1 and Q were related to savings and investment. In particular, he first defined current savings S as the difference between income (defined, it will be recalled, as exclusive of abnormal profits) and consumption, or

$$S = E - PR,$$

where all variables are defined in current money terms. From this definition and those listed above, it follows that (abnormal) profits in the consumption sector are

$$Q_1 = PR - (E - I') = I' - S,$$

whereas total (abnormal) profits in the economy are

$$Q = (PR + I) - E = I - S.$$

Thus one of the distinguishing features of the *Treatise* is that as a result of its special definition of income, savings and investment need not be equal even ex post. The fundamental equations can then be written as

$$P = E / O + (I' - S) / R \quad \text{(i)}'$$

and

$$\Pi = E / O + (I - S) / O \quad \text{(ii)}'$$

– and this, indeed, is their primary form in the *Treatise* (I, pp, 122–3). In this way a change in the general price level – which for Keynes of the *Treatise* (like other monetary economists of that time and earlier, such as Knut Wicksell, Irving Fisher, A.C. Pigou) was the central concern of monetary theory – was directly related to the excess of investment over savings. When $I' = I = S$, the second terms of (i)′ and (ii)′ respectively, disappear, so that price=cost of production (including normal return on capital), and the economy is in equilibrium.

It must be emphasized that though the relation between savings and investment plays a central role in the *Treatise*, this relation served there (in sharp contrast with the subsequent *General Theory*) to analyse in the first instance not changes in output, but changes in prices. Correspondingly, though as indicated, Keynes does discuss changes in output in the *Treatise*, he considers these to be derivative from the changes in prices.

Keynes recognized that his equations were identities, and indeed said so; but he also claimed that they were identities that were useful for classifying causal relationships (*TM* I, p. 125; see also p. 120). In particular, the causal relationship to which he assigned a crucial role in his theory was that connected with the rate of interest. Thus, if we start from a position of equilibrium, a (say) decrease in this rate would cause investment to increase and savings to decrease, thus generate an excess of the former over the latter, thus generate profits, and thus – as indicated by the second term of the second fundamental equation – cause prices to rise. In this way, says Keynes, a decrease in the rate of interest would ‘in itself’ cause a price rise – and not only (as in the traditional quantity theory) as the result of its first generating an increase in the quantity of money (*TM* I, pp. 167–76, esp. p. 171). Conversely, an increase in the rate of interest would directly cause prices to fall. Explicitly following Wicksell, Keynes denotes the rate of interest that would equate savings and investment (and thus generate equilibrium in the system) ‘the natural rate of interest’; and the rate which actually prevails, ‘the market rate’ (*TM* I, p. 139).

Keynes made use of the causal interrelationship of interest and prices to provide a dynamic analysis of the change in the price level generated by a change in the quantity of money –

by which Keynes meant currency *plus* total bank deposits, which because of the relative unimportance of the former in a modern economy can be conveniently approximated by these deposits alone (*TM* I, p. 27). For this purpose he first decomposes total deposits into ‘the industrial circulation’ (roughly, demand deposits) and ‘the financial circulation’ (roughly, savings or time deposits) (*TM* I, chs 15 and 17). These in turn roughly correspond to what were to become the transactions and precautionary-speculative balances of the *General Theory* (pp. 167 n.l, 194–6).

Similarly, the *Treatise* contains some of the major features of what was to become the liquidity-preference theory of the *General Theory*. The presentation in the *Treatise* is less precise in that it does not adequately analyse the nature of the ‘liquidity premium’ and explicitly present the corresponding *functional* relationship between the demand for money and the rate of interest. On the other hand, it is more precise with respect to the distinction between stocks and flows: between the stock of wealth on whose asset composition the individual must decide; and the flow of income, with respect to which the individual decides on how much to consume and how much to save, i.e. to add to his wealth (*TM* I, p. 127). (The emphasis on the distinction between stocks and flows and the specification of a functional relationship are the two major features which distinguish the liquidity-preference theory of the *Treatise* and *General Theory* from the Cambridge cash-balance theory which Keynes espoused in his *Tract*; cf. Patinkin, 1974.) In any event, Keynes explains that the volume of savings deposits (i.e. the financial circulation) is determined by the decision of individuals as to what proportion of their wealth to hold in the form of such deposits as compared with the alternative of holding securities, a decision that depends (*inter alia*) on the rate of interest (*TM* I, ch. 10, s.3). Insofar as the industrial circulation is concerned, this is determined by the basic relationship $M_1V_1 = E$, where M_1 is the volume of demand deposits, V_1 their velocity of circulation, and E the level of aggregate money income=aggregate money costs of production (or W_1O). In the real world, V_1 is largely determined by institutional factors and hence remains more or less constant in the short run.

Let us now start from an initial position of equilibrium in which, by definition, the market rate of interest equals the natural rate, so that $I' = I = S$. Assume that this equilibrium is disturbed by an increase in the quantity of money. Initially, only part of this increase will be absorbed in the industrial circulation; part will be used to bid up the price of securities and thus lower the rate of interest. Furthermore, the increase in the quantity of money will have increased the reserves of the banks, thus inducing them to lower the rate of interest at which they lend. As a result, entrepreneurs will increase their borrowings in order to finance the undertaking of new projects, so that investments will begin to exceed savings, thus generating excess profits and an increase in the price of output. But as a result of these profits, firms will begin to expand their outputs, thus generating an increased demand for labour inputs, hence an increase in the wage rate and thereby in the per-unit cost of production. That is, $E=W_1O$ will increase, and with it the need for the industrial circulation. This process will continue until money wages have risen sufficiently to eliminate excess profits and until all of the new money has been absorbed in the increased demand for the industrial circulation generated by the increase in W_1 and hence in E. In Keynes's words:

> This [process] must continue until $(M_1V_1)/O$ has settled down at a higher figure, which is in equilibrium with the new total quantity of money and also with values of P and P' which are enhanced relatively to their old values in a degree corresponding to the amount by which $(M_1V_1)/O$ has been increased (*TM* I, p. 241).

This conclusion has the unmistakable ring of the quantity theory. And indeed Keynes explains that his second fundamental equation can be rewritten as

$$\Pi = (M_1V_1)/O + (I - S)/O \qquad \text{(ii)}''$$

which in equilibrium (i.e. when $I = S$) reduces to the Fisherine

$$M_1V_1 = \Pi O.$$

Thus (emphasizes Keynes) for the purpose of comparing equilibrium positions (i.e. for purposes of comparative statics), the traditional quantity theory does indeed remain valid. The purpose of the *Treatise* in this context, however, is, first, to extend this theory to an economy with a developed banking system, and then to analyse the dynamics of the movement from one equilibrium position to another in such an economy. And this is the role of the interest-rate savings-investment mechanism as it manifests itself in the fundamental equations (*TM* I, pp. 120, 131–3, 137–8). Indeed, at the beginning of Volume II of the *Treatise*, Keynes summarizes the dynamic workings of his second fundamental equation by first writing

the quantity equation in the form $M_1V' = \Pi O$ and then stating that the purpose of his new theory is to explain how 'during the transition from one position of equilibrium to another' the overall velocity of circulation V' deviates upwards or downwards from its normally constant level, V_1, in accordance with whether $I - S > 0$ or $I - S < 0$, respectively (*TM* II, pp. 4–5; Patinkin 1976a, p. 46, n.2). Thus Keynes regarded his *Treatise* not as a refutation of the quantity theory, but as an extension of it.

The general policy proposal of the *Treatise* follows directly from its theoretical analysis: if the 'credit cycle' is generated by the alteration of prices with respect to costs, thus generating profits (losses) and hence increases (decreases) in output and employment, then, claimed Keynes (as had Wicksell, Fisher and Pigou before him – and the Chicago School of the 1930s afterwards: Patinkin, 1969), the way to stabilize the economy was to stabilize the price level. And, continued Keynes, the major policy variable for achieving this objective is the Bank Rate as fixed by the central bank, which should be raised when prices tend to rise and lowered when they tend to fall.

At the same time, Keynes recognized that in the gold-standard world which then existed, an undue lowering of the rate of interest in one country relative to others might generate a capital outflow and consequent dangerous loss of gold reserves; hence such 'international complications' might prevent the central bank from lowering the rate of interest sufficiently to deal with a depression. And Britain – which was a major centre of international trade and finance – was particularly vulnerable in this respect. For this reason, in the *Treatise* (II, pp. 337–38), as in the 'private evidence' that he gave before the Macmillan Committee when he was in the final stages (February–March 1930) of preparing this book (*JMK* XX, pp. 71, 125–32), and as in his earlier political pamphlet *Can Lloyd George Do It?: An Examination of the Liberal Pledge* (1929; *JMK* IX, pp. 118–19, 123–4) – Keynes's policy advice for Britain at that time was to combat the depression that beset it not by further reductions in the rate of interest, but by an increase in government expenditures on public works. On the other hand, the United States – which was in much less danger of loss of gold reserves due to international capital movements – should indeed combat its depression by means of a central-bank policy of lowering the rate of interest. This policy difference between Britain and the United States was repeatedly and most explicitly stressed by Keynes in his contributions to the round-table discussions at the 1931 Harris Foundation lectures in Chicago (1931b, pp. 84, 92, 303; see Patinkin, 1979a, pp. 292–3). Accordingly, when in September 1931 Britain abandoned the gold standard, Keynes immediately advocated that it reduce the rate of interest, thus laying the basis for the well-known 'cheap-money' policy of subsequent years (Moggridge and Howson, 1974; Howson and Winch, 1977, pp. 57–8; Patinkin, 1979b).

3. Keynes had great hopes for the *Treatise*. Thus shortly after its publication, in his June 1931 Harris Foundation lecture on 'An Economic Analysis of Unemployment', he explicitly made use of the analysis of this book and proclaimed, 'That is my secret, the clue to scientific explanation of booms and slumps (and of much else, as I should claim) which I offer you' (*JMK* XIII, p. 354). But these hopes were not to be fulfilled. For it rapidly became clear that the theoretical part of the book was not a success and was indeed subjected to severe criticism. To a certain extent this was due to the fact (which Keynes had only in part and somewhat grudgingly recognized (see *TM* I, pp. 176–8, especially p. 177, n.3, and p. 178, n.2) that this theory, as well as the corresponding policy proposal, had been largely adumbrated at the turn of the century by Wicksell (1898, 1906, 1907) – which brought on Gunnar Myrdal's (1933, pp. 8–9) chiding remark about 'the attractive Anglo-Saxon kind of unnecessary originality, which has its roots in certain systematic gaps in the knowledge of the German language on the part of the majority of English economists'. (In point of fact, Keynes – at least before World War I – knew German well enough to review in the *Economic Journal* several books written in that language (see the reviews reprinted in *JMK* XI, pp. 400–403, 562–74); it is, however, not difficult to believe that in the course of fifteen years, Keynes might have lost a good deal of his proficiency in that language). But the most telling criticism of the *Treatise* was that, on the one hand, its 'fundamental equations' were actually tautologies, and, on the other, that the book had explained the forces that caused output to expand or contract, but had not explained what determines its actual level during any period. (See the end of section 8 below for a discussion of circumstances connected with the writing of the *Treatise* that also contributed to its lack of success.)

As a result of this criticism, Keynes began within a relatively short time after the appearance of the *Treatise* to work on a new book which ultimately developed into the *General Theory* (1936). The chronology of this development can in part be traced by means of the materials (including correspondence, fragments of earlier drafts, and galleys of successive proofs) that Moggridge has reproduced and annotated in *JMK* XIII–XIV and XXIX. There can, however, be legitimate differences of opinion about the dating of some of these fragments (cf. Patinkin, 1976a, p. 71, n.7; 1980, pp. 14–15, n.22 and n.23, and pp. 18–19); so we are extremely fortunate to be able to supplement them with the precisely dated materials in the unique 'archaeological' record of the successive 'strata' of Keynes's thought provided by Robert Bryce's notes on Keynes's weekly lectures during the autumn terms of the years 1932, 1933, 1934 and Lorie Tarshis's notes for these years as well as 1935 (reproduced in Rymes (ed.), 1988). The first year after the publication of the *Treatise* (viz., 1930–31) was devoted to a criticism of this book, greatly aided by the detailed comments of Ralph Hawtrey and the extensive discussions that took place in the so-called 'Cambridge Circus' (in the sense of 'circle') – or what today would probably be called the 'Cambridge Colloquium'. The major participants of this legendary 'Circus' were Keynes's younger colleagues, Richard Kahn, James Meade, Austin Robinson, Joan Robinson and Piero Sraffa, with his former student Kahn serving as the channel of communication between Keynes and the group (*JMK* XIII, pp. 337–43; Kahn, 1984, pp. 105–11: Keynes at that time was in his late forties, whereas the members of the 'Circus' were mostly in their mid-twenties). The aforementioned lecture notes, however, show that the central message of the *General Theory* (explicated below) was not fully developed until some time in 1933, well after the activities of the 'Circus' as such had come to an end (Patinkin, 1976a, chs 7–8; 1977; 1982, ch. 1). However, from some of the younger members of the 'Circus' (especially Kahn and Joan Robinson) – as well as from his contemporaries, Ralph Hawtrey and Dennis Robertson – Keynes continued to seek out and benefit from criticisms throughout the process of working through and revising the successive drafts of the *General Theory* (cf. *JMK* XIII, ch. 5; *JMK* XXIX, ch. 3; Patinkin and Leith (eds), 1977, *passim*).

Like the *Treatise*, the *General Theory* is – in Keynes words of his preface – 'chiefly addressed to … fellow economists'. It differs from the *Treatise* in being almost exclusively concerned with theory. Indeed, this is the whole purpose of the book, as indicated by its very title. Thus the *General Theory* contains practically no description of institutional details.

And for a work that is credited with having initiated a revolution in fiscal policy, it contains surprisingly few explicit discussions of the policy implications of its analysis. Indeed, the major new policy conclusion of the *General Theory* as compared with the *Treatise* – namely, that monetary policy directed at lowering the interest rate, though an essential component of a full-employment policy, might not be enough even in the absence of 'international complications' to achieve this goal, so that an effective policy for this purpose may well require direct government spending – this conclusion is never developed systematically and in detail. Indeed, it is only referred to on one or two occasions in passing (e.g. *GT*, p. 164) and in brief 'Concluding Notes' of a general nature (*GT*, pp. 372–84). Thus, the advocacy per se of public-works expenditure was not the purpose of the *General Theory*; rather it was to provide a theory which would, among other things, rationalize such a policy – with the actual advocacy of the policy being left for Keynes's public activities of the period (see section 11 below).

Similarly, the problem of the relation between internal price levels and exchange rates – and indeed the whole problem of the international monetary system and its relation to domestic policies, which were a major concern of Keynes in the *Treatise*, as they had been in the *Tract*, and were again to be at Bretton Woods toward the end of World War II – are not discussed in the *General Theory*. The explanation for this fact too probably lies in the situation that prevailed in the Western world during the period that the *General Theory* was being written. In particular, this was the new world ushered in by England's abandonment of the gold standard: a world of flexible exchange rates and/or severe restrictions on the flow of international trade, in which the aforementioned problems had accordingly largely lost their relevance. Correspondingly, the analysis of the *General Theory* is carried out almost entirely on the implicit assumption of a closed economy.

I should, however, emphasize that if from these viewpoints the *General Theory of Employment, Interest and Money* was more narrowly conceived than the *Treatise on Money*, from another viewpoint it is – as its title indicates – much broader. For 'monetary theory' in the *Treatise* means, first and foremost, a theory that explains the determination of the price level. Accordingly, if the argument of the *Treatise* revolves about Keynes's 'fundamental equations', these are (as the title of its chapter 10 makes clear) 'The Fundamental Equations *for the Value of Money*' (*TM* I, p. 151, italics added). Again, Keynes prefaces Book VI of the *Treatise*, 'The Rate of Investment and Its Fluctuations', with the statement that it is 'in the nature of digression, which is doubtfully in place in a treatise on money' (*TM* II, p. 85). In conformity with this view – and in sharp contrast with the systematic attempt of the *General Theory* to base its analysis on the marginal concepts of value theory and thus integrate monetary and value theory (*GT*, pp. 292–3) – the term 'marginal productivity' (of labour or of capital) does not appear in the *Treatise*. Thus though, as noted above, Keynes attributes the term 'natural rate of interest' to Wicksell, he does not follow the latter in associating this term with the marginal productivity of capital (Wicksell, 1898, pp. 102–4, 171; 1906, pp. 192–3; 1907, pp. 214–19). Finally, and as a corollary of the primary concern of the *Treatise* with prices, whereas that book deals with output only as derivative from changes in price and in this context indicates only the direction of change of output and employment, the *General Theory* presents a theory of the determination of the equilibrium levels of these variables.

A more precise specification of the basic contention of the *General Theory* can be obtained by letting Keynes speak for himself, as he did in a letter to Roy Harrod in August 1936, commenting on a draft of the latter's review article of the *General Theory* – a letter whose first and most important point largely repeats what Keynes had written to Abba Lerner two months earlier on his review (see *JMK* XXIX, pp. 214–16):

> You don't mention *effective demand* or, more precisely, the demand schedule for output as a whole, except in so far as it is implicit in the multiplier. To me the most extraordinary thing, regarded historically, is the complete disappearance of the theory of demand and supply for output as a whole, i.e., the theory of employment, *after* it had been for a quarter of a century the most discussed thing in economics [presumably, the quarter-century between the beginning of the Ricardo–Malthus debate on the possibility of a 'general glut in the market' in 1820 and the appearance of J.S. Mill's *Principles of Political Economy* in 1848; see also the reference to this period in the *General Theory* (pp. 32–4)]. One of the most important transitions for me, after my *Treatise on Money* had been published, was suddenly realizing this. It only came after I had enunciated to myself the psychological law that, when income increase, the gap between income and consumption will increase, – a conclusion of vast importance to my own thinking but not apparently, expressed just like that, to anyone else's. Then, appreciably later, came the notion of interest being the measure of liquidity preference, which became quite clear in my mind the moment I thought of it. And last of all, after an immense amount of mudding and many drafts, the proper definition of the marginal efficiency of capital linked up one thing with another (cited from the 'Editorial Introduction' to the *General Theory JMK* VII, p. xv, italics in original; there are significant errors of transcription in this passage in the full text of this letter as reproduced in *JMK* XIV, pp. 83–6: see Patinkin, 1976a, p. 66, n.3).

Now, in the *General Theory* (p. 141) Keynes himself had attributed priority for the notion of the marginal efficiency of capital to Irving Fisher. Insofar as the theory of liquidity preference is concerned, this is clearly a contribution of Keynes, but (as noted above) it is one whose basic features had already been presented in the *Treatise*. This leaves the theory of effective demand as the distinctive analytical contribution of the *General Theory* and its central message (on the meaning and significance of this last term, see Patinkin 1982, chs 1 and 4).

That this is its central message is also clear from the *General Theory* itself. Thus Keynes tells us in its preface that, in contrast with his earlier *Treatise*, his new work is 'primarily a study of the forces which determine changes in the scale of output and employment as a whole'; gives chapter 3 of 'Book I: Introduction' the title 'The Principle of Effective Demand', and presents in it a 'summary of the theory of employment' that he will develop in the book (*GT*, p. 27); and devotes most of the remaining chapters of the *General Theory* to this development.

Figure 1 reproduces the familiar diagram which has served to transmit the central message of the *General Theory* to generations of economics students. I wish, however, to refine the usual analysis which accompanies this diagram in one respect. In particular, what I mean by the theory of effective demand is not only that the intersection of the aggregate-demand curve $E = F(Y)$ with the 45° line determines equilibrium real output Y_0 at a level that may be below that of full employment Y_F not only (as Leijonhufvud (1968) has also emphasized) that disequilibrium between aggregate demand and supply causes a change in output and not price; but also (and this is the distinctively novel feature) that the change in output (and hence income) itself acts as an equilibrating force. That is, if the economy is in a state of excess aggregate supply at (say) the level of output Y_1, then the resulting decline in output, and hence income, will depress supply more than demand and thus eventually bring the economy to equilibrium at Y_0. Or, in terms of the equivalent savings=investment equilibrium condition, the decline in income will decrease savings and thus eventually eliminate the excess of savings over investment that exists at Y_1. In Keynes's words,

> The novelty in my treatment of saving and investment consists, not in my maintaining their necessary aggregate equality, but in the proposition that it is, not the rate of interest, but the level of incomes which (in conjunction with certain other factors) ensures this equality (1937, p. 211; cf. also *GT*, p. 31, lines 16–23; p. 179, lines 2–6).

In more formal terms (which Keynes himself did not use), the theory of effective demand is concerned not only with the mathematical solution of the equilibrium equation $F(Y) = Y$, but with demonstrating the stability of this equilibrium as determined by the dynamic adjustment equation $dY/dt = G[F(Y) - Y]$, where $G' > 0$.

Figure 1

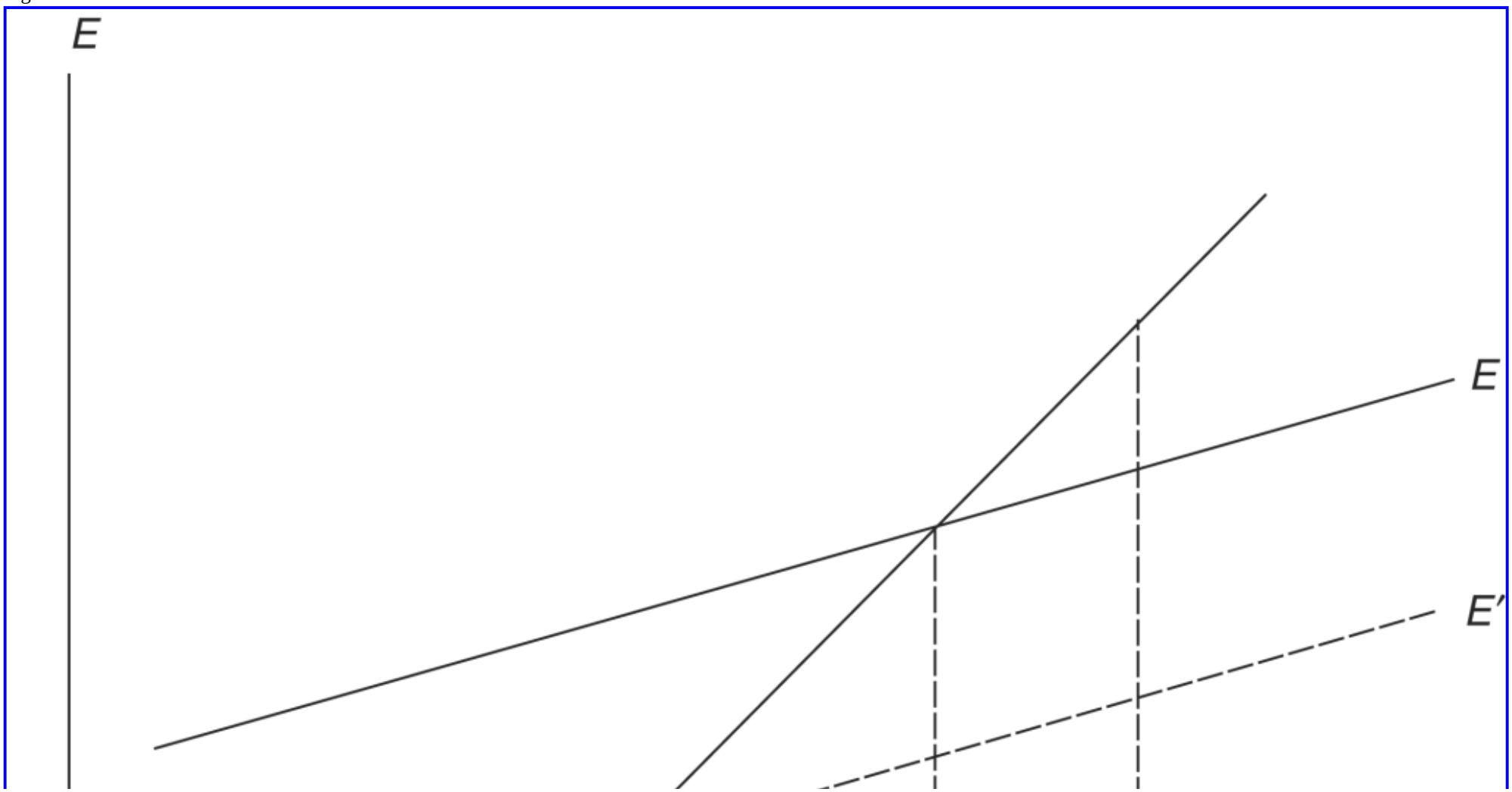

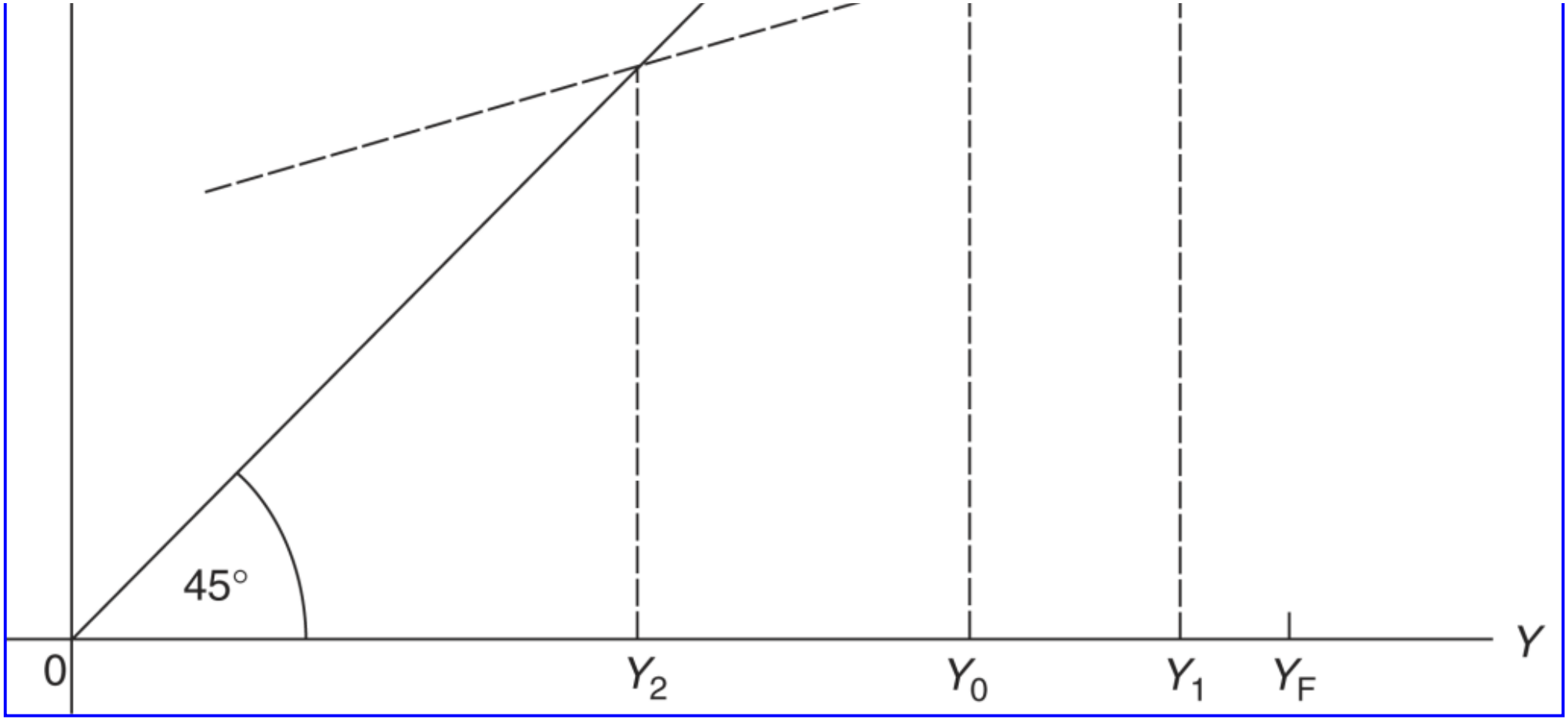

Correspondingly, as Keynes emphasizes in his letter to Harrod and elsewhere, a crucial assumption of his (Keynes's) analysis is that the marginal propensity to consume is less than unity, which in turn implies that the marginal propensity to save is greater than zero. For, if the marginal propensity to consume were equal to unity, no equilibrating mechanism would be activated by the decline in output. Specifically, as income (output) decreased, spending would decrease by exactly the same amount, so that any initial difference between aggregate demand and supply would remain unchanged. Alternatively, as income decreased, the initial excess of desired saving over investment would remain unchanged. Thus the system would be unstable. This is the major novel feature of the *General Theory* and its central message: the theory of effective demand as a theory which depends on the equilibrating effect of the decline in output itself to explain why ‘the economic system may find itself in stable equilibrium with *N* [employment] at a level below full employment, namely at the level given by the intersection of the aggregate demand function with the aggregate supply function’ (*GT*, p. 30).

Since most economists today probably learned the theory of effective demand as just another chapter in their introductory course in economics, it may be difficult for them to conceive of the intellectual shock wave that this theory created when Keynes first presented it. Testimony to this impact has, however, been given by many elders of our profession who (in Samuelson's words) were ‘born as economists prior to 1936’ (1946, p. 315). And though my ‘birthyear’ was about a decade after this date, I began my studies before the theory of effective demand had percolated down to the introductory course in the field. So I, too, can still remember how strange and even difficult it was during my later graduate studies to have to learn to think in terms of a demand for aggregate output as a whole – a demand that was in some way conceptually different from actual aggregate income, as if national income expended could somehow differ from national income received!

Similarly, under the influence of Marshall's *Principles* (which was then still being used as a textbook), it had been thoroughly ingrained into us that the demand function for a good could be defined only under the assumption of ‘ceteris paribus’. Indeed, in order to insure that this assumption was fulfilled in practice, the more punctilious economists of those days were only willing to speak of the demand function for a good the total expenditure on which was small, so that variations in these expenditures as price varied would not significantly affect the ‘marginal utility of money’ (i.e. the marginal utility of money expenditures: see ibid., Bk. III, chs iii and vi). How then could one validly speak of a demand function for the aggregate of all goods? How was it possible for ‘other things to be held constant’ in such a case?

(The foregoing diagram does not appear in the *General Theory* – a fact which has in recent years led certain circles to contend that it does not represent Keynes's theory. This,

however, is an invalid inference: for with one exception, Keynes did not use analytical diagrams in any of his writings. And that one exception is the diagram which appears on p. 180 of the *General Theory* – a diagram which in the accompanying footnote, Keynes attributes to Harrod. Furthermore, in his later 'How to Pay for the War' (1940; *JMK* IX, pp. 416–17), Keynes analysed the expected inflationary gap in Britain by means of the C+I=Y rubric, which is of course the arithmetical counterpart of the 45° diagram. See also section 5 below for a conjecture about why Keynes presented his theory of effective demand in terms of the level of employment, and not of national income, as in the diagram.)

Needless to say, there are other interpretations of the novelty and central message of the *General Theory*. The preceding and following discussions implicitly (and sometimes explicitly) explain why I do not accept some of the leading ones: namely, the interpretations which contend that this message is the analysis of an economy caught in the 'liquidity trap' (Hicks, 1937) and/or one in which money wages are completely inflexible downwards (Modigliani, 1944); that it is the proposition that unemployment is caused by the inadequacy of aggregate demand; that it is the analysis of the way expectations are formed and influence behaviour in an uncertain world whose uncertainty is not subject to the probability calculus (Shackle, 1967, ch. 11; Davidson, 1972); that it is the multiplier; that it is the crucial role of fluctuating investment in generating business cycles; that it is the theory of effective demand (and particularly of the aggregate supply function) as a determinant of the wage and price levels (Weintraub, 1961); and that it is the advocacy of public works as a means of combatting unemployment (the implicit interpretation of various writers who have regarded such advocacy as an anticipation of the *General Theory*; cf. e.g., Garvy, 1975 and Backhaus, 1985). Insofar as Leijonhufvud (1968) is concerned, he himself has subsequently admitted that his book was about 'theoretical problems that were current problems in the early or mid-sixties … . What Keynes might have meant etc. was not one of the problems. Doctrine history was not what the book was about' (Leijonhufvud, 1978). (For further details, see Patinkin, 1976a, pp. 141–2; 1982, pp. 5–7, 84 fn.8, 153–8; 1984, pp. 101–2.)

To bring out the central message of the *General Theory* more sharply, let me contrast Keynes's discussion in this book with the corresponding one of the *Treatise*. In the *General Theory*, a decrease in consumption – or, equivalently, an increase in savings – is represented by a downward shift of the aggregate-demand curve in Figure 1 to E'; the resulting decline in output will then cause a corresponding decline in the amount consumed – and hence in the amount saved – until a new equilibrium is necessarily reached at Y_2 (cf. *GT*, pp. 82–5, 183–4). Contrast this with Keynes's 'parable' in the *Treatise* of a simple 'banana plantation' economy in an initial position of full-employment equilibrium which is disturbed because (in Keynes's words) 'into this Eden there enters a thrift campaign'. Making use of the analytical framework of the *Treatise*, Keynes explains that the resulting increased savings, unmatched by increased investment, will cause entrepreneurs to suffer losses (i.e. $Q = I - S < 0$) and they

> will seek to protect themselves by throwing their employees out of work or reducing their wages. But even this will not improve their position, since the spending power of the public will be reduced by just as much as the aggregate costs of production. By however much entrepreneurs reduce wages and however many of their employees they throw out of work, they will continue to make losses so long as the community continues to save in excess of new investment. Thus there will be no position of equilibrium until either (a) all production ceases and the entire population starves to death, or (b) the thrift campaign is called off or peters out as a result of the growing poverty; or (c) investment is stimulated by some means or other so that its cost no longer lags behind the rate of saving (*TM* I, pp. 159–60).

In brief, it seems to me that – to make anachronistic use of a concept of the *General Theory* – Keynes is implicitly assuming here that the marginal propensity to spend is unity, so that a decline in output cannot reduce the excess of saving over investment and thus cannot act as an equilibrating force. Instead, the decline in output continues indefinitely; or alternatively, the decline might end as the result of some exogenous force that closes the gap between saving and investment – 'the thrift campaign is called off', or 'investment is stimulated by some means or another'. In brief, none of these alternatives indicates that Keynes of the *Treatise* understood that the decline in output itself acts directly as a systematic endogenous equilibrating force.

4. The foregoing is the essence of the theory of effective demand as presented in 'Book I: Introduction' of the *General Theory* under the explicit simplifying assumptions of a constant level of investment (which presupposes a constant rate of interest) and a constant money wage-rate (*GT*, pp. 27–9). (For deficiencies in this presentation – and particularly in that of the aggregate supply function – stemming primarily from Keynes's failure to apply the marginal concept correctly, see Patinkin, 1982, pp. 142–57. In this connection it should be noted that according to Joan Robinson's own testimony (1969, p. xi), 'Keynes was not much interested in the theory of imperfect competition' that she was developing in the early 1930s, and in which marginal analysis played a central role (J. Robinson, 1933a). See also the similar statement by Austin Robinson in Patinkin and Leith, 1977, p. 79.) After a 'digression' from the 'main theme' (*GT*, p. 37) in 'Book II: Definitions and Ideas' for the purpose of clarifying various concepts, Keynes then devotes most of the reminder of the book to an elaboration of the theory of effective demand which (inter alia) is free of these restrictive assumptions.

In 'Book III: The Propensity to Consume' he elaborates upon the determinants of the consumption component of aggregate demand and also discusses the related multiplier (*GT*, pp. 114–15), referring in this context to the 1931 article of his former student, Kahn. (This article was actually the successful outcome of Kahn's efforts – with his mentor's encouragement – to provide a precise formula for measuring the 'indirect effects' of an increase in government expenditures, effects which Keynes in his 1929 election pamphlet *Can Lloyd George Do It?*, had described as of 'immense importance', but impossible of measurement 'with any sort of precision' (*JMK* IX, pp. 106–7; cf. Howson and Winch, 1977, pp. 48–9; Patinkin, 1978)).

In 'Book IV: The Inducement to Invest', Keynes drops the assumption of a constant level of investment and explains how this level is determined by the marginal-efficiency-of-

capital schedule in conjunction with the rate of interest, which rate is determined in turn by the liquidity-preference schedule in conjunction with the quantity of money. I might note that Keynes's liquidity-preference function $-M = L_1(Y) + L_2(r)$, where *M* and *Y* respectively represent nominal money and nominal income (*GT*, p. 199) – actually (though in all probability, inadvertently) reflects money illusion (see Patinkin 1956 and 1965, chapter XI:1 and Supplementary Note K:2).

Chapter 12 (‘The State of Long-Term Expectations’) elaborates upon the argument of Book II, chapter 5 (‘Expectations as Determining Output and Employment’). The crucial influence of uncertainties on both the aforementioned schedules – and hence the necessity to make decisions with respect to them on the basis of expectations – is emphasized. As Samuelson (1946, p. 320) has however noted, Keynes's discussion ‘paves the way for a theory of expectations, but it hardly provides one’ (see also the detailed critique by Hart, 1947). In any event, Keynes emphasizes that the uncertainties in question are not subject to a probability calculus, so that long-run investment decisions in particular may instead be the result of ‘animal spirits’ (*GT*, p. 161; see also Keynes's 1937 *QJE* article as reproduced in *JMK* XIV, p. 114). (The distinction between risk, which is subject to such a calculus, and uncertainty which is not, was the major point of Knight's classic 1921 work on *Risk, Uncertainty, and Profit*; there may also be a hint of this distinction in chapter 6 of Keynes's *Treatise on Probability*, published the same year, to which Keynes refers (*GT*, p. 148, n.1); see also Lawson and Pesaran, 1985.) These uncertainties are a major source of the effectively low interest-elasticity of the first of these schedules, as well as the source of the speculative demand for money, and hence the effectively high (though not infinite) interest-elasticity of the second of them. (Keynes does not always distinguish between a movement along a demand curve and a shift of the curve itself, and it is the combined result of these two changes that I denote by ‘effective elasticity’.)

Thus the many interpretations to the contrary notwithstanding, Keynes did not base his theory on the so-called ‘liquidity trap’. In his words, ‘whilst this limiting case might become practically important in future, I know of no example of its hitherto’ (*GT*, p. 207. See also Keynes's brief description of the way in which, after Britain abandoned the gold standard in 1931 (see concluding paragraph of section 2 above), the monetary authorities had succeeded in gradually driving down the rate of interest. But see Patinkin, 1976a, pp. 111–13 for some indications of ambivalence in the *General Theory* about the relevance of the ‘liquidity trap’.) It is because of these elasticities that monetary policy may well be inadequate to the task of eliminating unemployment: for an increase in the quantity of money will not significantly reduce the rate of interest; and to the extent that there is such a reduction, it will not generate a significant increase in investment and hence in aggregate demand (cf. *GT*, pp. 164, 168–70). Book IV also includes chapter 17 on ‘The Essential Properties of Interest and Money’, with all of its confusions and obscurities (see Lerner, 1952; see also Hart, 1947, p. 416 and Hansen, 1953, p. 159).

Keynes concludes Book IV with a summary chapter (18) entitled ‘The General Theory of Employment Re-Stated’. In substance, though not in form, and certainly not with intent (see section 9 below and Patinkin, 1976a, pp. 98–100), this chapter (like the diagram on p. 180 of chapter 14) provides a general equilibrium analysis of the determination (as of a given money-wage rate and nominal quantity of money) of the equilibrium level of national income by the interactions between the commodity (consumption- and investment-goods) and money markets (*GT*, pp. 246–7). Thus a basic contribution of the *General Theory* is that it is in effect the first practical application of the Walrasian theory of general equilibrium: ‘practical’, not in the sense of empirical (though the *General Theory* did provide a major impetus to empirical work), but in the sense of reducing Walras's formal model of *n* simultaneous equations in *n* unknowns to a manageable model from which implications for the real world could be drawn. Furthermore, like Walras's model in the *Eléments* (1926, lessons 29–30), Keynes's model in the *General Theory* is one that integrates the real and monetary sectors of the economy. It is this general-equilibrium aspect of the *General Theory* that Hicks (1937) was subsequently to develop and formalize in his influential IS–LM interpretation of the book – with respect to which Keynes wrote him that ‘I found it very interesting and really have next to nothing to say by way of criticism’ (*JMK* XIV, p. 79).

Finally, in ‘Book V: Money-Wages and Prices’, Keynes drops the assumption of a constant money-wage rate and applies the theory of effective demand that he had developed in Books I–IV to an analysis (in the first chapter of this Book, ‘Chapter 19: Changes in Money Wages’) of the effects of a decline in this rate. It should be emphasized that Keynes regarded such a decline not as an abstract theoretical possibility, but as what had actually happened to money wages in the years immediately preceding the *General Theory*. Thus from 1925–33, money wages had declined in Britain by 7 per cent, whereas in the United States they had declined over the much shorter period 1929–33 by 28 per cent (sic!) (see Keynes's allusion to the former on p. 276 of the *General Theory*, and to the latter on p. 9; on the sources of the above data, see Patinkin, 1976a, pp. 17 and 121). During these periods, however, real wages in both countries actually rose, which was the background of Keynes's oft-cited enigmatic statement (to which I shall return below) that ‘there may exist no expedient by which labour as a whole can reduce its *real* wage to a given figure by making revised *money* bargains with the entrepreneurs’ (*GT*, p. 13, italics in original).

Keynes's basic argument in chapter 19 is that a decline in money wages (which in practice would, because of the resistance of workers, take place only very slowly: *GT*, p. 267; see also ibid., pp. 9, 251, 303) can increase the level of employment only by first increasing the level of effective demand; that the primary way it can generate such an increase is through its effect in increasing the quantity of money in terms of wage units, thereby decreasing the rate of interest and stimulating investment; that accordingly the policy of attempting to eliminate unemployment by reducing money wages is equivalent to a policy of attempting to do so by increasing the quantity of money at an unchanged wage rate and is accordingly subject to the limitations as the latter; namely, that a moderate change ‘may exert an inadequate influence over the long-term rate of interest’, while an immoderate one (‘even if it were practicable’) ‘may offset its other advantages by its disturbing effect on confidence’ (*GT*, pp. 266–7).

Indeed, the possible adverse effect on confidence is greater in the case of a wage (and price) decline than in that of a monetary expansion, and this for two reasons: first, the decline may create the expectation of still further declines, thus leading firms to postpone carrying out any decision to increase their demand for labour; second, ‘if the fall of wages and prices goes far, the embarrassment of those entrepreneurs who are heavily indebted may soon reach the point of insolvency – with severely adverse effects on investment’ (*GT*, p.

264). This adverse effect will be reinforced by the fact that the ‘expectation that wages are going to sag by, say, 2 per cent in the coming year will be roughly equivalent to the effect of a rise of 2 per cent in the amount of interest payable for the same period’ (*GT*, p. 265). (This use of what is essentially Fisher's distinction between the real and nominal interest rates is somewhat inconsistent with reservations that Keynes expressed about it earlier in the *General Theory*, pp. 141–3.) Hence Keynes's major conclusion – and indeed the negative component of his central message – that ‘the economic system cannot be made self-adjusting along these lines’ (*GT*, p. 267). In this way Keynes finally supplies the theoretical basis for his claim in chapter 2 of ‘Book I: Introduction’ that, contrary to the ‘classical’ view, ‘a willingness on the part of labour to accept lower money-wages is not necessarily a remedy for unemployment’ – a claim he had promised would be ‘fully elucidated … in Chapter 19’ (*GT*, p. 18).

The analysis of chapter 19, together with Keynes's acceptance in chapter 2 of the ‘classical postulate’ that ‘the wage is equal to the marginal product of labour’ (*GT*, p. 5), enables us to understand the enigmatic statement cited three paragraphs above. Specifically, if the effect of a decline in the money wage rate on the level of effective demand, hence output, and hence employment is indeterminate, then so too is its effect on the marginal product of labour and hence real wages. Thus Keynes's statement is simply a reflection of his basic view that

> the propensity to consume and the rate of new investment determine between them the volume of employment, and the volume of employment is uniquely related to a given level of real wages – not the other way round (*GT*, p. 30).

And since Keynes also accepts the classical law of diminishing returns (*GT*, p. 17), he contends that if a sharp decline in money wages should generate only a slight increase in the level of employment – hence only a slight decrease in the real wage rate – then it must also generate a sharp (though proportionately smaller) decrease in the price level (however, Keynes never explains the dynamic market forces that bring this about; see the discussion below of chapter 21). In Keynes's words at the end of chapter 19:

> It follows, therefore, that if labour were to respond to conditions of gradually diminishing employment by offering its services at a gradually diminishing money-wage, this would not, as a rule, have the effect of reducing real wages and might even have the effect of increasing them, through its adverse influence on the volume of output. The chief result of this policy would be to cause a great instability of prices, so violent perhaps as to make business calculations fultile in an economic society functioning after the manner of that in which we live (*GT*, p. 269).

Accordingly, Keynes concludes chapter 19 with the policy recommendation that ‘the money-wage level as a whole should be maintained as stable as possible, at any rate in the short period’ (*GT*, p. 270).

This is an appropriate point to note that though in Book III, Keynes take account of what might be called the capital-gains effect on consumption (*GT*, pp. 92–4), he does not do so with reference to the wealth effect as such, and in particular does not do so with reference to the real-balance component of this effect. Correspondingly, his analysis in chapter 19 does not take account of the positive real-balance effect generated by a wage and price decline. But since the operation of this effect in this deflationary context suffers from the same limitations described in this chapter, I do not believe that taking account of it would have affected Keynes's basic conclusion about the inefficacy of a wage decline as a means of increasing employment (Patinkin, 1951, pp. 272–8; 1956, pp. 234–7; 1965, pp. 336–40; 1976a, pp. 110–11).

Thus chapter 19 is the climax of the *General Theory*. And it is clear from it that, the many contentions to the contrary notwithstanding, the analysis of this book does not depend on the assumption of absolutely rigid money wages. What is, however, true is that, because of the aforementioned adverse effects of flexibility, the relative stability of money wages is the concluding policy recommendation of the chapter. I must also emphasize that were the *General Theory* to depend on the assumption of wage rigidity, there would be no novelty to its message: for the fact that such a rigidity can generate unemployment was a commonplace of classical economics. Needless to say, this does not mean that Keynes went to the opposite extreme of assuming wages to be perfectly flexible. Instead, his view of the real world was that ‘moderate changes in employment are not associated with very great changes in money-wages’ (*GT*, p. 251). At the same time, Keynes emphasizes that there exists an ‘asymmetry’ between the respective degrees of upward and downward wage flexibility: that, in particular, ‘workers are disposed to resist a reduction in their money-rewards, and that there is no corresponding motive to resist an increase’ (*GT*, p. 303).

I might note that Keynes's lack of faith in the efficacy of the market-equilibrium process in a macroeconomic context also manifests itself in such earlier writings as *The Economic Consequences of Mr Churchill* (1925; *JMK* IX, pp. 227–9 *et passim*) and the *Treatise* (I, pp. 141, 151, 244–5, 265). Nor (I conjecture) would Keynes have been impressed by the contention of some exponents of the ‘new classical macroeconomics’ that the market would not permit a situation of unemployment to persist because contracts could then be made which would make everyone better off. Indeed, I would conjecture that, as one who had seen how the most civilized countries of the world had engaged for four long years of stalemated trench warfare in the mutual slaughter of the best of their young men, Keynes was not predisposed to believe in natural forces that always brought agents to generate a mutually beneficial situation. Because of the uncertainty of how other react to our actions, the actual world for Keynes was one that – in a macroeconomic context – could readily lead to the ‘globally irrational’ results of the prisoner's dilemma; not to the rational results of the Walrasian auctioneer.

Book V also contains ‘Chapter 21: The Theory of Prices’. In ‘Book I: Introduction’, Keynes had stated that ‘we shall find that the Theory of Prices falls into its proper place as a matter which is subsidiary to our general theory’ (*GT*, p. 32). In particular, as already noted, the level of effective demand determines the level of employment, hence the marginal

productivity of labour, and hence the real wage rate; for any given money wage rate, then, the price level is determined. In the words of chapter 21, ‘The general price-level (taking equipment and technique as given) depends partly on the wage-unit [i.e., on the money wage rate] and partly on the volume of employment’ (*GT*, p. 295). It should again be noted that Keynes's discussion here is completely mechanical and provides no explanation of the dynamic market forces that cause the price level to change as a consequence of a change in money wages.

Chapter 21 also includes a discussion of the quantity theory of money. In the *Treatise*, as noted above, Keynes regarded this theory to be deficient only because of the absence of a dynamic analysis – which he then supplied. In the *General Theory*, however, Keynes saw himself as providing a new theory that replaced the quantity theory entirely. For, he claimed, the quantity theory holds only on two unrealistic conditions: first, that the speculative demand for money ‘will always be zero in equilibrium’ (actually, this is not a necessary condition; see Patinkin 1956 and 1965, ch. XII:1); second, that the level of output is constant at full employment (*GT*, pp. 208–9). Thus Keynes may well have regarded the *General Theory* as the culminating chapter in The Saga of Man's Struggle for Freedom from the Quantity Theory. Indeed, in his preface to the French edition of the *General Theory*, Keynes wrote that ‘the following analysis [of money and prices] registers my final escape from the confusions of the Quantity Theory, which once entangled me’ (*JMK* VII, p. xxxiv).

The last Book of the *General Theory* – ‘Book VI: Short Notes Suggested by the General Theory’ – is, as its title indicates, essentially an appendage to it, one that could have been omitted without affecting the logical integrity of the book as a whole. The Book begins with ‘Chapter 22: Notes on the Trade Cycle’. Here Keynes contends that the cycle is generated by changes in the marginal efficiency of capital – which changes, for reasons discussed in this chapter, ‘have had cyclical characteristics’. He claims no novelty for this interpretation (‘these reasons are by no means unfamiliar either in themselves or as explanations of the trade cycle’) and explains that the purpose of the chapter is ‘to link [these reasons] up with the preceding theory’ (*GT*, pp. 314–15). Chapter 23 is entitled ‘Notes on Mercantilism, the Usury Laws, Stamped Money and Theories of Under-Consumption’ – whose omnibus title is a further indication that the material of Book VI is not an integral part of the book. The last chapter of the Book – and of the *General Theory* as a whole – is ‘Chapter 24: Concluding Notes on the Social Philosophy towards Which the General Theory Might Lead’. Only to a minor extent, however, is this chapter concerned with the question of short-run, full-employment policy – and in this context Keynes reiterates his scepticism of sole reliance on monetary policy and his corresponding belief ‘that a somewhat comprehensive socialisation of investment will prove the only means of securing an approximation to full employment’ (*GT*, p. 378). Most of the chapter is devoted to the long-run implications of a successful full-employment policy for the accumulation of capital, hence the rate of interest and the distribution of income; for the future of laissez-faire versus state socialism; and for the prospects of war and peace.

In chapter 24, Keynes also expresses his belief in the efficacy of the market mechanism, once the ‘socialisation of investment’ has assured the maintenance of full employment. Under these conditions, says Keynes,

> there is no objection to be raised against the classical analysis of the manner in which private self-interest will determine what in particular is produced, in what proportions the factors of production will be combined to produce it, and how the value of the final product will be distributed between them. Again, if we have dealt otherwise with the problem of thrift, there is no objection to be raised against the modern classical theory as to the degree of consilience between private and public advantage in conditions of perfect and imperfect competition respectively. Thus, apart from the necessity of central controls to bring about an adjustment between the propensity to consume and the inducement to invest, there is no more reason to socialise economic life than there was before (*GT*, pp. 378–9).

(In a similar way, Keynes was to argue in his posthumously published article on ‘The Balance of Payments in the United States’ (1946) that it was important to establish a framework for international trade and finance ‘which allows the classical medicine to do its work’ in establishing equilibrium in this context (*JMK* XXVII, pp. 444–5; see also Cairncross, 1978. But see Keynes's 1926 essay on ‘The End of Laissez-Faire’ (reproduced in *JMK* IX, pp. 272–94) for some reservations à la Knight's classic 1923 paper on ‘The Ethics of Competition’ about the workings of the market economy.)

5. From the foregoing it is clear that the primary concern of the *General Theory* is theory and not policy, though Keynes does make brief use of the theory to explain the necessity for public-works expenditures to combat severe unemployment; that the primary concern of its theory is output (or employment) and not prices; and that the primary concern of its theory of output is the explanation of equilibrium at less-than-full-employment and not cyclical variations in output.

Another point which is clear from this summary is that Keynes's repeated use of the term ‘unemployment equilibrium’ (*GT*, pp. 28, 30, 242–3, 249) in the first 18 chapters of the *General Theory* must, strictly speaking, be understood as referring to a Marshallian short-period equilibrium (*Principles*, Book V, ch. v) that is attained under the provisional assumption of a constant money-wage rate (*GT*, pp. 27, 247). Clearly, such an equilibrium no longer obtains once Keynes drops this assumption in the climactic chapter 19, proceeds to analyse the effects on the economy of a decline in the money wage rate, and shows that such a decline will not necessarily lead to an increase in employment and *a fortiori* not to the establishment of full-employment equilibrium (see above). Thus in the strict sense of the term, the *General Theory* is a theory of unemployment *dis*equilibrium: it analyses the workings of an economy in which money wages and hence the rate of interest may be slowly falling, but in which ‘chronic unemployment’ (*GT*, p. 249) nevertheless continues to prevail, albeit with an intensity that may be changing over time (cf. Patinkin, 1951, part III; 1956, chs XIII:1, XIV:1, and Supplementary Note K:3, reproduced unchanged in the 1965 edition; 1976a, pp. 113–19).

This interpretation would seem to be in contradiction to Keynes's emphasis that one of his major accomplishments in this book was to have demonstrated the possible existence of

‘unemployment equilibrium’ (*GT*, pp. 30, 242–3). I would like to suggest that the answer lies in a letter that Keynes wrote to Roy Harrod in August 1935, in reply to the latter's criticism that Keynes's discussions of the classical position were carried out in an unduly polemical style that exaggerated the differences between the two positions. In Keynes's words:

> the general effect of your reaction … is to make me feel that my assault on the classical school ought to be intensified rather than abated. My motive is, of course, not in order to get read. But it may be needed in order to get understood. I am frightfully afraid of the tendency, of which I see some signs in you, to appear to accept my constructive part and to find some accommodation between this and deeply cherished views which would in fact only be possible if my constructive part has been partially misunderstood. That is to say, I expect a great deal of what I write to be water off a duck's back. I am certain that it will be water off a duck's back unless I am sufficiently strong in my criticism to force the classicals to make rejoinders. I *want*, so to speak, to raise a dust; because it is only out of the controversy that will arise that what I am saying will get understood (*JMK* XIII, p. 548; italics in original).

And what could ‘raise more dust’ than a seemingly frontal attack on the ‘deeply cherished’ classical proposition that there could not exist a state of unemployment equilibrium? Conversely, what could be more easily ‘accommodated’ within the classical framework than the statement that a sharp decline in aggregate demand would, despite the resulting decline in the wage-unit, generate a prolonged period of disequilibrium which would be marked by a continuous state of unemployment?

It also seems to me that it is precisely the attempt to interpret the *General Theory* as presenting a theory of unemployment equilibrium in the fullest sense of the term that has led to its interpretation (despite the internal evidence to the contrary, and despite the facts to the contrary that existed at the time that the book was being written) as being based on the special assumptions of absolutely rigid money wages and/or the ‘liquidity trap’. For by definition there cannot be a state of long-run unemployment equilibrium in the sense that nothing in the system tends to change unless wages are rigid. Alternatively, if money wages are not rigid, then a necessary condition for equilibrium – in the sense of the level of employment remaining constant over time – is that the rate of interest remain constant; and a necessary condition for the rate of interest to remain constant in the face of an ever-declining money-wage and hence ever-increasing real quantity of money is that the economy be caught in the ‘liquidity trap’. Correspondingly, once we recognize that the *General Theory* is concerned, strictly speaking, with a situation of unemployment disequilibrium, we also understand that the validity of its analysis does not depend on the existence of either one of these special assumptions.

Three further observations about the *General Theory*: First, I have already noted that the exposition of the theory of effective demand in Book I is carried out, not in terms of national income – to which concept Keynes even expresses what he regards as methodological objections (*GT*, pp. 38, 40) – but in terms of the level of employment. In part, this was undoubtedly due to the fact that the level of employment was indeed his major concern. But I also feel that this provides an instructive instance in our discipline of a basic characteristic of the physical sciences: namely, the relationship between the development of theory and the development of tools of measurement. In particular, I conjecture that Keynes's ambivalence toward the use of the national-income concept in the *General Theory* (for he did make use of it in his chapters on the consumption function (ch. 10) and liquidity-preference function (ch. 15), respectively) was not unrelated to the fact that at the time national-income estimates had not yet become the household concept they are today; indeed there did not then even exist current official estimates of British national income. In contrast, ever since the early 1920s, estimates of British employment – or rather unemployment, as measured by the ‘Number of Insured Persons Recorded as Unemployed’ – were being published monthly in the *Ministry of Labour Gazette*. Similarly, I conjecture that the change in Keynes's view as manifested in his 1940 *How to Pay for the War* (*JMK* IX, pp. 416–17, 429; see the discussion of Figure 1 in section 3 above) – and his willingness (albeit with reservations) to make use in it of Colin Clark's national-income estimates, about which he had earlier expressed much scepticism – reflected in part the exigencies of wartime, and in part the increased respectability and acceptability of national-income estimates as a result of their publication (based on the work of Simon Kuznets) on an official, current annual basis by the United States beginning with 1935 (cf. Patinkin 1976b, pp. 129–30, 243–5, 248–54; cf. also the discussion of Keynes and national-income statistics in section 7 below).

Second, in the *General Theory*, Keynes also appears as a historian of economic thought. Thus chapter 2 is entitled ‘The Postulates of the Classical Economics’ and references to ‘classical theory’ are strewn throughout the book. Similarly, most of chapter 23 is devoted to his ‘Notes on Mercantilism’, which are largely based on Heckscher's (1935) classic work. In a comment thirty-odd years later on his 1936 review of the *General Theory*, Viner, (1964, p. 254) – who had in 1930 published what was essentially a monograph on mercantilism (reprinted in Viner, 1937, chs 1–2; see ibid., p. xiv) – explained that the terms of reference of his original review did not include the doctrinal aspects of the book, and went on to express reservations about the ‘objectivity and judiciousness’ of Keynes ‘as a historian of thought in areas in which he was emotionally involved as a protagonist and prophet’. Viner did not specify the areas he had in mind, but Heckscher (1946) explicitly referred to Keynes's treatment of mercantilsm and charged him with citing from his (Heckscher's) work ‘only … those parts of mercantilist theory that happen to coincide with his own analysis of economic behaviour’ (ibid., p. 340; actually, most of Heckscher's article is devoted to a criticism of Keynes's theory itself). However, Hutchison (1978, pp. 127–35) and Walker (1986, part IV), basing themselves on more recent studies of mercantilism and its period, have largely supported Keynes's treatment, particularly with respect to his emphasis on the mercantilists’ concern with the problem of unemployment, and his corresponding contention that they advocated a positive balance of trade and resulting inflow of gold not as a fetish, but as a rational means of dealing with this problem (*GT*, pp. 346–48). But Hutchison (1978, p. 128) also cites Blaug's (1962, p. 15; 1964, pp. 114–15) dissenting opinion, and Walker (1986, p. 28) notes that Keynes was nevertheless guilty

of ‘excessively broad generalizations’ about the mercantilist literature.

Insofar as Keynes's treatment of ‘classical economics’ is concerned, both Hutchison and Walker conclude that Keynes's discussion of Ricardo and Say's Law, on the one hand, and Malthus's concern with the possibility of the inadequacy of aggregate demand, on the other (*GT*, pp. 18–21, 32–4) constitute important contributions to the history of economic thought, though here too they indicate some inaccuracies (see also Patinkin, 1956 and 1965, Supplementary Note L, on Keynes's misrepresentation of the passage in Mill's discussion of Say's Law which Keynes cites on p. 18 of the *General Theory*). At the same time, both Hutchison and Walker reject Keynes's contention that classical economics in this sense continued unchallenged through the second half of the 19th century on into the 20th. In particular, Hutchison (1978, pp. 165–6, 175–99) conclusively shows that Keynes was not justified in including Pigou among the ‘classical economists’ (*GT*, p. 3, n1; see also Corry, 1978, pp. 8–11; see also Walker, 1985, for a favourable view (though it too with some reservations) of Keynes as a historian of thought in his 1933 *Essays in Biography*).

In sum, though Keynes in the *General Theory* provided valuable and stimulating insights with respect to certain points in the history of economic thought, Viner did not err in saying that the balanced scholarly treatment of this subject was not Keynes's forte (cf. also Hutchison, 1978, p. 173 and Walker, 1986, p. 29).

My third and last observation is that in order to understand why the *General Theory* had such a revolutionary impact on the profession – and indeed on the general public – we must take account of the circumstances that prevailed when it burst on the scene. In the early 1930s, the Western world was desperately searching for an explanation of the bewildering and seemingly endless depression that was creating untold misery for millions of unemployed and even threatening the viability of its democratic institutions. Indeed, largely as a result of the widespread social unrest caused by the mass unemployment, a totalitarian government had already taken power in Italy and a far more evil and oppressive one was doing so in Germany. And the appearance of the *General Theory* in 1936 offered not only an explanation, but also a confident and theoretically supported prescription for ending depressions within a democratic framework by proper government policies. Thus the *General Theory* provided an answer not only to a theoretical problem, but to a burning political and social one as well. I might also add that the fact that the theoretical revolution embodied in the Keynes's *General Theory* took place concurrently with the Colin Clark–Simon Kuznets revolution in national-income measurement further increased its impact on the profession: for those measurements made possible the quantification of the analytical categories of the *General Theory*, hence the empirical estimation of its functional relationships, and hence its application to policy problems (cf. Patinkin, 1976b).

Despite the many criticisms and discussions of the *General Theory* that followed its publication (cf. e.g., the review articles by Harrod, Hicks, Leontief, Lerner, Meade, Pigou, Viner et al. reprinted in Lekachman, 1964 and Wood, 1983), its basic analytical structure not only remained intact, but also defined the research programme for both theoretical and empirical macroeconomics for the following three decades and more. Truly a scientific achievement of the first order. And as with the passage of time we gain a more critical view of the accomplishments – and deficiencies – of ‘monetarism’ and of ‘the new classical macroeconomics’ of the last two decades, an appropriately modified Keynesian model that will take advantage of what we have learned from these developments may yet regain its place as the leading one for macroeconomic analysis (Howitt, 1986; for some conjectures about what Keynes might have thought of these developments, see Patinkin, 1984).

6. Any great work brings in its wake claims of priority for other writers – and the *General Theory* was no exception. Thus within a year after its publication, Bertil Ohlin (1937) claimed that there were ‘surprising similarities’ between the analysis in this book and that which had been developed in the writings (in Swedish) of what he called the ‘Stockholm school’, under which rubric he included Erik Lindahl and Gunnar Myrdal as well as himself. Similarly, in a review article on the *General Theory*, the Polish economist Michal Kalecki (1936) claimed that he had anticipated its main arguments in a 1933 monograph in Polish on the business cycle (the ‘essential part’ of which was published many years later in English translation in Kalecki, 1966, pp. 1, 3–16). Ohlin's claim was presented in the *Economic Journal*, then the leading journal of the economics profession, and gained immediate attention – so much so that the claim of the Stockholm School became a ‘perennial of doctrinal history’ (in Gustafsson's, 1973, apt phrase). In contrast, Kalecki's claim was published in *Ekonomista* – the professional journal of Poland's economists, published, of course, in their own language – and so received no attention outside that country. (An English translation of this review has only recently been published; see Targetti and Kinda-Hass, 1982.) Fifteen years later, however, the claim of Kalecki was brought to the attention of the profession as a whole by Lawrence Klein (1951) and Joan Robinson (1952), and has in certain quarters received increasing support ever since.

A detailed examination of these claims, however, has led me to reject them on the grounds that the respective central messages of these writers were different from that of the *General Theory* (Patinkin, 1982, chs 1–4). In particular, the central message of the Stockholm school (like that of Keynes's *Treatise*) was a further development of that of Wicksell, and had to do with the interrelationships of the rate of interest and prices, and only indirectly with output. And though Kalecki's central message had to do with output, its concern was not with the forces that generate equilibrium at low levels of output, but with the forces that generate cycles of investment and hence output: more specifically, not with the feedback mechanism of the *General Theory* that equilibrates planned saving and investment via declines in output, but with the cyclical behaviour of investment in a capitalist economy on the implicit assumption that there always exists equality between planned savings and investment. At the same time I must emphasize that in his primary concern with quantities as against prices; in his concentration on national-income magnitudes and functional relations among them; and in his corresponding emphasis on analysing the relationship between investment and other macroeconomic variables, Kalecki came significantly closer to the *General Theory* than did the Stockholm School, and this was particularly true of his semi-popular 1935 paper ‘The Mechanism of the Business Upswing’.

7. The foregoing discussion has highlighted the differences between the respective volumes of Keynes's trilogy. There are, however, also important similarities. Thus a common element of these books is their concern with practical policy problems, and their related concern with the empirical aspects of these problems. At the same time I must emphasize that

Keynes (like the great majority of his contemporaries) largely used empirical data for illustrative purposes, or at most as a basis for rather impressionistic observations about the relations between the variables described by the data. Though there are partial exceptions (see the second paragraph below), Keynes practically never carried out a systematic statistical analysis of empirical data as a basis for conclusions.

Thus, for example, Keynes's excellent presentation of the purchasing-power-parity theory in the *Tract* is supported by charts and diagrams showing the generally corresponding movements of the actual exchange rates of England, France and Italy with those respectively predicted by the theory (*Tract*, pp. 81–6). Similarly, Keynes's aforementioned analysis of inflation as a tax on real cash balances – and his explanation that this tax will decrease the volume of these balances that individuals will be willing to hold – is illustrated by data from the postwar hyperinflations of Germany, Austria, and Russia (*Tract*, pp. 45–6). Similarly, in the second, ‘applied’ volume of his *Treatise*, Keynes presents empirical estimates of the variables that play a key role in the theory he developed in the first volume: namely, the quantity of money, the velocity of circulation, the volume of working capital – and he even adds a long chapter (30) providing historical illustrations of his theory.

Though there is less emphasis on empirical data in the *General Theory*, it is noteworthy that Keynes was quick to make use in it (though somewhat carelessly; see the correspondence reproduced in *JMK* XXIX, pp. 187–206) of Simon Kuznets's (1934) preliminary estimates of net investment in the United States in order to illustrate his (Keynes's) basic contention about the critical role of wide fluctuations in this variable in generating business cycles (*GT*, pp. 102–5). What is even more noteworthy is Keynes's use of these data in order to make an empirical estimate (crude as it was) of the magnitude of the multiplier in the United States – and thence of the marginal propensity to consume of that country (*GT*, pp. 127–8). Thus Keynes not only made the marginal propensity to consume a central component of macroeconomic theory, but also provided the first estimate of its magnitude that was based on an examination of statistical time series!

I must, however, immediately add that there are many problematic aspects of this estimate, not least of which is the mystery of the source of the national-income data which Keynes used (together with Kuznets’ aforementioned data on investment) to estimate the multiplier. Furthermore, despite the fact that he was one of the founding members of the Econometric Society in 1933 and even served as its President during 1944–45, Keynes was actually extremely skeptical of econometric methods. Thus his oft-cited critical review (1939) of Tinbergen's classic work was devoted not to the much better known second volume of this study on *Business Cycles in the United States of America, 1919–1932* (1939), but to the first volume (published a few months earlier), *A Method and Its Application to Investment Activity*, in which Tinbergen set out and exemplified the principles of multiple-correlation analysis. Accordingly, the criticisms Keynes presented in this review were levelled not at Tinbergen's ambitious 46-equation model of the United States economy, but at the use of correlation analysis to estimate a regression for even a single equation! It should, however, be noted that though not all of Keynes's criticisms were well taken, some raised problems that continue to trouble econometricians: namely (though obviously not in the terms that Keynes used), the problems of specification bias and of simultaneous-equation bias (Patinkin, 1976b, sections 1, 3; cf. also Lawson and Pesaran, 1985).

Another aspect of Keynes's interest in the empirical aspects of our discipline was his concern with improving the scope and reliability of economic data. Thus in the course of presenting the aforementioned estimates in the second volume of the *Treatise*, Keynes repeatedly complains about the inadequacy of the data (*TM* II, pp. 78, 87). Keynes was also responsible for the final chapter in the Macmillan Report (1931), which was devoted to proposals for extending and improving available economic statistics in Britain. It is, however, noteworthy that these proposals did not include one for the construction of current national-income statistics. Similarly, in the years that followed, Keynes failed to support Colin Clark's pioneering work in this field (1932, 1937). It was only after the outbreak of World War II that this attitude changed, and then Keynes played an important role in promoting the publication of the famous 1941 White Paper, *Analysis of the Sources of War Finance and an Estimate of the National Income and Expenditure in 1938 and 1940* (Cmd. 6261), for which James Meade and Richard Stone were primarily responsible, and which marked the beginning of official British national-income statistics (Patinkin, 1976b, pp. 230–31, 244–5, 248–54).

Though all three of Keynes's books are concerned with policy issues, they nevertheless differ in the extent and sense of immediacy with which their policy discussions are presented. In view of the origin of the *Tract* in articles in the *Manchester Guardian*, it is not surprising that discussions on current policy issues are paramount in it. Indeed, having only a short time before dealt so successfully with prime ministers in his *Economic Consequences of the Peace* (1919, *JMK* II) and in his *Revision of the Treaty* (1922, *JMK* III), Keynes had no hesitations in dispensing advice on current problems directly from the pages of the Tract to the finance ministers (or their equivalent), not only of England and the United States, but also of Czechoslovakia (p. 120), Germany (pp. 50–52), and France (pp. xxi–xxii).

In contrast – as befits a comprehensive, scientific work – Keynes's policy recommendations of the *Treatise* are for the most part of a more general nature, though here too there are references to specific, immediate issues (e.g., *TM* II, pp. 270 ff, 348 ff). Least specific in its policy proposals, for reasons indicated in section 3 above, is the *General Theory*.

What were the policy problems that concerned Keynes? The major one was obviously unemployment. This had plagued Britain in the two years that preceded the publication of the *Tract* (1923) and it continued to be a serious problem in the five years that he was writing the *Treatise* (1930). In contrast, those were years of boom and prosperity in the US; and when in the early 1930s – the period of writing the *General Theory* (1936) – prosperity gave way to depression in the US as well, unemployment in Britain became even more serious. A common characteristic of all three of these books is Keynes's opposition to attempts to combat unemployment by reducing the nominal wage rate. However, it seems to me that there is a difference between the *Treatise* and the *General Theory* on this point: for my impression is that in the *Treatise*, Keynes believed that such a reduction could theoretically help but practically could not be carried out; whereas in the *General Theory*, he opposed it on theoretical grounds as well. In part this difference may have stemmed from the fact that Keynes in 1930 was writing under the influence of the relative inflexibility of British money wages in the years that had preceded, whereas in 1936 he also had before

him the United States experience of the sharp reduction in money wages during 1929–33 that had not succeeded in solving the unemployment problem (note again Keynes's allusion to this experience on p. 9 of the *General Theory*).

At the same time, a recurrent theme of Keynes's discussion of unemployment was that if by agreement or decree money wages could be instantaneously and uniformly reduced in all sectors of the economy, then the problem would be solved (cf. *Economic Consequences of Mr. Churchill*, 1925, *JMK* IX, pp. 211, 228–9; *TM* I, pp. 141, 151, 244–5, 265, and 281; *GT*, pp. 265, 267, and 269). For such an instantaneous reduction would be accomplished before it could create adverse expectations, and it would also not change relative wage rates as between workers in different industries (see *JMK* IX, p. 211 and *GT*, p. 14 for Keynes's emphasis on the resistance of workers to such relative changes). Thus in the *General Theory* Keynes writes:

> To suppose that a flexible wage policy is a right and proper adjunct of a system which on the whole is one of *laissez-faire*, is the opposite of the truth. It is only in a highly authoritarian society, where sudden, substantial, all-round changes could be decreed that a flexible wage-policy could function with success. One can imagine it in operation in Italy, Germany or Russia, but not in France, the United States or Great Britain (*GT*, p. 269).

This is a somewhat naive notion of what even a totalitarian government can do. In any event, this passage – and the context in which it and the other passages cited above appear – makes it clear that Keynes's purpose was not to advocate the policy of wage flexibility, but to provide a ‘negative proof’ of its impracticability for a democratic society. (Today's version of Keynes's statement in the foregoing passage would be that if equilibrium prices and wages were established by means of a stable recontracted *tâtonnement* carried out by a Walrasian auctioneer, then, by definition, full employment would always obtain).

At the other extreme from the problem of unemployment was that of avoiding inflation. It is not surprising that this was a basic concern of Keynes during the period of the disastrous hyperinflations in Europe that followed World War I, which experience led him in his *Economic Consequences of the Peace* (1919, p. 148) to write that ‘Lenin is said to have declared that the best way to destroy the capitalist system was to debauch the currency’ (a statement that was actually due to Preobrazhensky; see Fetter, 1977, p. 78). Similarly, the adverse effects of inflation was a theme which Keynes most eloquently and forcefully presented in his *Tract on Monetary Reform* (1923). Thus in the preface to this book Keynes wrote: ‘Unemployment, the precarious life of the worker, the disappointment of expectation, the sudden loss of savings, the excessive windfalls to individuals, the speculator, the profiteer – all proceed in large measure from the instability of the standard of value.’ I must, however, add that in this book, as well as in his subsequent writings, Keynes consistently regarded the harm caused by deflation and its accompanying unemployment to be significantly greater than that of inflation.

It was probably the traumatic post-World War I experience – still fresh in his mind – that led Keynes, even in his 1930 *Treatise*, after more than five years of deflation and unemployment in Britain, to continue to be concerned with the dangers of inflation. It is also noteworthy that in his *Essays in Persuasion* (*JMK* IX, pp. 57–75) – published a year later – Keynes reproduced excerpts of the discussion of the destructive effects of inflation that had appeared in his *Economic Consequences of the Peace* and in his *Tract* – including the alleged statement of Lenin's (ibid., p. 57).

Perhaps because of the increasing severity of the depression in Britain in the years between the *Treatise* and the *General Theory* – and, even more so, because the depression had then become world-wide – the latter work is little concerned with the problems of inflation, though it does emphasize the undesirability of ‘great instability of prices’ (*GT*, p. 269; see the discussion of chapter 19 in section 4 above). It should also be noted that a recurrent theme of the *General Theory* (pp. 173, 249, 253, 296 and 301) is that as the level of employment in an economy increases as a result of an increase in effective demand, the money wage rate begins to rise even before full employment is reached. This view may be interpreted as something of an adumbration of one aspect of the later Phillips-curve analysis: namely the co-existence of inflation and unemployment.

It is also significant that after Britain began its rearmament programme early in 1937 – and when unemployment was still around 12 per cent – Keynes expressed concern with the possible inflationary outcome of such a programme that might be generated by the geographical immobility of labour. In particular, in two articles in the *Times* in the spring of 1937, Keynes argued (inter alia) that in order to avoid such pressures, the increased defence expenditures should be directed toward the distressed areas of the economy (*JMK* XXI, p. 407; see also ibid., pp. 385–6; cf. also Hutchison, 1977, pp. 10–14). And once war broke out, Keynes wrote his influential pamphlet on *How to Pay for the War* (1940), whose major purpose was to present a programme for financing the war without generating inflation – the main component of the programme being a proposal to adopt compulsory savings.

Two points should be made about the relationship between theory and policy in the *Treatise* and in the *General Theory*. First, in both cases the major contribution of the book is with respect to theory – and the purpose of the theory is to provide a rigorous underpinning for a policy position which already had many adherents. As Keynes himself indicated in chapter 13 of the *Treatise*, this was certainly true for the bank-rate policy he advocated in that work. And it is also true of the public-works-expenditure policy advocated in the *General Theory*, a policy which had been advocated by other British and American economists as well during the 1920s and early 1930s (cf. Hutchison, 1953, pp. 409–23 and 1978, pp. 175–99; Patinkin, 1969; Stein, 1969, chs 2, 7; Winch, 1969, pp. 104–46; and Davis, 1971). Indeed, as noted above, Keynes himself had already advocated this policy in his 1929 *Can Lloyd George Do It?*, and even here he was basically repeating views he had expressed five years earlier in the *Nation and Athenaeum* (*JMK* XIX, pp. 221–3). Accordingly, as also noted above, the major revolution effected by the *General Theory* was in the field of theory, and not of policy. And if (unlike the *General Theory*) the *Treatise* did deal at length with policy, it was not because it made any basic, new contribution to this question (at least in a domestic context), but because it was – as its name indicated – a comprehensive treatise, designed, inter alia, to describe the state of the art with respect to both theory and practice.

Second, and relatedly, it seems to me that the change in Keynes's policy views between the *Treatise* and the *General Theory* stemmed less from the transition from the fundamental equations to the $C + I + G = Y$ equation than from British economic developments in the quinquennium between the appearance of those two books. For, as we have seen, Keynes advocated public-works expenditures for the purpose of combating unemployment even in the *Treatise*, albeit as a second-best policy to be carried out in special circumstances. And what caused him to advocate such expenditures as a necessary addition to interest-rate policy (which, as in the *Treatise*, he continued to regard as an essential component of full-employment policy; cf. *GT*, p. 316) was the experience of five additional years of deep depression in the face of a 'cheap-money' policy that had brought the rate of interest down to unprecedented lows. In brief, I conjecture that it was this experience that led Keynes of the *General Theory* to conclude:

> For my own part I am now somewhat sceptical of the success of a merely monetary policy directed towards influencing the rate of interest. I expect to see the State, which is in a position to calculate the marginal efficiency of capital-goods on long views and on the basis of the general social advantage, taking an ever greater responsibility for directly organizing investment; since it seems likely that the fluctuations in the market estimation of the marginal efficiency of different types of capital, calculated on the principles I have described above, will be too great to be offset by any practicable changes in the rate of interest (*GT*, p. 164).

8. Just as Keynes's trilogy is bound together by a common concern with the problem of unemployment, so is it bound by a common lack of concern with the problem of economic growth. With respect to the *Treatise* and the *General Theory*, this omission is an understandable characteristic of the economic literature of the depression years. For at a time when a dismaying percentage of the existing productive potential was idle, it would have taken an unrealistic soul indeed to have concerned himself with the problem of assuring the further growth of this potential. But I think that this lack of concern reflected an additional element in Keynes's thought – and probably in that of many of his contemporaries as well.

In particular, I think that Keynes originally viewed economic growth as a process that would emerge naturally – and at a satisfactory pace – from a free-market system in which households saved, and then used these savings to purchase the securities which firms issued in order to finance their expansion. 'For a hundred years [before World War I] the system worked, throughout Europe, with an extraordinary success and facilitated the growth of wealth on an unprecedented scale' (*Tract*, p. 6) – and Keynes, like his contemporaries, was not much concerned with things outside Europe, in the broad sense of Western civilization. Now, what had seriously interfered with the growth process of Europe after the World War were the disastrous inflations, which had wiped out the real value of past savings and had accordingly discouraged further saving. Correspondingly, a necessary – and sufficient – condition to reactivate the growth process at a satisfactory pace was to reestablish the confidence of the public in the future real value of its savings (*Tract*, pp. 16–17).

The *General Theory* introduced another factor that interferes with steady growth: unemployment. And parallel to his view in the *Tract*, Keynes felt that once this disturbing factor was eliminated, growth would again proceed at a satisfactory pace. Indeed, if full employment could be maintained, 'a properly run community equipped with modern technical resources, of which the population is not increasing rapidly, ought to be able to bring down the marginal efficiency of capital in equilibrium approximately to zero within a single generation' (*GT*, p. 220): the 'zero' of the classical stationary state.

In brief, I would conjecture that in Keynes's view at this time there was no need for any special analysis of the process of economic growth. All that one had to do was to ensure the maintenance of two necessary preconditions: a stable value of money and full employment. And growth – to the extent that the economy was interested in it (cf. *GT*, p. 377) – would take care of itself.

(Though Keynes did not concern himself with the problem of growth, the analytical framework of the *General Theory* served as the point of departure for the growth models which were subsequently developed. In this context it is interesting to note the transformation that took place over the years in the attitude toward saving: whereas the spirit of the *General Theory* hovers over the early contributions by Harrod (1939) and Domar (1946), which regard the increase in potential savings generated by increasing income as a threat to full employment, and growth as the means (via the acceleration principle) of generating the level of investment necessary for absorbing these savings and thus eliminating this threat, the later contributions regard savings as a desirable act necessary for financing the additional investment required for the growth process. Correspondingly, growth was transformed from being a means to an end to being an end in itself.)

Another common bond of the *Treatise* and *General Theory*, in quite a different plane, is the fact that the highly novel theoretical developments which mark both works were first presented to the profession at large as finished products, i.e. in the form of published books. In neither case did Keynes attempt to exploit the relatively long period of preparation that was involved (roughly, five years) in order to publish articles in the leading scientific journals on the salient features of his new theories and thus to benefit from the exposure of these theories to the criticism of the profession at large before formulating them in final book form. It is true that such a 'research strategy' was much less customary at the time Keynes wrote than it became later. But I would conjecture that Keynes's failure to follow such a strategy also reflected his belief that the quintessence of economic knowledge was in Cambridge – which geographical point need at most be extended to a triangle that would include London and Oxford. So why bother publishing articles in order to benefit from criticism, if the most fruitful criticisms could be reaped more conveniently and efficiently simply by circulating draft-manuscripts and galley proofs among his colleagues in this fertile triangle?

And as the materials in *JMK* XIII show us, this is indeed the procedure that Keynes employed in the writing of the *General Theory*. On the other hand, there is little if any evidence that the *Treatise* was subjected to much effective prepublication criticism even within this triangle. And this is particularly true for what Keynes considered to be its major theoretical innovation – the fundamental equations (Patinkin, 1976a, pp. 20–21, 29–32). Correspondingly, there are many serious deficiencies in the *Treatise* which were pointed out

immediately after its publication and which (I conjecture) would have been avoided if only it had been subjected to such criticism. I would also conjecture that it was precisely this unfortunate experience with the *Treatise* that Keynes had in mind when in the preface to the *General Theory* he wrote that ‘It is astonishing what foolish things one can temporarily believe if one thinks too long alone, particularly in economics (along with the other moral sciences), where it is often impossible to bring one's ideas to a conclusive test either formal or experimental’ – and that accordingly made him so eager to seek out criticism at every stage of the writing of the *General Theory*.

Might I also digress to suggest that another cause of the deficiencies in the *Treatise* was the simple but frequently neglected fact that Keynes too was of flesh and blood, subject like all mortals to the inexorable constraint that there are only 24 hours in the day; and there can be little doubt that Keynes just did not have enough hours to devote to the writing of the book, and especially of its final version. In particular, in August 1929, Keynes informed his publisher that he felt he had to ‘embark upon a somewhat drastic rewriting’ of what was then a one-volume book, for the most part already in galley and page proof (*JMK* XIII, pp. 117–18). But three months later Keynes was appointed to the famous Macmillan Committee and proceeded to play a leading role in its deliberations. Then at the beginning of 1930, he became a member – and a most active one – of the newly appointed Economic Advisory Council (see section 11 below). All this makes it difficult to believe that Keynes could have had enough time during 1930 to devote to the rewriting of the *Treatise* that he deemed necessary.

Another indication of this pressure of time is the fact that though Hawtrey had provided Keynes with basic criticisms of the *Treatise* before its publication (specifically, in the spring and summer of 1930), Keynes did not take account of them and did not even answer Hawtrey until a month after the book was published in October 1930. Keynes apologised then for this delay by explaining that he was, as we can well believe, ‘overwhelmed’ with work of the Macmillan Committee, the Economic Advisory Council ‘and a hundred other matters’ (*JMK* XIII, p. 133). And I suspect that this was also the reason that in 1930 Keynes did not give the series of lectures on monetary economics that it was his custom to give every autumn term at Cambridge (see section 11 below), and that in autumn 1931 he deferred his lectures to the following spring.

And though it may sound like a morality play – like a didactic reaffirmation of the victory of good scientific procedures over bad – I would like to point out that in the writing of the *General Theory* this pressure of time was much less evident. In particular, after the completion of the Macmillan Report in June 1931, Keynes seems to have been much less occupied than before with activities on behalf of the government. Similarly, after 1933 there was (to judge from an enumeration of the relevant entries in Hudson's unpublished and admittedly incomplete bibliography of Keynes's writings) a falling-off in the intensity of his journalistic activities. Correspondingly, I would conjecture that in the last two years before their respective publication, Keynes was able to concentrate far more on the writing of the *General Theory* than he had been able to on the writing of the *Treatise*.

9. I turn now to some observations on Keynes's style – both analytical and literary. Insofar as the analytical style is concerned, let me again note Keynes's failure to make use in his writings of graphical techniques – and this despite the fruitful precedent on this score set by his teacher Marshall, and despite the many passages (see, e.g., the reference on pp. 25 and 30 of the *General Theory* to the ‘intersection of the aggregate demand function with the aggregate supply function’) that almost cry out for a diagram. Here and there in the trilogy there are diagrams of a statistical or schematic nature (*Tract*, pp. 83, 87; *TM* I, pp. 290–91; II, p. 317). But, as noted above, in all of these books there is only one diagram of an analytical nature – and that diagram is due to Harrod (*GT*, p. 180, n. 1). Similarly – to judge from the student notes that have survived (reproduced in Rymes (ed.), 1988) – Keynes made practically no use of diagrams in his lectures.

Keynes's failure to use graphical techniques in the *General Theory* is even more puzzling in light of the fact that his chief disciples and critics during the formative period of writing the book – namely, Richard Kahn and Joan Robinson – played a leading role in the breakthrough that was then taking place in the use of such techniques! I am, of course, referring to Joan Robinson's *Economics of Imperfect Competition* (1933a), in the writing of which she acknowledged the ‘constant assistance of Mr R.F. Kahn’ (ibid, p. v).

Marshall's influence on Keynes did, however, manifest itself in the fact that the analysis of both the *Treatise* and the *General Theory* is carried out in terms of ‘demand price’ and ‘supply price’ (see sections 2 and 3 above). It has also been contended in section 5 above that Keynes's ‘unemployment equilibrium’ in the *General Theory* must be understood in terms of Marshall's short-period equilibrium. A more subtle manifestation of Marshall's influence is the fact that the formal organization of the argument of the *General Theory* is that of partial-equilibrium analysis. In particular, if this argument had been organized in accordance with the Walrasian general-equilibrium approach, then (as in present-day textbooks of macroeconomics), Book III of the *General Theory* would have been devoted to the market for goods (both consumption and investment) and Book IV in a parallel fashion to that for money, and there would then follow a discussion of the interaction between these two markets. In point of fact, however, both Book III (‘The Propensity to Consume’) and Book IV (‘The Inducement to Invest’) are formally devoted to the market for goods, with the market for money being discussed in Book IV not as an equal partner, but as the source of an influence (via the rate of interest) on the market for investment goods. Nevertheless, as emphasized in the discussion in section 4 above of chapter 18 of the *General Theory*, the analysis of this book is essentially that of general equilibrium. The voice is that of Marshall, but the hands are those of Walras. And in his IS–LM interpretation of the *General Theory*, Hicks quite rightly and quite effectively concentrated on the hands.

In connection with Keynes's analytical style, I should also note his oft-cited criticism in the *General Theory* of ‘symbolic pseudo-mathematical methods of formalizing a system of economic analysis … which allow the author to lose sight of the complexities and interdependencies of the real world in a maze of pretentious and unhelpful symbols’ (*GT*, pp. 297–8). Let us, however, not take this statement too seriously. First of all, Keynes's own analysis in his earlier *Treatise on Money* (1930) was, in fact, largely based on fairly mechanical applications of the so-called fundamental equations. Similarly, an entire chapter (20) of the *Treatise* is devoted to ‘An Exercise in the Pure Theory of the Credit Cycle’, in which Keynes explored in a very formalistic manner – and under a variety of alternative assumptions – the mathematical properties of his model of the cycle. Thus, if ever an author made use of ‘a maze of pretentious and unhelpful symbols’, that author was Keynes of the *Treatise*.

Furthermore, I strongly suspect that a comparison of the *General Theory* (and a fortiori the *Treatise*) with other works on economic theory that were written during that period would

actually show Keynes's works to be among the more mathematical of them. Indeed, in his review of the *General Theory*, Austin Robinson commented that ‘even for the ordinary economist, the argument, being largely in mathematical form, is difficult’ (1936, p. 472).

It may have been Keynes's lack of success with formal model building in the *Treatise* that led him to the more critical attitude expressed in the passage from the *General Theory* just cited. In any event, it is significant that in the *General Theory* – in contrast with the *Treatise* – Keynes did not attempt to provide a formal mathematical model of the theory of employment that constitutes the central message of the book. This was left for the subsequent exegeses of such writers as Hicks (1937) and Lange (1938). Instead, to the extent that Keynes made use of mathematical analysis in the *General Theory*, he did so with respect to such secondary themes as the relationship between the own-rates of interest of different goods (ch. 17, section II) and the theory of prices (ch. 21, section VI). And even in these instances, the mathematical formulation adds little to the exposition, and so could be deleted without much loss of continuity. Indeed, in a letter he wrote a year after the publication of the book in response to criticisms of the formulas in the first section of his chapter on ‘The Employment Function’ (chapter 20), Keynes himself admitted:

> I have got bogged [sic] in an attempt to bring my own terms into rather closer conformity with the algebra of others than the case really permits. When I come to revise the book properly, I am not at all sure that the right solution may not lie in leaving out all this sort of stuff altogether, since I am extremely doubtful whether it adds anything at all which is significant to the argument as a whole (*JMK* XXIX, p. 246).

Actually, the *General Theory* reveals an ambivalent attitude toward the role of mathematical analysis in economics; for with all his reservations about the usefulness of such analysis, Keynes (as one who had once been bracketed Twelfth Wrangler; see Harrod, 1951, p. 103) could not resist the temptation to show that he too could employ it. Thus the foregoing quotation from the *General Theory* so critical of mathematical analysis actually occurs in section III of the same chapter 21 that I have just cited as providing an instance of the use of such analysis – and indeed this quotation appears as part of Keynes's apologia for nevertheless going ahead and resorting to it in section VI of that chapter!

Furthermore, judging from the critical literature that subsequently grew up around chapters 17 and 21, I think it fair to say that the mathematical analysis that appears in these chapters is not only not essential to the argument, but sometimes even incorrect (thus see Palander (1942) as cited by Borch (1969), as well as Naylor (1968, 1969), on the incorrect elasticity formula used to analyse the implications of the quantity theory in chapter 19 of the *General Theory* (p. 305); see also Patinkin (1982, p. 151, n.33) on the erroneous formula in n.2 on p. 126). And this fact, together with the ineffectualness of the fundamental equations of the *Treatise*, makes it clear that whatever may have been Keynes's attitude toward the proper role of mathematical methods in economic analysis, his strength did not lie in the use of such methods.

Nor in general did Keynes's analytical strength lie in rigour and precision: indeed, we run the risk of distorting the original intention of Keynes's writings – and reading meaning into them – if we try to view them through analytical lenses that are more sophisticated and more finely ground than those that he was wont to use. Thus in both the *Treatise* and the *General Theory* Keynes frequently failed to specify the exact nature of the assumptions that underlay his argument. Furthermore, there are many ambiguities in these books. And the best evidence of the existence of such ambiguities and obscurities is the fact that fifty years later disagreements continue about the role played in the *General Theory* by such crucial assumptions as wage rigidities, the liquidity trap, the interest elasticity of investment, unemployment equilibrium, and the like – not to speak of the protracted debate about the meaning of Keynes's aggregate supply function.

Instead, Keynes's analytical strength lay in his creative insights about fundamental problems that led him to make major breakthroughs, leaving for those that followed him to correct, formalize, and complete his initial achievements. In the *Treatise*, Keynes thought (erroneously, as it turned out) that his fundamental equations constituted such a breakthrough. In the *General Theory*, he saw his breakthrough as lying in his theory of effective demand – and this time he was undeniably right.

In view of this basic aspect of Keynes's analytical style, I should in all fairness also emphasize that the aforementioned lack of rigour and completeness in part reflects the natural deficiency of many a pathbreaking work. As Keynes wrote to Joan Robinson: ‘My own general reaction to criticisms always is that of course my treatment is obscure and sometimes inaccurate, and always incomplete, since I was tackling completely unfamiliar ground, and had not got my own mind by any means clear on all sorts of points’ (*JMK* XIII, p. 270). Keynes made this comment in 1932 with reference to the *Treatise*; it is even more relevant for the *General Theory*.

Another characteristic of Keynes's style that should be noted is his constant striving to present the conclusions of his analysis in the form of paradoxes. Sometimes this is very effective, as in the case of the ‘paradox of thrift’ in the *General Theory*. Sometimes, however, Keynes's love for the paradoxical tempts him into extreme statements that do not stand up under critical scrutiny, as in the case of the paradox of the widow's cruse in the *Treatise* (I, p. 125; see Joan Robinson, 1933b). And sometimes it tempts him into delphic pronouncements, such as his oft-cited contention that ‘there may exist no expedient by which labour as a whole can reduce its real wage to a given figure by making revised money bargains with entrepreneurs’ (*GT*, p. 13, italics in original; but see the discussion of chapter 19 of the *General Theory* in section 4 above for an interpretation).

A related characteristic of his style are occasional seemingly profound statements that upon closer examination lose much (if not all) of their profundity and are sometimes even involved in error. Thus consider the following passage from the *Treatise*:

> We have claimed to prove in this treatise that the price level of output depends on [1] the level of money incomes relatively to efficiency, on [2] the volume of investment (measured in cost of production) relatively to saving, and on [3] the ‘bearish’ or ‘bullish’ sentiment of capitalists relatively to the supply of savings deposits

> available in the banking system (*TM* II, p. 309, bracketed numbers added).

This is simply a verbal rendition of the second fundamental equation (itself a tautology) written as the weighted average of the respective prices of consumption goods (P) and investment goods (P')

$$\Pi = (P \cdot R + P' \cdot C) / O$$

where, by definition,

$$O = R + C$$

(*TM* I, p. 123). More specifically, the first fundamental equation in section 2 above can be written as

$$P = (W / e) + (I' - S) / R \qquad \text{(i)}''$$

(*TM* I, p. 122); expressions [1] and [2] in the foregoing passage thus correspond to the first and second terms, respectively, of this equation. And expression [3] in turn is a brief summary of Keynes's explanation of the determination of P' (*TM* I, pp. 127–9, 229–30). (For other instances of obscure statements in the *Treatise* which are simply verbal renditions of the fundamental equations, see *TM* I, pp. 144 and 248–9; for further details, see Patinkin, 1976a, ch. 6).

Or consider the following well-known passage at the end of chapter 19 of the *General Theory*:

> If, as in Australia, an attempt were made to fix real wages by legislation, then there would be a certain level of employment corresponding to that level of real wages; and the actual level of employment would, in a closed system, oscillate violently between that level and no employment at all, according as the rate of investment was or was not below the rate compatible with that level; whilst prices would be in unstable equilibrium when investment was at the critical level, racing to zero whenever investment was below it, and to infinity whenever it was above it (*GT*, pp. 269–70).

As at other points in the *General Theory*, Keynes assumes here that there is a fixed consumption function, so that the level of effective demand and hence employment is determined by that of investment. In the case where that level is greater than the level of employment corresponding to the fixed real wage rate, the argument is a straightforward application of the analytical framework of the book: viz, there will then be an excess demand for goods which will drive their price higher; but since the real wage rate is being held constant, the money wage rate must increase in the same proportion. Thus prices will ‘race to infinity’, unless (Keynes goes on to say) the resulting decrease in the real quantity of money and consequent increase in the rate of interest will decrease investment, and hence effective demand and employment to the level corresponding to the fixed real wage rate.

It is, however, not clear why – in the case where the level of effective demand and hence employment is less than that corresponding to the fixed real wage rate – the economy should be driven down to a situation of ‘no employment at all’. For the firms’ marginal productivity of labour corresponding to that lower level of employment is higher than the fixed real wage rate; on the other hand, that fixed rate is higher than the minimum one upon which workers insist in order to provide that level of employment. Hence this lower level can constitute a stable equilibrium in Keynes's sense of the term. Correspondingly, there is no reason in this situation for prices to ‘race to zero’. (See the discussion at the beginning of section 5 above of Keynes's use of the term ‘unemployment equilibrium’.)

Note the key to interpreting the above passages: each is a mechanical application of the basic formula of the book in question (the fundamental equations in the case of the *Treatise*,

and the theory of effective demand which determines employment hence the real wage rate in the case of the *General Theory*) – combined with Keynes's propensity to shock (see his letter to Harrod cited at the beginning of section 5 above).

Obscurities such as these, as well as those mentioned above, frequently impede the flow of the reading. But despite these difficulties, there are constant reminders throughout the trilogy that we are in the presence of a master of English style. The language is generally rich and incisive, enhanced occasionally by well-turned phrases and apt literary allusions. For Keynes's objective is to appeal not only to the intellect but also to the sense of literary appreciation.

This is particularly true of the *Tract*, and for two related reasons: because it is the least technical of the three books and because of its origin as a series of articles on current policy in the *Manchester Guardian*, where Keynes could give full expression to his brilliant journalistic style.

Least enjoyable as a reading experience is the *Treatise*, whose generally heavy and constrained style reflects the stately scientific objective that Keynes set for himself in it. Indeed, when one reads the *Treatise* against the background of Keynes's other writings, one cannot escape the feeling that it represents a Keynes out of character, a Keynes attempting to act the role of a Professor, and a Germanic one at that.

In the *General Theory* we once again find the true Keynes. Here (as in so many of Keynes's writings) is the stirring voice of a prophet who has seen a new truth and who is convinced that it – and only it – can save a world deep in the throes of crisis. It is a sharp, polemical voice directed at converting economists all over the world to the new dispensation and combating the false prophets among them who perversely continue with the erroneous teachings of the gods of classical mythology whom Keynes had already abandoned.

And so it is that these writings of Keynes are famous not only for their basic scientific contributions but also for having become part of the literary heritage of every economist. For who does not know that ‘in the long run we are all dead’ (*Tract*, p. 65)? Or that

> The ideas of economists and political philosophers, both when they are right and when they are wrong, are more powerful than is commonly understood. Indeed the world is ruled by little else. Practical men, who believe themselves to be quite exempt from any intellectual influences, are usually the slaves of some defunct economist. Madmen in authority, who hear voices in the air, are distilling their frenzy from some academic scribbler of a few years back …. The power of vested interests is vastly exaggerated compared with the gradual encroachment of ideas (*GT*, p. 383).

10. The foregoing discussion of similarities and differences among the volumes of Keynes's trilogy brings us finally to the question of the justification for reading them today. From the substantive viewpoint, all of these volumes are now in the domain of the history of monetary doctrine: their basic scientific contributions have long since been incorporated in the current literature, so that, by definition, the volumes themselves are of importance only to students of this history.

From a broader viewpoint, however, there are sharp differences among these volumes in this respect too. Thus, in these times of worldwide inflation, one can still read with both pleasure and profit Keynes's brilliant discussion of this problem in the *Tract*. On the other hand, the recent revival of interest in the *Treatise* notwithstanding, I can (from the viewpoint of macroeconomic theory) see little profit (and certainly no pleasure) in reading it today. Nor do I think that the *Treatise* is important as a key to an understanding of the major innovation of the *General Theory*, namely, the theory of effective demand. What the *Treatise* does help us understand are certain terminological aspects of Keynes's presentation of this theory (viz., his exposition in terms of ‘demand price’ and ‘supply price’; cf. *GT*, pp. 24–6 and *TM* I, pp. 186, 189); but it contributes little towards an understanding of the substance of the theory itself, which differs so fundamentally from that of the *Treatise*.

As for the *General Theory*: the work over the years of students of Keynes's thought has deepened our understanding of this book, but has also brought to light deficiencies and errors. Some of these are due to the stylistic excesses described in section 9 above; some are inconsequential mathematical ones, like those noted in the same section; but some (e.g., the ambiguities and errors in Keynes's discussion of the aggregate supply curve referred to in section 3 above) are more significant. But even these last should be regarded as the kind that naturally occur in a pioneering work that breaks new ground and develops a radically different analytical framework. We do no service to the place of Keynes in the history of economic thought – and *a fortiori* not to the history itself – by ignoring these errors. At the same time, they do not change the basic fact that this is the book that made the revolution which has continued to mould our basic ways of thinking about macroeconomic problems. And so the reading of it – at least in part – is an intellectual experience that no aspiring economist even today can afford to forego.

To this I must add the following related plea. In reading the *General Theory*, let us do so in order to acquaint ourselves with one of the classics of our discipline, and, more generally, in order to enjoy the pleasures of intellectual history: not in order to invoke Keynes's alleged authority with respect to further developments in macroeconomic theory. Thus, for example, if we feel that this theory should provide a more detailed analysis of the way expectations and hence behaviour decisions are formed under conditions of uncertainty; or of the role of money wages and prices in the equilibriating process generated by the interaction between aggregate demand and supply; or of the influence of the structure of interest rates on the respective markets for money and commodities – then let us by all means devote ourselves to the analysis of these important questions. At the same time, let us make a clear distinction between this objective and that of the history of thought – and thereby do a service both to Keynes and to the further development of macroeconomic theory: for we then permit the study of Keynes's thought to concern itself not with what Keynes might have said or should have said about current theoretical questions, but with what he actually did say; and we permit the attempts to improve upon the current state of macroeconomic theory to be judged substantively, on their own merits, without confusing the issue with arguments about ‘what Keynes really meant’. As Keynes said in concluding a long and tiresome correspondence in 1938 on a note that some economist had sent him on an aspect of the *General Theory*, ‘… the enclosed, as it stands looks to me more like theology than economics! …I am really driving at something extremely plain and simple which cannot

possibly deserve all the exegesis' (*JMK* XXIX, p.282; cf. also Patinkin, 1984, pp. 100–101).

11. Having devoted so much attention to Keynes's trilogy, I must emphasize that it would be a serious mistake to think of Keynes as devoting his major efforts in the interwar period to writing these books in the quiet halls of academe. On the contrary, after he became a public figure in the wake of his *Economic Consequences of the Peace* (1919), he resigned his lectureship at Cambridge (though he continued as an active Fellow of King's College) and earned his living from his publicistic writings and from speculation on the stock market (Johnson and Johnson, 1978, pp. 1–37; Harrod, 1951, pp. 288, 294–304). Correspondingly, Keynes's normal routine became one in which he divided his time between London and Cambridge, living in the former during most of the week and coming down to Cambridge for long weekends. In London he was absorbed in his publicistic and political activities; during the weekends at Cambridge he dealt with both academic and (as bursar of King's) business matters. On Monday mornings of the autumn term during most of the interwar years he also gave a course of lectures on monetary economics which were widely attended by students, faculty and visitors, and in the process of which he expounded his new theories as he developed them. It is of these lectures that we have the notes of Bryce, Tarshis and others mentioned at the beginning of section 3 above. On Monday evenings Keynes would then preside over his famous Political Economy Club, whose participants were drawn from the most promising undergraduates, and at which one of them would read a paper which would then be discussed (Harrod, 1951, pp. 149–52, 327–30; see also the reminiscences of Bryce and Tarshis of both the lectures and the Club in Patinkin and Leith, (eds.), 1977, pp. 39–63, 73–74). And the following morning he would be back in London.

Keynes's intensive public activity with respect to the policy discussions of the interwar period was reflected in the more than three-hundred articles he wrote for the 'highbrow' news magazines of the time (particularly the *Nation and Athenaeum* – of whose board Keynes was chairman in the 1920s – and its successor, *The New Statesman and Nation*) as well as for the popular press. Many of the latter articles were syndicated in newspapers all over the world. A selection from these and similar writings was reissued by Keynes in 1931 under the title *Essays in Persuasion*. These are marked by a brilliant style, truly the work of a literary craftsman.

There was one pressing and recurrent politico-economic issue of the postwar world of the 1920s – German reparations – which Keynes discussed not only in books addressed to the general public (1919, 1922) and in numerous magazine articles (reproduced in *JMK* XVII–XVIII), but also in the pages of the *Economic Journal* (which Keynes edited from 1912 to 1944; some of the interesting correspondence which he carried out in this capacity is reproduced in *JMK* XII, pp. 784–868). The reference is, of course, to Keynes's 1929 debate with Ohlin about the possibility of Germany's carrying out the payments imposed upon it by the Versailles Treaty: the famous debate about the 'transfer problem'. In light of the central role that the notion of effective demand was a few years later to play in the *General Theory*, it is ironic to note that in this debate it was Ohlin who emphasized the role of 'buying power' in carrying out the reparations, and Keynes who overlooked it. One cannot help suspecting that Keynes's thinking here was coloured by his violent objections to the Treaty itself (see introductory section of this essay). It should, however, be noted that a similar neglect of 'buying power' characterizes Keynes's other writings of this period: namely, his discussion of the effects of public-works expenditures in both *Can Lloyd George Do it?* (1929) and the *Treatise* (1930) (see Patinkin, 1976a, p. 129).

Keynes's accomplished literary style also characterizes his *Essays in Biography* (1933b), in which Keynes reprinted his impressions of the leading political figures he had known, as well as his biographical essays on various British economists. Most notable among the latter are his stimulating essay on Thomas Malthus and his perceptive and evocative memorial essay on his teacher, Alfred Marshall.

At various critical junctures in the interwar period, Keynes also published influential pamphlets in which he analysed the questions at issue and proclaimed his prescriptions. Such were his *Economic Consequences of Mr Churchill* (1925), in which he criticized the decision of the then Chancellor of the Exchequer to return to the gold standard at prewar parity, claiming that the resulting overvaluation of the pound generated depression in British export industries which then spread to the rest of the economy; *Can Lloyd George Do It?* (1929) (written with Hubert Henderson), in support of the Liberal Party's pledge in the 1929 election campaign to reduce unemployment by means of public works; *The Means to Prosperity* (1933a), in further support of public works (this time making use of the newly developed notion of the multiplier) as the depression deepened in the early 1930s; and *How to Pay for the War*, as in 1940 the problems of depression gave way to those of wartime inflationary pressures. (All of these pamphlets have been reproduced in *JMK* IX.)

I should, however, note that already in 1943 Keynes also began to concern himself with post-war problems and wrote a memorandum on 'The Long-Term Problem of Full Employment' advocating a programme in which 'two-thirds or three-quarters of total investment is carried out or can be influenced by public or semi-public bodies' (*JMK* XXVII, p. 322). And in reply to a comment on it by James Meade, he wrote (letter of 27 May 1943):

> It is quite true that a fluctuating volume of public works at short notice is a clumsy form of cure and not likely to be completely successful. On the other hand, if the bulk of investment is under public or semi-public control and we go in for a stable long-term programme, serious fluctuations are enormously less likely to occur (*JMK* XXVII, p. 326).

Similar views were expressed by Keynes in an unpublished February 1944 'Note on Postwar Employment' and in a December 1944 letter to Beveridge (*JMK* XXVII, pp. 365, 381). Thus to the end of his days, Keynes continued to advocate public-works expenditures as a necessary component of a full-employment policy. It should, however, also be emphasized that – as in the *General Theory* and *a fortiori* the *Treatise* – Keynes also continued to stress the essential role of a low rate of interest in carrying out this policy. Indeed, in a series of articles in the *Times* which he published in 1937 entitled 'How to Avoid the Slump', he wrote that 'we must avoid it [i.e., 'dear money'] as we would hell-fire' (*JMK* XXI, p. 389). Keynes influenced policy not only through his publicistic activities, but also by his active membership in various official government bodies. Thus he was the leading figure of the

Committee on Finance and Industry (the Macmillan Committee, 1929–31) and of the Economic Advisory Council (1930–39), and he also served as chairman of the Committee of Economists (1930) – all of which were charged with advising the British government on different aspects of the policies it should follow in order to overcome the serious depression in which Britain, together with the rest of the Western world, then found itself (cf. Howson and Winch, 1977). Similarly, at the outbreak of World War II, Keynes was appointed adviser to the Chancellor of the Exchequer, a position he held until his death. He also played a leading role in the negotiations with the United States government, first for lend-lease support in 1941 and again in 1944, and then for a special postwar loan in 1945. Keynes was also one of the architects of the Bretton Woods agreement (1944), which established the International Monetary Fund and the International Bank for Reconstruction and Development (the World Bank). Indeed, the Fund's original policy of fixing par values for the various exchange rates, but permitting fluctuations of up to 10 per cent about them, is clearly reminiscent of Keynes's advocacy in the *Treatise* (II, p. 303) of maintaining the fixed exchange rates of the international gold standard, but widening the gold points so as to permit fluctuations of the rates within a range of two per cent. In the foregoing capacities, Keynes wrote countless letters, memoranda, reports, draft proposals, and the like, the major ones of which are reproduced in the relevant Activities volumes of his *Collected Writings* (*JMK* XX–XXVI; see also Kahn (1976), Williamson (1983) and Moggridge (1986) on Keynes's views on the international monetary system from his earliest writings up to and including the IMF).

As indicated above, Keynes's concern with policy questions also exerted a strong influence on the direction of his scientific writings. This was clearly the case for his *Tract on Monetary Reform* (1923), which had its origins in newspaper articles that Keynes had written on current economic problems. Similarly, the predominant emphasis of the *Treatise on Money* (1930) on the problems of unemployment and of the workings of the international gold standard reflected the major economic concerns of the period. By the time the *General Theory* (1936) was being written, however, the gold standard had collapsed, while the problem of unemployment had become increasingly severe. Correspondingly, the *General Theory* is concerned almost exclusively with the problem of mass, long-run unemployment in a closed economy: that is, one not subject to the restrictions imposed by the gold standard.

Keynes's interests ranged far beyond the confines of economics. He was for many years a member of the famous Bloomsbury Circle. His cultural activities included the theatre, dance, paintings, and rare-book collecting. He was instrumental in establishing the Arts Council, which provided state patronage of the arts. In all these ways Keynes played a prominent role in the cultural and intellectual life of the Britain of his day (see Harrod, 1951; White, 1974; Milo Keynes (ed.), 1975; Crabtree and Thirlwall (eds.), 1980; and Skidelsky, 1983 and 1988).

See Also

- Keynes, John Maynard (new perspectives)

Selected works

1911. Review of *The Purchasing Power of Money*, by I. Fisher. *Economic Journal* 21, September, 393–98. As reprinted in Keynes, *Collected Writings*, Vol XI, 275–81.

1913. *Indian Currency and Finance*. As reprinted in Keynes, *Collected Writings*, Vol. I.

1919. *The Economic Consequences of the Peace*. As reprinted in Keynes, *Collected Writings*, Vol. II.

1921. *A Treatise on Probability*. As reprinted in Keynes, *Collected Writings*, Vol. VIII.

1922. *A Revision of the Treaty*. As reprinted in Keynes, *Collected Writings*, Vol. III.

1923., ed. *Reconstruction in Europe. Manchester Guardian Commercial*, appeared in twelve instalments by different authors over the period 20 April 1922–4 January 1923.

1923. *A Tract on Monetary Reform*. As reprinted in Keynes, *Collected Writings*, Vol. IV.

1925. *The Economic Consequences of Mr Churchill*. As reprinted in Keynes, *Collected Writings*, Vol. IX, 207–30.

1926. ‘The end of laissez-faire’. As reprinted in Keynes, *Collected Writings*, Vol. IX, 272–94.

1929. The German transfer problem. *Economic Journal* 39, March, 1–7; The reparation problem: a rejoinder. *Economic Journal* 39, June, 179–81; A reply. *Economic Journal* 39, September, 404–8. As reprinted in Keynes, *Collected Writings*, Vol. XI, 451–9, 468–72, 475–80.

1929. (With Hubert Henderson.) *Can Lloyd George Do It?: An Examination of the Liberal Pledge*. As reprinted in Keynes, *Collected Writings*, Vol. IX, 86–125.

1930. *A Treatise on Money*. Vol. I: *The Pure Theory of Money*. As reprinted in Keynes, *Collected Writings*, Vol. V.

1930. *A Treatise on Money*. Vol. II: *The Applied Theory of Money*. As reprinted in Keynes, *Collected Writings*, Vol. VI.

1931a. An economic analysis of unemployment. In Harris Memorial Foundation (1931a), 3–42. As reprinted in Keynes, *Collected Writings*, Vol. XIII, 343–67.

1931b. *Essays in Persuasion*. As reprinted with additions in Keynes, *Collected Writings*, Vol. IX.

1933a. *The Means to Prosperity*. As reprinted in Keynes, *Collected Writings*, Vol. IX, 335–66.

1933b. *Essays in Biography*. As reprinted with additions in Keynes, *Collected Writings*, Vol. X.

1936. *The General Theory of Employment, Interest and Money*. As reprinted in Keynes, *Collected Writings*, Vol. VII.

1937. The general theory of employment. *Quarterly Journal of Economics* 51, February, 209–23. As reprinted in Keynes, *Collected Writings*, Vol. XIV, 109–23.

1939. Professor Tinbergen's method. *Economic Journal* 49, September, 558–70. As reprinted in Keynes, *Collected Writings*, Vol. XIV, 306–18.

1940. *How to Pay For the War*. As reprinted in Keynes, *Collected Writings*, Vol. IX, 367–439.

1946. The balance of payments in the United States. *Economic Journal* 56, June, 172–89. As reprinted in Keynes, *Collected Writings*, Vol. XXVII, 427–46.

1973a. *The General Theory and After*. Part I, *Preparation*. Ed. by D. Moggridge, Vol. XIII of Keynes, *Collected Writings*.

1973b. *The General Theory and After*. Part II, *Defence and Development*. Ed. D. Moggridge, Vol. XIV of KeynesKeynes, *Collected Writings*.

1979. *The General Theory and After: A Supplement*. Ed. D. Moggridge, Vol. XXIX of Keynes, *Collected Writings*.

1980. *Activities 1940–1946: Shaping the Post-War World: Employment and Commodities*. Ed. D. Moggridge, Vol. XXVII of Keynes, *Collected Writings*.

1981. *Activities 1922–1929: The Return to Gold and Industrial Policy*. Ed. by D. Moggridge, Vol. XIX of Keynes, *Collected Writings* (in two parts).

1982. *Activities 1931–1939: World Crises and Policies in Britain and America*. Ed. D. Moggridge, Vol. XXI of Keynes, *Collected Writings*.

1983a. *Economic Articles and Correspondence: Academic*. Ed. D. Moggridge, Vol. XI of Keynes, *Collected Writings*.

1983b. *Economic Articles and Correspondence: Investment and Editorial*. Ed. D. Moggridge, Vol. XII of Keynes, *Collected Writings*.

1971–88. *The Collected Writings of John Maynard Keynes*. Vols I–VI (1971), Vols VII–VIII (1973), Vols IX–X (1972), Vols XI–XII (1983), Vols VIII–XIV (1973), Vols XV–XVI (1971), Vols XVII–XVIII (1978), Vols XIX–XX (1981), Vol. XXI (1982), Vol. XXII (1978), Vols XXIII–XXIV (1979), Vols XXV–XXVII (1980), Vol. XXVIII (1982), Vol. XXIX (1979), Vol. XXX (Bibliography, Index, 1989). London: Macmillan for the Royal Economic Society.

Bibliography

Reprinted or translated works are cited in the text by year of original publication; the page references to such works in the text are, however, to the pages of the reprint or translation in question.

Backhaus, J. 1985. Keynesianism in Germany. In Lawson and Pesaran (1985), 209–503.

Blaug, M. 1962. *Economic Theory in Retrospect*. Homewood, Ill.: Richard D. Irwin.

Blaug, M. 1964. Economic Theory and Economic History in Great Britain, 1650–1776. *Past and Present* 28, 111–16.

Borch, K. 1969. Another note on Keynesian mathematics. *Economic Journal* 79, March, 182–3.

Clark, C. 1932. *The National Income: 1924–1931*. London: Macmillan.

Clark, C. 1937. *National Income and Outlay*. London: Macmillan.

Cairncross, A. 1978. Keynes and the planned economy. In Thirwall (1978), 36–58.

Corry, B. 1978. Keynes in the history of economic thought. In Thirwall (1978), 3–34.

Crabtree, D. and Thirwall, A.P., eds. 1980. *Keynes and the Bloomsbury Group*. London: Macmillan.

Davidson, P. 1972. *Money and the Real World*. London: Macmillan.

Davis, J.R. 1971. *The New Economics and the Old Economists*. Ames, Iowa: Iowa State University Press.

Domar, E. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14, April, 137–47.

Fetter, F.W. 1977. Lenin, Keynes and inflation. *Economica* 44, February, 77–80.

Garvy, G. 1975. Keynes and the economic activists of pre-Hitler Germany. *Journal of Political Economy* 83, March–April, 391–405.

Gustafsson, B. 1973. A perennial of doctrinal history: Keynes and ‘the Stockholm School’. *Economy and History* 16, 114–28.

Hansen, A.H. 1953. *A Guide to Keynes*. New York: McGraw-Hill.

Harris Memorial Foundation. 1931a. *Unemployment as a World Problem*. Edited by Quincy Wright, Chicago: University of Chicago Press.

Harris Memorial Foundation. 1931b. *Reports of Round Tables: Unemployment As a World Problem*. 2 vols, Chicago (mimeographed).

Harrod, R.F. 1937. Mr Keynes and traditional theory. *Econometrica* 5, January, 74–86. As reprinted in Lekachman (1964), 124–52.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49, March, 14–33.

Harrod, R.F. 1951. *The Life of John Maynard Keynes*. London: Macmillan. Reprinted, New York: Augustus M. Kelley, 1969.

Hart, A.G. 1947. Keynes' analysis of expectations and uncertainty. In *The New Economics: Keynes' Influence on Theory and Economic Policy*, ed. S.E. Harris. New York: Knopf, 415–24.

Heckscher, E.F. 1935. *Mercantilism*. London: George, Allen & Unwin.

Heckscher, E.F. 1946. Något om Keynes' 'General Theory' ur ekonomisk-historisk synpunkt. *Ekonomisk Tidskrift* 48, September, 161–83. As translated under the title 'Keynes and Mercantilism' in E.F. Heckscher, *Mercantilism*, revised edition, London: George, Allen & Unwin, 1955, Vol. 2, 340–58.

Hicks, J.R. 1937. Mr Keynes and the 'Classics'; a suggested interpretation. *Econometrica* 5, April, 147–59. As reprinted in American Economic Association, *Readings in the Theory of Income Distribution*, ed. W. Fellner and B. Haley, Philadelphia: Blakiston for the American Economic Association, 1946, 461–76.

Howitt, P. 1986. The Keynesian recovery. *Canadian Journal of Economics* 19, November, 626–41.

Howson, S. and Winch, D. 1977. *The Economic Advisory Council 1930–1939*. Cambridge: Cambridge University Press.

Hutchison, T.W. 1953. *A Review of Economic Doctrines 1870–1929*. Oxford: Clarendon Press.

Hutchison, T.W. 1977. Keynes versus the 'Keynesians' …? An essay in the thinking of J.M. Keynes and the accuracy of its interpretation by his followers. London: Institute of Economic Affairs.

Hutchison, T.W. 1978. *On Revolutions and Progress in Economic Knowledge*. Cambridge: Cambridge University Press.

Johnson, E.S. and Johnson, H.G. 1978. *The Shadow of Keynes: Understanding Keynes, Cambridge and Keynesian Economics*. Chicago: University of Chicago Press.

Kahn, R.F. 1931. The relation of home investment to unemployment. *Economic Journal* 41, June, 173–98.

Kahn, R.F. 1976. Historical origins of the International Monetary Fund. In *Keynes and International Monetary Relations*, ed. A.P. Thirwall. London: Macmillan, 3–35.

Kahn, R.F. 1984. *The Making of Keynes' General Theory*. Cambridge: Cambridge University Press.

Kalecki, M. 1933. *Próba teorii koniunktury* (An essay on the theory of the business cycle). Warsaw: Institute of Research on Business Cycles and Prices. As reprinted in Kalecki (1979), 95–157.

Kalecki, M. 1935. Istota poprawy koniunkturalnej. *Polska Gospodarcza* 43, 1320–24. Reprinted in Kalecki (1979), 225–32. As translated by Ada Kalecki under the title 'The Mechanism of the Business Upswing' in Kalecki (1966), 26–33.

Kalecki, M. 1936. Par• uwag o teorii Keynesa (Some remarks on Keynes' theory). *Ekonomista* 36(3), 18–26. As reprinted in Kalecki (1979), 263–73. (For an English translation, see Targetti and Kinda-Hass, 1982.)

Kalecki, M. 1966. *Studies in the Theory of Business Cycles: 1933–1939*. Translated from the original Polish by Ada Kalecki, Oxford: Basil Blackwell.

Kalecki, M. 1979. *Kapitalizm: Koniunktura i zatrudnienie. Dziela*, tom 1 (Capitalism: Business cycles and employment. Works, Vol. 1). Ed. J. Osiatynski, Warsaw: Panstwowe Wydawnictwo Ekonomiczne (State Publishers for Economics) for Polska Akademia Nauk (Polish Academy of Sciences).

Keynes, J.N. 1891. *The Scope and Method of Political Economy*. 4th edn, London: Macmillan. Reprinted, New York: Kelley and Millman, 1955.

Keynes, M., ed. 1975. *Essays on John Maynard Keynes*. Cambridge: Cambridge University Press.

Klein, L. 1947. *The Keynesian Revolution*. New York: Macmillan.

Klein, L. 1951. The life of John Maynard Keynes. *Journal of Political Economy* 59, October, 443–51.

Knight, F.H. 1921. *Risk, Uncertainty and Profit*. New York: Houghton Mifflin.

Knight, F.H. 1923. The ethics of competition. *Quarterly Journal of Economics* 37, August, 579–624. As reprinted in Knight, *The Ethics of Competition and other Essays*, New York: Harper, 1935, >41–75.

Kuznets, S. 1934. *Gross Capital Formation 1919–1933*. New York: National Bureau of Economic Research, Bulletin No. 52.

Lange, O. 1938. The rate of interest and the optimum propensity to consume. *Economica* 5, February, 12–32. As reprinted in American Economic Association, *Readings in Business Cycle Theory*, ed. G. Haberler, Philadelphia: Blakiston for the American Economic Association, 1944, 169–92.

Lawson, T. and Pesaran, H. 1985. *Keynes' Economics: Methodological Issues*. Armonk, NY: M.E. Sharpe.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press.

Leijonhufvud, A. 1978. Foreword to the Japanese edition of Leijonhufvud (1968). Tokyo: Toyo Keizai Shinposha.

Lekachman, R., ed. 1964. *Keynes' General Theory: Reports of Three Decades*. New York: St. Martin's Press.

Lerner, A.P. 1936. The General Theory. *International Labour Review* 34, October, 435–54. As reprinted in Lekachman (1964), 203–22.

Lerner, A.P. 1952. The essential properties of interest and money. *Quarterly Journal of Economics* 66, May, 172–93.

Marshall, A. 1920. *Principles of Economics*. 8th edn, London: Macmillan.

Marshall, A. 1923. *Money, Credit and Commerce*. London: Macmillan.

Marshall, A. 1926. *Official Papers*. London: Macmillan.

Mill, J.S. 1848. *Principles of Political Economy*. Ed. W.A. Ashley, London: Longman Green, 1909.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12, January, 45–88. As reprinted in American Economic Association, *Readings in Monetary Theory*, ed. Lutz and L.W. Mints, Philadelphia: Blakiston for the American Economic Association, 1951, 186–240.

Moggridge, D.E. 1980. *Keynes*. 2nd edn, London: Macmillan.

Moggridge, D.E. 1986. Keynes and the international monetary system 1909–46. In *International Monetary Problems and Supply Side Economics: Essays in Honour of Lorie Tarshis*, ed. J.S. Cohen and G.C. Harcourt. London: Macmillan, 56–83.

Moggridge, D.E. and Howson, S. 1974. Keynes on monetary policy, 1910–1946. *Oxford Economic Papers* 26, July, 226–47.

Myrdal, G. 1933. Der Gleichgewichtsbegriff als Instrument der geldtheoretischen Analyse. In *Beiträge zur Geldtheorie*, ed. F.A. Hayek. Vienna: Julius Springer, 365–485. As translated with revisions under the title *Monetary Equilibrium*, London: W. Hodge, 1939.

Naylor, T.H. 1968. A note on Keynesian mathematics. *Economic Journal* 78, March, 172–3.

Naylor, T.H. 1969. A third note on Keynesian mathematics. *Economic Journal* 79, March, 183–4.

Ohlin, B. 1929. Transfer difficulties, real and imagined. *Economic Journal* 39, September, 172–8; A rejoinder. *Economic Journal* 39, September, 400–4.

Ohlin, B. 1937. Some notes on the Stockholm theory of saving and investment. *Economic Journal* 47, March, 53–69; June, 221–40. As reprinted in American Economic Association, *Readings in Business Cycle Theory*, Philadelphia: Blakiston for the American Economic Association, 1944, 87–130.

Palander, T. 1942. Keynes' allmänna teori. *Ekonomisk Tidskrift* 44, December, 233–71.

Patinkin, D. 1951. Price flexibility and full employment. In *Readings in Monetary Theory*, ed. F.A. Lutz and L.W. Mints. Philadelphia: Blakiston, for the American Economic Association, 252–83. (A revised and enlarged version of the article which originally appeared in *American Economic Review* 38, September 1948, 543–64.)

Patinkin, D. 1956. *Money, Interest, and Prices*. Evanston, Ill.: Row, Peterson.

Patinkin, D. 1965. *Money, Interest, and Prices*. 2nd edn, New York: Harper & Row.

Patinkin, D. 1969. The Chicago tradition, the quantity theory, and Friedman. *Journal of Money, Credit and Banking* 1, February, 46–70. As reprinted in Patinkin (1981), 241–64.

Patinkin, D. 1974. Keynesian monetary theory and the Cambridge school. In *Issues in Monetary Economics*, ed. H.G. Johnson and A.R. Nobay. Oxford: Oxford University Press, 3–30. As reprinted in Patinkin (1982), ch. 6.

Patinkin, D. 1975. The Collected Writings of John Maynard Keynes: from the *Tract* to the *General Theory*. *Economic Journal* 85, June, 249–70.

Patinkin, D. 1976a. *Keynes' Monetary Thought: A Study of Its Development*. Durham, NC: Duke University Press.

Patinkin, D. 1976b. Keynes and econometrics: on the interaction between the macroeconomic revolutions of the interwar period. *Econometrica* 44, November, 1091–123. As reprinted with minor changes and a Postscript in Patinkin (1982), ch. 9.

Patinkin, D. 1977. The process of writing the *General Theory*: a critical survey. In Patinkin and Leith (1977), 3–24.

Patinkin, D. 1978. Keynes and the multiplier. *The Manchester School* 46, September, 209–23. As reprinted in a shorter version in Patinkin (1982), ch. 7.

Patinkin, D. 1979a. Keynes and Chicago. *Journal of Law and Economics* 22, October, 213–32. As reprinted with minor additions in Patinkin (1981), 289–308.

Patinkin, D. 1979b. The development of Keynes' policy thinking. In *Theory for Economic Efficiency: Essays in Honor of Abba P. Lerner*, ed. H.I. Greenfield et al., Cambridge, Mass.: MIT Press, 151–66. As reprinted with minor changes and a Postscript in Patinkin (1982), ch. 8.

Patinkin, D. 1980. New materials on the development of Keynes' monetary thought. *History of Political Economy* 12, Spring, 1–28.

Patinkin, D. 1981. *Essays On and In the Chicago Tradition*. Durham, NC: Duke University Press.

Patinkin, D. 1982. Anticipations of the General Theory? And Other Essays on Keynes. Chicago: University of Chicago Press.

Patinkin, D. 1984. Keynes and economics today. *American Economic Review* 74, May, 97–102.

Patinkin, D. and Leith, J.C., eds. 1977. *Keynes, Cambridge and the General Theory: The Process of Criticism and Discussion Connected with the Development of the General Theory*. London: Macmillan.

Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32, November, 38–65. As reprinted in American Economic Association, *Readings in Monetary Theory*, Philadelphia: Blakiston for the American Economic Association, 1951, 162–83.

[Robinson, E.A.G.] E.A.G.R. 1936. Mr Keynes on money. *The Economist* 122, 29 February, 471–2.

Robinson, E.A.G. 1947. John Maynard Keynes, 1883–1946. *Economic Journal* 57, March, 1–68. As reprinted in Lekachman (1964), 13–86.

Robinson, J. 1933a. *The Economics of Imperfect Competition*. London: Macmillan.

Robinson, J. 1933b. A parable on savings and investment. *Economica* 13, February, 75–84.

Robinson, J. 1952. *The Rate of Interest: And Other Essays*. London: Macmillan.

Robinson, J. 1969. *Introduction to the Theory of Employment*. 2nd edn, London: Macmillan.

Rymes, T.K., ed. 1988. *Keynes' Lectures 1932–35: Notes of a Representative Student*. London: Macmillan.

Samuelson, P.A. 1946. Lord Keynes and the General Theory. *Econometrica* 14, July, 187–200. As reprinted in Lekachman (1964), 315–31.

Schefold, B. 1980. The General Theory for a totalitarian state? A note on Keynes's Preface to the German Edition of 1936. *Cambridge Journal of Economics* 4, June, 175–6.

Shackle, G.L.S. 1967. *The Years of High Theory: Invention and Tradition in Economic Thought*. Cambridge: Cambridge University Press.

Skidelsky, R. 1983. *John Maynard Keynes*. Vol. 1, *Hopes Betrayed 1883–1920*. London: Macmillan.

Skidelsky, R. 1992. *John Maynard Keynes*. Vol. 2, *The Economist as Saviour, 1920–1937*. London: Macmillan.

Skidelsky, R. 2000. *John Maynard Keynes*. Vol. 3, *Fighting for Britain, 1937–1946*. London: Macmillan.

Stein, H. 1969. *The Fiscal Revolution in America*. Chicago: University of Chicago Press.

Targetti, F. and Kinda-Haas, B. 1982. Kalecki's review of Keynes' *General Theory*. *Australian Economic Papers* 21, December, 245–53. (A translation of Kalecki (1936).)

Thirwall, A.P., ed. 1978. *Keynes and Laissez-Faire*. London: Macmillan.

Tinbergen, J. 1939a. *A Method and Its Application to Investment Activity*. Vol. I of *Statistical Testing of Business Cycle Theories*, Geneva: League of Nations, Economic Intelligence Service.

Tinbergen, J. 1939b. *Business Cycles in the United States of America 1919–1932*. Vol. II of *Statistical Testing of Business Cycle Theories*, Geneva: League of Nations, Economic Intelligence Service.

Viner, J. 1936. Mr. Keynes on the causes of unemployment. *Quarterly Journal of Economics* 51, November, 147–67. As reprinted in Lekachman (1964), 235–53.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

Viner, J. 1964. Comment on my 1936 review of Keynes' *General Theory*. In Lekachman (1964), 253–66.

Walker, D.A. 1985. Keynes as a historian of economic thought: the biographical essays on neoclassical economists. *History of Political Economy* 17, Summer, 159–86.

Walker, D.A. 1986. Keynes as a historian of economic thought: the perspectives of the *General Theory*. In *Research in the History of Economic Thought and Methodology*, ed. W.J. Samuels. Greenwich, Conn.: JAI Press, Vol. 4, 1–36.

Walras, L. 1926. *Eléments d'économie politique pure*. Edition définitive. Paris: F. Pichon. As translated by William Jaffé as *Elements of Pure Economics*, London: George Allen & Unwin, 1954.

Weintraub, S. 1961. *Classical Keynesianism, Monetary Theory and the Price Level*. Philadelphia: Chilton.

White, E.W. 1974. Keynes – architect of the Arts Council. In *Keynes: Aspects of the Man and His Work*, ed. D. Moggridge. London: Macmillan, 22–48.

Wicksell, K. 1898. *Interest and Prices: A Study of the Causes Regulating the Value of Money*. Trans. R.F. Kahn, London: Macmillan, 1936.

Wicksell, K. 1906. *Lectures on Political Economy*. Vol. II: *Money*. Translated from the original Swedish by E. Classen, ed. L. Robbins, London: Routledge, 1935. (Second Swedish edition, 1911; English translation from posthumous 1938 Swedish edition, ed. Emil Sommarin.)

Wicksell, K. 1907. The influence of the rate of interest on prices. *Economic Journal* 17, June, 213–20.

Williamson, J. 1983. Keynes and the international economic order. In *Keynes and the Modern World*, ed. D. Worswick and J. Trevithick. Cambridge: Cambridge University Press, 87–113.

Winch, D. 1969. *Economics and Policy: A Historical Study*. London: Hodder & Stoughton.

Wood, J.C., ed. 1983. *John Maynard Keynes: Critical Assessments*. Vol. II, London: Croom Helm.

Government publications

[Great Britain.] 1931. Committee on Finance and Industry (Macmillan Committee). *Report*. London: HMSO.

[Great Britain.] 1941. An Analysis of the Sources of War Finance and an Estimate of the National Income and Expenditure in 1938 and 1940. Parliamentary Papers, Cmd. 6261. London: HMSO.

Unpublished works

Hudson, R.J.S. 1950. Towards a bibliography of John Maynard Keynes. Ms. in the Library of King's College, Cambridge.

How to cite this article

Patinkin, Don. "Keynes, John Maynard (1883–1946)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000016> doi:10.1057/9780230226203.0889

The New Palgrave Dictionary of Economics Online

Keynes, John Maynard (new perspectives)

Roger E. Backhouse and Bradley W. Bateman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since Don Patinkin's article on Keynes appeared in the first edition of *The New Palgrave*, two major biographies have increased our understanding of Keynes's life. There has developed a large literature on the relation between his work on probability and his involvement in the Bloomsbury group, and the relation to both of these to his economics. This article reviews and assesses that literature along with more recent work on his economics.

Keywords

aggregate demand; behaviourial economics; Clarke, P. F.; expectations; forced saving; gold standard; Harrod, R. F.; induction; inflation; International Monetary Fund; Keynes and Bloomsbury; Keynes and philosophy; Keynes, J. M.; liberalism; liquidity preference; Meade, J. E.; Mill, J. S.; Moggridge, D.; Moore, G. E.; New Keynesian macroeconomics; Ohlin, B. G.; Patinkin, D.; Pigou, A. C.; Post Keynesian economics; probability; psychology and economics; public works; Ramsey, F. P.; Robertson, D.; rules of conduct; Skidelsky, R.; statistics and economics; Stone, J. R. N.; subjective probability; uncertainty; unemployment; utilitarianism; utility maximization

Article

Patinkin on Keynes

Don Patinkin's article on John Maynard Keynes, reproduced here from the first edition of *The New Palgrave*, is a classic. Patinkin had first grappled with Keynesian theory as a student in Chicago in the 1940s, going on to write *Money, Interest and* Prices (1956), which was the leading graduate textbook in macroeconomics from the 1960s until the early 1970s. Apart from offering a theoretical interpretation of Keynesian economics, as the economics of disequilibrium, the book contained detailed appendices on the history of many of the concepts with which he was working. When the publication of the volumes of

Keynes's *Collected Writings* (cited here as *JMK*) made available extensive correspondence and other previously unpublished material, Patinkin turned to this to provide an account of how Keynes had reached the ideas that he (Patinkin) and others had struggled so long to understand. Patinkin's scholarship was meticulous, and his ideas on Keynes were documented in a series of books published between 1974 and 1982. It was a case of one outstanding theorist exploring the mind of another, which explains the questions he chose to ask: how did the Marshallian quantity theorist of the early 1920s become the author of the *General Theory*? What were the theoretical innovations that marked the *General Theory* apart from works by others seen as having ‘anticipated’ Keynesian ideas? Patinkin's article sums up the results of this thinking, which explains why, at the start, he wrote that he was not going to offer a biography but would reflect on the development of Keynes's thinking on economic theory and policy.

When Patinkin made that remark, the only biography available was the official biography by Sir Roy Harrod (1951). The first volume of Skidelsky's biography (1983) had appeared, but it covered only the period up to 1920, a long way short of the work for which Keynes is now famous. Since then, Skidelsky has concluded his biography with two further volumes (1992; 2000) and has produced a one-volume abridgement, of a mere 1,000 pages (2003). The co-editor of the *Collected Writings*, Don Moggridge, has also written an important biography (1992: for two useful comparisons that explain why both need to be read, see Dimand, 1993, and Blaug, 1994). Apart from these, Keynes has been the subject of an earlier short biography by Moggridge (1976) as well as ones by Charles Hession (1984), David Felix (1999), and a short study by Skidelsky (1996). There has also been extensive research on the Keynesian revolution and Keynes's role in it, some prompted by Patinkin's conclusions, others by other concerns. Here, the work of the Peter Clarke (1988; 1998) is worth noting, being the result of a political historian taking the time to get to grips with circumstances that produced Keynesian economics.

A further development was that in the mid-1980s scholars were beginning to see Keynes not simply as an economist but as a philosopher, exploring the philosophical grounding of his King's College Fellowship Dissertation which later became the *Treatise on Probability* (1921; *JMK*, VIII) and the relationship of these ideas to his discussion of uncertainty in the *General Theory* (notably Chapter 12) and in the article in 1937 where he argued that his main thesis was the fact that we know virtually nothing about the future (*JMK*, X, pp. 108–23). The appearance, within a very short period, of three studies of Keynes's philosophical development (Carabelli, 1988; O'Donnell, 1989; Bateman, 1987; 1988; 1996) was an important factor behind the rapid growth of a very detailed ‘Keynes and philosophy’ literature during the late 1980s and early 1990s. The first volume of Skidelsky's biography also located Keynes, much more firmly than Harrod's had done, in the artistic and literary environment of the Bloomsbury group, stimulating further reassessments of the context in which Keynes's economics should be interpreted.

During the 1970s, the foundations were laid for a transformation in macroeconomics by the work of Phelps, Lucas, Sargent, Barro and others. Keynes was reinterpreted, many times, and the theoretical framework within which Patinkin had worked ceased to define the questions that became of primary concern to macroeconomists, which now concerned problems such as dynamics, expectations, and strategic decision-making. Though Keynes still had some relevance (witness the New Keynesian macroeconomics), it became much easier to see Keynesian economics as a historical episode than had been the case for members of Patinkin's generation. Studies such as Dimand (1988), Young (1987),

Meltzer (1988), Littleboy (1991) and Laidler (1999) reflected this new perspective: it had become much easier to see Keynes apart from the theoretical context that Patinkin and his contemporaries had developed.

Keynes's life – Bloomsbury, art and economics

When Patinkin wrote his account of Keynes's economics, he had two contrasting accounts of Keynes's early life on which he could have drawn had he considered it relevant to his task. Harrod's biography had been constrained not only by older Victorian conventions about biography but also by his view that it was important that nothing should be revealed about Keynes that would threaten the acceptance of Keynesian economics. He thus ignored Keynes's homosexuality and minimized Keynes's involvement with Bloomsbury. In contrast, Skidelsky's first volume presented a Bloomsbury Keynes. His was a psychological biography in the Bloomsbury tradition, in which Keynes's sexuality and personal relationships were to the fore. Where Harrod had portrayed Keynes's sense of duty – the famous ‘presuppositions of Harvey Road’, combining a strong sense of duty and a belief in the power of social science – as constraining any immoralism, Skidelsky emphasized the influence on Keynes of G. E. Moore, mentor to the group of Cambridge undergraduates who would ultimately form the male corps of the Bloomsbury group. Skidelsky attached much more importance than Harrod to Keynes's essay ‘My Early Beliefs’ (JMK, X, ch. 39, first written in 1938), creating a picture of someone dominated by a very private set of values, significantly insulated in his youth from the world of public affairs into which he was later thrust. The story ended in 1919, with the book that Keynes wrote when he resigned from the Treasury in protest over the impending outcome of the Paris peace negotiations, *The Economic Consequences of the Peace* (JMK, II). Not only did this thrust Keynes into the public arena, but it marked the shift from a Victorian belief in automatic economic progress to a world in which prosperity would need to be fought for. Skidelsky writes of Keynes fearing not the material but the organizational and moral destruction wrought by the war: prior to the war, it had been liberating for Keynes and his friends to be freed from their parents’ belief that God could be relied on to maintain the social order. After the war, that was no longer the case. Thus Skidelsky (1983, p. 402) concluded his first volume, by arguing that ‘In the last resort Keynes's post-war fear of the future of capitalism was profoundly influenced by the Victorian fear of a godless society’. The prospect of civilisation briefly opened up by Moore's *Principia Ethica* (1903) had receded over the horizon. The rest of Keynes's life was spent in trying to bring it back into sight. *Economic Consequences of the Peace*, Skidelsky (1983, p. 384) claimed, was Keynes's ‘best book’, in which, more than in any other, he brought ‘*all*’ his gifts to bear on the subject in hand. Harrod could never have raised doubts about the merits of the *General Theory* in this way.

Subsequent work has continued the process, begun with Skidelsky's volume 1, of correcting many of the errors in Harrod's biography. These were significant because Harrod had used his sources selectively: he was ‘a master of selective quotation’, aiming to cleanse Keynes's statements for public consumption. For example, in the quotation from a 1905 letter, ‘I want to manage a railway or organise a Trust …’, the omitted words denoted by the dots were ‘or at least swindle the investing public’, a sentiment that Harrod would not have wished his readers to encounter (Skidelsky,1983, p. xviii). Harrod had carefully not discussed Keynes's views on conscientious objection (where he supported his Bloomsbury friends),

even though surviving documents clearly attest to his position (Johnson, 1960). Neither had he discussed the ‘search for love’ that Keynes himself described as his main preoccupation before 1908, now carefully documented (‘boy by boy’, according to Keynes's own records) by Moggridge. Both Skidelsky and Moggridge explain that Keynes's search for love eventually was calmed for several years through his intimate relationship with the painter Duncan Grant that lasted from 1908 to 1912 and that would become a lifelong friendship. Moving beyond this sexual biography, however, both Skidelsky's subsequent volumes and other work such as Moggridge's biography have served to narrow the gap between the pictures of ‘Keynes the economist’ and ‘Keynes the member of Bloomsbury’. Thus, although as late as 1921 Keynes was expressing doubts about his vocation (the values of Bloomsbury were still important to him and he lamented the lack of a true artistic talent), his turn to economics had deep roots: though his first formal training was after graduation, when preparing for his Civil Service examination, he came to this having studied a formidable list of books on economics. His philosophical views, discussed separately below, also serve to bridge the two interpretations.

The early 1920s were a crucial period for Keynes in several respects. *The Economic Consequences of the Peace* had made him a celebrity, and given him the financial basis for his earliest speculative activities (which came to grief in 1920 when the currencies on which he had speculated moved in the wrong direction). His Treatise on Probability (1921; JMK, VIII) was published, although he left behind any ideas of further work in philosophy. Though he had left the Treasury (something he had planned to do even had he not resigned in protest) he chose not to resume his heavy pre-war teaching duties in Cambridge, and spent much time offering policy advice from the position of an outsider to the Whitehall system. In giving up his income as a university lecturer at Cambridge and his income as a fellow of King's College following the First World War, Keynes now needed new streams of income to support his lifestyle. Journalism and attempts to influence public opinion became major activities. He also embarked on a long-term career as an investor, losing his investments (and that of his friends) through commodity speculation when the post-war boom collapsed, but then began to rebuild them again. There was also the change in his private life. To the horror of his Bloomsbury friends, he fell in love with a ballerina from Diaghilev's company, Lydia Lopokhova, whom he eventually married in 1925. In the opinion of his family and many of his friends, this gave him a new lease of life and may, at least in part, account for the enthusiasm with which he was willing to entertain and explore new ideas. Their relationship (Hill and Keynes, 1989) clearly refutes the allegation that Keynes's emphasis on the short run stemmed from a belief that he would not have children.

With his marriage to Lydia, he became more distant from Bloomsbury, though contacts remained strong, even though Lydia and Vanessa Bell were never on good terms. Keynes (as Moggridge has shown) certainly found relaxation from his hectic schedule as a professional economist amongst his Bloomsbury friends, both in London and in the houses they acquired in the Sussex countryside. He also built professional and social contacts outside Bloomsbury. His experience as an investor grew, and his position as a director of the Provincial Insurance Company as well as Chairman of the National Mutual Insurance Company gave him regular contact with the City. He made money during the 1920s, but lost it a second time in 1929, and had a further setback in 1937. His speculative activities were *not* a success, though he regularly extricated himself from disaster. At one point he had to take delivery of a contract for wheat because the price had fallen too low for him to wish to sell it. Though he went so far, on one occasion, as to estimate the capacity of King's College Chapel to serve as a place to store wheat, he never had to try to use it for that purpose because he ingeniously realized that he could forestall the

delivery of the wheat by demanding that it be cleaned before delivery. When he eventually established his fortune, it was through a very different strategy: of holding, long term, a small number of stocks, in companies that he understood and in which he had confidence.

Keynes had suffered from recurrent bouts of appendicitis and influenza while in the Treasury during the First World War, both problems to which he was made vulnerable because of overwork and exhaustion. In 1937 he had a major heart attack and, though he was gradually brought back to health, the last decade of his life, as demanding as any in his career, were plagued by heart problems. However, recent work, notably by Craufurd Goodwin (1998; 2006) has argued that, despite this broadening out beyond Bloomsbury, strong intellectual links can be found between Keynes, Roger Fry, and the artists and novelists of Bloomsbury.

These intellectual links bear on Keynes's economics in both broad and narrow ways. One of the links that Goodwin uncovers is a consistent concern throughout the works of the Bloomsbury novelists (such as Virginia Woolf, E.M. Forster, David Garnett), art critics (such as Roger Fry, Clive Bell), and political theorists (such as Leonard Woolf) with the problem of inconsistent and disappointed expectations, a theme that would run through the *General Theory*. Another shared theme with his fellow Bloomsburies was a fascination with the emerging study of psychology and the varieties of human motivation. Again, it is difficult to read the *General Theory* without realizing the extent to which Keynes was decades ahead of the discipline in using psychological explanations of economic behaviours such as investment and consumption. And this points to another area of shared interest amongst the Bloomsburies crucial to understanding Keynes's mature theoretical work: like their mentor G.E. Moore, they completely eschewed utilitarian explanations of human behaviour and did not take utilitarian ethics as a reasonable guide for behaviour or policy.

Keynes and philosophy

Beginning in his years at Eton, Keynes demonstrated a keen interest in political and philosophical questions. As an elected member of College Pop, a debating club, Keynes frequently spoke in support of Liberal positions. In his final year, he wrote a long essay on the poet and monk Bernard of Cluny, which he would revise and read again several times later in his life. The central fascination of Bernard for Keynes was the tension between the following a path of contemplation and the path of active engagement with the world.

The skills that Keynes had begun to develop at Eton came into full bloom when he matriculated to Cambridge in the autumn of 1902. At the centre of his many engagements at Cambridge was his membership in the secret society known as the Apostles, one of Cambridge's most distinguished societies at the time, which the years immediately preceding Keynes's matriculation had begun to develop a concern with philosophical questions. Both Bertrand Russell and G.E. Moore had been active in the society in the years before Keynes's election and on occasion both would still attend the group's Saturday evening meetings. Moore's classic, *Principia Ethica*, was published at the end of Keynes's first year and it became the most important single text for Keynes in his undergraduate years. ‘[I]ts effect on *us*, and the talk which preceded and followed it, dominated, and perhaps still dominate, everything else’ (*JMK*, X, p. 435).

For several reasons, this early devotion to Moore and philosophy was still not generally understood as a part of Keynes's life and work when Patinkin wrote his entry on Keynes. On the one hand, as mentioned

above, Harrod chose to minimize this influence on Keynes's work because he feared that it would point toward his sexuality and so risked diminishing the status of his economic work. On the other hand, when the decisions were made to publish Keynes's *Collected Writings*, the editors concluded that, apart from including his *Treatise on Probability* so as to have all his published books in the edition, no effort would be made to include his philosophical correspondence or any of the early philosophical essays written for the Apostles. Following Harrod, the Keynes of the *Collected Writings* was to be an economist only. Thus, when Patinkin undertook his project he would have had little or no idea that this philosophical work even existed. As the early volumes of the *Collected Writings* began to appear, there was a flood of new material to absorb, and little or no indication of this additional trove of untapped material. By the late 1980s, however, as scholars began to explore the Keynes Papers deposited in the King's College Modern Archive, the importance of this philosophical material began to become clear. The times were perhaps propitious for such a discovery since macroeconomics in the 1980s had taken a sharp turn to questions of probability and uncertainty, and much of Keynes's early philosophical work involved the philosophy of probability. The discovery of this previously unexplored part of Keynes's work offered the possibility to see what Keynes himself had said about uncertainty and expectations, a topic that had become central to many critiques of his work (for example, rational expectations).

Two early Cambridge University dissertations on the topic by Anna Carabelli and Roderick O'Donnell opened this field, and became the basis, if somewhat altered, of Carabelli (1988) and O'Donnell (1989). These two early interpreters took diametrically opposed approaches to Keynes's work in probability, with Carabelli arguing that Keynes had authored a subjective theory of probability while O'Donnell argued that, quite to the contrary, Keynes had authored an objective theory of probability. Both agreed that Keynes's contribution hinged on his articulation of a logical theory of probability, in which Keynes argued that probability represented the logical degree of belief in a proposition rather than a frequency distribution of outcomes. This much, of course, was indisputable, but the question remained of how to interpret what Keynes had said. Was the logical relation an objectively known (and identical) value for all rational persons with the same knowledge, or was the logical relation subjectively known, unique for each individual?

The source of the possible confusion came from the well-known critique of Keynes's work in probability made in the 1930s by the great Cambridge philosopher Frank Ramsey (1931). Ramsey argued that there were no such things as Keynes's logical relations of probability, or that he at any rate could not identify them; rather, Ramsey argued, we form our own subjective probabilities, subject only to the consistency required of them by the Dutch book, gambling argument (that they should not be willing to accept combinations of bets that guarantee that they will lose money). Ramsey's work was published posthumously, after his tragic early death, and Keynes's review of it seems to make clear his acceptance of Ramsey's criticism. Thus, after recapping Ramsey's criticism, Keynes would report, ‘So far I yield to Ramsey – I think he is right’ (*JMK*, X, p. 339). Against this, Carabelli seemed to argue that Keynes had always had a subjective theory of probability, and that Ramsey's criticism amounted to an argument over whether the subjective probabilities represented a logical entity. O'Donnell, on the other hand, argued that Keynes had held an objective theory of probability in *Probability*, and that he had never capitulated to Ramsey.

In Bateman (1987; 1988; 1996) a very different argument is made that takes at face value both Keynes's statements about objectivity in *Probability*, as well as his capitulation to Ramsey's critique. This work approaches Keynes initial position through the same route that he took himself, his engagement with and

critique of Moore's ethical theory. In *Principia Ethica*, Moore had made one argument that Keynes did not accept, namely, that in a person should follow the general rules of conduct (for example, do not murder, do not steal, do not commit sodomy) because of the uncertainty of the outcome of one's actions. Keynes began his critique of this position while he was a second-year undergraduate in one of the earliest papers he read to the Apostles; and he developed the insights in that paper into his fellowship dissertation (submitted in 1909) and eventually into *Probability* (*JMK*, VIII).

The crux of Keynes's critique of Moore's ethical position was that he had defined probability incorrectly. In his argument, Moore depended on a frequency theory of probability and assumed the impossibility of knowing long-run frequencies with any certainty. In the face of this radical uncertainty about the possible outcomes of committing murder or committing adultery, for instance, Moore argued that the best course of action was represented by the general rules of conduct, which he argued gave the highest frequency of good outcomes. Moore felt that, since we would not know when an act of murder might turn out to have a good outcome, the best course of action was not to murder at all.

For Keynes, who was a practising homosexual and who generally enjoyed the youthful pleasure of making up his own mind about when rules were reasonable (and when not), this argument needed to be proven wrong. His method was to posit that probabilities are logical relations that are, indeed, capable of being known when we act. His argument had the added twist of drawing on Moore's Platonic treatment of the good. Like Moore, who argued that we know the good through intuition of an indefinable quality, Keynes argued that the logical relations of probability are Platonic entities, not reducible to anything else, and known through intuition. On this argument, any bright young man with the same knowledge could intuit the probabilities of the various possible outcomes of an action, as well as the amount of goodness that would attach to each one. On this argument there was no need to follow traditional rules of conduct. The argument seemed to be persuasive, as Moore dropped his argument for following rules in *Ethics* (1912) and included a logical theory of probability in the same book.

But it was exactly the idea that probability was a Platonic entity that Ramsey criticized. He had written about this as early as 1922 in a review of *Probability* in which he talked about fog-shrouded mountains which were not visible to the human eye. And his posthumously published essay, ‘Truth and Probability’ (Ramsey, 1931) says simply that, if these logical relations exist, he is unable to recognize them and certainly does not act on them. Perhaps the most convincing explanation of Keynes's ultimate position comes from Donald Gillies and Grazia Ietto-Gillies (1991) who describe Keynes as embracing ‘intersubjective’ probabilities. Gillies and Ietto-Gillies accept Keynes's capitulation to Ramsey but go on to point out that the positive result of this in the *General Theory* was a world in which most people formed their subjective probabilities by guessing what others were thinking. Thus in Chapter 12 of the *General Theory*, in Keynes's well-known description of how stock markets function, investors seek stability in an uncertain world by depending on the mass psychology of the market. This same idea is borne out in his description elsewhere in his magnum opus in his description of liquidity preference and the ways that bond traders make their portfolio decisions. This entire line of thought is most strongly emphasized in Keynes's famous restatement of his book's argument in response to his critics in his 1937 article in the *Quarterly Journal of Economics*.

There are many dimensions to Keynes's early work in ethics, probability, and political philosophy, but perhaps the most crucial for understanding his work in relation to mainstream economics as it evolved in the 20th century is his refusal to embrace utilitarianism. Bateman (1988) and Goodwin (2006) have noted this, but the philosopher Tom Baldwin (2006) has put it in a form that makes it particularly

striking for an economist. This comes in his observation that, despite his place at the centre of a long tradition of British liberal political philosophers, Keynes never seriously engaged John Stuart Mill's work during his long and prolific career. There is perhaps no Briton in the first half of the 20th century who so acutely explored the meaning of liberalism as John Maynard Keynes. Both in his economic work on the appropriate role of the state in the economy and his essays on political liberalism (reprinted in JMK, IX), Keynes explored the frontier questions of the autonomy of the individual, the possibilities for human freedom, and the role of the state. And yet, as Baldwin makes clear, Keynes's acceptance of Moore's argument that Mill had misidentified utility as good kept Keynes from ever seriously tackling Mill's work. For much of the 20th century, this meant a kind of deep misunderstanding and misapprehension of Keynes's work on the part of those economists who were unable to work outside a framework of individual utility maximization as both a positive description of human behaviour and a normative goal of policy. Much of Keynes's writing and his modelling are opaque to mainstream neoclassical economists for this reason. Perhaps with the recent shift to behaviourial economics Keynes's work will seem less difficult to understand and appreciate on its own terms. But, whichever is true, much of the time since the publication of the *General Theory* has been marked by a deep chasm caused by the fact that Keynes did not assume that people seek to maximize utility in most situations, or that they should.

Keynes's economics

Where Skidelsky painted a brilliant portrait of the economist as operating in a broader intellectual environment, encompassing not simply Cambridge economics but also the philosophical and political concerns of the Apostles and Bloomsbury, Moggridge sought to redress the balance with what his subtitle described as ‘an economist's biography’ (cf. Moggridge, 2002). Keynes's complete dismissal of utilitarianism distanced him from much economic theory, but Keynes was an original thinker of the first rank, especially in areas such as international finance and monetary economics where utilitarian thinking has always had a limited influence. Thus, what emerges from Moggridge's account is an economist who, to an extent greater than implied by Patinkin, with his focus on theory, was primarily an applied economist whose career was *dominated* by issues of international finance (see Mundell, 2008). Though politicians may not have always accepted his arguments, throughout his career he had the ear of governments and his ideas were avidly sought. Sometimes he was invited to serve on committees, but in other cases he sent a memorandum to the Treasury that caused him to be brought into the discussion. He may have been an outsider, but even in the 1920s he had considerable access to officials and government ministers. Thus he was never an economic theorist as the term is now understood, but was for ever analysing institutions, estimating rough magnitudes and formulating policy proposals on the basis of those estimates. When, during the Second World War he acted as mentor to James Meade and Richard Stone in constructing the British national accounts, this came after a lifetime of promoting the collection of economic statistics, a passion that he developed as a young man when he developed his ideas on probability to include an extended treatment of induction (Bateman, 1990).

His first appointment in the Civil Service was in the India Office, but it was after he had left and returned to Cambridge that, in 1913, he was invited to sit on the Royal Commission on Indian Finance and Currency. Several of his fellow commissioners worked in the Treasury or had done so recently, and

so his distinguished service on this appointment helped him to make the contacts that eventually would lead to his appointment to the Treasury during the First World War. Keynes's work on the Royal Commission was a harbinger for much of his later government service; he was appointed to the Commission largely on the basis of the proofs of his book *Indian Currency and Finance* that he was circulating in 1913. The book appeared within a month of the start of the Commission's work and Keynes used the analytical framework of the book to drive his questioning and to bring the other Commission members around to his own views.

This pattern of writing, demonstrating expertise and being invited into the inner circles of policymaking, would repeat itself over and over again during his career. It was the general pattern of his next assignment advising the government, when he entered the Treasury in January 1915 as an adviser to Sir George Paish, a special assistant to the Chancellor of the Exchequer, David Lloyd George. In summer and autumn 1914 Keynes had been asked his opinion on the crisis that initially rocked the London financial markets at the outset of the war. Using the information he had gained while consulting at the Treasury, he wrote and published articles that autumn in the *Economic Journal* and the *Quarterly Journal of Economics* explaining what had happened when the joint stock banks had, in his opinion, unnecessarily called loans and so restricted credit. These two pieces caused some consternation in the banking community, but they won him more invitations to write memoranda for the Treasury and eventually led to his appointment. Once inside the Treasury, Keynes moved through several committee assignments and served on inter-ally financial working groups to help keep the finances of France, Britain, and Russia coordinated and well functioning; after America's entry into the war in 1917, the group of allies expanded. Keynes rose to a position of considerable stature for such a young man, being made the head of A Division, the senior person in charge of Britain's external financial relations during the war.

As the war wound down, Keynes was asked to write a memorandum on German indemnity and the limits of what Germany could be expected to pay. This in turn led to more committee assignments and his eventual appointment to the Treasury team that was sent to Paris for the peace negotiations, where he served as the senior Treasury official in Paris. With this pattern of mastering his brief quickly and advancing in the Treasury well established, Keynes became the lead financial negotiator for Britain in many circumstances during the negotiations. The work was dispiriting to him because of the political machinations and the lack of any goodwill towards Germany. He seemed to see only avarice, revenge and political gain as the means and end of the negotiations, with little or no concern for the starving peoples of Europe. Keynes was also working to exhaustion in poor conditions in Paris, and eventually he felt compelled to submit his resignation. In June 1919, he left Paris unable to see any good in the outcome of the negotiations.

Resignation from the Treasury did not mean an end to his involvement as a policy adviser at the highest level. As early as February 1920, he was advising the Chancellor of the Exchequer on whether to raise interest rates, and was kept informed of discussions in official circles. In 1921, he began his career in journalism, the primary issue initially being reparations, on which international negotiations continued throughout the 1920s. This took him into questions of post-war reconstruction and exchange rates. Though his was not the dominant voice, his arguments were not without influence, as when, in 1924, his views and those of Reginald McKenna helped steer the Chamberlain Committee away from the idea that deflation would be the likely result of a return to gold. In coming to his decision to return to power, Churchill listened to Keynes and had his advisers respond to his arguments. His position in policy circles

was recognized by his being appointed to the Macmillan Committee on Finance and Industry in November 1929, and the newly formed Economic Advisory Council in January 1930, both of which were established to address the unfolding problems of the slump.

It was against this background of extensive journalism and advising on policy that Keynes's economics evolved. His attempt to write a major work on monetary economics started in 1924, whilst he was involved in debates over the return to gold. In the next two years he worked closely with his fellow Cambridge monetary economist, Dennis Robertson, whose *Banking Policy and the Price Level* (1926) strongly reflected his discussions with Keynes; but Keynes's own book, *A Treatise on Money* (1930; *JMK*, V and VI) did not appear for several years, its final drafting coming while Keynes's life was dominated by the Macmillan Committee, in whose deliberations Keynes played a dominant role, and by the attempt by the economists on the Economic Advisory Council to produce a unanimous report on measures to combat the slump. From November 1929 to April 1931, Keynes attended over 100 meetings of the Macmillan Committee alone, this at the time when he was preparing his *Treatise on Money* for publication.

Though Keynes's writings early in the 1920s were concerned with exchange rates and inflation, and were conducted within an essentially Marshallian framework, the basis for his advice changed. In the *Tract on Monetary Reform* (1923) he had argued for the importance of monetary management, but by 1925 he was focusing on the overvaluation of sterling involved in returning to gold at $4.86 to the pound. In 1924 he began to write specifically on unemployment, asking whether unemployment needed a drastic remedy, and this became an increasingly important theme in his writings as the decade went on, taking him into issues ostensibly far from international finance, such as the need for restructuring in the cotton industry. However, international monetary issues never went away. Of particular significance, given that it revealed his failure at this time to attach importance to the income effects central to the *General Theory*, was his 1929 exchange with Bertil Ohlin on the transfer mechanism, still in the context of reparations. Also, his policy recommendations concerning unemployment, though focused more on domestic issues, were never far from questions of international finance. In 1924, his argument for action against unemployment rested on arguments about diverting resources ‘from relatively barren foreign investment into state-encouraged productive enterprises at home’ (quoted in Moggridge, 1992, p. 421). In his writings later in the 1920s, the possibility of raising employment became more prominent, especially in the confrontation with the Treasury view in 1929, but the background was the constraint imposed by the restored gold standard. It was the fact of the restored gold standard that led Keynes, in early 1931, to support a limited introduction of tariffs.

The story of Keynes's theoretical development, discussed in detail by Patinkin, took place against this background. The *Treatise*, with its Wicksellian approach, taking Keynes away from the Marshallian framework of his earlier work, was written at a time when his public commitments left him precious little time for more academic pursuits. For example, Hawtrey offered detailed criticisms of drafts of the *Treatise*, but Keynes did not have time to read them until after publication. This may account for the unsatisfactory nature of the resulting book and hence the rapidity with which he moved on. Furthermore, when sterling was allowed to float in 1931, the case for public works made in the *Treatise* was no longer relevant, and, now that the need to defend sterling was removed, Keynes began to argue for low interest rates. Though he remained active in both journalism and policy advice, he made sure that what was to become the *General Theory* was subject to much more systematic academic criticism from his colleagues in Cambridge.

The process whereby Keynes made the crucial step taking him from the ‘classical’ framework of the *Treatise on Money* to the *General Theory* has been examined in great detail. Evidence for this comes from his publications, public statements on policy, recollections of scholars who visited Cambridge, and students’ notes of his lectures. Patinkin's conclusion was that Keynes reached a ‘full understanding’ of the ideas in the book some time in 1933: not until then did Keynes have in place the three crucial elements (the multiplier, equilibration through changes in output, and an unemployment equilibrium). Some scholars have supported this conclusion; others have argued for dates in 1932 and 1934 (for a concise survey see Skidelsky, 1992, pp. 443–4, including note 48; see also Clarke, 1998, ch. 4). What is of interest there is as much the reasoning underlying such arguments as the conclusions themselves. For Patinkin, an idea existed when it could be written down as a model. On the other hand, for Clarke, the intuition was more important, a conclusion endorsed by Skidelsky (1992, p. 444, quoting student lecture notes), who has pointed to Keynes's belief that one can think ‘accurately and effectively’ even before being able to formalize ideas. A further dimension is the weight to attach to practical theorizing in the context of policy arguments in relation to theorizing at a more abstract level. There is also the issue of whether an idea has to have been conceived in someone's mind, written down privately, or placed in the public domain. Thus it is because he believes that Keynes did not confine himself to what followed rigorously from the *Treatise* framework that Clarke (1998, pp. 92–5) is willing to date Keynes's understanding of aggregate demand to the summer of 1932: not only were his lectures that autumn significantly different from those the previous session (so that students well versed in the *Treatise* found them hard to follow), but by November 1932 he was writing about the contrast between Malthus and Ricardo in ways that clearly anticipated the break with Ricardian orthodoxy found in the *General Theory*. The Second World War took him into the Treasury again, though as an unpaid adviser rather than a salaried official, where he had a possibly unparalleled combination of access to officials and freedom to pursue whatever he thought important. Once again, it was his brilliant writing that helped him to be called into the Treasury: *How to Pay for the War* (1940; *JMK*, IX), perhaps his clearest success in the policy arena, providing the theoretical and statistical framework for controlling wartime inflation. Though this has been seen as an application of Keynesian theory to problems of inflation at full employment, it is worth noting that it relies more on the doctrine of forced saving, discussed by other economists in the years before the *General Theory*, a fact that acquires significance given that this is the theory of inflation first used by Milton Friedman (Laidler, 2002, pp. 103–4). However, because the path from Keynesian ideas to their implementation in the 1941 budget was relatively smooth, this did not serve as a drain on Keynes's time and health in the same way as his international negotiations. Here he was up against the Americans who, for a variety of reasons, were much harder to convert. Isolationist tendencies were strong in American politics, and there was widespread hesitation about using American resources to support a country that was still the centre of a worldwide empire. There was even a belief in some circles that a contributory factor behind America's over-expansion of 1928–9 and hence the subsequent collapse had been pressure on the Federal Reserve to keep interest rates low, due to the pressure on sterling after the return to gold in 1925. Britain was bankrupt and the burden of negotiating American financial support under difficult conditions fell substantially on Keynes. Given this weak bargaining position, it is thus not surprising that, in negotiating the post-war economic order, he failed to persuade the Americans to give the new international monetary authority greater resources to assist countries with balance of payments problems, for they were the ones who would, at least in the early years, be paying. Nonetheless, his success in helping to establish the International Monetary Fund was

significant, given that, after the previous war, the Americans had entered into negotiations only to walk away before agreement was reached.

Conclusions

The debate between Patinkin and Clarke over the exact dating of the birth of Keynes's new ideas in the *General Theory* is a classic in the history of ideas. As Moggridge (1992, p. 559) and Skidelsky (1992, pp. 443–4) have noted, however, there will never be a clear answer to the question of exact dates because of Keynes's lifelong style of work. Keynes always depended on working from fundamental intuitions that he would sketch out, sometimes years ahead, sometimes days ahead, of the final, formalized version of his newest work. Thus Clarke can be correct in the precise dating of Keynes's initial working out of the intuitions that might later become the concepts of aggregate demand and liquidity preference, and Patinkin may be correct in his dating of when Keynes published a theoretically satisfactory account of his theory; neither need be wrong in establishing two important points along the trajectory of one of Keynes's ideas. Although there is clear historical value in fixing the arc of these trajectories in Keynes's work, the real value of the recent work on Keynes to economists does not lie in the exact dating of his various contributions. Rather the value to economists of the work of historians like Clarke, biographers like Moggridge and Skidelsky, and commentators on his life in Bloomsbury such as Goodwin lies in seeing the full range and complexity of his ideas and understanding the nature of the influences that led him to his breakthroughs.

In his entry on Keynes, Patinkin argued that ‘a basic contribution of the *General Theory* is that it is in effect the first practical application of the Walrasian theory of general equilibrium’, and likewise that, ‘The voice is that of Marshall, but the hands are those of Walras’. This reflected Patinkin's own effort in *Money, Interest and Prices* (1956) to capture Keynes's ‘central message’ in the Walrasian general equilibrium framework and the effort throughout all of his historical writing on Keynes to identify the origins of elements of that Walrasian version of Keynes in Keynes's own writings. Whilst this remains a legitimate interpretation, for that is how Keynesian economics was conceived during the Keynesian era, understanding Keynes himself requires paying attention to other possibilities. Thus Leijonhufvud (2006) and others have reinterpreted the Marshallian dimension to his work, arguing that the gap between Keynesian and Walrasian theorizing was deeper than architects of the neoclassical synthesis believed. Rather than seeing Keynes as endorsing a Walrasian interpretation of his work when he responded favourably to the efforts of John Hicks and others to translate his work into simultaneous equation systems, it is better to see Keynes as concerned with his basic intuitions, content for them to be developed in different ways. The intuitions were more important to him than any specific model (Backhouse and Bateman, 2008). This helps justify both his radical statements about uncertainty and his endorsement of analysis that subsequent generations of Post Keynesians have found excessively orthodox and have needed to explain (for references to the Post Keynesian literature on Keynes see King, 2002; Chick, 1983; Dostaler, 2007; Lawlor, 2006; and a significant proportion of the 40 chapters in Harcourt and Riach, 1997). At least in part, this was because Keynes was not interested in the elaboration of his fundamental insights beyond the form in which he needed them for making successful policy arguments. In both the *Treatise* and the *General Theory*, he sought to formalize his insights enough to persuade fellow economists, but his main concern was to provide a workable basis for policy:

he was, in the words of Hoover (2006), primarily a physician creating a ‘diagnostic science’.

Keynes's focus on fundamental assumptions was the approach of the pre-war Apostles, and his intuitions about the economic system arose out of the Bloomsbury understanding that the modern world was built upon a set of inconsistent and easily disappointed expectations. This insight first appears in the *Economic Consequences of the Peace* and suffuses the *General Theory*. It was neither an argument for the impossibility of economic modelling, as some Post Keynesians have argued (Shackle, 1972; 1974; Davidson, 1972), nor was it, as Patinkin argued, of no central importance. Rather, it was a key insight that Keynes built into the *General Theory* and that he believed would have to be a part of any analysis relevant to formulating policies that would help to support the type of civilization that he believed was possible in a humane, well-managed capitalism.

However, though Keynes's involvement in Bloomsbury was fundamental, the intuition about the importance of uncertainty and expectations that he encountered amongst his artistic friends did not carry over into his economics in a simple and straightforward way. Though it animated the *Economic Consequences of the Peace* and the *General Theory*, it was not a feature of his work in the intervening years (Bateman, 1996). In the *Treatise* Keynes denied that expectations were an important factor in explaining the business cycle. When, in response to Keynes's questioning before the Macmillan Committee, Pigou put forward the standard Cambridge trade-cycle theory in which expectations were important, Keynes dismissed the idea, insisting that it was interest rates alone that explained the behaviour of the cycle. To understand the story of how he came, once again, to see the relevance of the old Bloomsbury concern with inconsistent expectations to economic modelling, one has to work carefully through his management of the Provincial bond portfolio (Westall, 1992), his work as bursar of King's College, his policy advice to the government at the time of Britain's abandonment of the gold standard in 1931, and his own investments in the stock market during the Great Depression (Moggridge, 1992).

There is also no straight line from Keynes's early work in the philosophy of probability to the *General Theory*. By the time he came back round to the view that expectations were central to the workings of a capitalist economy, he had abandoned his earlier conception of probability and was left with an adaptation of Ramsey's work in the form of intersubjective probabilities that are shaped by the mass psychology of investors (Gillies and Ietto-Gillies, 1991; Davis, 1994; Gillies, 2006). It is necessary to look at the whole of Keynes's life – as Apostle and Bloomsbury, as student of Marshall, as journalist, government adviser and City investor – to see all the pieces that came together in *General Theory*.

The part of Keynes's philosophical background that perhaps most consistently influenced his economic theorizing throughout his career was not his work in probabilities but his rejection of utilitarian thinking. Keynes never used utility maximizing in a thoroughgoing or consistent way during his long career as an economist. It was, in his view, neither an adequate description of human behaviour nor the desideratum of policy analysis. This, no doubt, is a significant reason why his work has been such a puzzle to economists, leading to misunderstanding of his motivations, modelling, and policy advice. Perhaps in the emerging era of behaviourial economics, this part of his work should seem less puzzling and disturbing. It is not the case, of course, that Keynes foreshadowed work in behaviourial economics, but rather that, like the behavioural economists, he looked for alternative explanations of suboptimal outcomes and behaviours that were clearly not driven by utility maximization.

See Also

- Keynes, John Maynard

Bibliography

Backhouse, R.E. and Bateman, B.W. 2006. *The Cambridge Companion to Keynes*. Cambridge: Cambridge University Press.

Backhouse, R.E. and Bateman, B.W. 2008. Whose Keynes? In *Keynes's General Theory: A Reconsideration After Seventy Years*, ed. R.W. Dimand, R. Mundell and A. Vercelli. London and Basingstoke: Palgrave Macmillan for the International Economic Association.

Baldwin, T. 2006. Keynes and ethics. In Backhouse and Bateman (2006).

Bateman, B.W. 1987. Keynes's changing conception of probability. *Economics and Philosophy* 3, 97–120.

Bateman, B.W. 1988. G.E Moore and J.M Keynes: a missing chapter in the history of the expected utility model. *American Economic Review* 78, 1098–106.

Bateman, B.W. 1990. Keynes, induction and econometrics. *History of Political Economy* 22, 359–79.

Bateman, B.W. 1996. *Keynes's Uncertain Revolution*. Ann Arbor: University of Michigan Press.

Blaug, M. 1994. Recent biographies of Keynes. *Journal of Economic Literature* 32, 1204–15.

Carabelli, A. 1988. *On Keynes' Method*. London: Macmillan.

Chick, V. 1983. *Macroeconomics after Keynes*. Deddington: Philip Allan.

Clarke, P. 1988. *The Keynesian Revolution in the Making, 1924–1936*. Oxford: Clarendon Press.

Clarke, P. 1998. *The Keynesian Revolution and its Economic Consequences*. Cheltenham: Edward Elgar.

Davidson, P. 1972. *Money and the Real World*. London: Macmillan.

Davis, J.B. 1994. *Keynes's Philosophical Development*. Cambridge: Cambridge University Press.

Dimand, R.W. 1988. *The Origins of the Keynesian Revolution*. Aldershot: Edward Elgar.

Dimand, R.W. 1993. Review of Moggridge (1992) and Skidelsky (1992). *Canadian Journal of Economics* 26, 993–9.

Dostaler, G. 2007. *Keynes and his Battles*. Cheltenham: Edward Elgar.

Felix, D. 1999. *Keynes: A Critical Life*. Westport, CT: Greenwood Press.

Gillies, D. 2006. Keynes and probability. In Backhouse and Bateman (2006).

Gillies, D.A. and Ietto-Gillies, G. 1991. Intersubjective probability and economics. *Review of Political Economy* 3, 393–417.

Goodwin, C.D.W. 1998. *Art and the Market: Roger Fry on Commerce in Art*. Ann Arbor: University of Michigan Press.

Goodwin, C.D.W. 2006. The art of an ethical life: Keynes and Bloomsbury. In Backhouse and Bateman (2006).

Harcourt, G.C. and Riach, P.A. 1997. *A 'Second Edition' of the General Theory*, 2 vols. London: Routledge.

Harrod, R. 1951. *The Life of John Maynard Keynes*. London: Macmillan.

Hession, C. 1984. *John Maynard Keynes: A Personal Biography of the Man Who Revolutionized Capitalism and the Way We Live*. New York: Macmillan.

Hill, P. and Keynes, R. 1989. *Lydia and Maynard: Letters between Lydia Lopokova and John Maynard Keynes*. London: Andre Deutsch.

Hoover, K.D. 2006. Doctor Keynes: economic theory in a diagnostic science. In Backhouse and Bateman (2006).

Johnson, E. 1960. Keynes' attitude to compulsory military service. *Economic Journal* 70, 160–5.

JMK (John Maynard Keynes). 1971–89. *The Collected Writings of John Maynard Keynes*, vols 30, ed. A. Robinson and D. Moggridge. London: Macmillan for the Royal Economic Society.

King, J.E. 2002. *A History of Post Keynesian Economics since 1936*. Cheltenham: Edward Elgar.

Laidler, D.W. 1999. *Fabricating the Keynesian Revolution: Studies of the Inter-War Literature on Money, the Cycle, and Unemployment*. Cambridge: Cambridge University Press.

Laidler, D.W. 2002. Skidelsky's Keynes: a review essay. *European Journal of the History of Economic Thought* 9, 97–110.

Lawlor, M.S. 2006. *The Economics of Keynes in Historical Context: An Intellectual History of the General Theory*. Basingstoke: Palgrave.

Leijonhufvud, A. 2006. Keynes as a Marshallian. In Backhouse and Bateman (2006).

Littleboy, B. 1991. *On Interpreting Keynes: A Study in Reconciliation*. London: Routledge.

Meltzer, A. 1988. *Keynes's Monetary Theory: A Different Interpretation*. Cambridge: Cambridge University Press.

Moggridge, D.E. 1976. *Keynes*. London: Fontana.

Moggridge, D.E. 1992. *Maynard Keynes: An Economist's Biography*. London: Routledge.

Moggridge, D.E. 2002. Rescuing Keynes from the economists? The Skidelsky trilogy. *European Journal of the History of Economic Thought* 9, 111–23.

Moore, G.E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Moore, G.E. 1912. *Ethics*. Oxford: Oxford University Press.

Mundell, R. 2008. Keynes and the international monetary system from *Indian Currency and Finance* to the IMF. In *Keynes's General Theory: A Reconsideration After Seventy Years*, ed. R.W. Dimand, R. Mundell and A. Vercelli. London and Basingstoke: Palgrave Macmillan for the International Economic Association.

O'Donnell, R.M. 1989. *Keynes: Philosophy, Economics and Politics*. London: Macmillan.

Patinkin, D. 1956. *Money, Interest and Prices*. Evanston, IL: Row, Peterson.

Ramsey, F.P. 1931. In *The Foundations of Mathematics and Other Logical Essays*, ed. R.B. Braithwaite. London: Routledge & Kegan Paul.

Robertson, D.H. 1926. *Banking Policy and the Price Level*. London: Macmillan.

Shackle, G.L.S. 1972. *Epistemics and Economics*. Cambridge: Cambridge University Press.

Shackle, G.L.S. 1974. *Keynesian Kaleidics*. Edinburgh: Edinburgh University Press.

Skidelsky, R. 1983. *John Maynard Keynes*, vol. 1: *Hopes Betrayed, 1883–1920*. London: Macmillan.

Skidelsky, R. 1992. *John Maynard Keynes*, vol. 2: *The Economist as Saviour, 1920–1937*. London: Macmillan.

Skidelsky, R. 1996. *Keynes*. Oxford: Oxford University Press.

Skidelsky, R. 2000. *John Maynard Keynes*, vol. 3: *Fighting for Britain, 1937–1946*. London: Macmillan.

Skidelsky, R. 2003. *John Maynard Keynes, 1883–1946: Economist, Philosopher, Statesman*. London: Macmillan.

Westall, O.M. 1992. *The Provincial Insurance Company, 1903–38: Family, Markets, and Competitive Growth*. Manchester: Manchester University Press.

Young, W. 1987. *Interpreting Mr Keynes: The IS–LM Enigma*. Boulder, CO: Westview Press.

How to cite this article

Backhouse, Roger E. and Bradley W. Bateman. "Keynes, John Maynard (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000074> doi:10.1057/9780230226203.0890

The New Palgrave Dictionary of Economics Online

Keynes, John Neville (1852 – 1949)

Phyllis Deane

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Keynes, J. N.; Marshall, A.; *Palgrave's Dictionary of Political Economy*

Article

John Neville Keynes was born in Salisbury and died in Cambridge, outliving his famous son, John Maynard Keynes, by over three years. He was a promising early pupil of Alfred Marshall, who on leaving Balliol in 1885 persuaded Oxford University to hire Neville, then a Cambridge lecturer and Fellow of Pembroke College, to fill the gap in its economics lectures, and then backed him strongly for the Oxford professorship when it became vacant in 1888. Keynes, however, was unwilling to leave Cambridge; he lectured for only two terms in Oxford, was not sorry when the Drummond professorship went to Thorold Rogers, and refused all offers of posts elsewhere (including the offer of a Chicago chair in 1894), devoting himself increasingly to his beloved family and to university administration – in which he held the top bureaucratic post, University Registrar, from 1910 to 1925.

Keynes published only two books, both textbooks, arising out of his university lectures: one in 1884 on formal logic, and the other, on which his reputation as an economist depends, *The Scope and Method of Political Economy* (1891), which grew out of the lectures he gave in Oxford in 1885. He also wrote a number of (mainly methodological) articles for Inglis Palgrave's *Dictionary of Political Economy*.

The importance of Keynes's *Scope and Method* lay in its becoming the standard text on economic method for the new Cambridge school led by Alfred Marshall. Its later drafts were composed at the same time as the later drafts of Marshall's *Principles of Economics* (1890), on which Keynes was commenting for Marshall while the latter was performing a similar critical service for *Scope and Method*. For the proponents of the new orthodoxy, the main contribution of Keynes's monograph was that it signalled the end of the methodological debates of the 1870s and 1880s, which had seemed to many, inside and outside the discipline, to call into question the scientific credentials of classical political economy.

It did so in three main ways: (1) by its lucid, judicious, low-key mode of exposition, deliberately occupying the middle ground in the methodological disputes and shifting the most controversial

arguments to appendices; (2) by redefining the hard scientific core of economic theory so as to insulate it from the charges of ideological bias, or immorality, or relativity, as well as from failures in practical economic policies; Keynes's threefold classification of economic enquiry claimed positive scientific status only for pure theory; the normative aspects and the policy aspects (the ethics and the art of political economy) constituted a protective belt which could absorb the attacks of historicists, socialists or nationalists, and so shift the doctrinal battle-ground away from fundamental principles; (3) by systematically minimizing the differences between the old economics and the new, by suggesting that the latter was a synthesis of the most fruitful of the conflicting views which characterized the period of methodological crisis and by stressing the continuity of economic ideas – so depicting a cumulative advance in economic knowledge analogous to the progressive improvements in knowledge claimed by researchers in the natural sciences.

In each of these ways, Neville Keynes successfully reflected the spirit of a new economic age, emergent not only in England but also in Europe and the USA, whose activists were bored with methodological argument and confident that they were in at the start of an exciting new research programme. At this point one might have expected the ambitious young academic to develop new research interests of his own in that programme.

The evidence of Keynes's diaries and letters leaves no doubt that he would have preferred to commit himself to political economy rather than logic, but the university's needs dictated otherwise. His appointment to a university lectureship coincided with Alfred Marshall's return to Cambridge as professor of political economy (and Mary Marshall's return as director of studies in economics at Newnham College), so that henceforth Keynes's opportunities to teach economics to Cambridge undergraduates were confined to an elementary course for Indian Civil Service candidates. One can thus date Keynes's loss of active interest in political economy from the completion of his *Scope and Method*. He was evidently bored by its long gestation, often depressed by Marshall's criticisms of successive draft chapters, but delighted with the flattering reviews or letters which its publication evoked from Marshall, Edgeworth, Palgrave, Taussig and Cossa, among other leading economists. Yet he made no attempt to embark on another book and was deaf to Edgeworth's pressing invitations to write for the new *Economic Journal* on topics outside the methodological field (though ready enough to rehash chunks of *Scope and Method* for Palgrave's *Dictionary*). Nor by then did he show any interest in being elected to an economics chair outside Cambridge, though in 1896 he seriously considered standing for the vacant registrarship in the University of London. The fact is that by the 1890s Keynes was fully committed elsewhere. His diaries show that he spent his working days trying to inject common sense into university politics (from the infighting on the Moral Sciences and Economics Boards to the perennial issue of women's degrees), while his leisure hours were absorbed in the ambitions and hobbies he shared with a lively, intelligent wife and three remarkable children – Maynard, Margaret and Geoffrey. These proved sufficiently demanding and fascinating concerns to distract him from the new economics which his colleagues (but not he) were then teaching in Cambridge.

Selected works

1884. *Studies and Exercises in Formal Logic*. London: Macmillan.

1891. *The Scope and Method of Political Economy*. London: Macmillan.

Bibliographic addendum

Deane, P. 2001. *The Life and Times of J. Neville Keynes: A Beacon in the Tempest*. Cheltenham: Edward Elgar.

How to cite this article

Deane, Phyllis. "Keynes, John Neville (1852–1949)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000017> doi:10.1057/9780230226203.0891

The New Palgrave Dictionary of Economics Online

Keynesian revolution

David Laidler
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘Keynesian Revolution’ suggests that Keynes's *General Theory* (1936) overthrew a defective and discredited classical orthodoxy and created a new understanding of how economies work. However, Keynes's critique of earlier work seriously misrepresented it, and his new system was, in fact, a synthesis of components drawn from it. Keynes's analysis nevertheless embodied an original and radical vision of how a monetary economy functions. The widespread adoption of the IS–LM interpretation of his system began a process of obscuring that vision, and economics has now largely lost sight of it

Keywords

Beveridge, W.; effective demand; expectations; Fisher, I.; Great Depression; Hansen, A.; IS–LM model; Kahn, R.; Keynes, J. M.; Keynesian Revolution; Keynesianism; Klein, L.; Leijonhufvud, A.; liquidity preference; macroeconomics, origins and history of; Malthus, T.; marginal efficiency of capital; marginal propensity to consume; Mill, J. S.; monetarism; multiplier; New Classical revolution; New Deal; Pigou, A.; rate of return over cost; Ricardo, D.; Robinson, J.; Say's Law; underemployment equilibria; Warming, J.

Article

The words ‘Keynesian Revolution’ conjure up the story of how a new intellectual framework, created by John Maynard Keynes and set out in his *General Theory of Employment, Interest and Money* (1936), challenged and quickly replaced an old-established but discredited classical orthodoxy.
The term, first popularized by Klein (1947), now has a permanent place in the vocabulary of economics, but this article follows Laidler (1999) (where references to relevant literature more extensive than space here permits are given) in suggesting that this revolution was *fabricated*, in two senses. First, Keynes to an extent *invented* the classical orthodoxy that he claimed to be overthrowing; but, second, he nevertheless *constructed* a radically new vision of how the economy functions, using components drawn

from his predecessors' work. It then suggests that the ideas underlying Keynes's revolution, always less influential than he intended, and currently barely recognized by economists, might nevertheless not be quite beyond revival.

Keynesian influences on economics

Economics saw enormous changes between the mid-1930s and the 1950s. To highlight only developments relevant to this essay, a new sub-discipline, *macroeconomics*, emerged; econometric modelling came into its own; and economic policy found a new anchor in the idea that ongoing government intervention in economic life can do more to promote economic and social well-being than the unregulated workings of the market. This whole apparatus is often labelled 'Keynesian', but to do so attributes too much to one man's influence.

Keynes was a sharp though perceptive critic, rather than a pioneer, of econometric modelling (Patinkin, 1976). His main contribution to the area was indirect and posthumous, lying in the adoption of the IS–LM model, extracted by some of his followers from the *General Theory*, as the template around which were built the forecasting models which became ubiquitous among central banks and departments of finance throughout the non-socialist world from the 1950s onwards.

As to economic policy, governments had begun to abandon laissez-faire long before the First World War, and the experiences of the inter-war years hugely accelerated this process, quite independently the influence of economics in general or of the *General Theory* in particular. The American New Deal was already well advanced before that book appeared; the famous Swedish model of a social-democratic mixed economy was mainly home-grown; and the government's major presence in the post-war British economy owed at least as much to the Fabian Society and William Beveridge as to Keynes. Even the use of fiscal and monetary policy for controlling the overall levels of economic activity and employment had many advocates before 1936, though the *General Theory* did significantly clarify the theoretical case for such macro-activism.

Keynes's intended revolution

Even so, this clarification was incidental to the book's purpose, which was, as its full title makes clear, to expound a *theory*, *general* in character, that would explain the determination of the level of *employment*, by referring to the roles played in economic life by the phenomena of *interest and money*. And, as Keynes told George Bernard Shaw in an oft-quoted letter dated 1 January 1935, he expected his book to 'largely revolutionise … the way the world thinks about economic problems'.

The central tenet of Keynes's intended revolution was that 'A monetary economy … is essentially one in which changing ideas about the future are capable of influencing the quantity of employment and not merely its direction' (1936, p. vii). By this Keynes meant that private investment decisions were heavily dependent upon expectations about the future profitability of the projects involved, and that, in a money economy, when those expectations varied, the result would be fluctuations in the level of employment. The outcome of similar shocks to a barter economy would be different: labour would simply be reallocated between the investment goods and consumption goods sectors.

Say's Law of Markets (that no one offers to sell goods except with the intention of buying goods with the

proceeds, so that their general oversupply is impossible) guaranteed a barter economy against unemployment, apart from that generated by frictions as workers moved among sectors; and variations in the rate of interest would induce changes in saving to match those in investment no matter how large and irregular the latter might be.

But, Keynes argued, Say's Law did not hold in a monetary economy, because the monetary system itself could, and often would, prevent the rate of interest from playing its equilibrating role. Since the days of David Ricardo, his predecessors, with a few honourable heterodox exceptions, had been blind to this, and his contemporaries remained so. They routinely, albeit often unself-consciously, applied Say's Law to monetary economies characterized by large-scale unemployment, thereby blundering into fundamental analytic inconsistency.

Not only did Keynes claim to have identified a ubiquitous error in the dominant economic theory of the preceding century and a quarter, but in the *General Theory* he also showed how an alternative system that avoided it could explain the mass unemployment plaguing market economies, and inform the design of a new policy framework. The latter's salient feature would be a government that took responsibility for guiding long-term investment decisions – whether only their overall volume or also their make-up was left unclear – this being an activity that both Keynes's theorizing and recent economic experience suggested were beyond the capacity of the private sector. Keynes provided only a sketch of this new policy framework, but emphasized its compatibility with a liberal-democratic political order.

Offered at a time when some important protagonists of economic and political liberalism, notably the Austrian school (Hayek, 1931; Robbins, 1934), were arguing that activist policy would only make the Great Depression worse, and when the apparent successes of the Soviet Union, Nazi Germany and Fascist Italy in conquering mass unemployment were tempting many in Britain, the United States, and elsewhere to embrace political totalitarianism as a necessary prerequisite for the conduct of effective economic policies, the political importance of Keynes's message is hard to overestimate. One cannot read Tarshis's (1987) first-hand account of his exposure to Keynes's radical and wide-ranging assault on received economic theory without sympathizing with the enormous excitement that it generated.

But such sympathy must be tempered by the knowledge that Keynes himself, perhaps inadvertently, presented a very particular and sometimes very inaccurate view of his *General Theory*'s place in the history of economics. He claimed to have overturned more than a century of economics and to be rebuilding the subject on new foundations, but it is more accurate to say that the *General Theory* first seriously misrepresented earlier work, and then selected from it components to be synthesized into what was, nevertheless, a strikingly original framework.

Keynes on classical economics

In the depressed period following the Napoleonic Wars, Ricardo had attacked Malthus's argument that too rapid a rate of capital accumulation had created a state of affairs in which the economy was no longer able to consume all that it was capable of producing. Deploying Say's Law, Ricardo had argued that a *general glut* of commodities, such as Malthus postulated, did not, indeed could not, exist. There was merely a temporary, though serious, mismatch between the compositions of output and demand as the economy continued to switch from wartime to peacetime patterns of production and consumption. Ricardo's logic initially carried the day, but it was soon subjected to an important qualification.

Specifically, John Stuart Mill (1844) pointed out that Say's Law, when applied to a money economy, ruled out only a general glut of everything *including money*, but left open the possibility that an oversupply of everything *except money* would appear at times when agents were trying to build up their stocks of the latter. Mill associated such behaviour with financial crises, and thought it would be short-lived, but as the 19th century progressed his insight became incorporated into accounts of the upper turning point of the business cycle.

There, in tandem with the hypothesis of nominal wage stickiness, it helped to provide a monetary explanation for the onset of cyclical unemployment which still found an influential expression among Keynes's contemporaries in the work of Ralph Hawtrey (for example, 1919). In Hawtrey's view, cyclical interactions of the supply and demand for money led to fluctuations in what he (like Malthus, and later Keynes) called *effective demand* – the rate of flow of money spending on goods and services in the aggregate – which would impinge upon income and employment to the extent that money wage and price stickiness prevented those fluctuations from being absorbed by movements in the general price level.

Keynes was thus wrong to assert (1936, p. 33) that Ricardo's version of Say's Law had for more than a century dominated an orthodox economics that had lost sight of the very concept of effective demand, and, as a corollary, he was also wrong to claim novelty for his own account of how cuts in money wages, and hence prices, might affect employment for the better through an indirect channel involving the interaction of the supply and demand for money (1936, p. 266).

These errors are arguably mere slips when viewed in the context of the *General Theory*'s central claim that large-scale unemployment was not, after all, due to money-wage stickiness, but to fundamental problems posed by the nature of a monetary economy for the capacity of the rate of interest to coordinate saving and investment; but Keynes was by no means the first to explore the latter issues either. From Wicksell (1898) onwards, increasing attention was paid to the influence of the rate of interest set by the central bank on saving and investment decisions, and to its capacity to disrupt their coordination by the capital market. Swedish, Austrian and British economists (for example, Myrdal, 1931; Hayek, 1931; Robertson, 1926) all investigated ways in which monetary mechanisms might create fluctuations in output and employment, albeit none of them with complete success.

Keynes's contribution, then, was not to emphasize that monetary factors could disrupt the smooth allocation of resources over time, but to show precisely how they might do so, what the consequences would be, and how economic institutions could be adapted to cope with such problems. Before the *General Theory* there was no coherent and widely accepted analytic framework in terms of which these issues could be discussed, but after it there was.

Keynes's new framework

That framework had three major components: a theory emphasizing the role of expectations in driving investment decisions, a theory of the demand for money that explained why the rate of interest could not be relied on to coordinate these with saving decisions, and a theory of saving–consumption behaviour that implied a self-limiting multiplier process. Each already existed before 1936, but they had not been brought together. When they were, output and employment variations were revealed to be what equilibrated saving and investment when, as would happen in a monetary economy, the interest rate

failed to do so.

Keynes called the central concept of his theory of investment the *marginal efficiency of capital*, the rate of discount that would equate the present value of the profits expected from a unit of investment to the cost of producing the capital equipment involved. Forward-looking maximizing firms would push the flow of investment to the point at which this rate of return was just equal to the rate of interest at which the expenditure was to be financed.

There was nothing revolutionary in this concept, which Keynes acknowledged to be essentially identical to Irving Fisher's (1907) *rate of return over cost*, but his views on the relationship between profit expectations and the real productivity of investment that, in Fisher's analysis, underlay them, were less conventional. For Keynes, this connection was essentially non-existent: the passage of time was a fundamental fact of economic life, and the future was simply too uncertain for calculations of productivity to be made in a rational fashion. Profit expectations therefore were inevitably driven by the essentially irrational *animal spirits* of investors. When these were high or low, so would be the marginal efficiency of capital.

Though arguably radical, this idea was not particularly new. It restated, using a new vocabulary, the view of other Cambridge economists – not least the ‘classical’ *bête noire* of the *General Theory*, A. C. Pigou (for example, 1927) – that investment decisions were largely driven by contagious and cumulative waves of errors of optimism or pessimism on the part of businessmen. Keynes's account of how the development of stock markets had deepened the wedge between long-run economic fundamentals and the short-term and ill-informed expectations on which investment decisions were actually based was also foreshadowed in earlier Cambridge work.

So too was his treatment of money-holding behaviour. Cambridge monetary analysis had been conducted since the 1870s in stock supply and demand terms, with demand being driven by money's role as the economy's means of exchange. In 1921 Keynes's former student Frederick Lavington had extended this analysis to an economy characterized by sophisticated financial markets. Here, he noted, money had the particular virtue of being always readily tradable at a market value that was subject to less uncertainty than that of other assets. Lavington concluded that, while remaining its means of exchange, money would also serve as a store of value in such an economy, providing a hedge against the uncertainty inherent in its financial markets.

Keynes himself considerably elaborated this idea in his *Treatise on Money* (1930), and under the label *liquidity preference*, it in due course appeared in the *General Theory*, where it provided a rationale for the incapacity of the rate of interest to maintain equilibrium between saving and investment. Holding money to obtain security against financial market uncertainty involved forgoing the interest yielded by alternative assets. In a monetary economy, then, interest was the price of liquidity, and its performance of this, its major role, would sometimes fatally interfere with its capacity to coordinate the allocation of resources over time.

When animal spirits were up and expectations optimistic, to be sure, financial market uncertainty would be a minor matter, liquidity preference would be both weak and relatively insensitive to the rate of interest, and these considerations would not be crucial to the economy's functioning. But when animal spirits and expectations were depressed they would be. Under these circumstances, which Keynes believed to be chronic in the market economies of the 1930s, the marginal efficiency of capital would be low and a low rate of interest would be needed to match it; but liquidity preference would simultaneously be strong with only very small movements in the interest rate being needed to induce

large changes in money holding.

The very nature of a modern monetary economy thus inhibited the interest rate's accommodating itself to a depressed marginal efficiency of capital, and such an economy could, and very likely would, settle into an equilibrium characterized by large-scale unemployment. This final implication was established by the logic of the *multiplier* relationship, which Keynes took over, not from earlier theoretical work, but from a much more practical literature dealing with the use of public works expenditures to fight unemployment.

Long before the First World War it had been recognized that prosperity and depression in one sector of the economy could spill over into others, and this idea helped establish the desirability of using public expenditure to put the unemployed to work. Sometimes, however, the implications here seemed too good to be true: if spending geared to employing previously unemployed workers would generate further expenditures on their part that would then put others to work, and so on, what was there to stop one small injection of public expenditure eliminating any level of unemployment?

A satisfactory answer to this awkward question was finally provided in 1931 by Keynes's student Richard Kahn: there would be some leakage of expenditure at each round in the process, and each successive increment to employment would be smaller than its predecessor. The effects of public expenditure on employment would be multiplied beyond their immediate impact, to be sure, but not infinitely so. They would converge to a limit that would be smaller the greater the size of the aforementioned leakages.

The Danish economist Jens Warming (1932) then offered a crucial modification to this analysis. Instead of applying it to employment, he applied it to output, and hence real income; and instead of invoking a number of possible leakages from the circular flow of expenditure as Kahn had done, he emphasized one, namely, saving. Postulating in his illustrative numerical example that consumers would spend 75 per cent of each increment to their income and save the balance, Warming showed that a given injection of public expenditure would generate a fourfold increase in national income.

Warming's version of the multiplier was incorporated in the *General Theory*, though only Kahn was cited there. In Keynes's hands, however, the stable fraction spent out of any increment to income – still 75 per cent in his own numerical example – became the *marginal propensity to consume*, the embodiment of a ‘fundamental psychological law’ of consumer behaviour, and the multiplier itself elucidated not merely the practical consequences of public works expenditures, but the fundamental theoretical links between investment and the economy's overall level of effective demand – consumption plus investment – and hence output and employment. When the rate of interest was unable to offset fluctuations in the marginal efficiency of capital because of its role as the price of liquidity, the multiplier would ensure that output and employment moved so that savings matched investment. Hence it lay at the very heart of Keynes's revolutionary message that, ‘a monetary economy … is essentially one in which changing views about the future are capable of influencing the quantity of employment…’ (1936, p. vii).

IS – LM and after

By the 1950s, what was still called Keynesian economics had largely lost track of this message. In part, this was because the Second World War and its aftermath had seen the restoration of high employment as the economy's apparently normal state of affairs, but more importantly it was the consequence of the

way in which Keynes himself had developed his results.

Before 1936, many economists had paid attention to expectations and their evolution over time, but available analytic techniques were not up tackling such problems, and their efforts, though sometime yielding valuable insights along the way, routinely ground to a halt in confusing, not to say confused, complexity. The critical step enabling Keynes to tell an analytically tractable story where others had failed was to treat expectations as exogenous to the mechanisms he analysed, and this simplification also made it possible for others to extract a formal comparative-static model from the *General Theory* that could be expressed either in simple algebra or in equally simple geometry.

This, the IS–LM (investment equals saving – liquidity preference equals the supply of money) model, was, in its simplest form, a set of simultaneous equations that linked consumption to income, investment to the rate of interest, and the demand for money to income and the rate of interest, and characterized the money stock as exogenous.

These components were all to be found in the pages of the *General Theory*, and the IS–LM system could also be manipulated to demonstrate some of that book's central conclusions – for example, about how a fall in the rate of investment spending at any level of the rate of interest would put downward pressure on the equilibrium values of output as well as the rate of interest, and about how a high degree of interest sensitivity on the part of the demand for money would force more of this adjustment onto income. It was, furthermore, easily extended to accommodate analyses of monetary and fiscal policy.

All of this gave IS–LM a strong claim to be a legitimate representation of Keynesian economics, as Alvin Hansen (1953), its most influential exponent, would in due course claim. By and large, this is how it came to be treated, despite the protests of some who had been close to Keynes when he had tried set his revolution in motion, not least his younger Cambridge colleague Joan Robinson, who memorably characterised IS–LM as ‘bastard Keynesianism’.

It is not necessary to take sides in this debate here. It will suffice to note that IS–LM proved remarkably flexible. In the 1960s it accommodated versions of not just Keynesian but also monetarist doctrine, and provided a framework in which some of the issues separating them could be debated. Hence it dominated both research and teaching within macroeconomics for close to two decades; but its dominance came at a cost (Backhouse and Laidler, 2004). In particular, under this model's influence macroeconomics lost sight of the importance of time in economic life. Ideas, including those that had lain at the heart of Keynes's intended revolution, about the crucial role played by expectations and uncertainty in inter-temporal coordination mechanisms, and the essential differences between the ways in which money and barter economies coped with such matters, were pushed into the background.

They have never quite disappeared, however. For example, Axel Leijonhufvud's (1968) restatement of Keynes's economics attracted much attention. So did his suggestion that Keynes had been forced by his analytic framework to treat as equilibria the disequilibria that coordination failures in fact created, though his suggestion that a new and explicitly disequilibrium dynamic economics be built on Keynesian foundations was not taken up. Instead, in the 1970s economists in large numbers embraced analytically more tractable New Classical analysis, built on the principle that markets are continuously in equilibrium.

This technically convenient clearing-markets assumption in fact had huge substantive significance, implying the total irrelevance of Mill's (1844) critique of the application of Say's Law to a monetary economy, and all that had followed from it. When Lucas and Sargent (1978) announced the demise of Keynesian economics, therefore, more than activist policies supported by macroeconometrics was under

attack. The discipline's very comprehension that a monetary economy might suffer coordination failures, let alone of Keynes's specific analysis of these issues, already weakened by IS–LM, was threatened. It is just as well, then, that the New Classical revolution, like its Keynesian forerunner, has been less than totally successful, perhaps leaving room for these old problems to be debated once more.

See Also

- Fisher, Irving
- Great Depression
- Kahn, Richard Ferdinand
- Keynes, John Maynard
- Keynesianism
- macroeconomics, origins and history of
- Pigou, Arthur Cecil
- Robertson, Dennis
- underemployment equilibria
- Warming, Jens

Bibliography

Backhouse, R. and Laidler, D. 2004. What was lost with IS–LM. In *The IS–LM Model: Its Rise, Fall and Strange Persistence*, ed. M. de Vroey and K. Hoover. Durham, NC: Duke University Press.

Fisher, I. 1907. *The Rate of Interest: Its Nature, Determination, and Relation to Economic Phenomena*. New York: Macmillan.

Hansen, A. 1953. *A Guide to Keynes*. New York: McGraw Hill.

Hawtrey, R. 1919. *Currency and Credit*. London: Longmans Group.

Hayek, F.A. von. 1931. *Prices and Production*. London: Routledge.

Kahn, R. 1931. The relation of home investment to unemployment. *Economic Journal* 41, 173–98.

Keynes, J. M. 1930. *A Treatise on Money*. London: Macmillan.

Keynes, J. M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Klein, L. 1947. *The Keynesian Revolution*. New York: Macmillan.

Laidler, D. 1999. *Fabricating the Keynesian Revolution*. Cambridge: Cambridge University Press.

Lavington, F. 1921. *The English Capital Market*. London: Methuen.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. London: Oxford University Press.

Lucas, R. Jr. and Sargent, T. 1978. After Keynesian economics. In *Rational Expectations and Econometric Practice*, ed. R. Lucas Jr. and T. Sargent. London: Allen and Unwin.

Mill, J. S. 1844. On the influence of consumption upon production. In *Essays on Unsettled Questions in Political Economy*. London: Longmans, Green, Reader and Dyer.

Myrdal, G. 1931. *Monetary Equilibrium*. London: W. Hodge, 1939.

Patinkin, D. 1976. Keynes and econometrics: on the interaction between the macroeconomic revolutions of the interwar period. *Econometrica* 44, 1091–123.

Pigou, A. 1927. *Industrial Fluctuations*. London: Macmillan.

Robertson, D. 1926. *Banking Policy and the Price Level*. London: Macmillan.

Robbins, L. 1934. *The Great Depression*. London: Macmillan.

Tarshis, L. 1987. Keynesian Revolution. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Warming, J. 1932. International difficulties arising out of the financing of public works during depression. *Economic Journal* 42, 211–24.

Wicksell, K. 1898. *Interest and Prices*, trans. R. Kahn. London: Macmillan, 1936.

How to cite this article

Laidler, David. "Keynesian revolution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000019> doi:10.1057/9780230226203.0892

The New Palgrave Dictionary of Economics Online

Keynesianism

Roger E. Backhouse and Bradley W. Bateman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the post-war years, Keynesianism became the label for the mixed economy, for an approach to fiscal policy that entailed fine-tuning the economy, and for the revolution in economic theory that brought macroeconomic analysis to the fore. This article debunks many of the myths that grew up around Keynes's legacy by examining his attitude to fiscal and monetary policy over the course of his career. By differentiating what Keynes said from what his followers and his critics said after his death, it is possible to understand the broad switch to demand management during the 20th century in a clearer light.

Keywords

aggregate demand; Beveridge, W.; Buchanan, J.; budget deficits; cheap money; command economy; counter-cyclical policy; deficit spending; fine tuning; fiscal policy; full employment; functional finance; general equilibrium; Great Depression; Hansen, A.; Harrod, R.; Hayek, F.; Hicks, J.; inflation; IS–LM model; Keynes, J. M.; Keynesianism; Keynesian Revolution; Lerner, A.; market frictions; Meade, J.; microfoundations; mixed economy; monetary policy; monetary targets; national debt; New Deal; public works; Robinson, J.; Salant, W.; Samuelson, P.; sinking fund; Social Security; stagflation; sticky wages; unemployment; welfare state

Article

Keynesianism has many meanings. It is the label for the political philosophy that dominated most Western countries in the 30 years after the Second World War, embracing a mixed economy and the welfare state, steering a course between what were believed to be the dead hand of socialism and the social injustices of free-market capitalism. Keynesianism is also used to refer to something narrower: to the use of macroeconomic policy to stabilize the economy and to maintain low levels of unemployment. In this usage, Keynesianism is associated with fine-tuning the level of government spending and taxation so as to use variations in the budget deficit to counteract shocks that would otherwise cause high

unemployment. Beneath all of these meanings, of course, lies Keynesian economic theory, which provides the theoretical foundation for these policies. (For further discussions of Keynes, Keynesianism, and the Keynesian Revolution, see Backhouse and Bateman, 2006.)

According to popular mythology there are clear historical links between these various meanings of Keynesianism. This mythology comprises the claim that Keynes's *General Theory* (1936) provided the basis for a new economics, marking a revolutionary break with previous orthodoxy, justifying the use of debt-financed government budget deficits to stimulate the economy and cure unemployment. Governments turned to Keynes to provide a way out of the Great Depression and to justify maintaining high levels of demand after the Second World War. In doing this they ensured that mass unemployment would not recur, thereby making possible the development of the welfare state. However, this policy had two unintended effects. It undermined an unwritten fiscal constitution, according to which governments would normally balance their budgets, except in wartime. Also, by removing the fear of unemployment, it undermined the willingness of workers to restrain their wage demands. The combination of budget deficits and high wage demands eventually caused the stagflation of the 1970s, thereby bringing about the demise of Keynesianism.

Fortunately for Keynes's reputation, modern scholarship has shown that most of the claims on which this account is based are mythical. Things simply did not happen this way. As the Keynesian Revolution in economic theory is discussed elsewhere in this dictionary, the focus here is on Keynesianism as economic policy and political philosophy. We start by showing how ‘Keynesian’ ideas actually entered policy, in many cases before 1936, the year that Keynes's *General Theory* was published. From there we outline Keynes's own views on fiscal policy, differentiating them from what came to be known as Keynesianism. In conclusion, we suggest that Keynesianism in economic policy is more alive than is often assumed.

The spread of Keynesianism

In a path-breaking comparative study, *The Political Power of Economic Ideas: Keynesianism Across Nations* (1989, p. 367), Peter Hall was surprised to discover ‘the degree to which Keynes's ideas about demand management were resisted or ignored in many nations’. Demand management was adopted in many countries, but often without reference to Keynes's ideas. The United States provides an excellent example. President Franklin Delano Roosevelt, in his 1932 and 1936 election campaigns, ran on the promise of balancing the budget. In his first administration he managed to limit deficits to what was spent on relief projects. It was only in his second term, as the economy slid into recession again in 1937, that Roosevelt submitted a budget that was purposely in deficit to stimulate the economy. He did this, not because he was influenced by Keynes, but because other options to raise prices and stimulate recovery had failed. Attempts to buy up gold had raised the price of gold but not prices of basic commodities. The National Industrial Recovery Act had been declared unconstitutional. The idea of running a deficit came from a group of economists (Laughlin Currie, Leon Henderson and Isador Lubin), recruited by Harry Hopkins, a New Deal administrator about to become Secretary of Commerce. They noted that the fortunes of the economy in 1936–7 had exactly mirrored the change in the government's fiscal position, caused by the ending of the First World War Veterans’ bonus and the imposition of taxes to support the new Social Security system. It was this evidence, not Keynesian theory, that was used to make the case for fiscal stimulus. Though young Keynesian economists did

eventually enter the government, the initial arguments for using fiscal policy to stimulate the economy were without reference to or influence from Keynes. Herbert Stein (1969, p. 131) thus went so far as to say, ‘it is possible to describe the evolution of fiscal policy in America up to 1940 without reference to him [Keynes]’.

In Germany and France, too, deficits came independently of Keynesian ideas. In Germany, the formation of a democratic government in the economic chaos that followed the First World War meant satisfying various interest groups, which implied deficits: businesses demanded tax cuts and workers and farmers demanded higher spending. Deficit spending was the ‘social cement’ (James, 1989, p. 234). In France, the *General Theory* was not translated until 1942, and few people at this time read it in English. Britain and Canada, where Keynes and young Keynesians were involved in government, were exceptional cases. However, even in Britain the precise nature of Keynes's influence is far from clear (Peden, 1988; 2006). His influence was clearest in the 1941 budget, which he believed marked a revolution in public finance; but his ideas about balancing aggregate supply and demand were used to control inflation, not to ensure full employment. In general there was considerable resistance to Keynes's ideas, often on grounds of administrative practicality. Even after the Second World War, it has been argued that Keynesian ideas were not applied till 1947, and then, as in 1941, to control inflation. Furthermore, though demand management was certainly in vogue after that, high employment was not achieved by running government deficits. If deficits are calculated on a traditional Gladstonian basis, excluding separately funded capital expenditure from the budget, the British government ran a surplus in every year from 1948 to 1972 (with the possible exception of 1965; see Clarke, 1998, pp. 210–11). Even if the result was sometimes an overall deficit, this hardly justifies the charges of profligacy and undermining an implicit fiscal constitution levelled by critics such as Buchanan and Wagner. And this was the country in which Keynes's influence was strongest.

The early arguments for counter-cyclical fiscal policy were independent of Keynes. They were what Hall (1989) described as ‘proto-Keynesianism’ – Keynesianism without the theoretical foundations provided by Keynes's *General Theory*. However, during the 1940s Keynes's name eventually came to be attached to such policies. One reason for this was simply that Keynes's model of aggregate demand was the most advanced economic theory at the time and it could easily be used to provide an *ex post* imprimatur to a change that was already under way. Keynesian economic theory, combined with national accounts constructed along Keynesian lines, provided a common language in which economists, in government and outside, could talk about macroeconomic problems. The combination of mathematical economic theory and statistical data analysis opened up an apparent gulf with pre-Keynesian work on money and the cycle. The credibility of this new economics, comprising Keynesian theory and counter-cyclical policy, was given an enormous boost from the apparent success of wartime demand management: Walter Salant (1989, pp. 45–6) has claimed that ‘The elimination of unemployment during World War II was one of the greatest influences on post-war views about the role of government in attaining and maintaining high employment’. It has, however, been argued that this was based on a misperception, in that the war economy in the United States was an example of a successful command and control economy, not successful demand management (Higgs, 1992).

The association of Keynes with counter-cyclical policy was actively fostered by Keynes's disciples, Alvin Hansen and young Keynesians in Britain and the United States, James Meade, Joan Robinson, Abba Lerner, Paul Samuelson and Walter Salant. Lerner's *The Economics of Control* (1944), ironically a

Ph.D. thesis supervised by Friedrich Hayek, then still at the London School of Economics, used Keynesian theory to advocate a high degree of fiscal fine tuning. In *A Guide to Keynes* (1953), Hansen wove together an exposition of Keynesian economic theory with his own policy recommendations on deficit spending, creating an indelible link between the two.

Keynes and Keynesianism

Though Keynes's views on economic policy were famously fluid, there was little change in his views on deficit spending. From the late 1920s till his death, he supported using public works projects to stimulate aggregate demand at appropriate points in the cycle. However, this did not imply support for government deficits. New housing or investment in the transport infrastructure were capital projects that would generate revenue streams capable of paying back any money that had been borrowed to finance their construction. Such spending would therefore appear on the capital budget, meaning that it would not affect the government's regular budget or the government's deficit. Keynes's opposition to funding such projects through the government budget, partly on the grounds that it might frighten businessmen, was strong enough for him to argue, from 1924, for the creation of a separate capital budget into which such funding could be placed. He also argued that payments to the sinking fund (the fund destined to repay the national debt) could be diverted into the capital budget, obviating the need to raise any new funds to undertake public works projects. A major part of Keynes's battles with the Treasury, therefore, was to argue for new accounting procedures so that such projects could be undertaken without unbalancing the budget.

Keynes continued his opposition to budget deficits to the end of his life. In working on the White Paper on full employment (Great Britain: Ministry of Reconstruction, 1944 – a set of proposals published by the government as the basis for legislation) and the National Debt Inquiry (Keynes, 1945), he insisted repeatedly that he was not arguing for deficits in the ordinary budget. Historians have even taken the failure to separate the capital and ordinary budgets in the White Paper as evidence on Keynes's limited impact on Treasury thinking at this time, so closely was he associated with the idea. He also argued against the young Keynesians, who were advocating the adjustment of social security taxes to regulate demand.

In these decades, Keynes's emphasis was on increasing investment, not consumption. Public investment might be increased directly, and private investment could be increased through a policy of cheap money. In the *General Theory* he argued that interest rates should be kept low, and maintained that view for the next ten years of his life. Nowhere did he support the view, commonly associated with Keynesianism, that monetary policy was ineffective.

Neither was Keynes especially enthusiastic about the welfare state. William Beveridge was the person who championed the development of Britain's post-war welfare state, and Keynes was never a close collaborator with Beveridge in this work. Keynes did look at several drafts of Beveridge's draft reports as a Treasury official, and he once wrote to Beveridge praising the plans. The fact remains, however, that his work within the wartime government on the implementation of the Beveridge plan consisted largely of efforts to trim the size of Beveridge's plan, limit child payments so that they did not cover the first-born in any family, and to delay implementation of the plan. For Keynes, the most important issue was policy to ensure full employment, and he never linked that objective directly to the welfare state in his own writing.

How, then, did Keynesian economics come to be associated with policies that Keynes clearly rejected throughout his life, such as the fiscal fine tuning of Lerner's ‘functional finance’? The answer is that his ideas came to be seen through the work of the young Keynesians, who claimed Keynes's imprimatur for their ideas. Lerner (1936, p. 435), in his review of the *General Theory*, wrote ‘this article has been read in manuscript by Mr. Keynes himself, who has expressed his approval of it’. Though this article was about model building, not policy, Lerner subsequently wrote as if Keynes's approval extended to all his work. Hansen, too, wove his interpretation of Keynesian economics together with his own policy recommendations on deficit spending. As a staple of undergraduate and graduate education in the 1950s and 1960s, this helped create the view that Keynes supported deficit spending (despite a caveat that Keynes had not endorsed such policies hidden away towards the end of Hansen's book). In Britain, young Keynesians such as James Meade had the advantage that they had actually worked with Keynes: even though they had disagreed on some policy recommendations, this personal link meant that their ideas came quickly to be associated with Keynes. Keynesianism became the policies of the Keynesians, not of Keynes himself.

Blame for the mislabelling of what came to be known as ‘Keynesian’ policies also rests with Keynes's critics. In the same way that supporters of using counter-cyclical policy to maintain a high level of aggregate demand wanted to claim Keynes's authority for their ideas, critics wanted to make the same identification, to give more significance to their own attacks on such ideas: had it been known that Keynes himself was critical of these policies, questions would have been raised concerning whether there were alternative policies to those being proposed. While Keynes was alive, Hayek was well aware of the differences between Keynes's own views and those of the young Keynesians; he recalled having a conversation in which Keynes agreed with him when he complained about the dangerous things that the young Keynesians were saying on Keynes's behalf (Hayek, 1995, p. 232). Yet, after Keynes's death, and in particular in the 1970s when he returned to writing about money and inflation after a decades doing other things, he too spoke of Keynesianism as though it reflected the ideas of Keynes. Buchanan and Wagner were even more explicit, in their *Democracy in Deficit: The Political Legacy of Lord Keynes* (1977), in claiming that it was Keynes who had called forth a world of ever-increasing budget deficits, never pointing out that Keynes was avowedly hostile to this idea (for an assessment of this view, see Bateman, 2005).

At the level of economic theory, Keynesian economics came to be associated with a theory that led to conclusions that were different from those advocated by Keynes himself. Within a year of the *General Theory*'s appearance, economists began reformulating its ideas as a simple two or three simultaneous equation system that eventually became known as the IS–LM model. This model was then given more formal microfoundations and seen as a miniature general equilibrium model. As with Lerner's review of his book, Keynes was willing to encourage such models, writing positively to Roy Harrod and John Hicks about their early papers that provided the initial foundation for the IS–LM model. However, while these models captured something of what Keynes was doing, there was much in the *General Theory* that they left out, and as a result the theory became simplified. The theory became shorn of many of the elements that related most closely to time, resulting in a static model in which the only mechanism that would ensure the economy did not return to full employment was inflexibility of the money wage rate (see Backhouse and Laidler, 2004). For example, Keynes's dynamic arguments about how wage cuts might have perverse effects through causing expectations of future wages to change were ignored. Keynesian economics thus came to be seen as ‘the economics of sticky wages’, even though Keynes had

written the *General Theory* to combat the view that this was the cause of the depression.
This theoretical reinterpretation probably happened because of another change that took place after the Second World War: economists began conducting their arguments in terms of mathematical models (either algebraic or geometric). Ideas, even if they were important to Keynes, that could not be forced into a model were forgotten. This reinterpretation of Keynesian theory tied in with the emerging Keynesianism in economic policy in that in these models, which analysed the determinants of a homogeneous aggregate output, the distinction between current and capital expenditure played a minor role. Government spending, in most models, was equated with current expenditure, and investment was seen as a private sector activity. These assumptions were largely unquestioned, at least in the mainstream of the discipline, until Axel Leijohnufvud's *On Keynesian Economics and the Economics of Keynes* (1968). This argued explicitly that Keynesian economics was very different from the economics of Keynes: Keynes's followers had come to the conclusion that Keynesian economics was a special case of the more general classical theory only because they had misrepresented his ideas.

Are we all Keynesians now?

In the 1970s governments moved away from Keynesianism, as the term had by then come to be understood. This parallels developments in economic theory, and had much to do with the apparent breakdown of the Phillips curve, and the failure of Keynesianism, as it was then understood, to provide guidance appropriate to a time of stagflation. However, governments were mainly moving away from doctrines that were associated with Keynes's followers, not with Keynes himself. Fiscal fine tuning to maintain full employment was abandoned in favour of focusing on money and inflation. Paradoxically, in Britain at least, it was only in the 1970s that the government began to run budget deficits (defined in the Gladstonian manner), perhaps because of its policy of drastically cutting public investment. Keynesianism appeared dead. And yet, as governments learned about the problems with using monetary targets, there was a move, by the end of the century, towards policies that were much more in line with what Keynes himself had advocated. Interest rates were kept as low as was consistent with reasonably stable prices (a slightly positive interest rate). Monetary policy, not fiscal policy, came to be seen as central to stabilizing the economy. There is much here that is not in Keynes – hardly surprising given the changes that took place during the preceding 60 years – but he would probably have had much sympathy with the broad framework of policy: targeting domestic prices, creating the conditions for high levels of investment, and perhaps even limiting the role of the welfare state. There were parallels in economic theory, where Keynesianism came to be used to refer to those who believed various market frictions opened up a role for activist policy. Perhaps the abandoning of ‘Keynesian’ political philosophy was what paved the way for implementing some of the ideas that were important to Keynes.

See Also

- functional finance
- Keynes, John Maynard
- Keynes, John Maynard (new perspectives)
- Keynesian revolution

- monetary and fiscal policy overview
- monetary policy, history of

Bibliography

Backhouse, R. and Bateman, B. 2006. *The Cambridge Companion to Keynes*. Cambridge: Cambridge University Press.

Backhouse, R. and Laidler, D. 2004. What was lost with IS–LM. *History of Political Economy* 30 (Supplement on 'The IS–LM model: its rise, fall and strange persistence'), 25–56.

Bateman, B. 2005. Scholarship in deficit: Buchanan and Wagner on John Maynard Keynes. *History of Political Economy* 37, 185–90.

Buchanan, J. and Wagner, R. 1977. *Democracy in Deficit: The Political Legacy of Lord Keynes*. New York: Academic Press.

Clarke, P. 1998. *The Keynesian Revolution and its Economic Consequences*. Cheltenham: Edward Elgar.

Great Britain: Ministry of Reconstruction. 1944. *Employment Policy*. Cmd. 6527. London: HMSO.

Hall, P., ed. 1989. *The Political Power of Economic Ideas: Keynesianism Across Nations*. Princeton, NJ: Princeton University Press.

Hansen, A. 1953. *A Guide to Keynes*. London: McGraw Hill.

Hayek, F.A., ed. 1995. *Contra Keynes and Cambridge*, ed. B. Caldwell. The Collected Works of F.A. Hayek, vol. 9, ed. S. Kresge. London: Routledge.

Higgs, R. 1992. Wartime prosperity: a reassessment of the US economy in the 1940s. *Journal of Economic History* 52, 41–60.

James, H. 1989. What is Keynesian about deficit financing? The case of interwar Germany. In Hall (1989).

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1945. National debt inquiry. In *The Collected Writings of John Maynard Keynes*, vol. 27, ed. D. Moggridge. London: Macmillan and Cambridge University Press, 1980.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. Oxford: Oxford

University Press.

Lerner, A. 1936. Mr. Keynes' General Theory of Employment, Interest and Money. *International Labour Review* 34, 435–54.

Lerner, A. 1944. *The Economics of Control*. New York: Macmillan.

Peden, G. 1988. *Keynes, the Treasury and British Economic Policy*. London: Macmillan.

Peden, G.C. 2006. Keynes and British economic policy. In *The Cambridge Companion to Keynes*, ed. R. E. Backhouse and B.W. Bateman. Cambridge: Cambridge University Press.

Salant, W. 1989. The spread of Keynesian doctrines and practices in the United States. In *The Political Power of Economic Ideas: Keynesianism Across Nations*, ed. P. Hall. Princeton, NJ: Princeton University Press.

Stein, H. 1969. *The Fiscal Revolution in America*. Chicago: University of Chicago Press.

How to cite this article

Backhouse, Roger E. and Bradley W. Bateman. "Keynesianism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000020> doi:10.1057/9780230226203.0893

The New Palgrave Dictionary of Economics Online

Kindleberger, Charles P. (1910 – 2003)

Peter Temin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International economist and economic historian of the late 20th century, Charles Kindleberger was an astute observer of the world around him and a master prose stylist. His most influential works made the case for irrationality in capital markets and the need for a lender of last resort to minimize the damage from bubbles and mania. His work was narrative rather than abstract, but no less convincing for that. His most famous book is *Manias, Panics and Crashes*.

Keywords

Bagehot, W.; bubbles; gold standard; Great Depression; Kindleberger, C.; labour supply

Article

Charles P. Kindleberger was born in New York City. He received his B.A. at the University of Pennsylvania in 1932 and his Ph.D. at Columbia University in 1937. He had a distinguished career in public service (including the Federal Reserve and the Office of Strategic Services during the Second World War) before going to teach international trade at MIT. His wartime experiences directed his interests towards the interaction of countries and gave him a keen sense of how academic ideas play out among real people and governments. His scholarship was characterized by its realism and willingness to consider actual – as opposed to idealized – behaviour.

Kindleberger made his mark on the field of international trade through his textbook and through papers and books about the recovery of Europe after the Second World War. He was active in the analysis of the dollar scarcity and then the dollar glut that characterized the short life of the Bretton Woods System. He also wrote a prescient book, *Europe's Postwar Growth: The Role of Labor Supply* (1967), on the role of immigrants and guest workers from eastern and southern Europe in alleviating the labour scarcity of western Europe. Kindleberger's emphasis on the evolution of labour supply has been echoed in many subsequent studies. The legacy of these post-war policies has been evident in political and economic conflict between the children and even grandchildren of these immigrants and other residents. As

Kindleberger said (1967, p. 213), the short-run benefits of labour migration are clear, but there are dangers in the long run: ‘To rely heavily on foreign labor in one's economy constitutes a positive risk.’
Kindleberger made his entry into economic history with *Economic Growth in France and Britain, 1851–1950* (1964). He surveyed the extensive literature on these two countries and concluded that there was no single convincing explanation for the differences between them. He ended the book with the following famous words: ‘Economic history, like all history, is absorbing, beguiling, great fun. But, for scientific problems, can it be taken seriously?’ This ironic comment set the tone for Kindleberger's future work in economic history. His books and papers are distinguished by his command of the previous literature. His reasoning is informed by an intelligent, if sceptical, use of economic theory. His prose is sprightly. And his conclusions are clear, forcefully presented, and always worth debating.
Kindleberger's impact on economics and economic history comes primarily from two books first published in the 1970s. The first, *The World in Depression, 1929–1939* (1973), provided a comprehensive narrative of the Great Depression from an international perspective. Instead of seeing the Depression as a succession of national stories, Kindleberger argued persuasively that it was the result of a failure of the international economic system. The economic structure built around the gold standard had allowed the pre-war industrial economies to weather various economic shocks in the late 19th and early 20th centuries, but it proved unable to contain or offset the shocks arising in the period after the First World War.
Why so? Kindleberger argued that the inter-war economy lacked a hegemon, a dominant leader. The hegemonic power in the pre-war period was the United Kingdom, more specifically the Bank of England, which acted to contain crises wherever they started. But England was exhausted by the effort to defeat Germany in the First World War, and the Bank of England was in no shape to continue this role. Although the United States was the obvious candidate to pick up the baton, Americans were isolationist after their wartime efforts and declined to act. In the shortest summary: no longer London, not yet New York. Without a hegemon, the shocks to the world economy in the late 1920s were allowed to drag the world into the Great Depression.
The costs of encouraging immigration of foreign workers after the Second World War emerged only slowly; the costs of poor macroeconomic policies in the early 1930s became evident more quickly. Kindleberger recounted the abortive efforts of central bankers and government officials to organize some kind of cooperative solution to the economic shocks. Failing in this endeavour, the world was subjected to competing devaluations and deflations. Among the costs was extensive damage to financial institutions and to the operation of those economies that held on to the gold standard.
Kindleberger generalized his argument in *Manias, Panics, and Crashes: A History of Financial Crises* (1978). He surveyed financial crises in the past two centuries that were important enough to have macroeconomic effects. He described the various irrationalities that preceded crises, as suggested in his title, and synthesized a vast literature in a small and engaging book. He argued that irrationally optimistic expectations frequently emerge among investors in the late stages of major economic booms, differing sharply from most modern models of finance and relying on a more impressionistic theory of financial crises. When these optimistic expectations appear, investors grossly overestimate the future profitability of some promising firms. These overestimates lead unscrupulous managers to over-promote their firms vigorously and to issue bogus debt and equity with abandon. They may lead even well-meaning, sober managers to issue unsupportable amounts of debt. The more a firm's managers sincerely overestimate their firm's growth opportunities or successfully promote a Ponzi-style fraud, the more

securities they try to issue. When the unrealistically high profits fail to develop as predicted, debt and stock values collapse. Markets for over-promoted financial assets may even dry up. The more severe the price decline, the more the collapsing value of previously high-flying assets spreads insolvency to creditors of both the over-expanded firms and their stockholders.

Kindleberger observed that speculation in a bubble often develops in two stages. In the first, sober stage of investment, seasoned professional investors and analysts are gradually persuaded that bubble assets offer a good chance of high returns. In the second stage, ‘professional company promoters – many of them rogues interested only in quick profits – tempted a different class of investors, including ladies and clergymen’. It is of course hard for any market participant or observer to know when the bubble has progressed from the first stage to the second.

Kindleberger concluded that stability is promoted when a lender of last resort exists and follows the recommendations of Walter Bagehot over a century ago in his *Lombard Street* (1873) to lend freely at punitive rates during a crisis. This is what a hegemonic power – the United States government internationally and the Federal Reserve domestically – should have done in the 1930s, in Kindleberger's view; it is what the International Monetary Fund should do today. His book has proved exceedingly popular with a varied audience: economists, investors and the general public alike. It was revised and expanded several times; the fourth edition was published shortly before Kindleberger's death, when he was 90 years old.

See Also

- bubbles
- Great Depression
- Great Depression (mechanisms)

Selected works

1964. *Economic Growth in France and Britain, 1851–1950*. Cambridge, MA: Harvard University Press.

1967. *Europe's Postwar Growth: The Role of Labor Supply*. Cambridge, MA: Harvard University Press.

1973. *The World in Depression, 1929–1939*, rev. edn. Berkeley: University of California Press, 1986.

1978. *Manias, Panics, and Crashes: A History of Financial Crises*. 4th edn. New York: Wiley, 2000.

1993. *A Financial History of Western Europe*, 2nd edn. New York: Oxford University Press.

2000. *Comparative Political Economy: A Retrospective*. Cambridge, MA: MIT Press.

Bibliography

Bagehot, W. 1873. *Lombard Street: A Description of the Money Market*. London: H. S. King.

How to cite this article

Temin, Peter. "Kindleberger, Charles P. (1910–2003)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000073> doi:10.1057/9780230226203.0894

The New Palgrave Dictionary of Economics Online

King, Gregory (1648 – 1712)

Phyllis Deane
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Davenant, C.; demand analysis; Gregory King's Law; King, G.; national income accounting; population estimation

Article

King was born in Lichfield in 1648, the son of a jobbing technician who had acquired a sufficient practical competence in mathematics to earn a modest livelihood practising as a surveyor, a sundial maker, a landscape gardener and even a teacher of bookkeeping. King himself, according to his autobiography, was educated partly at the local Free School and partly at home (for example, in bookkeeping and surveying) until he became clerk to Sir William Dugdale, then Norray King of Arms, whom he served for five years. This appointment set the course of his professional career as a herald, though he was to work for several years in London as a cartographer and engraver before being appointed Rouge Dragon in 1677, Registrar to the College of Arms in 1684 and eventually Lancaster Herald in 1688. After the accession of Queen Anne, when King's Tory bias ceased to be an obstacle to advancement in the public service, he held several appointments of an accounting nature, for example the secretaryship of the Commission of Public Accounts and secretary to the Controller of Army Accounts.

It was in the mid-1690s that King began to take an active interest in political arithmetic, or what contemporaries called ‘the art of reasoning upon things relating to government’, mainly as a result of his friendship with Charles Davenant, who was then playing a major role in the current debate on how to pay for the war. Davenant's *Essay upon Ways and Means of Supplying the War* appeared in print in 1695 and in the following year King wrote his *Natural and Politicall Observations and Conclusions upon the State and Condition of England*, the work that established him as the leading political arithmetician of his day. Henceforth, he and Davenant, a prolific pamphleteer, systematically exchanged ideas and statistical estimates on the main policy issues facing government, and King owed his considerable early 18th-century reputation largely to the polemical use Davenant made of his estimates. Indeed, King's most famous pamphlet was not published in full until 1802, when George Chalmers printed it as an

appendix to his bestseller, *An Estimate of the Comparative Strength of Great Britain During the Present and Four Preceding Reigns*, which had already gone through several editions. It then directly inspired Patrick Colquhoun to produce a comparable estimate of social income, first for England and Wales in *Treatise on Indigence* (1803) and later for the United Kingdom as a whole in *Treatise on the Wealth Power and Resources of the British Empire* (1812). Much later still, in 1936, King's *Observations* was published again, this time together with his only other tract, *Of the Naval Trade of England 1688 and the National Profit Arising Thereby* (MS dated 1697), and with an introduction by the Professor of Statistics at The Johns Hopkins University. His estimates then acquired fresh importance as part of the evidence used by the growing army of mid-20th-century economic statisticians researching long-term trends in population and national income.

Gregory King's claim to fame as a demographer and a national income statistician rests on the imaginative skill, methodical consistency and intellectual integrity with which he compiled and applied the severely limited statistical raw material available to him. His deep respect for the truth and his readiness to respond fully and frankly to those of his contemporaries who doubted the validity of his estimates is exemplary and illuminating. Modern statisticians may argue with his results, but they cannot fail to take them seriously as informed estimates of the dimensions of population, national income and national capital of England at the end of the 17th century and of immediate past trends in those dimensions. Similarly, modern social historians may query the details of social structure implicit in King's oft-reprinted ‘Scheme of the Income and Expence of the several Families of England Calculated for the year 1688’, but the overall picture it presents and the notion of distinguishing between those groups which he conceived of as ‘increasing the wealth of the kingdom’ and those decreasing it, remains a pioneering and instructive analytical device.

The other field to which King is generally believed to have made an original pioneer contribution as an economic statistician is that of demand analysis. The law which is generally referred to as ‘Gregory King's Law’ postulates a systematic relationship between downward deviations from the normal corn harvest and upward deviations in the price of corn. It was attributed to King by Lauderdale in 1804 (though not by Davenant, who first spelt it out with the aid of a numerical example in 1699) and, among others, by Tooke in his *History of Prices*. Jevons estimated the equation implicit in the Davenant exposition in his *Theory of Political Economy* (1871) and so did G. Udny Yule in 1915. Whether it really did originate with King has not yet been established, but the attribution to him rather than to Davenant is highly plausible.

See Also

- political arithmetic

Selected works

1936. *Two Tracts by Gregory King*. Edited with an introduction by G.E. Barnett, Baltimore: Johns Hopkins Press.

Bibliography

Chalmers, G. 1802. *Estimate of the Comparative Strength of Great Britain*. London.

Dallaway, J. 1793. *Inquiries into the Origin and Progress of the Science of Heraldry*. London. (For King's autobiography up to 1694.)

Evans, G.H., Jr. 1967. The Law of Demand – the role of Gregory King and Charles Davenant. *Quarterly Journal of Economics* 81, 483–92.

Glass, D.V. 1965. Two papers on Gregory King. In *Population in History*, ed. D.V. Glass and D.E.C. Eversley. London: Edward Arnold.

Holmes, G.S. 1977. Gregory King and the social structure of pre-industrial England. *Transactions of the Royal Historical Society* 27, 41–68.

Yule, G.U. 1915. Crop production and price: a note on Gregory King's Law. *Journal of the Royal Statistical Society* 78(March), 296–8.

How to cite this article

Deane, Phyllis. "King, Gregory (1648–1712)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000025> doi:10.1057/9780230226203.0895

The New Palgrave Dictionary of Economics Online

kinked demand curve

V. Bhaskar

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The kinked demand curve, one of the staples of oligopoly theory, was originally formulated as a theory of price rigidity. We review dynamic game-theoretic reformulations, which give rise to a theory of collusive price determination.

Keywords

Bertrand competition; collusion; Cournot competition; duopoly; folk theorem; kinked demand curve; Markov perfect equilibria; oligopoly; price rigidity; quick response equilibria; repeated games

Article

The kinked demand curve (Sweezy, 1939; Hall and Hitch, 1939) has been one of the staples of oligopoly theory. It was originally formulated as a theory of price rigidity. A firm conjectures that its rivals will match its price if it reduces the price, but will not match its price if it initiates a price increase. This gives rise to a kink in the firm's perceived demand curve, at the prevailing price. The consequent discontinuity in its marginal revenue curve implies that the firm will not adjust its price in response to small changes in costs, giving rise to price rigidity.

In contrast with the standard Cournot or Bertrand models, the theory represents one of the first attempts at a dynamic model of oligopoly. However, this modelling has been criticized. Implicit in the analysis is the assumption that the firm is motivated by its profits after all price adjustments have taken place. That is, profits in the time interval, where a firm has cut its price and before its rivals have responded, are insignificant. However, if this is so, why does a firm in a symmetric oligopoly not initiate a price increase? If its rivals fail to respond in kind, it can rescind the original increase. Knowing this, its rivals would have an incentive to match its price increase, as long as the original price was below the monopoly price.

To address these questions, one needs to formulate oligopolistic interaction as an explicit dynamic game.

The first option is the standard repeated game model, where one obtains an embarrassment of riches – the ‘folk theorem’ states that every individually rational feasible payoff is an equilibrium payoff, as long as firms are sufficiently patient. (Anderson, 1988, provides a foundation for the kinked demand curve in terms of ‘quick response equilibria’ of a repeated game, where the period length shrinks to zero.) Second, one can model price setting as a dynamic ‘pre-game’ with profits depending only on the profile of final prices that results. This is the modelling choice adopted by Bhaskar (1988) and Kalai and Satterthwaite (1986). Third, Maskin and Tirole (1988) analyse the Markov perfect equilibria of a repeated game where firms take turns in choosing price. These theories of the kinked demand curve are not theories of price rigidity. In all these models, a firm is deterred from undercutting price by the knowledge that its rivals can respond. In consequence, they may be thought of as models of oligopolistic collusion.

We set out a variant of the model of Kalai and Satterthwaite, possibly the simplest of these models. Consider a homogeneous good oligopoly with n firms, where firm i has constant marginal costs c_i. Let $D(p)$ denote market demand when p is the lowest price in the market, and assume that the revenue function, $p.D(p)$, is strictly concave. The game played by the firms has two stages, as follows. In stage 1, firms simultaneously choose prices. Given the vector of prices chosen, $(p_1, p_2, \ldots, p_n)$, let $\bar{p}$ denote the smallest of the prices chosen. In stage 2, firms may choose any price greater than or equal to $\bar{p}$. Our focus is on subgame perfect equilibria where firms do not use weakly dominated strategies in stage 1, given subgame perfect continuation play in stage 2. Let p_i^* denote firm i's optimal common price, that is the unique maximizer of firm i's profits when all firms choose the same price, $p_i^* = \operatorname{argmax}_p \frac{1}{n}(p - c_i)D(p)$. Without loss of generality we may assume that firm 1 has the *minimum optimal common price*. If the cost asymmetries between firms are not too large, then this game has a unique equilibrium. In the first stage, each firm chooses p_i^*, and in stage 2 all firms reduce their prices to p_1^*. That is, the equilibrium outcome is at the minimum optimal common price. The intuition for this result is as follows. In stage 2, one has Bertrand competition with a price floor at $\bar{p}$, the smallest price chosen at stage 1, and all firms will choose $\bar{p}$ as long as it is not too low. Given this, a firm knows that it influences the common equilibrium price only in the event that its price is lower than everyone else's. This ensures that it is weakly dominant at stage 1 for the firm to choose p_i^*.

The model set out here incorporates a restriction on stage two behaviour, namely, that no firm can price below the lowest price chosen at stage 1. To avoid this restriction on undercutting one must formulate a dynamic game without a last stage, since otherwise the Bertrand outcome is irresistible. Bhaskar (1988) sets out a duopoly formulation where firms may repeatedly revise prices downward, and the pre-game ends when no firm seeks to reduce its price. This game produces a similar equilibrium outcome to the one set out above. The theory does not imply price rigidity – if costs increase for firm 1, then this will increase the equilibrium price. The theory also has a flavour of price leadership, since the lowest-cost firm effectively selects the equilibrium price, with the follower firms having to follow suit. Indeed, the follower firms perceive a kinked demand curve at the equilibrium price. If a follower firm were to choose a higher price, firm 1 would not follow suit, thus ensuring that no other firm does so, while if it reduces price, all firms would match this.

Maskin and Tirole (1988) analyse a repeated duopoly where a firm's price is kept fixed for two periods,

and where firms alternate in choosing price. They find multiple Markov perfect equilibria, with the unique symmetric renegotiation proof equilibrium giving rise to a kinked demand curve at the monopoly price.

The traditional kinked demand theory has been criticized on empirical grounds (Stigler, 1947; Primeaux and Bomball, 1974) since oligopoly prices do not appear to be excessively rigid, nor do they show the predicted asymmetry. However, this is not a prediction of the reformulated theories. These theories do predict that in any market, $n - 1$ firms (that is all firms except the leader) should *expect* their rivals to respond asymmetrically to their price changes, at the equilibrium price. Bhaskar, Machin and Reid (1991) analyse survey evidence, where firms were asked how they expected their rivals to respond if they changed price. The survey data finds evidence of asymmetry in expected responses that is consistent with the prediction.

See Also

- oligopoly

Bibliography

Anderson, R.M. 1988. Quick response equilibrium. Mimeo, University of California at Berkeley.

Bhaskar, V. 1988. The kinked demand curve: a game theoretic approach. *International Journal of Industrial Organization* 6, 373–84.

Bhaskar, V., Machin, S. and Reid, G. 1991. Testing a model of the kinked demand curve. *Journal of Industrial Economics* 39, 241–54.

Hall, R. and Hitch, C. 1939. Price theory and business behavior. *Oxford Economic Papers* 2, 12–45.

Kalai, E. and Satterthwaite, M.A. 1986. The kinked demand curve, facilitating practices and oligopolistic competition. In *Imperfections and Behavior in Economic Organizations*, ed. R.P. Gilles and P.H.M. Ruys. Boston: Kluwer.

Maskin, E. and Tirole, J. 1988. A theory of dynamic oligopoly, II: price competition, kinked demand curves, and Edgeworth cycles. *Econometrica* 56, 571–99.

Primeaux, W.J. and Bomball, M.R. 1974. A reexamination of the kinky oligopoly demand curve. *Journal of Political Economy* 82, 851–62.

Stigler, G. 1947. Kinky oligopoly demand and rigid prices. *Journal of Political Economy* 55, 432–49.

Sweezy, P. 1939. Demand under conditions of oligopoly. *Journal of Political Economy* 47, 568–73.

How to cite this article

Bhaskar, V. "kinked demand curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000026> doi:10.1057/9780230226203.0896

The New Palgrave Dictionary of Economics Online

Kitchin, Joseph (1861 – 1932)

S.N. Solomou

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Kitchin, J.; trade cycles

Article

In 1888 Kitchin joined the staff of the *Financial News* as a compiler of statistics, with particular reference to the South African goldfields. In May 1897 he took up a business career in the South African mining industry. As a businessman Kitchin established a wide reputation for his statistical compilations and came to be regarded as a leading authority on the statistics of precious metals. He produced numerous articles on the theme and provided evidence before the Indian Currency Commission (1926), the Committee on Finance and Industry (1930) and the Gold Delegation of the Financial Committee of the League of Nations (1930).

Kitchin's work on money and gold gave him an interest in the study of trade cycles. His first study, which was a description of trade cycles since 1783, was published in *The Times Financial Review* in early 1921. In 1923 he published a study of British and American cycles during 1890–1922. Kitchin distinguished minor cycles of 40 months, major cycles of between 7 and 11 years, and trends dependent on the movement of world money supply. Although the existence of major cycles and secular trends was well established, the existence of a 40-month cycle was original to Kitchin. This cycle was seen to result from the psychological reactions to capitalistic production.

Selected works

1923. Cycles and trends in economic factors. *Review of Economics and Statistics* 5(1), 10–17.

1932. Gold production and consumption. *Review of Economics and Statistics* 14, 126–31.

How to cite this article

Solomou, S. N. "Kitchin, Joseph (1861–1932)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000028> doi:10.1057/9780230226203.0897

The New Palgrave Dictionary of Economics Online

Klein, Lawrence R. (born 1920)

Roberto S. Mariano

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Lawrence Robert Klein, pioneer in economic model building and in econometric forecasting and policy analysis in industry and government, was awarded the Nobel Memorial Prize in economic sciences in 1980. He has established the directions and accelerated the development of the theory, methodology and practice of econometric modelling since the 1940s. He has provided a training ground in applied econometrics for academicians and practitioners worldwide. Lawrence Klein continues to develop and apply econometric methodology for high-frequency forecasting, using weekly and daily information, and for analysing current world economic issues.

Keywords

aggregation; capacity utilization; Clark, C. G.; cost-of-living indexes; Cowles Commission; econometric forecasting; econometrics; estimation; expenditure side model; flow of funds; forecasting; growth models; high-frequency forecasts; income side model; input–output analysis; Keynes, J. M.; Klein, L. R.; Klein–Goldberger model; linear expenditure system; macroeconometric models; mathematics and economics; Morishima, M.; National Bureau of Economic Research; Phillips curve; Phillips, A. W.; principal components model; rational expectations; representative agent; sample survey method; Samuelson, P. A.; simultaneous equations model; statistics and economics; Theil, H.; Tinbergen, J.; two-stage least squares estimator

Article

Lawrence Robert Klein, 1980 Nobel laureate in economics, has been a pioneer in economic model building and in developing a worldwide industry in econometric forecasting and policy analysis. As Klein's Nobel citation states, ‘Few, if any, researchers in the empirical field of economic science have had so many successors and such a large impact as Lawrence Klein.’ When one thinks of macroeconometric models, his name is the first that comes to mind. Spanning six decades, his research

achievements have been broad, covering economic and econometric theory, methodology, and applications. In emphasizing the integration of economic theory with statistical methods and practical economic decision-making, he played a key role in establishing the directions and in accelerating the development of the theory, methodology and practice of econometric modelling.

His pioneering efforts in the 1940s built on the earlier works of Tinbergen (in the League of Nations Secretariat in 1936–8) and Haavelmo (1943), the seminal treatise of Keynes (1936), and the then emerging toolkit in mathematics and statistics for economic analysis. He was one of the first to establish an operational paradigm for macroeconometric models, and he developed statistical techniques for the estimation and application of these models. Always willing to give generously of himself as he interacted with students and colleagues, he has provided a training ground in applied econometrics for an impressive and long list of academicians, government officials and corporate executives from all over the world. Lawrence Klein continues to contribute in developing and applying econometric methodology for high-frequency forecasting, using weekly and daily information, and for analysing current world economic issues.

Lawrence Klein was born in Omaha, Nebraska on 14 September 1920. He obtained his undergraduate degree from the University of California at Berkeley in 1942 and completed his Ph.D. in Economics at Massachusetts Institute of Technology (MIT) in 1944. He has been professor of economics at the University of Pennsylvania since 1958. He founded Wharton Econometric Forecasting Associates (WEFA) and, as a principal investigator at the University of Pennsylvania, helped with Bert Hickman and Aaron Gordon to establish Project LINK. Together with Michio Morishima, he founded the *International Economic Review* as a joint publishing endeavour of Osaka University and the University of Pennsylvania. He has been President of the American Economic Association (1977), President of the Econometric Society (1960), editor-in-chief of *International Economic Review* (1959–65), and John Bates Clark Medalist (1959).

In 1980, he was awarded the Nobel Prize in Economics ‘for the creation of econometric models and their application to the analysis of economic fluctuations and economic policies’ (prize citation in the Alfred Nobel Memorial Prize in Economic Sciences, 1980; also in Lindbeck, 1992, p. 411).

Klein's experience as a youth in the Great Depression and his intense desire to understand what was going on led him to the study of economics. After spending two years in Los Angeles City College, Klein completed his last two undergraduate years at the University of California in Berkeley. With a keen interest in seeing how mathematics and statistics can be used in analysing economic problems, he worked with students of pioneers like Griffith Evans (professor of mathematics and a founding member of the Econometric Society) and Jerzy Neyman (professor of statistics and key developer of statistical theory). He also worked as a summer research assistant of George Kuznets in the Giannini Foundation. This summer work exposed him to perhaps his first foray into applied econometrics – estimating demand functions for Californian lemons! It was also during this time that Klein was introduced to the early scholarly works of Paul Samuelson, a serendipitous preparation for a long-time relationship that was to blossom as Klein moved to MIT for his doctoral studies.

On graduate scholarship at MIT, Klein was Paul Samuelson's research assistant from the outset. As Klein himself remarked:

> Working with Samuelson, who was at the forefront of interpreting Keynesian theory for teaching and policy applications, I was put immediately in the midst of two challenging

> contests – one to gain acceptance for a way of thinking about macroeconomics and another to gain acceptance for a methodology in economics, namely, the mathematical method. Later, both challenges were to be overcome, but for ten or twenty years, opposition was fierce. (Breit and Hirsch, 2004, p. 18)

He would elaborate further on this:

> working as an assistant for Samuelson was something that is very hard to duplicate anywhere in the world. He generates ideas so fast. At that time, there was a whole succession of ideas concerning Keynesian macroeconomics and econometrics and the development of mathematical methods in economics. It was a very exciting time, and I felt very fortunate to be in that background. (Mariano, 1987, p. 411; and Klein, 2006)

At that time, when Haavelmo's celebrated *Econometrica* paper (Haavelmo, 1943) was circulating as a working paper, the treatment of identification in econometric models led Samuelson to ask Klein to investigate the mathematical equivalence between the problems of identification in supply–demand models and in saving–investment analysis.

It was during his graduate student days that he started working on two papers that were later published in *Econometrica* and the *Journal of Political Economy*. The first, published in 1943, studied the specification of the investment function, while the second (1947) dealt with alternative theories of effective demand. Considered a seminal paper in the debate between the Keynesians and the classical economists, this latter paper formulated the Keynesian system in mathematical terms and argued that the specification of the liquidity preference function and determination of money wages are keys to the Keynesian system.

Klein completed his degree in two years, as Samuelson's first Ph.D. student. His thesis, dealing with Keynesian economics, led in 1947 to the publication of *The Keynesian Revolution*, which was to become one of Klein's best-known works. The book provided the mathematical specification of Keynes's ideas that served as the foundation for the economic models that Klein formulated subsequently.

After finishing at MIT, Klein accepted Jacob Marschak's invitation in 1944 to become a research associate in the Cowles Commission at the University of Chicago. This turned out to be a defining period for Klein's professional career. His interactions with an unusually talented group that included J. Marschak, T.W. Anderson, H. Rubin, M. Girschick, T. Haavelmo, T. Koopmans, D. Patinkin, L. Hurwicz, K. Arrow, H. Simon, R. Leipnik, H. Chernoff, and visitors such as J. Tinbergen, R. Frisch and M. Kalecki proved to be a catalyst for his development into an applied econometrician par excellence. His MIT work on Keynesian economics began to evolve at this time into applied econometric modelling. While teams were formed to work on various aspects of an emerging econometrics field, Klein focused his energies on what was to become his lifelong endeavour. He described his task in the Cowles Commission as follows:

> The central problem posed for research at Cowles by Jacob Marschak and Tjalling Koopmans was a fresh attempt at U.S. model building by using Haavelmo's new ideas about econometric theory. … Jacob Marschak insisted that I base my econometric

> modelling on received economic theory and that I justify macroeconomic specifications on the basis of reasoning about individualistic decision making, with proper attention to the problem of aggregation. … It turned out to be an exciting time for me and enabled me to build on the Keynesian lessons that were taught to me at MIT by Paul Samuelson. … That was the beginning of my long association with the problems of macroeconometrics that were then being tackled afresh at the Cowles Commission. (Marwah, 1997, pp. xx–xxii)

At Cowles, Klein completed his first series of macroeconometric models. The celebrated Klein Model 1 was part of this series. It was initially developed as a compact prototype model of the US economy to study computational methods. It has now become a standard reference in most introductory econometrics textbooks. Klein also put his models to work to answer pressing questions about the post-war US economy posed by professional colleagues like Albert Hart from the Committee for Economic Development, Theodore Yntema (former director of the Cowles Commission), and Alfred Cowles himself, who was a member of the Budget Committee of the Community Fund in Chicago. Klein's models proved useful in forecasting what was in store after the war – predicting that the US economy would not return to the Great Depression.

Klein's interactions with his peers at Cowles deepened his interest in statistical methodology, especially in estimation and prediction in simultaneous equations models. He developed a keen attention to detail in estimation in empirical work and a firm belief in the value of ‘high technology’ estimation procedures. At this juncture, he also started his joint work with Herman Rubin on the linear expenditure system for studying cost-of-living indexes in the context of a neoclassical demand model (see Klein and Rubin, 1947) as well as on aggregation issues and demand systems. During this period, Klein completed most of the material for another major work, *Economic Fluctuations in the United States, 1921–1941*, which was published in 1950.

In the summer of 1947, Klein left the Cowles Commission, briefly to help in the initial econometric model building effort in Canada, then to spend the greater part of the year to visit Ragnar Frisch's Institute in Oslo and Jan Tinbergen's office in the Central Planning Bureau in the Netherlands.

Klein then joined the National Bureau of Economic Research (NBER), at the invitation of Arthur Burns, to undertake econometric studies of production functions. Interested in investigating the influence of liquid assets on saving behaviour, Klein moved to the University of Michigan in 1949, initially as researcher in the Survey Research Center for one year, then lecturer in economics from 1950 to 1954. Having become involved with the sample survey studies in the Center, Klein produced a number of publications on savings and consumption behaviour using survey data, culminating in a book on the contributions of survey methods in economics (Katona, Klein, Lansing and Morgan, 1954).

While at Michigan, Klein noticed a considerable interest in the forecasts about the state of the economy and the use of econometric models and also resumed the econometric modelling work that he started at the Cowles Commission. With Arthur Goldberger, his doctoral student at Michigan, he developed what has come to be called the Klein–Goldberger model of the US economy. As the first substantial effort at an empirical representation of a large economy with a theoretical Keynesian structure (see Klein and Goldberger, 1955), this model has become an important reference to students and researchers in econometrics. This model has become a standard example in econometric textbooks, making very realistic simulation projections about the small recession following the Korean War. Klein and

Goldberger initially published this result in the *Manchester Guardian* to challenge a very pessimistic econometric forecast by Colin Clark. Another major piece of work during this time was his textbook on econometrics, the first to provide a blend of theoretical, methodological and applied developments in econometrics (Klein, 1953).

The University of Michigan was to promote Klein to full professorship but then reneged when Klein testified in a Detroit hearing that he had been a member of the Communist Party for about six months in 1946. (Subsequently, in 1978, the university awarded Klein an honorary doctoral degree in which the citation stated that he would probably be a Nobel laureate.) Oxford University's Institute of Statistics quickly invited him to join its staff, which he accepted, becoming first Senior Research Officer from 1954 to 1955 and Reader in Econometrics from 1956 to 1958. As Klein himself explained in Breit and Hirsch (2004): ‘In the McCarthy era, I left Michigan for the peace and academic freedom in Oxford.’ At that time, the Oxford Institute of Statistics was undertaking the Oxford Savings Surveys in partnership with the UK government, an enterprise in which Klein played a substantial role. He also developed an econometric model of the UK economy, which was published in Klein, Ball, Hazlewood and Vandome (1961).

In Oxford, Klein was ‘given the green light to do what he thought could be done within the confines of the Oxford system in teaching, attracting attention in seminars, and doing research activities in econometrics’ (Mariano, 1987, p. 422). It was in this period that he produced his more intuitive instrumental variable interpretation of Theil's two-stage least squares estimator (Klein, 1955). Carrying over research initiatives from Cowles Commission work, Klein also looked into the statistical efficiency gains from imposing a priori restrictions on an economic system. Klein had numerous productive discussions with colleagues, including Peter Vandome and Michio Morishima, and with A. W. Phillips about the Phillips curve and how it relates to his own ideas about closing the Keynesian system for the determination of absolute prices and wages. His discussions with Jim Ball and Peter Newman about growth theory and growth models led to his idea of constructing a total growth model of the economy in terms of stable ratios as limiting conditions in economics. Some of these ideas that circulated in Oxford were eventually refined in Klein's early years at the University of Pennsylvania (for example, Klein and Kosobud, 1961).

Klein returned to the United States in 1958 – partly under family pressure, to help ageing parents – and joined the economics faculty of the University of Pennsylvania. University President Gaylord Harnwell and Provost Jonathan Rhoads told Klein that they did not have any interest in his political beliefs and simply wanted him at the University of Pennsylvania to teach econometrics: the University of Pennsylvania remains his main base of operations. He produced a host of academic publications and contributions to both economics and econometrics and he created innovative ways of financing major economic research with fresh linkages with industry and government. He also played a key role in shaping the Economics Department into its position today as one of the top economics departments in the United States.

Klein's academic research at the University of Pennsylvania returned to favourite themes such as estimation and prediction in simultaneous equations models of economic systems (Johnston, Klein and Shinjo, 1974; Klein, 1969; Klein, Dhrymes and Steiglitz, 1970; Klein and Howrey, 1972; Klein and Nakamura, 1962; and Klein and Young, 1980). Some work opened up new issues such as the theoretical and empirical difficulties involved in measuring and tracking capacity utilization (Klein, 1960a; Klein and Summers, 1967; Klein and Preston, 1967; Klein and Su, 1979). Klein's subsequent research themes

delved into economic techniques, analysis, and policy, dealing with diverse topics in economic theory, econometric methodology and forecast uncertainty, microfoundations and linkages of the macro Keynesian paradigm, the role of expectations in empirical economic models, anticipations and forecasting, the Phillips curve, international economics and finance, economic growth, and policy formulation. At the same time, in a synergistic fashion, he continued and sustained his work on macroeconometric modelling, developing numerous econometric models for a vast array of applications. Prominent examples of these are the SSRC–Brookings model of the US economy, the Wharton School models of the US economy (medium-term and long-term), and the Project LINK world model.

Klein's methodological approach in econometrics blends economic analysis, statistical method and mathematics. Many times in his writings he would strongly recommend that the best approach to applied economic modelling is to first develop an underlying theory, then move on to observation and the preparation of a database with the statistical methodology to construct, test and apply the empirical model.

He is wary of oversimplification in economic modelling, because, he says, the problems are complicated and can be understood only in the context of large complex systems. His starting premise is that ‘the real world is very complicated and cannot be effectively understood or guided by simple rules, such as those that underlie monetarism or those that can be treated by single equation time series methods or even those that can be treated by vector autoregression (VAR) methods’ (Marwah, 1997, p. xxiv). It is his long-standing conviction that detailed structural modelling is the best kind of system for understanding the macroeconomy through its causal dynamic relationships, specified by received economic analysis. However, with more time-series information becoming available on weekly, daily, hourly, and real-time basis, he also feels that there are related approaches, based on indicator analysis, that are complementary, especially for use in high-frequency analysis.

Klein disagrees with the notion that macroeconomics is simply an adding up of the propositions of microeconomics (for example, see Klein, 1993). He argues that macroeconomics stands on its own as a separate subject and cannot be entirely derived from microeconomics. In his view, there are important concepts and analyses that are inherently macroeconomic. And, of course, there are also important macroeconomic propositions that can be derived from microeconomics, but only after paying painstaking attention to the formulas and processes of aggregation. And on the issue of aggregation, which he had studied since his undergraduate days in Berkeley, he maintains in subsequent analyses that ‘macromodeling in terms of unweighted aggregates or, even worse, in terms of the “representative agent”, fails to deal with the relevant distribution issues’ (Marwah, 1997, p. xxi). There are two dimensions to aggregation, over commodities and services and over economic units (firms and households). Specific and narrow market analyses, involving intricate aggregation over economic units, are important in price determination; yet they are not purely microeconomic since they involve aggregation in various dimensions.

Klein believes that the market system cannot provide adequate self-regulatory responses in an economy. The economy definitely needs guidance and Klein looks to professional economists to provide policymakers with the right information for appropriate decision-making and leadership. On methods for doing this, according to Klein, there is no alternative to the quantitative approach of econometrics, but with the realization that not all policy issues are quantitative and measurable, and that subjective decisions must also be made (see Klein, 1992). Furthermore, econometric information must be detailed if it is to be useful in policy formation. In general, there is a need to move in the direction of preparation

of large-scale complex systems in order to help policymakers. Significant advances in computer technology and the provision of detailed information through associated telecom processes make it possible to push econometrics in the direction of truly serving policymakers (see Klein, 1986).

Klein also sought to extend the narrow Keynesian model serially to the supply side, the open economy and the developing economy by integrating his conception of Keynesian theory with other branches of economics such as international trade and economic development. And he enhanced his model structure further not only with the flow-of-funds accounts but also with the introduction of input–output analysis. He felt strongly that the economic model structure must interface with the social accounts especially when supply side, industry, and longer-term analysis are of major concerns in the study. Thus, Klein's modelling team at the University of Pennsylvania produced standardized procedures for combining input–output analysis with macroeconometric modelling in a feedback mode (see Klein, 1989; Klein, Bodkin and Marwah, 1991).

Klein draws upon explicit surveys of consumer and manufacturer expectations to develop a powerful and meaningful way of dealing with expectations in macroeconometrics. He believes that the ‘rational expectations’ approach – where expectations are treated to be fully consistent with the model being estimated – is ‘unrealistic and singularly unhelpful in guiding economic policy or in forecasting’ (Marwah, 1997, p. xxiii). The most important analyses of expectations will come through the in-depth use of the sample survey method. Along these lines, Klein's research sought to endogenize measured expectations and to include anticipatory variables in his macroeconometric models (orders, investment intentions, housing starts, building permits, survey responses about future spending, incomes, or price movements: see Klein, 1972; Adams and Klein, 1972; and Klein and Ozmucur, 2007).

In the early stages of his career, Klein considered the real sector as the key focus of the analysis and that a good understanding of the economy is possible without careful reference to the monetary sector. But in studying the macroeconomy, he has increasingly come to appreciate the role of money and of the whole monetary sector. For example: ‘Monetarism is fundamentally flawed, and dangerous when used as a doctrinaire policy approach, but I do believe that money matters; it is not everything but it does matter’ (Klein, 1992, p. 188; also in Marwah, 1997, p. xliv). But he remarks further that science, technology, development, and innovation play important roles in the dynamics of the economy and that this interpretation of the supply side is different from and far more important than the simplistic and populist approaches through tax cuts.

In using his empirical models to forecast, Klein has always been concerned with how to adjust the model so that it would start on a forecast extrapolation at prevailing initial values. A particular concern was the frequent data revisions and new information flows about exogenous and endogenous variables at the very moment of forecast calculation. One approach – subjective adjustments to initialize the extrapolation process – is not replicable and is not satisfactory. Since the 1980s, Klein has been using time series methods to extrapolate higher frequency indicators for purposes of initializing the empirical macromodel (Klein and Young, 1980; Klein and Sojo, 1989; Klein and Park, 1993; Klein and Ozmucur, 2007).

Klein has always been conscious of the statistical uncertainties involved in the results of econometric model building and application. Consequently, coping with forecast errors from macroeconometric models is a constantly recurring theme in his research agenda. And he has developed methods for assessing the degree of uncertainty in econometric inference, in particular model simulation techniques to evaluate forecast errors of large-scale models and to perform sensitivity analyses of these models

(Klein and Howrey, 1972; Johnston, Klein and Shinjo, 1974; Klein and Marquez, 1989; and Klein, 1994).

Construction of the Wharton models started with a Rockefeller Foundation grant in the early 1960s. This was followed by the establishment of the Wharton Econometric Forecasting Unit as a research group at the Wharton School for general quantitative studies in economics, financed by the Ford Foundation and the National Science Foundation. Subsequent funding came from several major corporations that sought Klein's help in econometric model building to assist their economic research departments. Since this activity thrived, the unit was formally incorporated in 1969 as a non-profit entity, fully owned by the University of Pennsylvania, and with the name Wharton Econometric Forecasting Associates (WEFA). Through WEFA, Klein was able to tap major sources of funding and channel these towards the establishment of the University of Pennsylvania as a premier centre of academic research in applied econometric analysis. Through the 1960s, 1970s, and 1980s, WEFA earnings from research consultancies for private companies, and public agencies were ploughed back to support economics faculty, graduate students, and visiting scholars in the University of Pennsylvania. The commercial work within WEFA itself also pioneered the logical development and computer handling of large-scale systems.

While at the University of Pennsylvania, Klein also pursued his interests in international model building. In the 1960s, he started his work on modelling the economies of Japan, the Organisation for Economic Co-operation and Development (OECD), and Latin American countries, starting with Mexico, Brazil and Argentina. These model-building efforts then spread out into many developing countries in Asia, including China, and some in the Middle East. Klein headed the first delegation of academic economists from the US to China in 1979. The following year, in collaboration with Lawrence Lau, Klein convinced the Chinese Academy of Social Sciences to host an econometric workshop at the Summer Palace in Beijing. The workshop staff, consisting of Klein, T.W. Anderson, Albert Ando, Lawrence Lau, Gregory Chow, Cheng Hsiao and Vincent Su, introduced econometrics and related aspects of empirical economic model building to the then nascent community of Chinese economists. Klein also directed related efforts to socialist nations. All these efforts naturally led to Project LINK.

Project LINK was one of the biggest and most ambitious projects that Klein mounted, with initial funding support in 1968 from the Ford Foundation, National Science Foundation, the International Monetary Fund and the Federal Reserve Board. The project sought to integrate the macroeconometric models of different countries, which eventually included Third World and socialist nations, into a total simultaneous system through international trade and financial flows (for example, see Klein, 1983; and Klein and Hickman, 1984). The main objective was to improve understanding of international economic linkages and to make improved forecasts of world trade. Over the years, this worldwide project has provided an important research forum that brings together model builders from many countries to share each other's developments and to discuss in a systematic way world economic prospects and pressing economic policy challenges. Project LINK also has provided a critical impetus for the development of economic and econometric analysis in socialist and Third World countries. Today, Project LINK is headquartered at the University of Toronto and the United Nations office in New York. It involves approximately 100 countries worldwide.

One of Klein's research works now deals with forecasting with high-frequency data through the University of Pennsylvania Current Quarter Model (CQM) (see Klein and Sojo, 1989; Klein and Park, 1993; and Klein and Ozmucur, 2007). This model embodies a constant effort to improve forecasts by

combining results from different methods, namely the expenditure side model, income side model, and the principal components model of an economy. It combines data at different frequencies to enable use of all available information. High-frequency forecasts are useful not only for studying the short-term developments of the economy but also for adjusting lower-frequency macroeconometric models so that they are solved from up-to-date initial conditions. The University of Pennsylvania Current Quarter Model has generated a great deal of interest in high-frequency macroeconometric models. Klein has authored or advised the building of similar models for Russia and China (see Klein and Mak, 2005) and he is in the process of contributing to building a model for India. There are also efforts, mostly by Klein's students, to build high-frequency models for other countries such as Japan, Mexico, Hong Kong, France and the European Union.

Klein's efforts to improve forecasting accuracy also have moved in the related direction of building models that include survey results (households, investors, and managers: see Klein and Ozmucur, 2007). And his constant attempt to answer pressing substantive issues has led to recent applied and technical papers on using input–output tables with econometric models (Klein, 2003; on information technology and productivity, see Klein, Duggal and Saltzman, 1999; 2003; 2004; on estimating China's economic growth rate, Klein and Ozmucur, 2002/2003; and on financial crises' challenges and cures, see Klein, Mariano and Ozmucur, 2007; and Klein and Shabbir, 2007).

In addition to the academic activities that this article has focused on, Klein has been active throughout his career in non-academic pursuits as well. He chaired the economic task force of Jimmy Carter in the US presidential election campaign in 1976. He served on the Finance Committee of the National Academy of Sciences and on the Board of Directors of W.P. Carey Co. He found time to contribute journalistic pieces about economic affairs to the *Los Angeles Times*, *Manchester Guardian*, and *Banker's Magazine*; and to be a founding officer and active moving force of the Economists Allied for Arms Reduction (ECAAR), now Economists for Peace and Security (EPS). All these activities have produced an illustrious line-up of students and colleagues who have benefited from their collegial training and/or collaboration with Klein. Many of them are in the highest levels in major academic institutions, leading companies in the private sector, multinational organizations and government agencies all over the world.

Klein has developed and moulded macroeconometric models for over six decades in his own inimitable way, addressing the soundness of the theoretical basis for the model specification, using empirical evidence and data and appropriate methodology to estimate and validate and apply the model for the purposes that drove its creation. Over the years, these models have evolved in terms of complexity, breadth, and new econometric methodologies. And these models have been constructed and applied to address a wide variety of issues such as post-war economic policy formation in the 1940s (his mission at the Cowles Commission), the oil price shocks and the ensuing stagflation in the United States in the 1970s, impact and policy implications of the financial crisis in the 1990s, impact of tariff and non-tariff barriers on regional trade flows, policy analysis of regional trade groupings and various international agreements (for example, Uruguay Round, NAFTA, APEC, WTO), capital flows, economic development all over the world, and explaining endogenous exchange rates after the demise of Bretton Woods parities.

Though Klein had the benefit of Tinbergen's early work as well as the post-war effort at the Central Planning Bureau in the Netherlands, Klein's modelling work was akin to the original model T Ford out of which many other models developed in different parts of the world. Klein's own work has shaped developments in the field quite uniquely and has influenced model builders on a worldwide scale.

‘His principal achievement has been in pioneering an activity in the field of economic model building which has required foresight, persistence and great technical skill and which has been translated into a paradigm of research activity that has spread wherever statistical economics is taught and wherever models are built’ (Ball, 1981, p. 92). Undoubtedly, he continues to inspire, teach, lead, explore and push the intellectual and academic frontiers of the pursuits that continue to define his lifetime of pioneering work.

See Also

- econometrics
- IS–LM in modern macro
- Keynesian revolution
- Lucas critique
- macroeconomic forecasting
- Phillips curve (new views)
- simultaneous equations models
- structural vector autoregressions
- two-stage least squares and the *k*-class estimator
- United States, economics in (1945 to present)

Selected works

1943. Pitfalls in the statistical determination of the investment schedule. *Econometrica* 11, 246–58.

1947. Theories of effective demand and employment. *Journal of Political Economy* 55, 108–31.

1947. *The Keynesian Revolution*, 2nd edn. New York: Macmillan, 1966.

1947. (With H. Rubin.) A constant-utility index of the cost of living. *Review of Economic Studies* 15, 84–7.

1950. *Economic Fluctuations in the United States 1921–1941*, Cowles Commission Monograph 11. New York: Wiley.

1953. *A Textbook of Econometrics*, 2nd edn. New Jersey: Prentice-Hall, 1974.

1954. (With G. Katona, J. Lansing and J. Morgan.) *Contributions of Survey Methods to Economics*. New York: Columbia University Press.

1955. On the interpretation of Theil's method of estimating economic relationships. *Metroeconomnica* 7,

147–53.

1955. (With A.S. Goldberger.) *An Econometric Model of the United States 1929–1952*. Amsterdam: North-Holland.

1958. The estimation of distributed lags. *Econometrica* 26, 553–65.

1960a. Some theoretical issues in the measurement of capacity. *Econometrica* 28, 272–86.

1960b. Single equation vs. equation system methods of estimation in econometrics. *Econometrica* 28, 866–71.

1961. (With R.J. Ball, A. Hazlewood and P. Vandome.) *An Econometric Model of the United Kingdom*. Oxford: Basil Blackwell.

1961. (With R.F. Kosobud.) Some econometrics of growth: great ratios of economics. *Quarterly Journal of Economics* 75, 173–98.

1962. (With M. Nakamura.) Singularity in the equation systems of econometrics: some aspects of the problem of multicollinearity. *International Economic Review* 3, 274–99.

1967. (With R. Preston.) Some new results in the measurement of capacity utilization. *American Economic Review* 57, 34–58.

1967. (With R. Summers.) *The Wharton Index of Capacity Utilization*. Philadelphia: Wharton School, University of Pennsylvania.

1969. Estimation of interdependent systems in macroeconometrics. *Econometrica* 37, 171–92.

1970. (With P.J. Dhrymes and K. Steiglitz.) Estimation of distributed lags. *International Economic Review* 11, 235–50.

1972. The treatment of expectations in econometrics. In *Uncertainty and Expectations in Economics*, ed. C.F. Carter and J.L. Ford. Oxford: Blackwell.

1972. (With F.G. Adams.) Anticipations variables in macroeconometric models. In *Human Behavior in Economic Affairs: Essays in Honor of George Katona*, ed. B. Strumpel, J.N. Morgan and E. Zahn. Amsterdam: Elsevier.

1972. (With E.P. Howrey.) Dynamic properties of nonlinear econometric models. *International Economic Review* 13, 599–618.

1973. The treatment of undersized samples in econometrics. In *Econometric Studies of Macro and Monetary Relations*, ed. A.A. Powell and R.A. Williams. Amsterdam: North-Holland.

1974. (With H.N. Johnston and K. Shinjo.) Estimation and prediction in dynamic econometric models. In *Econometrics and Economic Theory*, ed. W. Sellekaerts. London: Macmillan.

1975. (With K. Johnston.) Stability in the international economy: the LINK experience. In *International Aspects of Stabilization Policies*, ed. A. Ando, R. Herring and R. Marston. Boston: Federal Reserve Bank of Boston.

1979. (With V. Su.) Direct estimates of unemployment rate and capacity utilization in macroeconometric models. *International Economic Review* 20, 725–40.

1980. (With R.M. Young.) *An Introduction to Econometric Forecasting and Forecasting Models*. Lexington, MA: D.C. Heath.

1981. The LINK project. In *International Trade and Multi-Country Models*, ed. R. Courbis. Paris: Economica.

1981. (With G. Fromm.) Scale of macro-econometric models and accuracy of forecasting. In *Large Scale Macroeconometric Models*, ed. J. Kmenta and J. Ramsey. Amsterdam: North-Holland.

1983. International models for forecasting and policy formation. *Prévision et Analyse Economique* (Cahiers du Gama) 4, 93–113. Repr. in Marwah (1997).

1984. (With B. Hickman.) Wage-price behavior in the national models of project LINK. *American Economic Review* 74, 150–4.

1986. Economic policy formation: theory and implementation (applied economics in the public sector). In *Handbook of Econometrics*, vol. 3, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

1989. Econometric aspects of input-output analysis. In *Frontiers of Input-Output Analysis*, ed. R.E. Miller, K.R. Polenski and A.Z. Rose. Oxford: Oxford University Press.

1989. (With E. Sojo.) Combinations of high and low frequency data in macroeconometric models. In *Economics in Theory and Practice: An Eclectic Approach*, ed. L.R. Klein and J. Marquez. Dordrecht: Kluwer.

1989. (With J. Marquez.) *Economics in Theory and Practice: An Eclectic Approach*. Dordrecht: Kluwer.

1991. (With R. Bodkin and K. Marwah.) *A History of Macroeconometric Model-Building*. Aldershot:

Edward Elgar.

1992. My professional life philosophy. In *Eminent Economists: Their Life Philosophies*, ed. M. Szenberg. Cambridge: Cambridge University Press. Repr. in Marwah (1997).

1993. What is macroeconomics? In *Monetary Theory and Thought*, ed. H. Barkai, S. Fischer and N. Liviatan. London: Macmillan. Repr. in Marwah (1997).

1993. (With J.Y. Park.) Economic forecasting at high frequency intervals. *Journal of Forecasting* 12, 301–19.

1994. Economic forecasting and decision-making under uncertainty. In *Money, Inflation, and Employment, Essays in Honor of James Ball*, ed. S. Holly. Aldershot: Edward Elgar. Repr. in Marwah (1997).

1999. (With V.J. Duggal and C. Saltzman.) Infrastructure and productivity: a nonlinear approach. *Journal of Econometrics* 92, 47–74.

2002/2003. (With S. Ozmucur.) The estimation of China's economic growth rate. *Journal of Economic and Social Measurement* 28, 187–202.

2003. The use of input-output tables to estimate the productivity of IT. *Journal of Policy Modeling* 25, 471–5.

2003. (With V.J. Duggal and C. Saltzman.) Information technology and productivity: the case of the financial sector. *Survey of Current Business*, August, 32–7.

2004. (With V.J. Duggal, and C. Saltzman.) Contributions of input-output analysis to the understanding of technological change: the information sector in the United States. In *Wassily Leontief and Input-Output Economics*, ed. E. Dietzenbacher and M.L. Lahr. Cambridge: Cambridge University Press.

2005. (With W. Mak.) Initial steps in the high-frequency modeling of China. *Business Economics* 40, 11–14.

2006. Paul Samuelson as a ‘Keynesian’ economist. In *Samuelsonian Economics and the Twenty-First Century*, ed. M. Szenberg, L. Ramrattan, and A. Gottesman. Oxford and New York: Oxford University Press.

2007. (With S. Ozmucur.) The University of Pennsylvania models for high-frequency macroeconomic modeling. In *Econometric Forecasting and High-Frequency Data Analysis*, ed. R.S. Mariano and Y.K. Tse. Singapore: World Scientific.

2007. (With R.S. Mariano and S. Ozmucur.) Capital controls, financial crises and cures: simulations with an econometric model for Malaysia. In *Recent Financial Crises, Challenges and Implications*, ed. L.R. Klein and T. Shabbir. Cheltenham: Edward Elgar.

2007. (With T. Shabbir.) *Recent Financial Crises: Analysis, Challenges and Implications*. Cheltenham: Edward Elgar.

Bibliography

Ball, R.J. 1981. On Lawrence R. Klein's contributions to economics. *Scandinavian Journal of Economics* 83, 81–103.

Breit, W. and Hirsch, B.T. 2004. *Lives of the Laureates: Eighteen Nobel Economists*, 4th edn. Cambridge, MA: MIT Press.

Courbis, R. 1981. *International Trade and Multi-Country Models*. Paris: Economica.

Haavelmo, T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1–12.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. New York: Harcourt Brace.

Lindbeck, A. 1992. *Nobel Lectures: Economic Sciences 1969–1980*. Singapore: World Scientific.

Mariano, R.S. 1987. The ET interview: Professor L.R. Klein. *Econometric Theory* 3, 409–60.

Mariano, R.S. and Tse, Y.K., eds. 2007. *Econometric Forecasting and High-Frequency Data Analysis*. Singapore: World Scientific.

Marquez, J. 1985. *Economic Theory and Econometrics*. Oxford: Blackwell.

Marwah, K. 1997. *Selected Papers of Lawrence R. Klein: Theoretical Reflections and Econometric Applications*. Singapore: World Scientific.

Sichel, W. 1989. *The State of Economic Science, Views of Six Nobel Laureates*. Michigan: W.E. Upjohn.

Szenberg, M. 1992. *Eminent Economists: Their Life Philosophies*. Cambridge: Cambridge University Press.

How to cite this article

Mariano, Roberto S. "Klein, Lawrence R. (born 1920)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000068> doi:10.1057/9780230226203.0898

The New Palgrave Dictionary of Economics Online

Knapp, Georg Friedrich (1842 – 1926)

B. Schefold

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; feudalism; gold standard; hereditary bondage; Knapp, G. F.; monetary theory; money; Mortality; paper money; peasants; state theory of money

Article

Knapp was born in Giessen, the son of a professor of technological chemistry who was also temporarily director of the Königliche Porzellanmanufaktur. Justus von Liebig, the famous chemist, was his uncle. Knapp studied in Munich, Berlin and Göttingen, and in 1867 became head of the statistical office of the municipality of Leipzig, in 1869 extraordinary professor of economics in Leipzig and in 1874 professor in Strassburg. He was one of the leading German ‘Kathedersozialisten’ (socialists of the chair), and cofounder of the Verein für Socialpolitik.

At the beginning of his career he carried out some important work in statistics: he was the first to develop a systematic theory of mortality measurement (1868), and he applied mathematical methods to demographical problems (1874). After his appointment to Strassburg in 1874, his research shifted to German agricultural history. He compared the economic organization of agriculture in the different parts of Germany (1925–7, vol. 1, ch. 3), his special interest being focused on the agrarian conditions in the German East. In a work now regarded as classic (1925–7, vols 2 and 3), Knapp described the peasant liberation and the rise of a class of rural workers in the long-settled provinces of Prussia. Around the turn of the 19th century, property relations in the rural parts of Eastern Prussia were dominated by the estate economy (Gutsherrschaft), which had arisen out of medieval landlordship (Grundherrschaft). It was characteristic of the estate economy that the peasants, who could own land, were obliged to do compulsory service on the land of the Junker, and remained in hereditary bondage (Erbunterthänigkeit). In Knapp's view, the latter had to be distinguished from slavery/serfdom and was not medieval at all, but inherently modern. Knapp perceived the estate economies as the first large capitalist enterprises and regarded hereditary bondage as the earliest capitalist labour constitution, a very controversial view. Thus Knapp emphasized that the origins of capitalism should be sought in agriculture (1925–7, vol. 1, ch. 2, pp. 91–106).

This system was to be changed by the reforms under Stein and Hardenberg (1925–7, vol. 1, ch. 2, pp. 107–23), which aimed at the abolition of hereditary bondage. However, under the pressure of the Junker, who were interested in maintaining an abundant labour supply for their estates, only wealthy peasants were allowed to own landed property; the others were transformed into wage labourers working on the Junker estates. Compensation payments for the Junker allowed them to absorb many peasant holdings. This implied a loss of the former feudal protection of the labourers. Thus the Prussian agrarian reforms resulted in a restructuring of the organization of agriculture, which worsened the social situation of large parts of the peasantry and strengthened the position of the Junker as the dominant class of Prussia and later of Imperial Germany. In order to avoid such failures for the future, in order to curb profit interests and prevent the harmful effects of class struggle, Knapp advocated strong state intervention and a typical German solution to the problem: a state ruled by civil servants (Beamtenstaat) (1925–7, vol. 1, p. 122).
At a later stage in his career Knapp became interested in monetary theory. His *Staatliche Theorie des Geldes* (State Theory of Money) (1905, ch. 4) was the counter-revolution against the traditional classical and neoclassical theories of money. These theories regarded it as a logical necessity for money to consist of (or to be ‘covered’ by) a commodity, generally gold, silver, or both, whose exchange value or purchasing power would then determine the exchange value or purchasing power of money. Knapp defined money independently of its material value as the creation of the legal order of the state. Consequently, he was able to explain theoretically the existence of ‘paper money’. Contrary to many of his followers, this did not lead Knapp to oppose the gold standard. He generally refrained from discussing monetary policy and tried to concentrate on the conceptual problems of monetary theory. In fact, the state theory of money did not really constitute a monetary theory, but was rather an analysis of the legal and historical aspects of money. In this sense it was supposed to be a precondition of monetary theory.
Knapp's approach aroused stormy controversies. It was extremely popular among those German economists who associated the gold standard with the international supremacy of the London money market. More importantly, both Knapp's institutional approach and his rejection of the quantity theory of money, his theoretical assessment of price increases being independent of the quantity of money and determined by ‘real’ phenomena such as wages and incomes (1905, pp. 436–48), constituted a first step towards the later theories of Keynes and his school.

Selected works

1868. *Über die Ermittlung der Sterblichkeit aus Aufzeichnungen der Bevölkerungsstatistik*. Leipzig: Hinrichs.

1874. *Theorie des Bevölkerungswechsels: Abhandlungen zur angewandten Mathematik*. Braunschweig: Vieweg.

1905. *Staatliche Theorie des Geldes*, 4th edn. Munich and Leipzig: Duncker & Humblot, 1923. Trans. as *The State Theory of Money*, London: Macmillan, 1924.

1925–7. *Ausgewählte Werke*. 3 vols. Munich: Duncker & Humblot. Vol. 1: *Einführung in einige*

Hauptgebiete der Nationalökonomie. This contains, in addition to a number of articles, the works *Landarbeiter in Knechtschaft und Freiheit* (Leipzig, 1891) and *Grundherrschaft und Rittergut* (Leipzig, 1897). Vols. 2 and 3: *Die Bauernbefreiung und der Ursprung der Landarbeiter in den älteren Teilen Preussens* (1887).

1958. (With F. Bendixen.) *Zur Staatlichen Theorie des Geldes, Ein Briefwechsel: 1905–1920*. Veröffentlichungen der List-Gesellschaft, vol. 10. Basle: Kyklos.

How to cite this article

Schefold, B. "Knapp, Georg Friedrich (1842–1926)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000030> doi:10.1057/9780230226203.0899

The New Palgrave Dictionary of Economics Online

Knies, Karl Gustav Adolf (1821 – 1898)

B. Schefold

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

causality in economics; culture and economics; Knies, K. G. A.; labour theory of value; Marx, K. H.; Ricardo, D.; Roscher, W. G. F.; Walras, L.; Weber, M.

Article

Karl Knies was born in Marburg, the son of a police employee. He studied history and political science in Marburg, and in 1846 was appointed university lecturer. In 1855, after a break in his career due to political problems, he was appointed professor in Freiburg. He transferred to Heidelberg in 1865, where he taught until 1896.

Knies was a progressive liberal with an outstanding sense of political integrity. His refusal to sign a declaration of loyalty to a reactionary state minister prevented his appointment to a professorship after the failure of the revolution of 1848, and compelled him to emigrate to Switzerland. From 1861 to 1865 he was a member of the Diet of Baden, where he actively opposed the control of the school system by the Catholic Church.

Knies, who had a profound influence on Max Weber, was one of the most important economists of the German ‘older’ historical school. He favoured the inductive method (1853, pp. 321–55), regarding facts derived from experience as more important than logical postulates. Strongly opposed to any ‘absolutism of theory’, to any theoretical assessments that claimed to be valid for all times and all people, Knies rejected the existence of general economic laws (1853, pp. 235–49), and strongly objected both to the abstract deductive reasoning of Ricardo and to the mathematical approach of Walras. Political economy has to do with the permanently changing habits and behaviour of human beings. Therefore economic analysis has to be oriented towards practical life, taking account of the peculiarities of different people and nations and different historical circumstances – Knies puts the emphasis on the historical relativism. By comparing economic relations of different countries and different historical times we may find laws of analogy, but certainly no laws of the same causal nexus. In his opposition to such laws Knies went further than other exponents of the older historical school such as Roscher.

Knies's main work on political economy (1853) was totally different from traditional textbooks. The

reader will look in vain for separate chapters on prices, wages and rents. Rather, he will find a treatise which is strongly history-oriented and focuses on the impact of history and geography on the characteristics of different people and economies, and on problems of method. Knies attacked the classical notion of self-interest as the central regulating mechanism of economic behaviour and emphasized the equal importance of the sense of membership in a community, justice and fairness (1853, pp. 147–68). He was interested in the interdependence of economics with general cultural and political life and therefore objected to an isolated study of political economy. He focused closely on the national character, and on the peculiarities and uniquenesses of different peoples, nations and races (pp. 57–70).

He also provided an analysis of money, capital, credit and interest (1873; 1879). He outlined a concept for a world currency as an international means of payment (1874). However, his analysis followed conventional methodological patterns; he did not succeed in applying the historical method to the analysis of concrete economic problems.

Knies was one of the rare bourgeois economists of 19th-century Germany who discussed Marx. He took special interest in the Marxian labour theory of value, which he opposed because of its neglect of the centrality of use value (1873, pp. 117–43).

Knies had a flexible approach towards state interventionism, which he regarded as necessary in certain cases. On the tariff question he took a stance similar to List: for an industrializing country tariffs are necessary to protect its young industry against the competition of more advanced foreign industries. Knies was also concerned with the economic and cultural implications of railroad transportation and with new systems of communication.

In an outstanding contribution to statistics Knies (1850) attempted to develop statistics as an independent discipline based on exact mathematical methods.

Selected works

1850. *Die Statistik als selbständige Wissenschaft*. Kassel: Verlag der J. Luckhardt'schen Buchhandlung.

1853. *Die politische Ökonomie vom Standpunkt der geschichtlichen Methode*. Braunschweig: Schwetschke. 2nd edn, *Die politische Ökonomie vom geschichtlichen Standpunkt*, 1883.

1862. *Zur Lehre vom volkswirtschaftlichen Güterverkehr, vom Geld, und vom Kredit*. Freiburg.

1873. *Das Geld*. Berlin: Weidmann.

1874. *Weltgeld und Weltmünze*. Berlin.

1879. *Der Kredit*. Berlin.

Bibliography

Cohn, G. 1899. Prof. K. Knies (Obituary). *Economic Journal* 9, 489.

Weber, M. 1903–6. *Roscher und Knies und die logischen Probleme der historischen Nationalökonomie*. Tübingen: Gesammelte Aufsätze zur Wissenschaftslehre, 1922.

How to cite this article

Schefold, B. "Knies, Karl Gustav Adolf (1821–1898)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000031> doi:10.1057/9780230226203.0900

The New Palgrave Dictionary of Economics Online

Knight, Frank Hyneman (1885 – 1962)

George J. Stigler

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A founder of the Chicago School, Frank Knight in his 1921 classic text *Risk, Uncertainty and Profit* defined perfect competition and distinguished risk from uncertainty in that under uncertainty the probability of events was unknowable. He criticized Pigou's proposal that increasing-cost industries should be taxed. His work on capital theory refuted Böhm-Bawerk's use of the period of production concept. Yet he conceived of economics as applying to only a small part of human activity; he criticized competitive enterprise as intrinsically unethical and unfair and debasing in practice, and feared freedom would be undermined by increasing monopoly and income inequality.

Keywords

Böhm-Bawerk, E. von; capital theory; comparative costs; curves of diminishing returns; dimensionality; diminishing returns; ethics of competition; external economies; Friedman, M.; Graham, F.; Hayek, F. von; Hutchison, T. W.; inequality of income; Kaldor, N.; Knight, F. H.; Lange, O.; Machlup, F.; Marshall, A.; monopoly; moral hazard; perfect competition; period of production; Pigou, A.C.; positivism; price theory; probability; rent; risk; Robbins, R.; Sombart, W.; Spencer, H.; Stigler, G. J.; time periods in production; uncertainty; Weber, M.; Young, A.

Article

Knight was born in McLean County, Illinois, on 7 November 1885, the first of eleven children of Winton Cyrus Knight and Julia Ann Hyneman Knight, farmers of Irish descent residing in southern Illinois. Two of Frank Knight's brothers, Melvin Moses and Bruce Winton, also became economists. Bruce once recounted an episode characteristic of his oldest brother. Under the suasion of their deeply religious parents, the children signed pledges at church to attend church the rest of their lives. Returning home, Frank (then 14 or 15) gathered the children behind the barn, built a fire, and said, ‘Burn these things because pledges and promises made under duress are not binding.’

Knight pursued his education through a series of schools and small colleges in the Midwest (see Dewey, 1986). His academic work was unfailingly marked by hard work, high intelligence and excellent grades, and one suspects that he was unfair to both himself and the poverty of his family when he once remarked that it would have been difficult to have chosen these institutions more unwisely. This preparatory period ended with two years at the University of Tennessee, and in 1913 Knight went to Cornell University, first to study philosophy and a year later (with the eager assistance of the philosophy department) he transferred to economics. His main teachers were Alvin S. Johnson and Allyn A. Young. He wrote a dissertation, ‘A Theory of Business Profit’ (1916), which displayed an astonishing depth and breadth of knowledge of the theory of value and distribution to have been acquired so quickly. With significant revision, the thesis appeared in 1921 as the classic *Risk, Uncertainty and Profit*.

Knight's subsequent academic career is easily summarized. After a year of teaching at Cornell and two (1917–19) at the University of Chicago, he went to the University of Iowa where he was an associate and then a full professor for eight years. In 1927 he returned to the University of Chicago, where he taught until 1958 and remained for the rest of his life. (Cornell in 1928 and Harvard in 1929 unsuccessfully attempted to lure him away.) The main courses he taught were in value and distribution and the history of economic thought, although occasionally he offered different topics (the present writer was one of a small number of students in a seminar on Max Weber in the mid-thirties). He was clearly the dominant intellectual influence upon economics students at Chicago in the 1930s (on his teaching, see Patinkin, 1973 and Stigler, in *Journal of Political Economy*, 1973).

He received the major honours that his profession could give him: the presidency of the American Economic Association in 1950, after he refused to be nominated in 1936 and 1937; and the Association's highest award, the Walker Medal, in 1957.

In 1911 he married a classmate at Milligan College, Minerva O. Shelburne, and they had three daughters and a son. They were divorced in 1928. In 1930 Knight married Ethel Verry, a social worker who was for many years the director of the Chicago Child Care Society, and they had two sons, Frank Bardsley, a mathematician, and Charles Alfred, a geologist. Knight died in Chicago on 15 April 1972.

The economist

Knight's dissertation, ‘A Theory of Business Profit’ was presented to Cornell University in June 1916. This was a short two years after he transferred to economics from philosophy, although evidently his interest in economics had begun earlier. (In 1913 he was already purchasing Marxist, Fabian and syndicalist pamphlets on a visit to London.) One can find much of Knight's mature thought in the thesis, which was completed when he was almost 31 years old.

The revisions of the thesis which appeared as *Risk, Uncertainty and Profit* in 1921 were substantial but not radical. Allyn Young reviewed the manuscript for the book and repeatedly asked him to ‘avoid the *appearance* of bumptiousness’ (Knight Papers, Box 54, folder 14), but the suggestions went unheeded. The three chapters in the thesis on the nature of perfect competition under stationary conditions became the four chapters of Part II of *Risk, Uncertainty and Profit* (hereafter *RUP*) with significant additions: the famous Knightian curves of diminishing returns (*RUP*, pp. 96ff.) made their first appearance, and the essence of the theory of the dominant firm was now mentioned (p. 193n). This section continued to present a clear, succinct statement of neoclassical price theory, and one can readily understand why

Lionel Robbins made it a basic text at the London School of Economics.

Knight said in this thesis that ‘The definition of perfect competition … is our principal task in this essay’ (p. 8), and it was certainly an enormously influential part of the book. Knight's conditions must have seemed extraordinarily severe to his readers: he required infinite numbers of independent traders, free and instantaneous mobility of resources and communication of knowledge, perfect knowledge and fore-knowledge, and infinite divisibility of traded goods (*RUP*, pp. 76ff.). Even today we do not normally find it useful to postulate such extreme simplicity in the economy, so that even time and space are eliminated. Some of the subtle conditions, such as that the individual ‘must be free to social wants, prejudices, preferences, or repulsions’ (p. 78), are not developed sufficiently to reveal their relevance or implications.

The treatment of risk and uncertainty quickly became Knight's ‘contribution’. Risk was characterized by the reliability of the estimate of its probability and therefore the possibility of treating it as an insurable cost. The reliability of the estimate came from either knowledge of the theoretical law it obeyed or from stable empirical regularities:

The crux of the whole question of probability, whether pure or empirical, for purposes of economic theory, is that in so far as the probability can be numerically evaluated by either method, it can be eliminated and disregarded (Thesis, p. 186).

In economic life of course the empirical probabilities are the important ones.

True uncertainty is to be ‘radically distinguished’ from calculable risks: here ‘there is *no valid basis of any kind* for classifying instances’ (*RUP*, p. 225, his italics; also p. 231). Knight believed that uncertainty cannot be explicitly and exactly defined, but one could read Bayesian elements into his discussion of probability (compare Thesis, ch. 6, with *RUP*, ch. VII).

The latter part of both the thesis and the book lack substantive structure. There is fertile, unsystematic attention to the use of combination (of which one form is specialization) to reduce uncertainty as well as risk, despite the assertion just quoted that this cannot be done for uncertainty. Considerable emphasis is placed upon intuitive knowledge in dealing with uncertainty: ‘knowledge of men's capacities to know [how to deal with uncertainty] turns out to be more accurate than direct knowledge of things’ (*RUP*, p. 298). Pure profit and pure ‘rent’ (his term for an accurately imputed income) are never found in real life: every income contains elements of both. Moral hazard makes an explicit and potentially major appearance (*RUP*, pp. 249–54) but then surprisingly vanishes from the subsequent discussion.

Several characteristics of Knight's writing were already well established in the first book:

1. (1) He looked upon received theory with a strongly sceptical eye. For example, the traditional distinctions between capital and labour are vigorously – and properly – criticized (*RUP*, pp. 126ff.). He was equally critical of both Clark's concept of the stationary economy (*RUP*, pp. 32ff.) and of Marshall's treatment of time periods in production (*RUP*, pp. 142ff.). He had already re-thought a large part of standard value theory by 1916.
2. (2) He was extremely dogmatic in his empirical generalizations – all without a trace of proof. Here are a few examples: ‘The normal rate of interest is one-half to two-thirds of the normal rate of return in fairly successful businesses’ (Thesis, p. 333). ‘There is little question that in fact speculators in land make on the whole less than the competitive rate of return on their investment’; but he has the rare qualm to add, ‘though this is difficult to prove conclusively’ (*RUP*, p. 337). ‘… Laborers show themselves ready to engage in hazardous

enterprises at their own risk for an increase in wages which is a fraction of an adequate compensation for the chances they take' (*RUP*, p. 301).

3. (3) He recurred time and time again to the same central thoughts. Once he defended the practice by quoting Herbert Spencer: 'Only by varied iteration can alien conceptions be forced on reluctant minds.' A lasting, and important, example of the tenacity of his beliefs is the view that a competitive enterprise system inherently leads to a cumulative increase in the inequality of the distribution of income. In later years at countless lunches this was challenged on both analytical and empirical grounds by Milton Friedman, each time leading Knight to make temporary concessions, only to return to his standard position by the next lunch. Knight must have felt that luncheons are doubly unfree.

A rather modest part of Knight's later writings fall within contemporary economic theory: chiefly two important articles in price theory and the series of articles on capital theory.

The first article, 'Cost of Production and Price over Long and Short Periods' (1921; reprinted in *The Ethics of Competition* [*EOC*]), offers an emendation of Marshall's analysis of time periods. Knight distinguishes a 'momentary' price which represents the supply and demand for a commodity in a speculative market: it is essentially an analysis of the prices of stocks of goods. His second, closely related period is that within which the supply of a commodity is (initially) fixed, perhaps the pricing of a given periodic crop during the crop year. Knight's third period, long run normal price, is a merging of Marshall's short and long run normal prices, a distinction which is criticized as an unnecessarily rigid classification of what is truly a continuum of time periods. The neglect of external economies will be explained shortly. There cannot be many articles in price theory that read so well after sixty-five years.

The second great article in price theory was 'Fallacies in the Interpretation of Social Cost' (1924; reprinted in *EOC*). The article contains an attack upon Pigou's celebrated error in wishing to tax increasing cost industries and upon Frank Graham's criticisms of the doctrine of comparative costs. (For a discussion of Knight's criticisms of the latter's work, see Viner, 1937, pp. 475–82.) Knight gave a lucid analysis of the role of intra-marginal transfers (rents) in achieving an efficient use of resources. It was in this article that Knight explicitly dismissed external economies: 'External economies in one business unit are internal economies in some other, within the industry' (*EOC*, p. 229). The last three words of this dismissal are inappropriate: the activities subject to increasing returns may fall in separate industries. Even if these activities subject to increasing returns are monopolized, that need not prevent the buyers of their products or services from experiencing external economies.

The major later work in theory was the series of articles on capital theory, directed against both the time preference theorists ('Professor Fisher's Interest Theory: A Case in Point', 1931) and, in a round dozen additional articles, the Austrian theory of capital. The chief of these are 'Capital, Time, and the Interest Rate' (1934), 'The Quantity of Capital and the Rate of Interest' (1936), and 'Diminishing Returns from Investment' (1944).

The first major theme of these articles is that the Böhm-Bawerkian theory of capital and interest is fatally flawed. In that theory labour joins with natural resources to produce capital goods (in the Wicksell extension of Böhm-Bawerk, sustenance for labourers and landlords). The process of producing further goods is time-consuming, and as a fundamental empirical law, the longer the production period, the larger the product. Knight denies the existence of any 'primary' factors of production which contain no capital, and equally he denies the possibility of measuring the period of production of a society or an

industry, although he would concede the possibility of measuring the period of construction or investment of a specific capital good. It is fair to claim victory for Knight over his adversaries (including Hayek, Machlup, Lange and Kaldor) on this score: the period of production concept, which had never been fertile in real applications of capital theory, has virtually vanished from the literature.

On the constructive side, Knight placed much emphasis on the correct treatment of dimensionality, with particular attention to the differences in magnitude of the stock of capital and its growth (savings) in a period such as a year. Knight believed that the long run substitution possibilities of capital for labour or for any specific form of capital such as land were immense, so diminishing marginal returns to capital either did not exist or acted extremely weakly. Accordingly, no truly long run equilibrium (such as received so much attention in classical economics) might exist:

> The peculiarity of the capital market, viewing capital service as a commodity, and the interest rate as its price, is twofold: (a) the stock of the commodity is enormously large in comparison with reasonably possible additions or subtractions in any moderate interval of time and (b) under anything like normal conditions in the real world the price is definitely above any theoretical equilibrium level (as proved by the fact that the supply does increase), and the very possibility of such a level is so problematic that it really has no interpretative value whatever (*JPE*, 1935, p. 813).

This work encountered much more criticism (see, for example, F. Lutz, *The Theory of Interest*, 1966, ch. 8, and Paul Samuelson, 1943).

Throughout his career at Chicago, Knight taught a highly idosyncratic course on the history of economics, and it is suitably represented by the famous article ‘The Ricardian Theory of Production and Distribution’ (1935). Knight's interest in intellectual history is not in the process by which it evolves but rather in the lessons it has for modern scholars; for example,

> The classical theory of wages and profits contrasts with that of rent in that it continued to be controversial, while the rent doctrine was, from the beginning, accepted as definitive. This, at least, is a good sign, for the theory sheds no light whatever on the economic principles of distribution and is an amazing tissue of inconsistency and irrelevance. These reasonings are interesting and important, not merely because they illustrate the workings of the best minds in one of the most important fields of thought and have, needless to say, some relation to facts and to real problems, but especially because they serve to warn against types of fallacy which seem to be perennially natural to minds not trained to be on guard against them (*History and Method of Economics* [*HME*], 1956, p. 75).

If Knight was quite unhistorical in treating with *Dogmengeschichte*, he was unusually widely read and perceptive in his rare appearances as an economic historian. ‘Historical and Theoretical Issues in the Problem of Modern Capitalism’ (1928) is a fascinating commentary on Werner Sombart and the related literature on capitalism, and Knight was also the translator of Max Weber's *General Economic History* (1927).

This is perhaps as appropriate a place as any to point out the unceasing intellectual curiosity Knight

displayed throughout his life. He was an inveterate and usually disappointed attendant at a vast number of lectures at the university. His wide-ranging reading never ceased. On our voyage to the first meeting of the Mt Pelerin Society in 1947, a voyage made in astonishingly powerful and persistent storms, he spent the whole time in his berth re-reading Jacob Burkhardt. It was a fundamental element of his character that his intellectual explorations were directed to the question of how ‘right’ the subject of these explorations was.

The philosopher

For most present-day economists, the primary purpose of their study is to increase our knowledge of the workings of the enterprise and other economic systems. For Knight, the primary role of economic theory is rather different: it is to contribute to the understanding of how by consensus based upon rational discussion we can fashion liberal society in which individual freedom is preserved and a satisfactory economic performance achieved. This vast social undertaking allows only a small role for the economist, and that role requires only a correct understanding of the central core of value theory. That is why the larger part of Knight's writings are outside of technical economics; indeed, that is why Knight did not return to the subjects constituting the main contributions of *Risk, Uncertainty and Profit*.
Economic theory prescribes the efficient ways of achieving given ends: this to Knight was a pathetically small part of human activity. The effects of acts often diverge grotesquely from the desires which led to them. Wants themselves are highly unstable, and it is their essential nature to change and grow. ‘The chief thing which the common-sense individual actually wants is not satisfactions for the wants he had, but more, and *better* wants’ (*EOC*, p. 22). So man is an explorer and experimenter, a seeker for unknown and perhaps unknowable truths, a creature better understood thought the study of literature than by scientific method.
It is easy, then, for Knight to castigate the competitive enterprise economy as essentially amoral, as he does in the famous essay ‘The Ethics of Competition’ (1923). Knight does not specify the nature of the ethical principles on which he bases his severe criticisms of a competitive economic system, beyond saying they are ‘the common-sense ideals of absolute ethics in modern Christendom’ (*EOC*, p. 44). That is a surprising criterion for him to employ, partly because he believed that ‘the Christian conception of goodness is the antithesis of competitive’ (*EOC*, p. 72) but also because he believed that Christian ethics had undergone great changes over time.
In the event, he bases his criticisms of those who praise the competitive system on three general grounds. The first ground is that the defence assumes perfect competition, which is certainly not even closely approximated in real life, and indeed the competitive economy instils in people crass and vulgar tastes (including placing a ‘premium on deceit and corruption’; *EOC*, p. 50). The second ground is that, viewed as a game, which is what business actually is in good part, the competitive system lacks most elements of fairness (*EOC*, p. 60). Finally, a competitive system is triply damned because competition itself is not ethically admirable (*EOC*, p. 64).
Knight's argument is subject to severe limitations. Because he avoids almost all questions of *quantity*, he often bases his argument on polar cases. Most of men's wants, for example, are stable, and at most only a small part of men's activities are devoted to the search for new wants or the exercise of curiosity. Again, he judges actual competitive enterprise by the criterion of perfect competition, yet this would be an incongruous criterion to judge other types of economic systems. (I offer some additional comments in

The Economist as Preacher, pp. 18–19.)
Yet he was even-handed in his criticisms, and when the historians criticize the competitive organization of economic life he laments their ignorance:

> Few critics of capitalism see clearly enough that the entrepreneur in his ‘control’ of production is relatively helpless as to what he shall produce, and where and when and by what instrumentalities and methods – and in particular as to what he shall pay for labour … . If one considers the range within which the manager can actually choose arbitrarily and remain in business, and averages out over a reasonable area and time period, it is evident that impersonal competition is after all overwhelmingly dominant (*HME*, p. 92).

The exploratory nature of man's goals, the infinite variety and changeability of tastes, and the mutuality of the relationship between scientist and subject in the social sciences, all led Knight to believe that positivism and behaviourism were grossly inappropriate to the study of man. (See, for example, ‘What Is “Truth” in Economics?’, 1940, reprinted in *Freedom and Reform* [*FR*], and the temperate reply of T. Hutchison, *Journal of Political Economy*, 1941, pp. 732–50.) The communication between individuals introduced a dimension wholly absent from the physical sciences, so the root fallacy ‘is to believe that social science should or can be a science in the same sense as in natural science’ (*FR*, p. 226).
On the basis of Knight's assignment of a narrow role to science in the study, let alone the control, of human behaviour, and of Knight's ethical axiom that one person should influence another only by rational discourse, he launched a series of powerful attacks on important exponents of social planning. Knight was a pungent writer and a skillful phrase maker. Instructive examples of these attacks are ‘The Newer Economics and the Control of Economic Activity’ (1932, *Journal of Political Economy*, pp. 433–76), ‘Bertrand Russell on Power’ (1939), *Ethics*, 253–85), and ‘Salvation by Science: The Gospel According to Professor Lundberg’ (1947, *HME*).
Although the main principles of economics are obvious, ‘even insultingly obvious’ (*FR*, 325), Knight despaired that they would ever be (or even could be) recognized in political life. A parable he contrived in an unpublished lecture presents this fatalistic outlook in a typical manner:

> As for telling the truth in political matters – well there is a popular story of a small boy who told the truth. Not George and the cherry tree story, but the equally famous boy who made the simple observation that an emperor had no clothes on. Scientifically, there is one fault in that story; it is unfinished. I think the author was a kindly, sensitive soul, and hadn't the heart. In the story, as a story, it is of course a merit. But in a scientific lecture if should be finished, and will only take a few sentences: That evening the people awoke to the realization that they had no emperor and the wise men were anxiously discussing what to do. You can't imagine a man as emperor after he had solemnly paraded the streets as his bare self, can you? The wise men couldn't agree, of course, and the next day there was a war. And in a year a prosperous, happy nation had been destroyed and a civilization reduced to barbarism. All because a child made an innocent remark about a plain matter of fact. And back of that, because the emperor was fool enough to let people see the human being inside an emperor's togs – which certainly everyone knew was there. Truth in society is like strychnine in the individual body, medicinal in special conditions and

minute doses; otherwise and in general, a deadly poison. …

And yet, Knight did not believe that the age of liberalism was doomed by man's incapacity to engage in and abide by rational discourse in the formation of social policy. Time and again he returned to the two forces which made liberalism intolerable: the cumulative growth of monopoly and increasing inequality of income (e.g., *EOC*, pp. 291, 310; *FR*, p. 31n). Perhaps there is no paradox here: perhaps a master of theory must become a servant of casual empiricism.

See Also

- Chicago School

Selected works

1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

1935. *The Ethics of Competition, and Other Essays*. London: Allen & Unwin.

1941. Anthropology and economics. *Journal of Political Economy* 49, 547–68. Reprinted in *Economic Anthropology: A Study in Comparative Economics*, ed. M.J. Herskovits, New York: Alfred A. Knopf, 1952.

1947. *Freedom and Reform: Essays in Economics and Social Philosophy*. New York: Harper and Brothers. Reprinted, ed. J.M. Buchanan, Indianapolis: Liberty Press, 1982.

1951. *The Economic Organisation*. New York: Augustus M. Kelley. This famous introduction to economics, first published in 1933 and based on an earlier draft, includes the English version of his articles, ‘Notes on utility and costs’, *Zeitschrift für Nationalökonomie*, 1935.

1956. *On the History and Method of Economics*. Chicago: University of Chicago Press.

1960. *Intelligence and Social Action*. Cambridge, MA: Harvard University Press.

Bibliography

American Economic Association. 1973. In memoriam: Frank H. Knight, 1885–1972. *American Economic Review* 63, 1047–8.

Buchanan, J.M. 1968. Frank H. Knight. In *International Encyclopaedia of the Social Sciences*, vol. 8, ed. D. Sills. New York: Macmillan and Free Press.

Dewey, D. 1986. Frank Knight before Cornell; some light on the dark years. In *Research in the History of Economic Thought and Methodology*, vol. 4. Greenwich, CT: JAI Press.

Howey, R.S. 1983. Frank Hyneman Knight and the history of economic thought. In *Research in the History of Economic Thought and Methodology*, vol. 1, ed. W.J. Samuels. Greenwich, CT: JAI Press. Contains a fairly full treatment of Knight's early life.

Journal of Political Economy. 1973. Vol. 81(3). ‘In Memoriam: Frank Knight (1885–1972)’, 81; Schultz, T.W., ‘Frank Knight as colleague’, 516–17; Stigler, G.J., ‘Frank Knight as teacher’, 518–20; Wick, W., ‘Frank Knight, philosopher at large’, 513–15.

Patinkin, D. 1973. Frank Knight as teacher. *American Economic Review* 63, 787–810.

Samuelson, P.A. 1943. Dynamics, statics, and the stationary state. *Review of Economics and Statistics* 25, 58–68.

Samuelson, P.A. 1966. Frank Knight's theorem in linear programming. In *The Collected Scientific Papers of Paul A. Samuelson*, Vol. 1, ed. J.E. Stiglitz. Cambridge, MA: MIT Press, 505–12.

Shils, E. 1981. Some academics, mainly in Chicago. *American Scholar*, Spring, 179–96.

Stigler, G.J. 1982. *The Economist as Preacher, and Other Essays*. Chicago: University of Chicago Press.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

How to cite this article

Stigler, George J. "Knight, Frank Hyneman (1885–1962)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000033> doi:10.1057/9780230226203.0901

The New Palgrave Dictionary of Economics Online

Kondratieff cycles

S.N. Solomou

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Kondratieff cycles are defined as regular variations in economic growth and price movements with a periodicity of 50 to 60 years. Although the balance of the evidence suggests that such regular cycles probably do not exist, this conclusion does not amount to a dismissal of the idea of long-term cyclicality. In fact, it is quite clear that various forms of long-term cycles are an observed empirical phenomenon.

Keywords

Aftalion, A.; Austrian economics; capital accumulation; Clarke, H.; De Wolff, S.; Garvy, G.; general purpose technologies; Helphand, A.; innovation; Juglar trade cycle; Kondratieff cycles; Kuznets swings; Lenoir, M.; long cycles; long waves; Marx, K.; product life cycle; Schumpeter, S.; spectral analysis; Spiethoff, A.; technical change; Tugan-Baranovsky, M.; Van Gelderen, J.; Wicksell, J.

Article

Kondratieff cycles are defined as regular variations in economic growth and price movements with a periodicity of 50 to 60 years. Over time the Kondratieff cycle has also been referred to as, *inter alia*, major cycle, long wave, long cycle, trend cycle, secular trend, secondary secular movement, secondary deviation, trend period and *mouvements de longue durée*. Kondratieff (1925) argued that the pre-1920 data he analysed imply that prices move pro-cyclically with output changes; periods of inflation were associated with rapid economic growth and periods of deflation with slow economic growth. Although Kondratieff was one of the first economists to provide a thorough statistical analysis of the long cycle, he was not the first to recognize its existence. The Russian Marxist Alexandre Helphand, writing under the pseudonym of Parvus, pointed to the existence of a long cycle as early as 1901. He drew on Marx's notion of *sturm-und-drang periode* of capital accumulation within a long-wave perspective. In 1913 the Dutch Marxist Van Gelderen, writing under the pseudonym of J. Fedder, gave an outline of ‘springtide’ and ‘ebbtide’ long cycle phases in the socialist monthly, *De Nieuwe Tijd*. A

long cycle in prices was also observed in the work of Wicksell (1898), Aftalion (1913), Lenoir (1913) and Tugan-Baranovsky (1894). In *Capital* Marx referred to ‘fluctuations extending over very long periods’ and implicitly related them to investment in buildings and other fixed capital with a low turnover period. In fact, as early as 1847 Hyde Clarke referred to a 54-year economic cycle associated with astronomical and meteorological variations.

In his first study of long cycles, Kondratieff (1922) referred exclusively to literature dealing with price movements. However, in his later work Kondratieff attempted to study long cycles as a more generalized phenomenon, observed in both nominal and real variables. Kondratieff's 1925 study is the most well known in the English-speaking world, having been first translated into English in 1935. The study is mainly an empirical exercise to test for the *existence* of long cycles. Kondratieff fitted ordinary least squares trend lines to per capita data and then used a nine-year moving average of the deviations in an attempt to eliminate the Juglar trade cycle. These filtered deviations were used to describe the historical time profile of long cycles. The statistical methodology employed is an application of Kondratieff's 1924 paper on static and dynamic equilibrium, which distinguished between reversible (wavelike movements) and non-reversible (trend) processes: ‘The wavelike fluctuations are processes of alternating disturbances of the equilibrium of the capitalistic system; they are increasing or decreasing deviations from the equilibrium levels’ (Garvy, 1943, p. 207).

Such a methodology implies that the equilibrium structure of capitalist economies remained unchanged over the period covered by his empirical work (*c*. 1780–1920). The periodization developed in Kondratieff (1925) is as follows:

1st long wave:	upswing	1780s – 1810/17
	downswing	1810/17 – 1844/51
2nd long wave:	upswing	1844/51 – 1870/75
	downswing	1870/75 – 1890/96
3rd long wave:	upswing	1890/96 – 1914/20
	downswing	1914/20 – ?

Kondratieff's early work had little to say about the generating processes for long waves; the emphasis was on describing five stylized facts:

1. 1. During the upswing phase years of prosperity are more numerous, whereas years of depression predominate during the downswing phases.
2. 2. The problems of agriculture are particularly severe during long wave downswings.
3. 3. Innovations (what he called inventions) cluster during the downswing phases, and their large-scale application during the next long upswing.
4. 4. Gold production increases during the beginning of the long upswing, and the world market for goods is generally enlarged by the assimilation of new and especially of colonial countries.
5. 5. Wars and revolutions occur during upswing phases.

All of these five aspects of long cycles are part of an endogenous process and not exogenous causal explanations. Even war is part of an endogenous long cycle; wars originate from the ‘increased tension

of economic life, in the heightened economic struggle for markets and raw materials' (Kondratieff, 1925, p. 539). The waves are supposed to reflect the development path of the capitalist world economy:

> The long cycles of the very important elements of life established above are international in nature and for the European capitalist countries the periods of these cycles are almost coincident in time. Based on the information given above, we conjecture that the same applies in the USA. (Kondratieff, 1926, translated in Makasheva, Samuels and Barnett, 1998, vol. 1, p. 38)

Kondratieff's theory of the generating process for long waves was developed in a paper read before the Economic Institute in Moscow in 1926. The theory was a long duration investment cycle, similar to Marx's ten-year investment cycle:

> … it may be asserted that the material basis for long cycles is the deterioration, replacement and extension of the main capital goods, with long production times and vast production costs. *The replacement and extension of the stock of these items is not a smooth process but a discontinuous one, which also finds* expression *in long cycles of conjuncture*. (Kondratieff, 1926, translated in Makasheva, Samuels and Barnett, 1998, vol. 1, p. 56)

To explain the discontinuities in re-investment, Kondratieff introduced Tugan-Baranovsky's (1894) theory of free loanable funds. Lumpy investments require large amounts of loanable capital and, therefore, the following preconditions are needed for the upswing: (*a*) a high propensity to save; (*b*) a large supply of loan capital at low rates of interest; (*c*) the accumulation of loan capital at the disposal of powerful entrepreneurial and financial groups; and (*d*) a low price level to induce saving. The expansion has its limits in the increased interest rate and the resulting capital shortage. Thus, Kondratieff has a monetary over-investment theory of the upper turning point, similar to that of Spiethoff (1925). The lower turning point was not explained (Garvy, 1943). Kondratieff's generating process for the long wave is similar to that of De Wolff (1924), who perceived the long wave as an echo wave, caused by the replacement of capital goods of a long lifetime, averaging 38 years. The replacement cycle was seen to be endogenous once set in motion by the Industrial Revolution of the 18th century.

Schumpeter (1939) diffused Kondratieff's ideas in the English-speaking world and significantly refined the explanatory framework. A major refinement was to view the Kondratieff cycle in a four-phase schema of prosperity, recession, depression and recovery around an equilibrium path. With respect to the price long cycle, the classification of the four phases can be interpreted using modern economic terminology. However, since Schumpeter worked within the Austrian theoretical economic framework, the pattern of real economic growth differs significantly from that postulated by Kondratieff (Solomou, 1987, pp. 6–8). In Schumpeter the economy is modelled as consisting of two sectors (producer goods and consumer goods); during the prosperity phase of the cycle output growth remains unchanged – only the structure of production changes, with the producer goods sector expanding relative to the consumer goods sector. Aggregate output expands only during the recession phase as the gestation of the new

investment generates increased productivity. Thus, over time, because of the impact of technical progress in a competitive environment, the price long wave is centred along a deflationary trend while output growth follows discrete upward steps.

The Schumpeterian long wave is an innovation-induced cycle. Schumpeter saw long cycles as resulting from the effects of lumpy, long gestation investments. Such investment was made possible by clusters of major innovations, such as railway and electricity networks. Many recent studies have developed Schumpeter's theory of long waves by linking the concept of product life cycles to the Schumpeterian idea of innovation clusters. For example, Mensch (1979) provides a modern restatement of Schumpeter's ideas. Mensch describes economic growth as being characterized by a series of intermittent innovative impulses that take the form of ‘S’-shaped growth trajectories. He postulates a *metamorphosis model*, depicting long periods of stable economic growth and relatively shorter intervals of economic turbulence. He begins with the following working hypothesis of basic innovations:

> A technological event is a technological basic innovation when the newly discovered material or newly developed technique is being put into regular production for the first time, or when an organised market for the new product is first created. (Mensch, 1979, p. 123)

Mensch argues that there is limited interest in implementing basic innovation during prosperous phases of growth; in such periods only minor improvements are introduced. In contrast, during major depression phases, when the old technologies have outlived their usefulness in sustaining profitability and economic growth, there is greater pressure for introducing basic innovations induced by low profit rates on the old technology and high potential profitability on new technology.

The empirical validity of Mensch's framework is dependent on proving the existence of *regularly recurring* clusters in basic innovations. Mensch rationalizes innovation clusters in terms of the pressures on profitability during periods of major depressions. However, without the assumption of the long wave pattern of major depressions as a macroeconomic conditioning factor, it is difficult to see why basic innovations should cluster in the interval of a regular 50-year cycle. The explanation for regular clusters has remained a major theoretical problem in the long wave literature (Garvy 1943; Kuznets, 1940; Rosenberg and Frischtak, 1983).

Do Kondratieff cycles exist?

Most economists find the empirical evidence for Kondratieff cycles to be weak. Garvy (1943) concluded that the waves identified by Kondratieff are, in part at least, statistical artefacts resulting from the techniques he employed to analyse long-run time-series data. Lewis (1978) concluded that long waves in production are not observed for the four major industrial economies (Britain, France, Germany and the United States) or for the weighted sum of these economies. Using spectral analysis (which is a statistical technique for analysing the existence of cycles of different durations), Van Ewijk (1981) found no evidence for the existence of a Kondratieff cycle in aggregate production. Beenstock (1983) examined Kondratieff's original data with the technique of spectral analysis and found no evidence of long cycles in either nominal or real variables.

Kondratieff mainly analysed price and production data from Britain and France. The focus of recent research has been on the major industrial countries of the period. In the case of Britain, most quantitative studies fail to find a pattern of Kondratieff long waves since 1850. Matthews, Feinstein and Odling-Smee (1982) recognized that British economic growth has shown long-run variations but did not observe Kondratieff cycles. Others (Van Duijn, 1983; Kleinknecht, 1987; Solomou, 1987) found a similar result. Lewis (1978) analysed industrial production trends over the period 1850–1913 and failed to find evidence of a Kondratieff cycle growth pattern.

Most studies have also failed to find evidence of Kondratieff cycles in the production trends of the US economy. Lewis (1978) focused on the period 1860–1913 and failed to find a long wave in industrial production. Solomou (1987) analysed data for the period 1870–1973 and failed to find a Kondratieff cycle in GDP growth: the most significant long-run growth variations are associated with the growth stagnation of the 1930s and the resurgence of growth in the 1940s. More recent examinations of economic growth in the USA point to one ‘big wave’ over the century after c.1870 (Gordon, 1999). Studies that have used the Kondratieff cycle to model the path of US economic growth have done so under restrictive assumptions: Bieshaar and Kleinknecht (1986) found a Kondratieff cycle in US GDP growth after 1890. A fast growth phase during 1890–1913 or 1890–1929 gives way to stagnation in the 1930s and is followed by a strong revival of economic growth after 1940. Metz (1992) also found a Kondratieff cycle over the period 1889–1979; however, this result is dependent on excluding the world war shocks and neglecting the available evidence for the 1870s and 1880s; with the war years included, the period of the long cycle is reduced significantly to the Kuznets swing periodicity. Given the importance of historical shocks to the growth process, it is difficult to justify a procedure that interpolates the war years. Similarly, neglecting the information on the 1870s and 1880s leads to a distorted picture of the US growth process during the period 1870–1913. Taking both sets of information into account results in the conventional picture that the US economy manifested a Kuznets swing growth process both during the classical gold standard period (Abramovitz, 1968) and in the period since 1913 (Hickman, 1974; Solomou, 1987). Although there are interesting long-term cyclical features, they are much longer than the Kondratieff wave period.

Similar results have been reported for France and Germany. During the period 1850–1938 the dominant long fluctuation in both economies is a Kuznets swing pattern of 20–25 year cycles (Lévy-Leboyer, 1978; Solomou, 1987; Van Duijn, 1983; Metz, 1992). Only after the Second World War is there evidence of Kondratieff-type trend periods.

Van Duijn (1983) has argued a case for the existence of Kondratieff waves in the world economy. Van Duijn found that, although the evidence for long cycles is weak, when we examine the growth path of individual countries there is stronger evidence for long cycles in the growth path of the world economy:

> Great Britain, the USA, Germany and France each have their own histories, in which the S-shaped life cycle of economic development may be more conspicuous than long wave fluctuations. The industrialized world as a whole, or even the four core countries taken together, moves forward along a long wave path. (Van Duijn, 1983, p. 154)

The production trends of the world economy provide some support to this view. The long-run growth

pattern in Maddison's (1982; 1995) world GDP series (a weighted average of GDP in the 16 major economies) suggests a pattern of long-run economic growth that is consistent with Kondratieff cycles since 1870. World exports show a similar pattern to world production trends (Lewis, 1981). It is important to emphasize, however, that this evidence cannot *prove* the existence of a propagation mechanism that generates long cycles as an endogenous economic process in the world economy. A number of shocks have played an important role in generating the phases of upswing and downswing in world economic performance. Solomou (1986) accounted for the upswing in world economic growth during 1890–1913 as the outcome of two main influences. First, countries were growing at differential rates during 1870–1913. Thus, while GDP in Britain and France grew at two per cent or less annually, the German rate averaged three per cent and the US rate four per cent. As the weight of the fast-growing economies increased over time, the world economy saw a stepping up of long-run economic growth. Second, many smaller countries started growing at a higher rate after 1890. Thus, to understand the ‘upswing’ of 1890–1913 we need to understand why countries industrialize when they do rather than why economic growth follows a long cycle. Both these effects are outcomes of one-off historical processes rather than being part of a cyclical structure in world economic growth, generated by technological developments. A similar historical perspective can be argued as explanations for the episodes of growth of the inter-war period and the post-war golden age.

Based on the evidence considered, Kondratieff cycles of a regular period probably do not exist. In reaching this conclusion it is important to stress that this is not a dismissal of the idea of long-term cyclicality. *In fact, it is quite clear that various forms of long-term cycles are an observed empirical phenomenon*. The idea that technological change is an important determinant of modern economic growth is a general truth among economists. However, determining the details of this hypothesis raises important questions for the long-cycle literature. One hypothesis is the idea that the path of major technological change is depicted as a series of general purpose technologies (GPTs). This idea has its roots in Schumpeter's theory of Kondratieff cycles – because in Schumpeter major innovation is clustered in time, the effects of clustering are at the heart of growth swings. Although the idea that there are regular 50-year Kondratieff cycles, resulting from major innovation clusters, remains a questionable empirical hypothesis, the characterization of technological change as a series of GPTs that appear episodically and have a profound effect on the growth process has received much attention in the literature on economic growth since the 1990s. Much of this work has been theoretical in nature, informing us of possible outcomes but offering no insights on actual historical economic growth. A number of macroeconomic growth hypotheses that work with fairly simple prototype models of GPTs have been accepted in the literature. For example, much of the literature argues that the diffusion of a new GPT will be correlated with a productivity slowdown in the early diffusion stage and, with long lags, will be followed by a productivity acceleration or bonus. Studies that have seen modern economic history as displaying a sequence of GPTs have used this idea as a basis of a theory for long cycles of the type that Kondratieff discussed (Freeman and Louçã, 2001). However, such models are, at present, simple thought experiments and have serious limitations when used to capture historical paths. For example, Lipsey, Bekar and Carlaw (2005, p. 384) argue that all models to date share the common problem that they deal with a complex historical economic system inappropriately, seeing economic growth as the outcome of a single GPT. In reality, at any point in historical time the growth process is the outcome of different GPTs at different stages of their life cycles, and as such the link between a new GPT and economic growth is not uniquely determined. Hence, although *episodic* long cycles are a

feasible historical outcome, we cannot assume that they will give rise to 50-year cycles. However, the concept of episodic and ‘stochastic’ long cycles may end up being a useful tool in understanding long-run economic growth and economic cycles more broadly understood.

See Also

- Kuznets swings
- technology

Bibliography

Abramovitz, M. 1968. The passing of the Kuznets cycle. *Economica* 35, 349–67.

Aftalion, A. 1913. *Les crises periodiques de surproduction*. Paris: Marcel Rivière.

Beenstock, M. 1983. *The World Economy in Transition*. London: George Allen & Unwin.

Bieshaar, H. and Kleinknecht, A. 1986. Kondratieff long waves in aggregate output? An econometric test. *Konjunkturpolitik* 30, 279–303.

Clarke, H. 1847. Physical economy – a preliminary inquiry into the physical laws governing the periods of famine and panics. *Railway Register*.

De Wolff, S. 1924. Prosperitäts- und Depressionsperioden. In *Der lebendige Marxismus: Festgabe zum 70 Geburtstage von Karl Kautsky*, ed. O. Jensen. Jena: Thuringer Verlagsanstalt.

Freeman, C. and Louçã, F. 2001. *As Time Goes By: From the Industrial Revolutions to the Information Revolution*. Oxford: Oxford University Press.

Garvy, G. 1943. Kondratieff's theory of long cycles. *Review of Economics and Statistics* 25(4), 203–19.

Gordon, R. 1999. U.S. economic growth since 1870: one big wave? *American Economic Review* 89(2), 123–8.

Hickman, B. 1974. What became of the building cycle? In *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, ed. P. David and M. Reder. New York and London: Academic Press Inc.

Kleinknecht, A. 1987. *Innovation Patterns in Crisis and Prosperity: Schumpeter's Long Cycle Reconsidered*. London: Macmillan.

Kondratieff, N. 1922. *Mirovoe khoziastvo i ego kon'iunktury vo vremia i posle viony*. [The world economy and its cycles during and after the war.] Vologda: Oblastnoe Otdelenie Gosudartsvennogo Izdatelstva.

Kondratieff, N. 1924. On the notion of economic statics, dynamics, and fluctuations. *Sotsialisticheskoe Khoziaistvo* No. 2, 349–82. An abridged English translation appears as ‘The static and dynamic view of economics’, *Quarterly Journal of Economics* 39 (1925), 575–83.

Kondratieff, N. 1925. The major economic cycles. *Voprosy kon'iunktury* vol. 1, 28–79. An abridged English translation appears as ‘The long waves in economic life’, *Review of Economic Statistics* 17(6) (1935), 105–15. A complete translation can be found in *Review: A Journal of the Fernand Braudel Center* 2(1979) (1935), 519–62.

Kondratieff, N. 1926. Long cycles of economic conjuncture. Translated in *The Works of Nikolai D Kondratiev*, vol. 1, ed. N. Makasheva, W. Samuels and V. Barnett. London: Pickering & Chatto., 1998.

Kuznets, S. 1940. Schumpeter's business cycles. *American Economic Review* 30, 257–71.

Lenoir, M. 1913. *Etudes sur la formation et le mouvement des prix*. Paris: M. Giard.

Lévy-Leboyer, M. 1978. Capital investment and economic growth in France. In *The Cambridge Economic History of Europe*, ed. P. Mathias and M. Postan. Cambridge: Cambridge University Press.

Lewis, W. 1978. *Growth and Fluctuations 1870–1913*. London: George Allen & Unwin.

Lewis, W. 1981. The rate of growth of world trade, 1830–1973. In *The World Economic Order: Past and Prospects*, ed. S. Grassman and E. Lundberg. London and Basingstoke: Macmillan.

Lipsey, R., Bekar, C. and Carlaw, K. 2005. *Economic Transformations*. Oxford: Oxford University Press.

Maddison, A. 1982. *Phases of Capitalist Development*. Oxford: Oxford University Press.

Maddison, A. 1995. *Monitoring the World Economy, 1820–1992*. Paris: OECD.

Makasheva, N., Samuels, W. and Barnett, V., eds. 1998. *The Works of Nikolai D. Kondratiev*, 4 vols. London: Pickering & Chatto.

Matthews, R., Feinstein, C. and Odling-Smee, J. 1982. *British Economic Growth, 1856–1973*. Oxford: Clarendon Press.

Mensch, G. 1979. *Stalemate in Technology: Innovations Overcome the Depression*. Cambridge, MA: Ballinger.

Metz, R. 1992. A re-examination of long waves in aggregate production series. In *New Findings in Long-Wave Research*, ed. A. Kleinknecht, E. Mandel and I. Wallerstein. London: Macmillan.

Rosenberg, N. and Frischtak, C. 1983. Long waves and economic growth: a critical appraisal. *American Economic Review* 73(2), 146–51.

Rostow, W. 1980. *Why the Poor Get Richer and the Rich Slow Down: Essays in the Marshallian Long Period*. London: Macmillan.

Schumpeter, J. 1939. *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process*. New York: McGraw-Hill.

Solomou, S. 1986. Non-balanced growth and Kondratieff waves in the world economy, 1850–1913. *Journal of Economic History* 46, 165–71.

Solomou, S. 1987. *Phases of Economic Growth 1850–1973: Kondratieff Waves and Kuznets Swings*. Cambridge: Cambridge University Press.

Spiethoff, A. 1925. Krisen. In *Handwörterbuch der Staatswissenschaften*. Jena: G. Fischer. An abridged translation appears as ‘Business cycles’, *International Economic Papers* 3: London and New York: Macmillan, 1953.

Tugan-Baranovsky, M. 1894. *Promyshlennye krizisy v sovremennoi Anglii*. St Petersburg. Translated into French as *Les Crises industrielles en Angleterre*, Paris: M.Giard & E. Brière, 1913.

Van Duijn, J. 1983. *The Long Wave in Economic Life*. London: George Allen & Unwin.

Van Ewijk, C. 1981. The long wave – a real phenomenon? *De Economist* 129, 324–72.

Van Gelderen, J. alias J. Fedder. 1913. Springtide: reflections on industrial development and price movements. *De Nieuwe Tijd* 18. Reprinted in C. Freeman, ed., *Long Wave Theory*. Cheltenham: Edward Elgar, 1996.

Wicksell, K. 1898. *Geldzins und Guterpreise*. Jena: G. Fischer. Translated into English as *Interest and Prices: A Study of the Causes Regulating the Value of Money*. New York: Augustus Kelley, 1965.

How to cite this article

Solomou, S. N. "Kondratieff cycles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000037> doi:10.1057/9780230226203.0902

The New Palgrave Dictionary of Economics Online

Kondratieff, Nikolai Dmitrievich (1892 – 1938)

S.N. Solomou

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Conjuncture Institute (Moscow); Econometric Society; Five Year Plans; Frisch, R.; Kondratieff cycles; Kondratieff, N.; Kuznets swings; Kuznets, S.; long cycles; New Economic Policy (NEP); peasant indices; Schumpeter, J.; scissors crisis; Tinbergen, J.; Tugan-Baranovsky, M.

Article

Kondratieff (also transliterated Kondrat'ev) was born in Russia on 4 March 1892. At the age of 13 he joined the Socialist Revolutionary Party. In 1915 he graduated from St Petersburg University with a First Class degree, having followed courses given by, among others, Tugan-Baranowsky. In the Soviet Union he established his reputation with his studies of the domestic economy, particularly agriculture. In October 1917, at the age of 25, he was appointed Deputy Minister for Food in the provisional (Kerensky) government, although this appointment lasted for only a few days. Kondratieff's professional career was associated with the Moscow Conjuncture Institute, which he founded and directed between 1920 and 1928.

Much of Kondratieff's work at the Conjuncture Institute consisted of obtaining accurate statistical information on the agricultural sector, including the sectoral terms of trade faced by farmers, the so-called ‘peasant indices’. These indices provided disaggregated information about the prices faced and received by the farming sector, and were calculated for the Soviet Union as a whole and for regions with different types of farming. The indices allowed Kondratieff to address the ‘scissors crisis’ by showing that the prices of farm products, relative to the cost of goods purchased by farmers, had declined during the 1920s. During 1923–5 his detailed knowledge of the agricultural sector allowed Kondratieff to prepare the first Five Year Plan for agriculture, proposing policies that did not place undue burdens on farmers. As a proponent of the New Economic Policy (NEP), he advocated a development strategy that emphasized the primacy of agriculture and the consumer goods sectors over the development of heavy industry. The abandonment of the NEP and the power struggles in the Communist Party saw Kondratieff's influence decline and in 1928 he was removed from the directorship of the Conjuncture Institute, which was closed in 1929. He was arrested in July 1930, accused of heading the ‘Working

Peasants' Party' and given an eight-year prison sentence. Kondratieff's daughter, Elena Kondratieva, has confirmed that his arrest in 1930 came after he had organized a meeting of 'dissidents' in his home (Makasheva, Samuels and Barnett, 1998, p. xiiv). At the end of this sentence he was tried again and sentenced to be executed. In fact, as early as August 1930 Stalin wrote a letter to Molotov asking that Kondratieff be executed (Barnett, 1995, p. 437).

In the West, Kondratieff is mainly known as an applied economist working on long cycles. The Kondratieff cycle became an aspect of Joseph Schumpeter's three-cycle schema, whereby the economic system was seen to display a short nine-year Juglar cycle, a medium-term Kuznets swing of 20 years and a long Kondratieff cycle of 55 years. Over the interwar period Kondratieff became a respected economist in the West, and his ideas on long cycles generated discussion from leading economists, including Schumpeter, Kuznets, Frisch and Tinbergen. Respect for his work is shown by the fact that Kondratieff became one of the founding Fellows of the Econometric Society in 1933 (Freeman and Loucã, 2001).

Kondratieff worked on long waves between 1919 and 1928. His interest in long waves may have been inspired by Tugan-Baranovsky, whom Kondratieff regarded as the 'greatest Russian economist of all time' (Jasny, 1972, p. 159). Kondratieff first outlined his theory of long waves in 1922. The manuscript was 'lost' by its Soviet publisher in 1921 but was rewritten from notes. The evidence was drawn exclusively from price trends and the conclusions were tentative: 'We consider the long cycles in the capitalistic economy only as probable' (1922, p. 255). A fuller analysis of long waves was offered in 'The Major Economic Cycles', which first appeared in 1925. In this paper Kondratieff analysed both price and production trends in Britain, France and Germany, and concluded that long waves are 'at least very probable'. This paper was purely descriptive and did not offer a theory to explain the cycle. The explanation for long cycles, in terms of reinvestment cycles of capital goods with a long lifetime, was given in a paper read before the Economic Institute in Moscow in 1926 and published in 1928.

Soon after Kondratieff was removed from the Conjuncture Institute, the official Soviet Encyclopaedia (*Bolshaya Sovetskaya Entsiklopediya*) referred to his theory on the major cycle in a single sentence: 'This theory is wrong and reactionary.' This claim to knowledge is clearly unscientific. Kondratieff was an applied economist who used historical evidence to pose questions that remain as relevant today as they were in the interwar period.

See Also

- Kondratieff cycles
- Schumpeter, Joseph Alois

Selected works

1922. *Mirovoe khoziastvo i ego kon'iunktury vo vremia i posle voiny*. [The world economy and its cycles during and after the war.] Vologda: Oblastnoe Otdelenie Gosudartsvennogo Izdatelstva.

1924. On the notion of economic statics, dynamics, and fluctuations. *Sotsialisticheskoe knoziastvo* 2,

349–82. An abridged English translation appears as ‘The static and dynamic view of economics’, *Quarterly Journal of Economics* 39 (1925), 575–83.

1925. The major economic cycles. *Voprosy kon'iunktury* I, 28–79. An abridged English translation appears as ‘The long waves in economic life’, *Review of Economics and Statistics* 17(6) (1935), 105–15. A complete translation can be found in *Review: A Journal of the Fernand Braudel Center* 2 (1979), 519–62.

1928. (With D. Oparin.) *Bol'shie tsikly kon'iunktury*. Moscow. Trans. as *The Long Wave Cycle*. New York: Richard and Snyder, 1984.

Bibliography

Barnett, V. 1995. *Kondratiev and the Dynamics of Economic Development: Long Cycles and Industrial Growth in Historical Context*. London: Macmillan.

Freeman, C. and Loucã, F. 2001. *As Time Goes By: From the Industrial Revolutions to the Information Revolution*. Oxford: Oxford University Press.

Jasny, N. 1972. *Soviet Economists of the Twenties: Names to be Remembered*. Cambridge: Cambridge University Press.

Makasheva, N., Samuels, W. and Barnett, V., eds. 1998. *The Works of Nikolai D. Kondratiev*, 4 vols. London: Pickering & Chatto.

How to cite this article

Solomou, S. N. "Kondratieff, Nikolai Dmitrievich (1892–1938)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000038> doi:10.1057/9780230226203.0903

The New Palgrave Dictionary of Economics Online

Koopmans, Tjalling Charles (1910 – 1985)

Carl F. Christ and Leonid Hurwicz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

activity analysis; additivity; American Economic Association; business cycles; Cowles Commission; Dantzig, G.; Econometric Society; econometrics; efficient allocation; estimation; exhaustible resources; felicity function; Fisher, R. A.; Frisch, R. A. K.; identification; impatience; infinite horizons; Koopmans, T. C.; limited information maximum likelihood estimator; linear programming; linear regression; maximum likelihood; optimal economic growth; renewable resources; shadow pricing; simplex method for solving linear programs; simultaneous equations models; two-stage least squares estimator; utility representation of preferences

Article

T.C. Koopmans was born in 1910 at ‘s Graveland, The Netherlands, and died in 1985 at New Haven, USA. His MA was from Utrecht in 1933 in mathematics and theoretical physics, and his Ph.D. was from Leiden in 1936 in mathematical statistics with applications to economics. His career was rather peripatetic for eight years: the Netherlands School of Economics, the League of Nations in Geneva, Princeton University, New York University, the Penn Mutual Life Insurance Company, and the Combined Shipping Adjustment Board in Washington (where he worked on the problem of optimizing the allocation of ships during the Second World War). Then in 1944 he joined the Cowles Commission for Research in Economics at the University of Chicago, where he remained until 1955 when the entire Cowles group moved to Yale. He retired from Yale in 1981. He was Research Director of the Cowles Commission at Chicago from 1948 to 1954, and of the Cowles Foundation at Yale from 1961 to 1967. He was a member of the economics faculty at Chicago from 1946 to 1955, and at Yale from 1955 to 1981.

His work earned him great honours. He was elected President of the Econometric Society in 1950, Distinguished Fellow of the American Economic Association in 1971, and its President in 1978. He received the Nobel Prize in economics jointly with Leonid Kantorovich in 1975.

He was a gentle and quiet man. His avocations were chess and music, both composing and playing the piano and the violin. He was dedicated to the search for knowledge, so much so that in the late 1960s, when the American Economic Association asked him to be its President, he declined on the ground that he had too much research he wanted to do. It was only after his friend and colleague Jacob Marschak died while president-elect that Koopmans accepted the association's second call to its presidency.

He was a theorist by nature, but a theorist interested in real problems. He made fundamental contributions not to just one but to three areas of economics: econometric methods, activity analysis (including linear programming), and the theory of optimization over time (including optimizing the use of energy and natural resources). His work was marked by precise statements of postulates and theorems, with rigorous proofs.

His doctoral dissertation (1937), now a classic, foreshadowed the style of his important later contributions to econometric methods. In it he brought the insight of Frisch and the rigour of R.A. Fisher to linear regression when all variables are subject to errors of measurement. He postulated that these errors are serially independent, and jointly normally distributed with a covariance matrix $\sigma^2 \varepsilon$. He derived maximum likelihood estimators of σ^2 and of the equation's coefficients, but showed that they depend upon ε, which in general cannot be estimated. Hence, if an incorrect ε is used in place of ε an 'error of weighting' results which does not go to zero as the sample size increases. He then showed (inter alia) that under favourable conditions this error is confined to an easily calculated region: If ε is diagonal (that is, if the errors of measurement are independent not only across time but across variables), and if the elementary regressions (each obtained by minimizing parallel to a coordinate axis) agree as to the signs of the coefficients, then the weighted regression vector based on any diagonal *e* must lie in the closed space angle that is defined by the elementary regressions and that includes the orthogonal regression vector.

Koopmans later turned to the simultaneous-equations case, assuming that each equation contains a stochastic disturbance, but that variables are measured without error. He and his co-workers made a fundamental contribution to econometric methods by solving two closely related problems that arise in such models. The identification problem concerns conditions under which simultaneous equations can be estimated at all. The estimation problem concerns how to avoid the bias inherent in least squares estimators of simultaneous equations, and obtain good estimators.

Economists had struggled with these problems for decades, with varying degrees of success or generality. Then Mann and Wald (1943) and Haavelmo (1943; 1944) in seminal pieces laid the foundations for the solution, by formulating an explicit stochastic simultaneous-equations model and considering the joint distribution of the jointly dependent variables as a function of the stochastic disturbances and the predetermined variables.

Koopmans, Rubin and Leipnik (1950c) worked out the solution to both the identification and the estimation problems in a long and technically difficult paper, which was first presented at a conference at the Cowles Commission in 1945. Much more readable papers about it are Koopmans (1945) and (1949c). Related pieces are Koopmans (1950b) and Koopmans and Hood (1953b).

An equation is said to be identified if, sampling variation aside, a unique vector of values for its parameters (up to multiplication by a nonzero constant) can be deduced from data for the variables in the model to which the equation belongs. Otherwise it is unidentified. A familiar example of unidentified equations occurs in a two-variable price-quantity model of supply and demand, where data for price and quantity allow one to estimate the intersection point of the supply and demand equations, but not the slope or intercept of either equation.

Koopmans and his colleagues considered a linear simultaneous-equations system of *G* equations in *G* jointly dependent variables, with *T* observations. In modern notation it can be written as

$$Y\Gamma + XB = E \quad \text{or} \quad ZA = E \tag{1}$$

Here *Y* and *X* are matrices of data for jointly dependent and predetermined variables, of order $T \times G$ and $T \times K$, respectively. *E* is a $T \times G$ matrix of unobservable stochastic disturbances, serially independent, with mean zero and $G \times G$ unknown covariance matrix Σ. Γ and *B* are matrices of unknown parameters, $G \times G$ and $K \times G$ respectively. *Z* is (*Y X*) and *A* is $(\Gamma' \; B')'$. Necessary and sufficient conditions were derived for the identifiability of the parameters Γ and *B*, that is, of *A*. Consider first the simple case where the only a priori information about *A* and Σ consists in the knowledge that certain elements of *A* are zero (meaning that certain variables do not appear in certain equations). Then a necessary and sufficient condition for the identification of the *i*th equation in the system (called the rank condition) is that the rank of a certain criterion matrix be equal to $G - 1$. This criterion matrix is obtained from *A* by omitting just those rows of *A* in which the element in the *i*th column is not required to be zero. Thus the criterion matrix can never have rank greater than $G - 1$, because its *i*th column is zero. Hence a necessary condition for the identification of the *i*th equation (called the order condition) is that its criterion matrix have at least $G - 1$ rows, that is, that at least $G - 1$ of the model's variables be excluded from it. In practice, if an equation satisfies the order condition, it is likely to satisfy the rank condition as well.

Koopmans and his colleagues also derived identifiability conditions for a type of a priori restriction that deals with two or more elements of *A* which may be in the same or different equations of the model.

In general, least squares estimators of simultaneous equations are biased if their expectations exist at all, and are inconsistent. Koopmans and his co-authors defined the reduced form of the model (1) as its algebraic solution for the jointly dependent variables,

$$Y = -XB\Gamma^{-1} + E\Gamma^{-1} \equiv X\Pi + V \tag{2}$$

where the reduced form's coefficients and disturbances Π and *V* are defined by the last equality in (2). Then the covariance matrix of *V* is $\Gamma^{-1'}, \Sigma\Gamma^{-1}$, denoted by Ω. They showed under their assumptions that the reduced form parameters Π and Ω are identified, and that when the disturbances are normally distributed, least squares estimators are maximum likelihood estimators and are consistent.

They also derived maximum likelihood estimators of the model's identified parameters Γ, *B* and Σ, and showed that they are consistent. This was done as follows. The likelihood function of the normally distributed reduced-form disturbances *V* is

$$(2\pi)^{-GT/2}\det{}^{-T/2}\mathbf{\Omega}\exp[-(1/2)\mathrm{tr}(V\mathbf{\Omega}^{-1}V')]. \tag{3}$$

By substituting for *V* from the reduced form (2), and subsequently substituting for Π and Ω from (2), this is transformed in two steps to

$$= (2\pi)^{-GT/2}\det{}^{-T/2}\mathbf{\Omega}\exp[-(1/2)\mathrm{tr}(Y - X\Pi)\mathbf{\Omega}^{-1}(Y' - \Pi'X')] = (2\pi)^{-GT/2}\det{}^{-T/2}(\Gamma^{-1'}, \Sigma\Gamma^{-1}) \times \exp[-(1/2)\mathrm{tr}(Y\Gamma + XB)\Sigma^{-1}(\Gamma'Y' + B'X')]. \tag{4}$$

Then the logarithm of this likelihood function is maximized with respect to the parameters *B*, Γ, and Σ, subject to the identifying restrictions. The result is the full-information maximum likelihood estimator of *B*, Γ, and Σ.

These developments created a revolution in the theory and practice of econometrics. Subsequent work by associates of Koopmans led to the limited information maximum likelihood estimator, which is much simpler than the computationally demanding full-information estimator. Later work by Theil, Zellner, and many others led to the still simpler two-stage least squares estimator and related estimators such as three-stage least squares. Koopmans can fairly be said to be the father of simultaneous-equations econometric methods, though it is clear that there were grandfathers and great-grandfathers too.

In addition to econometrics, Koopmans made outstanding contributions in several areas of economic analysis, both theoretical and applied. The extent to which the two aspects complement each other is particularly striking in his work in activity analysis and linear programming. On the applied side, Koopmans's interest in this area seems to stem from an investigation into tanker freight rates and tankship building (1939). The book on this subject, apparently his earliest published

work in economic (as distinct from econometric) analysis, has the subtitle *An Analysis of Cyclical Fluctuations*, and is indicative of his interest in the major macroeconomic issue of the 1930s, the business cycle. Other contributions dealing with business cycles appear in the early 1940s (1940; 1941; 1949b), although mainly concerned with econometric orientation. Subsequently Koopmans's macroeconomic interests shifted to the prevention of threatened post-Second World War inflation, a concern prevalent in the early 1940s (1942a; 1943). But his responsibilities with the Combined Shipping Adjustment Board pushed him in the direction of efficient resource allocation problems, applied to Allied freight shipping during the Second World War period. A memorandum (1942b), published for the first time in the *Scientific Papers of T.C. Koopmans* (1970), lays the foundation for what subsequently would be called activity analysis and linear programming. More specifically, the 1942 memo and more elaborate treatment (1949a) presented in 1947 at the International Statistical Conference deal with efficient utilization of transportation systems, a problem treated again (jointly with Reiter) in a more general setting in a paper presented at a memorable conference held in Chicago at the Cowles Commission in 1949 (1951). Koopmans' 1942 work on the transportation problem was done without awareness of the earlier (1941) study by Hitchcock and of the contributions due to Kantorovich (1939; 1942), von Neumann (1935), and Dantzig (1951a; 1951b; 1951c). The product of their insights and analyses, named *activity analysis*, is a model of production involving not only commodities (inputs and outputs), but also explicit recognition of the processes used in the course of production. With each *process* is associated a non-negative (scalar) variable, called the *level of activity* representing that process. Let x_k denote the level of kth activity (from among K possible ones), and let y_n be the net output of the nth commodity (from among N commodities present), with a negative value corresponding to an input. Technological information defines relations specifying the net outputs as a function of activity levels, say

$$y_{nk} = f_{nk}(x_k),\ n = 1, \ldots, N; \quad k = 1, \ldots, K. \tag{5}$$

Thus for a given level x_k of the kth activity, the function f_{nk} specifies the amount of y_{nk} of resulting net output (positive, negative, or zero) of the nth commodity. The equation system (5) allows for non-linear production relations, but most early work postulates linearity, in the sense that the ratio of net output to activity level is independent of the level of that activity. So (5) is specialized to

$$y_{nk} = a_{nk}x_k,\ n = 1, \ldots, N; \quad k = 1, \ldots, K \tag{6}$$

where a_{nk} is a constant independent of x_k. Furthermore, as was already implicit in our notation, when several processes are carried on simultaneously, it is assumed that they do not interfere with each other. Hence the aggregate output, say y_n, of the nth commodity is formed additively from the amounts contributed by the simultaneous activities, and so

$$y_n = \sum_{k=1}^{K} y_{nk} = \sum_{k=1}^{K} a_{nk}x_k,\ n = 1, \ldots, N\ . \tag{7}$$

In this setting a number of problems have been studied. *Optimization* calls for maximization or minimization of a function, say $g(x_1, \ldots, x_k)$ of the activity levels. In linear programming models this function (like the technological relations (6)) is linear so that

$$g(x_1, \ldots, x_k) \equiv \sum_{k=1}^{K} c_k x_k . \tag{8}$$

Formally, then, we have the problem of maximizing

$$\sum_{k=1}^{K} c_k x_k$$

with respect to the non-negative variables $x_k \geq 0$, subject to the technological relations (7), as well the resource constraints

$$y_n \geq -\eta_n, \quad n = 1, \ldots, N, \tag{9}$$

where η_n is the amount of the *n*-th good initially available. The most important technique for solving such a linear programming problem, known as the simplex method, is due to Dantzig. Dantzig's pioneering contributions in formulating the model itself were repeatedly stressed by Koopmans, who felt that Dantzig should have shared in the Nobel prize. This recognition in no way detracts from Koopmans's own role in formulating and developing the activity analysis model and analysing its properties. Most importantly, he built a bridge between activity analysis and the conceptual framework of classical economics. This involved distinguishing between primary, intermediate, and final commodities, the analysis of the efficiency concept, and the role of prices and profits. In the latter areas, again, Koopmans was careful to recognize the relationship of his analysis to the earlier contributions, especially those of Lange (1938) and Lerner (1944). In particular, Koopmans dealt with efficiency in production by defining it in terms of the vectorial ordering in the commodity space as follows. A commodity vector $y = (y_1, \ldots, y_N)$ is called *possible* if it satisfies the technological constraints, without taking into account the limitations due to resource availability. Then a possible point *y* is called *efficient* if there is no other possible point y' such that y' vectorially dominates *y*, that is, such that $y'_n \geq y_n$ for all $n = 1$ and $y'_r \geq y_r$ for some *r*.

An important contribution of Koopmans is the characterization of efficient points with the help of ‘accounting’ (also called ‘shadow’) prices, the condition being that no activity permits a positive profit and that the profit on activities actually carried out to be zero. The relationship of these conditions to those for

competitive profit maximization under constant returns to scale is evident. Indeed, Koopmans formulated a resource allocation ‘game’ whose equilibria would be efficient when ‘players’ follow specified behaviour rules, with activity managers expanding profitable activities, avoiding activities yielding losses, and keeping constant levels for activities yielding zero profit. Other participants for this ‘game’ are ‘commodity custodians’, whose function is to adjust prices according to the difference between demand and supply, and a ‘helmsman’, choosing prices of final goods according to specified objectives (tastes). As Koopmans pointed out, only static properties of the game follow from the rules.

Of particular significance is Koopmans's emphasis on the informational decentralization of this ‘game’: a manager only needs to know the technology of his own process; a custodian needs only to know the availability and demand for the commodity he is in charge of. Koopmans stressed the applicability of the model both in a competitive economy (where the role of the ‘helmsman’ would be played by the consumers’ competitive bidding) and in planned economies (where prices are an accounting rather than market phenomenon).

Koopmans’ contribution in the area of activity analysis are of significance not only for their content, but also for their form and style. Their rigour and clarity became a standard, or at least an ideal, for later mathematical economics with emphasis on explicit definitions, postulates, and theorem formulations. His meticulous attention to (and acknowledgement of the work of) predecessors as well as generosity in evaluation of the contributions of others invite emulation. (See, for instance, the Introduction to the *Activity Analysis* volume and his two notes on Kantorovich's work, 1960b; 1962.)

In a brief article, it is impossible to do justice to Koopmans’ own accomplishments. His expository talents are particularly striking in the first of his *Three Essays* [1957], which is a model for exposition of classical welfare economics, making particularly clear which propositions depend on which assumptions – for example, absence of convexity postulates in proving the Pareto optimality of competitive equilibria.

Perhaps the most important of Koopmans's theoretical contributions are those dealing with problems involving infinite horizon economies. A number of papers deal with optimal economic growth (for example, 1965a; 1965b, jointly with Beals; and 1973). But of particular interest are the papers (1960a, and 1964, jointly with Diamond and Williamson) concerning preferences and their representation by numerical (real-valued) utility functions over infinite horizons. Koopmans formalized the concept of impatience (introduced by Böhm-Bawerk into the theory of the rate of interest) and showed, surprisingly, that impatience is a necessary logical consequence of a set of postulates concerning utilities over infinite time horizons. Among the postulates are those of continuity and stationarity (that is, independence of calendar time).

An additional postulate is required to imply the discounted form of the utility function,

$$U(x_1, x_2, \ldots) = \sum_{t=1}^{\infty} a^{t-1} u(x_1)$$

where $0 < a < 1$ and $U(\cdot)$ is the utility function for the infinite programme $(x_1, x_2, \ldots)$. Here x_t denotes the choice x_t at time t, and $u(x_t)$ is the instantaneous ‘felicity’ experienced at time t. For instance, x_t may be the consumption vector, $x_t = (x_{t1}, \ldots, x_{tl})$, in a l-dimensional commodity space. More generally, all the x_t are assumed to be drawn from the fixed choice space X, a connected subset of n-dimensional Euclidean space.

Koopmans returned to the problem of utility representation of preferences over infinite programmes in two papers (1972a; 1972b), differing from the earlier work by its formulation of the underlying postulates in terms of the preference relations rather than of a utility representation, whose existence was assumed previously in the hypotheses. In the later work, it is the preferences that are assumed to be continuous and stationary as well as to satisfy a condition of independence over time. Under these postulates the utility function of an infinite programme is shown to have certain additivity properties. A stronger conclusion is obtained, for the space of all programmes that are ‘bounded in utility.’ [A programme $x = (x_1, \ldots)$, x_t in the choice space X, is said to be bounded in utility if there exist vectors x^* and x^{**} in X such that $x^{**} \geq x_t \geq x^*$ for all $t = 1, 2, \ldots$, where the symbol $\geq$ represents the (weak, that is, reflexive) preference relation. (We may note that ‘bounded in preference’ would have been a better term, since numerical utility is not involved in the definition.)] On the

space of all programmes bounded in utility preferences can again be represented by the discounted from

$$U(x_1, x_2, \ldots) \equiv \sum_{t=1}^{\infty} a^{t-1} u(x_t), \quad 0 < a < 1.$$

Among other contributions involving the infinite time horizon, we shall only mention Koopmans's work on exhaustible resources, in particular the problem, of such interest in the 1970s and early 1980s, of transition from exhaustible to renewable resources. Closely related to this area of interest was Koopmans's work on the modelling of alternative energy futures reflected in (1980) and the guidance he provided as chairman of the Modeling Resources Group of the Committee on Nuclear and Alternative Energy Systems (CONAES) of the National Academy of Sciences (1975–8).

The latter study exploited Koopmans' lifelong interest in the relationship between economics and physical sciences. His first published papers were in physics, and his presidential (American Economic Association) address in 1978 was entitled 'Economics among the Sciences' (1979).

This talk was in part based on observations made in the course of the energy modelling study and dealt with the difficulties in communication between physical scientists, engineers, and economists, illustrated by several examples, including the problems of discounting future costs and benefits. The address also pointed to the university procedures for academic appointment and promotion as barriers to interdisciplinary contacts.

While concerned about the relationship to physical sciences, Koopmans did not neglect the ethical issues implicit in the various criteria of optimal growth, especially the problems of balancing the consumption levels of successive generations (1967a; 1967b). Technically, the problem arises because an 'optimal' solution may fail to exist in an infinite horizon setting. When future enjoyments are discounted but the discount factor falls below a critical value, it turns out that a further postponement of some future consumption raises the utility of the overall programme. Since, with an infinite horizon, a postponement is always conceivable, no programme is 'best'. Koopmans' conclusion was that 'one cannot adopt ethical principles without regard to the anticipated population growth and to the anticipated technological possibilities', and that 'ethical principles … need mathematical screening to determine whether in given circumstances they are capable of implementation.'

In some models, given the discount rate, the existence of an optimum depends on the shape of the instantaneous 'felicity' function $u(\cdot)$, and so does the shape of the optimal path when existence conditions are satisfied. Koopmans regards as debatable whether the choice of the felicity function $u(\cdot)$ is an empirical or ethical question. He points out the paucity of empirical evidence concerning the asymptotic elasticity of marginal felicity at high consumption levels which is critical for the existence of an optimum. He then expresses concurrence with a remark due to Malinvaud that ethical judgments may be easier to base on comparison of optimal paths generated by alternative felicity assumptions than on 'direct and aprioristic' comparisons of the felicity functions themselves.

The depth and breadth of his scientific contributions as well as his influence on others (including these writers) amply justify the judgment of the Nobel Committee, as well as Scarf's (1985) characterization of Koopmans as the 'leader of a scientific revolution', a revolution ranging over econometrics and economic analysis.

Selected works

1937. *Linear Regression Analysis of Economic Time Series*. Haarlem: De Erven F. Bohn.

1939. *Tanker Freight Rates and Tankship Building*. Publication No. 27 of the Netherlands Economic Institute, Haarlem: De Erven Bohn.

1940. The degree of damping in business cycles. *Econometrica* 9(January), 79–89.

1941. The logic of economic business-cycle research. *Journal of Political Economy* 49(2), 157–81.

1942a. The dynamics of inflation. *Review of Economics and Statistics* 24(1), 53–65. Reply, *Review of Economics and Statistics* 24(4), 190.

1942b. Exchange ratios between cargoes on various routes (nonrefrigerating dry cargoes). Memorandum for the Combined Shipping Adjustment Board, Washington, DC, 1942, 1–12. Reprinted in Koopmans (1970).

1943. Durable consumers' goods and the prevention of post war inflation. *American Economic Review* 33, 882–8.

1945. Statistical estimation of simultaneous economic relations. *Journal of the American Statistical Association* 40, 448–66. Reprinted in Koopmans (1970), 93–111.

1949a. Optimum utilization of the transportation system. *Proceedings of the International Statistical Conferences*, Washington, DC, 1947, Vol. 5. Reprinted in Supplement to *Econometrica* 17 (July 1949), 136–46. Errata, *Econometrica* 19(2) (1951), 228.

1949b. The economic approach to business fluctuations. *American Economic Review, Papers and Proceedings* 39(3), 64–72.

1949c. Identification problems in economic model construction. *Econometrica* 17, 125–44. Reprinted with minor revisions and an added example in Hood and Koopmans (1953) and in Koopmans (1970).

1950a. (ed.) *Statistical Inference in Dynamic Economic Models*. Cowles Commission Monograph No. 10. New York: Wiley.

1950b. When is an equation system complete for statistical purposes? In Koopmans (1950a).

1950c. (With H. Rubin and R.B. Leipnik.) Measuring the equation systems of dynamic economics. In Koopmans (1950a).

1951. (With S. Reiter.) A model of transportation. In *Activity Analysis of Production and Allocation, Proceedings of a Conference*, ed. T.C. Koopmans. Cowles Commission Monograph No. 13, New York: John Wiley, 222–59.

1953a. (With W.C. Hood, ed.) *Studies in Econometric Method*. Cowles Commission Monograph No. 14. New York: Wiley.

1953b. (With W.C. Hood.) The estimation of simultaneous linear economic relationships. In Hood and Koopmans (1953a).

1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

1960a. Stationary ordinal utility and impatience. *Econometrica* 28, 287–309. Reprinted in *Readings in Mathematical Economics*, vol. 2, ed. P. Newman. Baltimore: Johns Hopkins Press, 1968.

1960b. A note about Kantorovich's papers, 'Mathematical Methods of Organizing the Planning Production'. *Management Science* 6, 363–5.

1962. On the evaluation of Kantorovich's work of 1939. *Management Science* 8, 264–5.

1964. (With P.A. Diamond and R.E. Williamson.) Stationary utility and time perspective. *Econometrica* 32(1–2), 82–100.

1965a. On the concept of optimal economic growth. Paper [4] of *[Study Week on the] Econometric Approach to Development Planning*. Pontificae Academiae Scientiarum Scripta Varia 28. Amsterdam: North-Holland; Chicago: Rand McNally.

1965b. (With R. Beals.) Maximizing stationary utility in a constant technology. *SIAM Journal of Applied Mathematics* 17, 1001–15.

1967a. Objectives, constraints, and outcomes in optimal growth models. *Econometrica* 35(1), 1–15. Reprinted in *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*, ed. K.A. Fox, G.V.L. Narasimhan and J.K. Sengupta. New York: Springer-Verlag, 1969.

1967b. Intertemporal distribution and ‘optimal’ aggregate economic growth. In W. Fellner et al., *Ten Economic Studies in the Tradition of Irving Fisher*. New York: John Wiley.

1970. *Scientific Papers of Tjalling C. Koopmans*. Berlin: Springer. Contains a bibliography through September 1969.

1972a. Representation of preference orderings with independent components of consumption. In *Decision and Organization: A Volume in Honor of Jacob Marschak*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

1972b. Representation of preference orderings over time. In *Decision and Organization: A Volume in Honor of Jacob Marschak*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

1973. Some observations on ‘optimal’ economic growth and exhaustible resources. In *Economic Structure and Development: Essays in Honour of Jan Tinbergen*, ed. H.C. Bos, H. Linnemann and P. de Wolff. Amsterdam: North-Holland.

1979. Economics among the sciences. Presidential Address to the American Economic Association. *American Economic Review* 69, 1–13.

1980. The transition from exhaustible to renewable or inexhaustible resources. In *Economic Growth and Resources. Vol. 3: Natural Resources*, ed. C. Bliss and M. Boserup. London: Macmillan.

1985. *Scientific Papers of Tjalling C. Koopmans*, vol. 2. Cambridge, MA: MIT Press.

Bibliography

Dantzig, G.B. 1951a. The programming of interdependent activities: mathematical model. In *Activity Analysis of Production and Allocation, Proceedings of a Conference*, ed. T.C. Koopmans. New York: John Wiley.

Dantzig, G.B. 1951b. Maximization of a linear function of variables subject to linear inequalities. In *Activity Analysis of Production and Allocation, Proceedings of a Conference*, ed. T.C. Koopmans. New York: John Wiley.

Dantzig, G.B. 1951c. Application of the simplex method to a transportation problem. In *Activity Analysis of Production and Allocation, Proceedings of a Conference*, ed. T.C. Koopmans. New York: John Wiley.

Haavelmo, T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1–12.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement), 1–115.

Hildreth, C. 1976. *The Cowles Commission in Chicago, 1939–1955*. Lecture Notes in Economics and Mathematical Systems No. 271. New York: Springer-Verlag.

Hitchcock, F.L. 1941. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics, Massachusetts Institute of Technology* 20, 224–30.

Jungenfelt, K.G. 1976. Tjalling Koopmans' contribution to economics. II: Koopmans and the recent development of growth theory. *Scandinavian Journal of Economics* 78(1), 94–102.

Kantorovich, L.V. 1939. Mathematical methods of organizing and planning production. *Management Science* 64 (1960), 363–422. (Originally published in Leningrad, 1939.)

Kantorovich, L.V. 1942. On the translocation of masses. *Doklady Akademii Nauk SSSR* 37, 199–201. Trans. in *Management Science* 5(1) (1958), 1–4.

Lange, O. 1938. On the economic theory of socialism. In F. Taylor and O. Lange, *On the Economic Theory of Socialism*, ed. B.E. Lippincott. Minneapolis: University of Minnesota Press.

Lerner, A.P. 1944. *The Economics of Control*. New York: Macmillan.

Malinvaud, E. 1972. The scientific papers of Tjalling C. Koopmans: a review article. *Journal of Economic Literature* 10, 798–802.

Mann, H.B. and Wald, A. 1943. On the statistical treatment of linear stochastic difference equations. *Econometrica* 13, 245–59.

Neumann, J. von. 1935. A model of general equilibrium. In *Ergebnisse eines Mathematischen Kolloquiums*, ed. Karl Menger, 8 (1935–6). Leipzig: Verlag, 1937. Trans. in *Review of Economic Studies* 13 (1945), 1–9.

Royal Academy of Sciences, Stockholm. 1976. The official announcement of the Royal Academy of Sciences. *Scandinavian Journal of Economics* 78(1), 59–60.

Scarf, H.E. 1985. Foreword to *Scientific Papers of Tjalling C. Koopmans*, vol. 1. Cambridge, MA: MIT Press.

Werin, L. 1976. Tjalling Koopmans' contribution to economics. I: Activity analysis, methodology and econometrics. *Scandinavian Journal of Economics* 78(1), 81–93.

How to cite this article

Christ, Carl F. and Leonid Hurwicz. "Koopmans, Tjalling Charles (1910–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000040> doi:10.1057/9780230226203.0904

The New Palgrave Dictionary of Economics Online

Korea, economics in

Young Back Choi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The study of economics in Korea is a modern development, starting around the turn of the 20th century, influenced by the West, at first through Japan and since the end of the Second World War mainly through the United States. Over time, it has become a clone of economics in the USA, with an overriding concern for mathematization and econometrics. Even so, the Korean term for economics, *kyung-je-hak*, which derives from a classical Chinese expression for good governance, reveals the prevailing conflation of science and governing techniques.

Keywords

classical economics; German Historical School; Japan, economics in; Korea, economics in; Korean Economic Association; marginal economics; Marxist economics; neoclassical synthesis; protectionism; United States, economics in (after 1945)

Article

The 19th century

Before the turn of the 20th century, there was not much economic analysis to speak of in Korea. Of course, as a country settled continuously over 2,000 years, Koreans had a variety of administrative techniques concerning tax collection, budgeting, coinage, government monopolies and so forth. But there was hardly an attempt to organize observations on economic matters along general principles. The situation in Korea was not very different from that in most pre-industrial economies.
In the 18th century, in the aftermath of a series of foreign invasions from Japan and Manchu, a group of reform-minded thinkers, Shil-Hak-Pa, wrote pamphlets proposing measures to ameliorate economic devastation and social instability. But there was no economic analysis beyond the simplest form of common sense; and the school was subsequently suppressed as subversive. At the end of the 19th

century, when Korea was encircled by imperialist powers and its independent status became precarious, a few Korean students were sent to Japan to study the secrets of a ‘prosperous nation and strong army’. Some thereby encountered classical economics and viewed free competition as a means of creating a prosperous nation. The attempts to learn from the West and reform the country, however, came too late; Korea became a Japanese colony in 1910.

The Japanese colonial period

During this period, Korean economists, mostly trained in Japan, were few in number and predominantly Marxists, doing economic history from a Marxian perspective. The reasons had to do with contemporary economics in Japan and what many colonial people took Marxism to be at the time.

Earlier, in the mid-19th century, classical economics was studied in Japan. From the 1880s, however, the German Historical School, critical of classical economics and providing a rationale for protectionism and nationalism, became the vogue among Japanese economists. (Note that economists in the United States were also heavily influenced by the Historical School, giving rise ultimately to American Institutionalism.) Another import from Germany was Marxism. By the 1920s, Marxism and the Historical School had become the dominant traditions in Japan. Protectionism and Keynesianism were later introduced, but they had a limited influence on the Japanese during this time.

Broadly speaking, Korean students had to choose between the two dominant traditions in Japanese economics. As the Historical School was seen as providing a rationale for Japanese fascism, most Korean students of economics were attracted to Marxism, which represented anti-imperialism and anti-capitalism. As the Japanese did not appoint Koreans to faculty posts in Japanese universities (not even those located in Korea), Korean economists taught at a handful of private colleges founded by Koreans. However, from the early 1940s the Japanese repression of Korean nationalism and communism meant that most Korean academics were imprisoned and students sent either to the front line or to armaments factories. The surrender of Japan at the end of the Second World War and the liberation of Korea abruptly ended this state of affairs.

Post-Second World War to the Korean War

In the liberated Korea Marxist economists were restored and became dominant, as in Japan. Korea was partitioned and occupied by the two victors of the Second World War, the northern half by the USSR and the southern half by the United States. Ideological conflicts ensued as the occupiers delegated power to the locals and Korean Marxists abandoned all academic pretensions. During the subsequent internecine Korean War (1950–3), Marxists and their sympathizers were completely eradicated from South Korea; they all moved to Communist North Korea, where they were subsequently purged. What was left in South Korea at the end of the war, therefore, was only a handful of non-Marxist Korean economists who had been trained in Japan.

Post-Korean War to the present

Since the post-war era economics in South Korea has by and large become Americanized. Although a

worldwide phenomenon, this is particularly pronounced in Korea, since the country became completely dependent on US military and economic aid after the destructive war.

In the late 1950s and the early 1960s the US government and UN organizations sent advisers, many of whom were New Dealers and/or Keynesians. During this period some Korean economists formerly trained in Japan visited American universities, or received additional training there. The first fruit of the initial contacts with the United States was a flurry of translations of economic literature, ranging from classical, neoclassical and Keynesian economics to development economics.

In the 1970s there was a noticeable slowdown in translations and the concomitant publication of a number of popular economics textbooks in Korean, which closely resembled the then popular American economics textbooks in the tradition of the neoclassical–Keynesian synthesis. From the late 1980s, popular American textbooks began to be used, untranslated, in elite Korean universities. Since the 1980s translations of scholarly books in economics have become popular again, but not to the same degree as in the late 1950s. This in part reflects the rapid increase in the number of American-trained economists who can directly access economic literature in English. The paucity of translations also reflects the shallow roots of economics in Korea.

From the late 1960s, Koreans with doctorates in economics from American universities began to return to Korea in significant numbers, increasing from the 1970s and peaking in 1990, when the figure reached over 60. In 1993, the Korean Economic Association had over 1,800 members with a Ph.D. in economics. Of these, 55 per cent had a Ph.D. from a foreign university and 43 per cent from an American university. The proportion alone understates the impact of Americanization, as the US-trained economists have come to occupy a disproportionately large share of key positions in academia, research institutes and central government in Seoul. For example, in 1993 at the top three Korean universities, the proportion of American-trained economists in the economics faculty reached over 70 per cent. In one of them, all but one had an economics Ph.D. from the United States.

The process of Americanization was greatly aided by the job market for economists in Korea, which can best be characterized by institutional inbreeding and credentialism. The former is the practice of hiring graduates of one's own department, a legacy of the Japanese colonial period, and the latter the practice of hiring based on the quality of credentials. Initially, the custom of obtaining an academic post in Korea was first to get a BA (or even an MA) at an economics department in an elite Korean university, and then get an American Ph.D.

Competition for credentials has become more intense. In the 1960s, an American Ph.D. was sufficient to gain an academic post. As the number of American Ph.D.s increased over time, the prestige of the school awarding the degree became significant. As more and more Koreans get their Ph.D.s from elite American universities, some teaching experience in the United States and even publications in English-language journals have become crucial for employment prospects. Economics in Korea, through the process of competition for better credentials, has come to reflect the prevalent practices of the economics department at elite American universities, with the overriding concern for publication in top journals, which necessarily implies emphasis on mathematical economics, model building, and econometrics. From the late 1980s, increasing competition for credentials has extended the period of scientific endeavours well beyond graduate school, and a few Korean economists have managed to gain mobility across the national border.

Yet there is no great man in economics *in* Korea. Of course, there have been economists who have been celebrated on account of their contributions in the popular media, or economists who have authored

popular economic textbooks, or commanded much respect on account of having taught many able students. But no economist in Korea has attained eminence in economic science as one might have expected from the investment of so many resources. The reason is only in part that economics in Korea is essentially a post-Korean War development, or that many Korean economists decided not to return to Korea after their graduate training. The main reason is the pragmatic orientation of the majority of Korean economists (which has only very recently begun to shift).

Economics in Korea gained its prestige under the economic dirigisme of President Park (1961–79). The administration and legitimization of planning development programmes required the services of economists, which many were eager to supply. For aspiring economic advisors or policymakers, the primary concern is political expedience; during the period 1961–79, it was as if Korean economists entertained scientific concerns only while their credentials were being established, that is, during their graduate training in the United States. Afterwards, even as they taught students what they had learned in the USA, the majority of them ended up taking on diverse extra-curricular activities, including extra teaching, consulting for government bureaux and cultivating political connections. This pragmatic approach to economics has produced an army of economists who are very competent in importing techniques, but quickly cease to be members of the scientific community. The pragmatism of Korean economists is reflected in the absence of doctrinal disputes or distinctive schools of thought. Academic fashions have come and gone, largely reflecting, if with a time lag, debates that took place in the United States during students' graduate training.

One exception has been doctrinaire Marxists whose number has increased in South Korea since the late 1980s, precisely when Marxism was being discarded elsewhere. This surprising development against the worldwide trend has been an outgrowth of reactions to the authoritarian rules of Presidents Park and Chun, in which anti-authoritarianism, pro-democracy, socialism, Marxism and nationalism had been all conflated. There is little debate between Marxists and non-Marxists, however. Marxists are more than willing to be engaged in doctrinal debate, though their concern largely focuses on who has a more faithful reading of the canon. Non-Marxist economists, the majority of Korean economists who are either US-trained or trained by other US-trained Korean economists, are pragmatists in their teaching and advice, generally seconding the popular preference for a welfare state of the European variety.

Outwardly, economics in Korea has become fully internationalized. Over 50 per cent of Korean economists with a Ph.D. have been trained overseas, the overwhelming majority in the United States. Most are competent in the techniques of modern economics, familiar with the relevant literature, an increasing number of them are publishing in international journals, and a few have even gained international job mobility. Yet if economics in Korea is to progress beyond the stage of competently importing the latest academic trends and techniques, more Korean economists will have to become less pragmatic and begin to examine basic questions.

See Also

- Japan, economics in

Bibliography

Bronfenbrenner, M. 1956. Economic thought and its application and methodology in the east: the state of Japanese economics. *American Economic Review: Papers and Proceedings* 46, 389–98.

Choi, Y.B. 1994. Industrial policy as the engine of economic growth in South Korea: myth and reality. In *The Collapse of Development Planning*, ed. P. Boettke. New York: New York University Press.

Choi, Y.B. 1997. The Americanization of economics in Korea. In *The Post-1945 Internationalization of Economics*, ed. A.W. Coats. Durham, NC: Duke University Press.

Coats, A.W. 1960. The first two decades of the American economic association. *American Economic Review* 50, 555–74.

Colander, D. and Klamer, A. 1987. Making of an economist. *Journal of Economic Perspectives* 1(2), 95–111.

Kang, M.K. 1983. Critical reflections on Korean economics [in Korean]. In *Introspections on Korean Economics*, ed. J.H. Kim. Seoul: Korean Economic Research Institute.

Shin, T.H. 1983. Annals of Korean economics for 50 years [in Korean]. In *Introspections on Korean Economics*, ed. J.H. Kim. Seoul: Korean Economic Research Institute.

How to cite this article

Choi, Young Back. "Korea, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000072> doi:10.1057/9780230226203.0905

The New Palgrave Dictionary of Economics Online

Kravis, Irving B. (1916 – 1992)

Alan W. Heston
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bureau of Labor Statistics (BLS) (USA); country-product dummy (CPD) method; exchange rates; income distribution; International Comparison Programme (ICP) (UN); international price competitiveness; international trade: and economic growth; Kravis, I.; Kuznets, S.; Laspeyeres index; law of one price; Maddison, A.; Organisation for European Economic Co-operation (OECD); Paasche index; Penn World Table; purchasing power parity; World Bank

Article

Irving Kravis is best known for his pioneering empirical estimates of country purchasing power parities (PPPs) and real products based on detailed price and expenditure comparisons. This research began in the early 1950s at the Organisation for European Economic Co-operation (OEEC, now OECD) in collaboration with Milton Gilbert. It was continued on a world scale when he jointly directed the United Nations International Comparison Project from 1968 to 1982. However, these contributions were just the most notable in a career that included yeoman service to the University of Pennsylvania: in building up the Department of Economics, in serving as Associate Dean of the Wharton School and chair of the University Senate, and in active participation in many important committees of the School of Arts and Sciences and the university.

With the exception of the Second World War, Kravis's career was all at the University of Pennsylvania, where he studied as both undergraduate and postgraduate and to which he returned as a faculty member in 1947. His mentor at Penn was Simon Kuznets, whose influence on Kravis's work shows in many ways, including the strong belief in and practice of making research replicable. Like many of his generation with economic and statistical training, he worked for the War Production Board during the Second World War, but only partly in Washington. Raymond Bye at Penn and Kravis wrote *Economic Problems of War* in 1942. Kravis also served in Kunming, China, as a logistics officer and worked with Claire Chenault's Flying Tigers involved in tracking supply missions.

Kravis joined the Bureau of Labor Statistics (BLS) prior to returning to Penn. With Irwin Friend of Wharton's finance department, Kravis directed an 18-volume BLS–Wharton Study of Consumer

Expenditures culminating in a major conference in 1959. As part of this research Kravis explored facets of income distribution, which remained one of his long-term intellectual interests.

Kravis maintained a long collaboration with Robert Lipsey, focusing on international price competitiveness. Their 1985 article ‘Towards an Explanation of National Price Levels’ continues to be influential in making clear the many frictions that lead to persistent divergence of both levels and changes in the relationship between purchasing power parities and exchange rates. The law of one price is a fundamental insight in spatial economics, but, as Kravis and Lipsey make clear in much of their joint work, the *flaw of one price* is that, in the world in which we live, the exceptions are frequent, persistent and often systematic.

Kravis had very broad interests in the global economy, including contributions to productivity comparisons, the role of multinationals in international trade and the construction of export and import price indexes. His important 1970 article ‘Trade as a Handmaiden to Growth’ sets forth his view that expanded international trade is better viewed as accompanying rather than causing economic growth. This article continues to capture the attention of economic historians and development economists.

Kravis began his work on PPPs at the OEEC in collaboration with Milton Gilbert, Director of Economics and Statistics, and other staff members including Angus Maddison. They undertook systematic binary purchasing power comparisons between the United States and the four largest European economies, comparing prices of items with written specifications for consumption and investment. These were combined with indirect estimates of government expenditure to make international real product comparisons at the GDP level. Gilbert and Kravis concentrated on binary comparisons with the United States that produced both Paasche and Laspeyeres indexes, as was sometimes done for price indexes at the time. The spreads in these indexes were much larger than had been anticipated, especially for Italy; but it turned out to be a characteristic of such comparisons that was even more pronounced as the range of countries compared increased.

Parallel purchasing power studies were being carried out in the 1950s by the then Economic and Social Commission of Latin America, the European Economic Commission, and the Council for Mutual Economic Assistance (CMEA). These purchasing power studies were important landmarks leading to the establishment of the International Comparison Programme (ICP) of the United Nations in 1968, where Kravis served as a joint director, based at Penn, with a counterpart at the UN Statistical Office. Under his direction benchmark comparisons were carried out for 1970 for 10 and 16 countries, and for 34 countries in 1975, all involving monographs by Kravis and others. In these studies multilateral methods for purchasing power comparisons were worked out, including the country-product dummy (CPD) method of Robert Summers, which has been extended widely in recent BLS and ICP work. An extension of the benchmark work to a total of 100 countries was published in 1978; it became the basis for the current Penn World Table of Alan Heston, Robert Summers and Bettina Aten.

Acceptance of the results of the ICP was slow in coming, but by the late 1980s many economics textbooks were using PPP comparisons. After the 1975 benchmarks responsibility for this work became international, and the European Union and the OECD now routinely carry out comparisons for their 50 member and associate countries. Currently the World Bank is coordinating a global 2005 ICP benchmark comparison involving about 150 countries, a major legacy of Kravis's leadership. While the methods used in the initial ICP work are being modified, the basic multilateral framework pioneered by Kravis remains in place. During a heated debate on methods following presentation of the results of the first ICP benchmark, the late Nancy Ruggles observed that the important thing was not which method

was used but rather that a multilateral PPP comparison was actually completed. It took the focus, patience, wisdom and good humour of Kravis to produce these initial multilateral PPP comparisons in a timely manner and of a quality that has led to their adoption on a global basis.

See Also

- Balassa, Béla
- Cassel, Gustav
- Geary, Robert Charles
- Laspeyres, Ernst Louis Etienne
- poverty lines
- purchasing power parity

Selected works

1942. (With R. Bye.) *Economic Problems of War*. New York: F. S. Crofts.

1954. (With M. Gilbert.) *An International Comparison of National Products and the Purchasing Power of Currencies: A Study of the United States, the United Kingdom, France, Germany, and Italy*. Paris: OEEC.

1970. Trade as a handmaiden of growth. *Economic Journal* 80, 850–72.

1975. (With Z. Kenessey, A. Heston and R. Summers.) *A System of International Comparisons of Gross Product and Purchasing Power*. Baltimore, MD: Johns Hopkins University Press.

1978. (With A. Heston and R. Summers.) *International Comparisons of Real Product and Purchasing Power*. Baltimore, MD: Johns Hopkins University Press.

1978. (With A. Heston and R. Summers.) Real GDP per capita for more than one hundred countries. *Economic Journal* 88, 215–42.

1982. (With A. Heston and R. Summers.) *World Product and Income: International Comparisons of Real GDP*. Baltimore, MD: Johns Hopkins University Press.

1983. (With R. Lipsey.) Towards an explanation of national price levels. Princeton Studies in International Finance No. 52, Department of Economics, Princeton University.

Bibliography

International Comparison Programme (ICP). Online. Available at http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/ICPEXT/0,pagePK:62002243~theSitePK:270065,00.html, accessed 26 April 2006.

Klein, L. 1993. Irving B. Kravis: memoir of a Distinguished Fellow. *Journal of Economic Perspectives* 7 (3), 175–84.

Penn World Table. Online. Available at http://pwt.econ.upenn.edu/php_site/pwt_index.php, accessed 26 April 2006.

How to cite this article

Heston, Alan W. "Kravis, Irving B. (1916–1992)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000070> doi:10.1057/9780230226203.0906

The New Palgrave Dictionary of Economics Online

Kuznets swings

S.N. Solomou

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Kuznets swings refer to variations in economic growth with an average cyclical period of 20 years. The evidence suggests that for the period before the First World War long swings were a dominant and pervasive aspect of national economic growth. Abramovitz (1968) argued that Kuznets swings were a feature only of the pre-1913 era, partly because the migration restrictions introduced in the New World during the interwar period changed the causal process. We consider this idea and argue that, although the features of long swings are time-varying, the idea of a ‘passing of the Kuznets cycle’ needs to be re-evaluated.

Keywords

currency unions; economic and monetary union (EMU); fixed exchange rates; gold standard; international capital flows; international migration; Kuznet swings; long swings in economic growth; migration; multiplier–accelerator interaction; population; real exchange rates; terms of trade

Article

Kuznets swings refer to variations in economic growth with an average cyclical period of 20 years. The average period of the swings found varies depending on the specific smoothing and trend elimination techniques employed (Bird et al., 1965). Kuznets (1958), Abramovitz (1959) and Lewis and O'Leary (1955) found mean swings of 22, 14 and 19 years respectively. The existing historical evidence suggests that for the period before the First World War long swings were a pervasive aspect of national economic growth, being observed in a wide set of economies, including the United States, Argentina, Australia, Brazil, Britain, Canada, France and Germany (Solomou, 1987). Similar results have been reported for Japan (Ohkawa and Rosovsky, 1973; Ohkawa, 1979; Shinohara, 1962).

Long swings have often been explained as the outcome of the pre-1913 economic structure, with population-sensitive investments playing a central role. The emphasis has been on the Anglo-American

economies, with the aim of explaining the economic impact of international migration. Kuznets (1958) suggested that internal and international migration responded to development opportunities in the American economy, inducing multiplier–accelerator effects via the building sector. Abramovitz (1959; 1961) and Easterlin (1968) have offered similar explanations. Working from the migration perspective, Thomas (1973) attempted to explain migration and aggregate long swings in terms of the framework of the ‘Atlantic Economy’ (of Britain and America). The high degree of economic integration in the Atlantic economy implies that the availability of factors of production was a constraint on economic growth within the region. An increase of investment in one region was assumed to result in a decrease of investment in the other region. Since construction activity was greatly influenced by population changes, which, in turn, were influenced by migration movements, migration was seen to be the main force generating inverse swings in output and investment in the Atlantic economy.

That *exogenous* migration movements can generate macroeconomic swings is theoretically plausible. However, to the extent that migration patterns are influenced by economic considerations, Thomas's (1973) model is misleading; the description of endogenous economic processes has been confused with an exogenous explanation of economic change. Moreover, the emphasis on migration as the *causal* variable has led to a neglect of other important determinants of long swings in economic growth. Cairncross (1953) focused on the variation of the international sectoral terms of trade between manufacturing and agriculture in the world economy. Britain, France and Germany were representative of industrial economies producing manufactured commodities while much of the rest of the world was taken to represent the primary-producing sector. Investment flows in the international economy were determined by the relative profitability of these two sectors. Migration flows were not an exogenous force generating long swings but were merely a response to these underlying economic variations. In Cairncross's framework, sectoral terms of trade changes reflected long-run sectoral imbalances in the world economy:

> One would expect to find, therefore, that during, or immediately after, a fairly long period in which the terms of trade were relatively unfavourable to Britain there would be heavy investment in the countries supplying her with imports … On the other hand, when capital goods were expensive and foodstuffs were in over-supply, the continuance of a rapid opening up of agricultural countries would be distinctly surprising. (Cairncross, 1953, p. 189)

Cairncross argued that a similar experience is also observed for the other major capital exporters. All these early studies of pre-1913 long swings emphasize monocausality, partly due to the limited macroeconomic and sectoral data available to the early researchers. Solomou (1987) has shown that long swings were observed for, *inter alia*, aggregate investment, profitability, output, productivity, agricultural output, construction output, weather variables, monetary growth, sectoral terms of trade and migration flows. The swings also have international dimensions; they are observed for overseas investment, international terms of trade and international relative profitability movements, and the balance of trade. Such evidence raises strong doubts about the simple migration and terms of trade explanations for these swings – instead, these variables are best seen as part of a broader causal structure for the observed swings.

Abramovitz (1968) argued that Kuznets swings were a feature only of the pre-1913 era, partly because the migration restrictions introduced in the New World during the interwar period changed the causal process for long swings. However, in arguing the case for ‘the passing of the Kuznets swing’ Abramovitz is relying on the validity of the prior hypothesis that growth swings of this duration were the outcome of *international* migration swings. As argued above, this offers a partial perspective on pre-1913 growth swings. Abramovitz's idea of the passing of the Kuznets swing is inconsistent with evidence on historical economic growth that suggests that long swings were observed after 1913, implying that a much broader causal framework is needed. For example, in the US economy the long-swing pattern of macroeconomic growth continues into the interwar and post-war eras (Hickman, 1974), with *domestic* migration swings playing an important role in the period after 1914 (Easterlin, 1968; Hickman, 1963). In Japan long swings of growth are observed throughout the period before the Second World War (Shinohara, 1962; Solomou and Shimazaki, 2007), accounted for by a broader causal structure than is emphasized in the traditional long-swing literature. This suggests that seeking an explanation for change in the features of long swings is more useful than seeking an explanation for a unique ‘passing’.

After surveying American long swings in the period 1840–1914, Abramovitz (1959, p. 462) concluded,

> It is not yet known whether they are the result of some stable mechanism inherent in the structure of the US economy, or whether they are set in motion by the episodic occurrence of wars, financial panics, or other unsystematic disturbances.

What Abramovitz seems to have had in mind is that the question of endogeneity is an open one. A useful general perspective on long swings is to view the features of this cycle as being a product of the specific policy framework. During the rules-based policy framework of the gold standard, a necessary outcome was the need for cyclical adjustment. This type of adjustment was manifested in a number of ways. In the case of the core industrial countries before the First World War, they were able to sustain the gold standard rule as their policy framework and were able to use migration, capital flow, trade and *real* exchange rate adjustment to cope with a changing and stochastic economic environment (Catão and Solomou, 2005). The slow-relaxing nature of these variables meant that most of the cyclical movement is observed in the long-swing frequency rather than the shorter business-cycle frequency. To this extent Abramovitz is right to say that interwar swings were of a different nature from previous ones; however, this does not constitute a passing of the cycle. In fact, we could argue that during the interwar period, as the conventional adjustment processes of the gold standard epoch disintegrated, discretionary policy became a new adjustment tool determining growth outcomes.

This analysis of Kuznets swings suggests that an understanding of the observed swings requires us to understand the episodic changes that have been observed in different historical periods. This emphasis on policy framework suggests that Kuznets swings could become of increasing relevance in the future. As different policy blocs attempt to establish fixed exchange-rate regimes and single currency areas, the adjustment to shocks in the future will once again become policy-constrained. For example, Europe's commitment to a single currency under economic and monetary union must imply that a number of adjustment mechanisms (such as migration and capital flows) will have to be activated as equilibrating mechanisms to national-specific shocks if we are not to observe persistent divergence across different

countries. Such flows could generate economic–demographic interactions on the cyclical process, which has clear homologies with the gold-standard pattern of international adjustments.

See Also

- Kondratieff cycles
- Kuznets, Simon

Bibliography

Abramovitz, M. 1959. Historical and comparative rates of production, productivity and prices. Statement in US Congress. Joint Economic Committee, *Employment, Growth and Price Levels, Hearings*, 86th Congress, Pt II, Washington, DC, 411–66.

Abramovitz, M. 1961. The nature and significance of Kuznets cycles. *Economic Development and Cultural Change* 9, 225–48.

Abramovitz, M. 1968. The passing of the Kuznets cycle. *Economica* 35, 349–67.

Bird, R.C., Desai, M.J., Enzler, J.J. and Taubman, P.J. 1965. ‘Kuznets cycles’ in growth rates: the meaning. *International Economic Review* 6, 229–39.

Cairncross, A.K. 1953. *Home and Foreign Investment, 1870–1913: Studies in Capital Accumulation*. Cambridge: Cambridge University Press.

Catão, L.A.V. and Solomou, S.N. 2005. Effective exchange rates and the classical gold standard adjustment. *American Economic Review* 95, 1259–75.

Easterlin, R.A. 1966. Economic–demographic interactions and long swings in economic growth. *American Economic Review* 56, 1063–104.

Easterlin, R.A. 1968. *Population, Labor Force and Long Swings in Economic Growth: The American Experience*. New York: Columbia University Press.

Hickman, B.G. 1963. Postwar growth in the US in the light of the long swing hypothesis. *American Economic Review* 53, Papers and Proceedings, 490–507.

Hickman, B.G. 1974. What became of the building cycle? *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, ed. P. David and M. Reder. New York: Academic Press.

Kuznets, S.S. 1930. *Secular Movements in Production and Prices: Their Nature and their Bearing upon*

Cyclical Fluctuations. Boston and New York: Houghton Mifflin.

Kuznets, S.S. 1958. Long swings in the growth of population and in related economic variables. *Proceedings of the American Philosophical Society* 102, 25–52.

Lewis, W.A. and O'Leary, P.J. 1955. Secular swings in production and trade, 1870–1913. *Manchester School of Economics and Social Studies* 23(2), 113–52.

Ohkawa, K. 1979. Aggregate growth and product allocation. In *Patterns of Japanese Economic Development*, ed. K. Ohkawa, and M. Shinohara, with L. Meissner. New Haven: Yale University Press.

Ohkawa, K. and Rosovsky, H. 1973. *Japanese Economic Growth: Trend Acceleration in the Twentieth Century*. Stanford, CA: Stanford University Press.

Shinohara, M. 1962. *Growth and Cycles in the Japanese Economy*. Tokyo: Kinokuniya Bookstore Co. Ltd.

Solomou, S.N. 1987. *Phases of Economic Growth 1850–1973: Kondratieff Waves and Kuznets Swings*. Cambridge: Cambridge University Press.

Solomou, S.N. and Shimazaki, M. 2007. Japanese episodic long swings in economic growth. *Explorations in Economic History* 44, 224–41.

Thomas, B. 1973. *Migration and Economic Growth*, rev. edn. Cambridge: Cambridge University Press.

How to cite this article

Solomou, S. N. "Kuznets swings." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000045> doi:10.1057/9780230226203.0908

The New Palgrave Dictionary of Economics Online

Kuznets, Simon (1901 – 1985)

Richard A. Easterlin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; American Statistical Association; business cycles; Koopmans, T. C.; Kuznets cycles; Kuznets, S.; measurement; Mitchell, W. C.; modern economic growth; National Bureau of Economic Research; national income accounting; national product accounting; social accounting

Article

Kuznets was born in Pinsk, Russia, on 30 April 1901 and died in Cambridge, Massachusetts, on 9 July 1985. After a brief period as youthful head of a statistical office in the Ukraine under the early Soviet regime, Kuznets emigrated to the United States, where he received his BA in 1923, MA in 1924, and Ph. D. in economics in 1926, all from Columbia University. He was a member of the research staff of the National Bureau of Economic Research in New York from 1927 to 1961, and held professorial appointments in economics at the University of Pennsylvania (1930–54), Johns Hopkins University (1954–60), and Harvard University (1960–71). After his retirement, Kuznets continued an active research career for another decade. During the Second World War he was associate director from 1942 to 1944 of the Bureau of Planning and Statistics of the US War Production Board. Kuznets was elected president of the American Economic Association in 1954 and the American Statistical Association in 1949, and was the 1971 recipient of the Nobel prize in economics.

Kuznets's foremost contribution, for which he received the Nobel prize, is an empirically founded comparative study of the economic growth of nations. In this work Kuznets identifies, documents and analyses the emergence of a new epoch in economic history, which he calls ‘modern economic growth’ (Kuznets, 1966). Modern economic growth first makes its appearance in north-western Europe in the latter half of the 18th century. In the course of the 19th century it diffuses southward and eastward throughout Europe, and by the end of the century its beginnings can be identified in Russia and Japan. Mirroring the diffusion pattern within Europe is a somewhat parallel development in overseas areas settled by Europeans. Modern economic growth appears first in areas initially settled by migrants from north-western Europe – the United States in the first part of the 19th century, followed by Canada, Australia, and New Zealand – and subsequently in parts of Latin America where migration from

southern and eastern Europe was especially important. In the 20th century, especially since the Second World War, the initial signs of modern economic growth have become more widespread in parts of Asia, and, to a lesser extent, Africa.

Three conditions especially set off the epoch of modern economic growth from prior forms of economic organizations – the growth rate of real per capita income, the industrial and occupational distribution of the labour force, and the form of population settlement. In economies experiencing modern economic growth the rate of increase of real per capita income has typically averaged around 15 per cent or more per decade over periods of a century or more. In this epoch, there may be shorter or longer fluctuations in the growth rate (such as business cycles, Kuznets cycles, or Kondratieff cycles) but there is no clear evidence of systematic long-term retardation or acceleration. A sustained growth rate of 15 per cent or more per decade is unprecedented in economic history.

In prior epochs of economic organization, economic activity was concentrated in the primary, extractive, sector of the economy, and took the form of agriculture, or, at an earlier time, hunting, gathering and fishing. The era of modern economic growth has witnessed a vast diversification and proliferation of industries and occupations. In today's developed economies, extractive pursuits often account for as little as five per cent or less of the labour force; the secondary sector, chiefly manufacturing and construction, may account for around another third; and the tertiary service sector for the remainder. Within the service sector, a sizeable share of the labour force, approaching the importance of manufacturing in magnitude, is employed in the transportation and distribution of goods, while the remainder is engaged in activities such as personal and professional services, and government. On the occupational side, white collar jobs (managerial, clerical, professional, and sales), of small importance in prior epochs, grow significantly in proportion to blue collar (manual) labour.

Associated with the shift out of agriculture is a major transformation in place of residence of the population. In prior epochs nomadic or village life was the overwhelming form; in contrast, the epoch of modern economic growth has seen the emergence and dominance of spatial concentration, in cities and surrounding suburbs. As a consequence, in many developed economies rural depopulation has been a pervasive phenomenon.

Underlying the acceleration in the growth rate of real per capita incomes and associated reallocation of resources by industry, occupation and location has been a technological revolution, most easily identified by the increased flow since the 18th century of inventions and innovations in economic activity. At bottom, this new technology stems from the emergence of modern science in the 16th and 17th centuries and the empirical outlook to which it gave rise. Modern technology is distinct from the technology of prior epochs in its reliance on inanimate sources of power, the growth in importance of minerals relative to fibres as raw material, the spread of mechanization and an associated increase in optimum scale of manufacturing production leading to replacement of artisanal by factory organization, and new forms of transportation and communication.

To some the epoch of modern economic growth is identified with industrialization and capitalism, but in Kuznets's view this is a misconception. ‘Industrial’ work in the sense of manufacturing and construction, accounts, as has been noted, for a minority share of economic activity in presently developed countries, and, in some of these countries, modern economic growth has been based primarily on the commercialization and technological modernization of agriculture rather than on industry. Moreover, the phenomenon of modern economic growth transcends specific institutional forms such as capitalism, socialism or communism. As Kuznets demonstrated in numerous works, the dramatic rise in the growth

rate of real per capita income, the immense reallocation of resources among economic activities, the spatial concentration of population, the adoption of a common modern technology, and many other features of contemporary economic development have been essentially similar in the United States and the Soviet Union, western and eastern Europe, and, in incipient stages today, China, India and Brazil.
The analysis of modern economic growth was the logical culmination of Kuznets's earlier work. The organizing framework for his comparative study was national income and its components, in whose conceptualization and measurement Kuznets was the foremost pioneer. Indeed, so great are his accomplishments in the measurement of national income and so close the identification of national income with his name that American economists frequently cite this work as the basis for his Nobel award. Certainly, there is no doubt that this work too is a landmark in the evolution of economic science. Today figures of gross national product (GNP) are taken for granted, but before the First World War there was almost total ignorance of such elementary facts of the economy's size and structure. Kuznets was not the first to seek to close this gap, but his work on national income and product was so distinctive and comprehensive that it became the benchmark in the field. It encompassed estimates of total output and income by final product, industry of origin, and type of income; capital formation and savings; and the distribution of income between rich and poor. This work, coinciding with the new demands for economic information generated, first, by the Great Depression of the 1930s and then by the mobilization requirements of the Second World War, laid the foundation for the establishment of official estimates of total GNP and its components by the federal government, a task in which Kuznets played a leading role. As mentioned, it also provided the basis for Kuznets's subsequent programme of research on economic growth, which was built upon historical series of national income and product for as many countries as possible.
Kuznets's work on national income played a crucial role in the transition of economics from a deductive to a quantitative science. This transition required a union of theory, economic measurement and statistical methodology. In the 1930s the new macroeconomic theories of John Maynard Keynes had aroused much interest because of their relevance to the worldwide economic crisis. Kuznets's concurrent and independent effort to develop measures of the consumption, savings, and investment components of national income provided the empirical counterparts of the Keynesian concepts. This advance in economic measurement and its concordance with new theoretical formulations was a key step in the development of econometrics, the statistical techniques for systematic quantitative modelling of the economic system pioneered by Ragnar Frisch and Jan Tinbergen.
In one of his earliest works, on secular movements in production and prices, Kuznets identified fluctuations of 15–25 years’ duration in a number of economic time series in the United States. Subsequently he returned to this subject several times, widening the range of observation to other developed countries and incorporating demographic as well as economic time series. These movements, although still somewhat controversial, are commonly referred to today as ‘Kuznets cycles’, in recognition of his pioneering contribution.
In the history of economic thought, Kuznets stands in a line of descent tracing back through the American institutional school to the German historical school and thence to Karl Marx. The common thread is a search for laws or generalizations about long-term economic development based on comparative study of historical experience. The unique feature of Kuznets's work, which endows it with the prospect of more enduring success, is its foundation in quantitative measurement. In using national income as the key organizing principle of his comparative studies, Kuznets made possible the replication

and extension of his work by others and thus the cumulation of a body of systematic knowledge about economic development forming the basis for tested generalizations.

Although Kuznets identified modern economic growth as a distinctive epoch in economic organization, he did not extend his empirical approach to the study of earlier epochs. In this respect, Kuznets was much like Marx and Joseph Schumpeter, in that he focused his attention on the study of a single, contemporary stage of economic evolution, although, unlike them, he did not see this stage as confined to the boundaries of capitalism.

Kuznets's approach to economics and to economic research may best be understood in terms of his intellectual heritage. Although Kuznets' convictions about the importance of quantitative measurement and knowledge of economic history antedated his emigration to the United States, they received strong reinforcement from his mentor at Columbia, Wesley C. Mitchell. Mitchell's scepticism about the reliance of economics on deductive economic theory and his belief in the need for quantitative facts had been instrumental in the establishment in 1920 of the National Bureau of Economic Research, a non-profit research organization devoted to basic economic science, the first of its kind in America. Mitchell brought Kuznets into the NBER, where he conducted his national income studies and came to head the bureau's programme in that field. This project, and that on business cycles, headed by Mitchell and Arthur F. Burns, were the central pillars of the bureau's work, and the basis for the national and international reputation that the bureau established in the 1930s and 1940s. Two themes of the NBER's work – a respect for facts and the cumulation of economic knowledge through quantitative measurement – were key ingredients in Kuznets' research strategy.

The quintessential criticism of the NBER approach was captured in the title of Tjalling Koopmans' famed 'Measurement without Theory' review of the Burns–Mitchell treatise on business cycles (Koopmans, 1947). In Kuznets's view, however, measurement can never be divorced from economic theory, and is necessarily guided by theory. Indeed, in the late 1940s Kuznets broke with the official estimators of GNP, because he considered their increasing emphasis on 'social accounting' to be an abandonment of the fundamental Marshallian and Pigovian theoretical concept of national income as a measure of economic welfare. Throughout his career, Kuznets's monographs, although structured around tables of data, were infused with 'tentative' interpretations and explanations based on economic theory. Admittedly Kuznets was reserved in his use of economic theory and sceptical of formal mathematical and econometric models. This arose, however, not from a rejection of theory but from another feature of his approach that he shared in common with the historical school. This notion is the historical relativity of economic theory. To Kuznets, much economic writing and theorizing was geared to current conditions and claimed validity far beyond the limits that would be revealed by an empirical test. His reservations about economic theory also stemmed from what he felt was its limited coverage of social reality. Particularly in the study of economic growth was an expansion of disciplinary boundaries necessary. Much of Kuznets' work on modern economic growth was carried out under the auspices of the Committee on Economic Growth of the Social Science Research Council. This committee, which Kuznets chaired and directed for the two decades of its life, included important representation from anthropology, sociology, and political science, and organized 16 interdisciplinary conferences. Some of Kuznets' own research involved interdisciplinary cooperation, notably in his collaboration with the distinguished sociologist, Dorothy S. Thomas, in their study of population redistribution and economic growth in the United States.

Simon Kuznets was a devoted family man and warm with intimates. Perhaps his happiest moments,

however, were those frequent long mornings spent over a calculator, bending the diverse facts of reality to manageable size. A brief anecdote may capture the spirit of the man. Once, after his retirement, Kuznets was invited to prepare a paper for a forthcoming conference, and, in reply, he asked how long it would be before the conference proceedings were published. When his interrogator quipped, ‘Why, Simon, you're not still interested in “publish or perish”, are you?’, he responded with a twinkle, ‘Well, in a sense, I am.’

Though one of the first Nobel prize winners in economics, Kuznets was in important respects a maverick. In a discipline where deductive analysis is the hallmark of accomplishment, Kuznets, though himself a creative and original thinker, was notable for his insistence on facts and measurement. In a field that prides itself as ‘queen of the social sciences’, Kuznets reached out to other disciplines both in teaching and research. And in a subject where sweeping ideological prescriptions for reform abound, Kuznets was in both words and example a passionate believer in the ultimate value of science.

Selected works

1926. *Cyclical Fluctuations, Retail and Wholesale Trade, United States, 1919–1925*. New York: Adelphi.

1930. *Secular Movements in Production and Prices: Their Nature and Their Bearing upon Cyclical Fluctuations*. Boston: Houghton Mifflin.

1933. *Seasonal Variations in Industry and Trade*. New York: NBER.

1934. *National Income, 1929–1932*. Senate Document No. 124, 73rd Congress, 2nd Session. Washington DC: Government Printing Office.

1937. *National Income and Capital Formation, 1919–1935*. New York: NBER.

1938. *Commodity Flow and Capital Formation*, vol. 1, New York: NBER.

1941. (Assisted by L. Epstein and E. Jenks.) *National Income and Its Composition, 1919–1938*, 2 vols. New York: NBER.

1945. *National Product in Wartime*. New York: NBER.

1945. (With M. Friedman.) *Income from Independent Professional Practice*. New York: NBER.

1946. (Assisted by L. Epstein and E. Jenks.) *National Product Since 1869*. New York: NBER.

1946. *National Income: A Summary of Findings*. New York: NBER.

1953. (Assisted by E. Jenks.) *Shares of Upper Income Groups in Income and Savings*. New York: NBER.

1953. *Economic Change: Selected Essays in Business Cycles, National Income and Economic Growth*. New York: W.W. Norton.

1957–64. (ed., with D. S. Thomas, and co-author.) *Population Redistribution and Economic Growth: United States, 1870–1950*, 3 vols. Philadelphia: American Philosophical Society.

1959. *Six Lectures on Economic Growth*. Glencoe, IL: Free Press.

1961. (Assisted by E. Jenks.) *Capital in the American Economy: Its Formation and Financing*. Princeton: Princeton University Press.

1964. *Postwar Economic Growth: Four Lectures*. Cambridge, MA: The Belknap Press of Harvard University Press.

1965. *Economic Growth and Structure: Selected Essays*. New York: W.W. Norton.

1966. *Modern Economic Growth: Rate, Structure, and Spread*. New Haven: Yale University Press.

1968. *Toward a Theory of Economic Growth, with Reflections on the Economic Growth of Modern Nations*. New York: W.W. Norton.

1971. *Economic Growth of Nations: Total Output and Production Structure*. Cambridge, MA: The Belknap Press of Harvard University Press.

1973. *Population, Capital, and Growth: Selected Essays*. New York: W.W. Norton.

1979. *Growth, Population, and Income Distribution: Selected Essays*. New York: W.W. Norton. See also ‘Bibliography of Simon Kuznets’, *Economic Development and Cultural Change* 31(2) (1983), 433–54.

Bibliography

Koopmans, T.C. 1947. Measurement without theory. *Review of Economics and Statistics* 29(3), 161–72.

How to cite this article

Easterlin, Richard A. "Kuznets, Simon (1901–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000046> doi:10.1057/9780230226203.0907

The New Palgrave Dictionary of Economics Online

Kydland, Finn Erling (1943 –)

Carlos E.J.M. Zarazaga
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Finn Kydland's contributions to economics science have changed the terms of the debate in two important and related counts: the theory of policymaking and of business cycles. In his Ph.D. dissertation, Finn showed that a complex ‘credibility problem’, inherent to the policymaking process, prevented the evaluation of economic policies with the optimal control theory techniques applied until then. His work with Edward Prescott on business cycles identified supply shocks as one of the primary causes of economic fluctuations, with the counter-intuitive and therefore resisted implication that the perfect smoothing of the business cycle may not be a sensible policy objective.

Keywords

assignment problem; Bellman's principle of optimality; business cycles; Friedman, M.; Keynesianism; Kydland, F.; Lucas, R.; monetarism; monetary shocks; neoclassical growth theory; Norwegian School of Economics and Business Administration; optimal control theory; output-consumption correlation puzzle; Phillips curve; Prescott, E.; rules vs. discretion; system-of-equation approach; targets; technology shocks

Article

Finn was born in Bjerkreim, Norway, in the southern tip of the country, and grew up in nearby Søyland with his mother Johanna, his five younger siblings, and his father Martin, who ran a family business that hauled milk and sheep.
Finn was the only one among his classmates to further his education beyond elementary school: at the early age of 15, he moved by himself to Bryne, 20 miles away from home, to be able to attend the nearest high school. His excellent grades there allowed him to apply to the Norwegian School of Economics and Business Administration (abbreviated NHH in Norwegian) at Bergen. He was at first rejected because the secondary school he had attended didn't have a business orientation. With his usual determination to fight adversity, Finn studied some more to acquire the necessary qualifications and

finally started his business degree at the NHH in August 1965.

There, in the winter of 1968 he would be hit by a major real shock that would derail him for ever from a business career track. The source of the shock was Sten Thore, the professor of economics teaching a managerial sciences course, in which Finn wrote his first computer program (in FORTRAN) doing dynamic programming, a tool he would use repeatedly throughout his career. According to Sten Thore's testimony (2005), ‘Finn quickly established himself as the smartest kid in the class’. It became obvious to him that this low-key farmer boy with a scientific mind had not been born for the business world. Secretly hoping that Finn would find his true calling in life, in the winter of 1968 Sten asked Finn to be his research assistant, ‘grasping him in the nick of time after his graduation, before he had time to disappear into the commercial world’.

A few months later Sten was given the opportunity to spend a year as a visiting professor at the Graduate School of Industrial Administration (GSIA), at Carnegie-Mellon University, and asked Finn to accompany him as his research assistant. When Finn joined Sten in Pittsburgh in the summer of 1969, his fate was sealed.

In that exciting intellectual atmosphere, it didn't take Finn long to catch the economics bug, apply formally to the graduate programme, and be promptly admitted. There, in the following spring, while attending a course by young professor Robert Lucas, Finn would see unfold on the blackboard, as it was conceived, Lucas's seminal paper ‘Expectations and the Neutrality of Money’ (1972). Without either of them knowing it, two future Nobel prizewinners were tuning their minds in the same classroom.

Another fortuitous encounter took place in August 1971, when Finn ran into Edward Prescott, a newly hired professor who asked him what he was working on.

At that time, Finn had become interested in the so-called assignment problem. The problem, formulated in the context of the system-of-equation framework dominating macroeconomics at the time, was to determine the most effective policy instrument for targeting each of multiple, potentially conflicting policy objectives (such as maintaining low inflation and full employment).

Finn was taking an out-of-the-box approach, not in terms of the usual system-of-equations but of a game between the monetary and fiscal authorities. Prescott was curious because he had been doing related research with Robert Lucas. But he was taken aback by the discovery that Finn was about to make: that the outcome of the game was different, depending on whether the players were forced to make all their decisions for the entire future at the beginning of the planning period (the open-loop solution in the so-called sequence space) or allowed to choose their actions one period at a time (the feedback solution in the so-called policy space).

Engineers and physicists routinely using optimal control theory had not considered that possibility because, by Bellman's principle of optimality, both solutions are equivalent in their ‘mechanical’ world. Intuitively, the reason is that a mechanical device implements instructions exactly as written in the program currently fed into it, regardless of whether in the future it will be asked to implement a different program. Put differently, machines are incapable of conditioning current behaviour on the future.

It was not surprising that economists at the time thought that the results from optimal control theory carried over automatically to economic policy questions. They had been trained in the system-of-equations tradition, which modelled the behaviour of economic agents as if they were machines in the sense that households and firms were allowed to make decisions using only information from the past, even if humans, unlike machines, can anticipate the future and therefore condition current actions on the economic programs (or policies) they expect to be implemented in the future. That artificial assumption

ensured the absence of feedback from the future to the present necessary for the principle of optimality to hold.

It is not by chance that Finn had his first Nobel-calibre insight at the GSIA and at that time. After all, it was there, in the spring of 1970, that he witnessed Lucas produce a path-breaking paper showing the serious shortcomings of evaluating alternative policies with the mechanical behavior of economic agents imbedded in the system-of-equations approach. However, in Lucas's 1972 paper the principle of optimality held because he could make his point about the constant money growth policies advocated by Milton Friedman, keeping intact the ‘single-player against nature’ structure underlying the system-of-equations approach with the policymaker – a random money growth process – instead of households behaving like a machine.

In his dissertation, Finn took Lucas's contribution one step further by assuming that the policymaker (a dominant fiscal authority) didn't behave mechanically but was forward-looking as well, picking policies strategically, depending on the reactions of the other optimizing agents (households or a follower monetary authority). With both participants in the policy game reacting strategically to each other's future decisions, the condition of no feedback from the future to the present required for the principle of optimality to hold was not met. As a consequence, decisions made one at a time are different from decisions made once and for the entire future.

That result had far-reaching consequences for the theory of policymaking. It implied that the reason governments around the world seemed to be unable to implement policies that were the best according to optimal control theory was not necessarily, as it was widely believed, myopic or incompetent policymakers. Rather, it was the inherently dynamic nature of the policymaking process, when there is feedback from the future to the present and societies lack commitment mechanisms to bind the decisions of not-yet-born policymakers. The profession at large would become fully aware of the importance of this revelation only later, after Finn joined forces with Ed Prescott to address the issue of optimal selection of policies in uncertain, dynamic environments that Prescott had been exploring earlier with Lucas.

But first Finn had to finish his degree, which he did in May 1973 with a gold medal. He then returned to Bergen as an assistant professor at the NHH, to fulfil the conditions of the fellowship he had received from that school. There he managed to publish the stunning discovery of his dissertation in a 1975 International Economic Review paper.

He also got Ed Prescott to visit the school for the 1974–5 academic year. In the spring of 1975 they came out with a paper with the provocative title ‘On the Inapplicability of Optimal Control for Policy Making’. The paper was received with the same scepticism as Finn's dissertation's first draft: everyone was expecting the principle of optimality to hold and trying to spot the error. Those difficulties persuaded Ed Prescott to add a conventional Phillips curve example to their paper before resubmitting it to the *Journal of Political Economy*, where it appeared in 1977.

Aware of the theoretical result in the paper, Finn was nevertheless surprised by the quantitative finding reported in it: that the time-consistent plan (the ‘decisions-one-period-at-a-time’ solution), arguably the one that governments will be most tempted to implement, represented a sizable loss of welfare relative to the optimal plan (the ‘decisions-for-the-entire-future’ solution). The theoretical contribution in his dissertation was not just an intellectual curiosity: it had concrete implications for the real world. This result surely accounts for the huge impact that the paper had in the profession and may help to explain the quantitative focus of Finn's subsequent research.

The rules-rather-than-discretion paper reopened the US market to Finn. In 1976–7 he spent the academic year as visiting faculty at the University of Minnesota. In 1978, the economics department at Carnegie-Mellon appointed Finn as associate professor at the GSIA, breaking an almost inviolable rule among US universities of not offering permanent faculty positions to their own Ph.D. graduates. Finn became full professor there in 1982.

That was also the year in which the other seminal contribution by Kydland and Prescott appeared in *Econometrica*. ‘Time-to-build and Aggregate Fluctuations’ is a testimony of the quantitative focus that the rules-rather-than-discretion paper impinged on Finn's research agenda. The question that the time-to-build paper set out to answer was: ‘If total factor productivity (or technology) shocks were the only source of impulse, what portion of business-cycle fluctuations could they account for?’ The answer was more than one-half, later raised to around 70 per cent.

That finding sent shockwaves throughout the profession because it undermined fundamental tenets of the monetarist and Keynesian rival schools of thought then dominating the profession. The monetarists couldn't come to terms with the finding that technology shocks, and not monetary shocks, were the most significant source of economic fluctuations. The Keynesians weren't amused either: they had been attributing economic fluctuations to demand-side shocks, and not to supply side shocks like the ones Kydland and Prescott had just unveiled.

Adding insult to injury, the time-to-build paper seriously challenged both schools’ long-held presumption that only models with nominal rigidities in prices and/or wages would be capable of producing fluctuations like the ones observed in the real world. Kydland and Prescott proved that presumption wrong: a neoclassical growth model with flexible prices fed with productivity shocks like the ones observed in the United States was perfectly capable of accounting for about two-thirds of that country's post-war cycles. Therein lay the methodological beauty and conceptual significance of the paper: it derived the quantitative business-cycle implications of well-established growth theory. It is hard to understand in retrospect why such a sensible research project ruffled so many feathers in the profession.

The answer lies in the policy implications: Kydland and Prescott's model, calibrated to US long-run economic growth features, suggests that that country's business-cycle fluctuations since the Second World War can be attributed mostly to the optimal responses of the private sector to exogenous (independent of economic policies) productivity shocks. Under that interpretation, the perfect smoothing of the business cycle that monetarists and Keynesians had been advocating not only is not a sensible policy objective, but can also result in large welfare losses. Neither camp has surrendered, but interestingly enough their subsequent attempts to overturn the finding of the time-to-build paper have preserved many of its distinctive features: money is not explicitly included in the analysis (as in Woodford's ‘cashless’ economy, 2003), agents exhibit forward-looking and optimizing behaviour and, more ironically, prices are a lot more flexible than in pre-1982 monetarist or Keynesian models. Unlike with the rules-rather-than-discretion paper, the controversy generated by the time-to-build paper rages on, eloquent testimony of the indelible mark it has left in the profession.

That paper was by no means Finn's last accomplishment. He has continued publishing consistently in the top journals of the profession, recruiting coauthors who share his methodological views and taste for quantitative questions motivated by anomalies and puzzles. For example, international economists have been trying to account for the ‘output-consumption correlation puzzle’ ever since he and coauthors David Backus and Patrick Kehoe uncovered it in a 1992 *Journal of Political Economy* paper.

Recognizing the calibre of his contributions, the Bank of Sweden awarded to Finn the Nobel Prize in Economics on 11 October 2004, jointly with another giant of the profession, his coauthor Edward Prescott. Shortly before, Finn had accepted the Jeffrey Henley Chair in Economics at the University of California at Santa Barbara. Shortly after, he married Tonya Engstler at their home in Vancouver, Canada.

Inevitably a busy scholar, Finn nevertheless finds time to see his four children from a previous marriage (Martin, Eirik, Camilla and Kari), listen to blues over beers with his colleagues, ride his Ducati, and watch or play soccer. In fact, no award exhibits him more unashamedly than the lifetime membership bestowed upon him in November 2004 by Club Atlético Boca Juniors, the Argentine soccer team he became an unruly fan of while watching, in my company, famous Diego Maradona play for that team in Buenos Aires the last game of his professional career.

See Also

- calibration
- central bank independence
- international real business cycles
- Keynesianism
- Lucas, Robert
- monetarism
- neoclassical synthesis
- Prescott, Edward Christian
- rational expectations
- real business cycles
- time consistency of monetary and fiscal policy
- welfare costs of business cycles

Selected works

1973. Decentralized macroeconomic planning. Doctoral dissertation, Carnegie-Mellon University.

1975. Noncooperative and dominant player solutions in discrete dynamic games. *International Economic Review* 16, 321–35.

1977. (With E. Prescott.) Rules rather than discretion: the time inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

1982. (With E. Prescott.) Time-to-build and aggregate fluctuations. *Econometrica* 50, 1345–70.

1992. (With D. Backus and P. Kehoe.) International real business cycles. *Journal of Political Economy*

100, 745–75.

1996. (With E. Prescott.) The computational experiment: an econometric tool. *Journal of Economic Perspectives* 10(1), 69–85.

1997. (With E. Prescott.) Response to Milton Friedman. *Journal of Economic Perspectives* 11(1), 210–11.

2005. Quantitative aggregate theory. Nobel Prize lecture. In *Les Prix Nobel. The Nobel Prizes 2004*, ed. T. Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/2004/kydland-lecture.pdf, accessed 27 June 2006.

Bibliography

Friedman, M. 1997. Computational experiments. *Journal of Economic Perspectives* 11(1), 209–10.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R. 1994. Review of Milton Friedman and Anna J. Schwartz's ‘A Monetary History of the United States, 1867–1960’. *Journal of Monetary Economics* 34, 5–16.

Njølstad, O., Steigum, E. and Thonstad, T. 2006. Conversation with Finn Kydland. In *Norske Nobel prisvinnere, Fra Bjørnson til Kydland* [Norwegian Nobel laureates: from Bjørnson to Kydland], ed. O. Njølstad. Oslo: Universitetsforlaget.

Prescott, E. 2005. The transformation of macroeconomic policy and research. Nobel Prize lecture. In *Les Prix Nobel. The Nobel Prizes 2004*, ed. T. Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/2004/prescott-lecture.pdf, accessed 27 June 2006.

Thore, S. 2005. *The Road to the Nobel: A Travelogue and Retrospect on the Occasion of Finn Kydland Receiving the Nobel Prize in Economics on December 10, 2004*. Online. Available at http://www.stenthore.info/The%20Road%20to%20the%20Nobel.doc Accessed 1 July 2006.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

Zarazaga, Carlos E. J. M. "Kydland, Finn Erling (1943–)." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000066> doi:10.1057/9780230226203.0909

The New Palgrave Dictionary of Economics Online

Kyrk, Hazel (1886 – 1957)

Robert W. Dimand and Richard A. Lobdell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

consumer economics; consumer price index; family economics; Kyrk, H.; new home economics; Reid, M. G.

Article

The University of Chicago economist and home economist Hazel Kyrk was a pioneer in the study of consumption decisions and of the allocation of time in households. Born in Ashley, Ohio, Kyrk was the only child of Elmer Kyrk, a drayman, and Jane Benedict Kyrk, a homemaker who died while her daughter was a teenager. After finishing high school, Hazel Kyrk taught for three years before entering Ohio Wesleyan University in 1904, where she supported herself by working as a mother's helper in the household of Leon Carroll Marshall, an economics professor. When he was hired by the University of Chicago, Kyrk went with the family. She graduated from the University of Chicago in 1910 with a Ph.B. in economics and a Phi Beta Kappa key. After a year as an instructor in economics at Wellesley College, Kyrk returned to the University of Chicago to study for a Ph.D. in economics, writing her dissertation with the economic demographer James A. Field. From 1914, she also taught at Oberlin College, first as an instructor, then as an assistant professor. Taking leave from Oberlin in 1918–19 to work on her thesis, she followed her adviser to London, where she served as a statistician for the American Division of the Allied Maritime Transport Council. Her dissertation, accepted in 1920, was published as *A Theory of Consumption* (1923) and won the prestigious, thousand-dollar Hart, Schaffner and Marx Prize for economic research. In that book and in *The Economic Problems of the Family* (1929), Kyrk discussed how social psychology shapes consumer choice and how the economic role of the housewife was moving beyond household production to being a ‘director of consumption’.

Hazel Kyrk worked at the Food Research Institute of Stanford University in 1923–4, co-authoring a study of the American baking industry, and taught at Iowa State College (1924–5). From 1925 until her retirement in 1952, she taught at the University of Chicago, appointed to both the Departments of Economics and Home Economics, with promotion to full professor in 1941. She made the University of Chicago the leading centre of consumer and family economics, supervising many dissertations, notably

Margaret Reid's *The Economics of Household Production* (1934). Reid, the first female Distinguished Fellow of the American Economic Association, returned from Iowa State to Chicago as a full professor of economics and home economics in 1951, a year before the retirement of her mentor, Kyrk. Active in consumer economics beyond the university, from 1938 to 1941 Kyrk spent the summers as principal economist in the Bureau of Home Economics of the Department of Agriculture, working on the 20-volume Consumer Purchases Study which, among other contributions, established base-year prices for the cost of living index. From 1943 she chaired the Consumer Advisory Committee of the wartime Office of Price Administration. In 1945–6, she returned to Washington to chair the Technical Advisory Committee of the Bureau of Labor Statistics, helping to create a ‘standard family budget’ and to revise the consumer price index. Kyrk was also active on the boards of the Chicago Women's Trade Union League and a consumer cooperative in Chicago's Hyde Park neighbourhood, and from 1922 to 1925 she taught at the Bryn Mawr Summer School for Women Workers. Never married, Kyrk took charge of bringing up and educating a teenage cousin.

The economic analysis of household time-allocation and production and consumption decisions by Kyrk and her Chicago students (most notably Reid) prefigured the later Chicago ‘new home economics’ of Gary Becker (see Reid, 1977). Kyrk's *Theory of Consumption* is also recognized as a landmark in the history of marketing thought (Zuckerman and Carsky, 1990).

See Also

- Becker, Gary S.
- family economics
- household production and public goods
- Reid, Margaret Gilpin

Selected works

1923. *A Theory of Consumption*. New York and Boston: Houghton Mifflin.

1929. *The Economic Problems of the Family*. New York and London: Harper & Brothers.

1950. The income distribution as a measure of economic welfare. *American Economic Review* 40, 342–55.

1953. *The Family in the American Economy*. Chicago: University of Chicago Press.

Bibliography

Hirschfeld, M.L. 1997. Methodological stance and consumption theory: a lesson in feminist methodology. In *New Economics and its History*, ed. J.B. Davis. Annual Supplement to *History of Political Economy 29*. Durham, NC: Duke University Press.

Nelson, E. 1980. Kyrk, Hazel. In *Notable American Women: The Modern Period*, ed. B. Sicherman and C.H. Green. Cambridge, MA: Harvard University Press.

Reid, M.G. 1934. *The Economics of Household Production*. New York: Wiley.

Reid, M.G. 1972. Hazel Kyrk. In *History of the Department of Home Economics at the University of Chicago*, ed. M. Dye. Chicago: Department of Home Economics, University of Chicago.

Reid, M.G. 1977. How new is the ‘New Home Economics’? *Journal of Consumer Research* 4, 181–3.

Zuckerman, M.E. and Carsky, M.L. 1990. The contribution of women to U.S. marketing thought: the consumers’ perspective, 1900–1940. *Journal of the Academy of Marketing Science* 18, 313–8.

How to cite this article

Dimand, Robert W. and Richard A. Lobdell. "Kyrk, Hazel (1886–1957)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_K000075> doi:10.1057/9780230226203.0910

The New Palgrave Dictionary of Economics Online

laboratory financial markets

Daniel Friedman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Small-scale financial markets have been studied in the laboratory for more than two decades. Typically, 6–20 human subjects buy and sell units of a single asset whose dividends extend over several periods and/or are uncertain. Such markets permit direct observation of informational efficiency, and allow sharp tests of theoretical predictions. They also provide test beds for policy initiatives, new market formats and automated trading strategies.

Keywords

arbitrage; asset price formation; behavioural finance; bid–ask markets; bubbles; call markets; continuous double auction; efficient markets hypothesis; equity premium; experimental economics; financial market anomalies; information aggregation; Iowa Electronic Market; laboratory financial markets; learning; market institutions; posted offer; prediction markets; prospect theory; South Sea bubble; stationary repetition; tulipmania

Article

Laboratory financial markets allow human subjects to trade assets under conditions controlled by the researcher. By varying the conditions – such as the trading format, or the timing and content of private information – the researcher can make direct and sharp inferences.

Such inferences are crucial to achieve insight into the ongoing debate about the importance of behavioural anomalies in financial markets (see behavioural finance). Efficient markets and related theories provide a satisfying explanation for many of the properties of modern financial markets, but they are hard to reconcile with well documented ‘market anomalies’ such as home bias, the large equity premium and excessive volatility. Should financial economists force a reconciliation, or should they embrace prospect theory and other behavioural theories?

These issues are not just academic. Since the collapse of the Soviet bloc around 1990, a dominant share of the world economy has relied on financial markets to choose its economic future. If the efficient markets theory is wrong, and asset prices do not necessarily reflect all available information, then major restructuring may be in order. Perhaps the global economy would be stronger with information disclosures that cater to our behavioural idiosyncrasies, or even with non-market allocation of investment.

Laboratory asset markets inform the debate by offering evidence that complements field data. The strength of experimental methodology is that the researcher can precisely control information, public and private, and can elicit beliefs as well as track offers, transactions and allocations. Thus, in a simplified setting, researchers can systematically dissect the process of asset price formation. In conjunction with theory and field empirical work, laboratory investigations help us understand how financial markets really work.

Early laboratory markets

Experimental economics cut its teeth on laboratory commodity markets. Reacting to Edward Chamberlin's casual classroom experiments, Vernon Smith pioneered the scientific study of markets in the laboratory. He refined the idea of *induced value and cost*: the experimenter promises to pay a subject the amount v if she buys a unit, and charges another subject the amount c if he sells a unit. If they transact at price p, she earns $v–p$ and he earns $p–c$, generating surplus of $v–c$. The payments are in cash and large enough for the subjects to take

seriously.

Smith introduced *stationary repetition* – several consecutive trading periods with the same endowed values and costs but no carry-over from one period to the next, so that subjects have the opportunity to adapt to the trading environment. He also brought the *continuous double auction* (CDA) market (sometimes referred to as the double oral auction) format into the laboratory: traders can make public, committed offers to buy and to sell and can accept others' offers at any time during a trading period. Variants of the CDA format predominate in modern financial markets, including the New York Stock Exchange (NYSE), NASDAQ, and the Chicago Mercantile Exchange.

Numerous laboratory studies, beginning with Smith (1962), show that CDA markets with only a few buyers and sellers (say, four of each) reliably produce highly efficient outcomes, where efficiency is defined as the fraction of potential surplus in the market that is captured by the buyers and sellers. Typically, over 95 per cent of total surplus is realized after a few periods of stationary repetition.

Such perishable commodity markets provide no interesting role for time or uncertainty, both important dimensions of financial assets. Laboratory financial markets should allow two-way traders who can both buy and sell, and who trade assets with a payout that is uncertain and/or carries over several periods. Experimenters at Caltech first introduced such markets in the early 1980s. For example, Plott and Sunder (1982) created a single period asset that was traded by six uninformed traders, who knew only that one of two states would occur with given probabilities independently each period, and six informed traders, who knew the realized state. Both informed and uninformed traders were distributed evenly across three types of state-contingent dividend schedules. Within a few periods, prices became highly efficient, and the trading patterns demonstrated that the market fully disseminated the private information. About the same time, several teams of researchers found very efficient asset prices in laboratory markets with assets paying individual- and state-contingent dividends over several trading periods. These and other early laboratory experiments demonstrated that futures and options contracts can speed convergence towards efficient asset prices. See Sunder (1995) for a thorough survey.

The main lesson from these studies is that financial markets can process information very efficiently. As Hayek (1945) conjectured, markets can fully aggregate and disseminate dispersed private information, and can do so quite rapidly. A few bids and asks in the CDA suffice to fully inform experienced traders, dealing appropriate assets, in moderately complex environments.

Dissecting financial markets

These positive early results encourage us to look more deeply at how financial markets process information. The process has several logical stages. Investors and other participants acquire relevant information from diverse sources, public and private. Individual investors incorporate the information into their beliefs about future asset prices. Acting on their beliefs, investors try to buy assets they expect to appreciate relatively rapidly and to sell assets that they expect to do less well. Their buy and sell orders in turn produce observable market outcomes such as asset price and trading volume. The market outcomes provide further public information for investors, other new information arrives from time to time, and so the process continues. We now know that the process can work quite well in favourable circumstances. But even the early laboratory studies show that it is sometimes fallible. When and where might it go wrong?

Each stage of the process can be examined in the laboratory and compared with theoretical predictions. Cognitive scientists focus on the first stage, the formation of beliefs given arriving information, and have documented many biases that might distort beliefs. Examples include overconfidence, the gambler's fallacy (believing that a coin that has come up 'heads' many times in succession is the more likely to come up 'tails') and the hot-hand fallacy (believing that basketball players who have made ten free throws in succession are especially likely to make the next). In the next stage, investors may make decision errors when they buy and sell assets, even when their beliefs are realistic. There are numerous examples, including hyperbolic (or quasi-hyperbolic) discounting, the disposition effect, and the sunk-cost fallacy.

It is often tempting to explain financial market anomalies simply by pointing to one or more of these biases and errors. But such explanations are incomplete and potentially erroneous. One problem is that there are so many documented biases and errors; indeed, a complete list seems not to exist. Given any market anomaly A, a diligent student can always find some decision error or bias B that superficially seems connected, whether or not B really causes A. Even more important, investors' biases and decision errors never translate directly into financial market imperfections. Asset prices are non-trivial functions of investors' buy and sell orders, and they provide information that affects subsequent orders and prices. These later stages of the process depend on the market format, and they can attenuate or amplify investors' biases and errors.

Attenuating biases and errors

Three different market forces can greatly attenuate the financial market impact of erratic investors. First, it is a powerful learning experience to lose money in a financial market, or even to see other investors do better when they have no informational advantage. Friedman (1998) and later studies demonstrate that people can overcome even the strongest biases and errors in a suitable learning environment. To the extent that a bias or error leads to clearly inferior performance, an investor will learn to do better over time. Subjects in most laboratory financial markets commit fewer errors and trade more efficiently in later periods than in earlier periods, and subjects with previous experience in a particular laboratory market do better yet.

Second, the market shares of investors with inferior trading strategies tend to shrink over time, reducing their influence on market performance. Blume and Easley (1992) demonstrate theoretically that wealth redistribution eventually eliminates all but the most effective investors. Laboratory studies routinely cancel out this force via stationary repetition, but it can easily be inferred by compounding relative profits across periods.

Third, persistent costly errors and biases create profit opportunities for entrepreneurs whose efforts attenuate (or even eliminate) the market impact. For example, yellow pages and speed dials help us overcome our cognitive limitations in remembering phone numbers. Similarly, mutual funds and a host of investor advisory services allow investors to sidestep their personal biases. Such entrepreneurs can create new problems but, as noted below, those problems also can be studied in the laboratory. Arbitrage is the most direct form of such entrepreneurship. If error-prone investors create an asset price discrepancy, this will attract profit-seeking arbitrageurs whose buy and sell orders tend to make it disappear. Laboratory studies, including those of Plott and Sunder (1982), confirm the power of arbitrage.

Amplifying biases and errors

There are also three strong forces that can amplify the market impact of errant investors. First, raw information is often gathered, analysed and released by individuals who have major personal stakes in the market reaction. Despite oversight by authorities such as the US Securities and Exchange Commission, these individuals may use their discretion to distort the market reaction. Bloomfield and O'Hara (1999) and subsequent laboratory studies confirm the possibility.

Second, professional fund managers typically are compensated (directly or indirectly, via competing job offers) for returns that rank highly relative to their peers. It is difficult to infer from field data whether such incentives have an impact, but inference is straightforward in the laboratory. James and Isaac (2000) find major distortions of laboratory asset prices when traders have rank-based performance incentives, and the distortions disappear in otherwise identical markets when traders are paid only their own realized returns.

Third, and most intriguingly, investors may go astray when they try to glean information from the trades of informed investors. Information mirages (for example, Camerer and Weigelt, 1991) can arise as follows. Uninformed trader A observes trader B attempting to buy (due to some slight cognitive bias, say) and mistakenly infers that B has favorable inside information. Then A tries to buy. Now trader C infers that A (or B) is an insider and tries to mimic their trades. Other traders follow, creating a price bubble.

Several research teams (including the author's) have occasionally observed such episodes in the laboratory. They cannot be produced consistently, because incurred losses teach traders to be cautious when they suspect the presence of better-informed traders. The lesson does not necessarily improve market efficiency, since excessive caution impedes information aggregation.

Price bubbles deserve longer discussion, as bubbles have produced important distortions in market prices. Asset prices seemed to disconnect from fundamental value in Japan in the late 1980s, in the dot.com bubble and crash of 1997–2002, and in a number of other episodes since the famous 17th and 18th century events now known as tulipmania and the South Sea bubble. Do such episodes indicate dysfunctional financial markets? Perhaps, but the field data also can be interpreted merely as unusual movements in fundamental value (Garber, 1989). By contrast, in the laboratory the experimenter can always observe (or more typically, control) the fundamental value, so bubbles can be detected and measured precisely.

Smith, Suchanek and Williams (1988) found large positive bubbles, and subsequent crashes, for long-lived laboratory assets and inexperienced traders. Figure 1 shows a representative example. The expected dividend is constant, so the fundamental value (the sum of expected remaining dividends) declines steadily over the 15 trading periods. Ask (‘offer’) and bid prices start low, but by the second period the transaction prices (indicated by lines connecting accepted bids and asks) rise above fundamental value. The bubble inflates rapidly until late in period 4. In period 9, prices crash below fundamental value.

Figure 1

A bubble and crash in the laboratory. *Source*: Smith, Suchanek and Williams (1988, Figure 9).

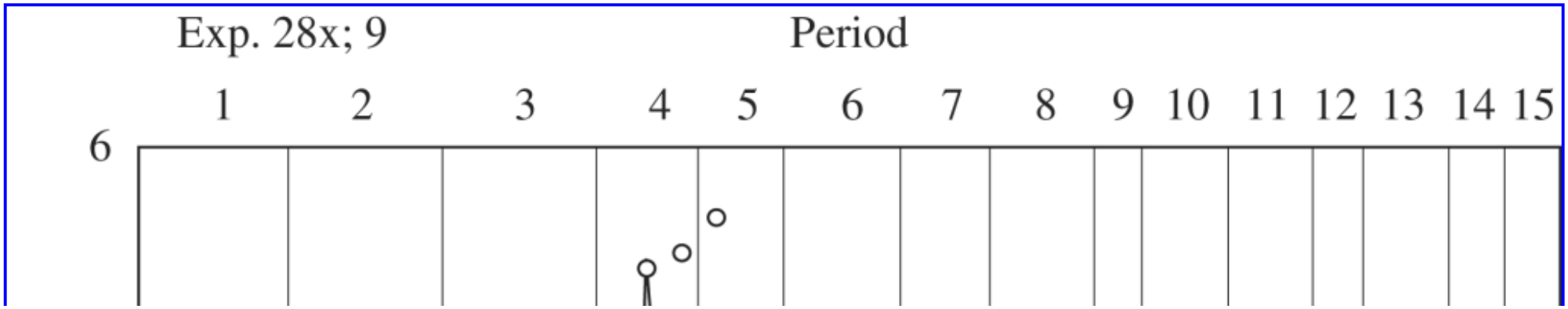

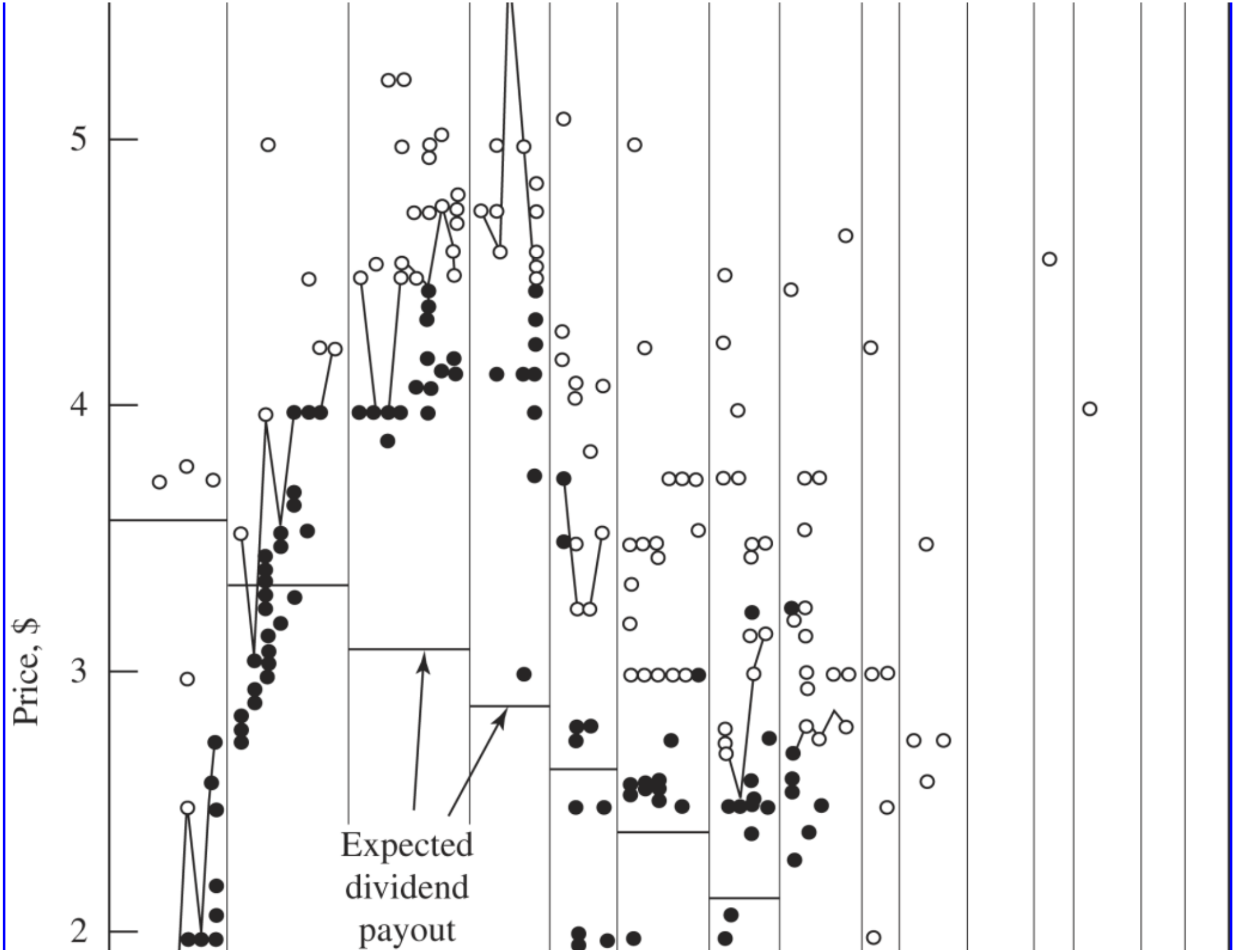
Price, $
5
4
3
2
Expected
dividend
payout

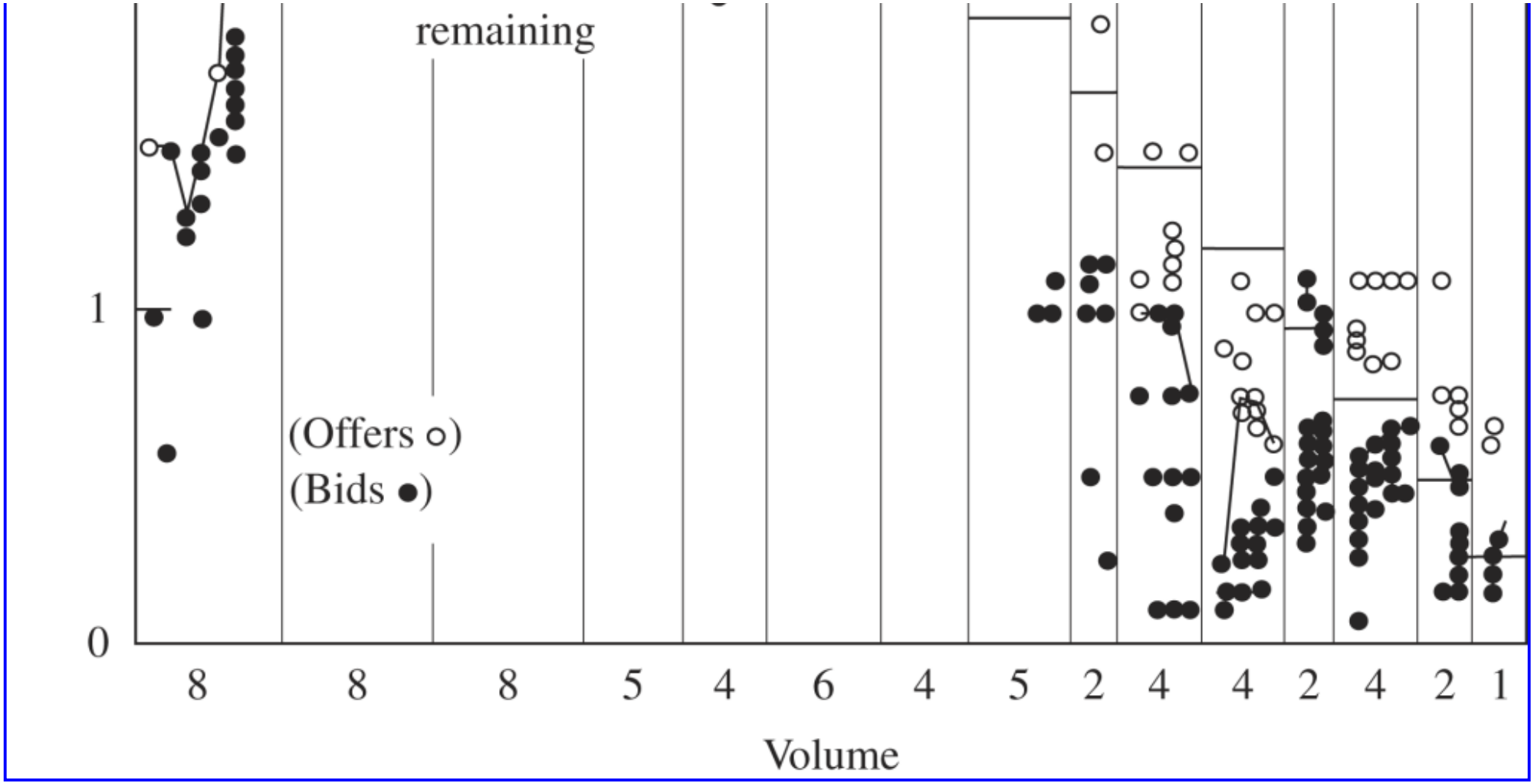

Keynes's ‘greater fool’ theory provides a possible interpretation. Traders who themselves have no cognitive bias might be willing to buy at a price above fundamental value because they expect to sell later at even higher prices to other traders dazzled by rising prices. Subsequent studies confirm that such dazzled traders do exist, and that bubbles are more prevalent when traders are less experienced (individually and as a group), have larger cash endowments, and have less conclusive information.

Current frontiers: market formats, agents, and prediction markets

Which underlying biases and errors are most important? When does attenuation predominate, and when does amplification? Accumulating laboratory evidence inspires new theoretical and empirical field work as well as follow-up laboratory studies.

It is increasingly clear that answers hinge on the market format or institution – the rules that transform bids and asks into transactions. In particular, the CDA format allows all traders to observe other traders’ attempts to buy and sell in real time, and thereby encourages information dissemination. The CDA format attenuates the impact of erratic traders because the closing price is not set by the most biased trader or even by a random trader. The most optimistic traders buy (or already hold) and the most pessimistic traders sell (or never held) the asset, so the closing price reflects the moderate expectations of marginal traders (see Smith, Vernon).

Other traditional formats include the call market (CM), in which bids and asks (or limit orders) are gathered and executed simultaneously at a uniform price, and the posted offer (PO), in which one side (usually sellers) simultaneously announces prices and the other side (buyers) choose transaction quantities at the given prices. Many other formats and hybrids are possible in the Internet age. Which formats are most efficient? Which can attract market share from other formats? Work so far indicates that the CM format does relatively well

for thinly traded assets and the PO format works best when the posting side is more concentrated; but the questions remain far from settled.

Related new work blurs the line between computer simulations and laboratory markets. Computer algorithms for artificial agents, or bots, incorporate specified cognitive limitations, and simulations examine the market level impact (for example, Arthur et al., 1997). Gode and Sunder (1993) showed that simple perishables CDA markets are quite efficient even when populated by zero intelligence (ZI) agents, bots that are constrained not to take losses but are otherwise quite random. Current work puts ZI and more intelligent bots into the same asset markets as human traders, and compares efficiency and the distribution of surplus. Such work should help inform regulators, reformers, and entrepreneurs creating new asset markets. Early published examples of policy-oriented research includes performance assessment of (*a*) trader privileges such as price posting and access to order flow information (for example, Friedman, 1993), and (*b*) transaction taxes, price change limits and trading suspensions intended (typically ineffectively) to mitigate price bubbles and panics (for example, Coursey and Dyl, 1990).

Prediction markets, which use the information-aggregation property of markets to forecast events such as election outcomes, are gaining increased attention. The Iowa Electronic Market, designed and operated by experimental economists (Berg et al., 2008), offers various assets that pay the holder ten dollars if (and only if) a specified event occurs by a specified date. Participants self-select, are not representative of the general public, and their trades exhibit partisan bias – for example, self-styled Democrats are more likely to buy assets that pay off when the Democratic Party candidates win. Nevertheless, political event asset prices have consistently outperformed opinion polls and all other available predictors. Prediction markets are a growing presence on the Internet, for example tradesports.com, and some corporations such as HP are beginning to rely on them when making business decisions. The line between laboratory and field financial markets is beginning to blur.

See Also

- behavioural finance
- Smith, Vernon

Bibliography

Arthur, W.B., Holland, J.H., LeBaron, B., Palmer, R. and Taylor, P. 1997. Asset pricing under endogenous expectations in an artificial stock market. In *The Economy as an Evolving Complex System II*, ed. W.B. Arthur, S.N. Durlauf and D.A. Lane. Reading, MA: Addison-Wesley.

Berg, J., Forsythe, R., Nelson, F. and Rietz, T. 2008. Results from a dozen years of election futures markets research. In *Handbook of Experimental Economics Results*, ed. C. Plott and V. Smith. Amsterdam: North-Holland (forthcoming).

Bloomfield, R.J. and O'Hara, M. 1999. Market transparency: who wins and who loses? *Review of Financial Studies* 12, 5–35.

Blume, L. and Easley, K. 1992. Evolution and market behavior. *Journal of Economic Theory* 58, 9–40.

Camerer, C. and Weigelt, K. 1991. Information mirages in experimental asset markets. *The Journal of Business* 64, 463–93.

Coursey, D.L. and Dyl, E.A. 1990. Price limits, trading suspension, and the adjustment of prices to new information. *Review of Futures Markets* 9, 343–60.

Friedman, D. 1993. How trading institutions affect financial market performance: some laboratory evidence. *Economic Inquiry* 31, 410–35.

Friedman, D. 1998. Monty Hall's three doors: construction and deconstruction of a choice anomaly. *American Economic Review* 88, 933–46.

Garber, P.M. 1989. Tulipmania. *Journal of Political Economy* 97, 535–60.

Gode, D.K. and Sunder, S. 1993. Allocative efficiency of markets with zero intelligence traders: market as a partial substitute for individual rationality. *Journal of Political Economy* 101, 119–37.

Hayek, F.A. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

Holt, C.A. 1999. Y2•K bibliography of experimental economics and social science asset market experiments. Online. Available at http://people.virginia.edu/~cah2k/assety2k.htm, accessed 19 February 2007.

James, D. and Isaac, R.M. 2000. Asset markets: how they are affected by tournament incentives for individuals. *American Economic Review* 90, 995–1004.

Plott, C.R. and Sunder, S. 1982. Efficiency of experimental security markets with insider information: an application of rational-expectations models. *Journal of Political Economy* 90, 663–98.

Smith, V.L. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70, 111–37.

Smith, V.L., Suchanek, G.L. and Williams, A.W. 1988. Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica* 56, 1119–51.

Sunder, S. 1995. Experimental asset markets: a survey. In *The Handbook of Experimental Economics*, ed. J.H. Kagel and A.E. Roth. Princeton, NJ: Princeton University Press.

How to cite this article

Friedman, Daniel. "laboratory financial markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000238> doi:10.1057/9780230226203.0911

The New Palgrave Dictionary of Economics Online

labour discipline

Peter Hans Matthews
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Although economists in different fields, or from different schools, use different words to describe the phenomenon, there is widespread agreement that workers can, and sometimes do, ‘contest’ the sale of their labour power to employers. The question of how employers maintain ‘labour discipline’ in such an environment has intrigued economists since at least Marx's time.

Keywords

asymmetric information; capitalism; contested transactions; effort extraction problem; employment rent; Hobbes, T.; incentive compatibility; Kalecki, M.; labour discipline; labour market contracts; loyalty; Marx, K.; supervision; unemployment

Article

Because it is difficult to write and enforce complete contracts in labour markets, transactions are often ‘contested’ (Bowles and Gintis, 1993) and labour discipline must somehow be enforced.

Recent formalizations of the ‘effort extraction problem’, for example, are premised on the notion that it is difficult for firms to monitor the effort levels of all workers at all times. How much effort workers expend will then depend on, among other things, the cost of job loss. It follows that, as the unemployment rate or, to be more precise, the expected duration of unemployment decreases, the wage at which workers will expend a particular effort level will increase. In many such models, the ‘employment rent’ consistent with near-full employment is not feasible, and it is equilibrium unemployment that ‘solves’ the labour discipline problem.

To fix ideas, consider a discrete time variant of the influential Shapiro and Stiglitz (1984) model. There are *N* identical, infinite-lived and risk-neutral workers, each of whom maximizes the expected value of $\sum_{i=0}^{\infty} \theta^i u(w_i, e_i)$, where:

$$u(w_i, e_i) = \begin{cases} w_i - e_i & \text{if the worker is employed in period } i \\ \bar{w} & \text{if the worker is unemployed in period } i \end{cases}$$

and where w_i and e_i are the real wage and effort level in period i, θ is the common rate of time preference and $\bar{w}$ is an unemployment benefit, financed, for the sake of convenience, with a lump-sum tax on profits. Workers must choose one of two effort levels, 0 or $\bar{e}$, each period, and there is some likelihood d that a worker who expends no effort in a particular period will be detected and then dismissed. Furthermore, at the end of each period a fraction q of all employed workers enters the jobless pool for other reasons.

In a stationary equilibrium, the lifetime utility, V_1, of an employed worker who expends $\bar{e}$ each period will be:

$$V_1 = \frac{w - \bar{e} + q\theta V_3}{1 - \theta(1 - q)}$$

where V_3 is the lifetime utility of a worker who is currently unemployed. (The worker receives $w - \bar{e} + \theta V_3$ and $w - \bar{e} + \theta V_1$ with likelihoods q and $1 - q$, respectively, which implies that $V_1 = q(w - \bar{e} + \theta V_3) + (1 - q)(w - \bar{e} + \theta V_1)$.) In a similar vein, the lifetime utility, V_2, of an employed worker who expends no effort each period will be:

$$V_2 = \frac{w + (d + q(1 - d))\theta V_3}{1 - \theta(1 - q)(1 - d)}$$

Workers will therefore not expend effort $\bar{e}$ unless $V_1 \geq V_2$ or, after substitution and simplification:

$$w \geq \left(\frac{1 - \theta(1 - q)(1 - d)}{\theta(1 - q)d}\right)\bar{e} + (1 - \theta)V_3 \qquad (1)$$

Consistent with intuition, firms will find it more expensive to achieve labour discipline (that is, the

incentive-compatible wage will be higher) the costlier effort is to workers, whether this is because the required effort level $\bar{e}$ has increased or the disutility of such effort has. Discipline will also be more expensive when either the likelihood of detection *d* or the discount rate θ is lower. When, for example, workers care less about the future, the prospect of eventual dismissal will be less salient. An increase in the separation rate *q* also causes the threshold in (1) to rise: as labour markets become more turbulent, workers have less incentive, *ceteris paribus*, to invest in a particular employment relationship.

To understand the full implications of (1), however, the lifetime utility of unemployed workers must be further decomposed. If *a* is the fraction of the jobless pool that is (re)hired at the start of each period in equilibrium, the value of V_3 will be:

$$V_3 = \frac{(1-a)\bar{w} + aV_1}{1-\theta(1-a)}$$

when employed workers find it in their interest to expend effort. It is then tedious but not difficult to show that (1) can be written:

$$w \geq \bar{w} + \left(\frac{1-\theta(1-a)(1-q)(1-d)}{\theta(1-a)(1-q)d}\right)\bar{e} \qquad (2)$$

In a provocative choice of words, Shapiro and Stiglitz (1984) called this now familiar incentive constraint the ‘no shirking condition’. As the likelihood of rehire *a* tends toward 1, labour discipline becomes impossible to achieve because the incentive-compatible real wage increases without limit. In more intuitive terms, workers are certain to ‘contest the exchange’ if the expected duration of unemployment, in this case $\frac{1-a}{a}$, and therefore the punishment value of dismissal, are small.

This model and the dozens, perhaps hundreds, of subsequent variations are sometimes viewed as mainstream restatements of the radical position that persistent joblessness is a characteristic feature of capitalism. In Volume I of *Capital*, for example, Karl Marx (1867, p. 701) saw the ‘industrial reserve army of the unemployed’ as a ‘condition of existence of the capitalist mode of production’, one which ‘[held the] pretentions of the active labor army in check’ in ‘periods of over-production and paroxysm’. Writing almost 80 years later, at the dawn of the Keynesian Revolution, Michal Kalecki (1943, p. 326) would claim that capitalists were ‘consistently opposed to creating employment by subsidizing consumption’, even if meant a reduction in profits, so that ‘discipline in the factories’ could be preserved. The similarities should not be overstated, however. For Bowles (1985), for example, the difference between ‘Marxian’ and ‘neo-Hobbesian’ models is the difference between those in which the nature of capitalist production is central and those in which simple ‘malfeasance’ is the issue. Furthermore, while

there is no doubt that Marx believed that the reserve army served to constrain the demands of workers, its existence owes more to the dynamics of accumulation and technological change than to asymmetric information. And, unlike Shapiro and Stiglitz, or for that matter Marx, Kalecki believed the impediments to full employment were largely political, not economic.

The enforcement of labour discipline involves more than reserve armies, however. Levine (1989), for example, extends the Shapiro–Stiglitz model to show that, when firms cannot be sure that low output is the result of low effort, dismissal policies will violate the just-cause principle, and that the (forced) adoption of this principle leads to more efficient outcomes. In other contributions to the literature, enforcement is more subtle. The slope of the representative wage-tenure profile, for example, which some labour economists believe is too steep to be explained in terms of human capital accumulation alone, could also reflect firms' pursuit of labour discipline: in this case, deferred compensation mimics the properties of a performance bond, and so increases the cost of job loss for recently hired workers.

The substantial variation in the ratio of supervisory to production workers across otherwise similar economies (and over time, for that matter) hints that, in practice, firms can influence the likelihood of detection or, in broader terms, decide how much, and in what form, workers will be monitored. Furthermore, there is reason to believe that, from an efficiency standpoint, firms will spend too much on supervision: if the size of the employment rent were increased at the expense of supervision, the same output could be produced with fewer inputs.

Both the choice of technique and the search for new methods of production influence, and are influenced by, the enforcement of discipline. In some cases, the most salient characteristic of a particular innovation is its effect on effort extraction. As the historian E. P. Thompson (1967) reminds us, for example, the spread of reliable mechanical clocks in production more than two centuries ago represented a watershed in the evolution of enforcement mechanisms, in much the same sense, perhaps, that computerization has, whatever its other effects, forever altered the power to monitor.

Braverman (1974) and others follow this line even further, arguing, in effect, that the widespread adoption of methods of mass production – in particular, the routinization of labour – owed much to how these methods simplified the extraction of effort and reduced replacement costs for dismissed workers. Even if mainstream economists are sceptical, few doubt that the 'rise of the factory' involved 'substantial investment in fixed capital with strict supervision and rigid discipline' (Mokyr, 2002, p. 2).

Finally, recent advances in behavioural and experimental economics have revitalized interest in 'bureaucratic control' (Edwards, 1977) of the workplace, in which the means to achieve labour discipline are often more subtle. There is considerable experimental evidence, for example, to support the view that workers and firms sometimes exchange 'gifts' of effort and wages, and that this relationship is 'socially embedded' (Gachter and Fehr, 2002), one consequence of which is that intrinsic motivation (a sense of loyalty, for example) can also contribute to labour discipline.

See Also

- Kalecki, Michal
- labour market institutions
- Marx, Karl Heinrich
- moral hazard

- underemployment equilibria

Bibliography

Bowles, S. 1985. The production process in a competitive economy: Walrasian, neo-Hobbesian and Marxian models. *American Economic Review* 75, 16–36.

Bowles, S. and Gintis, H. 1993. The revenge of homo economicus: contested exchange and the revival of political economy. *Journal of Economic Perspectives* 7(1), 83–102.

Braverman, H. 1974. *Labor and Monopoly Capital*. New York: Monthly Review Press.

Edwards, R. 1977. *Contested Terrain*. New York: Basic Books.

Gachter, S. and Fehr, E. 2002. Fairness in labor markets: a survey of experimental results. In *Surveys in Experimental Economics*, ed. F. Bolle and M. Lehmann-Waffenschmidt. Heidelberg: Physica-Verlag.

Kalecki, M. 1943. The political aspects of full employment. *Political Quarterly* 14, 322–31.

Levine, D. 1989. Just-cause employment policies when unemployment is a worker discipline device. *American Economic Review* 79, 902–5.

Marx, K. 1867. *Capital: A Critique of Political Economy*. New York: The Modern Library, 1906.

Mokyr, J. 2002. The rise and fall of the factory system: technology, firms and households since the Industrial Revolution. *Carnegie-Rochester Conference Series on Public Policy* 55, 1–55.

Shapiro, C. and Stiglitz, J. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Thompson, E. 1967. Time, work-discipline and industrial capitalism. *Past and Present* 38, 56–97.

How to cite this article

Matthews, Peter Hans. "labour discipline." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000231> doi:10.1057/9780230226203.0912

The New Palgrave Dictionary of Economics Online

labour economics (new perspectives)

Christopher Taber and Bruce A. Weinberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since Richard Freeman wrote labour economics for the first (1987) edition of *The New Palgrave: A Dictionary of Economics*, labour economics has become increasingly empirical, with less emphasis on theory. The most noticeable change in empirical work is an increased emphasis on the plausibility of identification assumptions such as the validity of instrumental variables. Among the areas growing or receiving the greatest attention are changes in the wage structure, the economics of education, social interactions and personnel economics. The range of topics studied by labour economists today has broadened far beyond those of traditional labour economics.

Keywords

education production functions; fixed effects; group selection; human capital; identification; instrumental variables; labour economics; labour market search; matching; natural experiments; personnel economics; returns to schooling; Roy model; sample selection problem; skill-biased technical change; wage differentials; wage inequality, changes in

Article

When Richard Freeman wrote his excellent article labour economics for the first (1987) edition of *The New Palgrave: A Dictionary of Economics*, which is reproduced in the present edition, labour economics had changed dramatically with the development of the human capital paradigm and the use of large-scale data-sets. In many ways labour economics has continued along the trends Freeman discussed, but in other important ways its focus has shifted, in terms of both topics and interests. The goal of this article is to describe major trends in this dynamic field of applied microeconomics since the 1980s. We begin with an overview of methodological trends that are common to much of the field and then talk about specific research questions within labour economics. We will direct readers to the appropriate *New Palgrave* articles for a more complete discussion of those topics.

One important way in which labour economics has changed since Freeman's article is that it has become increasingly empirical. Presumably, this trend is due at least in part to improvements in large-scale computing and ease of access to data sources. Along with this trend has come a decreased emphasis on theory in all but a few areas. The decreased emphasis on institutional factors, discussed in labour economics, has certainly continued (even the study of labour unions has declined substantially). Along with the trend towards increased empirical work has come a much stronger emphasis on the plausibility of identification assumptions. In many labour contexts, there are substantial unexplained variation in the dependent variables being studied, leading to interest in strategies for dealing with sample selection and endogeneity. In the case of earnings regressions, for instance, the vast majority of the variation in earnings cannot be explained by observable worker characteristics. While the presence of important unmeasured factors does not invalidate a regression model, it raises a concern that the coefficients on the variables of interest may be biased if the substantial unexplained component is correlated with the variables of interest. In the earnings regression case, one worries that workers who are more able or motivated (in ways that are unmeasured by the analyst) may obtain more school, biasing estimates of the return to school upwards. Labour economists have increasingly focused on these selection or endogeneity issues and this emphasis has spilled over from labour economics into other fields in economics.

The most noticeable change in approach is much greater emphasis on the plausibility of identification strategies. The two most noticeable examples in this vein are an increased use of fixed effect approaches (including ‘difference in differences’) and much more attention being paid to the validity of instrumental variables. A classic example is Angrist's (1990) study of Vietnam veterans. Estimating the effect of veteran status on earnings is plagued by the classic sample selection problem. Angrist solves this problem by using the Vietnam draft lottery number as an instrument for veteran status. This number is mechanically related to veteran status, but since it is random it will be unrelated to earnings again by construction. Studies along these lines are typically referred to as experimental or natural-experiment studies (depending on whether the variation arises from an explicit randomized experiment or policy or institutional factors that are plausibly, but not explicitly, random).

Structural estimation has also received substantial attention since the 1980s, although in relative terms, substantially less than during the previous 20 years. Different people may define structural in different ways. For example, simple linear models estimated by ordinary least squares or two-stage least squares can be considered structural if the researcher is explicit about the interpretation of the parameters. We have witnessed a large increase in popularity of a more ambitious approach in which a researcher formally models an individual's decision process and estimates the underlying parameters of say the utility function or production process by choosing the parameters that minimize the difference between observed outcomes and those implied by the model. For instance, a young individual has the option to attend school, or work in a variety of jobs, or remain in the household sector in each year of his or her life. One structural approach would be to estimate an individual's value function from the terminal period backwards at each node on the decision tree by matching observed behaviours to those implied by utility maximization. Unobserved individual factors can be addressed by including them in the value functions and integrating them out when trying to match the data. This approach is computationally demanding. Substantial advances in computational methods for these models and improvements in computer technology have allowed researchers to estimate considerably richer models. This approach has benefited from the year-by-year extension of longitudinal (panel) data sets.

The literature on returns to schooling provides a nice example of the evolution of empirical approaches. The goal of this literature is to estimate the causal effect of schooling. Willis and Rosen (1979) is a classic paper in this literature and an excellent illustration of empirical approaches prior to 1987. These authors consider a model with two schooling choices, high school and college, in which students make decisions to maximize the present value of earnings. They allow for individual heterogeneity in college and high-school earnings, college and high-school earnings growth, and interest rates. Their empirical approach consists of a three-stage method in which the first stage is a reduced form probit for college attendance. The second stage is a series of wage regressions including inverse Mills ratios. The third is a ‘structural probit’ that allows one to estimate the effects of earnings on schooling choices. The key for semiparametric identification in models like this is a variable that affects schooling choices but does not affect earnings directly (see Roy model for discussion of semiparametric identification in this type of model). Willis and Rosen use family background as their exclusion restriction. Family background is relatively strongly related to schooling and might not directly affect earnings. However, subsequent researchers have been sceptical about this exclusion restriction. The biggest concern in using regression analysis is that schooling is probably related to unobservable ability, but for similar reasons one may expect family background to be related to unobserved ability. Either through genetics, parenting skills, or simply resources one might worry that children from privileged backgrounds have more unobserved ability than their less fortunate peers.

Since 1990 or so many papers have tried to develop more credible exclusion restrictions to estimate the return to schooling. One of the most well known is Angrist and Krueger (1991), who use quarter of birth as an instrument. They argue that a combination of truancy laws and school starting ages will lead students born late in a calendar year to obtain more education than a student born early in a year. To see why, suppose that the cut-off date for starting school is 1 January. As a result, an individual born on 31 December 1962 will begin school a year earlier than a student born a day later, on 1 January 1963. However, if both of these students drop out of high school as soon as the truancy law says that they can, say on their 16th birthday, then the student born in December will have attained an extra year of schooling. Unfortunately, data-sets are not sufficiently large to focus only on these two days, so Angrist and Krueger (1991) use quarter of birth instead. Furthermore, there is a fair amount of slippage in that neither truancy laws nor age cut-off dates are strictly adhered to. Note that this last feature does not invalidate the instrument but reduces its power. As an example of a fixed effect approach, Ashenfelter and Krueger (1994) collect data on both earnings and educational attainment of twins. By using family fixed effects they can obtain an estimate of the returns to schooling, differencing out genetic ability.

At the same time we have seen a large structural literature emerge that has generalized Willis and Rosen (1979) by allowing for more complex educational choices and selection. A classic example of this approach is Keane and Wolpin (1994), who estimate a dynamic model of labour-market decisions. They generalize Willis and Rosen (1979) by allowing for many more than two schooling choices (high school versus college), by allowing students to go back and forth from the labour market and school, and by allowing the payoff to schooling to be sector specific. Another example is Heckman, Lochner and Taber (1998a), which estimates a general equilibrium version of the Willis and Rosen (1979) model that estimates not just the pricing equation for schooling but the determinants of both the supply and demand for college. This additional structure included in these papers allows one to simulate substantially more

complicated policy experiments than one can perform with the Willis and Rosen (1979) framework. There are substantial disagreements over the relative merits of different empirical approaches, and a full discussion goes well beyond the scope of this article. With that in mind, some of the instrumental variable and difference in differences approaches have the benefit of placing identification and the source of identification at the forefront of the analysis. For example, the source of identification in the Angrist and Krueger (1991) case is transparent. Estimation of intricate structural models typically requires substantial assumptions, whose validity is frequently unclear, but because the underlying parameters of the problem (preferences, the technology and so on) can be estimated, it is frequently possible to evaluate a wide range of policies that are not represented in the data. For example, identification in Keane and Wolpin (1994) is much less transparent. While it may appear that the reduced-form, natural experiment approach requires fewer or weaker assumptions, work of this type usually implicitly makes a number of important assumptions, particularly if one wants to apply these results to some other context. For example, Heckman, Lochner and Taber (1998b) demonstrate that one can severely underestimate policy effects if one ignores general equilibrium (GE) effects in their model. When researchers ignore GE effects in drawing policy predictions from their work, they implicitly assume that the demand for educated workers is perfectly elastic. The work of Heckman, Lochner and Taber (1998b) suggests that this is a very strong assumption. It seems likely that a large variety of approaches will continue to be used and that results that are robust across a wide range of approaches will be most convincing.

As indicated, labour economics has become increasingly empirical as emphasis on identification has increased. Personnel economics the study of incentives within firms, is a notable exception. This literature is discussed in more detail in personnel economics. Another exception is work on search and matching, which spans labour economics and macroeconomics and which tends to be more theoretical. Because it requires explicit statements of the decision problem, structural work tends to be more theoretical than reduced form work, although deriving explicit theoretical results is rarely the focus of such studies.

Another notable recent development in labour economics is that the scope of problems that labour economists address has broadened considerably since the 1980s. However one wants to define ‘labour economics’ – as the study of the determinants of individual earnings, the demand and supply for labour, and the functioning of labour markets or as whatever labour economists do – what is noteworthy is that much of the work being done by labour economists falls well outside a traditional definition of the field. Similarly, other fields have increasingly drawn on ideas developed in labour economics, and the lines between labour economics and closely related fields, including development, urban, and public economics, are blurring. To some extent this reflects the general applicability of traditional labour theory (for example, human capital, which has played an important role in growth economics) and to some extent it reflects the widespread applicability of the econometric techniques developed by labour economists.

One of the most influential areas in labour economics since the 1970s has been the changes in the wage structure (see wage inequality, changes in). Much of this work focuses on the increase in inequality and the increase in the returns to education in the United States. This literature has emphasized demand-side factors, and skill-biased technological change in particular, as the primary explanation for the recent trends. Most recent work on the demand-side of labour economics has been in this area. Whether it is a

result of this literature, or coincidental, the whole field has shifted towards trying to understand wage differentials and human capital accumulation.

Another increasingly active area in labour economics is the economics of education, which perhaps can be considered its own field rather than a subfield of labour. The increased interest in education probably has arisen both from the literature on the changing wage structure, which emphasizes human capital, and from the increasing attention to education in the policy world. Within labour economics, understanding the economic value of education is one of the most studied empirical questions (see returns to schooling for a summary of this literature). Economists have also moved from wanting a general understanding of the effects of education on wages to a more specific understanding of what aspects of school are most important in forming human capital. Specifically, researchers have tried to uncover these factors in the ‘education production function’ literature discussed in education production functions.

We have also seen increased research on private schools and school choice (see school choice and competition for a discussion of this literature). Another branch of this literature (which is really much more of a subfield of public economics rather than labour economics) studies the complicated system under which schools are financed and how changes in these schemes influence students (see educational finance for a description of this literature).

Empirically, race and gender are also important determinants of wages. There is a long literature in labour economics on the economics of discrimination, which tries to understand why these differences arise. The two most studied effects have been the male/female and black/white gap in the United States. While the raw log wage differentials are of similar magnitude (approximately 20 per cent), the effects are very different from each other. As more controls are included in the analysis the black–white gap declines substantially; see black–white labour market inequality in the United States. This has led researchers to focus on pre-market forces as the primary cause. By contrast, men and women look much more similar when they enter the labour market. Thus the difference seems to be related to post-market entry factors. women's work and wages discusses this literature.

The traditional field of labour supply has probably received less attention since the last Palgrave than in the decades preceding it. Much of the work on this subject has focused on the lower end of the earnings distribution. Perhaps most importantly, a large literature has arisen that attempts to measure the effects of transfer programmes on labour supply of low-income individuals, especially on single mothers.

Another important policy area has focused on understanding the effects of minimum wages on employment. Related to the literature on the changing wage structure, there has also been a substantial literature studying labour-force participation among low-skilled workers who are likely to be close to the margin to work and whose wages have fallen considerably (in the case of the United States). While it is well known that labour-force participation among women has increased substantially, participation among men has declined (see for example, Juhn, Murphy and Topel, 2002). This literature is discussed more thoroughly in labour supply.

Drawing on research in sociology, labour economists have also become increasingly interested in how people are affected by the groups (for example, schools or neighbourhoods) to which they belong (see social interactions (empirics) and social interactions (theory)). Such studies span a number of the topics already discussed – how students’ educational outcomes (and other behaviours such as substance use) depend on those of their peers; or how labour market activity (for example, employment or welfare participation) depends on that of neighbours. Naive estimates indicate that people's behaviours and

outcomes are highly correlated with those of their groups, but researchers have been concerned that the groups that people choose (or are ‘forced’ into) are similar to themselves. A substantial literature has developed using quasi-experiments and explicit experiments to estimate the effect of groups on the people who are in them controlling for the selection processes into groups. Estimates that control for the selection processes are considerably lower than those that do not.

While individual characteristics are very important for wage determination, characteristics of the firm may matter as well. There have been an increasing number of data-sets that allow researchers panels on both firms and workers. These types of data-set allow one to use procedures such as estimating both firm and worker fixed effects (see for example, Abowd, Kramarz and Margolis, 1999). These papers show that firm effects are an important component of wages. The most obvious explanation for this type of result is that there is some type of friction in the labour market. Perhaps as a result there has been an increased interest in labour market friction and its importance in explaining inequality: see labour market search for a discussion of this search literature and matching for a discussion of the matching literature.

There has also been increased attention on the economics of the household, which lies at the intersection of labour economics and other fields such as demography. This work includes studies of bargaining between members of the household on intra-household resource allocations and the effect of household behaviours on children's human capital. Related to these topics are marriage and fertility decisions and household labour supply.

In recent years, the theoretical concepts and empirical methods of labour economics have proven useful across a wide range of topics. Consequently, labour economics has influenced work in a wide variety of other areas and the topics studied by labour economists have expanded considerably.

See Also

- labour economics
- personnel economics
- returns to schooling
- Roy model
- wage inequality, changes in

Bibliography

Abowd, J., Kramarz, H.F. and Margolis, D. 1999. High wage workers and high wage firms. *Econometrica* 67, 251–334.

Angrist, J.D. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review* 80, 313–35.

Angrist, J. and Krueger, A. 1991. Does compulsory schooling attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979–1014.

Ashenfelter, O. and Krueger, A. 1994. Estimates of the economic return to schooling from a new sample twins. *American Economic Review* 84, 1157–73.

Heckman, J., Lochner, L. and Taber, C. 1998a. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics* 1, 1–58.

Heckman, J., Lochner, L. and Taber, C. 1998b. General equilibrium treatment effects. *American Economic Review* 88, 381–6.

Juhn, C., Murphy, K. and Topel, R. 2002. Current unemployment historically contemplated. *Brookings Papers on Economic Activity* 2002(1), 79–136.

Keane, M.P. and Wolpin, K.I. 1994. The career decisions of young men. *Journal of Political Economy* 105, 473–522.

Willis, R.J. and Rosen, S. 1979. Education and self selection. *Journal of Political Economy* 87, S7–S36.

How to cite this article

Taber, Christopher and Bruce A. Weinberg. "labour economics (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000241> doi:10.1057/9780230226203.0914

The New Palgrave Dictionary of Economics Online

labour economics

Richard B. Freeman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

affirmative action; Becker, G.; comparable worth; decentralized wage-setting; dual labour markets; efficient contracts; elasticities of complementarity; elasticity of substitution; forecasting; Harris–Todaro hypothesis; human capital; immigration and the city; internal labour markets; labour demand; labour economics; labour market institutions; labour supply; layoffs; long-term employment; Mincer, J.; minimum wages; negative income tax; period of production; personnel economics; profit sharing; search models of unemployment; trade unions; unemployment; wage differentials; wage dispersion; wage rigidity

Article

Labour economics studies the demand and supply for the most important factor of production, human beings. Since the days of Marshall and indeed of Smith, if not earlier, economists have recognized that one cannot analyse the market for labour, without taking account of such issues as social relations of production, long-term contractual arrangements, problems of effort and motivation, as well as institutions like unions and internal labour markets, which differentiate the labour market from a bourse. For many years recognition of these factors made labour economics an area in which economic theory was applied sparingly and in which institutional analyses dominated.

This is no longer the case. Sparked in part by theoretical advances and in part by the availability of computerized data-sets with observations on hundreds, (thousands, tens of thousands) of individuals, labour economics underwent a dramatic revolution beginning in the 1960s and accelerating thereafter. As a result modern labour economics diverges notably from its past in two respects: creative use of theory to cast light on the aforementioned aspects of reality and detailed empirical investigations of the behaviour of individuals using advanced econometrics. In addition, in contrast to earlier labour economics, which dealt largely with firms' behaviour from a demand perspective, there has been a pronounced interest in labour supply issues in much of the modern work.

Human capital

Conceptually the most important development in the rise of modern labour economics has been the '*human capital*' revolution associated with Gary Becker and Jacob Mincer, among others. Human capital analyses concentrate on individual decision-making, particularly with respect to labour supply and related areas of behaviour often associated with sociology rather than economics. Prior to Becker's *Human Capital*, many labour economists tended to regard labour supply decisions as being only loosely based on economic rationality and therefore as a poor subject area for rigorous theory and analysis. By putting decisions regarding education and other forms of improving skills in an investment framework and developing implications for wages, time worked, and diverse other forms of behaviour, the human capital analysis fundamentally changed the way in which economists see labour supply. The simple investment concept – that individuals, like enterprises, 'invest' early in life (through schooling, and on-the-job-training) and reap rewards later, thereby producing an upward tilt to the age-earnings profile – has proved valuable in interpreting wages, and in directing attention to lifetime considerations in labour supply (for example, use of deferred compensation to motivate workers). Equally important, the view that diverse forms of decision-making can be fruitfully analysed by economic models of rational behaviour has illuminated not only traditional areas of labour supply behaviour such as labour participation, hours worked, job search, career choice, and the like, but has also extended the boundary of analysis to issues ranging from crime to marriage, fertility, and health.

At roughly the same time that human capital theory directed attention at individual behaviour, *computerized data-sets* providing information on the economic and demographic characteristics of individuals became available to analysts. The conjunction of theory and data produced a massive outpouring of studies on the effect of individual as opposed to market or employer factors on wages, and on the supply decisions of individuals. As a result of these factors the labour economist of the 1980s differed substantively in his or her orientation and analytic approach from the labour economist of earlier decades. Whereas in the 1950s labour economists generally studied wages and mobility at the level of industry, area, or in some cases establishments, in the 1970s and 1980s they tended to focus on individuals, first with cross-sectional data comparing different people, then with longitudinal (or panel) data that follow the same person over time. Whereas in the 1950s labour economics was heavily concerned with case studies, in the 1970s and 1980s labour economics had become pre-eminently the field of applied econometrics and statistical analyses of large data types.

In addition to use of modern theoretical and econometric tools, labour economics had been intimately involved in development and analysis of 'controlled experiments' to explore labour supply responses to alternative tax or welfare systems. The most famous of these experiments, the New Jersey and Seattle–Denver experiments, used a control methodology to explore the potential effects of a negative income tax, finding labour supply elasticities that ranged from modest (men) to significant (women) and also uncovering some forms of behaviour relatively hard to explain by standard economics theories (notably in family behaviour). Despite problems with the experimental approach, it marks a striking advance in the set of tools which are employed to explore supply issues.

While there will be some disagreement among economists about the contribution of the human capital and human capital-inspired analysis to explanation of social phenomena, a reasonable assessment is that the analysis has done a good job in illuminating a broad area of social behaviour but at the same time has not explained most of what goes on in the labour market. Changes in behaviour and in structural

relations for reasons of tastes, technology, or whatever, create variation at a point in time and changes over time that are not readily explicable by standard models. For example, in the area of female labour participation, studies find that income effects (reflected in husband's income) and substitution effects (reflected in the wages of the woman) and various indicators of the shadow price of time, such as number of young children, have the sorts of impacts on participation one would expect, but that these factors cannot readily account for the magnitude of upward trends in participation or for cross-country differences in trends or levels. Similarly, while the magnitude and probability of punishment and rates of unemployment and related labour market factors affect crime, they do not account for the high rates of crime in the US relative to other countries not for the time series pattern of change in crime in the US. Even in terms of wage determination, while the variables associated with human capital enter equations with high significance, they are not the dominant factor in variations in wages among individuals: in a typical log-earnings equation, education may explain five per cent of the variation and education and years of experience may explain 15 per cent in total, with job tenure (whose effect is partly the outcome of on-the-job training and partly the result of institutional seniority rules) dominating the experience component; additional important contributors to wage variation include such factors as industry and firm (or establishment) of work that cannot be readily interpreted solely by supply-side factors.

Labour demand

The theoretical and empirical thrust of modern labour economics has had less impact on analyses and understanding of *demand for labour and firms' behaviour* than it has had on the supply of labour. One reason is that previous generations of scholars had devoted considerable effort to analysing the demand side, dealing with such issues as internal labour markets, hiring, promotion, and wage policies, and the structure of wages in various markets, yielding a body evidence on behaviour which has stood up to further analysis. Another reason is that cross-section and longitudinal data on firms and establishments comparable to that on individuals have not been readily available. The computerization of personnel records of firms provides the best potential for major empirical advances in analysis of their labour demand and personnel policy, but as yet work on these records has been rather sparse.

The modern analysis of labour demand has taken the key facts established by the previous generation – that labour markets are far from 'spot markets' – and sought to develop a consistent theory of economic behaviour, in which the firm is viewed as choosing a particular wage and personnel policy to optimize its profits, given the likely response of workers to the policy. Since firms will do best if they offer a labour compensation package that workers desire (at a given cost), some analysts look upon the firm as implicitly maximizing the utility of workers. Others pay greater attention to areas of conflict between the two sides, dealing with issues of shirking, (which makes deferred compensation especially valuable) and effort.

Thus far, the success of this approach has been more on the theoretical than empirical front. Analysts have developed models for such phenomena as deferred compensation, piece rates and related 'prize' systems for rewarding workers, and for such policies as mandatory retirement, but the ability of these 'stories' to account for the bulk of observed variation has not been demonstrated. To take one example, these are unquestionable differences in pay among firms to local labour markets: some firms pay what appear to be 'above-market' rates, while others pay less than the going rate. One can tell efficiency wage stories (firms pay high wages to reduce turnover and shirking); rent-sharing stories (firms share their

economic rents with workers); or union-threat stories (firms pay to keep unions out) about such policies; but labour economics has yet to determine the relative empirical relevance of these stories. In that sense, progress beyond the work of the generation of the 1940s and 1950s that stressed the firms' wage policies has been limited.

Another area of work on demand, more grounded in the neoclassical model of the firm, has examined the magnitude of elasticities and cross-elasticities of labour demand for workers of different skills and the effect of administered wages (minimum wages) on employment. Since the basic parameters in labour demand analysis are elasticities of demand one would hope that empirical work would pin down their magnitude with some certainty. Such has not always been the case. In the US most studies, including those focused on the minimum wage, yield relatively modest elasticities for low-wage workers and manufacturing labour, usually considerably below unity. Analysis of demand for women workers in Australia, exploiting an exogeneous change in female wages due to comparable-worth-type rulings, has also found relatively moderate demand responses. Work on the UK and some European countries, by contrast, has yielded larger estimates of elasticities, which is puzzling given the widespread belief that the United States has a more flexible labour market with employers able to adjust employment more freely than in Europe.

Analyses of elasticities of substitution (which measure the effect of changes in relative wages on changes in relative employment) and of elasticities of complementarity (which measure the effect of changes in relative employment on relative wages) for narrowly defined skill, age, or education groups tend to find higher elasticities, implying that a large exogenous increase in the relative number of persons in a group can significantly affect the relative wages. Two cases in point are the 1970s increased number of young workers ('baby boomers') and of young college graduates in the United States, which greatly reduced the earnings of those groups relative to older and less educated groups.

As a general rule, shifts in demand schedules tend to account for more observed changes in employment than do movements along demand schedules. Work on factors shifting demand for labour (technology, changes in consumer tastes, income elasticities for the goods produced by particular groups of workers) has, however, been rather limited. One body of work has focused on relative demand for minorities, where the development of specific programmes to raise demand provides the same sort of exogeneous shift in the curve as minimum wages provide movements along the curves. The available evidence here suggests that affirmative action and similar programmes have played a role in raising demand for minority labour in the United States, though here as elsewhere changes in the market cannot be solely attributed to one demand-shift factor. Another body of work, associated more with governmental agencies than with academic economists, has projected future labour 'requirements' in an input–output framework.

Comparing the theoretical and empirical work on demand, one is struck by the failure of the empirical analysis to take appropriate account of the potential importance of the long-term employment arrangements and internal labour markets stressed in the theory. A major cause of the difficulty is a data problem: until analysts of labour demand have available detailed longitudinal data on employment by establishment or firm, and on firms' personnel and wage policies, it is exceedingly difficult to marry the advances in theory to the data.

The contrast with the supply side, where theory and data came together, highlights the complementarity of the two 'blades' of the research scissors for a field to develop rapidly.

Institutions

In the area of institutions labour economics has tended to focus on unions as the major worker institution in modern capitalist economies. A massive body of work has examined the effects of unions on wages, beginning first with comparisons of wages in union and non-union sectors of the economy (industries, occupations across cities, and so forth), and then moving on to analysis of the computerized data-sets with information on individuals, classified by union status. While the question that motivates this work is ‘what do unions do to the economy?’ the empirical analyses have, of necessity, been devoted to measuring differences between union and non-union workers (firms).

Following a massive outburst of work on union–non-union wage differentials, labour economists turned to a wide variety of behaviour by individuals and firms likely to be affected by unions. Analysts found quits to be lower and job tenure higher under unionism; temporary layoffs (which occur when workers are laid off for short periods of time, then recalled) to be largely a union sector phenomenon: and the dispersion of wages to be lower in union plants, as well as finding effects of unions on profitability and productivity. This work has paralleled the human capital analysis by continually expanding the set of outcome variables under study and the labour demand analysis by focusing on issues dealt with by the earlier generation of labour economists.

On the theoretical front the thrust of modern work on unions has explored the idea of ‘efficient contracts’ in which unions and management eliminate potential inefficiencies due to monopoly through joint wage and employment determination. Efforts have also been made to develop models of unions as maximizing institutions, following the path laid out by Dunlop in the 1940s, in which unions are concerned with both wages and membership or job security.

Markets

Demand, supply and institutions interact in market settings, and labour economics contains numerous studies of the operation of labour markets for various types of labour. Attention has shifted from markets for blue-collar labour to markets for white-collar labour, and from case studied to more econometric investigations of wage, employment, and unemployment.

One strand of work, closely related to human capital analysis, has been to investigate markets for highly educated workers, where the time period of ‘production’ (college takes four years) allows one to differentiate supply and demand forces in the market. The first generation of such models used relatively simple cobweb structures; a later generation examined more complex rational expectations market-clearing models. The general tone of the results has been sufficiently successful to change the issue from whether markets follow readily understandable economic principles to which type of model best explains patterns of change. Even so, here as elsewhere in economics, the models have not done an especially good job in forecasting, in large because of our inability to project *shifts* in demand schedules, noted earlier.

Another stream of market analysis had dealt with such topics as geographic and industrial mobility, and unemployment and related wage patterns. Observed patterns of wages and mobility make it clear that in the United States decentralized wage setting across a huge geographic area produces separate local labour markets which experience different patterns of change, with costs of mobility sufficiently large as to produce significant ‘losses’ to some (particularly older) displaced workers. An important empirical

finding has been that high-wage cities tend to have high unemployment, providing some support for ‘job search’ as a factor in unemployment. Across industries, the United States evidence shows falling dispersion in wages in periods of economic boom (as low-wage employers raise pay while high-wage employers do not) and also an upward trend in dispersion of wages among industries. Other countries do not appear to have experienced such a trend over time, possibly because of centralized wage setting.

The question of whether unemployment is a long-term or transitory phenomenon has been analysed in the context of models which differentiate between completed and uncompleted spells and between the duration and incidence of unemployment. Perhaps the most important finding, which appears to hold for a large number of countries, is that the bulk of unemployment at any one time is due to a small number of people who are unemployed for long periods, rather than to short-term unemployed people.

Finally, an important area of labour research which diverges substantively from the micro-orientation of much of modern labour economics has involved analysis of macro-change in wages, employment and unemployment over time within a country and across countries. To some extent, labour economics has played a ‘devil's advocate’ role with respect to proposed macro-explanations of problems like unemployment and wage inflation. Macroeconomists have suggested that unemployment is due to such factors as rigid wages associated with three-year contract cycles, intertemporal substitution of time, shocks that require mobility across sectors; labour economists have tested and, in general, rejected these models in a macro-context.

In addition, however, studies suggest that different labour market institutions in different countries may affect macro-outcomes as well. An important hypothesis has been that ‘corporatist’ or centralized free/market economies have an advantage in adjusting to stagflation because all workers can jointly agree to lower rates of increases in wages, avoiding Prisoner's Dilemma problems. Another hypothesis has been that ‘flexibility’ in labour markets is the key to the differential performances of the European and the American economy in employment generation in the 1970s and through the mid-1980s. In the area of theory the notion that ‘a share economy’ (where workers are paid in part via profit or revenue sharing) may produce less unemployment than a ‘wage economy’ has directed attention at alternative modes of paying workers, particularly over the business cycle. Whether comparative analysis focusing on different wage-setting mechanisms across countries becomes a major part of the field, however, remains to be seen.

Another area of comparative labour market studies that proliferated in the 1960s and 1970s focused on labour markets in developing countries. The Harris–Todaro model, which interpreted urban unemployment in terms of migration to cities and queuing for high-wage jobs, directed attention at mobility issues and institutional forces causing ‘dual labour markets’. A variety of studies dealing with the effect of education and human capital on earnings and behaviour revealed patterns similar to those in developed lands, suggesting that some aspects of markets function similarly across levels of development.

Conclusion

In the span of two decades labour economics has moved from a largely institutional field into the mainstream of economics, while maintaining its empirical bent. It has widened the subject of discourse, particularly on the supply side, and struggled to synthesize the ‘facts’ of the labour market with economic principles. It is the interplay of detailed micro data and economic analysis which currently is

the hallmark of the field, differentiating it from more abstract theoretical and less factually based parts of the discipline.

See Also

- human capital
- industrial relations
- labour economics (new perspectives)
- strikes
- women's work and wages

Bibliography

Abraham, K. and Medoff, J. 1980. Experience, performance, and earnings. *Quarterly Journal of Economics* 95, 703–36.

Ashenfelter, O. 1984. Macroeconomic analyses and microeconomic analyses of labour supply. Working Paper No. 1500. Cambridge, MA: NBER.

Ashenfelter, O. and Heckman, J. 1974. The estimation of income and substitution effects in a model of family labour supply. *Econometrica* 42, 73–85.

Ashenfelter, O. and Layard, R., eds. 1984. *Handbook of Labour Economics*. Amsterdam: North-Holland.

Becker, G. 1964. *Human Capital*. New York: Columbia University Press for the NBER.

Becker, G. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Becker, G. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Brown, C., Gilroy, C. and Cohen, A. 1982. The effect of the minimum wage on employment and unemployment: a survey. *Journal of Economic Literature* 20, 487–528.

Bruno, M. and Sachs, J. 1985. *Economics of World wide Stagflation*. Cambridge, MA: Harvard University Press.

Clark, K. and Summers, L. 1979. Labor market dynamics and unemployment: a reconsideration. *Brookings Papers on Economic Activity* 1979(1), 13–72.

Doeringer, P. and Piore, M. 1971. *Internal Labor Markets and Manpower Analysis*. Lexington, MA:

Heath.

Ellwood, D. 1986. The spatial mismatch hypothesis are there teenage jobs missing in the ghetto? In *The Black Youth Job Crisis*, ed. R. Freeman and H. Holzer. Chicago: Chicago University Press.

Farber, H. 1984. The analysis of union behavior. In *Handbook of Labor Economics*. Amsterdam: North-Holland.

Freeman, R. 1971. *The Market for College-Trained Manpower*. Cambridge, MA: Harvard University Press.

Freeman, R. 1983. Crime and unemployment. In *Crime and Public Policy*, ed. J. Wilson. San Francisco: Institute for Contemporary Studies.

Freeman, R. and Medoff, J. 1984. *What Do Unions Do*? New York: Basic Books.

Gregory, R.G. and Duncan, R.C. 1981. Segmented labor market theories and the Australian experience of equal pay for women. *Journal of Post Keynesian Economics* 3, 403–28.

Hall, R. 1975. The rigidity of wages and the persistence of unemployment. *Brookings Papers on Economic Activity* 1975(2), 301–49.

Hamermesh, D. and Grant, J. 1979. Econometric studies of labour–labour substitution and their implications for policy. *Journal of Human Resources* 14, 518–42.

Harris, J.R. and Todaro, M.P. 1970. Migration, unemployment and development: a two sector analysis. *American Economic Review* 60, 126–42.

Hausman, J. and Wise, D. 1985. *Social Experimentation*. Chicago: University of Chicago.

Heckman, J. 1974. Life cycle consumption and labor supply. *American Economic Review* 64, 188–94.

Killingsworth, M. 1983. *Labour Supply*. Cambridge: Cambridge University Press.

Lazear, E. 1979. Why is there mandatory retirement? *Journal of Political Economy* 87, 1261–84.

Leonard, J. 1985. The effectiveness of equal employment law and affirmative action regulation. Working Paper No. 1745. Cambridge, MA: NBER.

Lewis, H.G. 1963. *Unionism and Relative Wages in the United States*. Chicago: University of Chicago Press.

Lewis, H.G. 1986. *Union Relative Wage Effects: A Survey*. Chicago: University of Chicago Press.

Mincer, J. 1962. Labor force participation of married women. In *Aspects of Labor Economics*. J. Mincer. Princeton: Princeton University Press.

Mincer, J. 1968. Labor force participation. In *International Encyclopedia of the Social Sciences*, vol. 8. New York: Macmillan.

Rees, A. 1962. *The Economics of Trade Unions*. Chicago: University of Chicago Press.

Rosen, S. 1984. Distribution of prizes in a match-play tournament with single eliminations. Working Paper No. 1516. Cambridge, MA: NBER.

Rosen, S. 1985. Implicit contracts: a survey. *Journal of Economic Literature* 23, 1144–75.

Segal, M. 1986. Post-institutionalism in labor economics: the forties and fifties revisited. *Industrial Labor Relations Review* 39, 388–403.

US Department of Labor. 1985. Projections of the economy, labor force, industrial and occupational change to 1995. *Monthly Labor Review*, November.

Watts, H. and Rees, A., eds. 1978. *The New Jersey Income Maintenance Experiment*. New York: Academic Press.

Weitzman, M. 1985. *The Share Economy*. Cambridge, MA: Harvard University Press.

How to cite this article

Freeman, Richard B. "labour economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000002> doi:10.1057/9780230226203.0913

The New Palgrave Dictionary of Economics Online

labour market institutions

Richard B. Freeman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Labour market institutions – unions, collective bargaining, government regulations – that help determine wages and working conditions differ greatly across countries. Advanced European countries rely extensively on institutions while the United States relies more on market forces. Labour institutions reduce the dispersion of pay and income inequality but have problematic effects on other aggregate economic outcomes, such as unemployment. The weak or inconclusive link between institutions and outcomes beyond wage dispersion could reflect different institutional effects under different economic conditions; efficient bargaining that balances the adverse and positive effects of institutions on those outcomes, or weaknesses in data and modelling.

Keywords

Coase Theorem; collective bargaining; employment at will; employment protection; Gini coefficient; health insurance; labour market institutions; layoffs; minimum wages; rent seeking; reservation wages; trade union density; trade unions; unemployment; unemployment insurance; wage differentials; wage dispersion; wage drift

Article

Labour market institutions – the organizations and procedures through which workers, firms, and the government affect wages, employment and working conditions – vary widely across countries and among firms and industries within a country. In some countries or settings within a country, trade unions, employer federations, personnel and human resource departments of firms and various forms of collective bargaining, or government regulations greatly affect how firms and workers interact at work places and help determine the hours, wages, occupational health and safety conditions, rules for promotion, and other conditions of work life. In other settings and countries these institutions have little impact. In those situations the market rules.

Once a minor tributary of economic analysis, the study of labour market institutions moved to the mainstream of discourse in the 1990s and 2000s as economists focused on differences in labour institutions as a possible cause of the varying economic performances among countries that had roughly similar macro-economic policies. The Organisation for Economic Co-operation and Development's influential 1994 Jobs Study (OECD, 1994a; 1994b) spurred research in advanced countries with its claim that many institutional interventions in the labour market reduced employment and that OECD countries should deregulate labour markets and weaken welfare state protections to achieve full employment. Ensuing analyses questioned the evidentiary basis for this diagnosis, producing a wide-ranging debate about how labour institutions affect advanced market economies. In developing countries, the analogous claim has been that institutionally determined wages and rules of work in the formal sector of economies reduce job creation in that sector and thus contribute to a dual labour market that harms economic growth and worsens the distribution of income. This also has generated considerable debate, pitting analysts who see institutions largely as creating distortions in competitive markets against those who see them as mechanisms for resolving market failures and shifting income distribution to workers.

Institutional differences

The starting fact for the debate is the wide variation of institutional arrangements in both advanced countries and developing countries. Table 1 summarizes the institutional architecture of the labour market in the United States and in advanced European countries – defined as European Union (EU) countries exclusive of the United Kingdom and Ireland, whose institutions are often closer to those of the United States than to the rest of the advanced Europe (Freeman, Boxall and Haynes, 2007) and inclusive of Norway and Switzerland, which are outside the European Union. The exhibit shows that the percentage of workers in unions is three times greater in the advanced European countries than in the United States. It notes a large difference in the organization of firms into employer associations. In advanced Europe many firms join employer associations that negotiate with unions, whereas in the United States employers negotiate separately with unions or with individual employees in the absence of collective bargaining. In addition, many advanced European governments extend the terms of a contract between an employer federation or major employer to all firms and workers in a sector, including those who were not party to the agreement, on the grounds that collective bargaining should produce a single wage just as supply and demand should produce a single wage in a competitive labour market. As a result of mandatory extension of contracts, the rate of collective bargaining coverage in advanced Europe (80 per cent in the table) exceeds the rate of unionization (38 per cent in the table); whereas the rates of union density and collective bargaining coverage are about the same in the United States. As a result, the gap in coverage between the United States and advanced Europe exceeds that in union density. The effect of mandatory extension on wage setting is most dramatic for France, where approximately 90 per cent of workers are covered by collective bargaining even though union density is six per cent or so – the lowest among advanced countries.

Labour market institutions in the market-driven United States versus institution-driven advanced Europe

	USA	Advanced Europe*

Union density, 2003	12%	38%
Extent of employer federation	Negligible	Substantial, bargain regularly
Percentage of workers covered by collective bargaining, 2000	14%	80%
Extension of collective contracts	none	Widespread by law
Employment protection legislation (higher values imply more protection, from 0 to 4)	0.7	2.7
Works councils	None	Mandated
Social dialogue	None	Widespread
Ratio of unemployment insurance to past wage, 2004	54	69
Months of unemployment insurance coverage, 2004	6 months	22 months
Social expenditures as share of national income, 2003	18.7%	28.9%
Rating of labour market in market orientation, 2003 Fraser (1=most market oriented), 103 countries	10	76
Rating of labour market in market orientation, 2003 Global Labor Survey (1=most market oriented), 33 countries	6	26

*Excluding the United Kingdom and Ireland. Italy, France, Spain, Greece, Sweden, Portugal, Germany, Belgium, Austria, Denmark, Finland, Norway, Netherlands, Switzerland.

Source: Union density from Visser (2006), OECD (2004, Table 3.3; 2004, Table 2.A2.4, version 2; 2006a, Table 3.2; 2006b, Figure GE1.2, p 41), Gwartney, Lawson and Gartzke (2005), and Freeman and Chor (2005).

At the enterprise level, all countries in the European Union require that firms above a specified size introduce a works council of democratically elected employee representatives and that the firm consult with the council on key decisions that affect workers. In Germany firms must reach agreement with the council on some issues or go to arbitration to resolve disagreements. By contrast, the United States outlaws non-union employee organizations at the workplace for fear that they will become company-dominated barriers to independent unions. Many US firms set up employee involvement committees to deal with issues regarding workplace productivity, but these committees cannot legally represent workers’ interests to management. Going beyond enterprises, the EU relies extensively on *social dialogue* among employer federations, unions, and in many cases, governments to determine labour and other economic policies. Social dialogue produced Ireland's 1987 Solidarity Wage Agreement in which the government agreed to lower taxes on workers, unions agreed to moderate wage demands, and employers agreed to seek to increase employment. The ensuing economic boom in Ireland suggested to some that the social pact contributed positively to Irish economic performance.

There are also large country differences in hiring and firing practices. Firms in the United States operate largely by employment at will, which means that the firm can replace workers for any business or other (non-discriminatory) reason. By contrast, many EU countries have employment protection legislation that requires firms to give substantial severance pay to laid off workers and to negotiate 'social contracts' with works councils to help laid off workers obtain training and new employment. In addition, European welfare states pay higher unemployment insurance in relation to wages for longer periods of time than does the United States, and provide national health insurance that US firms and workers must fund for themselves. These policies produce higher government social expenditures as a share of national income in the EU countries than in the United States, and commensurately higher taxes as a share of national income to pay for the benefits.

Taking these and related differences in the labour market together, analysts have created aggregate thermometer style indices of the institutional versus market orientation of country labour markets, in which higher scores reflect greater reliance on markets than on institutions. The Fraser Institute – a conservative think tank that produces an index of economic freedom based on metrics for 'personal choice, voluntary exchange, freedom to compete, and protection of person and property' (Gwartney, Lawson and Gartzke, 2005, p. 5) – codes countries that have extensive legal protection of labour and high levels of collective bargaining as having less economic freedom than those without these institutions. The Global Labor Survey has created a comparable index by asking union leaders, labour relations professors and other experts to report on the *actual situation* of labour in their country (Chor and Freeman, 2005). The difference in ideological persuasion between the Fraser Institute and most respondents to the Global Labor Survey notwithstanding, the two indices tell a similar story about cross-country differences. They give the United States and the other English-speaking advanced countries higher scores in using markets than European Union economies, and give the Scandinavian countries, which rely extensively on collective bargaining to determine pay and working conditions, particularly low scores in reliance on markets. While analyses of labour institutions in developing countries are less plentiful, the Fraser Economic Freedom Index and Global Labor Survey show a similar wide variation in the institutional framework for those countries. Botero et al. (2004) provide additional information on labour institutions across countries in terms of their labour laws. The indices of labour laws measure *de jure* labour institutions, whose impact on the labour market depends on the extent to which countries enforce their laws.

Institutions and outcomes

To see how institutions affect economic outcomes, analysts compare the economic outcomes for firms and workers *within countries* whose pay and work conditions are set by unions or regulations with the outcomes of firms and workers whose pay and conditions are set by market forces; compare the outcomes when institutional rules change (for instance, through an increase in minimum wages); and when the workers or firms move from market determination of wages and conditions of work to having an institution determine wages and conditions, or vice versa (for instance from moving from union to non-union status or non-union status to union status). To analyse how differences in institutions affect outcomes across countries, analysts contrast labour market outcomes *between countries* that rely more on institutions and those that rely more on markets; and compare outcomes before and after a country

changes its institutions with outcomes in countries that maintain their institutions over the same time period. The goal is to use the experiences of countries that do not change institutions as a counterfactual to predict what might have happened to countries that change institutions, and, conversely, to use the experience of the country that changed institutions to predict what might have happened in countries with stable institutions if they were to change.

Constructing a counterfactual to assess the impacts of institutions is difficult. One difficulty is that changes in institutional arrangements can affect the behaviour and outcomes for the group that is not covered by the changes as well as the covered group. A decline in union density, for example, might lower the wages of union and non-union workers equally, so that the differential between them was constant, which an analyst could misinterpret as implying no change in the wages of union workers. Another reason is that persons involved with institutions learn from past experiences, so that they may respond differently in the future to a given change in conditions than they might have done in the past. British unions made different decisions in the 1990s from those they made in the 1970s, in part because of their experiences in the earlier period. Finally, to the extent that one institutional rule affects another, a counterfactual analysis of a change in a single institution can be misleading if it does not allow for how the change interacts with other regulations and rules. When Spain enacted a law permitting firms to hire workers on temporary contracts, there was a huge increase in the proportion of workers hired under those contracts. When Germany enacted such a law, firms continue to hire apprentices for permanent jobs.

Difficulties of developing a valid counterfactual notwithstanding, virtually all analyses find that labour institutions reduce the dispersion of hourly earnings and the inequality of income (which depends on hours worked, and streams of income outside of work in addition to hourly pay) compared to market-pay setting. Studies that compare the distribution of earnings and incomes across countries find, for example, that the pay of persons in the 90th percentile of wages and salaries in relation to the pay of persons in the 10th percentile is lower in the advanced European countries that rely more on collective bargaining than in the market-driven United States and other English-speaking countries; and that the Gini coefficient of inequality for total income is also markedly lower in countries where labour institutions dominate wage-setting (Table 2). The US has the largest 90/10 earnings ratio of wages and the largest Gini coefficient for total income among advanced countries. By contrast, the Nordic countries, where collective bargaining sets wages for the vast majority of workers, have the lowest dispersion of pay and low Gini coefficients. Other advanced European countries and Japan also have relatively low pay dispersion and Gini coefficients. Centralized collective bargaining arrangements are sufficiently effective to narrow pay gaps even though most centralized agreements allow for ‘wage drift’ – higher or lower wages for some firms and workers than the negotiated central agreement due to variations in local market conditions.

90/10 Wage differentials and Gini coefficients for advanced countries, circa 2000

	Dispersion	Gini
US	4.59	40.8
Other English-speaking	3.46	35.2
Advanced Europe	3.10	32.2
Japan	2.99	24.9
Scandinavia	2.18	25.6

Source: 90/10 ratios averaged from data from OECD (2004, Table 3.2), where the data are from 1995–99 with figures from Austria, Belgium, Denmark, Portugal are for 1990–4; data for Spain and Greece from Martins and Pereira (2004, Table 1). Gini coefficients from United Nations, Human Development Report (2005, Table 15) Other English-speaking countries are: the United Kingdom, New Zealand, Canada, Ireland and Australia. Advanced Europe countries are: Belgium, Netherlands, Italy, Switzerland, France, Austria, Germany, Spain, Portugal, Greece; Scandinavia are: Norway, Finland, Sweden and Denmark.

Looking at earnings when country institutions change, increased reliance on institutions narrows the distribution of earnings while increased reliance on market-wage setting widens the distribution. Declines in collective bargaining coverage in the United States, Canada, United Kingdom and New Zealand contributed to greater inequality in those countries. Similarly, the decline in the real value of the US minimum wage added to inequality, while the introduction of the minimum wage in the United Kingdom limited the rise of inequality in that country. The breakdown of centralized negotiations between the major union federation and major employer association in Sweden raised inequality modestly in that country. But perhaps the most compelling evidence comes from the rise and fall of Italy's Scala Mobile mode of pay setting. The Scala Mobile was a national agreement that gave larger percentage increases in pay to low-wage workers than to high-wage workers. When the Scala Mobile determined wages, the dispersion of earnings in Italy fell sharply – towards Scandinavian levels. When Italy abandoned this mode of pay setting, in part because the distribution of pay seemed to have narrowed wage differentials beyond what made economic sense, the dispersion of earnings increased (Erickson and Iquino, 1995; Manacorda, 2004).

Studies within countries that contrast the inequality of pay among workers whose pay is set by institutions and those whose pay is set by markets also find that institutions are associated with lower dispersion of pay. Dispersion is less among unionized workers than among otherwise comparable non-union workers and less among government employees than among private sector employees whose pay is market-determined. Moreover, although the wage differential between union and non-union workers raises inequality between organized and non-organized workers, the net effect of unions on earnings is to reduce inequality. The overall distribution of earnings is dominated by the compression of wages within the union sector and by difference the reduced earnings between management and other high-paid non-union workers and union workers within firms. Consistent with this, studies that contrast the inequality of pay among workers who shift from non-union jobs to union jobs or the converse find that dispersion among a group of job changers falls when workers enter the union sector and rises when they leave the unionized setting (Freeman, 1984).

Is the institution-induced reduction in the dispersion of pay good or bad for the economy? To the extent that real world labour markets perform largely as ideal competitive markets, the reduced dispersion of pay distorts economic decisions on both the supply and demand sides of the market. By contrast, to the extent that real-world labour markets fall short of the competitive ideal, institutions can improve the efficiency of markets. There are plausible arguments and evidentiary support for both interpretations of what institutions do.

The arguments

The claim that *labour institutions adversely affect economic performance* begins with the assumption that in the absence of institutional interventions, real labour markets produce wage, employment, and working conditions that approach those of an ideal competitive labour market. In this case, institutions can only distort incentives and reduce the efficient allocation of resources. For instance, union-induced wages above the market rate induce unionized firms to reduce employment, which reallocates labour to lower paid less productive activities. The following statement from the World Bank expresses the view that institutions distort the demand for labour in developing countries and slow down the shift of labour from agriculture and informal sector work to more highly productive and better-paid formal sector jobs:

> Labor market policies – minimum wages, job security regulations, and social security – are usually intended to raise welfare or reduce exploitation. But they actually work to raise the cost of labor in the formal sector and reduce labor demand … increase the supply of labor to the rural and urban informal sectors, and thus depress labor incomes where most of the poor are found. (World Bank, 1990, p. 63)

The arguments against labour institutions in advanced countries are similar. On the demand side, institutionally driven increases in wages for the low-paid raise their cost to employers, which lowers their employment, and distort the allocation of the workforce among sectors, squeezing in particular low wage service industries. On the supply side, institutionally driven reductions in earnings inequality reduce pecuniary incentives to make efficient economic decisions. All else the same, reductions in the earnings premium paid to more-skilled workers will reduce investments in skills. And high unemployment insurance benefits will induce laid off workers to raise the reservation wage at which they will accept a new job and to search less intensely for jobs, producing longer spells of joblessness and higher rates of unemployment. In addition, the reduction in job search will lessen supply side pressures towards modest wage settlements that help job creation.

The magnitude of the distortions depends on the responsiveness of decision-makers to the institutionally determined incentives. In the standard ‘welfare triangle’ analysis, the economic loss from raising a wage above the market rate depends on the magnitude of the wage change and the elasticity of demand, which determines the magnitude of the distortion in the allocation of labour. (The formula for a welfare loss is $\frac{1}{2}$ (change in wages) x (change in employment), where the change in employment is the elasticity of demand times the change in wages.) The higher the elasticity of demand, the greater will be the welfare loss from wages above the market rate. Similarly, on the supply side, the higher the elasticity of supply to the returns to skills, the greater will be the welfare loss from decisions to forgo investments in skill due to the compression of wages, and the higher the elasticity of supply to unemployment benefits, the greater will be the welfare loss, due to the decision to search less intensely for a new job due to unemployment insurance.

Finally, institutional determination of labour outcomes can impose two additional costs on the economy. The first is the political lobbying and related resources that labour and management spend to affect labour regulations and the rules governing union and employer interactions. These are sometimes pejoratively labelled as the costs of rent seeking, though if institutions help to solve economic problems they could just as well be called the costs of problem-solving. The second are the resources involved in implementing institutional arrangements. These range from establishment of union and employer

federations, time spent in negotiations and dialogue at the workplace and at national levels. Discussion reduces the speed of decision-making, so that institutionally driven systems are likely to respond more slowly to economic changes than market-driven systems.

On the other side of the debate, the argument that *labour institutions improve economic performance* begins with the belief that real labour markets fall short of competitive equilibrium. Analysts view the high dispersion of pay for workers with observationally equivalent skills as reflecting the failure of the market to establish a single price of labour for similar workers. If this is a correct reading of the data, institutionally determined reductions in dispersion could create outcomes closer to the competitive ideal just as institutionally determined increases in wages can induce firms that are monopsonies to raise employment to competitive levels. Looking at the dynamics of wage changes, in an ideal competitive system, improvements in productivity in a given sector are supposed to show up in lower prices to consumers, not in higher wages (Salter, 1960; Council of Economic Advisors, 1962); while changes in the prices of products due to changes in demand are supposed to induce firms to change output and employment but not to change wages. The reason wages are not expected to respond to these shocks is that the competitive model posits that firms face a perfectly elastic supply of labour at the market wage rate. In fact, changes in wages are highly related to changes in productivity and prices among industries in the United States but not in the Nordic and other countries where institutions determine wages (Holmlund and Zetterberg, 1991; Teulings and Hartog, 2002). At the national level, some analysts argue that the union and employer federations that negotiate national wage agreements adjust wages more rapidly to macroeconomic developments such as balance of payments or inflation than local labour markets that respond to the macroeconomy less directly.

Finally, inside firms, labour institutions can facilitate the flow of information from workers to management and from management to workers. Workers are more likely to provide information to management when they can influence how management uses the information. Regulations or union pressure that force management to open its books to workers gives them or their representative access to the same information that guides management. Increasing the flow of information and communication can in turn lead management and workers to make better decisions. Workers will be more likely to give wage concessions when the firm is truly in crisis and avoid being snookered when the firm cries ‘wolf’ while continuing to earn profits (Freeman and Lazear, 1995). In addition, workers who have an institutional voice for dealing with problems are less likely to quit their employer and more likely to invest in firm-specific skills and seek to resolve problems by bringing them to the attention of management.

Evidence

The *OECD Jobs Study* contains two volumes of research and references to research that buttressed its claim that labour institutions explained some of the job market problems of OECD countries. Since the *Jobs Study* many other analysts have examined the link between those institutions and outcomes, generally using cross-country time series data that the OECD provides. Each year the OECD reviews the latest findings on particular issues regarding the impact of labour institutions in its *Employment Outlook*. As economists inside and outside the OECD have critically examined the data and models that link outcomes to institutions, they have moved to a more cautious stance about the evidentiary support for

the *Jobs Study* conclusions. Assessing the time series models that the OECD and others used in their analyses, Baker et al. (2005) found that the estimated coefficients on labour institutions were not robust to changes in specification. They found that models that covered more years, additional countries or used different measures of the institutions than the early studies ‘provide little support for those who advocate comprehensive deregulation of OECD labour markets’ (2005, p. 106). Baker et al. conclude that there is a ‘yawning gap between the confidence with which the case for labour market deregulation has been asserted and the evidence that the regulating institutions are the culprits’ (2005, p. 198). Assessing results in the mid-2000s, Howell et al. (2007) and Baccaro and Rei (2005) come to a similar conclusion.

For its part, the OECD has recognized that the evidence is more equivocal than first claimed. The 2004 *OECD Employment Outlook* noted that ‘the evidence of the role played by employment protection legislation (EPL) on aggregate employment and unemployment rates remains mixed’ (2004, p. 81). It argued for ‘the *plausibility* (my italics) of the Jobs Strategy diagnosis that excessively high aggregate wages and/or wage compression have been impediments’ to jobs, while admitting that ‘this evidence is somewhat fragile’. With respect to unionism, it summarized research as showing the effect of collective bargaining ‘to be contingent upon other institutional and policy factors that need to be clarified to provide robust policy advice’ (2004, p. 165). In a similar vein, the IMF (2003) reported that ‘Institutions … hardly account for the growing trend observed in most European countries and the dramatic fall in U. S. unemployment in the 1990s.’ German unemployment, for example, rose by about six percentage points in the 1990s while US unemployment fell, even though labour institutions were broadly unchanged in both countries. But the IMF still concluded that the route to full employment rested with deregulating labour markets. The strong priors and commitment to the case that institutions are the problem overrode the actual evidence.

The 2006 *OECD Employment Outlook* went a step further in assessing the impact of institutions on outcomes. It highlighted that countries with low unemployment had very different modes of wage-setting, ranging from some smaller European countries that relied on collective bargaining to the more market-determined United States and United Kingdom (2006a, Table 6.3). If different institutions can reach similar market outcomes, there may be no ‘peak’ form of labour market institutions to which each country should strive (Freeman, 2002). But this does not resolve the debate over the impact of institutions. In a study that took account of criticisms of the non-robust findings of earlier cross-country time series data, Bassanini and Duval (2006) found that changes in tax and labour policies explain about half the 1982–2003 changes in unemployment among countries, with changes in tax policies playing a particularly large role.

The potential effect of employment protection legislation on unemployment has attracted considerable attention. Countries pass these laws to reduce layoffs and raise job security for existing workers. But the laws make it more expensive to hire workers since firms must factor in the greater expense of laying them off if business dictates reductions of output. The net effect of employment protection laws on aggregate employment thus depends on the degree to which they reduce layoffs compared to the degree to which they reduce hires. An alternative perspective predicts that on net the employment protection laws should have little or no impact on aggregate employment or unemployment. If employers and unions bargain efficiently, then the Coase Theorem predicts that they should bargain so that the firm makes the efficient layoff regardless of the employment protection law. What differs is the division of the profits from the efficient choice. With employment protection the firm pays some of the profit from

a layoff to the worker to get the worker to leave. With employment at will, the firm gets all the profit from the decision. Studies of unemployment and employment between countries with greater or lesser employment protection are broadly consistent with this view. They show that the regulations have little effect on the overall rate of unemployment but shift unemployment from older workers to younger job-seekers (OECD, 2004). In developing countries as well, job security regulations appear to shift employment from the unskilled youth to the skilled and older workers protected by the legislation. (Montenegro and Pages, 2003).

In summary, there is no clear consensus from the empirical analyses that labour institutions have adverse or positive effects on aggregate economic outcomes beyond their distributional effects on earnings or employment.

Alternative interpretations

There are three possible interpretations of the empirical evidence that institutions reduce the dispersion of earnings and income but do not have clear or easily identified effects on other aggregate economic outcomes.

The first interpretation is that extant measures of institutions and models of their impact are too crude to pin down the hypothesized effects on other outcomes. Better cross-section time series data on countries and more sophisticated statistical modelling might produce statistically significant impacts of institutions on outcomes beyond dispersion of pay. Most economists believe that disaggregated data that cover thousands of observations on individuals or firms has a greater likelihood of pinning down responses of individuals and firms to changes in labour policies and institutions than further analysis of short time series across countries. But these analyses are insufficient in themselves to capture what might happen when a country changes its institutions. What might better illuminate the impacts of institutions at the national level would be to combine estimated response parameters from microeconomic studies with artificial agent models that simulate labour markets under different institution.

The second interpretation is that the effects of institutions vary over time as the economic environment changes. Given that the labour market institutions in the United States and advanced Europe were largely unchanged between the 1960s and 1990s, the only way for institutional factors to explain lower European unemployment in the former period and higher European unemployment in the latter period would be that the impact of the institutions changed over time (Blanchard and Wolfers, 2000; Lundquist and Sargent, 1998; OECD, 2006). Perhaps EU institutions were well suited to produce low unemployment in the economic conditions of the 1960s–1980s while US institutions were better suited to produce low unemployment in the globalized digital economy of the 1990s and 2000s. This interpretation is appealing. But it is difficult to test since it makes great demands on data. Allowing institutions to affect outcomes differently in different time periods reduces the number of observations with which to test the hypothesized impact and risks creating epicycles of interactions to account for observed patterns.

The third interpretation is that in fact labour institutions have first-order effects on income distribution but only modest second-order effects on other outcomes. Perhaps the hypothesized adverse effects that institutions can have on economic efficiency are balanced by their hypothesized positive effects, giving a net effect around zero. This is consistent with efficient bargaining theory, in which parties strive to

reach efficient outcomes but battle over distribution. This interpretation is appealing. But there are enough situations in which unions, firms and governments do not reach efficient solutions to raise questions about it. As Sir John Hicks pointed out in *The Theory of Wages* (1934), efficient bargaining implies that strikes, which in most cases harm workers and firms, should vary randomly across industries, regions, firms and time as a result of random errors of judgment or communication. In fact strikes occur frequently in some sectors (for instance coal mining) and not in others, in some firms but not in others, and vary over the business cycle in ways that conflict with the efficient bargaining model. In conclusion, we need to learn much more about how labour institutions affect the economy and how they operate for us to resolve the debate over whether institutions are part of the problem facing economies or part of the solution, or, more likely, which institutions and issues fall more into the former category and which fall into the latter category under particular economic conditions.

See Also

- labour market search
- labour supply
- minimum wages
- unemployment
- unemployment and hours of work, cross country differences
- unemployment insurance

Bibliography

Baccaro, L. and Rei, D. 2005. Institutional determinants of unemployment in OECD countries: a time series cross-section analysis (1960–98). Discussion Paper No. DP/160/2005. Geneva: International Institute for Labor Studies.

Baker, D., Glyn, A., Howell, D. and Schmitt, J. 2005. Labor market institutions and unemployment: a critical assessment of the cross-country evidence. In *Fighting Unemployment: The Limits of Free Market Orthodoxy*, ed. D. Howell. Oxford: Oxford University Press.

Bassanini, A. and Duval, R. 2006. Employment patterns in OECD countries: reassessing the role policies and institutions. Working Paper No. 486. Paris: Economic Department, OECD.

Blanchard, O. and Wolfers, J. 2000. Shocks and institutions and the rise of European unemployment: the aggregate evidence. *Economic Journal* 110, 1–33.

Botero, J., Djankov, S., La Porta, R. and López de Silanes, F. 2004. The regulation of labor. *Quarterly Journal of Economics* 119, 1339–82.

Chor, D. and Freeman, R. 2005. The 2004 global labor survey: workplace institutions and practices

around the world. Working Paper No. 11,598. Cambridge, MA: NBER.

Council of Economic Advisors. 1962. *Economic Report of the President*. Washington, DC: Government Printing Office.

Erickson, C. and Iquino, A.C. 1995. Wage differentials in Italy: market forces, institutions and inflation. In *Differences and Changes in Wage Structure*, ed. R.B. Freeman and L.F. Katz. Chicago: University of Chicago Press for NBER.

Freeman, R.B. 1984. Longitudinal analyses of the effects of trade unions. *Journal of Labor Economics* 2, 1–26.

Freeman, R.B. 1993. Labor market institutions and policies: help or hindrance to economic development? *Proceedings of the World Bank Annual Conference on Development Economics 1992*. Washington, DC: World Bank.

Freeman, R.B. 2002. Single peaked vs. diversified capitalism: the relation between economic institutions and outcomes. In *Advances in Macroeconomic Theory*, ed. J. Drèze. London: Palgrave.

Freeman, R.B. 2005. Labor market institutions without blinders: the debate over flexibility and labour market performance. *International Economic Journal* 19, 129–45.

Freeman, R.B., Boxall, P. and Haynes, P. 2007. *What Workers Say: Employee Voice in the Anglo-American World*. Ithaca, NY: Cornell University Press.

Freeman, R.B. and Lazear, E. 1995. An economic analysis of works councils. In *Works Councils: Consultation, Representation, and Cooperation in Industrial Relations*, ed. J. Rogers and W. Streeck. Chicago: University of Chicago Press for NBER.

Gwartney, J., Lawson, R.A. and Gartzke, E. 2005. *Economic Freedom of the World: 2005 Annual Report*. Vancouver, BC: Fraser Institute.

Hicks, J. 1934. *The Theory of Wages*. London: Macmillan.

Holmlund, B. and Zetterberg, J. 1991. Insider effects in wage determination: evidence from five countries. *European Economic Review* 35, 1009–34.

Howell, D.R., Baker, D., Glyn, A. and Schmitt, J. 2007. Are protective labor market institutions at the root of unemployment? A critical review of the evidence. *Capitalism and Society* 2, Article 1. Abstract online. Available at http://www.bepress.com/cas/vol2/iss1/art1, accessed 12 June 2007.

IMF (International Monetary Fund). 1999. Chronic unemployment in the Euro area: causes and cures. In *World Economic Outlook*, May. Washington, DC: IMF.

IMF. 2003. Unemployment and labor market institutions: why reforms pay off. In *World Economic Outlook*, April. Washington, DC: IMF.

Kruse, D. 1993. *Profit-Sharing: Does It Make A Difference?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Layard, R., Nickell, S. and Jackman, R. 1994. *The Unemployment Crisis*. Oxford: Oxford University Press.

Lundqvist, L. and Sargent, T.J. 1998. The European unemployment dilemma. *Journal of Political Economy* 106, 514–50.

Manacorda, M. 2004. Can the Scala Mobile explain the fall and rise of earnings inequality in Italy? A semiparametric analysis, 1977–1993. *Journal of Labor Economics* 22, 585–613.

Martins, P.S. and Pereira, P.T. 2004. Does education reduce wage inequality? Quantile regression evidence from 16 countries. *Labour Economics* 11, 355–71.

Milner, H. and Wadensjö, E. 2001. *Gösta Rehn, the Swedish Model and Labour Market Policies: International and National Perspectives*. Aldershot: Ashgate.

Montenegro, C.E. and Pages, C. 2003. Who benefits from labor market regulations? Chile 1960–1998. Working Paper No. 9850 Cambridge, MA: NBER.

Nickell, S.J. and Bell, B. 1996. Changes in the distribution of wages and unemployment in the OECD countries. *American Economic Review* 86, 302–8.

Nickell, S., Nunziata, L. and Ochel, W. 2005. Unemployment in the OECD since the 1960s: what do we know? *Economic Journal* 115, 1–27.

Olson, M. 1990. *How Bright Are the Northern Lights? Some Questions about Sweden (Crafoord Lectures)*. Lund: Institute of Economic Research, Lund University.

OECD (Organisation for Economic Co-operation and Development). 1994a. *OECD Jobs Study, Evidence and Explanations, Part I: Labor Market Trends and Underlying Forces of Change*. Paris: OECD.

OECD. 1994b. *OECD Jobs Study, Evidence and Explanations, Part II: The Adjustment Potential of the*

Labor Market. Paris: OECD.

OECD. 1997. Economic performance and the structure of collective bargaining. *OECD Employment Outlook*. Paris: OECD.

OECD. 2004. *OECD Employment Outlook*. Paris: OECD.

OECD. 2006a. *OECD Employment Outlook*. Paris: OECD.

OECD. 2006b. *Society at a Glance: OECD Social Indicators*. Paris: OECD.

Salter, W.E.G. 1960. *Productivity and Technical Change*. Cambridge: Cambridge University Press.

Teulings, C. and Hartog, J. 2002. *Corporatism or Competition? Labour Contracts, Institutions and Wage Structures in International Comparison*. Cambridge: Cambridge University Press.

UN (United Nations). 2005. *Human Development Report*. New York: Oxford University Press.

Visser, J.J. 2006. Union membership statistics in 24 countries. *Monthly Labor Review* 129, 38–49.

World Bank. 1990. *World Development Report*. New York: Oxford University Press.

World Bank. 1995. *World Development Report*. Washington, DC: World Bank.

How to cite this article

Freeman, Richard B. "labour market institutions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000006> doi:10.1057/9780230226203.0916

The New Palgrave Dictionary of Economics Online

labour market search

Dale Mortensen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Time and other resources are required in the process by which workers and jobs are matched: this process is referred to as labour market search. Models of the search process have made contributions to our understanding of unemployment incidence and duration, labour turnover, earnings growth and wage dispersion. These models, which are based on the assumption that agents act in their own best interest, are designed characterize market equilibria in environments complicated by imperfect information and uncertainty. Consequently, they are also useful in the analysis of labour market policy.

Keywords

Beveridge curve; labour market search; law of one price; market tightness; matching; productivity shocks; reservation wage; search costs; search theory; social networks; unemployment; unemployment insurance; wage dispersion

Article

Labour market search refers to the process by which workers and employers find and match with one another in the labour market. The theoretical models that have been developed to understand the process explicitly account for the fact that search and matching are time consuming activities. The models have been used to interpret empirical data on phenomena that include unemployment duration and incidence, unemployment fluctuations, labour turnover, earnings growth, and wage dispersion and discrimination. As with other equilibrium models in economics, these are based on the assumption that participants in the labour market act in their own self-interest and that the market phenomena observed are explained by outcomes of the market participant interaction. Hence, they can and are used to address the consequence of labour market policy on labour market statistics and on the welfare of labour market participants.
Equilibrium search theory extends the standard competitive model of the labour market. In spite of the

obvious usefulness and elegance of the competitive market equilibrium for the analysis of many questions, the framework in its simplest form excludes much of the phenomena of interest to a labour economist. To give but two examples, there is no unemployment in competitive equilibrium and workers with identical skills earn the same wage. Equilibrium labour market search theory was developed to explain these facts as well as other phenomena having to do with the dynamics of the employment and earning experiences of individual workers that cannot be accommodated in the standard model.

More generally, the labour market search framework has proven to be a useful tool for thinking about markets with ‘friction’, those that function without the market clearing auctioneer invoked in the competitive market framework. How do the participants in such markets come together? How are prices and the quantities exchanged at these prices determined? Search theory attempts to answer these questions.

Modelling the fact that the labour market experiences of individual workers take place in real time is an essential ingredient of the labour market search approach. After workers leave school and enter the labour market, most spend time seeking a job. Once employed, young workers seem to ‘job-shop’ by trying several employers and occupations before settling down to an extended period of employment with one. Later, employment spells are punctuated by interruptions attributable to changes in the individual's desire for employment, on the one hand, and the termination of the worker's current job, on the other. It is fair to say that virtually all the recent theoretical treatments of these phenomena are based on labour market search theory, and that theory informs the interpretation of most empirical studies that focus on them.

Individual search behaviour

The reservation wage

Labour market search theory began as a model of how a worker might gather information about employment opportunities. In the real world, there are many sources of such information. Economists and sociologists distinguish between formal and informal search channels. Formal channels are information sources provided by market institutions such as newspaper advertisements, public employment services, private employment agents and the Internet. Informal channels include friends, relatives, and neighbours, anyone in the workers’ extended social network. It is well known that most workers find their jobs through these informal channels, a fact that underlines the decentralized nature of the labour market.

The first important paper in search theory, by George Stigler (1961), was an attempt to formalize the economic problem posed by the need to gather information about trading opportunities in a non-auction market where different prices for the same or close substitutes can coexist. He modelled the problem as one of choosing the size of a sample of prices drawn randomly from the available set. Given that the agent would purchase the good from the lowest-priced seller in the sample and must pay a fixed cost per price sample, how many quotes should the buyer seek?

Although this well-known problem in sampling theory provides some interesting insights, it does not serve well as a model of job search. In that context, it is the worker's time rather than money that is the principal cost incurred. Furthermore, the length of time spent by an unemployed worker is an observable quantity that is measured in both survey and administrative data. The models focused on the duration of

search, which were simultaneously introduced by McCall (1970), Mortensen (1970), and Gronau (1971), became the basis for further work on the subject.

The idea underlying these models is that the duration of job search by an individual worker is usefully viewed as a random variable with the length determined by the worker's decision to accept or refuse offers as they arrive. In other words, instead of gathering a sample of job opportunities and selecting the one most preferred as in Stigler's formulation, these authors argued that it was more realistic to think of the search process as sequential in time. Offers arrive one at a time and the unemployment period ends when the worker accepts one of them. As McCall (1970) pointed out, this is formally an optimal stopping problem in the theory of decisions under uncertainty. It is well known that a reservation strategy is optimal: accept the first offer above some critical value.

Formally, let $F(w)$ characterize the distribution of offers and suppose that the number of offers received in a time period of unit length is a Poisson random variable characterized by the arrival rate λ. Assume that the worker is a risk neutral with an indefinite future life span. When the distribution of wages is known, the optimal strategy is to accept the first wage offered above a *reservation wage*. In other words, the reservation wage, denoted as R, is the lower bound on the set of wages that are acceptable to the worker. Accepting employment at the reservation wage must just compensate for any income forgone by becoming employed, which one can think of an unemployment benefit denoted by b, plus the option of continued search. Formally, if the worker does not plan to search while employed, the reservation wage is the implicit solution to the indifference condition.

$$R = b + \lambda \int_R^{\bar{w}} \left(\frac{w - R}{r + \delta} \right) dF(w), \tag{1}$$

where r is the rate at which worker's discount future income, δ is the rate at which the worker can expect to lose a job, and $\bar{w}$ is the upper support of the wage distribution (see Mortensen and Pissarides, 1999a; 1999b for a derivation of the equation). The second term, the option value of continued search, is the product of the offer arrival rate and the expected present value of the future gains in income attributable to the possibility of receiving an offer in the future.

Empirical application

Labour economists later exploited the empirical implications of the original stopping model. In one of the first such papers, Ehrenberg and Oaxaca (1976) pointed out that the expected duration of an unemployment spell and the expected post-spell wage were both increasing in unemployment insurance benefit. Specifically, since a spell ends only when an offer is received and it is found acceptable, the hazard rate of the unemployment duration distribution is the product $\lambda[1-F(R)]$. Hence, the duration distribution is exponential with expectation equal to the inverse of the hazard that is increasing in R. As the post-spell distribution of wages is the distribution of offers truncated on the left by the reservation wage, the expected post-spell wage,

$$E\left\{w|w \geq R\right\} = \frac{\int_R^{\overline{w}} w dF(w)}{1 - F(R)},$$

is increasing in R. Since the reservation wage, the solution to (1), is increasing in unemployment income b, the expected duration of an unemployment spell and the average wage earned once employed should both increase with the generosity of the unemployment benefit.

Subsequently, Kiefer and Neumann (1979) used the fact that the model specifies the form of the statistical likelihood function for the observed length of unemployment spells and accepted wages. Formally, for a sample of n completed spells of unemployment followed by employment at an observable wage, a set of pairs denoted by (t_i, w_i), i=1,…, n, the likelihood of the observed sample is given by

$$L = \Pi_{i=1}^{n} \lambda [1 - F(R_i)] e^{-\lambda[1 - F(R_i)]} \times \left(\frac{F'(w_i)}{1 - F(R_i)}\right),$$

for a set of workers who all sample from the same wage offer distribution. In this equation, R_i is the reservation wage of worker i which varies with the entitled unemployment insurance payment as determined by eq. (4). Given observed values of b_i one can estimate both the offer arrival rate parameter and the distribution of acceptable wage offers using this structure, at least in principle. For a review of the early empirical literature that uses the duration analysis approach to estimation and search theory to interpret the results see Devine and Kiefer (1991). Wolpin (1995) provides an excellent treatment of the structural approaches to estimation of decision theoretic search models.

Equilibrium wage dispersion

The wage dispersion assumed in the stopping formulation of the job search problem is obviously inconsistent with the ‘law of one price’ that characterizes competitive equilibrium. Idiosyncratic match productivity is the simplest way to justify the assumption that a worker's employment opportunities can be described by a distribution of alternative wages. Although this justification may be sufficient for the purpose, Rothschild (1973) asks the following question: are there reasonable conditions under which wage dispersion, different wages paid to workers of identical skill, exists in equilibrium?

Consider the following simple one-shot game. All workers are identical and the common value of their marginal product is p in every firm. Assume that each worker receives a finite sample of job offers, say

of size *n*, chosen at random from the set of all offers. Since the worker has no future in this formulation, his or her best strategy is to accept the highest offer in the set provided that it exceeds the opportunity cost of employment, denoted above by *b*. Of course, positive gain from trade requires that *p*>*b*. Given this strategy, what will profit maximizing employers offer?

Suppose that *n*=2, that is, every worker receives offers from exactly two different firms. Because each worker will accept only the higher of the two offers, any employer paying a wage strictly less than all the others will hire no workers. Hence, a strictly positive fraction of employers must offer the lowest wage in the market: denote it by $\underline{w}$. It follows that the expected profit earned is $\underline{\pi} = (1-\alpha)\frac{1}{2}(p-\underline{w})$ where α is the fraction of workers that receive a strictly larger offer. In other words, the other wage drawn by the worker must also be the smallest in the market, an event that occurs with probability $1-\alpha$. If so, the worker chooses one of the two offers at random, with probability 1/2.

But it will always pay an individual employer to break ties by offering slightly more. That is, because the profit obtained by doing so is $\pi = (1-\alpha)(p-\underline{w}-\varepsilon) > \underline{\pi} = (1-\alpha)\frac{1}{2}(p-\underline{w})$ for all sufficiently small $\varepsilon>0$ if $p > \underline{w}$, it follows that the smallest offer is $\underline{w} = p$ in any equilibrium. Hence, all offers equal the competitive equilibrium wage, *w*=*p*. Obviously, this argument holds for any value of $n \geq 2$.

The fact that Bertrand competition obtains when every worker receives at least two offers would seem to rule out wage dispersion. However, this conclusion is false because the price-gathering process embodies an information externality as Rothschild (1973) points out. Suppose for the sake of argument that the first wage quote is costless but there is a small cost of finding a second. In this case, all workers sampling twice is not a non-cooperative equilibrium strategy in the game of wage search. Namely, if all workers see two prices, there would be no dispersion as we have just shown. But, if there is no dispersion, no worker has an incentive to pay the cost of obtaining a second price quote. It follows immediately that a single common wage equal to the opportunity cost of employment, *w*=*b*, is an equilibrium, a result due to Diamond (1971). However, Burdett and Judd (1983) demonstrate that another equilibrium generally exists in which a fraction of the workers seeks two offers while the complementary fraction obtains only one. The equilibrium in this case can be characterized by a unique continuous distribution of wage offers.

The details of the Burdett–Judd argument are beyond the scope of this article. However, the reason why an equilibrium of a wage posting game of the kind outlined above can be characterized by a distribution of offers is easily understood in the context of the sequential search model outlined above extended to allow for search on-the-job as in Burdett and Mortensen (1998).

If search costs are not too large, it is obvious that the employed as well as unemployed workers have an incentive to search when wage offers are dispersed. Hence, in the model there will be two kinds of workers: the unemployed, who only see one wage offer at a time, and employed workers, who will be able to choose between continuing employment at the same wage or moving to alternative employment when the opportunity arises. In short, at any point in time a strict subset of the workers have two offers while another fraction has one.

Given search on-the-job, employers who pay more attract a larger fraction of applicants and suffer less turnover. This trade-off between wage and turnover costs provides the reason for dispersion. Formally, let *V*(*w*) represent the value of meeting a prospective employee to an employer who pays wage *w*. It is the product of two terms, the probability that an applicant will accept the wage offered and the present

value of the future stream of profit that the employer can expect to earn if he or she accepts.
Under the assumption that all the employed workers accept any wage above a common reservation value, *R*, but an employed worker accepts if and only if the wage offer exceeds that currently earned, the acceptance probability is equal to

$$A(w) = u + (1 - u)G(w)$$

where *u* is the unemployment rate and *G*(*w*) is the fraction of workers who currently earn less than the wage offered, *w*. Given that job separations occur at rate δ for exogenous reasons and a worker will quit when ever a higher-paying job is located, the expected present value of future profit is

$$J(w) = \frac{p - w}{r + \delta + \lambda[1 - F(w)]}$$

where the production of λ, the rate at which the worker generates outside offers, and 1–*F*(*w*), the probability that an alternative offer exceeds the worker's current wage, is the rate at which an employed worker can be expected to quit. Hence, the expected value of meeting a worker contingent on the wage offered is

$$V(w) = A(w)J(w) = \frac{[u + (1 - u)G(w)](p - w)}{r + \delta + \lambda[1 - F(w)]}.$$

Because a higher wage increases the acceptance rate and reduces the quit rate, a trade-off between wage and turnover costs is evident in this relationship. It is natural to assume that each individual employer will choose the wage to maximize the expected present value of future profit, the function *V*(*w*), given the wage offers of all the other workers. However, because all the employers are identical by assumption and the offer distribution *F*(*w*) is endogenously determined by all their wage-setting decisions, it follows that profits must be both maximal and equal to the support of any equilibrium distribution. Furthermore, because unemployed workers accept all offers and an employed worker accepts an offer only if it exceeds that currently earned, the acceptance probability is the unemployment rate and the quit rate is the offer arrival rate λ at the lowest equilibrium offer, which is the common reservation wage *R*. Hence, the equilibrium distribution is the unique solution for *F*(*w*) to the following equal profit condition:

$$V(w) = \frac{[u + (1-u)G(w)](p-w)}{r+\delta+\lambda[1-F(w)]} = \frac{u(p-R)}{r+\delta+\lambda} = V(R)\ \forall w \in [R, \bar{w}] \text{ where } F(\bar{w}) = 1.$$

As no worker will accept employment at a wage below *R* and any wage offer above $\bar{w}$ yields less profit, this condition is also sufficient for profit maximizing.
Variations and extensions of this model have been used to study the link between wages, labour turnover, the return to education, discrimination, and the duration of employment spells. The approach has also proven to be a valuable tool for the analysis of firm data on employment and workers flows. Eckstein and van den Berg (2007) provide a review of the literature that uses the model as the basis for parameter estimation. See Mortensen (2003) for a more complete development of the theory and a review of the empirical applications of the approach.

Equilibrium unemployment

Search and matching model of unemployment, those based on the original two-sided search models of Diamond (1982), Mortensen (1982) and Pissarides (1985) have focused on the time required to find employment. In this family of models, match rent exists after worker and employer meet because finding an alternative is costly. When worker and employer meet, they are assumed to bargain over their joint output. According to Nash (1950), the outcome of the bargaining problem yields a wage equal to the flow value of unemployment, represented by the reservation wage *R*, plus some share of the rent attributable to the current job–worker match, when search for an alternative partner is assumed to be the outside option. Formally,

$$w = R + \beta(p - R),\ \beta \in (0, 1) \tag{2}$$

where *p* represents match output and the value share parameter β reflects the worker's relative ‘bargaining power’. When all matches are identical, the wage is the same for all job–worker matches. Furthermore, one can show all matches are acceptable if and only if match product *p* exceeds the opportunity cost of employment *b*.
As wages are the same in all jobs in this model, there are no quits, which implies that the expected present value of the future profit attributable to employing a worker is $J(w) = \frac{p-w}{r+\delta}$. Hence, an employer has an incentive to create a job whenever *J*(*w*) exceeds the cost of doing so. For example, if the cost of advertising a job opening is *c* and the advertisement will attract applicants at frequency η per period, then it pays to post a vacancy whenever the expected cost of filling it, c/η, is less than the expected return to doing so as represented by *J*(*w*).
At this point, it may have occurred to the reader that the rate at which workers are matched with jobs,

denoted above as λ, and the rate at which vacant jobs are matched with worker, η, above must be related. In the literature these are determined by a *matching function*, a market relationship between the flow of matches that form, and the number of workers and jobs seeking, a match. In this view, the matching function, denoted as $M(u, v)$, is a kind of 'production function' that relates the match inputs to match output. Given this function, it follows that $\lambda u \equiv M(u,v) \equiv \eta v$ since λu and ηv are both equal to the total match flow in the aggregate.

It is natural to suppose that the matching function, like an aggregate production function, is increasing, concave and homogenous of degree one. There is now a relatively extensive empirical literature, reviewed by Petrongolo and Pissarides (2001), which for the most part confirms these assumptions. When they hold, the vacancy-filling rate $\eta = M(u, v) / v = M(u / v, 1)$ is decreasing function of the ratio of vacancies to unemployment. Hence, if the expected return to filling a vacancy exceeds the cost of posting it, more vacancies will be created, driving down the expected return. Under the assumption of free entry, then, the market equilibrium number of vacancies posted at any point in time satisfies the free entry condition

$$\frac{c}{\eta} = \frac{c\theta}{M(1, \theta)} = J(w) = \frac{p - w}{r + \delta} \tag{3}$$

where the vacancy–unemployment ratio, $\theta = v/u$, is referred to as *market tightness*.

In summary, an equilibrium solution to the model is a wage, reservation wage, and market tightness triple (w, R, θ) that joint satisfies the eqs (1), (2), and (3). Finally, because existing jobs are destroyed at rate δ and the flow of unemployed workers who find jobs is $\lambda u = M(1, \theta) u$, the steady state value of unemployment rate, that which equate the flows in and out of the unemployment state, is

$$u = \frac{\delta}{\delta + M(1, \theta)}. \tag{4}$$

In other words, unemployment tends over time to a steady state value that increases with the rate of job destruction, δ, and decreases with market tightness, θ. Furthermore, market tightness depends on the incentive to create new jobs, the profit an employer can expect to earn in the future after the match forms. From eq. (3), it follows that labour productivity, represented by the parameter p, is a major contributor to that incentive. Indeed, a positive shock to p first increases vacancies and, consequently, market tightness. Over time unemployment falls in response until its new steady state value is realized as characterized in eq. (4). As a consequence, shocks to productivity trace out a downward sloping relationship between vacancies and unemployment, known in the empirical literature as the Beveridge

curve. The effect of productivity shocks on unemployment is amplified by the fact that the rate of job destruction, δ in the model, falls with *p*. This channel of influence is incorporated into an extended version of the formal model by Mortensen and Pissarides (1994).

The theory has clear implications for labour market policy. For example, unemployment insurance is common in all developed economies and is either enacted or under consideration in many developing countries. As unemployment benefit is income contingent on being unemployed, it can be represented in the model by the parameter *b*. From eqs (1) and (2) it follows that any increase in the benefit will raise wages, though its effect on the worker's bargaining threat point. In turn, the increase in wages will decrease future expected profit which will lead to a reduction in vacancies and market tightness according to the free entry condition (3). These facts together with eq. (4) imply that a higher unemployment insurance benefit will raise the steady state level of unemployment. By clarifying this mechanism, the theory has played an important role in the debate over labour market policy reform in Europe.

For an extensive discussion of the matching model of unemployment and its implications see Mortensen and Pissarides (1999a; 1999b), Pissarides (2000), and the recent review article by Rogerson, Shimer and Wright (2005).

Summary

The development of the search-theoretic approach to the analysis of labour markets has focused on two different issues, wage dispersion and unemployment, and the models used in each case are not fully consistent with one another. For example, in one branch wages are set by the employer while in the other wages are the outcome of a bilateral bargain between worker and employer This and other specification differences are subjects of current theoretical and empirical research designed to collect the features of each approach that best explain all the phenomena of interest. The ongoing research, designed among other purposes to integrate the two approaches, is reviewed by Rogerson, Shimer and Wright (2005).

See Also

- matching
- microfoundations
- search models of unemployment
- search theory
- search theory (new perspectives)
- social networks in labour markets

Bibliography

Burdett, K. and Judd, K. 1983. Equilibrium price distributions. *Econometrica* 51, 955–70.

Burdett, K. and Mortensen, D.T. 1998. Equilibrium wage differentials and employer size. *International Economic Review* 40, 889–914.

Devine, T.J. and Kiefer, N.M. 1991. *Empirical Labor Economics: A Search Approach*. Oxford: Oxford University Press.

Diamond, P. 1971. A model of price adjustment. *Journal of Economic Theory* 3, 156–68.

Diamond, P.A. 1982. Wage determination and efficiency in search equilibrium. *Review of Economic Studies* 49, 217–27.

Eckstein, Z. and van den Berg, G.J. 2007. Empirical labor search. *Journal of Econometrics* 136, 531–64.

Ehrenberg, R. and Oaxaca, R. 1976. Unemployment insurance, duration of unemployment, and subsequent wage gain. *American Economic Review* 66, 754–66.

Gronau, R. 1971. Information and frictional unemployment. *American Economic Review* 61, 290–301.

Kiefer, N.M. and Neumann, G.R. 1979. An empirical job search model with a test of the constant reservation wage hypothesis. *Journal of Political Economy* 87, 89–107.

McCall, J.J. 1970. Economics of information and job search. *Quarterly Journal of Economics* 84, 113–26.

Mortensen, D.T. 1970. Job search, the duration of unemployment and the Phillips curve. *American Economic Review* 60, 847–62.

Mortensen, D.T. 1982. The matching process as a noncooperative/bargaining game. In *The Economics of Information and Uncertainty*, ed. J.J. McCall. Chicago: University of Chicago Press.

Mortensen, D.T. 2003. *Wage Dispersion: Why are Similar People Paid Differently*? Cambridge, MA: MIT Press.

Mortensen, D.T. and Pissarides, C.A. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61, 397–415.

Mortensen, D.T. and Pissarides, C.A. 1999a. Job reallocation, employment fluctuations, and unemployment differences. In *Handbook of Macroeconomics*, ed. M. Woodford and J. Taylor. Amsterdam: North-Holland.

Mortensen, D.T. and Pissarides, C.A. 1999b. New developments in models of search in the labor market.

In *Handbook in Labor Economics*, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Petrongolo, B. and Pissarides, C.A. 2001. Looking into the black box: a survey of the matching function. *Journal of Economic Literature* 38, 390–431.

Pissarides, C.A. 1985. Short-run equilibrium dynamics of unemployment, vacancies and real wages. *American Economic Review* 75, 676–90.

Pissarides, C.A. 2000. *Equilibrium Unemployment Theory*, 2nd edn. Cambridge, MA: MIT Press.

Rogerson, R., Shimer, R. and Wright, R. 2005. Search-theoretic models of the labor market: a survey. *Journal of Economic Literature* 43, 959–88.

Rothschild, M. 1973. Models of market organization with imperfect information: a survey. *Journal of Political Economy* 81, 1283–308.

Stigler, G. 1961. The economics of information. *Journal of Political Economy* 69, 213–25.

Wolpin, K.I. 1995. *Empirical Methods for the Study of Labor Force Dynamics*. Luxembourg: Harwood Academic.

How to cite this article

Mortensen, Dale. "labour market search." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000220> doi:10.1057/9780230226203.0917

The New Palgrave Dictionary of Economics Online

labour supply

Richard Blundell and Thomas MaCurdy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The analysis of labour supply is placed in a general framework within which empirical models and their resulting elasticity estimates can be interpreted. An explicitly intertemporal life-cycle structure is developed for the choice of hours and participation. The relationship between economic substitution effects found in the labour supply literature and wage impacts on different concepts of employment is considered. We provide a separate discussion of the main issues surrounding the analysis of family labour supply and the analysis of the impact of taxation. We conclude with a discussion on the interpretation of labour supply elasticities for policy analysis.

Keywords

benefit take-up; collective models of the household; cost functions; dynamic programming; employment; Engel curve; Euler equations; Frisch specification; Hicksian effect; hours worked; indirect utility function; labour supply; linear expenditure system; Marshallian effect; optimal taxation; reservation wage; retirement; Slutsky effect; tax credits

Article

The formal analysis of labour supply in economic research extends back to the 1960s, in the work of Becker (1965), Cain (1966), Hanoch (1965) and Mincer (1960), among others. It was developed further in the 1970s, most importantly in the work of Ashenfelter and Heckman (1974), Burtless and Hausman (1978), Gronau (1974) and Heckman (1974a). It would seem reasonable to ask why interest continues in the study of labour supply and what unanswered questions and puzzles remain.
Policy interest in labour supply continually motivates research on all aspects of the subject. One area of active inquiry evaluates the consequences of the new ideas in tax and welfare reform, especially those related to the growing focus on work requirements in the design of welfare reform and on the supply of effort by top-rate tax payers. Another important topic concerns the impacts of reforms of pension and

health-care systems on labour supply decisions in later life. Yet another involves gender inequality and the role of female labour supply in removing gender earnings differences and in supporting family incomes. If in addition to these policy motivations, understanding hours-of-work behaviour lies at the heart of explaining the reasons underlying a variety of key trends in the economy. One is the unprecedented growth in female labour supply across many developed economies since the 1970s; a second is the decline in labour supply among older men over the same period, again a phenomenon common to many developed economies; and a third is the labour supply impact of the growth in the disparity between the labour market returns of the educated and those with little formal training. Add to these questions the importance of labour supply in understanding employment over the business cycle and over the life cycle, and it becomes clear why labour supply has maintained a prominent position in economic research.

Having established its importance, what does the study of labour supply involve? Although the parameter (s) of interest in a labour supply model may seem obvious, on closer inspection it is not so clear-cut. We are typically interested in examining the reaction of labour supply to a change in the wage. But what measure of labour supply and what measure of the wage? Is it employment – the extensive margin – or hours of work for workers – the intensive margin – that is of key interest? Is it the impact that of an anticipated change in the wage or an unanticipated change in the wage? Are we simply concerned with individual labour supply or does family labour supply matter too?

What labour supply elasticity should be used? The wealth of empirical studies on labour supply has produced a plethora of estimated elasticities and response parameters. Differences between estimates can often be attributed to data measurement issues but, as documented in Blundell and MaCurdy (1999), more often than not, a large component of the differences can be explained by the economic framework within which each the estimates is derived. Apart from hourly wages and other income, are controls for lifetime wages included? What about expected changes in other income sources? The precise conditioning variables included in a labour supply model critically change the interpretation and therefore the comparability of estimated elasticities and response coefficients. It is also clear that labour supply responses differ according to the extensive or intensive margin, especially for women. To understand differences across these margins, the specification of effective budget constraints and the nature of fixed costs matter. For men it may well be that the retirement margin could be a margin of growing importance.

An important role of a review of this type is to provide a coherent framework within which different labour supply models can be compared. It is clearly useful to have an explicitly intertemporal framework, although, as we shall see, perfectly interpretable estimates of some important parameters of interest can be recovered from models that look essentially static. Much of the difference across empirical models reflects differences in data availability, and this provides another argument for this approach. The precise form of income, hours or wage variables available will vary wildly across data sources, but this does not necessarily imply incomparable results. Some data provides longitudinal information on individual wages and hours; other data is repeated cross-section but may have more detailed information on asset or consumption levels.

To set the scene we start with a brief discussion of the standard ‘static’ labour supply model. We then go on to ask what is meant by employment and how one translates estimates of economic substitution effects found in the labour supply literature into wage impacts relevant for the employment concept. Next we look at the extension to a life-cycle setting. The objective is always to present a framework

within which empirical models and their resulting elasticity estimates can be interpreted. We provide a separate discussion of the main issues surrounding the analysis of family labour supply and for analysing the impact of taxation and welfare reform. If the literature in respect to all of these topics is too rich to include all of the key references in the text, but a list of some of the leading references is provided at the end of this review. The review ends with a discussion on the interpretation of labour supply elasticities for policy analysis.

1 Setting the scene

In the standard labour supply model as applied to individual decisions at a point in time, choices are made over consumption and leisure hours. In each period of time t each individual i, defined by characteristics ν_{it}, has preferences over consumption and leisure hours described by a (within) period utility

$$U(c_{it}, l_{it}; \nu_{it}) \tag{1}$$

in which c_{it} and l_{it} are within-period consumption and leisure hours respectively. (The important extension to family labour supply is considered below.) The elements of the vector ν_{it} alter preferences, both through observed characteristics of the individual and through this person's unobserved factors influencing ‘tastes’. This utility is assumed to be maximized subject to the budget constraint

$$c_{it} + w_{it}l_{it} = y_{it} + w_{it}T \tag{2}$$

in which w_{it} is the hourly wage rate, y_{it}, is non-labour income and T is the total time available for work and leisure.

Non-labour income is made up of two components: asset income and other unearned income. Assuming beginning of period assets, denoted A_{it}, earn a return r_{it} during period t, the former is $r_{it}A_{it-1} + \Delta A_{it}$, in which ΔA_{it} denotes capital gains. Other unearned income is primarily benefit or transfer income and denoted g_{it}. The r.h.s. of (2) is often defined as ‘full income’ and we denote this income concept as m_{it} throughout, so that

$$m_{it} = y_{it} + w_{it}T. \tag{3}$$

First-order conditions take the familiar form:

$$U_c(c_{it}, l_{it}; v_{it}) = \lambda_{it} \tag{4}$$

and

$$U_l(c_{it}, l_{it}; v_{it}) \geq \lambda_{it} w_{it} \tag{5}$$

where λ_{it} is the marginal utility of income. The inequality in (5) determines the reservation wage rule for labour market participation.
Solving for λ_{it} using the budget constraint (2) yields the (Marshallian) decision rule and

$$l_{it} = l(w_{it}, m_{it}; v_{it}) \leq T_{it} \tag{6}$$

where m_{it} is full-income defined in (3) above. Equivalently we have the hours of work rule

$$h_{it} \begin{cases} = h^s(w_{it}, y_{it}; v_{it}) \ \forall \{w_{it}, y_{it}; v_{it}\} \ni U_L(c_{it}, l_{it}; v_{it}) \geq w_{it} U_c(c_{it}, l_{it}; v_{it}) \\ = 0 \ otherwise \end{cases} \tag{7}$$

where y_{it} is defined as in (2).
Preferences over hours of work can, of course, be written analogously to direct utility (1) as

$$U(y_{it}, T - h_{it}; v_{it});$$

(8)

or by the expenditure (that is, cost) function

$$\xi_{it} = \xi(w_{it}, V_{it}; v_{it});$$
(9)

or by the indirect utility function

$$V_{it} = V(w_{it}, y_{it}; v_{it}).$$
(10)

The expenditure function solves the problem

$$\xi_{it} = \xi(w_{it}, V_{it}; v_{it}) = \min c_{it} + w_{it}(T - h_{it}) \textit{subject to} V = U(y_{it}, T - h_{it}; v_{it}, \theta);$$
(11)

and the indirect utility inverts the expenditure function to obtain a solution for V_{it}. Whether analysis is conducted with the direct utility, expenditure function, indirect utility or the labour supply equation will depend largely on the approach to estimation.

The inequality (7) represents a corner solution for hours of work and can be stated as a reservation wage condition for participation $w_{it} \geq w_{it}^{*}$, where w_{it}^{*} is derived by inverting $h^{S}(w_{it}, y_{it}; v_{it}) = 0$. The key econometric problem that follows from this corner solution is that *w* will not be observed when *h*=0. Consequently a specification for wages is also required and together they create the *selection problem* addressed by Gronau (1974) and Heckman (1974a, 1979).

1.1 Substitution and income effects

In a static framework the literature typically cites two types of substitution effects when describing how labour supply responds to changes in the wage rate. First, the uncompensated (or Marshallian) effect refers to the following derivative of labour supply function (7):

$$\frac{\partial h^s}{\partial w} \tag{12}$$

which holds non-labour income y_{it} constant when measuring how much hours of work respond to a shift in wages. If second, one can derive an expression for the compensated labour supply function by computing the derivative of the expenditure function ξ_{it} with respect to w_{it}, and then constructing a function defined as T minus this derivative. This compensated function holds utility constant, and its derivative with respect to w_{it} measures the compensated (or Slutsky or Hicksian) effect. A familiar relationship linking compensated and uncompensated substitution effects is the Slutsky decomposition given by:

$$\left.\frac{\partial h^s}{\partial \omega}\right|_u = \frac{\partial h^s}{\partial w} + h\frac{\partial h^s}{\partial y}, \tag{13}$$

where the derivative $\frac{\partial h^s}{\partial y}$ shows the impact of changing income on hours of work holding wages constant.
Regular integrability conditions from optimization theory imply that the compensated substitution effect is non-negative

$$\left.\frac{\partial h^s}{\partial \omega}\right|_u \geq 0. \tag{14}$$

In sharp contrast, the compensated effect $\frac{\partial h^s}{\partial w}$ can be negative or positive depending on the strength of the income effect on labour supply. When $\frac{\partial h^s}{\partial w}$ is negative labour supply is said to be ‘backward bending’.

1.2 Empirical evidence

The empirical analysis of the standard labour supply model described here tends to distinguish

individuals by gender and by whether there are children at home, finding rather different elasticities across these groups (see Johnson and Pencavel, 1984). Allowing for a separate impact of the way the market wage affects the employment and the hours decision has proven to be essential. This partly reflects fixed costs of work and the workings of the welfare system, to be discussed below, but it also highlights the strong evidence that labour supply responses at the extensive margin dominate those at the intensive margin; see Blundell and MaCurdy (1999) for a review of this evidence.

1.3 Some popular labour supply specifications

In discussing particular specifications it is useful to be able to move between all three representations of preferences over labour supply (8)–(10). For example, if the focus is on taxation and welfare participation it is typical to express decisions as a multinomial choice problem over discrete hours choices and work with the direct utility specification. This will be discussed below.
To complete this brief review of the standard labour supply model we consider four popular specifications. The *linear expenditure system* assumes the direct utility function

$$U(c_{it}, T - h_{it}; v_{it}) = \beta_h(v_{it}) \ln [T - h_{it} - \gamma_h(v_{it})] + \beta_c(v_{it}) \ln [c_{it} - \gamma_c(v_{it})], \tag{15}$$

where the notation $\beta_h(\nu_{it})$, $\beta_c(\nu_{it})$, $\gamma_h(\nu_{it})$ and $\gamma_c(\nu_{it})$ indicates that the preference parameters β_h, β_c, γ_h and γ_c are functions of individual attributes ν_{it} and therefore can vary across members of the population. (Imposing the restriction $\beta_h(\nu_{it})+\beta_c(\nu_{it})=1$ identifies these coefficients.) Abstracting from the dependence on heterogeneous tastes v_{it}, the expenditure function (9) implied for the *linear expenditure system* takes the form:

$$\xi(w, V) = \gamma_h w + \gamma_c + w^{\beta_h} V;$$

and the uncompensated labour supply function is:

$$h^s(w, y) = T - \gamma_h - \frac{\beta_h}{w}(m - \gamma_h w - \gamma_c). \tag{16}$$

A second popular preference specification is the *linear labour supply*

$$h = \alpha + \beta w + \gamma y \tag{17}$$

(for example, see Hausman, 1981; 1985a), which comes from the indirect utility function:

$$V(w, y) = e^{\gamma w}\left(y + \frac{\beta}{\gamma}w - \frac{\beta}{\gamma^2} + \frac{\alpha}{\gamma}\right) \text{with } \gamma \leq 0 \text{ and } \beta \geq 0. \tag{18}$$

Note that since $\partial h / \partial y = \gamma > 0$, the Slutsky condition (13) all but requires $\beta > 0$, ruling out backward bending labour supply. It is arguable that this linear specification allows too little curvature with wages. Alternative *semilog specifications* and their generalizations are also popular in empirical work. For example, the semilog specification

$$h = \alpha + \beta \ln w + \gamma y \tag{19}$$

with indirect utility

$$V(w, y) = \frac{e^{\gamma w}}{\gamma}(\alpha + \beta \ln w + \gamma y) + \frac{\beta}{\gamma}\int_{-\gamma w} \frac{e^{-t}}{t}\, dt \text{ with } \gamma \leq 0 \text{ and } \beta \geq 0. \tag{20}$$

Moreover, the linearity of (19) in α and ln *w* makes it particularly amenable to an empirical analysis with unobserved heterogeneity, endogenous wages and non-participation as discussed below (see Blundell, Duncan and Meghir, 1998).

Neither (17) nor (19) allows backward bending labour supply behaviour, although it is easy to generalize (19) by including a quadratic term in ln *w*. Note that imposing integrability conditions at zero hours for either (17) or (19) implies positive wage and negative income parameters. A simple specification that does allow backward bending behaviour, while retaining a three parameter linear in variables form, is

that used in Blundell, Duncan and Meghir (1992):

$$h = \alpha + \beta \ln w + \gamma \frac{y}{w} \tag{21}$$

with indirect utility

$$V(w, y) = \frac{w^{1+\gamma}}{1+\gamma}\left(\alpha - \frac{\beta}{1+\gamma} + \beta \ln w + (1+\gamma)\frac{y}{w}\right) \text{with } \gamma \leq 0 \text{ and } \beta \geq 0; \tag{22}$$

see Stern (1986). This form has similar properties to the specification of Heckman (1974). Further empirical specifications are described in Blundell, MaCurdy and Meghir (2007), where the econometric issues of dealing with the extensive margin and missing wages are discussed in detail.

2 The impact of wages and income on hours of work and employment

Addressing many of the questions asked by policymakers about labour supply involves evaluating the extent to which employment in a population can be expected to change in response to a shift in the returns to work. Relying on existing empirical work to answer such questions requires resolution of two issues: (1) what is meant by employment?; and (2) how does one translate estimates of economic substitution effects found in the labour supply literature into wage impacts relevant for the relevant employment concept?

2.1 Three concepts of employment and labour supply

There are three distinct concepts of labour supply or expected hours of work, which are often confused in the literature. Consider a population of consumers all of whom receive a common wage *w* and non-labour income *y*, but who have different tastes ν_{it}'s. Let the density function $f(\nu)$ denote the distribution of ‘preferences for work’ over the population.
One measure of labour supply is the fraction of the population who works:

$$P(w, y) = \Pr(h^S(w, y, \nu_{it}) > 0) = \int_{\Theta} f(\nu)\, d\nu \text{ where } \Theta = \left\{\nu_{it} : h^S(w, y, \nu_{it}) > 0\right\}. \tag{23}$$

A second concept is the average hours worked among those employed:

$$E(h^s(w_{it}, y_{it}; v_{it}) | h^s_{it} > 0) = \frac{\int_{\Theta} h^s(w, y, v_{it}) f(v)\, dv}{P(w, y)}. \tag{24}$$

Yet a third measure of labour supply is the average hours worked in the entire population:

$$E(h^s(w_{it}, y_{it}; v_{it})) = \int_{\Theta} h^s(w, y, v_{it}) f(v)\, dv. \tag{25}$$

While these three measures of labour supply depend on many of the same parameters, they are clearly distinct concepts. If a researcher is interested in the effect of wages on employment, then the derivative of (23) with respect to *w* measures the appropriate quantity. If, instead, one wants to know how much an increase in the wage rate affects total aggregate hours of work, then the derivative of (25) with respect to *w* gives the relevant measure.

There is also some confusion in the literature concerning the appropriate interpretation of the partial derivatives of these different measures of labour supply. The partial derivatives of the hours of work function given by (7), h^s_w and h^s_y, produce the textbook uncompensated wage and income effects. Casual inspection of (23) reveals that the derivatives of $P(w, y)$ with respect to *w* and *y* do not correspond to h^s_w and h^s_y (Lewis, 1967; Ben-Porath, 1973). Whereas P_w must be positive, h^s_w need not be. Moreover, the partial derivatives of (24) or (25) with respect to *w* and *y* do not correspond to the uncompensated substitution and income effects, h^s_w and h^s_y, unless the inequality condition (7) is satisfied for everyone in the population and the labour supply function h^s takes a special form. These simple points have been ignored in much of the literature. For example, Hall (1973) and Boskin (1973) interpret the partial derivative of estimates of eq. (25) with respect to *w* and *y* as estimates of h^s_w and h^s_y respectively. Others interpret partial derivatives of (24) (estimated from labour supply functions fit on samples of working individuals) as estimates of the Marshallian–Hicks–Slutsky parameters. If non-participation is a significant phenomenon in the population being sampled, estimates of (23), (24) nor (25) do not generate meaningful structural labour supply parameters.

2.2 Aggregate labour supply

Conditions have been established for utility functions that enable one to aggregate micro labour supply functions to obtain economically meaningful market functions. Satisfaction of these conditions implies equivalency of micro and macro substitution effects. In the case when consumers face a common set of prices and have different incomes, Gorman's (1961; 1976) seminal contributions specify those sets of preference consistent with linear Engel curves, which he shows are required properties of preference to carry out exact aggregation of micro demand functions to macro formulations. The macro specification is a ‘representative consumer’ version of the original individual preference relationship. Gorman's conditions are insufficient for aggregation of labour supply functions since wages, in contrast to prices, vary considerably across individuals in any interesting empirical application. Muellbauer (1981) refines Gorman's aggregation conditions to apply to the labour supply case allowing for wages along with income to different across individuals.

For a market labour supply function to have a form consistent with the underlying micro specifications aggregated to derive its construction, the expenditure function (9) must necessarily take the general form:

$$\xi(w_{it}, V_{it}; \nu_{it}) = \alpha_t(\nu_{it}) + w_{it}\beta_t + w_{it}^{\delta} b_t V_{it}. \tag{26}$$

(Inspection of the specification – the equation above eq. (16) – for $\xi(w_{it}, V_{it}; \nu_{it})$ for the *linear expenditure system* reveals that it has the form required by (26) when $\beta_h(\nu_{it}) = \beta_h$, $\beta_c(\nu_{it}) = \beta_c$, and $\gamma_h(\nu_{it}) = \gamma_h \ \forall \ \nu_{it}$.) The uncompensated labour supply function implied by (26) is given by:

$$h^s(w_{it}, y_{it}; \nu_{it}) = \pi_t - \frac{\delta}{w_{it}}(y_{it} - \alpha_t(\nu_{it})) \tag{27}$$

where

$$\pi_t = (1 - \delta)(T - \beta_t). \tag{28}$$

In this specification, only the preference components $\alpha(\nu_{it})$ can vary across individuals in the static

setting. Rather than expressing this relationship as hours of work, one typically finds (9) it written as the earning function:

$$w_{it}h_{it}^{S} = \pi_t w_{it} - \delta y_{it} + \delta\alpha_t(v_{it}). \tag{29}$$

Given its linear structure, one clearly sees that estimation of the micro and aggregate substitution and income effects corresponds to the same preference parameters. Viewed in a pooled cross-section time-series context, the preference components α_t, β_t, and b_t typically will be functions of prices in period t which are common across individuals in the cross section corresponding to the period, but these prices do change over time. To create a valid form for preferences, the α_t and β_t must be homogeneous of degree 1 in prices, and b_t must be homogeneous of degree zero.

What concept of labour supply does this aggregate relationship represent? In a world where everyone works, the average of (27) corresponds to both the expected values of hours worked among the employed (24) and overall populations (25); after all, these are exactly the same samples. Moreover, the economic concept of the uncompensated substitution effect directly measures the response one would estimate using an empirical specification based on either eq. (24) or eq. (25).

These nice relationships, however, entirely break down when one recognizes that the employment decision is typically influenced by a change in wages, be it across people or a shift in the distribution that occurs over time. With the no-work/work decision being affected for some people, impacts now critically depend on the properties of distribution of preferences determined by the density function $f(\nu)$, which could itself shift over time. The effects of wages on the three concepts of labour supply given by (23), (24) and (25) again become distinct, and none directly measures the economic notions of substitution effects outlined above. When labour market participation is a choice in the population, *no* conditions exist for consistently aggregating micro labour supply function to obtain a macro function that can be given a coherent ‘representative agent’ interpretation. Substitution effects estimated in an aggregate setting cannot be interpreted coming from a single agent-optimizing framework, and the wage effects estimated from micro data considered alone will typically provide insufficient information to project aggregated impacts.

3 Labour supply over the life cycle

Although its study is often placed in an effectively static framework as in (1) and (2), labour supply is clearly part of a lifetime decision-making process. Individuals attend school early in life, accumulate wealth while in the labour force, and make retirement decisions late in life; each of these activities can only be understood in a life-cycle framework. We know that savings from labour earnings are often required to sustain individuals, or their dependants, during periods when they are out of the labour market. In addition, variations in health status, family composition and real wages provide incentives for individuals to vary the timing of their labour market earnings for income-smoothing and insurance

purposes.
To keep things simple we assume life-cycle utility at time *t* has the form

$$u_{is} = E_t\left\{\sum_{t=s}^{L} \frac{1}{1+\delta_t} U(c_{it}, l_{it}; v_{it})\right\}$$
(30)

in which E_t is the expectations operator conditional on information up to and including period *t* and where δ_t is the subjective discount rate. Maximization of (30) takes place subject to at intertemporal budget constraint. For this we need to write down the path of assets:

$$A_{it+1} = A_{it} + r_t A_{it} + b_{it} + w_{it} h_{it} - c_{it}$$
(31)

where A_{it} is the assets held at the beginning of period *t* and r_t is the return on assets earned in period *t*. The form of life-cycle preferences and of the budget constraint in (30) and (31) is not innocuous. The time-separability of (30) rules out habits and slow adjustment. The *rA* term in (31) assumes that individuals can borrow and lend via the simple credit market at rate *r* and consequently rules out borrowing constraints. Nevertheless, under these assumptions the first-order conditions (4) and (5) continue to hold and to determine within-period allocations of time and consumption. Intertemporal allocations are determined through the choice of the marginal utility of consumption λ_t in (4). Consequently allocations over the life cycle will be summarized through the evolution of λ_t.
To understand these conditions in an inter-temporal context we can use the knowledge that λ_{it}, the marginal utility of wealth, evolves over time according to

$$\lambda_{it} = \frac{1}{1+\delta_t} E_t\{\lambda_{it+1}(1 + r_{it})\}$$
(32)

where the real interest rate r_{it} is allowed to be stochastic. Relationship (32) is often referred to as the stochastic Euler equation (see Hansen and Singleton, 1983).

3.1 Frisch (λ -constant) labour supply equations

Frisch, or marginal-utility-of-wealth ‘ λ ’ constant, labour supply functions provide an extremely useful method for analysing life-cycle maximization problems (see Browning, Deaton and Irish, 1985). In this framework, the marginal utility of wealth, λ , serves as the sufficient statistic which captures all information from other periods that is needed to solve the current-period maximization problem. The time-separable form of the utility maximizing model implies that the marginal within-period decisions depend on the past and future through the single ‘sufficient statistic’ λ_{it}. Even though the marginal utility of wealth λ_{it} is not observable to the empirical economist, the rule for its evolution (32) enables a method of moments estimation of the labour supply parameters.

To briefly see how estimation takes place in this framework, consider the simple parametric form for preferences chosen in MaCurdy (1981). The utility specification MaCurdy used does not allow for corner solutions and takes the form

$$U_t = \theta_t c_t^{\gamma} - \varphi_t h_t^{\alpha} \quad 0 < \gamma < 1, \alpha > 1 \tag{33}$$

where h_t corresponds to hours of work and c_t to consumption. The range of parameters ensures positive marginal utility of consumption, negative marginal utility of hours of work and concavity in both arguments. The Frisch labour supply is

$$\log h_t = \theta_t^* + \log \lambda + \frac{1}{\alpha - 1} \ln w_t + \frac{\rho - r}{\alpha - 1} t \tag{34}$$

where the use of log hours of work presumes that all individuals work and hence $h>0$. In (34) λ is the shadow value of the lifetime budget constraint and t is the age of the individual. Finally θ_t^* reflects preferences and is defined by $\theta_t^* = -\frac{1}{\alpha - 1} \log \theta_t$. This equation has a simple message: Hours of work are higher at the points of the life cycle when wages are high ($\frac{1}{\alpha - 1} > 0$). Moreover if the personal discount rate is lower than the interest rate, hours of work decline over the life cycle. Finally, hours of work will vary over the life cycle with θ_t^*, which could be a function of demographic composition or other taste shifter variables.

The MaCurdy (1981) paper set out the first analysis of issues to do with estimating intertemporal labour supply relationships. However the approach did not deal with corner solutions and the extensive margin,

which is particularly relevant for women. The first attempt to do so, in the context of a life-cycle model of labour supply and consumption is the paper by Heckman and MaCurdy (1980). In this model women are endowed with an explicitly additive utility function for leisure *l* and consumption *c* in period *t*, of the form:

$$U_t = \theta_t \frac{l_t^{\alpha} - 1}{\alpha} + \varphi_t \frac{c_t^{\gamma} - 1}{\gamma} \alpha, \gamma < 1. \tag{35}$$

Optimization is assumed to take place under perfect foresight. Solving for the first-order conditions we obtain the following equation for leisure

$$\ln l_t \begin{cases} = \theta_t^* + \frac{1}{\alpha - 1} \ln w_t + \frac{\rho - r}{a - 1} t + \lambda^* \textit{ when the woman works} \\ = \ln \bar{l} \textit{ otherwise} \end{cases} \tag{36}$$

where

$$\lambda^* = \frac{1}{\alpha - 1} \ln \lambda \text{ and } \theta_t^* = -\frac{1}{\alpha - 1} \ln \theta_t. \tag{37}$$

3.2 Two-stage budgeting and Marshallian labour supply equations

In this time-separable optimizing problem there are alternative ‘sufficient statistics’ to the marginal utility of wealth that completely summarize the past and future as it impacts on the period *t* labour supply decision. From Gorman (1959; 1968), intertemporal separability implies that the decision rule can be thought of in two stages. First allocate to period *t* according to

$$m_{it} = M(w_{it}, y_{it}, A_{it-1}, r_t, v_{it}, z_{it}) \tag{38}$$

where z_{it} represents the information used to form expectations of future real wages and other household attributes that are uncertain at time t. At the second stage, given m_{it}, the within-period first-order conditions (4) and (6) remain valid. Moreover, the estimation of 'm-conditional' labour supply functions are robust to liquidity constraints and other capital market imperfections.

3.3 Marginal rate of substitution equations

Eliminating λ_{it} from the first-order conditions (4) and (6) yields the marginal rate of substitution function

$$MRS_l(c_{it}, l_{it}; v_{it}) \geq w_{it} \tag{39}$$

where

$$MRS_l(c_{it}, l_{it}; v_{it}) = \frac{U_l}{U_c}. \tag{40}$$

Again, (39) is robust to liquidity constraints and other capital market imperfections. As we know from our general discussion of elasticities, the constant marginal utility of wealth (Frisch) elasticity is greater than the Slutsky-compensated (within-period) elasticity which is again greater than the standard uncompensated Marshallian elasticity, see Blundell (1998).

3.4 Relationships among the life-cycle elasticities

The Frisch specification treats the individual marginal utility of wealth as a 'fixed effect' and allows the researcher to estimate only the intertemporal substitution elasticity. Given that appropriate methods are employed to account for the fixed effect (generally first differencing in panel data), the relevant independent variables, apart from the wage, are simply within-period characteristics and age. The Frisch elasticity, by ignoring this (unexpected) shift in wealth from a once-and-for-all change in real wages, is larger than the policy-relevant elasticity and overestimates the impact of a reform.
Direct estimation of the simple parameterization of the full life-cycle model, required to recover policy-relevant elasticity, relies on specifications for both within-period utility and the individual marginal-utility-of-wealth effect. As a result, controls are needed for all of the following: 'start of life'

characteristics, current-period characteristics which affect the within-period utility function, age, expected wages and initial wealth. Expected wages are typically unobservable and initial wealth is generally not included in data-sets, so these should be replaced with the parameters governing the time path of wages and property income, which must be jointly estimated with the labour supply equation. Estimation of this full framework allows computation of both the intertemporal substitution elasticity and the elasticity of labour supply in reaction to a full, parametric wage profile shift. However, it is also the most demanding in terms of data.

It is worth noting that the elasticity derived from the static specification which uses unearned income to compute virtual income can be placed in an intertemporal setting but is economically meaningful only under a strong assumption of either complete myopia or perfectly constrained capital markets. Otherwise, this elasticity confuses movements along wage profiles with shifts of these profiles and, thus, yields response parameters which are a mixture of these. Such hybrid estimates lack an economic interpretation and are not generally useful in policy evaluation.

To illustrate the challenges encountered with inferring the different substitution effects from one another, consider a life-cycle extension of the *linear expenditure system* (LES) in a deterministic setting. A multi-period expansion of the static LES utility function given by (15) takes the form

$$u = \sum_{t=1}^{T} \varphi_t \cdot U(c_{it}, l_{it}; v_{it}) = \sum_{t=1}^{T} \varphi_t [\beta_h \ln(T - h_t - \gamma_h) + \beta_c \ln(c_{it} - \gamma_c)], \tag{41}$$

where the normalization $\Sigma_{t=1}^{T} \varphi_t = 1$ (in addition to $\beta_h + \beta_c = 1$) identifies preference parameters. The specification implied for the life-cycle uncompensated labour supply function for hours of work in period *t* is:

$$h_t^S(\omega, R, M; v) = T - \gamma_h - \frac{\varphi_t \beta_h}{\omega_t}\left(M - \sum_{k=1}^{T} \gamma_h \omega_k - \sum_{k=1}^{T} \gamma_c R_k\right) \tag{42}$$

where the quantities ω_t denote the discounted value of the period-*t* wage rate; R_t represents the discounted price of consumption in period *t*; and *M* designates the ‘full income’ equivalent of the individual's wealth. The period-*t* marginal-utility-of-wealth ‘ λ ’ constant labour supply function takes the form:

$$h_{it} = h^{\lambda}(\omega_{it}, R_{it}, \lambda_{it}) = T - \gamma_h + \frac{\varphi_t \beta_h}{\lambda_{it}\omega_{it}}. \tag{43}$$

Accordingly, the uncompensated substitution effect associated with a change in wage rate ω_t on hours of work h_t is given by:

$$\frac{\partial h_t^S}{\partial \omega_t} = \frac{\varphi_t \beta_h}{\omega_t^2}\left(\gamma - \sum_{k=1}^{T} \gamma_h \omega_k - \sum_{k=1}^{T} \gamma_c R_k + \gamma_h \omega_t\right) = \frac{(T - \gamma_h)(1 - \varphi_t \beta_h)}{\omega_{it}} - \frac{h_t}{\omega_t}; \tag{44}$$

and the intertemporal substitution effect corresponding to change in ω_t on h_t is:

$$\frac{\partial h_t^{\lambda}}{\partial \omega_t} = -\frac{\varphi_t \beta_h}{\lambda_{it}\omega_{it}^2} = \frac{(T - \gamma_h)}{\omega_{it}} - \frac{h_t}{\omega_t}. \tag{45}$$

The following relationship links these two hour-of-work responses:

$$\frac{\partial h_t^{\lambda}}{\partial \omega_t} = \frac{\partial h_t^S}{\partial \omega_t} + \frac{(T - \gamma_h)\varphi_t \beta_h}{\omega_{it}}. \tag{46}$$

Finally, if one were to estimate an uncompensated substitution effect relying on a two-stage-budgeting variant of a labour supply function based on LES utility function (41), then one would compute values for:

$$\frac{\partial h_t^S}{\partial \omega_t} = \frac{\beta_h}{\omega_t^2}(\gamma - \gamma_h) = \frac{(T - \gamma_h)(1 - \beta_h)}{\omega_t} - \frac{h_t}{\omega_t}.$$

(47)

While inspection of these expressions not surprisingly reveals that the different substitution effects depend on common preference parameters, it also clearly indicates that one must exercise serious caution when attempting to infer values of one type of elasticity from any of the others. Relationship (46) shows that how one can vary endowments and preferences to change intertemporal substitution effects while not changing the uncompensated response. Of course, the above discussion has already described the additional complications encountered in any attempt to relate these economic notions of substitution effects to concepts of labour supply relevant for market measures of wage impacts on employment and hours of work which are the core concepts required for policy analyses.

3.5 Retirement and pension incentives

The study of retirement incentives and labour supply has typically focused on the dynamic effects of benefit entitlement that occur in many pension and social security schemes (Hurd and Boskin, 1984). This has resulted in the more formal use of dynamic programming tools; see Blau (1994) and Rust and Phelan (1997), for example. An important area for current research is the incorporation of these incentives into a life-cycle labour supply model.

4 Family labour supply

For the purposes of this discussion we are concerned with a family or household as comprising two working-age individuals, referred to as husband and wife below. These are the decision-making individuals in the family. Families with a single parent are subsumed in the discussion of the regular labour supply model. The central issue then becomes one of the mechanism whereby labour supply decisions are made within the household. Are they taken in a fully coordinated way as if by a single decision maker – the unitary model – or are they the result of some collective bargain – the collective model?

4.1 The unitary model of family labour supply

Suppose we can take a family or household as being made up of two working-age individuals, referred to as husband and wife below. Children and any other dependants will be included in the vector of observable household characteristics ν_{it}. For such a household, within period utility may be written

$$U_{it} = U(c_{it}, l_{it}^{h}, l_{it}^{w}; v_{it})$$
(48)

and budget constraint

$$c_{it} + w_{it}^{h} l_{it}^{h} + w_{it}^{w} l_{it}^{w} = m_{it}$$
(49)

where w_{it}^{h} and w_{it}^{w} refer to the hourly wage of the husband and wife respectively.
The marginal conditions for the λ-constant (Frisch), Marshallian and marginal rate of substitution labour supply equations described in the previous section follow naturally from the first-order conditions

$$U_c(c_{it}, l_{it}^{h}, l_{it}^{w}; v_{it}) = \lambda_{it},$$
(50)

$$U_h(c_{it}, l_{it}^{h}, l_{it}^{w}; v_{it}) \geq \lambda_{it} w_{it}^{h}$$
(51)

and

$$U_w(w_{it}, l_{it}^{h}, l_{it}^{w}; v_{it}) \geq \lambda_{it} w_{it}^{w}$$
(52)

where the subscripts *h* and *w* refer to derivatives with respect to the non-market hours of husband and wife respectively. See Ashenfelter and Heckman (1974), Wales and Woodland (1976) and Blundell and Walker (1982), for example.
Notice that there is still only a single marginal utility of wealth λ_{it} and therefore the extension to the life-cycle framework of the previous section is straightforward. There remains only one life-cycle condition (32). Consequently allocations to each individual in this time-separable model satisfy equality of marginal utility of wealth; see Blundell and Walker (1986), for example.

4.2 Collective family labour supply

The advantages of the unitary model are well known: it allows the direct utilization of consumer theory, recovering preferences from observed behaviour in an unambiguous way, and provides a coherent intertemporal framework for interpretation of empirical results. An argument against this approach is that it treats individuals in the family as a single decision-maker rather than as if they were a collection of individuals. Although true, this can be weakened through a simple decentralization argument. Suppose we let c^h and l^h refer to the private consumption of the husband and his own leisure time respectively. Defining the private consumption of the wife in the same way, we may write the within-period household utility as

$$U(c_{it}, l_{it}^h, l_{it}^w; v_{it}) = \tilde{U}(F_h(c_{it}^h, l_{it}^h; v_{it}), F_w(c_{it}^w, l_{it}^w; v_{it})) \tag{53}$$

where $F_h(c_{it}^h, l_{it}^h; v_{it})$ is the sub-utility for the husband and $F_w(c_{it}^w, l_{it}^w, v_{it})$ is the sub-utility of the wife. Family utility has a ‘weakly separable’ form and decentralization follows: allocations of total household (full) income are made between each household member and then individuals act as if they are making their labour supply and consumption decisions conditional on this initial-stage outlay. Of course, even if consumption goods are privately consumed, they are typically only measured at the household level – so that the individual consumptions are ‘latent’ to the economist.

So what is it that collective models offer? They effectively relax the income allocation rule between individuals so that this allocation can depend on relative wages and other variables in a way that reflects the bargaining position of individuals within the family rather than reflecting the symmetry assumption underlying the joint optimizing framework of the traditional approach. Individuals within the family can be altruistic and allocations Pareto efficient, but still the allocation rule can deviate from the optimal rule in the traditional model.

The most lucid statement of this argument can be found in the papers on household labour supply by Chiappori (1988; 1992). He states the family labour supply problem as one of

$$\max \theta U^h + (1-\theta) U^w s.t.\ c_{it} + w_{it}^h l_{it}^h + w_{it}^w l_{it}^w (= m_{it}) = (w_{it}^h + w_{it}^w) T + y$$

with some non-negative function $\theta = f(w_{it}^h, w_{it}^w, x_{it}, m_{it})$ representing the weight given to utility U^h. What Chiappori shows is that this is equivalent to a sharing rule solution in which U^h gets income $\phi(w_{it}^h, w_{it}^w, x_{it}, m_{it})$ out of y, and then allocates according to the rule:

$$\max U^h s.t. \quad c_{it}^h + w_{it}^h l_{it}^h = w_{it}^h T + \phi(w_{it}^w, w_{it}^w, x_{it}, m_{it})$$

where x_{it} may be a distribution factor.

Conditions for the identification of preferences and the sharing rule (up to a linear translation) simply require an observable private good – here assumed to be the individual's leisure. The intuition behind identification is simple: under the exclusive good assumption the spouse's wage can only have an effect through the sharing rule. Variation of income and wage will then provide an estimate of the marginal rate of substitution in the sharing rule. The same can be done for both spouses, and since the sharing rule must sum to 1, the partial derivatives of the sharing rule can be recovered.

The empirical implementation of the collective model has been slow but is growing in recent years; see Donni (2003) and Fortin and Lacroix (1997), for example. Generalizing the collective model to allow for non-participation and corner solutions requires additional care (see Blundell et al., 2006). The generalization to an intertemporal framework is still in its infancy.

The collective approach is not the only way to conceive of bargaining in family labour supply; see Kooreman and Kapteyn (1990), Lundberg (1988) and McElroy (1981) for important alternatives.

5 Labour supply with taxation and welfare participation

The tax and welfare system leads to well documented nonlinearities and non-convexities in the budget constraint facing any individual. This considerably complicates the labour supply problem and, even in the static setting, discrete choice programming methods are required. The basic nonlinear budget constraint problem has been described in detail in Hausman (1985a), Moffitt (1986), MaCurdy, Green and Paarsch (1990) among others.

To further address the issues encountered with nonlinear budget sets, there has been a steady expansion in the use of sophisticated statistical models characterizing distributions of discrete-continuous variables that jointly describe both interior choices and corner solutions in demand systems. These models offer a natural framework for capturing irregularities in budget constraints, including those induced by the institutional features of tax and welfare programmes. Typically the overall stochastic specification is represented by a mixed-multinomial specification across discrete choices over ranges of hours, for example in the work of Hoynes (1996) and Keane and Moffitt (1998). In this research, individuals are assumed to maximize their (stochastic) utility subject to a budget constraint, determined by a fixed hourly wage and the tax and benefit system. The utility function (8) is often approximated with a second-degree polynomial in hours of work and net income. A common feature of these models is the introduction of unobserved preference heterogeneity in the marginal rate of substitution between work and consumption. Further unobserved heterogeneity in the ‘costs’ of programme participation and in fixed costs of work is also now commonplace; see Blundell and MaCurdy (1999).

5.1 Discrete hours choices

In view of the large number of non-convexities, it is common to discretize hours into hours bands, and consider the choice across these intervals. For example, in Keane and Moffitt (1998) the utility function is modelled as

$$U^*_{H^j} = U(Y_{H^j}, T - H^j; x) + \varepsilon_{H^j} \tag{54}$$

where ε_{H^j} represents an unobserved preference component relating to the particular hours choice $h \equiv H^j$, assumed to be distributed as an extreme value random variable. Household disposable income, when supplying H^j hours, is defined by

$$Y_{H^j} = wH^j + b - R(H^j, w, g; x) \tag{55}$$

where w is the pre-tax hourly wage rate, g is other income (not including benefits and transfers) and $R(H^j, w, g; x)$ is the tax payable (positive or negative) when working H^j hours and having demographic composition x. Thus R will reflect both tax payments and credits or welfare payments received. This expression reflects the fact that the tax and benefit system may be nonlinear and may give rise to non-convexities; in these cases it is no longer possible to express the impact of the tax system simply by a marginal tax rate.

5.2 Fixed costs of work

Fixed costs are the costs that an individual has to pay to get to work; see Cogan (1980; 1981) and Hausman (1980). For parents, they are made up in part by childcare costs. In particular, childcare induces both fixed and variable costs that effectively act as a marginal tax rate. However, there are additional costs, for example, transport, which will vary by household type and by region. These are typically modelled as a once-off weekly cost and are subtracted directly from net income for any choices that involve work. They enter the utility comparisons in each individual's work–non-work choice.

5.3 Missing wages

For non-workers gross wages are not observed. As in the discussion of corner solutions and non-

participation in Section 1, for each individual we could write the logarithm of hourly wages as

$$\ln w = z'\gamma + \omega \tag{56}$$

where ω has density $g(\omega)$ and where z will include education, cohort and time dummies and their interactions. In principle the wage equation and the labour supply model can be estimated jointly. However, for computational reasons it is common to pre-estimate the marginal density of wages and then treat it as known at the estimation stage. This method can account for the endogeneity of gross wages and also allows for the complex relationship between gross wages and marginal wages in the tax and benefit system.

5.4 Programme participation, stigma and benefit take-up

Since the important work of Moffitt (1983) and Ashenfelter (1983), the formal analysis of welfare stigma and programme participation has been a key component of the labour supply impacts of tax and welfare programmes. Suppose P=1 indicates that an eligible individual participates in a welfare programme. Eligibility at any hours point H^j will typically depend on earnings, other income sources, family characteristics, and the rules of the tax and benefit system. Suppose that the hassle cost and stigma is given by η, an unobservable random variable. Then we may express utility for combination $\{H^j, P\}$ as

$$U^* \equiv U^*(Y_{H^j,P} - F, T - H^j, |x) - \eta P \tag{57}$$

where F is fixed costs of work. The stigma cost variable η may be modelled as a single unknown parameter representing a common cost across all individuals. More usefully it can be modelled as a random process with unknown mean μ_η and distribution $f_\eta(\eta)$. The parameters of its distribution are then recovered during estimation. Notice that net income $Y_{H^j,F}$ also depends directly on P through the working of the benefit and credit system. For any distribution of stigma costs an increase in the generosity of the benefit will increase the probability of take-up. Consequently, other things equal, take-up will be higher among those eligible for a larger benefit.

As documented in Blundell and MaCurdy (1999), for each hours H^j where the family is eligible to participate in the programme, utility function (57) defines a reservation stigma cost $\eta^*_{H^j}$ above which the

family would prefer not to participate at that hours level (note that the same family may choose to participate for some other hours level where it is also eligible for the programme). Given the family characteristics and the tax/benefit rules, the eligibility of each family at each level of hours can be determined, and the likelihood used in estimating the unknown parameters of labour supply, wages, fixed costs and programme participation can be fully specified.

5.5 Family labour supply and taxation

The modelling structure for couples requires but few modifications provided a ‘unitary’ model of family labour supply is adopted. The important difference in practice, as far as taxation and welfare is concerned, is that now we have to take into account the interaction of the welfare benefits that individuals may receive; see Hausman and Ruud (1984), Hoynes (1996) and van Soest (1995). Thus, the options facing each spouse are typically very different depending on whether the other family members work. Tax credit systems tend to lead to complex interactions between the effective tax rates for spouses (see Blundell et al., 2000; Eissa and Hoynes, 2004).

5.6 Optimal taxation and labour supply

One of the key developments in the use of labour supply elasticities has been in the design of ‘optimal’ tax and transfer systems following the innovative work of Saez (2001; 2002) and Laroque (2004). This has established a close link between the empirical analysis of labour supply responses and the early literature on optimal taxation (Mirrlees, 1971); see for example the implementation of these ideas in Immervol et al. (2007).

5.7 Randomized control trials and quasi-experimental approaches

Focusing purely on the reduced form impact of tax reform on labour supply, there have been several influential studies that have sidestepped the labour supply choice model and attempted to recover the impact of reforms on labour supply using randomized control experiments and quasi-experiments. The leading pure experiments are the Seattle–Denver Income Maintenance Experiment documented in Ashenfelter and Plant (1990) and the more recent Canadian Self Sufficiency Program for single mothers on welfare analysed in Card and Robins (1998). These provide a direct impact of a specific reform and also provide a useful basis from which to judge estimates from structural models.
Quasi-experimental methods, which compare an eligible and a comparison group before and after a reform, have also been influencial – for example the Eissa and Liebman (1996) study of the 1986 expansion of the Earned Income Tax Credit in the United States and the impact of tax rate changes on the taxable earnings of higher-income earners; see, in particular, the study by Feldstein (1995) and the further analysis by Gruber and Saez (2002). However, these quasi-experimental approaches require strong assumptions to be interpretable as measuring behavioural responses; see Blundell and MaCurdy (1999).

6 Conclusions: which labour supply elasticities for policy evaluation?

An argument has been made for an explicitly intertemporal framework, although, as we have seen, perfectly interpretable estimates of some important parameters of interest can be recovered from models that look essentially static. Much of the difference across empirical models reflects differences in data availability, and this provides another motivation for our approach. Precisely what form of income, hours or wage variables is available will vary widely across data sources, but this doesn't necessarily imply incomparable results. Some data provides longitudinal information on individual wages and hours; other data is repeated cross section but may have more detailed information on asset or consumption levels.

In whatever context the analysis of labour supply takes place, estimation will benefit from exogenous wage and income variation. One thing is clear: the type of trends that have occurred in many economies since the 1970s and the wide range of policy reforms designed to change labour supply incentives do strengthen the case for exploiting time-series information and avoiding complete reliance on purely cross-section data.

Four basic elasticities have been described which cover the main wage elasticities estimated in empirical labour supply analysis. Two are within-period elasticities: the first relating to the purely static formulation and the second relating to the two-stage budgeting specification. Two are life-cycle elasticities: the first being the intertemporal elasticity of substitution relating to the Frisch specification and measuring responses to evolutionary movements along the life-cycle wage profile, and the second relating to a full life-cycle specification and measuring responses to parametric shifts in the life-cycle profile itself. As most tax and benefit reforms are probably best described as once-and-for-all unanticipated shifts in net-of-tax real wages today and in the future, the most appropriate elasticity for describing responses to this kind of shift is the last of these. For the standard business cycle model it is the anticipated change that is of importance. As we have noted, these two elasticities can be substantially different due to income and wealth effects.

If a researcher regresses log hours of work on age; all age-invariant characteristics determining lifetime wages, preferences, and initial permanent income; and log wage, then the coefficient on the current wage rate is the Frisch elasticity. Intuitively, this approach controls for differences in the initial value of the marginal utility of wealth across consumers and leaves higher-order age variables as instruments to identify wage variation. Hence, only evolutionary wage variation along the age–wage path is included.

If, alternatively, a researcher regresses log hours worked on property income, age, age squared, and log wage, the coefficient on wage is the response of labour supply to a parametric wage shift – including both the intertemporal substitution effect and the reallocation of wealth across periods captured by a change in the marginal utility. Intuitively, this approach controls for age effects and leaves individual characteristics as instruments for wage. Changes in these characteristics capture full profile shifts rather than movements along the age–wage path.

The standard static labour supply representations fit neither of these patterns, as they include property income together with personal characteristics rather than age and age squared. Hence, given the existence of life-cycle effects they confuse the effect of movements along the wage profile with shifts in the profile and, thus, yield parameters without an economic interpretation.

See Also

- collective models of the household
- elasticity of intertemporal substitution
- hours worked (long run trends)
- indirect utility function
- retirement
- substitutes and complements
- taxation of income
- taxation of the family

Bibliography

Abbott, M. and Ashenfelter, O. 1976. Labor supply, commodity demand and the allocation of time. *Review of Economic Studies* 43, 389–411.

Altonji, J.G. 1982. The intertemporal substitution model of labour market fluctuations: an empirical analysis. *Review of Economic Studies* 49, 783–824.

Altonji, J.G. 1986. Intertemporal substitution in labor supply: evidence from micro data. *Journal of Political Economy* 94, S176–S215.

Altug, S. and Miller, R. 1990. Household choices in equilibrium. *Econometrica* 58, 543–70.

Altug, S. and Miller, R. 1998. The effect of work experience on female wages and labour supply. *Review of Economic Studies* 65, 45–85.

Apps, P.F. and Rees, R. 1997. Collective labor supply and household production. *Journal of Political Economy* 105, 178–90.

Arellano, M. and Meghir, C. 1992. Female labour supply and on-the-job search: an empirical model estimated using complementary data sets. *Review of Economic Studies* 59, 537–59.

Arrufat, J.L. and Zabalza, A. 1986. Female labour supply with taxation, random preferences, and optimization errors. *Econometrica* 54, 47–63.

Ashenfelter, O. 1983. Determining participation in income-tested social programs. *Journal of the American Statistical Society* 78, 517–25.

Ashenfelter, O. and Ham, J. 1979. Education, unemployment and earnings. *Journal of Political Economy* 87, S99–S166.

Ashenfelter, O. and Heckman, J.J. 1974. The estimation of income and substitution effects in a model of family labor supply. *Econometrica* 42, 73–85.

Ashenfelter, O. and Plant, M.W. 1990. Nonparametric estimates of the labor-supply effects of negative income tax programs. *Journal of Labor Economics* 8, S396–S415.

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–517.

Ben-Porath, Y. 1973. Economic analysis of fertility in Israel: point and counterpoint. *Journal of Political Economy* 81, S202–S233.

Bingley, P. and Walker, I. 1997. The labour supply, unemployment and participation of lone mothers in in-work transfer programs. *Economic Journal* 107, 1375–90.

Blau, D.M. 1994. Labor force dynamics of older men. *Econometrica* 62, 117–56.

Blau, D. and Robins, P. 1988. Child-care costs and family labor supply. *Review of Economics and Statistics* 70, 374–81.

Blomquist, N.S. 1983. The effect of income taxation on the labour supply of married men in Sweden. *Journal of Public Economics* 22, 169–97.

Blomquist, S. and Newey, W. 2002. Nonparametric estimation with nonlinear budget sets. *Econometrica* 70, 2455–80.

Blundell, R. 1998. Consumer Demand and Intertemporal Allocations: Engel, Slutsky and Frisch, In Stienar Strom (ed), *The Raynar Frisch Centennial Symposium*, Econometric Society Monographs 31, Cambridge University Press, Cambridge.

Blundell, R. 2006. Earned income tax credit policies: impact and optimality: the Adam Smith lecture. *Labour Economics* 13(4), 423–43.

Blundell, R., Chiappori, P.-A. and Meghir, C. 2005. Collective labour supply with children. *Journal of Political Economy* 113, 1277–306.

Blundell, R., Chiappori, P.A., Magnac, T. and Meghir, C. 2006. Collective labor supply: heterogeneity and non-participation. *Review of Economic Studies* 74, 417–45.

Blundell, R.W., Duncan, A. and Meghir, C. 1992. Taxation in empirical labour supply models: lone mothers in the UK. *Economic Journal* 102, 265–78.

Blundell, R., Duncan, A. and Meghir, C. 1998. Estimating labor supply responses using tax reforms. *Econometrica* 66, 827–61.

Blundell, R., Duncan, A., McCrae, J. and Meghir, C. 2000. The labour market impact of the working families' tax credit. *Fiscal Studies* 21, 75–104.

Blundell, R.W., Ham, J. and Meghir, C. 1987. Unemployment and female labour supply. *Economic Journal* 97, 44–64.

Blundell, R. and Hoynes, H. 2004. Has in-work benefit reform helped the labour market? In *Seeking a Premier League Economy*, ed. R. Blundell, D. Card and R.B. Freeman. Chicago: University of Chicago Press.

Blundell, R.W. and MaCurdy, T. 1999. Labor supply: a review of alternative approaches. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Blundell, R.W., MaCurdy, T. and Meghir, C. 2007. Estimation and specification in labor supply models. In *Handbook of Econometrics*, vol. 7, ed. J.J. Heckman and E. Leamer. Amsterdam: North-Holland (forthcoming).

Blundell, R., Meghir, C. and Neves, P. 1993. Labour supply and intertemporal substitution. *Journal of Econometrics* 59, 137–60.

Blundell, R.W. and Walker, I. 1982. Modelling the joint determination of household labour supplies and commodity demands. *Economic Journal* 92, 58–74.

Blundell, R.W. and Walker, I. 1986. A life-cycle consistent empirical model of family labour supply using cross-section data. *Review of Economic Studies* 53, 539–58.

Blundell, R.W., Meghir, C., Symons, E. and Walker, I. 1988. Labour supply specification and the evaluation of tax reforms. *Journal of Public Economics* 36, 23–52.

Boskin, M. 1973. Economics of the labor supply. In *Labor Supply and Income Maintenance*, ed. G. Cain and H. Watts. Chicago: Rand McNally.

Brewer, M., Duncan, A., Shephard, A. and Suárez, M.J. 2006. Did working families' tax credit work? The impact of in-work support on labour supply in Great Britain. *Labour Economics* 13, 699–720.

Browning, M., Deaton, A. and Irish, M. 1985. A profitable approach to labor supply and commodity demands over the life-cycle. *Econometrica* 53, 503–44.

Burtless, G. and Hausman, J.A. 1978. The effect of taxation on labor supply: evaluating the Gary negative income tax experiment. *Journal of Political Economy* 86, 1103–30.

Cain, G. 1966. *Married Women in the Labor Force*. Chicago: University of Chicago Press.

Card, D. 1994. Intertemporal labor supply: an assessment. In *Advances in Econometrics, Sixth World Congress*, ed. C. Sims. New York: Cambridge University Press.

Card, D. and Robins, P.K. 1998. Do financial incentives encourage welfare recipients to work? Evidence from a randomized evaluation of the self-sufficiency project. In *Research in Labor Economics*, vol. 17, ed. S. Polachek. Greenwich, CT: JAI Press.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56, 63–90.

Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100, 437–67.

Chiappori, P.-A. 1997. Introducing household production in collective models of labor supply. *Journal of Political Economy* 105, 191–209.

Chone, P.-A. and Laroque, G. 2005. Optimal incentives for labor force participation. *Journal of Public Economics* 89, 395–425.

Cogan, J.F. 1980. Labor supply with costs of labor market entry. In *Female Labor Supply: Theory and Estimation*, ed. J. Smith. Princeton, NJ: Princeton University Press.

Cogan, J.F. 1981. Fixed costs and labor supply. *Econometrica* 49, 945–64.

Donni, O. 2003. Collective household labor supply: nonparticipation and income taxation. *Journal of Public Economics* 87, 1179–98.

Eckstein, Z. and Wolpin, K. 1989. Dynamic labour force participation of married women and endogenous work experience. *Review of Economic Studies* 56, 375–90.

Eissa, N. and Liebman, J. 1996. Labor supply response to the earned income tax credit. *Quarterly Journal of Economics* 111, 605–37.

Eissa, N. and Hoynes, H. 2004. Taxes and the labor market participation of married couples: the earned income tax credit. *Journal of Public Economics* 88, 1931–58.

Feldstein, M. 1995. The effect of marginal tax rates on taxable income: a panel study of the 1986 tax reform act. *Journal of Political Economy* 103, 551–72.

Fortin, B. and Lacroix, G. 1997. A test of the unitary and collective models of household labour supply. *Economic Journal* 107, 933–55.

Fraker, T. and Moffitt, R. 1988. The effect of food stamps on labor supply: a bivariate selection model. *Journal of Public Economics* 35, 25–56.

Gorman, W.M. 1959. Separable utility and aggregation. *Econometrica* 27, 469–81.

Gorman, W.M. 1961. On a class of preference fields. *Metroeconomica* 13, 53–56. Repr. in *Separability and Aggregation: Collected Works of M. Gorman*, vol. 1, ed. C. Blackorby and A.F. Shorrocks. Oxford: Clarendon Press, 1995.

Gorman, W.M. 1968. The structure of utility functions. *Review of Economic Studies* 35, 367–90.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in Economic Analysis*, ed. M. Artis and R. Nobay. Cambridge: Cambridge University Press.

Gronau, R. 1974. Wage comparisons: a selectivity bias. *Journal of Political Economy* 82, 1119–44.

Gruber, J. and Saez, E. 2002. The elasticity of taxable income: evidence and implications. *Journal of Public Economics* 84, 1–32.

Hall, R. 1973. Wages, income and hours of work in the US labor force. In *Income Maintenance and labor supply*, ed. G. Cain and H. Watts. Chicago: Chicago University Press.

Ham, J. 1986a. Testing whether unemployment represents intertemporal labour supply behaviour. *Review of Economic Studies* 53, 559–78.

Hanoch, G. 1965. The ‘backward-bending’ supply of labor. *Journal of Political Economy* 73, 636–42.

Hanoch, G. 1980. A multivariate model of labor supply. In *Female Labor Supply: Theory and Estimation*, ed. J. Smith. Princeton, NJ: Princeton University Press.

Hanoch, G. and Honig, M. 1983. Retirement, wages, and labor supply of the elderly. *Journal of Labor Economics* 1, 131–51.

Hansen, L.P. and Singleton, K. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249–65.

Hausman, J. 1980. The effect of wages, taxes and fixed costs on women's labor force participation. *Journal of Pubic Economics* 14, 161–94.

Hausman, J. 1981. Labor supply. In *How Taxes Affect Economic Behavior?* ed. H. Aaron and J. Pechman. Washington, DC: Brookings Institution.

Hausman, J. 1985a. The econometrics of nonlinear budget sets. *Econometrica* 53, 1255–82.

Hausman, J. 1985b. Taxes and labor supply. In *Handbook of Public Economics*, vol. 1A, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Hausman, J. and Ruud, P. 1984. Family labor supply with taxes. *American Economic Review* 74, 242–8.

Heckman, J.J. 1974a. Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–94.

Heckman, J.J. 1974b. Life cycle consumption and labor supply: an explanation of the relationship between income and consumption over the life cycle. *American Economic Review* 64, 188–94.

Heckman, J.J. 1974c. Effects of child-care programs on women's work effort. *Journal of Political Economy* 82, S136–S163.

Heckman, J.J. 1976. Life-cycle model of earnings, learning and consumption. *Journal of Political Economy* 84, S11–S44.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–61.

Heckman, J.J. 1993. What has been learned about labor supply in the past twenty years. *American Economic Review* 83, 116–21.

Heckman, J.J. and MaCurdy, T.E. 1980. A life-cycle model of female labour supply. *Review of Economic Studies* 47, 47–74.

Hotz, V.J., Kydland, F.E. and Sedlacek, G.L. 1988. Intertemporal preferences and labor supply. *Econometrica* 56, 335–60.

Hoynes, H. 1996. Welfare transfers in two-parent families: labor supply and welfare participation under the AFDC-UP program. *Econometrica* 64, 295–332.

Hoynes, H. 2000. The employment, earnings and income of less skilled workers over the business cycle. In *Finding Jobs: Work and Welfare Reform*, ed. D.E. Card and R. Blank. New York: Russell Sage Foundation.

Hurd, M.D. and Boskin, M.J. 1984. The effect of social security on retirement in the early 1970s.

Quarterly Journal of Economics 99, 767–90.

Immervol, H., Kleven, H.J., Kreiner, C.T. and Saez, E. 2007. Welfare reform in European countries: a microsimulation analysis. *Economic Journal* 117, 1–44.

Johnson, T.R. and Pencavel, J.H. 1984. Dynamic hours of work functions for husbands, wives and single females. *Econometrica* 52, 363–89.

Keane, M.P. and Moffitt, R. 1998. A structural model of multiple welfare program participation and labor supply. *International Economic Review* 39, 553–89.

Killingsworth, M.R. 1983. *Labor Supply*. Cambridge: Cambridge University Press.

Killingsworth, M.R. and Heckman, J.J. 1986. Female labor supply: a survey. In *Handbook of Labor Economics*, vol. 1, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

Kooreman, P. and Kapteyn, A. 1990. On the empirical implementation of some game theoretic models of household labor supply. *Journal of Human Resources* 25, 584–98.

Krueger, A.B. and Pischke, J.-S. 1992. The effect of social security on labor supply: a cohort analysis of the notch generation. *Journal of Labor Economics* 10, 412–37.

Laroque, G. 2004. Income maintenance and labour force participation. *Econometrica* 73, 341–76.

Low, H. 2005. Self-insurance and unemployment benefit in a life-cycle model of labour supply and savings. *Review of Economic Dynamics* 8, 945–75.

Lewis, H.G. 1967. On income and substitution effects in labor force participation. Unpublished manuscript, University of Chicago.

Lundberg, S. 1988. Labor supply of husbands and wives: a simultaneous equations approach. *Review of Economics and Statistics* 70, 224–35.

MaCurdy, T.E. 1981. An empirical model of labour supply in a life-cycle setting. *Journal of Political Economy* 89, 1059–85.

MaCurdy, T.E. 1983. A simple scheme for estimating an intertemporal model of labor supply and consumption in the presence of taxes and uncertainty. *International Economic Review* 24, 265–89.

MaCurdy, T.E. 1985. Interpreting empirical models of labour supply in an intertemporal framework with uncertainty. In *Longitudinal Analysis of Labor Market Data*, ed. J.J. Heckman and B. Singer.

Cambridge: Cambridge University Press.

MaCurdy, T.E. 1992. Work disincentive effects of taxes: a reexamination of some evidence. *American Economic Review* 82, 243–9.

MaCurdy, T.E., Green, D. and Paarsch, H. 1990. Assessing empirical approaches for analyzing taxes and labour supply. *Journal of Human Resources* 25, 415–90.

McElroy, M.B. 1981. Empirical results from estimates of joint labor supply functions of husbands and wives. In *Research in Labor Economics*, vol. 4, ed. R.G. Ehrenberg. Greenwich, CT: JAI Press.

Mincer, J. 1960. Labor supply, family income, and consumption. *American Economic Review* 50, 574–83.

Mincer, J. 1962. Labor force participation of married women: a study of labor supply. In *Aspects in Labor Economics*, ed. H.G. Lewis. Princeton, NJ: Princeton University Press.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Moffitt, R. 1983. An economic model of welfare stigma. *American Economic Review* 73, 1023–35.

Moffitt, R. 1986. The econometrics of piecewise-linear budget constraints: a survey and exposition of the maximum likelihood method. *Journal of Business and Economic Statistics* 4, 317–27.

Moffitt, R. 2002. Economic effects of means-tested transfers in the US. In *Tax Policy and the Economy*, vol. 16, ed. J.M. Poterba. Cambridge, MA: MIT Press.

Moffitt, R.A. 2002. Welfare programs and labor supply. In *Handbook of Public Economics*, vol. 4, ed. A. J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Moffitt, R. 2006. Welfare work requirements with paternalistic government preferences. *Economic Journal* 116, F441–458.

Mroz, T.A. 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55, 765–99.

Muellbauer, J. 1981. Linear aggregation in neoclassical labour supply. *Review of Economic Studies* 48, 21–36.

Pencavel, J. 1986. Labor supply of men: a survey. In *Handbook of Labor Economics*, vol. 1, ed. O.

Ashenfelter and R. Layard. Amsterdam: North-Holland.

Rust, J. and Phelan, C. 1997. How Social Security and Medicare affect retirement behavior in a world of incomplete markets. *Econometrica* 65, 781–831.

Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68, 205–39.

Saez, E. 2002. Optimal income transfer programs: intensive versus extensive labor supply responses. *Quarterly Journal of Economics* 117, 1039–73.

Scholz, J.K. 1996. In-work benefits in the United States: the Earned Income Tax Credit. *Economic Journal* 106, 156–69.

Shaw, K. 1989. Life-cycle Labour Supply with Human Capital Accumulation. *International Economic Review* 30, 431–57.

Stern, N. 1986. On the specification of labor supply functions. In *Unemployment, Search and Labour Supply*, ed. R.W. Blundell and I. Walker. Cambridge: Cambridge University Press.

van Soest, A. 1995. Structural models of family labor supply: a discrete choice approach. *Journal of Human Resources* 30, 63–88.

Wales, T.J. and Woodland, A. 1976. Estimation of household utility functions and labor supply response. *International Economic Review* 17, 397–410.

How to cite this article

Blundell, Richard and Thomas MaCurdy. "labour supply." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000219> doi:10.1057/9780230226203.0918

The New Palgrave Dictionary of Economics Online

labour surplus economies

Gustav Ranis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In some sectors with a large endowment of unskilled labour and without sufficient cooperating land or capital, given technology and a wage level bounded from below, labour markets cannot clear. A full employment solution would drive remuneration below socially acceptable, possibly subsistence, levels of consumption. Consequently, a labour surplus exists in that much of the labour force contributes less to output than it requires: its marginal product falls below its remuneration, set by bargaining. A reallocation of such workers to other, competitive, sectors would eliminate the inefficiency and enhance total output. Open economy dimensions, extensions and critiques are dealt with.

Keywords

agriculture and economic development; Arrow, K.; balanced growth; Becker, G.; behavioural economics; dual economies; Engel's Law; family networks; informal sector; Kuznets, S.; labour surplus; marginal productivity; minimum subsistence level of existence (MSL); neoclassical economics; population growth; agricultural vs non-agricultural productivity; Sen, A.; solidarity networks

Article

Labour surplus economies are closely associated with the concept of economic dualism, that is, the existence of organizational heterogeneity as between major sectors of an economy. The basic premise is that there exist some sectors or sub-sectors in which, in the presence of a large endowment of unskilled labour and the absence of sufficient cooperating land or capital, and with a given technology and a wage level bounded from below, labour markets cannot clear. A full employment, neoclassical ‘wage equals marginal product’ solution would drive remuneration below socially acceptable, possibly subsistence, levels of consumption. Consequently, a labour surplus exists in the sense that a substantial portion of the labour force contributes less to output than it requires, that is, its marginal product falls below its remuneration, set by bargaining. The ‘labour surplus’ designation then arises from the fact that a reallocation of such workers to other, competitive, or neoclassically functioning sectors would eliminate the aforementioned inefficiency and thus materially enhance the total output of the system.

The prime location for such surplus labour has traditionally been developing countries’ agricultural sectors, concentrated especially in subsistence agriculture, characterized by family farms, that is, excluding commercialized plantation agriculture which consists of profit maximizing entities able to hire and fire workers following well-known neoclassical principles. Surplus labour makes its appearance in the context of owner-operated extended family networks, communes, villages or similar tenurial arrangements, all configurations in which income or output shares are determined via bargaining in relation to (though not necessarily equal to) the average rather than the marginal product of labour. Wage determination is thus based on a sharing principle, a function of the fact that, when high man–land ratios are among the initial conditions, low marginal-productivity workers cannot be dismissed or otherwise eliminated.

Here, we first present the static version of the labour surplus economy. Next we describe the conditions for balanced growth. Then open economy dimensions are introduced. Finally, some extensions are cited and rejoinders offered to some critiques.

The static labour surplus economy

Figure 1 illustrates the situation of relatively scarce land, intensively cultivated, yielding extremely low increments of output at the margin. Labour is measured on the horizontal and land on the vertical axis, with production contour lines indexed as M, M', and M'' in Figure 1. Given technology, fixed land at ON, and labour endowment at $OS = OS' = OS''$ in Figures 1, 2 and 3, the total product curve is ODQ_A in Figure 2 and the marginal product of labour, depicted by curve ABC in Figure 3, approaches very low levels, substantially below the bargaining or institutional wage or income share OW_a which is related to (again, not necessarily equal to) the average product (slope of OQ_A in Figure 2). Under these conditions, we can locate the proportion of the total agricultural labour force which is 'in surplus' in the sense that it is 'disguisedly unemployed' or 'underemployed' as $S''T$ in Figure 3. This includes all those whose marginal product lies below their consumption or income share. They represent the 'labour surplus' phenomenon or what Rosenstein-Rodan (1943) and Nurkse (1953) long ago designated as 'hidden rural savings' which could be mobilized via reallocation to higher-productivity activities elsewhere in the economy.

Figure 1

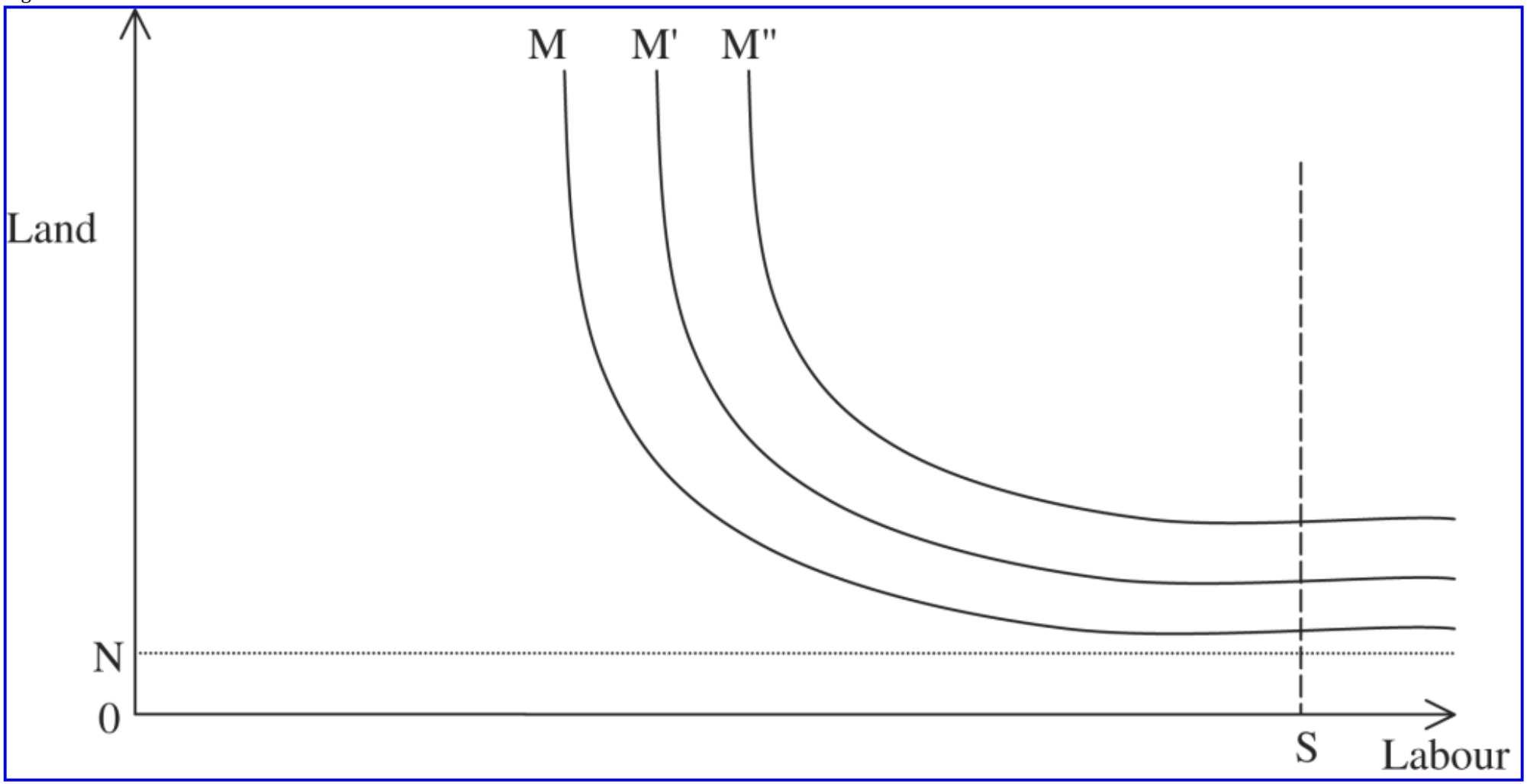

Figure 2

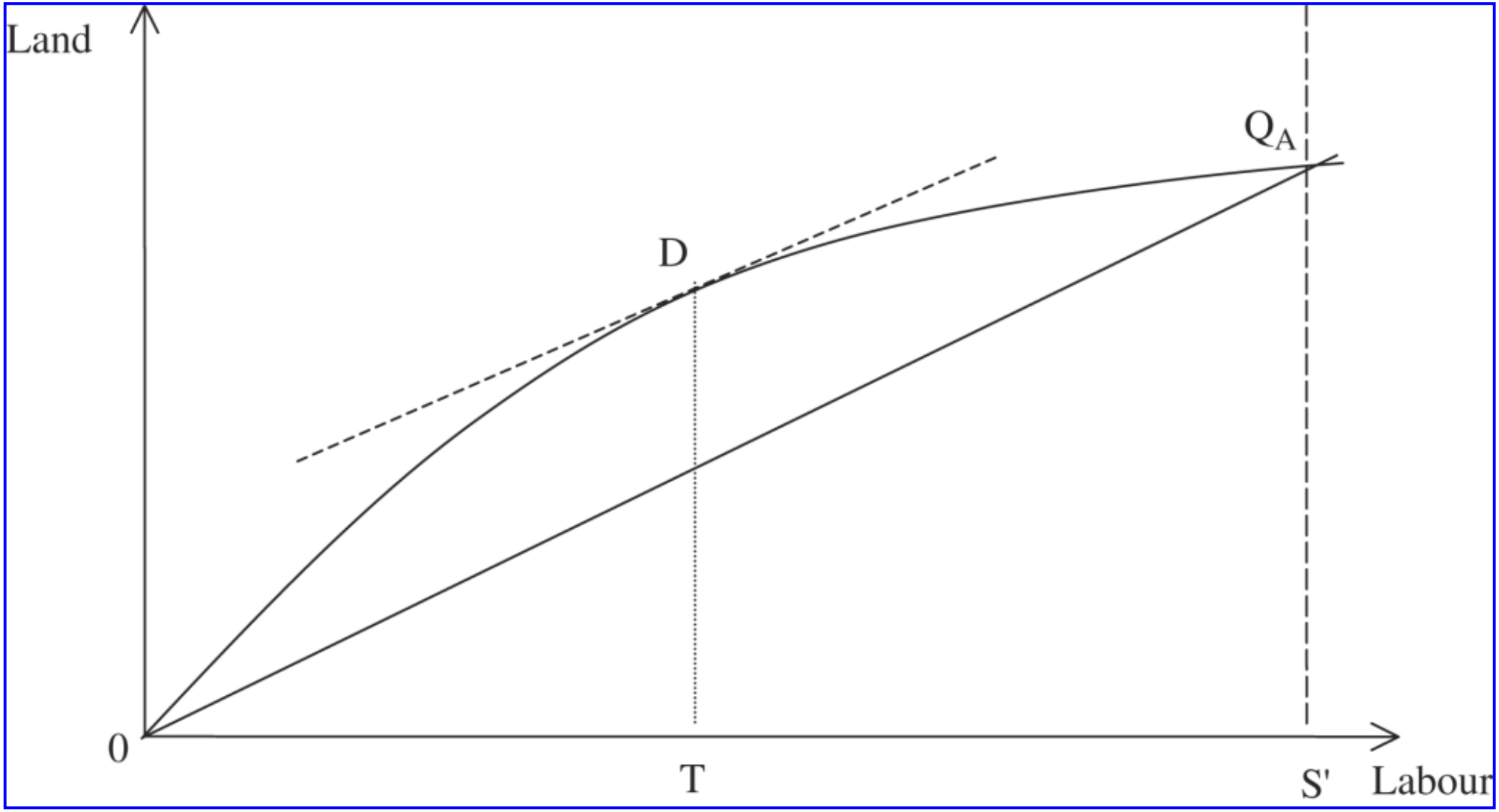

Figure 3

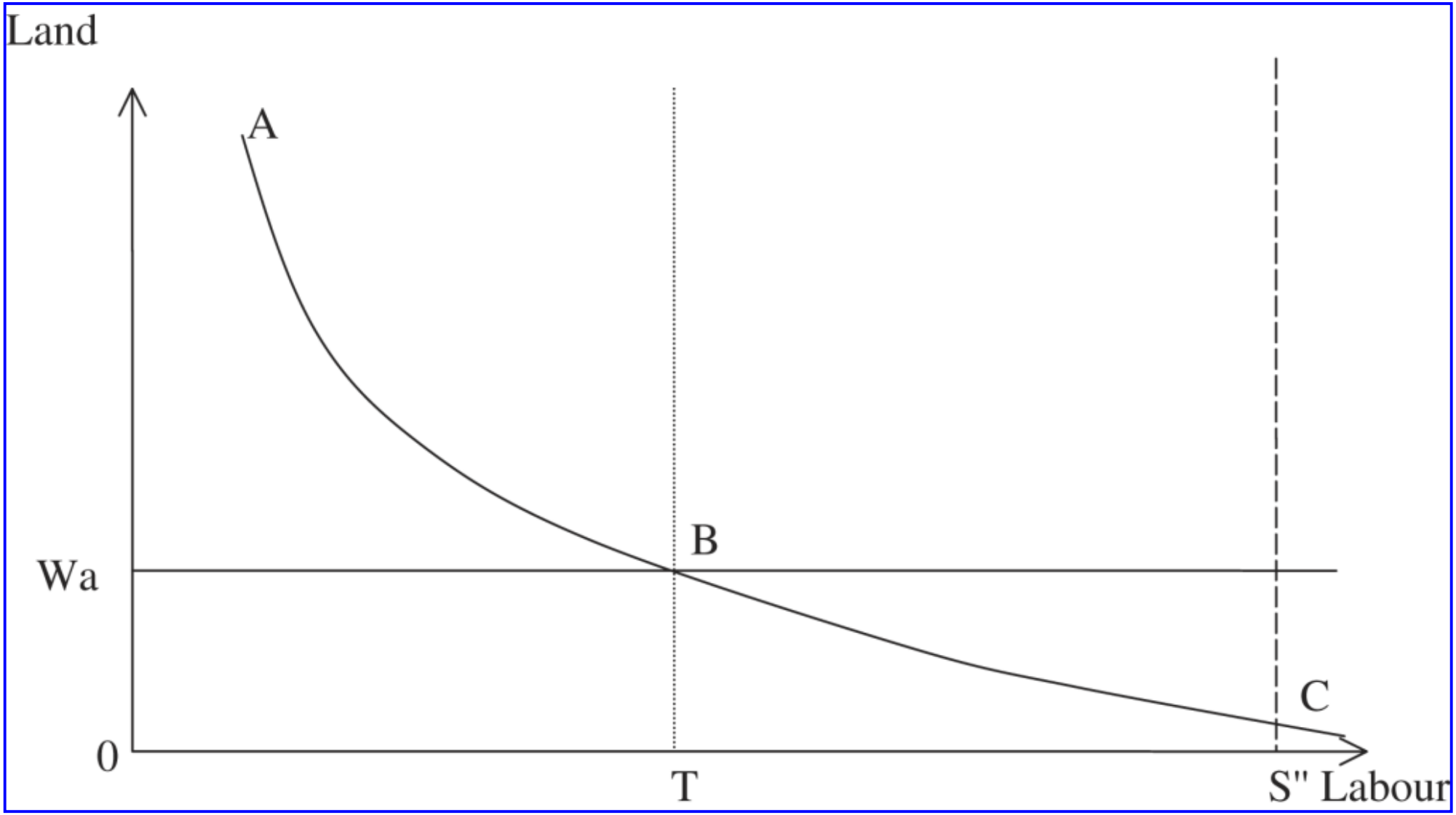

It should be emphasized that ‘labour surplus’ therefore does not mean, as has often been asserted, that a substantial portion of the agricultural labour force can be withdrawn without loss of output. Such a zero marginal product condition constitutes a statistically highly unlikely razor's edge event but, partly because it has been assumed for purely diagrammatic and/or mathematical convenience by Lewis (1972), by Fei and Ranis (1964), and by others, it has drawn extensive and often intemperate critical comment in the literature. Schultz (1964, p. 70), for example, cited the fact that output in India fell with a decline in the agricultural working population due to an influenza epidemic as proof that surplus labour was a ‘false doctrine’. As Sen (1967) pointed out in rebutting Schultz on this point, when some workers with low (or even zero) marginal productivity are withdrawn, some of those left behind are likely to adjust by working harder. Or, put more broadly, any withdrawal of labour from agriculture is very likely to be accompanied by a reorganization of production arrangements on the part of those left behind, that is, by technology change. This would be equivalent to an upward shift of the ODQ_A curve in Figure 2 and of the *ABC* curve in Figure 3.

Balanced growth in the labour surplus economy

Dynamically, the labour surplus condition can thus be seen as permitting an increasing number of agricultural workers and an increasing volume of agricultural surplus, defined as the

difference between total agricultural output and what is needed to satisfy the remaining agricultural population's consumption requirements, to move out and support the expansion of commercialized activities, industry and services, rural and urban. This labour surplus condition of the economy then ultimately comes to an end when increases in agricultural productivity, which free up workers and generate agricultural surpluses and, accompanied by increases in productivity in the expanding commercialized sector, enhance the demand for workers, have proceeded in a more or less ‘balanced’ fashion long enough, and at a rate exceeding population growth, to mop up the disguisedly unemployed, that is, all those whose marginal product lies below their wage or consumption standard.

This critical concept of the need for ‘balance’ between the non-commercialized and commercialized components of the labour surplus economy has really three ingredients. One, the most obvious, is that the release of labour from non-commercialized agriculture is roughly in balance with its absorption by commercialized non-agriculture. Another, focused on the product rather than the organizational dimension of dualism, suggests that relative advances in productivity in the two sectors proceed in such a fashion that the inter-sectoral terms of trade are not substantially affected, that is, that the system does not encounter food shortages or, less likely, food surpluses in the course of the development process. Third, the financial intermediation network, primitive at first, more sophisticated later, represents a crucial link as it must be capable of transforming non-commercialized sector surpluses, joined by commercialized sector profits, into efficient investment, mainly in the commercialized sector.

To turn first to more specifics on the inter-sectoral labour market, it should be noted that the unskilled real wage in the commercialized sector will tend to be tied to, though certainly not equal to, the non-commercialized agricultural real wage. A substantial unskilled labour wage gap is indeed likely to be required, partly to induce the typical agricultural worker to overcome her attachment to soil and family, partly to meet transport costs, and partly as a consequence of such institutional factors as commercialized sector minimum wage legislation, unionization, the public sector wage setting, and so forth, all of which usually do not extend into non-commercialized activities. Once these two wage levels are given within a general equilibrium context, the release of labour by the non-commercialized sector and its absorption by the commercialized sector represents an essential ingredient of balanced growth in the labour surplus economy.

It should also be noted that both wages may be expected to rise over time, in part because, as agricultural sector labour productivity increases, there is also likely to be some upward adjustment of the bargaining wage which is tied to the rising average product. Moreover, the inter-sectoral wage gap may rise as a consequence of a change in the extent of commercialized sector interventions via minimum wage increases, enhanced union bargaining power, and so on. The two unskilled real wage patterns over time may thus be conceived of as a step function, horizontal at any point in time, reflecting the labour surplus condition, but at a slightly higher level, again horizontal, in the next period. All this will, of course, yield a gently rising labour supply curve over time, giving way to a sharply rising pattern once the labour surplus has been exhausted and remuneration is determined neoclassically, that is to say, by the marginal product. Meanwhile, the existence of a relatively constant or gently upward-sloping real wage over time in both sectors, with a possibly growing gap between them, can be expected to induce labour-intensive technology choices and, more importantly, labour-using technological change in both the non-commercialized and commercialized sectors of the labour surplus economy.

Second, an understanding of the workings of the inter-sectoral commodity market is required for an assessment of the contribution of the non-commercialized sector to the rest of the economy. This can be seen in terms of the net real resources transferred, that is, the difference between the shipments of food and raw materials delivered to the commercialized sector and the shipments of goods and services sent in the opposite direction. The agricultural sector's export surplus may thus be viewed as the contribution of that sector to both the labour reallocation and overall growth process over time.

The main participants in the dualistic commodity market are thus, on the one hand, the owners of the agricultural surplus and, on the other, the newly allocated workers who may be thought of as receiving wage income in the form of non-agricultural goods and anxious to trade some of these for the food ‘left behind’. Once this transaction is completed, the reallocated worker finds herself in possession of the agricultural goods needed to at least maintain her consumption standard – most likely to increase it because of the aforementioned inter-sectoral wage gap. In this fashion the dualistic commodity market is indispensable for transforming the consumption bundle of the agricultural labour force into a wages fund for the newly allocated non-agricultural workers. At the same time the owners of the agricultural surplus, such as the landlords and/or the government via land taxes, obtain a claim against a portion of the newly formed non-agricultural capital stock; the other portion results from the reinvestment of profits by commercialized sector entrepreneurs. The above underlines the importance of the product, along with the organizational dimension of balanced growth in the closed labour surplus economy, rooted in the fact that food and non-agricultural products cannot readily be substituted for each other. Agriculture is thus a necessary condition for non-agriculture, while the converse does not strictly hold. In the open economy, food imports, of course, become possible, thus helping the system avoid premature food shortages, as illustrated by Japan's historical experience in the early decades of the 20th century (see Hayami and Ruttan, 1970).

Third, the financial counterpart of the real resources contribution of the non-commercialized to the commercialized sector over time is effected through the workings of the intersectoral financial market. As we have seen, the savings of the agricultural sector become a claim against non-agriculture, the magnitude of which is determined by the size of its export surplus. These savings must somehow be channeled into non-agricultural investment; that is, what is left of the agricultural surplus that is not siphoned off by consumption or intermediate input requirements must find its way into capital formation in the rest of the economy.

The dynamics reflecting all the main facets of such a balanced growth path can be illustrated by reference to Figure 4 within a simplified setting, that is, without intermediate input flows between the two sectors. Total population *L* is shown on the horizontal axis in quadrant II, moving from right to left, with agricultural output and the institutional consumption standard $c = w_a^0$, measured in terms of agricultural goods, on the vertical axis. The curve OQ_A^{*0} describes per capita food availability for the total population, or *Q/L*, at a given level

of technology, for various possible proportions, θ, of the total population already allocated to other activities, *B*, that is, $(\theta = B / L \geq 0)$. One equilibrium point along a balanced growth path may then be defined as follows: let initial consumption $c = w_a^0$, and the terms of trade between w_a^0, the 'wage in terms of agricultural goods' (Q_A), and w_{na}^0 the 'wage in terms of non-agricultural goods' (Q_{NA}) be given. For simplification only, we assume that there is no wage gap between unskilled agricultural and non-agricultural workers. The price–consumption curve (*PC*) in quadrant I of Figure 4 then indicates all possible points of tangency between changing terms of trade and a given typical worker's consumer preference between agricultural and non-agricultural goods. Point *e* is the consumption equilibrium point for the typical worker, given the terms of trade shown, regardless of whether she is engaged in agricultural or non-agricultural activities. *B* is the population outside of agriculture and the remaining agricultural population *V* $(L = B + V)$ produces enough food to meet everyone's consumption requirements at the institutional wage.

Figure 4

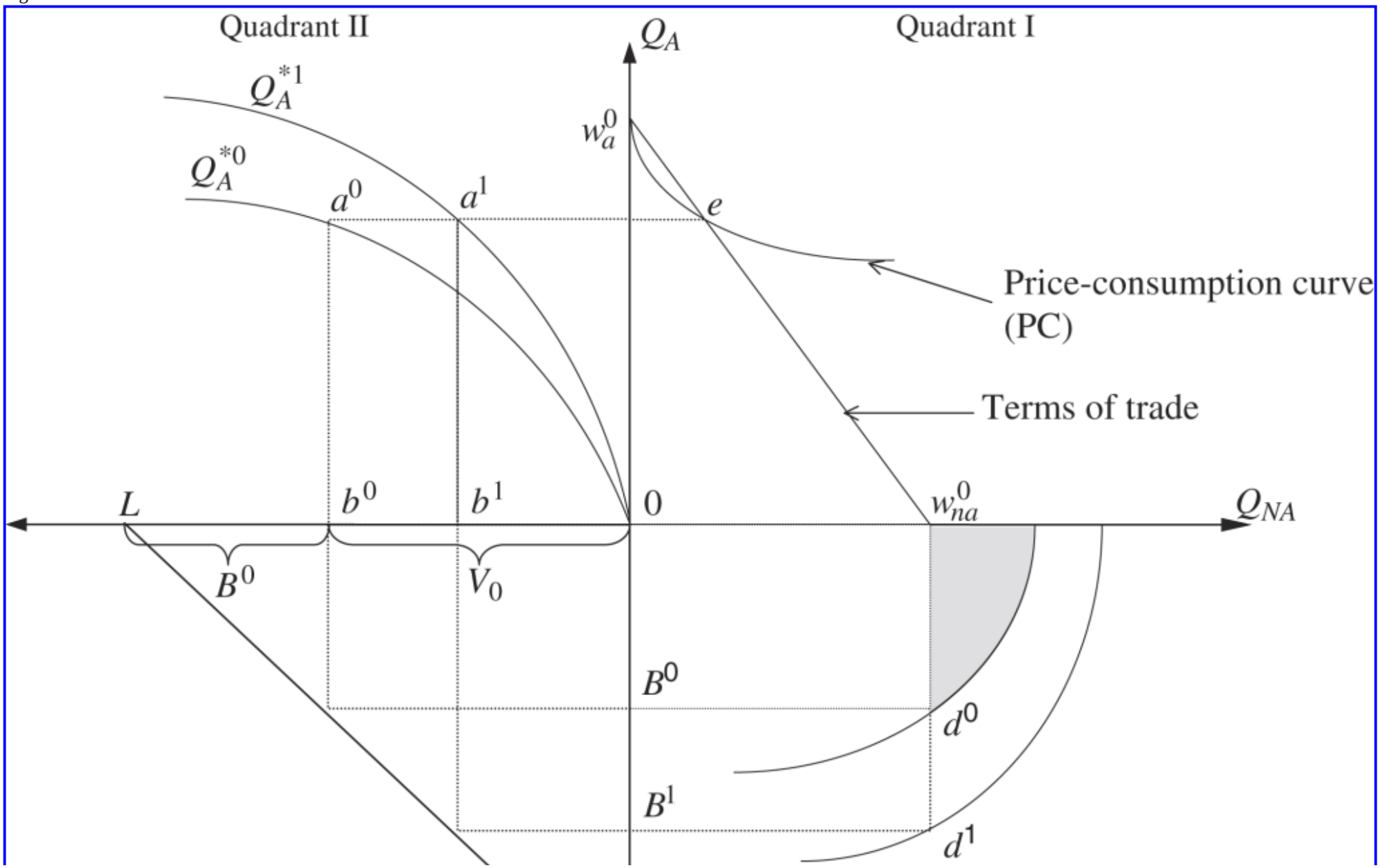

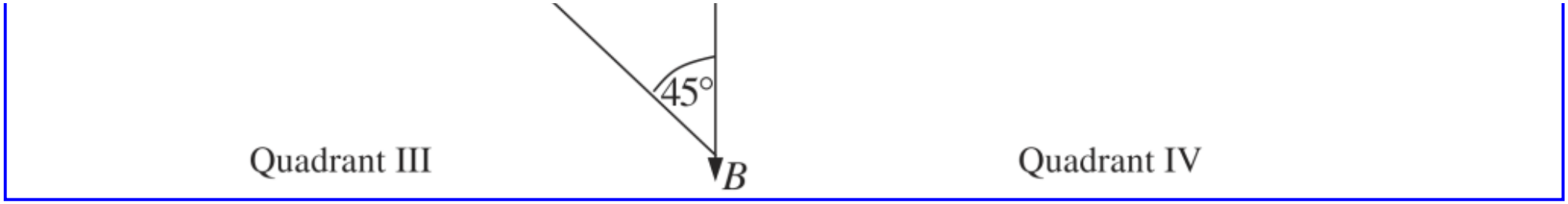

The auxiliary 45° line in quadrant III transposes workers $B°$, already allocated to non-agricultural work, onto the vertical axis, that is, $OB°$. The consistent equilibrium point for employment in the non-agricultural sector is then point $d°$, located at the intersection between the 'horizontal' supply curve of non-agricultural labour, at wage level w^0_{na}, and the demand curve for non-agricultural labour, or the marginal productivity curve corresponding to a particular level of the capital stock and technology in that sector. This describes an equilibrium position $a°b°d°$ in both the intersectoral labour and commodity markets.

To turn to the definition of balanced growth over time, and on the assumption of no upward adjustment of the agricultural real wage and, thus, of the non-agricultural real wage which is 'tied' to it, balanced increases in agricultural and non-agricultural productivity resulting from capital accumulation and technology change can be shown by a shift of the per capita food availability curve to OQ^{*1}_A in quadrant II, with Lb^1 or OB^1 workers now allocated, as well as of the marginal productivity of non-agricultural labour curve to d^1 in quadrant IV. This would result in a new equilibrium position $a^1b^1d^1$ where, once again, the two intersectoral markets clear. Such a growth path would clearly meet the labour market equilibrium condition, and a little more work would permit us to demonstrate that equilibrium in the commodity market sense, as previously defined, also continues to be achieved, permitting agricultural and non-agricultural workers to exchange some of the goods they produce for the goods they need, at the given terms of trade, enabling everyone to remain at the same equilibrium point e.

To turn to the inter-sectoral financial market, the landlords and/or the government, whoever owns the agricultural surplus, would end up with a claim against some part of the non-agricultural capital stock. This, plus the reinvested industrial profits represented by the shaded area in quadrant IV of Figure 4 would be invested in the non-agricultural sector, causing, along with technology change, the indicated shift of the marginal productivity curve. The investment fund for the next period is thus composed of this period's savings out of the agricultural surplus plus the savings out of non-agricultural profits. For the sake of convenience, we have made the assumption of no leakage into consumption by either landlords or capitalists. The allocation of the society's investment fund plus its innovative energies, as between the sectors, would then be guided by the relative shortages of agricultural and non-agricultural goods, as reflected, in the case of a market economy, by changes in the inter-sectoral terms of trade. In a non-market economy the role of changes in the terms of trade as a signalling device would be taken over by evidence of unplanned shortages or surpluses in the material balances sense. We have here again made a simplifying, but not critical, assumption that technology change is responsible for agricultural productivity change, while all the investment funds are allocated to non-agriculture.

As we have already noted, the entire transition process must not only be balanced but also proceed at a pace in excess of population growth if the initial reservoir of surplus labour is to ultimately be exhausted and neoclassical wage determination is to take over. Moreover, if balanced growth, as indexed by the rate of labour reallocation, only marginally exceeds the rate of population growth on average, the length of time it takes to arrive at the commercialization point, marking the end of labour surplus, must also be politically acceptable. The real world, of course, does not quite operate in such a smooth fashion. There are times when, under the impetus of an 'industry first' strategy, non-agricultural productivity increases for some time at a rate in excess of agricultural productivity growth, leading to food shortages, the shifting of the terms of trade in favour of agriculture, and an increase in the non-agricultural real wage. The reverse can also occur, although empirically there seems to be less danger of that. Most successful labour surplus societies (such as historical Japan and post-war South Korea, Taiwan and Thailand) have, in fact, experienced something approaching constancy in the terms of trade.

In any case, progress along a balanced growth path at a rate in excess of population growth – and sufficiently in excess to guarantee a politically acceptable time perspective – is essential to a society's successful transition into a modern growth regime. Success is defined as the end of labour surplus, that is, the end of organizational dualism in the labour market. Once balanced growth has proceeded long enough and fast enough labour surplus gives way to labour shortage in both sectors, which means that the marginal productivity calculus of wage determination takes over. At this point organizational dualism disappears; and, given considerable increases in per capita incomes and the workings of Engel's Law, product dualism also atrophies over time as agriculture gradually becomes an appendage to the economy, or just another symmetrical sector within the system's input–output matrix. Increasingly the economy is then ready to perform according to the rules of modern economic growth as described by Simon Kuznets (1966).

Open economy dimensions

Thus far we have discussed the development of the labour surplus economy mainly in a closed economy context. The open economy or trade-related dimensions of development in the labour surplus economy are, of course, important enough to warrant substantial amendment of the analysis presented here. During the early colonial, or open agrarian, phase of development, the economy may well be tied to foreign markets by virtue of some of the labour force being weaned away from food production and into land-based export-oriented

activity: for example, minerals and other primary products of interest to foreign investors. This typically leads to a triangular relationship among the cash-crop export sector, the foreign sector, and the food producing domestic agricultural sector. But once the economy moves out of its colonial or ‘overseas territory’ phase and into a national development-oriented effort, our analysis must be amended to take ‘openness’ into account.

To do so, we must, first, recognize that the export-oriented cash crop agricultural sub-sector continues to generate foreign exchange earnings but that these are now used, in addition to possible food imports, to assist in the construction of a new, domestically oriented, non-agricultural sector producing previously imported non-durable consumer goods, that is, to fuel so-called primary or ‘easy’ industrial import substitution. These raw material-intensive exports thus provide a second source of agricultural surplus which, converted into industrial capital goods imports, and possibly supplemented by the inflow of foreign savings, helps finance non-agricultural growth in the same balanced growth context. In this way a new triangular relationship between two kinds of commercialized activities, one agricultural and the other non-agricultural, plus the food producing non-commercialized agricultural hinterland, replaces the colonial triangle.

What happens at the end of this primary import substitution phase is critical; that is, once domestic markets for the non-durable consumer goods are exhausted, it is apparent that relatively natural resources rich labour surplus countries have a tendency to continue with import substitution, now shifting from labour-intensive light industries to the more capital-intensive durable consumer goods, the processing of raw materials, and the production of capital goods. At the same time, in the minority of countries which have a relatively poor natural resources base we observe a shift from a domestic to an export-market orientation for the same labour-intensive non-durable consumer goods. In that case the export sector now constitutes a powerful new production function available to the economy through which traditional and, later, non-traditional exports can be converted into imported capital goods and raw materials. Moreover, the openness of the economy permits foreign capital to provide additional finance in support of the balanced growth process. Finally, an important potential advantage of the economy's openness is, of course, the whole range of additional technological alternatives now made available, which, hopefully with modifications and adaptations, can help increase the efficiency and speed of the balanced growth process.

The open economy, in other words, not only permits the labour surplus economy to harvest the normal gains from trade, to benefit from the vent for surplus of previously underutilized resources – in this case not only raw materials but also unskilled labour – but also, dynamically, to affect the direction of technology change and thus introduce competitive forces and ideas from abroad which are able to diffuse throughout the economy and are undoubtedly of considerable importance in determining the success of the labour surplus economy's transition efforts.

Extensions and critiques

Up to now we have focused exclusively on owner-operated agriculture as the typical representative of the non-commercialized sector of the labour surplus economy. It should, however, be recognized that there are very likely to exist substantial portions of non-agricultural activities, both rural and urban, and both industry and services-oriented, which are labour surplus in the way we have defined the condition. This time, the cooperating factor in short supply is capital. Most relevant is the so-called informal sector – both rural but most heavily urban – which occupies a large, often dominant, position in many developing countries. Family and cooperative ventures in this setting are characterized by the same sharing of total income, that is, a bargaining wage, coupled with low marginal productivity, that we encountered in subsistence agriculture. We are here including not only the substantial portions of both the rural and urban populations engaged in distributive trades and services – ranging from the vendors of tea, flowers and cigarettes to barbers, bootblacks and car watchers – but also to blacksmiths, metal workers, and repair shops that dominate the landscape in most labour surplus developing countries. Some portions of this informal sector, especially its urban branch, are likely to be static and of the labour-absorptive ‘sponge’ variety; others may be capable of technology change, of subcontracting arrangements with the urban formal or commercialized sector as well as of generating surpluses for investment in that sector. Thus, organizational dualism is quite pervasive in both rural and urban non-agriculture, even as product dualism now loses its distinctive characteristic.

As development since the 1950s has proceeded apace, some initially labour surplus countries, including Taiwan, South Korea and Thailand, have graduated from their initial labour surplus condition, evidenced by gently rising unskilled wages in both sectors, finally giving way to rapid and sustained increases as secular labour shortages make their appearance. Such a turning point was reached around 1968 in the case of Taiwan, around 1973 in the case of South Korea and around 1993 in the case of Thailand. It is also true that many developing countries, starting with up to 80 per cent of their population and 50 per cent of their output in food producing agriculture, have gradually shifted substantially into non-agricultural pursuits, with services retaining their dominant position, even as their composition has changed radically, in the commercialized direction. As a consequence, the number of contemporary developing countries with typical initial labour surplus characteristics has been declining. Nevertheless, a large preponderance of the developing world, certainly by weight of population, continues to find itself in a labour surplus condition. This holds, for example, for China and India, huge countries both currently engaged in a vigorous balanced growth effort, as well as for other parts of South Asia, much of Central America, the Caribbean and parts of South America. Even some countries of sub-Saharan Africa, once considered land surplus by some observers, may, as a consequence of population growth and the loss of land to the Sahara, be approaching labour surplus status—though, given the AIDS epidemic, this remains a more controversial issue.

It should, finally, be noted that the fundamental concept of the labour surplus economy has come under increasing attack by the dominant neoclassical school of economics. While still viewed as relevant in the South and wherever heavy population pressure on scarce cultivable land remains a feature of the landscape, most Northern economists in the Becker microeconometric tradition find it difficult to accept the notion of an exogenous or bargaining wage in the non-commercialized sectors instead of one determined endogenously by the

customary interaction between demand and supply. The crux of the critique is based on the rejection of the notion that initial conditions, that is, a highly unfavourable ratio of people to cooperating land or capital, can lead to the subsidization of some members of the society by others, in lieu of ejecting them.

The work of Rosenzweig and associates (for example, Rosenzweig, 1988), presenting evidence of rising labour supply curves in a cross-section of such heavily populated agricultural sectors as India's, typifies current mainstream rejection of the ‘unlimited supply of labour’ condition underlying the labour surplus economy construct. Yet we would contend that such efforts capture an expressly static snapshot picture, addressing cross-sectional labour–leisure decisions across households already working at full capacity (that is, with little leisure to spare), while labour surplus models are concerned with the conditions governing inter-sectoral labour reallocation over time.

The exogenous agricultural wage assumption underpinning labour surplus economies, so troubling to neoclassical economists, gets support from anthropologists like Geertz (1963) and Scott (1976), as well as from economists like Lewis (1972), Ishikawa (1975), Fei and Ranis (1964), Osmani (1991), Ohkawa (1972) and others. Fafchamps (1992) provides an overview of the principles underlying the ‘solidarity network’ among peasants as depicted in anthropological evidence. Ishikawa (1975), long an astute observer of Asian economic development, endorses the concept of a ‘minimum subsistence level of existence’ (MSL), one version of the institutional real wage. His work indicates the prevalence of a ‘community principle of employment and income distribution’. This principle promises all member MSL families… an income not less than MSL’ (Ishikawa, 1975, p. 474). Hayami and Kikuchi (1982, p. 217), basically neoclassical in outlook, find that in Indonesia

> … wage rates cannot adjust directly to changes in labor's marginal productivity. Adjustments in wage rates are allowed only through modification of institutional arrangements themselves … In other words, ‘institutional wages’ based on a system of community-wide work and income-sharing similar to the classical concept can adjust to the neoclassical equilibrium through institutional innovations.

Only over time is there a tendency to adjust, but even then it does not necessarily occur by altering wages to equal the marginal product, which could reduce the wage below subsistence. Instead, in Java harvest contracts are adjusted to include weeding duties without a complementary rise in the wage rate, thereby not threatening the MSL but moving institutionally towards equilibrium. Even Kenneth Arrow (1988), one of the high priests of neoclassical economics, states that it may take a considerable period of time before equilibrium is reached. Osmani (1991) presents a model of downward rigidity of the sharing rule insisted on by the workers themselves. Current work in what is called behavioural economics may also prove to be of help in developing a theoretical structure to rationalize cross-worker subsidization in the absence of assured reciprocity – especially as some members of the group are likely to be leaving agriculture over time.

Perhaps even more relevant, there is evidence, not only for Taiwan, Korea, and Thailand but also for post-enclosure England between 1780 and 1840 and for post-Restoration Japan between 1870 and 1920, indicating substantial increases in agricultural labour productivity while both agricultural and non-agricultural unskilled real wages were rising only gently, until commercialization was reached and wages began to rise steeply in line with rising marginal productivity. Thus, both historical and 20th-century development patterns are inconsistent with the neoclassical school's one-sector full-employment equilibrium assumptions.

In the final analysis, what is relevant is whether the labour surplus model provides a better fit for the observed empirical pattern of successful labour-abundant developing countries; whether the model is better suited to analysing relative agricultural neglect in failure cases; whether it is better able to explain changing patterns of technology choice and the direction of technology change; whether, in sum, it makes better sense than to assume away the initial existence of underemployment and disequilibrium before the one-sector, fully commercialized modern growth epoch can be reached.

See Also

- agriculture and economic development
- classical growth model
- dual economies

Bibliography

Arrow, K. 1988. General economic theory and the emergence of theories of economic development. Presidential address, 8th World Economic Congress of the International Economic Association, New Delhi, India, 1986.

Fafchamps, M. 1992. Solidarity networks in preindustrial societies: rational peasants with a moral economy. *Economic Development and Cultural Change* 41, 147–74.

Fei, J. and Ranis, G. 1964. *Development of the Labor Surplus Economy: Theory and Policy*. Homewood, IL: Richard A. Irwin, Inc.

Geertz, C. 1963. *Agricultural Involution: The Process of Ecological Change in Indonesia*. Berkeley: University of California Press.

Hayami, Y. and Kikuchi, M. 1982. *Asian Village Economy at the Crossroads*. Baltimore, MD: Johns Hopkins University Press.

Hayami, Y. and Ruttan, V. 1970. Korean rice, Taiwan rice, and Japanese agricultural stagnation: an economic consequence of colonialism. *Quarterly Journal of Economics* 84, 562–89.

Ishikawa, S. 1975. Peasant families and the agrarian community in the process of economic development. In *Agriculture in Development Theory*, ed. L. Reynolds. New Haven: Yale University Press.

Kuznets, S. 1966. Modern Economic Growth: Rate, Structure and Spread. New Haven: Yale University Press.

Lewis, W. 1972. Reflections on unlimited labor. In *International Economics and Development: Essays in Honor of Raul Prebisch*, ed. L. DiMarco. New York: Academic Press.

Nurkse, R. 1953. *Problems of Capital Formation in Underdeveloped Countries*. New York: Oxford University Press.

Ohkawa, K. 1972. *Differential Structure and Agriculture: Essays on Dualistic Growth*. Tokyo: Kinokuniya.

Osmani, S. R. 1991. Wage determination in rural labor markets: the theory of implicit cooperation. *Journal of Development* 34, 3–23.

Rosenstein-Rodan, P. 1943. The problem of industrialization of eastern and south-eastern Europe. *Economic Journal* 53, 202–11.

Rosenzweig, M. 1988. Labor markets in low income countries. In *Handbook of Development Economics*, vol. 1, ed. H. Chenery and T. N. Srinivasan. Amsterdam: North Holland.

Schultz, T. 1964. *Transforming Traditional Agriculture*. New Haven: Yale University Press.

Scott, J. C. 1976. *The Moral Economy of the Peasant*. New Haven: Yale University Press.

Sen, A. 1967. Surplus labor in India: a critique of Schultz' statistical test. *Economic Journal* 77, 154–61.

How to cite this article

Ranis, Gustav. "labour surplus economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000170> doi:10.1057/9780230226203.0919

The New Palgrave Dictionary of Economics Online

labour theory of value

Fernando Vianello
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The only instance in which Adam Smith makes the value of commodities depend on the quantity of labour required to produce them is where ‘the whole produce of labour belongs to the labourer’ (Smith, 1776, vol. 1, p. 54; see ibid., p. 72). ‘In that early and rude state of society which precedes both the accumulation of stock and the appropriation of land’, he asserts ‘the proportion between the quantities of labour necessary for acquiring different objects seems to be the only circumstance which can afford any rule for exchanging them for one another’ (ibid., p. 53).
This contention is illustrated by the famous example of the beaver and the deer:

> If among a nation of hunters, for example, it usually costs twice the labour to kill a beaver which it does to kill a deer, one beaver should naturally exchange for or be worth two deer. It is natural that what is usually the produce of two days or two hours labour, should be worth double of what is usually the produce of one day's or one hour's labour (ibid., p. 53).

According to Smith, when profit and rent make their appearance alongside the labourer's income, the above rule is no longer applicable. The price of a commodity is then obtained by adding up its ‘component parts’: wage, profit and rent. These revenues, which Smith calls ‘the three original sources … of all exchangeable value’ (ibid., p. 59), enter into the ‘natural price’ of each commodity at their respective ‘natural rates’, such that ‘the natural price itself varies with the natural rate of each of its component parts, of wages, profit and rent’ (vol. I, p. 71).
The ‘adding-up’ theory of prices must be distinguished from Smith's claim that the price of every commodity ‘resolves itself’ entirely into wage, profit and rent (see vol. I, p. 57). The latter was accepted by Ricardo and rejected by Marx. The former was rejected by both.
1. Against the ‘adding-up’ theory Ricardo sets the labour theory of value extended to the capitalist mode of production:

> All the implements necessary to kill the beaver and deer might belong to one class of men,

> and the labour employed in their destruction might be furnished by another class; still their comparative prices would be in proportion to the actual labour bestowed, both on the formation of the capital, and on the destruction of the animals. (Ricardo, 1821, p. 24)

The value of the product would go partly to the labourers and partly to the capitalists; yet

> this division could not affect the relative value of these commodities, since whether the profits of capital were greater or less, whether they were 50, 20 or 10 per cent. or whether the wages of labour were high or low, they would operate equally on both employments. (ibid.)

As gold, the standard of value, is a commodity like any other, the above argument makes the price of commodities – the exchange-ratio between each of them and gold – independent of the level of the wage, a change in which is exactly offset by a change in the opposite direction of the rate of profits: the relative weight of the two ‘component parts’, wages and profits, varies, but their sum remains the same.
According to Ricardo the value of a commodity produced from natural resources in short supply is regulated by the quantity of labour expended to produce it ‘under the most unfavourable circumstances … under which the quantity of produce required, renders it necessary to carry on the production’ (ibid., p. 73). Thus the quantity of labour governing the value of the entire quantity produced of a commodity is not that actually expended on its production, but that which would need to be expended if the entire production took place under the most unfavourable circumstances. That portion of the value which is absorbed by rent corresponds to the difference between this fictitious quantity of labour and the one actually expended on the production of the commodity. The portion of value corresponding to the quantity of labour actually expended is split up into wages and profits.
Thus the labour theory of value enables Ricardo to conceive the different revenues as resulting from the breakdown of a known magnitude, rather than that magnitude (value) as resulting from the adding up of ‘component parts’ (the different revenues) determined independently of each other. The contrast between these two conceptions is fixed by Marx in a highly effective image:

> If I determine the lengths of three different straight lines independently, and then form out of these three lines as ‘component parts’ a fourth straight line equal to their sum, it is by no means the same procedure as when I have some given straight line before me and for some purpose divide it, ‘resolve’ it, so to say, into three different parts. In the first case, the length of the line changes throughout with the lengths of the three lines whose sum it is; in the second case, the lengths of the three parts of the line are from the outset limited by the fact that they are parts of a line of given length (Marx, 1885, p. 387).

2. If gold is produced by an unchanging quantity of labour, a rise in the price of a commodity can only stem from a process of ‘extensive’ or ‘intensive’ diminishing returns (only the former, however, will be considered in what follows). In discussing the consequences of an increasing ‘difficulty of procuring the necessaries on which wages are expended’, Ricardo takes the quantities consumed by each labourer as given. It follows that, as the price of corn (a typical necessary) rises, the wage in terms of gold also rises,

and the profits of the manufacturers fall:

> suppose corn to rise in price because more labour is necessary to produce it; that cause will not raise the prices of manufactured goods in the production of which no additional quantity of labour is required. If, then, wages continued the same, the profits of manufacturers would remain the same; but if, as is absolutely certain, wages should rise with the rise of corn, then their profits would necessarily fall (Ricardo, 1821, pp. 48, 110–11).

Let us assume that the entire production of corn is initially obtained from land of uniform quality, and that thereafter, in order to increase the quantity produced, land of an inferior quality be brought into cultivation. The value of the quantity of corn produced on the second quality of land is governed by the quantity of labour actually expended on its production and ‘is divided into two portions only: one constitutes the profits of stock, the other the wages of labour’ (ibid., p. 110). The increase in the value of the quantity of corn obtained from the first quality of land is wholly swallowed up by the rent, which now begins to be paid for the use of this quality of land.

In the production of corn both expenses and proceeds per unit of produce increase. But the result is the same as in manufacturing (where only expenses increase) since the farmer ‘will not only have to pay, in common with the manufacturer, an increase of wages to each labourer he employs, but he will be obliged either to pay rent, or to employ an additional number of labourers to obtain the same produce; and the rise in the price of raw produce will be proportioned only to that rent, or that additional number, and will not compensate him for the rise of wages’ (ibid., p. 111).

What causes the ratio of profits to wages to fall is not the rise of rent, but – in agriculture as well as in manufacturing – the increase in wages consequent upon the increased expenditure of labour required to produce necessaries in the most unfavourable circumstances. If the commodities which increase in value are not among those purchased by labourers, the ratio of profits to wages remains unchanged (even though a part of the capitalist's purchasing power is transferred to the landowners).

3. What is true of the ratio of profits to wages is also true, in Ricardo's opinion, of the rate of profits, which forms his main concern. Indeed, what he does is simply to refer to the latter his conclusions regarding the former, so that the two concepts appear to shade into one another. ‘In his observations on profit and wages’, says Marx, taking up a remark of G. Ramsay's (1836, p. 174n.), ‘Ricardo … treats the matter as though the entire capital were laid out directly in wages’ (Marx, 1905–10, vol. II, p. 373).

Marx traces this confusion back to ‘the absurd dogma pervading political economy since Adam Smith, that in the final analysis the value of commodities resolves itself completely into … wages, profit and rent’ (Marx, 1894, p. 841).

Smith's teaching is that, while the price of a commodity includes – along with the revenues derived from its direct production – the value of its means of production, the latter value can be broken down in the same way, and so on, going backwards, until an *initial stage of production* is reached, in which the means of production of the stage following are produced without the aid of any other means of production. Only the value of the output in the initial stage of production resolves itself immediately into wage, profit and rent. But the output in each stage, whose value equals the sum of the revenues obtained in that stage as well as in all the preceding ones, supplies the means of production for the next stage, so

that ‘the whole price still resolves itself either immediately or ultimately into the same three parts of rent, labour, and profit’ (Smith, 1776, vol. I, p. 57; here ‘labour’ obviously stands for ‘wages’).

Marx's criticism of Smith's thesis of complete ‘resolution’ of prices into revenues is made up of two parts, which should be kept strictly distinct. The first is of a factual nature. In moving back from a commodity to its means of production, and from these to their own means of production, and so on, one will never – in Marx's view – reach an initial stage of production, since sooner or later one is bound to encounter commodities that, either directly or indirectly, participate in the production of themselves. Since one can never get rid of these commodities, however far back one goes, ‘it is [of] no avail for Adam Smith to send us from pillar to post’ (Marx, 1905–10, vol. I, p. 99).

The conception according to which commodities are produced in a finite number of stages does not, of itself, lead to a confusion between the rate of profits and the ratio of profits to wages. Since, however, in this conception the value of the means of production employed in each stage resolves itself into the revenues obtained in all the previous stages, ‘one may … imagine along with Adam Smith’ – this being the second part of Marx's criticism – ‘that constant capital is but an apparent element of commodity-value, which disappears in the total pattern’ (Marx, 1894, p. 845; by ‘constant capital’ Marx means the value of the means of production).

That in dealing with the economy as a whole Smith and Ricardo fall into this error emerges clearly, for example, from Smith's statement, repeated almost *verbatim* by Ricardo, according to which ‘what is annually saved is as regularly consumed as what is annually spent, and nearly in the same time too; but it is consumed by a different set of people’ (Smith, 1776, vol. I, p. 359; see Ricardo, 1821, p. 151n.).

The funds devoted to accumulation are here treated as wholly employed in producing the necessaries for the labourers. This may help explaining how, when Ricardo approaches the problem from the point of view of the economy as a whole, he does not seem to make any distinction between the rate of profits and the ratio of profits to wages, referring to the former as depending only on the ‘proportion of the annual labour of the country [which] is devoted to the support of the labourers’ (Ricardo, 1821, p. 49; see Sraffa, 1951, p. xxxiii).

4. Although it is the labour theory of value that makes it possible for Ricardo to determine the rate of profits, his adherence to this theory appears anything but firm. Indeed, ‘the principle that the quantity of labour bestowed on the production of commodities regulates their relative value’ turns out to be, as Ricardo puts it, ‘considerably modified’ (ibid., p. 30) by the influence of other factors.

To show this Ricardo makes use of a numerical example which deserves to be quoted in full:

> Suppose I employ twenty men at an expense of £1,000 for a year in the production of a commodity, and at the end of the year I employ twenty men again for another year, at a further expense of £1,000 in finishing or perfecting the same commodity, and that I bring it to market at the end of two years, if profits be 10 per cent., my commodity must sell for £2,310; for I have employed £1,000 capital for one year, and £2,100 capital for one year more. Another man employs precisely the same quantity of labour, but he employs it all in the first year; he employs forty men at an expense of £2,000, and at the end of the first year he sells it with 10 per cent. profit, or for £2,200. Here then are two commodities having precisely the same quantity of labour bestowed on them, one of which sells for £2,310 – the other for £2,200 (ibid., p. 37).

Let *w* be the wage (equal in the example to £50 per labourer) and *r* be the rate of profits (equal to 10 per cent). For the sake of simplicity, we shall further suppose that the quantity produced of each of the two commodities be one unit. The price of commodity *a*, the first commodity in the example, is then

$$20w(1 + r)^2 + 20w(1 + r) = P_a$$

The price of the second commodity, *b*, is instead

$$40w(1 + r) = P_b$$

Although Ricardo does not deal systematically with the subject, here, as well as in other numerical examples, he does offer a theory in embryo, which – for any given rate of profits – makes natural prices depend not only on the quantity of labour directly or indirectly expended on each commodity, but also on what we may call the *distribution over time* of that quantity of labour.

5. Since in the foregoing example the prices of the two commodities are determined on the basis of prior knowledge of the wage and the rate of profits, one may be inclined to think, with Marshall, that according to Ricardo value is regulated by the cost of production, which includes ‘Time or Waiting as well as Labour’; and that Marx wrongly interpreted his doctrine ‘to mean that interest does not enter into that cost of production which governs … value’ (Marshall, 1920, p. 672 and pp. 672–3, n. 1). That this is not the case will emerge clearly if we look at Ricardo's approach to the problem of relative price variation as set forth in a numerical example contained in his 1823 paper on *Absolute Value and Exchangeable Value* (Ricardo, 1823, pp. 383–4); an example which closely follows the one we have just examined (the only differences, which we shall ignore, being that the prices of commodities *a* and *b* corresponding to *r*=10 per cent are said to be £231 and £220 respectively, rather than £2,310 and £2,200, and that a third commodity is also considered).

Ricardo supposes ‘labour to rise in value and profits to fall – that from 10 pc^t they fall to 5 pc^t. He further supposes that commodity *b* be the standard of value. Making the two examples into a single one, we shall suppose that gold is produced in a single stage. If, then, the price of commodity *b* is £2,200, that is not because the wage is £50 and the rate of profits 10 per cent, but rather because it has been produced, like gold, in a single stage, employing a quantity of labour equal to 2,200 times that required to produce the quantity of gold corresponding to £1. The fall in the rate of profits from 10 to 5 per cent will thus leave the price of commodity *b* unchanged; which amounts to saying that in its production (as in that of gold) the increase in wages and the fall in profits offset each other.

However, the same increase in *w* and fall in *r* cannot bring about a similar offsetting in the case of commodity *a*, whose price must fall from £2,310 to £2,255 (from £231 to £225.5 in Ricardo's 1823 example). This result is obtained by applying the rate of profits of 5 per cent (instead of 10 per cent) to

the value of the means of production employed in the second stage of production of commodity *a*. The latter value, £1,100, does not vary, since the means of production are produced, like gold (and commodity *b*), in a single stage. The value of the term $20w(1+r)^2$ in the equation of commodity *a* falls, therefore, from £1,210 to £1,155. The value of the second term in the sum, $20w(1+r)$=£1,100, can be assimilated to the unchanging value of a commodity produced in a single stage.

It is evident that, if gold were produced in two years, with the same proportional distribution of labour between the two corresponding stages of production as commodity *a*, the new ratio P_a/P_b would emerge from a rise in P_b with P_a constant. It is also evident that, if all commodities were produced with the same proportional distribution of labour over time, they would all be in the same situation as gold, in whose production an increase (fall) in wages is exactly offset by the corresponding fall (increase) in profits, and the labour theory of value would stand in no need of 'modification'.

The 'modifications' have, therefore, nothing to do with the alleged necessity of adding to the labour what is depicted as a second element of the cost of production. The misunderstanding may be traced back to Malthus, who ascribes to Ricardo the very fault that Marshall seeks to acquit him of, shifting the blame onto Marx. 'We have the power indeed', Malthus remarks:,

> 'arbitrarily, to call the labour which has been employed upon a commodity its real value, but in so doing, we use words in a different sense from that in which they are customarily used; we confound at once the very important distinction between *cost* and *value*; and render it almost impossible to explain with clearness, the main stimulus to the production of wealth, which in fact depends upon that distinction'.

To which Ricardo counters:

> Mr Malthus appears to think that it is part of my doctrine, that the cost and value of a thing should be the same; – it is, if he means by cost, 'cost of production' including profits. In the above passage, this is what he does not mean, and therefore he has not clearly understood me (Ricardo, 1821, p. 47n.).

What Ricardo makes clear in this passage (which, surprisingly enough, Marshall quotes as evidence in support of his reading of the matter: see Marshall, 1920, p. 672) is that the labour theory of value, in its 'unmodified' as well as its 'modified' form, takes full account of 'the very important distinction between cost and value'; that is, of the existence of profits ('the main stimulus to the production of wealth'). What equals value according to this theory is not, Ricardo argues, 'cost' as commonly understood, but 'cost of production including profits', profits being what is left of the value of a commodity once wages have been deducted. (Reference to the most unfavourable circumstances under which production is carried on has been dropped since the preceding section, land being now supposed to be abundant and all of the same quality.)

6. The reader will perhaps have noted how Ricardo omits to specify by how much the wage must increase in order to cause a fall from ten to five per cent in the rate of profits (elsewhere, again when dealing with the problem of relative price variation, he postulates 'such a rise of wages as should occasion a fall of one per cent. in profits': Ricardo, 1821, p. 36). Even though Ricardo continues to

express himself as if, in the relation between *w* and *r*, the independent variable were represented by the wage, in actual fact he reverses the roles, and makes *w* depend on *r*. The value of *w* when *r*=10 per cent is, as we know, *w*=£50. Its value when *r*=5 per cent can be calculated from the equation of commodity *b* (whose price remains £2,200). This value is slightly less than *w*=£52. 8•s. 0d.

As a matter of fact, Ricardo's argument is made up of two distinct stages. In the first of these the rate of profits is determined on the basis of the ‘unmodified’ labour theory of value; in this stage the necessaries consumed by each labourer are taken as given (see section 2 above). The second stage takes the rate of profits as given, the problem being now to determine the prices which make the rate of profits uniform throughout the economy. These prices, as Ricardo realizes, are not regulated by the quantities of labour expended on the production of the commodities, as they were assumed to be for the purpose of determining the rate of profits. And the wage (the £52. 8•s. 0d or so of the example) will in general turn out to be different from the value of the necessaries it was assumed to purchase in the first stage of the argument.

It does not escape Ricardo that the rate of profits should be determined on the basis of the ‘modified’ theory, and therefore of prices which, in turn, cannot be determined before the rate of profits is known. But he is unable to provide a theoretical construction capable of coping with this interdependence. Thus he does not see any other solution but that of continuing to base his analysis of income distribution on the ‘unmodified’ labour theory of value, which he defends as ‘the nearest approximation to truth as a rule for measuring relative value, as any I have ever heard’ (Letter to Malthus of 9 October 1820, in Ricardo, 1951–73, vol. VIII, p. 279).

7. A major difference between the Ricardian version of the labour theory of value and its Marxian version, to which we must now turn, lies precisely here: that the former can be described as an approximation, whereas the latter cannot. According to Marx the values of commodities exactly (not approximate) reflect the quantities of labour expended on their production, although this is not true, in general, of the ‘prices of production’ (Marx's name for ‘natural prices’), which coexist with values.

In discussing Marx's position we shall reckon the value of commodities directly in units of labour (say, man-years). The value of the means of production which assist one labourer in the annual cycle of production of any particular commodity, or *constant capital* per unit of labour (*c*), and the value of one labourer's necessaries, or *variable capital* per unit of labour (*v*), are thus made equal to the quantities of labour expended on the production of those means of production and of those necessaries respectively. If only circulating capital is used, the value of the output per unit of labour of any commodity is (c+1), or *c* plus the value added per unit of labour. Since *v* is uniform throughout the economy (each labourer being assumed to consume the same bundle of commodities), the *surplus-value* per unit of labour ($1-v$) will also be uniform. The same is obviously true of the ratio of surplus-value to variable capital (the *rate of surplus-value*), but not, in general, of the ratio of surplus-value to total (i.e. constant plus variable) capital. The latter ratio will be the higher, in any particular branch of production, the lower is the ratio c/v (the *organic composition of capital*).

Competition, however, redistributes the overall surplus-value of the economy among the various branches of production in such a way as to render it proportional not to the variable, but to the total capital. Thus a general rate of profits comes to be established, equal to the weighted average of the ($1-v$) to ($c+v$) ratios in the different branches of production – or, which amounts to the same thing, to the ratio of the overall surplus-value of the economy to the overall capital employed. The same mechanism establishes the prices of production, which make that rate of profits uniform throughout the economy.

Unlike Ricardo's Marx's argument is *explicitly* framed in two stages. Since the prices of production differ from the values only on account of the different distribution of the overall surplus-value of `the economy, according to Marx the rate of profits is accurately determined, for the economy as a whole, on the basis of the labour theory of value. The prices of production are then obtained from the values by replacing the surplus-value produced in each branch of production with the part of the overall surplus-value of the economy belonging to that branch according to the general rate of profits.

8. ‘Surplus-value and the rate of surplus-value’, says Marx, ‘are, relatively, the invisible and unknown essence that wants investigating, while rate of profit and therefore the appearance of surplus-value in the form of profit are revealed on the surface of the phenomenon’ (Marx, 1894, p. 43). To reveal the invisible: herein lies the task of science. But Marx's theoretical programme also involves explaining just *why* the intimate essence of things in invisible, why it does not reveal itself ‘on the surface of the phenomenon’. Marx's explanation is that those ‘who are entrapped in bourgeois production relations’ (ibid., p. 817) witness the *result* of the redistribution of surplus-value–the profit proportional to capital – but not the *process* leading up to this result:

> The actual difference of magnitude between profit and surplus-value … in the various spheres of production now completely conceals the true nature and origin of profit not only from the capitalist, who has a special interest in deceiving himself on this score, but also from the labourer (ibid., p. 168).

Thus it comes about that

> the splitting of the value of commodities after subtracting the value of the means of production consumed in their creation; the splitting of this given quantity of value, determined by the quantity of labour incorporated in the produced commodities into three component parts … appears in a perverted form on the surface of capitalist production,

wage, profit and rent taking on the aspect of ‘independent revenues in relation to one another, and as such related to three very dissimilar production factors, namely labour, capital and land’, from which ‘they seem to arise’ (ibid., pp. 867–8; we shall, however, continue to assume the absence of rent). ‘To have destroyed this false appearance and illusion’ represents ‘the great merit of [classical] political economy’ (ibid., p. 830). Against classical political economy – of which Ricardo is the ‘last great representative’ (Marx, 1873, p. 24) – Marx sets ‘vulgar’ economy: the first of these studied ‘the real relation of production in bourgeois society’, whereas the second ‘deals with appearances only’ (Marx, 1867, p. 85, n.1).

But even Ricardo cannot be completely acquitted, in Marx's opinion, of having taken as the *starting-point* of the argument the *result* of the redistribution of surplus-value. Indeed, it is the natural prices themselves that Ricardo claims are regulated (even if only approximately; but, as will be remembered, it is the nearest approximation to truth’ among those available; see section 6 above) by the quantities of labour expended on the production of commodities. Hence Marx's allegation that Ricardo confuses values and prices of production.

If Ricardo is compelled to presuppose what he should explain (the profit proportional to capital, as it

emerges from the redistribution of surplus-value), this is – according to Marx – because his unsatisfactory treatment of non-wage capital (see section 3 above) blinds him to the distinction between surplus-value and profit:

> Ricardo wrongly identifies surplus-value with profit … these are only identical in so far as the total capital consists of variable capital or is laid out directly in wages … Ricardo evidently shares Smith's view that the *total value* of the annual product resolves itself into revenues. Hence also his confusion of value with cost-price (Marx, 1905–10, vol. II, p. 426; as so often in *Theories of Surplus-Value*, ‘cost-price’ here stands for ‘price of production’).

Here, in Marx's opinion, lies the origin of the analytical difficulties with which Ricardo had to wrestle and which Marx himself claims to have overcome, thanks to his discovery of the redistribution mechanism.

9. On 24 August 1867, a few days after correcting the proofs of the first volume of *Capital*, Marx wrote to Engels:

> The best points in my book are: (1) the *double character of labour*, according to whether it is expressed in use value or exchange value (*all* understanding of the facts depends upon this …) (2) the treatment of *surplus-value independently of its particular forms* as profit, interest, ground rent, etc (Marx and Engels, 1942, pp. 226–7).

The second of these two contributions has been dealt with in sections 7 and 8 above (and something more on the subject will be said in section 11 below), within the limits of the hypothesis that all surplus-value is received in the form of profit. We must now turn to the first contribution – the one on which ‘*all* understanding of the facts’ is based: the ‘double character of labour’.

In the production of commodities the distribution of labour in a society among its various productive activities is not regulated *a priori*, through some form of agreement or coercion, but only *a posteriori*, through the exchange of products (Marx, 1867, p. 336). The labour of individuals is therefore not, immediately, the labour *of society* – as is the case in, say, a peasant family, within which ‘the labour-power of each individual, by its very nature, operates … merely as a definite portion of the whole labour-power of the family’ (Marx, 1867, p. 82; see Marx, 1859, p. 33). On the contrary, we are dealing here with ‘the labour of private individuals or groups of individuals who carry on their work independently of each other’; this labour ‘asserts itself as a part of the labour of society, only by means of the relation which the act of exchange establishes directly between the products, and indirectly, through them, between the producers’ (Marx, 1867, pp. 77–8). It is only when the social division of labour takes this particular form that the products of labour become commodities, or acquire the quality of possessing value.

In the first chapter of *Capital* (as well as in the first chapter of *A Contribution to the Critique of Political Economy*) Marx emphasizes how in the eyes of producers commodities count not for their ability to satisfy this or that human want, but rather for their ability to find a purchaser: not for their use-value but for their (exchange-) value. Of these two qualities of commodities, use-value is the one abstracted from

in the exchange, which cancels the difference between the products, in the sense that in the exchange different products are equated, or treated as equal, and reduced to their quality of possessing value. Labour participates in the two-fold character of commodities, as useful things and things possessing value. On the one hand, ‘it must, as a definite useful kind of labour, satisfy a definite social want, and thus hold its place as a part and parcel of the collective labour of all, as a branch of a social division of labour’ (Marx, 1867, p. 78). On the other hand, just as ‘in viewing the coat and linen as values, we abstract from their different use-values, so it is with the labour represented by those values: we disregard the difference between its useful forms, weaving and tailoring’ (ibid., p. 52); which is what producers themselves actually do, production of commodities being *production for value* – production, therefore, of abstract wealth, indifferent to its material content. What remains is a uniform, undifferentiated labour, which ‘counts only quantitatively’, having been ‘reduced to human labour, pure and simple’ (p. 52), to ‘abstract human labour’ (p. 81). Such is the labour which, embodied in commodities, figures as their value.

‘Whenever, by an exchange’, Marx writes, ‘we equate as values our different products, by that very act, we also equate, as human labour, the different kinds of labour expended upon them. We are not aware of this, nevertheless we do it’ (ibid., pp. 78–9). The reduction of a commodity to its mere quality of possessing value and the reduction of labour to abstract labour are thus in Marx's conception the outcome of one and the same real process (see Colletti, 1968, sect. 8). And it is only by being reduced to abstract labour and assuming the form of a quality of commodities, their value, that the *private* labour of the weaver and the *private* labour of the tailor enter into relation with each other, becoming part of a *social* division of labour. This is, in Marx's words, ‘the specific manner in which the social character of labour is established’ (Marx, 1859, p. 32) in the production of commodities. ‘But what is the value of a commodity?’, Marx enquires. ‘The objective form of the social labour expended on its production’ (Marx, 1867, p. 501). Or, to put it another way, abstract labour (social only in so far as abstract) represents ‘the substance of value’ (ibid., p. 46).

10. The picture is now complete, and we can attempt to gather together the threads of Marx's position. As we have just seen, the thesis of the reduction of labour to abstract labour is put forward by Marx in close connection with his theory of value. Indeed, the two merge into one, abstract labour being indicated as the substance of value and value as the form that labour must assume in order to acquire a social character. It remains to be added that the conception of abstract labour as the substance of value presupposes the sort of redistribution mechanism described in section 7 above. What constitutes the substance of value cannot, in fact, but constitute the substance of revenues, as the latter stem from the breakdown of the value of a given set of commodities. It follows that the conception of abstract labour as the substance of value necessitates that the whole of this substance be found in the prices of production, having merely been partly diverted away from some commodities and channelled into others (see the enlightening comparison with the ‘conservation of energy’ in Lippi, 1976, pp. 50–52). If this is not the case, then the aforesaid substance is not the ‘substance’ of anything real, and ‘value’ is merely a name for the quantity of labour directly and indirectly expended on the production of a commodity.

11. In the Afterword to the second (German) edition of *Capital* we read that ‘the method of presentation must differ in form from that of inquiry’ (Marx, 1873, p. 28). We are, now in a position to understand this celebrated (as much as hermetic) warning. If we attend to the ‘method of inquiry’, the theory of the rate of profits and of the prices of production (contained in the manuscripts published posthumously as

the third volume of *Capital*) represents – as stated in the preceding section – a premise for the conception of abstract labour as the substance of value, and the cornerstone of the whole theoretical structure of *Capital*. (From a chronological point of view, it has been remarked that ‘once Marx had attained – at the beginning of 1858 – what he regarded as the correct solution of the problem of how to determine the rate of profit, various elements in his thinking seem to have found an organic unity in the concept of value – the concept of a “substance” to be redistributed’ (Ginzburg, 1985, pp. 105–6); the ‘various elements’ being basically Marx's analysis of the social division of labour and his theory of income distribution and prices.)

But if, instead, we attend to the ‘method of presentation’, things take on a rather different aspect. Marx calls his own presentation of the argument ‘genetical’, meaning by this that it consists in ‘elaborating how the various forms come into being’ (Marx, 1905–10, vol. III, p. 500), proceeding from the form of value that labour assumes in the act of acquiring a social character, to arrive at surplus-value, the redistribution mechanism and the establishment of a general rate of profits.

The two ‘methods’, or procedures, reflect the two different aims mentioned in section 8 above: the aim (proper to scientific analysis) of tearing away the veil of appearances, and the aim (proper to genetical presentation) of showing how that veil is woven together. The latter aim is not regarded by Marx as less important than the former, to explain how appearances are produced being in his opinion the only sure way of evading their deceptions.

As we have already seen, Ricardo himself is believed by Marx to be partly the victim of such deceptions, even while he contributed so greatly towards dispelling them. In conceiving the labour theory of value as a theory of natural prices, Ricardo ‘omits some essential links and *directly* seeks to prove the congruity of economic categories with one another’ (Marx, 1905–10, vol. II, p. 165). He does so by taking ‘the rate of profits as something pre-existent which, therefore, even plays a part in the determination of *value*’ (ibid., p. 434), thus missing the inner connection of forms which is reflected in Marx's genetical presentation, and according to which ‘the determination of value is the primary factor, antecedent to the rate of profits and to the establishment of production prices’ (ibid., vol. III, p. 377; see Gajano, 1979, ch. 3).

12. If, however, the presentation must proceed from value to the rate of profits and the prices of production, it must assume (at least provisionally) that the foundation of value be independent of what comes after, as a result of the redistribution of surplus-value. Marx thus finds himself in an *impasse*, no such independent foundation being provided by his analysis.

So it comes as no surprise that value is introduced in *Capital* in a rather sketchy way. Marx starts by declaring, as something self-evident, that in two commodities equated in exchange ‘there exists in equal quantities something common to both’ (Marx, 1867, p. 45). He then goes on to enquire wherein this common element consists. It is at this point that we meet the argument according to which exchange involves an abstraction from the use-value of the commodities exchanged (‘the exchange of commodities is evidently an act characterised by a total abstraction from use-value’: (ibid., p. 45; see section 9 above). But, Marx pursues, ‘if then we leave out of consideration the use-value of commodities, they have only one common property left, that of being products of labour’ (ibid., p. 45). Thus he does his best to lead the reader into thinking that the prices of commodities are regulated by the quantity of labour expended on their production (otherwise the common element would not be ‘in equal quantities’). Only later on does Marx put the reader on his guard with sporadic and obscure hints. (‘Average prices do not directly coincide with the values of commodities, as Adam Smith, Ricardo, and

others believe': ibid., p. 163n.; see ibid., p. 212n., where the reader is referred to vol. III – unpublished – and ibid., p. 290, where Marx mentions the 'many intermediate terms' wanted to resolve the 'apparent contradiction' between the labour theory of value and the existence of a uniform rate of profits.)
'Analysis', writes Marx, 'is the necessary prerequisite of genetical presentation' (Marx, 1905–10, vol. III, p. 500). But it is a prerequisite which cannot be openly declared if presentation is to remain genetical. This limitation has given birth to two opposite and equally wrong interpretations. The one holds that Marx's theory of value has no foundation whatsoever, and treats that theory and the theory of prices of production as two mutually incompatible theories of prices (this is the thesis of the 'contradiction' between the first and the third volumes of *Capital*, put forward in Böhm-Bawerk, 1896). The other interpretation tries to defend the labour theory of value on the basis of Marx's analysis of the social division of labour, making no appeal to the redistribution mechanism and maintaining, in the last analysis, that labour forms the substance of value because it is through the exchange of commodities that the various labours, performed outside any conscious coordination, enter into relation with one another (this traditional Marxist reply to Böhm-Bawerk's criticism first appears in Hilferding, 1904, and finds its best expression in Colletti, 1968).
Obviously the labour theory of value cannot be defended on the grounds indicated by Hilferding and Colletti (as the latter has acknowledged: see Colletti, 1979). But, just as obviously, Böhm-Bawerk's grounds for dismissing it are not good ones. Actually, the reason why the labour theory of value must be rejected is not that it is devoid of foundation, but rather that what in Marx's view represents its foundation – his theory of the rate of profits and of prices of production – proves untenable in the light of the subsequent work of Tugan-Baranovsky (1905), Bortkiewicz (1907) and others, up to Sraffa (1960).

See Also

- British classical economics
- Marxian value analysis
- natural price
- Ricardo, David

Bibliography

Böhm-Bawerk, E. von. 1896. *Karl Marx and the Close of His System*. In Sweezy (1949).

Bortkiewicz, L. von. 1907. *On the Correction of Marx's Fundamental Theoretical Construction*. In Sweezy (1949).

Colletti, L. 1968. Bernstein and the Marxism of the Second International. In L. Colletti, *From Rousseau to Lenin: Studies in Ideology and Society*, London: New Left Books, 1972.

Colletti, L. 1979. *Tra marxismo e no*. Bari: Laterza.

Gajano, A. 1979. *La dialettica della merce. Introduzione allo studio di 'Per la critica dell'economia politica' di Marx*. Napoli: il Laboratorio.

Ginzburg, A. 1985. A journey to Manchester: a change in Marx's economic conceptions. *Political Economy* 1.

Hilferding, R. 1904. *Böhm-Bawerk's criticism of Marx*. In Sweezy (1949).

Lippi, M. 1976. *Value and Naturalism in Marx*. London: New Left Books, 1979.

Marshall, A. 1920. *Principles of Economics*. 8th edn. London: Macmillan, 1964.

Marx, K. 1859. *A Contribution to the Critique of Political Economy*. Moscow: Progress Publishers, 1978.

Marx, K. 1867. *Capital: A Critique of Political Economy*, Vol. I. London: Lawrence & Wishart, 1977.

Marx, K. 1873. *Afterword to the 2nd German edition*. In Marx (1867).

Marx, K. 1885. *Capital: A Critique of Political Economy*, Vol. II. London: Lawrence & Wishart, 1974.

Marx, K. 1894. *Capital: A Critique of Political Economy*, Vol. III. London: Lawrence & Wishart, 1974.

Marx, K. 1905–10. *Theories of Surplus-Value*. Vol. I, 1978; Vol. II, 1975; Vol. III, Moscow: Progress Publishers, 1975.

Marx, K. and Engels, F. 1942. *Selected Correspondence 1846–1895*. New York: International Publishers.

Ramsay, G. 1836. *An Essay on the Distribution of Wealth*. Edinburgh.

Ricardo, D. 1821. *On the Principles of Political Economy and Taxation*. 3rd edn, in Ricardo (1951–73), Vol. I.

Ricardo, D. 1823. *Absolute Value and Exchangeable Value*. In Ricardo (1951–73), Vol. IV.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*. Ed. P. Sraffa with the collaboration of M.H. Dobb, Cambridge: Cambridge University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. E. Cannan.

London: Methuen, 1961.

Sraffa, P. 1951. *Introduction*. In Ricardo (1951–73), Vol. I.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Sweezy, P.M., ed. 1949. *Karl Marx and the Close of His System* by Eugen von Böhm-Bawerk & ‘Böhm-Bawerk's Criticism of Marx’ by Rudolf Hilferding, together with an appendix consisting of an article by Ladislaus von Bortkiewicz on the Transformation of Values into Prices of Production in the Marxian System. New York: Augustus M. Kelley, 1966.

Tugan-Baranovsky, M. 1905. *Theoretische Grundlagen des Marxismus*. Leipzig: Duncker & Humblot.

How to cite this article

Vianello, Fernando. "labour theory of value." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000012> doi:10.1057/9780230226203.0920

The New Palgrave Dictionary of Economics Online

labour-managed firms

Louis Putterman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Labour-managed firms (LMFs) are enterprises over which suppliers of labour hold full control rights. Theoretical analysis suggests that such firms will behave in a distinctive and sometimes ‘perverse’ manner in response to short-run changes, but richer models can reverse the more problematic results, and the simple model indicates that LMFs behave no differently from capitalist firms in long-run competitive equilibrium. Empirical studies indicate that LMFs, while uncommon in most market economies, can achieve high productive efficiency. The search for an understanding of why LMFs are relatively rare has contributed to both positive and normative economic analysis.

Keywords

codetermination; democracy; efficiency wages; firm, theory of; free-rider problem; labour-managed firms; Mill, J. S.; ownership and control; partnerships; profit sharing; socialism; total factor productivity; wage differentials

Article

Although the traditional theory of the firm gave little attention to institutional detail, the common assumption about the units that engage in the production and sale of goods and services was that they are owned and controlled by individuals who provide risk-bearing capital and who hire the services of workers as one among several variable inputs. Worker-run cooperatives had existed in small numbers at least since the Industrial Revolution, but the study of such firms using formal analytical tools awaited the added stimuli provided by the challenge of understanding collective farm performance in the Soviet Union and China, and Yugoslavia's experiment with worker-managed market socialism. The models developed in the late 1950s and thereafter were subsequently applied not only to those cases but also to understanding worker-owned firms in industrial market economies, to investigating hypothetical economies consisting exclusively of worker-run firms, and to attempting to explain why worker control

is relatively rare. As studies on the topic multiplied, the term ‘labour-managed firm’ (LMF) came to be used by economists to describe an enterprise that operates under the ultimate control of those who work in it.

Such a definition of an LMF permits considerable variation in other dimensions. To qualify as an LMF, for example, an enterprise's workers must have control in the sense that managers are appointed and can be removed by them or by their representatives. But the degree of direct worker involvement in decision-making can vary, from the more direct democracy of small cooperatives to the representative structures of large Mondragon cooperatives or the now-defunct Yugoslav firms. A frequent assumption is that the exercise of worker control follows ‘one worker one vote’ lines, but the LMF concept has sometimes been extended to firms that include a class of workers lacking control rights. Most importantly, perhaps, the term LMF has been applied both to firms in socialist economies, in which the private ownership of capital is prohibited and the enterprise's capital is the property of ‘society’ or of a collective, and to worker-owned firms in capitalist economies, in which individual workers can hold property rights in their enterprise's assets, for example through ‘partnership deeds’, ‘individual capital accounts’ or shares. The principal example of an LMF with ‘social capital’ was the Yugoslav social enterprise, which arose from the application of new laws and principles to that country's Soviet-style state enterprises. Collective property was the prevailing legal notion applied to the land and equipment of collective farms in the Soviet Union, China, and other Communist states, and has also accounted for a portion of the assets of some Western worker-run firms. The canonical example of ‘partnership deeds’ is provided by worker-owned plywood companies in the United States. The capital account model was adopted by the group of worker-owned enterprises centered in the town of Mondragon in the Basque province of Spain. More hybrid cases with only elements of worker control, such as (*a*) the partial employee ownership of many American companies, (*b*) legal, medical, and other professional partnerships, (*c*) co-determination in Western Europe, and (*d*) the widespread employee ownership resulting from privatization programmes in many transition economies, also continue to stimulate interest in the economic analysis of firms run by workers.

Although the economic analysis of worker-run firms was stimulated by the cases mentioned, interest in the concept appears to be explained by other factors as well. Normative dissatisfaction with the capitalist employment relationship, in which workers assume a subordinate role in the production process and lack claims on enterprise profits, can be found among leading economists ranging from John Stuart Mill and Leon Walras to James Meade and Jacques Drèze. In his *Principles of Political Economy*, Mill, who dominated English political-economy in the mid-19th century, wrote

> To work at the bidding and for the profit of another, without any interest in the work – the price of their labour being adjusted by hostile competition, one side demanding as much and the other paying as little as possible – is not, even when wages are high, a satisfactory state to human beings of educated intelligence, who have ceased to think themselves naturally inferior to those whom they serve. (Mill, 1848, pp. 760–1, n. 1)

He predicted the extinction of the capitalist firm (‘There can be little doubt … that the relation of masters and workpeople will be gradually superseded by partnership’, pp. 763–4) and opined that the result ‘would be the nearest approach to social justice, and the most beneficial ordering of industrial affairs for the universal good, which it is possible at present to foresee’ (p. 792). Modern political

theorists such as Carole Pateman (1970) and Robert Dahl (1985) have argued that self-government of the workplace by workers is an implied requirement of the principle of control of government by the governed, and that it would help to deepen democracy in more traditional political spheres.

Another source of interest in LMFs is the fact that the theoretical analysis of such firms promises insights into why the large majority of firms in market economies are established and controlled by investors rather than workers (Dow, 2003). Whether that fact is to be attributed to social custom, to the exercise of economic power by the wealthy, to aversion to risk by the poor, or to other factors, seems important for judging policies such as the expansion of codetermination or the use of worker ownership in future privatizations. It also has an important part to play in the ethical evaluation of the economic system as a whole.

The first wave of models of worker-management abstracted from issues of ownership and financing by assuming a fixed charge for capital or land, presumed to be rented by the firm but fixed in quantity in the short run. By contrast, the number of worker-members was taken to be variable, and the firm's main decision problem was to select a level of this input. In the seminal model of Ward (1958) and in subsequent treatments by Domar (1966), Vanek (1970), Meade (1972) and others, the objective was taken to be maximizing revenue per worker net of capital, land, or other charges. The first and most frequently noted finding of such models was that, with the maximand being the (endogenous or firm-specific) net earnings of a variable input, output might not respond normally to changes in the product price. In particular, Ward showed that, if labour is the only variable input, workers share net revenue on an equal basis, and the firm's objective is to maximize the earnings of each worker employed (without concern for workers who might have to be expelled to achieve earnings maximization for those remaining), then an increase in the product price would reduce optimal employment and thus the firm's output level. An industry consisting entirely of worker-run firms would accordingly exhibit a downward- rather than upward-sloping short-run supply curve, so that output would go down, rather than up, in response to increased demand (on the assumption that a short-run equilibrium is even possible). Labour would be misallocated among firms in the short-run equilibrium of a labour-managed economy, since those with high marginal product of labour would have no incentive to accept workers from those with low marginal product. As an added oddity, the firm would seek more workers if the cost of its fixed factor or a lump sum tax rose, and it would reduce its membership if the opposite occurred.

Long-run outcomes are less peculiar. Abnormal returns would attract new capital investments by existing firms and entry of other firms into the industry, giving the long-run supply curve a more conventional shape. In the very long run, with both the number of firms and their utilization of all factors being variables, equilibrium behaviour of labour-managed and conventional firms would be identical (Drèze, 1976). Even short-run perverse supply responses would be rendered unlikely by a variety of factors. For example, Domar (1966) showed that the tendency of hypothetical LMFs to take on additional workers, as output prices fell or as net revenue was reduced by higher charges for fixed factors, could be annulled by incorporating in the model the supply of labour facing a firm. Other factors tending to weaken or reverse the ‘perverse output supply response’ include (*a*) use of variable inputs additional to labour, (*b*) flexibility of working hours, (*c*) reallocation of labour between product lines in multi-product firms, (*d*) reluctance to vote for the expulsion of incumbent members, perhaps because the voters face similar probabilities of being selected for expulsion, and (*e*) tradable membership rights. Empirical research failed to provide evidence for backward supply responses by LMFs. Chinese

collective farms were found to increase their output in response to higher government-set prices. Yugoslav firms were sometimes argued to be reluctant to take on new workers, in line with Ward model predictions, but no evidence has been adduced that they had insufficient flexibility over work hours or an inability to allocate workers among tasks and product lines so as to respond positively to better market conditions for a given product. In what is probably the most rigorous study of the supply response of worker-owned firms, that on US plywood cooperatives by Craig and Pencavel (1992), the authors concluded that the firms' output was significantly less responsive to product price changes than that of conventionally owned competitors, but they rejected backward bending supply at high levels of significance.

Property rights and investment incentives were another major concern of the LMF literature beginning in the late 1960s. In Yugoslavia, workers were empowered to elect councils which selected and had governing authority over their companies' managers, but the capital stock of the company was legally owned 'by society', with workers having rights to current revenue but obligations to maintain and ideally to add to that stock. Furubotn and Pejovich (1970) demonstrated theoretically that with this rights structure self-interested workers would privately value new investments in their company only in so far as they expected to remain employed there and have their pay enhanced by the resulting higher productivity. For capital goods having a useful life exceeding the expected employment horizon of a worker, the privately appropriable rate of return must be adjusted downward to take into account truncation of the future earnings stream from the standpoint of the worker. Furubotn and Pejovich argued that Yugoslavia avoided an otherwise predicted dearth of investment only because government and Communist authorities continued to have considerable leverage over managers, and because the government encouraged companies to finance their investments with low-cost loans from the state banks, although this had the effect of pumping money into the economy and thereby fuelling inflation (Pejovich, 1969).

Most economists studying the issue agreed that firms with social ownership of capital would suffer from a horizon problem of the sort that Furubotn and Pejovich identified. More generally, Vanek (1977) argued that failure to consider the scarcity price of capital can lead to inappropriate choice of technology, a factor that he viewed as being of sufficient importance to explain the historical failure of experiments with workers' management. He noted, however, that this need not be a general feature of LMFs. The truncation of the revenue stream that is considered when evaluating investments is a result not of worker control but of assuming that workers are deprived of any and all rights to their investments' returns after separation from their firm. The problem could thus be ameliorated or eliminated entirely by several methods, for instance the calculation of a severance payment based on the capitalized value of each worker's past contributions to their company's capital stock. Another possibility is for the worker to sell his position as a partner or member of the firm in a market. In a perfectly functioning membership market, the estimated remaining productivity or marketable value of physical and other assets created during the incumbent worker's career with the firm would be incorporated in the sale price of the membership right. Sertel (1982), Dow (1986), and Fehr (1993) demonstrated the theoretical ability of a membership market to eliminate the inefficiencies of worker control in other dimensions as well. Pencavel (2001) and Dow (2003), however, point out the rarity of such markets and evidence of their imperfect functioning, suggesting this as another place to search for possible explanations of why LMFs are not more common.

A much-discussed dimension of worker control and ownership is that of work incentives. Vanek argued that, as a means of motivating workers to give their full energies to their jobs, sharing profits is likely to be far superior to paying a fixed wage, since the worker on fixed pay receives the contractual wage regardless of how intensively she works and regardless of how the firm fares. At a theoretical level, such a claim can be disputed. On the one hand, the short-run insulation of the worker from the effects of her varying quantity or quality of effort need not imply the total absence of a connection, since the wage can be adjusted over time, including by performance-contingent promotions. Efficiency wage models also demonstrate the potential to elicit effort through the threat of firing for sub-par performance. A company's very survival may depend on the effort it obtains from its workforce. On the other hand, if workers share equally or according to predetermined proportions in the same pool of profit, then the incentive provided by profit-sharing suffers from the profit's dilution among many workers, and the prediction of a static or finitely repeated model of effort choice is that rational workers will choose to free ride.

Despite this inconclusiveness of theory, empirical studies have given Vanek's intuition about profit-sharing and motivation more support than refutation. Profit-sharing has often appeared to boost work incentives, in part because it changes the dynamics of worker–worker interactions – each worker now being far more inclined to show disapproval at a co-worker's slackness. The prevalence of mutual monitoring in worker-run firms is associated with concrete cost-saving from using fewer hired supervisors. Craig and Pencavel (1995) found total factor productivity to be between 6 and 14 per cent higher in worker-owned than in conventional plywood firms. Weitzman and Kruse (1990) found a positive effect of profit-sharing on productivity in a meta-analysis of studies of both worker-owned and conventional firms linking pay to profit. A similar finding is recorded by Doucouliagos (1995) in a meta-analysis of studies focusing on the effect of worker participation in decision-making.

If worker-run firms don't actually suffer from dysfunctional responses to changes in their economic environments, if they aren't dissuaded from investing by horizon problems, and if they motivate work effort at least as effectively as do conventional firms, why aren't they as common as Mill predicted they would one day be? Among the answers that have been proposed is that control by investors is superior to control by workers because investors’ representatives can reach decisions more easily, the idea being that investors share a uniform objective of maximizing the firm's market value, whereas workers have multiple interests (job security, pleasant working conditions, higher earnings) upon which each may place a different weight, thus defying easy consensus (Hansmann, 1990). Another answer, suggested by Kremer (1997), is that less productive workers tend to use the firm's internal decision process to obtain a flatter wage dispersion, which weakens incentives for the more productive workers to stay with the firm. Still another possibility, formalized by Ben-Ner (1984) and Miyazaki (1984) based on an earlier suggestion by Mikhail Tugan-Baranovsky (1921), is that successful LMFs have an incentive to replace retiring members with non-member hired workers, concentrating the profits in the hands of a smaller member group which, in the limit, collapses to contain only one member, a proprietor. Studies of the life cycle of cooperatives, from creation to dissolution, find few cases following precisely this scenario, but situations in which workers sell their firm to private owners and become their employees are reported, for example, in the U.S. plywood sector.

Possibly the most promising place to search for explanations is in the area of financing. Because inputs are committed before output value is certain, and because time passes between the utilization of input

services and the realization of revenue from product sales, firms typically need the services of both risk-bearers and financiers. There is no technical reason why all input suppliers, including workers, could not share in providing these services by accepting payments in the future and by working for shares of an uncertain total revenue, rather than for fixed wages. What is observed, however, is consistent with the view that the supply of risk-bearing and financing services follows comparative advantage: specialists with greater willingness to bear risk and/or ability to pay for inputs up front become the suppliers of equity and debt finance, while workers are paid within short intervals in amounts promised in advance and not contingent on the firm's results. The fact that workers typically have less wealth and thus both less ability to supply funds or to finance their consumption from savings, as well as less willingness to bear risk, is likely to play an important part in explaining this (Putterman, 1993). The thinness of potential markets for worker partnership shares and thus the absence or imperfection of the partnership market may add to the burden that financing their own firm imposes on workers (Dow, 2003).

Although workers do accumulate substantial assets in pension funds in the United States, risk aversion (and pension fund regulations) may deter them from investing too much of it in their own company or in any other single project. In a world in which wealth was quite equally distributed and was held mainly by workers, workers as principal owners of their own firms might still remain rare because workers might prefer to hold diversified portfolios containing shares of many firms other than their own.

If control (by managers) and ownership (by shareholders) are in any case separated in modern corporations, why not worker control with (outside) shareholder ownership? The fact that the de-linking of ownership and control remains incomplete even in those firms where ownership is most diffuse (in other words, the fact that shareholders retain ultimate control rights in publicly traded corporations) suggests an answer. Presumably ownership and control are almost universally linked in a market economy because the owner, the return on whose investment is subject to so many uncertainties, is unwilling to cede control over key decisions affecting that return. Until worker desires for control of their enterprises are strong enough that they are willing to bear considerable financial risk, or until market outcomes are altered by government interventions facilitating the de-linking of control rights from financial risk-bearing, LMFs appear likely to remain the exception to the rule in market economies.

See Also

- Domar, Evsey David
- Meade, James Edward
- Mill, John Stuart
- socialism
- worker participation and profit sharing

Bibliography

Ben-Ner, A. 1984. On the stability of the cooperative type of organization. *Journal of Comparative Economics* 8, 247–60.

Bonin, J. and Putterman, L. 1987. *Economics of Cooperation and the Labor-Managed Economy*. London: Harwood.

Bonin, J., Jones, D. and Putterman, L. 1993. Theoretical and empirical studies of producer cooperatives: will ever the twain meet? *Journal of Economic Literature* 31, 1290–320.

Craig, B. and Pencavel, J. 1992. The behavior of worker cooperatives: the plywood companies of the Pacific Northwest. *American Economic Review* 82, 1083–105.

Craig, B. and Pencavel, J. 1995. Participation and productivity: a comparison of worker cooperatives and conventional firms in the plywood industry. *Brookings Papers on Economic Activity – Microeconomics* 1995, 121–60.

Dahl, R. 1985. *A Preface to Economic Democracy*. Berkeley: University of California Press.

Domar, E. 1966. The Soviet collective farm as a producers' cooperative. *American Economic Review* 56, 734–57.

Doucouliagos, C. 1995. Worker participation and productivity in labor-managed firms and participatory capitalist firms: a meta-analysis. *Industrial and Labor* Relations *Review* 49, 58–77.

Dow, G. 1986. Control rights, competitive markets, and the labor management debate. *Journal of Comparative Economics* 10, 48–61.

Dow, G. 2003. *Governing the Firm: Workers' Control in Theory and Practice*. New York: Cambridge University Press.

Drèze, J. 1976. Some theory of labor management and participation. *Econometrica* 44, 1125–39.

Drèze, J. 1989. *Labor Management, Contracts and Capital Markets: A General Equilibrium Approach*. Oxford: Blackwell.

Fehr, E. 1993. The simple analytics of a membership market in a labor-managed economy. In *Markets and Democracy: Participation, Accountability, and Efficiency*, ed. S. Bowles, H. Gintis and B. Gustafsson. Cambridge: Cambridge University Press.

Furubotn, E. and Pejovich, S. 1970. Property rights and the behavior of the firm in a socialist state: the example of Yugoslavia. *Zeitschrift fur Nationalökonomie* 30, 431–54.

Hansmann, H. 1990. The viability of worker-ownership: an economic perspective on the political structure of the firm. In *The Firm as a Nexus of Treaties*, ed. M. Aoki, B. Gustaffson and O. Williamson.

London: Sage.

Kremer, M. 1997. Why are worker cooperatives so rare? Working Paper No. 6118. Cambridge, MA: NBER.

Meade, J. 1972. The theory of labor-managed firms and of profit-sharing. *Economic Journal* 82, 401–28.

Meade, J. 1989. *Agathotopia: The Economics of Partnership*. Aberdeen: Aberdeen University Press.

Mill, J.S. 1848. *Principles of Political Economy*. London: Longmans, Green and Co., 1936.

Miyazaki, H. 1984. On success and dissolution of the labor-managed firm in the capitalist environment. *Journal of Political Economy* 92, 909–31.

Pateman, C. 1970. *Participation and Democratic Theory*. Cambridge: Cambridge University Press.

Pejovich, S. 1969. The firm, monetary policy and property rights in a planned economy. *Western Economic Journal* 7, 193–200.

Pencavel, J. 2001. *Worker Participation: Lessons from the Worker Co-ops of the Pacific Northwest*. New York: Russell Sage Foundation.

Putterman, L. 1993. Ownership and the nature of the firm. *Journal of Comparative Economics* 17, 243–63.

Sertel, M. 1982. *Workers and Incentives*. Amsterdam: North-Holland.

Tugan-Baranovsky, M. 1921. *Sotsialnyia Osnovy Kooperatsii*. Berlin: Slowo Verlagsgesellschaft.

Vanek, J. 1970. *The General Theory of Labor-Managed Market Economies*. Ithaca: Cornell University Press.

Vanek, J. 1977. The basic theory of financing of participatory firms. In *The Labor-Managed Economy: Essays by Jaroslav Vanek*, ed. J. Vanek. Ithaca: Cornell University Press.

Ward, B. 1958. The firm in Illyria: market syndicalism. *American Economic Review* 68, 566–89.

Weitzman, M. and Kruse, D. 1990. Profit-sharing and productivity. In *Paying for Productivity: A Look at the Evidence*, ed. A. Blinder. Washington, DC: Brookings Institution.

How to cite this article

Putterman, Louis. "labour-managed firms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000014> doi:10.1057/9780230226203.0915

The New Palgrave Dictionary of Economics Online

labour's share of income

Douglas Gollin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists have long studied labour's share of national income as a crude indicator of income distribution. More recently, labour's share has also been seen as offering insights into the shape of the aggregate production function. This has made labour's share a parameter of interest for macroeconomics, growth economics, and international economics, among other fields. Recent studies support the long-standing observation that labour's share of national income is relatively constant over time and across countries. Measurement of labour income, however, can be difficult in economies where many people are self-employed or work in family enterprises.

Keywords

aggregation; balanced growth; Cobb–Douglas functions; constant-returns production function; entrepreneurial income; factor shares; labour's share of income; national income accounting

Article

At least since the time of Adam Smith, economists have been interested in the shares of production accruing to the owners of different factors. In the era before formalized national income and product accounts, factor shares were observed primarily at the firm or industry level. But Smith himself recognized that national product could similarly be divided into the income received by owners of land, labour and capital (the last of which he termed ‘stock’). Early in Book I of *The Wealth of Nations*, Smith (1776, p. 155) notes that

> the exchangeable value … of all the commodities which compose the whole annual produce of the labour of every country, taken complexly, must resolve itself into … three parts and be parcelled out among different inhabitants of the country, either as the wages of their labour, the profits of their stock, or the rent of their land … Wages, profit, and

rent, are the three original sources of all revenue as well as of all exchangeable value.

Smith and other early economists viewed the distribution of income among factors of production as intimately related to the level of wages and the degree of income inequality within a country. This was probably a reasonable assumption, given that, outside of agriculture and certain types of self-employment, most individuals probably subsisted entirely on wage income.

Factor shares were, in fact, one of the few available sources of data on the size distribution of income – a subject that was viewed as crucial for policymaking, but about which little was known. As late as 1912, a prominent US labour economist wrote (Streightoff, 1912, p. 155), ‘Knowledge of the distribution of incomes is vital to sane legislative direction of progress. In a form definite enough for practical use, this knowledge does not exist. No time should be wasted in obtaining this knowledge.’

Labour's share of national income was seen as a particularly sensitive issue – intimately related to the supposed struggle of labour against capital. Simon Kuznets (1933, p. 30) referred to ‘[t]he significant political and social conflicts that center about the relative share of these productive factors’. Because of the importance of the topic, and because factor shares could be estimated reasonably well from micro data, a considerable literature emerged to document cross-section and time series observations on factor shares. In fact, the literature on factor shares eventually served as one of the foundations for the emergence of national income and product accounts.

From the beginning, the measurement of factor shares has been complicated by the difficulty of disentangling individual incomes into their functional components. Certain categories of income are easily assigned to land, labour, or capital. For example, wages and salaries are generally classifiable as labour income – although for some high-skill workers (such as hedge fund managers, star athletes), they may also embody some rents. Dividends and interest must be forms of capital income. Land rents are easily classified. But Kuznets (1933) pointed out that entrepreneurial income – which was about one fourth of national income in the 1920s – represented a mix of wages, salaries, interest, rent, and profits. As national income accounting evolved over the succeeding decades, there were few improvements to the categorization of income according to factors of production. Irving Kravis (1962, p. 122) noted that ‘the theory of distribution remains in a parlous state’, largely because ‘the components of income for which we have data has not been determined by the requirements of the economists but by the legal and institutional arrangements of our society’.

Nevertheless, by the 1950s a striking empirical regularity had begun to emerge. Labour's share of national income in the United States appeared to have remained roughly constant over a long period of time. Modest increases in the share of wages and salaries in national income appeared to have come at the expense of declines in entrepreneurial income – consistent with a structural shift away from self-employment and towards wage work. The regularity was sufficiently pronounced that Charles Cobb and Paul Douglas, writing in 1928, suggested that a simple constant-returns production function in the now familiar form $Y=AK^{1/4}L^{3/4}$ would provide an accurate representation of the US time series for aggregate output as a function of aggregate capital stock and labour. They considered a value for labour's share as low as two-thirds to be plausible.

As national income accounting became more systematic, evidence on factor shares accumulated over succeeding decades. John Maynard Keynes, writing in 1939 (p. 48), referred to the ‘stability of the proportion of the national dividend accruing to labour, irrespective apparently of the level of output as a

whole and of the phase of the trade cycle'. He went on to refer to this (p. 48) as 'one of the most surprising, yet best-established facts in the whole range of economic statistics, both for Great Britain and for the United States'.

D. Gale Johnson (1954) constructed and analysed data for the US economy going back over a century, to 1850, and concluded (p. 175) that there had been no 'significant secular change' in labour's share of income over that period. Robert Solow's paper (1957) on the sources of growth in the US economy noted that the data for the US economy seemed consistent with a Cobb–Douglas representation for the aggregate production function, with a capital share of 0.35 (and thus, implicitly, a labour share of 0.65). (However, Solow, 1958, professed scepticism over the proposition that factor shares were actually constant, suggesting instead that variation within sectors was balanced out at the aggregate level.) Nicholas Kaldor (1961) characterized the phenomenon as one of the stylized facts of modern economic growth.

This apparent consensus soon began to unravel, however. A major challenge to the hypothesis of constant factor shares appeared in comparisons of factor shares across countries. Kuznets, in an influential 1959 paper, further argued that the cross-country evidence did not support the view that factor shares were constant across countries or over time. Kuznets argued that data for other countries – and in particular for poor countries – revealed very different levels for labour's share in other countries. In particular, Kuznets suggested that labour's share of income was systematically lower in poor countries than in rich countries, while the share of unincorporated enterprises in national income was higher in poor countries than in rich countries. Kuznets concluded that the concept of a labour share lacked useful meaning – particularly as a proxy for discussions of the size distribution of income. His scepticism over constant factor shares was echoed by Solow (1958) and by Kravis (1962), among others.

To a large degree, scholarly interest in the labour share waned in succeeding years, although quantitative studies in both international trade and growth continued to rely on Cobb–Douglas aggregate production functions. In the trade literature, it was commonplace to assume that rich countries had a relatively high labour share, while poor countries had lower shares. Macro and growth studies of advanced economies typically assumed a Cobb–Douglas production function with a labour share of about two-thirds, often based on the employee compensation share of GNP for the United States, but this parametrization was seen as problematic for models that were intended to characterize both poor countries and rich ones. This apparent discrepancy between cross-country and time series observations on labour's share was largely unaddressed in the literature until Gollin (2002) revisited the question. Drawing on the earlier work of Kuznets and others, he noted the potential significance of self-employment in skewing 'naive' calculations of factor shares. Gollin argued that poor countries typically have far higher levels of self-employment than do rich countries; as a result, cross-country comparisons of the employee compensation share (or wage share) will tend to yield large differences between rich and poor countries. Gollin showed that, after adjusting labour's share to account for differences in self-employment rates, no systematic patterns remained in the cross-country data between a country's income and its imputed labour share. Gollin reported labour shares in most countries, adjusted for self-employment, between 0.6 and 0.8. Similar results were obtained by Ben Bernanke and Refet Gürkaynak (2002), who used a different approach to adjust for the fraction of output produced by unincorporated enterprises.

Recent and preliminary work by Rodrigo García-Verdú (2005) for Mexico found that labour's share falls into this range when estimated from household survey data, rather than from national income accounts

might suggest. However, Daniel Ortega and Francisco Rodríguez (2006) present evidence from industrial census data that labour shares are lower in poor countries than in rich countries. And Samuel Bentolila and Gilles Saint-Paul (2003) show that labour's share within OECD countries is not constant, but rather moves in parallel with changes in the capital–output ratio.

Econometric studies of aggregate production functions, such as those by John Duffy and Chris Papageorgiou (2000) and Pol Antràs (2004), often reject the Cobb–Douglas specification of the aggregate production function. This suggests that, if factor shares are indeed (approximately) constant, there must be a different underlying mechanism. At the simplest level, any constant returns production function with labour-augmenting technical progress can give rise to constant factor shares if the rate of return on capital is constant over time – as, for example, on a balanced growth path. To see this, consider a simple Solow model with the constant returns aggregate production function $Y = F(K, AL)$. The productivity parameter A grows at a constant rate *g*, and there is an exogenous savings rate, *s*. This economy will converge to a balanced growth path; assuming no population growth, the condition for balanced growth is given by

$$k^* = \frac{sf(k^*)}{\delta + g},$$

where δ is the depreciation rate and

$$k \equiv \frac{K}{AL}.$$

But the balanced growth path implies that the capital share is

$$\frac{rk^*}{f(k^*)} = \frac{sr}{\delta + g},$$

which will necessarily be constant because the rate of return is constant along the balanced growth path.

An alternative way to generate constant factor shares is through aggregation. Charles I. Jones (2005) reproduces and generalizes a result of Houthakker (1955) in which an aggregate Cobb–Douglas technology can be derived from firm-level or industry-level Leontief techniques. Jones shows that the same intuition can be applied more generally to a world in which the underlying production technologies have almost any form, and the ‘aggregation’ can simply occur across ideas or techniques within a firm.

Jones's result is consistent with factor shares that are constant, but it also allows for movement in the factor shares and for differences across countries. In general, it appears to offer a useful theoretical framework for reconciling the different features of the data.

See Also

- Cobb–Douglas functions
- economic growth
- economic growth, empirical regularities in
- factor prices in general equilibrium
- growth accounting
- level accounting
- production functions

Bibliography

Antràs, P. 2004. Is the US aggregate production function Cobb-Douglas? New estimates of the elasticity of substitution. *Contributions to Macroeconomics* 4(1), Article 4.

Bentolila, S. and Saint-Paul, G. 2003. Explaining movements in the labor share. *Contributions to Macroeconomics* 3(1), Article 9.

Bernanke, B.S. and Gürkaynak, R.S. 2002. Is growth exogenous? Taking Mankiw, Romer and Weil seriously. In *NBER Macroeconomics Annual 2001*, ed. B.S. Bernanke and K.S. Rogoff. Cambridge, MA: MIT Press.

Cobb, C.W. and Douglas, P.H. 1928. A theory of production. *American Economic Review* 18, 139–65.

Duffy, J. and Papageorgiou, C. 2000. A cross-country empirical investigation of the aggregate production function specification. *Journal of Economic Growth* 5, 87–120.

García-Verdú, R. 2005. Factor shares from household survey data. Working Paper No. 2005-05, Dirección General de Investigación Económica, Banco de México.

Gollin, D. 2002. Getting income shares right. *Journal of Political Economy* 110, 458–74.

Houthakker, H.S. 1955. The Pareto distribution and the Cobb–Douglas production function in activity analysis. *Review of Economic Studies* 23, 27–31.

Johnson, D.G. 1954. The functional distribution of income in the United States, 1850–1952. *Review of*

Economics and Statistics 36, 175–82.

Jones, C.I. 2005. The shape of production functions and the direction of technical change. *Quarterly Journal of Economics* 120, 517–49.

Kaldor, N. 1961. Capital accumulation and economic growth. In *The Theory of Capital*, ed. F. Lutz. London: Macmillan, 1961.

Keynes, J.M. 1939. Relative movements of real wages and output. *Economic Journal* 49(193), 34–51.

Kravis, I.B. 1962. *The Structure of Income: Some Quantitative Essays*. Philadelphia: University of Pennsylvania Press.

Kuznets, S. 1933. National income. In *Encyclopedia of the Social Sciences*, vol. 11. New York: Macmillan. Repr. in *Readings in the Theory of Income Distribution*, selected by a committee of the American Economic Association. Philadelphia: Blakiston, 1946.

Kuznets, S. 1959. Quantitative aspects of the economic growth of nations IV: distribution of national income by factor shares. *Economic Development and Cultural Change* 7(3,Part II), 1–100.

Ortega, D. and Rodríguez, F. 2006. Are capital shares higher in poor countries? Evidence from industrial surveys. Wesleyan Economics Working Papers No. 2006-023, Wesleyan University.

Smith, A. 1776. *The Wealth of Nations: Books I–III*. Harmondsworth: Penguin Books, 1986.

Solow, R.M. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

Solow, R.M. 1958. A skeptical note on the constancy of relative shares. *American Economic Review* 48, 618–31.

Streightoff, F.H. 1912. *The Distribution of Incomes in the United States*. New York: Columbia University Press, 1912.

How to cite this article

Gollin, Douglas. "labour's share of income." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000228> doi:10.1057/9780230226203.0921

The New Palgrave Dictionary of Economics Online

Laffer curve

Don Fullerton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A Laffer curve is a hump-shaped curve showing tax revenue as a function of the tax rate. Revenue initially increases with the tax rate but then can decrease if taxpayers reduce market labour supply and investments, switch compensation into non-taxable forms, and engage in tax evasion. The revenue-maximizing tax rate can be calculated from an estimate of the elasticity of taxable income with respect to the after-tax share. Some studies find this elasticity to be near zero, and others find it to exceed 1. The mid-range for this elasticity is around 0.4, with a revenue peak around 70 per cent.

Keywords

capital supply; elasticity of labour supply; elasticity of taxable income; excess burden of taxation; home production; income effect; labour supply; Laffer curve; leisure; marginal and average tax rates; progressive and regressive taxation; revenue maximization; substitution effect; supply side economics; tax avoidance; tax compliance; tax evasion; tax revenue; taxation of corporate profits; taxation of income

Article

On a napkin in a Washington restaurant in 1974, Arthur Laffer famously drew his hump-shaped curve showing tax revenue as a function of the tax rate (see Figure 1). Revenue is zero both when the tax rate is zero and when the tax rate is 100 per cent or more. In between must be some t^* that maximizes revenue. The point is that taxes discourage supply of labour, especially by secondary workers in the family who have elastic behaviour, and they discourage supply of capital over time. Thus, proponents became known as ‘supply siders’. So far, these points were well accepted, as economists are quite familiar with the idea of supply as well as demand. Even as far back as 1776, Adam Smith understood that ‘High taxes, sometimes by diminishing the consumption of the taxed commodities, and sometimes by encouraging smuggling, frequently afford a smaller revenue to government than what might be drawn from more moderate taxes’ (Smith, 1776, V, II).

Figure 1
The Laffer curve

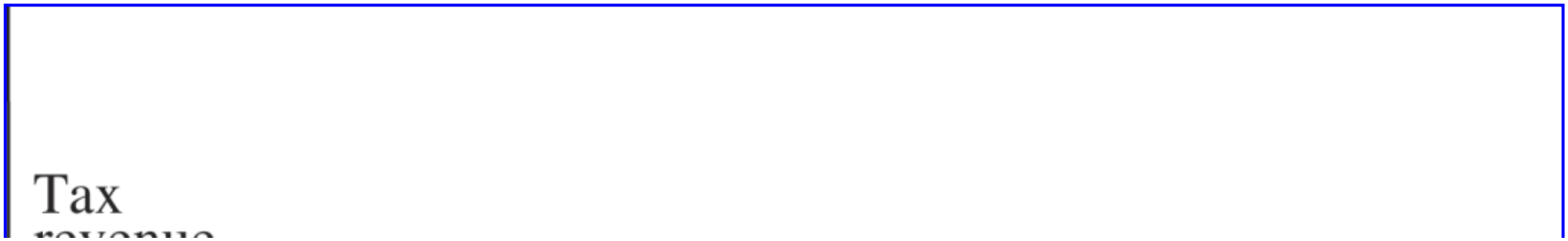

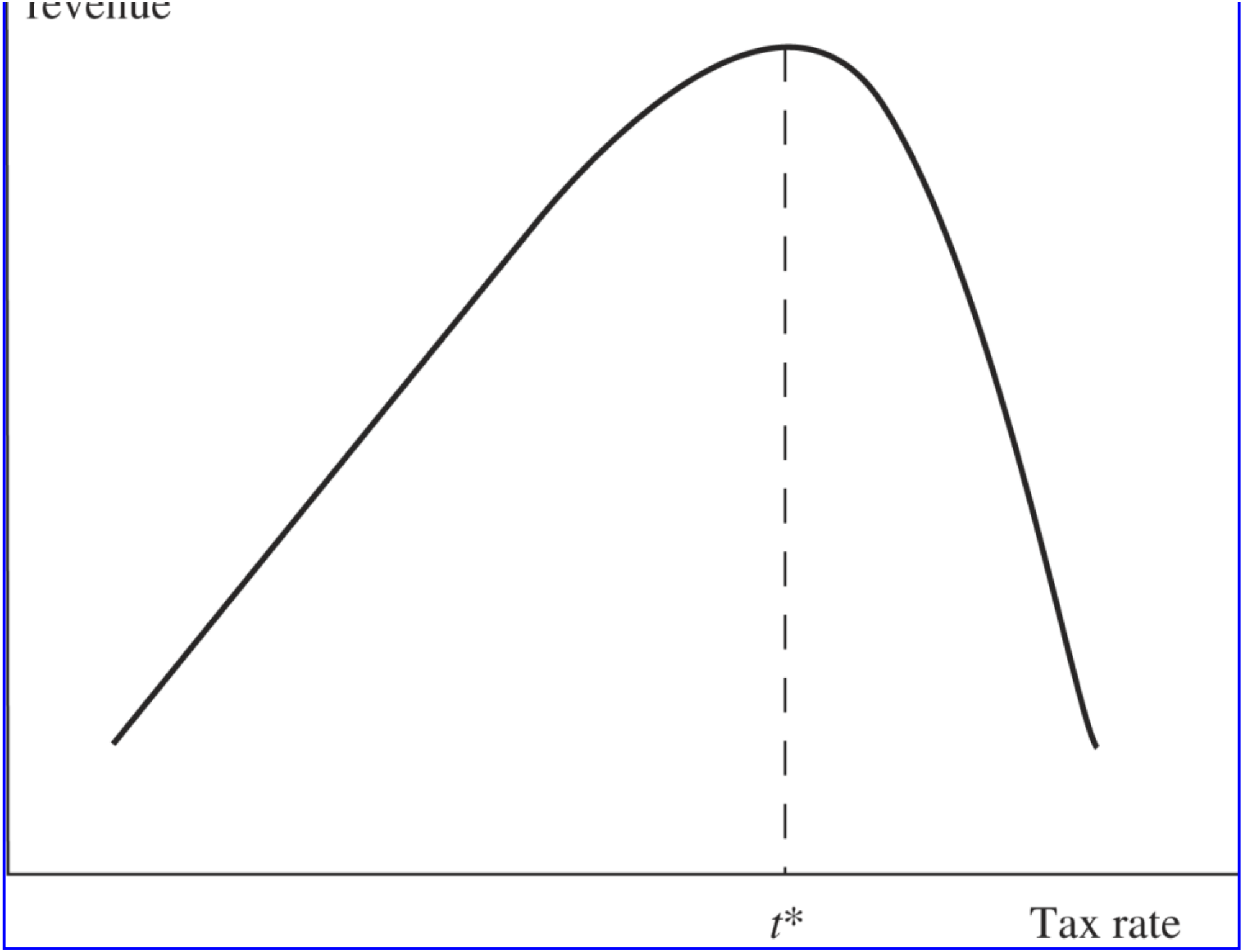
revenue
t*
Tax rate

The more controversial claim was that the US tax rate was greater than t^*, on the 'prohibitive range' where no rational government would knowingly operate, meaning that a reduction in tax would actually increase government revenue.

Initial research focused on static models of labour supply. Stuart (1981) builds a simple analytical model with a taxed sector and an untaxed sector, and he chooses parameters to represent Sweden. The untaxed sector includes illicit tax evasion as well as leisure and home production such as painting your own home, growing your own vegetables, cooking your own meals, and cleaning your own house. He finds a peak at 70 per cent, which is fairly high, but then he also finds that Sweden has an overall effective marginal tax rate of 80 per cent! Then Fullerton (1982) describes two models. First, in a simple partial equilibrium model where η is the labour demand elasticity and ε is the labour supply elasticity, it is easy to show that $t^* = (\eta - \varepsilon) / [\eta(1 + \varepsilon)]$. If the labour demand curve is flat $(\eta = -\infty)$, to focus on supply, then $t^* = 1 / (1 + \varepsilon)$. Thus, higher ε implies lower t^*. The second model is a multi-sector computable general equilibrium model of the United States, but one that still requires an overall labour supply elasticity (ε). Based on estimates that are zero or negative for men and positive for women, the choice of $\varepsilon = 0.15$ in this model yields $t^* = 79$ per cent.

This research faces a number of problems. First, we do not really know the labour supply elasticity, and heterogeneity means we have no such thing as 'the' elasticity anyway. Second, we do not know the current tax rate either, since actual tax systems are complicated combinations of income, payroll, and sales taxes. For example, the payroll tax does not apply for workers whose tax payments are offset at the margin by additional expected social security benefits, and it also does not apply for those above the cap. Third, the income tax is progressive, which means different rates for different individuals. All this heterogeneity means no such thing as 'the' tax rate.

Fourth, even if we ignore heterogeneity, a progressive system means that the marginal tax rate (which affects incentives) exceeds the average tax rate (which affects revenue). Then the question of how a change in marginal tax rate affects revenue is not well defined, because one must also specify how the reform affects average rates. Even if an increase in all marginal rates raised revenue, for example, an increase in only the top marginal rate may not. Also, if a change in progressivity transfers money between groups, then the outcome depends on different *income* elasticities of labour supply. A reduction of the top marginal tax rate may seem to have the best potential for a Laffer effect if both (*a*) the rate is high and (*b*) those workers are elastic. But if part of the increased revenue comes from redistribution between taxpayers with different elasticities, then it is not a true Laffer effect.

Fifth, the Laffer curve itself is not well defined, with revenue on the vertical axis, because it matters how that revenue is spent. Interestingly, Malcomson (1986) shows that the Laffer curve may continue to slope upwards, all the way to a tax rate of 100 per cent, which would mean no prohibitive range at all! Yet Gahvari (1989) shows how this result depends on the assumption that revenue is used to provide a public good that is separable in utility. Then the tax hike has an income effect that increases work effort, and revenue may continue to rise. If the increased revenue is used for lump-sum transfers, however, then this cash tends to offset the income effect, leaving only the substitution effect that is so emphasized by the supply siders in the first place.

So far, these models are static models of labour supply. Agell and Persson (2001) build a one-sector endogenous growth model with capital as the only input, and no labour at all, yet they obtain a strikingly similar result. They allow for separable government spending *G* or cash transfers *T*. One of their alternative definitions of a 'dynamic Laffer effect' is when government can reduce a tax rate and still increase at least one future year's *G* or *T* . They then show that a world with no transfers can never have a dynamic Laffer effect. The revenue-maximizing tax rate is 100 per cent, confiscating capital (so the growth rate is negative). With sizable transfers that are set to grow at some fixed rate, however, then a tax cut that increases the economy's growth rate means that transfers shrink as a fraction of GDP. Then, that negative wealth effect makes people save more, which increases the future tax base and may yield a dynamic Laffer effect.

The initial emphasis of the supply siders themselves was on supply of labour and capital, since these responses to a tax cut can increase income, growth, the tax base, and government revenue. Indeed, estimates of the labour supply elasticity mentioned above are estimates of the *hours'* elasticity, the effect of the tax cut on hours worked. Yet what matters for tax revenue is the effect of the tax cut on 'taxable income'. Feldstein (1995) points out that a 'change in individuals' marginal income tax rates can induce them to alter their taxable income in a wide variety of ways, including changes in labour supply, in the form in which employee compensation is taken, in portfolio investments, in itemized deductions and other expenditures that reduce taxable income, and in taxpayer compliance' (1995, pp. 552–3). Thus begins a large empirical literature trying to estimate *e*, defined as the elasticity of taxable income with respect to a change in the marginal net-of-tax share $(1 - t)$. If the economy really had only a single tax rate *t*, then the revenue-maximizing tax rate is $t^* = 1 / (1 + e)$.

Most of this literature takes a natural experiment approach that looks at years before and after a change in the income tax rate schedule, while comparing the top-bracket income group to the next-bracket income group. On the assumption that all other time trends affect the two groups similarly, then their *e* can be calculated by taking the difference between the two groups' change in reported taxable incomes compared with the difference between their changes in after-tax shares. Lindsey (1987) begins this literature by using cross-section data from the early 1980s for various income groups. The Economic Recovery Tax Act of 1981 reduced the top rate most, and the top bracket's reported taxable income increased the most. The implied elasticity is around 1.5, so the implied revenue-maximizing overall tax rate is around $t^* = 1 / (1 + e) = 40$ per cent. This result stands in stark contrast to estimates mentioned above where t^* was 70–80 per cent.

This type of research also faces a number of problems. First, income inequality was trending upwards during these years, which might mean rising incomes at the top, relative to other groups, irrespective of the tax change. Second, random shocks to income mean that the top bracket may not contain the same individuals across years. Feldstein (1995) deals with this

problem by use of panel data, tracking the same individuals before and after the top bracket rate cut of the Tax Reform Act of 1986. He also finds taxable income elasticities in excess of 1.0 (and sometimes 2.0 or 3.0).

Third, any given tax reform usually involves changes in the definition of taxable income, and not just changes in rates. Thus, these studies try to adjust their measure of income to use the same definition across years. Fourth, any change in the top personal income tax rate relative to the corporate tax rate might induce shifting: a change in personal taxable income that is offset by an opposite change in corporate taxable income. Fifth, the increase in taxable income in a single year after the tax change may be temporary rather than permanent. Sixth, the first few papers in this literature looked only at tax rate cuts in the 1980s, where other periods may have tax rate increases. Finally, each tax rate reform may involve a different set of income tax rules that determine the ease of tax avoidance. In other words, there is no such thing as ‘the’ taxable income elasticity.

To deal with several of these problems, Goolsbee (1999) applies the natural experiment approach to *six* different tax reforms from 1920 to 1975, including both tax rate cuts and increases, and including periods with different trends in income inequality. He finds that the 1980s are atypical: ‘the largest regression estimates of the taxable income elasticity from all of the previous historical periods are lower than the smallest estimates in the literature based on the 1980s’ (1999, p. 43). Other studies find *e* around zero, as reviewed by Gruber and Saez (2002). They use a 1979–90 panel of tax returns to analyse *all* state and federal tax reforms during the 1980s, and they ‘find that the overall elasticity of taxable income is 0.4, well below the original estimates of Feldstein but roughly at the mid-point of the subsequent literature’ (2002, p. 3).

Finally, Kopczuk (2005) adds a measure of the tax base, relative to total income for each individual, and finds that it affects the estimate of the taxable income elasticity. In other words, that elasticity is not just a taxpayer's behavioural parameter, but depends on the tax code. The rich have a narrower tax base, and thus a higher elasticity. This also means that reforms to broaden the base can raise t^* itself (and reduce excess burden).

In summary, if you choose to oversimplify the world by using a single elasticity and a single tax rate, and if you ignore other problems above with the whole concept of the Laffer curve, then the recent mid-point estimate of $e = 0.4$ implies that tax revenue is maximized at $t^* = 1/(1+e) = 71$ per cent.

See Also

- labour supply
- tax compliance and tax evasion
- taxation of income

Bibliography

Agell, J. and Persson, M. 2001. On the analytics of the dynamic Laffer curve. *Journal of Monetary Economics* 48, 397–414.

Feldstein, M. 1995. The effect of marginal tax rates on taxable income: a panel study of the 1986 Tax Reform Act. *Journal of Political Economy* 103, 551–72.

Fullerton, D. 1982. On the possibility of an inverse relationship between tax rates and government revenues. *Journal of Public Economics* 19, 3–22.

Gahvari, F. 1989. The nature of government expenditures and the shape of the Laffer curve. *Journal of Public Economics* 40, 251–60.

Goolsbee, A. 1999. Evidence on the high-income Laffer curve from six decades of tax reform. *Brookings Papers on Economic Activity* 1999(2), 1–64.

Gruber, J. and Saez, E. 2002. The elasticity of taxable income: Evidence and implications. *Journal of Public Economics* 84, 1–32.

Kopczuk, W. 2005. Tax bases, tax rates and the elasticity of reported income. *Journal of Public Economics* 89, 2093–119.

Lindsey, L. 1987. Individual taxpayer response to tax cuts, 1982–1984: with implications for the revenue maximizing tax rate. *Journal of Public Economics* 33, 173–206.

Malcomson, J.M. 1986. Some analytics of the Laffer curve. *Journal of Public Economics* 29, 263–79.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: J.M. Dent & Sons, 1975.

Stuart, C.E. 1981. Swedish tax rates, labor supply, and tax revenues. *Journal of Political Economy* 89, 1020–38.

How to cite this article

Fullerton, Don. "Laffer curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000015> doi:10.1057/9780230226203.0922

The New Palgrave Dictionary of Economics Online

Laffont, Jean-Jacques (1947 – 2004)

Jean Tirole
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Jean-Jacques Laffont was one of the great economists of the last quarter of the 20th century, with an encyclopedic mind in a time of intense specialization. He won widespread respect and recognition for his breakthroughs in both theory (including public goods, contract theory, and the regulation of natural monopoly) and econometrics. In addition, he was energetically engaged in institution-building not only in Europe but also in Africa, Asia and Latin America.

Keywords

access prices; adverse selection; asymmetric information; auctions; Bayesian implementation; budget balance; Coase Theorem; collective choice; collusion; contract theory; disequilibrium macroeconomic models; Econometric Society; econometrics; engineering cost models; European Economic Association; industrial organization; instrumental variables; Laffont, J.-J.; less developed countries; nonlinear simultaneous equations; occupational choice; partial equilibrium; productivity; public goods; regulation of natural monopolies; regulatory capture; revealed preferences; risk aversion; theory of entrepreneurs

Article

Jean-Jacques Laffont was born in 1947 and died in 2004 in Toulouse. He was one of the great economists of the last quarter of the 20th century. He made breakthroughs in many fields within both theory and econometrics, which made him perhaps the last encyclopedic mind in the economics profession at a time when the rapid growth of knowledge pushes most researchers into intense specialization. His creative and prolific contributions brought him widespread respect and recognition, from presidencies of learned societies (Econometric Society, European Economic Association) to numerous prizes (including the Yrö-Jahnsson prize), honorary memberships in foreign learned societies, honorary degrees from several universities and invitations to give numerous prestigious lectures. Besides his academic contributions – the topic of this contribution – Jean-Jacques Laffont will also be long

remembered for his selfless contributions to institution building in Europe and in particular Toulouse, where his warmth, devotion and energy allowed him, starting nearly from scratch, to create an enthusiastic and congenial research environment. In Africa, Asia and Latin America also, he encouraged young economists to work with him on frontier economics and helped build research centres.

Public goods

After completing his Ph.D. at Harvard University in 1974, Jean-Jacques Laffont embarked on a celebrated research agenda on public goods, in collaboration with Jerry Green (culminating in their 1979 book) and later with Eric Maskin. A collective decision-making problem with *n* economic agents $(i = 1, \ldots, n)$ who have quasi-linear preferences of the form:

$$u_i = v_i(a, \theta_i) + t_i$$

consists in selecting a policy *a* and transfers t_i for each configuration of taste parameters $\theta = (\theta_1, \ldots, \theta_n)$. An efficient policy $a^*(\theta)$ solves

$$\max_{\{a\}} \sum_{i=1}^{n} v_i(a, \theta_i) .$$

A central issue is how to implement this efficient action through appropriate transfers when agents privately know their own taste parameters. Clarke (1971), Groves (1973) and Vickrey (1961) (CGV) had defined ‘mechanisms’, in which agents announce ‘types’ $\hat{\theta}_i$, the collective decision is $a^*(\hat{\theta})$ and agent *i* receives a transfer of the form

$$t_i(\hat{\theta}) = \sum_{j \neq i} v_j(a^*(\hat{\theta}), \hat{\theta}_j) .$$

They had shown that such schemes would induce each agent to truthfully reveal her preferences $\hat{\theta}_i = \theta_i$, as she internalizes the consequences of her choices on the welfare of others. Green and Laffont (1977) showed that these mechanisms were, up to the addition of a function $t_i^o(\hat{\theta}_{-i})$ which is independent of the announcement of the others, the only schemes in which truthful revelation is a dominant strategy.

Laffont and Maskin (1980), pioneering the 'differentiable approach' to mechanism design, then showed that the transfers t_i^o were but constants of integration when the v_i are differentiable in *a* and θ_i.
A consequence of Green and Laffont's characterization was that dominant strategy public good schemes are inconsistent with budget balance ($\Sigma_i t_i = 0$). This negative result shifted the profession's attention to the weaker requirement of Bayesian implementation, in which truth telling is an agent's best response to the other agents' truth telling. Laffont and Maskin's (1979) pioneering work showed that inefficiency necessarily resulted from the stricter requirement that the budget be balanced for each configuration of preferences; their paper led the way to the equally pioneering paper of Myerson and Satterthwaite (1983) stating the generic inefficiency of bargaining processes under asymmetric information. These two papers thereby identified one important limitation of the Coase theorem.

Contract theory

More generally, during the decade following his Ph.D. Laffont was involved in many of the developments of contract theory, from adverse selection to moral hazard, from single-agent partial-equilibrium to general equilibrium settings. Examples of this work include the definitive treatment of adverse selection with Guesnerie (1984), the first model of occupational choice in which Kihlstrom and Laffont (1979) built a theory of entrepreneurs based on heterogeneity in risk aversion, and the prescient piece with Green (1986) on limited scopes for misreporting (the report $\hat{\theta}_i$ is restricted to belong to a subset of types that depends on the true θ_i), in which they showed how to amend the revelation principle and derived some implications for the magnitude of distortions brought about by private information.

Regulation

A common application of incentive theory is to the regulation of natural monopolies. The first experiments with price caps in the mid-1980s and later with deregulation raised questions about what could be expected from such reforms and about their potential pitfalls. Starting with the 1986 paper on the power of incentive schemes and up to their 1993 book, Laffont and Tirole focused on these issues, modelling the objective of the regulated firm as (variants of)

$$u = t - C(\theta, e, q) - \psi(e),$$

where *t* is the firm's budget, *C* its monetary cost, $\psi(e)$ an increasing and convex non-monetary function of the effort *e*, θ a technology parameter unknown to the regulator and *q* the vector of outputs. While costs and outputs are observable, the firm can transform naturally low costs into shirking (or private benefits). For any abstract regulatory mechanism $\{q(\hat{\theta}), t(\hat{\theta})\}$, expressing, as a function of productivity, the effort needed to reach a given cost level for given outputs and applying the envelope theorem, the

regulated firm's rent's sensitivity to the productivity parameter is given by

$$\frac{du}{d\theta} = \psi'(e)\left|\frac{\partial e}{\partial \theta}\right|.$$

where $\delta e/\delta\theta$ measures the firm's ability to transform productivity gains into private benefits (for example, for a single output q and $C(\theta, e, q) = (C_0 - \theta - e)q$, $|\delta e/\delta\theta| = 1$). This condition provides the intuition for the incentive-rent extraction trade-off: high-powered incentives schemes – that is, schemes for which the firm bears a high share of its cost (inducing a high effort and therefore a high $\psi'(e)$) – necessarily leave large rents (large $u(\theta)$ s) on the table (this is the reason why price caps are often subject to political pressure for renegotiation). The 1986 paper provided sufficient conditions for a menu of linear contracts to be optimal.

Subsequent work focused on how the power of the incentive scheme is affected by concerns for quality, auctioning of incentive contracts, dynamics (the ratchet effect), and regulatory capture. Laffont and Tirole argued that a key enabler of political capture of the regulatory process is the asymmetry of information with the political principal (perhaps Congress, and certainly the citizens), and that the regulatory response to the threat of capture was low-powered incentives, as these reduce rents and therefore make the concerted manipulation of information by the firm and its regulator less attractive to them.

Later, Laffont and Tirole derived theoretical principles for the design of access prices, a key ingredient of the liberalization policy, in the case of one-way access to a bottleneck such as a local loop, an electricity grid or a railroad network (1994) and, in collaboration with Rey (1998a; 1998b), two-way access, that is, access to mutual termination bottlenecks present in telecommunications or the internet.

Jean-Jacques Laffont was adamant about the ability of economic theory to help guide economic development, provided that the theory is properly adapted to reflect the specificities of the developing world. In his posthumous (2005) book, he did just that in the context of regulation. Characterizing less developed countries as countries with easy side transfers within families, ethnic groups and social networks, a lack of a constitutional control of government, a weak rule of law, a high cost of public funds, politically dependent regulators, and weak accounting structures, he systematically drew the implications for the design of regulation, from the power of incentive schemes to universal service obligations and a positive theory of privatization.

More contract theory

Convinced that collusion was a key determinant of economic outcomes and institutions, Jean-Jacques Laffont engaged in a thoughtful and seminal line of research on the methodology and implications of models of collusion, in particular in collaboration with David Martimort. Their 1997 paper developed a general approach for the analysis of collusion among n agents against a principal; an upper bound on the potential damage of collusive activities is obtained by introducing a fictitious coordinator (or cartel ringmaster in an auction) who (a) privately elicits the n agents' types $(\theta_1, \ldots, \theta_n)$, (b) dictates the

agents' behaviours in the game designed by the principal, and (c) breaks even. This 'side mechanism' must be incentive compatible as well as individually rational (the agents must be willing to collude).
In their 2000 paper, Laffont and Martimort point at the dual impact of the 'commonness' of information among agents. A fundamental insight due to Maskin (1999) is that information held by multiple agents can often be elicited at very low cost by having economic agents compete, challenge each other's reports, exercise options, and so on. Maskin's insight has wide-ranging consequences for the use of the informational content of financial and labour markets, auctions, options and other commonly used elicitation mechanisms for the design of contracts and organizations. Laffont and Martimort argue that Maskin's insight is most potent when the schemes have integrity, that is, they are not vulnerable to collusion among agents; for it is precisely when agents have the same information that it is easy for them to collude. Put differently, informational asymmetries among agents hinder collusion. Faure-Grimaud, Laffont and Martimort (2003) show that delegation is an optimal response to collusion.
On the more applied aspects of side-contracting, Laffont and Martimort (1999) showed that the separation of regulators may make capture more difficult. Laffont and Meleu (1997) provided one of the first endogenizations of side transfers, and showed that reciprocal supervision provides an undesirable conduit for collusion.

Econometrics

Quite remarkably, Laffont also made key contributions to theoretical and applied econometrics. As a Harvard student, he collaborated with Jorgenson to produce one of the first methods for estimating nonlinear simultaneous equations, in particular extending and studying the efficiency of minimum distance and instrumental variable estimators, paving the way for Hansen and Hansen and Singleton's 1982 pioneering contributions. Gouriéroux, Laffont and Monfort (1980) is another important illustration of Laffont's contributions to nonlinear econometrics, this time motivated by the identification of simultaneous equation models with latent variables, and in particular disequilibrium macroeconomic models.
Later, Laffont was one of the pioneers of the new empirical industrial economics. He firmly believed in the importance of theory for imposing structural constraints in econometric estimation, and in the continuous back-and-forth interaction between industrial organization theory and empirics. His first research along these lines (with Gasmi and Vuong, 1992) is on the study of tacit collusion in price and advertising in the Coca–Pepsi duopoly. He then found in auctions and their clear extensive form a most favorable ground for structural econometrics in IO. Positing Bayesian equilibrium strategies and adding parametric restrictions allows the researcher to identify the underlying distribution of types and thus the structure of the model. For example, Laffont, Ossard and Vuong (1995) develop a simulated nonlinear least-squares method to estimate auctions with independent private values for a range of first- and second-bid mechanisms and apply it to eggplant auctions in the south-west of France.
Last, Jean-Jacques Laffont's was also interested in the engineering cost models (with Gasmi et al., 2002) as he viewed these as enabling a better regulation of, say, universal service obligations or access prices.

See Also

- agent-based models
- auctions (empirics)
- cartels
- contract theory
- development economics
- public goods

I am grateful to Jacques Crémer, Marc Ivaldi, David Martimort and Eric Maskin for helpful comments.

Selected works

1974. (With D. Jorgenson.) Efficient estimation of nonlinear simultaneous equations with additive disturbances. *Annals of Social and Economic Measurement* 3, 615–640.

1977. (With J. Green.) Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica* 45, 427–38.

1979. (With J. Green.) *Incentives in Public Decision-Making*. Amsterdam: North-Holland.

1979. (With R. Kihlstrom.) A general equilibrium entrepreneurial theory of the firm based on risk aversion. *Journal of Political Economy* 87, 719–48.

1979. (With E. Maskin.) A differentiable approach to expected utility maximizing mechanisms. In *Aggregation and Revelation of Preferences*, ed. J.-J. Laffont. Amsterdam: North-Holland.

1980. (With C. Gouriéroux and A. Monfort.) Disequilibrium econometrics in simultaneous equations systems. *Econometrica* 48, 75–96.

1980. (With E. Maskin.) A differentiable approach to dominant strategy mechanisms. *Econometrica* 48, 1507–20.

1984. (With R. Guesnerie.) A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm. *Journal of Public Economics* 25, 329–69.

1986. (With J. Green.) Partially verifiable information and mechanism design. *Review of Economic Studies* 53, 447–56.

1986. (With J. Tirole.) Using cost observation to regulate firms. *Journal of Political Economy* 94, 614–41.

1992. (With F. Gasmi and Q. Vuong.) Econometric analysis of collusive behavior in a soft drink market.

Journal of Economics and Management Strategy 1, 277–311.

1993. (With J. Tirole.) *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

1994. (With J. Tirole.) Access pricing and competition. *European Economic Review* 38, 1673–710.

1995. (With H. Ossard and Q. Vuong.) Econometrics of first-price auction. *Econometrica* 63, 953–80.

1996. (With Q. Vuong.) Structural analysis of auction data. *American Economic Review* 86, 414–20.

1997. (With D. Martimort). Collusion under asymmetric information. *Econometrica* 65, 875–911.

1997. (With M. Meleu.) Reciprocal supervision, collusion and organizational design. *Scandinavian Journal of Economics* 99, 519–40.

1998a. (With P. Rey and J. Tirole.) Network competition: I. Overview and nondiscriminatory pricing. *RAND Journal of Economics* 29, 1–37.

1998b. (With P. Rey and J. Tirole.) Network competition: II. Price discrimination. *RAND Journal of Economics* 29, 38–56.

1999. (With D. Martimort.) Separation of regulators against collusive behavior. *RAND Journal of Economics* 30, 232–62.

1999. (With J. Tirole.) *Competition in Telecommunications*. Cambridge, MA: MIT Press.

2000. *Incentives and Political Economy*. Oxford: Oxford University Press.

2000. (With D. Martimort.) Mechanism design with collusion and correlation. *Econometrica* 68, 309–42.

2002. (With F. Gasmi, M. Kennet and W. Sharkey.) *Cost Proxy Models and Telecommunications Policy: A New Empirical Approach to Regulation*. Cambridge, MA: MIT Press.

2002. (With D. Martimort.) *The Theory of Incentives: The Principal–Agent Model*. Princeton, NJ: Princeton University Press.

2003. (With A. Faure-Grimaud and D. Martimort.) Collusion, delegation and supervision with soft information. *Review of Economic Studies* 70, 253–80.

2005. *Regulation and Development*. Cambridge: Cambridge University Press.

Bibliography

Clarke, E. 1971. Multipart pricing of public goods. *Public Choice* 2, 19–33.

Groves, T. 1973. Incentives in teams. *Econometrica* 41, 617–31.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–48.

Hansen, L. and Singleton, K. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–86.

Maskin, E. 1999. Nash equilibrium and welfare optimality. *Review of Economic Studies* 66, 23–8.

Maskin, E. 2004. Jean-Jacques Laffont: a look back. *Journal of the European Economic Association* 2, 913–23.

Myerson, R. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 28, 265–81.

Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16, 8–37.

How to cite this article

Tirole, Jean. "Laffont, Jean-Jacques (1947–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000211> doi:10.1057/9780230226203.0923

The New Palgrave Dictionary of Economics Online

Lagrange multipliers

S.N. Afriat

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Lagrange's ‘method of undetermined multipliers’ applies to a function of several variables subject to constraints, for which a maximum is required. Lagrange's procedure avoids the arbitrary distinction between independent and dependent variables. The method involves further variables, the ‘multipliers’ associated with the constraints, which have importance in application to economic problems. Beside the value obtainable from a given resource, one might also wish to know the ‘marginal value’ obtainable when a unit of it is added. The Lagrangian method is therefore a natural tool of the ‘marginalist revolution’, and the multiplier concept underlies ‘shadow price’, ‘implicit value’ and similar expressions.

Keywords

chain rule; convex programming; implicit function theorem; Kuhn–Tucker conditions; Lagange multipliers; Lagrangian function; marginal revolution; separating hyperplane theorem

Article

Lagrange's ‘method of undetermined multipliers’ applies to a function *f* of several variables *x* subject to constraints, for which a maximum is required. The constraints can be stated as $g(x) = q$ where the vector *q* is constant. Ordinarily one might distinguish independent and dependent variables under the constraints, and then by substitution for the dependent variables in *f* one has a function of independent variables whose derivatives must vanish. Instead Lagrange offered a procedure elegantly without the arbitrary distinction between variables and more suitable for some applications. The idea of it has other ramifications, such as for analytical mechanics, calculus of variations and control theory, beside the economic optimization dealt with here. The method involves introduction of further variables *u*, the ‘multipliers’ associated with the constraints. With *n* function variables and *m* constraints we then have $m + n$ variables (*x*, *u*). Lagrange's method depends on $m + n$ relations he obtained to determine these, and so the *n* function variables *x* which are among them and should give the required maximum. The

remaining *m* variables *u*, the ‘undetermined multipliers’, really are just as well determined. But originally they were just part of this device for determining a maximum and their values had no interest even if they could be determined.

The multipliers in fact have a further significance, as derivatives that tell how the maximum value varies as the constraints have variation from a variation of *q*. They therefore have importance in application to economic problems. For, beside the value obtainable from given resources, one might also wish to know the ‘marginal value’ of any resource, the extra value obtainable when a unit of it is added. The Lagrangian method is therefore a natural tool of the ‘marginalist revolution’ and the multiplier has become a part of economic language; it is also the concept that underlies ‘shadow price’, ‘implicit value’ and similar expressions.

The most typical economic maximum problem is formulated differently from that dealt with by Lagrange. Rather, the constraints have the form of inequalities, expressing that some resource availability must not be exceeded; also, functions involved have convexity properties required by diminishing marginal returns. The theory of such problems is different and does not depend on what we have for Lagrange's classical problem. Yet despite the essential difference there is an impressive similarity, from the role of ‘multipliers’, so one can think that here again is Lagrange's method in another shape. But about these multipliers in the new context quite new things can be said. In either case, classical or new, the required maximum is associated with multipliers enabling certain conditions to be satisfied. Here is similarity, but premises and conclusions related to such conditions in each case are different.

Though form brings the two lines together it is altogether a mistake to see coincidence, and rather it is proper to make the treatments entirely separate, instead of trying to deduce one from the other. The difference is well appreciated from the complete difference in proofs of main points. One requires the implicit function theorem, at least in a certain approach, or more simply just the chain rule, as here. The other, convex programming, requires instead the theorem of the separating hyperplane. Again, one is entirely concerned with differentiable functions while the other in its main part is not, though the differentiable case treated by H.W. Kuhn and A.W. Tucker is very familiar. Reassuring for the connection, there are special problems where both lines are applicable, and then the multipliers involved are identical. But even then more can be said about the multipliers than would come simply from the classical case. Our review of the classical and new multiplier theories will make clear the cleavages and connections. We will also see peculiar, and remarkable, features of the matter in the special context of linear programming. Following the ordinary method of distinguishing independent variables and eliminating dependent variables we can, without any other thought about it, arrive at Lagrange's method from consideration of the derivatives the multipliers happen to represent. In that way, beside other possible merit, the multipliers become at the same time identified with those derivatives. Though this is not a usual procedure, it is a counterpart for classical multipliers of an argument that is essential for the new multipliers of optimal programming theory.

It is convenient now to denote the *n* function variables by *z*, reserving *x* for independent variables among these. Lagrange's problem is to determine a maximum of $f(z)$ subject to *m* constraints, stated

$$g(z) = q. \tag{i}$$

Variables are column vectors, and all functions are understood to be differentiable, so for instance g has an $m \times n$-derivative matrix denoted g_z with elements $g_{ij} = \partial g_i / \partial z_j$. As necessary for z to be a maximum (or minimum, in any case a stationary point) Lagrange concluded that

$$f_z = u g_z \text{ for some } u, \tag{ii}$$

in other words the n conditions

$$f_j = \sum_i u_i g_{ij} (j = 1, \ldots, n).$$

Together with the m conditions

$$g_i = a_i (i = 1, \ldots, m)$$

provided by the constraints (i) we have $m + n$ *Lagrange conditions* on the $m + n$ variables

$$u_i (i = 1, \ldots, m), z_j (j = 1, \ldots, n).$$

Lagrange's method depends on the idea that these $m + n$ conditions can be solved to determine the $m + n$ variables, and so the n variables z_j which are among these. Put in another way, the multipliers u_i can be eliminated (and so left ‘undetermined’) and the conditions obtained then solved for the z_j.
With independent and dependent variables x and y under the constraints, the variables have a partition $z = (x, y)$, and we have a function $f(x, y)$ under constraints $g(x, y) = a$ that determine y as a function $y = Y(x, a)$. Then $g[x, Y(x, a)] = a$ is an identity and so, by differentiation with respect to x,

$$g_x + g_y Y_x = 0,$$
(iii)

and with respect to q, $g_y Y_q=1$, and since from here g_y and Y_q are inverse matrices we also have

$$Y_q g_y = 1.$$
(iv)

For any q the constrained values of f are described by $f[x, Y(x, q)]$ as x varies without restriction. The x-derivatives must vanish for a stationary point, that is

$$f_x + f_y Y_x = 0.$$
(v)

On the assumption that this condition determines a unique point x for any q, the stationary points for various q are described by a function $x = X(q)$. Then the corresponding stationary values of f are given by the function

$$F(q) = f[X(q), Y(X(q), q)],$$

with derivatives

$$F_q = f_x X_q + f_y(Y_x X_q + Y_q) = (f_x + f_y Y_x) X_q + f_y Y_q \text{ by (v)}.$$

Hence

$$F_q g_x = (f_y Y_q) g_x = (f_y Y_q)(-g_y Y_x) \text{ by (iii)} = -f_y(Y_q g_y) Y_x = -f_y Y_x \text{ by (iv)} = f_x \text{ by (v)},$$

and also

$$F_q g_y = (f_y Y_q) g_y = f_y (Y_q g_y) = f_y b_y \text{(iv)}.$$

It has now been seen that

$$F_q g_x = f_x, \; F_q g_y = f_y,$$

that is, $f_z = F_q g_z$, which is (i) with $u = F_q$. Thus we have Lagrange's conditions, together with the identification $u = F_q$ for the multipliers.

For any *x*, the existence of *u* so that (*x*, *u*) satisfy Lagrange's conditions (i) and (ii) is the condition for *x* to be a stationary point. It is therefore necessary for *x* to be a maximum, or a minimum, and on its own not sufficient for *x* to be either. Solutions of Lagrange's conditions, if there are any, therefore provide all stationary points, possibly many, without information that any should be a maximum. However, should a maximum be known to exist and the conditions be found to have a unique solution (*x*, *u*) then *x* is known to be that maximum. This is a common circumstance with many applications and where the method has strength.

Given any stationary point *x*, such as could be found from a solution of Lagrange's conditions, and so obtained by a condition on first derivatives at *x*, one can possibly find out if it is a local maximum, or a maximum in some neighbourhood of *x* under the constraints, by an examination of further conditions bringing in higher derivatives of *x*. However, no conditions on derivatives simply at the point *x* will tell anything about *x* except in the local sense. There is no way of telling *x* is a global maximum simply from a satisfaction of some condition on derivatives at *x*, of any order. Of course in economics a maximum is significant only in the global sense. Fortunately, typical functions of economics have convexity properties that enable one to go further on the basis of local conditions. Connected with this, any stationary point of a convex, or concave, function is necessarily a global minimum, or maximum, so in such cases first order conditions are enough. This matter has a part in further theory of Lagrange multipliers in the more typically economic context of convex programming.

Lagrange's method can be described with reference to the ‘Lagrangian function’

$$L(x, u) = f(x) - u[g(x) - g],$$

as requiring the *x* and *u* derivatives to be set to 0. This way of putting it is without significance except as

a cook-book statement. One first learns about setting derivatives to zero when there are no constraints, and now even though there are constraints one can with confident familiarity do it again, even with the impression that the Lagrangian function should be at a maximum as if the recipe had that sense. There is better occasion for something like this in convex programming, where *u* is fixed so as to make *x* a maximum. A problem with inequality constraints is stated

$$(M)\text{Max } f(x) : g(x) \leq q,$$

functions being defined in a set *A*. It can be imagined that *A* is an *activity* set, and the performance of any $x \in A$ gives a return *f*(*x*) and has a cost in terms of various resources given by the vector *g*(*x*), so for *feasibility* this must not exceed the available stock *q*, so $g(x) \leq q$ is required. The problem is to find an *optimal solution*, an activity *x* that gives the greatest return attainable with the available resources, as asserted by the condition

$$M(x) = : g(x) \leq q,\ g(y) \leq q \Rightarrow f(y) \leq f(x).$$

The *limit function* associated with the problem is

$$F(z) = \text{Sup}\ [f(x) : g(x) \leq z],$$

and a *support solution u* is defined by the condition

$$D(u) = : F(z) - F(q) \leq u(z - q) \text{for all } z,$$

equivalent to *u* being a support gradient of *F* at the point $z = q$.
Support solutions correspond to Lagrange multipliers in that they are variables associated with the constraints that give a means for characterizing optimal solutions. Thus, for a pair (*x*, *u*), *complementary slackness* is defined by

$$C(x, u) = : g(x) \leq q,\ u \geq 0,\ ug(x) = uq,$$

and a *shadow solution* by

$$S(x, u) = : f(x) - ug(x) \geq f(y) - ug(y) \text{for all } y.$$

An important proposition, not requiring any assumptions whatsoever about the set *A* or the functions *f* and *g* defined in it, is that any given pair (*x*, *u*) is a shadow solution with complementary slackness if and only if *x* is an optimal solution and *u* a support solution, that is,

$$M(x) \,\&\, D(u) \Leftrightarrow C(x, u) \,\&\, S(x, u).$$

For characterizing optimal solutions by means of the condition on the right, the outstanding issue therefore is the existence of a support solution. We will find this guaranteed under conditions natural for economics at least.

A *convex problem* is one where *f* is a concave function and the elements of *g* are convex. The only importance is to make the limit function *F* concave. Then it has a linear support, and so a support gradient providing a support solution, at any interior point of the region where it is finite. Now with *F*(*q*) finite, Slater's condition which requires $g(x) < q$ for some *x* assures that *q* is exactly such a point. Thus for a convex problem with Slater's condition, and with *F*(*q*) finite, as it must be if an optimal solution exists, we do have the existence of a support solution, and so a characterization of all optimal solutions by means of shadow solutions with complementary slackness.

It is a short step from here to the characterization by means of Kuhn–Tucker conditions. These apply to a problem where the activity set *A* is a space of non-negative vectors, and the functions are differentiable. All that has to be known further is that for a differentiable concave function $\phi(x)$ subject to $x \geq 0$ to be a maximum it is necessary and sufficient that

$$\varphi_x \leq 0,\ x \geq 0,\ \varphi_x x = 0.$$

Applied to the Lagrangian $f(x) - u[g(x) - q]$, with *u* fixed and non-negative, and *x* restricted non-negative, the conditions *S*(*x*, *u*) for a shadow solution become

$$f_x - ug_x \leq 0,\ x \geq 0,\ (f_x - ug_x)x = 0,$$

and so now, with complementary slackness $C(x, u)$, we have the Kuhn–Tucker conditions. In case $x > 0$ the conditions just obtained reduce to $f_x = ug_x$, in other words, ordinary Lagrange conditions, the support solution u providing the multipliers.
With F concave, it is differentiable at a point q if and only if it has a unique support gradient u there, and then the support gradient coincides with the differential gradient, $u = F_q$. Thus uniqueness of support solutions is associated with differentiability of the limit function F at the point q. The identification $u = F_q$ that can be made in this case is comparable with the identity of classical Lagrange multipliers with derivatives of the stationary value function. But this new multiplier theory, even for the Kuhn–Tucker case, in no way depends on differentiability of the limit function. Also, for the linear approximation near q that is available in the differentiable case, we know more about it in that the error is always positive, or that it overestimates the limit function, not just locally but everywhere. Consider now a standard linear programming problem.

$$(M)\ \text{Max}\ px: ax \le q,\ x \ge 0.$$

Another characterization for support solutions of LP problems can be noted, coming from the homogeneity. Thus, with F as the limit function of (M), the condition for u to be a support solution becomes

$$F(q) = uq,\ F(z) \le uz \text{ for all } z.$$

Since (M) is a convex problem the foregoing will apply to it. Also it has the required form for application of the Kuhn–Tucker conditions, which, following the way we put them before with some rearrangement in the second line, become

$$\begin{aligned} ax &\le q, & u &\ge 0, & uax &= uq, \\ ua &\ge p, & x &\ge 0, & uax &= px. \end{aligned}$$

We know from the foregoing that (x, u) is a solution of these conditions if and only if x is an optimal solution and u a support solution of the problem (M).
There is a symmetry in the situation that enables these conditions to be read differently. With an exchange of role between x and u they become Kuhn–Tucker conditions for the problem

$$(W)\ \text{Min}\ uq;\ ua \geq p,\ u \geq 0,$$

and so they hold if and only if *u* is an optimal solution and *x* a support solution of (*W*). It follows that support solutions of either problem are identical with optimal solutions of the other.
Of course, (*M*) and (*W*) are a standard dual pair of LP problems, and so by the LP duality theorem one has an optimal solution if and only if the other does. Hence an LP problem has a support solution if and only if it has an optimal solution. Most remarkable is the way for finding support solutions for an LP problem, as it were differentiating the limit function or finding the 'Lagrange multipliers', by finding optimal solutions for another LP problem – and we know how to do that.

See Also

- convex programming
- Hamiltonians
- nonlinear programming

Bibliography

Afriat, S.N. 1969. The output limit function in general and convex programming and the theory of production. 36th National Meeting of the Operations Research Society of America, Miami Beach, Florida, November 1969. Reprinted, *Econometrica* 39 (1971), 309–39.

Afriat, S.N. 1970. The progressive support method for convex programming. 7th Mathematical Programming Symposium, The Hague, 1970. *Journal of Numerical Analysis* 7(3), 44–57.

Afriat, S.N. 1971. Theory of maxima and the method of Lagrange. *SIAM Journal of Applied Mathematics* 20, 343–57.

Afriat, S.N. 1986. *Logic of Choice and Economic Theory*. Part V: Optimal Programming. Oxford: Clarendon Press.

Dantzig, G. 1963. *Linear Programming and Extensions*. Princeton: Princeton University Press.

Kuhn, H.W. and Tucker, A.W. 1950. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Lagrange, J.L. 1762. Essai sur une nouvelle méthode pour determiner les maxima et minima des

formules intégrales indéfinies. *Miscellanea Taurinensia* 2, 173–95. (Also *Théorie des fonctions analytiques*, 1797.)

Slater, M. 1950. Lagrange multipliers revisited: a contribution to non-linear programming. Cowles Commission Discussion Paper, Math. 403, November. University of Chicago.

How to cite this article

Afriat, S. N. "Lagrange multipliers." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000016> doi:10.1057/9780230226203.0924

close print preview

The New Palgrave Dictionary of Economics Online

laissez-faire, economists and

Roger E. Backhouse and Steven G. Medema

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article traces economists's attitudes towards government intervention since the term ‘laissez-faire’ was first used in late 17th- or early 18th-century France. Understanding of the term has changed significantly since then. Adam Smith, popularly associated with laissez-faire, had a much more nuanced and pragmatic view of the role of the state, as did many of the classical economists and their neoclassical successors. Dissatisfaction with certain aspects of industrial capitalism led to a more interventionist stance during the 20th century, though the second half of the century saw something of a reversion towards the classical approach.

Keywords

Banking School; Bastiat, C. F.; behavioural economics; Boisguilbert, P.; Bright, J.; Cairnes, J. E.; Cambridge School; capitalism; Chalmers, T.; Chicago School; Cobden, R.; consumer surplus; corporatism; Cowles Commission; currency competition; Currency School; division of labour; economic freedom; efficient allocation; free banking; gold standard; government failure; Great Depression; Hayek, F. A. von; Hobson, J. A.; information, economic of; institutionalism; Keynes, J. M.; Keynesianism; laissez-faire; Lange, O. R.; List, F.; Manchester School; marginal revolution; market failure; market socialism; markets; Marshall, A.; McCulloch, J. R.; mercantilism; Mill, J. S.; minimum wages; Mises, L. E. von; Mont Pèlerin Society; natural liberty; new classical macroeconomics; New Deal; North, D.; optimal resource allocation; planning; public choice; rational behaviour; rational choice; Ricardo, D.; role of government; Samuelson, P. A.; Sidgwick, H.; Smith, A.; socialist calculation debate; underconsumptionism; Utilitarianism; Viner, J.; Walras, L.

Article

Origins

The maxim ‘laissez-faire’ is commonly attributed to Vincent de Gournay (1712–1759), on the basis of a claim made the Physiocrat Du Pont de Nemours. However, his likely source, Turgot's ‘Eloge de Gournay’, did not attribute the phrase to Gournay, but implied that Gournay agreed with a well-known remark, made to Louis XIV's minister, Colbert, ‘laissez-nous faire’. This remark was apparently made around 1680 by François Legendre, a merchant and author of a text on commercial mathematics, and may well have been an unpremeditated answer to a question (Castelot, 1987). However, Legendre's contemporary, Pierre de Boisguilbert (1646–1714), repeatedly used the phrase ‘laisse faire la nature’ (‘leave nature alone’), arguing that interference in business spoiled everything, even if it was well-intentioned (Faccarello, 2000). The same ideas appear in English writings of the same period, though the phrase itself does not occur in the writings of commentators such as Nicholas Barbon and Dudley North, writing in the early 1690s, and Henry Martyn and Bernard Mandeville in the early 1700s. North (1691, p. 37), for example, wrote, ‘no people ever yet grew rich by policies; but it is peace, industry, and freedom that brings trade and wealth, and nothing else.’ By the mid-18th century, the idea of laissez-faire was well known, perhaps most clearly stated by the Marquis d'Argenson, in 1858: ‘Laissez faire ought to be the motto of every public authority’ (Castelot, 1987; see also Oncken, 1886). It was Adam Smith who became associated, more than any other economist, with laissez-faire during the 19th and 20th centuries, even though he neither invented the idea, nor used the phrase. The elements from which his *Wealth of Nations* was constructed may not have been original, but the vision of society that he presented, with its emphasis on natural liberty, resonated widely. Smith used the idea of liberty as a radical idea that, though cautiously expressed, placed him alongside radicals such as Tom Paine and Condorcet. Liberty had a political as well as an economic dimension, involving freedom from being oppressed by guilds and monopolies as much as freedom from government interference in one's affairs. However, to those for whom such talk of liberty smacked of Jacobinism and the threat to property posed by the French Revolution, Smith could be reinterpreted as advocating a narrower economic freedom, more conservative in its political implications. Such a reinterpretation happened within a decade of his death (Rothschild, 2001).

The case for laissez-faire

Smith's case for the market did not rest on any claim that it would produce an optimal allocation of resources. Instead, he argued that the system of natural liberty would produce a better outcome than would intervention by the state. There were hints concerning efficient allocation of resources, as on the only occasion when he used the phrase ‘invisible hand’ in the *Wealth of Nations*: in seeking his own advantage, ‘every individual necessarily labours to render the annual revenue of the society as great as he can’ (Smith, 1776, p. 456). Smith opposed mercantilist policies so strongly, not because they prevented an efficient or optimal allocation of resources, or because state action was inherently less efficient than private, but because mercantilist policies were typically the result of using state power to serve the interests of a privileged minority. Merchants conspired to restrain trade, using the state where they could. Smith supported laissez-faire because removal of mercantilist restrictions on trade would help to undermine monopoly, enabling individuals to bring their capital into competition with those who were earning high profits and allowing labour to flow freely between industries and regions. But Smith's support for laissez-faire was not for laissez-faire *in vacuo*: his system presumed a framework of justice

and morality, the basis for which he had analysed in his *Theory of Moral Sentiments* (1759), a book to which he continued to attach great importance, revising it long after the *Wealth of* Nations was published, and in his lectures on jurisprudence delivered in the 1760s (1978).

The classical economists' case for laissez-faire was substantially Smithian, though more narrowly focused on economic freedom. Their consumption-oriented view led them to the belief that freedom of choice was desirable for consumers, and that freedom for producers was the most effective means of satisfying these consumer desires. It was thought that the impersonal forces of the market, working through the system of natural liberty, would then serve to harmonize these interests – or at least would do so to a greater and more beneficial extent than would other systems. The case was comprised of some arguments, such as David Ricardo's theory of international trade, that could be interpreted in terms of optimal resource allocation, but it centred on raising the growth rate. However, some economists saw the case for laissez-faire as primarily a moral one, linked to arguments from evangelical Christian theology. Where Ricardo and many other economists focused on the link between laissez-faire and economic growth, economists such as Thomas Chalmers endorsed laissez-faire because it allowed individualistic capitalism to have its full educative, retributive and purgative effects. There has even been debate over whether this moral case for laissez-faire was in practice more influential than the economic one (cf. Hilton, 1988; Gash, 1989). Certainly in America during this period, the belief in laissez-faire could not be separated from the Protestant spirit of the times, and a belief in its virtues was considered a necessary identifying mark of an economist. When it came to free trade, there was the additional dimension, emphasized by John Bright and Richard Cobden, arguably the most influential advocates of laissez-faire in Victorian Britain, that free international commerce held out the prospect of harmony between nations. The most outspoken supporter of laissez-faire, however, was probably the French writer Frederic Bastiat, a brilliant economic journalist whose vivid examples (for example, candle-makers petitioning for protection against unfair competition from the sun) were influential in making the case for free trade. Standing in a French laissez-faire tradition going back to the 18th century, he linked laissez-faire with harmony between classes, in contrast with the class conflict seen by many English economists. In the United States, laissez-faire was also more than simply an economic doctrine, as is shown by the implications of the slogan of 'free labour' in a society divided over slavery. Along with the sanctity of private property it was part of a moral order that was believed to produce a harmonious society: free enterprise was strongly associated with the virtues of hard work and republican democracy (see United States, economics in (1776–1885 and 1885–1945)).

It was only towards the end of the century, with the developments commonly known as the marginal revolution, that economists began to argue that free competition might produce an optimal allocation of resources, thereby opening up a new defence of laissez-faire. Léon Walras (1954, p. 255) showed that if two conditions – that each product had only one price in the market and that prices equal corresponding costs of production – were satisfied, free competition would produce 'the greatest possible satisfaction of wants'. Marshall offered a doctrine of 'maximum satisfaction' that wedded his demand–supply apparatus with the concept of consumer surplus. However, they did not use these arguments to make a case for laissez faire, for their arguments showed much more clearly than did those of their predecessors why laissez-faire might in practice fail to produce such an optimal allocation. For example, immediately after stating his theorem, Walras pointed out that economists typically exaggerated the implications of the principle of laissez-faire: the conditions of a uniform price and equality of price and cost of production would often not be satisfied and, in any case, it the theorem did not apply to the question of

property.

The limits to laissez-faire

Viner (1960, p. 45), in one of the classic studies of the history of laissez-faire, started by saying that he understood laissez-faire to mean:

> the limitation of government activity to the enforcement of peace and of ‘justice’ in the restricted sense of ‘commutative justice,’ [justice in exchange] to defense against foreign enemies, and to public works regarded as essential and as impossible or highly improbably of establishment by private enterprise or, for special reasons, unsuitable to be left to private operation.

However, whilst Viner is correct to argue that laissez-faire did not mean anarchy and a complete absence of government intervention, his definition begs the question of how much intervention should be allowed: of what are the limits to laissez-faire.

Smith's view of the role of government is close to Viner's view of laissez-faire. The duties of the sovereign included maintaining justice, police, defence and such beneficial public works as would not otherwise be provided. This included support for transport and education – both of which Smith thought essential contributors to the wealth of a nation. It is important to note, though, that Smith's conception often went beyond modern views. For example, his support for education and for the arts was grounded in part in his concerns about the stultifying effects of the division of labour. His analysis of national defence led him to advocate a standing army rather than a militia because of a concern about the problems of attracting the right sort of people to military service in an increasingly wealthy commercial society. Smith's view of the appropriate sphere for state action also went significantly beyond the traditional public goods categories. He supported regulations dealing with public hygiene, legal ceilings on interest rates (to prevent excessive flows of financial capital into high-risk ventures), light duties on imports of manufactured goods, the mandating of quality certifications on linen and plate, certain banking and currency regulations to promote a stable monetary system, and the discouragement of the spread of drinking establishments through taxes on liquor (this being one of various regulations Smith advocated to compensate for the imperfect knowledge – or diminished telescopic faculty – of individuals). He also argued for measures that came within what Viner described as commutative justice. For example, he supported regulations that restricted wages in the interests of the labourer (that is, minimum wages) on the grounds that these redressed the imbalance between worker and employer.

The 19th-century classical economists, while holding to a strong belief in the market as an allocation mechanism, also believed that the market could only operate satisfactorily – harmonizing actions of self-interested agents with the interests of society as a whole – within a framework of legal, political, and moral measures that facilitated certain forms of action while restricting others. They were, in essence, pragmatic reformers, inclined towards laissez-faire but in practice willing to consider each case on its merits. We see this reflected in John Ramsay McCulloch's assertion in 1848 that ‘The principle of *laisser-faire* may be safely trusted to do in some things but in many more it is wholly inapplicable; and to appeal to it on all occasions savours more of the policy of a parrot than of a statesman or a

philosopher' (McCulloch, quoted in Robbins, 1952, p. 43). Over two decades later, John Elliot Cairnes (1870, p. 244) was even more forthright, asserting that the maxim of laissez-faire had 'no scientific basis whatever' but was a 'mere handy rule of practice'. In terms of specific policies, they were willing to support an increasing range of interventions from factory legislation to the state provision of education, the poor laws and measures to promote public health (Robbins, 1952; O'Brien, 2004).

However, whilst the classical economists, like Smith, saw many cases where government action could improve on what would result from laissez-faire, they remained suspicious of government and were vociferously opposed to policies – like those of mercantilism, but also many others – that they believed served the interests of particular groups at the expense of the larger population. They were optimistic that the insights of political economy could be used to point the discipline in a direction that would be beneficial to society and help mitigate the negative effects of partisan advocacy within that process (for example, Mill, 1859; 1861; 1862).

More radical objections to laissez-faire were found outside Britain. The name 'Manchesterism' was widely used, particularly in Germany, to denote British laissez-faire doctrines, and was allegedly the ideology of Manchester's manufacturing classes. The most penetrating critique came from Friedrich List in *The National System of Political Economy* (1856). As Britain had industrialized first, free trade was in her interests, because other countries could not compete; until they were in a position to do so, tariffs were needed. List's ideas were particularly influential in the United States, where economists such as Henry Carey were able to combine commitment to individualism and free enterprise with support for protective tariffs.

One of the additional elements introduced after Smith was the utilitarianism of Bentham and his followers. Though utilitarianism has, on account of the prominence of Philosophic Radicals within political economy, been equated with laissez-faire individualism, this is not correct. On the one hand, there was an authoritarian streak in utilitarianism, from Bentham to reformers such as Edwin Chadwick. On the other hand, there were many supporters of laissez-faire, of whom Gladstone is perhaps the outstanding example, alongside many evangelical political economists, who would have no truck with Benthamite anti-religious rationalism.

The trend away from laissez-faire has its roots in the utilitarian tradition, for utilitarianism provided a basis on which exceptions could be justified. John Stuart Mill (1848, Book V, ch. XI), in what became the dominant textbook on political economy, laid out an extensive list of cases where the system of natural liberty failed to generate outcomes in the best interests of society. He argued that government interference was justified when individuals' actions had spillover effects on others, when individuals did not have the capacity properly to judge the consequences of their own actions or when what would now be called principal-agent problems were present. Prominent here, too, was the distribution question: the classical period witnessed increasing concern about poverty but saw attempts at reducing it as at best futile (owing to natural laws governing distribution) and possibly even counterproductive (because redistributive measures could exacerbate the population problem). Mill challenged the received view here by positing that the laws of distribution were, in fact, mutable, and that state action had the potential to significantly improve the lot of the poor. However, his starting point remained the maxim that '*Laisser-faire*, in short, should be the general practice: every departure from it, unless required by some great good, is a certain evil' (Mill, 1965, p. 945). It was not just Mill who used utilitarianism as a means of justifying departures from laissez-faire. Robert Lowe, a controversial Liberal politician, at one time

Chancellor of the Exchequer, had very much a Smithian view of the merits of laissez-faire but used utilitarian arguments to justify an increasing number of exceptions to this rule (Maloney, 2005). William Stanley Jevons sought to move economic theory sharply away from the framework laid down by Ricardo and Mill, but used utilitarian arguments to justify extensive state intervention.

These utilitarian defences of state intervention were part of a much broader move away from laissez-faire from around the 1870s and 1880s when there developed widespread consciousness of what was, in Britain, called ‘the social problem’ at a time when the electorate in many European countries was widening to include the members of the working class (see Hutchison, 1953; 1978). One reason for the timing was the long recession that followed the collapse of the worldwide boom of 1873 and the severe, and in some countries prolonged, unemployment that resulted. Questioning of laissez-faire was particularly strong in Germany, where the Verein für Sozialpolitik was founded, essentially as an interventionist think tank (see Historical School, German). Its members, of whom Gustav Schmoller was pre-eminent, were known as the ‘Socialists of the lectern’. These attitudes carried over to the United States: many of the founders of the American Economic Association were exposed to them whilst taking their doctorates in Germany, only to find, on their return, a conflict with traditional laissez-faire attitudes. Their challenge to laissez-faire affected not just economic analysis, but economic policy: in the United States, the rise of big business was associated with the development of numerous and very obvious anti-competitive practices, which resulted in the government developing policies of industrial regulation not found in other countries, at least in relation to inter-state trade, culminating in the anti-trust acts of 1890 and 1913. Economists supported such measures with analysis of phenomena such as ‘cut-throat’ competition that went beyond anything found in, say, Jevons, Walras or Marshall (see United States, economics in (1885–1945)).

The British approach was dominated by the Cambridge School, at the headwaters of which was Henry Sidgwick, the author of one of the classic defences of utilitarianism ethics (1907). Sidgwick (1904) took Mill's analysis further: *all* outcomes that constituted departures from social utility maxima were potential candidates for government intervention. Sidgwick's optimism about the prospects for state action marked a significant turn. He was convinced that recent reforms in governance structures – such as the establishment of boards and commissions staffed by experts – portended great things for the ability of state action to improve on market performance.

Sidgwick's perspective signalled what was to become a distinctive Cambridge approach to issues of laissez-faire, continued by Marshall (1890) and A.C. Pigou (1912; 1920). Marshall wedded his demand–supply analysis with the concept of consumer's surplus to provide a tool with which the welfare implications of laissez-faire and government intervention could be analysed. In analysis since seen as flawed through its neglect of producer's surplus, Marshall argued that subsidies to industries characterized by increasing returns and taxes on industries operating under decreasing returns could enhance efficiency. Pigou took all this a step further with his analysis of private and social net products, which proved to be a very effective tool for illustrating both the nature of market failures and the means by which government corrective actions could prod markets toward efficiency. He argued that divergences between private and social net products constituted a ‘*prima facie* case’ for government intervention, but he also allowed that the state will not necessarily be capable of improving on market performance. Like his predecessors, Pigou was optimistic that governmental reforms held great promise, but he was also concerned about many of the governance problems that we now associate with public

choice analysis. The policy conclusions of the Cambridge economists, including the case for free trade, rested as much on beliefs about the competence of government to implement beneficial policies as on the results of formal economic theory.

The First World War and its aftermath

Laissez-faire was far from universally accepted before the First World War, but the move towards the welfare state and towards regulation of industry was generally gradual (free trade had never become universal, some countries never having abandoned protective tariffs). Economists made frequent concessions to socialism (this was easy because the term had such an elastic meaning) but could maintain the idea that laissez-faire should remain the general rule. After 1918, that confidence was harder to maintain. The Bolshevik revolution and the establishment of the Union of Soviet Socialist Republics (USSR) presented the challenge of an alternative economic system. Economic dislocation was widespread in Europe in the 1920s and worldwide after the onset of the Great Depression. To an extent unparalleled before 1914, laissez-faire and even capitalism were called into question, in the writings of economists as much as among politicians and policymakers.

Of particular significance was the extension of discussions of laissez-faire to what would now be considered macroeconomic issues – money and the business cycle. ‘Free banking’ might exist in some American states, but the need for some sort of monetary policy had been generally accepted since the bullion debates during the Napoleonic Wars. Though there were exceptions, it became accepted that paper money should have a fixed value in terms of precious metal. There were several reasons why this was seen as consistent with laissez-faire. To allow the value of paper to fall below par was to defraud those who had entered into contracts denominated in terms of money. A metallic standard, which increasingly meant the gold standard, facilitated trade. Most important, though there were underconsumptionists (more than are often recognized), they were in a clear minority among economists. The parallel with 20th-century debates over laissez-faire in macroeconomics is found in the debate between the Currency and Banking Schools in the 1840s (see Banking School, Currency School, Free Banking School). The Currency School sought to prevent the emergence of financial crises by making paper currency behave like a metallic one, removing discretion from central bankers. In contrast, the Banking School argued that, in times of depression, a central bank should pursue an accommodating monetary policy, lending according to sound banking principles. Strictly speaking, this was a debate about the type of policy to be pursued, not whether or not to intervene, but it posed the issue of discretion in monetary policy that came to be associated, in the 20th century, with debates over laissez-faire. Such ideas framed much of the discussions of central bank policy as late as the inter-war period, when the appropriate policy for the US Federal Reserve system was being debated (Laidler, 1999, chs 8–9).

The extent to which such a way of thinking carried over into the 20th century is illustrated by the ‘Austrian’ theories of money. Though in general ardent supporters of laissez-faire, Ludwig von Mises and Friedrich Hayek argued for the implementation of what they considered appropriate monetary policy. Mises (1912, pp. 456–63) supported the gold standard on the grounds that it rendered the value of money independent of political influence. Management of the currency meant inflation, a policy inevitably doomed to eventual failure. Hayek, though theoretically innovative, maintained this emphasis

on sound, or ‘neutral’ money; the problem of the business cycle was caused by the supply of money being too elastic. Despite his otherwise impeccable credentials as a supporter of laissez-faire, it was only in the 1970s that he turned to completely free banking and competition in the supply of currency (Hayek, 1999).

Though his target was the British authorities, this was the mindset that John Maynard Keynes attacked in his *Tract on Monetary Reform* (1923). He argued that to regard the gold standard as a fact of nature was to perpetuate an illusion. ‘There is,’ he wrote, ‘no escape from a “managed” currency, whether we wish it or not’ (Keynes, 1971, p. 136). He continued (1971, p. 138):

> A regulated non-metallic standard has slipped in unnoticed. *It exists*. Whilst the economists dozed, the academic dream of a hundred years, doffing its cap and gown, clad in paper rags, has crept into the real world by means of the bad fairies – always so much more potent than the good – the wicked ministers of finance.

It was but a short step from this to announcing ‘[t]he end of laissez-faire’ (Keynes, 1972, pp. 272–94). His account of laissez-faire focuses on the philosophical and political rather than the economic, his point being that it cannot rest on ideas of natural liberty, for there is no such thing. It was necessary, he argued, to work out the agenda and non-agenda of the state without the Benthamite prior assumption that interference was likely to be ‘generally pernicious’ (1972, p. 288). The agenda for the state should comprise those things that are otherwise not done, which he identified as regulation of currency and credit, management of investment, and policy in relation to population size (1926, p. 292). His *General Theory of Employment, Interest and Money* (1936) provided a new theoretical justification for such ideas, but the idea that the state's main agenda item was the maintenance of the level of investment, remained.

Most of Keynes's arguments were far from novel. J.A. Hobson and other underconsumptionists had long questioned the ability of unregulated capitalism to produce the appropriate level of saving. Not only had it been argued, even before 1914, that government spending could raise the level of employment, but schemes for doing so had been worked out. The significance of his arguments, which became clear only from the 1940s, lay in the fact that an economist at the heart of the establishment was arguing against laissez-faire from a macroeconomic point of view. Furthermore, his attack on the philosophical foundations of laissez-faire, by someone who was far from being a socialist, indicated a changing climate of opinion towards one in which management of the economy came to be seen as a central role of government.

Planning was also becoming more acceptable in the United States throughout the inter-war period. It has been argued that this marked a radical departure from previous attempts to reform the economy because it was ‘predicated on the assumption that intervention … was necessary for a well-functioning, dynamic economy’ (Balisciano, 1998, p. 154; see also Barber, 1985; 1996). The First World War had shown that planning could raise output above what had been thought possible, and economic fluctuations in the immediate post-war period suggested that government intervention might be desirable. There was a move to create a new economics, appropriate to a new age, exemplified in the White House by Herbert Hoover, an engineer who turned readily to experts. The move towards a scientific economics that could perform this task was represented by institutionalism (see institutionalism, old) but planning took many

forms, from the social planning of Rexford Tugwell and John Maurice Clark to the macroeconomic planning of Laughlin Currie (see Balisciano, 1998). The move towards providing a scientific foundation for policy extended beyond institutionalism: for example, both Wesley Mitchell and Irving Fisher called for quantitative research. These various strands of thought came together in the New Deal, with its mixture of microeconomic planning, macroeconomic management and extensive quantitative research. In continental Europe, planning was observed in the Soviet Union and in Germany under National Socialism. In other countries, corporatist ideas were highly influential. When placed alongside experience of the Great Depression, this raised the question of whether capitalism itself, let alone laissez-faire, was a viable alternative to planning. The socialist calculation debate, initiated by Otto Neurath and von Mises immediately after the First World War, tackled the question of whether a planned economy could be as efficient as a capitalist one. The significance of this controversy is twofold. In making the case that it was theoretically possible to design a socialist economy that was as efficient as a capitalist one, Oskar Lange and the so-called market socialists were shifting the climate of opinion in favour of planning. However, perhaps more significant in the longer term is the fact that planning was defended using arguments about the optimality of a perfectly competitive equilibrium. This took the arguments of Walras and Marshall a stage further, towards the post-war welfare theorems of Kenneth Arrow and Gerard Debreu. A defence of socialism could, with a small twist, be turned into an argument for laissez-faire.

The most theoretically innovative critic of the central planners was Hayek, who developed the idea that the market could be seen as an information-processing mechanism (Hayek, 1937; see also Gamble, 2006). The information possessed by modern societies was necessarily limited, imperfect and dispersed among many individuals, so to assume, as did the market socialists, that this knowledge could be available to central planners was a mistake. Markets enabled prices and economic activities to reflect the knowledge held by millions of distinct individuals and organizations. The significance of this theory is that it reinforces the point that arguments for laissez-faire do not need to rest on any claim that it produces an optimal outcome. If knowledge is imperfect, as Hayek claimed, it is not meaningful to argue in terms of optimality.

The Second World War and after

During the Second World War, planning was widely practised, not just in Germany and the Soviet Union but also in Britain and the United States, perhaps inevitably when military uses accounted for around 40 per cent of national production. Unlike in the First World War, careful attempts were made to plan for the post-war order and although this was to be a liberal world order, based on free trade and free movement of capital, it was to be a planned order, with appropriate national and international institutions to support it. Experience of the First World War was taken as demonstrating that a well-functioning free market economy would not occur spontaneously. The degree and nature of planning and commitment to laissez-faire varied from country to country: the United States may have been at one end of the spectrum, with suspicion that planning might be tainted by Communism, and with government accounting for a lower share of national output than in Europe, but the importance of the defence sector during the Cold War meant that the role of government was far-reaching. Though there was a retreat from the level of planning achieved during the war, and even compared with the New Deal, in favour of a free market economy, government remained very significant.

Economics had also changed, becoming more technical, more mathematical (see, for example, United States, economics in (1945 to present)). However, the relationship of this change to thinking about laissez-faire was complex. Many of the techniques used in this more technical economics had roots in economists' wartime activities, and were linked with planning. The Cowles Commission, the main centre of mathematical economics in the 1940s, was closely associated with these developments and was also linked, through Oskar Lange, with socialism. A case can be made that microeconomic theory in this period strengthened the case against laissez-faire by developing theories of market failure. General equilibrium theory may have been seen by outsiders as demonstrating rigorously the efficiency of competitive markets, but the restrictive assumptions needed could equally be taken as demonstrating that an efficient allocation of resources required conditions that could never be satisfied in the real world.

It was in macroeconomics that the challenge to laissez-faire was strongest, the nearest to a consensus view being the neoclassical synthesis articulated in the third edition of Paul Samuelson's *Economics* (1955). This proposed that if demand management could maintain full employment, the allocation of resources between economic activities could be undertaken by the market. Laissez-faire was rejected at the macroeconomic level in favour of a 'Keynesian' policy of demand management (see Keynesianism). At a microeconomic level, laissez-faire was limited by the need to provide public goods, deal with externalities and control monopoly. This left much scope for debate over precisely where the limits to laissez-faire lay, from those who favoured extensive intervention to the Chicago School, which challenged the need for active competition policies and, increasingly, the Keynesian consensus.

The pervasiveness of planning in the late 1930s and early 1940s provoked a response from some scholars who believed that classical liberal values were threatened. The most prominent such response was by Hayek, whose *The Road to Serfdom* (1944) became a best seller. In 1947 he helped establish the Mont Pèlerin Society, which became the centre of a network of economists committed to free-market ideas. This network encompassed research institutes aimed at influencing policy and academic economists, of which the most significant was a group centred on Chicago. This offered a much more optimistic, and even radical, view of what could be achieved under laissez-faire than was generally accepted by economists in the 1950s and 1960s. Laissez-faire was as much an end as a means, exemplified in Milton and Rose Friedman's *Free to Choose* (1979).

The 1960s and 1970s saw the beginnings of a major shift in the way that economists approached issues related to laissez-faire. At the heart of this shift was an extension in the scope of the theory of rational choice to the point where it could encompass all aspects of behaviour (see rationality, history of the concept). Two developments were particularly important in moving economists towards laissez-faire. The first was the application of rational choice theory to government and bureaucracies, resulting in the development of a theory of government failure to parallel the earlier theory of market failure. Rent-seeking, legislative vote trading and bureaucratic waste took their places alongside externalities and public goods as phenomena to be taken into account. This was most visible in public choice theory, but spread much more widely. The second was the transformation of macroeconomics associated with the new classical macroeconomics. Rational behaviour was taken to imply that markets would clear and that agents would form expectations rationally, which led to a presumption that attempts to stabilize economic activity would be counterproductive; that laissez-faire applied at the macroeconomic level. This was believed to explain the apparent breakdown of Keynesian policies in the 1970s. This did not go unchallenged, but there was a clear shift in the weight of economists' opinions on laissez-faire at both

microeconomic and macroeconomic levels.

However, other developments worked in the opposite direction. There was much work on the economics of information, which added to the weight of the evidence for why free markets might not be efficient. These involved questioning some of the most basic ideas of supply and demand on which much of the traditional case for laissez-faire rested, a point made most forcefully by Joseph Stiglitz. Market failures can occur in both the production and dissemination of information due to the informational asymmetries and uncertainty that result. A lack of effective futures markets causes intertemporal inefficiencies (for example, on the environmental front), moral hazard and adverse selection problems can cause insurance markets to fail, and the use of education as a signalling and screening device can lead to over-investment in education. At the macroeconomic level, problems associated with risk and information can cause financial markets to react in ways that are destabilizing. Game theory, too, presented problems, showing how strategic behaviour had a propensity to generate market outcomes that departed – sometimes substantially – from the dictates of optimality.

By the new millennium, some of the assumptions underlying this resurgence of laissez-faire thinking were being challenged. A form of Keynesianism re-emerged in the form of inflation targeting through interest rates, a development that reflected both macroeconomic theory and lessons learned from experience. Behavioural economics raised questions about human motivation and opened up the possibility of new ways of analysing economic behaviour. It is, however, too soon to tell what the implications of this will be for attitudes towards laissez-faire.

However, despite the resurgence of laissez-faire thinking, the context is radically different from that prevailing at the beginning of the 20th, let alone the 19th, century. In macroeconomics, the case for central banks operating according to rules so as to stabilize economic activity is, in some sense, almost universally accepted. Debates centre on what those rules should be, not whether there should be rules. At the micro level, there has been a significant expansion in the sphere of market activity since the 1980s, as a result of the deliberate creation of new markets, from financial options to CO_2 emissions. These markets are not simply heavily regulated: many of them are designed by government, usually on the basis of economists' advice. Furthermore, the scale of government is now such that government contracts are an inherent part of the activities of many businesses. In such an environment, it can be questioned whether the traditional distinction between laissez-faire and government intervention has become out of date.

A further complication in discussions of laissez-faire results from the enormous expansion of international organizations, from the International Monetary Fund (IMF) and the World Bank to the World Trade Organisation (WTO) and the United Nations. These have made it meaningful to discuss alternatives to laissez-faire at an international level at the same time that the so-called globalization of economic activity has raised new questions about its benefits and costs to different groups. If trade is to take place within rules laid down by organizations such as the WTO and the IMF, should these rules allow governments to protect industries or workers from what they perceive to be unfair international competition? Does laissez-faire apply to national governments or simply to private organizations? This is a complicated question in a world where many private companies are substantially shaped by their relations with governments.

Conclusions

It has been argued that the developments of recent decades have taken us back to Adam Smith and a laissez-faire welfare economics asserting the efficacy of the market in channelling individual self-interest towards actions that benefit society; and to a pre-Keynesian era when the need for active macroeconomic management was not recognized. But this is not accurate. The 19th- and early 20th-century exponents of laissez-faire, from Mill to Pigou (and perhaps even to Lange or Samuelson), saw an ever-widening range of exceptions to the general rule. Their policy prescriptions reflected well-articulated ideas about market failure and much less completely theorized views about the capacity of government to remedy such problems. As a result of recent developments there is, in general, awareness that neither the market nor government is perfect – that the choice is between two highly imperfect alternatives. Theory cannot settle the matter unless reasons are adduced to play down the importance of market failure (the ‘libertarian’ response) or government failure (the ‘socialist’ response). Because of this, and because of the transformed role of government, there is a strong case for arguing that notions such as ‘laissez-faire’ and ‘state action’, especially if this is seen as an either/or choice, are not particularly helpful. However, there is a reversion to Smithian ideas in one sense: economists increasingly recognize, as did Smith, that markets do not exist apart from an institutional structure that includes the state and its legal system. Discussions of state action are not usually about *replacing* the market; rather, they are about nudging markets this way or that in order to obtain a more desirable outcome than would obtain otherwise.

See Also

- Banking School, Currency School, Free Banking School
- Historical School, German
- institutionalism, old
- Keynesianism
- marginal revolution
- rationality, history of the concept
- United States, economics in (1776–1885)
- United States, economics in (1885–1945)

Bibliography

Balisciano, M.L. 1998. Hope for America: American notions of economic planning between pluralism and neoclassicism. *History of Political Economy* 30(Supp.), 153–78.

Barber, W.J. 1985. *From New Era to New Deal: Herbert Hoover, the Economists, and American Economic Policy, 1921–1933*. Cambridge: Cambridge University Press.

Barber, W.J. 1996. *Designs within Disorder: Franklin D. Roosevelt, the Economists, and the Shaping of American Economic Policy, 1933–1945*. Cambridge: Cambridge University Press.

Cairnes, J.E. 1870. Political economy and laissez-faire. In *Essays on Political Economy*. London: Macmillan, 1873.

Castelot, E. 1987. Laissez-faire, laissez-passer: history of the maxim. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Faccarello, G. 2000. *Foundations of Laissez-Faire: The Economics of Pierre de Biosguilbert*. London: Routledge.

Friedman, M. and Friedman, R. 1979. *Free to Choose*. New York: Harcourt Brace.

Gamble, A. 2006. Hayek on knowledge, economics and society. In *The Cambridge Companion to Hayek*, ed. E. Feser. Cambridge: Cambridge University Press.

Gash, N. 1989. Review of Hilton 1988. *English Historical Review* 104, 136–40.

Hayek, F.A. 1937. Economics and knowledge. *Economica N.S.* 4(13), 33–54.

Hayek, F.A. 1944. *The Road to Serfdom*. London: Routledge.

Hayek, F.A. 1999. Good money. In *The Collected Works of F.A. Hayek*, ed. S. Kresge. London: Routledge.

Hilton, B. 1988. *The Age of Atonement: The Influence of Evangelicalism on Social and Economic Thought, 1785–1865*. Oxford: Clarendon Press.

Hutchison, T.W. 1953. *A Review of Economic Doctrines, 1870–1929*. Oxford: Oxford University Press.

Hutchison, T.W. 1978. *On Revolutions and Progress in Economic Knowledge*. Cambridge: Cambridge University Press.

Keynes, J.M. 1923. A tract on monetary reform. In *The Collected Works of John Maynard Keynes*, vol. 4. London: Macmillan, 1971.

Keynes, J.M. 1926. The end of laissez faire. In *Essays in Persuasion*. In *The Collected Writings of John Maynard Keynes*, vol. 9. London: Macmillan, 1972.

Keynes, J.M. 1936. The general theory of employment, interest and money. In *The Collected Writings of John Maynard Keynes*, vol. 7. London: Macmillan, 1973.

Laidler, D.W. 1999. *Fabricating the Keynesian Revolution: Studies of the Inter-War Literature on*

Money, the Cycle, and Unemployment. Cambridge: Cambridge University Press.

List, F. 1856. *The National System of Political Economy*. New York: Kelley, 1966.

Maloney, J. 2005. *The Political Economy of Robert Lowe*. London: Palgrave Macmillan.

Marshall, A. 1890. *The Principles of Economics*. 8th edn. London: Macmillan, 1920.

Mill, J.S. 1848. Principles of political economy, with some of their applications to social philosophy. In *The Collected Works of John Stuart Mill*, vols 2–3, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Mill, J.S. 1859. *On Liberty*. New York: Classics of Liberty Library, 1992.

Mill, J.S. 1861. Considerations on representative government. In *The Collected Works of John Stuart Mill: Essays on Politics and Society*, vol. 19, ed. J.M. Robson. Toronto: University of Toronto Press, 1977.

Mill, J.S. 1862. Centralisation. In *The Collected Works of John Stuart Mill: Essays on Politics and Society*, vol. 19, ed. J.M. Robson. Toronto: University of Toronto Press, 1977.

Mises, L. von. 1912. *The Theory of Money and Credit*, tr. H.E. Bateson. Indianapolis, IN: Liberty Classics. 1980.

North, D. 1691. *Discourses upon Trade*. London: Thomas Basset.

O'Brien, D.P. 2004. *The Classical Economists Revisited*. Princeton, NJ: Princeton University Press.

Oncken, A. 1886. *Die Maxime Laissez Faire et Laissez Passer*. Bern: Wyss.

Pigou, A.C. 1912. *Wealth and Welfare*. London: Macmillan.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Robbins, L.C. 1952. *The Theory of Economic Policy in English Classical Political Economy*. London: Macmillan.

Rothschild, E. 2001. *Economic Sentiments: Adam Smith, Condorcet and the Englightenment*. Cambridge, MA: Harvard University Press.

Samuels, W.J. 1966. *The Classical Theory of Economic Policy*. Cleveland, OH: World.

Samuelson, P.A. 1955. *Economics*, 3rd edn. New York: McGraw Hill.

Sidgwick, H. 1904. *Principles of Political Economy*. London: Macmillan.

Sidgwick, H. 1907. *The Methods of Ethics*, 7th edn. London: Macmillan. Repr. Indianapolis: Hackett, 1981.

Smith, A. 1759. The theory of moral sentiments. In *The Works and Correspondence of Adam Smith*, vol. 1, ed. D.D. Raphael and A.L. Macfie. Oxford: Oxford University Press, 1759.

Smith, A. 1776. An inquiry into the nature and causes of the wealth of nations. In *The Works and Corresponence of Adam Smith*, vol. 2, ed. R.H. Campbell, A. Skinner and W.B. Todd. Oxford: Oxford University Press, 1976.

Smith, A. 1978. Lectures on jurisprudence. In *The Works and Correspondence of Adam Smith*, vol. 5, ed. R.L. Meek, P.G. Stein and D.D. Raphael. Oxford: Oxford University Press, 1978.

Viner, J. 1960. An intellectual history of laissez faire. *Journal of Law and Economics* 3, 49–69.

Walras, L. 1954. *Elements of Pure Economics: Or the Theory of Social Wealth*, tr. W. Jaffe. Philadephia, PA: Orion.

How to cite this article

Backhouse, Roger E. and Steven G. Medema. "laissez-faire, economists and." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000232> doi:10.1057/9780230226203.0925

The New Palgrave Dictionary of Economics Online

Lancaster, Kelvin John (1924 – 1999)

Ronald Findlay

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Kelvin Lancaster made at least three major original contributions to economic theory. The first, together with Richard G. Lipsey, is ‘The General Theory of the Second Best’ in the area of welfare economics. The other was his ‘characteristics’ approach to the pure theory of consumer behaviour. The third, based on this new approach to consumer behavior was a solution to the problem of ‘socially optimal product differentiation’, which showed how to balance the consumer's desire for more variety in the choice of goods to consume against economies of scale in the production of each good.

Keywords

American Economic Association; characteristics; consumer choice; Econometric Society; Johnson, H. G.; Lancaster, K. J.; monopolistic competition; new trade theory; Pareto efficiency; product differentiation; scale economies

Article

Kelvin Lancaster was born in Sydney, Australia, on 10 December 1924. He volunteered for the Royal Australian Air Force at the age of 18 and was trained as a bombardier. The war fortunately ended before this kindest and gentlest of men was required to release any bombs using the new Norden bombsight on which he had been trained. He graduated from the University of Sydney with a BSc in mathematics and geology (1948) and a BA (1949) an MA (1953), both in English literature. A growing interest in economics took him to the London School of Economics in 1953, where he obtained the BScEcon degree with First Class Honours as an external student without ever having taken a single course in economics, and his Ph.D. in 1958. He was on the faculty of the LSE from 1954 to1962, and immediately became one of the brightest stars of the famous seminar led since the early 1930s by Lionel Robbins, whose participants over the years included the likes of Hayek, Hicks, Kaldor, Lerner, Meade and many others of comparable stature.

Lancaster and Richard Lipsey, then also at the LSE, each submitted a paper to the *Review of Economic Studies*, edited by the indefatigable Harry Johnson, Lipsey's on tariffs and customs unions and Lancaster's on monopoly and nationalized industries. Johnson noted that they were both making the same general point, namely, that if one of the necessary conditions for a Pareto optimum failed to hold it was not in general desirable to make the remaining conditions hold. In other words, the Paretian conditions had to be fulfilled in their entirety for a ‘first-best’ optimum to be reached. If one condition failed to hold a ‘second-best’ optimum would in general involve departures from some or all of the others. Johnson suggested that the two papers be merged, making this fundamental general point and giving the customs union and nationalized industry problems as illustrative examples. The result was the celebrated paper on ‘The General Theory of the Second Best’ by Lipsey and Lancaster (1956) that has changed the way economists have since thought about economic policy in every field.

Lancaster moved to the United States in 1962, first to Johns Hopkins (1962–66) and then to Columbia, following his wife Dvora, who had been admitted to Columbia Law School. He remained at Columbia for the rest of his career, becoming the John Bates Clark Professor of Economics in 1978. In 1966 he published ‘A New Approach to Consumer Theory’ in the *Journal of Political Economy*, following it in 1971 with a more detailed treatment in the book *Consumer Demand: A New Approach*. His attempt at a new approach to the classic problem of consumer choice was motivated by the desire to make this most parsimoniously elegant of all economic theories more operational and relevant to the modern industrial world of an almost infinite variety of products. The standard theory involved considering the consumer as maximizing a utility function $U(x)$ subject to a budget constraint $px = I$, where x is an n-dimensional vector of goods, p the corresponding vector of prices and I the income of the consumer. The basic idea of the alternative approach he proposed is to regard the arguments of the utility function not as goods but the *characteristics* or attributes of these goods that they provide to the consumer in varying amounts and proportions, the goods themselves being merely the means whereby the consumer satisfies his essential wants. A simple version of the Lancaster approach therefore regards the consumer as maximizing $U(z)$, where z is an m-dimensional vector of characteristics, subject to $z = Bx$, where B is an (m x n) matrix representing the ‘technology of consumption’ or the amount of each characteristic embodied in each good, and the budget constraint $px = I$ as before. Lancaster regards the number of characteristics m as much smaller than the number of goods n in a modern economy.

Suppose, for purposes of illustration, that n is five and m is two. Given I and p we can find how much of each good can be obtained if all of the income is spent on that good alone. The amounts of each of these five goods obtained this way yield a pair of the amounts of the two characteristics that they embody. Number these goods from one to five in descending order of the ratio of the first characteristic to the second that they provide. Each of these five pairs can be plotted as a point in a diagram with the first characteristic on the vertical and the second on the horizontal axis. These points form the five vertices and the straight lines connecting them the four edges or flats of the ‘characteristics-possibility frontier’ or CPF available to the consumer, given his income and the prices of the goods that he is facing. Superimposing the map of convex indifference curves between the two characteristics specified by $U(z)$, we can find the optimal choice of the two characteristics for the consumer by the point at which the highest attainable indifference curve is tangent to the CPF. If the optimal point is on a flat the consumer will demand the convex combination of the two goods spanning the flat yielding that point; the only other possibility is for the optimal point to be at a vertex, in which case only the corresponding good is demanded. Each consumer will therefore demand at most two of the five goods available to him. Any

other consumer will face the same price vector *p* for the goods and the same objective technology of consumption represented by the matrix *B*. Differences in income will result in radial expansions or contractions of the CPF, leaving its structure unchanged. The utility functions $U(z)$ of the consumers will all in general differ, leading to different choices of the at most two goods that each demands, so that each of the five goods will have a positive demand in the market as a whole if the tastes for characteristics are sufficiently diverse. Adding together the amounts of each good demanded by all the consumers, we obtain a point on the market demand function for that good at the given price vector *p*. Repeating the analysis described for all possible price vectors, we can generate the market demand functions for all five goods by the Lancaster method and then proceed as usual.

The power of this alternative approach is perhaps best revealed by the problem of new goods. In the standard theory we would have to recast the entire utility function *U* as a function of six instead of five arguments in our example, with almost no restrictions capable of being placed on the properties of the new function in comparison with the old. In the Lancaster model, however, the utility function in characteristics space $U(z)$ is entirely unaffected by the introduction of the new good. Given its price the new good will appear in the budget constraint with an additional sixth term and in the matrix *B* as an additional sixth column, leading to a sixth vertex and a fifth 'flat' for the new CPF. The new good thus leads only to a change in the CPF, which is common for all consumers, with all individual utility functions in the space of characteristics unchanged, instead of each having to be altered in its own particular way in the space of goods. By looking at the CPF we can see exactly which consumers will be affected and which not by the introduction of the new good. If we consider the cases of electric light and candles, automobiles and the horse and buggy, compact discs and vinyl LP records, it is clear that the new goods altered the technology of consumption for all consumers by providing a 'dominant' new good that drove out the competing old one on efficiency grounds, rather than leading to a simultaneous subjective shift in tastes by all consumers. Though developed for the analysis of consumer demand, the characteristics approach is also clearly applicable to portfolio selection between alternative financial assets, occupational choice problems in labour economics, provision of public goods and services (see Lancaster, 1991, Part 3) and many other areas.

The characteristics approach also led Lancaster naturally to the problem of 'socially optimal product differentiation' that he investigated initially in an article, Lancaster (1975), and with considerably more depth and detail in the 1979 major treatise entitled *Variety, Equity and Efficiency*. To explain the essentials of this problem, consider once again the concept of the CPF introduced earlier. Suppose that we have a unit of 'resources', which can be used to produce many alternative goods, each yielding as before a set of characteristics. As the number of potential goods gets increasingly large we can think of the CPF in two dimensions as a continuous curve concave to the origin in characteristics space, like the familiar transformation curve in goods space. Which of these infinitely many alternative goods would be the one most preferred by a particular consumer, given his utility function $U(z)$ over the two characteristics? This 'most preferred good' or MPG would obviously be defined by the point of tangency between the CPF and the highest attainable indifference curve, with the slope of a ray from the origin to the optimal point indicating the ratio of the two characteristics provided by the MPG. Other consumers with different tastes would have different MPGs. What should a social planner do if he wants to attain the objective of putting each consumer at a specified utility level with the minimum use of overall resources? In particular, how many and which goods should be produced? With constant returns to scale it is clear that each consumer should be provided with his MPG, in whatever amount is needed

to place him at the desired welfare level. With increasing returns to scale, however, we have to trade off the provision of more variety against the sacrifice of less economies of scale. Most of Lancaster (1979) is devoted to a deep and subtle analysis of this fundamental problem under a wide range of alternative technological possibilities, market structures and compensation schemes for the attainment of equitable outcomes, with both first and second-best optima considered. This book, Lancaster's magnum opus, is undoubtedly a major landmark of economic theory that will continue to be an inspiration to the profession for decades to come.

The theory of international trade was another major area that attracted Lancaster's attention and benefited greatly from his application of these novel ideas to it. Early papers on the Heckscher–Ohlin model and the Stolper–Samuelson theorem (see Lancaster, 1996, chs 6 and 7) were followed by a pioneering paper (1980) on ‘Intra-industry Trade under Perfect Monopolistic Competition’, that together with and independently of Paul Krugman (1979; 1980), who was inspired by Dixit and Stiglitz (1977), launched what came to be known as the ‘new trade theory’, supplementing the standard Ricardian and Heckscher–Ohlin models of perfect competition with models involving economies of scale, differentiated products and monopolistic competition. Unlike the standard models it was easy to show that even identical economies could gain from trade and specialization by providing more variety for consumers in both countries and at lower prices for each differentiated product. The use of the convenient but highly restrictive Dixit–Stiglitz ‘love of variety’ utility function enabled Krugman to obtain this key result more easily and compactly than the more general framework used by Lancaster; but the latter offers additional insights not available in the former. Later papers considered tariff protection and monopoly policy in open economies in the context of the new trade theory (see Lancaster, 1996, Part 1).

At Columbia Lancaster regularly taught in the graduate theory sequence, a course built around his *Mathematical Economics*, an early advanced text published in 1968 the success of which around the world is attested by its translation into Spanish, Japanese, Russian and Rumanian. He also taught a popular undergraduate seminar with the noted philosopher Sidney Morgenbesser. He twice served as chairman of the Economics Department, first from 1973 to 1976 and then from 1989 to 1990. He was elected a Fellow of the Econometric Society, a Distinguished Fellow of the American Economic Association and a Fellow of the American Academy of Arts and Sciences. His death of cancer on 23 July 1999 deprived his university, colleagues, friends and family of a deeply original thinker and a wonderfully warm and compassionate human being. He is survived by his wife Dvora, sons Cliff and Gil, as well as by five grandchildren.

See Also

- demand theory
- product differentiation
- second best
- welfare economics

Selected works

1956. (With R. G. Lipsey.) The general theory of second best. *Review of Economic Studies* 24, 11–32.

1966. A new approach to consumer theory. *Journal of Political Economy* 74, 132–57.

1968. *Mathematical Economics*. New York: Dover, 1987.

1971. *Consumer Demand: A New Approach*. New York: Columbia University Press.

1975. Socially optimal product differentiation. *American Economic Review* 65, 567–85.

1979. *Variety, Equity and Efficiency*. New York: Columbia University Press.

1980. Intra-industry trade under perfect monopolistic competition. *Journal of International Economics* 10, 151–75.

1991. *Modern Consumer Theory*. Aldershot: Edward Elgar.

1996. *Trade, Markets and Welfare*. Cheltenham: Edward Elgar.

Bibliography

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product variety. *American Economic Review* 67, 297–308.

Krugman, P. 1979. Increasing returns, monopolistic competition and international trade. *Journal of International Economics* 9, 858–64.

Krugman, P. 1980. Scale economies, product differentiation and the pattern of trade. *American Economic Review* 70, 151–75.

How to cite this article

Findlay, Ronald. "Lancaster, Kelvin John (1924–1999)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000230> doi:10.1057/9780230226203.0926

close print preview

The New Palgrave Dictionary of Economics Online

land markets

Klaus W. Deininger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

While earlier research highlighted the potential market inefficiencies that can result from the particular characteristics of land, recent empirical evidence suggests that these may be small, not always amenable to policy intervention, and outweighed by the contribution of land markets to broader structural transformations, like population movements out of agriculture. High levels of transaction costs pose, however, still considerable obstacles to land market operation, suggesting that measures to reduce them through greater security and formalization of property rights, a streamlined regulatory framework, and ready availability of information may significantly improve functioning of, and enhance benefits from, land markets.

Keywords

access to land; agricultural economics; banking crises; fixed-rent contracts; land markets; land reform; land registries; land tax; land use rights; land use regulation; property rights; rent control; repeated games; sharecropping; structural change

Article

Land rental markets

In a world of perfect information, complete markets and zero transaction costs, the distribution of land ownership will affect welfare but will not matter for efficiency as everyone will operate his or her optimum farm size. However, in most empirical settings, the productivity of land use, and thus the impact of market-mediated transfers of land, will be affected by technology, producers' ability, potential scale (dis)economies of agricultural production, risk, and imperfections in labour and credit markets. The range of possible contracts will, furthermore, depend on potential tenants' endowments, their reservation utility, and the transaction costs associated with transferring land. Key questions include whether, with a given ownership distribution of land, rental markets will achieve socially desirable

outcomes, and which factors will enable participants to attain outcomes closer to the optimum.

By varying the share and a fixed payment to the tenant, landowners who wish to rent can achieve any combination of contractual forms, from a wage labour contract or a share contract to a fixed-rent contract. While all contracts will lead to equivalent outcomes if output is certain and tenants' effort can be enforced (Cheung, 1969), relaxation of this assumption gives way to a number of scenarios.

If effort cannot be monitored and agents are risk neutral, only the fixed-rent contract is optimal. The reason is that, in all other cases, equalizing the marginal disutility of effort to their marginal benefit will lead tenants to exert less than the socially optimal amount of effort, thus resulting in lower total production. The optimum outcome will require a trade-off between the risk-reducing properties of the fixed-wage contract, under which the tenant's residual risk is zero, and the incentive effects of the fixed-rent contract, which would result in optimal effort supply but no insurance. Limited tenant wealth has a similar effect because in case of a negative shock tenants with insufficient wealth are likely to default on rent payments. This implies that landlords will tend to enter into fixed-rent contracts only with tenants who are wealthy enough to pay the rent under all possible output realizations, implying that poorer tenants will be offered only a share contract (Shetty 1988). Finally, a dynamic setting opens up a number of additional perspectives, in addition to the scope for using the repeated game context and the threat of eviction to reduce the efficiency losses of sharecropping. A rental contract that provides tenants with adequate incentives to maximize production in any given time period may lead to over-exploitation of the land if (dis)investment is considered, implying that a share contract with lower-powered incentives and possibly compensation may be more appropriate (Ray, 2005).

A large literature has focused on testing the extent of inefficiency of sharecropping contracts, although often with mixed results and inappropriate methods (Otsuka and Hayami, 1988). Use of within-household variation suggests that, in India, share tenancy is associated with an average loss of productivity of 16 per cent (Shaban, 1987) although part of the losses may have been policy-induced. More recent studies fail to find support for inefficiency of sharecropping (Pender and Fafchamps, 2006), suggesting that agents' choice of contractual arrangements is rational given the constraints faced in a given situation and that the scope for government to bring about more effective outcomes may be limited.

While potential inefficiencies, if they exist at all, will thus be modest, productivity gains from land rental can be large. Analysis of the same plot before and after being rented in China points towards productivity gains of some 80 per cent, leading to a significant increase in welfare of tenants as well as landlords, in addition to helping the latter to migrate and gain access to non-agricultural income (Deininger and Jin, 2006). Although less direct, empirical analysis of determinants for rental market participation in a large number of countries suggests that the ability of those renting in is generally higher than that of those renting out (Deininger, 2003), implying a positive productivity impact of land rental which, at least in the case of China, is much superior to what is achieved by a social planner (Deininger and Jin, 2005).

The potentially important contribution of land rental to structural change is also illustrated by the fact that rental markets equalize the distribution of per capita operated land area and transfer land to those with lower levels of assets but higher levels of education, and that rental activity increases in settings where wage rates and thus non-agricultural opportunities are higher. Land rental is widespread in developing economies; 71 per cent of farmland is rented in Belgium, and 48 per cent, 47 per cent, and

43 per cent respectively in the Netherlands, France, and the United States (Swinnen and Vranken, 2006). Rental markets can emerge rapidly; for example, in Vietnam the share of participants in land rental increased from 3.8 per cent of rural households in 1992 to 15.8 per cent in 1998. They were also of great importance in the countries of eastern Europe and the former Soviet Union during the initial phases of economic transition, especially where radical individualization of land was pursued, such as in Albania and Moldova. As long as transaction costs arising from fragmentation were not too high, rental was critical where land had been restored to original owners, many of whom had little intention to use it but also did not want to part with their asset. In West Africa, long-term sharing arrangements did historically provide important incentives for long-term investment and, even though increased population density has shifted contractual parameters in favour of landlords, rental continues to be important in providing land access and increasing productivity.

Of course, a high incidence of rental transactions, and the fact that observed transactions had a positive impact, do not imply that the level of rental activity is optimal. Qualitative and quantitative evidence points towards considerable rationing in rental markets. For example in India, farmers are able to realize only about 75 per cent of their desired level of land transactions (Skoufias, 1995), implying that transaction costs or land rental remain high. Two key factors contributing to these are limited security of property rights, which makes renting out too risky, and implicit or explicit restrictions on rental markets in the form of either rent ceilings or the award of property rights to tenants.

Even if it leads to only a small decrease of the probability that landlords who rent out their land will get it back upon termination of the contract, insecurity of property rights can significantly reduce the supply of land to the rental market. This is confirmed by econometric studies in countries as diverse as the Dominican Republic, Nicaragua, China, Ethiopia, Vietnam, and Bulgaria. While insecure tenure may not prevent landlords from renting out completely, it often prompts them to rent only to close kin, where enforcement is easier even if, due to the limited pool of renters to choose from, productivity will be lower than from renting to outsiders, as is indeed observed in the case of Vietnam (Deininger and Jin, 2007).

Beyond tenure insecurity, rent ceilings or regulations that aim to confer de facto property rights on tenants by preventing landlords from evicting them and giving heritable use rights to tenants after a certain period of time are a frequent source of inefficiency in rental markets. Although the original intent was to improve equity, such measures led in many cases to self-cultivation by landlords or the adoption of wage labour contracts, both modes of production that are inferior to tenancy in terms of production incentives and outcomes. Analysis shows that, while rent controls can transfer resources to sitting tenants, they tend to make those who are not lucky enough to already sit on tenanted land worse off by restricting the supply of land available to the rental market, undermining tenure security, and reducing investment (Basu and Emerson, 2000). Similarly, conferring heritable (but often non-transferable) use rights on tenants, subject to the requirement that they continue paying rent, can increase welfare in the short term but will – in the medium to long term – reduce investment incentives and supply of land to rental markets in a way that is particularly detrimental to the poor and landless, as in the case of India (Deininger, Jin and Nagarajan, 2006) where such legislation has driven a large number of contracts into informality.

Land sales markets

Land sales markets provide an opportunity to obtain land for permanent use which will be associated with higher investment incentives than renting. In addition, markets for land sales are a precondition for using land as collateral in credit markets. If all markets were perfect, the sale price of land would equal the net present value of the stream of profits that can be derived from a given land use, and potential buyers would be indifferent between renting land and purchasing it. However, land sales markets will be affected by a number of factors that include (*a*) the ability to use land as a collateral in credit markets and thus overcome credit constraints; (*b*) expectations about future increases in land values due to infrastructure construction or population growth; (*c*) the risk–return profile and liquidity implications of holding land as compared with other assets; and (*d*) the level of transaction costs in land sales markets. In economies where risk is high, land is important as a store of wealth, and access to outside credit is limited, land prices can fluctuate significantly over time (Zimmerman and Carter, 1999). The reason is that, because returns from agricultural production are highly covariate, demand for land, and therefore land prices, will be high in good crop years when savings are high, sellers are few, and potential buyers of land are many. At the same time, households' need to satisfy basic subsistence needs can give rise to a large supply of land by people who are forced to engage in distress sales of their land in bad years, often to individuals with incomes or assets from outside the local rural economy (Cain, 1981). Such distress sales rarely enhance productivity, and improved functioning of markets for insurance and credit to avoid them will be important.

If covariance of asset prices is observed, those who sell off land during crises will not be able to repurchase it during subsequent periods of recovery, creating a potential for successive decline of asset endowments (Zimmerman and Carter, 2003). In high-risk environments this may lead the poor to prefer assets with a lower but more stable returns to land even if they had access to credit, implying that, in situations where land is very unequally distributed as in Latin America, land sales markets will not be a good way to achieve asset redistribution, and other measures, such as grants, may be needed to increase land access by the poor on a broader scale.

With macroeconomic instability, an expectation of future land price increases, or lack of sufficiently attractive alternative assets, land may be acquired for speculative rather than productive purposes. For example, inflation and changes in real returns on alternative uses of capital were shown to be key factors explaining changes in land prices in the United States. In eastern European countries, the expectation of large capital inflows due to EU accession was a major factor underlying real estate booms that propelled land prices far beyond the net present value of the flow of services that could be derived from the land. Credit or tax preferences, together with weak regulatory oversight, can reinforce such trends which, in the extreme, can lead to bank crises with far-reaching consequences.

Although empirical study of the functioning of land sales markets is more limited than for rental markets, evidence from India over the 1982–99 period supports the notion that distress sales are important, but that options to insure against risk – for example, the presence of safety net programmes or access to bank branches – helped to reduce or eliminate the adverse impact of climatic shocks. Moreover, there is little evidence of a negative impact of land sales markets on productivity or of speculative land accumulation, partly because of land ownership ceilings and partly because of increased availability of other stores of wealth. Although the number of landless who were able to purchase land remained modest, land sales markets constituted the most important avenue to access land by the poor (Nagarajan, Deininger and Jin, 2007).

Well-intended land sales restrictions in a number of countries failed to prevent distress sales but instead drove them into informality. Safety nets and measures to increase access to savings and insurance may be more effective to prevent socially undesirable land loss by the poor. One possible exception is in the transition from customary to more individualized forms of tenure whereby the potential for opportunistic behaviour and land sales by local chiefs is high. To counter this risk, a decision at the local level to maintain a customary land tenure regime that outlaws land transfers outside the community, similar to what was done in the Mexican *ejido* reforms (World Bank, 2002), may be an appropriate second-best solution (Andolfatto, 2002). As long as it results from a conscious choice and there are transparent mechanisms for changing the tenure regime, such a rule is unlikely to be harmful because, once potential advantages exceed the cost at the local level, communities are likely to change the rules to allow sales.

Policy options to improve the functioning of land markets

Land registries to make information on property rights available publicly in a cost-effective way have many advantages. They reduce the risk of land loss by landlords renting out, and provide the basis for credit market transactions. While informal rights can provide security within a well-defined and socially cohesive group, they preclude trade and exchange beyond this realm. Once gains from transactions with outsiders became sufficiently high, informal rights are likely to be replaced by formalized property right systems and associated enforcement institutions, leading eventually to abstract representation and the impersonal exchange of rights that allows the emergence of more abstract instruments such as mortgages based on the existing rights system (de Soto, 2000). Making information on private as well as public land ownership widely available would also reduce the potential for opportunistic behaviour and appropriation of public land by powerful interests as resource values rise.

While disputes among private parties can limit the propensity to rent out land, threats of expropriation without (or with only very limited and delayed) compensation and for a very broadly defined public purpose, which in many countries includes transfer of land to private investors, will limit incentives for investment and can prompt informal pre-emptive land transactions at very low prices that improve neither efficiency nor equity. To prevent this, it is critical to have a restrictive definition of public interest, to ensure compensation at market values if expropriation is unavoidable. If for political reasons ceilings cannot be abandoned altogether, they should be limited to preventing speculative land accumulation. Similarly, land use regulations should be used only if needed to avoid undesirable externalities and if capacity for cost-effective implementation is available.

As public investment in infrastructure and other amenities will be capitalized in land values, taxing land comes close to a benefit tax, and is less distorting than taxes on sales or income. It has thus been considered to be an ideal revenue source for local governments. Land taxes that effectively tax resource rents, that is, that are based on the normal potential yield from a certain plot, will discourage speculation and encourage land owners who are not able to make the most efficient use of their land to rent it out to others. Local land taxes are used effectively in the United States where they have been shown to induce land development. Although underexploited in the past, their potential to intensify land use – which is greater than that of other instruments – has provided a motivation for reforms in a number of countries (Bird, 2004).

The often limited ability of the poor to access land through purchase implies that market forces may be

unable to correct highly unequal and often inefficient distribution of land, thereby moving the economy towards an equilibrium with a more equal distribution of opportunities and higher overall output. Land reforms in Asia, such as in Japan, Korea, and Taiwan (China), or the abolition of intermediaries in India, and some of the immediate post-independence efforts in Africa – all of which were accomplished under external pressure or immediately after independence – illustrate that land reform can improve household well-being and productive efficiency. At the same time, in many other countries, including virtually all of Latin America, success often remained elusive because, among other things, such measures were guided by short-term political objectives, insufficient effort was devoted to ensuring access to complementary inputs and the competitiveness of producers, and the mechanisms adopted to implement land reform, like ceilings or rent controls, often undermined the functioning of land markets, thus limiting the potential for synergies. Together with multiple restrictions on beneficiaries' ability to transfer the land received, this often limited the scope for land reforms to bring about sustained improvement in beneficiaries' living conditions.

In countries where an unequal distribution of land or incomplete past reforms imply that land reform remains on the agenda, there is broad agreement on a number of common principles (Deininger, 2003). These include (*a*) the need to have programmes integrated into a broader development strategy that includes training and capacity building, as well as provisions for complementary investment to make the land productive so as to help put households on a viable trajectory of development; (b) a design based on clear and transparent rules that aims to maximize productivity gains; (*c*) a multiplicity of paths to land access needed to underpin land reform, including, in addition to state-sponsored land transfers, progressive land taxation to increase the supply of underutilized land, divestiture of suitable state land, foreclosure of mortgaged land, and rental and sales markets; (*d*) secure and unconditional rights for beneficiaries, including the right to rent or sell their land, perhaps after some initial period; and (*e*) an undistorted policy environment supportive of smallholder agriculture, decentralized implementation, and respect for the rule of law, in particular existing property rights.

See Also

- access to land and development
- agricultural markets in developing countries
- common property resources
- credit rationing

Bibliography

Andolfatto, D. 2002. A theory of inalienable property rights. *Journal of Political Economy* 110, 382–93.

Basu, K. and Emerson, P.M. 2000. The economics of tenancy rent control. *Economic Journal* 110, 939–62.

Bird, R.M. 2004. *International Handbook of Land and Property Taxation*. Cheltenham, UK and

Northampton, MA: Edward Elgar.

Cain, M. 1981. Risk and insurance: perspectives on fertility and agrarian change in India and Bangladesh. *Population and Development Review* 7, 435–74.

Cheung, S.N. 1969. Transaction costs, risk aversion, and the choice of contractual arrangements. *Journal of Law and Economics* 12, 23–42.

de Soto, H. 2000. *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. New York: Basic Books.

Deininger, K. 2003. *Land Policies for Growth and Poverty Reduction. A World Bank Policy Research Report*. New York and Oxford: World Bank and Oxford University Press.

Deininger, K. and Jin, S. 2005. The potential of land markets in the process of economic development: evidence from China. *Journal of Development Economics* 78, 241–70.

Deininger, K. and Jin, S. 2006. Productivity and equity effects of rental markets: evidence from China. Policy Research Working Paper, World Bank.

Deininger, K. and Jin, S. 2007. Does greater tenure security allow more efficiency-enhancing land transactions? Evidence from Vietnam. Policy Research Working Paper, World Bank.

Deininger, K., Jin, S. and Nagarajan, H.K. 2006. Equity and efficiency impacts of rural land market restrictions: Evidence from India. Policy Research Working Paper, World Bank.

Nagarajan, H.K., Deininger, K. and Jin, S. 2007. Market vs. non-market sales transactions in India: evidence over a 20-year period. *Economic and Political Weekly*.

Otsuka, K. and Hayami, Y. 1988. Theories of share tenancy: a critical survey. *Economic Development and Cultural Change* 37(1), 31–68.

Pender, J. and Fafchamps, M. 2006. Land lease markets and agricultural efficiency in Ethiopia. *Journal of African Economies* 15, 251–84.

Ray, T. 2005. Sharecropping, land exploitation and land-improving investments. *Japanese Economic Review* 56(2), 127–43.

Shaban, R.A. 1987. Testing between competing models of sharecropping. *Journal of Political Economy* 95, 893–920.

Shetty, S. 1988. Limited liability, wealth differences and tenancy contracts in agrarian economies. *Journal of Development Economics* 29, 1–22.

Skoufias, E. 1995. Household resources, transaction costs, and adjustment through land tenancy. *Land Economics* 71, 42–56.

Swinnen, J.F.M. and Vranken, L. 2006. Patterns of land market development in transition. Policy Research Working Paper, World Bank.

World Bank. 2002. *Mexico – Land Policy A Decade After the Ejido Reforms*. Washington, DC: Rural Development and Natural Resources Sector Unit, World Bank.

Zimmerman, F.J. and Carter, M.R. 1999. A dynamic option value for institutional change: marketable property rights in the Sahel. *American Journal of Agricultural Economics* 81, 467–78.

Zimmerman, F.J. and Carter, M.R. 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. *Journal of Development Economics* 71, 233–60.

How to cite this article

Deininger, Klaus W. "land markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000225> doi:10.1057/9780230226203.0927

The New Palgrave Dictionary of Economics Online

land tax

Mason Gaffney
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Clark, C. G.; double taxation; Ely, R. T.; excess burden of taxation; George. H.; land tax; neutral taxation; progressive and regressive taxation; rent; stewardship

Article

Taxation is the form of socialization used in market economies. Choosing what to tax is choosing what to socialize. Rather than socialize labour or repel capital it is possible to tax land.

Land holds a unique place in the distributional ethic because it is (by definition) of natural origin. Man did not create Earth with its resources but rather fights over it. Land is also (with exceptions) more nearly permanent than man or his works. Thus, rent as private income neither elicits the supply nor preserves it. Its main function is to allocate the fixed supply among uses, but it is arguable that land taxes, when based on land's capacity-to-serve, are at worst neutral to this function and at best improve on it.

The philosophical rationale for land taxes is strongest under an organic theory of the polity. It is no accident that Henry George (prominent protagonist of land taxes) crystallized his ideas after reading Andrew Bisset's *Strength of Nations* on feudal levies. Landholders have a privilege from the state and in return are liable for taxes in perpetuity.

The entire value of land, now and for ever, is here regarded as a benefit received from government. This is consistent with Alfred Marshall's concept, ‘the public value of land’, where value is the product of three things: nature; government; and spillover values from development of adjoining and linked lands. All these values, being unearned by the individual landholder, are fit to be taxed.

The organic view distinguishes the land from its holder. Land taxes may be paid by income the land earns, not by the holder as a person unless we identify him with the land and regard him as having a prior right to own land free of liabilities to the public from which he holds title. The contractual theory, by contrast, treats government as a kind of business, extending services to specific lands whose holders need pay only for recent benefits received, construed narrowly.

The rationale for land taxes presumes a functional attitude toward distribution, regarding property not as

an end in itself but a means to get things done. A land tax based on market value, not varying with actual use, is a fixed cost that sharpens marginal incentives. Critics today seldom argue otherwise, but oppose land taxes precisely because they do force landholders to respond to the market, which may have its own faults in a world of ‘second-best’.

Land taxes are *in rem* and so disregard the holder's personal circumstances, a drawback in some opinions. On the other hand landholdings are much more concentrated than the receipt of income or taxable consumption or payrolls, and land taxes are not shifted, making the tax inherently progressive even though but loosely correlated with taxable income. Avoiding land taxes is next to impossible, even though collection enforcement is limited to seizing the land, not the person or any other asset.

The rationale includes a concept of landholder stewardship. A limited number of land titles were issued in order to get land under tenure to assure best use. So far so good, but those not receiving or inheriting land need a counterpoise to assure they receive their share. Land taxes do so in three ways: by supporting government; by pressing landholders to produce goods and services; and pressing them to hire workers to do so. Land taxes act as a kind of social audit and performance standard of stewardship to promote equity towards those excluded.

There is also more equity among landholders, which in turn promotes efficiency. Absent land taxes there is pressure on government to do as much for A's land as for B's. Efficiency, however, calls for specialization and differentiation, meaning high values for some land and low values for other, with windfalls and wipeouts. Land taxes automatically compensate the losers from the gains of the winners, thus freeing land planners to maximize the joint benefits.

The rationale of equity for the excluded says that lands with open general access like parks and roadways should be exempted in whole or in part. But such exemption can lead to overcrowding, to meet which it is clear that some user charges on such land can be construed as special kinds of land taxes. An obvious example is a charge on large trucks in downtown streets. Lacking any such constraint the crowding might in turn lead to indefinite expansion of the exempt land use.

The rationale is only partly consonant with personal ability to pay. Landholding confers potential ability to pay, but that is only realized upon one's using the land well. And earned cash is not tapped at all. A land tax is a fixed periodic charge. It is based on qualities inherent in the land with few concessions to the landholder's personal illiquidity, weakness, setbacks or ageing. ‘Use it or sell it’ is the message, which many consider too harsh.

What is harsh for the distressed holder, however, is accommodating to frustrated buyers, and it boils down to which group shall be accommodated. Since liquidity is known not to increase in step with total wealth, imposing taxes on landed but illiquid holders has a strong progressive effect. The regular flow of land taxes also accommodates governments, especially small local ones needing steady revenues that are not turned on and off at the convenience of others.

It is not always a question of selling complete units. Land around homes and enterprises is subject to sharply diminishing marginal utility or productivity and a function of land taxes is to constrain horizontal extension of holdings, to the end that the nucleus of each holding may be closer to others to facilitate trade, cooperation, linkages, sharing common costs, and other synergies. The ‘highest and best use’ of land is usually that which most relates to and complements its neighbours and trading partners, who must not be held too far distant.

There is also a diminishing return to time as buildings age, and a function of land taxes, in conjunction with building exemption, is to advance (and/or stop retarding) renewal of sites, neighbourhoods, cities,

regions and whole economies.

Locke, Quesnay, Adam Smith and others have shown a tendency to shift all taxes to land, whatever the nominal base or event, assuming elastic supplies of labour and capital. This leads some to conclude that all taxes alike just tap land rent. But one cannot tap rent where there is none. Taxes on other bases simply abort the taxed input or activity at the no-rent margins of land use, both extensive and intensive. This excess burden in turn puts an upper limit on the possible tax rate, thus sparing much rent from being taxed at all while destroying other rent completely. The only way to tap much rent is to tax land directly.

Land value and capital are not convertible into one another (excepting exhaustible minerals, not treated here). From this it follows that efficiency does not require equal tax rates on the two, but only uniformity within each class. Uniformity is impossible with capital because of differential concealability. But land is uniformly non-concealable. The case for neutrality of land taxes is stronger under uniformity, but mainly requires that the tax not be a function of use.

A land tax may be based on the current potential rent, or on value. In practice, it is the latter. Values are not simply proportional to rents because many land values are elevated above that by expected higher future rents. In such cases taxes rise high relative to cash flow, and at a stiff rate may even be higher. This subjects the holders to a cash drain. The extra tax may be shown, however, in general to tax the unrealized increment, in the manner advocated by Haig–Simons, at the time it accrues. There is some recent falling-away from Haig–Simons, and to one school now this is ‘double taxation’, an issue currently mooted.

The most controversial question in land taxation is the effect on appreciating land. Most hands agree the land tax advances conversion to the higher use. To Henry George this ‘sovereign remedy’ would correct a market failure and unlock speculative holdings with profoundly beneficial effects. To several modern writers following Richard T. Ely the advance of conversion is unneutral and somewhat wasteful. Speculation is seen as efficiently keeping land from premature commitments. To this writer it seems mathematically obvious that an efficient adaptation to rising future incomes would result in advancing, not retarding conversion. But the issue is now moot.

Land taxation at the local level has a natural cap in local particularism as expressed in ‘Don't swamp the lifeboat’. Land taxation by a central national government might go much heavier, and accordingly statesmen like Austen Chamberlain in Britain and James Madison in America have contrived to divert land taxation to local governments. Colin Clark, on the other hand, published a plan to nationalize land through taxation without depriving the poorer localities. He would rank the local jurisdictions in order of land value per capita, and apply a central government surtax starting from zero but graduated upwards according to this ratio. The scheme basically had central government apply to local ones the same principle of direct land taxation that local governments can apply to individuals, tapping the rich rents without destroying marginal rents. Clark, like George, may have been reading Bisset's *Strength of Nations*.

See Also

- George, Henry

How to cite this article

Gaffney, Mason. "land tax." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000020> doi:10.1057/9780230226203.0928

close print preview

The New Palgrave Dictionary of Economics Online

Landry, Adolphe (1874 – 1956)

E. Malinvaud

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Böhm-Bawerk, E. von; fertility; income distribution theory; inequality; Landry, A.; marginal productivity of labour; mortality; optimal economic growth; optimal population size; production process; risk aversion; uncertainty; unemployment; value theory

Article

Born on 29 September 1874 in Ajaccio, Corsica, Landry died in Paris on 28 August 1956. A graduate of the Ecole Normale Supérieure, he began as a philosopher, then turned to economics and demography. From 1907 on he held the chair of economic history and history of economics at the Ecole Pratique des Hautes Etudes, Paris. He was elected as a Deputy for Corsica in 1910 for the Radical Socialist party, serving as Minister of Navy in 1920, of Public Instruction in 1924, and of Labour in 1932. As a member of Parliament he was particularly effective in promoting family legislation and family allowances which, intended to stimulate fertility, became quite substantial (subsidies to large families in 1913, the Code de la famille in 1939, and the law on family allowances in 1946).

Very early in his study of economics, Landry revealed himself as a gifted theoretician. His approach was purely literary but analytical and rigorous. He was able to master fully technical arguments and, for instance, early exposed in France the definition and relevance of the new demographic indicators proposed by Lotka and Kuczynski. His culture was quite broad and up to date. He was an explicit proponent of the deductive methodology.

His initial concern was with the theory of income distribution. In his dissertation (1901), which made him known as a socialist, he argued that individual ownership and the subsequent unequal distribution of property rights could not be considered as socially optimal and was responsible for a smaller national output than was feasible. His 1904 book was an excellent presentation of the theory of interest in continuation of Böhm-Bawerk, showing why interest was just an aspect of the general theory of value, paying particular attention to the productivity of capital and criticizing Böhm-Bawerk for his overemphasis on the length of the production process. His two articles on the theory of pure profits (1908b and 1938) discussed the role of uncertainty, the idea of risk aversion being already explicit in 1908. However, Landry refrained from making this the first determinant, arguing that this was rather the

scarcity of entrepreneurs, who must simultaneously have capital, abilities and will.
Also interesting are his two long articles. Starting in 1910 from a discussion introduced in 1755 by Cantillon on the demographic impact of a change in landlords' consumption behaviour, Landry finally explains why the returns to primary factors indeed vary with exogenous shifts of individual preferences. Discussing unemployment in 1935, he explains that it reveals an excess of the wage rate over the marginal productivity of labour but is mainly due to a depression of this productivity and can be cured by measures that will raise it again.
From his first writings, Landry always paid attention to population, which later became his main concern. His 1909 article introduced the distinction between three demographic regimes, population being regulated by mortality and the minimum of subsistence in the first case, by fertility behaviour and the wish to achieve some standard of living in the other two, but whereas in the 18th century the objective was a stationary standard of living, it shifted to a permanently progressive one in the late 19th century, 'social capillarity' making this progress feasible for everybody's children. His 1929 article on the optimal size of the population is interesting since it introduces an objective function that was also preferred in the theory of optimal economic growth in the 1960s: a sum of annual terms in which each term is the product of population size and a utility of average consumption per person. His main thesis, developed in his 1934 book, was that a decreasing population leads to decadence, this thesis being substantiated by a study of Ancient Greece and of the cultural centres of the Roman empire.

Selected works

1901. *L'utilité sociale de la propriété individuelle*. Paris: G. Bellais.

1904. *L'intérêt du capital*. Paris: Giard.

1908a. *Manuel d'économique à l'usage des Facultés de Droit*. Paris: Giard.

1908b. Le problème du profit. *Revue d'économie politique*, January, 24–62.

1909. Les trois théories de la population. *Scientia*. Reprinted in (1934).

1910. Une théorie négligée. De l'influence de la direction de la demande sur la productivité du travail, les salaires, la population. *Revue d'économie politique* 24, 314–28, 364–84, 747–57, 773–85.

1929. Le maximum et l'optimum de population. *Scientia*. Reprinted in (1934).

1934. *La révolution démographique*. Paris: Sirey.

1935–6. Réflexions sur les théories du salaire et du chômage. *Revue d'économie politique*, November–December 1935, 1652–90; March 1936, 327–57.

1938. Sur la théorie du profit. *Revue d'économie politique*, November–December, 1473–504.

1945. *Traité de démographie*. (In collaboration with H. Bunle, P. Depoid and A. Sauvy.) Paris: Payot.

A fuller bibliography, as well as biographical details appeared in: A. Sauvy, Adolphe Landry. *Population* 11 (1956), 609–20.

How to cite this article

Malinvaud, E. "Landry, Adolphe (1874–1956)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000023> doi:10.1057/9780230226203.0929

The New Palgrave Dictionary of Economics Online

Lange, Oskar Ryszard (1904 – 1965)

Tadeusz Kowalik

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Oskar Lange was a well-known economist, socialist thinker and politician. His special position in economics rested on his profound knowledge of its main currents, of both Marxist economics and Western academic economics (above all the neoclassical) and later of both capitalist and centrally planned eastern European economies. With Abba P. Lerner he was one of the founders of the theory of market socialism. This induced him to make several attempts at a ‘major synthesis’ and to undertake political actions aiming for a rapprochement between the West and the Communist world, for peaceful coexistence, economic cooperation and systemic convergence.

Keywords

accounting prices; Allen, R. D. G.; Breit, M.; Brus, W.; business cycle; cash balances, demand for; command economy; concentration; corporations; democracy; depreciation; econometrics; economic calculation in socialist economies; efficient allocation; forecasting; general equilibrium; Hicks, J. R.; industrialization; innovation; interest, theory of; interpersonal utility comparisons; inventories; Kalecki, M.; Keynesian Revolution; Knight, F. H.; Lange, O. R.; Lange–Lerner mechanism; •aski, K.; Leontief, W.; Lerner, A. P.; liquidity preference; loanable funds; marginal analysis; marginal cost pricing; marginal efficiency of capital; market socialism; Marx, K. H.; Miller, D.; mixed economy; money; money supply; monopoly capitalism; neutrality of money; optimal resource allocation; planning; price flexibility; propensity to consume; public ownership; public trusts; Roemer, J. E.; Samuelson, P. A.; Say's Law; Schultz, H.; Schumpeter, J. A.; Stiglitz, J.; Sweezy, P.M.; wage differentials; Yunker, J. A.

Article

Lange was born on 27 July 1904 in Tomaszow Mazowiecki, near Lodz, Poland, into the family of a German-born, assimilated textile manufacturer, and died on 2 October 1965 in a London hospital following thigh surgery. He studied law and economics in Poznan and Cracow. His main tutor was

Adam Krzyzanowski, liberal and Anglophile. In 1929, Lange studied in London and in 1934–5 in the United States, mostly at Harvard and Berkeley. He lectured in statistics and economics in Cracow (1927–37), Chicago (1938–45) and Warsaw (1948–65). Politically involved since his youth, he was active at the Independent Socialist Youth Union in the interwar period. During the Second World War he pushed the cause of Soviet–American rapprochement and socialist–communist cooperation. He served as the first ambassador of the Polish People's Republic in Washington (1945–6) and as the Polish delegate to the UN Security Council (1946–7). Later he was a member of parliament and a member of the State Council in Poland.

Lange's special position in economic theory rested on his profound knowledge of its main currents, of both Marxist economics and Western academic economics (above all the neoclassical) and later of both capitalist and centrally planned Eastern European socialist economies. This induced him to make several attempts at a ‘major synthesis’ and to undertake political actions for a rapprochement between the West and the Communist world, for peaceful coexistence and economic cooperation.

Capitalism and economics

The capitalist economy was Lange's chief research concern from his early youth until the end of the Second World War. His primary interests included the study of business cycles and the evolution of capitalism. His Ph.D. thesis was a study of business cycles in the Polish economy 1923–7 (1928a), and won the title of docent (assistant professor) for a statistical study of the business cycle (1931a). These were among the chief topics of his lectures at US universities, mainly in Chicago. Early in the war he studied, together with L. Hurwicz, ways of empirical verification of business cycle theories. Although he became a leading authority on this subject (see his review, 1941a, of Schumpeter's book and, 1941b, on Kalecki's cycle theory), he never produced a complete theory of his own. His studies of the business cycle led him to econometrics, a discipline he helped create (during the Second World War he edited the quarterly *Econometrica*). His textbook of econometrics (1959), the first of its kind in eastern European countries, recapitulates his studies of business cycle and of market mechanisms, in addition to an outline of programming theory based on Leontief's input–output tables and on Marxian reproduction schemata.

The evolution of capitalism was a close interest both as a scholar and as a political writer. Initially he believed that the development of large corporations marked a transition from ‘the anarchical freemarket capitalist economy to a consciously planned economy’ (1929[1973, p. 70]), that is, to an organized capitalism. But with the Great Depression those hopes vanished. Monopolies and government intervention cause chaos and disarray in the economy and led eventually to a collapse of capitalism and the victory of socialism (1931b[1973]). Soon, however, he came to the conclusion that ‘it was not capitalism but the worker movement which collapsed during the crisis’ resulting in a ‘stabilization of capitalism’ (1933[1973, p. 63]).

Just before and during the war, Lange often argued that capitalism cannot possibly be reconciled with economic progress in the long run. But at the same time he looked for ways of reforming capitalist structures to turn them into mixed-type economies – calling for a socialization of the monopolies which he regarded as threats to political democracy and which he blamed for generating unemployment.

During his stay in the United States, Lange published a number of contributions exploring and developing, as well as criticizing, the standard economics which was, and continues to be, taught at most universities in the West. Those studies fall roughly into two categories: the first was ‘pre-Keynesian’

from the point of view of general approach, while the other was closely connected with the absorption of the ‘Keynesian Revolution’ by traditional economics.

In one major study (1936b; 1937b), Lange tried to explore the relationships between interest theory and the theory of production factor cost. Using a strongly simplified model (one final commodity produced by labour and one capital good, free competition, ‘neutral’ role of money, risk is neglected), Lange unfolded a theory of interest which in many of its points came close to that of Frank Knight, even though in his concept of money capital (‘as a general command over means of production’) he was influenced more strongly by Schumpeter and Marx.

Lange is regarded as one of the founders of ‘modern welfare economics’ (Graaff, 1957). Following Bergson's pioneering study (Burk, 1938), Lange listed (1942a) theorems, which do not require interpersonal comparability of utility as well as those which do. The study of optimal distribution of incomes must be based on a priori hypotheses concerning marginal utility of incomes for different persons. For welfare economics propositions it is not necessary that utilities of individuals must be measurable as long as these utilities can be ordered.

The next and probably most important group of studies concern Keynesian theory's relationship to the mainstream of Western economic thinking. In a (1938b) study, Lange explores the internal logic of Keynes's theory investigating the mutual relations between interest rate, propensity to consume, marginal efficiency of capital, investment and national income. In Lange's model, elasticity is the all-decisive concept. Using this concept and some of Walras's ideas, Lange outlined a ‘general theory’ of which the Keynesian theory was one particular case. That special case occurs when elasticity of liquidity preference to income is close to zero or when it is infinitely great in relation to the rate of interest. Then, the rate of interest does not depend on marginal efficiency of capital or on propensity to consume. When the elasticity of liquidity preference to the rate of interest is close to zero, then the classical and neoclassical theory, stressing the dependence of money demand on income alone, holds. Keynes approved Lange's interpretation of his theory as following ‘closely and accurately my line of thought’ (Keynes, 1973, p. 232n). Lange's exposition of the notion of multiplier (1943a) was more modest in its intention.

Analysing Say's Law (1942b), Lange made one of the first ever attempts to overcome what was called the dichotomy of the pricing process. In traditional neoclassical theory, commodity prices were determined under the assumption that money is just ‘a worthless medium of exchange and a standard of value’ (1942b, p. 64), and hence of a barter economy. Only later on, prices determined in this way, were pecuniary prices ‘superimposed’. Accordingly, the substitution of money for commodities and vice versa was ignored completely. That was the gist of the assumption that total demand is identically equal to the total supply of commodities. Thus, the theory of money must start with the rejection of this contention (of Say's Law) and investigate conditions and processes leading to equilibrium of total demand with total supply. For this purpose, money must be included in the theory of general equilibrium.

These studies prepared the ground for a more ambitious synthesis. In his previous studies, Lange had already studied questions and problems asked by Keynes (this partly holds also for the theorists of imperfect competition and for Schumpeter) and tried to resolve them in his own fashion, relying on mathematical tools of general economic equilibrium as developed and modified by Henry Schultz, R.G. D. Allen and Paul Samuelson, but especially by J.R. Hicks.

That undertaking found its most complete and systematic exposition in Lange's (1944a) book, which sums up his theoretical work during his American period. The book is something like a restatement of

the theory of general economic equilibrium in which money is incorporated explicitly as part of this theory. Substitution between money and goods is the key concept for understanding processes of equilibrating and disequilibrating the national economy. As Lange puts it, ‘The interest in the problem and the recognition of the crucial importance of substitution between money and goods were inspired by Lord Keynes. For the tools of analysis the author is heavily indebted to Professor J.R. Hicks’ (1944a, p. vii).

But Lange's book was an outcome as much of theoretical as of practical disputes over general economic policy. His main point of interest was the belief, which survived repeated attacks from Keynesians, that price flexibility – and in particular flexible prices of production factors, mainly of labour – is a condition of full utilization of production factors. Defending the Keynesians’ position on this matter, Lange intended to reach both the general public and sophisticated, mathematically minded economists who refuted Keynes's language of aggregate concepts as too unscientific.

With a view of such different audiences, Lange composed his exposition at two or even three levels of difficulty. The main body of the book is ‘as simple as possible’ and in colloquial non-mathematical language full of socio-political corollaries. Only in the numerous footnotes did he present technical details. The final part of the book, called ‘The Stability of Economic Equilibrium’ and published as an appendix, is in rigid mathematical language and is addressed to the narrower group of specialists.

The book's main message can be summarized in the following way. There are three ways in which money can affect economic equilibrium under flexible prices:

1. 1. If the overall amount of money is constant, the fall in prices of a factor leads at first to a fall in other prices and to a growth in purchasing power of the existing stock of money. An excess supply of money arises. This, in turn, drives up demand for goods and checks prices from falling further. As other prices are falling less quickly than that of the factor under consideration, demand for this factor increases. Along with that, the amount of loanable funds grows, which causes a fall of the interest rate. This, then, encourages investment and results in employment growth. This is the case of the effect of money being *positive*.
2. 2. When the overall amount of money is determined by credit creation and changes in step with the changing demand for money (cash balances), the effect of money can be said to be *neutral*. In this case, the mechanism of automatic maintenance and restoration of equilibrium no longer works. The stock of money shrinks in proportion to the falling demand for cash balances and an excess money supply develops. The purchasing power of the stock of money remains unchanged. In consequence, the fall in prices is not checked by a rise in the purchasing power of the stock of money and interest rates do not fall. The excess supply of the production factor under consideration is not being absorbed.
3. 3. Money has a negative effect when its amount shrinks more than proportionately to falling demand for cash balances. Banks, for example, react to the fall in prices by demanding loan repayment. A shortage of money is then felt in the market. Pessimism, growing uncertainty, and so on fosters this development. Then, a fall in the given production factor's price (for example, wages) causes an even more dramatic fall in prices of other goods, which leads to an even larger excess supply of the production factor than was the case originally (for example, to even higher unemployment).

Lange's general conclusion from his analysis was quite pessimistic:

> Only under very special conditions does price flexibility result in the automatic maintenance of restoration of equilibrium of demand for and supply of factors of production. These conditions require the combination of such a responsiveness of the monetary system and such elasticities of price expectations as produce a positive monetary effect, sensitivity of intertemporal substitution to changes in interest rates …, absence of highly specialized factors with demand or supply dependent on strongly elastic price expectations, and finally, absence of oligopolistic or oligopsonistic rigidities of output and input. To a certain extent, the absence of a positive monetary effect may be replaced by the stabilizing influence of foreign trade … (1944a, p. 83)

On the whole, Lange regarded price flexibility as ‘a workable norm’ of long-run but not necessarily short-run economic policy during the long period of between the 1840s and 1914. However, the favourable conditions which prevailed during that period belong to the remote past. The oligopolization process, the deteriorating investment opportunities, the tendency towards money supply caused by new technology applications, along with the bad experiences of the two world wars and the Great Depression – all these made any automatic attainment of equilibrium and stability a very unlikely prospect.

This conclusion prompts two questions. First, what significance does the general economic equilibrium theory have for economic theory and for economic policy? Several years later, Lange compared that theory, which deals with very unlikely contingencies, to the case of an ape trying to write the Encyclopaedia Britannica. While probability calculus does not preclude such a possibility, we should ask ourselves if dealing with such an unlikely case is not an utterly futile exercise.

Price flexibility was the last fruit of Lange's study of the general equilibrium theory. To what extent his subsequent silence on this subject was due to the fact that, after 1945, he found himself in an entirely different environment, and to what extent due to his disenchantment with the theory, is difficult to say. Anyway, his economic thinking in later years took an unexpected turn. Contrary to his attitude in public life, as a philosopher of science Lange was rather conservative-minded, believing that ‘science does not progress … by the wholesale rejection of old theories and the devising of new ones, but by arduous work of enriching and improving existing scientific achievements’ (1970, pp. 80–1). Accordingly, he put a great deal of effort into showing that the so-called Keynesian Revolution was no revolution at all; and that it should be viewed as a contribution merely ‘enriching and improving scientific achievements’. But when he accomplished that job, Lange dropped the synthesis he had worked out with such a great expense of effort only to choose an alternative paradigm.

After the Second World War, however, Lange only sporadically resumed his study of capitalism, mainly to consider whether capitalism is able to resolve economic problems of backward countries (to which his answer was emphatically negative, 1957) or prospects for disarmament and economic cooperation between the Council for Mutual Economic Assistance countries and the capitalist West.

Lange – Breit model of socialist economy

Lange first manifested himself as a socialist writer in his book (1928b) on Edward Abramowski (1868–1918), whose ideology Lange called ‘constructive anarchism’. In those ideas, Lange emphasized

Abramowski's resentment of government interventionism, pitting it against the ideas of English Guild Socialism and of Austro-Marxism, both of which had strongly influenced Lange himself. Lange advocated especially the idea of industrial self-government, of separating the economy from political power, and the decay of the state as an institution of class domination though not of an instrument of coercion.

Together with Marek Breit (1907–42), he wrote the first outline of a socialist economy's functioning in the chapter of a collective book, *Economy–Polity–Tactics–Organization of Socialism* (1934[1973]). It was the product of a group of left-wing socialists, led by Lange, and committed to the revolutionary reconstruction of a system in Poland, which would be different from the Soviet model of polity and economy.

The Lange–Breit model, or the 1934 model (see Kowalik, 1970; 1974; Chilosi, 1986, 2005; Toporowski, 2003) is one version of a corporate market economy under socialism. It rests on the following rules.

Plants should go public, or be ‘socialized’, in his terminology, by transferring private ownership titles to a Public Bank and by organizing the national economy into public trusts by industrial branches. Trusts would be the basic units of the economy and endowed with a great deal of autonomy. The decisive say in their boards would belong to workers, who would be organized into ‘an appropriate system of worker councils’. Trusts autonomy is limited by the Public Bank's supervision and coordination functions or, more exactly, by the functions performed by a uniform and monopolistic bank system. Basic planning instruments would include accumulation fund management and trust financing. The Public Bank would also watch if trusts and companies subordinate to them abided by management rules, in particular by rules of ‘rigorous’ price and cost accounting. Plants run at a loss would be closed down. Plants failing to record an average surplus would forfeit their right to get loans not only for expansion but even for ordinary capital replacement, and hence they would decline. Both trusts and plants would be obliged not only to remit their production costs but also to achieve a certain accumulation, the rate of which would be established by the Public Bank and subsequently redistributed for investment and for subsidizing public utilities (which may be run at a loss).

Since trusts would hold virtually monopoly power in the market, as all public plants would by law belong to some trust, Lange and Breit perceived the danger of charging excessive prices and cutting output rates. They realized that such a policy might become quite popular among employees of any given trust, who might hope to get their wages increased. To forestall monopoly practices, they therefore proposed to oblige trusts to take on all job-seekers applying to them. If price increases resulted in higher wages in any given trust, employees from other trusts would swarm to it so that the increased wage fund would have to be redistributed among a larger number of employees. The underlying purpose of that obligation, then, was to deter trusts from driving up prices.

As the two authors did not consider the question of inflation, they did not say why excessive wage increases by one trust should not set off an avalanche of price increases if other trusts attempted to forestall an exodus of their own workforce. Nor did they envisage possible consequences of the indivisible nature of means of production and of possible consequences of delays in market adaptation.

Moreover, the Public Bank's investment policy would be based on workforce migration in reaction to changing demand, price fluctuations and subsequently price changes. This was to be something like an automatic indicator of demand intensity for individual goods.

The Public Bank would further control capital imports and exports, whereas a ‘foreign trade office’ created by the trusts concerned would be in charge of goods sales and purchases abroad. The Public

Bank would also be authorized to transfer capital assets from trust to trust.
The private sector, which is consistently referred to as the ‘non-socialized’, that is, non-public, sector of the economy, was to remain ‘broad’, consisting of private farms holding less than 20 hectares of land, crafts shops, business enterprises with less than 20 people on their payrolls, as well as retail trade shops. However, because economies of scale were expected to impart higher efficiency to larger companies, the private sector would be ‘a relic on the way out’. The two authors said nothing about credit policies towards this sector, but the Public Bank would conduct a discriminatory kind of policy towards profit-making small capitalist businesses (up to 20 employees) designed eventually to bring about their demise through taxes. Lange and Breit recommended that the Public Bank should levy taxes equal to the accumulation rate, which was supposed to reduce owners’ incomes to the level of manager's salaries. The two authors failed to take account of the role of risk and innovation.
Nor is it clear how the two authors thought plants (which they preferred not to call enterprises) would be managed, or how trusts would be organized and what prerogatives the latter would have. They merely said workers organized in a system of worker councils would have the decisive say and that trade unions and worker cooperatives were best suited to create trusts. Nor did they propose any clear procedure for appointing the Public Bank's board of management, which was expected to make the socialist economy a planned economy.
Designed as an alternative model to the command–planning system then existing in the Soviet Union, the Lange–Breit concept was largely reminiscent of Bolshevik concepts from before the period of wartime communism or right after it (trusts, worker councils, a single state-owned bank, a long-run policy of farm collectivization), modified by an emphasis on separating political authority from economic organization, on impartial economic criteria, and on recognizing consumer preferences as the foundation of investment policies.

The theory of market socialism

The next model of socialist economy, which I propose to call the classical, Lange presented in a study (originally published as two articles, 1936a; 1937a), and in a book form (with Taylor, 1938b). It was devised only two or three years after publishing Lange–Breit model. But this period brought an immense improvement of Lange's analytical expertise.
On a Rockefeller Foundation Grant, Lange studied at Harvard, Berkeley and Chicago, and at the London School of Economics. He was strongly influenced by Schumpeter, under whose tutorship he worked at Harvard during most of his two-year scholarship, and he took part in a famous seminar (The Economics Club) led by the Austrian-American economist. That influence surfaces in many of Lange's studies, including his study *On the Economic Theory of Socialism*, especially in the economic justification of socialism. That study, or at least its main body, was written at Harvard and must have been heatedly discussed there. At that time he also became intellectually involved with the brothers Alan and Paul Sweezy, economists and socialists of a similar orientation to that of the visitor from Poland. He also had working contact with W. Leontief.
On the Economic Theory of Socialism expresses Lange's long-lasting conviction that neoclassical economics, especially welfare economics, is best suited to serve as a foundation of a theory of socialist economy.
The classical model, of course, is theoretically more sophisticated and more accurate in its purely

economic aspect, but perhaps at the cost of giving less specific treatment to institutional aspects than the 1934 model. That was probably due to the chief purpose of that study, namely, to disprove Mises' argument about a theoretical and practical (practical, according to Hayek and Robbins) infeasibility of economic calculus in socialism because of the absence of a genuine market (prices) for capital.
Many formulations in that classical study indicate that a socialist society's general outlines of economic organization were similar or identical in both the early and classical models. In particular, this is true of the separation of political power from economic management, of its three-level structure – the centre, the branches organized in trusts, individual plants – and of the similar powers of the Central Planning Board (CPB) and the Public Bank. In both models, the centre is expected to react to changes in market factors (prices and wages) and, correspondingly, to changes in employment in the early model or to changing inventories and emerging shortages in the classical one. The CPB, basically, is to imitate the market. The early model was clearly more 'market-oriented' because all prices of goods and services were to be determined by the market. Accordingly, there would be no difference between actual market prices and calculated prices as set by the CPB.

Lange – Lerner mechanism

This is a designation commonly used to denote a market-oriented socialism model devised by Lange, who later amended it after public discussion with Lerner. The first, fundamental part of Lange's study was published together with A.P. Lerner's (1936) critical remarks in the same issue of the *Review of Economic Studies*, while the second part appeared together with Lange's reply to Lerner (1937). Later on, Lange made the changes necessary to publish his study (together with F.M. Taylor's essay) in book form (1938b). The term is occasionally used in a less restricted sense, to bring out the similarity of Lange's and Lerner's views on other matters concerning socialist economy.
The mechanism of socialist economy in the Lange–Lerner blueprint was based on the following assumptions. It has its institutional framework in the public ownership of means of production (for simplicity, the private sector is omitted) and in the free choice of consumption and employment (job and workplace), while consumer preferences – 'through demand prices' – are the all-decisive criterion of both production and resource allocation. Under these assumptions, an authentic market (in the institutional sense) exists for consumer goods and labour services. But prices of capital goods and 'all other productive resources except labour' are set by a CPB as indicators of existing alternatives established for the purpose of economic calculation. So, apart from market prices, there are also 'accounting prices'. In order to make their choices, both categories of prices are used by enterprise and industry managers, who are public officials.
Production managers in charge of individual enterprises or entire industries make autonomous decisions about what and how much should be produced and how it should be done, while prices are set as parameters outside the enterprises or industries. But since profit maximization has by definition ceased to be a direct goal of economic activity, to ensure that they can achieve effects close to those achieved in free-market economy, production managers must obey two rules. First, they must pick a combination of production factors under which average cost is minimized; and second, they must determine a given industry's total output at a level at which marginal cost is equal to product price. The first rule was expected to eliminate all less efficient alternatives. In combination with the second rule, in so far as it concerns plant managers, it performs the same function as the free-market economy desire to maximize

profit. This leads to minimization of production costs. The second rule compels production managers to increase or cut the output of a whole industry in accordance with consumer preferences, which is a substitute for free entry in a free competitive economy.

These rules lead to an economic equilibrium by the trial-and-error method first described by Fred M. Taylor (1929). The CPB acts like an auctioneer, initially watching the behaviour of economic actors in reaction to a price system it picks at random or – perhaps the best solution – to the historically inherited prices. The behaviour of the system is measured by the movement of inventories of goods. If there is too much of some product at a given price, then its inventory grows, and vice versa. This is regarded as information that the product price should be cut or increased, respectively. This procedure is applied as many times as is necessary to reach equilibrium, providing that this process does in fact converge to the system of equilibrium prices. Accounting prices, then, are objective in character, just like market prices in a competitive system, the difference being that in this case the CPB performs the role of the market.

The same trial-and-error way towards equilibrium could also be applied in two other models of socialist economy, one providing for a decreased consumer influence on production programme, the other presupposing none at all.

In its extreme version, which for sociopolitical reasons Lange deems untenable, the model might provide no freedom of choice for either consumption or employment. Production plans would be decided by the CPB officials’ scale of preferences. In such a version all prices are basically accounting prices. Consumer goods are rationed, while the place and kind of employment are imposed by command. If production managers keep to the above-mentioned rules, and if the CPB keeps to the parametric price system, then economic calculus is possible even in this version, while prices are not arbitrary but reflect the relative scarcity of factors of production.

There is an intermediate model, which provides for freedom of consumption decisions but only within a production plan established on the ground of CPB preferences. In this case, accounting prices of producer and consumer goods reflect the CPB's preference scale, while production managers would rely on them in their decision-making. Market prices for consumer goods would be set by supply and demand. But Lange rejects even this system as undemocratic, saying that the dual system of prices could be applied only when there is widespread agreement that checking the consumption of some products (say, alcohol) while promoting the consumption of other goods (say, cultural services) is in the public interest.

But the CPB might conceal its preferences and resort to rationing production goods and resources. Society can defend itself against such practices by creating a supreme economic court, which would be entitled to declare any unconstitutional CPB decision as null and void. In Lange's view, any decision introducing rationing would be unconstitutional.

Interestingly, Lange rejects these two versions of socialist economy on account of the potential hazards they carry for democracy, and says not a word about democracy's possible link with economic efficiency.

Lange considers the distribution of national income in three aspects.

Wages would be differentiated by seeking a distribution of labour services that would maximize society's wealth in general. This happens when differences in marginal disutility of work in different trades and workplaces are offset by wage differences. Wage differentials can be treated as converses of prices paid by employees for differing work conditions, as a simplified form of buying free time, safety or pleasant work (which is easy to imagine assuming that all employees get the same earnings but pay different prices for doing different jobs; the easier and safer a given job, the more one has to pay for it).

In this sense, the wage differentiation rule can be brought into harmony with egalitarianism.
Apart from wages paid by employees, each consumer is paid a public dividend as his or her share of capital and natural resources. At first Lange was inclined to distribute such dividends proportionally to wages. But as Lerner pointed out that such a policy would impart added attractiveness to the hardest jobs, Lange changed his mind, saying there should be no link between procedures for public dividend distribution and wage differentials.
The distribution of national income between consumption and accumulation, said Lange, would not be arbitrary when only consumers' individual savings decide the rate of accumulation. But if savings are 'corporately' determined – and Lange at first thought that was typical of a socialist economy – then there would be no way of preventing the CPB from being at least partly arbitrary in its decisions.
Emphasizing that resource allocation is guided by formally analogous rules in both socialist and free competitive economies, Lange argued that real allocation in socialism would be different from and more rational than that in capitalism. In his static analysis, he considered the following factors as decisive in judging the relative performance of the two systems. Greater equality of income distribution enhances society's well-being (in the subjective sense, that is, as a sum total of individual satisfactions). Second, socialist economy makes allowances in its calculus for all the services rendered by producers and for all the costs involved, while a private entrepreneur does not care for benefits that do not flow into his own pocket nor for costs he does not have to pay: 'Most important alternatives, like life, security, and health of the workers, are sacrificed without being accounted for as a cost of production' (1938b, p. 104).
Even the possible flaws that Lange conceded might appear in a socialist economy, such as the arbitrary setting of the rate of accumulation or the danger of bureaucratization of economic life, would be milder than under capitalism, he argued.
But the ultimately decisive economic argument in favour of socialism, Lange believed, was the general waste and endogenous tendency towards stagnation generated by modern capitalism's monopolistic tendencies. This question, though, goes beyond the scope of the often-criticized static analysis underlying Lange's classical model. Leaving aside the now enormous critical literature, let us try to answer the question of what Lange himself saw as his model's limitations.
Lange anticipated possible charges by critics in the second part of his study, in his discussion of 'The Economist's Case for Socialism':

> The really important point in discussing the economic merits of socialism is not that of comparing the equilibrium position of a socialist and of a capitalist economy with respect to social welfare. Interesting as such a comparison is for the economic theorist, it is not the real issue in the discussion of socialism. The real issue is *whether the further maintenance of the capitalist system is compatible with economic progress*. (1938b, p. 110)

But as he develops this general idea, Lange clearly uses an asymmetrical kind of argument. Having presented free competitive capitalism as the system that generated 'the greatest economic progress in human history', Lange proceeds to show (among other things, by referring to Keynes) that the source of that progress is drying up because of the progressive concentration and monopolization of production. His main point is that corporations, which are capable of controlling the market, attempt to avoid losses due to capital depreciation caused by innovation, and hence they try to check progress in technology.

Neither a return to free competition nor government control can effectively eliminate this tendency. The only effective solution, then, is the socialization of big capital, the introduction of socialism.
But will socialism ensure rapid technical progress? Will the abolition, via socialization, of capitalist monopolies' well-known tendency to check technological progress automatically dismantle all the barriers to innovation? Or will it amount to substituting new barriers for old? Will the two rules for managers be sufficient to guarantee the adoption of state-of-the-art production techniques? In his classic study, Lange never even asked such questions and only much later did he become aware of them. Towards the end of his life (in a letter to the present writer dated 14 August 1964), Lange wrote:

> What is called optimal allocation is a second-rate matter, what is really of prime importance is that of incentives for the growth of productive forces (accumulation and progress in technology). This is the true meaning of, so to say, 'rationality'.

It seems that he must have lacked the indispensable tools to solve this question or even to present it in detail.

Towards a mixed economy

Perhaps, the most important difference between the early (Lange–Breit) and the classical models was his new emphasis that 'the real danger of socialism is that of a bureaucratization of economic life, and not the impossibility of coping with the problem of allocation of resources'. He reassured himself by pointing out that the same danger existed in monopolistic capitalism and that 'officials subject to democratic control seem preferable to private corporation executives who practically are responsible to nobody' (Lange, 1938b, pp. 127–8).
When he became aware of that danger, which would exist even in a market-dominated brand of socialism, he embarked on a long quest for what he called in the title of one article (1943b), 'The Economic Foundations of Democracy in Poland'. In the classical study he had already put forward the idea of a Supreme Economic Court whose function would be to safeguard the use of the nation's productive resources in accordance with the public interest, in particular to declare as null and void any CPB decision which was incompatible with adopted management rules.
During the Second World War, Lange suggested a number of ideas for better safeguards for democracy, either by substantiating the injunction to take account of consumer preferences (and hence limiting the central economic authority's prerogatives) or by devising institutional guarantees for democratic control of decision-making bodies, or by indicating limits to the socialization of property.
There were a number of highlights of the evolution of Lange's views during that period.
In his letter to Hayek in 1940 (Kowalik, 1984) Lange gave a more accurate, and perhaps slightly different, description of the CPB's prerogatives for pricing goods and services:

> Practically, I should, of course, recommend the determination of prices by a thorough market process whenever this is feasible, i.e. whenever the number of selling and purchasing units is sufficiently large. Only where the number of these units is so small that a situation of oligopoly, oligopsony, or bilateral monopoly would obtain, would I

> advocate price fixing by public agency•…•.

Accordingly, he recommends socialization of industries only in areas where there is not automatic competitive market process.

Later in 1942–3, he departed even further from his classical model towards a mixed economy. In his review of Dickinson's book (1942c), he had the following idea of how to prevent the central authority's arbitrariness in determining the accumulation rate. With reference to Lerner's observation of the dependence of interest rates not only on the quantity of capital involved but also on investment rates, Lange thought that, if saving was ceded to individual consumers, accumulation rates could be made to reflect consumers' preference. His 1936–8 model should be improved in this way, he said.

In his two public lectures delivered in Chicago in 1942 on 'The Economic Operation of a Socialist Society' (1975), Lange tacitly dropped what was perhaps the chief feature of his classical model, namely, the central authority's prerogative of setting and reviewing prices as a road towards equilibrium. He made only a passing remark about such a possibility, and only in reference to *future* prices the centre may impose on production managers in order to ensure stable forecasting (which is as a rule erratic in capitalist economy).

But perhaps the greatest change in his concept of the desired shape of socialism can be found in his above-mentioned article on economic foundations of democracy in Poland (1943b). The title alone shows that a commitment to furnish solid economic foundations for 'Poland's democratic order' was the point of departure in designing future political transformations. In that article, Lange envisaged the socialization only of key industries (which necessarily include banks and transport). This would put an end to the power of 'the socially irresponsible monopolistic capitalism'. Having said this, he cautions that care should be taken to prevent the socialized key industries from becoming a foundation for 'an equally dangerous' threat to democracy in the form of too much economic power being concentrated in the state bureaucracy along with privileges arising from this.

But private farms, crafts shops and minor but also medium-sized industries were all to remain areas of private initiative and enterprise. So broad a field of action for private entrepreneurship was, on the one hand, to be one foundation of democracy, and, on the other, it was to preserve 'the kind of flexibility, pliability and adaptiveness that private initiative alone can achieve'. This is the reason for which the development of private sector is to be one of the chief guidelines for the socialized financial policy. The private sector then appears to have been a permanent element of the new model Lange proposed for Poland.

This proposal had its counterpart for the United States in the lengthy essay written with Abba P. Lerner on a democratic programme for full employment (1944b).

The changes in Lange's views of socialist economy during the war years were evidently so substantial that they could be used to compose from them an alternative version of a market socialism, compared with which his classical model can indeed be described as 'quasi-centralistic' (Pryor, 1985). The extent of those changes may have been the reason why he dropped his previous plan to revise his classical study:

> The essay is so far removed from what I would write on the subject today that I am afraid that any revision would produce a very poor compromise, unrepresentative of my thoughts. Thus, I am becoming inclined to let the essay go out of print and express my

> present views in entirely new form. I am writing a book on economic theory in which a chapter will be devoted to this subject. This may be better than trying to rehash old stuff. (Letter to M. Harding, 25 May 1945: 1986, p. 553)

Towards a major synthesis

Lange's lifelong ambition to produce a synthesis can be seen to have differed in scope, so that a ‘minor’ and a ‘major’ synthesis can be distinguished in it. His earliest endeavours included an attempt to incorporate the Marshall's method of partial equilibrium into the general equilibrium theory developed by the mathematical school (1932). In later years, he wrote a series of studies commenting on various aspects of the Keynesian theory to include it in and reduce to a particular case of general equilibrium theory.

Several times during his life Lange prepared himself to create his major synthesis. He did have the indispensable background for such a job, not only on account of his economic versatility (he was intimately familiar with all the main currents and schools in economic theory, and with the ‘three economic worlds’) but also because he felt at home in several other disciplines such as statistics and econometrics, history and sociology, praxeology and cybernetics.

The first outline for a major synthesis came in his article ‘Marxian Economics and Modern Economic Theory’ (1935). His chief argument was that these two currents are in fact complementary. Their advantages and drawbacks arose from the different specific tasks each of them was supposed to do. Marxian economics was designed to furnish the revolutionary movement with guidance for rational policies, defining as it did the lines and limitations of the evolution of capitalism. Modern economic theory, for its part, was expected to provide a foundation for capitalist management. But equilibrium theory, which was designed to serve precisely this purpose, was actually universal in character, so after some adaptation it could be used for day-to-day management of a socialist economy, a job Marxist economics was ill-suited to do. For some time Lange thought his synthesis should be based on marginalist economics, the categories of which seemed even useful for presenting problems of class structure. Clinging to ‘Marxist semantics’ was to him a sign of traditionalism and conservative attitudes.

In the late 1950s, he began to work on a three-volume treatise on political economy that would rest on two tiers – historical materialism and the principle of rationality. On a lower level of abstraction he attempted, rather unsuccessfully, to synthesize Marxian political economy with the neoclassical economics. He managed to finish the first volume (1959; 1963) on scope and method of economics and half of the second one (1966; 1971a). However, Poland was at that time only at the beginning of shedding its isolation straitjackets and thus of rapidly changing political, ideological and scientific perspectives and possibilities. That is why, only four years after the publication of the first volume, Lange came to a conclusion that, after having written the two next ones, it would need a substantial reworking.

From idea to reality

Having returned to Poland after the Second World War Lange gave an entirely new expression to his view of socialist economy. But by an ironic twist of history (to which he was fond of referring) he articulated his new approach only when his views changed in an entirely different direction from what

he was pursuing during the wartime: namely, Lange embarked on the search for a rationale for the command-type economy and subsequently for ways of reforming it.

The evolution of Lange's views of socialism in the post-war years is much harder to follow because he became so deeply involved in political activity. Not only the form but also the substance of his views was often influenced by tactical considerations and by the changing scope of freedom of expression accorded to scholars in social science. The freedom was broad prior to 1948, virtually extinct in the early 1950s, considerable in the latter half of the 1950s, and gradually curtailed later on.

The main change in Lange's theoretical approach was that he switched over from a micro to a macroeconomic approach. Whereas he had previously based his argument on the general equilibrium theory, after 1945 he relied on a Marxian reproduction model. The new approach was first presented in the report he submitted to the International Statistical Conference (1947) on practical economic planning and optimal resource allocation. In this report he tried to confront eastern European economic practices with welfare economics. His point was that the centre's main decisions resulted from a desire to industrialize the country as rapidly as possible. The economic successes those countries had scored up to then were due to full employment and to the liquidation of private monopolies, which worked as powerful checks on their national economies in the past. Economic choices were a second-rate matter in the period of reconstruction, but as those countries were moving into a phase of development more sophisticated choices may have to be made. Marginal analysis may in such events prove useful, provided it is carried out in categories adequately reflecting reality. Although, Lange talked about practical planning in descriptive rather than theoretical terms and although he did not reject marginal analysis, F. Perroux said:

> Je note que le théoricién socialiste a complétement changé de méthode. Il a autrefois essayé de montrer qu'une économie socialiste peut fonctionner à peu près comme isolée des unités économie de marché, sur la base de calcul. … Il fonde aujourd'hui sa thèse sur les macro-décisions de l'Etat. Il le fait paradoxalement au moment précisément où tout le monde est d'accord sur la necessité du ‘breakdown of the aggregate quantities’. (1947, p. 172)

The new theoretical approach was given more clear-cut contours in a booklet (1953) in which Lange commented on Stalin's famous work on socialist economy in the USSR. The reasons for which Lange wrote that book, in which he extolled the Stalin work as ‘a momentous event in the history of science with far-reaching practical consequences’, are somewhat puzzling. He did it, probably, for two reasons. First, he was convinced that the Stalin work marked a turn from economic voluntarism towards respect for the inexorable laws governing economic life, towards a rehabilitation of efficiency and greater consideration of social needs. Indeed, the first studies written by Polish theorists who later became known as revisionists did find some support in Stalin's work.

The second reason that prompted him to write this booklet must have been his view of the evolution the Communist economies were undergoing due to industrialization. He believed that not only the Stalinist terror but also the main body of practical devices applied then, as well as the functioning of the economy itself at that time, were all determined by political considerations, specifically by militarization and the forceful industralization bid (1943c). Lange often defined the centralistic command model as wartime economy. But he hoped that industrialization, with the subsequent emergence of an educated working

class and socialist intelligentsia, creates a good social base for democracy and decentralization of management. Presuming that industrialization entailed democratization, he believed the future of the ‘Polish economic model’ depended on how mature and experienced society will be. This is why he was unwilling ‘to design any new model from behind the desk’. In 1956–7 he refused to give his permission for the publication of an already finished translation of his classical work of 1936–8 because he did not want to lend his support to the ‘socialist free-marketers’. But it is unclear whether he regarded the market-oriented model of socialism as premature or as invalidated by the progress made in economic theory and practice (1967).

Late in his life, cybernetics and mathematical programming became his fascination. Using the theory of systems self-regulation and self-control, Lange gave an interpretation of the chief categories, wholes and parts, of dialectical materialism ([1962] 1965a). He also wrote an introduction to economic cybernetics (1965b), and to the theory of optimal decisions (1971b). This fascination was born from a belief in a great role of the computer as a most powerful device for central planning (sometimes called ‘computopia’). The strongest expression of this fascination contains his last publication on *The Computer and the Market* (1967). Recalling his polemics with Hayek and Mises, he confesses, that:

> Were I to rewrite my essay today my task would be much simpler. My answer to Hayek and Robbins would be: so what's the trouble? Let us put the simultaneous equations on an electronic computer and we shall obtain the solution in less than a second. The market process with its cumbersome *tatonnements* appears old-fashioned. Indeed, it may be considered as a computing device of the pre-electronic age.

It is rather obvious that such a view sharply contradicts his strong attachment to Marxian economics as economic sociology regarded by him as a seminal step in explaining a structure and evolution of capitalism.

This was, however, one side of his views. The other one, expressed rather in private communications, stems from his everyday observation and was truly pessimistic. Above we mentioned his opinion about the prime importance of incentives for the growth of productive forces termed by him as a true meaning of rationality. A couple of months before his death he did appreciate sociological factors of economic development: ‘Poland became a completely parochial country. It is going to become the Portugal of the socialist block. The sociological setting generates an enduring stagnation, while an “explosive” solution of her problems stands no chance of success (nor does it seem really desirable). A change, if it comes, may be touched off by external developments, namely when Poland falls too far behind the capitalist world and the socialist world’ (O. Lange's letter to T. Kowalik of 19 February 1965, in his possession). Even if a comparison of Poland with the Salazar-time Portugal may be shocking, Lange's prophecy has proved to be quite realistic. Fifteen years later an ‘explosive solution’ in a form of a ten-million mass movement, ‘Solidarity’, brought first a lot of hopes, ended with martial law, but in another ten years Poland entered upon a track of peaceful dismantling of a Communist system. Theoretically, this could have opened a freedom for a democratic choice of socio-economic system, based if not on Lange's classical model literally then on his general democratic and egalitarian principles. It happened to the contrary. A wild form of capitalism emerged. Ronald Reagan, Margaret Thatcher, Milton Friedman and F.A. Hayek became prophets. The works of Oskar Lange and another eminent economist, Micha• Kalecki, were rejected, their followers marginalized. At least five Polish politically engaged historians

accused Lange of being a secret agent for the Soviet Union during his stay in the USA, although 13 volumes of declassified FBI documents clearly contradict this slander.

The post-Langean concepts of market socialism

Not all Western intellectuals have treated the sudden collapse of the Soviet bloc as a final victory of liberal capitalism. Even the *Washington Post* published (on 14 January 1990) an article entitled ‘In Eastern Europe Social Democracy – not Capitalism of “1984” is winning’. Some of them saw the possibility of creating a new economic system, which would not simply emulate Western-type capitalism, and elaborated proposals using some of Lange's ideas as a starting point.
One of the first of them was Joseph Stiglitz, who as early as spring 1990 sent the following remarkable message to the post-Communist countries:

> The answer that socialism provided to the age-old question of the proper balance between the public and the private can now (…) be seen to have been wrong. But if it was based on wrong, or at least incomplete, economic theories (…) it was also based on ideals and values many of which are eternal. It represented a quest for a more humane and a more egalitarian society (…). As the former socialist countries embark on their journey, they see many paths diverging. There are not just two roads. Among these there are many that are less traveled by – where they end up no one yet knows. One of the large costs of the socialist experiment of the past seventy years is that it seemed *to foreclose exploring many of the other roads*. As the former socialist economies set off on this journey, let us hope that they keep in mind not only the narrower set of economic questions that I have raised (…) but the broader set of social ideals that motivated many of the founders of the socialist tradition. Perhaps some of them will *take the road less traveled by, and perhaps that will make all the difference, not only for them, but for the rest of us as well*. (Stiglitz, 1990, p. 70; 1994, p. 279, emphasis added)

Stiglitz was very critical about the Lange–Lerner model of market socialism as based on wrong premises of the neoclassical paradigm. However, inspired by more general ideas of Lange, Micha• Kalecki and the experience of Chinese gradual reforms, he suggested to the post-Communist countries several (for an American mainstream economist) very unconventional recommendations: not shock therapy as favoured by the IMF experts and particularly by Jeffrey Sachs, but evolutionary systemic changes; not market versus the state, but a search for the proper balance between market and government, the private and the state sector; not imitation of Anglo-Saxon capitalism, but a search for people's capitalism. He stressed that the post-Communist countries had most probably a chance to create a more egalitarian socio-economic system than any Western country. It was to be a mixed (market-cum-state) economy striving for social justice.
The efforts of a British philosopher and political scientist David Miller (1989) went in a different direction. Trying to create the theoretical foundations of market socialism, he explicitly says that his model would involve even ‘more extensive use of markets’ then the classical model of Oskar Lange. Several economists also presented different concepts as alternatives to capitalism directly referring to

some of Lange's ideas. The best known among them is John E. Roemer's (1994) proposal. Searching for an alternative system, which would be at least as efficient as present-day capitalism, he proposes to organize corporations in groups, operated according to the rules of the Japanese corporations called *keiretsu* with main banks crediting and monitoring them. Corporations would have to transfer after-tax profits to a state agency which would distribute it among all citizens as social dividend. This idea of a social dividend borrowed from Lange would be the main socialist feature of Roemer's model, which was nevertheless criticized as closer to capitalism than to socialist ideas.

Another American economist, James A. Yunker (1992), declared himself to be an enthusiast of Lange, not so much as an author of a classical model of socialism, but rather as a socialist thinker and particularly as a pioneer of reconciliation of conflicting theories. In this vein, he was arguing at the beginning of the 1990s for ‘East–West ideological convergence’ (Yunker, 1993) based on his ‘pragmatic market socialism’ presented in many publications. He took over from his master only certain ideas, such as the social dividend, the interest rate as the main regulator of investment, and the scope of public ownership to be limited to firms where management was separated from ownership. But in other respects Yunker's model was quite far from its original inspiration. The institutional crux of his concept was to be – as he writes – the Bureau of Public Ownership, which would take over all rights inherent in stocks, bonds and other financial instruments owned by private households. The operation of this public sector would be based on institutional investing, which would proceed much as it does in present-day capitalism.

Contrary to Stiglitz, both Roemer's and Yunker's models are based on fully fledged market mechanism and the neoclassical paradigm.

Different character of a book is that by W•odzimierz Brus (Oxford) and Kazimierz •aski (Vienna) (1989), both Polish emigrants as a result of the anti-Semitic campaign of March 1968. Earlier, while in Poland, Brus was a close collaborator of Lange and an eminent and very influential reform economist in the central European debates. Already in the beginning of the 1980s he became sceptical about the viability of market socialism. As a result of the analysis of its theoretical foundations and particularly a summary of the outcomes of reforms in the Soviet bloc, Yugoslavia and China, Brus and •aski are inclined to abandon the very concept of socialism meant as an economic organization radically different from capitalism. They do not reject, however, socialist ideals, but see them as possibly realized rather in Scandinavian-type reforms of capitalism. After the collapse of Communism they saw some sort of market socialism rather as a necessary stage of transition to a new socio-economic system, when a coexistence of public and private ownership will be tolerated.

Needless to say, in all countries of central and eastern Europe the above-mentioned concepts, even as cautious and moderate as that of Brus and •aski, fell on deaf ears.

See Also

- decentralization
- economic calculation in socialist countries
- efficient allocation
- planning

Selected works

1928a. Koniunktura w zyciu gospodarczym Polski 1923–27 [Business performance in Poland's economic life]. In *Przewroty walutowe i gospodarcze po wielkiej wojnie* [Currency and economic upheavals after the Great War]. Kraków.

1928b. *Socjologia i idee spo•eczne Edwarda Abramowskiego* [Sociology and social ideas of Edward Abramowski]. Kraków.

1929. ‘Wrastanie w socjalizm’ czy nowa faza kapitalizmu? [‘Growing into’ socialism, or a new phase of capitalism?]. *Robotniczy Przegl•d Gospodarczy* 3, 69–74.

1931a. *Statystyczne badanie koniunktury gospodarczej* [Statistical investigation of the business cycle]. Kraków.

1931b. Kryzys socjalizmu [Crisis of socialism]. In *Historia, ideologia, zadania Zwi•zku Niezale•nej M•odzie•y Socjalistycznej* [History, ideology, goals of the Association of Independent Socialist Youth]. Kraków.

1932. Die allgemeine Interdependenz der Wirtschaftsprognosen und die Isolierungsmethode. *Zeitschrift für Nationalökonomie* 4(1).

1933. Od kryzysu do stabilizacji kapitalizmu [From crisis to the stabilization of capitalism]. *P•omienie* 17/18, 218–25.

1934. (With M. Breit.) Droga do socjalistycznej gospodarki planowej [The road to the socialist planned economy]. In *Gospodarka–polityka–taktyka–organizacja socjalizmu* [Economy–polity–tactics–organization of socialism]. Warsaw.

1935. Marxian economics and modern economic theory. *Review of Economic Studies* 2(3), 189–201.

1936a. On the economic theory of socialism, Part I. *Review of Economic Studies* 4(1), 53–71.

1936b. The place of interest in the theory of production. *Review of Economic Studies* 3(3), 159–92.

1937a. On the economic theory of socialism, Part II. *Review of Economic Studies* 4(2), 123–42.

1937b. Professor Knight's note on interest theory. *Review of Economic Studies* 4(4), 231–5.

1938a. The rate of interest and the optimum propensity to consume. *Economica* 5, 12–32.

1938b. (With F.M. Taylor.) *On the Economic Theory of Socialism*. With an introduction by Benjamin Lippincott. Minneapolis: University of Minnesota Press.

1941a. Review of J. Schumpeter. *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process. Review of Economic Statistics* 23(4), 190–3.

1941b. Review of M. Kalecki. *Essays in the Theory of Economic Fluctuations. Journal of Political Economy* 49, 279–85.

1942a. The foundations of welfare economics. *Econometrica* 10, 215–28.

1942b. Say's Law: a criticism and restatement. In *Studies in Mathematical Economics and Econometrics*, ed. O. Lange et al. Chicago: University of Chicago Press.

1942c. Review of H.D. Dickinson. *Economics of Socialism. Journal of Political Economy* 50(2), 299–303.

1943a. The theory of the multiplier. *Econometrica* 11, 227–45.

1943b. Gospodarcze podstawy demokracji w Polsce [Economic foundations of democracy in Poland]. In *Ku gospodarce planowej* [Towards a centrally planned economy]. London.

1943c. *Working Principles of the Soviet Economy*. New York: Russian Economic Institute.

1944a. *Price Flexibility and Employment*. Bloomington: Principia Press.

1944b. (With A.P. Lerner.) The American way of business. In *Problems in American Lif: Teaching Aids for the Social Studies*. Washington, DC: National Council for the Social Studies.

1947. The practice of economic planning and the optimum allocation of resources. In *Proceedings of the International Statistical Conference*, vol. 5, Washington, DC. Published in *Econometrica* 17 (1949), 166–71.

1953. *Zagadnienia ekonomii politycznej w swietle pracy J. Stalina 'Ekonomiczne problemy socjalizmu w ZSRR'* [Problems of political economy in the light of J. Stalin's work, 'Economic Problems of Socialism in the USSR']. Warsaw.

1957. *Dlaczego kapitalizm nie potrafi rozwiaza• problemu krajów gospodarczo zacofanych* [Why capitalism is unable to solve the problems of backward countries]. Warsaw.

1959. *Introduction to Econometrics*. Oxford and London: Pergamon Press.

1963. *Political Economy. Vol. 1: General Problems*. Oxford and London: Pergamon Press.

1965a. *Wholes and Parts: A General Theory of System Behaviour*. Oxford and London: Pergamon Press.

1965b. *Introduction to Economic Cybernetics*. Warsaw: Pa•stwowe Wydawnictwo Naukowe.

1966. *Ekonomia polityczna*, vol. 2. Warsaw: Pa•stwowe Wydawnictwo Naukowe.

1967. The computer and the market. In *Socialism, Capitalism and Economic Growth: Essays Presented to M. Dobb*, ed. C.H. Feinstein. Cambridge: Cambridge University Press.

1970. *Papers in Economics and Sociology*, trans. P.F. Knightsfield. Oxford: Pergamon Press.

1971a. *Political Economy*, vol. 2. Warsaw: Pa•stwowe Wydawnictwo Naukowe.

1971b. *Optimal Decision: Principles of Programming*. Oxford: Pergamon Press.

1973. *Dziela* [Works], vol. 1. Warsaw: Pa•stwowe Wydawnictwo Ekonomiczne.

1975. *Dziela* [Works], vol. 3. Warsaw: Pa•stwowe Wydawnictwo Ekonomiczne.

1986. *Dzie•a* [Works], vol. 8. Warsaw: Pa•stwowe Wydawnictwo Ekonomiczne.

1994. *Economic Theory and Market Socialism: Selected Essays of Oskar Lange*, ed. T. Kowalik. Aldershot: Edward Elgar Publishing Limited.

2003. (With M. Breit.) *The Way to the Socialist Planned Economy (1934)*, trans. J. Toporowski. *History of Economics Review* 37, 51–70.

Bibliography

Bergson, A. 1967. Market socialism revisited. *Journal of Political Economy* 75, 655–73.

Brus, W. and •aski, K. 1989. *From Marx to the Market: Socialism in Search of an Economic System*. Oxford: Clarendon Press.

Burk, A. [pen-name of Abraham Bergson] 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52, 310–34.

Chilosi, A. 1986. Self-managed market socialism with ‘free mobility of labor’. *Journal of Comparative*

Economics 10, 237–54.

Chilosi, A. 2005. The Lange–Breit model, the right to employment and the outside option. In *Oskar Lange a wspó•czesno* Warsaw: Polskie Towarzystwo Ekonomiczne.

Graaff, J. de V. 1957. *Theoretical Welfare Economics*. Foreword by P.A. Samuelson. Cambridge: Cambridge University Press.

Keynes, J.M. 1973. *Collected Writings, Vol. XIV: The General Theory and After*. London: Macmillan.

Kowalik, T. 1970. Oskara Langego Wczesne modele socjalizmu [O. Lange's early models of socialism]. *Ekonomista* 5, 965–1000.

Kowalik, T. 1974. Zur klassischem Modell des Sozialismus. In *Sozialismus, Geschichte und Wirtschaft: Festschrift fur Eduard Marz*. Vienna: Europaverlag.

Kowalik, T. 1984. Review of A. Nove. *The Economics of Feasible Socialism. Contributions to Political Economy* 3, 91–7.

Lavoie, D. 1985. *Rivalry and Central Planning: The Socialist Calculation Debate Reconsidered*. Cambridge: Cambridge University Press.

Lerner, A.P. 1936. A note on socialist economics. *Review of Economic Studies* 4, 72–6.

Miller, D. 1989. *Market, State and Community. Theoretical Foundations of Market Socialism*. Oxford: Clarendon Paperbacks.

Perroux, F. 1947. Comments on Lange's paper. In *Proceedings of the International Statistical Conference*, vol. 5. Washington, DC. Published in *Econometrica* 17 (1949), 172–8.

Pryor, F.L. 1985. *A Guidebook to the Comparative Study of Economic Systems*. Englewood Cliffs, NJ: Prentice-Hall.

Roemer, J.A. 1994. *A Future for Socialism*. Cambridge, MA: Harvard University Press.

Stiglitz, J. 1990. Whither socialism? Perspectives from the economics of information. Mimeo, Stockholm.

Stiglitz, J. 1994. *Whither Socialism? Perspectives from the Economics of Information*. Cambridge, MA: MIT Press.

Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19, 1–8.

Toporowski, J. 2003. Marek Breit and Oskar Lange's financial model of a socialist economy. *History of Economics Review* 37, 41–50.

Yunker, J.A. 1992. *Socialism Revised and Modernized: the Case for Pragmatic Market Socialism*. New York: Praeger Publishers.

Yunker, J.A. 1993. New prospects for East–West ideological convergence: a market socialist viewpoint. *Coexistence* 30, 237–67.

How to cite this article

Kowalik, Tadeusz. "Lange, Oskar Ryszard (1904–1965)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000024> doi:10.1057/9780230226203.0930

The New Palgrave Dictionary of Economics Online

Lardner, Dionysius (1793 – 1859)

Robert B. Ekelund, Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Jevons, W. S.; Lardner, D.; mathematical economics; price discrimination; profit maximization; spatial economics

Article

Scientific popularizer and railway economist, Lardner was born in Dublin on 3 April 1793 and died on 29 April 1859. He was educated at Trinity College, Dublin, between 1817 and 1827 and is probably best known for his *Cabinet Cyclopaedia* of 133 volumes, published between 1829 and 1849. Although Lardner's series was graced by a number of distinguished contributors, he was satirized in the scientific community as ‘Dionysius Diddler’. An astronomer as well as an essayist on numerous scientific topics, Lardner often took side trips into other fields. He studied railway engineering in Paris, and was probably well acquainted with the econo-engineering work at the Ecole des Ponts et Chaussées at a time when Jules Dupuit was actively pursuing economic topics. His sole work relating to economics, *Railway Economy* (1850), was filled with the kind of factual work and analysis being undertaken by the French engineers and by an American pupil of the Ecole, Charles Ellet. Lardner's work caught the eye of W.S. Jevons, who claimed that a reading of *Railway Economy* in 1857 led him to investigate economics in mathematical terms.

There is little doubt that Lardner's book contains important and creative insights into economic theory. An authority on Belgian railroads of the time, Lardner drew up a vast array of facts to develop a theory of the railway firm's costs and revenues. His theory of profit maximization derived from ‘empirical’ firm's costs and revenues may be set out graphically (see Figure 1).

Figure 1

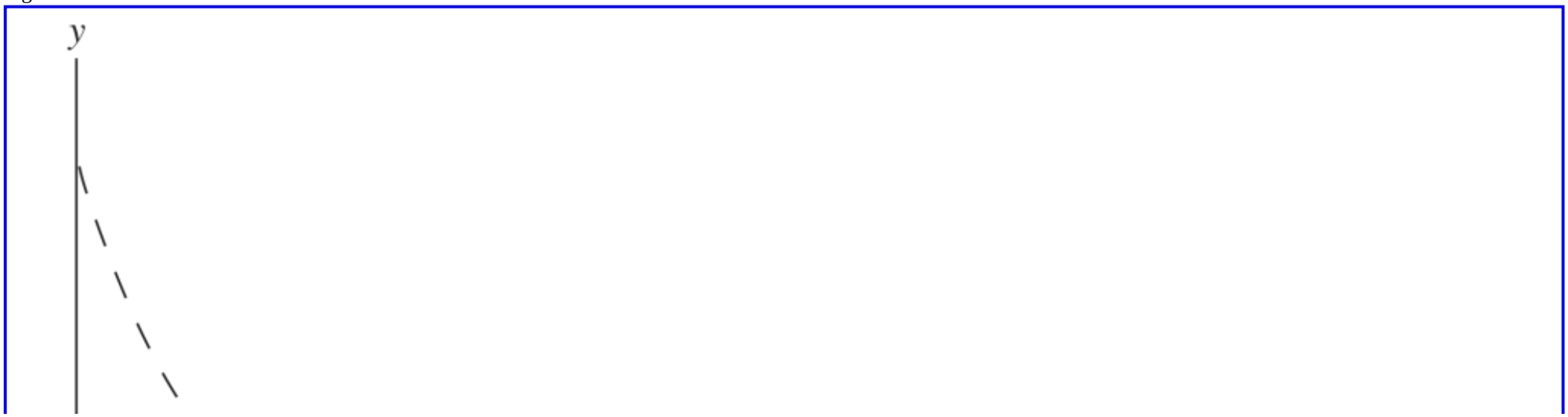

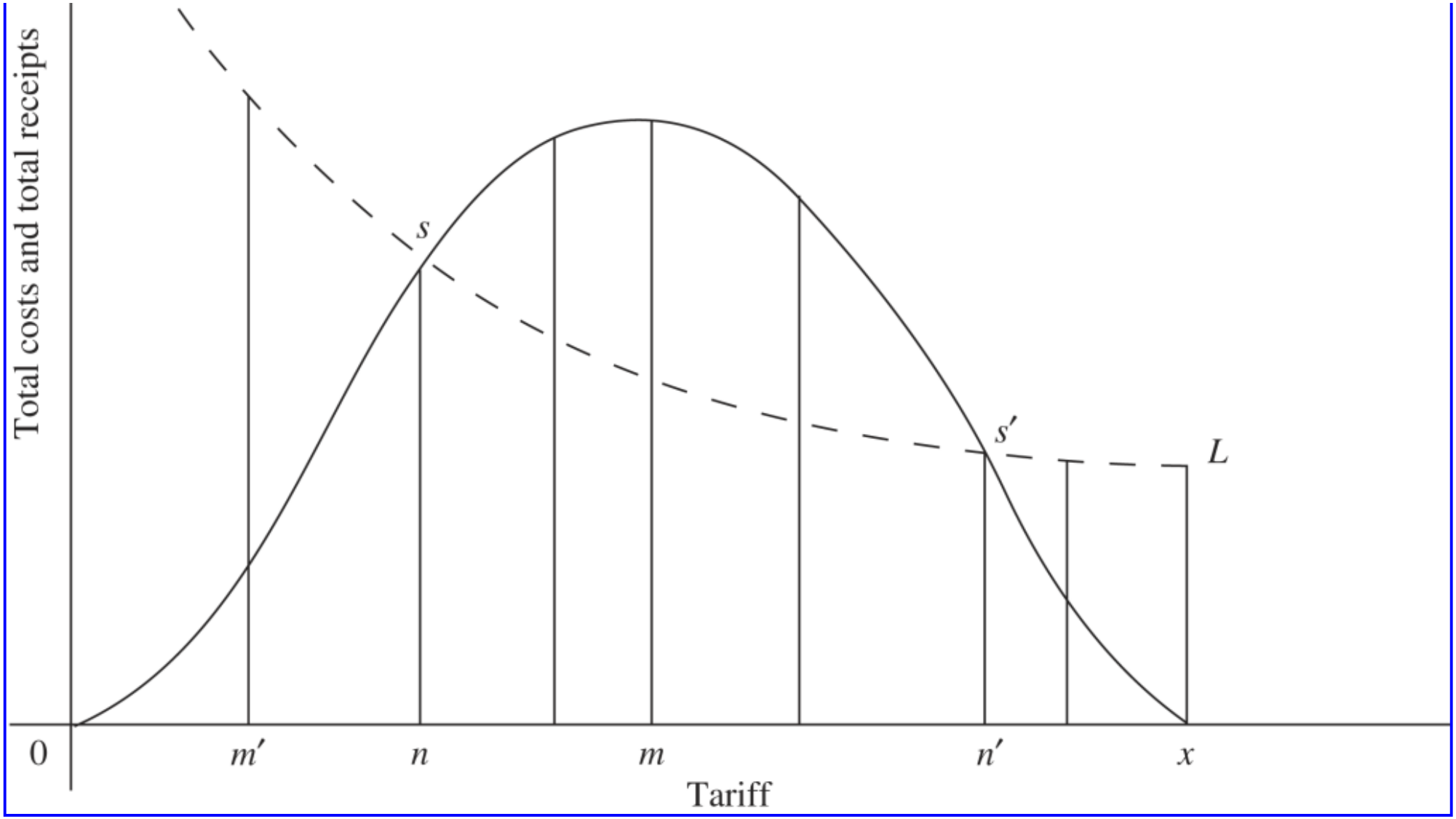

The railway tariff, which Lardner identified as the independent variable, is displayed on the horizontal axis of the figure while total cost and receipts are measured on the horizontal. The total cost curve shows costs increasing as the tariff is lowered. At a prohibitive tariff *Ox*, that is, where no traffic would be transported, costs are some positive amount. Fixed costs, which exist whether traffic is carried or not, are an amount *xL*. As the tariff is lowered, increases in traffic carried cause total costs to increase until they reach maximum at a zero tariff. Both fixed and variable components of cost, then, are considered by Lardner.

Lardner formalized his conception of total receipts in the following terms. If, with Lardner, we let r = the tariff imposed per mile on each ton of goods carried; D = the average distance in miles to which each ton of goods is carried; N = the number of tons booked, and; R = the gross receipts from goods transport, then total receipts may be expressed as

$$R = NDr.$$

As the tariff is lowered from *Ox*, the average distance of each ton carried, *D*, and the number of tons booked, *N*, increase. With reference to Figure 1, lowering the tariff from *Ox* causes receipts, *R* in Lardner's equation, to increase to some maximum *mp*. Tariff reductions below *Om*, however, cause total receipts to fall, so that at a tariff of zero, total receipts are zero (demand is inelastic for tariffs below *Om*).

Tariffs *On′* and *On* are 'break even' tariffs in Figure 1 and, significantly, Lardner argued that the profit-maximizing tariff would fall somewhere between the break-even tariff *On′* and the revenue-maximizing tariff *Om*. In modern terminology Lardner identified, if implicitly, the profit maximizing *quantity* as being where marginal cost equals marginal revenue. It is noteworthy that Lardner's analysis of profit maximization, which so impressed Jevons, is nowhere to be found in Jevons's writings.

In addition to a fine model of the profit-maximizing firm, Lardner presented a fairly complete theory of price discrimination related to location in his *Railway Economy*. Specifically, Lardner called for a reduction in long-haul rail rates and for the increase in short-haul rates in order to increase the aggregate profits of the railroad. The differing elasticities of demand for transport which made this discriminatory pricing structure possible were explained on the basis of spatially distributed demanders.

Selected works

1850. *Railway Economy*. Reprinted, New York: A.M. Kelley, 1968.

Bibliography

Ekelund, R.B., Jr., Furubotn, E.G. and Gramm, W.P. 1972. *The Evolution of Modern Demand Theory*. Lexington, MA: Heath.

Hooks, D.L. 1971. Monopoly price discrimination in 1850: Dionysius Lardner. *History of Political Economy* 3, 208–23.

Robertson, R.M. 1951. Jevons and his precursors. *Econometrica* 19, 229–42.

How to cite this article

Ekelund, Robert B., Jr. "Lardner, Dionysius (1793–1859)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000026> doi:10.1057/9780230226203.0931

The New Palgrave Dictionary of Economics Online

large economies

John Roberts
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

contract curve; convergence; convexity; Cournot, A. A.; large economies; non-standard analysis; oligopoly; Richter's theorem

Article

Economists have often claimed that our theories were never intended to describe individual behaviour in all its idiosyncrasies. Instead, in this view, economic theory is supposed to explain only general patterns across large populations. The prime example is the theory of competitive markets, which is designed to deal with situations in which the influence of any individual agent on price formation is ‘negligible’.
As in so many aspects of economics, Cournot (1838) was the first to make the role of large numbers explicit in his analysis. Cournot provided a theory of price and output which, as the number of competing suppliers increases without bound, asymptotically yields the competitive solution of price equals marginal and average cost. However, for any given finite number of competitors, an imperfectly competitive outcome results.
It took over a century for Cournot's insights on the role of large numbers to be fully appreciated. Edgeworth (1881) argued the convergence of his contract curve as the economy grew, and increasing numbers of authors assumed that the number of agents was ‘sufficiently large’ that each one's influence on quantity choices was negligible, but it was not until the contributions of Shubik (1959) and Debreu and Scarf (1963) to the study of the asymptotic properties of the core that the number of agents took a central role in economic analysis.
The crucial step in this line of analysis was taken by Aumann (1964). Arguing that, in terms of standard models of behaviour, an individual agent's actions could be considered to be negligible only if the individual were himself arbitrarily small relative to the collectivity, Aumann modelled the set of agents as being (indexed by) an atomless measure space. In this context, an individual agent corresponds to a set of measure zero, while aggregate quantities are represented as integrals (average, per capita amounts). Then changing the actions of a single individual (or any finite number) actually has no

influence on aggregates.

The non-atomic measure space formulation brings three mathematical properties that have proven important. The first is that it provides a consistent modelling of the notion of individual negligibility: only in such a context is an individual truly able to exert no influence on prices. Thus, this model correctly represents the primary reason for appealing to ‘large numbers’: in it, competitive price-taking behaviour is rational. Moreover, this individual negligibility, when combined with an assumption that individual characteristics are sufficiently ‘diffuse’, means that discontinuities in individual demand disappear under aggregation (Sondermann, 1975).

The second property is that a (non-negligible) subset of agents drawn from an economy with a non-atomic continuum of agents is essentially sure to be a representative sample of the whole population. This property has proven crucial in the literature relating the core and competitive equilibrium. (See Hildenbrand, 1974, for a broad-ranging treatment of these issues.) It is also used in showing equivalence of core and value allocations (Aumann, 1975).

The other important property of the non-atomic continuum model is the convexifying effect. Even though individual entities (demand correspondences, upper-contour sets, production sets) may not be convex, Richter's theorem implies that the aggregates of these are convex sets when the set of agents is a non-atomic continuum. This property yields existence of competitive equilibrium in large economies even when the individual entities are ill behaved and no ‘diffuseness’ is assumed.

In the non-atomic continuum modelling, the individual agent formally disappears. Instead, one has coalitions (measurable sets of agents), and an individual is formally indistinguishable from any set of measure zero. The irrelevance of individuals is made very clear in the model of Vind (1964), where only coalitions are defined and individual agents play no part. Debreu (1967) showed the equivalence of Vind's and Aumann's approaches. A further extension of this line is to consider economies in terms only of the distributions of individual characteristics and allocations in terms of distributions of commodities. The strengths of this approach are shown in Hildenbrand (1974).

This disappearance of the individual is intuitively bothersome: economists are used to thinking about individual agents being negligible, but not about individuals having no existence whatsoever. Brown and Robinson (1972) provided an escape from this dilemma by their modelling of a large set of agents via non-standard analysis. This approach gives formal meaning to such notions as an infinitesimal that had been swept out of mathematics and replaced by ‘epsilon-delta’ arguments. In interpreting non-standard models, one distinguishes between how things appear from ‘inside the model’ and what they look like from ‘outside’. From outside, these models may have an infinity of (individually negligible, infinitesimal) agents, yet from inside each agent is a well-defined, identifiable entity. Using this mathematical modelling eases the interpretation of large economies and also allows formalization of some very intuitive arguments that otherwise could not be made. Unfortunately, the difficulties of mastering the mathematics of non-standard analysis have limited the number of economists using this approach.

While these formal models capture the essential intuition about the nature of economic behaviour of large economies, results obtained in this context should be of interest only to the extent that these models provide a good approximation to large but finite economies. This point was first emphasized by Kannai (1970), and its elaboration was the central issue confronting mathematical general equilibrium theory through the 1960s and early 1970s. The issue is one of continuity: in what sense are infinite

economy models the limits of finite economies as the economy grows, and do the various constructs of interest (competitive or Lindahl allocations, cores, value allocations, and so on) of the finite economies approach those of the limit, infinite economies? These questions are extremely subtle. A good introduction to them is Hildenbrand (1974).

The study of the limiting, asymptotic properties of various economic concepts represents an alternative, more direct (but often less tractable) approach to large economy questions than does working with infinite economies. This line begins with Cournot's (1838) treatment of the convergence of oligopoly to perfect competition, the general equilibrium development of which has been a major focus of recent activity (see Mas-Colell, 1982 and the references there). The work growing out of Edgeworth (1881) and Debreu and Scarf (1963) on the core-competitive equilibrium equivalence noted above also follows this line.

Once such convergence is established, the crucial question becomes that of the rate of convergence because asymptotic results are of limited interest if convergence is too slow. This question was first addressed for the core by Debreu (1975), who showed convergence at a rate of at least 1 over the number of agents.

A more direct approach to this issue of how large a market must be for its outcomes to be approximately competitive is to employ a model in which price formation is explicitly modelled. (Note that this is not a property of the Cournot or Arrow–Debreu analyses.) In a partial equilibrium context the Bertrand (1883) model of price-setting homogeneous oligopoly indicates that ‘two is large’, in that duopoly can yield price equal to marginal cost. Recent striking results in the same line for the double auction are due to Gresik and Satterthwaite (1985), who show that, even with individual reservation prices being private information, equilibrium under this institution can yield essentially competitive, welfare-maximizing volumes of trade with as few as six sellers and buyers.

This work is very heartening, for it tends to justify the profession's traditional reliance on competitive models which make formal sense only with an infinite set of agents. Another basis for optimism on this count comes from experimental work which shows strong tendencies for essentially competitive outcomes to be attained with quite small numbers. The further study of such institutions is clearly indicated.

See Also

- non-standard analysis
- perfect competition
- Shapley–Folkman theorem

Bibliography

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Aumann, R.J. 1975. Values of markets with a continuum of traders. *Econometrica* 43, 611–46.

Bertrand, J. 1883. Theorie mathématique de la richesse sociale. *Journal des Savants* 48, 499–508.

Brown, D.J. and Robinson, A. 1972. A limit theorem on the cores of large standard exchange economies. *Proceedings of the National Academy of Sciences of the USA* 69, 1258–60.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Debreu, G. 1967. Preference functions on measure spaces of economic agents. *Econometrica* 35, 111–22.

Debreu, G. 1975. The rate of convergence of the core of an economy. *Journal of Mathematical Economics* 2, 1–8.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Gresik, T. and Satterthwaite, M. 1985. The rate at which a simple market becomes efficient as the number of traders increases: an asymptotic result for optimal trading mechanisms. Discussion Paper No. 641, Center for Mathematical Studies in Economics and Management Science, Northwestern University.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Kannai, Y. 1970. Continuity properties of the core of a market. *Econometrica* 38, 791–815.

Mas-Colell, A., ed. 1982. *Non-cooperative Approaches to the Theory of Perfect Competition*. New York: Academic Press.

Shubik, M. 1959. *Strategy and Market Structure: Competition, Oligopoly, and the Theory of Games*. New York: Wiley.

Sondermann, D. 1975. Smoothing demand by aggregation. *Journal of Mathematical Economics* 2, 201–24.

Vind, K. 1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5, 165–77.

How to cite this article

Roberts, John. "large economies." The New Palgrave Dictionary of Economics. Second Edition. Eds.

Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000027> doi:10.1057/9780230226203.0932

The New Palgrave Dictionary of Economics Online

large games (structural robustness)

Ehud Kalai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In strategic games with many semi-anonymous players all the equilibria are structurally robust. The equilibria survive under structural alterations of the rules of the game and its information structure, even when the game is embedded in bigger games. Structural robustness implies *ex post* Nash conditions and a stronger condition of information-proofness. It also implies fast learning, self-purification and strong rational expectations in market games. Structurally robust equilibria may be used to model games with highly unspecified structures, such as games played on the web.

Keywords

herding; large games; mixed-strategy equilibrium; pure-strategy equilibrium; purification; rational expectations equilibrium; structural robustness

Article

Earlier literature on large (many players) cooperative games is surveyed in Aumann and Shapley (1974). For large strategic games, see Schmeidler (1973) and the follow-up literature on the purification of Nash equilibria. There is also substantial literature on large games with special structures, for example large auctions as reported in Rustichini, Satterthwaite, and Williams (1994).
Unlike the above, this survey concentrates on the structural robustness of (general) Bayesian games with many semi-anonymous players, as developed in Kalai (2004; 2005). (For additional notions of robustness in game theory, see Bergemann and Morris, 2005.)

Main message and examples

In simultaneous-move Bayesian games with many semi-anonymous players, all Nash equilibria are structurally robust. The equilibria survive under structural alterations that relax the simultaneous-play

assumptions, and permit information transmission, revisions of choices, communication, commitments, delegation, and more.

Large economic and political systems and distributive systems such as the Web are examples of environments that give rise to such games. Immunity to alterations means that Nash equilibrium predictions are valid even in games whose structure is largely unknown to modellers or to players.

The next example illustrates immunity of equilibrium to revisions, or being *ex post* Nash, see Cremer and McLean (1985), Green and Laffont (1987) and Wilson (1987) for early examples.

Example 1: Ex post stability illustrated in match pennies

Simultaneously, each of *k* males and *k* females chooses one of two options, *H* or *T*. The payoff of every male is the proportion of females his choice matches and the payoff of every female is the proportion of males her choice *mis*matches. (When *k*=1 this is the familiar match-pennies game.) Consider the mixed-strategy equilibrium where every player chooses *H* or *T* with equal probabilities.

Structural robustness implies that the equilibrium must be *ex post* Nash: it should survive in alterations that allow players to revise their choices after observing their opponents' choices. Clearly this is not the case when *k* is small. But as *k* becomes large, the equilibrium becomes arbitrarily close to being *ex post* Nash. More precisely, the Prob[some player can improve his payoff by more than ε *ex post*] decreases to zero at an exponential rate as *k* becomes large.

Example 2: Invariance to sequential play illustrated in a computer choice game

Simultaneously, each of *n* players chooses one of two computers, *I* or *M*. But before choosing, with 0.50–0.50 i.i.d. probabilities, every player is privately informed that she is an *I*-type or an *M*-type. The payoff of every player is 0.1 if she chooses the computer of her type (zero otherwise) plus 0.9 times the proportion of opponents whose choices she matches. (Identical payoffs and prior probabilities are assumed only to ease the presentation. The robustness property holds without these assumptions.)

Consider the favourite-computer equilibrium (FC) where every player chooses the computer of her type. Structural robustness implies that the equilibrium must *be invariant to sequential play*: it should survive in alterations in which the (publicly observed) computer choices are made sequentially. Clearly this is not the case for small *n*, where any equilibrium must involve herding. But as *n* becomes large, the structural robustness theorem below implies that FC becomes an equilibrium in all sequential alterations. More precisely, the Prob[some player, by deviating to her non favorite computer, can achieve an ε-improvement at her turn] decreases to zero at an exponential rate.

The general definition of structural robustness, presented next, accommodates the above examples and much more.

Structural robustness

A mixed-strategy (Nash) equilibrium $\sigma = (\sigma_1, \ldots, \sigma_n)$ of a one-simultaneous-move *n*-person strategic game *G* is structurally robust if it *remains an equilibrium* in *every* structural *alteration* of *G*. Such an alteration is described by an extensive game, $\mathcal{A}$, and for σ to remain an equilibrium in $\mathcal{A}$ means that *every adaptation* of σ to $\mathcal{A}$, $\sigma^{\mathcal{A}}$, must be an equilibrium in $\mathcal{A}$.

Consider any *n*-person one-simultaneous-move Bayesian game *G*, like the Computer Choice game above.

Definition 1: A (structural) alteration of G is any finite extensive game $\mathcal{A}$ with the following properties:

1. 1. $\mathcal{A}$ *includes the (original) G-players*: The players of $\mathcal{A}$ constitute a superset of the *G*-players (the players of *G*).
2. 2. *Unaltered type structure*: At the first stage of $\mathcal{A}$, the *G*-players are assigned a profile of types by the same prior probability distribution as in *G*. Every player is informed of his own type.
3. 3. *Playing* $\mathcal{A}$ *means playing G*: with every final node of $\mathcal{A}$, *z*, there is an associated unique profile of *G* pure-strategies, $a(z) = (a_1(z), \ldots, a_n(z))$.
4. 4. *Unaltered payoffs*: the payoffs of the *G*-players at every final node *z* are the same as their payoffs in *G* (at the profile of realized types and final pure-strategies *a*(*z*)).
5. 5. *Preservation of original strategies*: every pure-strategy a_i of a *G*-player *i* has at least one $\mathcal{A}$ adaptation. That is, an $\mathcal{A}$-strategy $a_i^{\mathcal{A}}$ that guarantees (w.p. 1) ending at a final node *z* with $a_i(z) = a_i$ (no matter what strategies are used by the opponents).

In the computer choice example, every play of an alteration $\mathcal{A}$ must produce a profile of computer allocations for the *G*-players. Their preferences in $\mathcal{A}$ are determined by their preferences over profiles of computer allocations in *G*. Moreover, every *G*-player *i* has at least one $\mathcal{A}$-strategy $I_i^{\mathcal{A}}$ (which guarantees ending at a final node where she is allocated *I*), and at least one $\mathcal{A}$-strategy $M_i^{\mathcal{A}}$ (which guarantees ending at a final node where she is allocated *M*).

Definition 2: An $\mathcal{A}$ *(mixed) strategy-profile,* $\sigma^{\mathcal{A}}$*, is an adaptation of* a *G* (mixed) strategy-profile σ, if for every *G*-player i, every $\sigma_i^{\mathcal{A}}$ is an $\mathcal{A}$-adaptation of σ_i. That is, for every *G* pure-strategy a_i, $\sigma_i(a_i) = \sigma_i^{\mathcal{A}}(a_i^{\mathcal{A}})$ for some $\mathcal{A}$-adaptation $a_i^{\mathcal{A}}$ of a_i.

In the computer choice example, for a *G*-strategy where player *i* randomizes 0.20 to 0.80 between *I* and *M*, an $\mathcal{A}$ adaptation must randomize 0.20–0.80 between a strategy of the type $I_i^{\mathcal{A}}$ and a strategy of the type $M_i^{\mathcal{A}}$.

Definition 3: An equilibrium σ of *G is structurally robust* if in every alteration of *G*, $\mathcal{A}$, and in every adaptation of σ, $\sigma^{\mathcal{A}}$, the strategy of every *G*-player *i*, $\sigma_i^{\mathcal{A}}$, is best response to $\sigma_{-i}^{\mathcal{A}}$.

Remark 1: The structural robustness theorem, discussed later, presents an asymptotic result: the equilibria are structurally robust up to two positive numbers (ε , ρ), which can be made arbitrarily small as *n* becomes large. The notion of approximate robustness is the following.

An equilibrium is (ε , ρ)-structurally robust if in every alteration and every adaptation as above, Prob [visiting an information set where a *G*-player can improve his payoff by more than ε] $\leq \rho$. (ε -improvement is computed conditional on being at the information set. To gain such improvement the player may coordinate his deviation: he may make changes at the information set under consideration together with changes at forthcoming ones.)

For the sake of brevity, the next section discusses full structural robustness. But all the observations presented there also hold for the properly defined approximate counterparts. For example, the fact that structural robustness implies *ex post* Nash also implies that approximate structural robustness implies approximate *ex post* Nash. The implications of approximate (as opposed to full) structural robustness are important, due to the asymptotic nature of the structural robustness theorem.

Implications of structural robustness

Structural robustness of an equilibrium σ in a game *G* is a strong property, because the set of *G*-alterations that σ must survive is rich. The simple examples below are meant to suggest the richness of its implications, with the first two examples showing how it implies the notions already discussed (see Dubey and Kaneko, 1984 for related issues).

Remark 2: Ex post Nash and being information-proof

G with revisions, $\mathcal{GR}$, is the following *n*-person extensive game. The *n* players are assigned types as in *G* (using the prior type distribution of *G* and informing every player of his own type). In a first round of simultaneous play, every player chooses one of his *G* pure strategies; the types realized and pure strategies chosen are all made public knowledge. Then, in a second round of simultaneous play, the players again choose pure strategies of *G* (to revise their first round choices). The payoffs are as in *G*, computed at the profile of realized types with the profile of pure strategies chosen in the *second* round. Clearly $\mathcal{GR}$ satisfies the definition of an alteration (with no additional players), and every equilibrium σ of *G* has the following $\mathcal{GR}$ adaptation, σ^{NoRev}: in the first round the players choose their pure strategies according to σ, just as they do in *G*; in the second round nobody revises his first round choice. Structural robustness of σ implies that σ^{NoRev} must be an equilibrium of $\mathcal{GR}$, that is, σ is *ex post* Nash. Moreover, the above reasoning continues to hold even if the information revealed between the two rounds is partial and different for different players. The fact that σ^{NoRev} is an equilibrium in all such alterations shows that σ is *information-proof*: no revelation of information (even if strategically coordinated by *G*-players and outsiders) could give any player an incentive to revise. Thus, structural robustness is substantially stronger than all the variants of the *ex post* Nash condition. (In the non-approximate notions, being *ex post* Nash is equivalent to being information proof. But in the approximate notions information proofness is substantially stronger.)

Remark 3: Invariance to order of play

G played sequentially, $\mathcal{GS}$, is the following *n*-person extensive game. The *n* players are assigned types as in *G*. The play progresses sequentially, according to a fixed publicly known order. Every player, at his turn, knows all earlier choices.

Clearly, $\mathcal{GS}$ is an alteration of *G*, and every equilibrium σ of *G* has the following $\mathcal{GS}$ adaptation: At his turn, every player *i* chooses a pure-strategy with the same probability distribution σ_i as he does in the simultaneous-move game *G*. Structural robustness of σ implies that this adaptation of σ must be an equilibrium in every such $\mathcal{GS}$.

Moreover, the above reasoning continues to hold even if the order of play is determined dynamically, and even if it is strategically controlled by *G*-players and outsiders. Thus, a structurally robust equilibrium is invariant to the order of play in a strong sense.

Remark 4: Invariance to revelation and delegation

G with delegation, $\mathcal{GD}$, is the following (*n*+1)-players game. The original *n* *G*-players are assigned types as in *G*. In a first round of simultaneous play, every *G*-player chooses between (1) self-play and (2) delegate-the-play and report a type to an outsider, player *n*+1. In a second round of simultaneous play all the self-players choose their own *G* pure strategies, and the outsider chooses a profile of *G* pure strategies for all the delegators. The payoffs of the *G* players are as in *G*; the outsider may be assigned any payoffs.

Clearly, $\mathcal{GD}$ is an alteration of *G*, and every equilibrium σ of *G* has adaptations that involve no delegation.

In the computer choice game, for example, consider an outsider with incentives to coordinate: his payoff equals one when he chooses the same computer for all delegators, zero otherwise. This alteration has a new (more efficient) equilibrium, not available in *G*: everybody delegates and the outsider chooses the most-reported type.

Nevertheless, as structural robustness implies, FC remains an equilibrium in $\mathcal{GD}$ (nobody delegates in the first round and they choose their favorite computers in the second). Moreover, FC remains an equilibrium under any scheme that involves reporting and voluntary delegation of choices.

Remark 5: Partially specified games

Structurally robust equilibria survive under significantly more complex alterations than the ones above. For example, one could have multiple opportunities to revise, to delegate, to affect the order of play, to communicate, and more. Because of these strong invariance properties, such equilibria may be used in games which are only partially specified as illustrated by the following example.

Example 3: A game played on the Web

Suppose that instead of being played in one simultaneous move, the Computer Choice game has the following instruction: ‘Go to Web site xyz before the end of the week, and click in your computer choice.’ This instruction involves substantial structural uncertainty: In what order would the players choose? Who can observe whom? Who can talk to whom? Can players sign binding agreements? Can players revise their choices? Can players delegate their choices? And so forth.

Because it is unaffected by the answers to such qsts, a structurally robust equilibrium σ of the one-simultaneous-move game can be played on the Web in a variety of ways without losing the equilibrium property. For example, players may make their choices according to their σ_i probabilities prior to the beginning of the click-in period, then go to the Web and click in their realized choices at individually selected times.

Remark 6: Competitive prices in Shapley–Shubik market games

For a simple illustration, consider the following *n*-trader market game (see Shapley and Shubik, 1977, and later references in Dubey and Geanakoplos, 2003, and McLean, Peck and Postlewaite, 2005). There are two fruits, apples and bananas, and a finite number of trader types. A type describes the fruit a player owns and the fruit he likes to consume. The players’ types are determined according to individual independent prior probability distributions. Each trader knows his own type, and his payoff depends on his own type and the fruit he ends up with, as well as on the distribution of types and fruit ownership of his opponents (externalities are allowed, for example, a player may wish to own the fruit that most opponents like). In one simultaneous move, every player has to choose between (1) keeping his fruit and (2) trading it for the other kind.

The banana/apple price is determined proportionately (with one apple and banana added in to avoid division by zero). For example, if 199 bananas and 99 apples are traded, the price of bananas to apples would be (199+1)/(99+1)=2, that is, every traded apple brings back two bananas and every traded banana brings back 0.5 apples.

With a small number of traders, the price is unlikely to be competitive. If players are allowed to re-trade after the realized price becomes known, they would, and a new price would emerge.

However, when *n* is large, approximate structural robustness implies being approximately information-proof. So even when the realized price becomes known, no player has significant incentive to re-trade,

that is, the price is approximately competitive (Prob[some player can ε -improve his expected payoff by re-trading at the observed price] $\leq \rho$).

This is stronger than classical results relating Nash equilibrium to Walras equilibrium (for example, Dubey, Mas-Colell and Shubik, 1980). First, being conducted under incomplete information, the above relates Bayesian equilibria to rational expectations equilibria (rather than Walras). Also the competitive property described here is substantially stronger, due to the immunity of the equilibria to alterations represented by extensive games. If allowance is made for spot markets, coordinating institutions, trade on the Web, and so on, the Nash-equilibrium prices of the simple simultaneous-move game are sustained through the intermediary steps that may come up under such possibilities.

Remark 7: Embedding a game in bigger worlds

Alterations allow the inclusion of outside players who are not from *G*. Moreover, the restrictions imposed on the strategies and payoffs of the outsiders are quite limited. This means that alterations may describe bigger worlds in which *G* is embedded. Structural robustness of an equilibrium means that the small-world (*G*) equilibrium remains an equilibrium even when the game is embedded in such bigger worlds.

Remark 8: Self-purification

Schmeidler (1973) shows that in a normal-form game with a continuum of anonymous players, every strategy can be purified, that is, for every mixed-strategy equilibrium one can construct a pure-strategy equilibrium (Ali Khan and Sun, 2002 survey some of the large follow-up literature).

The *ex post* Nash property above constitutes a stronger (but asymptotic) result. Since the resulting play of a mixed strategy equilibrium yields pure-strategy profiles that are Nash equilibria (of the perfect information game), one does not need to construct pure-strategy equilibria: simply playing a mixed-strategy equilibrium yields pure-strategy profiles that are equilibria.

The approximate statement is: for every (ε , ρ) for sufficiently large *n*, Prob[ending at a pure strategy profile that is not an ε Nash equilibrium of the realized perfect information game] $\leq \rho$. Since both ε and ρ can be made arbitrarily small, this is asymptotic purification. Note that the model of Schmeidler, with a continuum of players, requires non-standard techniques to describe a continuum of independent random variables (the mixed strategies of the players). The asymptotic result stated here, dealing always with finitely many players, does not require any non-standard techniques.

Remark 9: 'As if' learning

Kalai and Lehrer (1993) show that in playing an equilibrium of a Bayesian repeated game, after a sufficiently long time the players best-respond as if they know their opponents' realized types and, hence, their mixed strategies.

But being information-proof, at a structurally robust equilibrium (even of a one shot game) players' best respond (immediately) as if they know their opponents' realized types, their mixed strategies and even the pure-strategies they end up with.

Sufficient conditions for structural robustness

Theorem 1: Structural Robustness (rough statement): the equilibria of large one-simultaneous-move Bayesian games are (approximately) structurally robust if

1. (a) the players' types are drawn independently, and
2. (b) payoff functions are anonymous and continuous.

Payoff anonymity means that in addition to his own type and pure-strategy, every player's payoff may depend only on aggregate data of the opponents' types and pure-strategies. For example, in the computer choice game a player's payoff may depend on her own type and choice, and on the *proportions* of opponents in the four groups: *I*-types who chose *I*, *I*-types who chose *M*, *M*-types who chose *I*, and *M*-types who chose *M*.

The players in the games above are only semi-anonymous, because there are no additional symmetry or anonymity restrictions other than the restriction above. In particular, players may have different individual payoff functions and different prior probabilities (publicly known).

The continuity condition relates games of different sizes and rules out games of the type below.

Example 4: Match the expert

Each of *n* players has to choose one of two computers, *I* or *M*. Player 1 is equally likely to be one of two types: 'an expert who is informed that *I* is better' (*I*-better) or 'an expert who is informed that *M* is better' (*M*-better). Players 2, … ,*n* are of one possible 'non-expert' type. Every player's payoff is one if he chooses the better computer, zero otherwise. (Stated anonymously: choosing computer *X* pays one, if the proportion of the *X-better* type is positive, zero otherwise.)

Consider the equilibrium where player 1 chooses the computer he was told was better and every other player chooses *I* or *M* with equal probabilities. This equilibrium fails to be *ex post* Nash (and hence, fails structural robustness), especially as *n* becomes large, because after the play approximately one-half of the players would want to revise their choices to match the observed choice of player 1. (With a small *n* there may be 'accidental *ex post* Nash', but it becomes extremely unlikely as *n* becomes large.)

This failure is due to discontinuity of the payoff functions. The proportions of *I*-better types and *M*-better types in this game must be either $(1/n, 0)$ or $(0, 1/n)$, because only one of the *n* players is to be one of these types. Yet, whatever *n* is, every player's payoff is drastically affected (from 0 to 1 or from 1 to 0) when we switch from $(1/n, 0)$ to $(0, 1/n)$ (keeping everything else the same).

As *n* becomes large, this change in the type proportions becomes arbitrarily small, yet it continues to have a drastic effect on players' payoffs. This violates a condition of uniform equicontinuity imposed simultaneously on all the payoff functions in the games with $n=1, 2, \ldots$ players.

See Also

- Internet, economics of the
- large economies
- purification
- rational expectations

Bibliography

Aumann, R.J. and Shapley, L.S. 1974. *Values of Nonatomic Games*. Princeton, NJ: Princeton University Press.

Bergemann, D. and Morris, S. 2005. Robust mechanism design. *Econometrica* 73, 1771–813.

Cremer, J. and McLean, R.P. 1985. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica* 53, 345–61.

Dubey, P., Mas-Colell, A. and Shubik, M. 1980. Efficiency properties of strategic market games: an axiomatic approach. *Journal of Economic Theory* 22, 339–62.

Dubey, P. and Kaneko, M. 1984. Information patterns and Nash equilibria in extensive games: 1. *Mathematical Social Sciences* 8, 111–39.

Dubey, P. and Geanakoplos, J. 2003. From Nash to Walras via Shapley–Shubik. *Journal of Mathematical Economics* 39, 391–400.

Green, J.R. and Laffont, J.J. 1987. Posterior implementability in a two-person decision problem. *Econometrica* 55, 69–94.

Kalai, E. 2004. Large robust games. *Econometrica* 72, 1631–66.

Kalai, E. 2005. Partially-specified large games. *Lecture Notes in Computer Science* 3828, 3–13.

Kalai, E. and Lehrer, E. 1993. Rational learning leads to Nash equilibrium. *Econometrica* 61, 1019–45.

Khan, A.M. and Sun, Y. 2002. Non-cooperative games with many players. In *Handbook of Game Theory with Economic Applications*, vol. 3, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

McLean, R., Peck, J. and Postlewaite, A. 2005. On price-taking behavior in asymmetric information economies. In *Essays in Dynamic General Equilibrium: Festschrift for David Cass*, ed. A. Citanna, J. Donaldson, H. Polemarchakis, P. Siconolfi and S. Spear. Berlin: Springer. Repr. in *Studies in Economic Theory* 20, 129–42.

Rustichini, A., Satterthwaite, M.A. and Williams, S.R. 1994. Convergence to efficiency in a simple market with incomplete information. *Econometrica* 62, 1041–64.

Schmeidler, D. 1973. Equilibrium points of nonatomic games. *Journal of Statistical Physics* 17, 295–300.

Shapley, L.S. and Shubik, M. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 85, 937–68.

Wilson, R. 1987. Game-theoretic analyses of trading processes. In *Advances in Economic Theory: Fifth World Congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

How to cite this article

Kalai, Ehud. "large games (structural robustness)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000239> doi:10.1057/9780230226203.0933

The New Palgrave Dictionary of Economics Online

Laspeyres, Ernst Louis Etienne (1834 – 1913)

W.E. Diewert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Drobisch price index; index numbers; Laspeyres index; Laspeyres, E. L. E.

Article

Laspeyres was born at Halle, Germany, on 28 November 1834 and died on 4 August 1913 at Giessen, Germany.
From 1853 to 1857, he studied at the universities of Tübingen, Berlin, Göttingen and Halle. He received a law degree from the University of Halle in 1857. He studied at the University of Heidelberg from 1857 to 1859, and in 1860 he obtained his Ph.D. from Heidelberg for the thesis, ‘The Correlation between Population Growth and Wages’.
From 1860 until 1864 he worked as a lecturer at Heidelberg, where he wrote a history of the economic views of the Dutch (1863). In the following ten years, he taught at four different universities: 1864 – Basel; 1866 – the Polytechnic at Riga; 1869 – Dorpat; 1873 – Karlsruhe. Finally, from 1874 to 1900, he taught at the Justus-Liebig University at Giessen.
Laspeyres’ main contribution to economics was his development of the index number formula that bears his name. Let the price and quantity of commodity *n* in period *t* be p_n^t and q_n^t respectively for $n = 1, \ldots, N$ and $t = 0, 1, \ldots, T$. Then the Laspeyres price index of the *N* commodities for period *t* (relative to the base period 0) is defined as

$$P_L \equiv \sum_{n=1}^{N} p_n^t q_n^0 \Big/ \sum_{n=1}^{N} p_n^0 q_n^0.$$

Laspeyres wrote his classic paper (1871), which suggested the above formula partly as an outgrowth of

his empirical work on measuring price movements in Germany and partly to criticize the index number formula of Drobisch (1871). Using the notation defined above, the Drobisch price index for period *t* is defined as

$$P_D \equiv \left(\sum_{n=1}^{N} p_n^t q_n^t \Big/ \sum_{n=1}^{n} q_n^t\right) \Big/ \left(\sum_{n=1}^{N} p_n^0 q_n^0 \Big/ \sum_{n=1}^{N} q_n^0\right).$$

Laspeyres criticized this formula by showing that the index generally changed even if all prices remained constant (that is, P_D does not satisfy an identity test, to use modern terminology). An even more effective criticism of P_D is that it is not invariant to changes in the units of measurement (whereas P_L is invariant).

Laspeyres did not write any further papers on index number theory. He wrote papers on economic history, the history of economic thought and on topical economic issues of his time; see Rinne (1981).

Selected works

1863. *Geschichte der volkswirtschaftlichen Anschauungen der Niederländer und ihrer Literatur zur Zeit der Republik*. Leipzig.

1871. Die Berechnung einer mittleren Waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik* 16, 296–315.

Bibliography

Drobisch, M.W. 1871. Über die Berechnung der Veränderungen der Waarenpreise und des Geldswerths. *Jahrbücher für Nationalökonomie und Statistik* 16, 143–56.

Rinne, H. 1981. Ernst Louis Etienne Laspeyres 1834–1913. *Jahrböcher für Nationalökonomie und Statistik* 196, 194–216.

How to cite this article

Diewert, W. E. "Laspeyres, Ernst Louis Etienne (1834–1913)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000028> doi:10.1057/9780230226203.0934

The New Palgrave Dictionary of Economics Online

Lassalle, Ferdinand (1825 – 1864)

Tom Bottomore
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cooperative production; Engels, F.; iron law of wages; Lassalle, F.; Malthus's theory of population; Marx, K. H.; Socialism

Article

Born in Breslau, 13 April 1825; died in Geneva, 31 August 1864. The only son of a prosperous Jewish silk merchant, Lassalle studied philosophy and history at the University of Breslau and subsequently at the University of Berlin, where he encountered the radical ideas of the ‘Young Hegelians’ and of the French socialist thinkers. During the 1848 revolution he was associated with Marx and the *Neue Rheinische Zeitung*, and was arrested for his activities but acquitted by a jury in 1849. In the course of his short and turbulent life (which ended as a result of an absurd duel with the former fiancé of a woman he wished to marry), Lassalle became known primarily as a political and economic theorist, and as a leading figure in the radical and working-class movements, who organized in 1863 the first socialist party in Germany (the General Union of German Workers).

Lassalle's economic ideas were derived to a large extent from Marx, often without acknowledgement, but he diverged from the latter in important respects. As Bernstein (1891) observed: ‘Lassalle was much more indebted to Marx than he admitted in his writings, but he was a disciple of Marx only in a restricted sense.’ The main divergence can be summarized as the substitution of an evolutionary conception of the movement from capitalism to socialism for Marx's idea of a revolutionary transition. In his ‘Workers’ Programme’ (1862) and his ‘Open Letter’ (1863), Lassalle advocated a course of political action for the working-class movement with two principal aims: first, the achievement of universal and equal suffrage; second, the development, with state aid, of workers’ cooperatives that would lead to a gradual socialization of the economy. His reliance upon the action of the state (conceived in the manner of Hegel rather than Marx) was very great, and in the ‘Open Letter’ he adduced an ‘iron law of wages’, derived from classical political economy, to show that neither individually nor collectively could workers improve their conditions of life except by replacing the wage system with self-employment (cooperative production), for which the necessary capital must be

provided by the state. It was in this context that Lassalle responded to Bismarck's invitation (11 May 1863) to express his views on ‘working class conditions and problems’ and subsequently had several meetings with him; a course of action which Engels (letter to Kautsky, 23 February 1891) later assessed harshly as a step towards allying the workers’ movement with German nationalism and the monarchy. Marx had a low opinion of Lassalle's abilities as an economist and political thinker, and in his *Critique of the Gotha Programme* (1875) on the occasion of the unification of the two existing German workers’ parties (the Social Democratic Workers’ Party and the General Union of German Workers) he strongly criticized the Lassallean ideas which were embodied in the draft programme; in particular, the erroneous restriction of ownership of the instruments of labour to the capitalist class, excluding landowners, and the confused notion of an ‘iron law of wages’, which is simply, Marx argued, ‘the Malthusian theory of population’.

Selected works

1919–20. *Gesammelte Reden und Schriften*, 12 vols, edited with an Introduction by E. Bernstein. Berlin: Paul Cassirer.

Bibliography

Bernstein, E. 1891. *Ferdinand Lassalle as a Social Reformer*. London: Swan Sonnenschein, 1893. Reprinted, New York: Greenwood Press, 1969.

Footman, D. 1946. *The Primrose Path: A Life of Ferdinand Lassalle*. London: Cresset Press.

How to cite this article

Bottomore, Tom. "Lassalle, Ferdinand (1825–1864)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000029> doi:10.1057/9780230226203.0935

The New Palgrave Dictionary of Economics Online

Lauderdale, Eighth Earl of [James Maitland] (1759 – 1839)

Morton Paglin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Böhm-Bawerk, E. von; consumer choice; Corn Laws, free trade and protectionism; division of labour; labour theory of value; Lauderdale, Eighth Earl of; Malthus, T. R.; monopoly; natural harmony of interests; over-investment; parsimony; period of production; profit and profit theory; Ricardo, D.; saving and investment; sinking fund; Smith, A.; total output theory; value

Article

Born into a Scottish aristocratic family, Lauderdale entered the House of Commons at the age of 21 as a supporter of the Liberal Whig leader Charles Fox. Following the death of his father, he entered the House of Lords in 1790, where he became known for his defence of civil liberties. After a visit to France in 1792 he publicly expressed sympathy for the ideals of the French Revolution and supported a motion in Parliament (1795) to make peace with the new government of France. In his middle years he swung over to the Tory side and adamantly opposed most economic and political reform measures, especially bills to protect labour (even one which would restrict the use of young children in cleaning chimney flues). His views covered the political spectrum: in 1792 he flirted with Jacobinism, becoming a founding member of the Friends of the People; 40 years later he worked against the Reform Bill of 1832. He died in 1839 at 80, a ripe age indeed for a man known for his apoplectic temper.

Lauderdale had a sustained interest in trade policy, but here he also shifted ground. In 1804 he argued ‘that all impediments thrown in the way of commercial communication, obstruct the increase of wealth’ (1804, p. 365). Yet in his pamphlet *A Letter on the Corn Laws* (1814) he claimed Adam Smith was in error, and advocated protection for agriculture, a position which he strongly held in the House of Lords for some 20 years.

Apart from some tracts on currency questions and debt policy, Lauderdale's contributions to economic thought are found in one major work, *An Inquiry into the Nature and Origin of Public Wealth* (1804). A second edition (1819) contained only minor revisions. This suggests that Lauderdale's involvement with economic theory was a one-time affair. The intellectual ferment generated by Ricardo's *Principles* (1st and 2nd editions), and the earlier tracts by Malthus, Edward West, and Ricardo on rent and profits seems

to have passed him by: no mention of his contemporaries or the theoretical issues which they raised appeared in his new introduction or in the footnotes to the 1819 edition. The focus of both editions is the *Wealth of Nations*, and a large part of the *Inquiry* is given over to a negation of Smith's conclusions. Specifically, Lauderdale asserts that: (1) the maximization of private riches does not lead to maximum public wealth and welfare; (2) labour is not the cause of value or an adequate measure of value; (3) division of labour is not a major factor in economic growth; (4) parsimony and saving are frequently a public detriment as they may lead to over-investment and a capital glut; and (5) government tax revenues applied to rapid debt reduction (‘a forced conversion of revenue into capital’) will reduce aggregate consumption, deflate profits and capital values, and result in economic distress.

In developing these ideas Lauderdale exposes his deficiencies as a thinker. His analysis is sketchy, his style prolix and repetitious, and his conclusions based on weak or incomplete reasoning occasionally seem pretentious. Not surprisingly, his contemporaries focused on these flaws. Henry Brougham wrote a long very critical commentary on Lauderdale's *Inquiry* in the *Edinburgh Review* (July 1804), to which Lauderdale responded with an acerbic but not too effective pamphlet. Ricardo exposed several of his logical errors (Ricardo, 1823, pp. 267–77, 37ln., 384–5), and Malthus, who on a number of issues (capital glut, value theory and agricultural protectionism) was his intellectual heir, failed to acknowledge his intellectual debt; instead, he accused Lauderdale of ‘going too far’ in his condemnation of parsimony and savings (Malthus, 1836, p. 314), even though, as we shall see, their arguments were quite similar.

Despite the negative opinions of his contemporaries, and his modest theoretical ability, Lauderdale now occupies a firm, albeit secondary, place in the history of economic doctrine. We may ask why.

The answer I believe lies in the fact that Lauderdale had a number of valuable insights into the workings of the economy which later economists thought important. Böhm-Bawerk considered Lauderdale's theory of profit a limited but significant step towards the true and complete explanation of interest and profit (that is, his own theory). Following the appearance of Keynes's *General Theory* there was a re-examination of earlier writers who might have anticipated Keynesian ideas on saving, investment and employment. Malthus obviously was placed in the centre of this pantheon of economists, and Lauderdale as an earlier thinker espousing similar ideas was accorded lesser status. This is not a wholly satisfactory way of evaluating past intellectual contributions, but there is no doubt that each age searches for harmonious resonances in the historical literature. Here I shall try to broaden the perspective.

In the *Inquiry* Lauderdale challenged the natural harmony of interests propounded by Smith; namely, that individuals seeking private riches would lead a nation to maximize public wealth. To destroy this identity, Lauderdale tried to prove that the sum of private riches could increase while public wealth and welfare declined. Unfortunately, Lauderdale obfuscated the problem by treating the individual riches occasionally produced by monopoly or a sudden scarcity of supply as a *net* addition to aggregate riches when it was clear that Adam Smith meant aggregate riches in real terms, so the scarcity-induced gains of some are more than offset by real losses of others. Furthermore, Lauderdale overlooked Smith's postulate of free competition as a necessary condition for the coincidence of private and public interest. Ricardo came to Smith's defence and cleared up Lauderdale's ten pages of confusion in a couple of succinct paragraphs (Ricardo, 1823, p. 276).

But something positive came out of Lauderdale's discussion of value and riches. His examination of the effect of monopoly on total revenue led to an early and fairly sophisticated discussion of demand curves. Lauderdale reviews empirical estimates of the relationship between a percentage change in the price of a good and the percentage change in the quantity demanded, and notes that for various kinds of consumer

goods elasticities may differ. In addition to the concept of price elasticity, Lauderdale gave us the beginnings of a theory of consumer choice, noting the utility sacrifices involved in giving up alternative bundles of goods when consumers make new choices in response to price changes (1804, pp. 59–86). Not surprisingly, Lauderdale rejected the labour theory of value, both as a cause of value and a measure of value (1804, p. 12). Although he related consumer preferences to demand, and was aware of demand in the schedule sense, he failed to relate costs to supply, and hence his theory of value suffered the inadequacies of all the early supply and demand theories, a weakness which Ricardo pointed out (1823, pp. 384–5).

We now come to the section of the *Inquiry* which has been of most interest in the post-Keynesian period: that dealing with saving, investment and fiscal policy. Lauderdale argues that the social benefits from savings have distinct limits: ‘In every state of society, a certain quantity of capital, proportioned to the existing state of knowledge of mankind, may be usefully and profitably employed.’ Invention may enlarge the scope for the application of capital, but outlets for profitable investment are still limited by the demand for consumer goods (Lauderdale, 1804, p. 227).

Individual parsimony may be misguided, but the harm it does tends to be offset by the prodigality of others. However, when a belief in parsimony leads to bad legislation such as a mandated sinking fund, which forces an increase in public parsimony through taxation and debt reduction, then the results may be ‘fatal to the progress of wealth’ (Lauderdale, 1804, pp. 228–30, 271). But there remains the question of what is the mechanism by which high savings rates or forced parsimony become ‘fatal to the progress of wealth’. Superficially this discussion of the evils of parsimony has a Keynesian air to it, but actually Lauderdale (and Malthus) go on to describe a situation in which savings *are* invested, and it is over-investment relative to restricted consumption (made lower by taxation) which finally produces a collapse in profitability.

It is noteworthy that both writers developed a model in which productive applications of net additions to the capital stock are dependent on increases in consumption. They both also failed to recognize that for long periods a nation can use part of its investment for further investment – a deepening of the capital structure or, in Böhm-Bawerk's terms, a lengthening of the period of production, certainly an attribute of 19th-century capitalism. Whatever their limitations, it seems clear that the macroeconomic contributions of Lauderdale and Malthus are more closely related to the growth models of the Harrod–Domar type than to a short-run Keynesian analysis in which output drops because savings are *not* invested.

Nevertheless, there is a tenuous connection with Keynes when we look at their descriptions of the late phase of the over-investment cycle. For Lauderdale over-investment reduces profits and the value of capital, and the resulting low prices ‘discourage reproduction’. When we observe such deflation we ‘must be cautious not to mistake for the effects of abundance that which in reality may be only the effect of failure of demand’ (Lauderdale, 1804, pp. 263–4). Malthus wrote in a similar vein when he pointed to owners of floating capital vainly seeking investment outlets in the glutted capital markets of Europe (Malthus, 1836, p. 420).

We may conclude that the Lauderdale–Malthus theory of total output was not for the most part in the Keynesian mould, but surely that is no reason to downgrade it. Both men saw defects in the Smith–Say–Ricardo theory of total output and employment, and they recognized that restricted consumption and high rates of saving and investment could lead to a sectoral imbalance – a glut of capital, falling profits and, finally, a drop in the inducement to invest. In the policy arena, Lauderdale used these insights to oppose tax surpluses and debt reduction in a period of recession (Paglin, 1961, pp. 98–107; Lauderdale,

1829).

See Also

- Malthus, Thomas Robert

Selected works

1797. *Thoughts on Finance*. 3rd edn. London: G.G. & J. Robinson.

1804. *An Inquiry into the Nature and Origin of Public Wealth*. Edinburgh. Reprinted, New York: Augustus M. Kelley, 1962.

1814. *A Letter on the Corn Laws*. Edinburgh: A. Constable.

1829. *Three Letters to the Duke of Wellington*. London: J. Murray.

Bibliography

Malthus, T.R. 1836. *Principles of Political Economy Considered with a View to Their Practical Application*. 2nd edn. London: William Pickering. Reprinted, New York: Augustus Kelley, 1964.

Paglin, M. 1961. *Malthus and Lauderdale: The Anti-Ricardian Tradition*. New York: Augustus Kelley.

Ricardo, D. 1823. *Principles of Political Economy and Taxation*. Ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

How to cite this article

Paglin, Morton. "Lauderdale, Eighth Earl of [James Maitland] (1759–1839)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000032> doi:10.1057/9780230226203.0936

The New Palgrave Dictionary of Economics Online

Laughlin, James Laurence (1850 – 1933)

Milton Friedman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost-push inflation; free silver; Laughlin, J. L.; quantity theory of money; Veblen, T.

Article

Scholar, teacher, monetary reformer and university administrator, Laughlin was born in Deerfield, Ohio of middle-class parents of modest means. A scholarship plus outside work, largely tutoring, enabled him to attend Harvard. After completing his undergraduate study in history, he did graduate work under Henry Adams, receiving a Ph.D. for a thesis on ‘The Anglo-Saxon Legal Procedure’. His subsequent academic career, however, was entirely in economics.

From 1878 to 1888 he taught at Harvard, from 1888 to 1890 he was successively Secretary and President of the Philadelphia Manufacturers’ Mutual Fire Insurance Company, from 1890 to 1892 Professor of Political Economy and Finance at Cornell, and in 1892 was persuaded by President Harper to become Head Professor of Political Economy at the new University of Chicago, the position he held until he retired in 1916. From 1916 until his death in 1933 he continued his scientific writing and public activities.

Laughlin's scholarly work was almost entirely in the field of money and banking. Much of it, notably his *History of Bimetallism in the United States* (1885), consisted of a thorough and extremely careful presentation of historical evidence on the development of money and monetary institutions. But Laughlin also wrote extensively on monetary and banking theory, and on proposals for monetary reform. His work on these topics was marred by a dogmatic and rigid opposition to the quantity theory of money, an opposition that developed out of his public activities opposing the free silver movement. The proponents of free silver used a crude form of the quantity theory to support their position, which sufficed to render the theory anathema to Laughlin.

Laughlin's attack on the quantity theory had much in common with recent cost-push or structural or supply shock theories of inflation, in emphasizing the role of factors affecting specific goods and services rather than general monetary influences. Then, as now, such theories ran against the major stream of monetary analysis as exemplified in Laughlin's time by the work of Irving Fisher. As a result,

his writings on theory have had no lasting influence on economic thought.
According to Wesley C. Mitchell, one of his students,

> Professor Laughlin's indubitable success as a teacher puzzled many who did not pass through his classroom. He was not an original thinker of great power. He did not enrich economics … . He did not even keep abreast of current developments in economic theory … . He had a prim and tidy mind, which he kept in perfect order by admitting nothing that did not harmonize with the furnishings installed in the 1880's … . Yet he held that a teacher's aim should be ‘the acquisition of independent power and methods of work, rather than specific beliefs’.
>
> The very limitations I have listed helped Professor Laughlin to accomplish this aim. … [His] honesty of purpose impelled others to be honest, which meant that doubting students had to work out the reasons for their dissent … . Laughlin forced one to face intellectual conflicts in his own mind and find out where he stood in the world of ideas. That, I have long believed, was the secret of his success in helping so many students of such diverse capacities to make the most of their several gifts. (1941, pp. 879–80)

As monetary reformer, Laughlin was a leading opponent of the advocates of free silver. He wrote, lectured, and campaigned extensively in favour of ‘hard money’. In his widely circulated free-silver pamphlet, *Coin's Financial School*, William Hope Harvey used Laughlin as a hard-money foil for the fictional Coin's free-silver argument. That episode terminated in a widely reported public debate in Chicago in 1885 between Laughlin and Harvey.
After the defeat of William Jennings Bryan and the free-silver forces in the presidential election of 1896, financial and commercial interests in the country organized the Indianapolis Monetary Commission to develop proposals for reform of the monetary and banking system. One of the 11 members of the commission, Laughlin was also the author of its extensive final report, which served as an important stepping-stone en route to the Aldrich–Vreeland Act of 1908 and the Federal Reserve Act of 1913. In addition, Laughlin served for nearly two years from 1911 to 1913, on leave from the University of Chicago, as full-time chairman of the executive committee of the National Citizens League, an organization formed to mobilize public opinion in favour of banking reform.
Laughlin's close links with the Republican Party prevented him from playing any public role in the final preparation of the Federal Reserve Act under a Democratic administration. However, he exerted considerable influence behind the scenes through extensive private correspondence with his former student and assistant, H. Parker Willis, who, as banking expert for the House Banking and Currency Committee, has been regarded as primarily responsible for drafting the Act.
Laughlin's most important and lasting contribution was as head of the Department of Political Economy of the new University of Chicago. Though himself a hard-money man of rigidly conservative views, he demonstrated an extraordinary degree of tolerance for divergent views in staffing and guiding the department. At the very outset, he brought with him from Cornell Thorstein Veblen, who remained in the department for 14 years, the longest period Veblen spent at any single university during his stormy career. Veblen served as managing editor of the *Journal of Political Economy*, which Laughlin founded as one of his first acts at Chicago. Laughlin himself was the editor. As John U. Nef wrote in his obituary

notice of Laughlin, ‘his wide cultural interests combined with his other qualities to enable him to gather about him a more remarkable group of younger men than was to be found in any other economics department in the country and to help these men in making the most of their own gifts.’ Nef notes that

> a very considerable portion of all the men who have made an important mark in economic thought between 1895 and 1930, beginning with Thorstein Veblen and coming down to Jacob Viner (Laughlin's last appointment) were connected at one time or another, as members or students, with the department of political economy. … Laughlin frequently chose the best men when they were of very different persuasions from his own. … And so it came about that one of the most conservative heads of an economics department in the country had politically the most liberal and economically the least orthodox department. (1934, p. 2)

Laughlin's emphasis on quality rather than ideology was combined with an emphasis on research by his faculty, as well as by graduate students as part of their training. A corollary was his belief in personal teaching as opposed to formal lecturing. These have remained key characteristics of the Chicago Department of Economics from that day to this. In more recent years, as in his day, the department has been widely regarded as a stronghold of proponents of a free-market economy. That reputation was justified in the sense that throughout the period the department had prominent members who held these views and presented them effectively. But they were always a minority. The department has been characterized by heterogeneity of policy views, not homogeneity. The economists at Chicago who held the generally fashionable views – who were ‘liberal’ in the 20th-century sense – could be matched at other institutions; the ones who were ‘liberal’ in the 19th-century sense could not be. That, plus the emphasis on economics as a serious scientific subject, capable of being tested by empirical and historical evidence, and of being used to illuminate important practical issues of conduct and policy, made Chicago economics unique. These were Laughlin's bequest to the department he built.

Selected works

1885. *The History of Bimetallism in the United States*. New York: D. Appleton & Co.

1898. *Report of the Monetary Commission of the Indianapolis Convention of Boards of Trade, Chambers of Commerce, Commercial Clubs, and Other Similar Bodies of the United States*. Chicago: University of Chicago Press.

1903. *The Principles of Money*. New York: Charles Scribner's Sons.

1909. *Latter Day Problems*. New York: Charles Scribner's Sons.

1931. *A New Exposition of Money, Credit and Prices*, 2 vols. Chicago: University of Chicago Press.

1933. *The Federal Reserve Act: Its Origin and Problems*. New York: Macmillan.

Bibliography

Bornemann, A. 1940. *J. Laurence Laughlin*. Washington, DC: American Council on Public Affairs.

Mitchell, W.C. 1941. J. Laurence Laughlin. *Journal of Political Economy* 49, 875–81.

Nef, J.U. 1934. James Laurence Laughlin (1850–1933). *Journal of Political Economy* 42, 1–5.

How to cite this article

Friedman, Milton. "Laughlin, James Laurence (1850–1933)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000033> doi:10.1057/9780230226203.0937

The New Palgrave Dictionary of Economics Online

Launhardt, Carl Friedrich Wilhelm (1832 – 1918)

Jürg Niehans
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Banking School; Exchange; Hotelling, H.; investment criteria; Jevons, W. S.; Launhardt, C. F. W.; marginal cost pricing; marginal revolution; marginal utility of money; monopoly; optimal tariffs; price discrimination; quantity theory of money; rate of interest; rent; transportation economics; wage differentials; Walras, L.

Article

Launhardt was born on 4 April 1832 in Hannover, where he died on 14 May 1918. His work is Germany's most important and in fact only significant contribution to the ‘marginal revolution’ in the last three decades of the 19th century. In the economic analysis of transportation and location, this contribution was not surpassed until the 1930s. Available only in German, some of it in publications that are hard to find, it still has not found the recognition it deserves, and Schumpeter's references in the *History of Economic Analysis* are inadequate.

Like Dupuit, Launhardt began his professional life as a civil engineer, working for the public road administration. In 1869 he joined the faculty of the Hannover Polytechnic Institute as a professor for roads, railways and bridges. This was the beginning of a distinguished academic career, in the course of which he served as the director of the institute and, when it became the Technische Hochschule Hannover, its first rector. He was made a member of the Königliche Akademie des Bauwesens and of the Preussische Herrenhaus. Dresden gave him an honorary degree for his contributions to the technology and economics of transportation.

Practical problems of highway planning led Launhardt to the gradually more general analysis of efficient transportation networks. This work was later systematized in *Theorie des Trassirens* (Theory of Network Planning). Part I, entitled ‘Commercial Network Planning’, contains the derivation of efficiency criteria without regard to topography. This part is the second edition, much revised and enlarged, of the 1872 publication, and also incorporates sections from the 1885 book. Part II, entitled ‘Technical Network Planning for Railroads’, applies economic efficiency criteria to curves and gradients imposed by topography; an earlier version was published in 1877.

The contributions to economics are found in Part I. This begins with a discussion of investment criteria. From a social point of view, networks should be planned in such a way that the sum of operating and capital costs is a minimum. Private capitalists, however, try to maximize the internal rate of return on their capital. Under perfect competition the two criteria would coincide, since the internal rate of return, if duly maximized, would equal the market rate of interest. In reality, however, since the railroad industry is inherently non-competitive, rates of return can be pushed above market rates of interest by keeping railroad investment below the social optimum. This was one of Launhardt's basic arguments for government ownership of railroads. For his own analysis he uses, of course, the social criterion.

Using geometry and calculus, Launhardt derives rules, depending on freight costs and volumes, for the optimal direction and density of highways connecting given market centres. He shows that highways of different quality (and thus with different freight costs) should meet at angles analogous to those of refracted light, a rule later popularized by Stackelberg as the ‘law of refraction’. According to the ‘law of nodes’, transport costs on a star-shaped transportation network connecting three cities are minimized if the sines of the angles between its rays bear the same proportions as the total transportation costs per mile along the rays. The efficient combination of different modes, like highways, waterways and railways, is also considered.

Applying his analysis of network nodes to the location of plants, Launhardt produced the first substantial theory of industrial location (1882). In this basic contribution he determines the efficient location of a plant with given sources of supplies and given sales outlets by minimizing transportation costs. The optimum is found by an ingenious geometrical construction which became known as the ‘pole principle’, later amplified by Palander. It is given a mechanical interpretation as the centre of gravity of forces, representing freight rates, acting at the different input and output locations. After first assuming that the network of routes is being planned from scratch, Launhardt also derives rules for optimal additions to existing networks. The analysis is far superior to that in Alfred Weber's later book on the location of industries, in which Launhardt is not mentioned, and whose only claim to attention is the appendix by Georg Pick.

Launhardt's main contribution to the theory of railway rates is found in chapter 32 of (1885). It was elaborated in (1887) and further detail was added in (1890a) and (1890b), but these extensions add nothing for more general economic interest. The paper on ‘Economic Problems of the Railway Industry’ provides an extensive analysis, based on consumer surplus, of the social rate of return of railroads, both theoretical and numerical, including a cost–benefit analysis of future railway development.

For railway rates, Launhardt establishes the principle that the maximization of social welfare requires – in modern terminology – marginal cost pricing. But this, in turn, requires competition, while profit maximization by monopolistic railway firms implies that rates exceed marginal cost. In particular, if a railway transports homogeneous goods from a uniform plain to a market centre, the monopoly price is calculated to exceed marginal cost by 50 per cent (because, in modern terminology, freight volume reacts to the freight rate with an elasticity of – 2 and ton-miles thus with an elasticity of – 3). As a consequence, the freight volume is suboptimal. By perfect discrimination according to ‘what the traffic will bear’ over each distance, both railway profits and general welfare can be increased compared with simple monopoly. This, however, is only a second-best solution. For Launhardt, the efficiency of marginal cost pricing is another basic argument for government ownership.

Launhardt's main claim to a prominent place in the history of economic analysis is his slender treatise *Mathematische Begründrung der Volkswirtschaftslehre* (Mathematical Foundations of Economics) of

1885. It was written in the light of Walras's *Mathematische Theorie der Preisbestimmung wirtschaftlicher Güter* (1881) and the second edition of Jevons's *Theory of Political Economy* (1879). At the same time, it is clearly pre-Marshall and pre-Edgeworth (though *Mathematical Psychics* had appeared in 1881). Two other books by Walras, sent by the author, arrived too late to be of use, nor was Launhardt acquainted with Cournot at that time. He reports that the copy he finally obtained from a library had apparently never been read, and Gossen could nowhere be found (because virtually no copies had been sold). Launhardt shows what a competent engineer with an economic turn of mind and a little calculus could do (and also what he could not do) in economics a hundred years ago. Launhardt's addiction to special functional forms, particularly quadratic utility functions, often results in spurious precision, limited generality and reduced lucidity, but the basic contributions are sound, important and original.

In his theory of exchange (Part I), Launhardt rightly criticizes Walras for believing (if taken literally) that there is no way for a trader to improve his position relative to free competition at uniform prices. His counter-examples relate to monopoly and price discrimination, leading him to the idea of an optimal tariff. While valid in principle, this analysis falls short of Edgeworth's. The discussion of the total gain from trade and its distribution, whose shortcomings were pointed out by Wicksell, was soon obsolete because of its dependence on the interpersonal additivity of utility.

In his discussion of distributive shares, Launhardt recognizes the backward-bending supply curve of labour and the effect of property incomes on labour supply and thus on wages. He also recognizes that the inter-occupational mobility of labour tends to equalize relative wage rates with both the ratios of the marginal products of labour and (to the extent an individual can choose between occupations) the ratio of its marginal disutilities. For profits, Launhardt's ‘basic equation’ expresses, substantially, the familiar optimality condition that the profit margin, as a percentage of price, is the inverse of the elasticity of demand (though this concept is not used, of course). It is clearly explained that the entrepreneur, in setting his price, considers only marginal costs, while prices are equalized to the average costs of the marginal firm by exit and entry. The profits of intra-marginal firms are correctly interpreted as rents, and the same principle is used to explain wage differentials.

Launhardt's theory of interest is Jevonian in spirit. Though brief and somewhat sketchy, it anticipates all the basic elements of Fisher's theory. In many respects Launhardt achieves more in 20 pages than Böhm-Bawerk in about 500. Using modern terminology, the rate of interest is explained by the interplay between a psychological preference for present consumption, modified by variations in expected income, and the marginal productivity of capital (ch. 24). Saving is interpreted as a sacrifice of current consumption for the sake of an infinite stream of additions to future consumption. It is shown mathematically that, with a rising rate of interest, given the rate of time preference, saving first rises to a maximum and then declines, because at high interest rates small savings are enough to buy a lot of future income. According to the ‘basic principle of accumulation’, the present value of the future marginal utility of income is made equal to the current marginal utility of income. In the course of time, optimal saving, if initially positive, will decline until a steady state is reached (ch. 15). Investments will be made up to the point where the marginal saving in operating costs is equal to the rate of interest.

The subject of Part III is the effect of transportation on production and consumption. Launhardt starts out by determining production and prices of a single seller supplying an unlimited market of uniform density. Delivered prices are seen to rise towards the periphery in the shape of a hollow cone, known as the ‘Launhardt Funnel’ (ch. 27). If sellers of differentiated products compete in a uniformly populated

plain, their market areas are shown to be polygons, whose sides, depending on circumstances, are pieces of ellipses, hyperbolas or straight lines. In this context there emerges what Palander later called the Launhardt–Hotelling solution for heterogeneous duopoly. Forty-four years before Hotelling, Launhardt already used the paradigm of two competing suppliers, located at different points along a street, each maximizing his profits on the assumption that the price of his competitor is given. His solution, forgotten for half a century, is substantially identical to Hotelling's. An analogous analysis is provided for suppliers of differentiated products at the same location, showing how their ring-shaped market areas depend on transportation costs (ch. 29).

From the market areas of given suppliers, Launhardt shifts his attention to the supplying areas of given markets, which brings rent to the foreground. His description of the product ‘rings’ surrounding a single market city in an unlimited plain adds nothing to Von Thünen (ch. 30). The analysis is then extended to a number of markets, each with its limited supplying area. If identical cities are located in a pattern of regular triangles, the supplying areas are, of course, hexagonal. While this foreshadows Lösch's later work, Launhardt's triangular pattern is based on intuition and not on explicit optimality conditions. It is shown, however, how the mutual limitation of adjoining supplying areas raises rent and product prices (ch. 31). Much of this material was later incorporated in the second edition of *Commercial Network Planning* (1887).

Launhardt's monetary theory is far inferior to his microeconomics. Its centrepiece is the rejection of the quantity theory of money. In part, this is based, in the tradition of Senior and the Banking School, on the argument that under a gold standard an increase in the quantity of paper money just leads to an external (and/or internal) gold drain, while commodity prices remain tied to international prices or, in a closed economy, the gold price. To this extent, Launhardt is on firm ground. He went much further, however. In the theory of relative prices he had assumed that the marginal utility of money is constant. When first introduced, this was an innocuous simplification, but in the theory of money it became the source of fatal confusion, for it induced Launhardt to treat money incomes, which he chose as the proximate determinant of absolute prices, as if they were ‘real’ variables, independent of the money supply. After that, one is hardly surprised to read that higher interest rates result in higher prices and that gold discoveries have no influence on prices. The basic argument is found in *Mathematische Begründung* (1885); later elaborations (1889; 1894) and historical illustrations and applications.

Selected works

1868. *Bestimmung der zweckmässigsten Steigungsverhältnisse der Chausseen*. Hannover.

1869. *Ueber Rentabilität and Richtungsfeststellung der Strassen*. Hannover.

1872. *Kommercielle Tracirung der Verkehrswege*. Hannover.

1877. Die Betriebskosten der Eisenbahnen in ihrer Abhängigkeit von den Steigungs- und Krümmungsverhältnissen der Bahn. *Handbuch für specielle Eisenbahn-Technik*, vol. 4, Supplement. Leipzig.

1882. Die Bestimmung des zweckmässigsten Standortes einer gewerblichen Anlage. *Zeitschrift des Vereines deutscher Ingenieure* 26, 106–15.

1883. Wirtschaftliche Fragen des Eisenbahnwesens. *Centralblatt der Bauverwaltung*, vol. 3.

1885. *Mathematische Begründung der Volkswirtschaftslehre*. Leipzig.

1887–8. *Theorie des Trassirens*. 2 parts. Hannover.

1889. *Die Quantitätshtheorie. Ein Beitrag zur Lehre vom Wesen des Geldes*. Hannover.

1890a. *Theorie der Tarifbildung der Eisenbahnen*. Berlin.

1890b. Zur Frage einer besseren Feststellung des Personenfahrgeldes. *Organ für die Fortschritte des Eisenbahnwesens*.

1894. *Mark, Rubel und Rupie Erläuterungen zur Währungsfrage und Erörterungen über das Wesen des Geldes*. Berlin.

1900. Am sausenden Webstuhl der Zeit. *Aus Natur und Geisteswelt*, vol. 23. Leipzig.

How to cite this article

Niehans, Jürg. "Launhardt, Carl Friedrich Wilhelm (1832–1918)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000035> doi:10.1057/9780230226203.0938

The New Palgrave Dictionary of Economics Online

law of demand

Michael Jerison and John K.-H. Quah

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We formulate several laws of individual and market demand and describe their relationship to neoclassical demand theory. The laws have implications for comparative statics and stability of competitive equilibrium. We survey results that offer interpretable sufficient conditions for the laws to hold and we refer to related empirical evidence. The laws for market demand are more likely to be satisfied if commodities are more substitutable. Certain kinds of heterogeneity across individuals make the laws more likely to hold in the aggregate even if they are violated by individuals.

Keywords

asymmetric information; Bernoulli utility function; Cobb–Douglas preferences; comparative statics; compensated demand; Engel curve; Giffen effects; Giffen goods; income effect; Jacobian matrix; law of demand; Lyapunov's second theorem; marginal utility of income; metonymy; non-decreasing dispersion of excess demand; portfolio choice; risk aversion; Slutsky matrix; stability of equilibrium; substitution effect; tâtonnement; uniqueness of equilibrium

Article

The most familiar version of the law of demand says that as the price of a good increases the quantity demanded of the good falls. The principal use of the law of demand in economic theory is to provide sufficient and, in some contexts, necessary conditions for the uniqueness and stability of equilibrium, and for intuitive comparative statics. To guarantee such properties in equilibrium models with more than one good, the familiar one-good law of demand just stated is not sufficient – some multi-good version of the law is needed. In its multi-good form, the law of demand is said to hold for a particular change in prices if the prices and the quantities demanded move in opposite directions; in formal terms, the vector of price changes and the vector of resulting demand changes have a negative inner product.
In this article, we examine different formulations of the law of demand. They differ principally in the

domain of price changes over which the law applies. It is not always the case that the law of demand is required to hold for *all* price changes: the version of the law which is required for stability analysis and comparative statics varies from one context to another. For each formulation of the law of demand, we discuss the conditions which are sufficient to guarantee that it is satisfied.

To point out the obvious, the law of demand, in whatever form, is not a universal law at all but a condition which may hold in some situations and not others. It is well known that, in transactions where asymmetric information is an important consideration, violations of the law can occur. For example, lowering the price of a set of used cars does not necessarily lead to higher demand if potential buyers think that the lower price reflects the quality of the cars being offered. (For a discussion of violations of the law of demand and other issues which arise when price has an impact on the perceived quality of the good being exchanged, see Stiglitz, 1987.) In this article we make the classical assumption that the features of the good being transacted are commonly known and independent of the price. As we shall see, even in this classical setting various forms of the law of demand will hold only under conditions which are often neither obviously onerous nor obviously innocuous; in these cases, one must necessarily turn to empirical work to ascertain whether or not the law holds.

We use the notation and terminology of Mas-Colell, Whinston and Green (1995, chs. 2, 3, 5) and assume that the reader is familiar with the basic consumer and producer theory described there. We assume that there are L commodities and that consumers are price-takers. The demand of a consumer of type α with income w at price vector $p = (p_\ell)_{\ell=1}^{L} \gg 0$ is the vector $x(p, w, \alpha) = (x_\ell(p, w, \alpha))_{\ell=1}^{L}$ in $\mathbb{R}^L_+$, satisfying the budget identity $p \cdot x(p, w, \alpha) = w$ for all p and w. Unless stated otherwise, we assume the demand function $x(\cdot,\cdot,\alpha)$ to be C^1. Then it has a Slutsky matrix of substitution effects $S(p,w,\alpha)$ with •j element $S_{\ell j}(p, w, \alpha) = \partial x_\ell(p, w, \alpha) / \partial p_j + [\partial x_\ell(p, w, \alpha) / \partial w] x_j(p, w, \alpha)$. The Slutsky matrix $S(p,w,\alpha)$ is the Jacobian matrix of the Slutsky-compensated demand function x^*, defined by $x^*(q) = x(q, q \cdot x(p, w, \alpha), \alpha)$, evaluated at $q = p$. The term $[\partial x_\ell(p, w, \alpha) / \partial w] x_j(p, w, \alpha)$ is called an income effect since it approximates the effect on the demand for good • when income rises enough to compensate for a unit increase in the price of good j. If the consumer chooses demand bundles by maximizing a well-behaved utility function, then the Slutsky matrix is symmetric and negative semidefinite. The latter means that $v \cdot S(p, w, \alpha) v \leq 0$ for all $v \in \mathbb{R}^L$; in particular, the diagonal terms of the Slutsky matrix are non-positive.

One-good and multi-good laws of demand

The term ‘law of demand’ most often refers to the effect of price changes on consumers with fixed incomes. The law for a single good • and a single consumer of type α is

$$(p_\ell - \bar{p}_\ell)(x_\ell(p, w, \alpha) - x_\ell(\bar{p}, w, \alpha)) \leq 0, \tag{1}$$

for *p* and $\bar{p}$, with $p_i = \bar{p}_i$ for all $i \neq \ell$ and income *w* fixed. (In the *strict* version of the law, the weak inequality in (1) is replaced by strict inequality when $p \neq \bar{p}$; all the laws of demand discussed in this article can be stated in their corresponding strict forms, though we generally do not do so.) The inequality (1) is equivalent to

$$0 \geq \frac{\partial x_\ell}{\partial p_\ell}(p, w, \alpha) = S_{\ell\ell}(p, w, \alpha) - x_\ell(p, w, \alpha)\frac{\partial x_\ell}{\partial w}(p, w, \alpha), \quad \forall (p, w).$$

It holds if the substitution effect $S_{\ell\ell}$ is negative and larger in magnitude than the income effect $x_\ell(p, w, \alpha)\frac{\partial x_\ell}{\partial w}(p, w, \alpha)$. If the consumer is utility-maximizing, then $S_{\ell\ell} \leq 0$, so a sufficient condition for good • to obey the law of demand is that the demand for this good is normal $(\partial x_\ell(p, w, \alpha) / \partial w \geq 0)$. If the demand for good • is not normal, the price effect $\partial x_\ell / \partial p_\ell$ may be positive. This is called a Giffen effect and good • is called a Giffen good. All goods are normal and Giffen effects are ruled out if the demand function is generated by homothetic preferences or by a concave additive utility function $(u(x) = \sum_{\ell=1}^{L} u_\ell(x_\ell))$, or, more generally, by a supermodular concave function u, that is, one in which all commodity pairs are Auspitz–Lieben–Edgeworth–Pareto complements: $\partial^2 u(x) / \partial x_j \partial x_\ell \geq 0$ for all $j \neq \ell$ (Chipman, 1977).

Giffen goods are rarely observed. Sometimes demand for a durable good like oil may increase with its current price if traders expect an even higher price in the future. However, if commodities are distinguished by date, this is not a Giffen effect since a future price changes along with the current price. A possible example of a Giffen good is proposed by Baruch and Kannai (2002). They give evidence suggesting that, in Japan of the 1970s, shochu, a cheap (and, by some accounts, nasty) alcoholic drink, fits the definition. One may explain the demand for shochu in the following way. A consumer chooses between sake (good 1) and shochu (good 2). He always prefers sake to shochu, but he also *must have* a minimum alcohol intake (which we fix at 1). Formally, his utility is $u(x_1, x_2) = x_1$, subject to the 'survival' constraint $x_1 + x_2 \geq 1$. If the consumer is sufficiently poor, both the budget and survival constraints bind, with the consumer consuming as much sake – and as little shochu – as possible. A fall in the price of shochu allows him to buy less shochu and more sake and still meet his alcohol requirement; this he chooses to do since he always prefers sake to shochu.

Turning now to multi-good laws of demand, let $P \subseteq \mathbb{R}^L_{++}$ be a set of prices and let $X: P \to \mathbb{R}^L$ be a function representing individual or aggregate demand of firms or of consumers. The natural multi-good generalization of the one-good law in (1) is

$$(p - p') \cdot (X(p) - X(p')) \leq 0 \tag{2}$$

for all (p, p') in some subset of $P \times P$. If P is convex and open and X is C^1, (2) holds on $P \times P$ if and only if the Jacobian matrix $\partial X(p)$ is negative semidefinite at each p (Hildenbrand and Kirman, 1988). Suppose that the supply vector of the L goods changes from ω to ω'. Let p and p' be corresponding equilibrium prices so $X(p) = \omega$ and $X(p') = \omega'$. Then, if X obeys (2) for all prices, we obtain $(p - p') \cdot (\omega - \omega') \leq 0$. It is clear that this comparative statics property and the law of demand on X are essentially two sides of the same coin. Note also that, according to this property, an increase in the supply of good k, with the supply of all other goods held fixed, will lead to a fall in the price of k.
Suppose that P is open and X obeys the *strict* law of demand, that is, X satisfies (2) with strict inequality for all distinct p and p' in P. This implies in particular that X is 1–1 and that, for each $\overline{\omega}$ in $X(P)$, there is a unique equilibrium price vector $\overline{p} = X^{-1}(\overline{\omega})$. A tâtonnement path for the function $X - \overline{\omega}$ is the solution to $dp / dt = X(p(t)) - \overline{\omega}$ for some initial condition $p(0) = p'$ in P. We say that $X - \overline{\omega}$ is monotonically stable for $\overline{\omega}$ if each of its tâtonnement paths satisfies $d|p(t) - \overline{p}|^2 / dt < 0$ whenever $p(t) \neq \overline{p}$. It is easy to check that $X - \overline{\omega}$ is monotonically stable for all $\overline{\omega}$ in $X(P)$ if and only if X obeys the strict law of demand. Furthermore, because P is open, a tâtonnement path for $X - \overline{\omega}$ which begins at a price sufficiently close to $\overline{p} = X^{-1}(\overline{\omega})$ stays in P for all $t > 0$. Lyapunov's second theorem then guarantees that the tâtonnement path converges to $\overline{p}$.
Laws of demand are thus useful as intuitive sufficient conditions for the uniqueness and stability of equilibrium and for comparative statics. We will examine, in different contexts, circumstances under which they hold.

Law of demand for competitive firms and consumers with quasilinear utility

For a firm with production set Y, profit maximizing net output vector y at price vector p and $\overline{y}$ at $\overline{p}$ satisfy $p \cdot y \geq p \cdot \overline{y}$ and $\overline{p} \cdot \overline{y} \geq \overline{p} \cdot y$. The net demand vectors $x = -y$ and $\overline{x} = -\overline{y}$ satisfy $p(x - \overline{x}) \leq 0$ and $\overline{p}(x - \overline{x}) \geq 0$, hence satisfy the law of demand: $(p - \overline{p}) \cdot (x - \overline{x}) \leq 0$. Similarly, a consumer with utility function $u(x_0, x) = x_0 + \varphi(x_1, \ldots, x_L)$ (*quasilinear* with respect to good 0) and with sufficiently high income w satisfies the law of demand on a restricted domain, where the price of good 0 is fixed (say at 1). This is a special case of the law for firms. The consumer's optimal demand for goods 1 through L at p (the price vector for goods 1 to L) and income w maximizes $w - p \cdot x + \varphi(x)$. This is equivalent to profit maximization with x an input vector and $\phi(x)$ the value of output.
Bewley (1977) shows that a long-lived consumer with a random income stream and a random but stationary time-separable utility function, who is constrained from borrowing, will accumulate savings so that the marginal utility of income is nearly constant. In the short run, this consumer acts (nearly) as if its utility is quasilinear with respect to money, and its short run demands for other goods satisfy the law of demand. Vives (1987) formalizes Marshall's idea (in his *Principles*) that consumer demands for goods with small expenditure shares are close to demands generated by quasilinear utility.

Multi-good laws of demand for a consumer

Suppose the demand of a consumer of type α is determined by maximizing a utility function u^{α}. The Hicksian compensated demand $h(p, \bar{u}, \alpha)$ is a bundle that minimizes $p \cdot x$ subject to $u^{\alpha}(x) \geq \bar{u}$. Keeping the utility level fixed at $\bar{u}$, this Hicksian demand function satisfies the multi-good law of demand: (2) holds for $X(p) = h(p, \bar{u}, \alpha)$. Utility maximization also guarantees that $x(\cdot,\cdot,\alpha)$ satisfies the *weak weak axiom of revealed preference*: $p \cdot x(p', w', \alpha) \leq w \Rightarrow p' \cdot x(p, w, \alpha) \geq w'$. Equivalently, for any fixed w, $X(p) = x(p, w, \alpha)$ satisfies (2) on the restricted domain with $p \cdot X(p') = w$. This is also called the compensated law of demand since the demand vector $X(p')$ remains barely affordable when the price vector changes from p' to p. The weak weak axiom is satisfied so long as the consumer maximizes a *complete* preference relation; the preferences need not be transitive. When $x(\cdot,\cdot,\alpha)$ is C^1, the following are equivalent: (i) $x(\cdot,\cdot,\alpha)$ obeys the weak weak axiom; (ii) its Slutsky matrix $S(p,w,\alpha)$ is negative semidefinite (but not necessarily symmetric); (iii) its Jacobian matrix $\partial_p x(p,w,\alpha)$ is negative semidefinite on the hyperplane orthogonal to $x(p,w,\alpha)$ (Kihlstrom, Mas-Colell and Sonnenschein, 1976; Brighi, 2004).

When we say that $x(\cdot,\cdot,\alpha)$ obeys the unrestricted law of demand (or law of demand, for short) we mean that for each w, $X(p) = x(p, w, \alpha)$ satisfies (2) for *all* price changes. Since this is equivalent to negative semidefiniteness of the Jacobian $\partial_p x(p,w,\alpha)$ for all p, it is stronger than simply saying that the diagonal terms of the matrix are non-positive. Thus it is not equivalent to the one-good law of demand for every good and does not follow from the assumption that the demand for every good is normal.

Let $M(p, w, \alpha)$ be the income effects matrix, with •j component $[\partial_w x_i(p, w, \alpha)]\, x_j(p, w, \alpha)$. From the Slutsky decomposition, $\partial_p x(p, w, \alpha) = S(p, w, \alpha) - M(p, w, \alpha)$, we see that type α satisfies the law of demand if it satisfies the weak weak axiom and $M(p, w, \alpha)$ is positive semidefinite at each p. However, the latter condition is strong; it occurs if and only if demand is linear in income for all goods, which excludes the possibility of luxuries or necessities.

A more promising approach is to find conditions under which the Slutsky matrix always ‘dominates’ the income effects matrix even when the latter ‘misbehaves’. On the assumption that type α has a concave utility function u^{α}, a sufficient and (in a sense) necessary condition for the law of demand is $-[x^T \partial^2 u^{\alpha}(x) x] / (\partial u^{\alpha}(x) x) \leq 4, \ \forall x$. This result was obtained independently by Milleron (1974) and Mitjuschin and Polterovich (1978) (see also Mas-Colell, Whinston and Green, 1995, p. 145, and an alternative formulation in Kannai, 1989).

An important application of this result is in the theory of portfolio choice. In that case, the demand bundle is the consumer's contingent consumption over L states of the world; it is standard to assume that the consumer has a von Neumann–Morgenstern utility function $u^{\alpha}(x) = \sum_{i=1}^{l} \pi_i v^{\alpha}(x_i)$, where π_i is the subjective probability of state i and $v^{\alpha}: R_{++} \rightarrow R$ is the Bernoulli utility function. Suppose the coefficient of relative risk aversion, $-y v^{\alpha''}(y) / v^{\alpha'}(y)$, does not *vary* by more than four on the domain of v^{α}. Then the consumer's demand for contingent consumption at different state prices will obey the law of demand; this in turn implies that the law of demand holds for the consumer's demand for

securities, whether or not the market is complete (Quah, 2003).

Laws of market demand when the income distribution is independent of price

Consider a large economy with consumers drawn at random from a probability space $A \times \mathbb{R}_+$ of consumer types and their incomes, with distribution μ. The expected aggregate (market) demand vector at prices p is $X(p) = \int_{A \times \mathbb{R}_+} x(p, w, \alpha)\, d\mu$. We are interested in conditions under which X obeys the unrestricted law of demand, that is, (2) holds for all price changes; equivalently, $\partial X(p)$ is negative semidefinite for all p. If $x(\cdot,\cdot, \alpha)$ obeys the law of demand for all α, then, clearly, so will X. One justification for studying the law of demand at the individual level is that it is preserved by aggregation. Aggregating the Slutsky decomposition across all agents, the law of demand requires

$$v \cdot \partial X(p) v = v \cdot \left[\int_{\alpha \in A} x(p, w, \alpha)\, d\mu \right] v = v \cdot \bar{S}(p) v - v \cdot \bar{M}(p) v \leq 0, \ \forall v \tag{3}$$

where $\bar{S}(p) = \int S(p, w, \alpha)\, d\mu$ is the mean Slutsky matrix, and $\bar{M}(p)$ is the mean income effects matrix, with •j element $\int [\partial x_g(p, w, \alpha) / \partial w]\, x_j(p, w, \alpha)\, d\mu$. (We assume here and below that these integrals exist.) If all consumers obey the weak weak axiom, which they do if they are utility maximizers, then $S(p,w, \alpha)$ and hence $\bar{S}(p)$ are negative semidefinite; so $\partial X(p)$ is negative semidefinite if $\bar{M}(p)$ is positive semidefinite.

The matrix $\bar{M}(p)$ is determined by the consumers' Engel curves $x(p\cdot,\cdot, \alpha)$ at p. Positive semidefiniteness of this matrix is known as *increasing spread* (Hildenbrand, 1994). To see why, note that

$$2 v \cdot \bar{M}(p) v = \partial_t \int [v \cdot x(p, w + t, \alpha)]^2 d\mu(\alpha, w) |_{t=0}. \tag{4}$$

We can interpret $v \cdot x(p,w, \alpha)$ as α's demand for a commodity (call it T_v), which is consumed when the other goods are consumed; specifically, the consumption of one unit of good j requires v_j units of T_v. Then $\int [v \cdot x(p, w, \alpha)]^2 d\mu$ measures the spread of the consumers' demands for T_v around the origin. By (4), $\bar{M}(p)$ is positive semidefinite if and only if for every v the consumers' demands for T_v spread out from 0 as their incomes rise. This is the multi-good generalization of normality, where the consumers' demands for a single good increase (spread from 0) as their incomes rise.

We now consider various interpretable conditions on the distribution of consumer characteristics which

guarantee increasing spread (and thus the law of demand). This property holds if consumers have the same demand function and income is distributed with a non-increasing density function ρ on $[0, \bar{w}]$ (Hildenbrand, 1983). In that case, integrating by parts, (4) becomes

$2v \cdot \bar{M}(p)v = [v \cdot x(p, \bar{w}, \alpha)]^2 \rho(\bar{w}) - \int [v \cdot x(p, w, \alpha)]^2 \rho'(w) dw \geq 0$. While the non-increasing density condition is strong, imposing some weak restrictions on the Engel curves will guarantee increasing spread for a significantly larger class of income density functions (Chiappori, 1985). However, to guarantee increasing spread for *every* non-trivial income distribution requires stringent conditions on the consumers' Engel curves: $x(p,\cdot,\alpha)$ must lie in a single plane (depending on p) and the demand for each good is either a concave or convex function of income (Freixas and Mas-Colell, 1987; Jerison, 1999).

Increasing spread is also implied by certain kinds of behavioural heterogeneity across consumers. We consider consumers with the same income w and demands of the form

$x_\ell(p, w, \alpha) = e^{\alpha_\ell} \hat{x}(e^{\alpha_1} p_1, \ldots, e^{\alpha_L} p_L, w)$, where $\hat{x}$ is an arbitrary demand function and $\alpha = (\alpha_1, \ldots, \alpha_L) \in R^L$. If $\hat{x}$ is generated by some utility function $\hat{u}$, then $x(\cdot,\cdot,\alpha)$ is generated by the utility function $u^\alpha(x) = \hat{u}(e^{-\alpha_1} x_1, \ldots, e^{-\alpha_L} x_L)$. Increasing spread is guaranteed if α has a sufficiently flat density over R^L. This condition also ensures that the mean Slutsky matrix $\bar{S}(p)$ is negative semidefinite even if $\hat{x}$, hence each $x(\cdot,\cdot,\alpha)$, violates the weak weak axiom (and so is *not* generated by a utility function). Thus when α has a sufficiently flat density, X satisfies the law of demand; in fact it can be shown that X is nearly generated by Cobb–Douglas preferences (Grandmont, 1992). Whether flatness of the α density implies heterogeneity (in some meaningful sense) of the consumers' demands depends on the behaviour of $\hat{x}$ (Giraud and Quah, 2003).

Even when $\bar{M}(p)$ is not positive semidefinite, that is, $v \cdot \bar{M}(p)v < 0$ for some v, it is clear from (3) that $v \cdot \partial X(p)v < 0$ can hold provided the substitution effects are large enough, that is, $v \cdot \bar{S}(p)v$ is sufficiently negative. This feature can be exploited; for example, one can substantially weaken the non-increasing density condition in Hildenbrand (1983; described above) and still obtain the law of demand if substitution effects are accounted for through restrictions on the utility function (Quah, 2000). Similarly, a large enough positive income effect can compensate for consumers' violations of the weak weak axiom, that is, situations where, for some v, $v \cdot \bar{S}(p)v > 0$.

Whether the substitution effect $v \cdot \bar{S}(p)v$ dominates the income effect $v \cdot \bar{M}(p)v$ is an empirical question. The sizes of the effects must be estimated. Härdle, Hildenbrand and Jerison (1991) show how this can be done with cross-section data under standard econometric assumptions, without restrictions on the functional forms of the consumer demands. In most empirical demand analyses, consumers are grouped according to observable attributes other than income, and within a group, a, the consumers' budget share vectors are assumed to have the form $b^a(p, w) + \varepsilon$, where ε is a mean 0 random variable with distribution independent of income w. Under this assumption, a consumer's type is its attribute group and a realized value of ε. Within group a, the distribution of types with income w, denoted $\mu^a(\alpha|w)$, does not vary with w. Thus, if the income distribution in the group has a density ρ^a, then

$$\int\{\partial_w[v\cdot x(p,w,\alpha)]^2\}d\mu^a = \int\left\{\partial_w\int[v\cdot x(p,w,\alpha)]^2 d\mu^a(\alpha|w)\right\}\rho^a(w)dw,\ \forall v\in R^L. \tag{5}$$

The left side of (5) equals $2v\cdot M^a(p)v$, where $M^a(p)$ is the mean income effect matrix of the consumers in group a. The right side of (5) is the mean of the derivative of $\int[v\cdot x(p,w,\alpha)]^2 d\mu^a(\alpha|w)$ with respect to w. It can be efficiently estimated by the nonparametric method of average derivatives (Härdle and Stoker, 1989). The mean income effect matrix $\overline{M}(p)$ is a weighted average of the matrices $M^a(p)$, weighted by the shares of the population in the groups a. Condition (5), called metonymy, is weaker than the assumption that the budget shares have the form $b^a(p,w)+\varepsilon$, so weak, in fact, that it is not potentially refutable with infinite cross-section data (Evstigneev, Hildenbrand and Jerison, 1997; Jerison, 2001). Income effect matrices estimated in this way using cross-section expenditure data from several countries are all positive semidefinite (Härdle, Hildenbrand and Jerison, 1991; Hildenbrand and Kneip, 1993).

Laws of demand in private ownership economies

In the previous section, we assumed consumer incomes to be exogenously given independently of prices. This is plainly not true in general equilibrium. For example, consider a private ownership economy with consumers drawn randomly from a distribution μ over types, where type α has the demand function $x(\cdot,\cdot,\alpha)$ and an endowment vector ω^α. If the consumers receive no profits, the income of type α at price vector p is $p\cdot\omega^\alpha$. We are interested in laws of demand that can be satisfied by the consumer sector's aggregate demand $\ddot{X}(p)=\int x(p,p\cdot\omega^\alpha,\alpha)\,d\mu$ or aggregate excess demand $\zeta(p)=\ddot{X}(p)-\overline{\omega}$, where $\overline{\omega}=\int\omega^\alpha d\mu$ is the aggregate endowment.

The first thing to note is that under standard assumptions, both $\ddot{X}$ and ζ are zero-homogeneous and, essentially for this reason, satisfy the unrestricted law of demand only in exceptional cases (Hildenbrand and Kirman, 1988). However, if the consumers' endowments are collinear (that is, if for each α there is some $k\geq 0$ with $\omega^\alpha=k\overline{\omega}$) then the sufficient conditions for the law of market demand given in the previous section are also sufficient for $\ddot{X}$ (and hence ζ) to satisfy (2) for p and p' in $P=\left\{p\in R^L_{++}:p\cdot\overline{\omega}=1\right\}$; in other words, the law of demand holds for mean income preserving price changes. This is so because, when endowments are collinear, a price change which preserves mean income also preserves the income of *every* agent.

When we drop the strong assumption of collinear endowments, this restricted form of the law of demand is not guaranteed even if all consumers have homothetic preferences (Mas-Colell, Whinston and Green, 1995, p. 598). However, it *does* hold when the consumer sector has two properties: (*a*) all agents have homothetic preferences and (*b*) the preferences and endowments are independently distributed. Quah (1997) shows that this scenario can be understood as the idealization of a more general situation. The

crucial feature of homothetic preferences here is that they generate demand functions which are linear in income. Retaining the independence assumption (*b*), one can show that, when substitution effects are non-trivial (in some specific sense), $\ddot{X}$ obeys the restricted law of demand provided the mean demand of agents with identical endowments is not 'too non-linear' in income. This last property can arise from an appropriate form of heterogeneity in demand behaviour, which can be modelled using the parametric framework employed by Grandmont (1987; 1992).

It is interesting to ask when aggregate consumer excess demand ζ satisfies the weak weak axiom: $p \cdot \zeta(p') \leq 0 \Rightarrow p' \cdot \zeta(p) \geq 0$. This condition ensures that the set of equilibrium prices is convex in all competitive production economies with convex technology and constant returns to scale; furthermore, it is the weakest restriction on ζ guaranteeing this conclusion (Mas-Colell, Whinston and Green, 1995, p. 609). The sufficiency of this condition hinges on the fact that the production side of the economy satisfies the law of demand. Since the equilibrium set is generically discrete, its convexity implies generic uniqueness of equilibrium (up to scalar multiple). When ζ satisfies the weak weak axiom it also satisfies the law of demand (2) on the restricted set with $p \cdot \zeta(p') = 0$. If (2) holds strictly on this set when *p* and *p* $'$ are not collinear, then the unique equilibrium is globally stable under tâtonnement, and there are natural comparative statics.

With the use of a Slutsky decomposition, it can be shown that ζ satisfies the weak weak axiom if the mean Slutsky matrix *S*(*p*) is negative semidefinite (as it is if the consumers are utility maximizing) and the consumers' excess demand vectors spread apart on average when their incomes rise. The latter condition is called non-decreasing dispersion of excess demand (NDED). To formalize it, define $z(p, t, \alpha) \equiv x(p, t + p \cdot \omega^{\alpha}) - \omega^{\alpha}$, the excess demand of type α with income transfer *t*. The corresponding aggregate excess demand is $Z(p, t) \equiv \int z(p, t, \alpha) d\mu$. NDED holds if $\partial_t \int \{v \cdot [z(p, t, \alpha) - Z(p, t)]\}^2 d\mu|_{t=0} \geq 0$ for every $p \in R^L_{++}$ and every *v* with $v \cdot p = 0$ and $v \cdot \zeta(p) = 0$; in other words, the income transfers raise the variance of the composite excess demands $v \cdot z(p, t, \alpha)$ (Jerison, 1999). Quah's 1997 model (described above) is an example of an economy where NDED is satisfied approximately.

See Also

- comparative statics
- Engel curve
- general equilibrium
- Giffen's paradox
- revealed preference theory
- risk aversion
- tâtonnement and recontracting

Bibliography

Baruch, S. and Kannai, Y. 2002. Inferior goods, Giffen goods, and shochu. In *Economics Essays*, ed. G. Debreu, W. Neuefeind and W. Trockel. Berlin: Springer-Verlag.

Bewley, T. 1977. The permanent income hypothesis: a theoretical formulation. *Journal of Economic Theory* 16, 252–92.

Brighi, L. 2004. A stronger criterion for the weak weak axiom. *Journal of Mathematical Economics* 40, 93–103.

Chiappori, P.-A. 1985. Distribution of income and the ‘law of demand’. *Econometrica* 53, 109–28.

Chipman, J. 1977. An empirical implication of Auspitz–Lieben–Edgeworth–Pareto complementarity. *Journal of Economic Theory* 14, 228–31.

Evstigneev, I.V., Hildenbrand, W. and Jerison, M. 1997. Metonymy and cross-section demand. *Journal of Mathematical Economics* 28, 397–414.

Freixas, X. and Mas-Colell, A. 1987. Engel curves leading to the weak axiom in the aggregate. *Econometrica* 55, 515–31.

Giraud, G. and Quah, J.K.-H. 2003. Homothetic or Cobb–Douglas behavior through aggregation. *Contributions to Theoretical Economics* 3(1), Article 8.

Grandmont, J.M. 1987. Distribution of preferences and the ‘law of demand’. *Econometrica* 55, 155–61.

Grandmont, J.M. 1992. Transformations of the commodity space, behavioral heterogeneity, and the aggregation problem. *Journal of Economic Theory* 57, 1–35.

Härdle, W., Hildenbrand, W. and Jerison, M. 1991. Empirical evidence on the law of demand. *Econometrica* 59, 1525–49.

Härdle, W. and Stoker, T. 1989. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84, 986–95.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51, 997–1019.

Hildenbrand, W. 1994. *Market Demand*. Princeton: Princeton University Press.

Hildenbrand, W. and Kirman, A. 1988. *Equilibrium Analysis*. Amsterdam: North-Holland.

Hildenbrand, W. and Kneip, A. 1993. Family expenditure data, heterscedasticity and the law of demand.

Ricerche Economiche 47, 137–65.

Jerison, M. 1999. Dispersed excess demands, the weak axiom, and uniqueness of equilibrium. *Journal of Mathematical Economics* 31, 15–48.

Jerison, M. 2001. Demand dispersion, metonymy and ideal panel data. In *Economics Essays*, ed. G. Debreu, W. Neuefeind and W. Trockel. Berlin: Springer-Verlag.

Kannai, Y. 1989. A Characterization of monotone individual demand functions. *Journal of Mathematical Economics* 18, 87–94.

Kihlstrom, R., Mas-Colell, A. and Sonnenschein, H. 1976. The demand theory of the weak axiom of revealed preference. *Econometrica* 44, 971–78.

Mas-Colell, A. 1991. On the uniqueness of equilibrium once again. In *Equilibrium Theory and Applications*, ed. W. Barnett et al. Cambridge: Cambridge University Press.

Mas-Colell, A., Whinston, M.D. and Green, J.R. 1995. *Microeconomic Theory*. Oxford: Oxford University Press.

Milleron, J.C. 1974. Unicité et stabilité de l’équilibre en économie de distribution. Unpublished seminar paper, Séminaire d'Econométrie Roy-Malinvaud. Paris: CNRS.

Mitjuschin, L.G. and Polterovich, W.M. 1978. Criteria for monotonicity of demand functions. *Ekonomika i Matematicheskie Metody* 14, 122–8.

Quah, J.K.-H. 1997. The law of demand when income is price dependent. *Econometrica* 65, 1421–42.

Quah, J.K.-H. 2000. The monotonicity of individual and market demand. *Econometrica* 68, 911–30.

Quah, J.K.-H. 2003. The law of demand and risk aversion. *Econometrica* 71, 713–21.

Stiglitz, J.E. 1987. The Causes and consequences of the dependence of quality on price. *Journal of Economic Literature* 25(1), 1–48.

Vives, X. 1987. Small income effects: a Marshallian theory of consumer surplus and downward sloping demand. *Review of Economic Studies* 54, 87–103.

How to cite this article

Jerison, Michael and John K.-H. Quah. "law of demand." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000215> doi:10.1057/9780230226203.0939

The New Palgrave Dictionary of Economics Online

law of indifference

F.Y. Edgeworth
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

law of indifference

Article

A designation applied by Jevons to the following fundamental proposition: ‘In the same open market, at any one moment, there cannot be two prices for the same kind of article.’
This proposition, which is at the foundation of a large part of economic science, itself rests on certain ulterior grounds: namely, certain conditions of a perfect market. One is that monopolies should not exist, or at least should not exert that power in virtue of which a proprietor of a theatre, in Germany for instance, can make a different charge for the admission of soldiers and civilians, of men and women. The indivisibility of the articles dealt in appears to be another circumstance which may counteract the law of indifference in some kinds of market, where price is not regulated by cost of production.
[Jevons (1875), *Theory of Exchange*, 2nd edn, p. 99 (statement of the law). Walker (1886), *Political Economy*, art. 132 (a restatement). Mill (1848), *Political Economy*, bk. ii. ch. iv. § 3 (imperfections of actual markets). Edgeworth (1881), *Mathematical Psychics*, pp. 19, 46 (possible exceptions to the law of indifference).]

Bibliography

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Jevons, W.S. 1875. *Money and the Mechanism of Exchange*. London: C. Kegan Paul & Co.

Mill, J.S. 1848. *Principles of Political Economy*. London: J.W. Parker.

Walker, F.A. 1886. *A Brief Textbook of Political Economy*. London.

How to cite this article

Edgeworth, F. Y. "law of indifference." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000064> doi:10.1057/9780230226203.0941

The New Palgrave Dictionary of Economics Online

law(s) of large numbers

Werner Ploberger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

It is a well-known fact that averages of most random variables converge. The laws of large numbers are mathematical theorems which explain this phenomenon. We discuss the various forms of this theorem. Generalizations to dependent variables (ergodic ths) are introduced. We also mention uniform laws of large numbers, which are quite indispensable tools to prove consistency of estimators.

Keywords

Bernoulli experiments; Bernoulli, J.; ergodic theorems; law of large numbers; maximum likelihood; Poisson, S. D.; probability; strong law of large numbers; variance; weak law of large numbers

Article

When we have a large number of independent replications of a random experiment, we observe that the frequency of the outcomes can be very well approximated by the probabilities of the corresponding events. The profits of many commercially successful enterprises – like casinos or insurance companies – are based on random events obeying some laws.
Mathematically, this idea was first formulated by Jacob Bernoulli, for experiments with only two outcomes (‘Bernoulli experiments’). The terminology ‘law of large numbers’ was introduced by S.D. Poisson in 1835.
In the most basic version, LLN (the standard abbreviation for ‘law(s) of large numbers’) describes results of the following type. We assume that we have given a sequence of random variables X_1, X_2,•…• We say we have a LLN if

$$\frac{1}{N}(X_1 + \ldots X_N) \tag{1}$$

converges for $N \to \infty$, preferably to a constant.
For stating our results, we have to state the nature of the convergence in our LLN and impose some restrictions on the X_i. The more we restrict our X_i, the stronger our convergence results will be.

The weak law of large numbers

The 'weak law of large numbers' states that averages like (1) converge in a 'weak' sense (like for example convergence in probability) to a limit. In most cases, the requirements for the random variables involved are not very restrictive. A typical weak LLN is the following theorem.
Theorem 1: Assume that the random variables X_i satisfy

$$EX_i = 0, \tag{2}$$

$$\sup EX_i^2 < \infty \tag{3}$$

and

$$\lim_{M \to \infty} \sup_{|i-j|>M} \left| EX_i X_j \right| < \infty . \tag{4}$$

Then for $N \to \infty$

$$\frac{1}{N}(X_1 + \ldots X_N) \to^P 0,$$

where $\xrightarrow{P}$ *denotes convergence in probability.*
Our random variables have to be centred, of bounded variance, and condition (4) requires that the correlation of random variables ‘far apart’ converges to zero uniformly. This is a very general and important result. Another advantage is the simplicity of its proof: it is an elementary task to show that the variance of the average converges to zero. Then the theorem is an immediate consequence of Chebyshev's inequality. Moreover, the assumptions of the theorem can easily be checked, and only depend on the second moments of the X_i.

The strong law of large numbers

In some cases, we want to have more than convergence in probability of the averages. For this purpose, we have strong laws of large numbers. We do need, however, stricter requirements. The following theorem is a typical strong LLN. A more stringent discussion of this type of theorems
Theorem 2: Assume that the random variables X_i *satisfy* (2),(3). *Let* $\mathfrak{F}_i$ *be an increasing sequence of* σ*-algebras* (*for example* $\mathfrak{F}_{i-1} \subset \mathfrak{F}_i$) *so that* X_i *is* $\mathfrak{F}_i$ *–measurable. Then let us assume that*

$$E(X_i / \mathfrak{F}_{i-1}) = 0. \tag{5}$$

Then

$$\frac{1}{N}(X_1 + \ldots X_N) \to 0 \ P - \textit{almost surely}.$$

Heuristically, we can interpret $\mathfrak{F}_i$ as information available at time i. Then (5) postulates that we cannot predict X_i given the information at time $i - 1$. One important special case where (5) is fulfilled is the case of independent. In this case, we can choose $\mathfrak{F}_i$ to be the σ-algebra generated by X_1,•…•X_i,. Then, assuming the X_i to be independent, we have $E(X_i / \mathfrak{F}_{i-1}) = E(X_i)$.
Hence (5) is more general than the requirement of independence, but still far more restrictive than (4).

Ergodic theorems

We can easily see that (5) implies that our X_i are uncorrelated. In many applications, this requirement is unrealistic. Fortunately, there is a theory guaranteeing convergence of sums like (1) at least for

stationary processes X_i. A process X_i, $i \in \mathbf{Z}$ is called (strictly) stationary if for all $n \in \mathbf{Z}$ the distributions of $(X_1, X_2, \ldots X_m)$ and $(X_{n+1}, X_2, \ldots X_{n+m})$ are the same. To describe the limits of our process, we need to introduce the transition operator T: This operator is a mapping defined on the space of random variables measurable with respect to the σ-algebra generated by the X_i, $i \in \mathbf{Z}$. For random variables

$$Y = f(X_{t_1}, X_{t_2}, \ldots X_{t_n}) \tag{6}$$

we define the random variable TY by

$$TY = f(X_{t_1+1}, X_{t_2+1}, \ldots X_{t_n+1}). \tag{7}$$

So the transition operator T shifts every random variable 'one step in the future'. (T can be considered as the inverse of the usual lag operator). One can show that the definition based on (6), (7) can be uniquely extended to the space of all X_i, $i \in$ Z measurable random variables. Then an event A is called *invariant* if

$$TI_A = I_A \text{ almost surely,}$$

where I_A is the indicator of the event A. It can be easily seen that the invariant events form a σ–algebra, which we denote by $\mathfrak{F}$. Then the ergodic theorem states that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = E(X_i / \mathfrak{F}). \tag{8}$$

(Since we are taking the conditional expectation with respect to $\mathfrak{F}$, it can easily be seen that $E(X_1 / \mathfrak{F}) = E(X_2 / \mathfrak{F}) = \ldots$).

The ergodic theorem is included in most of advanced textbooks on probability theory (see, for example, Billingsley, 1995). A more detailed exposition can be found in Gray (2007).

We now can take various conclusions from our theorem. First of all, we can regardless of the nature of the σ–algebra $\mathfrak{F}$ conclude that the limit of $\frac{1}{n}\sum_{i=1}^{n} X_i$ *exists*. In econometric theory, one often postulates the existence of limits of certain averages (that is, in regression theory we often assume that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ exists). In case of stationary processes, the theorem here makes assumptions of this type very plausible.
If the σ–algebra $\mathfrak{F}$ is trivial (that is, consists only of events of probability 0 and 1), then the right-hand side of (8) is constant. One sufficient criterion for this property is that the process is a causal function of i.i.d. random variables. So if

$$X_i = f(e_i, e_{i-1}, \ldots)$$

where e_i are i.i.d., $\mathfrak{F}$ is trivial.

Applications and uniform laws of large numbers

For many statistical applications, we need stronger results. As a first example, consider the asymptotic of the maximum likelihood estimator. As a simplest case, let us discuss the case of i.i.d. random variables X_i, distributed according to densities f_θ for parameters $\theta \in \Theta$, and let θ_0 be the true parameter. Then the LLN guarantees that for every fixed θ

$$\frac{1}{n}\sum \ln(f_\theta(X_i) \to \int \ln(f_\theta) f_{\theta_0}, \tag{9}$$

and the function on the right-hand side is maximized if $\theta = \theta_0$. Since the maximum likelihood estimator maximizes the right-hand side, it seems reasonable to exploit this relation for a proof of consistency of the maximum likelihood estimator. The LLN guarantees only convergence for *fixed* θ, from our LLN we cannot say anything about the limiting behaviour of

$$\sup_{\theta\in\Theta} \ln\left(\frac{1}{n}\sum \ln(f_\theta(X_i)\right).$$

This problem would go away if one could establish that the convergence in (9) is *uniform* in θ. This

strategy was first realized in a path breaking paper by A. Wald (Wald, 1949), where he first established the consistency of the maximum likelihood estimator. Today the techniques are a little more sophisticated. Nevertheless, consistency proofs for M-estimators still rely to good extend on Wald's idea.
Another application of uniform LLN is the consistency of ‘plug-in’ estimators. In many cases, the asymptotic variance of certain estimators can be expressed as a function of the expectations of certain random functions, possibly depending on the parameter to be estimated (for example, the well-known ‘sandwich formula’ derived by H. White; see for example Hayashi, 2000). A standard strategy is to estimate the parameter, then replace the expectation by an average (and hope that – due to the LLN – average and expectation are close together) and use the estimated parameter as an argument. One can easily see that only a uniform law of large numbers can justify procedures of this type.
Fortunately, there exist a lot of criteria to establish uniform laws of large numbers. For most cases of interest to econometricians, the papers by Andrews (1992) and P•tscher and Prucha (1989) will be sufficient.
A more general and abstract theory can be found in van der Vaart and Wellner (1996). These theories allow us also to estimate the cumulative distribution function of random variables directly. Suppose we have given random variables $X_i,\ldots X_n$. Then the empirical distribution function F_n is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

(that is F_n jumps $1/n$ in X_i and is constant in between the jumps). Then the theorem of Glivenko-Cantelli (see van der Vaart and Wellner, 1996) states that if the X_i are i.i.d. with cumulative distribution function F, then

$$\sup|F_n(x) - F(x)| \to 0.$$

It should be noted that there are generalizations to multivariate or even more general X_i. In these cases, however, one has to use slightly more sophisticated techniques. Instead of the ‘empirical distribution function’, one has to use the ‘empirical measure’ (a random measure, which puts mass $1/n$ in the points X_i, and instead of the maximum difference of the distribution functions one has to consider the maximal difference of the measures over certain classes (‘VC-classes’).

Bibliography

Andrews, D.W.K. 1992. Generic uniform convergence. *Econometric Theory* 8, 241–57.

Billingsley, P. 1995. *Probability and Measure*, 3rd edn. New York: Wiley.

Gray, R.M. 2007. *Probability, Random Processes, and Ergodic Properties*. Online. Available at http://ee.stanford.edu/~gray/arp.html, accessed 29 April 2007.

Hall, P. and Heyde, C.C. 1980. *Martingale Limit Theory and its Application*. San Diego: Academic Press.

Hayashi, F. 2000. *Econometrics*. Princeton: Princeton University Press.

P•tscher, B.M. and Prucha, I.R. 1989. A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica* 57, 675–83.

van der Vaart, A.W. and Wellner, J.A. 1996. *Weak Convergence and Empirical Processes*. New York: Springer.

Wald, A. 1949. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* 20, 595–601.

How to cite this article

Ploberger, Werner. "law(s) of large numbers." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000041> doi:10.1057/9780230226203.0943

The New Palgrave Dictionary of Economics Online

law, economic analysis of

A. Mitchell Polinsky and Steven Shavell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys the economic analysis of five primary fields of law: property law; liability for accidents; contract law; litigation; and public enforcement and criminal law. It also briefly considers some criticisms of the economic analysis of law.

Keywords

adverse possession; asymmetric information; Becker, G.; Bentham, J.; bona fide purchase rule; Coase, R.; collective action; compensated takings; contract formation; contractual interpretation; copyright; corrective tax; criminal law; damage measures; Demsetz, H.; deterrence; direct regulation; disclosure of information; due care; eminent domain; expectation measure of damages; fairness; fault-based liability; finders-keepers rule; fines; first-party insurance; good title; hold-out problem; imprisonment; incapacitation; incomplete contracts; injunction; justice; labels; land registries; law and economics; liability; liability insurance; litigation; moral hazard; negligence rule; original ownership rule; patents; Posner, R.; possessory rights; product liability; property rights; public enforcement of law; public property; punishment; risk; risk aversion; risk-bearing; settlement vs trial; social norms; social welfare; specific performance; strict liability; suit; takings; tort law; trade secret law; trademarks; unilateral accident model; utilitarianism

Article

Economic analysis of law seeks to identify the effects of legal rules on the behaviour of relevant actors and to determine whether these effects are socially desirable. The approach employed is that of economic analysis generally: the behaviour of individuals and firms is described on the assumption that they are forward looking and rational, and the framework of welfare economics is adopted to assess the social desirability of outcomes. The field may be said to have begun with Bentham (1789), who

systematically examined how actors would behave in the face of legal incentives (especially criminal sanctions) and who evaluated outcomes with respect to a clearly stated measure of social welfare (utilitarianism). His work was left essentially undeveloped until four important contributions were made: Coase (1960) on externalities and liability, Becker (1968) on crime and law enforcement, Calabresi (1970) on accident law, and Posner (1972) on economic analysis of law in general. (Calabresi's book was the culmination of a series of articles, the first of which was published in 1961; see Calabresi, 1961.) Our focus here is on the analytical foundations of five basic legal subjects: property, torts, contracts, civil litigation, and crime and law enforcement (on these, see generally Cooter and Ulen, 2003; Posner, 2003; Miceli, 1997; and Shavell, 2004). We do not treat more particular areas of law, such as antitrust, corporate and tax law, nor do we cite empirical work; for surveys of these and other areas of law and economics, including empirical studies, see Polinsky and Shavell (2007).

1 . Property law

Justification and emergence of property rights

A beginning question is why there should be property rights in things. A number of arguments have been stressed, especially by early writers, including that property rights furnish incentives to work and to maintain durable things; that the rights make trade possible; and that, if such rights were absent, individuals would spend effort trying to take things from each other and protecting their things. Property rights would be expected to emerge when their advantages become sufficiently great. For example, Demsetz (1967) explains the development of property rights in land among Indians as a way of preventing overly intensive hunting of valuable animals. Umbeck (1981) shows that when gold was discovered in California in 1848 property rights in gold-bearing land and river beds developed, as this encouraged individuals to pan for gold and to build sluices; it also curbed wasteful efforts to grab land from others. For a survey, see Libecap (1986).

Division of property rights

Property rights can be viewed as composed of *possessory* rights – rights of use – and rights to *transfer* possessory rights. Thus, what we commonly conceive of as ownership (say, of land) entails both a large swath of possessory rights (rights to build on land, plant on it, under most contingencies, and into the infinite future) and associated rights to transfer them. Property rights in things are generally held in substantially agglomerated bundles, but there is also significant partitioning of rights contemporaneously, over time and contingencies, and according to whether the rights are possessory or are for transfer. For example, an owner of land may not hold complete possessory rights, in that others may possess an easement giving them the right of passage upon his land, or the right to take timber, or the right to extract oil if found (thus a contingent right). A rental agreement constitutes a division of property rights over time. Trust arrangements, such as those under which an adult manages property for a child, divide possessory rights and rights to transfer.

The division of property rights may be valuable when different parties derive different benefits from them, because gains can then be achieved if rights are allocated to those who obtain the most from them.

There may, however, be disadvantages to the division of rights, including that externalities may arise (a person with a right of passage might trample crops).

Public property and its acquisition; takings and compensation

An important class of property is that owned by the public. As is well known, the main justification for public property concerns the difficulty that private providers would experience in charging for certain goods and services.

When it is desirable for the state to acquire property for public use, the state can either purchase it or take it through the exercise of the power of *eminent domain*. In the latter case, the law typically provides that the state must compensate property owners for the value of what has been taken from them.

A difference between purchase and compensated takings is that the amounts owners receive are determined by negotiation in the former case but unilaterally by the state in the latter. Because of errors in state determination of value, as well as concern about the behaviour of government officials, purchase would ordinarily be superior to compensated takings. When, however, the state needs to assemble many contiguous parcels, such as for a road, acquisition by purchase might be stymied by hold-out problems, making the power to take socially advantageous.

On the assumption that there is a reason for the state to take property, a requirement to pay compensation may curb problems of overzealousness or abuse of authority by public officials, yet it may also exacerbate potential problems of insufficient public activity, because public authorities do not directly receive the benefits of takings (Kaplow, 1986). Payment of compensation also may lead property owners to invest excessively in property (see Blume, Rubinfeld and Shapiro, 1984).

Acquisition of property in unowned things

The law must determine the conditions under which a person will become a legal owner of previously unowned things, such as wild animals, fish, and mineral and oil deposits. Under the *finders-keepers rule*, incentives to invest in capture (such as to hunt for animals or explore for oil) are optimal if only one person is making the effort. However, if many individuals seek unowned things, they will invest a socially excessive amount of resources in search: one person's investment usually will come, at least partly, at the expense of other person's likelihood of finding unowned things. Various aspects of the law ameliorate this problem of excessive search effort. For example, regulations may limit the quantities of fish and wild animals that can be taken; the right to search for minerals on the ocean floor may be auctioned; and oil extraction rights may be assigned to a single party.

Acquisition of good title when property is sold

A basic difficulty associated with sale of property that a legal system must solve is establishing validity of ownership or *title*. Good title is important for trade, since buyers want to be assured that they have property rights in what they purchase. But, if any sale gives a buyer good title, theft is encouraged, since thieves could then easily sell stolen goods. Under a *registration system*, good title means that one's name is listed in the registry as the owner, and title passes at the time of sale by an authorized change in the

registry. Hence, buyers can clearly determine whether they are obtaining good title by checking the registry, and a thief could not easily sell stolen property by claiming that he has good title. Registries, however, are expensive to establish and maintain.

In the absence of registries, the law may employ the *original ownership rule*, under which the buyer does not obtain good title if the seller did not have good title. Alternatively, under the *bona fide purchase rule*, a buyer acquires good title as long as he had reason to think that the sale was legitimate, even if the item sold was in fact wrongfully obtained. This rule makes theft more attractive because thieves will often be able to sell their property to buyers who will be motivated to 'believe' that sales are bona fide.

Adverse possession

The legal doctrine of adverse possession allows involuntary transfer of land: a person is deemed to become the legal owner of land if he takes possession of it and uses it openly for at least a prescribed period, such as ten years. It may appear that this rule could be desirable because it encourages productive use of idle land. But this overlooks the possibility that a prospective adverse possessor could always bargain with the owner to rent or buy the land, and that there may be good reasons for allowing the land to remain idle. Additionally, the rule induces owners to expend resources policing incursions, and potential adverse possessors to attempt possession. A historical justification for the rule is that, before reliable land registries existed, it allowed a seller of land to establish good title to a buyer relatively easily: the seller need only show that he was on the land for the prescribed period.

Constraints on sale of property

Legal restrictions are often imposed on the sale of goods and services. One standard justification is externalities. For example, the sale of fireworks might be banned because of the externality that their ownership creates, namely, putting others at risk of injury. The other standard justification for legal restrictions on sale is lack of consumer information. For instance, a drug may not be sold without a prescription because of fear that otherwise buyers would not use it properly. Rather than restrict sales, however, the government could supply relevant information to consumers, such as by indicating that the drug has dangerous side effects, or that it should be taken only on the advice of a medical expert.

Externalities

When individuals use property, they may cause externalities, namely, harm or benefit to others. Generally, it is socially desirable for individuals to do more than is in their self-interest to reduce detrimental externalities and to act so as to increase beneficial externalities. The socially optimal resolution of harmful externalities often involves the behaviour of victims as well as that of injurers. If victims can do things to reduce the amount of harm more cheaply than injurers (say, install air filters to avoid pollution), it is optimal for victims to do so. Moreover, victims can sometimes alter their locations to reduce their exposure to harm.

Legal intervention can ameliorate problems of externalities. A major form of intervention that has been studied is *direct regulation*, under which the state restricts permissible behaviour, such as requiring

factories to use smoke arrestors. Closely related is the *injunction*, whereby a potential victim can enlist the power of the state to force a potential injurer to take steps to prevent harm or to cease his activity. Society can also make use of financial incentives to induce injurers to reduce harmful externalities. Under the *corrective tax*, a party pays the state an amount equal to the expected harm he causes – for example, the expected harm due to a discharge of a pollutant into a lake. There is also *liability*, a privately initiated means of providing financial incentives, under which injurers pay for harm done if sued by victims. These methods differ in the information that the state needs to apply them, in whether they require or harness information that victims have about harm, and in other respects, such that each may be superior to the other in different circumstances (Shavell, 1993).

Parties affected by externalities will sometimes have the opportunity to make mutually beneficial agreements with those who generate the externalities, as Coase (1960) stressed. But bargaining may not occur, for many reasons: cost; collective action problems (such as when many victims each face small harms); and lack of knowledge of harm (such as from an invisible carcinogen). If bargaining does occur, it may not be successful, owing to asymmetric information. These difficulties often make bargaining a problematic solution to externality problems and imply that liability rules are needed, as discussed by Calabresi and Melamed (1972).

Property rights in information

The granting of property rights in information, notably the award of *patents* for inventions and *copyrights* for written works and certain other compositions, involves a major social benefit – the provision of incentives to create intellectual works – but also a social disadvantage – the creation of power to price above marginal cost. Patent and copyright law have been examined to ascertain how they reflect the trade-off between this benefit and disadvantage. A distinct form of legal protection is *trade secret law*, comprising various doctrines of contract and tort law that serve to protect a range of commercially valuable information that is not (or cannot be) protected by patent or copyright, such as customer lists. On property rights in information, see generally Landes and Posner (2003).

An alternative to property rights in information is for the state to offer *rewards* to creators of information, and for information that is developed to be made available to all who want it. Thus, an author of a book would receive a reward from the state for writing the book, possibly based on sales, but anyone who wanted to print it and sell it could do so. This system would create incentives for the creation of information without distorting prices, but requires the state to choose the magnitude of rewards.

Property rights in labels

Many goods and services are identified by labels, which have substantial social value because the quality of goods and services may be hard for consumers to determine directly. Labels enable consumers to purchase goods and services on the basis of product quality without requiring consumers to independently determine quality; a person who wants to stay at a high-quality hotel in another city can choose such a hotel merely by its label, such as ‘Ritz Hotel’. In addition, sellers who label their output will have an incentive to produce goods and services of quality because consumers will recognize

quality through sellers' labels. This basic reasoning is used to justify property rights in *trademarks*, as discussed by Landes and Posner (1987b).

2 . Liability for accidents

Legal liability for accidents, which is governed by tort law, is a means by which society can reduce the risk of harm by threatening potential injurers with having to pay for the harms they cause. Liability is also frequently viewed as a device for compensating victims of harm, though we emphasize that insurance can provide compensation more cheaply than the liability system. There are two basic rules of liability. Under *strict liability*, an injurer must always pay for harm due to an accident that he causes. Under the *negligence rule*, an injurer must pay for harm caused only when he is found negligent, that is, only when his level of care was less than a standard of care chosen by the courts, often referred to as *due care*. (There are various versions of these rules that depend on whether victims' care was insufficient.) In practice, the negligence rule is the dominant form of liability; strict liability is reserved mainly for certain especially dangerous activities. On economic analysis of liability for accidents, see generally Calabresi (1970), Landes and Posner (1987a), and Shavell (1987a).

Incentives to take care

In order to focus on how liability affects the incentive to prevent harm, assume first that parties are risk neutral and that accidents are *unilateral* – only injurers (not victims) influence risk by their choice of *care x*. Let $p(x)$ be the probability of an accident that causes harm h, where p is declining in x. Assume that the social objective is to minimize total expected costs, $x + p(x)h$, and let x^* denote the optimal x. Under strict liability, injurers pay damages equal to h whenever an accident occurs, and they naturally bear the cost of care x. Thus, they minimize $x + p(x)h$; accordingly, they choose x^*.
Under the negligence rule, suppose that the due care level is set equal to x^*, meaning that an injurer who causes harm will have to pay h if $x < x^*$, but will not have to pay anything if $x \geq x^*$. Then it can be shown that the injurer will choose x^*: clearly, the injurer will not choose x greater than x^*; and he will not choose $x < x^*$, for then he will be liable (in which case the analysis of strict liability shows that he would not choose $x < x^*$). Thus, under both forms of liability, injurers are led to take optimal care. Note that to apply the negligence rule courts need sufficient information to calculate x^* and to observe x, whereas under strict liability they only have to observe x.
The analysis of incentives and liability has been undertaken as well for *bilateral* accidents, in which victims also take care, and when there is uncertainty in the determination of negligence (such as due to imperfect observation of x). On incentives and liability for unilateral and bilateral accidents, see originally Brown (1973) and also Diamond (1974).

Level of activity

An important extension allows for injurers to choose their *level of activity z*, which is interpreted as the (continuously variable) number of times they engage in their activity (or, if injurers are firms, their

output). Let $b(z)$ be the benefit (or profit) from the activity, and assume the social objective is to maximize $b(z) - z(x + p(x)h)$; here $x + p(x)h$ is assumed to be the cost of care and expected harm each time an injurer engages in his activity. Let x^* and z^* be optimal values. Note that x^* minimizes $x + p(x)h$, so x^* is as described above, and that z^* is determined by $b'(z) = x^* + p(x^*)h$, which is to say, the marginal benefit from the activity equals the marginal social cost.

Under strict liability, an injurer will choose both the level of care and the level of activity optimally, as his objective will be the same as the social objective, to maximize $b(z) - z(x + p(x)h)$, because damage payments equal h whenever harm occurs. Under the negligence rule, an injurer will choose optimal care x^* as before, but his level of activity z will be socially excessive. In particular, because an injurer will escape liability by taking care x^*, he will choose z to maximize $b(z) - zx^*$, so that z will satisfy $b'(z) = x^*$. The injurer's cost of raising his level of activity is only his cost of care x^*, which is less than the social cost, which also includes $p(x^*)h$. On liability and the level of activity, see Shavell (1980b).

The failure of the negligence rule to control the level of activity arises because negligence is defined here (and also generally in practice) in terms of care alone. A justification for this restriction is the difficulty courts would face in determining the optimal activity level z^* and the actual z. The failure of the negligence rule to control the injurer's level of activity is applicable to any aspect of injurer behaviour that would be difficult to regulate directly (including, for example, research and development activity). If, however, courts were able to incorporate all aspects of injurer behaviour into the definition of due care, the negligence rule would result in optimal behaviour in all respects. (Note that the variable x in the original problem could be interpreted as a vector, with each element corresponding to a dimension of behaviour.)

Product liability

Another extension of the model of liability and incentives concerns product liability, the liability of firms for harms suffered by their customers. Here the degree to which liability creates incentives to reduce risk depends on customer knowledge of risk. If their knowledge is perfect, liability does not affect incentives since customers will recognize risky products and pay appropriately less for them. If their knowledge is imperfect, there is a role for liability, in many respects similar to what has been discussed above.

Risk-bearing and insurance

In addition to affecting incentives to reduce harm, the socially optimal resolution of the accident problem involves the spreading of risk to lessen risk-bearing by risk-averse parties. Risk-bearing is relevant not only because potential victims may face the risk of accident losses, but also because potential injurers may face the risk of liability. The former risk can be mitigated through so-called first-party insurance that covers losses suffered in accidents, and the latter through liability insurance. Because risk-averse individuals tend to purchase insurance, the incentives associated with liability do not function in the direct way discussed above, but instead are mediated by the terms of insurance

policies. To illustrate, consider strict liability in the unilateral accident model with care alone allowed to vary, and assume that insurance is sold at actuarially fair rates. If injurers are risk averse and liability insurers can observe their levels of care, injurers will purchase full liability insurance coverage and their premiums will depend on their level of care; their premiums will equal $p(x)h$. Thus, injurers will want to minimize their costs of care plus premiums, or $x+p(x)h$, so they will choose the optimal level of care x^*. In this instance, liability insurance eliminates risk for injurers, and the situation reduces to the previously analysed risk-neutral case. (Victims do not bear risk either because, in the present case, they are fully compensated for their losses.)

If, however, liability insurers cannot observe levels of care, insurance policies with full coverage could create severe moral hazard, and so might not be purchased. Instead, as we know from the theory of insurance, the typical amount of coverage purchased will be partial, for that leaves injurers with an incentive to reduce risk. In this case, therefore, the liability rule results in some direct incentive to take care because injurers are left bearing some risk after their purchase of liability insurance. But levels of care will still tend to be less than first-best.

This last observation raises the question of whether the sale of liability insurance is socially desirable. (We note that because of concern about diluted incentives, liability insurance was delayed for decades in many countries and is sometimes forbidden today, such as for punitive damages.) Notwithstanding the moral hazard problem, the sale of liability insurance is socially desirable, at least in basic models of accidents and some variations of them. This is because, if the liability insurer and the injurer together have to pay for the harm caused, the insurance policy will appropriately balance the social desire to reduce harm and the social desire to reduce risk-bearing.

Parallel observations apply under the negligence rule, where the focus of concern is on the bearing of risk by victims since injurers generally will take due care and not be liable. Risk-averse potential victims will tend to purchase first-party accident insurance.

The presence of insurance implies that the liability system cannot be justified primarily as a means of compensating risk-averse victims against loss. Rather, the justification for the liability system must lie in significant part in the incentives that it creates to reduce risk. To amplify, although both strict liability and the insurance system can compensate victims, the liability system is much more expensive than the insurance system (see below). Accordingly, if there were not a social need to create incentives to reduce risk, it would be best to dispense with the liability system and to rely on insurance to accomplish compensation. On liability and insurance, see Shavell (1982a).

Administrative costs

The administrative costs of the liability system – the legal costs and effort of litigants involved in suit, settlement and trial – are substantial, generally exceeding the amounts received by victims. Consideration of administrative costs affects the comparison of liability rules, but it is not clear which rule involves greater expense: more cases are brought under strict liability than under the negligence rule (victims will not sue under the negligence rule if they believe the injurer was not negligent), but the cost of resolving a case should be greater under the negligence rule (because due care and the injurer's care level need to be ascertained). The presence of administrative costs raises the questions of whether the incentive benefits of the liability system justify incurring these costs, and whether the private incentive to sue is socially optimal. These questions are discussed in section 4.

3 Contracts

A contract is a specification of the actions that named parties are supposed to take at various times, as a function of the conditions that then obtain. A contract is said to be completely detailed, or simply *complete*, if the contract provides *explicitly* for all possible conditions. An incomplete contract may well cover all conditions by implication. A contract stating merely that a specified price will be paid for a bushel of wheat is incomplete because it does not mention many contingencies that might affect the parties. Note that such an incomplete contract has no *gaps*, as it stipulates what the parties are to do in all circumstances. Typically, incomplete contracts do not include conditions which, were they easy to include, would allow both parties to be made better off in an expected sense.

Contracts are here assumed to be enforced by a tribunal, which will usually be interpreted to be a state-authorized court, but it could also be another entity, such as an arbitrator or the decision-making body of a trade association or a religious group. (Reputation and other non-legal factors may also serve to enforce contracts, but we do not discuss these.) Enforcement refers to actions taken by the tribunal when one or more of the parties to the contract decide to come before it.

General reasons for contracts

Broadly speaking, parties make contracts when they have a need to make plans. They also want contracts enforced to prevent opportunistic behaviour that otherwise might occur during the course of the contractual relationship and stymie fulfilment of their plans.

There are two basic contexts in which parties make enforceable contracts. The first concerns virtually any kind of financial arrangement. The necessity of contract enforcement here is transparent. In financial arrangements, there is often a party who extends credit to another for some time period, and contract enforcement prevents his credit from being appropriated, which otherwise would render the arrangements impossible. For example, if borrowers were not forced to repay loans, loans would be unworkable. In addition, financial contracts that allocate risk would generally be useless without enforcement because, once the risky outcome became known, one of the parties would not wish to honour the contract.

The second context in which parties make enforceable contracts involves the supply of customized or specialized goods and services which cannot be purchased on a spot market with a simultaneous exchange for money. The need for enforcement of agreements for supply of customized goods and services inheres in several advantages: averting problems of hold-up, which might distort incentives to invest in the contractual enterprise; allocation of risk; and prevention of inappropriate breach or performance, which can result from imperfect bargaining due to sheer cost or asymmetric information.

Contract formation

The formation of contracts is of interest, in several respects. One issue concerns search effort (Diamond and Maskin, 1979). Parties expend effort in finding contractual partners, and it is apparent that their search effort will not generally be socially optimal. On the one hand, they might not search enough:

because the joint gain from contracting will generally be divided between the parties through the bargaining process, the private return to search may be less than the social return. On the other hand, parties might search more than is socially desirable because of a negative externality associated with discovery of a contract partner: when one party finds and contracts with a second, other parties are thereby prevented from contracting with that party.

A basic question that a tribunal must answer is: at what stage of interactions between parties does a contract become legally recognized? The general legal rule is that contracts are recognized if and only if both parties give a clear indication of assent, such as signing their names on a document. This rule allows parties to make enforceable contracts when they so desire, and it also protects parties from becoming legally obliged against their wishes, such as from one party's reliance on the other's statements (Bebchuk and Ben-Shahar, 2001; Wils, 1993). Mutual assent sometimes is not simultaneous; one party will make an offer and time will pass before the other agrees. An issue that this raises is how long, and under what circumstances, the offeror will want to be held to his offer, and whether he should be held to it. If an offeror is held to his terms, offerees will often be led to invest effort in investigating contractual opportunities. Otherwise, offerees might be taken advantage of by offerors if the offerees expressed serious interest after costly investigation (the offeror could change to less favourable terms). The anticipation of such offeror advantage-taking would reduce offerees' incentive to engage in investigation and thus diminish mutually beneficial contract formation (see, for example, Craswell, 1996; Katz, 1990; 1996).

Another issue of note is disclosure of information at the time of contract formation. Disclosure may be socially beneficial because the disclosed information may be desirably employed by one of the parties; for example, a buyer of a house may learn from the seller that the basement leaks and thus decide not to store valuables there. However, a disclosure obligation discourages parties from investing in acquisition of information (Kronman, 1978). For instance, an oil company contemplating buying land might decide against conducting a geological analysis of it to determine its oil-bearing potential if the company would be required to disclose its findings to the seller of the land, as the seller would then demand a price reflecting the value of the land. The social welfare consequences of the effect of a disclosure obligation on the motive to acquire information depend on whether the information is socially valuable or mere foreknowledge, on whether the party acquiring information is the buyer or the seller, and on inferences that would be drawn from silence (Shavell, 1994).

Even if both parties have given their assent, a contract will not be recognized if it was made when one of the parties was put under undue pressure – for example, if a party was physically or otherwise threatened by another. This legal rule has virtues similar to those of laws against theft; it reduces individuals' incentives to expend effort making threats and defending themselves against threats.

In addition, contracts may not be legally recognized if they are made in emergency situations, such as when the owner of a ship in distress promises to pay an exorbitant amount for rescue. Non-enforcement in such situations beneficially provides potential victims with implicit insurance against having to pay high prices, but it also reduces incentives for rescue.

Incomplete nature of contracts and their less-than-rigorous enforcement

Contracts are commonly observed to be significantly incomplete, leaving out all manner of variables and

contingencies that are of potential relevance to contracting parties. Moreover, contracts are not enforced with high sanctions, and breach is not an uncommon event.

There are three reasons for the incompleteness of contracts. The first is the cost of writing more complete contracts. The second is that some variables (effort levels, technical production difficulties) cannot be verified by tribunals. The third is that the expected consequences of incompleteness may not be very harmful to contracting parties. Incompleteness may not be harmful because a tribunal might interpret an imperfect contract in a desirable manner. Also, as will be seen, the prospect of having to pay damages for breach of contract may serve as an implicit substitute for more detailed terms. Furthermore, the opportunity to renegotiate a contract often furnishes a way for parties to alter terms in the light of circumstances for which contractual provisions had not been made.

Interpretation of contracts

Contractual interpretation, which includes a tribunal's filling gaps, resolving ambiguities, and overriding literal language, can benefit parties by easing their drafting burdens or reducing their need to understand contractual detail. For example, if it is efficient to excuse a seller from having to perform if his factory burns down, the parties need not incur the cost of specifying this exception in their contract if they can trust the tribunal to interpret their contract as if the exception were specified. A method of interpretation can be viewed formally as a function that transforms the contract individuals write into the effective contract that the tribunal will enforce. Given a method of interpretation, parties will choose contracts in a constrained-efficient way. Notably, if the parties are concerned that an aspect of their contract would not be interpreted as they want, they could either bear the cost of writing a more explicit term that would be respected by the tribunal, or they could simply accept the expected loss from having a less-than-efficient term. The socially optimal method of interpretation will take this reaction of contracting parties into account and can be regarded as minimizing the sum of the costs the parties bear in writing contracts and the losses resulting from inefficient enforcement. (See Ayres and Gertner, 1989; Hadfield, 1994; Schwartz, 1992; Shavell, 2006.)

Damage measures for breach of contract

When parties breach a contract, they often have to pay damages in consequence. The damage measure, the formula governing what they should pay, can be determined by the tribunal or it can be stipulated in advance by the parties to the contract. One would expect parties to specify their own damage measure when it would better serve their purposes than the measure the tribunal would employ, and otherwise to allow the tribunal to select the damage measure. In either case, we now examine the utility of different damage measures to contracting parties, assuming initially that there is no renegotiation of contracts. Clearly, the prospect of having to pay damages provides an incentive to perform contractual obligations, and thus generally promotes enforcement of contracts and the goals of the parties. Under the commonly employed *expectation measure*, damages equal the amount that compensates the victim of breach for his losses. Under this measure, a seller contemplating breach will be induced to perform if the cost of performance to the seller is less than the value of performance to the buyer, and to breach otherwise. Because the expectation measure leads to maximization of joint value, it would be chosen by the parties

(ignoring consideration of investment incentives and risk bearing), as emphasized by Shavell (1980a).
Another commonly employed measure of damages is the *reliance measure*: damages equal to the amount spent by the victim relying on contract performance, such as expenditures on advertising an entertainer who has contracted to appear at one's nightclub.
The point that the expectation measure of damages induces efficient performance of parties sheds light on the view of many legal commentators that breach is immoral. This view fails to account for the fact that contracts that are breached are generally incomplete, and that breach constitutes behaviour that the parties truly want and would have provided for in a complete contract.
Damage measures not only affect performance, they also influence the *ex ante* motive to make investments in reliance on contract performance. Under the expectation measure, reliance investments tend to exceed efficient levels: the buyer will treat an investment (like advertising an entertainer) as one with a sure payoff, since he will receive either performance or expectation damages, whereas the actual return to the investment is uncertain, due to the possibility of breach (advertising will be a waste if the entertainer does not appear); see Shavell (1980a). This tendency toward over-reliance stands in contrast to the problem of inadequate reliance investment associated with lack of contract enforcement.
Damage measures affect risk-bearing as well as incentives. Notably, because the expectation measure compensates the victim of a breach, the measure might be mutually desirable as a form of insurance if the victim is risk averse (Polinsky, 1983). However, the prospect of having to pay damages also constitutes a risk for a party who might commit breach (such as a seller whose costs suddenly rise), and he might be risk averse as well. The latter consideration may lead parties to want to lower damages or to employ damages less frequently by writing more detailed contracts (for instance, the parties could go to the expense of specifying in the contract that a seller can be excused from performance if his costs are unusually high).

Specific performance as a remedy for breach

An alternative to use of a damage measure for breach of contract is specific performance: requiring a party to satisfy his contractual obligation. Specific performance can be accomplished with a sufficiently high threat or by exercise of the state's police powers, such as by a sheriff removing a person from the land that he promised to convey. (Note that, if a monetary penalty can be employed to induce performance, then specific performance is equivalent to a damage measure with a high level of damages.)
It is apparent from what has been said about incomplete contracts and damage measures that parties should not want specific performance of many contracts that they write, for they do not wish their incomplete contracts always to be performed. It is therefore not surprising that, in fact, specific performance is not used as the remedy for breach for most contracts for production of goods or for provision of services. Additionally, specific performance might be peculiarly difficult to enforce in these contexts because of problems in monitoring and controlling parties’ effort levels and the quality of production.
However, specific performance does have advantages for parties in certain contexts, such as in contracts for the transfer of things that already exist, like land, and specific performance is the usual legal remedy for sellers’ breaches of contracts for the sale of land.

Renegotiation of contracts

Parties often have the opportunity to renegotiate their contracts when problems arise. Indeed, the assumption that they will do this has appeal because, having made an initial contract, the parties know of each other's existence and of many particulars of the contractual situation. For this reason, much of the economics literature (as opposed to law and economics literature) on contracts assumes that renegotiation always occurs and that, due to symmetric information between the parties, it always results in efficient performance. Hence, damage measures for breach of contract, or more generally, the mechanisms that the parties stipulate in their contracts, establish the threat points for renegotiation. If properly designed, the mechanisms can foster beneficial incentives to invest *ex ante* for both parties. On this extensive literature, see, for example, Rogerson (1984), Hart (1987), Hart and Moore (1988), and Bolton and Dewatripont (2005).

Legal overriding of contracts

A basic rationale for legislative or judicial overriding of contracts is the presence of externalities. Contracts that are likely to harm third parties are often not enforced, including, for example, agreements to commit crimes, price-fixing compacts, liability insurance policies against fines, and certain sales contracts (such as for machine guns).

Another general rationale for non-enforcement of contracts is to prevent a loss in welfare to one or both of the parties to a contract. This concern may justify non-enforcement when a party is incompetent, lacks relevant information, or is in an emergency situation. The rationale also applies in the context of contract interpretation by tribunals. As noted, contract interpretation may amount to the overriding of a written contractual term, and this practice may promote the welfare of contracting parties by allowing them to save writing costs, given that courts will step in and correct inefficient terms.

Additionally, contracts sometimes are not enforced because they involve the sale of things said to be inalienable, such as human organs, babies, and voting rights. In many of these cases, the inalienability justification for lack of enforcement can be recognized as involving externalities or the welfare of the contracting parties.

4 . Litigation

We here consider the bringing and adjudication of lawsuits: the decision of a party who has suffered a loss whether to sue; the choice of the litigants whether to settle with each other or instead go to trial; and the choice of litigants, before or during trial, of how much to spend on litigation.

Suit

As a general rule, a party who has suffered loss, the plaintiff, will sue when the cost of suit c_P is less than the expected benefits from suit. The expected benefits from suit incorporate potential settlements or trial outcomes, but assume for simplicity that, if suit is brought, the plaintiff obtains for sure a judgment equal to harm suffered, h. Thus the plaintiff will sue when his litigation cost, c_P, is less than h.

(Obviously, if there is only a probability p of winning this amount, a risk-neutral plaintiff would sue when $c_P<ph$; and a risk-averse plaintiff would be less likely to sue.)

The private incentive to sue is fundamentally misaligned with the socially optimal incentive to sue, as emphasized by Shavell (1982b; 1997). The deviation could be in either direction. On the one hand, there is a divergence between private and social costs that can lead to socially excessive suit: when a plaintiff contemplates bringing suit, he bears only his own costs; he does not take into account the defendant's costs or the state's costs that his suit will engender. On the other hand, there is a difference between the private and social benefits of suit that can either lead to a socially inadequate level of suit or reinforce the cost-related tendency towards excessive suit. Specifically, the plaintiff considers his private benefit from suit (the gain he would obtain from prevailing) but not the social benefit (the deterrent effect on the behaviour of injurers generally). The private gain could be larger or smaller than the social benefit.

To illustrate, suppose that liability is strict. As stated, victims will sue if and only if $c_P<h$. Let x be the precaution expenditures that injurers will be induced to make if there is suit, q the probability of harm if suit is not brought, and q' the probability of harm if suit is brought. (Thus, q' will be less than q if x is spent on precautions.) Suit will be socially worthwhile if and only if

$q'(c_P+c_D+c_S)<(q-q')h-x$, where c_D is the defendant's litigation cost and c_S is the state's cost.

In other words, suit is socially worthwhile if the expected litigation costs are less than the deterrence benefits of suit net of the cost of precautions. The condition for victims to sue and the condition for suit to be socially optimal are very different. Whether victims will sue does not depend on the costs c_D and c_S. Moreover, the private benefit of suit is what the victim will receive as a damages award, h; in contrast, the social benefit is the harm weighted by the reduction in the accident probability, $q-q'$, net of the cost of precautions, x. It is evident, therefore, that victims might sue when suit is not socially desirable, or that victims might not sue even when suit would be socially beneficial.

The main implication of the private-social divergence is that state intervention may be desirable, either to correct a problem of excessive suit (notably by taxing suit or barring it in some domain) or a problem of inadequate suit (by subsidizing suit in some way). For the state to determine optimal policy, however, requires it to estimate the effects of suit on injurer behaviour and weigh them against the social costs of suit.

The importance of the private-social divergence in incentives to sue may be substantial. This is suggested by the high costs of using the legal system; indeed, legal costs may on average actually equal the amounts received by those who sue. Hence, the incentives created by the legal system must be significant to justify its use. Regardless of whether the legal system creates valuable incentives, however, the private motive to bring suit may be great, giving rise to a reason for social intervention. Conversely, in some domains the incentive to sue may be low (say, damages per plaintiff are not great) even though the value of deterrence is significant. This might justify the state's encouraging litigation.

Settlement versus trial

Assuming that a suit has been brought, we now consider whether parties will reach a settlement or go to trial. A settlement is a legally enforceable contract, usually involving a payment from the defendant to the plaintiff, in return for which the plaintiff agrees not to pursue his claim further. If the parties do not

reach a settlement, we assume that they go to trial, that is, that some tribunal determines the outcome of their case. In fact, the vast majority of cases settle.

One model of the settlement-versus-trial decision presumes that the parties have somehow each come to a belief about the probability of the trial outcome (Posner, 2003, ch. 21; Shavell, 2004, ch. 17). Let p_P represent the plaintiff's opinion about his probability of prevailing, and let p_D be the defendant's opinion about that same probability. Let *w* be the amount that would be won (for simplicity assume that they agree about *w*). Assume also that the parties are risk neutral. The plaintiff's expected gain from trial, net of his litigation costs, is $p_P w - c_P$. The defendant's expected loss from trial, including his litigation costs, is $p_D w + c_D$. Hence, a settlement is possible if and only if $p_P w - c_P > p_D w + c_D$, in which case the settlement amount will be in the settlement range $[p_P w - c_P, p_D w + c_D]$. Note that, if the parties agree on the plaintiff's probability of prevailing, a settlement is feasible. A settlement range does not exist, and therefore trial will occur, if $p_P w - p_D w > c_P + c_D$. Risk aversion of the parties increases the size of the settlement range and thus, one presumes, makes settlement more likely: if the plaintiff is risk averse, he will be willing to settle for less than $p_P w - c_P$; and if the defendant is risk averse, she will be willing to pay more than $p_D w + c_D$.

The model just discussed does not explain the origin of the parties' beliefs and does not include a description of rational bargaining between them. Subsequently, standard asymmetric information models of settlement versus litigation were examined (Bebchuk, 1984; Reinganum and Wilde, 1986; Schweizer, 1989; Spier, 1992; Hay and Spier, 1998; Daughety, 2000). In a simple model of this type, there is one-sided asymmetry of information and the party without private information makes a take-it-or-leave-it settlement proposal. For example, the plaintiff makes a demand *x* to the defendant, who has private information about the probability *p* that he will lose at trial. If $pw + c_D < x$, the defendant will reject the demand and the plaintiff will therefore obtain only $pw - c_P$, but if $pw + c_D > x$, the defendant will accept and pay *x*. The plaintiff chooses *x* to maximize his expected payoff from settlement or trial. The higher his demand *x*, the more he will obtain if it is accepted, but the greater the likelihood of rejection and thus of his bearing trial costs. At the optimal demand for the plaintiff, there will generally be a positive probability of trial and also of settlement.

The virtues of such asymmetric information models are twofold. First, they include an explicit account of bargaining and thus of the probability of settlement and the magnitude of the settlement offer or demand. (The outcomes of these models depend, however, on essentially arbitrary modelling choices, such as whether the informed or the uninformed party makes the settlement proposal.) Second, the models explain differences of opinion that give rise to trial in terms of differences in possession of information. (However, the models do not account for why there should be differences in information, given that the parties have incentives to share information and may be forced to do so through legal discovery.)

The private and social incentives to settle generally diverge for several reasons. First, because the litigants do not bear all of the costs of a trial (such as the salaries of judges and the forgone value of juror time), they save less by settling than society does, which tends to make the private incentive to settle socially inadequate. Second, when there is asymmetric information, parties will fail to settle when the plaintiff's demand turns out to have been too high or the defendant's offer too low. But their desire to obtain from each other a greater share of the benefit from settling does not itself translate into any social benefit. Third, the prospect of settlement may reduce deterrence because defendants gain from

settlement.

Litigation expenditures

A plaintiff will continue spending on litigation as long as this raises his expected return from settlement or trial (net of litigation costs), and a defendant will make such expenditures as long as this lowers his expected total outlays. The effects of each litigant's expenditures will generally depend on what the other does, and the two will often be spending to rebut one another.

There are several reasons why the private and social incentives to spend on litigation diverge. First, to the extent that their expenditures simply offset each other, without altering trial or settlement outcomes, the expenditures constitute a social waste. Second, the litigants' trial expenditures may mislead the tribunal rather than enhance the accuracy of the outcome, which has negative social value. Third, even if trial expenditures do improve the accuracy of outcomes, they may not be socially optimal in magnitude, for the parties consider only how their expenditures influence the litigation outcome, without regard to their influence (if any) on deterrence.

Because private and social incentives to spend on litigation may diverge, it may be beneficial for expenditures to be either curtailed or encouraged. In practice, courts often restrict the legal effort that parties can undertake, for example by limiting the extent of discovery and the number of testifying experts.

Other topics

A number of other topics that relate to litigation and the legal process have been studied, including the selection of suits for litigation (Priest and Klein, 1984); the accuracy of adjudication (Kaplow, 1994; Png, 1986); 'discovery', that is, mandated disclosure of information during litigation (Shavell, 1989); and the appeals process (Daughety and Reinganum, 2000; Shavell, 1995; Spitzer and Talley, 2000).

5 . Public law enforcement and criminal law

Law enforcement often is the result of the efforts of public agents, such as inspectors, tax auditors, and police. We here discuss certain characteristics of optimal public law enforcement. As noted, this subject was first analysed by Bentham (1789) and Becker (1968) (for a survey, see Polinsky and Shavell, 2000).

Rationale of public enforcement

A basic question is why there is a need for public enforcement of law in the light of the availability of private suits brought by victims (Becker and Stigler, 1974; Landes and Posner, 1975; Polinsky, 1980). The answer depends importantly on the locus of information about the identity of injurers. When victims of harm naturally possess knowledge of the identity of injurers, allowing private suits for damages will motivate victims to sue and thus harness the information they have for purposes of law enforcement. This may help to explain why the enforcement of contractual obligations and of accident law is primarily private. When victims do not know who caused harm, however, or when finding injurers is

difficult, society tends to rely instead on public investigation and prosecution; this is broadly true of crimes and of many violations of environmental and safety regulations.

Basic framework for analysing public enforcement

Suppose that, if an individual commits a harmful act, he obtains a gain and also faces the risk of being caught and sanctioned. The sanction could be a fine or a prison term. Fines will be treated as socially costless because they are mere transfers of money, whereas imprisonment is socially costly because of the expense of operating prisons and the disutility suffered by those imprisoned (which is not offset by gains to others). The higher the probability is of detecting and sanctioning violators, the more resources the state must devote to enforcement.

We assume that social welfare equals the sum of individuals' expected utilities. If individuals are risk neutral, social welfare can be expressed as the gains individuals obtain from committing their harmful acts, minus the harms caused and the costs of law enforcement. The enforcement authority's problem is to maximize social welfare by choosing enforcement expenditures, or, equivalently, a probability of detection, the form of sanctions, and their level.

Fines

Suppose that the sanction is a fine and that individuals are risk neutral. Then the optimal level of the fine is maximal, f_M, as emphasized in Becker (1968). If the fine were not maximal, society could save enforcement costs by simultaneously raising the fine and lowering the probability without affecting the level of deterrence. Formally, if $f<f_M$, then raise the fine to f_M and lower the probability from p to $(f/f_M)p$; the expected fine is still pf, so that deterrence is maintained, but expenditures on enforcement are reduced, implying that social welfare rises. Moreover, the optimal probability is such that there is some under-deterrence; in other words, at the optimal p the expected fine pf_M is less than the harm h. The reason for this result is that, if pf_M equals h, behaviour will be ideal, in which case decreasing p must be socially beneficial because the individuals thereby induced to commit the harmful act cause no net social losses (because their gains essentially equal the harm), but reducing p saves enforcement costs.

If individuals are risk averse, the optimal fine may well be below the maximal fine, as stressed in Polinsky and Shavell (1979). This is because the use of a very high fine would impose a substantial risk-bearing cost on individuals who commit harmful acts.

Imprisonment

Now suppose that the sanction is imprisonment and that individuals are risk neutral in imprisonment. Then the optimal imprisonment term is maximal. The reasoning is similar to that employed above with respect to fines: if the imprisonment term were not maximal, it could be raised and the probability of detection lowered so as to keep the expected prison term constant; neither individual behaviour nor the costs of imposing imprisonment are affected (because the expected prison term is the same), but enforcement expenditures fall.

If, instead, individuals are risk averse in imprisonment (the disutility of each additional year of

imprisonment grows with the number of years in prison), there is a stronger argument for setting the imprisonment sanction maximally than when individuals are risk neutral. Now, when the imprisonment term is raised, the probability of detection can be lowered even more than in the risk-neutral case without reducing deterrence. Thus, not only are there greater savings in enforcement expenditures, but the social costs of imposing imprisonment sanctions decline because the expected prison term falls. Last, suppose that individuals are risk preferring in imprisonment (the disutility of each additional year of imprisonment declines with the number of years in prison). This possibility seems particularly important: the first years of imprisonment may create unusually high disutility, due to brutalization of the prisoner or due to the stigma of having been imprisoned at all. In addition, individuals generally have positive time discount rates, which are thought to be especially significant for criminals. In the case of risk-preferring individuals, the optimal prison term may well be less than maximal: if the sentence were raised, the probability that maintains deterrence could not be lowered proportionally, implying that the expected prison term would rise. Thus, although there would be enforcement-cost savings, they might not be great enough to offset the increased sanctioning costs.

Fines versus imprisonment

Fines generally are preferable to prison terms as a means of deterrence, since fines are socially cheaper sanctions to impose (Becker, 1968). Hence, fines should be employed to the greatest extent possible – until a party's wealth is exhausted – before imprisonment is imposed. Further, imprisonment should be used as a sanction only if the harm prevented by the added deterrence is sufficiently great.

Fault-based liability

Our discussion so far has presumed that liability is strict, but liability may also be based on fault, an assessment of whether the act that caused harm was socially undesirable (analogous to the negligence rule and due-care standard discussed above in the accident context). Fault-based liability, like strict liability, can induce individuals to behave properly, but fault-based liability possesses an advantage when individuals are risk averse: if they act responsibly, they will not be found at fault, so will not bear the risk of being sanctioned. Similarly, fault-based liability is advantageous when the form of the sanction is imprisonment, for then, again, individuals may be led to behave optimally without the actual imposition of sanctions, and thus without social costs being incurred (Shavell, 1987b). To the extent that mistakes are made in determining fault, however, these two advantages are reduced because risk is imposed and sanctioning costs are incurred. Note, too, that fault-based liability is more difficult to implement, because it requires the state to determine optimal behaviour.

Incapacitation

Society may reduce harm not only through deterrence but also by imposing sanctions that remove parties from positions in which they are able to cause harm, that is, by incapacitating them. Imprisonment is the primary incapacitative sanction, although there are other examples: individuals can lose their drivers' licences, businesses can lose their right to operate in certain domains, and the like.

Suppose that the sole function of imprisonment is to incapacitate. Then it will be desirable to keep someone in jail as long as the reduction in crime from incapacitating him exceeds the costs of imprisonment (Shavell, 1987c). Although this condition could hold for a long period, it is unlikely to unless the harm prevented is very high, because the proclivity to commit crimes apparently declines sharply with age.

Note that, as a matter of economic logic, the incapacitation rationale might imply that a person should be imprisoned even if he has not committed a crime – because the danger he poses to society makes incapacitating him worthwhile. In practice, however, the fact that a person has committed a harmful act may be the best basis for predicting his future behaviour, in which case the incapacitation rationale would suggest imprisoning an individual only if he has committed such an act.

Two observations are worth noting about optimal enforcement when incapacitation is the goal as opposed to when deterrence is the goal. First, when enforcement is based on incapacitation, the optimal magnitude of the sanction is independent of the probability of apprehension, which contrasts with the case when enforcement is based on deterrence. Second, when enforcement is deterrence-oriented, the probability and magnitude of sanctions depend on the ability to deter, and, if this ability is limited (as, for instance, with the insane), a low expected sanction may be optimal, whereas a high sanction still might be called for to incapacitate.

Other issues

A number of other topics have been studied in the economic analysis of public law enforcement, including mistake, marginal deterrence (the effect of sanctions in reducing the severity of harm a party causes), self-reporting of violations (Kaplow and Shavell, 1994a; Innes, 1999), repeat offences, plea bargaining (Reinganum, 1988), general enforcement (when detection resources simultaneously influence the deterrence of a range of harmful acts) (Mookherjee and Png, 1992; and Shavell, 1991), and corruption of law-enforcement agents (Shleifer and Vishny, 1993; Rose-Ackerman, 1999; and Polinsky and Shavell, 2001).

Criminal law

The subject of criminal law may be viewed in the light of the theory of public law enforcement (Posner, 1985; Shavell, 1985). First, the fact that the acts in the core area of crime (robbery, murder, rape, and so forth) are punished by the sanction of imprisonment makes basic sense. Were society to rely on fines alone, deterrence of the acts in question would be grossly inadequate. Notably, the probability of detecting many of these acts is low, making the money sanction necessary for deterrence high, but the assets of individuals who commit these acts often are insubstantial. Hence, the threat of prison is needed for deterrence. Moreover, the incapacitative aspect of imprisonment is valuable because of the difficulty of deterring individuals who are prone to commit criminal acts.

Second, many of the doctrines of criminal law appear to enhance social welfare. This seems true of the basic feature of criminal law that punishment is not imposed on all harmful acts, but instead is usually confined to those that are undesirable. (For example, murder is subject to criminal sanctions, but not all accidental killing is.) As we have stressed, when the socially costly sanction of imprisonment is

employed, the fault system is desirable because it results in less frequent imposition of punishment than strict liability. Also, the focus on intent in criminal law as a precondition for imposing sanctions may be sensible with regard to deterrence because those who intend to do harm are more likely to conceal their acts, and may be harder to discourage because of the benefits they anticipate. That unsuccessful attempts to do harm are punished in criminal law is an implicit way of raising the likelihood of sanctions for undesirable acts. Study of specific doctrines of criminal law seems to afford a rich opportunity for economic analysis.

6 . Criticism of economic analysis of law

Many observers, and particularly non-economists, view economic analysis of law with scepticism. We consider several such criticisms here.

Description of behaviour

It is sometimes claimed that individuals and firms do not respond to legal rules as rational maximizers of their well-being. For example, it is often asserted that decisions to commit crimes are not governed by economists’ usual assumptions. Some sceptics also suggest that, in predicting individuals’ behaviour, certain standard assumptions are inapplicable. For example, in predicting compliance with a law, the assumption that preferences be taken as given would be inappropriate if a legal rule would change people's preferences, as some say was the case with civil rights laws and environmental laws. In addition, laws may frame individuals’ understanding of problems, which could affect their probability assessments or willingness to pay. The emerging field of behavioural economics, as well as work in various disciplines that address social norms, is beginning to examine these sorts of issues (Jolls, Sunstein and Thaler, 1998).

Distribution of income

A frequent criticism of economic analysis of law concerns its focus on efficiency to the exclusion of the distribution of income. The claim of critics is that legal rules should be selected in a manner that takes into account their effects on the rich and the poor. But achieving sought-after redistribution through income tax and transfer programmes tends to be superior to redistribution through the choice of legal rules. This is because redistribution through legal rules and the tax-transfer system both will distort individuals’ labour-leisure decisions in the same manner, but redistribution through legal rules often will require choosing an inefficient rule, which imposes an additional cost (Shavell, 1981; Kaplow and Shavell, 1994b).

Moreover, it is difficult to redistribute income systematically through the choice of legal rules. Many individuals are never involved in litigation; and for those who are there is substantial income heterogeneity among plaintiffs as well as among defendants. Additionally, in contractual contexts the choice of a legal rule often will not have any distributional effect because contract terms, notably the price, will adjust, so that any agreement into which parties enter will continue to reflect the initial distribution of bargaining power between them.

Concerns for fairness

An additional criticism is that the conventional economic approach slights important concerns about fairness, justice and rights. Some of these notions refer implicitly to the appropriateness of the distribution of income and, accordingly, are encompassed by our preceding remarks. Also, to some degree, the notions are motivated by instrumental concerns. For example, the attraction of paying fair compensation to victims must derive in part from the beneficial risk reduction effected by such payments, and the appeal of obeying contractual promises must rest in part on the desirable consequences contract performance has on production and exchange. To some extent, therefore, critics' concerns are already taken into account in standard economic analysis.

However, many who promote fairness, justice and rights do not regard these notions merely as some sort of proxy for attaining instrumental objectives. Instead, they believe that satisfying these notions is intrinsically valuable. This view also can be partially reconciled with the economic conception of social welfare: if individuals have a preference for a legal rule or institution because they regard it as fair, that should be credited in the determination of social welfare, just as any other preference should.

But many commentators take the position that conceptions of fairness are important as ethical principles in themselves, without regard to any possible relationship the principles may have to individuals' welfare. This opinion is the subject of long-standing debate among moral philosophers. Some readers may be sceptical of normative views that are not grounded in individuals' well-being because embracing such views entails a willingness to sacrifice individuals' well-being. Indeed, consistently pursuing any non-welfarist principle must sometimes result in everyone being made worse off (see Kaplow and Shavell, 2001; 2002).

Efficiency of judge-made law

Also criticized is the contention of some economically oriented legal academics, notably Posner (1972), that judge-made law tends to be efficient (in contrast to legislation, which is said to reflect the influence of special interest groups). Some critics believe that judge-made law is guided by notions of fairness, or is influenced by legal culture or judges' biases, and thus will not necessarily be efficient. Whatever is the merit of the critics' claims, they are descriptive assertions about the law, and their validity does not bear on the power of economics to predict behaviour in response to legal rules or on the value of normative economic analysis of law.

See Also

- Coase theorem
- law, public enforcement of
- property law, economics and
- uncertainty
- welfare economics

Bibliography

Ayres, I. and Gertner, R. 1989. Filling gaps in incomplete contracts: an economic theory of default rules. *Yale Law Journal* 99, 87–130.

Bebchuk, L. 1984. Litigation and settlement under imperfect information. *RAND Journal of Economics* 15, 404–15.

Bebchuk, L. and Ben-Shahar, O. 2001. Pre-contractual reliance. *Journal of Legal Studies* 30, 423–57.

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Becker, G. and Stigler, G. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3, 1–18.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*, in *The Utilitarians*. Garden City, NY: Anchor Books, 1973.

Blume, L., Rubinfeld, D. and Shapiro, P. 1984. The taking of land: when should compensation be paid? *Quarterly Journal of Economics* 99, 71–92.

Bolton, P. and Dewatripont, M. 2005. *Contract Theory*. Cambridge, MA: MIT Press.

Brown, J. 1973. Toward an economic theory of liability. *Journal of Legal Studies* 2, 323–49.

Calabresi, G. 1961. Some thoughts on risk distribution and the law of torts. *Yale Law Journal* 70, 499–553.

Calabresi, G. 1970. *The Costs of Accidents: A Legal and Economic Analysis*. New Haven, CT: Yale University Press.

Calabresi, G. and Melamed, A. 1972. Property rules, liability rules, and inalienability: one view of the cathedral. *Harvard Law Review* 85, 1089–128.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Cooter, R. and Ulen, T. 2003. *Law and Economics*, 4th edn. Reading, MA: Addison-Wesley.

Craswell, R. 1996. Offer, acceptance, and efficient reliance. *Stanford Law Review* 48, 481–553.

Daughety, A. 2000. Settlement. In *Encyclopedia of Law and Economics*, vol. 5, ed. B. Bouckaert and G. De Geest. Cheltenham: Edward Elgar.

Daughety, A. and Reinganum, J. 2000. Appealing judgments. *RAND Journal of Economics* 31, 502–25.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review: Papers and Proceedings* 57, 347–59.

Diamond, P. 1974. Single activity accidents. *Journal of Legal Studies* 3, 107–64.

Diamond, P. and Maskin, E. 1979. An equilibrium analysis of search and breach of contract, I: steady states. *Bell Journal of Economics* 10, 282–316.

Hadfield, G. 1994. Judicial competence and the interpretation of incomplete contracts. *Journal of Legal Studies* 23, 159–84.

Hart, O. 1987. Incomplete contracts. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. New York: Macmillan.

Hart, O. and Moore, J. 1988. Incomplete contracts and renegotiation. *Econometrica* 56, 755–8.

Hay, B. and Spier, K. 1998. Settlement of litigation. In *The New Palgrave Dictionary of Economics and the Law*, vol. 3, ed. P. Newman. London: Macmillan.

Innes, R. 1999. Remediation and self-reporting in optimal law enforcement. *Journal of Public Economics* 72, 379–93.

Jolls, C., Sunstein, C. and Thaler, R. 1998. A behavioral approach to law and economics. *Stanford Law Review* 50, 1471–550.

Kaplow, L. 1986. An economic analysis of legal transitions. *Harvard Law Review* 99, 509–617.

Kaplow, L. 1994. The value of accuracy in adjudication: an economic analysis. *Journal of Legal Studies* 23, 307–401.

Kaplow, L. and Shavell, S. 1994a. Optimal law enforcement with self-reporting of behavior. *Journal of Political Economy* 102, 583–606.

Kaplow, L. and Shavell, S. 1994b. Why the legal system is less efficient than the income tax in redistributing income. *Journal of Legal Studies* 23, 667–681.

Kaplow, L. and Shavell, S. 2001. Any non-welfarist method of policy assessment violates the Pareto principle. *Journal of Political Economy* 109, 281–86.

Kaplow, L. and Shavell, S. 2002. *Fairness versus Welfare*. Cambridge, MA: Harvard University Press.

Katz, A. 1990. The strategic structure of offer and acceptance: game theory and the law of contract formation. *Michigan Law Review* 89, 215–95.

Katz, A. 1996. When should an offer stick? The economics of promissory estoppel in preliminary negotiations. *Yale Law Journal* 105, 1249–1309.

Kronman, A. 1978. Mistake, disclosure, information, and the law of contracts. *Journal of Legal Studies* 7, 1–34.

Landes, W. and Posner, R. 1975. The private enforcement of law. *Journal of Legal Studies* 4, 1–46.

Landes, W. and Posner, R. 1987a. *The Economic Structure of Tort Law*. Cambridge, MA: Harvard University Press.

Landes, W. and Posner, R. 1987b. Trademark law: an economic perspective. *Journal of Law and Economics* 30, 265–309.

Landes, W. and Posner, R. 2003. *The Economic Structure of Intellectual Property Law*. Cambridge, MA: Harvard University Press.

Libecap, G. 1986. Property rights in economic history: implications for research. *Explorations in Economic History* 23, 227–52.

Miceli, T. 1997. *Economics of the Law: Torts, Contracts, Property, Litigation*. New York: Oxford University Press.

Mookherjee, D. and Png, I. 1992. Monitoring vis-à-vis investigation in enforcement of law. *American Economic Review* 82, 556–65.

Png, I. 1986. Optimal subsidies and damages in the presence of judicial error. *International Review of Law and Economics* 6, 101–5.

Polinsky, A.M. 1980. Private versus public enforcement of fines. *Journal of Legal Studies* 9, 105–27.

Polinsky, A.M. 1983. Risk sharing through breach of contract remedies. *Journal of Legal Studies* 12, 427–44.

Polinsky, A.M. and Shavell, S. 1979. The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* 69, 880–91.

Polinsky, A.M. and Shavell, S. 2000. The economic theory of public enforcement of law. *Journal of Economic Literature* 38, 45–76.

Polinsky, A.M. and Shavell, S. 2001. Corruption and optimal law enforcement. *Journal of Public Economics* 81, 1–24.

Polinsky, A.M. and Shavell, S., eds. 2007. *Handbook of Law and Economics*, vol. 1. Amsterdam: North-Holland.

Posner, R. 1972. *Economic Analysis of Law*. Boston: Little, Brown and Company.

Posner, R. 1985. An economic theory of the criminal law. *Columbia Law Review* 85, 1193–231.

Posner, R. 2003. *Economic Analysis of Law*, 6th edn. New York: Aspen Publishers.

Priest, G. and Klein, B. 1984. The selection of disputes for litigation. *Journal of Legal Studies* 13, 1–55.

Reinganum, J. 1988. Plea bargaining and prosecutorial discretion, *American Economic Review* 78, 713–28.

Reinganum, J. and Wilde, L. 1986. Settlement, litigation, and the allocation of litigation costs. *RAND Journal of Economics* 17, 557–66.

Rogerson, W. 1984. Efficient reliance and damage measures for breach of contract. *RAND Journal of Economics* 15, 39–53.

Rose-Ackerman, S. 1999. *Corruption and Government: Causes, Consequences and Reform*. New York: Cambridge University Press.

Schwartz, A. 1992. Relational contracts in the courts: an analysis of incomplete agreements and judicial strategies. *Journal of Legal Studies* 21, 271–318.

Schweizer, U. 1989. Litigation and settlement under two-sided incomplete information. *Review of Economic Studies* 56, 163–78.

Shavell, S. 1980a. Damage measures for breach of contract. *Bell Journal of Economics* 11, 466–90.

Shavell, S. 1980b. Strict liability versus negligence. *Journal of Legal Studies* 9, 1–25.

Shavell, S. 1981. A note on efficiency vs distributional equity in legal rulemaking: should distributional equity matter given optimal income taxation? *American Economic Review: Papers and Proceedings* 71, 414–18.

Shavell, S. 1982a. On liability and insurance. *Bell Journal of Economics* 13, 120–32.

Shavell, S. 1982b. The social versus the private incentive to bring suit in a costly legal system. *Journal of Legal Studies* 11, 333–9.

Shavell, S. 1985. Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* 85, 1232–62.

Shavell, S. 1987a. *Economic Analysis of Accident Law*. Cambridge, MA: Harvard University Press.

Shavell, S. 1987b. The optimal use of nonmonetary sanctions as a deterrent. *American Economic Review* 77, 584–92.

Shavell, S. 1987c. A model of optimal incapacitation. *American Economic Review: Papers and Proceedings* 77, 107–110.

Shavell, S. 1989. Sharing of information prior to settlement or litigation. *RAND Journal of Economics* 20, 183–95.

Shavell, S. 1991. Specific versus general enforcement of law. *Journal of Political Economy* 99, 1088–108.

Shavell, S. 1993. The optimal structure of law enforcement. *Journal of Law and Economics* 36, 255–87.

Shavell, S. 1994. Acquisition and disclosure of information prior to sale. *RAND Journal of Economics* 25, 20–36.

Shavell, S. 1995. The appeals process as a means of error correction. *Journal of Legal Studies* 24, 379–426.

Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26, 575–612.

Shavell, S. 2004. *Foundations of Economic Analysis of Law*. Cambridge, MA: Harvard University Press.

Shavell, S. 2006. On the writing and interpretation of contracts. *Journal of Law, Economics, &*

Organization 22, 289–314.

Shleifer, A. and Vishny, R. 1993. Corruption. *Quarterly Journal of Economics* 108, 599–617.

Spier, K. 1992. The dynamics of pretrial negotiation. *Review of Economic Studies* 59, 93–108.

Spitzer, M. and Talley, E. 2000. Judicial auditing. *Journal of Legal Studies* 29, 649–83.

Umbeck, J. 1981. *A Theory of Property Rights with Application to the California Gold Rush*. Ames: Iowa State University Press.

Wils, W. 1993. Who should bear the costs of failed negotiations? A functional inquiry into precontractual liability. *Journal des Economistes et des Etudes Humaines* 4, 93–134.

How to cite this article

Polinsky, A. Mitchell and Steven Shavell. "law, economic analysis of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000038> doi:10.1057/9780230226203.0940

The New Palgrave Dictionary of Economics Online

Law, John (1671 – 1729)

Michael D. Bordo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Banque Royale; commodity money; convertibility; disequilibrium theory of money; fiat money; Law, J.; medium of exchange; Mississippi Bubble; money supply; national banks; paper money; real bills doctrine; speculation

Article

John Law of Lauriston has been regarded by some observers as a monetary crank, by others as a precursor of modern schemes of managed money and Keynesian full-employment policies. He was the originator of the Mississippi Bubble, perhaps the greatest speculative bubble of all time.

Born in Edinburgh, the son of prosperous parents, Law was well educated in political economy. A fugitive from justice in 1694 for killing a man in a duel in England, Law travelled extensively throughout Europe, observing and gaining experience in banking, insurance and finance. He proposed a number of unsuccessful schemes to set up a national bank of issue – in Paris in 1702, Edinburgh in 1705 and Savoy in 1712 – finally attaining success in France with the establishment in 1718 of the Banque Royale.

Law's theories on money and banking are contained in *Money and Trade Considered: With a Proposal for Supplying the Nation With Money* (1705) and other works (Hamilton, 1968; Harsin, 1934). Like other 18th-century writers Law adopted a disequilibrium theory of money, viewing it as a stimulant to trade. In a state of unemployment, Law maintained that an increase in the nation's money supply would stimulate employment and output without raising prices since the demand for money would rise with the increase in output. Moreover, once full employment was attained the monetary expansion would attract factors of production from abroad, so output would continue to increase.

According to Law, a paper-money standard was preferable to one based on precious metals. Suitable candidates for the money supply included government fiat, banknotes, stocks and bonds. Since the primary function of money was as a medium of exchange, it could best be served by a commodity (paper) not subject to considerable fluctuation in value and high resource costs. Thus Law advocated the establishment of note-issuing national banks that would extend productive loans (real bills), providing

sufficient currency to guarantee prosperity. Two proposals for such banks, in Paris 1702 and Edinburgh 1705, would have had the note issues based on land initially valued in terms of silver.

From 1716 to 1720 John Law had the unique opportunity to apply his theories to the French economy. In 1715, the heritage of two exhausting wars was depression and deflation. Law succeeded in convincing the Regent (the Duke of Orleans) that a bank of issue would alleviate the problem of financing the national debt. Accordingly, he established in Paris on 2 May 1716 a private bank, the Banque Générale. In its 31 months of operation, the bank was remarkably successful; its notes (convertible into specie and payable as taxes) were issued in moderation and gained national circulation. On 4 December 1718, the Banque Générale was nationalized and renamed the Banque Royale, with Law in control, and in January 1719 it began to issue notes denominated in *livres tournois*, the unit of account, replacing the previously issued *écus de banque* representing fixed amounts of specie.

Alongside the bank, in August 1717, Law established the Compagnie d'Occident after obtaining the franchise on Louisiana and the monopoly of the Canadian fur trade. This company in the succeeding 22 months acquired the tobacco monopoly, the East India Company and the trading monopolies to Africa and China. Law changed its name in June 1719 to the Compagnie des Indes, and the following winter obtained the farm of the royal mints and of the indirect taxes. In October 1719 he refunded the national debt of 1.5 million *livres tournois*, and in January 1720 became Finance Minister.

The stock of the Compagnie des Indes, initially selling at a par value of £500, within half a year in an unprecedented speculative mania was bid up to many times its original price. The bubble burst in January 1720 after the price of the stock reached a peak of £18,000. To support the price Law made the mistake of pegging it at £9,000, thereby monetizing it and engendering a rapid expansion of notes (125 per cent in two months). In May 1720, in a desperate attempt to salvage his system Law issued a deflationary decree depreciating the stock and reducing the denomination of notes by stages. This decree led to a panic as the public, fearful of further capital losses, sold off both notes and stock. Law's dismissal by the Regent worsened the panic. He was quickly reinstated but his final attempt to restore confidence by reducing the outstanding note issue proved unsuccessful. By December 1720 the ‘system’ collapsed. Law fled to Belgium and payments quickly reverted to a specie basis. The collapse of the system ruined many in all walks of life and made the word ‘bank’ anathema in France for well over a century.

Though Law's system reduced unemployment and stimulated output, it was at the expense of doubling the price level. His system was undermined by his actions breaking the link between the note issue and specie convertibility; by retiring the national debt with bank notes convertible into stock; and by encouraging speculation in stock by declaring dividends unrelated to the company's true prospects. Monetizing the stock by pegging its price in the end destroyed the public's confidence in his system. Law was aware of many of the principles of sound money and banking, but by equating money with stock and relying on the real bills doctrine he sowed the seeds of disaster.

Selected works

1705. *Money and Trade Considered: With a Proposal for Supplying the Nation With Money*. New York: Augustus Kelley, 1966.

Bibliography

Blaug, M. 1978. *Economic Theory in Retrospect*. 3rd edn. Cambridge: Cambridge University Press.

Hamilton, E.J. 1936. Prices and wages at Paris under John Law's system. *Quarterly Journal of Economics* 51, 42–70.

Hamilton, E.J. 1968. Law, John. In *International Encyclopedia of the Social Sciences*, vol. 9. New York: The Free Press.

Hamilton, E.J. 1969. The political economy of France at the time of John Law. *History of Political Economy* 1, 123–49.

Harsin, P., ed. 1934. *John Law: Oeuvres complètes*. 3 vols. Paris: Sirey.

Kindleberger, C.P. 1984. *A Financial History of Western Europe*. London: Allen & Unwin.

Neal, L. and Schubert, E. 1985. The first rational bubble: a new look at the Mississippi and South Sea schemes. Faculty Working Paper No. 1188, Bureau of Economic and Business Research, University of Illinois at Urbana-Champaign.

Rist, C.W. 1940. *A History of Monetary and Credit Theory from John Law to the Present Day*. London: Allen & Unwin.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

How to cite this article

Bordo, Michael D. "Law, John (1671–1729)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000045> doi:10.1057/9780230226203.0942

The New Palgrave Dictionary of Economics Online

law, public enforcement of

A. Mitchell Polinsky and Steven Shavell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys the economic analysis of public enforcement of law – the use of public agents (inspectors, tax auditors, police, prosecutors) to detect and to sanction violators of legal rules. We first discuss the basic elements of the theory: the probability of imposition of sanctions, the magnitude and form of sanctions (fines, imprisonment), and the rule of liability. We then examine a variety of extensions, including the costs of imposing fines, mistakes, marginal deterrence, settlement, self-reporting, repeat offences, and incapacitation.

Keywords

audits; Becker, G.; Bentham, J.; bribery; corruption; criminal law; deterrence; efficiency wages; fairness; fault-based liability; fines; harm; imprisonment; incapacitation; principal and agent; public enforcement of law; safety regulations; sanctions; self-reporting; settlements; social norms; strict liability; time discount rates

Article

In this article we consider the theory of public enforcement of law – the use of public agents (inspectors, tax auditors, police, prosecutors) to detect and to sanction violators of legal rules. After briefly discussing the rationale for public (as opposed to private) enforcement, we present the basic elements of the theory: the probability of imposition of sanctions, the magnitude and form of sanctions (fines, imprisonment), and the rule of liability. We then examine a variety of extensions of the central theory, including the costs of imposing fines, mistakes, marginal deterrence, settlement, self-reporting, repeat offences, and incapacitation. (For a fuller treatment of the material in this entry, see Polinsky and Shavell, 2007.)

Before proceeding, we note that economically oriented analysis of public law enforcement dates primarily from the 18th century contribution of Jeremy Bentham (1789), whose analysis of deterrence

was sophisticated and expansive. After Bentham, the subject of enforcement lay essentially dormant in economic scholarship until Gary Becker (1968) published a highly influential article, which has led to a voluminous literature.

Rationale of public enforcement

A basic question is why there is a need for public enforcement of law (see generally Becker and Stigler, 1974; Landes and Posner, 1975; Polinsky, 1980a). In particular, why not rely solely on private suits brought by victims? The answer depends importantly on the locus of information about the identity of injurers. When victims of harm naturally possess knowledge of the identity of injurers, allowing private suits for damages will motivate victims to sue and thus harness the information they have for purposes of law enforcement. This may explain why the enforcement of contractual obligations and of accident law is primarily private. When victims do not know who caused harm, however, or when finding injurers is difficult, society may need to rely instead on public investigation and prosecution; this is broadly true of crimes and of many violations of environmental and safety regulations.

Basic framework for analysing public enforcement

An individual who commits a harmful act obtains a gain and also faces the risk of being caught and sanctioned. The form of sanction could be a fine or a prison term. Fines generally will be treated as socially costless because they are mere transfers of money, whereas imprisonment will be considered as socially costly because of the expense of operating prisons and the disutility suffered by those imprisoned. The higher the probability of detecting violators, the more resources the state must devote to enforcement.

We assume that social welfare equals the sum of individuals' expected utilities. If individuals are risk neutral, social welfare can be expressed as the gains individuals obtain from committing their harmful acts, minus the harms caused and the costs of law enforcement. The enforcement authority's problem is to maximize social welfare by choosing enforcement expenditures (or, equivalently, a probability of detection), the form of sanctions, and their level.

Fines

Suppose that the sanction is a fine and that individuals are risk neutral. If the probability of detection p is taken as fixed, then the optimal fine is the harm h divided by the probability, that is, h/p; for then the expected fine $p(h/p)$ equals h. This fine is optimal because, facing it, an individual will commit a harmful act if, and only if, the gain he would derive exceeds the harm he would cause. Such behaviour is first-best. The fundamental formula h/p essentially was noted by Bentham (1789) and it has been observed by many others since.

If the probability of detection can be varied, the optimal fine is maximal, f_M, as emphasized by Becker (1968). If the fine were not maximal, society could save enforcement costs by simultaneously raising the fine and lowering the probability without affecting the level of deterrence. If $f < f_M$, then raise the fine

to f_M and lower the probability from p to $(f/f_M)p$; the expected fine is still pf, so that deterrence is maintained but expenditures on enforcement are reduced, implying that social welfare rises.

The optimal probability p of imposing a fine is low in the sense that it results in some under-deterrence; that is, the optimal p is such that the expected fine pf_M is less than the harm h (Polinsky and Shavell, 1984). The reason is to economize on enforcement resources. In particular, if pf_M equals h, behaviour will be ideal, meaning that the individuals who are just deterred obtain gains essentially equal to the harm. These are the individuals who would be led to commit the harmful act if p were lowered slightly. That in turn must be socially beneficial because these individuals cause no net social losses (their gains essentially equal the harm), but reducing p saves enforcement costs. How much pf_M should be lowered below h depends on the saving in enforcement costs from reducing p compared with the net social costs of under-deterrence that will result if p is lowered non-trivially.

If individuals are risk averse, the optimal fine may be well less than the maximal fine, as first shown in Polinsky and Shavell (1979); see also Kaplow (1992). This is because a high fine would impose substantial risk-bearing costs on individuals who commit harmful acts. If $f < f_M$, it is still true that f can be raised and p lowered so as to maintain deterrence, but because of risk aversion this now implies that pf falls, meaning that fine revenue falls. The reduction in fine revenue reflects the disutility caused by imposing greater risk on risk-averse individuals. The decline in fine revenue could more than offset the savings in enforcement expenditures, causing social welfare to be lower.

Imprisonment

Now suppose that the sanction is imprisonment. If the probability of detection is fixed, there is no simple formula for the optimal imprisonment term (see Polinsky and Shavell, 1984). The optimal term could be such that there is either under-deterrence or over-deterrence. On the one hand, a relatively low imprisonment term, implying under-deterrence, might be socially desirable because imprisonment costs are reduced for those individuals who commit harmful acts. On the other hand, a relatively high term, implying over-deterrence, might be socially desirable because imprisonment costs are reduced due to fewer individuals committing harmful acts, even if some of these deterred individuals would have obtained gains exceeding the harm.

If the probability of detection can be varied and individuals are risk neutral in imprisonment, then the optimal imprisonment term is maximal. The reasoning is similar to that employed above: if the imprisonment term were not maximal, it could be raised and the probability of detection lowered so as to keep the expected prison term constant; neither individual behaviour nor the costs of imprisonment are affected, but enforcement expenditures fall.

If, instead, individuals are risk averse in imprisonment (the disutility of each additional year of imprisonment grows with the number of years in prison), there is a stronger argument for setting the imprisonment sanction maximally (Polinsky and Shavell, 1999). Now when the imprisonment term is raised, the probability of detection can be lowered more than in the risk-neutral case without reducing deterrence. Thus, not only are there greater savings in enforcement expenditures, but also the costs of imposing imprisonment sanctions decline because the expected prison term falls.

Last, suppose that individuals are risk preferring in imprisonment (the disutility of each additional year of imprisonment declines with the number of years in prison). This possibility seems particularly

important: the first years of imprisonment may create unusually high disutility, due to brutalization of the prisoner or to the stigma of having been imprisoned at all. Individuals' positive time discount rates, which are thought to be especially significant for criminals, also make the disutility of later years less significant. In the case of risk-preferring individuals, the optimal prison term may well be less than maximal: if the sentence were raised, the probability that maintains deterrence could not be lowered proportionally, implying that the expected prison term would rise. Thus, although there would be enforcement-cost savings, they might not be great enough to offset the increased sanctioning costs.
When the sanction is imprisonment, the optimal probability of detection may be such that there is either under-deterrence or over-deterrence. On the one hand, the motive to lower the probability is reinforced relative to the case of fines because imprisonment costs, as well as detection costs, decline if fewer offenders are caught. On the other hand, raising the probability of detection results in fewer offenders, which, everything else equal, decreases imprisonment costs because fewer are imprisoned. Either effect may dominate.

Fines versus imprisonment

Fines generally are preferable to prison terms as a means of deterrence, since fines are socially cheaper sanctions to impose (Becker, 1968; Polinsky and Shavell, 1984). Hence, fines should be employed to the greatest extent possible – until a party's wealth is exhausted – before imprisonment is imposed. Further, imprisonment should be used as a sanction only if the harm prevented by the added deterrence is sufficiently great.

Fault-based liability

Our discussion thus far has presumed that liability is strict (imposed whenever harm occurs), but liability may instead be based on fault (imposed only when behaviour was found to be socially undesirable). Fault-based liability, like strict liability, can induce individuals to behave properly, but fault-based liability possesses an advantage when individuals are risk averse: if they act responsibly, they will not be found at fault, so will not bear the risk of being sanctioned. Similarly, fault-based liability is advantageous when the sanction is imprisonment, for then again individuals may be led to behave optimally without the actual imposition of sanctions, and thus without social costs being incurred (Shavell, 1987b). To the extent that mistakes are made in determining fault, however, these two advantages are reduced.
Fault-based liability is more difficult to implement because it requires more information than strict liability. To apply fault-based liability, the enforcement authority must be able to determine the proper fault standard – that is, socially desirable behaviour – and it must ascertain whether the defendant's conduct was in compliance with the fault standard. Under strict liability, the authority need only measure harm. (Moreover, for reasons we discuss below, strict liability encourages better decisions by injurers regarding their level of participation in harm-creating activities.)
This concludes the presentation of the basic theory of public enforcement of law. We now turn to various extensions and refinements of the analysis.

Accidental harms

We have been implicitly assuming that individuals decide whether or not to commit acts that cause harm with certainty, that is, they decide whether or not to cause intentional harms. In many circumstances, however, harms are accidental – they occur only with a probability. Essentially all that we have said above applies in a straightforward way when harms are accidental.

There is, however, an additional issue that arises when harm is uncertain: a sanction can be imposed either on the basis of the commission of an act that increases the chance of harm (such as storing chemicals in a substandard tank) or on the basis of the actual occurrence of harm (if the tank ruptures and results in a spill). In principle, either approach can achieve optimal deterrence – by setting the (expected) sanction equal to expected harm if liability is imposed whenever a dangerous act is committed, or equal to actual harm if liability is imposed only if harm occurs.

Several factors are relevant to the choice between act-based and harm-based sanctions (Shavell, 1993). First, act-based sanctions need not be as high as harm-based sanctions to accomplish a given level of deterrence (expected harm is less than actual harm), and thus offer an advantage because of parties' limited assets. Second, because act-based sanctions can accomplish a given level of deterrence with lower sanctions, they are preferable when parties are risk averse. Third, either act-based sanctions may be simpler to impose (it might be less difficult to determine whether an oil shipper properly maintains its vessels' holding tanks than to detect whether one of the vessels leaked oil), or harm-based sanctions may be easier to implement (a driver who causes harm might be caught without difficulty, but not one who speeds). Fourth, it may be hard to calculate the expected harm due to an act, but relatively easy to ascertain the actual harm if it eventuates, favoring harm-based sanctions.

Costs of imposing fines

The costs borne by enforcement authorities in imposing fines should be reflected in the fine. Recall that, if the probability of detection is taken as fixed and individuals are risk neutral, the optimal fine is h/p, the harm divided by the probability of detection. Now suppose there is a public cost k of imposing a fine. The optimal fine then becomes $h/p + k$; the cost k should be added to the fine that would otherwise be desirable (Becker, 1968; Polinsky and Shavell, 1992). The explanation is that, if an individual commits a harmful act, he causes society to bear not only the immediate harm h but also, with probability p, the cost k of imposing the fine – that is, his act results in an expected total social cost of $h + pk$. If the fine is $h/p + k$, the individual's expected fine is $p[h/p + k] = h + pk$, leading him to commit the harmful act if and only if his gain exceeds the expected total social cost of his act.

Not only does the state bear costs when fines are imposed, so do individuals who pay the fines (such as legal defence expenses). The costs borne by individuals, however, do not affect the formula for the optimal fine. Individuals properly take these costs into account because they bear them.

Level of activity

In many settings in which harm may occur, an individual chooses not only whether to commit a harmful act when engaging in an activity, but also the level at which to engage in the activity. Drivers decide

how careful to be while driving, as well as how many miles to drive; similarly, firms choose safety precautions as well as their level of output. The socially optimal activity level is such that the actor's marginal utility from the activity just equals the marginal expected harm caused by the activity (we assume that optimal care is taken). Thus, the optimal number of miles driven is the level at which the marginal utility of driving an extra mile just equals the marginal expected harm per mile driven.

Under strict liability parties will choose the optimal level of activity because they will pay for all harm done. They will choose the optimal number of miles to drive because they will pay for all harm per mile driven. Under fault-based liability, however, parties generally do not pay for the harm they cause because they tend to behave so as not to be found at fault. As a consequence, they will choose an excessive level of activity (Shavell, 1980). Driving more miles increases expected harm, but this effect generally will be ignored under fault-based liability.

The interpretation of the preceding points in relation to firms is that under strict liability the product price will reflect the expected harm caused by production. Hence, the amount purchased, and thus the level of production, will tend to be socially optimal. However, under fault-based liability the product price will not reflect harm, but only the cost of precautions; thus, the level of output will be excessive (Polinsky, 1980b).

Relatedly, safety regulations and other regulatory requirements are often framed as standards of care that have to be met, but which, if met, free the regulated party from liability. Hence, regulations of this sort are subject to the criticism that they lead to excessive levels of the regulated activity. Making parties strictly liable for harm would be superior to safety regulation with respect to inducing socially correct activity levels.

Mistakes

An individual who should be found liable might mistakenly be acquitted. Conversely, an individual who should not be found liable might mistakenly be convicted. For an individual who has been detected, let the probabilities of these errors be ε_A and ε_C, respectively. Given the probability of detection *p* and the chances of these types of error, an individual will commit the wrongful act if and only if his gain *g* net of his expected fine if he does commit it exceeds his expected fine if he does not commit it, that is, when $g - p(1-\varepsilon_A)f > -p\varepsilon_C f$, or, equivalently, when $g > (1-\varepsilon_A-\varepsilon_C)pf$.

As emphasized by Png (1986), both types of error reduce deterrence: the term $(1-\varepsilon_A-\varepsilon_C)pf$ is declining in both ε_A and ε_C. The first type of error diminishes deterrence because it lowers the expected fine if an individual violates the law. The second type of error lowers deterrence because it reduces the difference between the expected fine from violating the law and not violating it — the greater is ε_C, the smaller is the increase in the expected fine if one violates the law.

Because mistakes dilute deterrence, they reduce social welfare. Specifically, to achieve any level of deterrence, the probability *p* must be higher to offset the effect of errors. Mistaken convictions have the additional effect of discouraging socially desirable participation in the activity. Consequently, expenditures made to reduce errors may be socially beneficial (Kaplow and Shavell, 1994a).

Two other points regarding the implications of mistake are worth noting. First, if individuals are risk averse, the possibility of mistakes of either type generally lowers optimal sanctions (Block and Sidak,

1980). Second, as stressed by Craswell and Calfee (1986), individuals will often have a motive to take excessive precautions under fault-based liability in order to reduce the chance of being found erroneously at fault.

General enforcement

In many settings, enforcement may be said to be general in the sense that several different types of violations will be detected by an enforcement agent's activity. For example, a police officer waiting at the roadside may notice a driver who litters as well as one who goes through a red light or who speeds, and a tax auditor may detect a variety of infractions when he examines a tax return. (In contrast, if enforcement is specific, the probability is chosen independently for each type of harmful act.)
When enforcement is general, the optimal sanction rises with the level of harm, and is maximal only for relatively high harms (Shavell, 1991; Mookherjee and Png, 1992). To see why, assume that liability is strict, the sanction is a fine, and injurers are risk neutral. Let $f(h)$ be the fine given harm h. Then, for any general probability of detection p (that is, p applies regardless of h), the optimal fine schedule is h/p, provided that h/p is feasible; otherwise the optimal fine is maximal. This schedule is obviously optimal given p because it implies that the expected fine equals harm, thereby inducing ideal behaviour whenever that is possible. That sanctions should rise with the severity of harm up to a maximum when enforcement is general also holds if the sanction is imprisonment and if liability is fault-based.

Marginal deterrence

In many circumstances a person may consider which of several harmful acts to commit: for example, whether to release only a small amount of a pollutant into a river or a large amount, or whether to kidnap a person or also to kill the kidnap victim. In such contexts, sanctions influence which harmful acts individuals choose to commit (as well as whether to commit any harmful act). Marginal deterrence is said to occur when a more harmful act is deterred because its sanction exceeds that for a less harmful act (Stigler, 1970; Shavell, 1992; Wilde, 1992; Mookherjee and Png, 1994).
Other things being equal, it is socially desirable that enforcement policy creates marginal deterrence so that, when harmful acts do occur, less harm is done. One way to accomplish marginal deterrence is for sanctions to rise with the magnitude of harm, which means that sanctions generally will not be maximal. However, fostering marginal deterrence may conflict with achieving overall deterrence: in order for the schedule of sanctions to rise steeply enough to accomplish marginal deterrence, sanctions for less harmful acts may have to be so low that individuals are not deterred from committing some harmful act. Note that marginal deterrence also can be promoted by increasing the probability of detection. Kidnappers can be better deterred from killing their victims if more police resources are devoted to apprehending kidnappers who murder their victims than to those who do not.

Principal – agent relationship

Although we have assumed that an injurer is a single actor, injurers often are more appropriately characterized as collective entities, and specifically as a principal and the principal's agent. For example,

the principal could be a firm and the agent an employee, or the principal could be a contractor and the agent a subcontractor.

When harm is caused by the behaviour of principals and agents, many of our prior conclusions carry over to the sanctioning of principals. Notably, if a risk-neutral principal faces an expected fine equal to harm done, he will behave socially optimally in controlling his agents, and in particular will contract with them and monitor them in ways that will give the agents appropriate incentives to reduce harm (Newman and Wright, 1990; but see Arlen, 1994).

An issue that arises when there are principals and agents concerns the allocation of financial sanctions between the two parties. It is apparent that the particular allocation of sanctions does not matter when the parties can reallocate the sanctions through their own contract. For example, if the agent finds that he faces a large fine but is more risk averse than the principal, the principal can assume it; conversely, if the fine is imposed on the principal, he will retain it and not impose an internal sanction on the agent. Thus, the post-contract sanctions that the agent bears are not affected by the particular division of sanctions initially selected by the enforcement authority.

The allocation of monetary sanctions between principals and agents would matter, however, if some allocations allow the pair to reduce their total burden. An important example is when a fine is imposed only on the agent and he is unable to pay it (Sykes, 1981; Kornhauser, 1982). Then, he and the principal (who often would have higher assets) would jointly escape part of the fine, diluting deterrence. The fine therefore should be imposed on the principal rather than on the agent (or at least the part of the fine that the agent cannot pay).

A closely related point is that the imposition of imprisonment sanctions on agents may be desirable when their assets are less than the harm that they can cause, even if the principal's assets are sufficient to pay the optimal fine (Polinsky and Shavell, 1993). That an agent's assets are limited means that the principal may be unable to control him adequately through the use of contractually determined penalties, which can only be monetary. In such circumstances it may be socially valuable to use the threat of a jail sentence to better control agents’ misconduct.

Settlements

It is common for lawbreakers to settle with public enforcement authorities prior to being found liable in a trial. (In the criminal context, the settlement usually takes the form of a plea bargain, an agreement in which the injurer pleads guilty to a reduced charge.) Both parties might prefer an out-of-court settlement to avoid the cost of a trial and to eliminate the risks inherent in the trial outcome (Cooter and Rubinfeld, 1989; on plea bargaining, see Reinganum, 1988, and Miceli, 1996).

These advantages suggest that settlement is socially valuable, but the effect of settlement on deterrence is a complicating factor. Specifically, settlements dilute deterrence: for if injurers desire to settle, it must be because the expected disutility of sanctions is lowered for them (Polinsky and Rubinfeld, 1988). The state may be able to offset this effect by increasing the level of sanctions.

Settlements may have other socially undesirable consequences. First, they may result in sanctions that are not as well tailored to harmful acts as would be true of court-determined sanctions. For example, if injurers have private information about the harm that they have caused, settlements will tend to reflect the average harm caused, resulting in high-harm (low-harm) injurers being under-deterred (over-

deterred), whereas trial outcomes may better approximate the actual harm. Second, settlements hinder the amplification and development of the law through the setting of precedents. Third, if the sanction is imprisonment and defendants are risk averse, settlements necessitate longer terms than the expected sentence at trial in order to maintain deterrence, and thus increase public expenditures. On the social welfare evaluation of settlement, see, for example, Shavell (1997) and Spier (1997).

Self-reporting

We have assumed that individuals are subject to sanctions only if they are detected by an enforcement agent, but in fact parties sometimes disclose their own violations. For example, firms often report infractions of environmental and safety regulations, individuals usually notify police of their involvement in traffic accidents, and even criminals occasionally turn themselves in.

Self-reporting can be induced by lowering the sanction for individuals who disclose their own violations (Kaplow and Shavell, 1994b). Moreover, the reward for self-reporting can be made small, so that deterrence is only negligibly reduced. For example, if a risk-neutral individual commits a violation and does not self-report, his expected fine is *pf*. If he self-reports, the fine can be set just below *pf*, say at $pf - \varepsilon$, where $\varepsilon > 0$ is small. Then the individual will want to self-report but the deterrent effect of the sanction will be essentially the same as if he did not self-report.

There are several social advantages of self-reporting. First, self-reporting reduces enforcement costs because the enforcement authority does not have to identify and prove who the violator was. Second, self-reporting reduces risk (a relatively high sanction imposed with a relatively low probability is replaced by a certain punishment), and thus is advantageous if injurers are risk averse. Third, self-reporting may allow harm to be mitigated (early notice of an oil spill may facilitate its containment).

Repeat offenders

In practice, the law often sanctions repeat offenders more severely than first-time offenders. This policy cannot be socially advantageous if deterrence always induces first-best behaviour. For if the expected sanction for an offence equals its harm, then raising the sanction because an offender has a record of sanctions would over-deter him. Only if deterrence is inadequate is it possibly desirable to condition sanctions on offence history to increase deterrence. But, as we observed above, it usually will be worthwhile for the state to tolerate some under-deterrence in order to reduce enforcement expenses.

If there is under-deterrence, making sanctions depend on offence history may be beneficial. First, the use of offence history may create an additional incentive not to violate the law: if detection results not only in an immediate sanction but also in a higher sanction for any future violation, an individual will, everything else equal, be deterred to a greater extent (Polinsky and Shavell, 1998). Second, making sanctions depend on offence history allows society to take advantage of information about the dangerousness of individuals and the need to deter them: individuals with offence histories may be more likely than average to commit future violations, which might make it desirable to impose higher sanctions on them (Rubinstein, 1979; Polinsky and Rubinfeld, 1991). In addition, if repeat offenders have higher propensities to commit violations, they are more likely to be worth incapacitating by imprisonment (see below).

Imperfect knowledge about the probability and magnitude of sanctions

Individuals might not know the true probability of a sanction because the enforcement authority refrains from publishing information about the probability (perhaps hoping that individuals will believe it to be higher than it is in fact); or because the probability depends on factors that individuals do not fully understand; or because probabilities are difficult to assess. Also, individuals may have incomplete knowledge of the true magnitude of sanctions, particularly if the levels of sanctions are discretionary.
The implications of injurers' imperfect knowledge are straightforward. First, to predict how individuals behave, what is relevant, of course, is not the actual probability and magnitude of a sanction but the perceived levels or distributions of these variables. Second, to determine the optimal probability and magnitude of a sanction, account must be taken of the relationship between the actual and the perceived variables (Bebchuk and Kaplow, 1992; Kaplow, 1990). For example, if enforcement resources are increased in order to raise the probability of detection, there might be a delay before this increase is perceived by individuals, making such an investment less worthwhile.

Incapacitation

Society may reduce harm not only through deterrence, but also by imposing sanctions that remove parties from positions in which they are able to cause harm, that is, by incapacitating them.
Imprisonment is the primary incapacitative sanction, although there are other examples: individuals can lose their driver's licences, businesses can lose their rights to operate in certain markets, and the like.
Suppose that the sole function of imprisonment is to incapacitate. Then it will be desirable to keep someone imprisoned as long as the reduction in criminal harm from incapacitating him exceeds the cost of imprisonment (Shavell, 1987c). Although this condition could hold for a long period, it often will not because the proclivity to commit crimes appears to decline sharply with age.
As a matter of economic logic, the incapacitation rationale might imply that a person should be imprisoned even if he has not committed a crime, because the danger he poses to society makes incapacitating him worthwhile. In practice, however, the commission of a harmful act may be a good basis for predicting a person's future behaviour, in which case the incapacitation rationale would suggest imprisoning an individual only if he has committed such an act.
Two observations are worth noting about the relationship between the incapacitation goal and the deterrence goal. First, when enforcement is based on incapacitation, the optimal magnitude of the sanction is independent of the probability of apprehension, which contrasts with the case when enforcement is based on deterrence. Second, when enforcement is deterrence-oriented, the probability and magnitude of sanctions depend on the ability to deter, and if this ability is limited (as, for instance, with the insane), a low expected sanction may be optimal, whereas a high sanction still might be called for to incapacitate.

Corruption

One form of corruption in the enforcement process is bribery, in which an enforcer accepts a payment in

return for not reporting a violation (or for reducing the mandated sanction for the violation). A second form of corruption is framing and framing-related extortion, in which an enforcement agent may frame an innocent individual or threaten to frame him in order to extort money from him. On corruption of law enforcement, see Bowles and Garoupa (1997) and Polinsky and Shavell (2001) (and on corruption more generally, see, for example, Shleifer and Vishny, 1993, and Rose-Ackerman, 1999).

Bribery dilutes deterrence of violations of law because it results in a lower payment by an individual than the sanction for the offence. Framing and framing-related extortion also dilute deterrence. The reason is that framing and extortion imply that those who act innocently face an expected sanction, so that the difference between the expected sanction if an individual commits a violation and if he does not is lessened. (This point is essentially the same as the earlier observation that mistaken convictions dilute deterrence.)

One way to reduce corruption is to impose fines (or imprisonment sentences) on individuals caught engaging in bribery, extortion or framing. Corruption also can be reduced by paying enforcers rewards for reporting violations. Such payments will reduce their incentive to accept bribes because they will sacrifice their rewards if they fail to report violations. But high rewards give enforcers a greater incentive to frame innocent individuals. A third way to control corruption is to pay enforcers more than their reservation wage (that is, to pay them an efficiency wage). Then they would have more to lose if punished for corrupt behaviour and denied future employment.

A natural question is whether the deterrence-diluting effects of corruption can be offset by raising the fine on offenders. In the basic risk-neutral model of enforcement, it is not possible to raise the fine because the optimal fine is maximal. More realistically, however, the optimal fine is less than maximal for a variety of reasons, including those related to risk aversion, marginal deterrence, and general enforcement. While it would then be possible to raise the fine to offset the deterrence-diluting effects of corruption, doing so would lead to social costs (for example, by imposing greater risk).

Costly observation of wealth

Individuals and firms may be able to hide assets from government enforcers, including by hoarding cash, transferring assets to relatives or related legal entities, or moving money to offshore bank accounts. Consequently, an individual's level of wealth might not be able to be observed at all, or only after a costly audit.

Suppose first that the enforcement authority employs fines as sanctions and can audit an individual who claims that he cannot pay the fine (Polinsky, 2006). The optimal fine for misrepresenting one's wealth level equals the fine for the offence divided by the audit probability, and therefore generally exceeds the fine for the offence. This is a natural generalization of the formula for the optimal fine when the probability of detection is fixed, which is the harm divided by the probability. Auditing is valuable because it reduces misrepresentation of wealth and thereby increases deterrence.

Next, suppose that the enforcement authority cannot observe wealth because the cost of an audit is prohibitively high (Levitt, 1997; Polinsky, 2006). If the authority would have used fines alone if it could have observed wealth at no cost, it would have imposed a higher fine on higher-wealth individuals. It obviously cannot do this when wealth is unobservable. Instead, it may be desirable to use the threat of an imprisonment sentence to induce individuals capable of paying a higher fine to do so. Alternatively, the

enforcement authority might have used both fines and imprisonment if it could have observed wealth at no cost. Perhaps surprisingly, the inability to observe wealth might not be detrimental in this case. The reason is that the mix of fines and imprisonment that would be chosen when wealth is observable might impose a higher burden (though a lower fine) on low-wealth individuals. Then, high-wealth individuals will naturally want to identify themselves. Specifically, they will prefer to pay a higher fine and bear a shorter imprisonment sentence than to masquerade as low-wealth individuals, who will bear longer imprisonment sentences and a higher overall burden.

Social norms

To some extent, social norms and morality are substitutes for public law enforcement because they encourage in significant ways the attainment of desired behaviour (McAdams and Rasmusen, 2007; Posner, 1997; Shavell, 2002). Social norms influence behaviour partly through internal incentives: when a person obeys a moral rule, he will tend to feel virtuous, and if he disobeys the rule, he will tend to feel guilty. Social norms also affect behaviour through external incentives: when a person is observed by another party to have obeyed a moral rule, that party may bestow praise on the first party, who will enjoy the praise; and if the person is observed by the other party to have disobeyed the rule, the second party will tend to disapprove of the first party, who will dislike the disapproval. Because social norms channel behaviour in this way, some socially desirable conduct can be encouraged reasonably well without employing the legal system.

Notwithstanding these observations, there will, of course, often be a need for formal law enforcement. First, much conduct that society desires cannot be controlled through moral incentives alone. One reason is that the private gains from undesirable conduct are often large and dominate the moral incentives. Another reason is that external moral sanctions might be imposed only with a low probability (the robber, tax cheat or polluter might not be spotted by others). A second rationale of formal law enforcement is that the social harm from failing to control an act through moral incentives may be large. This makes the expense of law enforcement worth incurring (as in the case of controlling robbery, but not of breaking into a queue at a movie theatre).

Fairness

So far we have not considered the possibility that individuals have opinions about the fairness of sanctions or the arbitrariness of enforcement (Polinsky and Shavell, 2000b; Kaplow and Shavell, 2002). Suppose, first, that individuals believe that the magnitude of sanctions should reflect the gravity of the acts. As discussed previously, if individuals are risk neutral, the usual solution to the enforcement problem consists of the highest possible sanction and a relatively low probability of detection. When the issue of fairness is added to the analysis, however, the usual solution generally is not optimal because a very high sanction will be seen as unfair.

A consequence of the desire to keep sanctions at fair levels, meaning at quite constrained levels for acts that are not very harmful, is that the socially optimal probability of detection changes. The optimal probability could be higher than the conventionally optimal probability: to achieve a desired level of deterrence with a lower fairness-restricted sanction, the probability has to rise, perhaps significantly.

Alternatively, the optimal probability could be lower than in the conventional case: the additional deterrence from raising the probability might be relatively low because the sanction is relatively low; and the lower the deterrent benefit from raising the probability, the lower would be the social incentive to devote resources to enforcement.

Another aspect of fairness concerns the probability of detection rather than the magnitude of sanctions. Suppose that individuals consider it unfair for some lawbreakers to be sanctioned when others, who were lucky enough not to be caught, are not sanctioned. Then the optimal probability would be higher, and therefore the optimal sanction would be lower, than in the absence of this fairness concern.

A further notion of fairness involves the form of liability, whether liability is strict or based on fault. Individuals might prefer fault-based liability because sanctions are imposed on parties only if they behaved in a socially inappropriate way.

A final issue concerns the relevance of fairness considerations when firms, as opposed to individuals, are sanctioned. If what matters in terms of fairness is that the individuals responsible for harmful acts bear sanctions, as opposed to the artificial legal entity of a firm, one would want to identify the sanctions actually suffered by such persons within a firm if the firm bears a sanction. Note, too, that the imposition of sanctions on firms often penalizes individuals who are unlikely to be considered responsible for the harm, namely, shareholders and customers.

Criminal law

The subject of criminal law may be viewed in the light of the theory of public law enforcement (Posner, 1985; Shavell, 1985). First, the fact that the acts in the core area of crime (robbery, murder, rape, and so forth) are punished by the sanction of imprisonment makes basic sense. Were society to rely on fines alone, deterrence of the acts in question would be grossly inadequate. This is because the probability of detecting many of these acts is low, making the money sanction necessary for deterrence high, but the assets of individuals who commit these acts often are insubstantial. Hence, the threat of prison is needed for deterrence. Moreover, the incapacitative aspect of imprisonment is valuable because of the difficulty of deterring individuals who are prone to commit criminal acts.

Second, many of the doctrines of criminal law appear to enhance social welfare. This seems true of the basic feature of criminal law that punishment is not imposed on all harmful acts, but instead is usually confined to those that are especially undesirable. (For example, murder is subject to criminal sanctions, but some accidental killing is not.) As we have stressed, when the socially costly sanction of imprisonment is employed, the fault system is desirable because it results in less frequent imposition of punishment than strict liability. Also, the focus on intent in criminal law as a precondition for imposing sanctions may serve to foster deterrence because those who intend to do harm are more likely to conceal their acts, and may be harder to discourage because of the benefits they anticipate. An additional example of a welfare-enhancing doctrine in criminal law concerns attempts. That attempts to do harm are punished is an implicit way of raising the likelihood of sanctions for undesirable acts.

See Also

- deterrence (theory), economics of

- externalities
- law, economic analysis of
- pecuniary versus non-pecuniary penalties

A. Mitchell Polinsky's research was supported by the John M. Olin Program in Law and Economics at Stanford Law School. Steven Shavell's research was supported by the John M. Olin Center for Law, Economics, and Business at Harvard Law School.

Bibliography

Arlen, J. 1994. The potentially perverse effects of corporate criminal liability. *Journal of Legal Studies* 23, 833–67.

Bebchuk, L. and Kaplow, L. 1992. Optimal sanctions when individuals are imperfectly informed about the probability of apprehension. *Journal of Legal Studies* 21, 365–70.

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Becker, G. and Stigler, G. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3, 1–18.

Bentham, J. 1789. An Introduction to the Principles of Morals and Legislation. In The Utilitarians, Garden City, NY: Anchor Books, 1973.

Block, M. and Sidak, J. 1980. The cost of antitrust deterrence: why not hang a price fixer now and then? *Georgetown Law Journal* 68, 1131–9.

Bowles, R. and Garoupa, N. 1997. Casual police corruption and the economics of crime. *International Review of Law and Economics* 17, 75–87.

Cooter, R. and Rubinfeld, D. 1989. Economic analysis of legal disputes and their resolution. *Journal of Economic Literature* 27, 1067–97.

Craswell, R. and Calfee, J.E. 1986. Deterrence and uncertain legal standards. *Journal of Law, Economics, & Organization* 2, 279–303.

Kaplow, L. 1990. Optimal deterrence, uninformed individuals, and acquiring information about whether acts are subject to sanctions. *Journal of Law, Economics, & Organization* 6, 93–128.

Kaplow, L. 1992. The optimal probability and magnitude of fines for acts that definitely are undesirable.

International Review of Law and Economics 12, 3–11.

Kaplow, L. and Shavell, S. 1994a. Accuracy in the determination of liability. *Journal of Law and Economics* 37, 1–15.

Kaplow, L. and Shavell, S. 1994b. Optimal law enforcement with self–reporting of behavior. *Journal of Political Economy* 102, 583–606.

Kaplow, L. and Shavell, S. 2002. *Fairness versus Welfare*. Cambridge, MA: Harvard University Press.

Kornhauser, L. 1982. An economic analysis of the choice between enterprise and personal liability for accidents. *California Law Review* 70, 1345–92.

Landes, W. and Posner, R. 1975. The private enforcement of law. *Journal of Legal Studies* 4, 1–46.

Landes, W. and Posner, R. 1987. *The Economic Structure of Tort Law*. Cambridge, MA: Harvard University Press.

Levitt, S. 1997. Incentive compatibility constraints as an explanation for the use of prison sentences instead of fines. *International Review of Law and Economics* 17, 179–92.

McAdams, R. and Rasmusen, E. 2007. Norms in law and economics. In *Handbook of Law and Economics*, vol. 2, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland.

Miceli, T. 1996. Plea bargaining and deterrence: an institutional approach. *European Journal of Law and Economics* 3, 249–64.

Mookherjee, D. and Png, I. 1992. Monitoring vis-à-vis investigation in enforcement of law. *American Economic Review* 82, 556–65.

Mookherjee, D. and Png, I. 1994. Marginal deterrence in enforcement of law. *Journal of Political Economy* 102, 1039–66.

Newman, H. and Wright, D. 1990. Strict liability in a principal–agent model. *International Review of Law and Economics* 10, 219–31.

Png, I. 1986. Optimal subsidies and damages in the presence of judicial error. *International Review of Law and Economics* 6, 101–5.

Polinsky, A. 1980a. Private versus public enforcement of fines. *Journal of Legal Studies* 9, 105–27.

Polinsky, A. 1980b. Strict liability vs. negligence in a market setting. *American Economic Review* 70, 363–70.

Polinsky, A. 2006. The optimal use of fines and imprisonment when wealth is unobservable. *Journal of Public Economics* 90, 823–35.

Polinsky, A. 2006. Optimal fines and auditing when wealth is costly to observe. *International Review of Law and Economics* 26, 232–35.

Polinsky, A. and Rubinfeld, D. 1988. The deterrent effects of settlements and trials. *International Review of Law and Economics* 8, 109–16.

Polinsky, A. and Rubinfeld, D. 1991. A model of optimal fines for repeat offenders. *Journal of Public Economics* 46, 291–306.

Polinsky, A. and Shavell, S. 1979. The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* 69, 880–91.

Polinsky, A. and Shavell, S. 1984. The optimal use of fines and imprisonment. *Journal of Public Economics* 24, 89–99.

Polinsky, A. and Shavell, S. 1992. Enforcement costs and the optimal magnitude and probability of fines. *Journal of Law and Economics* 35, 133–48.

Polinsky, A. and Shavell, S. 1993. Should employees be subject to fines and imprisonment given the existence of corporate liability? *International Review of Law and Economics* 13, 239–57.

Polinsky, A. and Shavell, S. 1998. On offense history and the theory of deterrence. *International Review of Law and Economics* 18, 305–24.

Polinsky, A. and Shavell, S. 1999. On the disutility and discounting of imprisonment and the theory of deterrence. *Journal of Legal Studies* 28, 1–16.

Polinsky, A. and Shavell, S. 2000a. The economic theory of public enforcement of law. *Journal of Economic Literature* 38, 45–76.

Polinsky, A. and Shavell, S. 2000b. The fairness of sanctions: some implications for optimal enforcement policy. *American Law and Economics Review* 2, 223–37.

Polinsky, A. and Shavell, S. 2001. Corruption and optimal law enforcement. *Journal of Public Economics* 81, 1–24.

Polinsky, A. and Shavell, S. 2007. The theory of public enforcement of law. In *Handbook of Law and Economics*, vol. 1, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland.

Posner, R. 1985. An economic theory of the criminal law. *Columbia Law Review* 85, 1193–231.

Posner, R. 1997. Social norms and the law: an economic approach. *American Economic Review: Papers and Proceedings* 87, 365–9.

Reinganum, J. 1988. Plea bargaining and prosecutorial discretion. *American Economic Review* 78, 713–28.

Rose-Ackerman, S. 1999. *Corruption and Government: Causes, Consequences and Reform*. New York: Cambridge University Press.

Rubinstein, A. 1979. An optimal conviction policy for offenses that may have been committed by accident. In *Applied Game Theory*, ed. S. Brams, A. Schotter and G. Schwodiauer. Wurzburg: Physica-Verlag.

Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 9, 1–25.

Shavell, S. 1982. On liability and insurance. *Bell Journal of Economics* 13, 120–32.

Shavell, S. 1985. Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* 85, 1232–62.

Shavell, S. 1987a. The optimal use of nonmonetary sanctions as a deterrent. *American Economic Review* 77, 584–92.

Shavell, S. 1987b. A model of optimal incapacitation. *American Economic Review: Papers and Proceedings* 77, 107–10.

Shavell, S. 1987c. *Economic Analysis of Accident Law*. Cambridge, MA: Harvard University Press.

Shavell, S. 1991. Specific versus general enforcement of law. *Journal of Political Economy* 99, 1088–108.

Shavell, S. 1992. A note on marginal deterrence. *International Review of Law and Economics* 12, 345–55.

Shavell, S. 1993. The optimal structure of law enforcement. *Journal of Law and Economics* 36, 255–87.

Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26, 575–612.

Shavell, S. 2002. Law versus morality as regulators of conduct. *American Law and Economics Review* 4, 227–57.

Shleifer, A. and Vishny, R. 1993. Corruption. *Quarterly Journal of Economics* 108, 599–617.

Spier, K. 1997. A note on the divergence between the private and the social motive to settle under a negligence rule. *Journal of Legal Studies* 26, 613–21.

Stigler, G. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78, 526–36.

Sykes, A. 1981. An efficiency analysis of vicarious liability under the law of agency. *Yale Law Journal* 91, 168–206.

Wilde, L. 1992. Criminal choice, nonmonetary sanctions, and marginal deterrence: a normative analysis. *International Review of Law and Economics* 12, 333–44.

How to cite this article

Polinsky, A. Mitchell and Steven Shavell. "law, public enforcement of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000316> doi:10.1057/9780230226203.0944

The New Palgrave Dictionary of Economics Online

layoffs

John Haltiwanger

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Layoffs reflect employer-initiated job separations that play an important role in frictional and cyclical unemployment. The relative importance of temporary and permanent layoffs and layoffs themselves has varied over time, and understanding the factors underlying this variation is important for understanding fluctuations in frictional and cyclical unemployment over time. Modern models of labour market dynamics often emphasize the layoffs associated with endogenous job destruction at the firm level induced by the interaction of aggregate and firm-specific shocks.

Keywords

business cycles; cyclical unemployment; hold-up problem; implicit contracts; information capital; labour market search; layoffs; search and matching models; job creation and destruction

Article

The term ‘layoff’ is controversial in itself. For some the term connotes a temporary employer-initiated discharge, for others it represents any employer-initiated discharge that is without prejudice to the worker. The data on layoffs collected by the Bureau of Labor Statistics (BLS) in the United States (see, for example, various issues of the journal *Employment and Earnings*) takes the alternative types of layoffs into account across its firm and household surveys. Layoff data from the BLS survey of firms (the Job Openings and Labor Turnover Survey, JOLTS) provide data on employer-initiated discharges making no distinction as to whether the layoff is temporary or permanent. According to JOLTS, layoffs average about 1.1 per cent of US non-farm employment each month, which is about one-third of all worker separations. The BLS survey of households (the Current Population Survey, CPS) distinguishes between ‘temporary layoffs’ and ‘permanent job losers’ in tracking unemployment, where the former are layoffs for which recall is expected within six months and the latter are layoffs where employment ended involuntarily and the workers have begun looking for work. According to the CPS, about 50 per

cent of all unemployed are classified as job losers and temporary layoffs account for one-third of the job losers.

The controversy over the terminology is dwarfed by the controversy over the occurrence of layoffs. When General Motors announces that it is laying off 20,000 of its workers indefinitely there is widespread press coverage. This attention is well deserved since substantial variation in layoffs (both temporary and permanent) is frequently observed, and layoffs play an important role in cyclical unemployment. Empirical studies of unemployment (for example, Davis, Haltiwanger and Schuh, 1996; Bleakley, Ferris and Fuhrer, 1999) indicate that the typical increase in unemployment during a business cycle slump is primarily due to an increase in employer-initiated discharges, that is, layoffs. For example, in the sharp 1982 recession in the USA, the fraction of the unemployed due to job loss peaked at 63 per cent while in the 2001 recession this fraction peaked at 56 per cent.

The increase in layoff unemployment during recessions is closely tied to the increase in gross job destruction in recessions. Davis, Haltiwanger and Schuh (1996) show that job destruction rises substantially during recessions and is increasingly driven by establishments contracting substantially (for example, with contractions greater than 25 per cent). In turn, Davis, Faberman and Haltiwanger (2006) show that establishments that are contracting intensively use layoffs as the primary means of contraction.

The structure of temporary and permanent layoffs over the cycle has varied over time. Groshen and Potter (2003) show that, in the four recessions in the USA between 1967 and 1990, both temporary layoff and permanent layoff unemployed surged in each of the recessions. However, starting with the 1990–1 recessions, temporary layoffs have played a much smaller role and the rise in job loss has been driven almost entirely by permanent layoffs.

The theory of layoff unemployment has evolved with the relative importance of temporary versus permanent layoffs. Given the important role for temporary layoffs in the 1970s, the so-called ‘implicit contract models’ (see, for example, Azariadis, 1975; Baily, 1974; and Burdett and Mortensen, 1980) were developed during that time to help account for the role of temporary layoffs. The temporary layoff models provide a basis for understanding how in a long-term employer–employee relationship it may be optimal for firms and workers to use temporary layoffs to respond to transitory shocks. However, the increased understanding and role of permanent job destruction and associated permanent job loss has pushed theoretical developments in new directions.

Recent theories that incorporate the evidence on permanent job destruction adopt the premise that the economy is subject to a continuous stream of allocative shocks – shocks that cause idiosyncratic variation in profitability among job sites and worker–job matches (see Davis and Haltiwanger, 1999; Mortensen and Pissarides, 1999; and Shimer, Rogerson and Wright, 2005 for an extensive survey of these theories). The continuous stream of allocative shocks generates the large-scale job and worker reallocation observed in the data. To explicitly model the job and worker reallocation process, these theories incorporate heterogeneity among workers and firms along one or more dimensions. Various theories also emphasize search costs, moving costs, sunk investments and other frictions that impede or otherwise distort the reallocation of factor inputs. The combination of frictions and heterogeneity gives rise to potentially important roles for allocative shocks and the reallocation process in aggregate economic fluctuations.

Theories of cyclical fluctuations in job and worker flows with such reallocation frictions can be classified into two broad types. One type treats fluctuations over time in the intensity of allocative

shocks as an important driving force behind aggregate fluctuations and the pace of reallocation activity. A second type maintains that while allocative shocks and reallocation frictions are important, aggregate shocks drive business cycles and fluctuations in the pace of worker and job reallocation. Although different in emphasis, the two types of theories offer complementary views of labour market dynamics and business cycles, and both point toward a rich set of interactions between aggregate fluctuations and the reallocation process.

One can think of allocative shocks as events that alter the closeness of the match between the desired and actual characteristics of labour and capital inputs. Adverse aggregate consequences can result from such events because of the time and other costs of reallocation activity. In considering this view, it is important to emphasize that allocative shocks affect tangible inputs to the production process (labour and physical capital) and intangible inputs. These intangible inputs include the information capital embodied in an efficient sorting and matching of heterogeneous workers and jobs, knowledge about how to work productively with co-workers, knowledge about suitable locations for particular business activities and about idiosyncratic attributes of those locations, the information capital embodied in long-term customer–supplier and debtor–creditor relationships, and the organization capital embodied in sales, product distribution and job-finding networks. These remarks make clear why the economic adjustments to these shocks are often costly and time consuming. It follows that sharp time variation in the intensity of allocative shocks can cause large fluctuations in gross job flows and in turn unemployment dynamics and layoffs in particular.

The connection between cyclical fluctuations in job destruction and layoffs may also stem from responses to adverse aggregate shocks. An adverse aggregate shock can push many declining and dying plants over an adjustment threshold. During boom times, a firm may choose to continue operating a plant that fails to recover its long-run average cost, because short-run revenues exceed short-run costs, or because of a sufficiently large option value to retaining the plant and its work force. A closely related mechanism emphasizes the changes in the incentives for reallocation over the cycle. The reallocation of specialized labour and capital inputs involves forgone production due to lost work time (for example, unemployment or additional schooling), worker retraining, the retooling of plant and equipment, the adoption of new technology, and the organization of new patterns of production and distribution. On average across firms and workers, the value of forgone production tends to fluctuate procyclically, rising during expansions and falling during recessions. This cyclical pattern generates incentives for both workers and firms to concentrate costly reallocation activity during recessions, when the opportunity cost of the resulting forgone production is relatively low. This mechanism is highlighted in the models of Davis and Haltiwanger (1999), Mortensen and Pissarides (1994) and Caballero and Hammour (1994).

A key question is whether the cyclical fluctuations in job destruction and layoffs reflect efficient or inefficient responses to shocks. Caballero and Hammour (1996) highlight the potential for labour markets to malfunction because of appropriability or hold-up problems. These problems arise whenever investment in a new production unit or the formation of a new employment relationship involves some degree of specificity for workers or employers, and there are difficulties in writing or enforcing complete contracts. In their model, Caballero and Hammour (1996) show that efficient restructuring involves synchronized job creation and destruction and relatively little unemployment. In contrast, the inefficient equilibrium restructuring process that emerges under incomplete contracts involves the decoupling of creation and destruction dynamics and relatively large unemployment responses to negative shocks. As discussed in Mortensen and Pissarides (1999), appropriability problems arise

naturally in many search and matching models. Malcomson (1999) provides a broad discussion of hold-up problems in the labour market.

Overall, understanding layoffs requires understanding of the underlying dynamics of job and worker reallocation. New theories and new data sets have emerged that provide a rich new perspective on the dynamics of the labour market at the micro level and in turn the implications of these dynamics for aggregate fluctuations. Much work remains to be done on both theoretical and empirical questions, particularly on understanding the role of market imperfections in these dynamics. Along these lines, one continuing open question is not only to understand the driving forces of job loss but also the closely related forces of the job gains. After all, the loss of a job has much lower costs to the individual and the economy if the worker in question moves quickly to another job.

See Also

- natural rate of unemployment
- search models of unemployment

Bibliography

Ashenfelter, O. and Card, D., eds. 1999. *Handbook of Labor Economics*. Amsterdam: North-Holland.

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83, 1183–202.

Baily, M.N. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41, 37–50.

Bleakley, H., Ferris, A. and Fuhrer, J. 1999. New data on worker flows during business cycles. *Federal Reserve Bank of Boston Review* July/August, 49–76.

Burdett, L. and Mortensen, D. 1980. Search, layoffs, and labor market equilibrium. *Journal of Political Economy* 88, 652–72.

Bureau of Labor Statistics. Current Population Survey. Online. Available at http://www.census.gov/cps, accessed 8 March 2007.

Bureau of Labour Statistics. Job Openings and Labor Turnover Survey. Online. Available at http://www.bls.gov/jlt, accessed 8 Match 2007.

Caballero, R. and Hammour, M. 1994. The cleansing effect of recessions. *American Economic Review* 84, 1350–68.

Caballero, R. and Hammour, M. 1996. On the timing and efficiency of creative destruction. *Quarterly Journal of Economics* 111, 805–52.

Davis, S.J., Haltiwanger, J. and Schuh, S. 1996. *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Davis, S.J. and Haltiwanger, J. 1999. Gross job flows. In Ashenfelter and Card (1999).

Davis, S.J., Faberman, J. and Haltiwanger, J. 2006. The flow approach to labor markets: new data sources and micro-macro links. *Journal of Economic Perspectives* 20(3), 3–26.

Groshen, E. and Potter, S. 2003. Has structural change contributed to a jobless recovery? *Current Issues in Economics and Finance* 9(8), 1–7.

Malcomson, J.M. 1999. Individual employment contracts. In Ashenfelter and Card (1999).

Mortensen, D.T. and Pissarides, C. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61, 397–415.

Mortensen, D.T. and Pissarides, C. 1999. New developments in models of search in the labor market. In Ashenfelter and Card (1999).

Shimer, R., Rogerson, R. and Wright, R. 2005. Search theoretic models of the labor market. *Journal of Economic Literature* 43, 959–88.

U.S. Department of Labor. 2007. *Employment and Earnings*, various issues, Washington, DC: Government Printing Office.

How to cite this article

Haltiwanger, John. "layoffs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000047> doi:10.1057/9780230226203.0945

The New Palgrave Dictionary of Economics Online

Le Chatelier principle

Eugene Silberberg

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the field of economics, the Le Chatelier principle refers to the differences in the responses of decision variables to changes in parameters when additional constraints are imposed on the system. In the context of demand theory, for example, the Le Chatelier principle is the ‘second law of demand’, that demand curves are more elastic in the long run than in the short run. In many models, additional constraints reduce the absolute response of a decision variable to a change in a parameter.

Keywords

comparative statics; conjugate pairs theorem; envelope theorem; Le Chatelier principle; parameter values; Samuelson, P.

Article

Henri Louis Le Chatelier was a French chemist born in Paris in 1850. In 1884, he offered the following observation:

> Any system in stable chemical equilibrium, subjected to the influence of an external cause which tends to change either its temperature or its condensation (pressure, concentration, number of molecules in unit volume), either as a whole or in some of its parts, can only undergo such internal modifications as would, if produced alone, bring about a change of temperature or of condensation of opposite sign to that resulting from the external cause. (Oliver and Kurtz, 1992)

Later writers produced a more heuristic simplification: ‘If the external conditions…are altered, the equilibrium … will tend to move in such a direction so as to oppose the change in external conditions’ (Fermi, 1937, p. 111, cited in Samuelson, 1949, p. 639), or even more simply: if a stress is

applied to a system at equilibrium, then the system readjusts, if possible, to reduce the stress. The Le Chatelier principle is a firmly established proposition in classical thermodynamics, though its verbal statement is somewhat vague in operational content. In the field of economics, the law of demand, which states that as a price increases, ceteris paribus, consumers will decrease their consumption of that good, is in fact a direct application of the Le Chatelier principle. Consumers (or firms) mitigate the adverse effects of the price increase by utilizing less of that good or input.

Following up a suggestion by his professor and mentor E.B. Wilson at Harvard, Paul Samuelson showed that this principle was a simple application of maximizing behaviour (see especially Samuelson, 1949; 1960a; 1974.) Moreover, physicists and economists – among economists, principally Samuelson – came to realize that the Le Chatelier principle was being used to describe two separate phenomena. The first referred to first-order changes in response to a change in a parameter value, such as a price. The second, which is what the Le Chatelier principle is now generally understood to mean, refers to *differences* in the changes as additional constraints are imposed on the system.

The general case

First-order effects

The most general comparative statics model with explicit maximizing behaviour is *maximize* $y = f(x, \alpha)$ subject to $g(x, \alpha) = 0$, where $x(x_1, \ldots, x_n)$ is a vector of decision variables, $\alpha = (\alpha_1, \ldots, \alpha_m)$ is a vector of parameters (though for simplicity, we treat α as a scalar in the discussion below), and $g(\cdot)$ represents one or more constraints. Models at this level of generality, however, imply no refutable implications and are hence largely uninteresting. In particular, there are never refutable implications for parameters that enter the constraint (see, for example, Silberberg and Suen, 2000). Thus we restrict the analysis to models of the form

$$\text{maximize } y = f(x, \alpha) \tag{1}$$

$$\text{subject to } g(x) = 0. \tag{2}$$

Since it has no effect on the analysis to follow, we consider the case of only one external constraint. Also, parameters β, which enter the constraint but which do not enter the objective function, also do not affect the analysis, and hence we suppress them in the notation. The Lagrangian for this model is

$L = f(x, \alpha) + \lambda g(x)$ producing the necessary first-order conditions (NFOC)

$$L_i = f_i(x, \alpha) + \lambda g_i(x) = 0 i = 1, \ldots, n$$
(3)

$$L_\lambda = g(x) = 0$$
(4)

Assuming the sufficient second-order conditions hold, we can in principle 'solve' for the $n + 1$ explicit choice functions $x = x^*(\alpha)$ and $\lambda^*(\alpha)$. Of course, since these choice functions are the result of solving the NFOC simultaneously, each individual x_i is a function of *all* the parameters, not just the ones which appear in L_i.

Substituting the x_i^*'s into the objective function yields the *indirect objective function* $\phi(\alpha) = f(x^*(\alpha), \alpha)$, the maximum value of *f* for given α, subject to the constraint. Since $\varphi(\alpha)$ is by definition a maximum value, $\phi(\alpha) \geq f(x, \alpha)$, but $\phi(\alpha) = f(x, \alpha)$ when $x = x^*$. Thus the function $F(x, \alpha) = f(x, \alpha) - \phi(\alpha)$ has a (constrained) maximum of zero, with respect to both *x* and α. Thus we consider the *primal-dual* model

$$\textit{maximize}\, F(x, \alpha) = f(x, \alpha) - \phi(\alpha)$$
(5)

$$\textit{subject to}\, g(x) = 0$$
(6)

where the maximization runs over *x* and also α. (In the latter instance, we ask, for given x_i's, what values of the parameters would make these x_i's the maximizing values?) The Lagrangian for this model

is

$$L = f(x, \alpha) - \phi(\alpha) + \lambda g(x) \tag{7}$$

The first-order conditions with respect to *x* are the same as in the original model. With respect to α, the NFOC yield the famous 'envelope theorem'

$$L_\alpha = f_\alpha - \phi_\alpha = 0 \tag{8}$$

When α enters the constraint also, we get the envelope theorem in its most general form,

$$\phi_\alpha = L_\alpha = f_\alpha + \lambda g_\alpha \tag{8a}$$

Importantly, however, since we have restricted the model so that the parameters α do not enter the constraint, the primal-dual model is *an unconstrained maximization in* α. Hence in the α dimensions, the second-order conditions are simply

$$F_{\alpha\alpha} = f_{\alpha\alpha} - \phi_{\alpha\alpha} \leq 0. \tag{9}$$

This inequality says that in the α dimensions, *f* is relatively more concave than φ. This is the fundamental geometrical property that underlies all comparative statics relationships and also the 'second-order' Le Chatelier relationships.

The NFOC (8) are identities when $x = x^*$. That is,

$$\phi_\alpha(\alpha) \equiv f_\alpha(x^*(\alpha), \alpha) \tag{10}$$

Differentiating with respect to α,

$$\Phi_{\alpha\alpha} \equiv \sum_{1}^{n} f_{\alpha i} \frac{\partial x_i^*}{\partial \alpha} + f_{\alpha\alpha} \tag{11}$$

Rearranging terms, using (9) and invariance to the order of differentiation,

$$\Phi_{\alpha\alpha} - f_{\alpha\alpha} \equiv \sum_{1}^{n} f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0 \tag{12}$$

This is the fundamental relation of comparative statics. From it, we can derive Samuelson's famous ‘conjugate pairs’ theorem, namely, that refutable implications occur in maximization models when and only when a parameter enters one and only one first-order condition. For in that case, where say α enters only $L_i = 0$, $f_{j\alpha} \equiv 0$, $j \neq i$, and so (12) reduces to one term:

$$f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0 \tag{13}$$

In this case we can say that the response of x_i is in the same direction as the disturbance to the equilibrium (or, in the case of minimization models, in the opposite direction). These relationships constitute the ‘first-order’ Le Chatelier effects. Note that these results are identical to those in models with no constraints at all, or with multiple constraints, as long as those constraints do not contain the parameter that is changing.

Second-order effects

Suppose now the NFOC hold at the parameter value α^0 and consider now the imposition of an

additional constraint, $h(x) = 0$, with the important restriction that this constraint does not change the original equilibrium, for example, a constraint holding some input fixed at the previous profit maximizing level. Then the new NFOC are solved for new explicit choice functions, $x_i = x_i^s(\alpha)$, where the superscript '*s*' stands for 'short run'. Substituting these short run choice functions into the objective function produces a new indirect objective function, $\psi(\alpha)$. Since the new constraint did not disturb the equilibrium, $\psi(\alpha^0) = \phi(\alpha^0)$ at that point. However, since the objective function is now more constrained, for $\alpha \neq \alpha^0$, $\psi(\alpha) \leq \phi(\alpha)$. Thus the function $G(\alpha) = \psi(\alpha) - \phi(\alpha)$ has an unconstrained maximum (of zero) at $\alpha = \alpha^0$. The NFOC are

$$G_\alpha(\alpha) = \psi_\alpha(\alpha) - \phi_\alpha(\alpha) = 0 \tag{14}$$

We note that $\psi_\alpha(\alpha) = \phi_\alpha(\alpha) = f_\alpha$ using the same analysis leading to eq. (8), since α appears in neither constraint. The second-order conditions are

$$G_{\alpha\alpha}(\alpha) = \psi_{\alpha\alpha}(\alpha) - \phi_{\alpha\alpha}(\alpha) \leq 0 \tag{15}$$

That is, the more constrained indirect objective function $\psi(\alpha)$ is tangent to $\varphi(\alpha)$ at $\alpha = \alpha^0$, but it is relatively more concave, or less convex. Using $\psi_\alpha(\alpha) \equiv \phi_\alpha(\alpha) \equiv f_\alpha$ expressed as identities, and proceeding as in eqs. (10) through (12), inequality (15) yields the general second-order Le Chatelier effects:

$$\sum_1^n f_{i\alpha}\left(\frac{\partial x_i^*}{\partial \alpha} - \frac{\partial x_i^s}{\partial \alpha}\right) \geq 0 \tag{16}$$

In the empirically important case where α enters only the *i*th first-order condition, this summation reduces to one term, producing

$$f_{i\alpha}\frac{\partial x_i^*}{\partial \alpha} \geq f_{i\alpha}\frac{\partial x_i^S}{\partial \alpha}$$
(17)

Thus $\partial x_i^* / \partial \alpha \geq \partial x_i^S / \partial \alpha \geq 0$ when $f_{i\alpha} > 0$, and $\partial x_i^* / \partial \alpha \leq \partial x_i^S / \partial \alpha \leq 0$ when $f_{i\alpha} < 0$. In either case, $\left|\partial x_i^* / \partial \alpha\right| \geq \left|\partial x_i^S / \partial \alpha\right|$.

Examples

Profit maximization

Consider the profit-maximization model *maximize* $\pi = f(x, w, p) = p\theta(x_1 \ldots, x_n) - \Sigma w_i x_i$ Each parameter w_i enters only the ith NFOC, and $f_{x_i w_i} = -1$, so that (13) yields the negative slope property $\partial x_i / \partial w_i \leq 0$. Moreover, (17) yields, in addition, for any additional constraint (not involving w_i) imposed on the initial equilibrium,

$$\frac{\partial x_i^*}{\partial w_i} \leq \frac{\partial x_i^S}{\partial w_i} \leq 0$$
(18)

The ‘long-run’ factor demand functions are more elastic than any short-run factor demands defined as above.

In the case where the additional constraint is simply $x_n = x_n^0$, an analysis based on ‘conditional demands’ (Pollak, 1969) is available. If we substitute this constraint directly into the objective function, the ‘short-run’ demand functions are $x_i = x_i^S(w_1, \ldots w_{n-1}, p, x_n^0)$. These functions are related to the long-run demands by the identity

$$x_i^*(w_1, \ldots, w_n, p) \equiv x_i^S(w_1, \ldots w_{n-1}, p, x_n^*(w_1, \ldots, w_n, p))$$
(19)

Differentiating both sides of this identity with respect to w_i and w_n,

$$\frac{\partial x_i^*}{\partial w_i} \equiv \frac{\partial x_i^s}{\partial w_i} + \frac{\partial x_i^s}{\partial x_n^0}\frac{\partial x_n^*}{\partial w_i} \tag{20}$$

$$\frac{\partial x_i^*}{\partial w_n} \equiv \frac{\partial x_i^s}{\partial x_n^0}\frac{\partial x_n^*}{\partial w_n} \tag{21}$$

Substituting (21) into (20) and using a well-known reciprocity condition yields

$$\frac{\partial x_i^*}{\partial w_i} \equiv \frac{\partial x_i^s}{\partial w_i} + \frac{(\partial x_i^* / \partial w_n)^2}{\partial x_n^* / \partial w_n} \tag{22}$$

Since the last term in (22) is negative, we get the Le Chatelier result (18).

Cost (expenditure) minimization

The cost functions in production theory are derived from the model, *minimize* $C = \sum w_i x_i$ subject to $f(x_1, \ldots, x_n) = y$, where *y* is now a parameter, that is, it is an arbitrary fixed level of output. This model is directly related to the profit maximization model. Write the profit maximization model as *maximize* $py - \sum w_i x_i$ *subject to* $f(x_1, \ldots, x_n) = y$. When output *y* is a variable, this model is the profit-maximization model. If *y* is parametric, it is the constrained cost minimization model. Thus we see that the cost minimization model is the profit maximization model with an added constraint. Denoting the factor demands derived from cost minimization as $x_i = x_i^y(w_1, w_n, y)$, we apply (13) and (17) to derive $\partial x_i^* / \partial w_i \leq \partial x_i^y / \partial w_i \leq 0$. The profit maximizing factor demand function, which incorporates an output effect, is always more elastic with respect to its own price than the constant output factor demand functions, regardless of whether the output effect is positive or negative. We can also show by this method that, if another constraint is imposed on the factors, these cost-minimizing demand functions

become less elastic. When the additional constraint takes on the form of holding some factor fixed, as in the above profit-maximization model, a similar conditional demand process is available (see Silberberg and Suen, 2000).

Marginal cost functions

Many – perhaps most – important economic models incorporate a constraint of the form $g(x_1, \ldots, x_n) = k$. The cost minimization model is an example; so are the various two-factor two-good models in which endowment levels are fixed. The Lagrangian for the cost minimization model is $L = \Sigma w_i x_i + \lambda(y - f(x_1, \ldots, x_n))$. The indirect objective function is the cost function $C = C^*(w_1, \ldots, w_n, y)$. The envelope theorem (8a) identifies $\lambda^*(w_1, \ldots, w_n, y)$ as the marginal cost function: $C_y^* = \lambda^*$. We know from the above comparative statics discussion that cost minimization does not imply a sign for the slope of the marginal cost function, that is, $\partial\lambda^* / \partial y / 0 \rightarrow \partial\lambda^* / \partial y \gtrless 0$. Nonetheless, we can still derive a Le Chatelier result for the marginal cost function.
Adding a new constraint $h(x) = 0$ to the cost minimization model consistent with the original equilibrium produces a new ‘short run’ cost function $C^S(w_1, \ldots w_n, y)$. Since this is more constrained than C^*, it must be the case that $C^* \leq C^S$, but the two are equal at the original equilibrium. Thus the function $F = C^* - C^S$ has an unconstrained maximum (of zero) with respect to all the parameters, and in particular, *y*. Thus $F_y = C_y^* - C_y^S = 0$ and $F_{yy} = C_{yy}^* - C_{yy}^S \leq 0$. But this latter inequality is $\partial\lambda^* / \partial y \leq \partial\lambda^S / \partial y$. That is, the long-run marginal cost function either falls faster or rises slower than the short-run marginal cost function. This is the mathematical foundation for the famous article by Viner (1932) and his draftsman Wong that started it all.

Extensions

The Le Chatelier principle is a local result. Even with the usual sufficient second-order conditions, if some price changes by a finite amount, it is not an implication of the model that the long-run effects are absolutely larger than the short-run effects. However, Milgrom and Roberts (1996) showed, using lattice theory, that, for example, for the profit-maximizing firm model, if all the cross-partials of the production function are everywhere non-negative, the Le Chatelier results hold in the large. A few years later, Suen, Silberberg and Tseng (2000) provided an easier proof of this result, showing also that the global Le Chatelier result held when the factors of production and the fixed factor do not switch from being substitutes to being complements (or vice versa) over the relevant price range.
Samuelson (1960a) analysed Le Chatelier phenomena for equilibrium systems not resulting from an explicit maximization hypothesis, using the ‘well-known’ theorem of reciprocal determinants of Jacobi. (I used to joke to my classes that the theorem was well-known to Jacobi and to Samuelson.) Lady and Quirk (2004) have analysed non-maximizing systems using a theory of cycles in determinants; they prove the Le Chatelier principle applies to systems identified by Morishima (1952), which allows substitutes and complements.

See Also

- comparative statics
- envelope theorem

Bibliography

Fermi, E. 1937. *Thermodynamics*. New York: Dover Publications, p. 111.

Lady, G. and Quirk, J. 2004. The scope of the Le Chatelier principle. Online. Available at http://optima-com.com/LeChat/The%20Scope%20of%20the%20LeChatelier%20Principle.doc, accessed 1 February 2006.

Milgrom, P. and Roberts, J. 1996. The Le Chatelier principle. *American Economic Review* 86, 173–9.

Morishima, M. 1952. On the laws of change in the price system in an economy which contains complementary commodities. *Osaka Economics Papers* 1, 101–13.

Oliver, J. and Kurtz, J. 1992. Henri Louis Le Chatelier, a man of principle. Woodrow Wilson Fellowship Foundation. Online. Available at http://www.woodrow.org/teachers/chemistry/institutes/1992/LeChatelier.html, accessed 31 December 2005.

Pollak, R. 1969. Conditional demand functions and consumption theory. *Quarterly Journal of Economics* 83, 60–78.

Samuelson, P. 1974. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P. 1949. *The Le Chatelier principle in linear programming*. RAND Corporation Monograph. Reprinted in *The Collected Scientific Papers of Paul Samuelson*, vol. 1, ed. J. Stiglitz. Cambridge, MA: MIT Press, 1966.

Samuelson, P. 1960a. Structure of a minimum equilibrium system. In *Essays in Economics and Econometrics: A Volume in Honor of Harold Hotelling*, ed. R. Pfouts. Chapel Hill: University of North Carolina Press.

Samuelson, P. 1960b. An extension of the le Chatelier principle. *Econometrica* 28, 368–79. Reprinted in *The Collected Scientific Papers of Paul Samuelson*, vol. 1, ed. J. Stiglitz. Cambridge, MA: MIT Press, 1966.

Silberberg, E. and Suen, W. 2000. *The Structure of Economics*, 3rd edn. New York: Irwin/McGraw-Hill.

Suen, W., Silberberg, E. and Tseng, P. 2000. The Le Chatelier principle: the long and the short of it. *Economic Theory* 16, 471–76.

Viner, J. 1932. Cost curves and supply curves. *Zeitschrift fur Nationalokonomie* 3, 23–46. Reprinted in American Economic Association, *Readings in Price Theory*. Chicago: Richard D. Irwin, 1952.

How to cite this article

Silberberg, Eugene. "Le Chatelier principle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000049> doi:10.1057/9780230226203.0946

The New Palgrave Dictionary of Economics Online

Le Trosne, Guillaume François (1728 – 1780)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Le Trosne, G. F.; Physiocracy; single tax; value

Article

French lawyer and economist. Born in Orléans, Le Trosne studied natural law philosophy with Pothier in preparation for work as a magistrate. In 1753 he was appointed Royal Councillor at the Orléans Presidial Court, whence he retired in 1773. Le Trosne joined the Physiocrats in 1764 by publishing a book defending the free trade in grain (1765) and articles in *Ephémérides* and other journals. His major economic work, *De l'ordre social*, appeared in 1777, its second volume, *De l'intérêt social*, having major economic content with its discussion of value, circulation, money, industry, and domestic, foreign and colonial trade, partly by way of criticism of Condillac's (1776) anti-physiocratic views on these subjects. Le Trosne died in Paris in 1780.

De l'ordre social sets out the laws required for good government designed to ensure and enhance the reproduction of subsistence and wealth. Two major laws are identified. The first demands freedom for economic activity and security of property (Le Trosne, 1777a, p. 38). The second seeks to secure sufficient government revenue to defray public expenses in providing not only security of property and defence but also public works in communication and transport most favourable to reproduction (1777a, p. 122). The second law entails an appropriate tax system ensured by gradual implementation of the single tax on net product (1777a, p. 147). The remaining discourses of the first volume develop the absolute necessity of these laws from historical examples and from their undesirable consequences when transgressed. Constitutional issues of good government defended in part by standard physiocratic arguments in lengthy footnotes (for example, on luxury, 1777a. pp. 214–19, and free trade, pp. 347–50) form the thrust of the argument in the first volume.

Le Trosne's second volume (1777b) is particularly noted for its theory of value (Meek, 1962, p. 389, n. 1), which distinguishes its various determinants such as usefulness, tastes, relative scarcity and competition but which identifies necessary expenses of production as the major influence on value, hence the name fundamental price (pp. 503–4). To analyse value effects on production and wealth Le

Trosne distinguishes various value forms linking, for example, the excess of the price received for produce by the farmer over costs, to accumulation and the increase of wealth. Other roles for these complex value relationships are illustrated in Le Trosne's perceptive discussions of exchange, money, circulation, the sterility of industry and the benefits of trade for an agricultural nation. This analysis clearly confirms the value foundations of physiocratic theory, crystallized in his demonstration of the special productivity of agriculture by means of a simple example where all payments are assumed to be in kind (‘*en nature*’), thereby demonstrating the inaccuracy of interpretations which neglect the sophisticated physiocratic value analysis (p. 590).

Selected works

1765. *La liberté du commerce des grains, toujours utile et jamais nuisible*. Paris.

1777a. *De l'ordre social*. Paris. Reprinted, Munich: Kraus, 1980.

1777b. *De l'intérêt social, par rapport à la valeur, à la circulation, à l'industrie, & au commerce intérieur & extérieur*. Paris. Reprinted, Munich: Kraus, 1980.

Bibliography

Condillac, E.B. de. 1776. *Le commerce et le gouvernment considérés relativement l'un à l'autre*. Paris.

Meek, R.L. 1962. *The Economics of Physiocracy*. London: George Allen & Unwin.

How to cite this article

Groenewegen, Peter. "Le Trosne, Guillaume François (1728–1780)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000050> doi:10.1057/9780230226203.0947

The New Palgrave Dictionary of Economics Online

learning and evolution in games: adaptive heuristics

H. Peyton Young

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A ‘heuristic’ is a method or rule for solving problems; in game theory it refers to a method for learning how to play. Such a rule is ‘adaptive’ if it is directed towards higher payoffs and is reasonably simple to implement. This article discusses a variety of such rules and the forms of equilibrium that they implement. It turns out that even sophisticated solution concepts, like subgame perfect equilibrium, can be achieved by relatively simple and intuitive methods.

Keywords

adaptive heuristics; commitment; correlated equilibrium; learning; Nash equilibrium; probability; regret; repeated games; strategic learning; subgame perfection

Article

‘Adaptive heuristics’ are simple behavioural rules that are directed towards payoff improvement but may be less than fully rational. The number and variety of such rules are virtually unlimited; here we survey several prominent examples drawn from psychology, computer science, statistics and game theory. Of particular interest are the informational inputs required by different learning rules and the forms of equilibrium to which they lead. We shall begin by considering very primitive heuristics, such as reinforcement learning, and work our way up to more complex forms, such as hypothesis testing, which still, however, fall well short of perfectly rational learning.

One of the simplest examples of a learning heuristic is *cumulative payoff matching*, in which the subject plays actions next period with probabilities proportional to their cumulative payoffs to date. Specifically, consider a finite stage game G that is played infinitely often, where all payoffs are assumed to be strictly positive. Let $a_{ij}(t)$ denote the cumulative payoff to player *i* over all those periods $0 \leq t' \leq t$ when he played action *j*, including some *initial propensity* $a_{ij}(0) > 0$. The cumulative payoff matching rule stipulates that in period *t*+1, player *i* chooses action *j* with probability

$$p_{ij}(t+1) = a_{ij}(t) / \sum_k a_{ik}(t). \quad (1)$$

Notice that the distribution has full support given the assumption that the initial propensities are positive. This idea was first proposed by the psychologist Nathan Herrnstein (1970) to explain certain types of animal behaviour, and falls under the more general rubric of *reinforcement learning* (Bush and Mosteller, 1951; Suppes and Atkinson, 1960; Cross, 1983). The key feature of a reinforcement model is that the probability of choosing an action increases monotonically with the total payoff it has generated in the past (on the assumption that the payoffs are positive). In other words, taking an action and receiving a positive payoff *reinforces* the tendency to take that same action again. This means, in particular, that play can become concentrated on certain actions simply because they were played early and often, that is, play can be *habit-forming* (Roth and Erev, 1995; Erev and Roth, 1998).

Reinforcement models differ in various details that materially affect their theoretical behaviour as well as their empirical plausibility. Under cumulative payoff matching, for example, the payoffs are not discounted, which means that current payoffs have an impact on current behaviour that diminishes as $1/t$. Laboratory experiments suggest, however, that recent payoffs matter more than those long past (Erev and Roth, 1998); furthermore, the rate of discounting has implications for the asymptotic properties of such models (Arthur, 1991).

Another variation in this class of models relies on the concept of an *aspiration level*. This is a level of payoffs, sometimes endogenously determined by past play, that triggers a change in a player's behaviour when current payoffs fall below the level and inertial behaviour when payoffs are above the level. The theoretical properties of these models have been studied for 2×2 games, but relatively little is known about their behaviour in general games (Börgers and Sarin, 2000; Cho and Matsui, 2005).

Next we turn to a class of adaptive heuristics based on the notion of minimizing *regret*, about which more is known in a theoretical sense. Fix a particular player and let $\alpha(t)$ denote the average per period payoff that she received over all periods $t' \leq t$. Let $\alpha_j(t)$ denote the average payoff she *would have* received by playing action j in every period through t, on the assumption that the opponents played as they actually did. The difference $r_j(t) = \alpha_j(t) - \alpha(t)$ is the subject's *unconditional regret* from not having played j in every period through t. (In the computer science literature this is known as *external regret*; see Greenwald and Gondek, 2002.)

The following simple heuristic was proposed by Hart and Mas-Colell (2000; 2001) and is known as *unconditional regret matching*: play each action with a probability that is proportional to the positive part of its unconditional regret, that is,

$$p_j(t+1) = [r_j(t)]_+ / \sum_k [r_k(t)]_+. \quad (2)$$

This learning rule has the following remarkable property: when used by any one player, his regrets become non-positive almost surely as *t* goes to infinity *irrespective of the behaviour of the other players*. When all players use the rule, their time average behaviour converges almost surely to a generalization of correlated equilibrium known as the *Hannan set* or the *coarse correlated equilibrium set* (Hannan, 1957; Moulin and Vial, 1978; Hart and Mas-Colell, 2000; Young, 2004). In general, a *coarse correlated equilibrium* (CCE) is a probability distribution over outcomes (joint actions) such that, given a choice between (*a*) committing *ex ante* to whatever joint action will be realized, and (*b*) committing *ex ante* to a fixed action, given that the others are committed to playing their part of whatever joint action will be realized, every player weakly prefers the former option. By contrast, a *correlated equilibrium* (CE) is a distribution such that, after a player's part of the realized joint action has been disclosed, he would just as soon play it as something else, given that the others are going to play their part of the realized joint action. It is straightforward to show that the coarse correlated equilibria form a convex set that contains the set of correlated equilibria (Young, 2004, ch. 3).

The heuristic specified in (2) belongs to a large family of rules whose time-average behaviour converges almost surely to the coarse correlated equilibrium set; equivalently, that assures no long-run regret for all players simultaneously. For example, this property holds if we let $p_j(t+1) = [r_j(t)]_+^\theta / \Sigma_k [r_k(t)]_+^\theta$ for some exponent $\theta > 0$; one may even take different exponents for different players. Notice that these heuristics put positive probability only on actions that would have done strictly better (on average) than the player's realized average payoff. These are sometimes called *better reply rules*. Fictitious play, by contrast, puts positive probability only on action(s) that would have done *best* against the opponents' frequency distribution of play.

Fictitious play does not necessarily converge to the coarse correlated equilibrium set (CCES); indeed, in some 2×2 coordination games fictitious play causes perpetual miscoordination, in which case both players have unconditional long-run regret (Fudenberg and Kreps, 1993; Young, 1993). By choosing θ to be very large, however, we see that there exist better reply rules that are arbitrarily close to fictitious play and that do converge almost surely to the CCES. Fudenberg and Levine (1995; 1998; 1999) and Hart and Mas-Colell (2001) give general conditions under which stochastic forms of fictitious play converge in time average to the CCES.

Without complicating the adjustment process too much, one can construct rules whose time average behaviour converges almost surely to the *correlated equilibrium set* (CES). To define this class of heuristics we need to introduce the notion of conditional regret. Given a history of play through time *t* and a player *i*, consider the change in per period payoff if *i* had played action k in all those periods $t' \leq t$ when he actually played action *j* (and the opponents played what they did). If the difference is positive, player *i* has conditional regret – he wishes he had played *k* instead of *j*. Formally, *i*'s *conditional regret* at playing *j* instead of *k* up through time *t*, $r^i_{jk}(t)$, is 1/*t* times the increase in payoff that would have resulted from playing *k* instead of *j* in all periods $t' \leq t$. Notice that the average is taken over all *t* periods to date; hence, if *j* was not played very often, $r^i_{jk}(t)$ will be small.

Consider the following *conditional regret matching* heuristic proposed by Hart and Mas-Colell (2000):

if a given agent played action *j* in period *t*, then in period *t*+1 he plays according to the distribution

$$q_k(t+1) = \varepsilon r_{jk}(t)_+ \text{ for all } k \neq j, \text{ and } q_j(t+1) = 1 - \varepsilon \sum_{k \neq j} r_{jk}(t)_+. \quad (3)$$

In effect 1– ε is the degree of inertia, which must be large enough that $q_k(t+1)$ is non-negative for all realizations of the conditional regrets $r_{jk}(t)$. If all players use conditional regret matching and ε is sufficiently small, then almost surely the joint frequency of play converges to the set of correlated equilibria (Hart and Mas-Colell, 2000). Notice that *pointwise* convergence is not guaranteed; the result says only that the empirical distribution converges to a convex *set*. In particular, the players' time-average behaviour may wander from one correlated equilibrium to another. It should also be remarked that, if a single player uses conditional regret matching, there is no assurance that his conditional regrets will become non-positive over time unless we assume that the other players use the same rule. This stands in contrast to unconditional regret matching, which assures non-positive unconditional regret for any player who uses it irrespective of the behaviour of the other players. One can, however, design more sophisticated updating procedures that unilaterally assure no conditional regret; see for example Foster and Vohra (1999), Fudenberg and Levine (1998, ch. 4), Hart and Mas-Colell (2000), and Young (2004, ch. 4).

A natural question now arises: do there exist simple heuristics that allow the players to learn *Nash* equilibrium instead of correlated or still coarser forms of equilibrium? The answer depends on how demanding we are about the long-run convergence properties of the learning dynamic. Notice that the preceding results on regret matching were concerned solely with time-average behaviour; no claim was made that period-by-period behaviour converges to any notion of equilibrium. Yet surely it is period-by-period behaviour that is most relevant if we want to assert that the players have 'learned' to play equilibrium. It turns out that it is very difficult to design adaptive learning rules under which period-by-period behaviour converges almost surely to Nash equilibrium in any finite game, unless one builds in some form of coordination among the players (Hart and Mas-Colell, 2003; 2006). The situation becomes even more problematic if one insists on fully rational, Bayesian learning. In this case it can be shown that there exist games of incomplete information in which no form of Bayesian rational learning causes period-by-period behaviours to come close to Nash equilibrium behaviour even in a probabilistic sense (Jordan, 1991, 1993; Foster and Young, 2001; Young, 2004; see also learning and evolution in games: belief learning).

If one does not insist on full rationality, however, one can design stochastic adaptive heuristics that cause period-by-period behaviours to come close to Nash equilibrium – indeed close to subgame perfect equilibrium – most of the time (without necessarily *converging* to an equilibrium). Here is one approach due to Foster and Young (2003); for related work see Foster and Young (2006) and Germano and Lugosi (2007). Let G be a finite *n*-person game that is played infinitely often. At each point in time, each player thinks that the others are playing i.i.d. strategies. Specifically, at time *t* player *i* thinks that *j* is playing the i.i.d strategy $p_j(t)$ on *j*'s action space, and that the opponents are playing independently; that is, their

joint strategies are given by the product distribution $p_{-i}(t) = \prod_{j \neq i} p_j(t)$. Suppose that *i*'s best response is to play a smoothed best response to $p_{-i}(t)$. Specifically, assume that *i* plays each action *j* with a probability proportional to $e^{\beta u_i(j, p_{-i})}$, where $u_i(j, p_{-i})$ is *i*'s expected utility from playing *j* in every period when the opponents play p_{-i}, and $\beta > 0$ is a *response parameter*. This is known as a *quantal* or *log linear* response function. For brevity, denote *i*'s response in period *t* by $q_i^{\beta}(t)$; this depends, of course, on $p_{-i}(t)$. Player *i* views $p_{-i}(t)$ as a hypothesis that he wishes to test against data. After first adopting this hypothesis he waits for a number of periods (say *s*) while he observes the opponents' behaviour, all the while playing $q_i^{\beta}(t)$. After *s* periods have elapsed, he compares the empirical frequency distribution of the opponents' play during these periods with his hypothesis. Notice that both the empirical frequency distribution and the hypothesized distribution lie in the same compact subset of Euclidean space. If the two differ by more than some tolerance level τ (in the Euclidean metric), he rejects his current hypothesis and chooses a new one.

In choosing a new hypothesis, he may wish to take account of information revealed during the course of play, but we shall also assume he engages in some *experimentation*. Specifically, let us suppose that he chooses a new hypothesis according to a probability density that is uniformly bounded away from zero on the space of hypotheses. One can show the following: given any $\varepsilon > 0$, if the response parameter β is sufficiently large, the test tolerance τ is sufficiently small (given β), and the amount of data collected *s* is sufficiently large (given β and τ), then the players' *period-by-period* behaviours constitute an ε-equilibrium of the stage game G at least $1 - \varepsilon$ of the time (Foster and Young, 2003). In other words, classical statistical hypothesis testing is a heuristic for learning Nash equilibria of the stage game. Moreover, if the players adopt hypotheses that condition on history, they can learn complex equilibria of the repeated game, including forms of subgame perfect equilibrium.

The theoretical literature on strategic learning has advanced rapidly in recent years. A much richer class of learning models has been identified since the mid-1990s, and more is known about their long-run convergence properties. There is also a greater understanding of the various kinds of equilibrium that different forms of learning deliver. An important open question is how these theoretical proposals relate to the empirical behaviour of laboratory subjects. While there is no reason to think that any of these rules can fully explain subjects' behaviour, they can nevertheless play a useful role by identifying phenomena that experimentalists should look for. In particular, the preceding discussion suggests that weaker forms of equilibrium may turn out to be more robust predictors of long-run behaviour than is Nash equilibrium.

See Also

- behavioural game theory
- learning and evolution in games: belief learning

Bibliography

Arthur, W.B. 1991. Designing agents that act like human agents: a behavioral approach to bounded rationality. *American Economic Association, Papers and Proceedings* 81, 353–9.

Börgers, T. and Sarin, R. 2000. Naïve reinforcement learning with endogenous aspirations. *International Economic Review* 31, 921–50.

Bush, R.R. and Mosteller, F. 1951. A mathematical model for simple learning. *Psychological Review* 58, 313–23.

Cho, I.-K. and Matsui, A. 2005. Learning aspiration in repeated games. *Journal of Economic Theory* 124, 171–201.

Cross, J. 1983. *A Theory of Adaptive Economic Behavior*. Cambridge: Cambridge University Press.

Erev, I. and Roth, A.E. 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88, 848–81.

Foster, D.P. and Vohra, R. 1999. Regret in the on-line decision problem. *Games and Economic Behavior* 29, 7–35.

Foster, D.P. and Young, H.P. 2001. On the impossibility of predicting the behavior of rational agents. *Proceedings of the National Academy of Sciences of the USA* 98(222), 12848–53.

Foster, D.P. and Young, H.P. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45, 73–96.

Foster, D.P. and Young, H.P. 2006. Regret testing: learning Nash equilibrium without knowing you have an opponent. *Theoretical Economics* 1, 341–67.

Fudenberg, D. and Kreps, D. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5, 320–67.

Fudenberg, D. and Levine, D. 1995. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control* 19, 1065–90.

Fudenberg, D. and Levine, D. 1998. *The Theory of Learning in Games*. Cambridge MA: MIT Press.

Fudenberg, D. and Levine, D. 1999. Conditional universal consistency. *Games and Economic Behavior* 29, 104–30.

Germano, F. and Lugosi, G. 2007. Global Nash convergence of Foster and Young's regret testing. *Games and Economic Behavior* 60, 135–54.

Greenwald, A. and Gondek, D. 2002. On no-regret learning and game-theoretic equilibria. *Journal of*

Machine Learning 1, 1–20.

Hannan, J. 1957. Approximation to Bayes risk in repeated plays. In *Contributions to the Theory of Games*, vol. 3, ed. M. Dresher, A.W. Tucker and P. Wolfe. Princeton, NJ: Princeton University Press.

Hart, S. and Mas-Colell, A. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 1127–50.

Hart, S. and Mas-Colell, A. 2001. A general class of adaptive strategies. *Journal of Economic Theory* 98, 26–54.

Hart, S. and Mas-Colell, A. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93, 1830–6.

Hart, S. and Mas-Colell, A. 2006. Stochastic uncoupled dynamics and Nash equilibrium. *Games and Economic Behavior* 57, 286–303.

Herrnstein, R.J. 1970. On the law of effect. *Journal of the Experimental Analysis of Behavior* 13, 243–66.

Jordan, J.S. 1991. Bayesian learning in normal form games. *Games and Economic Behavior* 3, 60–81.

Jordan, J.S. 1993. Three problems in learning mixed-strategy equilibria. *Games and Economic Behavior* 5, 368–86.

Moulin, H. and Vial, J.P. 1978. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory* 7, 201–21.

Roth, A.E. and Erev, I. 1995. Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* 8, 164–212.

Suppes, P. and Atkinson, R. 1960. *Markov Learning Models for Multiperson Interaction*. Stanford CA: Stanford University Press.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

Young, H.P. 2004. *Strategic Learning and Its Limits*. Oxford: Oxford University Press.

How to cite this article

Young, H. Peyton. "learning and evolution in games: adaptive heuristics." The New Palgrave Dictionary

of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000256> doi:10.1057/9780230226203.0948

The New Palgrave Dictionary of Economics Online

learning and evolution in games: an overview

William H. Sandholm
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We provide a taxonomy and brief overview of the theory of learning and evolution in games.

Keywords

adaptive learning; belief learning; convergence; deterministic evolutionary dynamics; Dirichlet distributions; evolutionarily stable strategies; evolutionary game theory; fictitious play; forward induction; learning and evolution in games; Markov processes; mathematical biology; natural selection; purification; rational learning; regret; regret matching; repeated games; replicator dynamic; revision protocols; risk dominant equilibrium; stochastic evolutionary dynamics; subgame perfection

Article

The theory of learning and evolution in games provides models of disequilibrium behaviour in strategic settings. Much of the theory focuses on whether and when disequilibrium behaviour will resolve in equilibrium play, and, if it does, on predicting which equilibrium will be played. But the theory also offers techniques for characterizing perpetual disequilibrium play.

1 A taxonomy

Models from *evolutionary game theory* consider the behaviour of large populations in strategic environments. In the biological strand of the theory, agents are genetically programmed to play fixed actions, and changes in the population's composition are the result of natural selection and random mutations. In economic approaches to the theory, agents actively choose which actions to play using simple myopic rules, so that changes in aggregate behaviour are the end result of many individual decisions. *Deterministic evolutionary dynamics*, usually taking the form of ordinary differential equations, are used to describe behaviour over moderate time spans, while *stochastic evolutionary*

dynamics, modelled using Markov processes, are more commonly employed to study behaviour over very long time spans.

Models of *learning in games* focus on the behaviour of small groups of players, one of whom fills each role in a repeated game. These models too can be partitioned into two categories. Models of *heuristic learning* (or *adaptive learning*) resemble evolutionary models, in that their players base their decisions on simple myopic rules. One sometimes can distinguish the two sorts of models by the inputs to the agents' decision rules. In both the stochastic evolutionary model of c, Kandori, Mailath and Rob (1993) and the heuristic learning model of Young (1993), agents' decisions take the form of noisy best responses. But in the former model agents evaluate each action by its performance against the population's current behaviour, while in the latter they consider performance against the time averages of opponents' past play.

In models of *coordinated Bayesian learning* (or *rational learning*), each player forms explicit beliefs about the repeated game strategies employed by other players, and plays a best response to those beliefs in each period. The latter models assume a degree of coordination of players' prior beliefs that is sufficient to ensure that play converges to Nash equilibrium. By dropping this coordination assumption, one obtains the more general class of *Bayesian learning* (or *belief learning*) models. Since such models can entail quite naive beliefs, belief learning models overlap with heuristic learning models – see Section 3 below.

2 Evolutionary game theory

The roots of evolutionary game theory lie in mathematical biology. Maynard Smith and Price (1973) introduced the equilibrium notion of an *evolutionarily stable strategy* (or ESS) to capture the possible stable outcomes of a dynamic evolutionary process by way of a static definition. Later, Taylor and Jonker (1978) offered the *replicator dynamic* as an explicitly dynamic model of the natural selection process. The decade that followed saw an explosion of research on the replicator dynamic and related models of animal behaviour, population ecology, and population genetics: see Hofbauer and Sigmund (1988).

In economics, evolutionary game theory studies the behaviour of populations of strategically interacting agents who actively choose among the actions available to them. Agents decide when to switch actions and which action to choose next using simple myopic rules known as *revision protocols* (see Sandholm, 2006). A population of agents, a game, and a revision protocol together define a stochastic process – in particular, a Markov process – on the set of population states.

2.1 Deterministic evolutionary dynamics

How the analysis proceeds depends on the time horizon of interest. Suppose that for the application in question, our interest is in moderate time spans. Then if the population size is large enough, the idiosyncratic noise in agent's choices is averaged away, so that the evolution of aggregate behaviour follows an almost deterministic path (Benaïm and Weibull, 2003). This path is described by a solution to an ordinary differential equation. For example, Björnerstedt and Weibull (1996) and Schlag (1998) show

that if agents use certain revision protocols based on imitation of successful opponents, then the population's aggregate behaviour follows a solution to Taylor and Jonker's (1978) replicator dynamic.

This argument provides an alternative, economic interpretation of this fundamental evolutionary model.

Much of the literature on deterministic evolutionary dynamics focuses on connections with traditional game theoretic solution concepts. For instance, under a wide range of deterministic dynamics, all Nash equilibria of the underlying game are rest points. While some dynamics (including the replicator dynamic) have additional non-Nash rest points, there are others under which rest points and Nash equilibria are identical (Brown and von Neumann, 1950; Smith, 1984; Sandholm, 2006).

A more important question, though, is whether Nash equilibrium will be approached from arbitrary disequilibrium states. For certain specific classes of games, general convergence results can be established (Hofbauer, 2000; Sandholm, 2007). But beyond these classes, convergence cannot be guaranteed. One can construct games under which no reasonable deterministic evolutionary dynamic will converge to equilibrium – instead, the population cycles through a range of disequilibrium states forever (Hofbauer and Swinkels, 1996; Hart and Mas-Colell, 2003). More surprisingly, one can construct games in which nearly all deterministic evolutionary dynamics not only cycle for ever, but also fail to eliminate strictly dominated strategies (Hofbauer and Sandholm, 2006). If we truly are interested in modelling the dynamics of behaviour, these results reveal that our predictions cannot always be confined to equilibria; rather, more complicated limit phenomena like cycles and chaotic attractors must also be permitted as predictions of play.

2.2 Stochastic evolutionary dynamics

If we are interested in behaviour over very long time horizons, deterministic approximations are no longer valid, and we must study our original Markov process directly. Under certain non-degeneracy assumptions, the long-run behaviour of this process is captured by its unique stationary distribution, which describes the proportion of time the process spends in each population state.

While stochastic evolutionary processes can be more difficult to analyse than their deterministic counterparts, they also permit us to make surprisingly tight predictions. By making the amount of noise in agents' choice rules vanishingly small, one can often ensure that all mass in the limiting stationary distribution is placed on a single population state. This *stochastically stable state* provides a unique prediction of play even in games with multiple strict equilibria (Foster and Young, 1990; Kandori, Mailath and Rob, 1993).

The most thoroughly studied model of stochastic evolution considers agents who usually play a best response to the current population state, but who occasionally choose a strategy at random. Kandori, Mailath and Rob (1993) show that if the agents are randomly matched to play a symmetric 2×2 coordination game, then taking the probability of 'mutations' to zero generates a unique stochastically stable state. In this state, called the *risk dominant equilibrium*, all agents play the action that is optimal against an opponent who is equally likely to choose each action.

Selection results of this sort have since been extended to cases in which the underlying game has an arbitrary number of strategies, as well as to settings in which agents are positioned on a fixed network, interacting only with neighbours (see Kandori and Rob, 1995; Blume, 2003; Ellison, 1993; 2000).

Stochastic stability has also been employed in contexts where the underlying game has a nontrivial

extensive form; these analyses have provided support for notions of backward induction (for example, subgame perfection) and forward induction (for example, signalling game equilibrium refinements): see Nöldeke and Samuelson (1993) and Hart (2002).

Still, these selection results must be interpreted with care. When the number of agents is large or the rate of 'mutation' is small, states that fail to be stochastically stable can be coordinated upon for great lengths of time (Binmore, Samuelson and Vaughan, 1995). Consequently, if the relevant time span for the application at hand is not long enough, the stochastically stable state may not be the only reasonable prediction of behaviour.

3 Learning in games

3.1 Heuristic learning

Learning models study disequilibrium adjustment processes in repeated games. Like evolutionary models, heuristic learning models assume that players employ simple myopic rules in deciding how to act. In the simplest of these models, each player decides how to act by considering the payoffs he has earned in the past. For instance, under reinforcement learning (Börgers and Sarin, 1997; Erev and Roth, 1998), agents choose each strategy with probability proportional to the total payoff that the strategy has earned in past periods.

By considering rules that look not only at payoffs earned, but also at payoffs foregone, one can obtain surprisingly strong convergence results. Define a player's *regret* for (not having played) action *a* to be the difference between the average payoff he would have earned had he always played *a* in the past, and the average payoff he actually received. Under *regret matchingt*, each action whose regret is positive is chosen with probability proportional to its regret. Hart and Mas-Colell (2000) show that regret matching is a *consistent* repeated game strategy: it forces a player's regret for each action to become nonpositive. If used by all players, regret matching ensures that their time-averaged behaviour converges to the set of *coarse correlated equilibria* of the underlying game. (*Coarse correlated equilibrium* is a generalization of correlated equilibrium under which players' incentive constraints must be satisfied at the *ex ante* stage rather than at the interim stage: see Young, 2004.)

Some of the most striking convergence results in the evolution and learning literature establish a stronger conclusion: namely, convergence of time-averaged behaviour to the set of *correlated equilibria*, regardless of the game at hand. The original result of this sort is due to Foster and Vohra (1997; 1998), who prove the result by constructing a calibrated procedure for forecasting opponents' play. A *forecasting procedure* produces probabilistic forecasts of how opponents will act. The procedure is *calibrated* if in those periods in which the forecast is given by the probability vector *p*, the empirical distribution of opponents' play is approximately *p*. It is not difficult to show that if players always choose myopic best responses to calibrated forecasts, then their time-averaged behaviour converges to the set of correlated equilibria.

Hart and Mas-Colell (2000) construct simpler procedures – in particular, procedures that define conditionally consistent repeated game strategies – also ensure convergence to correlated equilibrium. A repeated game strategy is *conditionally consistent* if for each frequently played action *a*, the agent would not have been better off had he always played an alternative action a' in place of *a*. As a matter of

definition, the use of conditionally consistent strategies by all players leads time-averaged behavior to converge to the set of correlated equilibria.

Another variety of heuristic learning models, based on *random search and independent verification*, ensures a stochastic form of convergence to Nash equilibrium regardless of the game being played (Foster and Young, 2003). However, in these models the time required before equilibrium is first reached is quite long, making them most relevant to applications with especially long time horizons.

In some heuristic learning models, players use simple rules to predict how opponents will behave, and then respond optimally to those predictions. The leading examples of such models are *fictitious play* and its stochastic variants (Brown, 1951; Fudenberg and Kreps, 1993): in these models, the prediction about an opponents' next period play is given by the empirical frequencies of his past plays. Beginning with Robinson (1951), many authors have proved convergence results for standard and stochastic fictitious play in specific classes of games (see Hofbauer and Sandholm (2002) for an overview). But as Shapley (1964) and others have shown, these models do not lead to equilibrium play in all games.

3.2 Coordinated Bayesian learning

The prediction rule underlying two-player fictitious play can be described by a belief about the opponent's repeated game strategy that is updated using Bayes's rule in the face of observed play. This belief specifies that the opponent choose his stage game actions in an i.i.d. fashion, conditional on the value of an unknown parameter. (In fact, the player's beliefs about this parameter must come from the family of Dirichlet distributions, the conjugate family of distributions for multinomial trials.) Evidently, each player's beliefs about his opponent are wrong: player 1 believes that player 2 chooses actions in an i.i.d. fashion, whereas player 2 actually plays optimally in response to his own (i.i.d.) predictions about player 1's behaviour. It is therefore not surprising that fictitious play processes do not converge in all games.

In models of *coordinated Bayesian learning* (or *rational learning*), it is not only supposed that players form and respond optimally to beliefs about the opponent's repeated game strategy; it is also assumed that the players' initial beliefs are coordinated in some way. The most studied case is one in which prior beliefs satisfy an absolute continuity condition: if the distribution over play paths generated by the players' actual strategies assigns positive probability to some set of play paths, then so must the distribution generated by each player's prior. A strong sufficient condition for absolute continuity is that each player's prior assigns a positive probability to his opponent's actual strategy.

The fundamental result in this literature, due to Kalai and Lehrer (1993), shows that under absolute continuity, each player's forecast along the path of play is asymptotically correct, and the path of play is asymptotically consistent with Nash equilibrium play in the repeated game. Related convergence results have been proved for more complicated environments in which each player's stage game payoffs are private information (Jordan, 1995; Nyarko, 1998). If the distributions of players types are continuous, then the sense in which play converges to equilibrium can involve a form of purification: while actual play is pure, it appears random to an outside observer.

How much coordination of prior beliefs is needed to prove convergence to equilibrium play? Nachbar (2005) proves that for a large class of repeated games, for any belief learning model, there are no prior beliefs that satisfy three criteria: learnability, consistency with optimal play, and diversity. Thus, if

players can learn to predict one another's behaviour, and are capable of responding optimally to their updated beliefs, then each player's beliefs about his opponents must rule out some seemingly natural strategies a priori. In this sense, the assumption of coordinated prior beliefs that ensures convergence to equilibrium in rational learning models does not seem dramatically weaker than a direct assumption of equilibrium play.

For additional details about the theory of learning and evolution in games, we refer the reader to the entries on specific topics listed in the cross-references below.

See Also

- deterministic evolutionary dynamics
- learning and evolution in games: adaptive heuristics
- learning and evolution in games: belief learning
- learning and evolution in games: ESS
- stochastic adaptive dynamics

The author thanks John Nachbar for a number of helpful conversations and for sharing his expertise on coordinated Bayesian learning. Financial support under NSF Grants SES-0092145 and SES-0617753 is gratefully acknowledged.

Bibliography

Benaïm, M. and Weibull, J.W. 2003. Deterministic approximation of stochastic evolution in games. *Econometrica* 71, 873–903.

Binmore, K., Samuelson, L. and Vaughan, R. 1995. Musical chairs: modeling noisy evolution. *Games and Economic Behavior* 11, 1–35.

Björnerstedt, J. and Weibull, J.W. 1996. Nash equilibrium and evolution by imitation. In *The Rational Foundations of Economic Behavior*, ed. K.J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt. New York: St. Martin's Press.

Blume, L.E. 2003. How noise matters. *Games and Economic Behavior* 44, 251–71.

Börgers, T. and Sarin, R. 1997. Learning through reinforcement and the replicator dynamics. *Journal of Economic Theory* 77, 1–14.

Brown, G.W. 1951. Iterative solutions of games by fictitious play. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: Wiley.

Brown, G.W. and von Neumann, J. 1950. Solutions of games by differential equations. In *Contributions*

to the Theory of Games I, ed. H.W. Kuhn and A.W. Tucker, vol. 24 of *Annals of Mathematics Studies*. Princeton: Princeton University Press.

Ellison, G. 1993. Learning, local interaction, and coordination. *Econometrica* 61, 1047–71.

Ellison, G. 2000. Basins of attraction, long run equilibria, and the speed of step-by-step evolution. *Review of Economic Studies* 67, 17–45.

Erev, I. and Roth, A.E. 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88, 848–81.

Foster, D.P. and Vohra, R. 1997. Calibrated learning and correlated equilibrium. *Games and Economic Behavior* 21, 40–55.

Foster, D.P. and Vohra, R. 1998. Asymptotic calibration. *Biometrika* 85, 379–90.

Foster, D.P. and Young, H.P. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38, 219–32.

Foster, D.P. and Young, H.P. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45, 73–96.

Fudenberg, D. and Kreps, D.M. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5, 320–67.

Fudenberg, D. and Levine, D.K. 1998. *Theory of Learning in Games*. Cambridge: MIT Press.

Hart, S. 2002. Evolutionary dynamics and backward induction. *Games and Economic Behavior* 41, 227–64.

Hart, S. and Mas-Colell, A. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 1127–50.

Hart, S. and Mas-Colell, A. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93, 1830–6.

Hofbauer, J. 2000. From Nash and Brown to Maynard Smith: equilibria, dynamics, and ESS. *Selection* 1, 81–8.

Hofbauer, J. and Sandholm, W.H. 2002. On the global convergence of stochastic fictitious play. *Econometrica* 70, 2265–94.

Hofbauer, J. and Sandholm, W.H. 2006. Survival of dominated strategies under evolutionary dynamics. Working paper. University of Vienna and University of Wisconsin.

Hofbauer, J. and Sigmund, K. 1988. *Theory of Evolution and Dynamical Systems*. Cambridge: Cambridge University Press.

Hofbauer, J. and Swinkels, J.M. 1996. A universal Shapley example. Working paper. University of Vienna and Northwestern University.

Jordan, J.S. 1995. Bayesian learning in repeated games. *Games and Economic Behavior* 9, 8–20.

Kalai, E. and Lehrer, E. 1993. Rational learning leads to Nash equilibrium. *Econometrica* 61, 1019–45.

Kandori, M., Mailath, G.J. and Rob, R. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61, 29–56.

Kandori, M. and Rob, R. 1995. Evolution of equilibria in the long run: a general theory and applications. *Journal of Economic Theory* 65, 383–414.

Maynard Smith, J. and Price, G.R. 1973. The logic of animal conflict. *Nature* 246, 15–8.

Nachbar, J.H. 2005. Beliefs in repeated games. *Econometrica* 73, 459–80.

Nöldeke, G. and Samuelson, L. 1993. An evolutionary analysis of backward and forward induction. *Games and Economic Behavior* 5, 425–54.

Nyarko, Y. 1998. Bayesian learning and convergence to Nash equilibria without common priors. *Economic Theory* 11, 643–55.

Robinson, J. 1951. An iterative method of solving a game. *Annals of Mathematics* 54, 296–301.

Sandholm, W.H. 2006. Pairwise comparison dynamics and evolutionary foundations for Nash equilibrium. Working paper. University of Wisconsin.

Sandholm, W.H. 2007. Population games and evolutionary dynamics. Cambridge, MA: MIT Press.

Schlag, K.H. 1998. Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory* 78, 130–56.

Shapley, L.S. 1964. Some topics in two person games. In *Advances in Game Theory*, ed. M. Dresher, L. S. Shapley and A.W. Tucker, vol. 52 of *Annals of Mathematics Studies*. Princeton: Princeton University

Press.

Smith, M.J. 1984. The stability of a dynamic model of traffic assignment: an application of a method of Lyapunov. *Transportation Science* 18, 245–52.

Taylor, P.D. and Jonker, L. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–56.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

Young, H.P. 2004. *Strategic Learning and Its Limits*. Oxford: Oxford University Press.

How to cite this article

Sandholm, William H. "learning and evolution in games: an overview." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000241> doi:10.1057/9780230226203.0949

The New Palgrave Dictionary of Economics Online

learning and evolution in games: belief learning

John Nachbar

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the context of learning in games, *belief learning* refers to models in which players are engaged in a dynamic game and each player optimizes with respect to a *prediction rule* that gives a forecast of next-period opponent behaviour as a function of the current history. This article focuses on the most studied class of dynamic games, namely, two-player discounted repeated games with finite stage game action sets and perfect monitoring.

Keywords

belief learning; best-response dynamics; convergence; Cournot, A. A.; distributional strategies; dynamic games; fictitious play; folk theorem; learnable best-response property; mixed strategy equilibrium; myopia; perfect monitoring; prediction rules; purification; repeated games

Article

In the context of learning in games, *belief learning* refers to models in which players are engaged in a dynamic game and each player optimizes, or ε optimizes, with respect to a *prediction rule* that gives a forecast of next period opponent behaviour as a function of the current history. This article focuses on the most studied class of dynamic games, two-player discounted repeated games with finite stage game action sets and perfect monitoring. An important example of a dynamic game that violates perfect monitoring and therefore falls outside this framework is Fudenberg and Levine (1993). For a more comprehensive survey of belief learning, see Fudenberg and Levine (1998).

The earliest example of belief learning is the *best-response dynamics* of Cournot (1838). In Cournot's model, each player predicts that her opponent will repeat next period whatever action her opponent chose in the previous period.

The most studied belief learning model is *fictitious play* (Brown, 1951), and its variants. In fictitious play, each player predicts that the probability that her opponent will play an action, say *L*, next period is

a weighted sum of an initial probability on L and the frequency with which L has been chosen to date. The weight on the frequency is $t/(t+k)$, where t is the number of periods thus far and $k>0$ is a parameter. The larger is k, the more periods for which the initial probability significantly affects forecasting.
The remainder of this article discusses four topics: (1) belief learning versus Bayesian learning, (2) convergence to equilibrium, (3) special issues in games with payoff uncertainty, and (4) sensible beliefs.

Belief learning versus Bayesian learning

Recall that, in a repeated game, a behaviour strategy gives, for every history, a probability over the player's stage game actions next period. In a Bayesian model, each player chooses a behaviour strategy that best responds to a *belief*, a probability distribution over the opponent's behaviour strategies.
Player 1's prediction rule about player 2 is mathematically identical to a behaviour strategy for player 2. Thus, any belief learning model is equivalent to a Bayesian model in which each player optimizes with respect to a belief that places probability 1 on her prediction rule, now reinterpreted as the opponent's behaviour strategy.
Conversely, any Bayesian model is equivalent to a belief learning model. Explicitly, for any belief over player 2's behaviour strategies there is a degenerate belief, assigning probability 1 to a particular behaviour strategy, that is equivalent in the sense that both beliefs induce the same distributions over play in the game, no matter what behaviour strategy player 1 herself adopts. This is a form of Kuhn's theorem (Kuhn, 1964). I refer to the behaviour strategy used in the degenerate belief as a *reduced form* of the original belief. Thus, any Bayesian model is equivalent to a Bayesian model in which each player's belief places probability 1 on the reduced form, and any such Bayesian model is equivalent to a belief learning model.
As an example, consider fictitious play. I focus on stage games with just two actions, L and R. By an i.i.d. strategy for player 2, I mean a behaviour strategy in which player 2 plays L with probability q, independent of history. Thus, if $q=1/2$, then player 2 always randomizes 50:50 between L and R. Fictitious play is equivalent to a degenerate Bayesian model in which each player places probability 1 on the fictitious play prediction rule, and one can show that this is equivalent in turn to a non-degenerate Bayesian model in which the belief is represented as a beta distribution over q. The uniform distribution over q, for example, corresponds to taking the initial probability of L to be 1/2 and the parameter k to be 2.
There is a related but distinct literature in which players optimize with respect to *stochastic* prediction rules. In some cases (for example, Foster and Young, 2003), these models have a quasi-Bayesian interpretation: most of the time, players optimize with respect to fixed prediction rules, as in a Bayesian model, but occasionally players switch to new prediction rules, implicitly abandoning their priors.

Convergence to equilibrium

Within the belief learning literature, the investigation of convergence to equilibrium play splits into two branches. One branch investigates convergence within the context of specific classes of belief learning models. The best-response dynamics, for example, converge to equilibrium if the stage game is solvable by the iterated deletion of strictly dominated strategies. See Bernheim (1984) and, for a more general

class of models, Milgrom and Roberts (1991). For an ε optimizing variant of fictitious play, convergence to approximate equilibrium play obtains for all zero-sum games, all games with an interior ESS, and all common interest games, in addition to all games that are strict dominance solvable, with the approximation closer the smaller is ε. Somewhat weaker convergence results are available for supermodular games. These claims follow from results in Hofbauer and Sandholm (2002).

In the results surveyed above, convergence is to repeated play of a single-stage game Nash equilibrium; in the case of ε fictitious play, this equilibrium may be mixed. There is a large body of work on convergence that is weaker than what I am considering here. In particular, there has been much work on convergence of the empirical marginal or joint distributions. For mixed strategy equilibrium, it is possible for empirical distributions to converge to equilibrium even though play does not resemble repeated equilibrium play; play may exhibit obvious cycles, for example. The study of convergence to equilibrium play is relatively recent and was catalysed by Fudenberg and Kreps (1993).

There are classes of games that cause convergence problems for many standard belief learning models, even when one considers only weak forms of convergence, such as convergence of the empirical marginal distributions (see Shapley, 1962; Jordan, 1993). Hart and Mas-Colell (2003; 2006) (hereafter HM) shed light on non-convergence by investigating learning models, including but not limited to belief learning models, that are *decoupled*, meaning that player 1's behaviour does not depend directly on player 2's stage game payoffs. A continuous time version of fictitious play fits into the framework of Hart and Mas-Colell (2003). The HM results imply that universal convergence is impossible for large classes of decoupled belief learning models: for any such model there exist stage games and initial conditions for which play fails to converge to equilibrium play.

The second branch of the literature, for which Kalai and Lehrer (1993a) (hereafter KL) is the central paper, takes a Bayesian perspective and asks what conditions on beliefs are sufficient to give convergence to equilibrium play. I find it helpful to characterize this literature in the following way. Say that a belief profile (giving a belief for each player) has the *learnable best-response property* (LBR) if there is a profile of best-response strategies (LBR strategies) such that, if the LBR strategies are played, then each player learns to predict the play path.

A player *learns to predict the play path* if her prediction of next period's play is asymptotically as good as if she knew her opponent's behaviour strategy. If the behaviour strategies call for randomization then players accurately predict the distribution over next period's play rather than the realization of next period's play. For example, consider a 2×2 game in which player 1 has stage game actions *T* and *B* and player 2 has stage game actions *L* and *R*. If player 2 is randomizing 50:50 every period and player 1 learns to predict the path of play, then for every ε there is a time, which depends on the realization of player 2's strategy, after which player 1's next period forecast puts the probability of *L* within ε of 1/2. (This statement applies to a set of play paths that arises with probability 1 with respect to the underlying probability model; I gloss over this sort of complication both here and below.) For a more complicated example, suppose that in period *t* player 2 plays *L* with probability $1-\alpha$, where α is the frequency that the players have played the profile (*B*, *R*). If player 1 learns to predict the play path, then for any ε there is a time, which now depends on the realization of both players' strategies, after which player 1's next period forecast puts the probability of *L* within ε of $1-\alpha$.

Naively, if LBR holds, and players are using their LBR strategies, then, in the continuation game, players are optimizing with respect to posterior beliefs that are asymptotically correct and so

continuation behaviour strategies should asymptotically be in equilibrium. This intuition is broadly correct, but there are three qualifications.

First, in general, convergence is to Nash equilibrium play in the *repeated* game, not necessarily to repeated play of a single stage game equilibrium. If players are myopic (meaning that players optimize each period as though their discount factors were zero), then the set of equilibrium play paths comprises all possible sequences of stage game Nash equilibria, which is a very large set if the stage game has more than one equilibrium. If players are patient, then the folk theorem applies and the set of possible equilibrium paths is typically even larger.

Second, convergence is to an equilibrium play path, not necessarily to an equilibrium of the repeated game. The issue is that LBR implies accurate forecasting only along the play path. A player's predictions about how her opponent would respond to deviations may be grossly in error, for ever. Therefore, posterior beliefs need *not* be asymptotically correct and, unless players are myopic, continuation behaviour strategies need *not* be asymptotically in equilibrium. Kalai and Lehrer (1993b) shows that behaviour strategies can be doctored at information sets off the play path so that the modified behaviour strategies are asymptotically in equilibrium yet still generate the same play path. This implies that the play path of the original strategy profile was asymptotically an equilibrium play path.

Third, the exact sense in which play converges to equilibrium play depends on the strength of learning. See KL and also Sandroni (1998).

KL shows that a strong form of LBR holds if beliefs satisfy an absolute continuity condition: each player assigns positive probability to any (measurable) set of play paths that has positive probability given the players’ actual strategies. A sufficient condition for this is that each player assigns positive, even if extremely low, probability to her opponent's actual strategy, a condition that KL call *grain of truth*. Nyarko (1998) provides the appropriate generalization of absolute continuity for games with type space structures, including the games with payoff uncertainty discussed below.

Games with payoff uncertainty

Suppose that, at the start of the repeated game, each player is privately informed of his or her stage game payoff function, which remains fixed throughout the course of the repeated game. Refer to player *i*'s stage game payoff function as her *payoff type*. Assume that the joint distribution over payoff functions is independent (to avoid correlation issues that are not central to my discussion) and commonly known. Each player can condition her behaviour strategy in the repeated game on her realized payoff type. A mathematically correct way of representing this conditioning is via distributional strategies (see Milgrom and Weber, 1985).

For any belief about player 2, now a probability distribution over player 2's distributional strategies, and given the probability distribution over player 2's payoff types, there is a behaviour strategy for player 2 in the repeated game that is equivalent in the sense that it generates the same distribution over play paths. Again, this is essentially Kuhn's theorem. And again, I refer to this behaviour strategy as a *reduced form*.

Say that a player *learns to predict the play path* if her forecast of next period's play is asymptotically as good as if she knew the reduced form of her opponent's distributional strategy. This definition specializes to the previous one if the distribution over types is degenerate. If distributional strategies are

in equilibrium then, in effect, each player is optimizing with respect to a degenerate belief that puts probability one on her opponent's actual distributional strategy and in this case players trivially learn to predict the path of play.

One can define LBR for distributional strategies and, as in the payoff certainty case, one can show that LBR implies convergence to equilibrium play in the repeated game with payoff types. More interestingly, there is a sense in which play converges to equilibrium play of the *realized* repeated game – the repeated game determined by the realized type profile. The central paper is Jordan (1991). Other important papers include KL (cited above), Jordan (1995), Nyarko (1998), and Jackson and Kalai (1999) (which studies recurring rather than repeated games).

Suppose first that the realized type profile has positive probability. In this case, if a player learns to predict the play path, then, as shown by KL, her forecast is asymptotically as good as if she knew both her opponent's distributional strategy *and* her opponent's realized type. LBR then implies that actual play, meaning the distribution over play paths generated by the realized behaviour strategies, converges to equilibrium play of the realized repeated game. For example, suppose that the type profile for matching pennies gets positive probability. In the unique equilibrium of repeated matching pennies, players randomize 50:50 in every period. Therefore, LBR implies that, if the matching pennies type profile is realized, then each player's behaviour strategy in the realized repeated game involves 50:50 randomization asymptotically.

If the distribution over types admits a continuous density, so that no type profile receives positive probability, then the form of convergence is more subtle. Suppose that players are myopic and that the realized stage game is like matching pennies, with a unique and fully mixed equilibrium. Given myopia, the unique equilibrium of the realized repeated game calls for repeated play of the stage game equilibrium. In particular, it calls for players to randomize. It is not hard to show, however, that in a type space game with a continuous density, optimization calls for each player to play a pure strategy for almost every realized type. Thus, for almost every realized type profile in a neighbourhood of a game like matching pennies, actual play (again meaning the distribution over play paths generated by the realized behaviour strategies) cannot converge to equilibrium play, *even if the distributional strategies are in equilibrium*. Foster and Young (2001) provides a generalization for non-myopic players.

There is, however, a weaker sense in which play nevertheless does converge to equilibrium play in the realized repeated game. For simplicity, assume that each player knows the other's distributional strategy and that these strategies are in equilibrium. One can show that to an outsider observed play looks asymptotically like equilibrium play in the realized repeated game. In particular, if the realized game is like repeated matching pennies then observed play looks random. Moreover, to a player in the game, opponent behaviour looks random because, even though she knows her opponent's distributional strategy, she does not know her opponent's type. As play proceeds, each player in effect learns more about her opponent's type, but never enough to zero in on her opponent's realized, pure, behaviour strategy. Thus, when the distribution over types admits a continuous density, convergence to equilibrium involves a form of purification in the sense of Harsanyi (1973), a point that has been emphasized by Nyarko (1998) and Jackson and Kalai (1999).

Sensible beliefs

A number of papers investigate classes of prediction rules that are sensible in that they exhibit desirable properties, such as the ability to detect certain kinds of patterns in opponent behaviour (see Aoyagi, 1996; Fudenberg and Levine, 1995; 1999; Sandroni, 2000).

Nachbar (2005) instead studies the issue of sensible beliefs from a Bayesian perspective. For simplicity, focus on learning models with known payoffs. Fix a belief profile, fix a subset of behaviour strategies for each player, and consider the following criteria for these subsets.

- *Learnability* – given beliefs, if players play a strategy profile drawn from these subsets then they learn to predict the play path.
- *Richness*. Informally (the formal statement is tedious), richness requires that if a behaviour strategy is included in one of the strategy subsets then certain variations on that strategy must be included as well. Richness, called CSP in Nachbar (2005), is satisfied automatically if the strategy subsets consist of all strategies satisfying a standard complexity bound, the same bound for both players. Thus richness holds if the subsets consist of all strategies with *k*-period memory, or all strategies that are automaton implementable, or all strategies that are Turing implementable, and so on.
- *Consistency* – each player's subset contains a best response to her belief.

The motivating idea is that, if beliefs are probability distributions over strategy subsets satisfying learnability, richness, and consistency, then beliefs are sensible, or at least are candidates for being considered sensible. Nachbar (2005) studies whether any such beliefs exist.

Consider, for example, the Bayesian interpretation of fictitious play in which beliefs are probability distributions over the i.i.d. strategies. The set of i.i.d. strategies satisfies learnability and richness. But for any stage game in which neither player has a weakly dominant action, the i.i.d. strategies violate consistency: any player who is optimizing will not be playing i.i.d.

Nachbar (2005) shows that this feature of Bayesian fictitious play extends to all Bayesian learning models. For large classes of repeated games, for *any* belief profile there are *no* strategy subsets that simultaneously satisfy learnability, richness, and consistency. Thus, for example, if each player believes the other is playing a strategy that has a *k*-period memory, then one can show that learnability and richness hold but consistency fails: best responding in this setting requires using a strategy with a memory of more than *k* periods. The impossibility result generalizes to ε optimization and ε consistency, for ε sufficiently small. The result also generalizes to games with payoff uncertainty (with learnability, richness, and consistency now defined in terms of distributional strategies) (see Nachbar, 2001).

I conclude with four remarks. First, since the set of all strategies satisfies richness and consistency, it follows that the set of all strategies is not learnable for *any* beliefs: for any belief profile there is a strategy profile that the players will not learn to predict. This can also be shown directly by a diagonalization argument along the lines of Oakes (1985) and Dawid (1985). The impossibility result of Nachbar (2005) can be viewed as a game theoretic version of Dawid (1985). For a description of what subsets *are* learnable, see Noguchi (2005).

Second, if one constructs a Bayesian learning model satisfying learnability and consistency then LBR holds and, if players play their LBR strategies, play converges to equilibrium play. This identifies a

potentially attractive class of Bayesian models in which convergence obtains. The impossibility result says, however, that if learnability and consistency hold, then player beliefs must be partially equilibrated in the sense of, in effect, excluding some of the strategies required by richness.

Third, consistency is not *necessary* for LBR or convergence. For example, for many stage games, variants of fictitious play satisfy LBR and converge even though these learning models are inconsistent. The impossibility result is a statement about the ability to construct Bayesian models with certain properties; it is not a statement about convergence per se.

Last, learnability, richness, and consistency may be too strong to be taken as necessary conditions for beliefs to be considered sensible. It is an open question whether one can construct Bayesian models satisfying conditions that are weaker but still strong enough to be interesting.

See Also

- deterministic evolutionary dynamics
- learning and evolution in games: adaptive heuristics
- learning and evolution in games: an overview
- learning and evolution in games: ESS
- purification
- repeated games
- stochastic adaptive dynamics

Bibliography

Aoyagi, M. 1996. Evolution of beliefs and the Nash equilibrium of normal form games. *Journal of Economic Theory* 70, 444–69.

Bernheim, B.D. 1984. Rationalizable strategic behavior. *Econometrica* 52, 1007–28.

Brown, G.W. 1951. Iterative solutions of games by fictitious play. In *Activity Analysis of Production and Allocation*, ed. T.J. Koopmans. New York: Wiley.

Cournot, A. 1838. Researches into the Mathematical Principles of the Theory of Wealth. Trans. N.T. Bacon, New York: Kelley, 1960.

Dawid, A.P. 1985. The impossibility of inductive inference. *Journal of the American Statistical Association* 80, 340–1.

Foster, D. and Young, P. 2001. On the impossibility of predicting the behavior of rational agents. *Proceedings of the National Academy of Sciences* 98, 12848–53.

Foster, D. and Young, P. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45, 73–96.

Fudenberg, D. and Kreps, D. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5, 320–67.

Fudenberg, D. and Levine, D. 1993. Steady state learning and Nash equilibrium. *Econometrica* 61, 547–74.

Fudenberg, D. and Levine, D. 1995. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control* 19, 1065–89.

Fudenberg, D. and Levine, D. 1998. *Theory of Learning in Games*. Cambridge, MA: MIT Press.

Fudenberg, D. and Levine, D. 1999. Conditional universal consistency. *Games and Economic Behavior* 29, 104–30.

Harsanyi, J. 1973. Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory* 2, 1–23.

Hart, S. and Mas-Colell, A. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93, 1830–6.

Hart, S. and Mas-Colell, A. 2006. Stochastic uncoupled dynamics and Nash equilibrium. *Games and Economic Behavior* 57, 286–303.

Hofbauer, J. and Sandholm, W. 2002. On the global convergence of stochastic fictitious play. *Econometrica* 70, 2265–94.

Jackson, M. and Kalai, E. 1999. False reputation in a society of players. *Journal of Economic Theory* 88, 40–59.

Jordan, J.S. 1991. Bayesian learning in normal form games. *Games and Economic Behavior* 3, 60–81.

Jordan, J.S. 1993. Three problems in learning mixed-strategy Nash equilibria. *Games and Economic Behavior* 5, 368–86.

Jordan, J.S. 1995. Bayesian learning in repeated games. *Games and Economic Behavior* 9, 8–20.

Kalai, E. and Lehrer, E. 1993a. Rational learning leads to Nash equilibrium. *Econometrica* 61, 1019–45.

Kalai, E. and Lehrer, E. 1993b. Subjective equilibrium in repeated games. *Econometrica* 61, 1231–40.

Kuhn, H.W. 1964. Extensive games and the problem of information. In *Contributions to the Theory of Games*, vol. 2, ed. M. Dresher, L.S. Shapley and A.W. Tucker. Princeton: Princeton University Press.

Milgrom, P. and Roberts, J. 1991. Adaptive and sophisticated learning in repeated normal form games. *Games and Economic Behavior* 3, 82–100.

Milgrom, P. and Weber, R. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10, 619–32.

Nachbar, J.H. 2001. Bayesian learning in repeated games of incomplete information. *Social Choice and Welfare* 18, 303–26.

Nachbar, J.H. 2005. Beliefs in repeated games. *Econometrica* 73, 459–80.

Noguchi, Y. 2005. Merging with a set of probability measures: a characterization. Working paper, Kanto Gakuin University.

Nyarko, Y. 1998. Bayesian learning and convergence to Nash equilibria without common priors. *Economic Theory* 11, 643–55.

Oakes, D. 1985. Self-calibrating priors do not exist. *Journal of the American Statistical Association* 80, 339–42.

Sandroni, A. 1998. Necessary and sufficient conditions for convergence to Nash equilibrium: the almost absolute continuity hypothesis. *Games and Economic Behavior* 22, 121–47.

Sandroni, A. 2000. Reciprocity and cooperation in repeated coordination games: the principled-player approach. *Games and Economic Behavior* 32, 157–82.

Shapley, L. 1962. On the nonconvergence of fictitious play. Discussion Paper RM-3026, RAND.

How to cite this article

Nachbar, John. "learning and evolution in games: belief learning." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000305> doi:10.1057/9780230226203.0950

The New Palgrave Dictionary of Economics Online

learning and evolution in games: ESS

Ross Cressman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The ESS concept, developed in the 1970s to predict through static fitness comparisons the evolutionary outcome of individual behaviours in a biological species, emerged as the cornerstone of evolutionary game theory. This theory is now as central to the analysis of strategic interactions in the social and management sciences as in the life sciences. The ESS also addresses stability questions for dynamics describing how individual behaviours evolve over time. Here, we summarize ESS theory as originally developed for symmetric two-player games and then discuss generalizations to population games, extensive form games, games with continuous strategy spaces, asymmetric and bimatrix games.

Keywords

asymmetric games; asymptotic stability; backward induction; best response dynamics; bimatrix games; buyer–seller game; common interest games; continuously stable strategy; direct ESS; ESS (evolutionarily stable strategy); ESSet; evolutionary dynamics; evolutionary game theory; extensive form games; invasion barrier; local superiority; Nash demand game; Nash equilibrium (NE); NE component; neighbourhood invader strategy; neighbourhood superiority; neutrally stable strategies; normal form games; owner–intruder game; partnership games; pervasive strategy; playing-the-field models; population games; prisoner's dilemma game; replicator equation; rock–scissors–paper game; Selten, R.; strictly stable game; symmetrical games; two-species ESS; war of attrition game

Article

1 . Introduction

According to John Maynard Smith in his influential book *Evolution and the Theory of Games* (1982, p.10), an ESS (that is, an *evolutionarily stable strategy*) is ‘a strategy such that, if all members of the population adopt it, then no mutant strategy could invade the population under the influence of natural selection’. The ESS concept, based on static fitness comparisons, was originally introduced and developed in the biological literature (Maynard Smith and Price, 1973) as a means to predict the eventual outcome of evolution for individual behaviours in a single species. It avoids the complicated dynamics of the evolving population that may ultimately depend on spatial, genetic and population size effects.
To illustrate the Maynard Smith (1982) approach, suppose individual fitness is the expected payoff in a random pairwise contest. The ESS strategy p^* must then do at least as well as a mutant strategy p in their most common contests against p^* and, if these contests yield the same payoff, then p^* must do better than p in their rare contests against a mutant. That is, Maynard Smith's definition applied to a symmetric two-player game says p^* is an ESS if and only if, for all $p \neq p^*$,

$$\text{(i)} \quad \pi(p, p^*) \leq \pi(p^*, p^*) \qquad \text{(equilibrium condition)}$$

$$\text{(ii)} \quad \text{if } \pi(p, p^*) = \pi(p^*, p^*), \ \pi(p, p) < \pi(p^*, p) \qquad \text{(stability condition)}$$

(1)

where $\pi(p, \hat{p})$ is the payoff of p against $\hat{p}$. One reason the ESS concept has proven so durable is that it has equivalent formulations that are equally intuitive (see especially the concepts of invasion barrier and local superiority in Section 2.1).

By (1) (i), an ESS is a Nash equilibrium (NE) with the extra refinement condition (ii) that seems heuristically related to dynamic stability. In fact, there is a complex relationship between the static ESS conditions and dynamic stability, as illustrated throughout this article with specific reference to the replicator equation. It is this relationship that formed the initial basis of what has come to be known as 'evolutionary game theory'.

ESS theory (and evolutionary game theory in general) has been extended to many classes of games besides those based on a symmetric two-player game. This article begins with ESS theory for symmetric normal form games before briefly describing the additional features that arise in each of several types of more general games. The unifying principle of local (or neighborhood) superiority will emerge in the process.

2 . ESS for symmetric games

In a symmetric evolutionary game, there is a single set S of pure strategies available to the players, and the payoff to pure strategy e_i is a function π_i of the system's strategy distribution. In the following subsections we consider two-player symmetric games with S finite in normal and extensive forms (Sections 2.1 and 2.2 respectively) and with S a continuous set (Section 2.3).

2.1 . Normal form games

Let $S \equiv \{e_1, \ldots, e_n\}$ be the set of pure strategies. A player may also use a mixed strategy $p \in \Delta^n \equiv \{p = (p_1, \ldots, p_n) | \Sigma p_i = 1, p_i \geq 0\}$ where p_i is the proportion of the time this individual uses pure strategy e_i. Pure strategy e_i is identified with the ith unit vector in Δ^n. The population state is $\hat{p} \in \Delta^n$ whose components are the current frequencies of strategy use in the population (that is, the strategy distribution). We assume the expected payoff to p is the bilinear function $\pi(p, \hat{p}) = \Sigma_{i,j=1}^{n} p_i \pi(e_i, e_j) \hat{p}_j$ resulting from random two-player contests.

Suppose the resident population is monomorphic at p^* (that is, all members adopt strategy p^*) and a monomorphic sub-population of mutants using p appears in the system. These mutants will not invade if there is a positive *invasion barrier* $\varepsilon_0(p)$ (Bomze and Pötscher, 1989). That is, if the proportion ε of mutants in the system is less than $\varepsilon_0(p)$, then the mutants will eventually die out due to their lower replication rate. In mathematical terms, $\varepsilon = 0$ is a (locally) asymptotically stable rest point of the corresponding resident-mutant invasion dynamics. For invasion dynamics based on replication, Bomze and Pötscher show p^* is an ESS (that is, satisfies (1)) if and only if every $p \neq p^*$ has a positive invasion barrier.

Important and somewhat surprising consequences of an ESS p^* are its asymptotic stability for many evolutionary dynamics beyond these monomorphic resident systems invaded by a single type of mutant. For instance, p^* is asymptotically stable when simultaneously invaded by several types of mutants and when a polymorphic resident system consisting of several (mixed) strategy types whose average strategy is p^* is invaded (see the 'strong stability' concept developed in Cressman, 1992). In particular, p^* is asymptotically stable for the replicator equation (Taylor and Jonker, 1978; Hofbauer, Schuster and Sigmund, 1979; Zeeman, 1980)

$$\dot{p}_i = p_i(\pi(e_i, p) - \pi(p, p))$$

(2)

when each individual player is a pure strategist.

Games that have a completely mixed ESS (that is, p^* is in the interior of Δ^n) enjoy further dynamic stability properties since these games are *strictly stable* (that is,

$\pi(p-\hat{p},\ p-\hat{p})<0$ for all $p\neq\hat{p}$) (Sandholm, 2006). The ESS of a strictly stable game is also globally asymptotically stable for the best response dynamics (the continuous-time version of fictitious play) (Hofbauer and Sigmund, 1998) and for the Brown–von Neumann–Nash dynamics (related to Nash's, 1951, proof of existence of NE) (Hofbauer and Sigmund, 2003).

The preceding two paragraphs provide a strong argument that an ESS will be the ultimate outcome of the evolutionary adjustment process. The proofs of these results use two other equivalent characterizations of an ESS p^* of a symmetric normal form game; namely,

1. (a) p^* has a *uniform* invasion barrier (i.e. $\varepsilon_0(p)>0$ is independent of p)
2. (b) for all p sufficiently close (but not equal) to p^*

$$\pi(p,\ p)<\pi(p^*,\ p). \tag{3}$$

It is this last characterization, called ‘local superiority’ (Weibull, 1995), that proves so useful for other classes of games (see below). Heuristically, (3) suggests p^* will be asymptotically stable since there is an incentive to shift towards p^* whenever the system is slightly perturbed from p^*.

Unfortunately, there are many normal form games that have no ESS. These include most three-strategy games classified by Zeeman (1980) and Bomze (1995). No mixed strategy p^* can be an ESS of a symmetric zero-sum game (that is, $\pi(\hat{p},\ p)=-\pi(p,\ \hat{p})$ for all $p,\ \hat{p}\in\Delta^n$) since $\pi(p^*,\ p)=\pi(p^*-p,\ p)\leq 0=\pi(p,\ p)$ for all $p\in\Delta^n$ in some direction from p^*. Thus, the classic zero-sum Rock–Scissors–Paper Game in Table 1 has no ESS since its only NE $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ is interior. An early attempt to relax the ESS conditions to rectify this replaces the strict inequality in (1) (ii) by $\pi(p,\ p)\leq\pi(p^*,\ p)$. The NE p^* is then called a *neutrally stable strategy* (NSS) (Maynard Smith, 1982; Weibull, 1995). The only NE of the Rock–Scissors–Paper Game is a NSS.

The payoff matrix for the Rock–Scissors–Paper Game

$$\begin{array}{c} \text{Rock} \\ \text{Scissors} \\ \text{Paper} \end{array}\begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

Each entry is the payoff to the row player when column players are listed in the same order.

Also, the normal forms of most interesting extensive form games have no ESS, especially when NE outcomes do not specify choices off the equilibrium path and so correspond to NE components. In general, when NE are not isolated, the *ESSet* introduced by Thomas (1985) is more important. This is a set E of NSS so that (1) (ii) holds for all $p^*\in E$ and $p\notin E$. An ESSet is a finite union of disjoint NE components, each of which must be an ESSet in its own right. Each ESSet has setwise dynamic stability consequences analogous to an ESS (Cressman, 2003). The ES structure of a game refers to its collection of ESSs and ESSets.

There are then several classes of symmetric games that always have an ESSet. Every two-strategy game has an ESSet (Cressman, 2003) which generically (that is, unless $\pi(\hat{p},\ \hat{p})=\pi(p,\ \hat{p})$ for all $p,\ \hat{p}\in\Delta^2$) is a finite set of ESSs. All games with symmetric payoff function (that is, $\pi(\hat{p},\ p)=\pi(p,\ \hat{p})$ for all $p,\ \hat{p}\in\Delta^n$) have an ESSet corresponding to the set of local maxima of $\pi(p,p)$ which generically is a set of isolated ESSs). These are called partnership games (Hofbauer and Sigmund, 1998) or common interest games (Sandholm, 2006).

Symmetric games with payoff, $\pi_i(\hat{p})$, of pure strategy e_i nonlinear in the population state $\hat{p}$ are quite common in biology and in economics (Maynard Smith, 1982; Sandholm, 2006), where they are called playing-the-field models or population games. With $\pi(p,\ \hat{p})=\Sigma_i p_i\pi_i(\hat{p})$, nonlinearity implies (1) is a weaker condition than (3), as examples in Bomze and Pötscher (1989) show. Local superiority (3) is then taken as the operative definition of an ESS p^* (Hofbauer and Sigmund, 1998) and it is equivalent to the existence of a uniform

invasion barrier for p^*.

2.2 . Extensive form games

The application of ESS theory to finite extensive form games has been less successful (see Figure 1). Every ESS can have no other realization equivalent strategies in its normal form (van Damme, 1991) and so, in particular, must be *pervasive strategy* (that is, it must reach every information set when played against itself). To ease these problems, Selten (1983) defined a *direct ESS* in terms of behaviour strategies (that is, strategies that specify the local behaviour at each player information set) as a b^* that satisfies (1) for any other behaviour strategy b. He showed each such b^* is subgame perfect and arises from the backward induction technique applied to the ES structure of the subgames and their corresponding truncations.

Figure 1

The extensive form tree of the van Damme example. For the construction of the tree of a symmetric extensive form game, see Selten (1983) or van Damme (1991)

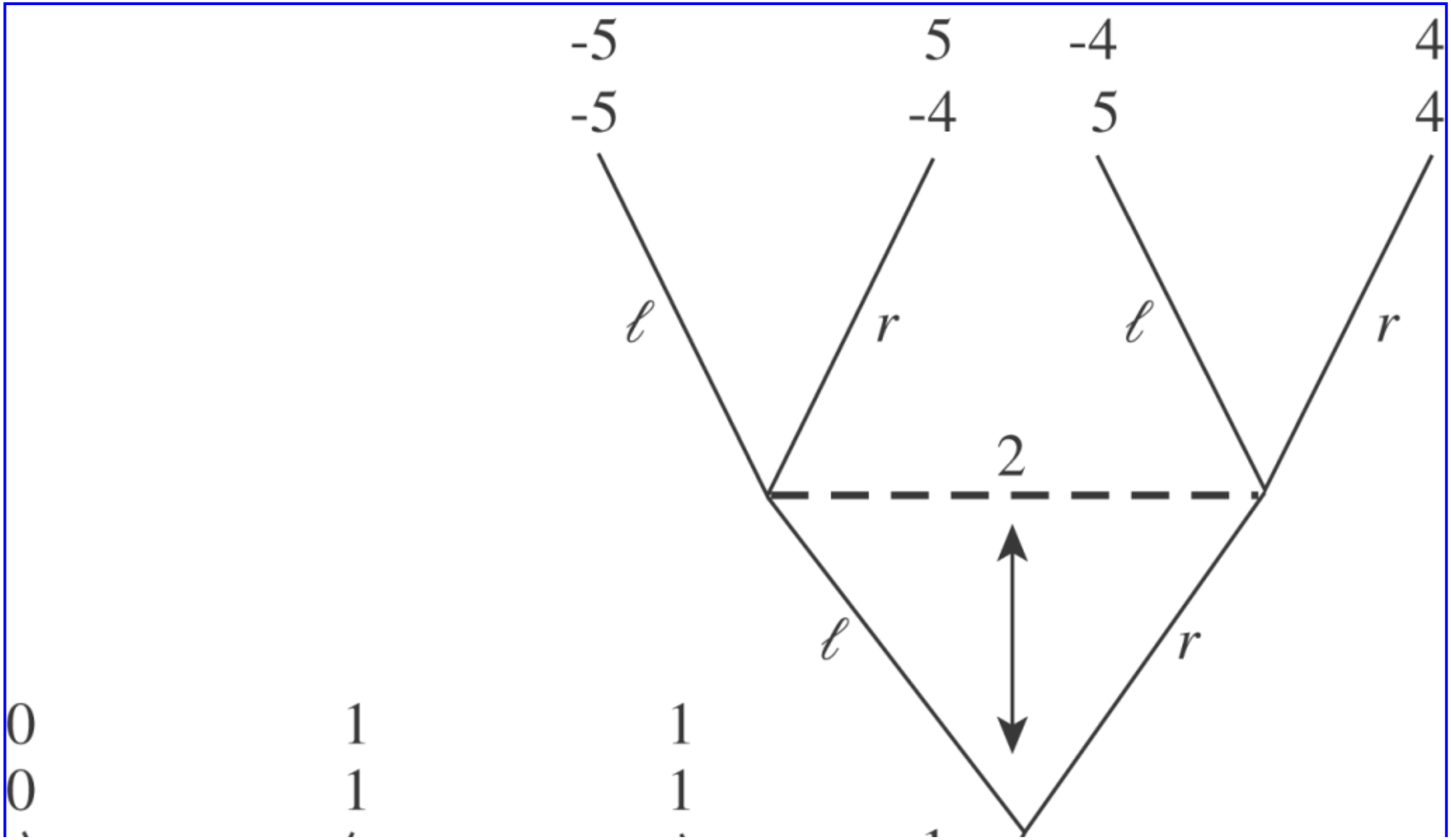

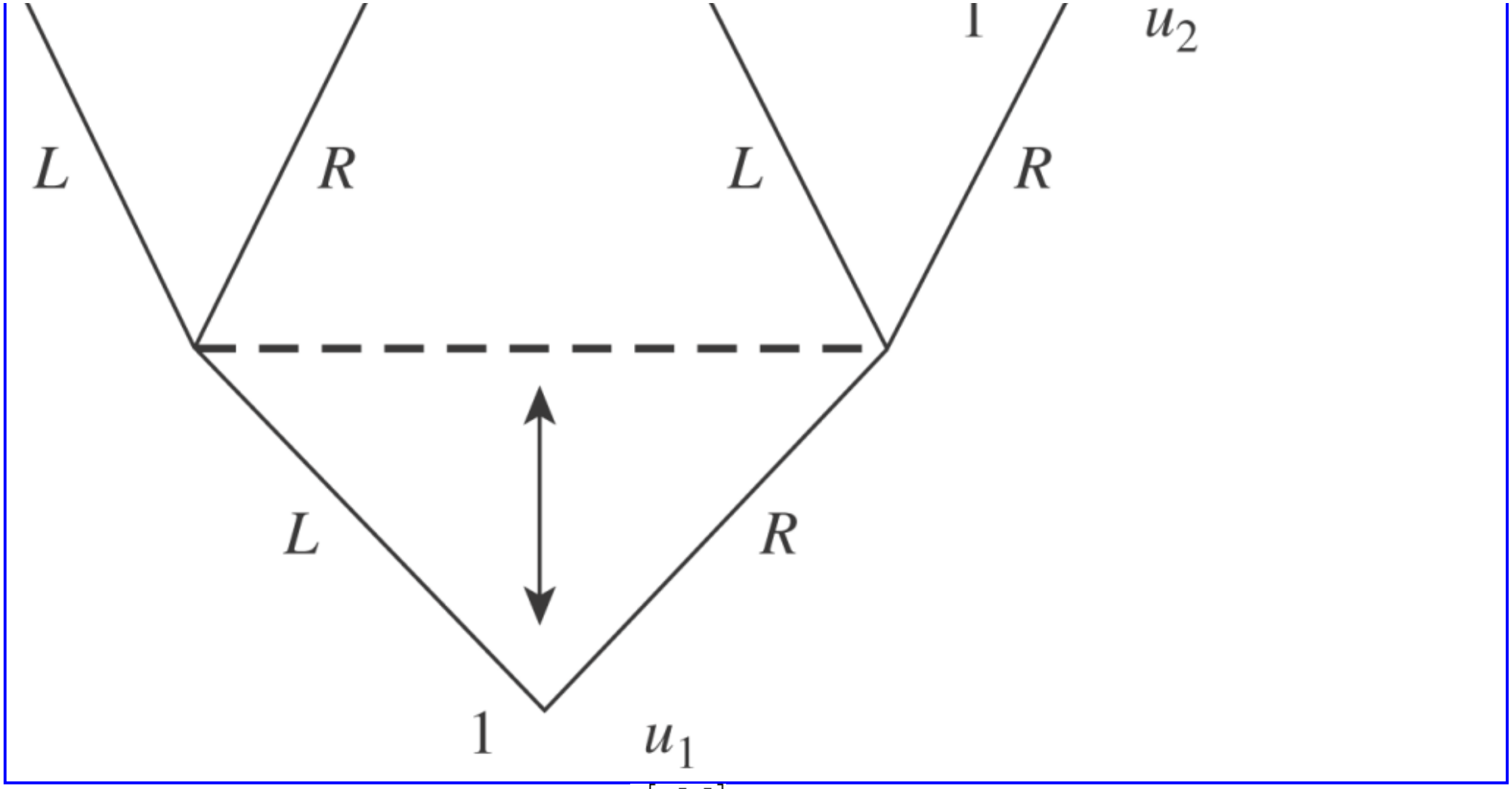

Consider backward induction applied to Figure 1. Its second-stage subgame $\begin{array}{c} \ell \\ r \end{array}\begin{bmatrix} -5 & 5 \\ -4 & 4 \end{bmatrix}$ has mixed ESS $b_2^* = \left(\frac{1}{2}, \frac{1}{2}\right)$ and, when the second decision point of player 1 is replaced by the payoff 0 from b_2^*, the truncated single-stage game $\begin{array}{c} L \\ R \end{array}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ also has a mixed ESS $b_1^* = \left(\frac{1}{2}, \frac{1}{2}\right)$. Since both stage games have a mixed ESS (and so a unique NE since they are strictly stable), (b_1^*, b_2^*) is the only NE of Figure 1 and it is pervasive. Surprisingly, this example has no direct ESS as Selten originally hoped since (b_1^*, b_2^*) can be invaded by the pure strategy that plays *Rr* (van Damme, 1991).

The same technique applied to Figure 1 with second-stage subgame replaced by $\begin{array}{c} \ell \\ r \end{array}\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$ yields $b_2^* = \left(\frac{1}{2}, \frac{1}{2}\right)$ and truncated single-stage game $\begin{array}{c} L \\ R \end{array}\begin{bmatrix} 0 & 1 \\ 1 & -1/2 \end{bmatrix}$ with $b_1^* = \left(\frac{3}{5}, \frac{2}{5}\right)$.

This is an example of a two-stage War of Attrition with base game $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ where a player remains (*R*) at the first stage in the hope the opponent will leave (*L*) but incurs a waiting cost of one payoff unit if both players remain. This (b_1^*, b_2^*) is a direct ESS since all *N*-stage War of Attrition games are strictly stable (Cressman, 2003).

The examples in the preceding two paragraphs show that, although backward induction determines candidates for the ES structure, it is not useful for determining which candidates

are actually direct ESSs. The situation is more discouraging for non-pervasive NE. For example, the only NE outcome of the two-stage repeated Prisoner's Dilemma game (Nachbar, 1992) with cumulative payoffs is mutual defection at each stage. This NE outcome cannot be an isolated behaviour strategy (that is, there is a corresponding NE component) and so there is no direct ESS. Worse, for typical single-stage payoffs such as $\begin{array}{c} \text{Defect} \\ \text{Cooperate} \end{array}\begin{bmatrix} -1 & 10 \\ -2 & 5 \end{bmatrix}$, this component does not satisfy setwise extensions of the ESS (for example, it is not an ESSet).

Characterization of NE found by backward induction with respect to dynamically stable rest points of the subgames and their truncations shows more promise. Each direct ESS b^* yields an ESSet in the game's normal form (Cressman, 2003) and so is dynamically stable. Furthermore, for the class of simultaneity games where both players know all player actions at earlier stages, Cressman shows that, if b^* is a pervasive NE, then it is asymptotically stable with respect to the replicator equation if and only if it comes from this backward induction process. In particular, the NE for Figure 1 and for the *N*-stage War of Attrition are (globally) asymptotically stable. Although the subgame perfect NE for the *N*-stage Prisoner's Dilemma game that defects at each decision point is not asymptotically stable, the eventual outcome of evolution is in the NE component (Nachbar, 1992; Cressman, 2003).

2.3 . Continuous strategy space

Evolutionary game theory for symmetric games with a continuous set of pure strategies *S* has been slower to develop. Most recent work examines static payoff comparisons that predict an $x^* \in S$ is the evolutionary outcome. There are now fundamental differences between the ESS notion (1) and that of local superiority (3) as well as between invasion by monomorphic mutant sub-populations and the polymorphic model of the replicator equation. Here, we illustrate these differences when *S* is a subinterval of real numbers and $\pi(x, y)$ is a continuous payoff function of $x, y \in S$.

First, consider an $x^* \in S$ that satisfies (3). In particular,

$$\pi(x, x) < \pi(x^*, x) \tag{4}$$

for all $x \in S$ sufficiently close (but not equal) to x^*. This is the *neighbourhood invader strategy* (NIS) condition of Apaloo (1997) that states x^* can invade any nearby monomorphism *x*. On the other hand, from (1), x^* cannot be invaded by these *x* if it is a *neighbourhood strict NE*, that is

$$\pi(x, x^*) < \pi(x^*, x^*) \tag{5}$$

for any other *x* sufficiently close to x^*. Inequalities (4) (5) are independent of each other and combine to assert that x^* strictly dominates *x* in all these two-strategy games $\{x^*, x\}$.

In the polymorphic model, populations are described by a *P* in the infinite dimensional set $\Delta(S)$ of probability distributions with support in *S*. When the expected payoff $\pi(x, P)$ is given through random pairwise contests, Cressman (2005) shows that strict domination implies x^* is *neighbourhood superior* (that is,

$$\pi(x^*, P) > \pi(P, P) \tag{6}$$

for all other $P \in \Delta(S)$ with support sufficiently close to x^*) and conversely, neighbourhood superiority implies weak domination. Furthermore, a neighborhood superior monomorphic population x^* (that is, the Dirac delta probability distribution δ_{x^*}) is asymptotically stable for all initial P with support sufficiently close to x^* (and containing x^*) under the replicator equation. This is now a dynamic on $\Delta(S)$ (Oechssler and Riedel, 2002) that models the evolution of the population distribution.

In the monomorphic model, the population is a monomorphism $x(t) \in S$ at all times. If a nearby mutant strategy $y \in S$ can invade x, the whole population is shifted in this direction. This intuition led Eshel (1983) to define a *continuously stable strategy* (CSS) as a neighbourhood strict NE x^* that satisfies, for all x sufficiently close to x^*,

$$\pi(y, x) > \pi(x, x) \tag{7}$$

for all y between x^* and x that are sufficiently close to x. Later, Dieckmann and Law (1996) developed the canonical equation of adaptive dynamics to model the evolution of this monomorphism and showed a neighbourhood strict NE x^* is a CSS if and only if it is an asymptotically stable rest point. Cressman (2005) shows x^* is a CSS if and only if it is *neighbourhood half-superior* (that is, there is a uniform invasion barrier of at least $\frac{1}{2}$ in the two-strategy games $\{x^*, x\}$) (see also the half-dominant concept of Morris, Rob and Shin, 1995).

For example, take $S = \mathbf{R}$ and payoff function

$$\pi(x, y) = -x^2 + bxy \tag{8}$$

that has strict NE $x^* = 0$ for all values of the fixed parameter b. x^* is a NIS (CSS) if and only if $b < 1$ $(b < 2)$ (Cressman and Hofbauer, 2005). Thus, there are strict NE when $b > 2$ that are not ‘evolutionarily stable’.

3 Asymmetric games

Following Selten (1980) and van Damme (1991), in a two-player asymmetric game with two roles (or species), pairwise contests may involve players in the same or in opposite roles. First, consider ESS theory when there is a finite set of pure strategies $S = \{e_1, \ldots, e_n\}$ and $T = \{f_1, \ldots, f_m\}$ for players in role 1 and 2 respectively. Assume payoff to a mixed strategist is given by a bilinear payoff function and let $\pi_1(p; \hat{p}, \hat{q})$ be the payoff to a player in role one using $p \in \Delta^n$ when the current state of the population in roles 1 and 2 are $\hat{p}$ and $\hat{q}$ respectively. Similarly, $\pi_2(q; \hat{p}, \hat{q})$ is the payoff to a player in role 2 using $q \in \Delta^m$. For a discussion of resident-mutant invasion dynamics, see Cressman (1992), who shows the monomorphism (p^*, q^*) is uninvadable by any other mutant pair (p, q) if and only if it is a *two-species ESS*, that is, for all (p, q) sufficiently close (but not equal) to (p^*, q^*),

$$\text{either } \pi_1(p; p, q) < \pi_1(p^*; p, q) \text{ or } \pi_2(q; p, q) < \pi_2(q^*; p, q). \tag{9}$$

The ESS condition (9) is the two-role version of local superiority (3) and has an equivalent formulation analogous to (1) (Cressman, 1992). This ESS also enjoys similar stability properties to the ESS of Subsection 2.1 such as its asymptotic stability under the (two-species) replicator equation (Cressman, 1992; 2003).

A particularly important class of asymmetric games consists of truly asymmetric games that have no contests between players in the same role (that is, there are no intraspecific contests). These are bimatrix games (that is, given by an $n \times m$ matrix whose *ij*th entry is the pair of payoffs $(\pi_1(e_i, f_j), \pi_2(e_i, f_j))$ for the interspecific contest between e_i and f_j). The ESS concept is now quite restrictive since Selten (1980) showed that (p^*, q^*) satisfies (9) if and only if it is a strict NE. This is also equivalent to asymptotic stability under the (two-species) replicator equation (Cressman, 2003). Standard examples (Cressman, 2003), with two strategies for each player include the Buyer–Seller Game that has no ESS since its only NE is in the interior. Another is the Owner–Intruder Game that has two strict NE Maynard Smith (1982) called the bourgeois ESS where the owners defend their territory and the paradoxical ESS where owners retreat.

Asymmetric games with continuous sets of strategies have recently received a great deal of attention (Leimar, 2006). For a discussion of neighbourhood (half) superiority conditions that generalize (6) and (7) to two-role truly asymmetric games with continuous payoff functions, see Cressman (2005). He also shows how these conditions are related to NIS and CSS concepts based on (9) and to equilibrium selection results for games with discontinuous payoff functions such as the Nash Demand Game (Binmore, Samuelson and Young, 2003).

See Also

- deterministic evolutionary dynamics
- learning and evolution in games: an overview

Bibliography

Apaloo, J. 1997. Revisiting strategic models of evolution: the concept of neighborhood invader strategies. *Theoretical Population Biology* 52, 71–7.

Binmore, K., Samuelson, L. and Young, P. 2003. Equilibrium selection in bargaining models. *Games and Economic Behavior* 45, 296–328.

Bomze, I. 1995. Lotka–Volterra equation and replicator dynamics: new issues in classification. *Biological Cybernetics* 72, 447–53.

Bomze, I. and Pötscher, B. 1989. *Game Theoretical Foundations of Evolutionary Stability*. Lecture notes in economics and mathematical systems 324. Berlin: Springer-Verlag.

Cressman, R. 1992. *The Stability Concept of Evolutionary Games (A Dynamic Approach)*. Lecture notes in biomathematics 94. Berlin: Springer-Verlag.

Cressman, R. 2003. *Evolutionary Dynamics and Extensive Form Games*. Cambridge, MA: MIT Press.

Cressman, R. 2005. Continuously stable strategies, neighborhood superiority and two-player games with continuous strategy space. Mimeo.

Cressman, R. and Hofbauer, J. 2005. Measure dynamics on a one-dimensional continuous trait space: theoretical foundations for adaptive dynamics. *Theoretical Population Biology* 67, 47–59.

Dieckmann, U. and Law, R. 1996. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of Mathematical Biology* 34, 579–612.

Eshel, I. 1983. Evolutionary and continuous stability. *Journal of Theoretical Biology* 103, 99–111.

Hofbauer, J., Schuster, P. and Sigmund, K. 1979. A note on evolutionarily stable strategies and game dynamics. *Journal of Theoretical Biology* 81, 609–12.

Hofbauer, J. and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Hofbauer, J. and Sigmund, K. 2003. Evolutionary game dynamics. *Bulletin of the American Mathematical Society* 40, 479–519.

Leimar, O. 2006. Multidimensional convergence stability and the canonical adaptive dynamics. In *Elements of Adaptive Dynamics*, ed. U. Dieckmann and J. Metz. Cambridge University Press.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, J. and Price, G. 1973. The logic of animal conflicts. *Nature* 246, 15–18.

Morris, S., Rob, R. and Shin, H. 1995. Dominance and belief potential. *Econometrica* 63, 145–57.

Nachbar, J. 1992. Evolution in the finitely repeated Prisoner's Dilemma. *Journal of Economic Behavior and Organization* 19, 307–26.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

Oechssler, J. and Riedel, F. 2002. On the dynamic foundation of evolutionary stability in continuous models. *Journal of Economic Theory* 107, 223–52.

Sandholm, W. 2006. *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press.

Selten, R. 1980. A note on evolutionarily stable strategies in asymmetrical animal contests. *Journal of Theoretical Biology* 84, 93–101.

Selten, R. 1983. Evolutionary stability in extensive two-person games. *Mathematical Social Sciences* 5, 269–363.

Taylor, P. and Jonker, L. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–156.

Thomas, B. 1985. On evolutionarily stable sets. *Journal of Mathematical Biology* 22, 105–15.

van Damme, E. 1991. *Stability and Perfection of Nash Equilibria*, 2nd edn. Berlin: Springer-Verlag.

Weibull, J. 1995. *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

Zeeman, E. 1980. Population dynamics from game theory. In *Global Theory of Dynamical Systems*, ed. Z. Nitecki and C. Robinson. Lecture notes in mathematics 819. Berlin: Springer.

How to cite this article

Cressman, Ross. "learning and evolution in games: ESS." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000230> doi:10.1057/9780230226203.0951

The New Palgrave Dictionary of Economics Online

learning and information aggregation in networks

Douglas Gale and Shachar Kariv

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Social learning’ is a process whereby economic agents learn by observing the behaviour of others. ‘Social learning in networks’ requires sophistication because individuals draw inferences from the behaviour of agents they cannot directly observe. Theoretical research suggests that, even if networks are very incomplete, social learning leads to uniform behaviour. Experimental evidence suggests that learning in networks conforms quite well to theoretical predictions. It also illustrates how the network architecture influences the pattern of learning and the efficiency of information aggregation.

Keywords

Bala–Goyal model; bounded rationality; circle network; complete network; connected graph; directed graph; Griliches, Z.; herd behavior; hubs; imitation principle; information aggregation; informational cascades; perfect information; pure information externality; scale-free networks; social experimentation; social learning; star network

Article

‘Social learning’ is a process whereby economic agents learn by observing the actions (but not the payoffs) of others; ‘social learning in networks’ applies this idea to situations in which individuals observe the other individuals to whom they are connected in a social network.

Griliches (1957) first studied the gradual adoption of corn planted with hybrid seed in the USA, a new agricultural technique, from the early 1930s to mid-1950s. He observed that at first farmers learned from salespersons; later they learned from their neighbours. The result was an S-shaped time profile of adoption. A number of recent papers, including Foster and Rosenzweig (1995), Conley and Udry (2001), Kremer and Miguel (2003) and Munshi (2004) examine how agents in developing countries learn from their social contacts when deciding whether to adopt new technologies.

The classical model of social learning, first studied by Banerjee (1992) and Bikhchandani, Hirshleifer

and Welch (1992), and extended by Smith and Sørensen (2000), assumes a *pure information externality*. An agent's payoff $u(a,w)$ depends only on his own action a and an unknown state of nature w. Each agent i has private information about the state and his choice of action a will reflect that information. By observing an agent's action, it is possible to learn something about his information and make a better decision. The problem is that agents may rationally ignore their own information and ‘follow the herd’, that is, imitate the actions they see others choose. So-called *herd behaviour* and *informational cascades* can arise very rapidly, before much information has been revealed, and often result in inefficient choices. A number of experimental studies replicate herd behaviour in the laboratory.

The classical models assume that agents make decisions sequentially and observe the action chosen by each of their predecessors. In reality, individuals are bound together by a *social network*, the complex of relationships that brings them into contact with other agents, such as neighbours, co-workers, family, and so on. A specific framework, introduced by Gale and Kariv (2003), henceforth GK, assumes that individuals are bound together by a social network and can observe the agents to whom they are connected only through the network. The social network is represented by a *directed graph* in which nodes correspond to agents and agent i can observe agent j if there is an edge leading from node i to node j. In order to model the diffusion of information through the network, GK assume that agents choose actions simultaneously and revise their decisions as new information is received. More precisely, an agent whose current information is I chooses an action a to maximize his short-run payoff $E[u(a, w)|I]$. GK rationalize non-strategic behaviour by assuming there is a large number of agents of each type, so a single agent's decision has no impact on the future play of the game.

An agent's beliefs can be represented by a random sequence of probability distributions $P_t(\omega)$. At date t, an agent derives a posterior $P_{t+1}(w)$ from the prior $P_t(w)$ and the new information received. These beliefs satisfy the martingale property $E[P_{t+1}|I_t] = P_t$, and the martingale convergence theorem implies that these beliefs converge to a constant with probability one. The limiting beliefs are not necessarily uniform (different agents may have different beliefs) and need not be fully revealing. However, in *connected* networks, where every agent is connected directly or indirectly with every other agent, the initial diversity of actions is eventually replaced by uniformity. More precisely, except in cases of indifference, agents will choose the same action. This is the network-learning analogue of the herd behaviour found in the classical social learning model. The proof of uniformity makes use of the *imitation principle*. If agent i can observe the actions of agent j, agent i must be able to do as well as j on average (because one feasible strategy is to choose the same action j). In a connected network, all agents get the same payoff on average and this implies that they choose different actions only if they are indifferent.

Learning in a network is ‘simply’ a matter of Bayesian updating, but a rational agent must take account of the network architecture in order to update correctly. For example, suppose there are three (types of) agents, A, B, and C, arranged in a circle: A observes B, B observes C, and C observes A. At the first decision, A has not yet had a chance to observe B, so he makes his decision based on his private information. Before the second decision, A observes B's first decision and uses it to update his beliefs about the true state of nature. Before the third decision, A observes B's second decision and realizes that any change from the first must be based on B's observation of C's first decision. So now A can make some inference about C's private information and update his beliefs accordingly. This learning can go on for some time. Eventually, A may observe changes in B's action that were prompted by changes in C's

action that were prompted by *C*'s observation of *A*. Even this is informative because it reveals how strong *C*'s information is relative to *A*'s. In any case, exploiting fully the information revealed in a network requires agents to consider not only what they observe, but also what their neighbours observe, what their neighbours' neighbors observe, and so on. The chains of inferences that rational individuals make naturally involve hierarchies of beliefs, that is, beliefs about a neighbour's beliefs about a neighbour's beliefs about ….

The complexity of Bayesian learning in networks has led some authors to suggest that models of *bounded rationality* are more appropriate for describing learning in networks. Bala and Goyal (1998) examine the decisions of boundedly rational agents, who try to extract information from the actions and payoffs of the agents they observe, but without taking account of the fact that those agents also observe other agents. Hence, there is private information in the Bala–Goyal model, but agents are assumed to ignore it. In the Bala–Goyal model, at each date, an agent chooses one of several available actions with unknown payoff distributions. Agents can observe the actions and payoffs of their neighbours (those to whom they are directly connected by the network) and use this information to update their beliefs about the payoff distribution. Thus, agents learn by observing the outcome (payoff) of an experiment (choice of action) rather than by inferring another agent's private information from his action. This is a model of *social experimentation*, in the sense that it generalizes the problem of a single agent experimenting with a multi-armed bandit to a social setting, rather than social learning. A model of social experimentation is quite different from a model of social learning because there is an informational externality but there is no informational asymmetry. As with Bayesian learning, boundedly rational learning implies convergence of beliefs and uniformity of actions in the limit.

Laboratory experiments provide the cleanest test for the theory since subjects' neighbourhoods and private information can be controlled. Choi, Gale and Kariv (2004; 2005) describe the results of an experimental investigation of learning in networks based on the model of GK. The experiments involve three-person, connected social networks. The experimental design uses three representative networks: the *complete network*, in which each agent can observe the actions chosen by the other agents; the *star network*, in which one agent, the centre, can observe the actions of the other two peripheral agents, and the peripheral agents can observe only the centre; and the *circle network*, in which each agent can observe only one other agent and each agent is observed by one other agent. Despite the small number of players in each game, it can be shown that myopic payoff maximization is rational: there is no gain to strategic behaviour. Nonetheless, larger-scale experiments might be informative.

The experimental data from these studies exhibit a strong tendency toward herd behaviour, but despite this tendency the efficiency of information aggregation is quite good. Although convergence to a uniform action is quite rapid, frequently occurring within two to three turns, there are significant differences between the behaviour of different networks. Most herds entail correct decisions, which is consistent with the predictions of the parametric model underlying the experimental design. Comparing the behaviour of different individuals indicates that there is indeed high variation in individual behaviour across subjects, but the error rates (the proportion of times a subject deviates from the best response) are uniformly fairly low.

These results suggest that the theory adequately accounts for large-scale features of the data, but in some situations the theory does less well in accounting for subjects' behaviour. It is likely that the theory fails in those situations because the complexity of the decision problem exceeds the bounded rationality of the subjects. Clearly, because of the lack of common knowledge in the networks, the decision problems

faced by subjects require quite sophisticated reasoning. Subjects' success or failure in the experiment results from the appropriateness of the heuristics they use as much as the inherent difficulty of the decision-making. Thus, an important subject for future research is to identify 'black spots' where the theory does least well in interpreting the data and ask whether additional 'behavioural' explanations might be needed to account for the subjects' behaviour.

Many important questions about social learning in networks remain to be explored. While small networks can be very insightful, especially in experimental contexts, the development of the theory depends on properties of networks that can be generalized. The recent discovery of Barabási and Albert (1999) that many networks are *scale-free*, in the sense that a few nodes are *hubs*, which have a very large number of links to other nodes whereas most nodes have just a few, has significant implications for the efficiency of information aggregation. Once information reaches a hub it passes to numerous other nodes and spreads rapidly throughout the entire population, but if the hub's information is of poor quality its disproportionate influence becomes a disadvantage. Thus, the impact of hubs on the efficiency of information aggregation is not clear. Perhaps the most important subject for future research is to identify the impact of network architecture on the efficiency and dynamics of social learning. Progress in this area requires both new theory and new experimental data.

See Also

- behavioural game theory;
- experimental economics, history of;
- Griliches, Zvi;
- information cascades;
- learning and evolution in games: an overview;
- logit models of individual choice;
- network formation.

Bibliography

Bala, V. and Goyal, S. 1998. Learning from neighbors. *Review of Economic Studies* 65, 595–621.

Banerjee, A. 1992. A simple model of herd behavior. *Quarterly Journal of Economics* 107, 797–817.

Barabási, A. and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509–12.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1992. A theory of fads, fashion, custom, and cultural change as informational cascade. *Journal of Political Economy* 100, 992–1026.

Choi, S., Gale, D. and Kariv, S. 2004. Learning in networks: an experimental study. Mimeo. New York: Center for Experimental Social Science (CESS), New York University.

Choi, S., Gale, D. and Kariv, S. 2005. Behavioral aspects of learning in social networks: an experimental study. *Experimental and Behavioral Economics Advances in Applied Microeconomics*, vol. 13, ed. J. Morgan. Oxford: JAI Press.

Conley, T. and Udry, C. 2001. Social learning through networks: the adoption of new agricultural technologies in Ghana. *American Journal of Agricultural Economics* 83, 668–73.

Foster, A. and Rosenzweig, M. 1995. Learning by doing and learning from others: human capital and technical change in agriculture. *Journal of Political Economy* 103, 1176–209.

Gale, D. and Kariv, S. 2003. Bayesian learning in social networks. *Games and Economic Behavior* 45, 329–46.

Griliches, Z. 1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25, 501–22.

Kremer, M. and Miguel, T. 2003. Networks, social learning, and technology adoption: the case of deworming drugs in Kenya. Mimeo. Berkeley: Department of Economics, University of California.

Munshi, K. 2004. Social learning in a heterogeneous population: technology diffusion in the Indian Green Revolution. *Journal of Development Economics* 73, 185–213.

Smith, L. and Sørensen, P. 2000. Pathological outcomes of observational learning. *Econometrica* 68, 371–98.

How to cite this article

Gale, Douglas and Shachar Kariv. "learning and information aggregation in networks." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000218> doi:10.1057/9780230226203.0952

The New Palgrave Dictionary of Economics Online

learning in macroeconomics

George W. Evans and Seppo Honkapohja

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Expectations play a key role in macroeconomics. The assumption of rational expectations has been recently relaxed by explicit models of forecasting and model updating. Rational expectations can be assessed for stability under various types of learning, with least squares learning playing a prominent role. In addition to assessing the plausibility of an equilibrium, learning also provides a selection criterion when there are multiple equilibria. Monetary policy should be designed to avoid instability under learning and to facilitate coordination on desirable equilibria. Learning can also help to explain macroeconomic fluctuations as arising through either instabilities, stable indeterminacies or persistent learning dynamics.

Keywords

actual law of motion; adaptive learning; animal spirits; asset pricing; bounded rationality; Cagan model; calibration; constant gain learning; eductive learning; expectational stability; expectations; game theory; hyperinflation; imperfect knowledge; increasing social returns; indeterminacy of equilibrium; interest-rate rule; learning in macroeconomics; liquidity trap; monetary policy; multiple equilibria; New Keynesian macroeconomics; overlapping contracts model; overlapping generations model; perceived law of motion; Phillips curve; rational expectations; rational expectations equilibrium; rational learning; real business cycles; recursive least squares; seigniorage; stationary sunspot equilibria; sunspot equilibrium; Taylor rule; uniqueness of equilibrium; vector autoregression

Article

Learning in macroeconomics refers to models of expectation formation in which agents revise their forecast rules over time, for example in response to new data. Expectations of future income, prices and sales play key roles in theories of saving and investment. Many other examples of the central role of expectations could be given.

1 Introduction

The current standard methodology for modelling expectations is to assume that the economy is in a rational expectations equilibrium (REE). REE is a model-consistent equilibrium in the two-way relationship between the influence of expectations on the economy and the dependence of expectations on the time path of the economy.

The standard formulation of REE makes strong assumptions on the information of economic agents. The true stochastic process of the economy is assumed known, with unforecastable random shocks constituting the remaining uncertainty. This assumption presupposes that the economic agents know much more than, say, the economists who in practice do not know the true stochastic structure and instead must estimate its parameters.

Recently, macroeconomic theory has been moving beyond the strict rational expectations (RE) hypothesis. Explicit models of imperfect knowledge and associated learning processes have been developed. In models of learning economic agents try to improve their knowledge of the stochastic process of the economy over time as new information becomes available.

Different approaches to modelling learning behaviour have been employed. Perhaps the most common has been ‘adaptive learning’, which views economic agents as econometricians who estimate the parameters of their model and make forecasts using their estimates. In adaptive learning economic agents have limited common knowledge since they estimate their own perceived laws of motion.

A second approach, called ‘eductive learning’, assumes common knowledge of rationality: economic agents engage in a process of reasoning about the possible outcomes knowing that other agents engage in the same process. Eductive learning takes place in logical time. A third approach has been ‘rational learning’, which employs a Bayesian viewpoint. Full knowledge of economic parameters is then replaced by priors and Bayesian updating under a correctly specified model, including common knowledge that all agents share this knowledge. Rational learning thus retains a form of REE at each point of time.

Basic theories of learning were developed largely in the 1980s and 1990s. See Sargent (1993, 1999), Evans and Honkapohja (2001), Guesnerie (2005) and Beck and Wieland (2002) for references. Recently, models of learning have been applied to issues of macroeconomic, and especially monetary, policy. In this overview, we focus on adaptive learning as it has been the most widely used approach. (For references to the pre-2001 literature, see Evans and Honkapohja, 2001.)

2 Least squares learning

In adaptive learning it is commonly assumed that agents estimate their model of the dynamics of economic variables, called the *perceived law of motion* (PLM), by *recursive least squares* (RLS), arguably the most common estimation method in econometrics.

2.1 Overview

We illustrate the key concepts using the Cagan model of the price level

$\hat{m} - p_t = -\psi(p^e_{t+1} - p_t) + \phi' w_t + \varepsilon_t$, where p_t and $\hat{m}$ are logarithms of the price level and (constant) nominal money supply. Here $\psi > 0$ and p^e_{t+1} denotes the expectations of p_{t+1} formed at time t. w_t is a vector of observable exogenous variables, assumed to follow a stationary vector autoregression (VAR) process $w_t = Fw_{t-1} + e_t$, in which F is taken as known for simplicity. ε_t is an unobservable i.i.d. shock.

The reduced form of the Cagan model is

$$p_t = \alpha_0 + \alpha_1 p^e_{t+1} + \beta' w_t + v_t, \tag{1}$$

where $v_t = -(1+\psi)^{-1}\varepsilon_t$ and α_0, α_1 and β depend on $\hat{m}$, ψ and φ. The model has a unique REE of the form $p_t = \bar{a} + \bar{b}' w_t + v_t$, where $\bar{a} = (1-\alpha_1)^{-1}\alpha_0$, $\bar{b} = (I - \alpha_1 F')^{-1}\beta$.

Agents are assumed to use the PLM $p_t = a + b' w_t + \eta_t$, where η_t is a disturbance term. The PLM has the same functional form as the REE but possibly different coefficients since agents do not know the REE. To estimate the PLM, agents use data $\{p_i, w_i\}_{i=0}^{t-1}$ and forecast using the estimated model $E^*_t p_{t+1} = a_{t-1} + b'_{t-1} F w_t$.

These forecasts lead to a temporary equilibrium or *actual law of motion* (ALM) $p_t = T(\varphi_{t-1})' z_t + v_t$, where $T(\varphi)' = (\alpha_0 + \alpha_1 a, \alpha_1 b' F + \beta')$. The REE $(\bar{a}, \bar{b}')$ is a fixed point of the mapping $T(\phi)$ from the PLM to the ALM. If we let $\varphi'_t = (a_t, b'_t)$ and $z'_t = (1, w'_t)$, RLS estimation is given by

$$\varphi_t = \varphi_{t-1} + t^{-1} R_t^{-1} z_t (p_t - \varphi'_{t-1} z_t) R_t = R_{t-1} + t^{-1}(z_t z'_t - R_{t-1}). \tag{2}$$

where p_t is given by the ALM. We say that the REE is *stable under RLS learning* if $(a_{t-1}, b'_{t-1}) \to (\bar{a}, \bar{b}')$ over time.

This model of learning involves bounded rationality. Each period agents maximize their objective, given their forecasts. However, agents treat the economy as having constant parameters, which is true only in the REE. Outside the REE the PLMs are misspecified, but misspecification vanishes as learning converges to the REE.

A key result, which holds in numerous models, is that RLS learning converges to RE under certain

conditions on model parameters. Thus, the REE can be learned even though economic agents initially have limited knowledge and are boundedly rational.

Expectational stability (*E-stability*) is a convenient way for establishing the convergence conditions for RLS learning. Define the differential equation $d\varphi / d\tau = T(\varphi) - \varphi$, which describes partial adjustment in virtual time τ. The REE is *E-stable* if it is locally stable under the differential equation. For models of the form (1), convergence is guaranteed if $0 < \alpha_1 < 1$, which is satisfied in the Cagan model since $\alpha_1 = \psi(1 + \psi)^{-1}$. Evans and Honkapohja (2001) contains a detailed discussion of convergence of RLS learning.

2.2 The roles of learning

Adaptive learning has several other important roles besides being a stability theory for REE. RE models can have multiple stationary equilibria, that is, *indeterminacy of equilibrium*. In such situations learning stability acts as a *selection criterion* to determine the plausibility of a particular REE.

As an example consider the non-stochastic Cagan model with government spending financed by seigniorage, with nonlinear reduced form $x_t = G(x^e_{t+1})$, where x_t denotes inflation (see Evans and Honkapohja, 2001, chs. 11 and 12, for details). This model has two (interior) steady state solutions $\hat{x} = G(\hat{x})$. The low-inflation steady state x_L is stable under learning and the high-inflation steady state x_H is not. Learning selects a unique REE x_L in this model. In more general models, learning stability does not necessarily select a unique REE, but the set of ‘plausible’ REE is usually significantly smaller than the set of all REE.

The roles of RLS learning are not restricted to stability of REE and equilibrium selection. Learning can also provide new forms of dynamics as discussed below.

3 Monetary policy design

Indeterminacy of equilibria and instability of REE under RLS learning mean that the economy can be subject to persistent fluctuations. These instabilities can arise in the New Keynesian (NK) model (Woodford, 2003), which is widely used for studying monetary policy. Policy design has an important role in eliminating these instabilities and facilitating convergence to ‘desirable’ equilibria.

Consider the linearized NK model. The IS and PC curves $x_t = -\phi(i_t - E^*_t \pi_{t+1}) + E^*_t x_{t+1} + g_t$ and $\pi_t = \lambda x_t + \beta E^*_t \pi_{t+1} + u_t$ summarize private sector behaviour. Here x_t, π_t and i_t denote the output gap, inflation and the nominal interest rate. φ and λ are positive parameters while $0 < \beta < 1$ is the discount factor. The shocks g_t and u_t are assumed to be observable and follow a known *VAR*(1) process.

Central bank (CB) behaviour is described by an interest-rate rule. CB may use an instrument rule that is not based on explicit optimization. Examples are Taylor rules that depend on current data or forecasts, $i_t = \chi_\pi \pi_t + \chi_x x_t$ or $i_t = \chi_\pi E^*_t \pi_{t+1} + \chi_x E^*_t x_{t+1}$, where $\chi_\pi, \chi_x > 0$.

The IS and PC equations, together with either Taylor rule, lead to a bivariate reduced form in (x_t, π_t),

which can be examined for determinacy (uniqueness of equilibrium) and E-stability. Bullard and Mitra (2002) show that current-data Taylor rules yield both E-stability and determinacy iff $\lambda(\chi_\pi-1)+(1-\beta)\chi_x>0$. Under forward-looking rules $\chi_\pi>1$ and small χ_x yield E-stability and determinacy.

Optimal monetary policy under discretion and commitment has been examined by Evans and Honkapohja (2003a; 2003b; 2006). Various ways to implement optimal policy have been suggested. Some commonly suggested interest-rate rules, based on fundamental shocks and variables, can lead to E-instability and/or indeterminacy. Evans and Honkapohja advocate appropriate expectations-based rules that deliver both E-stability and determinacy.

Other aspects of learning are also important for monetary policy. One practical concern is the observability of private forecasts needed for forecast-based rules. Results by Honkapohja and Mitra (2005) show that using internal CB forecasts in place of private sector expectations normally delivers E-stability.

Another difficulty for optimal monetary policy is that it requires knowledge of structural parameters, which are in practice unknown. CB can learn the values of φ and λ by estimating IS and PC equations. Expectations-based optimal rules continue to deliver stability under simultaneous learning by private agents and the CB (see Evans and Honkapohja, 2003a; 2003b).

4 Fluctuations

A major issue in macroeconomics is economic fluctuations, for example, business cycles and asset price movements. Can learning help to explain these phenomena?

4.1 Stable sunspot fluctuations

One theory of macroeconomic fluctuations interprets them as rational ‘sunspot’ equilibria. Although many macroeconomic models – for example, the real business cycle (RBC) model or Taylor's overlapping contracts model – have a unique stationary solution under RE, other models can have indeterminacy. Examples include the overlapping generations (OLG) model and RBC models with increasing returns and monopolistic competition or tax distortions.

When multiple equilibria are present, some solutions may depend on variables, ‘sunspots’, that are completely extraneous to the economy. Such stationary sunspot equilibria (SSEs) exhibit self-fulfilling prophecies with the sunspot acting as a coordinating device: if expectations depend on a sunspot variable, then the actual economy, since it depends on expectations, can also depend rationally on the sunspot.

As already noted, learning stability is a selection device. Suppose agents’ forecasts are a linear function of both the macroeconomic state and a sunspot variable. If the forecast functions have coefficients close to but not equal to SSE values, and if agents update the estimated coefficients using RLS, can the coefficients converge to SSE values? If not, this casts doubt on the plausibility of SSEs.

SSEs appear not to be stable under learning in indeterminate RBC models but are learnable in some other models. We first describe results for the NK model and then discuss the possibility of stable SSE in other models.

4.1.1 SSEs in the NK model

Consider again the linearized NK model augmented by either the current-data or forward-looking Taylor rule. As noted above, indeterminacy is likely when the 'Taylor principle' $\chi_\pi > 1$ is violated.
In practice CBs are said to use forward-looking rules, and Clarida, Gali and Gertler (2000) argue that empirical estimates of χ_π are less than 1 in the period before 1984, while they are greater than 1 for the subsequent period. Could SSEs explain the higher economic volatility in the earlier period?
Honkapohja and Mitra (2004) and Evans and McGough (2005) approach this question by asking when SSEs are stable under learning in the NK model. Surprisingly, SSEs appear never to be stable under learning for current-data Taylor rules. When the forward-looking Taylor rule is employed, stable SSEs occur not when $\chi_\pi < 1$, but rather when $\chi_\pi > 1$ and χ_π and χ_x are sufficiently large, that is, *overly* aggressive rules lead to learnable SSEs. However, this does not rule out the Clarida, Gali, Gertler explanation for pre-1984 instability because, if $\chi_\pi < 1$ leads to indeterminacy, *no* REE is stable under learning and aggregate instability would presumably result.

4.1.2 Stable SSEs in other models

Stability under learning is a demanding test for SSEs that is met in only some cases in the NK model. There are, however, other examples of stable SSEs, such as the basic OLG model.
Some nonlinear models can have multiple steady states that are locally stable under RLS learning. In this case there can also be SSEs that take the form of occasional random shifts between neighbourhoods of the distinct stable steady states. Examples of this are the 'animal spirits' model of Howitt and McAfee (1992), based on a positive search externality, and the 'growth cycles' model of Evans, Honkapohja and Romer (1998) based on monopolistic competition and complementarities between capital goods.
Two stable steady states also play a role in some important policy models. This can arise in a monetary inflation model with a fiscal constraint, developed by Evans, Honkapohja and Marimon (2001), and in the liquidity trap model of Evans and Honkapohja (2005). In these set-ups policy has an important role in eliminating undesirable steady states.

4.2 Dynamics with constant gain learning

An alternative route to explaining economic fluctuations is to modify RLS learning so that more recent observations are given a higher weight. A natural way to motivate this is to assume that agents are concerned about the possibility of structural change. In the RLS formula (2) this can be formally accomplished by replacing t^{-1} with a small 'constant gain' $0 < \gamma < 1$, yielding weights that geometrically decline with the age of observations.
This apparently small change leads to 'boundedly rational' fluctuations, with sometimes dramatic effects. Three main phenomena have emerged. First, as shown by Sargent (1999) and Cho, Williams and Sargent (2002), even when there is a unique equilibrium, occasional 'escape paths' can arise with learning dynamics temporarily driving the economy far from the equilibrium. Sargent shows how the reduction of inflation in the 1982–99 period might be due to such an escape path in which policymakers

are led to stop attempting to exploit a perceived (but misspecified) Phillips curve trade-off.
Second, in models with multiple steady states, learning dynamics can take the form of periodic shifts between regimes as a result of intrinsic random shocks interacting with learning dynamics. This is seen in the ‘increasing social returns’ example of Evans and Honkapohja (2001), the hyperinflation model of Marcet and Nicolini (2003), the exchange rate model of Kasa (2004) and the liquidity trap model of Evans and Honkapohja (2005).
Third, even when large escapes do not arise, there can be policy implications, because constant gain learning differs in small but persistent ways from full rationality. Orphanides and Williams (2005) show that policymakers attempting to implement optimal policy should be more hawkish against inflation than under RE.

5 Other developments

There continue to be many new applications of learning dynamics in macroeconomics, with closely related work in asset pricing and game theory.
One recent topic concerns the possibility that agents use a misspecified model. Under RLS learning agents may still converge, but to a restricted perceptions equilibrium, rather than to an REE (see Evans and Honkapohja, 2001). Another recent development is to allow agents to select from alternative predictors. In the Brock and Hommes (1997) model agents choose, based on recent past performance, between a costly sophisticated and a cheap naive predictor. This can lead to complex nonlinear dynamics. Branch and Evans (2006) combine dynamic predictor selection with RLS learning and show the existence of ‘misspecification equilibria’ when all forecasting models are underparameterized.
Other topics and applications include empirical work on expectation formation, calibration and estimation of learning models to data, interaction of policymaker and private-sector learning, learning and robust policy, experimental studies of expectation formation, the role of calculation costs, expectations over long horizons, alternative learning algorithms, expectational and structural heterogeneity, transitional learning dynamics, consistent expectations and near-rationality.
Current interest in learning dynamics is evidenced by five recent Special Issues devoted to learning and bounded rationality, in *Macroeconomic Dynamics* (2003), *Journal of Economic Dynamics and Control* (two in 2005), *Review of Economic Dynamics* (2005), and *Journal of Economic Theory* (2005).

See Also

- animal spirits
- determinacy and indeterminacy of equilibria
- expectations
- learning and evolution in games: an overview
- multiple equilibria in macroeconomics
- new Keynesian macroeconomics
- rational expectations
- rationality, bounded

- sunspot equilibrium
- two-stage least squares and the *k*-class estimator

Bibliography

Beck, G. and Wieland, V. 2002. Learning and control in a changing environment. *Journal of Economic Dynamics and Control* 26, 1359–77.

Branch, W. and Evans, G. 2006. Intrinsic heterogeneity in expectation formation. *Journal of Economic Theory* 127, 264–95.

Brock, W. and Hommes, C. 1997. A rational route to randomness. *Econometrica* 65, 1059–95.

Bullard, J. and Mitra, K. 2002. Learning monetary policy rules. *Journal of Monetary Economics* 49, 1105–29.

Cho, I.-K., Williams, N. and Sargent, T. 2002. Escaping Nash inflation. *Review of Economic Studies* 69, 1–40.

Clarida, R., Gali, J. and Gertler, M. 2000. Monetary policy rules and macroeconomic stability: evidence and some theory. *Quarterly Journal of Economics* 115, 147–80.

Evans, G. and Honkapohja, S. 2001. *Learning and Expectations in Macroeconomics*. Princeton, NJ: Princeton University Press.

Evans, G. and Honkapohja, S. 2003a. Expectations and the stability problem for optimal monetary policies. *Review of Economic Studies* 70, 807–24.

Evans, G. and Honkapohja, S. 2003b. Adaptive learning and monetary policy design. *Journal of Money Credit and Banking* 35, 1045–72.

Evans, G. and Honkapohja, S. 2005. Policy interaction, expectations and the liquidity trap. *Review of Economic Dynamics* 8, 303–23.

Evans, G. and Honkapohja, S. 2006. Monetary policy, expectations and commitment. *Scandinavian Journal of Economics* 108, 15–38.

Evans, G., Honkapohja, S. and Marimon, R. 2001. Convergence in monetary inflation models with heterogeneous learning rules. *Macroeconomic Dynamics* 5, 1–31.

Evans, G., Honkapohja, S. and Romer, P. 1998. Growth cycles. *American Economic Review* 88, 495–

515.

Evans, G. and McGough, B. 2005. Monetary policy, indeterminacy and learning. *Journal of Economic Dynamics and Control* 29, 1809–40.

Guesnerie, R. 2005. *Assessing Rational Expectations 2: 'Eductive' Stability in Economics*. Cambridge, MA: MIT Press.

Honkapohja, S. and Mitra, K. 2004. Are non-fundamental equilibria learnable in models of monetary policy? *Journal of Monetary Economics* 51, 1743–70.

Honkapohja, S. and Mitra, K. 2005. Performance of monetary policy with internal central bank forecasting. *Journal of Economic Dynamics and Control* 29, 627–58.

Howitt, P. and McAfee, R. 1992. Animal spirits. *American Economic Review* 82, 493–507.

Kasa, K. 2004. Learning, large deviations and recurrent currency crises. *International Economic Review* 45, 141–73.

Marcet, A. and Nicolini, J. 2003. Recurrent hyperinflations and learning. *American Economic Review* 93, 1476–98.

Orphanides, A. and Williams, J. 2005. Imperfect knowledge, inflation expectations and monetary policy. In *The Inflation-Targeting Debate*, ed. B. Bernanke and M. Woodford. Chicago: University of Chicago Press.

Sargent, T. 1993. *Bounded Rationality in Macroeconomics*. Oxford: Oxford University Press.

Sargent, T. 1999. *The Conquest of American Inflation*. Princeton, NJ: Princeton University Press.

Woodford, M. 2003. *Interest and Prices*. Princeton, NJ: Princeton University Press.

How to cite this article

Evans, George W. and Seppo Honkapohja. "learning in macroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000229> doi:10.1057/9780230226203.0954

The New Palgrave Dictionary of Economics Online

learning-by-doing

Spyros Vassilakis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capital–labour ratio; externalities; firm size, theory of; learning spillovers; learning-by-doing; partial equilibrium

Article

Empirical studies of the production process in various industries have demonstrated a positive association between current labour productivity and measures of past activity like past cumulative output or investment (see Wright, 1936; Hirsch, 1956; Alchian, 1963; Hollander, 1965; Sheshinski, 1967; Boston Consulting Group, 1972; 1974; 1978; and Lieberman 1984). A hypothesis advanced to explain this is that labour learns through experience and that experience is obtained during the production process. In other words, learning-by-doing is one of the reasons giving rise to dynamic economies of scale, because a firm knows that increasing current production reduces future average costs. If knowledge obtained within one firm cannot be communicated to other firms, we speak of learning without spillovers. There is some empirical evidence, though, that firms cannot totally exclude outsiders from their stock of knowledge, mainly because of labour turnover (see Boston Consulting Group, 1978, and Lieberman, 1984). Learning spillovers are a special case of positive externalities. The study of learning-by-doing, therefore, is a special case in the study of economies characterized by dynamic economies of scale and positive externalities.

Empirical studies of growth have demonstrated that increases in per capita output cannot be attributed solely, or even mainly, to increases in the capital–labour ratio (see Abramovitz, 1956; Solow, 1957; Kendrick, 1976). On the other hand, Verdoorn (1956) observed a positive relationship between past cumulative output and current labour productivity in the aggregate. This seems to suggest that the part of growth unexplained by increases in the capital–labour ratio could be accounted for by learning-by-doing. Once income per head increases for any reason, say because of an increase in the capital–labour ratio, it will keep on increasing for ever because the initial increase will improve labour productivity and income per head in the next period; after that, the chain of output increases resulting in productivity

increases and vice versa is repeated for ever and for the right values of the coefficients it will generate unbounded growth even with stationary population.

Formal models of this process were constructed by Arrow (1962), Levhari (1966), Romer (1986), and Stokey (1986). None of these authors model the process of learning explicitly but consider a world of perfect information with features that are supposed to emerge from a process of learning going on behind the scenes. Here we analyse Romer's model, which is the more general and pays particular attention to existence.

Romer considers a continuous time model with infinitely lived agents who produce and consume a single final good out of a fixed, inexhaustible supply of primary factors. At any given moment in time, the final good can be either consumed or added to the indestructible stock of capital. There is an exogenously given number of firms; each firm's output at any moment in time depends on the amount of capital accumulated by the firm up to that moment, on the amount of natural resources it employs and on the total amount of capital accumulated by all firms up to that moment. In other words, knowledge can be communicated across firms and is incorporated in the capital stock. There are diminishing returns in private capital accumulation but increasing returns when the effect of a firm's accumulation on the total capital stock is taken into account. Notice that the assumption of diminishing returns in private capital accumulation implies that the technical process generated by new learning is capital-augmenting, not land-augmenting. Firms maximize profit and consumers maximize utility taking prices as given.

Existence of Walrasian equilibria is demonstrated under the following assumptions: (*a*) firms do not recognize that their accumulation affects the total capital stock; (*b*) the growth rate of each firm's capital stock is uniformly bounded above and is a concave function of each firm's investment; (*c*) the production function is majorized by a constant plus a constant-elasticity function of the capital stock; (*d*) the discount factor is larger than the product of the above elasticity times the upper bound in the growth rate of the capital stock. Under some additional conditions, the equilibrium capital stock and consumption per head grow without bound. Equilibria are Pareto inefficient because firms do not take into account the fact that their private accumulation adds to the aggregate capital stock and therefore reduces everybody's future costs. Romer and Sasaki (1985) have generated unbounded growth with constant population and a fixed supply of exhaustible resources under more restrictive conditions on the coefficients.

Clearly, the main achievement of Romer's competitive model is to generate unbounded growth without assuming exogenous improvements in technology. The applicability and generality of the model, though, are restricted by the price-taking assumption and the related assumptions that all economies of scale are external to the firm and that the number of firms is fixed. Suppose for a moment that we accept the last two assumptions. Then, as Fudenberg and Tirole (1983) showed, if returns with respect to private capital accumulation are constant, no price-taking equilibrium exists; in other words, the assumption that technical progress is not land-augmenting is crucial. Also, there is no reason why firms should fail to recognize the effect of their actions on the capital stock. Spence (1981) and Fudenberg and Tirole (1983) constructed dynamic partial equilibrium models of learning without spillovers in which a fixed number of firms compete in quantities with Cournot expectations. Industry output may decline over time at a subgame perfect equilibrium, depending on how large is the discount factor relative to the number of firms. Stokey (1986) investigated the same model with spillovers and found that industry output increases over time. Given the importance of spillovers in generating growth, therefore, it seems

worthwhile to study their determinants.

The next step is to remove the assumption that all dynamic economies of scale are external to the firm and that the number of firms is fixed. Even if one begins with a situation of purely external economies of scale, there are powerful economic incentives to internalize these economies by reductions in the number of firms, either by collusion or by competition that drives some firms out of business. The tendency to collude to internalize externalities is checked by the incentive of each firm to shirk (underinvest in learning) given that others have done their share of investment. The tendency of competition to reduce the number of firms is checked by entry when the number of firms is so small that a new entrant's gains by wiping out excess profits exceed losses due to lost economies of scale. The learning-by-doing model coupled with such a theory of firm size could generate more predictions about growth, concentration and distribution of income over time.

Finally, one has to address the issue of the evolution of the externalities themselves over time. The extent of learning spillovers is limited by concentration and by the creation of markets in order to transform the external effects into ordinary goods; both of these magnitudes are endogenous, and so the extent of learning spillovers should also be an endogenous variable. The current formulation of learning spillovers assumes a stable, exogenous relationship between measures of past activity and future productivity, but in a long-run model one would not expect such a relationship to hold, exactly because the number of firms and completeness of markets are variable in the long run.

Bibliography

Abramovitz, M. 1956. Resource and Output Trends in the U.S. since 1870. Occasional Paper No. 52, NBER.

Alchian, A. 1963. Reliability of progress curves in airframe production. *Econometrica* 31, 679–93.

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 29, 155–73.

Boston Consulting Group. 1972. Perspectives in experience. Technical Report.

Boston Consulting Group. 1974. The experience curve reviewed: price stability. Technical Report.

Boston Consulting Group. 1978. Cross sectional experience curves. Technical Report.

Fudenberg, D. and Tirole, J. 1983. Learning-by-doing and market performance. *Bell Journal of Economics* 14, 522–30.

Hirsch, W.Z. 1956. Firm progress ratios. *Econometrica* 24(2), 136–44.

Hollander, S. 1965. *The Source of Increased Efficiency: A Study of DuPont Rayon Plants*. Cambridge, MA: MIT Press.

Kendrick, J.W. 1976. *The Formation and Stocks of Total Capital*. New York: NBER.

Levhari, D. 1966. Extensions of Arrow's learning by doing. *Review of Economic Studies* 33, 117–32.

Lieberman, D. 1984. The learning curve and pricing in the chemical processing industries. *RAND Journal of Economics* 15, 213–28.

Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Romer, P. and Sasaki, H. 1985. Monotonically decreasing natural resource prices under perfect foresight. Working Paper No. 19, Rochester Center for Economic Research, New York.

Sheshinski, E. 1967. Tests of the learning-by-doing hypothesis. *Review of Economics and Statistics* 49, 568–78.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

Spence, A.M. 1981. The learning curve and competition. *Bell Journal of Economics* 12, 49–70.

Stokey, N. 1986. The dynamics of industry-wide learning. In *Equilibrium Analysis: Essays in Honor of K.J. Arrow*, vol. 2, ed. W.P. Heller, R.M. Starr and D.A. Starrett. Cambridge: Cambridge University Press.

Verdoorn, P.J. 1956. Complementarity and long-range projections. *Econometrica* 24, 429–50.

Wright, T.P. 1936. Factors affecting the cost of airplanes. *Journal of Aeronautical Sciences* 3(4), 122–8.

How to cite this article

Vassilakis, Spyros. "learning-by-doing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000056> doi:10.1057/9780230226203.0953

The New Palgrave Dictionary of Economics Online

Lederer, Emil (1882 – 1939)

Robert A. Dickler
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; bureaucracy; business cycles; cartels; class; creative destruction; division of labour; entrepreneurship; excess capacity; Great Depression; increasing returns; Lederer, E.; mixed economy; planning; structural unemployment; technical change; uncertainty; vertical integration

Article

Emil Lederer was a prominent economist and sociologist in the German Weimar Republic. He was born in the Bohemian town of Pilsen in 1882 and died in political exile in New York City in 1939. In Vienna and Berlin, where he studied both law and economics, Lederer participated in advanced seminars conducted by Menger, Böhm-Bawerk and Schmoller. From 1918 to 1931 he served as professor in Heidelberg and then succeeded Sombart in Berlin from 1931 to 1933. In collaboration with E. Jaffé, Schumpeter and Sombart as well as Max and Alfred Weber, he edited the *Archiv für Sozialwissenschaft und Sozialpolitik*, the renowned social science journal which ceased publication under the Nazi regime. After emigrating to New York in 1933 be became the first Dean of the New School for Social Research's Faculty of Political and Social Sciences, which was comprised of outstanding Continental scholars who had also sought asylum in the United States.

Lederer made pioneering contributions towards understanding the social, political and economic significance of large-scale, bureaucratic private enterprise. In a major theoretical and empirical study based on his *Habilitation*, Lederer undertook the first comprehensive analysis of the working conditions and political attitudes of salaried employees (Lederer, 1912). Subsequent work together with Jacob Marschak showed how rationalization of production along with bureaucratic division of labour in administration formed the basis for the rise of the new middle class (Lederer and Marschak, 1926). They concluded that the evolution of class structure in advanced capitalist societies undermines political stability and raises the spectre of fascism. Anxiety stemming from economic insecurity and abhorrence of collective action with organized labour weakens the growing middle class's support for democratic forms of government and strengthens its tolerance of authoritarian institutions to suppress the demands of the proletariat.

Lederer's advanced economics textbook contains an authoritative exposition and critique of objective and subjective value theories (Lederer, 1931a). The laws of the market economy, as depicted in the marginalist doctrine, are no longer in effect, since economies of large-scale production prevail. Adoption of modern technologies requires vertical integration, high proportion of fixed capital and substantial fixed costs for sales and general administrative overhead. Complementarities and decreasing marginal costs are the rule in basic industries (coal, steel, chemicals and utilities).

Increasing returns to scale is not only an anomaly which cannot be subsumed under the marginalist paradigm; it forms the starting point for business cycle theory (Lederer, 1924; 1927).

Disproportionalities in growth of demand for investment and consumer goods are due to unavoidable price inflexibilities and absence of strong equilibrating tendencies. Cartels which administer prices and set production quotas are the natural outcome of the technically determined drive to realize economies of scale. The self-contained planning of separate industrial bureaucracies lacks inter-industry coordination and thus cannot prevent misallocation, underutilization and periodic decumulation of capital.

Rapid labour-saving technical change is regarded by Lederer as a key factor in explaining the severity of unemployment during the Great Depression (Lederer, 1931b; 1936a). In an upswing, dynamic enterprises exploit opportunities to realize above-normal returns on investment offered by introduction of highly mechanized techniques. Labour is displaced not only by rationalization of operations but also by diversion of capital from static enterprises which do not employ the new techniques. As productivity and productive capacity in dynamic enterprises increases, monopolistic market structures prevent prices from falling faster than wages. Redistribution of income from labour to capital decreases consumer goods demand, which in turn reduces the derived demand for capital goods and brings about excess capacity in capital goods production. Without incentives for accelerating the form of technical progress which creates new products, opens up new markets and stimulates labour-absorbing investment, technological unemployment persists.

Stressing the distinction between labour-saving and labour-absorbing forms of technical progress, Lederer criticized Keynes for his failure to analyse long-run dynamics (Lederer, 1936b). Investment in plant and equipment embodies new techniques. Not the lack of profitable investments, but rather an abundance of abnormally profitable rationalization investments creates structural unemployment in addition to the cyclical unemployment treated by Keynes. Government spending is necessary to stimulate the economy, but it is not sufficient to overcome mass unemployment. Democratic national planning is also necessary to attract capital to new industries offering additional employment opportunities.

Lederer's conviction that a mix of market and planned economies based on political consensus is practicable may be traced to his close association with the industrialist and statesman, Walther Rathenau, who was the architect of German economic mobilization in the First World War (Lederer, 1933; 1934).

Along with Schumpeter, Lederer cultivated an undogmatic Austrian style of theorizing. Both emphasized the significance of uncertainty, entrepreneurship (or its absence as a consequence of bureaucratization), disequilibrating forces, such as technical change, and underlying instability of capitalism. Schumpeter (1939; 1942) defended neoclassical equilibrium theory by asserting that the price system it represents moves automatically, but not without friction, towards a new equilibrium following the ‘creative destruction’ of an old equilibrium. Similarly, Lederer (1931b) wrote: ‘The capitalist dynamic is not only “development” but also “destruction” ’. However, Lederer combined neo-

Ricardian (von Bortkiewicz, 1907) and Austro-Marxian (Hilferding, 1910) approaches to focus on the production system; accordingly, in his view there was no automatic mechanism to assure that investment brings about a rate and direction of technical change consistent with full employment equilibrium.

Selected works

1912. *The Problem of the Modern Salaried Employee: Its Theoretical and Statistical Basis*. Trans. E.E. Warburg, New York: State Department of Social Welfare and the Department of Social Science, Columbia University, 1937.

1924. Konjunktur und Krise. *Grundriss der Sozialökonomik* 4(1). Tübingen: J.C.B. Möhr.

1926. (With J. Marschak.) *The New Middle Class*. Trans. S. Ellison, New York: State Department of Social Welfare and the Department of Social Science, Columbia University, 1937.

1927. Monopol und Konjunktur. *Vierteljahreshefte zur Konjunkturforschung*, vol. 2, Supplement 2. Berlin: Duncker & Humblot.

1931a. *Aufriss der ökonomischen Theorie*. 3rd edn. Tübingen: J.C.B. Möhr.

1931b. *Technischer Fortschritt und Arbeitslosigkeit*. Tübingen: J.C.B. Möhr.

1933. National economic planning. In *Encyclopaedia of the Social Sciences*, vol. 11. New York: Macmillan.

1934. Rathenau, Walther. In *Encyclopaedia of the Social Sciences*, vol. 13. New York: Macmillan.

1936a. *Technical Progress and Unemployment. An Enquiry into the Obstacles to Economic Expansion*. Studies and Reports, Series C, No. 22. Geneva: International Labor Office. Translation of revised edition of (1931a).

1936b. Commentary on Keynes. *Social Research* 3, 478–87.

1979. *Kapitalismus, Klassenstruktur und Probleme der Demokratie in Deutschland 1910–1940*. Göttingen: Vandenhoeck & Ruprecht. (Collection of essays, edited by J. Kocka with biographical essay by Hans Speier.)

Bibliography

Bortkiewicz, L. von. 1907. On the correction of Marx's fundamental theoretical construction in the third volume of *Capital*. In *Karl Marx and the Close of His System by E. von Böhm-Bawerk and Böhm-*

Bawerk's Criticism of Marx by R. Hilferding, ed. P.M. Sweezy, New York: Kelley, 1949.

Hilferding, R. 1910. *Das Finanzkapital*. Vienna: I. Brand.

Schumpeter, J.A. 1939. *Business Cycles*. New York: McGraw-Hill.

Schumpeter, J.A. 1942. *Capitalism, Socialism and Democracy*, New York: Harper. 2nd edn, 1947.

How to cite this article

Dickler, Robert A. "Lederer, Emil (1882–1939)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000059> doi:10.1057/9780230226203.0955

The New Palgrave Dictionary of Economics Online

legal institutions and the ancient economy

Dennis P. Kehoe

From The New Palgrave Dictionary of Economics Online, 2008,
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This brief survey suggests some of the issues that can be investigated by a careful analysis of the relationship between legal institutions and the economy in the ancient world. By investigating legal institutions, we can better understand the relationships that shaped the economy and the likely implications of these relationships for economic performance. It covers institutions in the Ancient Greek world, in the Ancient Roman world, and more briefly in Ptolemaic Egypt.

Keywords

agency; ancient Greece; ancient Rome; ancient Egypt; law and economics; new institutional economics

Article

Within the broad constraints imposed by population and technology (Scheidel, Morris and Saller, 2007), law and legal institutions played an important role in ancient economies. The overriding question concerns how formal institutions, including courts and contractual types, and informal ones, such as social conventions or ideology, affected incentives to enter into mutually beneficial contractual arrangements. The alternative is that the laws and legal institutions surrounding an ancient economy served primarily to protect the privileges or interests of certain well-connected groups. Understanding the role of legal institutions in an ancient economy is complex because the available evidence usually makes it impossible to verify hypotheses about the likely incentives resulting from various property rights regimes. Still, analysing ancient legal institutions can shed light on the basic relationships among the principal actors in an ancient economy, including the state, elite property owners, urban residents, and farmers.

Legal institutions in the Greek world

One key issue is the role that legal institutions played in promoting commerce. The Greek world in the classical (480–323 BCE) and Hellenistic periods (323–31 BCE) was politically fragmented, and individual city-states (*poleis*) had their own legal systems. Consequently, we can speak of a unified system of Greek law only to a limited degree. This made it difficult to develop governance structures to enforce the types of contractual arrangements essential to commerce. In the Hellenistic period, the emergence of larger monarchies may have promoted a more unified system of commercial law between states. Eventually, the incorporation of the Greek world into the Roman Empire greatly enhanced the possibilities for developing a more uniform set of legal institutions. In the absence of unified formal institutions to govern commerce, we should expect merchants to have developed their own private ways of enforcing contractual obligations and resolving disputes (cf. Greif, 2006).

At the level of the polis, we are best informed about the way in which commercial law functioned in classical Athens, particularly in the fourth century BCE (Todd, 1993; Cohen, 2005). At this time, Athens had become a commercial hub, and its involvement in commerce was vital to its survival, since it depended on imported grain from the Black Sea region. Certainly the economy of Athens, as much as any place in the ancient world, required legal mechanisms to develop and enforce complex commercial arrangements. The state intervened in commerce directly only to protect the grain supply by imposing severe sanctions on Athenians who exported grain to other cities. Even so, institutions developed in Athens to promote trade. Banks played a crucial role in assembling the capital necessary for maritime commerce. Often these commercial undertakings might be complex, with multiple investors supplying cargoes to the same ship, so that a single voyage might involve a wide variety of contracts and loans (Cohen, 1992). The question is how merchants involved in this commerce, many of whom came from locations overseas, could enforce the obligations of their trading partners. In most city-states, the local courts were open only to citizens, unless two states negotiated a bilateral commercial treaty. The Athenians endeavored to meet a more general need for a forum to resolve disputes by developing courts in which lawsuits involving overseas commerce, *dikai emporikai*, could be heard (Todd, 1993, pp. 333–7; Cohen, 2005, pp. 299–300). These courts were open to anyone doing business in Athens, not just Athenian citizens. Their success in encouraging commerce depended on their treating foreign traders in Athens impartially. Foreign traders, when sued in the court, had to post bond, but at the same time the courts discouraged frivolous lawsuits by imposing financial penalties on plaintiffs who failed to gain at least one-sixth of the jury's votes (Cohen, 2005, pp. 299–302).

Another issue is the role that private contract law played in the economy. In contrast to Roman law, the mutual consensus of the two parties to a contract did not in and of itself create contractual obligations; rather, a real act, such as the exchange of property, was required (Todd, 1993, pp. 262–8; Rupprecht, 2005, p. 337). The apparent simplicity of this type of contract would seem to preclude certain complex commercial arrangements, such as sales of real estate on credit or sales of crops in advance of the harvest, but this was manifestly not the case. In Athens, one part of the solution to the problem was the freedom of procedure in courts; this flexibility made it possible to sue regardless of whether a business arrangement corresponded to an accepted contractual form.

The Hellenistic period saw legal developments potentially significant for the economy (Rupprecht, 2005). Typically, multiple legal systems functioned side by side. In Egypt, for example, the Ptolemies, a Macedonian dynasty, introduced Greek law for the immigrant Greek population, while the native Egyptians continued to rely on their own legal traditions and contract forms. The substantial Jewish

population in Egypt could also use its own laws. In Greek law, written documents were increasingly common in private business arrangements. They tended to be written in standard language, and so they would have served to make contracting in business simpler. The widespread use of written contracts means that there were scribes well versed in the basics of commercial law. The trend of using written documents to record what had originally been oral contracts accelerated in Roman times (Meyer, 2004).
A second development in Ptolemaic Egypt was the increasing registration of documents in state archives. In the early Roman imperial period in Egypt, the state developed a registry of real property and the rights assigned to it, the *bibliotheke enkteseon*, which helped to eliminate some of the uncertainty surrounding the ownership of real property that is characteristic of pre-modern economies.
The development of commerce in the Greek world, and in the Roman world later, depended on property owners having reliable agents to manage their businesses. Part of the solution in both the Greek and Roman worlds was to employ agents who were social dependants. In fourth-century Athens, this can be seen especially in the banking industry. The general prohibition against the ownership of land in Attica by non-Athenian citizens surely made banking an attractive business undertaking for resident aliens (metics), many of whom were quite wealthy. The foreign owners of these banks commonly employed slaves as their managers. A highly trained slave could operate a bank independently, but there was no threat that he would take advantage of his training to set up a rival bank to compete with that of his former employer (Cohen, 1992, pp. 61–110, 133–6). In Rome, property owners employed slaves and freedmen in similar functions, as will be discussed below.

Legal institutions in the Roman world

The development of Roman law as a legal system with wide application in the Mediterranean world had potentially enormous consequences for the Roman economy. Roman society had a professional class of jurists who interpreted the law in a rigorous fashion and, in effect, created a science of jurisprudence. The jurists originally provided legal advice in private trials, but beginning with the reign of Augustus (31 BCE – 14 CE) they gained a state-sanctioned role in providing authoritative interpretations of the law. In economic matters, one of the jurists' main contribution was to interpret contract law. By the second century BCE, the Roman praetors (the officials in charge of the administration of private law) had developed the concept of consensual contracts, including sale (*emptio-venditio*), lease and hire (*locatio-conductio*), mandate (*mandatum*), and partnership (*societas*). The contract types defined legal relationships crucial for the Roman economy, and they provided a basis for Roman commercial law for centuries to come. Although this is a controversial subject, it is now increasingly accepted that the jurists endeavored to respond to social needs as they interpreted contract law.
The Roman Empire was also successful in developing legal institutions that were accessible to a broad segment of the population. One key to this was the petition process. The Roman emperor received such a volume of petitions that the Roman government had an office, headed by an official of equestrian rank, the *a libellis*, whose responsibility was to receive petitions and issue answers, or rescripts, in the emperor's name (Peachin, 1996). Petitioners would receive an authoritative response about the law applicable to their case, and they could then take these responses to local courts, whose judges would be obliged to follow them. People also sent petitions to officials of lower rank, from local magistrates to provincial governors. The petition process was so widespread that it suggests that the empire's subjects

viewed it as a reasonably reliable way to protect their interests. Responding to petitions, moreover, provided the state with one way, albeit reactive, to intervene in the economy. Such intervention can be discerned in legal policies concerning farm tenancy, in such issues as the tenant's security of tenure and the allocation of the risks associated with agriculture (Kehoe, 2007).

To consider agency again, the Romans, like the Greeks, often relied on social dependants, particularly slaves and freedmen, to serve as business agents. To some extent, this resulted from the basic organizing principles of the Roman household. In Roman society, the head of a Roman household, the *pater familias*, exercised a great deal of power over the members of his *familia*. These included his agnatic descendants as well as his slaves and freedmen. In economic terms, he was the ultimate owner of all the property in the hands of anyone in his power, or *patria potestas* (Saller, 1994: 102–32). The *familia* provided the basic structure for organizing much of economic life in the Roman world. It was a setting in which people were trained in specialized skills important for the economy, and it also influenced the organization of commercial enterprises. When employing social dependants as agents, Roman property owners tended to give them a great deal of freedom. The slave would operate with a *peculium*, funds and property under his control but ultimately belonging to the owner. The slave agent had every incentive to manage the business well, since he could earn his freedom in doing so, whereas the owner could impose sanctions in the event of his misbehaviour more easily than would be possible with a free employee (Frier and Kehoe, 2007, pp. 130–4). Often freedmen who gained their initial training as slaves could establish businesses of their own, training their own slaves, and continuing the cycle.

Merchants dealing with agents had to be assured that they would be able to enforce their claims in the event of a dispute. Part of the solution was a series of remedies, the so-called *actiones adiecticiae qualitatis*, created in the late third or second centuries BCE. These established the circumstances under which a property owner could be liable for obligations taken on by an agent. In many cases, the principal's liability was limited to the size of the *peculium* granted the slave agent. This legal regime may have carried a substantial social cost, since in theory at least, the limited liability of the principal will have deterred some people from entering into otherwise productive business arrangements. At the same time, it responded to the needs of an upper class that was cautious in its approach to investing wealth (Kehoe, 1997). The formal regime surrounding agency in Roman law can be contrasted with the type of agency that characterized Ptolemaic Egypt (Von Reden, 2007, pp. 239–50). There, property owners who also held official posts relied on private, individual agents, who collected debts or made loans on their behalf. The activities of the agent, however, created no formal legal relationship between the property owner and a third party who was either a debtor or a creditor. This system of agency clearly revolved around the personal reputations of the individuals involved.

In interpreting Roman contract law, the Roman jurists seem to envision a class of independent contractors who had sufficient resources to undertake major jobs, such as leasing farms or construction projects. In the contractual relationship covering major construction projects, called *locatio-conductio operis* (Martin, 1989), the builders were expected to organize tasks and finance operations until they were paid by their principals. Again, this situation can be contrasted usefully with the corresponding contract arrangement in Ptolemaic Egypt, called *ergolabia*. In such contracts from the third century BCE, for example, the property owner employing the contractor generally had to pay the latter up front. The contractor still had a great deal of responsibility, but the payment up front created potential monitoring problems, and it was probably necessary because at this time contractors did not have ready

access to cash (von Reden, 2007, pp. 146–50).

This brief survey suggests some of the issues that can be investigated by a careful analysis of the relationship between legal institutions and the economy in the ancient world. By investigating legal institutions, we can better understand better the relationships that shaped the economy and the likely implications of these relationships for economic performance.

See Also

- agency problems
- economy of ancient Greece
- new institutional economics

Bibliography

Cohen, E.E. 1992. *Athenian Economy and Society: A Banking Perspective*. Princeton: Princeton University Press.

Cohen, E.E. 2005. Commercial law. In Gagarin and Cohen 2005, pp. 290–302.

Frier, B.W. and Kehoe, D.P. 2007. Law and economic institutions. In Scheidel et al., 2007, pp. 113–43.

Gagarin, M. and Cohen, D. (eds.) 2005. *The Cambridge Companion to Ancient Greek Law*. Cambridge: Cambridge University Press.

Greif, A. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge: Cambridge University Press.

Kehoe, D.P. 1997. *Investment, Profit, and Tenancy: The Jurists and the Roman Agrarian Economy*. Ann Arbor: University of Michigan Press.

Kehoe, D.P. 2007. *Law and the Rural Economy in the Roman Empire*. Ann Arbor: University of Michigan Press.

Martin, S.D. 1989. *The Roman Jurists and the Organization of Private Building in the Late Republic and Early Empire*. Collection Latomus 204. Brussels.

Meyer, E.A. 2004. *Legitimacy and Law in the Roman World: Tabulae in Roman Belief and Practice*. Cambridge: Cambridge University Press.

Peachin, M. 1996. *Iudex vice Caesaris: Deputy Emperors and the Administration of Justice during the Principate*. Heidelberger Althistorische und Epigraphische Beiträge 21. Stuttgart: Steiner.

Rupprecht, H.-A. 2005. Greek law in foreign surroundings: continuity and development. In Gagarin and Cohen 2005, pp. 328–42.

Saller, R.P. 1994. *Patriarchy, Property and Death in the Roman Family*. Cambridge: Cambridge University Press.

Scheidel, W., Morris, I. and Saller, R. (eds.) 2007. *The Cambridge Economic History of the Greco-Roman World*. Cambridge: Cambridge University Press.

Todd, S.C. 1993. *The Shape of Athenian Law*. Oxford: Clarendon Press.

Von Reden, S. 2007. *Money in Ptolemaic Egypt: From the Macedonian Conquest to the End of the Third Century BC*. Cambridge: Cambridge University Press.

How to cite this article

Kehoe, Dennis P. "legal institutions and the ancient economy." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000319> doi:10.1057/9780230226203.1883

The New Palgrave Dictionary of Economics Online

leisure class

F. Stankovi•
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The idea of a leisure class was popularized by Thorstein Veblen, whose *Theory of the Leisure Class* (1899) developed the social categories of pecuniary competition, conspicuous leisure and conspicuous consumption. Bukharin's *Economic Theory of the Leisure Class* (1919) argued that marginal utility theory was the theoretical expression of the class of rentiers who had been eliminated from the process of production and were interested only in disposing of their incomes. In *The Age of Uncertainty* (1977) J. K. Galbraith argued for the continuing relevance of Veblen's analysis. Modern sociologists, however, show little interest in the idea of a leisure class.

Keywords

absentee ownership; Bukharin, N.; class; conspicuous consumption; conspicuous leisure; Galbraith, J. K.; labour; labour theory of value; leisure; leisure class; marginal utility theory; Marxism; private ownership; rentiers; social status; socialism; subjective value theory; Veblen, T.

Article

This term became popular after Thorstein Veblen's book, *The Theory of the Leisure Class* (1899). In that book the author gives a historical and socio-economic explanation of the development of that wealthy class in the society of his time whose main characteristic was leisure. By ‘leisure’ Veblen means the non-productive spending of time which originates from a sense of the worthlessness of productive work and from the need to show pecuniary ability to afford a life of idleness. The basic social categories of Veblen's theory of the leisure class are pecuniary competition, conspicuous leisure and conspicuous consumption.
The leisure class is an old institution. The emergence of a leisure class coincides, according to Veblen, with the beginning of ownership. These two institutions (leisure class and ownership) are different aspects of the same general facts of social structure. The conditions for the appearance of the leisure

class as a permanent form are: (1) the community has to be of a predatory type; and (2) the means of life must be affordable in a relatively easy way, so that part of the population can be liberated from routine work. It is necessary to make a proper distinction between the leisure and the labouring class. For the lower classes, as Veblen explains, since labour is their accepted and only mode of life, an emulative pride in a reputation for efficiency in their work becomes the only emulation that is open to them. For the ‘superior pecuniary class’ the most imperative secondary demand for emulation is the abstention from productive work. It is not sufficient to possess wealth or power. It becomes important to show that you have no need to do productive work. From the days of the Greek philosophers to the present a life of leisure is, as Veblen says, in a great part of secondary and derivative value (1936, p. 231).

Leisure does not usually bring about a material product; the result takes the form of non-material goods. Good examples of occupations that members of the leisure class choose to pursue are the knowledge of dead languages and the occult sciences, of correct spelling, of the various forms of music and other household arts, fashionable dress, furniture, games, sports, dogs and racehorses. Elegant speech shows the level of a speaker's emancipation from productive work. The leisure class is not interested in technological innovations and it is an obstruction to social and economic progress.

Veblen's critique of ‘absentee ownership’ is the next step in his analysis of modern civilization. The members of the leisure class do not want to have any connection with a production process and they leave the managing and guiding of this process to so-called ‘captains of industry’. The last group is for Veblen the only positive power interested in technological development.

Veblen's critique of ownership and his opinion that the modern type of ownership is not compatible with industrial efficiency has been and still is unacceptable to many of his fellow economists. Like Marx, Max Weber and Karl Polanyi, Veblen had demonstrated the importance of studying primitive economies for general economic history and the interaction between economics and the society in general.

Veblen's leisure class seems very much like Marx's ruling capitalist class. Both of them see the bitter struggle between capital and labour. But they differ very much according to the methods of solving this contradiction. Marx's solution is, as is well known, a socialist revolution. In *The Theory of the Leisure Class* Veblen was not explicit about this issue. He was not sure about the end of this struggle, but he was very positive about the existence of the struggle. This goes very well with his Darwinian philosophy. A kind of trade unionism could be closer to Darwinism and is more acceptable for Veblen than neo-Hegelianism or Marxism.

Although his theory is based on Darwinian philosophy and although he argues for evolutionary socialism, in some of his later writings he was closer to the attitude that certain radical social movements could be the solution for breaking with old social institutions. In his search for emancipation of man, Veblen, together with other great humanists, had to become a socialist. His analysis of private ownership and the leisure class necessarily pushed him into this direction.

Nikolai Bukharin's *Economic Theory of the Leisure Class* (1919) is quite different from Veblen's book. This is not an economic account of the conditions which give rise to the existence of a leisure class in the manner of Veblen. The leisure class is not the central category in this work, as the title may suggest. Bukharin mentions the leisure class in the context of his critique from a Marxist point of view of the theory of marginal utility, especially that of the Austrian school. He gives no reference to Veblen. Bourgeois political economy, according to Bukharin, seeks to justify the capitalist system and therefore it loses its scientific role, contrary to the Marxist theory which claims its general validity precisely for the reason that it is the theoretical expression of the most advanced class – the working class.

In the critique of marginal utility theory Bukharin points out that this theory is the theoretical expression of the class of rentiers who have been eliminated from the process of production and are interested in disposing of their income from holdings of securities and bonds only. This marginal utility theory is, according to Bukharin, unhistorical: it starts with consumption not the production process. His critique of the logic and the method of subjective value theory is settled in direct confrontation with the labour theory of value.

J.K. Galbraith refers to Veblen's work, suggesting that the concepts ‘conspicuous leisure’ and ‘conspicuous consumption’ are still of significance (Galbraith, 1977). In the United States, as Galbraith explains, class as described by Veblen still exists: the members of the leisure class are still buying their social status.

It might have been expected that contemporary sociologists would have been more concerned with the concept of the leisure class. But this is not the case. They are very much occupied with leisure itself: it is common for sociologists to define leisure as the portion of time which remains when time for work and the basic requirements for existence have been satisfied, and this issue is related to the problem of how to spend leisure time. But this is a different problem and does not have a substantial relation to the problem of the leisure class.

See Also

- Bukharin, Nikolai Ivanovitch
- conspicuous consumption
- Veblen, Thorstein Bunde

Bibliography

Bukharin, N. 1919. *The Economic Theory of the Leisure Class*. New York: Monthly Review Press, 1972. (Originally published in Russian and first published in English in 1927.)

Galbraith, J.K. 1977. *The Age of Uncertainty*. Boston: Houghton Mifflin.

Veblen, T. 1899. *The Theory of the Leisure Class*. London: George Allen & Unwin.

Veblen, T. 1936. What Veblen taught. In *Selected Writings of Thorstein Veblen*, ed. W.C. Mitchell. New York: Viking Press.

How to cite this article

Stankovi•, F. "leisure class." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000064> doi:10.1057/9780230226203.0957

The New Palgrave Dictionary of Economics Online

leisure

Lars Osberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists have typically defined ‘leisure’ residually, as equal to ‘non-work time’, and, despite the problematic classification of enjoyable jobs, commuting time and unemployment, presumed that individuals derive utility from non-work time and disutility from working time. However, a recent literature now emphasizes ‘social leisure’ and coordination problems in leisure time. Since longer working hours by some individuals make arranging a social life more difficult for others (thereby decreasing the utility of their non-work time), externalities in time use may create multiple possible equilibria in time use, which may explain the sharp divergence in working hours between Europe and the United States.

Keywords

Becker, G.; external economies; goods-intensive commodities; individualism; labour supply; leisure; marriage; social capital; social interaction; time use; time-intensive commodities; unemployment; well-being

Article

What is ‘leisure’? The Merriam-Webster Online Dictionary defines it as ‘freedom provided by the cessation of activities; *especially*: time free from work or duties’, while the Oxford English Dictionary suggests it is ‘The state of having time at one's own disposal; time which one can spend as one pleases; free or unoccupied time’. (Both note that the adjective ‘leisurely’ describes an action that is done without haste, in a relaxed way.) In common parlance, attendance at a relative's funeral or time spent voting would therefore not generally be seen as ‘leisure’, because time spent on an activity due to a sense of civic or familial duty cannot qualify.
‘Leisure’ is therefore a problematic concept for economists, because the context and subjective interpretation of an activity is crucial to deciding whether it should be counted as work, duty or leisure –

cooking or driving are, for example, activities that may be performed as parts of a paid occupational role, as a duty or for personal enjoyment. It is, in fact, not easy to think of an activity or time use that is not done sometimes for pay, sometimes for duty and sometimes for pleasure – perhaps by different people, but sometimes also by the same people. In many universities, the subtleties of such distinctions are explored in departments of ‘Leisure Studies’, which is now a recognized area of academic teaching and research. Peer-reviewed journals such as *Annals of Leisure Research* or *Leisure Sciences* report the latest research on leisure activities, and conferences are organized on such topics as ‘Serious and Casual Leisure’.

Leisure as a residual category: the standard approach

However, for many economists, ‘leisure’ is simply the *L* in labour supply theory. This approach starts, in a one-period model, with each individual maximizing a utility function, where *U* is the individual's utility level, *C* represents consumption goods and *L* is leisure time, as in eq. (1):

$$Max\, U = u(C, L)\,[u' > 0 u'' < 0] \tag{1}$$

The wage rate available in the paid labour market (*w*) and total time (*T*) are seen as the fundamental constraints facing individuals. In this framework, the problem of utility maximization can be equivalently seen as one of ‘labour supply’ or ‘leisure demand’ since total time is divided between hours of paid work (*H*) and leisure time (*L*).

$$H + L = T \tag{2}$$

$$C \leq wH. \tag{3}$$

From this perspective, ‘leisure’ is whatever ‘work’ isn't – that is, leisure is a residual category, which is rarely examined directly or defined explicitly. Standard practice in economics journals is to focus on the hours of work decision – and ‘work’ is usually interpreted to mean ‘paid employment’. In the JSTOR

database of the top 26 economics journals, a keyword search, conducted in July 2005, for ‘leisure’ in archived articles published since 1995 yielded 823 ‘hits’. Of the top 100, sorted for ‘relevance’, only 25 had an explicit verbal definition of leisure – in most cases leisure was defined implicitly, as in eq. (2). If one discards the three articles discussing consumer demand for ‘leisure goods’ and focuses on time use, one finds the overwhelming majority of articles used leisure as a synonym for ‘non-market time’ – only three per cent recognized the possibility of ‘on the job leisure’ (but the definition was similarly residual – a lack of work effort – and implicit – for example, Dickinson, 1999, p. 639). Relatively few articles (about 15 per cent) considered the possibility that home production (such as shopping time) may be a form of ‘work’, while a similar number (about 13 per cent) argued that time spent in schooling or training preparatory to paid employment is not leisure. For a very few articles (three per cent), leisure was the residual time available after paid work and some other alternative, such as criminal activity. When working time is defined as equal to hours of paid employment, commuting time is implicitly defined as part of leisure, although it is plausibly an intermediate input into paid employment. Commuting time is an important percentage of time use in modern societies – Putnam (2000, p. 212), for example, has ascribed much of the decline in civic engagement in the United States to increased commuting time and commented that ‘American adults average seventy-two minutes every day behind the wheel….more than we spend cooking or eating and more than twice as much as the average parent spends with the kids'.’ However, commuting time is strangely absent from most labour–leisure models. As well, although ‘retirement’ is the particular form of non-work time consumed at the end of the life cycle, most economics articles implicitly exclude it from analysis, by concentrating on the working-age population.

All the same, although $L = T - H$ remains the dominant approach in economics, it has long been recognized that classifying time use as ‘work’ (painful) or ‘leisure’ (pleasurable) can be a bit oversimplified. A large body of research indicates, for example, that the unemployed are typically quite unhappy (Frey and Stutzer, 2002; Di Tella, McCulloch and Oswald, 2003) – time spent in unemployment seems to be qualitatively different from non-work time spent in other ways (that is, unpleasant). In general, people tend to rank their jobs fairly highly when asked to compare the satisfaction derived from specific activities (including jobs and types of housework and leisure). Juster and Stafford (1985) argued long ago that, in general, activities that involve social interaction – whether paid or unpaid – tend to be highly valued by individuals. Gary Becker (1965, p. 504) commented even earlier that ‘Not only is it difficult to distinguish leisure from other non-work, but also even work from non-work'.’

‘Time-intensive commodities’ and the disappearance of ‘leisure’: the Becker approach

Becker's solution to the time classification problem was to posit that ‘commodities’ (like dinner, or a sailing excursion) are what enters individuals’ utility functions, and that the production of these commodities requires the input of both material goods and time. In this approach, ‘leisure’ therefore disappears as a distinct category, somewhat replaced by the concept of a ‘time-intensive commodity’. The Becker perspective has important implications for the type of leisure activities that people are predicted to choose. Personal time is, essentially, the only input into commodities like contemplation or conversation or the pure enjoyment of peace and quiet – so their cost is just the opportunity cost of time

(that is, the wage rate). The cost of goods-intensive non-work commodities (like speedboat racing) depends partly on the cost of those material goods. When (if) the wage rate rises, time-intensive leisure commodities increase in relative price compared with goods-intensive commodities. Hence, the Becker prediction is for greater materialism over time.

As well, consuming more ‘commodities’ in the same time period – for example, squeezing a tennis game and a sail and dinner and a night at the opera into the same day – is seen in the Becker model as representing an increase in the ‘productivity of consumption time’ (and more is always better), but some would also describe this as a more frenetic lifestyle. Winston (1987, p. 160) has commented that ‘the most serious casualty [in Becker's approach] was loss of the sense of a leisurely and controlled pace that produces genuine satisfaction'.’

However, Becker's approach has not, in fact, been much used. The straightforward work–leisure dichotomy continues to dominate economics journals. The pleasures of non-work time and the marginal disutility of labour were stressed by Marshall (1920, p. 117) many decades ago, and they continue to be the dominant framework today. Can one – should one – expect this constancy of perspective among economists to persist?

Social leisure and the coordination problem

One of the peculiarities of the traditional ‘leisure demand–labour supply’ perspective is its individualism. If utility really did depend only on the quantity of consumption goods and number of non-work hours experienced by individuals, a person's level of utility would be unaffected by solitary confinement, or by any other configuration of social interaction. However, time spent in isolation is, for most people, pleasurable only in small doses. Although one can choose to be alone, relatively few leisure activities are intrinsically asocial. Most leisure activities can be arranged on a continuum of ‘teamness’, and the vast majority of them are distinctly more pleasurable if done with others. Playing softball or soccer are activities that make no sense if done alone. Singing to oneself may be something done in the shower, but singing with a choir is generally a different level of experience. Travelling to exotic foreign places or going for a walk are activities which are usually more pleasurable if done with a companion. Reading a novel is certainly solitary, but many people also like to talk about it afterwards, either formally in a book club or informally with friends over dinner.

To list these different possible leisure activities is to underscore the variety of leisure tastes that individuals have. This variety creates, for each individual, the problem of locating somebody congenial to play with, and scheduling the simultaneous free time to do so. The basic problem with wanting to have a social life is that individuals cannot do it unilaterally – arranging a social life involves a search process which is constrained by the social contacts available to each person, and by the availability of other people. This interdependence of leisure has generated a new literature, with a set of new insights. Corneo (2005), for example, contrasts privately consumed leisure time (watching television) and socially enjoyed leisure (which requires investment in relationships). Across nations, average hours of television watching are *positively* correlated with average working time. Corneo explains this in terms of the strategic complementarities that arise in the organization of social leisure. If these complementarities are strong enough, equilibria with little social leisure but long hours of work and television viewing, and equilibria in which there is much social leisure along with short hours of work and television viewing,

are both possible. Although workers will prefer the higher wages and lower hours of work of the latter, capitalists will prefer the former, since they realize a higher rate of return on their capital stock when total hours of work increase. And if desired working hours are conditional on what others do, individuals need coordination devices to ensure that social leisure is feasible – such as public holidays, a common weekend or working hours regulation – which implies a potentially crucial role for the state and for the relative power of workers and capitalists in influencing public policy.

Jenkins and Osberg (2005) argue that, although solo television watching is certainly feasible, companionship may nonetheless increase the utility derived from the activity. Their emphasis is on modelling more explicitly the constraints involved in locating leisure companions. They argue that the leisure time choices of household members depend on the opportunities for associational life that exist outside the household, and they show that the likelihood of associational activity for persons of a given age group depends on the percentage of persons in other age groups that also engage in that activity. They note that economic models of marriage have discussed the interdependence of spouses in income and material consumption, but it is also plausible that an important reason for marriage is that couples may like spending time together. Like Hamermesh (1998; 2002), they provide evidence on the synchronization and scheduling of spousal work and leisure time.

What are the implications of these new models of social leisure? From a theoretical perspective, the emphasis on the social nature of leisure opens up a whole new set of coordination issues – there is certainly no presumption that individualistic decision making will automatically produce a socially optimal equilibrium. However, the new models of social leisure nest the old labour–leisure choice perspective, since the option of ‘solo leisure’ is always there (albeit now one of several alternatives).

Kuhn (1970) argued that paradigms are replaced when they confront an important empirical anomaly that they are unable to resolve and when a more encompassing alternative theoretical perspective becomes available. The empirical fact which is now forcing a reconsideration of the analysis of leisure is the huge size of cross-national differences in the trend and level of non-work time. From 1980 to 2000, for example, average annual working hours per adult (ages 15–64) rose by 234 hours in the United States to 1,476 hours, but fell by 170 hours in Germany to 973, and by 210 hours in France to 957 (see Osberg, 2003a). By 2000, the cross-sectional difference was huge – non-work time per adult per week was some 9.7 hours greater in Germany, and 9.9 more hours greater in France, than in the United States. In principle, an increase in hourly wages increases both potential income and the opportunity cost of leisure, so the demand for a normal good (like leisure) may rise or fall depending on the relative size of income and substitution effects. However, why should one be larger in Europe and the other larger in America? It is just not very satisfactory to say that ‘tastes differ’.

Cross-country differences in average leisure time are due in part to inter-country differences in probability of employment, in part to differences in common entitlements to paid vacations and public holidays, and in part to differences in the usual hours of work of employees. Trends in these three components are driven by distinctly different processes – the number of paid public holidays is, for example, determined by a set of political processes quite different from the determinants of individual decisions to enter the workforce and to work specific hours. A robust debate has emerged over the causes of these differences in total leisure time (for example, Bell and Freeman, 2001; Alesina, Glaeser and Sacerdote, 2005) – but it is clear that these differences are large enough to motivate both a concern over their implications and a discontent with the traditional labour–leisure choice model.

It has long been acknowledged that one reason why GDP per capita is a poor measure of economic well-being is that it does not recognize that leisure time has any value at all. If, as in the comparison of the United States with Germany or France, greater per capita GDP is obtained primarily from greater average working time, a comparison of economic well-being should measure both the cost of forgone individual leisure and the cost of the externality on the marginal utility of each individual's leisure as the decrease in the leisure time of everyone else impedes the feasibility of leisure time matches.

When (by increasing the availability of potential leisure matches) the choice of more leisure time by some individuals has a positive externality for other persons, there can be multiple equilibria in labour supply, in which the ‘high work’ equilibrium has unambiguously lower total utility. Societies which are better able to coordinate the level and timing of paid working hours may be better off in aggregate, because they enable their citizens to enjoy more satisfying social lives. To be specific, the leisure externality hypothesis suggests that Americans may work more hours than Europeans partly because they are more likely to have less satisfying social lives – because other Americans are also working more hours – and that they are worse off as a result.

Moreover, if authors such as Putnam (1993; 2000) and the OECD (2001) are correct in stressing the dependence of social capital on associational life and the importance of social capital for social and economic development, the costs of a high-work/low-social life equilibrium may be substantial, in terms of market income as well as utility. Knack and Keefer (1997) are representative of an empirical literature which argues that localities with an active civic society and associational life (and more generally a dense network of social ties among individuals, and a high level of trust) have higher growth rates of GDP per capita. This relationship has been argued to be due to a number of possible influences: for example lower transactions costs in capital, labour and product markets, more effective governance, lower costs of crime, labour conflict and political uncertainty, better health outcomes, and so on (see Osberg, 2003b). Whatever the channel of influence, it suggests that, although working longer hours may accelerate growth in GDP per capita in the short run, both income and social life may suffer in the longer run. There may be some wisdom in the old saying that: ‘All work and no play makes Jack a dull boy.’

See Also

- external economies
- labour supply
- social capital
- time use

Bibliography

Alesina, A., Glaeser, E. and Sacerdote, B. 2005. Work and leisure in the U.S. and Europe: why so different? Working Paper No. 11278. Cambridge, MA: NBER.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–517.

Bell, L. and Freeman, R. 2001. The incentive for working hard: explaining hours worked differences in the US and Germany. *Labour Economics* 8, 181–202.

Corneo, G. 2005. Work and television. *European Journal of Political Economy* 21, 99–113.

Di Tella, R., MacCulloch, R. and Oswald, A. 2003. The macroeconomics of happiness. *Review of Economics and Statistics* 85, 809–27.

Dickinson, D. 1999. An experimental examination of labor supply and work intensities. *Journal of Labor Economics* 17, 638–70.

Frey, B. and Stutzer, A. 2002. *Happiness and Economics: How the Economy and Institutions Affect Well-Being*. Princeton: Princeton University Press.

Hamermesh, D. 1998. When we work. *American Economic Review* 88, 321–5.

Hamermesh, D. 2002. Timing, togetherness and time windfalls. *Journal of Population Economics* 15, 601–23.

Jenkins, S. and Osberg, L. 2005. Nobody to play with? The implications of leisure co-ordination. In *The Economics of Time Use*, ed. D. Hamermesh and G. Pfann. Amsterdam: Elsevier.

Juster, T. and Stafford, F., eds. 1985. *Time, Goods, and Well-Being*. Ann Arbor: Institute for Social Research, University of Michigan.

Knack, S. and Keefer, P. 1997. Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics* 112, 1251–88.

Kuhn, T. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan, 1961.

OECD (Organization for Economic Co-operation and Development). 2001. *The Well-being of Nations: The Role of Human and Social Capital*. Paris: OECD.

Osberg, L. 2003a. Understanding growth and inequality trends: the role of labour supply in the U.S.A. and Germany. *Canadian Public Policy* 29, Supplement 1, S163–S183.

Osberg, L. ed. 2003b. *The Economic Implications of Social Cohesion*. Toronto: University of Toronto Press.

Putnam, R. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Putnam, R. 2000. *Bowling Alone: The Collapse and Revival of the American Community*. New York: Simon and Schuster.

Winston, G. 1987. Leisure. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. Basingstoke: Palgrave.

How to cite this article

Osberg, Lars. "leisure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000063> doi:10.1057/9780230226203.0956

The New Palgrave Dictionary of Economics Online

Lenin, Vladimir Ilyich [Ulyanov] (1870 – 1924)

Meghnad Desai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; falling rate of profit; Hobson, J. A.; imperialism; Lenin. V. I.; Marx's analysis of capitalist production; New Economic Policy (USSR); planning; socialism; state capitalism

Article

Vladimir Ilyich Ulyanov, who wrote and gained fame under the pseudonym Lenin, was born in April 1870, the second son of a Russian provincial official in Simbirsk (now Ulyanovsk). After the arrest and execution of his elder brother Alexander, in 1887 for alleged terrorist activity, Lenin became increasingly active in political study groups at Kazan, Samara and St Petersburg. He came to identify himself with the Marxist rather than the populist (Narodniki) stream in these study groups. He played an active part in the early theoretical debates between these two streams on the future course of Russia's economic and political development. At the time of the founding of the Russian Social Democratic Labour Party (RSDLP) in 1898, he was already known as its best young theorist. A split in the RSDLP took place in 1902 and Lenin became identified as the leader of the majority (Bolshevik) faction. He spent much of the early years of the 20th century in exile in London, Paris and Zurich. He returned to Russia in April 1917 after the February Revolution had initiated the post-tsarist phase of Russian politics. Lenin, unlike his fellow party members, correctly foresaw the instability of the political situation in which an unelected liberal democratic cabinet uneasily shared power with the federation of popularly elected factory committees (Soviets). He launched the Bolsheviks on a strategy of revolutionary rejection of the government and a platform of peace in the World War at any price. His analysis proved correct when in November 1917 the Bolsheviks won a majority in the All Russian Congress of Soviets and took power. Lenin led the communist government from that day until illness forced his withdrawal from active politics in March 1923. He died in January 1924.

Lenin's economic writing is extensive, comprising books, pamphlets, newspaper articles and occasional speeches (see Desai, 1986, for a full bibliography). His contributions can be placed under three headings: analysis of Russia's capitalist development in the period 1880–1900; the analysis of the developments in world capitalism in the period 1900–1916, where his concept of imperialism as a form

of monopoly capitalism was an innovation; and lastly as a Marxist policy maker during the period 1917–23.

The development of capitalism in Russia

Lenin's book of this title published in 1898 is a substantial piece of work which traces the growth of commercial relations and specialization in agriculture leading to an erosion of the traditional communal forms. On the industrial side, Russia's late arrival entailed an active role for the tsarist state in fostering industrialization and an influx of foreign capital to finance the development. This meant that Russia, although a newly industrializing country in the 1890s, had a larger proportion of its industrial labour force in large factories than older industrialized countries like Britain. Lenin saw these as predictable consequences of rapid capitalist growth which made any going back to pre-capitalist communal forms of village organization impossible. The growth of large factories also meant concentration of workers in a few places, facilitating their combination in trade union activities. These economic circumstances – the growth of commercial relations in the countryside and of concentrations of the urban proletariat – dictated for Lenin the political strategy of a socialist party which hoped to win power by mass organization. Lenin's theory of the development of the democratic political movement follows the economic stages quite closely. In this sense he can be said to have developed an economic framework for a Marxist political theory. *The Development of Capitalism in Russia* is even to this day the only comprehensive economic history of a country from a Marxist perspective.

Imperialism, the highest stage of capitalism

In 1916, Lenin wrote his well-known economic pamphlet of this title. The background was provided by the First World War, which had broken out two years previously with enthusiastic participation by the working people of various combatant nations and the connivance of the socialist parties. The ‘betrayal’ by the workers and their political leaders was one factor in Lenin's urge to explain these events. The second urge was perhaps provided by a desire to integrate the facts of a war into a Marxist theory of the long-run development and eventual breakdown of capitalism.

Marx had predicted a tendency for the rate of profit to fall as capitalist development proceeded. Among the forces which may counteract this tendency was an increasing concentration in industry and the emergence of larger industrial units. In 1907, Hilferding in his *Finance Capital* had provided a theory and empirical evidence for the increasing integration of bank finance and industrial capital. The formation of trusts and cartels was helped by banks willing to finance mergers and controlling and interlocking equity holdings. Marxist economists saw the 20th century as entering a monopoly phase of capitalism in contrast to the competitive phase that Marx had written about.

Lenin's achievement is to add to the Marx–Hilferding account an international economic and political element. One part of his theory came from Hobson's *Imperialism*. As an underconsumptionist, Hobson linked the fight over African and Asian territory in the last decades of the 19th century among European nations to the search for outlets for surplus which could not be sold at home. Hobson took the view that this imperial search was irrational. Lenin, as a Marxist, saw the irrationality as a systematic functional element in a world of monopoly capital economies each of which was trying to stave off the falling rate of profit by exporting. The battle for markets could not however take place in a politically neutral

context as envisaged by competitive economic theory. Large cartels and monopolies gave a few leading bankers and industrialists influence with the political governments of their country. The battle for markets thus became a struggle between developed capitalist nations for territory. It was the struggle for territory as a surrogate for markets which led to military confrontation between the major industrial nations and hence war. War was not however predicted to be a satisfactory solution to the problem of markets or of profitability. It was likely in Lenin's view to be the harbinger of proletarian uprising against the system in these countries which would end it.

Thus Lenin blends international political developments into a Marxian theory of capitalist development. Imperialism in Lenin's definition is the entire set of unequal economic relations between capitalist countries – between rival mature capitalist countries fighting for markets as well as between mature countries and developing economies which become their markets. Formal political control by one nation over another is not a necessary element in Lenin's view of imperialism. Although immensely influential in the interwar years due to Comintern orthodoxy, this theory has come under some attack recently (Warren, 1980). It lacks a coherent analytical theory of how monopoly capital differs from competitive capitalism and its empirical predictions proved only temporarily true when a series of political uprisings took place in Europe after the First World War. These uprisings did not mature into a full-scale collapse of capitalism, which continues many decades after Lenin foresaw its highest phase as having been achieved.

Socialist economic policy

As the first Marxist to lead a government, Lenin had to formulate practical economic policy. Given the notorious lack of discussion of socialist economic policy in Marx's writings, Lenin had to improvise. Two notions stand out as his distinctive contribution to this area. First, in his description of the post-revolutionary Russia as a transitional state from capitalism to socialism. During this transition, state capitalism was seen by Lenin as an advance upon private capitalism in as much as the political state was not a capitalist one but a workers’ state. Lenin used the wartime German economic organization as the ideal of a fully integrated single economic unit which a planned socialist economy could beneficially emulate. Second, in the return to normality after the Civil War – in his pamphlet ‘The Tax in Kind’ – Lenin sketched a theory for the role of trade in reviving economic activity. The key was to move from a forced requisition of food surpluses to a policy of tax in kind and encouraging exchange. A revival of agriculture was required for an industrial revival but the terms of trade between the two sectors was a crucial policy variable in this respect. Trade is seen as an antidote to economic bureaucracy in this pamphlet. It was this pamphlet that inaugurated the New Economic Policy which could be said to have lasted from 1921 to 1929.

Selected works

1898. *The Development of Capitalism in Russia*. In *Collected Works*, vol. 3.

1916. *Imperialism; the Highest Stage of Capitalism*. In *Collected Works*, vol. 22.

1921. *The Tax in Kind*. In *Collected Works*, vol. 32.

1960–70. *Collected Works*, 45 vols. Trans. of the 4th enlarged Russian edn, Moscow: Progress Publishers.

Bibliography

Desai, M. 1986. *Lenin as an Economist*. London: Lawrence & Wishart.

Warren, B. 1980. *Imperialism: Pioneer of Capitalism*. London: New Left Books.

Bibliographic addendum

Studies of aspects of Lenin's life and thought continue to be produced because of his importance in world history. Within this massive literature, valuable studies of his ideas include N. Harding, *Lenin's Political Thought*, 2 vols, New York: St. Martins Press, 1977 and 1981; and N. Harding, *Leninism*. Durham, NC: Duke University Press, 1996. An extensive general biography is R. Service, R. *Lenin: A Political Life*, 3 vols, Bloomington: Indiana University Press, 1985, 1991, and 1995. This massive study is synthesized and updated in R. Service, *Lenin*, Cambridge, MA: Harvard University Press, 2000.

How to cite this article

Desai, Meghnad. "Lenin, Vladimir Ilyich [Ulyanov] (1870–1924)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000068> doi:10.1057/9780230226203.0958

The New Palgrave Dictionary of Economics Online

Leontief paradox

Edward E. Leamer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Using 1947 US input–output tables and data on exports and imports, Leontief (1953) found, to the surprise of the profession, that the capital per worker of US exports was less than the capital per worker of US import substitutes. The response to this empirical ‘paradox’ was the formulation of theory that might explain why a capital abundant country had labour-intensive exports. These were the first (confused) steps in an ongoing process of making the theory and data conform sufficiently to enable us comfortably to claim to understand the basis for international trade.

Keywords

factor content of trade; factor mobility; Heckscher–Ohlin–Samuelson trade model; Leontief paradox

Article

The Heckscher–Ohlin–Samuelson (HOS) model of international trade with two factors of production and two commodities implies that a country will export the commodity that is produced intensively with the relatively abundant factor. Leontief (1953) discovered, to the surprise of the profession, that 1947 US exports were more labour-intensive than US imports in the sense that the capital per man required to produce a $1 million of exports was less than the capital per man required to produce a $1 million in import substitutes. This seemed to conflict sharply with the presupposition that the USA was abundant in capital compared with labour. Leontief's finding was so startling that it has been called a ‘paradox’, even though the result amounted to at most a single contradiction of the theory and even though no alternative model could be said to conform better with the facts.

Leontief's finding preceded and apparently stimulated a search of great breadth and intensity for a new theory of trade that could account for his result. It is in fact difficult to find another empirical result that has had as great an impact on the intellectual development of the discipline. Among the explanations of the finding are: (*a*) high productivity of US workers; (*b*) capital-biased consumption; (*c*) factor-intensity

reversals; (*d*) tariffs; (*e*) abundance of natural resources; (*f*) abundance of human capital; (*g*) technological differences. These developments are surveyed in Chacholiades (1978, pp. 298–306).

It is surprising in retrospect that no one thought to examine the theoretical foundation for Leontief's inference that the factor content of US trade revealed the United States to be scarce in capital compared with labour, though a clear theory of the factor content of trade was not laid out until Vanek (1968). Vanek's model of the factor content of trade was first used in an overlooked article by Williams (1970) to criticize Leontief's inference. The very simple theoretical foundation for the Leontief calculation was clearly laid out in Leamer (1980), which shows that Leontief's data in fact reveal the United States to have been abundant in capital compared with labour.

Theoretical relationships that can serve as a foundation for studying the relative factor abundance revealed by international trade are the Heckscher–Ohlin–Vanek equations. These equations are derived from the simple identity that net exports of the services of a factor f are the difference between home supply and home demand: $T_f = X_f - M_f = S_f - D_f$, where T_f is the amount of factor *f* embodied in net exports, X_f is the amount of factor f required to produce the exported commodities, M_f is the amount required to produce the imported commodities, S_f is the domestic supply and D_f is the domestic demand. This identity is given empirical content by assuming identical homothetic tastes which implies that domestic demand for factor *f* is proportional to world supply, $D_f = sW_f$, where W_f is the world supply and *s* is the country's share of world consumption. With the use of this assumption, the net export equation can be written as

$$T_f / S_f = 1 - s(W_f / S_f).$$

In words, net exports as a share of domestic supply is positively related to factor abundance defined as the share of the world's total supply S_f / W_f. Accordingly, the relative scarcity of the factors is revealed by the ordering of the net export ratios T_f / S_f. Leamer (1980) shows that, although the net export of both capital and labour services were positive in 1947, the share of domestic supply of capital that was exported exceeded the share of labour exported, and consequently the United States was revealed by trade to be relatively abundant in capital compared with labour. In addition, Leamer (1980) shows that Leontief's finding that the exports were more capital intensive than imports is compatible with either ordering of factor abundance.

This fully resolves the apparent paradoxical ordering of capital and labour abundance, but a new problem arises. Brecher and Choudhri (1982) note that, if net exports are positive, the overall consumption share *s* must be less than the abundance ratio S_f / W_f. If trade is balanced, the consumption share is the ratio of home to world GNP, s=GNP/GNP$_w$. The inequality $S_f / W_f > s = GNP / GNP_w$ can be rewritten as $GNP_w / W_f > GNP / S_f$. Thus the United States is revealed by its positive net exports of labour services embodied in commodities to have had a per-capita GNP that is less than the rest of the world. Even after adjusting for the trade surplus, this is impossible to

square with the facts. Another way of expressing this new paradox is that the positive export of labour services reveals that labour is abundant compared with other resources on the average since the consumption share *s* is an average of the abundance ratios.

It is ironic that this is one of the few empirical findings that can be said to have had a decided impact on the course of the profession and at the same time is based on a simple conceptual misunderstanding. The error that is implicit in Leontief's paradox is the use of an intuitive but false theorem which states that the ordering of capital per man in exports compared with imports reveals the relative abundance of capital and labour. This is true for the simple two-good model, but it is not the case for a multi-commodity reality. There is a lesson to be learned from this experience. Empirical work requires a fully articulated theoretical foundation. Intuition alone is not enough.

Although the precise form that Leontief's calculations took is inappropriate, the calculation of flows of factor services embodied in trade remains an interesting activity since these flows can be used to form a proper test of the Heckscher–Ohlin–Samuelson theorem and since the net effect of trade on the demand for factors of production can be an important input into trade policy that is intended to affect the distribution of income.

As it turns out, measurements of 1967 factor contents of trade reported in Bowen, Leamer and Sveikauskas (1987) rather badly violate the HOS model, thus reinvigorating the message of the Leontief paradox: there is something wrong with this model. One thing that is wrong is emphasized by Trefler's (1995) title: ‘The Case of the Missing Trade’. Given the world's apparent unequal geographic distribution of capital, labour and land, the HOS model suggests that there should be much more trade than actually occurs. Trefler's solution to this puzzle is to allow in the model both home bias in consumption and also international productivity differences (for example, the United States is not so labour-scarce when allowance is made for the intensity of work). Also, Conway (2002) finds problems with the measurement of factor scarcity and calls for the model to include factor-specific differences in domestic factor mobility. It seems likely that we have not seen the end of the search for a model that most fully explains the nature of international trade.

See Also

- Heckscher–Ohlin trade theory
- input–output analysis

Bibliography

Bowen, H., Leamer, E. and Sveikauskas, L. 1987. Multicountry, multifactor tests of the factor abundance theory. *American Economic Review* 77, 791–809.

Brecher, R. and Choudhri, E. 1982. The Leontief Paradox, continued. *Journal of Political Economy* 90, 820–3.

Chacholiades, M. 1978. *International Trade Theory and Policy*. New York: McGraw-Hill.

Conway, P. 2002. The case of the missing trade and other mysteries: comment. *American Economic Review* 92, 394–404.

Leamer, E. 1980. The Leontief Paradox, reconsidered. *Journal of Political Economy* 88, 495–503.

Leontief, W. 1953. Domestic production and foreign trade: the American capital position re-examined. *Proceedings of the American Philosophical Society* 97, 332–49. Reprinted in *Readings in International Trade*, ed. H. Johnson and R. Caves. Homewood, IL: R.D. Irwin, 1968.

Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85, 1029–46.

Vanek, J. 1968. The factor proportions theory: the N-factor case. *Kyklos* 21, 749–56.

Williams, J. 1970. The resource content in international trade. *Canadian Journal of Economics* 3, 111–22.

How to cite this article

Leamer, Edward E. "Leontief paradox." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000179> doi:10.1057/9780230226203.0959

The New Palgrave Dictionary of Economics Online

Leontief, Wassily (1906 – 1999)

R. Dorfman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cobweb cycles; concentration; foreign aid; implicit theorizing; index numbers; input–output analysis; Leontief paradox; Leontief, W.

Article

Wassily Leontief was born on 5 August 1906 in St Petersburg, the only child of an academic family. He studied first at the University of Leningrad, earning the degree of ‘Learned Economist’ in 1925, and then at the University of Berlin (Ph.D., 1928). While working on his doctorate, he was appointed a research economist at the University of Kiel, where he remained for about three years, with a year out to serve as adviser to the Chinese Ministry of Railways in Nanking.

In 1931 he went to the United States to join the staff of the National Bureau of Economic Research, but after only a few months accepted an appointment at Harvard University, where he remained for the following 44 years. During those years he attained worldwide eminence, particularly for the invention and application of input–output analysis. Prominent among the honours he received during those years were election as President of the American Economic Association in 1970 and the Nobel Memorial Prize in Economics in 1973. In 1975 he accepted a chair at New York University, where he spent the remainder of his career.

Leontief had an exceptionally strong training in mathematics and a marked flair for mathematical and geometric reasoning. These qualities were displayed in his earliest papers, in the late 1920s and early 1930s, in which he applied his technical talents to a variety of topics including the estimation of elasticities of supply and demand, the measurement of industrial concentration, the use of indifference maps at a time when they were still novelties to explain patterns of international trade in a two-commodity, two-country model, analysis of the conditions under which cobweb cycles would converge or would expand explosively, and several others. These papers established his reputation as an economic theorist of first rank.

During this same period, he struck a theme that he was to emphasize repeatedly throughout his career: the thesis that economic concepts were meaningless and misleading unless they could be observed and

measured. Thus, in 1936 he studied the significance of index numbers that purported to measure composite concepts such as the aggregate output of an economy or the general price level, and the following year published his famous diatribe against ‘implicit theorizing’, that is, explaining phenomena by introducing ill-defined concepts (the economist's version of Molière's doctor who attributed the effect of sleeping potions to their dormative propensities). Eleven years later, he returned to the measurement of aggregates much more profoundly and fruitfully in his ‘Introduction to a Theory of the Internal Structure of Functional Relationships’, which developed the mathematical conditions in which a single aggregate or index could replace a mass of detailed data without loss of information. And much later he devoted his presidential address to the American Economic Association to decrying ‘Theoretical Assumptions and Nonobserved Facts’ (1971).

These two characteristics – adroitness at mathematical expression and analysis and insistence that theoretical concepts be implementable – congealed in Leontief's major achievement, the invention, development, and application of input–output analysis. As a purely theoretical construct, input–output analysis had a long genealogy before Leontief began his work on it, around 1933. In the 18th century, François Quesnay used his *Tableau économique* to illustrate the relationships between agriculture and other sectors of the economy. A hundred years later, Marx demonstrated the relationships between the capital-goods and consumers’ goods departments of an economy by a very similar two-sector table. The most important predecessor, however, was Walras's formulation of the general equilibrium of an economy, which employed a concept that is very similar to Leontief's input–output coefficients. In addition, as Leontief discovered after input–output analysis was well known, H.E. Bray had published essentially the same equations in 1922, and R. Remak had discovered them again in 1929.

The algebraic theory of input–output analysis had been explored by a number of late 19th-century algebraists, particularly by E. Frobenius and O. Perron, for whom the basic theorems have come to be named. All of these preceding theories expressed fundamental, abstract theoretical concepts; none could be used to specify the relationships among the sectors of an actual economy.

But throughout his career, Leontief has insisted that the task of a theorist only begins with the proposal of a well-formulated theory; the central task is to show that the theory can be applied to real economies, that it leads to interesting predictions about the behaviour of those economies, and that those predictions can be checked and found to be reasonably accurate. This radically operational point of view led Leontief to his critical contribution: the perception that the coefficients that express the relationships among the sectors of an economy can be estimated statistically, and that they are sufficiently stable so that they can be used in comparative static analyses to give quantitative estimates of the effects of different economic policies, taking into account their reverberations throughout the economy along with their effects on the industries affected in the first instance.

It is almost impossible now to appreciate the task of confirming these conjectures in the early 1930s. Input–output computations depend on inverting large matrices; the most powerful computing machines in existence then were punch-card machines that could multiply, after a fashion, but could not divide. Solving a half-dozen simultaneous linear equations was a formidable calculation; Leontief envisaged systems that numbered in the hundreds.

Input–output analysis also required data of an unfamiliar type – coefficients specifying the amounts of various raw materials and intermediate goods required per unit of product in each sector. The US Census of Manufactures included many of these coefficients, but by no means all. The remainder had to be compiled laboriously from trade journals and scattered sources.

Furthermore, the underlying assumption of the method, that the input–output coefficients remained essentially constant for substantial periods, was hard to reconcile with one of the main tenets of the theory of production – that factors of production were substituted for one another quite sensitively in response to price changes.

Beginning around 1933–4, Leontief concentrated on overcoming these difficulties by compiling coefficients for a 44-sector input–output table – about 2,000 coefficients – and making plans for their analysis. Since the solution of 44 simultaneous equations was far beyond the realm of the possible, the 44 sectors were consolidated into a scant ten for computational purposes. To check on the stability of the coefficients, tables were to be compiled for 1919 and 1929.

The first result of this study, ‘Quantitative Input and Output Relations in the Economic System of the United States’, appeared in 1936. Its centrepiece was a 41-sector input–output table for the United States in 1919, presenting the intersectoral flow coefficients along with sources and methods of estimation. The next year, Leontief published ‘Interrelation of Prices, Output, Savings and Investment’. In the interim, he had made contact with Professor John B. Wilbur of the Massachusetts Institute of Technology, who had just invented an analog computer that could solve systems of up to nine linear equations. Accordingly, Leontief aggregated his 41-sector table into ten sectors and used Wilbur's computer to calculate the inverse. This was the first Leontief-inverse ever computed, and probably the first use of a large computer in economics or other social science.

By 1941, a parallel 41-sector table had been compiled for 1929 and the inverse of a ten-sector aggregation of it had been computed. The two tables were presented and compared in Leontief's first monograph, *The Structure of American Economy, 1919–1929*. The comparisons were intended to test whether the input coefficients were stable enough to yield useful empirical predictions. The comparisons were indecisive, in part for lack of a clear standard for judging the stability of the estimated coefficients. The monograph did establish, however, that it was feasible to compile the raw data needed for an input–output table and to compute coefficients and an inverse table that appeared to make good economic sense. The importance of such tables for economic planning was recognized almost immediately. Within a few years, the US Bureau of Labour Statistics, with Leontief as a consultant, constructed a 400-sector table for projecting post-war employment by major industries, and the method was being applied all over the world for constructing economic development plans.

Leontief remained in the forefront of these developments. By 1944 he had calculated a table of input coefficients for 1939, comparable with the earlier two tables, and found a satisfactory degree of stability for most of the coefficients extending over two decades. Using this up-to-date table, he published a sequence of three important papers in the *Quarterly Journal of Economics* for 1944 and 1946 exemplifying the use of input–output analysis for estimating the effects of exogenous disturbances on output, employment, wages, and prices in individual sectors.

In 1948, Leontief established the Harvard Economic Research Project as a centre for applying and extending input–output analysis. He became director of the Project, and headed it for the next 25 years. He was particularly active in developing interregional input–output analysis and in introducing capital–coefficient matrices to derive the investment implications of changes in final demand and, thereby, to use input–output analysis to generate growth paths as well as static equilibriums of economic systems. This work led to two books, *The Structure of American Economy 1919–1939*, in 1951, and *Studies in the Structure of the American Economy*, in 1953, as well as several international conferences and a score of papers and articles. Probably the most striking discovery of this period of work has come to be called

‘the Leontief paradox’, the finding that, when indirect as well as direct input requirements are taken into account, American exports are more labour-intensive and less capital-intensive than American imports, although the United States is exceptionally well endowed with capital and has exceptionally high real wages.

Leontief and the staff of the Harvard Economic Research Project devised and implemented numerous other applications of input–output analysis. They included estimates of the inflationary impact of wage settlements, calculations of the direct and indirect effects of armament expenditures on the individual sectors of the economy, and methods for projecting the growth-paths of the sectors in a developing economy and for estimating capital requirements for economic development.

In the middle 1970s, Leontief became persuaded that, while competitive markets might guide an economy to a socially efficient equilibrium if given sufficient time, the process would be likely to be very protracted and unduly wasteful of mistakenly invested resources. Economic growth and efficient adjustments would be promoted better by establishing an economic planning board that would work out a number of detailed growth possibilities based on input–output analyses. The ultimate choice among these possibilities would be made by a political process. He advocated this type of indicative planning in a number of articles in *The New York Review of Books*, the *New York Times* ‘op-ed page’, and other general interest periodicals.

Leontief subsequently turned to the problems of worldwide economic growth, its environmental impact, its demands on the world's base of natural resources, and particularly on its implications for relations between the economies of the so-called First and Third Worlds. Under the sponsorship of the United Nations, he directed a study of the evolution of the world economy until the year 2000, based on a multiregional input–output model consisting of 15 regions, each comprised of 45 sectors, and linked by balanced trading relationships. This is, perhaps, the most ambitious input–output study yet undertaken. The results were published as *The Future of the World Economy* (1977). It found that, under a wide range of plausible assumptions, little progress would be made in closing the gap between the industrial and the developing regions unless current policies concerning international trade and finance were changed drastically in the directions of increased multinational aid and an increased flow of imports from the Third World to the First.

Leontief was a leader in improving the computational methods of economics, beginning with his use of Wilbur's analog equation solver in 1936. Subsequently he inverted input–output matrices on Howard Aiken's early Mark I and Mark II computers, the immediate predecessors of the electronic computer. In the 1980s the very large matrices required by his world economic models led him to be the first economist to use the so-called supercomputers and to apply parallel-processors and other highly efficient methods of computation.

Throughout his career, Leontief took an active interest in the education of the next generation of scholars. While at Harvard he served for 11 years as chairman of the Society of Fellows, the foundation that provides three-year, duty-free fellowships to promising young scholars, to enable them to reside at Harvard and pursue whatever interests they choose. He delighted in presiding over the weekly dinner meetings of the Society and leading conversations that range over all the fields of interest represented at the table.

See Also

- input–output analysis

Selected works

Leontief's principal scientific contributions can be found in four volumes:.

1941. *The Structure of American Economy, 1919–1929*. Cambridge, MA: Harvard University Press and later editions.

1953. (With others.) *Studies in the Structure of the American Economy*. New York: Oxford University Press.

1966. *Essays in Economics: Theories and Theorizing*. New York: Oxford University Press, and later editions.

1977. *Essays in Economics*, Vol. 2. White Plains, NY: M.E. Sharpe.

Articles exemplifying Leontief's wide range of interests and activities can be found in:

Bulletin of the American Mathematical Society, April 1947.

New York Review of Books, 10 October 1968, 21 August 1969, 4 June 1970, 7 January 1971, 20 July 1972, 4 December 1980, 12 August 1982.

New York Times, op-ed page, 14 March 1974, 24 March 1977, 6 March 1979, 5 April 1981, 19 September 1983.

Scientific American, April 1961, September 1963, April 1965, September 1980.

Bibliographic addendum
A range of essays evaluating and extending Leontief's work may be found in

Dietzenbacher, E. and Lahr, M. 2004. *Wassily Leontief and Input–Output Economics*. Cambridge: Cambridge University Press.

Wood, J. 2001. *Wassily Leontief: Critical Assessments*. London and New York: Routledge.

How to cite this article

Dorfman, R. "Leontief, Wassily (1906–1999)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000070> doi:10.1057/9780230226203.0960

The New Palgrave Dictionary of Economics Online

Lerner, Abba Ptachya (1905 – 1982)

T. Scitovsky

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost-push inflation; demand-pull inflation; distributional optimality; excess-claims inflation; excess-demand inflation; expectational inflation; expected and unexpected inflation; factor-price equalization; full employment; functional finance; incomes policy; inflation; international trade theory; Keynes, J. M.; Lerner, A. P.; low full employment; marginal cost pricing; Market Anti-Inflation Plan (MAP); market pricing; market socialism; monopoly; natural rate of unemployment; optimality; optimum currency areas; Organization of Petroleum Exporting Countries (OPEC); Pareto efficiency; perfect competition; socialist free enterprise; stagflation; unemployment; welfare economics

Article

Lerner was one of the last of the great non-mathematical economists and certainly one of the most original, versatile and prolific members of the profession. Born in Rumania, raised from early childhood in the Jewish immigrant quarter of London's East End, he went to rabbinical school, started work at 16, working as tailor, capmaker, Hebrew School teacher, typesetter, and then founded his own printing shop. When that went bankrupt at the onset of the Great Depression, he enrolled as an evening student at the London School of Economics to find out the reason for his shop's failure. There, his outstanding logical faculties soon became evident and won him all the available prizes and fellowships, one of which took him to Cambridge to study with Keynes. He published many major articles already as an undergraduate, was appointed temporary assistant lecturer at the London School of Economics in 1935, assistant lecturer in 1936, and in 1937 a Rockefeller fellowship took him to the United States, where he remained, although his restlessness kept him from settling at any one university for more than a few years.
Lerner was a lifelong socialist, advocate of market pricing for its allocative efficiency, and believer in private enterprise, whose offer of private employment he considered an essential safeguard of individual freedom. That unusual combination of principles accounts for Lerner's loneliness and political isolation. In his economics, however, he knew how to reconcile those principles. His reconciliation of the first two made him into one of the founders (along with Oskar Lange) of the theory of market pricing in the

decentralized socialist economy, and he sought to reconcile the first and third principles by advocating what he called socialist free enterprise: ‘the freedom of both public and private enterprise to enter any industry on fair terms which, in each particular case, permit that form to prevail which serves the public best.’

Although Lerner's ambition was to improve the economy, not economics, he made many, often fundamental contributions to economic theory, mainly in the fields of welfare economics, international trade and macroeconomics but also in the theories of production, capital, monopoly, duopoly, spatial competition and index numbers. Furthermore, and hardly less important, he made generous use of his geometrical skill and genius for exposition in tidying up and clarifying other people's ideas. As a result, a number of important economic theorems and ideas, though first stated by others, became the profession's common property in Lerner's simpler and clearer formulations. An important example of that is the well-known rule that marginal cost pricing is a condition of welfare optimality. Another example is his definitive proof (Lerner, 1936a) that in the two-country, two-commodity model, export and import duties have identical consequences if their proceeds are spent in the same way.

In welfare economics, one of his first articles (Lerner, 1934a) not only introduced the notion that monopoly is a matter of degree, whose extent is best measured by the excess of price over marginal cost, but in the process also provided the first complete, comprehensive and clear statement and discussion of the nature and limitations of Pareto optimality, and of the equality between price and marginal cost and between price and marginal value product as necessary conditions of optimality. All that, along with Lerner's many papers on market pricing under socialism, was restated, elaborated and extended in his 1944 *The Economics of Control: Principles of Welfare Economics*.

That work, Lerner's best book, became and remains the most comprehensive non-mathematical text on welfare economics. Although written in the style of a handbook, with its propositions presented as rules for the planners and plant managers of a decentralized socialist economy to follow, the book is better described by the second than by the first half of its title. For most of those rules are nothing but the first-order conditions of optimality, presented with great care, clarity and completeness but without a hint at the practical obstacles in the way of putting them into actual practice. As a text on welfare economics, however, it is exceptionally meticulous and complete, it extends the scope of the welfare principle from resource allocation narrowly defined to taxation, macroeconomics and international trade and finance, and it contains the first logically based analysis of distributional optimality. Moreover, since a socialist economy, for Lerner, meant the use of private enterprise in some sectors, state-owned plants in others, depending on which was the more efficient in each, his guidebook for socialist planners also discusses why and when perfect competition leads to optimality and why and when real-life competition falls short of being perfect.

In the field of international trade theory, Lerner derived Samuelson's celebrated factor-price equalization theorem 15 years before Samuelson in a 1933 unpublished seminar paper printed only 19 years later (Lerner, 1952a). His elegant and ingenious resolution of a 19th-century controversy over the identity of import and export duties has already been mentioned; he devised (Lerner, 1932; 1934b) the standard geometry of the two-country, two-commodity model, which is well known from a whole generation of textbooks; and he was the first to raise and deal with the question of ‘optimum currency areas’ in his 1944 *Economics of Control*.

Most of Lerner's innovations in microeconomics and international trade theory were so basic and so useful that they promptly became integral parts of every economist's standard equipment. That is why it

is hard to appreciate, at this late stage, the striking originality and elegant simplicity of his logic. One gets a glimpse of that by looking at his almost unknown proposal of how to counter OPEC's raising of the price of oil (Lerner, 1980a). He proposed the imposition of a variable import duty on oil (which he called extortion tax), whose level would always match the producer's profit margin, thereby rising and falling with the oil price and being higher on imports from high-priced and lower on those from low-priced producers. Since such a tariff would make consumers face much larger price changes than those decided upon by OPEC and much greater price differentials than those set by the different oil exporters, it would also make consumers' responses to those price changes and differentials correspondingly greater, thereby raising the price elasticity of demand for oil as it appears to producers. That would lower OPEC's monopoly power and so its profit maximizing monopoly price, and it would increase the rewards and the temptation for OPEC members to break up the coalition by defecting from it.

In macroeconomics, Lerner did as much as anyone to clarify, extend and popularize Keynes's *General Theory*; he was the first to recognize the inflationary implications of employment policies, the first to analyse in depth and in detail the causes and nature of inflation, and to propose a remedy for stagflation. Lerner wrote the first article (1936b) to make Keynes's employment theory simple and generally intelligible, and in two short papers clarified Keynes's 'user cost' and 'marginal efficiency of capital' concepts (1943b; 1953). He wrote an interesting book (1951) to summarize and significantly extend Keynes's employment theory; he published an enlightening paper to explain the *General Theory*'s obscure Chapter 17 (1952b), thereby clearing up the complex role wage rigidity plays in rendering underemployment equilibrium possible; and he was the person best to elucidate the relation between macroeconomics and microeconomics by representing them as the two limiting cases of a more general type of economic analysis (1962).

Next to his work on welfare economics and international trade theory, Lerner's best known and most shockingly new contribution was his introduction of the idea of 'functional finance' (1943a; also restated in 1951, and in his 1944 *Economics of Control*), whose advocacy of Keynesian employment policies exposed the latter's logical implications and revolutionary nature. To careless readers, it also seemed like a wildly inflationary doctrine, although Lerner's concern over inflation and over the inflation effects of employment policies antedate everybody else's by many years.

Lerner's extensive work on inflation began with his distinguishing between low and high full employment (Lerner, 1951). High full employment is that beyond which further demand expansion presses against supply limitations and creates overspending (demand-pull) inflation; low full employment is the employment level below which the price level is stable. Levels of employment between the low and high full-employment levels create administered (cost-push) inflation, owing to labour's excessive bargaining strength. His 'low full employment' therefore is a forerunner (by 17 years) of Friedman's 'natural rate of unemployment'.

Lerner's theoretical papers on inflation contain many pioneering insights. One is his sharp analytic distinction between overspending or excess-demand inflation and administered or excess-claims inflation (1958; 1972), of which the former does, but the latter (according to him) does *not*, call for fiscal and/or monetary restraint. He later added a third category, expectational inflation (1972), which he also called defensive inflation to differentiate it from the aggressive nature of excess-claims inflation – arguing that incomes policy is effective against the former but ineffective against the latter. Another and well-known distinction which Lerner was the first to draw was that between expected and unexpected inflation (1949).

Since Lerner's heart was in reform, not in analytic niceties, his many discussions of inflation were just a preamble for working out a plan to control the main economic problem of his time, stagflation, that is, the combination of unemployment and inflation, which he considered characteristic of administered or excess-claims inflation. Restrictive policies were to him an inadmissible cure for that type of inflation, because he considered the creation of unemployment a prohibitive cost. Incomes policies he judged ineffective against all but expectational inflation, and he was too ardent a believer in the pricing mechanism to argue for wage and price controls. He wanted to stabilize the general price level without impeding the free movement of individual prices and wages. To accomplish that, he devised and, with David Colander's help, worked out in detail a scheme, called Market Anti-Inflation Plan, better known as MAP (1980b), for rationing the right of firms to raise the ‘effective price’ of their output, that is, the sum of profits and wages entering the price of their products (value added). The scheme would give every firm the right to increase its value added in the proportion of the estimated rise in the economy's overall productivity, but it would also allow them to sell their unused rights or the unused portion of their rights (in a market created for the purpose) to those other firms that want to increase their wages and/or profits (value added) in greater proportion.

Lerner developed his Market Anti-Inflation Plan gradually and published it at several stages and in several versions before it reached its final form in 1980. It was his last major contribution to economics and a fitting end to his career, because it well illustrates both the strengths and the weaknesses of his extraordinarily fertile and original mind. It is bold, elegant, ingenious and impeccably logical, with meticulous attention to every conceivable detail and exception, but combines those qualities with a slightly utopian flavour, all of which have characterized just about all of Lerner's many proposals for reform.

For the sheer novelty and stark logic of Lerner's arguments and policy proposals usually took people aback, but he was utterly unwilling and perhaps also unable to soften their impact in the interests of their easier acceptability. He was well aware of the reasons for the hostile reception of virtually all his recommendations but believed, with some justification, that, as time wore off their shocking novelty, they would become more acceptable and politically feasible. Lerner's MAP could well be the best remedy for stagflation but many less good remedies will first have to be tried and prove ineffective in order to render MAP politically acceptable.

Selected works

1932. The diagrammatical representation of cost conditions in international trade. *Economica* 12, 346–56.

1934a. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1(June), 157–75.

1934b. The diagrammatical representation of demand conditions in international trade. *Economica*, NS 1, 319–34.

1936a. The symmetry between import and export taxes. *Economica*, NS 3, 306–13.

1936b. Mr. Keynes' 'General Theory of Employment, Interest and Money'. *International Labour Review* 34, 435–54.

1943a. Functional finance and the federal debt. *Social Research* 10(February), 38–51.

1943b. User cost and prime user cost. *American Economic Review* 33(March), 131–2.

1944. *The Economics of Control: Principles of Welfare Economics*. New York: Macmillan.

1949. The inflationary process: some theoretical aspects. *Review of Economics and Statistics* 31 (August), 193–200.

1951. *Economics of Employment*. New York: McGraw-Hill.

1952aFactor prices and international trade. *Economica*, NS 19, 1–15.

1952b. The essential properties of interest and money. *Quarterly Journal of Economics* 66(May), 172–93.

1953. On the marginal product of capital and the marginal efficiency of investment. *Journal of Political Economy* 61(February), 1–14.

1958Inflationary depression and the regulation of administered prices. *Joint Economic Committee Print*, Conference on Economic Stability and Growth, March.

1962. Macro-economics and micro-economics. In *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, ed. E. Nagel, P. Suppes and A. Tarski. Stanford: Stanford University Press.

1972. *Flation: Not Inflation of Prices, Not Deflation of Jobs*. New York: Quadrangle Books.

1980a. OPEC – a plan – if you can't beat them, join them. *Atlantic Economic Journal* 8(3), 1–3.

1980b(With D.C. Colander.) *MAP, A Market Anti-inflation Plan*. New York: Harcourt, Brace and Jovanovich.

For a representative collection of Lerner's best writings, see D.C. Colander, ed., *Selected Economic Writings of Abba P. Lerner*, New York: New York University Press, 1983. That volume contains most of the articles cited here and also has Lerner's complete bibliography.

For other, more detailed appraisals of Lerner's contribution to economics, see:

Samuelson, P.A. 1964. A.P. Lerner at sixty. *Review of Economic Studies* 31(June), 169–78.

Scitovsky, T. 1984. Lerner's contribution to economics. *Journal of Economic Literature* 22, 1547–71.

Sobel, I. 1979. Abba Lerner on employment and inflation: a post-Keynesian perspective. In *Essays in Post-Keynesian Inflation*, ed. J.H. Gapinski and C.E. Rockwood. Cambridge, Mass.: Ballinger.

How to cite this article

Scitovsky, T. "Lerner, Abba Ptachya (1905–1982)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000071> doi:10.1057/9780230226203.0961

The New Palgrave Dictionary of Economics Online

Leroy-Beaulieu, Pierre-Paul (1843 – 1916)

R.F. Hébert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

iron law of wages; Leroy-Beaulieu, P.-P.; mathematical method; public finance; Ricardian theory of rent

Article

French economist and journalist, Leroy-Beaulieu was born at Paris in 1843; he died there in 1916. His father was a Prefect and a Deputy under Louis-Philippe, his older brother a famous historian and a director of the Ecole des Sciences Politiques. His son Pierre, with whom he is sometimes confused, was also an economist. Initially trained in law, Paul Leroy-Beaulieu turned to economics in his early twenties, launching this new career with a prize-winning essay in 1867 on the effects of the moral and intellectual conditions of the working class on the rate of wages. Soon thereafter, he began collaborating on the *Revue des deux mondes*, and in 1871 he became editor of the *Journal des débats*. Two years later he founded the *Economiste française*, for which, as editor, he wrote weekly articles, missing only once in 43 years.

When Emile Boutmy established the Ecole Libre des Sciences Politiques in 1872, Leroy-Beaulieu accepted the chair of public finance. He later succeeded his father-in-law, Michel Chevalier, in the chair of political economy at the Collège de France. His ideas found wide exposure in countless journal articles and over a dozen books. A member of the French Institute and of the American Philosophical Society, he also received honorary degrees from the universities of Cambridge, Edinburgh, Dublin and Bologna.

Leroy-Beaulieu belonged to the French Liberal School of individualism and free trade. His major work, the *Traité théorique et pratique d'économie politique* (1896) is largely an exposition of classical theory. However, he rejected the pessimistic conclusions of Ricardo and Malthus, having argued in his *Essai sur la répartition des richesses* (1881) that there was no factual basis to either the Ricardian theory of rent or the 'iron law of wages'. Moreover, he sought to defuse the population bomb by arguing that the progress of civilization must always bring a declining birth rate because the altered demands and increased expenditures that accompany it are incompatible with the duties and responsibilities of parentage. In value theory, he followed the marginal analysis of the Austrians. Even as Walras was proselytizing on

its behalf, however, Leroy-Beaulieu reviled the mathematical method as ‘pure delusion and a hollow mockery … [without] scientific foundation and … practical use’. Showing equally poor judgement, he rejected the demand curve on frivolous grounds.

Leroy-Beaulieu's most enduring work was his treatise on public finance (1877), an effort that examines both public revenues and public credit. The second volume of this work rose somewhat above the first, remaining authoritative well into the 20th century.

Selected works

1877. *Traité de la science des finances*, 2 vols. Paris: Guillaumin.

1881. *Essai sur la répartition des richesses et sur la tendance à une moindre inégalité des conditions*. Paris: Guillaumin.

1890. *L’état moderne et ses* functions. Paris: Guillaumin.

1896. *Traité théorique et pratique d’économie politique*, 4 vols. Paris: Guillaumin.

1913. *La question de la population*. Paris: F. Alcan.

Bibliography

Pirou, G. 1925. *Les doctrines économiques en France depuis 1870*. Paris: A. Colin.

Stourm, R. 1917. Paul Leroy-Beaulieu. *Revue des deux mondes*, Series VI 38, 532–53.

How to cite this article

Hébert, R. F. "Leroy-Beaulieu, Pierre-Paul (1843–1916)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000072> doi:10.1057/9780230226203.0962

The New Palgrave Dictionary of Economics Online

level accounting

Francesco Caselli
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Level accounting (more recently known as development accounting) consists of a set of calculations whose purpose is to find the relative contributions of differences in inputs and differences in the efficiency with which inputs are used to cross-country differences in GDP. It is therefore the cross-country analogue of growth accounting.

Keywords

development accounting; growth accounting; level accounting; technical change; total factor productivity

Article

Suppose that country A is observed to produce more output than country B: is this because it employs a larger amount of labour, a larger amount of capital or a larger amount of some other input? Or because it somehow succeeds (or endeavours) to make more effective use of given inputs? Level accounting refers to a particular approach to attacking these questions. In this approach, one computes indices of the quantities of each input participating in production in different countries, as well as the shares of each input in total income. The contribution of inputs (or of a subset of the inputs) to differences in output is then given by a geometric average of the inputs, with the shares acting as weights. The difference between the cross-country difference in output and the cross-country differences in inputs, a residual, is interpreted as a cross-country difference in the efficiency with which the inputs are employed, or in total factor productivity (TFP). Level accounting is therefore the cross-country analogue of growth accounting.

The earliest level-accounting exercises are a five-country study by Denison (1967) and a two-country comparison by Walters (1968). In the late 1970s Jorgenson and Nishimizu (1978) and Christensen, Cummings, and Jorgenson (1981) adapted the growth-accounting framework of Jorgenson's work with Griliches and Christensen to level comparisons between the United States and eight other advanced

economies. They found substantial TFP differences.

More recently, level accounting has been a popular technique in addressing the sources of the enormous differences in income observed between the richest and poorest economies of the world (King and Levine, 1994; Klenow and Rodriguez-Clare, 1997; Hall and Jones, 1999). This trend has caused several authors to begin referring to it as ‘development accounting’. While details vary, a consensus emerging from the development-accounting literature is that observed inputs of labour and capital account for at best 50 per cent of the observed variation in aggregate value added across a large sample (numbering about 100) of developed and developing countries. It is often argued that this evidence points to the need for developing countries to underemphasize saving and investment, and emphasize technical change and technology adoption.

Unfortunately, residual variation in development accounting poses at least as many problems of interpretation as residual variation in growth accounting. The problems are compounded by the appalling coarseness of the data. Instead of accounting for compositional differences amongst a large number of education, gender, race, and age categories, as mandated by the Jorgensonian framework, development accountants to date have mostly had to limit themselves to a rough correction for average years of schooling. Perhaps more importantly, instead of allowing for imperfect susbstitutability among different types of capital, again as prescribed by best accounting practice, measures of the capital stock are based on linear aggregation. Caselli and Wilson (2004) show that this could be a fatal flaw. Finally, most development-accounting exercises assume constant capital (and hence labour) shares across countries.

Creative improvements in the measurement of labour quality have recently been proposed by Weil (2007) and Jones and Schneider (2007). Weil proposes a way to account for differences in the productive capacity of the labour force caused by differences in health, while Jones and Schneider bring to bear cross-country differences in IQ. Both succeed in reducing residual variation considerably. These appear to be two (rare) instances where level accounting has introduced innovations that could potentially also be usefully incorporated into growth accounting, instead of the other way around.

Another recent extension of the development-accounting framework is due to Caselli and Coleman (2006), who show how to decompose the cross-country residual into differences in the efficiency with which different inputs are used. Caselli (2005) uses this technique to show that most differences in efficiency are differences in the efficiency with which labour is used. Caselli and Coleman (2006) further trace these differences to differences in the efficiency of skilled labour.

Cross-country level accounting can also be performed at the industry level, and indeed this seems a necessary step towards shedding light on the sources of large residual variation at the aggregate level. Conrad and Jorgenson (1985), and Jorgenson, Kuroda and Nishimizu (1987) presented industry-level productivity comparisons for the United States, Japan, and Germany. Despite the richness of their data they found surprisingly large TFP differences. The more recent development-accounting literature has only attempted an agriculture–nonagriculture decomposition. The most convincing effort to date is possibly due to Vollrath (2006), who appears to be able to eliminate a significant amount of residual variation in aggregate GDP by accounting for the allocation of factors across these two sectors.

See Also

- growth accounting

Bibliography

Caselli, F. 2005. Accounting for cross-country income differences. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Caselli, F. and Coleman, J. 2006. The world technology frontier. *American Economic Review* 96, 499–522.

Caselli, F. and Wilson, D. 2004. Importing technology. *Journal of Monetary Economics* 51, 1–32.

Christensen, L., Cummings, D. and Jorgenson, D. 1981. Relative productivity levels, 1947–1973: an international comparison. In *New Developments in Productivity Measurement and Analysis, Studies in Income and Wealth*, vol. 41, ed. J. Kendrick and B. Vaccara. Chicago: University of Chicago Press.

Conrad, K. and Jorgenson, D. 1985. Sectoral productivity gaps between the United States, Japan, and Germany, 1960–1979. In *Probleme und Perspektiven der Wirtschaftlichen Entwicklung*, ed. H. Giersch. Berlin: Duncker and Humblot.

Denison, E. 1967. *Why Growth Rates Differ?* Washington, DC: Brookings Institution.

Hall, R.E. and Jones, C.I. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114, 83–116.

Jones, G. and Schneider, J. 2007. IQ and productivity. Working paper, George Mason University.

Jorgenson, D., Kuroda, M. and Nishimizu, M. 1987. Japan–U.S. industry-level productivity comparisons, 1960–1979. *Journal of the Japanese and International Economies* 1, 1–30.

Jorgenson, D. and Nishimizu, M. 1978. U.S. and Japanese economic growth, 1952–1974. *Economic Journal* 88, 707–26.

King, R.G. and Levine, R. 1994. Capital fundamentalism, economic development, and economic growth. *Carnegie-Rochester Conference Series on Public Policy* 40, 259–92.

Klenow, P.J. and Rodriguez-Clare, A. 1997. The neoclassical revival in growth economics: has it gone too far? In *NBER Macroeconomics Annual 1997*, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Vollrath, D. 2006. *How Important are Dual Economy Effects for Aggregate Productivity?* Houston:

University of Houston Press.

Walters, D. 1968. *Canadian Income Levels and Growth: An International Perspective*. Ottawa: Economic Council of Canada.

Weil, D. 2007. Accounting for the effect of health on economic growth. *Quarterly Journal of Economics* 122 (forthcoming).

How to cite this article

Caselli, Francesco. "level accounting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000242> doi:10.1057/9780230226203.0963

The New Palgrave Dictionary of Economics Online

Lewis, W. Arthur (1915 – 1991)

Ronald Findlay

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides an outline of the career and contributions of W. Arthur Lewis, who was awarded the Nobel Prize for Economics in 1982. Born in 1915 on the Caribbean island of St Lucia, Lewis began his career at the London School of Economics before moving to Manchester University and then to Princeton University. While *The Theory of Economic Growth* (1955) and *Growth and Fluctuations 1870–1913* (1978) have both been regarded as classics since they were first published, it is the 1954 article, ‘Economic Development with Unlimited Supplies of Labour’, that will probably remain as Lewis's most famous and influential single contribution.

Keywords

development economics; dual economies; Lewis, W. A.; Nurkse, R.; periphery; surplus labour; terms of trade

Article

W. Arthur Lewis was born on the island of St Lucia in the British West Indies on 23 January 1915. His early education was at St Mary's College on the island, where he completed a rigorous high school curriculum by the age of 14. This school is remarkable for having been attended not only by Lewis but also, 15 years later, by St Lucia's other Nobel Laureate, the poet Derek Walcott. A scholarship took Lewis to the London School of Economics in 1933, where he obtained a BA in Commerce with first class honours in 1937 and then went on to do a Ph.D. under the supervision of Arnold Plant, who incidentally was also the supervisor of Ronald Coase. In 1938 he was appointed as a junior member of the faculty, the first black man to receive such a position in the history of the institution. His very active teaching at the LSE on a very broad range of subjects undoubtedly prepared him well for his future work on economic development. He moved to Manchester University in 1947, where he held the Stanley Jevons Chair, previously occupied by J.R. Hicks, and where he was himself to be succeeded by Harry Johnson. It was here that he did some of his most seminal work on development economics, the

Manchester School article on ‘Economic Development with Unlimited Supplies of Labor’ (1954) and the treatise on *The Theory of Economic Growth* (1955). In the 1950s he was a senior official in agencies of the United Nations, and was for a time Vice Chancellor of the University of the West Indies. He went to Princeton in 1963, where he remained until his retirement in 1983: just as at the LSE, he was the first person of African descent ever to be appointed to the faculty. He held many part-time advisory positions with international organizations and governments in developing countries, particularly in West Africa and the Caribbean. He was awarded the Nobel Prize for Economics in 1979, together with T.W. Schultz, for their contributions to economic development. He died at his summer home in Barbados on 15 June 1991.

His earliest original research, including his Ph.D. thesis, was on the application of price theory to problems of industrial organization and public utilities. A number of studies published during the 1940s, such as ‘The Two Part Tariff’ (1941), ‘Competition in Retailing’ (1945), ‘Fixed Costs’ (1946), and other related topics, were brought together in a volume entitled *Overhead Costs*, published in 1949. Two other books published in the same year, based on his LSE lectures, were *Economic Survey 1919–1939* and *Principles of Economic Planning*. The first of these was an examination of the troubled economic history of the world economy in the interwar period, notable in particular for the way in which he linked together the experiences of the ‘core’ industrial countries with those of the primary producing ‘periphery’ of the world economy. The pessimism about the possibility of international trade to serve as a sustained ‘engine of growth’ for the developing countries, that has marked his subsequent writings on development economics down to his Nobel Prize Lecture in 1980 (entitled the ‘The Slowing Down of the Engine of Growth’), can perhaps be traced to his study of the inter-war period, an interesting parallel with the case of Ragnar Nurkse, who also came to the study of development problems after writing his *International Currency Experience* on the breakdown of the international monetary system in the 1930s. The book on planning, though written at an introductory level, was a penetrating early examination of the problems of coordinating government intervention and the market in a mixed economy.

Lewis's most famous and influential contribution to economics is undoubtedly the 1954 paper on development with ‘unlimited supplies’ of labour. He presents a stylized model in which the typical poor country is divided into a ‘traditional’ and a ‘modern’ sector. The former consists of peasant agriculture as well as self-employment of various sorts in urban areas, where the primary objective of economic activity is to maintain consumption. The ‘modern’ sector comprises commercial farming, plantations and mines and manufacturing, in all of which there is hired labour and profit is the motive for production organized by a class of capitalists and entrepreneurs. Lewis adopts a strictly classical viewpoint on two crucial features of his model. First, the real wage of unskilled labour in the modern sector is exogenously given, with employment and profits then being determined by the demand for labour corresponding to the fixed stock of capital in the short run. The second classical feature is that the accumulation of capital is governed by saving out of profits. The process of economic development is viewed as the expansion of the modern relative to the traditional sector until such time as the ‘surplus labour’ pool in the traditional sector is drained and an integrated labour market emerges with a neoclassically determined equilibrium real wage, rising steadily over time as growth proceeds. The model as a whole thus has two distinct phases, an initial ‘classical’ one with a fixed real wage, that is the main focus of the analysis, and a subsequent ‘neoclassical’ one with a rising real wage. The concept of a ‘dual economy’ in the first phase of the model has generated considerable controversy and an extensive polemical literature, to which references can be found in Findlay (1980), together with an appraisal, extensions and critique of

the model itself. The most sophisticated and thorough theoretical defence of the dual economy and the associated notion of ‘surplus labour’ remains that provided by Sen (1966). The *Manchester School* in 2004 appropriately marked the 50th anniversary of the most celebrated article it ever published by a special issue, which contains a valuable survey of subsequent developments by Kirkpatrick and Barrientos (2004).

Another notable, but much less well-known, contribution of this seminal (1954) paper, in a neglected section on the open economy, is a model of the terms of trade between manufactures and primary products that is developed further, with empirical applications, in his (1969) Wicksell Lectures. The key idea is that the world price of manufactures, relative to the prices of tropical products such as coffee, tea, sugar, rubber and jute, is determined by the relative opportunity costs of labour in food production. Thus the Pittsburgh steel worker's wage is governed by the Kansas farmer's productivity, while the Brazilian coffee plantation wage is determined by the much lower productivity of peasant subsistence agriculture, which explains why a unit of steel in the world market commands so many more units of coffee. Since the transformation curves between steel and food and coffee and food are assumed to be linear, demand only determines quantities produced, consumed and traded, not relative prices, exactly as in the approach of the classical economists. Lewis applied this model in a very imaginative way to illuminate several key aspects of the history of the world economy in his last major work, *Growth and Fluctuations 1870–1913*, published in (1978). This volume extended his examination of the world economy in the inter-war period in *Economic Survey 1919–1939* back to the ‘golden age’ of globalization from 1870 to 1913, and is a deeply original piece of theoretical, statistical and historical research in the manner of Schumpeter and Kuznets. Both volumes are still essential reading for any serious student of the evolution of the world economy.

The reader can find an extensive collection of Lewis's articles and shorter monographs in the volume edited by Mark Gersovitz (1983). A measure of his influence on the field of development economics can be gathered from the volume of essays in his honour edited by Gersovitz and others (1982). Robert L. Tignor (2006) is a very valuable account of the life and inspiring achievements of this great pioneer of development economics, rightly drawing attention to the stoic courage and steely resolution with which he confronted and overcame the racial prejudice that was so virulent even in Western academic circles during his early career. The effect of these experiences may have made him appear to many as reserved, aloof and ‘prickly’ but to all who knew him well he was always kind, courteous and considerate, with a puckish sense of humour. The writer Pico Iyer (1997) described Derek Walcott, the other Nobel Laureate of St Lucia, as a ‘Tropical Classical’ because of the deep influence of Homer and other classical authors on his poetry. The designation fits Arthur Lewis admirably as well, and not only because of the influence of Ricardo and other classical authors on his economics.

See Also

- dual economies

Selected works

1941. The two part tariff. *Economica*, NS 8, 240–70.

1945. Competition in retail trade. *Economica*, NS12, 202–34.

1946. Fixed costs. *Economica*, NS13, 231–58.

1949. *Economic Survey 1919–1939*. London: Allen & Unwin.

1949. *Overhead Costs*. London: Allen & Unwin.

1949. *The Principles of Economic Planning*. London: Allen & Unwin.

1954. Economic development with unlimited supplies of labor. *Manchester School* 22, 139–91.

1955. *The Theory of Economic Growth*. London: Allen & Unwin.

1969. *Aspects of Tropical Trade 1883–1965*. Wicksell Lectures, Stockholm: Almqvist & Wiksell.

1978. *Growth and Fluctuations 1870–1913*. London: Allen & Unwin.

1980. The slowing down of the engine of growth (Nobel Lecture). *American Economic Review*, 70, 555–64.

Bibliography

Findlay, R. 1980. On W. Arthur Lewis' contributions to economics. *Scandinavian Journal of Economics* 82(1), 62–76.

Gersovitz, M., Diaz-Alejandro, C.F., Ranis, G. and Rosenzweig, M.R. 1982. *The Theory and Experience of Economic Development: Essays in Honour of W. Arthur Lewis*. London: George Allen & Unwin.

Gersovitz, M. 1983. *Selected Economic Writings of W. Arthur Lewis*. New York: New York University Press.

Iyer, P. 1997. *Tropical Classical: Essays from Several Directions*. New York: Vintage.

Kirkpatrick, C. and Barrientos, A. 2004. The Lewis model after 50 years. *Manchester School* 72, 679–90.

Nurkse, R. 1944. *International Currency Experience*. N.p.: League of Nations.

Sen, A.K. 1966. Peasants and dualism with or without surplus labor. *Journal of Political Economy* 74,

425–50.

Tignor, R.L. 2006. *W. Arthur Lewis and the Birth of Development Economics*. Princeton: Princeton University Press.

How to cite this article

Findlay, Ronald. "Lewis, W. Arthur (1915–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000079> doi:10.1057/9780230226203.0964

The New Palgrave Dictionary of Economics Online

lexicographic orderings

C. Blackorby

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

interpersonal utility comparisons; lexicographic orderings; maximin; social choice

Article

Lexicographic orderings are orderings in which certain elements of the space being ordered have been selected for special treatment. I begin with an example. Suppose an agent has an ordering over commodities *a* and *b*. Although he or she likes both *a* and *b*, any bundle which has more of *a* is *preferred* to any bundle which has less of *a*. Of course among bundles which have the same amount of *a*, bundles with more *b* are preferred to those with less. Thus, there are no trade-offs between *a* and *b* and each indifference set is a single point. The name ‘lexicographic’ comes from the way words are ordered in a dictionary, alphabetically by the first letter and then the second and so on.

Lexicographic orderings were known chiefly as simple examples of orderings which *could not* be represented by a continuous real-valued function; see Debreu (1954) for the first discussion of this issue in economics. It is, however, in social choice theory and welfare economics where these orderings have come to prominence. To demonstrate their role a lexicographic maximin rule (leximin) follows. Let $u = (u_1, \ldots, u_N)$ be an element of a Euclidean *N*-space where u_n is the utility of person *n*. In each possible state of the world, say $\bar{u}_1 = (\bar{u}_1, \ldots, \bar{u}_n)$, let $r(\bar{u})$ be the person who is the *r*th best off. For example, if *N*=3 and $\bar{u} = (2, 7, 3)$ then $1(\bar{u}) = 2$ as person 2 has the highest utility, $2(\bar{u}) = 3$, and $3(\bar{u}) = 1$; ties are broken arbitrarily. An ordering *R* is a leximin rule if and only if for all $(u, \bar{u})$, $\bar{u}P\bar{u}$ if and only if there exists a k, $1 \leq k \leq N$, such that $\bar{u}_{k(u)} > \bar{u}_{k(u)}$ and for all $j > k$, $\bar{u}_{j(u)} = \bar{u}_{j(u)}$ where *P* is the strict preference relation, the asymmetric factor of *R*. That is, if the worst-off $N - k$ people have the same utility levels in $\bar{u}$ and $\bar{u}$ and the next worst-off person, *k*, is better off in $\bar{u}$ than in $\bar{u}$, then $\bar{u}$ is preferred to $\bar{u}$. Continuing the numerical example above let $\bar{u} = (2, 7, 2.5)$ so that $1(\bar{u}) = 2$, $2(\bar{u}) = 3$, and $3(\bar{u}) = 1$. Then $k = 2$, $\bar{u}_{2(u)} = 3 > \bar{u}_{2(u)}$ and $\bar{u}_{3(u)} = 2 = \bar{u}_{3(u)}$ hence $\bar{u}P\bar{u}$.

It is important to notice that if each person's utility function were subjected to the same increasing transformation the above ordering would not change. This is a case where utility is ordinally measurable

but fully comparable as levels of utility can be compared across individuals. That the leximin rule satisfies all of the original axioms of Arrow (1951; 1963) except for the comparability of levels of utility was first worked out by D'Aspremont and Gevers (1977).

Other types of lexicographic orderings appear frequently in social choice theory; see Sen (1986, section 6).

See Also

- orderings

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley. 2nd edn, 1963.

D'Aspremont, C. and Gevers, L. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44, 199–209.

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision Processes*, ed. R. Thrall, C. Coombs and R. Davis. New York: Wiley.

Sen, A.K. 1986. Social choice theory. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M. Intriligator. Amsterdam and New York: North-Holland.

How to cite this article

Blackorby, C. "lexicographic orderings." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000080> doi:10.1057/9780230226203.0965

The New Palgrave Dictionary of Economics Online

liability for accidents

Steven Shavell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Legal liability for accidents determines the circumstances under which injurers must compensate victims for harm. The effects of liability on incentives to reduce risk, on risk-bearing and insurance (both direct coverage for victims and liability coverage for injurers), and on administrative expenses are considered. Liability is also compared with other methods of controlling harmful activities, notably, with regulation and corrective taxation.

Keywords

accident insurance; contributory negligence; corrective taxes; damages; due care; judgment-proof problem; liability for accidents; liability insurance; moral hazard; negligence rule; product liability; risk aversion; safety regulation; strict liability

Article

Legal liability for accidents governs the circumstances under which parties who cause harm to others must compensate them. There are two basic rules of liability. Under *strict liability*, an injurer must always pay a victim for harm due to an accident that he causes. Under the *negligence rule*, an injurer must pay for harm caused only when he is found negligent, that is, only when his level of care was less than a standard of care chosen by the courts, often referred to as *due care*. (There are various versions of these rules that depend on victims' care, as will be discussed.) In fact, the negligence rule is the dominant form of liability; strict liability is reserved mainly for certain especially dangerous activities (such as the use of explosives). The amount that a liable injurer must pay a victim is known as *damages*. Our discussion of liability begins by examining how liability rules create incentives to reduce risk. The allocation of risk and insurance is next addressed, and, following that, the factor of administrative costs. Then a number of topics are reviewed. Comprehensive economic treatments of accident liability are presented in Landes and Posner (1987) and Shavell (1987); an early, insightful informal, economically

oriented treatment of liability is presented in Calabresi (1970). Empirical literature is surveyed in Kessler and Rubinfeld (2007) and is not considered here.

Incentives

In order to focus on liability and incentives to reduce risk, we assume in this section that parties are risk neutral. Further, we suppose that there are two classes of parties – injurers and victims – who do not have a contractual relationship. For example, injurers might be drivers and victims pedestrians, or injurers might be polluting firms and victims affected residents.

Unilateral accidents and the level of care

Here we suppose that injurers alone can reduce risk by choosing a level of *care*. Let x be expenditures on care (or the money value of effort) and $p(x)$ be the probability of an accident that causes harm h, where p is declining in x. Assume that the social objective is to minimize total expected costs, $x + p(x)h$, and let x^* denote the optimal x.

Under strict liability, injurers pay damages equal to h whenever an accident occurs, and they naturally bear the cost of care x. Thus, they minimize $x + p(x)h$; accordingly, they choose x^*.

Under the negligence rule, suppose that the due care level $\hat{x}$ is set equal to x^*, meaning that an injurer who causes harm will have to pay h if $x < x^*$ but will not have to pay anything if $x \geq x^*$. Then the injurer will choose x^*: he will not choose $x > x^*$, for that will cost him more and he escapes liability by choosing merely x^*; he will not choose $x < x^*$, for then he will be liable (in which case the analysis of strict liability shows that he would not choose $x < x^*$).

Thus, under both forms of liability, injurers are led to take optimal care, as first shown in Brown (1973). Note that under the negligence rule courts need to be able to calculate optimal care x^* and to observe actual care x, in addition to observing harm. Under strict liability courts need only to observe harm.

It should also be noticed that, under the negligence rule with due care $\hat{x}$ equal to x^*, negligence is never found, because injurers are induced to be non-negligent. Findings of negligence may occur, however, under a variety of modifications of our assumptions. Courts might make errors in observing injurers' care, so that an injurer whose true x is at least x^* might mistakenly be found negligent because his observed level of care is below x^*. Similarly, courts might err in calculating x^* and thus might set due care $\hat{x}$ above x^*. If so, an injurer who chooses x^* would be found negligent (even though care is accurately observed) because $\hat{x}$ exceeds x^*. As emphasized by Craswell and Calfee (1986), error in the negligence determination leads injurers to choose incorrect levels of care, and under some assumptions, to take excessive care in order to reduce the risk of being found negligent by mistake. Other explanations for findings of negligence are that individuals may not know x^* and thus take too little care, the judgment-proof problem (see below), which may lead individuals to choose to be negligent, and the inability of individuals to control their behaviour perfectly at every moment or of firms to control their employees.

Bilateral accidents and levels of care

We now assume that victims also choose a level of care y, that the probability of an accident is $p(x,y)$ and is declining in both variables, that the social goal is to minimize $x + y + p(x, y)h$, and that the optimal levels of care x^* and y^* are positive.

Under strict liability, injurers' incentives are optimal conditional on victims' level of care, but victims have no incentive to take care because they are fully compensated for their losses. However, the usual strict liability rule that applies in bilateral situations is strict liability with a defense of *contributory negligence*, meaning that an injurer is liable for harm only if the victim's level of care was not negligent, that is, his level of care was at least his due care level $\hat{y}$. If victims' due care level is y^*, then it is a unique equilibrium for both injurers and victims to act optimally: victims choose y^* in order to avoid having to bear their losses, and injurers choose x^* since they will be liable because victims are non-negligent.

Under the negligence rule, optimal behaviour is also the unique equilibrium. Injurers choose x^* to avoid being liable, and, since victims therefore bear their losses, they choose y^*. Two other variants of the negligence rule are negligence with the defence of contributory negligence (under which a negligent injurer is liable only if the victim is not negligent) and the comparative negligence rule (under which a negligent injurer is only partially liable if the victim is also negligent). These rules also induce optimal behaviour.

Thus, all of the negligence rules, and strict liability with the defence of contributory negligence, support optimal care, on the assumption due care levels are chosen optimally. Courts need to be able to calculate optimal care levels for at least one party under any of the rules, and in general this requires knowledge of the function $p(x,y)$. The main conclusions of this section were first proved by Brown (1973) (see also Diamond, 1974, for closely related results).

Unilateral accidents, level of care, and level of activity

Now let us reconsider unilateral accidents, allowing for injurers to choose their level of *activity* z, which is interpreted as the (continuously variable) number of times they engage in their activity (or, if injurers are firms, the scale of their output). Let $b(z)$ be the benefit from the activity, and assume the social object is to maximize $b(z) - z(x + p(x)h)$; here $x + p(x)h$ is assumed to be the cost of care and expected harm each time an injurer engages in his activity. Let x^* and z^* be optimal values. Note that, as before, x^* minimizes $x + p(x)h$, and that z^* satisfies $b'(z) = x^* + p(x^*)h$, the marginal benefit from the activity equals the marginal social cost, comprising the sum of the cost of optimal care and expected accident losses.

Under strict liability, injurers choose both the level of care and the level of activity optimally, as their objective is the social objective.

Under the negligence rule, injurers choose optimal care x^* as before, but their activity is socially excessive. Because an injurer escapes liability by taking care of x^*, he chooses z to maximize $b(z) - zx^*$, so that z satisfies $b'(z) = x^*$. The injurer's cost of raising his activity level is only his cost of care x^*, which is less than the social cost, as that also includes $p(x^*)h$. The excessive level of activity under the negligence rule is more important the larger is the expected harm $p(x^*)h$ from the activity.

The failure of the negligence rule to control the level of activity arises because negligence is defined here (and for the most part in reality) in terms of care alone. A justification for this assumption is that courts might face informational difficulties were they to include the activity level in the definition of negligence. The problem with the activity level under the negligence rule is applicable to any aspect of behaviour that would be difficult to incorporate into the negligence standard (including, for example, research and development activity). The distinction between levels of care and levels of activity was developed in Shavell (1980).

Bilateral accidents, levels of care, and levels of activity

If we consider levels of care and of activity for both injurers and victims, then none of the liability rules that we have considered leads to full optimality (on the assumption that activity levels are unobservable). The reason that full optimality cannot be achieved is in essence that injurers must bear full accident losses to induce them to choose the right level of their activity, but this means that victims will not choose the optimal level of their activity.

Risk-bearing and insurance

We next examine the implications of risk aversion and the role of insurance in the liability system (see Shavell, 1982a). A number of general points may be made.

First, the socially optimal resolution of the accident problem now involves not only the reduction of losses from accidents but also the protection of risk-averse parties against risk. Risk bearing is relevant for two reasons: not only because potential victims may face the risk of accident losses, but also because potential injurers may face the risk of liability. The former risk can be mitigated through accident insurance, and the latter through liability insurance.

Second, the incentives associated with liability do not function in the direct way discussed in the previous section, but instead are mediated by the terms of insurance policies. To illustrate, consider strict liability in the unilateral accident model with care alone variable, and assume that insurance is sold at actuarially fair rates. If injurers are risk averse and liability insurers can observe their levels of care, injurers will purchase full liability insurance coverage and their premiums will depend on their level of care; their premiums will equal $p(x)h$. Thus, injurers will want to minimize their costs of care plus premiums, or $x + p(x)h$, so they will choose the optimal level of care x^*. In this instance, liability insurance eliminates risk for injurers, and the situation reduces to the previously analysed risk-neutral case.

If, however, liability insurers cannot observe levels of care, ownership of full coverage could create severe moral hazard, so would not be purchased. Instead, as is known from the theory of insurance, the typical amount of coverage purchased will be partial, for that leaves injurers with an incentive to reduce risk. In this case, therefore, the liability rule results in some direct incentive to take care because injurers are left bearing some risk after their purchase of liability insurance, but their level of care tends to be less than first best.

This last situation, in which liability insurance dilutes incentives, leads to a third point, concerning the question whether the sale of liability insurance is socially desirable. (We note that, because of fears

about incentives, the sale of liability insurance was delayed for decades in many countries and that it was not allowed in the Soviet Union; further, in the United States liability insurance is sometimes forbidden against certain types of liability, such as against punitive damages.) The answer to the question is that, even though it may dilute incentives, sale of liability insurance is socially desirable, at least in basic models of accidents and some variations of them. In the case just considered, for example, injurers are made better off by the presence of liability insurance, as they choose to purchase it, and victims are indifferent to its purchase by injurers because victims are fully compensated for any harm suffered. This argument must be modified in other cases, such as when the damages injurers pay are less than harm because injurers are judgment-proof.

Fourth, consider how the comparison between strict liability and the negligence rule is affected by risk bearing. The immediate effect of strict liability is to shift the risk of loss from victims to injurers, whereas the immediate effect of the negligence rule is to leave the risk on victims (as injurers tend to act non-negligently). However, the presence of insurance means that victims and injurers can substantially shield themselves from risk, attenuating the relevance of risk bearing for the comparison of strict liability and negligence.

Finally, the presence of insurance implies that the liability system cannot be justified primarily as a means of compensating risk-averse victims against loss. Rather, the justification for the liability system must lie in significant part in the incentives that it creates to reduce risk. To amplify, although both the liability system and the insurance system can compensate victims, the liability system is much more expensive than the insurance system (see the next section). Accordingly, were there no social need to create incentives to reduce risk, it would be best to dispense with the liability system and to rely on insurance to accomplish compensation.

Administrative costs

The administrative costs of the liability system are the legal and other costs (notably the time of litigants) involved in bringing suit and resolving it through settlement or trial. These costs are substantial; a number of estimates suggest that, on average, administrative costs of a dollar or more are incurred for every dollar that a victim receives through the liability system (Shavell, 2004, p. 281).

Strict liability versus negligence

The factor of administrative costs affects the comparison of liability rules. On one hand, we would expect the volume of cases – and thus administrative costs – to be higher under strict liability than under the negligence rule. On the other hand, given that there is a case, we would anticipate administrative costs to be higher under the negligence rule because due care will be at issue. Hence, it is not clear which liability rule is administratively cheaper.

Social desirability of the liability system and private motives to sue

The existence and the surprisingly high magnitude of administrative costs raise rather sharply the question whether the liability system is socially worthwhile. Moreover, the private motive to sue is not

in alignment with the social reasons for using the liability system. First, the private benefit of suit is the amount of money that would be obtained from it, whereas the social benefit is the deterrence that would be created. Second, the private cost of suit is the victim's cost, whereas the social cost includes also the injurer's and the state's cost. These differences give rise to the possibility of socially excessive or socially insufficient suit. To illustrate the former, suppose that care has no effect on the accident probability, so that it is socially undesirable for suit to be brought. Yet under strict liability a victim will bring suit as long as his cost is less than the harm suffered, so the volume of litigation activity could be high. To illustrate the possibility of socially inadequate suit, suppose that an expenditure on care of only one hundredth of harm will eliminate the possibility of otherwise certain harm, and suppose also that the magnitude of harm is less than the cost of suit. Then no suit will be brought. However, it would be desirable for victims to have an incentive to bring suit, for that would induce care to be taken, and, since no harm would then occur, no suit would ever occur. The private versus the social motive to make use of the legal system was first developed in Shavell (1982b, 1997); see also Polinsky and Rubinfeld (1988).

Topics

Damages

Under strict liability, damages must equal harm h for incentives to be optimal. Under the negligence rule, however, damages higher than h also would induce injurers to take optimal care of x^*. Higher damages will increase the incentive to be non-negligent; they will not lead injurers to take excessive care because injurers can escape liability merely by taking care of x^*. But when there is uncertainty in the negligence determination, damages higher than h may lead to problems of excessive care.
Damages exceeding h are desirable if injurers sometimes escape liability, as when injurers may be hard to identify (the origin of pollution may be difficult to trace). If the probability of liability for harm is q, then, if damages are raised to $(1/q)h$, expected liability will be h. Thus, the more likely an injurer is to escape liability, the higher should be damages. On these points and others about punitive damages, see Cooter (1989) and Polinsky and Shavell (1998).

Causation

A fundamental principle of liability law is that a party cannot be held liable unless he was the cause of losses. For example, if cancer occurs in an area where a firm has polluted, the firm will be liable only for the cancer that it caused, not for cancer due to other carcinogens. This principle is necessary to achieve social efficiency under strict liability, because otherwise incentives would be distorted. Socially desirable production might be rendered unprofitable if the firm were held responsible for all cases of cancer. Under the negligence rule, restricting liability to accidents caused by an actor may be less important than under strict liability: if negligent actors were held liable for harms they did not cause, they would only have greater reason to act non-negligently. On causation and incentives, see Calabresi (1975), Kahan (1989), and Shavell (1987).

Judgment-proof problem

The possibility that injurers may not be able to pay in full for the harm they cause is known as the judgment-proof problem and is of substantial importance, for individuals and firms often cause harms significantly exceeding their assets. When injurers are unable to pay fully for the harm they may cause, their incentives to reduce risk are inadequate, and their incentives to engage in risky activities excessive. Policy responses to the judgment-proof problem include vicarious liability (imposed on a party who has some control over the judgment-proof party), minimum asset requirements for participation in harmful activities, safety regulation, and criminal liability. On the judgment-proof problem and responses to it, see Kornhauser (1982), Pitchford (1995), Shavell (1986; 2005), and Sykes (1984).

Product liability

When victims are customers of firms, the role of liability in providing incentives may be attenuated or even non-existent. If customers have perfect knowledge of product risks, then they will pay less for risky products, and incentives to reduce risk will be optimal without liability. If, however, customer knowledge of risk is imperfect, liability is potentially useful in reducing risk. In the latter case, a question of interest is whether court-determined liability or market-determined liability, namely, warranties, is likely to be better, on which see Priest (1981), Rubin (1993), and Spence (1977).

Liability versus other means of controlling risk

Liability is only one method of controlling harm-causing behaviour; safety regulation and corrective taxes are among the alternatives. Liability harnesses the information that victims have about the occurrence of harm, and thus may be advantageous when victims, rather than the state, naturally observe how harm comes about; whereas when harm-causing behaviour and its occurrence requires state effort to be ascertained, regulation and taxation may be advantageous. In order for liability to function well as an incentive device, injurers must have assets approximating the harm they might cause, whereas regulation and taxation (based on expected harm rather than actual harm) do not require injurers to have substantial assets. Liability, however, may enjoy an administrative cost advantage over regulation and taxation, in that administrative costs are incurred under the liability system only when harm comes about, whereas such costs generally are incurred more often under regulation and taxation. On the comparison of the liability system and other means of controlling risk, see Calabresi and Melamed (1972), Kolstad, Ulen and Johnson (1990), and Shavell (1993).

See Also

- externalities
- law, economic analysis of

Bibliography

Brown, J.P. 1973. Toward an economic theory of liability. *Journal of Legal Studies* 2, 323–49.

Calabresi, G. 1970. *The Costs of Accidents*. New Haven, CT: Yale University Press.

Calabresi, G. 1975. Concerning cause and the law of torts. *University of Chicago Law Review* 43, 69–108.

Calabresi, G. and Melamed, A.D. 1972. Property rules, liability rules, and inalienability: one view of the cathedral. *Harvard Law Review* 85, 1089–128.

Cooter, R.D. 1989. Punitive damages for deterrence: when and how much? *Alabama Law Review* 40, 1143–96.

Craswell, R. and Calfee, J.E. 1986. Deterrence and uncertain legal standards. *Journal of Law, Economics, & Organization* 2, 279–303.

Diamond, P.A. 1974. Single activity accidents. *Journal of Legal Studies* 3, 107–64.

Kahan, M. 1989. Causation and incentives to take care under the negligence rule. *Journal of Legal Studies* 18, 427–47.

Kessler, D. and Rubinfeld, D.R. 2007. Empirical study of the common law and legal process. In *Handbook of Law and Economics*, vol. 1, ed. A.M. Polinsky and S. Shavell. Amsterdam: North-Holland.

Kolstad, C.D., Ulen, T.S. and Johnson, G.V. 1990. Ex post liability vs. ex ante regulation: substitutes or complements? *American Economic Review* 80, 888–901.

Kornhauser, L. 1982. An economic analysis of the choice between enterprise and personal liability for accidents. *California Law Review* 70, 1345–92.

Landes, W.M. and Posner, R.A. 1987. *The Economic Structure of Tort Law*. Cambridge, MA: Harvard University Press.

Pitchford, R. 1995. How liable should a lender be? The case of judgment-proof firm and environmental risk. *American Economic Review* 85, 1171–86.

Polinsky, A.M. and Rubinfeld, D.R. 1988. The welfare implications of costly litigation for the level of liability. *Journal of Legal Studies* 17, 151–64.

Polinsky, A.M. and Shavell, S. 1998. Punitive damages: an economic analysis. *Harvard Law Review* 111, 869–962.

Priest, G.L. 1981. A theory of the consumer warranty. *Yale Law Journal* 90, 1297–352.

Rubin, P.H. 1993. *Tort Reform by Contract*. Washington, DC: AEI Press.

Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 9, 1–25.

Shavell, S. 1982a. On liability and insurance. *Bell Journal of Economics* 13, 120–32.

Shavell, S. 1982b. The social versus the private incentive to bring suit in a costly legal system. *Journal of Legal Studies* 11, 333–9.

Shavell, S. 1986. The judgment proof problem. *International Review of Law and Economics* 6, 45–58.

Shavell, S. 1987. *Economic Analysis of Accident Law*. Cambridge, MA: Harvard University Press.

Shavell, S. 1993. The optimal structure of law enforcement. *Journal of Law and Economics* 36, 255–87.

Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26, 575–612.

Shavell, S. 2004. *Foundations of Economic Analysis of Law*. Cambridge, MA: Harvard University Press.

Shavell, S. 2005. Minimum asset requirements and compulsory liability insurance as solutions to the judgment-proof problem. *Rand Journal of Economics* 36, 63–77.

Spence, M. 1977. Consumer misperceptions, product failure, and producer liability. *Review of Economic Studies* 44, 561–72.

Sykes, A. 1984. The economics of vicarious liability. *Yale Law Journal* 93, 1231–80.

How to cite this article

Shavell, Steven. "liability for accidents." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000215> doi:10.1057/9780230226203.0966

The New Palgrave Dictionary of Economics Online

liberalism and economics

Ralf Dahrendorf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

anarchism; Beveridge, W. H.; Buchanan, J. M.; capitalism; classical liberalism; conservatism; democracy; Friedman, M.; Hayek, F. A. von; invisible hand; Keynes, J. M.; liberalism; liberty; Locke, J.; market; minimal state; natural rights; neoliberalism; Nozick, R.; Popper, K.; rule of law; Smith, A.; social contract; social democracy; social liberalism; social market economy; social rights; socialism; stagflation; supply-side economics; totalitarianism; uncertainty; unemployment; welfare state

Article

Liberalism is the theory and practice of reforms which has inspired two centuries of modern history. It grew out of the English Revolutions of the 17th century, spread to many countries in the wake of the American and French Revolutions of the 18th century, and dominated the better part of the 19th century. At that time, it also underwent changes. Some say it died, or gave way to socialism, or allowed itself to be perverted by socialist ideas; others regard the social reforms of the late 19th and 20th centuries as achievements of a new liberalism. More recently, interest in the original ideas of liberals has been revived. Thus, classical liberals, social liberals and neoliberals may be distinguished.

Classical liberalism is a simple, dramatic philosophy. Its central idea is liberty under the law. People must be allowed to follow their own interests and desires, constrained only by rules which prevent their encroachment on the liberty of others. Early liberals before and after John Locke (1690) liked to use the metaphor of a social contract to express this view. Society can be thought of as emerging from an agreement among its members to protect themselves against the selfish desires of others. Man's ‘unsociable sociability’ (Kant, 1784) makes rules necessary which bind all, but requires also the maximum feasible space for competition and conflict.

In fact, of course, early liberals were not concerned with building societies from scratch. They were concerned with forcing absolute rulers to yield to demands for liberty. The rule of law envisaged by liberals was a revolutionary force which heralded the enlightened phase of modernity.

The notion, rule of law, is not without ambiguity. It is, in the first instance, largely formal. One thinks of

rules of the game applying to all and regulating the social, economic and political process. In theory, such rules are intended not to prejudge the outcome of the game itself. Still, even their formal conditions, equality before the law and due process, involved fundamental changes which justify speaking of a movement of reform. Throughout the history of liberalism, however, the question of certain substantive rights of man has been an issue. The inviolability of the person and the rights of free expression have been liberal causes along with constitutional rules. Liberals have rarely found it easy to reason for such substantive rights to their own satisfaction. A certain tension between liberal thought and the notion of natural rights is unmistakable.

The modern debate of these issues began in Scotland and England. John Locke, David Hume (1740) and Adam Smith (1776) are but three of many names to consider. From Britain, the ideas spread to the United States and to continental Europe. Montesqieu and Kant borrowed some of their ideas from British liberals. The American Declaration of Independence and the Constitution, the Declaration of the Rights of Man three years after the French Revolution are only two practical illustrations of the effect of the new ideas. If one wants to, one can distinguish, with Friedrich von Hayek, between a British ‘evolutionary’ and a continental ‘constructivist’ concept of liberalism. Either or both however became the dominant reform movements of the early 19th century and determined the dynamics of Europe and North America between the 1780s and the 1840s or 1850s.

Liberalism had consequences for economic, social and political thought. Its economic application was the most obvious and remains the most familiar. If rules of the game are all that can be justified whereas otherwise interests should be allowed a free reign, the scene is set for the operation of the market. It is the forum where equal rights of access and participation but divergent and competing interests lead, through the operation of an ‘invisible hand’ (Adam Smith), to the greatest welfare for all. Liberalism and market capitalism are inseparable, much as later European theorists (notably in Germany and Italy) have tried to dissociate the two.

The social application of liberalism analogously leads to the emergence of the public, if by ‘public’ we understand the meeting place of divergent views from which a ‘public opinion’ emerges. On the Continent, a more emphatic language is often preferred; here, one likes to speak of the emergence of society from under the state. Either way, the basic idea involves the same departure from an all-embracing system of domination by traditional authorities to one in which public authority is confined to certain tasks of regulation, and thus bound to grant and defend the freedom of individuals to express their views.

This is the point at which classical liberalism was not only instrumental for the promotion of market capitalism and social participation, but also for the development of what is called today, democracy. Again, the term is anything but clear. It can be understood to mean a system of government which is based on the competition of divergent views – individual views or group views – for power, constrained by rules which limit the instruments used in the process, and stipulate the possibility for change. In this sense, a variety of constitutional forms of democracy respond to liberal views, including versions of representative government as well as forms of plebiscite. Liberalism is not anarchism, but anarchism is in some ways an extreme form of liberalism. The law has a key role in liberal thinking, but for a long time the prevalent interest of liberals was that of liberating people from the fetters of control imposed by the tangible force of the state (and the church) or the abstract force of tradition. Not surprisingly, some authors took this intention of liberation to its extreme. If they believed in the essential goodness of man, they advocated the abolition of all social restraint; at times, Jean-Jacques Rousseau seems to argue this

way. If on the other hand they believed in the ambivalence of human nature, they were not afraid to demand unlimited room for manoeuvre for ‘the singular one and his property’ (Max Stirner, 1845).

Perhaps this anarchist strain in early liberal thinking can be said to have been one of the reasons for the counter-reaction of the 19th century. Marx was the first to point out the historical advance brought about by ‘bourgeois’ equality before the law, including the contractual basis of economic action, but also the price paid by many for the ‘anarchic’ quality of the resulting market. The market – it was increasingly argued – was in fact not neutral, but favoured certain players to the systematic disadvantage of others. Mass poverty, conditions of labour, the state of industrial cities were cited as examples. Nor was this merely a view of anti-liberals. The great ambiguities in the thinking of John Stuart Mill tell the story.

There are two ways of describing the resulting history of thought and of social movements. One is to say that as the 19th century progressed, and certainly in the early decades of the 20th century, liberalism was replaced by socialism as a dominant force. People began to shrink back from the unconstrained market and sought new kinds of intervention. Today, authors would add that the ‘structural change of the public’ (J. Habermas, 1962) and the bureaucratization of democracy followed suit. Liberalism died a ‘strange death’; it ceased to be a source of reform and became a defence of class interest.

Another view ascribes the new reforms to liberals also, albeit to a different kind of liberalism. In his Alfred Marshall Lectures of 1949, T.H. Marshall (1950) argued that the progress of citizenship rights had to involve, from a certain point onwards, their extension from the legal and the political to the social realm. Social citizenship rights turned out to be a necessary prerequisite for the exercise of equality before the law and universal suffrage. Thus, the social, or welfare state was no more than a logical extension of the process which began with the revolutions of the 18th century.

There is much to be said for this line of argument if one considers that the two men who above all determined the climate of political thought and action from the 1930s to the 1970s, John Maynard Keynes and William Beveridge, were both self-declared liberals. In effect if not in intention, they advanced ideas which led to restrictions on the operation of markets. One will be remembered as the author of economic policy as a deliberate effort by governments, the other has contributed much to the creation of transfer systems which are operated by governments in the light of an assumed common interest. In other words, these were liberals who pursued policies which led to strengthening rather than limiting the power of public authorities. Theirs was a substantive, a social liberalism.

Liberal parties have found it difficult to follow the twists of theoretical liberalism. Before the First World War, when socialist parties were still in their infancy and unable to determine policy in any major country, they were often the spokesmen of the deprived and underprivileged. At least one strand of the liberal tradition continued to be reformist. However, after the First World War, socialists or social democrats came to form governments in many countries. Their gain was the liberals’ loss. Liberal parties declined to the point of insignificance, unless they merely kept the name and changed their policies out of recognition, either in the direction of social democracy (Canada) or in that of conservatism (Australia). Indeed, as a practical political movement, liberalism came to present such a confused picture that Hayek could argue that liberalism has become a mere intellectual, and not a political force.

The experience of totalitarianism interrupted this process without stopping it altogether. To the dismay but also to the surprise of many, basic human rights and the rules of the game of civil government became an issue again in the 1930s and 1940s. This gave rise to an important literature in which the underlying values of liberal thought were spelt out anew. Hayek's *Road to Serfdom* is one example, but

the most important one is probably Karl Popper's *Open Society and Its Enemies* (1952). Popper developed above all what might be called the epistemology of liberalism. We are living in a world of uncertainty. Since no one can know all answers, let alone what the right answers are, it is of cardinal importance to make sure that different answers can be given at any one time, and especially over time. The path of politics, like that of knowledge, must be one of trial and error. The principle can be applied to economy and society as well.

The liberal revolt against totalitarianism waned with the memory of totalitarianism itself. While the term ‘social market economy’ was coined for Germany in the 1950s, the quarter-century of the economic miracle was in fact a social-democratic quarter-century. In it, economic growth was combined almost everywhere with a growing role of government and with the extension of the social state. Entitlements came to matter as much as achievements. Consensus counted for more than competition or conflict. Despite variations, this was a very successful period in the countries of the First World. But by the 1970s, the side effects of success had become major problems in their own right. These were not only obvious problems like environmental and social ‘limits to growth’, but systematic ones arising from the role of the state. Both Keynes and Beveridge gave rise to new questions. Neither stagflation in the 1970s nor boom unemployment in the 1980s seemed amenable to government intervention. The social state had got out of hand; it became harder and harder to finance, and its bureaucracies robbed it of much of its plausibility. There were demands for a reversal of trends.

Where such a reversal happened, it remained bitty, halting and inconsistent. However, the new climate gave rise also to elements of a new theory of liberalism. In one sense, this was, and is a return to the original project of asserting society against the state, the market against planning and regulation, the right of the individual against overpowering authorities and collectivities. American authors in particular restated the theory. Milton Friedman tried to show in a series of arguments that the role of government is usually contrary to the interests of people. Robert Nozick made a strong case for the ‘minimal state’ and against the arrogance of modern state power. James Buchanan (1975) and the ‘constitutional economists’ reconstructed the social contract and argued for severely limited rules and regulations, using the fiscal system as one of their main examples. This trend, more than the notion of supply-side economics (which in some ways is merely Keynes stood on his head) signifies the revival of liberalism. There are other facets of the many-faceted term. For many, the extension of civil rights to hitherto disadvantaged groups is a liberal programme. Others still concentrate on the separation of church and state and the reduction of church influence. Again others regard liberalism as an advocacy of cultural values, including pluralism and creativity. It is not difficult to see the connection of such preferences with the mainstream of liberal thought.

This mainstream has three elements. Liberalism is a theory and a movement of *reform* to advance *individual liberties* in the horizon of *uncertainty*. This means by the same token that the prevailing theme of liberalism cannot be the same at all times. In the face of absolutism, it is liberty under the law; in the face of market capitalism, it is the full realization of citizenship rights; in the face of the ‘cage of bondage’ (Max Weber, 1922) of modern bureaucratic government, it is the optimal, if not the minimal state. The struggle for the social contract has become virulent in the advanced free societies. The crisis of the social state, the new unemployment, issues of law and order all raise basic questions of what is Caesar's and what are therefore the proper limits of individual desires. It is no accident that constitutional questions have come to the fore in several countries. At such a time, liberalism is gaining new momentum. It will not solve all issues, but it will remain a source of dynamism and progress

towards more life chances for more people.

See Also

- invisible hand
- libertarianism
- property rights
- utilitarianism and economic theory

Bibliography

Buchanan, J. 1975. *The Limits of Liberty*. Chicago: University of Chicago Press.

Habermas, J. 1962. *Strukturwandel der Öffentlichkeit*. Neuwied: Luchterhand.

Hume, D. 1740. *A Treatise of Human Nature*. Ed. L.A. Selby-Bigge, Oxford: Clarendon Press, 1888.

Kant, I. 1784. Idee zu einer allgemeinen Geschichte inweltbürgerlicher Absicht. In *Kants Populäre Schriften*, ed. P. Menzer, Berlin: Georg Reimer, 1911.

Locke, J. 1690. *Second Treatise of Government*. Ed. T.P. Peardon, New York: Liberal Arts Press, 1952.

Marshall, T.H. 1950. *Citizenship and Social Class*. Cambridge: Cambridge University Press.

Popper, K.R. 1952. *The Open Society and its Enemies*, 2nd edn. London: Routledge & Kegan Paul.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Oxford University Press, 1976.

Stirner, M. 1845. *Der Einzige und sein Eigentum*. Leipzig: D. Wigand.

Weber, M. 1922. *Wirtschaft und Gesellschaft*, 4th edn. Töbingen: Mohr/Siebeck, 1956.

How to cite this article

Dahrendorf, Ralf. "liberalism and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000085> doi:10.1057/9780230226203.0967

The New Palgrave Dictionary of Economics Online

libertarianism

David Friedman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Libertarians favour coordination by voluntary decentralized mechanisms such as private property and trade. In response to economic arguments for government intervention in the market, they point to the existence in the real world of private solutions to many problems of market failure and the ubiquity of market failure in political markets. Libertarians differ among themselves in the degree to which they rely on rights-based or consequentialist arguments and on how far they take their conclusions, ranging from classical liberals, who wish only to drastically reduce government, to anarcho-capitalists who would replace all useful government functions with private alternatives.

Keywords

anarchism; anarcho-capitalism; antitrust; Arrow, K.; Bork, R.; Coase, R.; consequentialism; democracy; depletable resources; efficiency; externalities; first efficiency theorem; flat tax; Food and Drug Administration (USA); free trade; Friedman, M.; George, H.; Harsanyi, J.; higher education; Hotelling, H.; human capital; immigration; imperfect competition; intellectual property; land tax; laissez-faire; left libertarianism; libertarianism; liberty; limited liability; lobbying; Locke, J.; market failure; negative income tax; Nozick, R.; objectivism; paternalism; positive and negative rights; pressure groups; professional licensing; property rights; public choice; public enforcement of law; public goods; Rand, A.; rational ignorance; redistribution of income and wealth; rent seeking; rights; Rothbard, M.; schooling; status; tariffs; Tiebout, C.; utilitarianism; victimless crimes; welfare state

Article

Libertarians, in current American usage and in this essay, are those who prefer to organize the world through the decentralized mechanisms of private property, trade, and voluntary cooperation rather than through government. Their position is thus a modern variant of the liberalism of the 19th century. Libertarians are likely to be critical of eminent domain, government regulation of business, paternalistic

social policies, income redistribution, laws banning ‘victimless crimes’ such as drug use, gambling and prostitution, and much else. Since there are good arguments for government as well as good arguments against, only a minority of libertarians carry their position all the way to anarchism. Most accept some level of taxation to pay for the production of public goods such as national defence. Some accept government production or subsidy of things well short of pure public goods, such as schooling.

The term ‘libertarian’ is also sometimes applied to left anarchists, usually outside of the United States; its original meaning seems to have been believers in free will. The current American usage is largely a response to the shift in the meaning of ‘liberal’ over the first half of the 20th century. Since believers in what used to be called liberalism could no longer use that term without confusion, many adopted ‘libertarian’ as a substitute.

One reason for libertarians to support a less than perfectly libertarian society is the belief that, in terms of individual liberty, it is the best we can do. A second is the belief that, while liberty is important, it is not the only thing that is important. Support by many libertarians for government funding of some public goods – scientific research and public health are examples – is based on the idea not that their production makes us freer but that it makes us better off in other ways.

In this article I sketch the general arguments for a libertarian position, discuss libertarian views on particular issues, and finally consider different forms of libertarianism and the internal disagreements that define them.

Why liberty is right

Libertarian conclusions may be supported either by showing that restraints on individual liberty are wrong or by showing that they lead to undesirable consequences. The former approach is often put in terms of individual rights. Each person has a right to control his own body, a right violated by laws against using drugs, by a military draft, and by many other government acts. Each person has a right to control his legitimately acquired property, a right violated by taxation, regulation, price controls,….

Putting the argument in this form raises an obvious question: how to justify such claims. Libertarians offer a variety of answers, ranging from Objectivists, who believe that individual rights can be logically deduced from the nature of man, to intuitionists, who induce them by trying to generalize their moral intuitions (Rand, 1964; Den Uyl and Rasmussen, 1991; Rothbard, 1978; Lester, 2000; Nozick, 1974; Boaz, 1997; 1998).

It also raises questions about how rights are acquired and how far they extend. Almost nobody argues that my right to control my body includes the right to punch you in the nose. Whether it includes the right to make noise on my property that keeps you awake or burn coal in my fireplace whose smoke makes you cough is less clear.

Robert Bork, in the article (Bork, 1971) explaining why he was not a libertarian, argued that my disutility from knowing that you are doing something I disapprove of is just as real an externality as my disutility from breathing your smoke, hence that there is no rights-based case for individual freedom as libertarians understand it. If we treat everything I do that affects others without their consent as a trespass liable to be enjoined, we are left with no self-regarding actions and no liberty – the exception swallows the rule. A response from the standpoint of moral philosophy depends on some way of deriving rights that distinguishes between those sources of disutility to me that do and those that do not violate my rights – hitting me over the head versus living your life in a way I disapprove of.

The economic response starts by observing that the enforcement cost of a rule giving me control over my own body is low, since I already control my body. The enforcement cost of giving you control over my body is substantial. Hence the latter alternative is an inefficient definition of property rights, at least unless my use of my body clearly imposes substantial and measurable costs on you that cannot be dealt with by voluntary transactions along Coasean lines. Although your disutility from knowing that I am reading pornography may be just as real as your disutility from breathing my smoke, it is considerably harder to demonstrate to a court, so a liability rule awarding you damages for the disutility you suffer from my reading pornography is likely to result in inefficient outcomes and substantial litigation costs. Alternatively, a property rule giving you rather than me a property right in my behaviour – requiring me, before doing anything, to get permission from everyone who objects – imposes transaction costs due to the hold-out problem sufficient to guarantee that nobody ever does anything, which is unlikely to be the efficient outcome. Following out this line of argument provides a defence of libertarian conclusions on consequentialist grounds.

‘Liberty’ and ‘rights’ are rhetorically powerful words, so it is not surprising that libertarians are not the only ones who claim them. Competing uses can be clarified by distinguishing between negative rights (‘the area within which a man can act unobstructed by others’, Berlin, 1969, p. 122) and positive rights. A negative right is a right to be left alone. A positive right is the right to some outcome. The right not to be killed is a negative right, the right to live – implying the right to be provided with what you need to live, such as food – a positive right. Other positive rights sometimes claimed include the right to decent housing, adequate food, medical care and equal treatment.

One problem with positive rights is that they contradict negative rights, including some that many find persuasive. If I have the right to decent housing and medical care, someone else must have the obligation to produce them, which is inconsistent with his right to control his own body. If I have the right to equal treatment, the right not to have an employer or homeowner decide whether to deal with me on the basis of my race or religion, someone else does not have the right of freedom of association, since he is required to deal with me even if he prefers not to. If I have the right not to be hated or despised for my sexual preferences, that means that I have a claim over the inside of your head, that being where your emotions are to be found. Thus the assertion of positive rights can be seen, and by libertarians often is seen, as the claim that some people are to some degree the slaves of others, required to serve them without having consented to do so – the violation of a deeply held negative right.

A second problem with positive rights is that they are more prone to internal inconsistency than negative rights. There is no conflict between my not killing or enslaving you and your not killing or enslaving me. But there is a conflict between my having adequate food, housing and medical care and your having them, if one or another of those goods happens to be in short supply.

Why liberty is useful

Large parts of the consequentialist argument for individual freedom go back to Adam Smith and should be familiar to every economist. Private property, exchange, prices provide a decentralized coordination mechanism that makes it possible for individuals with different objectives, knowledge and abilities to cooperate while pursuing their separate ends. In the limiting case of perfect competition, the result is provably efficient in the usual economic sense – cannot be improved by even a perfectly intelligent central planner with unlimited control over the actions of the planned. (For both the classical and

modern versions of the First Efficiency Theorem, see Arrow, 1983, and references therein. For a non-technical sketch of the classical version, see D. Friedman, 1997, ch. 16.)

The fact that this argument is correct, non-obvious, and included in the professional training of any economist is part of the reason why libertarianism is more popular with economists than with most other academics and why even non-libertarian economists tend to be sympathetic to market approaches. To put it differently, one important reason for the rejection of libertarian conclusions by non-economists is the failure to understand price theory – how markets solve the coordination problem.

The case against

Yet not all economists, not even all good economists, are libertarians. The economic counter-argument starts with the facts that real markets are imperfectly competitive and real individuals are limited by, at least, imperfect information, transaction costs, and limited calculating ability. Once we drop the assumptions of the ideal model we are faced with the possibility of market failure, situations where individual rationality fails to lead to group rationality and hence where it is possible for restrictions on the actions of each to produce a better outcome for all. Familiar examples include the underproduction of public goods, the overproduction of negative externalities, and potentially beneficial transactions blocked by adverse selection.

These are real problems, but not always insoluble ones. A market failure results in an outcome inferior, for all concerned, to some alternative outcome. A sufficiently ingenious entrepreneur may be able to create that alternative and collect a share of the net benefit as his reward; a market failure is also a profit opportunity. Radio broadcasts are a pure public good produced privately. So are the services that Google provides to its users. Other forms of market failure may be dealt with by the development of systems of private norms (Ellickson, 1991; Posner, 2000). Where market failure exists we can expect private arrangements to produce imperfect outcomes, but less imperfect than casual consideration might suggest. (For an interesting example of a real world solution to a theoretically intractable market failure, see Cheung, 1973.)

A second objection to the argument for laissez-faire is that efficiency as defined in economics in the sense of Marshall or Hicks–Kaldor (D. Friedman, 1997, ch. 15) is inadequate as a normative criterion, so that a less efficient outcome may be preferable to a more efficient one. What is maximized by the market is value defined by willingness to pay, measured in dollars not utiles, so a transfer from rich to poor might decrease value measured in dollars but increase total utility.

This utilitarian argument for redistribution can be seen as a special case of the argument from market failure. Declining marginal utility is not merely a conjecture of philosophers; it is observed, in the form of risk aversion, in individual choices under uncertainty. In a perfect market, individuals would buy insurance against the risk of being born poor up to the point where the marginal utility costs of any resulting disincentives or transactions costs just balanced the marginal utility gain of transferring income from states of the world where they were rich to ones where they were poor. Thus the outcome of a perfect market would mirror the welfare programme that would be proposed by a utilitarian. It is merely our inconvenient inability to negotiate and sign insurance contracts prior to being born that prevents the market from solving the problem. The argument for utilitarianism in Harsanyi, 1955 – that it is what individuals would choose if they were designing a society behind a veil of ignorance with an equal

probability of living any of its lives – makes it possible to view redistribution of income either as a way of increasing total utility or as a correction for market failure.

Other objections to market outcomes come from egalitarians who see equality as good in itself and from those who put substantial weight on values unrelated to individual humans achieving their objectives. If what really matters is the preservation of endangered species, whether or not of any value to human beings, there is no guarantee that the market to achieve it. The same is true if what really matters is behaving according to God's will, producing great art and literature, or doing justice whatever the consequences.

A libertarian response

It follows that one can imagine outcomes that improve, in one sense or another, on the outcome of pure laissez-faire. It does not follow that one can construct institutions that predictably produce such outcomes.

Consider the case of market failure. It exists because actions taken by A sometimes have effects on B. If A is free to ignore those effects he may make the pair on net worse off by taking actions that increase his welfare by less than they decrease B's or failing to take actions that would increase B's welfare by more than they decrease A's. A well-designed legal structure can sometimes make it in A's interest to take account of those effects, whether through property rules, liability rules, or bargaining between the parties. But sometimes, for reasons explored by Coase (1960) and others (D. Friedman, 2000, pp. 39–45), no legal structure can be constructed that makes it in the interest of all parties to make the efficient choices.

All this is true in private markets. But it is true far more often in the political markets that control the political institutions that are proposed as a solution to market failure in private markets.

Consider the naive model of democracy – politicians doing good because if they do not they will lose the next election. In order for it to work, individual voters have to acquire the information needed to know what politicians are doing and whether it is good. No politician campaigns on the slogan ‘I'm the bad guy’. No farm bill is labelled ‘An act to make farmers richer and city folk poorer’.

If I correctly identify the better candidate, vote for him, and – improbably – my vote proves decisive, the benefit is shared with everyone in the polity. The cost is borne by me alone. Time and energy spent acquiring the information necessary for informed voting produce something very close to a pure public good. Public goods are underproduced; one with a public of many millions is likely to be very badly underproduced. The implication is rational ignorance, voters failing to acquire the information they need to judge politicians because its value to them is less than its cost. That eliminates the simple argument for why politicians will find it in their political interest to act as we would wish them to.

A similar problem arises with a more sophisticated model in which political outcomes are driven by interest group pressure. The more an interest group stands to gain by passing or blocking a piece of legislation, the more it will offer politicians in order to support or oppose it. If that were the only relevant factor, the market for legislation would produce something close to an efficient outcome. If a bill produced net benefits, its supporters would spend more supporting it than its opponents spent to block it, and the bill would be likely to pass.

It is not the only relevant factor. An interest group lobbying for legislation is producing a public good for its members and faces an internal public good problem in doing so, since members that refuse to

contribute will still benefit if the bill passes. Some interest groups are much better able than others to solve their internal public good problem. A concentrated interest group such as the auto industry – a handful of firms and one union – can raise a substantial fraction of the benefit it expects from an auto tariff in order to lobby for it. A dispersed interest group such as consumers of automobiles and producers of export goods, the people that bear most of the burden of such a tariff, can raise a negligible fraction of the cost to lobby against. Hence we would expect the political market to consistently redistribute from dispersed interest groups to concentrated ones, even when the benefit to the latter is much smaller than the cost to the former – as demonstrated by the continued existence of tariffs nearly two centuries after Ricardo demonstrated that they are, under most circumstances, injurious to the nation that imposes them.

In a private market, a producer receives a price that measures the value to consumers of what he produces, pays a cost that measures the cost to the suppliers of his inputs of producing them, and pockets the difference. It is only when special circumstances arise – externalities that cannot be dealt with by the market, information asymmetry, and the like – that his actions impose net costs or benefits on others. In the political market, in contrast, almost all decisions are made by people who bear few of the costs and receive few of the benefits those decisions produce. A legislator who passes an auto tariff imposes net costs of many billions of dollars on those affected, but all that comes out of his pocket is the extra cost of the car he buys. A judge whose precedent establishes a seriously inefficient legal rule might reduce national income by, say, a tenth of a percentage point – a staggering amount of damage for a single human being to do. But not only will he not pay any of the cost, he will never even know he made a mistake.

Consider, for example, *Davis v. Wyeth Laboratories, Inc*., 399 F.2d 121 (9th Cir. (Idaho) Jan 22, 1968), where the court found Wyeth liable for the failure to adequately warn of the risk of polio vaccination. Their argument hinged on whether, if warned, Davis might reasonably have chosen not to be vaccinated. The court wrote: ‘Thus appellant's risk of contracting the disease without immunization was about as great (or small) as his risk of contracting it from the vaccine. Under these circumstances we cannot agree with appellee that the choice to take the vaccine was clear.’ They reached this conclusion by comparing the 0.9 in a million chance of getting polio from the vaccination with the 0.9 in a million *annual* rate of adult polio from natural causes. Since vaccination provided protection for many years, possibly a lifetime, the proper comparison was to the risk over many years, not one. The court made a mathematical error of more than an order of magnitude, set a precedent which substantially discouraged the development of new vaccines, caused many, perhaps thousands, of unnecessary deaths, and suffered no penalty for doing so.

Market failure is a real problem. It is a problem in ordinary private markets and a much more severe problem in political markets. That is an argument for shifting decisions, so far as possible, from political to private markets – an argument for, not against, the libertarian position.

A possible response is that decisions should be shifted to public markets only where private markets fail. But some degree of market failure can be alleged for almost any activity. Under legal rules permitting government intervention to correct any alleged market failure, intervention can be expected whenever it is politically profitable.

Libertarians vary in how far they are willing to push the arguments that I have just sketched. Consider the case of national defence, a public good with a very large public. The failure to produce it privately at an adequate level is likely to lead to a drastic reduction in liberty. That is an argument sufficiently strong to convince many, although not all, libertarians to include it in the proper functions of government.

So far I have been dealing with arguments based on market failure, but similar point cans be made with regard to other criticisms of market outcomes. It is true that the market takes account of values only to the extent that individuals do; if nobody cares about the survival of the oldest tree in the world or some threatened species of birds, there is no reason to expect the market to preserve it. But the same is true of the political system. It too is driven by the desires of individuals. It just does a much clumsier job of satisfying them.

Indeed, there are some reasons to expect the market to do a better job of serving ‘non-economic’ values than the political system. Many are things, not that nobody cares about, but only that most people don't, and the market is generally better at providing for small minorities than the political system. A religion followed by a per cent or two of the population has no difficulty getting the market to produce copies of its scriptures. If it is sufficiently unpopular with the majority, it may have problems getting the government to permit them to be printed. A minority in power might be able to do a better job of diverting resources to serve its values, whether religious or environmental, through the political system than through the market. But shifting decisions to the political system for that reason could be a risky gamble.

Another common criticism, but a mistaken one, is that the market ignores the interest of future generations. Future as well as present demand counts. It is worth planting hardwoods today for harvest a century hence as long as the return is at least as great as from alternative investments. Markets allocate resources over time, as Hotelling (1931) showed, in an economically efficient fashion. If it can be predicted that petroleum will be very valuable a century hence, it is profitable to leave it unpumped now so as to sell it then.

This argument depends on secure property rights. It breaks down if oil saved or a tree planted today is likely to be expropriated tomorrow, making holding it for future use a poor gamble. The alternative to decisions by the market is decisions by political mechanisms. Property rights in the political marketplace are much less secure than those in the private marketplace. A president who accepts costs today for benefits 10 or 20 years in the future can be reasonably confident that neither he nor his party will receive credit for those benefits. A dictator, unlike an entrepreneur, rarely has the opportunity to collect the benefit from investments expected to pay off in the future by transferring his long-term assets to a successor in exchange for immediate payment. Hence we would expect political institutions to be much more inclined to sacrifice the future to the present than market institutions, a conclusion supported by the evidence of environmental policy in the Soviet Union and Social Security in the United States.

What about income redistribution? Here again, the question is not whether there is an outcome that some would prefer to that produced by the market but whether there are institutions that predictably create such an outcome. The equal distribution of votes gives the poor some advantage on the political marketplace, but it may easily be outweighed by the very unequal distribution of other politically relevant resources. Modern governments are observed to redistribute from rich to poor via welfare, from poor to rich by subsidies for art, music, and – the big one – higher education, paid for mostly by state and local taxes and consumed mostly by people from the upper part of the income distribution. (The median family income of US college freshmen in 2001 was $67,200, compared with a median family income for all households of $42,228 – US Census Bureau, 2003, Tables 284 and 683. See Gwartney and Stroup, 1986, for a discussion of theory and evidence of the consequences of redistributional policies.) Similarly, farm policy provides a subsidy mostly to wealthy farmers and pays for it mainly by a regressive tax in the form of higher food prices.

A second problem with redistribution is rent seeking. In a polity that redistributes, it is in the interest of nearly everyone to spend resources trying to shift the redistribution in his favour, opposing redistribution from him and promoting redistribution to him (Tullock, 1967; D. Friedman, 1973, ch. 38; Krueger, 1974). The resulting deadweight cost might easily outweigh any utility gain from redistribution.

Issues

Libertarians differ in how far they are willing to carry their libertarianism. In the following discussion I present libertarian positions and the arguments for them while recognizing that in many cases the libertarian position is not supported by all who consider themselves libertarians.

The easy cases

Most of the arguments against price control, wage control, rent control, usury laws, and similar restrictions on the terms of market exchange are familiar to any economist. Many libertarians also argue that such restrictions violate individual rights. If I own my body, it is up to me to decide on what terms I will sell my labour to you. If I own my house, it is up to me to decide what terms I am willing to offer to potential tenants and up to them to decide what terms they are willing to accept. Thus many libertarians would reject not only rent and wage control but also legal restrictions on private discrimination in home sales, employment, and the like. (Nozick, 1974, ch. 7, provides an extended discussion and defence of a libertarian view of self-ownership.)

Libertarians taking that position may defend it either in terms of individual rights or by arguing that minorities are worse off in a world where such decisions are controlled by government than in one where they are controlled by private contract. State intervention in the US South during the first half of the 20th century provides an obvious example. A prejudiced majority can do a great deal more harm to the minority it is prejudiced against where decisions are made by the government than where they are made privately.

Free trade is another easy case. If building cars in Detroit costs more than growing grain, putting it on ships, sending them out into the Pacific, and having them come back with Hondas on them, we are better off growing our cars instead of building them. A tariff forces us to use the more expensive technology instead of the less expensive; it protects American auto workers from the competition of American farmers, making Americans on the whole worse off. While economists can construct special circumstances in which a trade restriction might benefit the nation that imposed it, such as infant industries that require temporary protection, the restrictions we observe are not those suggested by such arguments: In the U.S., steel and auto are not infant industries. We observe instead the restrictions predicted by the public choice analysis offered earlier, policies that benefit concentrated interest groups at the expense of dispersed interest groups. (For an explanation of why tariff protection is particularly likely for declining industries such as steel, see D. Friedman, 1997, p. 294.)

Many libertarians find paternalism another easy case, since it contradicts the idea that each individual owns his own body and is free to make choices regarding it. As a practical matter, paternalistic regulations substitute for each individual's decisions about his own welfare the decisions of someone else. The regulator may have expert information the individual lacks, but he lacks both the individual's

specialized knowledge about his own circumstances and the individual's incentive to act in that individual's interest. Thus professional licensing, justified as a paternalistic protection of the consumer, is in practice used by professions to reduce competition and so benefit themselves at the expense of their customers. (The classic discussion is M. Friedman, 1962, ch. 9). Similar arguments apply to laws against victimless crimes – the War on Drugs, laws against prostitution and gambling. Individuals might make the wrong decisions for themselves; others should be free to warn them against doing so. But the final decision ought to be made by each individual for himself.

A familiar example of the dangers of such regulation in the United States is the Food and Drug Administration (FDA). Letting a dangerous drug onto the market ends the regulator's career. Keeping a drug off the market for a few more years can do enormous damage – arguably an excess mortality on the order of a hundred thousand lives in the case of beta-blockers (Gieringer, 1985. For a webbed discussion, see FDAReview.org.) But damage that appears only in the mortality statistics is very nearly irrelevant, politically speaking. And the connection between over-regulation, higher prices and fewer new drugs is still less visible. (See Peltzman, 1973, for a classic examination of the effect of regulation on quality and rate of introduction of new drugs.)

Antitrust

There are legitimate arguments, widely supported by economists, in favour of government intervention against monopolies. Even libertarians are troubled by hypotheticals in which one firm owns the only well in the desert and insists on thirsty travellers giving all they own and indenturing their labour for decades into the future in exchange for a drink. Government regulation of monopoly, however, has its own problems. The regulator needs information he is unlikely to have – cost curves and demand curves – in order to force the firm to follow welfare-maximizing rather than profit-maximizing strategies (D. Friedman, 1997, pp. 238–43). And it is far from clear why a real-world regulator, driven by political rather than altruistic incentives, would attempt to regulate in the public interest rather than letting himself be captured by the regulated industry, a concentrated interest well positioned to reward politicians with money and regulators with future jobs (Stigler, 1971). An industry that is imperfectly competitive may be imperfectly efficient, but the situation is not improved by giving firms the opportunity to use government regulation, as the US railroad industry used the Interstate Commerce Commission (ICC), to exclude competitors and restrict competition (Kolko, 1977).

Such considerations persuade many libertarians that antitrust, both as a legal doctrine and as a basis for regulation, does more harm than good – that we would be better off putting up with any ills private monopoly may produce, since the cure is likely to be worse than the disease (M. Friedman, 1962, pp. 128–9). Others argue that the state need not prevent monopoly but ought not to support it, and can avoid doing so by refusing to enforce contracts in restraint of trade.

Immigration

The economic arguments for free movement of goods apply to capital and labour as well, implying that immigration produces net benefits for the country that permits it, just as free trade produces net benefits for the country that practises it. Freer immigration also produces what many would consider a desirable

redistribution, since its major beneficiaries, the immigrants, are much poorer than those who might be made worse off by their move: workers in the country the immigrants go to, capitalists and landowners in the countries they come from.

This assumes a context of voluntary transactions. Some immigrants may come in order to profit by involuntary transactions, private or political – to commit robbery or collect welfare. And new immigrants, once they become citizens and voters, might use the political mechanism to advantage themselves at the cost of the rest of us. Such arguments help explain why not all libertarians support free immigration – despite empirical evidence that, at least under current circumstances, immigrants pay more in taxes than they collect in benefits (Simon, 1989; 1995).

The flip side to the 'immigrant as welfare recipient' argument is that, while the existence of a welfare state makes the desirability of free immigration less clear, free immigration makes it more difficult to maintain a welfare state. Free movement of people imposes limits on the ability of governments to exploit those they rule, similar to the limits that market competition imposes on the ability of firms to take advantage of their customers (Tiebout, 1956). For libertarians, that is an additional advantage to freer immigration.

Schooling

The usual argument for government provision or subsidy of schooling is that a democracy requires educated voters and an economy educated workers, hence that money spent educating my children benefits you and your children, hence that leaving education to the free market will result in too little. The first part of that argument might be true, although it is hard to find evidence to support it. The second is simply bad economics. To the extent that education makes a worker more productive, the additional productivity is reflected in his wages; investing in human capital is no more a public good than investing in physical capital. In both cases the investor may receive less than the full value of his investment due to the distorting effect of taxation – some of my additional productivity goes, not to me, but to the Internal Revenue Service. But subsidizing the investment merely shifts the inefficiency to whoever pays the taxes that fund the subsidy.

There may be indirect externalities to subsidized education – a cure for cancer, say. But not all such externalities are positive. By educating my children I make them better able to use the political system to advantage themselves at the expense of your children. By sending my son to Harvard I give him an opportunity to feel superior to your son, who went to Podunk U. That is a benefit to me and my son, a cost to you and yours, and a negative externality produced by my expenditure on education. As Robert Frank (1986) has persuasively argued, one of the things humans care about and economists ought to take account of is relative status.

This example illustrates a common problem with arguments based on externalities. Those making them usually count only externalities that lead to the conclusion they want – positive if they want to subsidize something, negative if they want to ban it. If an activity produces both positive and negative externalities, as many do, and if we are unable to measure them accurately enough to determine the sign of their sum, we do not know whether we should be encouraging the activity or discouraging – in which case it might be wiser to do neither (D. Friedman, 1971).

Another argument for government involvement in schooling is that, since parents act in their own

interest rather than that of their children, they may fail to pay the cost of schooling even when it produces a benefit larger than its cost. But shifting the decision to the political system means shifting it, not to children, but to other adults. Adults routinely make large sacrifices on behalf of their children, much more rarely on behalf of other people's children. So while a parent is not a perfect proxy for his children, he may be the best proxy available – a much better one than either the legislature or the teachers' unions.

Other government activities can be supported, and opposed, with similar arguments. Subsidies for basic research can be defended as producing a public good, rejected on the grounds that enough of the benefits can be privatized to make subsidy unnecessary (Kealey, 1997), that government involvement diverts too many smart people into whatever field is currently in fashion, and that it subverts the scientific enterprise by converting the search for truth into a search for grants.

The relevance of public good theory is less clear for police and courts, government activities traditionally accepted by believers in a minimal government. Law enforcers can choose to pursue criminals who commit crimes against those who have paid for their services and not those who have not; England survived with private thieftakers but without police in the modern sense until well into the 19th century. (Davies, 2002; D. Friedman, 1995. Both argue that there is no clear evidence that failure of the traditional system was the reason why it was eventually replaced.) Courts can refuse to settle disputes among those unwilling to pay for the service, and some – both private arbitrators and government courts – do. Many libertarians accept the conventional arguments for state provision of police and courts, paid for by taxation; others do not (D. Friedman 1973, part 3).

There are a few issues where libertarians disagree among themselves about which side is more libertarian. Intellectual property is one example. Some argue that a book or an invention, as the pure creation of a human mind, deserves strong protection. Others regard all intellectual property as coercive, a restriction on how individuals are permitted to use their own material property. Limited liability for corporations is another such. Many libertarians reject it on the grounds that individuals ought to be liable for their actions. Others see it as a legitimate consequence of freedom of association and contract and observe that, while it is possible for a corporation to impose costs it does not have the resources to compensate for, the same is true for an individual.

Foreign policy provides a particularly divisive example. Opponents of the United States in recent decades have been strikingly unfree societies – Hitler's Germany, Stalin's Russia, Mao's China, Ho Chi Minh's Vietnam – making a policy of overthrowing, or at least containing, them attractive to many libertarians. But such a policy is conducted by a government whose competence and motives libertarians find suspect – and badly done interventionism may well be worse than no interventionism (D. Friedman, 1989, ch. 45). Hence many libertarians favour the non-interventionist policy famously advocated by George Washington – peace and friendship with all, entangling alliances with none.

Libertarian: yes/no or more/less

Some libertarians propose a bright line definition of who is a libertarian, often along the lines of 'one who believes in never initiating force against another'. One problem with this is that libertarians do not have an entirely satisfactory account of what determines who owns what – in particular, of how unproduced resources, such as land, become property. Without a clear answer to that question, it is

sometimes hard to distinguish the initiation of force from the use of force to defend what you justly own. A second problem is that the bright line definition, taken literally, eliminates almost everyone, including almost all libertarians. Consider a scenario popularized by the late R.W. Bradford, editor of *Liberty Magazine*. You have carelessly fallen out of a 50th storey window. By good luck, you catch hold of the flagpole of the apartment immediately below you and start trying to climb in the window. The owner of the apartment objects that you are violating his property rights – not only by climbing in his window, but by using his flag pole without his permission. Do you let go and fall to your death? Such arguments suggest that 'libertarian' is more usefully defined as a continuum – more libertarian or less rather than libertarian or not.

An issue which has attracted a good deal of attention within the libertarian movement is whether there ought to be any government at all. One faction, sometimes labeled 'minarchist', supports a government that provides, at least, for courts, police, and national defence. The other – anarchists or anarcho-capitalists – argues that, with suitable institutions, voluntary cooperation in a free market can adequately provide all government services worth providing (D. Friedman, 1989, part 3; Rothbard, 1978). The latter position can be defended either on the (rights-based) grounds that all other alternatives involve violations of rights or on the (consequentialist) grounds that, just as the free market does a better job than government of building cars and growing food, it could also do a better job of producing laws and defending rights. While the latter claim seems obviously false to many when they first encounter it, it has proved sufficiently persuasive to be adopted by a significant minority of those seriously involved with libertarian ideas and libertarian argument. (Liberty Magazine Editors, 1999.)

Varieties of libertarianism

Does 'individuals have the right not to be coerced' mean that one should never initiate coercion or that one should act to minimize coercion? If rights are best protected by a tax-supported system of police and courts, should one support such taxes as a way of minimizing rights violations or oppose them as a violation of rights? (Nozick, 1974, pp. 28–35, discusses the distinction between rights as side constraints and a 'utilitarianism of rights' and offers arguments for the former.) One answer makes anarchism something close to a moral imperative, the other decides the anarchist/minarchist issue in terms of how well either alternative works.

It is useful for land to be treated as private property. But how does a claimant get ownership? Locke (1689, ch. 5, section 27) famously argued that he did it by mixing his labour with the land – clearing trees, plowing, removing boulders. But that argument included the proviso that there be as much land and as good available for other claimants, since otherwise the first claimants deprive others of the opportunity to claim land themselves. The value of the land is in part site value and in part value due to human effort; how does the owner get a just claim to the former?

Many libertarians avoid these questions by simply accepting existing titles to land. Others argue that such claims are legitimate only if based on a chain of voluntary transfer back to a legitimate appropriation, whether by Lockean mixing of labour with land or some other mechanism. A few, 'geolibertarians' or, more confusingly, 'left libertarians', reject unqualified private ownership of land entirely, arguing for the land tax of Henry George or something similar (Brody, 1983; D. Friedman, 1983; Valentyne and Steiner (2000a; 2000b); George, 1879.)

For a final variant on libertarianism, consider someone who accepts both the utilitarian argument for redistribution from rich to poor and libertarian arguments against government intervention in the market. He might favour a laissez-faire society combined with some very simple system of redistribution – say a flat tax used to finance a modest demogrant. (The best-known proposal along these lines is the negative income tax; M. Friedman, 1962, pp. 191–5. A more recent version is Murray, 2006.) Making the redistribution simple reduces the opportunity for individuals to spend resources trying to shift it in their favour. Putting all redistribution in one form eliminates arguments for other government interventions defended – often implausibly – as helping the poor. While many, perhaps most, libertarians would be reluctant to consider this a fully libertarian position, it provides a possible compromise for those who accept large parts of the consequentialist argument for libertarian policies while remaining unconvinced by libertarian arguments about rights.

See Also

- antitrust enforcement
- externalities
- Friedman, Milton
- international migration
- laissez-faire, economists and
- public choice
- public goods
- rent seeking
- Rothbard, Murray N.

Bibliography

Arrow, K. 1983. An extension of the basic theorems of classical welfare economics. In *Collected Papers of Kenneth J. Arrow, Volume 2: General Equilibrium*. Cambridge, MA: Belknap.

Berlin, I. 1969. *Four Essays on Liberty*. Oxford: Oxford University Press.

Boaz, D. 1997. *Libertarianism: A Primer*. New York: Free Press.

Boaz, D. 1998. *The Libertarian Reader: Classic and Contemporary Writings from Lao Tzu to Milton Friedman*. New York: Free Press.

Bork, R. 1971. Neutral principles and some First Amendment problems. *Indiana Law Journal* 47, 1–35.

Brody, B. 1983. Redistribution without egalitarianism. *Social Philosophy and Policy* 1, 71–87.

Cheung, S. 1973. The Fable of the Bees: an economic investigation. *Journal of Law and Economics* 16, 11–33.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Davies, S. 2002. The private provision of police during the eighteenth and nineteenth centuries. In *The Voluntary City: Choice, Community, and Civil Society*, ed. D. Beito, P. Gordon and A. Tabarrok. Ann Arbor: University of Michigan Press.

Den Uyl, D. and Rasmussen, D. 1991. *Liberty and Nature*. Chicago: Open Court.

Ellickson, R. 1991. *Order Without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

FDAReview.org. Theory, evidence and examples of FDA harm. Online. Available at http://www.fdareview.org/harm.shtml, accessed 2 August 2006.

Frank, R. 1986. *Choosing the Right Pond: Human Behavior and the Quest for Status*. New York: Oxford University Press.

Friedman, D. 1971. Laissez-faire in population: the least bad solution. Occasional paper, Population Council. Online. Available at http://www.daviddfriedman.com/Academic/Laissez-Faire_In_Popn/L_F_in_Population.html, accessed 3 August 2006.

Friedman, D. 1973. *The Machinery of Freedom: Guide to a Radical Capitalism*. Chicago: Open Court, 1989.

Friedman, D. 1983. Comment on Brody ‘Redistribution Without Egalitarianism’. *Social Philosophy and Policy* 1, 88–93.

Friedman, D. 1995. Making sense of English law enforcement in the eighteenth century. *University of Chicago Law School Roundtable* Spring/Summer, 475–505.

Friedman, D. 1989. *The Machinery of Freedom: Guide to a Radical Capitalism*. Chicago: Open Court.

Friedman, D. 1997. *Hidden Order: The Economics of Everyday Life*. New York: Collins.

Friedman, D. 2000. *Law's Order: What Economics Has to Do With Law and Why It Matters*. Princeton: Princeton University Press.

Friedman, M. 1962. *Capitalism and Freedom*. Chicago: University of Chicago Press.

George, H. 1879. *Progress and Poverty*. New York: Robert Schalkenbach Foundation, 2003.

Gieringer, D. 1985. The safety and efficacy of new drug approval. *Cato Journal* 5(1), 177–201.

Gwartney, J. and Stroup, R. 1986. Transfers, equality, and the limits of public policy. *Cato Journal* 6(1), 111–37.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 64, 309–21.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Kealey, T. 1997. *The Economic Laws of Scientific Research*. New York: Palgrave Macmillan.

Kolko, G. 1977. *Railroads and Regulation, 1877–1916*. Westport, CT: Greenwood.

Krueger, A. 1974. The political economy of the rent seeking society. *American Economic Review* 64, 291–303.

Lester, J. 2000. *Escape From Leviathan: Liberty, Welfare, and Anarchy Reconciled*. New York: Palgrave Macmillan.

Liberty Magazine Editors. 1999. The liberty poll. *Liberty Magazine*, February, 11–22.

Locke, J. 1689. *Two Treatises of Government*, 2nd edn. Cambridge: Cambridge University Press, 1967.

Murray, C. 2006. *In Our Hands: A Plan To Replace The Welfare State*. Washington, DC: AEI Press.

Nozick, R. 1974. *Anarchy, State and Utopia*. Malden, MA: Blackwell.

Peltzman, S. 1973. An evaluation of consumer protection legislation: the 1962 drug amendments. *Journal of Political Economy* 81, 1049–91.

Posner, E. 2000. *Law and Social Norms*. Cambridge, MA: Harvard University Press.

Rand, A. 1964. Man's rights. In *The Virtue of Selfishness*. New York: Signet.

Rothbard, M. 1978. *For a New Liberty: The Libertarian Manifesto*, rev. edn. Lanham, MD: University Press of America.

Simon, J. 1989. *The Economic Consequences of Immigration*. Cambridge, MA: Blackwell.

Simon, J. 1995. *Immigration: The Demographic and Economic Facts*. Washington, DC: Cato Institute. Online. Available at http://www.cato.org/pubs/policy_report/pr-immig.html, accessed 3 August 2006.

Stigler, G. 1971. The theory of economic regulation. *Bell Journal of Economics and Management Science* 2, 3–21.

Tiebout, C. 1956. A pure theory of local public expenditures. *Journal of Political Economy* 64, 416–24.

Tullock, G. 1967. The welfare costs of tariffs, monopoly and theft. *Western Economic Journal* 5, 224–32.

US Census Bureau. 2003. *Statistical Abstract of the United States*. Online. Available at http://www.census.gov/prod/www/statistical-abstract-2001_2005.html, accessed 3 August 2006.

Valentyne, P. and Steiner, H., eds. 2000a. *The Origins of Left Libertarianism: An Anthology of Historical Writings*. New York: Palgrave.

Valentyne, P. and Steiner, H., eds. 2000b. *Left Libertarianism and Its Critics: The Contemporary Debate*. New York: Palgrave.

How to cite this article

Friedman, David. "libertarianism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000227> doi:10.1057/9780230226203.0968

The New Palgrave Dictionary of Economics Online

Lieben, Richard (1842 – 1919)

Jürg Niehans
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Auspitz, R.; consumer's rent; gold standard; Lieben, R.; mathematical economics; reciprocal demand curves; Walras, L.

Article

Lieben was born on 6 October 1842, in Vienna; he died there on 11 November 1919. After studying mathematics and engineering sciences, he became a partner in the Jewish family bank and a respected member of the Viennese business community. In 1892 he advocated the adoption of a gold standard. He married late and had no children. He seems to have been of scholarly and artistic tastes, more contemplative than active.

Together with his cousin and brother-in-law Rudolf Auspitz, Lieben wrote the ‘Researches on the Theory of Price’ (1889), the only Austrian contribution to mathematical economics and one of the outstanding contributions in the last two decades of the 19th century. (This book is discussed in the dictionary entry on Auspitz, Rudolf.)

As a correspondent to the *Economic Journal*, Lieben provided a lucid summary of their views on consumer's rent (1894). After his collaborator's death he concluded the controversy with Walras by a complex three-dimensional analysis of reciprocal demand curves (1908), gracefully acknowledging their original misunderstanding. Appropriate corrections were made in the French translation of the ‘Researches’ (1914). While it is impossible to separate Auspitz's and Lieben's contributions, these papers suggest that Lieben was more than a junior partner.

Selected works

1887. (With R. Auspitz.) *Zur Theorie des Preises*. Leipzig: Duncker & Humblot.

1889. (With R. Auspitz.) *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot.

French translation by Louis Suret, Paris: M. Giard & E. Brière, 1914.

1890. (With R. Auspitz.) Reply. (To article by L. Walras in same publication.) *Revue d'économie politique* 4.

1894. On consumer's rent. *Economic Journal* 4, 716–19.

1897. Ueber die weitere Ausdehnung des Wasserstrassennetzes in Oesterreich in der Zukunft. *Verbands-Schriften des Deutsch-Oesterreichisch-Ungarischen Verbandes für Binnenschiffahrt*.

1898. Indian currency. *Economic Journal* 8, 151–2.

1908. Die mehrfachen Schnittpunkte zwischen der Angebots- und der Nachfragekurve. *Zeitschrift für Volkswirtschaft, Sozialpolitik und Verwaltung* 17, 607–16.

Bibliography

Weinberger, O. 1931. Rudolf Auspitz und Richard Lieben. *Zeitschrift für die gesamte Staatswissenschaft* 91.

Winter, J. 1927. *Fünfzig Jahre eines Wiener Hauses*. Vienna: F. Jasper.

How to cite this article

Niehans, Jürg. "Lieben, Richard (1842–1919)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000088> doi:10.1057/9780230226203.0969

The New Palgrave Dictionary of Economics Online

life tables

Nathan Keyfitz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Life tables present the age incidence of mortality in a population. The population may be all those people in a country or other area, or some category within a country; it may be all persons counted at a particular moment or period of time, say 1980 (period table); or it may be those born at a particular time and followed through life (cohort table).

The abridged life table officially calculated for the United States deaths and population of 1983 (National Center for Health Statistics, 1983) is shown as Table 1. It is based on population estimated to mid-year ($_5P_x$ for age x to x+4 at last birthday) and extrapolated from the 1980 census, and corresponding deaths to residents occurring during the year 1980 ($_5D_x$). Unregistered deaths are few in developed countries, but population censuses tend to under count, and give a life table of too high mortality unless a correction is made. In most less developed countries registration of deaths is incomplete, and model (e.g. Coale and Demeny, 1983 or UN, 1982) tables fill the gap.

Abridged life tables by race and sex: United States, 1980

Age interval	Proportion dying	Of 100,000 born alive		Stationary population		Average remaining lifetime
Period of life between two exact ages stated in years (1)	Proportion of persons alive at beginning of age interval dying during interval (2)	Number living at beginning of age interval (3)	Number dying during age interval (4)	In the age interval (5)	In this subsequent age intervals (6)	Average number of years of life remaining of age interval (7)
X to x+n	$_nq_x$	L_x	$_nd_x$	$_nL_x$	T_x	$\$_x$
All races						
•0–1	0.0127	100,000	1,266	98,901	7,371,986	73.7
•1–5	.0025	98,734	250	394,355	7,273,085	73.7

•5–10	.0015	98,484	150	492,017	6,878,730	69.8
10–15	.0015	98,334	152	491,349	6,386,713	64.9
15–20	.0049	98,182	482	489,817	5,895,364	60.0
20–25	.0066	97,700	648	486,901	5,405,547	55.3
25–30	.0066	97,052	638	483,665	4,918,646	50.7
30–35	.0070	96,414	672	480,463	4,434,981	46.0
35–40	.0091	95,742	875	476,663	3,954,518	41.3
40–45	.0139	94,867	1,321	471,250	3,477,855	36.7
45–50	.0222	93,546	2,079	462,857	3,006,605	32.1
50–55	.0351	91,467	3,209	449,811	2,543,748	27.8
55–60	.0530	88,258	4,676	430,230	2,093,937	23.7
60–65	.0794	83,582	6,638	402,081	1,663,707	19.9
65–70	.1165	76,944	8,965	363,181	1,261,626	16.4
70–75	.1694	67,979	11,517	312,015	898,445	13.2
75–80	.2427	56,462	13,702	248,534	586,430	10.4
80–85	.3554	42,760	15,197	175,192	337,896	7.9
85 and over	1.0000	27,563	27,563	162,704	162,704	5.9

Source: National Center for Health Statistics (1983).

Having the age-specific death rates, ${}_5M_x={}_5D_x/{}_5P_x$, the important step is calculating the probability that a person living at the beginning of the age interval will survive to the end. If the death rate within the interval can be assumed constant then the exact probability is $l_{x+5}/l_x = e^{-5\,{}_5M_x}$. In fact, for ages from about 10 onwards the death rate rises within as well as between intervals, and this is partly taken into account by the alternative more precise expression

$$\frac{l_{x+5}}{l_x} = \frac{1 - 5\,{}_5M_x/2}{1 + 5\,{}_5M_x/2}.$$

Greville (1943) gives a more general expression. More generally yet, if $p(x+t)$ is the continuous age distribution within the interval x to $x+5$, and $\mu(x+t)$ the continuous death rate, then we have the equation

$${}_5M_x = \frac{\int_0^5 p(x+t)\mu(x+t)\,dt}{\int_0^5 p(x+t)\,dt}$$

from which it is required to extract the quantity

$$l_{x+5} / l_x = \exp\left[-\int_0^5 \mu(x+t)\mathrm{d}t \right].$$

A solution (Keyfitz, 1985, p. 39) is obtained by expanding the *p*'s and the μ's by Taylor's theorem. Having obtained the probability of surviving from one point of age to the next, the life table is completed by cumulating these probabilities from age 0; with an arbitrary starting point ('radix') of 1 or 100000 the l_x column is obtained by successive multiplication

$$l_{x+5} = l_x \left(\frac{l_{x+5}}{l_x} \right), \text{ etc.}$$

The l_x column has three interpretations: (1) The probability of a person just born surviving to age *x*. (2) The number of survivors in a hypothetic cohort (say starting with 100,000 births) by the time age *x* is reached. (3) The number of persons aged *x* in the stationary population.
For this last interpretation one integrates over one- or five-year age intervals, and so obtains ${}_5L_x = \int_0^5 l(x+t)\,\mathrm{d}t$, the number of individuals in a stationary population (say one in which there are exactly 100,000 births per year) at age *x* to *x*+4 at last birthday.
What makes possible the simultaneous representation of these three quite different entities is a central assumption of the life table: that the actual number of deaths to occur will be the probability multiplied by the initial number exposed. In short, the life table is a deterministic model: if there are a million people, each with a probability of 0.01 of dying during the following year, there will be exactly 10,000 deaths. It also assumes that every individual of a given age and sex has the same probability of dying.
The estimators above do not make explicit allowance for withdrawals, nor for the individual times at death. With small populations, for instance those used in follow-up studies after a diagnosis of cancer, or after a particular treatment, more refined methods are needed. One such, called the product-limit method and using maximum likelihood, is due to Kaplan and Meier (1958). This and ways of dealing with withdrawals and censoring are taken up in Elandt-Johnson and Johnson (1980, ch. 6).
From the probability of surviving the life expectancy is calculated as $e_x^0 = \int_0^{\omega - x} l(x+t)\,\mathrm{d}t$. In the deterministic model this traces a (usually synthetic) cohort consisting of l_x individuals, who will live ${}_5L_x$ person-years over the next five years; ${}_5L_x$+5 over the five years after that, etc. These future years may be thought of as divided among the l_x persons, giving each of them an average of $e_x^0 = \Sigma L_{x+h} / l_x$.

An original purpose of the life table was to calculate annuities and life insurance, and much of the modern notation has been developed by actuaries. If money carried no interest then the value of an annuity starting at age 65, say, would be $\int_{65}^{\omega} l(t)\,\mathrm{d}t$, and if this was to be paid for by yearly payments from age 20, the annual premium would be $\int_{65}^{\omega} l(t)\,\mathrm{d}t \,/\, \int_{20}^{65} l(t)\,\mathrm{d}t$. If money carries interest we need to discount this (say back to birth), and the premium will be less, being calculated as $\int_{65}^{\omega} \mathrm{e}^{-it} l(t)\,\mathrm{d}t \,/\, \int_{20}^{65} \mathrm{e}^{-it} l(t)\,\mathrm{d}t$, where *i* is the rate of interest compounded momently (Jordan, 1967).

The expectation of life at age 0 is a common measure of mortality, for comparing countries and other population aggregates: in the United States the life expectancy was 75 years in 1983, compared with 66 years for Mexico. Mexico's crude rate (1000×D/P) is 6 per thousand against the US 9, a comparison that does not reflect true mortality because Mexico's population is much younger.

The third meaning of the life table can be generalized to represent the age distribution of a population that is increasing at a steady rate *r*; in this generalization the number of persons aged *x* to *x*+4 at last birthday is proportional to $\int_{0}^{5} \mathrm{e}^{-(x+rt)} l(x+t)\,\mathrm{d}t$.

The life table idea is readily extended to more than two states of exit. One can work out the chance of dying from the several possible causes of death – cancer, heart disease, etc.; this is still a decrement table, but now with several causes of decrement.

While the notation and the concepts of the life tables were worked out for mortality, it is applied to many processes other than living and dying. A woman has a certain probability month by month of becoming pregnant; the probabilities can be cumulated to give the probability of still not being pregnant by the *x*th month, from which the expected months to pregnancy can be calculated for women who are fertile. An aircraft engine has a certain probability of breaking down in the first month, the second month, etc.; a life table shows the expected number of months of service, and by an extension the number of engines that will have to be kept in reserve for replacements up to a given level of security. Biological ecologists calculate life tables for many species of animals and insects. Probability of divorce in the first year, the second year, etc. after marriage can be worked out in the same two-state model, except that now first marriage and divorce rather than living and dead are the states in question. A table can be made for survival within the school system, in which the states are attending school and dropping out.

In increment–decrement tables persons can re-enter some of the states. For instance they can enter the labour force, then leave it, then enter again. The same applies to marriage, or to migration among regions of a country. For this a multi-dimensional analogue of the life table is available, and has been extensively used (Rogers, 1975, 1984; Schoen, 1975). The relevant formulas are matrix analogues of the ordinary life table formulas given above.

A main use of life tables is for population projection. If the population age *x* to *x*+4 at the jumping-off point is ${}_5P_x$, then 5 years later it will be ${}_5P_x {}_5L_{x+5}/{}_5L_x$ if the life table is appropriate and random variation and migration can be disregarded. (For the birth component and other aspects of projection, see Brass, 1974.)

In pursuing these and other purposes, one often deals with populations for which mortality data are deficient or altogether lacking. A common procedure in the past was to substitute a suitable member of a series (for example England and Wales at an appropriate date). Today it is more convenient to use one of the sets of model tables calculated for the purpose, based not on one country, but on all the countries for

which reliable data are available (UN, 1982; Coale and Demeny, 1983).

Life tables are calculated on the (unrealistic) assumption that the population is homogeneous in respect of all unmeasured variables. Because the observed population is constantly being selected towards persons of greater robustness, the true expectation for a person initially of average robustness is less (by something of the order of one year) than that shown by published tables (Vaupel and Yashin, 1985).

See Also

- economic demography
- fertility in developing countries
- Graunt, John
- historical demography
- mortality
- stable population theory

Bibliography

Brass, W. 1971. On the scale of mortality. In *Biological Aspects of Demography*, ed. W. Brass. London: Taylor & Francis, 69–110.

Brass, W. 1974. Perspectives in population prediction, illustrated by the statistics of England and Wales. *Journal of the Royal Statistical Society*, Series A 137, 532–83.

Chiang, C.L. 1984. *The Life Table and Its Applications*. Malabar, Florida: Robert E. Krieger.

Coale, A.J. 1984. Life table construction on the basis of two enumerations of a closed population. *Population Index* 50(2), Summer, 193–213.

Coale, A.J. and Demeny, P. 1983. *Regional Model Life Tables and Stable Populations*. 2nd edn. New York: Academic Press.

Elandt-Johnson, R.C. and Johnson, N.L. 1980. *Survival Models and Data Analysis*. New York: John Wiley & Sons.

Greville, T.N.E. 1943. Short method of constructing abridged life tables. *Record American Institute Actuaries* 32, 29–43.

Jordan, C.W. 1967. *Life Contingencies*. Chicago: Society of Actuaries.

Kaplan, E.L. and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of*

the American Statistical Association 53, 457–81.

Keyfitz, N. 1985. *Applied Mathematical Demography*. 2nd edn. New York: Springer-Verlag.

National Center for Health Statistics. 1983. *Advance Report of Final Mortality Statistics, 1980* 32(4), Supplement, US Department of Health and Human Services.

Rogers, A. 1975. *Introduction to Multiregional Mathematical Demography*. New York: John Wiley & Sons.

Rogers, A. 1984. *Migration, Urbanization, and Spatial Population Dynamics*. Boulder, Colorado: Westview Press.

Schoen, R. 1975. Constructing increment-decrement life tables. *Demography* 12, 313–24.

United Nations. 1982. *Model Life Tables for Developing Countries*. New York: United Nations.

Vaupel, J.W. and Yashin, A.I. 1985. The deviant dynamics of death in heterogeneous populations. In *Sociological Methodology 1985*, ed. N.B. Tuma. San Francisco: Jossey-Bass.

How to cite this article

Keyfitz, Nathan. "life tables." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000093> doi:10.1057/9780230226203.0970

The New Palgrave Dictionary of Economics Online

limit pricing

Stephen Martin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The central idea of limit pricing is that an incumbent monopolist or collusive group will or can forestall entry by charging some price below that which maximizes static own profit. But a strategic price response is only one possible incumbent response to entry. A full understanding of the determinants of equilibrium market structure in inherently oligopolistic industries must take the full range of possible responses into account.

Keywords

advertising; commitment; entry; excess capacity; limit pricing; potential competition; predatory pricing

Article

Modern economists generally trace models of limit pricing to Modigliani (1958). The idea of limit pricing is closely related to, and often not distinguished from, the much older idea that potential competition will induce a profit-maximizing incumbent monopolist or dominant group to set a price that would allow it (or, in some formulations, an entrant) only a normal rate of return (Giddings, 1887; Gunton, 1888; Liefmann, 1915; Marshall, 1890, p. 270; 1919, pp. 397–8, 524; Kaldor, 1935). With this second idea, it is the presence of potential entrants that constrains the options of incumbents, not the other way around.

Modigliani's (1958) more-than-a-book-review of Bain (1956) and Sylos-Labini (1957) offered a formal model based on what Modigliani called the Sylos postulate (1958, p. 217) ‘that potential entrants behave as though they expected existing firms to adopt the policy … of maintaining output’ in the face of entry. Given such beliefs, if incumbents produce a sufficiently large output that the best post-entry price a profit-maximizing entrant could expect would be below its average cost, entry would not occur.

Gaskins (1971) generalizes the static limit price model to a dynamic context, with a model in which incumbent pricing determines the rate of expansion of a fringe of price-taking suppliers. This might also

be regarded as a dynamic generalization of the familiar Forchheimer–Auspitz–Lieben model of a dominant firm in a market with a price-taking fringe.

Friedman (1979) points out that, under conditions of complete and perfect information, profit-maximizing incumbents would not, in general, maintain post-entry output at pre-entry levels, and entrants would not expect them to do so. Much the same point had been made, less formally, by Bain (1949, p. 452). Without commitment, a low price fails as an entry-limiting device if entrants believe that in the post-entry market incumbents will act in their own self-interest.

One line of research that seeks to finesse the unsatisfactory nature of the Modigliani–Sylos postulate can be traced to Spence (1977) and Dixit (1979). They offer models in which an incumbent's pre-entry investment (in capacity) alters the incumbent's post-entry incentives, and by so doing gives credibility to post-entry conduct that renders entry unprofitable. See Allen et al. (2000) for careful discussion. The vast literature on strategic entry deterrence (Salop and Scheffman, 1983; Fudenberg and Tirole, 1984) springs from this root.

An alternative approach is taken by Kreps and Wilson (1982) and Milgrom and Roberts (1982). They give up the assumption of complete information and model entry-limiting behaviour based on an incumbent firm's reputation or entrants’ uncertainty about an incumbent's costs. The modelling techniques employed here have been generalized to analyse predation and the conduct of regulation/competition policy under conditions of uncertainty.

The development of internally consistent theoretical models in which entry-limiting behaviour might occur as an equilibrium phenomenon was a major step in laying the game-theoretic foundation for modern industrial economics. The assembling of empirical evidence on the occurrence of limit pricing and other strategic reactions to entry has similarly followed the general trend of empirical research in industrial economics, studying particular markets for specific instances of entry-deterring behaviour. There are case studies of limit-pricing behaviour (Blackstone, 1972). But empirical studies of entry suggest that theoretical models of entry and entry deterrence abstract from essential aspects of the phenomena (Simon, 2005, p. 1230).

Some real-world entry, no doubt, is like the entry of limit price and other entry-deterrence models – entry at large-scale into production of a standardized product hitherto offered by a small number of firms themselves aware of their oligopolistic interdependence. Archer Daniel Midland's well-known 1991 entry into lysine production is a case in point. Much more often, however, entry seems like the act of beginning small-scale production at a point on a Hotelling line, when location in characteristic space is largely fixed after entry and neither the entrant nor incumbents have a terribly good idea of the distribution of consumers in the region near the entrant's location.

Geroski (1995, p. 433) concludes in his careful survey that ‘price is not frequently used by incumbents to deter entry, but that marketing activities are’, and (1995, p. 434, fn. 7) ‘work that has tried to test for the presence of limit pricing … in general … has produced somewhat ambiguous results.… Studies of the strategic use of excess capacity to block entry have also generally produced weak and fairly unpersuasive evidence on its importance.’

Empirical work suggests that the incumbent response to entry will vary with entrant and incumbent characteristics. The response to entry will sometimes be by lowering price, sometimes by other rival strategies, and sometimes by accommodating entry.

Scott Morton (1997) finds that longer-established entrants into turn-of-the-19th-century shipping cartels

were less likely to evoke a hostile response, as were entrants with substantial financial resources. Podolny and Scott Morton (1999) suggest that a predatory response was less likely if social factors were present that would allow incumbents and entrants to judge each others' 'types'. Thomas (1999) finds that incumbent US breakfast cereal manufacturers are more likely to respond with advertising to entry into a product group by other incumbents, and more likely to lower price in response to entry by a new firm. Yamawaki (2002) finds a price response to Japanese entry by German manufacturers of luxury cars for the US market, but no such response by British manufacturers. In a study of entry into the US magazine industry, Simon (2005) finds that multi-market and single-market incumbents respond differently to entry. Multi-market incumbents are more likely to cut price in response to entry by a new firm, single-market incumbents more likely to cut price in response to entry by an established publisher. He also finds that a hostile response to entry is more likely the more concentrated the target market. Conlin and Kadiyali (2006) find some evidence of the use of excess capacity as an entry-deterring device in the Texas hotel market, and also that the maintenance of excess capacity is more likely by larger firms and by firms in more concentrated markets.

It thus appears that, while a strategic price response is one possible incumbent response to entry, it is only one. A full understanding of the determinants of equilibrium market structure in inherently oligopolistic industries must take the full range of possible responses into account.

See Also

- barriers to entry
- market structure
- monopoly
- oligopoly
- predatory pricing

Bibliography

Allen, B., Deneckere, R., Faith, T. and Kovenock, D. 2000. Capacity precommitment as a barrier to entry: a Bertrand–Edgeworth approach. *Economic Theory* 15, 501–30.

Auspitz, R. and Lieben, R. 1889. *Untersuchen über die Theorie des Preises*. Leipzig: Duncker & Humblot.

Bain, J.S. 1949. A note on pricing in monopoly and oligopoly. *American Economic Review* 39, 448–64.

Bain, J.S. 1954. Conditions of entry and the emergence of monopoly. In *Monopoly and Competition and their Regulation*, ed. E.H. Chamberlin. London: Macmillan.

Bain, J.S. 1956. *Barriers to New Competition*. Cambridge, MA: Harvard University Press.

Blackstone, E.A. 1972. Limit pricing and entry in the copying machine market. *Quarterly Review of Economics and Business* 12, 57–65.

Conlin, M. and Kadiyali, V. 2006. Entry-deterring capacity in the Texas lodging industry. *Journal of Economics and Management Strategy* 15, 167–85.

Dixit, A. 1979. A model of duopoly suggesting a theory of entry barriers. *Bell Journal of Economics* 10, 20–32.

Forchheimer, K. 1908. Theoretisches zum unvollstandigen Monopole. *Schmoller's Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft*, vol. 32. Munich, Leipzig: Duncker & Humblot.

Friedman, J.W. 1979. On entry preventing behavior and limit price models of entry. In *Applied Game Theory*, ed. S.J. Brams and G. Schwodiauer. Wurzburg and Vienna: Physica-Verlag.

Fudenberg, D. and Tirole, J. 1984. The fat-cat effect, the puppy-dog ploy, and the lean and hungry look. *American Economic Review* 74, 361–6.

Gaskins, Jr., D.W. 1971. Dynamic limit pricing: optimal limit pricing under threat of entry. *Journal of Economic Theory* 3, 306–22.

Geroski, P.A. 1995. What do we know about entry? *International Journal of Industrial Organization* 13, 421–40.

Giddings, F.H. 1887. The persistence of competition. *Political Science Quarterly* 2, 62–78.

Gunton, G. 1888. The economic and social aspect of trusts. *Political Science Quarterly* 3, 385–408.

Kaldor, N. 1935. Market imperfection and excess capacity. *Economica* NS 2, 33–50.

Kreps, D.M. and Wilson, R. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27, 253–79.

Liefmann, R.L. 1915. Monopoly or competition as the basis of a government trust policy. *Quarterly Journal of Economics* 29, 308–25.

Marshall, A. 1890. Some aspects of competition. Presidential Address to the Economic Science and Statistics Section of the British Association, Leeds. In *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan, 1925.

Marshall, A. 1919. *Industry and Trade*, 4th edn. London: Macmillan, 1923.

Milgrom, P. and Roberts, J. 1982. Limit pricing and entry under incomplete information: an equilibrium analysis. *Econometrica* 50, 443–66.

Modigliani, F. 1958. New developments on the oligopoly front. *Journal of Political Economy* 66, 215–32.

Podolny, J.M. and Scott Morton, F.M. 1999. Social status, entry and predation: the case of British shipping cartels 1879–1929. *Journal of Industrial Economics* 47, 41–67.

Salop, S.C. and Scheffman, D.T. 1983. Raising rivals' costs. *American Economic Review* 73, 267–71.

Scott Morton, F. 1997. Entry and predation: British shipping cartels 1879–1929. *Journal of Economics and Management Strategy* 6, 679–724.

Simon, D. 2005. Incumbent pricing responses to entry. *Strategic Management Journal* 26, 1229–48.

Spence, A.M. 1977. Entry, capacity, investment oligopolistic pricing. *Bell Journal of Economics* 8, 534–44.

Sylos-Labini, P. 1957. *Oligopolio e Progresso Tecnico*. Milano: Giuffrè.

Sylos-Labini, P. 1962. *Oligopoly and Technical Progress*. Cambridge, MA: Harvard University Press.

Thomas, L.A. 1999. Incumbent firms' response to entry: price, advertising, and new product introduction. *International Journal of Industrial Organization* 17, 527–55.

Yamawaki, H. 2002. Price reactions to new competition: a study of US luxury car market, 1986–1997. *International Journal of Industrial Organization* 20, 19–39.

How to cite this article

Martin, Stephen. "limit pricing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000097> doi:10.1057/9780230226203.0971

The New Palgrave Dictionary of Economics Online

Lindahl equilibrium

John Roberts

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Lindahl equilibrium embodies a market solution to the problem of providing public goods. Each individual faces personalized prices at which he or she may buy total amounts of the public goods. In equilibrium, these prices are such that everyone demands the same levels of the public goods and thus agrees on the amounts that should be provided. Since individuals buy the total production of public goods, the price to producers is the sum of the prices paid by individuals, and equilibrium involves the supply at these prices equalling the common demand, with costs being shared in proportion to (marginal) benefits.

Keywords

bargaining; efficient allocation; externalities; incentive compatibility; joint production; Lindahl equilibrium; Lindahl, E. R.; misrepresentation of preferences; missing markets; Nash equilibrium; optimality; property rights; public goods; pure public goods; revealed preferences; tax incidence; Walras equilibrium

Article

Lindahl equilibrium attempts to solve the problem of determining the levels of public goods to be provided and their financing by adapting the price system in a way that maintains its central feature of an efficient allocation being the outcome of voluntary market activities within the context of private property rights. Instead of some political choice mechanism and coercive taxation, under the Lindahl approach each individual faces personalized prices at which he or she may buy total amounts of the public goods. In equilibrium, these prices are such that everyone demands the same levels of the public goods and thus agrees on the amounts of public goods that should be provided. Since each individual buys and consumes the total production of public goods, the price to producers is the sum of the prices paid by individuals, and equilibrium involves the supply at these prices equalling the common demand.

Thus, Lindahl equilibrium brings unanimity about the level of public goods provision, with costs being shared in proportion to (marginal) benefits.

The basic idea of a market solution to the problem of providing public goods is due to Erik Lindahl (1919). In its modern formulation, Lindahl equilibrium has come to play a benchmark role in the study of economies with public goods, externalities, and government expenditure which parallels that played by Walrasian competitive equilibrium in the analysis of questions where these factors are absent. For example, tax incidence can be measured relative to the Lindahl equilibrium. On the other hand, the Lindahl concept does not share the competitive equilibrium's centrality of position as a predictor of the actual outcomes of economic activity.

This latter point involves some irony, because Lindahl's original exposition of the idea treats it as having both normative and descriptive/predictive value.

Lindahl considered a legislature in which two parties represent the two homogeneous classes that constitute the electorate. (He also indicates how to extend the analysis to more classes and their representatives.) The issue is how much government activity should be carried out and how the costs of this activity should be shared between the two groups.

Lindahl identified two functions, say $f_A(s)$ and $f_B(s)$, which give, respectively, the expenditure on public activity that group A would want if it had to pay a fraction s of the corresponding costs and the level that B would want if it had to pay the complementary fraction $1 - s$. The value $x = f_A(s)$ is just the solution to the problem of maximizing the utility of after-tax income and public expenditure for group A, given that it will pay 100s% of the costs, while f_B solves the corresponding problem for B. Ignoring income effects, Lindahl obtained $s = v_A'(f_A(s))$, where v_A is A's utility for public expenditure, and, correspondingly, $1 - s = v_B'(f_B(s))$. Note that f_A is decreasing and f_B is increasing. Thus, assuming $f_A(0) > f_B(0)$ or $f_A(1) < f_B(1)$, so that a group bearing all the costs wants less expenditure than does the group paying nothing, there is a unique value s^* strictly between zero and one at which the two groups agree on the desired level of expenditure, that is, $x^* = f_A(s^*) = f_B(s^*)$.

Much of Lindahl's analysis is in terms of bargaining between the two groups over x and s under the assumption that, at any partition of the costs, the smaller of the two proposed quantities will be implemented. (This reflects the connection to voluntary exchange, where no one is forced to transact.) He recognized that such bargaining would not automatically lead to s^*, x^*. However, he claimed that if both groups were equally adept at defending their interests, this outcome would result.

Foley (1970) provided the basic general equilibrium treatment of Lindahl's idea in the context of an Arrow–Debreu private ownership economy with both private and pure public goods (no rivalry in consumption and no possibility of exclusion) where there are zero endowments of public goods, these goods are never used as inputs, and production takes place under constant returns to scale. See Milleron (1972), Roberts (1973), and Kaneko (1977) for extensions and Roberts (1974) for a survey.

Foley's model focuses on prices for the public goods rather than cost shares. Individual demand functions for public goods, as depending on the prices of both private and public goods, are defined (exactly as for private goods) as the choices of quantities to consume that maximize utility subject to the budget constraint defined by the prices and the agent's endowment. Thus, the quantity demanded of any public good at a particular price vector differs with individual preferences and endowments. However,

the nature of pure public goods requires that all agents' consumption of any of these goods be equal. Thus, if prices are to lead different individuals all to demand the same quantities of public goods, it is clear that the prices charged to consumers must be personalized, differing across individuals to reflect differences in preferences and incomes. The price received by a producer of public goods is then the sum of the price paid by individuals, because each unit of each public good is allocated to and paid for by every individual. Meanwhile, private goods markets involve standard competitive pricing. With this, Lindahl equilibrium is a vector p of private goods prices, a vector q_i of public goods prices for each consumer i, an allocation of private goods x_i to each i and a vector of public goods y such that: (x_i, y) is the most preferred consumption bundle for consumer i from those affordable at prices (p, q_i), given i's wealth as determined by p and i's initial endowment of private goods ω_i; and also such that the net input–output vector $(\Sigma_i x_i - \omega_i, y)$ is profit maximizing at the producer prices $(p, \Sigma_i q_i)$. Note that both consumers and producers are following standard, competitive, price-taking behaviour just as in the Walrasian equilibrium.

Further appreciation of the connection between Lindahl and Walrasian equilibria can be gained using Arrow's insight (1970) that externalities (and the public goods problem in particular) can be viewed as a phenomenon of missing markets. Given a public goods economy with I consumers, M private goods and N public goods such as studied by Foley, consider an associated economy with I consumers, $(M + K)$ private goods, and no public goods, where $K = IN$. In this economy, each public good n in the original economy is replaced by a collection of I private goods, each of which is of interest to and consumable by only one consumer and which together are joint products in production. A net input–output vector in this economy of the form

$$(z, \tilde{y}), \quad z \in R^M, \tilde{y} = (y^1, \ldots, y^{IN}) = (y_1, y_2, \ldots, \ldots y_N, y_1, y_2, \ldots, y_N, \ldots, y_1, y_2, \ldots, y_N) \in R_+^{IN}$$

is producible if and only if $(z, y_1, \ldots, y_N)$ is in the production set of the original public goods economy. A Walras equilibrium in this economy is a price vector $(p, q^1, \ldots, q^{IN}) \in R_+^{MK}$ and consumption vectors $(x_i, y_i^1, \ldots, y_i^{IN}) \in R^{M+K}, i = 1, \ldots, I$, where $(x_i, y_i^1, \ldots, y_i^{IN})$ is the most preferred bundle for i from among those costing no more than $p\omega_i$ and where $(\Sigma_i x_i - w_i, \Sigma_i y_i^1, \ldots, \Sigma_i y_i^{IN})$ is profit maximizing at prices $(p, q^1, \ldots, q^{IN})$. Clearly, these conditions imply $y_i^{jn} = 0$, for $i \neq j$ so that no consumer receives positive amounts of another's personalized goods, and $y_i^{in} = y_j^{in}$ for all i, j, and n, so that each individual consumes the same quantities of these personalized goods. Thus, Walras equilibria of the artificial economy exactly correspond to the Lindahl equilibria of the original economy, with a parallel correspondence between the feasible allocations in the two economies and between the Pareto optima.

This construction, which was used by Foley to prove existence of Lindahl equilibrium, illuminates the claim that the Lindahl equilibrium involves voluntary exchange in the context of maintaining private

property rights. It also makes clear that Lindahl equilibria are Pareto optimal and that any optimum can be supported as an equilibrium with a reallocation of resources. (In fact, Silvestre, 1984, has characterized Lindahl allocations in terms of optimality plus a condition that no agent wants to reduce his or her contribution to paying for public goods if the level of provision would be proportionately reduced.) The Lindahl equilibrium's role as a benchmark is largely attributable to its having these properties, plus the fact that the Lindahl equilibrium allocations belong to the core if blocking is defined by a group being able to produce a more preferred consumption bundle for each of its members, even if non-members contribute nothing to public goods production (Foley, 1970). However, this construction also suggests some of the problems with the Lindahl equilibrium which prevent it from having great appeal as a positive prediction.

In particular, the usual complaint against a price-based solution to the public goods problem is that there would be no reason for an individual to take the Lindahl prices as given: misrepresentation of preferences should be profitable. Of course, as long as there are only a finite number of participants in a market, the behaviour of each typically has some influence on price formation, and so the assumption of price-taking in Walrasian, private goods equilibrium is questionable too.

Progress on this incentives question requires being more specific about the mechanism used to determine the allocation as a function of the initially dispersed information about the economic environment. In this context, Hurwicz (1972) formalized the idea that there must be incentive problems even with only private goods by showing that if a mechanism always yields Pareto optima and, if participation is voluntary, so that its outcomes must be unanimously preferred to the no-trade point, then it cannot be a dominant strategy always to report one's preferences (demand) correctly. The exactly parallel result for public goods was achieved by Ledyard and Roberts (see Roberts, 1976). Thus neither Walrasian nor Lindahl equilibria can be the outcome of a mechanism which is incentive compatible in this dominant-strategy sense.

Of course, the standard case in which the Walrasian equilibrium seems appealing is a ‘large numbers’ one where each individual's influence is small. This intuition has been formalized in a number of ways: revealing one's true demand for private goods generically is asymptotically a dominant strategy as the number of participants in the economy becomes large; only competitive allocations are in the core of large economies; Nash equilibria of various models in which individuals recognize their influence on prices converge to the competitive solution as the economy grows. However, with public goods the situation is much different: increasing the size of the economy makes price-taking less attractive. This too has been shown in various ways. Roberts (1976) showed that increasing numbers can worsen the incentives for correct revelation of preferences for public goods and that as the numbers grow, the departure of the outcome from efficiency can also increase. Muench (1972) showed that the core and Lindahl equilibria do not coincide in large economies, and Champsaur, Roberts and Rosenthal (1975) demonstrated that the core of a public goods economy may actually expand when the number of consumers increases. In terms of the artificial economy, the essential intuition is that the market for each of the personalized goods is monopsonized, and the joint-product interaction constrains the bargaining power of the producer which otherwise might permit an efficient outcome to the bilateral monopoly situation. Thus, it seems that in the large numbers situations that have been the traditional concern of economics, the price-taking assumption renders the Lindahl solution of little predictive or descriptive value.

These essentially negative results are in some contrast with the results on incentives for correct revelation in iterative planning procedures for determining public goods. This literature was begun by Malinvaud (1971a; 1971b) and Drèze and de la Vallée Poussin (1971) and is surveyed in Roberts (1986). In this context, the notion of incentive-compatible behaviour is Nash equilibrium: each agent selects his/her responses to the central planning authority's proposals so as to maximize his/her payoff, given the strategies being used by the other agents to determine their responses. Such behaviour typically involves misrepresentation of preferences. However, various authors (Roberts, 1979, Champsaur and Laroque, 1982, and Truchon, 1984, for example), have shown that this misrepresentation need not prevent convergence to a Pareto optimum and, in particular, to the Lindahl allocation.
However, as argued in Roberts (1986), these results are of limited interest because they rely on the implausible assumption that each agent is perfectly informed about the other's preferences. (A similar criticism can be laid against the static mechanisms for obtaining Walrasian or Lindahl allocations as Nash equilibria; Hurwicz, 1979.) Moreover, once the (self-selection or truthful reporting) constraints associated with preferences being private information are recognized, it is not clear that any mechanism can achieve Lindahl allocations (see Laffont and Maskin, 1979; d'Aspremont and Gerard-Varet, 1979). This gives a further reason for doubting the empirical relevance of Lindahl equilibrium.

See Also

- duality
- incentive compatibility
- public goods

Bibliography

Arrow, K.J. 1970. The organization of economic activity: issues pertinent to the choice of market versus non-market allocation. In *Public Expenditures and Policy Analysis*, ed. R.H. Haveman and J. Margolis. Chicago: Markham.

Champsaur, P. and Laroque, G. 1982. Strategic behavior in decentralized planning procedures. *Econometrica* 50, 325–44.

Champsaur, P., Roberts, J. and Rosenthal, R. 1975. Cores in economies with public goods. *International Economic Review* 16, 751–64.

d'Aspremont, C. and Gerard-Varet, L.A. 1979. On Bayesian incentive compatible mechanisms. In *Aggregation and Revelation of Preferences*, ed. J. Laffont. Amsterdam: North-Holland.

Drèze, J. and de la Vallée Poussin, D. 1971. A tâtonnement process for public goods. *Review of Economic Studies* 38, 133–50.

Foley, D. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38, 66–72.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and Organization: A Volume in Honor of Jacob Marschak*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Hurwicz, L. 1979. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *Review of Economic Studies* 46, 217–27.

Kaneko, M. 1977. The ratio equilibrium and a voting game in a public goods economy. *Journal of Economic Theory* 16, 123–36.

Laffont, J. and Maskin, E. 1979. A differential approach to expected utility maximizing mechanisms. In *Aggregation and Revelation of Preferences*, ed. J. Laffont. Amsterdam: North-Holland.

Lindahl, E. 1919. *Die Gerechtigkeit der Besteuering*. Lund: Gleerup. Part I, ch. 4, ‘Positive Lösung’, trans. E. Henderson and reprinted as ‘Just taxation – a positive solution’. In *Classics in the Theory of Public Finance*, ed. R.A. Musgrave and A.T. Peacock, London: Macmillan, 1958.

Malinvaud, E. 1971a. A planning approach to the public goods problem. *Swedish Journal of Economics* 11, 96–112.

Malinvaud, E. 1971b. Procedures for the determination of a program of collective consumption. *European Economic Review* 2, 187–217.

Milleron, J. 1972. Theory of value with public goods: a survey article. *Journal of Economic Theory* 5, 419–77.

Muench, T. 1972. The core and the Lindahl equilibrium of an economy with a public good: an example. *Journal of Economic Theory* 4, 241–55.

Roberts, J. 1973. Existence of Lindahl equilibrium with a measure space of consumers. *Journal of Economic Theory* 6, 355–81.

Roberts 1974. The Lindahl solution for economies with public goods. *Journal of Public Economics* 3, 23–42.

Roberts, J. 1976. The incentives for correct revelation of preferences and the number of consumers. *Journal of Public Economics* 6, 359–74.

Roberts, J. 1979. Incentives in planning procedures for the provision of public goods. *Review of*

Economic Studies 46, 283–92.

Roberts, J. 1986. Incentives, information and iterative planning. In *Information, Incentives, and Economic Mechanisms*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Silvestre, J. 1984. Voluntariness and efficiency in the provision of public goods. *Journal of Public Economics* 24, 249–56.

Truchon, M. 1984. Nonmyopic strategic behavior in the MDP planning procedure. *Econometrica* 52, 1179–89.

How to cite this article

Roberts, John. "Lindahl equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000102> doi:10.1057/9780230226203.0972

The New Palgrave Dictionary of Economics Online

Lindahl, Erik Robert (1891 – 1960)

Otto Steiger

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Erik Lindahl's writings between 1919 and 1959 covered four major areas. In public finance his pioneering contribution is today known as the ‘Wicksell–Lindahl paradigm of just taxation’. In dynamic analysis he was first to develop the methods of intertemporal equilibrium and temporary equilibrium. In macroeconomics Lindahl anticipated many of the insights of Keynes's *General Theory*, and in his discussion of the concepts of income and capital he laid the foundations of the theory of national accounting. His contributions in the four fields have been acknowledged internationally step by step only since the 1950s – in the third area since the 1970s.

Keywords

accelerationist hypothesis; aggregate demand and supply; benefit rule; budget balance; capital theory; central banking; Davidson, D.; Debreu, G.; Domar, E. D.; European Central Bank; European Monetary Union; Eurosystem; *ex ante* and *ex post* ; expectations; financial equilibrium; Friedman, M.; Frisch, R. A. K.; general equilibrium; Hansen, B.; Hayek, F. A.; Hicks, J. R.; imperfect foresight; income; International Economic Association; intertemporal equilibrium; Kaldor, N.; Keynes, J. M.; Leijonhufvud, A.; Lindahl, E. R.; Lundberg, E. F.; monetary equilibrium; monetarist(s); Myrdal, G.; Musgrave, R. A.; national accounting; normal rate of interest; Neumann, J. von; Ohlin, B. G.; Peacock, A. T.; Phelps, E. S.; paradox of saving; perfect foresight; public debts; public expenditure; public finance; public works; quantity theory of money; Riksbank; Samuelson, P.A.; sequence analysis; Stockholm School; stock and flow; Swedish Unemployment Committee; taxation; temporary equilibrium; time; total capital; unemployment equilibrium; Wicksell–Lindahl paradigm of just taxation; Wicksell, J. G. K.

Article

Lindahl was born on 21 November 1891 in Stockholm and died on 6 January 1960 in Uppsala, Sweden. He is now reckoned one of the great economists who were at work between the two world wars, and earned his reputation above all as a leading member within a group of Swedish economists during the 1930s consisting, besides himself, of Gunnar Myrdal, Bertil Ohlin, Dag Hammarskjöld, Alf Johansson, Erik Lundberg and Ingvar Svennilsson – a body which Ohlin (1937) had baptised the ‘Stockholm School’.

The son of a prison governor, Lindahl grew up in Jönköping, the capital of a province in southern Sweden. After passing the studentexamen at a Stockholm Secondary School in spring 1910, he enrolled the following autumn as a student at the University of Lund, where economics soon became the favourite subject in his studies of humanities and law, which he passed with the degrees of the fil*osofie kandidatexamen* (BA) in 1912 and the *juris kandidatexamen* (LLB) in 1914. Although Knut Wicksell was professor of economics and fiscal law in Lund at that time (1901–16), Lindahl did not have any personal contact with him during this period. However, Emil Sommarin, the successor to Wicksell's chair (1916–39) and at the time of Lindahl's student years *docent* (reader) in economics and a great admirer of Wicksell, succeeded in encouraging Lindahl to study the former's works to such an extent that the latter became in effect the first pupil of Wicksell. As Lindahl's dissertation of 1919, *Die Gerechtigkeit der Besteuerung*, was largely based on Wicksell's theory of public finance (1896), Sommarin let Wicksell read and comment on it, and, at the public defence of the thesis at Lund University on 13 December 1919, Wicksell officiated as the official ‘challenger’ appointed by the faculty of law (Lindahl, 1951, pp. 26–7).

With his doctoral thesis Lindahl had earned the title *docent* in public finance at Lund University (1920) and later also in economics and fiscal law at Uppsala University (1924), but not yet the position of a professor. In 1926 he became responsible for the planning of the voluminous investigations on *Wages, Cost of Living and National Income in Sweden 1860–1930* (see Lindahl, 1937a; and Benny Carlson, 1982, pp. 11–20) carried on in the following decade at the Institute for Social Sciences in Stockholm University and financed by the Rockefeller Foundation. In his attempts to obtain a chair in economics Lindahl failed twice: in 1924 he lost the competition for a professorship at the University of Copenhagen to Bertil Ohlin, later his colleague in the Stockholm School, and in 1930 he was ranked as number two only for a chair in political economy and sociology at Gothenburg University, this time defeated by Gustaf Åkerman, like Lindahl an early pupil of Wicksell.

Only two years later, however, in 1932, Lindahl obtained the chair in political economy at the Gothenburg School of Business Economics without application, and from this time onwards Swedish universities competed to call him to their departments of economics. In 1939, the year of publication of his most famous work, *Studies in the Theory of Money and Capital*, he succeeded Sommarin at Lund University, and in 1942 he became professor at the University of Uppsala, where he retired in 1958. Internationally, Lindahl's outstanding position as economist was honoured by his election as President of the International Economic Association in 1956.

Lindahl's growing reputation from the early 1930s onwards also led to numerous calls for economic expertise by Sweden's governments and official institutions. When Sweden left the gold standard in 1931 he became an adviser to Riksbanken, the Swedish central bank. When as a result of the Great Depression the final report of the Swedish Unemployment Committee had to be given a theoretical foundation of its proposal for public works as remedy against unemployment (see Hammarskjöld, 1935, p. ix and ch. 1; Otto Steiger, 1971, p. 40; and Bent Hansen, 1981, pp. 266–7) and when, therefore, the

character of Sweden's budget system had to be superseded in 1937 by a system deliberately designed to operate in a countercyclical manner (see Lindahl, 1935; cf. 1939a, app.), his expertise was sought by the Minister of Finance. Lindahl also became an economic adviser to the League of Nations (1936–9) and on two occasions to the United Nations (1949–50 and 1952–4).

Lindahl's work can be said to cover four major areas: (*a*) public finance; (*b*) methods of dynamic analysis; (*c*) monetary and macroeconomic theory; and (*d*) concepts of income and capital. Although Lindahl did not neglect empirical research, especially in public finance and national accounting, his contributions concentrated mainly on pure economic theory (see the detailed bibliography by Gertrud Lindahl and Olof Wallmén, 1960).

Public finance

Lindahl started his scientific career with a treatise on ‘just taxation’, his doctoral dissertation of 1919, which built on Wicksell (1896) and which, together with two re-examinations in 1928 and 1959, made it a pioneering contribution to the economic theory of the public household, today known as the ‘Wicksell–Lindahl paradigm of just taxation’ (Heinz Grossekettler, 2006, p. 557; for more detail see Peter Bohm, 1987, and John Roberts, 1987). It can be characterized as a culmination of the neoclassical reformulation of the classical version of the benefit approach to the simultaneous determination of public revenue and expenditure – a reformulation which applied a new interpretation of the benefit rule as a condition of equilibrium instead of as a standard of justice as in the classical version.

Lindahl formulated this condition in a partial equilibrium framework, where ‘financial equilibrium’, that is, the equilibrium of *public* finance, is determined by equalization of the ratio of prices paid by each taxpayer for public and for private goods to his marginal benefits derived from public and from private goods, the equilibrating financial process brought about by the political mechanism in a parliamentary democracy (cf. Roberts, 1987). Lindahl was convinced that his model could explain voting behaviour and the influence of pressure groups on decisions of the government concerning public expenditures and taxes (Bohm, 1987, p. 201). However, this ‘voluntary exchange approach’ (Richard A. Musgrave, 1959, pp. 73–8) for a long time failed to meet with much understanding. The importance of Lindahl's path-breaking contribution was first acknowledged in the 1950s via the works on the pure theory of public expenditure of Paul A. Samuelson (1954) and Musgrave (1959) as well as by the English translation of important parts of his dissertation of 1919 in 1958 in the volume *Classics in Public Finance*, edited by Musgrave and Alan T. Peacock. In the 1970s and the 1980s, however, Lindahl's model came under attack. It was criticized for relying on the ‘implausible assumption that each agent is perfectly informed about the other's preferences’ (Roberts, 1987, p. 200), and also for lacking empirical relevance in face of today's, unlike in Lindahl's time, ‘considerable amount of taxes raised for income distribution purposes’ (Bohm, 1987, p. 201).

Methods of dynamic analysis

Lindahl's contributions to dynamic method were formulated as part of the theoretical core of his macroeconomic ideas, culminating in 1939 in his *Studies in the Theory of Money and Capital*. As has

been shown by Björn Hansson (1982; cf. 1987; 1991, pp. 168–202; and Jan Petersson 1987), Lindahl's dynamic theory was developed by mutual influence within the Stockholm School, with himself and Myrdal as the key figures and mainly independent not only from influences from other contemporary economists but also – contrary to William P. Yohe (1959) – from Wicksell.

Already in his first macroeconomic treatise, the first edition of *Penningpolitikens mål* ([The aims of monetary policy], 1924, ch. 3), Lindahl stressed the *time* factor as a problem for economic analysis and used the notion of ‘subjective calculations of the future’ and also the term *ex post* (p. 33). A first coherent dynamic method was formulated in his treatise on capital theory (1929a; cf. 1939a, pt. III), where Lindahl developed the famous notion of *intertemporal equilibrium*, that is, the analysis of the sequential character of an economy by a sequence of periods with equilibrium in each period as a consequence of the assumption of perfect foresight. This approach has been praised by Gérard Debreu (1959, p. 35) as being ‘the first mathematical study of an economy whose activity extends over a finite number of elementary time-intervals’. However, as has been pointed out later (Murray Milgate, 1982, pp. 133–5), Friedrich A. Hayek had been moving on similar lines one year earlier. But this does not disturb the claim of Lindahl's originality, because a comparison of the 1929 and 1930 editions of his *Penningpolitikens medel* [The means of monetary policy] clearly shows that Lindahl became aware of Hayek's approach first after having worked out his own concept – Hayek's 1928 paper is referred to only in the second (1930, p. 11), not in the first edition (1929c, p. 10).

As has been shown by Hansson (1982, ch. 4, pp 59–67; 1987, pp. 504–5), Lindahl's formulation of intertemporal equilibrium, however, does not really represent a sequential process, since all prices and quantities are determined simultaneously at the beginning of the process for all periods. Lindahl became aware of this weakness when, under the influence of Myrdal's explicit introduction of expectations in equilibrium theory (1927, ch. 1), in the last section of his treatise on capital theory (1929a, pp. 80–1; cf. 1939a, pt. III, pp. 348–50) he substituted imperfect for perfect foresight. In *Penningpolitikens medel* (1930, pp. 18–24, 31–2; cf. 1939a, pt. II, pp. 158–9) Lindahl abandoned therefore, for the case of imperfect foresight, the method of intertemporal equilibrium for the notion of *temporary equilibrium*, that is, the analysis of the sequence of an economy as a series of very short periods of temporary equilibria with changes allowed only at the transition points of the periods. This notion looks closely akin to John R. Hicks's dynamic analysis in *Value and Capital* (1939, ch. 9) which in fact had been influenced decisively by Lindahl via personal contacts in 1934 and 1935, as later acknowledged by Hicks (cf. 1973, p. 8; 1985, pp. 66, 69; 1991, pp. 372–6; and Claes-Henric Siven, 2002, pp. 142–5).

More important in a historical perspective, however, is the striking fact that general equilibrium theorists, from the late 1960s onwards, began to give up their mathematically more elaborated intertemporal equilibrium models of Arrow–Debreu type and to develop different notions of temporary equilibrium for very much the same reason as Lindahl in 1929 and 1930: recognition of the fact that intertemporal equilibrium does not reflect the sequential character of an economy in an essential way and the impossibility of handling imperfect foresight, that is, problems involving uncertainty and money.

However, under the influence of the criticism of his approach by Lundberg in 1930 and Myrdal in 1932 and 1933, Lindahl realized that even with the notion of temporary equilibrium there was no real causation between the periods when he applied this dynamic method to the analysis of the saving–investment mechanism during a Wicksellian cumulative process, since the equilibrium approach in the construction of temporary equilibrium cannot handle unforeseen events *during* a period. Therefore, he

abandoned this notion and formulated instead the method of *sequence analysis*. This was done in the first part, Section 1, of his *Studies* (1939b, pp. 21–69), but a fully developed sequence analysis had already been presented in two unpublished papers of 1934a (published in Steiger, 1971, pp. 204–11) and 1935 (cf. Hansson, 1982, ch. 9). Furthermore, in Section 2 of the first part of his book Lindahl was the first economist who, in an extensive algebraic discussion of the relations between fundamental economic concepts (1939b, pp. 74–136), made the methodologically important distinction between ‘micro-economic’ and ‘macro-economic terms’ (p. 74) by which he tried to base the relations between macro values on some kind of microeconomic behaviour (pp. 111, 125; cf. Svennilsson, 1938, ch. 1; Siven, 1991, pp. 155–6; and Jens Christopher Andvig, 1991, p. 414). As shown by Hal R. Varian (1987, p. 461), this innovation has been wrongly attributed to Ragnar Frisch who, in an article of 1933 (pp. 172–3), had used the related terms ‘micro-dynamic’ and ‘macro-dynamic analysis’ in which he, however, ‘was uninterested in the problems of microeconomic roots’ of macroeconomics (Andvig, 1991, p. 415). Incorporating the method of *ex ante* and *ex post* (cf. Steiger, 1987a) developed by Myrdal (1932; 1933) in his disequilbrium analysis and adopted by Ohlin (1934, ch. 1), where the former had criticized Lindahl's method of temporary equilibrium, and taking care of the sequence analysis of consecutive periods formulated by Hammarskjöld (1933a; 1933b, chs 1–5) and Svennilsson (1938, ch. 1), Lindahl's dynamic method in 1939b consisted of two parts: (*a*) a single-period analysis where *ex ante* plans determine *ex post* results; and (*b*) a continuation analysis where these *ex post* events lead to revised *ex ante* plans of a subsequent period. While Lindahl allowed for disequilibrium as long as he analysed a single period only, his analysis for several periods demanded equilibrium within each period. Because of this assumption Lindahl's sequence analysis – although it can be regarded as the first dynamic method with a meaningful sequential character, that is, not relying on the mutual interdependence of all events – did not imply the solution to the dynamic problem of establishing a convincing explanation of the causal connection between successive periods. In the end, while acknowledging Myrdal's plea for disequilibrium analysis, Lindahl hesitated to rely on the ‘cumbersome *ex ante* and *ex post* terminology’ (1939b, p. 68; cf. 1939c, pp. 264–5) because of its ‘analytical complexity’ (Siven, 2006b, p. 694; cf. 1985, p. 590; and Hansen, 1981, p. 274).

It was left to Lundberg's sequence analysis of 1937 (ch. 9) to overcome this limitation by allowing for disequilibrium within the different periods with the help of the assumption of constant expectation functions (cf. Hansson, 1982, ch. 10). Lindahl accepted Lundberg's method in his Studies (1939b, pp. 57–9), but was not keen on the time-related model sequences based on difference equations which were incorporated in the latter's construction. On the contrary, this dynamic method was rejected by Lindahl because of its mechanical character resulting from the assumption that expectations need not enter explicitly. However, it was exactly this approach which came to dominate dynamic theory until the late 1960s when general equilibrium theorists reintroduced the notion of temporary equilibrium and developed dynamic models which look very similar to Lindahl's sequence analysis (for example, Frank H. Hahn, 1980).

Monetary and macroeconomic theory

While Lindahl's contributions to dynamic analysis were formulated independently of Wicksell, his work

on monetary and macroeconomic theory was clearly derived from the latter (1898; 1906). This influence can be traced back as far as Lindahl's first treatise on monetary matters, *Penningpolitikens mål* (1924; 1929b), where he systematized and extended the concepts used in the Swedish controversy between Wicksell and David Davidson before and after the First World War on the aim of monetary policy being to preserve the real value of contracts, that is, Wicksell's desideratum of a constant price level versus Davidson's proposal of price level variations in inverse proportion to changes in productivity (cf. Hammarskjöld, 1944; Carl G. Uhr, 1960, pp. 270–305; Klas Fregert, 1993; and Siven, 2002, 124–9).

While Lindahl's analysis is worked out along Wicksellian lines, he ends up with Davidson's and not Wicksell's solution by showing that the latter's proposition of a *'normal' rate of interest*, that is, the particular level of the money rate which is equal to the 'natural' rate, determined by the marginal productivity of capital, does not hold for a constant price level in the face of productivity variations.

A more systematic treatment of Wicksell's concept of the normal rate was given in *Penningpolitikens medel* (1930, pp. 121–30; cf. 1939a, pt. II, pp. 245–57), where Lindahl was the first to show that this notion implies three different conditions for equilibrium: '(1) it corresponds to the *natural* or … the *real rate of interest*; (2) it establishes *equilibrium between the demand for and supply of saving* [that is, investment and savings]; and (3) it is *neutral in* relation to *the price level* – whereas a rate of interest above or below "normal" will influence the price level in a downward or upward direction' (1939a, pt. II, p. 246; cf. 1930, p. 122). It was this formulation which inspired Myrdal's famous reconstruction of Wicksell's normal rate (1932; 1933; 1939) in which the three conditions were characterized as *monet*ary equilibrium (cf. Steiger, 1987c, p. 507; Siven, 2006a, pp. 11–12; 2006b, pp. 672–4).

However, in the central part of *Penningpolitikens medel*, the analysis of the relation between the rate of interest and the price level, Lindahl (1930, pp. 131–4; cf. 1939a, pt. II, pp. 257–60; and 1939c, pp. 260–8) did not employ the notion of the normal rate, because his explicit consideration of expectations showed him the impossibility of a unique equilibrium rate irrespective of the rate of change of the level of prices – a reasoning very similar to John Maynard Keynes's emphasis in the *General Theory* (1936, pp. 242–4) that there are different normal rates for different levels of employment. Instead, Lindahl explained changes in the general price level with the help of another concept introduced by Wicksell (1906, p. 159): the approach of *aggregate demand and supply*. In Lindahl's formulation of this approach changes in the price level were determined by changes in the relation between the total demand for and the total supply of consumption goods, the total demand for consumption goods being defined as 'the portion of the total nominal income which is not saved', $E\ (1 - s)$ where E denotes total nominal income and s the ratio of saving to income, and the total supply defined as PQ, where P denotes the price level and Q the quantity of consumer goods of a certain period (1930, pp. 12–13; cf. 1939a, pt. I, pp. 142–3). In general, he never analysed imbalances in macroeconomic variables 'caused by "wrong" relative prices' (Siven, 2002, p. 141). Using Wicksell's suggestion of a perfect credit system, this approach left no room for the quantity of money either, and it was indeed, in Lindahl's analysis of the issue of money by the central bank, directed against the quantity theory of money, although he did not deprive it of all significance for the theory of money (cf. 1929d, p. 18; 1955, pp. 32–4). In fact, as has been first emphasized by Hansen (1979, p. 123) and later confirmed by Axel Leijonhufvud (1991, p. 464) and Lars Werin (1991, p. 178) as well as Mauro Boianovsky and Hans-Michael Trautwein (2006, pp. 881–2, 888–95), Lindahl in his later writings (cf. 1957, pp. 13–15, 19–21) was 'a true monetarist' and the first one to formulate the 'accelerationist' hypothesis in inflation theory offered a decade later by Edmund S. Phelps

(1967) and Milton Friedman (1968). This hypothesis lies also at the heart of Lindahl's critical reformulation of *The General Theory*, in which he criticized Keynes's method of comparative statics in the equilibration of savings and investment by changes in income and employment, because it presupposes ‘correct anticipations’ (Lindahl, 1953, p. 27; cf. already 1934a, pp. 209–10; 1939c, pp. 264–5). Instead, Keynes should have relied on the dynamic analysis of changes of monetary and real variables where, like in his own analysis, expectations are allowed to adapt to the changes, but where expectational errors may nevertheless have effects that alter the equilibrium rates of interest and (un) employment (cf. Boianovsky and Trautwein, 2006, p. 897). It is most interesting to note that Keynes (1934), in a correspondence with Lindahl (1934b) on the latter's paper of 1934a, rejected Lindahl's method because its ‘dealing with time leads to undue complications and will be very difficult either to apply or to generalise about’.

Although Lindahl did not attempt to explicitly explain changes in output and employment, his aggregate analysis resulted in achievements which paved the way for Myrdal's (1932; 1933; 1939) and Ohlin's (1933; 1934, chs 1–3) monetary approaches, and which are still important for modern macroeconomics: (*a*) the use of the savings ratio *s* in the expression E(1 – s) which related saving to expected income and which can be regarded as an alternative formulation of Keynes's (1936) propensity to consume, led to a definite distinction between saving and investment, with Lindahl (1929c, pp. 11–12, written in 1927–28; cf. Hansen, 1981, p. 261) being the first economist to see the independence of the latter from the former variable; (*b*) their distinction allowed him to divide aggregate income into saving and consumption demand and aggregate output into investment and consumer goods; (*c*) the ‘paradox of savings’ could be solved according to which a reduction in savings results in increased savings; (*d*) the assumption of unused resources was introduced (1930, pp. 42–51; cf. 1939a, 176–9 and 185–6), and unemployment was explained by deflation caused by a fall of aggregate demand where even the possibility of a stable unemployment equilibrium was visualized (1929c, pp. 43–4; 1930, p. 44); however, deleted in 1939a, pt. II; cf. Hansen, 1981, pp. 261–3). Compared with Keynes's principle of effective demand determining the equilibrium level of (un)employment there are, however, certain limitations in Lindahl's aggregate demand/supply approach: (*a*) unemployment equilibrium was considered only as an ‘exceptional case’, with – like in the other Stockholm School analyses on the relation between unemployment and wages (esp. Alf Johansson, 1934, ch. 5) – ‘rigid money wages as a necessary condition for’ and no ‘complete macro model of unemployment’ (Hansen, 1981, pp. 268–9; cf. Siven, 2002, p. 141); (*b*) the rate of interest was treated in its orthodox role as equilibrator of savings and investment in the long run; (*c*) saving and investment were not equilibrated by changes in aggregate income but by variations in its *distribution* (cf. the discussion initiated by Karl-Gustav Landgren, 1960, ch. 6:3; and followed up by Steiger, 1971, pp. 173–9; 1978, pp. 424–5; 1991, 129–30; Hansen, 1981, pp. 261–3; Don Patinkin, 1982, pp. 44–6; and Johan Myrman, 1991, pp. 272–6). In the analysis of this adjustment process, however, Lindahl was able to anticipate the whole neo-Keynesian or Kaldor–Pasinetti theory of distributive shares (cf. Guglielmo Chiodi and Kumaraswamy Velupillai, 1983; Velupillai, 1988).

On the other side, the equilibrating role of changes in total income with respect to saving and investment is implicit in Lindahl's sequence analysis of 1934 (1934a, pp. 208–11), where he showed how a difference between investment and saving *ex ante* leads to their *ex post* equality, and it was clearly visualized in his discussion of loan-financed public works as a means against unemployment (1932, pp.

136–7; 1935, pp. 1–5; cf. 1939a, app., pp. 356–67). As has been pointed out by Hansen (1955, p. 41), Lindahl in *Penningpolitikens medel* (1930, pp. 63–8; however, deleted in 1939a, pt. II) was the first economist to consider the possibility of a systematic use of variations in the relation between public expenditures and public incomes, that is, the budget balance, as a means to stabilize economic fluctuations because, as he recognized, a surplus in the balance can be defined as equivalent to state saving and a deficit as state investment (1930, p. 65). With this analysis he paved the way for Ohlin's (1934, ch. 5) and Myrdal's (1934, pts III–IV) more detailed analysis of loan-financed public works as remedies against unemployment. Unlike Keynesian economists, however, Lindahl in this discussion did not neglect the effects of such a fiscal policy for the national debt which he analysed in detail in 1944 (cf. 1946). There he formulated rules governing state borrowing which are very similar to those developed at the same time by Evsey D. Domar (1944), that is, that the problem of public debts is first and foremost a problem of the growth of national income.

Another innovation in Lindahl's monetary and macroeconomic theory was his discussion of how to organize the central banking system in a monetary union of independent nations (cf. 1930, pp. 170–9; however, deleted in 1939a, pt. II). As recently recognized by Gunnar Heinsohn and Steiger (2003, p. 13; cf. Steiger, 2007, pp. 43–5), Lindahl was the first economist to develop the model of a decentralized, two-stage central banking system for a common currency consisting of a main central bank and the national central banks, where the latter would receive the banknotes in the same way from the former like the domestic commercial banks of their national central bank. With this model he hoped to open the possibility for the main central bank to equilibrate differences in real rates of interest due to different rates of inflation between the union's members by allowing for differences in nominal rates of interest. With this proposal, Lindahl anticipated the central Achilles’ heel of the Eurosystem, the central banking system of the European Monetary Union (EMU), where its Council can decide only on a unique nominal rate of interest (cf. Dieter Spethmann and Steiger, 2005, pp. 55–8). Furthermore, in spite of its name, the European Central Bank is not designed for issuing money and, therefore, not the central monetary authority of the EMU that could push through such a differentiation of credit. In his model, Lindahl (1930, p. 171) also demonstrated the necessity of a central fiscal authority in a monetary union to support the main central bank – another problem that has not been solved in the EMU (cf. Heinsohn and Steiger, 2003, pp. 13, 39).

Concepts of income and capital

While Lindahl's contributions to macroeconomic theory have been discussed extensively in the literature on the Stockholm School, his work on the macroeconomic concepts of income and capital have not received much attention (cf. Yohe, 1962).

The discussion of the notions of capital and income in Lindahl's theoretical framework stemmed from two different roots: (*a*) the approach to capital theory conceiving capital goods as stored services of land and labour, originally formulated by Eugen von Böhm-Bawerk (1889) and developed by Wicksell (1893; 1901) and Gustaf Åkerman (1923–4); and (*b*) the approach to the concept of income regarding income as a flow of benefits from the stock of capital and introduced into economic theory by Irving Fisher (1906). Both approaches had in common that *time* was included in a decisive manner and in

concentrating on this element, Lindahl made the notions of income and capital essentially correlative. In his piece on capital theory where he had introduced the method of intertemporal equilibrium (1929a; 1939a, pt. III), Lindahl avoided the theoretical difficulties of working with the concept of total capital in a world of heterogeneous capital goods by developing a completely disaggregated stationary equilibrium system – a ‘Walrasian model with capital à la Wicksell’ and with ‘a striking similarity’ to John von Neumann's model of equilibrium growth of 1937 (Hansen, 1970, pp. 199, 207–8). Although Lindahl did not make use of the concept of total capital in his equilibrium system, it can be shown that its total capital value can be determined on the basis of the term ‘income’ employed in his model – an insight which Lindahl formulated unequivocally in *Penningpolitikens medel* (1930, pp. 13–15; cf. 1939a, pt. II, pp. 143–6) and in more detail in his contributions on the concept of income (1933; 1937b, pp. 76–111). Starting from Irving Fisher's basic premise that income consists of the services obtained from capital goods during a certain period, whereas capital is a stock existing at a given point of time, Lindahl (1933, pp. 400–1) looked upon *income as interest* accruing on the value of capital goods, and considered capital value as future income discounted. With this concept of income, implying that income is equal to the sum of consumption *and* saving, he solved the inconsistencies in Fisher's analysis of capital and income where saving, a flow term expressing the increase in capital value, had been excluded from income and incorporated into capital, a stock term. Consequently, Lindahl's discussion also made clear that changes in capital value, contrary to what had been the premises of Böhm-Bawerk, Wicksell and Åkerman, are not determined by changes in the use of capital but in the use of income, that is, the part which is not consumed: saving.

However, as Lindahl realized, this thesis holds true only under the assumption of perfect foresight. Following Myrdal's analysis of expectations of 1927, he showed that as soon as uncertainty about future events is introduced changes in anticipation of owners of capital assets lead to changes in capital value by *gains* and *losses* which cannot be regarded as positive or negative income, because like the stock of capital they refer not to a certain period but to a point of time (1929a, p. 75; 1939a, pt. III, p. 341; cf. Myrdal, 1927, p. 44, and the further discussion in Lindahl, 1939b, pp. 101–10). This insight led Lindahl to abandon the most practical concept of income – income as earnings – because it included gains and losses. Although Lindahl's concept of income – like his abstract classification of capital goods (Hansen, 1970, p. 200) – has been criticized for not being capable of empirical application and measurement, his contributions – together with Myrdal's approach of 1927 – have been acknowledged as ‘the fundamental theoretical work concerning the notion of income’ (Nicholas Kaldor, 1955, p. 162). This work should also become fundamental to Lindahl's research on national accounting (1937b; 1939b, pp. 74–136; 1954) which made him ‘the father of Social Accounting *theory*’ (Hicks, 1973, p. 8; cf. Carlson, 1982, 33–6).

During his lifetime – and until the late 1980s – Lindahl never earned a reputation comparable to that of his colleagues in the early Stockholm School, Gunnar Myrdal and Bertil Ohlin, and it has been argued by one of his younger colleagues (Lundberg, 1982, p. 275) that his contributions to economics were not distinguished by ‘ingenious ideas’ like those of Myrdal and Ohlin. As has been shown in this survey of Lindahl's work, however, the numerous original ideas in each of the four fields covered by his writings argue for quite different judgement. This holds especially true for Lindahl's monetary and macroeconomic theory, as has been demonstrated since the late 1980s by Boianovsky and Trautwein (2006), Leijonhufvud (1991), Siven (1991; 2002; 2006a; 2006b), Steiger (1987a; 1987b; 1987c; 2006a;

2006b; 2007), and Velupillai (1988).

The author wishes to thank Claes-Henric Siven (Stockholms Universitet) and Hans-Michael Trautwein (Universität Oldenburg) for valuable suggestions

See Also

- ex ante and ex post
- Lindahl equilibrium
- Myrdal, Gunnar
- Ohlin, Bertil Gotthard
- Stockholm School

Selected works

1919. *Die Gerechtigkeit der Besteuerung. Eine Analyse der Steuerprinzipien auf der Grundlage der Grenznutzentheorie*. Lund: Gleerupska Universitetsbokhandeln and H. Ohlsson. Ch. 4 trans. as ‘Just taxation – a positive solution’, in Musgrave and Peacock (1958).

1924. *Penningpolitikens mål och medel. Del I* [The aims and means of monetary policy. Part I]. Lund: C. W.K. Gleerup; Malmö: Försäkringsaktiebolaget. 1st edn of Lindahl (1929b).

1928. Einige strittige Fragen der Steuertheorie. In *Die Wirtschaftstheorie der Gegenwart*, vol. 4, ed. H. Mayer. Vienna: J. Springer; abridged trans. as ‘Some controversial questions in the theory of taxation’, in Musgrave and Peacock (1958).

1929a. Prisbildningsproblements uppläggning från kapitalteoretisk synpunkt [The formulation of the theory of prices from the viewpoint of capital theory]. *Ekonomisk Tidskrift* 31, 31–81. Revised version trans. as Pt. III of Lindahl (1939a).

1929b. *Penningpolitikens mål* [The aims of monetary policy]. zund: C.W.K. Gleerup; Malmö: Försäkringsaktiebolaget. 2nd edn of Lindahl (1924).

1929c. *Penningpolitikens medel* [The means of monetary policy]. Lund: C.W.K. Gleerup; Malmö: Försäkringsaktiebolaget. 1st edn of Lindahl (1930).

1929d. *Om förhållandet mellan penningmängd och prisnivån* [On the relation between the quantity of money and the price level]. Uppsala: Lundequistska and Almqvist & Wiksell.

1930. *Penningpolitikens medel* [The means of monetary policy]. Lund: Gleerup; Malmö: Försökringsaktiebolaget. 2nd edn of Lindahl (1929c); revised version trans. as Pt. II of Lindahl (1939a).

1932. Offentliga arbeten i depressionstider [Public works in times of depression]. *Nationalekonomiska Föreningens Förhandlingar* [Proceedings of the Swedish Economic Association], Meeting of 25 November, pp. 127–37, 163–4.

1933. The concept of income. In *Economic Essays in Honour of Gustav Cassel*. October 20th, 1933. London: Allen & Unwin.

1934a. A note on the dynamic pricing problem. Mimeo, Gothenburg, 13 October. Quoted from the corrected version published in Steiger (1971).

1934b. Letter to John Maynard Keynes. Gothenburg, 7 November. In Steiger (1971).

1935. Arbetslöshet och finanspolitik [Unemployment and fiscal policy]. *Ekonomisk Tidskrift* 37, 1–36. Revised version trans. as App. to Lindahl (1939a).

1937a. (with E. Dahlgren and K. Kock.) *National Income of Sweden 1861–1930. In Two Parts*. Vol. III of *Wages, Cost of Living and National Income in Sweden 1860–1930*, ed. the Staff of the Institute for Social Sciences, University of Stockholm. London: P.S. King & Son; Stockholm: P.A. Norstedt & Söner.

1937b. ‘National income’, the concept and methods of estimation. In Lindahl (1937a), Pt. I.

1939a. *Studies in the Theory of Money and Capital*. London: Allen & Unwin.

1939b. The dynamic approach to economic theory. Pt. I of Lindahl (1939a).

1939c. Additional note (1939) (to Lindahl, 1939a, Pt. II). In Lindahl (1939a).

1944. Teorien för den offentliga skuldsättningen [The theory of the public debt]. In *Studier i ekonomi och historia tillägnade Eli F. Heckscher* [Studies in economics and history in honour of Eli F. Heckscher]. Uppsala: Almqvist & Wiksell. Abridged version trans. as Lindahl (1946).

1946. The problem of the growing national debt. *Skandinaviska Banken Quarterly Review* 27(2), 43–8. Abridged version of Lindahl (1944).

1951. Till hundraårsminnet av Knut Wicksells födelse [On the centenary of Knut Wicksell's birth]. *Ekonomisk Tidskrift* 53, 197–243. Quoted from the abridged version trans. as ‘Wicksell's Life and Work’, in K. Wicksell, *Selected Papers on Economic Theory*, ed. E. Lindahl. London: Allen & Unwin, 1958.

1953. Om Keynes’ ekonomiska system. *Ekonomisk Tidskrift* 55, 186–243. Quoted from and trans. as ‘On Keynes’ economic system I–II’, *Economic Record* 30 (1954), 19–32 and 159–71.

1954. Nationalbokföringens grundbegrepp. *Ekonomisk Tidskrift* 56, 87–138. Trans. as ‘The basic concepts of national accounting’, *International Economic Papers* 7 (1957), 71–100.

1955. Penningteoretiska utgångspunkter [Starting points in monetary theory]. In *Om Riksbankens sedelutgivningsrätt och därmed sammanhängende frågor* [On the Riksbanken's right to issue banknotes and related questions]. Stockholm: Statens offentligar utredningar 1955:43, 32–57.

1957. *Spelet om penningvärdet*. Stockholm: Kooperativa Förbundet. Quoted from and trans. as ‘Das Spiel mit dem Geldwert’, *Weltwirtschaftliches Archiv* 87 (1961), 7–53.

1959. Om skatteprinciper och skattepolitik. In *Ekonomi, politik, samhälle. En bok tillägnad Bertil Ohlin* [Economics, politics, society. A publication in honour of Bertil Ohlin], Stockholm: Folk & Samhälle and Esselte. Trans. as ‘Tax principles and tax policy’, *International Economic Papers* 10 (1960), 7–23.

Bibliography

Åkerman, J.G. 1923–4. *Realkapital und Kapitalzins*. 2 vols. Stockholm: Centraltryckeriet.

Andvig, J.C. 1991. Ragnar Frisch and the Stockholm School. In Jonung (1991).

Böhm-Bawerk, E. von. 1889. *Kapital und Kapitalzins. Zweite Abteilung: Positive Theorie des Kapitales*. Jena: G. Fischer; 4th edn, 1921. Trans. as *Capital and Interest*, vols 2 and 3, South Holland, IL: Libertarian Press, 1959.

Bohm, P. 1987. Lindahl on public finance. In *The New Palgrave. A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Boianovsky, M. and Trautwein, H.-M. 2006. Price expectations, capital accumulation and employment. Lindahl's macroeconomics from the 1920s to the 1950s. *Cambridge Journal of Economics* 30, 881–900.

Carlson, B. 1982. Bagge, Lindahl och nationalinkomsten. Om ‘National Income of Sweden 1861–1930’ [Bagge, Lindahl and the national income. On ‘National Income of Sweden 1861–1930’]. *Meddelande från Ekonomisk–historiska institutionen, Lunds Universitet*, No. 27.

Chiodi, G. and Velupillai, K. 1983. A note on Lindahl's theory of distribution. *Kyklos* 36, 103–11.

Debreu, G. 1959. *Theory of Value. An Axiomatic Analysis of Economic Equilibrium*. New York: Wiley.

Domar, E.D. 1944. The ‘burden of the debt’ and the national income. *American Economic Review* 34, 798–827.

Fisher, I. 1906. *The Nature of Capital and Income*. New York: Macmillan. 2nd edn, 1912.

Frisch, R. 1933. Propagation problems and impulse problems in economic dynamics. In *Economic Essays in Honour of Gustav Cassel. October 20th, 1933*. London: Allen & Unwin.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Fregert, K. 1993. Erik Lindahl's norm for monetary policy. In *Swedish Economic Thought. Explorations and Advances*, ed. L. Jonung. London: Routledge.

Grossekettler, H. 2006. Wicksell, Knut: Finanztheoretische Untersuchungen (1896). In *Lexikon der ökonomischen Werke. 650 wegweisende Schriften von der Antike bis ins 20. Jahrhundert*, ed. D. Herz and V. Weinberger. Düsseldorf: Verlag Wirtschaft und Finanzen.

Hahn, F.H. 1980. Unemployment from a theoretical viewpoint. *Economica* NS 47, 285–98.

Hammarskjöld, D. 1933a. Utkast till en algebraisk metod för dynamisk prisanalys [Outline of an algebraic method for the dynamic analysis of prices]. *Ekonomisk Tidskrift* 34(5–6) (1932, printed 1933), 157–76.

Hammarskjöld, D. 1933b. *Konjunkturspridningen. En teoretisk och historisk undersökning* [The propagation of business cycles. A theoretical and historical investigation]. Stockholm: Statens offentliga utredningar 1933:29.

Hammarskjöld, D. 1935. *Åtgärder mot arbetslöshet. Arbetslöshetsutredningens betänkande, del II* [Measures against unemployment. Report of the Swedish Unemployment Committee, Part II]. Stockholm: Statens offentliga utredningar 1935:6.

Hammarskjöld, D. 1944. Den svenska diskussionen om penningpolitikens mål. In *Studier i ekonomi och historia tillägnade Eli F. Heckscher* [Studies in economics and history in honour of Eli F. Heckscher]. Uppsala: Almqvist & Wiksell. Trans. as ‘The Swedish discussion on the aims of monetary policy’, *International Economic Papers* 5 (1955), 145–54.

Hansen, B. 1955. *Finanspolitikens ekonomiska teori*. Stockholm: Statens offentliga utredningar, 1955:25. Quoted from and trans. as *The Economic Theory of Fiscal Policy*, London: Allen & Unwin, 1958.

Hansen, B. 1970. *A Survey of General Equilibrium Systems*. New York: McGraw-Hill.

Hansen, B. 1979. Review of Erik Lundberg (ed.): Inflation Theory and Anti-Inflation Policy (1977). *Scandinavian Journal of Economics* 81, 119–25.

Hansen, B. 1981. Unemployment, Keynes, and the Stockholm School. *History of Political Economy* 13, 256–77.

Hansson, B. 1982. *The Stockholm School and the Development of Dynamic Method*. London: Croom Helm.

Hansson, B. 1987. Stockholm School. In *The New Palgrave. A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Hansson, B. 1991. The Stockholm School and the development of dynamic method. In *The History of Swedish Economic Thought*, ed. B. Sandelin. London: Routledge.

Hayek, F.A. 1928. Das intertemporale Gleichgewichtssystem der Preise und die Bewegungen des 'Geldwertes'. *Weltwirtschaftliches Archiv* 28, 33–76. Trans. as 'Intertemporal price equilibrium and movements in the value of money', in F.A. von Hayek, *Money, Capital and Fluctuations. Early Essays*, ed. R. Cloughry. London: Routledge & Kegan Paul, 1984.

Heinsohn, G. and Steiger, O. 2003. The European Central Bank and the Eurosystem. An analysis of the missing central monetary institution in the European Monetary Union. ZEI Working Paper No. B03-09. Center for European Integration Studies (ZEI), Universität Bonn. Repr. in *The Euro, the Eurosystem, and the European Economic and Monetary Union*, ed. D. Ehrig and O. Steiger. Hamburg: LIT-Verlag, 2007.

Hicks, J.R. 1939. *Value and Capital. An Inquiry into Some Fundamental Principles of Economic Theory*. Oxford: Clarendon Press. 2nd edn, 1946.

Hicks, J.R. 1973. Recollections and documents. *Economica* NS 40, 2–11.

Hicks, Sir J. 1985. *Methods of Dynamic Economics*. Oxford: Oxford University Press.

Hicks, Sir J. 1991. The Swedish influence on *Value and Capital*. In Jonung (1991).

Johansson, A. 1934. *Löneutvecklingen och arbetslösheten* [The development of wages and unemployment]. Stockholm: Statens offentliga utredningar 1934:2.

Jonung, L., ed. 1991. *The Stockholm School of Economics Revisited*. Cambridge: Cambridge University Press.

Kaldor, N. 1955. *An Expenditure Tax*. London: Allen & Unwin. Quoted from the reprint of pp. 54–78, 'The concept of income in economic theory', in *Readings in the Concept and Measurement of Income*, ed. R.H. Parker and G.C. Harcourt. Cambridge: Cambridge University Press, 1969.

Keynes, J.M. 1934. Letter to Erik Lindahl. Bloomsbury, 8 December. Quoted from the reproduction in Steiger (1971).

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Landgren, K.-G. 1960. *Den 'nya ekonomien' i Sverige. J.M. Keynes, B. Ohlin och utvecklingen 1927–39* [The 'new economics' in Sweden. J.M. Keynes, B. Ohlin and the development 1927–39). Stockholm and Uppsala: Almqvist & Wiksell.

Leijonhufvud, A. 1991. Roundtable discussion. In Jonung (1991).

Lindahl, G. and Wallmén, O. 1960. Erik Lindahl. Bibliografi 1919–1960 [Erik Lindahl. Bibliography 1919–1960]. *Ekonomisk Tidskrift* 62, 59–74.

Lundberg, E. 1930. Om begreppet ekonomisk jämvikt [On the notion of economic equilibrium]. *Ekonomisk Tidskrift* 32, 133–60.

Lundberg, E. 1937. *Studies in the Theory of Economic Expansion*. Stockholm: Norstedt & Söner.

Lundberg, E. 1982. Lindahl, Erik. In *Svenskt biografiskt lexikon* [The Swedish Dictionary of Bibliography] 23. Stockholm: Norstedts Tryckeri.

Milgate, M. 1982. *Capital and Employment. A Study of Keynes's Economics*. London: Academic Press.

Musgrave, R.A. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

Musgrave, R.A. and Peacock, A.T., eds. 1958. *Classics in the Theory of Public Finance*. London: Macmillan. 2nd edn, 1967.

Myrdal, G. 1927. *Prisbildningsproblemet och föränderligheten* [The problem of price formation and changeability]. Uppsala and Stockholm: Almqvist & Wiksell.

Myrdal, G. 1932. Om penningteoretisk jämvikt. En studie över den 'normala räntan' i Wicksells penninglära [On monetary equilibrium. A study on the 'normal rate of interest' in Wicksell's monetary doctrine]. *Ekonomisk Tidskrift* 33 (1931, printed 1932), 191–302. Revised version trans. as Myrdal (1933).

Myrdal, G. 1933. Der Gleichgewichtsbegriff als Instrument der geldtheoretischen Analyse. In *Beiträge zur Geldtheorie*, ed. F.A. Hayek. Vienna: J. Springer. 1st rev. version of Myrdal (1932); 2nd rev. version trans. as Myrdal (1939).

Myrdal, G. 1934. *Finanspolitikens ekonomiska verkningar* [The economic effects of fiscal policy]. Stockholm: Statens offentliga utredingar 1934: 1.

Myrdal, G. 1939. *Monetary Equilibrium*. London: Hodge. Revised version of Myrdal (1933).

Myrman, J. 1991. The monetary economics of the Stockholm School. In Jonung (1991).

Neumann, J. von. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Browerschen Fixpunktsatzes. In *Ergebnisse eines mathematischen Kolloquiums* 8, ed. K. Menger. Trans. as ‘A model of general equilibrium’, *Review of Economic Studies* 13 (1945–6), 1–9.

Ohlin, B. 1933. Till frågan om penningteoriens uppläggning [On the question of the method and structure in monetary theory]. *Ekonomisk Tidskrift* 35, 45–81. Quoted from and translated as ‘On the formulation of monetary theory’, *History of Political Economy* 10 (1978), 353–88.

Ohlin, B. 1934. *Penningpolitik, offentliga arbeten, subventioner och tullar som medel mot arbetslöshet. Bidrag till expansionens teori* [Monetary policy, public works, subsidies and tariffs as remedies for unemployment. A contribution to the theory of expansion]. Stockholm: Statens offentliga utredningar 1934:12.

Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment. I–II. *Economic Journal* 47, 53–69 and 221–240.

Patinkin, D. 1982. *Anticipations of the General Theory? And Other Essays on Keynes*. Chicago: University of Chicago Press.

Petersson, J. 1987. *Erik Lindahl och Stockholsskolans dynamiska metod* [Erik Lindahl and the dynamic method of the Stockholm School]. Lund: Universitetsförlaget Dialogus.

Phelps, E.S. 1967. Phillips Curves, expectations of inflation and optimal unemployment over time. *Economica* NS 34, 254–81.

Roberts, J. 1987. Lindahl equilibrium. In *The New Palgrave. A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan; New York: Stockton Press.

Samuelson, P.A. 1954. The pure theory of public expenditures. *Review of Economics and Statistics* 36, 387–9.

Siven, C.-H. 1985. The end of the Stockholm School. *Scandinavian Journal of Economics* 87, 577–93.

Siven, C.-H. 1991. Expectation and plan. The microeconomics of the Stockholm School. In Jonung

(1991).

Siven, C.-H. 2002. Analytical foundations of Erik Lindahl's monetary analysis, 1924–1930. *History of Political Economy* 34, 111–53.

Siven, C.-H. 2006a. Stockholmsskolan och Keynes [The Stockholm School and Keynes]. *Økonomisk Forum* 60(6), 10–7.

Siven, C.-H. 2006b. Monetary equilibrium. *History of Political Economy* 38, 665–709.

Spethmann, D. and Steiger, O. 2005. The four Achilles' heels of the Eurosystem. Missing central monetary institution, different real rates of interest, nonmarketable securities, and missing lender of last resort. *International Journal of Political Economy* 34(2), (2004, printed Summer 2005) 46–68.

Steiger, O. 1971. *Studien zur Entstehung der Neuen Wirtschaftslehre in Schweden. Eine Anti-Kritik*. Berlin: Duncker & Humblot.

Steiger, O. 1978. Prelude to the theory of a monetary economy. Origins and significance of Ohlin's 1933 approach. *History of Political Economy* 10, 420–46.

Steiger, O. 1987a. *Ex ante* and *ex post*. In *The New Palgrave. A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan; New York: Stockton Press.

Steiger, O. 1987b. Lindahl, Erik Robert (1891–1960). In *The New Palgrave. A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan; New York: Stockton Press.

Steiger, O. 1987c. Monetary equilibrium. In *The New Palgrave. A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan; New York: Stockton Press.

Steiger, O. 1991. Comment on Eskil Wadensjö: 'The Committee of Unemployment and the Stockholm School'. In Jonung (1991).

Steiger, O. 2006a. Lindahl, Erik Robert: *Penningpolitikens medel* [The means of monetary policy] (1930). In *Lexikon der ökonomischen Werke. 650 wegweisende Schriften von der Antike bis ins 20. Jahrhundert*, ed. D. Herz and V. Weinberger. Düsseldorf: Verlag Wirtschaft und Finanzen.

Steiger, O. 2006b. Lindahl, Erik Robert: *Studies in the Theory of Money and Capital* (1939). In *Lexikon der ökonomischen Werke. 650 wegweisende Schriften von der Antike bis ins 20. Jahrhundert*, ed. D. Herz and V. Weinberger. Düsseldorf: Verlag Wirtschaft und Finanzen.

Steiger, O. 2007. Erik Lindahl och Eurosystemet [Erik Lindahl and the Eurosystem]. *Ekonomisk Debatt*

35(2), 42–5.

Svennilsson, I. 1938. *Ekonomisk planering. Teoretiska studier* [Economic planning. Theoretical studies]. Uppsala: Almqvist & Wiksell.

Uhr, C.G. 1960. *Economic Doctrines of Knut Wicksell*. Berkeley and Los Angeles: University of California Press.

Varian, H.R. 1987. Microeconomics. In *The New Palgrave. A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Velupillai, K. 1988. Somer Swedish stepping stones to modern macroeconomics. *Eastern Economic Journal* 14, 87–98.

Werin, L. 1991. There were *two* Stockholm Schools. In Jonung (1991).

Wicksell, K. 1893. *Über Wert, Kapital und Rente*. Jena: G. Fischer. Trans. as *Value, Capital and Rent*, London: Allen & Unwin, 1954.

Wicksell, K. 1896. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens*. Jena: G. Fischer, iv–vi, 76–87, 101–59. Trans. as ‘A new principle of just taxation’, in Musgrave and Peacock (1958).

Wicksell, K. 1898. *Geldzins und Güterpreise. Eine Studie über die den Tauschwert des Geldes bestimmenden Ursachen*. Jena: G. Fischer. Trans. as *Interest and Prices. A Study of the Causes Regulating the Value of Money*. London: Macmillan, 1936.

Wicksell, K. 1901. *Föreläsningar i nationalekonomi*. Vol. 1, Stockholm: Fritzes; Lund: Berlingska. Trans. of the 3rd Swedish edn (1928) as *Lectures on Political Economy. Volume 1: General Theory*, London: Routledge & Sons, 1934.

Wicksell, K. 1906. *Föreläsningar i nationalekonomi*. Vol. 2: *Om penningar och kredit*. Stockholm: Fritzes; Lund: Berlingska. Quoted from the trans. of the 3rd Swedish edn (1929), *Lectures on Political Economy*. *Volume 2*: *Money*, London: Routledge & Sons, 1935.

Yohe, W.P. 1959. *The Wicksellian Tradition in Swedish Macroeconomic Theory*. Ann Arbor, MI: University Microfilms.

Yohe, W.P. 1962. A note on some lesser known works of Erik Lindahl. *Canadian Journal of Economics and Political Science* 28, 274–80.

How to cite this article

Steiger, Otto. "Lindahl, Erik Robert (1891–1960)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000103> doi:10.1057/9780230226203.0973

The New Palgrave Dictionary of Economics Online

linear models

Edwin Burmeister

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Von Neumann's linear economic model was first published in German in 1938 and translated into English in 1945. Since then there have been numerous economic and mathematical refinements, most of which are summarized in Burmeister and Dobell (1970) and/or Morishima (1969). The original von Neumann formulation did not admit either primary factors or final consumption. However, in the generalized von Neumann model described below, one primary factor, labour, is allowed, as well as a vector of final consumption goods. A further generalization allowing a vector of different primary factors is possible. Accordingly, linear models of Leontief–Sraffa (1960) type become a special case.
Assume there exist *m* alternative production activities for producing *n* different commodities, where *m*•*n*. Activity *j* operating at a unit intensity level requires a labour input a_{0j} and a vector of commodity inputs $(a_{1j}, a_{2j}, \ldots, a_{nj})$ to produce a vector of commodity outputs $(b_{1j}, b_{2j}, \ldots, b_{nj})$. Production takes one time period, so inputs at time *t* result in outputs at time *t*+1. Constant returns to scale is implied by linearity, and inputs λa_{0j} and $(\lambda a_{1j}, \lambda a_{2j}, \ldots, \lambda a_{nj})$ yield outputs $(\lambda b_{1j}, \lambda b_{2j}, \ldots, \lambda b_{nj})$ for all $\lambda \geq 0$.
The vector of labour requirements for the *m* alternative production activities is written as $A_0 = (a_{01}, a_{02}, \ldots, a_{0m})$. The input matrix is

$$\left[\frac{A_0}{A}\right] = \begin{bmatrix} \frac{a_{01}}{a_{11}} & - & \frac{a_{0m}}{a_{1m}} \\ \vdots & & \vdots \\ a_{n1}, & \ldots & , a_{nm} \end{bmatrix}$$

the output matrix is

$$B = \begin{bmatrix} b_{11} & \dots & b_{1m} \\ \vdots & & \vdots \\ b_{11} & \dots & b_{nm} \end{bmatrix}$$

and the intensity levels at which each of the *m* production activities is operated is given by the column vector

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

Although in general some of the a_{0j}'s, a_{ij}'s and b_{ij}'s may be zero, here we assume that they are all positive; we thereby avoid some technical difficulties and gain expositional simplicity.
Assume that labour grows at the rate $g \geq 0$,

$$L(t+1) = (1+g)L(t), \tag{1}$$

and is fully employed with

$$L(t) = A_0(t)x(t). \tag{2}$$

For all *t* production must satisfy the resource constraint

$$Ax \leq Bx - C \tag{3}$$

where *C* denotes the column vector of commodities consumed

$$C = \begin{bmatrix} C_1 \\ \vdots \\ C_m \end{bmatrix}$$

The economy is capable of producing positive final consumption and balanced growth at the rate *g* provided the inequalities

$$(1+g)Ax \leq Bx - C, \quad C \geq 0, \quad C \neq 0$$
$$x \geq 0, \quad x \neq 0$$
(4)

are satisfied.
Prices for a unit of labour services and the *n* commodities are given by $p_0 \equiv w$ and the row vector $p = (p_1, p_2, \ldots, p_n)$, respectively, where

$$\sum_{i=0}^{n} p_i = 1$$

is one convenient normalization. The steady-state (or balanced growth) rate of interest (or profit rate) is denoted by *r*.
The economy can achieve a steady-state equilibrium at a given value of $r \geq 0$ if the von Neumann price system has a solution satisfying the inequalities

$$wA_0 + (1+r)pA \geq pB, \quad w \geq 0, \quad w \neq 0$$
$$p \geq 0, \quad p \neq 0 .$$
(5)

The quantity system (4) is dual to the price system (5) when *r*=*g*.
The von Neumann solution to (5) involves three economically essential inequalities:

(i) If the cost of operating an activity exceeds the revenue from that activity, then that activity is not used in a steady-state equilibrium solution, that is that activity is operated at a zero intensity level. Formally,

$$x_j = 0 \text{ if } wa_{0j} + (1 + r)\sum_{i=1}^{n} p_i a_{ij} > \sum_{i=1}^{n} p_i b_{ij},\ j = 1, \ldots, m. \tag{6}$$

(ii) If a commodity has a positive price, then its supply and demand are equal:

$$(1 + g)\sum_{i=1}^{m} a_{ij} x_j = \sum_{j=1}^{m} b_{ij} x_j - C_i \text{ if } p_i > 0,\ i = 1, \ldots, n. \tag{7}$$

(iii) The price of a commodity is zero if it is in excess supply:

$$p_i = 0 \text{ if } (1 + g)\sum_{i=1}^{m} a_{ij} x_j < \sum_{j=1}^{m} b_{ij} x_j - C_i,\ i = 1, \ldots, n. \tag{8}$$

The above generalized von Neumann model is not a general equilibrium model because there is one missing equation. Some behavioural equation involving the rate of interest and consumption is required to form a general equilibrium model. Nevertheless, this incomplete specification can be used to confirm and generalize many well known results.
For example, in steady-state equilibrium equations (5), (6), (7) and (8) imply that

$$(1 + g)Ax + pC = pBx = \text{value of output} = wA_0 + p(1 + r)Ax. \tag{9}$$

Using (2), the per capita value of commodity inputs or capital is given by

$$V = \frac{pAx}{A_0 x};$$
(10)

similarly the per capita value of consumption is

$$pc = \frac{pC}{A_0 x}.$$
(11)

Substituting (10) and (11) into (9) and rearranging yields

$$pc = w + (r - g)V.$$
(12)

The Golden Rule result that the value of per capita consumption is equal to the wage rate at the Golden Rule point where *r*=*g* is an immediate consequence of (12). Other such results are easily derived; thus allowing for the possibility of joint production does not invalidate many economic results.
Two familiar classes of models are special cases of this generalized von Neumann model. First, Leontief–Sraffa models result simply by setting *m*=*n* and *B*=*I*, which implies that the technology is free of joint production. Then if *p*>0 from (7)

$$C = [I - (1 + g)A]x$$
(13)

where now the column vector *x* is interpreted as the output vector for commodities $1, \ldots, n$. Provided (4) has a solution, (13) may be solved for

$$x = [I - (1 + g)A]^{-1}C,$$
(14)

and premultiplying (14) by the vector A_0 gives the consumption possibility frontier

$$A_0 x = L = [I - (1 + g)A]^{-1} C. \tag{15}$$

Similarly, steady-state equilibrium prices for the Leontief–Sraffa model are given by

$$p = wA_0 [I - (1 + r)A]^{-1}. \tag{16}$$

Second, most neo-Austrian models of the type studied by Hicks (1973a) are a special case of this generalized von Neumann model. The latter fact is most easily demonstrated by considering the simple numerical example due to Burmeister (1974). A neo-Austrian process is a time sequence of input–output vectors

$$\{(a_t, b_t)\}_{t=0}^{T} \tag{17}$$

where a_t is the input of a commodity and b_t is the output (of the same commodity) in period *t*. Consider a process

$$\{(a_t, b_t)\}_{t=0}^{2} = \{(a_0, 0), (a_1, b_1), (a_2, 1)\}. \tag{18}$$

The neo-Austrian model (18) is *equivalent* to a von Neumann specification with

$$\left[-\frac{A_0}{A}-\right]=\begin{bmatrix} a_1 & a_1 & a_2 & \\ - & - & - & - \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & \end{bmatrix} \tag{19}$$

and

$$B=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & b_1 & 1 \end{bmatrix}; \tag{20}$$

see Burmeister (1974, pp. 441–4).

We see, therefore, that the generalized von Neumann model is extraordinarily useful for unifying several apparently different ways of describing the production technology. However, when we do not restrict our attention to steady-state equilibria, the dynamic evolution of the model becomes extremely complex. The inequalities (4) and (5) must be satisfied for each *t*, as well as some additional equation to determine the interest rate *r*. Known results on the dynamics of models with heterogeneous capital goods – see, for examples the discussion and references cited in Chapters 5 and 6 of Burmeister (1980) – warn us that the task of completely characterizing the dynamic properties of von Neumann models will not be easy. The fact that the von Neumann formulation admits joint production makes the task even harder.

See Also

- Hawkins–Simon conditions
- input–output analysis
- linear programming
- Marxian value analysis
- non-substitution theorems
- Perron–Frobenius theorem
- Sraffian economics

Bibliography

Burmeister, E. 1974. Synthesizing the neo-Austrian and alternative approaches to capital theory: a survey. *Journal of Economic Literature* 12(2), June, 413–56.

Burmeister, E. and Dobell, A.R. 1970. *Mathematical Theories of Economic Growth*. New York: Macmillan.

Burmeister, E. 1980. *Capital Theory and Dynamics*. New York: Cambridge University Press.

Burmeister, E. and Kuga, K. 1970. The factor-price frontier, duality and joint production. *Review of Economic Studies* 37(109), 11–19.

Hicks, J.R. 1973a. *Capital and Time. A Neo-Austrian Theory*. Oxford: Clarendon Press.

Hicks, J.R. 1973b. The Austrian theory of capital and its rebirth in modern economics. In *Carl Menger and the Austrian School of Economics*, ed. J.R. Hicks and W. Weber. New York and London: Oxford University Press.

Morishima, M. 1969. *Theory of Economic Growth*. Oxford: Clarendon Press.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

von Neumann, J. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums*, Vol. 8, ed. K. Menger, Leipzig: Verlag. Trans. by G. Morgenstern as ‘A model of general economic equilibrium’, *Review of Economic Studies* 13, (1945–6), 1–9. Reprinted in *Readings in Mathematical Economics*, ed. P. Newman, Vol. II, Baltimore: Johns Hopkins University Press, 1968, 221–9.

von Weizsacker, C.C. 1971. *Steady State Capital Theory*. New York: Springer-Verlag.

How to cite this article

Burmeister, Edwin. "linear models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000105> doi:10.1057/9780230226203.0974

The New Palgrave Dictionary of Economics Online

linear programming

George B. Dantzig
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An article by George Dantzig, the ‘father of linear programming’. The problem of minimizing or maximizing a function of several variables subject to constraints when all the functions are linear is called a ‘linear program’. Linear programs can be used to approximate the broad class of convex functions commonly encountered in economic planning. Thousands of linear programs are efficiently solved with the simplex method, an algorithm. Solving a model with alternative activities requires software not only for solving on computers large systems of equations but also for selecting the best combination from an astronomical number of possible combinations of activities.

Keywords

bimatrix games; convex program; Dantzig, G.; decomposition principle; Dantzig, G. B.; Kantorovich, L. V.; Koopmans, T. C.; Kuhn–Tucker conditions; Lagrange multipliers; Leontief input–output model; linear programming; mathematical programs; mini-max theorem; mixed strategies; simplex method for solving linear programs; von Neumann, J.

Article

A list of applications of linear programming, since it was first proposed in 1947 by G. Dantzig, could fill a small volume. Both J. von Neumann and L. Kantorovich made important contributions prior to 1947. Its first use by G. Dantzig and M. Wood was for logistical planning and deployment of military forces. A. Charnes and W. Cooper in the early 1950s pioneered its use in the petroleum industry. S. Vajda and E.M.L. Beale were early pioneers in the field in Great Britain. In socialist countries, it is used to determine the plan for optimal growth of the economy. Thousands of linear programs are efficiently solved each day all over the world using the simplex method, an algorithm, also first proposed in 1947. Many problems which once could only be solved on high-speed mainframe computers can now be solved on personal computers.

The problem of minimizing or maximizing a function f_0 of several variables $X = (X_1, X_2, \ldots, X_n)$ subject to constraints $f_i(X) \leq 0$, $i = 1, \ldots, n$ is called a *mathematical program*. When all the functions f_i are linear, it is called a *linear program*; otherwise a *non-linear program*. If all f_i are convex functions, it is called a *convex program*. At first glance, linear inequality systems appear to be a very restricted class. However, as pointed out by T. C. Koopmans as early as 1948, linear programs can be used to approximate the broad class of convex functions commonly encountered in economic planning.

Linear programs may be viewed as a generalization of the Leontief Input–Output Model, one important difference being that alternative production processes (activities) are allowed to compete; another being the representation of capacity as an input that becomes available at a later point in time as an output (possibly depreciated). Solving a model with alternative activities requires not only software for efficiently solving on computers large systems of equations as in the Leontief case, but also software for selecting the best combination from an astronomical number of possible combinations of activities. (See the entry simplex method for solving linear programs.)

Formulating a linear program

Finding an optimal product mix (for example blend of gasoline, or metals, or mix of nuts, or animal feeds) is a typical application. For example, a manufacturer wishes to purchase at minimum total cost a number of solder alloys *A*, *B*, *C*, *D* which are available in the market-place in order to melt them down to make a blend of 30 per cent lead, 30 per cent zinc, and 40 per cent tin. Their respective costs per pound are shown in Table 1.

	Alloy				
Composition	A	B	C	D	Desired blend
% Lead	10	10	40	60	30%
% Zinc	10	30	50	30	30%
% Tin	80	60	10	10	40%
Cost/lb	4.1	4.3	5.8	6.0	Minimize cost per pound

Suppose 100 pounds of blend is desired and X_A, X_B, X_C, X_D are the unknown number of pounds of *A*, *B*, *C*, *D* to be purchased. The problem to be solved is clearly: find *Z* and $(X_A, X_B, X_C, X_D) \geq 0$, such that:

$$\begin{aligned} 0.1X_A + 0.1X_B + 0.4X_C + 0.6X_D &= 30 \\ 0.1X_A + 0.3X_B + 0.5X_C + 0.3X_D &= 30 \\ 0.8X_A + 0.6X_B + 0.1X_C + 0.1X_D &= 40 \\ \hline 4.1X_A + 4.3X_B + 5.8X_C + 6.0X_D &= Z(\min) \end{aligned}$$

This example can be solved in a few seconds on a personal computer.
The standard form of a linear program is: find min *z*, $x = (x_1, \ldots, x_n) \geq 0$:

$$Ax = b,\ cx = z(\min)$$

where *A* is a *m* by *n* matrix, *b* a column vector of *m* components and *c* a row vector of *n* components. The matrix *A* of coefficients is referred to as the *technology matrix*.

One way to formulate a linear program is to begin by (*a*) listing various constraints such as resources availability, demand for various goods by consumers, known bounds on productive capacity; (*b*) listing variables to be determined representing the levels of activities whose net inputs and outputs must satisfy the constraints, and finally (*c*) tabulating the coefficients of the various inequalities and equations.

Since linear programming models can be very large systems with thousands of inequalities and variables, it is necessary to use a special software, called *matrix generators*, to facilitate the model building process. Such systems have millions of coefficients, fortunately most of them are zero. Matrices with very few nonzero elements are called *sparse*. The World Bank uses software called GAMS to generate moderate-size sparse matrices *A* by rows. Another type of software called OMNI has been developed by Haverly Systems and has been used to generate very large sparse matrices by columns. When a model is formulated by columns, it is called *Activity Analysis*: the column of coefficients of a variable is the same as a recipe in a cook book – these are the input and output flows required to carry out one unit of an activity (or process). The variables, usually non-negative, are the unknown levels of activity to be determined. For example the activity of 'putting one unit (pound) of solder alloy *A* in the blend' has an input of $4.10 and outputs to the blend of 0.10•lb of lead, 0.10•lb of zinc, 0.80•lb of tin.

In economic applications, *output* coefficients are typically stated with + signs and *input* coefficients with – signs. Under this convention, the signs of the coefficients of the *Z* equation in the blending example should be reversed and, net revenues, (–*Z*) maximized. In practice, instead of equations in non-negative variables, there can be a mix of equations and inequalities. Simple algebraic steps allow one to pass from one form of the system to another.

Primal and dual statements of the linear program

John von Neumann in 1947 was the first to point out that associated with a linear program is another called its *Dual*, formed by transposing the matrix *A* and interchanging the role of the RHS *b* and the 'cost' vector *c*. The original problem is called the *Primal*. Von Neumann expressed both of these LP in inequality form:

$$\text{Primal}: \min\ \bar{z} = cX: AX \geq b,\ X \geq 0,\ \text{(P)}\quad \text{Dual}: \min\ \underline{z} = Y'b: Y'A \leq c,\ Y \geq 0,\ \text{(D)}$$

where Y' is the transpose of column vector *Y*.

If we denote the *j*th column of *A* by *A* (*,•*j*), the *n* inequalities $Y'A \leq c$ may be rewritten as $Y'A(*, j) \leq c_j$ for $j = 1, \ldots, n$.

(P) expresses the physical constraints of the system under study. The variables *Y* of the dual (D) can be interpreted as *prices*. Mathematicians call them *Lagrange multipliers*. The dual conditions, $Y'A(*, j) - c_j \leq 0$ for $j = 1, \ldots, n$, may appear strange and just the opposite from what one would expect. They state that levels X_j of all activities *j* that show profit in the economy will rise to the point that all 'price out' nonprofitable. It turns out that when the value $\underline{z} = Y'b$ in (D) is maximized, all activities *j* that are operated at *positive* levels will just *break even*, i.e., just 'clear their books' and that all activities *j* operating at a strict loss will be operating at zero levels.

The famous duality theorem of von Neumann states that, when there exist 'feasible' solutions $AX \geq b$, $X \geq 0$ to (P) and $Y'A \leq c$, $Y \geq 0$ to (D),

$$\text{Max}\ \underline{z} = \text{Min}\ \bar{z}$$

It is easy to prove that *any* feasible solutions to (P) and (D) not necessarily optimum satisfy

$$\underline{z} = Y'b \leq cX = \bar{z},$$

so that if it happens that $Y^*b = cX^*$, for some feasible $X = X^*$, $Y = Y^*$ then by the duality theorem we know that such a pair (X^*, Y^*) are optimal solutions to (P) and (D).
This makes it possible to combine the primal and dual problems into the single problem of finding a feasible solution to the following: find $(X, \bar{X}, Y, \bar{Y}, \theta) \geq 0$:

$$\begin{bmatrix} 0 & A & -b \\ -A' & 0 & c' \\ b' & -c & 0 \end{bmatrix} \begin{bmatrix} Y \\ X \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \bar{X} \\ \theta \end{bmatrix}$$

(P,D)

where we have introduced two *slack* vectors $\bar{Y} \geq 0$ and $\bar{X} \geq 0$ which turn the inequality relations (P) and (D) into equality relations $AX - \bar{Y} = b$ and $Y'A + \bar{X}' = c$. The last relation is the single equation $Y'b - cX = \theta$ where $\theta \geq 0$ is a scalar.
It we multiply (P,•D) by the vector $(Y', X', 1)$ on the left and perform all the matrix multiplications, everything on the left side cancels out because of the skew symmetry of the matrix and we are left with

$$0 = [Y'X'1][\bar{Y}\bar{X}\theta]' = \sum_i Y_i\bar{Y} + \sum_j X_j\bar{X}_j + \theta.$$

Because all terms are non-negative, it follows that

$$X_j\bar{X}_j = 0,\ Y_i\bar{Y}_i = 0,\ \theta = 0 \text{for all } i \text{ and } j.$$

These are called *complementary slackness* or *Kuhn–Tucker* conditions for optimality.

Zero-sum matrix games

These games can be formulated as a special class of linear programs. The ‘row’ player chooses row *i* of a matrix while his opponent, the ‘column’ player, simultaneously chooses column *j*. Column player wins an amount a_{ij} if $a_{ij} \geq 0$ otherwise he pays the other player $-a_{ij}$. The *payoff* matrix is $A = [a_{ij}]$. It is called a *zero-sum* game because the sum of the payments each player receives adds up to zero. Von Neumann analysed this game in 1928 and introduced the notion of a *mixed strategy* $(Y_1, Y_2, \ldots, Y_m)$, $(X_1, X_2, \ldots, X_n)$ which are the probabilities of the players choosing any particular row and column. He showed that there exist optimal mixed strategies, $Y = Y^*$ for the row player and $X = X^*$ for the column player, such that if a player's mixed strategy is discovered by his opponent, it will have no effect on his expected payoff and hence no effect on the expected payoff of his opponent which is the negative of his.
The column player, if he plays conservatively and assumes his mixed strategy will become known to his opponent, will choose his probabilities $X_j \geq 0$ so as to maximize *L* where Max *L* and $X \geq 0$ are chosen so that

$$\sum_j a_{ij} X_j \geq L,\ X_j \geq 0,\ i = 1, \ldots, m \sum_j X_j = 1.$$

(C)

Likewise the row player's-optimal mixed strategy, if he plays conservatively, will choose his probabilities $Y_i \geq 0$ so as to minimize *K* where Min *K* and $Y \geq 0$ are chosen so that

$$\sum_i Y_i a_{ij} \leq K,\ Y_i \geq 0,\ j = 1, \ldots, n \sum_i Y_i = 1.$$

(R)

It is not difficult to prove that (C) and (R) are feasible linear programs and each is the dual of the other. Let $(Y_1^*, \ldots, Y_m^*)$ and $(X_1^*, \ldots, X_m^*)$ be optimal solutions to (R) and (C). Applying the duality theorem, we obtain von Neumann's famous *mini-max theorem* for zero-sum bimatrix games:

$$\max L = \min K = \sum_i \sum_j Y_i^* a_{ij} X_j^*,$$

the expected payoff to the column player.

Decomposition principle

Linear programming can be used in an iterative mode to aid a Central Authority to allocate scarce resources to factories in an optimal way without having to have detailed knowledge about each factory. Specifically the Central Authority proposes prices on the scarce commodities that induce the factories to submit a summary plan for approval of their requirement for scarce resources. The Central Authority blends these proposed plans with earlier ones submitted and uses them to generate new proposed prices. The entire cycle is then iterated. This method, first proposed by Dantzig and Wolfe in 1960, is known as the D–W or *Primal Decomposition Principle*.

The dual form of the Decomposition Principle is known as Benders Decomposition and was proposed by Benders in 1962. We illustrate it here in the context of a two-period planning problem.

$$\text{find min } Z = c_1X_1 + c_2X_2 \text{ subject to: } b_1 = A_1X_1 \quad (X_1, X_2) \geq 0, b_2 = -B_1X_1 + A_2X_2$$

where A_1, B_1, A_2 are matrices, b_1, b_2, c_1, c_2 vectors and $X_t \geq 0$ are the vectors of activity levels to be determined in periods $t = 1$ and 2.

The first period planners determine a feasible plan (p) that satisfies $b = AX_l^p, X_l^p \geq 0$ (augmented by certain necessary conditions, called ‘cuts’), which they submit to the second period planners in the form of a vector $B_1X_1^p$ which is used by them to solve the second period *sub* problem:

$$A_2X_2 = b_2 + B_1X_1^p, X_2 \geq 0 c_2X_2 = Z_2(\text{min}).$$

The second period planners respond with a vector of *optimality prices* π_2^k corresponding to the second period if the sub-problem is feasible, or with *infeasibility prices* σ_2^l (obtained at the end of phase 1 of the simplex method) if it is infeasible.

The first period planners then iteratively resolve, their problem augmented by $k' + l'$ additional necessary conditions (cuts) shown below:

$$\text{Find } c_1X_1 + \theta = Z(\text{min}) A_1X_1 = b_1, X_1 \geq 0, \text{optimality cuts: } -(\pi_2^kB_1)X_1 + \theta \geq \pi_2^kb_2, k = 1, \ldots, k' \text{infeasibility cuts: } -(\sigma_2^lB_1)X_1 \geq \sigma_2^lb \quad l = 1, \ldots, l'$$

where $\theta = (c_2X_2)$ is treated as an unknown variable. The interative process stops if $\theta = Z_2$, or $Z_2 - \theta = \Delta > 0$ is small enough.

Note that the additional conditions imposed on Period 1 are expressed in terms of Period-1 variables and θ only. These serve as surrogates for future

periods (in this example for only one future period). The decomposition principle allows one to solve a multi-time-period problem one period at time and pass the ending conditions of one period on to initiate the next and to pass back price vectors to earlier periods that are translated into policy constraints called cuts. Applying this same approach to a multi-stage production line, one obtains an iterative process that can be viewed as an intelligent control system with learning.

See Also

- efficient allocation
- nonlinear programming
- simplex method for solving linear programs

Bibliography

Beale, E.M.L. 1954. Linear programming by the method of leading variables. Report of the Conference on Linear Programming, May, arranged by Ferranti Ltd, London.

Benders, J.F. 1962. Partitioning procedures for solving mixed-variable programming problems. *Numerische Mathematik* 4, 238–52.

Charnes, A., Cooper, W.W. and Mellon, B. 1952. Blending aviation gasolines – a study in programming interdependent activities in an integrated oil company. *Econometrica* 20(2), 135–59.

Dantzig, G.B. 1948. Programming in a linear structure. Comptroller, USAF, Washington, DC, February.

Dantzig, G. 1949, 1951. Programming of interdependent activities, II, mathematical model. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley, 330–35; also published in *Econometrica* 17(3 and 4), 1949, 200–211.

Dantzig, G.B. 1963. *Linear Programming and Extensions*. Princeton: Princeton University Press.

Dantzig, G.B. and Wolfe, P. 1960. A decomposition principle for linear programs. *Operations Research* 8(1), 101–11.

Kantorovich, L.V. 1939. *Mathematical Methods in the Organization and Planning of Production*. Publication House of the Leningrad State University. Translated in *Management Science* 6 (1960), 366–422.

Koopmans, T.C. (ed.) 1951. *Activity Analysis of Production and Allocation*. New York: John Wiley.

Kuhn, H.W. and Tucker, A.W. 1951. [The symposium was held in 1950, but the proceedings volume was published in 1951.] Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. Neyman. Berkeley: University of California Press, 481–92; also in *Econometrica* 19(1) (1951), 50–51 (abstract).

Leontief, W. 1951. *The Structure of the American Economy, 1919–1939*. New York: Oxford University Press.

Vajda, S. 1956. *The Theory of Games and Linear Programming*. New York: John Wiley.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320. Translated by Sonya Bargmann in *Contributions to the Theory of Games*, vol. 4, ed. A.W. Koopmans and R.D. Luce, Annals of Mathematics Study No 40, Princeton, NJ: Princeton University Press, 1959, pp. 13–42.

von Neumann, J. 1937. *Über ein ökonomisches Gleichungssytem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes*. *Ergebnisse eines Mathematischen Kolloquiums* No. 8. Translated in *Review of Economic Studies* 13(1), (1945–46) 1–9.

von Neumann, J. 1947. Discussion of a maximization problem. Manuscript, Institute for Advanced Study, Princeton, New Jersey, November.

von Neumann, J. *Collected Works*, Vol.VI, ed. A.H. Taub. Oxford: Pergamon Press. 1963, pp. 89–95.

Wood, M.K. and Dantzig, G.B. 1949, 1951. The programming of interdependent activities: general discussion. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley, 1951, 15–81; also in *Econometrica* 17(3 and 4) (1949), 193–9.

How to cite this article

Dantzig, George B. "linear programming." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000106> doi:10.1057/9780230226203.0975

The New Palgrave Dictionary of Economics Online

linkages

Antonio Ciccone

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic activities in different industries are linked to each other through aggregate income (horizontal linkages) and input–output relationships (vertical linkages). Could such linkages give rise to vicious circles of underdevelopment or virtuous circles of development when there are increasing returns to scale at the firm level?

Keywords

cost linkages; demand linkages; horizontal and vertical linkages; increasing returns to scale; industrialization; input chains; linkages; multiple equilibria; pre-industrial production methods; underdevelopment traps

Article

1 Introduction

Economic activities in different industries are linked to each other through aggregate income (horizontal linkages) and input–output relationships (vertical linkages). Could such linkages give rise to vicious circles of underdevelopment or virtuous circles of development when there are increasing returns to scale at the firm level? A standard account of a vicious circle goes as follows. Small-scale production methods in industry *A* lead to low output and income. This translates into low demand for industry *B*, which therefore also ends up using small-scale production methods and generating low output and income. The result is low demand for industry *A*, which justifies the small-scale production methods used in this industry. Low aggregate output and income are seen as the result of a vicious circle because the same economic environment is thought to be compatible with a high-income equilibrium where all industries use technologies that achieve high productivity at large scale. This high-income equilibrium is sustained by a virtuous circle. Large-scale production methods in industry *A* are profitable because of

high income in industry *B*, and vice versa.

We will show that vicious or virtuous circles based on demand linkages are subject to a simple fallacy if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. Still, vertical demand linkages will give rise to vicious or virtuous circles if increasing-returns-to-scale technologies use intermediate inputs more intensively than the technologies they replace. And horizontal demand linkages will do so if firms adopting increasing-returns-to-scale technologies must pay a compensating wage differential. Moreover, when there are both vertical demand and cost linkages, underdevelopment traps can be consistent with economic principles even if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. We first discuss the role of horizontal demand linkages, then that of vertical demand linkages, and finally turn to vertical cost linkages.

Horizontal demand linkages. Imagine an economy populated by households and by firms in different industries. Suppose that each industry sells only to households. Assume also that the amount households spend on each industry is independent of prices (industry demand functions are unit elastic). In this case, demand linkages among industries are said to be *horizontal*. This simply means that economic activity in one industry affects spending on other industries only through the aggregate income of households. Could horizontal demand linkages lead to economies being trapped into a situation of low income due to a vicious circle of low income and output? Rosenstein-Rodan (1943) and Nurkse (1953) thought so. They imagined a situation where low aggregate income was an obstacle to the adoption of technologies that achieve high productivity at large scale. But large-scale production methods would be profitable if all industries adopted them, because incomes generated in one industry would create demand for other industries.

The elements necessary for underdevelopment traps to be consistent with economic principles have always been subject to debate. Increasing returns to scale appeared to be crucial. But Fleming (1955) made clear that this was not enough. He imagined a situation where, because of low aggregate income, industry *A* cannot make a profit from adopting the increasing-returns-to-scale technology and that the same is true for industry *B*. Is it possible that the increasing-returns-to-scale technology becomes profitable if both *A* and *B* adopt it? Consider forcing *A* to adopt. In this case, the loss made in industry *A* will lower aggregate income. As a result, industry *B* will now face even lower demand and therefore make an even greater loss if it adopts the increasing-returns-to-scale technology. This means that aggregate income will fall further if we also force industry *B* to adopt the increasing-returns-to-scale technology. Hence, if the adoption of increasing-returns-to-scale technologies is unprofitable for any single industry, adoption in all industries will not be profitable either. Increasing returns alone can therefore not explain why industrialization does not take place although it would ultimately be profitable. All accounts of underdevelopment traps did in fact feature (several) additional elements. In particular, Rosenstein-Rodan maintained that firms using large-scale production methods had to pay a compensating wage differential (partly because of the higher costs of living in urban areas, where industrial firms were located). Section 2 follows Murphy, Shleifer and Vishny (1989) in showing that underdevelopment traps may emerge when firms adopting the increasing-returns-to-scale technologies must pay a compensating wage premium.

Vertical demand linkages. Suppose now that industries sell goods to households and each other (to be used as intermediate inputs). Economic activity in one industry can then affect demand in another industry even if aggregate income remains unchanged. As a result, there are said to be *vertical* linkages.

For example, consider the situation where industry *B* buys from *A* (industry *A* is *upstream* of *B*). In this case there is a *vertical demand* linkage as demand for the upstream industry *A* will depend on the economic activity in downstream industry *B*. There could also be a *vertical cost* linkage because the cost of production in downstream industry *B* is partly determined by the cost of goods produced in upstream industry *A*.

While the effects of horizontal demand linkages on economic development have always been subject to some controversy, there appears to be a consensus among early contributors that vertical demand linkages can lead to underdevelopment traps when technologies are subject to increasing returns to scale (Fleming, 1955; Scitovsky, 1954; Hirschman, 1958). It is simple to show however that this is not the case if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. To see this, note that with vertical demand linkages the adoption of increasing-returns-to-scale technologies affects aggregate income directly and indirectly: directly through the profits made in the adopting industry, and indirectly through the profits made in supplying (upstream) industries. It would therefore seem that increasing-returns-to-scale technologies could be unprofitable in the adopting industry but still increase aggregate income. But this cannot happen when the increasing-returns-to-scale and the pre-industrial technologies use upstream inputs with the same intensity. In this case, the increase in the value of upstream goods demanded by a firm adopting increasing-returns-to-scale technologies is always a fraction of the (absolute value of the) loss that it makes. Moreover, as profits cannot exceed revenues, the increase in profits in supplying industries is necessarily smaller than the increase in the value of goods they sell. It therefore follows that the increase in profits in supplying industries (the positive indirect effect) can never compensate for the loss made in the industry adopting the increasing-returns-to-scale technology.

The empirical evidence indicates that the intermediate-input intensity of production increases with a country's level of industrialization. Increasing-returns-to-scale technologies may therefore be using intermediate inputs more intensively than the production methods they replace. Section 3 draws on Ciccone's (2002) model of input chains to show that vertical linkages can in this case explain why countries may be trapped into a vicious circle of underdevelopment, and why escaping this trap may be associated with large gains in aggregate income and productivity.

The interplay of vertical cost and demand linkages. The greater demand for intermediate inputs brought about by industrialization (vertical demand linkages) may partly be caused by falling intermediate input prices (vertical cost linkages). Falling intermediate input prices, on the other hand, are possible because of the higher productivity of large-scale production methods. Vertical cost and demand linkages therefore feed on each other (Young, 1928; Okuno-Fujiwara, 1988; Rodriguez-Clare, 1996). For example, Rodriguez-Clare considers a small open economy framework where the entry of new intermediate input varieties lowers the cost of intermediate inputs relative to labour, which leads final-good producers to substitute towards intermediate inputs. When this substitution effect is strong enough, it translates into greater revenues and profits for intermediate-input producers, which may validate intermediate-input producers’ decision to start up new varieties in the first place. Rodriguez-Clare shows that this interplay of vertical demand and cost linkages may lead to two equilibria: a low-income equilibrium where final-good producers use labour-intensive production methods because of the limited range of intermediate inputs available, and a high-income equilibrium where a large variety of intermediate inputs leads final-good producers to use intermediate-input intensive production methods. Okuno-Fujiwara (1988) considers a situation where vertical demand and cost linkages interact because

greater demand for intermediate inputs leads to lower prices due to competition among a larger number of Cournot oligopolists. The final section of this entry uses the model with input chains to show that the interplay between vertical demand and cost linkages can result in underdevelopment traps even if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale.

2 A model of horizontal demand linkages

We will now examine the role of horizontal demand linkages for economic development using the model of Murphy, Shleifer, and Vishny (1989) (for a historical and methodological perspective on the horizontal-linkages literature, see Krugman, 1993; 1994). The first step is to describe the model set-up – the household sector, the production sector, and market structure. The second step is to characterize equilibrium prices and equilibrium allocations.

Households. There are *L* households, each of whom supplies one unit of labour inelastically (labour is the only production factor in this model and serves as the numeraire). Households spend an equal share of their incomes on each of the *N* goods produced in the economy.

Production. Each of the *N* goods demanded by households can be produced using two different production methods: a *pre-industrial* method requiring one unit of labour for each unit of output produced, and an *industrial* or increasing-returns-to-scale method, which is more efficient at the margin but subject to a fixed labour requirement (f). Formally, the increasing-returns-to-scale production method requires

$$l_i = f + cq_i \tag{1}$$

units of labour to produce q_i units of good *i*, where $f > 0$ and $1 > c > 0$.

Industry wage premium. Working in the industrial sector generates a disutility $v \geq 0$ for households. Hence, relative to pre-industrial firms, industrial firms will have to pay a wage premium $v \geq 0$ as a compensating wage differential.

Market structure. Many firms are assumed to know the pre-industrial method to produce good *i*. As a result, the pre-industrial sector (also called competitive fringe) will be characterized by perfect competition. By contrast, only a single firm is taken to have the ability to produce each good in the industrial sector. These firms set prices optimally, taking the prices of all other firms as given. The labour market is taken to be perfectly competitive.

What keeps this model simple to analyse is that the equilibrium price of each good is unity whether the good is produced by the pre-industrial or the industrial sector. To see this, note that perfect competition and constant returns to scale in the pre-industrial sector imply that the price of goods produced in this sector must be equal to unity. A higher price would mean strictly positive profits and therefore further entry of pre-industrial producers, while a lower price would mean that no pre-industrial producer could break even. Now consider goods produced in the industrial sector. Clearly, the industrial producer will

not set a price above unity, as she would lose the entire market to pre-industrial producers in this case. Moreover, industrial producers do not have an incentive to set a price below unity either, as households spend the same fraction of income on their good irrespectively of the price. Hence, industrial producers find it optimal to use a limit pricing strategy, setting prices exactly equal to the marginal cost of pre-industrial producers. As a result, the price of each of the *N* goods is equal to unity independently of the production method.

Pre-industrial equilibrium. Under what conditions will there be an equilibrium where all goods are produced with the pre-industrial method? In such an equilibrium, firms just break even, and aggregate income *Y* in the economy is therefore equal to aggregate labour income *L*. Because households spread income equally among all *N* goods, the quantity of good *i* demanded and supplied is $q_i = L / N$. The remaining question is whether firms in the industrial sector have an incentive to adopt the increasing-returns-to-scale method. The potential profit of such firms is $\pi_i = q_i^m - (f + cq_i^m)(1 + v)$, where q_i^m is the demand faced by the industrial producer of good *i*. As industrial and pre-industrial producers set the same price, the first industrial producer faces exactly the same demand as the pre-industrial producers she replaces, $q_i^m = L / N$. Her profits are therefore

$$\pi_i = L / N - (f + cL / N)(1 + v). \tag{2}$$

If $\pi_i < 0$, an industrial producer has no incentive to adopt the increasing-returns-to-scale method, and it will be an equilibrium for all goods to be produced with the pre-industrial method. Hence, (2) implies that there is an equilibrium where all goods are produced with the pre-industrial method if

$$L(1 - c(1 + v)) < F(1 + v), \tag{3}$$

where $F \equiv fN$.

Industrial equilibrium. What about equilibria where all goods are produced using the industrial method? We already know that prices of all goods will be equal to unity in this equilibrium also. Moreover, households will keep spending the same share of income on all goods. Hence, all industries will employ the same amount of labour, *L*/*N*, in equilibrium. (1) therefore implies that the value of production in each industry is $(L / N - f) / c$. Summing across the *N* industries in the economy yields a value for gross domestic product, and hence aggregate household income, of $Y = (L - F) / c$ (recall that $F \equiv fN$).

Do firms make the profit necessary to sustain the industrial production method when all production takes place in the industrial sector? Profits of firms in the industrial sector are $\pi_i = q_i^m - (f + cq_i^m)(1 + v) \geq 0$, where q_i^m is the demand faced by the industrial producer of good *i*,

$q_i^m = Y / N = (L - F) / cN$. Hence, there will be an equilibrium where firms using the increasing-returns-to-scale method make a profit if

$$L(1 - c(1 + v)) \geq F. \quad (4)$$

Efficient allocation. When is the adoption of increasing-returns-to-scale technologies efficient? The aggregate value of production is $Y = (L - F) / c$ when industrial production methods are used and $Y = L$ with pre-industrial methods. The amount of goods necessary to pay the compensating wage differential when all workers are employed in the industrial sector is vL. Hence aggregate welfare will be higher with industrial production methods if and only if $(L - F) / c - vL \geq L$, or

$$L(1 - c(1 + v)) \geq F. \quad (5)$$

Note that (4) and (5) coincide. Hence, an industrial equilibrium exists if and only if it is efficient.
Multiple equilibria and underdevelopment traps. Only one of the two inequalities in (3) and (4) can hold if there is no industry wage premium ($v = 0$). Hence, the equilibrium is unique in this case and, as a result, there cannot be development traps. Moreover, because an industrial equilibrium exists if and only if it is efficient, economies in a pre-industrial equilibrium actually do the best they can given the economic environment.
But when there is an industry wage premium ($v > 0$) there may be multiple equilibria as the inequalities in (3) and (4) can both be satisfied. When this is the case, economies may be stuck in a pre-industrial equilibrium, although the same economic environment would be compatible with an (efficient) industrial equilibrium. To understand why, suppose the economy is in a pre-industrial equilibrium when we force an industry to adopt the increasing-returns-to-scale technology. If (3) holds, then the adopting firm will make a loss. Still, its contribution to aggregate income is strictly positive. To see this, note that demand for this industry is *L*/*N*, and that this is also the amount of labour required to produce the amount of goods demanded using the pre-industrial production methods. Production with the increasing-returns-to-scale technology requires $cL / N + f$ units of labour, which is strictly smaller than *L*/*N* if (4) holds. Hence, the adoption of the increasing-returns-to-scale technology saves labour in the adopting industry, and therefore increases aggregate output and income. This increases demand faced by other industries and therefore raises the profitability of further adoption of the increasing-returns-to-scale technology. Eventually, industrialization raises aggregate income enough for increasing-returns-to-scale industries to break even. Hence, the industrial equilibrium can be seen as the result of a virtuous circle. The adoption of increasing-returns-to-scale technologies raises aggregate income and therefore the profitability of adopting increasing-returns-to-scale technologies. At the same time, the economic environment also

allows for a development trap where low aggregate income is both the cause and the consequence of the failure to adopt increasing-returns-to-scale technologies.

3 Vertical demand linkages in an input chain model

The economic activity of different industries is linked to each other because the output of some industries is used as input in other industries. Can such vertical linkages give rise to vicious circles of underdevelopment or virtuous circles of development when there are increasing returns at the firm level? We will show that – just as for horizontal linkages – this cannot happen if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. Chenery, Robinson and Syrquin's (1986) comparative study of industrialization shows, however, that the industrialization of countries has typically been accompanied by an increase in the intermediate-input intensity of production. This suggests that industrial technologies may use intermediate inputs more intensively than the technologies they replace. We will therefore start by analysing a model of development where increasing-returns-to-scale technologies use intermediate inputs more intensively than pre-industrial technologies.

It will be useful to analyze the consequences of vertical linkages for industrialization in a framework that is as close as possible to the model of horizontal linkages of Murphy, Shleifer and Vishny. In particular, the aggregate amount of labour supplied by households continues to be *L* and households spend an equal share of their incomes on each of the *N* goods produced in the economy. On the production side, we continue to assume that each good can be produced using two different production methods, namely, a pre-industrial method and an industrial (increasing-returns-to-scale) method. The pre-industrial method requires one unit of labour for each unit of output. The increasing-returns-to-scale method will turn out to be cheaper at the margin but subject to a fixed labour requirement *f*. Many firms know the pre-industrial method, but for each good there is only a single firm with the ability to produce in the industrial sector.

Input chains and industrial production. The key difference with the horizontal linkages model is that now the increasing-returns-to-scale method is taken to be more intermediate-input intensive than the pre-industrial method. One way to model the intermediate-input structure of the economy is to think of goods being produced in *S* different locations along a river. Each location produces *H* different goods (the total number of goods is $N \equiv HS$). Goods at location 1 are produced using labour only. Goods at any location $s > 1$, on the other hand, are produced using all goods at location $s - 1$. This implies that all goods at locations $s < S$ may face intermediate-input demand from downstream industries in addition to consumption-goods demand from households (the exception are the *H* goods furthest downstream, at location *S*, which face consumption-goods demand only). In particular, we assume that, after having incurred the overhead labour cost, one unit of any good *j* located at $s > 1$ can be produced with *c* units of an intermediate-input composite $z_{j,s}$ that combines all *H* goods produced at location $s - 1$,

$$z_{j,s} = \prod_{i=1}^{H} \left(Hq_{i,s-1}\right)^{1/H}, \tag{6}$$

where $q_{i,s-1}$ is the input of good i at location $s-1$. This formulation implies that industrial firms spend the same amount on all upstream inputs. As a result, the marginal cost of the intermediate-input composite necessary for industrial production at location $s > 1$ is simply a geometrically weighted average of prices $p_{i,s-1}$ of the *H* upstream goods,

$$MC_s = \prod_{i=1}^{H} p_{i,s-1}^{\frac{1}{H}}. \tag{7}$$

Industrial production for goods at location $s = 1$ requires *f* units of overhead labour and *c* units of labour for each unit of output. (The assumption that the industrial overhead requires labour only while production at the margin requires intermediate inputs only simplifies the analysis considerably. Ciccone (2002) analyses the case where production of the overhead and at the margin use both labour and intermediate inputs.)

Just as in the horizontal linkages model, industrial firms find it optimal to use a limit pricing strategy for consumption goods vis-à-vis the competitive fringe. Their intermediate-input pricing strategy is potentially more complicated but also simplifies to a limit pricing strategy vis-à-vis the competitive fringe when *H* is sufficiently large.

Pre-industrial equilibrium. When will there be an equilibrium where all goods are produced with the pre-industrial method? It turns out that if *H* is sufficiently large the condition is

$$L(1-c) < F, \tag{8}$$

which coincides with the condition for a pre-industrial equilibrium in the Murphy, Shleifer, and Vishny model of horizontal linkages. To see this, suppose that all goods are produced with the pre-industrial technology and their price is unity. When (8) holds, any single firm adopting the increasing-returns-to-scale method to produce consumption goods will make a loss. Moreover, when *H* is sufficiently large, (7) also implies that single industrial firms are unable to generate intermediate-input demand for their good even if they lower their price to the marginal cost of production. To see this, suppose that one industrial firm at location $s-1$ is considering selling its good at marginal cost to firms at location s in order to generate intermediate-input demand. In this case, one of the *H* inputs of potential industrial firms at s would become available at price *c* and (7) implies that the marginal cost of production would

therefore fall from c to $c^{(1+H)/H}$(recall that the remaining $H-1$ inputs are available at price of unity). Goods at S face demand L/N, which comes exclusively from households as there are no upstream industries. Hence, profits of the potential industrial firm at *S* producing at marginal cost $c^{(1+H)/H}$ would be $(1-c^{(1+H)/H})L/N-f$, which is strictly negative if (8) holds and *H* is large enough. Potential industrial firms at location *S* would therefore find it unprofitable to start production even after the price cut, which implies that potential industrial firms at location $S-1$ must break even on consumption-goods demand only. Applying the same argument sequentially to potential industrial firms in locations $S-2, S-3$, …, 1 yields that pre-industrial production of all goods is an equilibrium when (7) holds and *H* is sufficiently large.

Industrial equilibrium. To determine the conditions for the existence of an industrial equilibrium, it is necessary to determine aggregate income when all goods at location σ and upstream of location σ are produced with the increasing-returns-to-scale technology. This turns out to be straightforward. If aggregate income is Y, the quantity of each good demanded by households is Y/N. The intermediate-input structure implies that industrial production of Y/N units of each of the *H* goods at location σ requires cY/N units of each of the *H* goods at location $\sigma-1$. Hence, as Y/N units of good $\sigma-1$ are demanded by households, production of each good at $\sigma-1$ must be $Y/N+cY/N$. Production of this quantity of goods at $\sigma-1$ requires $c(Y/N+cY/N)$ units of each good at $\sigma-2$. Adding the Y/N units of goods at $\sigma-2$ demanded by households, yields that production at $\sigma-2$ must be $Y/N+cY/N+c^2Y/N$. Continuing all the way upstream yields that the total production of each of the *H* goods at location 1 must be

$$q_1 = Y/N + cY/N + c^2Y/N + \ldots + c^{\sigma-1}Y/N = \frac{1-c^{\sigma}}{1-c}Y/N. \tag{9}$$

To turn to the labour market, *f* units of labour must be used as overhead in the production of each good produced with the industrial technology. Moreover, Y/N units of labour are required for the production of each good produced with the pre-industrial technology. Hence, the amount of labour available for production at the margin of the *H* goods at $s=1$ is $L-\sigma Hf-(N-\sigma H)Y/N$. Labour market clearing requires $cHq_1 = L-\sigma Hf-(N-\sigma H)Y/N$. Substituting (9) yields aggregate income in an economy where the σ industries furthest upstream have industrialized:

$$Y(\sigma) = \frac{L-F(\sigma H/N)}{c\theta[\sigma](\sigma H/N)+(1-(\sigma H/N))} = \frac{L-F(\sigma H/N)}{1-(\sigma H/N)(1-c\theta[\sigma])}, \tag{10}$$

where

$$\theta[\sigma] \equiv \frac{1 - c^{\sigma}}{(1 - c)\sigma}.$$

$c\theta[\sigma]$ has a simple interpretation. It is the amount of labour required to produce one additional unit of goods located at σ if all industries upstream of (including) σ have adopted the industrial technology. Note that the amount of labour required to produce one additional unit of goods at location σ falls the longer the industrial input chain ($\theta[\sigma]$ is strictly decreasing in σ).

The intermediate-input structure implies that the demand for goods is greater the further upstream they are located. Hence, profits from adopting the increasing-returns-to-scale technology fall the further downstream industries are located. An equilibrium where all industrial firms make a profit will therefore exist if goods produced furthest downstream (at location *S*) can be produced using the increasing-returns-to-scale technology without a loss. Because firms furthest downstream sell to households only, their sales are equal to aggregate income divided by the number of goods, $Y[S] / N$ (recall that all firms set prices optimally at unity). As a result, their profits are positive if and only if $\pi_S = (1 - c)(Y[S] / N) - f \geq 0$ or, to make use of (10),

$$(1 - c)L \geq (c\theta[S] + (1 - c))F. \tag{11}$$

Multiple equilibria and underdevelopment traps. Comparison of (8) and (11) yields that, with input chains ($S > 1$), it is possible for the pre-industrial equilibrium and the industrial equilibrium to exist side by side. (When $S = 1$ then $\theta = 1$ and the model is that of Murphy, Shleifer, and Vishny without an industry wage premium.) This is because the adoption of increasing-returns-to-scale technologies now has a direct and indirect effect on income. The direct effect is given by the profit or loss in the adopting industry. The indirect effect is equal to the profits generated upstream of the adopting industry. When the indirect profits generated by the increased intermediate-input demand more than offset direct losses of industrial technologies, then industrialization increases aggregate income. As a result, further industrialization becomes more profitable. When (7) and (10) hold simultaneously, this effect is strong enough to ensure that all industrial firms make a profit once all goods are produced with increasing-returns-to-scale technologies.

The pre-industrial and industrial equilibrium can exist side by side even if aggregate income is much greater in the industrial equilibrium. Note that aggregate income in the industrial equilibrium is $Y[S] = (L - F) / c\theta[S]$, see (10). As intermediate-input chains become longer, $\theta[S]$ in (10) tends to zero, and aggregate income in the industrial equilibrium increases. Aggregate income in the pre-industrial equilibrium, on the other hand, is independent of *S* as production does not rely on intermediate

inputs. Moreover, the range of parameter values for which the industrial equilibrium exists increases. Hence, long input chains imply that equilibrium multiplicity is more likely and also that the aggregate income difference between industrial and pre-industrial equilibria may be very large.

Vertical linkages and equilibrium uniqueness. To see that the equilibrium is unique when increasing-returns-to-scale technologies use intermediate inputs as intensively as pre-industrial technologies, note that costs of production plus profit must add up to the value of firms' sales, $COST + \pi = q$. Suppose that intermediate inputs are a share α of costs of production for both the pre-industrial and the industrial production method. In this case, the demand for goods produced at $s - 1$ is equal to $\alpha COST_s = \alpha(q_s - \pi_s)$. Now suppose that all goods upstream of σ are produced with the increasing-returns-to-scale technology. Is it possible that aggregate income increases with the adoption of the increasing-returns-to-scale technology at σ even if the adopting firm makes a loss? A switch to industrial production at σ does not affect the value of goods produced at this location (q_σ is unchanged). Hence, the adoption of the increasing-returns-to-scale technology at σ increases demand for each good produced at $\sigma - 1$ by $-\alpha\pi_\sigma / H$. Loss-making industrialization at σ therefore leads to greater demand at $\sigma - 1$. But the profits generated by this input demand can never be greater than the initial loss π_σ. To see this, notice that total profits at location $\sigma - 1$ increase by $-(1 - c)\alpha\pi_\sigma$. Total profits at $\sigma - 2$ increase by $-(1 - c)\alpha^2 c\pi_\sigma$, where $-\alpha^2 c\pi_\sigma / H$ is the increase in demand for each good produced at $\sigma - 2$. The general formula is that total profits at location $\sigma - i$ increase by $-(1 - c)\alpha^i c^{i-1}\pi_\sigma$. Summing profits across all locations yields $-(1 - c)\pi_\sigma\alpha\left[1 + \alpha c + (\alpha c)^2 + \ldots + (\alpha c)^{\sigma-1}\right]$, which is smaller than $-(1 - c)\pi_\sigma\alpha\left[1 + \alpha c + (\alpha c)^2 + \ldots\right] = -\pi_\sigma(\alpha - \alpha c) / (1 - \alpha c)$. Hence, $\alpha \leq 1$ implies that the sum of profits generated upstream of *s* by loss-making industrialization at *s* is always smaller than the initial loss (π_σ). Loss-making industrialization necessarily lowers aggregate income. The aggregate demand externality necessary for multiple equilibria is therefore absent when increasing-returns-to-scale technologies are no more intermediate-input intensive than pre-industrial technologies.

4 Vertical demand and cost linkages with input chains

So far firms adopting increasing returns to scale technologies did not have an incentive to cut prices. This eliminated virtuous circles of development where lower intermediate-input prices (vertical cost linkages) and greater intermediate-input demand (vertical demand linkages) feed on each other. A simple way to capture the interplay between vertical demand and cost linkages is to suppose that firms in the competitive fringe can produce one unit of goods at location $s > 1$ with $1 + \varepsilon > 1$ units of the intermediate-input composite in (6) or one unit of labour. That is, firms have access to two modes of production, a labour-intensive mode and an intermediate-input intensive mode. The exception continues to be goods at location 1, for which there is a labour-intensive mode of production only. Industrial firms at locations $s > 1$ also have access to a labour-intensive and an intermediate-input intensive mode of production, but are more efficient than pre-industrial firms at the margin. Once they have incurred the overhead labour requirement *f*, industrial firms can produce one unit of output with $c(1 + \varepsilon) < 1$ of the intermediate-input composite in (6) or $c < 1$ units of labour. Industrial firms producing goods at location 1 have access to the labour-intensive mode of production only. The assumption that the overhead is

produced using labour only continues to simplify the analysis considerably. A new by-product of this assumption is that industrial firms now actually use intermediate inputs less intensively than pre-industrial firms at the same factor prices – the opposite of what we assumed in the previous section.

Pre-industrial equilibrium with labour-intensive production. Can there be an equilibrium where all goods are produced with the pre-industrial technology using labour only? The marginal cost of production with the pre-industrial technology in the labour-intensive mode is unity. Hence, the price of all goods would be equal to unity. To see that these prices make it optimal to use the labour-intensive mode of production, note that they imply that the marginal cost of intermediate-input composites in (7) is unity. The marginal cost of production using the intermediate-input intensive mode compared with the labour-intensive mode is therefore $1 + \varepsilon > 1$ (in the pre-industrial as well as the industrial sector). Hence, all firms will find it optimal to use the labour-intensive mode of production.

In a pre-industrial equilibrium, the adoption of the increasing-returns-to-scale technology by a single firm must lead to losses. If industrial firms can count on consumption-goods demand only, this will be the case if $L(1 - c) < F$. But an industrial firm may be able to generate additional demand by getting industries just downstream to switch to an intermediate-input intensive mode of production. While this can happen in principle, it will not happen if *H* is sufficiently large. To see this, consider the case where a single industrial firm supplies its good to downstream industries at marginal cost. In this case, (7) yields that the marginal cost of the intermediate input-intensive mode of production relative to the labour-intensive mode becomes $c^{1/H}(1 + \varepsilon)$, which will be greater than unity when *H* is sufficiently large (recall that $1 + \varepsilon > 1$). Hence, a single industrial firm cannot generate downstream intermediate-input demand even if it reduces its price to marginal cost. For *H* sufficiently large, a pre-industrial labour-intensive equilibrium will therefore exist if $L(1 - c) < F$.

Industrial equilibrium with intermediate-input intensive production. When is there an industrial equilibrium where all firms use the intermediate-input intensive mode of production? To simplify the analysis, suppose that industrial firms can price discriminate between households and industrial users of their goods. As before, industrial firms will find it optimal to follow a limit pricing strategy when it comes to sales to households. Industrial firms will therefore price consumption goods at unity. When it comes to intermediate-input sales to downstream industries, industrial firms must also take into account that users will switch to the labour-intensive mode of production if the cost of the intermediate-input composite is greater than $1 / (1 + \varepsilon)$. Hence, each industrial firm will find it optimal to set a limit price of $1 / (1 + \varepsilon)$ for intermediate inputs if other industrial intermediate-input suppliers do the same. Aggregate income in the industrial equilibrium where all firms use the intermediate-input intensive mode of production can be determined following the argument that led to (10). The only difference is that an additional unit of all goods at location $s > 1$ now translates into a demand of $c(1 + \varepsilon)$ units of each good at location $s - 1$. Aggregate income when all goods are produced with the industrial technology in the intermediate-input intensive mode is therefore $Y[S] = (L - F) / c\hat{\theta}[S]$ where

$$\hat{\theta}[S] \equiv \frac{1 - (c(1 + \varepsilon))^S}{(1 - c(1 + \varepsilon))S}. \tag{12}$$

An industrial equilibrium exists if the firm furthest downstream can break even given the demand for consumption goods, $\pi_S = (1 - c(1+\varepsilon))(Y[S]/N) - f \geq 0$ or, to make use of the expression for aggregate income just above, $(1 - c(1+\varepsilon))L \geq \left(c\hat{\theta}[S] + (1 - c(1+\varepsilon))\right)F$.

Multiple equilibria with vertical demand and cost linkages. There will be multiple equilibria if both $L(1-c) < F$ and $(1 - c(1+\varepsilon))L \geq \left(c\hat{\theta}[S] + (1 - c(1+\varepsilon))\right)F$. This implies that the pre-industrial equilibrium with labour-intensive production and the industrial equilibrium with intermediate-input intensive production may exist side by side if and only if there are input chains ($\hat{\theta}[S] < 1$). The virtuous circle sustaining industrial equilibria now consists of an interplay between vertical demand and cost linkages. The increase in the intermediate-input intensity of production necessary for increasing-returns-to-scale technologies to be profitable (vertical demand linkages) comes about because the adoption of increasing-returns-to-scale technologies translates into falling intermediate-input prices (vertical cost linkages). Note that, for this virtuous circle to be operative, the elasticity of substitution between intermediate inputs and labour in industrial production must be greater than unity (our model assumed that this elasticity is infinity for simplicity). In a pre-industrial equilibrium, on the other hand, pre-industrial technologies are both the cause and the consequence of labour-intensive modes of production.

5 Conclusion

Neither horizontal nor vertical demand linkages across industries lead to underdevelopment traps if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. Nevertheless, theories of underdevelopment based on vicious circles of low demand and low productivity are consistent with economic principles. For example, in the case of vertical demand linkages, there can be development traps if increasing-returns-to-scale technologies use intermediate inputs more intensively than the technologies they replace. More generally, multiple equilibria in our models exist under assumptions that do not appear to be in contradiction by empirical evidence. The exception is that all our model economies were taken to be closed to international trade, but we could have assumed instead that only some goods are non-tradable or that all goods are tradable at some cost (for example, Okuno-Fujiwara, 1988; Rodriguez-Clare, 1996; Krugman and Venables, 1995). Still, it remains to be seen what part of international income differences can be attributed to development traps (for steps in this direction, see Fafchamps and Helms, 1996; Graham and Temple, 2006).

See Also

- balanced growth
- development economics
- external economies
- externalities

- multiple equilibria in macroeconomics
- new economic geography
- returns to scale
- supermodularity and supermodular games

Bibliography

Chenery, H., Robinson, S. and Syrquin, M. 1986. *Industrialization and Growth: A Comparative Study*. New York: Oxford University Press.

Ciccone, A. 2002. Input chains and industrialization. *Review of Economic Studies* 69, 565–87.

Fafchamps, M. and Helms, B. 1996. Local demand, investment multipliers, and industrialization: theory and application to the Guatemalan highlands. *Journal of Development Economics* 49, 61–92.

Fleming, M. 1955. External economies and the doctrine of balanced growth. *Economic Journal* 65, 241–56.

Graham, B.S. and Temple, J.R.W. 2006. Rich nations, poor nations: how much can multiple equilibria explain? *Journal of Economic Growth* 11, 5–41.

Hirschman, A. 1958. *The Strategy of Economic Development*. New Haven, CT: Yale University Press.

Krugman, P. 1993. Toward a counter-counterrevolution in development theory. In *Proceedings of the World Bank Annual Conference on Development Economics*, ed. L.H. Summers and S. Shah. Washington, DC: World Bank.

Krugman, P. 1994. The fall and rise of development economics. In *Rethinking the Development Experience: Essays Provoked by the Work of Albert O. Hirschman*, ed. L. Rodwin and D.A. Schon. Washington, DC: Brookings Institution.

Krugman, P. and Venables, A.J. 1995. Globalization and the inequality of nations. *Quarterly Journal of Economics* 110, 857–80.

Murphy, K.M., Shleifer, A. and Vishny, R.W. 1989. Industrialization and the big push. *Journal of Political Economy* 97, 1003–26.

Nurkse, R. 1953. *Problems of Capital Formation in Underdeveloped Countries*. New York: Oxford University Press.

Okuno-Fujiwara, M. 1988. Interdependence of industries, coordination failure and strategic promotion

of an industry. *Journal of International Economics* 25, 25–43.

Rodriguez-Clare, A. 1996. The division of labor and economic development. *Journal of Development Economics* 49, 3–32.

Rosenstein-Rodan, P. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53, 202–11.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62, 143–51.

Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

How to cite this article

Ciccone, Antonio. "linkages." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000107> doi:10.1057/9780230226203.0976

The New Palgrave Dictionary of Economics Online

Lintner, John Virgil (1916 – 1983)

J. Fred Weston

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

asset pricing; capital asset pricing model; capital formation under inflation; dividend policy; financial markets theory; Lintner, J. V.; merger analysis

Article

Lintner was born in Lone Elm, Kansas. He received the Ph.D. at Harvard University in 1946, becoming a member of the faculty a year earlier. He remained a member of the Harvard faculty throughout his career and was designated the George Gund Professor of Economics and Business Administration in 1964, with a joint appointment in the Business School and the Faculty of Arts and Sciences in Economics.

The contributions by John Lintner that are most frequently cited in the economic literature involve asset pricing, dividend policy, mergers, and capital formation under inflation. Along with others, Lintner was one of the independent creators of the modern theory of asset pricing. This model is usually referred to as the capital asset pricing model (CAPM) which holds that the equilibrium rates of return on all risky assets are a function of their covariance with the returns on the market portfolio.

In addition to his major contribution to the creation of the modern theory of financial markets, Lintner wrote the seminal articles on dividend policy which provided the foundations for further research and remain the basic references on the subject.

Mergers represented the third area of important contributions. An early study focused on the impact of taxes on mergers (Butters, Lintner and Cary, 1951). One important impact of taxes documented was the sale of companies to convert an earnings stream that would otherwise be subject to personal income tax rates to capital gains which would be taxed at lower rates. His later studies of mergers developed an analysis of the historical influences on mergers during the major merger movements of the United States. In addition, a theoretical rationale for pure conglomerate mergers was also developed (1971).

Many aspects of Lintner's interest in capital formation under inflation were brought together in his Presidential Address to the American Finance Association in December 1974 (1975). His subsequent work sought to develop further the major themes which he had set forth in his Presidential Address.

Selected works

1951. (With J.K. Butters and W.L. Cary.) *Effects of Taxation on Corporate Mergers*. Boston: Harvard Business School.

1956. Distribution of incomes of corporations among dividends, retained earnings and taxes. *American Economic Review* 46(May), 97–113.

1964. Optimal dividends and corporate growth under uncertainty. *Quarterly Journal of Economics* 78 (February), 49–95.

1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47(February), 13–37.

1965. Security prices, risk, and maximal gains from diversification. *Journal of Finance* 20, 587–615.

1969. The aggregation of investors' diverse judgments and preferences in purely competitive security markets. *Journal of Financial and Quantitative Analysis* 4, 347–400.

1971. Expectations, mergers, and equilibrium in purely competitive securities markets. *American Economic Review* 61(2), 101–11.

1975. Presidential address. American Finance Association Annual Meeting, San Francisco, California, 29 December 1974. *Journal of Finance* 30, 259–80.

How to cite this article

Weston, J. Fred. "Lintner, John Virgil (1916–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000108> doi:10.1057/9780230226203.0977

The New Palgrave Dictionary of Economics Online

liquidity constraints

Stephen D. Williamson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Liquidity constraints affect the ability of an economic agent to exchange his or her existing wealth for goods and services or for other assets. These constraints arise because of frictions, including private information, limited commitment, transactions costs, and spatial considerations.

Keywords

cash-in-advance model; consumption smoothing; contingent claims markets; endowments; expected utility; incomplete markets; law of large numbers; limited commitment; liquidity constraints; precautionary savings; private information; risk sharing; search and matching models of monetary exchange

Article

A benchmark model

To explain what liquidity constraints are, and their implications for economic activity, it is useful to start with a simple benchmark model. Suppose a world with a continuum of households having unit mass. Time is indexed by $t = 0, 1, 2, \ldots,$ and household i has preferences given by

$$E_0 \sum_{t=0}^{\infty} \beta^t u(c_{it}),$$

where E_0 is the expectation operator conditional on period 0 information, $0 < \beta < 1$, c_{it} is consumption,

and $u(\cdot)$ is twice continuously differentiable, strictly concave, and has the property that $u'(0) = \infty$. Each household receives a random endowment of the perishable consumption good at the beginning of each period. That is, household *i* receives an endowment y_{it} in period *t* where y_{it} is assumed to be independent and identically distributed across households and over time. Assume that $\underline{y} \leq y_{it} \leq \bar{y}$, where $0 < \underline{y} < \bar{y}$. The law of large numbers then implies that the aggregate endowment is a constant, which we will denote by *y*. Therefore, this is an economy with no aggregate risk, but each household faces idiosyncratic risk associated with its endowment shocks.

Now, suppose that this economy has a complete set of markets. One market structure that gives completeness is contingent claims markets that open at $t = 0$ before households receive their period 0 endowments. All households trade on these markets, and a particular contingent claims market involves trade in claims to the consumption good deliverable at a particular date only under a particular realization for the path of endowment shocks for all households up to that date. Given this complete set of markets, what will be the equilibrium allocation of consumption across households at each date? All households are identical at the first date, and the result will be that, in equilibrium, $c_{it} = y$ for all *i* and *t*. The complete set of contingent claims markets provides perfect insurance for households. That is, they are able to share their risk efficiently, in that each household can shed the idiosyncratic risk associated with its endowment shocks. Indeed, the resulting equilibrium allocation of consumption is Pareto optimal.

Models with complete markets have proved to be very useful in economics, for example in the theory of asset pricing and in business cycle modelling. However, there are many applications where it is necessary that we depart from the complete markets paradigm, and the liquidity constraints literature is one such set of applications. To think about liquidity constraints we need to seriously address the frictions that will cause markets to work differently than in the complete markets case, and in some instances will cause some markets to shut down altogether. In the following sections we will explore some key departures from our benchmark model that illustrate the role of liquidity constraints.

Incomplete markets: a Bewley model

One approach to studying market incompleteness is to simply eliminate markets in the model under consideration, without asking questions about the underlying frictions which would cause incomplete markets. Bewley (1977) was a pioneer in this area, and Aiyagari (1994) provides a particularly clear treatment of the implications of incomplete markets.

As an example of the Bewley approach, suppose in our benchmark model that there is only one asset market, a market for non-contingent bonds on which trading occurs each period. Households can borrow and lend on this bond market. Assume that each bond is a one-period financial instrument. In period *t*, a bond sells for one unit of consumption goods, and is a promise to pay $1 + r_{t+1}$ units of consumption goods in period $t + 1$. Since there is no aggregate risk, there will exist a steady state competitive equilibrium where $r_{t+1} = r$, a constant, for all *t*.

We now need to write down the series of constraints that a household faces in the steady state equilibrium. The first of these is the sequence of budget constraints

$$c_{it} + b_{i,t+1} = y_{it} + (1 + r)b_{i,t},$$

for $t = 0, 1, 2, \ldots$, where b_{it} is the quantity of bonds acquired by household *i* in period *t*, and $b_{i0} = 0$ for all *i*. Typically in models of this type, there is also a borrowing constraint added, which could take the form

$$b_{it} \geq \underline{b}. \quad (1)$$

Constraint (1) serves a technical purpose, in that it prevents a household from borrowing an infinite amount so as to finance infinite consumption. Further, the constraint will affect the household's ability to smooth consumption over time in the face of fluctuating income. Constraint (1) is a kind of liquidity constraint, as it potentially prevents the household from borrowing against its lifetime wealth.

A competitive equilibrium will have the property that the bond market clears, that is the net stock of bonds in the population is zero in each period. This model is a special case of Aiyagari (1994), and so his results apply here. With Aiyagari's regularity conditions on $u(\cdot)$, a steady state competitive equilibrium will have the property that $r < \frac{1}{\beta} - 1$, that is, the equilibrium real interest rate is less than the rate of time preference. This reflects a precautionary savings motive, in that households wish to hold bonds to self-insure against having a string of bad luck, which in this case would be a string of low endowment shocks. Over time, a household will tend to increase its stock of bonds when its endowment is large, and to decrease the stock of bonds when its endowment is small. What we will observe in equilibrium is some distribution of bonds and consumption across the population of households. Households who have had good luck will tend to have a larger stock of bonds and higher consumption than those households who have had bad luck. The competitive equilibrium is therefore not in general Pareto optimal.

Another related application, from Bewley (1980), is to suppose that the single asset that is traded is money. For example, suppose that there is a fixed stock of money, *M*, for all *t*. Let P_t denote the price level in period *t*, and consider the steady state equilibrium where $P_t = P$, a constant, for all *t*. For the household, we can just reinterpret its constraints, in that $b_{i,t+1}$ is the real quantity of money carried over by the household into period $t + 1$, and $\underline{b} = 0$ as the household's money balances cannot fall below zero. An individual household in this set-up is even more severely liquidity-constrained than was the household in the Bewley model with borrowing and lending above. This is because the household cannot borrow at all, and cannot hold interest-bearing assets. Note that, in this monetary model, a household need only use money to buy consumption goods if it wishes to consume more than its endowment. Money is essentially held for insurance purposes, so as to smooth consumption over time.

Cash-in-advance

The idea for the basic cash-in-advance model seems to come from Clower (1967), but the important initial modelling work was done mainly by Robert Lucas, with a key contribution being Lucas (1980). Most cash-in-advance applications begin with the view that the basic frictions that might give rise to cash-in-advance constrained households need not be modelled, and that it is useful to proceed from the premise that money is necessary to purchase some goods and services.

Here, suppose in our basic model that there are no assets other than money, and that the only exchanges are trades of money for goods. Assume that a household's purchases of goods during the current period must be financed with money carried over from the previous period, and also suppose that the household cannot consume its own endowment. Let m_{it} denote the nominal money balances that household i has at the beginning of period t, and let P_t denote the price level. Then, the household's budget constraint in period t is

$$P_t c_{it} + m_{i,t+1} = P_t y_{it} + m_{it}. \tag{2}$$

The cash-in-advance constraint for the household is

$$P_t c_{it} \leq m_{it}. \tag{3}$$

Thus, constraint (3) is another type of liquidity constraint. In this case, the interpretation is that some class of assets, which we refer to here as money, is necessary to carry out goods market spot exchanges. Now, suppose that there is a fixed nominal stock of money M. Also, suppose that in equilibrium constraint (3) binds for each household i. Then, since in equilibrium the entire stock of money is held by households at the beginning of period and is spent to purchase the aggregate endowment, y, the equilibrium price level is

$$P_t = \frac{M}{y}$$

for all t. Then, given (2) and (3) with equality, we have

$$m_{i,t+1} = M\frac{y_{it}}{y},$$

which then implies, from (3) with equality, that

$$c_{it} = y_{i,t-1}.$$

Therefore, in this environment, households have essentially no ability to smooth consumption relative to income, as a result of this extreme type of liquidity constraint. The distribution of consumption across households in period *t* is determined by the distribution of income across households in the previous period.

Economists who are serious about monetary theory often treat cash-in-advance models with some disdain (see, for example, Wallace, 1996). As they see it, the problem is not that one cannot write down a model that is explicit about frictions and gives rise to cash-in-advance as an endogenous phenomenon. For example, suppose that we modify our benchmark model to permit an absence-of-double-coincidence friction of the type considered by early monetary theorists such as Jevons (1875). That is, assume that households are of *N* types, with measure $\frac{1}{N}$ households of each type, where type is indexed by $j = 1, 2, \ldots, N$. Type *j* households are endowed with good *j*, and consume the good which is endowed to type $j + 1$, modulo *N*. Further, suppose that a household has two members, a shopper that takes money from the household to buy goods in another market each period, and a seller who stays at home to sell the household's endowment. There are *N* distinct markets, and in a given period a shopper from a household of type *j* goes to market $j + 1$, modulo *N*, with money to buy goods, while a the seller stays behind and sells goods in market *j*. Note that this is still not enough to give us cash-in-advance, as we need to close off the possibility of credit arrangements among households which could take place through centralized communication, as is made clear in Kocherlakota (1998). Credit can be shut down by assuming that no communication is possible across markets, with buyers and sellers in a given market having no information about each other, beside the fact that sellers have identifiable goods and buyers have identifiable money balances. With competitive pricing in each of the *N* markets, we get exactly the set-up outlined above in this section, with a cash-in-advance constraint for each household. Given symmetry, there is an equilibrium where prices are the same in every market, and so the equilibrium allocation of consumption is identical to what was specified above.

The key problem that must be addressed in cash-in-advance environments involves what happens when there are other assets than money. For example, if we permit borrowing and lending by households, why is it that goods cannot be purchased with credit? How can money be dominated in rate of return by other assets? Why is it that government bonds, for example, are not used in transactions rather than money? Many cash-in-advance applications leave these questions unanswered.

Random matching

A useful way to extend our benchmark model at this point is to expand on the explicit cash-in-advance environment above to relate it more directly to the literature on monetary search and matching. The seminal work in this literature is by Jones (1976) and Kiyotaki and Wright (1989).

Suppose as above that there is a double coincidence problem, but here assume that there is one agent in a household, and that each household is randomly matched with one other household each period. Households produce different goods, and no household can consume its own endowment. Now, for a given household, assume that the probability is α that it is matched with another household whose goods it consumes, with the other household not wanting its goods (a single coincidence meeting). As well, assume that there is a probability γ that a household is matched with another household and there is a double coincidence of wants – each household consumes the other's goods. Suppose that $\alpha > 0$, $\gamma > 0$, and $2\alpha + \gamma < 1$. Suppose that a household in a bilateral match has no information about the other household, except that it can observe its quantity of money balances and its endowment. Thus, exchange can only involve bilateral exchanges of goods and money.

Now, suppose that household *i* and household *k* are matched. There is probability α that household *i* is a seller and *k* is a buyer. In this case, we have $c_{it} = 0$, $c_{kt} = y_{it}$, and

$$m_{i,t+1} = m_{it} + m(y_{it}, m_{it}, m_{kt}), \quad m_{k,t+1} = m_{k,t} - m(y_{it}, m_{it}, m_{kt}),$$

where $m(y_{it}, m_{it}, m_{kt})$ is the quantity of money exchanged for the y_{it} units of goods given up by the seller when the seller has m_{it} units of money and the buyer has m_{kt} units of money. As money balances must be non-negative, we have

$$-m_{it} \leq m(y_{it}, m_{it}, m_{kt}) \leq m_{kt} \tag{4}$$

and these constraints are essentially liquidity constraints. Similarly, with probability α, household *i* is the buyer and *k* is the seller, in which case $c_{it} = y_{kt}$, $c_{kt} = 0$, and

$$m_{i,t+1} = m_{it} - m(y_{kt}, m_{kt}, m_{it}), \quad m_{k,t+1} = m_{k,t} + m(y_{kt}, m_{kt}, m_{it}),$$

with

$$-m_{kt} \leq m(y_{kt}, m_{kt}, m_{it}) \leq m_{it}. \tag{5}$$

Finally, with probability γ there is a double coincidence, and household *i* and *k* exchange goods, so that $c_{it} = y_{kt}$, $c_{kt} = y_{it}$, and

$$m_{i,t+1} = m_{it} + b(y_{it}, y_{kt}, m_{it}, m_{kt}), \quad m_{k,t+1} = m_{k,t} - b(y_{it}, y_{kt}, m_{it}, m_{kt}),$$

where

$$-m_{it} \leq b(y_{it}, y_{kt}, m_{it}, m_{kt}) \leq m_{kt}. \tag{6}$$

Here, $b(y_{it}, y_{kt}, m_{it}, m_{kt})$ is the quantity of money passed from household *k* to household *i*, which depends on the money balances and endowments of each household.

Note that this environment will give a clear sense in which money improves the equilibrium allocation. If money is not valued, then households can trade only when there is a double coincidence of wants, and this could severely limit exchange possibilities. In principle, the constraints (4), (5) and (6) will matter for the equilibrium allocation in important ways. However, the model as we have laid it out is quite intractable. It is possible to use a bargaining approach, as for example in Trejos and Wright (1995) or Shi (1995), to determine how much money is transferred in each type of match, but the key problem is in tracking the distribution of money balances in the population over time.

In some of the monetary search and matching literature, tractability is achieved through assuming that money and goods are indivisible (Kiyotaki and Wright, 1989) or that money is indivisible and goods are divisible (Trejos and Wright, 1995; Shi, 1995), and that there is an inventory constraint on money holdings. If a household can hold only one unit of money or nothing, and money is never disposed of, then the quantity of money outstanding tells us how many households have it and how much, and how many do not have it. Models with indivisible money yield some insights, but they are extremely awkward for dealing with some types of policy questions, such as those involving money growth and the effects of inflation. Some recent progress in the development of tractable search models of divisible money was achieved by Lagos and Wright (2005), who use a quasilinear utility setup with labour supply and alternating periods of centralized meeting and search. This type of model yields a result where, in the periods when centralized meeting takes place, economic agents optimally redistribute money among

themselves in such a way that the distribution of money balances becomes degenerate. Recent research using this type of model (for example, Williamson, 2006; Berentsen, Camera and Waller, 2005) has been quite productive.

Private information and limited commitment

As an alternative to shutting down markets in an ad hoc fashion, imposing borrowing constraints, assuming cash-in-advance constraints, or making extreme informational assumptions that shut down all trade except monetary exchange, there are available approaches to facing frictions head-on that lead to incomplete insurance and imperfect credit. These approaches involve economies with private information and limited commitment.

A well-developed approach to dealing with private information frictions in large economies follows the pioneering work of Green (1987), Atkeson and Lucas (1993) and others. Extending our benchmark model, suppose now that endowments are private information. In our baseline environment, we know that if endowments are public information, then a Pareto optimal allocation that treats households identically has $c_{it} = \bar{y}$ for all *i*,*t*. What is optimal from a social planner's point of view under private information?

It is clear that private information implies that the $c_{it} = \bar{y}$ allocation cannot be implemented by the social planner. To see this, note that to achieve this allocation requires that household *i* make a transfer of $y_{it} - \bar{y}$ to the planner in period *t*. But it would then be incentive compatible for every household in every period to report that its endowment was $\underline{y}$, and so the planner could not achieve this allocation.

Following Green (1987) and Atkeson and Lucas (1995), one can solve for an optimal private information allocation by recursive methods. The state variable for any household is w_{it}, which is the level of expected utility promised to the household as of the beginning of period *t*. At the beginning of period *t*, the household reports its endowment y_{it} to the social planner, and it must be optimal for the household to report the truth (that is, the allocation must be incentive compatible). The planner delivers consumption $c(w_{it}, y_{it})$ to the household, which depends on its state and reported endowment, and promises expected utility $w(w_{it}, y_{it})$ for next period. There is a functional equation that solves for a cost function $V(w_{it})$, which is the cost to the social planner of delivering expected utility w_{it} to a particular household. On the right-hand side of this functional equation is a cost minimization problem, and the minimization is subject, first, to a promise-keeping constraint, which is

$$w_{it} = \int \{u[c(w_{it}, y_{it})] + \beta w(w_{it}, y_{it})\} dF(y_{it}),$$

where $F(y_{it})$ is the distribution function for y_{it}. The remaining constraints are incentive compatibility constraints, written as

$$u[c(w_{it}, y_{it})] + \beta w(w_{it}, y_{it}) \geq u[c(w_{it}, \tilde{y}) + y_{it} - \tilde{y}] + \beta w(w_{it}, \tilde{y}),$$

for all y_{it}, $\tilde{y} \in [\underline{y}, \overline{y}]$. The optimal allocation will typically have the property that some incentive compatibility constraints bind. For efficient risk sharing, we want households with high (low) endowments to be making positive (negative) transfers to the social planner. To accomplish this in an incentive compatible manner requires that households with high (low) endowments receive increases (decreases) in their future expected utility promises. Thus, the distribution of consumption will tend to fan out over time. Under some conditions, a vanishing fraction of households will ultimately consume the entire endowment. However, under other conditions there will be a limiting distribution of expected utility promises with mobility and a lower bound on expected utilities. If a household hits this lower bound (which is not absorbing), then this is much like having a borrowing constraint bind for this household. Thus, this type of set-up can yield what are essentially endogenous borrowing constraints or liquidity constraints.

An alternative approach to modeling frictions in a serious way is to assume some form of limited commitment. One approach to limited commitment is that of Kehoe and Levine (1993), which has elements of competitive equilibrium. Extending our benchmark model to illustrate the flavour of this modelling approach, suppose that there is only one type of intertemporal trade, involving one-period bonds, and that we wish to study a steady state where the real interest rate is a constant, *r*. Suppose that the key friction here is that a household may decide strategically to repudiate its debt, in which case it would be barred from the credit market for ever and would then consume its own endowment for ever. Thus, if a household does not repudiate its debt, then its budget constraint is given by

$$c_{it} + b_{i,t+1} = y_{it} + (1 + r)b_{it}, \tag{7}$$

where $b_{i,t+1}$ is the quantity of one-period bonds acquired in period *t* that each pay off $1 + r$ units of consumption in period $t + 1$. Let $v(b_{it}, y_{it})$ denote the expected utility of the household at the beginning of period *t* as a function of the household's asset position and endowment, determined by the functional equation

$$v(b_{it}, y_{it}) = \max_{c_{it}, b_{i,t+1}} \left[u(c_{it}) + \beta \int v(b_{i,t+1}, y_{i,t+1}) dF(y_{i,t+1}) \right]$$

subject to (7). To insure that the household does not repudiate its debt in equilibrium requires that the

value of not repudiating is no smaller than the value of repudiating, or

$$v(b_{it}, y_{it}) \geq u(y_{it}) + \frac{\beta}{1-\beta}\int u(\hat{y})\,dF(\hat{y}). \tag{8}$$

Note that constraint (8) is another type of borrowing constraint or liquidity constraint. Typically, $v(b_{it}, y_{it})$ must be strictly increasing in b_{it} and so, given y_{it}, there will be some critical value of b_{it} for which the constraint binds. Thus, lenders cannot lend too much to a particular household, as doing so would imply debt repudiation.

Kocherlakota (1996) takes a somewhat different approach by examining a two-agent problem with limited commitment. In his set-up, two infinite-lived agents work out a risk-sharing arrangement subject to limited commitment. Kocherlakota's problem does not have some of the loose ends found in Kehoe and Levine (1993). In the Kehoe and Levine model, we are forced to accept an incomplete markets view of the world with no explanation for why the markets are missing, and it is not clear how credit market participants coordinate to discipline agents who repudiate their debts.

Aiyagari and Williamson (2000) integrate private information and limited commitment with a Bewley model of monetary exchange to study the relationship between money and credit. Credit arrangements are constrained by private information considerations, and if agents defect from credit arrangements their alternative is to be liquidity constrained in the manner of a Bewley-type consumer.

See Also

- Aiyagari, S. Rao
- incomplete markets
- Lucas, Robert
- money
- money and general equilibrium

Bibliography

Aiyagari, R. 1994. Uninsured idiosyncratic risk and aggregate saving. *Quarterly Journal of Economics* 109, 659–84.

Aiyagari, R. and Williamson, S. 2000. Money and dynamic credit arrangements with private information. *Journal of Economic Theory* 91, 248–79.

Atkeson, A. and Lucas, R. 1993. On efficient distribution with private information. *Review of Economic*

Studies 59, 427–53.

Atkeson, A. and Lucas, R. 1995. Efficiency and inequality in a simple model of unemployment insurance. *Journal of Economic Theory* 66, 64–88.

Berentsen, A., Camera, G. and Waller, C. 2005. The distribution of money balances and the nonneutrality of money. *International Economic Review* 46, 465–88.

Bewley, T. 1977. The permanent income hypothesis: a theoretical formulation. *Journal of Economic Theory* 16, 252–92.

Bewley, T. 1980. The optimum quantity of money. In *Models of Monetary Economies*, ed. J. Kareken and N. Wallace. Minneapolis, MN: Federal Reserve Bank of Minneapolis.

Clower, R. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6, 1–8.

Green, E. 1987. Lending and the smoothing of uninsurable income. In *Contractual Arrangements for Intertemporal Trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

Jevons, W.S. 1875. *Money and the Mechanism of Exchange*. London: Appleton.

Jones, R. 1976. The origin and development of media of exchange. *Journal of Political Economy* 84, 757–75.

Kehoe, T. and Levine, D. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60, 865–88.

Kiyotaki, N. and Wright, R. 1989. On money as a medium of exchange. *Journal of Political Economy* 97, 927–54.

Kocherlakota, N. 1996. Implications of efficient risk-sharing without commitment. *Review of Economic Studies* 63, 595–610.

Kocherlakota, N. 1998. Money is memory. *Journal of Economic Theory* 81, 232–51.

Lagos, R. and Wright, R. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 113, 463–84.

Lucas, R. 1980. Equilibrium in a pure currency economy. In *Models of Monetary Economies*, ed. J. Kareken and N. Wallace. Minneapolis, MN: Federal Reserve Bank of Minneapolis.

Shi, S. 1995. Money and prices: a model of search and bargaining. *Journal of Economic Theory* 67, 467–96.

Trejos, A. and Wright, R. 1995. Search, bargaining, money, and prices. *Journal of Political Economy* 103, 118–41.

Wallace, N. 1996. A dictum for monetary theory. In *Foundations of Research in Economics: How do Economists do Economics?* ed. S.G. Medema and W.J. Samuels. Cheltenham: Edward Elgar.

Williamson, S. 2006. Search, limited participation, and monetary policy. *International Economic Review* 47, 107–28.

How to cite this article

Williamson, Stephen D. "liquidity constraints." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000235> doi:10.1057/9780230226203.0978

The New Palgrave Dictionary of Economics Online

liquidity effects, models of

Chris Edmond and Pierre-Olivier Weill

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An exogenous increase in the money supply is typically followed by a temporary fall in nominal interest rates. Flexible price macroeconomic models argue that this *liquidity effect* arises because asset markets are segmented. That is, only a fraction of the agents are present in the bond market when the central bank conducts an open market operation. However, to be quantitatively successful, segmented markets models assume frictions that are too large to be interpreted literally in terms of constraints faced by real-world firms and households. An important open question is: can a complicated array of microeconomic frictions imply one large aggregate friction of this kind?

Keywords

asset market frictions; asset market segmentation; cash-in-advance models; Fisher effect; inflation; inflation expectations; liquidity effects; long-horizon interest rates; monetary transmission mechanism; money supply; nominal interest rates; open-market operations; real business cycle; real interest rates; short-horizon liquidity effects; velocity of circulation

Article

In macroeconomics, the term *liquidity effect* refers to a fall in nominal interest rates following an exogenous persistent increase in narrow measures of the money supply. According to the classical *Fisher effect*, however, an exogenous persistent increase in money is predicted to increase expected inflation and so increase nominal interest rates. Friedman (1968) argues that, in practice, both forces operate: a persistent increase in the money supply both reduces nominal interest rates and increases expected inflation so that the real rate – nominal minus expected inflation – also falls. Friedman (1968, pp. 5–7) speculates that nominal and real rates may fall below their typical levels for up to a year, but, over time, rates will then tend to increase before tending to the levels consistent with the inflation generated by the original monetary impulse.

Empirical macroeconomists have interpreted Friedman (1968) as follows. At long horizons real interest rates are determined by ‘fundamentals’ including the rate at which households discount the future and average productivity growth. Consequently, we should expect that long-horizon real interest rates are relatively stable and are unaffected by transitory monetary disturbances. Long-horizon nominal interest rates are this stable real rate plus expected inflation. At short horizons, however, Friedman's (1968) argument suggests that real and nominal interest rates are both volatile and positively correlated. His argument also suggests that short-horizon real rates and expected inflation are negatively correlated (Barr and Campbell, 1997, provide evidence consistent with this interpretation and Cochrane, 1989, provides specific evidence for liquidity effects at short horizons).

Perhaps the easiest way to interpret Friedman (1968) is in terms of the following market equilibrium scenario. Suppose that a monetary authority increases the money supply by conducting an unexpected outright purchase of bonds (an *open market operation*). At short horizons, nominal interest rates fall so that households are willing to hold a smaller quantity of bonds and a larger quantity of money. But this is only a partial equilibrium effect. As households spend their increased money holdings on goods, the price level increases and so real balances do not rise as fast as nominal balances. This general equilibrium effect mitigates the need for the nominal interest rate to fall. In many simple monetary models, households tend to spend money so ‘fast’ that the general equilibrium price level effect can completely overturn the partial equilibrium effect.

A textbook cash-in-advance (CIA) model with a constant aggregate endowment of goods (‘output’) and identically and independently distributed (IID) money growth shocks provides a stark example. In this model, households immediately spend an unexpected increase in money on a fixed quantity of goods. This increases the price level one-for-one with the increase in the money supply so that real balances are unchanged. In addition, because money growth is serially uncorrelated, expected inflation is constant. Taken together, constant real balances and constant expected inflation imply that the money market clears at a constant nominal interest rate. If instead monetary growth shocks are persistent then a positive shock increases expected inflation and nominal interest rates increase. In short, there is a Fisher effect but no liquidity effect. CIA models that are carefully calibrated to empirical processes for money growth and output, such as Hodrick, Kocherlakota and Lucas (1991) and Giovannini and Labadie (1991), lead to similar conclusions, as do studies of conceptually similar production economies, such as Cooley and Hansen (1989).

We now turn to departures from the standard CIA model in which a liquidity effect dominates at short horizons while a Fisher effect dominates at long horizons. Although models with nominal rigidities are in principle capable of generating these liquidity effects, we instead focus on flexible price models in which a liquidity effect is generated by an *asset market friction* of one form or another. Each of the models we discuss – Lucas (1990), Grossman and Weiss (1983), and Alvarez, Atkeson and Kehoe (2002) – captures, albeit in different ways, some of the spirit of Friedman's (1968) intuition.

Lucas (1990) modifies the standard CIA endowment economy with a simple timing assumption: households have to allocate cash between a goods market and an asset market *before* observing the size of an open-market operation. Once that allocation has been made, there is a fixed quantity of cash sitting in the bond market. Now consider an unexpected purchase of bonds. Relative to the supply of bonds, there is now an unexpectedly large amount of cash available to purchase assets, so bond prices increase and the nominal interest rate falls.

Fuerst (1992) and Christiano and Eichenbaum (1995) integrate Lucas's (1990) timing assumption into otherwise standard real business cycle (RBC) models. The key innovation of these papers is that, in each period, firms have to borrow cash from financial intermediaries in order to pay their workers. After a positive monetary shock, the nominal interest rate decreases so that firms find it optimal to borrow the unexpected increase in money balances. This increases firms' labour demand and increases output. Thus, these models are consistent with the commonly held view that positive monetary shocks have a positive, albeit temporary, effect on output.

A limitation of models that use Lucas's (1990) timing assumption is that the liquidity effect is very transitory even when monetary shocks are persistent. Households can adjust their allocation of cash every period. Therefore, the liquidity effect is entirely driven by serially uncorrelated 'expectational errors' in cash allocation.

We now turn to Grossman and Weiss (1983) and Alvarez, Atkeson and Kehoe (2002). These are general equilibrium models inspired by Baumol (1952) and Tobin's (1956) 'inventory-theoretic' analyses of money demand. In this class of models, two key forces influence short-horizon liquidity effects. First, at any point in time, there are always some households that participate in asset markets and some households that do not. Second, because households do not acquire cash every period, they choose to spend their money holding slowly over time. The first force alone is sufficient to generate a liquidity effect; the second force provides an *amplification* mechanism.

In this setting, an open-market increase in the money supply must, in equilibrium, be held by the subset of households that are currently participating in asset markets. Therefore, even if the price level responds one-for-one with the increase in money supply, the *share* of aggregate real balances that must be held by these households increases. Hence, the nominal interest rate falls to clear the market. Also, because they hold a larger share of real balances, these households are able to increase their share of aggregate consumption and this drives down real interest rates. So, at short horizons, there is a liquidity effect. Moreover, if households spend their money slowly over several subsequent periods then the price level does not respond one-for-one to an increase in the money supply. Instead, the price level responds slowly. This implies that aggregate real balances rise (equivalently, in a model with constant output, *velocity* falls) and this provides a second force driving down nominal interest rates. The liquidity effect is amplified.

The influential model of Grossman and Weiss (1983) is a deterministic CIA endowment economy that exhibits both effects. Households are imperfectly synchronized and only participate in asset markets every second period. They spend money on consumption goods over two periods (Rotemberg, 1984, studies a production version of essentially the same environment).

Alvarez, Atkeson and Kehoe (2002) endogenize the fraction of households that participate in asset markets. They assume that households can participate if they pay a fixed cost. If a household's individual real balances are neither too high nor too low, they do not pay the cost, do not participate in asset markets, and end up consuming their individual real balances. If their real balances are high, they pay the cost and invest money in the asset market. Similarly, if their real balances are low, they pay the cost in order to purchase goods with money invested in the asset market. The equilibrium amount of participation ends up depending on the curvature of the utility function, the expected growth rate of money and on the size of the fixed cost. For example, in a high-inflation economy almost all households pay the cost to participate in asset markets. Hence, increases in the money supply raise expected

inflation and nominal interest rates as in a basic CIA model. By contrast, in a low inflation economy, more households choose not to participate and the effects of incomplete participation are larger and may be big enough to cause a liquidity effect (that is, to dominate the Fisher effect at short horizons).

To simplify their analysis, however, Alvarez, Atkeson and Kehoe (2002) set up the model so that both active and inactive households spend all their money each period. No households save money to spend on consumption over multiple periods. Therefore, velocity is constant and the price level responds one-for-one with increases in the money supply. Alvarez, Atkeson and Kehoe (2002) can therefore generate a liquidity effect but without the amplification that is provided by a (transitory) fall in velocity. Alvarez, Atkeson and Edmond (2003) provide a stochastic counterpart to Grossman and Weiss (1983) where both forces are operative (but at the cost of reverting to an exogenous timing of transactions).

Limited participation models of the liquidity effect provide a number of important qualitative insights into the co-movements of money, interest, and prices (and, to a lesser extent, output). The *quantitative* insight provided by these models is, however, more debatable. To generate realistic co-movements of money, interest and prices, calibrated models of liquidity effects need ‘large’ asset market frictions. It is typically difficult to interpret the calibrated friction literally in terms of constraints faced by real-world firms and households (making it difficult, in the words of Manuelli and Sargent, 1988, p. 524, to ‘find the people’). For example, the most successful parameterizations in Alvarez, Atkeson and Edmond (2003) require the representative household to make withdrawals of money (broadly defined) from an asset market account once every 24–36 months. Alvarez, Atkeson and Edmond (2003) defend this with an appeal to the low frequency of asset market participation observed in the *cross-section* by Vissing-Jorgensen (2002). Thus, the size of the friction is defended by appealing to the likely size of the friction facing a household representative of the US economy rather than by appealing to direct evidence of the heterogeneous frictions facing individual observations of US households.

Cole and Ohanian (2002) provide another demonstration of the difficulty of interpreting such models literally. They note that the distribution of money holdings between US firms and households has been quite unstable over the post-war period. When this observation is embedded in a model of liquidity effects, it implies a corresponding instability in the effects of money shocks on output – an instability that seems to be counterfactual.

In our opinion, these limitations should not be interpreted as reasons for rejecting models of asset market segmentation. If anything, these limitations are instead reasons for rejecting an implicit aggregation hypothesis. Traditional macro models work with relatively crude frictions that are intended to summarize a complicated array of micro frictions facing individual households and firms. For example, the literature on models of liquidity effects assumes only one level of market segmentation – either between households and asset markets, or between firms and asset markets. However, asset market segmentation seems to occur at numerous levels of financial intermediation. A large body of empirical evidence shows that phenomena consistent with market segmentation arise within the financial system – a system that might best be viewed as a collection of partially integrated and relatively specialized ‘local’ asset markets (see, among many others, Collin-Dufresne, Goldstein and Martin, 2001).

This evidence motivates us to ask how a collection of small segmentation frictions cumulates in the aggregate, and whether they add up to a quantitatively significant macro friction. If they do, then the models of liquidity effects that we have discussed here would indeed be natural laboratories for the analysis of the monetary transmission mechanism.

In short, we conjecture that addressing segmentation at a disaggregative level is likely to provide important empirical and theoretical insights into the relationship between patterns of intermediation in financial markets and traditional macro questions – including the size and stability of liquidity effects at short horizons and the monetary policy transmission mechanism more generally.

See Also

- finance
- Fisher, Irving
- Friedman, Milton
- inflation expectations
- Lucas, Robert
- money supply

Bibliography

Alvarez, F., Atkeson, A. and Edmond, C. 2003. On the sluggish response of prices to money in an inventory- theoretic model of money demand. Working Paper No. 10016. Cambridge, MA: NBER.

Alvarez, F., Atkeson, A. and Kehoe, P.J. 2002. Money, interest rates and exchange rates with endogenously segmented markets. *Journal of Political Economy* 110, 73–112.

Barr, D.G. and Campbell, J.Y. 1997. Inflation, real interest rates and the bond market: a study of UK nominal and index-linked government bond prices. *Journal of Monetary Economics* 39, 361–83.

Baumol, W.J. 1952. The transactions demand for cash: an inventory-theoretic approach. *Quarterly Journal of Economics* 66, 545–56.

Christiano, L.J. and Eichenbaum, M. 1995. Liquidity effects, monetary policy, and the business cycle. *Journal of Money, Credit and Banking* 27, 1113–36.

Cochrane, J.H. 1989. The return of the liquidity effect: a study of the short-run relation between money growth and interest rates. *Journal of Business and Economic Statistics* 7(1), 75–83.

Cole, H.L. and Ohanian, L.E. 2002. Shrinking money: the demand for money and the nonneutrality of money. *Journal of Monetary Economics* 49, 653–86.

Collin-Dufresne, P., Goldstein, R.S. and Martin, J.S. 2001. The determinants of credit spread changes. *Journal of Finance* 56, 653–86.

Cooley, T.F. and Hansen, G.D. 1989. The inflation tax in a real business cycle model. *American*

Economic Review 79, 733–48.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Fuerst, T.S. 1992. Liquidity, loanable funds, and real activity. *Journal of Monetary Economics* 29, 3–24.

Giovannini, A. and Labadie, P. 1991. Asset prices and interest rates in cash-in-advance models. *Journal of Political Economy* 99, 1215–51.

Grossman, S. and Weiss, L. 1983. A transactions-based model of the monetary transmission mechanism. *American Economic Review* 73, 871–80.

Hodrick, R.J., Kocherlakota, N. and Lucas, D. 1991. The variability of velocity in cash-in-advance models. *Journal of Political Economy* 99, 358–84.

Lucas, R.E., Jr. 1982. Interest rates and currency prices in a two-country world. *Journal of Monetary Economics* 10, 335–59.

Lucas, R.E., Jr. 1990. Liquidity and interest rates. *Journal of Economic Theory* 50, 237–64.

Manuelli, R. and Sargent, T.J. 1988. Models of business cycles: a review essay. *Journal of Monetary Economics* 22, 523–42.

Rotemberg, J.J. 1984. A monetary equilibrium model with transactions costs. *Journal of Political Economy* 92, 40–58.

Tobin, J. 1956. The interest-elasticity of the transactions demand for cash. *Review of Economics and Statistics* 38, 241–7.

Vissing-Jorgensen, A. 2002. Towards an explanation of household portfolio choice heterogeneity: nonfinancial income and participation cost structures. Working Paper No. 8884. Cambridge, MA: NBER.

How to cite this article

Edmond, Chris and Pierre-Olivier Weill. "liquidity effects, models of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000378> doi:10.1057/9780230226203.0979

The New Palgrave Dictionary of Economics Online

liquidity preference

Carlo Panico

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Keynes's notion of liquidity preference stems from the fact that he made some specific sources of demand for monetary instruments depend upon the expected variations of the interest rate, and consequently on the expected variations in the capital value of financial assets. This source of demand was considered to be *the* cause of variations in the rate of interest. Economists close to Keynes realized that in the *General Theory* he had turned the analysis of liquidity preference into a new theory of the interest rate. Robertson defended the marginalist theory, while Hicks paved the way for the ‘neoclassical synthesis’.

Keywords

Cambridge equation; central bank money; finance motive; Hicks, J. R.; IS–LM models; Keynes, J. M.; liquidity preference; loanable funds; marginalist theory of the rate of interest; monetary instruments; natural rate and average rate of interest; neoclassical synthesis; precautionary motive; reserves–loans ratio; Robertson, D.; saving and investment; simultaneous equilibrium; speculation; speculative motive; subjective probability; temporary equilibrium; Tobin, J.; transaction motive; uncertainty; Walras's Law

Article

The notion of ‘liquidity preference’ has become generally used in the literature on monetary issues (particularly that concerned with the interest rate) following Keynes's contributions in the 1930s. It concerns the motives for demanding monetary instruments or other close substitutes. Earlier, the analysis of the demand for monetary instruments was based on other motives and concepts and led to different conclusions.

The analysis of the motives for demanding monetary instruments plays a specific role within monetary theory. The literature dealing with the interest rate, for instance, has always distinguished two different analytical steps. The first deals with the *variations* in the ‘market interest rate’, that is, that actually

observed everyday: it *describes* how a change in this rate (or in the structure of the interest rates) comes about. To do that, it provides an analytical scheme which describes the behaviour of the money markets, by considering one after the other all different sources of demand for and supply of monetary instruments, pointing out the main causes of their variations. The second step deals with the *level* of the interest rate. It *explains* why this rate tends to remain, over a specific period of time, at a certain level, pointing out the factors determining it. The way in which these factors operate is then described by using the scheme provided in the first step of analysis. This clarifies the market mechanisms (that is, changes in the different components of demand and supply in the money markets) through which the prevailing level of the interest rate asserts itself.

The analysis of the motives for demanding monetary instruments thus properly belongs to the first step: it cooperates to describe the working of the money markets and the way in which variations in the interest rate (or in the structure of interest rates) occur.

This approach was followed by Smith, Ricardo, Tooke, J.S. Mill, Marx, Marshall, Wicksell, J.M. Keynes, Robertson, and so on, independently of the particular theory they proposed, that is whether the level of the ‘average interest rate’ (that prevailing over a specific period of time) was determined by the ‘forces of productivity and thrift’ or by other factors.

Prior to Keynes's contributions in the 1930s, it was assumed that monetary instruments (in most cases, central bank money) are demanded for two reasons. First, they are demanded by the household sector for the ‘circulation of income’. Households, that is, hold in the form of currency a certain fraction of their income to carry out their daily consumption expenditure.

The second source of demand for central bank money, it was assumed, comes from the banking sector which requires liquid reserves to make payments to depositors and to meet the demand for bank loans of different maturity. Banks’ decisions, it was argued, are concerned with protecting themselves against the risk of running out of liquid means while minimizing cost. In such analyses, which did not use modern portfolio choice tools, the amount of reserves banks demand depends upon the composition of their portfolio (particularly the maturity of their loans) and upon the degree of uncertainty they feel as to the smooth operation of the credit payment system. On the basis of these two elements, banks fix the desired ratio between their reserves in central bank money and the amount of loans they can supply.

As some authors noticed, the presence of uncertainty among the elements affecting the decisions of financial operators makes the credit payment system unstable. The desired ratio of reserves to loans changes continuously and sometimes sharply. Financial markets become tighter precisely when more liquid means are required. A higher degree of uncertainty as to the smooth operation of the system, for instance, leads the business and the banking sectors to desire to ‘become more liquid’. The former tend to discount a larger amount of bills of exchange (that is, demand more short-term bank loans), while the latter set at a higher level the desired reserves–loans ratio, so supplying a smaller amount of bank loans. The instability of the system and the variability of the interest rates were therefore recognized by some economists (a minority) and ascribed to the uncertainty felt by banks and business as to their ability to solve cash-flow problems.

This analysis of the demand for monetary instruments was dominant from Adam Smith onwards. Its basic points were still reflected in the famous ‘Cambridge equation’ presented by Pigou in his article ‘The value of money’ (1917) and in Keynes's *A Tract on Monetary Reform* (1923).

Keynes's analysis of liquidity preference

The analysis of the motives for demanding monetary instruments was considerably refined by J.M. Keynes in the 1930s. Developing the analysis inherited from Marshall and Pigou, Keynes distinguished three motives for demanding monetary instruments (by which was now meant member banks' money, that is, bank deposits).

First, monetary instruments are demanded for transaction purposes. The amount demanded due to this motive is a stable function of the level of income.

The second source of demand for monetary instruments is for precautionary purposes, defined as the demand coming from different sectors as a protection against the possibility that some unexpected payment has to be made, or that some expected receipts cannot be realized. This definition has been differently interpreted. Some authors (and the majority of textbooks) have interpreted it in a restrictive way, by identifying it with the households' holding of bank deposits as a precaution against extraordinary events (for example, payment of hospital bills). The precautionary demand for monetary instruments was typically lumped together with the transaction demand, both being an increasing function of the level of income. Other authors have instead given more extensive interpretation of this motive by including in it the demand coming from all financial operators feeling highly uncertain as to the future level of the interest rate. R. Kahn (1954) explained that people prefer holding part of their wealth in liquid means when their knowledge as to how the rate of interest is going to behave in the near future is so limited as to make it impossible to consider some future levels of this rate more probable than others.

This way of interpreting the precautionary motive makes it close to the third motive for demanding monetary instruments identified by Keynes: the speculative motive. Speculation in financial assets occurs because some agents expect with sufficient conviction that the rate of interest will move in a certain direction. The existence of uncertainty (that is, that lack of 'complete knowledge') is not denied. Yet the 'limited knowledge' available allows some agents to consider some future levels of the rate of interest more probable than others. Monetary instruments are so demanded (to avoid a loss in the capital value of financial assets) because a rise in the rate of interest is expected, and not because of the lack of any conviction as to the future of the rate of interest (as in the case of precautionary motive).

The novelty introduced by Keynes (some authors claim that it had been anticipated by Lavington, 1921) lies not in the fact that he recognized that money is also a 'store of value' (an element already present in previous literature), but in the fact that he made some specific sources of demand for monetary instruments depend upon the expected variations of the interest rate, and consequently on the expected variations in the capital value of financial assets.

On account of its magnitude, but principally on account of its high variability, which is due to the uncertain character of expectations about future events, this latter source of demand played a central role in Keynes's writing. It was considered to be *the* cause of variations in the rate of interest. Indeed, in subsequent years, some authors even identified the notion of liquidity preference with speculative motive, while many others put it at the centre of the intense debates on interest rate after the publication of the *General Theory of Employment, Interest and Money* (1936).

Keynes's innovations stimulated many controversies dealing with different aspects of the theory of interest and money. A central point in these debates was the evaluation of Keynes's own contribution: had he really presented a new theory of the rate of interest, alternative to the dominant marginalist one? In the preparatory works and in the *General Theory* itself, Keynes had so characterized his contribution.

He had tried to show the existence of logical inconsistency in the dominant real theory and, in opposition to it, had argued in favour of a *monetary* theory of the rate of interest based on historical and conventional factors.

The essential elements of the analysis of liquidity preference had already been introduced in *A Treatise of Money*, where the marginalist theory determining the ‘natural’ level of the interest rate on the basis of functions of demand for investment and supply of saving was still accepted. Here liquidity preference was integrated within the marginalist theory. In the *General Theory*, instead, the notion of a ‘natural’ interest rate was rejected. The ‘average’ level of the interest rate over a specific period of time was now determined by those factors able to affect the ‘common opinion’ as to the prevailing value of this rate in the future, and among these factors some importance was given to the policy of the monetary authority. Thus, while in *A Treatise on Money* the novelty of liquidity preference referred to the first step of the analysis of the interest rate (that describing how variations in this rate come about), in the *General Theory*, the novelty regarded the second step of analysis, that is the theory determining the level of this rate.

Robertson's critique after the General Theory

The group of economists close to Keynes in those years, with whom he discussed the proofs of the *General Theory*, fully realized that only in this book had he turned the analysis of liquidity preference, already present in the *Treatise*, into a new theory of the interest rate. Not all of them, however, agreed with him. Robertson, brought up in the same Marshallian tradition as Keynes, defended the marginalist theory, claiming that Keynes was in the *General Theory* overstating the role played by monetary factors (see Keynes, 1973a, pp. 499, and Robertson, 1936; 1940). He invited Keynes to attribute to monetary and real forces their proper place, as he had done in *A Treatise on Money*. The abandonment of the ‘forces of productivity and thrift’, when dealing with the determination of the ‘average’ interest rate over long periods of time, left the ‘expected normal value’ of this rate unexplained. Robertson could not accept that ‘the common opinion’ as to the future value of the interest rate should be explained in terms of factors changing from one historical period to the others, rather than by referring to one specific set of factors able to affect the course of events in different historical contexts. If we ask, Robertson stated, ‘what ultimately governs the judgement of wealthowners as to why the rate of interest should be different in the future from what it is today, we are surely led straight back to the fundamental phenomena of productivity and thrift’ (Robertson, 1940, p. 25).

To clarify his view, Robertson translated Keynes's arguments into a different analytical framework based on ‘flow’ concepts. The determination of the ‘market interest rate’ (that actually observed daily) and of the ‘average interest rate’ (the one prevailing over long periods of time) was analysed in terms of ‘loanable funds’, to show that in both cases (but especially in the long-period case) the influence of the demand function for investment and the supply of saving could not be ignored. Within this discussion, Robertson also pointed out the need for extra funds to finance new investment.

The debate with Robertson was intense. Other economists also joined in to discuss the three issues raised: whether Keynes's theory left the determination of the average interest rate ‘hanging in the air’ (or ‘hanging by its own bootstraps’); the role of speculative motive and saving and investment within a ‘loanable-funds’ approach; the ‘finance’ motive.

Hicks and the rise of the ‘neoclassical synthesis’

While the debate with Robertson moved on the common ground of the Marshallian tradition, those with other economists were characterized, from the beginning, by greater problems of understanding and communication.

A major figure in these debates was J.R. Hicks, whose reviews of the *General Theory* (Hicks, 1936; 1937) were discussed with Keynes in an exchange of correspondence (see Keynes, 1973b, pp. 71–83). This correspondence reveals Keynes's insistence on his inability to understand the meaning and the aim of Hicks's claim that the validity of Keynes's theory of interest did not prove other theories to be wrong. Hicks's aim was to integrate Keynes's ideas within an approach, different from the Marshallian one, based on a new version of the neoclassical theory of value which used the notion of *temporary* general equilibria. The rate of interest was determined, with the other distributive variables, relative prices and the level of activity, within an analysis characterized by interdependence between different markets and the simultaneous attainment of equilibrium between supply and demand in all of them. Equilibrium between saving and investment decisions was reached simultaneously with equilibrium between supply of and demand for monetary instruments. The application of ‘Walras's Law’ then made it possible to argue that the claim that the rate of interest is determined in the money market and the claim that it is determined in the market for saving and investment are equivalent.

Hicks's writings had a great impact on the literature. They opened the way to the interpretation of Keynes's work known as the ‘neoclassical synthesis’ and to the wide use of the famous IS–LM apparatus. Indeed, orthodox ‘Keynesian economics’ was derived from this line of development, rather than from Keynes's own writings, as the debate on interest rate shows.

The distinction between the two steps of an analysis of the interest rate was now obscured. In spite of Keynes's explicit claim to the contrary (Keynes, 1937, p. 215), the analysis of liquidity preference, which was intended as a means of describing the market mechanisms through which changes in the interest rate occur, became a theory determining the level of the interest rate. This theory was counter-posed to others – the ‘loanable-funds theory’ and the ‘investment–saving theory’ – in a long debate which in the end established what Hicks had hinted in his reviews of the *General Theory*, that is, that in a general equilibrium analysis to attribute the determination of a price or of a distributive variable to the attainment of equilibrium in one specific market makes no sense.

Now, none of the orthodox Keynesian literature mentioned any more what Keynes had emphasized: the instability of the speculative demand for money due to the uncertain character of the expectations about the future level of the interest rate. The integration of the market for monetary instruments within a general equilibrium analysis requires that the data determining the functions of demand for and supply of money have to be as stable as those determining the demand and supply functions in other markets. The abandonment of Keynes's view of an unstable speculative demand for money was achieved by moving along two lines. First, the notion of an expected normal value of the interest rate was gradually abandoned. Second, the issue of stability was moved from a theoretical to an empirical level.

Already in *Value and Capital* (1939), Hicks had moved along the first line. After him, Modigliani (1944) derived a stable function of demand for money by referring to the risk of future increases in the interest rate, taking this risk as independent of people's specific expectations. The risk is thus *in general* low when the interest rate is high and high when the interest rate is low. Reference to specific

expectations of the future value of interest rate could, instead, make the risk high when the rate is high and low when the rate is low. Finally, Tobin (1958) with the explicit aim of making the theoretical treatment of uncertainty more precise in Keynesian analysis, proposed deriving the demand function for money by including, among the data, subjective probability distributions of the future level of the interest rate, not considering any particular variation in this rate more probable than others. (The similarity with Kahn's precautionary motive mentioned above is clear.) In this analysis, stability of the demand function for money can be achieved by adding one more assumption: any new piece of information acquired by agents does not change their subjective probability distribution. The meaning of this hypothesis is that agents have ‘complete knowledge’ of all relevant information, which amounts to assuming uncertainty away from the analysis. In his subsequent writings, Tobin did not return to this particular point, preferring to consider the issue of ‘stability’ an empirical, rather than a theoretical one. This line has been adopted by most followers of the orthodox Keynesian approach, thus avoiding complex theoretical problems. As a result, the possibility of reaching satisfactory conclusions on this issue appears more difficult.

Theories of the interest rate, which imply a departure from the dominant neoclassical tradition, whether Marshallian or modern general equilibrium versions, can also be found in the literature. They were held by authors close to Keynes during the preparation of the *General Theory*, like Joan Robinson and Kahn, and appear to reflect Keynes's original intentions more than other theories. Robinson and Kahn themselves, in subsequent years (see Robinson, 1937; 1951; Kahn, 1954) contributed to developing these analyses, which were also put forward by Kaldor (1939; 1970; 1982), and re-elaborated by a large group of economists, including Shackle (1967), Pasinetti (1974), Minsky (1975), Davidson (1978), Eatwell (1979), and Garegnani (1979).

Although there are some points of difference between these authors, they seem to agree on the instability of the speculative demand for money due to the uncertain character of the expectations about future level of the interest rate, and on the need to reject the neoclassical theory, for being either analytically inconsistent or for being based on the assumption of a simultaneous achievement of equilibrium in all markets, an assumption which neglects the different ways in which these markets are organized and operate.

The analyses of these authors have contributed to the development of a treatment of monetary issues which breaks with the traditional causal links between ‘monetary’ and ‘real’ variables, and where institutional elements, such as the way financial markets are organized over a certain period of time, play a central role. These analyses make it possible to argue in favour of a ‘monetary’ determination of the interest rate, based on historical and conventional factors, thus supporting Robinson's claim that *any* opinion ‘that is widely believed tends to verify itself, so that there is a large element of “thinking makes it so” in the determination of the interest rates’ (Robinson, 1951, p. 258).

The instability of the financial system and the variability of the interest rates are therefore recognized today, too, by some economists, who also allow for the influence of monetary factors on the level of activity and within the theory of value and distribution, in opposition to the dominant marginalist approach.

See Also

- finance
- Keynes, John Maynard

Bibliography

Davidson, P. 1978. *Money and the Real World*. London: Macmillan.

Eatwell, J.L. 1979. Theories of value, output and employment. In *Keynes's Economics and the Theory of Value and Distribution*, ed. J. Eatwell and M. Milgate. London: Duckworth, 1983.

Garegnani, P. 1979. Notes on consumption, investment and effective demand: II. *Cambridge Journal of Economics* 3(1), 63–82.

Hicks, J.R. 1936. Mr Keynes's theory of employment. *Economic Journal* 46, 238–53.

Hicks, J.R. 1937. Mr Keynes and the ‘classics’; a suggested interpretation. *Econometrica* 5(April), 147–59.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Kahn, R.F. 1954. Some notes on liquidity preference. *Manchester School of Economics and Social Studies* 22, 229–57.

Kaldor, N. 1939. Speculation and economic stability. *Review of Economic Studies* 7(October), 1–27.

Kaldor, N. 1970. The new monetarism. *Lloyds Bank Review* No. 97, July, 1–18.

Kaldor, N. 1982. *The Scourge of Monetarism*. Oxford: Oxford University Press.

Keynes, J.M. 1923. A tract on monetary reform. In *Collected Writings of J.M. Keynes*, vol. 4, ed. D.E. Moggridge. London: Macmillan, 1971.

Keynes, J.M. 1930a. A treatise on money, vol. 1: The pure theory of money. In *Collected Writings of J. M. Keynes*, vol. 5, ed. D.E. Moggridge. London: Macmillan, 1971.

Keynes, J.M. 1930b. A treatise on money, Vol. 2: The applied theory of money. In *Collected Writings of J.M. Keynes*, vol. 6, ed. D.E. Moggridge. London: Macmillan, 1971.

Keynes, J.M. 1936. The general theory of employment, interest and money. In *Collected Writings of J. M. Keynes*, vol. 7, ed. D.E. Moggridge. London: Macmillan, 1973.

Keynes, J.M. 1937. Alternative theories of the rate of interest. *Economic Journal* 47, 241–52. In *The General Theory and After; Part II: Defence and Development, Collected Writings of J.M. Keynes*, vol. 14, ed. D.E. Moggridge, London: Macmillan, 1973.

Keynes, J.M. 1973a. The general theory and after; Part I: Preparation. In *Collected Writings of J.M. Keynes*, vol. 13, ed. D.E. Moggridge. London: Macmillan.

Keynes, J.M. 1973b. The general theory and after; Part II: defence and development. In *Collected Writings of J.M. Keynes*, vol. 14, ed. D.E. Moggridge. London: Macmillan.

Keynes, J.M. 1979. The general theory and after: a supplement. In *Collected Writings of J.M. Keynes*, vol. 19, ed. D.E. Moggridge. London: Macmillan.

Lavington, F. 1921. *The English Capital Market*. London: Methuen.

Minsky, H.P. 1975. *John Maynard Keynes*. New York: Columbia University Press.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12 (January), 45–88.

Pasinetti, L.L. 1974. *Growth and Income Distribution*. Cambridge: Cambridge University Press.

Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32(November), 38–65. Correction (February 1918), 209.

Robertson, D.H. 1936. Some notes on Mr Keynes' 'General Theory of Employment'. *Quarterly Journal of Economics* 51(November), 168–91.

Robertson, D.H. 1940. *Essays in Monetary Theory*. London: P.S. King.

Robinson, J.V. 1937. *Introduction to the Theory of Employment*. London: Macmillan.

Robinson, J.V. 1951. The rate of interest. *Econometrica* 19(April), 92–111. In *Collected Economic Papers of Joan Robinson*, vol. 2, Oxford: Blackwell, 1960.

Shackle, G.L.S. 1967. *The Years of High Theory*. Cambridge: Cambridge University Press.

Tobin, J. 1958. Liquidity preference as behavior risk. *Review of Economic Studies* 25(February), 65–86.

How to cite this article

Panico, Carlo. "liquidity preference." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000114> doi:10.1057/9780230226203.0980

The New Palgrave Dictionary of Economics Online

liquidity trap

Gauti B. Eggertsson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A liquidity trap is defined as a situation in which the short-term nominal interest rate is zero. The old Keynesian literature emphasized that increasing money supply has no effect in a liquidity trap so that monetary policy is ineffective. The modern literature, in contrast, emphasizes that, even if increasing the current money supply has no effect, monetary policy is far from ineffective at zero interest rates. What is important, however, is not the current money supply but managing expectations about the future money supply in states of the world in which interest rates are positive.

Keywords

arbitrage; Bank of Japan; central banks; commitment to optimal policy; deflation bias; Euler equation; expectations; foreign exchange interventions; general equilibrium; Great Depression; incentive compatibility; inflation bias; inflation targeting; inflationary expectations; Keynesianism; Krugman, P.; Kydland, F.; liquidity trap; monetary policy; money supply; New Keynesian Phillips curve; nominal interest rate; output gap; Prescott, E.; public debt; quantity theory of money; real government spending; rules versus discretion; Taylor rule; zero bound (on short-term nominal interest rate)

Article

A liquidity trap is defined as a situation in which the short-term nominal interest rate is zero. In this case, many argue, increasing money in circulation has no effect on either output or prices. The liquidity trap is originally a Keynesian idea and was contrasted with the quantity theory of money, which maintains that prices and output are, roughly speaking, proportional to the money supply.

According to the Keynesian theory, money supply has its effects on prices and output through the nominal interest rate. Increasing money supply reduces the interest rate through a money demand equation. Lower interest rates stimulate output and spending. The short-term nominal interest rate, however, cannot be less than zero, based on a basic arbitrage argument: no one will lend 100 dollars unless she gets at least 100 dollars back. This is often referred to as the ‘zero bound’ on the short-term nominal interest rate. Hence, the Keynesian argument goes, once the money supply has been increased to a level where the short-term interest rate is zero, there will be no further effect on either output or prices, no matter by how much money supply is increased.

The ideas that underlie the liquidity trap were conceived during the Great Depression. In that period the short-term nominal interest rate was close to zero. At the beginning of 1933, for example, the short-term nominal interest rate in the United States – as measured by three-month Treasuries – was only 0.05 per cent. As the memory of the Great Depression faded and several authors challenged the liquidity trap, many economists begun to regard it as a theoretical curiosity.

The liquidity trap received much more attention again in the late 1990s with the arrival of new data. The short-term nominal interest rate in Japan collapsed to zero in the second half of the 1990s. Furthermore, the Bank of Japan (BoJ) more than doubled the monetary base through traditional and non-traditional measures to increase prices and stimulate demand. The BoJ policy of ‘quantitative easing’ from 2001 to 2006, for example, increased the monetary base by over 70 per cent in that period. By most accounts, however, the effect on prices was sluggish at best. (As long as five years after the beginning of quantitative easing, the changes in the CPI and the GDP deflator were still only starting to approach positive territory.)

The modern view of the liquidity trap

The modern view of the liquidity trap is more subtle than the traditional Keynesian one. It relies on an intertemporal stochastic general equilibrium model whereby aggregate demand depends on current and expected future real interest rates rather than simply the current rate as in the old Keynesian models. In the modern framework, the liquidity trap arises when the zero bound on the short-term nominal interest rate prevents the central bank from fully accommodating sufficiently large deflationary shocks by interest rate cuts.

The aggregate demand relationship that underlies the model is usually expressed by a consumption Euler equation, derived from the maximization problem of a representative household. On the assumption that all output is consumed, that equation can be approximated as:

$$Y_t = E_t Y_{t+1} - \sigma(i_t - E_t \pi_{t+1} - r_t^e) \quad (1)$$

where Y_t is the deviation of output from steady state, i_t is the short-term nominal interest rate, π_t is inflation, E_t is an expectation operator and r_t^e is an exogenous shock process (which can be due to host of factors). This equation says that current demand depends on expectations of future output (because spending depends on expected future income) and the real interest rate which is the difference between the nominal interest rate and expected future inflation (because lower real interest rates make spending today relatively cheaper than future spending). This equation can be forwarded to yield

$$Y_t = E_t Y_{T+1} - \sigma \sum_{s=t}^{T} E_t(i_s - \pi_{s+1} - r_s^e)$$

which illustrates that demand depends not only on the current short-term interest rate but on the entire expected path for future interest rates and expected inflation. Because long-term interest rates depend on expectations about current and future short-term rates, this equation can also be interpreted as saying that demand depends on long-term interest rates.

Monetary policy works through the short-term nominal interest rate in the model, and is constrained by the fact that it cannot be set below zero,

$$i_t \geq 0. \quad (2)$$

In contrast to the static Keynesian framework, monetary policy can still be effective in this model even when the current short-term nominal interest rate is zero. In order to be effective, however, expansionary monetary policy must change the public's expectations about future interest rates at the point in time when the zero bound will no longer be binding. For example, this may be the period in which the deflationary shocks are expected to subside. Thus, successful monetary easing in a liquidity trap involves committing to maintaining lower future nominal interest rates for any given price level in the future once deflationary pressures have subsided (see, for example, Reifschneider and Williams, 2000; Jung, Teranishi and Watanabe, 2005; Eggertsson and Woodford, 2003; Adam and Billi, 2006).

This was the rationale for the BoJ's announcement in the autumn of 2003 that it promised to keep the interest rate low until deflationary pressures had subsided and CPI inflation was projected to be in positive territory. It also underlay the logic of the Federal Reserve announcement in mid-2003 that it would keep interest rates low for a ‘considerable period’. At that time, there was some fear of deflation in the United States (the short-term interest rates reached one per cent in the spring of 2003, its lowest level since the Great Depression, and some analysts voiced fears of deflation).

There is a direct correspondence between the nominal interest rate and the money supply in the model reviewed above. There is an underlying demand equation for real money

balances derived from a representative household maximization problem (like the consumption Euler equation 1). This demand equation can be expressed as a relationship between the nominal interest rate and money supply

$$\frac{M_t}{P_t} \geq L(Y_t, i_t) \tag{3}$$

where M_t is the nominal stock of money and P_t is a price level. On the assumption that both consumption and liquidity services are normal goods, this inequality says that the demand for money increases with lower interest rates and higher output. As the interest rate declines to zero, however, the demand for money is indeterminate because at that point households do not care whether they hold money or one-period riskless government bonds. The two are perfect substitutes: a government liability that has nominal value but pays no interest rate. Another way of stating the result discussed above is that a successful monetary easing (committing to lower *future* nominal interest rate for a given price level) involves committing to higher money supply *in the future* once interest rates have become positive again (see, for example, Eggertsson, 2006a).

Irrelevance results

According to the modern view outlined above, monetary policy will increase demand at zero interest rates only if it changes expectations about the future money supply or, equivalently, the path of future interest rates. The Keynesian liquidity trap is therefore only a true trap if the central bank cannot to stir expectations. There are several interesting conditions under which this is the case, so that monetary easing is ineffective. These ‘irrelevance’ results help explain why BoJ's increase in the monetary base in Japan through ‘quantitative easing’ in 2001–6 may have had a somewhat more limited effect on inflation and inflation expectations in that period than some proponents of the quantity theory of money expected.

Krugman (1998), for example, shows that at zero interest rates if the public expects the money supply in the future to revert to some constant value as soon as the interest rate is positive, quantitative easing will be ineffective. Any increase in the money supply in this case is expected to be reversed, and output and prices are unchanged.

Eggertsson and Woodford (2003) show that the same result applies if the public expects the central bank to follow a ‘Taylor rule’, which may indeed summarize behaviour of a number of central banks in industrial countries. A central bank following a Taylor rule raises interest rates in response to above-target inflation and above-trend output. Conversely, unless the zero bound is binding, the central bank reduces the interest rate if inflation is below target or output is below trend (an output gap). If the public expects the central bank to follow the Taylor rule, it anticipates an interest rate hike as soon as there are inflationary pressures in excess of the implicit inflation target. If the target is perceived to be price stability, this implies that quantitative easing has no effect, because a commitment to the Taylor rule implies that any increase in the monetary base is reversed as soon as deflationary pressures subside.

Eggertsson (2006a) demonstrates that, if a central bank is discretionary, that is, unable to commit to future policy, and minimizes a standard loss function that depends on inflation and the output gap, it will also be unable to increase inflationary expectations at the zero bound, because it will always have an incentive to renege on an inflation promise or extended ‘quantitative easing’ in order to achieve low *ex post* inflation. This deflation bias has the same implication as the previous two irrelevance propositions, namely, that the public will expect any increase in the monetary base to be reversed as soon as deflationary pressures subside. The deflation bias can be illustrated by the aid of a few additional equations, as illustrated in the next section.

The deflation bias and the optimal commitment

The deflation bias can be illustrated by completing the model that gave rise to (1), (2) and (3). In the model prices are not flexible because firms reset their price at random intervals. This gives rise to an aggregate supply equation which is often referred to as the ‘New Keynesian’ Phillips curve. It can be derived from the Euler equation of the firm's maximization problem (see, for example, Woodford, 2003)

$$\pi_t = \kappa(Y_t - Y_t^n) + \beta E_t \pi_{t+1} \tag{4}$$

where Y_t^n is the natural rate of output (in deviation from steady state), which is the ‘hypothetical’ output produced if prices were perfectly flexible, β is the discount factor of the household in the model and the parameter $\kappa > 0$ is a function of preferences and technology parameters. This equation implies that inflation can increase output above its natural level because not all firms reset their prices instantaneously.
If the government's objective is to maximize the utility of the representative household, it can be approximated by

$$\sum_{t=0}^{\infty} \beta^t \left\{ \pi_t^2 + \lambda_y (Y_t - Y_t^e)^2 \right\} \tag{5}$$

where the term Y_t^e is the target level of output. It is also referred to as the ‘efficient level’ or ‘first-best level’ of output. The standard ‘inflation bias’ first illustrated by Kydland and Prescott (1977) arises when the natural level of output is lower than the efficient level of output, that is, $Y_t^n < Y_t^e$.
Eggertsson (2006a) shows that there is also a deflation bias under certain circumstances. While the inflation bias is a steady state phenomenon, the deflation bias arises to temporary shocks. Consider the implied solution for the nominal interest rate when there is an inflation bias of $\bar{\pi}$. It is

$$i_t = \bar{\pi} + r_t^e.$$

This equation cannot be satisfied in the presence of sufficiently large deflationary shocks, that is, a negative r_t^e. In particular if $r_t^e < -\bar{\pi}$ this solution would imply a negative nominal interest rate. It can be shown (Eggertsson, 2006a) that a discretionary policymaker will in this case set the nominal interest rate to zero but set inflation equal to the ‘inflation bias’ solution $\bar{\pi}$ as soon as the deflationary pressures have subsided (that is, when the shock is $r_t^e \geq -\bar{\pi}_t$). If the disturbance r_t^e is low enough, the zero bound frustrates the central bank's ability to achieve its ‘inflation target’ $\bar{\pi}$ which can in turn lead to excessive deflation. (While deflation and zero interest rates are due to real shocks in the literature discussed above, an alternative way of modelling the liquidity trap is that it is the result of self-fulfilling deflationary expectations; see, for example, Benhabib, Schmitt-Grohe and Uribe, 2001.)
To illustrate this consider the following experiment. Suppose the term r_t^e is unexpectedly negative in period 0 $(r_t^e = r_L < 0)$ and then reverts back to its steady state value $\bar{r} > 0$ with a fixed probability α in every period. For simplicity assume that $\bar{\pi} = 0$. Then it is easy to verify from eqs. (1), (4), the behaviour of the central bank described above and the assumed process for r_t^e that the solution for output and inflation is given by (see Eggertsson, 2006a, for details)

$$\pi_t = \frac{1}{\alpha(1 - \beta(1-\alpha)) - \sigma\kappa(1-\alpha)} \kappa\sigma r_L^e \text{ if } r_t^e = r_L^e \text{ and } \pi_t = 0 \text{ otherwise} \tag{6}$$

$$Y_t = \frac{1-\beta(1-\alpha)}{\alpha(1-\beta(1-\alpha)) - \sigma\kappa(1-\alpha)}\sigma r_L^e \text{ if } r_t^e = r_L^e \text{ and } Y_t = 0 \text{ otherwise} \tag{7}$$

Figure 1 shows the solution in a calibrated example for numerical values of the model taken from Eggertsson and Woodford (2003). (Under this calibration $\alpha = 0.1$, $\kappa = 0.02$, $\beta = 0.99$ and $r_L = -\frac{0.02}{4}$ but the model is calibrated in quarterly frequencies.) The dashed line shows the solution under the contingency that the natural rate of interest reverts to positive level in 15 periods. The inability of the central bank to set negative nominal interest rate results in a 14 per cent output collapse and 10 per cent annual deflation. The fact that in each quarter there is a 90 per cent chance of the exogenous disturbance to remaining negative for the next quarter creates the expectation of future deflation and a continued output depression, which creates even further depression and deflation. Even if the central bank lowers the short-term nominal interest rate to zero, the real rate of interest is positive, because the private sector expects deflation. The same results applies when there is an inflation bias, that is, $\bar{\pi} > 0$, but in this case the disturbance r_t^e needs to be correspondingly more negative to lead to an output collapse.

Figure 1

Response of the nominal interest rate, inflation and the output gap to a shocks that lasts for 15 quarters. *Note*: The dashed line shows the solution under policy discretion, the solid line the solution under the optimal policy commitment.

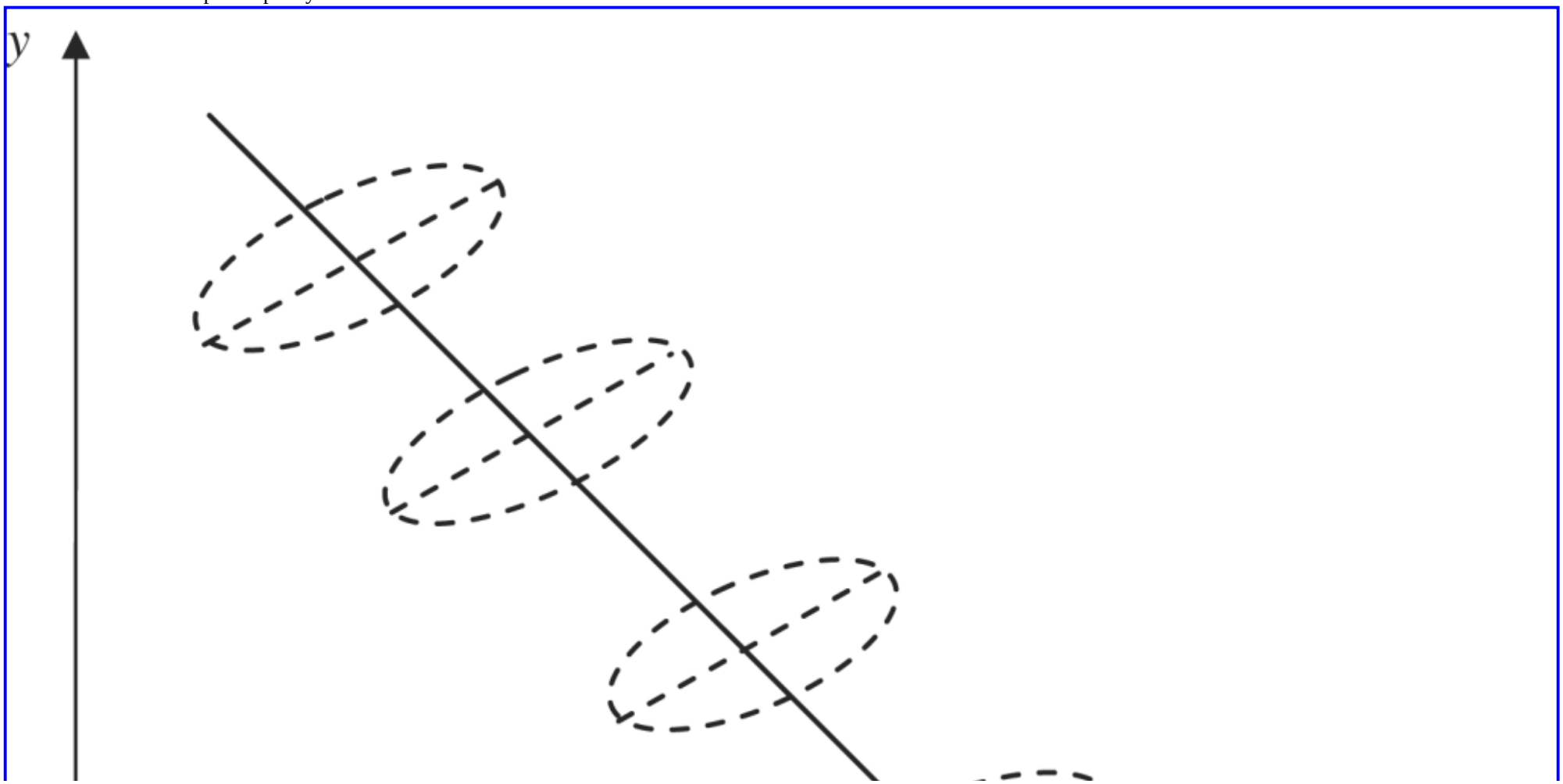

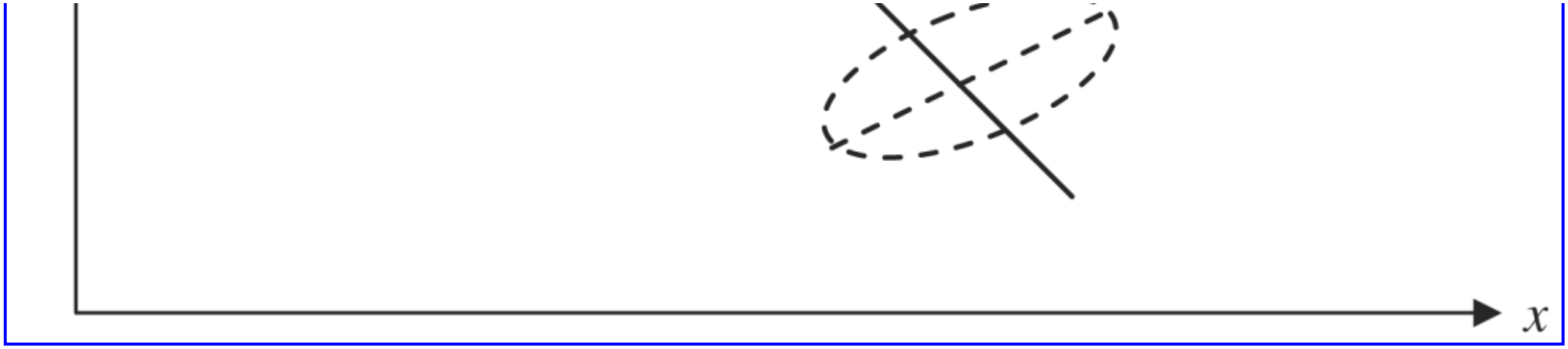

The solution illustrated in Figure 1 is what Eggertsson (2006a) calls the deflation bias of monetary policy under discretion. The reason why this solution indicates a deflation bias is that the deflation and depression can largely be avoided by the correct *commitment* to optimal policy. The solid line shows the solution in the case that the central bank can commit to optimal future policy. In this case the deflation and the output contraction are largely avoided. In the optimal solution the central bank commits to keeping the nominal interest at zero for a considerable period beyond what is implied by the discretionary solution; that is, interest rates are kept at zero even if the deflationary shock r_t^e has subsided. Similarly, the central bank allows for an output boom once the deflationary shock subsides and accommodates mild inflation. Such commitment stimulates demand and reduces deflation through several channels. The expectation of future inflation lowers the real interest rate, even if the nominal interest rate cannot be reduced further, thus stimulating spending. Similarly, a commitment to lower future nominal interest rate (once the deflationary pressures have subsided) stimulates demand for the same reason. Finally, the expectation of higher future income, as manifested by the expected output boom, stimulates current spending, in accordance with the permanent income hypothesis (see Eggertsson and Woodford, 2003, for the derivation underlying this figures. The optimal commitment is also derived in Jung, Teranishi and Watanabe, 2005, and Adam and Billi, 2006, for alternative processes for the deflationary disturbance).

The discretionary solution indicates that this optimal commitment, however desirable, is not feasible if the central bank cannot commit to future policy. The discretionary policymaker is cursed by the deflation bias. To understand the logic of this curse, observe that the government's objective (5) involves minimizing deviations of inflation and output from their targets. Both these targets can be achieved at time $t = 15$ when the optimal commitment implies targeting positive inflation and generating an output boom. Hence the central bank has an incentive to renege on its previous commitment and achieve zero inflation and keep output at its optimal target. The private sector anticipates this, so that the solution under discretion is the one given in (6) and (7); this is the deflation bias of discretionary policy.

Shaping expectations

The lesson of the irrelevance results is that monetary policy is ineffective if it cannot stir expectations. The previous section illustrated, however, that shaping expectations in the correct way can be very important for minimizing the output contraction and deflation associated with deflationary shocks. This, however, may be difficult for a government that is expected to behave in a discretionary manner. How can the correct set of expectations be generated?

Perhaps the simplest solution is for the government to make clear *announcements* about its future policy through the appropriate ‘policy rule’. This was the lesson of the ‘rules vs. discretion’ literature started by Kydland and Prescott (1977) to solve the inflation bias, and the same logic applies here even if the nature of the ‘dynamic inconsistency’ that gives rise to the deflation bias is different from the standard one. To the extent that announcements about future policy are believed, they can have a very big effect. There is a large literature on the different policy rules that minimize the distortions associated with deflationary shocks. One example is found in both Eggertsson and Woodford (2003) and Wolman (2005). They show that, if the government follows a form of price level targeting, the optimal commitment solution can be closely or even completely replicated, depending on the sophistication of the targeting regime. Under the proposed policy rule the central bank commits to keep the interest rate at zero until a particular price level is hit, which happens well after the deflationary shocks have subsided.

If the central bank, and the government as a whole, has a very low level of credibility, a mere announcement of future policy intentions through a new ‘policy rule’ may not be sufficient. This is especially true in a deflationary environment, for at least three reasons. First, the deflation bias implies that the government has an incentive to promise to deliver future expansion and higher inflation, and then to renege on this promise. Second, the deflationary shocks that give rise to this commitment problem are rare, and it is therefore harder for a central bank to build up a reputation for dealing with them well. Third, this problem is even further aggravated at zero interest rates because then the central bank cannot take

any direct actions (that is, cutting interest rate) to show its new commitment to reflation. This has led many authors to consider other policy options for the government as a whole that make a reflation credible, that is, make the optimal commitment described in the previous section ‘incentive compatible’.

Perhaps the most straightforward way to make a reflation credible is for the government to issue debt, for example by deficit spending. It is well known in the literature that government debt creates an inflationary incentive (see, for example, Calvo, 1978). Suppose the government promises future inflation and in addition prints one dollar of debt. If the government later reneges on its promised inflation, the real value of this one dollar of debt will increase by the same amount. Then the government will need to raise taxes to compensate for the increase in the real debt. To the extent that taxation is costly, it will no longer be in the interest of the government to renege on its promises to inflate the price level, even after deflationary pressures have subsided in the example above. This commitment device is explored in Eggertsson (2006a), which shows that this is an effective tool to battle deflation.

Jeanne and Svensson (2007) and Eggertsson (2006a) show that foreign exchange interventions also have this effect, for very similar reasons. The reason is that foreign exchange interventions change the balance sheet of the government so that a policy of reflation is incentive compatible. The reason is that, if the government prints nominal liabilities (such as government bonds or money) and purchases foreign exchange, it will incur balance-sheet losses if it reneges on an inflation promise because this would imply an exchange rate appreciation and thus a portfolio loss.

There are many other tools in the arsenal of the government to battle deflation. Real government spending, that is, government purchases of real goods and services, can also be effective to this end (Eggertsson, 2005). Perhaps the most surprising one is that policies that temporarily reduce the natural level of output, Y_t^n, can be shown to increase equilibrium output (Eggertsson, 2006b). The reason is that policies that suppress the natural level of output create actual and expected reflation in the price level and this effect is strong enough to generate recovery because of the impact on real interest rates.

Conclusion: the Great Depression and the liquidity trap

As mentioned in the introduction, the old literature on the liquidity trap was motivated by the Great Depression. The modern literature on the liquidity trap not only sheds light on recent events in Japan and the United States (as discussed above) but also provides new insights into the US recovery from the Great Depression. This article has reviewed theoretical results that indicate that a policy of reflation can induce a substantial increase in output when there are deflationary shocks (compare the solid line and the dashed line in Figure 1: moving from one equilibrium to the other implies a substantial increase in output). Interestingly, Franklin Delano Roosevelt (FDR) announced a policy of reflating the price level in 1933 to its pre-Depression level when he became President in 1933. To achieve reflation FDR not only announced an explicit objective of reflation but also implemented several policies which made this objective credible. These policies include all those reviewed in the previous section, such as massive deficit spending, higher real government spending, foreign exchange interventions, and even policies that reduced the natural level of output (the National Industrial Recovery Act and the Agricultural Adjustment Act: see Eggertsson, 2006b, for discussion). As discussed in Eggertsson (2005; 2006b) these policies may greatly have contributed to the end of the depression. Output increased by 39 per cent during 1933–7, with the turning point occurring immediately after FDR's inauguration, when he announced the policy objective of reflation. In 1937, however, the administration moved away from reflation and the stimulative policies that supported it – prematurely declaring victory over the depression – which helps explaining the downturn in 1937–8, when monthly industrial production fell by 30 per cent in less than a year. The recovery resumed once the administration recommitted to reflation (see Eggertsson and Puglsey, 2006). The modern analysis of the liquidity trap indicates that, while zero short-term interest rates made static changes in the money supply irrelevant during this period, expectations about the future evolution of the money supply and the interest rate were key factors determining aggregate demand. Thus, recent research indicates that monetary policy was far from being ineffective during the Great Depression, but it worked mainly through expectations.

See Also

- expectations
- inflation expectations
- optimal fiscal and monetary policy (with commitment)
- optimal fiscal and monetary policy (without commitment)

Bibliography

Adam, K. and Billi, R. 2006. Optimal monetary policy under commitment with a zero bound on nominal interest rates. *Journal of Money, Credit and Banking* (forthcoming).

Benhabib, J., Schmitt-Grohe, S. and Uribe, M. 2001. Monetary policy and multiple equilibria. *American Economic Review* 91, 167–86.

Calvo, G. 1978. On the time consistency of optimal policy in a monetary economy. *Econometrica* 46, 1411–28.

Eggertsson, G. 2005. Great expectations and the end of the depression. Staff Report No. 234. Federal Reserve Bank of New York.

Eggertsson, G. 2006a. The deflation bias and committing to being irresponsible. *Journal of Money, Credit and Banking* 38, 283–322.

Eggertsson, G. 2006b. Was the New Deal contractionary? Working paper, Federal Reserve Bank of New York.

Eggertsson, G. and Pugsley, B. 2006. The mistake of 1937: a general equilibrium analysis. *Monetary and Economic Studies* 24(SI), 151–90.

Eggertsson, G. and Woodford, M. 2003. The zero bound on interest rates and optimal monetary policy. *Brookings Papers on Economic Activity* 2003(1), 212–19.

Jeanne, O. and Svensson, L. 2007. Credible commitment to optimal escape from a liquidity trap: the role of the balance sheet of an independent central bank. *American Economic Review* 97, 474–90.

Jung, T., Teranishi, Y. and Watanabe, T. 2005. Zero bound on nominal interest rates and optimal monetary policy. *Journal of Money, Credit and Banking* 37, 813–36.

Krugman, P. 1998. It's baaack! Japan's slump and the return of the liquidity trap. *Brookings Papers on Economic Activity* 1998(2), 137–87.

Kydland, F. and Prescott, E. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

Reifschneider, D. and Williams, J. 2000. Three lessons for monetary policy in a low inflation era. *Journal of Money, Credit and Banking* 32, 936–66.

Wolman, A. 2005. Real implications of the zero bound on nominal interest rates. *Journal of Money, Credit and Banking* 37, 273–96.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

Eggertsson, Gauti B. "liquidity trap." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000237> doi:10.1057/9780230226203.0981

The New Palgrave Dictionary of Economics Online

List, Friedrich (1789 – 1846)

K. Tribe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic nationalism; infant-industry protection; international division of labour; List, F.; national economy; productive powers; Zollverein

Article

Known chiefly as a proponent of economic nationalism and protection to ‘infant industries’, List's career followed a colourful, not to say disorderly course, from his engagement on behalf of a customs union in the early 1820s to exile and residence in the United States, agitation on behalf of railway construction, energetic economic journalism, and finally to his death by suicide in November 1846, depressed by his lack of success in promoting a commercial agreement between Prussia and Britain and also by chronic financial insecurity. Born into the family of a tanner on or about 6 August 1789 in Reutlingen, Württemberg, List's early life was unremarkable. After briefly working in his father's business, he entered service in the state administration as a clerk and in 1811 secured a position in Tübingen. There he began attending the occasional law lecture, giving up his appointment in 1813 to concentrate on his legal studies. He never sat for the final lawyers’ examination, instead taking and passing the actuaries’ examination in September 1814.

Re-entering the administration as an accountant, he was promoted in 1816 to the position of Chief Examiner of Accounts. At the same time he became involved in the publication of a reformist journal, contributing articles on the reform of local administration. Through his connections in Stuttgart he also became involved in proposals for the creation of a new faculty for state economy at the University of Tübingen; teaching began in January 1818 and List was appointed full professor of administrative practice. List seems to have made little effort to compensate for his lack of formal academic qualification for the post, and he was dismissed in mid-1819 for absenteeism.

It is at this point that List's ‘life’ begins; for it transpired that his absence during April 1819 was on account of his attendance at the founding meeting of the German Association for Trade and Commerce, a body dedicated to the abolition of internal barriers to trade and which appointed List consular secretary. During the following year List travelled on behalf of the Association, and was also elected to

the Württemberg representative assembly as Deputy for Reutlingen. As a result of his activities in the latter role he was tried and sentenced for sedition in 1822; appealing from the sanctuary of Baden, he failed to get the verdict altered and began a life of exile, travelling in 1823 to Paris where he made the acquaintance of Lafayette. In May 1824, believing that he had been reprieved, List returned to Stuttgart, was promptly imprisoned and, in January 1825, exiled.

Acting on a suggestion of Lafayette, List set sail for America with his family in April 1825. Taking advantage of a tour that Lafayette was undertaking at the time, List travelled and studied, making the acquaintance of several leading political figures. Settling in Pennsylvania, where he briefly tried his hand at farming, he assumed in 1826 the editorship of a German-language newspaper, the *Readinger Adler*, and became closely associated with the Pennsylvania Society for the Encouragement of Manufactures and Mechanic Arts. Through this involvement he became a supporter of the ‘American system’ of protective tariffs, and published in late 1827 his first serious economic work, *Outlines of American Political Economy*, which was a critique of Thomas Cooper's free-trade *Elements of Political Economy*. Such was the success of this that the Pennsylvania Society asked List to write a school textbook on political economy, but only the first chapter of this work was ever written.

As a result of an interest in coal deposits List became involved in a railway construction company which eventually opened its railroad in 1831. By this time, List had supported the presidential campaign of Jackson in 1828 and had become an American citizen; he returned to Europe, settling there permanently in late 1832 and in 1834 was appointed American Consul in Leipzig. There he became involved with the construction of the Leipzig–Dresden railway and founded the *Eisenbahnjournal* (1835), but he parted with his fellow projectors in 1837 and moved to Paris, where he spent the next three years writing a prize essay and pursuing various journalistic projects.

After his period in Paris, he moved to Augsburg and then resumed his agitation on behalf of German economic unity and south German protectionism. As before, this was largely conducted through the medium of newspapers, one of these being the *Zollvereinsblatt*, founded in 1843. These last restless years brought literary success with the publication of his *National System of Political Economy*, but little effective influence on the formation of contemporary commercial policy.

List's contribution to the formation of the Zollverein was limited to the period between 1819 and 1820, when he travelled German courts representing the cause of tariff reform. His theoretical proposals concerning protection and ‘infant industries’ date from his American period and are indeed a direct result of his American experience of tariff debates in the later 1820s. Much of his writing is repetitive of simple themes, as one would expect of work produced in haste for newspapers, journals and pamphlets arguing for specific reforms. However, the general logic of his position can be summarized in the following terms.

The Smithian principle of ‘natural liberty’ and commercial freedom was a ‘cosmopolitan doctrine’ which erroneously generalized the situation of Britain to the rest of the world. Commercial freedom in this sense was a freedom for Britain to dominate the world economy, thanks to the degree of development of the British economy. Free trade and economic liberty were highly desirable for a true world economy, but were only appropriate to a world of economic equals. Such a world could be created only if those countries which were in the process of development could protect their key industries against premature competition. On the international front it was necessary to create a system of treaties and agreements which would regulate trade and competition in such a way that protective tariffs and other protectionist measures would one day be redundant. On the national level, it was important to

abolish internal limitations to development, such as duties between German states which hindered trade and communication. A powerful device for the creation of strong national economies was the railway, perceived not so much for its freight capacity as for its role in promoting the freedom of movement of active populations. While the abolition of internal duties opened up the fiscal geography of an economy, this space was to be given shape by a railway network which would link major centres of population – and it is this emphasis on a communications *network* that distinguishes List's work in the 1830s.

List's writing on railway development is scattered in several articles and was never presented systematically, but his conception of economic liberty and world economic development is developed in the two books he published, and the prize essay which he wrote in Paris. His *Outlines of American Political Economy* clearly contrasts a ‘Smithian’ economy of individuals and of mankind with ‘national economy’. The error of Smith was to believe that the promotion of ‘individual economy’ – the satisfaction of individual wants – would lead to ‘the economy of mankind’ or cosmo-political economy – securing the necessities and comforts of life to the whole human race. List argued that this would not happen; the true path to the economy of mankind lay through national economy, the consideration of measures and conditions appropriate to actually existing nations. The general laws of economics outlined by Smith and his followers could manifest themselves only through these nations, which necessarily modified the operation of these laws by force of their specific ‘productive powers’. The strength and independence of a national economy was secured through the control of the interior market, enabling the economy to flourish on the basis of its natural and human endowments.

The Natural System of Political Economy was written in 1837 as a response to questions concerning the ways of reconciling the interests of producers and consumers on the introduction of commercial liberty. This recapitulates the argument on individual and cosmopolitan economy already developed in the *Outlines*, but goes further in elaborating a general theory of economic development as a series of stages of agricultural, manufacturing and commercial activity. While the first stage involves a basic reliance on agriculture, by the fourth and final stage raw materials are imported for manufacture and re-export, while food is also imported.

The National System of Political Economy was published in 1841 and represents a rounding out of arguments already exposed in his earlier writing. Importantly, List now placed his arguments in a general conception of the civilizing process of international trade, underlining the fact that his opposition to free trade was by no means a narrowly nationalistic one. Also added to the original arguments is a conception of the international division of labour elaborated on the basis of the distinction of manufacture and agriculture. List divided the world into temperate zones naturally oriented towards manufacture, and hot zones with a natural advantage in the production of agricultural goods. A balanced development of the world economy, or in other words the civilizing process, requires that the nations in the temperate zone be in equilibrium with each other and that they neither singly nor jointly exploit the lands of the hot zone, which would otherwise become dependent on manufacturing powers.

Much of the *National System* is given over to a historical account of economic development which today is very dated, while List's critique of classical economics is likewise limited by the primarily non-academic readership to which he appealed. Nonetheless, his emphasis on productive powers rather than ‘value and capital’, and his insistence on the specificity of national endowments and conditions in considering world economic development remain of interest. While List's primary interest lay in political and economic reform, and his audience was emergent ‘informed popular opinion’, he nevertheless developed conceptions of economic space and economic development that have lasting

intellectual merits.

Selected works

1827. *Outlines of American Political Economy*. Reprinted in *The Life of Friedrich List and Selection from his Writings*, ed. M.E. Hirst. London: Smith, Elder & Co., 1909.

1837. *The Natural System of Political Economy*. Trans. and ed. W.O. Henderson, London: Cass, 1983.

1841. *Das nationale System der politischen Oekonomie*. Stuttgart and Tübingen: J.G. Cotta. Trans. G.A. Matile as *National System of Political Economy*, Philadelphia: J.B. Lippincott & Co., 1856. Trans. S.S. Lloyd as *The National System of Political Economy*, London: Longmans & Co., 1885.

1927–36. *Friedrich List. Schriften, Reden, Briefe*, 10 vols. Berlin.

How to cite this article

Tribe, K. "List, Friedrich (1789–1846)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000118> doi:10.1057/9780230226203.0982

The New Palgrave Dictionary of Economics Online

litigation, economics of

Kathryn E. Spier
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article begins by introducing the basic economic framework for studying litigation and out-of-court settlement. One set of issues addressed is positive (or descriptive) in nature. Under what conditions will someone decide to file suit? When do cases settle out of court? Normative issues are also addressed. Are these private litigation decisions in the interest of society more broadly? Next, the article surveys some of the more active areas in the litigation literature including rules of evidence, loser-pays rules, appeals, contingent fees for attorneys, alternative dispute resolution, class actions, and plea bargaining.

Keywords

adverse selection; alternative dispute resolution; asymmetric information; bargaining; contingent fees; dispute resolution; English Rule; Environmental Protection Agency; insurance contracts; liability; litigation, economics of; moral hazard; most-favoured-nation clauses; negative expected value; patents; Posner, R.; Superfund; symmetric information; transaction costs

Article

Litigation refers to the process of taking an argument to a court of law where a decision will be made. The discipline of economics has provided researchers – economists and legal scholars alike – with useful tools and frameworks for thinking about litigation. Is there too much litigation or too little? Why do some lawsuits go to trial while many others settle before trial? Should the losing party be required to reimburse the winning party's legal expenses? The first part of this article presents the main frameworks for studying the economics of litigation. The second part surveys just some of the active topics in the literature.
This article is largely a condensed version of Spier (2005). Previous surveys of this topic include Cooter and Rubinfeld (1989), Hay and Spier (1998), and Daughety (2000).

Basic framework

The decision to litigate

Suppose there are two *litigants*: one *plaintiff* and one *defendant*. The plaintiff is the injured party who seeks compensation; the defendant is the party who is potentially responsible for the plaintiff's injuries. A plaintiff will rationally choose to bring suit when the expected gross return from litigation, x, exceeds the cost of pursuing the case, c_p. The gross return, x, represents the expected judgment at the end of a long and costly trial or a settlement that takes place at some time prior to the trial. It could also reflect other issues, such as the impact that a court decision will have on future cases or the plaintiff's concern for her business reputation. In general, the plaintiff's cost of pursuing the case, c_p, and the defendant's cost of fighting back, c_d, would influence the gross return, x, and could be modelled in a similar way to other economic contests (Dixit, 1987). For the moment, however, we will treat them as exogenous.
The plaintiff's incentive to bring suit typically diverges from what is best for society as a whole (see Shavell, 1982b; 1997). Consider a situation where accidents are totally avoided if the defendant makes a small investment in precautions. If the plaintiff were expected to sue following an accident, the defendant would rationally take the precautions. No accidents would occur and no litigation costs would be incurred. If $c_p > x$, however, then the plaintiff lacks a credible threat to sue. Knowing this, the defendant has no incentive to take the precautions (however inexpensive). In this example, the plaintiff's private incentive to sue is *socially insufficient*. This is not always the case, however. Suppose that the defendant's investment is totally ineffective: accidents occur whether or not the defendant takes precautions. Following an accident, the plaintiff will sue the defendant when $c_p < x$. The plaintiff's incentive to bring suit is *socially excessive* in this example. Litigation is a socially wasteful activity here because there is nothing the defendant could have done to avoid the accident.

Settlement

Not surprisingly, the overwhelming majority of lawsuits settle before trial. (Fewer than four per cent of civil cases that are filed in the US State Courts go to trial; see Ostrom, Kauder and La Fountain, 2001, p. 29). To use our earlier notation, the plaintiff will receive a net payoff of $x - c_p$ if the case goes to court and the defendant will receive $-x - c_d$. Although x represents a simple transfer from the defendant to the plaintiff, the litigation costs, $c_p + c_d$, represent a deadweight loss. Any out-of-court transfer $S \in (x - c_p, x + c_d)$ from the defendant to the plaintiff would be a Pareto improvement. The precise outcome of settlement negotiations will hinge on a variety of factors, including the timing of offers and counteroffers, the information and beliefs of the two litigants, and the nature of the broader legal and strategic environment.

Settlement with symmetric information

Suppose that the litigants are symmetrically informed and play an alternating-offer game with T–1 rounds of bargaining before trial in round T. At trial, the defendant pays x to the plaintiff and the

litigation costs, c_p and c_d, are incurred. The litigants share a common discount factor, δ.

This game is easily solved by backwards induction. Suppose that the plaintiff is designated to make the last settlement offer in period *T*–1. The defendant will accept any offer that is better than going to trial, so the plaintiff will offer $S_{T-1} = \delta(x + c_d)$, minus a penny perhaps. If the case hasn't settled earlier, it will certainly settle on the courthouse steps. If we work backwards, the litigants are willing to settle for $S_{T-2} = \delta^2(x + c_d)$ in period *T*–2, and (by an extension of this logic) are willing to settle for $S_1 = \delta^{T-1}(x + c_d)$ in period 1.

Two observations about this example are in order. First, the allocation of the bargaining surplus is sensitive to the timing of the settlement offers. If the defendant were the one to make the last offer instead, then the case would settle for $S_{T-1} = \delta(x - c_p)$ in the last round and, working backwards, we would have $S_1 = \delta^{T-1}(x - c_p)$. In other words, the party who makes the last offer succeeds in extracting all of the bargaining surplus. The bargaining surplus would, of course, be more evenly allocated in a random-offer or framework where the two litigants flip a coin to determine who makes an offer.

Second, this simple example does not predict exactly *when* settlement will take place. The litigants are, in fact, indifferent between settling for $S_1 = \delta^{T-1}(x + c_d)$ in period 1 and for $S_{T-1} = \delta(x + c_d)$ on the courthouse steps. The reason for this is straightforward: there is no inefficiency associated with delay when the litigation costs are entirely borne at trial. (Settlement models differ from the related models of bilateral trade. There, discounting causes the pie to shrink. Here, discounting by itself does not affect the size of the pie.) If the costs of litigation were incurred gradually over time instead, so the first *T*–1 rounds of bargaining were costly as well, then there would be a unique subgame-perfect equilibrium with settlement in period *1* (Bebchuk, 1996).

Settlement with asymmetric information

Asymmetric information is common in litigation settings. Plaintiffs often have first-hand knowledge about the damages they have suffered; defendants often have first-hand knowledge about their degree of involvement in the accident. Litigants also receive private signals concerning the credibility of their witnesses and the quality and work ethic of their lawyers. Some of this information will become commonly known over time – the parties surely learn a great deal through pretrial proceedings and discovery. Other information may never come to light at all, but can nevertheless affect trial outcomes. Suppose that the defendant has private information about *x*, the expected judgment at trial. A similar analysis would follow if the plaintiff were privately informed instead. Formally, suppose *x* drawn from a nicely behaved probability density function *f*(*x*) on $[\underline{x}, \overline{x}]$ with cumulative density *F*(*x*). Starting with P'ng (1983) and Bebchuk (1984), many papers assume that the uninformed player – the plaintiff in our example – makes a single take-it-or-leave-it settlement offer, *S*, before trial. The defendant accepts *S* if it is lower than what he would expect to pay at trial, $S < \delta(x + c_d)$. The offer generates a ‘cut-off,’ $\hat{x} = \delta^{-1}S - c_d$, where defendant types above the cut-off accept the offer and those below the cut-off reject the offer and go to court.

The plaintiff's optimization problem may be written as a function of the cutoff, $\hat{x}$:

$\underset{\hat{x}}{Max}\int_{\underline{x}}^{\hat{x}}\delta(x-c_p)f(x)dx+[1-F(\hat{x})]\delta(\hat{x}+c_d)$. The first term represents the plaintiff's net payoff associated with those types who reject the settlement offer, and the second term reflects the settlement payments from the defendant types above the cut-off, $\hat{x}$, who accept the offer. Any interior solution is characterized by the following first-order condition:

$$1-F(\hat{x})-(c_p+c_d)f(\hat{x})=0.$$

At least some cases will settle – the plaintiff will certainly make a settlement offer that is accepted by the most liable defendants – and an interior solution exists when (c_p+c_d) is not too high.

Bebchuk's basic model has been extended in a variety of ways. Nalebuff (1987) argues that the plaintiff may no longer have a credible commitment to take the case to trial following the rejection of the settlement offer, and explicitly incorporates a credibility constraint. Spier (1992) allows the plaintiff to make a sequence of settlement offers before trial. When litigation costs are all borne at trial (so there is no efficiency loss from delay), the plaintiff waits until the very last moment to offer $S_{T-1}=\delta(\hat{x}+c_d)$, where $\hat{x}$ is defined above. (The deadline effect is less pronounced when there are pretrial costs as well.)

Reinganum and Wilde (1986) let the informed litigant made a single take-it-or-leave-it offer before trial and characterize a perfect Bayesian equilibrium – unique under the D1 refinement of Cho and Kreps (1987). The defendant's equilibrium offer $S(x)=\delta(x-c_p)$ perfectly reveals his type. Making the correct inference, the plaintiff is indifferent and accepts the settlement offer with probability

$$\pi(x)=e^{-(\bar{x}-x)/(c_p+c_d)}.$$

Note that this probability is increasing in the defendant's expected liability, *x*. This is implied by incentive compatibility; the defendant must be rewarded in equilibrium for making higher settlement offers with a higher rate of acceptance by the plaintiff.

Some scholars have used mechanism-design techniques to study settlement and have shown, among other things, that some cases will *necessarily* go to trial when the litigation costs are not too large (Spier, 1994a). In contrast to Myerson and Satterthwaite's (1983) analysis of bilateral trade, settlement bargaining breaks down with one-sided incomplete information and despite common knowledge that gains from trade exist. (Schweizer, 1989, and Daughety and Reinganum, 1994, explore extensive form games with two-sided asymmetric information.) Finally, it is important to mention an older literature where litigants have different priors about the outcome at trial. Landes (1971), Posner (1973), and Gould (1973) show that settlement negotiations may fail when the two sides are sufficiently optimistic. (See Loewenstein et al., 1993, for empirical evidence on self-serving biases.)

Normative implications

There are strong normative arguments in favour of settlement. Through a private settlement, the parties can avoid their litigation costs and (if they are risk averse) the risk premium associated with trials. *All else equal*, private settlement serves society's interest. What makes this topic more interesting – and sometimes exceptionally challenging – is that *all else is not equal*. First, settlement dilutes a defendant's incentives to avoid accidents. Following an accident, the defendant is better off if he has the option to settle his claim. Anticipating settlement on relatively advantageous terms, the defendant has less incentive to take precautions to avoid the lawsuit to begin with (Polinsky and Rubinfeld, 1988). (This not necessarily a bad thing: when cases settle out of court the litigations costs are avoided so the social cost of an accident is lower. Therefore, the defendant *should* be taking less care than if all cases went to trial.) Spier (1997) shows that the defendant's incentives are diluted even further if the defendant has private information. Second, the plaintiff is made better off through settlement than she would be going to trial and is therefore more likely to bring the suit. Therefore, the anticipation of settlement raises the *overall volume of cases* that are pursued.

Topics

Accuracy

Several papers present formal analyses of the social value of accuracy in legal settings. Kaplow and Shavell (1996) argue that the *ex post* accurate verification of the victim's damages is socially valuable if the injurer knew the victim's damages at the time when he chose his precaution level. Accuracy is not valuable, however, if the victim's damages could not have been known by the injurer *ex ante*. The ‘scheduling’ of damages, or standardizing awards for injuries that fall into particular categories (as in workers’ compensation), may be desirable in these cases. Scheduling also makes the future outcome of the case more transparent – there is less to argue about – and can help to promote settlement (Spier, 1994b). Kaplow and Shavell (1992) argue that accuracy gives injurers an incentive to learn about the injuries that their activities might cause and will subsequently fine-tune their precautions. (Accurate information created by earlier trials may also help future actors fine-tune their actions; Hua and Spier, 2005.)

Alternative dispute resolution

Alternative dispute resolution (ADR) refers to the formal and informal proceedings that help parties resolve their disputes *outside* of formal litigation. Unlike settlement, which is typically achieved by the litigants themselves (and their lawyers), ADR proceedings often involve third parties who offer opinions and/or advice. Many of these systems are part of the court system, but many others are designed by the parties themselves (for example, ADR clauses in commercial contracts). In either case, ADR reflects the need to reduce the transaction costs of litigation and to make accurate decisions (Shavell, 1994; Mnookin, 1998). Farber and White's (1991) empirical study of medical malpractice claims suggests that

non-binding arbitration provides an informative signal and encourages subsequent settlement. Yoon (2004) confirms this result, but finds that ADR neither reduces litigation costs nor significantly shortens the delay. The importance of this topic and the relative dearth of research – both theoretical and empirical – makes ADR a ripe topic for further investigation.

Appeals

In most legal systems, a litigant who is dissatisfied with a lower court's decision can appeal to a higher court. In Shavell (1995), appeals can be an efficient means of correcting the errors made at the lower-court level. Appeals harness the private information of the litigants themselves: an incorrectly convicted defendant is more likely to appeal an earlier ruling since the probability of reversal is higher. In this way, resources are saved relative to random auditing. (See also Spitzer and Talley, 2000.) Daughety and Reinganum (2000a) consider a Bayesian model of appeals where the upper court perceives the private decision to appeal as informative and tries to rule ‘correctly’ given its posterior beliefs.

Bifurcation

Landes (1993) was the first to formally analyse ‘bifurcated’ trials where the court establishes the defendant's negligence before determining the plaintiff's damages. One benefit of bifurcation is that, once the defendant is absolved of liability, no further costs are incurred. The effect on the settlement rate is ambiguous, however. Chen, Chien and Chu (1997) consider these issues in a model with asymmetric information. Daughety and Reinganum (2000b) endogenize the level of litigation spending. White (2002), in her empirical analysis of asbestos trials, shows bifurcation raises the plaintiffs’ expected returns and increases the number of cases that are filed.

Case selection

The cases that go to trial are the tip of the iceberg – the vast majority of cases are settled before trial. These tried cases are likely to differ – perhaps systematically – from the cases that never reach the courtroom. Suppose the defendant is privately informed about the expected judgment at trial. Both the screening (Bebchuk, 1984) and signalling (Reinganum and Wilde, 1986) approaches discussed earlier predict that defendants with weak cases are more likely to settle out of court than defendants with strong cases. Intuitively, a defendant who expects an adverse judgment is more likely to accept a settlement offer. This result would be reversed if the plaintiff has private information instead. Many authors have explored case selection using models with non-common priors instead of asymmetric information. Most notably, Priest and Klein (1984) predicted that, for tried cases, the plaintiff win rate will tend towards 50 per cent. This stark result depends on the symmetry of the litigants, among other things. (With asymmetric information, Shavell, 1996, shows that any plaintiff win rate is possible.) More generally, however, the Priest–Klein framework suggests ways that trial rates may be systematically related to plaintiff win rates. Waldfogel (1995) estimates a structural model and finds results roughly consistent with the Priest–Klein theory.

Class actions

When an injurer has harmed a group of victims, these victims may (under some circumstances) join their claims for the purpose of litigation and/or settlement. One advantage of consolidation is the scale economies associated with common proceedings and legal representation. Che (1996) assumes that plaintiffs who join a class forgo a fine-tuned award and receive instead the average damage of the group. Absent settlement, it is clear that plaintiffs with weak cases are more likely to join a class. This adverse selection problem is mitigated when plaintiffs are privately informed. Weak plaintiffs have an incentive to remain independent, too, in an attempt to ‘signal’ that they have strong cases and, in equilibrium, fewer weak plaintiffs join the class. Che (2002) argues that classes may form to increase the members’ bargaining power via information aggregation. The defendant is more generous when bargaining with the class as a whole than when bargaining with individuals.

Contingent fees

In the United States, plaintiffs’ attorneys are often paid on a contingent basis, receiving a third (say) of any settlement or judgment but nothing if the case is lost. The use of contingent fees is regulated in the US. In particular, lawyers are prohibited from purchasing cases from their clients (Santore and Viard, 2001). Many European countries prohibit contingent fees altogether. There are many economic rationales for contingent fees. First, they give liquidity-constrained plaintiffs a way to finance their cases and shift some of the risk to the attorney. They also mitigate moral hazard (Danzon, 1983) and adverse selection problems. In Rubinfeld and Scotchmer (1993), attorneys have private information about their abilities and signal high quality through a willingness to accept contingent payment. Menus of contingent fees also arise when the clients have private information. (See also the mechanism-design model of Klement and Neeman, 2004.) In Dana and Spier (1993), the attorney has private information about the merits of the plaintiff's case. With contingent fees, the plaintiff can rest assured that the attorney will decline cases that are sure to lose. Finally, contingent fees can also be used strategically to make plaintiffs into ‘tougher’ negotiators (Hay, 1997; Bebchuk and Guzman, 1996). In empirical studies, Danzon and Lillard (1983) show a higher drop rate with contingent fees, and Helland and Tabarrok (2003) find that contingent fees are associated with higher-quality cases and faster case resolution.

Decoupling

It may be socially desirable to tax or subsidize the plaintiff's damage award. In Polinsky and Che (1991), a defendant chooses his level of precautions and, if injured, the plaintiff decides whether to bring suit. The optimal decoupled scheme taxes the plaintiff's award so that only a handful of cases are brought, but, at the same time, it makes the award very large so that the defendant's incentives are maintained. Since the defendant's stakes are large relative to the plaintiff's, the defendant will tend to spend more at trial (Kahan and Tuckman, 1995; Choi and Sanchirico, 2004). Daughety and Reinganum (2003)

consider these issues in a model with asymmetric information.

Disclosure and discovery

Litigants may voluntarily share information before trial. Indeed, the ‘unravelling’ logic of Grossman (1981) implies that all private information would come to light because an adverse inference would be drawn from silence. Full unravelling cannot occur, however, when hard evidence is simply unavailable. Guilty defendants have an incentive to pool with the innocent defendants who are unable to prove their innocence, for example. This suggests an important role for laws that require litigants to share information before trial. ‘Discovery’ can improve the accuracy of later court decisions (Hay, 1994; Cooter and Rubinfeld, 1994) and facilitate settlement negotiations before trial by narrowing the scope of asymmetric information (Shavell, 1989). (In contrast, Schrag, 1999, argues that discovery can lead to higher litigation costs and longer delays.) In Farber and White's (1991) sample of medical malpractice cases, many lawsuits are settled or dropped following discovery. Using a survey of attorneys in federal civil cases, Shepherd (1999) finds defendants increase their discovery efforts, ‘tit-for-tat’, in response to heightened discovery requests by the plaintiff.

The English Rule

In the United States, litigants bear their own costs of litigation – the ‘American Rule’. In contrast, the ‘English Rule’ shifts the winner's costs to the loser. Shavell (1982a) and Katz (1990) show that the English Rule discourages the filing of low-probability-of-prevailing cases but encourages high-probability-of-prevailing cases. (Kaplow, 1993, and Polinsky and Rubinfeld, 1998, discuss the normative implications.) The English Rule also tends to raise the litigation rate when parties disagree about the probability of winning (Bebchuk, 1984; Shavell, 1982a). Intuitively, the scope for disagreement is even higher because the parties have different beliefs about who will bear the litigation costs. Finally, the English Rule tends to raise the level of litigation spending (Braeutigam, Owen and Panzar, 1984; Hause, 1989; Katz, 1987). Intuitively, the marginal cost associated with spending is lower since the costs are partially externalized.

Inquisitorial versus adversarial systems

In adversarial systems, each side gathers and processes information separately. In inquisitorial systems – such as those found in continental Europe – these activities are more centralized and often presided over by a judge (see the discussion in Parisi, 2002). Adversarial systems are often criticized for giving litigants an incentive to hide relevant information from each other and from the court. They also can lead to the wasteful duplication of effort. On the other hand, adversarial systems may provide better incentives for information gathering (Dewatripont and Tirole, 1999). Milgrom and Roberts (1986) present a persuasion game where the parties have equal access to all of the relevant evidence and show that accuracy is not compromised in equilibrium. This stark result may no longer hold when parties have asymmetric access to evidence or when evidence is costly to gather and disclose; see also Shin (1998),

Daughety and Reinganum (2000b) and Froeb and Kobayashi (1996).

Insurance contracts

It is common for insurance contracts to place an upper bound on the level of coverage. This creates a potential conflict between the defendant and his insurer when deciding to settle a case (Meurer, 1992; Sykes, 1994). The insurance company is averse to settling because the defendant will bear the downside of a very large judgment at trial. Nevertheless, the defendant may delegate settlement authority to his insurer as a strategic commitment to be ‘tough’ in settlement negotiations. By reducing the most that the insurer is willing to pay in settlement, the insurance contract serves to extract value from the plaintiff. These contracts may be undesirable from a social welfare perspective, however, since the toughness of the insurer can increase the litigation rate (and the associated litigation costs). Formally, these ideas are related to Aghion and Bolton's (1987) analysis of contracts as a barrier to entry. (Spier and Sykes, 1998, show that corporate debt has a similar strategic value.)

Joint and several liability

There are many situations where a single victim is harmed by the actions of many injurers (for example, toxic-tort and price-fixing cases). Common rules for allocating responsibility include non-joint liability, where each losing defendant is responsible for his own share of damages, and joint and several liability, where a single losing defendant can be held responsible for the entirety of the plaintiff's damages. Kornhauser and Revesz (1994) analyse settlement incentives when the liability of a non-settling defendant is reduced, dollar for dollar, by the value of the previous settlements. (If the plaintiff's damages are $80 and one defendant settles for S, the remaining defendant may be responsible for $80 – S.) This rule encourages settlement when the cases are positively correlated but discourages settlement when the cases are independent. Some empirical support has been found in disputes between the Environmental Protection Agency (EPA) and Superfund defendants (Chang and Sigman, 2000).

Most-favoured-nation clauses

Settlement contracts in environments with multiple plaintiffs sometimes include ‘most-favoured-nation’ (MFN) clauses. They work in the following way: if an early settlement agreement includes an MFN clause and the defendant settles later with another plaintiff for more money, the early settlers receive the better terms, too. Spier (2003a) argues that MFN clauses economize on delay costs when a single defendant makes repeated offers to privately informed plaintiffs. MFNs may also be used to extract value from future plaintiffs (Spier, 2003b; Daughety and Reinganum, 2004). Intuitively, an MFN commits the defendant to be tough in future negotiations, allowing the defendant and the early plaintiffs to capture a greater share of the future bargaining surplus. The welfare effects of most-favoured-nation clauses are ambiguous. They can make early settlement negotiations more efficient but may lead later negotiations to fail.

Negative expected value claims

Suppose that a plaintiff has a negative expected value (NEV) claim – he stands to lose money if the case proceeds all the way to trial. Could this plaintiff succeed in extracting a settlement from the defendant? Interestingly, the *divisibility* of litigation costs over time can make the plaintiff's threat to litigate the NEV claim credible (Bebchuk, 1996). Here is the intuition. With divisibility, the bulk of the costs are sunk once the case reaches the courthouse steps. At that point, the plaintiff's threat to litigate is credible, so the defendant will settle. If we work backwards, the plaintiff's threat to continue may be credible at all stages of the game. Furthermore, a privately informed plaintiff with a NEV claim may mimic a plaintiff with a positive expected value claim and the defendant (not knowing for sure) may capitulate (Bebchuk, 1988; Katz, 1990). Finally, Rosenberg and Shavell (1985) present a model where the defendant must sink some defence costs or risk a summary judgment before trial.

Offer-of-judgment rules

Under Rule 68 of the United States Rules of Civil Procedure, if a plaintiff rejects a settlement offer and later receives a judgment that is less favourable, then the plaintiff is forced to bear the defendant's post-offer costs. Other rules allow for two-sided cost shifting. Spier (1994a) shows that these rules raise the settlement rate when liability is acknowledged but there is private information about damages. Intuitively, the rule serves to discipline aggressive settlement tactics (but see Farmer and Pecorino, 2000, and Miller, 1986). Bebchuk and Chang (1999) show that offer-of-judgment rules level the playing field in bargaining and lead to settlements that more accurately reflect the expected judgment at trial.

Patent litigation

Suppose that a patentee and an imitator are trying to settle a dispute. At trial, the patent may be invalidated, in which case the imitator will compete on equal footing with the patentee. Settlement provides an opportunity for collusion. Shapiro (2003) discusses these mechanisms and proposed criteria for judicial approval of patent settlements; see also Meurer (1989). Marshall, Meurer and Richard (1994) argue that the mere threat of patent litigation may be enough to soften competition in a patent race; see also Choi (1998). Lanjouw and Schankerman (2001) document interesting correlations between litigation decisions and the characteristics of the patents. In particular, a patent is more likely to be litigated if it serves as the ‘base of a cumulative chain’ or, in other words, there are more rents to be captured from future innovators.

Plea bargaining

In criminal cases in the United States, the prosecutor and the defendant often negotiate a guilty plea in exchange for a lighter sentence – a process known as plea bargaining. Landes (1971), in the first formal analysis of plea bargaining, assumes that the prosecutor maximizes the sum of expected sentences subject to a resource constraint. Grossman and Katz (1983) assume that the defendant privately observes

his guilt and the uninformed prosecutor makes a single take-it-or-leave-it offer of a reduced sentence in exchange for a guilty plea. In the screening equilibrium, the guilty defendants accept the offer and the innocent defendants reject the offer and go to trial. This is, of course, similar to Bebchuk's (1984) analysis of civil settlement. In Reinganum (1988), the prosecutor's offer signals the prosecutor's private information and, as in Reingaum and Wilde's (1986) analysis of civil settlement, the offers with high sentences are rejected more. In contrast to Grossman and Katz (1983), trials are more likely when the defendant is guilty. (In Reinganum, 2000, an informed defendant makes an offer to an uninformed prosecutor.)

Precedent

In Anglo-American legal systems, laws can be created and changed by judges over time. Cooter, Kornhauser and Lane (1979) present an early formal model where the courts learn about – and subsequently adjust – standards of care for injurers and victims. Landes and Posner (1976) consider the possibility of judicial bias, but argue that the threat of being overruled mitigates a judge's incentive to pursue his own agenda. Gennaioli and Shleifer (2005) present a formal model with a different conclusion. Rasmusen (1994) formalizes strategic interactions among a sequence of judges in a dynamic framework and shows that judges may cooperate in equilibrium and follow past precedents because violations would lead to future breakdowns where their own precedents would be violated by others; see also Schwartz (1992), Daughety and Reinganum (1999b) and Kornhauser (1992). Levy (2005) presents a model where judges have career concerns and go against precedent to signal their abilities. (A set of related rules and doctrines, ‘collateral estoppel’, applies when at least one litigant is involved in multiple suits; see Spurr, 1991, and Che and Yi, 1993.)

Secret settlement

It is not uncommon lawsuits to settle secretly, where neither the existence of the suit nor the terms of the settlement are observed by the public. Secrecy may be facilitated through ‘gag orders’ or through private contracts. In Daughety and Reinganum (1999a; 2002), open settlements publicize the defendant's involvement in a case and increase the likelihood that other plaintiffs will file suit in the future. They also provide future plaintiffs with information about the expected value of their claims. Daughety and Reinganum (1999a) show that, because of the publicity effect, early plaintiffs can extract ‘hush money’ from defendants, enriching themselves at the expense of later plaintiffs. Importantly, secrecy can compromise firms’ behaviour and product safety choices in a market setting (Daughety and Reinganum, 2005).

Standards of proof

How confident should a judge or jury be before convicting a defendant or finding in favour of a plaintiff? Rubinfeld and Sappington (1987) present a framework where the defendant can manipulate the signal received by the court, and shows how the optimal standard of proof balances litigation costs and

ex ante deterrence concerns. Sanchirico (1997) presents a model where plaintiffs, as well as defendants, make investments in their cases. Demougin and Fluet (2006) explores the trade-offs when the defendant's wealth is limited. See Bernardo, Talley and Welch (2000) and Hay and Spier (1997) for discussions of the burden of proof.

See Also

- dispute resolution
- law, economic analysis of

The author thanks the Searle Fund for financial support.

Bibliography

Aghion, P. and Bolton, P. 1987. Contracts as a barrier to entry. *American Economic Review* 77, 388–401.

Bebchuk, L. 1984. Litigation and settlement under imperfect information. *RAND Journal of Economics* 15, 404–15.

Bebchuk, L. 1988. Suing solely to extract a settlement offer. *Journal of Legal Studies* 17, 437–50.

Bebchuk, L. 1996. A new theory concerning the credibility and success of threats to sue. *Journal of Legal Studies* 25, 1–25.

Bebchuk, L. and Chang, H. 1999. The effect of offer-of-settlement rules on the terms of settlement. *Journal of Legal Studies* 28, 489–513.

Bebchuk, L. and Guzman, A. 1996. How would you like to pay for that? The strategic effects of fee arrangements on settlement terms. *Harvard Negotiation Law Review* 1, 3–63.

Bernardo, A., Talley, E. and Welch, I. 2000. A theory of legal presumptions. *Journal of Law, Economics, and Organization* 16, 1–49.

Braeutigam, R., Owen, B. and Panzar, J. 1984. An economic analysis of alternative fee shifting systems. *Law and Contemporary Problems* 47, 173–204.

Chang, H.F. and Sigman, H. 2000. Incentives to settle under joint and several liability: an empirical analysis of superfund litigation. *Journal of Legal Studies* 24, 205–36.

Che, Y.-K. 1996. Equilibrium formation of class action suits. *Journal of Public Economics* 62, 339–61.

Che, Y.-K. 2002. The economics of collective negotiations in pretrial bargaining. *International Economic Review* 43, 549–76.

Che, Y.-K. and Yi, J.G. 1993. The role of precedents in repeated litigation. *Journal of Law, Economics, and Organization* 9, 399–424.

Chen, K.-P., Chien, H.-K. and Chu, C.Y.C. 1997. Sequential versus unitary trials with asymmetric information. *Journal of Legal Studies* 26, 239–58.

Cho, I.-K. and Kreps, D. 1987. Signalling games and stable equilibria. *Quarterly Journal of Economics* 102, 179–221.

Choi, A.H. and Sanchirico, C.W. 2004. Should plaintiffs win what defendants lose? Litigation stakes, litigation effort, and the benefits of decoupling. *Journal of Legal Studies* 33, 323–54.

Choi, J.P. 1998. Patent litigation as an information-transmission mechanism. *American Economic Review* 88, 1249–63.

Cooter, R., Kornhauser, L. and Lane, D. 1979. Liability rules, limited information, and the role of precedent. *Bell Journal of Economics* 10, 366–73.

Cooter, R. and Rubinfeld, D. 1989. Economic analysis of legal disputes and their resolution. *Journal of Economic Literature* 27, 1067–97.

Cooter, R. and Rubinfeld, D. 1994. An economic model of legal discovery. *Journal of Legal Studies* 23, 435–64.

Dana, J. and Spier, K. 1993. Expertise and contingent fees: the role of asymmetric information in attorney compensation. *Journal of Law, Economics, and Organization* 9, 349–67.

Danzon, P. 1983. Contingent fees for personal injury litigation. *Bell Journal of Economics* 14, 213–23.

Danzon, P. and Lillard, L. 1983. Settlement out of court: the disposition of medical malpractice claims. *Journal of Legal Studies* 12, 345–78.

Daughety, A. 2000. Settlement. In *Encyclopedia of Law and Economics*, vol. 5, ed. B. Bouckaert and G. De Geest. Cheltenham: Edward Elgar.

Daughety, A. and Reinganum, J. 1994. Settlement negotiations with two-sided asymmetric information: model duality, information distribution, and efficiency. *International Review of Law and Economics* 14, 283–98.

Daughety, A. and Reinganum, J. 1999a. Hush money. *RAND Journal of Economics* 30, 661–78.

Daughety, A. and Reinganum, J. 1999b. Stampede to judgment: persuasive influence and herding behavior by courts. *American Law and Economics Review* 1, 158–89.

Daughety, A. and Reinganum, J. 2000a. Appealing judgments. *RAND Journal of Economics* 31, 502–25.

Daughety, A. and Reinganum, J. 2000b. On the economics of trials: adversarial process, evidence, and equilibrium bias. *Journal of Law, Economics, and Organization* 16, 365–94.

Daughety, A. and Reinganum, J. 2002. Information externalities in settlement bargaining: confidentiality and correlated culpability. *RAND Journal of Economics* 334, 587–604.

Daughety, A. and Reinganum, J. 2003. Found money? Split-award statutes and settlement of punitive damages cases. *American Law and Economics Review* 5, 134–64.

Daughety, A. and Reinganum, J. 2004. Exploiting future settlements: a signaling model of most-favored-nation clauses in settlement bargaining. *RAND Journal of Economics* 35, 467–85.

Daughety, A. and Reinganum, J. 2005. Secrecy and safety. *American Economic Review* 95, 1074–91.

Demougin, D. and Fluet, C. 2006. Preponderance of evidence. *European Economic Review* 50, 963–76.

Dewatripont, M. and Tirole, J. 1999. Advocates. *Journal of Political Economy* 107, 1–39.

Dixit, A. 1987. Strategic behavior in contests. *American Economic Review* 77, 891–98.

Farber, H. and White, M. 1991. Medical malpractice: an empirical examination of the litigation process. *RAND Journal of Economics* 22, 199–217.

Farmer, A. and Pecorino, P. 2000. Conditional cost shifting and the incidence of trial: pretrial bargaining in the face of a Rule 68 offer. *American Law and Economics Review* 2, 318–40.

Froeb, L. and Kobayashi, B. 1996. Naive, biased, yet Bayesian: can juries interpret selectively produced evidence? *Journal of Law, Economics, and Organization* 12, 257–76.

Gennaioli, N. and Shleifer, A. 2005. The evolution of precedent. Working Paper No. 11265. Cambridge, MA: NBER.

Gould, J. 1973. The economics of legal conflicts. *Journal of Legal Studies* 2, 279–300.

Grossman, S. 1981. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics* 24, 461–83.

Grossman, G. and Katz, M. 1983. Plea bargaining and social welfare. *American Economic Review* 73, 749–57.

Hause, J. 1989. Indemnity, settlement, and litigation, or ‘I'll be suing you’. *Journal of Legal Studies* 18, 157–80.

Hay, B. 1994. Civil discovery: its effects and optimal scope. *Journal of Legal Studies* 23, 481–517.

Hay, B. 1997. Optimal contingent fees in a world of settlement. *Journal of Legal Studies* 26, 259–78.

Hay, B. and Spier, K. 1997. Burdens of proof in civil litigation. *Journal of Legal Studies* 26, 413–33.

Hay, B. and Spier, K. 1998. Settlement of litigation. In *The New Palgrave Dictionary of Economics and the Law*, ed. P. Newman. London: Macmillan.

Helland, E. and Tabarrok, A. 2003. Contingency fees, settlement delay, and low-quality litigation: empirical evidence from two datasets. *Journal of Law, Economics, and Organization* 19, 517–42.

Hua, X. and Spier, K. 2005. Information and externalities in sequential litigation. *Journal of Institutional and Theoretical Economics* 161, 215–32.

Kahan, M. and Tuckman, B. 1995. Special levies for punitive damages. *International Review of Law and Economics* 15, 175–85.

Kaplow, L. 1993. Shifting plaintiffs’ fees versus increasing damage awards. *RAND Journal of Economics* 24, 625–30.

Kaplow, L. and Shavell, S. 1992. Private versus socially optimal provision of ex-ante legal advice. *Journal of Law, Economics, and Organization* 8, 306–20.

Kaplow, L. and Shavell, S. 1996. Accuracy in the assessment of damages. *Journal of Law and Economics* 39, 191–209.

Katz, A. 1987. Measuring the demand for litigation: is the English Rule really cheaper? *Journal of Law, Economics, and Organization* 3, 143–76.

Katz, A. 1990. The effect of frivolous lawsuits on the settlement of litigation. *International Review of Law and Economics* 10, 3–27.

Klement, A. and Neeman, Z. 2004. Incentive structures for class action lawyers. *Journal of Law, Economics, and Organization* 20, 102–24.

Kornhauser, L. 1992. Modeling collegial courts, 2: Legal doctrine. *Journal of Law, Economics, and Organization* 8, 441–70.

Kornhauser, L. and Revesz, R. 1994. Multidefendant settlements: the impact of joint and several liability. *Journal of Legal Studies* 23, 41–76.

Landes, W. 1971. An economic analysis of the courts. *Journal of Law and Economics* 14, 61–107.

Landes, W. 1993. Sequential versus unitary trials: an economic analysis. *Journal of Legal Studies* 22, 99–134.

Landes, W. and Posner, R. 1976. Legal precedents: a theoretical and empirical analysis. *Journal of Law and Economics* 19, 249–307.

Lanjouw, J. and Schankerman, M. 2001. Characteristics of patent litigation: a window on competition. *RAND Journal of Economics* 32, 129–51.

Levy, G. 2005. Careerist judges. *RAND Journal of Economics* 36, 275–97.

Loewenstein, G., Issacharoff, S., Camerer, C. and Babcock, L. 1993. Self-serving assessments of fairness and pretrial bargaining. *Journal of Legal Studies* 22, 135–58.

Marshall, R., Meurer, M. and Richard, J. 1994. Litigation settlement and collusion. *Quarterly Journal of Economics* 109, 211–39.

Meurer, M. 1989. The settlement of patent litigation. *RAND Journal of Economics* 20, 77–91.

Meurer, M. 1992. The gains from faith in an unfaithful agent: settlement conflict between defendants and liability insurer. *Journal of Law, Economics and Organization* 8, 502–22.

Milgrom, P. and Roberts, J. 1986. Relying on the information of interested parties. *RAND Journal of Economics* 17, 18–32.

Miller, G. 1986. An economic analysis of Rule 68. *Journal of Legal Studies* 15, 93–125.

Miller, G. 1987. Some agency problems in settlement. *Journal of Legal Studies* 161, 189–215.

Mnookin, R. 1998. Alternative dispute resolution. In *The New Palgrave Dictionary of Economics and the Law*, ed. P. Newman. London: Macmillan.

Myerson, R. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29, 265–81.

Nalebuff, B. 1987. Credible pretrial negotiation. *RAND Journal of Economics* 18, 198–210.

Ostrom, B., Kauder, N. and LaFountain, R. 2001. *Examining the Work of the State Courts, 1999–2000*. Williamsburg, VA: National Center for State Courts.

Parisi, F. 2002. Rent seeking through litigation: adversarial and inquisitorial systems compared. *International Review of Law and Economics* 22, 193–216.

P'ng, I.P.L. 1983. Strategic behavior in suit, settlement, and trial. *RAND Journal of Economics* 14, 539–50.

Polinsky, A. and Che, Y.-K. 1991. Decoupling liability: optimal incentives for care and litigation. *RAND Journal of Economics* 22, 562–70.

Polinsky, A. and Rubinfeld, D. 1988. The deterrent effects of settlements and trials. *International Review of Law and Economics* 8, 109–16.

Polinsky, A. and Rubinfeld, D. 1998. Does the English Rule discourage low-probability-of-prevailing plaintiffs? *Journal of Legal-Studies* 27, 519–35.

Posner, R. 1973. An economic approach to legal procedure and judicial administration. *Journal of Legal Studies* 2, 399–458.

Priest, G. and Klein, B. 1984. The selection of disputes for litigation. *Journal of Legal Studies* 13, 1–55.

Rasmusen, E. 1994. Judicial legitimacy as a repeated game. *Journal of Law, Economics, and Organization* 10, 63–83.

Reinganum, J. 1988. Plea bargaining and prosecutorial discretion. *American Economic Review* 78, 713–28.

Reinganum, J. 2000. Sentencing guidelines, judicial discretion, and plea bargaining. *RAND Journal of Economics* 31, 62–81.

Reinganum, J. and Wilde, L. 1986. Settlement, litigation, and the allocation of litigation costs. *RAND*

Journal of Economics 17, 557–68.

Rosenberg, D. and Shavell, S. 1985. A model in which lawsuits are brought for their nuisance value. *International Review of Law and Economics* 5, 3–13.

Rubinfeld, D. and Sappington, D. 1987. Efficient awards and standards of proof in judicial proceedings. *RAND Journal of Economics* 18, 308–15.

Rubinfeld, D. and Scotchmer, S. 1993. Contingent fees for attorneys: an economic analysis. *RAND Journal of Economics* 24, 343–56.

Sanchirico, C. 1997. The burden of proof in civil litigation: a simple model of mechanism design. *International Review of Law and Economics* 17, 431–47.

Santore, R. and Viard, A.D. 2001. Legal fee restrictions, moral hazard, and attorney rights. *Journal of Law and Economics* 44, 549–72.

Schrag, J. 1999. Managerial judges: an economic analysis of the judicial management of legal discovery. *RAND Journal of Economics* 30, 305–23.

Schwartz, E. 1992. Policy, precedent, and power: a positive theory of supreme-court decision making. *Journal of Law, Economics, and Organization* 8, 219–52.

Schweizer, U. 1989. Litigation and settlement under two sided incomplete information. *Review of Economic Studies* 56, 163–77.

Shapiro, C. 2003. Antitrust limits to patent settlements. *RAND Journal of Economics* 34, 391–411.

Shavell, S. 1982a. Suit, settlement and trial: a theoretical analysis under alternative methods for the allocation of legal costs. *Journal of Legal Studies* 11, 55–82.

Shavell, S. 1982b. The social versus the private incentive to bring suit in a costly legal system. *Journal of Legal Studies* 11, 333–9.

Shavell, S. 1989. The sharing of information prior to settlement or litigation. *RAND Journal of Economics* 20, 183–95.

Shavell, S. 1993. Suit versus settlement when parties seek nonmonetary judgments. *Journal of Legal Studies* 22, 1–14.

Shavell, S. 1994. Alternative dispute resolution: an economic analysis. *Journal of Legal Studies* 24, 1–

28.

Shavell, S. 1995. The appeals process as a means of error correction. *Journal of Legal Studies* 24, 379–426.

Shavell, S. 1996. Any probability of plaintiff victory at trial is possible. *Journal of Legal Studies* 25, 493–501.

Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26, 575–613.

Shepherd, G. 1999. An empirical study of the effects of pretrial discovery. *International Review of Law and Economics* 19, 245–63.

Shin, H. 1998. Adversarial and inquisitorial procedures in arbitration. *RAND Journal of Economics* 29, 378–405.

Spier, K. 1992. The dynamics of pretrial negotiation. *Review of Economic Studies* 59, 93–108.

Spier, K. 1994a. Pretrial bargaining and the design of fee-shifting rules. *RAND Journal of Economics* 25, 197–214.

Spier, K. 1994b. Settlement bargaining and the design of damage awards. *Journal of Law, Economics, and Organization* 10, 84–95.

Spier, K. 1997. A note on the divergence between the private and social motive to settle under a negligence rule. *Journal of Legal Studies* 26, 613–23.

Spier, K. 2003a. The use of most-favored-nation clauses in settlement of litigation. *RAND Journal of Economics* 34, 78–95.

Spier, K. 2003b. Tied to the mast: most-favored-nation clauses in settlement contracts. *Journal of Legal Studies* 32, 91–120.

Spier, K. 2005. Litigation. In *The Handbook of Law and Economics*, ed. A. Mitchell Polinsky and S. Shavell. Amsterdam: North-Holland.

Spier, K. and Sykes, A. 1998. Capital structure, priority rules, and the settlement of civil claims. *International Review of Law and Economics* 18, 187–200.

Spitzer, M. and Talley, E. 2000. Judicial auditing. *Journal of Legal Studies* 24, 649–83.

Spurr, S. 1991. An economic analysis of collateral estoppel. *International Review of Law and Economics* 11, 47–61.

Sykes, A. 1994. ‘Bad faith’ refusal to settle by liability insurers: some implications of the judgment-proof problem. *Journal of Legal Studies* 23, 77–110.

Waldfogel, J. 1995. The selection hypothesis and the relationship between trial and plaintiff victory. *Journal of Political Economy* 103, 229–60.

White, M. 2002. Explaining the flood of asbestos litigation: consolidation, bifurcation, and bouquet trials. Working Paper No. 9362. Cambridge, MA: NBER.

Yoon, A. 2004. Mandatory arbitration and civil litigation: an empirical study of medical malpractice litigation in the west. *American Law and Economics Review* 6, 95–134.

How to cite this article

Spier, Kathryn E. "litigation, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000217> doi:10.1057/9780230226203.0983

The New Palgrave Dictionary of Economics Online

Lloyd, William Forster (1794 – 1852)

Barry Gordon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

concentration of ownership; justice; labour power; Lloyd, W. F.; marginal utility theory of value; class; Poor Law; poverty; subsistence wage

Article

Lloyd was Drummond Professor of Political Economy at the University of Oxford from 1832 to 1837. During those years he delivered a series of lectures which display marked originality and willingness to differ from the current canons of received wisdom among political economists. Twelve of the lectures were published. The manuscripts of the remaining lectures, approximately 24 in number, have not been found (Romano, 1977).

Among the published lectures, that of 1833 and the second set of 1836 are quite outstanding. His lecture on Value (1833) has moved some leading historians of economic thought to hail Lloyd as one of the first writers to articulate the marginal utility theory of value. Less celebrated, but equally notable, is his analysis of the manner in which the operations of the contemporary British economy condemn unskilled labourers to poverty. Against the popular Malthusianism of his day, he argues in favour of the principle of poor laws and of the proposition that relief of the poor is a matter of social justice (rather than individual charity).

In the course of his 1836 lectures Lloyd constructs a model of the British economy which, he believes, demonstrates that the present situation of unskilled labourers is akin to that of slaves. Further, he observes, contemporary British society is dividing progressively into two mutually exclusive classes, and the degree of concentration of ownership and control of capital in the nation is increasing. Under existing circumstances, the unskilled worker is obliged to give ever greater quantities of his ‘power of labouring’ in order to obtain in return a subsistence wage.

As a person, Lloyd remains an elusive, even enigmatic, figure. He followed an older brother Charles (later, Regius Professor of Divinity and Bishop of Oxford) to Christ Church in 1812. There he studied mathematics and classics, took an MA in 1818 and was ordained in 1822. Before Lloyd succeeded Richard Whately in the Drummond Chair, he was Reader in Greek (1823) and lecturer in mathematics

(1824). In 1834 he was elected a Fellow of the Royal Society. At the end of his period as Professor of Political Economy, Lloyd left Oxford to live at Prestwood, Great Missenden, Buckinghamshire, where he died in 1852. During his last 15 years Lloyd appears to have lived very quietly and published nothing. There is as yet no satisfactory explanation as to why this able and well-connected scholar chose to remain silent.

Selected works

Twelve of Lloyd's lectures, 1834–36, were published collectively as *Lectures on Population, Value, Poor Laws and Rent*, London, 1837; reprinted, New York: A.M. Kelley, 1968. The collection includes: Two Lectures on the Checks to Population, delivered in 1834; A Lecture on the Notion of Value as Distinguishable not only from Utility, but also from Value in Exchange, delivered in 1833; Four Lectures on Poor Laws, delivered in Hilary term, 1836; Two Lectures on the Justice of Poor-Laws, and One Lecture on Rent, delivered in Michaelmas term, 1836. Earlier, Lloyd had published *Prices of Corn in Oxford in the Beginning of the Fourteenth Century: Also from the Year 1583 to the Present Time*, Oxford, 1830.

Bibliography

Bowley, M. 1972. The predecessors of Jevons – the revolution that wasn't. *Manchester School of Economic and Social Studies* 40(1), 9–29.

Gordon, B.J. 1966. W.F. Lloyd: a neglected contribution. *Oxford Economic Papers*, NS 18(1), 64–70.

Harrod, R.F. 1927. An early exposition of final utility: W.F. Lloyd's lecture on the notion of value (1833) reprinted. *Economic History* (supplement to the *Economic Journal*), May, 168–83.

Romano, R.M. 1971. W.F. Lloyd – a comment. *Oxford Economic Papers* 23, 285–90.

Romano, R.M. 1977. William Forster Lloyd – a non-Ricardian? *History of Political Economy* 9, 412–41.

How to cite this article

Gordon, Barry. "Lloyd, William Forster (1794–1852)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000121> doi:10.1057/9780230226203.0984

The New Palgrave Dictionary of Economics Online

loanable funds

S.C. Tsiang

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The term ‘loanable funds’ was used by the late D.H. Robertson, the chief advocate of the loanable funds theory of the interest rate, in the sense of what Marshall used to call ‘capital disposal’ or ‘command over capital’, (Robertson, 1940, p. 2). In a money-using economy where money is the only accepted means of payment, however, loanable funds are simply sums of money offered and demanded during a given period of time for immediate use at a certain price.

The loanable funds theory of interest is the theory which maintains that the interest rate, i.e. the price for the use of such funds per unit of time, must be determined by the supply and demand for such funds.

The insistence on the *flow* nature of loanable funds is based upon the crucial conception that in a money-using world the major bulk of money normally exists in a continuous circular flow. It is constantly passing out of the hands of one person as the means of payment for his expenditures into the hands of others as the embodiment of their incomes and sales proceeds, which will in turn be expended, and so on *ad infinitum*. A part of the money in this endless circular flow, however, is observed to be constantly being diverted into a side stream leading to the money market, where it constitutes the supply of loanable funds. From there borrowers of loanable funds would then take them off and in general would put them back into the main circular flow of expenditures and incomes (receipts).

This emphasis on the flow nature of loanable funds does not imply that the loanable funds theory would be unaware that there are sometimes money balances held inactive, like stagnant puddles lying off the main stream of the money flow. The loanable funds theory, however, would maintain that the stocks of money off the circular flow, as well as the stock of money inside the circular flow, have no direct influence on the money market. It is only when people attempt to divert money from the circular flow into the money market (saving), or into the stagnant puddles (hoarding), or conversely try to withdraw the inactive money from the stagnant puddles for re-injection into the circular flow or into the money market (dishoarding), that the interest rate will be directly affected. In other words, only *adjustments* in the idle balances (hoarding or dishoarding) together with the flows of savings and investment exert direct influences on the interest rate.

Since flows must be measured over time, we must choose a convenient unit to measure time. To take account of the fact that money does not circulate with infinite velocity, Robertson defined the unit period

as one ‘during which, at the outset of our inquiry, the stock of money changes hands once in final exchange for the constituents of the community's real income or output’ (Robertson, 1940, p. 65). In my opinion, however, it would be more consistent and convenient to define the unit period as one during which, at the outset, the stock of money changes hands once in exchange for all commodities and services instead of restricting the objects of exchange to final products only (Tsiang, 1956, esp. pp. 545–7). The reason for this will be clear later. Based on our new definition of the unit period, all gross incomes and sales proceeds from goods and services received during the current period cannot be spent on anything until the next period when they are then said to be ‘disposable’.

The definition of the unit period, however, does not preclude the funds borrowed or realized from sales of financial assets from being expendable during the same period. This differential treatment of the proceeds of sales of financial assets as distinguished from the proceeds of sales from goods and services is also an attempt to simulate the real situation in our present world; for the velocity of circulation of money against financial assets is in fact observed to be many times faster than that against goods and services. Assuming that there is a fixed unit period in our short period analysis does not necessarily imply that we are *ipso facto* assuming the invariability of the velocity of circulation of money; for short period variations in the velocity of money can be taken care of in terms of increases or decreases in the idle balances held.

Under this definition of the unit period and the implicit assumptions behind it, each individual, therefore, faces a financial constraint in that during a given unit period he can spend only his disposable income and his idle balances (the sum of the two constitutes the entire stock of money he possesses at the beginning of the period) plus the money he can currently borrow on the money market. Buying on credit is to be treated as first borrowing the money and then spending it. Thus when he plans to spend more than his disposable income and the amount he is willing to dishoard from his idle balances, he must borrow the excess from the money market to satisfy his total demand for finance. Since additions to the demand side are equivalent to deductions from the supply side, and vice versa, we need not dispute with Robertson when he classifies the demand for, and the supply of, loanable funds on the money market as follows (Robertson, 1940, p. 3).

On the demand side, he lists, with terminology slightly changed:

- D1 funds required to finance current expenditures on investment of fixed or working capital;
- D2 funds required to finance current expenditures on maintenance or replacement of existing fixed or working capital (note here that if our unit period were defined in the way Robertson defined it, i.e., as the period during which the total stock of money changes hands only once in the final exchange for the constituents of the community's real income, then the current expenditure on maintenance and replacement, i.e., on intermediate products, cannot be said to require a dollar for dollar provision of finance as would expenditures on final products);
- D3 funds to be added to inactive balances held as liquid reserves;
- D4 funds required to finance current expenditures on consumption in excess of disposable income. Correspondingly, on the supply side, he gives:
- S1 current savings defined as disposable income minus planned current consumption expenditure;
- S2 current depreciation or depletion allowances for fixed and working capital taken out of the gross sales proceeds of the preceding period;

- S3 dishoarding withdrawn from previously held inactive balances of money;
- S4 net creation of additional money by banks.

The function of the money market is to match the flow demands for loanable funds to the flow supplies, and the instrument with which it operates to achieve equilibrium between the two sides is the vector of interest rates. It is to be noted that in the flow equilibrium condition the total stock of money does not figure at all.

Nevertheless, it must be pointed out that the flow equilibrium condition of the money market as conceived by the loanable funds theorists can imply the stock equilibrium condition as conceived by the liquidity preference theorists, provided two necessary conditions are satisfied. Of the four demands for loanable funds listed above, D1, D2 and D4 are the additional demands for transactions balances (or what Keynes in 1937 called the finance demand for liquidity) needed by some firms and consumers to finance their current planned expenditures. And of the four sources of supply of loanable funds, S1 and S2 are but the reductions in demand for finance which other consumers of firms can spare during the current period. Therefore, D1, D2 and D4•minus S1 and S2 must be equal to the net aggregate increase which the community as a whole would want to add to their transaction balances.

Similarly, D3•minus S3 is the net increase which the community would want to add to their inactive balances (including precautionary, speculative, and investment balances).

Thus the equilibrium condition of the demand for and supply of loanable funds, i.e.,

$$D1 + D2 + D3 + D4 = S1 + S2 + S3 + S4,$$

which can be rearranged as:

$$[D1 + D2 + D4 - (S1 + S2)] + (D3 - S3) = S4,$$

implies that the total increases in aggregate demand for transaction balances (finance) and for inactive balances equal the net current increases in money supply created by banks. Provided it may be presumed (*a*) that the previous stock supply of and demand for money were originally equal to each other, and (*b*) that the current increases (or decreases) in supply and demand for money (treated above as flow supply and demand for loanable funds) represent the full unlagged adjustments of the previous stock supply and demand to their new equilibrium values, the flow equilibrium of the loanable funds should necessarily imply a new stock equilibrium (Tsiang, 1982).

The two necessary provisos used to be taken for granted by the liquidity preference theorists, who generally think that full stock equilibrium can be achieved instantaneously at any point in time. However, Professor James Tobin, in his Nobel lecture given in 1981 (Tobin, 1982), has come to recognize that the money market cannot operate within a dimensionless point of time, but must operate

in finite time periods, which he called slices of time. Furthermore, he recognized that the equilibrium which can be expected in such a short slice of time can only be that between the adjustments in the stock demanded and in the stock supplied during the period. Since adjustments in stocks per time period are flows, Tobin's new approach is thus really a sort of flow equilibrium analysis.

Moreover, Tobin, at the same time, also admitted that in such a short period as a slice of time, portfolios of individual agents cannot adjust fully to new market information. Lags in response are inevitable and rational in view of the costs of transactions and decisions. Thus neither of the two necessary conditions is satisfied in the real world. Consequently, even when the money market has brought the flow demand for and supply of loanable funds to equality, the stock demand for money and the total money stock need not have reached mutual equilibrium, which the Keynesians and the stock-approach economists used to assume as being attainable at every point of time.

Finally, it should be realized that the demand for finance for planned investment expenditure, which Keynes (1937, p. 667) admitted he should not have overlooked in his *General Theory*, is of the nature of a flow generated by a flow decision to invest. It is not just a partial adjustment of the stock demand for money towards its new equilibrium value as treated in Tobin's new theory (Tobin, 1982). As Keynes put it in his reply to Ohlin (1937)>, ‘“Finance” is a revolving fund•…•. As soon as it is used in the sense of being expended, the lack of liquidity is automatically made good and the readiness to become temporarily unliquid is available to be used over again’ (Keynes, 1937, p. 666). This is essentially a reaffirmation of the traditional conception of the circular flow of money, which loanable funds theorists had emphasized from the outset, but which Keynes himself had pushed into the dark background with his emphasis that the entire stock of money is being held voluntarily in portfolio allocation.

The rediscovery of the demand for finance by Keynes and the more recent unheralded switch on the part of Tobin towards the flow approach from his usual stock approach indicate that the loanable funds theory is perhaps the more appropriate approach at least for short period dynamic analysis.

See Also

- liquidity preference

Bibliography

Keynes, J.M. 1937. The ex-ante theory of the rate of interest. *Economic Journal* 47(December), 663–9.

Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment, I. *Economic Journal* 47 (March), 53–69.

Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment, II. *Economic Journal* 47 (June), 221–40.

Robertson, D.H. 1940. *Essays in Monetary Theory*. London: P.S. King.

Tobin, J. 1982. Money and finance in the macroeconomic process. *Journal of Money, Credit and Banking* 14(May), 171–204.

Tsiang, S.C. 1956. Liquidity preference and loanable funds theories, multiplier and velocity analysis: a synthesis. *American Economic Review* 46(September), 539–64.

Tsiang, S.C. 1982. Stock or portfolio approach to monetary theory and the neo-Keynesian school of James Tobin. *IHS-Journal* 6, 149–71.

How to cite this article

Tsiang, S. C. "loanable funds." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000126> doi:10.1057/9780230226203.0985

The New Palgrave Dictionary of Economics Online

local public finance

John M. Quigley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The mobility of consumers and producers in response to fiscal incentives gives the study of local public finance its distinctive character. Households and firms are partitioned into spatial units on the basis of preferences, costs and the incentives provided by local tax and expenditure policies. These fiscal incentives are, in turn, chosen by the members of each of these jurisdictions or clubs. Externalities within and between these localities greatly affect the efficiency of taxation and the provision of public goods and services.

Keywords

clubs; congestion; efficient allocation; excise taxes; exclusionary zoning; intergovernmental grants; inter-jurisdictional competition; Lindahl tax structure; local public finance; local public goods; lump-sum taxes; marginal rate of substitution; marginal rate of transformation; median voter theorem; poll tax; property taxation; public finance; public goods; residential mobility; revenue sharing; stabilization; tax price; technical change; Tiebout hypothesis

Article

Economic analysis of the taxation and expenditure policies of local public authorities has become far more sophisticated as theoretical enquiry has directed attention towards the uniquely local aspects of public finance and as national policies have increased the importance of the local public sector.
Many of the issues that arise in the analysis of the local public sector are familiar reflections of the important questions in public finance that have been addressed at the national level; for example, the incidence of taxation and the welfare losses from revenue instruments; the effect of government expenditures on consumer welfare and the distribution of well-being; the effect of public sector distortions on resource allocation and relative prices.
However, the principal difference between the economic analysis of public finance at the national and at

the local levels is the potential for mobility among jurisdictions by the transport of final products and inputs, and especially by residents who finance local government and consume public output. Critically, this mobility may be endogenous to the revenue or expenditure actions taken by the local public authority, and this must be considered in any economic analysis of local finance.

This insight, as it affects efficiency in the allocation of local public output and the incidence of local taxes, goes back at least to the fifth edition of Marshall's *Principles* (1907, Appendix G). Marshall presented a lucid discussion of the effect of local public expenditures on residential mobility (‘A high rate spent on providing good primary and secondary schools may attract artisan residents while repelling the well-to-do’ – Marshall, 1920, p. 794). He also noted the effects of mobility upon the incidence of local taxes.

Given the increased complexity of decentralized taxation and expenditure patterns when compared to national government policies, one may begin by asking which economic functions of government ought to be undertaken by the central (national) government rather than by local authorities. Consider the original Musgrave (1959) taxonomy of public sector functions: distribution, stabilization and allocation. It seems clear that a system of local taxes and expenditures is inappropriate for achieving distributional or stabilization goals. After the adoption of any system of taxation and redistribution by a locality, even one which reflects a unanimous view of the citizens, it will be in the interests of those bearing the burden of the tax to relocate in other jurisdictions and in the interests of potential beneficiaries of the redistribution to move into the jurisdiction. Similarly, locally adopted monetary and fiscal policies are unlikely to further stabilization objectives, even if such objectives are uniformly held by local citizens. Import leakages are so large that the local benefits of stabilization policies (for example, local public employment programmes) are almost certain to be less than their costs.

It is precisely the mobility of households, goods and factors across jurisdictions that defeats local stabilization and redistribution policies. Conversely, however, the same ‘openness’ of the local economy means that the decentralized local provision of public goods will in many cases improve the allocative efficiency of the economy. In particular, the smaller and more homogeneous a community in a system of local government, the more likely is it that the provision of public goods by any community will be consistent with the demands of its citizens. In the limit, of course, if public goods are financed by a head tax, and if there are neither economies of scale in production nor externalities in consumption, then provision by a system of small jurisdictions, each with citizens of homogeneous tastes and incomes, will result in an efficient allocation.

If, however, there are economies of scale in production, it makes sense to have larger jurisdictions. But when the public good is produced by a larger entity, ‘congestion’ may result; that is, the quality of the good may decline as it is shared with more people. In larger jurisdictions, moreover, citizen demands may be more heterogeneous. The problem of balancing the benefits of cost-sharing in production, on the one hand, with the sacrifice in well-being by compromising individual consumers’ demands or by introducing ‘congestion’ in public goods consumption, on the other, has been central to the normative analysis of the local provision of public goods.

Consider, for example, a ‘club’ providing some collective benefit to identical individuals (Buchanan, 1965). Suppose an organization supplies some public output Q subject to congestion, or equivalently, suppose it supplies a good whose standardized cost $C(N)$ increases with population N. Individuals of income Y are assessed the average cost of service provision and allocate their remaining income to some numeraire good X. A community of N identical individuals will choose public output to maximize

utility, $U(Q,X)$, subject to the individual budget constraint, $Y = X + [C(N) / N] Q$. This implies the familiar Samuelson (1954) condition:

$$N[(\partial U / \partial Q) / (\partial U / \partial X)] = C(N). \quad (1)$$

The level of public good provision is chosen by the club of fixed size N so that the *sum* of the individual marginal rates of substitution (MRS) between private and public goods equals the marginal rate of transformation (MRT) in production. Given this level of public output, from the budget constraint it also follows that choice of club size to maximize utility is:

$$C'(N) = C(N) / N. \quad (2)$$

The optimum size of the club is the membership at which the average cost of public output is equal to the marginal cost of adding another member. From equations (1) and (2) it follows that for a pure public good, that is, $C'(N) = 0$, the optimal size of the club is unbounded, while for a private good, where $C(N) = PN$, the *individual* MRS is equal to the MRT and the size of the club is indeterminate.

Applied to local public finance, the model indicates that a system of communities, each with identical individuals and of that size which minimizes average cost, would be a stable and efficient mechanism for public service provision. Homogeneity of demands is necessary for efficiency even if the tax structure (or club dues) is of the Lindahl variety. Each group in a heterogeneous community would be better off by moving to a jurisdiction with identical tax shares.

Theoretical analyses of local public economies are much more complicated when the partitioning of individuals into political jurisdictions is ‘non-anonymous’, that is, when the characteristics of the other members (in addition to their incomes) matter to those in the club. In many cases, an equilibrium allocation of residents to jurisdictions may not exist at all (Scotchmer, 1997). As noted below, non-anonymous crowding may also affect the costs of public goods provision and the interpretation of demands for local public goods.

The ‘club’ model of the provision of local public goods is a special case of the so-called Tiebout (1956) model, probably the most influential idea in the modern analysis of local public finance. Tiebout's stylized and informal analysis assumes that residential mobility is costless, that local jurisdictions provide public goods at minimum average cost and that local government is financed by non-distortionary lump-sum taxes. Under these circumstances, Tiebout argues that the provision of public goods by a system of competitive local governments may be no less efficient than the allocation of private goods by the market economy. The conclusion of this argument also depends crucially upon the

availability to citizens of a sufficiently large number of jurisdictions offering differing packages of local public goods and upon the absence of inter-jurisdictional externalities, as well as more conventional assumptions about full information. In reality, in most metropolitan areas, local public output is supplied by a small number of communities (small, at least, relative to the number of types of demanders); local mobility is quite costly and is motivated by many non-fiscal concerns. Individuals often live in one jurisdiction and work in another, and there are externalities among jurisdictions. Finally, revenues are raised, not by head taxes but by a variety of local levies, especially *ad valorem* taxes on real property. Each of these factors limits the economic efficiency of the local public sector in important ways.

The externalities or ‘spillouts’ of the benefits of public service provision mean that such goods will be underprovided without coordination by local communities – since each community will only consider the benefits accruing to its own citizens in choosing the level of service provision. For public goods and services with substantial spillouts of benefits, efficient levels of production can be stimulated by a system of open-ended matching grants to localities by the central government. As Pigou (1932) originally demonstrated, if the matching rate (the fraction of local spending reimbursed by higher government) corresponds to the fraction of local public output, which spills out to non-residents, then the externality will be internalized. It is, of course, rather difficult to implement this maxim of local public finance (Oates, 1972).

The heavy reliance upon local property taxes for financing the local public sector, especially in Britain, Canada and the United States, is another source of allocative inefficiency in local finance. Clearly, a property tax alters the housing consumption decision and leads to underconsumption of housing as well as to inefficiency in public goods consumption. Until rather recently the system of local property taxes was viewed as a system of excises (Netzer, 1966), regressive levies on property and housing consumption, in contrast to the original Henry George (1879) position on land taxes. Modern theoretical analyses (following Mieszkowski, 1972), which assume that capital is mobile across jurisdictions and that the supply of capital is insensitive to its rate of return, have led to a reconsideration of the regressive nature of the tax. The inelastic supply of aggregate capital means that a national system of local property taxes will reduce returns to capitalists by the average level of the tax. The geographical mobility of capital implies that capital will flee from high-tax jurisdictions, raising marginal productivity and pre-tax returns, to low-tax jurisdictions, depressing pre-tax returns. Thus the incidence of the system of property taxes depends upon the magnitude of the average level of the tax, relative to the deviations from that average, as well as distribution of households among high-tax and low-tax jurisdictions. Despite the ambiguities in resolving these detailed empirical issues, this theoretical argument suggests that the burden of property taxation is heavily skewed towards the owners of capital. Empirically, this conclusion is probably modified by regressive appraisal and administrative procedures. It should be noted, moreover, that from local governments’ perspective an increase in the level of the property tax to finance service provision is an excise on property users (since a change in any one community's property tax rate can have only a negligible effect on the average level of rates for the nation).

The distortion inherent in property tax financing may lead to local policies of exclusionary zoning. If, for example, the benefits of the local public sector were roughly equal per household, then it would be in the interests of current residents to force incoming households to consume more housing than the average household. Current residents may attempt to enforce this by imposing minimum lot-size restrictions or by other exclusionary practices to increase the housing consumption of newcomers. Of course, as noted before, unless there are sufficient communities so that the households residing within a

jurisdiction are literally identical, those who chose to consume less housing will typically enjoy a fiscal residual.

Despite these clear examples of allocative inefficiency in the system of local public finance and service provision, there is a substantial body of evidence that variations in property tax rates are reflected in property values and that variations in public services (for example, school quality) are capitalized into the sale prices of residential property. These findings are certainly consistent with the process of ‘voting with one's feet’ implied by the Tiebout model, but the capitalization of taxes and services is not necessary to efficiency in local government, nor does efficient service provision necessarily imply capitalization.

The observation that individuals register their demands for publicly financed services in their choices of community has other important implications, however. Specifically, information about the public goods provided by different jurisdictions, together with information about the characteristics of the residents of those jurisdictions, may be sufficient to identify consumer demands for public services. Extensive analyses of these issues have been undertaken, combining economic theories of the local political process with aggregate data on local public finance and choice of output. Under rather restrictive assumptions, the political process which determines the level of service provision can be modelled as the choice of the median voter of the community. Given the characteristics of that individual (or rather, estimates obtained from aggregate information), the ‘tax price’ that individual confronts, and the level of public output chosen, the parameters of the demand curve are estimated econometrically. The ‘tax price’ is the marginal cost to the individual of purchasing an additional dollar of public output. With property tax financing, this is typically approximated by the median voter's house value as a fraction of the community's taxable real property per household.

As noted above, the residents of localities may ‘care’ about the characteristics of other residents simply because their characteristics affect the cost of producing public services. One example may involve local schools, which absorb the largest share of local government spending on public services. To the extent that peers ‘matter’ in the production of educational outputs in primary school, policies of matching grants to local governments based on disadvantaged residents are called for (see Nechyba, 2003). The specification of empirical models of the demand for local public services is much more problematic when the demographic characteristics reflect either tastes for public goods or the costs of supplying them, or both.

Nevertheless, the results of these empirical investigations have proven useful in the positive analysis of citizen demands for public services and in the analysis of local finance. Nevertheless, the underlying economic model of local government behaviour is open to questions, both technical (for example, the requirement that preferences exhibit single peakedness) and substantive (for example, the neglect of the role of bureaucracy in government decisions). For example, if the median voter determines the demand for local public output, then the propensity for a community to spend out of lump-sum aid from higher government ought to be no different from the propensity to spend out of income generated by local taxation. Yet empirical evidence suggests that the propensity of communities to spend out of untied grant income greatly exceeds the propensity to spend out of ordinary income. A variety of alternative models of local finance have been espoused to help explain this ‘flypaper’ effect (‘money sticks where it lands’) in the context of bureaucratic decision-making. Chief among them are the so-called Leviathan models of a government that exploits its citizens by maximizing revenues extracted by taxation (Brennan and Buchanan, 1980). Clearly, however, more theoretical work needs to be done to resolve the

contradictions between mobile consumers of local public output and sluggish suppliers.
Finally, it has been suggested that the inherent nature of local output and the traditional financing mechanisms of local government combine to exacerbate the economic and administrative problems of the local public sector (Baumol, 1967). Local output consists largely of labour-intensive services, where technical change is inherently slow, and is typically financed by income-inelastic tax instruments. Under reasonable demand conditions, these may produce a more or less continuous ‘crisis’ in local public finance, as service costs escalate more rapidly than revenue increments. Given these characteristics of the local financing mechanism, as well as the redistributive nature of many local services, there may thus be a strong case for revenue or tax-base sharing at the national level.

See Also

- fiscal federalism
- public finance
- public goods
- Tiebout hypothesis
- urban economics

Bibliography

Baumol, W.J. 1967. Macroeconomics of unbalanced growth: the anatomy of urban crisis. *American Economic Review* 62, 415–26.

Brennan, G. and Buchanan, J.M. 1980. *The Power to Tax: Analytical Foundations of a Fiscal Constitution*. Cambridge: Cambridge University Press.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* 32, 1–14.

George, H. 1879. *Progress and Poverty*. New York: Appleton.

Inman, R.P. 1979. The fiscal performance of local governments: an interpretive review. In *Current Issues in Urban Economics*, ed. P. Mieszkowszki and M. Straszheim. Baltimore, MD: Johns Hopkins University Press.

Marshall, A. 1907. *Principles of Economics*, 5th edn. 8th edn, London: Macmillan, 1920.

Mieszkowski, P. 1972. The property tax: an excise tax or a property tax? *Journal of Public Economics* 1, 73–96.

Musgrave, R.A. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

Nechyba, T. 2003. School finance, spatial income segregation, and the nature of communities. *Journal of Urban Economics* 54, 61–88.

Netzer, D. 1966. *Economics of the Property Tax*. Washington, DC: Brookings Institution.

Oates, W.E. 1972. *Fiscal Federalism*. New York: Harcourt Brace Jovanovich.

Pigou, A.C. 1932. *The Economics of Welfare*, 4th edn. London: Macmillan.

Rubinfeld, D.L. 1985. The economics of the local public sector. In *Handbook of Public Economics*, vol. 2, ed. J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36, 387–9.

Scotchmer, S. 1997. On price-taking equilibria in club economies with nonanonymous crowding. *Journal of Public Economics* 65, 75–87.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

How to cite this article

Quigley, John M. "local public finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000130> doi:10.1057/9780230226203.0986

The New Palgrave Dictionary of Economics Online

local regression models

Oliver B. Linton

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses local regression models, that is, regression models where the parameters are allowed to vary with some covariates either in a completely unrestricted fashion or in an intermediate way with some exclusion restrictions that make some parameters vary only with some covariates. Special cases are nonparametric regression and additively separable nonparametric regression.

Keywords

additive models; Cobb–Douglas parametric model; conditional expectation; conditional variance; GARCH models; generalized additive models; identification; linear models; local regression models; parametric models

Article

Local regression models are regression models where the parameters are ‘localized’, that is, they are allowed to vary with some or all of the covariates in a general way. Suppose that (Y, X) are random variables and let

$$E(Y|X = x) = m(x) \tag{1}$$

when it exists. The regression function $m(x)$ is of primary interest because it describes how X affects Y. One may also be interested in derivatives of m or averages thereof or in derived quantities like conditional variance $var(Y|X = x) = E(Y^2|X = x) - E^2(Y|X = x)$. In cases of heavy-tailed distributions,

the conditional expectation may not exist, in which case one may instead work with other location functionals like trimmed mean or median. The conditional expectation is particularly easy to deal with but a lot of what is done for the mean can also be done for the median or other quantities.

A parametric regression model for $m(x)$ is a family of functions $M(x; \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$, where for each θ, $M(,; \theta)$ is a known function. The true parameter θ_0 for which $M(x; \theta_0) = m(x)$ for all $x \in \mathcal{X}$ is unknown and has to be estimated from data. For example, $M(x; \theta) = x^T \theta$ would correspond to the linear regression case, which is the central model of econometrics. A key concept is that of identifiability: M is identifiable when distinct parameter values lead to different values of M for at least some x values. See Rothenberg (1971) for discussion. Parametric models arise frequently in economics and are of central importance. However, such models arise only when one has imposed specific functional forms on utility or production functions. Without these ad hoc assumptions one only gets much milder restrictions on functional form like concavity, symmetry, homogeneity and so on. The nonparametric approach is based on the belief that parametric models are usually mis-specified and may result in incorrect inferences. In this approach one treats the regression function $m(x)$ as being of unknown functional form. One usually assumes that m is a continuous function or even differentiable, although there are cases of interest where $m(x)$ is, say, continuous only from the right (left) with limits on the left (right), that is, there may be jumps at certain known or unknown locations in the support $\mathcal{X}$ of X (see Delgado and Hidalgo, 2000). By not restricting the functional form one obtains valid inferences for a much larger range of circumstances. In practice, the applicability depends on the sample size and the quality of data available. The theory and methods for carrying out such estimation are well understood, and are reviewed elsewhere (Härdle and Linton, 1994). Local regression models are one way of interpreting the nonparametric approach.

A local regression model is a family of functions

$$M(x; \theta(x)), \quad \theta \in \Theta = \{\theta : \mathcal{X} \to \mathbb{R}^p\}, \tag{2}$$

where $M(x; \theta)$ is a known function of both arguments. The true (functional) parameter $\theta_0(\cdot)$ for which $M(x; \theta_0(x)) = m(x)$ for all $x \in \mathcal{X}$ is unknown. It is usually assumed to be smooth. In other words this is a standard parametric regression model except that the parameters vary with the covariate value. There are a number of special cases. At one extreme lies the parametric model in which $\theta(x) = \theta$ for all $x \in \mathcal{X} \subset \mathbb{R}^d$, but the true θ_0 is unknown. At the other extreme lies the fully nonparametric case where $\theta(\cdot)$ is not subject to any exclusion restrictions.

Many different M functions will generally do. For example, the local constant case corresponds to $M(x; \theta) = \theta$ and the local linear case corresponds to $M(x; \theta) = \theta_0 + \theta_1 x$. These cases along with higher-order polynomials have been widely studied (see, for example, Fan and Gijbels, 1996). There are also other possibilities. Consider the Cobb–Douglas parametric model

$$M(x;\ \theta) = \theta_0 x_1^{\theta_1} \cdots x_d^{\theta_d}, \tag{3}$$

which is widely used in studies of production. By making $\theta = (\theta_0, \theta_1, \ldots, \theta_d)$ vary freely with x one can match with any function $m(x)$ so long as the supports coincide (see, for example, Charnes, Cooper and Schinnar, 1976). For binary data where it is known that $m(x) \in [0, 1]$ it is appropriate to take $M(x;\ \theta) = F(\theta_0 + \theta_1 x)$ for some given c.d.f. F like the normal or logit. In that case, for a given x, there exists $\theta_0(x)$, $\theta_1(x)$ such that $m(x) = F(\theta_0(x) + \theta_1(x)x)$. This example illustrates some pitfalls; for example, when $m(x) > 1$ for some x of interest. In that case, taking $M(x;\ \theta) = F(\theta_0 + \theta_1 x)$ will not be satisfactory.

The statistical justification for using local constant, local linear, and more generally local polynomial models is that any smooth function $m(x)$ can be approximated near the point x_0 by Taylor series expansions, so for p-times continuously differentiable scalar functions we have

$$m(x) = \sum_{j=0}^{p} \frac{1}{j!} \frac{d^j m}{dx^j}(x_0)(x - x_0)^j + R(x, x_0), \tag{4}$$

where the remainder term satisfies $R(x, x_0) / |x - x_0|^p \to 0$ as $x \to x_0$. Thus the function m is locally well approximated by a polynomial of order p, $\sum_{j=0}^{p} \alpha_j (x - x_0)^j$, where α_j can be identified with $j!^{-1} d^j m(x_0) / dx^j$. This justifies using local polynomial regression. But why should one ever work with local regression models outside the local polynomial class? First, any other local parametric model $M(x, \theta)$ that is p-times continuously differentiable in x at x_0, satisfies a similar expansion to (4), $\sum_{j=0}^{p} \beta_j (x - x_0)^j$, where β_j are functions of θ. By equating coefficients one obtains the same leading terms as long as there are ‘enough’ parameters in θ. Therefore, the same approximating objectives are reached by any such model. In some cases other equivalent classes may provide better approximations. Polynomials can sometimes violate some known features, like for example $m(x) \in [0, 1]$. In that case, taking $M(x; \theta)$ to be a c.d.f. of a polynomial provides the same approximation (so long as the c.d.f. chosen is also smooth enough) but imposes the boundedness restriction. Second, the local parameters may also be of interest in themselves. In the Cobb–Douglas case, the $\theta_j(x)$ can be interpreted as local elasticities. A third benefit is that the local model nests the parametric model. This leads to better statistical properties for estimators and test statistics when the model is true or approximately true, the ‘home turf’ case (see Hjort and Glad, 1995). When the default parametric model in the area of interest is

nonlinear, as is true in many fields, there are some advantages to taking a localization of this in the nonparametric approach.

The issue of identification in local regression models is not well explored but some results are known (see Gozalo and Linton, 2000). The expansion (4) is clearly crucial for identification. If the function *m* is continuous but not differentiable, then only a single parameter is identifiable, which corresponds to the first term in (4); additional parameters remain unidentified. It is also necessary that there is a neighbourhood of the estimation point that contains enough observations (this is guaranteed when the marginal density exists and is positive).

Estimation of local regression models can be carried out by localization of the usual estimation criteria adopted for estimation of the corresponding parametric model like maximum likelihood or the method of moments where the localization is carried out by multiplying the contribution of observation *i* to the sample average objective function by the weight $w_{ni} = K((x - X_i) / h)$, where *K* is called the kernel and usually satisfies at least $\int K(u)du = 1$, while $h = h(n)$ is the bandwidth, a sequence designed to go to zero with sample size. The effect of the weighting factor w_{ni} is to emphasize observations close to the point of interest *x* and to de-emphasize observations far from *x*, whence the appellation ‘localization’.

In the multivariate case, the expansion (4) becomes much more complicated: there are *d* first order partial derivatives, $d(d - 1) / 2$ second order partial derivatives, and so on. With $p = 5$ and $d = 10$ the local parametric model would have over 1,000 parameters, which is too many for practical use. There are many interesting and important cases lying between the two extremes of parametric and fully nonparametric models, where some of the θ_j vary with only a subset of *x*. In this case, the local parametric model is imposing exclusion restrictions on the function *m* and the expansion is reduced. We next give some examples.

A function $m(x)$ is additively separable if

$$m(x) = \sum_{j=1}^{d} m_j(x_j)$$

for some functions m_j. In terms of the framework of the previous section $p = d$ and

$$M(x; \theta) = \sum_{j=1}^{d} M_j(x_j, \theta_j); \ \theta_j(x) = \theta_j(x_j).$$

The functions $\theta_j(x_j)$ are one-dimensional but of unknown form. This implies that $m(x) = \Sigma_{j=1}^{d} m_j(x_j)$, where $m_j(x_j) = \theta_j(x_j)$. In this case, each function $\theta_j(x)$ has $d - 1$ exclusion

restrictions. This is consistent with strong separability as defined in Goldman and Uzawa (1964). A generalization of this is to the so-called generalized additive models where $M(x, \theta) = G\left(\sum_{j=1}^{d} M_j(x_j, \theta_j)\right)$, where $\theta_j(x) = \theta_j(x_j)$, in which G is a known 'link' function, while θ_j are univariate functions as before. For example, G could be the c.d.f. of a random variable like the normal or logit. Linton and Nielsen (1995) discuss estimation of additive models.

In time series one is often interested in the relationship

$$E[y_t | I_{t-1}] = m(I_{t-1}),$$

where the information set $I_{t-1} = \{y_{t-1}, \ldots\}$ includes all past variables, either for estimation or forecasting purposes. This situation is complicated because I_{t-1} contains infinitely many variables and apart from the important class of Markov models m generally depends on all of them. A common assumption here is some kind of mixing condition that guarantees that the effect of y_{t-k} on y_t dies out as $k \to \infty$. For example, an invertible $MA(1)$ process has $m(I_{t-1}) = \sum_{j=1}^{\infty} \theta^{j-1} y_{t-j}$ for some $|\theta| < 1$. A natural generalization of this is the model $m(I_{t-1}) = \sum_{j=1}^{\infty} m_j(y_{t-j})$, where m_j is a sequence of functions such that the sum is well defined, that is, $m_j(\cdot)$ must decline in importance as $j \to \infty$. This model is hard to analyse and to estimate. Instead, consider the more restrictive version

$$m(I_{t-1}) = \sum_{j=1}^{\infty} \theta^{j-1} m(y_{t-j}) \tag{5}$$

for some unknown function $m(\cdot)$ and parameter θ. When $m(y) = y$ this includes the MA(1) process as a special case, but includes many other nonlinear models. By taking a local parametric model $M(y) = a_0(y) + a_1(y)y + a_2(y)y^2$ for m one can nest the GARCH(1, 1) model of Bollerslev (1986). Linton and Mammen (2005) have recently developed a theory of estimation for this class of models. Another popular approach is the locally stationary models pioneered by Dahlhaus (1997). A locally stationary AR(1) process is $y_t = \rho(t/T)y_{t-1} + \varepsilon_t$, where ε_t is i.i.d. and $\rho(\cdot)$ is a smooth but unknown form. By taking the local parametric model $M(y) = a_0$ one can nest the conventional autoregression, although there are other possibilities. Dahlhaus actually deals with a more general class of linear processes with $y_t = \sum_{j=0}^{\infty} c_j(t/T)\varepsilon_{t-j}$, where $c_j(\cdot)$ are unknown but smooth functions.

See Also

- kernel estimators in econometrics
- nonparametric structural models
- semiparametric estimation

Bibliography

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–27.

Charnes, A., Cooper, W. and Schinnar, A. 1976. A theorem on homogeneous functions and extended Cobb–Douglas forms. *Proceedings of the National Academy of Science, USA* 73, 3747–4748.

Dahlhaus, R. 1997. Fitting time series models to nonstationary processes. *Annals of Statistics* 25, 1–37.

Delgado, M. and Hidalgo, F. 2000. Nonparametric inference on structural breaks. *Journal of Econometrics* 96, 113–44.

Fan, J. and Gijbels, I. 1996. *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

Goldman, S. and Uzawa, H. 1964. A note on separability and demand analysis. *Econometrica* 32, 387–98.

Gozalo, P. and Linton, O. 2000. Local nonlinear least squares estimation: using parametric information nonparametrically. *Journal of Econometrics* 99, 63–106.

Härdle, W. and Linton, O. 1994. Applied nonparametric methods. In *The Handbook of Econometrics*, vol. 4, ed. D. McFadden and R. Engle. Amsterdam: North- Holland.

Hjort, N. and Glad, I. 1995. Nonparametric density estimation with a parametric start. *Annals of Statistics* 23, 882–904.

Linton, O. and Mammen, E. 2005. Estimating semiparametric ARCH models by kernel smoothing methods. *Econometrica* 73, 771–836.

Linton, O. and Nielsen, J. 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.

Rothenberg, T. 1971. Identification in parametric models. *Econometrica* 39, 577–91.

How to cite this article

Linton, Oliver B. "local regression models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000223> doi:10.1057/9780230226203.0987

The New Palgrave Dictionary of Economics Online

location theory

Jacques-François Thisse

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Location theory deals with *what* is *where*. ‘What’ refers to any possible type of economic activity involving stores, dwellings, plants, offices, or public facilities. ‘Where’ refers to areas such as regions, cities, political jurisdictions, or custom unions. The objective of location theory is to explain why particular economic activities choose to establish themselves in particular places. Here we focus on spatial competition theory between firms, where locations are subject to attracting and repelling forces. We then extend this framework in order to account for the residential choices made by consumers.

Keywords

agglomeration forces; clusters; collusion; dispersion forces; general equilibrium; globalization; Hotelling, H.; land use; local labour markets; location theory; Nash equilibrium; new economic geography; oligopolistic competition; partial equilibrium; Principle of Differentiation; product differentiation; spatial competition; Stigler, G.; transportation costs; urban economics

Article

From a historical perspective, location theory has been at both the centre and the periphery of economic theory. It has been at the centre to the extent that it has followed the tradition taking its roots in Hotelling's classical paper ‘Stability in Competition’ (1929) and has used the spatial framework as a metaphor to explain issues involving heterogeneity and diversity across agents (Rosen, 2002). Examples include the supply of differentiated products, electoral competition between political parties, the matching process on the labour market, competition between communities to attract residents or firms, and the number and size of jurisdictions. Location theory has been at the periphery to the extent that space has not been a major concern for most economists. Indeed, it is rare to find a principles textbook in which location issues are covered, let alone mentioned. This is despite their obvious importance for the way actual markets function, as shown, for instance, by the debate raging in many industrialized

countries about the consequences of globalization for the location of jobs.

The theory of optimal location for a firm has long been dominated by the minisum model in which the firm aims at minimizing its total transportation costs (Weber, 1909). Formally, this is achieved by minimizing the weighted sum of distances to a finite number of points, which represent input and output markets. When the length of the shortest path connecting any two points of a transportation network measures the distance between these points, the firm's optimal location is an input/output market, or a node of the network, or both (Hurter and Martinich, 1988). Hence, *the locational choice of a firm is either sluggish or catastrophic*. Another interesting feature of that model is that the firm's optimal location is the outcome of the interplay of a system of forces pulling the firm in different directions. When several competing firms are to be located, the system of forces becomes richer in that it involves what are called ‘agglomeration’ and ‘dispersion’ forces.

Spatial competition between firms

To see how such a system of forces works, we consider the framework developed by Hotelling (1929). The market of a homogeneous good is made up of consumers who request one unit of the good. Because any single consumer is negligible to firms, Hotelling assumes that consumers are continuously distributed along a linear and bounded segment: think of Main Street. For simplicity, consumers are also supposed to be uniformly distributed along the linear segment. Two stores, aiming to maximize their respective profits, seek a location along the same segment. Because they are dispersed across locations, consumers differ in their access to the same store. In such a context, firms anticipate correctly that each consumer will buy from the store posting the lower full price, namely, the price at the firm's gate, called ‘mill price’, augmented by the travel costs that consumers must bear to go to the store they patronize. Accordingly, once they are located firms have some monopoly power over the consumers located in their vicinity, which enables them to choose their price. Of course, this choice is restricted by the possibility that consumers have to supply themselves from the competing firm. Note that any firm is supposed to have a single location – that is, an *address* – because increasing returns and indivisibilities do not allow it to run a large number of outlets dispersed along Main Street without incurring major losses (Koopmans, 1957).

Since each firm is aware that its price choice affects the consumer segment supplied by its rival, *spatial competition is inherently strategic*. This is one of the main innovations introduced by Hotelling, who uses a two-stage game to model the process of spatial competition. In the first stage, stores choose their location non-cooperatively; in the second, these locations being publicly observed, firms select their selling price. The use of a sequential procedure means that firms anticipate the consequences of their locational choices on their subsequent choices of prices, thus imparting to the model an implicit dynamic structure. The game is solved by backward induction. For an arbitrary pair of locations, Hotelling starts by solving the price subgame corresponding to the second stage. The resulting equilibrium prices are introduced into the profit functions, which then depend only upon the locations chosen by the firms. These functions stand for the payoffs that firms will maximize during the first stage of the game. Such an approach anticipates by several decades the concept of subgame perfect Nash equilibrium introduced by Selten (1965).

Whereas the individual purchase decision is discontinuous – a consumer buying only from one firm –

Hotelling finds it reasonable to suppose that firms' aggregated demands are continuous with respect to prices. Supposing that each consumer is negligible solves the apparent contradiction between discontinuity at the individual level and continuity at the aggregated level. In other words, when consumers are continuously distributed across locations aggregated demands are 'often' continuous. The hypothesis of the continuum that had been popularized much later by Aumann (1964) is found here to represent the idea that competitive agents have a negligible impact on the market outcome. However, Hotelling considers a richer setting involving both 'dwarfs' – consumers – whose behaviour is competitive and 'giants' – firms – whose behaviour is strategic because they can manipulate the market outcome.

Hotelling's claim was that the process of spatial competition leads firms to agglomerate at the market centre. If true, this provides us with a rationale for the observed spatial concentration of firms selling similar goods (such as restaurants, movie theatres, or fashion clothes shops). But Hotelling's analysis is undermined by a mistake that invalidates his main conclusion: when firms are sufficiently close, the corresponding subgame does not have a Nash equilibrium in pure strategies, so that the payoffs used by Hotelling in the first stage are wrong (d'Aspremont, Gabszewicz and Thisse, 1979). This negative conclusion has led d'Aspremont et al. to slightly modify the Hotelling setting by assuming that the travel costs borne by consumers are quadratic in the distance covered, instead of being linear as in Hotelling. This new assumption captures the idea that the marginal cost of time increases with the length of the trip to the store. In this modified version, d'Aspremont et al. show that any price subgame has one and only one Nash equilibrium in pure strategies. Plugging these prices into the profit functions, they show that firms choose to set up at the two extremities of the linear segment. Firms do so because this allows them to relax price competition and to restore their profit margins. Indeed, when prices are fixed and equal the quest for customer proximity – or, equivalently, for a larger market area – leads the two firms to agglomerate at the market centre. The tendency for firms to choose distinct locations or products has been confirmed by many works, and has led Tirole (1988) to call it the 'Principle of Differentiation'. Consequently, *price competition is a dispersion force*, whereas *the market area effect is an agglomeration force*. What the Principle of Differentiation tells us is that the dispersion force always dominates the agglomeration force, at least when firms sell a homogeneous product and compete in price. Hence, the Hotelling setting is to be enriched if we want to be able to understand why firms selling similar products often form spatial clusters. This has been accomplished by following two different research strategies. In the first, the purpose is to identify market mechanisms allowing firms to relax price competition without being spatially separated. From this perspective, the most natural approach is to assume that firms sell products that are differentiated in the space of characteristics. It combines both spatial and product differentiation per se. An alternative approach, however, is to appeal to some form of collusion between firms that permits them to avoid the devastating effects of price competition. This is especially relevant when products can hardly be differentiated. The second research strategy is based on Stigler (1961) and develops the idea that consumers are imperfectly informed about the places where the existing varieties are made available. In such a context, consumers must undertake some search before finding a good match.

Product differentiation and collusion

Several papers have shown that firms selling differentiated varieties choose to agglomerate at the market centre when products are sufficiently differentiated, transportation costs borne by the consumers are low enough, or both (de Palma et al., 1985). This can be understood as follows. When consumers have different tastes and when residential locations and tastes are not correlated (or, alternatively, when individuals exhibit a love of variety), each firm supplies what is the best match for consumers who are otherwise dispersed across all locations. Price competition is relaxed by product differentiation, so that firms may afford to set up at the place offering the best accessibility to their potential customers. Such a place is obviously the market centre when the consumer distribution along Main Street is uniform. In addition, it is never profitable for a firm to leave the cluster when transportation costs are low because the benefit of a good match dominates the additional transportation costs that the consumer must bear to buy her best match. All of this seems to fit modern economies characterized by more and more variety and decreasing travel costs. In a nutshell, we may then safely conclude that one of the main reasons for agglomeration is that *firms substitute product differentiation for spatial separation*, very much as *Newsweek* and *Time* are supplied in the same stores but differentiated by their cover stories (Irmen and Thisse, 1998).

The welfare analysis of such an outcome is somewhat unexpected. At the optimum, prices are set equal to the common marginal cost so that consumers' well-being depends only upon firms' locations. In the case of a homogeneous good, maximizing total welfare boils down to minimizing aggregate transportation costs. However, once we introduce differentiation across varieties, consumers no longer patronize the nearest firm on each trip because they now benefit from intrinsic differentiation between stores. In this context, one needs a more general approach accounting for both distance and product diversity effects, the appropriate measure being the consumers' indirect utility. As a result, the formation of a cluster need not be socially sub-optimal. Quite the opposite: when products are sufficiently differentiated, transportation costs are low, or both, it is socially desirable to have all firms agglomerated within a cluster. Hence, unlike what Hotelling thought, such an extreme concentration may be socially optimal.

Under what became known as 'semi-collusion', it has been shown that firms that anticipate some form of collusion in the price stage, which is typically repeated, will choose to locate together at the market centre (Jehiel, 1992; Friedman and Thisse, 1993). In this case, selling a homogenous product makes it easier to sustain price collusion because the punishment for a defecting firm is more severe. Of course, collusion is not easy to maintain in the long run, so that firms face a positive probability that price collusion will break down. In this case, firms select separated locations but do not seek to maximize their spatial differentiation. Specifically, Jehiel, Friedman and Thisse (1995) have established that the higher the probability that the price agreement will break down, the larger is the distance between firms.

Search

When firms sell differentiated products, it is reasonable to assume that consumers are incompletely informed about the varieties that are supplied. Even though the typical consumer knows which varieties are available in the market, she is unsure about which variety is offered where (and at which price). If consumers have to compare alternatives before buying, they must undertake *search* among firms. Stated differently, when the only way for consumers to find out which variety is on offer in a particular store is

to visit this store, they must bear the corresponding travel cost. Gathering information being costly, each consumer must compare the cost of an additional bit of information with the expected gain in terms of surplus. In a spatial setting, both the cost and the gain vary with consumers' and firms' locations.
When several stores are located together, it is reasonable to assume that the typical consumer knows the location and size of the cluster but not its composition. Once she arrives at the cluster, the travel costs are sunk and she can visit any store at a very low cost. But she must pay the transportation cost to each isolated store she visits. Spatial clustering of stores is, therefore, a particular means by which firms can facilitate consumer search. Indeed, a consumer is more likely to visit a cluster of stores than an isolated one because of the higher probability she faces of finding there a good match and a good price. When firms realize this fact, each of them understands that it might be in its own interest to form a marketplace with others. When a firm considers the possibility of joining competitors within the same marketplace, it thus faces a trade-off between a negative competition effect and a positive market area effect, both being generated by the pooling of firms selling similar products.
In the case of a market with a fixed size, Wolinsky (1983) has shown that the market outcome involves all firms forming a single marketplace once transportation costs are sufficiently low and when there are enough stores to make the cluster attractive. It is worth noting that the agglomeration may arise away from the market centre. Any point such that no single firm is able to find an alternative location far enough to induce some consumers to visit it before the cluster is a spatial equilibrium. Of course, the cluster cannot be too far from the market centre because stores need to offer a good accessibility to *all* consumers. Accordingly, once the urban area extends far away into the same direction, this implies that some firms will want to create a new cluster away from the original one.
Schulz and Stahl (1996) show that it is possible to uncover additional and surprising results by considering a market of variable size. To this end, they consider an unbounded space that allows them to capture the idea that more competition within the cluster may attract new customers coming from more distant locations, thus allowing the demand for each variety to increase. More concretely, the entry of a new variety may lead to an increase in the cluster's demand that outweighs the decrease in market area inflicted on existing varieties. Although price competition becomes fiercer, it appears here that firms may take advantage of the extensive margin effect to increase their prices in equilibrium. Clearly, when the number of varieties is not too large, such positive effects associated with the gathering of firms strengthen the agglomeration force that lies behind the cluster. Though collectively several firms might want to form a new market, it may not pay an individual firm to open a new market in the absence of a coordinating device. Consequently, a new firm entering the market will choose instead to join the incumbents, thus leading to a larger agglomeration. In this case, *the entry of a new firm creates a positive externality for the existing firms by making total demand larger*. This in turn explains the common fact that department stores encourage the location of competing firms within the shopping centre.

The relationship with new economic geography

It appears that location theory and new economic geography have a lot in common, a fact that has been overlooked in the literature. Such a relationship between the two domains is worth noting because economic geography models are developed in general equilibrium frameworks involving monopolistic competition on the product market, whereas location theory uses partial equilibrium models under

oligopolistic competition. Indeed, one of the main conditions identified in spatial competition for a cluster of firms to emerge corresponds with the main finding established in ‘new economic geography’, that is, firms agglomerate when trade costs are sufficiently low (Krugman, 1991; Fujita, Krugman and Venables, 1999). Likewise, product differentiation fosters agglomeration whereas, by its mere existence, a cluster generates a lock-in effect similar to those encountered in economic geography. In both settings, the absence of increasing returns would lead to the emergence of ‘backyard capitalism’ in which each household produces its own consumption bundle. Finally, the market size effect uncovered in search models is similar to the agglomeration effect identified by Krugman and others.

Spatial competition and urban economics

So far, consumers have been able to seek where to buy but not where to live. Yet it is reasonable to assume that consumers adjust their residential choices to the locations selected by large firms and/or by public facilities. For the resulting distribution to be non-degenerate, a land market must be introduced in which consumers compete for land use. In such a context, the demand of a consumer for the firm's output becomes in turn endogenous in that it depends on the income left after the land rent is paid. This brings into the picture some general equilibrium ingredients in that firms and households locations are interdependent. Fujita and Thisse (1986) consider a setting in which firms choose their locations, anticipating consumers’ residential choices, this sequence reflecting the fact that firms have market power whereas consumers adjust their locational choices to those made by firms. Because they compete for land, consumers are spread around firms in a way such that no consumer can find a better place to live. In the case of two firms selling a homogeneous good at a common given price, the agglomeration of the two firms is always a Nash equilibrium. However, dispersed equilibria may also coexist when travel costs are sufficiently high. This is because the decrease in individual consumption resulting from a move toward the rival dominates the market area effect. In other words, firms may not find it advantageous to agglomerate, thus showing that *competition for land acts a major dispersion force*.

Public facilities

Cities provide a large variety of local public goods. Because its location interacts with the locational choices of firms and households, a large public facility which consumers wish to access influences the nature of the urban structure. In particular, one expects the presence of a major equipment to act as an agglomeration force on the private sector (Thisse and Wildasin, 1992). When topographical boundaries have no impact on the location of the public facility, this one is always established at the centre of the urban area and there is a tendency for this facility to draw the private firms together as income rises with respect to transportation costs. When the facility is set up near the edge of the area available for urban use – think of an urban area on the coast of a body of water – the resulting asymmetry has a significant impact on the locational interactions between firms: the two private firms are located together at the centre of the urban area. Hence, *the public facility may serve as the center of a dispersed spatial configuration, or it may induce the agglomeration of firms in a location different from that of the public facility itself*. In both cases, it vastly contributes to the shaping of the city structure.

Local labour markets

Due to the evolution of technological progress and the concomitant expansion of metropolitan areas, the urban labour force has become more heterogeneous whereas the labour market has been segmented in thinner sub-markets. The force inducing the formation of local labour markets finds its origin, at least partially, in the skill and geographical heterogeneity of workers (Brueckner, Thisse and Zenou, 2002). When workers have heterogeneous skills, firms have different job requirements because they have incentives to differentiate their job offers in order to gain market power in the labour market. This in turn implies that the labour market works as an *oligopsony* in which firms with different skill needs and different urban locations compete for mobile and skill-heterogeneous workers. In terms of urban economics, each firm may be considered as a company town attracting workers who also choose to reside near this firm. As in the case of firms selling consumption goods, firms are separated in the geographical space because this allows them to enjoy market power over the workers situated in their vicinity. Consequently, the economy may be viewed as a system of cities in which each firm/city competes to attract workers who are also residents. *The fact that each firm is anchored in a distinct location is a fundamental reason for the emergence of local labour markets*.
When workers bear the training cost that allows them to erase any mismatch between their innate skills and the skill needs of their employer, the net wage is lower for workers whose ‘skill distance’ from their employer is larger. Firms understand that, in the residential equilibrium, commuting distance is positively related to a worker's skill distance from the firm. In such a context, the equilibrium residential location of workers is governed by the quality of their match in the labour market. Knowledge of the connection between skill and commute distances affects the firm's interaction with its rivals as it competes for labour. The critical issue is that the equilibrium wage depends on the commuting cost parameters, yielding a link between the urban structure and the labour market. More precisely, low-skill workers incur high commuting costs, which may in turn lead low-skill workers not to take a job. Unemployment may arise, therefore, because some workers turn out to be too ‘distant’ from firms in *both* the skill and urban spaces. As in the foregoing, two different spatial components interact to shape the social structure of cities.

See Also

- new economic geography
- product differentiation
- spatial economics
- systems of cities
- urban agglomeration
- urban economics

Bibliography

Aumann, R. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Brueckner, J., Thisse J.-F. and Zenou, Y. 2002. Local labor markets, job matching and urban location. *International Economic Review* 43, 155–71.

d'Aspremont, C., Gabszewicz, J. and Thisse, J.-F. 1979. On Hotelling's ‘Stability in Competition’. *Econometrica* 47, 1045–50.

de Palma, A., Ginsburgh, V., Papageorgiou, Y.Y. and Thisse, J.-F. 1985. The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica* 53, 767–81.

Friedman, J. and Thisse, J.-F. 1993. Partial collusion fosters minimum product differentiation. *RAND Journal of Economics* 24, 631–45.

Fujita, M., Krugman, P. and Venables, A. 1999. *The Spatial Economy: Cities, Regions and International Trade*. Cambridge, MA: MIT Press.

Fujita, M. and Thisse, J.-F. 1986. Spatial competition with a land market: Hotelling and von Thünen unified. *Review of Economic Studies* 53, 819–41.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39, 41–57.

Hurter, A. and Martinich, J. 1988. *Facility Location and the Theory of Production*. Dordrecht: Kluwer.

Irmen, A. and Thisse, J.-F. 1998. Competition in multi-characteristics spaces: Hotelling was almost right. *Journal of Economic Theory* 78, 76–102.

Jehiel, P. 1992. Product differentiation and price collusion. *International Journal of Industrial Organization* 10, 633–41.

Jehiel, P., Friedman, J. and Thisse, J.-F. 1995. Collusion and antitrust detection. *Japanese Economic Review* 46, 226–46.

Koopmans, T. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99, 483–99.

Rosen, S. 2002. Markets and diversity. *American Economic Review* 92, 1–15.

Schulz, N. and Stahl, K. 1996. Do consumers search for the highest price? Equilibrium and monopolistic optimum in differentiated products markets. *RAND Journal of Economics* 27, 542–62.

Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft* 121, 301–24.

Stigler, G. 1961. The economics of information. *Journal of Political Economy* 69, 213–25.

Thisse, J.-F. and Wildasin, D. 1992. Public facility location and urban spatial structure. *Journal of Public Economics* 48, 83–118.

Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

Weber, A. 1909. *Ueber den Standort der Industrien*, Tübingen, J.C.B. Mohr. English translation: *The Theory of the Location of Industries*. Chicago: Chicago University Press, 1929.

Wolinsky, A. 1983. Retail trade concentration due to consumers' imperfect information. *Bell Journal of Economics* 14, 275–82.

How to cite this article

Thisse, Jacques-François. "location theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000213> doi:10.1057/9780230226203.0988

The New Palgrave Dictionary of Economics Online

logit models of individual choice

Thierry Magnac

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The logit model was named by Berkson after probit, its close competitor; the two are the most popular econometric methods used in applied work to estimate models for binary variables. It can be easily extended to the treatment of multinomial variables and enjoys specific properties in panel data binary models. Increasingly flexible logit models have also been elaborated for demand analyses. Their development has been stimulated by the increasing availability of databases on individual discrete choices. Because generalized logit models belong to the class of random utility models, their use has promoted sound applied economic research in demand analysis.

Keywords

asymptotic least squares; conditional likelihood; discrete choices; general extreme value distributions; generalized linear models; incidental parameters; independence of irrelevant alternatives; logit models of individual choice; maximum likelihood; method of moments; minimum distance; mixture models; probit models; random utility models; simulation methods; spatial statistics; statistical mechanics; two-level nested logit

Article

The logit function is the reciprocal function to the sigmoid *logistic* function. It maps the interval [0,1] into the real line and is written as:

$$logit(p) = \ln(p / (1 - p)).$$

Two traditions are involved in the modern theory of logit models of individual choices. The first one concerns *curve fitting* as exposed by Berkson (1944), who coined the term ‘logit’ after its close competitor ‘probit’ which is derived from the normal distribution. Both models are by far the most popular econometric methods used in applied work to estimate models for binary variables, even though the development of semiparametric and nonparametric alternatives since the mid-1970s has been intensive (Horowitz and Savin, 2001).

In the second strand of literature, models of discrete variables and discrete choices as originally set up by Thurstone (1927) in psychometrics have been known as ‘random utility models’ (RUM) since Marschak (1960) introduced them to economists. As the availability of individual databases and the need for tools to forecast aggregate demands derived from discrete choices were increasing from the 1960s onwards, different waves of innovations, fostered by McFadden (see his Nobel lecture, 2001) elaborated more and more sophisticated and flexible logit models. The use of these models and of simulation methods has triggered burgeoning applied research in demand analysis in recent years.

Those who wish to study the subject in greater detail are referred to Gouriéroux (2000), McFadden (2001) or Train (2003), where references to applications in economics and marketing can also be found.

Measurement models

As Berkson (1951, p. 327) put it, logit (or probit) models may be seen as ‘merely a convenient way of graphically representing and fitting a function’. They are used for any empirical phenomenon delivering a binary random variable Y_i, taking values 0 and 1, to be analysed. In a logit model, it is postulated that its probability distribution conditional on a vector of covariates X_i is given by:

$$\Pr(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

where β is a vector of parameters. This model can also be derived from more general frameworks in statistical mechanics or spatial statistics (Strauss, 1992).

With the use of cross-sectional samples, the parameter of interest is estimated using maximum likelihood or by generalized linear models (GLM) methods where the link function is logit (McCullagh and Nelder, 1989). Under the maintained assumption that it is the true model and other standard assumptions, the maximum likelihood estimator (MLE) is consistent, asymptotically normal and efficient (Amemiya, 1985). Nevertheless, the MLE may fail to exist, or more exactly be at the bounds of the parameter space, when the samples are uniformly composed of 0•s or 1•s, for instance (Berkson, 1955).

When repeated observations are available, the method of Berkson delivers an estimator close to MLE since they are asymptotically equivalent. Observe first that the logit function of the true probability obeys the linear equation:

$$logit(\Pr(Y_c = 1|X_c)) = X_c\beta$$

where the covariates X_c now take a discrete number of values defining each cell, c. Second, use the observed frequency in each cell, $\hat{p}_c$, and contrast it with the theoretical probability, p_c, as:

$$logit(\hat{p}_c) = X_c\beta + (logit(\hat{p}_c) - logit(p_c)) = X_c\beta + \varepsilon_c.$$

The random term ε_c properly scaled by the square root of the number of observations in cell c is asymptotically normally distributed with variance equal to $1/(p_c(1-p_c))$. The method of Berkson then consists in using minimum chi-square, that is, a method of moments, to estimate β, an instance of what is know as minimum distance or asymptotic least squares (Gouriéroux, Monfort and Trognon, 1985).

When measurements for a single individual are repeated, Rasch (1960) suspected that individual effects might be important and proposed to write:

$$logit(\Pr(Y_{it} = 1|X_{it})) = X_{it}\beta + \delta_i$$

where t indexes the different items that are measured and δ_i is an individual specific intercept or fixed effect. Items can be different questions in performance tests or different periods. In the original Rasch formulation, parameters were allowed to be different across items, β_t, and there were no covariates. Given that the number of items is small, it is well known that the estimation of such a model runs into the problem of incidental parameters (see Lancaster, 2000). As the number of parameters δ_i increases with the cross-section dimension, the MLE is inconsistent (Chamberlain, 1984). Nevertheless, the nuisance parameters δ_i can be differenced out using conditional likelihood methods (Andersen, 1973) because:

$$logit(\Pr(Y_{it} = 1|X_{it}, Y_{it} + Y_{it'} = 1)) = (X_{it} - X_{it'})\beta.$$

The conditional likelihood estimator of β is consistent and root n asymptotically normal but it is not

efficient, although no efficient estimator is known. Furthermore, when binary variables Y_{it} are independent, conditionally on X_i, the only model where a root n consistent estimator exists is a logit model (Chamberlain, 1992). Extensions of Rasch rely on the fact that root n consistent estimators exist if and only if $Y_{it} + Y_{it'}$ is a sufficient statistic for the nuisance parameters δ_i (Magnac, 2004). When the number of items or periods becomes large, profile likelihood methods where individual effects are treated as parameters seem to be accurate in Monte Carlo experiments as soon as the number of periods is four or five (Arellano, 2003).

Multinomial logit (or in disuse ‘conditional logit’) is to binary logit what a multinomial is to a binomial distribution (Theil, 1969). Given a vector Y_i consisting of K elements which are binary random variables and lie in the $\mathbb{R}^K$– simplex (their sum is equal to 1), it is postulated that:

$$\Pr(Y_i^{(k)} = 1|X_i) = \frac{\exp(X_i\beta^{(k)})}{1 + \Sigma_{k=2}^{K}\exp(X_i\beta^{(k)})}$$

where by normalization, $\beta^{(1)} = 0$. Ordered logit has a different flavour since it applies to rank-ordered data such as education levels (Gouriéroux, 2000).

As probits, logit models are very tightly specified parametric models and can be substantially generalized. Much effort has been exerted to relax parametric and conditional independence assumptions, starting with Manski (1975). Manski (1988) analyses the identifying restrictions in binary models, and Horowitz (1998) reviews estimation methods. In some cases, Lewbel (2000) and Matzkin (1992) offer alternatives.

Random utility models

The theory of discrete choice is directly set up in a multiple alternative framework. A choice of an alternative k belonging to a set C is assumed to be probabilistic either because preferences are stochastic or heterogenous, or because choices are perturbed in a random way. By definition, choice probability functions map each alternative and choice sets into the simplex of $\mathbb{R}^K$.

A strong restriction on choices is the axiom of Independence of Irrelevant Alternatives (IIA, Luce, 1959). The axiom states that the choice between two alternatives is independent of any other alternative in the choice set. The version that allows for zero probabilities (McFadden, 2001) states that for any pair of choice set C,C' such that $\{k, k'\} \in C$ and $C \subset C'$:

$$\Pr(k \text{ is chosen in } C') = \Pr(k \text{ is chosen in } C).\Pr(\text{An element of } C \text{ is chosen in } C').$$

Under this axiom, choice probabilities take a multinomial generalized logit form.
Moreover, assume that choices are associated with utility functions, $\{u^{(k)}\}_k$ that depend on determinants X_i and random shocks:

$$u^{(k)} = X\beta^{(k)} + \varepsilon^{(k)},$$

and that the actual choice of the decision maker yields maximum utility to her. Then, the IIA axiom is verified if and only if $\varepsilon^{(k)}$ are independent and extreme value distributed (McFadden, 1974). Extensions of decision theory under IIA were proposed in the continuous case (Resnick and Roy, 1991) or in an intertemporal context (Dagsvik, 2002).
The IIA axiom is a strong restriction as in the famous red and blue bus example where, if IIA is assumed, the existence of different colours affects choices of transport between bus and other modes while introspection suggests that colours should indeed be irrelevant. Several generalizations which proceed from logit were proposed to bypass IIA. Hierarchical or tree structures were the first to be used. At the upper level, the choice set consists of broad groups of alternatives. In each of these groups, there are various alternatives which can consist themselves of subsets of alternatives, and so on. The best-known model is the two-level nested logit, where alternatives are grouped by similarities. For instance, the first level is the choice of the type of the car, the second level is the make of the car. The formula of choice probabilities for nested logit,

$$p^{(k)} = \frac{\exp(X\beta^{(k)} / \lambda_{B_s})\left(\Sigma_{j\in B_s}\exp(X\beta^{(j)} / \lambda_{B_s})\right)^{\lambda_{B_s}-1}}{\Sigma_{t=1}^{T}\left(\Sigma_{j\in B_t}\exp(X\beta^{(j)} / \lambda_{B_s})\right)^{\lambda_{B_t}}},$$

where alternative k belongs to B_s, is not illuminating but the logic of construction is clear. Choices at each level are modelled as multinomial logit (Train, 2003).
General extreme value distributions (McFadden, 1984) provide more extensions, although they do not generate all configurations of choice probabilities. In contrast, mixed logit does, as shown by McFadden and Train (2000). Instead of considering that parameters are deterministic, make them random or heterogeneous across agents. The result is a mixture model where individual probabilities of choice are obtained by integrating out the random elements as in

$$p^{(k)} = \int p^{(k)}(\beta)\, f(\beta)\, d\beta.$$

Integrals are computed using simulation methods (MacFadden, 2001). The same principle is used by Berry, Levinsohn and Pakes (1995) with a view to generalizing the aggregate logit choice models using market data. Logit models are still very much in use in applied settings in demand analysis and marketing, and are equivalent to a representative consumer model (Anderson, de Palma and Thisse, 1992). Mixed logits permit much more general patterns of substitution between alternatives and should probably become the standard tool in the near future.

See Also

- categorical data
- econometrics
- hierarchical Bayes models
- maximum likelihood
- McFadden, Daniel
- mixture models
- nonlinear panel data models
- product differentiation
- rational behaviour
- utility

Bibliography

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Andersen, E.B. 1973. *Conditional Inference and Models for Measuring*. Copenhagen: Mentalhygiejnisk Forlag.

Anderson, S.P., de Palma, A. and Thisse, J.F. 1992. *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.

Arellano, M. 2003. Discrete choices with panel data. *Investigaciones Economicas* 27, 423–58.

Berkson, J. 1944. Application of the logistic function to bioassay. *Journal of the American Statistical Association* 39, 357–65.

Berkson, J. 1951. Why I prefer logits to probits. *Biometrics* 7, 327–39.

Berkson, J. 1955. Maximum likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association* 50, 130–62.

Berry, S.T., Levinsohn, J.A. and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–90.

Chamberlain, G. 1984. Panel data. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Chamberlain, G. 1992. Binary response models for panel data: identification and information. Unpublished manuscript, Harvard University.

Dagsvik, J. 2002. Discrete choice in continuous time: implications of an intertemporal version of IAA. *Econometrica* 70, 817–31.

Gouriéroux, C. 2000. *Econometrics of Qualitative Dependent Variables*. Cambridge: Cambridge University Press.

Gouriéroux, C., Monfort, A. and Trognon, A. 1985. Moindres carrés asymptotiques. *Annales de l'INSEE* 58, 91–121.

Horowitz, J. 1998. *Semiparametric Methods in Econometrics*. Berlin: Springer.

Horowitz, J.L. and Savin, N.E. 2001. Binary response models: logits, probits and semiparametrics. *Journal of Economic Perspectives* 15(4), 43–56.

Lancaster, T. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95, 391–413.

Lewbel, A. 2000. Semiparametric qualitative response model estimation with unknown heteroskedasticity or instrumental variables. *Journal of Econometrics* 97, 145–77.

Luce, R. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Magnac, T. 2004. Panel binary variables and sufficiency: generalizing conditional logit. *Econometrica* 72, 1859–77.

Manski, C.F. 1975. The maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C.F. 1988. Identification of binary response models. *Journal of the American Statistical*

Association 83, 729–38.

Marschak, J. 1960. Binary choice constraints and random utility indicators. In *Mathematical Methods in the Social Sciences*, ed. K. Arrow. Stanford: Stanford University Press.

Matzkin, R. 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica* 60, 239–70.

McCullagh, P. and Nelder, J.A. 1989. *Generalized Linear Models*. London: Chapman and Hall.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1984. Econometric analysis of qualitative response models. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

McFadden, D. 2001. Economic choices. *American Economic Review* 91, 351–78.

McFadden, D. and Train, K. 2000. Mixed MNL models for discrete responses. *Journal of Applied Econometrics* 15, 447–70.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Denmark Paedagogiske Institut.

Resnick, S.I. and Roy, R. 1991. Random USC functions, max stable process and continuous choice. *Annals of Applied Probability* 1, 267–92.

Strauss, D. 1992. The many faces of logistic regression. *American Statistician* 46, 321–27.

Theil, H. 1969. A multinomial extension of the linear logit model. *International Economic Review* 10, 251–9.

Thurstone, L. 1927. A law of comparative judgement. *Psychological Review* 34, 273–86.

Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

How to cite this article

Magnac, Thierry. "logit models of individual choice." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_L000222> doi:10.1057/9780230226203.0989

The New Palgrave Dictionary of Economics Online

lognormal distribution

P.E. Hart

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

central limit theorems; Gibrat's Law; lognormal distribution

Article

If there is a number, θ, such that $Y = \log_e(X - \theta)$ is normally distributed, the distribution of *X* is *lognormal*. The important special case of θ =0 gives the two parameter lognormal distribution, $X \sim \Lambda(\mu, \sigma^2)$ with $Y \sim N(\mu, \sigma^2)$, where μ and σ^2 denote the mean and variance of $\log_e X$. The classic work on the subject is by Aitchison and Brown (1957). A useful survey is provided by Johnson, Kotz and Balakrishnan (1994, ch. 14). They also summarize the history of this distribution: the pioneer contributions by Galton (1879) on its genesis, and by McAlister (1879) on its measures of location and dispersion, were followed by Kapteyn (1903), who studied its genesis in more detail and also devised an analogue machine to generate it. Gibrat's (1931) study of economic size distributions was a most important development because of his law of proportionate effect. Since then there has been an immense number of applications of the lognormal distribution in the natural, behavioural and social sciences.
Why does the lognormal distribution appear to occur so frequently? One plausible answer is based on the central limit theorems used to explain the genesis of a normal curve. If a large number of random shocks, some positive, some negative, change the size of a particular variable, *X*, in an additive fashion, the distribution of that variable will tend to become normal as the number of shocks increases. But if these shocks act multiplicatively, changing the value of *X* by randomly distributed proportions instead of absolute amounts, the central limit theorems apply to $Y=\log_e X$ which tends to be normally distributed. Hence *X* has a lognormal distribution.
The substitution of multiplicative for additive random shocks generates a positively skew, leptokurtic, lognormal distribution instead of the symmetric, mesokurtic normal curve. But the degree of skewness and kurtosis of the two-parameter lognormal curve depends solely on σ^2, so if this is low enough, the lognormal approximates the normal curve. The important difference is that *X* cannot take zero or negative values which may make the lognormal distribution a more appropriate representation of

variables, such as height and weight, which must take positive values. Clearly, the widespread occurrence of positive variables in practice, coupled with the great flexibility of the shape of the lognormal, provide further reasons for its frequent application.

See Also

- Gini ratio
- inequality (measurement)
- Lorenz curve
- Pareto distribution

Bibliography

Aitchison, J. and Brown, J.A.C. 1957. *The Lognormal Distribution*. Cambridge: Cambridge University Press.

Galton, F. 1879. The geometric mean in vital and social statistics. *Proceedings of the Royal Society of London* 29, 365–7.

Galton, F. 1889. *Natural Inheritance*. London: Macmillan.

Galton, F. 1892. *Hereditary Genius*. London: Macmillan.

Gibrat, R. 1931. *Les inégalites économiques*. Paris: Libraire du Recueil Sirey.

Johnson, N., Kotz, S. and Balakrishnan, L. 1994. *Continuous Univariate Distributions*, vol. 1. New York: John Wiley.

Kapteyn, J.C. 1903. *Skew Frequency Curves in Biology and Statistics*. Astronomical Laboratory, Groningen: Noordhoff.

McAlister, D. 1879. The law of the geometric mean. *Proceedings of the Royal Society of London* 29, 367–75.

How to cite this article

Hart, P. E. "lognormal distribution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000137> doi:10.1057/9780230226203.0990

The New Palgrave Dictionary of Economics Online

long memory models

P.M. Robinson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Time series exhibiting varying forms of strong dependence are considered. Stationary parametric and semiparametric models, and their estimation, are first discussed. We go on to review nonlinear, nonstationary and multivariate models.

Keywords

ARMA processes; cointegration; Fourier frequencies; fractional autoregressive integrated moving average (FARIMA); fractional noise; generalized method of moments (GMM); long memory models; maximum likelihood; multivariate models; nonlinear models; nonstationary models; semiparametric estimation; statistical inference; time series analysis; Whittle estimates

Article

Much analysis of economic and financial time series focuses on stochastic modelling. Deterministic sequences, based on polynomials and dummy variables, can explain some trending or cyclic behaviour, but residuals typically exhibit serial dependence. Stochastic components have often been modelled by stationary, weakly dependent processes: parametric models include stationary and invertible autoregressive moving average (ARMA) processes, while a non-parametric approach usually focuses on a smooth spectral density. In many cases, however, we need to allow for a greater degree of persistence or ‘memory’. This is characterized by stationary time series whose autocorrelations are not summable or whose spectral densities are unbounded, or by non-stationary series evolving over time. The latter are partly covered by unit root processes, but considerably greater flexibility is possible.

Basic models

Early empirical evidence of slowly decaying autocorrelations emerged long ago, in analyses of

astronomical, chemical, agricultural and hydrological data, and then in economics and finance. A stationary parametric model which attracted early interest is ‘fractional noise’. Let x_t, $t = 0, \pm 1, \ldots$, be a covariance stationary discrete time process, so its autocovariance $cov(x_t, x_{t+u})$ depends only on u, and thus may be denoted by γ_u. Then fractional noise x_t has autocovariance

$$\gamma_u = \gamma_0 \left\{ |u+1|^{2d+1} - 2|u|^{2d+1} + |u-1|^{2d+1} \right\}, \; u = 0, \pm 1, \ldots, \tag{1}$$

where the parameter d is called the ‘memory parameter’, and satisfies $-\frac{1}{2} < d < \frac{1}{2}$. When $d = 0$ (1) implies that $\gamma_u = 0$ for $u \neq 0$, so x_t is white noise. But if $0 < d < \frac{1}{2}$, we have

$$\gamma_u \sim 2d\left(d + \frac{1}{2}\right)\gamma_0 |u|^{2d-1}, \text{ as } |u| \to \infty, \tag{2}$$

where ‘~’ means that the ratio of left- and right-hand sides tends to one. It follows from (2) that γ_u does decrease with lag u, but so slowly that

$$\sum_{u=-\infty}^{\infty} \gamma_u = \infty. \tag{3}$$

In the frequency domain, when x_t has a spectral density $f(\lambda)$, $\lambda \in (-\pi, \pi)$ given by

$$f(\lambda) = (2\pi)^{-1} \sum_{u=-\infty}^{\infty} \gamma_u \cos(u\lambda), \; \lambda \in (-\pi, \pi),$$

the property (3) is equivalent to

$$f(0) = \infty, \tag{4}$$

and more precisely a fractional noise process x_t has spectral density satisfying

$$f(\lambda) \sim C\lambda^{-2d}, \text{ as } \lambda = 0+ . \tag{5}$$

In general we can regard (3) and (4) as basic indicators of a ‘long memory’ process x_t, and (2) and (5) as providing more detailed description of autocorrelation structure at long lags, or spectral behaviour at low frequencies. By contrast, if x_t were a stationary ARMA, γ_u would decay exponentially and $f(\lambda)$ would be analytic at all frequencies. The structure (5) is similar to Granger's (1966) ‘typical spectral shape of an economic variable’.

The model (1) is connected with the physical property of ‘self-similarity’, and, so far as economic and financial data are concerned, found early application in work of Mandelbrot (1972) and others. However, (1) imposes a very rigid structure, with autocorrelations decaying monotonically and depending on a single parameter. In addition, though a formula for $f(\lambda)$ corresponding to (1) can be written down, it is complicated, and (1) does not connect well mathematically with other important time series models, and does not lend itself readily to forecasting.

An alternative class of ‘fractionally integrated’ processes leads to a satisfactory resolution of these concerns. This is conveniently expressed in terms of the lag operator L, where $Lx_t = x_{t-1}$. Given the formal expansion

$$(1-s)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} s^j,$$

we consider generating x_t from a zero-mean stationary sequence u_t, $t = 0, \pm 1, \ldots$, by

$$(1-L)^d (x_t - \mu) = v_t, \tag{6}$$

where $\mu = Ex_t$ and

$$|d| < \frac{1}{2}.$$

If v_t has absolutely summable autocorrelations, that satisfy some mild additional conditions, both the properties (2) and (5) hold. In the simplest case of (6), v_t is a white noise sequence. Then γ_u decays monotonically when $d \in (0, \frac{1}{2})$ and indeed behaves very much like (1). This model may have originated in Adenstedt (1974), though he stressed the case $d \in (-\frac{1}{2}, 0)$, where x_t is said to have 'negative dependence' or 'antipersistence'. Taking v_t to be a stationary and invertible ARMA process, with autoregressive order *p* and moving average order *q*, gives the FARIMA (*p*, *d*, *q*) process of Granger and Joyeux (1980). In principle, the short memory process v_t in (6) can be specified in any number of ways so as to yield (2) and/or (5); a process satisfying this condition is sometimes called *I*(*d*).

Statistical inference

Given observations x_t, $t = 1, \ldots, n$ there is interest in estimating *d*. If v_t has parametric autocorrelation, as when x_t is a FARIMA (*p*, *d*, *q*), one can form a Gaussian maximum likelihood estimate of *d* and any other parameters. This estimate has the classical properties of being $n^{1/2}$ – consistent and asymptotically normal and efficient. Computationally somewhat more convenient estimates, called Whittle estimates, have the same asymptotic properties. Indeed, for standard FARIMA (*p*, *d*, *q*) parameterizations, say, the estimates of *d* and of ARMA coefficients have asymptotic variance matrix that is unaffected by many departures from Gaussianity. Though these asymptotic properties are of the same type as one obtains for estimates of short memory processes, such as ARMAs, their proof is considerably more difficult (see Fox and Taqqu, 1986), due to the spectral singularity (4). In econometrics, generalized method of moments (GMM) estimation has become very popular, and GMM estimates have been proposed for long memory models. However, unless a suitable weighting is used, they are not efficient under Gaussianity, are not more robust asymptotically to non-Gaussianity, and are not even asymptotically normal when $d > \frac{1}{4}$.

If the parametric autocorrelation is mis-specified, for example if in the FARIMA (*p*, *d*, *q*) *p* or *q* are chosen too small or both are chosen too large, then the procedures described in the previous paragraph will generally produce inconsistent estimates of *d*, as well as of other parameters. Essentially, the attempt to model the short memory component of x_t damages estimation of the long memory component. This difficulty can be tackled by a 'semiparametric' approach, if one regards the local or asymptotic specifications (2) or (5) as the model, and estimates *d* using only information in low

frequencies or in long lags. Frequency domain versions are by far the more popular here, having the nicest asymptotic statistical properties. In the log periodogram estimate of *d*, logged periodograms are regressed on a logged local approximation to $f(\lambda)$, over the *m* Fourier frequencies closest to the origin (Geweke and Porter-Hudak, 1983), *m* having the character of a bandwidth number similar to those used in smoothed nonparametric functional estimation. An alternative approach optimizes a local Whittle function, again based on the lowest *m* Fourier frequencies (Künsch, 1987). In the asymptotics for both types of estimate (see Robinson, 1995a; 1995b) *m* must increase with *n*, but more slowly (to avoid bias); both the log periodogram and local Whittle estimates are $m^{1/2}-\text{consistent}$ and asymptotically normal, with the latter the more efficient (though it is computationally more onerous, being only implicitly defined). Because both converge more slowly than estimates of correctly specified parametric models, a larger amount of data may be necessary for estimates to be reasonably precise. Moreover, estimates are sensitive to the choice of *m*. However, automatic and other rules are available for determining *m*; and semiparametric methods of estimating memory parameters have become very popular not only because of the robust character of the asymptotic results, but because of their relative simplicity.

The long memory processes we have been discussing exhibit an excess of low frequency power (5). But one can also consider parametric or semiparametric models for a spectral density with one or more poles at non-zero frequencies. These models can be used to describe seasonal or cyclic behaviour (see Arteche and Robinson, 2000). It is also possible to estimate the unknown location of a pole, that is, cycle (see Giraitis, Hidalgo and Robinson, 2001).

Nonlinear models

In non-Gaussian series, not all information is contained in first and second moments. In particular, in many financial series observations x_t may appear to have little or no autocorrelation, but instantaneous nonlinear functions, such as squares x_t^2, exhibit long memory behaviour. We can develop models to describe such phenomena. For example, let

$$x_t = \varepsilon_t h_t, \tag{7}$$

where x_t is a sequence of independent and identically distributed random variables with unit variance, whereas h_t is a stationary autocorrelated sequence, such that ε_s and h_t are independent for all *s*,•*t*. Then for all $u \neq 0$, $cov = (x, x_{t+u}) = 0$ but $cov(x_t^2, x_{t+u}^2) = cov(h_t^2, h_{t+u}^2)$, which in general can be non-zero. In particular, if h_t^2 has long memory, so has x_t^2. In a more fundamental modelling we can take h_t to be a nonlinear function of an underlying long memory Gaussian processes, with the functional form of *h* determining the extent of any long memory in h_t^2; these issues were discussed in some generality by

Robinson (2001). The models form a class of long memory stochastic volatility models, whose estimation has been discussed by Hurvich, Moulines and Soulier (2005), for example.

The fractional class (6) can be modified or extended to describe a wide class of nonstationary behaviour. For $d \geq \frac{1}{2}$ the variance of x_t (6) explodes, but we can consider truncated versions such as

$$x_t = (1 - L)^{-d}\{v_t 1(t \geq 1)\}$$

where $1(\cdot)$ is the indicator function, or

$$x_t = (1 - L)^{-k}\{w_t 1(t \geq 1)\}$$

for integer $k \geq 1$, where w_t is a stationary $I(c)$ process, $|c| < \frac{1}{2}$, and $d = k + c$. In either case we might call x_t a (nonstationary) $I(d)$ process, for $d \geq \frac{1}{2}$. Both models include the unit root case $d = 1$ that has proved so popular in econometrics. However, the fractional class $I(d)$, for real-valued d, bridges the gap between short memory and unit root processes, allowing also for the possibility of arbitrarily long memory d. The ‘smoothness’ of the $I(d)$ family is associated with classical asymptotic theory, which is not found in autoregressive based models around a unit root. Robinson (1994) showed that Lagrange multiplier tests for the value of d, and any other parameters, have asymptotic null χ^2 distributions for all real d. Also, under nonstationary suitably modified parametric and semiparametric methods of estimating d, extending those for the stationary case, tend still to be respectively $n^{1/2}-$ and $m^{1/2}-$consistent, and asymptotically normal, unlike, say, the lag-one sample autocorrelation of a unit root series.

Multivariate models

Often in economics and finance we are concerned with a vector of jointly dependent series, so x_t is vector-valued. Such series can be modelled, either parametrically or semiparametrically, to have long memory, with different elements of x_t possibly having different memory parameters, and being stationary or nonstationary. Methods of statistical inference developed for the univariate case can be extended to such settings. However, multivariate data introduces the possibility of (fractional) cointegration, where a linear combination of x_t (the cointegrating error) can have smaller memory parameter than the elements of x_t. Cointegration has been extensively developed for the case x_t is $I(1)$ and cointegrating errors are $I(0)$, and methods developed for this case can fail to detect fractional

cointegration. Moreover, it is possible for stationary series, not only nonstationary ones, to be fractionally cointegrated, as seems relevant in financial series. In either case, methods of analysing cointegration that allow memory parameters of observables and cointegrating errors to be unknown (see, for example, Hualde and Robinson, 2004) afford considerable flexibility.

See Also

- central limit theorems
- econometrics
- nonparametric structural models
- semiparametric estimation
- time series analysis

Research supported by ESRC Grant R000239936.

Bibliography

Adenstedt, R. 1974. On large-sample estimation for the mean of a stationary random sequence. *Annals of Statistics* 2, 1095–107.

Arteche, J. and Robinson, P. 2000. Semiparametric inference in seasonal and cyclic long memory processes. *Journal of Time Series Analysis* 21, 1–25.

Fox, R. and Taqqu, M.S. 1986. Large sample properties of parameter estimates of strongly dependent stationary Gaussian time series. *Annals of Statistics* 14, 517–32.

Geweke, J. and Porter-Hudak, S. 1983. The estimation and application of long memory time series models. *Journal of Time Series Analysis* 4, 221–38.

Giraitis, L., Hidalgo, J. and Robinson, P. 2001. Gaussian estimation of parametric spectral density with unknown pole. *Annals of Statistics* 29, 987–1023.

Granger, C. 1966. The typical spectral shape of an economic variable. *Econometrica* 34, 150–67.

Granger, C. and Joyeux, R. 1980. An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 15–39.

Hualde, J. and Robinson, P. 2004. Semiparametric estimation of fractional cointegration. Mimeo, London School of Economics.

Hurvich, C., Moulines, E. and Soulier, P. 2005. Estimating long memory in volatility. *Econometrica* 73, 1283–328.

Künsch, H. 1987. Statistical aspects of self-similar processes. In *Proceedings of the First World Congress of the Bernoulli Society*, vol. 1, ed. Y. Prohorov and V. Sazonov. Utrecht: VNU Science Press.

Mandelbrot, D. 1972. Statistical methodology for non-periodic cycles: from the covariance to R/S analysis. *Annals of Economic and Social Measurement* 1, 259–90.

Robinson, P. 1994. Efficient tests of nonstationary hypotheses. *Journal of the American Statistical Association* 89, 1420–37.

Robinson, P. 1995a. Log-periodogram regression of time series with long range dependence. *Annals of Statistics* 23, 1048–72.

Robinson, P. 1995b. Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* 5, 1630–61.

Robinson, P. 2001. The memory of stochastic volatility models. *Journal of Econometrics* 101, 192–218.

How to cite this article

Robinson, P. M. "long memory models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000234> doi:10.1057/9780230226203.0991

The New Palgrave Dictionary of Economics Online

long run and short run

Carlo Panico and Fabio Petri

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The notion of long-run and short-run equilibrium was introduced by Marshall in 1890 and reflected the ‘long-period method’ of analysis in use among classical political economists since the 18th century. In the early 1930s, dissatisfaction with some of the neoclassical conclusions led to a shift to different methods of analysis and to the introduction of new equilibrium notions. These changes, together with the tendency to use old terminology for new equilibrium concepts, have deprived the terms ‘short-period’ and ‘long-period’ of a uniform meaning and have been a source of confusion and misunderstandings in recent debates on theoretical and applied work.

Keywords

average interest rate; capital endowment; circulating capital; diminishing marginal returns; effective demand; fixed capital; general equilibrium; intertemporal equilibria; Keynes, J. M.; long-period method; long-run and short-run; market price; Marshall, A.; Marx, K. H.; natural price; natural rate and market rate of interest; partial equilibrium; secular equilibrium; short-run and long-run equilibrium; stationary equilibria; stationary state; supply and demand; temporary equilibrium; underemployment equilibrium; value and distribution, neoclassical theory of

Article

The distinction between long-run and short-run (or long-period and short-period) equilibrium, introduced by Marshall (see Marshall, 1890, pp. 363–80; hints at this distinction are also to be found in some of Marshall's early works, dated 1870–71, recently re-presented in Whitaker, 1975, pp. 119–64), reflected a method which was the generally accepted one at the time, and essentially the same as the method of the classical political economists and of Marx. The use of the method was not affected by the deep change undergone by the theory of value and distribution around the 1870s with the advent of what is nowadays called the ‘neoclassical’ school. This method, called ‘method of long-period

positions' (Garegnani, 1976), however, has been abandoned in much of the modern mainstream work on value. Further, there is no uniform meaning attributed to the terms 'short-period' and 'long-period', but rather a variety of usages depending on the theoretical framework of the writer, a situation responsible for many misunderstandings and debates at cross purposes.

The classical political economists

Since its origin in the writings of 18th-century authors, economic theory has used what has been subsequently named the 'long-period method' of analysis to investigate how production, distribution and accumulation take place within a market economy. According to Quesnay and A. Smith, the system 'market economy' produces results which are 'independent of men's will' (Quesnay, 1758). Competition, Smith thought, tends to establish uniformity in the 'average' or 'natural' rates of wages, profits and rent. 'Market' prices, that is, observed prices, thus tend to gravitate towards their 'natural' levels (also called 'average prices' or 'prices of production'), defined as those which allowed the payment of wages, profits and rents at their average or natural rates (Smith, 1776, pp. 57–61). According to the classical political economists, a divergence between the 'market' and the 'natural' price of a commodity is caused by a divergence between the amount supplied by producers and the 'effectual demand' for it, that is, 'the demand of those who are willing to pay the natural price of the commodity, or the whole value of rent, labour and profit, which must be paid in order to bring it thither' (Smith, 1776, p. 58). This divergence implies windfall profits or losses for that commodity. If supply coincides with 'effectual demand', 'market' price corresponds to 'natural' price. The rate of profit earned in that sector is equal to the one which is uniformly earned in the whole economy. Equilibrium conditions are said to prevail. Within this approach, therefore, fluctuations of supply and demand explain nothing but the deviations of 'market' prices from 'natural' prices.

The idea that the interaction of competitive market forces pushes the actual level of economic variables towards their 'natural' or 'average' level was applied to different fields of economic theory. Marx, for instance, applied it to the analysis of the 'market' and the 'average' interest rate (see Marx, 1972, pp. 355–66). The latter rate, according to Marx, was determined by 'the average conditions of competition, the balance between lender and borrower' (Marx, 1972, p. 363) in the money market over a certain historical period (Marx, 1972, p. 363). He rejected previous views determining this rate in terms of 'natural' laws, like the rate of growth of timber in central Europe forests (Marx, 1972, p. 363 n.) or in terms of the rate of return on capital invested in the productive sectors depending upon the material or technological conditions of production of commodities (Marx, 1972, p. 363). In his historically relative determination, the 'average' interest rate, being constrained by no 'natural' or 'material' law, can be at any level. At the same time, the interaction of demand and supply determines the daily variations of the 'market' interest rate and makes it converge towards its 'average' level.

The application of the 'long-period method' to the analysis of the interest rate makes it clear that the essential element of the method is the reference to an 'average' or 'normal' position around which the actual values of the variable considered gravitate. Reference to the attainment of a uniform rate of profit in all sectors is not strictly necessary if the theory does not determine the variable considered on the basis of the technological conditions of production. In Marx's analysis, since the 'average' interest rate is independent of the rate of profits, it is possible to separate the study of the factors determining the

former rate from the study of the technological links between distributive variables and commodity prices, where competitive forces set in motion a gravitation process when windfall profits or losses appear in particular industries. The notion of ‘average’ interest rate, which may be used to identify a position of long-period equilibrium for this variable, can thus be introduced and analysed by referring to a normal position of this variable, which has actually prevailed over a certain period, without making reference to a uniform rate of profits. In a theory determining the ‘natural’ interest rate on the basis of technological conditions of production, instead, no separation can be made between the analysis of the average interest rate and that of the links between commodity prices and distributive variables. In this case, the condition of a uniform rate for return on capital defines the ‘long-period equilibrium’ position for both commodity prices and interest rate.

The rise of neoclassical economics

The long-period method was also used by those economists (like Walras, Menger, Jevons, Böhm-Bawerk, J.B. Clark, Wicksell, et al.) who some years later introduced and developed the ‘neoclassical’ theory of value and distribution. *No question was raised by these authors as to the use of this method*. The new theory, unlike the previous one, determined prices, output and distribution simultaneously. The ‘natural’ or ‘equilibrium’ values of all these variables (including the interest rate and the level of activity in the economy, which turns out to be a full employment level) depended, among other things, upon the technological conditions of production and were thus associated with the attainment of a uniform rate of profits in the economy.

Among the earlier neoclassical economists, Marshall deserves special consideration, since he introduced the notion of short- and long-period equilibrium (see Marshall, 1890, pp. 80). In his writings, Marshall tried to show how the neoclassical principles of price determination in terms of supply and demand functions could be applied to analytical levels which were closer to actual events. He thus analysed price determination for each single market (partial equilibrium) and within this analysis he referred to three different notions of equilibrium (temporary, short-period and long-period), which differed as to the conditions determining the supply functions. In a temporary equilibrium, it was supposed, there is no time to change the supply of the commodity. The amount supplied is fixed and the equilibrium price is that which allows that quantity to be demanded.

Analyses of short-period equilibrium assume that there is time to change supply through production, but there is no time to change the structure of *fixed capital* goods existing in *that* industry. This assumption constrains the technological possibilities of production. As in the case of temporary equilibrium, short-period equilibrium is compatible with windfall profits or losses.

In long-period analyses, it is assumed instead that there is time to adapt the structure of fixed capital goods of the industry so that quasi-rents (that is, entrepreneurial net profits) disappear. The price then guarantees just the ‘normal rate of profits’ (that is, the ‘equilibrium’ real rate of return on capital which is uniform in the whole economy).

Marshall's partial equilibrium analysis appears to rely on general equilibrium analysis for the determination of the ‘equilibrium’ rate of return on capital and of ‘ceteris paribus’ prices. The view that the ‘general equilibrium’ analysis was logically prior appears accepted in some major contributions of the debate on Marshall's theory of value of the 1920s and the early 1930s (see Sraffa, 1925 and 1926,

and Pigou's reply, 1927). Marshall's starting point thus was the same as that of Walras, Wicksell, and of the other neoclassical economists mentioned above.

Long-period general equilibrium must not be confused with ‘secular’ equilibrium, which results from allowing enough time for factor endowments to change under the influences of demographic factors and propensity to save, so as to cause the economy to reach ‘stationary’ or ‘steady growth’ conditions (see Robbins, 1930).

Short- and long-period in Keynes

By the end of the 1920s, dissatisfaction with the neoclassical conclusions as to the level of activity of the economy and with the analysis of capital led some economists to new analytical developments, which affected for the first time the method used too.

J.M. Keynes criticized the neoclassical conclusion that the market economy has an inherent tendency towards full employment. In the preparatory works and in the introduction to the *General Theory* he insisted that his concern was not the analysis of the temporary and cyclical fluctuations of the level of activity, but the theory dealing with the more fundamental forces which tend to prevail in the economic system (see Keynes, 1936, pp. 4–5; 1973; pp. 405–7; and 1979, pp. 54–7). He wanted thus to replace the neoclassical long-period theory of the level of output with a new one. Yet the way he presented his new theory has raised many problems of interpretation also related to the method used.

First of all, Keynes stated in his book that he assumed as given the structure of *fixed capital* goods existing in the economy. This can lead to consider his theory as a short-period one, arguing that it would determine the level of capacity utilization in the economy. It is difficult, however, to support this interpretation also with the argument that in the *General Theory* Keynes was following Marshall's definition of short-period, which was confined to partial equilibrium analysis. Marshall knew that the time required for adjustment of the structure of fixed capital goods differed from one industry to the other, so that it would have been unreasonable to extend the hypothesis of a fixed structure from one industry to the whole economy, as Keynes did. This element of ambiguity as to the use of the concepts has raised many puzzling questions among the interpreters of Keynes.

At the same time, Keynes explicitly stated that his theory was meant to explain why the level of employment, over a specific historical period, oscillates round an intermediate or average position (often not a full-employment one), whereas in other periods it oscillates round a different one (Keynes, 1936, p. 254). This reference to ‘specific historical periods’ and to ‘average or normal positions’ can lead to consider Keynes's theory as a long-period one, in the same way as Marx's theory of the ‘average’ interest rate. The assumption of a fixed structure of capital goods would thus play a secondary role in Keynes's theory.

Besides, Keynes hinted towards an analysis of accumulation which emphasizes the role played by effective demand (Keynes, 1936, p. 372–80). The trend followed by a growing economy in which adjustment in the structure of fixed capital goods has occurred, is affected by the level of effective demand. The possibility of assuming in this analysis an adjusted structure of fixed capital goods (to which a uniform rate of profits corresponds) can lead to consider this as the *long-period theory* present in the *General Theory*.

Finally, the maintenance in the *General Theory* of elements belonging to the neoclassical tradition, like

the acceptance of the principle of diminishing marginal returns for capital from which the existence of a full-employment level of the rate of interest is derived (see Keynes, 1936, pp. 147–8, 178, 203, 235 and 243; Keynes, 1973, pp. 456, 615, 630) has allowed some interpreters to consider Keynes's ‘underemployment equilibrium’ as a situation in which market forces have not yet worked out their effects fully, consequently defining it as a position of ‘short-period equilibrium’ (see Patinkin, 1976, pp. 116–19; Winch, 1969, p. 167).

The presence of several lines of development of its basic principle (that of effective demand) and the lack of precision and coherence as to the concepts and the analytical elements used appear to be an endless source of discussion as to the interpretation of Keynes's work. The existing evidence does not seem to support, however, the view that the *General Theory* wanted to move along the same lines as Hayek, Hicks and others, who in those years were proposing the neoclassical theory of value, distribution and the level of output on the basis of a method of analysis different from the long-period one.

Post-Walrasian developments

In the same years, dissatisfaction with the neoclassical analysis of capital was leading to a shift in method, owing to the adoption of what may be called ‘post-Walrasian’ notions of general equilibrium, elaborated by Hayek and Lindahl around the 1930s, but first proposed to a wider audience in 1939 by Hicks's *Value and Capital* (see Garegnani, 1976; Milgate 1979). The change in method derives from the change in the treatment of the capital endowment.

In the traditional neoclassical treatment, dominant up to the 1950s, the conception of equilibrium as a centre of gravitation of time-consuming adjustments (a conception incompatible with taking as given the equilibrium endowments of the several capital goods) had been reconciled with the supply-and-demand approach to factor pricing by conceiving capital as a *single* factor of production, capable of changing ‘form’ (that is, of embodying itself into different vectors of heterogeneous capital goods) without changing in ‘quantity’, so that its ‘form’ (that is, composition) could be left to be determined by the equilibrium condition of a uniform rate of return on the supply price of capital goods – the distinguishing element of long-period positions. Capital so conceived had ultimately to be measured as an amount of value, because in equilibrium different capital goods earn rewards proportional to their values. Within the neoclassical framework, therefore, the reference to a homogeneous factor ‘capital’, a value magnitude, was a logical necessity, entailed by the attempt to explain distribution through the equilibrium between demand for and supply of ‘factors of production’, without abandoning the traditional method of long-period positions (Petri, 2004). With one exception, this conception of capital was in fact more or less explicitly adopted by all founders of neoclassical theory and it was the target of the Cambridge critique of the 1960s (Harcourt, 1969; Garegnani, 1970). The only exception had been Walras, who intended as well to determine a long-period equilibrium and accordingly maintained the uniform-profit-rate condition, but took as data the endowment of each kind of capital goods, with the result that his model was generally devoid of solutions.

Walras's treatment of the capital endowment as a given vector is maintained in post-Walrasian general equilibrium analyses, but the condition of uniform profit rate on supply price is dropped. Existing capital goods are treated like natural resources; commodities are dated, so prices of future commodities are

distinguished from prices of currently available commodities; and the current composition of the production of new capital goods is determined in either of two ways: by assuming the existence of complete futures markets (intertemporal equilibria, see for example Debreu, 1959), or through the introduction of expectations among the data (temporary equilibria, see for example Hicks, 1939 and Grandmont, 1977).

The difference between the notion of equilibrium entailed by such a treatment of capital and that entailed by the long-period method of analysis warrants emphasis (Garegnani, 1990). The latter attempts to represent states of the economy which have the role of centres of gravitation of observed day-to-day magnitudes: chance movements away from such a state set off forces tending to bring the economy back to it. Changes in the economy can then be studied by comparing the long-period positions corresponding to the situation before and after the change. Post-Walrasian equilibria cannot have such a role, because they rely on data some of which (the endowments of capital goods and, where futures markets are not complete, expectations) would be altered by any chance deviation from the equilibrium: thus the forces set off by this deviation would not tend to bring the economy back to the same equilibrium. For the same reason, stability questions relative to post-Walrasian equilibria can only be asked for imaginary atemporal adjustment processes which exclude the implementation of disequilibrium production decisions before the equilibrium is reached.

A variety of usages

The introduction of new equilibrium concepts, together with the tendency to overlook the existence of differences with previous ones and to use the same terminology for the former and for the latter, has been a source of confusion and misunderstandings in recent debates on theoretical and applied work. The term ‘short-period equilibria’ has been sometimes applied to post-Walrasian equilibria (including ‘fix price’ equilibria with quantity adjustments, which share the same impermanence of data). On other occasions, Keynes's notion of equilibrium has been identified with temporary equilibrium. In both cases, the very great difference between Marshall's and Keynes's analyses on one side and post-Walrasian analyses on the other side has been neglected: in post-Walrasian models, *all* capital goods, including circulating capital goods, are given, while in Marshall's short-period analyses only the fixed plant of a single industry is a datum, and in Keynes's work only the fixed capital goods of the whole economy are given.

At the same time, the term ‘long-period equilibrium’ has been used in recent years to refer (*a*) to post-Walrasian intertemporal equilibria with futures markets extending far into the future; (*b*) to sequences of temporary equilibria; (c) to stationary or steady-growth equilibria. In all these cases, an incomplete grasp of the changes introduced in the notion of equilibrium appears to emerge.

Finally, modern neoclassical economists sometimes develop applied analyses using the traditional method of long-period positions, although rejecting, as their theoretical foundations, the traditional versions of neoclassical theory in favour of the post-Walrasian ones, which are not compatible with that method.

See Also

- Marshall, Alfred

Bibliography

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37, 407–36.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in Modern Capital Theory*, ed. M. Brown, K. Sato and P. Zarembka. Amsterdam: North-Holland.

Garegnani, P. 1990. Quantity of capital. In *Capital Theory*, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45, 535–72.

Harcourt, G.C. 1969. Some Cambridge controversies in the theory of capital. *Journal of Economic Literature* 7, 369–405.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Keynes, J.M. 1936. The General Theory of Employment, Interest and Money. In *The Collected Writings of J.M. Keynes*, vol. VII, ed. D. Moggridge. London: Macmillan.

Keynes, J.M. 1973. The General Theory and After. Part I: Preparation. In *The Collected Writings of J.M. Keynes*, vol. XIII, ed. D. Moggridge. London: Macmillan.

Keynes, J.M. 1979. The General Theory and After: A Supplement. In *The Collected Writings of J.M. Keynes*, vol. XXIX, ed. D. Moggridge. London: Macmillan.

Marshall, A. 1890. *Principles of Economics*. 8th edn, London: Macmillan, 1920.

Marx, K. 1972. *Capital*, vol. 3. London: Lawrence & Wishart.

Milgate, M. 1979. On the origin of the notion of ‘intertemporal equilibrium’. *Economica* 46, 1–10.

Patinkin, D. 1976. *Keynes’ Monetary Thought*. Durham, NC: Duke University Press.

Petri, F. 2004. *General Equilibrium, Capital and Macroeconomics. A Key to Recent Controversies in*

Equilibrium Theory. Cheltenham: Edward Elgar.

Pigou, A.C. 1927. The laws of diminishing and increasing cost. *Economic Journal* 37(June), 188–97.

Quesnay, F. 1758. *Tableau économique*. Reproduced and trans. as *Quesnay's Tableau économique*, ed. M. Kuczynski and R.L. Meek, London: Macmillan, 1972.

Robbins, L. 1930. On a certain ambiguity in the conception of stationary equilibrium. *Economic Journal* 40(June), 194–214.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. E. Cannan, London: Methuen, 1904.

Sraffa, P. 1925. Sulle relazioni fra costo e quantità prodotta. *Annali di Economia* 2, 277–328.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36(December), 535–50.

Whitaker, J.K., ed. 1975. *The Early Economic Writings of Alfred Marshall, 1867–1890*. London: Macmillan.

Winch, D. 1969. *Economics and Policy: A Historical Study*. London: Hodder & Stoughton.

How to cite this article

Panico, Carlo and Fabio Petri. "long run and short run." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000192> doi:10.1057/9780230226203.0992

The New Palgrave Dictionary of Economics Online

Longfield, Mountifort (1802 – 1884)

R.D. Collison Black

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Longfield was born at Desertserges, Country Cork, Ireland, in 1802. Although he graduated from Trinity College, Dublin, in 1823 with first class honours in natural sciences, he was elected a Fellow of his college in 1825 as ‘jurist’. His subsequent career was primarily in real property law, but when Archbishop Whately founded the professorship of political economy at Trinity College, Dublin, in 1832, Longfield was the successful candidate and became the first holder of the chair, from 1832 until 1836. In 1834 he was appointed Regius Professor of Feudal and English Law and in 1849 became one of the first Commissioners of the newly established Irish Incubered Estates Commission. When this was transmuted into the Landed Estates Court in 1858, Longfield was appointed a Judge of that court, retiring in 1867. He died in Dublin in 1884.

In 1847 he was one of the founder members of the Dublin Statistical Society (later re-named the Statistical and Social Inquiry Society of Ireland) and followed Whately as its President in 1863, but his many other public services derived primarily from his positions as advocate and judge. In his later years Longfield never returned to political economy but continued to write on questions of Irish land tenure and social reform.

The three volumes of lectures which Longfield published during his tenure of the Whately chair attracted little attention at the time, but have since been recognized as containing contributions to economic theory of outstanding originality. In his *Lectures on Political Economy* (1834a) Longfield dealt with the central issues of classical theory, those of value and distribution, in a manner which displayed a very clear grasp of the structure of Ricardian theory, but which in content diverged fundamentally from Ricardo's approach. He laid stress on the determination of market rather than natural values and presented remarkably complete demand-and-supply theory supplemented by elements of utility analysis. Perhaps his most original contribution was made in the area of distribution, where he formulated a theory of profits as determined by the marginal productivity of physical capital and a theory of wages as determined by the specific productivity of the labourer.

Longfield rejected the idea that the ‘natural price’ of labour was determined by subsistence, arguing that the ‘wages of the labourer depend upon the value of his labour and not upon his wants’ (1834a, p. 206). Although, like Ricardo, Longfield predicted a rise in rents, a fall in profits and a rise in wages in the progress of society, his view of the long-term prospects for economic growth was optimistic. He

expected the effects of increased population to be offset by technical progress in agriculture, and foresaw many benefits from the increased accumulations of capital which would lower profits, not least among them the increased productivity of labour, which would raise wages.

Longfield's two other published courses of lectures are more concerned with current economic problems, but his *Lectures on Commerce* (1835) contained several anticipations of later developments in international trade theory. His analysis of the causes of international specialization extended to all variations in factor endowments and he specifically treated the case of trade in more than two commodities, showing that each country would tend to export those commodities in which the productivity of its labour was above average and import those in which it was below average.

In his *Lectures on Poor Laws* (1834b) Longfield endorsed Senior's stern principle that assistance to the able-bodied should be confined to the barest subsistence – perhaps, ironically, because of the very optimism of his views about the likely trends of profits and wages. On the other hand, he favoured generous public assistance to those unable, through age or disability, to fend for themselves – even to the extent of advocating non-contributory old-age pensions. Longfield repeated this proposal in 1872, when he specifically considered state interference with the distribution of wealth; unlike most of his contemporaries he was then prepared also to advocate public dispensaries and hospitals to which access would not be means-tested, improved sanitary regulation of housing standards, free public education and improved public recreation facilities.

Longfield's economic writings appear to have had little influence on his contemporaries, but since his rediscovery by Seligman (1903) the originality of his contributions has come to be generally recognized.

Selected works

1834a. *Lectures on Political Economy, delivered in Trinity and Michaelmas Terms, 1833*. Dublin: R. Milliken & Son. Reprinted, 1971.

1834b. *Four Lectures on Poor Laws Delivered in Trinity Term, 1834*. Dublin: R. Milliken & Son. Reprinted, 1971.

1835. *Three Lectures on Commerce, and one on Absenteeism, delivered in Michaelmas Term, 1834*. Dublin: William Curry, Junior & Co. Reprinted, 1971.

1840. Banking and currency. *Dublin University Magazine* 15, 3–15, 218–33, 369–89, 609–20. Reprinted, 1971.

1870. Tenure of land in Ireland. *Systems of Land Tenure*. Cobden Club, London.

1872a. The limits of state interference with the distribution of wealth in applying taxation to the assistance of the public. *Journal of the Statistical and Social Inquiry Society of Ireland* 6, 105–14.

1872b. *Elementary Treatise on Series*. Dublin.

1971. *The Economic Writings of Mountifort Longfield*. Ed. R.D. Collison Black, New York: A.M. Kelley.

Bibliography

Black, R.D.C. 1945. Trinity College, Dublin, and the theory of value 1832–1863. *Economica*, NS 12, 140–48.

Black, R.D.C. 1984. The Irish dissenters and nineteenth-century political economy. In *Economists and the Irish Economy*, ed. A.E. Murphy. Dublin: Irish Academic Press, 120–37.

Moss, L.S. 1976. *Mountifort Longfield, Ireland's First Professor of Political Economy*. Ottowa, III.: Green Hill Publishers.

Murphy, A.E. 1984. Mountifort Longfield's appointment to the chair of political economy in trinity College, Dublin, 1832. In *Economists and the Irish Economy*, ed. A.E. Murphy. Dublin: Irish Academic Press, 13–24.

Seligman, E.R.A. 1903. On some neglected British economists. *Economic Journal* 13, 335–63, 511–35. Revised version published in E.R.A. Seligman, *Essays in Economics*, New York: Macmillan, 1925.

How to cite this article

Black, R. D. Collison. "Longfield, Mountifort (1802–1884)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000148> doi:10.1057/9780230226203.0993

The New Palgrave Dictionary of Economics Online

longitudinal data analysis

Cheng Hsiao
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The advantages and fundamental methodological issues of statistical inference using data sets that contain time series observations of a number of individuals are discussed.

Keywords

central limit theorems; discrete choice models; generalized method of moments; instrumental variables; laws of large numbers; least squares; linear models; logit models; maximum likelihood; panel data; tobit models; unit roots

Article

1 Why panel data?

‘Longitudinal data’ (or ‘panel data’) refers to data-sets that contain time series observations of a number of individuals. In other words, it provides multiple observations for each individual in the sample. Compared with cross-sectional data, in which observations for a number of individuals are available only for a given time, or time-series data, in which a single entity is observed over time, panel data have the obvious advantages of more degrees of freedom and less collinearity among explanatory variables, and so provide the possibility of obtaining more accurate parameter estimates. More importantly, by blending inter-individual differences with intra-individual dynamics, panel data allow the investigation of more complicated behavioural hypotheses than those that can be addressed using cross-sectional or time-series data.

For instance, suppose a cross-sectional sample yields an average labour-participation rate of 50 per cent for married women. Given that the standard assumption for the analysis of cross-sectional data is that, conditional on certain variables, each woman is a random draw from a homogeneous population, this would imply that each woman has a 50 per cent chance of being in the labour force at any given time. Hence, a married woman would be expected to spend half of her married life in the labour force and half

out of it. The job turnover would be frequent, and the expected average job duration would be just two years (Ben-Porath, 1973). However, the cross-sectional data could be drawn from a heterogeneous population in which 50 per cent of the sample was drawn from the population that always works and 50 per cent from the population that never works. In this situation, there is no turnover and a woman's current work status is a perfect predictor of her future work status. To discriminate between these two possibilities, we need information on individual labour-force histories in different sub-intervals of the life cycle, which can be provided only if information is available on the intertemporal dynamics of individual entities. On the other hand, although time series data provide information on dynamic adjustment, variables over time tend to move collinearly, hence making it difficult to identify micro-dynamic or macro-dynamic effects. Often, estimation of distributed lag models has to rely on strong prior restrictions like the Koyck or Almon lag, with very little empirical justification (for example, Griliches, 1967). With panel data, the inter-individual differences can often lessen the problem of multicollinearity and provide the possibility of estimating unrestricted time adjustment patterns (for example, Pakes and Griliches, 1984).

By utilizing information on both the intertemporal dynamics and the individuality of the entities, panel data may also allow an investigator to control the effects of missing or unobserved variables. For instance, MaCurdy's (1981) life-cycle labour supply of prime-age males with perfect foresight model assumes that the logarithm of hours worked is a linear function of the real wage rate and the logarithm of the worker's marginal utility of initial wealth, which is unobserved. Since the wage rate and the marginal utility of initial wealth are correlated, any instrument that is correlated with the wage rate will be correlated with the marginal utility of initial wealth. There is no way one can obtain a consistent estimate of the coefficient of the wage rate with cross-sectional data. But, if panel data are available and since marginal utility of initial wealth stays constant over time, one can take the difference of the labour supply model over time to get rid of the marginal utility of initial wealth as an explanatory variable. Regressing change in hour on change in wage rate and other socio-demographic variables can yield consistent estimates of the coefficient of the wage rate and other explanatory variables.

Panel data may also provide microfoundations for aggregate data analysis. Aggregate data analysis often invokes the ‘representative agent’ assumption. If micro units are heterogeneous, the time series properties of aggregate data may be very different from those of disaggregate data (for example, Granger, 1990; Lewbel, 1994) and policy evaluation based on aggregate data could also be grossly misleading (for example, Hsiao, Shen and Fujiki, 2005). By providing time series observations for a number of individuals, panel data are ideal for the investigation of the homogeneity issue.

Panel data involve observations of two or more dimensions. In normal circumstances, one would expect the computation and inference of panel data models to be more complicated than those of cross-section or time series data. However, in certain situations the availability of panel data actually simplifies inference. For instance, statistical inference for non-stationary panel data can be complicated (for example, Phillips, 1986). But, if observations are independently distributed across cross-sectional units, central limit theorems applied across cross-sectional units lead to asymptotically normally distributed statistics (for example, Levin, Lin and Chu, 2002; Im, Pesaran and Shin, 2003).

2 Issues of panel data analysis

Standard statistical methodology is based on the assumption that the outcomes, say *y*, conditional on certain variables, say $\underset{\sim}{x}$, are random outcomes from a probability distribution that is characterized by a fixed dimensional parameter vector, $\underset{\sim}{\theta}$, $f(y|\underset{\sim}{x};\underset{\sim}{\theta})$. For instance, the standard linear regression model assumes that $f(y|\underset{\sim}{x};\underset{\sim}{\theta})$ takes the form that $E(y|\underset{\sim}{x}) = \alpha + \underset{\sim}{\beta}'\underset{\sim}{x}$, and $\mathrm{Var}(y|\underset{\sim}{x}) = \sigma^2$, where $\underset{\sim}{\theta}' = (\alpha, \underset{\sim}{\beta}', \sigma^2)$. Panel data, by their nature, focus on individual outcomes. Factors affecting individual outcomes are numerous. It is rare to be able to assume a common conditional probability density function of *y* conditional on $\underset{\sim}{x}$ for all cross-sectional units, *i*, at all time, *t*. If the conditional density of *y* given $\underset{\sim}{x}$ varies across *i* and over *t*, the fundamental theorems for statistical inference, the laws of large numbers and central limit theorems, will be difficult to implement. Ignoring the heterogeneity across *i* and over *t* that are not captured by $\underset{\sim}{x}$ can lead to severely biased inference. For instance, suppose that the data is generated by

$$y_{it} = \alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + v_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array}$$
(2.1)

as depicted by Figure 1 in which the broken-time ellipses represent the point scatter of individual observation around the mean, represented by the broken straight lines. If an investigator ignores the presence of unobserved individual-specific effects, α_i, and mistakenly estimates a model of the form

$$y_{it} = \alpha + \underset{\sim}{\beta}'x_{it} + v_{it}^*$$
(2.2)

the following equation solid line in Figure 1 would depict the pooled least squares regression result which could completely contradict the individual relation between *y* and $\underset{\sim}{x}$.

Figure 1
Scatter diagram of (y_{it}, x_{it})

Figure 1 Scatter diagram of (yit, xit)

One way to restore homogeneity across *i* and/or over *t* is to add more conditional variables, say $\underset{\sim}{z}$,

$$f(y_{it} | \underset{\sim}{x}_{it}, \underset{\sim}{z}_{it}; \underset{\sim}{\theta}). \tag{2.3}$$

However, the dimension of $\underset{\sim}{z}$ can be large. A model is a simplification of reality, not an exact representation of reality. The inclusion of $\underset{\sim}{z}$ may confuse the fundamental relationship between *y* and *x*, in particular when there is a shortage of degrees of freedom or multicollinearity, and so on. Moreover, $\underset{\sim}{z}$ may not be observable. If an investigator is interested only in the relationship between *y* and $\underset{\sim}{x}$, one approach to characterize the heterogeneity not captured by $\underset{\sim}{x}$ is to assume that the parameter vector varies across *i* and over *t*, $\underset{\sim}{\theta}_{it}$, so that the conditional density of *y* given $\underset{\sim}{x}$ takes the form $f(y_{it} | \underset{\sim}{x}_{it}; \underset{\sim}{\theta}_{it})$. However, without a structure being imposed on $\underset{\sim}{\theta}_{it}$, such a model has only descriptive value; it is not possible to draw any inference on $\underset{\sim}{\theta}_{it}$ from observed data.

One primary focus of methodological panel data literature is to suggest possible structures for $\underset{\sim}{\theta}_{it}$. One way to impose some structure on $\underset{\sim}{\theta}_{it}$ is to decompose $\underset{\sim}{\theta}_{it}$ into $(\underset{\sim}{\beta}, \underset{\sim}{\gamma}_{it})$, where $\underset{\sim}{\beta}$ is the same across *i* and over *t*, referred to as *structural parameters*, and $\underset{\sim}{\gamma}_{it}$ as *incidental parameter*s because when observations in cross-sectional units and/or time series units increase, there are rising numbers of $\underset{\sim}{\gamma}_{it}$ to be estimated. The focus then will be on how to make valid inference on $\underset{\sim}{\beta}$ after controlling the impact of $\underset{\sim}{\gamma}_{it}$. Without imposing structure for $\underset{\sim}{\gamma}_{it}$, again it is not possible to make any inference on $\underset{\sim}{\beta}$ because the unknown $\underset{\sim}{\gamma}_{it}$ will exhaust all available sample information. On the assumption that the impacts of observable variables, $\underset{\sim}{x}$, are the same across *i* and over *t*, represented by the structure parameters, $\underset{\sim}{\beta}$, the incidental parameters $\underset{\sim}{\gamma}_{it}$ represent the heterogeneity across *i* and over *t* that are not captured by $\underset{\sim}{x}_{it}$. They can be considered as composed of the effects of omitted individual time-invariant, α_i, period individual-invariant, λ_t, and individual time-varying variables, δ_{it}. The individual time-invariant variables are variables that are the same for a given cross-sectional unit through time but that vary across cross-sectional units, such as individual-firm management, ability, gender, and socio-economic background. The period individual-invariant variables are variables that are the same for all cross-sectional units at a given time but that vary though time, such as prices, interest rates, and widespread optimism or pessimism. The individual time-varying variables are variables that vary across cross-sectional units at a given point in time and also exhibit variations through time, such as firm profits, sales and capital stock. The unobserved heterogeneity as represented by the individual-specific effects, α_i and time specific effects, λ_t, or individual time-varying effects, δ_{it} can be assumed to be either random variables (referred to as the *random effects* model) or fixed parameters (referred to as the *fixed effects* model).

3 Linear static models

A widely used panel data model assumes that the effects of observed explanatory variables, $\underset{\sim}{x}$, are identical across cross-sectional units, *i*, and over time, *t*, while the effects of omitted variables can be decomposed into the individual-specific effects, α_i, time-specific effects, λ_t, and individual time-varying effects, $\delta_{it} = u_{it}$, as follows:

$$y_{it} = \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i + \lambda_t + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix}$$

(3.1)

In a single equation framework, individual time effects, *u*, are assumed random and uncorrelated with $\underset{\sim}{x}$, while α_i and λ_t may or may not be correlated with $\underset{\sim}{x}$. When α_i and λ_t are treated as fixed constants, they are parameters to be estimated, so whether they are correlated with $\underset{\sim}{x}$ is not an issue. On the other hand, when α_i and λ_t are treated as random, they are typically assumed to be uncorrelated with $\underset{\sim}{x}_{it}$. For ease of exposition, we assume that there are no time-specific effects, that is, $\lambda_t = 0$ for all *t* and u_{it} are independently, identically distributed (i.i.d) across *i* and over *t*. Stack an individual's *T* time series observations of (y_{it}, x_{it}') into a vector and a matrix, (3.1) may alternatively be written as

$$\underset{\sim}{y}_i = X_i \underset{\sim}{\beta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i, \quad i = 1, \ldots, N,$$

(3.2)

where $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (\underset{\sim}{x}_{i1}, \ldots, \underset{\sim}{x}_{iT})'$, $\underset{\sim}{u}_i = (u_{i1}, \ldots, u_{iT})'$, and $\underset{\sim}{e}$ is a $T \times 1$ vector of 1's. Let *Q* be a $T \times T$ matrix satisfying the condition that $Q\underset{\sim}{e} = \underset{\sim}{0}$. Pre-multiplying (3.2) by *Q* yields

$$Q\underset{\sim}{y}_i = QX_i\underset{\sim}{\beta} + Q\underset{\sim}{u}_i, \quad i = 1, \ldots, N.$$

(3.3)

Equation (3.3) no longer involves α_i. The issue of whether α_i is correlated with $\underset{\sim}{x}_{it}$ or whether α_i should be treated as fixed or random is no longer relevant for (3.3). Moreover, since X_i is exogenous, $E(QX_i\underset{\sim}{u}_i'Q') = QE(X_i\underset{\sim}{u}_i')Q' = \underset{\sim}{0}$ and $EQ\underset{\sim}{u}_i\underset{\sim}{u}_i'Q' = \sigma_u^2 QQ'$. An efficient estimator of $\underset{\sim}{\beta}$ is the generalized least squares estimator (GLS),

$$\hat{\underset{\sim}{\beta}} = \left[\sum_{i=1}^{N} X_i'Q'(QQ')^{-}QX_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'Q'(QQ')^{-}Q\underset{\sim}{Y}_i\right], \tag{3.4}$$

where $(Q'Q)^{-}$ denotes the Moore–Penrose generalized inverse (for example, Rao, 1973). When $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$, Q is idempotent. The Moore–Penrose generalized inverse of $(Q'Q)^{-}$ is just $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$ itself. Pre-multiplying (3.3) by Q is equivalent to transforming (3.1) into a model

$$(y_{it} - \bar{y}_i) = \underset{\sim}{\beta}'(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i) + (u_{it} - \bar{u}_i), \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{3.5}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{\underset{\sim}{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\underset{\sim}{x}_{it}$ and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. The transformation is called *covariance transformation*. The least squares estimator (LS) (or a generalized least squares estimator, GLS) of (3.5),

$$\hat{\underset{\sim}{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)(y_{it} - \bar{y}_i)\right], \tag{3.6}$$

is called *covariance estimator* or *within* estimator because the estimation of $\underset{\sim}{\beta}$ only makes use of within (group) variation of y_{it} and $\underset{\sim}{x}_{it}$ only. The covariance estimator of $\underset{\sim}{\beta}$ turns out to be also the least squares

estimator of (3.1) when $\lambda_t = 0$. It is the best linear unbiased estimator of $\underset{\sim}{\beta}$ if α_i is treated as fixed and u_{it} is i.i.d.

If α_i is random, transforming (3.2) into (3.3) transforms *T* independent equations (or observations) into $(T-1)$ independent equations, hence the covariance estimator is not as efficient as the efficient generalized least squares estimator if $E\alpha_i \underset{\sim}{x}_{it}' = \underset{\sim}{0}'$. When α_i is independent of $\underset{\sim}{x}_{it}$ and is independently, identically distributed across *i* with mean $\underset{\sim}{0}$ and variance σ_α^2, the best linear unbiased estimator (BLUE) of $\underset{\sim}{\beta}$ is GLS,

$$\hat{\underset{\sim}{\beta}} = \left[\sum_{i=1}^{N} X_i' V^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' V^{-1} \underset{\sim}{y}_i\right], \tag{3.7}$$

where $V = \sigma_u^2 I_T + \sigma_\alpha^2 \underset{\sim}{e}\,\underset{\sim}{e}'$, $V^{-1} = \frac{1}{\sigma_u^2}\left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2} \underset{\sim}{e}\,\underset{\sim}{e}'\right]$, Let $\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$, the GLS is equivalent to first transforming the data by subtracting a fraction $(1-\psi^{1/2})$ of individual means $\bar{y}_i$ and $\bar{\underset{\sim}{x}}_i$ from their corresponding y_{it} and $\underset{\sim}{x}_{it}$, then regressing $[y_{it} - (1-\psi^{1/2})\bar{y}_i]$ on $[\underset{\sim}{x}_{it} - (1-\psi^{1/2})\bar{\underset{\sim}{x}}_i]$. (for detail, see Baltagi, 2001; Hsiao, 2003).

When α_i is treated as fixed, the covariance estimator is equivalent to applying LS to the transformed model (3.5). If a variable is time-invariant, like a gender dummy, $x_{kit} = x_{kis} = \bar{x}_{ki}$, the transformation eliminates the corresponding variable from the specification. Hence, the coefficients of time-invariant variables cannot be estimated. On the other hand, if α_i is random and uncorrelated with $\underset{\sim}{x}_i$, $\psi \neq 1$, the GLS can still estimate the coefficients of those time-invariant variables.

4 Dynamic models

When the regressors of a linear model contains lagged dependent variables, say, of the form (for example, Balestra and Nerlove, 1966)

$$\underset{\sim}{y}_i = \underset{\sim}{y}_{i,-1}\gamma + X_i\underset{\sim}{\beta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i = Z_i\underset{\sim}{\theta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i,\ i = 1, \ldots, N. \tag{4.1}$$

where $\underset{\sim}{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $Z_i = (\underset{\sim}{y}_{i,-1}, X_i)$ and $\underset{\sim}{\theta} = (\gamma, \underset{\sim}{\beta}')'$. For ease of notation, we assume that y_{i0} are observable. Technically, we can still eliminate the individual-specific effects by pre-multiplying (4.1) by the transformation matrix $Q(Q\underset{\sim}{e} = \underset{\sim}{0})$,

$$Q\underset{\sim}{y}_i = QZ_i\underset{\sim}{\theta} + Q\underset{\sim}{u}_i. \tag{4.2}$$

However, because of the presence of lagged dependent variables, $EQZ_i\underset{\sim}{u}_i'Q' \neq 0$ even with the assumption that u_{it} is independently, identically distributed across i and over t. For instance, the covariance transformation matrix $Q = I_T - \frac{1}{T}\underset{\sim}{e}\,\underset{\sim}{e}'$ transforms (4.1) into the form

$$(y_{it} - \bar{y}_i) = (y_{i,t-1} - \bar{y}_{i,-1})\gamma + (\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)'\underset{\sim}{\beta} + (u_{it} - \bar{u}_i), \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{4.3}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{y}_{i,-1} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}$ and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Although, $y_{i,t-1}$ and u_{it} are uncorrelated under the assumption of serial independence of u_{it}, the covariance between $\bar{y}_{i,-1}$ and u_{it} or $y_{i,t-1}$ and $\bar{u}_i$ is of order $(1/T)$ if $|\gamma|<1$. Therefore, the covariance estimator of $\hat{\underset{\sim}{\theta}}$ creates a bias of order $(1/T)$ when $N \to \infty$ (Anderson and Hsiao, 1981; 1982; Nickell, 1981). Since most panel data contain large N but small T, the magnitude of the bias can not be ignored (for example, with $T = 10$ and $\gamma = 0.5$, the asymptotic bias is –0.167).

When $EQZ_i\underset{\sim}{u}_i'Q' \neq \underset{\sim}{0}$, one way to obtain a consistent estimator for $\hat{\underset{\sim}{\theta}}$ is to find instruments W_i that satisfy

$$EW_i\underset{\sim}{u}_i'Q' = \underset{\sim}{0}, \tag{4.4}$$

and

$$\mathrm{rank}(W_i Q Z_i) = k, \tag{4.5}$$

where k denotes the dimension of $(\gamma, \underset{\sim}{\beta}')'$, then apply the generalized instrumental variable or generalized method of moments (GMM) estimator by minimizing the objective function

$$\left[\sum_{i=1}^{N} W_i(Q\underset{\sim}{y}_i - QZ_i\underset{\sim}{\theta})\right]'\left[\sum_{i=1}^{N}(W_i Q\underset{\sim}{u}_i\underset{\sim}{u}_i' Q' W_i')\right]^{-1}\left[\sum_{i=1}^{N} W_i(Q\underset{\sim}{y}_i - QZ_i'\underset{\sim}{\theta})\right], \tag{4.6}$$

with respect to $\underset{\sim}{\theta}$ (for example, Arellano, 2003; Ahn and Schmidt, 1995; Arellano and Bond, 1991; Arellano and Bover, 1995). For instance, one may let Q be a $(T-1)\times T$ matrix of the form

$$D = \begin{bmatrix} -1 & 1 & 0 & . & . \\ 0 & -1 & 1 & . & . \\ 0 & . & . & . & . \\ . & . & . & -1 & 1 \end{bmatrix}, \tag{4.7}$$

then the transformation (4.2) is equivalent to taking the first difference of (4.1) over time to eliminate α_i for $t = 2, \ldots, T$,

$$\Delta y_{it} = \Delta y_{i,t-1}\gamma + \Delta \underset{\sim}{x}_{it}'\underset{\sim}{\beta} + \Delta u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 2, \ldots, T, \end{matrix} \tag{4.8}$$

where $\Delta = (1 - L)$ and L denotes the lag operator, $Ly_t = y_{t-1}$. Since $\Delta u_{it} = (u_{it} - u_{i,t-1})$ is uncorrelated with $y_{i,t-j}$ for $j \geq 2$ and $\underset{\sim}{x}_{is}$, for all s, when u_{it} is independently distributed over time and $\underset{\sim}{x}_{it}$ is exogenous, one can let W_i be a $T(T-1)[K + \frac{1}{2}] \times (T-1)$ matrix of the form

$$W_i = \begin{bmatrix} \underset{\sim}{q}_{i2} & 0 & . & . \\ 0 & \underset{\sim}{q}_{i3} & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & \underset{\sim}{q}_{iT} \end{bmatrix},$$

(4.9)

where $\underset{\sim}{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2}, \underset{\sim}{x}_i'), \underset{\sim}{x}_i = (\underset{\sim}{x}_{i1}', \ldots, \underset{\sim}{x}_{iT}')'$, and $K = k - 1$. Under the assumption that $(\underset{\sim}{y}_i', \underset{\sim}{x}_i')$ are independently, identically distributed across i, the Arellano–Bover (1995) GMM estimator takes the form

$$\hat{\underset{\sim}{\theta}}_{AB,GMM} = \left\{\left[\sum_{i=1}^{N} Z_i' D' W_i'\right]\left[\sum_{i=1}^{N} W_i A W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i D Z_i\right]\right\}^{-1}\left\{\left[\sum_{i=1}^{N} Z_i' D' W_i'\right]\left[\sum_{i=1}^{N} W_i A W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i D \underset{\sim}{y}_i\right]\right\},$$

(4.10)

where A is a $(T-1) \times (T-1)$ matrix with 2 on the diagonal elements, –1 on the elements above and below the diagonal elements, and 0 elsewhere.

The GMM estimator has the advantage that it is consistent and asymptotically normally distributed whether α_i is treated as fixed or random because it eliminates α_i from the specification. However, the number of moment conditions increases at the order of T^2, which can create severe downward bias in finite sample (Zilak, 1997). An alternative is to use a (quasi-) likelihood approach which has the advantage of having a fixed number of orthogonality conditions independent of the sample size. It also has the advantage of making use of all the available samples, hence can yield a more efficient estimator than (4.10) (for example, Hsiao, Pesaran and Tahmiscioglu, 2002; Binder, Hsiao and Pesaran, 2005). Since there is no reason to assume the data-generating process of initial observations, y_{i0}, to be different

from the rest of y_{it}, the likelihood approach has to formulate the joint likelihood function of $(y_{i0}, y_{i1}, \ldots, y_{iT})$ (or the conditional likelihood function $(y_{i1}, \ldots, y_{iT} | y_{i0})$). However, y_{i0} depends on previous values of $\underset{\sim}{x}_{i,-j}$ and α_i, which are unavailable. Bhargava and Sargan (1983) suggest circumscribing this missing data problem by conditioning y_{i0} on $\underset{\sim}{x}_i$ and α_i if α_i is treated as random, while Hsiao, Pesaran and Tahmiscioglu (2002) propose conditioning $(y_{i1} - y_{i0})$ on the first difference of $\underset{\sim}{x}_i$ if α_i is treated as a fixed constant.

5 Random vs. fixed effects specification

The advantages of random effects (RE) specifications are as follows:

1. 1. The number of parameters stays constant when sample size increases.
2. 2. It allows the derivation of efficient estimators that make use of both within- and between-(group) variation.
3. 3. It allows the estimation of the impact of time-invariant variables.

The disadvantages of RE specification are that it typically assumes that the individual- and/or time-specific effects are randomly distributed with a common mean and are independent of $\underset{\sim}{x}_{it}$. If the effects are correlated with $\underset{\sim}{x}_{it}$ or if there is a fundamental difference among individual units, that is, conditional on $\underset{\sim}{x}_{it}$, y_{it} cannot be viewed as a random draw from a common distribution, the common RE model is mis-specified and the resulting estimator is biased.
The advantages of fixed effects (FE) specification are that it allows the individual-and/or time-specific effects to be correlated with explanatory variables $\underset{\sim}{x}_{it}$. Neither does it require an investigator to model their correlation patterns.
The disadvantages of the FE specification are as follows:

1. 1. The number of unknown parameters increases with the number of sample observations. In the case when T (or N for λ_t) is finite, it introduces the classical incidental parameter problem (for example, Neyman and Scott, 1948).
2. 2. The FE estimator does not allow the estimation of the coefficients that are time-invariant.

In other words, the advantages of RE specification are the disadvantages of FE specification, and the disadvantages of RE specification are the advantages of FE specification. To choose between the two specifications, Hausman (1978) notes that the FE estimator (or GMM), $\hat{\underset{\sim}{\theta}}_{FE}$, is consistent whether α_i is fixed or random. On the other hand, the commonly used RE estimator (or GLS), $\hat{\underset{\sim}{\theta}}_{RE}$, is consistent and efficient only when α_i is indeed uncorrelated with $\underset{\sim}{x}_{it}$. If α_i is correlated with $\underset{\sim}{x}_{it}$, the RE estimator is

inconsistent. Therefore, Hausman (1978) suggests using the statistic

$$\left(\hat{\theta}_{\sim FE} - \hat{\theta}_{\sim RE}\right)' \left[\operatorname{cov}\left(\hat{\theta}_{\sim FE}\right) - \operatorname{cov}\left(\hat{\theta}_{\sim RE}\right)\right]^{-} \left(\hat{\theta}_{\sim FE} - \hat{\theta}_{\sim RE}\right) \tag{5.1}$$

to test RE vs FE specification. The statistic (5.1) is asymptotically chi-square distributed with degrees of freedom equal to the rank of $[\operatorname{cov}(\hat{\theta}_{\sim FE}) - \operatorname{cov}(\hat{\theta}_{\sim RE})]$.

6 Nonlinear models

The introduction of individual-specific effects, α_i, and/or time-specific effects, λ_t, provides a simple way to capture the unobserved heterogeneity across *i* and over *t*. However, the likelihood functions are in terms of observables, $(\underset{\sim}{y}_i, \underset{\sim}{x}_i), i = 1, \ldots, N$. Therefore, we will have either to treat α_i as unknown parameters (fixed effects) and consider the conditional likelihood,

$$f(\underset{\sim i}{y} \mid \underset{\sim i}{x}, \underset{\sim}{\beta}, \alpha_i), i = 1, \ldots, N, \tag{6.1}$$

or to treat α_i as random and consider the marginal likelihood

$$f\left(\underset{\sim i}{y} \mid \underset{\sim i}{x}, \underset{\sim}{\beta}\right) = \int f(y_i \mid \underset{\sim i}{x}, \underset{\sim}{\beta}, \alpha_i) f(\alpha_i \mid \underset{\sim i}{x}) d\alpha_i, i = 1, \ldots, N, \tag{6.2}$$

where $f(\alpha_i \mid \underset{\sim i}{x})$ denotes the conditional density of α_i given $\underset{\sim i}{x}$.
When the unobserved individual specific effects, α_i, (and or time-specific effects, λ_t) affect the outcome, y_{it}, linearly, one can avoid the consideration of random versus fixed effects specification by eliminating them from the specification through some linear transformation such as the covariance transformation (3.3) or first difference transformation (4.8). However, if α_i affects y_{it} nonlinearly, it is

not easy to find a transformation that can eliminate α_i. For instance, consider the following binary choice model where the observed y_{it} takes the value of either 1 or 0 depending on the latent response function

$$y_{it}^* = \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i + u_{it}, \tag{6.3}$$

and

$$y_{it} = \begin{cases} 1, & \text{if } y_{it}^* > 0, \\ 0, & \text{if } y_{it}^* \leq 0, \end{cases} \tag{6.4}$$

where u_{it} is independently, identically distributed with density function $f(u_{it})$. Let

$$y_{it} = E(y_{it} | \underset{\sim}{x}_{it}, \alpha_i) + \varepsilon_{it}, \tag{6.5}$$

then

$$E(y_{it} | \underset{\sim}{x}_{it}, \alpha_i) = \int_{-(\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i)}^{\infty} f(u)\,du = [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]. \tag{6.6}$$

Since α_i affects $E(y_{it} | \underset{\sim}{x}_{it}, \alpha_i)$ nonlinearly, α_i remains after taking successive difference of y_{it},

$$y_{it} - y_{i,t-1} = [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)] - [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{i,t-1} - \alpha_i)] + (\varepsilon_{it} - \varepsilon_{i,t-1}).$$

(6.7)

The likelihood function conditional on $\underset{\sim}{x}_i$ and α_i takes the form,

$$\prod_{i=1}^{N}\prod_{t=1}^{T}[F(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i)]^{1-y_{it}}[1-F(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i)]^{y_{it}}. \tag{6.8}$$

If T is large, a consistent estimator of $\underset{\sim}{\beta}$ and α_i can be obtained by maximizing (6.8). If T is finite, there is only limited information about α_i no matter how large N is. The presence of incidental parameters, α_i, violates the regularity conditions for the consistency of the maximum likelihood estimator of $\underset{\sim}{\beta}$.
If $f(\alpha_i|\underset{\sim}{x}_i)$ is known, and is characterized by a fixed dimensional parameter vector, a consistent estimator of $\underset{\sim}{\beta}$ can be obtained by maximizing the marginal likelihood function,

$$\prod_{i=1}^{N}\int\prod_{t=1}^{T}[F(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i)]^{1-y_{it}}[1-F(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i)]^{y_{it}}f(\alpha_i|\underset{\sim}{x}_i)d\alpha_i. \tag{6.9}$$

However, maximizing (6.9) involves T-dimensional integration. Butler and Moffitt (1982), Chamberlain (1984), Heckman (1981), and others have suggested methods to simplify the computation.
The advantage of RE specification is that there is no incidental parameter problem. The problem is that $f(\alpha_i|\underset{\sim}{x}_i)$ is in general unknown. If a wrong $f(\alpha_i|\underset{\sim}{x}_i)$ is postulated, maximizing the wrong likelihood function will not yield a consistent estimator of $\underset{\sim}{\beta}$. Moreover, the derivation of marginal likelihood through multiple integration may be computationally infeasible. The advantage of FE specification is that there is no need to specify $f(\alpha_i|\underset{\sim}{x}_i)$. The likelihood function will be the product of individual likelihood (for example, (6.8)) if the errors are assumed i.i.d. The disadvantage is that it introduces incidental parameters.
A general approach to estimating a model involving incidental parameters is to find transformations to transform the original model into a model that does not involve incidental parameters. Unfortunately, there is no general rule available for nonlinear models. One has to explore the specific structure of a nonlinear model to find such a transformation. For instance, if $f(u)$ in (6.3) is logistic, then

$$\text{Prob}(y_{it} = 1 | \underset{\sim}{x}_{it}, \alpha_i) = \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i}}{1 + e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i}}. \tag{6.10}$$

Since, in a logit model, the denominators of $\text{Prob}(y_{it} = 1 | \underset{\sim}{x}_{it}, \alpha_i)$ and $\text{Prob}(y_{it} = 0 | \underset{\sim}{x}_{it}, \alpha_i)$ are identical and the numerator of any sequence $\{y_{i1}, \ldots, y_{iT}\}$ with $\sum_{t=1}^{T} y_{it} = s$ always equal to $\exp(\alpha_i s) \cdot \exp\left\{\sum_{t=1}^{T} (\underset{\sim}{\beta}' \underset{\sim}{x}_{it}) y_{it}\right\}$, the conditional likelihood function conditional on $\sum_{t=1}^{T} y_{it} = s$ will not involve the incidental parameters α_i. For instance, consider the simple case that $T = 2$, then

$$\text{Prob}(y_{i1} = 1, y_{i2} = 0 | y_{i1} + y_{i2} = 1) = \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i1}}}{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i1}} + e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i2}}} = \frac{1}{1 + e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}} \tag{6.11}$$

and

$$\text{Prob}(y_{i1} = 0, y_{i2} = 1 | y_{i1} + y_{i2} = 1) = \frac{e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}}{1 + e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}}, \tag{6.12}$$

(Chamberlain, 1980; Hsiao, 2003).

This approach works because of the logit structure. In the case when $f(u)$ is unknown, Manski (1987) exploits the latent linear structure of (6.3) by noting that, for given i,

$$\begin{aligned}
\underset{\sim}{\beta}'\underset{\sim}{x}_{it} > \underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1} &\Leftrightarrow E(y_{it}|\underset{\sim}{x}_{it}, \alpha_i) > E(y_{i,t-1}|\underset{\sim}{x}_{i,t-1}, \alpha_i),\\
\underset{\sim}{\beta}'\underset{\sim}{x}_{it} = \underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1} &\Leftrightarrow E(y_{it}|\underset{\sim}{x}_{it}, \alpha_i) = E(y_{i,t-1}|\underset{\sim}{x}_{i,t-1}, \alpha_i),\\
\underset{\sim}{\beta}'\underset{\sim}{x}_{it} < \underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1} &\Leftrightarrow E(y_{it}|\underset{\sim}{x}_{it}, \alpha_i) < E(y_{i,t-1}|\underset{\sim}{x}_{i,t-1}, \alpha_i),
\end{aligned}$$

(6.13)

and suggests maximizing the objective function

$$H_N(b) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T} sgn(\underset{\sim}{b}'\Delta\underset{\sim}{x}_{it})\Delta y_{it},$$

(6.14)

where $sgn(w) = 1$ if $w > 0$, $= 0$ if $w = 0$, and -1if $w<0$. The advantage of the Manski (1987) maximum score estimator is that it is consistent without the knowledge of $f(u)$. The disadvantage is that (6.13) holds for any $c\underset{\sim}{\beta}$ where $c > 0$. Only the relative magnitude of the coefficients can be estimated with some normalization rule, say $\|\underset{\sim}{\beta}\| = 1$. Moreover, the speed of convergence is considerably slower ($N^{1/3}$) and the limiting distribution is quite complicated. Horowitz (1992) and Lee (1999) have proposed modified estimators that improve the speed of convergence and are asymptotically normally distributed.

Other examples of exploiting specific structure of nonlinear models to eliminate the effects of incidental parameters α_i include dynamic discrete choice models (Chamberlain, 1993; Honoré and Kyriazidou, 2000; Hsiao et al., 2005), symmetrically trimmed least squares estimator for truncated and censored data (tobit models) (Honoré, 1992), sample selection models (or type II tobit models) (Kyriazidou, 1997), and so on. However, often they impose very severe restrictions on the data such that not much of it can be utilized to obtain parameter estimates. Moreover, there are models that do not appear to yield consistent estimator when T is finite.

An alternative to consistent estimators is to consider bias-reduced estimators. The advantage of such an approach is that the bias-reduced estimators may still allow the use of all the sample information so that, from a mean square error point of view, the bias-reduced estimator may still dominate consistent estimators because the latter often have to throw away a lot of the sample, and thus tend to have large variances.

Following the ideas of Cox and Reid (1987), Arellano (2001) and Carro (2006) propose to derive the modified MLE by maximizing the modified log-likelihood function

$$L^*(\underset{\sim}{\beta}) = \sum_{i=1}^{N}\left[\mathcal{L}_i^*(\underset{\sim}{\beta}, \widehat{\alpha}_i(\underset{\sim}{\beta})) - \frac{1}{2}\log \mathcal{L}^*_{i,\alpha_i\alpha_i}(\underset{\sim}{\beta}, \widehat{\alpha}_i(\underset{\sim}{\beta}))\right] \tag{6.15}$$

where $\mathcal{L}_i^*(\underset{\sim}{\beta}, \widehat{\alpha}_i(\underset{\sim}{\beta}))$ denotes the concentrated log-likelihood function of $\underset{\sim}{Y}_i$ after substituting the MLE of α_i in terms of $\underset{\sim}{\beta}, \widehat{\alpha}_i(\underset{\sim}{\beta})$ (that is, the solution of $\frac{\partial \log L}{\partial \alpha_i} = 0$ in terms of $\underset{\sim}{\beta}, i = 1, \ldots, N$) into the log-likelihood function and $\mathcal{L}^*_{i,\alpha_i,\alpha_i}(\underset{\sim}{\beta}, \widehat{\alpha}_i(\underset{\sim}{\beta}))$ denotes the second derivative of $\mathcal{L}_i^*$ with respect to α_i. The bias correction term is derived by noting that to the order of (1/*T*) the first derivative of $\mathcal{L}_i^*$ with respect to $\underset{\sim}{\beta}$ converges to $\frac{1}{2}\frac{E[\mathcal{L}^*_{i,\beta\alpha_i\alpha_i}(\underset{\sim}{\beta}, \alpha_i)]}{E[\mathcal{L}^*_{i,\alpha_i\alpha_i}(\underset{\sim}{\beta}, \alpha_i)]}$. By subtracting the order (1/T) bias from the likelihood function, the modified MLE is biased only to the order of ($1/T^2$), without increasing the asymptotic variance.
Monte Carlo experiments conducted by Carro (2006) have shown that, when $T = 8$, the bias of modified MLE for dynamic probit and logit models is negligible. Another advantage of the Arellano–Carro approach is its generality. For instance, a dynamic logit model with time dummy explanatory variable does not meet the Honoré and Kyriazidou (2000) conditions for generating consistent estimators, but will not affect the asymptotic properties of the modified MLE.

7 Modelling cross-sectional dependence

Most panel studies assume that, apart from the possible presence of individual invariant but period-varying time-specific effects, λ_t, the effects of omitted variables are independently distributed across cross-sectional units. However, often economic theory predicts that agents take actions that lead to interdependence among themselves. For example, the prediction that risk-averse agents will make insurance contracts allowing them to smooth idiosyncratic shocks implies dependence in consumption across individuals. Ignoring cross-sectional dependence can lead to inconsistent estimators, in particular when *T* is finite (for example, Hsiao and Tahmiscioglu, 2005). Unfortunately, contrary to the time series data in which the time label gives a natural ordering and structure, general forms of dependence for cross-sectional dimension are difficult to formulate. Therefore, econometricians have relied on strong parametric assumptions to model cross-sectional dependence. Two approaches have been proposed to model cross-sectional dependence: economic distance (or a spatial approach) and a factor approach.
In regional science, correlation across cross-section units is assumed to follow a certain spatial ordering, that is, dependence among cross-sectional units is related to location and distance, in a geographic or more general economic or social network space (for example, Anselin, 1988; Anselin and Griffith, 1988; Anselin, Le Gallo and Jayet, 2006). A known spatial weights matrix, $W = (w_{ij})$, an $N \times N$ positive matrix in which the rows and columns correspond to the cross-sectional units, is specified to express the prior

strength of the interaction between individual (location) *i* (in the row of the matrix) and individual (location) *j* (column), w_{ij}. By convention, the diagonal elements, $w_{ii} = 0$. The weights are often standardized so that the sum of each row, $\sum_{j=1}^{N} w_{ij} = 1$.

The spatial weight matrix, *W*, is often included into a model specification to the dependent variable, to the explanatory variables, or to the error term. For instance, a *spatial lag* model for the $NT \times 1$ variable $\underset{\sim}{y} = (\underset{\sim}{y}_1', \ldots, \underset{\sim}{y}_N')'$, $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iT})'$, may take the form

$$y = \rho(W \otimes I_T)\underset{\sim}{y} + X\underset{\sim}{\beta} + \underset{\sim}{u} \tag{7.1}$$

where *X* and $\underset{\sim}{u}$ denote the $NT \times 1$ explanatory variables and $NT \times 1$ vector of error terms, respectively, and ⊗ denotes the Kronecker product. A *spatial error* model may take the form

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{v} \tag{7.2}$$

where $\underset{\sim}{v}$ may be specified as in a *spatial autoregressive* form,

$$\underset{\sim}{v} = \theta(W \otimes I_T)\underset{\sim}{v} + \underset{\sim}{u}, \tag{7.3}$$

or a spatial moving average form,

$$\underset{\sim}{v} = \gamma(W \otimes I_T)\underset{\sim}{u} + \underset{\sim}{u}. \tag{7.4}$$

The spatial model can be estimated by the instrumental variables (GMM estimator) or the maximum

likelihood method. However, the approach of defining cross-sectional dependence in terms of ‘economic distance’ measure requires that the econometricians have information regarding this ‘economic distance’. Another approach to model cross-sectional dependence is to assume that the error of a model, say model (7.3), follows a linear factor model,

$$v_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + u_{it}, \tag{7.5}$$

where $\underset{\sim}{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of random factors, $\underset{\sim}{b}_i' = (b_{i1}, \ldots, b_{ir})$, is $r \times 1$ non-random factor loading coefficients, u_{it}, represents the effects of idiosyncratic shocks which is independent of $\underset{\sim}{f}_t$ and is independently distributed across *i*. (for example, Bai and Ng, 2002; Moon and Perron, 2004; Pesaran, 2006). The conventional time-specific effects model is a special case of (7.5) when $r = 1$ and $b_i = b_\ell$ for all *i* and ℓ.

The factor approach requires considerably less prior information than the economic distance approach. Moreover, the number of time-varying factors, *r*, and factor load matrix $B = (b_{ij})$ can be empirically identified if both *N* and *T* are large. However, when *T* is large, one can estimate the covariance between *i* and *j*, σ_{ij}, by $\frac{1}{T}\sum_{t=1}^{T} \hat{v}_{it}\hat{v}_{jt}$ directly, then apply the generalized least squares method, where $\hat{v}_{it}$ is some preliminary estimate of v_{it}.

8 Large-N and large-T panels

Our discussion has been mostly focusing on panels with large *N* and finite *T*. There are panel data sets, like the Penn-World tables, covering different individuals, industries and countries over long periods. In general, if an estimator is consistent in the fixed-*T*, large-*N* case, it will remain consistent if both *N* and *T* tend to infinity. Moreover, even in the case that an estimator is inconsistent for fixed *T* and large *N* (say, the MLE of dynamic model (4.1) or fixed effects probit or logit models (6.6)), it can become consistent if *T* also tends to infinity. The probability limit of an estimator, in general, is identical irrespective of how *N* and *T* tend to infinity. However, the properly scaled limiting distribution may depend on how the two indexes, *N* and *T*, tend to infinity.

There are several approaches for deriving the limits of large-*N*, large-*T* panels:

1. 1. *Sequential limits*. First, fix one index, say *N*, and allow the other, say *T*, to go to infinity, giving an intermediate limit, then let *N* go to infinity.
2. 2. *Diagonal-path limits*. Let the two indexes, *N* and *T*, pass to infinity along a specific diagonal path, say $T = T(N)$ as $N \to \infty$.

3. 3. *Joint limits*. Let N and T pass to infinity simultaneously without placing specific diagonal path restrictions on the divergence.

In many applications, sequential limits are easy to derive. However, sometimes sequential limits can give misleading asymptotic results. A joint limit will give a more robust result than either a sequential limit or a diagonal-path limit, but will also be substantially more difficult to derive and will apply only under stronger conditions, such as the existence of higher moments. Phillips and Moon (1999) have given a set of sufficient conditions that ensures that sequential limits are equivalent to joint limits.

When T is large, there is a need to consider serial correlations more generally, including both short-memory and persistent components. For instance, if unit roots are present in y and x (that is, both are integrated of order 1) but are not cointegrated, Phillips and Moon (1999) show that, if N is fixed but $T \to \infty$, the least squares regression of y on x is a non-degenerate random variable that is a functional of Brownian motion that does not converge to the long-run average relation between y and x, but it does if N also tends to infinity. In other words, the issue of spurious regression will not arise in a panel with large N (for example, Kao, 1999).

Both theoretical and applied researchers have paid a great deal of attention to the unit root and spurious regression properties of variables. When N is finite and T is large, standard time-series techniques can be used to derive the statistical properties of panel data estimators. When N is large and cross-sectional units are independently distributed across i, central limit theorems can be invoked along the cross-sectional dimension. Asymptotically normal estimators and test statistics (with suitably adjustment for finite T bias) for unit roots and cointegration have been proposed (for example, Baltagi and Kao, 2000; Im, Pesaran and Shin, 2003; Levin, Lin and Chu, 2002). They, in general, gain statistical power over their standard time series counterpart (for example, Choi, 2001).

When both N and T are large and cross-sectional units are not independent, a factor analytic framework of the form (7.5) has been proposed to model cross-sectional dependency and variants of unit root tests are proposed (for example, Moon and Perron, 2004). However, the implementation of those panel unit root tests is quite complicated. When $N \to \infty$, $\frac{1}{N}\sum_{i=1}^{N} u_{it} \to 0$, (7.5) implies that $\bar{v}_t = \bar{\underset{\sim}{b}}' \underset{\sim}{f}_t$, where $\bar{\underset{\sim}{b}}'$ is the cross-sectional average of $\underset{\sim}{b}_i' = (b_{i1}, \ldots, b_{ir})$. Approximating $\underset{\sim}{b}_i' \underset{\sim}{f}_i$ by its cross-sectional mean function, Pesaran (2005; 2006) suggests a simple approach to filter out the cross-sectional dependency by augmenting the cross-sectional means, $\bar{y}_t$ and $\bar{\underset{\sim}{x}}_t$ to the regression model (7.2),

$$y_{it} = \underset{\sim}{x}_{it}' \underset{\sim}{\beta} + \alpha_i + \bar{y}_t c_i + \bar{\underset{\sim}{x}}_t' \underset{\sim}{d}_1 + e_{it}, \tag{8.1}$$

or $\bar{y}_t, \Delta\bar{y}_{t-j}$ to the Dickey–Fuller (1979) type regression model,

$$\Delta y_{it} = \alpha_i + \delta_i t + \gamma_i y_{i,t-1} + \sum_{\ell=1}^{p_i} \varphi_{i\ell} \Delta y_{i,t-\ell} + c_i \bar{y}_{t-1} + \sum_{\ell=1}^{p_i} d_{i\ell} \Delta \bar{y}_{t-\ell} + e_{it}, \tag{8.2}$$

for testing of unit root, where $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$, $\bar{\underset{\sim}{x}}_t = \frac{1}{N}\sum_{i=1}^{N} \underset{\sim}{x}_{it}$, $\Delta\bar{y}_{t-j} = \frac{1}{N}\sum_{i=1}^{N}\Delta y_{i,t-j}$ and $\Delta = (1 - L)$, L denotes the lag operator. The resulting pooled estimator will again be asymptotically normally distributed.

When cross-sectional dependency is of unknown form, Chang (2002) suggests using nonlinear transformations of the lagged level variable, $y_{i,t-1}$, $F(y_{i,t-1})$, as instrumental variables (IV) for the usual augmented Dickey–Fuller (1979) type regression. The test static for the unit root hypothesis is simply defined as a standardized sum of individual IV *t*-ratios. As long as $F(\cdot)$ is regularly integrable, say $F(y_{i,t-1}) = y_{i,t-1}e^{-c_i|y_{i,t-1}|}$, where c_i is a positive constant, the product of the nonlinear instruments $F(y_{i,t-1})$ and $F(y_{i,t-1})$ from different cross-sectional units *i* and *j* are asymptotically uncorrelated, even the variables $y_{i,t-1}$ and $y_{i,t-1}$ generating the instruments are correlated. Hence, the usual central limit theorems can be invoked and the standardized sum of individual IV *t*-ratios is asymptotically normally distributed.

For further review of the literature on unit roots and cointegration in panels, see Breitung and Pesaran (2006) and Choi (2006).

9 Concluding remarks

In this paper we have tried to provide a summary of the advantages of using panel data and the fundamental issues of panel data analysis. Assuming that the heterogeneity across cross-sectional units and over time that is not captured by the observed variables can be captured by period-invariant individual specific and/or individual-invariant time-specific effects, we surveyed the fundamental methods for the analysis of linear static and dynamic models. We have also discussed difficulties in analysing nonlinear models and modelling cross-sectional dependence. There are many important issues, such as the modelling of joint dependence or simultaneous equations models, time-varying parameter models (for example, Hsiao, 1996; 2003; Hsiao and Pesaran, 2006), unbalanced panel, measurement errors (Griliches and Hausman, 1986; Wansbeek and Koning, 1989), and so on, that were not discussed, but can be found in Arellano (2003), Baltagi (2001) or Hsiao (2003).

Although panel data offer many advantages, they are no panacea. The power of panel data to isolate the effects of specific actions, treatments or more general policies depends critically on the compatibility of the assumptions of statistical tools with the data-generating process. In choosing the proper method for exploiting the richness and unique properties of the panel, it might be helpful to keep the following questions in mind. First, in investigating economic issues what advantages do panel data offer us over data-sets consisting of a single cross section or time series? Second, what are the limitations of panel data

and the econometric methods that have been proposed for analysing such data? Third, when using panel data, how can we increase the efficiency of parameter estimates? Fourth, are the assumptions underlying the statistical inference procedures and the data-generating process compatible?

I would like to thank Steven Durlauf for helpful comments.

Bibliography

Ahn, S.C. and Schmidt, P. 1995. Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68, 5–27.

Anderson, T.W. and Hsiao, C. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76, 598–606.

Anderson, T.W. and Hsiao, C. 1982. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* 18, 47–82.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Boston: Kluwer.

Anselin, L. and Griffith, D.A. 1988. Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association* 65, 11–34.

Anselin, L., Le Gallo, J. and Jayet, H. 2006. Spatial panel econometrics. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. 3rd edn, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Arellano, M. 2001. Discrete choice with panel data. Working Paper No. 0101. Madrid: CEMFI.

Arellano, M. 2003. *Panel Data Econometrics*. Oxford: Oxford University Press.

Arellano, M. and Bond, S.R. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58, 277–97.

Arellano, M. and Bover, O. 1995. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68, 29–51.

Bai, J. and Ng, S. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70, 91–121.

Balestra, P. and Nerlove, M. 1966. Pooling cross-section and time series data in the estimation of a dynamic model: the demand for natural gas. *Econometrica* 34, 585–612.

Baltagi, B.H. 2001. *Econometric Analysis of Panel Data*. 2nd edn. New York: Wiley.

Baltagi, B.H. and Kao, C. 2000. Nonstationary panels, cointegration in panels and dynamic panel: a survey. In *Nonstationary Panels Panel Cointegration, and Dynamic Panels*, ed. B. Baltagi. Amsterdam: JAI Press.

Ben-Porath, Y. 1973. Labor force participation rates and the supply of labor. *Journal of Political Economy* 81, 697–704.

Bhargava, A. and Sargan, J.D. 1983. Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51, 1635–59.

Binder, M., Hsiao, C. and Pesaran, M.H. 2005. Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory* 21, 795–837.

Breitung, J. and Pesaran, M.H. 2006. Unit roots and cointegration in panels. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. 3rd edn, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Butler, J.S. and Moffitt, R. 1982. A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica* 50, 761–4.

Carro, J.M. 2006. Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics* (forthcoming).

Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Review of Economic Studies* 47, 225–38.

Chamberlain, G. 1984. Panel data. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M. Intriligato. Amsterdam: North-Holland.

Chamberlain, G. 1993. Feedback in panel data models. Mimeo, Department of Economics, Harvard University.

Chang, Y. 2002. Nonlinear IV unit root tests in panels with cross-sectional dependency. *Journal of Econometrics* 110, 261–92.

Choi, I. 2001. Unit root tests for panel data. *Journal of International Money and Finance* 20, 249–72.

Choi, I. 2006. Nonstationary panels. In *Palgrave Handbooks of Econometrics*, vol. 1, ed. T.C. Mills and K.D. Patterson. Basingstoke: Palgrave Macmillan.

Cox, D.R. and Reid, N. 1987. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, B* 49, 1–39.

Dickey, D.A. and Fuller, W.A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.

Granger, C.W.J. 1990. Aggregation of time-series variables: a survey. In *Disaggregation in Econometric Modeling*, ed. T. Barker and M.H. Pesaran. London: Routledge.

Griliches, Z. 1967. Distributed lags: a survey. *Econometrica* 35, 16–49.

Griliches, Z. and Hausman, J.A. 1986. Errors-in-variables in panel data. *Journal of Econometrics* 31, 93–118.

Hausman, J.A. 1978. Specification tests in econometrics. *Econometrica* 46, 1251–71.

Heckman, J.J. 1981. Statistical models for discrete panel data. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

Honoré, B. 1992. Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60, 533–67.

Honoré, B. and Kyriazidou, E. 2000. Panel data discrete choice models with lagged dependent variables. *Econometrica* 68, 839–74.

Horowitz, J.L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–31.

Hsiao, C. 1996. Random coefficient models. In *The Econometrics of Panel Data*. 2nd edn, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Hsiao, C. 2003. *Analysis of Panel Data*, 2nd edn. Cambridge: Cambridge University Press.

Hsiao, C. and Pesaran, M.H. 2006. Random coefficients models. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. 3rd edn, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Hsiao, C., Pesaran, M.H. and Tahmiscioglu, A.K. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109, 107–50.

Hsiao, C., Shen, Y. and Fujiki, H. 2005. Aggregate vs disaggregate data analysis – a paradox in the estimation of money demand function of Japan under the low interest rate policy. *Journal of Applied*

Econometrics 20, 579–601.

Hsiao, C., Shen, Y., Wang, B. and Weeks, G. 2005. Evaluating the effectiveness of Washington State repeated job search services on the employment rate of prime-age female welfare recipients. Mimeo, University of Southern California.

Hsiao, C. and Tahmiscioglu, A.K. 2005. Estimation of dynamic panel data models with both individual and time specific effects. Mimeo.

Im, K., Pesaran, M.H. and Shin, Y. 2003. Testing for unit roots in heterogeneous panels. *Journal of Econometrics* 115, 53–74.

Kao, C. 1999. Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics* 90, 1–44.

Kyriazidou, E. 1997. Estimation of a panel data sample selection model. *Econometrica* 65, 1335–64.

Lee, M.J. 1999. A root-N-consistent semiparametric estimator for related effects binary response panel data. *Econometrica* 67, 427–33.

Levin, A., Lin, C. and Chu, J. 2002. Unit root tests in panel data: asymptotic and finite sample properties. *Journal of Econometrics* 108, 21–24.

Lewbel, A. 1994. Aggregation and simple dynamics. *American Economic Review* 84, 905–18.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life cycle setting. *Journal of Political Economy* 89, 1059–85.

Manski, C.F. 1987. Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* 55, 357–62.

Moon, H.R. and Perron, B. 2004. Testing for a unit root in panels with dynamic factors. *Journal of Econometrics* 122, 81–126.

Neyman, J. and Scott, E.L. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.

Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49, 1399–416.

Pakes, A. and Griliches, Z. 1984. Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs. *Review of Economic Studies* 51, 243–62.

Pesaran, M.H. 2005. A simple panel unit root test in the presence of cross-section dependence. DAE Working Paper No. 0346, Cambridge University.

Pesaran, M.H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74, 967–1012.

Phillips, P.C. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–40.

Phillips, P.C. and Moon, H.R 1999. Linear regression limit theory for nonstationary panel data. *Econometrica* 67, 1057–111.

Rao, C.R. 1973. *Linear Statistical Inference and Its Applications*, 2nd edn. NewYork: Wiley.

Wansbeek, T.J. and Koning, R.H. 1989. Measurement error and panel data. *Statistica Neerlandica* 45, 85–92.

Zilak, J.P. 1997. Efficient estimation with panel data when instruments are predetermined: an empirical comparison of moment-condition estimators. *Journal of Business and Economic Statistics* 15, 419–31.

How to cite this article

Hsiao, Cheng. "longitudinal data analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000221> doi:10.1057/9780230226203.0994

Economic Growth Centre
Working Paper Series

Longitudinal Data Analysis

by

Cheng HSIAO

Economic Growth Centre
Division of Economics
School of Humanities and Social Sciences
Nanyang Technological University
Nanyang Avenue
SINGAPORE 639798

Website: http://www.hss.ntu.edu.sg/egc/

Working Paper No: 2005/10

Copies of the working papers are available from the World Wide Web at:

Website: http://www.hss.ntu.edu.sg/egc/

The author bears sole responsibility for this paper. Views expressed in this paper are those of the author(s) and not necessarily those of the Economic Growth Centre.

Longitudinal Data Analysis

Cheng Hsiao
Nanyang Technological University, Singapore

and

University of Southern California
Department of Economics
KAP300
3620 S. Vermont Ave.
Los Angeles, CA 90089-0253
email: chsiao@usc.edu

July 28, 2005

1. Why Panel Data?

Panel data, or longitudinal data, refers to data set that contains observations of a number of individuals over time. In other words, it provides multiple observations for each individual in the sample. Compared to the cross-sectional data in which observations for a number of individuals are available only for a given time, or the time series data, in which a single entity is observed over time, panel data has the obvious advantages of having more degrees of freedom and less collinearity among explanatory variables, hence provides the possibility of obtaining more accurate parameter estimates. More importantly, panel data by blending inter-individual differences with intra-individual dynamics, allows the investigation of more complicated behavioral hypotheses than those that can be addressed using cross-sectional or time series data. For instance, standard assumption for the analysis of cross-sectional data is that conditional on certain variables, each woman is a random sample from a homogeneous population. Therefore, if a cross-sectional sample yields an average labor-participation rate of 50 percent for married women, it would imply that each woman has a 50 percent chance of being in the labor force at any given time, hence a married woman would be expected to spend half of her married life in the labor force and half out of the labor force. The job turnover would be frequent, and the average job duration would be expected just two years (Ben-Porath (1973)). However, the cross-sectional data could be drawn from a heterogeneous population in which 50 percent of the sample coming from the population that always work and 50 percent from the population that never work. In this situation, there is no turnover and current work status about a woman is a perfect predictor of her future work status. To discriminate between these two possibilities, we need information on individual labor-force histories in different subintervals of the life cycle, which can only be provided if information on intertemporal dynamics of individual entities are available. On the other hand, although time series data provide information on dynamic adjustment, variables over time tend to move collinearly, hence makes it difficult to identify microdynamic or macrodynamic effects. Often estimation of

distributed lag models has to rely on strong prior restrictions like Koyck or Almon lag with very little empirical justification. With panel data, the interindividual differences often can reduce or lessen the problem of multicollinearity and provide the possibility of estimating unrestricted time adjustment patterns (e.g. Pakes and Griliches (1984)).

By utilizing information on both the intertemporal dynamics and the individuality of the entities, panel data may also allow an investigator to control the effects of missing or unobserved variables. For instance, MaCurdy's (1981) life cycle labor supply of prime-age males under certainty model assumes that the logarithm of hours worked is a linear function of the real wage rate and the logarithm of the worker's marginal utility of initial wealth, which is unobserved. Since wage rate and marginal utility of initial wealth are correlated, any instrument that is correlated with the wage rate will be correlated with the marginal utility of initial wealth. There is no way one can obtain consistent estimate of the coefficient of the wage rate with cross-sectional data. But if panel data are available, one can transform the labor supply model by taking first difference to get rid of the marginal utility of initial wealth as an explanatory variable. The resulting regression can yeild consistent estimates of the coefficient of wage rate and other explanatory varibles.

Panel data may also provide micro foundations for aggregate data analysis. Aggregate data analysis often invokes the "representative agent" assumption. If micro units are heterogeneous, the time series properties of aggregate data may be very different from those of disaggregate data (e.g. Granger (1990), Lewbel (1992, 94), Pesaran (1999)) and policy evaluation based on aggregate data could also be grossly misleading (e.g. Hsiao, Shen and Fujiki (2004)). Panel data by providing time series observations for a number of individuals is ideal for the investigation of homogeneity issue.

Panel data involve observations of two or more dimensions. In normal circumstance, one would expect that the computation and inference of panel data models be more complicated than cross-section or time series data. However, in certain situations, the availability of panel data actually simplify inference. For instance, statistical inference for nonstation-

ary panel data can be complicated (e.g. Phillips (1986)). But, if observations are independently distributed across cross-sectional units, central limit theorems applied across cross-sectional units lead to asymptotically normally distributed statistics (e.g. Levin, Lin and Chu (2002), Pesaran, Shin and Smith (2002)).

2. Issues of Panel Data Analysis

Standard statistical methodology is based on the assumption that the outcomes, say y, conditional on certain variables, say $\underset{\sim}{x}$, are random outcomes from a probability distribution that is characterized by a fixed dimensional parameter vector, $\underset{\sim}{\theta}$, $f(y \mid \underset{\sim}{x}; \underset{\sim}{\theta})$. For instance, the standard linear regression model assumes that $f(y \mid \underset{\sim}{x}; \underset{\sim}{\theta})$ takes the form that

$$E(y \mid \underset{\sim}{x}) = \alpha + \underset{\sim}{\beta}' \underset{\sim}{x}. \tag{2.1}$$

and

$$\mathrm{Var}(y \mid \underset{\sim}{x}) = \sigma^2, \tag{2.2}$$

where $\underset{\sim}{\theta}' = (\alpha, \underset{\sim}{\beta}', \sigma^2)$. Panel data, by its nature, focus on individual outcomes. Factors affecting individual outcomes are numerous. It is rare to be able to assume a common conditional probability density function of y conditional on $\underset{\sim}{x}$ for all cross-sectional units, i, at all time, t. If the conditional density of y given $\underset{\sim}{x}$ varies across i and over t, the fundamental theorems for statistical inference, the laws of large numbers and central limit theorems, will be difficult to implement. Blindly imposing a homogeneity assumption of $f(y \mid \underset{\sim}{x}; \underset{\sim}{\theta})$ across i and over t can lead to severely biased inference. For instance, suppose that the data is generated by

$$y_{it} = \alpha_i + \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + v_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{2.3}$$

as depicted by Figure 1 in which the broken-time ellipses represent the point scatter of individual observation around the mean, represented by the broken straight line. If an investigator mistakenly estimate a model of the form

$$y_{it} = \alpha + \underset{\sim}{\beta}' \underset{\sim}{x} + v_{it}^*. \tag{2.4}$$

The solid line in Figure 1 would depict the pooled least squares regression result which could be completely contradict the individual relation between y and $\underset{\sim}{x}$.

One way to restore homogeneity across i and/or over t is to add more conditional variables, say $\underset{\sim}{z}$,

$$f(y_{it} \mid \underset{\sim}{x}_{it}, \underset{\sim}{z}_{it}; \underset{\sim}{\theta}). \tag{2.3}$$

However, the dimension of $\underset{\sim}{z}$ can be large. A model is a simplification of reality, not a mimic of reality. The inclusion of $\underset{\sim}{z}$ may confuse the fundamental relationship between y and $\underset{\sim}{x}$, in particular, when there is a shortage of degrees of freedom or multicollinearity, etc. Moreover, $\underset{\sim}{z}$ may not be observable. If an investigator is only interested in the relationship between y and $\underset{\sim}{x}$, a common approach to characterize the heterogeneity not captured by $\underset{\sim}{x}$ is to assume that the parameter vector varies across i and over t, $\underset{\sim}{\theta}_{it}$, so that the conditional density of y given $\underset{\sim}{x}$ takes the form $f(y_{it} \mid \underset{\sim}{x}_{it}; \underset{\sim}{\theta}_{it})$. However, without a structure being imposed on $\underset{\sim}{\theta}_{it}$, such a model only has descriptive value, it is not possible to draw any inference.

One way to impose some structure on $\underset{\sim}{\theta}_{it}$ is to decompose $\underset{\sim}{\theta}_{it}$ into $(\underset{\sim}{\beta}, \underset{\sim}{\gamma}_{it})$, where $\underset{\sim}{\beta}$ is the same across i and over t, referred to as *structural parameters*, and $\underset{\sim}{\gamma}_{it}$ as *incidental parameters* because when cross-units, N and/or time series observations, T increases, so is the dimension of $\underset{\sim}{\gamma}_{it}$. The focus of panel data literature is to make inference on $\underset{\sim}{\beta}$ after controlling the impact of $\underset{\sim}{\gamma}_{it}$.

Without imposing structure for $\underset{\sim}{\gamma}_{it}$, again it is not possible to make any inference on $\underset{\sim}{\beta}$ because the unknown $\underset{\sim}{\gamma}_{it}$ will exhaust all available sample information. Assuming that the impacts of observable variables, $\underset{\sim}{x}$, are the same across i and over t, represented by the structure parameters, $\underset{\sim}{\beta}$, the incidental parameters $\underset{\sim}{\gamma}_{it}$ represent the heterogeneity across i and over t that are not captured by $\underset{\sim}{x}_{it}$. They can be considered as composed of the effects of omitted individual time-invariant, α_i, period individual-invariant, λ_t, and individual time-varying variables, u_{it}. The individuals time-invariant variables are variables that are the same for a given cross-sectional unit through time but that vary across cross-sectional

units such as individual-firm management, ability, gender, and socio-economic background variables. The period individual-invariant variables are variables that are the same for all cross-sectional units at a given time but that vary though time such as prices, interest rates, and wide spread optimism or pessimism. The individual time-varying variables are variables that vary across cross-sectional units at a given point in time and also exhibit variations through time such as firm profits, sales and capital stock. In a single equation frmaework, it is a common practice to assume that the effects of omitted individual time-varying variables, u_{it} as random and uncorrelated with $\underset{\sim}{x}$. The individual-specific effects, α_i and time specific effects, λ_t can either be assumed as random variables-referred to as the *random effects* model, or fixed parameters-referred to as the *fixed effects* model.

3. Linear Static Models

A widely used panel data model is to assume that the effects of observed explanatory variables, $\underset{\sim}{x}$, are identical across cross-sectional units, i, and over time, t, while the effects of omitted variables can be decomposed into the individual-specific effects, α_i, time-specific effects, λ_t, and individual time-varying effects, u_{it}, as follows:

$$y_{it} = \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i + \lambda_t + u_{it}, \begin{array}{c} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{3.1}$$

In a single equation framework. individual time effects, u, are assumed to be uncorrelated with $\underset{\sim}{x}$, while α_i and λ_t may or may not correlated with $\underset{\sim}{x}$. When α_i and λ_t are treated as fixed constants, they are parameters to be estimated so whether they are correlated with $\underset{\sim}{x}$ is not an issue. On the other hand, when α_i and λ_t are treated as random, they are typically assumed to be uncorrelated with $\underset{\sim}{x}_{it}$.

For ease of exposition, we shall assume that there are no time specific effects, i.e., $\lambda_t = 0$ for all t and u_{it} are independently, identically distributed (i.i.d) across i and over t. Stack an individuals T time series observations of $(y_{it}, \underset{\sim}{x}'_{it})$ into a vector and a matrix, (3.1) may alternatively be written as

$$\underset{\sim}{y}_i = X_i \underset{\sim}{\beta} + \underset{\sim}{e} \alpha_i + \underset{\sim}{u}_i, i = 1, \ldots, N, \tag{3.2}$$

where $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iT})', X_i = (\underset{\sim}{x}_{i1}, \ldots, \underset{\sim}{x}_{iT})', \underset{\sim}{u}_i = (u_{i1}, \ldots, u_{iT})'$, and $\underset{\sim}{e}$ is a $T \times 1$ vector of 1's.

Let Q be a $T \times T$ matrix satisfying the condition that $Q\underset{\sim}{e} = \underset{\sim}{0}$. Premultiplying (3.2) by Q yields

$$Q\underset{\sim}{y}_i = QX_i\underset{\sim}{\beta} + Q\underset{\sim}{u}_i, \quad i = 1. \ldots, N. \tag{3.3}$$

Equation (3.3) no longer involves α_i. The issue of whether α_i is correlated with $\underset{\sim}{x}_{it}$ or whether α_i should be treated as fixed or random is no longer relevant for (3.3). Moreover, since X_i is exogenous, $E(QX_i\underset{\sim}{u}_i'Q') = QE(X_i\underset{\sim}{u}_i')Q' = \underset{\sim}{0}$ and $EQ\underset{\sim}{u}_i\underset{\sim}{u}_i'Q' = \sigma_u^2 QQ'$. An efficient estimator of $\underset{\sim}{\beta}$ is the generalized least squares estimator (GLS),

$$\hat{\underset{\sim}{\beta}} = \left[\sum_{i=1}^{N} X_i'(Q'Q)^- X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i'(Q'Q)^- \underset{\sim}{y}_i\right], \tag{3.4}$$

where $(Q'Q)^-$ denotes the Moore-Penrose generalized inverse (e.g. Rao (1973)).

When $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}', Q$ is idempotent. The Moore-Penrose generalized inverse of $(Q'Q)^-$ is just $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$ itself. Premultiplying (3.3) by Q is equivalent to transforming (3.1) into a model

$$(y_{it} - \bar{y}_i) = \underset{\sim}{\beta}'(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i) + (u_{it} - \bar{u}_i), \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{3.5}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \bar{\underset{\sim}{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \underset{\sim}{x}_{it}$ and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. The transformation is called *covariance transformation.* The least squares estimator (LS) (or a generalized least squares estimator (GLS)) of (3.5),

$$\hat{\underset{\sim}{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)'\right]^{-1} \left[\sum_{t=1}^{N}\sum_{t=1}^{T}(\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)(y_{it} - \bar{y}_i)\right], \tag{3.6}$$

is called *covariance* estimator or *within* estimator because the estimation of $\underset{\sim}{\beta}$ only makes use of within (group) variation of y_{it} and $\underset{\sim}{x}_{it}$ only. The covariance estimator of $\underset{\sim}{\beta}$ turns out to be also the least squares estimator of (3.1) when $\lambda_t = 0$. It is the best linear unbiased estimator of $\underset{\sim}{\beta}$ if α_i is treated as fixed and u_{it} is i.i.d.

If α_i is random, transforming (3.2) into (3.3) transforms T independent equations (or observations) into $(T-1)$ independent equations, hence the covariance estimator is not as efficient as the efficient generalized least squares estimator if $E\alpha_i \underset{\sim}{x}_{it} = \underset{\sim}{0}$. When α_i is independent of $\underset{\sim}{x}_{it}$ and is independently, identically distributed across i with mean $\underset{\sim}{0}$ and variance σ_α^2, the best linear unbiased estimator (BLUE) of $\underset{\sim}{\beta}$ is GLS,

$$\hat{\underset{\sim}{\beta}} = \left[\sum_{i=1}^{N} X_i' V^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' V^{-1} \underset{\sim}{y}_i\right]. \tag{3.7}$$

where $V = \sigma_u^2 I_T + \sigma_\alpha^2 \underset{\sim}{e}\underset{\sim}{e}', V^{-1} = \frac{1}{\sigma_u^2}\left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}\underset{\sim}{e}\underset{\sim}{e}'\right]$. The GLS is equivalent to first transforming the data by subtracting a fraction $(1-\psi^{1/2})$ of individual means $\bar{y}_i$ and $\bar{\underset{\sim}{x}}_i$ from their corresponding y_{it} and $\underset{\sim}{x}_{it}$, then regressing $[y_{it} - (1-\psi^{1/2})\bar{y}_i]$ on $[\underset{\sim}{x}_{it} - (1-\psi^{1/2})\bar{\underset{\sim}{x}}_i]$, where $\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$. (for detail, see Baltagi (2001), Hsiao (2003)).

When α_i is treated as fixed, the covariance estimator is equivalent to applying LS to the transformed model (3.5). If a variable is time-invariant, like gender dummy, $x_{kit} = x_{kis} = \bar{x}_{ki}$, the transformation eliminates the corresponding variable from the specification. Hence, the coefficients of time-invariant variables cannot be estimated. On the other hand, if α_i is random and uncorrelated with $\underset{\sim}{x}_i, \psi \neq 0$, the GLS can still estimate the coerfficients of those time-invariant variables.

4. Dynamic Models

When the regressors of a linear model contains lagged dependent variables, say, of the form

$$\underset{\sim}{y}_i = \underset{\sim}{y}_{i,-1}\gamma + X_i\underset{\sim}{\beta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i = Z_i\underset{\sim}{\theta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i, \quad i = 1, \ldots, N. \tag{4.1}$$

where $\underset{\sim}{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $Z_i = (\underset{\sim}{y}_{i,-1}, X_i)$ and $\underset{\sim}{\theta} = (\gamma, \underset{\sim}{\beta}')'$. For ease of notation, we assume that y_{i0} are observable. Technically, we can still eliminate the individual-specific effects by premultiplying (4.1) by the transformation matrix Q $(Q\underset{\sim}{e} = \underset{\sim}{0})$,

$$Q\underset{\sim}{y}_i = QZ_i\underset{\sim}{\theta} + Q\underset{\sim}{u}_i. \tag{4.2}$$

However, because of the presence of lagged dependent variables, $EQZ_i\underset{\sim}{u}_i'Q' \neq 0$ even with the assumption that u_{it} is independently, identically distributed across i and over t. For instance, the covariance transformation matrix $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$ transforms (4.1) into the form

$$(y_{it} - \bar{y}_i) = (y_{i,t-1} - \bar{y}_{i,-1})\gamma + (\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)'\underset{\sim}{\beta} + (u_{it} - \bar{u}_i), \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{4.3}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \bar{y}_{i,-1} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}$ and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Although, $y_{i,t-1}$ and u_{it} are uncorrelated under the assumption of serial independence of u_{it}, the covariance between $\bar{y}_{i,-1}$ and u_{it} or $y_{i,t-1}$ and $\bar{u}_i$ is of order (1/T) if $\mid \gamma \mid < 1$. Therefore, the covariance estimator of $\underset{\sim}{\theta}$ creates a bias of order (1/T) when $N \to \infty$ (Anderson and Hsiao (1981, 1982), Nickell (1981)). Since most panel data contain large N but small T, the magnitude of the bias can not be ignored (e.g. with T=10 and γ=0.5, the asymptotic bias is -0.167).

When $EQZ_i\underset{\sim}{u}_i'Q' \neq \underset{\sim}{0}$, one way to obtain a consistent estimator for $\underset{\sim}{\theta}$ is to find instruments W_i that satisfy

$$EW_i\underset{\sim}{u}_i'Q' = \underset{\sim}{0}, \tag{4.4}$$

and

$$\text{rank } (W_iQZ_i) = k, \tag{4.5}$$

where k denotes the dimension of $(\gamma, \underset{\sim}{\beta}')'$, then apply the generalized instrumental variable or generalized method of moments estimator (GMM) by minimizing the objective function

$$\left[\sum_{i=1}^{N} W_i(Q\underset{\sim}{y}_i - QZ_i\underset{\sim}{\theta})\right]' \left[\sum_{i=1}^{N} W_iQ\underset{\sim}{u}_i\underset{\sim}{u}_i'Q'W_i'\right]^{-1} \left[\sum_{i=1}^{N} W_i(Q\underset{\sim}{y}_i - Q\underset{\sim}{Z}_i\underset{\sim}{\theta})\right], \tag{4.6}$$

with respect to $\underset{\sim}{\theta}$. (e.g. Ahn and Honoré (2003), Ahn and Schmidt (1995), Arellano and Bond (1991), Arellano and Bover (1995)). For instance, one may let Q be a $(T-1) \times T$ matrix of the form

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdot & \cdot \\ 0 & -1 & 1 & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -1 & 1 \end{bmatrix}, \tag{4.7}$$

then the transformation (4.2) is equivalent to taking the first difference of (4.1) over time to eliminate α_i for $t = 2, \ldots, T$,

$$\Delta y_{it} = \Delta y_{i,t-1}\gamma + \Delta \underset{\sim}{x}'_{it}\underset{\sim}{\beta} + \Delta u_{it}, \begin{array}{l} i = 1, \ldots, N, \\ t = 2, \ldots, T, \end{array} \tag{4.8}$$

where $\Delta = (1 - L)$ and L denotes the lag operator, $Ly_t = y_{t-1}$. Since $\Delta u_{it} = (u_{it} - u_{i,t-1})$ is uncorrelated with $y_{i,t-j}$ for $j \geq 2$ and $\underset{\sim}{x}_{is}$, for all s, when u_{it} is independently distributed over time and $\underset{\sim}{x}_{it}$ is exogenous, one can let W_i be a $T(T-1)[K + \frac{1}{2}] \times (T-1)$ matrix of the form

$$W_i = \begin{bmatrix} \underset{\sim}{q}_{i2} & \underset{\sim}{0} & \cdot & \cdot \\ \underset{\sim}{0} & \underset{\sim}{q}_{is} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \underset{\sim}{q}_{iT} \end{bmatrix}, \tag{4.9}$$

where $\underset{\sim}{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2}, \underset{\sim}{x}'_i)'$, $\underset{\sim}{x}_i = (\underset{\sim}{x}'_{i1}, \ldots, \underset{\sim}{x}'_{iT})'$, and $K = k - 1$. Under the assumption that $(\underset{\sim}{y}'_i, \underset{\sim}{x}'_i)$ are independently, identically distributed across i, the Arellano-Bover (1991) GMM estimator takes the form

$$\begin{aligned} \hat{\underset{\sim}{\theta}}_{AB,GMM} = & \left\{ \left[\sum_{i=1}^{N} Z'_i D' W'_i \right] \left[\sum_{i=1}^{N} W_i A W'_i \right]^{-1} \left[\sum_{i=1}^{N} W_i D Z_i \right] \right\}^{-1} \\ & \left\{ \left[\sum_{i=1}^{N} Z'_i D W'_i \right] \left[\sum_{i=1}^{N} W_i A W'_i \right]^{-1} \left[\sum_{i=1}^{N} W_i D \underset{\sim}{y}_i \right] \right\} \end{aligned} \tag{4.10}$$

where A is a $(T-1) \times (T-1)$ matrix with 2 on the diagonal elements, -1 on the elements above and below the diagonal elements and 0 elsewhere.

The GMM estimator has the advantage that it is consistent and asymptotically normally distributed whether α_i is treated as fixed or random because it eliminates α_i from the specification. However, the number of moment conditions increases at the order of T^2 which can create severe downward bias in finite sample (Ziliak (1997)). An alternative is to use a (quasi-) likelihood approach which has the advantage of having a fixed number of orthogonality conditions independent of the sample size. It also has the advantage of making use all the available sample, hence can yield more efficient estimator than (4.10) (e.g.

Hsiao, Pesaran and Tahmiscioglu (2002), Binder, Hsiao and Pesaran (2004)). Since there is no reason to assume that the data generating process of initial observations, y_{i0}, to be different from the rest of y_{it}, the likelihood approach has to formulate the joint likelihood function of $(y_{i0}, y_{i1}, \ldots, y_{iT})$ (or the conditional likelihood function $(y_{i1}, \ldots, y_{iT} \mid y_{i0})$). However, y_{i0} depends on previous values of $\underset{\sim}{x}_{i,-j}$ and α_i which are unavailable. Bhargava and Sargan (1983) suggest to circumscribe this missing data problem by conditioning y_{i0} on $\underset{\sim}{x}_i$ and α_i if α_i is treated as random while Hsiao, Pesaran and Tahmisciogulu (2002) propose conditioning $(y_{i1} - y_{i0})$ on the first difference of $\underset{\sim}{x}_i$ if α is treated as fixed constants.

5. Random vs Fixed Effects Specification

The advantages of random effects (RE) specifications are:

1. The number of parameters stay constant when sample size increases.
2. It allows the derivation of efficient estimators that make use of both within and between (group) variation.
3. It allows the estimation of the impact of time-invariant variables.

The disadvantages of RE specification is that it typically assumes that the individual- and/or time-specific effects are randomly distributed with a common mean and are independent of $\underset{\sim}{x}_{it}$. If the effects are correlated with $\underset{\sim}{x}_{it}$ or if there is a fundamental difference among individual units, i.e., conditional on $\underset{\sim}{x}_{it}, y_{it}$ cannot be viewed as a random draw from a common distribution, common RE model is misspecified and the resulting estimator is biased.

The advantages of fixed effects (FE) specification are that it allows the individual- and/or time specific effects to be correlated with explanatory variables $\underset{\sim}{x}_{it}$. Neither does it require an investigator to model their correlation patterns.

The disadvantages of the FE specification are:

1'. The number of unknown parameters increases with the number of sample observations. In the case when T (or N for λ_t) is finite, it introduces the classical

incidental parameter problem (e.g. Neyman and Scott (1948)).

2'. The FE estimator does not allow the estimation of the coefficients that are time-invariant.

In other words, the advantages of RE specification are the disadvantages of FE specification and the disadvantages of RE specification are the advantages of FE specification. To choose between the two specifications, Hausman (1978) note that the FE estimator (or GMM), $\hat{\underset{\sim}{\theta}}_{FE}$, is consistent whether α_i is fixed or random. On the other hand, the commonly used RE estimator (or GLS), $\hat{\underset{\sim}{\theta}}_{RE}$, is consistent and efficient only when α_i is indeed uncorrelated with $\underset{\sim}{x}_{it}$. If α_i is correlated with $\underset{\sim}{x}_{it}$, the RE estimator is inconsistent. Therefore, Hausman (1978) suggests using the statistic

$$\left(\hat{\underset{\sim}{\theta}}_{FE} - \hat{\underset{\sim}{\theta}}_{RE}\right)' \left[\mathrm{Cov}(\hat{\underset{\sim}{\theta}}_{FE}) - \mathrm{Cov}(\hat{\underset{\sim}{\theta}}_{RE})\right]^{-} \left(\hat{\underset{\sim}{\theta}}_{FE} - \hat{\underset{\sim}{\theta}}_{RE}\right) \tag{5.1}$$

to test RE vs FE specification. The statistic (5.1) is asymptotically chi-square distributed with degrees of freedom equal to the rank of $\left[\mathrm{Cov}(\hat{\underset{\sim}{\theta}}_{GMM}) - \mathrm{Cov}(\hat{\underset{\sim}{\theta}}_{RE})\right]$.

6. Nonlinear Models

The introduction of individual-specific effects, α_i, and/or time-specific effects, λ_t, provide a simple way to capture the unobserved heterogeneity across i and over t. However, the likelihood functions are in terms of observables, $(\underset{\sim}{y}_i, \underset{\sim}{x}_i), i = 1, \ldots, N$. Therefore, we will have to either treat α_i as unknown parameters (fixed effects) and consider the conditional likelihood,

$$f(\underset{\sim}{y}_i \mid \underset{\sim}{x}_i, \underset{\sim}{\beta}, \alpha_i), i = 1, \ldots, N, \tag{6.1}$$

or to treat α_i as random and consider the marginal likelihood

$$f(y_i \mid \underset{\sim}{x}_i; \underset{\sim}{\beta}) = \int f(y_i \mid \underset{\sim}{x}_i, \underset{\sim}{\beta}, \alpha_i) f(\alpha_i \mid \underset{\sim}{x}_i) d\alpha_i, i = 1, \ldots, N, \tag{6.2}$$

where $f(\alpha_i \mid \underset{\sim}{x}_i)$ denotes the conditional density of α_i given $\underset{\sim}{x}_i$.

When the unobserved individual specific effects, α_i, (and or time-specific effects, λ_t) affect the outcome, y_{it}, linearly, one can avoid the consideration of random versus fixed

effects specification by eliminating them from the specification through some linear transformation such as the covariance transformation (3.3) or first difference transformation (4.8). However, if α_i affects y_{it} nonlinearly, it is not easy to find transformation that can eliminate α_i. For instance, consider the following binary choice model where the observed y_{it} takes the value of either 1 or 0 depending on the latent response function

$$y_{it}^* = \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i + u_{it}, \tag{6.3}$$

and

$$y_{it} = \begin{cases} 1, & \text{if } y_{it}^* > 0, \\ 0, & \text{if } y_{it}^* \leq 0, \end{cases} \tag{6.4}$$

where u_{it} is independently, identically distributed with density function $f(u_{it})$. Let

$$y_{it} = E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i) + \epsilon_{it}, \tag{6.5}$$

then

$$\begin{aligned} E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i) &= \int_{-(\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i)}^{\infty} f(u) du \\ &= [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]. \end{aligned} \tag{6.6}$$

Since α_i affects $E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i)$ nonlinearly, α_i remains after taking successive difference of y_{it},

$$\begin{aligned} y_{it} - y_{i,t-1} &= [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)] \\ &\quad - [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{i,t-1} - \alpha_i)] + (\epsilon_{it} - \epsilon_{i,t-1}). \end{aligned} \tag{6.7}$$

The likelihood function conditional on $\underset{\sim}{x}_i$ and α_i takes the form,

$$\Pi_{i=1}^N \Pi_{t=1}^T [F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{1-y_{it}} [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{y_{it}}. \tag{6.8}$$

If T is large, consistente estimator of $\underset{\sim}{\beta}$ and α_i can abe obtained by maximizing (6.8). If T is finite, there is only limited information about α_i no matter how large N is. The presence of incidental parameters, α_i, violates the regularity conditions for the consistency of the maximum likelihood estimator of $\underset{\sim}{\beta}$.

If $f(\alpha_i \mid \underset{\sim}{x}_i)$ is known, and is characterized by a fixed dimensional parameter vector, consistent estimator of $\underset{\sim}{\beta}$ can be obtained by maximizing the marginal likelihood function,

$$\Pi_{i=1}^{N} \int \Pi_{t=1}^{T} [F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{1-y_{it}} [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{y_{it}} f(\alpha_i \mid \underset{\sim}{x}_i) d\alpha_i. \tag{6.9}$$

However, maximizing (6.9) involves T-dimensional integration. Butler and Moffit (1982), Chamberlain (1984), Heckman (1981), etc., have suggested methods to simplify the computation.

The advantage of RE specification is that there is no incidental parameter problem. The problem is that $f(\alpha_i \mid \underset{\sim}{x}_i)$ is in general unknown. If a wrong $f(\alpha_i \mid \underset{\sim}{x}_i)$ is postulated, maximizing the wrong likelihood function will not yield consistent estimator of $\underset{\sim}{\beta}$. Moreover, the derivation of marginal likelihood through multiple integration may be computationally infeasible. The advantage of FE specification is that there is no need to specify $f(\alpha_i \mid \underset{\sim}{x}_i)$. The likelihood function will be the product of individual likelihood (e.g. (6.8)) if the errors are assumed i.i.d. The disadvantage is that it introduces incidental parameters.

A general approach of estimating a model involving incidental parameters is to find transformations to transform the original model into a model that does not involve incidental parameters. Unfortunately, there is no general rule available for nonlinear models. One has to explore the specific structure of a nonlinear model to find such a transformation. For instance, if $f(u)$ in (6.3) is logistic, then

$$\text{Prob } (y_{it} = 1 \mid \underset{\sim}{x}_{it}, \alpha_i) = \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i}}{1 + e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i}}. \tag{6.10}$$

Since, in a logit model, the denominators of Prob$(y_{it} = 1 \mid \underset{\sim}{x}_{it}, \alpha_i)$ and Prob$(y_{it} = 0 \mid \underset{\sim}{x}_{it}, \alpha_i)$ are identical and the numerator of any sequence $\{y_{i1}, \ldots, y_{iT}\}$ with $\sum_{t=1}^{T} y_{it} = s$ is always equal to $\exp(\alpha_i s) \cdot \exp\{\sum_{t=1}^{T} (\underset{\sim}{\beta}' \underset{\sim}{x}_{it}) y_{it}\}$, the conditional likelihood function conditional on $\sum_{t=1}^{T} y_{it} = s$ will not involve the incidental parameters α_i. For instance, consider the

simple case that $T = 2$, then

$$\begin{aligned}\text{Prob}(y_{i1} = 1, y_{i2} = 0 \mid y_{i1} + y_{i2} = 1) &= \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i1}}}{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i1}} + e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i2}}} \\ &= \frac{1}{1 + e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}},\end{aligned} \tag{6.11}$$

and

$$\text{Prob}(y_{i1} = 0, y_{i2} = 1 \mid y_{i1} + y_{i2} = 1) = \frac{e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}}{1 + e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}}, \tag{6.12}$$

(Chamberlain (1980), Hsiao (2003)).

Alternatively, Manski (1987) exploits the latent linear structure of (6.3) by noting that for given i,

$$\underset{\sim}{\beta}' \underset{\sim}{x}_{it} \gtreqless \underset{\sim}{\beta}' \underset{\sim}{x}_{i,t-1} \Longleftrightarrow E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i) \gtreqless E(y_{i,t-1} \mid \underset{\sim}{x}_{i,t-1}, \alpha_i), \tag{6.13}$$

and suggests maximizing the objective function

$$H_N(b) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T} sgn(\underset{\sim}{b}' \Delta \underset{\sim}{x}_{it}) \Delta y_{it}, \tag{6.14}$$

where $sgn(w) = 1$ if $w > 0, = 0$ if $w = 0$, and -1 if $w < 0$. The advantage of the Manski (1987) maximum score estimator is that it is consistent without the knowledge of $f(u)$. The disadvantage is that (6.13) holds for any $c\underset{\sim}{\beta}$ where $c > 0$. Only the relative magnitude of the coefficients can be estimated with some normalization rule, say $\| \underset{\sim}{\beta} \| = 1$. Moreover, the spped of convergence is considerably slower ($N^{1/3}$) and the limiting distribution is quite complicated. Horowitz (19) and Lee ( ) have proposed modified estimators that improve the speed of convergence and are asymptotically normally distributed.

Other examples of exploiting specific structure of nonlinear models to eliminate the effects of incidental parameters α_i include dynamic discrete choice models (Chamberlain (1993), Honoré and Kyriazidou (2000), Hsiao, Shen, Wang and Weeks (2004)), symmetrically trimmed least squares estimator for truncated and censored data (Tobit models) (Honoré (1992)), sample selection models (or type II Tobit models) (Kyriazidou (1997)),

etc. However, often they impose very severe restrictions on the data such that not much information of the data can be utilized to obtain parameter estimates. Moreover, there are models such that there does not appear to possess consistent estimator when T is finite.

An alternative to consider consistent estimators is to consider bias reduced estimator. The advantage of such an approach is that the bias reduced estimators may still allow the use of all the sample information so that from a mean square error point of view, the bias reduced estimator may still dominate a consistent estimators because the latter often have to throw away a lot of sample, thus tend to have large variances.

Following the idea of Cox and Reid (1987), Arellano (2001) and Carro (2004) propose to derive the modified MLE by maximizing the modified log-likelihood function

$$L^*(\underset{\sim}{\beta}) = \sum_{i=1}^{N} \left[\ell_i^*(\underset{\sim}{\beta}, \hat{\alpha}_i(\underset{\sim}{\beta})) - \frac{1}{2} \log \ell_{i,d_i d_i}^*(\beta_1 \hat{\alpha}_i(\underset{\sim}{\beta}) \right], \tag{6.15}$$

where $\ell_i^*(\underset{\sim}{\beta}, \hat{\alpha}_i(\underset{\sim}{\beta}))$ denotes the concentrated log-likelihood function of $\underset{\sim}{y}_i$ after substituting the MLE of α_i in terms of $\underset{\sim}{\beta}, \hat{\alpha}_i(\underset{\sim}{\beta})$, (i.e., the solution of $\frac{\partial \log L}{\partial \alpha_i} = 0$ in terms of $\underset{\sim}{\beta}, i = 1, \ldots, N$), into the log-likelihood function and $\ell_{i,\alpha_i\alpha_i}^*(\underset{\sim}{\beta}, \hat{\alpha}_i(\underset{\sim}{\beta}))$ denotes the second derivative of ℓ_i^* with respect to α_i. The bias correction term is derived by noting that to the order of $(1/T)$ the first derivative of ℓ_i^* with respect to $\underset{\sim}{\beta}$ converges to $\frac{1}{2} \frac{E[\ell_{i,\beta\alpha_i\alpha_i}^*(\underset{\sim}{\beta}, \alpha_i)]}{E[\ell_{i,\alpha_i\alpha_i}^*(\underset{\sim}{\beta}, \alpha_i)]}$. By subtracting the order (1/T) bias from the likelihood function, the modified MLE is biased only to the order of $(1/T^2)$, without increasing the asymptotic variance.

Monte Carlo experiments conducted by Carro (2004) have shown that when $T = 8$, the bias of modified MLE for dynamic probit and logit models are negligible. Another advantage of the Arellano-Carro approach is its generality. For instance, a dynamic logit model with time dummy explanatory variable can not meet the Honoré and Kyriazidou (2000) conditions for generating consistent estimator, but will not affect the asymptotic properties of the modified MLE.

7. Modeling Cross-Sectional Dependence

Most panel studies assume that apart from the possible presence of individual in-

variant but period varying time specific effects, λ_t, the effects of omitted variables are independently distributed across cross-sectional units. However, often economic theory predicts that agents take actions that lead to interdependence among themselves. For example, the prediction that risk averse agents will make insurance contracts allowing them to smooth idiosyncratic shocks implies dependence in consumption across individuals. Ignoring cross-sectional dependence can lead to inconsistent estimators, in particular when T is finite (e.g. Hsiao and Tahmiscioglu (2005)). Unfortunately, contrary to the time series data in which the time label gives a natural ordering and structure, general forms of dependence for cross-sectional dimension are difficult to formulate. Therefore, econometricians have relied on strong parametric assumptions to model cross-sectional dependence. Two approaches have been proposed to model cross-sectional dependence: economic distance or spatial approach and factor approach.

In regional science, correlation across cross-section units is assumed to follow a certain spatial ordering, i.e. dependence among cross-sectional units is related to location and distance, in a geographic or more general economic or social network space (e.g. Anselin (1988), Anselin and Griffith (1988), Anselin, Le Gallo and Jayet (2005)). A known spatial weights matrix, $W = (w_{ij})$ an $N \times N$ positive matrix in which the rows and columns correspond to the cross-sectional units, is specified to express the prior strength of the interaction between individual (location) i (in the row of the matrix) and individual (location) j (column), w_{ij}. By convention, the diagonal elements, $w_{ij} = 1$. The weights are often standardized so that the sum of each row, $\sum_{j=1}^{N} w_{ij} = 1$.

The spatial weight matrix, W, is often included into a model specification to the dependent variable, to the explanatory variables, or to the error term. For instance, a *spatial lag* model for the $NT \times 1$ variable $\underset{\sim}{y} = (\underset{\sim}{y}_1', \ldots, \underset{\sim}{y}_N')', \underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iT})'$, may take the form

$$y = \rho(W \otimes I_T)\underset{\sim}{y} + X\underset{\sim}{\beta} + \underset{\sim}{u} \tag{7.1}$$

where X and $\underset{\sim}{u}$ denote the $NT \times K$ explanatory variables and $NT \times 1$ vector of error terms,

respectively, and $\otimes$ denotes the Kronecker product. A *spatial error* model may take the form,

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{v}, \tag{7.2}$$

where $\underset{\sim}{v}$ may be specified as in a *spatial autoregressive* form,

$$\underset{\sim}{v} = \theta(W \otimes I_T)\underset{\sim}{v} + \underset{\sim}{u}, \tag{7.3}$$

or a *spatial moving average* form,

$$\underset{\sim}{v} = \gamma(W \otimes I_T)\underset{\sim}{u} + \underset{\sim}{u}. \tag{7.4}$$

The spatial model can be estimated by the instrumental variables (generalized method of moments estimator) or the maximum likelihood method. However, the approach of defining cross-sectional dependence in terms of "economic distance" measure requires that the econometricians have information regarding this "economic distance". Another approach to model cross-sectional dependence is to assume that the error of a model, say model (7.3) follows a linear factor model,

$$v_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + u_{it}, \tag{7.5}$$

where $\underset{\sim}{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of random factors, $\underset{\sim}{b}_i' = (b_{i1}, \ldots, b_{ir})$, is a $r \times 1$ nonrandom factor loading coefficients, u_{it}, represents the effects of idiosyncratic shocks which is independent of $\underset{\sim}{f}_t$ and is independently distributed across i. (e.g. Bai and Ng (2002), Moon and Perron (2004), Pesaran (2004)). The conventional time-specific effects model is a special case of (7.5) when $r = 1$ and $b_i = b_\ell$ for all i and ℓ.

The factor approach requires considerably less prior information than the economic distance approach. Moreover, the number of time-varying factors, r, and factor load matrix $B = (b_{ij})$ can be empirically identified if both N and T are large. However, when T is large, one can estimate the covariance between i and j, σ_{ij}, by $\frac{1}{T}\sum_{t=1}^{T} \hat{v}_{it}\hat{v}_{jt}$ directly, then apply the generalized least squares method, where $\hat{v}_{it}$ is some preliminary estimate of v_{it}.

8. Concluding Remarks

In this paper we have tried to provide a summary of advantages of using panel data and the fundamental issues of panel data analysis. Assuming that the heterogeneity across cross-sectional units and over time that are not captured by the observed variables can be captured by period-invariant individual specific and/or individual-invariant time specific effects, we surveyed the fundamental methods for the analysis of linear static and dynamic models. We have also discussed difficulties of analyzing nonlinear models and modeling cross-sectional dependence. There are many important issues such as the modeling of joint dependence or simultaneous equations models, time-varying parameter models (e.g. Hsiao (1992, 2003), tests of unit root or cointegration (e.g., Levin, Lin and Chu (2003), Pesaran, Shin and Smith (2004), Hsiao and Pesaran (2004)), the asymptotics for panels with large N and T (e.g. Phillips and Moon (1999)), unbalanced panel, measurement errors (Griliches and Hausman (1986), Wansbeek and Konig (1989)), etc. that were not discussed, but could be found in Baltagi (2001) or Hsiao (2003).

Although panel data offer many advantages, they are not panacea. The power of panel data to isolate the effects of specific actions, treatments or more general policies depends critically on the compatibility of the assumptions of statistical tools with the data generating process. In choosing the proper method, for exploiting the richness and unique properties of the panel, it might be helpful to keep the following factors in mind: First, what advantages do panel data offer us in investigating economic issues over data sets consisting of a single cross section or time series? Second, what are the limitations of panel data and the econometric methods that have been proposed for analyzing such data? Third, when using panel data, how can we increase the efficiency of parameter estimates? Fourth, are the assumptions underlying the statistical inference procedures and the data-generating process compatible.

References

Ahn, S.C. and P. Schmidt (1995), " Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics*, 68, 5-27.

Anderson, T.W. and C. Hsiao (1981), "Estimation of Dynamic Models with Error Components", *Journal of the American Statistical Association*, 767, 598-606.

_____ (1982), "Formulation and Estimation of Dynamic Models Using Panel Data", *Journal of Econometrics*, 18, 47-82.

Anselin, L., J. Le Gallo and H. Jayet (2005), "Spatial Panel Econometrics", *mimeo*.

Arellano, M. and S.R. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations", *Review of Economic Studies*, 58, 277-297.

Arellano, M. and O. Bover (1995), "Another Look at the Instrumental Variable Estimation of Error-Components Models", *Journal of Econometrics*, 68, 29-51.

Bai, J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models", *Econometrica*, 70. 91-121.

Balestra, P. and H. Nerlove (1966), "Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas", *Econometrica*, 34, 585-612.

Baltagi, B.H. (2001), "Econometric Analysis of Panel Data", Second edition, New York: Wiley.

Ben-Porath, Y. (1973), "Labor Force Participation Rates and the Supply of Labor", *Journal of Political Economy*, 81, 697-704.

Bhargava, A. and J.D. Sargan (1983), "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods", *Econometrica*, 51, 1635-1659.

Binder, M., C. Hsiao and M.H. Pesaran (2004), "Estimation and Inference in Short Panel Vector Autoregressions with Unit Roots and Cointegration", *Econometric Theory* (forthcoming).

Butler, J.S. and R. Moffitt (1982), "A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model", *Econometrica*, 50, 761-764.

Carro, J.M. (2004), "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", *mimeo*.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitatiave Data", *Review of Economic Studies*, 47, 225-238.

_____ (1993), "Feedback in Panel Data Models", *mimeo*, Department of Economics, Harvard University.

Granger, C.W.J. (1990), "Aggregation of Time-Series Variables: A Survey", in *Disag-*

gregation in Econometric Modeling, ed. by T. Barker and M.H. Pesaran, London: Routledge.

Griliches, Z. and J.A. Hausman (1986), "Errors-in-Variables in Panel Data", *Journal of Econometrics*, 31, 93-118.

Hausman, J.A. (1978), "Specification Tests in Econometrica", *Econometrica*, 46. 1251-71.

Heckman, J.J. (1981a), "Statistical Models for Discrete Panel Data", in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C.F. Manski and D. McFadden, Cambridge, Mass., MIT Press, 114-178.

Honoré, Bo and E. Kyriazidou (2000),. "Panel Data Discrete Choice Models with Lagged Dependent Variables", *Econometrica*, 68, 839-874.

Honoré, Bo (1992), "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects", *Econometrica*, 60, 533-567.

Hsiao, C., (1992), "Rnadom Coefficient Models" in *The Econometrics of Panel Data*, ed. by L. Matyas and P. Sevestres, Kluwer, 1st edition, 223-241, 2nd ed. (1996), 410-428.

_____ (2003), *Analysis of Panel Data*, 2nd edition, Cambridge: Cambridge University Press (Econometric Society monograph no. 34).

Hsiao, C., M.H. Pesaran and A.K. Tahmiscioglu (2002), "Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Time Periods", *Journal of Econometrics*, 109, 107-150.

Hsiao, C., Y. Shen and H. Fujiki (2004), "Aggregate vs Disaggregate Data Analysis — A Paradox in the Estimation of Money Demand Function of Japan Under the Low Interest Rate Policy", *Journal of Applied Econometrics* (forthcoming).

Kyriazidou, E. (1997), "Estimation of a Panel Data Sample Selection Model", *Econometrica*, 65, 1335-1364.

Levin, A., C. Lin, and J. Chu (2002), "Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties", *Journal of Econometrics*.

MaCurdy, T.E. (1981), "An Empirical Model of Labor Supply in a Life Cycle Setting", *Journal of Political Economy*, 89, 1059-85.

Manski, C.F. (1987), "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55, 357-362.

Moon, H.R. and B. Perron (2004), "Testing for a Unit Roots in Panels with Dynamic Factors", *Journal of Econometrics*, 122, 81-126.

Neyman, J. and E.L. Scott (1948), "Consistent Estimates Based on Partially Consistent Observations", *Econometrica*, 16, 1-32.

Nickell, S. (1981), "Biases in Dynamic Models with Fixed Effects", *Econometrica*, 49, 1399-1416.

Pakes, A. and Z. Griliches (1984), "Estimating Distributed Lags in Short Panels with

an Application to the Specification of Depreciation Patterns and Capital Stock Constructs", *Review of Economic Studies*, 51, 243-262.

Pesaran, M.H. (2003), "A Simple Panel Unit Root Test in the Presence of Cross-Section Dependence", *mimeo*.

Phillips, P.C.B. (1986), "Understanding Spurious Regressions in Econometrics", *Journal of Econometrics*, 33, 311-340.

Phillips, P.C.B. and H.R. Moon (1999), "Linear Regression Limit Theory for Nonstationary Panel Data", *Econometrica*, 67, 1057-1111.

Rao, C.R., (1973), *Linear Statistical Inference and Its Applications*, 2nd ed., New York: Wiley.

Wansbeek, T.J. and R.H. Koning (1989), "Measurement Error and Panel Data", *Statistica Neerlandica*, 45, 85-92.

Zilak, J.P. (1997), "Efficient Estimation with Panel Data When Instruments are Predetermined: An Empirical Comparison of Moment-Condition Estimators", *Journal of Business and Economic Statistics*, 15, 419-431.

The New Palgrave Dictionary of Economics Online

Lorenz curve

Nanak Kakwani
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The Lorenz curve is the most widely used technique to represent and analyse the size distribution of income and wealth. The curve plots cumulative proportion of income units and the cumulative proportion of income received when income units are arranged in ascending order of their income. Max Otto Lorenz, a statistician (born 19 September 1876 in Burlington, USA; retired 1944), proposed this curve in 1905 in order to compare and analyse inequalities of wealth in a country during different epochs, or in different countries during the same epoch – and since then, the curve has been widely used as a convenient graphical device to summarize the information collected about the distributions of income and wealth.
The Lorenz curve may be represented by a function $L(p)$, which is interpreted as the fraction of total income received by the lowest pth fraction of income units. It satisfies the following conditions (Kakwani, 1980):

1. (a) if $p=0$, $L(p)=0$
2. (b) if $p=1$, $L(p)=1$
3. (c) $L'(p)=(x/\mu)\geq 0$ and $L''(p)=(1/\mu f(x))>0$
4. (d) $L(p)\geq p$

where income x of a unit (which can be negative for some units but is assumed to be non-negative here for notational convenience) is a random variable with the probability density function $f(x)$ with mean μ and $L'(p)$ and $L''(p)$ are the first and second derivatives of $L(p)$ with respect to p, respectively.
A hypothetical Lorenz curve is illustrated in Figure 1. The ordinate and abscissa of the curve are $L(p)$ and p, respectively. The slope of the Lorenz curve is positive and increases monotonically, in other words, the curve is convex to the p-axis. From this it follows that $L(p)<p$. The straight line represented by the equation $L(p)=p$, is called the egalitarian line. The curve lies below this line. If, however, the curve coincides with the egalitarian line, it means that each unit receives the same income, which is the case of perfect equality of incomes. In the case of perfect inequality of incomes, the Lorenz curve coincides with OA and AB, which implies that all income is received by only one unit.
Figure 1
The Lorenz curve

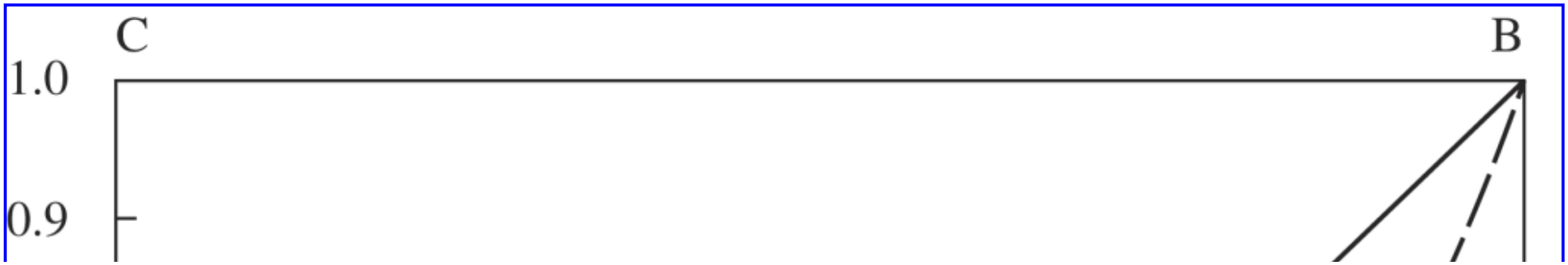

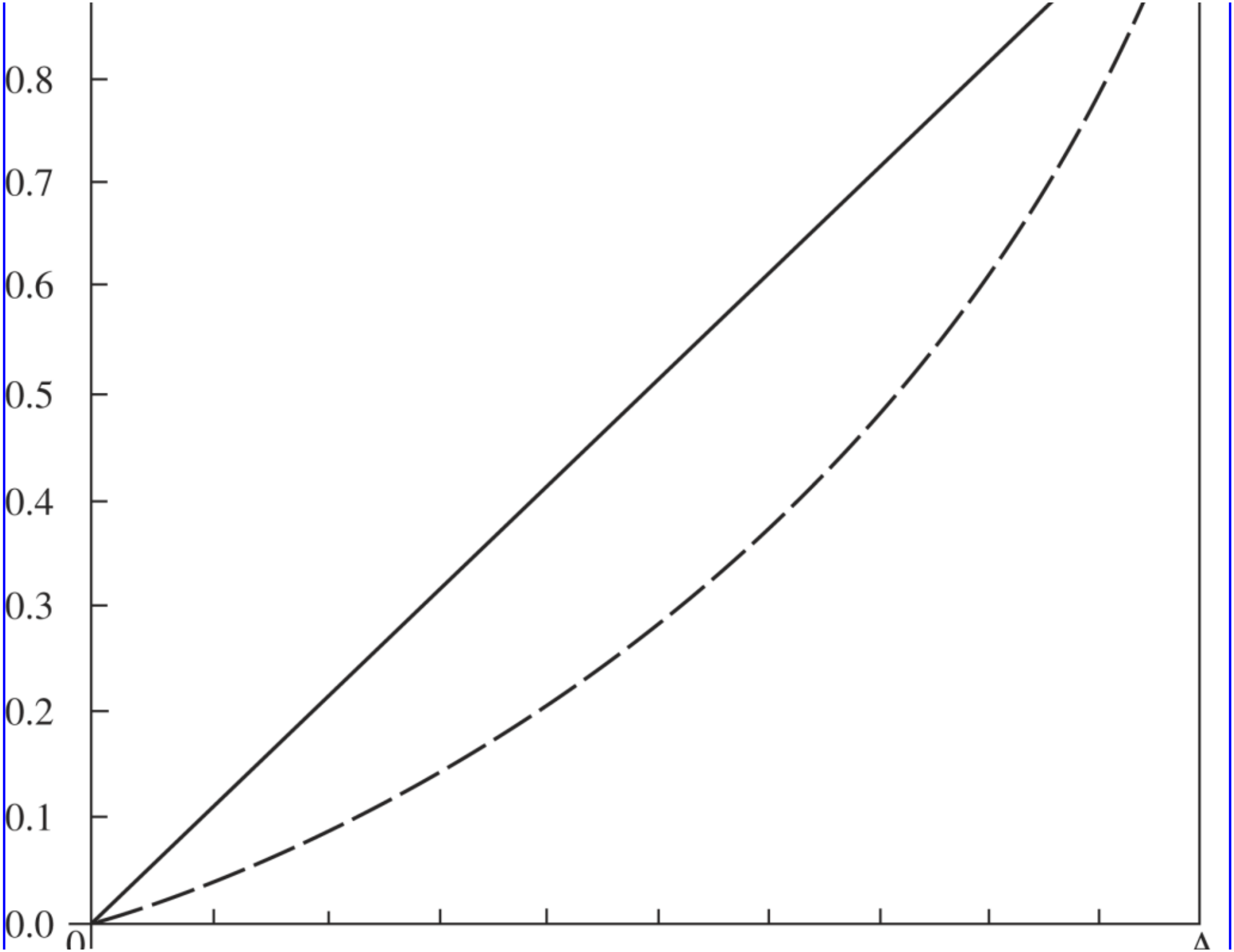
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
A

Since the Lorenz curve displays the deviation of each individual income from perfect equality, it captures, in a sense, the essence of inequality. The nearer the Lorenz curve is to the egalitarian line, the more equal the distribution of income will be. Consequently, the Lorenz curve could be used as a criterion for ranking income distributions: for if the Lorenz curve for one distribution, *X*, lies everywhere above that for another distribution, *Y*, then the distribution *X* may be said to be more equal than the distribution *Y*. However, the ranking provided by the curve is only partial – when two Lorenz curves intersect, neither distribution can be said to be more equal than the other. This partial ranking (or quasi-ordering as Sen (1973) calls it) need not, however, be considered a weakness of the Lorenz curve. In fact Sen (1973) criticizes the inequality measures that provide complete orderings on the grounds that ‘the concept of inequality has different facets which may point in different directions and sometimes a total ranking can not be expected to emerge’. According to him, the concept of inequality is essentially a question of partial ranking and the Lorenz curve is consistent with such a notion of inequality.

Is there any relation between the Lorenz curve ranking of distributions and social welfare? The answer has been provided by Atkinson (1970) who proved a theorem which shows that if social welfare is the sum of the individual utilities and every individual has an identical utility function which is concave, the ranking of distributions according to the Lorenz curve criterion is identical to the ranking implied by the social welfare function, provided the distributions have the same mean income and their Lorenz curves do not intersect. This theorem implies that one can judge between the distributions without knowing the form of the utility function except that it is increasing and concave. If the Lorenz curves do intersect, however, two utility functions that will rank the distributions differently can always be found.

Atkinson's theorem is based on the assumption that the social welfare function is equal to the sum of individual utilities and that every individual has the same utility function. These assumptions are somewhat limited and have been criticized by DasGupta, Sen and Starrett (1973) as well as by Rothschild and Stiglitz (1973), who have demonstrated that the result is, in fact, more general and would hold for any symmetric welfare function that is quasi-concave.

The Lorenz curve makes distributional judgements independently of the size of income, which as Sen (1973) points out, ‘will make sense only if the relative ordering of welfare levels of distributions were strictly neutral to the operation of multiplying everybody's income by a given number’. This is rather an extreme requirement because social welfare depends on both size and the distribution of income.

Working independently on extensions of the Lorenz partial ordering, Shorrocks (1983) and Kakwani (1984) arrived at a criterion which would rank any two distributions with different mean incomes. The new criterion is given by $L(\mu, p)$, which is the product of the mean income μ and the Lorenz curve $L(p)$, whereas the Lorenz curve ranking is based only on $L(p)$. Ranking the distributions according to $L(\mu, p)$ will be identical to the Lorenz ranking if the distributions have the same mean income. This criterion of ranking has been justified from the welfare point of view in terms of several alternative classes of social welfare functions. Kakwani (1984) has used this criterion for international comparison of welfare using data from 72 countries.

As pointed out in the beginning, the Lorenz curve technique was devised as a convenient graphical method to represent and analyse the size distributions of income and wealth. The technique has proved to be extremely powerful and its applications in many areas of applied economics have recently been explored. In analysing data on consumer expenditures Mahalanobis (1960) developed a new technique ‘Fractile Graphical Analyses’ for comparison of socioeconomic groups at different places or points of time. In this paper, he proposed to extend and generalize the concept of the Lorenz curve to deal with problems of consumer behaviour patterns with respect to different commodities. He suggested that generalized Lorenz curves be called concentration curves, and in fact, used them as a convenient graphical device to describe consumption patterns for different commodities based on data from the National Sample Survey of India.

Kakwani (1977, 1980) provided, however, a more general and rigorous treatment of concentration curves in order to study the relationships among the distributions of different economic variables. He proved theorems which have many applications, particularly in the field of public finance where the effect of taxation and public spending of income distribution is analysed. Other areas in which concentration curves can be applied are inflation as it affects income distribution, estimation of Engel elasticities, disaggregation of total inequality by factor components, and economic growth and income distribution. In a later contribution he used concentration curves to explore how the sense of envy felt by individuals affects the optimal tax structure (Kakwani, 1985).

See Also

- Gini ratio
- Pareto distribution
- poverty

Bibliography

Atkinson, A.B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 244–63.

Dasgupta, P., Sen, A.K. and Starrett, D. 1973. Notes on the measurement of inequality. *Journal of Economic Theory* 6, 180–87.

Kakwani, N. 1977. Applications of Lorenz curves in economic analysis. *Econometrica* 45, 719–27.

Kakwani, N. 1980. *Inequality and Poverty: Methods of Estimation and Policy Applications*. New York: Oxford University Press.

Kakwani, N. 1984. Welfare ranking of income distributions. *Advances in Econometrics*, 191–213.

Kakwani, N. 1985. Applications of concentration curves to optimal negative income taxation. *Journal of Quantitative Economics* 1(1).

Lorenz, M.O. 1905. Methods for measuring concentration of wealth. *Journal of the American Statistical Association* 9, 209–19.

Mahalanobis, P.C. 1960. A method of fractile graphical analysis. *Econometrica* 28, 325–51.

Rothschild, M. and Stiglitz, J.E. 1973. Some further results on the measurement of inequality. *Journal of Economic Theory* 6(2), April,188–204.

Sen, A. 1973. *On Economic Inequality*. Oxford: Clarendon Press.

Shorrocks, A.F. 1983. Ranking income distributions. *Economica* 50, 3–18.

How to cite this article

Kakwani, Nanak. "Lorenz curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000149> doi:10.1057/9780230226203.0995

The New Palgrave Dictionary of Economics Online

Lösch, August (1906 – 1945)

Wolfgang F. Stolper
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; distance; Great Depression; location; Lösch, A.; partial equilibrium; population cycles; regional economics; secular stagnation; transfer problem

Article

Lösch was born on 15 October 1906 in Oehringen (Württ), though he considered Heidenheim (Brenz) his home. He went to school there, studied in Freiburg with Eucken and in Bonn with Schumpeter and Spiethoff. He was twice a Rockefeller Fellow in the United States, where he did most of the theoretical and empirical work on *Die räumliche Ordnung der Wirtschaft* (1939a), published in the United States as the *Economics of Location* in 1954. His Habilitation (that is, his qualification to teach at a university) on population waves and business cycles was accepted but its unpopular conclusions and his known anti-Nazi views prevented him from getting the *venia legendi*, the actual permission to teach. He found refuge with the Kiel Institut für Weltwirtschaft, where he became chief of his own research group while at the same time suffering from political interference. He wrote a number of reports for the institute, one of which was published with his conclusions reversed. He kept his personal integrity at great personal cost. He died on 30 May 1945 in Ratzeburg (Holstein) of scarlet fever, which his weakened condition could not tolerate. In 1971, the City of Heidenheim honoured his memory by sponsoring biennial international conferences on location problems, establishing a prize for the best theses in the field and, a few years later, a special honour for older scholars in the field.

Although Lösch's first published paper dealt with the transfer problem, and he continued to be interested in international monetary problems, his only other publications in that field are two discussions of the transfer problem and an extensive fragment in the posthumously published ‘Theory of Foreign Exchanges’. The two major subjects of his published work were the relation of population and business cycles and, of course, his highly original *Räumliche Ordnung der Wirtschaft*.

The discussions of population problems anticipate many later developments. Waves of population increase were neither sufficient nor necessary for the explanation of business cycles. With detailed statistics, some going back to the 17th century, Lösch showed that any relation went from business

cycles to population waves, much as recent theory suggests. Though Lösch can claim priority there is no evidence that he actually influenced later developments.

The investigations about a declining and ageing population resulted, however, in quite different conclusions from what was then either politically or academically acceptable. The ageing of the population (the German '*Vergreisung*' has sinister overtones absent from the English equivalent) had its economic compensations. It allowed the better training of the younger generation and increased capital accumulation and productivity. Even in military terms, fewer but better trained and better equipped people were preferable to more but less skilled individuals. In short, fewer young people allowed greater savings and investments leading to increased productivity and growth. This differed substantially from the then prevalent secular stagnation thesis and is much more in keeping with the warnings of present-day development economists of the dangers of rapid population growth. Lösch's earlier *Was ist vom Geburtenrückgang zu halten*? (1932) was later put on the index by the Nazis and his doctoral thesis on the same topic was effectively suppressed.

Lösch's greatest contribution dealt, in most general terms, with general equilibrium theory applied to space. Distance itself becomes the central phenomenon. Lösch's intellectual predecessors dealt with this problem essentially in two ways. They either solved a partial equilibrium system (Alfred Weber) or they substituted a series of smaller regions for one large one (Ohlin).

Going from partial to general equilibrium, and investigating the structure of the region instead of taking it as given, involved the substitution of a very general set of assumptions for the usual *ceteris paribus* assumptions made. In Weber (and practically everyone else) the locations of markets, raw materials and populations are assumed. In Lösch the basic assumption is a perfectly even distribution of population and of all raw materials. With these extraordinarily general and brilliantly unrealistic assumptions Lösch succeeds in showing that competitive forces alone will establish a system of locations which, in turn, can be understood either as agglomerations of productions or the intersection of fewer or more crossroads, all being simultaneously determined.

Lösch presents a Walrasian model with distance built in as a system of coordinates of location. His most famous contribution, however, is the analysis of the *structure* of an economic landscape on the basis of the simple generalized assumptions mentioned. The empirical work related mostly to the American Mid-West, where the assumptions are approximately realistic. One test of the genius of the model is that, unlike with most theoretical models, the introduction of more realistic assumptions simplifies rather than complicates the model.

In the 'ideal' Lösch landscape the basic unit is a hexagon. This follows from the condition that consumers are initially equidistant from each other, that each producer and consumer must lie within the market area of each good and that there must be no empty corners. Modifications introduced are rectangular areas on the model of, say, the layout of American counties; or the effect of different resource endowments of different areas; or of a border separating what might otherwise be one market area.

The work does not exhaust itself with equilibrium analysis or the structure of economic landscapes. There is a dynamic analytical and empirical study of how business cycles spread over the economic landscape or how transfers are made over and between areas through intra-regional adjustments in connected areas and from one sub-market to another. Thus the initial impact of a change in demand in one landscape capital might first be felt in the capital in the centre of another landscape and spread from there in declining ripples to the border. There is a study of how the Great Depression spread in time and

geographically through an area. The usual multiplier is supplemented by a spatial one.
The Lösch analyses the *Gestalt* of a region rather than defining it by such criteria as the immobility of factors of production between but not within regions: all factors are mobile at a cost which varies with distance, even land whose physical immobility is substituted for by changes in its utilization. The case of completely specific resources is investigated, though considered rare.
Lösch left a number of unfinished studies, and plans for many more. His is probably the most original book published on economics in the German language between the two world wars. Most scholars would consider themselves lucky if they had added a layer of bricks to an existing wall. Only few scholars can claim to have started a new wall, and even fewer to have started a new building. Lösch is one of those few scholars.

See Also

- location theory

Selected works

1930. Eine Auseinandersetzung über das Transfer Problem. *Schmollers Jahrbuch* 54, 1193–206.

1932. *Was ist vom Geburtenrückgang zu halten?* 2 vols, Heidenheim: privately published.

1936a. *Bevölkerungswellen und Wechsellagen*. Jena: Gustav Fischer.

1936b. Die Vergreisung wirtschaftlich gesehen. *Schmollers Jahrbuch* 60, 577–685.

1936–7. Population cycles as a cause of business cycles. *Quarterly Journal of Economics* 51, 649–62.

1938. The nature of economic regions. *Southern Economic Journal* 5(1), 71–8.

1939a. *Die räumliche Ordnung der Wirtschaft. Eine Untersuchung über Standort, Wirtschaftsgebiete und Internationalen Handel*. 2nd revised edn, 1944. 3rd edn (reprint of the 2nd edn), Jena: Gustav Fischer, 1962. 2nd edn trans. as *The Economics of Location*, New Haven: Yale University Press, 1954.

1939b. Eine neue Theorie des Internationalen Handel. *Weltwirtschaftliches Archiv* 50, 308–28. Trans. as ‘A new theory of international trade’, *International Economic Papers* No. 6, London: Macmillan, 1956.

1949. Theorie der Währung. Ein Fragment. *Weltwirtschaftliches Archiv* 62, 35–88.

Bibliography

Riegger, R., ed. 1971. *August Lösch in Memoriam*. Heidenheim: Verlag der Buchhandlung Meuer.

Contains eight contributions and a bibliography of 78 items, including literature about Lösch.

Valavanis, S. 1955. Lösch on location. *American Economic Review* 45, 637–44.

Zottmann, A. 1949. Dr. Habil. August Lösch, gestorben am 30. Mai 1945. *Weltwirtschaftliches Archiv* 62 (1), 28–31. Appended bibliography, 32–4.

How to cite this article

Stolper, Wolfgang F. "Lösch, August (1906–1945)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000151> doi:10.1057/9780230226203.0996

The New Palgrave Dictionary of Economics Online

Lowe, Adolph (1893 – 1995)

Edward J. Nell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic growth; freedom; German hyperinflation; instrumental analysis; Lowe, A.; planning; technical change

Article

Born on 4 March 1893 in Stuttgart, Adolph Lowe was educated at Berlin and Tübingen and received the Dr. Juris. from Tübingen in 1918. From 1919 to 1924 he was Section Head in the Ministries of Labour and Economics of the Weimar Republic, and was largely responsible for the practical planning and management of the currency reforms that brought the great hyperinflation to an end. From 1924 to 1926 he was Head of the International Division of the Federal Statistical Bureau, a politically sensitive post in the light of disputes over reparations payments. In 1926 he became Director of Research at the Institute of World Economics at the University of Kiel, where he established an important centre for research into business cycles and their control and regulation through planning. In 1931 he was appointed Professor of Political Economy at the University of Frankfurt, where he joined the leaders of a major renaissance in social and socialist thinking. But in March 1933 he became the first professor in the social sciences to be fired by Hitler. He moved immediately to England, where he held a post at Manchester until 1940, when he moved to the New School for Social Research in New York, where he was Professor of Economics, Director of Research at the Institute of World Affairs, and then Professor Emeritus, remaining active in the Department until his return to Germany, in March 1983, 50 years after his forced departure. In 1984 he was awarded the Dr. *honoris causa* by the University of Bremen.

His publications include ‘Wie Ist Konjunkturtheorie Überhaupt Moglich?’ (1926), *Economics and Sociology* (1935), *The Price of Liberty* (1937), ‘The Classical Theory of Economic Growth’ (1954), *On Economic Knowledge* (1965; 1977) and *The Path of Economic Growth* (1976). *Economic Means and Social Ends*, edited by Robert L. Heilbroner, was published in 1969 in honour of Professor Lowe's 75th birthday.

Unlike many economists, Lowe considered economics inseparable from social inquiry in general. In his view, the central question of economics is the determination of the path of economic growth and its

relation to technical progress and social change. Lowe developed a strikingly simple three-sector model in which structural changes during expansion could be displayed. Growth will normally not take place in a balanced manner; more commonly the actual path will be a ‘traverse’ from one desired path to another, which is likely to shift again before it is reached. But the problem has to be understood in the light of what Lowe calls ‘instrumental analysis’. Conventional economic theory begins with knowledge of the prevailing situation and a set of well-defined behavioural laws, based on maximizing. From these two givens one can deduce/predict the future configuration of the economy. This approach worked well in the early stages of capitalism, when the pressure of poverty on labour and competition on capital ensured stable patterns of behaviour. But mass production and economies of scale undermine competition, while affluence and unionization, together with the growth of the middle class, lead both to unpredictable wage bargaining and to unstable consumer spending. Tastes become volatile, while consumption can be postponed or redirected, and businesses plan strategically, often in cooperation with their rivals, instead of maximizing on a short horizon – so the traditional approach is no longer appropriate. The historical conditions do not constrain behaviour sufficiently for maximizing models, even complex ones, to picture it accurately, so that the conventional method must be set aside. (Which means, as well, that the forces of the market cannot be relied upon; they are no longer determinate.) Instead, the givens should be the existing conditions and the *desired terminal position*, and the job of economic analysis then becomes to find the ‘goal-adequate’ sequences of change, together with the stimuli and/or constraints that will create the necessary behaviour patterns. Such stimuli and constraints must be imposed by government. Economic analysis becomes a form of planning, and Lowe's work in his last years analysed the relation of planning to freedom.

Selected works

1926. Wie ist Konjunkturtheorie überhaupt möglich? *Weltwirtschaftliches Archiv* 24(2), 165–97.

1935. *Economies and Sociology: A Plea for Co-operation in the Social Sciences*. London: G. Allen & Unwin.

1937. *The Price of Liberty: A German on Contemporary Britain*. London: L. and Virginia Woolf at the Hogarth Press.

1954. The classical theory of economic growth. *Social Research* 21(July), 127–58.

1965. *On Economic Knowledge*. New York: Harpers; London: Longmans. Enlarged edn, New York and London: M.E. Sharpe, 1977.

1969. *Economic Means and Social Ends: Essays in Politics Economics*, ed. R. L. Heilbroner. Englewood Cliffs, NJ: Prentice-Hall.

1976. *The Path of Economic Growth*. Cambridge: Cambridge University Press.

1988. *Has Freedom a Future?* New York: Praeger.

How to cite this article

Nell, Edward J. "Lowe, Adolph (1893–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000157> doi:10.1057/9780230226203.0997

The New Palgrave Dictionary of Economics Online

low-income housing policy

Edgar O. Olsen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Low-income housing assistance is an important part of the welfare system in many countries. This article discusses the rationale for this government activity, describes the most important differences between different low-income housing programmes, explains why economic theory has limited implications for the effects of these programmes, and summarizes the evidence on their most important effects. The most important finding of the empirical literature on the effects of different housing programmes from the viewpoint of housing policy is that recipient-based housing assistance has provided equally good housing at a much lower total cost than any type of unit-based assistance.

Keywords

crowding out; low-income housing policy; neighbourhood effects; public housing; recipient-based housing assistance; residential segregation; unit-based housing assistance

Article

Low-income housing assistance is an important part of the welfare system in many countries.

Rationales

The most compelling rationale for this government activity is that some taxpayers care about low-income households and think that the decision makers in some of these households spend too little of their income on housing for their own good. Another important argument is that some taxpayers are particularly concerned about the well-being of the children in low-income households and prefer housing subsidies to unrestricted cash grants in order to better target assistance to the objects of their concern. These rationales imply that a successful housing programme induces its recipients to occupy better housing and consume less of other goods than they would choose in response to an unrestricted

cash grant in an amount equal to the housing subsidy.

Programme types

Governments have tried many methods of providing housing assistance. The most important distinction between rental housing programmes is whether the subsidy is attached to the dwelling unit or to the assisted household. If the subsidy is attached to a rental dwelling unit, each family must accept the particular unit offered in order to receive assistance and loses its subsidy when it moves. Each family offered recipient-based rental assistance has a choice among many units in the private market that meet the programme's standards, and the family can retain its subsidy when it moves. The analogous distinction for homeownership programmes is between programmes that require eligible families to buy from selected sellers in order to receive a subsidy and programmes that provide subsidies to eligible families that are free to buy from any seller that provides housing meeting the programme's standards. There are two broad types of unit-based rental assistance, namely, public housing and privately owned subsidized projects. Public housing projects are owned and operated by government entities. In public housing programmes, civil servants make all of the decisions made by private owners of unsubsidized housing. Governments also contract with private parties to provide unit-based assistance in subsidized housing projects. In the United States, the majority of these private parties are for-profit firms, but non-profit organizations have a significant presence. Under most programmes, these private parties agree to provide rental housing meeting certain standards at restricted rents to households with particular characteristics for a specified number of years. The overwhelming majority of the projects were newly built under a subsidized construction programme. Almost all of the rest were substantially rehabilitated as a condition for participation in the programme. None of the programmes that subsidize privately owned projects provide subsidies to all suppliers who would like to participate.
In 2004, the United States government spent about $15 billion on its housing voucher programme, more than $15 billion to subsidize private projects for low-income households, and about $7.5 billion to subsidize public housing projects. The US Department of Housing and Urban Development's Section 8 New Construction and Substantial Rehabilitation Program and the Internal Revenue Service's Low-Income Housing Tax Credit Program are the two largest programmes that subsidize private rental projects, accounting for about 75 per cent of public expenditure on programmes of this type. In total, these rental programmes served about seven million households. During the same year, the US government spent only $4 billion to subsidize low-income homeowners. These programmes tend to provide shallower subsidies to households with substantially higher incomes than the rental programmes.

Theory

Economic theory that accounts for the most rudimentary features of real housing programmes does not have strong implications about their effects. For example, these programmes may induce households to occupy worse housing even if housing is a normal good. Such counterintuitive outcomes result from the nonlinear budget frontiers facing households offered housing assistance. For instance, a household offered a unit in a subsidized housing project is offered an all-or-nothing choice of a particular dwelling unit at a below-market rent. This unit might be worse than the household's current unit, but the household may accept the offer because the reduction in its rent enables it to consume more of other

goods.

Evidence

The remainder of this article summarizes the evidence on the effects of the major rental housing programmes in the United States. The United States has rental programmes of each broad type, and a disproportionate share of the evidence on the performance of low-income housing programmes throughout the world pertains to these programmes. Homeownership programmes are a small part of the current system, and little is known about their effects.

Different rental housing programmes have different effects. Indeed, the same programme has different effects in different circumstances. Olsen (2003) provides a more detailed account of the evidence on the performance of individual programmes, and the bibliography to this article contains references to some of the more important recent studies. This article endeavours to characterize what is typical of these programmes and the differences in the average effect of programmes of different types.

The most important finding of the empirical literature on the effects of different housing programmes from the viewpoint of housing policy is that recipient-based housing assistance has provided equally good housing at a much lower total cost than any type of unit-based assistance. The reasons for this result suggest that it would apply generally. These reasons include the absence of a financial incentive for good decisions on the part of civil servants who operate public housing, the excessive profits that inevitably result from allocating subsidies to selected developers of private subsidized projects, and the distortions in usage of inputs resulting from the subsidy formulas. Another reason for the excess cost of unit-based assistance is that this assistance is usually tied to the construction of new units. The least expensive approach to improving the housing conditions of low-income households involves heavy reliance on upgrading the existing housing stock.

Since housing programmes are intended to produce particular changes in consumption of housing services compared with consumption of other goods, knowledge of these changes is important for evaluating these programmes. The overwhelming majority of recipients of housing assistance occupy better housing than they would occupy in the absence of assistance. More importantly, they typically occupy better housing than they would occupy if they were given cash grants in amounts equal to their housing subsidies. Most recipients of rental housing assistance pay significantly less for their housing and hence have more to spend on other goods.

One aspect of the housing bundle broadly conceived that has attracted considerable attention is its neighbourhood. Recipients of tenant-based vouchers and occupants of privately owned subsidized projects typically live in somewhat better and less racially segregated neighbourhoods than in the absence of housing assistance. Occupants of public housing typically live in noticeably worse and more racially segregated neighbourhoods.

A careful theoretical analysis that accounts for a key feature of low-income housing programmes has shown that, even if the subsidy under the programme declines with increases in earnings and leisure is a normal good, the programme will not necessarily induce the recipient to work less (Schone, 1992). Nevertheless, evidence based on a controlled experiment indicates that voucher recipients reduce their earnings about 13 per cent on average (Patterson et al., 2004). Other evidence indicates that programmes of unit-based assistance have somewhat larger work disincentive effects (Olsen et al., 2005).

Low-income housing programmes differ substantially from unrestricted cash grants in their effects. The mean value of project-based housing assistance as judged by recipients is much less than 75 per cent of the mean housing subsidy (that is, the difference between the market rent of the subsidized unit and the tenant's contribution). The mean value of tenant-based housing assistance as judged by recipients is about 80 per cent of the mean housing subsidy.

Consistent with their intentions, the mean benefit to recipients in these programmes is greater for poorer and larger households among households that are the same in other respects. Mean benefit varies little with the age, race and sex of the head of the household after other household characteristics are accounted for. The variance in benefit among recipients with the same characteristics is large under construction programmes that have produced new units for many years. In these mature construction programmes, there is an enormous difference between the best and the worst units, and a tenant with specified characteristics would pay the same rent for these units.

Unit-based or recipient-based housing programmes can make the neighbourhoods into which subsidized households move better or worse places to live. Neighbourhood property values capture these effects. On average across all units in a programme, the evidence indicates that no programme has had a significant effect on neighbourhood property values.

Housing programmes affect the rents of unsubsidized units with unchanging characteristics. Evidence from the Housing Assistance Supply Experiment indicates that an entitlement housing voucher programme for which the poorest 20 per cent of the population is eligible will have small effects on market rents (Lowry, 1983). No evidence is available for construction programmes. However, economic theory suggests that, if a construction programme leads to a larger housing stock, it will result in higher market rents because it will drive up the prices of inputs used heavily in the housing industry. This effect might be small, however, because the evidence indicates that subsidized construction crowds out unsubsidized construction to a considerable extent (Malpezzi and Vandell, 2002; Sinai and Waldfogel, 2005; and references in Olsen, 2003).

An important recent literature estimates a wide range of impacts of offering portable vouchers to families living in the worst public housing projects or in public housing projects in the poorest neighbourhoods. The larger strand of this research is based on data from a controlled experiment called Moving to Opportunity, in which one experimental group was offered a housing voucher without any restriction on the neighbourhood where it could be used and another experimental group had to move for at least a year to a neighbourhood where the poverty rate was less than ten per cent prior to the experiment (Orr et al., 2003). These treatments led their recipients to live in better housing and neighbourhoods without a reduction in expenditure on other goods. However, they did not lead to some expected outcomes. After four to seven years in the experiment, the treatment groups did not increase their earnings and their children's educational performance did not improve. With a few notable exceptions such as the mental health of girls and their mothers, the treatments had minimal effects on health outcomes. The treatments generally had effects in opposite directions on the delinquency and risky behaviour of boys and girls. The effects on boys were negative, though these effects were not usually statistically significant. A smaller strand of this literature is based on data on natural experiments such as when public housing tenants must move because their project is torn down (Jacob, 2004).

See Also

- crowding out
- housing policy in the United States
- housing supply
- welfare state

Bibliography

Jacob, B. 2004. Public housing, housing vouchers, and student achievement: evidence from public housing demolitions in Chicago. *American Economic Review* 94, 233–58.

Lowry, I.S., ed. 1983. *Experimenting with Housing Allowances: The Final Report of the Housing Assistance Supply Experiment*. Cambridge, MA: Oelgeschlager, Gunn & Hain.

Malpezzi, S. and Vandell, K. 2002. Does the low-income housing tax credit increase the supply of housing? *Journal of Housing Economics* 11, 360–80.

Olsen, E.O. 2003. Housing programs for low-income households. In *Means-Tested Transfer Programs in the United States*, ed. R. Moffitt. Chicago: University of Chicago Press.

Olsen, E.O., Tyler, C.A., King, J.W. and Carrillo, P.E. 2005. The effects of different types of housing assistance on earnings and employment. *Cityscape* 8, 163–87.

Orr, L. et al. 2003. *Moving to Opportunity for Fair Housing Demonstration Program: Interim Impacts Evaluation*. Washington, DC: US Department of Housing and Urban Development.

Patterson, R. et al. 2004. *Evaluation of the Welfare to Work Voucher Program: Report to Congress*. Washington, DC: US Department of Housing and Urban Development.

Schone, B.S. 1992. Do means tested transfers reduce labor supply? *Economics Letters* 40, 353–8.

Sinai, T. and Waldfogel, J. 2005. Do low-income housing subsidies increase the occupied housing stock? *Journal of Public Economics* 11–12, 2137–64.

How to cite this article

Olsen, Edgar O. "low-income housing policy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000236> doi:10.1057/9780230226203.0998

The New Palgrave Dictionary of Economics Online

Lucas critique

Lars Ljungqvist

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The ‘Lucas critique’ is a criticism of econometric policy evaluation procedures that fail to recognize that optimal decision rules of economic agents vary systematically with changes in policy. In particular, it criticizes using estimated statistical relationships from past data to forecast the effects of adopting a new policy, because the estimated regression coefficients are not invariant but will change along with agents’ decision rules in response to a new policy. A classic example of this fallacy was the erroneous inference that a regression of inflation on unemployment (the Phillips curve) represented a structural trade-off for policy to exploit.

Keywords

expectations; Lucas Critique; macroeconomic policy evaluation; optimization behaviour; Phillips curve; rational expectations; rational expectations econometrics; real vs. nominal shocks

Article

The ‘Lucas Critique’ is a criticism of econometric policy evaluation procedures that fail to recognize the following economic logic:

> [G]iven that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any changes in policy will systematically alter the structure of econometric models. (Lucas, 1976, p. 41)

At the time of his writing, Robert E. Lucas, Jr. (1976) was criticizing the prevailing approach to quantitative macroeconomic policy evaluation for ignoring this logic and, hence, as being fundamentally inconsistent with economic theory. To fully appreciate Lucas's critique, we first consider a general

theoretical argument and then turn to a particular example.

At each date t there is a vector s_t of state variables summarizing all aspects of the history that are relevant to the economy's future evolution; for example, the vector might include the economy's capital stock. The economy is also described by a vector x_t of government policy variables and a vector ε_t of random shocks – for example, shocks to technology or to government policy. For given specifications of the processes governing x_t and ε_t, it is common in macroeconomic theory to analyse models that yield an equilibrium law of motion in form of a difference equation,

$$s_{t+1} = f(s_t, x_t, \varepsilon_t). \tag{1}$$

(For many textbook examples of stochastic rational expectations models that yield such a recursive equilibrium representation; see Ljungqvist and Sargent, 2004.) Equation (1) is also the point of departure for the econometric policy evaluation procedures criticized by Lucas, who argued that their approach failed to recognize the optimization behaviour of economic agents that is implicit in eq. (1). Specifically, the criticized approach proceeds as follows. First, historical data are used to estimate the equation

$$s_{t+1} = F(\theta, s_t, x_t, \mu_t), \tag{2}$$

where F is specified in advance, θ is a fixed parameter vector to be estimated, and μ_t is a vector of random disturbances. Second, with the use of the estimated eq. (2), policy evaluations are performed by comparing economic outcomes for different paths of government policy variables $\{x_t\}$. The policy choice that produces the most desirable economic outcome is deemed to be the best policy. But, as argued by Lucas, this approach violates the premises for economic theory because the parameter vector θ depends partly on agents' decision rules that are not invariant to the conduct of government policy. That is, if the government changes its policy, the parameter θ will also change, so that the consequences of a new policy cannot be evaluated on the basis of the historical relationship in eq. (2).

Lucas's argument is best illustrated with an example. Consider the classic example of the so-called 'Phillips curve'. Phillips (1958) had estimated a negative relationship between wage inflation and unemployment using British data for the period 1861–1957. Samuelson and Solow (1960) and others interpreted this and related empirical findings as evidence of a structural trade-off between an economy's inflation rate and its unemployment rate. That is, the parameter θ in eq. (2), estimated with historical data, was considered to be fixed and to describe how unemployment would respond to inflation outcomes associated with different monetary policies. Friedman (1968) and Phelps (1968) argued

against the existence of such an exploitable trade-off because it was inconsistent with economic theory based on rational agents. To understand the fallacy of the Phillips curve and its extension – the fallacy of the econometric policy evaluation procedures criticized by Lucas – consider the monetary model of Lucas (1972). Exchange in the economy takes place in physically separated markets. Producers in a market base their output decisions on the local market-clearing price level without knowing the current economy-wide price level. The price in a market varies stochastically because there are exogenous random shocks both to the distribution of producers across markets and to the aggregate quantity of nominal money, none of which is directly observable to the agents. Hence, information on the current state of these real and monetary shocks is transmitted to agents only through the price in the market where each agent happens to be. In an equilibrium, producers in a market would like to increase their output in response to a high price driven by real but not nominal shocks. A high price due to a real shock means that the ratio of producers to consumers is low in that market and, therefore, profits on sales are high in real terms (when evaluated in terms of the economy-wide price level). But a high price in a market due to an expansion of the aggregate quantity of nominal money means that prices tend to be high in all markets and, therefore, profits on sales are high in nominal but not real terms. The inference and decision problems solved by the agents in this model are shown to give rise to a Phillips curve, as had been estimated with real-world data, but where the model's apparent trade-off between inflation and output cannot be systematically exploited by the government in its choice of monetary policy.

To further convey the insights from this general equilibrium model of the Phillips curve, we adopt a version of Lucas's (1976) simplified model that does not spell out all the details of the economic environment but instead postulates three equations that capture the forces at work in the fully articulated model. The economy-wide price level (in logs), p_t, is given by

$$p_t = \bar{p}_t + m_t, \tag{3}$$

where $\bar{p}_t$ reflects a systematic component of monetary policy that is known to all agents, and m_t reflects an i.i.d. shock to monetary policy. It is assumed that the random variable m_t is normally distributed with mean zero and variance σ_m^2. The price (in logs) in market i at time t, p_{it}, is given by

$$p_{it} = p_t + z_{it}, \tag{4}$$

where z_{it} is a deviation from the economy-wide price level because of shocks to the distribution of producers across markets. The real shock z_{it} is assumed to be a normal, i.i.d. random variable with mean

zero and variance σ_z^2. Finally, let y_{it} denote the log-deviation of output from its 'natural rate' in market i at time t which varies with the perceived, relative price:

$$y_{it} = \alpha [p_{it} - E(p_t | I_{it})], \tag{5}$$

where $\alpha > 0$ reflects intertemporal substitution possibilities in supply (determined by technological factors and tastes for substituting labour over time), and $E(\cdot|I_{it})$ denotes the mathematical expectation conditioned upon information I_{it} available in market i at time t. The agents' prediction problem in eq. (5) is straightforward to solve (see, for example, Ljungqvist and Sargent, 2004, ch. 5):

$$E(p_t | I_{it}) = E(p_t | p_{it}, \bar{p}_t) = (1 - \Omega) p_{it} + \Omega \bar{p}_t, \tag{6}$$

where $\Omega = \sigma_z^2 / (\sigma_m^2 + \sigma_z^2)$. The substitution of eqs. (3), (4) and (6) into eq. (5) yields

$$y_{it} = \alpha\Omega(m_t + z_{it}). \tag{7}$$

Thus, output in market i varies with the sum of nominal and real shocks, $(m_t + z_{it})$, because producers cannot perfectly disentangle these shocks but must make inferences based on the observed price p_{it}. Producers' willingness to vary output from its natural rate depends on how likely observed price variations are due to real rather than nominal shocks, as captured by the magnitude of $\Omega \in [0, 1]$. Under the assumption of a large number N of markets, the real shocks, $\{z_{it}\}$, cancel each other out when averaged over markets, and the economy's deviation from its natural rate of output, y_t, becomes

$$y_t = \frac{1}{N}\sum_{i=1}^{N} y_{it} = \alpha\Omega m_t = \alpha\Omega(p_t - \bar{p}_t), \tag{8}$$

where the last equality invokes eq. (3) and, hence, the economy exhibits a positive relationship between unanticipated inflation and output.

If estimations were performed using data on output and inflation from the described economy, we would find a Phillips curve along which increases in inflation are associated with higher output realizations. However, any attempts by the government to exploit that relationship would fail. For example, a government that permanently increases the growth rate of the money supply to generate higher inflation in order to stimulate output will ultimately see no real effects from that change in policy. The reason for this is that, after agents have become aware of the higher underlying inflation rate in the economy, they will change their expectations when making predictions about relative price movements due to real disturbances. Formally, the change in monetary policy represents an increase in the component $\bar{p}_t$ and, when that systematic change becomes known to the agents, it will not affect unanticipated inflation, $(p_t - \bar{p}_t) = m_t$, so output is left unaffected in eq. (8).

This example illustrates Lucas's general criticism of econometric policy evaluation procedures that fail to recognize that the estimated eq. (2) depends partly on agents' decision rules and is therefore not invariant to changes in government policy. For a proper policy evaluation procedure, we need to revise the econometric formulation in eq. (2) so that it becomes consistent with equilibrium outcomes as represented by eq. (1). Recall that the latter equation is derived for given specifications of the processes governing x_t and ε_t. In particular, to analyse agents' optimization behaviour, we need to specify the environment in which they live, including their perceptions about future government policy. As Lucas (1976, p. 40) remarked, 'one cannot meaningfully discuss optimal decisions of agents under arbitrary sequences $\{x_t\}$ of future shocks'. Instead, Lucas suggested that one proceeds by viewing government policy as a function of the state of the economy,

$$x_t = G(\lambda, s_t, \eta_t), \tag{9}$$

where λ is a parameter vector that characterizes government policy, and η_t is a vector of random disturbances. Then the new version of eq. (2) becomes

$$s_{t+1} = F(\theta(\lambda), s_t, x_t, \mu_t), \tag{10}$$

and the econometric problem is that of estimating the function $\theta(\lambda)$. A change in government policy is viewed as a change in the parameter λ affecting the behaviour of the system in two ways: first, by altering the time series behaviour of $\{x_t\}$, and second, by leading to modification of the parameter θ

governing the rest of the system, which reflects changes in agents' decision rules in response to the new policy.

A constructive response to the Lucas critique has been the development of rational expectations econometrics. A goal of that approach has been to estimate the 'primitives' of dynamic rational expectations models, in the form of parameters describing tastes and technologies. If historical data can be used to obtain such estimates, the economic model can in principle be used to evaluate alternative government policies that could be without precedent, as explained by Lucas and Sargent (1981). That is, knowledge about the primitives of a model enables us to derive agents' decision rules and equilibrium outcomes for any specified policy process. In terms of eq. (10), this explains how the function $\theta(\lambda)$ could conceivably be estimated even if the historical data have been generated under a single government policy λ.

Though one of the key contributors to the methodology of rational expectations econometrics, Sargent (1984) has raised a philosophical conundrum with this approach to policy evaluation (as earlier discussed by Sargent and Wallace, 1976). Suppose that the primitives of an economic model have been estimated during an estimation period in which government policy was specified to be λ, and then the estimated model is used to compare alternative policies in order to find the best future policy λ^*. But such a procedure leads to an internal contradiction under the assumption of rational expectations, because, if the procedure were in fact likely to be persuasive in having the policy recommendation actually adopted soon, it would mean that the original econometric model with it specified policy λ had been mis-specified. As pointed out by Sargent (1984, p. 413): 'A rational expectations model during the estimation period ought to reflect the procedure by which policy is thought later to be influenced, for agents are posited to be speculating about government decisions into the indefinite future.'

Given its fundamental impact on questions of economic policy both in practice and in theory, the Lucas critique figured prominently in the list of contributions when the Royal Swedish Academy of Sciences (1995) awarded Robert E. Lucas, Jr. the Nobel Prize in economics 'for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy.'

See Also

- Phillips curve
- rational expectations

Bibliography

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Ljungqvist, L. and Sargent, T.J. 2004. *Recursive Macroeconomic Theory*, 2nd edn. Cambridge, MA: MIT Press.

Lucas, R.E., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R.E., Jr. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and the Labor Market*, ed. K. Brunner and A. Meltzer, Vol. 1 of Carnegie-Rochester Conference on Public Policy. Amsterdam, North-Holland.

Lucas, R.E., Jr. and Sargent, T.J. 1981. *Rational Expectations and Econometric Practice*. Minneapolis: University of Minnesota Press.

Phelps, E.S. 1968. Money wage dynamics and labor market equilibrium. *Journal of Political Economy* 76, 687–711.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Econometrica* 25, 283–99.

Royal Swedish Academy of Sciences 1995. The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for 1995. Press release, 10 October. Online. Available at http://nobelprize.org/economics/laureates/1995/press.html, accessed 4 October 2006.

Samuelson, P.A. and Solow, R.M. 1960. Analytical aspects of anti-inflation policy. *American Economic Review* 50, 177–94.

Sargent, T.J. 1984. Autoregressions, expectations, and advice. *American Economic Review* 74, 408–15.

Sargent, T.J. and Wallace, N. 1976. Rational expectations and the theory of economic policy. *Journal of Monetary Economics* 2, 169–83.

How to cite this article

Ljungqvist, Lars. "Lucas critique." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000159> doi:10.1057/9780230226203.1000

The New Palgrave Dictionary of Economics Online

Lucas, Robert (born 1937)

Levon Barseghyan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robert E. Lucas, Jr is one of the most influential economists of our time. His work on rational expectations offered a truly new way of thinking about economics and policy that led to most of the recent successes in macroeconomics. Lucas's path breaking research on so many issues of vital importance has advanced the frontier of science and set the stage for new exciting discoveries.

Keywords

business cycles; cash-in-advance constraint; dynamic stochastic general equilibrium models; economic growth; endogenous growth theory; human capital; inflation; Lucas Critique; Lucas trees; Lucas, R.; monetary business cycle models; monetary shocks; money growth; neoclassical growth theory; neutrality of money; overlapping generations models; Phillips curve; rational expectations; time consistency of monetary and fiscal policy

Article

In 1995, Robert E. Lucas, Jr received the Nobel Prize in Economic Sciences ‘for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy’ (Press Release announcing the Nobel Prize, 1995; repr. in Svensson, 1996, p. 1).

Robert Lucas was born in Yakima, Washington on 15 September 1937. He received his BA in History in 1959, and his Ph.D. in Economics in 1964, both from the University of Chicago. He began his career as an assistant professor at Carnegie Mellon University, where he became an associate professor in 1967 and a full professor in 1970. He joined the Department of Economics at the University of Chicago as a full professor in 1975, and since 1980 has served as John Dewey Distinguished Service Professor of Economics there. He is a fellow of the Econometric Society, the American Academy of Arts and Sciences, and the American Finance Association; a member of the National Academy of Sciences and

the American Philosophical Society and a titular member of the European Academy of Arts, Sciences and Humanities. Lucas served as the President of the Econometric Society in 1997 and as the President of the American Economic Association in 2002.

Robert Lucas's seminal contributions in the early 1970s led to a paradigm shift in macroeconomics: the rational expectations revolution. By the late 1970s–early 1980s, due to the efforts of Robert Lucas and others (including Robert Barro, William Brock, Edward Prescott, Thomas Sargent and Neil Wallace) the frontier of macroeconomic research had moved away from models with static or adaptive expectations towards models in which agents act in their best interest, utilizing all available information about past, present and future. As a result, dynamic stochastic general equilibrium models with rigorous microfoundations have been developed to understand economic fluctuations and growth and to analyse the effects of monetary and fiscal policies. While these models have become increasingly complex in an effort to better understand the economy, almost all of them are built on the principles set forth by Robert Lucas.

The beginning of the rational expectations revolution: expectations and the neutrality of money

Robert Lucas's ‘Expectations and Neutrality of Money’, published in 1972 in the *Journal of Economic Theory*, was the first paper to incorporate the idea of rational expectations into a dynamic general equilibrium model. (rational expectations were introduced by Muth, 1961. In their ground-breaking study of investment under uncertainty, Lucas and Prescott, 1971 applied the notion of rational expectations in a dynamic partial equilibrium model of a competitive industry facing stochastic demand.) The agents in Lucas's (1972b) model are fully rational: based on the available information, they form expectations about future prices and quantities, and based on these expectations they act to maximize their expected lifetime utility. This paper also was the first to provide sound theoretical underpinnings to Milton Friedman's (1968) and Edmund Phelps's (1968) view of the long-run neutrality of money, and at the same time to provide an explanation of the observed positive correlation between output and inflation, famously depicted by the Phillips curve.

Lucas's model is built on Paul Samuelson's (1958) overlapping generations model. Agents live for two periods. In each period the young generation works, consumes and saves. The old generation consumes its savings. Goods are perishable and there is only one savings instrument in the economy, money.

The population in the economy is allocated into two distinct markets (islands) across which no communication is possible. The old generation is equally divided between the islands. The allocation of the young generation across the islands is a random variable. The amount of money holdings by the old generation is also a random variable, because it depends on the realization of a random shock in the money growth rate: each dollar carried from one period to another is multiplied by the realized money growth rate between these two periods. Agents do not observe the current allocation of young across islands and the money growth rate, but know their underlying probability distributions. To solve for the optimal amount of labour supply and savings, the young must form expectations about the future value of money, that is, the future price level. How does one form such expectations? Lucas's answer is, rationally. He defined and explicitly solved for the rational expectations equilibrium, in which agents correctly predict how the price level depends on the state of the economy. Of course, to do so each agent also must correctly understand the actions of all other agents in the current and future generations and

how these actions affect prices (and quantities).

In the model, the positive correlation between the money growth rate and output arises because the young, when faced with a high demand for their goods, are unable to distinguish its source: the demand could be high because of a higher money growth rate, or because of a lower fraction of the young workers on the island. Due to their inability to infer exactly the source of the high demand, the young find it optimal to produce whenever they face a high demand. Consequently, a positive money growth shock leads to an economic expansion on both islands. Without uncertainty about the money growth rate, the neutrality of money is immediately attained. Any pre-announced proportional money growth rule – for example, the *k%* rule advocated by Milton Friedman – results in the same real outcomes. Lucas showed that invariance of real outcomes to the pre-announced part of the money growth rule holds also when there are shocks to money growth. This finding is often characterized as a ‘policy ineffectiveness’ result, because it implies that, although there is a positive correlation between output and money growth, this correlation cannot be exploited by the monetary authority to influence real economic activity.

Prior to Lucas's (1972b) work, economists often emphasized that a distinction should be drawn between the long-run and the short-run effects of monetary shocks. An important corollary of Lucas's work is that this distinction often is misleading. The true distinction must be made between anticipated and unanticipated monetary disturbances, because their effects on real economic activity are likely to be very different. Most of the subsequent monetary business cycle literature embraces this distinction.

Econometric policy evaluation: the Lucas Critique

Lucas (1976), known as the Lucas critique, marked the turning point in how economists approached econometric policy evaluation. Thomas Sargent's (1996) account of the events following the Lucas critique gives a sense of its tremendous impact:

> [W]e didn't understand what was going on until, upon reading Lucas's ‘Econometric Policy Evaluation’ in Spring of 1973, we were stunned into terminating our long standing Minneapolis Fed research project to design, estimate and optimally control a Keynesian macroeconometric model. We realized that Kareken, Muench, and Wallace's (1973) defense of the ‘look-at-everything’ feedback rule for policy – which was thoroughly based on ‘best responses’ for the monetary authority exploiting a ‘no response’ private sector – could not be the foundation of a sensible research program, but was better viewed as a memorial plaque to the Keynesian tradition in which we had been trained to work. (Sargent, 1996, p. 539)

The essence of the Lucas critique stems naturally from the concept of rational expectations. Indeed, rationality of the private sector implies that it cannot be modelled as a ‘no response’ entity. Rather, any observed or anticipated change in monetary policy, including the ‘best response’ of the monetary authority, will induce the ‘best responses’ from the agents in the private sector. This, in turn, implies that the effects of a new policy cannot be assessed according to econometric estimation of the private sector's behaviour under the old policy.

Other major contributions

Robert Lucas has made several other major contributions in different areas of economics. A small subset of them is presented below, in chronological order.

Lucas (1978a) elegantly introduced the first general equilibrium model of asset pricing. In the model economy, physical assets are represented by what nowadays typically is referred to as ‘Lucas trees’: infinitely lived objects that generate stochastic dividends (fruits). Lucas explicitly derived asset prices as functions of the economy's state variables. The logic of Lucas's asset pricing equation forms the foundation of many models in macro and financial economics.

Lucas (1980b) and Lucas and Stokey (1987) helped to lay the foundations of monetary economics. The ideas and the methodology developed in these papers continue to guide monetary economists, particularly in applied research. Lucas (1980b) is the first general equilibrium study of the determination of prices in an economy in which the use of money arises from a cash-in-advance constraint. The model in Lucas and Stokey (1987), which is the prototype for a number of widely used dynamic stochastic general equilibrium monetary models, features both real and nominal shocks. Methods developed by Lucas and Stokey for establishing the existence of, characterizing and solving for the equilibrium of such models have proven to be powerful tools in applied and theoretical research.

Lucas (1982) extended the logic of his earlier contributions, Lucas (1978a) and Lucas (1980b), to a two-country stochastic general equilibrium model with infinitely lived agents, in which he explicitly derived formulas for pricing real assets and nominal bonds as well as for determining exchange rates. The framework developed in this paper serves as a point of departure for many models in international economics.

Lucas and Stokey (1983) is a major contribution to modern public finance. Lucas and Stokey studied the Ramsey (1927) problem – the problem of optimal taxation when non-distortionary tax instruments are unavailable – in dynamic stochastic economies without physical capital. Their paper provided a number of important insights about the structure and time consistency of optimal fiscal and monetary policies. Lucas and Stokey showed that a sufficiently rich debt maturity structure could allow for time consistency of the optimal fiscal policy.

Lucas (1988) is a seminal contribution in the economic development and growth literature (see also the 1991 Fisher and Shultz Lecture at the European Meetings of Econometric Society, published as Lucas, 1993). Lucas (1988) and an earlier paper by Paul Romer (1986) heralded the birth of endogenous growth theory and the resurgence of research on economic growth in the late 1980s and the 1990s. These papers offered an escape from ‘the straightjacket of the neoclassical growth model, in which the long term per capita (output) growth is pegged by the rate of exogenous technological progress’ (Barro and Sala-i-Martin, 2004, p. 19), by showing that factor accumulation does not need to run into diminishing returns to scale and, therefore, could lead to perpetual growth. In particular, Lucas emphasized the role of human capital, and externalities generated by it, as important sources of long-run economic growth.

Robert Lucas has written a number of seminal books. Among them are *Models of Business Cycles* (1987) and, with N. Stokey and E. Prescott, *Recursive Methods in Economic Dynamics* (1989). The former presents a critical assessment of the business cycle literature of the 1970s and the early 1980s and offers novel insights about economic fluctuations. This monograph contains Lucas's famous calculation of the cost of business cycles, which he argued to be insignificant. (In a similar spirit, Lucas, 2000a

provided a quantitative assessment of the welfare cost of inflation. In this paper, he found that the gains from reducing inflation could be non-negligible. Subsequent research often has taken his calculations of the cost of business cycles and of the cost of inflation as benchmarks.) Another indispensable volume, *Recursive Methods in Economic Dynamics* (1989), deals with stochastic dynamic programming. It has been widely used as a textbook in graduate macroeconomics courses and as a guide for formulating and solving dynamic stochastic general equilibrium models.

See Also

- Lucas critique
- monetary business cycles (imperfect information)
- neutrality of money
- Phillips curve
- rational expectations

In writing this article I have drawn from Fisher (*1996*), *Hall* (*1996*), *Svensson* (*1996*), *Sargent* (*1996*), *Lucas* (*1996*) *and Chari* (*1998*).

Selected works

1962. (With Z. Griliches, G.S. Maddala, and N. Wallace). Notes on estimated aggregate quarterly consumption functions. *Econometrica* 30, 491–500.

1967a. Optimal investment policy and the flexible accelerator. *International Economic Review* 8, 78–85.

1967b. Tests of a capital-theoretic model of technological change. *Review of Economic Studies* 34, 175–89.

1967c. Adjustment costs and the theory of supply. *Journal of Political Economy* 75, 321–34.

1968. (With T. McGuire, J. Farley and W. Ring.) Estimation and inference for linear models in which subsets of the dependent variable are constrained. *Journal of the American Statistical Association* 63, 1201–13.

1969a. (With L. Rapping.) Real wages, employment, and inflation. *Journal of Political Economy* 77, 721–54.

1969b. (With L. Rapping.) Price expectations and the Phillips curve. *American Economic Review* 59, 342–50.

1970a. Capacity, overtime and empirical production functions. *American Economic Review* 60, 23–7.

1970b. (With L.A. Rapping et al.). Real wages, employment and inflation. In *The New Microeconomics in Employment and Inflation Theory*, ed. E.S. Phelps et al. New York: W.W. Norton.

1971. (With E.C. Prescott.) Investment under uncertainty. *Econometrica* 39, 659–81.

1972a. (With E.C. Prescott.) A note on price systems in infinite dimensional space. *International Economic Review* 13, 416–22.

1972b. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

1972c. (With L. Rapping.) Unemployment in the great depression: is there a full explanation?*Journal of Political Economy* 80, 186–91.

1972d. Econometric testing of the natural rate hypothesis. In *The Econometrics of Price Determination Conference*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.

1973. Some international evidence on output-inflation trade-offs. *American Economic Review* 63, 326–34.

1974. (With E.C. Prescott.) Equilibrium search and unemployment. *Journal of Economic Theory* 7, 188–209.

1975. An equilibrium model of the business cycle. *Journal of Political Economy* 83, 1113–44.

1976. Econometric policy evaluation: a critique. *Carnegie-Rochester Conference Series on Public Policy*, vol. 1. First publ. in *The Phillips Curve and Labor Markets*, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland, 1975.

1977. Understanding business cycles. In *Stabilization of the Domestic and International Economy*, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

1978a. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

1978b. On the size distribution of business firms. *Bell Journal of Economics* 508–23.

1978c. Unemployment policy. *American Economic Review* 68, 353–57.

1979(With T.J. Sargent.) After Keynesian macroeconometrics. In *After the Phillips Curve*, Federal Reserve Bank of Boston, Conference Series No. 19: 49–72. Repr. in *Federal Reserve Bank of*

Minneapolis Quarterly Review 3, 1–6.

1980a. Rules, discretion and the role of the economic advisor. In *Rational Expectations and Economic Policy*, ed. S. Fischer. Chicago: University of Chicago Press for the NBER.

1980b. Equilibrium in a pure currency economy. In *Models of Monetary Economics*, ed. J.H. Karaken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

1980c. Two illustrations of the quantity theory of money. *American Economic Review* 1970, 1005–14.

1980d. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12, 696–717.

1981a(With T.J. Sargent.) *Rational Expectations and Econometric Practice*. Minneapolis: University of Minnesota Press.

1981b. *Studies in Business-Cycle Theory*. Cambridge, MA: MIT Press.

1981c. Distributed lags and optimal investment policy. In Lucas and Sargent (1981a).

1981d. Optimal investment with rational expectations. In Lucas and Sargent (1981a).

1982. Interest rates and currency prices in a two-country world. *Journal of Monetary Economics* 10, 335–60.

1983. (With N.L. Stokey.) Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12, 55–94.

1984a. (With N.L. Stokey.) Optimal growth with many consumers. *Journal of Economic Theory* 32, 139–71.

1984b. Money in a theory of finance. In *Essays on Macroeconomic Implications of Financial and Labor Markets and Political Processes*, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

1986a. Principles of fiscal and monetary policy. *Journal of Monetary Economics* 17, 117–34.

1986b. Adaptive behavior and economic theory. *Journal of Business* 59, S401–S426.

1987. (With N.L. Stokey.) Money and interest in a cash-in-advance economy. *Econometrica* 55, 491–514.

1987. *Models of Business Cycles*. Y. Jahnsson Lectures.Oxford: Basil Blackwell.

1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

1989. (With N.L. Stokey and E.C. Prescott.) *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

1990a. Liquidity and interest rates. *Journal of Economic Theory* 50, 237–64.

1990b. Why doesn't capital flow from rich to poor countries? *American Economic Review* 80, 92–6.

1990c. Supply side economics: an analytical review. *Oxford Economic Papers* 42, 293–316.

1992a. (With A.G. Atkeson.) On efficient distribution with private information. *Review of Economic Studies* 59, 427–53.

1992b. On efficiency and distribution. *Economic Journal* 102, 233–47.

1993. Making a miracle. *Econometrica* 61, 251–72.

1995. (With A.G. Atkeson.) Efficiency and equality in a simple model of efficient unemployment insurance. *Journal of Economic Theory* 66, 64–88.

1996. Nobel lecture: monetary neutrality. *Journal of Political Economy* 104, 661–82.

2000a. Inflation and welfare. *Econometrica* 68, 247–74.

2000b. Some macroeconomics for the 21st century. *Journal of Economic Perspectives* 14(1), 159–68.

2001a. *Lectures on Economic Growth*. Cambridge, MA: Harvard University Press.

2001b. Externalities and cities. *Review of Economic Dynamics* 4, 245–74.

2002. (With E. Rossi-Hansberg.) On the internal structure of cities. *Econometrica* 70, 1445–76.

2003. Macroeconomic priorities. *American Economic Review* 93, 1–14.

2004. Life earnings and rural-urban migration. *Journal of Political Economy* 112(S1), S29–S59.

2007. (With M. Golosov.) Menu costs and Phillips curves. *Journal of Political Economy* 115, 171–99.

2007. (With F. Alvarez.) General equilibrium analysis of the Eaton–Kortum model of international trade. *Journal of Monetary Economics*.

Bibliography

Barro, R. and Sala-i-Martin, X. 2004. *Economic Growth*, 2nd edn. Cambridge, MA: MIT Press.

Chari, V.V. 1998. Nobel Laureate Robert E. Lucas, Jr.: architect of modern macroeconomics. *Journal of Economic Perspectives* 12(1), 171–86 Repr. in *Federal Reserve Bank of Minneapolis Quarterly Review* 23, 2–12, 1999.

Fisher, S. 1996. Robert Lucas's Nobel memorial prize. *Scandinavian Journal of Economics* 98, 11–31.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–15.

Hall, R.E. 1996. Robert Lucas, recipient of the 1995 Nobel memorial prize in economics. *Scandinavian Journal of Economics* 98, 33–48.

Kareken, J.A., Muench, T. and Wallace, N. 1973. Optimal open market strategy: the use of information variables. *American Economic Review* 63, 156–72.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Phelps, E. 1968. Money-wage dynamics and labor market equilibrium. *Journal of Political Economy* 76, 687–711.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Romer, P. 1986. Increasing returns and long run growth. *Journal of Political Economy* 94, 1002–37.

Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the contrivance of money. *Journal of Political Economy* 66, 467–82.

Sargent, T.J. 1996. Expectations and nonneutrality of Lucas. *Journal of Monetary Economics* 37, 535–48.

Svensson, L.E.O. 1996. The scientific contributions of Robert E. Lucas, Jr. *Scandinavian Journal of Economics* 98, 1–10.

How to cite this article

Barseghyan, Levon. "Lucas, Robert (born 1937)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000216> doi:10.1057/9780230226203.0999

The New Palgrave Dictionary of Economics Online

lump sum taxes

J. de V. Graaff

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

excess burden of taxation; initial endowments; lump sum taxes; poll tax; redistribution of income of wealth

Article

A lump sum tax is fixed in amount and of such a nature that no action by the victim (short of emigration or suicide) can alter his or her liability. An example would be a poll tax, perhaps differentiated on the basis of sex and age.

It is difficult to find other examples. Differentiation on the basis of ability, wealth, income or expenditure would clearly lead to taxes that were not lump sum. Ability can be disguised. Wealth can be consumed. Leisure can be substituted for income, and saving for spending. All such actions would reduce tax. This implies that the principal criteria one might like to use as a basis for redistributive taxation are ruled out if one is confined to lump sum taxes. It also implies that it may be difficult to relate lump sum taxes to ability to pay. A feature of lump sum taxation is that what taxpayers bear is exactly balanced (in monetary terms) by what the fisc gains. That is because there is no tax at the margin. (If there were tax at the margin, taxpayers could vary their liabilities by varying their activities, and the tax would not be lump sum.) The absence of tax at the margin means that no transaction is killed off by the driving of a wedge between what one party pays and the other receives. When there is such a wedge (caused by a tax that is not lump sum) transactions are not entered into which, but for the tax, would have been mutually advantageous to the parties; and the loss to the parties is not balanced by any gain to the fisc. This is the ‘excess burden’ of taxation. It can never occur when taxes are lump sum.

In general equilibrium analysis the imposition of a set of lump sum taxes and bounties is equivalent to an adjustment of initial endowments. The attainment of equilibrium is not impaired, but its position will usually be altered. In welfare economics the conditions are investigated under which such an equilibrium may also represent a general optimum of production and exchange (in the sense of Pareto). If these conditions are met, it will not be possible to make one person better off without making someone else worse off. But the distribution of wealth may be very unequal: in an extreme case one person could end

up with everything, the others with nothing. Lump sum taxes can, in theory, correct this situation without impairing the general optimum. In this sense, they are an ideal form of taxation.

Lump sum taxes are thus of some importance in theoretical work. But in the real world, poll taxes being their only viable form, they are rarely encountered precisely because they cannot in practice be matched to ability to pay or used to achieve a redistribution of income of wealth without ceasing to be lump sum. At most they are a benchmark against which the less than perfect taxes we normally encounter can be measured.

See Also

- compensation principle
- neutral taxation
- optimal taxation

How to cite this article

Graaff, J. De V. "lump sum taxes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000160> doi:10.1057/9780230226203.1001

The New Palgrave Dictionary of Economics Online

Lundberg, Erik Filip (1907 – 1987)

Assar Lindbeck

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost inflation; Horndal effect; inventory cycles; inventory theory; Lundberg lag; Lundberg wage-multiplier; Lundberg, E. F.; Metzler, L. A.; redistribution of income; stabilization policy

Article

Lundberg was born in Stockholm and obtained a Ph.D. in economics in 1937 at the Stockholms Högskola. From 1937 to 1955 he was director of the Government Economic Research Institute (Konjunkturinstitutet), and from 1946 to 1965 he was professor of economics at the University of Stockholm; he held the same post at the Stockholm School of Economics from 1965 to 1970. He was president of the Royal Swedish Academy of Science from 1973 to 1976, and chairman of the Nobel Prize Committee for Economics from 1975 to 1980. He held numerous visiting professorships throughout the world.

Lundberg's main contributions to economic theory are his models of macroeconomic fluctuations and his analysis of the problems of economic policy, in particular the conflicts between stabilization policy and policies for the allocation of resources and the distribution of income.

His *Studies in The Theory of Economic Expansion* (1937) is an early work of high originality about the instability of growth, the main analytical technique being systems of difference equations of multiplier and accelerator mechanisms (with some consideration to the possibilities of flexible coefficients), embedded in a simple macroeconomic framework. Lags between inputs, output, income formation and spending play strategic roles (the lag between output and income-formation is often referred to as ‘the Lundberg lag’). Rather than providing reduced-form solutions to the system, Lundberg presented numerical sequences of various macroeconomic variables and their relations, so-called ‘sequence analysis’.

The part of the book which had the strongest immediate influence on other theorists is perhaps the inventory model. Non-anticipated increases in sales, while first resulting in a fall in inventories, later on, due to attempts by firms to restore the initial relation of inventory stocks to production levels, result in various kinds of inventory cycles. Lundberg's inventory analysis inspired, for instance, Lloyd Metzler's

inventory model, as well as the inventory analysis, with more elaborate microeconomic underpinnings, by Holt and Modigliani.

Among Lundberg's contributions to the analysis of economic policy, *Business Cycles and Economic Policy* (1953 in Swedish; 1957 in English), stands out as a particularly important piece of work. The analysis is characterized by rather informal theorizing, though using concepts of traditional economic theory, both for the ‘international’ and the Swedish economy. Calculations of *ex ante* inflation and deflation gaps, by way of excess demand (supply) in the goods market and/or the labour market, are important instruments of analysis.

Lundberg was also a pioneer in analysing the role of taxation for ‘cost inflation’, a point formalized by an equation expressing how much nominal wage rates would have to rise to guarantee a one per cent increase in after-tax real wage rates, after considering both the marginal tax rates and the price effects of wage increases (the Lundberg wage-multiplier).

When analysing long-term growth problems, Lundberg also discovered the so-called ‘Horndal effect’, expressing how labour productivity can go on rising over long periods of time without new investment, hence providing an indication of disembodied productivity growth (1961). Lundberg also made interesting comparative studies of growth, fluctuations and economic policy in various countries, for instance in *Instability and Economic Growth* (1968).

He also participated frequently in Swedish economic policy discussion, emphasizing the importance of avoiding overvalued exchange rates. His own policy recommendations, in addition, built on combining *general*, market-orientated stabilization policies with rather selective social policies to achieve economic security and desired income redistributions.

Selected works

1930. Om ekonomisk jämvikt. *Ekonomisk Tidskrift* 32(4), 133–60.

1937. *Studies in the Theory of Economic Expansion*. London: P.S. King & Sons. Oxford: Blackwell, 1955.

1953. *Konjunkturer och Ekonomisk Politik*. Stockholm: SNS. English edn, *Business Cycles and Economic Policy*, London: Allen & Unwin, 1957.

1959. The profitability of investment. *Economic Journal* 69, 653–77.

1961. *Produktivitet och räntabilitet*. Stockholm: SNS.

1968. *Instability and Economic Growth*. New Haven: Yale University Press.

1972. Productivity and structural change – a policy issue in Sweden. *Economic Journal* 82, March, Supplement, 465–85.

1985. The rise and fall of the Swedish model. *Journal of Economic Literature* 23, 1–36.

How to cite this article

Lindbeck, Assar. "Lundberg, Erik Filip (1907–1987)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000161> doi:10.1057/9780230226203.1002

The New Palgrave Dictionary of Economics Online

Lyapunov functions

C. Henry
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Within twelve years, from Poincaré's *Mémoire sur les courbes définies par une équation différentielle* (1881–6) to Lyapunov's thesis *Obsh•aya zada•a ob unstoi•ivosti dviženiya* (1892), the qualitative theory of differential equations emerged almot from scratch as the core of a new field in mathematics; both Poincaré and Lyapunov were motivated by problems in mechanics, celestial mechanics above all. Even if he did not match Poincaré's prodigious creativity between 1880 and 1883, Lyapunov developed from 1888 to 1892 a theory of dynamical stability which makes his 1892 thesis both a pioneering piece of work and a classic; in particular he developed a general stability criterion which now bears his name: the Lyapunov function.
Consider a system of ordinary differential equations

$$\dot{x} = f(x, t)$$

Where x is a vector in R^n and depends on t (t is in general interpreted as time), where $\dot{x} = dx / dt$ is the derivative of x with respect to t and where f is a function from R^{n+1} to R^n. A trajectory of the system is a function x from an interval T in R to R^n

$$x: T \to R^n: t \to x(t)$$

$$\dot{x}(t) \equiv f(x(t), t)$$

which is a solution of the system, i.e. such that $\dot{x}(t)$ and $f[x(t),•t]$ are identical on T; T is often of the form $[t_0,+\infty]$.
In what follows we shall limit ourselves to autonomous systems, i.e. systems of the form $\dot{x} = f(x)$, where f is dependent on t only through x. However, our whole presentation is easily generalized to non-autonomous systems, as is done in Rouche et al. (1977) and Rouche and Mawhin (1980).
It would appear at first sight that the system $\dot{x} = f(x)$ suffers from another restriction: it is a first-order system, in the sense that its equations include first-order derivatives only. This

might be seen a serious restriction indeed; think for example of the system formalizing the dynamics of the simple frictionless pendulum

$$\dot{x}_1 + \sin x_1 = 0$$

where x_1 is the angular distance from the vertical line. However, it is always possible to transform a system including derivatives of order higher than one into a first-order system with a higher number of equations. For example, the former system consisting of one second-order equation is equivalent to the following system of two first-order equations:

$$\dot{x}_1 = x_2 \dot{x}_2 = -\sin x_1$$

To investigate the stability properties of the pendulum, it is thus immediately possible to make use of the general concepts and methods available for first-order systems.
In order to introduce these concepts and methods in the spirit of Lyapunov, and then to see how they operate in economic models, we first have to be slightly more precise in defining an autonomous differential system as a system of first-order differential equations

$$(\mathrm{DS}) \quad \dot{x} = f(x)$$

where f is a continuous Lipschitzian function from an open subset Ω of R^n to R^n, i.e.

$$f : \Omega \to R^n : x \to f(x).$$

'Lipschitzian' means that there exists a constant α such that

$$\forall x^1, x^2 \in \Omega, \ \| f(x^1) - f(x^2) \| \leq \alpha \| x^1 - x^2 \|;$$

this assumption is very convenient because it ensures (see Coddington and Levinson, 1955) that through any point in Ω there passes one and only one trajectory of (DS); hence trajectories do not cross. However this assumption is not strictly necessary for what follows (see Aubin and Cellina, 1984 and Rouche et al. 1977).
We may now introduce the basic concepts of stability and attractivity for equilibria of dynamic systems. A point x^e in Ω is an equilibrium of (DS) if $f(x^e)=0$; in other words, x^e is an equilibrium if for any t_0 in R the function

$$x : [t_0, +\infty] \to R^n : t \to x^e$$

is a trajectory.

Stability: an equilibrium x^e is stable if a trajectory which comes sufficiently close to x^e never after recedes too far from x^e.

More precisely an equilibrium x^e is stable if, for any neighbourhood B_ε of x^e included in Ω, there exists a neighbourhood $B_\eta \subset B_\varepsilon$ such that any trajectory passing through B_η remains in B_ε ever after (see Figure 1).

Figure 1

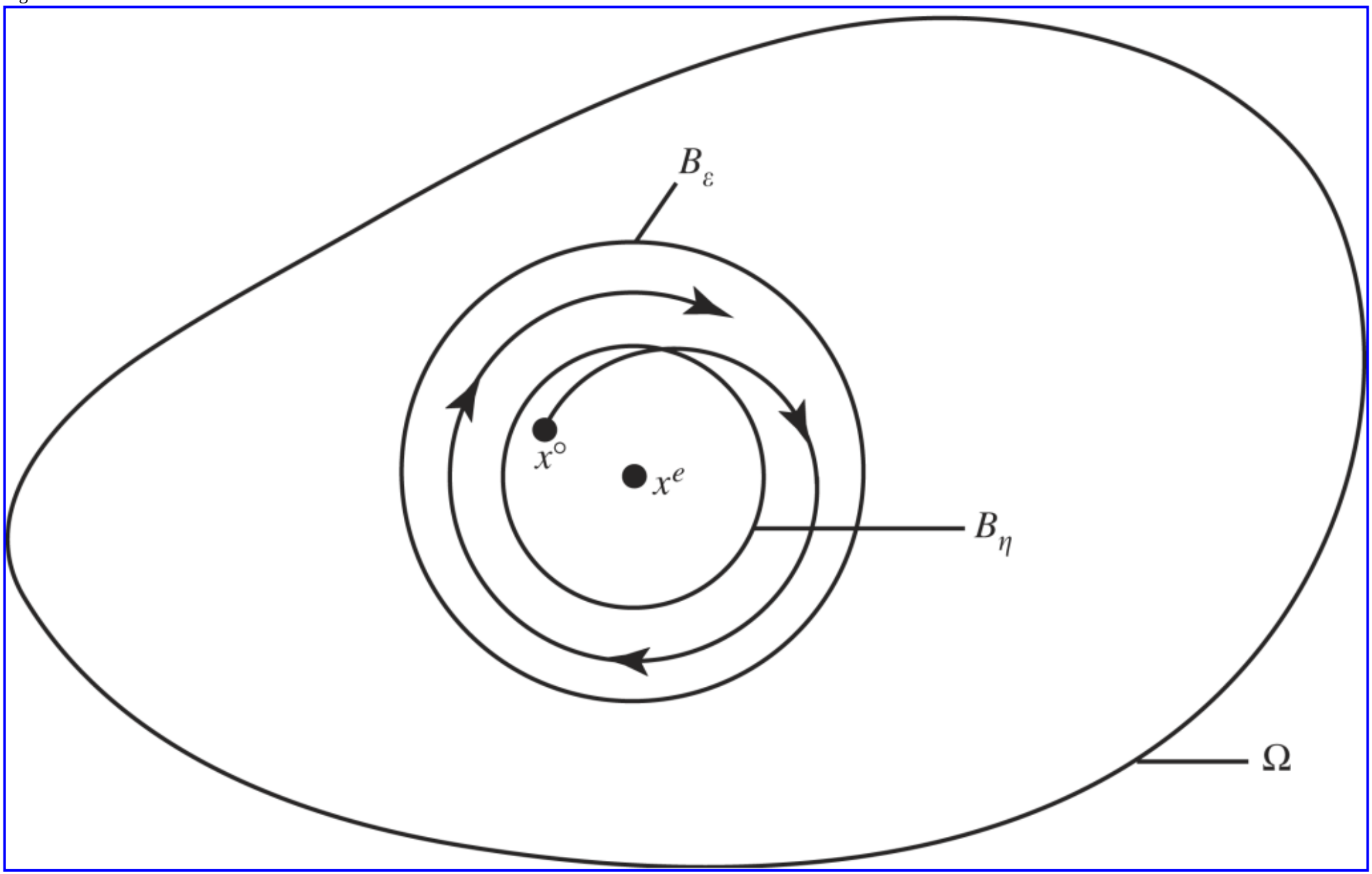

Local attractivity: an equilibrium x^e is a local attractor if a trajectory which comes sufficiently close to x^e later on tends to x^e.

More precisely an equilibrium x^e is a local attractor if there exists a neighbourhood B_δ of x^e included in Ω such that any trajectory which passes through B_δ tends to x^e as $t \to +\infty$; this does not mean that the trajectory always remains in B_δ (see Figure 2).

Figure 2

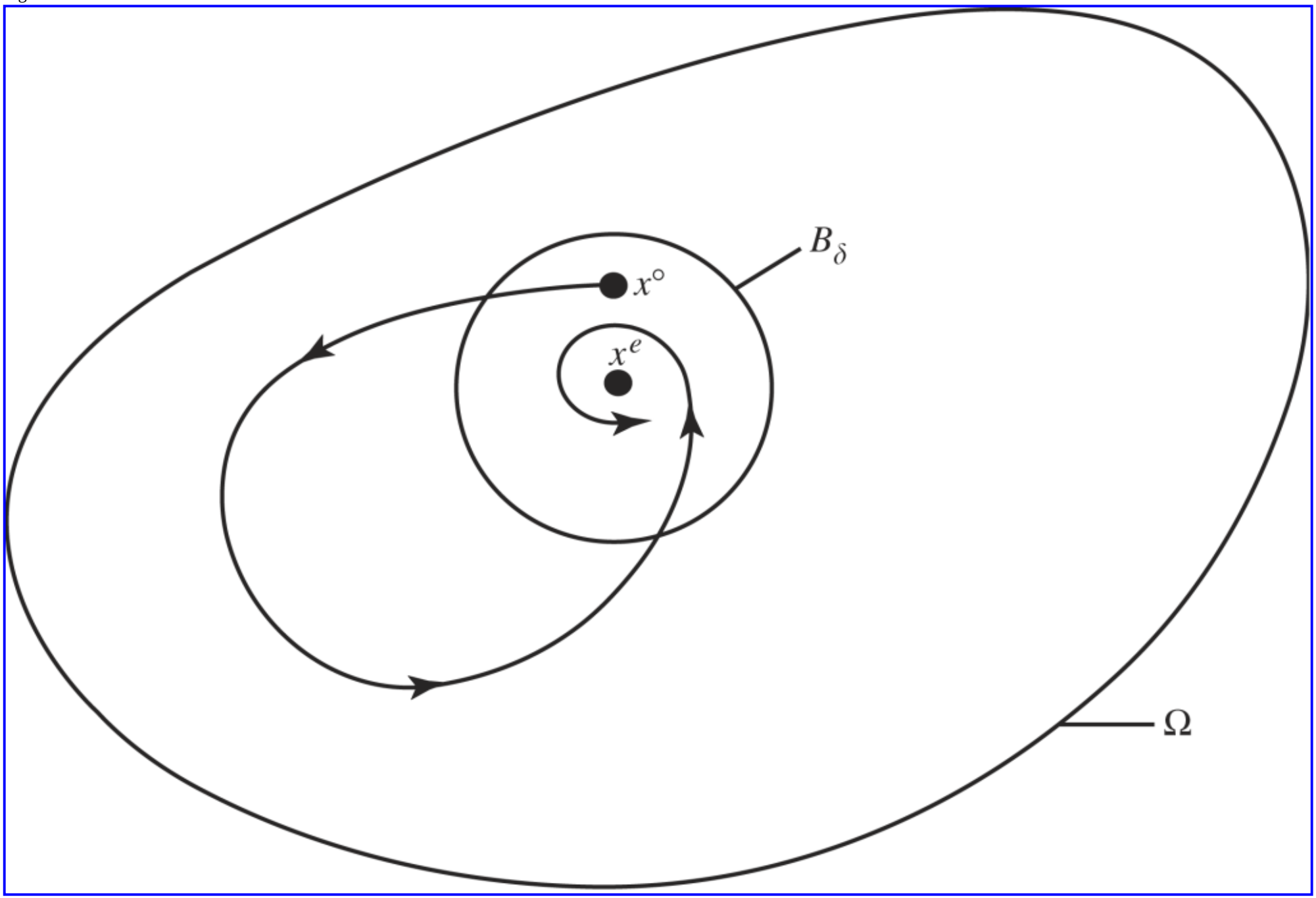

An equilibrium may be stable without being a local attractor, as in the case of the frictionless pendulum. It is also true that an equilibrium may be a local attractor, without being stable, but this is much more difficult to illustrate (see section 40 in Hahn, 1967). An equilibrium which is both stable and a local attractor is often called asymptotically stable.

Global attractivity: given a subset Ω_s of Ω, an equilibrium x^e in Ω_s is a global attractor with respect to Ω_s if any trajectory which passes through Ω_s tends to x^e as $t \to +\infty$.
An equilibrium which is both stable and a global attractor with respect to some Ω_s is often called globally asymptotically stable (globally with respect to Ω_s).
The convenient way, often the sole way, to deal with stability and attractivity as defined above, is in general to find a suitable Lyapunov function.
Lyapunov function: consider a subset Δ of Ω and a function of class C^1 (i.e. continuous and having continuous first-order partial derivatives)

$$W: \Delta \to R: x \to W(x).$$

W is a Lyapunov function if it satisfies the following requirements:

1. (i) it is bounded below on Δ, i.e.

$$\exists a \in R \text{ such that, } \forall x \in \Delta,\ W(x) \geq a.$$

(ii)it tends to infinity as *x*does, i.e.

$$\text{if } \|x\| \to +\infty, \text{ then } W(x) \to +\infty.$$

(iii)its time derivative $\dot{W}(x)$is nonpositive on Δ, i.e.

$$\forall x \in \Delta,\ \dot{W}(x) \leq 0$$

where the time derivative is defined as

$$\dot{W}: \Delta \to R: x \to \dot{W}(x) = \sum_{k-1}^{x} \frac{\partial W(x)}{\partial x_k} f_h(x).$$

The name ‘time derivative’ is warranted as

$$\dot{W}[x(t)] = \sum_{k-1}^{x} \frac{\partial W[x(t)]}{\partial x_k} \dot{x}_k(t) = \frac{dW[x(t)]}{dt}$$

along any trajectory in Δ.

It is possible for n=2 to draw level curves of W in the subset Δ of the plane (x_1, x_2). On Figure 3 two level curves are drawn, with $k' < k''$, as well as a trajectory. W is non-increasing along any trajectory in Δ; hence, as soon as a trajectory reaches a level curve, say k', it never again comes back to points on level curves k with $k>k'$, for example $k=k''$. This suggests deriving properties of stability and attractivity from the existence of a Lyapunov function; we indeed have:

Figure 3

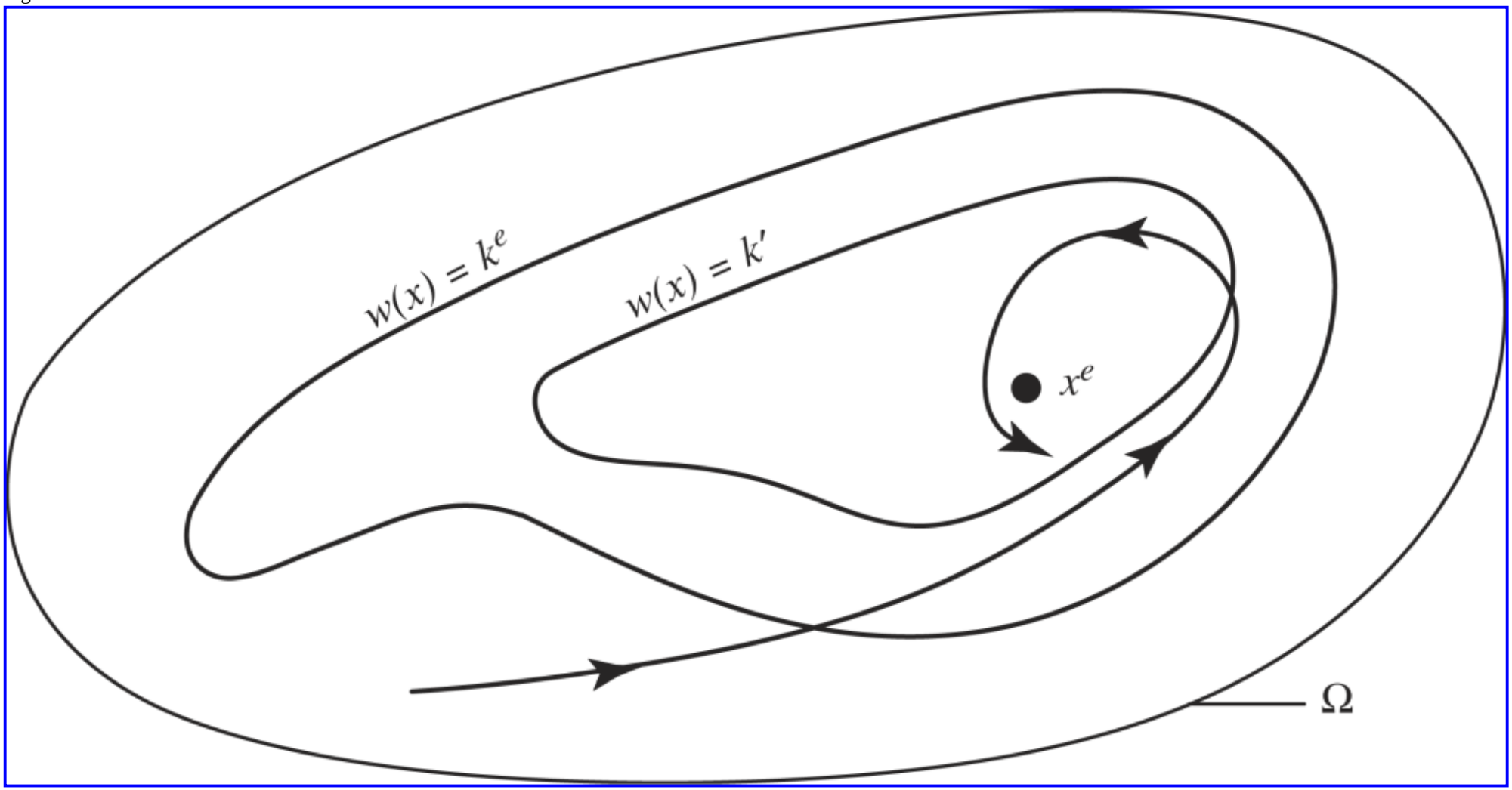

Proposition 1: : if there exists on some neighbourhood B of the equilibrium x^e a Lyapunov function W and if x^e is an isolated minimum of W on B, then x^e is a stable equilibrium. If moreover x^e is the only point in B where $\dot{W} = 0$, then x^e is also a local attractor.

These are sufficient conditions for stability and local attractivity. It turns out that the existence of a Lyapunov function is also a necessary condition (for a general exposition and complete proofs which are valid even for nonautonomous systems, see Rouche et al. (1977) and Rouche and Mawhin (1980)).

The first part of Proposition 1, but not the second part, applies to the frictionless pendulum, the Lyapunov function being here the total energy $\frac{1}{2}x_2^2 - \cos x_1 + 1$. Both parts of Proposition 1 apply to the tâtonnement process in a competitive economy where all goods are gross substitutes for all prices. Let *n* be the number of goods; let *p* be the price vector normalized in such a way that it is in the *n* – 1 dimensional unit simplex $\overline{\Sigma}$ defined by $\Sigma_{j=1}^{n} p_j = 1;$ and let $z_j(p)$ be the aggregate excess demand function for good *j*. Let Σ be the interior of $\overline{\Sigma}$. Gross substitutability implies that there exists one and only one general competitive equilibrium price vector p^e and that p^e is in Σ, i.e. p_j^e is strictly positive for all goods *j* (for more details see Arrow and Hahn, 1971).

Consider then the well-known tâtonnement process

$$(\mathrm{TP})\ \dot{p} = z(p).$$

It is a (DS) system, with $\Omega = \Sigma$; being the unique general competitive equilibrium price vector, p^e is also the unique equilibrium of the differential system (TP). Consider on Σ the function

$$W(p) = \| p - p^e \|^2.$$

Its time derivative is

$$\dot{W}(p) = \sum_{j-1}^{x} \frac{\partial W(p)}{\partial p_j} z_j(p) = 2\sum_{j-1}^{x} (p_j - p_j^e) z_j(p) = -2p^e \cdot z(p),$$

because of Walras's law. On the other hand, it is a consequence of gross substitutability that

$$\forall p \neq p^e,\ p^e \cdot z(p) > 0.$$

It is thus clear that *W* is a Lyapunov function, that p^e is its unique minimum on Σ and that $\dot{W}$ is zero only at p^e; hence p^e is stable and is a local attractor for the tâtonnement process. Is it a global attractor with respect to Σ? The answer is not within the range of proposition 1. Something more is needed.

Proposition 2: : consider a system (DS) and a bounded subset Δ of Ω which is such that there exists a Lyapunov function *W*

$$W: \Delta \to R: x \to W(x)$$

satisfying the additional requirement that $\dot{W}$ is zero only at equilibria of the system, i.e.

$$\dot{W}(x) = 0 \Rightarrow f(x) = 0.$$

Then, if all the limit points of a trajectory are in Δ, they are equilibria of the system; if moreover all the equilibria of the system are isolated, this trajectory tends to one of them as $t \to +\infty$.

Corollary: : if the system has a unique equilibrium, if Δ is open and if any trajectory which passes through Δ has all its limit points in Δ, then the equilibrium is a global attractor with respect to Δ and it is stable.

This corollary has no general counterpart when there are several isolated equilibria, because a trajectory starting in the neighbourhood of one equilibrium may tend to another one. However, if an equilibrium is an isolated minimum of *W*, prpt 1 ensures that it is stable and is a local attractor.

Proposition 2 and its corollary allow us to answer the qst, left unanswered above, about the tâtonnement process: as Σ is bounded and as gross substitutability prevents any trajectory from having a limit point on the boundary of Σ, all conditions in prpt 2 and in the corollary are met; so p^e is a global attractor with respect to Σ.

Another well-known application of Lyapunov functions is in the theory of public goods. Consider an economy with *N* consumers (*i*=1,…, *N*), *m* public goods (*k*=1,…, *m*) and one private good used as numeraire. Let $x \in R_+^m$ denote the bundle of public goods made available to the consumers, and let y^i denote the amount of numeraire consumed by *i*=1,…, *N*; this means that $(x, y^i) \in R_+^{m+1}$ describe the total consumption of *i*. His preferences are formalized by a utility function

$$u^i: R_+^{m+1} \to R: (x, y^i) \to u^i(x, y^i);$$

this function is of class C^1, quasi-concave, nondecreasing with respect to each of its arguments, and is strictly increasing with respect to the consumption of the numeraire, i.e.

$$\frac{\partial u^i}{\partial y^i} > 0 \text{ on } R_+^{m+1}.$$

Let *Z* be the set of feasible allocations, i.e. the set of all $z = (x, y^1, \ldots, y^N)$ in R_+^{m+N} which can be made available for consumption, given the technical possibilities and the initial resources of the economy. *Z* is of course bounded; it is reasonable to consider that it is closed and convex; hence it is a compact convex subset of R_+^{m+N}.

How to reach a Pareto-optimal feasible allocation? The MDP (for Malinvaud–Drèze–Poussin, see Champsaur et al., 1977) planning procedure gives the following answer: starting from any feasible allocation, revise *z* continuously according to the following differential system, which is a (DS) system:

$$\text{(MDP)}\ \frac{\mathrm{d}x_k}{\mathrm{d}t} = \sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z),\ k = 1, \ldots, m \quad \frac{\mathrm{d}y^i}{\mathrm{d}t} = -\sum_{k=1}^{m} \pi_k^i(z)\frac{\mathrm{d}x_k}{\mathrm{d}t} + \delta^i \sum_{k=1}^{m} \frac{\mathrm{d}x_k}{\mathrm{d}t}\left[\sum_{j=1}^{N} \pi_k^j(z) - \gamma_k(z)\right],\ i = 1, \ldots, N$$

where $\pi_k^i(z)$ is the marginal willingness to pay of consumer *i* for public good *k*, and $\gamma_k(z)$ is the marginal cost of public good *k*. The δ_i, *i*=1,…, *N*, are non-negative weights summing up to 1: $\Sigma_{i=1}^N \delta^i = 1$. These differential equations mean that the quantity made available of each public good is revised according to the difference between the total marginal willingness to pay for that public good and its marginal cost; simultaneously every consumer pays an amount of numéraire equal to his willingness to pay for this set of revisions, and receives a fraction of the total surplus that the revisions generate.
Consider the function

$$W: Z \to R: z \to W(z) = -u^i(x, y^i)$$

where *i* is chosen among those *i* for which δ^i>0. Straightforward calculations lead to

$$\dot{W}(z) = -\delta^i \sum_{k-1}^{m} [\pi_k^i(z) - \gamma_k(z)]^2 \frac{\partial u^i}{\partial y^i},$$

which is nonpositive everywhere on *Z* and is zero if and only if *z* is an equilibrium of (MDP). *W* is a Lyapunov function and Proposition 2 applies with Δ=*Z*. As *Z* is compact it is even possible to conclude that any limit point of any trajectory which is included in *Z* is an equilibrium of (MDP). If all the utility functions u^i, *i*=1,…, *N*, are strictly quasi-concave, the result is sharpened in the sense that any trajectory which is included in *Z* tends to an equilibrium as $t \to +\infty$. The economic significance of these results proceeds from the fact that all the equilibria of (MDP) are Pareto optima.
However, it is not guaranteed that all trajectories of (MDP) starting in *Z* are included in *Z*, as the revisions generated by the equations

$$\frac{dx_k}{dt} = \sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z),\ k = 1, \ldots, m$$

may lead to negative values of the public goods. We would then have a meaningless procedure. In order to avoid this possibility the above equations must be replaced in (MDP) by:

$$\frac{dx_k}{dt} = \begin{cases} \sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z) & \text{for } x_k > 0 \\ \max\left[\sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z), 0\right] & \text{for } x_k = 0\,. \end{cases}$$

It is then immediate that any trajectory of (MDP) starting in *Z* is included in *Z*. But do trajectories still exist? and if they exist, do they actually tend to equilibria of (MDP)? The

answers are not trivial, as there are significant discontinuities in the right-hand sides of the new equations. These answers nevertheless turn out to be positive; this is a by-product of the extension of existence and stability theorems to multivalued dynamical systems

$$\frac{dz}{dt} \in F(z)$$

where F is an upper hemicontinuous correspondence such that the image $F(z)$, of any point z in an open subset Ω of R^n, is a compact convex subset of R^n. For such systems, Lyapunov functions have been defined with the same purposes as for ordinary systems (see Champsaur et al., 1977; Aubin and Cellina, 1984).

Lyapunov functions are used in many other economic models, to prove the convergence of non-tâtonnement processes for example (see Arrow and Hahn, 1971) or to investigate the stability properties of a process of free entry and exit of firms, facing random demand and guided by expected profits (see Drèze and Sheshinski, 1984). Of particular interest is the use of a Lyapunov function of the form $(Q - Q^e) \cdot (k - k^e)$, where k is the vector of capital stocks in the economy and Q is the vector of current prices for investment goods, to show that any optimal growth path tends to the (suitably modified) golden rule capital stock k^* when the discount rate is not too large (see Brock and Scheinkman, 1976; Cass and Shell, 1976).

Till now we have dealt only with dynamical stability, i.e. with qsts typically like the following one: two trajectories happen to pass through two neighbouring points; does it imply that they will ever after remain close to each other? Around 1970, G. Debreu and S. Smale introduced structural stability into economic theory, i.e. stability with respect to parameters of the system; a typical question is here: is the configuration of competitive equilibria of an economy (for example the fact that they are isolated) stable when the initial endowments of the agents in the economy change? Almost a century before, Poincaré introduced and systematically explored the concept of bifurcation in mathematics (see Poincaré, 1881); the word came to him as a natural comparison with daily experience:

> On voit que les deux catégories d'ellipsoïdes forment deux séries continues de figures d'équilibre. Mais il y a une figure qui est commune aux deux séries et qui est, si l'on veut me permettre cette comparaison, un point de bifurcation. (Poincaré, 1892, p. 810)

Bifurcation is a basic concept for the study of structural stability; even if the latter expression was to come much later, the essence of the approach is in Poincaré's works.

In the introduction (Poincaré, 1882), Poincaré refers to 'les recherches ultérieures parmi lesquelles les plus importantes sont, sans contredit, celles de M. Liapounoff'. It seems indeed that no other mathematician of the time saw better than Lyapunov did the significance of Poincaré's new concepts and methods. In the three volumes (Lyapunov, 1906–1912), Lyapunov explored in great detail the bifurcation of the equilibrium configurations of a rotating homogeneous mass of liquid. The ultimate goal was to explain the evolution of stars. As long as the angular velocity Ω of the rotating mass is less than or equal to a critical value ω^c, there is one and only one equilibrium configuration for each velocity ω, and it is an ellipsoid. But at ω^c a bifurcation appears: at ω^c the equilibrium configuration is still unique and is an ellipsoid but, in Lyapunov's own words, 'C'est l'ellipsoïde, par lequel on entre dans la série des figures d'équilibre que M. Poincaré a appelé pyriformes' (Lyapunov, 1906–1912, vol. 3, p. 6). It is indeed shown (Lyapunov, 1906–1912, vol. 1, pp. 216–17 and vol. 3, p. 106) that there exists an interval $[\omega^c, \overline{\omega}]$ such that, for every ω in this interval there are two equilibrium configurations: the usual ellipsoid and a pear-shaped configuration, whose symmetry and stability properties were systematically investigated by Lyapunov. This is a study in structural stability, the last one to appear before 1937, at which time Andronov and Pontrjagin (1937) picked up the subject, which has been exploding since then.

It has recently been shown that in (strictly deterministic) economic growth models, bifurcation phenomena can take place which are strikingly similar to those explored by Lyapunov: ω is replaced by the discount rate r, the ellipsoids by steady states and the pear-shaped configurations by closed cycles that bifurcate from the steady state for some value r_0 of r (see Benhabib and Nishimura, 1979). Bifurcations even appear in stationary competitive monetary economies: at critical values of some parameters of the economy – for example the degree of concavity of utility functions – a stationary equilibrium bifurcates towards a line (a 'série', in Poincaré's words) of stationary equilibria on one hand, and a simultaneous line of closed cycles; the latter are the business cycles of the model, and their stability under suitable assumptions has been shown using Poincaré–Lyapunov methods (see Grandmont, 1985).

Economists tend to know Lyapunov for his celebrated functions, it appears that there is even more to interest them in the various approaches to stability that Lyapunov has developed during his lifelong study of dynamical systems.

See Also

- correspondence principle
- gross substitutes

Bibliography

Andronov, A.A. and Pontrjagin, L.S. 1937. Systèmes grossiers. *Doklady Akademii Nauk* 14, 247–51.

Arrow, K. and Hahn, F. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Aubin, J.P. and Cellina, A. 1984. *Differential Inclusions*. Berlin: Springer.

Benhabib, J. and Nishimura, K. 1979. The Hopf bifurcation and the existence and stability of closed orbits in multisector models of optimal economic growth. *Journal of Economic Theory* 21, 421–44.

Brock, W.A. and Scheinkman, J.A. 1976. Global asymptotic stability of optimal control systems with applications to the theory of economic growth. *Journal of Economic Theory* 12, 164–90.

Cass, D. and Shell, K. 1976. The structure and stability of competitive dynamical systems. *Journal of Economic Theory* 12, 31–70.

Champsaur, P., Drèze, J. and Henry, C. 1977. Stability theorems with economic applications. *Econometrica* 45, 273–94.

Coddington, E.A. and Levinson, N. 1955. *Theory of Ordinary Differential Equations*. New York: McGraw-Hill.

Drèze, J.H. and Sheshinski, E. 1984. On industry equilibrium under uncertainty. *Journal of Economic Theory* 33, 88–97.

Grandmont, J.M. 1985. On endogenous competitive business cycles. *Econometrica* 53, 995–1045.

Hahn, W. 1967. *Stability of Motion*. Berlin: Springer.

Lyapunov, A. 1892. Obsh•aya zada•a ob ustoi•ivosti dviženiya (The general problem of the stability of motion). *Soc. Math. Kharkov*. The 1907 French translation has been reproduced in *Annals of Mathematics Studies* 17, Princeton: Princeton University Press, 1949.

Lyapunov, A. 1906–12. *Sur les figures d'équilibre peu différentes des ellipsoïdes d'une masse liquide homogène douée d'un mouvement de rotation*. 3 vols, Académie impériale des Sciences, St. Petersburg.

Poincaré, H. 1881. Mémoire sur les courbes définies par une équation différentielle. *Journal de Mathématiques Pures et Appliquées* 7(3), 1881, 375–422; 8(3), 1882, 251–96; 1(4), 1885, 167–244; 2(4), 1886, 151–217.

Poincaré, H. 1882. Les formes d'équilibre d'une masse fluide en rotation. *Revue Générale des Sciences* 3, 809–15.

Rouche, N., Habets, P. and Laloy, M. 1977. *Stability Theory by Lyapunov's Direct Method*. Applied Mathematical Sciences No. 22, Berlin: Springer.

Rouche, N. and Mawhin, J. 1980. *Ordinary Differential, Equations: Stability and Periodic Solutions*. Surveys and Reference Works in Mathematics, London: Pitman.

How to cite this article

Henry, C. "Lyapunov functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_L000165>
doi:10.1057/9780230226203.1003

The New Palgrave Dictionary of Economics Online

Macfie, Alec Lawrence (1898 – 1980)

Andrew Skinner

From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Article

Alec Macfie was born in Partick on 29 May 1898. He went first to school at Hillhead but later joined his brother at the High School of Glasgow where he had a distinguished career.

When he left school, Macfie, too young to enlist, worked in a munitions factory for a few months. But in 1916 he joined the Second Battalion, the Gordon Highlanders, and was commissioned as a lieutenant. He saw action at Passchendaele, and was badly wounded during an action on the Asiego Plateau in the early summer of 1918. After recovering to some extent from his wounds, Macfie entered Glasgow University where he graduated with first class honours in philosophy and English literature in 1922. Thereafter he entered a law office and took his LL.B. But having opted for an academic career, he returned to the university and took a first in economics and politics in 1927.

In the years that followed Macfie held temporary teaching posts in Nottingham and St Andrews (where he deputized for the professor of moral philosophy) before returning to Glasgow in 1932 as lecturer in the Department of Political Economy under W.R. Scott. Scott (the ‘chief’) died in 1940 and Macfie was invited to take the Adam Smith Chair in 1945.

Side by side with his teaching, Macfie produced three books in the period up to 1945, all of which reflect his interest in philosophy and psychology as well as in economics: *Theories of the Trade Cycle* (1934), *An Essay on Economy and Value* (1936) and *Economic Efficiency and Social Welfare* (1943). It was not until the mid-1950s, only a few years before retiring, that he embarked on a serious study of Adam Smith with special reference to the *Theory of Moral Sentiments*. Following the acquisition of the manuscripts, discovered by J.M. Lothian in 1958, Macfie became one of the driving forces behind the Glasgow edition of the Works and Correspondence of Adam Smith, and acted as co-editor (with D.D. Raphael) of the *Theory of Moral Sentiments* (1976). He also produced a little book which has exerted an enormous influence in this field, *The Individual in Society: Papers on Adam Smith* (1967).

Few modern scholars have been better equipped for the study of Smith. Macfie was a qualified lawyer, with degrees in philosophy, literature, and economics, while Smith was writing at a time when it was possible to think in terms of a global system of thought which might embrace these separate disciplines.

Selected works

1934. *Theories of the Trade Cycle*. London: Macmillan. Reprinted, New York: Augustus M. Kelley, 1971.

1936. *An Essay on Economy and Value: Being an Enquiry into the Real Nature of Economy*. London: Macmillan.

1940. (ed.) *Studies Relating to Adam Smith During the Last Fifty Years*, by W.R. Scott. Proceedings of the British Academy.

1943. *Economic Efficiency and Social Welfare*. Oxford: Oxford University Press.

1967. *The Individual in Society: Papers on Adam Smith*. London: Allen & Unwin.

1976. (ed., with D.D. Raphael.) *The Theory of Moral Sentiments*, by Adam Smith. Oxford: Clarendon Press; Vol. I of the Glasgow edn. of the Works and Correspondence of Adam Smith.

How to cite this article

Skinner, Andrew. "Macfie, Alec Lawrence (1898–1980) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000002> doi:10.1057/9780230226203.1885

The New Palgrave Dictionary of Economics Online

machinery question

S. Rashid

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

circulating and fixed capital; fixed capital; machinery question; Mill, J.S.; Ricardo, D.; Say's Law; Stewart, D.; Tozer, J.E.; Tucker, J.; wages fund; Wicksell, J.G.K.

Article

That machinery is of benefit to the manufacturer who introduces it has never been a point of discussion in the history of economics and the machinery question is solely a dispute over whether society benefits from the introduction of machinery, the most pressing social issue being the displacement of labour by machinery and the consequent threat of widespread unemployment. In general terms, the social benefits of machinery were well appreciated by the middle of the 18th century. However, the greatly increased use of machinery at the end of the 18th century gave a new intensity to the debate at the beginning of the 19th century. The analytical tools used by classical economists to tackle this general equilibrium problem were however quite inadequate and it is doubtful whether a deeper understanding of the issue was achieved by the heroic abstractions of the 19th century.

The earliest explicit discussions of machinery appear to be in the pamphlets of John Cary (1695), *A Discourse on Trade*. It was a time when the competitiveness of English industry was being much discussed and John Cary pointed out that England retained her business advantage because of the ability of English manufacturers to invent.

> Tobacco is cut by Engines: Books are printed; Deal Boards are sawn with Mills; Lead is smelted by Wind-Furnaces; all which save the Labour of many Hands, so the Wages of those employed need not be fallen. …
> New Projections are every Day set on Foot to render the making our Woollen Manufactures easy, which should be rendered cheaper by the Contrivance of the Manufacturers, not by falling the Price of Labour: Cheapness creates Expence, and gives fresh Employments, whereby the Poor will be still kept at Work. (Cary, 1695, pp. 99–100)

A few years later, in his *Considerations of the East-India Trade* (1701), Henry Martin advocated the import of cheaper cloth from the East Indies by comparing it with the effects of machinery:

> Arts, and Mills, and Engines, which save the labour of Hands, are ways of doing things with less labour, and consequently with labour of less price, tho' the Wages of Men imploy'd to do them shou'd not be abated. The *East-India* Trade procures things with less and cheaper labour than would be necessary to make the like in *England*; it is therefore very likely to be the cause of the invention of Arts, and Mills, and Engines, to save the labour of Hands in other Manufactures. Such things are successively invented to do a great deal of work with little labour of Hands; they are the effects of Necessity and Emulation; every Man must be still inventing himself, or be still advancing to farther perfection upon the invention of other Men … (Martin, 1701, pp. 589–90)

At this stage the effect of machinery in preserving competitiveness receives primary emphasis. There is as yet no link drawn between high wages and the incentive to create machinery. In the years that followed only the prolific Daniel Defoe paid serious attention to the role of machinery without making any substantive analytical contribution. Indeed, Defoe even wondered whether machinery were not sometimes an evil because it displaced labour. In parliamentary debates in 1738 on the making of buttons by loom instead of by hand, Henry Archer implicitly subscribed to the full employment and sustainability thesis in a speech of considerable eloquence:

> As to the honourable gentleman's other arguments, drawn from the number of hands employed in the needle-work manufacture … it is, in my humble opinion, a very good argument for dismissing this Bill; because, as the manufacture may be carried on by a much fewer number of hands, with equal advantage to our trade in general, those who are employed in the needle-work way, are so many hands taken from other arts and other manufactures, in which they might be employed to much better purpose.

Archer goes on to use an example that was repeated often by classical economists:

> There was a time, Sir, when all the learning of this kingdom, and the rest of Europe, was contained in manuscripts, the writing of which employed great numbers of hands, and took up a vast deal of time in re-copying. But, Sir, how ridiculous would it have been, if on the discovery of the art of printing, the transcribers and copyers of those manuscripts had joined in a petition to the legislature, that it would be pleased to prohibit the art of printing, for the same reason which the honourable gentleman now uses, because great numbers would thereby be deprived of bread! (Archer, 1742)

The next advance was stimulated by Montesquieu's claim in *The Spirit of the Laws* that the introduction of machinery was not necessarily beneficial. This provoked Josiah Tucker to provide one of the best statements on the effects of machinery:

> What is the Consequence of this Abridgment of Labour, both regarding the Price of the Goods, and the Number of Persons employed? The Answer is very short and full, *viz*. That the Price of Goods is thereby prodigiously lowered from what otherwise it must have been; and that a much greater Number of Hands are employed….
> And the first Step is, that Cheapness, *ceteris paribus* is an Inducement to buy, – and that many Buyers cause a great Demand, – and that a great Demand brings on a great Consumption; – which great Consumption must necessarily employ a vast Variety of Hands, whether the original Material is considered, or the Number and Repair of Machines, or the Materials out of which those Machines are made, or the Persons necessarily employed in tending upon and conducting them: Not to mention those Branches of the Manufacture, Package, Porterage, Stationary Articles, and Book-keeping, &c. &c. which must inevitably be performed by human Labour.…
> That System of Machines, which so greatly reduces the Price of Labour, as to enable the Generality of a People to become Purchasers of the Goods, will in the End, though not immediately, employ more Hands in the Manufacture, than could possibly have found Employment, had no such machines been invented. And every manufacturing Place, when duly considered, is an Evidence in this Point. (Tucker, 1757, pp. 241–2)

The subject received little further impetus in the half-century that followed. The tangential discussion of machinery by Adam Smith in the *Wealth of Nations* perhaps contributed to this state of affairs. The only notable treatment is in the lectures of Dugald Stewart at Edinburgh (1858–78), which were very influential as part of an oral tradition, but which were not available in print till the 1860s. Stewart's contribution lay in seeing the machinery question as part of a much larger and more fundamental policy issue – the trade-off between individual losses and social gains. He therefore links together three issues that had hitherto been separately discussed – the creation of large farms, the benefits of enclosures and the use of machinery. In each case Stewart grants that the hardships imposed on individuals were undeniable. He then continues;

> In judging of the policy of such innovations … it is absolutely necessary to abstract from the individual hardships that may fall under our notice, and to fix our attention on those general principles which influence the national prosperity. (Stewart, 1856, vol. 8, p. 131)

In deciding upon the benefits of introducing machinery, Stewart observes that the material improvement of mankind and the use of machinery are practically inseparable. The policy recommendation was thus unequivocal.

> It is hardly possible to introduce suddenly the smallest innovation into the Political Economy of a State, let it be ever so reasonable, nay, ever so profitable, without incurring some inconveniences. But temporary inconveniences furnish no objection to solid improvements. Those which may arise from the sudden introduction of a machine cannot possibly be of long continuance. The workmen will, in all probability, be soon able to turn their industry into some other channel; and they are certainly entitled to every assistance

> the public can give them, when they are thus forced to change their professional habits. (1856, vol. 8, p. 193)

The severe post-Napoleonic depression contributed greatly to a reconsideration of the effects of machinery and John Barton should perhaps be given most credit for the new interest with his pamphlet, *Observations on the Circumstances which Influence the Condition of the Labouring Classes of Society* (1817). Commenting on the distinction, inherited from Adam Smith, between circulating and fixed capital, Barton pointed out that only the former serves to employ labour – the latter is embodied in machinery. Since it appeared empirically undeniable that progress involved a greater proportionate use of fixed capital Barton argued that the funds for employing labour, or circulating capital, must be subject to proportionate decrease and lead to greater unemployment. Barton was very clear about the role of high wages in inducing the adoption of machinery.

> It is the proportion which the wages of labour at any particular time bear to the whole produce of that labour, which appears to me to determine the appropriation of capital in one way or the other. For if at any time the rate of wages should decline, while the price of goods remained the same, or if goods should rise, while wages remained the same, the profit of the employer would increase, and he would be induced to hire more hands. If, on the other hand, wages should rise, in proportion to commodities, the labour's share in the produce of his own industry would be increased at the expense of his master, who would of course keep as few hands at possible. – He would aim at performing every thing by machinery, rather than by manual labour. (Barton, 1817, pp. 17–18)

How far David Ricardo was directly influenced by Barton in reversing his initial optimistic position on the benefits of machinery is unclear, but in the third edition of his *Principles of Political Economy and Taxation* (1821, pp. 388–95) Ricardo tried to justify some of the pessimistic attitudes to machinery by means of a numerical example. To begin with, we have a farmer whose yearly activities can be summarized as follows:

Fixed Capital	7,000
Wages (Circulating)	13,000
	– – – – – –
	20,000
Profits (10 per cent)	2,000 (used for consumption)
	– – – – – –
Total	22,000
	– – – – – –

The circulating capital is said to ‘replace the value of 15,000’, that is, to provide the required profit of 2,000. In year 1, the capitalist sets half his workers to construct machines. As surplus value arises from circulating capital, the profits of 2,000 arises in equal parts from the workers in farming and the workers in machines:

Fixed Capital (Old)	7,000

Wages (Farming)	6,500
Profits (Farming)	1,000
Wages (Machines)	6,500 (Embodied in machines)
Profits (Machines)	1,000
Total	22,000

If the farmer still spends 2,000 for his own consumption, he is left with 5,500 to spend on wages the next year. In other words, the wage bill falls from 13,000 to 5,500 because of the construction of machines. The gross produce consists of profits, rent and wages, while the net produce consists of profits and rent only. In our case, there is no rent, so the gross produce falls from 15,000 to 7,500 while the net produce stays at 2,000. Ricardo concludes as follows:

> In this case, then, although the net produce will not be diminished in value, although its power of purchasing commodities may be greatly increased, the gross produce will have fallen from a value of 15,000•*l* to a value of 7,500•*l*, and as the power of supporting a population, and employing labour, depends always on the gross produce of a nation, and not on its net produce, there will necessarily be a diminution in the demand for labour, population will become redundant, and the situation of the labouring classes will be that of distress and poverty.

Subsequently, Ricardo concedes that more workers can be employed in producing goods that the capitalist may wish to consume with his increased real power of consumption, but this may not be strong enough to compensate for the initial loss of employment.

> All I wish to prove, is, that the discovery and use of machinery may be attended with a diminution of gross produce; and whenever that is the case, it will be injurious to the labouring class, as some of their number will be thrown out of employment, and population will become redundant, compared with the funds which are to employ it.

There are a number of curious features about Ricardo's analysis which, though based on a simple numerical example, is claimed to have some relevance. First, it is not at all clear whether Say's Law, which Ricardo adhered to so vehemently on other occasions, also operates when labour is displaced by machinery. Secondly, Ricardo simply presents the initial disruption of new machinery without saying anything about the nature of the new equilibrium or the adjustment process leading to it. This contrasts sharply with his usual emphasis upon permanent effects – indeed, in assuming that the new machines made will be able to realize 1,000 units of profit Ricardo is implicitly assuming some sort of pervasive equilibrium. Thirdly, Ricardo appears to deny the practical importance of his example at the end of the chapter when he emphasizes that his argument holds only when the new machinery is introduced suddenly.

> The statements which I have made will not, I hope, lead to the inference that machinery should not be encouraged. To elucidate the principle, I have been supposing, that improved machinery is *suddenly* discovered, and extensively used; but the truth is, that these discoveries are gradual, and rather operate in determining the employment of the capital which is saved and accumulated, than in diverting capital from its actual employment.

This point gains additional force from Ricardo's insistence that the state take no action to discourage technological progress. Most subsequent economists, from Malthus onwards, took exception to the collection of assumptions necessary to produce Ricardo's result.

Of the classical economists who followed, only Nassau Senior and John Stuart Mill tried to justify Ricardo's reasoning, sometimes with the same surprising pattern of argument that characterized Ricardo. For example, John Stuart Mill (1848) begins by asserting that workers suffer temporarily when circulating capital is converted to fixed capital; almost immediately however he adds that this is a case which scarcely ever occurs in practice. An attempt by J.E. Tozer to provide a mathematical formulation of the question does not go beyond the framework set of by Ricardo. Tozer (1838) grants that there is an initial deduction from the wages fund but points out that the fund is replenished over time as the additional output from the machinery is produced. There does not appear to be a serious effort at going beyond Ricardo's analytical schema until the writings of Knut Wicksell.

With his usual clarity, Wicksell begins his section on production and distribution by setting forth the technical conditions necessary for the validity of the marginal productivity theory of distribution. He recognizes that the distributive impact of machinery depends upon the manner in which machinery alters the marginal productivities of labour and capital and that simple answers to such a question are unlikely. One issue which he analyses with considerable acumen is the position of Ricardo that machinery may actually diminish the gross product. Wicksell takes issue with Ricardo's conclusion and claims that Ricardo did not follow his premises to their logical conclusion – under free competition, changes in technique cannot lead to a diminution of gross product. This is proved as follows:

> Let x and y denote the number of labourers per acre using the old and new methods of cultivation, and let $f(x)$ and $\phi(y)$ denote the production functions of these lands. If m acres are cultivated by the old method and n acres by the new method, then the problem of maximizing total product is
> Maximize

$$mf(x) + n\varphi(y)$$

> subject to

$$m + n = Bmx + ny = A$$

where *B* is the total number of acres and *A* is the total amount of labour available. The first order conditions for a maximum are,

$$f'(x) = \varphi'(y) \text{ and } f(x) - xf'(x) = \varphi(y) - y\varphi'(y)$$

where the prime denotes differentiation. The first condition states that total product is maximized when the marginal product of labour is equal, under both methods and the second condition states the equality of rents per acre. Wicksell now observes that these are precisely the conditions achieved by pure competition and hence that Ricardo was wrong to claim that a diminution of gross product was possible. Modern readers will note that Wicksell assumes throughout the validity of interior maxima. Subject to this qualification, Wicksell's analysis is a considerable improvement on anything produced before him.

The problem just discussed considered labour and land as the only explicit factors of production. Even here, Wicksell feels that ‘It is scarcely possible to discover a simple and intelligible criterion which will indicate whether a change in the technique of production is in itself likely to raise or lower wages’. When Wicksell goes on to add capital as a factor of production, he has to concede that inventions may reduce the marginal product and share of labour. This leads him to say that ‘The capitalist saver is … fundamentally, the friend of labour, though the technical inventor is not infrequently its enemy’ (Wicksell, 1911, pp. 140, 143, 164).

A satisfactory treatment of the machinery question depends upon modelling the general equilibrium of an economy and of following its transition from an initial equilibrium to the new equilibrium after the introduction of machinery. Even today, such a treatment is by no means easily achieved. Perhaps the classical economists would have done best to accept the general benefits of machinery, subject to transitional difficulties, as expounded by economists such as Tucker and Stewart, and wait until the proper analytical tools to discuss the issue satisfactorily had been developed.

See Also

- Ricardo, David

- Tozer, John Edward

Bibliography

Archer, H. 1742. Second Parliament of George II: Fourth session (8 of 9, begins 7/4/1738). *The History and Proceedings of the House of Commons: Volume 10: 1737–1739* (1742), pp. 258–92. Online. Available at http://www.british-history.ac.uk/report.asp?compid=37804, accessed 19 July 2007.

Barton, J. 1817. *Observations on the Circumstances which Influence the Conditions of the Labouring Classes of Society*. London.

Cary, J. 1695. *A Discourse on Trade*. 3rd edn, London, 1745.

Defoe, D. 1704. *Giving Alms no Charity*. London.

Martin, H. 1701. *Considerations on the East-India Trade*. Reprinted in *A Select Collection of Scarce and Valuable Economic Tracts*, ed. J.R. McCulloch. London, 1856.

Mill, J.S. 1848. Principles of Political Economy with Some of Their Applications to Social Philosophy. In *Collected Works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press, 1970.

Ricardo, D. 1821. Principles of Political Economy and Taxation. 3rd edn. In *The Works and Correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Stewart, D. 1856. Lectures on Political Economy. In *Collected Works of Dugald Stewart*, vols 8 and 9, ed. Sir W. Hamilton. Edinburgh: Thomas Constable & Co., 1854–60.

Tozer, J.E. 1838. *A Mathematical Investigation of the Effect of Machinery*. ed. D. Collard. New York: A. M. Kelley, 1968.

Tucker, J. 1757. *Instructions for Travellers*. Repr. in R.L. Schuyler, *Josiah Tucker*. New York: Columbia University Press, 1931.

Wicksell, K. 1911. *Lectures on Political Economy*, 2nd edn. trans. E. Classen. London: G. Routledge & Sons, 2 vols, 1934–5.

How to cite this article

Rashid, S. "machinery question." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/

article?id=pde2008_M000005> doi:10.1057/9780230226203.1004

The New Palgrave Dictionary of Economics Online

Machlup, Fritz (1902 – 1983)

John S. Chipman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic integration; Haberler, G.; innovation; international monetary system; invention; Machlup, F.; multiplier analysis

Article

Fritz Machlup was born in Wiener Neustadt, south of Vienna, on 15 December 1902, and died in New York on 30 January 1983. He studied economics at the University of Vienna in the 1920s under Friedrich von Wieser and Ludwig von Mises, and wrote his doctoral dissertation on the gold-exchange standard (Machlup, 1925) under the latter. In the years 1922–32 he pursued a business career in the family cardboard-manufacturing partnership while continuing his intellectual interests in economics and philosophy of science in association with von Mises, Hayek, Haberler, Morgenstern, Felix Kaufmann and Alfred Schütz. During this period he wrote two more books including one (Machlup, 1931) dealing with the role of stock-market speculation in capital formation. As business conditions deteriorated in the 1930s he took leave from his partners to accept a Rockefeller fellowship, and spent 1933–35 in the United States. Upon receiving an appointment at the University of Buffalo in 1935 he liquidated his Austrian business interests, and following a brief stay in England began an academic career in the United States. He moved to Johns Hopkins in 1947, and to Princeton in 1960 to succeed Jacob Viner. At Hopkins he had a profound influence as a graduate teacher and in building up a first-rate graduate programme that achieved national prominence; a list of his students is contained in Machlup (1963), and tributes and testimonials from many of them will be found in Dreyer (1978). At Princeton he was extremely active in his direction of the International Finance Section. Upon his retirement in 1971 he resumed his active career at New York University until his death shortly after his 80th birthday. He was president of the Southern Economic Association (1959), the American Association of University Professors (1962–64), the American Economic Association (1966), and the International Economic Association (1971–4).

Machlup's two great areas of research were international monetary economics and industrial organization, the latter with special emphasis on the ‘knowledge industry’, an activity which began with

a study of the patent system (Machlup and Penrose, 1950; Machlup, 1958), continued with the development of a formal theory of invention, innovation, and the optimal lag of imitation behind innovation (see Bitros, 1976, pp. 439–502), and culminated in a monograph on the subject (1962), the multi-volumed second edition of which remained unfinished at the time of his death (Machlup, 1980b; 1982b; 1983). What was especially original in his contributions was his peculiar talent, resulting from his business background and study of philosophy of science, of being able to formulate a theory that took into account – in addition to the usual economic facts – the theories or rationalizations put forward by economic agents to justify their own actions. This was used to support his contention that economic agents engage in maximizing behaviour even though they may deny this. Such perceptions permeate his works on industrial organization (Machlup, 1949; 1952a; 1952b) and were developed in numerous articles collected in Machlup (1963).

Machlup's contributions to international economics were likewise characterized by a combination of clear logical thinking and intimate knowledge of the workings of economic institutions. His two-country extension of the theory of the multiplier (1943) was especially illuminating in bringing out the implicit financial assumptions of Keynesian theory. His work on the theory and policy of foreign-exchange markets and international economic adjustment (collected in Machlup, 1964) was very influential. His classic (1939–40) article developing Haberler's concepts of demand and supply of foreign exchange was required reading for a generation of graduate students. In his famous controversy with Sidney Alexander, while stressing the importance of relative prices he proved himself to be always the eclectic, never espousing one narrow ‘approach’ to the exclusion of all others. At first countering ‘elasticity pessimism’ and championing flexible exchange rates in his academic writings, he later became the prime architect of plans to reform the international monetary system in his organization of the ‘Bellagio group’ (Machlup and Malkiel, 1964). These activities have been recounted by Robert Triffin and John Williamson in Dreyer (1978). Machlup's last contributions to international economics included a series of penetrating analyses of the Eurodollar market (starting with Machlup, 1970) and his one foray into ‘real’ international trade, a work on the theory of economic integration (Machlup, 1977).

Machlup had a remarkable and unforgettable personality. He was brilliant and as sharp as a whistle in his keen analysis and grasp of economic issues; he was lucid and patient as a teacher, yet tough; he was charming and witty; he was a great music-lover and an avid sportsman to the end of his days. Above all he was a man of extraordinary energy and passion.

Most of Machlup's important articles have been reprinted in Machlup (1963; 1964) and Bitros (1976); the first and third of these contain bibliographies of his work. Further information concerning his life and work will be found in Dreyer (1978), Chipman (1979), Machlup (1980a; 1982a), and Haberler (1983). The latter concludes with an apt poetic tribute by Kenneth Boulding.

Selected works

1925. *Die Goldkernwährung*. Halberstadt: Meyer.

1931. *Börsenkredit, Industriekredit und Kapitalbildung*. Vienna: Julius Springer. Revised English edn, *The Stock Market, Credit and Capital Formation*, London: Hodge; New York: Macmillan, 1940.

1943. *International Trade and the National Income Multiplier*. Philadelphia: Blakiston.

1949. *The Basing-point System*. Philadelphia: Blakiston.

1950. (With E. Penrose.) The patent controversy in the nineteenth century. *Journal of Economic History* 10, 1–29.

1952a. *The Economics of Sellers' Competition*. Baltimore: Johns Hopkins Press.

1952b. *The Political Economy of Monopoly*. Baltimore: Johns Hopkins Press.

1958. *An Economic Review of the Patent System*. Study of the Subcommittee on Patents, Trademarks, and Copyrights of the Committee on the Judiciary, US Senate, Study No. 15. Washington, DC: Government Printing Office.

1962. *The Production and Distribution of Knowledge in the United States*. Princeton: Princeton University Press.

1963. *Essays on Economic Semantics*. Englewood Cliffs, NJ: Prentice-Hall.

1964. (With G. Malkiel, eds.) *International Monetary Arrangements: The Problem of Choice*. Princeton: International Finance Section, Department of Economics, Princeton University.

1964. *International Payments, Debts, and Gold*. New York: Scribner's. 2nd edn, New York: New York University Press, 1976.

1970. Euro-dollar creation: a mystery story. *Banca Nazionale del Lavoro Quarterly Review* 23, 219–60. Reprinted in Bitros (1976).

1977. *A History of Thought on Economic Integration*. New York: Columbia University Press.

1980a. My early work on international monetary problems. *Banca Nazionale del Lavoro Quarterly Review* 35, 115–46.

1980b, 1982b, 1983. *Knowledge: Its Creation, Distribution, and Economic Significance*. Vol. 1: *Knowledge and Knowledge Production*. Vol. 2: *The Branches of Learning*. Vol. 3: *The Economics of Information and Human Capital*. Princeton: Princeton University Press.

1982a. My work on international monetary problems, 1940–1964. *Banca Nazionale del Lavoro Quarterly Review* 35, 3–36.

Bibliography

Bitros, G. (ed.) 1976. *Selected Economic Writings of Fritz Machlup*. New York: New York University Press.

Chipman, J.S. 1979. Machlup, Fritz. In *International Encyclopedia of the Social Sciences*, vol. 18. New York: Free Press.

Dreyer, J.S. (ed.) 1978. *Breadth and Depth in Economics: Fritz Machlup – The Man and His Ideas*. Lexington, MA: D.C. Heath.

Haberler, G. 1983. Fritz Machlup zum Gedenken. *Neue Zürcher Zeitung*, 16 February, 19. Fritz Machlup: in memoriam. *Cato Journal* 3, Spring, 11–14.

How to cite this article

Chipman, John S. "Machlup, Fritz (1902–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000006> doi:10.1057/9780230226203.1005

The New Palgrave Dictionary of Economics Online

Macleod, Henry Dunning (1821 – 1902)

Murray Milgate and Alastair Levy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

banking system; bimetallism; catallactics; Gresham's Law; Macleod, H. D.

Article

Macleod warrants special mention in this *Dictionary* if only because in the late 1850s and early 1860s he undertook to produce single-handedly a dictionary of economics on a grand scale – and, what is more, one to which he was to be the sole contributor. In the event the task proved to be beyond him, as it would for any mortal, and all that appeared was the first volume covering the letters A–C. Macleod never held an academic appointment, though he applied unsuccessfully for chairs at Cambridge (1863), Edinburgh (1871), and Oxford (1888).

Macleod, the son of a Scottish landholder, was born in Edinburgh. After graduation from Cambridge (BA, Trinity, 1843) and admission to the Bar (1849), he wrote a report on the administration of poor relief in the nine local parishes of the district of Easter Ross in Scotland (1851). This report led directly to the establishment of a poor-house under Scotland's first Poor Law Union. In 1854, he joined the Royal British Bank and wrote a memorandum and opinion on that bank's legal position under the Joint Stock Banking Act of 1845. This first excursion into financial matters stimulated him to study the literature of economics on the subject, but he found that

> for the purpose of describing the actual principles and mechanisms of commerce they [Smith, Ricardo and Mill] were absolutely worthless. … I saw that the greatest opportunity that had come to any man since Galileo had come to me, and I then determined to devote myself to the construction of a real science of Economics on the model of the already established physical sciences. (1896b, pp. 142–3)

To his credit he stuck fast to his task. His detractors, however, have passed harsh judgement on its results (see, for example, the assessment of him in Higgs's edition of *Palgrave's Dictionary*); his sympathetic readers have been more generous (see, for example, Hayek, 1933). Given the sheer

magnitude of his project, there would seem to be more to be said for the position of the critics.
Macleod's employment in the banking system led to what is perhaps his most important book on that subject: *The Theory and Practice of Banking* (1855–6). Its two most interesting features for the modern reader are, perhaps, the discussion of discount policy and the insistence on the proposition that ‘the distinction between capital and currency … is of the most profound delusions that ever existed’ (1855, vol. 2, p. lxxii; see also the entry on ‘Credit’ in his *Dictionary*). Not surprisingly, for one who kept fast to the basic position of the Bullion Report, this latter notion introduced a number of ambiguities into the argument. However, not withstanding these peculiarities, the book was apparently quite successful, going through five editions by the 1890s and being reprinted soon after his death. Charles Rist referred to it as Macleod's ‘great book’ (1940, p. 261).
There followed many publications on monetary matters of which two may be singled out. In *Bimetallism* (1894), he criticized the proponents of a dual standard for advocating ‘an impossibility’; a position which put him at odds with many of his contemporaries. This polemic was continued in two short tracts issued by the Gold Standard Defence Association in 1895 under the titles ‘Gresham's Law’ and ‘Bimetallism in France’. Secondly, in 1898, he published two contributions to the debate surrounding the Fowler Commission on Indian currency arrangements: *Indian Currency* and *A Tentative Scheme for Restoring a Gold Currency to India*.
Macleod's project of reconstructing economic science continued on more general matters with *Elements of Political Economy* (1858). The book is interesting as an example of Macleod's advocacy of a definition of economics as the ‘science of exchanges’, or catallactics, which Marshall claimed ‘anticipated much both of the form and substance of recent criticisms on the classical doctrine of value in relation to cost, by Profs. Walras and Carl Menger’ (1920, p. 821), and for the fact that in it he introduces into the vocabulary of economics the phrase ‘Gresham's Law’.
One of Macleod's interesting habits was that of publishing the same material in different forms and under different titles. In this he reminds one of McCulloch. Thus, *Bimetallism* was itself an expanded version of the seventh chapter of his *Theory of Credit* (1889–91), his *Elements of Political Economy* (1858) appeared in successive editions under the titles *The Principles of Economic Philosophy* (1872–5) and *The Elements of Economics* (1881–6), and his *History of Economics* (1896b) seems to be made up of material from his unfinished dictionary.
Macleod died at Norwood on 16 July 1902.

See Also

- catallactics

Selected works

1851. *The Results of the Operation of the Poorhouse System in Ross*. Inverness: Courier Office, repr. from the *Inverness Courier*.

1855–6. *The Theory and Practice of Banking; with the Elementary Principles of Currency, Prices, Credit and Exchanges*. 2 vols. London: Longman, Brown, Green and Longmans.

1858. *The Elements of Political Economy*. London: Longman, Brown, Green, Longmans and Roberts.

1863. *A Dictionary of Political Economy: Biographical, Bibliographical, Historical and Practical: A–C*. London: Longmans, Green, Longman, Roberts, and Green.

1889–91. *The Theory of Credit*. 2 vols. London: Longmans, Green & Co.

1894. *Bimetallism*. London: Longmans, Green & Co.

1896a. *A History of Banking in all the Leading Nations*. 2 vols. London: Bliss, Sands & Co.

1896b. *The History of Economics*. London: Bliss, Sands & Co.

1898. *Indian Currency*. Longmans: Green & Co.

Bibliography

Hayek, F.A. 1933. Henry Dunning Macleod. In *Encyclopedia of the Social Sciences*, vol. 10, ed. E.R.A. Seligman. New York: Macmillan.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Rist, C. 1940. *History of Monetary and Credit Theory from John Law to the Present Day*. Trans. J. Degras from the French edn of 1938, New York: Macmillan; London: G. Allen & Unwin. Reprinted, New York: A.M. Kelley, 1966.

How to cite this article

Milgate, Murray and Alastair Levy. "Macleod, Henry Dunning (1821–1902)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000007> doi:10.1057/9780230226203.1006

The New Palgrave Dictionary of Economics Online

macroeconomic effects of international trade

Reuven Glick

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

International trade can affect the macroeconomy by helping to transmit disturbances from one economy to another and by muting or amplifying the impact of fiscal and monetary policies on economic activity. Representative open economy macro models are discussed, highlighting the role different theoretical features play in influencing the channels through which trade flows can have macro effects.

Keywords

Balassa–Samuelson effect; beggar-thy-neighbour; capital mobility; comparative advantage; consumption correlations puzzle; contagion; currency unions; elasticities approach to the balance of payments; elasticity of substitution; endogenous growth models; exchange rate puzzles; exchange rate regimes; externalities; Feldstein–Horioka puzzle; flexible exchange rates; general equilibrium; imperfect competition; income–expenditure models; income multiplier; inflation; international real business cycles; international trade; Keynesian model; macroeconomic effects of international trade; Marshall–Lerner–Robinson condition; microfoundations; monopolistic competition; Mundell–Fleming model; national income; new open-economy macroeconomics; production sharing; purchasing power parity; research and development; specialization; spillovers; technological progress; terms of trade; trade costs; trade frictions

Article

The field of open economy macroeconomics deals with the macroeconomic behaviour of economies that trade with each other. International trade can have macroeconomic effects by helping the transmission of disturbances from one economy to another as well as by affecting the impact of macroeconomic policies on economic activity. This article discusses several representative open economy macro models, highlighting the role different theoretical features play in influencing the channels through which trade flows can have macro effects.

Keynesian framework

At its simplest level, international trade is linked to macroeconomic activity through the national income relation. Consider the Keynesian income–expenditure model of a small open economy, in which prices and the interest rate are given, foreign demand for exports is exogenous, and domestic output is determined by demand. With these assumptions, an exogenous increase in domestic expenditures raises domestic income and worsens the current account balance; however, income rises less than in a closed economy because of leakages from the income stream through imports and through saving. In contrast, an exogenous increase in foreign demand for domestic goods leads to an increase in both exports and domestic income. Because the increased direct demand for exports is only partially offset by the expansion of imports induced by higher income, the current account improves overall. The resulting rise in domestic output implies positive cross-country transmission of the foreign disturbance.

Income multiplier effects through changes in trade also characterize open economy extensions of the Keynesian framework, such as the classic Mundell–Fleming model. This model also takes prices as given, but allows the income effects of monetary stimulus and exogenous expenditure changes to take account of interest rate changes depending on the degree of international capital mobility and of exchange rate changes, which in turn depend on the exchange rate regime. With a flexible exchange rate regime, exchange rate changes affect the relative demand for domestic and foreign goods. Thus, for example, domestic monetary stimulus that reduces the interest rate, raises income, and creates an excess demand for foreign exchange also depreciates the domestic currency. If the Marshall–Lerner–Robinson condition is satisfied, that is, the sum of price elasticities of domestic and foreign demands for imports exceeds unity, then the lower relative price of domestic goods switches demand from foreign to domestic goods and raises the current account balance, causing domestic income to increase and foreign income to decrease. Accordingly, the domestic income multiplier effect of the monetary stimulus is augmented by the expenditure-switching effect of the exchange rate; in addition, the trade transmission effect of domestic monetary shocks to foreign income is negative.

In these models crucial parameters affecting transmission effects include the marginal propensity to import and the elasticity of trade with respect to the exchange rate. Thus, for example, an increase in the marginal propensity to import out of income lessens the multiplier effects of domestic policy stimulus.

New open economy macro models

New open-economy macroeconomic models (NOEM) integrate older fixed-price Keynesian models of macroeconomic fluctuations with dynamic intertemporal analysis based on microeconomic foundations and optimizing agents. These models embed imperfect competition and short-run nominal rigidities in a general equilibrium framework and provide clear welfare criteria in the form of the utility of the representative consumer. They also assume that bond (but not equity) markets are integrated, providing a consumption-smoothing role for net trade flows via the current account. Thus, for example, a temporary productivity shock that raises domestic output induces higher saving and a temporary current account surplus (though with investment dynamics a current account deficit may result if the increase in investment exceeds the increase in saving).

In a seminal paper, Obstfeld and Rogoff (1995) use a two-country framework in which each country specializes in producing a subset of tradable goods, and domestic and foreign consumers have identical preferences over a basket of both domestic and foreign goods. They show that monetary shocks have a positive effect on domestic output and a negative transmission effect on foreign output, as in the Mundell–Fleming model. Because monetary stimulus depreciates the domestic currency, it lowers the domestic country's terms of trade, reduces the purchasing power of domestic residents and raises the purchasing power of foreign residents. This terms-of-trade effect makes foreign residents better off and domestic residents worse off, but not by enough to offset the domestic gains from greater output. A temporary current account surplus is generated as well via the intertemporal consumption-smoothing channel.

A key parameter in NOEM models is the elasticity of substitution between goods embedded in consumer preferences. Obstfeld and Rogoff assume that the elasticity of substitution between goods produced in the same country is the same as the elasticity of substitution between goods produced in different countries. Several papers show how the international transmission of shocks is affected by relaxing this assumption. Tille (2001) shows that, if the elasticity of substitution of domestic and foreign goods exceeds unity, the Marshall–Lerner–Robinson condition holds. In this case, a currency depreciation and decline in the terms of trade results in a large demand switch towards domestic goods and a rise in export revenue. Tille also shows that, if there is less substitutability between domestic and foreign goods across countries than within countries (the empirically more relevant case), the terms-of-trade effect of domestic monetary expansion may be large enough to lower domestic welfare (termed a ‘beggar-thyself’ effect), while raising foreign welfare. In contrast, greater fiscal expenditures on domestic output raise the domestic terms of trade and domestic welfare, while reducing relative demand for foreign goods and foreign welfare (a ‘beggar-thy-neighbour’ effect), particularly when domestic and foreign goods are poor substitutes.

Corsetti and Pesenti (2001) deal with the special case in which the elasticity of substitution between domestic and foreign goods is unity, implying constant expenditure shares on domestic and foreign goods. This specification implies that the current account is always in balance. The reason is that, with unit elasticity between domestic and foreign goods, an increase in the foreign price of foreign goods results in a proportionate decrease in the quantity of foreign demand for domestic goods, leaving expenditures on exports constant and the current account unaffected.

Other extensions to NOEM models that affect the transmission of policy include consumption bias for domestic over foreign goods (Warnock, 2003), pricing-to-market behaviour (Betts and Devereux, 1998), and non-traded distribution services (Burstein, Eichenbaum and Rebelo, 2006).

International real business cycle models

The tendency of macro aggregates, such as output, to move together in different countries is well documented (Backus, Kehoe and Kydland, 1992; Baxter, 1995). Cross-country business cycle correlations depend on the interaction of common international shocks, country-specific shocks, and the transmission of these shocks between countries. An important question in international macroeconomics is how much these comovements reflect the transmission of shocks across borders through international trade linkages. International real business cycle (IRBC) models analyse this issue within a dynamic

general equilibrium framework based on microfoundations. Unlike NOEM models, these models typically assume flexible prices and complete markets, though more recent work has introduced price rigidity and incomplete asset markets.

On theoretical grounds, the effect of international trade links on the comovement of national business cycles is ambiguous. On the one hand, greater integration can increase intra-industry specialization and production-sharing because of low elasticity of substitution between intermediate inputs produced in different countries; in addition, it may allow demand shocks to propagate more easily across national borders, which may lead to a higher correlation of business cycles when countries trade more. On the other hand, greater trade integration can increase inter-industry specialization if countries specialize more in the goods in which they have a comparative advantage in order to achieve gains from trade; this case, if industry-specific shocks are a dominant source of business cycle movements, may lead to a lower correlation of business cycles when countries trade more.

On balance, the empirical evidence suggests that the former effect dominates, and that countries with a lot of bilateral trade tend to have more synchronized business cycles (for example, Frankel and Rose, 1998; Baxter and Kouparitsas, 2005). However, since the early 1980s business cycle synchronization has not in fact increased among industrial countries despite increasing trade integration. Stock and Watson (2005) provide a partial explanation by showing that common international shocks experienced by G-7 countries have been smaller in the 1980s and 1990s than they were in the 1960s and 1970s. But they also show that cyclical comovements have increased for subgroups of countries, notably within Europe and North America. Burstein, Kurz and Tesar (2005) construct a model that is consistent with this development in which trade between core countries and their periphery (for example, the United States and Canada) involves more production sharing than does trade between core regions (for example, the United States and Europe). Consequently, one should observe higher output correlations between core and peripheral countries than between core regions. IRBC models have been less successful in explaining the quantitative magnitude of the relation between trade intensity and the cross-country correlation of business cycles; that is, a given change in bilateral trade intensity generates a much smaller change in output correlations than is apparent in the data; this is referred to as the ‘trade comovement gap puzzle’ (Kose and Yi, 2006).

The finding that greater trade intensity is associated with greater cross-country comovements in business cycles suggests that these comovements depend on policies that enhance international trade, such as lowering of trade barriers or reductions in exchange rate costs due to membership in currency unions. Frankel and Rose (2002) find that the positive effect of currency unions on trade in turn has a large effect on output in member countries. Since the main cost of joining a currency area is the cost of giving up monetary independence, this has the implication that a pair of countries with business cycles that are dissimilar *ex ante* (making the act of joining a currency union appear costly) might have more correlated business cycles *ex post* because the increase in trade stimulated by the currency union tends to synchronize business cycles.

Trade frictions and macro models

The international tradability of goods depends not just on the degree of substitutability in consumption, but also on transport costs and other trade frictions. In fact, Obstfeld and Rogoff (2000) argue that

introducing real trade costs helps explain a variety of puzzles in international economics, including the low cross-country correlation of consumption (consumption correlations puzzle), the limited magnitude of current account imbalances (Feldstein–Horioka puzzle), international price discrepancies (purchasing power parity puzzle), and home bias in trade and asset holdings.

Taken to the extreme, trade frictions play a role in explaining why some goods may not be traded at all. While open economy macroeconomics by definition analyses trade across national borders, the field has long found it useful to assume that a given exogenous set of goods is non-traded. This traded/non-traded distinction is essential to many well-known results in the field, such as the Balassa–Samuelson effect, which says that, as the productivity of traded goods rises relative to that of non-traded goods, there will be tendency for the real exchange rate to appreciate.

The international trade literature has explained non-tradedness as an outcome of trade frictions. For example, Dornbusch, Fischer and Samuelson (1977) show how a range of non-traded goods can arise in the presence of cross-country trade costs within a model in which differences in labour productivity across a continuum of goods determine the range of goods a country produces as well as the pattern of trade.

A growing field of international economics research tries to integrate models of trade and macroeconomics and treats the set of tradable goods not as exogenously given but rather as an endogenously determined characteristic of the analysis. Several authors (Ghironi and Melitz, 2005; Bergin, Glick, and Taylor, 2006) formulate open economy macro models with monopolistic competition and heterogeneously productive firms, in which firms face fixed costs of selling in domestic and export markets, to explain phenomena such as the Balassa–Samuelson effect. Since only relatively more productive firms are profitable enough to engage in trade, they endogenously satisfy the precondition of the Balassa–Samuelson story that productivity gains are concentrated in the traded goods sector.

Loose ends

International trade can influence macroeconomic activity through other channels. For example, as highlighted in endogenous growth models, technological progress may depend on incentives to undertake R&D and innovate, which, in turn, may depend on externalities or spillover effects from greater markets provided by international trade (Grossman and Helpman,1991). Greater openness to trade can also complicate the optimal conduct of monetary policy because of the impact of the exchange rate on real activity and inflation. Clarida, Gali and Gertler (2001) show how more openness to international trade can influence a central bank following an optimal policy feedback rule to raise the domestic interest rate more aggressively in response to inflation pressures. Lastly, trade may serve as a transmission channel through which financial crises may spread contagiously across countries (Glick and Rose, 1999).

See Also

- growth and international trade
- international real business cycles

- international trade and heterogeneous firms
- international trade theory
- Marshall–Lerner condition
- new open economy macroeconomics
- trade costs
- tradable and non-tradable commodities

Bibliography

Backus, D., Kehoe, P. and Kydland, F. 1992. International real business cycles. *Journal of Political Economy* 100, 745–75.

Baxter, M. 1995. International trade and business cycles. In *Handbook of International Economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: North-Holland.

Baxter, M. and Kouparitsas, M. 2005. Determinants of business cycle comovement: a robust analysis. *Journal of Monetary Economics* 52, 113–57.

Bergin, P., Glick, R. and Taylor, A. 2006. Productivity, tradability, and the long-run price puzzle. *Journal of Monetary Economics* 53, 2041–66.

Betts, C. and Devereux, M. 1998. Exchange rate dynamics in a model of pricing to market. *Journal of International Economics* 50, 215–44.

Burstein, A., Eichenbaum, M. and Rebelo, S. 2006. The importance of nontradable goods' prices in cyclical real exchange rate fluctuations. *Japan and the World Economy* 18, 247–53.

Burstein, A., Kurz, C. and Tesar, L. 2005. International trade, production sharing and the transmission of business cycles. Working paper, University of California at Los Angeles.

Clarida, R., Gali, J. and Gertler, M. 2001. Optimal monetary policy in open versus closed economies: an integrated approach. *American Economic Review* 91, 248–52.

Corsetti, G. and Pesenti, P. 2001. Welfare and macroeconomic interdependence. *Quarterly Journal of Economics* 116, 421–45.

Dornbusch, R., Fischer, S. and Samuelson, P. 1977. Comparative advantage, trade, and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67, 823–39.

Frankel, J. and Rose, A. 1998. The endogeneity of the optimum currency area criteria. *Economic Journal* 108, 1009–25.

Frankel, J.A. and Rose, A. 2002. An estimate of the effect of common currencies on trade and income. *Quarterly Journal of Economics* 117, 437–66.

Ghironi, F. and Melitz, M. 2005. International trade and macroeconomic dynamics with heterogenous firms. *Quarterly Journal of Economics* 120, 865–915.

Glick, R. and Rose, A. 1999. Contagion and trade: why are currency crises regional? *Journal of International Money and Finance* 18, 603–17.

Grossman, G. and Helpman, E. 1991. *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.

Kose, M. and Yi, K.-M. 2006. Can the standard international business cycle model explain the relation between trade and comovement? *Journal of International Economics* 68, 267–95.

Obstfeld, M. and Rogoff, K. 1995. Exchange rate dynamics redux. *Journal of Political Economy* 103, 624–60.

Obstfeld, M. and Rogoff, K. 2000. The six major puzzles in international macroeconomics: is there a common cause? In *NBER Macroeconomics Annual*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Stock, J. and Watson, M. 2005. Understanding changes in international business cycle dynamics. *Journal of the European Economic Association* 3, 968–1006.

Tille, C. 2001. The role of consumption substitutability in the international transmission of monetary shocks. *Journal of International Economics* 53, 421–44.

Warnock, F. 2003. Exchange rate dynamics and the welfare effects of monetary policy in a two-country model with home-product bias. *Journal of International Money and Finance* 22, 343–63.

How to cite this article

Glick, Reuven. "macroeconomic effects of international trade." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000401> doi:10.1057/9780230226203.1007

The New Palgrave Dictionary of Economics Online

macroeconomic forecasting

Mark W. Watson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank of England; dynamic adjustment; dynamic stochastic general equilibrium models; expectations; Federal Reserve System; forecasting; leading indicators index; macroeconomic forecasting; mean squared forecast error; Mitchell, W.C.; probability theory; structural econometric models; time series analysis; vector autoregressions

Article

Macroeconomic forecasts are ‘guesses’ of the future values of important macroeconomic aggregates such as GDP, inflation, or the unemployment rate. These forecasts inform the decisions of business, policymakers, investors, and consumers. Macroeconomic forecasts are regularly constructed by government agencies and private companies. For example, every quarter the Bank of England publishes its *Inflation Report*, which contains forecasts of inflation over the next three years. Federal Reserve policymakers also rely on forecasts from the Green Book; however, unlike the Bank of England, the Fed does not release its forecasts to the public. The Federal Reserve Bank of Philadelphia summarizes private sector macroeconomic forecasts for the United States in its quarterly *Survey of Professional Forecasters*.

Macroeconomic forecasts are constructed using a variety of methods. These methods can be grouped into four categories: (1) leading indicator indexes; (2) structural econometric models; (3) time series models; and (4) judgement.

The origin of leading indicator indexes can be traced to the 1930s when, at the request of the US Secretary of the Treasury, Wesley Mitchell proposed a set of variables that historically had moved in anticipation of the business cycle. Averages of these leading indicators are an index of leading indicators. Such an index was constructed in the United States for several years by the Department of Commerce and is now maintained and published monthly by the Conference Board, which also publishes leading indicator indexes for several other countries.

Structural econometric models construct forecasts using dynamic relationships suggested by economic theory and estimated by statistical methods. Work on these models by Tinbergen, Klein and Haavelmo

resulted in Nobel prizes for these researchers in 1969, 1980 and 1989 respectively. Large-scale structural models with hundreds of equations were developed in the 1960s and early 1970s, but forecast failures in the 1970s led researchers to question both the economic theory used in the models and the statistical procedures used to fit the models' equations. Refinements in theory (notably the importance of expectations and dynamic adjustment) and statistical methods (notably time series methods) are incorporated in the current generation of large-scale structural models. Currently, there is a significant research effort aimed at constructing small-scale structural models ('dynamic stochastic general equilibrium' models) for policy evaluation and forecasting.

Time series models use serial correlation (or persistence) in variables to construct forecasts. For example, a simple autoregressive model (AR) has the form $y_t = \alpha + \varphi y_{t-1} + \varepsilon_t$, where y_t is an economic variable of interest and ε_t is a zero-mean serially uncorrelated random shock. When ϕ is positive (negative), larger than average values of y_{t-1} tend to be associated with larger (smaller) than average values of y_t. Thus, an autoregressive forecast of y_{T+1} using data through time T is $y_{T+1/T} = \alpha + \varphi y_T$. Many macroeconomic variables display short-run dependence, and time series models typically produce more accurate short-run forecasts than other forecasting methods. Time series models have been developed to construct forecasts based on linear and nonlinear dependence properties in macroeconomic variables, and multivariate time series models, such as vector autoregressions (VARs), are widely used for short-horizon macroeconomic forecasting.

Professional forecasters also rely on judgement when constructing their forecasts. That is, while the macroeconomic forecasts published by the Bank of England or the Fed's Green Book forecasts rely on econometric models, the forecasts are not identical to model-based forecasts. Professional forecasters typically use judgement to adjust model-based forecasts. These adjustments – sometimes called 'add-factors' – allow forecasters (so they argue) to incorporate information that is not captured in the economic model. As an empirical matter, good judgement appears to improve the accuracy of model-based forecasts.

Much of the theory of forecasting can be derived from elementary concepts in probability theory. Let y_{T+1} denote the variable to be forecast and X_T denote a set of variables to be used for constructing the forecast. In general, X_T will include y_T, y_{T-1}, and longer lags, as well as current and lagged values of other series. Let $g(X_T)$ denote the forecast or 'guess' of y_{T+1} constructed from X_T, where good choices of $g(\,.\,)$ lead to more accurate forecasts. The forecast error is $e_{T+1} = y_{T+1} - g(X_T)$, and accuracy can be measured by mean squared forecast error (MSFE), where the conditional $\mathrm{MSFE} = E(e_{T+1}^2|X_T)$. A fundamental result from probability theory is that $E(e_{T+1}^2|X_T)$ is minimized using $g(X_T) = E(y_{T+1}|X_T)$; that is, the regression (conditional expectation) produces the minimum mean squared forecast error.

A key implication of this theoretical result is that more information is always better – that is, it never hurts to include more variables in X_T, and the information in these additional variables will often reduce the MSFE. But, this result assumes that the regression function $E(y_{T+1}|X_T)$ is known, and in practice this function must be estimated using sample data. Including many variables in X_T means that many parameters must be estimated to characterize the regression function, and estimating a large number of

parameters leads to statistical estimation error that increases the MSFE. This trade-off between including more variables in X_T to capture more information about y_{T+1} and the increased statistical error associated with estimating additional parameters for the forecasting model is one of the major practical problems in forecasting.

Another major problem is the temporal stability of the forecasting model. That is, the regression $E(y_{T+1}|X_T)$ might change over time, so that a regression estimated using past data might provide poor forecasts for future values of y_T. These two problems – developing methods for forecasting using many past variables and problems associated with instability – are active areas of current research. The relevant chapters in Elliott, Granger and Timmerman (2006) summarize current research on these and other important topics in economic forecasting.

See Also

- time series analysis
- vector autoregressions

Bibliography

Bank of England. 2007. *Inflation Report*. Online. Available at http://www.bankofengland.co.uk/publications/inflationreport/index.htm, accessed 19 February 2007.

Elliott, G., Granger, C.W.J. and Timmerman, A.C., eds. 2006. *Handbook of Economic Forecasting*, vol. 1. Amsterdam: North-Holland.

Federal Reserve Bank of Philadelphia. 2007. *Survey of Professional Forecasters*. Online. Available at http://www.phil.frb.org/econ/spf, accessed 19 February 2007.

How to cite this article

Watson, Mark W. "macroeconomic forecasting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000376> doi:10.1057/9780230226203.1008

The New Palgrave Dictionary of Economics Online

macroeconomics, origins and history of

Robert W. Dimand

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Macroeconomics, the analysis of economic aggregates, became a recognized field with Keynes's *General Theory* (1936) and its mathematical and diagrammatic reformulations, and the macroeconometric modelling pioneered by Tinbergen and Frisch. Macroeconomics grew out of two long-standing traditions: business cycle analysis from Jevons and Juglar to Mitchell, and monetary theory, building on the work of Hume, Thornton, Ricardo, Wicksell, and Fisher, supplemented by the circular flow analysis of Quesnay and Marx.

Keywords

Austrian economics; Bank of England; Banking School, Currency School, Free Banking School; Bauer, O.; Bergmann, Eugen von; Beveridge, W. H.; bimetallism; Bodin, J.; Burns, A. F.; business cycles; calculus of variations; Cantillon, R.; capital accumulation; central banking; Champernowne, D. G.; circular flow; Cobb–Douglas functions; consumption smoothing; covered interest parity; Cowles Commission; crowding out; deflation; depressions; Dunlop, J. T.; endogenous growth; equation of exchange; excess demand and supply; exploitation; Fel'dman, G. A.; Fisher ideal index; Fisher, I.; Fleming, J. M.; Friedman, M.; Frisch, R. A.; fundamental uncertainty; German hyperinflation; Gervaise, I.; gold standard; growth theory; Haavelmo, T.; Hammarskjold, D.; Harrod, R. F.; Hawtrey, R. G.; Hayek, F. A.; Hicks, J. R.; hoarding; Hume, D.; income–expenditure analysis; increasing returns; index numbers; index of leading indicators; inflation; innovations; Institute of World Economics (Germany); insurable risk; international trade, theory of; involuntary unemployment; IS–LM model; Jevons, W. S.; Jones, E. D.; Juglar, C.; Kaldor, N.; Kalecki, M.; Keynes, J. M.; Keynesian Revolution; Keynesianism; Kock, K.; Klein, L. R.; Kondratieff, N.; Koopmans, T. C.; labour theory of value; Lange, O. R.; Laspeyres indexes; law of markets; Lerner, A. P.; Lindahl, E. R.; liquidity preference; Locke, J.; Lundberg, E. F.; Luxemburg, R.; macroeconomics, origins and history of; Malthus, T. R.; Marshall, A.; Marx, K. H.; Meade, J. E.; mercantilism; Mill, J.; Mill, J. S.; Mises, L. E. von; Mitchell, W. C.;

Modigliani, F.; monetarism; money illusion; Mundell, R.; Myrdal, G.; National Bureau of Economic Research; National Institute of Economic and Social Research (UK); natural rate and market rate of interest; Neoclassical; new classical macroeconomics; New Deal; new Keynesian macroeconomics; Ohlin, B. G.; Paasche index; Patinkin, D.; Periodogram; Phillips curve; Physiocracy; Pigou, A. C.; post Keynesian economics; Price Revolution; profit sharing; public works; purchasing power parity; quantity theory of money; Quesnay, F.; Ramsey, F. P.; recessions; Reddaway, B.; representative agent; Ricardo, D.; Robbins, L. C.; Robertson, D. H.; rules versus discretion; Samuelson, P. A.; Say, J.-B.; Say's Equality; Say's Identity; Say's Law; Simons, H. C.; simultaneous equations models; Slutsky, E.; Smith, A.; specie-flow mechanism; spending multiplier; stabilization policy; Sunspots; Tarshis, L.; Taylor, F. I.; Thornton, H.; time series analysis; Timlin, M.; Tinbergen, J.; Tobin, J.; Tooke, T.; trade cycle theory; uncovered interest parity; underconsumptionism; Veblen, T.; vector autoregressions; velocity of circulation; Walras's Law; Warburton, C.; Wicksell, J. G. K.; Wilson, E. B.; Young. A. A.

Article

Macroeconomics analyses a whole economy or economies, dealing with aggregate output and employment, the price level and interest rate, rather than with the prices or quantities of particular commodities. It became a recognized field as textbooks and course offerings responded to John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936; 1971–89, vol. 7), to the mathematical and diagrammatic reformulations of Keynes by David Champernowne, Brian Reddaway, Roy Harrod, J. R. Hicks, James Meade, Oskar Lange, Mabel Timlin and Franco Modigliani (Hicks, 1937; Young, 1987), and to the first aggregate econometric models such as Tinbergen (1939). Ragnar Frisch (1933) introduced the terms ‘macrodynamics’ and ‘macroanalysis’, and his distinction between macroanalysis and microanalysis is the same as the subsequent distinction between macroeconomics and microeconomics. Michal Kalecki (1935) first used ‘macrodynamic’ in a title, and by the time that Lawrence Klein (1946) used ‘macroeconomics’ in the title of a journal article, he presumed that its meaning would be clear to his readers. But just as Molière's *bourgeois gentilhomme* spoke prose long before he knew he was doing so, economists wrote macroeconomics long before they called it by that name. Macroeconomics grew out of two long-standing traditions within economics: business cycle analysis and the theory of money.

Macroeconomic themes in pre-classical and classical political economy

The quantity theory of money is the oldest surviving theory in economics, yet remained, in David Laidler's (1991a) phrase, ‘always and everywhere controversial’ (primarily over whether changes in the quantity of money are exogenous or endogenous). Holding that a change in the money supply will ultimately change prices in the same proportion, the quantity theory was first used in the 16th century by Martin Navarro de Azpilcueta (writing in Latin as Navarrus) and other scholastics at the University of Salamanca (Grice-Hutchinson, 1952), and then by Jean Bodin in France, to explain the ‘Price Revolution’, the inflation following the inflow of silver from the Spanish colonies in the New World.

John Locke, Richard Cantillon and Isaac Gervaise contributed to understanding the velocity of circulation and the adjustment of international payments (Vickers, 1959). The economic essays in David Hume's *Political Discourses* (see Hume, 1752) mark a high-point of pre-classical monetary economics (see Humphrey, 1993). Hume's analysis of the specie-flow mechanism of adjustment under the gold standard showed that an increase in the quantity of gold in one country would increase prices and spending in that country, causing a trade deficit and gold inflow until balance of payments of equilibrium was restored with the world's gold distributed among countries in proportion to their demand for real money balances. Hume's specie-flow mechanism provided a crushing rejoinder to mercantilist schemes for increasing the amount of gold in a country by promoting exports and restricting imports. Such tariffs, quotas and subsidies would distort resource allocation without producing a lasting trade surplus, and would raise prices rather than the real wealth of a nation. Hume recognized that an increased money supply would provide a temporary stimulus to real output, which would fade as prices and wages adjusted. While Hume linked each country's price level to that country's money stock and emphasized relative price effects on trade balances, his younger contemporary and friend Adam Smith anticipated the monetary approach to the balance of payments by assuming purchasing power parity (with the world price level set by the world gold stock and world demand for real money balances) with adjustment taking place, not through relative price changes, but through the direct effect of a nation's excess demand for or supply of money on spending, hence on the balance of payments and on the country's stock of gold.

Keynes's *General Theory* revived interest in the debate in the years after the Napoleonic Wars about the possibility of a general glut of commodities. Keynes deplored the victory of David Ricardo's sharper analysis and endorsement of Say's (or James Mill's) Law of Markets over what Keynes regarded as Thomas Robert Malthus's deeper (but fuzzier) insight that insufficient effective demand could result in an excess supply of labour without an excess demand for any good (other than money). Malthus's insight was obscured by his failure to distinguish between a decision to save and a decision to invest, and hence to see the significance of hoarding. Statements of the Law of Markets by classical economists were more varied and complex, often subtler, and sometimes confused and contradictory than Keynes suggested in short quotations from the classics, which sometimes misled when taken out of context (see Link, 1959, and Corry, 1962, on the macroeconomics of English classical economists and their critics, and Sowell, 1972, on Say's Law). John Stuart Mill and others searched for a statement of the Law of Markets that would the be stronger truism that Oskar Lange later labelled as Say's Equality (if each and every commodity market is in equilibrium, then the sum of excess demand over all commodity markets much add to zero) but weaker than what Lange called Say's Identity, that excess demand for all commodity markets (that is, all markets except money) always sums to zero for any set of prices, regardless of whether any individual market is in equilibrium. Say's Identity, taken together with the adding up of budget constraints that Lange termed Walras's Law, implies that the money market always clears for any prices, leaving the absolute level of prices indeterminate. The policy implications that classical economists drew from their analysis are also more varied and pragmatic than the later textbook caricature: Jean-Baptiste Say recommended public works as a temporary response to unemployment during periods of adjustment, and criticized Ricardo for ignoring the possibility that savings might be hoarded if investment opportunities were inadequate. Ricardo, whose economic writings had begun with a pamphlet arguing that the premium on bullion demonstrated the wartime overissue and depreciation of

inconvertible banknotes, was willing after the end of the war to support restoration of gold convertibility at the depreciated parity, rather than deflation to restore the pre-war parity. Henry Thornton (1802) introduced the concept of the central bank as the lender of last resort to support solvent but illiquid banks against bank runs. The proper role, if any, of the Bank of England generated prolonged controversy among the Banking, Currency, and Free Banking Schools in the first three quarters of the 19th century, producing analyses of lasting significance for monetary economics (V. Smith, 1936; Fetter, 1965).

François Quesnay's *Tableau Economique*, the crowning achievement of Physiocratic economics in France at the time of Hume and Smith, represented the circular flow of income and spending. It was not taken up by the mainstream of British and French classical political economy, but, a century after Quesnay, the *Tableau Economique* inspired Karl Marx's schemes of simple and expanded reproduction in the second volume of *Capital* (published posthumously in 1885), relating output and reinvestment rates in Department I (capital goods) and Department II (wage goods). For decades, this pioneering two-sector growth model was used only by Marxist economists such as Rosa Luxemburg and Otto Bauer constructing models of the supposed inevitable breakdown of capitalism, and then in 1928 by G. A. Fel'dman, proposing a growth theory for a planned economy. Fel'dman's articles were part of a false dawn of modern growth theory, appearing in the same year as the December 1928 issue of the *Economic Journal* that contained Allyn Young on increasing returns and economic progress (inspired by Adam Smith) and Frank Ramsey's application of calculus of variations to optimal capital accumulation by a representative agent, but by 1930 Young and Ramsey were dead and Fel'dman had vanished in Stalin's purges (see Fel'dman, Ramsey and Young in Dimand, 2002, vol. 3, and Bauer in vol. 5). Neoclassical hostility to Marx's theory of value and exploitation led to neglect of his contribution to growth theory, just as classical rejection of the Physiocratic doctrine of the exclusive net productivity of agriculture diverted attention from the circular flow. Marx also analysed the cyclical fluctuation of the profit rate around a downward trend, with cyclical troughs in the profit rate causing layoffs that force down wages by swelling the reserve army of the unemployed and cyclical peaks in the profit rate leading to realization crises as redistribution away from wages reduces demand for output (since Marx rejected Say's Law). However, his analysis of the increasing severity of crisis (as of the downward trend of the profit rate) was conducted within the special terminology and assumptions of his labour theory of value, which limited its influence on the mainstream of economics.

Business cycles

Recognition of the more or less periodic recurrence of crises and prosperity goes back at least to Thomas Tooke's discussion in 1823 of ‘waves’ in prices (Arnon, 1991), the beginning of the vast literature on cyclical fluctuations most conveniently sampled in the multi-volume anthologies of O'Brien (1997), Hagemann (2001), and Boianovsky (2005) and in the encyclopedia of Glasner (1997). Clément Juglar (1862) and W. Stanley Jevons (1884, collecting essays written from 1862 to 1882) advanced the analysis of economic fluctuations as periodic oscillations to a higher level, surpassing earlier descriptive and classificatory works (such as Max Wirth's *Geschichte der Handelskrisen* in 1858) and displacing the perception of crises as the result of occasional events. Jevons built upon Hyde Clarke's 1847 suggestion of a meteorological cause for the recurrence of crises every ten years or so (Hyde Clarke also perceived

multiple, overlapping cycles, including a longer period of 54 years, anticipating Kondratiev). Jevons's sunspot theory of the trade cycle has so fallen out of favour that the term ‘sunspots’ in now used in the field of business cycles to refer to any intrinsically irrelevant variables (and even the term ‘business cycles’ is no longer taken to imply that fluctuations are in fact periodic cycles). This is unfair to Jevons, who was following the accepted meteorology of his era, which held that the cycle in solar activity affected weather. Cycles in weather would affect harvests, which, in a still largely agricultural world economy, would affect all economic sectors. Jevons's sunspot theory, together with his warnings about the impending exhaustion of coal, did much more than his marginal utility analysis of relative prices to persuaded the British Association for the Advancement of Science that economics was sufficiently scientific for Section F to remain in the Association. Nonetheless, as Wesley Mitchell (1927, p. 384) remarked,

> Jevons had an admirably candid mind; yet in 1875, when the sun-spot cycle was supposed to last 11.1 years, he was able to get from Thorold Rogers’ *History of Agriculture and Prices in England* a period of 11 years in price fluctuations, and when the sun-spot cycle was revised to 10.45 years he was able to make the average interval between English crises 10.466 years.

Jevons was misled by the belief that an economic cycle must have a cause that is itself cyclical, but, as Knut Wicksell put it, the motion of a rocking horse does not resemble the motion of the stick that started it rocking (cited by Frisch, 1933). Jevons's sunspot theory has distracted attention from such lasting contributions as the seasonal cycle (in his essay on the annual autumnal pressure on the Bank of England) and his use of index numbers to trace the effects of the Australian and California gold discoveries.

Wesley Mitchell (1913; 1927) was the leading figure in the statistical approach to business cycle analysis. In 1920, Mitchell founded the National Bureau of Economic Research (NBER), which was the model for institutes of business cycle or conjuncture research in Berlin, Vienna (directed first by Friedrich Hayek and then by Oskar Morgenstern), Belgium, Sofia, Moscow (directed by Nikolai D. Kondratiev, the theorist of long waves), and Warsaw (where Kalecki worked), Britain's National Institute of Economic and Social Research, and the Institute of World Economics in Kiel. Although his Columbia lectures on types of economic theory were famous, Mitchell was sceptical about taking any single explicit economic theory, such as the quantity theory of money or utility maximization, as a starting point, as he felt that many of the theories surveyed in Mitchell (1927) captured something of the truth, but none the whole truth. Mitchell was influenced by his teacher at the University of Chicago, the institutionalist Thorstein Veblen (1904), who coined the term ‘neoclassical’ to describe the sort of Marshallian economics of which he disapproved. Mitchell and Arthur F. Burns (his successor directing the NBER) concentrated on investigating the statistical properties of time series, looking for patterns of leads and lags and for superimposed cycles of different periods and amplitudes. The widely reported index of leading indicators continues the original NBER approach.

Sir William Beveridge, director of the London School of Economics, used the periodogram, an early version of spectral analysis, to decompose wheat prices into 19 cycles with periods varying from 2.735 years to 68 years (Beveridge, 1921; 1922). Finding so many cycles led sceptics, such as Harvard

statistician E.B. Wilson, to wonder whether there were any truly periodic oscillations in economic time series (apart from seasonality), since with enough cycles any series could be represented as a summation of cycles. Eugen Slutsky (1937, originally published in Russian in 1927), used a moving average of the last three digits of the winning Moscow lottery numbers to show that summation of random series could produce apparent cycles. Slutsky (1937) and Frisch (1933) influenced economists to consider fluctuations as oscillatory responses to random shocks (real or monetary), turning away from the emphasis of Jevons, Juglar, Mitchell, Beveridge and Kondratiev on underlying cycles. Cowles Commission director Tjalling Koopmans (1947) denounced Burns and Mitchell's *Measuring Business Cycles* (1946) as ‘Measurement without Theory’, and argued instead for simultaneous equation macroeconometric models, with the equations identified by exclusionary restrictions derived from a priori economic theory. Koopmans's Chicago colleague Milton Friedman (whose Columbia dissertation had been supervised by Burns) responded by writing down a formal model representing Mitchell's business cycle analysis (Friedman, 1952, Section III and Appendix, pp. 257–82). The vector autoregressions (VAR) of Christopher Sims (1980) marks a return (with more modern statistical techniques) to the NBER approach of investigating the statistical properties of macroeconomic time series with only limited reliance on a priori restrictions drawn from theory.

1886 and all that: the dawn of modern monetary macroeconomics, 1886 – 1914

Around 1886, during a period of depression, analysis of cycles and crises acquired a new emphasis on fluctuation of employment as the problem and variations in the general price level as a preventable cause. Carroll Wright (1886) devoted his first annual report as US Commissioner of Labor to a statistical study, *Industrial Depressions*, finding such depressions to be largely contemporaneous across manufacturing nations and advocating profit-sharing to mitigate the severity of fluctuations (a proposal independently rediscovered nearly a century later by Martin Weitzman). In the same year, Britain had a Royal Commission on the Depression of Trade and Industry, chaired by Lord Iddesleigh (Stafford Northcote) and including Professor Bonamy Price of Oxford but most notable for the evidence of Professor Alfred Marshall of Cambridge. In his evidence to that inquiry and to the Gold and Silver Commission of 1887–8 (both reprinted in Marshall, 1926, edited by Keynes), and in a paper to the Industrial Remuneration Conference of 1885, Marshall considered how far remediable causes adversely affect continuity of employment. This led him to suggest ‘Remedies for Fluctuations of General Prices’ in the *Contemporary Review* in March 1887 (reprinted in Pigou, 1925), revising Ricardo's ingot plan to make the monetary unit a claim on a fixed weight of gold plus a fixed weight of silver, a step toward pegging the monetary unit to a basket of commodities (Irving Fisher's compensated dollar). This symmetallism proved incomprehensible to bimetallists, who persisted in plans that would require pegging the relative price of gold and silver. Also in 1886 (two years after writing the introduction to the posthumous collections of Jevons's *Investigations in Currency and Finance*), Herbert Foxwell published a lecture on *Irregularity of Employment and Fluctuations of Prices* (in Dimand, 2002, vol. 1). Like his colleague Marshall (both were fellows of St John's College, Cambridge), Foxwell emphasized fluctuations in employment as the crucial challenge posed by economic instability, and argued that the problem ‘How to secure greater industrial stability’ could be reformulated as ‘How to diminish price fluctuations’. A young Swedish student named Knut Wicksell attended Foxwell's lectures at University

College London in 1886. As Wesley Mitchell (1927, p. 7) observed, ‘Before the end of the nineteenth century there had accumulated a body of observations and speculations sufficient to justify the writing of histories of the theories of crises’: Eugen von Bergmann's *Die Wirtschaftskrisen: Geschichte der nationalökonomischen Krisentheorien*, published in Stuttgart in 1895, and E. D. Jones's *Economic Crises*, published in New York in 1900 (see also Barnett, 1941). In 1909, the year of Beatrice and Sidney Webb's Minority Report of the Poor Law Commission and of the first edition of Beveridge's *Unemployment*, the London School of Economics published a 71-page bibliography of unemployment and the unemployed by F. Isabel Taylor.

The cover of David Laidler's *The Golden Age of the Quantity Theory* (1991b) shows the three economists who dominated monetary economics before the First World War, making the case for monetary shocks and imbalances as the avoidable source of fluctuations: Alfred Marshall, Knut Wicksell and Irving Fisher.

In *Interest and Prices* in 1898 and then in his *Lectures on Political Economy* (1915), Wicksell distinguished the market rate of interest, set by the banking system, from the natural rate, the interest rate at which desired saving and investment would balance and the price level would not change. If a technical innovation raises the natural rate, or the banking system lowers the market rate, it will be profitable for entrepreneurs to borrow for new investment projects as long as the natural rate exceeds the market rate, causing (in a pure credit economy with no cash drain) a cumulative inflation. If the market rate exceeded the natural rate, a cumulative deflation would ensue. Although he considered himself a quantity theorist following in the footsteps of Ricardo, Wicksell was a pioneer in analysing a pure credit economy, not anchored by gold or other base money, which is why Michael Woodford deliberately chose Wicksell's title *Interest and Prices* for his 2003 treatise analysing a world in which financial innovation has greatly reduced the role of cash and bank reserves. Wicksell's economic contributions (which included using what came to be called the Cobb–Douglas production function four years before Cobb and Douglas) were continued by a Stockholm School including Dag Hammarskjold, Karin Kock (1929), Erik Lindahl (1939), Erik Lundberg (1937), Gunnar Myrdal (1939) and Bertil Ohlin – a list including three Nobel laureates (two in economics, one in peace) and four Swedish cabinet ministers. The Stockholm economists later expressed confidence that, even if Keynes had never written *The General Theory*, they would have discovered it themselves, but Don Patinkin (1982) expressed doubt, because the focus of Wicksell's heirs was on price dynamics, not the equilibrium level of employment and national income. Keynes's earlier *Treatise on Money* in 1930 (1971–89, vols 5 and 6) was much more Wicksellian than *The General Theory* in its emphasis on cumulative inflation or deflation when the interest rate does not equate planned investment to planned saving.

J. Bradford De Long (2000) writes:

> The story of 20th century macroeconomics begins with Irving Fisher. In his books *Appreciations and Interest* (1896), *The Rate of Interest* (1907), and *The Purchasing Power of Money* (1911) [Fisher, 1997, vols. 1, 3, and 4], Fisher fueled the intellectual fire that much later became monetarism. To understand the determination of prices and interest rates and the course of the business cycle, monetarism holds, look first (and often last) at the stock of money – at the quantities in the economy of those assets that constitute readily spendable purchasing power. … It is true that the ideas that we see as necessarily

> producing the quantity theory of money go back to David Hume, if not before. But the equation-of-exchange and the transformation of the quantity theory of money into a tool for making quantitative analyses and predictions of the price level, inflation, and interest rates was the creation of Irving Fisher.

In *Appreciation and Interest*, Fisher argued that the difference between interest rates expressed in two standards (money and commodities, gold and silver, dollars and francs) is the expected rate of appreciation of one standard in terms of the other, deriving from this uncovered interest parity between two countries, the expectations theory of the term structure of interest rates, and the Fisher relation that nominal interest is the sum of real interest and expected inflation (plus a cross-product term). In *The Rate of Interest*, Fisher introduced the Fisher diagram, showing the optimal smoothing of consumption over two periods (assuming perfect credit markets) and an individual's saving or dissaving in each period. In *The Purchasing Power of Money*, Fisher (with his former student Harry G. Brown) upheld the quantity theory both against bimetallists who predicted permanent real benefits from expanding the money supply and against hard-money opponents of bimetallism (notably J. Laurence Laughlin of the University of Chicago), who denied the path of US prices could be explained by changes in the money supply. Fisher and Brown explained economic fluctuations by the slow adjustment of nominal interest to monetary shocks during ‘transition periods’ (lasting perhaps ten years), so that fluctuations could be avoided either by educating the public against what Fisher later termed ‘the money illusion’ (so that expected inflation and hence nominal interest would adjust to monetary shocks, leaving real interest unaltered) or by a monetary policy rule of varying the exchange rate (the dollar price of gold) to hold constant a price index (for which Fisher later proposed the Fisher ideal index, the geometric mean of the Paasche and Laspeyres indexes). Fisher's 1926 article, ‘A statistical relation between unemployment and price changes’ (in Fisher, 1997, vol. 8), correlated unemployment with a distributed lag of past price level changes (as a proxy for expected inflation), and was reprinted in the *Journal of Political Economy* in 1973 under the heading ‘Lost and Found: I Discovered the Phillips Curve – Irving Fisher’. Unlike Marshall in Cambridge and Wicksell in Stockholm, Fisher did not attract a school of disciples at Yale. Through his role in establishing the Econometric Society and the Cowles Commission, Fisher advanced his preferred economic methodology of formal theorizing using mathematical and statistical techniques, but his contributions to monetary economics and economic fluctuations (like those of Hayek, Hawtrey, and many others) were long overshadowed by Keynes's *General Theory*, notwithstanding Keynes's acknowledgement of Fisher as his intellectual great-grandparent in appreciating the real effects of monetary changes.

Although Alfred Marshall's *Money, Credit and Commerce* was not published until 1923, the year before his death, parts of it were drafted as early as the 1870s, and his ideas had long circulated through his lectures, his evidence to official inquiries (gathered by Keynes in Marshall 1926), and the ‘Cambridge oral tradition’ of monetary theory (Eshag, 1963; Bridel, 1987; Laidler; 1999). Marshall, his professorial successor A. C. Pigou, Pigou's successor D.H. Robertson (1926), the young J.M. Keynes, and Cambridge economics lecturers Frederick Lavington and J.R. Bellerby used a cash balance version of the quantity theory, relating the number of units of purchasing power the public wished to hold as cash to the level of income (in contrast to Fisher's logically equivalent version, which expressed the quantity theory in terms of the velocity of circulation of money).

Departure from the gold standard during the First World War and the post-war central European hyperinflations provided the occasion for the highest achievement of Marshallian monetary economics, Keynes's *Tract on Monetary Reform* in 1923 (Keynes 1971–89, vol. 4), an innovative work but one that innovated within the tradition established by Marshall. Keynes analysed inflation as a form of taxation of real money balances, identified as a social cost the consequent reduction in desired holdings of real money balances (M/P), and introduced covered interest parity (the spread between forward and spot exchange rates equals the difference between interest rates in the two currencies). Keynes calculated that real money balances had fallen by 92 per cent during the German hyperinflation, as a result of the soaring opportunity cost of holding money. Others had mistakenly argued that since the price level (P) was rising faster than the money supply (M), monetary expansion could not be the cause of the price inflation, and the Reichsbank president Rudolf Havenstein promised that, with 38 new high-speed printing presses, the Reichsbank would be able to print enough money to catch up with the prices. Robertson (1926), then collaborating closely with Keynes, examined forced saving (‘induced lacking’ in Robertson's terminology) caused by inflation. Turning from inflation to deflation, Keynes wrote *The Economic Consequences of Mr. Churchill* (in Keynes 1971–89, vol. 9) to oppose Britain's return to the gold standard at the pre-war parity in 1925, arguing that restoration of the pre-war parity would require a reduction of prices and money wages that could be achieved only through prolonged unemployment (see June Flanders, 1989, on the development of international monetary economics).

Keynesian Revolution and monetarist counter-revolution

The Great Depression of the 1930s helped provide a receptive audience for John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936; 1971–89, vol. 7), which argued that involuntary unemployment could persist unless the government intervened with appropriate management of aggregate demand (Clarke, 1988; Dimand, 1988; Backhouse, 1995). *The General Theory* challenged Lionel Robbins and Friedrich Hayek of the London School of Economics, who argued against expansionary fiscal and monetary policy and for letting the depression take its course, and William Beveridge, who held (until his conversion to Keynesianism) that the existing level of British unemployment could be fully accounted for by structural, frictional, and seasonal unemployment without invoking any deficiency of aggregate demand. To the rising generation of new economists, from Harvard students Paul Samuelson and James Tobin to LSE economists Abba Lerner and Nicholas Kaldor, Keynes offered a message of hope that depressions were curable and preventable without adopting a Soviet-style centrally planned economy. Attempts to dismiss or ignore Keynes (Burns and Mitchell, 1946, mentioned Keynes in one sentence, in a footnote) were futile. Keynes provided an agenda for economists providing public policy advice and a framework for empirical, policy-oriented modelling, at a time when depression and war greatly expanded the role of governments.
Keynes's success in winning over the next generation of economists obscured the extent to which his contemporaries in economics shared his policy views rather than those of Robbins and Hayek: although Keynes used Pigou (1933) as the target of his attack on classical theory, he recognized how close they stood on practical policy. Even Ralph Hawtrey, the Treasury economist associated with the ‘Treasury view’ about crowding out and the ineffectiveness of fiscal policy, was convinced of the effectiveness of (and need for) stabilizing monetary policy (and contributed intriguing numerical examples to the

development of the Kahn–Keynes spending multiplier; see Hawtrey, 1932). Keynes's caricature in *The General Theory* of ‘classical economists’ from Ricardo to Pigou as upholders of a rigid version of Say's Law, denying any role to aggregate demand in explaining unemployment (in contrast to the superior insight but fuzzier analytics of mercantilists, Malthus, and the underconsumptionists Hobson and Mummery), was more widely noted than his subsequent clarification that he did not consider Fisher or Hawtrey or Robertson or Wicksell as classical. However, support for expansionary fiscal or monetary policy during the Depression did not necessarily imply anticipation of Keynesian economics: proposals circulated for emergency public works financed by cutting other government spending and for domestic monetary expansion while keeping the exchange rate fixed, and in the United States the New Deal's National Recovery Administration was an attempt to raise price toward pre-Depression levels by restricting supply, rather than by stimulating demand. Keynes provided a framework within which the implications of such policies could be analysed. Independently of Keynes, starting from Marx and Rosa Luxemburg, Michal Kalecki in Poland developed a theory very close to Keynes's income–expenditure analysis, and in 1934 published in Polish a three-equation model of goods market equilibrium, money market equilibrium and aggregate supply. Patinkin (1982) argued that Kalecki was concerned with the dynamics of cyclical fluctuations, Keynes with determining the equilibrium level of income that equates saving to desired investment, and that Kalecki's 1934 essay (which Kalecki did not choose to be translated among his selected articles in 1966 and 1971, or refer to in other works) was not part of his central message.

The analytical framework that dominated macroeconomics for at least a quarter century after the Second World War was based on Keynes's aggregate supply and aggregate demand functions (generally with more attention to aggregate demand than to aggregate supply) and the small system of simultaneous equations behind the Hicks–Hansen IS/LM diagram, which included Keynes's money demand function (liquidity preference) and later substitutes for his consumption function (De Vroey and Hoover 2004). The system of equations representing Keynes's message in a form equivalent to IS/LM was a four-equation model in Keynes's Cambridge lectures in December 1933, attended by David Champernowne and Brian Reddaway, the first economists to use such a model in print, but Keynes did not include it in *The General Theory*, perhaps following Marshall's advice to use mathematics as a tool of inquiry but to then translate the analysis into English and burn the mathematics (Rymes, 1989; Dimand, 1988). The resulting framework (extended to open economies by Robert Mundell and J. Marcus Fleming in the 1960s) did not capture all of Keynes's message (or messages), notably his distinction between fundamental uncertainty and insurable risk. Econometric estimation of macroeconomic models was pioneered, independently of Keynes, by Ragnar Frisch, Jan Tinbergen and Trygve Haavelmo (and Keynes's review of the first volume of Tinbergen, 1939, expressed severe scepticism), but it was taken up with enthusiasm by such Keynesians as Lawrence Klein. The claim in Chapter 2 of *The General Theory* that real and money wages move in opposite directions over the course of the cycle (and by implication, that real wages vary counter-cyclically) was challenged empirically by John Dunlop and Lorie Tarshis and on theoretical grounds by Michal Kalecki, leading Keynes in 1939 to acknowledge the cyclical pattern of real wages as an open question, which it remains to this day.

Milton Friedman and his students (Friedman, 1956) offered a renewed quantity theory of money as a challenge to Keynesianism, claiming to follow a Chicago oral tradition of monetary theory. Certainly it drew on such Chicago landmarks as Henry Simons's 1936 argument for rules rather than authorities in

monetary policy (reprinted in Simons, 1948), but the intellectual inheritance from non-Chicago quantity theorists such as Irving Fisher and Clark Warburton (and even the young Keynes of *A Tract on Monetary Reform*) gradually came to be recognized. As Patinkin (1981) noted, a key element of Friedman's approach, the demand for money as a function of a small number of variables, originated in Keynes's *General Theory*. Although others had come close (in 1930, Fisher stated the marginal opportunity cost of holding cash balances), Keynes was the first to write the demand for money as a function of income and the interest rate. A further irony was that, though the spread of Keynesianism stemmed largely from its apparent ability to explain the Great Depression, the monetary interpretation of the Great Depression by Friedman and Schwartz (1963), as the consequence of mistaken Federal Reserve policy that permitted the US money supply to contract by a third, was crucial in persuading many economists of the explanatory power of monetarism, the revived form of the quantity theory. For an overview of the development of macroeconomics from Keynes through Friedman to the New Classical and New Keynesian research programs (and the non-mainstream Post Keynesian and Austrian schools, from Keynesian fundamental uncertainty and Mises–Hayek trade cycle theory, respectively), enlivened with interviews with leading participants (see Snowdon and Vane, 2005).

Recurring themes

Certain issues reappear throughout the history of economics. Do fluctuations result from monetary disturbances, as Hawtrey (1913; 1932) and Fisher argued, or from real productivity shocks such as Schumpeterian innovations? Is unemployment best analysed as the functioning or malfunctioning of the labour market (as in Beveridge, 1930, and Hutt, 1939) or in terms of the demand for and supply of output as a whole (Keynes)? Is there a role for demand management to offset instability caused by volatile private investment reflecting the fundamental uncertainty of future profitability (Keynes) or is government itself the source of instability (von Mises, Hayek)? Should a central bank follow a rule rather than having discretion (as Henry Simons asked in 1936), or need there even be a central bank (as Hayek's student Vera Smith asked the same year)? Are recessions undesirable and preventable disequilibrium phenomena, or, as Arthur Ellis (1879) and Friedrich Hayek (1931) held, are they a normal and necessary part of the equilibrium path of the economy? Should analysis of economic fluctuations should be primarily a study of the statistical properties of the fluctuations, as in Burns and Mitchell (1946) and decades later Sims's vector autoregressions, or should the analysis be explicitly grounded in formal economic theory? As macroeconomists continue to theorize, measure, test, and argue about these issues, they stand, knowingly or not, on the ‘shoulders of giants’ who discussed these questions before. De Long (2000, p. 83) notes that ‘The New Classical research program walks in the footprints of Joseph Schumpeter's *Business Cycles* (1939) [and of Schumpeter, 1912, and Robertson, 1915], holding that the key to the business cycle is the stochastic nature of economic growth [so that] the “cycle” should be analyzed with the same models used to understand the “trend”’ ,while the name of the New Keynesian research program (which emphasizes frictions that prevent instantaneous adjustment to nominal shocks) indicates its historical antecedents (although, as De Long points out, it also incorporates important features of Milton Friedman's contributions, such as emphasis on policy rules and on monetary rather than fiscal policy). Insights have sometimes long preceded the ability to formalize them;

even Adam Smith's famous increasing returns through the division of labour, revived in Allyn Young's 1928 essay on economic progress, did not make its mark on the theories of international trade and endogenous growth until the last decades of the 20th century, when ways were devised to incorporate increasing returns to scale in formal models. The field has experienced major changes, as when Keynes made determining the equilibrium level of national income the central issue, or when monetarism posed inflation as the central problem instead of unemployment, or when attention shifted from fluctuations to long-term growth, but in each case the change was a transformation of a rich heritage.

Bibliography

Arnon, A. 1991. *Thomas Tooke, Pioneer of Monetary Theory*. Aldershot, UK: Edward Elgar.

Backhouse, R.E. 1995. *Interpreting Macroeconomics: Explorations in the History of Macroeconomic Thought*. London and New York: Routledge.

Barnett, P. 1941. *Business-Cycle Theory in the United States, 1860–1900*. Chicago: University of Chicago Press.

Beveridge, W.H. 1921. Weather and harvest cycles. *Economic Journal* 21, 429–52.

Beveridge, W.H. 1922. Wheat prices and rainfall in Western Europe. *Journal of the Royal Statistical Society* 85, 412–59.

Beveridge, W.H. 1930. *Unemployment, a Problem of Industry (1909 and 1930)*. London: Longmans, Green.

Boianovsky, M., eds. 2005. *Business Cycle Theories: Selected Texts 1860–1939*, vols. 5–8, London: Pickering & Chatto.

Bridel, P. 1987. *Cambridge Monetary Thought: The Development of Saving–Investment Analysis from Marshall to Keynes*. Basingstoke: Macmillan.

Burns, A.F. and Mitchell, W.C. 1946. *Measuring Business Cycles*. New York: NBER.

Clarke, H. 1847. *Physical Economy: A Preliminary Inquiry into the Physical Laws Governing the Periods of Famines and Panics*. London (Cat. No. 34987.8, Kress Library, Harvard University).

Clarke, P. 1988. *The Keynesian Revolution in the Making 1924–1936*. Oxford: Oxford University Press.

Corry, B. 1962. *Money, Saving and Investment in English Economics 1800–1850*. London: Macmillan.

De Long, J.B. 2000. The triumph of monetarism? *Journal of Economic Perspectives* 14(1), 83–94.

De Vroey, M. and Hoover, K.D., eds. 2004. *The IS–LM Model: Its Rise, Fall, and Strange Persistence*. Durham, NC: Duke University Press. *Annual Supplement to History of Political Economy 36*.

Dimand, R.W. 1988. *The Origins of the Keynesian Revolution*. Aldershot, UK: Edward Elgar, and Stanford, CA: Stanford University Press.

Dimand, R.W., ed. 2002. *The Origins of Macroeconomics*, 10 vols. London and New York: Routledge.

Ellis, A. 1879. *The Rationale of Market Fluctuations*, 4th edn. London: Effingham Wilson.

Eshag, E. 1963. *From Marshall to Keynes: An Essay on the Monetary Theory of the Cambridge School*. Oxford: Basil Blackwell.

Fetter, F.W. 1965. *The Development of British Monetary Orthodoxy 1797–1875*. Cambridge, MA: Harvard University Press.

Fisher, I. 1997. *The Works of Irving Fisher*, 14 vols, ed. W. J. Barber assisted by R. Dimand and K. Foster, consulting ed. J. Tobin. London: Pickering & Chatto.

Flanders, M.J. 1989. *International Monetary Economics 1870–1960*. Cambridge: Cambridge University Press.

Friedman, M. 1952. The economist theorist. In *Wesley Clair Mitchell: The Economic Scientist*, ed. A.F. Burns. New York: NBER.

Friedman, M., ed. 1956. *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press for NBER.

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic Essays in Honour of Gustav Cassell*, London: George Allen & Unwin.

Glasner, D., ed. 1997. *Business Cycles and Depressions: An Encyclopedia*. New York: Garland.

Grice-Hutchinson, M. 1952. *The School of Salamanca: Readings in Spanish Monetary Theory 1544–1605*. Oxford: Clarendon Press.

Hagemann, H., ed. 2001. *Business Cycle Theories: Selected Texts 1860–1939*, 4 vols. London: Pickering

& Chatto.

Hawtrey, R.G. 1913. *Good and Bad Trade*. London: Constable. Reprinted New York: A. M. Kelley, 1970 (with 1962 preface by author).

Hawtrey, R.G. 1932. *The Art of Central Banking*. London: Longmans, Green.

Hayek, F.A. 1931. *Prices and Production*. London: Routledge.

Hicks, J.R. 1937. Mr. Keynes and the classics: a suggested interpretation. *Econometrica* 5, 147–59.

Hume, D. 1752. *Writings on Economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.

Humphrey, T.M. 1993. *Money, Banking, and Inflation: Essays in the History of Monetary Thought*. Aldershot, UK, and Brookfield, VT: Edward Elgar.

Hutt, W.H. 1939. *The Theory of Idle Resources*. London: Cape.

Jevons, W.S. 1884. *Investigations in Currency and Finance*, ed. H.S. Foxwell. London: Macmillan.

Juglar, C. 1862. *Des crises commerciales et de leur retour périodique en France, en Angleterre et aux Etats-Unis*. Paris: Guillaumin.

Kalecki, M. 1934. Trzy uktady [Three economic models]. *Ekonomista* 34, 54–70. Trans. C.A. Kisel. In M. Kalecki, Collected Works of Michal Kalecki, Volume 1: Capitalism, *Business Cycles and Full Employment*, ed. J. Ostiatynski. Oxford: Clarendon Press, 1990.

Kalecki, M. 1935. A macrodynamic theory of the business cycle. *Econometrica* 3, 327–44.

Keynes, J.M. 1971–89. *Collected Writings of John Maynard Keynes*, 30 vols., ed. D.E. Moggridge and E. A.G. Robinson. London: Macmillan, and New York: Cambridge University Press, for the Royal Economic Society.

Klein, L.R. 1946. Macroeconomics and the theory of rational behavior. *Econometrica* 14(2), 93–108.

Kock, K. 1929. *A Study of Interest Rates*. Stockholm Economic Studies No. 1. London: P.S. King.

Koopmans, T.C. 1947. Measurement without theory. *Review of Economic Statistics* 29, 161–72.

Laidler, D. 1991a. *The Golden Age of the Quantity Theory*. Princeton, NJ: Princeton University Press.

Laidler, D. 1991b. The quantity theory is always and everywhere controversial – why? *Economic Record* 67, 289–306.

Laidler, D. 1999. *Fabricating the Keynesian Revolution: Studies of the Inter-war Literature on Money, the Cycle, and Unemployment*. Cambridge: Cambridge University Press.

Laidler, D. 2003. *Macroeconomics in Retrospect: Selected Essays*. Cheltenham, UK, and Northampton, MA: Edward Elgar.

Lindahl, E. 1939. *Studies in the Theory of Money and Capital*. London: George Allen & Unwin.

Link, R. 1959. *English Theories of Economic Fluctuations 1815–1848*. New York: Columbia University Press.

Lundberg, E. 1937. *Studies in the Theory of Economic Expansion*. London: P.S. King.

Marshall, A. 1923. *Money, Credit and Commerce*. London: Macmillan.

Marshall, A. 1926. *Official Papers*, ed. J.M. Keynes. London: Macmillan.

Mises, L. von. 1935. *The Theory of Money and Credit*, trans. H. Batson. London: Cape.

Mitchell, W.C. 1913. *Business Cycles*. Berkeley: University of California Press.

Mitchell, W.C. 1927. *Business Cycles: The Problem and Its Setting*. New York: NBER.

Myrdal, G. 1939. *Monetary Equilibrium*, trans. R. Bryce and N. Stolper. London: W. Hodge.

O'Brien, D.P., ed. 1997. *Foundations of Business Cycle Theory*, 3 vols. Cheltenham, UK, and Brookfield, VT: Edward Elgar.

Patinkin, D. 1981. *Essays on and in the Chicago Tradition*. Durham, NC: Duke University Press.

Patinkin, D. 1982. *Anticipations of the General Theory? And Other Essays on Keynes*. Chicago: University of Chicago Press.

Pigou, A.C., ed. 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Pigou, A.C. 1933. *Theory of Unemployment*. London: Macmillan.

Robertson, D.H. 1915. *A Study of Industrial Fluctuation*. London: P.S. King.

Robertson, D.H. 1926. *Banking Policy and the Price Level*. London: P.S. King.

Rymes, T.K. 1989. *Keynes's Lectures, 1932–35: Notes of a Representative Student*. Basingstoke: Macmillan, and Ann Arbor: University of Michigan Press.

Schumpeter, J.A. 1912. *The Theory of Economic Development*, trans. R. Opie. Cambridge, MA: Harvard University Press, 1934.

Schumpeter, J.A. 1939. *Business Cycles*, 2 vols. New York: McGraw-Hill.

Simons, H. 1948. *Economic Policy for a Free Society*. Chicago: University of Chicago Press.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Slutsky, E. 1937. The summation of random causes as the source of cyclic processes. *Econometrica* 5, 105–46.

Smith, V.C. 1936. *The Rationale of Central Banking*. London: P.S. King.

Snowdon, B. and Vane, H.R. 2005. *Modern Macroeconomics: Its Origins, Development and Current State*. Cheltenham: Edward Elgar.

Sowell, T. 1972. *Say's Law: A Historical Analysis*. Princeton, NJ: Princeton University Press.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*, ed. F.A. Hayek. New York: A.M. Kelley, 1965.

Tinbergen, J. 1939. *Statistical Testing of Business Cycle Theories*, 2 vols. Geneva: League of Nations. Reprinted New York: Agathon Press, 1968.

Veblen, T. 1904. *The Theory of Business Enterprise*. New York: Charles Scribner's Sons.

Vickers, D. 1959. *Studies in the Theory of Money 1690–1776*. New York: Chilton.

Wicksell, K. 1898. *Interest and Prices*, trans. R. Kahn. London: Macmillan, 1936.

Wicksell, K. 1915. *Lectures on Political Economy*, vol. 2, trans. E. Claasen. London: Routledge, 1935.

Wirth, M. 1858. *Geschichte der Handelskrisen*. 4th edn, Frankfurt am Main, 1890.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, NJ: Princeton University Press.

Wright, C.D. 1886. *Industrial Depressions: The First Annual Report of the Commissioner of Labor, March, 1886*. Washington, DC: Government Printing Office. Reprinted New York: Augustus M. Kelley, 1968.

Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

Young, W. 1987. *Interpreting Mr. Keynes: The IS–LM Enigma*. Oxford: Polity Press and Boulder, CO: Westview.

How to cite this article

Dimand, Robert W. "macroeconomics, origins and history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000370> doi:10.1057/9780230226203.1009

The New Palgrave Dictionary of Economics Online

Maddala, G.S. (1933 – 1999)

Kajal Lahiri

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

G.S. Maddala's many publications covered almost every substantive areas of econometrics – distributed lags, generalized least squares, panel data, simultaneous equations, measurement errors, switching and disequilibrium models, qualitative and limited dependent variable models, selection and self-selection models, exact small sample distributions of estimators, outliers and bootstrap methods, robust estimators and more. G.S. became a veritable textbook himself – a pre-eminent teacher in econometrics and an authority on almost every econometrics topic. G.S. was a brilliant expositor – he could cut through the technical superstructure to reveal only essential details, while retaining the nerve centre of the subject matter he sought to explain.

Keywords

Maddala, G. S.; Griliches. Z.; Bayesian econometrics; bootstrap; cointegration; cross-section information; disequilibrium models; dummy variables; Econometric Society; econometrics; elasticity of substitution; error component estimator; exchange rate determination; heteroskedasticity; instrumental variables; limited dependent variables; maximum likelihood; Mundlak, Y.; Nerlove, M.; panel data; Pareto distribution; Pascal lag; production functions; qualitative variables; rational distributed lag models; rational expectations; recursive models; sample theory; simultaneous equations models; Singh–Maddala distribution; structural breaks; structural change; survey data; technical change; time series analysis; two-step procedures; unit roots; Wald test; Weibull distribution

Article

G.S. Maddala (universally known as ‘G.S.’) was born on 21 May 1933 in the south Indian state of Andhra Pradesh, where he had his high-school education. G.S. held the University Eminent Chair at the Ohio State University when he died on 4 June 1999 due to congestive heart failure.
G.S.'s father was a schoolteacher of modest means, and his mother, though having only an elementary

education, was well versed in Sanskrit and the works of the great Indian philosopher Sankara. After graduating from high school in 1947, G.S. had to drop out of college for a few years due to health and other reasons. In 1955 he graduated first in his class from Andhra University with a BA in mathematics, and went on to graduate in First Class from Bombay University with an MA in statistics in 1957. With a Fulbright Fellowship, G.S. travelled to the University of Chicago in 1960 and completed his Ph.D. in 1963 under the supervision of the late Zvi Griliches. In that year, he was offered the job of Assistant Professor of Economics at Stanford University. Before joining Ohio State in 1993, G.S. taught at the University of Rochester (1967–75) and at the University of Florida (1975–93). He also held visiting appointments at Cornell, Yale, CORE, Monash, Columbia, Caltech (as the Fairchild Distinguished Scholar), Emory and Oakridge Labs. The fascinating narration of his journey from an early college dropout in a remote Indian village in 1947 to a faculty position at Stanford in 1963 can be found in the Introduction (‘How I Became an Econometrician’) to the two-volume selected works of Maddala (1994). More detailed biographical information, his life story and philosophy can be found in Lahiri and Phillips (1999), Lahiri (1999), Griliches (1999), Rosen (2000) and Hsiao (2003).

Beginning with his first published paper (with Zvi Griliches, Robert Lucas, and Neil Wallace) in 1962, through the next four decades, G.S. published 12 books and more than 110 articles covering almost every emerging area of econometrics – distributed lags, generalized least squares, panel data, simultaneous equations, measurement errors, tests of significance, switching and disequilibrium models, qualitative and limited dependent variable models, selection and self-selection models, exact small sample distributions of estimators, outliers and bootstrap methods, robust estimators, and more. The list is practically endless. Throughout his career G.S. used sample theory and Bayesian techniques freely in his research, a rarity in the econometrics profession, and was one of the early proponents of Bayesian approach in econometrics. Through his many books and the breadth of his own research, G.S. became a veritable textbook himself – a pre-eminent teacher in econometrics and an authority on almost every econometrics topic. Not surprisingly, according to the *Social Science Citation Index*, G.S. was one of the top five most-cited econometricians during each of the years 1988–93, and he was cited more times in 1994 and 1996 than each of the six econometricians who won the Clark Medal during 1970–2000.

During the 1960s, G.S. contributed heavily towards the formulation and estimation of production functions and technical change. His doctoral dissertation was on productivity and technical change in the US bituminous coal industry. His two papers with Jay Kadane in 1966 and 1967 considered, respectively, the importance of alternative exogeneity assumptions in the estimation of the constant elasticity of substitution production functions parameters inclusive of the share equations; and the bias in the estimation of the returns to scale parameter when the production function is incorrectly specified as a Cobb Douglas. The rigour and depth in these papers were undoubtedly ahead of their time.

The early 1970s saw a flurry of activity on efficient estimation methods of alternative distributed lag models. One of G.S.'s widely cited papers (1971a) showed why certain commonly used two-step procedures are asymptotically less efficient than the maximum likelihood estimator in the presence of lagged dependent variables as regressors. This sort of problem is encountered also in dynamic panel data models with individual heterogeneity. The key result in this paper is that in these models the information matrix of the slope parameters and the parameters embedded in the covariance matrix of residuals are not diagonal. Using this as a starting point, Pagan (1986) developed a more thorough and modern characterization of numerous two-step procedures with estimated covariance matrix in the context of various econometric models.

With Dave Grether in 1973, G.S. studied the effects of errors in variables in distributed lag models with serial correlation. They showed analytically that the estimated speed of adjustment can be severely biased, and can give the spurious appearance of a long lag in adjustment. In two influential papers with A.S. Rao in 1971 and 1973, G.S. developed maximum likelihood procedures for Solow's Pascal lag and Jorgenson's rational distributed lag models, and compared the power of tests for serial correlation in regression models with lagged dependent variables. One important conclusion that emerged from the latter study was that the nature of the autocorrelation and trend in the exogenous variable is crucial in determining the small sample behaviour of the test statistics and the estimators – hinting at much of the work on integrated variables that would come in the 1980s.

During the early 1970s G.S. also produced a number of important papers on the use and estimation of panel data models, and rightfully became one of the three ‘fathers’ (together with Yair Mundlak and Marc Nerlove) of modern panel data analysis in econometrics. In his influential *Econometrica* (1971b) paper, G.S. demonstrated – with his characteristic clarity – that the error component estimator is a weighted combination of within and between estimators, and thus the use of dummies entails substantial loss of information by ignoring the ‘between’ variation in the data. In another *Econometrica* (1971c) paper, G.S. discussed the problem of pooling cross-section and time series data, and emphasized tests for consistency between time series and cross-section information. The paper contains a very deep analysis of an alternative Bayesian approach with diffuse priors and concludes that the two approaches should be complementary. (Publishing three full-length articles in *Econometrica* in a year has to be some kind of a record for an economist!) The profession quickly saw the enduring value of these publications and elected G.S. a fellow of the Econometric Society in 1975.

During the 1970s, like many other econometrics stalwarts of the period, G.S. was also involved in the development of econometric methodology in simultaneous equations models. He worked on appropriate estimation strategies in large and medium-size econometric models (1971d), and studied the power characteristics of alternative tests of significance associated with simultaneous equation estimation (1974a). His *Econometrica* (1974b) paper showed that ‘diffuse’ and ‘non-informative’ priors might lead to sharp posterior distributions even in under-identified models. Only recently have Chao and Phillips (2002) fully solved the so-called ‘Maddala paradox’ using Jeffreys prior. They interpret the pathological result in terms of a naive use of the diffuse prior that fails to downweight sufficiently that part of the parameter space where the rank condition either fails or nearly fails. In another potent contribution to an important recent work on weak instruments, Maddala and Jeong (1992) correctly showed that the bimodal distribution of the instrumental variable estimator obtained in the literature is merely due to the illustrative model used, where the correlation between the structural and the first-stage errors is perfect. Phillips (2006) gives a complete characterization of the bimodality problem when instruments are weak.

From the mid-1970s, G.S. was primarily focused on developing estimation and test procedures for qualitative and limited dependent variable models, and produced nearly 40 articles. This line of research also dealt with models with selection, self-selection, disequilibrium and controlled prices. His work at Rochester with Forrest Nelson (1974) on disequilibrium models and with Lung-Fei Lee (1976) on recursive models with qualitative endogenous variables and generalized selection models represents a long and very fruitful period of research on this topic. His 1983 Econometric Society monograph, *Limited Dependent and Qualitative Variables in Econometrics*, was an immediate best-seller and was declared a citation classic in *Current Contents* (vol. 30, 16 July 1993). It has fuelled much of the innovative applied and theoretical research using these tools since the mid-1980s, and has served as a

bible to empirical researchers in applied microeconomics. The strength of the book lies in its comprehensiveness, expositional simplicity, and depth. As of June 2006, the Google Scholar reports a record 3,721 citations of this advanced monograph. G.S. also wrote a number of theoretical and empirical papers analysing limited dependent and qualitative variable models with panel data, and wrote widely cited expository articles for use in other disciplines such as accounting, finance, transportation, and health.

It is notable that G.S. can jointly claim a statistical distribution – the Singh–Maddala (1976) distribution – a much better name than the Burr type 12 to which it is related. Maddala and Singh's proposed statistical distribution has triggered much research in describing the actual size distribution of incomes, and is a generalization of the Pareto distribution and the Weibull distribution used in analysis of equipment failures. As aptly noted by Sherwin Rosen (2000) while delivering the first Maddala lecture at Ohio State University on 26 April 2000, ‘Coase may have his Theorem, Stigler his Laws, Black and Scholes their Formula, and Lucas his critique, but what economist aside from Pareto (who was just as much a sociologist and political scientist and only one third economist) has half ownership of a distribution? And what an elegant economic derivation it has.’

G.S. had a deep interest in rational expectations models, in the validity of the hypothesis that can be gleaned from recorded survey data, and in how econometric disequilibrium models play out in this framework. Maddala, Fishe and Lahiri (1983) developed methods to estimate aggregate expectations when available survey data are partly qualitative and partly quantitative. He had done pioneering work (Maddala, 1983a) on the estimation for models with bounded price variation, and with Scott Shonkwiler (1985) applied the methodology to the corn market. With Steve Donald (1992), G.S. studied the disequilibrium model with upper and lower bounds on prices under rational expectations. The latter paper foreshadowed much work on exchange rate determination in a target zone in the 1990s. Undoubtedly, the full potential of this line of research initiated by G.S. is yet to be realized.

With failing health, G.S. spent much of the 1990s working primarily on bootstrap techniques and time series models with cointegration and structural breaks. During this period, he also wrote important papers on tests of unit roots in panel data models, robust inference, errors in variables problems in finance, Bayesian shrinkage estimation, outliers and influential observations, neural nets, and many others. Thus, ill health neither slowed down his research nor dampened his passion for mentoring and supervising Ph.D. students. In total G.S. supervised close to 60 doctoral students, co-authoring more than 65 published articles with them.

While testing the rationality of survey data on interest rate expectations in the context of a multiple-indicator single index model with heteroskedasticity, Maddala and Jeong in the mid-1990s used the weighted double bootstrap method to implement the Wald test in finite samples. His work with Hongyi Li in 1996 explored the use of different bootstrap techniques in cointegration regressions, financial and non-linear models. With Wu (1999) on panel data unit root test, G.S. suggested the use of a novel Fisher test that combines N individual tests with bootstrap-based critical values. Since much remains to be done to extend the Fisher approach to combining individual tests that are correlated, further generalizations of the Maddala–Wu test are certainly to come.

Much of his work on modern time series analysis has been summarized in his seminal book with In-Moo Kim (1998). This book also presents a comprehensive and lucid review of unit root and cointegration tests, and estimation with integrated variables. It discusses problems of unit root tests and cointegration under structural change, outliers, robust methods, the Markov switching model, and Harvey's structural

time series model. The book contains a welcome chapter on the Bayesian approach to many of these problems and bootstrap methods for small-sample inference.

G.S. contributed to a number of purely policy-oriented and applied areas. Some of these topics include consumption, production and cost functions, money demand, regulation, pseudo-data, returns to college education, housing markets, energy demand, stock prices, international macro, and cross-country growth analysis. In all these papers, G.S. made serious attempts to grapple with substantive and important issues of the day. However, one common characteristic that flows through all these papers is that they unfailingly reflect the discriminating judgement of a consummate econometrician.

G.S. had the gift of a brilliant expositor – the ability to cut through the technical superstructure to reveal only essential details, while retaining the nerve centre of the subject matter he sought to explain. He loved to write econometrics in plain English. There was magic in how he could cut to the core, strip away all the irrelevant details and illuminate the essence of the issue in a quiet and unassuming way. This exceptional expository capability made him revered by applied and theoretical econometricians alike. This skill was apparent in all his writing and was a central element in his textbook expositions. His 1977 econometrics text redefined the boundaries of econometrics that could be integrated into graduate teaching, and became a new standard for subsequent econometrics textbooks. His advanced undergraduate textbook *An Introduction to Econometrics* has gone into its third edition (2000), and all his textbooks have been translated into a number of foreign languages.

G.S.'s style was to take a critical but constructive look at evolving econometric techniques – in particular those that have little practical significance. In this, G.S. had something that was close to perfect pitch in econometrics. He was one of the few econometricians who constantly asked whether the questions being answered were worth asking – always maintaining a clear perspective on a wide range of issues in econometrics and their relationship to economic problems. In doing so, he never hesitated to go against the tide of the profession. While much of his work was undoubtedly constructive, much was also critical of many current fads in econometrics. That is also a very important contribution.

See Also

- Bayesian econometrics
- bootstrap
- categorical data
- distributed lags
- econometrics
- elasticity of substitution
- expectations
- fixed effects and random effects
- inequality (measurement)
- rational expectations
- Roy model
- selection bias and self-selection
- shrinkage-biased estimation in econometrics

- time series analysis

I am grateful to Anthony Davies, Cheng Hsiao, Kay, Tara and Vivek Maddala, Thad Mirer, Peter Phillips and others who have contributed to this biography.

Selected works

1962. (With Z. Griliches, R.E. Lucas and N. Wallace.) Notes on estimated aggregate consumption functions. *Econometrica* 30, 491–500.

1965. Productivity change in the bituminous coal industry. *Journal of Political Economy* 73, 352–65.

1966. (With J.R. Kadane.) Notes on the estimation of elasticity of substitution. *Review of Economics and Statistics* 48, 340–4.

1967. (With J.R. Kadane.) Estimation of returns to scale and elasticity of substitution. *Econometrica* 35, 419–23.

1971a. Generalized least squares with an estimated covariance matrix. *Econometrica* 39, 23–33.

1971b. On the use of variance component models in pooling cross-section and time series data. *Econometrica* 39, 341–58.

1971c. The likelihood approach to pooling cross-section and time series data. *Econometrica* 39, 939–53.

1971d. Simultaneous equations methods for large and medium-sized econometric models. *Review of Economic Studies* 38, 435–45.

1971. (With A.S. Rao.) Maximum likelihood estimation of Solow's and Jorgenson's distributed lag models. *Review of Economics and Statistics* 53, 80–8.

1972. (With D.M. Grether.) On the asymptotic properties of two-step procedures used in the estimation of distributed lag models. *International Economic Review* 13, 737–44.

1973. (With D.M. Grether.) Errors in variables and serially correlated residuals in distributed lag models. *Econometrica* 41, 255–62.

1973. (With A.S. Rao.) Tests for serial correlation in regression models with lagged dependent variables and serially correlated errors. *Econometrica* 41, 255–62.

1974a. Some small sample evidence on tests of significance in simultaneous models. *Econometrica* 42,

841–51.

1974b. Weak priors and sharp posteriors in simultaneous equation models. *Econometrica* 44, 345–51.

1974. (With F.D. Nelson.) Maximum likelihood methods for models of markets in disequilibrium. *Econometrica* 42, 1013–30.

1976. (With L.-F. Lee.) Recursive models with qualitative endogenous variables. *Annals of Social and Economic Measurement* 5, 525–45.

1976. (With S.K. Singh.) A function for size distribution of incomes. *Econometrica* 44, 963–70.

1977. *Econometrics*. New York: McGraw Hill.

1983a. Methods for models of markets with bounded price variation. *International Economic Review* 24, 361–78.

1983b. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

1983. (With R.P. Fishe and K. Lahiri.) A time series analysis of popular expectations data. In *Economic Applications of Time-Series Analysis*, ed. A. Zellner. Washington, DC: U.S. Census Bureau.

1985. (With J.S. Shonkwiler.) Modeling expectations of bounded prices: an application to the market for corn. *Review of Economics and Statistics* 67, 697–702.

1992. (With J. Jeong.) On the exact small sample distribution of the instrumental variable estimator. *Econometrica* 60, 181–3.

1992. (With S. Donald.) A note on the estimation of limited dependent variable models under rational expectations. *Economics Letters* 38, 17–23.

1994. *Econometric Methods and Applications: Selected Papers of G.S. Maddala*, 2 vols. Aldershot: Edward Elgar.

1996. (With H. Li.) Bootstrapping time series models (with discussion). *Econometric Reviews* 15, 115–95.

1998. (With I.-M. Kim.) *Unit Roots, Cointegration and Structural Change*. Cambridge: Cambridge University Press.

1999. (With S. Wu.) A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics* 61, 631–52.

2000. *An Introduction to Econometrics*, 3rd edn. Chichester: Wiley.

Bibliography

Chao, J.C. and Phillips, P.C. 2002. Jeffreys prior analysis with simultaneous equations models in the case with n+1 endogenous variables. *Journal of Econometrics* 111, 251–82.

Griliches, Z. 1999. Forward. In *Analysis of Panels and Limited Dependent Variable Models* (essays in honor of G.S. Maddala), ed. C. Hsiao et al. Cambridge: Cambridge University Press.

Hsiao, C. 2003. In Memoriam: G.S. Maddala. *Econometric Reviews* 22, vii–viii.

Hsiao, C., Lahiri, K., Lee, L.-F. and Pesaran, M.H. 1999. *Analysis of Panels and Limited Dependent Variable Models* (essays in honor of G.S. Maddala). Cambridge University Press.

Lahiri, K. 1999. ET Interview: Professor G.S. Maddala. *Econometric Theory* 15, 753–76.

Lahiri, K. and Phillips, P.C. 1999. Obituary: G.S. Maddala, 1933–1999. *Econometric Theory* 15, 639–41.

Pagan, A.R. 1986. Two stage and related estimators and their applications. *Review of Economic Studies* 53, 517–38.

Phillips, P.C. 2006. A remark on bimodality and weak instrumentation in structural equation estimation. *Econometric Theory* 22, 947–60.

Rosen, S. 2000. G.S. Maddala Memoir. First G.S. Maddala Lecture. Department of Economics, Ohio State University, Columbus, 26 April.

How to cite this article

Lahiri, Kajal. "Maddala, G.S. (1933–1999)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000349> doi:10.1057/9780230226203.1010

The New Palgrave Dictionary of Economics Online

Mahalanobis, Prasanta Chandra (1893 – 1972)

Sukhamoy Chakravarty

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

India, economics in; Mahalanobis, P. C.; optimal growth; planning; Prebisch, R.; statistical inference; two-sector models; Fel'dman, G. A.; dual development thesis

Article

Mahalanobis was born in Calcutta of a well-to-do Bengali middle-class family with a reformed outlook on Hindu religion. He was educated first at Presidency College, Calcutta, and then at Cambridge, where he graduated with a First in Natural Sciences from King's College in 1915. He became a Fellow of the Royal Society in 1946 and received many other scientific honours. While Mahalanobis served as Professor of Physics at Presidency College for nearly three decades, his scientific work consisted chiefly of developing statistical theory and techniques that had application to a wide range of subjects, beginning with meteorology and anthropology and ending in economics.

Mahalanobis established a firm international reputation on the basis of his work on the design of large-scale sample surveys (for example, 1944) and thus laid the basis for systematic collection of a large variety of data relating to socio-economic conditions. Mahalanobis's sense of realism was combined with a deep understanding of the problems of statistical inference. This led him to place stress on ‘non-sampling errors’ in addition to the standard preoccupation with sampling errors. He devised his system of ‘interpenetrating network of sub-samples’ to derive among other things, an idea of ‘non-sampling errors’ which are inherently associated with large-scale collection of data.

Mahalanobis's work on experimental designs developed with a view to estimating crop yields (1946) was highly influential in laying down the basis for collection of agricultural statistics in India. In multivariate statistics, Mahalanobis's measure of distance between two populations (1936), usually known as Mahalanobis's D^2 statistic, is a major contribution that is much used in anthropometry and elsewhere.

Mahalanobis maintained a keen interest in problems of national planning even before India had gained Independence. He recognized very early that such planning had to have a firm statistical base, and from the beginning of the 1950s, when the Indian Five Year Plan was launched, began to devote a very large

part of his time and attention to questions of estimating national income and the factors determining its rate of growth. His approach to planning issues, with its strong emphasis on quantification, was significantly different from the qualitative approach favoured by the Indian economists of his generation. However, Mahalanobis was no exclusive believer in narrowly conceived quantitative techniques. He developed an important blend between qualitative and quantitative considerations, which is reflected in his ‘Approach of Operational Research to Planning in India’ (1955).

The second Five Year Plan, whose analytical structure was largely the handiwork of Mahalanobis, stands out as a very distinguished document in the development of planning theory. Mahalanobis is generally regarded as one of the prominent advocates of the inward-looking strategy of industrialization, along with Raul Prebisch. But the analytical foundation of the Mahalanobis approach was derived from somewhat different premises. While Prebisch began his theoretical study from what he thought was a historical fact, that is, the secular decline in terms of the trade of primary producing countries, Mahalanobis developed a two-sector model of growth to deduce a strategy of industrial development which he thought was best suited to India. The classification of the economy into sectors resembled in some respects Marx's famous Departmental Schema, although they were not identical.

Mahalanobis's sector-schema (1953) distinguished between ‘capital goods’ and ‘consumer goods’, but the assumption of vertical integration made in the interest of simplicity made statistical implementation difficult. The essential point of the model is that the capacity of the capital goods sector determines the potential rate of expansion of the consumer goods sector, and not the other way round. Further, at any given instant, capacities are not directly transferable from one sector to the other. Labour is not considered to be a constraint on expanding production. The model was developed initially for a closed economy but has been subsequently extended to open economies, with an exogenously given profile of export earnings. Mahalanobis used the model to illustrate the nature of the trade-off between present and future consumption, given the objective characteristics of the two sectors.

For the dynamic closure of the model he used the ratio of the output of the capital goods sector that is ploughed back into itself (‘ λ_k’ in his notation), to deduce a ‘gradualist growth’ path of consumption. For any given value λ_k maintained over time, the rate of growth of aggregate output tends, over a sufficiently long period, to a magnitude $\lambda_k \beta_k$, where β_k is the output–capital ratio of the capital goods sector. The Mahalanobis model was subsequently freed from the assumption of an exogenously stipulated λ_k Exercises carried out by Stoleru (1965), Chakravarty (1969), Dasgupta (1969) and others introduced explicit intertemporal social utility functions along with a production technology of the Mahalanobis type. They deduced the characteristics of optimal growth paths with the help of variational calculus. $\lambda_k(t)$ was deduced as a solution of the optimizing exercise. It was shown that while the assumption of ‘non-shiftability’ critical to Mahalanobis's model could in several cases give rise to a preference for capital goods sector in early stages of growth (a strategy preferred by Mahalanobis himself), one could not obtain a universal rule of priority for capital goods irrespective of initial conditions, or the nature of social utility functions over time.

In all these exercises, the coefficients pertaining to the ‘capital goods sector’, sometimes identified as the ‘machine tool sector’, turned out to be an important determinant of the growth process. Earlier literature on business cycle theory originating with Marx, Tugan Baranovsky and Adolph Lowe had placed emphasis on the ‘machine tools sector’, without linking it up with an explicit growth model. In the growth-theoretic area Fel'dman alone appears to be the true predecessor of Mahalanobis, as is evident

from Domar's discussion (1957).

Mahalanobis extended the two-sector model to a four-sector model, to focus on issues of reduction in unemployment along with increases in income. Mahalanobis came to the ‘dual development thesis’, which consisted in assigning high weights to the capital goods sector in the interests of long-term growth, and emphasis on the highly labour-intensive consumer goods sector in the short run. In the literature on planning, this has on occasion been referred to as the strategy of ‘walking on two legs’, with authorship occasionally ascribed to Mao Tse Tung.

Towards the end of his life, Mahalanobis returned to issues of statistical methodology and concentrated on developing what he called ‘fractile graphical analysis’ (1960), which is based on a geometrical concept of error and can also provide a generalized measure of separation between two ‘different universes’ of study.

Mahalanobis's work remains important for economists who are working on quantitative approaches to problems of plan formulation, especially in the context of large-sized economies. His work on sample surveys has generated a very valuable literature to which economic statisticians from India and elsewhere have made notable contributions.

See Also

- development economics
- Fel'dman, Grigorii Alexandrovich

Selected works

1936. On generalized distance in statistics. National Institute of Science, India, *Proceedings* 2, 49–55.

1944. On large scale sample surveys. *Philosophical Transactions*, Series B 231, 392–451. London: The Royal Society.

1946. Sample surveys of crop yields. *Sankhya* 7, 269–80.

1953. Some observations on the process of growth of national income. *Sankhya* 12, 307–12.

1955. The approach of operational research to planning in India. *Sankhya* 16(1–2), 3–130.

1958. Science and national planning. *Sankhya* 20, 69–106.

1960. A method of fractile graphical analysis. *Econometrica* 28, 325–51.

1961. *Talks on Planning*. Bombay and London: Asia Publishing House.

Bibliography

Bhagwati, J.N. and Chakravarty, S. 1969. Contributions to Indian economic analysis – a survey. *American Economic Review* (Supplement) 59(4), 1–73.

Chakravarty, S. 1969. *Capital and Development Planning*. Cambridge, MA: MIT Press.

Dasgupta, P.S. 1969. Optimum growth when capital is non-transferable. *Review of Economic Studies* 36 (1), 77–88.

Domar, E.D. 1957. *Essays in the Theory of Economic Growth*. New York and Oxford: Oxford University Press.

Rao, C.R. 1973. Prasanta Chandra Mahalanobis. In *Biographical Memoirs of the Fellows of Royal Society* 19. London: The Royal Society.

Stoleru, L.G. 1965. An optimal policy for economic growth. *Econometrica* 33, 321–48.

How to cite this article

Chakravarty, Sukhamoy. "Mahalanobis, Prasanta Chandra (1893–1972)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000013> doi:10.1057/9780230226203.1011

The New Palgrave Dictionary of Economics Online

Makower, Helen (1910 – 1998)

K.J. Lancaster

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

consumer theory; household production; linear methods; Makower, H.

Article

Helen Makower was educated at Cambridge and obtained her doctorate from London University. From 1938 until her retirement in 1973 she taught at the London School of Economics and Political Science. In collaboration with Jacob Marschak she made a pioneering contribution to modern asset portfolio theory and to the study of labour mobility. After the Second World War her analytical insights and interest in work then being performed at the Cowles Commission in Chicago led to her being one of the important links through which such techniques as activity analysis entered the academic scene in Britain. Her 1957 book and other papers made original contributions to the application of linear methods in economic analysis. One of her important insights was into the analogy between production and consumption, a precursor of later work on the household production and characteristics approaches to consumer theory.

Selected works

1938. (With J. Marschak.) Assets, prices and monetary theory. *Economica* N.S. 5(August), 261–88.

1939–40. (With J. Marschak and H.W. Robinson.) Studies in the mobility of labour: analysis for Great Britain. Pt. I–II. *Oxford Economic Papers* 2(May 1939), 70–97; 4(September 1940), 39–62.

1950. The analogy between producer and consumer equilibrium analysis. *Economica* N.S. 17(February), 63–8.

1957. *Activity Analysis and the Theory of Economic Equilibrium*. London: Macmillan.

How to cite this article

Lancaster, K. J. "Makower, Helen (1910–1998)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000016> doi:10.1057/9780230226203.1012

The New Palgrave Dictionary of Economics Online

Malthus, Thomas Robert (1766 – 1834)

J.M. Pullen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A brief biographical sketch precedes a discussion of important methodological features of Malthus's work, especially his ‘doctrine of proportions’ and the need for moderation and balance in economic principles and policies. His principle of laissez-faire admitted exceptions; and, although his principle of population warned of over-population, he acknowledged the potential advantages of population growth. His ideas on the Poor Laws, the Corn Laws, Say's Law, and the relation between saving and investment are discussed; and the roles given to effective demand and to distribution as a factor of production, especially the distribution of landed property, are emphasized.

Keywords

classical economics; Cobbett, W.; Condorcet, M. de; Corn Laws; distribution theories: classical; doctrine of proportions; economic growth; effective demand; free trade; Godwin, W.; happiness; hoarding; Keynes, J. M.; laissez-faire; living standards; Malthus, T. R.; Malthus's theory of population; Marx, K. H.; Mill, J.; Old Poor Law; optimum; perfectibilism; population growth; protectionism; redistribution of wealth; saving–investment equality; Say, J.-B.; Say's Law; unproductive consumption

Article

Malthus has the unusual distinction not only of being a founder of classical economics – mainly because of his principle of population – but also of being instrumental in attempts to overthrow classical economics, mainly because of his principle of effective demand and its influence on John Maynard Keynes.

The most comprehensive and authoritative source of biographical information on Malthus is James (1979), from which the following brief details have been largely derived. Additional information can be found in the first edition of *The New Palgrave: A Dictionary of Economics* (Pullen, 1987), Malthus (1989b, pp. xv–lxix), and the *Oxford Dictionary of National Biography* (Pullen, 2004). Malthus was

born on 13 February 1766, near Wotton, in the county of Surrey, England, and died on 29 December 1834, at Bath. He was buried in Bath Abbey where there is a commemorative plaque. Although he was baptized Thomas Robert, he used his full name only in formal situations; in less formal correspondence he signed himself T. Robert Malthus or Robert Malthus, and was known to family and close friends as Robert or Bob.

He was the son of Henrietta (née Graham) (1733–1800) and Daniel Malthus (1730–1800). The latter, having inherited independent means, cultivated literary, artistic, scientific and theatrical interests. He was an admirer and correspondent of Rousseau, who once visited the family home soon after Malthus's birth. The extensive library of Daniel Malthus was eventually passed on to Malthus and, supplemented by acquisitions of his own and other family members, is now held in Jesus College, Cambridge.

Malthus graduated in 1788, and in 1789 was ordained deacon with title to a stipendiary curacy at the small chapel at Okewood in the parish of Wotton. He was ordained priest in 1791, was appointed non-resident Rector of Walesby in Lincolnshire in 1803, and succeeded to the perpetual curacy of Okewood in 1824. He married in 1804 and had three children, but no grandchildren. In 1805 he was appointed to the East India College as ‘Professor of General History, Politics, Commerce and Finance’, a title later altered to ‘Professor of History and Political Economy’. He held the post for the rest of his life, residing in the College at Haileybury, near Hertford. As well as performing his teaching duties, he preached regularly in the college chapel. The important collection of Malthus manuscripts held at Kanto Gakuen University in Japan (Malthus, 1997; 2004) contains four of his sermons. They corroborate the statement of his colleague William Empson: ‘Mr. Malthus was a clergyman – a most conscientious one, pure and pious. We never knew one of this description so entirely free of the vices of his caste’ (Empson, 1837, p. 481). His main publications were *An Essay on the Principle of Population*, first published in 1798, with five further editions in 1803, 1806, 1807, 1817 and 1826, and *Principles of Political Economy*, first published in 1820 with a posthumous second edition in 1836. He also published at least 20 smaller works – his authorship of a 21st is disputed – and evidence he gave at two public enquiries can be found in the published reports. There is a full list of his publications in *The New Palgrave* (Pullen, 1987) or in Malthus (1986, vol. 1, pp. 41–4). He engaged in extensive correspondence throughout his career, with Ricardo and many others. More than 230 letters to and from over 50 correspondents are known to have survived.

Malthus's methodology: ‘the doctrine of proportions’

Before considering particular aspects of Malthus's political economy, it is important to understand some of the peculiar features of his methodology. Failure to do so has resulted in many misunderstandings and unnecessary disagreements among commentators.

One of the most important, but one of the most unrecognized, aspects of Malthus's methodology was the principle that he called the ‘doctrine of proportions’. This was the traditional ethical notion of the just mean or middle way. As Leslie Stephen (1893) said, in the first *Dictionary of National Biography*, Malthus was always ‘a lover of the golden mean’. The distinctive innovation of Malthus lay in applying the concept to political economy, and in giving it such a prominent and consistent role.

He stated that his aim was to show ‘how frequently the doctrine of proportions meets us at every turn, and how much the wealth of nations depends upon the relation of parts’. It was his view that ‘all the

great results in political economy, respecting wealth, depend upon *proportions*', and warned that the 'tendency to extremes is one of the great sources of error in political economy, where so much depends upon proportions'. He added that 'It is not, however, in political economy alone that so much depends upon proportions, but throughout the whole range of nature and art' (1989b, vol. 1, pp. 352, 432; vol. 2, pp. 252, 269, 278).

Malthus's doctrine of proportions is thus essentially the same as the concept of the optimum. Although he did not use the term 'optimum', he must be recognized as having been one of the first to introduce the concept of the optimum into economics. In giving this central role to the doctrine of proportions, he has in effect said that *the* economic problem is the problem of balance, not the problem of choice.

But, despite his widespread use of the doctrine of proportions, Malthus recognized that precise determination of optimum points would be difficult. In discussing the optimum level of saving, he acknowledged that 'the resources of political economy may not be able to ascertain it' (1989b, vol. 1, p. 9), and, in discussing the just means for saving and the division of landed property, he said 'the extremes are obvious and striking, but the most advantageous mean cannot be marked' (1989b, vol. 1, p. 10).

The moderation and balance implied by the doctrine of proportions was evident in Malthus's personal temperament. Bishop Otter, who knew Malthus for nearly 50 years, said that he 'scarcely ever saw him ruffled, never angry, never above measure, elated or depressed', and that Malthus possessed 'a degree of temperance and prudence, very rare at that period, and carried by him even into his academical pursuits' (in Otter, 1836, pp. xxxii, xlix); and William Empson said in reference to the doctrine of proportions: 'The lesson which he sought to impress on others, he faithfully applied to himself; and so successfully, that few characters have ever existed of more perfect symmetry and order' (Empson, 1837, pp. 476–7).

Malthus has been given credit for introducing or propagating, either alone or with others, a number of key ideas in the history of economics; notably, the principle of population, the law of diminishing returns, and the role of effective demand. The doctrine of proportions could be added to the list.

Limitations and exceptions

Another facet of Malthus's methodology was his insistence on limitations and exceptions to the general principles in political economy. This could be seen either as a corollary of his doctrine of proportions or as another way of expressing the same doctrine. He believed that there are some general principles in political economy to which exceptions are 'most rare', but added 'yet there is no truth of which I feel a stronger conviction than that there are many important propositions in political economy which absolutely require limitations and exceptions' (1989b, vol. 1, p. 8). In this respect, he departed from the absolutist and universalist aspirations of some of his contemporaries, who, anxious to promote the scientific credentials of political economy, pretentiously declared them to be 'laws'. He was critical of the 'precipitate attempt to simplify and generalize', which he regarded as the 'principal cause of error, and of the differences which prevail at present among the scientific writers on political economy' (1989b, vol. 1, pp. 5–6).

Malthus has been accused, in his own day and now, of lacking in logic, especially by comparison with Ricardo. The accusation that his views did not constitute a logical and coherent system appears to have emanated from a failure to appreciate that his views were formulated in the context of his doctrine of

proportions, and that he believed exceptions and limitations frequently have to be admitted when principles are used to formulate policies for application to particular real-world circumstances.

Laissez-faire and government intervention

Malthus strongly supported the principle of laissez-faire or freedom of trade: ‘the wealth of nations is best secured by allowing every person, as long as he adheres to the rules of justice, to pursue his own interest in his own way’, and ‘governments should not interfere in the direction of capital and industry, but leave every person, so long as he obeys the laws of justice, to pursue his own interest in his own way’. He described this as a ‘great principle’ and as ‘one of the most general rules of political economy’ (1989b, vol. 1, pp. 3, 13, 518).

But he also argued that some exceptions to the principle of laissez-faire have to be recognized, and that the principle of non-interference is ‘necessarily limited in practice’ (1989b, vol. 1, 18–19, 525). He believed that there are certain duties that belong to the government – for example, in areas such as education; support of the poor; construction and maintenance of roads, canals, and public docks; colonization and emigration; and the support of forts and establishments in foreign countries – although he recognized that there may be differences of opinion about the extent to which government should share in such matters. In particular, the ‘necessity of taxation … impels the government to action, and puts an end to the possibility of letting things alone’ (1989b, vol. 1, pp. 18–19).

Thus, although Malthus strongly supported the principle of laissez-faire, his support, like that of Adam Smith, was pragmatic and conditional rather than dogmatic and absolute. There was, however, a major difference in their conception of the laissez-faire principle. In what Donald Winch has described as ‘an attack on a central feature of the *Wealth of Nations*’ and as ‘a major qualification to Smith's system of natural liberty’, Malthus doubted whether economic growth has always been, or will always be, advantageous to the mass of society. Malthus criticized Smith's view that the economic growth of Britain during the 18th century had improved the living standards of the labouring classes; he recognized that investments in trade and manufacturing had benefited individual capitalists, but argued that they were of less benefit to society as a whole. Thus Malthus raised the possibility of conflict between economic growth and human happiness, and implied that interventionist welfare policies by government might be justified. In this respect, as Winch has argued, ‘if general allegiance to the system of natural liberty, as interpreted by Smith and upheld under the different circumstances by some of his followers, is the hallmark of an orthodox political economist during the first half of the nineteenth century, Malthus occupies a decidedly ambivalent position’ (Winch, 1987, pp. 32, 59–61, 76–7).

Population

Malthus's first published work – *An Essay on the Principle of Population* (1798) – was written primarily to controvert the perfectibilist notions of Godwin and Condorcet. He believed that the growth of population presented a major obstacle to unlimited human progress. He argued that population will constantly tend to exceed the food supply, with the result that human progress will be neither rapid nor unlimited, and will be accompanied by sufferings and evils arising from the operation of unavoidable checks to population growth.

To support his views on the threat of overpopulation to human progress, Malthus introduced the notion of the two ratios. He argued that population will tend to increase in a geometrical ratio (1, 2, 4, 8 …) doubling every 25 years; but the food supply will increase only in an arithmetical ratio (1, 2, 3, 4 …). He believed that the population of ‘this Island’ was then about seven million, and that after the first 25 years it would reach 14 million, after 50 years 28 million, after 75 years 56 million, and so on. But the utmost that could be expected for the supply of food is that there would be sufficient to feed 14 million after 25 years, 21 million after 50 years, 28 million after 75 years, and so on. Thus, after the first 25 years, the food supply would become insufficient, and any further progress in the size of the population and the standard of living would be impossible. He concluded that this argument is conclusive against the perfectibility of the mass of mankind.

Opinions differ on whether Malthus's principle of population depends essentially on the empirical accuracy of these ratios or whether they were intended merely as approximate tendencies, or as a mathematical metaphor. Whatever his intention, there is no doubt that the ratios have exerted a powerful rhetorical influence in promoting his message and his fame.

Malthus was not the first writer to issue a warning about the dangers of overpopulation, as he himself acknowledged, but for a variety of reasons his arguments have become the best-known, and have exerted a great influence on human thought and human affairs. He alerted the world to the problem of overpopulation, and his views continue to affect the population policies of governments through the world today.

Having presented his basic arguments in the first two chapters of the *Essay*, Malthus then proceeded to discuss the ‘checks’ to population. As he said in his first postulate, people cannot live without food, and therefore it would be impossible to have a situation where 28 million people were in existence but the food supply was adequate for only 21 million. There must therefore be some mechanisms or checks whereby populations are prevented from exceeding the food supply. The bulk of the *Essay*, especially in the much enlarged later editions, was devoted to a detailed description of the checks that have operated in different countries and at different times.

He classified the checks as either positive checks that reduce normal life expectancy and increase the death rate, or preventive checks that reduce the birth rate. Among his list of positive checks he included common diseases, epidemics, wars, plagues, pestilence, famines, infanticide, unwholesome occupations and habitations, severe labour, exposure to the seasons, extreme poverty, bad nursing of children, great cities, and excesses of all kinds.

The preventive checks, described in circumspect language (‘vicious customs with respect to women’), included prostitution and birth control, but the only preventive check that he approved of and advocated was prudential restraint, by which he meant delaying marriage until sufficient resources of food, accommodation and other necessaries are available to provide the parents and the expected number of children with an acceptable standard of living. He noted that prudential restraint is practised, and should be practised, by those who want to maintain after marriage the social and economic status they enjoyed before marriage. The case for prudential restraint was even more vigorously argued in the later editions, where those who marry and raise children without ensuring that they have sufficient resources are accused of irresponsible and immoral behaviour.

He also classified the population checks as either vice or misery, but did not clearly show how the vice-and-misery classification is related to the positive-and-preventive. Presumably he meant that, among the positive checks, some, such as war and infanticide, are vices, and all lead to misery; and among the

preventive checks all except prudential restraint are vices, and all are likely to lead to misery; and although prudential restraint is a virtue, not a vice, it often leads to vice.

> Restraints upon marriage are but too conspicuous in the consequent vices that are produced in almost every part of the world; vices, that are continually involving both sexes in ‘inextricable unhappiness’. (1986, vol. 1, p. 28)

In admitting that prudential restraint might also be a cause of misery, he might have been speaking from personal experience, being in 1798 a 32-year-old bachelor with an income as a curate insufficient to support a wife and family in a socially acceptable manner.

In the second and later editions, he introduced the expression ‘moral restraint’, by which he meant prudential restraint conducted in accordance with Christian moral precepts regarding premarital sex, but the *concept* of moral restraint is *implicit* in the first edition. It is unlikely that, in advocating prudential restraint, Malthus as a Protestant clergyman would have intended to condone prudential restraint that was accompanied by immoral sexual behaviour.

In the second (and later) editions of the *Essay*, he softened some of the harshest conclusions of the first by arguing that, if people could be made aware of the harm done by improvident procreation, then moral restraint, though still a difficult challenge, could be practised without causing misery and without leading to vicious practices. He objected to contraception on moral grounds and also because, by facilitating control of the birth rate and reducing the pressure of population, it would remove one of the incentives needed to overcome our natural indolence, to promote economic growth, and to encourage the ‘growth of mind’. It is ironical that the expression ‘Malthusian practices’ became synonymous with contraception, and that contraception has become the method most commonly adopted throughout the world to control population. The world has responded to Malthus's warnings of the danger of overpopulation by adopting a remedy he strongly rejected.

Arguments in favour of population growth

The popular and superficial view of Malthus is that he was opposed to population growth. But there are numerous instances in his writings which show that he regarded an increase of population, under certain conditions, as desirable in itself, and as a necessary cause of economic growth. For example, he spoke of the ‘pursuit of the desirable object of population’ (1986, vol. 3, p. 455); and, referring to the possibility of a great increase of population in Ireland in the 19th century, he said ‘so great an increase of human beings, if they could be well supported, would be highly desirable’ (1986, vol. 4, p. 32). In a similar vein he said: ‘That an increase of population, when it follows in its natural order, is … a great positive good in itself, … I should be the last to deny’ (1989a, vol. 1, p. 439). And those who use Malthus to support a policy of population reduction forget that on one occasion he argued that a diminution of population would be harmful: ‘It is evidently therefore regulation and direction which are required with regard to the principle of population, not diminution or alteration’ (1989a, vol. 2, p. 94).

Some of his most forceful statements in favour of population growth occurred in the appendices added to the third (1806) and fifth (1817) editions of the *Essay*, in response to critics who had accused him of being anti-population. The fact that these appendices have been omitted from some modern reprints of the *Essay* might explain the limited awareness of his pro-population ideas.

Malthus's pro-population views can even be seen when he was advocating prudential restraint: ‘Prudential habits with regard to marriage carried to a considerable extent, among the labouring classes of a country mainly depending upon manufactures and commerce, might injure it’ (1989b, vol. 1, p. 236; vol. 2, p. 215). This is a surprising argument, given that he had said that the preventive check of prudential restraint should be a principal remedy for overpopulation. It shows that he wished the doctrine of proportions to be applied as a check to the preventive check!

Malthus's pro-population views can also be seen in his statements on population as a *necessary* cause of economic growth. He admitted that population growth alone will not promote economic growth; for example, he argued that ‘the increase of population alone … does not furnish an effective stimulus to the continued increase of wealth’ (1989b, vol. 1, pp. 347–8), ‘population alone cannot create an effective demand for wealth’ (1989b, vol. 1, p. 350) and ‘encouragements to population … will not alone furnish an adequate stimulus to the increase of wealth’ (1989b, vol. 1, p. 351). But he also stated:

> That a permanent increase of population is a powerful and necessary element of increasing demand, will be most readily allowed. (1989b, vol. 1, p. 347; in the second edition, ‘permanent’ was changed to ‘continued’)

and

> That an increase of population, when it follows in its natural order, is … absolutely necessary to a further increase in the annual produce of the land and labour of any country, I should be the last to deny. (1989a, vol. 1, p. 439)

In other words, although Malthus recognized that population growth is not a *sufficient* cause of economic growth, he nevertheless regarded it as a *necessary* cause.

In some circumstances, according to Malthus, an increase in population will bring about a decrease in living standards; but in other circumstances it will bring about an increase in living standards, and a decrease in population will bring about a decrease in living standards. Living standards can be both a direct and an inverse function of population. Some critics would regard this as self-contradictory, as proof of his lack of logic, and as a justification for William Cobbett's epithet ‘muddle-headed Malthus’. Others would see it as a reasonable, parabolic application of the doctrine of proportions.

Theological aspects of the principle of population

The early chapters of the first edition of the *Essay* have a rather pessimistic tone. They appear to be saying that the pressure of population against the food supply will keep the mass of the population at or near subsistence level, and that this struggle between food and population will be accompanied by miseries and vices. However, in the last two chapters of the first edition of the *Essay* Malthus explored the theological implications of his principle of population. His published contributions in theology are too limited for him to be considered as a theologian in a professional sense, but his theological views are interesting in their own right, because of their heterodox nature, and because they seem to have been presented, not as a mere afterthought or pious homily, but in an attempt to integrate his principle of

population into a comprehensive world view, in opposition to that of Godwin and Condorcet. As a Christian minister he would have been concerned to show that his view of population did not conflict with Christian ideas about the nature of God. He had due cause for concern. In saying that misery and vice can come from obeying the biblical injunction to go forth and multiply, he was accused by some critics of blasphemously denying the Creator's omnipotence, omniscience and benevolence.
However, in the last two chapters of the first edition of the *Essay* he argued, on the contrary, that population pressure is providentially ordained by God as a means whereby human development (‘the growth of mind’) is stimulated. He argued that the constant pressure of population against food supply, although it might produce some moral and physical evils, would also produce an overbalance of good. The first edition of the *Essay* thus finished on a note of moral and theological optimism.
The last two chapters were omitted from subsequent editions of the *Essay*. Comments contained in his correspondence (1997, pp. 73–7) and remarks from other contemporaries indicate that the omission occurred at the instigation of friends. Some commentators interpret the omission as a recantation. Others find traces of his theology in the later editions, and argue that his growth-of-mind theology remained an essential, if only implicit, framework throughout all editions of the *Essay*. They argue that to ignore its theological aspect is to ignore an essential element of his total population theory and, contrary to his intentions, to reduce the *Essay* to a mere economic or political tract.

Poor Laws

Although Malthus believed that population pressure was a phenomenon common to most societies, he argued that the problem had been exacerbated in England by the Poor Laws. They were intended to alleviate poverty, but only succeeded in creating the poor they sought to maintain. They encouraged people to marry too early and have large families, in the expectation that food and accommodation would be provided for them; and they discouraged hard work and the development of productive skills. In his earlier writings Malthus had argued for abolition of the Poor Laws, both as a principle and as a practical policy; but in later writings and in correspondence (see James, 1979, p. 450; Winch, 1996, pp. 320–1) his position moved from complete abolition to gradual abolition, and then to administrative reform, arguing that a fundamental change involving complete abolition would present practical and political difficulties, and that the most that could be achieved in the current circumstances would be an amelioration of the present system through improved administration.
This is an example of his insistence on the need for limitations and exceptions in the practical application of general principles. He did not see any contradiction in subscribing to the idea in principle while at the same time rejecting it as a practical policy for a particular place and time. Another example of this feature of his methodology occurred in his views on the Corn Laws.

Corn Laws

Although Malthus strongly supported the principle of laissez-faire, he published a pamphlet in 1815 supporting the retention of the Corn Laws which prohibited for example, the import of wheat when the home price fell below 80 shillings a quarter. This radical departure from laissez-faire caused dismay among other political economists and among his Whig friends who opposed the protectionist policy of

the Tory government. In admitting this exception to the principle of laissez-faire, he was in effect reaffirming his view that, unlike the laws of mathematics, the principles of political economy should not be applied in an absolutist and universalist manner.

It has been argued that in his later years Malthus changed his mind and recanted his earlier support for the Corn Laws. The arguments for and against this change-of-mind hypothesis have been elaborated elsewhere (Hollander, 1992; 1995; Pullen, 1995), and are too detailed to be repeated here. It would probably be fair to say that, on the basis of the textual and contextual evidence so far presented, there is no clear, unambiguous statement of a recantation by Malthus. But it should also be said that Malthus was strongly in favour of the principle of free trade, and that he strongly regretted the need for an exception in the case of the Corn Laws. It is obvious from his writings and correspondence that, if the circumstances that necessitated the exception were removed, he would have gladly removed his support for agricultural protection.

Economic growth, effective demand and Say's Law

Malthus's views on economic growth are to be found scattered throughout his many publications, with his most systematic (but not comprehensive) treatment of this topic in the final chapter of the *Principles*, namely. chapter 7, ‘On the Immediate Causes of the Progress of Wealth’. He divided the immediate causes of progress into two categories: ‘the powers of production’ and ‘the means of distribution’. On the production side he discussed four causes: population, accumulation, soil fertility and inventions (which, by combining the second and the fourth, could be reclassified as labour, capital and land). On the distribution side he discussed three causes: the division of landed property, commerce (internal and external), and unproductive consumers. His views on the production side were unremarkable at the time, and would be quite acceptable in standard texts today, but his views on the distribution side have proved to be controversial because of their emphasis on the role of effective demand.

By effective or effectual demand Malthus meant the *power* to purchase at a price sufficient to cover the vendor's costs and required profit, combined with the *willingness* to purchase. His distinction between power and will, or means and motives, was a recurring theme in his political economy. He stressed that production requires more than the power to produce; it requires also the motive to produce, which comes from effective demand.

> … the powers of production, to whatever extent they may exist, are not alone sufficient to secure the creation of a proportionate degree of wealth. Something else seems to be necessary in order to call these powers fully into action. This is an effectual and unchecked demand for all that is produced. (1989b, vol. 1, p. 413; vol. 2, pp. 263, 447)

The powers of production will be ‘called into action, in proportion to the effective demand for them’ and ‘General wealth, like particular portions of it will always follow effective demand (1989b, vol. 1, pp. 414, 417). In effect he was saying that demand-side forces are as powerful and as necessary as the supply-side forces of natural resources, capital accumulation, division of labour, and so on.

He believed that an important cause of an adequate level of effective demand was the existence of a body of ‘unproductive consumers’, who purchase material products but do not produce material

products. They would include menial servants, military personnel, actors, clergymen and other service providers. The concept was completely misunderstood by Ricardo who said that unproductive consumption is as useful as a fire in a warehouse or the destruction of war. Malthus later recognized that the term ‘unproductive’ had pejorative implications, and altered it to ‘the provision of services’. However, the concept could also include those who live on their investments in the national debt and those whose wealth enables them to consume without either producing material goods or providing services, thus inviting Marx's description of Malthus as a protector of the ruling classes and the idle rich.
Malthus's views on effective demand were largely rejected during his lifetime, and largely ignored for the next hundred years. It was generally believed with James Mill, Jean-Baptiste Say and others that the purchasing power generated during the production process would be sufficient for all the products to be sold, that aggregate demand deficiency would never be a cause of economic decline, and that a general glut of products would be impossible. This view, known as Say's Law or Mill's Principle, and popularly expressed as ‘supply creates its own demand’, became a standard theme of classical economics, and still finds its supporters, even though Malthus showed, and Say virtually admitted, that its validity relies on a tautological definition of ‘supply’ and ‘product’. The experience of the depression of the 1930s and the publication of J.M. Keynes's *General Theory* (1936) cast doubt on this conventional wisdom of Say's Law, and rescued Malthus's views on effective demand from oblivion.

Effective demand and the division of landed property

Malthus's views on effective demand as a stimulus to economic progress led him to advocate a wider distribution of wealth, because ‘Practically it has always been found that the excessive wealth of the few is in no respect equivalent, with regard to effective demand, to the mere moderate wealth of the many’ (1989b, vol. 1, p. 431). But this redistribution of property has often been neglected, and sometimes even denied, in the secondary literature. Karl Marx, in particular, misinterpreted Malthus in this regard, and some other commentators appear to have taken their views of Malthus from Marx; and, like Marx, have not bothered to test them against Malthus's text.
Admittedly, there are passages in some parts of Malthus's writings that support a pro-landlordism interpretation. But in other passages he was critical of the distribution of land and other property, and described the existing maldistribution as unjust and as an impediment to economic growth; for example

> A very large proprietor, surrounded by very poor peasants, presents a distribution of property most unfavourable to effective demand … Thirty or forty proprietors, with incomes answering to between one thousand and five thousand a year, would create a much more effective demand for wheaten bread, good meat, and manufactured products, than a single proprietor possessing a hundred thousand a year. (1989b, vol. 2, 373–4)

In his view, ‘the division of landed property is one of the great means of the distribution of wealth’, and without ‘an easy subdivision of landed property … a country with great natural resources might slumber for ages with an uncultivated soil, and a scanty yet starving population’ (1989b, vol. 1, pp. 439–40).
He did not propose that either private property in land or the class of landed proprietors should be abolished. He regarded both as necessary. But he did not regard ‘the present great inequality of property’

as ‘either necessary or useful to society’; and added that ‘On the contrary, it must certainly be considered as an evil, and every institution that promotes it is essentially bad and impolitic’ (Malthus, 1986, vol. 1, p. 102). Less inequality in land ownership would mean that rents would be enjoyed by a larger number of proprietors.
However, he did not wish the division of property to be pushed too far:

> The division and distribution of property, which is so beneficial when carried only to a certain extent, is fatal to production when pushed to extremity. (1989a, I, 372)

This argument is an excellent illustration of his characteristic middle-way methodology. He himself regarded the question of the division of property as the most important application of the doctrine of proportions.

> It will be found, I believe, true that all the great results in political economy, respecting wealth, depend upon *proportions* … But there is no part of the whole subject, where the efficacy of proportions in the production of wealth is so strikingly exemplified, as in the division of landed and other property; and where it is so very obvious that a division to a certain extent must be beneficial, and beyond a certain extent prejudicial to the increase of wealth. (1989b, vol. 1, pp. 432–3)

Distribution as a factor of production

Other writers, before and after Malthus, have discussed production and distribution, but generally their approach has been to regard distribution as the process whereby the proceeds of production are shared out after they have been produced by the factors of production. Their theory of distribution is worked out independently of their theory of production.
Malthus also looked at the problem of distribution in this way, with separate chapters analysing the way in which wages, profits and rents are determined. But in addition he looked at distribution from another direction. For Malthus, distribution is not merely concerned with sharing out the spoils of production. It has a further function. It is an essential determinant of production, and an integral part of the production process considered in its totality. Without a proper distribution, there would be no production – except at a self-subsistence level. He saw distribution as a problem to be resolved *before* (as well as after) production takes place. Whereas others were concerned mainly with how the distribution of the product between wages, profits and rent is affected by economic development, Malthus made a major contribution by stressing that the distribution of the product in turn affects economic development. He was in effect saying, if not in these precise words, that distribution must be regarded as a factor of production, along with the conventional listing of the other factors of production – land, labour and capital. They represent only the supply side of the production process; but, if production is to occur, there must be a motive to produce as well as the means. In an exchange system, there will be no motive for producers to produce unless there are prospects of profits, and there will be no profit prospects unless there is an adequate effective demand for the products. This effective demand from potential consumers will not be forthcoming unless there has been a proper distribution of spending power. As Malthus said,

‘there is certainly no indirect cause of production as powerful as consumption’ (1989b, vol. 2, p. 34). The effective demand generated by a proper distribution provides demanders with the power or means to demand, and this provides suppliers with the will or motive to supply. It is obvious that, unless there has been production, there can be no distribution. But Malthus insisted that maximum production will not be achieved unless an optimum spread of distribution is established.

The separation and dichotomy between production and distribution that is presented in typical textbooks would therefore have been unacceptable to Malthus, for whom any listing of the factors of production would have to include distribution. It is this relationship of reciprocal causation between distribution and production that makes Malthus's theory of distribution innovative and distinctive.

Saving, investment and hoarding

Some commentators have interpreted Malthus as holding that savings are always invested; and have concluded that in Malthus's theory saving is not a leakage from the circular flow, does not constitute a reduction in effective demand, is not an impediment to economic growth, and must always be beneficial. In this respect, they see a major difference between Malthus and Keynes. They deny the claim that Malthus was a precursor of Keynes, and argue that Keynes was mistaken in regarding Malthus as a precursor. This interpretation appears to have been based in part on statements such as:

> it is stated by Adam Smith, and it must be allowed to be stated justly, that the produce which is annually saved is as regularly consumed as that which is annually spent, but that it is consumed by a different set of people. (Malthus, 1989b, vol. 1, p. 31)

Malthus appears here to agree with Adam Smith that savings will always find an outlet in investment, and that there will never be a surplus of savings over investment. However, that interpretation is doubtful, given that ‘and it must be allowed to be stated justly’ was omitted from the second edition of the *Principles* (see Malthus 1989b, vol. 2, pp. 28, 300–1). Ricardo in his *Notes on Malthus* had said that a saving-equals-investment interpretation is inconsistent with the views expressed elsewhere by Malthus on saving. It would be reasonable to conclude that the omission was made by Malthus in response to Ricardo's note (Ricardo, 1951–73, vol. 2, p. 15, n. 4).

Another possible source for attributing a saving-equals-investment view to Malthus might be Malthus's statement ‘No political economist of the present day can by saving mean mere hoarding’ (1989b, vol. 1, p. 32). Some commentators have interpreted this to mean that, in Malthus's view, savings are always invested, never hoarded, and never intended to be hoarded. If correct, such an interpretation would also constitute a major difference between Malthus and Keynes.

But there is another, more plausible, interpretation. Malthus here was not saying that savings are never held as idle cash balances. He was not precluding the possibility that savings might remain uninvested and idle, not on purpose but because a satisfactory investment outlet cannot be found. This alternative interpretation negates a saving-equals-investment interpretation.

There are numerous instances in Malthus's writings that support this alternative interpretation. They clearly show that, in his view, savings will not always be invested, and that excessive savings are harmful. For example, Adam Smith had said that ‘every frugal man [appears to be] a publick benefactor’

and that the increase of wealth depends on a favourable balance of production over consumption (Smith, 1776, book II, ch. 3, para. 25; book IV, ch. 3, iii, para. c.15), but Malthus disagreed:

> That these propositions are true to a great extent is perfectly unquestionable … but it is quite obvious that they are not true to an indefinite extent. (1989b, vol. 1, 8)

To say that Malthus identified or equated saving and investment is to ignore his frequent use of expressions such as redundant capital, excessive capital, idle capital, spare capital, premature supply of capital, unemployed capital, vacant capital, capitalists at a loss where they can safely employ their capitals, capitals at a loss for employment, and so on. These expressions refer to funds that arise through savings and are intended for investment but for which an actual investment, at an acceptable degree of profit and risk, cannot be found. Malthus was thus recognizing the possible existence of an inequality between *ex ante* or intended investment and *ex post* or actual investment, because of the exhaustion of profitable investment outlets. This gap between savings intended for investment and savings actually invested could be described as unintended or residual hoarding – although Malthus did not use those terms – as distinct from the intended hoarding of a miser, or ‘mere hoarding’.

Malthus and Ricardo

The correspondence between Malthus and Ricardo provides a revealing insight into the minds and characters of two of the most important contributors to the development of political economy in England during its formative years in the early 19th century (see Ricardo, 1951–73, vols 6–9). They expressed their arguments forcefully but politely, although at times hints of frustration and exasperation began to appear, as they struggled to comprehend and to counter the other's point of view, especially when in an era without carbon copies and photocopies they seemed to forget what they had previously written. And despite their doctrinal and methodological differences, they remained close friends, with frequent visits to one another's homes. Ricardo's last letter to Malthus concluded with the statement: ‘I should not like you more than I do if you agreed in opinion with me’ (Ricardo, 1951–73, vol. 9, p. 382); and Malthus, after the death of Ricardo, was reported to have said: ‘I never loved any body out of my own family so much’ (Empson, 1837, p. 489). Ricardo had offered to assist Malthus financially by investing money for him in a stockbroking venture; and at one stage Malthus might have been seriously considering a personal involvement in international trading in commodities and bullion, using statistics and advice provided by Ricardo (see Malthus, 2004, ch. 3). After Ricardo's death, Malthus defended him against critics who Malthus considered had gone too far in their criticisms; and, in lectures read to the Royal Society of Literature in 1825 and 1827, Malthus developed a theory of value which, while maintaining his previous emphasis on demand and supply as determinants of value, gave greater recognition to Ricardo's emphasis on the cost of production (Malthus, 1986, vol. 7, pp. 301–23).
Opinions differ on who was the greater economist – Malthus or Ricardo. Who made the more significant contributions to the development of economics? On the one hand there are those who see Malthus as muddle-headed, and Ricardo as the better logician. On the other hand, there are those who reject the claim that Ricardo was a better logician, and who argue that Malthus's understanding of the multi-causal complexity of the real world was of far greater value to the progress of economics than Ricardo's

abstract theorizing. The most famous member of the latter group, J.M. Keynes, said that the world would be ‘a much wiser and richer place’ if ‘Malthus, instead of Ricardo, had been the parent stem from which nineteenth-century economics proceeded’; and that ‘the almost total obliteration of Malthus's line of approach and the complete domination of Ricardo's for a period of a hundred years has been a disaster to the progress of economics’ (Keynes, 1933, pp. 120, 117).

See Also

- Corn Laws, free trade and protectionism
- history of economic thought
- Keynes, John Maynard
- laissez-faire, economists and
- Malthusian economy
- Smith, Adam

Selected works

1986. *The Works of Thomas Robert Malthus*, 8 vols, ed. E. Wrigley and D. Souden. London: William Pickering.

1989a. *An Essay on the Principle of Population*, variorum edn, 2 vols, ed. P. James. Cambridge: Cambridge University Press for the Royal Economic Society.

1989b. *Principles of Political Economy*, variorum edn, 2 vols, ed. J. Pullen. Cambridge: Cambridge University Press for the Royal Economic Society.

1997. *T.R. Malthus: The Unpublished Papers in the Collection of Kanto Gakuen University*, vol. 1, ed. J. Pullen and T. Hughes Parry. Cambridge: Cambridge University Press for the Royal Economic Society.

2004. *T.R. Malthus: The Unpublished Papers in the Collection of Kanto Gakuen University*, vol. 2, ed. J. Pullen and T. Hughes Parry. Cambridge: Cambridge University Press for the Royal Economic Society.

Bibliography

Bonar, J. 1924. *Malthus and His Work*, 2nd edn. London: George Allen and Unwin.

Bonar, J. 1925–6. Thomas Robert Malthus. In *Palgrave's Dictionary of Political Economy*, 2nd edn, ed. H. Higgs. London: Macmillan.

Dupâquier, J., Fauve-Chamoux, A. and Grebenik, E. 1903. *Malthus Past and Present*. London:

Academic Press.

Empson, W. 1837. Life, writings and character of Mr Malthus. *Edinburgh Review* 64(January), 496–506.

Ghosh, R. 1963. ‘Malthus on emigration and colonization. Letters to Wilmot-Horton’. *Economica* 30, 45–62.

Hollander, S. 1992. Malthus's abandonment of agricultural protectionism: a discovery in the history of economic thought. *American Economic Review* 82, 650–9.

Hollander, S. 1995. More on Malthus and agricultural protection. *History of Political Economy* 27, 531–7.

Hollander, S. 1997. *The Economics of Thomas Robert Malthus*. Toronto: University of Toronto Press.

James, P. 1966. *The Travel Diaries of Thomas Robert Malthus*. Cambridge: Cambridge University Press.

James, P. 1979. *Population Malthus: His Life and Times*. London: Routledge and Kegan Paul.

Keynes, J.M. 1933. Robert Malthus. In *Essays in Biography*. London: Macmillan, 1961.

Keynes, J.M. 1936. *The General Theory of Employment,Interest and Money*. London: Macmillan.

McCleary, G. 1953. *The Malthusian Population Theory*. London: Faber and Faber.

Meek, R., ed. 1953. *Marx and Engels on Malthus*. London: Lawrence and Wishart.

Otter, W. 1836. Memoir of Robert Malthus. In T.R. Malthus, *Principles of Political Economy*, 2nd edn. New York: Augustus M. Kelley, 1951.

Pullen, J. 1987. Malthus, Thomas Robert. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Pullen, J. 1995. Malthus on agricultural protection: an alternative view. *History of Political Economy* 27, 517–29.

Pullen, J. 2004. Malthus, Thomas Robert. In *Oxford Dictionary of National Biography*. Oxford: Oxford University Press.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*, 11 vols, ed. P. Sraffa with the collaboration of M. Dobb. Cambridge: Cambridge University Press for the Royal Economic Society.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R. Campbell and A. Skinner; textual editor W. Todd. Oxford: Clarendon Press, 1976.

Stephen, L. 1893. Malthus, Thomas Robert. In *Dictionary of National Biography*, vol. 12. London: Oxford University Press.

Waterman, A. 1991. *Revolution, Economics and Religion: Christian Political Economy, 1798–1833*. Cambridge: Cambridge University Press.

Winch, D. 1983. Higher maxims: happiness versus wealth in Malthus and Ricardo. In *That Noble Science of Politics*, ed. S. Collini, D. Winch and J. Burrow. Cambridge: Cambridge University Press.

Winch, D. 1987. *Malthus*. Oxford: Oxford University Press.

Winch, D. 1996. *Riches and Poverty: An Intellectual History of Political Economy in Britain, 1750–1834*. Part 3, ‘Robert Malthus as Political Moralist’. Cambridge: Cambridge University Press.

How to cite this article

Pullen, J. M. "Malthus, Thomas Robert (1766–1834)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M100018> doi:10.1057/9780230226203.1013

The New Palgrave Dictionary of Economics Online

Malthusian economy

Gregory Clark
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Malthusian economy was the economic system that characterized almost all economies before the industrial revolution. In this regime fertility and mortality rates at different material income levels determined the average real income level and life expectancy at birth. Thus before 1800 the improvement of production technologies resulted only in population growth, and not in any gains in material living conditions beyond those that were found in the original hunter gatherer societies.

Keywords

anthropometric history; birth and death rates; fertility; historical demography; industrial revolution; Malthus, T.; Malthusian economy; population growth; technology schedule

Article

The Malthusian economy is the economic system which prevails whenever a society's production technology advances so slowly that population growth forces incomes down to the subsistence level. In such an economy material welfare is independent of natural resources, technology and capital accumulation, but instead depends solely on the factors governing fertility and mortality. The resulting subsistence income can, however, vary widely across societies. Some Malthusian economies were rich by the standards of most countries in modern Africa, for example.

Almost all societies until 1800 were Malthusian, from the original foragers of the African savannah 50,000 years ago down through settled agrarian societies of considerable sophistication such as England, France, China and Japan in 1800. The operation of all human societies through history up until the Industrial Revolution can thus seemingly be described by this one simple economic system. An implication of this is that there was most likely no gain in material welfare between the evolution of anatomically modern humans and the onset of the Industrial Revolution.

Government actions, in so far as they change fertility or mortality, can influence material welfare in the Malthusian economy, but in a contradictory fashion. Good governments that reduced mortality through order and security made people poorer. Bad governments that increased mortality through warfare and banditry made them wealthier.

The economic logic of these societies was first, though only partially, appreciated by Thomas Malthus in his famous *Essay on a Principle of Population* of 1798. Malthus's insights were elaborated by writers such as David Ricardo and James Stuart Mill into the system called classical political economy in the early 19th century. Ironically, this intellectual development happened just as for the first time the rate of technological advance was becoming sufficiently rapid to bring the Malthusian era to a close.

Insight into the Malthusian economy starts from the insight that the biological capacity of women to produce offspring is much greater than the number of births required to reproduce the population. If fertility is unrestricted women can have 12 or more children. Social institutions regulating marriage and contraceptive practices will determine the actual numbers of births per women. In modern societies these institutions and practices vary greatly, so the number of births per women varies greatly. Completed fertility now ranges across the world from a low of 1.15 in Spain to a high of 8.0 in Niger. Only where women happen on average to have two children who survive to adulthood will population be stable. Even small deviations from this number will cause rapid increases or decreases in population. Thus modern populations are not stable.

Despite this potential for explosive population growth, pre-industrial populations were remarkably stable over the long run. The average annual growth rate of world population from 10,000 years bc to ad 1800 was 0.05 per cent. The typical woman before 1800 thus had 2.02 children who survived to reproductive age. As an extreme case the population of Egypt,

for example, is estimated at between four million and five million at 1000 years bc. The population in Greek and Roman Egypt a millennium later is estimated at this same four million to five million. The first modern census in 1848 suggests a population of 4.5 million. Thus over nearly 3,000 years the Egyptian population growth rate was to a close approximation zero, and women on average had two surviving children. Yet it is estimated that in Roman Egypt the average woman gave birth to six children. Some mechanism kept fertility and mortality in balance in these pre-industrial economies.

The Malthusian equilibrium

The simple Malthusian model of how pre-industrial society functioned supplies an economic mechanism to explain its population stability. In its simplest version there are just three assumptions:

1. 1. The *birth rate*, the number of births per year per thousand people, is a socially determined constant, independent of material living standards. Birth rates will vary across societies, but in this simplest model they are assumed to be independent in any given society of material living conditions.
2. 2. The *death rate*, the number of deaths per year per thousand persons, declines as material living standards increase. Again, the death rate will differ across societies depending on climate and lifestyles, but it assumed that in all societies it will decline as material living conditions improve.
3. 3. *Material living standards* decline as population increases.

Figure 1 shows the first two assumptions of the simple Malthusian model in graphical form in the upper panel. The birth and death rates are plotted on the vertical axis, material income per capita on the horizontal axis. The first two assumptions of the simple Malthusian model imply that there is only one level of real incomes at which the birth rate equals the death rate, denoted as y^*. And this constitutes a stable equilibrium. Thus y^* is called the ‘subsistence income’ of the society: it is the income at which the population barely subsists, in the sense of just reproducing itself. This subsistence income is determined without any reference to the production technology. It depends only on the factors which determine birth and death rates. Once we know these factors we can determine the subsistence income.

Figure 1

Long-run equilibrium in the Malthusian economy

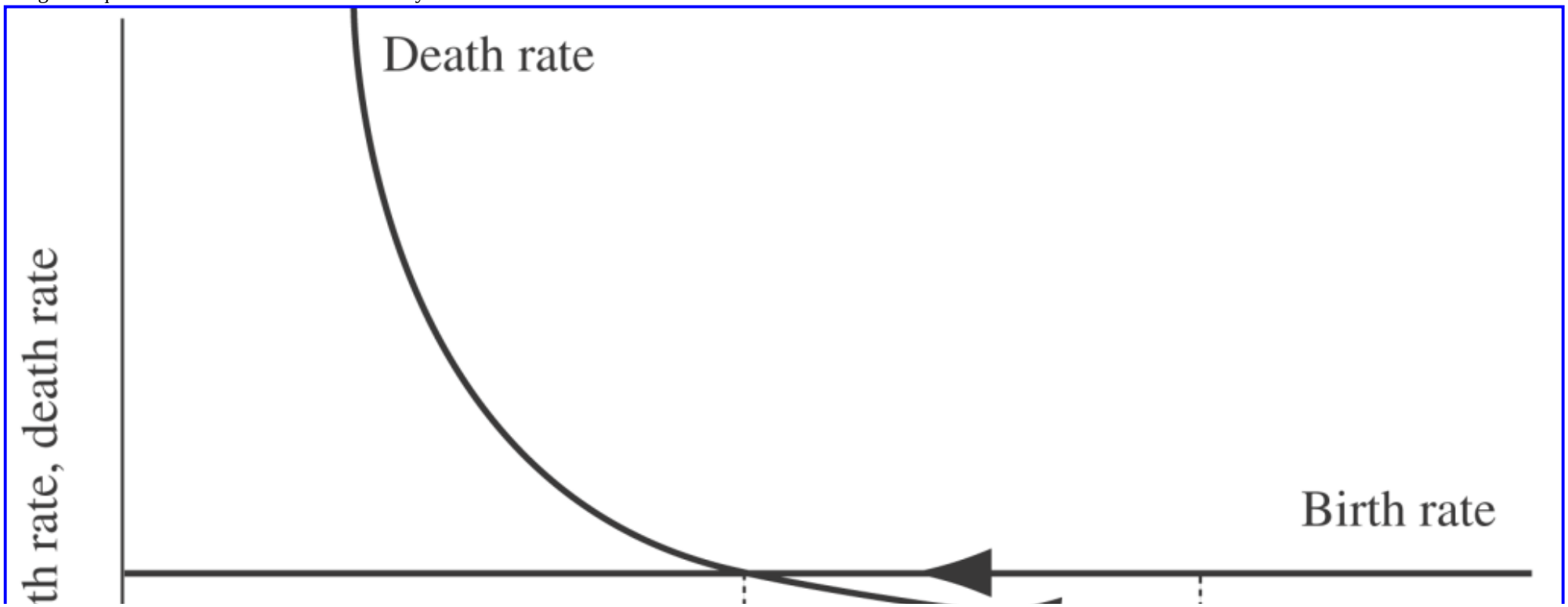

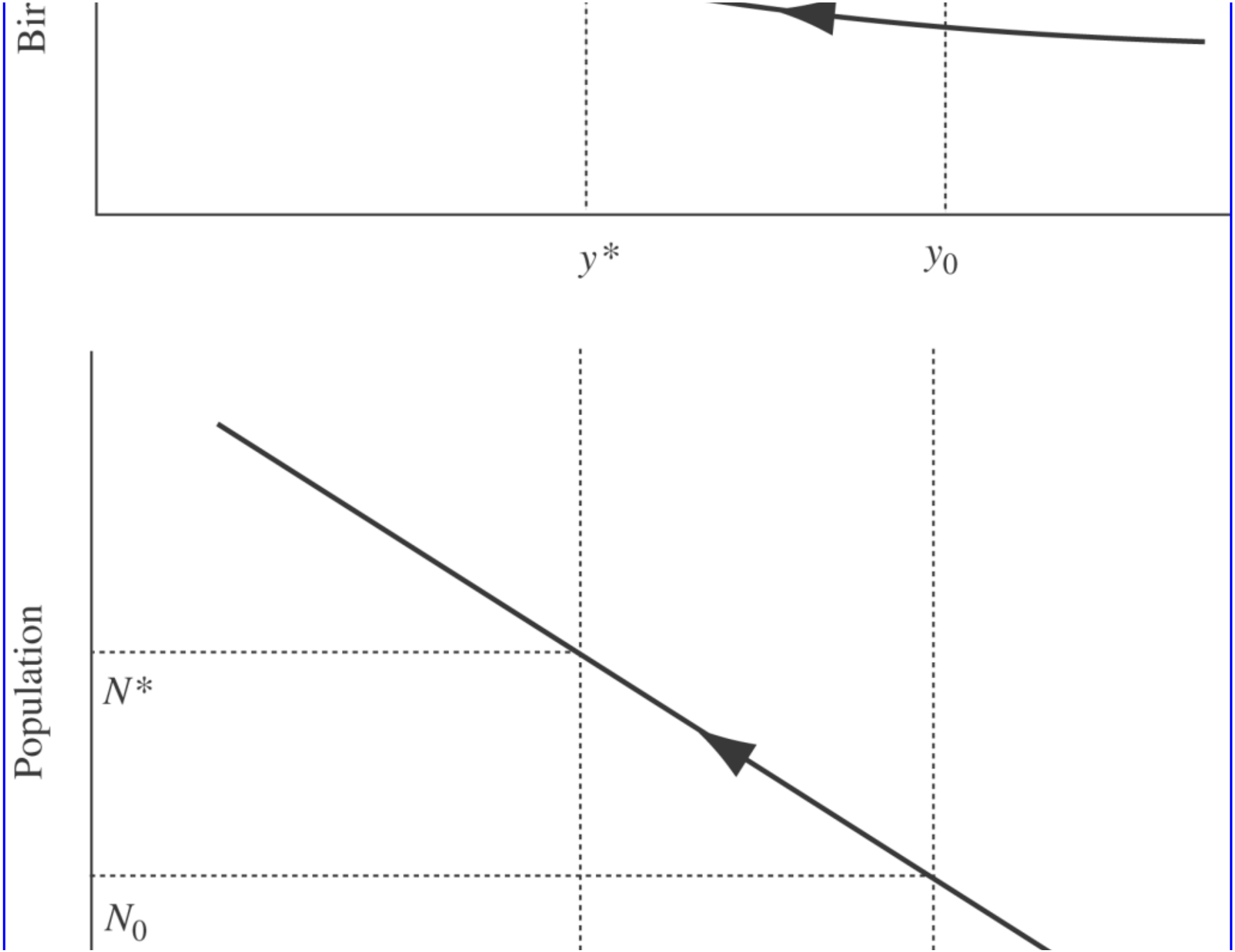
Bir
y*
y0
Population
N*
N0

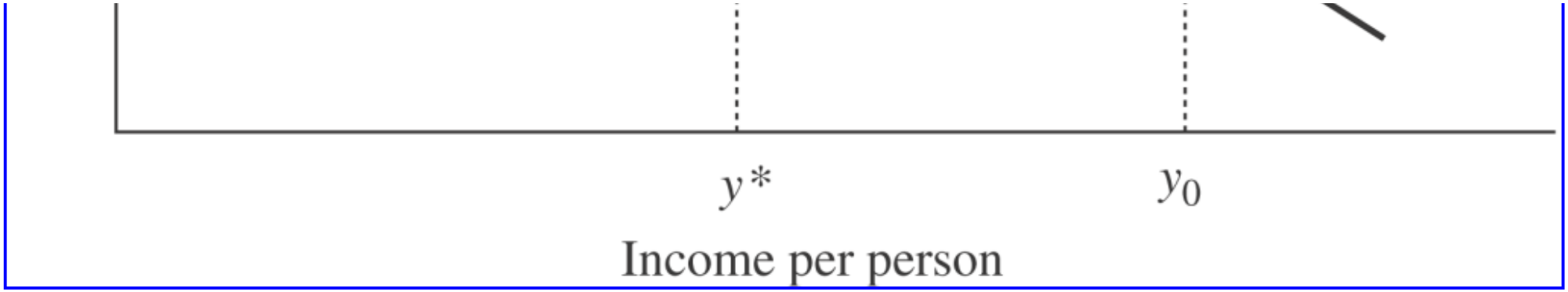

Another aspect of human welfare is life expectancy at birth, that is, the average number of years a person will live. In the Malthusian era life expectancy at birth also depended only on the factors determining birth and death rates. This is because with a *stable* population, where annual births have equalled deaths for a long time, life expectancy at birth is the inverse of the crude birth rate. With fertility not restricted in any way crude birth rates would be 50–60 per thousand in pre-industrial populations (based on modern experience). This would imply a life expectancy at birth of 20 years or less.

The term ‘subsistence income’ can lead to the confused notion that in the Malthusian economy people were always living on the edge of starvation. In fact, in almost all Malthusian economies the subsistence income was considerably above the income required for the physiological minimum daily diet. All pre-industrial societies for which we have good demographic records limited fertility below the biological maximum. Differences in the location of the mortality and fertility schedules generated subsistence incomes at very different levels. Thus, both 1450 and 1650 were periods of population stability in England, and hence periods where by definition income was at subsistence. But the wage of unskilled agricultural labourers was equivalent to about 6 lb of wheat flour per day in 1650, compared with 18 lb in 1450. Even the 1650 unskilled wage was well above the physiological minimum. A diet of about 1.33 lb of wheat flour per day would keep a labourer alive and fit for work (it would supply about 2,400 calories per day). Thus, pre-industrial societies, while they were subsistence societies, were not starvation regimes. England in 1450, indeed, was wealthy even by the standards of many modern societies such as those in sub-Saharan Africa.

The bottom panel of Figure 1 illustrates the third assumption. The panel has on the vertical axis the population, N, and on the horizontal axis the material income. As population increased material income per person by assumption declined. The justification for this assumption is the law of diminishing returns. Since one important factor of production, land, is always in fixed supply in pre-industrial economies, the law of diminishing returns implies that average output per worker fell as the labour supply increased as long as the technology remained static. Thus the average amount of material consumption available per person fell with population.

Figure 1 also shows how an equilibrium birth rate, death rate, population level and real income were arrived at in the long run in a pre-industrial economy. Suppose we start at an arbitrary initial population N_0 in the diagram, greater than N^*. This generates an income y_0, above the subsistence income. At this income the birth rate exceeds the death rate, so population grows until income falls to y^* and population equals N^*.

Changes in the birth rate, death rate and ‘technology’ schedules

Suppose that the birth rate schedule in Figure 1 was higher. Then at the equilibrium, real income would be lower, and the population greater. Thus any increase in birth rates in the Malthusian world drove down real incomes and reduced life expectancy. Conversely, anything which limited birth rates drove up real incomes and increased life expectancy. Thus in the pre-industrial era birth rates were a crucial determinant of material living conditions.

If the death rate schedule was higher, so that at each income there was a higher death rate, then the equilibrium real income would be higher. But if the birth rate was not responsive to income then a greater death rate increased real incomes but in the long run had no effect on the annual death rate or on life expectancy at birth. Thus in this simplest Malthusian model higher mortality risks at a given income were unambiguously a good thing, at least in the long run.

The simple Malthusian world thus exhibits an almost counter-intuitive logic. Anything that raised the death rate schedule, the death rate at a given income, such as war, disorder, disease or poor sanitary practices, increased material living standards without changing life expectancy at birth. Anything that reduced the death rate schedule, such as advances in medical technology, or better public sanitation, or public provision for harvest failures, or peace, reduced material living standards without any gain in life expectancy at birth.

While the real income was determined from the birth and death schedules, the population size depended on the schedule linking population and real incomes. Above I labelled this the ‘technology’ schedule, because in general the major cause of changes in this schedule has been technological advances. But other things could shift this schedule – a larger capital stock, improvements in the terms of trade, climate improvements, and a more productive organization of the economy. A shift upwards in this schedule, in the short run, since population can change only slowly, would have increased real incomes. But the increased real incomes reduced the death rate, so that births exceeded deaths and population began

growing. The growth of population ended only when the income returned to the subsistence level, y^*. At this new equilibrium the only effect of the technological change was to increase the population supported. There was no lasting change in the living standards of the average person.

More complicated Malthusian models

An issue that has exercised historical demographers is whether the birth rate in pre-industrial societies was ‘self-regulating’. What they mean by this is shown in Figure 2, which shows the birth and death schedules of a simplified Malthusian model, as well as a modified birth schedule, which slopes upwards with material incomes. In the modified Malthusian model it is assumed that in good times people married earlier and more people married, so that fertility increased, whereas in bad times fewer married, and they married later, so that fertility declined.

Figure 2

A Malthusian model where births increase with income

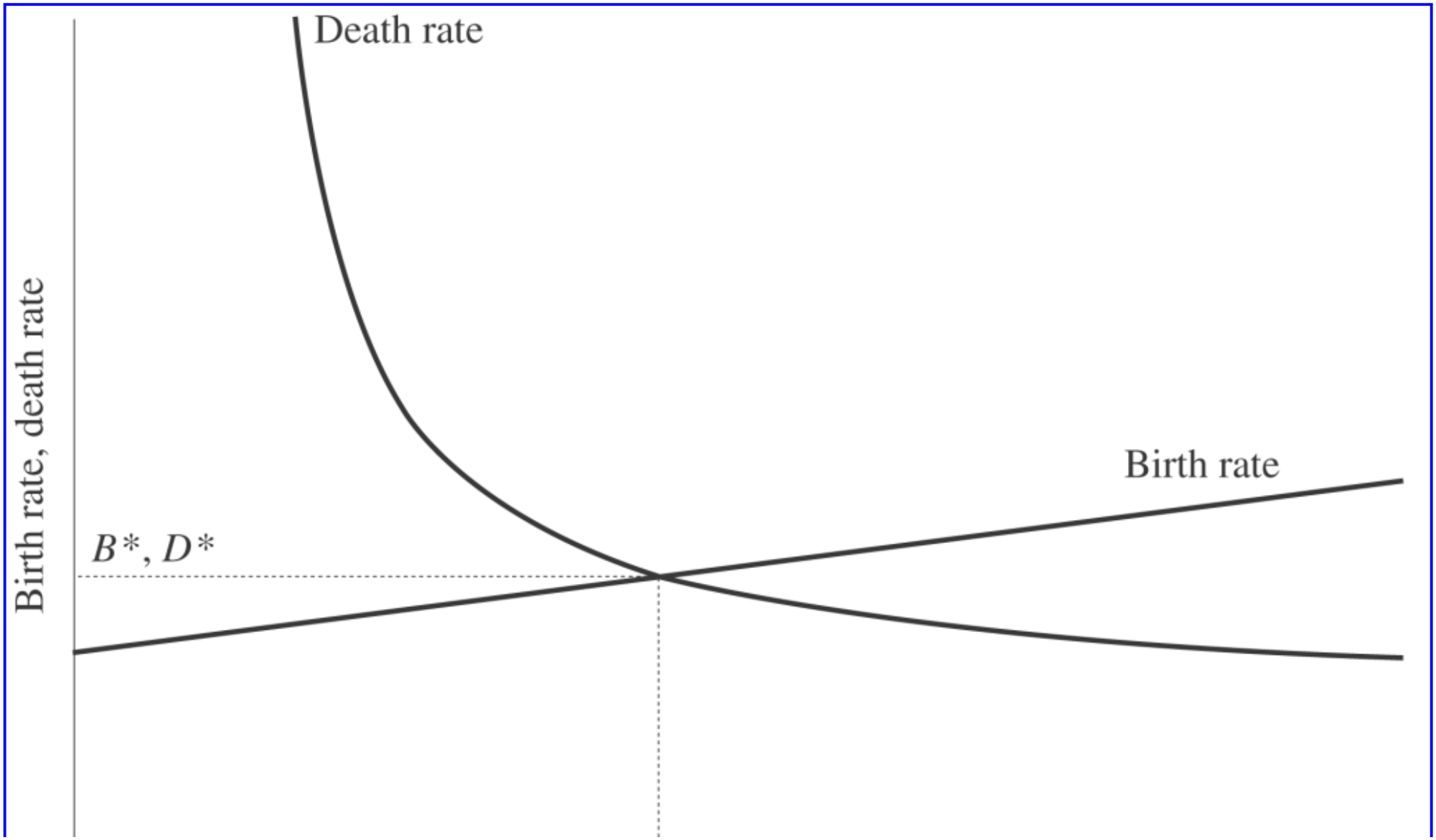

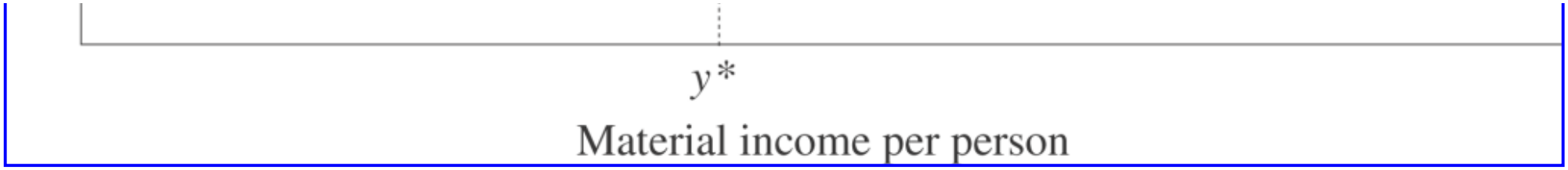

It should be clear that a positive association of fertility and income does not change the basic equilibrium of the model. The only difference is that increases in the death rate at any given material income are now not so unambiguously good, since they will be associated with higher fertility and mortality rates and hence lower incomes. The evidence for societies such as pre-industrial England, however, shows no response of fertility to income (Wrigley et al., 1997). Thus the simple model may well describe pre-industrial societies well.

What causes many more potential complications is a birth schedule that declines with material incomes. Suppose that as real incomes go up one of the responses of people is to desire fewer children. With a birth rate that declines with real incomes the model could have multiple crossings between the birth rate and death rate schedules. At those places where the birth rate schedule was declining more steeply than the death rate schedule the equilibrium would be unstable. Figure 3 gives a declining birth rate schedule that twice intersects the death rate schedule. The intersection at the lower real income, y_0, is a stable equilibrium. But the second higher income equilibrium at y_1 is unstable. If real incomes drop below this level by any amount then population starts to grow, leading real incomes all the way down to the stable equilibrium at y_0. Conversely if they increase at all above y_1 then deaths will exceed births and real incomes continue to grow indefinitely. The population will fall eventually to zero.

Figure 3

A Malthusian model where births decline with income

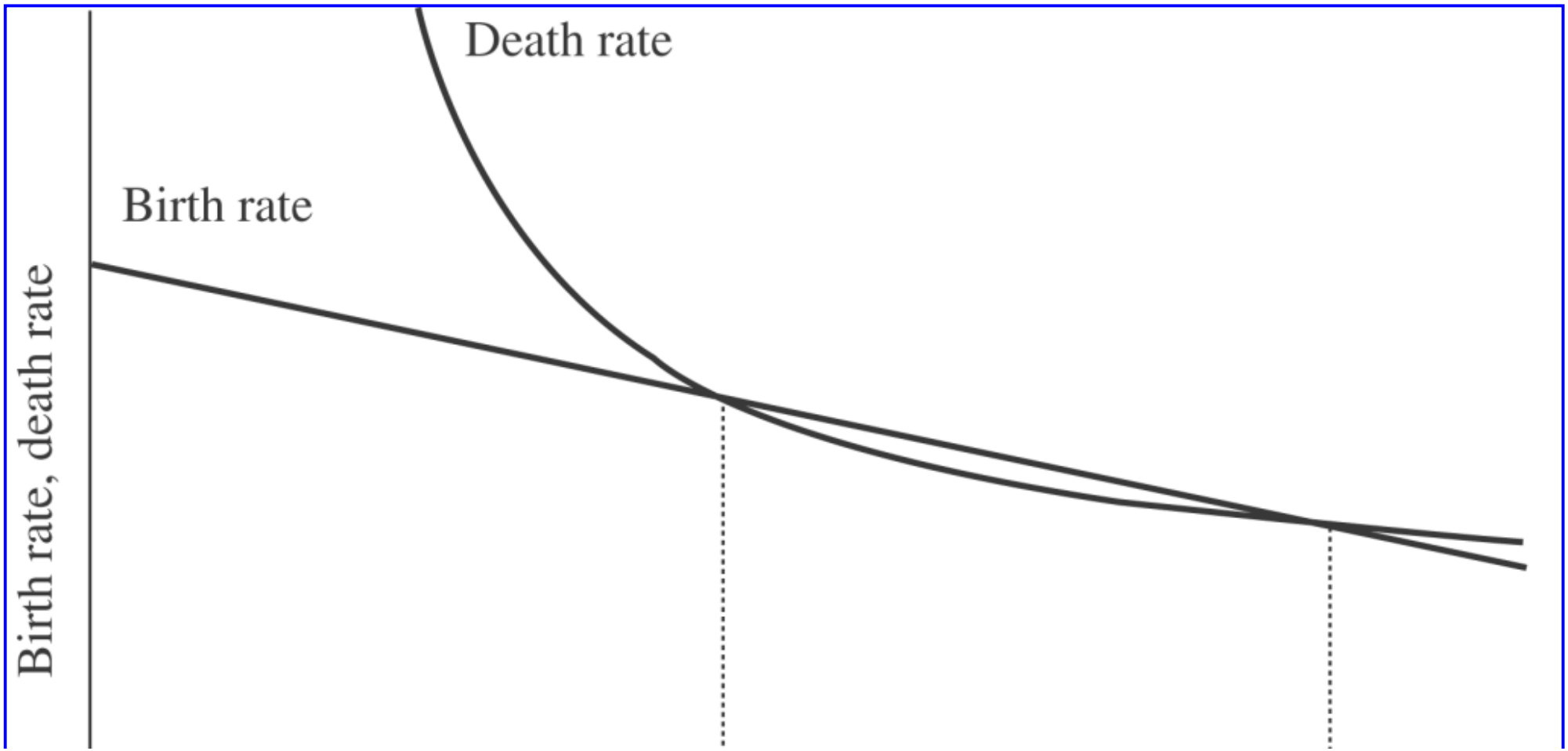

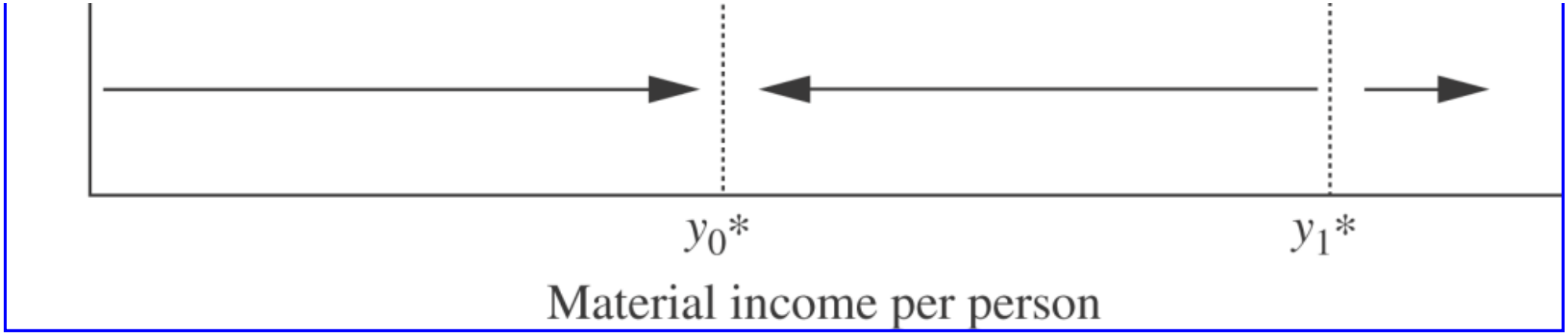

In this case there is a 'Malthusian trap' in the pre-industrial economy. A society can be stuck in the subsistence income equilibrium unless some jolt such as acquiring extra land, experiencing a much higher death rate, or experiencing faster technological progress pushes up wages enough so that fertility falls permanently. The shock of the Black Death, however, which tripled real incomes for the poorest workers in England by 1450, did not lead to any permanent movement towards lower fertility and the escape from the Malthusian trap. Again, the evidence for pre-industrial demography suggests no declines in fertility with higher incomes.

The empirical implications of the Malthusian model

The most interesting empirical implication of the Malthusian model is that material living conditions for people, including life expectancy at birth, may well have been unchanged between the dawn of humanity and ad 1800. Were the people in sophisticated societies such as England, France, the Netherlands, Japan and China in 1800 really no better off than the original hunter-gatherers? This seems particularly counter-intuitive for England, reckoned to be the richest country the world by 1800.

By then England was a society that would not seem that different from our own. The middle and upper classes in London breakfasted at coffee shops as they read the daily newspapers. They dwelled in homes of brick and glass with water supplied by lead pipes, lighted at night by oil harvested from sperm whales taken thousands of miles away in the oceans. There was extensive trade for luxury products from the tropics – cottons, silks, spices. How could the material condition of humanity not be better then than in the savage past when our ancestors faced the elements naked, and sought shelter at night in depressions in the ground or in crude lean-tos?

But even in England in 1800 the living conditions of the mass of the population were still primitive. The largest employment was still agriculture, where the average day wage in 1800–9 was the equivalent of 5.7 lb of wheat flour. This was enough to keep a family fed only if most of the income was spent on the cheapest forms of food such as bread. Farm labourers lived in simple structures little better than those of the medieval period. They slept when it was dark because they could not afford lighting. They could afford one new set of clothing per year. English farm labourers six hundred years before, in 1200–9, received a wage which was the equivalent of 12 lb of flour, significantly more than in 1800. And at the best time for pre-industrial workers in England, circa 1450, when the population losses of the plagues which ravaged Europe from 1348 on were their greatest, the real wage was much higher, equivalent to 18 lb of flour. In the years 1200–1800 in England there is no sign of long-run gains in real wages for the mass of workers. We know also the real day wage of farm workers in Roman Egypt circa ad 250 was the equivalent of 5 lb of flour, not much less than England in 1800.

How did English material living conditions around 1800 compare with hunter-gatherer societies such as those that constituted society through the great bulk of human history? We can obtain insight on this in two ways. The first is by comparing living conditions in England in 1800 with those of the few surviving hunter-gatherer groups. Since the diets were very different here we have to use measures such as the number of calories consumed per person per day. In 1787–96 for the families of English farm workers this was a meagre 1,508 calories. For a group of eight hunter-gatherer societies studied in the 1960s to 1980s the average consumption was 2,272 calories, much better than for England. On this measure the English on the eve of the Industrial Revolution seem to have lived less well than the average hunter-gatherer. Another aspect of the quality of life is life expectancy at birth. One measure of this is the fraction of infants that survived the first year of life. In England as a whole this is estimated at 83 per cent in the second half of the 18th century. For modern hunter-gatherer societies survival rates were a little lower at 79 per cent. But this is still not that much lower than for the richest society in the world in 1800. And survival rates for infants in London, the richest part of England, were only 70 per cent because of the health hazards of city life.

A second measure is the average stature of people. Height is a good index of material living conditions, since it depends on both food consumption and the amount of sickness people experience as they grow. Average heights for adult males in England circa 1800 were 67 inches or less. This was very good by the standards of societies just before industrialization. Average male heights in Japan in the late 19th century were 61 inches and in India in the early 19th century 64 inches. Yet these heights in England are little if any better than those recorded from skeletons of hunter-gatherers in the Mesolithic (10,000–5000 bc) and Neolithic (5000–1000 bc) in Europe. Average male height from these skeletons is estimated at 66 inches. So overall, if we look at agrarian societies across the world in 1800 ad, the stature evidence suggests a decline in living conditions from hunter-gatherer society.

Thus, the evidence is that for the mass of humanity on the eve of the Industrial Revolution living conditions were no better and probably worse than in the hunter-gatherer past.

See Also

- anthropometric history
- historical demography
- industrial revolution
- Malthus, Thomas Robert

Bibliography

Bennike, P. 1985. *Paleopathology of Danish Skeletons*. Copenhagen: Akademisk Forlag.

Clark, G. 2007. *A Farewell to Alms: A Brief Economic History of the World*. Princeton, NJ: Princeton University Press.

Koepke, N. and Baten, J. 2005. The biological standard of living in Europe during the last two millennia. *European Review of Economic History* 9, 61–95.

Malthus, T. 1798. *An Essay on a Principle of Population*. 6th edn. London: John Murray, 1826.

Miller, M. and Upton, C. 1986. *Macroeconomics: A Neoclassical Introduction*. Chicago: University of Chicago Press.

Steckel, R. and Rose, J., eds. 2001. *The Backbone of History: Health and Nutrition in the Western Hemisphere*. Cambridge: Cambridge University Press.

Wrigley, E., Davies, R., Oeppen, J. and Schofield, R. 1997. *English Population History from Family Reconstruction: 1580–1837*. Cambridge and New York: Cambridge University Press.

How to cite this article

Clark, Gregory. "Malthusian economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000199> doi:10.1057/9780230226203.1014

The New Palgrave Dictionary of Economics Online

Malynes, Gerard de (fl. 1586 – 1623)

Lawrence H. Officer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

exchange controls; Malynes, G. de; money supply; price level; specie-flow mechanism; terms of trade; usury

Article

A merchant of English parentage, born in Antwerp at an unknown date, Malynes was a commissioner of trade in the Low Countries about 1586. He came to London and was frequently consulted on commercial questions by the Privy Council in the reigns of Elizabeth I and James I. He became an assay master at the mint and obtained a patent to supply farthings; he was imprisoned for a time, complaining later that he had been ruined by being paid in his own coins. He also served as a spy for England. Called on by the standing commission on trade for evidence on the state of the coinage, he published a series of pamphlets on money and prices. A mercantilist and a bullionist, he was heavily influenced by Scholastic literature.

Malynes viewed individual commodity prices as determined by demand and supply. However, he was more interested in the price level, governed by the quantity of money (Malynes, 1601b; 1603). An expanding money supply, associated with a rising price level, decreased interest rates and stimulated the economy (1601b; 1622a). Therefore Malynes viewed usury as at best a necessary evil (see Muchmore, 1969, p. 346) and, above all, opposed any export of specie whatsoever.

Rejecting the balance of trade theory, Malynes charged that ‘bankers’ (exchange dealers) controlled the exchange rate (1601b; 1622a; 1622b; 1623). By their incorporation of usury in the price of a bill of exchange and through speculation, they conspired to undervalue sterling, leading to a deterioration in England's terms of trade (‘overbalancing’) and a specie outflow (1601b; 1622a; 1623). But overvalued sterling would not lead to a specie inflow, because the export proceeds would be spent on luxury imports (1601b). Yet Malynes (1601b) has a theory of price level changes in response to exchange rates differing from mint parity and money flowing between countries – a price specie-flow mechanism, marred only by the assumption of inelastic demand. His solution to the twin problems of specie outflow and terms of trade deterioration is comprehensive exchange control with enforced exchange dealings at rates fixed at

mint parities (Malynes, 1601b; 1622a; 1622b; Muchmore, 1969, pp. 347–8).

Selected works

1601a. *Saint George for England, allegorically described*. London: Richard Field for William Tymme.

1601b. *A Treatise of the Canker of England's Commonwealth*. London: Richard Field for William Iohnes. Reprinted in part in *Tudor Economic Documents*, vol. 3, ed. R.H. Tawney and E. Power. London: Longmans, Green, 1924.

1603. *England's View, in the Unmasking of Two Paradoxes*. London: Richard Field.

1622a. *Consuetudo, vel lex mercatoria, or the Ancient Law-merchant*. London: Adam Islip.

1622b. *The Maintenance of Free Trade*. London: I. Legatt for W. Sheffard.

1623. *The Centre of the Circle of Commerce*. London: William Iones.

Bibliography

Johnson, E. 1933. Gerard de Malynes and the theory of the foreign exchanges. *American Economic Review* 22, 441–55.

Muchmore, L. 1969. Gerrard de Malynes and mercantile economics. *History of Political Economy* 1, 336–58.

Officer, L.H. 1982. The purchasing-power-parity theory of Gerrard de Malynes. *History of Political Economy* 14, 251–5.

Roover, R. de 1974. *Business, Banking and Economic Thought*. Chicago: University of Chicago Press.

Schumpeter, J. 1954. *History of Economic Analysis*. Oxford: Oxford University Press.

Spiegel, H. 1971. *The Growth of Economic Thought*. Englewood Cliffs, NJ: Prentice-Hall.

Wu, C.-Y. 1939. *An Outline of International Price Theories*. London: George Routledge & Sons.

How to cite this article

Officer, Lawrence H. "Malynes, Gerard de (fl. 1586–1623)." The New Palgrave Dictionary of

Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000020> doi:10.1057/9780230226203.1015

The New Palgrave Dictionary of Economics Online

Manchester School

William D. Grampp
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Anti-Corn Law League; Bright, J.; Cobden, R.; Corn Laws; free trade; Manchester School; Navigation Laws; philosophic radicalism

Article

The Manchester School was the name given by Disraeli after the event to the leaders of the successful agitation conducted between 1838 and 1846 to abolish the Corn Laws. It is wrongly associated with the arch-advocacy of laissez-faire. The people of the School were not in fact united by any single idea, other than believing in the complete and immediate repeal of the tariff on grain.

Within the School there were five discernible groups in the sense of there being five different reasons why people wanted repeal or purposes that directed them. Some were compatible with others, and one group could agree with another over what was important but differ over how important it was. The arguments that each group made do not, when taken together, constitute a cogent or even coherent whole but taken separately could be both, and are always interesting. Moreover, the campaign for repeal is itself an instructive event in the history of economic policy.

(1) One group was the mill-owners of Lancashire who provided most of the money for the campaign and formed the National Anti-Corn Law League to conduct it. Some believed that repeal, by reducing the price of bread, would reduce money wages, hence the cost of production in their mills. The belief comes from the Ricardian principle that real wages are constant in the long run. It could have made the businessmen believe the export of grain should be protected, since that too could reduce its price, hence have placed them in the interesting but not unusual position of half-believing in the free market.

They in fact did not support protection because a greater reason for their wanting repeal was to increase the export of manufactured goods. The economic argument most often made was that importing more grain would provide foreigners with more income to spend on British exports, with the result that income and employment would increase at home. The mill-owners were repeatedly accused of simply wanting to cut wages. Cobden privately warned them to stay out of the repeal campaign if they could not come in with clean hands. Publicly he offered to support a Factory Bill of Lord Ashley – the ‘universal

syllabub of philanthropic twaddle', in Carlyle's description – if Ashley would pay his farm hands what the workers in Cobden's factory were paid. The offer was declined.

Another economic augument for repeal was that it would retard the growth of manufacturing abroad and so keep Britain in its leading industrial position. Why the owner of a small mill would profit by his country's having more mills than any other was not made clear (although he might by way of an externality of some sort). The argument is nevertheless noteworthy. It was revived after 1945, when the undeveloped countries hastened to industrialize in the belief that doing so was a necessary condition of their progress. The argument is also part of the curious notion, entertained by historians, that Britain's free trade was an instrument of its imperialism. They reason that Britain, by keeping others in a non-industrial condition, could dominate them, exploit them, and/or make them dependent on it. Why one country would choose to be mistreated by another when it could choose a trading partner that did not mistreat it, as in a system of free trade it could do, is not explained. Or is there an explanation of why a dollar's worth of manufactured goods adds more to total welfare than a dollar's worth of goods that are not manufactured?

(2) Among the businessmen working for repeal were those who believed it would make life better for the lower classes. They have been called the humanitarian employers. They did more for their workers than the market or the law required, providing schools for the children, reading rooms and meeting places for the men and women, helping them to form friendly societies, cooperatives and cultural groups. Some employed a 'salaried visitor' (social worker) to call at the homes. These business people also undertook to improve the communities where they were established. One such effort was the Manchester Statistical Society which collected information on living conditions and used it to improve them. The Greg family stood out in this group.

(3) The radical businessmen, working on a larger scale than the humanitarians, aspired to improve the nation and the world. In economic affairs their great end was free trade and after the repeal of the Corn Laws they had a part in the abolition of the Navigation Laws. In politics they looked toward democratic government and worked to extend the franchise until all adult males had the right to vote. The radicals believed free trade would first increase the influence of the business classes, increase their members in Parliament, then (in a way not fully explained) increase the power of the working classes.

John Bright was the leader of this group, which itself was the Manchester version of the middle-class radicalism of the time. It had a finger or a hand or more in most reform movements, great and small, from the abolition of slavery and removal of religious disabilities to the penny post and repeal of the taxes on knowledge. The radicals were disrespectful of authority, indifferent to custom, unmindful of the ridiculous figure they often cut, and they were meddlesome, tiresome, persistent and effective. Like Pancks, what they did, they did, they did indeed, and when they finished there were noble institutions in ruins.

(4) The Philosophic or London Radicals had a different place from that of the radical businessmen, grounding their reform on a considered application of Bentham's utilitarianism and conducting themselves in the mannerly, measured way that made them heard and respected but unheeded and ineffectual. They did not care for the rough and ready way of Manchester and had to be reminded of where they were before it took on the repeal of the Corn Laws. Before them, Charles Villiers, a leading Benthamite, had each year moved in the House that it constitute itself a committee of the whole to consider the repeal of the Corn Laws, and each year the motion was defeated. The leadership of the free trade bloc passed to Cobden when he became a Member of Parliament, an instance, his friends said, of

talent giving way to genius. Francis Place, who was on the edge of the London Radicals, put things plainly and said that when the Manchester people wanted something done they did it.

(5) Cobden represented the pacifists of the School. They believed that trading nations had a material interest in peace, an idea Ricardo had stated in his *Essay on Profits* in 1815, and that they were natural friends by virtue of meeting on the market, an idea Ricardo was too realistic to entertain. Oddly, the pacifists seem not to have noticed they could have drawn an argument from the *Wealth of Nations*. No pacifist himself, Smith said Britain should not engage in trade that would diminish its military power. The implication is that free trade makes nations unable to go to war as well as unwilling.

The pacifists, although not the largest group within the School, were even more influential than the radicals. Cobden wanted to graft the peace movement onto the repeal campaign although he would not permit the franchise to be so joined, as Bright wanted to do. After repeal, the franchise had more public support than the peace movement and grew until all adults had the vote. Nevertheless those who believe free trade is conducive to peace can and do point out that the 19th century was a time when trade was freer than ever and was the only century in recent history when there has not been a world war.

Cobden wanted free trade because it would bring peace, Bright because it would bring the franchise. Others in the School had each of them his own purpose. They made common cause for seven years until the Corn Laws were brought down, then returned to their separate ways.

The Manchester School was a coalition around a single issue. It was not a group of ideologues committed to laissez-faire, as historians have carelessly said, nor did it express the pure spirit of the middle class, as some contemporaries believed. It was not a rent-seeking force, as Public Choice economists are tempted to say, nor did it preach the principles of huckstering (Disraeli), nor were its leaders ‘bartering Jews’ (Engels), nor were they ‘the official representatives of the bourgeosie’ (Marx). If the Manchester School is to be described simply, it was a remarkably successful effort to remove a major obstacle in the way of the market.

See Also

- Bright, John
- Cobden, Richard

Bibliography

Grampp, W.D. 1960. *The Manchester School of Economics*. Stanford: Stanford University Press; London: Oxford University Press.

Hirst, F.W. (ed.) 1903. *Free Trade and other Fundamental Doctrines of the Manchester School*. London, New York: Harper & Bros.

McCord, N. 1958. *The Anti-Corn Law League 1838–1846*. London: Allen & Unwin.

Morley, J. 1881. *The Life of Richard Cobden*. London: Chapman & Hall.

Prentice, A. 1853. *History of the Anti-Corn Law League*. London: W. & F.G. Cash.

Proceedings of the Chamber of Commerce and Manufactures at Manchester. *1821–1865*. Mss. at the Manchester Central Library.

Students in the Honours School of History in the University of Manchester and Arthur Redford. 1934. *Manchester Merchants and Foreign Trade 1794–1858*. Manchester: Publications of the University of Manchester, Economic History Series.

Trevelyan, G.M. 1913. *The Life of John Bright*. London: Constable.

How to cite this article

Grampp, William D. "Manchester School." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000027> doi:10.1057/9780230226203.1016

The New Palgrave Dictionary of Economics Online

mandated employer provision of employee benefits

Jonathan Gruber

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mandated employer provision of social benefits is of rising importance in the United States. As highlighted by Summers (1989), the efficiency losses from such mandates may be much lower than those of taxation due to tax–benefit linkages. I review the theory underlying this observation and the empirical evidence which documents full shifting to wages (and therefore little efficiency cost) of mandated benefits. A host of important questions about mechanisms remains unanswered, however.

Keywords

adverse selection; anti-discrimination law; cost shifting; health insurance; labour supply; mandating employer provision of employee benefits; minimum wages; payroll taxation; unemployment insurance; workers' compensation insurance

Article

The provision of social benefits can be financed in a number of different ways: through broad income taxation, through taxation of payroll only, or through mandates on employers to provide those benefits for their employees. The last channel is one of sizable and growing importance in the United States, although less so in other nations that tend to rely more on tax-financed government provision. Yet, until the late 1980s, the impacts of mandates were not much studied. The implicit assumption in economic analysis was that such mandates could be analysed using the standard tools of tax incidence and efficiency.

A very influential article by Summers (1989) changed all that. Summers pointed out that mandating employer provision of benefits to their employees had two effects on labour market equilibrium. On the one hand, a reduction in labour demand naturally accompanies the imposition of extra costs on employers. On the other hand, however, mandates should also cause an outward shift in labour supply, since individuals are now being effectively compensated more highly for their labour; they are receiving

their previous compensation plus the mandated benefit. This shifts more of the costs of benefits to workers and reduces the deadweight loss from their provision. Indeed, as Summers pointed out, if employees valued the mandated benefit at its cost to the employer, then these supply and demand shifts would be equal. The end result would be ‘full shifting to wages’: a decline in wages by exactly the cost of the benefit with no impact on total labour supplied to the market and no efficiency consequences. Employees would simply be buying a benefit they value with their wages.

This article inspired a large follow-up literature, mostly empirical, investigating the equity and efficiency properties of mandates. I review that literature here, in three steps. First, I comment on the theoretical points made by Summers. Second, I discuss the empirical evidence available on the impacts of mandates. Finally, I discuss the key unanswered questions that must be addressed by future research.

Theoretical background

Summers’ analysis was as straightforward as it was insightful, highlighting the impacts of mandates in a simple demand and supply framework. The mathematics behind this analysis is explored in Gruber (1992), Gruber and Krueger (1991) and Anderson and Meyer (1997). These analyses show that the incidence of mandated benefits depends on the elasticities of supply and demand, as with any tax, along with a new parameter: the valuation of the benefit by employees. If valuation is equal to the cost paid by the employer for the benefit, then there is full shifting to wages.

But this analysis misses an important point: Summers’ analysis is in no way restricted to mandates. Indeed, the analysis is exactly the same for Unemployment Insurance, a US programme which provides tax-financed benefit to unemployed workers. The key to Summers’ analysis is not the form of provision (mandate or tax); rather, the key is that the benefits are *restricted to workers*, generating the labour supply increase that offsets some of the efficiency consequences of the intervention. For example, a payroll tax-financed expansion of health insurance to workers fits into this framework, but a payroll tax-financed expansion of health insurance to all individuals in society does not. In the latter case, there would not be the corresponding increase in labour supply, since individuals would not have to work to receive the benefit.

Another question raised by Summers’ analysis is this: if there were full incidence on wages, why wouldn't employers simply provide the benefit voluntarily? Why is government coercion necessary to promote employer provision of a benefit fully valued by employees? The best answer here, as pointed out by Summers, is that there may be market failures that lead employers to not reflect workers’ valuation of this programme without a government mandate. Most obviously, adverse selection in the market for benefits could cause employer reluctance to be, for example, the one employer in town that offered health insurance or paid maternity leave. This standard adverse selection problem may keep employers from offering benefits that are fully valued by employees. (Indeed, if there is such a market failure, it is feasible that a programme such as workers’ compensation could *raise* the quantity of labour in the market. If workers value workers’ compensation at more than its cost to employers – as might be the case if workers are risk averse – the labour supply curve would shift out by more than the demand curve shifted in; workers would be willing to accept a wage cut of *more* than the cost of workers’ compensation in order to have this benefit. This would actually raise employment.)

Empirical evidence

During the 1990s a large number of articles explored the empirical impact of mandates, in particular the extent to which mandated costs were shifted to wages. This literature is reviewed in detail in Gruber (2001); I provide an overview here. The consensus of this literature is that, over the medium to long term, the cost of mandates is fully reflected in wages.

Gruber and Krueger (1991) provide the first such analysis, dealing with increases in the employer costs of Workers' Compensation (WC) insurance across US industries and states over time. WC provides cash benefits and health coverage to workers injured on the job, and much of the variation in costs in the authors' data comes from increases in the health care component of this programme. They focus on workers in five industries for which WC costs are high and rapidly growing; in some industries and states, these costs amounted to over 25 per cent of payroll by 1987, the end of their sample period. They use both micro-data on wages and aggregate data on employment and wages by state/industry. They include state and industry fixed effects in their models, so that they are controlling for general differences in pay across industries and places, and estimating only how that pay changed when the costs of WC rose. In both data-sets, they find that for these sets of industries 85 per cent of increases in workers compensation costs were shifted to wages.

Anderson and Meyer (1997) undertake a similar analysis for Unemployment Insurance (UI), which provides cash benefits to unemployed workers. This programme is not a mandate, but it should operate in the same fashion as it levies payroll taxes on firms to provide benefits to their workers. Anderson and Meyer's conclusion is similar to Gruber and Krueger: general differences in UI payroll taxes appear to be fully reflected in wages with little effect on labour supply.

There is also a long literature on the impact of payroll taxation on wages that is reviewed by Hamermesh (1987). This literature is much more mixed in its conclusions, although the variation in payroll taxes mostly comes over time, and it is difficult to estimate its incidence separately from other time series factors in the United States where there is little variation across workers in payroll tax rates. More recent evidence is consistent, however, with the notion of full shifting to wages in other countries (for example, Gruber, 1997).

Labour supply is not simply a discrete choice, however, but rather a combination of participation and hours of work decisions. Increases in costs will effect both the supply of and the demand for work hours conditional on participation. From the employer perspective, increases in health insurance costs are an increase in the fixed cost of employment and are as a result more costly (as a fraction of labour payments) for low-hours employees. Employers will therefore desire increased hours by fewer workers, lowering the cost per hour of the health insurance for a given total labour supply. Of course, if the wage offset is lower for low-hours workers, workers will demand the opposite outcome: there will be increasing demand for part-time work, with hours falling and employment increasing. Moreover, since part-time workers may be more readily excluded from health insurance coverage, there may also be a countervailing effect on the employer side, as full-time employees are replaced with their (uninsured) part-time counterparts. In this case as well, hours would fall and employment would rise. Thus, the effect on hours of work is uncertain. Several studies have addressed this issue, and the general consensus is that mandating fixed costs of employment leads to rises in hours and falls in employment (Gruber, 1994; Cutler and Madrian, 1998).

Remaining questions

While there have been significant gains in our understanding of the impacts of mandates, important questions remain unanswered. The most important is the question of heterogeneity across workers: how do mandates affect workers differentially within the workplace? Consider the example of mandated health insurance. The cost of health insurance will not be uniform throughout the workplace; costs are higher for family insurance than for individual coverage, or for older workers than for younger workers. In the limit, with extensive experience rating, costs vary worker by worker, depending on their underlying health status. Gruber (1992) extends the model of Gruber and Krueger (1991) to the case of two groups of workers, where costs increase for one but not the other. If there is group-specific shifting, then the solution collapses to the one group model. If not, however, the substitutability of these groups will also determine the resulting labour market equilibrium; in general, there will be effects on both the group for which costs increase and the group for which they do not.

In practice, there may be a number of barriers to group-specific, and in particular individual-specific, shifting. Most obviously, there are anti-discrimination regulations which prohibit differential pay for the same job across particular demographic groups, or which prevent differential promotion decisions by demographic characteristic. Workplace norms which prohibit different pay across groups or union rules about equality of pay may have similar effects. Thus, a central question for incidence analysis is *how finely* firms can shift increased costs to workers' wages. If there is imperfect group- or worker-specific shifting, there may be pressure on employers to discriminate against costly workers in their hiring decisions.

Two studies suggest that there is within-workplace shifting to wages. Gruber (1994) studied the effect of state laws (and a follow-up federal law) that mandated in the mid-1970s that the costs of pregnancy and childbirth be covered comprehensively. Before this time, health insurance plans provided very little coverage for the costs associated with normal pregnancy and childbirth, while providing generous coverage for other medical conditions. This distinction was viewed as discriminatory by some state governments, leading to the state laws mandating that pregnancy costs be covered as completely as other medical costs. These laws significantly increased the insurance costs for women of childbearing age in those states, thereby raising the costs of employing a specific group of workers (or their husbands, who may provide them with insurance). I estimated full shifting to wages for these groups. Further corroborating evidence on this point is provided by Sheiner (1999), who found that, when health-care costs rise in a city, the wages of workers who have the highest costs (older and married workers) fall the most.

This research suggests that within-workplace shifting to wages is possible. The news here is good for efficiency: mandates which have differential effects across broad groups of workers will not necessarily lead to displacement of the high-cost group. The news is potentially bad for equity, however: other groups will not cross-subsidize the high costs imposed on one group in the workplace. In any case, neither of these studies addresses the extent to which within-firm shifting is possible; in the limit, it is hard to conceive that employers could shift costs to wages on a worker-by-worker basis.

Other questions have not been addressed at all by the literature. First, how rapidly does shifting to wages occur? Despite the evidence that mandates are fully reflected in wages, employers vociferously oppose mandated benefits as costing jobs. The reason for their opposition could be that wages cannot adjust

quickly enough in the short run to offset displacement effects; the studies cited earlier show full shifting only over several-year periods.

Second, what are the effects of existing constraints on compensation design in the labour market? For example, for workers already at the minimum wage, firms will be unable to shift to wages increases in the cost of health insurance. Similarly, union contract or other workplace pay norms may interfere with the adjustment of wages to reflect higher costs. These institutional features could increase the disemployment effects of rising health costs.

Third, what is the underlying structural mechanism behind a finding of full shifting to wages? In the simple labour market framework above, there are two reasons why increased costs might be shifted to wages: because individuals value the benefits that they are getting fully; or because labour supply is perfectly inelastic. Disentangling these alternatives is very important for future policy analysis. Consider the example of national health insurance, which is financed by a mandate, with an additional payroll tax to cover non-workers. If the full shifting documented earlier is due to full employee valuation with somewhat elastic labour supply, then national health insurance will have important disemployment effects, since labour supply will not increase in response to a benefit that is not restricted to workers. If full shifting is due to inelastic supply, however, then the population which is receiving benefits is irrelevant; in any case the costs will be passed on to workers' wages, so national health insurance will not cause disemployment. Existing evidence, as reviewed in Gruber (2001), is mixed on which of these channels is at work.

Conclusion

Since the early 1990s there has been a substantial growth in research on mandated benefits. The conclusions from the work to date are clear: the costs of mandated benefits are fully shifted onto wages, with little impact on total labour supply. But important questions about the mechanisms behind such shifting remain unanswered.

Bibliography

Anderson, P. and Meyer, B. 1997. The effect of firm specific taxes and government mandates with an application to the U.S. Unemployment Insurance program. *Journal of Public Economics* 65, 119–44.

Cutler, D. and Madrian, B. 1998. Labor market responses to rising health insurance costs: evidence on hours worked. *RAND Journal of Economics* 29, 509–30.

Gruber, J. 1992. The efficiency of a group-specific mandated benefit: evidence from health insurance benefits for maternity. Working Paper No. 4157. Cambridge, MA: NBER.

Gruber, J. 1994. The incidence of mandated maternity benefits. *American Economic Review* 84, 622–41.

Gruber, J. 1997. The incidence of payroll taxation: evidence from Chile. *Journal of Labor Economics* 15 (3, Part 2), S72–S101.

Gruber, J. 2001. Health insurance and the labor market. In *The Handbook of Health Economics*, ed. J. Newhouse and A. Culyer. Amsterdam: North-Holland.

Gruber, J. and Krueger, A. 1991. The incidence of mandated employer-provided insurance: lessons from Workers' Compensation insurance. In *Tax Policy and the Economy 5*, ed. D. Bradford. Cambridge, MA: MIT Press.

Hamermesh, D.S. 1987. Payroll taxes. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Sheiner, L. 1999. Health care costs, wages, and aging. Finance and Economics Discussion Series 1999–19, Board of Governors of the Federal Reserve System.

Summers, L. 1989. Some simple economics of mandated benefits. *American Economic Review* 79, 177–83.

How to cite this article

Gruber, Jonathan. "mandated employer provision of employee benefits." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000360> doi:10.1057/9780230226203.1017

The New Palgrave Dictionary of Economics Online

Mandel, Ernest (1923 – 1995)

Geoffrey M. Hodgson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; central planning; Kondratieff cycles; labour theory of value; long waves; Mandel, E.; Marx, K.; neo-liberalism; socialism; Soviet Union; Trotsky, L.; unemployment

Article

Ernest Mandel was born of Jewish parents in Frankfurt-am-Main on 5 April 1923. The family emigrated to Antwerp. By 1939 he was actively involved in socialist and trade union politics. When the Nazis invaded Belgium in 1940, he became a member of the resistance. On three occasions he was arrested and imprisoned, but each time he escaped. He was arrested for a final time in October 1944 and liberated by the Allies in March 1945. He obtained a higher education in Brussels and Paris. His name was prominent in academia in the 1960s and 1970s when Marxism and Trotskyism enjoyed significant popularity, particularly among university students. He died on 20 July 1995.

His *Marxist Economic Theory* was first published in French in 1962 and in English in 1968. When student revolts and labour unrest broke out in the late 1960s, Mandel's text and the much shorter *Introduction to Marxist Economic Theory* (1967) were available for the growing numbers interested in Marxist economics. His *Marxist Economic Theory* was widely praised and his *Introduction* sold over half a million copies and was translated into 30 languages.

His *Formation of the Economic Thought of Karl Marx* was published in French in 1967 and in English in 1971. It was one of the first works in English to analyse Marx's *Grundrisse*, which did not appear in complete form in English until 1973.

In his *Europe vs. America* – published in German in 1968 and in English in 1970 – he predicted relative economic decline and increasing ‘public squalor’ in the United States, sustained rapid economic growth in Japan, and the achievement of productivity levels in the western European ‘core’ regions to rival those in America.

In his *Late Capitalism* – published in German in 1972 and in English in 1975 – he revisited the idea that capitalism was subject to repeated waves of boom and stagnation in 45–60 year cycles. Not only did Mandel predict the downturn of the 1970s on the basis of this analysis, but also this work help to revive

academic interest in the study of long waves, which has continued to the present day. However, his analysis has been criticized for misunderstanding Trotsky's criticisms of Kondratiev (Day, 1976) and lacking a plausible mechanism to explain the complete long-wave cycle (Tylecote, 1992).

Mandel wrote introductions to the new English translations of the three volumes of Marx's *Capital*, published by Penguin (Marx, 1976; 1978; 1981). He was obliged to consider the stormy technical debates in the 1970s over the labour theory of value and Marx's theory of the tendency of the rate of profit to fall (Steedman, 1977). However, instead of addressing the detailed critical arguments, he simply brushed them aside.

In 1978 Mandel was invited by the University of Cambridge to give the prestigious Alfred Marshall Memorial lectures. These were published as *Long Waves of Capitalist Development* (1980): a restatement and development of ideas in *Late Capitalism*. Further weaknesses in his position emerged when it became clear that mass unemployment in the West was not leading to political advances for socialism. Instead the period saw a resurgent political individualism and neoliberalism.

Like Trotsky, Mandel opposed the view that the Soviet-type economies were another type of ‘capitalism’, envisaged a collapse of the Soviet regimes, and expected that the working class would rise up in defence of state planning and nationalized property. Even after their collapse in 1989–91, in his *Power and Money* (1992) he hoped for a new workers’ movement in eastern Europe and predicted that capitalism would not be easily re-established. Overall, the theoretical weakness of his outlook became increasingly clear in the last 15 years of his life.

See Also

- capitalism
- Kondratieff cycles
- Marx's analysis of capitalist production
- socialism
- socialism (new perspectives)
- Soviet economic reform
- Trotsky, Lev Davidovitch

Selected works

1967. *An Introduction to Marxist Economic Theory*. New York: USA.

1968. *Marxist Economic Theory*, 2 vols, trans. B. Pearce from the French edition of 1962. London: Merlin.

1970. *Europe vs. America: Contradictions of Imperialism*, trans. from the German edition of 1968. New York: Modern Reader.

1971. *The Formation of the Economic Thought of Karl Marx: 1843 to Capital*, trans. B. Pearce from the

French edition of 1967. London: NLB.

1975. *Late Capitalism*, trans. from the German edition of 1972. London: NLB.

1980. *Long Waves of Capitalist Development: The Marxist Interpretation*. Cambridge: Cambridge University Press.

1992. *Power and Money: A Marxist Theory of Bureaucracy*. London: Verso.

Bibliography

Day, R. 1976. The theory of long waves: Kondratiev, Trotsky, Mandel. *New Left Review* 99, 67–82.

Marx, K. 1976. *Capital*, vol. 1, trans. B. Fowkes from the fourth German edition of 1890. Harmondsworth: Pelican.

Marx, K. 1978. *Capital*, vol. 2, trans. D. Fernbach from the German edition of 1893. Harmondsworth: Pelican.

Marx, K. 1981. *Capital*, vol. 3, trans. D. Fernbach from the German edition of 1894. Harmondsworth: Pelican.

Steedman, I. 1977. *Marx After Sraffa*. London: NLB.

Tylecote, A. 1992. *Long Waves in the World Economy: The Present Crisis in Historical Perspective*. London: Routledge.

How to cite this article

Hodgson, Geoffrey M. "Mandel, Ernest (1923–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000384> doi:10.1057/9780230226203.1018

The New Palgrave Dictionary of Economics Online

Mandeville, Bernard (1670 – 1733)

N. Rosenberg

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

division of labour; laissez-faire; Mandeville, B.; self-interest; Smith, A.

Article

Mandeville was born in or near Rotterdam in 1670 and died in Hackney, London, in 1733. He was awarded the degree of Doctor of Medicine from the University of Leyden in 1961. He took up the practice of medicine, specializing in the ‘Hypochondriack and Hysterick Diseases’, a subject on which he later published a treatise. Mandeville travelled to England, married there in 1699, and lived in England for the rest of his life. He was very widely read in the 18th century. His writings have often led to his being referred to as a satirist, but that is an inadequate and misleading classification.

Although Mandeville was not an economist, his writings were influential in shaping the direction of economic thinking in the 18th century. In 1705 he published a pamphlet, in doggerel verse, under the title *The Grumbling Hive: Or Knaves turn'd Honest*. In 1714 it was republished under its better-known title, *The Fable of the Bees: or, Private Vices, Publick Benefits*. This and subsequent editions included extensive expansions, clarifications and ‘vindications’ of his earlier themes. The grumbling hive was originally a thriving and powerful community. When, however, its inhabitants were suddenly and miraculously converted from a vicious to a virtuous moral condition, the community was swiftly reduced to an impoverished and depopulated state.

Mandeville's central theme is that public benefits are the product of private vices and not of private virtues. His paradox, which was widely regarded as scandalous, was achieved by employing a highly ascetic and self-denying definition of virtue. Since behaviour that could be shown to be actuated by even the slightest degree of self-regarding motive – pride, vanity, avarice or lust – was classified as vice, Mandeville had little difficulty in concluding that a successful social order must inevitably be one where public benefits are built upon a foundation of private vices.

> … I flatter myself to have demonstrated that, neither the Friendly Qualities and kind Affections that are natural to Man, nor the real Virtues he is capable of acquiring by

> Reason and Self-Denial, are the Foundation of Society; but that what we call Evil in this World, Moral as well as Natural, is the grand Principle that makes us sociable Creatures, the solid Basis, the Life and Support of all Trades and Employments without Exception: That there we must look for the true Origin of all Arts and Sciences, and that the Moment Evil ceases, the Society must be spoiled, if not totally dissolved. (Mandeville, 1732, vol. 1, p. 369)

What was of more enduring significance in Mandeville's views was his forceful and unapologetic popularization of the belief that socially desirable consequences would flow from the individual pursuit of self-interest. It is an essential part of Mandeville's argument that a viable social order can emerge out of the spontaneous actions of purely egoistic impulses, requiring neither the regulation of government officials, on the one hand, nor altruistic individual behaviour, on the other.

> As it is Folly to set up Trades that are not wanted, so what is next to it is to increase in any one Trade the Numbers beyond what are required. As things are managed with us, it would be preposterous to have as many Brewers as there are Bakers, or as many Woolen-drapers as there are Shoe-makers. This Proportion as to Numbers in every Trade finds it self, and is never better kept than when nobody meddles or interferes with it. (Mandeville, 1732, vol. 1, pp. 299–300)

Thus, Mandeville enunciates a vision of an economy that organizes itself and that allocates resources through the market place. Although there is no serious analysis of the workings of the market mechanism, there is the clear assertion that the unregulated market provides a system of signals and inducements such that the interactions of purely egoistic motives will somehow produce results that will advance the public good.

In developing his views, Mandeville offered many acute observations on the causes as well as the consequences of the division of labour in society. He regarded the division of labour as the great engine of economic improvement over the ages. It is the most reliable way for ‘savage People’ to go about ‘meliorating their Condition’. For

> … if one will wholly apply himself to the making of Bows and Arrows, whilst another provides Food, a third builds Huts, a fourth makes Garments, and a fifth Utensils, they not only become useful to one another, but the Callings and Employments themselves will in the same Number of Years receive much greater Improvements, than if all had been promiscuously follow'd by every one of the Five. (Mandeville, 1729, vol. 2, p. 284)

Although one can identify a number of possible precursors to Adam Smith's celebrated views on the division of labour, it is well established that Smith had in fact read and digested Mandeville carefully. Smith's marvellous description of the extensive division of labour involved in the production of a day-labourer's coat, with which he closes the first chapter of the *Wealth of Nations*, may be traced to Mandeville's earlier treatment of the same subject – a treatment which, indeed, Smith extensively paraphrases. Moreover, the passage in the *Wealth of Nations* containing the often quoted statement that

‘It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest’ (Smith, 1776, p. 14) is in a direct lineage from Mandeville's earlier observation:

> … The whole Superstructure [of Civil Society] is made up of the reciprocal Services, which Men do to each other. How to get these Services perform'd by others, when we have Occasion for them, is the grand and almost constant Sollicitude in Life of every individual Person. To expect, that others should serve us for nothing, is unreasonable; therefore all Commerce, that Men can have together, must be a continual bartering of one thing for another. (Mandeville, 1729, vol. 2, p. 349)

Thus Mandeville was, in some important respects, an early advocate of laissez-faire (although this advocacy did not extend to foreign trade, where Mandeville's views were still distinctly Mercantilist). He articulated a vision of the role of the division of labour in society, and of the forces making for social change and evolution, as well as for social cohesion, that were in many respects distinctly precocious, and that exercised a powerful influence in shaping the intellectual agenda of economists and other social scientists later in the 18th century.

Selected works

1714; 1729; 1732. *The Fable of the Bees*, 2 vols, ed. F.B. Kaye. London: Oxford University Press, 1924. (This is the definitive edition.)

Bibliography

Hayek, F.A. 1978. Dr. Bernard Mandeville. In *New Studies in Philosophy, Politics, Economics and the History of Ideas*, ed. F.A. Hayek. Chicago: University of Chicago Press.

Primer, I. (ed.) 1975. *Mandeville Studies*. The Hague: Martinus Nijhoff.

Rosenberg, N. 1963. Mandeville and laissez-faire. *Journal of the History of Ideas* 24(2), 183–96.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. New York: Modern Library, 1937.

Stephen, L. 1876. *History of English Thought in the Eighteenth Century*, 2 vols. London: Smith, Elder.

Viner, J. 1953. *Introduction to Bernard Mandeville, A Letter to Dion*. Los Angeles: Augustan Reprint Society, Publication No. 41.

How to cite this article

Rosenberg, N. "Mandeville, Bernard (1670–1733)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000028> doi:10.1057/9780230226203.1019

The New Palgrave Dictionary of Economics Online

Mangoldt, Hans Karl Emil von (1824 – 1868)

H.C. Recktenwald

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

entrepreneurship; equilibrium; Historical School, German; joint supply and demand; Mangoldt, H. K. E. von; multiple equilibria; price theory; product life cycle; profit and profit theory; rents of differential ability

Article

Mangoldt was born in Dresden in 1824 and died in 1868 of a heart attack after a short life. He was an eminent theorist in economics, yet greatly underrated by his German contemporaries. He received his doctorate in Tübingen (1847) and was afterwards a civil servant in the Ministry of Foreign Affairs – a post he resigned for political reasons – and became editor of the official *Weimarer Zeitung* (1852). His academic career began in 1855 as Privatdozent in Göttingen and ended as Professor of Political Science and Political Economic after only six years (1862–8) at the University of Freiburg (Breisgau).
Mangoldt ranks among those pioneers in Germany, like von Thünen, von Buquoy, von Hermann, Gossen and Launhardt, who applied formal analysis to explain economic phenomena. Yet the predominant influence of the Historical School diminished the impact of his methods and ideas on German university economists. He shared this fate with Cournot and Walras. A second reason for this underrating of his pioneering achievements at home was his strong interest in classical economics. Thus it is not surprising that his reputation was much higher in England (via Edgeworth and Marshall) than in 19th-century Germany.
In his most important books, *Unternehmergewinn* (1855) and *Grundriss* (1863), he argues in the classic tradition, examines its hypotheses in the light of economic and political reality and modifies them considerably. Like Cournot (earlier) and Marshall (later) Mangoldt uses a novel apparatus of partial analysis – Frisch's microanalysis – to expound originally a mathematical theory of prices that goes far beyond Cournot. He describes in a very modern way the process from one equilibrium to another, analyses multiple equilibria and explains joint supplies and demands, a concept which Marshall would take up later on. Further, he has deeply influenced our theories of profit and rent by interpreting the entrepreneurial gain as rents of differential ability. Indeed, Mangoldt definitely anticipates Schumpeter's

theory of the entrepreneur. He clearly distinguishes profit as an independent category of income from interest (of the capitalist), by stressing different elements of gain such as the compensation for risk-bearing or for new goods or techniques of production and sale. The pioneer function of the entrepreneur, motivated also by intangibles, as in Smith's concept, is clearly expressed in the statement

> … die Auffindung und Verwirklichung der besten Produktionsmethoden … die Ausbeutung der von der Natur gegebenen Hilfsmittel, die Herstellung der Güter in der für das Bedürfnis dienlichsten Weise [the discovery and realization of the best methods of production, … the exploitation of natural resources, the manufacturing of goods in a way that is most appropriate for the need]. (1855, p. 68)

This means, of course, novel improvements as well.
Unfortunately, Mangoldt did not attempt to make these realistic and dynamic elements an essential part of his price theory via a notion of *evolutionary* competition as did Smith and Schumpeter. Thus he neglected the different properties and intensities of competition in different stages in the life cycle of a product.
Furthermore, his contribution to allocation theory was as pioneering as his analysis of coalitions on the labour market. Here he objected to profit participation by workers without risk-sharing. Finally, it is notable that Mangoldt originally extended Ricardo's theory of comparative costs by applying, although in rather vague terms, the notion of elasticity of demand and supply in the theory of international trade.
Mangoldt was, no doubt, one of the eminent theorists and rare pioneers in the 19th century whose achievements are still underrated and whose use of mathematics seems to be rather overrated. Though an abstract thinker, he seldom lost the binding ties to reality.

Selected works

1847. Über die Aufgabe, Stellung und Einrichtung der Sparkassen. Dissertation, Tübingen University.

1855. *Die Lehre vom Unternehmergewinn: ein Beitrag zur Volkswirtschaftslehre*. Leipzig: Teuber.

1863. *Grundriss der Volkswirtschaftslehre*, 2nd edn. Stuttgart: Maier. A chapter was translated as ‘The Exchange Ratio of Goods’, *International Economic Papers* 11, (1962), 32–59 11.

Bibliography

Recktenwald, H.C. 1951. Zur Lehre von den Marktformen. *Weltwirtschaftliches Archiv* 67(2), 298–326.

Recktenwald, H.C. 1985. Über das Selbstverständnis der ökonomischen Wissenschaft. *Jahrbuch der Leibniz-Akademie der Wissenschaften und der Literatur*. Wiesbaden: Steiner.

Recktenwald, H.C. and Samuelson, P.A. 1986. *Über Thünens ‘Der isolierte Staat’*. Darmstadt and

Dusseldorf: Wirtschaft und Finanzen.

How to cite this article

Recktenwald, H. C. "Mangoldt, Hans Karl Emil von (1824–1868)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000029> doi:10.1057/9780230226203.1020

The New Palgrave Dictionary of Economics Online

Maoist economics

Wei Li

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

During the Maoist era (1949–76), China attempted to adopt Soviet-style central planning in her first Five-Year Plan, but soon changed course. Under the heavy hand of Chairman Mao Zedong, China created a unique style of central planning where the centre enunciated broad policy directives in the form of slogans that could be easily passed down to local cadres, who were given strong incentives to find ways to implement them. This article outlines a consistent framework for analysing the policy changes during the Maoist era and their dramatic impact on the Chinese economy.

Keywords

agricultural productivity; agricultural taxation; collectivization; Cultural Revolution (China); famines; Great Leap Forward (China); Lysenkoism; Maoist economics; Marxist economics; nutrition; peasants; planning; tax compliance

Article

> Ten thousand years is too long; seize the day, seize the hour.
> (Mao Zedong, *Mengjianghong – A Reply to Comrade Guo Moruo*, 1963)
>
> Had Mao died in 1956, there would be no doubt that he was a great leader of the Chinese people, a respected, loved and outstanding great man in the proletarian revolutionary movement of the world. Had he died in 1966, his meritorious achievements would have been somewhat tarnished but still very good. Since he actually died in 1976, there is nothing we can do about it.
>
> (Chen Yun at the Central Party Work Conference, November–December 1978. Quoted

from Lardy and Lieberthal, 1983. *Ming-Pao* (Hong Kong) 15 January 1979)

‘Maoist economics’ refers to the collection of economic policies implemented by the Communist Party of China (CPC) during the Maoist era, which began with the founding of the People's Republic in 1949 and ended shortly after the demise of Chairman Mao Zedong in 1976. Thanks to the CPC's meticulous cultivation of Mao's personality cult, Mao was able to exploit his ‘mass line’ political strategy by exhorting the masses to follow his vision when the CPC hierarchy was unwilling. As a result, Mao could set major policy initiatives with few checks and balances. But to attribute all major decisions to Mao would be an oversimplification, especially before 1958. The leadership of the CPC in Beijing and local cadres, often split into factions with different policy agendas and preferences (for a detailed historical account of the policy debates within the leadership circle in China in the 1950s and 1960s, see for example, Lardy and Lieberthal, 1983; Riskin, 1987; and Bachman, 1997), contributed not only to policy implementation but also to policy formulation. Maoist economics is therefore not synonymous with ‘Mao Zedong Thought’ on economic matters. (‘Mao Zedong Thought’ is considered an extension of Marxism–Leninism derived from the teachings of Mao Zedong and the distillation of the experience of the Communist revolution in China. It has been enshrined in the Constitution of the CPC as part of the party's official ideology since 1945. As China has embarked on market-oriented reforms since 1978, ‘Deng Xiaoping Theory’, which advocated the pragmatic concept of ‘socialism with Chinese characteristics’, has served as the party's working doctrine.)

The aim of this article is to outline a consistent framework for organizing and understanding the economic policies that were formulated and implemented during the Maoist era.

Agricultural taxation and the Chinese-style central planning

When on 1 October 1949 Chairman Mao proclaimed that the Chinese people had finally stood up at the ceremony for the founding of the People's Republic, China was a desperately poor agrarian economy ravaged by more than a century of internal turmoil, foreign invasions and civil wars. With most of her industrial assets either destroyed or looted by the Soviet forces that occupied Manchuria at the end of the Second World War, or removed to Taiwan and Hong Kong ahead of advancing Communist troops, China was ‘poor and blank’, as Mao (1958) put it. With 90 per cent of her population of 550 million living in abject poverty in the countryside and toiling on small plots of land using traditional labour-intensive farming technology, China was barely able to feed and clothe her population.

Since poor peasants made up the vast majority of the population, the CPC under Mao had focused on building its support base among peasants by, among other things, promising to deliver what every peasant wanted: a private plot of land. Between 1946 and 1953, the CPC launched land reform, first in the territories under its control and then in all ethnically Chinese areas on the mainland after 1949. The process generally involved assigning each rural family a class status; motivating the poor, lower-middle, middle and initially the ‘rich’ peasants to engage in ‘class struggle’ against landlords; and expropriating land, draft animals, farm implements and property from landlords and redistributing them to landless peasants (Fairbank, 1992). The ‘class struggle’, which included public trials, denunciations and mass executions of landlords and counter-revolutionaries, created an atmosphere of terror. But the land reform solidified support for the CPC among the poor and middle peasantry. (For an on-the-ground observation

of the land reform in a Chinese village, see Hinton, 1967.)

As the CPC secured military and political control of the mainland, its priority shifted to managing and rebuilding the war-torn economy. With the economy rebounding quickly, Chinese leaders turned their attention to long-run economic development, aimed at building a socialist, industrial nation. Having secured material and technical support from the Soviet Union, they adopted a Soviet-style, heavy-industry-oriented development strategy in the first Five Year Plan (FYP 1953–7). The plan called for massive industrial investment, including the construction of 156 industrial plants outfitted with imported Soviet equipment. The Soviet contributions to this big push included loans amounting to about four per cent of the total investment, technology transfers and 10,000 Soviet specialists (Fairbank, 1992). The success of this ambitious plan therefore hinged on the ability of the government to mobilize investable surplus internally. Without a significant industrial and commercial sector, the government had to extract the needed surplus from the vast agricultural sector.

Throughout Chinese history, agricultural taxes, collected in kind, have been the primary source of government revenue. (Indeed the Chinese character for tax, *shui*, as a portmanteau of grain and convert, refers to a levy on the use of land payable in grain.) In the 1950s, China had a three-tiered agricultural tax structure. At the first tier was an in-kind levy on grain production, known as the ‘government grain’. Peasants received no compensation for turning over the government grain to the State Grain Bureau. At a statutory rate of 15 per cent in 1950, this tax accounted for 39 per cent of government revenue. (This figure is calculated using data posted on the official website of China's Ministry of Finance.) In later years, as the price scissors – the differences between the prices on industrial and agricultural goods – widened, and as the industrial sector grew rapidly because of the massive capital expenditure funded largely by agricultural taxes, the share of explicit agricultural taxes dropped to six per cent by 1976.)

At the second tier was an implicit tax, a grain procurement quota, which dictated how much each peasant household had to sell to the State Grain Bureau out of their after-tax grain at below-market procurement prices. After meeting these two obligations, peasants would usually be left with just enough grain to sustain a subsistent living. Markets still existed in the early 1950s, where peasants could exchange some of their surplus produce for other goods. At the third tier was an in-kind levy on rural labour. Under the traditional subsistence farming practice in China, peasants would take a break or work less intensively during agricultural off-seasons in order to conserve food energy. To the government, this idling was unacceptable. Dams, irrigation systems, roads and other large-scale infrastructure projects could be worked on more intensively during off-seasons by drafting peasants to carry out backbreaking manual labour. Utilizing a mixture of exhortation and coercion, the government mobilized tens of millions of peasants for large construction projects in the 1950s.

Collecting the three-tiered taxes from hundreds of millions of independent peasant households was a daunting task. Tax enforcement became even harder when market prices of grain rose substantially in 1952 as a result of increased demand caused by rapid industrialization and urbanization and by the need to export agricultural products in exchange for Soviet equipment. In response, the government in 1953 closed the grain market and monopolized grain trade by fiat, making it illegal for anyone other than the government to engage in large-scale grain trade. In 1954, it expanded the control to include oil seeds, cotton, pork, and other key agricultural commodities.

Extracting agricultural surplus was further hampered by the lower level of agricultural productivity in China than in more developed countries. With nearly 90 per cent of the population living in the countryside, China was producing barely enough food and wearable fibres to meet basic domestic needs.

Estimates by Ashton et al. (1984) suggest that the daily average food energy intake in China in the 1950s was around 2,000 calories per capita, below the 2,350 calories recommended by the United Nations.

To further improve its extractive capability and to raise agricultural productivity, the government turned to collectivization. By organizing peasant households into collectives, the CPC could extend its political control down to the village level. The grass-roots party organizations could effectively monitor production to further improve tax compliance. Rooted in the prevailing ideology, Chinese leaders also believed that collectivization would enable peasants to take advantage of economies of scale, to learn best practices in scientific farming, to accelerate the adoption of high-yield seeds and modern inputs, and therefore to realize a great leap in agricultural productivity. (Apparently influenced by Soviet propaganda, Chinese leaders were taken in by the miraculous claims of productivity-boosting farming techniques made by a group of pseudo-scientists who dominated the Soviet agricultural science establishment. Because these pseudoscientific techniques contradicted the farming experience of Chinese peasants, the only way to propagate them was to make it a political task for rural collectives and grass-roots party organizations. For an account of Lysenkoism in the Soviet Union and its influence in China during the Maoist era, see Becker, 1996.)

Collectivization, however, represented a radical reorganization of rural life in China. Given the importance of agriculture in the Chinese economy and the traumatic experience of forced collectivization in the Soviet Union (Becker, 1996), China's first FYP emphasized voluntary participation and set out a relatively conservative and flexible timetable, calling for socialist transformation in agriculture to be accomplished in 10–15 years. Between 1952 and 1954, collectivization proceeded gradually. By 1954, only 11 per cent of peasant households were enrolled in elementary Agricultural Producers’ Cooperatives (APCs), where members pooled their privately owned land, draft animals and large tools and used them jointly. APC members were paid wages for their labour as well as rents for their contributions in land and capital. While wage rates and rents were supposedly set at market levels, actual practice left many richer peasants complaining that the rents were insufficient. Reports of richer peasants exiting the cooperatives, selling and killing their draft animals, and downing trees on their plots in 1954 started to alarm leaders in Beijing. In January 1955, the CPC issued an urgent order for the protection of draft animals. The combination of state monopolization of grain trade and collectivization had, by the authorities’ own admission, dampened the ‘enthusiasm’ of the peasants for production. Emergency measures that the government implemented included fixing procurement quotas and putting on hold any further push for collectivization in the spring of 1955. But any reprieve that peasants got was short-lived.

By the summer of 1955, imbalances in the economy from implementing the aggressive first FYP had reached record levels. The supply of agricultural products, raw materials and consumer goods could not keep up with the growing demand. With tax revenues insufficient to meet the funding needs in the first FYP, the government was running a large fiscal deficit. (In 1955, debt issuance by the government reached a record high of 2.5 per cent of GDP.) Factors that contributed to the imbalances included the agricultural bottleneck exacerbated by the collectivization movement, the ambitious first FYP that allocated massive investment to heavy industry, and the inherent difficulties of managing a centrally planned economy.

Mao's own analysis, however, identified over-centralization as a serious problem of the Soviet-style central planning whereby the planners tried to do what could be done better by local cadres. The solution that Mao put forth was not to stop the expansion of the role of the state in the economy, but to limit the

role of the nascent central planning bureaucracy and expand the role of local governments. He faulted the planners in Beijing for not doing enough to harness the enthusiasm of local cadres, workers and peasants for socialist transformation both in industry and in agriculture. In a policy speech delivered on 31 July 1955, Mao made the argument for accelerating socialist transformation in general and collectivization in particular.

> [Some] comrades fail to understand that socialist industrialization cannot be carried out in isolation from the cooperative transformation of agriculture. In the first place, as everyone knows, China's current level of production of commodity grain and raw materials for industry is low, whereas the state's need for them is growing year by year, and this presents a sharp contradiction. If we cannot basically solve the problem of agricultural cooperation within roughly three five-year plans, that is to say, if our agriculture cannot make a leap from small-scale farming with animal-drawn implements to large-scale mechanized farming, … then we shall fail to resolve the contradiction between the ever-increasing need for commodity grain and industrial raw materials and the present generally low output of staple crops, and we shall run into formidable difficulties in our socialist industrialization and be unable to complete it. (Mao, 1977, pp. 196–7)

To ensure that Mao's vision was turned quickly into action, the CPC passed in October 1955 a resolution that reiterated the policy directive for accelerating collectivization and authorized the party hierarchy to criticize any party member who disagreed with the policy as a ‘right-leaning opportunist’. (‘The Resolution Regarding Agricultural Collectivization’ was passed in the 6th Plenary Meeting of the 7th CPC Congress held in Beijing from 4–11 October 1955.)

As local cadres who moved decisively and quickly to implement this policy directive were publicly praised, and laggards were publicly criticized, local cadres found themselves locked into a rat race on who could coerce peasants to form bigger collectives at a faster pace. By the end of 1956, 96.3 per cent of all peasant households had joined collectives, more than ten years ahead of the schedule set in the first FYP.

Mao's administrative decentralization was not a repudiation of the concept of central planning. It was an attempt to redefine central planning in the Chinese context with perhaps an implicit intent to enlarge the sphere of Mao's influence. By weakening the nascent central planning bureaucracy, Mao effectively strengthened his own influence in enunciating broad policy directives in the form of slogans that could be easily passed down to local cadres. To align the interests of local cadres with the centre, Mao offered high-powered incentives: those who found innovative ways to implement the centre's directives irrespective of economic consequences were rewarded with public praise and promotion, while those who ignored the centre's policy directives were punished with the humiliation of public criticism and denunciation. In more serious cases, those who resisted the centre's policy directives could be purged as ‘rightists’ or ‘counter-revolutionaries’. Mao also made frequent use of brutal political campaigns against nonconformists and instilled an atmosphere of terror. (One of the most notorious political campaigns was the 1957 ‘anti-rightist’ campaign; Fairbank, 1992.) The resulting political system was one in which Mao could exploit his personality cult in enunciating broad policy directives without the inconveniences of checks and balances. Mao's administrative decentralization thus marked the beginning of the politicization of economic policy formulation and implementation in China. When Mao launched the

Great Leap Forward (GLF) movement in 1958, the inaugural year of the second FYP, there was hardly any dissenting voice.

The Great Leap Forward

By setting production targets even more aggressively in the second FYP, the CPC hoped that China would grow out of the imbalances created during the first FYP by exhorting local cadres and the masses to make selfless sacrifices in order to transform China into an industrial, socialist nation. In March 1958, the CPC issued a new directive, calling local cadres to amalgamate smaller cooperatives into larger ones. Zealous local cadres in Henan province created township-sized collectives, dubbed ‘People's Communes’. Each of the communes was an all-encompassing institution that functioned as a local government, an agricultural collective, local government-owned industrial and commercial enterprises (one of the enduring legacies of Mao's administrative decentralization was the policy directive that encouraged the creation of local government-owned enterprises), local schools, and a militia integrated into the national defence system. In these collectives, communalization went beyond all means of production and invaded the private lives of peasants. For example, family kitchens were banned and were replaced by communal kitchens that offered members free meals (Li and Yang, 2005). ‘People's Communes are good because they are big and communal’, declared Mao. By early autumn, communes had spread across China.

Believing that collectivization significantly boosted agricultural productivity, the CPC created a new rat race for local cadres by exhorting them to ‘overcome reactionary conservatism’ (*People's Daily*, 10 September 1958). Unable to deliver the expected increase in grain output, local cadres started to outdo each other in statistical gamesmanship by making wild claims about grain output. An initial tally of the 1958 grain output after the autumn harvest pegged it at 525 million metric tons (MMTs), up by nearly 170 per cent from 1957. The figure was subsequently revised down to a more modest 375 MMTs. (The downward revisions did not stop here. Two more were made: first to 250 on 22 August 1959 and then to 200 in 1979; Li and Yang, 2005.)

With the numbers indicating that collectivization had permanently resolved China's agricultural bottleneck, the government raised agricultural taxes: grain procurement (including government grain) was increased from 46 million metric tons in 1957 to 64 in 1959; 16.4 million peasants, about twice the size of the industrial labour force in 1957, were relocated to cities in 1958 to support the expansion of industry and construction; and more than 100 million peasants were mobilized in the winter of 1957–8 to undertake large irrigation and land reclamation projects, and to operate millions of small ‘backyard iron furnaces’. (Built using mud and bricks, these furnaces melted scrap metal – for example, iron woks made obsolete by communal kitchens – to produce iron, which even the government admitted was of useless quality; Becker, 1996.) The increase in agricultural taxes allowed the government to raise national savings from 24.9 per cent of national income (measured by net material product) in 1957 to 43.8 per cent in 1959. These savings were almost exclusively invested in heavy industries (Riskin, 1987, p. 142). Grain export was raised from an average of 2.11 million tons between 1953 and 1957 to 3.95 million tons in 1959 to meet payment obligations for importing capital goods.

The collectivization miracle was, however, a mirage. Lin (1990) finds that incentive problems within large collectives had deleterious effects on agricultural productivity. With actual grain output

significantly lower than the falsified statistics, the agricultural taxes were excessive. Grain retained in rural areas fell sharply from 273•kg per capita in 1957 to 193•kg in 1959, and further down to 182•kg in 1960. Since grain was the primary source of food energy in China at the time, the drop in per capita food availability coincided with the onset of worst famine in human history. (Demographers who extrapolated mortality trends in China estimated the total number of premature deaths during the GLF famine at between 16.5 and 30 million; see Li and Yang, 2005.)

As the disastrous consequences of the GLF policies became known in 1959, Mao temporarily stepped aside to let his pragmatic colleagues, Liu Shaoqi and Deng Xiaoping, take responsibility in managing both government and party affairs. The pragmatic leaders started to reverse course: they reduced grain procurement by ten million tons, increased agricultural labour force by more than 50 million between 1958 and 1962 by sending back new industrial recruits back to the countryside, dismantled communal kitchens, downsized the collectives and started to import grain. More importantly, they allowed spontaneous, bottom-up experimentation with market-oriented reforms in 1961. Grain output began to recover in 1961, but did not surpass its pre-GLF level of 195 million metric tons (recorded in 1957) until 1966, the first year of yet another political upheaval – the Cultural Revolution.

The model

To better understand the trade-offs faced by Chinese policymakers and the key factors that contributed to the GLF disaster, I turn next to Mao's policy directive for accelerating collectivization with the aid of a simple two-sector dynamic model developed in Li and Yang (2005).

A key feature of the Li–Yang model is the explicit dependence of agricultural labour productivity on nutrition. For simplicity, assume that in the agricultural sector labour is the only factor and the technology exhibits constant returns to scale. If L_t is labour allocated to agriculture, the aggregate grain output in year t can be written as

$$Q_t = af(c_t)L_t \tag{1}$$

where $af(c_t)$ measures the contribution of nutrition to the labour productivity of an average worker who consumes c_t amount of grain in year t, and a is a productivity parameter. Experimental and empirical studies have found that $f(\cdot)$ tends to be an increasing, S-shaped function with $f''(c) \geq 0$ at a very low level of food intake, and $f''(c) < 0$ as food intake reaches a sufficiently high level. (For a survey on health, nutrition and economic development, see Strauss and Thomas, 1998.) If the government taxes away p_t amount of grain output from each agricultural worker after the harvest in year t, the amount of grain saved for consumption in year t+1 is then

$$c_{t+1} = af(c_t) - p_t. \tag{2}$$

The industrial sector uses a Leontief technology that produces one unit of industrial output by employing one unit of labour, *d* units of capital service and *m* units of grain as an intermediate input in fixed proportions. Assume that all capital goods must be imported and paid for by exporting grain, and the exchange rate is one unit of grain to one unit of capital service. With abundant grain supply and the economy's labour supply normalized to 1, the industrial output is simply $1–L_t$. The government is assumed to maximize a discounted flow of industrial output, $\sum_{t=0}^{\infty}\beta^t(1-L_t)$, subject to the following budget constraint:

$$p_t L_t \geq (d+m+n)(1-L_t), \tag{3}$$

where $\beta<1$ is the government's discount factor and *n* is the food entitlement of each industrial worker. (For more discussion on food entitlement, see Li and Yang, 2005. In 1956, the national average of monthly ration of grain for labourers assigned to the most physically demanding jobs was 25•kg. Retail prices of food items in stores were set by the government and played little role in resource allocation.) This constraint, which captures China's key bottleneck during the Maoist era, states that the amount of grain procured must be sufficient to meet export demand for the importation of capital goods, industrial demand for intermediate inputs, and food demand from industrial workers.

Given the government objective, the optimal solution calls for allocating just enough labour to grain production, so the constraint (3) is binding in each year. This implies that the optimal allocation of labour to grain production should be $L_t = (d+m+n)/(p_t+d+m+n)$. Substituting this binding constraint into eq. (2), one can show that the government's optimal policy is a solution to the following Euler equation for a given initial level of food consumption c_0:

$$a\beta f'(c_{t+1}) = \left(\frac{af(c_{t+1}) - c_{t+2} + d + m + n}{af(c_t) - c_{t+1} + d + m + n}\right)^2. \tag{4}$$

The optimal steady state policy is to set the food consumption per agricultural work $\bar{c}$ such that $f'(\bar{c}) = (a\beta)^{-1}$. The steady state is asymptotically stable if $f''(\bar{c}) < 0$. This stability condition is

satisfied if the productivity effect of nutrition exhibits diminishing returns around the steady state per capita food consumption, which is consistent with previous experimental findings.

Under the stability condition, the steady state grain procurement $\bar{p}$ and industrial output $1-\bar{L}$ are both increasing functions of agricultural productivity *a* and the discount factor β. The model therefore validates Mao's claim in his quoted policy speech that raising agricultural productivity would contribute to the relaxation of the agricultural bottleneck and hence permit a faster pace of industrialization. It also proves that patience is a virtue: a more patient government, one that uses a larger discount factor β (or a lower discount rate) in setting intertemporal policies, can sustain a higher level of steady state agricultural and industrial production. The intuition is as follows. A more patient government, one that discounts future industrial production at a lower rate and is content with a lower growth rate, would set a lower tax rate on peasants, allowing them to improve nutrition and labour productivity. The improved productivity would in turn increase the tax base sufficiently high to more than compensate for any revenue loss from lowering the tax rate. As a result, both the grain procurement and industrial production are higher in the steady state for a more patient government.

Like Stalin, Mao was impatient. (In a speech delivered in 1931, Stalin, 1952, used nationalistic rhetoric to demonstrate the imperative to press on with rapid industrialization regardless of the obstacles during the first Five Year Plan of 1928–32 in the USSR.) And, like Stalin, Mao saw collectivization as a means to achieve rapid industrialization. Expecting collectivization to raise *a*, the increasingly impatient planners exhorted local cadres to increase grain procurement and to divert more agricultural labour to industrial production and large infrastructure projects. Since collectivization actually caused *a* to fall, the GLF policies left many peasants with insufficient amount of grain for consumption. Malnutrition (and famine in several grain-producing provinces) significantly reduced labour productivity, leading to a collapse in grain production. The further reduction in grain output caused malnutrition, and famine to spread from the countryside into cities. The linkage between nutrition and productivity thus offers a dynamic explanation of why the negative incentive effect of collectivization could cascade into a major catastrophe. Empirical investigation by Li and Yang (2005) finds that the GLF policies were principally responsible for this disaster. As the GLF policies were reversed by Liu Shaoqi and Deng Xiaoping, the Chinese economy began to stabilize in 1962. In 1966, when the economy appeared to have fully recovered by 1966, Mao launched another political campaign – the Great Proletarian Cultural Revolution.

The Cultural Revolution

The post-GLF policies had some noteworthy features. First, the centre–province distribution of power was rebalanced in favour of the centre. The task of collecting reliable information on the prevalence and severity of famine was simply too important to be delegated to local cadres. Second, collectives were downsized by making village-level ‘production brigades’ responsible for their own finances, and communal kitchens were closed. More important, the policies permitted spontaneous experimentation with household responsibility schemes within collectives, allowed peasant families to keep small private plots, and reopened markets in which peasants could sell their surplus produce. These policies arrested the downward momentum and brought about a gradual recovery.

But they represented a humiliating retreat in the campaign towards socialism. As long as the retreat was

tactical, Mao was content standing on the sideline. However, Khrushchev's denunciation of Stalin's rule in 1956 and the subsequent de-Stalinization in the Soviet Union gave Mao reasons to be concerned about his own legacy. As soon as the economy recovered, Mao moved to reclaim power so that he could purge those who had the potential to become China's Khrushchev. In 1966, Mao turned against Liu and Deng. Exploiting his personality cult, Mao kicked off the ‘Great Proletarian Cultural Revolution’ in 1966 by exhorting the Red Guards, made up primarily of students and other urban youths, to rebel against the power base of Liu and Deng – the government and party hierarchy. Liu and Deng, along with many of their colleagues, were labelled ‘capitalist roaders’ and were purged in 1968.

Mindful of the fragile conditions in the Chinese countryside, moderate leaders did their best to keep the revolution from spreading into the countryside, preventing a rerun of the famine during the GLF. But the market-oriented reforms permitted under Liu and Deng were nullified. Agricultural productivity continued to stagnate until market-oriented reforms were restarted in 1978. The demand for food, however, continued to grow as a result of the post-war baby boom. Unable to raise grain procurement quotas to meet the growing demand for food rations in the cities, the government resorted to sending millions of urban youths to the countryside to grow their own food and to receive ‘re-education’.

The Cultural Revolution brought politicization to every facet of life in China. It was better to be revolutionary (that is, loyal to Mao) than productive. Intellectuals and experts, considered less loyal to Mao, were sent to re-education camps in the countryside. Colleges were closed at first and were reopened later to admit only students from ‘revolutionary families’ – families of workers, peasants and soldiers – based on recommendations from grass-roots party organizations. Seasoned bureaucrats and factory managers were purged by the Red Guards, and ‘Revolutionary Committees’, comprised of workers, peasants and students, took over government offices and state-owned enterprises.

The revolution paralysed the government and the nascent economic planning apparatus. With neither the plan nor the market to guide the allocation of resources, the economy fell into a state of anarchy. As coordination across regions fell by the wayside, regional self-sufficiency, a policy stance endorsed by Mao, became a necessity. Specialization based on regional comparative advantage gave way to the duplication of industrial structure across provinces. The economy stagnated until 1978, when a rehabilitated Deng Xiaoping restarted market-oriented reforms.

Discussion

One of the classic tenets of Marxian economics is that, with planning eliminating the ‘anarchy of production’, a planned economy can avoid or at least better manage large aggregate economic fluctuations (Ellman, 1989). The experience of the Chinese-style central planning offers little support for this claim. The analysis of the GLF disaster by Li and Yang (2005) suggests that, on the contrary, central planning as practised in China exposed the economy to a new systemic risk. Because policy directives formulated at the centre had to be carried out in all localities, policy failures had generated large economic imbalances, severe economic and political crises, and prolonged stagnation. The source of the risk is the concentration of economic and political power in the hands of the planners. In the case of China, Chairman Mao, a charismatic leader, maintained a near monopoly on economic and political policies. With no effective checks and balances during the Maoist era, the economic and political system in China was incapable of arresting the momentum of apparently deleterious policy directives.

The Maoist era was tumultuous. It saw spectacular post-war reconstruction, the build-up of a

rudimentary industrial economy aided by Soviet assistance, and the formation of a decentralized government administration that emphasized regional self-sufficiency on the one hand and economic collapse, stagnation, a personality cult and brutal ‘class struggle’ on the other. It conditioned a generation of pragmatic leaders who, after the demise of Mao, would restart market-oriented reforms through decentralized regional experimentation, disown the personality cult, ban mass movements and depoliticize economic policymaking, while resolutely maintaining the CPC's hold on power. The historical significance of Maoist economics may lie not in what it is but in what it is not.

See Also

- agriculture and economic development
- China, economics in
- Chinese economic reforms
- command economy
- famines
- planning

Bibliography

Ashton, B., Hill, K., Piazza, A. and Zeitz, R. 1984. Famine in China: 1958–61. *Population and Development Review* 10, 613–45.

Bachman, D.M. 1997. *Bureaucracy, Economy, and Leadership in China: The Institutional Origins of the Great Leap Forward*. Cambridge: Cambridge University Press.

Becker, J. 1996. *Hungry Ghosts: Mao's Secret Famine*. New York: Henry Holt and Company.

Communist Party of China. Constitution. Online. Available at http://news.xinhuanet.com/ziliao/2004-11/24/content_2255749.htm, accessed 13 June 2007.

Ellman, M. 1989. *Socialist Planning*. Cambridge: Cambridge University Press.

Fairbank, J.K. 1992. *China: A New History*. Cambridge, MA: Harvard University Press.

Hinton, W. 1967. *Fanshen: A Documentary of Revolution in a Chinese Village*. New York: Monthly Review Press.

Lardy, N.R. and Lieberthal, K. 1983. Introduction. In *Chen Yun's Strategy for China's Development: A Non-Maoist Alternative*, ed. N.R. Lardy and K. Lieberthal. The China Book Project. Armonk, NY: M.E. Sharpe.

Li, W. and Yang, D.T. 2005. The Great Leap Forward: anatomy of a central planning disaster. *Journal of Political Economy* 113, 840–77.

Lin, J.Y. 1990. Collectivization and China's agricultural crisis in 1959–61. *Journal of Political Economy* 98, 1228–52.

Mao, Z. 1958. Introducing a cooperative. In Mao (1977).

Mao, Z. 1977. *Selected Work of Mao Zedong*, vol. 5. Beijing: Foreign Language Press.

Ministry of Finance, People's Republic of China. Online. Available at http://www.mof.gov.cn, accessed 20 February 2007.

Riskin, C. 1987. *China's Political Economy*. Oxford: Oxford University Press.

Stalin, J.V. 1952. *Economic Problems of Socialism in the USSR*. Moscow: Foreign Language Publishing House.

Strauss, J. and Thomas, D. 1998. Health, nutrition, and economic development. *Journal of Economic Literature* 36, 866–17.

How to cite this article

Li, Wei. "Maoist economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000386> doi:10.1057/9780230226203.1021

The New Palgrave Dictionary of Economics Online

Marcet, Jane Haldimand (1769 – 1858)

Robert W. Dimand and Evelyn L. Forget

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

British classical economics; Malthus, T. R.; Marcet, J. H.; Martineau, H.; Say, J.-B.; utility

Article

The classical political economist Jane Haldimand Marcet was born in London, the eldest of ten children of Anthony (Antoine) Haldimand, a Swiss citizen who was a successful London banker and property developer, and his English wife, Jane Pickersgill. She was tutored at home, studying the same subjects as her brothers, and took charge of the household at the age of 15, when her mother died. In December 1799 she married Alexander Marcet, a London physician from Geneva. Since her father bequeathed all his children an equal share of the family fortune, regardless of gender, she was independently wealthy, with no need to write for money. Nonetheless, she wrote 30 educational books on chemistry, political economy, botany, mineralogy, grammar and history, many written in the form of conversations. Her first book, an introduction to experimental chemistry, was published in 1806 after attending Humphrey Davy's lectures at the Royal Institution and after repeating Davy's experiments at home in Alexander Marcet's laboratory. The book was adapted in the United States as a college text, and its tremendous commercial success is shown by the many plagiarized editions that emerged in a period with no effective international copyright law. It introduced the young Michael Faraday to science.
Jane Marcet encountered the ideas of Adam Smith through Sydney Smith's lectures on moral philosophy at the Royal Institution in 1804 and 1806. Alexander Marcet and David Ricardo were both elected to the Geological Society in 1808. Jane Marcet's younger brother, William Haldimand (who lived with the Marcets), was elected a director of the Bank of England in 1809 at the age of 25, and, like his sister, shared Ricardo's attribution of the rising price of bullion to the excessive issue of bank notes, which was very much a minority view among the directors of the Bank of England. James Mill and Thomas Robert Malthus were also friends of the Marcets. Jane Marcet's *Conversations on Political Economy*, published anonymously in 1816, attempted to make the economic ideas of Smith, Malthus, Ricardo and Jean-Baptiste Say accessible to a wider public. Robert Torrens declared her ‘one female, at least, fully competent to instruct the members of the present cabinet in Political Economy’, while J.R. McCulloch

considered her book ‘on the whole, the best introduction to the science that has yet appeared’. Ricardo's daughter read the book at her father's recommendation, and Say wrote for permission to ‘translate sizeable passages from her excellent book’ for his political economy class (Polkinghorn, 1993, p. 55). Jane Marcet (1816; 1833; 1851) was a successful popularizer of classical political economy, but she was also fully capable of independent judgement, sharing Ricardo's opposition to the Corn Laws rather than Malthus's support for them, and supporting the proposed Factory Act in 1833, contrary to the beliefs of her younger friend, Harriet Martineau. Marcet was more optimistic than Ricardo or Malthus about the prospects for economic growth, being less concerned that the working class would erode gains in the standard of living by heedlessly multiplying. Like Say, she placed more emphasis on utility than labour cost as a source of value: when Malthus, after high praise of her discussion of rent, protested that ‘I think you have given too much sanction to Mr. Say's opinion reflecting utility’, she cut out and discarded the rest of his letter (Polkinghorn, 1986). A talented educational writer and the first woman to expound the principles of economics, Jane Marcet succeeded in bringing classical political economy (and other disciplines such as chemistry, botany and mineralogy) to a wider public.

See Also

- British classical economics
- Corn Laws, free trade and protectionism
- Malthus, Thomas Robert
- Martineau, Harriet
- Ricardo, David
- Say, Jean-Baptiste

Selected works

1816. *Conversations on Political Economy*. 3rd edn. London: Longman, 1818.

1833. *John Hopkins's Notions on Political Economy*. London: Longman.

1851. *Rich and Poor*. London: Longman.

Bibliography

Bodkin, R.G. 1999. The issue of female agency in classical economic thought: Jane Marcet, Harriet Martineau, and the men. *Gender Issues* 17, 62–73.

Polkinghorn, B. 1986. An unpublished letter from Malthus to Jane Marcet. *American Economic Review* 76, 845–7.

Polkinghorn, B. 1993. *Jane Marcet: An Uncommon Woman*. Aldermaston, UK: Forestwood Publications.

Shackleton, J.R. 1990. Jane Marcet and Harriet Martineau: pioneers of economic education. *History of Education* 19, 283–97.

How to cite this article

Dimand, Robert W. and Evelyn L. Forget. "Marcet, Jane Haldimand (1769–1858)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000415> doi:10.1057/9780230226203.1022

The New Palgrave Dictionary of Economics Online

Marget, Arthur William (1899 – 1962)

Donald A. Walker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cash-balance approach; Keynesian Revolution; Marget, A. W.; quantity theory of money; velocity of circulation

Article

A leading monetary theorist during the first half of his career, Marget went on to make an even greater contribution by formulating and implementing government policies regarding international banking and finance. Born in Chelsea, Massachusetts, on 17 October 1899, he graduated from Harvard with AB (1920) and MA (1921) degrees in Semitics, and a Ph.D. in economics (1927). He taught economics at Harvard (1920–7) and the University of Minnesota (1927–43; resigned 1948). He died in Guatemala City on 5 September 1962.

As an academician, Marget's principal concern was with the central problems of monetary theory, and since these had been so strikingly shaped by John Maynard Keynes, much of Marget's work became a critique of his views. Marget regarded himself as building upon an enduring neoclassical tradition, and saw the Keynesian Revolution as a largely misdirected episode that had the merit, however, of making some genuine contributions, and especially of stimulating the sort of re-examinations and refinements of doctrine exemplified by his own writings. His most significant critical contributions were his evaluations of Keynes's *Treatise on Money*, of liquidity preference, of Keynes's treatment of expectations, and of the implications of Keynes's *General Theory of Employment, Interest and Money* for the theory of prices (Marget, 1938; 1942). Marget's principal positive contributions were an extension and refinement of the concepts of the velocity of circulation of money and of goods; his reformulation of the quantity equation relating prices, output and money; his argument that the cash-balance approach is useful only in connection with the analysis of changes in the velocity of the circulation of money; and his analysis of the relevance of particular demand curves and their elasticity to the structure of money prices (Marget, 1938; 1942). The valuable elements of his critique of 20th-century theory and of his constructive writings have been assimilated into the discipline and are no longer a focus of discussion. Marget also undertook some studies in the history of thought which are among the best work on the subjects with

which he dealt. Particularly worthy of note are his examinations of the monetary theory of 19th-century neoclassical economists (Marget, 1931; 1935; 1938; 1942).

As an applied economist, Marget was concerned with international financial policies. While a major (1943–5) and a lieutenant colonel (1945) in the US Army, he devoted himself to preparations to bring about the economic and financial rehabilitation of Austria. He then became chief of the finance division of the US Allied Command for Austria (1946–9); a member of the US delegation in London that prepared for the treaty with Austria (1947); and a member of the US delegation to the Council of Foreign Ministers, which was charged with negotiating that treaty in London and Moscow (1947). His subsequent career included the positions of Chief of the Finance Division at the headquarters of the Marshall Plan in Paris (July 1948–December 1949), consultant to the US Treasury (1948), and Director of the Division of International Finance of the Federal Reserve Board of Governors (January 1950–April 1961). Among other activities, he represented the Board at meetings of the central banks of the western hemisphere in Bogotá (1956) and Guatemala City (1958). He then became the US representative to the Central American Common Market in Guatemala City and an adviser to the Common Market Bank in Honduras (April 1961–September 1962). In the latter roles he was instrumental in promoting the effectiveness of the Common Market policies.

Marget's scholarly work was distinguished by an insistence on logical clarity and an amassing of scholarly detail in the presentation of his expositions. His bureaucratic work was distinguished by an outstanding ability to suggest workable new financial institutions and procedures.

Selected works

1931. Léon Walras and the ‘cash-balance approach’ to the problem of the value of money. *Journal of Political Economy* 39, 569–600.

1932. The relation between the velocity of circulation of money and the velocity of circulation of goods. Parts I and II. *Journal of Political Economy* 40, 289–313; 477–512.

1935. The monetary aspects of the Walrasian system. *Journal of Political Economy* 43, 145–86.

1938, 1942. *The Theory of Prices: A Re-Examination of the Central Problems of Monetary Theory*. 2 vols. New York: Prentice-Hall.

How to cite this article

Walker, Donald A. "Marget, Arthur William (1899–1962)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000040> doi:10.1057/9780230226203.1023

The New Palgrave Dictionary of Economics Online

marginal and average cost pricing

William Vickrey
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Under perfect competition, marginal cost and average cost of a product are equal to each other and to its price, an arrangement that is Pareto-optimal in the absence of neighbourhood effects. Technical progress is making it possible to vary the prices of some products (such as telephony and electricity) from moment to moment in accordance with marginal cost. Such responsive pricing would help guarantee essential services and reduce the cost of providing reserve capacity. Where there are economies of scale, prices set at marginal cost will fail to cover total costs, thus requiring a subsidy.

Keywords

airline industry; average cost pricing; congestion charges; depreciation; economies of scale; electricity markets; escrow funds; excise taxes; inverse elasticity rule; land rents; land tax; leakage ratio; long run and short run; marginal and average cost pricing; non-price competition; ‘one-horse-shay’ asset; price discrimination; price stability; quantity–volume interrelationships; Ramsey, F.; reservation price; responsive pricing; second-best pricing; spatial economics; sticky prices; subsidy; Vickrey, W.S.

Article

In a pure and simple static world of perfect competition, where production units purchase or rent all their inputs in competitive markets and each sells a single homogeneous product competitively, production takes place at a point of constant returns to scale where the marginal cost and average cost of the product are equal to each other and to its price. If in addition there are no neighbourhood effects or externalities operating outside the market, the result will be Pareto efficient, meaning that there is no feasible alternative arrangement that would be better for someone and no worse for anyone.

Difficulties with the concept of average cost

As soon as production takes place with durable capital facilities that must be adapted to the needs of an individual firm there may no longer be an effective market for these facilities and a cost of their use during any particular period must be determined by other means. In the rather extreme case of the ‘one-horse-shay’ asset that in a static environment yields a stream of identical services over a known lifetime, a constant periodic rental cost can be derived by the use of a ‘sinking fund’ method of depreciation in which the rent is the sum of an increasing depreciation charge and a decreasing interest charge on the net value. But where the value of the service varies over time, whether because of physical deterioration, an increasing cost of maintenance needed to keep the item in ‘as new’ condition, or shifts in demand, this would in principle cause depreciation charges to vary; in practice this is done in one of a number of arbitrary ways by using ‘straight-line’ or various forms of ‘accelerated’ depreciation. If these charges are used as a basis for pricing, where competition is imperfect enough to give some leeway, the results can be correspondingly arbitrary.

More serious problems arise in the increasingly widespread cases of joint production of several distinguishable products or services. Where competitive markets exist, the market conditions dictate the allocation of joint costs among the various products, as when a meat-packing establishment produces steaks, hides, glue and offals. There is no way in which one can determine a meaningful average cost of hides by considering only the production process. Where the products, though economically widely different, are physically similar, it is tempting to cut the Gordian knot and average over the entire output, often at the cost of serious impairment of economic efficiency. Even when elaborate rationales are concocted by cost accountants, unless demand conditions as well as production conditions are taken into account the results are essentially arbitrary.

One can do a little better with marginal cost, at least if one is seeking a short-run marginal social cost (hereafter SRMSC), which is the concept that would be relevant for efficiency-promoting pricing decisions. Unless a consumer is presented with a price that correctly represents the marginal social cost associated with the various alternatives open to him, he is likely to make inefficient decisions.

The importance of emphasizing the short run

One often finds in the literature proposals to use a ‘long-run marginal cost’ as a basis for setting rates. The trouble is that in an operation producing a multitude of products with interrelated costs it is not possible even to define in any precise way what could be meant by a ‘long-run marginal cost’, any more than one could define a relevant long-run marginal cost for the hides and steak that are derived from the same carcass in the face of fluctuations over time in relative demand.

The attempt to use a long-run concept seems to be motivated in part by the notion that in some sense the long-run concept is more inclusive in that it allows for variation in capital investment and would include a return on such investment, whereas short-run marginal costs would fail to cover the costs of capital investment. In the single-product steady-state case, however, which is the only case for which the long-run marginal cost can be clearly defined, if the investment in plant is at the optimal level, i.e. the level which will result in the given output being produced at the lowest total cost, short- and long-run average cost curves will be tangent to each other at the given output, and short- and long-term marginal costs will be equal. Short run marginal-cost prices will therefore cover just as much of the total cost as will prices based on ‘long-run marginal cost’. If short-run marginal cost is below the long-run marginal cost, this would indicate that the installed plant is larger than optimum, and conversely if plant is below

optimum size, short-run marginal cost will be above long-run marginal cost.

Flexible versus stable prices

A long-run approach is sometimes advocated on the ground that it results in more stable prices. Price rigidity, however, exacts a high toll in terms of reduced efficiency. It is sometimes argued that stable prices are required for intelligent planning for installations that commit the investor to the use of a given volume of service. There is nothing in a SRMSC pricing policy, however, that precludes providing the consumer with estimates of the probable course of prices in the longer term, or even entering into long-term contracts to purchase specified quantities of service. If they are not to interfere with efficiency, however, such contracts should allow for the possibility of purchasing additional amounts at the eventual going rates, or of selling back some of the contracted-for output if this should prove profitable for the consumer.

Lack of flexibility in pricing has, indeed, been a major source of inefficiency in the use of utility services, whether arising as a result of the cumbersomeness of the regulatory procedures in privately owned utilities, or of bureaucratic inertia in publicly owned ones. At times it has even appeared that it takes longer to carry out the bureaucratic procedures involved in altering a price than to install additional capacity, whereas in terms of the underlying capabilities prices can and should be altered on shorter notice than the time taken to adjust fixed capital installations.

Optimal decision-making sequence

The efficient pattern of decision-making consists of first establishing a pricing policy to be followed in the future (as distinct from the application of that policy to produce a specific set of prices), then planning adjustments to fixed capital installations according to a cost-benefit analysis based on predicted demand patterns and predicted application of the pricing policy, subject to whatever financial constraints may be applicable, and then eventually determining prices on a day-to-day or month-to-month basis in terms of conditions as they actually develop.

Too often a rigid adherence to inappropriate financial constraints results in a pattern of pricing over time that leads to gross inefficiency in the utilization of facilities that are added in large increments. In the setting of tolls on bridges, for example, a high fixed toll is often imposed from the start in an attempt to minimize early shortfalls of revenues below interest and amortization charges. When the indebtedness incurred to finance the facility is finally paid off, tolls are often eliminated, sometimes just at the time that they should be increased in order to check the growth of traffic and congestion and defer the necessity for the construction of additional facilities.

The forward-looking character of marginal cost

Since changes in present usage cannot affect costs incurred or irrevocably committed to in the past, it is only present and future costs that are of concern in the determination of marginal cost. Past recorded costs are relevant only as predictors of what current and future costs will turn out to be. The marginal cost of ten gallons of gasoline pumped into a car is not determined by what the service station paid for

that gasoline, but by the cost expected to be incurred to replace that gasoline at the next delivery. The substantial time-lag that often exists between a change in price at the raw material level and its reflection at the retail level is one of the pervasive failings that contributes to the inefficiency of the economic system.

Another more important case in which future impacts are of vital importance in the calculation of marginal cost is where congestion accumulates a backlog of demand that has to be worked off over a period of time. A particularly striking case of this occurs when traffic regularly accumulates in a queue during rush hours at a bottleneck such as a toll bridge. The consequence of adding a car to the traffic stream is that there will be one more car waiting in the queue from the time the car joins the queue until the queue is eventually worked off, assuming that the flow through the bottleneck will be unaffected by the lengthening of the queue.

The marginal cost of a vehicle trip will be measured in terms of a number of vehicle hours of delay equal to the interval from the time the car would have arrived at the choke point if there had been no delay, to the time the queue is finally worked off. This is not measured by the length of the queue at the moment, but will be determined by the subsequent arrival of traffic over an extended period. A car arriving at the queue after it began to accumulate at 7:30 may get through the bottleneck at 8:00, after being delayed by only 15•minutes, but if the bottleneck will not be worked off until 10:00 the marginal cost will be $2\frac{1}{4}$ vehicle hours of which only $\frac{1}{4}$ hour is borne by the added car itself. The remaining two hours, if evaluated at $5 per vehicle hour, would indicate that under these conditions the toll that would represent this externality would be $10. Marginal cost cannot be determined exclusively from conditions at the moment, but may well depend, often to an important extent, on predictions as to what the impact of current consumption will be on conditions some distance into the future.

Marginal cost of heterogeneous sets of uses

It will often happen, for various reasons, that the same price will have to be applied to a non-homogeneous set of uses. To set such a price properly, the marginal costs of the various uses within the set covered must be combined in some way to get a marginal cost relevant to this decision. It would be wrong, however, merely to average the marginal costs of all the uses for which this price is to be charged. Rather, the decision as to whether a decrease in a given price is desirable must consider the cost of the increments or decrements in the various outputs that will be bought as a result of the price change. In averaging the marginal costs of the various usage categories, the weighting will have to be in proportion to the responsiveness of each usage category to the change in price.

For example, if a price is to be set for electricity consumption on summer weekday afternoons, in a system where air-conditioning is an important load, consumption and marginal cost may be higher on hot days than on warm days, but it may be considered too difficult to differentiate in price between the two categories of days. An increase in the price for this entire set of periods may induce some customers to adjust the thermostat setting. But during hot days the equipment may work full tilt without reducing the temperature to the thermostat setting, whereas on warm days there will be a reduction in power consumption. The marginal cost relevant to the setting of the common price would then be determined predominantly by the lower marginal cost of the warm-day consumption, and relatively little, if at all, by the higher marginal cost hot-day consumption.

Anticipatory marginal cost

In many cases a customer will make his effective decision to consume an item some time in advance, and it will be the expected price as perceived by him at that time that determines his decision. If, as in services subject to reservation, a firm price must be quoted the time the reservation is made, it is the expected marginal cost as of that moment that should govern the price charged. In the case of a service where the demand is highly variable and to a considerable extent unpredictable, such an expected marginal cost would be an average of marginal costs that might arise under alternative possible developments, possibly ranging from a very low value, if there turns out to be unused capacity, to the possibly quite high value if another latecomer must be turned away. The respective probabilities of these outcomes, as estimated at a given time, will vary with the proportion of the total supply already sold, the time remaining to the delivery of the service, and the pricing policy to be followed in the interim.
At one extreme, for long-haul airline reservations where the unit of sale is large, one might find it worth while to have a fairly elaborate pricing scheme in which the price quoted would vary according to the proportion of seats on a given flight already sold and the time remaining to departure, in simulation of what an ideal speculators' market might produce, the price at any time being an estimate of the price which, if maintained thereafter, would result in all the remaining seats being just sold out at departure time. This would correspond to marginal cost in that the sale of a seat at any given time would slightly raise the price during the remaining period to decrease demand by one unit, at a price that would be expected to be on the average equal to the price at which the seat was sold, indicating that the price was equal to the value of the seat to the alternative passenger.

Quality-volume interrelationships

In principle, in the absence of barriers to entry, competition would induce the supply of just sufficient seats on the various routes to cause revenues produced by such pricing to just cover costs. Even this, however, would be optimal only on those routes where traffic is so heavy that even with planes of a size producing the lowest cost per seat, further increases in service frequency would be of negligible value. On most routes there will remain economies of scale in that either providing more seats at the same frequency of service with larger planes would reduce costs per seat, or providing more seats with the same size of planes would provide an increased frequency of service that would be of value to others than the additional riders. In the latter case the marginal cost of providing for the additional passengers would be calculated by deducting the increase in the value of the service by reason of increased frequency from the cost of providing the added seats.
If it were possible to adjust plane size and frequency in a continuous fashion, then if the situation is optimal the two marginal costs would be equal. In practice both plane size and service frequency can be varied only in discrete jumps, so that this relation would be only approximate. Optimal price would be above a downward marginal cost calculated on the basis of a reduction in service, and below an upward marginal cost calculated on the basis of an increase in service. The decreases and increases might involve a combination of frequency and plane size changes. To preserve the formulation that price should equal marginal cost it may be useful to define marginal cost in such cases as consisting of the range between these upward and downward values rather than as a single point.

In practice, between the existence of economies of scale and the imperfect cross-elasticity of demand between flights at various times and with different amenities, removal of regulation tends to result in an emphasis on non-price competition, attempts to subdivide the market by various devices and restrictions to permit discriminatory pricing, and a bunching of service schedules at salient times and places that provides a lower overall level of convenience than would be possible were the given number of seat miles distributed more efficiently.

Where the unit of sale is small it may not be worth while to incur the transaction costs of varying price in strict conformity with SRMSC. One could, in theory, apply the same principle to the sale of newspapers at a given outlet. The price of a newspaper would vary according to the number of unsold papers remaining and the time of day. This would result in less disappointment of customers having an urgent desire for a paper late in the day and encountering a sold-out condition, and fewer unsold papers returned. But unless some ingenious device can be found for executing such a programme at low transaction costs, it probably would not be considered worth while, even by the most sanguine advocate of marginal cost pricing.

Wear and tear, depreciation and marginal cost

Even in the absence of lumpiness or technological change, existing methods of charging for capital use often fail to give a proper evaluation of marginal cost. This is especially true where the useful life of a unit of equipment is determined more by amount of use than by lapse of time. In the extreme case of equipment that must be retired at the end of a given number of miles or hours of active service, or after the production of so many kwh of energy, and which, in one-horse-shay fashion, gives a uniform quality of service over its lifetime without requiring increasing levels of maintenance, the marginal cost of use at a given time will be the consequent advancing of the time of retirement of the equipment. The marginal cost of using the newest units will be the lowest, and will advance over time at a rate equal to the rate of interest as the equipment ages and the advancement of replacement consequent upon use becomes less and less remote.

In a service subject to daily and weekly peaks, the newest equipment will be allocated to the heaviest service, operating during both peak and off-peak hours. Equipment will be relegated to less and less intense service as it ages. The marginal cost of service at a particular moment will be that for the oldest unit that has to be pressed into service at that instant. The rental charge for the use of the unit will vary gradually over the entire range of demands, rather than dropping off to zero whenever the full complement of equipment is not required. At the other end, in this extreme case, the service provided would not necessarily be held constant by price variation over an extended peak period: under the conditions postulated it would be possible to provide for needle peaks by planning for the stretching out over time of the final service units of the oldest equipment. In this way the required peak capacity can be provided at a cost much lower than that which would be calculated by loading all the capital charges for the added equipment on this brief period of use.

Another way of looking at the matter is to appeal to the proposition that perfect competition under conditions of perfect foresight will produce optimal results. To this end one can suppose a situation in which vehicles are rented by the hour from a large number of lessors operating in a competitive market. For simplicity, initially, one can assume all vehicles to be of the one-horse-shay variety, being equivalent to bundles of hours of active service, with the quality of service being independent of age up

to a final ‘bubble-burst’ collapse. Also, for simplicity, assume a steady state in which vehicles are scrapped and replaced at a constant rate over time, so that at any given moment vehicles are evenly distributed by age.

A common market rental price for all vehicles at any given time of the week will emerge, being higher as the number of vehicles in service at the time is greater. During any given week, each renter will have a reservation price for his vehicle, such that he will rent his vehicle during those hours for which the market rental is above this reservation price and never when the market rate is lower. This reservation price will increase over time for any given vehicle at the market rate of interest, since a renter will rent his vehicle if and only if the net present value of the rental discounted back to the time of purchase exceeds some fixed amount. The owner would not want to rent his vehicle for a net present value less than he could have got by selling one of this stock of service units at some other time at or just below his reservation price. New buses will have the lowest reservation price and will be assigned to the schedules calling for the most hours of service per week, while old buses will be held idle during slack hours and used only for peak service. As each bus ages it will be assigned to less and less heavy service along the load-duration curve.

This pattern of usage can be regarded as resulting from a desire to recover the capital tied up in the usage units of each bus as rapidly as possible. It is related to the practice in electric utilities of using the newest units for peak service, in that case motivated in part by the tendency for the newer units to be more efficient in thermal terms. To be sure, occasionally new units are designed specifically for peaking service, with a correspondingly low capital cost, though this is a relatively recent phenomenon related to a slowing-down of secular increases in potential thermal efficiency.

In any case, where wear-and-tear is a factor, one cannot properly allocate depreciation charges primarily to peak service, however defined, nor should they be spread evenly over all service, much less spread evenly over hours of the week so that vehicle hours in off-peak periods would get higher charges than during the peak. Rather the depreciation charge per vehicle hour will vary gradually and in a positive direction with the intensity of use of the equipment at any given hour.

The analysis becomes a little complicated when equipment life is dependent on mileage or loading or intensity of use as well as hours of active service, so that different rentals would properly be chargeable according to the nature of the service for which the unit is being rented. Also further analysis is required if equipment is laid up between runs at isolated terminals rather than at a central depot where a market could be postulated, or if the fleet contains vehicles varying in size or other characteristics. It would even be theoretically appropriate to charge different fares for the same trip at the same time if made on vehicles with different origins or destinations. (In Hong Kong, indeed, the practice is to charge a flat fare on each route, but to differentiate the fare fairly elaborately as among routes. On segments where routes converge, this has the unfortunate result of unduly concentrating riding on buses with the lower fares, even where the higher fare buses have empty seats and are making stops in any case for other passengers.)

Costs of major overhauls that are performed at relatively long intervals would also complicate the picture. There are also problems associated with gradual or sudden changes in overall demand levels, or special events that can be anticipated sufficiently to present an opportunity for reacting in terms of a change in price. The picture can be further complicated if, as was discussed above, there are changes in available technologies or other changes in quality or cost. But the same method of analysis in terms of a hypothetical competitive market can be used to obtain appropriate results.

For the sake of simplicity the above analysis has been couched mainly in terms of a bus service, but the analysis is applicable wherever the useful life of equipment is in part a function of the intensity with which it is used.

Responsive pricing

In some cases, notably in telephone and electric power services, the technical possibility exists for conveying information as to the current price to customers at the instant of consumption, and for customers to respond to such information in a worthwhile manner at modest cost. In the case of telephone service the information as to the level of charges for local calls can be substituted for the dial tone, with information on rates for long distance calls provided to users who wait for it before dialling the final digits. If the charge exceeds what the customer is willing to pay the call can be aborted with little occupancy of equipment or inconvenience to the user. Prices can be varied from moment to moment in accordance with marginal cost, as estimated from the degree of busyness of the relevant sets of equipment.

In the case of electric power, the costs of providing for a variation of the price according to the conditions of the moment would be somewhat greater. But if the facilities take the form of remote meter reading, either by carrier current over the power lines or by a separate communications channel, much of the cost would be covered by the avoidance of costs involved in manual meter reading. A signal of rate changes can then be provided to the customer as a by-product of the signal required to initiate a new rate period. The customer can then respond either manually or by installing automatic equipment which will adjust the operation of such items as air-conditioning and refrigeration compressors, water heaters and the like, according to the level of rates in a manner determined by the customer himself. Retrofitting of existing meters by attaching a pulse-generating device such as a mirror and photo-electric cell to the rotor shaft of the existing meter and feeding the pulses to electronic counters and registers should be possible at relatively low cost.

Such responsive pricing would be especially valuable in dealing with emergencies, providing greater assurance of the maintenance of essential services than is possible with existing techniques, and making it possible to reduce substantially the cost of providing reserve capacity. In the case of floods, conflagrations, breakdowns in transit, or other emergencies that under present conditions tend to result in the overloading of telephone facilities and difficulties in completing calls of a vital nature, rates can be charged that are high enough to inhibit a sufficient number of less important calls so that the ability of the system to handle vital calls promptly is preserved. This is difficult to do with present techniques, for while it is relatively easy to give priority to calls originating at such points as police stations, hospitals, and the like, most emergency calls are calls to rather than from these points and it is much more difficult to distinguish such calls close to the point of origin. And there are always a certain number of vital calls not distinguishable in terms of either origin or destination.

Again, in the case of unscheduled power cuts, it would be possible to cause an almost instantaneous shedding of substantial water-heating and refrigeration loads, followed, in the case of an extended cut, by partial shedding of elevator, transit and batch process loads for which it is more inconvenient to respond quite so promptly, after which a sufficient refrigeration load can be picked up as needed to avoid food spoilage. Many of the serious consequences of major power blackouts could have been avoided had such a system been in place at the time. Reserve capacity might well be cut back to

provision for scheduled maintenance, leaving the load-shedding capability of responsive pricing to function as a reserve. In many cases the speed of response possible with responsive pricing would be faster than the reaction time within which reserve capacity can pick up load, leading to better voltage regulation and a higher quality of service to customers remaining on the line. And if, in spite of everything, areas must be cut off completely, responsive pricing would also be of considerable help in facilitating a smooth recovery from an outage: instead of having a whole army of motors trying to start up at once upon the restoration of power, with consequent load surges, voltage fluctuations, and malfunction of equipment, load could be picked up smoothly and gradually as the price is lowered from the inhibiting level.

Preserving incentives with escrow funds

With privately owned utilities the regulatory process is too slow to permit prices established directly by regulation to be constantly adjusted to changing current conditions, unless indeed the regulators were to assume a large part of what are normally the responsibilities of management. The problem thus arises of how to allow the prices to be paid by customers to be varied by the utility management without giving rise to incentives for behaviour contrary to the public interest. Even if a formula could be devised that would require the utility to adjust prices to track short-run marginal cost, if the utility were allowed to keep the revenues thus generated without restriction, this would set up undesirable incentives for the utility to skimp on the provision of capacity in order to drive up the marginal cost, price, revenues, and profits.

A resolution of this dilemma can be achieved by separating the revenue to be retained by the utility from the amounts to be paid by customers. We can have the ‘responsive’ prices paid by customers vary according to short-run marginal social cost, while the revenues to be retained by the utility are determined by a ‘standard’ price schedule fixed by regulation in the normal manner, the difference being paid into or out of an escrow fund. Failure of the utility to expand capacity adequately would drive marginal cost up, and with it the responsive price, causing revenues to flow into the escrow fund, but the only way the utility could draw on these funds would be to expand capacity sufficiently to drive marginal cost down, causing the responsive rate to fall below the standard rate on the average, entitling the utility to make up the difference from the escrow fund as long as it lasts. Excessive expansion would result in the escrow fund being exhausted, with a corresponding constraint on the revenues obtainable by the utility from the unaugmented low responsive rates.

The setting of the responsive rates would have to be to a large extent at the discretion of the operating utility, though the regulatory commission could monitor the process and even attempt to establish guidelines according to which the responsive price should be set. The utility would normally have no incentive to set the responsive rate below marginal cost, since this would merely increase sales and hence costs by more than any possible long-run increase in revenues to the utility. To be sure, in the short run it might be able to draw on the escrow fund to the extent of the excess, if any, of the standard rate over the responsive rate, but since from a long-run perspective there will normally be other more advantageous ways of drawing on this fund this will not be attractive.

When marginal cost is below the standard price, which would tend to be the usual situation, the utility would in general have an incentive to set the price between the marginal cost and the standard price, since each additional sale produced by the lower price will yield an immediate net revenue equal to the

difference between marginal cost and the standard price, offset only by the drawing down of the escrow fund by the difference between the responsive and the standard price. When marginal cost is above the standard price, which with a properly designed standard rate schedule with time-of-day variation should happen relatively rarely, the utility would have an incentive to set the price at least at the marginal cost level, since to set it lower would tend to increase output at a cost in excess of anything the utility could ever recover. How much higher than marginal cost the price might be set would in theory be limited by the condition that the price could not be high enough to curtail demand sufficiently to drive marginal cost below the standard price. If the standard price has an adequate time-of-day variation, this constraint, loose as it may seem, may be sufficient. Additional guidelines could of course be imposed by the regulatory commission for those rare occasions where this constraint might seem insufficient to keep prices within bounds.

Actual steps towards responsive pricing

Some actual practices of utility companies are steps in the direction of responsive pricing. Contracts for ‘interruptible’ power provide for load shedding at the discretion of the utility subject to some overall limits. As these are fairly long-term contracts that usually require ad hoc communication between the utility and the customer, their applicability is limited and there is no assurance that the necessary shedding will be done in the most economical manner. Many customers are reluctant to submit to load shedding that is not under their control at least to some extent, and that might be imposed under awkward circumstances. Where reserves are ample and interruption is highly unlikely, such contracts have been challenged as being a form of concealed discriminatory concession. On the other hand customers entering into such contracts in the expectation of not being interrupted may feel aggrieved if interruption actually takes place.

Another experimental provision applied by a company with a heavy summer air-conditioning load is for a special surcharge to be applied to the usage of larger customers on days when the temperature at some standard location exceeds a critical level. And another company bases its demand charge on the individual customer's demand recorded at the time that turns out to have been the monthly system peak load, supplying the customers with information as to moment-to-moment variations in the system load. This leads to interesting game-playing on the part of customers as they attempt to keep their own consumption down at times that look as though they might become the monthly peak, with the result that this action may itself shift the peak to another time.

Economies of scale, subsidy and second best pricing

Where there are economies of scale, prices set at marginal cost will fail to cover total costs, thus requiring a subsidy. One reason for wanting to avoid such a subsidy is that if an agency is considered eligible for a subsidy much of the pressure on management to operate efficiently will be lost and management effort will be diverted from controlling costs to pleading for an enhancement of the subsidy. This effect can be minimized by establishing the base for the subsidy in a manner as little susceptible as possible to untoward pressure from management. But it is unlikely that this can be as effective in preserving incentives for cost containment as a requirement that the operation be financially self-sustaining. To achieve this, prices must be raised above marginal cost, and in a multi-product

operation the question arises as to how these margins should vary from one price to another within the agency.

Another objection to subsidy is that it raises hard questions of who should bear the burden of the subsidy. More fundamentally the taxes imposed to provide the subsidy will often have distorting effects of their own, and minimizing the overall distortion would again require prices to be raised above marginal cost. One can, indeed, regard these excesses of price over marginal cost as excise taxes comparable to other excise taxes that might be levied to raise a specified amount of revenue.

The answer given to the problem of how to allocate excise taxes and other margins of price above marginal cost so as to minimize the overall loss of economic efficiency given by Frank Ramsey in 1927 can be expressed for the case of independent demands as the inverse elasticity rule, which says that the margin of price over marginal cost as a percentage of the price shall be inversely proportional to the elasticity of demand. A more general formulation is one that states that prices shall be such that consumption of the various services would be decreased by a uniform percentage from that which would have been consumed if price had been set at marginal cost and demand had been a linear extrapolation from the neighbourhood of the ‘second-best’ point.

A more transparent formulation, devised by Bernard Sobin in work for the US Postal Service, is the requirement of a uniform ‘leakage ratio’, leakage being the difference between the net revenue actually derivable from a small increment in a particular price and the hypothetical revenue that would have been obtained had there been no change in consumption as a result of this increment. Leakage is the algebraic sum of the products of the changes in consumption of the various related products induced by the small change in a given price, and the respective margins between their prices and marginal costs. Leakage is a measure of the loss of efficiency resulting from the change in the particular price, and the leakage ratio is the ratio of this loss of efficiency to the hypothetical gain in gross revenue if there had been no change in consumption. If one leakage ratio should be greater than another, the same net revenue could be obtained at greater economic efficiency by getting more revenue from the price with the smaller leakage ratio and less from the other. The second-best solution accordingly requires that all leakage ratios be equal.

This analysis can be extended to the case where the agency is being subsidized by taxes which involve an adverse impact on the economy, in terms of marginal distorting effects, compliance costs, and collection costs, which can be expressed as the ‘marginal cost of public funds’ (MCPF). For a net decrease in the subsidy derived by increasing a price, which can be considered to be equivalent to imposing a tax equal to the difference between the marginal cost and the price, $MCPF = LR / (1 - LR)$, where LR is the leakage ratio. A second-best optimum is then one where the MCPF's are equalized over both external and internal taxes.

Special sources of subsidy: land rents and congestion charges

In the case of goods and services with economies of scale that are provided primarily to consumers within a particular urbanized area, methods of financing may be available that involve no marginal cost of public funds or even result in an enhancement of efficiency. The existence of large cities, indeed, is to a predominant extent due to the availability in the city of goods and services produced under conditions of economies of scale: if there were no economies of scale, activity could be scattered about the landscape in hamlets, with great reduction in the high transportation costs involved in movement about a

large city. If prices of these services are reduced to marginal cost, the increased attractiveness of the city as a consequence would tend to drive up land rents within the city, and it appears quite appropriate that a levy on such rents should be used to finance the required subsidies. And while there are practical and conceptual difficulties in defining exactly how land rents or land values should be specified for purposes of levying a tax, it is generally considered that a tax on land values, properly defined, has negligible adverse impacts on the efficient allocation of resources.

Indeed, there is a theorem of spatial economics which states that in a system of perfect competition among cities, the availability in the city of services and products subject to economies of scale, priced at their respective marginal social costs, will generate land rents just sufficient to supply the subsidies required to permit prices to be lowered to marginal cost. Among the more important of these services are utility services such as electric power, telephone, cable communications, water supply, mail collection and delivery, sewers and waste disposal, and local transit. It is not clear just how broad the conditions are under which this theorem would hold, and there are difficulties in capturing all land rents for subsidy purposes, but steps in this direction are clearly desirable.

On a more intuitive level, one can note that a person who occupies or uses land that is provided with services such as the availability of transit, electricity, telephone, mail delivery and the like will be requiring that these services be carried past his property to serve others whether or not he himself uses them. The user of tennis courts located conveniently in a built-up area should no more be excused from contributing to the costs of carrying these services past the courts, even though no direct use is made of electric power, telephone, mail, or other services, than he should expect his auto dealer to cut the price of an automobile by the cost of the headlights and windshield wipers merely because he asserts that he will never drive at night or in bad weather. Tennis players will indeed pay a rent enhanced by the presence of these services and the consequent greater demand for the land for other purposes, but the rent will go to the landlord, not to the purveyor of the services, and the price of the services to those who do use them will be too high for efficiency, unless indeed they are subsidized by other taxes that have their own distorting effects.

It is a corollary of this theorem that it would be to the advantage of the landlords in the area, *faute de mieux*, to agree collectively to pay a tax based on their land values, in order to subsidize the various utility services to enable the prices to be set closer to marginal social cost. They could expect in the long run that this action would increase their rents by as much or more than the taxes. To be sure, they might do better by getting someone else to pull their chestnuts out of the fire, but they can do this only at considerable damage to the overall efficiency of the economy of the city, to say nothing of the inequity of such a parasitic relationship.

In addition to land rents in the conventional sense, there is the land used for city streets for the use of which no adequate rental is generally charged. Charging on the basis of SRMSC for the use of congested city streets would in most cases yield a revenue far in excess of the cost of maintaining such facilities, which could appropriately be used for the subsidy of other urban facilities. Properly adjusted, such charges would increase efficiency by bringing home to the users the costs that their use directly imposes on others.

Formerly it would have been considered impractical to attempt to charge for the use of city streets according to the amount of congestion caused: the collection of tolls by manual methods at a multitude of points within the city might well create more congestion than it averted. Advances in technology have, however, made it possible to do this at minimal interference with traffic flow and at modest cost.

One method, proposed as long ago as 1959 and recently carried to the point, is to require all vehicles using the congested facilities to be equipped with electronic response units which will permit individual vehicles to be identified as they pass scanning stations suitably distributed within and around the congested area so that the records thus generated can be processed by computer and appropriate bills sent to the registered owners at convenient intervals. If properly done, this would greatly improve traffic conditions so that the net cost of the revenue to the road users would be far less than the amount collected as revenue. A pilot installation has recently been tested in Hong Kong with satisfactory results, but full implementation appears to have been deferred, because of the political situation associated with the impending transfer of sovereignty.

Indeed, one can define ‘hypercongestion’ as a condition where so many cars are attempting to move in a given area that fewer vehicle miles of travel are being accomplished than could be if fewer vehicles were in the area but could move more rapidly; for example if 1000 vehicles in an area move at 8 mph and produce 8000 vehicle miles of travel per hour, reducing the number of cars in the area at a given time to 800 might raise speed to 11 mph producing 8800 vehicle miles of travel per hour. By restricting the flow of traffic in the period leading up to the hypercongestion period, road pricing could prevent hypercongestion from occurring, except possibly sporadically, and in any case so improve conditions that more movement would be accomplished during the peak period at faster speeds. The improvement during peak periods might even be such that total movement throughout the day would be increased, and where conditions are now severe users could find that they are better off than before, even inclusive of the payment of the congestion charge.

If there are bridges, tunnels, or other special facilities for which a toll is already being charged, and which regularly back up a queue during the morning rush-hour, substantial revenues can be obtained at no overall net cost to the users by adding a surcharge to the toll during the period where queueing regularly threatens, rising gradually from zero to a maximum and down again in such a way that by gradual adjustment regular queueing is substantially eliminated. The toll surcharge will then be taking the place of the queue in influencing decisions as to when to travel, and in general those who plan their trips in terms of time of arrival at their destination will be able to leave as many minutes later as they formerly wasted in the queue, pass the bottleneck at the same time as before, and arrive at their destinations at the same time as before. The extra toll will be roughly the equivalent of the value of the extra time enjoyed at the origin point, and the revenue will in effect be obtained at no net burden on the users. In practice the results may be even better than this as a result of the added encouragement to car-pooling, the reduction of obstruction to cross-traffic, and the expediting of emergency or other trips where the delay had been a particularly serious matter.

Gains in the evening may be not quite so dramatic. The situation is not symmetrical, as typically the timing of the trip will be determined in terms of time of departure, which is separated by the queue from the time at the bottleneck. On the other hand the risk of conditions approaching gridlock is greater, since the accumulation of queues inside circumferential bottlenecks is more likely to create congestion, and there is less of a physical barrier to the simultaneous emergence of large quantities of traffic from parking lots into the downtown streets than there is in the morning to the convergence on the congested area of traffic arriving from the outside.

Congestion charges should be imposed, at least notionally, without exception on all forms of traffic. Such charges would be a necessary element in the cost-benefit analysis by which decisions are made as to the level and pattern of bus service to be provided, even though they would not be directly relevant to

the determination of the price structure to be applied to that service.

Paradoxes in the behaviour of marginal social cost

A strict calculation of marginal social cost in particular circumstances may produce what may appear to be quite paradoxical results. For example, in many circumstances it will be optimal, and even essential, to maintain at least a minimum frequency of service in off-peak hours with buses of a standard size, resulting in there being practically always a large number of empty seats in each bus. Under these circumstances the cost of carrying additional passengers is predominantly the cost of boarding and alighting, including the time of the driver and the other passengers on the bus who are delayed in the process. This cost will be relatively higher if the bus is half full than if it is nearly empty. The result is likely to be that the cost of a trip from a point near one end of the run to a point near the other end, at both of which points the bus is likely to be lightly loaded, may be smaller than for a shorter trip between points near the middle of the run where the bus is likely to be more heavily loaded. This is not a trivial matter: if it were there would be no sense to the refusal of express buses with empty seats to pick up local passengers. It is highly unlikely, however, that fares based on such a seemingly perverse behaviour of cost would meet with popular approval. Indeed, the original US interstate commerce legislation contained prohibitions against higher rates being charged shorter hauls than for any longer hauls within which they might be included.

Another paradoxical example can occur in mixed hydrothermal electric power systems: an increase in fuel prices could result in the marginal cost of power at particular times being reduced rather than increased. If hydro dams are spilling water at certain seasons of the year, increased fuel costs may make it economical to increase the installed generating capacity to make use of the spilling water, even for a briefer period of time over the year than was previously worth while. If during the wet season installed hydro generating capacity is more than sufficient to meet trough demand, marginal cost during such periods will be substantially zero, or at most limited to a small element of wear and tear on equipment pressed into service. Installing more turbo-generators would expand the period during which this low marginal cost is effective, so that while increased fuel costs cause marginal cost to rise during the peak, the result could also be to lower marginal cost in these intervals into which the period of exclusive hydro supply expands.

In the case of long distance telephone service, the drastic reductions in the cost of bulk line-haul transmission have created a situation where distance, especially beyond the range where separate wire transmission is economical, is relatively unimportant as a cost factor, and where satellite transmission is involved, ground distance is indeed irrelevant. What remains important is the number of successive circuits, with their associated termination and switching equipment, involved in the making of a call. Thus a call between two small communities over a moderate distance, for which the volume of calling is insufficient to warrant the provision of a separate circuit, will generally cost substantially more than a call between important centres over a much longer distance, since the latter will involve only a single long-haul circuit, while the former will require patching through two or more long-haul circuits.

Another anomaly occurs when an innovation promising substantial reductions in costs appears on the horizon, such as has happened repeatedly in telecommunications. Any further installation of the old technology in the interim before the new technology is actually available will involve an investment which will have its capital value diminished over a brief period to that determined by its competition

with the new technology. High depreciation or obsolescence charges are in order, and the prospect of the new lower costs results in higher current prices which would serve to hold back current demand and lessen the amount of old technology required to be installed.

Marginal cost pricing is thus not a matter of merely lowering the general level of prices with the aid of a subsidy; with or without subsidy it calls for drastic restructuring of pricing practices, with opportunities for very substantial improvements in efficiency at critical points.

See Also

- congestion
- ideal output

Bibliography

Beckwith, B.P. 1955. *Marginal Cost Price-Output Control*. New York: Columbia University Press.

Mitchell, M., Manning, G. and Acton, J.P. 1978. *Peak-Load Pricing*. Cambridge, MA: Ballinger.

Nelson, J.R. (ed.) 1964. *Marginal Cost Pricing in Practice*. Englewood Cliffs, NJ: Prentice-Hall.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Vickrey, W. 1967. Optimization of traffic and facilities. *Journal of Transport Economics and Policy* 1 (2), 1–14.

Vickrey, W. 1969. Congestion theory and transport investment. *American Economic Review* 59, 251–60.

Vickrey, W. 1970. The city as a firm. In *The Economics of Public Services*, ed M.S. Feldstein and R.F. Inman, Proceedings of a conference held by the International Economic Association, Turin, Italy. London: Macmillan; New York: Wiley, 334–43.

Vickrey, W. 1971. Responsive pricing of public utility services. *Bell Journal of Economics and Management Science* 2(1), 337–46.

How to cite this article

Vickrey, William. "marginal and average cost pricing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000043> doi:10.1057/9780230226203.1024

The New Palgrave Dictionary of Economics Online

marginal productivity theory

Robert Dorfman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Marginal productivity theory is an approach to explaining the rewards received by the various factors or resources that cooperate in production. Broadly stated, it holds that the wage or other payment for the services of a unit of a factor is equal to the decrease in the value of commodities produced that would result if any unit of that factor were withdrawn from the productive process, the amounts of all other factors remaining the same.

The basic justification of this assertion is highly intuitive. It rests on three assumptions: that the product is sold and the factor services are purchased in competitive markets; that the firms in those markets operate so as to maximize their profits; and that the products sold are produced by technologies that satisfy the ‘law of variable proportions’, which holds that successive equal increments of one factor of production, the amounts of all other factors remaining unchanged, will yield successively smaller increments of physical output. It follows immediately from these assumptions that if the wage of any factor exceeds the value of the output that would be lost if a unit less of that factor were employed, then a unit less of that factor will be employed, and successive units will be released until the inequality is annihilated. Similarly, if the wage of any factor is less than the value of the output that an additional unit could produce, successive units of that factor will be employed until the inequality vanishes.

The motivating concept in the foregoing argument was the effect on the value of output of small changes in the quantities used of different factors of production. This idea is so important that a special vocabulary has developed in order to discuss it with precision. The marginal product of a factor of production is the ratio of the greatest change in the output of some product that can be obtained by a small change in the use of the factor to the change in the use of the factor. The marginal product multiplied by the price of the product is the value of the marginal product. Marginal productivity theory holds that the payment for any factor of production tends to be about equal to the value of its marginal product, where, in a multiproduct firm, the product used in the calculation is the one for which the value of marginal product is greatest.

Clearly, the marginal product and its value may depend on the size of the ‘small’ change in the amount of the factor that is used in the calculation. To avoid being ambiguous when the amount of the factor is a continuous variable, the concept of marginal productivity is used: the marginal productivity of a factor is the limit that its marginal product approaches as the change in the quantity of the factor approaches zero.

The result of multiplying the marginal productivity by the price of the product is called, somewhat inaccurately, the value of the marginal product; confusion rarely results.

Two of the assumptions made above to justify the marginal productivity doctrine can be relaxed. First, if the assumption that the firm produces for a competitive market is dropped, the conclusions of the theorem has to be weakened slightly. If a firm produces for a market that is not perfectly competitive, it will recognize that it cannot change the quantity of any of the commodities it sells without simultaneously changing the price. Consequently, it will take account of the fact that if it changes the amount used of any factor, the resulting change in sales revenue will not equal the value of the factor's marginal product, but that value adjusted for the induced change in price. The ratio of the change in sales revenue to the change in the employment of a factor, for ‘small’ changes, is called the factor's marginal revenue product. Then the reasoning used to deduce the marginal productivity doctrine leads to the conclusion that the firm will employ each factor at the level where its marginal revenue product equals its rate of pay, whether or not the firm sells in a competitive market. In competitive markets, the marginal revenue product of a factor equals the value of its marginal product, but not necessarily in other market types.

The assumption that firms operate so as to maximize their profits can also be weakened for some purposes. If the firm operates only so as to produce its outputs at the lowest possible total cost, the same line of argument shows that the rates of pay for any two factors used by the firm will be proportional to the marginal revenue products of the factors. This is a weaker conclusion than was found for profit-maximizing firms, and does not imply any particular relationship between a factor's rate of pay and its marginal revenue product or the value of its marginal product.

Development of the concept

Simple as it may appear, the marginal productivity principle was seen clearly only after a long, slow evolution. It was first presented in essentially its modern form around 1890, by J.B. Clark and Alfred Marshall, who apparently arrived at it independently. Their formulation built on the work of numerous predecessors, each of whom saw an important aspect of the principle but did not perceive its full generality.

The problem that gave rise to the marginal productivity principle – to explain the distribution of the national income among the great social classes and, especially to explain the shares claimed by the owners of capital and land – was at the top of the agenda of 19th-century economics. Thus, originally, only three very broad factors of production were considered: land, labour and capital, corresponding to the three social classes.

The first application of the principle occurred in the Malthus–Ricardo theory of rent, in particular in the concept of the intensive margin, which held that doses of labour and capital (in unspecified proportions) would be applied to each parcel of land until the value of the increase in product equalled the cost of the dose. The separate rewards to labour and capital were explained on other grounds.

In 1833, Longfield argued that the rate of interest was governed by the earnings of the least productive unit of capital, using a marginal argument. But he did not extend the reasoning to wages. At around the same time, von Thünen applied the principle to both wages and interest but did not publish his findings until much later, and then so obscurely that they had no influence. Jevons, in 1871, accounted for the rate of interest by a marginal argument, but explained wages as a residual after rent and interest were

paid. Indeed, Jevons's theory is remarkably similar to Longfield's.

The ingredient that all these applications of the marginal principle missed was that the equality of marginal product and factor reward applied to all factors. Walras in 1874 (and, indeed, J.-B. Say three-quarters of a century before) insisted on treating the various factors of production symmetrically, but he did not derive any of the factor shares by a marginal argument until the later editions of the *Elements*, and then only awkwardly. Thus the marginal insight was not applied symmetrically to all factors until Clark published the papers that led to his *The Distribution of Wealth* (1899), and Marshall published his *Principles of Economics* (1890), thereby introducing a unified theory of income distribution.

The achievement of the unified theory raised a puzzling question: if each unit of every factor was paid the value of that factor's marginal product, would the total value produced be neither more nor less than just sufficient to make all the factor payments? In 1894 Philip Wicksteed showed that the answer is affirmative for production processes with constant returns to scale, thus establishing the internal consistency of the marginal productivity principle. (Clark had believed it all along, but on inadequate grounds.) Wicksteed's proof amounted to an independent rediscovery of part of Euler's Theorem for Homogeneous Functions.

Beginning with the late 1880s, when the various partial glimpses of the doctrine congealed, marginal productivity theory became an essential part of the accepted explanation of the general level of wages and of the rate of interest, with important implications for practical economic issues. For example, it is often held that unions are powerless to raise the average level of wages because wages are governed by the marginal productivity of the labour force, which union activity cannot affect.

Although the marginal productivity concept was originally applied to explain the rewards of the broad social classes – the workers, landowners and entrepreneur–capitalists – beginning with Walras it became absorbed into the general theory of production and value. In that context it is used to explain the payments for the services of all the classes of factors that enter into production, and the definitions of these classes can be chosen freely to fit the problem under study. The tripartite classification continues to be used frequently, however.

Qualifications

The marginal productivity doctrine does not purport to be a complete explanation, even in principle, of the payments received by factors of production. As the simple, basic argument indicated, it explains only the amount of each factor that an enterprise will employ at different rates of payment for its services and in the presence of given quantities of the other factors used; that is, the demand curves for the factors. Supply curves also are needed to complete the explanation of the equilibrium level of use and rate of payment for the factors.

Furthermore, especially in the version that deals with numerous factors, rather than just two, a high degree of simultaneity arises. The demand curve for each factor depends on the amounts used of the other factors, but those amounts, depend on the amount used of the first factor, so, in the end, the rates of payment and the quantities used of all the factors are determined simultaneously. Consequently, the rates of payment for the various factors cannot be explained except in the context of a full-fledged general equilibrium model. Still, in such a model it often turns out that the payment received by each factor corresponds to its marginal productivity in each productive process in which it is used and in which marginal productivity is a well-defined concept. These complications will be clarified by considering a

more formal and rigorous derivation and statement of the principle.

Formal derivation of the marginal productivity thesis

The theory is based on the behaviour of a profit-maximizing firm in a competitive industry. To describe that behaviour, imagine a firm that produces *m* products by the use of *n* factors or inputs. Suppose that the price of the *i*th product is p_i and that the quantity produced (per year) is y_i. Then the gross revenues per year will be $R = \Sigma p_i y_i$. Similarly, let w_j be the price per unit of the *j*th factor used. If the *j*th factor is a kind of labour or a purchased input w_j is simply its price or wage, but if it is a kind of fixed capital, then w_j should be regarded as its rental cost, normally interest on its purchase or construction cost plus a depreciation allowance. The amount of the *j*th factor used will be denoted by x_j. Then the total cost incurred per year will be $C = \Sigma w_j x_j$. The profit that the firm seeks to maximize is *R*–*C*.
The quantities (per year) of output, y_i and input x_j, cannot be chosen freely. This basic presentation will be limited to the simplest situation, in which the choices are constrained only by an explicit, differentiable production function, which will be written $g(y_1, \ldots, y_m, x_1, \ldots, x_n) = c$. The implicit constraint that none of the arguments of *g*(.,.,…) can be negative has important consequences that will be noted below.
In this set-up, invoking the assumptions mentioned in the second paragraph of this entry, the necessary conditions for a choice of *y* and *x* to maximize the firm's profits are the familiar marginal equalities. Specifically:

1. (1) Marginal rates of substitution. The marginal rate of substitution between two factors, say the *j*th and the *k*th is the rate at which small amounts of the *j*th factor can be substituted for the *k*th with no effect on the rates of output in accordance with the production function constraint. Denote it by $\delta x_j / \delta x_k$. Mathematically, it is the ratio of the partial derivatives of the production function, or $\delta x_j / \delta x_k = -(\partial g / \partial x_k) / (\partial g / \partial x_k)$. Economically, when the firm's profits are being maximized $\delta x_j / \delta x_k = w_k / w_j$. The intuitive content is clearest when the maximizing condition is written as $w_j \delta x_j = w_k \delta x_k$, which requires that when profits are being maximized the amounts of factors that can be substituted for each other in accordance with the production function must have equal monetary value.
2. (2) Marginal rates of transformation. There is a similar relationship among the rates at which the outputs can be ‘transformed’ into one another in accordance with the production function constraint. Let $\delta y_i / \delta y_k$ denote the ratio at which production of the *i*th output can replace production of the *k*th. Mathematically, $\delta y_i / \delta y_k = -(\partial g / \partial y_k) / (\partial g / \partial y_i)$. Economically, when profits are being maximized $\delta y_i / \delta y_k = p_k / p_i$. Again, this result asserts that when profits are being maximized a small quantity of one of the outputs can be replaced by a quantity of equal value of any of the other outputs.
3. (3) Marginal productivity of a factor. The final necessary marginal equality relates small changes in the quantity of an input, say x_j, to the resulting changes in the quantity of any of the outputs, say y_i, in accordance with the production function. Mathematically,

$\delta y_i / \delta x_j = -(\partial g / \partial x_j) / (\partial g / \partial y_i)$. When profits are being maximized $\delta y_i / \delta x_j = w_j / p_i$. Economically, this is seen to require that if any output is increased by a small amount, the use of some factor must be increased by an amount of equal value.

This third differential equality, of course, is the marginal productivity doctrine, which is seen to be one of the consequences of the theory of profit maximization under competitive conditions. It is often written in the form $\mathrm{VMP}_j = p_i(\delta y_i / \delta x_j) = w_j$, for all values of *i*. This formula defines the value of the marginal product of the *j*th factor to be the increase in the value produced of any product for which that factor is used, per unit increase in the use of the factor, and holds that the price per unit of that factor's services will be equal to the VMP when profits are being maximized.

Evaluation

At present the marginal productivity principle is used to explain the demand for factors of production in both a two-factor version using aggregate capital and aggregate labour as the factors, and an *n*-factor version, where *n* is the number of distinguishable factors used in the production process.

To use the two-factor version it is necessary to establish quantitative measures of the aggregates of dissimilar objects that are given the names ‘capital’ and ‘labour’, a task that has never been performed to anyone's satisfaction. For a long time, until the publication of J. Robinson's disturbing paper, ‘The production function and the theory of capital’ (1953), the lack of satisfactory measures of the aggregates was regarded as a technicality that did not affect the essential insight. But that paper drove home the realization that in the absence of those measures the marginal productivities, i.e. $\partial g / \partial K$ and $\partial g / \partial L$, were essentially undefined. From that time forth, analyses that use the two-factor version have been regarded as simplified ‘parables’, useful for making an intuitive point, but not to be taken literally. Clarifying the meaning of ‘capital’ and ‘labour’ regarded as homogeneous factors of production continues to be one of the main problems of capital theory.

The *n*-factor version avoids the impossible task of aggregating apples and bulldozers, but has problems of its own. The formulation considered in this article is too simple to fit most industrial or commercial situations. It presumes that the constraints on choices can be described reasonably well by a single well-behaved, differentiable production function. This is generally not the case. Extreme examples are production functions in which factors are used or outputs are produced in fixed proportions. Any cooking recipe provides an example. More usual are cases in which the choices of input and output quantities are constrained by several functional relationships. The typical example is a firm or industry in which several different machines are each used in the production of several different products. Then there will be a functional relationship for each type of machine, to express the capacity of that type required for each combination of quantities of the different products. This is the sort of problem that has given rise to the use of linear programming in production planning and economic planning generally. Where there are several constraints, the formulation used above does not apply because, essentially, the amounts of the inputs and outputs cannot be varied two at a time if they are connected by more than a single constraint. Marginal productivity is still a well-defined concept, but it no longer satisfies simple formulas like $\mathrm{MP}_j = p_j(\delta y_i / \delta x_j)$ for any value of *i*. Instead, the marginal productivities, as defined above, are identified with the shadow prices in the solution to a mathematical programming problem,

which is a considerably less intuitive concept.
Very frequently, if the problem of finding the combination of factor inputs that maximizes profits is solved straightforwardly, some of the input levels in the solution turn out to be negative – which is nonsense. Then, again, resort must be had to mathematical programming types of formulation and interpretation. The essential perceptions of marginal productivity theory still apply, but they can no longer be expressed by equalities between price ratios and ratios of marginal changes.

See Also

- capital theory (paradoxes)
- Clark, John Bates
- classical distribution theories

Bibliography

Expositions of marginal productivity theory can be found in any standard text on microeconomics, for example Mansfield (1985). A famous and thoughtful presentation is contained in Hicks (1932). More advanced, and more mathematical, treatments, can be found in Baumol (1977) and Malinvaud (1972). For the relation between marginal productivity and mathematical programming, see Dorfman et al. (1958). The standard, and excellent, reference on the history of the doctrine is Stigler (1941).

Baumol, W.J. 1977. *Economic Theory and Operations Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Dorfman, R., Samuelson, P.A. and Solow, R.M. 1958. *Linear Programming and Economic Analysis*. New York: McGraw-Hill.

Hicks, J.R. 1932. *The Theory of Wages*. New York: Macmillan.

Malinvaud, E. 1972. *Lectures on Microeconomic Theory*. Amsterdam and London: North-Holland.

Mansfield, E. 1985. *Microeconomics: Theory and Applications*. New York: Norton.

Stigler, G.J. 1941. *Production and Distribution Theories: The Formative Period*. New York: Macmillan.

How to cite this article

Dorfman, Robert. "marginal productivity theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000046> doi:10.1057/9780230226203.1025

The New Palgrave Dictionary of Economics Online

marginal revolution

Roger E. Backhouse
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The marginal revolution saw the introduction of the idea of marginal utility into economics in the early 1870s by Jevons, Walras and Menger. This change in economic theory was a slower process than the word ‘revolution’ suggests, and, to understand the changes associated with it, it is necessary to explore the scientific, social and political context in which they occurred.

Keywords

Clark, J. B.; classical economics; Cliffe Leslie, T. E.; collectivism; Cournot, A. A.; Energetics; eugenics; evolution; evolutionary psychology; Fabian economics; Green, T. H.; income distribution; individualism; Ingram, J. K.; Jevons, W. S.; marginal revolution; marginal utility; Marshall, A.; Marxism; mathematics and economics; Menger, C.; Mill, J. S.; Mirowski, P.; poverty; Rau, K. H.; Say, J.-B.; Schumpeter, J. A.; simultaneous equations; Social Darwinism; socialism; statistics and economics; Steuart, J.; subjective theory of value; supply and demand theory of value; utilitarianism; Walras, L.; Wicksell, J. G. K.

Article

The marginal revolution (sometimes called the marginal utility revolution) refers to the introduction into economics, in 1870–1, of the concept of marginal utility by William Stanley Jevons, Léon Walras and Carl Menger and which has widely been seen as involving a revolutionary break with the ‘classical’ economics of David Ricardo, John Stuart Mill and many of their contemporaries (see Blaug, 1996, ch. 8). The value of a commodity was no longer explained in terms of its cost of production (possibly reducible to the labour required to produce it) but in terms of its value to the consumer. The concept of utility was used to explain consumer choices, marginal utility being seen by some (though not all) authors as replacing cost of production as the foundation on which the theory of value rested. In the 1890s marginal techniques were then applied systematically to the problem of income distribution. This

change, it is argued, revolutionized economics, laying the foundations on which modern economic theory is built. Its many dimensions – viewing behaviour as optimization, using utility to describe individual behaviour, focusing on individual agents, the use of mathematics – attest to its importance (Hutchison, 1978, provides a longer list; Maas, 2005, ch. 1, sketches more recent attempts to choose between them). There is disagreement over the extent to which the change should be described as a revolutionary or as an evolutionary change going back many decades, and over its exact significance; but the marginal revolution is firmly established in histories of economic thought. However, while it describes certain developments in economic theory, to understand the changes that took place in economics around this time one should place it in a broader historical context.

Varieties of marginalism

The most important qualification to the idea of a marginal revolution is the heterogeneity of economics during this period. Classical ideas were dominant in Britain, but even within classical economics there was great variety, and it has even been argued that marginalist ideas can be traced back as far as Steuart (see English School of political economy). At some time, virtually every element in the classical system outlined above had been challenged, many of these challenges leaving their mark. Much of this variety was captured within Mill's *Principles of Political Economy*, which went through seven editions between 1848 and 1873, and was undoubtedly the leading treatment of the subject: he worked with a very broad supply and demand theory of value and had accommodated many modifications to the Ricardian theory of income distribution. From Mill, the jump to marginalist theories was much easier than from Ricardo. Indeed, Alfred Marshall, though unfairly praising Ricardo at the expense of Mill (see O'Brien, 1990), derived his theory of value by translating Mill into mathematics during the 1860s; when he encountered Jevons's work, it was a simple matter to graft marginal utility on to a mathematical treatment of supply and demand (Whitaker, 1975).

There was also great variety across countries. In Ireland, it has been argued that an independent tradition of subjective value theory had been established at Trinity College Dublin, by successive holders of the Whately Chair (Black, 1945). Ireland also produced two leading exponents of a historical approach to economics, T.E. Cliffe Leslie and John Kells Ingram. Leslie's assault on deductive theorizing was a significant factor in the shaking of confidence in classical economics in Britain in the 1870s (see Hutchison, 1953). In Germany, supply and demand theories had a long history, a supply and demand diagram having been used in a textbook as early as 1843 by Heinrich Rau (see Streissler, 1990). In France, Smithian political economy had been mediated not through Ricardo but through Jean-Baptiste Say. The work of Augustin Cournot and the engineers of the École des ponts et chaussées, whose analysis rested on the concept of a demand curve, created an intellectual background very different from that prevailing in Britain.

These differences, together with profound differences in their personal backgrounds, meant that the work of Jevons, Walras and Menger, though often bracketed together, was far from homogeneous (see Jaffe, 1975). Though Jevons and Walras both advocated the importance of mathematical argument, their emphases differed. Walras, closer to French rationalism, saw his general equilibrium equations as an abstract system that could solve the same problem that was solved in the real world by other means. Jevons focused on mechanical analogies and the notion that the same methods could be applied to

physical and social sciences (Maas, 2005). The contrast was even more marked in their applied work, where Jevons was a pioneer in the use of statistics but Walras was not. Menger, in contrast with both of them, rejected the use of mathematics, seeing the use of simultaneous equations as incompatible with identifying the causal relations between human needs and the value of commodities.

The varieties of marginalism increase further when later marginalists are brought into the account. Jevons, Walras and Menger did have disciples, most of them took their analysis in new directions and many are in many ways originals, the best examples being Marshall, Joseph Alois Schumpeter (Austrian, geographically and intellectually, yet an admirer of Walras), Knut Wicksell (whose Swedish synthesis of Austrian and Walrasian ideas bore little resemblance to Schumpeter's) and John Bates Clark (who constructed a non-mathematical American version of marginalism). Given this variety, it is not surprising that the marginal revolution can also be portrayed as a very slow process. Even in the 1880s and 1890s, some economists were still writing textbooks organized on classical lines, marginalist ways of thinking co-existing with other lines of enquiry.

The wider context

While scholars might, at one time, have been content to explain the advent of marginalist ideas in terms of economists coming to see the truth about consumers and value, historians are no longer satisfied with such explanations, arguing that economic ideas have to be explained in terms of the context out of which they arose. One context is that of 19th-century science. The most widely discussed explanation has been Mirowski's (1984; 1989) argument that marginalist economics reflects developments in physics (see De Marchi, 1993). The 1860s saw the rise of energetics – the attempt to reduce all physical phenomena to energy. If physical phenomena could be reduced to energy, then so should social phenomena. More than that, adopting the methods of physical scientists and the mathematics of maximization and energy conservation offered economists the possibility of acquiring the status of physicists, adopting similar standards of rigour. Mirowski directed historians’ attention to the many passages where Jevons, Walras and others stated explicitly that this was what they were doing. Though Mirowski drew normative conclusions about which many historians have been sceptical, and though his interpretation clearly does not fit some of the most important marginalists (notably Menger and Marshall), historians have taken up the idea that a major dimension to the marginal revolution was seeing economics as amenable to the methods of the physical sciences rather than as something radically different (see Maas, 2005; Schabas, 2005).

Moreover, at this period, physics was not the only prestigious natural science: controversies over evolution were at their height, following the publication of Charles Darwin's *Origin of Species* and the application of evolutionary arguments to human society by Herbert Spencer. This cannot explain the advent of marginalism, but it represents an important additional connection between economics and contemporary science and helps explain why economics looked very different at the end of the 19th century from the way it looked in the 1860s. Raffaelli (2003) has pointed out that Marshall, perhaps the most significant figure in late-19th century marginalist economics, based his economics, not on the Benthamite utilitarianism used by Jevons, but on evolutionary psychology. Human nature was moulded by experience. Evolutionary ideas thus reinforced the notion that human beings had to be seen as different from one another and that they could be changed. This way of thinking could lead into

eugenics, a widely entertained body of ideas that developed towards the end of the century (see Peart and Levy, 2005). But such ideas also served to undermine the Malthusian bogey that had provided an argument against much social reform throughout the century (Stedman Jones, 1984). Marshall, for example, though he used the static, mechanical apparatus of supply and demand, used it to discuss dynamic processes. He saw industries evolving as biological species, and human character changing in response to human activities, a process that was too complicated to be represented mathematically, and as a result never worked with formal dynamic models: they would have been too mechanical. Against such arguments that evolution became influential at that time, Schabas (2005), though stressing that neoclassical economists were very interested in psychology, has recently questioned whether Darwin has as much influence as has been claimed.

The significance of evolutionary ideas points to another aspect of the context against which the advent of marginalist ideas needs to be set: the political climate. The late 19th century has been called the liberal age, when Europe moved towards freer trade and the franchise became more democratic. The progress of liberal ideas and policies varied greatly from country to country, but everywhere there was debate over the merits of liberalism and collectivism, with the latter taking many forms, ranging from Fabian ‘municipal socialism’ to Marxian socialism. In Britain, the mid-century radicals, amongst whom Mill was pre-eminent, were liberals who wanted to reform the institutions of society in ways consistent with their liberalism. But, by the end of the century, following the extension of the franchise to much of the working class in 1867 and 1884, radicalism became increasingly collectivist. Against Social Darwinist arguments for individualism were ranged ethical arguments for reform, from the American Social Gospel movement to the variety of movements for social reform inspired in Britain by the Oxford philosopher T.H. Green (see Richter, 1964). In the same way that the Great Depression motivated many who came into economics in the 1930s and 1940s, the problem of poverty affected this earlier generation. Economists’ attitudes towards policy changed (see Hutchison, 1978), as did the way they developed their theories, the most noticeable example being the development of welfare economics by the Cambridge School, J.A. Hobson, and others.

Though it was again a process the speed of which varied greatly from country to country, a further element of the context in which the marginal revolution took place was the professionalization of economics. By the middle of the 19th century, economics was being developed by a mixture of academics and members of a broader educated elite; those recognized as economists might be politicians or businessmen. Specialist journals existed in some countries, but original work in the subject was also published in journals read by non-specialists. By the end of the century economics, like many other disciplines, had changed, becoming an academic discipline in which the main communication was between specialists. This made possible a different type of discourse, more technical and addressing issues that might seem more tenuously related to issues of concern to lay people.

Conclusions

The marginal revolution, like other revolutions in economics, is associated with changes in economic theory that undoubtedly altered the way economics was conceived. However, picking out a single theoretical or methodological innovation that explains why the marginal revolution was apparently so important has proved difficult. The reason may be that, as in the case of the Keynesian Revolution,

though economics changed profoundly in the closing decades of the 19th century, these changes owed as much, if not more, to deeper changes in the social, political and intellectual context in which economists were working as to any specific innovation in economic theory.

See Also

- English School of political economy
- Jevons, William Stanley
- Marshall, Alfred
- Menger, Carl
- Walras, Léon

Bibliography

Black, R.D.C. 1945. Trinity College Dublin and the theory of value. *Economica* 47, 140–8.

Blaug, M. 1996. *Economic Theory in Retrospect*, 5th edn. Cambridge: Cambridge University Press.

De Marchi, N. (ed.) 1993. *Non-natural Social Science: Reflecting on the Enterprise of More Heat than Light*. Durham, NC: Duke University Press. Also in *History of Political Economy* 25 (supplement), 271–82.

Hutchison, T.W. 1953. *A Review of Economic Doctrines, 1870–1929*. Oxford: Oxford University Press.

Hutchison, T.W. 1978. *On Revolutions and Progress in Economic Knowledge*. Cambridge: Cambridge University Press.

Jaffe, W. 1975. Menger, Jevons and Walras dehomogenized. *Economic Inquiry* 14, 511–24.

Maas, H.B.J.B. 2005. *William Stanley Jevons and the Making of Modern Economics*. Cambridge: Cambridge University Press.

Mirowski, P. 1984. Physics and the marginalist revolution. *Cambridge Journal of Economics* 8, 361–79.

Mirowski, P. 1989. *More Heat Than Light: Economics as Social Physics, Physics as Nature's Economics*. Cambridge: Cambridge University Press.

O'Brien, D.P. 1990. Marshall's work in relation to classical economics. In *Centenary Essays on Alfred Marshall*, ed. J. Whitaker. Cambridge: Cambridge University Press.

Peart, S. and Levy, D. 2005. *The Vanity of the Philosopher: From Equality to Hierarchy in Post-Classical Economics*. Ann Arbor: University of Michigan Press.

Raffaelli, T. 2003. *Marshall's Evolutionary Economics*. London: Routledge.

Richter, M. 1964. *The Politics of Conscience: T.H. Green and His Age*. London: Weidenfeld & Nicolson.

Schabas, M. 2005. *The Natural Origins of Economics*. Chicago: University of Chicago Press.

Stedman Jones, G. 1984. *Outcast London: A Study in the Relationship between Classes in Victorian Society*, 2nd edn. New York: Pantheon Books.

Streissler, E.W. 1990. The influence of German economics on the work of Menger and Marshall. *History of Political Economy* 22 (annual supplement), 31–68.

Whitaker, J.K. (ed.) 1975. *The Early Economic Writings of Alfred Marshall, 1867–1890*. London: Macmillan.

How to cite this article

Backhouse, Roger E. "marginal revolution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000392> doi:10.1057/9780230226203.1026

The New Palgrave Dictionary of Economics Online

marginal utility of money

Eugene Silberberg

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Alfred Marshall identified the area to the left of a demand curve as consumer surplus, but he added that his discussion was valid only under the assumption of ‘constant marginal utility of money’. For much of the 20th century economists debated the meaning of that phrase and its relevance to consumer surplus. The analysis became clear only after the development of duality theory, particularly the properties of the expenditure function. Marshall's caution becomes unnecessary with a proper definition of consumer surplus.

Keywords

consumer surplus; envelope theorem; Hicksian and Marshallian demands; homothetic utility functions; indirect utility function; marginal utility of money; Marshall, A.; money; Roy's equality

Article

Interest in the marginal utility of money probably dates from Alfred Marshall's identification of consumer surplus as the area under the demand curve. Marshall went on to add a qualification to his analysis:

> In the same way if we were to neglect for the moment the fact that the same sum of money represents a different amount of pleasure to different people, we might measure the surplus satisfaction which the sale of tea affords, say, in the London market, by the aggregate of the sums by which the prices shown in a complete list of demand prices of tea exceeds its selling price. (Marshall, 1920, p. 106)

In the mathematical appendix (Note VI), Marshall identifies the ‘total utility of the commodity’ with the area under the demand curve, defined by an integral, but then qualifies that analysis by saying ‘we

assume that the marginal utility of money to the individual purchaser is the same throughout'. The meaning of these phrases is anything but clear. The text phrase seems to indicate that interpersonal comparisons of utility are a necessary prerequisite for the use of consumer's surplus; in the appendix, Marshall's concern is that, as more of a commodity is purchased, money will yield less satisfaction to the consumer, destroying any linear relationship between money and utility.

Later interpretation of 'constant marginal utility of money' was further complicated by the use of the word 'money' in two different contexts. To Marshall, money provided no direct utility to the consumer; it was a device solely for lowering the transactions cost of exchange. The concurrently developed general equilibrium theory of Walras, however, treated money as that one good which happened to have the additional property of serving as the medium of exchange, a numéraire commodity whose price was unity.

We now analyse the connection between the marginal utility of money and consumer's surplus. The consumer is assumed to maximize $U = U(x_1, \ldots, x_n)$ subject to $\Sigma\, p_i x_i = M$. We derive the Marshallian (money income held constant) demand functions $x_i = x_i^M(p_1, \ldots, p_n, M)$ along with $\lambda^M(p_1, \ldots, p_n, M)$ using the Lagrangian $L = U + \lambda(M - \Sigma\, p_i x_i)$.

The indirect utility function $U^*(p_1, \ldots, p_n, M) = U(x_1^M, \ldots, x_n^M)$ indicates the maximum utility for given prices and money income. Using the envelope theorem, $\partial U^* / \partial M = \lambda^M$, the marginal utility of money income. Also, $\partial U^* / \partial p_i = -\lambda^M x_i^M$ (Roy's identity). The Hicksian (utility held constant) or 'compensated' demand functions $x_i^U(p_1, \ldots, p_n, U)$ are derived from minimizing $M = \Sigma\, p_i x_i$ subject to $U(x_1, \ldots, x_n) = U^0$. The expenditure function $M^*(p_1, \ldots, p_n, U^0) = \Sigma\, p_i x_i^U$ indicates the minimum cost of maintaining utility level U^0 for arbitrary prices $p_1, \ldots, p_n$. By the envelope theorem, the Hicksian demands are the first partials of the expenditure function: $x_i^U = \partial M^* / \partial p_i$. (See Hicksian and Marshallian demands.)

The area to the left of a consumer's demand curve between two prices (where the initial price is higher than the final price), is $-\int x_i \partial p_i$. The units of this integral are that of money income, being price times quantity. Suppose this area equals some value *A*. The issue is: what question does *A* answer, and what is the relation between that answer and the marginal utility of money? Since a Hicksian demand function is the first partial of the expenditure function, the area to the left of this demand curve is simply a change in the expenditure function:

$$-\int x_i^U \mathrm{d}\, p_i = -\int (\partial M^* / \partial p_i)\, \mathrm{d}\, p_i = M^*\left(p^0, U^0\right) - M^*\left(p^1, U^0\right)$$

where p^0 and p^1 are the initial and final price vectors over which the integral is taken. The area to the left of Hicksian demand function therefore represents a change in expenditure with utility held constant; this area indicates the amount a consumer would be willing to pay (or have to be paid) to willingly

accept some change in the purchase price of some good. Moreover, there is no need to invoke any assumption at all about the marginal utility of money.

The area to the left of the Marshallian demand function, however, has no such easy interpretation, because unlike the Hicksian demands, the Marshallian demand functions are *not* in general the partial derivatives of some integral function; therefore the integrals of the Marshallian demands are not expressible in terms of changes in some well-defined function of the initial and final prices and income levels. From Roy's equality, the Marshallian demands are the first partials of the indirect utility function *divided by the marginal utility of income*. Thus

$$-\int x_i^M \mathrm{d}\, p_i = -\int \left(1/\lambda^M\right)\left(\lambda^M x_i^M\right)\mathrm{d}\, p_i = \int \left(1/\lambda^M\right)\left(\partial U^* / \partial p_i\right)\mathrm{d}\, p_i.$$

However, if the marginal utility of money term is ‘constant’, that is, independent of prices, it can be moved in front of the integral sign; only then can this expression be integrated to yield a function of the endpoint prices (and money income):

$$-\int x_i^M \mathrm{d}\, p_i = (1/\lambda^M)\int (\partial U^* / \partial p_i)\, \mathrm{d}\, p_i = (1/\lambda^M)\left[U(p^1) - U(p^0)\right].$$

Thus, in this case, the area to the left of the Marshallian demand function would equal a change in utility divided by the marginal utility of money, thus converting that change in utility into units of money. Marshall's claim that the area to the left of a demand curve may be interpreted as a change in utility under the assumption of constant marginal utility of money would thus be technically correct for the demand functions derived from utility maximization, though how much of the above discussion he had in mind can easily be debated.

The problem with this analysis is that λ^M cannot literally be a ‘constant,’ as shown by Samuelson (1942). Since $\lambda^M = U_i / p_i$ a proportionate change (for example, doubling) of prices and income leaves the amount of the goods consumed unchanged (since the Marshallian demand functions are homogeneous of degree zero in prices and income), and thus the numerator of this expression unchanged. However, the denominator has doubled, meaning λ^M has halved. Thus $\lambda^M = (p_1, \ldots, p_n, M)$ must be homogeneous of degree –1; it can be independent of at most *n* of its arguments. It can, for example, be independent of all prices, but not income also, or it can be independent of *n* – 1 prices and income.

Since $\partial U^* / \partial p_i = -\lambda^M x_i^M$ and $\partial U^* / \partial M = \lambda^M$, Young's theorem on invariance of partial derivatives to the order of differentiation yields (omitting superscripts)

$$M^*_{p_i M} = -[\lambda \partial x_i / \partial M + x_i \partial \lambda / \partial M] = \partial \lambda / \partial p_i = M^*_{M p_i}.$$

Suppose

$$\partial \lambda^M / \partial p_i = 0, i = 1, \ldots, n.$$

Then

$$(M / x_i)(\partial x_i / \partial M) = -(M / \lambda)(\partial \lambda / \partial M) \text{for} i = 1, \ldots, n.$$

That is, the income elasticities are all equal (necessarily to unity, from the budget constraint); thus the utility function must be homothetic. Denoting the Marshallian area *CS*, we have

$$CS = \left(1 / \lambda^M\right)\left[U^*\left(p^1, M\right) - U^*\left(p^0, M\right)\right].$$

Thus for homothetic utility functions, where the indifference curves are all radial blow-ups of each other, the Marshallian area represents the unique monetary equivalent of a change in utility; the coefficient which converts ‘utiles’ to money income is invariant over the price change.
Suppose now that λ^M is a function of one price only, say p_n. Then from the above equation, $\partial x_i^M / \partial M = 0, i = 1, \ldots, n-1$. Since there is no income effect for goods 1 to $n - 1$, the Marshallian demand functions for those goods coincide with the Hicksian demands. This is the famous case of ‘vertically parallel’ indifference curves. Therefore the interpretation of the area to the left of any of these Marshallian demand curves is identical to the case of the Hicksian demands, that is, the willingness to pay to face the lower price.

See Also

- consumer surplus
- Giffen's paradox
- Hicksian and Marshallian demands

- indirect utility function
- Marshall, Alfred

Bibliography

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Samuelson, P.A. 1942. Constancy of the marginal utility of money. In *Studies in Mathematical Economics and Econometrics: In Memory of Henry Schultz*, ed. O. Lange, F. McIntyre and T.O. Yntema. Chicago: University of Chicago Press.

Silberberg, E. and Suen, W. 2000. *The Structure of Economics*, 3rd edn. New York: McGraw-Hill.

How to cite this article

Silberberg, Eugene. "marginal utility of money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000047> doi:10.1057/9780230226203.1027

The New Palgrave Dictionary of Economics Online

market competition and selection

Lawrence E. Blume and David Easley

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There is a long history in economics of using market selection arguments in defence of rationality hypotheses. According to these arguments, rational investors drive irrational investors out of asset markets and profit maximizing firms drive non-maximizing firms out of goods markets. In this article we present the history of these arguments and discuss the literature that examines whether these arguments for market selection, and its impact on efficiency, are correct.

Keywords

rationality; market selection; Kelly rule; incomplete markets

Article

> Realized positive profits, not *maximum* profits, are the mark of success and viability. It does not matter through what process of reasoning or motivation such success was achieved. The fact of its accomplishment is sufficient. This is the criterion by which the economic system selects survivors: those who realize *positive profits* are the survivors; those who suffer losses disappear. (Alchian, 1950, p. 213)

Most economic models make use of extreme rationality hypotheses: firms maximize profits with full knowledge of their technology and prices, and investors are subjective expected utility maximizers whose beliefs are correct. Surely some firms and some investors do not always behave as these models hypothesize, but does this matter for predictions of market outcomes? It could be that the aggregation that takes place in supply and demand results in prices and market quantities that agree with the predictions of models using extreme versions of rationality. It could be that, over time, firms and investors learn to behave as these models predict and so market outcomes converge to those predicted by the models. Finally, it could be that markets select for firms and investors who behave ‘as if’ they are

rational. This last defence of the use of rationality is the essence of the quote from Alchian (1950). There is a long history in economics of using market selection arguments in defence of rationality hypotheses. The early literature focused on selection for profit maximizing firms. Among its best-known proponents is Friedman (1953, p. 22): ‘The process of natural selection thus helps to validate the hypothesis (of profit maximization) or, rather, given natural selection, acceptance of the hypothesis can be based largely on the judgment that it summarizes appropriately the conditions for survival.’ Of course, even if the selection reasoning is correct, selection can only work over those types of behaviours that are present in the economy. If no firm maximizes profits, then no profit-maximizing firm can be selected. Alchian was acutely aware of this:

> The pertinent requirement – positive profits through relative efficiency – is weaker than ‘maximized profits,’ with which, unfortunately, it has been confused. Positive profits accrue to those who are better than their actual competitors, even if the participants are ignorant, intelligent, skilful, etc. The crucial element is one's aggregate position relative to actual competitors, not some hypothetically perfect competitors. As in a race, the award goes to the relatively fastest, even if all the competitors loaf. (Alchain, 1950, p. 213)

Enke (1951) argued that, at least in competitive industries, the relatively fastest will in fact be profit maximizers, and so in this case selection will lead to the survival only of profit maximizing firms:

> In the long run, however, if firms are in active competition with one another rather than constituting a number of isolated monopolies, natural selection will tend to permit the survival of only those firms that either through good luck or great skill have managed, almost or completely, to optimize their position and earn the normal profits necessary for survival. In these instances the economist can make aggregate predictions *as if* each and every firm knew how to secure maximum long-run profits. (Enke, 1951, p. 567)

Similar market selection arguments have been proposed to justify strong rationality hypotheses on the part of investors. Fama argues that:

> dependency in the noise generating process would tend to produce ‘bubbles’ in the price series … If there are many sophisticated traders in the market, however, they may cause these ‘bubbles’ to burst before they have a chance to really get underway. (Fama, 1965, p. 38)

According to Fama, ‘A superior analyst is one whose gains over many periods of time are *consistently* greater than those of the market’. This is at least indirectly an argument for market selection and its affect on the efficiency of prices. Cootner was an early, clear proponent of this argument:

> Given the uncertainty of the real world, the many actual and virtual traders will have many, perhaps equally many, forecasts … If any group of traders was consistently better

> than average in forecasting stock prices, they would accumulate wealth and give their forecasts greater and greater weight. In this process, they would bring the present price closer to the true value. (Cootner, 1967, p. 80)

In this article we examine the more recent analyses of whether these arguments for market selection, and its impact on efficiency, are correct. We consider in turn, selection over firms and selection over investors.

Selection over firms

Alchain, Friedman and Enke argue that a profit dynamic will select for firms that, for whatever reason, maximize profits. Correspondingly, according to this argument, those that do not act as profit maximizers will be driven out of the market. But how is it that non-maximizers are driven out? The implicit idea is that, in the presence of maximizers, the non-maximizers experience losses that deplete their financial capital, which forces them out of the market. The literature has explored two avenues by which losses of financial capital could have this effect. One is that if the firm's operations are financed from retained earnings, then firms that consistently experience losses would eventually exhaust their retained earnings, causing them to vanish. A second argument is that unsuccessful firms will not be able to raise capital in the financial markets, and may not even be able to retain their initial capital. Thus, so this argument goes, the markets will punish unsuccessful firms, which will eventually vanish.

Winter (1964; 1971) and Nelson and Winter (1982) analyse a retained earnings dynamic. They argue that the retained earnings of profit maximizers will grow fastest, and thus these firms will eventually dominate the market. These authors construct a partial equilibrium model in which the ‘as if’ hypothesis of profit maximization describes the long-run steady state behaviour of firms. In their analysis, prices are fixed and all firms have access to the same technology. This structure leads to the existence of a uniformly most-fit firm, which is selected for by a retained earnings-based investment dynamic.

The early work on market selection was greatly concerned with the meaning of profit maximization when profits are random. Dutta and Radner (1999) directly take up the question of whether markets select for firms that maximize expected profits. Their answer is ‘no’: the decision rules that maximize the long probability of survival are not those that maximize expected profits. Dutta and Radner's firms are owned by investors who choose how much of the firm's earnings to reinvest in the firms and how much to withdraw as dividends. An expected profit maximizing firm is one that maximizes the expectation of present discounted value of dividends paid to its owners. This policy results in an upper bound on the retained earnings left in the firm, and from this level of retained earnings any firm can experience a string of losses that results in bankruptcy.

There are two parts to the argument for market selection of profit maximizers. First, there is the issue of whether the market selects for profit maximizers. Second, there is the issue of whether in the long run the economy behaves as if only profit mazimizing firms exist. The Dutta and Radner analysis casts doubt about a positive answer to the first question in stochastic settings. Koopmans (1957) cast doubts about a positive answer to the second question even in a deterministic setting. According to his analysis, appealing to an external dynamic process to defend the profit maximization assumption is not a satisfactory way to proceed. Instead, he believed that the dynamic process itself should be modelled.

Nelson and Winter (1982, p. 58) were also aware that the co-evolution of firm behaviour and the economic environment resulting from a complete model of the dynamic process could pose problems for the evolutionary defence of profit maximization. They observed that among the ‘less obvious snags for evolutionary arguments that aim to provide a prop for orthodoxy’ is ‘that the relative profitability ranking of decision rules may not be invariant with respect to market conditions’. They do not, however, go on to provide a general equilibrium analysis of the consequences of replacing static profit maximization with a selection dynamic.

Blume and Easley (2002) showed that Koopman's concern about the market selection dynamic in a general equilibrium setting is correct. They show that although only profit maximizers persist in any steady state of the retained earnings dynamic, the long run of the economy need not be well described by assuming that only profit maximizing firms exist. The difficulty arises because of the endogeneity of prices, which causes the relative profitability of firms to depend on the allocation of capital across the firms. As a result, the retained earnings dynamic need not settle down, and efficient firms can be driven out of the market by inefficient firms.

In addition to raising working capital through retained earnings, firms also enter the capital markets. Whether these markets reinforce the market selection hypothesis, as Friedman argues, or undermine it, depends on how well these markets function. If markets are complete (without the securities created by non-maximizing firms) and investors are expected utility maximizers with rational expectations, then investors would not allocate capital to non-maximizing firms. Such firms would never produce, and the selection hypothesis would be trivially, and instantly, correct. Alternatively, if some investors have incorrect expectations, then they could invest in non-maximizing firms. The fate of these firms depends on the fate of their investors. So, in this case, the question of selection for profit maximizing firms reduces to the question of selection for investors who act as expected utility maximizers with rational expectations.

Selection over investors

Friedman, Fama and Cootner argue that asset markets will select for rational investors, and that because of this selection, assets will eventually be priced efficiently. Two interesting approaches have been taken to the selection for rational investors question. First, suppose traders use a variety of portfolio rules. Is it the case that traders whose rules are not rational will lose their money to those who do act as if they are rational? Second, suppose that all traders are subjective expected utility maximizers. Is it the case that markets select for those whose expectations are correct, or most nearly correct?

In order to pose these questions precisely rationality has to be defined (see rationality). The selection literature has asked about selection for a very strong form of rationality – expected utility maximization with correct expectations about the payoffs to assets. This is the interesting question because in economies populated by subjective expected utility maximizers whose beliefs are not tied down by a rational expectations hypothesis we have little to say about asset prices. The mere assumption that investors are subjective expected utility maximizers (in the sense of Savage, 1951) places no restrictions on the stochastic process of Arrow security prices (Blume and Easley, 2005).

Selection over rules

Consider an intertemporal general equilibrium economy with a collection of Arrow securities and one physical good available at each date. Suppose traders are characterized by their stochastic processes of endowments of the good and by portfolio and savings rules. A savings rule describes the fraction of wealth the trader saves and invests at each date given any partial history of states. Similarly, a portfolio rule describes the fraction of savings the trader allocates to each Arrow security. The savings and portfolio rules that rational traders could choose form one such class of rules; but other, non-rationally motivated rules are also possible.

Three questions arise about the dynamics of wealth selection in this economy. First, is there any kind of selection at all? Second, is it possible to characterize the rules which win? Third, if selection does take place, does every trader using a rational rule survive, and in the presence of such a trader do all non-rational traders vanish?

In repeated betting, with exogenous odds, the betting rule that maximizes the expected growth rate of wealth is known as the Kelly rule (Kelly, 1956). The use of this formula in betting with fixed, but favourable odds was further explored by Breiman (1961). In asset markets the ‘odds’ are not fixed; instead they are determined by equilibrium asset prices, which in turn depend on traders’ portfolio and savings rules. Nonetheless, the market selects over rules according to the expected growth rate of wealth share they induce. Blume and Easley (1992) show that if there is a unique trader using a rule that is globally maximal with respect to this criterion, then this trader eventually controls all the wealth in the economy, and prices are set as if he is the only trader in the economy. A trader whose savings rate is maximal and whose portfolio rule is, in each partial history, the conditional probability of states for tomorrow has a maximal expected growth rate of wealth share. This rule is consistent with the trader having logarithmic utility for consumption, rational expectations and a discount factor that is as large as any trader's savings rate. Thus, if this trader exists, he is selected for. However, rationality alone does not guarantee a maximal expected growth rate of wealth share. There are rational portfolio rules that do not maximize fitness (even controlling for savings rates), and traders who use these rules can be driven out of the market by traders who use rules that are inconsistent with rationality.

Amir et al. (2005) and Evstigneev, Hens and Schenk-Hoppe (2006) take an alternative approach to selection over rules in asset markets. They consider general one-period assets and ask if there are simple portfolio rules that are selected for, or are evolutionarily stable, when the market is populated by other simple (not explicitly price dependent) portfolio rules. In this research, either all winnings are invested, or equivalently, all investors are assumed to invest an equal fraction of their winnings. So selection operates only over portfolio rules. Amir et al. (2005) find that an investor who apportions his wealth across assets according to their conditional expected relative payoffs drives out all other investors as long as none of the other investors end up holding the market. This result is consistent with Blume and Easley (1992) as the log optimal portfolio rule agrees with the conditional expected relative payoff rule when only these two rules exist in the market. Hence, both these rules hold the market in the limit. Evstigneev, Hens and Schenk-Hoppe (2006) use notions of stability from evolutionary game theory to show that the expected relative payoffs rule is evolutionarily stable.

Selection among subjective expected utility maximizers

DeLong et al. (1990; 1991) analyse selection over traders who are subjective expected utility maximizers with differing beliefs. In an overlapping generations model they show (1990) that traders with incorrect beliefs can earn higher expected returns, because they take on extra risk. But as survival is not determined by expected returns, this result does not answer the selection question. DeLong et al. (1991) argue that traders whose beliefs reflect irrational overconfidence can eventually dominate an asset market in which prices are set exogenously. This result appears to contradict Alchian's and Friedman's intuitions. But, as prices are exogenous, these traders are not really trading with each other; if they were, then were traders with incorrect beliefs to dominate the market, prices would reflect their beliefs and rational traders might be able to take advantage of them.

In an economy with complete markets and traders who have a common discount factor, Alchian and Friedman's intuition is correct. Sandroni (2000) shows, in a Lucas trees economy with some rational-expectations traders, that if traders have a common discount factor, then all traders who survive have rational expectations. Blume and Easley (2006) show that this result holds in any Pareto optimal allocation in any bounded classical economy and thus for any complete markets equilibrium. To see why the market selection hypothesis is true for these economies suppose that states are iid and that traders have differing, fixed iid beliefs. Then each trader assigns zero probability to almost all the infinite sample paths that any other trader believes to be possible. Each trader would be willing to give up all his endowment on the sample paths he believes to be impossible in order to obtain more consumption on those he believes to be possible. Since markets are complete, these trades are effectively possible. But, if only one trader has correct beliefs, then only one trader puts positive probability on the infinite sample paths that actually occur. So only this trader will have positive consumption, and thus positive wealth, in the limit.

For bounded complete market economies there is a survival index that determines which traders survive and which vanish. This index depends only on discount factors, the actual stochastic process of states, and, traders’ beliefs about this stochastic process. Most importantly, for these economies, attitudes towards risk do not matter for survival. The literature also provides various results demonstrating how the market selects among learning rules. The market selects for traders who learn the true process over those who do not learn the truth, for Bayesians with the truth in the support of their prior over comparable non-Bayesians, and among Bayesians according to the dimension of the support of their prior (assuming that the truth is in the support).

In economies with incomplete markets, the market selection hypothesis can fail to be true. Blume and Easley (2006) show that if markets are incomplete, then rational traders may choose either savings rates or portfolio rules that are dominated by those selected by traders with incorrect beliefs. If some traders are irrationally optimistic about the payoff to assets, then the price of those assets may be high enough for rational traders to choose to consume more now, and less in the future. Their low savings rates are optimal, but as a result of their low savings rates the rational traders do not survive.

An alternative version of the market selection hypothesis is that asset markets select for traders with superior information. The research discussed above asks about selection over traders with different, but exogenously given, beliefs. Alternatively, if traders begin with a common prior and receive differential information they will have differing beliefs, but now they will care about each others’ beliefs. In this case, the selection question is difficult because the information that traders have will be reflected in prices. If the economy is in a fully revealing rational expectations equilibrium, then there is no

advantage to having superior information; see Grossman and Stiglitz (1980). So the question only makes sense in the more natural, but far more complex, case in which information is not fully revealed by market statistics. Figlewski (1978) shows that traders with information which is not correctly reflected in prices have an advantage in terms of expected wealth gain over those whose information is fully impounded in prices. But as expected wealth gain does not determine fitness this result does not fully answer the question. Mailath and Sandroni (2003) consider a Lucas trees economy with log utility traders and noise traders. They show that the quality of information affects survival, but so does the level of noise in the economy. Scuibba (2005) considers a Grossman and Stiglitz (1980) economy in which informed traders pay for information and shows that in this case uninformed traders do not vanish.

Conclusion

The modern literature has shown that the market selection hypothesis needs to be qualified. For some economies it acts much as the earlier writers conjectured; in others it does not select for profit maximizers or rational traders. Much work remains to be done, however. Blume and Easley (2006) and Sandroni (2000) mostly discuss selection in complete markets. Sandroni, though, points out that even when markets are incomplete, traders with log utility and rational expectations are favoured, while Blume and Easley construct some examples to show that the outcome of market selection can depend on market completeness. The connection between market structure and market selection is not well understood. The implications of market selection for asset pricing are known only for complete markets in the long run and some examples. Most economists' intuition about market behaviour and asset pricing comes from the study of market models that allow little or no agent heterogeneity. Taking heterogeneity seriously and chasing down its implications for market performance promises to be a rich area for future research.

See Also

- general equilibrium
- rational expectations
- rationality

Bibliography

Alchian, A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58, 211–21.

Amir, R., Evstigneev, I., Hens, T. and Schenk-Hoppe, K.R. 2005. Market selection and survival of investment strategies. *Journal of Mathematical Economics* 41, 105–22.

Blume, L. and Easley, D. 1992. Evolution and market behavior. *Journal of Economic Theory* 58, 9–40.

Blume, L. and Easley, D. 2002. Optimality and natural selection in markets. *Journal of Economic Theory* 107, 95–130.

Blume, L. and Easley, D. 2005. Rationality and selection in asset markets. In *The Economy as an Evolving Complex System*, ed. L. Blume and S. Durlauf. Oxford: Oxford University Press.

Blume, L. and Easley, D. 2006. If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets. *Econometrica* 74, 929–66.

Breiman, L. 1961. Optimal gambling systems for favorable games. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Cootner, P. 1967. *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press.

DeLong, J.B., Shleifer, A., Summers, L. and Waldmann, R. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98, 703–38.

DeLong, J.B., Shleifer, A., Summers, L. and Waldmann, R. 1991. The survival of noise traders in financial markets. *Journal of Business* 64, 1–19.

Dutta, P. and Radner, R. 1999. Profit maximization and the market selection hypothesis. *Review of Economic Studies* 66, 769–98.

Enke, S. 1951. On maximizing profits: a distinction between Chamberlin and Robinson. *American Economic Review* 41, 566–78.

Evstigneev, I., Hens, T. and Schenk-Hoppe, K.R. 2006. Evolutionary stable stock markets. *Economic Theory* 27, 449–68.

Fama, E. 1965. The behavior of stock market prices. *Journal of Business* 38, 34–105.

Figlewski, S. 1978. Market ‘efficiency’ in a market with heterogeneous information. *Journal of Political Economy* 86, 581–97.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Grossman, S.J. and Stiglitz, J.E. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70, 393–408.

Kelly, J.L. 1956. A new interpretation of information rate. *Bell System Technical Journal* 35, 917–26.

Koopmans, T. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Mailath, G. and Sandroni, A. 2003. Market selection and asymmetric information. *Review of Economic Studies* 70, 343–68.

Nelson, R. and Winter, S. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA: Harvard University Press.

Sandroni, A. 2000. Do markets favor agents able to make accurate predictions? *Econometrica* 68, 1303–42.

Savage, L.J. 1951. The theory of statistical decision. *Journal of the American Statistical Association* 46, 55–67.

Scuibba, E. 2005. Asymmetric information and survival in financial markets. *Economic Theory* 25, 353–79.

Winter, S. 1964. Economic natural selection and the theory of the firm. *Yale Economic Essays* 4, 225–72.

Winter, S. 1971. Satisficing, selection and the innovating remnant. *Quarterly Journal of Economics* 85, 237–61.

How to cite this article

Blume, Lawrence E. and David Easley. "market competition and selection." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000355> doi:10.1057/9780230226203.1028

The New Palgrave Dictionary of Economics Online

market failure

John O. Ledyard

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Market failure occurs when there are too few markets, non-competitive behaviour, or non-existence, leading to inefficient allocations. Many suggested solutions for market failure, such as tax-subsidy schemes, property rights assignments, and special pricing arrangements, are simply devices for the creation of more markets. This remedy can be beneficial but, if the addition of markets creates either non-convexities or thin participation, then adding markets will simply lead to market failure from monopolistic behaviour. Examples are natural monopolies and informational monopolies. To achieve a more efficient allocation of resources in the presence of such fundamental failures one must explore non-market alternatives.

Keywords

asymmetric information; contingent claims markets; free rider problem; fundamental theorem of welfare economics; increasing returns to scale; Lindahl prices; market failure; mechanism design; monopoly; monopsony; natural monopoly; non-competitive behaviour; non-convexity; non-existence of equilibrium; Pareto efficiency; property rights reassignments; rational expectations

Article

The best way to understand market failure is first to understand market success, the ability of a collection of idealized competitive markets to achieve an equilibrium allocation of resources that is Pareto optimal. This characteristic of markets, which was loosely conjectured by Adam Smith, has received its clearest expression in the theorems of modern welfare economics. For our purposes, the first of these, named the first fundamental theorem of welfare economics, is of most interest. Simply stated it reads: (1) if there are enough markets, (2) if all consumers and producers behave competitively, and (3) if an equilibrium exists, then the allocation of resources in that equilibrium will be Pareto optimal (see Arrow, 1951; Debreu, 1959). Market failure is said to occur when the conclusion of this theorem is

false; that is, when the allocations achieved with markets are not efficient.

Market failure is often the justification for political intervention in the marketplace (for one view, see Bator, 1958, section V). The standard argument is that if market allocations are inefficient, everyone can and should be made better off. To understand the feasibility and desirability of such Pareto-improving interventions, we must achieve a deeper understanding of the sources of market failure. Since each must be due to the failure of at least one of the three conditions of the first theorem, we will consider those conditions one at a time.

The first condition requires there to be enough markets. Although there are no definitive guidelines as to what constitutes ‘enough’, the general principle is that if any actor in the economy cares about something that also involves an interaction with at least one other actor, then there should be a market for that something; it should have a price (Arrow, 1969). This is true whether the something is consumption of bread, consumption of the smoke from a factory, or the amount of national defence. The first of these examples is a standard private good, the second is an externality, and the third is a public good. All need to be priced if we are to achieve a Pareto-optimal allocation of resources; without these markets, actors may be unable to inform others about mutually beneficial trades which can leave both better off.

The informational role of markets is clearly highlighted by a classic example of market failure analysed by Scitovsky (1954). In this example, a steel industry, which must decide now whether to operate, will be profitable if and only if a railway industry begins operations within five years. The railway industry will be profitable if and only if the steel industry is operating when the railway industry begins its own operations. Clearly each cares about the other and it is efficient for each to operate; the steel industry begins today and the railway industry begins later. Nevertheless, if there are only spot markets for steel, the railway industry cannot easily inform the steel industry of its interests through the marketplace. This inability to communicate desirable interactions and to coordinate timing is an example of market failure and has been used as a justification for public involvement in development efforts; a justification for national planning. However, if we correctly recognize that there are simply too few markets, we can easily find another solution by creating a futures market for steel. If the railway industry is able to pay today for delivery of steel at some specified date in the future then both steel and railway industries are able to make the other aware of their interests through the marketplace. It is easy to show that as long as agents behave competitively and equilibrium exists, the addition of futures markets will solve this type of market failure.

A completely different example of the informational role of markets arises when actors in the marketplace are asymmetrically informed about the true state of an uncertain world. The classic example involves securities markets where insiders may know something that outsiders do not. Even if it is important and potentially profitable for the uninformed actor to know the information held by the informed actor, there may not be enough markets to generate an efficient allocation of resources. To see this most clearly, suppose there are only two possible states of the world. Further, suppose there are two consumers, one of whom knows the true state and one of whom thinks each state is equally likely. If the only markets that exist are markets for physical commodities, then the equilibrium allocation will not in general be Pareto optimal. One solution is to create a contingent claims market. An ‘insurance’ contract can be created in which delivery and acceptance of a specified amount of the commodity is contingent on the true state of the world. Assuming both parties can, *ex post*, mutually verify which is indeed the true state of the world, if both behave competitively and an equilibrium allocation exists, it will be

Pareto optimal, given the information structure. A more general and precise version of this theorem can be found in Radner (1968).

Analysing this example further we note that in equilibrium the prices of commodities in the state that is not true will be close to or equal to zero, since at positive prices the informed actor will always be willing to supply an infinite amount contingent on the false state, knowing delivery will be unnecessary. If the uninformed actor is clever and realizes that prices will behave this way in equilibrium then he can become informed simply by observing which contingency prices are zero. If he then uses this information, which has been freely provided by the market, the equilibrium will be Pareto optimal under full information. In a very simple form, this is the idea behind rational expectations (see Muth, 1961).

With clever competitive actors, it may not be necessary to create all markets in order to achieve a Pareto-efficient equilibrium allocation.

Completing markets seems to be an easy technique to correct market failure. The suggestions that taxes and subsidies (Pigou, 1932) or property rights reassignments (Coase, 1960) can cure market failure follow directly from this observation. However, an unintended consequence can sometimes occur after the creation of these markets. In some cases, adding more markets may cause conditions (2) and (3) of the first theorem to be false. Curing one form of market failure can lead to another. To understand how this happens and how the second condition requiring competitive behaviour can be affected, consider the informed consumer in our previous example. If he realizes that the uninformed consumer is going to make inferences based indirectly on his actions then he should not behave competitively because he could do better by pretending to be uninformed. He can, by strategically limiting the supply of information of which he is the monopoly holder, do better than if he behaved competitively. It is only his willingness to supply infinite amounts of the commodity in the false state that gives away his knowledge. Supplying only a little commodity contingent on that (false) state in return for a small payment today would not allow the uninformed agent to infer anything and would allow the informed agent to make a profit from his monopoly position. This is not very different from the standard example of a violation of condition (2), monopoly supply of a commodity.

A different example of this phenomenon of unintended outcomes arises when markets are created to allocate public goods. It is now well known that the introduction of personal, Lindahl prices to price individual demands for a public good does indeed lead to Pareto-optimal allocations if consumers behave competitively (see Foley, 1970). However, under this scheme, each agent becomes a monopsonist in one of the created markets and, therefore, has an incentive to understate demand and not to take prices as given. This is the phenomenon of ‘free riding’, often alluded to as the reason why the creation of markets may not be a viable solution to market failure. To understand why, let us now examine the second condition of the first theorem in more detail.

The second condition of the first theorem about market success is that all actors in the marketplace behave competitively. This means that each must act as if they cannot affect prices and, given prices, as if they follow optimizing behaviour. Consumers maximize preferences subject to budget constraints and producers maximize profits, each taking prices as fixed parameters. This condition will be violated when actors can affect the values that equilibrium prices take and in so doing be better off. The standard example of market failure due to a violation of this condition is monopoly, in which one actor is the sole supplier of an output. By artificially restricting supply, this actor can cause higher prices and make himself or herself better off even though the resulting equilibrium allocation will be inefficient.

Can we correct market failure due to non-competitive behaviour? To find an answer let us first isolate

those conditions under which agents find it in their interests to follow competitive behaviour. The work of Roberts and Postlewaite (1976) has established that if each agent holds only a small amount of resources relative to the aggregate available, then they will usually be unable to manipulate prices in any significant way and will act as price takers. It is the depth of the market that is important. This is also true when the commodity is information. If each agent is informationally small, in the sense that he either knows very little or what he does know is of little importance to others, then he loses little by behaving competitively (see Postlewaite and Schmeidler, 1986). On the other hand, if he is informationally important, as in the earlier example, he may have an incentive to behave non-competitively. The key is the size of the agent's resources, both real and informational, relative to the market.

The solution to market failure from non-competitive behaviour then seems to be to ensure that all agents are both resource and informationally small. Of course this must be accomplished through direct intervention as in the antitrust laws and the securities market regulations of the United States and may not be feasible. For example, it may not be possible to correct this type of market failure by simply telling agents to behave competitively. In such an attempt, one would try to enforce a public policy that all firms must charge prices equal to the marginal cost of output. But, unless the costs and production technology of the firm can be directly monitored, a monopolist can easily act as if he were setting price equal to marginal cost while using a false cost curve. It would be impossible for an outside observer to distinguish this non-competitive behaviour from competitive behaviour without directly monitoring the cost curve. If the monopolist were a consumer whose preferences were unobservable, then even monitoring would not help. In general, market failure from non-competitive behaviour is difficult to correct while still retaining markets. We will hint at some alternatives below.

Expansion of the number of markets can also lead to violations of the third condition of the first theorem. For illustration we consider three examples. The first and simplest of these is the case of increasing returns to scale in production. The classic case is a product that requires a fixed set-up cost and a constant marginal cost to produce. (More generally we could consider non-convex production possibilities sets.) If the firm acts competitively in this industry and if the price is above marginal cost the firm will supply an infinite amount. If the price is at or below marginal cost the firm will produce nothing. If the consumers’ quantity demand is positive and finite at a price equal to marginal cost, then there is no price such that supply equals demand. Equilibrium does not exist. The real implication of this situation is not that markets do not equilibrate or that trade does not take place, it is that a natural monopoly exists. There is room for at most one efficient firm in this industry. Again it is the assumption of competitive behaviour that is ultimately violated.

The next example, due to Starrett (1972), involves an external diseconomy. Suppose there is an upstream firm that pollutes the water and a downstream firm that requires clean water as an input into its production process. It is easy to show that if such a diseconomy exists and if the downstream firm always has the option of inaction (that is, it can use no inputs to produce no outputs at zero cost), then the aggregate production possibilities set of the economy when expanded to allow enough markets cannot be convex (see Ledyard, 1976 for a formal proof). If the production possibilities set of the economy is non-convex, then, as in the last example, it is possible that a competitive equilibrium will not exist. Expansion of the number of markets to solve the inefficiencies due to external diseconomies can lead to a situation in which there is no competitive equilibrium.

The last example, first observed by Green (1977) and Kreps (1977), arises in situations of asymmetric information. Recall the earlier example in which one agent was fully informed about the state of the world while the other thought each state was equally likely. Suppose preferences and endowments in each state are such that if both know the state then the equilibrium prices in each state are the same. Further, suppose that if the uninformed agent makes no inferences about the state from the other's behaviour then there will be different prices in each state. Then no (rational expectations) equilibrium will exist. If the informed agent tries to make inferences the prices will not inform him, and if the uninformed agent does not try to make inferences the prices will inform him. Further, it is fairly easy to show that if a market for information could be created (ignoring incentives to hide information) the resulting possibilities set is in general non-convex. In either case there is no equilibrium.

Most examples of non-existence of equilibrium seem to lead inevitably to non-competitive behaviour. In our example of non-existence due to informational asymmetries, it is natural for the informed agent to behave as a monopolist with respect to that information. In the example of the diseconomy, if a market is created between the upstream and the downstream firm, each becomes a monopoly. If there is a single polluter and many pollutees, the polluter holds a position similar to a monopsony. The non-existence problem due to the fundamental non-convexity caused by the use of markets to eliminate external diseconomies is simply finessed by one or more of the participants assuming non-competitive behaviour. An outcome occurs but it is not competitive and, therefore, not efficient.

Market failure, the inefficient allocation of resources with markets, can occur if there are too few markets, non-competitive behaviour, or non-existence problems. Many suggested solutions for market failure, such as tax-subsidy schemes, property rights assignments, and special pricing arrangements, are simply devices for the creation of more markets. If this can be done in a way that avoids non-convexities and ensures depth of participation, then the remedy can be beneficial and the new allocation should be efficient. On the other hand, if the addition of markets creates either non-convexities or shallow participation, then attempts to cure market failure from too few markets will simply lead to market failure from monopolistic behaviour. Market failure in this latter situation is fundamental. Examples are natural monopolies, external diseconomies, public goods and informational monopolies. If one wants to achieve efficient allocations of resources in the presence of such fundamental failures one must accept self-interested behaviour and explore non-market alternatives. A literature using this approach, sometimes called implementation theory and sometimes called mechanism design theory, was initiated by Hurwicz (1972) and is surveyed in Groves and Ledyard (1986). More recent results can be found at mechanism design and mechanism design (new developments).

See Also

- incentive compatibility
- incomplete contracts
- incomplete markets
- mechanism design
- mechanism design (new developments)
- Pareto efficiency
- welfare economics

Bibliography

Arrow, K. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K. 1969. The organization of economic activity: issues pertinent to the choice of market versus non-market allocation. In Joint Economic Committee, *The Analysis and Evaluation of Public Expenditures: The PPB System*, Washington, DC: Government Printing Office.

Bator, F. 1958. The anatomy of market failure. *Quarterly Journal of Economics*, 351–79.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. Cowles Foundation Monograph No. 17. New York: Wiley.

Foley, D. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38, 66–72.

Green, J. 1977. The nonexistence of informational equilibria. *Review of Economic Studies* 44, 451–63.

Groves, T. and Ledyard, J. 1986. Incentive compatibility ten years later. In *Information, Incentives, and Economic Mechanisms*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and Organization*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Kreps, D. 1977. A note on ‘fullfilled expectations’ equilibria. *Journal of Economic Theory* 14, 32–43.

Ledyard, J. 1976. Discussion of ‘on the nature of externalities’. In *Theory and Measurement of Economic Externalities*, ed. S. Lin. New York: Academic Press.

Muth, J. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Pigou, A. 1932. *The Economics of Welfare*, 4th edn. New York: Macmillan.

Postlewaite, A. and Schmeidler, D. 1986. Differential information and strategic behavior in economic environments: a general equilibrium approach. In *Information, Incentives, and Economic Mechanisms*,

ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36, 31–58.

Roberts, J. and Postlewaite, A. 1976. The incentives for price-taking behavior in large exchange economies. *Econometrica* 44, 115–27.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62, 70–82.

Starrett, D. 1972. Fundamental non-convexities in the theory of externalities. *Journal of Economic Theory* 4, 180–99.

How to cite this article

Ledyard, John O. "market failure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000056> doi:10.1057/9780230226203.1029

The New Palgrave Dictionary of Economics Online

market institutions

John McMillan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Market-supporting institutions ensure that property rights are respected, that people can be trusted to live up to their promises, that externalities are held in check, that competition is fostered, and that information flows smoothly. Evidence is reviewed here on some market institutions: property rights and contracting with and without the law, and mechanisms to sustain information flow in markets.

Keywords

adverse selection; Akerlof, G.; asymmetric information; Coase, R.; contracting; equity market; fisheries; Grameen Bank; group lending; India; information; information transmission; land titles; market institutions; markets for lemons; micro-credit; moral hazard; non-price information; North, D.; property rights; screening; signalling; trade credit; transaction costs

Article

In order to work as they should, markets need institutions. Defining the rules of the game, institutions consist of the constraints, formal and informal, on economic and political actors (North, 1991). Market institutions serve to limit transaction costs: the time and money spent locating trading partners, comparing their prices, evaluating the quality of the goods for sale, negotiating agreements, monitoring performance and settling disputes (McMillan, 2002).

The notion that institutions matter is as old as the study of economics. For markets to create gains from trade, as Adam Smith recognized, the state must define property rights and enforce contracts.

That institutions matter is also one of the chief insights from modern economics. In the presence of informational asymmetries, markets can falter. If buyer and seller have different information about the item to be exchanged, a ‘lemons market’ may arise. Unable to distinguish high-quality goods, buyers may be unwilling to pay a price that elicits supply of anything other than low-quality items. Potential gains from trade go unrealized (Akerlof, 1970). When information is distributed unevenly – as is

ubiquitous in the real world of economics, even if most of the textbooks have yet to bring it on board – prices do not incorporate all relevant information, and so non-price information is needed (Spence, 1973; Rothschild and Stiglitz, 1976). Limiting the inefficiencies from informational asymmetries requires mechanisms for signalling and screening: devices like reputation, warranties and credentials, as well as in some cases government-set rules and regulations. A more nuanced view of market processes is called for than the institution-free textbook account of price equilibration via supply and demand.
Evidence on the role of market-supporting institutions is accumulating. Much of the evidence comes from developing countries and countries in transition from communist central planning. Where markets work smoothly, in affluent countries, the market-supporting institutions are almost invisible. It is hard to find evidence of lemons markets in a country like the United States, because institutional solutions have evolved. By contrast, where markets work badly, in poor countries, the absence of institutions is conspicuous (Klitgaard, 1991). A few examples are given in what follows.

Property rights and contracting

Institutional innovation sometimes occurs even in affluent countries. An experiment in property rights has arisen in fisheries. Worldwide, fisheries are in crisis. Overfishing results from an externality: the costs of any one fisher's taking too many fish are mostly borne by others. Applying the idea of Ronald Coase (1960) of defining property rights to solve an externality, the New Zealand government has created, essentially, property rights in the fish. Fishers are assigned quotas that define, by species, their allowable fish catch. The quotas are tradable, so they end up with those fishers with the highest willingness to pay, which probably leads to an efficient allocation. Property rights in fish do not come for free, however, but require extensive, costly government monitoring (Grafton, Squires and Fox, 2000). Military aircraft patrol the oceans. Each step of every single fish's journey from landing to final sale is documented, with catch reports, buyers’ receipts, cold-storage records and export invoices being collated. Fishery inspectors police breaches. The costs of overseeing the quotas have yielded a return, as fish stocks have been successfully conserved.
Another property-rights experiment has occurred in residential land. In cities in every developing country there are squatters, poor people living on land to which they hold no legal rights. Ad hoc property rights exist even in the absence of formal legal protections, as neighbourhood associations and the squatters themselves guard the land. However, the inability to appeal to the law brings some inefficiencies. Hernando de Soto (2000) argued that, if the impoverished squatters held land titles, they would acquire access to capital markets, because they would then have collateral to offer. In Peru, following de Soto's advocacy, over a million squatter households were granted title to the land they occupied. The effects of this huge inauguration of property rights showed up, unexpectedly, not in the capital market but in the labour market. Householders’ borrowing increased little, but hours worked outside the home by adult household members increased and hours worked by their children decreased (Field, 2003; Field and Torero, 2004). Without land titles, householders stayed at home to watch over their property, sending their children out to work. Holding land titles, they felt secure enough to enter the workforce. Establishing the market institution brought instant welfare gains. However, the gains came in an unforeseen form, illustrating the difficulty in general of anticipating the effects of institutional reform (McMillan, 2004).

With contracting, as with property rights, informal substitutes operate in the absence of formal institutions. Small firms make deals with each other and get finance, using personal networks and ongoing relationships to substitute for missing laws of contract and using retained earnings and trade credit to make up for a lack of access to financial markets (Fafchamps, 2004; McMillan and Woodruff, 2002). Large firms also can prosper without institutions, coping instead by cultivating favours from politicians. Where the lack of institutions shows up is for small firms wishing to grow. Needing to make large, discrete investments, they can no longer rely on retained earnings and trade credit, so they may be unable to grow if the financial market is underdeveloped. Needing to deal with increasing numbers of trading partners, they cannot continue to rely on personal connections but must start to use the law of contract. The firm-size distribution in a typical developing country shows a missing middle, with a lot of employment in tiny firms and quite a lot in large firms, but not much in mid-sized firms (Snodgrass and Biggs, 1996). The missing middle is a symptom of weak legal and regulatory institutions.

Information transmission

An archetypical lemons market existed in India in the 1970s (Klitgaard, 1991). Quality fresh milk was hard to find because vendors routinely watered it down. Buyers could not assess the milk's butterfat content, and so the low-quality milk drove out the high-quality milk. Launching a campaign against adulterated milk, the National Dairy Development Board provided inexpensive machines to measure butterfat content as the milk moved from farmer to wholesaler to vendor. It also set up payment schemes making the price of milk reflect its measured quality and created brand names to give buyers trust in what they were getting. As a result of this coordinated initiative, quality improved and consumption rose.
The loan market is impeded by information asymmetries: both adverse selection (a lender may find it hard to distinguish whether any given loan applicant is a good credit risk) and moral hazard (a borrower, having received a loan, may have an incentive to default). Since these transaction costs are proportionately larger for small than for large loans, small lenders often pay exorbitant interest rates or are frozen out of the loan market. In Bangladesh's Grameen Bank and other microcredit banks, tiny loans are made to poor people via groups of borrowers. Each group member is held responsible for any other member's loan. Being neighbours, the group members know each other's business better than any banker, can monitor each other's use of the loans and can invoke social sanctions to discipline defaulters. Group lending is an elegant solution to the loan market's informational asymmetries.
The equity market relies heavily on institutions. For shareholders, who lack information about the firm's affairs, evaluating managers is difficult, and so a lemons market may arise. In many countries, lax oversights allow controlling shareholders to expropriate minority shareholders (Johnson et al., 2000). If the rules governing the financial markets are inadequate, investors are reluctant to buy stocks because they are unwilling to trust managers, and so firms do not get the finance they need. A well-functioning equity market relies on a complex set of interrelated institutions, formal and informal, to foster information flow (Black, 2001). First, reputations for honest dealings must be built up by auditors, law firms, investment banks and the business press. Second, there are self-regulating private-sector bodies such as industry associations as well as the stock exchange, with its rules on listing firms’ financial reporting and its sanction of delisting. Third, the equity market rests on state-provided mechanisms: not only laws requiring that investors receive accurate data, but also an activist regulator. The law's

transaction costs (Glaeser and Shleifer, 2003) mean that a regulator supplements the courts in setting and enforcing the rules of the game.

Conclusion

Market-supporting institutions ensure that property rights are respected, that people can be trusted to live up to their promises, that externalities are held in check, that competition is fostered and that information flows smoothly (McMillan, 2002). Without institutions, the promise of efficient markets goes unrealized.

See Also

- Akerlof, George Arthur
- Arrow, Kenneth Joseph
- development economics
- growth and institutions
- institutionalism, old
- micro-credit
- property law, economics and
- search theory (new perspectives)
- Spence, A. Michael
- Stiglitz, Joseph E.

Bibliography

Akerlof, G. 1970. The market for ‘lemons’. *Quarterly Journal of Economics* 84, 488–500.

Black, B. 2001. The legal and institutional preconditions for strong stock markets. *UCLA Law Review* 48, 781–855.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

de Soto, Hernando. 2000. *The Mystery of Capital*. New York: Basic Books.

Fafchamps, M. 2004. *Market Institutions in Sub-Saharan Africa*. Cambridge, MA: MIT Press.

Field, E. 2003. Entitled to work. Unpublished manuscript. Cambridge, MA: Harvard University.

Field, E. and Torero, M. 2004. Do property titles increase credit access among the urban poor? Unpublished manuscript. Cambridge, MA: Harvard University.

Glaeser, E. and Shleifer, A. 2003. The rise of the regulatory state. *Journal of Economic Literature* 41, 401–25.

Grafton, R., Squires, D. and Fox, K. J. 2000. Private property and economic efficiency. *Journal of Law and Economics* 43, 679–714.

Johnson, S., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. 2000. Tunneling. *American Economic Review Papers and Proceedings* 90, 22–7.

Klitgaard, R. 1991. *Adjusting to Reality*. San Francisco: ICS Press.

McMillan, J. 2002. *Reinventing the Bazaar*. New York: Norton.

McMillan, J. 2004. Avoid hubris. *Finance and Development* 41, 34–7.

McMillan, J. and Woodruff, C. 2002. The central role of entrepreneurs in transition economies. *Journal of Economic Perspectives* 16, 153–70.

North, D. 1991. Institutions. *Journal of Economic Perspectives* 5, 97–112.

Rothschild, M. and Stiglitz, J.E. 1976. Equilibrium in competitive insurance markets. *Quarterly Journal of Economics* 90, 629–50.

Snodgrass, D. and Biggs, T. 1996. *Industrialization and the Small Firm*. San Francisco: ICS Press.

Spence, A. 1973. Job market signaling. *Quarterly Journal of Economics* 87, 355–74.

How to cite this article

McMillan, John. "market institutions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000347> doi:10.1057/9780230226203.1030

The New Palgrave Dictionary of Economics Online

market microstructure

Maureen O'Hara

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Market microstructure research uses the rules and trading protocols of markets to analyse price formation in asset markets. Microstructure research shows how markets provide liquidity and price discovery, and how prices come to reflect information. Of particular importance to this process is how market participants learn from market data. Microstructure researchers often consider issues related to market structure, in particular how changing features of the market affect the price process. Empirical microstructure research uses high-frequency data-sets, and develops statistical approaches to deal with such data.

Keywords

adverse selection; asset pricing; asymmetric information; auto-conditional duration model; Bayes' rule; bid–ask markets; bonds; censored sampling; electronic commerce; market microstructure; inventory models; inventory risk; liquidity; market structure; noise traders; options; price discovery; rational expectations; sequential models; spreads

Article

Market microstructure studies the behaviour and formation of prices in asset markets. Whereas economic analyses of price formation generally abstract from any particular price-setting mechanisms, market microstructure relies on the specific rules and protocols of markets to analyse how prices are determined. This focus on the microstructure of the market provides insights into how the design of markets affects the price process, detailing both how individual prices are determined and how those prices evolve over time. Such insights are useful for a wide range of issues in asset pricing, as well as for guiding econometric investigations of high frequency data. In addition, microstructure research analyses structural issues in securities trading, such as the role and function of exchanges, the optimal design of trading systems, and the optimal regulation of securities markets.

Fundamental to microstructure research is the realization that asset prices are set in actual markets, and not by fictional auctioneers. Thus, while the forces of supply and demand ultimately underlie all asset prices, the specific formation and evolution of prices is much more complex. Buyers and sellers, for example, need not arrive synchronously, making the determination of a market-clearing price at a point in time problematic. When traders do arrive at markets, they may also face a range of market frictions such as transactions costs, search costs and the like (see Stoll, 2001). Furthermore, the value of assets may change over time, with some traders potentially knowing more about future values than other traders. Markets facilitate the trading of assets by providing liquidity and price discovery, and how they do so depends on the rules and structure of the market (see O'Hara, 2003).

Canonical models in microstructure

Early microstructure models focused on the specific market structure found in organized stock markets. In such markets, a designated market-maker or specialist quotes prices to buy or sell units of the asset. By serving as counter-party to buyers and sellers, the market-maker solves the asynchronicity problem noted above by standing ready to provide liquidity on either side of the market. The market-maker earns the ‘spread’, or the difference between the price at which he buys shares (the bid) and the price at which he will sell shares (the ask). In return, however, the market-maker has to bear inventory risk, essentially going long when traders wish to sell, and short when traders wish to buy.

There is an extensive literature analysing the market-maker's pricing problem in the presence of inventory risk (for a review of models, see O'Hara, 1995). In general, such models assume risk-averse market-makers facing exogenous holding costs in a setting in which all agents are symmetrically informed and ‘true’ asset prices are assumed fixed or, at least, stationary processes. An important feature of the equilibrium is that there is no single price: the price the market-maker sets depends upon whether the trader wishes to buy or sell, and on how much he wishes to trade. Prices change over time in response to the specialist's inventory position, his market power and parameters relating to the supply and demand for the asset. Such inventory models have been extended to a wide variety of market settings such as foreign exchange, bond markets, and options and futures markets. Empirical analyses find substantial support for the predictions of inventory models.

An alternative class of microstructure models considers price-setting when some agents have better information about the asset's true value than do other agents. The impetus for such models was an early paper by Treynor (1971), who noted that traders arriving at the market included those who needed to trade for liquidity reasons, those with better information about the asset's true value, and those who thought they had better information but were in fact incorrect. Treynor conjectured that the market-maker's prices were a balancing act offsetting his losses to the informed traders with his gains from the liquidity and noise traders. Viewed from this perspective, a spread arises naturally in security markets, independent of any inventory or transactions costs explanations. Fisher Black (1986) expanded on this notion to highlight the important role played by noise or liquidity traders in allowing markets to become efficient.

An intriguing implication of this research is that, if some traders do have better information about the asset's true value, then the nature of the order flow can be informative as to future asset values. Consequently, the market-maker's price-setting problem evolves from being a simple balancing of

expected gains and losses to that of learning how to extract information from the order flow. With the market-maker drawing inferences from the order flow, this sets the stage for traders to consider the impact of their trades as well, particularly if they are attempting to profit on private information.

There are two general approaches to modelling price-setting in the presence of asymmetric information, sequential trade models and Kyle (1985) models. Glosten and Milgrom (1985) consider a risk-neutral market-maker facing known populations of informed and uninformed traders, where traders arrive sequentially to the market. The market-maker knows these population parameters, but does not know the identity of any individual trader. The market-maker does know, however, that traders informed of good news will all want to buy, while those informed of bad news will all want to sell. Consequently, the market-maker's conditional expectation of the asset's value also differs with trade direction, and it is these conditional expectations that become his bid and ask prices. Based on the trade that actually occurs, the market-maker updates his beliefs regarding the asset's value using Bayes’ rule. The continued one-sided trading of the informed traders eventually forces prices to the true equilibrium level.

Sequential trade models provide an elegant means to characterize the relation between trades and prices on a tick-by-tick basis. Because the market-maker learns from trades, the evolution of prices depends on the order flow, as does the size and movement of the spread. More complex analyses demonstrate a role for other market information in affecting price behaviour. Trade size, for example, may be informative as informed traders prefer to trade larger rather than smaller amounts (Easley and O'Hara, 1987). The time between trades may also have information content as a signal of the existence of new information, and this, in turn, can impart information content to volume (Easley and O'Hara, 1992). Trade location, trade in correlated assets, and alternative order types can also have information content. Because of their tick-by-tick focus, sequential models are particularly useful for guiding empirical analysis of microstructure data, an issue we will return to shortly.

An alternative modelling approach is a Kyle (1985) model. Kyle models focus on the dual problems facing the market-maker, who must figure out what the informed traders know, and the informed trader, who wishes to exploit his private information for profit. The Kyle model uses a batch-auction framework in which the market-maker sees the aggregated trades of both the informed traders and the noise or liquidity traders, and based on this order flow he sets a single price. The market-maker conjectures a trading strategy for the informed trader that is linear in the asset's true value, while the informed trader conjectures a pricing strategy for the market-maker that is linear in the total order flow. In equilibrium, both conjectures must be correct, a feature typical of rational expectations equilibrium models. As in sequential trade models, the market-maker's price reflects his conditional expected value for the asset, this conditional expectation changes as he learns from the order flow, and prices eventually adjust to true values. Back and Baruch (2004) demonstrate conditions under which the Kyle and Glosten–Milgrom models essentially converge.

An important feature of Kyle models is their ability to characterize the trading strategy of the informed trader. The optimal strategy for the informed trader is essentially to hide his trades in the noise trade, and he varies his trades over time in response to the market-maker's growing precision of his beliefs about the asset's true value. Holden and Subrahmanyam (1992) show that, if there are many informed traders, then their combined trading actions force prices almost instantaneously to true values, a result again reminiscent of rational expectations models. A wide range of research has considered variants of the Kyle model allowing for different types of information structures, for uninformed traders to also act

strategically, and for the market-maker to have differential information.
These two asymmetric information-based modelling approaches allow researchers to address a broad range of issues in the trading of financial assets, and are particularly useful in demonstrating how markets perform their price discovery function. Because market-makers are risk neutral and unconstrained as to their inventory holdings, liquidity issues in these models reflect more difficulties induced by the potential information content of trades, rather than the risk-bearing considerations that arise in inventory models. As both effects are likely to be present in actual markets, a wide range of research has investigated empirically how spreads and price changes are influenced by information, inventory, and the fixed costs of making markets.

Research directions in microstructure research

The growth of financial asset markets worldwide, as well as the increasing availability of high frequency microstructure data from a wide array of markets, has allowed microstructure researchers to investigate a broad range of issues, both empirical and theoretical. I highlight here a few areas that are of particular importance.

Econometrics of high-frequency data

Microstructure data allows researchers to analyse the evolution of prices and market data on a second-by-second basis. Indeed, most microstructure data sets include millions of observations, raising a range of econometric issues. Of particular importance are the periodicity of the data, biases introduced by market structure protocols, optimal statistical models for evaluating the behaviour of prices and spreads, and data sampling issues. Hasbrouck (2006) discusses each of these topics.
Because prices arise only when there are trades, price data is not spaced uniformly throughout the trading day. This introduces a censored sampling problem as prices can be thought of as draws from the true asset value distribution, but where the timing of the draws may not be independent of evolution of the value process itself. Engle and Russell (1998) exploit this insight to develop the auto-conditional duration (ACD) model to analyse the evolution of intra-day volatility. A related problem is sampling across assets, as non-synchronicity of trading may result in price observations that lag true value innovations across stocks. A number of authors have considered the implications of non-synchronous trading for cross-sectional econometric analyses.
A variety of authors also consider the time-series properties of microstructure data, with a particular focus on decomposing price movements into those associated with the value process and those reflecting noise arising from the microstructure such as tick size constraints, bid/ask bounce, price continuity rules, and so on. These econometric issues are particularly important for asset pricing research.

Asset pricing – liquidity and information risk

Microstructure models analyse the liquidity and price discovery roles markets play in asset pricing. Recent research has focused on whether these two market roles also affect asset returns. Amihud and Mendelson (1986) first suggested that liquidity could influence asset returns by affecting an investor's

overall cost of trading. Numerous empirical researchers have investigated whether spreads, a proxy for these liquidity costs, are related to asset returns, but the empirical evidence has been mixed. More recent research by Pastor and Stambaugh (2004) using lagged volume measures of liquidity provides stronger evidence, and the authors propose a liquidity factor to explain asset returns. One reason why this effect may arise is commonality in liquidity. Chordia, Roll and Subrahmanyam (2000) find that liquidity measures appear to vary systematically across stocks, and these effects may be time-varying. Other researchers have found similar commonality effects in bond market liquidity measures.

A second research stream considers the price discovery process, and whether investors require higher returns to hold stocks for which a greater fraction of the available information is private rather than public. Easley, Hvidkjaer and O'Hara (2002) derive measures of information-based trading using a structural microstructure model, and demonstrate that asset returns are explained by these information measures. What generates this effect is the inability to diversify optimally, as uninformed traders always lose to informed traders, who are better able to shift their portfolio weights to reflect true values. Empirical research supports a distinct role for both liquidity and information risk in affecting asset returns.

Electronic markets and trading systems

Microstructure models have typically analysed price-setting on a centralized market with a designated market-maker (or makers). While such a setting corresponds well to an exchange or dealer market, it is less applicable to the wide variety of electronic markets now used to trade many financial assets. Of particular importance are electronic trading systems which rely on the aggregation of limit orders to effectuate trades. Orders to buy and sell at a specific price and quantity are collected in the ‘book’, with price and time priority rules dictating how such orders are handled. At any point in time, a spread exists between the highest (lowest) price at which someone is willing to buy (sell) the asset. In such systems, trades arise when orders cross, imparting an importance to the order decisions of individual traders. Traders face complex decision problems in placing orders due to the uncertainty of execution of any order. Of particular concern is that uninformed traders may face an adverse selection problem in that their trades are more likely to execute when there is new information, causing them to buy when there is bad news and sell when there is good news. This difficulty is further compounded by trading protocols that allow limit orders to ‘sweep the book’ and thereby trigger the execution of many individual orders as the opposite side of a large order. There is a substantial literature looking at the behaviour of such electronic markets, but the complexity of these markets leaves many important issues yet to be resolved.

Market structure

Microstructure research is traditionally concerned with issues related to the design and structure of markets. The rise of new markets and trading technologies has raised a plethora of market structure issues. Of particular interest to many researchers are questions relating to transparency, or what information is available to traders and when can they see it. Bond markets, for example, were traditionally opaque, but new reporting rules have increased their transparency. Numerous authors have investigated how this has changed the liquidity and efficiency of the bond market. Option markets

traditionally faced little competition, but the development of a national options market in the United States, along with the rise of electronic competitors, has changed this market structure. Regulatory changes in the United States and Europe have also dramatically affected market structure in equities, raising questions as to the efficacy of these new rules. Finally, the markets themselves are evolving from member-owned cooperatives to publicly traded firms, raising a host of issues relating to corporate governance and self-regulation. Microstructure research provides a means to evaluate the economic impact of these changes and to suggest alternative structures for the trading of financial assets.

See Also

- adverse selection
- noise traders

Bibliography

Amihud, Y. and Mendelson, H. 1986. Asset pricing and the bid-ask spread. *Journal of Financial Economics* 17, 223–49.

Back, K. and Baruch, S. 2004. Information in securities markets: Kyle meets Glosten and Milgrom. *Econometrica* 72, 433–65.

Black, F. 1989. Noise. *Journal of Finance* 41, 529–43.

Chordia, T., Roll, R. and Subrahmanyam, A. 2000. Commonality in liquidity. *Journal of Financial Economics* 56, 3–28.

Easley, D. and O'Hara, M. 1987. Price, trade size, and information in securities markets. *Journal of Financial Economics* 19, 69–90.

Easley, D. and O'Hara, M. 1992. Time and the process of security price adjustment. *Journal of Finance* 47, 577–605.

Easley, D., Hvidkjaer, S. and O'Hara, M. 2002. Is information risk a determinant of asset prices? *Journal of Finance* 56, 2185–221.

Engle, R. and Russell, J.R. 1998. Autoregressive conditional duration: a new model for irregularly spaced data. *Econometrica* 66, 1127–62.

Glosten, L. and Milgrom, P. 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14, 71–100.

Hasbrouck, J. 2006. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. New York: Oxford Economic Press.

Holden, C.W. and Subrahmanyam, A. 1992. Long-lived private information and imperfect competition. *Journal of Finance* 47, 247–70.

Kyle, A. 1985. Continuous auctions and insider trading. *Econometrica* 53, 1315–35.

O'Hara, M. 1995. *Market Microstructure Theory*. Boston: Blackwell.

O'Hara, M. 2003. Presidential address: liquidity and price discovery. *Journal of Finance* 58, 1335–54.

Pastor, L. and Stambaugh, R. 2004. Liquidity risk and expected stock returns. *Journal of Political Economy*, 111, 642–85.

Stoll, H. 2001. Market frictions. *Journal of Finance* 55, 1479–514.

Treynor, J. 1971. The only game in town. *Financial Analysts Journal* 27, 12–14, 22.

How to cite this article

O'Hara, Maureen. "market microstructure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000364> doi:10.1057/9780230226203.1031

The New Palgrave Dictionary of Economics Online

market period

D.R. Helm

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

expectations; imperfect competition; IS–LM; long run and short run; market period; new classic macroeconomics; perfect foresight; perfect markets; rational expectations; temporary equilibria

Article

The concept of market period was introduced by Marshall to define markets according to the time period over which they extended. It was thus an additional classification of markets to that of location or space (*Principles*, V.i.6). This distinction became the modern textbook one between the short period and the long, reducing Marshall's more complex three-period classification. As he put it,

> we shall find that if the period is short, the supply is limited to the stores which happen to be at hand; if the period is long, the supply will be influenced, more or less, by the cost of producing the commodity in question; and if the period is very long, this cost will in its turn be influenced, more or less, by the cost of producing the labour and material things required for producing the commodity.

Hence the short run is that period for which stocks are constant, the long run that period where price is determined by the costs of production (but factors are constant) and the very long run that period where all factors vary.

The Marshallian market period was, as Hicks pointed out (1965, ch. 5) one of the ways in which Marshall used his ‘static method’. For in the short period, Hicks goes on to say, Marshall could treat the industry as if it were in static equilibrium. Capital, fixed in the short period, is like land in Ricardo, and it earns a rent. In the longer run, the static method breaks down, as capital becomes variable, like labour. The concept of Marshallian short period has been used extensively in the theory of the firm, in terms of short- and long-run equilibria, and the defining of cost curves according to this classification. Harrod, in 1934, linked this Marshallian concept with the new theory of imperfect competition developed by Joan

Robinson (and Chamberlin), to look at the process of imperfect competition and the impact of entry on short- and long-run profit maximization.

The Marshallian short run concept was taken over into macroeconomics by Keynes as one of three components of Marshall's theory which he used to construct the *General Theory* (the others were partial equilibrium and thus exogenous expectations, and the representative firm aggregate which Keynes took over as the economy). However the Keynesian use of market period was not universally adopted in macroeconomics and it is Hicks's much more restrictive concept used in the IS–LM framework which is now much more familiar. The Hicksian ‘week’ is a market period in which fundamentally it is stocks that are constant, while the Keynesian ‘year’ allows for an element of ‘user cost’ whereby the utilization of capital affects the future demand for capital.

The Hicksian week and the concept of temporary equilibrium associated with it were first set out by Hicks in the middle chapters of *Value and Capital* in 1939, and were subsequently revised in an important, somewhat neglected essay entitled ‘Methods of Dynamic Analysis’ in 1956, reprinted in *Money, Interest and Wages* (1982). These concepts formed part of an attempt to construct a theory of dynamics, going beyond Marshall's static analysis. The alternative extreme hypothesis of allowing all factors to vary is a longer-run theory, and forms the basis of general equilibrium and growth theory. Both the Hicksian and Keynesian theories attempt to construct an intermediary period, and for each the corresponding problems were to decide which factors are to be allowed to vary, which to stay constant and what process of adjustment to vary, which to be employed by firms. But once these theoretical assumptions have been made, the individual periods become discrete rather than continuous (as in the longer-run case). Thus a theory of dynamics based on the Hicksian week requires an additional theory by which to link the discrete periods together to form a continuous model. We need to know how to get from one period to another. Both Keynes and Hicks resorted in one way or another to a link via expectations, though both provided what now appear to be inadequate explanations of their formation. The modern new classical theory of macroeconomic equilibrium avoids the short and the long run distinction by appealing both to the perfectibility of markets per se, and to a theory of expectations which is itself based on perfect markets. Thus although the rational expectations approach ‘solves’ the problem of linking market periods, it does so in a way which avoids rather than solves the problems of market period analysis. Perfect markets do not have dynamics with limited time horizons and rigidities and thus, in the rational expectations perfect foresight model, there is really no need for market period analysis. It is clear however that market imperfections in real variables and expectations do exist, and hence that short-run temporary equilibria cannot easily be linked together.

See Also

- long run and short run
- Marshall, Alfred
- reservation price and reservation demand
- temporary equilibrium

Bibliography

Harrod, R. 1934. Doctrines of imperfect competition. *Quarterly Journal of Economics* 48, 442–70.

Hicks, J.R. 1956. Methods of dynamic analysis. In *25 Essays in Honour of Erik Lindahl*, Stockholm. Repr. in J.R. Hicks, *Money, Interest and Wages*. Oxford: Basil Blackwell, 1982.

Hicks, J.R. 1965. The method of Marshall. In *Capital and Growth*, ed. J.R. Hicks. Oxford: Clarendon Press.

Marshall, A. 1920. *Principles of Economics*. 8th edn, London: Macmillan.

How to cite this article

Helm, D. R. "market period." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000063> doi:10.1057/9780230226203.1032

The New Palgrave Dictionary of Economics Online

market power and collusion in laboratory markets

Douglas D. Davis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Despite the robust tendency of laboratory markets to generate competitive outcomes, some market designs deviate persistently from competitive predictions. This article discusses the primary drivers of supra-competitive prices that have been observed in market experiments.

Article

The robustness of competitive market predictions stands as one of the most impressive results in experimental economics. Laboratory markets regularly generate competitive outcomes in environments populated by just two or three sellers. However, as in natural contexts, competitive outcomes do not always emerge. This article reviews results of laboratory markets in which price increases are driven by factors such as the exercise of unilateral market power or by collusion.

Before reviewing the main concepts and contributions in this area, I offer two observations. First, laboratory methods represent an important but limited complement to existing empirical tools for investigating market performance. Given the stark simplicity and limited duration of laboratory markets, experimentalists can aspire to say little about specific naturally occurring markets. Experiments can, however, provide important insights into the behavioural relevance of theories upon which antitrust policies are based.

Second, the trading rules defining negotiations and contracting can exert first-order effects on market competitiveness. For example, markets organized under the double auction trading rules used in many financial exchanges, are much more robustly competitive than markets organized under the posted-offer trading rules used in most retail exchanges: duopoly or even monopoly sellers are less able to increase market prices in double-auction than in posted-offer markets (Davis and Holt, 1993, chs 3, 4; Holt, 1995). Indeed, one of the motivating factors in the emerging field of institutional design was an interest in developing institutional rules that promoted efficient market outcomes.

For specificity I focus here on results from posted-offer markets, primarily because posted-offer markets allow a particularly intuitive illustration of the factors affecting market competitiveness. However, a host of other trading institutions exist, ranging from single and multi-unit auctions, to multi-sided computerized ‘smart’ markets, and again to institutions that exist primarily as theoretical constructs, such as quantity-setting Cournot mechanisms. The competitive implications of each of these institutions must be evaluated independently.

Posted-offer markets and unilateral market power

Unilateral market power is perhaps the most frequently observed reason why prices in laboratory markets deviate from competitive predictions. This market power exists when one or more sellers, acting on their own, find it profitable to raise prices above the competitive level. The supply and demand structures shown in the two panels of Figure 1 illustrate how capacity restrictions can create market power. In each panel, the market consists of three sellers, S1, S2 and S3, each of whom offers four units for sale, under the conditions that two units cost $2.00 and two units cost $3.00. A buyer will purchase a fixed number of units (seven in the left panel or ten in the right panel) at prices less than or equal to $6.00.

Figure 1
Supply and demand arrays for markets without and with unilateral market power

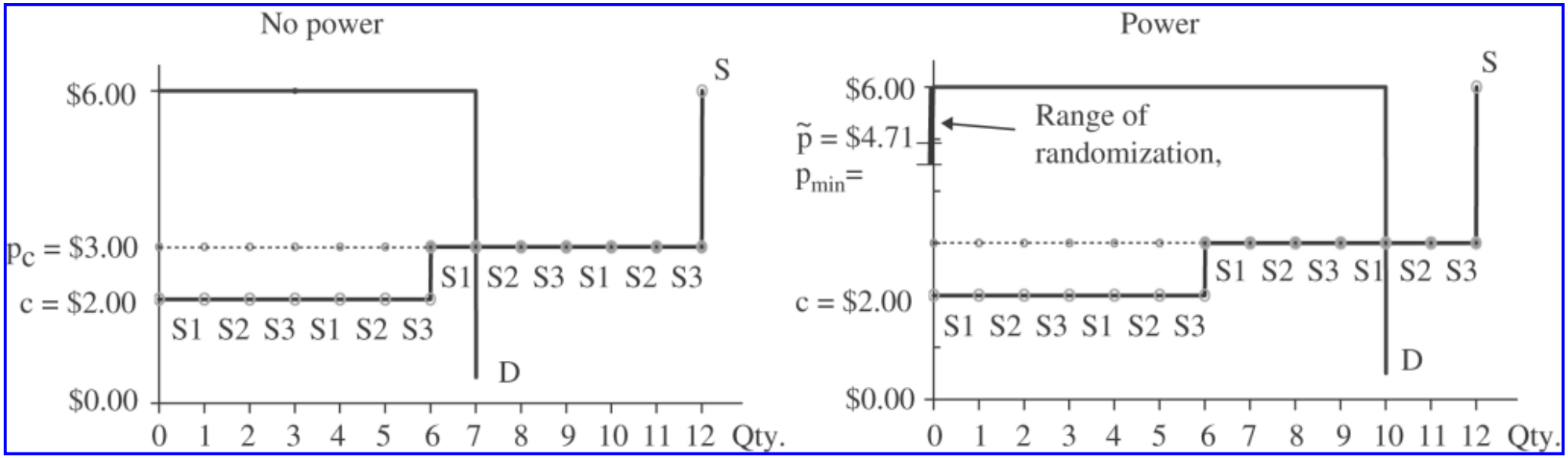

Exchange in these markets proceeds in a number of trading periods. At the outset of each period, sellers simultaneously make price decisions. Production is ‘to order’ in the sense that sellers incur costs only for the units that actually sell. Once all sellers post prices, a simulated fully revealing buyer makes all possible purchases, starting with the least expensive units first. In the case of a tie, the buyer rotates purchases among the tied sellers.

In the market shown in the left panel of Figure 1 the buyer will purchase at most seven units. Given an aggregate supply of 12 units, sellers in this market have no market power: at any common price above $3.00, each seller can increase sales from an expected 2.33 units to four units by posting a price just slightly below the common price. For any vector of heterogeneous prices above $3.00 only the seller posting the lowest price will sell all four units. The seller posting the second highest price will sell three units, while the high-pricing seller will sell nothing. The unique Nash equilibrium for the stage game has each seller posting the competitive price of $3.00, selling 2.33 units in expectation and earning $2.00.

Expanding demand to ten units, as shown in the right panel of Figure 1, limits excess supply, and thus creates market power. Given that the highest price seller is now certain to sell at least two units, the competitive price of $3.00 is no longer a Nash equilibrium for the stage game. At a common price of $3.00 each seller sells 3.33 units (in expectation) and earns $2.00. By posting a price of $6.00, any seller can sell two units and increase earnings to $8.00. A common price of $6.00 is not an equilibrium for the stage game, since any seller would find that deviating from $6.00 increases sales to four units. Sellers have similar incentives to undercut any common price down to a minimum p_{min}=$4.50, where the profits from selling four units as the lowest pricing seller equals earnings at the limit price. The equilibrium for this game involves mixing over the range from $4.50 to $6.00. As shown in the figure, the unique symmetric equilibrium is $4.71.

An extensive series of experiments show that sellers respond to unilateral market power by raising prices. Further, power drives pricing outcomes more powerfully than do changes in the number of sellers. For example, when they reallocated units among five sellers to create market power, Davis and Holt (1994) observed substantial price increases. However, reducing the number of sellers from five to three in a way that held market power conditions fixed, Davis and Holt observed only modest additional price increases. Market power of the sort illustrated in the right panel of Figure 1 has wide applications, ranging from distortions in markets for emissions trading (Godby, 2000) and for electricity transmission (Rassenti, Smith and Wilson, 2003), to price stickiness in the face of aggregate demand shocks (Wilson, 1998).

Tacit collusion

Experimentalists have also observed supra-competitive prices in repeated market games where sellers have no market power. This *tacit collusion* has been observed most frequently in duopolies (for example, Alger, 1987; Fouraker and Seigel, 1963). However, tacit collusion has also been observed in thicker markets where sellers possess no market power. For example, Cason and Williams (1990) observe persistently high prices in a four-seller design similar to that shown as the left panel of Figure 1. Experimentalists often measure tacit collusion as the difference between observed prices and prices consistent with the Nash equilibrium for the market analysed as a stage game. Importantly, other than exceeding equilibrium price predictions, tacitly collusive laboratory outcomes typically exhibit no obvious signs of coordinated activity.

Tacit collusion may coexist with market power. For example, prices in the market power sessions reported by Davis and Holt (1994) were significantly above prices consistent with the equilibrium mixing distribution. In this context, the difference between mean observed prices and the mean of the equilibrium mixing distribution may be reasonably taken as a

measure of tacit collusion.

Tacit collusion is not yet well understood, and isolating the causes of tacit collusion represents an important project for future experimental work. Price signalling activity at least partially explains tacit collusion (for example, Durham et al., 2004). However, evidence suggests that more than price signals and responses may be at play. Dufwenberg and Gneezy (2000) report an experiment where duopolists deviate from the static Nash (competitive) prediction for a game, even when sellers are rematched into different markets after each decision. In such a context price signalling is not possible.

Explicit collusion

Given opportunities to explicitly discuss pricing, laboratory sellers quite persistently organize profit-increasing cartels (Isaac, Ramey and Williams, 1984). However, a capacity to monitor agreements and prevent secret discounts appears critical to the success of these arrangements (Davis and Holt, 1998). Given the illegality of explicit agreements, the more interesting questions regarding explicit collusion concern the capacity of authorities to detect such arrangements through the actions of sellers in the market (Davis and Wilson, 2002).

Other factors affecting pricing

A host of experimental studies indicate that standard ‘facilitating practices’ can contribute to price increases. Experimental studies where supra-competitive prices have been attributed to facilitating practices include ‘most favoured nation’ and ‘meet-or-release’ clauses (Grether and Plott, 1984), non-binding price signals (Holt and Davis, 1990) and multi-market competition (Phillips and Mason, 1991).

Buyer behaviour can also affect market outcomes. When buyer decisions are simulated, details of the purchasing rules can have a large effect on prices (Kruse, 1993). Powerful human buyers can substantially undermine both market power and tacit collusion (Ruffle, 2000). However, the use of real rather than simulated buyers appears to generate more competitive prices even when the human buyers engage in no strategic behaviour (Coursey et al., 1984).

Finally, information conditions and even sellers’ expectations can significantly affect pricing outcomes. For example, Huck, Norman and Oechssler (2000) report that information regarding underlying supply and demand conditions facilitates the exercise of predicted market power (markets are drawn to static Nash predictions). However, information on rival sellers’ profits made markets more competitive in a market where the high-profit seller has the highest market share, so imitation by others will tend to expand quantity and reduce price. Also, in a Cournot context, Huck et al. (2007) report that seller aspirations for increased profits helped consolidated sellers maintain prices substantially above static Nash levels.

See Also

- antitrust enforcement
- Bertrand competition
- experimental economics
- market institutions

Bibliography

Alger, D. 1987. Laboratory tests of equilibrium predictions with disequilibrium price data. *Review of Economic Studies* 54, 105–45.

Cason, T.N. and Williams, A.W. 1990. Competitive equilibrium convergence in posted-offer markets with extreme earnings inequities. *Journal of Economic Behavior and Organization* 14, 331–52.

Coursey, D., Isaac, R.M., Luke, M. and Smith, V.L. 1984. Market contestability in the presence of sunk (entry) costs. *RAND Journal of Economics* 15, 69–84.

Davis, D.D. and Holt, C.A. 1993. *Experimental Economics*. Princeton: Princeton University Press.

Davis, D.D. and Holt, C.A. 1994. Market power and mergers in laboratory markets with posted prices. *RAND Journal of Economics* 25, 467–87.

Davis, D.D. and Holt, C.A. 1998. Conspiracies and secret discounts in laboratory markets. *Economic Journal* 108, 736–56.

Davis, D.D. and Wilson, B. 2002. An experimental investigation of methods for detecting collusion. *Economic Inquiry* 40, 213–30.

Dufwenberg, M. and Gneezy, U. 2000. Price competition and market concentration: an experimental study. *International Journal of Industrial Organization* 18, 7–22.

Durham, Y., McCabe, K., Olson, M.A., Rassenti, S. and Smith, V. 2004. Oligopoly competition in fixed cost environments. *International Journal of Industrial Organization* 22, 147–62.

Fouraker, L.E. and Siegel, S. 1963. *Bargaining Behavior*. New York: McGraw-Hill.

Godby, R. 2000. Market power and emission trading: theory and laboratory results. *Pacific Economic Review* 5, 349–64.

Grether, D.M. and Plott, C.R. 1984. The effects of market practices in oligopolistic markets: an experimental examination of the ethyl case. *Economic Inquiry* 24, 479–507.

Holt, C.A. 1995. Industrial organization: a survey of laboratory research. In *The Handbook of Industrial Organization*, ed. J.H. Kagel and A.E. Roth. Princeton: Princeton University Press.

Holt, C.A. and Davis, D.D. 1990. The effects of non-binding price announcements on posted-offer markets. *Economics Letters* 34, 307–10.

Huck, S., Normann, H. and Oechssler, J. 2000. Does information about competitors' actions increase or decrease competition in experimental oligopoly markets? *International Journal of Industrial Organization* 18, 39–57.

Huck, S., Konrad, K.A., M•ller, W. and Normann, H.T. 2007. The merger paradox and why aspiration levels let it fail in the laboratory. *Economic Journal* 117, 1073–95.

Isaac, R.M., Ramey, V. and Williams, A. 1984. The effects of market organization on conspiracies in restraint of trade. *Journal of Economic Behavior and Organization* 5, 191–222.

Kruse, J.B. 1993. Nash equilibrium and buyer rationing rules: experimental evidence. *Economic Inquiry* 31, 631–66.

Phillips, O.R. and Mason, C.F. 1991. Mutual forbearance in experimental congolomerate markets. *RAND Journal of Economics* 23, 395–414.

Rassenti, S.J., Smith, V.L. and Wilson, B.J. 2003. Controlling market power and price spikes in electricity networks: demand-side bidding. *Proceedings of the National Academy of Sciences* 100, 2998–3003.

Ruffle, B.J. 2000. Some factors affecting demand withholding in posted-offer markets. *Economic Theory* 16, 529–44.

Wilson, B.J. 1998. What collusion? Unilateral market power as a catalyst for countercyclical markups. *Experimental Economics* 1, 133–45.

How to cite this article

Davis, Douglas D. "market power and collusion in laboratory markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?

id=pde2008_E000307> doi:10.1057/9780230226203.1034

The New Palgrave Dictionary of Economics Online

market price

G. Vaggi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cantillon, R.; competition; market price; natural price

Article

The market price, or market value, is defined as the actual price paid for a commodity during a certain period of time, and may be contrasted with the natural or normal price, which is determined by the long-term forces and the permanent causes of the value of commodities (*see* natural price).

The distinction between market price and the intrinsic value of a good can be traced back to the origins of economic science. Before Adam Smith, Richard Cantillon had already analysed the causes which influence the temporary value of a commodity (Hollander, 1973, p. 41).

Many different causes can affect the market price of a commodity and it is difficult to explain the day-to-day changes in its value. However, economic theory has generally singled out the relationship between the demand for a product and its supply on the market as the main force determining the market value. For Smith, the existence of a positive difference between the effectual demand for a commodity and the quantity of it which has been produced and brought to the market leads to a high market price vis-à-vis the natural value, and vice versa. But if there is free competition between producers, market prices cannot be too different from natural prices for a long period of time. Market competition forces lead to the gravitation of market prices around the natural prices. Therefore in classical political economy the two concepts are carefully listed. In particular, the market price is continuously brought towards the natural price.

The concept of market price is an important feature of Adam Smith's description of the competitive mechanism and the way in which it leads to a uniform rate of profit in all sectors of the economy. For instance, when the supply of a commodity falls short of its effectual demand the market price is higher than the natural one, because of the competition between the buyers who are eager to purchase that good (Smith, 1776, pp. 73–4). Either one or more of the three component elements of value – wages, profits and rent – is paid at a rate higher than the natural rate. In a freely competitive economy producers compare their rate of profit with profit rates earned in other activities. Thus entrepreneurs invest their

capital in the sectors which yield the highest rates or profit. This leads to an increase in the output of the commodities whose market prices are higher than natural ones, and vice versa a decrease when market prices are lower than natural values. Therefore the concept of market price is part of Smith's explanation of the changes in output which occur from one production period to another in each sector.

Given the natural price, and the corresponding level of effectual demand, the increase in output leads to a market situation in which more consumers (willing to pay for the good at its natural price) can be satisfied. There is less competition than before between the consumers, and the market price tends to move towards the natural price. Again, the entrepreneurs compare market prices and profit rates in all sectors of the economy and capital will move if the rate of profit is not uniform. Only when all the demand is matched by an equal supply at a market price level which equals the natural one will competition stop; in this situation the market price is exactly the right amount to pay all the components of price at their natural values. The market price depends on excess demand (or supply) on the market at any moment in time, but cannot be too far away from the natural price for a long period of time, because competition tends to bring it towards this level.

Alfred Marshall's distinction between the market and normal value of commodities is similar to Adam Smith's. Normal value is related to the cost production of commodities, while market prices are mainly influenced by utility and demand (Marshall, 1920, pp. 289–90). Marshall also believes that it is difficult to work out a precise theory to determine the market values of commodities; they are affected by too many factors. Marshall argues that there are long and short period forces acting on prices. But he is much more sceptical than Smith about the existence of a precise mechanism, namely competition, which should prevent short-term market prices from moving too far away from the normal price of commodities.

See Also

- natural price

Bibliography

Hollander, S. 1973. *The Economics of Adam Smith*. London: Heinemann.

Marshall, A. 1920. *Principles of Economics*. 8th edn. Reprinted, London: Macmillan, 1972.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Oxford University Press, 1976.

How to cite this article

Vaggi, G. "market price." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/

article?id=pde2008_M000065> doi:10.1057/9780230226203.1035

The New Palgrave Dictionary of Economics Online

market structure

John Sutton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘market structure’ relates to the number and size distribution of firms in a market. Markets dominated by a few large firms are said to be ‘concentrated’. This article offers a brief review of the modern literature that sets out to explain differences in concentration levels across different industries.

Keywords

advertising; barriers to entry; endogenous sunk costs; first mover advantages; Herfindahl index; integer effects; k-firm concentration ratio; learning; market structure; mergers; minimum efficient scale; monopolistic competition; monopoly; multi-stage games; network effects; oligopoly; perfect competition; rate of return; research and development; scale economies

Article

Why is the world market for large commercial jet aircraft dominated by just two firms, while oil tankers are produced by a large number of firms spread over many countries? This is the kind of question addressed in the literature on ‘market structure’, a field once seen as a rather arcane area, in which explanatory theories were weak and in which discussion tended to focus on rival interpretations of ‘statistical regularities’ reported in empirical studies. The most famous of these ‘regularities’ related to a supposed link, across different industries, between the degree to which the industry was dominated by a few large firms (‘concentration’), and some average measure of the rate of return (profit) on fixed assets enjoyed by firms in the industry. (Popular summary measures of concentration include the ‘k-firm concentration ratio’, that is, the share of industry sales revenue accounted for by the top k firms, and the Herfindahl index, defined as the sum of squares of all firms’ market shares). Now the presence of a (positive) relation of this kind would raise the question, ‘why do industries with high rates of profit not attract entry, to the point where such differences are eroded?’ This question was countered in the older literature, following Bain (1956), by appealing to the supposed existence of ‘barriers to entry’ in various

industries. These barriers fell into three categories. The first related to factors intrinsic to the industry's methods of production (‘scale economies’). If the average cost of production falls sharply as output rises to a certain level, then we might regard that level as a ‘minimum efficient scale’, and postulate that the industry is large enough to accommodate only a small number of firms of this size. This point was, and remains, uncontroversial. The second category related to institutional barriers associated with legal or regulatory impediments, or poor access to financial markets, and so on, but, while barriers of this category may be important in some industries and for some countries, they are probably of secondary relevance to the general run of industries in market economies. The third type of barrier related to the role played by advertising and R&D, and it is here that some serious difficulties arise, a point to which we turn in what follows.

The series of ideas just set out came to be known as the Bain paradigm, or the Structure–Conduct–Performance paradigm. Expressed briefly, this view held that a more concentrated structure, however sustained, allowed firms to operate less intensive forms of price competition (‘conduct’), and this in turn led to high profits (‘performance’). This view was seriously undermined in the 1980s as a result of two developments in the literature. The first of these developments was empirical: it became clear, in the light of new empirical studies, that the claim for a positive relationship between concentration and profitability was not well-founded. (For a review of the evidence, see Schmalensee, 1989.) The second development was theoretical: it was clear that any successful explanation of differences in concentration across industries could not rely solely on ‘scale economies’ and ‘institutional barriers’; the role played by advertising and R&D in raising the stakes required of entrants to an industry seemed crucial. But here a problem arises: the levels of advertising and R&D, unlike the degree of scale economies, are matters that are under the control of the firms themselves. The levels of expenditure firms undertake in these areas are ground out as part of the competitive process – and so we cannot treat their levels as a given, and claim that, when we observe a high ratio of advertising and/or R&D to industry sales revenue, this constitutes a ‘barrier to entry’ that explains the industry's high level of concentration. Rather, an explanation of market structure must explain *both* the level of concentration *and* the levels of advertising and R&D intensity. The ‘given’ that distinguishes one industry from another must not be the observed (or ‘equilibrium’) *level* of advertising or R&D, but rather the underlying (industry-specific) relationship between any firm's level of spending on these fixed outlays and the resulting benefit (‘perceived product quality’, say, or, more generally, any effect leading to an outward shift in the firm's demand schedule or a fall in its unit cost of production).

These problems with the older literature led from the late 1980s onwards to the development of a new literature on market structure. (See, for example, Dasgupta and Stiglitz, 1980; Shaked and Sutton, 1986; Sutton, 1991; 1998. A full technical review of the literature will be found in Sutton, 2007.) The point of departure of this literature lies in modelling the evolution of structure by reference to a ‘free entry’ model, in which any one of a number of potential entrants is free to enter the industry, and to choose its level of outlays on advertising, R&D, and so on, in the light of the choices made by its rivals.

The modern game-theoretic literature

The models used in the modern literature take the form of ‘multi-stage games’. In the simplest example, a firm decides, at stage 1, to enter (and pay some positive, minimal, entry fee whose size is a given, and

which can, for example, be interpreted as the cost of building a production plant of ‘minimum efficient scale’). At stage 2, each firm, knowing the number of firms that have entered, chooses its level(s) of advertising and/or R&D. Its choices will depend inter alia on the (industry-specific) degree of effectiveness of these expenditures in influencing consumer demand for its product(s). In ‘commodity’ type industries, where this effectiveness is very low, these outlays will be close to zero. Finally, in stage 3, firms compete in price, taking as given the attributes of their respective products, and they realize corresponding levels of (gross) profits. (It is always assumed that firms have constant marginal costs of production, and that they face downward-sloping demand schedules.)

The central idea that emerges from these (‘endogenous sunk costs’) models is as follows: as the size of the market increases (in the sense of having a larger number of consumers, so that each firm's demand schedule shifts outwards) the industry may adjust to this in two ways: the number of firms may rise (‘entry’), and/or the spending level per firm on advertising, R&D, and so on, may rise – because, in the absence of a proportional rise in the number of firms, each firm now enjoys a higher level of demand, and so the marginal return it gains from being able to charge a given price premium for a higher-quality product rises (‘escalation’). Now the degree to which one or other of these effects operates depends inter alia on the effectiveness of advertising and/or R&D. It also depends on the degree to which high-quality products can draw customers away from rival products of a lower quality. Suppose, for example, that products differ not only in quality, but in other attributes also, and that customers differ in their preferences over these latter attributes. Then it will be correspondingly harder for a firm that raises its ‘perceived quality’ level to attract sales from rivals. An example may be helpful here: consider, for instance, the market for flowmeters. These devices are used to measure the rate of flow of liquids, and they come in a large number of types. An increase in R&D spending by a producer of ‘electromagnetic’ flowmeters will have only a limited impact in drawing consumers away from ‘ultrasonic’ flowmeters, since the latter type of meter has attributes that makes it better suited than the electromagnetic type in certain applications. By way of contrast, consider the case of the (civil) aircraft industry, as it developed since the late 1920s. At that period, there were many types of plane in operation (monoplanes/biplanes, metal/wood construction, land/seaplanes, and so on). Yet all makers faced a market where all buyers (airlines) sought to achieve the same objective: to minimize the carrying cost per passenger mile. As soon as it became clear which type of design best achieved this single aim, plane-makers converged on the solution (an all metal monoplane with a cantilever wing design, following the Douglas DC3). Thereafter, technical developments were focused on pushing forward the performance of this type of plane, and, as plane-makers escalated their efforts in this direction, the stakes required to keep up with rival firms’ innovations rose, and there was a ‘shake-out’ of all but a handful of firms. This story was repeated at the dawn of the jet age in the 1950s. Here a growing world market led, not to the entry of new plane-makers, but to an increasing flow of development outlays by the surviving firms, so that only Boeing and Airbus remain in the wide-body commercial jet business today. (For the details of this story, and the rise of Airbus, see Sutton, 1998, ch. 15.)

Where does this leave us? The kind of market profiles that emerge are these: (*a*) those with high R&D outlays and high global concentration (for example, wide-body commercial jets); (*b*) those with high R&D outlays, low concentration and a fragmented set of distinct product categories (for example, flowmeters); and (*c*) those with low R&D spending, where, once the size of the market is large, the level of concentration may become arbitrarily low.

Within advertising-intensive industries, a simpler picture emerges. Here, the fact that a firm can use a

single brand to span a range of product types in a market means that we can, with few exceptions, define markets in a way that avoids the complication posed by the presence of sub-markets for distinct product types, of the kind we encountered in the flowmeter example above. Here, the theory leads to a very simple prediction: if we take a cross-section of markets of different sizes (by looking, say, at a single industry across a number of countries of different sizes), then we will find a sharp difference in the market size–concentration relationship as between the 'advertising-intensive' industries and a control group of industries in which advertising plays an insignificant role. In the latter group, very low levels of concentration may be reached as market size increases. In the former group, concentration levels will necessarily remain above some critical level in all countries, no matter how large the size of the country (the 'non-convergence' property). This prediction, and related predictions of the theory, have been widely tested over the past decade and appears to be closely in line with what is found in the data (for a review, see Sutton, 2007).

One further comment is called for in relation to this prediction: what is predicted is not an actual or equilibrium level of concentration, but rather a *lower bound* to the level of concentration that can emerge under given circumstances. It is intrinsic to models of this kind that a range of different outcomes is possible, depending on such factors as the form of the entry process (simultaneous, sequential, and so on). The most graphic illustration of this point comes from thinking about the pattern of ownership of plants spread over some geographic region large enough to support many plants. There will be a 'fragmented' equilibrium in which every plant is owned by a different firm, and there will be other equilibria in which the number of plants will remain (roughly) the same, but several of these plants will be owned by the same firm. In other words, a range of outcomes can arise as equilibria, depending on the form of the entry process and the nature of price competition, and the theoretical focus of interest lies in asking, not about the actual outcome, but about the range of possible outcomes, or, more specifically, about the lower bound to the level of concentration that can emerge.

Extensions: learning effects and network externalities

Two further ('dynamic') mechanisms play a role in explaining high levels of concentration. First, if each firm's unit cost level falls over time as a function of its cumulated volume of output to date, then an early entrant may build a dominant market position by setting an initial low price – possibly below its current unit variable cost of production – with a view to achieving a high output volume, and so a relatively low level of unit cost in the future. In a small number of industries – aircraft, semiconductors and chemical fibres – this effect is quite large.

Second, if the attractiveness of a firm's product to new consumers increases with the number of consumers it has supplied in the past, then again a firm may use an initial low price to build up its early client base and so stimulate future demand (Katz and Shapiro, 1985). Examples of such effects abound in the information technology sector: as an item of hardware becomes more widely owned, more firms in the software industry will find it attractive to develop dedicated software for it, thus reinforcing its initial popularity.

What both these examples have in common with the endogenous sunk cost models described earlier becomes clear once we interpret the 'planned losses' incurred in the initial phase as a fixed outlay – analogous to an outlay on R&D or advertising – which yields a payoff in the later phase, either through lower unit costs or increased demand. The novel element which arises in these 'learning' or 'network

effects' models is that these effects can be cumulative over time, and so a small initial disparity in the costs or sales of two firms may in principle become amplified over time.

Structure, conduct and performance revisited

What does this imply for the Structure–Conduct–Performance paradigm? If high concentration is merely the natural outcome of the competitive process, should we still see high concentration in an industry as an indication that policy intervention might be warranted?

At a conceptual level, what remains is this: it is still true within the modern 'free entry' models discussed above that structure affects conduct. It is also true that conduct affects performance; but now there is a feedback loop through which high levels of profit may attract new entry, that is, structure is not a given, but is now determined as part of the market process. One consequence that emerges from this is that there is no simple and general link, of the kind central to the old literature, between high concentration and high profitability: it is possible, for example, to have industries with widely different levels of concentration that exhibit no difference in their rates of return on investment. High ('supernormal') profits can, however, arise in these 'free entry' models, through a number of channels. Most notably, they may arise because of asymmetries in the entry process ('first mover advantages'): an early entrant to the market may build up a level of investment in R&D, for example, and so enjoy both a high market share and a high rate of return on its investments, so that the industry-wide levels of concentration and profitability are relatively high. A second channel relates to the important but neglected role of 'integer effects', that is, if there is room in market at equilibrium for only a small number of entrants, then it may be, for example, that two firms can both make supernormal profits, but the entry of a third firm would drive the profit rate below a normal rate of return, so further entry does not occur. Finally, and most importantly, variations in productivity (unit costs) across firms associated with non-imitable advantages can lead to positive (supernormal) profits for (all) intra-marginal firms – a free entry condition implies 'zero profits' only for the marginal entrant (Demsetz, 1973).

One key issue remains: what of comparisons between alternative forms of market structure within any one industry? This is the question that lies at the heart of competition policy regarding mergers. As we have seen, the normal workings of the competitive process fix a lower bound to the level of concentration that must come about under free competition. This bound can, in the case of some industries, be very high in absolute terms, even in a large market; but in other industries it will be very low. Above this bound, varying levels of concentration can emerge since various patterns of market structure can be sustained as equilibria (as in the example of geographically dispersed plants mentioned above). What remains true of the Structure–Conduct–Performance story is that these different market structures may have different welfare properties; a proposed merger that moves us towards a more concentrated structure which will lead to reduced consumer welfare will be subject to the traditional objections. On the other hand, it may be that a merger arises merely as a response to changes in external conditions, and represents a shift away from a form of market structure that constituted an equilibrium outcome under the previous setting, but is no longer sustainable as an equilibrium in this changed environment. Distinguishing between these two possibilities in any specific instance is one of the (many) challenges in dealing with merger cases.

Why does it matter?

The traditional rationale for studying market structure was based on its link to profitability and to social welfare. There is, however, another line of argument that has gained considerable force as a result of the empirical success of the modern free entry models in explaining cross-industry differences in concentration. This line of argument rests on the claim that the success of these models provides convincing, though indirect, evidence for the workings of some key competitive mechanisms that appear to operate in a more or less uniform way across a wide range of industries.

To place this in perspective, it is worth noting that the conventional wisdom in economics from the 1950s to the late 1980s was deeply pessimistic in respect of models that lay between the two polar cases of perfect competition or (Chamberlinian) monopolistic competition, on the one hand, and monopoly on the other. This pessimism was typically expressed in the observation, ‘with oligopoly, anything can happen’. The new game-theoretic literature of the 1980s formalized oligopoly theory using the Nash equilibrium concept, and offered it as a general framework within which perfect competition and monopoly appeared as special (limiting) cases. While this new literature appeared to some critics to simply reinforce the negative view of oligopoly theory, the successful application of these game-theoretic models to the task of ‘explaining market structure’ suggests that the early pessimism is unwarranted: it seems that these models capture at least some ‘robust’ competitive mechanisms that operate in a more or less uniform way across the general run of industries.

See Also

- airline industry
- antitrust enforcement

Bibliography

Bain, J. 1956. *Barriers to New Competition*. Cambridge, MA: Harvard University Press.

Dasgupta, P. and Stiglitz, J.E. 1980. Industrial structure and the nature of innovative activity. *Economic Journal* 90, 266–93.

Demsetz, H. 1973. Industry structure, market rivalry, and public policy. *Journal of Law and Economics* 20, 113–24.

Katz, M.L. and Shapiro, C. 1985. Network externalities, competition and compatibility. *American Economic Review* 75, 424–40.

Schmalensee, R. 1989. Inter-industry studies of structure and performance. In *Handbook of Industrial Organization*, vol. 2, ed. R. Schmalensee and R. Willig. Amsterdam: North-Holland.

Shaked, A. and Sutton, J. 1986. Product differentiation and market structure. *Journal of Industrial*

Economics 36, 131–46.

Sutton, J. 1991. *Sunk Costs and Market Structure*. Cambridge, MA: MIT Press.

Sutton, J. 1998. *Technology and Market Structure*. Cambridge, MA: MIT Press.

Sutton, J. 2007. Market structure: theory and evidence. In *Handbook of Industrial Organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: North-Holland.

How to cite this article

Sutton, John. "market structure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000070> doi:10.1057/9780230226203.1036

The New Palgrave Dictionary of Economics Online

marketing boards

Christopher B. Barrett and Emelly Mutambatsere

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Marketing boards (state-controlled or state-sanctioned entities legally granted control over the purchase or sale of agricultural commodities) flourished in the 20th century. Since the mid-1980s they have declined in number under pressure from domestic liberalization and from international trade rules that increasingly cover agriculture. Where reforms have been widespread and successful, marketing boards have vanished or retreated to providing public goods, such as strategic grain reserves or insurance against extraordinary price fluctuations. Elsewhere, the weaknesses of private agricultural marketing channels have been revealed by the rollback of marketing boards, often leading to calls for reinstatement of powerful marketing boards.

Keywords

agricultural markets in developing countries; agriculture in economic development; contract enforcement; corruption; deregulation; Great Depression; international monetary fund; Lomé Convention; marketing boards; monopoly power; monopsony power; price control; price discrimination; price stabilization; privatization; protection; structural adjustment programmes; subsidies; World Bank; World Trade Organization

Article

Marketing boards are state-controlled or state-sanctioned entities legally granted control over the purchase or sale of agricultural commodities. They can be divided into two broad categories. Monopolistic marketing boards that create a single-commodity seller are found mainly in developed countries. Monopsonistic marketing boards concentrating buyer-side market power in one institution were commonplace for many years in developing countries. Monopolistic marketing boards were typically established with the main objective of maintaining or raising and stabilizing farm prices and incomes in an administratively practical and politically acceptable manner. By contrast, monopsonistic

marketing boards were typically established to give the state control over commodity prices – normally for the benefit of foreign and urban buyers – and capacity to tax agriculture so as to subsidize industrialization.

Marketing boards in developed countries

Marketing boards are state-sponsored trading enterprises legally invested with monopoly powers to organize the marketing of agricultural commodities. These statutory entities typically operate under direct or indirect producer control. Among the earliest boards were the New Zealand Meat Producers Board and the New Zealand Dairy Board, each established in 1922, the Australia Queensland Sugar Board of 1923, and the Australia Wheat Board, formed in 1939. In Australia, marketing boards used import protection and home consumption price schemes to stabilize producer prices. They initially received financial support from the state, although such support later declined as the focus of the boards changed. A number of state and commonwealth-level marketing boards were later established, with varying degrees of authority and responsibilities in the marketing of agricultural products such as wool, dairy, meat, wine and brandy, honey and horticultural products. The marketing boards in New Zealand evolved in a similar manner, with regulatory authority in export marketing and licensing but no direct financial support from the state. These boards, involved in the marketing of dairy, apple and pear, kiwi fruit, horticulture, meat and wool products, all used activities such as single-desk selling, price pooling, revenue pooling and preferential financing to seek higher producer prices.

The earliest major marketing schemes in Britain were the milk, potatoes and bacon marketing boards formed under the British Marketing Acts of 1931 and 1933. These acts enabled producers to set up marketing schemes that had the legislative power to ensure conformity by all producers. The core purpose of the marketing boards was to maintain or raise producer prices of basic agricultural commodities through acreage restrictions, direct or indirect limits on saleable quantities, and price discrimination, with higher prices in sheltered markets and lower prices in exposed ones. In addition, monopolies of processed products were legalized, leading to the organization of processor and distributor schemes. The marketing boards thus held the monopoly power to control supply, the terms of sale and the channels and conditions of sale (Bauer, 1948). By 1948 marketing boards had spread to include all major agricultural commodities. In Canada, marketing boards were also formed in response to the price fluctuations of the Great Depression. The Dominion Marketing Board, a federal agency established under the National Farm Products Act of 1934, exercised extensive market power over the sale of regulated products, transferable to provincial-level producer-organized boards. The Agricultural Marketing Acts of 1940 and 1956 delineated the powers of regulation and market control activities for the established and new federal and provincial marketing boards. The result was marketing boards with diverse market powers and scope of operations across provinces, and across boards within the same province. Some marketing boards act only in a supervisory capacity, whereas others wield more extensive powers in market regulation and control. Activities generally range from negotiating minimum prices, regulating quantity and quality of marketed products, collecting and distributing payments, as well as grading and quality control.

Several common features distinguish marketing boards in developed countries from those found in developing nations. First, marketing boards in developed countries tend to be specialized in both scale and scope of operations. For example, New Zealand currently runs strictly export monopolies, such as

the Dairy Board, that have control over the country's agricultural exports but negligible influence over domestic production, sales, imports or tariff rates.

Second, marketing boards in developed countries tend to subsidize farmers at the expense of consumers, as evidenced by their mandate to maintain high producer prices for farmers through limited supply. One result is that marketing boards in developed countries have tended to generate windfall profits for the owners of farm land and other sector-specific assets in agriculture.

Third, and following directly from their role in subsidizing farmers, state trading enterprises tend to encourage and support cartels at producer, processor and distributor levels. Developed country agricultural marketing boards have been a major issue in international trade because historically they dominated certain markets. For example, McCalla and Schmitz (1982) estimated that 95 per cent of world wheat trade in 1973–7 involved a state marketing board on at least one side of the transaction. Because marketing boards enjoy greater flexibility than private traders in pricing – for example, they can commonly delay payments to producers, pool payments so as to reduce producer price risk, and can practise discriminatory pricing among export or import markets – their operations are closely scrutinized by the World Trade Organization for prospectively anti-competitive practices.

Marketing boards in developing countries

Marketing boards in developing countries were typically begun during colonial times for purposes distinct from those of their counterpart marketing boards in developed economies. And they have followed a somewhat different trajectory from those of marketing boards in developed countries. European colonial powers formed marketing boards in large measure to facilitate the export of agricultural commodities to Europe and to stabilize prices faced by colonial elites (for food crops) and metropolitan buyers (for export crops). Post-independence governments generally maintained marketing boards because these were considered simpler to manage and more efficient in conducting organized trade than the traditional, decentralized private sector. More compellingly, marketing boards provided a convenient way for the governments to maintain control over the marketing of strategic commodities, such as the food staples and important export crops (Lele and Christiansen, 1989). The marketing boards system was most prevalent in the anglophone African and South Asian countries, but widespread as well in francophone and lusophone African countries and in Asia and Latin America.

Marketing boards were both state-owned and state-funded, based on centralized decision making systems. They possessed the sole legal authority to purchase commodities from farmers and to engage in trade. Through the boards, governments typically fixed official producer prices for all controlled commodities, often in a pan-seasonal and pan-territorial manner whereby a single price was set for the whole marketing season and for all regions of the country. Marketing boards provided a guaranteed market for the farmers, absorbing all marketed surplus at the official producer prices, and maintaining extensive buying networks and storage facilities throughout the production regions. Pan-seasonal and pan-territorial pricing practices eliminated any opportunities for arbitrage, discouraging private investment in commodity storage or transport capacity, and reinforcing the government's control over the marketing channel. Unlike marketing boards in developed countries, producer sales into the network were rarely rationed, because the marketing boards’ objective was normally to increase supply and lower prices for consumers, as opposed to controlling supply for the benefit of producers.

Two features of the export crop marketing boards – as distinct from those handling staple food

commodities – are worth noting. First, the marketing boards held the sole legal rights in commodity export, and had a mandate to generate income for the state. Therefore, storage costs were maintained at low levels through selling policies such as rapid evacuation and forward selling. In addition, local producer prices were typically set at levels lower than the international free-on-board prices, through price fixing or overvalued exchange rates. Essentially, export crop marketing boards were used as a means to tax agriculture in order to develop the industrial sector in these agrarian economies. The taxes were often quite severe. In Tanzania, for example, local producer prices for coffee and tobacco fell to 23 per cent and 15 per cent of international prices, respectively, by the mid-1980s.

Second, because export crop marketing boards served foreign demand, no price controls existed on the selling end. Marketing boards could trade on an open market for the highest possible selling prices. However, because most of the former European colonies enjoyed preferential access to European markets under the Lomé Convention, most commodities were sold to Europe. In addition, some export crops enjoyed commodity price stabilization through international commodity agreements such as the International Coffee Agreement or the International Rubber Agreement. In those cases where a country enjoys world market power, a state marketing board can, at least in theory, increase prices and thereby extract consumer surplus from foreign buyers to benefit the exporting country, including its producers. This is one of the concerns surrounding state trading enterprises within global trade policy fora.

Even though the export crop marketing boards were generally established first, in most developing countries staple food commodity marketing boards became at least as significant a part of the parastatal system. For food commodities, government control extended to every stage of the market chain, to include farm gate, wholesale and retail price controls. In-country commodity movement was restricted, especially the movement of strategic food commodities, and private trade was either illegal or legal only by licence. To achieve food security objectives, food subsidies were generally offered, mostly implicitly, in the form of fixed consumer prices set at levels lower than the market price. Although farm prices were generally set at a below-market level as well, the government often offered implicit subsidies to farmers, through price stabilization operations, and input and credit subsidies administered through the marketing boards (Lele and Christiansen, 1989). Moreover, pan-territorial pricing typically implied subsidies for farmers in more remote smallholder regions. In some countries and for some crops, these arrangements likely stimulated greater crop production than would have occurred under open market arrangements. Grain marketing boards commonly also handled the strategic food reserves for emergency situations, and had the responsibility to import food in shortage seasons. These parastatals therefore held most of their nations' inter-seasonal and inter-annual grain storage capacity, a legacy that would affect inter-seasonal commodity price movements after the liberalization of commodity marketing systems in the 1980s and 1990s. Although processing was not their core business, marketing boards, in some cases, were also involved in preliminary processing, such as milling rice or maize, or in licensing and monitoring the processing industry activities. This underscores an important difference from developed country marketing boards: the breadth of commodity marketing boards' mandate in most developing countries.

Over time, the fiscal sustainability of marketing boards in developing countries became questionable. The broad range of marketing operations handled by marketing boards and the politically charged manner in which these operations were typically handled led to massive inefficiencies and deficits that cash-strapped central governments had an increasingly difficult time covering. The subsidies embedded in grains pricing systems, coupled with heavy overhead costs associated with high administrative,

transportation and storage costs, soon created huge tax burdens. In an attempt to ensure food security, the state would generally increase producer prices with less than proportional increases in consumer prices, taking on responsibility for a significant share of the marketing costs associated with moving food from farm to table. The pan-territorial pricing system meant higher transportation and handling costs in moving commodities from some remote areas, and the management of large volumes of commodities in storage was costly. In addition, the monitoring of private trade was not only costly but generally ineffective, especially for food commodities in shortage seasons, when parallel markets flourished to meet local demand. In Mali, for example, even though private cereals trade was illegal before 1981, only 30–40 per cent of total grain trade was actually handled by the state trading agency, OPAM (Dembélé and Staatz, 2002). On the international market, marketing boards faced decreasing real commodity prices for export crops, further undermining their sustainability.

By the end of the 1970s budget deficits resulting from the management and mismanagement of parastatals had reached astronomical levels in most countries. In Mali, OPAM's annual deficit reached US$80 million by 1980, three times the board's annual grain sales. In Tanzania, the National Marketing Corporation's overdrafts were about $250 million in 1993, against total state expenditures on agriculture of $12million. The National Cereals and Produce Board (NCPB) of Kenya accumulated an estimated loss of about $300 million by 1993, in contrast with central government expenditure on agriculture of $33 million (Staaz et al., 2002; Lele and Christiansen, 1989). These patterns were by no means exclusive to Africa. Indonesia's price stabilization scheme for rice, managed by the National Logistics Supply Organization (BULOG), also proved a high price to pay for self-sufficiency, as did the Food Corporation of India.

In addition to budgetary complications, marketing boards also faced organizational challenges. Their susceptibility to bureaucracy and corruption increased both the inefficiency in their operations and the transactions costs for farmers and consumers. For example, Arhin, Hesp and van der Laan (1985) argue that by the mid-1970s the Ghana Cocoa Marketing Board had become little more than an instrument of the government for the purpose of mobilizing political support for the incumbent government.

Mounting deficits, poor management and the perverse incentives created by anti-competitive behaviour brought marketing boards and price stabilization systems under attack, based in part on seminal research into the welfare effects of government interventions to stabilize commodity prices (Newbery and Stiglitz, 1981). These deficit problems, coupled with the new economic insights, triggered widespread agricultural market reforms in the 1980s and 1990s throughout the developing world, implemented mainly but not exclusively, in the context of structural adjustment programmes (SAPs) of the World Bank and the International Monetary Fund.

Agricultural marketing reforms generally aimed to reduce the role of the public sector in marketing and to encourage private sector participation so as to let markets allocate scarce goods more efficiently. Marketing boards experienced major reforms under these programmes, comprising the elimination of price controls, termination of farm input and consumer food subsidies, removal of marketing boards' monopsony power and deregulation of private trade. In many cases, marketing boards were privatized or at least commercialized, the latter referring to cases where marketing boards remained government owned, but with autonomous decision-making power and an explicit objective to maximize profits. The logic was that, by removing political interference in the marketing process, market forces would lead to efficient resource allocation and price discovery. Market deregulation was thus expected to improve marketing efficiency by reducing transactions costs, increasing producer prices, thus inducing increased

production and potentially also lowering consumer prices.

The response of the market was immediate and quite dramatic in many cases. Entry into formerly controlled agricultural markets was massive in most countries, although with continued bottlenecks in functions requiring significant capital outlays, such as bulk inter-seasonal storage and long-haul motorized transport, entry was typically restricted to niches with low entry barriers (Barrett, 1997). Nonetheless, formal and informal private traders became a significant part of the marketing channel, performing most of the trade activities that the marketing boards previously performed.

In spite of widespread liberalization, marketing operations for most ‘strategic’ food and export crops changed little. Newly privatized or commercialized marketing boards were often replaced with ‘new’ marketing boards that were initially intended to provide public goods, but eventually and predictably became involved in crop marketing. In Zambia, for example, the government-owned Food Reserve Agency (FRA) that replaced the National Agricultural Marketing Board (NAMBOARD) in 1995, charged with maintaining the strategic grain reserve and acting as a buyer of last resort for smallholder farmers, in time took up prior NAMBOARD responsibilities such as fertilizer distribution. Moreover, some of the commercialized marketing boards did not significantly change their pricing systems and continued to use the power of the state to remain dominant players in the current market system. In Indonesia for example, even though the market was opened to private traders, BULOG remained a price leader by operating a major buffer stock, purchasing rice when rice prices fell below a stated floor price and releasing stocks when prices rose above a price ceiling. Similarly, in the Kenyan maize sector the NCPB continued to intervene directly in markets to support maize prices; and in Malawi ADMARC remains the dominant maize buyer and distributor of inputs. Zimbabwe went so far as to reinstate the monopsony power of the Grain Marketing Board and its pre-reform operations. Not surprisingly, the budget deficit of these marketing boards actually increased after reforms.

These trends reflect governments’ reluctance to relinquish control over marketing board operations, particularly the setting of prices for key food and export crops, given political sensitivity to these issues. As it turned out, such concerns were not completely unwarranted. In many developing countries the legacy of private underinvestment in storage and transport capacity, inadequate commercial trading skills in the nascent private sector, combined with limited access to finance, restricted entry into key niches of the marketing channel. These market conditions facilitated the emergence of new monopolies, often substituting private for public market power. Problems of weak contract enforcement, unreliable physical security and underdeveloped communications and transport infrastructure often impeded business expansion, market integration and price transmission. Despite increased private investment in transportation and storage infrastructure after reforms, the weaknesses of the existing systems implied considerable business risk. Consequently, private traders did not fully or quickly fill the voids left by the withdrawal of the marketing boards from core commodity market intermediation functions. Price volatility increased sharply in many countries. Moreover, farmers’ access to seasonal credit dropped significantly as market liberalization ended formerly monopsonistic marketing boards’ willingness to extend seasonal credit to growers that were collateralized by future sales. Reduced credit often led to fewer purchased inputs and lower crop output. In an attempt to restore market stability and production volumes, states often suspended or reversed reforms, reinstating price controls and trade restrictions, thereby further exacerbating instability and undermining investor confidence. The result has been incomplete reforms in most developing countries, where private sector involvement remains pervasive but small-scale and weak, while unprofitable commercialized marketing boards remain prominent and

prone to government interference.

The current state of play

Far fewer marketing boards exist than previously. Because they reduce or eliminate competition, marketing boards are widely believed to induce inefficiency in marketing and sluggishness in price discovery. Therefore, government involvement in agricultural marketing has been weakening in both developed and developing countries since the mid-1980s, a result of the adoption of more liberal domestic economic policies, coupled with global pressure to conform to international trade rules steadily expanding their coverage of agriculture. The monopoly or monopsony powers of all but a few marketing boards have been lifted, and the marketing and processing activities of the boards have been streamlined. Where reforms have been widespread and successful, marketing boards have vanished or retreated to providing public goods, such as strategic grain reserves or insurance against extraordinary price fluctuations. Where reforms have been halting or unsuccessful, the weaknesses of private agricultural marketing channels have been laid bare by the rollback of marketing boards.

See Also

- agricultural markets in developing countries
- agriculture and economic development
- international trade theory

Bibliography

Arhin, K., Hesp, P. and van der Laan, L. 1985. *Marketing Boards in Tropical Africa*. Nairobi: KIP Limited.

Barrett, C. 1997. Food marketing liberalization and trader entry: evidence from Madagascar. *World Development* 25, 763–77.

Bauer, P. 1948. A review of the agricultural marketing schemes. *Economica* 15, 132–50.

Dembélé, N. and Staatz, J. 2002. The impact of market reform on agricultural transformation in Mali. In *Perspectives on Agricultural Transformation: A View from Africa*, ed. T. Jayne, I. Minde and G. Argwings-Kodhek. Hauppauge, NY: Nova Science Publishers Inc.

Lele, U. and Christiansen, R. 1989. *Markets, Marketing Boards and Cooperatives in Africa, Issues in Adjustment Policy*. MADIA Discussion Paper 11. Washington, DC: World Bank.

McCalla, A. and Schmitz, A. 1982. State trading in grain. In *State Trading in International Markets*, ed. M. Kostecki. New York: St Martin's Press.

Newbery, D. and Stiglitz, J. 1981. *The Theory of Commodity Price Stabilization*. Oxford: Clarendon Press.

How to cite this article

Barrett, Christopher B. and Emelly Mutambatsere. "marketing boards." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000078> doi:10.1057/9780230226203.1037

The New Palgrave Dictionary of Economics Online

marketplaces

Winifred B. Rothenberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Marketplace’ was defined by the great 18th century jurist, William Blackstone, as ‘a spot of ground set apart by custom for the sale of particular goods.’ The definition is striking in its simplicity, but the simplicity is deceptive, for each word should give us pause. If a marketplace is a ‘spot’ it is both bounded and of little consequence. Having said it was a ‘spot’, ‘of ground’ would seem to be redundant were it not that this, the market as a *place* located in space is the feature that over time will change the most. ‘Set apart’ underscores the irreconcilability of Commerce and Community, and, by implication, bestows upon Community the greater authenticity. ‘Sale’– legally to divest a seller – presupposes a body of legal rules protecting title and the alienability of title, at least to ‘particular goods’. And ‘particular goods’, in turn, implies the inalienability of title to *other* goods. Finally, there is ‘custom’. In Blackstone's syntax, custom's role appears limited to the setting of the spot. But when Commerce threatens to overwhelm the barriers that Community has erected against it, it will be custom that writes the regulations into law. ‘Marketplace’ is not at all ‘simple,’ and reckoning with it in one or another of the protean forms it has assumed over the millennia deserves to engage, as indeed it has, the attention of archaeologists, historians, anthropologists, sociologists, and economists.

Keywords

Black Death; Blackstone, W.; Braudel, F.; Domar, E.; Durkheim, E.; law of one price; market institutions; marketplaces; Marshall, A.; non-market economies; peasant economy; peasants; regulation of markets; repeated games; Silk Road; social networks

Article

Marketplaces in the long-run

The history of marketplaces is traced here as a sequence of pivotal events that began at least 60,000

years ago when, archaeologists suggest, the earliest appearance of trade in Europe can be inferred from the discovery of made objects -- amber, mollusc shells and worked stone -- that had been moved to sites hundreds of kilometers from their sources. But the origins of trade may extend back even further in time, for ‘the archaeological record does not reveal *any* time in the history of man in Europe in which there was no movement of “manufactured” objects’ (Grantham, 1997, p. 18).

A giant leap takes us to Mesopotamia 10,000 years ago. Having by now collected abundant evidence of the deleterious effect of settled agriculture on human health in the Neolithic, it has been suggested by some paleopathologists that the Agricultural Revolution – the sedentary cultivation of grain, sugar, flax, wool, sisal, jute and hemp – may have been driven less by the quest for food security than by the high value these staple crops could earn in trade.

Study of the 4,000 year-old Heqanakht Papyri reveals Heqanakht himself to have been ‘obsessed’ with running his farm to make a profit and augment his wealth, and the economy of Pharaonic Egypt to have been one *not* of priestly redistribution, as had been thought, but of private property, money prices, cash crops, rental land, wage labour and marketplaces (Allen, 2002).

Ten centuries before the Christian era, the marketplaces rimming the Aegean Basin were specializing in the export of high-value wines, olives and oil to trade for imports of grains, raw materials and slaves from the ‘barbarians’ north of the Mediterranean.

In the eighth and seventh centuries BC, Phoenicia's legendary traders integrated the markets of the eastern and western Mediterranean; the Etruscans made contact with the Celts; and the merchants of India reached the northwest coast of England.

In the Bible, The Book of Ezekiel, written, it is thought, in the sixth century BC, describes in chapters 26–27 the prosperous city of Tyre whose marketplaces overflowed with an abundance of fir trees from Senir, cedars from Lebanon, ivory benches from the isle of Chittim, fine linen from Egypt, silver, iron, tin and lead from Tarshish, emeralds, purple, coral, agate and fine linen from Syria, wheat from Minnith, honey, oil and balm from Pannag, wine and white wool from Helbon, lambs, rams and goats from Arabia, spices from Sheba and Ra-amah, and skilled artisans, laborers seafaring men, and merchants from all over the eastern Mediterranean.

Most exotic of all were the marketplaces that grew up along the Silk Road, established by the Han Dynasty in the second century BC as China's only link to the West. Along much of its length, the Silk Road was less a ‘road’ than a hazardous path around the world's most unlivable desert, through the world's most inaccessible mountain passes, over the world's highest peaks, and among the world's most isolated and hostile peoples. The Road started in what is now Xian on China's northwest border, and made its way west through Kazakstan, Kyrgyzstan, Uzbekistan, Tajikistan, and Afghanistan, encircling the Hindu Kush, until crossing the Black Sea, it ended in Roman Syria – which is as much as to say, in Venice! The market-places of this commerce were the oases of central Asia and the fabled ‘Arabian Nights’ cities they became: Xian, Islamabad, Tehran, Kashi, Samarkand, Bukhara, Kabul, Kandahar, Tashkent, Aleppo, Lahore, Baghdad, Ankara, Istanbul. For 15 centuries, caravans of up to 1,000 camels each made the journey from oasis to oasis for months on end, under armed guard, bearing gold, metals, ivory, precious stones, myrrh, frankincense, ostrich eggs, horses, glass, silks, porcelain, guns, powder, jade, bronzes, lacquer – and the great religious civilizations of Buddhism and Islam.

On the European end of the Silk Road great ‘diaspora networks’ were founded by Greek, Armenian and Jewish merchant families of the eastern Mediterranean, who traded a world away with the diaspora networks of India, China, Japan and Indonesia (McCabe, Harlaftis and Minoglou, 2005).

In time, the Romans would discover the secret of making silk, Europeans would find a sea route to China, the Silk Road with its storied past would disappear beneath the sand, the Mongol hordes would burst out of the East, and Europe would reap the whirlwind. In the Black Death that ravaged the continent between 1348 and 1351, Europe lost as much as 40 per cent of its population.

To a demographic event so catastrophic no system, no institution, no mode of production could remain impervious. Especially transformed were Europe's labour markets. In England, those who survived the Black Death lived to enjoy ‘the sole golden age of the English peasantry’ (Hatcher 1987, p. 281). The Statute of Artificers notwithstanding, the general response of the manorial economy to the desperate scarcity of labour, the abundance of abandoned land, and rising prices in grain markets was to relax feudal constraints on the mobility of labour, to raise nominal wages, lower rents, and make concessions on customary dues and obligations. In a word, peasants under villein tenure were able to secure copyhold and leasehold tenancies. By the 17th century, England had become ‘a peasant-free zone’ (R.M. Smith in Scott, 1998, p. 346).

In Russia, in stunning contrast, the response to depopulation was the establishment of serfdom. In the face of acute labour scarcity and expanding grain markets, the Boyars demanded tighter and tighter legal restrictions on the mobility of peasants until, by an Act of the Duma in 1649, serfs and their households were made the personal property of the lord -- effectively slaves (Domar, 1970, pp. 13–32).

Thus, ‘peasant’ in the European context came to be a status that defies definition: some were well on the way to yeomanry by 1600, others were only a technicality away from a heritable condition of slavery.

Marketplaces and peasants

From the 14th century to the 18th, the history of marketplaces is linked to the history of peasants, as that of peasants is to marketplaces. In Brueghel's exuberant paintings of peasants tumbling about in overflowing marketplaces the link has become an icon of Flemish art. ‘Historically, peasants only exist when markets exist, even if they do not fully participate in them’ (Scott, 1998 p. 2).

Having for well over a century minutely observed a vast number of peasant societies, the characteristics of peasant markets, and the behaviour of peasants with respect to them, it has been anthropology, rather than economics, that has become, to use Grantham's phrase (1997, p. 19), ‘the referent social science’ of the peasant economy. Peasant villages, whether in Indo-China, West Africa, the Stone Age, or – improbably – colonial New England, are said to have this in common: that the village marketplace is ‘embedded in’ the social fabric of the community; the rules governing it, and the motivations of buyers and sellers and borrowers and lenders transacting in it are subordinate to and constrained by the non-market values of the ambient culture, and stand as epicentres of resistance to the encroachment of the ‘disembedded’, proto-capitalist, dynamic, hegemonic Market, with a capital M.

Influential as this model has been, there is a counter-narrative, accepted even by some anthropologists. Thus, for example, among the Panajachelenos of the Guatemalan highlands, ‘Commerce is the breath of life’ (Tax, 1963, p. 132). According to Sol Tax, both men and women spend more time buying and selling and talking about buying and selling than doing anything else. When they have nothing to sell, they buy something in a cheap market and carry it to a dear one to sell. The Maoris of New Zealand, the Ifugao of the Philippines, the Senegambian, Afikpo, Esusu, Yoruba, Hausa, Tiv and Dahomean peoples of West Africa, the northeastern Malays, the Javanese and the Trobriand Islanders have all been found to

engage in price- and quantity-bargaining, to 'seek maximal advantage' in marketplace transactions (Firth and Yamey, 1964), and to exhibit in their market-dependence the same U-shaped relationship to their disposable assets as is associated with risk-averse behaviour in developed economies.

A case in point is a new study of African markets by Marcel Fafchamps (2004) who argues that sub-Saharan Africa today is decisively 'market-oriented'. The development of Africa, Fafchamps insists, has been impeded not by the cultural incongruity of markets, or the absence of markets, or the lack of a market *mentalité*, but by the accommodation that traders in sub-Saharan markets have had to make to the ubiquity of market imperfections. In the absence of well-defined property rights, intermediary institutions, notaries public and contract enforcement, the sole bar against asymmetric information, adverse selection and moral hazard has been what might be called 'insider trading': repeated transactions among networks of friends and relations bound to one another by webs of trust. In the context of sub-Saharan Africa these webs, although intimate, are not cooperative, says Fafchamps, but strategic; the object is not to avoid the discipline of the market but more nearly to satisfy its assumptions.

Few peasant economies fit the 'moral economy' model (Rothenberg, 1992, chap. 2). And those that do may, like 15th-century French villages, have been torn by strife, the collective experience 'rubbed raw' (Hoffman, 1996, p. 77) by the face-to-face pursuit of what Avner Offer (1997) has called 'regard'.

But the principal critique of a non-market model is that it is a steady-state model, and that has tended to impart a Durkheimian steady-state bias to peasant studies. It is a bias very much 'at home' with a methodology in anthropology itself in which scrutiny ever closer is leveled at subjects ever narrower -- from tribe, to village, clan, household, extended family, nuclear family, gender – thereby surrendering to economists the trade *among* groups and the articulation *between* marketplaces that generates growth. It is in regulating that process that the community asserts its dominion over what Braudel (1982, p. 58) has called 'the insidious tentacles of the economy'.

Regulating the marketplace

Such interventions have a very long history, coming down to us, Blackstone thought, from Saxon times when no title to goods valued above 20 pence could change hands without witnesses. But such impediments to trade could not have persisted were it not for law. It is the province of any law worthy of the name to recognize in 'You have got what belongs to me' its sphere of action (Pollock and Maitland, 1895 p. 33), and nowhere more urgently than at the point of transferring the ownership of chattels by sale. For what but law can distinguish sale from theft, right from use, ownership from possession, *dominium* from *seisin*?

As the frequency of trade increased in the 13th century, exchange in an open market established by royal grant – a 'market ouvert' – sufficed, in lieu of witnesses, to divest a seller. With outdoor marketplaces came rules, regulations and ordinances establishing, locating, restricting, and supervising open markets, covered markets, fairs, merchant courts and piepoudre courts for itinerant merchants; appointing ringers of opening and closing bells, monitors of weights and measures, collectors of fees and fines, inspectors of quality, enforcers of Just Prices, and wardens to patrol the perimeters of the marketplace. This regulatory 'apparatus' was brought to the American colonies and given the force of law in the municipal marketplaces of all large towns in Massachusetts and throughout the colonies.

'The Laws and Liberties of Massachusetts', codified between 1641 and 1691, declared the taking of

excessive wages by mechanics and day labourers and the charging of unreasonable prices by shopkeepers and merchants punishable by double restitution or imprisonment. They set the weight of the pennyloaf of white bread, and authorized the selection of two able persons annually to enter into the houses of all bakers ‘as oft as they see cause’ to inspect and weigh all bread found there on pain of forfeiture. Two persons were appointed annually to ascertain the range of prevailing wheat prices each month and set the price at which bakers shall bake their bread. ‘The Laws and Liberties’ also set the days of the week when a marketplace shall be kept in Boston, Salem, Lynn and Charlestown, set the times of the ringing of the opening and closing bells, forbade all trade outside the perimeter of the marketplace, and set the two days a year when Boston, Salem, Watertown and Dorchester shall have fairs (Cushing, 1976).

In 1737, Faneuil Hall, Boston's beleaguered marketplace, was besieged by farmers from the surrounding hinterland who ‘donned the livery of heaven’ (disguised themselves as clergy) and burned it to the ground (Brown, 1900, chap. 8). By the early 1820s, the regulated ‘market ouvert’ and the legal doctrine of implied contract expressed by it were abandoned in favour of the rule of *caveat emptor*. Contract, not Community, would come increasingly to regulate markets.

Marshallian markets: the marketplace after 1900

With the publication of Alfred Marshall's *Principles of Economics* in 1900, economics became the referent social science of markets. It may therefore come as something of a surprise to discover that in this, the urtext of economics, marketplaces have vanished! Marshall's definition of a market is of an abstraction, outside of spatial coordinates, oblivious of cultural context, and functioning homeostatically in accordance with laws of its own making. ‘The distinction of locality is not necessary,’ he wrote. ‘Economists understand by the term “Market” not any particular market-place in which things are bought and sold, but the whole of any region in which buyers and sellers are in such free intercourse with one another that the prices of the same good tend to equality easily and quickly’ (Marshall, 1890, p. 324).

Thus, the market is not a place but a *process* that expands in space and unfolds in time, driven by the pace at which different prices for the same good converge toward a single price. Called the Law of One Price, that convergence is the unintended consequence of arbitrage between buyers seeking cheap markets and sellers seeking dear markets.

But as long as a wedge between ‘cheap’ and ‘dear’ markets persists, that is, as long as the convergence process is incomplete, marketplaces remain significant, not as ‘spots of ground’ but as transitional nodes of price-formation that become ‘folded in’ as the market process advances along its dendritic expansion-path. The story we have followed for 60,000 years has not, even in a Marshallian sense, become irrelevant.

At the same time, economics itself, as the referent social science of markets, is expanding its narrow field of vision beyond its Marshallian boundaries in an attempt to comprehend the wider social and psychological foundations of economic behaviour. I am reminded of the earthquake of 8 October 2005 which struck high up in the Kashmiri Himalayas. Within a week of the catastrophe, survivors set up a village marketplace (BBC News). Seventy-three thousand had been killed, three million were made homeless, none had necessities, none had surpluses, but within days they made a marketplace.

See Also

- Braudel, Fernand
- Domar, Evsey David
- market institutions
- peasant economy
- peasants
- Polanyi, Karl

Bibliography

Allen, J.P. 2002. *The Heqanakht Papyri*. New York: Metropolitan Museum of Art and Yale University Press.

Braudel, F. 1982. *The Wheels of Commerce: Civilization & Capitalism, 15th–18th Century*, vol. 2. New York: Harper and Row.

Brown, A.E. 1900. *Faneuil Hall and the Faneuil Hall Market, Or Peter Faneuil's Gift*. Boston: Lee and Shepard.

Curtin, P.D. 1975. *Economic Change in Precolonial Africa: Senegambia in the Era of the Slave Trade*. Madison: University of Wisconsin Press.

Cushing, D. 1976. *The Laws and Liberties of Massachusetts, 1641–1691*, 3 vols. Wilmington: Scholarly Resources.

Domar, E.D. 1970. The causes of slavery and serfdom: a hypothesis. *Journal of Economic History* 30, 18–32.

Ellis, F. 1988. *Peasant Economics: Farm Households and Agrarian Development*. Cambridge: Cambridge University Press.

Fafchamps, M. 2004. *Market Institutions in Sub-Saharan Africa*. Cambridge, MA: MIT Press.

Firth, R. and Yamey, B.S., eds. 1964. *Capital, Saving and Credit in Peasant Societies: Studies from Asia, Oceania, the Caribbean, and Middle America*. Chicago: Aldine.

Gonzales, R. 2005. The Geography of the Silk Road. Road. http://www.humboldt.edu/~geog3091/ideas/raysilk.html. Accessed 21 July 2005.

Grantham, G. 1997. The shards of trade: archaeology and the economic history of the super-long-run. Paper delivered at the Economic History Association Conference, August 1997, p. 18.

Hatcher, J. 1987. English serfdom and villeinage: towards a reassessment. In *Landlords, Peasants and Politics in Medieval England*, ed. T. Aston. Cambridge: Cambridge University Press.

Hoffman, P.T. 1996. *Growth in a Traditional Society: The French Countryside, 1450–1815*. Princeton: Princeton University Press.

Hughes, J.R.T. 1976. *Social Control in the Colonial Economy*. Charlottesville: University Press of Virginia.

Kohn, M. 2003. Organized Markets in Pre-Industrial Europe. Working paper. Dartmouth College, July 2003.

Marshall, A. 1890. *Principles of Economics*. 8th edn. London: Macmillan.

Mazower, M. 2004. *Salonica, City of Ghosts: Christians, Muslims, and Jews, 1430–1950*. New York: Alfred Knopf.

McCabe, I.B., Harlaftis, G. and Minoglou, I., eds. 2005. *Diaspora Entrepreneurial Networks: Four Centuries of History*. New York: Berg.

McNeill, W.H. 1976. *Plagues and Peoples*. Garden City: Anchor/Doubleday.

Offer, A. 1997. Between the gift and the market: the economy of regard. *Economic History Review* 50, 450–76.

Polanyi, K. 1944. *The Great Transformation: The Political and Economic Origins of Our Time*. Boston: Beacon Press.

Pollock, F. and Maitland, F. 1895. *The History of English Law before the Time of Edward*, vol. 2. Cambridge: Cambridge University Press, 1968.

Rothenberg, W.B. 1992. *From Market-Places to a Market Economy: The Transformation of Rural Massachusetts, 1750–1850*. Chicago: University of Chicago Press.

Scott, T., ed. 1998. *The Peasantries of Europe: From the Fourteenth to the Eighteenth Centuries*. New York: Addison Wesley Longman.

Tax, S. 1963. *Penny Capitalism: A Guatemalan Indian Economy*. Chicago: University of Chicago Press.

How to cite this article

Rothenberg, Winifred B. "marketplaces." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000064> doi:10.1057/9780230226203.1033

The New Palgrave Dictionary of Economics Online

markets

Geoffrey M. Hodgson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Until recently, and despite strong interest in market outcomes, economists have paid relatively little attention to the institutional structure of markets. This article considers the historical evolution of markets and poses some dilemmas concerning their definition. Several alternative definitions are considered, involving different degrees of historical specificity. It is argued that developments in economics and elsewhere since the 1980s point to a more nuanced view of markets, involving a recognition of different types of market mechanisms and institutions. These developments include work in experimental economics and auction theory. A definition of markets is offered that is consistent with these developments.

Keywords

agent-based models; Amsterdam Bourse; Auctions; auctions; Austrian economics; bonded labour; central planning; Clark, J. M.; Coase, R.; commodity exchange; Commons, J.; contemporary capitalism; contracts for service; cultural norms; Dutch East India Company; economic sociology; experimental economics; game theory; general equilibrium; German Historical School; gift-exchange; Globalization; Hayek, F.; Hobson, J.; imperfect information; information; labour market contracts; law of one price; learning; London Stock Exchange; market institutions; marketplaces; markets; marriage markets; Marxism; Mises, L. von; missing markets; new institutional economics; New York Stock Exchange; North, D.; Parsons, T.; Polanyi, K.; property rights; Rau, K. H.; relational exchange; Robbins, L.; slavery; Smith, V.; Stigler, G.; trust; Weber, M.

Article

Markets dominate modern life, and economists have for long been concerned about market prices, but, despite this ongoing preoccupation, until recently there has been little discussion of the nature and operation of markets themselves.

No fewer than three Nobel Laureates in economics have noted this paradox. George Stigler (1967, p. 291) wrote: ‘The efficacy of markets should be of great interest to the economist: Economic theory is concerned with markets much more than with factories or kitchens. It is, therefore, a source of embarrassment that so little attention has been paid to the theory of markets and that little chiefly to speculation.’ Stigler made a plea for the theoretical study of markets, which for a long time went unheard.

Ten years later Douglass North (1977, p. 710) similarly remarked: ‘It is a peculiar fact that the literature on economics and economic history contains so little discussion of the central institution that underlies neoclassical economics – the market’. Another 11 years had passed when Ronald Coase (1988, p. 7) observed that ‘in modern economic theory the market itself has an even more shadowy role than the firm’. Economists are interested only in ‘the determination of market prices’ whereas ‘discussion of the market place itself has entirely disappeared’.

Economists have had little to say about the nature of markets, other than classifying them by their degrees of competition and their numbers of buyers and sellers. Beyond this, the institutional aspects of markets have been widely neglected. For much of the 20th century there has been little discussion of how specific markets are structured to select and authenticate information, and of how specific prices are actually formed. Furthermore, ‘the market’ was treated as a relatively homogeneous and undifferentiated entity, with little consideration of different market mechanisms and structures. When market mechanisms were addressed this was typically confined within the framework of general equilibrium theory, with relatively little attention to the institutional details and alternative market structures.

Inspection of standard economics textbooks confirms these observations. While market outcomes such as prices are always central to the discussion, there is generally little consideration of the detailed rules and mechanisms through which prices are formed, and the concept of the market itself often goes undefined. Indeed, there is an entry on markets in neither the massive 1968 edition of the *Encyclopaedia of the Social Sciences* nor the otherwise comprehensive 1987 edition of *The New Palgrave: A Dictionary of Economics*.

Three questions arise. First, what briefly is the nature of markets and how can the market be defined? Second, why has the specific anatomy of markets been neglected by economists? Third, what recent developments in economics and elsewhere help to remedy the deficiency? After a brief historical discussion this article addresses these questions.

Historical background

Goods have changed hands within human societies for hundreds of thousands of years. However, much of this internal circulation was powered by custom and tradition. Transfers of goods often involved ceremony and personal, reciprocal actions. These personal and kin-based exchanges contrast with the organized and competitive pecuniary ambiance of modern markets. Ceremonial transfers involved ‘the continuous definition, maintenance and fulfilment of mutual roles within an elaborate machinery of status and privilege’ (Clarke, 1987, p. 4). Most of this internal circulation of goods was devoid of any conception of the voluntary, contractual transfer of ownership or property rights. These reciprocal transfers of goods were more to do with the validation of custom and social rank.

Something more akin to trade existed at least as far back as the last ice age. However, this trade was

largely peripheral and occurred at the meeting of different tribal groups. As Max Weber (1927, p. 195) attested, it did not take place ‘between members of the same tribe or of the same community’ but was ‘in the oldest social communities an external phenomenon, being directed only towards foreign tribes’. This contention that trade began externally and between communities rather than within them has withstood subsequent scholarly examination. Trade was typically a collective and *inter*-social enterprise between one tribe and another.

With the rise of the ancient civilizations, both external and internal trade increased substantially. The development of money and coinage facilitated its expansion. A definable internal commodity market (or *agora*), with multiple buyers and sellers, first appeared in a designated open space in Athens in the sixth century bc (Polanyi, 1971; North, 1977). The *agora* opened frequently and had strict trading rules. At around the same time there existed an annual auction market on Babylonia: young women were put on display and male bidders competed for marriage rights (Cassady, 1967). Nevertheless, some scholars have warned against the view that these ancient civilizations were generally and predominantly market economies (Finley, 1962; Polanyi, Arensberg and Pearson, 1957). By contrast, researchers such as Peter Temin (2006) have argued that the Roman Empire in particular contained developed and interlocking markets with variable prices, albeit without a highly developed banking system and with a relatively limited market for capital.

European and Mediterranean trade contracted after the fall of the Roman Empire. When commerce began to develop again in medieval times, internal markets then had a limited role in the medieval economy. ‘Strange though it may seem’, wrote the historian Henri Pirenne (1937, p. 140), ‘medieval commerce developed from the beginning not of local but of export trade.’ Although there are likely to have been other earlier organized markets in England, systematic evidence of the king enforcing his right to license all markets and fairs does not appear until the 13th century.

Markets for slaves existed in classical antiquity and persisted in some regions until the modern era. By contrast, feudal serfs were not owned as chattels, but they did not enjoy the right to choose their masters. Feudal institutions, driven by traditional obligations rather than voluntary contract, meant that the hiring of labourers was marginalized and markets for wage labourers were rare. With the decline of bonded labour, which began as early as the 14th century in England, employment contracts were limited largely to casual labourers, alongside a large number of self-employed producers and others in peasant family units. In England it was not until about the 18th century that a class of potentially mobile wage labourers emerged who constituted the most important source of labour power. Organized markets for employees, involving labour exchanges or employment agents, did not become prominent until the 19th century.

To turn to capital markets, an early market for debts was the French *courratier de change* in the 12th century. In the 13th century, after the development of a banking system in Venice, trade began in government securities in several Italian cities. In 1309 a ‘Beurse’ was organised in Bruges in Flanders, named after the Van der Beurse family, who had previously hosted regular commodity exchanges in their residence. Soon after, similar ‘Beurzen’ opened in Ghent and Amsterdam. In 1602 the Dutch East India Company issued the first shares on the Amsterdam Bourse or Stock Exchange. The London Stock Exchange, founded in 1801, traces its origins to 1697 when commodity and stock prices began to be published in a London coffee house. The origins of the New York Stock Exchange go back to 1792, when 24 stockbrokers organized a regular market for stocks in Wall Street. Accordingly, developed capital markets first appeared in the Netherlands in the 17th century and later spread to other countries.

Overall, in the last 400 years markets have expanded enormously in scope, volume and economic importance. Markets have come to pervade internal as well as external trade and to dominate the global economic system. The modern era of globalization is often identified with the growth of global commodity and financial markets since the middle of the 19th century.

This brief historical sketch is background for the task of defining markets. At least three options emerge, involving different degrees of historical specificity. The broadest definition would be to use the term ‘market’ to refer to all forms of transfer of goods or services between persons, including the age-old customary or ceremonial transfers within tribes and households, exchanges of property between tribes, and modern organized markets with multiple buyers and sellers. We consider this option and some more restrictive alternatives next.

The nature of markets

The Austrian school economist Ludwig von Mises (1949) is exceptional among economists in devoting a lengthy chapter to ‘the market’. He sees the market economy as ‘the social system of the division of labour under private ownership of the means of production’ (1949, p. 257). He explicitly excludes economies under social or state ownership of the means of production from this category, but nevertheless regards such systems as strictly ‘not realizable’. Consequently, the historical and territorial boundaries of his concept of the market depend very much on what is regarded as ‘private ownership’. He associates private ownership with the rise of civilization, and defines ownership in terms of full control of the services that derive from a good, rather than in legal terms. Together these specifications amount to a definition of the market that embraces all forms of trade or exchange that involve private property, defined loosely as assets under private control.

Although von Mises associates secure private property and exchange with the rise of civilization, these terms are defined in a manner that does not exclude their application to earlier periods of human history. It then becomes problematic whether or not ceremonial transfers and ritualistic gift-giving are regarded as ‘exchanges’ of ‘property’ and whether or not these activities come within the sphere of ‘the market’. Essentially, the historical compass of the latter term depends very much on what we mean by notions such as exchange and property.

In downplaying the legal aspects of property and exchange, von Mises also fails to probe the nature of the rights that form part of the exchange. Instead he sustains the notion that uncoerced and informed consent by the parties to the transaction is a sufficient basis to constitute the contractual and property rights involved. A problem with this idea is that mutual individual consent itself requires a legislative and institutional framework to legitimize, scrutinize and protect those individual rights. The importance of this legal and constitutional framework is widely recognized, including by other Austrian theorists such as Friedrich Hayek (1960). Several historical cases of the spontaneous evolution of systems of enforced property rights do exist, but they generally rely on reputational and other monitoring mechanisms that are more difficult to sustain in large-scale, complex societies (Sened, 1997).

An alternative intellectual tradition places more emphasis on the legal and statutory basis of individual rights. This approach pervaded the 19th-century German Historical School and their predecessors such as Karl Heinrich Rau, and continued into the 20th century in the original American Institutionalist School, particularly in the writings of John Rogers Commons. Both Rau and Commons (1924) argued

that exchange is more than a voluntary and reciprocal transfer of resources: it also involves the contractual interchange of statutory property rights. For them, exchange had to be understood and analysed in terms of the key institutions that are required to sustain it.

This narrower and more legalistic understanding of private property and contractual exchange confine them both in longevity and scale. Statutorily endorsed property rights, applied to moveable goods and services, were not codified until the ancient civilizations. In feudal times, much of the transfer of goods and services was achieved by custom or coercion rather than by contract and consent. Indeed, economic historians such as North (1981), who attempt to explore the origins of modern markets and commodity exchange, generally focus on the late medieval or early modern period as the era in which well-defined individual property rights began to spread widely from specific parts of the world.

A second important dilemma emerges. This is whether the market is regarded as coextensive with the exchange of private property per se or whether it is given an even narrower meaning and used to refer to forms of *organized* exchange activity. Two major factors lead us to consider an even narrower meaning for the term.

The first consideration is the commonplace use of the term ‘market’ itself and its equivalent in other languages. The word ‘market’ originally appeared as a noun to describe a specific place where people gathered and exchanges of a particular kind took place. The first market in Athens in the sixth century bc had rules concerning who could buy or sell, what could be bought or sold, and how trading should take place. In medieval England markets were permitted by royal charters and located in specific towns. In Europe and elsewhere in the last 300 years organized town and village markets have become commonplace. There are also permanent buildings that function as ‘exchanges’ for agricultural products, minerals, financial stocks, and so on. Although it has acquired additional meanings, the noun ‘market’ still refers to a place or gathering where trade is organized.

The second issue is the existence of a well-researched form of exchange that takes place in different contexts and involves different considerations. In three seminal and influential works, George B. Richardson (1972), Victor P. Goldberg (1980) and Ronald Dore (1983) point out that many real-world commercial transactions do not take place in the competitive arena of a market. Instead they involve firms in ongoing contact, which exchange relevant information before, during and after the contract itself. The relationship is durable and the contract is often renewed. This is most often described as ‘relational exchange’. A question in the derivative literature is to examine the reason for the mutual choice of an ongoing exchange relationship rather than the more competitive institution of the market. Among the explanations is the importance of establishing ongoing trust in circumstances of uncertainty where product characteristics are complex, relatively unique or involve continuous potential improvements. Whatever the reason for its existence, such relational contracting is very different from the more anonymous exchanges in organized markets. Relational exchanges are nevertheless still contractual exchanges of property rights, in their fullest and most meaningful sense. If they are distinguished by definition from market exchanges, then not all exchanges take place in markets. Furthermore, the exchange of goods or services that are strictly unique may be regarded as a non-market phenomenon, even if the exchange is not relational. The term ‘market’ is thus reserved for forms of exchange activity with many similar exchanges involving multiple buyers or sellers.

In part, it is the degree of organization of exchange activity that makes markets different from relational exchange. In financial markets, for example, there are typically strict rules concerning who can trade and how trading should be conducted. In such relatively volatile markets, specific institutions sift

information and present it to traders to help the formation of price expectations and norms. Market institutions in other contexts monitor the quality of goods and the instruments of weight and measure. Within these structures, trading networks emerge on the basis of business connections and reputations. Modern telecommunications mean that a market does not have to be organized in a specific location. Either bidders can communicate with the market centre over long distances, as with many financial markets, or the market *place* can itself disappear, as in the case of Internet-based markets, such as eBay. The latter case nevertheless remains a market, because it is an organized (virtual) forum, subject to specific procedures and rules.

We thus arrive at a definition of a market in the following terms. Markets involve multiple exchanges, multiple buyers or multiple sellers, and thereby a degree of competition. A market is defined as an institution through which multiple buyers or multiple sellers recurrently exchange a substantial number of similar commodities of a particular type. Exchanges themselves take place in a framework of law and contract enforceability. Markets involve legal and other rules that help to structure, organize and legitimize exchange transactions. They involve pricing and trading routines that help to establish a consensus over prices, and often help by communicating information regarding products, prices, quantities, potential buyers or possible sellers. Markets, in short, are organized and institutionalized recurrent exchange.

Of course, it is often difficult to draw the line between organized and relational exchange. There are many possible intermediate cases. However, such difficulties are typical when dealing with highly varied phenomena and are commonplace in some other sciences, notably biology. Similar difficulties exist in distinguishing other economic forms, such as making the important distinction between employment contracts and contracts for services. Nevertheless, such distinctions are important. The difficulty of defining a species does not mean that species should not be defined.

The operation of the law of one price is often taken as an indication of the existence of a market. Of course, imperfect information and quality variations can explain variations within a market from a single price. Nevertheless, the organized competition of the market and its associated information facilities are necessary institutional conditions for any gravitation by similar commodities to a single price level.

Taking stock, we may contrast the narrower definition of the market given above – as an institution with multiple buyers or multiple sellers, and recurrent exchanges of a specific type of commodity – with the much broader definitions raised earlier. These differences in definition do not simply affect the degree of historical specificity of ‘market’ phenomena, they also sustain different theoretical frameworks and promote different questions for research. Some explanations for this divergence arise in the next section.

Why have economists neglected the institutional character of markets?

For much of the 20th century, the institutional character of markets has been neglected by economists because institutions generally have been neglected. The exceptions consist of economists who placed a special emphasis on institutions. The institutional character of markets was emphasised by German historical economists such as Gustav Schmoller and Werner Sombart in the 19th century (Hodgson, 2001). The British dissident economist John A. Hobson (1902, p. 144) wrote: ‘A market, however crudely formed, is a social institution’. Likewise, for the American institutionalist John Maurice Clark (1957, p. 53): ‘the mechanism of the market, which dominates the values that purport to be economic, is

not a mere mechanism for neutral recording of people's preferences, but a social institution with biases of its own'. Coase, North and others have effectively revived an interest in the institutional structure of markets that was eclipsed by developments in mainstream economics during much of the 20th century.
A further clue to help explain why generations of economists have neglected the institutional character of markets lies in the preceding section, where the problem of defining the boundaries of key concepts such as property, exchange and market was raised. Many economists have maintained that the principles of the subject should be as universal as possible – like physics – to the extent that substantial consideration of historically or nationally specific institutional structures is lost. The idea that economics should be defined as a general 'science of choice' (Robbins, 1932) is part of this tradition. Consequently, terms such as property, exchange and market are given a wide meaning. Accordingly, many forms of human interaction have been regarded as 'exchange' and the summation of such 'exchanges' as 'markets'. In these terms, there is little difficulty in applying these concepts to many different types of system, from tribal societies through classical antiquity to the modern capitalist world.
Consequently, the idea of the market assumes a de-institutionalized form, as if it was the primeval and universal ether of all human interactions. Whenever people gather together in the name of self-interest, then a market somehow emerges in their midst. The market springs up simply as a result of these spontaneous interactions: it results neither from a protracted process of multiple institution-building nor from the full development of a historically specific commercial culture.
Incidentally, many sociologists have also assumed a de-institutionalized concept of the market. This is partly the result of the influence of a notion, promoted by Talcott Parsons and others, that sociology should also aspire to a high degree of historical generality. It is also a result of the influence of Marxism within sociology. Despite its emphasis on historical specificity, Marxism also treats markets as uniform entities, ultimately permeated by just one specific set of pecuniary imperatives and cultural norms.
From the 1940s to the 1970s, general equilibrium theory provided the framework in which economists attempted to understand the functioning of markets in wide-ranging terms. Even here, however, some significant attention had to be paid to institutional mechanisms and structures. Something special like the 'Walrasian auctioneer' had to be assumed in order to make the model work (Arrow and Hahn, 1971).
Some elemental institutional structures had to be brought in to make the model function in its own terms. The limits to this project of theoretical generalization became more apparent in the 1970s, when it was shown that few general conclusions could be derived. In particular, Hugo Sonnenschein (1972) and others demonstrated that within general equilibrium theory the aggregated excess demand functions can take almost any form (Rizvi, 1994).
The existence of 'missing markets' always poses a problem for the general equilibrium approach: a complete set of markets for all present and future commodities in all possible states of the world is typically assumed as a basis for general clearance in all markets. However, if market institutions are themselves scarce and costly to establish, then some may be missing for that reason. Furthermore, while capitalism has historically promoted market institutions, modern developed capitalism prohibits several types of market, such as markets for slaves, votes, drugs, or futures markets for labour. In so far as capitalism makes such prohibitions, 'missing markets' are inevitable within capitalism.
The technical problems exposed by Sonnenschein and others led economists to shift their attention away from general equilibrium theory. Instead, game theory became the cutting edge of theoretical analysis. By its nature, game theory tends to lead to less general propositions and points instead to more specific rules and institutions. As game theory became fashionable in the 1980s, it became a theoretical tool in

the ‘new institutionalist’ revival in economic theory.

The revival of the notion of markets as institutions

At least three further developments helped to promote the study of markets as social institutions. First, the basic theory of auctions emerged in the 1970s and 1980s (McAfee and McMillan, 1987; Wolfstetter, 1996). It was assumed from the outset that participants in an exchange did not have complete information, and on this basis it was shown that choices concerning auction forms and rules could significantly affect market outcomes. These ideas assumed centre stage in the 1990s with the use by governments of auction mechanisms in electricity and telecommunications deregulation, most notably in the selling of the electromagnetic spectrum for telecommunications services, and subsequently with the growth of auctions on the Internet (McAfee and McMillan, 1996).

A second and closely related development was the rise of experimental economics, which began to be recognized as an important subdiscipline in the 1980s. Modern experimental economists, in simulating markets in the laboratory, have found that they have had also to face the unavoidable problem of setting up its specific institutional structure. Simply calling it a market is not enough to provide the experimenter with the institutionally specific structures and procedural rules. As leading experimental economist Vernon Smith (1982, p. 923) wrote: ‘it is not possible to design a laboratory resource allocation experiment without designing an institution in all its detail’. Work within experimental economics has underlined the importance of these specific rules, by showing that market outcomes are sometimes relatively insensitive to the information processing capacities of the agents involved, because particular constraints govern the results (Gode and Sunder, 1993).

In reality, each particular market is entwined with other institutions and a particular social culture. Accordingly, there is not just one type of market but many different markets, each depending on its inherent routines, cultural norms and institutional make-up. Differentiating markets by market structure according to textbook typology – from perfect competition through oligopoly to monopoly – is not enough. Institutions, routines and culture have to be brought into the picture. Experimental economists have discovered an equivalent truth in laboratory settings, and have learned that experimental outcomes often depend on the tacit assumptions and cultural settings of participants. Different types of market institution are possible, involving different routines, pricing procedures, and so on. This has been acknowledged by a growing number of economists, as the notion of a single universal type of market has lost credibility (McMillan, 2002).

Third, these theoretical developments were dramatized by events. Following the collapse of the Eastern bloc in 1989–91, a number of economists presumed that many markets would emerge spontaneously in the vacuum created by the breakdown of central planning. This view turned out to be mistaken, as capital and other markets were slow to develop and their growth was thwarted by the lack of an appropriate institutional infrastructrure. Several formerly planned economics slipped back into recession. Critics such as Coase (1992, p. 718) drew attention to the necessary institutional foundations of the market system: ‘The ex-communist countries are advised to move to a market economy … but without the appropriate institutions, no market of any significance is possible’.

While sociologists, like economists, had previously paid relatively little attention to market institutions, the revitalization of the sub-discipline of economic sociology led to a series of studies by sociologists of

financial and other markets (Abolafia, 1996; Baker, 1984; Burt, 1992; Fligstein, 2001; Lie, 1997; Swedberg, 1994; White, 1981; 1988). These works show how specific networks and social relationships between actors structure exchanges, and how cultural norms govern market operations and outcomes. Similar considerations have emerged in empirical and simulation work by economists that stresses the importance of learning and previous experience in trading partner selection and in the decision to accept a transaction (Kirman and Vignes, 1991; Härdle and Kirman, 1995).

Taken as a whole, these literatures testify to a much more nuanced conception of market phenomena. As a result of all these developments, the treatment by economists and others of markets began to change. Both economists and sociologists are now paying detailed attention to the nature of specific market rules and mechanisms. A milestone paper by Alvin Roth (2002) challenges the view of a single universal theory of market behaviour. While those economists who had paid attention to different market mechanisms had typically been preoccupied with a search for ‘optimal’ rules and institutional forms, gradually this has become a will-o’-the-wisp with the realization that typical assumptions in the emerging literature concerning cognitive and information impairments have made this search difficult or impossible (Lee, 1998; Mirowski, 2007).

Nevertheless, while the search for optimal institutional blueprints is intractable, these theoretical developments have begun to provide an analytical framework within which the limits and potentialities of different types of market mechanism can be appraised. An outcome is to abandon the former widespread notion – shared by all kinds of theorists from Marxists to the Austrian School – that ‘the market’ is a singular type of entity entirely understandable in terms of the same principles or laws. While Hayek and his followers should be given inspirational credit for their emphasis on the informational limitations inherent in all complex economic systems, they stressed that markets are the most effective processors of information while downplaying or ignoring the differences between various types of market.

In this context, markets reappear as varied and historically specific phenomena. The general equilibrium approach has been overshadowed by an array of theoretical and empirical methodologies, including game theory, agent-based modelling, laboratory experimentation and real-world observation.

Conclusions

A number of options for defining a market have been outlined here. The broadest option is to regard the market as the universal ether of human interaction, depending on little more than the division of labour. A second option is to regard the market as synonymous with commodity exchange, in which case it dates at least as far back to the dawn of civilization.

By contrast, several considerations militate in favour of a narrower definition, and recent developments in economic theory point in this direction. In the narrower sense, markets are organized exchange. Where they exist, markets help to structure, organize and legitimize numerous exchange transactions. Pricing and trading procedures within markets help to establish a consensus over prices, and communicate information regarding products, prices, quantities, potential buyers or possible sellers. Variation in market rules and procedures means that markets differ substantially, especially when we consider markets in different cultures. The markets of 2,000 years ago were very different from (say) the electronic financial markets of today. In the real world, and even in a single country, we may come

across many different examples of the market. The market itself is neither a natural datum nor an ubiquitous ether, but is itself a social institution, governed by sets of rules restricting some and legitimizing other behaviours. Furthermore, the market is necessarily entwined with other social institutions, such as in many cases the local or national state. It can emerge spontaneously, but it can also be promoted or guided by conscious design.

A clear implication of this argument is that the unnuanced but familiar pro- and anti-market policy stances are both insensitive to the possibility of different types of market institution. Instead of recognizing the important role of different possible cultures and trading customs, both the opponents and the advocates of the market have focused exclusively on its general features. Thus, for instance, Marxists have deduced that the mere existence of private property and markets will themselves encourage acquisitive, greedy behaviour, with no further reference in their analysis to the role of ideas and culture in helping to form the aspirations of social actors. This is the source of their ‘agoraphobia’, or fear of markets. Obversely, overenthusiastic advocates of the market claim that its benefits stem simply and unambiguously from the existence of private property and exchange, without regard to possible variations in detailed market mechanism or cultural context. As strange bedfellows, both Marxists and some market advocates have underestimated the degree to which all market economies are unavoidably made up of densely layered social institutions.

See Also

- arbitrage
- auctioneer
- auctions (applications)
- Austrian economics
- capitalism
- competition
- competition and selection
- competition, Austrian
- competition, classical
- computing in mechanism design
- contemporary capitalism
- economic sociology
- econophysics
- efficient markets hypothesis
- electricity markets
- existence of general equilibrium
- experimental economics
- general equilibrium
- general equilibrium (new developments)
- institutionalism, old

- labour market institutions
- land markets
- market institutions
- market microstructure
- marketplaces
- marketing boards
- marriage markets
- Marx's analysis of capitalist production
- property rights
- tâtonnement and recontracting
- two-sided markets

Bibliography

Abolafia, M. 1996. *Making Markets: Opportunism and Restraint on Wall Street*. Cambridge, MA: Harvard University Press.

Arrow, K. and Hahn, F. 1971. *General Competitive Analysis*. Edinburgh: Oliver and Boyd.

Baker, W. 1984. The social structure of a national securities market. *American Journal of Sociology* 89, 775–811.

Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Cassady, R. 1967. *Auctions and Auctioneering*. Berkeley and Los Angeles: University of California Press.

Clark, J. 1957. *Economic Institutions and Human Welfare*. New York: Alfred Knopf.

Clarke, D. 1987. Trade and industry in barbarian Europe till Roman times. In *The Cambridge Economic History of Europe, Volume II: Trade and Industry in the Middle Ages*, 2nd edn, ed. M. Postan and E. Miller. Cambridge: Cambridge University Press.

Coase, R. 1988. *The Firm, the Market, and the Law*. Chicago: University of Chicago Press.

Coase, R. 1992. The institutional structure of production. *American Economic Review* 82, 713–19.

Commons, J. 1924. *Legal Foundations of Capitalism*. New York: Macmillan.

Dore, R. 1983. Goodwill and the spirit of market capitalism. *British Journal of Sociology* 34, 459–82.

Finley, M. (ed.) 1962. *Second International Conference of Economic History, Volume. I: Trade and Politics in the Ancient World*. New York: Arno.

Fligstein, N. 2001. *The Architecture of Markets: An Economic Sociology of Twenty-First Century Capitalist Societies*. Princeton, NJ: Princeton University Press.

Gode, D. and Sunder, S. 1993. Allocative efficiency of markets with zero-intelligence traders: market as a partial substitute for individual rationality. *Journal of Political Economy* 101, 119–37.

Goldberg, V. 1980. Relational exchange: economics and complex contracts. *American Behavioral Scientist* 23, 337–52.

Härdle, W. and Kirman, A. 1995. Nonclassical demand: a model-free examination of price quantity relations in the Marseille fish market. *Journal of Econometrics* 67, 227–7.

Hayek, F. 1960. *The Constitution of Liberty*. London and Chicago: Routledge and Kegan Paul, and University of Chicago Press.

Hobson, J. 1902. *The Social Problem: Life and Work*. London: James Nisbet.

Hodgson, G. 2001. *How Economics Forgot History: The Problem of Historical Specificity in Social Science*. London and New York: Routledge.

Kirman, Alan P. and Vignes, A. 1991. Price dispersion: theoretical considerations and empirical evidence from the Marseilles fish market. In *Issues in Contemporary Economics: Proceedings of the Ninth World Congress of the International Economic Association*, ed. K. Arrow. New York: New York University Press.

Lee, R. 1998. *What Is an Exchange? The Automation, Management, and Regulation of Financial Markets*. Oxford: Oxford University Press.

Lie, J. 1997. Sociology of markets. *Annual Review of Sociology* 23, 341–60.

McAfee, R.P. and McMillan, J. 1987. Auctions and Bidding. *Journal of Economic Literature* 25, 699–738.

McAfee, R. and McMillan, J. 1996. Analyzing the airwaves auction. *Journal of Economic Perspectives* 10(1), 199–76.

McMillan, J. 2002. *Reinventing the Bazaar: A Natural History of Markets*. New York and London: Norton.

McMillan, J. 2003. Market design: the policy uses of theory. *American Economic Review (Papers and Proceedings)* 93, 139–44.

Mirowski, P. 2007. Markets come to bits: evolution, computation and markomata in economic science. *Journal of Economic Behavior and Organization*. 63, 209–42.

Mises, L. von. 1949. *Human Action: A Treatise on Economics*. London and New Haven: William Hodge and Yale University Press.

North, D. 1977. Markets and other allocation systems in history: the challenge of Karl Polanyi. *Journal of European Economic History* 6, 703–16.

North, D. 1981. *Structure and Change in Economic History*. New York: Norton.

Pirenne, H. 1937. *Economic and Social History of Medieval Europe*. New York: Harcourt Brace.

Polanyi, K. 1971. *Primitive and Modern Economics: Essays of Karl Polanyi*. Boston: Beacon Press.

Polanyi, K., Arensberg, C. and Pearson, H., eds. 1957. *Trade and Market in the Early Empires*. Chicago: Henry Regnery.

Richardson, G. 1972. The organisation of industry. *Economic Journal* 82, 883–96.

Rizvi, S. 1994. The microfoundations project in general equilibrium theory. *Cambridge Journal of Economics* 18, 357–77.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Roth, A. 2002. The economist as engineer: game theory, experimentation, and computation as tools for design economics. *Econometrica* 70, 1341–78.

Sened, I. 1997. *The Political Institution of Private Property*. Cambridge: Cambridge University Press.

Smith, V. 1982. Microeconomic systems as an experimental science. *American Economic Review* 72, 923–55.

Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40, 549–63.

Stigler, G. 1967. Imperfections in the capital market. *Journal of Political Economy* 75, 287–92.

Swedberg, R. 1994. Markets as social structures. In *Handbook of Economic Sociology*, ed. N. Smelser and R. Swedberg. Princeton: Princeton University Press.

Temin, P. 2006. The economy of the early Roman Empire. *Journal of Economic Perspectives* 20(1), 133–51.

Weber, M. 1927. *General Economic History*. London: Allen and Unwin.

White, H. 1981. Where do markets come from? *American Journal of Sociology* 87, 517–47.

White, H. 1988. Varieties of markets. In *Social Structure: A Network Approach*, ed. B. Wellman and S. Berkowitz. Cambridge, MA: Harvard University Press.

Wolfstetter, E. 1996. Auctions: an introduction. *Journal of Economic Surveys* 10, 367–420.

How to cite this article

Hodgson, Geoffrey M. "markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000402> doi:10.1057/9780230226203.1038

The New Palgrave Dictionary of Economics Online

Markov chain Monte Carlo methods

Siddhartha Chib
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

MCMC methods, an important class of Monte Carlo methods, have played a major role in the growth of Bayesian statistics and econometrics. In an MCMC simulation, one samples a given distribution (say the posterior distribution in a Bayesian model) by simulating a suitably constructed Markov chain whose invariant distribution is the target distribution. The Metropolis–Hastings algorithm and its special case, the Gibbs sampler, are two common ways of devising an MCMC simulation. We discuss how these methods originate, discuss implementation issues and provide examples. The use of MCMC methods in Bayesian prediction and model choice problems is also discussed.

Keywords

autocorrelation; Bayesian econometrics; Bayesian prior–posterior analysis; Bayesian statistics; invariance; latent variables; marginal likelihood; Markov chain Monte Carlo methods; model choice; prediction; proposal densities; reversibility; transition density

Article

1 Introduction

Markov chain Monte Carlo methods, popularly called MCMC methods, are a class of Monte Carlo methods for sampling a given univariate or multivariate probability distribution (the target distribution). These methods play a central role in the theory and practice of modern Bayesian methods where they are used for the numerical calculation of quantities (such as the moments and quantiles of posterior and predictive densities) that arise in the Bayesian prior–posterior analysis. They have transformed the fields of Bayesian statistics and econometrics.

Suppose that in a given Bayesian model the prior density is $\pi(\boldsymbol{\theta})$ and the sampling density or likelihood function is $f(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y}$ is a vector of observations and $\boldsymbol{\theta} \in \Re^d$ is an unknown parameter. In the Bayesian context, inferences about θ are based on the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})$. Now suppose that one is interested in finding the mean of the posterior density

$$E(\boldsymbol{\theta}|\mathbf{y}) = \int \boldsymbol{\theta} \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$$

but that the integral cannot be computed analytically. In that case one can compute the integral by Monte Carlo sampling methods. The general idea is to calculate the integral from a sample

$$\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \sim \pi(\boldsymbol{\theta}|\mathbf{y}),$$

that is drawn from the posterior density. This sample can be used to estimate the posterior mean and other features of the posterior density. For instance, the posterior mean can be estimated by the average of the sampled draws, and the quantiles of the posterior density by the quantiles of the sampled output.

The requisite sampling of the target density is made possible by MCMC methods. In a MCMC simulation, one samples the target density in an indirect way: by simulating a suitably constructed Markov chain whose invariant distribution is the target density. Then the draws beyond some chosen burn-in period are taken as a (correlated) sample from the target density. The defining feature of Markov chains is the property that the conditional density of $\boldsymbol{\theta}^{(j)}$ (the *j*th element of the sequence) conditioned on the entire preceding history of the chain depends only on the previous value $\boldsymbol{\theta}^{(j-1)}$. Denote this conditional density, the transition density of the Markov chain, by $p(\boldsymbol{\theta}^{(j-1)}, \cdot|\mathbf{y})$. Then, in the MCMC framework, a sample is produced by simulating the transition density as

$$\boldsymbol{\theta}^{(1)} \sim p(\boldsymbol{\theta}^{(0)}, \cdot|\mathbf{y}) \vdots \boldsymbol{\theta}^{(j)} \sim p(\boldsymbol{\theta}^{(j-1)}, \cdot|\mathbf{y}) \vdots$$

If we let the first n_0 cycles represent the burn-in phase, for some choice of n_0, the draws

$$\boldsymbol{\theta}^{(n_0+1)}, \boldsymbol{\theta}^{(n_0+2)}, \ldots, \boldsymbol{\theta}^{(n_0+M)}$$

are treated as those from $\pi(\boldsymbol{\theta}|\mathbf{y})$. Even though the sampled variates are correlated, laws of large numbers for Markov sequences can be used to show that, under regularity conditions, the sample average of any integrable function $g(\boldsymbol{\theta})$ converges to its posterior expectation:

$$M^{-1}\sum_{j=1}^{M} g(\boldsymbol{\theta}^{(j)}) \to \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}, \tag{1}$$

as *M* becomes large.

There are two common ways of constructing a transition density $p(\boldsymbol{\theta}^{(j-1)}, \cdot|\mathbf{y})$ whose limiting distribution is the required target density. One way is by a method called the Metropolis–Hastings (M–H) algorithm, which was introduced by Metropolis et al. (1953) and Hastings (1970). Key references about this method are Tierney (1994), and Chib and Greenberg (1995). A second approach is by the so-called Gibbs sampling algorithm. This method was introduced by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990), and was the impetus for the current interest in Markov chain sampling methods. A summary of many aspects of MCMC methods is contained in Chib (2001) while textbook accounts include Gilks, Richardson and Spiegelhalter (1996), Chen, Shao and Ibrahim (2000), Liu (2001) and Robert and Casella (2004).

2 Metropolis – Hastings algorithm

Suppose that we are interested in sampling the target density $\pi(\boldsymbol{\theta}|\mathbf{y})$, where θ is a vector-valued parameter and $\pi(\boldsymbol{\theta}|\mathbf{y})$ is a continuous density. The idea behind the M–H algorithm is to simulate a proposal value $\boldsymbol{\theta}'$ from a transition density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ that is convenient to stimulate but does not necessarily have the correct limiting distribution and then to subject the proposal value to a specific randomization to ensure that the resulting Markov chain has the correct limiting distribution.
To define the M–H algorithm, let $\boldsymbol{\theta}^{(j-1)}$ be the current value. Then the next value $\boldsymbol{\theta}^{(j)}$ is produced by a two-step process consisting of a 'proposal step' and a 'move step'.

- *Proposal step*: Sample a proposal value $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}|\mathbf{y})$ and calculate the quantity

$$\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})}{\pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})} \frac{q(\boldsymbol{\theta}', \boldsymbol{\theta}^{(j-1)}|\mathbf{y})}{q(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})}\right\}. \quad (2)$$

- *Move step*: Let $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}'$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$; remain at the current value $\boldsymbol{\theta}^{(j-1)}$ with probability $1 - \alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$.

In terms of nomenclature, the source density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ is called the candidate generating density or proposal density, and $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$ the *acceptance probability* or, more descriptively, the *probability of move*. Note also that the function $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$ in this algorithm can be computed without knowledge of the norming constant of the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. In addition, if the proposal density is symmetric, satisfying the condition $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$, then the acceptance probability reduces to $\pi(\boldsymbol{\theta}'|\mathbf{y}) / \pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$; hence, if $\pi(\boldsymbol{\theta}') \geq \pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$, the chain moves to θ', otherwise it moves to θ' with probability given by $\pi(\boldsymbol{\theta}'|\mathbf{y}) / \pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$. The latter is the algorithm of Metropolis et al. (1953).
Remark 1: Derivation of the M–H algorithm: A question of some interest is the justification of this two-step approach. This question was tackled by Chib and Greenberg (1995) who derived the method from the logic of reversibility. A Markov transition density $p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ is said to be reversible for $\pi(\boldsymbol{\theta}|\mathbf{y})$ if the following condition holds for every $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in the support of the target distribution:

$$\pi(\boldsymbol{\theta}|\mathbf{y})\, p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \pi(\boldsymbol{\theta}'|\mathbf{y})\, p(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}). \quad (3)$$

The reversibility condition is important because reversible chains are invariant. Invariance refers to the property that

$$\pi(\boldsymbol{\theta}'|\mathbf{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\, \pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \quad (4)$$

which means that, if the transition density is invariant for the target density, then, once convergence is achieved, a subsequent value θ' drawn from the transition density is also from the target density. To see that reversibility implies invariance one simply integrates both sides of (3) over θ. This leads to the invariance condition since $\int p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\, d\boldsymbol{\theta}' = 1$

by virtue of being a transition density. Now consider the Markov chain induced by the proposal density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$. Because this was formulated without the reversibility condition in mind it is unlikely to satisfy reversibility. Suppose that for a pair of points $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ it is true that

$$\pi(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) > \pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}), \tag{5}$$

which means informally that the process moves from θ to θ' too frequently and too rarely in the reverse direction. This situation can be corrected by reducing the flow from θ to θ' by introducing probabilities $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ and $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ of making the moves in either direction so that

$$\pi(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}).$$

One now sets $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ to be as high as possible, namely, equal to 1. Solving for $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ one then gets

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\frac{q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})}.$$

Because one started from (5) this is clearly less than 1. On the other hand, if the inequality in (5) were reversed, the same argumentation leads to the conclusion that $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = 1$. Thus, on combining these two cases we reproduce the expression of $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ given in (2).

Remark 2: Transition density of the M–H chain: The transition density of the M–H chain has two components – one for the move away from θ and given by $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ and one for the probability of staying at θ given by $r(\boldsymbol{\theta}|\mathbf{y}) = 1 - \int\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\,d\boldsymbol{\theta}'$. In particular,

$$p_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) + \delta_{\theta}(\boldsymbol{\theta}')r(\boldsymbol{\theta}|\mathbf{y})$$

where $\delta_{\theta}(\boldsymbol{\theta}')$ is the Dirac-function at θ defined as $\delta_{\theta}(\boldsymbol{\theta}') = 0$ for $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ and $\int\delta_{\theta}(\boldsymbol{\theta}')\,d\boldsymbol{\theta}' = 1$. It is easy to check that the integral of the transition density over all possible values of θ is 1, as required.

Remark 3: Convergence properties: The theoretical properties of the M–H algorithm (in particular the ergodic behaviour of the chain from an arbitrarily specified initial value) depend crucially on the nature of the proposal density. One requirement is that the proposal density be everywhere positive in the support of the posterior density, which means that the M–H chain can make a transition to any point in its support in one step. Further discussion of the conditions is given in Tierney (1994) and Robert and Casella (2004).

Remark 4: Mixing: The sampled values from the M–H algorithm (as from any Markov chain) are correlated. The goal in any particular application is to ensure that the serial correlation is not excessive. One diagnostic to check for the degree of serial correlation in the sampled draws is the *autocorrelation time* or *inefficiency factor* of each component θ_k of θ defined as

$$a_k = \left\{1 + 2\sum_{s=1}^{M}(1 - \frac{s}{M})\rho_{ks}\right\},$$

where ρ_{ks} is the sample autocorrelation at lag *s* from the *M* sampled draws $\theta_k^{(n_0+1)}, \ldots, \theta_k^{(n_0+M)}$. One can interpret this quantity in terms of the *effective sample size*, or ESS, defined for the *k*th component of θ as $ESS_k = \frac{M}{a_k}$. With independent sampling the autocorrelation times are theoretically equal to 1, and the effective sample size is *M*. When the inefficiency factors are high, the effective sample size is much smaller than *M*.

2.1 Choice of proposal density

One family of candidate-generating densities is given by $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}' - \boldsymbol{\theta})$. The candidate θ' is thus drawn according to the process $\boldsymbol{\theta}' = \boldsymbol{\theta} + \mathbf{z}$, where **z** follows the distribution *q*, and is called the *random walk M–H* chain. The random walk M–H chain is quite popular in applications. One has to be careful in setting the variance of **z** because if it is too large the chain may remain stuck at a particular value for many iterations, while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution.

Another possibility is to let $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}'|\mathbf{y})$, an *independence M–H chain* in the terminology of Tierney (1994). One way to implement such chains is by tailoring the proposal density to the target at the mode by a multi-variate normal or multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode (Chib and Greenberg, 1994; 1995).

Yet another way to generate proposal values is through a Markov chain version of the accept–reject method (Tierney, 1994; Chib and Greenberg, 1995). To explain this method, suppose $c > 0$ is a known constant and $h(\theta)$ a source density. Let $C = \{\boldsymbol{\theta}: \pi(\boldsymbol{\theta}|\mathbf{y}) \leq ch(\boldsymbol{\theta})\}$ denote the set of value for which $ch(\theta)$ dominates the target density. Given $\boldsymbol{\theta}^{(j-1)} = \boldsymbol{\theta}$ the next value $\boldsymbol{\theta}^{(j)}$ is obtained as follows. First, a candidate value $\boldsymbol{\theta}'$ is obtained, independent of the current value θ, by applying the accept–reject algorithm with $ch(\cdot)$ as the 'pseudo-dominating' density. The candidates θ' that are produced under this scheme have density $q(\boldsymbol{\theta}'|\mathbf{y}) \propto \min\left\{\pi(\boldsymbol{\theta}'|\mathbf{y}), ch(\boldsymbol{\theta}')\right\}$. Then, the M–H probability of move is given by

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{\theta} \in C \\ 1 / w(\boldsymbol{\theta}) & \text{if} \quad \boldsymbol{\theta} \notin C, \boldsymbol{\theta}' \in C \\ \min\left\{w(\boldsymbol{\theta}') / w(\boldsymbol{\theta}), 1\right\} & \text{if} \quad \boldsymbol{\theta} \notin C, \boldsymbol{\theta}' \notin C \end{cases} \tag{6}$$

where $w(\boldsymbol{\theta}) = c^{-1}\pi(\boldsymbol{\theta}|\mathbf{y}) / h(\boldsymbol{\theta})$.

2.2 Example

To illustrate the M–H algorithm, consider the binary response data in Table 1, on the occurrence or non-occurrence of infection following birth by Caesarean section. The response variable *y* is 1 if the Caesarean birth resulted in an infection, and zero if not. There are three covariates: x_1, an indicator of whether the caesarean was non-planned; x_2, an indicator of

whether risk factors were present at the time of birth; and x_3, an indicator of whether antibiotics were given as a prophylaxis. The data in the table contains information from 251 births. Under the column of the response, an entry such as 11/87 means that there were 98 deliveries with covariates (1,1,1) of whom 11 developed an infection and 87 did not. Suppose that the probability of infection for the *i*th birth $(i \leq 251)$ is

$$\Pr(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}), \tag{7}$$

$$\boldsymbol{\beta} \sim N_4(\mathbf{0}, 5\mathbf{I}_4) \tag{8}$$

where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ is the covariate vector, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ is the vector of unknown coefficients, Φ is the cdf of the standard normal random variable and $\mathbf{I}_4$ is the four-dimensional identity matrix. The target posterior density, under the assumption that the outcomes $\mathbf{y} = (y_1, y_2, \ldots, y_{251})$ are conditionally independent, is

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{251} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i} \left\{1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})\right\}^{(1-y_i)}$$

where $\pi(\boldsymbol{\beta})$ is the density of the $N(0,10\mathbf{I}_4)$ distribution.

Caesarean infection data

y(1/0)	x_1	x_2	x_3
11/87	1	1	1
1/17	0	1	1
0/2	0	0	1
23/3	1	1	0
28/30	0	1	0
0/9	1	0	0
8/32	0	0	0

Source: Fahrmeir and Tutz (1994).

To define the Chib and Greenberg (1994) tailored proposal density, let

$$\hat{\boldsymbol{\beta}} = (-1.093022 \quad 0.607643 \quad 1.197543 \quad -1.904739)'$$

be the maximum likelihood estimate and let

$$\mathbf{V} = \begin{pmatrix} 0.040745 & -0.007038 & -0.039399 & 0.004829 \\ & 0.073101 & -0.006940 & -0.050162 \\ & & 0.062292 & -0.016803 \\ & & & 0.080788 \end{pmatrix}$$

be the symmetric matrix obtained by inverting the negative of the Hessian matrix (the matrix of second derivatives) of the log-likelihood function evaluated at $\hat{\boldsymbol{\beta}}$. To generate proposal values, we use a multivariate-t density with 15 degrees of freedom, location given by $\hat{\boldsymbol{\beta}}$ and dispersion given by **V**. The M–H algorithm is run for 5,000 iterations beyond a burn-in of 100 iterations. The prior–posterior summary is reported in Table 2. It contains the first two moments (the mean and the standard deviation) of the prior and posterior and the 2.5th (lower) and 97.5th (upper) percentiles of the marginal densities of β.

Caesarean data: prior–posterior summary based on 5,000 draws (beyond a burn-in of 100 cycles) from the tailored M–H algorithm

	Prior		Posterior			
	Mean	Std dev	Mean	Std dev	Lower	Upper
β_0	0.000	2.236	–1.080	0.220	–1.526	–0.670
β_1	0.000	2.236	0.593	0.249	0.116	1.095
β_2	0.000	2.236	1.181	0.254	0.680	1.694
β_3	0.000	2.236	–1.889	0.266	–2.421	–1.385

In addition, we plot in Figure 1 the four marginal posterior densities. These are derived by smoothing the histogram of the simulated values with a Gaussian kernel. In the same plot we also report the autocorrelation functions (correlation against lag) for each of the sampled parameter values. The serial correlations decline quickly to zero indicating that the algorithm is mixing well.

Figure 1

Caesarean data with tailored M–H algorithm: marginal posterior densities (top panel) and autocorrelation plot (bottom panel)

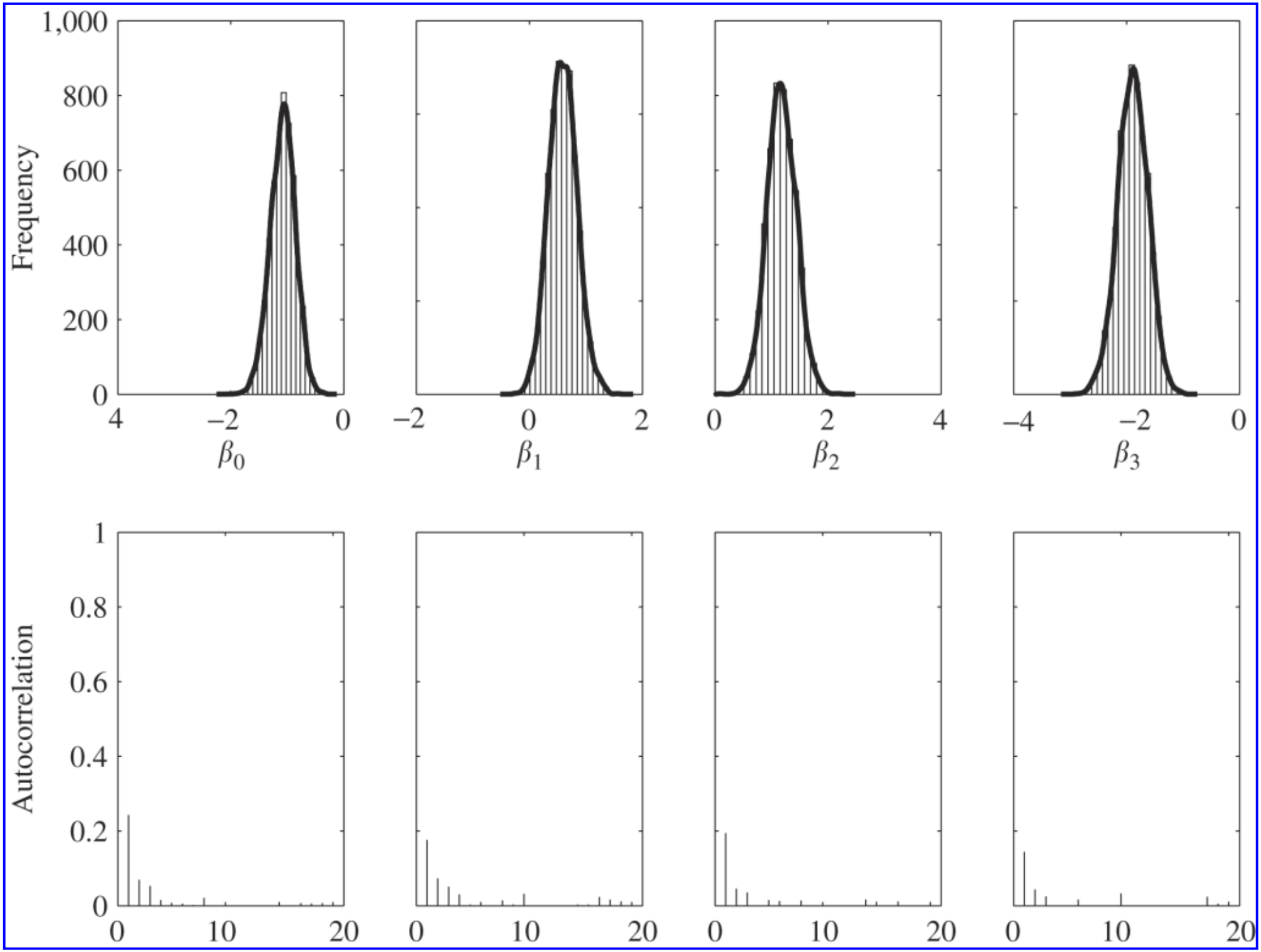
Frequency
1,000
800
600
400
200
0
4
−2
0
β_0
−2
0
2
β_1
0
2
4
β_2
−4
−2
0
β_3
Autocorrelation
1
0.8
0.6
0.4
0.2
0
0
10
20

2.3 Multiple-block M – H algorithm

When the dimension of θ is large it is often necessary to divide the parameters into smaller groups or blocks and then to sample the blocks in turn. For simplicity suppose that two blocks are adequate and that θ is written as $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, with $\boldsymbol{\theta}_k \in \Omega_k \subseteq \Re^{d_k}$. To sample these blocks let

$$q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2); \; q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1),$$

denote the two proposal densities, one for each block θ_k, where the proposal density q_k may depend on the current value of the remaining block. Also, define

$$\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) = \min\left\{\frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}) q_1(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2)}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2)}, 1\right\}$$

and

$$\alpha(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1) = \min\left\{\frac{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2' | \mathbf{y}) q_2(\boldsymbol{\theta}_2', \boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1)}, 1\right\},$$

as the probability of move for block θ_k conditioned on the other block. Then, in what may be called the multiple-block M–H algorithm, one updates each block using an M-H step with the above probability of move, given the most current value of the other block. The method can be extended to several blocks in the same way.

Remark 5: An important special case arises if each proposal density is the full conditional density of that block. Specifically, if we set

$$q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}), \quad q_1(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$$

and

$$q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1) \propto \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2' | \mathbf{y}), \quad q_2(\boldsymbol{\theta}_2', \boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1) \propto \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$$

then an interesting simplification occurs. The probability of move (for the first block) becomes

$$\alpha_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2|\mathbf{y})\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2|\mathbf{y})}\right\} = 1,$$

and similarly for the second block, implying that, if proposal values are drawn from their full conditional densities, then the proposal values are accepted with probability one. This special case is called the Gibbs sampling algorithm.

3 The Gibbs sampling algorithm

The Gibbs sampling was introduced by Geman and Geman (1984) in the context of image processing and then discussed in the context of missing data problems by Tanner and Wong (1987). It was brought into prominence by Gelfand and Smith (1990) who demonstrated its use in a range of Bayesian problems.

3.1 The algorithm

Suppose that the parameters are grouped into two *p* blocks $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)$ with the associated set of full conditional distributions

$$\{\pi(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p); \pi(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_p); \ldots, \pi(\boldsymbol{\theta}_p|\mathbf{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{d-1})\},$$

where each full conditional distribution is proportional to $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)|\mathbf{y})$. Then, one cycle of the Gibbs sampling algorithm is completed by simulating $\{\boldsymbol{\theta}_k\}_{k=1}^{p}$ from these distributions, recursively refreshing the conditioning variables.

3.2 Sufficient conditions for convergence

Under rather general conditions, the Markov chain generated by the Gibbs sampling algorithm converges to the target density as the number of iterations become large. Formally, if we let $p_G(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ represent the transition density of the Gibbs algorithm and let $p_G^{(M)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}|\mathbf{y})$ be the density of the draw $\boldsymbol{\theta}'$ after *M* iterations given the starting value $\boldsymbol{\theta}_0$, then

$$\| p_G^{(M)}(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}'|\mathbf{y}) - \pi(\boldsymbol{\theta}'|\mathbf{y}) \| \to 0, \quad \text{as} \quad M \to \infty. \tag{9}$$

Roberts and Smith (1994) (see also Chan, 1993) have shown that this convergence occurs under the following weak conditions: (i) $\pi(\boldsymbol{\theta}|\mathbf{y}) > 0$ implies there exists an open neighbourhood N_θ containing θ and $\varepsilon > 0$ such that, for all $\boldsymbol{\theta}' \in N_\theta$, $\pi(\boldsymbol{\theta}'|\mathbf{y}) \geq \varepsilon > 0$; (ii) $\int \pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}_k$ is locally bounded for all *k*, where θ_k is the *k*th block of parameters; and (iii) the support of θ is arc connected.

4 MCMC sampling with latent variables

MCMC sampling can involve not just parameters but also latent variables. This idea was called data augmentation by Tanner and Wong (1987) in the context of missing data problems.

To fix notations, suppose that **z** denotes a vector of latent variables and let the modified target distribution be $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. If the latent variables are tactically introduced, the conditional distribution of θ (or sub-components of θ) given **z** may be easy to derive. Then, a multiple-block M–H simulation is conducted with the blocks θ and **z** leading to the sample

$$\left(\boldsymbol{\theta}^{(n_0+1)}, \mathbf{z}^{(n_0+1)}\right), \ldots, \left(\boldsymbol{\theta}^{(n_0+M)}, \mathbf{z}^{(n_0+M)}\right) \sim \pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}),$$

where the draws on θ, ignoring those on the latent data, are from $\pi(\boldsymbol{\theta}|\mathbf{y})$, as required.

To demonstrate this technique in action, consider the probit regression example discussed in Section 2.2. Albert and Chib (1993) introduced a technique for this and related models that capitalizes on the simplifications afforded by introducing latent data into the sampling. The Albert–Chib method has found wide use and has made possible the routine analysis of models for categorical responses. To begin, let

$$z_i|\boldsymbol{\beta} \sim N(\mathbf{x}_i'\boldsymbol{\beta}, 1),$$

$$y_i = I[z_i > 0], \quad i \leq n,$$

$$\boldsymbol{\beta} \sim N_K(\boldsymbol{\beta}_0, \mathbf{B}_0). \tag{10}$$

This specification is equivalent to the probit model since $\Pr(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = \Pr(z_i > 0|\mathbf{x}_i, \boldsymbol{\beta}) = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$. Now the MCMC sampling is based on the full conditional distributions

$$\boldsymbol{\beta}|\mathbf{y}, \{z_i\}; \quad \{z_i\}|\mathbf{y}, \boldsymbol{\beta},$$

which are both tractable. In particular, the distribution of β conditioned on the latent data becomes independent of the observed data and has the same form as in the Gaussian linear regression model with the response data given by $\{z_i\}$ and is multivariate normal with mean $\hat{\boldsymbol{\beta}} = \mathbf{B}(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \Sigma_{i=1}^n \mathbf{x}_i z_i)$ and variance matrix $\mathbf{B} = (\mathbf{B}_0^{-1} + \Sigma_{i=1}^n \mathbf{x}_i\mathbf{x}_i')^{-1}$. Next, the distribution of the latent data conditioned on the data and the parameters factor into a set of n independent distributions with each depending on the data through y_i:

$$\{z_i\}|\mathbf{y}, \boldsymbol{\beta} \overset{d}{=} \prod_{i=1}^{n} z_i|y_i, \boldsymbol{\beta},$$

where the distribution $z_i|y_i, \beta$ is the normal distribution $z_i|\beta$ truncated by the knowledge of y_i; if y_i=0, then $z_i \leq 0$ and if y_i=1, then z_i>0. Thus, one samples z_i from $TN_{(-\infty,0)}(\mathbf{x}_i'\beta, 1)$ if y_i=0 and from $TN_{(0,\infty)}(\mathbf{x}_i'\beta, 1)$ if y_i=1, where $TN_{(a,b)}(\mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ distribution truncated to the region (a,b).

We apply this method to the example considered in Section 2.2 above and report the results in Figure 2. We see the close agreement between the two sets of results.

Figure 2

Caesarean data with Albert–Chib algorithm: marginal posterior densities (top panel) and autocorrelation plot (bottom panel)

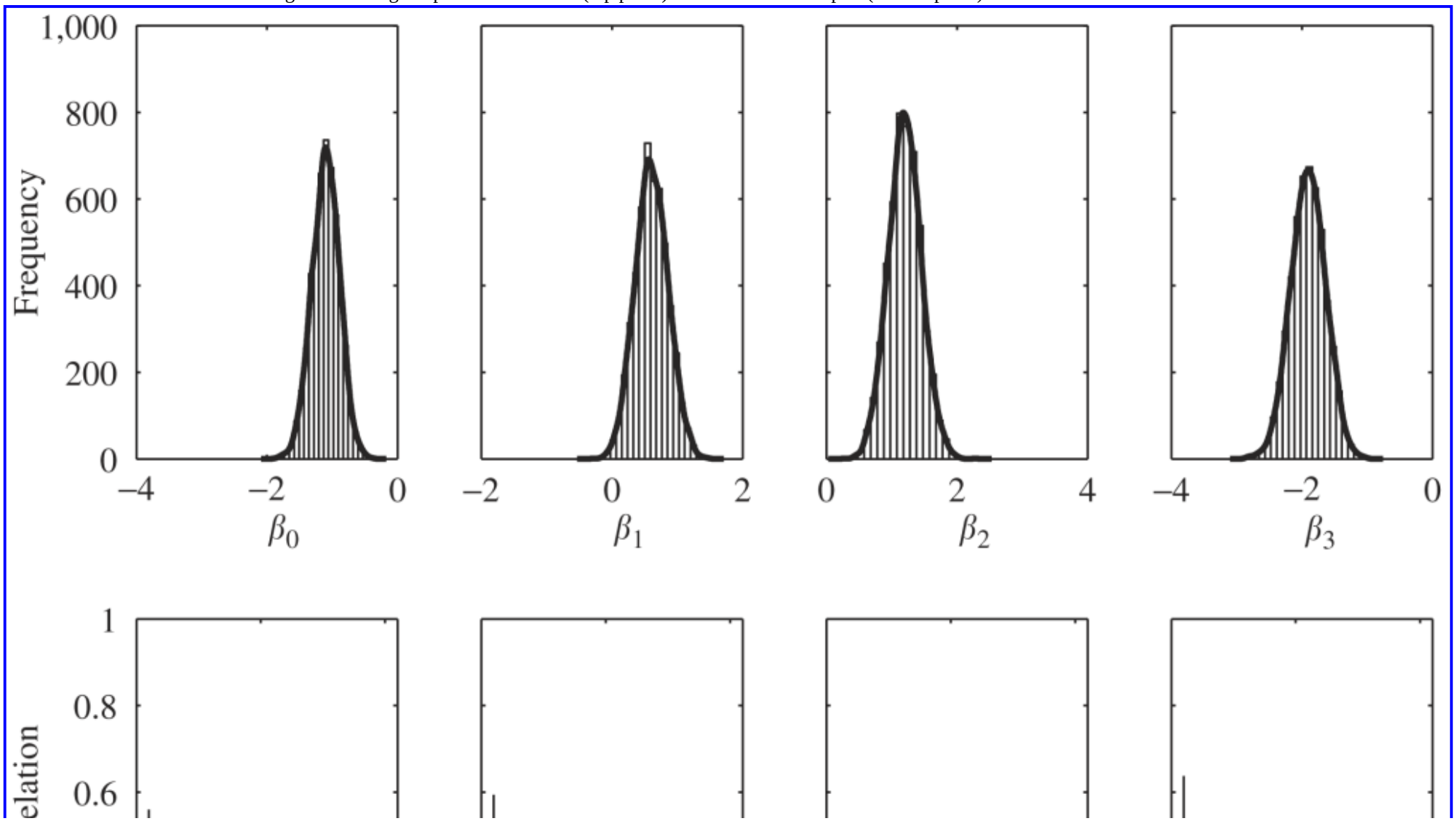

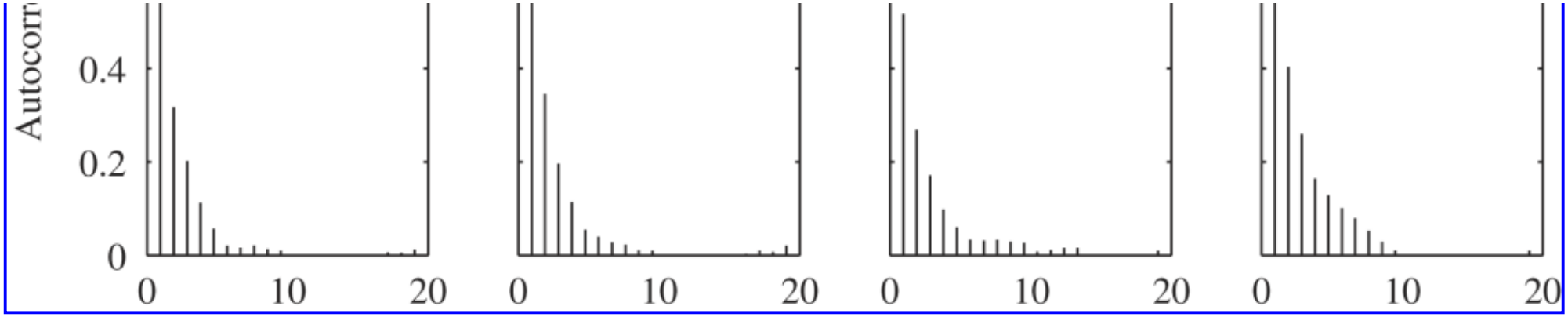

5 Strategies for improving mixing

In practice, while implementing MCMC methods it is important to construct samplers that mix well, where mixing is measured by the autocorrelation time, because such samplers can be expected to converge more quickly to the invariant distribution.

5.1 Choice of blocking

As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution.
Blocks can be combined by the method of composition. For example, suppose that $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ denote three blocks and that the distribution $\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_3$ is tractable (that is, can be sampled directly). Then, the blocks $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ can be collapsed by first sampling, $\boldsymbol{\theta}_1$ from $\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_3$ followed by $\boldsymbol{\theta}_2$ from $\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3$. This amounts to a two-block MCMC algorithm. In addition, if it is possible to sample $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ marginalized over θ_3 then the number of blocks is reduced to one. Liu (1994) discusses the value of these strategies in the context of a three-block Gibbs MCMC chain. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains used to sample multivariate normal target distributions.

5.2 Tuning the proposal density

The proposal density in an M–H algorithm has an important bearing on the mixing of the MCMC chain. Chib and Greenberg (1994; 1995), Tierney (1994), Tierney and Mira (1999) and Liu (2001) discuss various possibilities for formulating proposal density that can be helpful in a variety of problems.

6 Prediction and model choice

In some settings, for example in models for time series data, an important goal is prediction. In the Bayesian context, a future observation y_f is predicted through the (predictive) density defined as

$$f(y_f|\mathbf{y}) = \int f(y_f|\mathbf{y}, \boldsymbol{\theta})\,\pi(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta},$$

where $f(y_f|\mathbf{y}, \mathcal{M}, \boldsymbol{\theta})$ is the conditional density of y_f given $(\mathbf{y}, \boldsymbol{\theta})$. In general, the predictive density is not available in closed form. It can be shown, however, that, if one simulates

$y_f^{(j)} \sim f(y_f|\mathbf{y}, \boldsymbol{\theta}^{(j)})$ for each sampled draw $\boldsymbol{\theta}^{(j)}$ from the MCMC simulation, then the collection of simulated values $\left\{y_f^{(1)}, \ldots, y_f^{(M)}\right\}$ is a sample from $f(y_f|\mathbf{y})$. This simulated sample can be summarized in the usual way.

MCMC methods have also been widely applied to the problem of the model choice. Suppose that there are *K* possible models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ for the observed data defined by the sampling densities $\{f(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\}$ and proper prior densities $\{p(\boldsymbol{\theta}_k|\mathcal{M}_k)\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach this question is answered by placing prior probabilities $\Pr(\mathcal{M}_k)$ on each of the *K* models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathcal{M}_1|\mathbf{y}), \ldots, \Pr(\mathcal{M}_K|\mathbf{y})\}$ conditioned on the data but marginalized over the unknowns θ_k (Jeffreys, 1961). Specifically, the posterior probability of $\mathcal{M}_k$ is given by the expression

$$\Pr(\mathcal{M}_k|\mathbf{y}) = \frac{\Pr(\mathcal{M}_k)\, m(\mathbf{y}|\mathcal{M}_k)}{\sum_{l=1}^{K} \Pr(\mathcal{M}_l)\, m(\mathbf{y}|\mathcal{M}_l)} \propto \Pr(\mathcal{M}_k)\, m(\mathbf{y}|\mathcal{M}_k), \quad (k \leq K)$$

where $m(\mathbf{y}|\mathcal{M}_k)$ is the marginal likelihood of $\mathcal{M}_k$.

A problem in estimating the marginal likelihood is that it is an integral of the sampling density over the prior distribution of θ_k. Thus, MCMC methods, which deliver sample values from the posterior density, cannot be used to directly average the sampling density. One method for dealing with this difficulty is due to Chib (1995). The starting point is the expression

$$m(\mathbf{y}|\mathcal{M}_k) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{\pi(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k)}$$

which is an identity in θ_k. From here an estimate of the marginal likelihood on the log-scale is given by

$$\log \hat{m}(\mathbf{y}|\mathcal{M}_k) = \log f(\mathbf{y}|\boldsymbol{\theta}_k^*, \mathcal{M}_k) + \log p(\boldsymbol{\theta}_k^*|\mathcal{M}_k) - \log \hat{\pi}(\boldsymbol{\theta}_k^*|\mathbf{y}, \mathcal{M}_k)$$

where $\boldsymbol{\theta}_k^*$ denotes an arbitrarily chosen point and $\hat{\pi}(\boldsymbol{\theta}_k^*|\mathbf{y}, \mathcal{M}_k)$ is the estimate of the posterior density at that single point. To estimate the posterior ordinate one utilizes the Gibbs output in conjunction with a decomposition of the ordinate into marginal and conditional components. Chib and Jeliazkov (2001) extend this approach for output produced by the M–H algorithm while Basu and Chib (2003) show how the method can be applied in semiparametric models.

In some cases one is interested in a large number of candidate models, each with parameters $\boldsymbol{\theta}_k \in B_k \subseteq R^{d_k}$. In such cases one can get information about the relative support for the contending models from a model space-parameter space MCMC algorithm. In these algorithms, the models are represented by a categorical variable $\mathcal{M}$ which is then sampled along with the parameters of each model. The posterior distribution of $\mathcal{M}$ is computed as the frequency of times each model is visited. Methods for doing this have been proposed by Carlin and Chib (1995) and Green (1995). Both methods are closely related as shown by Dellaportas, Forster and Ntzoufras (2002) and Godsill (2001). Related methods for the problem of variable selection have also been developed starting with George and McCulloch (1993).

See Also

- Bayesian econometrics

- Bayesian statistics
- econometrics
- hierarchical Bayes models
- simulation-based estimation

Bibliography

Albert, J.H. and Chib, S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88, 669–79.

Basu, S. and Chib, S. 2003. Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association* 98, 224–35.

Carlin, B.P. and Chib, S. 1995. Bayesian model choice via Markov chain Monte Carlo. *Journal of Royal Statistical Society*, Series B 57, 473–84.

Chan, K.S. 1993. Asymptotic behavior of the Gibbs sampler. *Journal of the American Statistical Association* 88, 320–6.

Chen, M.H., Shao, Q.M. and Ibrahim, J.G. 2000. *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 1313–21.

Chib, S. 2001. Markov chain Monte Carlo methods: computation and inference. In *Handbook of Econometrics*, vol. 5, ed. J.J. Heckman and E. Leamer. Amsterdam: North-Holland.

Chib, S. and Greenberg, E. 1994. Bayes inference in regression models with ARMA (p,q) errors. *Journal of Econometrics* 64, 183–206.

Chib, S. and Greenberg, E. 1995. Understanding the Metropolis–Hastings algorithm. *American Statistician* 49, 327–35.

Chib, S. and Jeliazkov, I. 2001. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* 96, 270–81.

Dellaportas, P., Forster, J.J. and Ntzoufras, I. 2002. On Bayesian model and variable selection using MCMC. *Statistics and Computing* 12, 27–36.

Fahrmeir, L. and Tutz, G. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Berlin: Springer.

Gelfand, A.E. and Smith, A.F. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions, PAMI* 6, 721–41.

George, E.I. and McCulloch, R.E. 1993. Variable selection via Gibbs sampling. *Journal Of The American Statistical Association* 88, 881–9.

Gilks, W.K., Richardson, S. and Spiegelhalter, D.J. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Godsill, S.J. 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational And Graphical Statistics* 10, 230–48.

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–32.

Hastings, W.K. 1970. Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109.

Jeffreys, H. 1961. *Theory of Probability*, 3rd edn. Oxford: Oxford University Press.

Liu, J.S. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene-regulation problem. *Journal of the American Statistical Association* 89, 958–66.

Liu, J.S. 2001. *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. et al. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087–92.

Robert, C.P. and Casella, G. 2004. *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.

Roberts, G.O. and Sahu, S.K. 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B* 59, 291–317.

Roberts, G.O. and Smith, A.F.M. 1994. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and Their Applications* 49, 207–16.

Tanner, M.A. and Wong, W.H. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82, 528–50.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 21, 1701–62.

Tierney, L. and Mira, A. 1999. Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* 18, 2507–15.

How to cite this article

Chib, Siddhartha. "Markov chain Monte Carlo methods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000388> doi:10.1057/9780230226203.1039

The New Palgrave Dictionary of Economics Online

Markov equilibria in macroeconomics

Dirk Krueger and Felix Kubler

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We review the recent literature in macroeconomics that analyses Markov equilibria in dynamic general equilibrium model. After defining the Markov equilibrium concept we first summarize what is known about the existence and uniqueness of such equilibria in models where sequential equilibria can be obtained by solving a suitable social planner problem. We then discuss the existence problems of Markov equilibria in models where equivalence of equilibrium allocations and solutions to social planner problems cannot be established and review techniques the literature has developed to deal with the existence problem, as well as recent applications of these techniques in macroeconomics.

Keywords

backward induction; business cycles; commitment; convexity; dynamic programming; enforcement frictions; Euler equations; existence of a Markov equilibrium; fixed-point theorems; functional equations; generalized Markov equilibria; informational frictions; Lagrange multipliers; Markov chains; Markov equilibria in macroeconomics; Markov processes; neoclassical growth theory; optimal taxation; overlapping generations models; policy functions; principle of optimality; Recursive Markov equilibria; recursive techniques; reputation; sequential equilibria; social planner problem; state spaces; transition functions

Article

We say that a dynamic economy has a Markovian structure (or is Markovian, for short) if the stochastic processes that specify the fundamentals of the economy (such as endowments, preferences and technologies) are Markov processes. Note that deterministic economies are special cases in which the stochastic processes for the fundamentals have degenerate distributions. In many applications attention is restricted to first-order Markov processes in which the probability distributions over fundamentals today are functions exclusively of their values yesterday.

In dynamic economies sequential equilibria are sequences of functions mapping histories of realizations of the stochastic process of the fundamentals into allocations and prices such that all agents in the economy maximize their objectives, given prices, and all markets clear. Under fairly mild conditions (that is, convexity and continuity assumptions on the primitives) such equilibria exist. However, in order to characterize and compute equilibria it is often useful to look for equilibria of a different form.
Recursive Markov equilibria can be characterized by a state space, a policy function and a transition function. The policy function maps the state today into current endogenous choices and prices, and the transition function maps the state today into a probability distribution over states tomorrow (see, for example, the definition in Ljungqvist and Sargent's 2000 textbook). In most of this survey we will use the terms ‘Markov equilibria’ and ‘recursive Markov equilibria’ interchangeably; however, below we also consider Markov equilibria which are not recursive and refer to these as ‘generalized Markov equilibria’. This characterization leaves open, of course, what the appropriate state variables are that constitute the state space.
Most simply, the state space would consist of the set of possible exogenous shocks governing endowments, preferences and technologies. But, other than in exceptional cases (see, for example, Lucas's 1978 asset pricing application where asset prices are solely functions of the underlying shocks to technology), such a strongly stationary Markov equilibrium does not exist.
In addition to the exogenous shocks, endogenous variables have to be included in the state space to assure existence of a Markov equilibrium. We define as the minimal state space the space of all exogenous shocks and endogenous variables that are payoff-relevant today, in that they affect current production or consumption sets or preferences (see Maskin and Tirole, 2001).
We call Markov equilibria with this minimal state space ‘simple Markov equilibria’. In the remainder of this article we want to discuss what we know about the existence and uniqueness of such Markov equilibria, both in general and for important specific examples. As it turns out, when equilibria are Pareto efficient, and thus equilibrium allocation can be determined by solving a suitable social planner problem, simple Markov equilibria can be shown to exist under fairly mild conditions. We therefore discuss this case first. On the other hand, when equilibria are not Pareto efficient – for example, when markets are incomplete or economic agents behave strategically – forward-looking variables often have to be included for a Markov equilibrium to exist; therefore, simple Markov equilibria in the sense defined above do not exist in general. We discuss this case in Section 2.

1 Markov equilibria in economies where equilibria are Pareto optimal

In this section we discuss the existence and uniqueness of simple Markov equilibria in economies whose sequential market equilibrium allocations can be determined as solutions to a suitable social planner problem. In these economies the problem of proving the existence of a Markov equilibrium reduces to showing that the solution of the social planner can be written as a time-invariant optimal policy function of the minimal set of state variables, as defined above.
This is commonly done by reformulating the optimization problem of the social planner as a functional equation and showing that the optimal Markov policy function generates a sequential allocation which solves the original social planner problem; this is what Bellman (1957) called the principle of optimality. This principle can be established under weak conditions (see Stokey, Lucas and Prescott, 1989).

Equipped with this result, the existence of a Markov equilibrium then follows from the existence of a solution to the functional equation associated with the social planner problem.

If the functional equation can be shown to be a contraction mapping (sufficient conditions for this were provided by Blackwell, 1965), then it follows that there exists a unique value function solving the functional equation and an optimal policy correspondence. In addition, the contraction mapping theorem also gives an iterative procedure to find the solution to the functional equation from any starting guess, which is helpful for numerical work.

Under weaker conditions other fixed-point theorems may be employed to argue at least for the existence (if not uniqueness) of a solution to the functional equation, with associated optimal Markov policy correspondence. In order to establish that the policy correspondence is actually a function (and thus the Markov equilibrium is unique), in general strict concavity of the return function needs to be assumed. Stokey, Lucas and Prescott (1989) provide a summary of the main results in the general theory of dynamic programming.

This technique of analysing and computing dynamic equilibria in Pareto optimal economies is now widely used in macroeconomics. Its first application can be found in Lucas and Prescott (1971) in their study of optimal investment behaviour under uncertainty. Lucas (1978) used recursive techniques to study asset prices in an endowment economy and showed that the Markov equilibrium has a particularly simple form. Kydland and Prescott (1982) showed how powerful these techniques are for a quantitative study of the business cycle implications of the neoclassical growth model with technology shocks to production. The volume by Cooley (1995) provides a comprehensive overview over this line of research.

2 Generalized Markov equilibria

In models where the first welfare theorem is not applicable (for example, models with incomplete financial markets or with distorting taxes), in models where there are infinitely many agents (such as overlapping generations models) or in models with strategic interaction the existence of simple Markov equilibria (that is, Markov equilibria with minimal state space) cannot be guaranteed. See Santos, 2002; Krebs, 2004; Kubler and Schmedders, 2002; and Kubler and Polemarchakis, 2004, for simple counter-examples. An important exception is Bewley-style models with incomplete markets where simple recursive Markov equilibria exist; see, for example, Krebs, 2006. The functional equations characterizing equilibrium have no contraction properties, and more general fixed-point theorems than the contraction mapping theorem, such as Schauder's fixed-point theorem, cannot be applied because it is difficult to guarantee compactness of the space of admissible functions. Coleman (1991) is an important example where existence can be shown. However, his results rely on monotonicity conditions on the equilibrium dynamics which are not satisfied in general models.

In the applied literature a solution to this problem was suggested early on. For example Kydland and Prescott (1980) analyse a Ramsey dynamic optimal taxation problem. To make the problem recursive they add as a state variable last period's marginal utility.

On the theoretical side Duffie et al. (1994) were the first to rigorously analyse situations where recursive equilibria may fail to exist in general equilibrium models. Kubler and Schmedders (2003) and Miao and Santos (2005) refine their approach and make it applicable for computations. Miao and Santos (2005),

also give a clear explanation of how this approach relates to the work by Abreu, Pearce and Stacchetti (1990). We now present their basic idea.

Consider a Markovian economy where a date-event (or node) can be associated with a finite history of shocks, $S^t = (s_0, \ldots, s_t)$. The shocks follow a Markov chain with support $S = \{1, \ldots, S\}$. Denote by $z(s^t)$ the vector of all endogenous variables at node s^t. Typically this would include the vector of household asset holdings across individuals and the capital stock at the beginning of the period, but also prices and endogenous choices at note s^t, as well as shadow variables such as Lagrange multipliers. A competitive equilibrium is a process of endogenous variables $\{z(s^t)\}$ with $z(s^t) \in \mathcal{Z} \subset \mathbb{R}^M$, which solve the optimization problems of all agents in the economy, and clear markets. The set $\mathcal{Z}$ denotes the set of all possible values of the endogenous variables.

We focus on dynamic economic models where an equilibrium can be characterized by a set of equations relating current-period exogenous and endogenous variables to endogenous and exogenous variables next period. It is straightforward to incorporate inequality constraints into this framework. For expositional purposes we focus on equations. Examples of such equations are the Euler equations of individual households, first order conditions of firms, as well as market-clearing conditions for all markets. We assume that such a set of equations characterizing equilibrium is given and denote it by

$$h(\hat{s}, \hat{z}, z_1, \ldots, z_S) = 0.$$

The arguments $(\hat{s}, \hat{z})$ denote the exogenous state variables and endogenous variables for the current period. Note that the endogenous variables might contain variables which were determined in the previous period, such as the capital stock and individuals' assets. The variables $(z_s)_{s=1}^{S}$ denote endogenous variables in the subsequent period, in states $s = 1, \ldots, S$, respectively. We refer to $h(\cdot) = 0$ as the set of 'equilibrium equations'.

As explained above, to analyse Markov equilibria one needs to specify an appropriate state space. We assume that the equilibrium set $\mathcal{Z}$ can be written as the product $\mathcal{Y} \times \hat{\mathcal{Z}}$, where $\mathcal{Y}$ denotes the set into which the endogenous state variables fall. In the neoclassical growth model, $\mathcal{Y}$ would consist of the set of possible values of the capital stock; in models with heterogeneous agents one would need to add the set of possible wealth distributions across agents. Unfortunately, as the references cited above show, a recursive Markov equilibrium with this state space may not exist. We therefore require a more general notion of Markov equilibrium for these types of economies.

A *generalized Markov equilibrium* consists of a (non-empty valued) 'policy correspondence', *P*, that maps the state today into possible endogenous variables today, and a 'transition function', *F*, that maps the state and endogenous variables today into endogenous variables next period. Formally, the maps

$$P: \mathcal{S} \times \mathcal{Y} \rightrightarrows \hat{\mathcal{Z}} \text{ and } F: \text{graph}(P) \to \mathcal{Z}^S$$

should satisfy that for all shocks and endogenous variables in the current period, $(\hat{s}, \hat{z}) \in \text{graph}(P)$, the transition function prescribes values next period that are consistent with the equilibrium equations, that is,

$$h(\hat{s}, \hat{z}, F(\hat{s}, \hat{z})) = 0,$$

and lie in the policy correspondence, that is,

$$(s, F_S(\hat{s}, \hat{z})) \in \text{graph}(P) \text{ for all } s \in S.$$

It follows that a generalized Markov equilibrium is recursive, according to our earlier definition, if the associated policy correspondence is single valued. It is simple if the state space is the natural minimal state space.

It is easy to see that Markov equilibria are in fact competitive equilibria in the usual sense. Duffie et al. (1994) show that, under mild assumptions on the primitives of the model, generalized Markov equilibria exist whenever competitive equilibria exist. The basic idea of their approach is very similar to backward induction, using critically a natural monotonicity property of the inverse of the equilibrium equations. (See their original paper, Kubler and Schmedders (2003), or Miao and Santos (2005), for details.)

For practical purposes it is of course crucial that the chosen state space is relatively small and that the Markov equilibrium is recursive. In an asset pricing model with heterogeneous agents, Kubler and Schmedders choose the state space to consist of the beginning-of-period wealth distribution, but can show the existence only of a generalized Markov equilibrium. One cannot rule out the possibility that the equilibrium is not recursive; the same value of the state variables might occur with different values of the endogenous variables. The counter-examples to existence mentioned above show that this is precisely the problem. If for given initial conditions there exist multiple competitive equilibria, the one that realizes is pinned down by lagged variables. Without ruling out multiplicity of equilibria, it does not seem possible to prove the existence of recursive equilibria with the natural state space.

Miao and Santos (2005) enlarge the state space with the shadow values of investment of all agents and prove that with this larger state space a recursive Markov equilibrium exists. The basic insight of their approach is that one needs to add variables to the natural state space that uniquely select one out of several possible endogenous variables.

The main practical problem with the approach originated by Duffie et al. (1994) and refined by Miao and Santos (2005) is that it provides a method to construct all Markov equilibria. There might exist some recursive equilibria for the natural (minimal) state space, but this approach naturally solves for all other recursive Markov equilibria as well. Datta et al., 2005, provide ideas for solving for the one Markov

equilibrium with minimal state space.

In many recent applications of recursive methods to macroeconomics the focus of researchers studying non-optimal economies is to find a recursive equilibrium with minimal state space. Notable examples in which even this natural state space is large are Rios-Rull (1996), Heaton and Lucas (1996) and Krusell and Smith (1998). They mark the boundary of economies that currently can be analysed with recursive techniques.

In dynamic endowment economies with either informational frictions or limited enforceability of contracts, constrained-efficient (efficient, subject to the informational or enforcement constraints) consumption allocations usually display a high degree of dependence on past endowment shocks, even though the natural state space contains only the current endowment shock. Therefore, Markov equilibria with minimal state space do not exist. However, using ideas by Spear and Srivastava (1987) and Abreu, Pearce, and Stacchetti (1990), the papers by Atkeson and Lucas (1992) and Thomas and Worrall (1988) demonstrate that nevertheless the constrained social planner problem has a convenient recursive structure if one includes promised lifetime utility as a state variable into the recursive problem. This approach or its close alternative, namely, to introduce as an additional state variable Lagrange multipliers on the incentive or enforcement constraints (as in Marcet and Marimon, 1998), has seen many applications in macroeconomics, since it facilitates making a large class of dynamic models with informational or enforcement frictions recursive and hence tractable. Miao and Santos (2005) show how such problems with strategic interactions can be incorporated into the framework above.

In optimal policy problems in which the government has no access to a commitment technology, a discussion has emerged about the desirability of a restriction to Markov policies with minimal state space. Such restrictions rule out reputation if one confines attention to smooth policies. See Phelan and Stacchetti (2001) and Klein and Rios-Rull (2003) for examples of the two opposing views on this issue. However, as Krusell and Smith (2003) argue, if one allows discontinuous policy functions reputation effects can be generated even with Markov policies. (While Krusell and Smith discuss optimal decision rules in a consumption–savings problem with quasi-geometric discounting, their results carry over to optimal policy problems without commitment on the part of the policymaker.)

See Also

- computation of general equilibria
- decentralization
- Euler equations
- existence of general equilibrium
- functional analysis
- general equilibrium
- general equilibrium (new developments)
- income taxation and optimal policies
- incomplete markets
- Markov processes

- optimal fiscal and monetary policy (without commitment)
- Pareto efficiency
- recursive competitive equilibrium
- recursive contracts

Bibliography

Abreu, D., Pearce, D. and Stacchetti, E. 1990. Toward a theory of repeated games with discounting. *Econometrica* 58, 1041–63.

Atkeson, A. and Lucas, R. 1992. On efficient distribution with private information, *Review of Economic Studies* 59, 427–53.

Bellman, R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Blackwell, D. 1965. Discounted dynamic programming. *Annals of Mathematical Statistics* 36, 226–35.

Coleman, J. 1991. Equilibrium in a production economy with an income tax. *Econometrica* 59, 1091–104.

Cooley, T. 1995. *Frontiers of Business Cycle Research*. Princeton, NJ: Princeton University Press.

Datta, M., Mirman, L., Morand, O. and Reffett, K. 2005. Markovian equilibrium in infinite horizon economies with incomplete markets and public policy. *Journal of Mathematical Economics* 41, 505–44.

Duffie, D., Geanakoplos, J., Mas-Colell, A. and McLennan, A. 1994. Stationary Markov equilibria. *Econometrica* 62, 745–81.

Heaton, H. and Lucas, D. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104, 443–87.

Klein, P. and Rios-Rull, V. 2003. Time consistent optimal fiscal policy. *International Economic Review* 44, 1217–46.

Krebs, T. 2004. Non-existence of recursive equilibria on compact state spaces when markets are incomplete. *Journal of Economic Theory* 115, 134–50.

Krebs, T. 2006. Recursive equilibrium in endogenous growth models with incomplete markets. *Economic Theory* 29, 505–23.

Krusell, P. and Smith, A. 1998. Income and wealth heterogeneity in the macroeconomy. *Journal of*

Political Economy 106, 867–96.

Krusell, P. and Smith, A. 2003. Consumption–savings decisions with quasi–geometric discounting. *Econometrica* 71, 365–76.

Kubler, F. and Polemarchakis, H. 2004. Stationary Markov equilibria for overlapping generations. *Economic Theory* 24, 623–43.

Kubler, F. and Schmedders, K. 2002. Recursive equilibria in economies with incomplete markets. *Macroeconomic Dynamics* 6, 284–306.

Kubler, F. and Schmedders, K. 2003. Stationary equilibria in asset-pricing models with incomplete markets and collateral. *Econometrica* 71, 1767–95.

Kydland, F. and Prescott, E. 1980. Dynamic optimal taxation, rational expectations and optimal control. *Journal of Economic Dynamics and Control* 2, 79–91.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations, *Econometrica* 50, 1345–71.

Ljungqvist, L. and Sargent, T. 2000. *Recursive Macroeconomic Theory*. Cambridge, MA: MIT Press.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1426–45.

Lucas, R. and Prescott, E. 1971. Investment under uncertainty. *Econometrica* 39, 659–81.

Marcet, A. and Marimon, R. 1998. Recursive contracts. Working paper, University Pompeu Fabra, Barcelona.

Maskin, E. and Tirole, J. 2001. Markov perfect equilibrium. *Journal of Economic Theory* 100, 191–219.

Miao, J. and Santos, M. 2005. Existence and computation of Markov equilibria for dynamic non–optimal economies. Working paper, Department of Economics, Boston University.

Phelan, C. and Stacchetti, E. 2001. Sequential equilibria in a Ramsey tax model. *Econometrica* 69, 1491–518.

Rios-Rull, V. 1996. Life cycle economies with aggregate fluctuations. *Review of Economic Studies* 63, 465–90.

Santos, M. 2002. On non-existence of Markov equilibria for competitive-market economies. *Journal of Economic Theory* 105, 73–98.

Spear, S. and Srivastava, S. 1987. On repeated moral hazard with discounting, *Review of Economic Studies* 54, 599–617.

Stokey, N., Lucas, R. and Prescott, E. 1989. *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

Thomas, J. and Worrall, T. 1988. Self-enforcing wage contracts. *Review of Economic Studies* 55, 541–54.

How to cite this article

Krueger, Dirk and Felix Kubler. "Markov equilibria in macroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000377> doi:10.1057/9780230226203.1040

The New Palgrave Dictionary of Economics Online

Markov processes

Daniel W. Stroock
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In this article the theory of Markov processes is described as an evolution on the space of probability measures. Following a brief historical account of its origins in physics, a mathematical formulation of the theory is given. Emphasis has been placed on the ergodic properties of Markov processes, and their presence is checked in a simple example.

Keywords

Boltzmann, L.; Brownian motion; ergodic theory; Gibbs, G.; Markov processes; Markov property; Newton's equation; probability; statistical mechanics; Wiener process

Article

Unless one is clairvoyant, the only temporally evolving processes which are tractable are those whose future behaviour can be predicted on the basis of data which is available at the time when the prediction is being made. Of course, in general, the behaviour of even such an evolution will be impossible to predict. For example, if, in order to make a prediction, one has to know the detailed history of everything that has happened during the entire history of the entire universe, one's chance of making a prediction may be a practical, if not a theoretical, impossibility. For this reason, one tries to study evolutions mathematically with models in which most of the distant past can be ignored when one makes predictions about the future. In fact, many mathematical models of evolutions have the property that, for the purpose of predicting the future, the past becomes irrelevant as soon as one knows the present, in which case the evolution is said to be a ‘Markov process’, the topic at hand, after Andrei Andreyevich Markov (1856–1922).

The components of a Markov process are its *state space* $\mathbb{S}$ and its *transition rule T*. Mathematically, $\mathbb{S}$ is just some non-empty set, which in applications will encode all the possible states in which the evolving system can find itself, and $T:\mathbb{S}\to\mathbb{S}$ is a function from $\mathbb{S}$ into itself which gives the *transition rule*. More precisely, if now the system is in state *x*, it will be next in state *T*(*x*), from which it will go to $T^2(x) = T(T(x))$, and so on. (Here we are thinking of time being discrete. Thus, ‘next’ means after one unit of time has passed.)

To give a sense of the sort of reasoning required to construct a Markov process, consider a (classical) physical particle whose motion is governed by Newton's equation $\vec{F} = m\vec{a}$ (‘force equals mass times acceleration’). At least in theory, Newton's equation says that, on the assumption that one knows the mass of the particle and the force field $\vec{F}$ which acts on it, one can predict where the particle will be in the future as soon as one knows what its position and velocity are now. On the other hand, knowing only its present position is not sufficient by itself. Thus, even though one may care about nothing but its position, in order to produce a Markov process for a particle evolving according to Newton's equation it is necessary to adopt the attitude that the *state* of the particle consists of its position *and* velocity, not just its position alone. Of course, in that velocity is the derivative of position, the two are so inextricably intertwined that one might be tempted to concentrate on position on the grounds that one will be able to compute the velocity whenever necessary. However, this tack destroys the Markov property,

namely, there is no way of computing the velocity of a particle 'now' if all one knows is its position 'now'. For this reason, physicists consider the state of a particle to be a composite of its position and velocity, and the resulting state space $\mathbb{R}^6 = \mathbb{R}^3 \times \mathbb{R}^3$ (three coordinates for position and three for velocity) they call the *phase space* of the particle.

The same point may be clearer in the following example. Suppose that one has an evolution on a state space $\mathbb{S}$ which proceeds according to the rule that, if the present state is x_n and the preceding state was x_{n-1}, then the next state will be $x_{n+1} = T(x_{n-1}, x_n)$. This is *not* a Markov process. Nonetheless, it can be 'Markovized'. Indeed, replace the original state space by $\hat{\mathbb{S}} = \mathbb{S} \times \mathbb{S}$, the set of ordered pairs (x, y) with x and y from $\mathbb{S}$, and define $\hat{T}((x, y)) = (y, T(x, y))$. It is then an easy matter to check that, if the original system was in state x_{-1} at time –1 and state x_0 at time 0, then its state at time $n \geq 1$ will be x_n, the second component of the pair $(x_{n-1}, x_n) = \hat{T}^n((x_{-1}, x_0))$.

The moral to be drawn from these examples is that *the presence or absence of the Markov property is in the eye of the beholder*. That is, a change of venue (the state space) can make the Markov property appear in circumstances where it was not originally apparent. In fact, by making the state space sufficiently large, any evolution can be forced to be Markov. On the other hand, the more complicated the state space, the less useful is the Markov property. Thus, in practice, what one seeks is the 'simplest' state space on which one's evolution possesses the Markov property.

Stochastic Markov processes

Roughly speaking, Markov processes fall into one of two categories. Those in the first category are 'deterministic' in the sense that their state space is sufficiently detailed that the individual states give complete and unambiguous information. Both the examples given above are deterministic. The mathematical analysis of deterministic Markov processes has a proud history going back to Newton which includes major contributions by such luminaries as P. Chebyshev, A. Markov, A. Lyapounov, H. Poincaré, and J. Moser. The second category of Markov processes, and the one on which the rest of this article will concentrate, are 'probabilistic' or 'stochastic' Markov processes. To understand where and why these processes arise, consider the problem of describing the state of all the gas molecules in a room. Each litre of gas contains approximately Avogrado's number, $6.02214199 \times 10^{23}$, of molecules. Thus, even a small room will contain something on the order of 10^{26} molecules. Moreover, because, by Newton's laws of motion, the state of each individual molecule will lie in its individual phase space, the state of the entire system of molecules will have to specify the positions and velocities of all 10^{26} molecules. Stated mathematically, the state space of the system will be $\mathbb{R}^{6 \times 10^{26}}$, on which any sort serious analysis is too daunting to contemplate.

When one is confronted with a problem which is intractable as presented, the time-honoured procedure of choice is to reformulate the problem in a way which makes it more tractable. In the case just described, the reformulation was made by G. W. Gibbs (1902) and L. Boltzmann (1896; 1898), the fathers of statistical mechanics. They abandoned any hope of saying exactly where all the molecules will be and reconciled themselves to settling for a description of the statistics of the molecules. That is, instead of asking exactly where all the molecules would be, they asked what would be the probability of finding a molecule in various regions of phase space. From this point of view, the state of the system will not be an element of $\mathbb{R}^{6 \times 10^{26}}$ but of $\mathbf{M_1}(\mathbb{R}^6)$, the space probability distributions on the individual phase space $\mathbb{R}^6$. Of course, Gibbs and Boltzmann's reformulation only changes the problem, it does not solve it. Indeed, although Newton's equation determines how the system of molecules evolves and therefore how their distribution will evolve, the use of Newton's equation would remove the advantage which Boltzmann and Gibbs hoped to gain from their reformulation. Thus, they had to come up with an alternative way of describing the transition rule which governs the evolution of the distribution of the system as a Markov process on $\mathbf{M_1}(\mathbb{R}^6)$. The description proposed by Boltzmann is given by the famous Boltzmann equation. Unfortunately, Boltzmann's equation is itself so complicated that it is only recently that substantial progress has been made toward understanding it in any generality. On the other hand, Gibbs and Boltzmann's idea of studying Markov processes on the space of probability distributions is seminal and has proved to be both ubiquitous and powerful.

The abstract setting for a stochastic Markov process starts with a non-empty set $\mathbb{S}$, the deterministic state space, and the associated space $\mathbf{M_1}(\mathbb{S})$ of probability distributions on $\mathbb{S}$. The easiest and most commonly studied stochastic Markov processes are those for which the transition rule $T: \mathbf{M_1}(\mathbb{S}) \to \mathbf{M_1}(\mathbb{S})$ is a linear (more correctly, an affine) function. To be definite, suppose $\mathbb{S}$ is a finite set. Then $\mathbf{M_1}(\mathbb{S})$ is the set of all functions μ on $\mathbb{S}$ which assign each $x \in \mathbb{S}$ a number $\mu(\{x\}) \in [0, 1]$ (the

probability of $\{x\}$ under μ) in such a way that $\sum_{x\in S}\mu(\{x\}) = 1$. (The use of $\mu(\{x\})$ instead of $\mu(x)$ here is a little pedantic. However, one must remember that probabilities are assigned to *events* – that is, subsets of the sample space – and that $\{x\}$ is the event that '*x* occurred'.) Clearly, if μ and ν are in $\mathbf{M_1}(S)$ and $\theta\in[0, 1]$, then the convex combination $\theta\mu + (1-\theta)\nu$ is again an element of $\mathbf{M_1}(S)$. Sets with this property are said to be 'affine' (as distinguished from 'linear', which refers to sets which are closed under all linear, not just convex, combinations), and a function on an affine set is said to be affine if it commutes with convex combinations. Thus, for $\mathbf{M_1}(S)$, the transition rule *T* is affine if $T(\theta\mu + (1-\theta)\nu) = \theta T(\mu) + (1-\theta)T(\nu)$. Because S is finite, one can dissect such transition rules in the following way. First, for each $x\in S$, let δ_x denote the element of $\mathbf{M_1}(S)$ which assigns 1 to $\{x\}$ (and therefore 0 to $S\setminus\{x\}$). Next, set $\mathbf{P}(x, \cdot) = T(\delta_x)$. That is, $\mathbf{P}(x, \cdot)$ is the element of $\mathbf{M_1}(S)$ to which *T* takes $T(\delta_x)$, and so $\mathbf{P}(x, \{y\}) = [T(\delta_x)](\{y\})$. Because, for any $\mu\in\mathbf{M_1}(S)$ which is not equal to δ_x, $\mu = \mu(\{x\})\delta_x + (1-\mu(\{x\}))\mu^x$, where $\mu^x\in\mathbf{M_1}(S)$ is determined so that $\mu^x(\{y\})$ equals $(1-\mu(\{x\}))^{-1}\mu(\{y\})$ or 0 depending on whether $y\neq x$ or $y = x$, the affine property of *T* means that $T(\mu) = \mu(\{x\})\mathbf{P}(x, \cdot) + (1-\mu(\{x\}))T(\mu^x)$. Hence, after peeling off one *x* at a time, one concludes that

$$T(\mu) = \sum_{x\in S}\mu(\{x\})\mathbf{P}(x, \cdot) \tag{1}$$

when *T* is affine.

Probabilistic interpretation

The representation of *T* given by (1) admits an intuitively pleasing probabilistic interpretation: namely, $\mathbf{P}(x, \{y\})$ can be thought of as the probability that the system will next be in the state *y* given that is now in state *x*. With this interpretation in mind, probabilists call $x\in S\mapsto\mathbf{P}(x, \cdot)\in\mathbf{M_1}(S)$ a *transition probability* on the state space S. The terminology here is confusing. From the point of view adopted earlier, one might, and should, have thought that $\mathbf{M_1}(S)$ is the state space. However, the probabilistic interpretation is most easily appreciated if one thinks of S as the state space and $x\in S\to(x, \cdot)\in\mathbf{M_1}(S)$ as a random transition rule. To complete this picture, probabilists introduce random variables to represent the random points in S visited. More precisely, again assume that S is finite, and suppose that $\mu\in\mathbf{M_1}(S)$ describes the initial distribution of the process under consideration. Then probabilists construct a sequence $\{X_n: n\geq 0\}$ of random variables, called a *Markov chain*, in such a way that, for any $n\geq 0$,

$$\mathbf{P}(X_0 = x_0, \ldots, X_n = x_n) = \mu(\{x_0\})\mathbf{P}(x_0, \{x_1\})\cdots\mathbf{P}(x_{n-1}, \{x_n\}).$$

In words, this says that the right-hand side above is the probability that the chain with initial distribution μ starts at x_0 and then goes on to visit, successively, the points x_1 through x_n.
To see that the probabilistic interpretation is completely consistent in the deterministic case, observe that a deterministic Markov process can be formulated as a stochastic Markov process. That is, if *T* is the transition rule for the deterministic process, take $\mathbf{P}(x, \cdot) = \delta_{T(x)}$, and check that, with probability 1, the Markov chain with transition probability $\mathbf{P}(x, \cdot)$ follows the same path as the deterministic one with transition rule *T*. Equivalently, with probability 1, $X_n = T^n(X_0)$ for all $n\geq 1$.

Ergodic theory of Markov chains

Continue in the setting of the preceding section. One of the phenomena predicted by Gibbs in connection with his and

Boltzmann's study of gases was that, no matter what the initial distribution of the gas, after a long time the gas should equilibrate in the sense that it will achieve a *stationary distribution* (that is, a distribution that does not change with time) which does not depend on how it was distributed initially. One's experience with the behaviour of gases makes this prediction entirely plausible: place an opened bottle of perfume in the corner of a room; wait an hour, and confirm that the perfume will have become more or less equi-distributed throughout the room. Be that as it may, the prediction, which goes by the name of Gibbs's ‘ergodic hypothesis’, has been mathematically verified in only one physically realistic model. Nonetheless, as will be explained next, ergodicity is relatively easy to verify for most stochastic Markov processes on a finite state space.

To develop some intuition for what ergodicity means and why it might hold for a stochastic Markov process on a finite state space $\mathbf{s}$, it is best to first know how to recognize when a $\mu \in \mathbf{M_1}(\mathbf{s})$ is stationary. But, if μ is stationary, then it is left unchanged as the system evolves, and, in terms of the transition probability, this means that

$$\mu(\{y\}) = \sum_{x \in \mathbf{s}} \mu(\{x\})\mathbf{P}(x, \{y\}) \text{for all } y \in \mathbf{s}. \tag{2}$$

Now suppose that $\mathbf{s} = \{1, 2\}$, and consider the problem of finding a solution to (2). That is, we want to find $\mu \in \mathbf{M_1}(\{1, 2\})$ so that

$$\mu(\{1\}) = \mu(\{1\})\mathbf{P}(1, \{1\}) + \mu(\{2\})\mathbf{P}(2, \{1\})\mu(\{2\}) = \mu(\{1\})\mathbf{P}(1, \{2\}) + \mu(\{2\})\mathbf{P}(2, \{2\}). \tag{3}$$

At first sight, there appear to be too many conditions on μ: not only must it satisfy the two equations in (3), it also has to satisfy $\mu(\{1\}) + \mu(\{2\}) = 1$ as well as being non-negative. Even if one ignores the non-negativity, one suspects that three linear equations are just too many for a pair of numbers to satisfy. On the other hand, after a little manipulation, one sees (remember that $\mathbf{P}(1, \cdot)$ and $\mathbf{P}(2, \cdot)$ are probability distributions) that both the equations in (3) are equivalent to $\mu(\{1\})\mathbf{P}(1, \{2\}) = \mu(\{2\})\mathbf{P}(2, \{1\})$. Hence the two equations in (3) are equivalent, and so there are really only two equations to be satisfied: $\mu(\{1\})\mathbf{P}(1, \{2\}) = \mu(\{2\})\mathbf{P}(2, \{1\})$ and $\mu(\{1\}) + \mu(\{2\}) = 1$. There are two cases to be considered. The first case is when the chain never moves, or, equivalently, $\mathbf{P}(1, \{2\}) = 0 = \mathbf{P}(2, \{1\})$. In this case there are two solutions, namely, δ_1 and δ_2, which is exactly what one should expect for a chain which never moves. In the second case, the one corresponding to a chain which can move, either $\mathbf{P}(1, \{2\}) > 0$ or $\mathbf{P}(2, \{1\}) > 0$. In both these cases, one can easily check that the one and only solution to (3) is given by

$$\mu(\{1\}) = \frac{\mathbf{P}(2, \{1\})}{\mathbf{P}(1, \{2\}) + \mathbf{P}(2, \{1\})} \text{ and } \mu(\{2\}) = \frac{\mathbf{P}(1, \{2\})}{\mathbf{P}(1, \{2\}) + \mathbf{P}(2, \{1\})}.$$

Continuing in the setting of the preceding, we want to examine when Gibbs's ergodic hypothesis holds. Obviously, at the very least, ergodicity requires that there be only one stationary μ, otherwise we could start the chain with one of them as initial distribution, in which case it would never get to the other. Thus, we need to assume that $\mathbf{P}(1, \{2\}) + \mathbf{P}(2, \{1\}) > 0$, and, to simplify matters, we will assume more, namely, that $m \equiv m_1 + m_2 > 0$, where $m_1 = \min\{P(1, \{1\}), P(2, \{1\})\}$ and $m_2 = \min\{P(1, \{2\}), P(2, \{2\})\}$, and, under this assumption we (following Doeblin) will show that, for any $\nu \in \mathbf{M_1}(\{1, 2\})$

$$\|\nu\mathbf{P} - \mu\| \leq (1 - m)\|\nu - \mu\|, \tag{4}$$

where $\nu\mathbf{P} \in \mathbf{M_1}(\{1, 2\})$ is determined by

$$\nu\mathbf{P}(\{y\}) \equiv \sum_{x=1}^{2} \nu(\{x\})\mathbf{P}(x, \{y\})$$

and, for any pair $\nu_1, \nu_2 \in \mathbf{M_1}(\{1, 2\})$, $\|\nu_2 - \nu_1\| \equiv \Sigma_{x=1}^{2}|\nu_2(\{x\}) - \nu_1(\{x\})|$. To prove (4), first observe that, because μ is stationary, $\mu = \mu\mathbf{P}$, and therefore, since $\Sigma_{x=1}^{2}(\nu(\{x\}) - \mu(\{x\})) = 1 - 1 = 0$,

$$\nu\mathbf{P}(\{y\}) - \mu(\{y\}) = \sum_{x=1}^{2}(\nu(\{x\}) - \mu(\{x\}))\mathbf{P}(x, \{y\}) = \sum_{x=1}^{2}(\nu(\{x\}) - \mu(\{x\}))(\mathbf{P}(x, \{y\}) - m_y).$$

Next, take the absolute value of both sides, remember that the absolute value of a sum of numbers is dominated by the sum of their absolute values, and arrive at

$$\|\nu\mathbf{P} - \mu\| \leq \sum_{y=1}^{2}\left(\sum_{x=1}^{2}|\nu(\{x\}) - \mu(\{x\})|(\mathbf{P}(x, \{y\}) - m_y)\right) = \sum_{x=1}^{2}\left(\sum_{y=1}^{2}|\nu(\{x\}) - \mu(\{x\})|(\mathbf{P}(x, \{y\}) - m_y)\right) = (1 - m)\|\nu - \mu\|.$$

Given (4), it becomes an easy matter to check ergodicity. Indeed, $\nu\mathbf{P}$ is the distribution of the chain at time 1 when it is started with initial distribution ν. Similarly, its distribution at time 2 will be $\nu\mathbf{P}^2 = (\nu\mathbf{P})\mathbf{P}$, and so $\left\|\nu\mathbf{P}^2 - \mu\right\| \leq (1 - m)\|\nu\mathbf{P} - \mu\| \leq (1 - m)^2\|\nu - \mu\|$. Proceeding by induction, one sees that distribution $\nu\mathbf{P}^n = (\nu\mathbf{P}^{n-1})\mathbf{P}$ at time n will satisfy $\left\|\nu\mathbf{P}^n - \mu\right\| \leq (1 - m)^n\|\nu - \mu\|$. Hence, because $m > 0$, this implies that $\|\nu\mathbf{P}^n - \mu\|$ tends to 0 exponentially fast, which means that the chain possesses an extremely strong form of ergodicity.

Other directions

In this article we have discussed only the most elementary examples of Markov processes. In particular, in order to avoid technical difficulties, all our considerations have been about processes for which the time parameter is discrete. As soon as one moves into the realm of processes with a continuous time parameter, the theory becomes much more technically involved. However, the price which one has to pay in technicalities is amply rewarded by the richness of the continuous time theory. To wit, Brownian motion (also known as the Wiener process) is a continuous parameter Markov process which makes an appearance in a surprising, and ever growing, number of places: harmonic analysis in pure mathematics, filtering and separation of signal from noise in electrical engineering, the kinetic theory of gases in physics, price fluctuations on the stock market in economics, and so on. Thus, for the sake of the curious, the bibliography below gives a very brief and enormously inadequate list of places where one can learn more about Markov processes.

Bibliography

Elementary texts

Karlin, S. and Taylor, H. 1975. *A First Course in Stochastic Processes*, 2nd edn. New York: Academic Press.

Norris, J. 1997. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Stroock, D. 2005. *An Introduction to Markov Processes*. Graduate Text Series No. 230. Heidelberg: Springer-Verlag.

Advanced texts

Dynkin, E. 1965. *Markov Processes*, vols. 1 and 2. Grundlehren Nos. 121 and 122. Heidelberg: Springer-Verlag.

Ethier, S. and Kurtz, T. 1986. *Markov Processes: Characterization and Convergence*. New York: Wiley.

Revuz, D. 1984. *Markov Chains*. North-Holland Mathematical Library, vol. 11. Amsterdam and New York: North-Holland.

Stroock, D. 2003. *Markov Processes from K. Itô's Perspective*. Annals of Mathematical Studies No. 155. Princeton, NJ: Princeton University Press.

Physics texts

Boltzmann, L. 1896, 1898. *Lectures on Gas Theory*, 2 vols, trans. S. Brush. New York: Dover Publications, 1995.

Gibbs, J. 1902. *Elementary Principles in Statistical Mechanics*. New York: Scribner.

How to cite this article

Stroock, Daniel W. "Markov processes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000083> doi:10.1057/9780230226203.1041

The New Palgrave Dictionary of Economics Online

Markowitz, Harry Max (born 1927)

Donald D. Hester

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Harry M. Markowitz shared the 1990 Nobel Memorial Prize in Economics with Merton Miller and William Sharpe for their contributions to financial economics. He is principally known for his Cowles Foundation monograph, *Portfolio Selection: Efficient Diversification of Investments*, in which he developed and made accessible to general readers the concept of an efficient portfolio, that is, a collection of assets that has a maximum rate of return for an arbitrary rate of return variance. The monograph provided a rigorous justification for portfolio diversification. He has also developed important applied mathematical tools for working with sparse matrices and performing simulations.

Keywords

American Finance Association; capital asset pricing model (CAPM); Cholesky factorizations; customary wealth; expected utility hypothesis; linear programming; Markowitz, H.; portfolio selection; quadratic utility function; risk

Article

Harry M. Markowitz is a Nobel laureate who shared a 1990 prize with Merton Miller and William Sharpe for their contributions to financial economics. A native of Chicago, he received undergraduate and graduate degrees from the University of Chicago, culminating in a Ph.D. in 1954. His article on portfolio selection (1952a), drawn from his dissertation, was a path-breaking contribution that would be fully developed in his 1959 Cowles Foundation monograph, *Portfolio Selection: Efficient Diversification of Investments*. The monograph provided a strong case for receiving the Nobel Memorial Prize.

Markowitz is a gifted applied mathematical economist who responds creatively to observed behaviour and has a strong interest in providing tools that facilitate applications of economics. As a graduate student he published a second influential article (1952b), which extended and qualified an important

contribution by Friedman and Savage (1948) that proposed an explanation for why individuals both insure and gamble. Specifically, he transformed their argument to describe bets that involved deviations from an individual's ‘customary wealth’, which is wealth exclusive of recent windfall gains or losses, and imposed a third inflection point, which was needed to satisfy the expected utility hypothesis requirement that a utility function be bounded from below. By describing how the Friedman and Savage model could not account for some commonly observed behaviour, this article afforded a clear insight into the way Markowitz analysed decisions about risk. It takes only one simple division to transform deviations from an individual's customary wealth to rates of return on customary wealth.

Markowitz's article on portfolio selection lucidly explained why focusing on the expected rate of return (hereafter, ‘return’ means ‘rate of return’) was inadequate to account for widely observed portfolio diversification. By simultaneously considering expected return and the variance of return (E and V), he developed a set of efficient EV portfolios that would have a maximum return for an arbitrary variance of return. Further, almost all of these efficient portfolios would have more than one asset and thus be diversified. Using elegant geometric arguments, the article explained how in a problem involving N securities the set of efficient portfolios could be represented by a set of connected line segments. This insight underlies the algorithm for computing efficient portfolios that is presented in his monograph.

In that article and in his monograph, Markowitz was careful to emphasize that he was developing a method for using an investor's beliefs (or perhaps those of security analysts) about expected return and variance so that he or she could use them in an optimal way. In neither did he explain how expectations should be formed. Similarly, he was agnostic about whether the probabilities investors used in forming expectations were objective or subjective. Finally, he did not assume that returns were normally distributed or that an investor had a quadratic utility function, although one of these conditions is formally necessary to describe portfolio choice in terms of expected return and variance of return. The complications raised in the preceding three sentences are briefly considered in the final section of the monograph and would absorb many journal pages in the coming years. Levy and Markowitz (1979) addressed the limitations of restricting attention to expected return and variance and argued that by focusing on these two measures investors were not likely often to be misled.

The monograph was an expositional tour de force and consequently had an enormous impact on the theory and practice of finance. Its first chapters were quite intuitive and made no technical demands on the reader. The third and fourth chapters contain elementary discussions of the concepts of expected return and variance, the fifth and sixth generalize the discussion to cover large numbers of securities and aggregation over time, and the seventh provides a clear geometric interpretation of efficient portfolios. The eighth chapter presents the critical line method for isolating efficient portfolios and solving the underlying quadratic programming problem. The ninth chapter restates the argument using a semi-variance. The remaining four chapters describe rational portfolio behaviour and discuss how the expected utility hypothesis can be applied to the portfolio selection problem. They include the topics of portfolio choice over time and when objective and subjective probabilities differ.

Technical derivation of the critical line method is reported in Appendix A in the monograph, which generalizes its original exposition in Markowitz (1956). The method works because the set of efficient portfolios is convex, in part because there are assumed to be upper and lower bounds on the holdings of any asset. In the monograph no short sales are allowed, although this restriction can be relaxed. If we ignore some minor technical issues involving singularities that cannot be dealt with here, the method can be described intuitively. It is initiated by finding the security with the highest expected return. A

portfolio fully invested in this security is an element in the set of efficient portfolios. Then, find the security or linear combination of securities that can be substituted for that highest-yielding security in a manner which respects the balance sheet identity that the sum of asset shares equals unity and provides the minimum reduction in return per unit decrease in variance. This substitution is continued until one or more of the securities reach zero (or a lower bound) or until a security not in this combination can be beneficially introduced, at which point another linear combination is chosen. The algorithm stops when no substitution is possible that further lowers the variance of a portfolio.

The monograph and a contemporaneous paper by Tobin (1958) underlay the development of the capital asset pricing model (CAPM) by Sharpe (1964), which argued that an asset's return was determined by its correlation with the return of the market portfolio. This model greatly increased interest in EV models among practitioners and the academic community. In his presidential address to the American Finance Association, however, Markowitz (1983) expressed some reservations about the CAPM because it failed to take into account limits on borrowing.

Apart from his work on modelling portfolio decisions, Markowitz made significant contributions to management science. In Markowitz (1957), he developed sparse matrix techniques for simplifying the solution of linear programming problems, which continue to be used in present-day algorithms that employ Cholesky factorizations. In Markowitz and Manne (1957) an important set of applications, discrete programming problems, were analysed. In Manne and Markowitz (1963), applications of ‘process analysis’ are reported in which Markowitz was a co-author on several papers that studied metal-working industries. Process analysis examines production capabilities in an industry. Also, he made many contributions that led to improvements in simulations, including the construction of a programming language, SIMSCRIPT (see Markowitz, Hausner and Karr, 1963, and Dimsdale and Markowitz, 1999).

Markowitz spent much of his career outside academia. From 1952 through 1963 he was on the staff of the RAND Corporation and from 1974 through 1983 he was at IBM's T. J. Watson Research Center. His monograph was largely written when he was a visitor at Yale University in 1955–6, on leave from RAND. He joined the faculty of Baruch College of the City University of New York as a distinguished professor of finance and economics in 1982, and in 2004 was a research professor at the University of California at San Diego. In 1989 he was awarded the prestigious Von Neumann Prize in Operations Research by the Operations Research Society of America and The Institute of Management Science for his work on ‘portfolio selection, mathematical programming, and simulation’.

See Also

- computational methods in econometrics
- efficiency bounds
- expected utility hypothesis
- foreign direct investment
- risk
- risk-coping strategies
- Sharpe, William F.
- Tobin, James

- uncertainty

Selected works

1952a. Portfolio selection. *Journal of Finance* 7, 77–91.

1952b. The utility of wealth. *Journal of Political Economy* 60(2), 151–58.

1956. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly* 3, 111–33.

1957. The elimination form of the inverse and its application to linear programming. *Management Science* 3, 255–69.

1957. (With A. Manne.) On the solution of discrete programming problems. *Econometrica* 25, 84–110.

1959. *Portfolio Selection: Efficient Diversification of Investments*. Cowles Foundation Monograph 16. New York: John Wiley and Sons.

1963. (With B. Hausner and H. Karr.) *SIMSCRIPT: A Simulation Programming Language*. Englewood Cliffs, NJ: Prentice Hall.

1963. (Co-edited with A. Manne.) *Studies in Process Analysis: Economy-Wide Production Capabilities*. Cowles Foundation Monograph 18. New York: John Wiley and Sons.

1979. (With H. Levy.) Approximating expected utility by a function of mean and variance. *American Economic Review* 69, 308–17.

1983. Negative or not nonnegative: a question about CAPMs. *Journal of Finance* 38, 283–95.

1999. (With B. Dimsdale.) A description of the SIMSCRIPT language. IBM Systems Journal 38 (2/3), 151–61.

Bibliography

Friedman, M. and Savage, L. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56, 279–304.

Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, 65–86.

How to cite this article

Hester, Donald D. "Markowitz, Harry Max (born 1927)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000342> doi:10.1057/9780230226203.1042

The New Palgrave Dictionary of Economics Online

marriage and divorce

Yoram Weiss
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We document the increase in marital turnover and survey economic models of the marriage market. Couples match based on attributes but sorting is constrained by costs of search. Divorce is caused by new information on match quality, and remarriage requires further search. Although most men and women marry, they are single more often than before and more children live in one-parent household. The impact on children depends on child-support transfers. Such transfers may rise with the aggregate divorce (remarriage) rates.

Keywords

altruism; assortative matching; child care; collective models; commitment; comparative advantage; complementarity; division of labour; household production; increasing returns; leisure; marriage and divorce; marriage market; matching model; multiple equilibria; Poisson process; reservation utility; search models; sharing rules; stable sharing rule; time use; transferable utility; unitary models of the household

Article

This article summarizes the economic analysis of marriage markets. The first section provides a description of stylized facts that motivate the interest of economists in this problem. It is shown that marital status is closely tied with ‘economic’ variables such as work and wages. We illustrate these facts using mainly US data but the patterns are similar in all developed countries. The second section demonstrates how the tools of economists bear on ‘non-economic’ subjects such as marriage, fertility and divorce, often analysed by researchers from other fields. The final section highlights some connections between the theory and empirical evidence.

Basic facts

Marriage and divorce

The 20th century was characterized by substantial changes in family structure (Figure 1). More men and women are now divorced and unmarried or have alternative arrangements, such as cohabitation. Interestingly, the rise in divorce rates is associated with an increase in remarriage rates (relative to first marriage rates), reflecting higher turnover. Most people had a first marriage, and most divorces end in remarriage. Moreover, the remarriage rate is greater than the first marriage rate and far exceeds the divorce rate, suggesting that, despite the larger turnover, marriage is still a ‘natural’ state (Table 1). Women enter the first marriage faster than men. However, following divorce, men remarry at higher rates than women, especially at old ages. This pattern reflects the earlier marriage of women and their longer lives, which causes the ratio of men to women to decline with age.

Marital histories of men and women, United States, 1996

Age in 1966	Ever married by 1966 (%)		Divorced from first marriage by 1966 (%)		Remarried after first divorce by 1966 (%)	
	Men	Women	Men	Women	Men	Women

25	31.8	50.0	4.6	12.2	55.5	44.0
30	65.4	71.1	16.7	17.2	35.6	49.7
35	77.4	84.1	26.9	26.4	60.7	65.1
40	80.9	85.2	34.0	36.5	66.4	67.6
45	87.3	89.8	41.1	41.6	71.6	68.1
50	93.2	91.3	39.8	42.4	78.3	68.9
55	94.5	95.3	38.2	38.0	79.0	64.1
60	96.6	94.9	34.3	30.7	86.9	64.7

Source: US Census Bureau, Survey of Income and Program Participation (SIPP), 1996 Panel, Wave 2 Topical Module.

Figure 1

Annual numbers (per 1,000) of first marriage, divorce, and remarriage: United States, 1921–1989 (three-year averages). *Source*: National Center of Health Statistics.

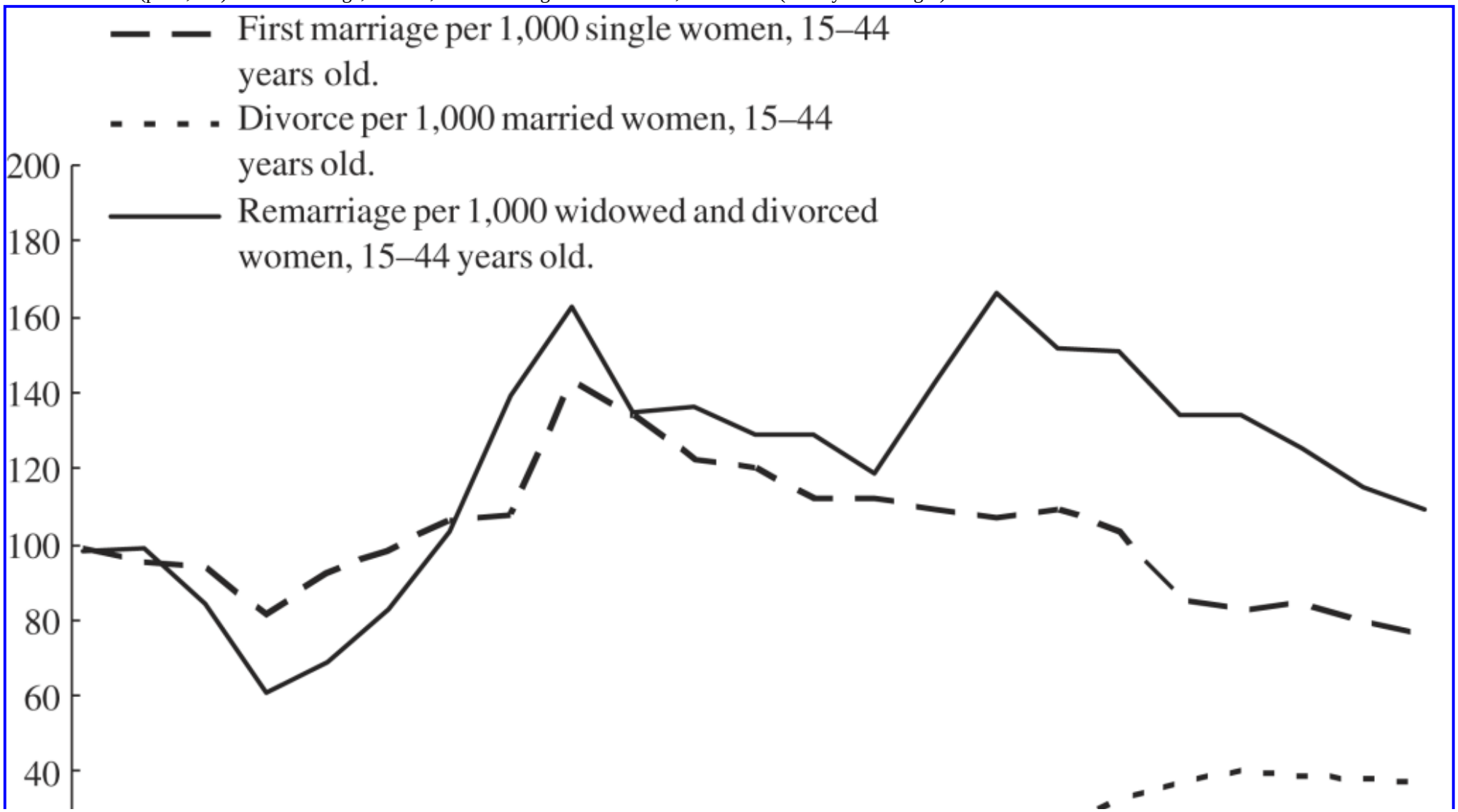

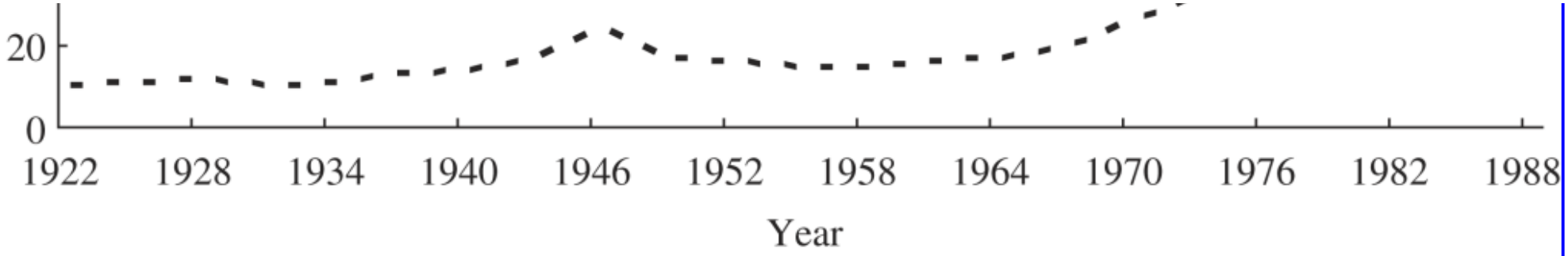

Source: National center of health statistics.

One consequence of higher marital turnover is the large number of children who live in single-parent and step-parent households. In 2002, 23 per cent of US children younger than 18 years lived only with their mother, and five per cent lived only with their father. Children of broken families are more likely to live in poverty and to underperform in school. Lower attainments of such children are observed also prior to the occurrence of divorce, suggesting that bad marriage rather than divorce may be the cause (Piketty, 2003).

Marriage and work

Time use data (Table 2) show that men work more than women in the market; women do more housework than men. Per day, single women work at home three hours while single men work less than two hours. These figures roughly double for married couples with young children, showing clearly that children require a substantial investment of time and that most of this load is carried by the mother. The total time worked and the corresponding amount of leisure is about the same for married men and women.

Daily hours of work of men and women (age 20–59) in the market and at home, by marital status, selected countries and years

	US 1985	Can. 1982	UK 1985	Ger. 1992	Italy 1989	Norw. 1990
Paid work						
Single men	5.5	5.6	4.2	6.4	4.9	4.7
Single women	4.6	4.3	3.3	5.0	3.3	4.0
Married men, no child	6.2	6.2	5.5	6.3	5.5	5.7
Married women, no child	3.3	4.0	3.8	3.3	2.0	4.2
Married men, child 5–17	6.1	5.9	5.7	6.7	6.1	6.0
Married women, child 5–17	3.5	3.7	2.6	3.2	2.2	3.6
Married men, child<5	6.9	6.2	6.1	6.8	6.2	5.7
Married women, child<5	1.9	2.4	2.0	2.2	1.9	2.1
Housework (including child care)						
Single men	1.6	1.7	2.2	1.6	0.7	1.7
Single women	2.8	3.3	3.9	3.4	3.1	2.9
Married men, no child	1.8	2.0	3.3	2.2	1.3	2.1
Married women, no child	4.1	3.9	3.8	4.8	6.4	3.5
Married men, child 5–17	2.3	2.5	2.1	2.3	1.2	2.4

Married women, child 5–17	4.4	4.7	5.5	5.5	7.0	4.5
Married men, child<5	2.3	3.2	2.3	2.8	1.5	3.2
Married women, child<5	6.4	6.8	3.8	6.9	7.6	6.1

Source: Multinational Time Use Study.

Figure 2 displays the work patterns within couples. The most common situation is that the husband works full-time and the wife works part-time or does not work at all. However, the proportion of such couples has declined and the proportion of couples where *both* partners work full-time has risen sharply, reflecting the increased entry of married women into the labour force.

Figure 2

Work patterns of husbands and wives (age 30–40), United States, 1964–2001. *Note*: A spouse is employed full-time-full-year (FTFY) if he/she works 50 weeks or more and hours exceed 34 per week. *Source*: Current Population Surveys.

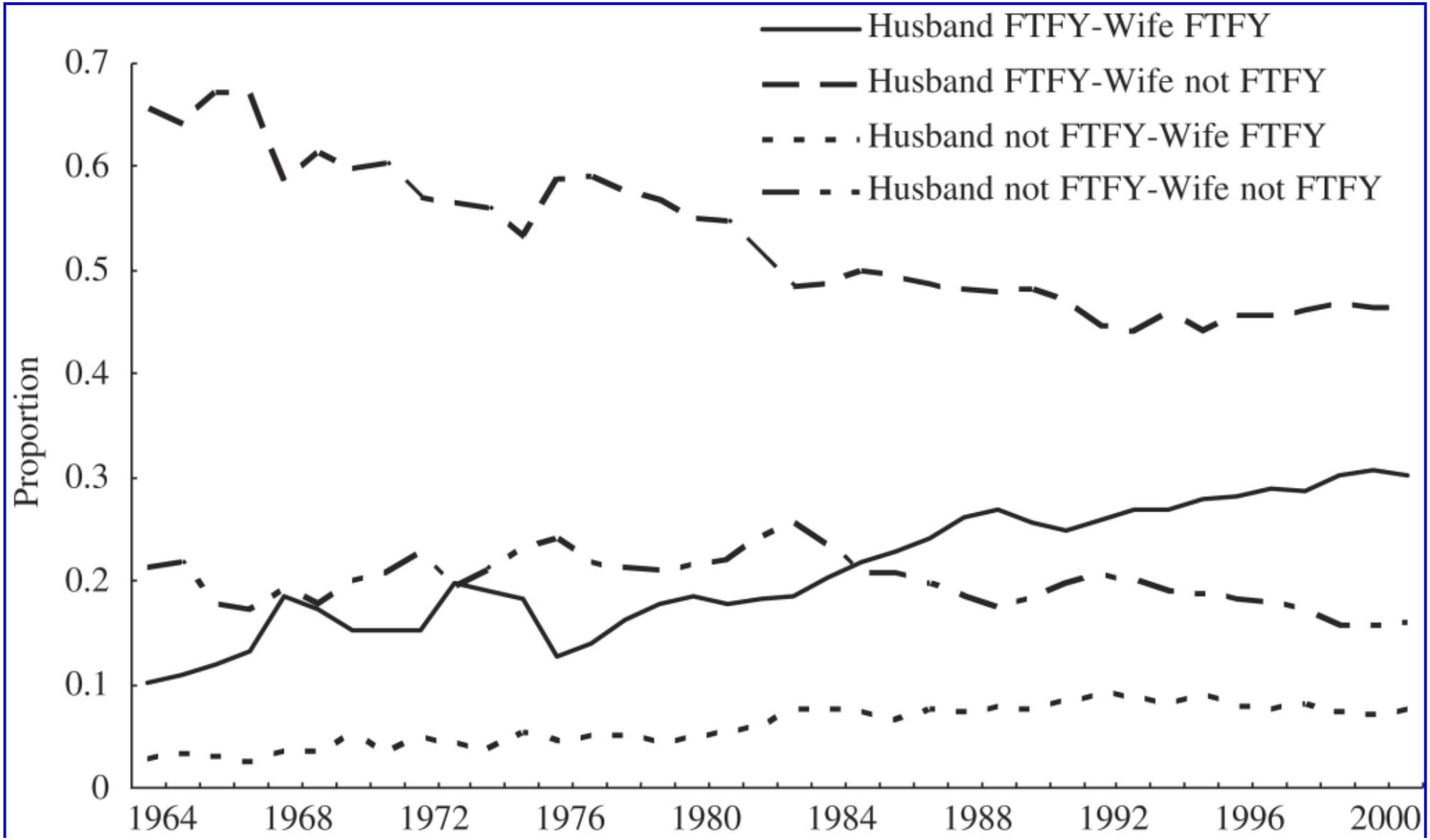

Year

Note: A spouse is employed full-time-full-year (FTFY) if he/she works 50 weeks or more and hours per week

Source: Current population surveys.

Marriage and wages

Male–female wage differences of full-time workers are larger among married than among single persons. Married men have consistently the highest wage among men, while never-married women have the highest wage among women. The wage gap between married men and women rises as the cohort ages, reflecting the cumulative effects of gender differences in the acquisition of labour market experience (Figures 3 and 4). The increased participation of married women, associated with the increase in their wages, has increased their wage relative to those of never-married women and their husbands (Table 3).

Relative wage gaps associated with marital status for fully employed men and women, by year and age, United States, 1965–2001

Years/age	Married–never married		Married–divorced		Mar. men–mar. women between groups	Husband–wife within couples
	Men	Women	Men	Women		
1965–74						
25–34	13.8	–8.8	9.6	4.5	37.2	32.5
35–44	21.5	–17.6	17.1	–1.6	52.1	42.7
1975–84						
25–34	15.6	–6.5	8.5	–.5	35.4	29.6
35–44	21.0	–17.5	12.4	–2.8	52.1	43.8
1985–94						
25–34	15.6	–2.0	15.4	7.7	23.6	21.1
35–44	21.3	–9.9	15.4	2.4	38.7	32.1
1995–2001						
25–34	13.6	2.3	13.6	2.3	17.0	18.1
35–44	23.7	–1.8	21.4	7.8	31.7	27.5

Source: Current Population Surveys.

Figure 3

Hourly wages (in logs) of fully employed married and never-married US men and women born in 1946–1950, by age. *Source*: Current Population Surveys.

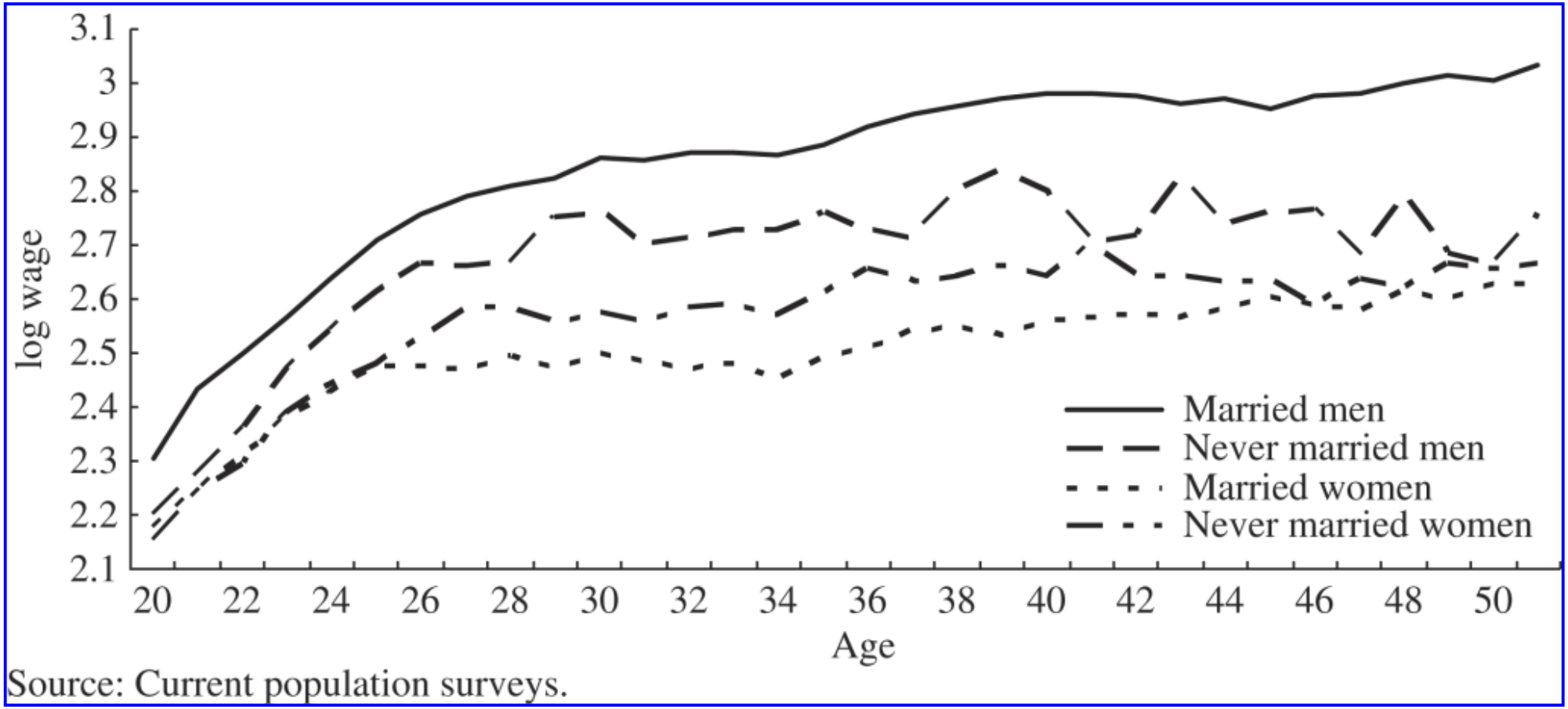

Figure 4
Hourly wages (in logs) of fully employed married and divorced US men and women born in 1946–1950, by age. *Source*: Current Population Surveys.

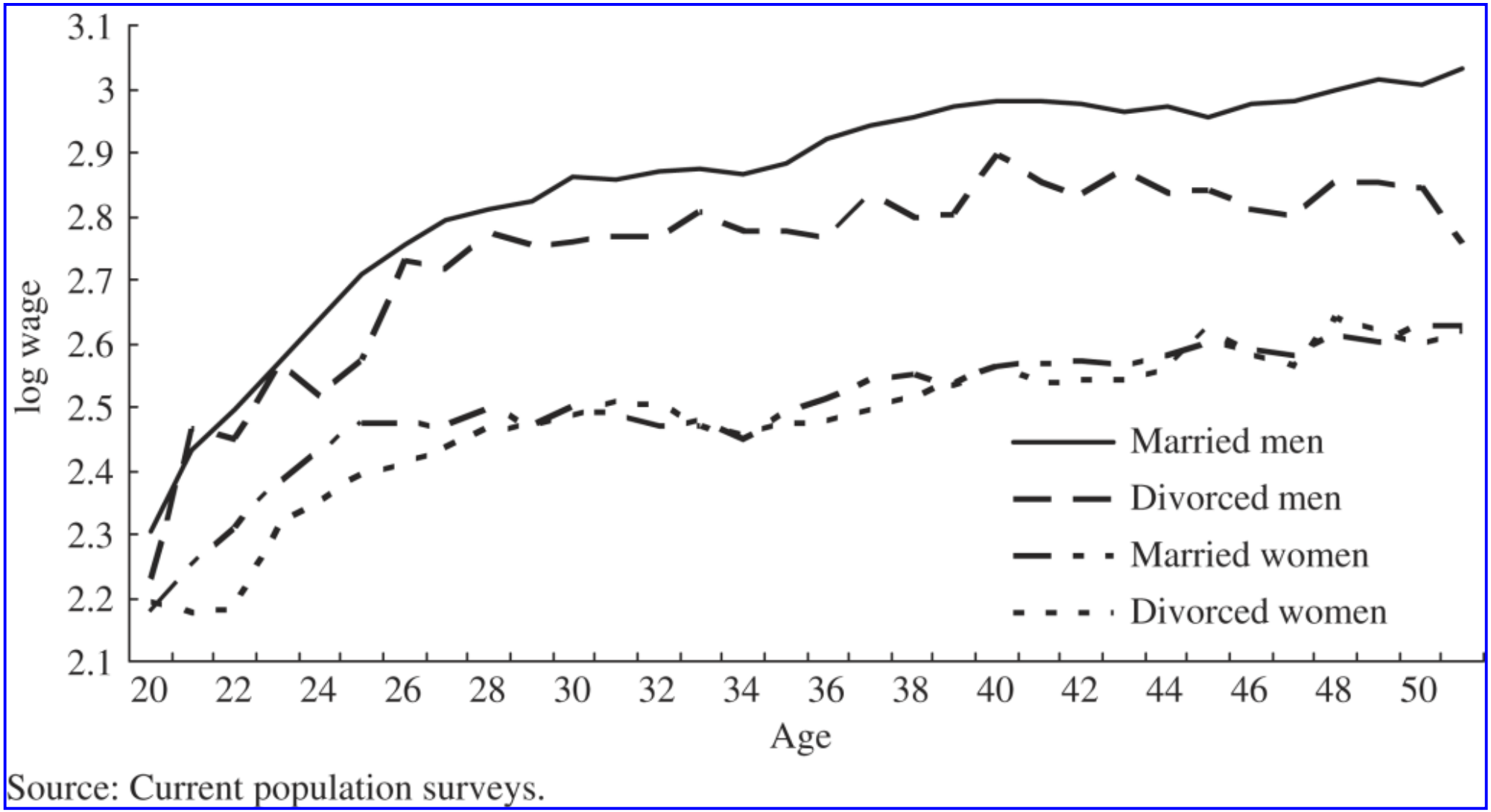

Source: Current population surveys.

Economic theory of marriage and divorce

From an economic point of view, marriage is a voluntary partnership for the purpose of joint production and joint consumption. As such, it is comparable to other economic organizations that aim to maximize some private gains but are subject to market discipline.

Gains from marriage

Consumption and production in the family are broadly defined to include non-marketable goods and services, such as companionship and children. Indeed, the production and rearing of children is the most commonly recognized role of the family. We mention here five broad sources of *economic* gain from marriage, that is, why ‘two are better than one’:

1. 1. Sharing of collective (non-rival) goods; both partners can equally enjoy their children, share the same information and use the same home.
2. 2. Division of labour to exploit comparative advantage or increasing returns; one partner works at home and the other works in the market.

3. 3. Extending credit and coordination of investment activities; one partner works when the other is in school.
4. 4. Risk-pooling; one partner works when the other is sick or unemployed.
5. 5. Coordination of child care, which is a collective good for the parents. Although children can be produced and raised outside the family, the family has a substantial advantage in carrying out these activities. Two interrelated factors cause this advantage: by nature, parents care about their own children and, because of this mutual interest, it is more efficient that the parents themselves determine the expenditure on their children. If the parents live separately, whether single or remarried, the non-custodian parent loses control of child expenditures. Lack of contact further reduces the incentive or ability to contribute time and money to the children. Together, these factors reduce the welfare of both parents and children when they live apart (Weiss and Willis, 1985).

Family decision making

The existence of potential gains from marriage is not sufficient to motivate marriage and to sustain it. Prospective mates are concerned whether the potential gains will be realized and how they are divided. Family members have potentially conflicting interests and a basic question is how families reach decisions. The old notion that families maximize a common objective appears to be too narrow. Instead of this *unitary* model, it is now more common to consider *collective* models in which partners with different preferences reach some binding agreement that specifies an *efficient* allocation of resources and a *stable sharing rule*. (Browning, Chiappori and Weiss, 2005, ch. 3).

In a special case, referred to as *transferable utility*, it is possible to separate the issues of efficiency and distribution. This situation arises if there is a commodity (say, money) that, upon changing hands, shifts utilities between the partners at a fixed rate of exchange. In this case, the family decision process can be broken into two steps: actions are first chosen to maximize a weighted *sum* of the individual utilities, and then money is transferred to divide the resulting marital output. In general, the problems of efficiency and distribution are intertwined. We may still describe the family as maximizing a weighted sum of the individual utilities, but the weights depend on the individual bargaining powers, and any shift in the weights will affect the family choice. The bargaining power may depend on individual attributes such as earning capacity, subjective factors such as impatience and risk aversion, and on market conditions, such as the sex ratio and availability of alternative mates (Lundberg and Pollak, 1993).

The question remains: what enforces the coordination between family members? One possibility is that the partners sign a formal ‘marriage contract’ that is enforced by law. However, such contracts are quite rare in modern societies, which can be probably ascribed to a larger reliance than in the past on emotional commitments and the presumption that too much contracting can ‘kill love’. In the absence of legal enforcement, efficient contracts may be supported by repeated interactions and the possibility to trade favours and punishments. This possibility arises because marriage is a durable relationship, forged by the long-term investment in children and the accumulation of marital specific capital, which is lost or diminished in value if separation occurs. However, repeated game arguments cannot explain unconditional giving, such as taking care of a spouse stricken by Alzheimers who would never be able to return the favour. Emotional commitments and altruism play a central role in enforcing family contracts (Becker, 1991, ch. 8).

The marriage market

Individuals in society have many potential partners. An undesired marriage can be avoided or replaced by a better one. This situation creates competition over the potential gains from marriage. In modern societies, explicit price mechanisms are not observed. Nevertheless, the assignment of partners and the sharing of the gains from marriage can be analysed within a market framework.

Matching models provide a starting point for such analysis. These models investigate the mapping from preferences over prospective matches into a stable assignment (Roth and Sotomayor, 1990). An assignment is said to be *stable* if no married person would rather be single and no two (married or unmarried) persons prefer to form a new union. To illustrate, assume that each male is endowed with a single trait, m, and each female is endowed with a single trait, f. Let

$$z = h(m, f). \tag{1}$$

be the *household production function* that summarizes the impact of traits of the matched partners on marital output, z, and assume that $h(m, f\bullet)$ is increasing in m and f.

Suppose, first, that z is a public good that the partners must consume jointly. Then, the only stable assignment is such that males with high m marry females with high f, and, if there are more (fewer) eligible men than women, the men (women) with the fewest endowments remain unmarried. All men want to marry the best woman, and she will accept only the best man. After this pair is taken ‘out of the game’, we can apply the same argument to the next-best couple and proceed sequentially. Such a matching pattern is called positive

assortative matching.

If one assumes, instead, that z can be divided between the two partners and that utility is transferable, then a man with low m may obtain women with high f by giving up part of his private share in the gains from marriage. The type of interaction in the gains from marriage determines the willingness to pay for the different attributes. Complementarity (substitution) means that the two traits interact in such a way that the benefits from a woman with high f are higher (lower) for a male with high m than for a male with low m. Thus, a positive (negative) assortative matching occurs if the two traits are complements (substitutes). An important lesson is that in a marriage market with sufficient scope for compensation within marriage, the best man is not necessarily the one married to the best women, because, with negative interaction, either one of them can be bid away by the second-best of the opposite sex (Becker, 1991, ch. 4).

What determines the division of marital gains? If each couple is considered in isolation then, in principle, any efficient outcome is possible, and one has to use bargaining arguments to determine the allocation. However, in an 'ideal' frictionless case, where partners are free to break marriages and swap partners at will, the outcome depends on the joint distribution of male and female characteristics in the market at large. Traits of the partners in a particular marriage have no direct impact on the shares of the two partners, because these traits are endogenously determined by the requirement of stable matching.

These features show up more clearly if one assumes a continuum of agents and continuous marital attributes. Let $F(m)$ and $G(f\bullet)$ be the cumulative distributions of the male and female traits, respectively, and let the measure of women in the total population be r, where the measure of men is normalized to 1. Assume that the female and male traits are complements and transferable utility. Then, if man m' is married to woman $f\bullet'$, the set of men with m exceeding m' must have the same measure as the set of women with f above $f\bullet'$. Thus, for all m and f in the set of married couples,

$$1 - F(m) = r(1 - G(f)). \tag{2}$$

This simple relationship determines a positively sloped matching function, $m = \varphi(f)$.

A *sharing rule* specifies the shares of the wife and husband in every marriage that forms. Let $v(m)$ be the *reservation utility* that man m requires in *any* marriage and let $u(\bullet f\bullet)$ be the reservation utility of woman f. Then the sharing rule that supports a stable assignment must satisfy

$$v(m) = \max_f (h(f, m) - u(f)),$$

and

$$u(f) = \max_y (h(z, y) - v(y)). \tag{3}$$

That is, each married partner gets the spouse that maximizes his or her 'profit' from the partnership over all possible alternatives. As we move across matched couples, the welfare of each partner changes according to the *marginal* contribution of his/her *own* trait to the *marital* output, irrespective of the potential impact on the partner whom one marries. With a continuum of agents, there are no rents in the marriage market because everyone receives roughly what can be obtained in the next-best alternative. Another condition for a stable assignment is that, if there are unmarried men, the least attractive married man cannot get any surplus from marriage. Otherwise, slightly less attractive men could bid away his match. A similar condition applies for unmarried women.

From these considerations, one can obtain a unique sharing rule, provided that $r \neq 1$. Basically, one first finds the sharing in the least attractive match, using the no-rent condition. Then the division in better marriages is determined sequentially, by using the condition that along the stable matching profile each partner receives his or her marginal contribution to the marital output. The sharing rule is fully determined by the sex ratio and the respective trait distributions of the two sexes. It can be shown that a marginal increase in the ratio of women to men in the marriage market improves (or leaves unchanged) the welfare of all men, and reduces (or leaves unchanged) the welfare of all women. From (2), it is seen that an upward (downward) first-order shift in the distributions of traits is equivalent (in terms of the effects on the sharing rule) to a marginal increase (decrease) in the female–male ratio. In this regard, there is close correspondence between the impact of changes in quality (that is, the average trait) and size of the two groups that are matched in the marriage market

(Browning, Chiappori and Weiss 2005, ch. 9).

Search

The process of matching in real life is characterized by scarcity of information about potential matches. Models of search add realism to the assignment model because they provide an explicit description of the sorting process that happens in real time.

Following Mortensen (1988), consider infinitely lived agents and assume that meetings are governed by a Poisson random process (these two assumptions are made to ensure a stationary environment). The total marital output is observed upon meeting and, on the assumption of transferable utility, marriage will occur whenever this marital output exceeds the *sum* of the values of continued search of the matched partners. This rule holds because it implies the existence of a division within marriage that makes both partners better off. Because meetings are random and sparse in time, those who actually meet and choose to marry enjoy a positive rent. The division of these rents between the partners is an important issue. Two considerations determine the division of the gains from marriage: outside options, reflected in the value of continued search, and the self-enforcing allocation that would emerge if the marriage continued without agreement (Wolinsky, 1987). If these two considerations are combined, the sharing rule is influenced by both the value of search as single and the value of continued search during the bargaining process, including the option of leaving when an outside offer arrives. In this way, a link is created between the division of marital output gains and market conditions.

Search models explain why, despite the gains from marriage, part of the population is not married and individuals move between married and single states. The steady state proportions of the population in each state are such that the flows into and out of each state are equalized. These two flows are determined by the search strategies that individuals adopt.

Search models may have significant externalities. For instance, it may be easier to find a mate if there are many singles searching for mates. There are several possible reasons for such *increasing returns* in the matching process. One reason is that the two sexes meet in a variety of situations (work, sport, social life and so on) but many of these meetings are ‘wasted’ in the sense that one of the individuals is already attached and not willing to divorce. A second reason is that the establishment of more focused channels, where singles meet only singles, is costly. These will be created only if the ‘size of the market’ is large enough. Third, the intensity of search by unattached decreases with the proportion of attached people in the population who are less likely to respond to an offer (Mortensen, 1988). In such a case, the marriage (divorce) rates will be above (below) their efficient levels, as each person fails to consider the effect of marriage or separation on the prospects of other participants in the marriage market.

Search and assortative matching

The presence of frictions modifies somewhat the results on assortative matching. Following Burdett and Coles (1999), consider a case of non-transferable utility with frictions. Assume that if man *m* marries women *f*, he gets *f* and she gets *m*. There is a continuum of types with continuous distributions and meetings are generated by a Poisson process with parameter λ. Upon meeting, each partner decides whether to accept the match or to continue the search. Marriage occurs only if both partners accept each other and, by assumption, a match cannot be broken.

Each man (woman) chooses a reservation policy that determines which women (men) to accept. The reservation values for men and women, R_m and R_f, respectively, depend on the individual's own trait. Agents at the top of the distribution of each gender can be choosier because they know that they will be accepted by most people on the other side of the market. Hence, continued search is more valuable for them. Formally, let

$$R_m = b_m + \frac{\lambda \mu_m}{r} \int_{R_m}^{\bar{f}} (f - R_m)\, dG_m(f), \tag{4}$$

$$R_f = b_f + \frac{\lambda \mu_f}{r} \int_{R_f}^{\bar{m}} (m - R_f)\, dF_f(m)$$

where the flow of benefits as single, *b*, the proportion of meetings that end in marriage, μ, and the distribution of 'offers' if marriage occurs, all depend on traits, as indicated by the *m* and *f* subscripts. The common discount factor, *r*, represents the cost of waiting.
In equilibrium, the reservation values of all agents must be a best response against each other, yielding a (stationary) Nash equilibrium. In particular, the 'best' woman and the 'best' man will adopt the policies

$$R_{\overline{m}} = b_{\overline{m}} + \frac{\lambda}{r}\int_{R_{\overline{m}}}^{\overline{f}} (f - R_{\overline{m}})\, dG(f), \tag{5}$$

$$R_{\overline{f}} = b_{\overline{f}} + \frac{\lambda}{r}\int_{R_{\overline{f}}}^{\overline{m}} (m - R_{\overline{f}})\, dF(m).$$

Thus, the best man accepts some women who are inferior to the best woman and the best woman accepts some men who are inferior to the best man, because a bird in the hand is worth two in the bush.
The assumption that the ranking of men and women is based on a single trait introduces a strong commonality in preferences whereby all men agree on the ranking of all women and vice versa. Because all individuals of the opposite sex accept the best woman and all women accept the best man, μ is set to 1 in eq. (5) and the distribution of offers equals the distribution of types in the population. Moreover, if the best man accepts all women with *f* in the range $[R_{\overline{m}}, \overline{f}]$, then all men who are inferior in quality will also accept such women. But this means that all women in the range $[R_{\overline{m}}, \overline{f}]$ are sure that all men accept them and therefore will have the *same* reservation value, $R_{\overline{f}}$, which in turn implies that all men in the range $[R_{\overline{f}}, \overline{m}]$ will have the same reservation value, $R_{\overline{m}}$.
These considerations lead to a *class structure* with a finite number of distinct classes in which individuals marry each other. Having identified the upper class, we can then examine the considerations of the top man and woman in the rest of the population. Lower-class individuals face $\mu < 1$ and a *truncated* distribution of offers because not all meetings end in marriage but, in principle, these can be calculated and then one can find the reservation values for the highest two types and all other individuals in the group forming the second class. Proceeding in this manner to the bottom, it is possible to determine all classes. This pattern is similar to the case without frictions and non-transferable utility except that, because of the need to compromise, low- and high-quality types mix within each class.
With frictions and transferable utility, there is still a tendency towards positive (negative) assortative matching based on the interaction in traits. If the traits are complements, individuals of either sex with a higher endowment will adopt a more selective reservation policy and will be matched, on the average, with a highly endowed person of the opposite sex. However, with sufficient friction it is possible to have negative assortative matching even under complementarity. This, again, is driven by the need to compromise. With low frequency of meetings and costs of waiting, males with low *m* expect some women with high *f* to accept them. If the gain from such a match is large enough, they will reject all women with low *f* and wait until a high *f* woman arrives.

Divorce and remarriage

Divorce is motivated by uncertainty and changing circumstances. Thus, individuals may enter a relationship and then break it if a better match is met. Or changing economic and emotional circumstances may dissipate the gains from marriage. As time passes, new information on match quality and outside options is accumulated, and each partner decides whether to dissolve the partnership. In making this choice, partners consider the expected value of each alternative, where the value of remaining married includes the option of later

divorce and the value of divorcing includes the option of later remarriage. Under divorce at will, divorce occurs endogenously whenever one partner has an alternative option that the current spouse cannot, or is unwilling to, match by a redistribution of the gains from marriage.

Following divorce, the options for sharing and coordination of activities diminish. The divorced partners may have different economic prospects, especially if children are present. Asymmetries arise because the mother usually loses earning capacity as a result of having a child. To mitigate these risks, the partners have a mutual interest in signing binding contracts that stipulate post-divorce transfers. Such contracts are negotiated ‘in the shadow of the law’ and are legally binding. Child support payments are mandatory but the non-custodial father may augment the transfer to influence child expenditures by the custodial mother. Payments made to the custodial mother are usually fungible and, therefore, the amount that actually reaches the children depends on the mother's marital status. If she remarries, child expenditures depend on the new husband's net income, including his child-support commitments to his ex-wife. Hence, the willingness of each parent to provide child support depends on commitments of others. These interdependencies can yield *multiple equilibria*, with and without children and correspondingly low and high divorce rates (Browning, Chiappori and Weiss, 2005, ch. 11).

Theory and evidence

There is a growing body of empirical research that addresses the testable implications of the models outlined above.

1. 1. The unitary model of the household implies that the consumption levels of husband and wife depend only on *total* family income. This, however, is rejected by the data (Lundberg, Pollak and Wales, 1997). Nevertheless, consumption and work patterns of married couples indicate that they act efficiently (Browning and Chiappori, 1994), implying that a collective model fits the data.
2. 2. Matching models with transferable utility imply positive assortative matching based on the spouses’ schooling but negative matching based on their wages (Becker, 1991, ch. 10). In fact, the correlation between the education levels of married partners (about .6) is substantially higher than the correlation between their wages (about .3).
3. 3. Because partners are matched based on their traits as observed at the time of marriage, both positive or negative *surprises* trigger divorce (Becker, 1991, ch. 10). Weiss and Willis (1997) find an impact of unexpected changes in husband's and wife's incomes on the probability of divorce.
4. 4. Unanticipated shocks are less destabilizing if partners are well matched. Anticipating that, couples would sort into marriage according to characteristics that enhance the stability of marriage. In fact, individuals with similar schooling are less likely to divorce and are more likely to marry. This pattern holds for religion and ethnicity, too (•Weiss and Willis, 1997).
5. 5. Individual types congregate into locations that facilitate matching; gays in San Francisco (Black et al., 2000) or Jews in New York (Bisin, Topa and Verdier, 2004). Such patterns suggest increasing returns in search. Higher wage variability among men induces women to search longer for their first or second husband, consistently with an optimal search strategy (Gould and Paserman, 2003).
6. 6. Marital choices and family decisions respond to aggregate marriage market conditions. Black women in the United States delay their marriage and have children out of wedlock because of a shortage of eligible black men (Willis, 1999); a higher male–female ratio reduces the hours worked by wives and raises the hours worked by husbands (Chaippori, Fortin and Lacroix, 2002).
7. 7. The sharp increase in divorce in the United States and other countries during 1965–75 seems to constitute a switch across two different equilibria. A marriage market is capable of such abrupt change because of inherent positive feedbacks in matching and contracting. Explanations for the timing of the change include the appearance of the contraceptive pill, the break-up of norms and legal reforms (Michael, 1988; Goldin and Katz, 2002).

See Also

- assortative matching
- collective models of the household
- marriage markets

Bibliography

Becker, G. 1991. *Treatise on the Family*. Cambridge, MA: Harvard University Press.

Bisin, A., Topa, G. and Verdier, T. 2004. Religious intermarriage and socialization in the US. *Journal of Political Economy* 112, 612–65.

Black, D., Gates, G., Sanders, S. and Taylor, L. 2000. Demographics of the gay and lesbian population in the United States: evidence from available systematic data sources. *Demography* 37, 139–54.

Browning, M. and Chiappori, P. 1994. Efficient intra-household allocations: characterization and empirical tests. *Econometrica* 66, 1241–78.

Browning, M., Chiappori, P. and Weiss, Y. 2005. *Family Economics*. Cambridge: Cambridge University Press.

Burdett, K. and Coles, M. 1999. Long-term partnership formation: marriage and employment. *Economic Journal* 109, F307–F334.

Chaippori, P., Fortin, B. and Lacroix, G. 2002. Marriage market, divorce legislation, and household labor supply. *Journal of Political Economy* 110, 37–72.

Goldin, C. and Katz, L. 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. *Journal of Political Economy* 110, 730–70.

Gould, E. and Paserman, D. 2003. Waiting for Mr Right: rising inequality and declining marriage rates. *Journal of Urban Economics* 53, 257–81.

Lundberg, S. and Pollak, R. 1993. Separate spheres bargaining and the marriage market. *Journal of Political Economy* 101, 988–1010.

Lundberg, S., Pollak, R. and Wales, T. 1997. Do husbands and wives pool resources? Evidence from UK Child Benefit. *Journal of Human Resources* 32, 463–80.

Michael, R. 1988. Why did the divorce rate double within a decade? *Research in Population Economics* 6, 367–99.

Mortensen, D. 1988. Matching: finding a partner for life or otherwise. *American Journal of Sociology* 94(supplement), s215–s240.

Piketty, T. 2003. The impact of divorce on school performance: evidence from France, 1968–2002. Discussion Paper No. 4146. London: CEPR.

Roth, A. and Sotomayor, M. 1990. *Two Sided Matching: A Study in Game-theoretic Modeling and Analysis*. Cambridge: Cambridge University Press.

Weiss, Y. and Willis, R. 1985. Children as collective goods. *Journal of Labor Economics* 3, 268–92.

Weiss, Y. and Willis, R. 1997. Match quality, new information, and marital dissolution. *Journal of Labor Economics* 15, S293–S329.

Willis, R. 1999. A theory of out-of-wedlock childbearing. *Journal of Political Economy* 107(6), S33–S64.

Wolinsky, A. 1987. Matching, search, and bargaining. *Journal of Economic Theory* 42, 311–33.

How to cite this article

Weiss, Yoram. "marriage and divorce." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000367> doi:10.1057/9780230226203.1043

The New Palgrave Dictionary of Economics Online

marriage markets

Andrew Foster

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘marriage market’ refers to the application of economic theory to the analysis of the process that determines how men and women are matched to each other through marriage and how this process influences other choices including human capital investment and the allocation of marital surplus. The specific sub-topics in this article include a characterization of stable assignment in marriage, consideration of the effects of marriage allocations on distribution within marriage, discussion of the extent to which partners who marry have similar characteristics, and a review of results on marriage timing.

Keywords

assortative mating; dowries; fertility; human capital investment; inequality; marriage markets; non-market production; stable assignment; transferable utility

Article

The marriage market is a term used by economists to characterize the process that determines how men and women are matched to each other through marriage. Formally the marriage market may be thought of as an allocative process that, given the preferences and endowments of two sets of individuals (men and women), yields a set of couples and unmatched individuals and a distribution of resources within each match. Marriage markets are generally distinguished from other sorting processes such as worker–firm matching by the assumption that each member of each set of individuals is matched to at most one member of the other set. However, the basic concept of the marriage market may also be applied to other cases such as polygamy or same-sex partnerships. It is also generally assumed that one's well-being within marriage is determined by the characteristics of one's partner and the distribution of resources within the marriage, but not the matches of other individuals in the marriage market conditional on these factors. The economics literature on the marriage market has built importantly on a two-part

foundational article on the economics of marriage published in 1973 and 1974 (Becker, 1973; 1974). However, the phrase ‘marriage market’ is considerably older, with a first citation in the *Oxford English Dictionary* of 1842.

Stable assignment

Central to the notion of a marriage market is the notion of stable assignment. A stable assignment may be characterized as a set of partner allocations and distributions of resources within marriage so that no individual of one sex would be willing to make an offer (in terms of partnership and a distribution of resources within that partnership) to an individual of the other sex which that individual strictly prefers to his or her equilibrium allocation.

An early and important divide in terms of economic models of marriage arises with respect to the question of transferable utility. Transferable utility arises when well-being within the household may be freely transferred between members of the household through a reallocation of household resources. Under these conditions the question of who marries whom can be importantly separated from the question of how resources are distributed within marriage and any stable marriage assignment can be characterized as the outcome of the maximization of a linear programme (Bergstrom, 1997).

At the other extreme from a transferable utility model is one in which there is no possibility of transferring resources within or across marriage. A key feature of such models is that there is generally a wide variety of possible stable equilibria. Gale and Shapley (1962) illustrate two such stable equilibria, by the construction of two matching algorithms based on who makes offers and who makes the decision to accept, tentatively accept, or reject those offers. Each man is at least as well off in the equilibrium in which men make offers relative to the equilibrium in which women makes offers and vice versa.

Distributive effects

Becker's (1973) pioneering analysis of the marriage market considered, among other things, the effects of the marriage market on household distribution. Consider, for example, a simplified version of this model in which there is heterogeneity in tastes for being single, transferable utility within marriage, and no heterogeneity across couples in total utility within marriage. The outcome of the model is a distribution parameter that characterizes the share of total marital utility going to each partner within marriage and a number of marriages, with those individuals of both sexes with the highest taste for being single remaining unmarried. Among other things the model illustrates how a rise in the female wage raises the utility of married females within marriage even when married women are not active in the labour market. The increased opportunities for women outside marriage implies that women must, at the margin, receive a higher share of marital utility in order to be willing to marry.

There is substantial debate about the importance of marriage market structure in influencing transfers between partners and their respective households of origin at the time of marriage. Of particular relevance is the evidence of a historical transition from bride-price to dowry in parts of South Asia and the very large levels of dowry relative to annual income that are sometimes observed in that region. A number of factors have been argued to play an important role in this regard, including changes in the relative sizes of female and male populations of marriageable age associated with population growth and

the gap in typical ages at marriage, changes in inequality and economic opportunity, and changes in the relative merits of different forms of parental transfers in their children.

Assortative mating

A second issue that has received significant theoretical and empirical scrutiny is the question of the extent to which the marriage market matches men and women with similar characteristics. This issue is thought to be important because of its implications for interhousehold inequality and for intergenerational transmission of inequality. If high-earning men match with high-earning women, and these high-earning couples transfer these resources to their children in the form of financial assistance and/or human capital, then inequality is likely to be more persistent across generations than would be the case otherwise. Assortative mating by religion and/or immigrant status is also thought to be both an indicator of and contributor to the process of assimilation. Finally, assortative mating on unobservable (to analysts) attributes can affect inferences about household behaviour that condition on household composition. For example, if men with a high unobservable taste for child human capital match with more educated women, then highly educated women will appear to have more educated children even if there is no direct effect.

A simple transferable utility model in which marital output is increasing in the product of male and female quality yields the prediction that there should be positive assortative mating on such attributes as intelligence, wealth and beauty. A possible exception arises with respect to market earnings capacity to the extent that, as postulated by Becker (1973), one member of the couple specializes in the production of non-market goods. Interestingly, the theoretical prediction of positive assortative mating across classes of individuals can arise within the marriage market with imperfect information (Burdett and Coles, 1997).

The evidence supports the prediction of positive assertive mating on partner attributes, although there have been changes over time in the degree to which this is observed. In particular, the degree of educational assortative mating fell between 1940 and 1960 in the United States but has increased subsequently, largely due to a decline in the share of low-education individuals marrying (Schwartz and Mare, 2005). There has also been a shift in the sign of the correlation in partner earnings from negative to positive since the 1960s (Schwartz, 2005), a pattern that has contributed to the overall increase in interhousehold inequality in income.

Marriage timing

A third set of marriage-market issues relates to the timing of marriage, particularly for women. It is argued that early marriage can result in higher fertility, lower rates of human capital investment, and an adverse bargaining position from the perspective of women. Boulier and Rosenzweig (1984), in an early contribution on this subject, showed how unobserved attractiveness could lead to incorrect inference about the role of education in delaying marriage and increasing spousal quality. Bergstrom and Bagnoli (1993) show how the process of uncertainty resolution with regard to the marital prospects may differentially affect the timing of marriage for men and women of different qualities. It also is the case that timing of marriage can play an important role in the equilibration of marriage markets given

substantial differences in the relative numbers of eligible men and women arising from sex differences in mortality or a gap in the age at marriage for men and women for a growing population. In particular, because of how changes in the timing of marriage for sequential cohorts of eligible men and women affect the number of marriages taking place at a particular point in time, a persistent ten per cent excess in the number of eligible females relative to males can be accommodated with an increase in the female relative to male age at marriage by just one year over a decade (Foster, Khan and Protik, 2004).

See Also

- assortative matching
- Becker, Gary S.
- family economics
- household production and public goods
- marriage and divorce
- matching and market design

Bibliography

Becker, G.S. 1973. A theory of marriage: part I. *Journal of Political Economy* 81, 813–46.

Becker, G.S. 1974. A theory of marriage: part II. *Journal of Political Economy* 82, S11–26.

Bergstrom, T. 1997. A survey of theories of the family. In *Handbook of Population and Family Economics*, vol. 1A, ed. M.R. Rosenzweig and O. Stark. New York: North-Holland.

Bergstrom, T.C. and Bagnoli, M. 1993. Courtship as a waiting game. *Journal of Political Economy* 101, 185–202.

Boulier, B. and Rosenzweig, M. 1984. Schooling, search and spouse selection: testing economic theories of marriage and household behavior. *Journal of Political Economy* 92, 712–32.

Burdett, K. and Coles, M. 1997. Marriage and class. *Quarterly Journal of Economics* 112, 141–68.

Foster, A., Khan, N. and Protik, A. 2004. Equilibrating the marriage market in a rapidly growing population: evidence from rural Bangladesh. Working paper, Department of Economics, Brown University.

Gale, D. and Shapley, L. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69, 9–15.

Schwartz, C.R. 2005. Earnings inequality among married couples and the increasing association between spouses' earnings. Working paper, Department of Economics, UCLA.

Schwartz, C.R. and Mare, R.D. 2005. Trends in educational assortative marriage from 1940 to 2003. *Demography* 42, 621–46.

How to cite this article

Foster, Andrew. "marriage markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000354> doi:10.1057/9780230226203.1044

The New Palgrave Dictionary of Economics Online

Marschak, Jacob (1898 – 1977)

Roy Radner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

decision theory; demand theory; econometrics; equilibrium; information, value of; joint demand; Marschak, J.; rationality; socialism; stochastic theory; teams, economic theory of; uncertainty

Article

The diversity of Jacob Marschak's education and early experience made it likely that he would approach the study of economic behaviour with more than the average breadth of interest and vision. He was born in Kiev on 23 July 1898, and studied mechanical engineering at the Kiev Institute of Technology. At the beginning of the Russian Revolution he served briefly as Minister of Labour in the Menshevik government of Georgia but was forced to escape to Germany. There he went first to the University of Berlin, where he studied economics and statistics with L.V. Bortkiewicz, and then to the University of Heidelberg, where he received his Ph.D. in economics in 1922. His professors at Heidelberg included E. Lederer in economics, A. Weber in sociology, K. Jaspers in philosophy and G. Anschuetz in public law. Following his doctoral studies at Heidelberg, he earned his living for the next eight years as an economic journalist and applied economist. He was economic editor for the *Frankfurter Zeitung* (1924–5), a research associate at the Research Centre for Economic Policy in Berlin (1926–8), and supervisor and editor of research for a Parliamentary Commission of Exporting Industries, at the Institute of World Economics of the University of Kiel (1928–30). Also, in 1926 he spent time in London on a travelling fellowship from the University of Heidelberg.

In 1930 he was appointed as a Privatdozent in economics at the University of Heidelberg, but three years later, once again the victim of political events, he left Germany and went to Oxford as a university lecturer. In 1935 he became Reader in Statistics and Director of the Oxford Institute of Statistics, where he remained until 1939. During this period he wrote extensively on theoretical and statistical aspects of demand analysis, a field in which he was a pioneer (Marschak, 1931).

In 1939 Marschak moved to the United States, where he lived the rest of his life, teaching at the New School for Social Research (1940–42), the University of Chicago (1943–55), Yale University (1955–60), and the University of California at Los Angeles (1960–77).

During the first dozen years Marschak was an active participant in the econometric revolution that is commonly associated with the Cowles Commission for Research in Economics. This revolution was nurtured at an early and crucial stage by the seminar on econometric methods and results that Marschak organized at the National Bureau of Economic Research, while he was on the faculty of the New School for Social Research. The intensive contacts fostered in this seminar led, in particular, to three fundamental papers on the statistical estimation of systems of simultaneous equations, by Haavelmo (1943), Mann and Wald (1943), and Marschak and Andrews (1944). Two further publication landmarks in this movement were the Cowles Commission Monographs No. 10 and No. 14, to which Marschak contributed the opening chapters (Marschak, 1950a; 1953).

Two other topics on which Marschak worked presaged his later work on decision and organization. First, he was, for a number of years, interested in the demand for money, and through his work and that of others the idea evolved that this demand could be better understood in the context of a more general theory of the joint demand for various assets (Marschak, 1938; 1949; 1950b). Furthermore, since the ultimate values of assets are rarely known with certainty at the time they are acquired, such a general theory needed to be based on a more systematic theory of decision in the face of uncertainty than was then available.

A second topic was the subject of his first scientific publication, a contribution to the debate on the efficiency, or even viability, of socialism. A central issue in that debate was whether the centralization of economic authority in a socialist state was compatible with the decentralization of information necessary in a complex economy.

From 1950 on, Marschak's research and writing was concerned with the general area of decision, information and organization. More specifically, one can identify at least three topics to which he made substantial contributions: (1) stochastic decision, (2) the economic value of information, and (3) the theory of teams.

Stochastic decision

In a series of articles (Marschak, 1959a; 1964a; Marschak and Block, 1960; Marschak and Davidson, 1959b; Marschak, Becker and DeGroot, 1963a; 1963b; 1963c; 1964b), Marschak proposed and elaborated the theory of stochastic decision and reported on a number of experiments. This work had its roots in the theory of rational economic choice or utility theory and in certain theories of psychological measurement.

Marschak developed a framework for describing the behaviour of economic decision makers who are approximately rational or consistent, or whose consistency of behaviour cannot be exactly verified through observation because of the observer's inability to control or identify all of the relevant factors in the decision-making situation.

It had long been recognized that economic decision makers did not exhibit exact consistency in their detailed choices. Economists were and remain loath to abandon the general framework of rational decision making that has appeared to be so fruitful in the analysis of the economic system as a whole, Marschak's theory provided a theoretical model that could be used for econometric studies of individual choice behaviour and that was connected in a coherent way with the general hypothesis of economic rationality. The work of Marschak and his co-authors was at first more appreciated by psychologists than by economists. His papers on this subject are still standard references in the theory of psychological

scaling (Luce, Bush and Galanter, 1963, vol. 3, ch. 19). More recently, this theory has provided the basis of statistical studies of individual choice behaviour (McFadden, 1982), as well as of a new approach to the theory of economic equilibrium that takes account of the uncertainty of individual behaviour (Hildenbrand, 1971; Bhattacharya and Majumdar, 1973).

Economic value of information

Marschak was probably the first to develop a systematic theory of the economic value of information. In this development he recognized that the measurement of quantity of information used by communication engineers, and associated with the work of Wiener and Shannon, was not adequate to measure the value of information. Indeed, it was not possible to identify a single measure of information such that more is always better.

Instead, Marschak turned to the newly developed theory of statistical decision for the source of his framework. For him, the value of a particular information system – or more generally, a system of information gathering, communication and decision – was related to the particular class of economic decision problems under consideration. His theoretical analysis of the value and cost of information pointed to the importance of more empirical knowledge concerning the technology of observation, information processing, communication and decision making, although he, himself, did not do any empirical work in this field. These ideas are elaborated in a long series of papers beginning with his contribution to *Decision Processes* (Marschak, 1954) and summarized in his paper ‘Economics of Information Systems’ (1971).

Economic theory of teams and organization

In an economic or other organization, the members of the organization typically differ in (1) the actions or strategies available to them, (2) the information on which their actions can be based, and (3) their preferences among alternative outcomes and their beliefs concerning the likelihoods of alternative outcomes given any particular organization action. Marschak recognized that the difficulty of determining a solution concept in the theory of games was related to differences of type 3. However, a model of an organization in which only differences of types 1 and 2 existed, which he called a team, presented no such difficulty of solution concept, and promised to provide a useful tool for the analysis of problems of efficient use of information in organizations. Such a model provided a framework for analysing the problems of decentralization of information so central to both the theory of competition and the operation of a socialist economy. The idea of a team was introduced in Marschak (1954; 1955), and a systematic development of the theory of teams is provided in Marschak and Radner (1972).

Towards the end of his career, Marschak returned to the theoretical issues concerning conflict of interest among the members of a decentralized organization. He approached this primarily in terms of the normative problem of devising incentives for the members of a ‘team’ to behave in accord with the goals of the organization. Of course, to the extent that such incentives are needed, the organization is no longer a team, in the technical sense of the term and the problem is back in the domain of the more general theory of games. It was left to others to make substantial progress on this set of problems. An important early effort in this direction was by T. Groves, who in his doctoral dissertation (1969) and his

subsequent article, ‘Incentives in Teams’ (1973) presented – in a particular case – a solution to the problem of providing incentives to decentralized decision makers to both send truthful messages and make optimal decisions. These ideas were further developed in the contexts of the theory of public goods, the allocation of resources in a divisionalized firm and the principal–agent relationship. (For references to the literature on these developments see Groves and Ledyard, 1987; Hurwicz, 1979; Radner, 1986.)

Besides the significance of Marschak's individual contributions to economic analysis, I would like to emphasize the cumulative significance of his life's work. Through his work ran the important message that economists must come to grips with problems of uncertainty. He led the way, not only through his own research, but through his indefatigable and successful efforts at explaining these problems to his colleagues in economics and related disciplines. His work drew from psychology, statistics and engineering, and in turn influenced research in those disciplines. Indeed, more than any other economist I know, Marschak typified the best in behavioural science.

Selected works

A bibliography of Marschak's publications (excluding most book reviews and all newspaper articles) can be found in McGuire and Radner (1986). A large number of his papers have been reprinted in Marschak (1974).

1923. Wirtschaftsrechnung und Gemeinwirtschaft. *Archiv für Sozialwissenschaft* 51, 501–20.

1931. *Elastizität der Nachfrage*. Tübingen: J.C.B. Mohr.

1938. Money and the theory of assets. *Econometrica* 6, 311–25.

1944. (With W.H. Andrews.) Random simultaneous equations and the theory of production. *Econometrica* 12, 143–205.

1949. Role of liquidity under complete and incomplete information. *American Economic Review* 39, 182–95.

1950a. Statistical inference in economics: an introduction. In *Statistical Inference in Dynamic Economic Models*, ed. T.C. Koopmans. New York: Wiley.

1950b. The rationale of the demand for money and ‘money illusion’. *Metroeconomica* 2, 71–100.

1953. Economic measurements for policy and prediction. In *Studies in Econometric Method*, ed. W.C. Hood and T.C. Koopmans. New York: Wiley.

1954. Towards an economic theory of organization and information. In *Decision Processes*, ed. R.M.

Thrall, C.H. Coombs and R.L. Davis. New York: Wiley.

1955. Elements for a theory of teams. *Management Science* 1, 127–37.

1959a. Binary-choice constraints and random utility indicators. In *Mathematical Methods in Social Sciences*, ed. K.J. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

1959b. (With D. Davidson.) Experimental tests of stochastic decision theory. In *Measurement: Definitions and Theory*, ed. C.W. Churchman and P. Ratoosh. New York: Wiley.

1960. (With H.D. Block.) Random orderings and stochastic theories of responses. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. I. Olkin et al. Stanford: Stanford University Press.

1963a. (With G. Becker and M. DeGroot.) Stochastic models of choice behavior. *Behavioral Science* 8, 41–55.

1963b. (With G. Becker and M. DeGroot.) An experimental study of some stochastic models for wagers. *Behavioral Science* 8, 199–202.

1963c. (With G. Becker and M. DeGroot.) Probability of choices among very similar objects: an experiment to decide between two models. *Behavioral Science* 8, 306–11.

1964a. Actual versus consistent decision behavior. *Behavioral Science* 9, 103–10.

1964b. (With G. Becker and M. DeGroot.) Measuring utility by a single-response sequential method. *Behavioral Science* 9, 226–32.

1971. Economics of information systems. In *Frontiers of Quantitative Economics*, ed. M. Intriligator. Amsterdam: North-Holland.

1972. (With R. Radner.) *Economic Theory of Teams*. New Haven: Yale University Press.

1974. *Economic Information, Decision, and Prediction*. 3 vols. Dordrecht: Reidl.

Bibliography

Bhattacharya, R.N. and Majumdar, M.K. 1973. Random exchange economies. *Journal of Economic Theory* 6, 37–67.

Groves, T. 1973. Incentives in teams. *Econometrica* 41, 617–31.

Groves, T. and Ledyard, J. 1987. Incentive compatibility ten years later. In *Information, Incentives, and Economic Mechanisms: Essays in Honor of Leonid Hurwicz*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Haavelmo, T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1–12.

Hildenbrand, W. 1971. Random preferences and equilibrium. *Journal of Economic Theory* 3, 414–29.

Hurwicz, L. 1979. On the interaction between information and incentives in organizations. In *Communication and Control in Society*, ed. K. Drittendorf. New York: Gordon Breach.

Luce, R.D., Bush, R. and Galanter, E., eds. 1963–5. *Handbook of Mathematical Psychology*. 3 vols. New York: Wiley.

Mann, H.B. and Wald, A. 1943. On the statistical treatment of linear stochastic difference equations. *Econometrica* 11, 173–220.

McFadden, D. 1982. Qualitative choice models. In *Advances in Economic Theory*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.

McGuire, C.B. and Radner, R., eds. 1986. *Decision and Organization*, 2nd edn. Minneapolis: University of Minnesota Press. Originally published Amsterdam: North-Holland, 1972.

Radner, R. 1986. The internal economy of large firms. *Economic Journal* 96, Supplement, 1–22.

How to cite this article

Radner, Roy. "Marschak, Jacob (1898–1977)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000087> doi:10.1057/9780230226203.1045

The New Palgrave Dictionary of Eco

Marshall Plan

Francisco Alvarez-Cuadrado
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Marshall Plan transferred over US$12.5 billion to Western European countries between 1948 and 1951. This article contrasts the main views on its impact on the post-war European performance. It concludes that, although the direct impact of the plan through private and public investment was rather limited, Marshall Aid provided the recipient economies with a temporary solution for the severe dollar constraint that posed a threat to the continuation of the European miracle. Furthermore the Plan played an important role in promoting collaboration among former adversaries.

Keywords

aid; Marshall Plan; OEEC; post-war economics

Article

In Europe during the Second World War the productive effort of more than an entire generation was lost, with per capita income returning to the levels of the turn of the century. This fall in output reflected not only the destruction of capacity but also the disruption of channels for obtaining inputs and distributing production. The reconstruction process began right after the war, with industrial production reaching pre-war levels as soon as 1947. The weather conditions in 1947 substantially depressed agricultural yields, leading to important food and energy shortages. The substantial trade deficits of the recovering economies combined with the negative experience of international investors after the First World War lead to a ‘dollar gap’ which posed a threat

to the continuation of the European miracle.

These were the circumstances that surrounded the development of the Marshall Plan, officially the European Recovery Program (ERP). The views on the motivations behind the ERP range from plain American imperialism (Kolko and Kolko, 1972) to pure altruism; in the words of Galbraith (1998), ‘the primary purpose of the Plan was compassionate good will, the notion that our former allies needed to have the help of the US’. Nonetheless most of the literature acknowledges that, beyond the concern for former allies, the political and social stability of non-communist Europe and the continuity of export markets for US products were two of the main concerns of the Truman administration.

The initial Plan proposal was presented in June 1947 by secretary of state George Marshall, asking European governments to design a coordinated aid programme to be funded by the United States. The offer included the Soviet Union and its allies, but the conditional terms on economic collaboration and disclosure of information guaranteed that the Soviet Union would never accept it. In response to the American offer the final aid recipients, Austria, Belgium, Denmark, France, West Germany, Great Britain, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Sweden, Switzerland and Turkey, formed the Organization for European Economic Cooperation (OEEC) to coordinate a proposal based on national needs and consistent with American objectives on trade and economic cooperation between recipient countries. US President Truman signed the Plan into a law on 3 April 948, establishing the American-led Economic Cooperation Administration (ECA) to administer the programme.

Over the four years that followed its approval by Congress, the Plan transferred $12.5 billion of US aid to Western Europe. The allocation of funds did not follow a simple rule, although it was mainly determined by the dollar balance of payment deficits of the recipient economies, taking also into account geopolitical considerations especially in the cases of France and the United Kingdom. Marshall Aid represented 2.1 per cent of US GNP in 1948, rose to 2.4 per cent in 1949 and then fell to 1.5 per cent in the remaining two years. In terms of national income of the recipient economies, the funds ranged from 0.3 per cent per year for Sweden to 14 per cent for Austria. For the large Western European economies it represented an average yearly transfer of 2.5 per cent of GDP for France, 2.2 per cent for Italy, 1.3 per cent for the United Kingdom and 1.2 per cent for Germany.

The OEEC took the leading role in allocating funds, conditional on the

approval of the ECA. The American supplier was paid in US dollars, which were debited against the ERP account corresponding to the European buyer. This buyer paid for the American imports in local currency, which was deposited by its own government in a counterpart fund. These additional resources were used for local investment projects and eventually were absorbed into the recipient's national budget.

The impact of the plan on the European recovery is not free of controversy. On the one hand, early triumphalist accounts (Jones, 1955; Mayne, 1970; Arkes, 1972) describe the Plan as vital for the reconstruction of productive capacity, the development of the necessary institutions for cooperation among former adversaries, and the restoration of European confidence in market capitalism. In the words of Mayne (1970), Marshall Aid ‘was a precondition of all later affluence and economic miracles, as well as moves toward European unity’. On the other hand, Milward (1984) discounts the importance of ERP transfers, arguing that the recovery was well under way before 1948 and the reconstruction of the damaged private and public capital stocks was almost completed. Somewhere in between, De Long and Eichengreen (1993) argue that Marshall Aid helped the recipient economies more in terms of political economy than macroeconomics. The ERP bought European governments the political space needed to avoid the attrition wars that characterized the interwar period, allowing an institutional environment conducive to growth.

Until the influential work of Milward (1984), the literature agreed on the vital importance of the ERP funds for Western European growth. According to this view Marshall Aid allowed for the reconstruction of the capital stock, the elimination of bottlenecks that obstructed production, the public provision of infrastructures and the surge in intra-European trade. In the words of Arkes (1972), ‘the plan was critical at the margins having a multiplier effect of three or four times its value’. A superficial analysis of the data suggests that this view is exaggerated. If we compute the growth rates of per capita GDP for the recipient economies, we find that in the three years that preceded the ERP, yearly growth averaged 6.5 per cent, while during the Plan it averaged 4.4 per cent, falling to four per cent between 1953 and 1956.

Eichengreen et al. (1992) argue that if the effects of the Plan worked mainly through private and public capital accumulation, there should be a significant correlation between output growth and ERP allotments as a share of GDP across recipient economies. Contrary to this view, their statistical analysis does not turn up a significant coefficient on Aid allotments. Alvarez-Cuadrado

and Pintea (2008) present a two-sector neoclassical growth model with public capital to explore the direct impact of the ERP. Their numerical analysis suggests that the transfers increased the rate of private investment by less than one percentage point, leading to no more than half a percentage point increase in the growth rate of output. Since average yearly transfers represented no more than half a year's worth of post-war growth, it is not surprising that the effects of the Plan through capital accumulation are rather limited. Along similar lines, Milward (1984) argues that the outstanding performance of Western European economies would have not been very different in the absence of the Plan. He discounts the direct impact of the Plan and convincingly documents that the leverage afforded by the ERP was insufficient for the United States of America to force through its vision of the United States of Europe. In Milward's view the primary role of the ERP was limited to sustaining the flow of capital imports necessary to prolong the recovery.

In a different spirit, De Long and Eichengreen (1993) argue that political economy considerations lie behind the true impact of the Plan. Marshall Aid provided the currency needed to relax the foreign exchange constraint, giving European policy makers extra room to manoeuvre. This political space, together with aid conditionality, induced European governments to balance their budgets, restore internal financial stability, and maintain their commitment to free markets. Their counterfactual vision of Western Europe suggests a permanent influence of Communist parties, an expansion of government controls and regulations, and a resurgence of economic nationalism and isolationism.

This argument, although it has its merits, does not seem to account for the variety of institutional arrangements present in the recipient economies. For instance, two of the fastest-growing economies, France and Germany, adopted rather different growth strategies. The French economy was characterized by major involvement of the state in key economic sectors, while the German approach illustrates the growth potential of relatively free markets. In my view, although the functioning of the market mechanism was constrained in many countries as a result of war priorities, there was a tacit consensus that this was only a temporary interruption of the long European experience with free markets, and therefore the influence of the Plan was rather limited in this respect.

To sum up, the direct impact of the Plan led to no more than half of a percentage point of growth per year. Along the lines of Milward (1984), I

believe the plan provided European governments the means to prolong the recovery process which began after the war. In some cases the transfers complemented export revenues, preventing a balance-of-payment crisis, while in others they only postponed its occurrence. The political economy argument is more difficult to evaluate, but given the prior European experience with free markets and the existence of a well-developed system of property rights, it is difficult fully to accept the counterfactual scenario drawn by De Long and Eichengreen (1993). Finally, the Marshall Plan played an important role by inducing British and French support for a strong Germany. Although the forces that led to the process of European integration responded more to internal political and economic developments in the European countries than to American pressure, the Marshall Plan helped promote collaboration among former adversaries.

See Also

- foreign aid

I would like to thank Markus Poschke for helpful comments and suggestions. Of course all remaining errors are mine.

Bibliography

Alvarez-Cuadrado, F. and Pintea, M. 2008. A quantitative exploration of the golden age of European growth. Florida International University, Department of Economics Working Paper no. 0805 (forthcoming in *Journal of Economic Dynamics and Control*).

Arkes, H. 1972. *Bureaucracy, the Marshall Plan, and the National Interest*. Princeton, N.J.: Princeton University Press.

De Long, J.B. and Eichengreen, B. 1993. The Marshall Plan: history's most successful structural adjustment program. In *Postwar Economic Reconstruction and Lessons for the East Today*, ed. R. Dornbusch, W. Nolling and R. Layard. Boston, Mass.: MIT Press.

Eichengreen, B., Uzan, M., Crafts, N. and Hellwig, M. 1992. The Marshall Plan: economic effects and implications for Eastern Europe. *Economic Policy* 7(14), 14–75.

Galbraith, J.K. 1998. Testimony in *The Cold War*, directed by T. Coombs.

Jones, J.M. 1955. *The Fifteen Weeks: An Inside Account of the Genesis of the Marshall Plan*. New York: Viking.

Kolko, J. and Kolko, G. 1972. *The Limits of Power: The World and United States Foreign Policy 1945–54*. New York: Harper & Row.

Mayne, R. 1970. *The Recovery of Europe*. Garden City, N.Y.: Anchor Press.

Milward, A.S. 1984. *The Reconstruction of Western Europe 1945–51*. London: Methuen.

How to cite this article

Alvarez-Cuadrado, Francisco. "Marshall Plan." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_M000442> doi:10.1057/9780230226203.1903

The New Palgrave Dictionary of Economics Online

Marshall, Alfred (1842 – 1924)

John K. Whitaker

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

English economist Alfred Marshall, founder of the Cambridge School of economics, was a leading and internationally prominent figure in the development of economic thought between 1870 and 1920. He played a significant role in professionalizing British economics, always stressing the social importance of wider economic understanding. His influential ideas on economic theory were conveyed primarily in his *Principles of Economics* (1890). Accounts of his life, career and general views on economics are followed by a more technical treatment of his contributions to various aspects of economic analysis. Guides to his writings and to the secondary literature on him are appended.

Keywords

Bernoulli hypothesis; Berry, A.; bimetallism; Böhm-Bawerk, E.; Bowley, A.; British classical economics; British Economic Association; Cambridge School; ceteris paribus; Chapman, S.; Clapham, J.; Clark, J.; combinations; comparative static method; conspicuous consumption; consumer surplus; cost and supply curves; cost of production; Cournot, A.; Cunynghame, H.; cyclical unemployment; deductive method; demand for money; demand price; demand theory; derived demand; differential rent; distribution theories (classical); Dupuit, J.; Edgeworth, F.; envelope theorem; equilibrium analysis; excess burden of taxation; external economies; Fay, C.; Flux, A.; Foxwell, H.; free trade; Gonner, E.; homo economicus; imperfect competition; institutions; internal economies; international trade (theory); interpersonal utility comparisons; Jenkin, H.; Jevons, W.; joint demand and supply; Keynes, J. M.; Keynes, J. N.; Lavington, F.; long-period equilibrium; long-period supply; Macgregor, D.; marginal cost pricing; marginal productivity theory; marginal utility of money; marginal utility theory; market demand elasticity; market interdependence; market value; Marshall, A.; Marshall, M. P.; Menger, C.; Methodenstreit; methodology of economics; Mill, J. S.; monetary theory; monopolistic competition; monopoly; multiple equilibria; multiplier; net product; Nicholson, J.; normal profit; normal value; offer curves; optimal taxation; Pantaleoni, M.; Pareto, V.; partial-equilibrium method; perfect competition;

period analysis; Pigou, A.; poverty; Price, L.; principle of substitution; producer surplus; product differentiation; quasi-rent; Ramsey price; rent; representative firm; Ricardo, D.; Royal Economic Society; Sanger, C.; Say's Law; self-interest; short-period equilibrium; Sidgwick, H.; Smith, A.; subjective theories of value; supply and demand; tariffs; temporary equilibrium; trade unions; utilitarianism; von Thünen, J.; Walker, F.; Walras, L.; welfare economics

Article

Alfred Marshall, Professor of Political Economy at the University of Cambridge from 1885 to 1908 and founder of the Cambridge School of Economics, was born in Bermondsey, a London suburb, on 26 July 1842. He died at Balliol Croft, his Cambridge home of many years, on 13 July 1924 at the age of 81. His magnum opus, *Principles of Economics* (1890a) evolved through eight editions in his lifetime, the final edition (1920) being most commonly cited today. It was one of the most influential treatises of its era and was for many years the Bible of British economics, introducing many still familiar concepts. The Cambridge School rose to great eminence in the 1920s and 1930s. A.C. Pigou and J.M. Keynes, the most important figures in this development, were among Marshall's pupils.
Marshall's biography and career are outlined initially, after which descriptions are given of his views on the social setting, aims and methods of economics, and his intellectual debts to others. An analysis of his fundamental ideas on theories of value and distribution, which were mainly set out in *Principles*, follows, after which his contributions to monetary and international-trade theory are considered briefly. A final section provides additional documentation and general suggestions for further reading. Also, some of the more technical sections have attached to them brief ‘bibliographic notes’ offering suggestions for further exploration. All bibliographic references lacking an author's name are to works by Marshall, and the bibliographic details of all his cited publications can be found in the list of ‘Selected works’ below. The bibliographic details for all cited works written or edited by others are listed in the concluding ‘Bibliography’.

Biography and career

Marshall grew up in the London suburb of Clapham, being educated at the Merchant Taylors’ School where he showed academic promise and a particular aptitude for mathematics. Eschewing the more obvious path of a closed scholarship to Oxford and a classical education, he entered St John's College, Cambridge, in 1862 on an open exhibition. There he read for the Mathematical Tripos, Cambridge University's most prestigious degree competition, emerging in 1865 in the exalted position of Second Wrangler, bettered only by the future Lord Rayleigh. This success ensured Marshall's election to a Fellowship at St John's. Supplementing his stipend by some mathematical coaching, and abandoning – doubtless because of a loss of religious conviction – half-formed earlier intentions of a clerical career, he became engrossed in the study of the philosophical foundations and moral bases for human behaviour and social organization. In 1868 he became a College Lecturer in Moral Sciences at St John's, coming to specialize in teaching political economy. By about 1870 he seems to have committed his career to developing this subject, seemingly ripe for reform, and helping to transform it into a new science of economics.

For several years he laboured persistently to develop and refine his economic ideas, and to deepen his understanding and grasp of both the existing economic literature and the economic reality that was its subject matter. In 1875 he visited the United States to probe economic conditions, and throughout his life he was tireless in his efforts to master the practicalities of the economic world. Prior to 1879 his publications were meagre. He had embarked on a book on international trade and problems of protectionism in the mid-1870s, and before that he had worked out many of his distinctive theoretical ideas in the form of short essays, many now reproduced in Whitaker (1975). But the only part of this material to be made public was four chapters from the theoretical appendices for the proposed international-trade volume. In 1879 Henry Sidgwick had these printed for private circulation under the title *The Pure Theory of Foreign Trade: The Pure Theory of Domestic Values* (1879a). (An amplified version together with surviving portions of the text of the abandoned trade volume is also reproduced in Whitaker, 1975.) The year 1879 also saw the publication of Marshall's first book, *The Economics of Industry* (1879b), written jointly with his wife Mary Paley Marshall.

Mary Paley had been one of the first group of students at Newnham Hall (later Newnham College) where Marshall, an early supporter of the informal scheme of Cambridge lectures for women, taught her political economy. Their marriage in 1877 required Marshall to give up his Cambridge position under the celibacy rules then in force. He found a new livelihood as principal of the recently established University College, Bristol, where he also became Professor of Political Economy. There *The Economics of Industry* was brought to completion and published by the house of Macmillan, which continued as Marshall's publisher thereafter. Ostensibly an elementary primer, this book contained the first general statement of Marshall's emerging theories, and a considerable sophistication lay beneath its deceptively simple surface. Together with the powerful *Pure Theory* chapters published by Sidgwick, a few copies of which circulated outside Cambridge, *The Economics of Industry* marked Marshall as a rising star in the economics firmament. With the death of W. S. Jevons in 1881, he moved into the public eye as the leader in Britain of the new scientific school of economics.

The duties of the Bristol principalship proved irksome to Marshall, especially as the college was struggling financially. He was anxious to proceed with his writing, having by 1877 conceived the plan for the book that was to become the *Principles*. His frustrations were increased by the onset in 1879 of a debilitating illness, diagnosed as kidney stones, which restricted his activities. He was persuaded to continue as principal until 1881, when he resigned both posts at the college. The next year was spent travelling, with an extended sojourn in Palermo, and it was in this year that composition of the new book began in earnest.

At Bristol, Marshall had got to know well Benjamin Jowett, the famed Master of Balliol, who was one of the governors of the struggling college. It was probably by Jowett's generosity that Marshall was able to return to Bristol in 1882 as Professor of Political Economy. And it was doubtless at Jowett's instigation that the Marshalls moved to Oxford in 1883, when a Balliol lectureship became vacant on the unexpected death of Arnold Toynbee. Marshall had considerable success as a teacher in Oxford and appeared settled in for an indefinite stay. But an ‘Oxford School of Economics’ was not to be. The sudden death of Henry Fawcett, who had been Professor of Political Economy at Cambridge since 1863, opened up the irresistible prospect of a return to Cambridge and a position with great potential for academic leadership. Marshall, the dominant candidate, was duly elected in December 1884, holding the chair until 1908, when he resigned to devote himself entirely to writing.

In many ways Cambridge's inviting prospects were to prove illusory. Economics was taught as part of

the Historical and Moral Sciences Triposes, but neither avenue provided a supply of able interested students, nor was there much scope for advanced work. Marshall struggled for many years, with limited success, to increase the scope for economic teaching. But it was not until 1903, with the establishment of a new Tripos in Economics and Politics, that his goal was achieved. Even then, few resources were made available by the university and colleges for the teaching of economics, and the staffing of the new Tripos relied heavily on Marshall's willingness to support two young lecturers from his own pocket. The flowering of the new school came about mainly after his retirement, but the seeds were certainly planted by his efforts.

Absorbed in the struggle for his own subject, Marshall took relatively little part in general university affairs. Indeed, his rather obsessive personality and proneness to magnify details would have made him ineffectual as a university statesman even if he had aspired in that direction. But he did play a prominent part in the successful campaign of 1896–7 against the granting of Cambridge degrees to students of the women's colleges – this despite his wife being at the time a lecturer at Newnham. He was not opposed to women's education, indeed had been a warm supporter in his early days, but was vehemently opposed to the assimilation of women into an educational system designed for men.

But the dominant fact in Marshall's life after his return to Cambridge, and certainly the aspect of greatest interest to posterity, is his long struggle to give adequate written expression to the stores of economic knowledge and understanding he had accumulated. The demands of teaching and administration left him little time or energy for sustained composition during term time and it was in the jealously guarded long vacations, usually spent away from Cambridge on the south coast of England or in the Tyrol of Austria, that the only real progress could be made. By 1887 the book commenced in 1881 had grown into a projected two-volume treatise. He hoped to complete the first volume in time for it to appear in the autumn of that year with the second volume appearing by 1889. In fact, the first volume (1890a) appeared as the *Principles of Economics, Volume One*, only in July 1890, when it was received with great and immediate acclaim and established Marshall firmly as one of the world's leading economists. The second volume never appeared. It was to have covered foreign trade, money, trade fluctuations, taxation, collectivism and aims for the future – a tall order!

Marshall struggled for the next 13 years with his intractable second volume, meanwhile spending much time on substantial, but not very substantive, recastings of the first volume in new editions of 1891, 1895 and 1898, and in preparing a digest of it to replace the earlier *Economics of Industry* which he had come to dislike intensely. (The digest, 1892, appeared under the title *Elements of the Economics of Industry, Volume One*. Like the earlier work it included material on trades unions that was never incorporated into *Principles*.) By 1903 much material had been accumulated for the second volume, but the scope was becoming unmanageable as Marshall became increasingly preoccupied with problems of trusts, trades unions, international trade, and comparative economic development, and decreasingly concerned with matters of pure theory. In that year, partly from the impetus of writing a private memorandum on trade policy for the use of the then Chancellor of the Exchequer, and partly because the tariff controversy was at full heat, Marshall was tempted into writing a short topical book on foreign trade questions, intending to publish it speedily. But this project too grew unmanageably in his hands. In 1907, the preface to the fifth edition of *Principles* (the last major rewriting) announced the abandonment of the proposed continuation and promised instead a volume, already partly in print, on ‘National Industry and Trade’, to be followed soon by a companion volume on ‘Money, Credit and Employment’ (Guillebaud, 1961, vol. 2, p. 46). To reflect this change, the title of the sixth and subsequent editions of *Principles* was changed

to *Principles of Economics: An Introductory Volume*. Retirement in 1908, at the age of 66, freed Marshall to concentrate on these projects, but progress continued to be slow. He appears to have suffered from recurrent dyspepsia and high blood pressure, necessitating a strict regimen and limiting his ability to work. But the more fundamental problem was that the world kept changing and the increasingly realistic and factual tone of his enquiry called for incessant recasting and revision. Nothing had been completed by the time war broke out in 1914, and then much rewriting was required to take into account the radical changes that were transforming the world economy and its post-war prospects. At last, when Marshall was 77 years old, *Industry and Trade* (1919), his second masterpiece, finally appeared. It was a magisterial, largely factual, consideration of trends in the British and international economy and of future economic prospects. But, lacking an obvious theoretical skeleton, it has not received from economists the kind of attention lavished on *Principles*, although interest in it is now beginning to stir among historians of economics.

In its final form, *Industry and Trade* was narrower in scope than had been intended earlier, while the proposed book on ‘Money, Credit and Employment’ still remained to be written. Over the next four years, by a remarkable effort, and despite rapidly waning powers, some of the mass of accumulated raw material remaining was pulled together in *Money, Credit and Commerce* (1923). This contains Marshall's fullest treatment of the theories of money and international trade, but it is an imperfect pastiche of earlier material, some dating back almost 50 years.

In the last months of his life, Marshall toyed with the occasional writings and the memoranda and evidence for governmental enquiries that he had prepared at various stages during his career, with the hope of editing them for publication in book form. This was not to be, but his plan was largely fulfilled after his death in two books sponsored by the Royal Economic Society (Pigou, 1925; Keynes, 1926).

Judged by what might have been, Marshall's authorial performance after 1890 was a sorry one, marked by repeated procrastination and inconstancy and by chronically over-optimistic expectations. The mantle of leadership that he had assumed on Jevons's death had proved a heavy one. Both temperamentally and by virtue of his acknowledged position as the doyen of British economists, Marshall was compelled to attempt the magisterial and to denigrate the kind of forceful direct essay of which he was eminently capable.

As Cambridge professor and unquestioned leader of British orthodox economists, Marshall could hardly avoid becoming a public figure whose pronouncements carried more than a personal weight. His consciousness of this, and of the precarious public standing of economics, as well as his own temperament, made him peculiarly reluctant to enter into public controversy, although he would on occasion fire off a letter to *The Times* on some issue of the day. He served as an expert witness for several government enquiries and was an influential member of the Royal Commission on Labour of 1890–94. As President of Section F of the British Association in 1890 he took the formal lead in the movement to found the British (later Royal) Economic Association, but he was not a prime mover. Indeed, he was not a clubbable or organizational man and relied on others to further whatever goals he desired for economics and the economics profession at large. But neither was he a recluse. Balliol Croft received a continuing stream of visitors, ranging from working class leaders to distinguished foreign economists, while students or young colleagues were always welcomed and offered generous advice mixed with exhortation.

Although able students interested in economics were in short supply, Marshall did over the years teach and influence several students who were to make contributions to the subject. From the early Cambridge

period H.S. Foxwell, H.H. Cunynghame, J.N. Keynes and J.S. Nicholson might be mentioned. The Oxford period brought L.L.F.R. Price and E.C.K. Gonner, while the period as Professor in Cambridge produced, among others, A. Berry, A.W. Flux, C.P. Sanger, A.L. Bowley, S.J. Chapman, A.C. Pigou, J. H. Clapham, D.H. Macgregor, C.R. Fay, and, last but not least, J.M. Keynes.

The undoubted fact of Marshall's professional leadership of British economics calls for some explanation. He was far from suited to such a role by temperament, and his fussiness and inflexibility could be irritating. For example, Sidgwick, J.N. Keynes, and Foxwell, the most important of his early allies in Cambridge, were all eventually alienated. Marshall's success can be attributed partly to sheer persistence. As in the case of the new Tripos, he had a clear idea of what he wanted to accomplish and worried away at it until he exhausted the opposition and was allowed to have his way. But it must also have been due to the lack of any alternative. The relevant question is not ‘Why Marshall?’ but ‘Who else?’ Economics was rapidly evolving as a profession around the turn of the 20th twentieth century, creating a leadership vacuum. Leadership was unlikely to emanate from outside Oxford, Cambridge or London, but F.Y. Edgeworth at Oxford was perhaps the last man capable of meeting the need, while E. Cannan at the new London School of Economics, although more suited than Marshall to the hurly-burly of professional politics, was too much the perennial critic and iconoclast to fill the bill. Moreover, whatever Marshall's foibles, the sheer power of his intellectual vision, his international standing as Britain's leading economic thinker, and his ability to inspire an impressive flow of budding scholars, all conspired to make him the only feasible contender.

Marshall's views on the social setting, aims and methods of economics

Marshall saw economics as concerned with those aspects of human behaviour open to pecuniary influences and sufficiently regular and ubiquitous to permit statements of broad scope and some persistence. While maintaining, especially in earlier work, that some heeded moral imperatives might be impervious to pecuniary considerations, he conceded that most behaviour lay within the ambit of the measuring rod of money. On the other hand, he emphasized that motivation was not merely a matter of pursuing pecuniary self-interest, even if broadly conceived to include interests of family and friends. He was anxious to lay the ghost of *homo economicus* and emphasized the human desires to obtain social approbation or distinction and to enjoy the pleasures of skilful activity. He saw actors as diverse as captains of industry and sculptors driven more by the joys of creative activity and the striving for the regard of peers than by the desire for material acquisition.

As well as not being pecuniary maximizers in any narrow sense, individuals were for the most part seen as imperfect optimizers. The working classes, especially, often lacked the knowledge and foresight to judge their long-term interests. Marshall's actors were not imbued with complete knowledge of their environment but had to acquire knowledge slowly, and often painfully, through experience. Nor were they endowed with fixed desires and an intrinsic, unchanging character. Indeed, character and preferences evolved as individuals were exposed to new possibilities and chose to enter into new activities. The workplace, in particular, was an important moulder of character. Self-improvement and character development induced by environmental changes, planned or unplanned, both figured largely in Marshall's world view. He believed that social institutions, such as land tenancy practices, were pliable and ultimately moulded themselves into conformity with the individual interests involved, rather than presenting a permanent constraint on mutually desired accommodations. (For this he was taken to task

by his most vehement critic, W. Cunningham, who denied the applicability of modern economic theory to medieval practices – see Cunningham, 1892.) But institutional change must be slow, slower even than changes in individual character and wants, because informal customs and tacit agreements are hard to change. Thus, while the institutions and informal understandings and prohibitions that constrain and mould economic behaviour might ultimately be endogenous they will often be ill adapted to current circumstances and thereby act as an independent constraint on the pursuit of mutually desired accommodations. Institutions, in the broad sense, are important and not always socially rational constraints on individual action.

Marshall was impelled to economics because ‘the study of the causes of poverty is the study of the causes of the degradation of a large part of mankind’ (1920, p. 3). For the bulk of the population, mired in poor living and working conditions, little progress in habits, aspirations and self-esteem could be expected without prior improvement in economic conditions. Such improvement was socially important not so much for its own sake, at least once the pangs of immediate want were assuaged, but because of its instrumental role in permitting and stimulating improvement in the quality and character of the population. What Marshall really valued was not improvement in the standard of living but the enhancement of the standard of life that this improvement made possible. And he entertained little doubt about what constituted a qualitative improvement here, even though – or perhaps because – his values may seem quite parochial and culture-bound.

Economic improvement required appropriate institutions, incentives and attitudes, and would be threatened by wide-scale government intrusions into economic affairs, although some forced income redistribution could be tolerated. But even if economic conditions were improved, the full yield of social betterment would be garnered only if enlarged consumption were turned to ennobling and horizon-expanding channels (rather than, say, to strong drink), involved a due consumption of beneficial leisure, and was accompanied by healthier and less stultifying conditions of working and town life. The government had a guiding role to play here. But even more important would be the assistance and example of employers and the upper and middle classes, who must first rid themselves of a frequent propensity to showy and ostentatious consumption and excessive materialism. The working-class leaders and skilled artisans who had already raised their own standard of life had an important leadership role too. Voluntary individual efforts to assist the rise of the underprivileged must rest on an adequate understanding of economic consequences. For this, as well as to secure an informed electorate, the diffusion of sound economic knowledge was an essential and integral element in the process of socio-economic transformation. Economics thus was itself a noble activity of high importance for the future of mankind.

The broad view of the economy suggested by the foregoing is of a complex evolutionary process of combined economic, social and individual change in which each individual's abilities, character, preferences and knowledge develop jointly, along with social institutions, markets and the technologies of production and communication. The pursuit of self-interest, broadly conceived, is ubiquitous in directing this evolutionary process, but is subject to inertia, ignorance and limited foresight, not to mention individual mutability.

Unfortunately, Marshall was able to bring little formal analysis to bear on this general ‘biological’ vision of the economy and could only evoke it descriptively. It might be true that ‘the Mecca of the economist lies in economic biology rather than in economic dynamics’ (1920, p. xiv). Nevertheless, the only available analytical tools were those of classical mechanics, tools that Marshall's early mathematical

training had equipped him to employ skillfully. In fact, chief reliance had to be put on that branch of classical mechanics dealing with statics. Dynamics, beyond a few qualitative applications, required more precise information than was likely to be available. Perforce then, much of Marshall's formal analysis, like that of W.S. Jevons or Léon Walras, was based on simple assumptions of individual optimization and market equilibrium, taking preferences, technology and market institutions for granted. Such provisional or tentative ‘statical’ treatments could often be valuable. Indeed Marshall viewed them as indispensable for the correct analysis of many questions. But he was always anxious to stress that the analysis was preliminary, and perhaps of only transitory validity. This awareness made him impatient of over-elaboration, so that, for example, he showed no interest at all in pushing the statical approach to its logical conclusion in the general equilibrium analysis of the stationary state. For him, equilibrium analysis was an indispensable but rough and ready instrument that needed to be employed with due caution and a continuing awareness of its limitations in the face of a complex ever-evolving reality. It was only a tool and did not itself constitute concrete knowledge.

Marshall had no great profundity as a philosopher of science and had little patience with metaphysics: ‘in a sense … he held no views on method’ (Coase, 1975, p. 27). Marshall's discussions of methodology largely reflect the philosophical presuppositions of his day. His method was in the general deductive tradition of John Stuart Mill, but he sought to emphasize the relativity of particular theories, as contrasted with the universality of the general theoretical ‘organon’ or economists’ toolbox. Anxious to present a public image of the unity of economics in the face of the Methodenstreit among economists in the late 19th century, he attempted to maintain an uneasy balance on method, decrying extended chains of deductive reasoning but denying the possibility of purely inductive inference unguided by a coherent conceptual framework. Economics had room for specialists in both deductive and inductive methods, but both must ultimately be co-workers. Assumptions must be selected with close regard to the facts of the case and potential disturbing causes must be kept prominently in mind and due allowance made for them. J.N. Keynes described Marshall's analytical method as ‘deductive political economy guided by observation’ (1891, p. 217n) and Keynes's chapter ‘On the Deductive Method in Political Economy’ (1891, pp. 204–35) is perhaps as good a rationalization of Marshall's method as one can find.

Intellectual debts

The intellectual background to Marshall's work in economics was established in the 1860s, partly in his stringent mathematical training, but perhaps more importantly in the heady mixture of utilitarianism, evolutionism and German idealism which he eagerly imbibed in the years immediately following his graduation. He seems to have started on economics from J.S. Mill's *Principles of Political Economy* (1848), moving on to the classic works of Smith and Ricardo. At a fairly early stage, probably around 1868, he discovered Cournot's *Récherches* (1838), which provided examples of the application of mathematics to economic questions. Acquaintance with J.H. von Thünen's work, which influenced Marshall's distribution theory, must have come somewhat later, in the early to mid-1870s. During the 1870s and early 1880s Marshall also read widely on economic development and socialism, including much literature in German, the only foreign language he mastered thoroughly. After that, his reading seems to have been concentrated mainly on factual and practical matters. Once his own theoretical views had crystallized, he appears to have been reluctant to do more than attempt to explain and clarify them to

others, and to have taken remarkably little interest in new theoretical issues or in the theoretical ideas of others.

In many ways, the list of Marshall's denials of theoretical indebtedness is more remarkable than that of his acknowledgments. He claimed to have developed his ideas on consumer surplus before learning of anticipations by J. Dupuit and H. Fleeming Jenkin. The grudging attitude to W.S. Jevons's marginal utility theory shown in his review (1872) of Jevons (1871), although subsequently relaxed, was never replaced by any acknowledgement of indebtedness. He showed little or no interest in the work of Walras, gave meagre credit to Carl Menger, whose work must have become known to him by the early 1880s, patronized Pantaleoni and Böhm-Bawerk, largely ignored Pareto, and so on. Even in the case of Edgeworth, one of his few intimates, Marshall felt that undoubted theoretical powers were guided by an unreliable judgement and refused to follow Edgeworth's subtle elaborations far. In fact, the only major theorist of the day to command Marshall's entire admiration and respect was J.B. Clark, and even here there was no acknowledgement of serious indebtedness. This tendency to denigrate the work of his contemporaries was matched by an equally strong tendency to overvalue the achievements of the British Classical School led by A. Smith, D. Ricardo and J.S. Mill. For one reason or another – perhaps a personality quirk, perhaps an effort to boost the public esteem of economics – Marshall was prone to exaggerate the intellectual continuity and maturity of his subject – see O'Brien (1990) on this.

A growing interest in wider intellectual influences on Marshall in his formative years 1865–70 has been sparked by the publication and analysis of his early philosophical manuscripts (Raffaelli, 1994, 2003), especially a paper entitled ‘Ye Machine’ that outlines a mechanism capable of learning new routines from experience, thus freeing its limited learning ability to gradually establish new and higher level routines, and so on. It appears that Marshall's ambitions in these early years lay in the area of ‘psychology’ or perhaps better in the ‘philosophy of mind’. Whether the world lost more than economics gained from his switch to economics remains an open and perhaps insoluble question. But it does appear that the pattern of a sequential routinizing of new methods, continually leading to new levels of individual or organizational complexity, continued to play a significant part in Marshall's economic thought. More generally it is clear that he read philosophical literature widely in his formative years: Kant, Hegel, H. Spencer, and others. But whether and how these sources influenced his economic thought remains uncertain, partly because evidence is slight or absent.

Demand theory

So far the discussion has remained on a very general level, dealing with broad aspects of Marshall's life and work. At this point there begins a much more detailed and technical consideration of various aspects of his theoretical contributions, commencing with his demand theory. Marshall's treatment of the theory of demand is sketchy and incomplete, concentrating on the demand for a single commodity, or commodity group, against a loosely defined background. A utility-maximizing individual's utility is defined by $u(x) + w(y)$ where x is the individual's consumption of the particular good X, while y is the individual's expenditure on all other goods. This expenditure is measured in money of constant purchasing power: that is, deflated by a general price index. How this index is defined and whether, as seems appropriate, the price of X is excluded from it, is left unclear. Such money can be treated as a composite good, Y, and y can be regarded as the amount of this composite good consumed. If m is the

individual's initial endowment of Y, then $y = m$ whenever $x = 0$, while if X can be freely purchased at a fixed price of *p* units of Y per unit of X then *x* and *y* must satisfy the constraint $px + y = m$. Marshall assumes that the utility functions *u*(*x*) and *w*(*y*) have positive but diminishing marginal utility so that $u'(x) > 0 > u''(x)$ and $w'(y) > 0 > w''(y)$, where single and double primes are used to denote first and second derivatives. The maximum expenditure, *e*, that the individual is willing to make to secure *x* units of X is implicitly defined as a function *e*(*x*, *m*) by $u(x) + w(m - e) - w(m) = 0$. Providing that *x* and *y* are both positive, the rate at which *e* increases with *x* is $u'(x) / w'(m - e)$ by the implicit function theorem. This ratio would be the demand price for the *x*th unit of X if all previous units had been acquired at their corresponding demand prices: that is, if the individual had faced perfect price discrimination in exchanging Y for X. Alternatively, if the individual had been able to obtain any amount of X at fixed per unit price, *p*, the resulting demand function *x*(*p*, *m*) for X would be implicitly defined (given *x* and *y* are both positive) by the first-order condition $u'(x) - pw'(m - px) = 0$. Partial differentiation of *x*(*p*, *m*) shows that *x* falls as *p* increases, while an increase in *m* increases both *x* and *y*: thus, the Giffen possibility of an increase in *p* increasing the quantity of X demanded is excluded. But an increase in *p* may lower or raise the value of *px*, so that demand for X may be price elastic or price inelastic at a given *p*.

The possibility of buying at a fixed price rather than facing perfect price discrimination creates a consumer surplus of $e(x(p, m), m) - px(p, m)$. This is the additional amount that could have been extracted by perfect price discrimination for all units up to the price-taking optimal one. That this surplus is positive follows from the fact that every infra-marginal unit of X acquired creates a surplus utility when the individual faces a fixed price (since $u'(x) > pw'(m - px)$ for each such *x*) but no surplus when the individual is faced with perfect price discrimination.

Marshall's mathematical notes (1920, pp. 838–42) on his general case are obscure and puzzling. Doubtless he felt this case was too dependent on unobservables to be of much practical value. He therefore emphasized the special case in which the marginal utility of money is treated as a constant. The rationale offered is that an individual's ‘expenditure on any one thing … is only a small part of his whole expenditure’ (1920, p. 842). This simplifies $e = e(x, m)$ above to $e = u(x) / w'(m)$ while *x*(*p*, *m*) is now defined implicitly by $u'(x) / w'(m) = p$. At the *x* value defined by the latter equation, consumer surplus arising from the ability to buy any amount of X at the per-unit price *p* can be expressed in utility terms as $u(x) - xu'(x)$ or in money terms as $u(x) / w'(m) - xu'(x) / w'(m)$. These formulae are exactly analogous to the standard formula for Ricardian land rent, with the first term the output obtained on a piece of land from the application of *x* doses of variable input, each dose remunerated at the common marginal product. Partly because of this analogy, Marshall used the term ‘consumer rent’ rather than ‘consumer surplus’ prior to 1898.

Although priority must go to Dupuit, Marshall's simple concept of consumer surplus based on the assumption of a constant marginal utility of money has been influential. But he was well aware of the complications arising from variation in the marginal utility of money: ‘Strictly speaking we ought to take account of the fact that if he spent less on tea the marginal utility of money to him would be less than it is, and he would get an element of consumers’ surplus from buying other things at prices which now yield him no such rent’ (1920, p.842). Although such influences may be ‘of the second order of

smallness' they raise the more disquieting issue of assessing the overall welfare effects of changes that affect many markets simultaneously. On this Marshall had little to say: 'the task of adding together the total utilities of all commodities, so as to obtain the aggregate of the total utility of all wealth, is beyond the range of any but the most elaborate mathematical formulae' (1920, p. 131n.). It was a task he chose not to pursue. Apart from generalizing for the possibility that a certain quantity of good X might be indispensable, Marshall elected not to develop his demand theory further, or even to generalize it to incorporate utility functions that were not additively separable (1920, p. 845). It is clear that each commodity in turn might take the spotlighted role of good X and that in certain circumstances simultaneous consumer surpluses for several goods might be added (1920, p. 842). An unpublished early manuscript note from the 1870s on the theory of taxation (Whitaker, 1975, vol. 2, pp. 285–305) had advanced matters considerably further by working formally with the maximization of utility under a budget constraint, but this lead was not followed up in print and some of its lessons for welfare economics were apparently forgotten. *Principles* gave a clear intuitive account of the consumer's overall optimization problem (1920, pp. 117–23), but failed to connect it to the resulting interrelated set of demand functions for the various goods consumed. Indeed, it is clear that for positive purposes Marshall was willing to treat market demand functions in a quite pragmatic way, admitting, for example, close substitutes or complements and the Giffen exception, all inconsistent with the simple formal theory set out above. In judging this, it must be borne in mind that consistency and generality of 'statical' analysis were not Marshall's real goal. Rather, 'fragmentary statical hypotheses are used as temporary auxiliaries to dynamical – or rather biological – conceptions' (1920, p. xv).

The market demand for a good that is offered to all actual or potential buyers at the same given price is of course obtained as a function of that price by summing the amounts demanded at that price by all the consumers. A sufficient but not necessary condition for market demand to fall as price increases is that each individual's demand decreases. The now familiar concept of market demand elasticity – proportional quantity change divided by proportional price change – was first introduced by Marshall, although several authors had come close to the idea previously. It appeared without flourish in (1885c), and appeared more prominently in *Principles*. But Marshall himself made relatively little use of it.

Bibliographic note: Marshall's treatment of demand is essentially contained in (1920, pp. 92–137, 838–43). An influential, although controversial, interpretation of Marshall's demand theory is given by Friedman (1949). Biswas (1977) gives another alternative to the orthodox reading provided by Stigler (1950) that is largely adopted here. An excellent overview is Aldrich (1996). On consumer surplus see Chipman (1990).

Production and long-period competitive supply

In deriving the long-period supply curve of a commodity in *Principles*, Marshall envisages production as organized by firms, typically family businesses. Each firm strives to minimize its production costs, substituting one productive factor or production method for another according to the 'Principle of Substitution'. In its simpler forms this involves marginalist adjustment to bring relative marginal value products into line with relative marginal costs. But more generally, the Principle of Substitution is akin to a natural selection process, being 'a special and limited application of the law of survival of the fittest' (1920, p. 597). Marshall's firms do not have costless access to a common production function, but

must grope and experiment their way to cost-reducing modifications. The long-period supply curve is defined for a given state of general scientific and technical knowledge. But each firm must explore this to some extent anew.

Although the distinction is not entirely clear – distinctions seldom are for Marshall – two polar cases may be distinguished within his theory of long-period competitive supply. These will be referred to as the ‘agricultural’ and the ‘industrial’ cases. The former is much the more straightforward and involves an industry in which production is relatively simple, internal economies of scale are minimal, and the product is homogeneous and easily marketed. The optimal firm size is small, and management is sufficiently routine to need no exceptional ability to keep a firm operating efficiently. As the overall market expands, new firms may be added, but changing composition of the population of firms is not an essential feature of this case.

The long-period supply price per unit of output at which such an industry can supply any quantity of output must just cover the cost of maintaining that level of output indefinitely. That is, it must just suffice to pay all the inputs (including management) needed to produce that level of output in a cost-minimizing way at rates that just ensure that the requisite input quantities will continue to be forthcoming indefinitely. In the case of skilled workers, in particular, the rate must just suffice to induce parents to apprentice new workers to the industry at a rate exactly offsetting the attrition through retirement and other causes. Similarly, the return to fixed capital must just suffice to induce replacement of the existing stock of fixed assets, while the return to management must keep up the necessary replacement flow of managers. On the other hand, the return to land services must just suffice to prevent these services from migrating elsewhere, replacement not being necessary. As the level of industry output being considered is increased, the supply price will probably rise, mainly because of the need to pay a higher return to land so as to attract a greater supply from other uses, but perhaps also because of the need to pay more for rare natural talents that, like land, must be attracted in greater quantity from other uses, not being capable of replication through education and training. Such a tendency for long-period supply price to rise with output may be mitigated though seldom eliminated by substitution against inputs whose supply price is rising, and by possible external economies that increase each firm's efficiency by influences that depend, not on its own output, but on the entire industry's output. A tendency for supply price to rise with output will imply that infra-marginal units of those inputs whose supply prices are rising receive rents, since all units will be remunerated at the rate necessary to induce continuing supply of the marginal unit. In the absence of external economies (or diseconomies) the total rent or producer surplus generated will be the ‘triangular’ area above the supply curve. That is, it will be

$$R = xg(x) - \int_0^x g(v)\mathrm{d}v$$

where $g(x)$ is the supply price of output quantity x, an increasing function of x. This result does not apply in the presence of external economies. In later editions of the *Principles*, Marshall introduced the device of the ‘particular expenses curve’ (1920, pp. 810–12) to display rent in such a case, but this *ex post* construction does not give an independent basis for determining rent.

It is clear that the long-period supply curve of an industry depends on the general economic background against which the industry is assumed to operate. As is the case with demand, Marshall does not consider this background in detail. He assumes prices to be expressed in money of constant purchasing power and recognizes on occasion that there may be close interrelations between two industries (for example, they may compete for the same specialized land). He also recognizes that ‘a theoretically perfect long period must give time enough to enable not only the factors of production of the commodity to be adjusted to the demand, but also for the factors of production of those factors of production to be adjusted and so on’ (1920, p. 379n.), and that this leads ultimately to the assumption of a stationary state. But he is not willing to follow this route far and is content in general to take the supply conditions of the factors of production for granted when analyzing long-period price determination.

In the ‘manufacturing’ case, to which we now turn, the product is differentiated, marketing is difficult, and each firm must build up and retain goodwill and a customer connection for its own specialized product. There are substantial internal economies of scale in production and successful management calls for business ability of a high and rare character. In this environment, a family business may be built up by an exceptional founder, but this build-up must be slow because of the difficulty of establishing a market and perhaps also because of constraints on financing. And when the founder passes on, his successors are unlikely to have equal talents or even the lesser talents required to prevent the firm's business from languishing. By the third generation of succession, the firm is likely to expire. Even a joint stock company (a case added rather as an afterthought) is likely to ossify into bureaucratic stagnation, and presumably the same is true of family businesses that rely on paid managers. Thus, the typical firm in the manufacturing case passes through a finite life cycle, and the industry is comprised of a population of such firms at various phases of the life cycle, some in the early expanding phase, others in decline.

The long-period supply price at which such an industry can supply a specified level of output must now be regarded as an index of the prices of all the different firms’ products. It must meet all the conditions required in the agricultural case. Thus, the price must allow for a continuing replacement flow of the various types of workers (including managers) and fixed assets, as well as the retention of the necessary ‘land’ services. But now there must also be a surplus sufficient to induce a replacement flow of new firms – a supply of ‘business organization’ that will just suffice to replace the expiring firms and keep the age distribution of firms constant.

Industry equilibrium does not require each firm to be in an unchanging equilibrium any more than the trees in the proverbial forest. A new firm will be established if the prospective earnings over the expected life cycle appear to justify the cost and trouble involved. The firm's initial earnings are likely to be negative as it slowly builds up its technical expertise and market connections, but these early losses can be regarded as investments to be recouped in the later stages of the firm's prospective life cycle.

It is here that Marshall's ‘representative firm’ enters the picture. It is best regarded as a parable that avoids the need to consider the entire distribution of firms. By definition, the long-period supply price of any level of industry output is the average cost of the representative firm at that level of output. Industry-level magnitudes may then be regarded as if they were generated by a fixed number of unchanging representative firms rather than by the actual heterogeneous body of ever-changing firms. In other words, the manufacturing case may be treated as if it were an agricultural case populated by representative firms only. Such arguments add nothing conceptually and are prone to confuse, although it might be noted that Marshall believed an acute well-informed observer could select an actual firm that

was close to being representative in this sense.

The average cost and size of the representative firm will change as industry output changes. There are two main reasons for this. A larger industry output is likely to generate more external economies, lowering the costs of every firm. But more importantly, the larger industry demand is, the easier it will be for a new firm to build up a market, and so the larger the size to which firms will grow before they begin to decline. This will bring about greater access on average to unexhausted internal economies of scale, again leading to lower costs on average. For both these reasons, long-period supply price is likely to decline as a larger industry output is considered, even though the opportunity cost of obtaining greater supplies of land services and rare natural talents may rise. Again, the particular expenses curve may be used to display the producer surpluses or rents accruing to such scarce factors at any given level of industry output, but the relationship of this family of curves to the long-period supply curve is tenuous and complex. Rent obviously cannot be represented by a ‘welfare triangle’ above the supply curve when the latter is falling.

The conception of competition in Marshall's manufacturing case is much closer to later ideas of imperfect or monopolistic competition than to modern notions of perfect competition. Products are differentiated and firms are not price takers, but face at any time downward-sloping demand curves in their special markets. Even if the difficulties of rapidly building up a firm's internal organization can be overcome, the resulting enlarged output cannot be sold at a price covering cost – even granted substantial scale economies in production – without going through the slow process of building up a clientele and shifting the firm's particular demand curve. The time this takes is assumed to be considerable relative to the duration of the firm's initial vitality. But in some cases the difficulties of rapid expansion may be overcome. They may not have been very severe, as when different firms’ products are highly substitutable, or the firm's founder may have unusual genius. In such cases the industry will pass into a monopoly or be dominated by a few, strategically interacting firms, or ‘conditional monopolies’ as Marshall termed them.

Marshall's reconciliation of persisting competition with increasing returns and falling supply price is complex and problematic, but it does not depend in any essential way on scale economies being external to the firm. The concept of external economies is one of his significant contributions, although his treatment of it can hardly be called pellucid. But it was added more for verisimilitude than because it was theoretically essential to the structure of his theory.

The issues surrounding Marshall's representative firm, and the problem of reconciling the persistence of competition with the presence of unexhausted internal economies of scale, continue to receive attention among historians of economic thought but no definitive reading has yet been attained, or perhaps ever will be. The account given above is well supported by Marshall's text, but as is often the case with Marshall, elements of ambiguity and vagueness remain.

Bibliographic note: Marshall's treatment of long-period competitive supply is to be found in (1920, pp. 314–22, 337–80, 455–61, 805–12) and (1919, pp. 178–96). The earliest version, dating from the early 1870s is reproduced in Whitaker (1975, vol. 1, pp. 119–59) and see also (1879a). Key early commentaries and criticisms of Marshall's theory of supply are Sraffa (1926), Robbins (1928), D.H. Robertson, Sraffa and Shove (1930), Viner (1931), Frisch (1950), Hague (1958) and Newman (1960).

Price determination and period analysis

The long-period supply curve for any good indicates for each market quantity the least price at which that quantity will continue to be supplied indefinitely. The long-period equilibrium price and quantity are determined by the intersection of this supply curve with the market demand curve, assumed to be negatively sloped, that indicates the highest uniform price at which any total quantity can be sold. In the agricultural case, equilibrium will be unique as the supply curve slopes positively. But in the manufacturing case, the supply curve, as well as the demand curve, may have negative slope, so that multiple equilibria can occur. Equilibrium is adjudged locally stable if demand price is above (below) supply price at a quantity just below (above) the equilibrium quantity. The intuitive justification for this is that the actual price of any available quantity is determined by the demand price, while quantity produced tends to increase (through both expansion of existing firms and entry of new firms) whenever an excess of market price over supply price promises high profits, while it tends to decrease in the opposite case.

This stability argument is sketchy and, in any case, there still remains the question of exactly how a new long-period equilibrium is attained following some change, such as a permanent shift in the demand curve. One possibility would be to consider explicitly the adjustment process through time, but Marshall preferred to approach the problem by another route – his period analysis, one of his most memorable and lasting contributions. (His passing claim (1920, p. 808) that the long-period supply curve may not be reversible, supply price depending upon past-peak output as well as current output, is something of an exception to this generalization. It appears to rest on some restriction of the degree of downward supply adjustment, and so not to involve a true long-period analysis, or else to invoke a kind of learning by doing that once attained is not readily lost.)

Period analysis is Marshall's most explicit and self-conscious application of the comparative-static, partial-equilibrium method with which his name will always be associated. As he observed,

> the most important among the many uses of this method is to classify forces with reference to the time which they require for their work; and to impound in *Caeteris Paribus* those forces which are of minor importance relatively to the particular time we have in view. (Guillebaud, 1961, vol. 2, p. 67)

Which forces or variables are to be hypothetically frozen or impounded, and which are to be determined by the requirements of equilibrium (an equilibrium contingent upon the contents of the *ceteris paribus* pound, of course), should be determined pragmatically in each case with the aim of focusing on the features deemed dominant in that case. As a general rule, those forces should be impounded which move very slowly, or else bounce around very rapidly, relative to the length of ‘the particular time we have in view’. This is well illustrated by Marshall's example of a fish market, where the focus may be on the determinants of price over a few days, a few months, or several years, or even decades (1920, pp. 369–71). As an expositional matter, however, and also to embody distinctions of wide (but not universal) applicability, Marshall emphasized three broad cases. Temporary or market equilibrium analysis proceeded on the assumption of a fixed stock of output already available or in the pipeline. Short-period normal equilibrium analysis permitted output to be varied, but not the stock of productive ‘appliances’ available to produce that output. ‘Appliances’ must be taken here to cover skilled labour and business organization as well as fixed capital assets, so that the existing set of firms is to be taken as given. Finally, long-period normal equilibrium, which has already been considered, allows the stock of

appliances, as well as the level of output, to be freely varied. In this case equilibrium incorporates the conditions necessary for inducing an exact replacement flow of each kind of appliance, including a replacement flow of new firms in the manufacturing case.

Temporary equilibrium for a perishable commodity is simply a matter of selling off the existing stock. Marshall recognizes the possibility of ‘false trading’ – sales at a non-equilibrium price – but argues that (*a*) this will not affect the eventual price if the marginal utility of money is constant, and (*b*) price will quickly settle close to the uniform price that would just clear the market if used in all transactions. With a storable good there is the further speculative possibility of holding back supply for future sale, and this gives expected future cost of production an indirect role in influencing current market price. Cost of production already incurred is an irrelevant bygone, however.

In short-period normal equilibrium, output is adapted to demand within the constraints set by the fixed supply of available ‘appliances’. High demand will raise equilibrium output, but only within the limits possible by working existing appliances more intensively or pulling in versatile unspecialized labour and land from elsewhere. Low demand will lead to low utilization of appliances, perhaps idleness of some, and migration of unspecialized inputs to elsewhere. In the agricultural case a firm will change output until marginal prime or variable cost equals market price. In the manufacturing case, a fear of spoiling the future market or invoking retaliation from competitors tends to make a firm's output more responsive to variation in market price, and hence to make market price less responsive to demand shifts. Otherwise, the two cases are similar, both involving a fixed population of firms and a rising supply curve.

The return received by an appliance will often exceed the minimum necessary to induce its operation at the chosen intensity (its prime cost) and this excess is a ‘quasi rent’. To the extent that land and rare natural talents are immobile in the short period, or less mobile in the short period than the long, their returns too will often have a quasi-rent element. Otherwise, they will receive only differential rents, though often at rates differing from their long-period values. It should be stressed that the concepts of quasi-rent and differential rent are relative to a specific use. The prime cost necessary to retain an input in this use may itself include rent or quasi-rent when viewed in the context of a more inclusive set of alternative uses. Thus, from the viewpoint of all possible uses in the economy, the return to any factor in fixed supply is entirely a rent or quasi-rent (the latter if fixity is only short-period).

Marshall paid little attention to the possibility that forces similar to those constraining the adjustment of supply when time is limited might also operate on the side of demand. Thus the same considerations underlie the market demand curve whether it is coupled with a temporary, short-period or long-period supply curve. In each case, market equilibrium price and quantity are determined by the intersection of the appropriate demand and supply curve. The stability of temporary equilibrium is directly asserted. The stability of short-period equilibrium depends on the same quantity-adjustment argument invoked for long-period equilibrium, but since the short-period supply curve is always positively sloped, uniqueness and stability are assured.

The theory of short-period normal equilibrium was designed as a tool for analysing unemployment and economic fluctuations in the never-completed second volume of *Principles*. But it also has use in explaining adjustment to a permanent disturbance. Suppose, for instance, that an industry is in long-period equilibrium when a permanent shift in demand occurs. The immediate or short-period effects can be analysed by freezing output and stocks of appliances at their initial levels. Insight into the actual adjustment through time can then be obtained by appropriately changing the output level assumed in the

temporary equilibrium, so that movement of temporary equilibrium towards short-period equilibrium can be traced out as output , but not stocks of appliances, adjusts. Similarly, the levels assumed for the stocks of appliances in this short-period equilibrium can be allowed to change and the movement of short-period equilibrium towards long-period equilibrium traced out. Such arguments are now a staple of elementary pedagogy. They clearly require additional assumptions about the adjustment of output and the way in which investment or disinvestment in appliances proceeds, and are only a poor and ambiguous substitute for an explicit dynamic analysis. But such ‘statical’ procedures, although imperfect, may, in Marshall's words, be ‘the first step towards a provisional and partial solution in problems so complex that a complete dynamical solution is beyond our attainment’ (Pigou, 1925, p. 312).

Marshall's period analysis, and more generally his partial-equilibrium approach to price determination, was designed in large part as a usable tool for the analysis of concrete issues. Its longevity amply testifies to its usefulness in this respect. But it was also meant to serve the more doctrinal purpose of clarifying the respective roles utility and cost of production play in determining value. The aim was to show that the greater the scope for supply adjustment permitted in the definition of equilibrium, the more dominant the supply side influence on price becomes. This doctrinal goal helps to account for the rather heavy weight given to long-period analysis in *Principles*. For, as Marshall recognized, its value as a tool of applied analysis is seriously qualified by the fact that ‘violence is required for keeping broad forces in the pound of Caeteris Paribus during, say, a whole generation, on the ground that they have only an indirect bearing on the question in hand’ (1920, p. 379n). That is, there is no good ground for assuming that background forces such as technology and tastes will remain constant for the length of time required for long-period equilibrium to be practically relevant. For concrete analysis of problems of such long duration it will often be necessary to transcend the period analysis, with its reliance on statical equilibrium, and undertake directly an analysis of secular change, of which Book 6, Ch. 12 of *Principles* on the ‘General Influence of Economic Progress’ (1920, pp. 668–88) offers the main example, but not a very impressive one.

In emphasizing the role that cost of production plays in the determination of long-period value, Marshall was not content to rest on money costs of production but sought to go behind these costs to the real costs – the efforts and abstinences – for which in a non-coercive economy the money costs are recompense. In doing so he purported to follow Ricardian tradition, but is more plausibly viewed as attempting to place the newer subjective value theories in broader (but still subjective) focus. Just as the price paid by a consumer serves as a measure of marginal utility, with a consumer surplus gained on infra-marginal units, so the unit price received by a worker or saver measures the real cost or disutility at the margin, with a producer surplus on the inframarginal units of effort or abstinence. But, as Marshall recognized, the parallel holds imperfectly in the long period when workers must be regarded as produced means of production as well as final consumers and cost bearers. In particular, parental sacrifice for raising and training offspring obtains little or no direct pecuniary reward.

Bibliographic note: Marshall's treatment of period analysis is concentrated in (1920, pp. 363–80) but see Whitaker (1975, vol. I, pp. 119–59) for the earliest version. For commentary and exposition see especially Viner (1931), Opie (1931), Frisch (1950), Whitaker (1982). On temporary equilibrium see (1920, pp. 331–6, 791–3, 844–5) and Walker (1969). On short-period normal value see Gee (1983).

Normal value and normal profit

Implicit in the preceding discussion are Marshall's conceptions of normal value and normal profit. Normal value is defined as the value that would result ‘if the economic conditions under view had time to work out undisturbed their full effect’ (1920, p. vii). It is contrasted with market value, which is ‘the actual value at any time’ (1920, p. 349). Normal value is hypothetical, resting on a *ceteris paribus* condition, its role being to indicate underlying tendencies. The normal value of a commodity may approximate its average value over periods sufficiently long for the ‘fitful and irregular causes’ (1920, pp. 349–50) that dominate market value to cancel out, but this should not be presupposed automatically outside a hypothetical stationary state.

The distinction between normal and market value is closely related to the distinction between natural and market value found in the work of Smith and the classical economists. In 1879 Marshall had identified normal value with ‘the results which competition would bring about in the long run’ (1879b, p. vii), but in *Principles* he switched to the view that ‘Normal does not mean Competitive’ (1920, p. 347) and admitted any kind of regular influence so long as it was sufficiently persistent. The economic forces hypothetically permitted to achieve full mutual accommodation could now be chosen appropriately for each case. In particular, the distinction between short-period and long-period normal (or ‘sub-normal’ and ‘true-normal’ in earlier editions) was emphasized.

Profit was viewed by Marshall as the residual income accruing to a firm's owner, a return on the investment of the owner's own capital and recompense for the pains of exercising ‘business power’ in planning, supervision and control. Normal profit is essentially an opportunity cost, the minimum return necessary to secure the owner's inputs to their current use, or rather to accomplish this for an owner of normal ability. Marshall presumes that there is a large and elastic supply of versatile actual or potential owner managers of normal ability. In long-period equilibrium each of these must just receive the same normal rates of return on investment and exercise of business power whatever the line of business. However, those who are exceptional may do better, essentially by exerting greater business power. These common rates of normal return are simultaneously determined, along with the normal returns to other kinds of effort and abstinence, by Marshall's macroeconomic theory of the long-period determination of factor incomes (see below). Although it is the case that profits are a residual, rather than a contractually agreed amount like other incomes, this difference is immaterial in long-period equilibrium. In particular, a long-period equilibrium analogy between ordinary wages and the normal earnings of business power is stressed. Normal profit is a necessary element in the costs that underlie the long-period normal supply curve, but actual profit is a quasi-rent or producer surplus for shorter periods. Normal profits are a return to ‘business power in command of capital’ and compensate for three distinct elements: ‘the supply of capital, the supply of the business power to manage it, and the supply of the organization by which the two are brought together and made effective for production’ (1920, p. 596). The combined compensation of the latter two components comprises ‘gross earnings of management’, the return to the second component being ‘net earnings of management’. In long-period equilibrium, the normal return to the first element is imputed at the market interest rate on default-free loans, and that to the second component at the rate paid to hired managers performing comparable tasks. The residual third element, the return to ‘organization’, is most straightforwardly interpreted as an extra return on owned capital equivalent to the premium for default risk, or ‘personal risk’, that would have to be paid on borrowed capital. In the manufacturing case, the annual level of normal profit for each firm in an

industry must be interpreted as the annualized equivalent of the expected stream of returns that is just sufficient to induce an individual of normal ability to found a firm in the industry rather than divert energies and capital elsewhere. Normal ability here is defined relative to other potential founders of firms, a group already exceptional relative to the population as a whole. By construction, such normal profits must be earned by the representative firm.

Bibliographic note: The most pertinent commentary is Frisch (1950). For Marshall's views on normal value see (1879b, pp. v–vii, 65–71, 146–9; 1920, pp. vii, 33–6, 337–50, 363–80). For his views on normal profit see (1879b, pp. 135–45; 1920, pp. 73–4, 291–313, 596–628). For the role of ‘personal risk’ see Guillebaud (1961, vol. 2, p. 672).

Welfare economics

To serve as a tool of welfare economics, monetary measures of consumer surplus, producer surplus and rent must be aggregated over individuals. But how are the resulting sums to be interpreted? Marshall's very limited and proximate attempts at formal welfare arguments are carried out within a utilitarian framework, for which the goal is maximizing aggregate utility. He implies that interpersonal utility comparisons are possible in principle and that utility functions will be similar for all members of any group that is homogeneous in terms of mental, physical and social attributes. Within such a group, the marginal utility of money will be the same for two individuals having the same income, and lower for the richer of two individuals with differing incomes, on the assumption in each case that both individuals face the same trading opportunities. A postulated government action may impose gains and losses on various individuals that can be measured and aggregated in money-equivalent terms. But how can these measures be translated into statements about aggregate gains and losses of utility? Marshall emphasizes two special cases. First, if the gains and losses are both proportionately distributed over income classes in exactly the same way, then net aggregate gain (positive or negative) in money will serve as an ordinal index for the net aggregate gain in utility. A corollary of this is that if two alternative actions affecting the same group have the same relative distributions of gains and losses over income classes then the alternative yielding the greater net aggregate gain in money must have the greater net aggregate gain in utility. Second, if some change makes for a zero net aggregate change in money terms, but the gains accrue to individuals of lower income than those bearing the costs, then the aggregate net utility gain must be positive – a warrant for certain redistributive policies. In other cases he sees that careful assessments of the marginal utility of money to the various injured and benefited groups would be needed, assessments that could be used to transform monetary gains and losses into utility measures. He toys (1920, pp. 135, 842–3) with using the Bernoulli hypothesis on the relation between wealth and utility as a basis for such calculations, but gives little indication as to how assessments might be made in practice.

Marshall's best known and most successful foray into formal welfare analysis was his proof that total welfare might be increased by using the proceeds of a tax on an ‘agricultural’ industry to subsidize a ‘manufacturing’ industry. All comparisons involved long-period equilibria and relied on the validity of aggregated money-equivalent measures of gains and losses. He demonstrated that the gain in consumer surplus in the expanded decreasing-cost manufacturing industry might exceed the combined loss in consumer and producer surplus in the contracted increasing-cost agricultural industry. No formal account was taken of a possible gain in producer surplus in the manufacturing industry as this merely

makes the argument hold *a fortiori*. The crucial point in this argument, as Marshall recognized, is that producers are not harmed by ‘a fall in price which results from improvements in industrial organization’ (1920, p. 472). It is immaterial whether the improved organization of the enlarged manufacturing industry is due to external economies or to internal economies resulting from an increase in the size of the representative firm. Contrary to much subsequent opinion, Marshall's tax-subsidy argument is not necessarily dependent upon external economies.

Another significant, but overlooked, welfare analysis provided by Marshall was that of a monopolistic public enterprise in a situation where taxation involves an excess burden (1920, pp. 487–93, 857–8). In the absence of this excess burden Marshall proposes that the enterprise seek to maximize ‘total benefit’, the sum of net profit and consumer surplus. This implies marginal cost pricing, since the area below the demand curve and above the marginal cost curve is maximized when the two curves intersect. But, given that taxation involves an excess burden, it may be desirable to augment tax receipts from monopoly revenue if the sacrifice of consumer surplus is small. Marshall proposes the alternative goal of maximizing ‘compromise benefit’, the sum of consumer surplus and monopoly revenue when the latter is in effect multiplied by the marginal cost of raising a unit of government revenue from other sources. Maximization of compromise benefit leads to the setting of what has come to be termed a ‘Ramsey price’.

The two examples of welfare analysis just described proceed within a partial equilibrium framework, treating each industry as negligible compared to the entire economy and regarding the marginal utility of money as approximately constant to each individual. Marshall's rather fragmentary remarks on optimal tax systems, income redistribution and the ‘doctrine of maximum satisfaction’ cannot be restricted in this way, and so raise serious unresolved analytical difficulties. On the other hand, his tax-subsidy argument was a valid counterexample to arguments that competition must lead to a social optimum, or that optimal indirect tax systems must involve uniform tax rates. It must also be borne in mind that utilitarian welfare economics was for Marshall only a first step towards a more evolutionary analysis of possible modes of improving the physical quality and the values and activities of mankind.

Bibliographic note: Marshall's treatment of welfare economics is to be found in (1920, pp. 18–19, 124–37, 462–76, 487–93). Ellis and Fellner (1943) is a good statement of the standard interpretation of the Marshall–Pigou tax-subsidy argument, emphasizing external effects. See also Bharadwaj (1972). On Marshall's treatment of compromise benefit see Whitaker (1986, pp.186–8). Myint (1948) gives a useful general perspective on Marshall's welfare theory. Albon (1989) offers an intriguing insight into Marshall's attempt to apply welfare analysis to issues surrounding the British Post Office monopoly.

Interrelated markets and distribution theory

Marshall was anxious to emphasize the interdependence of markets and introduced his treatment of joint and composite demand and supply largely for this purpose. A group of goods is jointly supplied if all are outputs of a single productive activity and jointly demanded if all are inputs. On the other hand, a particular good is compositely supplied or demanded if it is provided or acquired by several distinct productive activities. Marshall's formal treatment of joint demand and supply proceeded on the general assumption that the products involved were consumed or produced in fixed proportions, as did his related analysis of the ‘derived demand’ for any one of several jointly demanded inputs – ‘derived’ since

the demand for such inputs is derived from the demand for their joint product. The derived demand curve for a specific input can be constructed conceptually by supposing that its supply is perfectly elastic at an arbitrary price and that the markets for the output and all the inputs (including the specific input) adjust to equate quantity demanded to quantity supplied in each market. This gives a price quantity combination on the derived demand curve for the specific input. Other such combinations can be obtained by varying the arbitrarily chosen price and repeating the exercise, and so on. Marshall laid down four rules for inelasticity of derived demand. These were that the input should have no good substitutes, that the product it helps make should be inelastically demanded, that the input should account for only a small part of production costs, and that cooperating inputs should be inelastically supplied. Fixity of input ratios guaranteed the first condition, but the more general case was asserted rather than proven. The advantage of working with the derived demand curve for an input is that it permits a more transparent analysis of the effects of changes in the supply conditions of the singled-out input.

The prime example of joint demand is the demand for productive inputs, and Marshall's analysis of market interdependence was carried through more fully in this specific connection, the role of substitution among inputs receiving full acknowledgement. The principle of substitution ensured that input usage tended to be adjusted by firms so as to minimize the total production cost of any level of output. Thus, the value of the marginal product of an input (or the ‘net product’ as Marshall termed it) tended under competition to equal the unit price of the input. There has been some confusion about the relation between ‘net product’ and marginal product because the former allows usage of other inputs to adjust consequentially when the chosen input is increased while the latter does not. But, provided that the initial situation is cost-minimizing, the adjustment of other inputs (if small) has no effect on the change in output – an application of the envelope theorem. Marshall recognizes this explicitly (1920, p. 409n) and there is no good reason for refusing to classify him as a marginal productivity theorist.

Interdependence among input markets was further highlighted in the analysis of the competition of several industries for an input that is in temporarily or permanently fixed overall supply. A peculiarity of this last analysis was the insistence on excluding from the marginal cost of any industry the cost of bidding such fixed resources away from other uses. This is a perfectly legitimate application of the general envelope theorem: provided resource use is optimally adjusted, the marginal cost of increasing output will be the same whatever input or sub-group of inputs is increased. But Marshall's insistence on asymmetry where there is really symmetry can be accounted for only by his desire to legitimize, and extend to quasi-rent, the classical doctrine that rent is price determined rather than a price-determining element of cost.

Marshall's vision of market interdependence culminates in his treatment of income distribution, where he seeks to bring out the extents to which the interests of different factors of production are harmonious or conflicting. Distribution is determined by the interaction of the demands and supplies for the various inputs, the demands being essentially joint demands. Marginal productivity is a theory of input demand, not a complete theory of distribution, because the supplies of the various inputs cannot be viewed as fixed, at least in the long period. Indeed, in the long period the dominant influences on the prices of factors other than land are exerted by their supply conditions. The costs that then have to be met must ensure that various kinds of labour and capital continue to be replaced in their existing uses and quantities.

From an overall view ‘The net aggregate of all the commodities produced is itself the true source from

which flow the demand prices for all these commodities, and therefore for the agents of production used in making them' (1920, p. 536). This aggregate, 'the national dividend', is distributed among the factors of production. It is at once the

> aggregate net product of, and the sole source of payment for, all the agents of production within the country: it is divided up into earnings of labour; interest of capital; and lastly the producer's surplus, or rent, of land and of other differential advantages for production. It constitutes the whole of them, and the whole of it is distributed among them; and the larger it is, the larger, other things being equal, will be the share of each of them. (1920, p. 536)

The share going to any class of inputs will depend upon the need people have for its services: 'not the *total* need, but the *marginal* need' (p. 536, italics original). But a complicating influence for distribution theory, although one 'more full of hope for the future of the human race than any that is known to us' lies in the fact that 'highly paid labour is generally efficient and therefore not dear labour' (p. 510). Influenced by F.A. Walker, Marshall was a strong proponent of the 'economy of high wages' argument that high wages increase labour efficiency, not perhaps immediately, but cumulatively over time and perhaps over generations: effects that transcend simple theorizing in terms of static equilibrium.

All the different productive factors cooperating in production have a common interest in increasing the size of the pie to be shared, the national dividend or income, but each factor has a selfish interest in restrictive practices that increase its own share, even if they reduce the size of the pie slightly. A prime question of social policy for Marshall is how these divergent incentives can be reconciled: how combined action by various groups, such as unions, can be prevented from assuming forms that, while perhaps individually beneficial to any one group in isolation, are certainly mutually harmful if undertaken by all.

Marshall here enters into macroeconomic forms of argument, and it is indeed true that he did toy with the formal specification of macroeconomic models of growth and distribution (see Whitaker, 1975, vol. 2, pp. 305–16). But, with this exception, it should be emphasized that his treatment of market interdependence fell far short of a full theory of general equilibrium on Walrasian lines. Even when formalizing market interdependence in the mathematical appendix to *Principles* (1920, pp. 846–56), he simply treated the demand or supply of each commodity as a function of nothing but the price of the commodity itself. The links between the generation of income in factor markets and the expenditure of that income in product markets were left quite vague. Again, it must be recalled that the development of comprehensive fully articulated equilibrium theories was not Marshall's aim.

Bibliographic note: The key sections for Marshall's treatment of interrelated markets and distribution theory are (1920, pp. 381–54, 504–45, 660–67, 846–56). For general commentaries on Marshall's distribution theory see Stigler (1941), H.M. Robertson (1970), Whitaker (1974; 1988). On Marshall's treatment of labour supply see Walker (1974; 1975), Matthews (1990). On the economy of high wages see Petridis (1996).

Monopoly and combination

Marshall's analysis of price and output determination by a profit-maximizing monopolist, and of the effects of taxing such a monopolist, followed the lead of Cournot. The concept of marginal revenue was implicit in the mathematical statement, but Marshall's chosen vehicle was geometrical. Curves of average revenue and cost, and of their difference, average net revenue, y, (all functions of the quantity sold, x) were superimposed on a grid of iso-profit hyperbolae of form $xy = constant$. Profit was maximized when the average net revenue curve touched the highest such iso-profit curve. Weighting consumer surplus into the maximand, as well as net revenue, gave rise to the welfare analysis of ‘compromise benefit’ already mentioned.

Monopoly analysis was applied to trades unions, with the use of the concept of the derived demand for an input. A union controlling a labour input for which derived demand is inelastic can certainly raise wages – not only the wage rate but the total wages received – although at the price of unemployment of some members. Whether such a monopolistic restriction can be sustained for long is more doubtful, as there will be pressures both to enter the union and to evade its grasp by the relocation or reorganization of production.

A more problematic question was whether ‘labour's disadvantage in bargaining’ meant that combined action by workers could raise wages, even without any restriction of labour supply. Marshall believed that it did, but emphasized that the result might be less capital accumulation by non-workers, an outcome that could harm workers eventually.

The extremes of monopoly and competition were both covered by the theory of normal value, even though the competition might be more akin to later concepts of imperfect or monopolistic competition than to any ideal form of perfect competition. But ‘normal action falls into the background, when Trusts are striving for the mastery of a large market’ (1920, p. xiv). The incidents, tactics and alliances of oligopolistic conflict defied reduction to a simple general theory. They were to have been considered in the uncompleted second volume of *Principles* and were to some extent covered by *Industry and Trade*. The latter's treatment of entry-limiting behaviour by a ‘conditional monopolist’, who dominates the market but does not control entry, is of considerable interest in the light of much recent work on this class of problems.

Bibliographic note: Marshall's treatment of monopoly theory is to be found in (1879b, pp. 180–86; 1920, pp. 477–95, 856–8). For his views on trusts and conditional monopolies see (1890b; 1919, pp. 395–635, especially 395–422). For his views on trades unions see (1879b, pp. 187–213; 1892, pp. 362–402; 1920, pp. 689–722) and Petridis (1973). On ‘labour's disadvantage in bargaining’ see Hicks (1930). Liebhafsky (1955) summarizes the relevant arguments of *Industry and Trade*.

Monetary theory

Marshall was in full command of previous British discussions of monetary issues, but not himself a major contributor to the development of monetary theory. His evidence before royal commissions in 1887 and 1899 showed an impressive mastery of monetary analysis, both domestic and international, and was minutely examined by successive generations of Cambridge students, serving for many years virtually as a textbook. But it was not until 1923, with the appearance of *Money Credit and Commerce*, that Marshall put forward his monetary views in a systematic way. By then these had not the novelty, nor he the vigour, to advance contemporary discussion.

Marshall's most important contribution to monetary theory was to place the overall demand for money in

the context of individual choices as to the fraction of one's wealth to keep on hand as ready cash. This approach, set out clearly in a manuscript of the early 1870s (Whitaker, 1975, Vol. I, pp. 164–77), was developed by Marshall's Cambridge successors, especially A.C. Pigou and F. Lavington, into what is termed the ‘Cambridge k’ approach. It laid the background for the treatment of the demand for money in J.M. Keynes (1936). On international monetary theory, Marshall espoused a form of purchasing power parity.

Marshall's name is particularly associated with his proposals for ‘symmetalism’, the use of a fixed-weight combination of gold and silver as the monetary base, and for indexed contracts based on a ‘tabular standard of value’, or price index, to be maintained by the government. The former was offered as an improvement on fixed-ratio bimetallism, of which he was never more than a lukewarm adherent.

Marshall had interesting, if fragmentary, insights into business fluctuations and general unemployment, which he viewed as temporary disequilibrium consequences of credit market dislocations. These spilled over into general coordination failures, with unemployment in one market spreading to others by reducing demand in cumulative fashion – the germ at least of the multiplier concept. On the other hand, Say's Law was maintained as an equilibrium truth of great importance. He saw the remedies for cyclical unemployment in the ‘continuous adjustment of means to ends, in such a way that credit can be based on the solid foundation of fairly accurate forecasts’, and in curbs on reckless inflations of credit that are ‘the chief cause of all economic malaise’ (1920, p. 710).

Bibliographic note: Marshall's monetary evidence is reproduced in J.M. Keynes (1926, pp. 3–195, 265–326). Other sources for his monetary views are Whitaker (1975, vol. 1, pp. 164–77), and Marshall (1887; 1923, pp. 12–97, 140–54, 225–33, 264–320). The standard treatment of Marshall's monetary views is Eshag (1963). For Marshall's views on business fluctuations see his (1879b, pp. 150–57; 1885a; 1892, pp. 400–3; 1920, pp. 710–11; 1923, pp. 234–63). Also see Wolfe (1956), Laidler (1990).

International trade

Marshall's major contribution to international trade theory was his well-known geometrical analysis of the equilibrium and stability of two-country trade by means of intersecting offer curves. Each country's offer curve indicated the number of ‘bales’ of home goods it was prepared to exchange for a specified number of bales of foreign goods, demand being elastic or inelastic as an increase in the latter caused the former to increase or decrease. Possibilities of multiple and unstable offer-curve intersections were noted. The offer curves themselves were taken as data, although complex readjustments of production and consumption underlay them. The need for a separate theory of international trade was justified, in classical vein, by the supposed international immobility of factors of production that remain mobile domestically.

The main purpose of this theoretical apparatus was to examine the effects of tariffs. A country might gain by selfishly exploiting its monopoly power through restricting trade, and would certainly gain if trading equilibrium occurred on an inelastic portion of the foreign offer curve. But Marshall came to doubt increasingly the transferability of this result to a multi-country case, although admitting that it might apply to an export tax on an exceptional commodity (like British steam coal) lacking close substitutes and incapable of being produced elsewhere.

A related attempt to construct a theoretical measure of the ‘net benefit’ a country gains from foreign

trade, analogous to the measures of consumer and producer surplus, was not entirely satisfactory as the partial equilibrium context had clearly been transcended.

On matters of concrete trade policy for Britain, Marshall was a firm but cautious adherent of free trade, even unilateral free trade, but became increasingly concerned with the prospects for Britain's position in the world economy. The discussion in *Industry and Trade* of the links between foreign competition and domestic industrial organization and structure reflected this concern.

Bibliographic note: For Marshall's treatment of the theory of international trade by offer curves see Whitaker (1975, vol. 1, pp. 260–79; vol. 2, pp. 111–81), Marshall (1923, pp. 155–224, 330–60). For the net benefit measure see Whitaker (1975, vol. 1, pp. 379–81) and Marshall (1923, pp. 338–40). Commentaries on Marshall's theory are to be found in Viner (1937, pp. 527–92), Chipman (1965), Johnson and Bhagwati (1960) and Creedy (1990). For Marshall's views on trade policy and trends see Marshall (1919, pp. 1–177, 681–784; 1923, pp. 98–139, 201–24) and J.M. Keynes (1926, pp. 367–420).

A brief survey of Marshall's writings with suggestions for further reading

The first editions of Marshall's five books were (1879b), (1890a), (1892), (1919), (1923). *Economics of Industry* (1879b) had a new edition in 1881 and was reprinted with minor changes several times up to 1892. It is an important source for Marshall's views on distribution theory, trades unions and business fluctuations. His magnum opus, *Principles* (1890a), had new editions in 1891, 1895, 1898, 1907, 1910, 1916 and 1920. The title was changed to its final form (as in (1920)) in the fifth edition. *Principles* is the basic source for Marshall's views on the theories of value and distribution as well as his broader views on economics and social welfare. Since the rewritings between editions were substantial, the ninth variorum edition, edited by C.W. Guillebaud, Marshall's nephew (Guillebaud, 1961), is essential for serious study. The first of its two volumes is a facsimile of the eighth edition of 1920. The second volume contains deleted passages from earlier editions, editorial notes, and various supporting documents. Users of the differently paginated Macmillan paperback edition of the eighth edition should note that all page references to the eighth edition given above must be located by using the table of correspondences appended to the paperback version.

Elements of the Economics of Industry (1892) had new editions in 1896 and 1899 and frequent reprintings. The last preface is dated 1907. It is essentially an abridgement of *Principles*, designed to replace *Economics of Industry*, a book that Marshall had come to despise, quite unjustifiably. *Elements* contains Marshall's fullest treatment of trades unions. *Industry and Trade* (1919) had new editions in 1919, 1920 and 1923 but only the first of these involved significant changes. Its three books deal with ‘Some origins of present problems of industry and trade’, ‘Dominant tendencies of business organization’, and ‘Monopolistic tendencies: their relations to public well-being’. It adopts a largely historical and comparative approach and focuses on contemporary issues. Nevertheless, it contains many passages and insights of permanent interest and warrants closer attention by economists than it has received until recently. *Money Credit and Commerce* (1923) had only one edition and conveys Marshall's views on money, international trade and business fluctuations. Although blemished, it should not be dismissed.

An almost comprehensive annotated list of Marshall's occasional writings is found in Pigou (1925, pp. 500–8), which also reprints many of the texts of these writings. Guillebaud (1961) reproduces further

occasional pieces. The ‘Pure Theory’ chapters (1879a), privately printed by Sidgwick, were first published in reprint form in 1930. A corrected and amplified version is included in Whitaker (1975). The two volumes of the latter also reproduce Marshall's unpublished early manuscripts, mainly from the 1870s, including several manuscript chapters from the abandoned volume on foreign trade. Marshall's important contributions to official enquiries are collected in J.M. Keynes (1926), a book that is supplemented by Groenewegen (1996).

The literature on Marshall's life and thought is too extensive to allow for more than a highlighting of some significant contributions. The splendid memorial essay by J.M. Keynes (1924) is not to be missed, although outdated on some points, nor is the charming memoir by Marshall's wife (M.P. Marshall 1944). Pigou (1925) includes fascinating vignettes by several of Marshall's colleagues and friends and Guillebaud (1971) gives a nephew's reminiscences. A major scholarly biography (Groenewegen 1995) covers Marshall's life and thought exhaustively, while a comprehensive three-volume edition of Marshall's correspondence (Whitaker 1996) provides much new information. Additional primary material on Marshall is provided in Raffaelli, Biagini and McWilliams-Tullberg (1995), where notes on Marshall's 1873 lectures to women students are reproduced, and Raffaelli (1994), where Marshall's essays on philosophical and psychological manuscripts from the late 1860s are reproduced and analyzed. See also Harrison (1963). Newspaper reports on public lectures Marshall gave during his years in Bristol are reproduced in Coase and Stigler (1969), Whitaker (1972) and Butler (1995).

Valuable overall assessments of Marshall are provided by Cannan (1924), Schumpeter (1941), Viner (1941), Shove (1942) and O'Brien (1981). Maloney (1985) studies Marshall's involvement in the professionalization of British economics. An extensive body of detailed analysis and criticism of Marshall's thought, mainly conducted in academic journals, continues to expand, with growing tributaries from Italy and Japan in particular. Wood (1982; 1996) assembles in eight volumes a somewhat miscellaneous collection of 239 pieces on Marshall, but standard bibliographic aids such as EconLit are recommended for a comprehensive search. The 1990 centenary of the publication of *Principles* produced two books of essays on Marshall (Whitaker 1990, McWilliams-Tullberg 1990) and several symposia on Marshall in economics journals. Samples of recent research can be found in Arena and Quéré (2003). On Marshall's social and behavioural views see Parsons (1931; 1932), Whitaker (1977) and Chasse (1984). For Marshall's views on socialism and trades unions see, respectively, McWilliams-Tullberg (1975) and Petridis (1973).

See Also

- ceteris paribus
- consumer surplus
- demand price
- external economies
- Marshall, Mary Paley

Selected works

1872. Review of Jevons (1871). *Academy*, April. Reprinted in Pigou (1925).

1874. The future of the working classes. *The Eagle*. Reprinted in Pigou (1925).

1876. On Mr. Mill's theory of value. *Fortnightly Review*, April. Reprinted in Pigou (1925).

1879a. *The Pure Theory of Foreign Trade. The Pure Theory of Domestic Values*. Privately printed. Reprinted in 1930, London: London School of Economics, Scarce Works in Political Economy No. 1; and in amplified form in Whitaker (1975).

1879b. (With M.P. Marshall.) *The Economics of Industry*, 2nd edn. London: Macmillan, 1881.

1884. Where to house the London poor. *Contemporary Review*, March. Reprinted in Pigou (1925).

1885a. How far do remediable causes influence prejudicially (a) the continuity of employment (b) the rate of wages? with four appendices. In *Report of Proceedings and Papers of the Industrial Remuneration Conference*, ed. C. Dilke. London: Cassel. The important appendix on ‘Theories and facts about wages’ is also reproduced in Guillebaud (1961).

1885b. *The Present Position of Political Economy: An Inaugural Lecture delivered at the Senate House Cambridge in February 1885*. London: Macmillan. Reprinted in Pigou (1925).

1885c. On the graphic method of statistics. *Jubilee Volum*e, a supplement to *Journal of the [London] Statistical Society*. Reprinted in Pigou (1925).

1887. Remedies for fluctuations of general prices. *Contemporary Review*, March. Reprinted in Pigou (1925).

1889. Cooperation. Presidential address to the 21st annual Cooperative Congress, Ipswich. Reprinted in Pigou (1925).

1890a. *Principles of Economics, Volume One*. London: Macmillan.

1890b. Some aspects of competition. Presidential address to Section F of the British Association for the Advancement of Science. Reprinted in Pigou (1925).

1892. *Elements of Economics of Industry*, 3rd edn. London: Macmillan, 1899.

1893. On rent. *Economic Journal* 3, 74-90. Reprinted in Guillebaud (1961).

1897. The old generation of economists and the new. *Quarterly Journal of Economics* 11, 115–35. Reprinted in Pigou (1925).

1898. Distribution and exchange. *Economic Journal* 8, 37–59. Portions are reprinted in Pigou (1925) and Guillebaud (1961).

1902. *A Plea for the Creation of a Curriculum in Economics and Associated Branches of Political Science*. London: Macmillan. Reprinted in Guillebaud (1961).

1907. The social possibilities of economic chivalry. *Economic Journal* 17, 7–29. Reprinted in Pigou (1925).

1917. National taxation after the war. In *After-War Problems*, ed. W. Dawson. London: George Allen and Unwin. Partly reproduced in Pigou (1925).

1919. *Industry and Trade*, 4th edn. London: Macmillan, 1923.

1920. *Principles of Economics: An Introductory Volume*. London: Macmillan. The eighth edition of Marshall (1890a).

1923. *Money, Credit and Commerce*. London: Macmillan.

Bibliography

Albon, R. 1989. Alfred Marshall and the consumers' loss from the British Post Office monopoly. *History of Political Economy* 21, 679–88.

Aldrich, J. 1996. The course of Marshall's theorizing about demand. *History of Political Economy* 28, 171–218.

American Economic Association. 1953. *Readings in Price Theory*. Homewood, IL: Irwin.

Arena, R. and Quéré, M., eds. 2003. *The Economics of Alfred Marshall: Revisiting Marshall's Legacy*. Basingstoke and New York: Palgrave Macmillan.

Bharadwaj, K. 1972. Marshall on Pigou's Wealth and Welfare. *Economica* 39, 32–46.

Biswas, T. 1977. The Marshallian consumer. *Economica* 44, 47–56.

Butler, R.W. 1995. 'The economic condition of America': Marshall's missing speech at University

College, Bristol. *History of Political Economy* 27, 405–16.

Cannan, E. 1924. Alfred Marshall, 1842–1924. *Economica* 4, 257–61.

Chasse, J. 1984. Marshall, the human agent and economic growth: wants and activities revisited. *History of Political Economy* 16, 381–404.

Chipman, J. 1965. A survey of international trade: Part 2, the neoclassical theory. *Econometrica* 33, 685–760.

Chipman, J. 1990. Marshall's consumer's surplus in modern perspective. In Whitaker (1990).

Coase, R. 1975. Marshall on method. *Journal of Law and Economics* 18, 25–31.

Coase, R. and Stigler, G. 1969. Alfred Marshall's lectures on progress and poverty. *Journal of Law and Economics* 12, 181–226.

Cournot, A. 1838. *Mathematical Principles of the Theory of Wealth*. Trans. N. Bacon, New York: Macmillan, 1897.

Creedy, J. 1990. Marshall and international trade. In Whitaker (1990).

Cunningham, W. 1892. The perversion of economic history. *Economic Journal* 2, 491–506.

Ellis, H. and Fellner, W. 1943. External economies and diseconomies. *American Economic Review* 33, 493–511.

Eshag, E. 1963. *From Marshall to Keynes: An Essay on the Monetary Theory of the Cambridge School*. Oxford: Blackwell.

Friedman, M. 1949. The Marshallian demand curve. *Journal of Political Economy* 57, 463–95. Reprinted in M. Friedman, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Frisch, R. 1950. Alfred Marshall's theory of value. *Quarterly Journal of Economics* 64, 495–524.

Gee, J. 1983. Marshall's views on ‘short period’ value formation. *History of Political Economy* 15, 181–205.

Groenewegen, P. 1995. *A Soaring Eagle: Alfred Marshall 1842–1924*. Aldershot and Brookfield, VT: Edward Elgar.

Groenewegen, P., ed. 1996. *Official Papers of Alfred Marshall: A Supplement*. Cambridge: Cambridge University Press.

Guillebaud, C., ed. 1961. *Alfred Marshall: Principles of Economics: Ninth (Variorum) Edition*, 2 vols. London: Macmillan.

Guillebaud, C. 1971. Some personal reminiscences of Alfred Marshall. *History of Political Economy* 3, 1–8.

Hague, D. 1958. Alfred Marshall and the competitive firm. *Economic Journal* 68, 673–90.

Harrison, R. 1963. Two early articles by Alfred Marshall. *Economic Journal* 73, 2–30.

Hicks, J. 1930. Edgeworth, Marshall and the indeterminateness of wages. *Economic Journal* 40, 215–31.

Jevons, W. 1871. *The Theory of Political Economy*. London: Macmillan.

Johnson, H. and Bhagwati, J. 1960. Notes on some controversies in the theory of international trade. *Economic Journal* 70, 74–93.

Keynes, J. 1924. Alfred Marshall, 1842–1924. *Economic Journal* 34, 311–72. Reprinted with slight changes in Pigou (1925).

Keynes, J.M., ed. 1926. *Official Papers of Alfred Marshall*. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.N. 1891. *The Scope and Method of Political Economy*. London: Macmillan.

Laidler, D. 1990. Alfred Marshall and the development of monetary economics. In Whitaker (1990).

Liebhafsky, H. 1955. A curious case of neglect: Marshall's *Industry and Trade*. *Canadian Journal of Economics* 21, 339–53.

McWilliams-Tullberg, R. 1975. Marshall's ‘tendency to socialism’. *History of Political Economy* 7, 75–111.

McWilliams-Tullberg, R., ed. 1990. *Alfred Marshall in Retrospect*. Aldershot and Brookfield VT: Edward Elgar.

Maloney, J. 1985. *Marshall, Orthodoxy and the Professionalisation of Economics*. Cambridge:

Cambridge University Press.

Marshall, M. 1944. *What I Remember*. Cambridge: Cambridge University Press.

Matthews, R. 1990. Marshall and the labour market. In Whitaker (1990).

Mill, J.S. 1848. *Principles of Political Economy*. London: Parker. Many subsequent editions.

Myint, H. 1948. *Theories of Welfare Economics*. London: London School of Economics.

Newman, P. 1960. The erosion of Marshall's theory of value. *Quarterly Journal of Economics* 74, 587–600.

O'Brien, D.P. 1981. Alfred Marshall, 1842–1924. In *Pioneers of Modern Economics in Britain*, ed. D. O'Brien and J. Presley. London: Macmillan.

O'Brien, D. 1990. Marshall's work in relation to Classical economics. In Whitaker (1990).

Opie, R. 1931. Marshall's time analysis. *Economic Journal* 41, 199–215.

Parsons, T. 1931. Wants and activities in Marshall. *Quarterly Journal of Economics* 46, 101–40.

Parsons, T. 1932. Economics and sociology: Marshall in relation to the thought of his time. *Quarterly Journal of Economics* 46, 316–47.

Petridis, A. 1973. Alfred Marshall's attitudes to and economic analysis of trade unions: a case of anomalies in a competitive system. *History of Political Economy* 5, 165–98.

Petridis, A. 1996. Brassey's law and the economy of high wages in nineteenth century economics. *History of Political Economy* 28, 583–606.

Pigou, A., ed. 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Raffaelli, T. 1994. Alfred Marshall's early philosophical writings. *Research in the History of Economic Thought and Methodology: Archival Supplement* 4, 51–158.

Raffaelli, T. 2003. *Marshall's Evolutionary Economics*. London and New York: Routledge.

Raffaelli, T., Biagini, E. and McWilliams-Tullberg, R., eds. 1995. *Alfred Marshall's Lectures to Women: Some Economic Questions Directly Connected to the Welfare of the Laborer*. Aldershot and Brookfield, VT: Edward Elgar.

Robbins, L. 1928. The representative firm. *Economic Journal* 38, 387–404.

Robertson, D., Sraffa, P. and Shove, G. 1930. Increasing returns and the representative firm. *Economic Journal* 40, 76–116.

Robertson, H. 1970. Alfred Marshall's aims and methods illustrated from his treatment of distribution. *History of Political Economy* 2, 1–65.

Schumpeter, J. 1941. Alfred Marshall's *Principles*: a semi-centennial appraisal. *American Economic Review* 31, 236–48.

Shove, G. 1942. The place of Marshall's *Principles* in the development of economic theory. *Economic Journal* 52, 294–329.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36, 535–50.

Stigler, G. 1941. *Production and Distribution Theories: The Formative Period*. New York: Macmillan.

Stigler, G. 1950. The development of utility theory. *Journal of Political Economy* 58, 307–27; 373–96.

Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3, 23–46. Reprinted in American Economic Association (1953).

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

Viner, J. 1941. Marshall's economics in relation to the man and his times. *American Economic Review* 31, 223–35.

Walker, D. 1969. Marshall's theory of competitive exchange. *Canadian Journal of Economics* 2, 590–98.

Walker, D. 1974. Marshall on the long-run supply of labor. *Zeitschrift für die Gesamte Staatswissenschaft* 130, 691–705.

Walker, D. 1975. Marshall on the short-run supply of labor. *Southern Economic Journal* 41, 429–41.

Whitaker, J. 1972. Alfred Marshall: the years 1877 to 1885. *History of Political Economy* 4, 1–61.

Whitaker, J. 1974. The Marshallian system in 1881: distribution and growth. *Economic Journal* 84, 1–17.

Whitaker, J., ed. 1975. *The Early Economic Writings of Alfred Marshall, 1867–1890*, 2 vols. London:

Macmillan.

Whitaker, J. 1977. Some neglected aspects of Alfred Marshall's economic and social thought. *History of Political Economy* 9, 161–97.

Whitaker, J. 1982. The emergence of Marshall's period analysis. *Eastern Economic Journal* 8, 15–29.

Whitaker, J. 1986. The continuing relevance of Alfred Marshall. In *Ideas in Economics*, ed. R. Black. London: Macmillan.

Whitaker, J. 1988. The distribution theory of Marshall's Principles. In *Theories of Income Distribution*, ed. A. Asimakopulos. Boston, Dordrecht and Lancaster: Kluwer Academic.

Whitaker, J., ed. 1990. *Centenary Essays on Alfred Marshall*. Cambridge: Cambridge University Press.

Whitaker, J., ed. 1996. *The Correspondence of Alfred Marshall, Economist*, 3 vols. Cambridge: Cambridge University Press.

Wolfe, J. 1956. Marshall and the trade cycle. *Oxford Economic Papers* 8, 90–101.

Wood, J. 1982. *Alfred Marshall: Critical Assessments*, 4 vols. London: Croom Helm.

Wood, J. 1996. *Alfred Marshall: Critical Assessments, Second Series*, 4 vols. London: Routledge.

How to cite this article

Whitaker, John K. "Marshall, Alfred (1842–1924)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000089> doi:10.1057/9780230226203.1046

The New Palgrave Dictionary of Economics Online

Marshall, Mary Paley (1850 – 1944)

G. Becattini

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Marshall, A.; Paley, M. P.

Article

British economist, born in Ufford (Nottinghamshire) on 24 October 1850; died in Cambridge 7 March 1944. Great-granddaughter of the great theologian William Paley, she was brought up in a strictly evangelical faith in Ufford, her father's vicarage. Thomas Paley, had taken a good degree in mathematics (33rd wrangler) in 1833 at Cambridge, and had been, for a period, a fellow of St John's College. Mary had one elder sister and two younger brothers.

In 1871, with a scholarship, she went up to Cambridge to complete her education with studies at university level. Under the whimsical chaperon Anne J. Clough (sister of the poet), and the teaching of a handful of young voluntary dons committed to the cause of higher education of women (among them Henry Sidgwick and Alfred Marshall), she took the Moral Sciences Tripos. She graduated, albeit informally, in 1874 (the first woman to achieve such a distinction in Cambridge) but the board of examiners (W.S. Jevons was among them) was so bitterly divided that in the certificate they recorded, very unusually, that she had received two votes for a first class and two for a second.

Shortly after her degree Mary Paley began to teach and to tutor female students in the newly opened Newnham Hall. In 1876, on request, she began to write a small economic textbook for Extension Lectures, that eventually became *The Economics of Industry* (1879). In the same year she became engaged to Alfred Marshall. They married in Ufford in July 1877. From that date onwards, till the death of Alfred Marshall in 1924, her life was essentially devoted, first in Bristol, where they settled after marriage, then in Oxford (1883–4) and finally in Cambridge, to helping him in his scientific work and to saving him from all the normal nuisances of life.

For several decades Mary Paley Marshall taught and tutored female students of economics in Newnham college. A member of many associations (Charity Organization Society, Ethical Society, and so on) she participated in the founding group of the British Economic Association. After 1924 she became the first librarian of the newly founded Marshall Library, which she visited regularly until her 90th year. In 1928

the University of Bristol awarded her an honorary degree. She was a gifted amateur water colour painter and her posthumous Memoir, *What I Remember* (1947), shows glimpses of literary talent. Mary was not buried beside Alfred, but her ashes were scattered in the garden of her house.

Mary Paley Marshall's claims to be considered as an economist by herself are, strictly speaking, unassessable. Personally she signed only a few short notes in the early issues of the *Economic Journal*, which show a clear mind, a good style and a balanced judgement, but no more. Her only title to fame resides in the green-covered *Economics of Industry*, co-authored with Alfred Marshall. This small textbook, reprinted many times and translated into several foreign languages, was rated very highly by contemporaries. J.M. Keynes went so far as to say: ‘It was, in fact, an extremely good book; nothing more serviceable for its purpose was produced for many years, if ever.’ From the viewpoint of the development of economic analysis the book is relevant as a sort of half-way house between the *Principles* of J.S. Mill and the *Principles* of A. Marshall. Despite some hints to the contrary by J.M. Keynes, the respective positions (teacher and pupil) and ages (Alfred was older by eight years) suggest that Mary Paley's contribution was only secondary and subordinate.

Worthy of mention is the help she afforded Alfred Marshall in preparing and amending all his works. In a letter to John Neville Keynes, there is a hint of a substantial collaboration: ‘My wife and I’, writes Alfred Marshall, alluding to an article by J.L. Laughlin (*Quarterly Journal of Economics*, 1887), ‘find it very hard to see Laughlin's points & perhaps we underrate the strength of his attack.’

Had it not been for the suffocating influence of Alfred, Mary Paley, with her clear mind, earnestness and strong will, would have become herself, we can confidently guess, an economist of repute and not, as is the case, a minor figure in the shadow of Alfred Marshall.

Selected works

1879. (With Alfred Marshall.) *The Economics of Industry*. London: Macmillan. 2nd edn, 1881.

1896. Conference of women workers. *Economic Journal* 6(21), 107–9.

1947. *What I Remember*. Cambridge: Cambridge University Press.

Bibliography

Clough, B.A. and Strachey, J.P. 1944. Mrs Alfred Marshall (Mary Paley) 1850–1944. *Cambridge Review* 65(1597), 300.

Constable, W.G. 1960. Art and economics in Cambridge. *The Eagle*, April.

Keynes, J.M. 1944. Mary Paley Marshall 1850–1944. *Economic Journal* 54, June–September, 268–84. Reprinted in J.M. Keynes, *Essays in Biography*, ed. G. Keynes. London: Rupert Hart-Davis. 2nd edn, 1951.

Keynes, J.N. Letter 1(58), kept in the Marshall Library, Cambridge.

How to cite this article

Becattini, G. "Marshall, Mary Paley (1850–1944)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000090> doi:10.1057/9780230226203.1047

The New Palgrave Dictionary of Economics Online

Marshall – Lerner condition

Murray C. Kemp
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

correspondence principle; international trade theory; Marshall–Lerner condition; terms of trade; transfer problem

Article

In the comparative-statical calculations of the simplest two-commodity barter theory of international trade, the outcomes invariably depend on the magnitude of the sum of the two import-demand elasticities, one relating to the country under study (the ‘home’ country) and the other to the rest of the world (collectively, the ‘foreign’ country); in particular, the response of a variable to a disturbance will be in one direction if the sum of elasticities is less than minus 1 and in the opposite direction if the sum is greater than minus 1. On the other hand, for some dynamic or ‘disequilibrium’ models of international trade it is a necessary and sufficient condition of local stability that the same sum of elasticities be less than minus 1. Let us define Δ as one plus the sum of the two elasticities of import demand. Then the so-called Marshall–Lerner condition requires that Δ be negative. Evidently the condition provides a link between the comparative-statics of international trade and some forms of trade dynamics. That such a link exists is, of course, the essence of Samuelson's correspondence principle.
Proceeding to a more detailed account of the Marshall–Lerner condition, let us suppose that the home country imports the first commodity, the foreign country the second; and let us denote by p the world price of the second commodity in terms of the first and by α and α^* parameters of the home and foreign economies, respectively. Then, in the absence of trade impediments and autonomous international transfers, we may write the general-equilibrium or *mutatis mutandis* home import-demand function as $\phi(1/p, \alpha)$, the foreign import-demand function as $\phi^*(p, \alpha^*)$ and the condition of world equilibrium as

$$\varphi(1/p, \alpha) - p\varphi^*(p, \alpha^*) = 0.$$

(1)

Suppose now that an initial equilibrium is disturbed by small changes in α and α^*. Differentiating (1) totally we obtain

$$\left(p^{-2}\varphi_{1/p} + \varphi^* + p\varphi_p^*\right)\mathrm{d}p = \varphi_\alpha \mathrm{d}\alpha - p\varphi_{\alpha^*}^* \mathrm{d}\alpha^*$$

or, equivalently,

$$\Delta \mathrm{d}p \equiv \left(1 + \xi + \xi^*\right)\mathrm{d}p = \left(\varphi_\alpha \mathrm{d}\alpha - p\varphi_{\alpha^*}^* \mathrm{d}\alpha^*\right) / \varphi^*$$
(2)

where the subscripts indicate partial derivatives and where $\xi \equiv \varphi_{1/p} / (p\varphi)$ and $\xi^* \equiv p\varphi_p^* / \varphi^*$ are the price elasticities of home and foreign import demand, respectively.
On the other hand, we may consider the dynamic tâtonnement defined by the differential equation

$$\dot{p} \equiv \mathrm{d}p / \mathrm{d}t = f\left[\varphi^*\left(p, \alpha^*\right) - \varphi(1/p, \alpha) / p\right]$$
(3)

where *f* is a differentiable sign-preserving function of the world excess demand for the second commodity and *t* denotes time. Evidently (3) is a dynamic extension of (1). For the local stability of *p* at an equilibrium value it is necessary and sufficient that $\mathrm{d}f / \mathrm{d}p$ be negative in a sufficiently small neighbourhood of the equilibrium value. Now a little calculation shows that $\mathrm{d}f / \mathrm{d}p = f' \Delta \varphi^* / p$, where the prime indicates differentiation; moreover, since *f* is differentiable and sign-preserving, f' is necessarily positive sufficiently near the equilibrium value of *p*. For local stability, therefore, it is necessary and sufficient that Δ be negative.
By way of illustration we may consider the traditional ‘transfer problem’. Identifying $\alpha = \alpha^*$ with the amount transferred from the foreign to the home country, in terms of the numeraire, eq. (1) reduces to

$$\varphi(1/p,\alpha) - p\varphi^*(p,\alpha) - \alpha = 0$$

and, if α is initially zero, eq. (2) reduces to

$$\Delta dp = \left(\varphi_\alpha - 1 + p\varphi^*_\alpha\right)d\alpha / \varphi^*.$$

Evidently ϕ_α and $1 - p\varphi^*_\alpha$ are home and foreign marginal propensities to consume the first commodity. Thus, in stable systems, the terms of trade move in favour of the recipient of a small payment if and only if that country's marginal propensity to consume the imported commodity is less than the foreign country's marginal propensity to consume the same commodity.

The home and foreign economies have been specified in terms of the functions ϕ and ϕ^* only. Whether in any particular context the Marshall–Lerner condition is satisfied depends on the structure imposed on ϕ and ϕ^* by the context. Evidently the appropriate structure will be quite different for economies with and without chronic unemployment and, more generally, for economies with and without internal market-clearing. It will also be quite different for rural and industrial economies, for growing and declining economies and for rich and poor economies.

While the inequality $\Delta < 0$ is now widely known as the Marshall–Lerner condition, the label is inappropriate in a context of dynamic analysis. For Alfred Marshall developed a quite different stability condition (see Marshall, 1879, II; 1923, Appendix J; Samuelson, 1947, pp. 266–7); Amano, 1968; Kemp, 1964, pp. 89–90); and Abba Lerner was not at all concerned with disequilibrium dynamics (see Lerner, 1944).

Bibliography

Amano, A. 1968. Stability conditions in the pure theory of international trade: a rehabilitation of the Marshallian approach. *Quarterly Journal of Economics* 82(2), 326–39.

Kemp, M.C. 1964. *The Pure Theory of International Trade and Investment*. Englewood Cliffs, NJ: Prentice-Hall.

Lerner, A.P. 1944. *The Economics of Control*. New York: Macmillan.

Marshall, A. 1879. *The Pure Theory of Foreign Trade*. Published privately. Reprinted with *The Pure Theory of Domestic Values*. London: London School of Economics and Political Science, 1949.

Marshall, A. 1923. *Money, Credit and Commerce*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

How to cite this article

Kemp, Murray C. "Marshall–Lerner condition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000092> doi:10.1057/9780230226203.1048

The New Palgrave Dictionary of Economics Online

Martineau, Harriet (1802 – 1876)

Robert W. Dimand and Evelyn L. Forget
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

classical political economy; laissez-faire; Marcet, J.; Martineau, H.; property rights; slavery

Article

Harriet Martineau, the best-selling popularizer of classical political economy, was born in Norwich, England, on 12 June 1802, the sixth of eight children of Thomas Martineau, a Unitarian textile manufacturer, and Elizabeth Rankin Martineau. She was educated at home, except that from 1813 to 1815 she studied French, Latin, and English composition at a school run by the Reverend Isaac Perry. Her early writings were religious, beginning with an article on ‘Female Writers on Practical Divinity’ for the *Monthly Repository*. After her father's death in 1826, she became engaged, but her fiancé died before they could be married. She remained single for the rest of her life. Investment losses in 1829 forced her to support herself by writing: William Johnson Fox's *Monthly Review* hired her as a book reviewer for 15 pounds a year, and when the Central Unitarian Association offered prizes for essays to convert Catholics, Jews, and Muslims, Martineau won all three prizes, for 15 guineas each. These prizes enabled her to visit her brother in Dublin in 1831. While there, she planned the series *Illustrations of Political Economy*, stories that would expound (especially to the working classes) the principles of classical political economy, to which she had been introduced by Jane Marcet's *Conversations on Political Economy* (1816) and James Mill's *Elements of Political Economy* (1821). The first of the 34 tales of political economy, of the Poor Laws, and of taxation was published in February 1832 by Charles Fox (brother of William Fox), and distributed by the Society for the Diffusion of Useful Knowledge. By 10 February, the first printing of 1,500 copies had all been sold, and a second printing of 5,000 copies ordered.

The success of *Illustrations of Political Economy* (1832–4; 2004a) made Martineau a celebrity: Henry Brougham, the Lord Chancellor, provided her with private papers on the impending reform of the Poor Law, while Robert Owen attempted, without success, to convert her to socialism. Although poor health interrupted her work several times, Harriet Martineau refused offers of government pensions from Lord Grey in 1835, Lord Melbourne in 1841, and W.E. Gladstone in 1873, lest her independence be

compromised. Instead, her friends raised funds for an annuity for her in 1843, and, when her health permitted she worked with great diligence, writing more than 1,600 articles for the *Daily News* from 1852 to 1866. Although her income was modest, she was a philanthropist, founding a building society among other charitable projects.

Harriet Martineau disclaimed originality for *Illustrations of Political Economy*, insisting that its purpose was didactic, making established principles better known. Her presentation of the principles was original, and her work stood out for its recognition of women as rational economic agents (1985). In her *Illustrations*, Martineau upheld laissez-faire and property rights. In later life, Martineau endorsed married women's property rights, condemned slavery as an illegitimate form of property, gave qualified support to workers' cooperatives, and accepted the need for state intervention in certain very limited circumstances. Her lectures at working men's institutions stressed education rather than state intervention as the remedy for most social ills. Her subsequent writings on America and slavery (1837; 2002) demonstrated that she was an adept economic analyst, not just a popularizer. For example, Martineau recognized the limited demand for the services of prostitutes in the South as evidence of the sexual exploitation of slaves, and identified the inability of slaveholders to make credible long-term commitments to their slaves as a source of inefficiency in slave agriculture (Levy, 2003). Martineau's *Society in America* (1837) emphasized the incompatibility of slavery and of the legal, political, and economic position of American women (lacking votes and, if married, property rights) with America's founding rhetoric of liberty. Her study of the United States was accompanied by her *How to Observe Morals and Manners* (1838), a methodological manual on comparative sociology and ethnography. In 1853, Martineau published a translation and abridgement of the pioneer sociologist Auguste Comte's *Philosophie positive*, an adaptation that pleased Comte so much that he had it translated back into French.

In 1855, anticipating death from heart disease, Martineau wrote her autobiography for posthumous publication (1877), but she lived until 27 June 1876.

See Also

- British classical economics
- Marcet, Jane Haldimand
- slavery

Selected Works

1832–4. *Illustrations of Political Economy, Taxation, Poor Law and Paupers*, 13 vols. London: C. Fox; repr. with new introduction by C. Franklin Bristol: Thoemmes, 2001.

1837. *Society in America*. London: Saunders and Otley. Abridged edn, ed. S.M. Lipset, New Brunswick, NJ: Transaction, 1981.

1838. *How to Observe Morals and Manners*. London: C. Knight. Repr. 1989, ed. M.R. Hill New Brunswick, NJ: Transaction.

1877. *Harriet Martineau's Autobiography*, 2 vols. London: Smith, Elder; repr. London: Virago, 1983.

1985. *Harriet Martineau on Women*, ed. G.G. Yates. New Brunswick, NJ: Rutgers University Press.

2002. *Writings on Slavery and the American Civil War by Harriet Martineau*, ed. D.A. Logan. DeKalb, IL: Northern Illinois University Press.

2004a. *Illustrations of Political Economy: Selected Tales*, ed. D.A. Logan. Peterborough, ON: Broadview Press.

2004b. *Harriet Martineau's Writings on the British Empire*, 5 vols, ed. D.A. Logan. London: Pickering & Chatto.

2005. *Harriet Martineau: Writings on British History and Military Reform*, 6 vols, ed. D.A. Logan. London: Pickering & Chatto.

2006. *The Collected Letters of Harriet Martineau*, 5 vols, ed. D.A. Logan. London: Pickering & Chatto.

Bibliography

Henderson, W. 1992. Harriet Martineau or when political economy was popular. *History of Education* 21, 383–403.

Levy, D. 2003. Taking Harriet Martineau's economics seriously. In *The Status of Women in Classical Economic Thought*, ed. R. Dimand and C. Nyland. Cheltenham, and Northampton, MA: Edward Elgar.

Logan, D.A. 2002. *The Hour and the Woman: Harriet Martineau's ‘Somewhat Remarkable’ Life*. DeKalb: Northern Illinois University Press.

Marcet, J. 1816. *Conversations on Political Economy*. 3rd edn, London: Longman, 1818.

Mill, J. 1821. *The Elements of Political Economy*. London: Baldwin, Cradock & Joy.

Orazem, C. 1999. *Political Economy and Fiction in the Early Writings of Harriet Martineau*. Frankfurt: Peter Lang.

Webb, R.K. 1960. *Harriet Martineau, a Radical Victorian*. London: Heinemann, and New York: Columbia University Press.

How to cite this article

Dimand, Robert W. and Evelyn L. Forget. "Martineau, Harriet (1802–1876)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000096>
doi:10.1057/9780230226203.1049

The New Palgrave Dictionary of Economics Online

martingales

Alan F. Karr

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

A martingale is a mathematical model of a fair game, or of some other process that is incrementally random noise. The term, which also denotes part of a horse's harness or a ship's rigging, refers in addition to a gambling system in which every losing bet is doubled; it was introduced into probability theory by J.L. Doob. Among stochastic processes, martingales have particular constancy properties with respect to conditioning. The time parameter may be either discrete or continuous, but since the latter is more important in economic applications, we concentrate on it.
Suppose that on a basic probability space there is defined a *history* $\mathcal{H} = (\mathcal{H}_t)_{t \geq 0}$ representing observable events as a function of time. For each *t*, $(\mathcal{H}_t)$ is the σ-algebra comprising events determined by observations over the interval [0,*t*], so that $(\mathcal{H}_s) \subseteq (\mathcal{H}_t)$ when $s \leq t$. Then a stochastic process $M = (M_t)_{t \geq 0}$ is a *martingale* with respect to this history if

1. (a) For each *t*, M_t is $\mathcal{H}_t$ measurable (i.e., the state of the process at *t* is observable over [0,*t*]);
2. (b) $E[|M_t|]<\infty$ for each *t*;
3. (c) The ‘martingale property’ holds: whenever *s*•*t*,

$$E[M_t|\mathcal{H}_s] = M_s. \tag{1}$$

When no history is specified, it is usually understood that $\mathcal{H}_t = \sigma(M_s; s \leq t)$. One specific consequence is that $E[M_t] = E[M_0]$ for each *t*, so that a martingale is constant in the mean.
Written as

$$E[M_t - M_s|\mathcal{H}_s] = 0, \; s \leq t,$$

the martingale property implies that the optimal (in the sense of minimum mean squared error, or MMSE) predictor of a future increment of a martingale is zero. Thus, a martingale is indeed a mathematical idealization of a fair game. In some ways this property is clearest in differential form: assuming that the differential dM_t, which always extends *forward* in time from t, can be defined, then M is a martingale provided that

$$E[dM_t|\mathcal{H}_t] = 0 \tag{2}$$

for each t. Thus, a martingale can be interpreted as a 'noise' process, in which the MMSE prediction of the differential dM_t is simply zero; in many applications this interpretation becomes quite literal. Martingales are also analogous to the residuals in a regression problem, where what remains unexplained by the model should reduce, ideally, to chance variation.

One can also define *supermartingales*, for which (1) becomes

$$E[M_t|\mathcal{H}_t] \leq M_s, \tag{3}$$

and *submartingales*, in which the sense of the inequality in (3) is reversed. A supermartingale represents a less-than-fair game.

All martingales are in some sense convex combinations of (generalizations of) two key examples, namely the Wiener and Poisson processes. If (W_t) is a Wiener process (Brownian motion), then the processes W_t and $W_t^2 - t$ are both martingales; in fact, these properties characterize the Wiener process. In discrete time, martingales generalize sums of independent, mean zero random variables; the Wiener process, which has independent and stationary increments, is a continuous time counterpart of these partial sum processes.

If (N_t) is a point process (or counting process), with N_t the number of events occurring in $[0, t]$, then under quite general assumptions there exists a nonnegative, predictable (a technical term, which in practice means left-continuous) random process (λ_t), the stochastic intensity of N, such that the process $M_t = N_t - \int_0^t \lambda_s ds$ is a martingale. Since $\lambda_t dt = E[dN_t|\mathcal{H}_t]$, M represents the new information realized as a function of time and, because of this and applications in statistics and state estimation, is known as the innovation martingale. For a Poisson process, which like the Wiener process has independent and stationary increments, the stochastic intensity is deterministic and equal to the rate of the process.

Square integrable martingales are especially important. A martingale *M* is *square integrable* if $\sup_t E[M_t^2] < \infty$, and in this case there exists a predictable process ⟨*M*⟩, the *predictable variation* of *M*, such that $M_t^2 - \langle M \rangle_t$ is a martingale. That the predictable variation is incrementally a conditional variance is confirmed by the differential relationship

$$d\langle M \rangle_t = E[(dM_t - E[dM_t|\mathcal{H}_t])^2|\mathcal{H}_t] = E[(dM_t)^2|\mathcal{H}_t].$$

Here the second equality holds because *M* is a martingale.

For the Wiener process $\langle W \rangle_t \equiv t$ in particular, the predictable variation is deterministic, a property characteristic of processes with independent increments. For a point process *N* with stochastic intensity λ, the predictable variation of the innovation martingale $dM_t = dN_t - \lambda_t dt$ is given by $d\langle M \rangle_t = \lambda_t dt$, which implies that a point process is locally and conditionally Poisson, in the sense that the incremental conditional mean and variance coincide.

Existence of the predictable variation is proved via the Doob-Meyer decomposition theorem, a cornerstone of the theory. The principal theoretical results pertaining to martingales fall into three classes: inequalities, convergence theorems and optimal sampling theorems.

So-called maximal inequalities, which provide upper bounds for probabilities of the form $P\{\sup_{s \le t}|M_s| > c\}$, are not only of inherent interest, but also the key tools for proving convergence theorems. Moreover, these inequalities form the basis of a profound connection between martingales and classical mathematical analysis.

Under various assumptions, given a martingale *M* there exists a random variable M_∞ such that $M_t \to M_\infty$ almost surely as $t \to \infty$. Convergence obtains both almost surely and in L^1 if *M* is uniformly integrable, and in this case $M_t = E[M_\infty|\mathcal{H}_t]$ for each *t*. Not all martingles converge, however; those that fail to converge include, for example, the Wiener process and most innovation martingales.

Optional sampling theorems require the further concept of a stopping time. A random time *T* (a random variable with values in $[0, \infty]$, interpreted as the time at which some event occurs – with $T=\infty$ corresponding to its not occurring) is a *stopping time* of the history $\mathcal{H}$ if $\{T \le t\} \in \mathcal{H}_t$ for each *t*. Intuitively, whether a stopping time has occurred by *t* can be determined from observations over [0, *t*], and does not require prescient knowledge of the future. The rule by which a gambler quits a game must be a stopping time. Associated with a stopping time *T* is a σ-algebra $\mathcal{H}_T$ representing events determined by observations over the random time interval [0, *T*] in the same way that for deterministic *t*, $\mathcal{H}_t$ corresponds to the interval [0, *t*].

Martingale property extends from deterministic times to stopping times, and imply in particular that an unfair game cannot be made fair by means of a stopping time. More precisely, if *M* is a martingale and *S* and *T* are stopping times with $S \le T$, then under broad – albeit not universal – conditions,

$$E[M_T|\mathcal{H}_S] \le M_S.$$

(4)

With S=0 in (4), taking expectations yields $E[M_T] = E[M_0]$. The corresponding result for supermartingales,

$$E[M_T|\mathcal{H}_S] \leq M_S,$$
(5)

demonstrates that an unfair game cannot be made fair via a stopping time, and dooms gambling systems without infinite resources to eventual failure.
Significant applications of martingales include mathematical statistics (likelihood ratio processes are martingales), queueing theory, filtering and prediction (for example, in signal processing) and economics.
A common feature of these applications is that they involve a random system ‘driven’ by a martingale in precisely the same manner that a dynamical system is driven by a forcing function. Given a (square integrable) martingale M and a predictable process C fulfilling integrability restrictions, the stochastic integral process

$$(C * M)_t = \int_0^t C_s dM_s$$

is itself a martingale, for which M acts as driving term. (Since M may change state discontinuously, whether endpoints are included in the interval of integration must be specified; in this case, the integral is over the closed interval [0, t].) Construction of stochastic integrals is a difficult, subtle problem: none of the conventional definitions can be applied pathwise (typically the sample paths of M are not of bounded variation), and instead one must employ sophisticated probability theory. The predictable variations satisfy

$$d\langle C * M\rangle_t = C_t^2 d\langle M\rangle_t.$$

Economic applications include, e.g., models of securities prices.
In applications, the inclusion of a ‘dt’-integral is often desirable or necessary, leading to

semimartingales, which are random processes Z of the form

$$Z_t = \int_0^t A_s ds + \int_0^t C_s dM_s, \tag{6}$$

where M is a martingale, C is a predictable process and A fulfills a technical property known as progressive measurability. (Integrability conditions must be satisfied as well.) The differential version of (6) is

$$dZ_t = A_t dt + C_t dM_t. \tag{7}$$

If the processes A and C, rather than specified exogenously, are functionals of Z, then (7) becomes a stochastic differential equation

$$dZ_t = \mu(Z_t)dt + \sigma(Z_t)dM_t, \tag{8}$$

or, more generally,

$$dZ_t = \mu(Z_s;\ s \le t)dt + \sigma(Z_s;\ s \le t)dM_t. \tag{9}$$

These equations can be solved – however, not using pathwise methods – under a variety of assumptions, but essentially only when the driving term is a martingale. For example, if the martingale is the Wiener process, solutions to (8) and (9) are known as diffusions and It ō processes, respectively, and the resultant theory as the It ō calculus, after its principal inventor, K. It ō . Alternatively, if M is the innovation martingale associated with a point process N then, inter alia, solutions to (8) can be used to construct recursive methods for filtering to extract signals from noise.

Bibliography

Brémaud, P. 1981. *Point Processes and Queues: Martingale Dynamics*. Berlin: Springer-Verlag.

Hall, P. and Heyde, C.C. 1980. *Martingale Limit Theory and its Applications*. New York: Academic Press.

Kallianpur, G. 1980. *Stochastic Filtering Theory*. New York: Springer-Verlag.

Karr, A.F. 1986. *Point Processes and their Statistical Inference*. New York: Marcel Dekker.

Lipster, R.S. and Shiryayev, A.N. 1978. *Statistics of Random Processes*, I and II. Berlin: Springer-Verlag.

Metivier, M. and Pellaumail, J. 1980. *Stochastic Integration*. New York: Academic Press.

Shiryayev, A.N. 1981. Martingales: recent developments, results and applications. *International Statistical Review* 49, 199–233.

How to cite this article

Karr, Alan F. "martingales." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000094> doi:10.1057/9780230226203.1050

The New Palgrave Dictionary of Economics Online

Marx, Karl Heinrich (1818 – 1883)

Ernest Mandel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes the methodology and economics of Karl Marx. After a brief account of his life, it deals with his historical materialism, and then his labour theory of value, his theories of rent, money, surplus value, and crises, his account of the laws of motion of the capitalism mode of production, and his and Engels's conception of the economy of post-capitalist societies.

Keywords

accumulation of capital; Asiatic mode of production; Böhm-Bawerk, E. von; Bukharin, N.; capitalist mode of production; class; commodity theory of money; communism; concentration (centralization) of capital; constant and variable capital; credit cycles; economic crisis; economic laws; economic science; Engels, F.; exploitation; feudalism; Hilferding, R.; historical materialism; ideology; invisible hand; iron law of wages; labour power; labour theory of value; land rent; Lassalle, F.; laws of motion of capitalism; Lenin, V.; market value; Marx, K.; mobility of capital; modes of production; monopoly; organic composition of capital; overproduction; paper money; post-capitalist society; prices of production; proletariat; rate of profit; rent; reserve army of labour; Ricardo, D.; social polarization; socialism; Sraffa, P.; surplus value; technological change; trade unions; transformation problem; underconsumptionism

Article

Karl Marx was born on 5 May 1818, the son of the lawyer Heinrich Marx and Henriette Pressburg. His father was descended from an old family of Jewish rabbis, but was himself a liberal admirer of the Enlightenment and not religious. He converted to Protestantism a few years before Karl was born to escape restrictions still imposed upon Jews in Prussia. His mother was of Dutch-Jewish origin.

Life and work

Karl Marx studied at the *Friedrich-Wilhelm Gymnasium* in Trier, and at the universities of Bonn and Berlin. His doctoral thesis, *Differenz der demokritischen und epikurischen Naturphilosophie*, was accepted at the University of Jena on 15 April 1841. In 1843 he married Jenny von Westphalen, daughter of Baron von Westphalen, a high Prussian government official.

Marx's university studies covered many fields, but centred around philosophy and religion. He frequented the circle of the more radical followers of the great philosopher Hegel, befriended one of their main representatives, Bruno Bauer, and was especially influenced by the publication in 1841 of Ludwig Feuerbach's *Das Wesen des Christentums* (The Nature of Christianity). He had intended to teach philosophy at the university, but that quickly proved to be unrealistic. He then turned towards journalism, both to propagandize his ideas and to gain a livelihood. He became editor of the *Rheinische Zeitung*, a liberal newspaper of Cologne, in May 1942. His interest turned more and more to political and social questions, which he treated in an increasingly radical way. The paper was banned by the Prussian authorities a year later.

Karl Marx then planned to publish a magazine called *Deutsch-Französische Jahrbücher* in Paris, in order to escape Prussian censorship and to be more closely linked and identified with the real struggles for political and social emancipation which, at that time, were centred around France. He emigrated to Paris with his wife and met there his lifelong friend Friedrich Engels.

Marx had become critical of Hegel's philosophical political system, a criticism which would lead to his first major work, Zur Kritik des *Hegelschen Rechtsphilosophie* (A Critique of Hegel's Philosophy of Right). Intensively studying history and political economy during his stay in Paris, he became strongly influenced by socialist and working-class circles in the French capital. With his ‘Paris Manuscripts’ (*Oekonomisch-philosophische Manuskripte*, 1844), he definitely became a communist, i.e. a proponent of collective ownership of the means of production.

He was expelled from France at the beginning of 1845 through pressure from the Prussian embassy and migrated to Brussels. His definite turn towards historical materialism (see below) would occur with his manuscript *Die Deutsche Ideologie* (1845–6) culminating in the eleven *Theses on Feuerbach*, written together with Engels but never published during his lifetime.

This led also to a polemical break with the most influential French socialist of that period, Proudhon, expressed in the only book Marx would write in French, *Misère de la Philosophie* (1846).

Simultaneously he became more and more involved in practical socialist politics, and started to work with the Communist League, which asked Engels and himself to draft their declaration of principle. This is the origin of the *Communist Manifesto* (1848), *Manifest der Kommunistischen Partei* (1848).

As soon as the revolution of 1848 broke out, he was in turn expelled from Belgium and went first to France, then, from April 1848 on, to Cologne. His political activity during the German revolution of 1848 centred around the publication of the daily paper *Neue Rheinische Zeitung*, which enjoyed wide popular support. After the victory of the Prussian counter-revolution, the paper was banned in May 1849 and Marx was expelled from Prussia. He never succeeded in recovering his citizenship. Marx emigrated to London, where he would stay, with short interruptions, till the end of his life. For fifteen years, his time would be mainly taken up with economic studies, which would lead to the publication first of *Zur Kritik der Politischen Oekonomie* (1859) and later of *Das Kapital*, Vol. I (1867). He spent long hours at the British Museum, studying the writings of all the major economists, as well as the government Blue Books, Hansard and many other contemporary sources on social and economic conditions in Britain and the world. His reading also covered technology, ethnology and anthropology, besides political economy

and economic history; many notebooks were filled with excerpts from the books he read.

But while the activity was mainly studious, he never completely abandoned practical politics. He first hoped that the Communist League would be kept alive, thanks to a revival of revolution. When this did not occur, he progressively dropped out of emigré politics, but not without writing a scathing indictment of French counter-revolution in *Der 18. Brumaire des Louis Bonaparte* (1852), which was in a certain sense the balance sheet of his political activity and an analysis of the 1848–52 cycle of revolution and counter-revolution. He would befriend British trade-union leaders and gradually attempt to draw them towards international working class interests and politics. These efforts culminated in the creation of the International Working Men's Association (1864) – the so-called First International – in which Marx and Engels would play a leading role, politically as well as organizationally.

It was not only his political interest and revolutionary passion that prevented Marx from becoming an economist pure and simple. It was also the pressure of material necessity. Contrary to his hope, he never succeeded in earning enough money from his scientific writings to sustain himself and his growing family. He had to turn to journalism to make a living. He had initial, be it modest, success in this field, when he became European correspondent of the *New York Daily Tribune* in the summer of 1851. But he never had a regular income from that collaboration, and it ended after ten years.

So the years of his London exile were mainly years of great material deprivation and moral suffering. Marx suffered greatly from the fact that he could not provide a minimum of normal living conditions for his wife and children, whom he loved deeply. Bad lodgings in cholera-stricken Soho, insufficient food and medical care, led to a chronic deterioration of his wife's and his own health and to the death of several of their children; that of his oldest son Edgar in 1855 struck him an especially heavy blow. Of his seven children, only three daughters survived, Jenny, Laura and Eleanor (Tussy). All three were very gifted and would play a significant role in the international labour movement, Eleanor in Britain, Jenny and Laura in France (where they married the socialist leaders Longuet and Lafargue).

During this long period of material misery, Marx survived thanks to the financial and moral support of his friend Friedrich Engels, whose devotion to him stands as an exceptional example of friendship in the history of science and politics. Things started to improve when Marx came into his mother's inheritance; when the first independent working-class parties (followers of Lassalle on the one hand, of Marx and Engels on the other) developed in Germany, creating a broader market for his writings; when the IWMA became influential in several European countries, and when Engels’ financial conditions improved to the point where he would sustain the Marx family on a more regular basis.

The period 1865–71 was one in which Marx's concentration on economic studies and on the drafting of *Das Kapital* was interrupted more and more by current political commitments to the IWMA, culminating in his impassioned defence of the Paris Commune (*Der Bürgerkrieg in Frankreich* [*The Civil War in France*] 1871). But the satisfaction of being able to participate a second time in a real revolution – be it only vicariously – was troubled by the deep divisions inside the IMWA, which led to the split with the anarchists grouped around Michael Bakunin.

Marx did not succeed in finishing a final version of *Das Kapital* vols II and III, which were published posthumously, after extensive editing, by Engels. It remains controversial whether he intended to add two more volumes to these, according to an initial plan. More than 25 years after the death of Marx, Karl Kautsky edited what is often called vol. IV of *Das Kapital*, his extensive critique of other economists: *Theorien über den Mehrwert* (*Theories of Surplus Value*).

Marx's final years were increasingly marked by bad health, in spite of slightly improved living

conditions. Bad health was probably the main reason why the final version of vols II and III of *Capital* could not be finished. Although he wrote a strong critique of the Programme which was adopted by the unification congress (1878) of German social democracy (*Kritik Des Gothaer Program*), he was heartened by the creation of that united working-class party in his native land, by the spread of socialist organizations throughout Europe, and by the growing influence of his ideas in the socialist movement. His wife fell ill in 1880 and died the next year. This came as a deadly blow to Karl Marx, who did not survive her for long. He himself died in London on 14 March 1883.

Historical materialism

Outside his specific economic theories, Marx's main contribution to the social sciences has been his theory of historical materialism. Its starting point is anthropological. Human beings cannot survive without social organization. Social organization is based upon social labour and social communication. Social labour always occurs within a given framework of specific, historically determined, social relations of production. These social relations of production determine in the last analysis all other social relations, including those of social communication. It is social existence which determines social consciousness and not the other way around.

Historical materialism posits that relations of production which become stabilized and reproduce themselves are structures which can no longer be changed gradually, piecemeal. They are modes of production. To use Hegel's dialectical language, which was largely adopted (and adapted) by Marx: they can only change qualitatively through a complete social upheaval, a social revolution or counter-revolution. Quantitative changes can occur within modes of production, but they do not modify the basic structure. In each mode of production, a given set of relations of production constitutes the basis (infrastructure) on which is erected a complex superstructure, encompassing the state and the law (except in classless society), ideology, religion, philosophy, the arts, morality etc.

Relations of production are the sum total of social relations which human beings establish among themselves in the production of their material lives. They are therefore not limited to what actually happens at the point of production. Humankind could not survive, i.e. produce, if there did not exist specific forms of circulation of goods, e.g. between producing units (circulation of tools and raw materials) and between producing units and consumers. A priori allocation of goods determines other relations of production than does allocation of goods through the market. Partial commodity production (what Marx calls ‘simple commodity production’ or ‘petty commodity production’ – ‘einfache Warenproduktion’) also implies other relations of production than does generalized commodity production.

Except in the case of classless societies, modes of production, centred around prevailing relations of production, are embodied in specific class relations which, in the last analysis, overdetermine relations between individuals.

Historical materialism does not deny the individual's free will, his attempts to make choices concerning his existence according to his individual passions, his interests as he understands them, his convictions, his moral options etc. What historical materialism does state is: (1) that these choices are strongly predetermined by the social framework (education, prevailing ideology and moral ‘values’, variants of behaviour limited by material conditions etc); (2) that the outcome of the collision of millions of different passions, interests and options is essentially a phenomenon of social logic and not of individual

psychology. Here, class interests are predominant.

There is no example in history of a ruling class not trying to defend its class rule, or of an exploited class not trying to limit (and occasionally eliminate) the exploitation it suffers. So outside classless society, the class struggle is a permanent feature of human society. In fact, one of the key theses of historical materialism is that 'the history of humankind is the history of class struggles' (Marx, *Communist Manifesto*, 1848).

The immediate object of class struggle is economic and material. It is a struggle for the division of the social product between the direct producers (the productive, exploited class) and those who appropriate what Marx calls the social surplus product, the residuum of the social product once the producers and their offspring are fed (in the large sense of the word; i.e. the sum total of the consumer goods consumed by that class) and the initial stock of tools and raw materials is reproduced (including the restoration of initial fertility of the soil). The ruling class functions as ruling class essentially through the appropriation of the social surplus product. By getting possession of the social surplus product, it acquires the means to foster and maintain most of the superstructural activities mentioned above; and by doing so, it can largely determine their function – to maintain and reproduce the given social structure, the given mode of production – and their contents.

We say 'largely determine' and not 'completely determine'. First, there is an 'immanent dialectical', i.e. an autonomous movement, of each specific superstructural sphere of activity. Each generation of scientists, artists, philosophers, theologists, lawyers and politicians finds a given *corpus* of ideas, forms, rules, techniques, ways of thinking, to which it is initiated through education and current practice, etc. It is not forced to simply continue and reproduce these elements. It can transform them, modify them, change their interconnections, even negate them. Again: historical materialism does not deny that there is a specific history of science, a history of art, a history of philosophy, a history of political and moral ideas, a history of religion etc., which all follow their own logic. It tries to *explain* why a certain number of scientific, artistic, philosophical, ideological, juridical changes or even revolutions occur at a given time and in given countries, quite different from other ones which occurred some centuries earlier elsewhere. The nexus of these 'revolutions' with given historical periods is a nexus of class interests.

Second, each social formation (i.e. a given country in a given epoch) while being characterized by predominant relations of production (i.e. a given mode or production at a certain phase of its development) includes different relations of production which are largely remnants of the past, but also sometimes nuclei of future modes of production. Thus there exists not only the ruling class and the exploited class characteristic of that prevailing mode of production (capitalists and wage earners under capitalism). There also exist remnants of social classes which were predominant when other relations of production prevailed and which, while having lost their hegemony, still manage to survive in the interstices of the new society. This is for example the case with petty commodity producers (peasants, handicraftsmen, small merchants), semifeudal landowners, and even slave-owners, in many already predominantly capitalist social formations throughout the 19th and part of the 20th centuries. Each of these social classes has its own ideology, its own religious and moral values, which are intertwined with the ideology of the hegemonic ruling class, without becoming completely absorbed by that.

Third, even after a given ruling class (e.g. the feudal or semi-feudal nobility) has disappeared as a ruling class, its ideology can survive through sheer force of social inertia and routine (custom). The survival of traditional *ancien régime* catholic ideology in France during a large part of the 19th century, in spite of the sweeping social, political and ideological changes ushered in by the French revolution, is an

illustration of that rule.

Finally, Marx's statement that the ruling ideology of each epoch is the ideology of the ruling class – another basic tenet of historical materialism – does not express more than it actually says. It implies that other ideologies can exist side by side with that ruling ideology without being hegemonic. To cite the most important of these occurrences: exploited and (or) oppressed social classes can develop their own ideology, which will start to challenge the prevailing hegemonic one. In fact, an ideological class struggle accompanies and sometimes even precedes the political class struggle properly speaking. Religious and philosophical struggles preceding the classical bourgeois revolutions; the first socialist critiques of bourgeois society preceding the constitution of the first working-class parties and revolutions, are examples of that type.

The class struggle has been up to now the great motor of history. Human beings make their own history. No mode of production can be replaced by another one without deliberate actions by large social forces, i.e., without social revolutions (or counter-revolutions). Whether these revolutions or counter-revolutions actually lead to the long-term implementation of deliberate projects of social reorganization is another matter altogether. Very often, their outcome is to a large extent different from the intention of the main actors.

Human beings act consciously, but they can act with false consciousness. They do not necessarily understand why they want to realize certain social and (or) political plans, why they want to maintain or to change economic or juridical institutions; and especially, they rarely understand in a scientific sense the laws of social change, the material and social preconditions for successfully conserving or changing such institutions. Indeed, Marx claims that only with the discovery of the main tenets of historical materialism have we made a significant step forward towards understanding these laws, without being able to predict ‘all’ future developments of society.

Social change, social revolutions and counter-revolutions are furthermore occurring within determined material constraints. The level of development of the productive forces – essentially tools and human skills, including their effects upon the fertility of the soil – limits the possibilities of institutional change. Slave labour has shown itself to be largely incompatible with the factory system based upon contemporary machines. Socialism would not be durably built upon the basis of the wooden plough and the potter's wheel. A social revolution generally widens the scope for the development of the productive forces and leads to social progress in most fields of human activity in a momentous way. Likewise, an epoch of deep social crisis is ushered in when there is a growing conflict between the prevailing mode of production (i.e. the existing social order) on the one hand, and the further development of the productive forces on the other. Such a social crisis will then manifest itself on all major fields and social activity: politics, ideology, morals and law, as well as in the realm of the economic life properly speaking.

Historical materialism thereby provides a measuring stick for human progress: the growth of the productive forces, measurable through the growth of the average productivity of labour, and the number, longevity and skill of the human species. This measuring stick in no way abstracts from the natural preconditions for human survival and human growth (in the broadest sense of the concept). Nor does it abstract from the conditional and partial character of such progress, in terms of social organization and individual alienation.

In the last analysis, the division of society into antagonistic social classes reflects, from the point of view of historical materialism, an inevitable limitation of human freedom. For Marx and Engels, the real measuring rod of human freedom, i.e. of human wealth, is not ‘productive labour’; this only creates the

material pre-condition for that freedom. The real measuring rod is leisure time, not in the sense of ‘time for doing nothing’ but in the sense of time freed from the iron necessity to produce and reproduce material livelihood, and therefore disposable for all-round and free development of the individual talents, wishes, capacities, potentialities, of each human being.

As long as society is too poor, as long as goods and services satisfying basic needs are too scarce, only part of society can be freed from the necessity to devote most of its life to ‘work for a livelihood’ (i.e. of forced labour, in the anthropological/sociological sense of the word, that is in relation to desires, aspirations and talents, not to a juridical status of bonded labour). That is essentially what represents the freedom of the ruling classes and their hangers-on, who are ‘being paid to think’, to create, to invent, to administer, because they have become free from the obligation to bake their own bread, weave their own clothes and build their own houses.

Once the productive forces are developed far enough to guarantee all human beings satisfaction of their basic needs by ‘productive labour’ limited to a minor fraction of lifetime (the half work-day or less), then the material need of the division of society in classes disappears. Then, there remains no objective basis for part of society to monopolize administration, access to information, knowledge, intellectual labour. For that reason, historical materialism explains both the reasons why class societies and class struggles arose in history, and why they will disappear in the future in a classless society of democratically self-administering associated producers.

Historical materialism therefore contains an attempt at explaining the origin, the functions, and the future withering away of the state as a specific institution, as well as an attempt to explain politics and political activity in general, as an expression of social conflicts centred around different social interest (mainly, but not only, those of different social classes; important fractions of classes, as well as non-class social groupings, also come into play).

For Marx and Engels, the state is not existent with human society as such, or with ‘organized society’ or even with ‘civilized society’ in the abstract; neither is it the result of any voluntarily concluded ‘social contract’ between individuals. That state is the sum total of apparatuses, i.e. special groups of people separate and apart from the rest (majority) of society, that appropriate to themselves functions of a repressive or integrative nature which were initially exercised by all citizens. This process of alienation occurs in conjunction with the emergence of social classes. The state is an instrument for fostering, conserving and reproducing a given class structure, and not a neutral arbiter between antagonistic class interests.

The emergence of a classless society is therefore closely intertwined, for adherents to historical materialism, with the process of withering away of the state, i.e. of gradual devolution to the whole of society (self-management, self-administration) of all specific functions today exercised by special apparatuses, i.e. of the dissolution of these apparatuses. Marx and Engels visualized the dictatorship of the proletariat, the last form of the state and of political class rule, as an instrument for assuring the transition from class society to classless society. It should itself be a state of a special kind, organizing its own gradual disappearance.

We said above that, from the point of view of historical materialism, the immediate object of class struggle is the division of the social product between different social classes. Even the political class struggle in the final analysis serves that main purpose; but it also covers a much broader field of social conflicts. As all state activities have some bearing upon the relative stability or instability of a given social formation, and the class rule to which it is submitted, the class struggle can extend to all fields of

politics, from foreign policy to educational problems and religious conflicts. This has of course to be proven through painstaking analysis, and not proclaimed as an axiom or a revealed truth. When conducted successfully, such exercises in class analysis and class definition of political, social and even literary struggles become impressive works of historical explanation, as for example Marx's *Class Struggles in France 1848–50*, Engels' *The German Peasant War*, Franz Mehring's *Die Lessing-Legende*, Trotsky's *History of the Russian Revolution*, etc.

Marx's economic theory – general approach and influence

A general appraisal of Marx's method of economic analysis is called for prior to an outline of his main economic theories (theses and hypotheses).
Marx is distinct from most important economists of the 19th and 20th centuries in that he does not consider himself at all an 'economist' pure and simple.
The idea that 'economic science' as a special science completely separate from sociology, history, anthropology etc. cannot exist, underlies most of his economic analysis. Indeed, historical materialism is an attempt at unifying all social sciences, if not all sciences about humankind, into a single 'science of society'.
For sure, within the framework of this general 'science of society', economic phenomena could and should be submitted to analysis as specific phenomena. So economic theory, economical science, have a definite autonomy after all; but is only a partial and relative one.
Probably the best formula for characterizing Marx's economic theory would be to call it an endeavour to explain the social economy. This would be true in a double sense. For Marx, there are no eternal economic laws, valid in every epoch of human prehistory and history. Each mode of production has its own specific economic laws, which lose their relevance once the general social framework has fundamentally changed. For Marx likewise, there are no economic laws separate and apart from specific relations between human beings, in the primary (but not only, as already summarized) social relations of production. All attempts to reduce economic problems to purely material, objective ones, to relations between things, or between things and human beings, would be considered by Marx as manifestations of mystification, of false consciousness, expressing itself through the attempted reification of human relations. Behind relations between things, economic science should try to discover the specific relations between human beings which they hide. Real economic science has therefore also a demystifying function compared to vulgar 'economics', which takes a certain number of 'things' for granted without asking the question: Are they really only what they appear to be? From where do they originate? What explains these appearances? What lies behind them? Where do they lead? How could they (will they) disappear? *Problemblindheit*, the refusal to see that facts are generally more problematic than they appear at first sight, is certainly not a reproach one could address to Marx's economic thought.
Marx's economic analysis is therefore characterized by a strong ground current of *historical relativism*, with a strong recourse to the genetical and evolutionary method of thinking (that is why the parallel with Darwin has often been made, sometimes in an excessive way). The formula 'genetic structuralism' has also been used in relation to Marx's general approach to economic analysis. Be that as it may, one could state that Marx's economic theory is essentially geared to the discovery of specific 'laws of motion' for successive modes of production. While his theoretical effort has been mainly centred around the discovery of these laws of motion for capitalist society, his work contains indications of such laws –

different ones, to be sure – for precapitalist and postcapitalist social formations too.

The main link between Marx's sociology and anthropology on the one hand, and his economic analysis on the other, lies in the key role of social labour as the basic anthropological feature underlying all forms of social organization. Social labour can be organized in quite different forms, thereby giving rise to quite different economic phenomena (‘facts’). Basically different forms of social labour organization lead to basically different sets of economic institutions and dynamics, following basically different logics (obeying basically different ‘laws of motion’).

All human societies must assure the satisfaction of a certain number of basic needs, in order to survive and reproduce themselves. This leads to the necessity of establishing some sort of equilibrium between socially recognized needs, i.e. current consumption and current production. But this abstract banality does not tell us anything about the concrete way in which social labour is organized in order to achieve that goal.

Society can recognize all individual labour as *immediately social labour*. Indeed, it does so in innumerable primitive tribal and village communities, as it does in the contemporary *kibbutz*. Directly social labour can be organized in a despotic or in a democratic way, through custom and superstition as well as through an attempt at applying advanced science to economic organization; but it will always be immediately recognized social labour, in as much as it is based upon *a prior* assignment of the producers to their specific work (again: irrespective of the form this assignation takes, whether it is voluntary or compulsory, despotic or simply through custom etc.).

But when social decision-taking about work assignation (and resource allocation closely tied to it) is fragmented into different units operating independently from each other – as a result of private control (property) of the means of production, in the economic and not necessarily the juridical sense of the word – then social labour in turn is fragmented into private labours which are not automatically recognized as socially necessary ones (whose expenditure is not automatically compensated by society). Then the private producers have to exchange parts or all of their products in order to satisfy some or all of their basic needs. Then these products become commodities. The economy becomes a (partial or generalized) market economy. Only by measuring the results of the sale of his products can the producer (or owner) ascertain what part of his private labour expenditure has been recognized (compensated) as social labour, and what part has not.

Even if we operate with such simple analytical tools as ‘directly social labour’, ‘private labour’, ‘socially recognized social labour’, we have to make quite an effort at abstracting from immediately apparent phenomena in order to understand their relevance for economic analysis. This is true for all scientific analysis, in natural as well as in social sciences. Marx's economic analysis, as presented in his main books, has not been extremely popular reading; but then, there are not yet so many scientists in these circumstances. This has nothing to do with any innate obscurity of the author, but rather with the nature of scientific analysis as such.

The relatively limited number of readers of Marx's economic writings (the first English paperback edition of *Das Kapital* appeared only in 1974!) is clearly tied to Marx's scientific rigour, his effort at a systematic and all-sided analysis of the phenomena of the capitalist economy.

But while his economic analysis lacked popularity, his political and historical projections became more and more influential. With the rise of independent working-class mass parties, an increasing number of these proclaimed themselves as being guided or influenced by Marx, at least in the epoch of the Second and the Third Internationals, roughly the half century from 1890 till 1940. Beginning with the Russian

revolution of 1917, a growing number of governments and of states claimed to base their policies and constitutions on concepts developed by Marx. (Whether this was legitimate or not is another question.) But the fact itself testifies to Marx's great influence on contemporary social and political developments, evolutionary and revolutionary alike.

Likewise, his diffused influence on social science, including academic economic theory, goes far beyond general acceptance or even substantial knowledge of his main writings. Some key ideas of historical materialism and of economic analysis which permeate his work – e.g. that economic interests to a large extent influence, if not determine, political struggles; that historic evolution is linked to important changes in material conditions; that economic crises (‘the business cycle’) are unavoidable under conditions of capitalist market economy – have become near-platitudes. It is sufficient to notice how major economists and historians strongly denied their validity throughout the 19th century and at least until the 1920s, to understand how deep has been Marx's influence on contemporary social science in general.

Marx's labour theory of value

As an economist, Marx is generally situated in the continuity of the great classical school of Adam Smith and Ricardo. He obviously owes a lot to Ricardo, and conducts a current dialogue with that master in most of his mature economic writings.

Marx inherited the labour theory of value from the classical school. Here the continuity is even more pronounced; but there is also a radical break. For Ricardo, labour is essentially a numeraire, which enables a common computation of labour and capital as basic elements of production costs. For Marx, labour is value. Value is nothing but that fragment of the total labour potential existing in a given society in a certain period (e.g. a year or a month) which is used for the output of a given commodity, at the average social productivity of labour existing then and there, divided by the total number of these commodities produced, and expressed in hours (or minutes), days, weeks, months of labour.

Value is therefore essentially a social, objective and historically relative category. It is social because it is determined by the overall result of the fluctuating efforts of each individual producer (under capitalism: of each individual firm or factory). It is objective because it is given, once the production of a given commodity is finished and is thus independent from personal (or collective) valuations of customers on the market place; and it is historically relative because it changes with each important change (progress or regression) of the average productivity of labour in a given branch of output, including in agriculture and transportation.

This does not imply that Marx's concept of value is in any way completely detached from consumption. It only means that the feedback of consumers’ behaviour and wishes upon value is always mediated through changes in allocation of labour inputs in production, labour seen as subdivided into living labour and dead (dated) labour, i.e. tools and raw materials. The market emits signals to which the producing units react. Value changes after these reactions, not before them. Market price changes can of course occur prior to changes in value. In fact, changes in market prices are among the key signals which can lead to changes in labour allocation between different branches of production, i.e. to changes in labour quantities necessary to produce given commodities. But then, for Marx, values determine prices only basically and in the medium-term sense of the word. This determination only appears clearly as an explication of *medium and long-term price movements*. In the shorter run, prices fluctuate around values

as axes. Marx never intended to negate the operation of market laws, of the law of supply and demand, in determining these short-term fluctuations.

The ‘law of value’ is but Marx's version of Adam Smith's ‘invisible hand’. In a society dominated by private labour, private producers and private ownership of productive inputs, it is this ‘law of value’, an objective economic law operating behind the backs of all people, all ‘agents’ involved in production and consumption, which, in the final analysis, regulates the economy, determines what is produced and how it is produced (and therefore also what can be consumed). The ‘law of value’ regulates the exchange between commodities, according to the quantities of socially necessary abstract labour they embody (the quantity of such labour spent in their production). Through regulating the exchange between commodities, the ‘law of value’ also regulates, after some interval, the distribution of society's labour potential and of society's non-living productive resources between different branches of production. Again, the analogy with Smith's ‘invisible hand’ is striking.

Marx's critique of the ‘invisible hand’ concept does not dwell essentially on the analysis of how a market economy actually operates. It would above all insist that this operation is not eternal, not immanent in ‘human nature’, but created by specific historical circumstances, a product of a special way of social organization, and due to disappear at some stage of historical evolution as it appeared during a previous stage. And it would also stress that this ‘invisible hand’ leads neither to the maximum of economic growth nor to the optimum of human wellbeing for the greatest number of individuals, i.e. it would stress the heavy economic and social price humankind had to pay, and is still currently paying, for the undeniable progress the market economy produced at a given stage of historical evolution.

The formula ‘quantities of abstract human labour’ refers to labour seen strictly as a fraction of the total labour potential of a given society at a given time, say a labour potential of 2 billion hours a year (1 million potential producers supposedly capable of working each 2000 hours a year). It therefore implies making abstraction of the specific trade or occupation of a given male or female producer, the product of a day's work of a weaver not being worth less or more than that of a peasant, a miner, a housebuilder, a milliner or a seamstress. At the basis of that concept of ‘abstract human labour’ lies a social condition, a specific set of social relations of production, in which small independent producers are essentially equal. Without that equality, social division of labour, and therefore satisfaction of basic consumers’ needs, would be seriously endangered under that specific organizational set-up of the economy. Such an equality between small commodity owners and producers is later transformed into an equality between owners of capital under the capitalist mode of production.

But the concept of homogeneity of productive human labour, underlying that of ‘abstract human labour’ as the essence of value, does not imply a negation of the difference between skilled and unskilled labour. Again: a negation of that difference would lead to breakdown of the necessary division of labour, as would any basic heterogeneity of labour inputs in different branches of output. It would then not pay to acquire skills: most of them would disappear. So Marx's labour theory of value, in an internally coherent way, leads to the conclusion that one hour of skilled labour represents more value than one hour of unskilled labour, say represents the equivalent of 1.5 hours of unskilled labour. The difference would result from the imputation of the labour it costs to acquire the given skill. While an unskilled labourer would have a labour potential of 120,000 hours during his adult life, a skilled labourer would only have a labour potential of 80,000 hours, 40,000 hours being used for acquiring, maintaining and developing his skill. Only if one hour of skilled labour embodies the same value of 1.5 hours of unskilled labour, will the equality of all ‘economic agents’ be maintained under these circumstances, i.e. will it ‘pay’

economically to acquire a skill.

Marx himself never extensively dwelled on this solution of the so-called *reduction problem*. This remains indeed one of the most obscure parts of his general economic theory. It has led to some, generally rather mild, controversy. Much more heat has been generated by another facet of Marx's labour theory of value, the so-called *transformation problem*. Indeed, from Böhm-Bawerk writing a century ago till the recent contributions of Sraffa (1960) and Steedman (1977), the way Marx dealt with the transformation of values into ‘prices of production’ in *Capital* Vol. III has been considered by many of his critics as the main problem of his ‘system’, including being a reason to reject the labour theory of value out of hand.

The problem arises out of the obvious modification in the functioning of a market economy when *capitalist* commodity production substitutes itself for *simple* commodity production. In simple commodity production, with generally stable technology and stable (or easily reproduceable) tools, living labour is the only variable of the quantity and subdivision of social production. The mobility of labour is the only dynamic factor in the economy. As Engels pointed out in his Addendum to *Capital* Vol. III (Marx, *g*, pp. 1034–7) in such an economy, commodities would be exchanged at prices which would be immediately proportional to values, to the labour inputs they embody.

But under the capitalist mode of production, this is no longer the case. Economic decision-taking is not in the hands of the direct producers. It is in the hands of the capitalist entrepreneurs in the wider sense of the word (bankers – distributors of credit – playing a key role in that decision-taking, besides entrepreneurs in the productive sector properly speaking). Investment decisions, i.e. decisions for creating, expanding, reducing or closing enterprises, determine economic life. It is the *mobility of capital* and not the mobility of labour which becomes the motive force of the economy. Mobility of labour becomes essentially an epiphenomenon of the mobility of capital.

Capitalist production is production for profit. Mobility of capital is determined by existing or expected profit differentials. Capital leaves branches (countries, regions) with lower profits (or profit expectations) and flows towards branches (countries, regions) with higher ones. These movements lead to an equalization of the rate of profit between different branches of production. But approximately equal returns on all invested capital (at least under conditions of prevailing ‘free competition’) coexist with unequal proportions of inputs of labour in these different branches. So there is a disparity between the direct value of a commodity and its ‘price of production’, that ‘price of production’ being defined by Marx as the sum of production costs (costs of fixed capital and raw materials plus wages) and the average rate of profit multiplied with the capital spent in the given production.

The so-called ‘transformation problem’ relates to the question of whether a relation can nevertheless be established between value and these ‘prices of production’, what is the degree of coherence (or incoherence) of the relation with the ‘law of value’ (the labour theory of value in general), and what is the correct quantitative way to express that relation, if it exists.

We shall leave aside here the last aspect of the problem, to which extensive analysis has recently been devoted (Mandel and Freeman, 1984). From Marx's point of view, there is no incoherence between the formation of ‘prices of production’ and the labour theory of value. Nor is it true that he came upon that alleged difficulty when he started to prepare *Capital* III, i.e. to deal with capitalist competition, as several critics have argued (see e.g. Joan Robinson, 1942). In fact, his solution of the transformation problem is already present in the *Grundrisse* (Marx, d), before he even started to draft *Capital* Vol. I.

The sum total of value produced in a given country during a given span of time (e.g. one year) is

determined by the sum total of labour-inputs. Competition and movements of capital cannot change that quantity. The sum total of values equals the sum total of ‘prices of production’. The only effect of capital competition and capital mobility is to *redistribute* that given sum – and this through a redistribution of surplus value (see below) – between different capitals, to the benefit of some and at the expense of others.

Now this redistribution does not occur in a haphazard or arbitrary way. Essentially value (surplus-value) is transferred from technically less advanced branches to technologically more advanced branches. And here the concept of ‘quantities of socially necessary labour’ comes into its own, under the conditions of constant revolutions of productive technology that characterize the capitalist mode of production. Branches with lower than average technology (organic composition of capital, see below) can be considered as wasting socially necessary labour. Part of the labour spent in production in their realm is therefore not compensated by society. Branches with higher than average technology (organic composition of capital) can be considered to be economizing social labour; their labour inputs can therefore be considered as more intensive than average, embodying more value. In this way, the transfer of value (surplus-value) between different branches, far from being in contradiction with the law of value, is precisely the way it operates and should operate under conditions of ‘capitalist equality’, given the pressure of rapid technological change.

As to the logical inconsistency often supposedly to be found in Marx's method of solving the ‘transformation problem’ – first advanced by von Bortkiewicz (1907) – it is based upon a misunderstanding in our opinion. It is alleged that in his ‘transformation schemas’ (or tables) (Marx, *g*, pp. 255–6) Marx calculates inputs in ‘values’ and outputs in ‘prices of production’, thereby omitting the feedback effect of the latter on the former. But that feedback effect is unrealistic and unnecessary, once one recognizes that inputs are essentially data. Movements of capital posterior to the purchase of machinery or raw materials, including ups and downs of prices of finished products produced with these raw materials, cannot lead to a change in prices and therefore of profits of the said machinery and raw materials, on sales which have already occurred. What critics present as an inconsistency between ‘values’ and ‘prices of production’ is simply a recognition of *two different time-frameworks* (cycles) in which the equalization of the rate of profit has been achieved, a first one for inputs, and a second, later one for outputs.

Marx's theory of rent

The labour theory of value defines value as the socially necessary quantity of labour determined by the average productivity of labour of each given sector of production. But these values are not mathematically fixed data. They are simply the expression of a *process* going on in real life, under capitalist commodity production. So this average is only ascertained in the course of a certain time-span. There is a lot of logical argument and empirical evidence to advance the hypothesis that the normal time-span for essentially modifying the value of commodities is the business cycle, from one crisis of over-production (recession) to the next one.

Before technological progress and (or) better (more ‘rational’) labour organization etc. determines a more than marginal change (in general: decline) in the value of a commodity, and the crisis eliminates less efficient firms, there will be a coexistence of firms with various ‘individual values’ of a given commodity in a given branch of output, even assuming a single market price. So, in his step-for-step

approach towards explaining the immediate phenomena (facts of economic life) like prices and profits, by their essence, Marx introduces at this point of his analysis a new mediating concept, that of *market value* (Marx, *g*, ch. 10). The market value of a commodity is the ‘individual value’ of the firm, or a group of firms, in a given branch of production, around which the market price will fluctuate. That ‘market value’ is not necessarily the mathematical (weighted) average of labour expenditure of all firms of that branch. It can be below, equal or above that average, for a certain period (generally less than the duration of the business cycle, at least under ‘free competition’), according to whether social demand is saturated, just covered or to an important extent not covered by current output plus existing stocks. In these three cases respectively, the more (most) efficient firms, the firms of average efficiency, or even firms with labour productivity below average, will determine the market value of that given commodity. This implies that the more efficient firms enjoy *surplus profits* (profits over and above the average profit) in case 2 and 3 and that a certain number of firms work at less than average profit in all three cases, but especially in case 1.

The mobility of capital, i.e. normal capitalist competition, generally eliminates such situations after a certain lapse of time. But when that mobility of capital is impeded for long periods by either unavoidable scarcity (natural conditions not renewable or non-substitutable, like land and mineral deposits) or through the operation of institutional obstacles (private property of land and mineral resources forbidding access to available capital, except in exchange for payments over and above average profit), these surplus profits can be frozen and maintained for decades. They thus become *rents*, of which *ground rent* and *mineral rent* are the most obvious examples in Marx's time, extensively analysed in *Capital* vol. III (Marx, *g*, part 6).

Marx's theory of rent is the most difficult part of his economic theory, the one which has witnessed fewer comments and developments, by followers and critics alike, than other major parts of his ‘system’. But it is not obscure. And in contrast to Ricardo's or Rodbertus's theories of rent, it represents a straightforward application of the labour theory of value. It does not imply any emergence of ‘supplementary’ value (surplus value, profits) in the market, in the process of circulation of commodities, which is anathema to Marx and to all consistent upholders of the labour theory of value. Nor does it in any way suggest that land or mineral deposits ‘create’ value.

It simply means that in agriculture and mining less productive labour (as in the general case analysed above) determines the market value of food or minerals, and that therefore more efficient farms and mines enjoy surplus profits which Marx calls differential (land and mining) rent. It also means that as long as productivity of labour in agriculture is generally below the average of the economy as a whole (or more correctly: that the organic composition of capital, the expenditure in machinery and raw materials as against wages, is inferior in agriculture to that of industry and transportation), the sum-total of surplus-value produced in agriculture will accrue to landowners+capitalist farmers taken together, and will not enter the general process of (re)distribution of profit throughout the economy as a whole.

This creates the basis for a supplementary form of rent, over and above differential rent, rent which Marx calls absolute land rent. This is, incidentally, the basis for a long-term separation of capitalist landowners from entrepreneurs in farming or animal husbandry, distinct from feudal or semi-feudal landowners or great landowners under conditions of predominantly petty commodity production, or in the Asiatic mode of production, with free peasants.

The validity of Marx's theory of land and mining rents has been confirmed by historical evidence, especially in the 20th century. Not only has history substantiated Marx's prediction that, in spite of the

obstacle of land and mining rent, mechanization would end up by penetrating food and raw materials production too, as it has for a long time dominated industry and transportation, thereby causing a growing decline of differential rent (this has occurred increasingly in agriculture in the last 25–50 years, first in North America, and then in Western Europe and even elsewhere). It has also demonstrated that once the structural scarcity of food disappears, the institutional obstacle (private property) loses most of its efficiency as a brake upon the mobility of capital. Therefore the participation of surplus-value produced in agriculture in the general process of profit equalization throughout the economy cannot be prevented any more. Thereby absolute rent tends to wither away and, with it, the separation of land ownership from entrepreneurial farming and animal husbandry. It is true that farmers can then fall under the sway of the banks, but they do so as private owners of their land which becomes mortgaged, not as share-croppers or entrepreneurs renting land from separate owners.

On the other hand, the reappearance of structural scarcity in the realm of energy enabled the OPEC countries to multiply the price of oil by ten in the 1970s, i.e. to have it determined by the oilfields where production costs are the highest, thereby assuring the owners of the cheapest oil wells in Arabia, Iran, Libya, etc. of huge differential mineral rents.

Marx's theory of land and mineral rent can be easily extended into a general theory of rent, applicable to all fields of production where formidable difficulties of entry limit mobility of capital for extended periods of time. It thereby becomes the basis of a Marxist theory of monopoly and monopoly surplus profits, i.e. in the form of cartel rents (Hilferding, 1910) or of technological rent (Mandel, 1972). Lenin's and Bukharin's theories of surplus profit are based upon analogous but not identical reasoning (Bukharin, 1914, 1926; Lenin, 1917).

But in all these cases of general application of the Marxist theory of rent, the same caution should apply as Marx applied to his theory of land rent. By its very nature, capitalism, based upon private property, i. e. ‘many capitals’ – that is, competition – cannot tolerate any ‘eternal’ monopoly, a ‘permanent’ surplus profit deducted from the sum total of profits which is divided among the capitalist class as a whole. Technological innovations, substitution of new products for old ones including the field of raw materials and of food, will in the long run reduce or eliminate all monopoly situations, especially if the profit differential is large enough to justify huge research and investment outlays.

Marx's theory of money

In the same way as his theory of rent, Marx's theory of money is a straightforward application of the labour theory of value. As value is but the embodiment of socially necessary labour, commodities exchange with each other in proportion of the labour quanta they contain. This is true for the exchange of iron against wheat as it is true for the exchange of iron against gold or silver. Marx's theory of money is therefore in the first place a commodity theory of money. A given commodity can play the role of universal medium of exchange, as well as fulfil all the other functions of money, precisely because it is a commodity, i.e. because it is itself the product of socially necessary labour. This applies to the precious metals in the same way it applies to all the various commodities which, throughout history, have played the role of money.

It follows that strong upheavals in the ‘intrinsic’ value of the money-commodity will cause strong upheavals in the general price level. In Marx's theory of money, (market) prices are nothing but the expression of the value of commodities in the value of the money commodity chosen as a monetary

standard. If £1 sterling=$\frac{1}{10}$ ounce of gold, the formula ‘the price of 10 quarters of wheat is £1’ means that 10 quarters of wheat have been produced in the same socially necessary labour time as $\frac{1}{10}$ ounce of gold. A strong decrease in the average productivity of labour in gold mining (as a result for example of a depletion of the richer gold veins) will lead to a general depression of the average price level, all other things remaining equal. Likewise, a sudden and radical increase in the average productivity of labour in gold mining, through the discovery of new rich gold fields (California after 1848; the Rand in South Africa in the 1890s) or through the application of new revolutionary technology, will lead to a general increase in the price level of all other commodities.

Leaving aside short-term oscillations, the general price level will move in medium and long-term periods according to the relation between the fluctuations of the productivity of labour in agriculture and industry on the one hand, and the fluctuations of the productivity of labour in gold mining (if gold is the money-commodity), on the other.

Basing himself on that commodity theory of money. Marx therefore criticized as inconsistent Ricardo's quantity theory (Marx, *h*, part 2). But for exactly the same reason of a consistent application of the labour theory of value, the quantity of money in circulation enters Marx's economic analysis when he deals with the phenomenon of paper money (Marx, *c*).

As gold has an intrinsic value, like all other commodities, there can be no ‘gold inflation’, as little as there can be a ‘steel inflation’. Abstraction made of short-term price fluctuations caused by fluctuations between supply and demand, a persistent decline of the value of gold (exactly as for all other commodities) can only be the result of a persistent increase in the average productivity of labour in gold mining, and not of an ‘excess’ of circulation in gold. If the demand for gold falls consistently, this can only indirectly trigger off a decline in the value of gold through causing the closure of the least productive gold mines. But in the case of the money-commodity, such overproduction can hardly occur, given the special function of gold of serving as a universal reserve fund, nationally and internationally. It will always therefore find a buyer, be it not, of course, always at the same ‘prices’ (in Marx's economic theory, the concept of ‘price of gold’ is meaningless. As the price of a commodity is precisely its expression in the value of gold, the ‘price of gold’ would be the expression of the value of gold in the value of gold).

Paper money, bank notes, are a *money sign* representing a given quantity of the money-commodity. Starting from the above-mentioned example, a banknote of £1 represents $\frac{1}{10}$ ounce of gold. This is an objective ‘fact of life’, which no government or monetary authority can arbitrarily alter. It follows that any emission of paper money in excess of that given proportion will automatically lead to an increase in the general price level, always other things remaining equal. If £1 suddenly represents only $\frac{1}{20}$ ounce of gold, because paper money circulation has doubled without a significant increase in the total labour time spent in the economy, then the price level will tend to double too. The value of $\frac{1}{10}$ ounce of gold remains equal to the value of 10 quarters of wheat. But as $\frac{1}{10}$ ounce of gold is now represented by £2 in paper banknotes instead of being represented by £1, the price of wheat will move from £1 to £2 for 10 quarters (from two shillings to four shillings a quarter before the introduction of the decimal system).

This does not mean that in the case of paper money, Marx himself has become an advocate of a quantity theory of money. While there are obvious analogies between his theory of paper money and the quantity

theory, the main difference is the rejection by Marx of any mechanical automatism between the quantity of paper money emitted on the one hand, and the general dynamic of the economy (including on the price level) on the other.

In Marx's explanation of the movement of the capitalist economy in its totality, the formula *ceteris paribus* is meaningless. Excessive (or insufficient) emission of paper money never occurs in a vacuum. It always occurs at a given stage of the business cycle, and in a given phase of the longer-term historical evolution of capitalism. It is thereby always combined with given ups and downs of the rate of profit, of productivity of labour, of output, of market conditions (overproduction or insufficient production). Only in connection with these other fluctuations can the effect of paper money ‘inflation’ or ‘deflation’ be judged, including the effect on the general price level. The key variables are in the field of production. The key synthetic resultant is in the field of profit. Price movements are generally epiphenomena as much as they are signals. To untwine the tangle, more is necessary than a simple analysis of the fluctuations of the quantity of money. Only in the case of extreme runaway inflation of paper money would this be otherwise; and even in that border case, *relative* price movements (different degrees of price increases for different commodities) would still confirm that, in the last analysis, the law of value rules, and not the arbitrary decisions of the Central Bank, or any other authority controlling or emitting paper money.

Marx's theory of surplus-value

Marx himself considered his theory of surplus-value his most important contribution to the progress of economic analysis (Marx, *l*; letter to Engels of 24 August 1867). It is through this theory that the wide scope of his sociological and historical thought enables him simultaneously to place the capitalist mode of production in its historical context, and to find the roots of its inner economic contradictions and its laws of motion in the specific relations of production on which it is based.

As said before, Marx's theory of classes is based on the recognition that in each class society, part of society (the ruling class) appropriates the social surplus product. But that surplus product can take three essentially different forms (or a combination of them). It can take the form of straightforward unpaid surplus labour, as in the slave mode of production, early feudalism or some sectors of the Asian mode of production (unpaid corvée labour for the Empire). It can take the form of goods appropriated by the ruling class in the form of use-values pure and simple (the products of surplus labour), as under feudalism when feudal rent is paid in a certain amount of produce (produce rent) or in its more modern remnants, such as sharecropping. And it can take a money form, like money-rent in the final phases of feudalism, and capitalist profits. Surplus-value is essentially just that: the money form of the social surplus produce or, what amounts to the same, the money product of surplus labour. It has therefore a common root with all other forms of surplus product: unpaid labour.

This means that Marx's theory of surplus-value is basically a deduction (or residual) theory of the ruling classes’ income. The whole social product (the net national income) is produced in the course of the process of production, exactly as the whole crop is harvested by the peasants. What happens on the market (or through appropriation of the produce) is a distribution (or redistribution) of what already has been created. The surplus product, and therefore also its money form, surplus-value, is the residual of that new (net) social product (income) which remains after the producing classes have received their compensation (under capitalism: their wages). This ‘deduction’ theory of the ruling classes’ income is

thus *ipso facto* an exploitation theory. Not in the ethical sense of the word – although Marx and Engels obviously manifested a lot of understandable moral indignation at the fate of all the exploited throughout history, and especially at the fate of the modern proletariat – but in the economical one. The income of the ruling classes can always be reduced in the final analysis to the product of unpaid labour: that is the heart of Marx's theory of exploitation.

That is also the reason why Marx attached so much importance to treating *surplus-value as a general category*, over and above profits (themselves subdivided into industrial profits, bank profits, commercial profits etc.), interest and rent, which are all part of the total surplus product produced by wage labour. It is this general category which explains both the existence (the common interest) of the ruling class (all those who live off surplus value), and the origins of the class struggle under capitalism.

Marx likewise laid bare the economic mechanism through which surplus-value originates. As the basis of that economic mechanism is a huge social upheaval which started in Western Europe in the 15th century and slowly spread over the rest of the continent and all other continents (in many so-called underdeveloped countries, it is still going on to this day).

Through many concomitant economic (including technical), social, political and cultural transformations, the mass of the direct producers, essentially peasants and handicraftsmen, are separated from their means of production and cut off from free access to the land. They are therefore unable to produce their livelihood on their own account. In order to keep themselves and their families alive, they have to hire out their arms, their muscles and their brains, to the owners of the means of production (including land). If and when these owners have enough money capital at their disposal to buy raw materials and pay wages, they can start to organize production on a capitalist basis, using wage labour to transform the raw materials which they buy, with the tools they own, into finished products which they then automatically own too.

The capitalist mode of production thus presupposes that the producers' *labour power has become a commodity*. Like all other commodities, the commodity labour power has an exchange value and a use value. The exchange value of labour power, like the exchange value of all other commodities, is the amount of socially necessary labour embodied in it, i.e. its reproduction costs. This means concretely the value of all the consumer goods and services necessary for a labourer to work day after day, week after week, month after month, at approximately the same level of intensity, and for the members of the labouring classes to remain approximately stable in number and skill (i.e. for a certain number of working-class children to be fed, kept and schooled, so as to replace their parents when they are unable to work any more, or die). But the use value of the commodity labour power is precisely its capacity to create new value, including its potential to create more value than its own reproduction costs. Surplus-value is but that difference between the total new value created by the commodity labour power, and its own value, its own reproduction costs.

The whole Marxian theory of surplus-value is therefore based upon that subtle distinction between 'labour power' and 'labour' (or value). But there is nothing 'metaphysical' about this distinction. It is simply an explanation (demystification) of a process which occurs daily in millions of cases.

The capitalist does not buy the worker's 'labour'. If he did that there would be obvious theft, for the worker's wage is obviously smaller than the total value he adds to that of the raw materials in the course of the process of production. No: the capitalist buys 'labour power', and often (not always of course) he buys it at its *justum pretium*, at its real value. So he feels unjustly accused when he is said to have caused a 'dishonest' operation. The worker is victim not of vulgar theft but of a social set-up which condemns

him first to transform his productive capacity into a commodity, then to sell that labour power on a specific market (the labour market) characterized by institutional inequality, and finally to content himself with the market price he can get for that commodity, irrespective of whether the new value he creates during the process of production exceeds that market price (his wage) by a small amount, a large amount, or an enormous amount.

The labour power the capitalist has bought ‘adds value’ to that of the used-up raw materials and tools (machinery, buildings etc.). If, and until that point of time, this added value is inferior or equal to the workers’ wages, surplus-value cannot originate. But in that case, the capitalist has obviously no interest in hiring wage labour. He only hires it because that wage labour has the quality (the use value) to add to the raw materials’ value more than its own value (i.e. its own wages). This ‘additional added value’ (the difference between total ‘value added’ and wages) is precisely surplus-value. Its emergence from the process of production is the precondition for the capitalists’ hiring workers, for the existence of the capitalist mode of production.

The institutional inequality existing on the labour market (masked for liberal economists, sociologists and moral philosophers alike by juridical equality) arises from the very fact that the capitalist mode of production is based upon generalized commodity production, generalized market economy. This implies that a propertyless labourer, who owns no capital, who has no reserves of larger sums of money but who has to buy his food and clothes, pay his rent and even elementary public transportation for journeying between home and workplace, *in a continuous way* in exchange of money, is under the *economic compulsion* to sell the only commodity he possesses, to wit his labour power, also on a continuous basis. He cannot withdraw from the labour market until the wages go up. He cannot wait.

But the capitalist, who has money reserves, can temporarily withdraw from the labour market. He can lay his workers off, can even close or sell his enterprise and wait a couple of years before starting again in business. This institutional difference makes price determination of the labour market a game with loaded dice, heavily biased against the working class. One just has to imagine a social set-up in which each citizen would be guaranteed an annual minimum income by the community, irrespective or whether he is employed or not, to understand that ‘wage determination’ under these circumstances would be quite different from what it is under capitalism. In such a set-up the individual would really have the economic choice whether to sell his labour power to another person (or a firm) or not. Under capitalism, he has no choice. His is forced by economic compulsion to go through with that sale, practically at any price.

The economic function and importance of trade unions for the wage-earners also clearly arises from that elementary analysis. For it is precisely the workers’ ‘combination’ and their assembling a collective resistance fund (what was called by the first French unions *caisses de résistance*, ‘reserve deposits’) which enables them, for example though a strike, to withdraw the supply of labour power temporarily from the market so as to stop a downward trend of wages or induce a wage increase. There is nothing ‘unjust’ in such a temporary withdrawal of the supply of labour power, as there are constant withdrawals of demand for labour power by the capitalists, sometimes on a huge scale never equalled by strikes. Through the functioning of strong labour unions, the working class tries to correct, albeit partially and modestly, the institutional inequality on the labour market of which it is a victim, without ever being able to neutralize it durably or completely.

It cannot neutralize it durably because in the very way in which capitalism functions there is a powerful built-in corrective in favour of capital: the inevitable emergence of an industrial reserve army of labour.

There are three key sources for that reserve army: the mass of precapitalist producers and self-employed (independent peasants, handicraftsmen, trades-people, professional people, small and medium-sized capitalists); the mass of housewives (and to a lesser extent, children); the mass of the wage-earners themselves, who potentially can be thrown out of employment.

The first two sources have to be visualized not only in each capitalist country seen separately but on a world scale, through the operations of international migration. They are still unlimited to a great extent, although the number of wage-earners the world over (including agricultural wage labourers) has already passed the one billion mark. At the third source, while it is obviously not unlimited (if wage labour would disappear altogether, if all wage labourers would be fired, surplus-value production would disappear too; that is why ‘total robotism’ is impossible under capitalism), its reserves are enormous, precisely in tandem with the enormous growth of the absolute number of wage earners.

The fluctuations of the industrial reserve army are determined both by the business cycle and by long-term trends of capital accumulation. Rapidly increasing capital accumulation attracts wage labour on a massive scale, including through international migration. Likewise, deceleration, stagnation or even decline of capital accumulation inflates the reserve army of labour. There is thus an upper limit to wage increases, when profits (realized profits and expected profits) are ‘excessively’ reduced in the eyes of the capitalists, which triggers off such decelerated, stagnating or declining capital accumulation, thereby decreasing employment and wages, till a ‘reasonable’ level of profits is restored.

This process does not correspond to any ‘natural economic law’ (or necessity), nor does it correspond to any ‘immanent justice’. It just expresses the inner logic of the *capitalist* mode of production, which is geared to profit. Other forms of economic organization could function, have functioned and are functioning on the basis of other logics, which do not lead to periodic massive unemployment. On the contrary, a socialist would say – and Marx certainly thought so – that the capitalist system is an ‘unjust’, or better stated ‘alienating’, ‘inhuman’ social system, precisely because it cannot function without periodically reducing employment and the satisfaction of elementary needs for tens of millions of human beings.

Marx's theory of surplus-value is therefore closely intertwined with a *theory of wages* which is far away from Malthus's, Ricardo's or the early socialists’ (like Ferdinand Lassalle's) ‘iron law of wages’, in which wages tend to fluctuate around the physiological minimum. That crude theory of ‘absolute pauperization’ of the working class under capitalism, attributed to Marx by many authors (Popper, 1945, et al.), is not Marx's at all, as many contemporary authors have convincingly demonstrated (see among others Rosdolsky, 1968). Such an ‘iron law of wages’ is essentially a demographic one, in which birth rates and the frequency of marriages determine the fluctuation of employment and unemployment and thereby the level of wages.

The logical and empirical inconsistencies of such a theory are obvious. Let it be sufficient to point out that while fluctuations in the supply of wage-labourers are considered essential, fluctuations in the demand for labour power are left out of the analysis. It is certainly a paradox that the staunch opponent of capitalism, Karl Marx, pointed out already in the middle of the 19th century the potential for wage increases under capitalism, even though not unlimited in time and space. Marx also stressed the fact that for each capitalist wage increases of other capitalists’ workers are considered increases of potential purchasing power, not increases in costs (Marx, *d*).

Marx distinguishes two parts in the workers’ wage, two elements of reproduction costs of the commodity labour power. One is purely physiological, and can be expressed in calories and energy

quanta; this is the bottom below which the wage cannot fall without destroying slowly or rapidly the workers' labour capacity. The second one is historical-moral, as Marx calls it (Marx, *i*), and consists of those additional goods and services which a shift in the class relationship of forces, such as a victorious class struggle, enables the working class to incorporate into the average wage, the socially necessary (recognized) reproduction costs of the commodity labour power (e.g. paid holidays after the French general strike of June 1936). This part of the wage is essentially flexible. It will differ from country to country, continent to continent, and from epoch to epoch, according to many variables. But it has the upper limit indicated above: the ceiling from which profits threaten to disappear, or to become insufficient in the eyes of the capitalists, who then go on an 'investment strike'.

So Marx's theory of wages is essentially an *accumulation-of-capital theory of wages* which sends us back to what Marx considered the first 'law of motion' of the capitalist mode of production: the compulsion for the capitalists to step up constantly the rate of capital accumulation.

The laws of motion of the capitalist mode of production

Marx's theory of surplus-value is his most revolutionary contribution to economic science, his discovery of the basic long-term 'laws of motion' (development trends) of the capitalist mode of production constitutes undoubtedly his most impressive scientific achievement. No other 19th-century author has been able to foresee in such a coherent way how capitalism would function, would develop and would transform the world, as did Karl Marx. Many of the most distinguished contemporary economists, starting with Wassily Leontief (1938), and Joseph Schumpeter, (1942) have recognized this.

While some of these 'laws of motion' have obviously created much controversy, we shall nevertheless list them in logical order, rather than according to the degree of consensus they command.

(a) *The capitalist's compulsion to accumulate*. Capital appears in the form of accumulated money, thrown into circulation in order to increase in value. No owner of money capital will engage in business in order to recoup exactly the sum initially invested, and nothing more than that. By definition, the search for profit is at the basis of all economic operations by owners of capital.

Profit (surplus-value, accretion of value) can originate outside the sphere of production in a precapitalist society. It represents then essentially a *transfer of value* (so-called primitive accumulation of capital); but under the capitalist mode of production, in which capital has penetrated the sphere of production and dominates it, surplus-value is currently produced by wage labour. It represents a constant increase in value.

Capital can only appear in the form of many capitals, given its very historical-social origin in private property (appropriation) of the means of production. 'Many capitals' imply unavoidable competition. Competition in a capitalist mode of production is competition for selling commodities in an anonymous market. While surplus-value is produced in the process of production, it is *realized* in the process of circulation, i.e. through the sale of the commodities. The capitalist wants to sell at maximum profit. In practice, he will be satisfied if he gets the average profit, which is a percentage really existing in his consciousness (e.g. Mr. Charles Wilson, the then head of the US automobile firm General Motors, stated before a Congressional enquiry: we used to fix the expected sales price of our cars by adding 15% to production costs). But he can never be sure of this. He cannot even be sure that all the commodities produced will find a buyer.

Given these uncertainties, he has to strive constantly to get the better of his competitors. This can only

occur through operating with more capital. This means that at least part of the surplus-value produced will not be unproductively consumed by the capitalists and their hangers-on through luxury consumption, but will be accumulated, added to the previously existing capital.

The inner logic of capitalism is therefore not only to ‘work for profit’, but also to ‘work for capital accumulation’. ‘Accumulate, accumulate; that is Moses and the Prophets’, states Marx in *Capital*, Vol. I (Marx, *e*, p. 742). Capitalists are *compelled* to act in that way as a result of competition. It is competition which basically fuels this terrifying snowball logic: initial value of capital → accretion of value (surplus-value) → accretion of capital → more accretion of surplus-value → more accretion of capital etc. ‘Without competition, the fire of growth would burn out’ (Marx, *g*, p. 368).

(b) *The tendency towards constant technological revolutions*. In the capitalist mode of production, accumulation of capital is in the first place accumulation of productive capital, or capital invested to produce more and more commodities. Competition is therefore above all competition between productive capitals, i.e. ‘many capitals’ engaged in mining, manufacturing, transportation, agriculture, telecommunications. The main weapon in competition between capitalist firms is cutting production costs. More advanced production techniques and more ‘rational’ labour organization are the main means to achieve that purpose. The basic tend of capital accumulation in the capitalist mode of production is therefore a trend towards more and more sophisticated machinery. Capitalist growth takes the dual form of higher and higher value of capital and of constant revolutions in the techniques of production, of constant technological progress.

(c) *The capitalists’ unquenchable thirst for surplus-value extraction*. The compulsion for capital to grow, the irresistible urge for capital accumulation, realizes itself above all through a constant drive for the increase of the production of surplus-value. Capital accumulation is nothing but surplus-value capitalization, the transformation of part of the new surplus-value into additional capital. There is no other source of additional capital than additional surplus-value produced in the process of production.

Marx distinguishes two different forms of additional surplus-value production. *Absolute surplus-value* accretion occurs essentially through the extension of the work day. If the worker reproduces the equivalent of his wages in 4 hours a day, an extension of the work day from 10 to 12 hours will increase surplus-value from 6 to 8 hours. *Relative surplus-value* accretion occurs through an increase of the productivity of labour in the wage-goods sector of the economy. Such an increase in productivity implies that the equivalent of the value of an identical basket of goods and services consumed by the worker could be produced in 2 hours instead of 4 hours of labour. If the work day remains stable at 10 hours and real wages remain stable too, surplus-value will then increase from 6 to 8 hours.

While both processes occur throughout the history of the capitalist mode of production (viz. the contemporary pressure of employers in favour of overtime!), the first one was prevalent first, the second one became prevalent since the second half of the 19th century, first in Britain, France and Belgium, then in the USA and Germany, later in the other industrialized capitalist countries, and later still in the semi-industrialized ones. Marx calls this process the *real subsumption* (subordination) of *labour under capital* (Marx, *k*), for it represents not only an economic but also a physical subordination of the wage-earner under the machine. This physical subordination can only be realized through social control. The history of the capitalist mode of production is therefore also the history of successive forms of – tighter and tighter – control of capital over the workers inside the factories (Braverman, 1974); and of attempts at realizing that tightening of control in society as a whole.

The increase in the production of relative surplus-value is the goal for which capitalism tends to

periodically substitute machinery for labour, i.e. to expand the industrial reserve army of labour. Likewise, it is the main tool for maintaining a modicum of social equilibrium, for when productivity of labour strongly increases, above all in the wage-good producing sectors of the economy, real wages and profits (surplus-value) can both expand simultaneously. What were previously luxury goods can even become mass-produced wage goods.

(d) *The tendency towards growing concentration and centralization of capital*. The growth of the value of capital means that each successful capitalist firm will be operating with more and more capital. Marx calls this the tendency towards growing concentration of capital. But in the competitive process, there are victors and vanquished. The victors grow. The vanquished go bankrupt or are absorbed by the victors. This process Marx calls the centralization of capital. It results in a declining number of firms which survive in each of the key fields of production. Many small and medium-sized capitalists disappear as independent business men and women. They become in turn salary earners, employed by successful capitalist firms. Capitalism itself is the big ‘expropriating’ force, suppressing private property of the means of production for many, in favour of private property for few.

(e) *The tendency for the ‘organic composition of capital’ to increase*. Productive capital has a double form. It appears in the form of *constant* capital: buildings, machinery, raw materials, energy. It appears in the form of *variable* capital: capital spent on wages of productive workers. Marx calls the part of capital used in buying labour power variable, because only that part produces additional value. In the process of production, the value of constant capital is simply maintained (transferred *in toto* or in part into the value of the finished product). Variable capital on the contrary is the unique source of ‘added value’.

Marx postulates that the basic historic trend of capital accumulation is to increase investment in constant capital at a quicker pace than investment in variable capital; the relation between the two he calls the ‘organic composition of capital’. This is both a technical/physical relation (a given production technique implies the use of a given number of productive wage earners, even if not in an absolutely mechanical way) and a value relation. The trend towards an increase in the ‘organic composition of capital’ is therefore a historical trend towards *basically labour-saving technological progress*.

This tendency has often been challenged by critics of Marx. Living in the age of semi-automation and ‘robotism’, it is hard to understand that challenge. The conceptual confusion on which this challenge is mostly based is an operation with the ‘national wage bill’, i.e. a confusion between wages in general and variable capital, which is only the wage bill of productive labour. A more correct index would be the part of the labour costs in total production costs in the manufacturing (and mining) sector. It is hard to deny that this proportion shows a downward secular trend.

(f) *The tendency of the rate of profit to decline*. For the workers, the basic relation they are concerned with is the rate of surplus-value, i.e. the division of ‘value added’ by them between wages and surplus-value. When this goes up, their exploitation (the unpaid labour they produce) obviously goes up. For the capitalists however, this relationship is not meaningful. They are concerned with the relation between surplus-value and the *totality* of capital invested, never mind whether in the form of machinery and raw materials or in the form of wages. This relation is the *rate of profit*. It is a function of two variables, the organic composition of capital and the rate of surplus-value. If the value of constant capital is represented by *c*, the value of variable capital (wages of productive workers) by *v* and surplus-value by *s*, the rate of profit will be $s/(c+v)$. This can be rewritten as

$$\frac{s / v}{(c + v) / (v)} + 1$$

with the two variables emerging $((c + v) / (v)$ obviously reflects *c*).
Marx postulates that the increase in the rate of surplus value has definite limits, while the increase in the organic composition of capital has practically none (automation, robotism). There will therefore be a basic tendency for the rate of profit to decline.
This is however absolutely true only on a very long-term, i.e. essentially ‘secular’, basis. In other time-frameworks, the rate of profit can fluctuate under the influence of countervailing forces. Constant capital can be devalorized, through ‘capital saving’ technical process, and through economic crises (see below). The rate of surplus-value can be strongly increased in the short or medium term, although each strongly increase makes a further increase more difficult (Marx, *d*, pp. 335–6); and capital can flow to countries (e.g. ‘Third World’ ones) or branches (e.g. service sectors) where the organic composition of capital is significantly lower than in the previously industrialized ones, thereby raising the average rate of profit. Finally, the increase in the mass of *surplus-value* – especially through the extension of wage labour in general, i.e. the total number of workers – offsets to a large extent the depressing effects of moderate declines of the average rate of profit. Capitalism will not go out of business if the mass of surplus-value produced increases ‘only’ from £10 to £17 billion, while the total mass of capital has moved from 100 to 200 billion; and capital accumulation will not stop under these circumstances, nor necessarily slow down significantly. It would be sufficient to have the unproductively consumed part of surplus-value pass e.g. from £3 to £2 billion, to obtain a rate of capital accumulation of 15/200, i.e. 7.5%, even higher than the previous one of 7/100, in spite of a decline of the rate of profit from 10 to 8.5%.
(g) *The inevitability of class struggle under capitalism*. One of the most impressive projections by Marx was that of the inevitability of elementary class struggle under capitalism. Irrespective of the social global framework or of their own historical background, wage-earners will fight everywhere for higher real wages and a shorter work day. They will form elementary organizations for the collective instead of the individual sale of the commodity labour power, i.e. trade unions. While at the moment Marx made that projection there were less than half a million organized workers in at the most half a dozen countries in the world, today trade unions encompass hundreds of millions of wage-earners spread around the globe. There is no country, however remote it might be, where the introduction of wage labour has not led to the appearance of workers’ coalitions.
While elementary class struggle and elementary unionization of the working class are inevitable under capitalism, higher, especially political forms of class struggle, depend on a multitude of variables as to the rapidity with which they extend beyond smaller minorities of each ‘national’ working class and internationally. But there too the basic secular trend is clear. There were in 1900 innumerably more conscious socialists than in 1850, fighting not only for better wages but, to use Marx's words, for the abolition of wage labour (Marx, *i*) and organizing working class parties for that purpose. There are today many more than in 1900.
(h) *The tendency towards growing social polarization*. From two previously enumerated trends, the trend towards growing centralization of capital and the trend towards the growth of the mass of surplus-value, flows the trend towards growing social polarization under capitalism. The proportion of the active

population represented by wage-labour in general, i.e. by the modern proletariat (which extends far beyond productive workers in and by themselves) increases. The proportion represented by self-employed (small, medium-sized and big capitalists, as well as independent peasants, handicraftsmen, tradespeople and ‘free professions’ working without wage-labour) decreases. In fact, in several capitalist countries, the first category has already passed the 90 per cent mark, while in Marx's time it was below 50 per cent everywhere but in Britain. In most industrialized (imperialist) countries, it has reached 80–85 per cent.

This does not mean that the petty entrepreneurs have tended to disappear. Ten or 15–20 per cent out of 30 million people, not to say out of 120 million, still represent a significant social layer. While many small businesses disappear, especially in times of economic depression, as a result of severe competition, they also are constantly created, especially in the interstices between big firms, and in new sectors where they play an exploratory role. Also, the overall social results of growing proletarization are not simultaneous with the economic process in and by itself. From the point of view of class consciousness, culture, political attitude, there can exist significant time-lags between the transformation of an independent farmer, grocer or doctor into a wage-earner, and his acceptance of socialism as an overall social solution for his own and society's ills. But again, the secular trend is towards *growing homogeneity*, less and less heterogeneity, of the mass of the wage-earning class, and not the other way around. It is sufficient to compare the differences in consumer patterns, attitudes towards unionization or voting habits between manual workers, bank employees and government functionaries in say 1900 and today, to note that they have decreased and not increased.

(i) *The tendency towards growing objective socialization of labour*. Capitalism starts in the form of private production on a medium-sized scale for a limited number of largely unknown customers, on an uncontrollably wide market, i.e. under conditions of near complete fragmentation of social labour and anarchy of the economic process. But as a result of growing technological progress, tremendously increased concentration of capital, the conquest of wider and wider markets throughout the world, and the very nature of the labour organization inside large and even medium-sized capitalist factories, a powerful process of objective socialization of labour is simultaneously set in motion. This process constantly extends the sphere of economy in which not blind market laws by conscious decisions and even large-scale cooperation prevail.

This is true especially inside mammoth firms (inside multinational corporations, such ‘planning’ prevails far beyond the boundaries of nation-states, even the most powerful ones!) and inside large-scale factories; but it is also increasingly true for buyer/seller relations, in the first place on an inter-firm basis, between public authorities and firms, and more often than one thinks between traders and consumers too. In all these instances, the rule of the law of value becomes more and more remote, indirect and discontinuous. Planning prevails on a short and even medium-term basis.

Certainly, the economy still remains capitalist. The rule of the law of value imposes itself brutally through the outburst of economic crises. Wars and social crises are increasingly added to these economic crises to remind society that, under capitalism, this growing objective socialization of labour and production is indissolubly linked to private appropriation, i.e. to the profit motive as motor of economic growth. That linkage makes the system more and more crisis-ridden; but at the same time the growing socialization of labour and production creates the objective basis for a general socialization of the economy, i.e. represents the basis of the coming socialist order created by capitalism itself, within the framework of its own system.

(j) *The inevitability of economic crises under capitalism*. This is another of Marx's projections which has been strikingly confirmed by history. Marx ascertained that periodic crises of overproduction were unavoidable under capitalism. In fact, since the crisis of 1825, the first one occurring on the world market for industrial goods to use Marx's own formula, there have been twenty-one business cycles ending (or beginning, according to the method of analysis and measurement used) with twenty-one crises of overproduction. A twenty-second is appearing on the horizon as we are writing.

Capitalist economic crises are always crises of *overproduction of commodities* (*exchange values*), as opposed to pre- and post-capitalist economic crises, which are essentially crises of *underproduction of use-values*. Under capitalist crises, *expanded reproduction* – economic growth – is brutally interrupted, not because too few commodities have been produced but, on the contrary, because a mountain of produced commodities finds no buyers. This unleashes a spiral movement of collapse of firms, firing of workers, contraction of sales (or orders) for raw materials and machinery, new redundancies, new contraction of sales of consumer goods etc. Through this *contracted reproduction*, prices (gold prices) collapse, production and income is reduced, capital loses value. At the end of the declining spiral, output (and stocks) have been reduced more than purchasing power. Then production can pick up again; and as the crisis has both increased the rate of surplus-value (through a decline of wages and a more ‘rational’ labour organization) and decreased the value of capital, the average rate of profit increases. This stimulates investment. Employment increases, value production and national income expand, and we enter a new cycle of economic revival, prosperity, overheating and the next crisis.

No amount of capitalists’ (essentially large combines’ and monopolies’) ‘self-regulation’, no amount of government intervention, has been able to suppress this cyclical movement of capitalist production. Nor can they succeed in achieving that result. This cyclical movement is inextricably linked to production for profit and private property (competition), which imply periodic over-shooting (too little or too much investment and output), precisely because each firm's attempt at maximizing profit unavoidably leads to a lower rate of profit for the system as a whole. It is likewise linked to the separation of value production and value realization.

The only way to avoid crises of overproduction is to eliminate all basic sources of disequilibrium in the economy, including the disequilibrium between productive capacity and purchasing power of the ‘final consumers’. This calls for elimination of generalized commodity production, of private property and of class exploitation, i.e. for the elimination of capitalism.

Marx's theory of crises

Marx did not write a systematic treatise on capitalist crises. His major comments on the subject are spread around his major economic writings, as well as his articles for the *New York Daily Tribune*. The longest treatment of the subject is in his *Theorien über den Mehrwert*, subpart on Ricardo (Marx, *h*, Part 2). Starting from these profound but unsystematic remarks, many interpretation of the ‘Marxist theory or crisis’ have been offered by economists who consider themselves Marxists. ‘Monocausal’ ones generally centre around ‘disproportionality’ (Bukharin, Hilferding, Otto Bauer) – anarchy of production as the key cause of crises – or ‘underconsumption’ – lack of purchasing power of the ‘final consumers’ as the cause of crises (Rosa Luxemburg, Sweezy). ‘Non-monocausal’ ones try to elaborate Marx's own *dictum* according to which *all* basic contradictions of the capitalist mode of production come into play in the process leading to a capitalist crisis (Grossman, Mandel).

The question of determining whether according to Marx, a crisis of overproduction is first of all a crisis of overproduction of commodities or a crisis of overproduction of capital is really meaningless in the framework of Marx's economic analysis. The mass of commodities is but one specific form of capital, commodity capital. Under capitalism, which is generalized commodity production, no overproduction is possible which is not simultaneously overproduction of commodities and overproduction of capital (over-accumulation).

Likewise, the question to know whether the crisis ‘centres’ on the sphere of production or the sphere of circulation is largely meaningless. The crisis is a disturbance (interruption) of the process of enlarged reproduction; and according to Marx, the process of reproduction is precisely a (contradictory) unity of production and circulation. For capitalists, both individually (as separate firms) and as the sum total of firms, it is irrelevant whether more surplus-value has actually been produced in the process of production, if that surplus-value cannot be totally realized in the process of circulation. Contrary to many economists, academic and Marxist alike, Marx explicitly rejected any Say-like illusion that production more or less automatically finds its own market.

It is correct that in the last analysis, capitalist crises of overproduction result from a downslide of the average rate of profit. But this does not represent a variant of the ‘monocausal’ explanation of crisis. It means that, under capitalism, the fluctuations of the average rate of profit are in a sense the seismograph of what happens in the system as a whole. So that formula just refers back to the sum-total of partially independent variables, whose interplay causes the fluctuations of the average rate of profit.

Capitalist growth is always disproportionate growth, i.e. growth with increasing disequilibrium, both between different departments of output (Marx basically distinguishes department I, producing means of production, and department II, producing means of consumption; other authors add a department III producing non-reproductive goods – luxury goods and arms – to that list), between different branches and between production and final consumption. In fact, ‘equilibrium’ under capitalism is but a conceptual hypothesis practically never attained in real life, except as a border case. The above mentioned tendency of ‘overshooting’ is only an illustration of that more general phenomenon. So ‘average’ capital accumulation leads to overaccumulation which leads to the crisis and to a prolonged phenomenon of ‘underinvestment’ during the depression. Output is then consistently inferior to current demand, which spurs on capital accumulation, first to a ‘normal’ level and then to renewed overaccumulation, all the more so as each successive phase of economic revival starts with new machinery of a higher technological level (leading to a higher average productivity of labour, and to a bigger and bigger mountain of produced commodities. Indeed, the very duration of the business cycle (in average 7.5 years for the last 160 years) seemed for Marx determined by the ‘moral’ life-time of fixed capital, i.e. the duration of the reproduction cycle (in value terms, not in possible physical survival) of machinery.

The ups and downs of the rate of the profit during the business cycle do not reflect only the gyrations of the output/disposable income relation; or of the ‘organic composition of capital’. They also express the varying correlation of forces between the major contending classes of bourgeois society, in the first place the short-term fluctuations of the rate of surplus-value reflecting major victories or defeats of the working class in trying to uplift or defend its standard of living and its working conditions.

Technological progress and labour organization ‘rationalizations’ are capital's weapons for neutralizing the effects of these fluctuations on the average rate of profit and on the rate of capital accumulation.

In general, Marx rejected any idea that the working class (or the unions) ‘cause’ the crisis by ‘excessive

wage demands'. He would recognize that under conditions of overheating and 'full employment', real wages generally increase, but the rate of surplus-value can simultaneously increase too. It can, however, not increase in the same proportion as the organic composition of capital. Hence the decline of the average rate of profit. Hence the crisis.

But if real wages do not increase in time of boom, and as they unavoidably decrease in times of depression, the average level of wages during the cycle in its totality would be such as to cause even larger overproduction of wage goods, which would induce an even stronger collapse of investment at the height of the cycle, and in no way help to avoid the crisis.

Marx energetically rejected any idea that capitalist production, while it appears as 'production for production's sake', can really emancipate itself from dependence on 'final consumption' (as alleged e.g. by Tugan-Baranowsky). While capitalist technology implies indeed a more and more 'roundabout-way-of-production', and a relative shift of resources from department II to department I (that is what the 'growing organic composition of capital' really means, after all), it can never develop the productive capacity of department I without developing in the medium and long-term the productivity capacity of department II too, admittedly at a slower pace and in a lesser proportion. So any medium or long-term contraction of final consumption, or final consumers' purchasing power, increases instead of eliminates the causes of the crisis.

Marx visualized the business cycle as intimately intertwined with a *credit cycle*, which can acquire a *relative* autonomy in relation to what occurs in production properly speaking (Marx, *g*, pp. 570–73). An (over)expansion of credit can enable the capitalist system to sell temporarily more goods that the sum of real incomes created in current production plus past savings could buy. Likewise, credit (over)expansion can enable them to invest temporarily more capital than really accumulated surplus-value (plus depreciation allowances and recovered value of raw materials) would have enabled them to invest (the first part of the formula refers to net investments; the second to gross investment).

But all this is only true temporarily. In the longer run, debts must be paid; and they are not automatically paid through the results of expanded output and income made possible by credit expansion. Hence the risk of a *krach*, of a credit or banking crisis, adding fuel to the mass of explosives which cause the crisis of overproduction.

Does Marx's theory of crisis imply a theory of an inevitable final collapse of capitalism through purely economic mechanisms? A controversy has raged around this issue, called the 'collapse' or 'breakdown' controversy. Marx's own remarks on the matter are supposed to be enigmatic. They are essentially contained in the famous chapter 32 of volume I of *Capital* entitled 'The historical tendency of capitalist accumulation', a section culminating in the battle cry: 'The expropriators are expropriated' (Marx, *e*, p. 929). But the relevant paragraphs of that chapter describe in a clearly non-enigmatic way, an interplay of 'objective' and 'subjective' transformations to bring about a downfall of capitalism, and not a purely economic process. They list among the causes of the overthrow of capitalism not only economic crisis and growing centralization of capital, but also the growth of exploitation of the workers and of their indignation and revolt in the face of that exploitation, as well as the growing level of skill, organization and unity of the working class. Beyond these general remarks, Marx, however, does not go.

Marx and Engels on the economy of post-capitalist societies

Marx was disinclined to comment at length about how a socialist or communist economy would operate.

He thought such comments to be essentially speculative. Nevertheless, in his major works, especially the *Grundrisse* and *Das Kapital*, there are some sparse comments on the subject. Marx returns to them at greater length in two works he was to write in the final part of his life, his comments on the *Gotha Programme* of united German social-democracy (Marx, *j*), and the chapters on economics and socialism he wrote or collaborated with for Engels' *Anti-Dühring* (1878). Generally his comments, limited and sketchy as they are, can be summarized in the following points.

Socialism is an economic system based upon conscious planning of production by associated producers (nowhere does Marx say: by the state), made possible by the abolition of private property of the means of production. As soon as that private property is completely abolished, goods produced cease to be commodities. Value and exchange value disappear. Production becomes production for use, for the satisfaction of needs, determined by conscious choice (*ex ante* decisions) of the mass of the associated producers themselves. But overall economic organization in a postcapitalist society will pass through two stages.

In the first stage, generally called 'socialism', there will be relative scarcity of a number of consumer goods (and services), making it necessary to measure exactly distribution based on the actual labour inputs of each individual (Marx nowhere refers to different quantities and *qualities* of labour; Engels explicitly *rejects* the idea that an architect, because he has more skill, should consume more than a manual labourer). Likewise, there will still be the need to use incentives for getting people to work in general. This will be based upon strict equality of access for all trades and professions to consumption. But as human needs are unequal, that formal equality masks the survival of real inequality.

In a second phase, generally called 'communism', there will be plenty, i.e. output will reach a saturation point of needs covered by material goods. Under these circumstances, any form of precise measurement of consumption (distribution) will wither away. The principle of full needs satisfaction covering all different needs of *different* individuals will prevail. No incentive will be needed any more to induce people to work. 'Labour' will have transformed itself into meaningful many-fold activity, making possible all-round development of each individual's human personality. The division of labour between manual and intellectual labour, the separation of town and countryside, will wither away. Humankind will be organized into a free federation of producers' and consumers' communes.

Selected works

There is still no complete edition of all of Marx's and Engels's writings. The standard German and Russian editions by the Moscow and East Berlin Institutes for Marxism-Leninism, generally referred to as *Marx-Engels-Werke* (MEW), do not include hundreds of pages printed elsewhere (e.g. Marx's *Enthüllungen zur Geschichte der Diplomatie im 18. Jahrhundert* [Revelations on the History of 18th-century Diplomacy]), and several thousand pages of manuscripts not yet printed at the time these editions were published. At present, a monumental edition called *Marx-Engels-Gesamtausgabe* (MEGA) has been started, again both in German and in Russian, by the same Institutes. It already encompasses many of the unpublished manuscripts referred to above, in the first place a previously unknown economic work which makes a bridge between the *Grundrisse* and vol. 1 of *Capital*, and which was written in the years 1861–3 (published under the title *Zur Kritik der Politischen Oekonomie – Contribution to a Critique of Political Economy 1861–1863* in MEGA II/3/1–6, Berlin Dietz Verlag, 1976–1982). Whether it will include all of Marx's and Engels's writings remains to be seen.

In English, key works by Marx and Engels have been systematically published by Progress Publishers, Moscow, and Lawrence & Wishart, London; but this undertaking is by no means an approximation of the *Marx-Engels-Werke* mentioned above. The quality of the translation is often poor. The translations of Marx's and Engels's writings published by Penguin Books in the *Marx Pelican Library* are quite superior to it. We therefore systematically refer to the latter edition whenever there is a choice. Marx's and Engels's books and pamphlets referred to in the present text are mostly in chronological order:

- (Marx *a*) *Die Deutsche Ideologie* (1846), together with Friedrich Engels.
- (Marx *b*) *Manifest der Kommunistischen Partei* (1848), written in collaboration with Friedrich Engels. In English: *Manifesto of the Communist Party*, in *Marx: The Revolutions of 1848*, Harmondsworth: Penguin Books, 1973.
- (Marx *c*) *Zur Kritik der Politischen Oekonomie* (1858). In English: *Contribution to the Critique of Political Economy*, London: Lawrence & Wishart, 1970.
- (Marx *d*) *Grundrisse der Kritik der Politischen Oekonomie* (written in 1858–1859, first published in 1939). English edition: *Foundations of a Critique of Political Economy*, Harmondsworth: Penguin Books, 1972.
- (Marx *e*): *Das Kapital, Band I* (1867). In English: *Capital*, Vol. I, Harmondsworth: Penguin Books, 1976.
- (Marx *f*) *Das Kapital, Band II*, published by Engels in 1885. In English: *Capital*, Vol. II, Harmondsworth: Penguin Books, 1978.
- (Marx *g*) *Das Kapital, Band III*, published by Engels in 1894. In English: *Capital*, Vol. III, Harmondsworth: Penguin Books, 1981.
- (Marx *h*) *Theorien über den Mehrwert*, published by Karl Kautsky 1905–10. In English: *Theories of Surplus Value*, Moscow: Progress Publishers, 1963.
- (Marx *i*) *Lohn, Preis und Profit*, written in 1865. In English: *Wages, Price and Profits*, in *Marx-Engels Selected Works*, Vol. II, Moscow: Progress Publishers, 1969.
- (Marx *j*) *Kritik des Gothaer Programms*, written in 1878 in collaboration with Engels. In English: *Critique of the Gotha Programme*, in *Marx-Engels: The First International and After*, Harmondsworth: Penguin Books, 1974.
- (Marx *k*) *Resultate des unmittelbaren Produktionsprozesses* (unpublished section VII of Vol. I of *Capital*), first published in 1933. In English: *Results of the Immediate Process of Production, Appendix to Capital*, Vol. I, Harmondsworth: Penguin Books, 1976.
- (Marx *l*) *Marx-Engels: Briefwechsel (Letters)*. There is no complete English edition of the letters. Some are included in the *Selected Works* in 3 vols, published by Progress Publishers, Moscow.
- (Engels): *Anti-Dühring* (1878). The chapter on economy was written by Marx, who also read all the other parts and collaborated in their final draft. In English: *Anti-Dühring*, London: Lawrence & Wishart, 1955.

Bibliography

There are innumerable books and articles devoted to comments or elaborations on Marx's economic thought, or which criticize them. We list here those works we refer to in the above text, as well as those we consider the most important ones (based, needless to say, upon subjective judgement).

Baran, P.A. 1957. *The Political Economy of Growth*. London: John Calder.

Baran, P.A. and Sweezy, P.M. 1966. *Monopoly Capital*. New York and London: Monthly Review Press.

Böhm-Bawerk, E. von. 1896. *Zum Abschluss des Marxschen Systems*. English edn, *Karl Marx and the Close of this System*, including a reply by Rudolf Hilferding, *Böhm-Bawerk's Criticism of Karl Marx*, ed. P.M. Sweezy, New York: Augustus M. Kelley, 1949.

Bortkiewicz, L. von. 1907. *Zur Berichtigung der grundlegenden theoretischen Konstruktion von Marx im Dritten Band des 'Kapital'*. Trans. by P.M. Sweezy as 'On the Correction of Marx's Fundamental Theoretical Construction in the third volume of *Capital*'. In Böhm-Bawerk (1949), 197–221.

Braverman, H. 1974. *Labor and Monopoly Capital: the degradation of work in the twentieth century*. New York and London: Monthly Review Press.

Bronfenbrenner, M. 1970. *The Vicissitudes of Marxian Economics*. London.

Bukharin, N. 1914. *Imperialism and World Economy*. English trans., London: M. Lawrence, 1915.

Bukharin, N. 1926. *Imperialism and the Accumulation of Capital*. Trans. by R. Wichmann, ed. K.J. Tarbuck, London: Allen Lane, 1972.

Dobb, M. 1937. *Political Economy and Capitalism: some essays in economic tradition*. London: G. Routledge & Sons. Reprinted Westport, Conn.: Greenwood Press, 1972.

Emmanuel, A. 1969. *L'échange inégal. Essai sur les antagonismes dans les rapports économiques internationaux*. Paris: François Maspero. Trans. by B. Pearce as *Unequal Exchange: a study of the imperialism of trade*, London: New Left Books, 1972.

Grossman, H. 1929. *Das Akkumulations- und Zusammenbruchsgesetz des kapitalistischen Systems*. Leipzig: C.L. Hirschfeld.

Hayek, F.A. von. 1944. *The Road to Serfdom*. Chicago: University of Chicago Press: London: G. Routledge & Sons.

Hilferding, R. 1910. *Das Finanzkapital*. Vienna: Wiener Volksbuchhandlung. Trans. by M. Watnick and S. Gordon, ed. T. Bottomore as *Finance Capital*, London: Routledge & Kegan Paul, 1981.

Itoh, M. 1980. *Value and Crisis: Essays on Marxian Economics in Japan*. London: Pluto.

Kolakowski, L. 1976–8. *Main Currents of Marxism: its rise, growth and dissolution*, 3 vols. Trans. P.S. Falla, Oxford: Clarendon Press, 1978.

Lange, O. 1963. *Political Economy*, 2 vols. Trans., ed. P.F. Knightsfield, Oxford: Pergamon Press; Warsaw: PWN – Polish Scientific Publishers.

Lange, O. and Taylor, F.M. 1938. *On the Economic Theory of Socialism*. 2 vols, Minneapolis: University of Minnesota Press.

Lenin, V.I. 1917. *Imperialism, Last Stage of Capitalism*. Petrograd. English trans., *Imperialism, the Highest Stage of Capitalism*, Moscow: Foreign Languages Publishing House, 1947.

Leontief, W. 1938. The significance of Marxian economics for present-day economic theory. *American Economic Review*, Supplement, March. Reprinted in *Marx and Modern Economists*, ed. D. Horowitz, London: MacGibbon & Kee, 1968.

Luxemburg, R. 1913. *Akkumulation des Kapitals*. Trans. A. Schwarzschild, with an introduction by J. Robinson, as *The Accumulation of Capital*, London: Routledge & Kegan Paul, 1951.

Mandel, E. 1962. *Traité d'économie marxiste*. Paris: R. Juillard. Trans. B. Pearce as *Marxist Economic Theory*, 2 vols, London: Merlin Press, 1968.

Mandel, E. 1972. *Der Spätkapitalismus*. Frankfurt am Main: Suhrkampf. Trans. J. De Bres as *Late Capitalism*, London: New Left Books. Revised edn, 1975.

Mandel, E. 1980. *Long Waves of Capitalist Development: the Marxist interpretation*. Cambridge: Cambridge University Press.

Mandel, E. and Freeman, A., eds. 1984. *Ricardo, Marx, Sraffa*. London: Verso.

Mattick, P. 1969. *Marx and Keynes: The Limits of the Mixed Economy*. Boston: P. Sargent.

Mises, L. von. 1920. Die Wirtschaftsrechnung im sozialistischen Gemeinwesen. Trans. S. Adler as 'Economic calculations in the socialist commonwealth', in *Collectivist Economic Planning*, ed. F.A. von Hayek, London: G. Routledge & Sons, 1935.

Morishima, M. 1973. *Marx's Economics: A Dual Theory of Value and Growth*. Cambridge: Cambridge University Press.

Nutzinger, H.G. and Wolfstetter, E., eds. 1974. *Die Marx'sche Theorie und ihre Kritik*. 2 vols, Frankfurt am Main: Campus.

Pareto, V. 1966. *Marxisme et économie pure*. Geneva: Droz.

Popper, K. 1945. *The Open Society and its Enemies*. 2 vols, London: G. Routledge & Sons.

Robinson, J. 1942. *An Essay on Marxian Economics*. London: Macmillan. 2nd edn, 1966.

Rosdolsky, R. 1968. *Entstehungsgeschichte des Marxschen 'Kapital'*. Frankfurt am Main: Europäische Verlagsanstalt. Trans. P. Burgess as *The Making of Marx's Capital*, London: Pluto Press, 1977.

Rubin, I.I. 1928. *Essays on Marx's Theory of Value*. Moscow. English trans., Detroit: Black and Red, 1972.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. New York and London: Harper & Brothers.

Sraffa, D. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

Sternberg, F. 1926. *Der Imperialismus*. Berlin: Malik-Verlag.

Sweezy, P.M. 1942. *The Theory of Capitalist Development*. New York: Monthly Review Press.

Tugan-Baranowsky, M. von. 1905. *Theoretische Grundlagen der Marxismus* Leipzig: Duncker & Humblot.

Wygodsky, S. *Der gegenwärtige Kapitalismus*. Trans. by C.S.V. Salt as *The Story of a Great Discovery: How Karl Marx wrote 'Capital'*, Tunbridge Wells: Abacus Press, 1972.

Bibliographic addendum

Foley, D. 1986. *Understanding Capital: Marx's Economic Theory*. Cambridge, MA: Harvard University Press.

Roemer, J. 1981. *Analytical Foundations of Marxian Economic Theory*. Cambridge: Cambridge University Press.

Roemer, J. 1988. *Free to Lose: An Introduction to Marxist Economic Philosophy*. London: Radius.

Outside of economics, important studies include:

Buchanan, A. 1982. *Marx and Justice*. Totowa, NJ: Rowan and Allanhead.

Carver, T. 1998. *The Post-Modern Marx*. Manchester: Manchester University Press.

Cohen, G.A. 1988. *History, Labour, and Freedom: Themes from Marx*. Oxford: Oxford University Press.

Cohen, G.A. 2000. *Karl Marx's Theory of History: A Defence*. 2nd edn. Oxford: Clarendon Press.

Elster, J. 1985. *Making Sense of Marx*. Cambridge: Cambridge University Press.

Kain, P. 1989. *Marx and Ethics*. Oxford: Oxford University Press.

Roemer, J., ed. 1986. *Analytical Marxism*. Cambridge: Cambridge University Press.

Rockmore, T. 2002. *Marx After Marxism*. Oxford: Basil Blackwell.

Van den Berg, A. 1989. *The Immanent Utopia: From Marxism on the State to the State of Marxism*. Princeton: Princeton University Press.

Wolff, J. 2002. *Why Read Marx Today?* Oxford: Oxford University Press.

Wood, A. 2004. *Karl Marx*. 2nd edn. London: Routledge.

All of these are complemented by what remains a classic short introduction to Marx's ideas:

Berlin, I. 1985. *Karl Marx*. 4th edn. Oxford: Oxford University Press.

as well as the standard biography of Marx:

McLellan, D. 1996. *Karl Marx: A Biography*. 3rd revised edn. London: Palgrave Macmillan.

Finally,

Bottomore, T., ed. 1983. *A Dictionary of Marxist Thought*. Cambridge, MA: Harvard University Press.

is helpful both as a survey and as a mechanism for overcoming language differences between Marxist approaches and others.

How to cite this article

Mandel, Ernest. "Marx, Karl Heinrich (1818–1883)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000097> doi:10.1057/9780230226203.1051

The New Palgrave Dictionary of Economics Online

Marxian transformation problem

Duncan Foley and Gérard Duménil
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The transformation problem relates the labour theory of value and the competitive equalization of the rate of profit. Marx distinguishes the production of surplus-value from its redistribution through prices. Critics claim that the labour theory of value is an unnecessary detour to the determination of prices because total value and surplus-value are not conserved. The Single-System Labour Theory of Value (SS-LTV) argues that at any prices (1) the price of the net product expresses the labour expended, and (2) total profits are the price form of surplus-value, because the value of labour-power is the labour time equivalent of the wage.

Keywords

capitalist law of exchange; commodity law of exchange; constant and variable capital; dual system; exploitation; Fundamental Marxian Theorem; gravitation; labour power; labour theory of value; labour time; Marx, K. H.; Marxian transformation problem; monetary expression of value or labour time; natural price; net product; ‘new interpretation’ of labour theory of value; profit, rate of; Ricardo, D.; single-system labour theory of value; Smith, A.; surplus-value; transformation problem; value; value added; value of labour-power; variable capital

Article

Marx's framework: value, surplus-value, prices and competition

Marx consistently distinguishes the notions of *value* and *price*, in contrast to contemporary economic language, which uses the term ‘value’ to refer to prices in a situation of general equilibrium, though the use of the term is rather flexible; for example ‘value added’ is actually the value of net product measured in price terms. For Marx, value is a ‘social substance’ manifested in economic relations in the ‘form’ of prices, though prices are not necessarily proportional to values, as we will see.

Value and surplus-value

We first recall Marx's basic concepts (see also Marx's analysis of capitalist production). Central to Marx's framework of analysis in *Capital* is the *labour theory of value* (LTV), which defines the value of a commodity as the ‘socially necessary’ labour time required by its production, that is, the labour time required by average available techniques of production for workers of average skill.

The LTV is central to Marx's theory of exploitation, a term he uses to describe a situation in which one individual or group lives on the product of the labour of others. According to the LTV, when commodities are exchanged through sale and purchase, no value is created. But this principle does not apply to capitalists’ purchase of the

labour power of workers. Workers sell their labour power, that is, their capability to work, to a firm, owned by a capitalist. The buyer uses this labour power in production to add value to the commodity produced. The *value of labour power* is the labour time required by the production of the commodities the worker buys. But the worker can typically work more hours than are on average required to produce this bundle of commodities. For example, the goods the worker can buy may require eight hours of labour per day, when the labour-day lasts 12 hours. The difference, four hours, is unpaid labour time. If an hour of social labour on average produces a value whose price form is $10, four hours of unpaid labour time results in a surplus-value whose price form is $40, which is appropriated by the capitalist. The *rate of surplus-value* is the ratio of unpaid to paid labour time, in this case 4/8, that is, 50 per cent.

Two laws of exchange

Marx situates his discussion in the context of the distinction made by Adam Smith and David Ricardo between ‘market prices’ and ‘natural prices’. Market prices are the prices at which commodities actually exchange from day to day in the market. Smith and Ricardo, however, regarded market prices as fluctuating (or ‘gravitating’) around centres of attraction they called ‘natural prices’. (‘Gravitation’ means that the economy is in a permanent situation of disequilibrium, though in a vicinity of equilibrium where natural prices would prevail.)

In the above analysis, Marx assumes that commodities tend to exchange at their values (at *prices proportional to values*), that is, in proportion to the labour time embodied in them. ‘Tend’ means here that deviations are obviously possible, but that such prices will ‘regulate’ the market, in the sense that if the prevailing set of prices systematically under-compensates the labour used in the production of a commodity, labour will move to the production of better-paid commodities. As a result, the supply of the under-compensated commodity will decline, and its price will rise. In reality prices would gravitate around values, which would play the role of natural prices in such an economy. This is the *commodity law of exchange*.

In a capitalist economy, however, capitalists buy not only the labour power of workers (which Marx denotes as *variable capital*), but also non-labour inputs, such as raw materials, and fixed capital, such as machinery (which Marx denotes as *constant capital*). If natural prices were proportional to labour inputs, as the commodity law of exchange posits, capitalists using more constant capital per worker than the average would realize smaller profit in comparison to their total capital advanced, that is, lower profit rates. Marx accepts the idea that competition tends to equalize profit rates in various industries, despite differences in capital advanced per worker, which is the *capitalist law of exchange*. Marx uses the term ‘prices of production’ to describe a system of prices which guarantee to the capitalists of various industries a uniform profit rate. Capitalists will invest more where profit rates are larger, and conversely in the symmetrical case. They move their capital from one industry to another seeking maximum profit rates, and this movement results in a gravitation of market prices around prices of production. Marx regards prices of production as the centres of gravitation of market prices, and thus the natural prices relevant to a competitive capitalist economy.

Is the theory of surplus-value compatible with the theory of competition?

The problem is posed of the compatibility of the capitalist law of exchange at prices of production with the theory of exploitation as extraction of surplus-value. Marx's line of argument is that surplus-value is *created* in production through the exploitation of labour, that is, in proportion to labour expended, but *realized* proportionally to total capital invested. According to Marx, this separation between the locus of extraction and the locus of realization does not contradict the theory of exploitation so that capitalist competition is compatible with his theory of exploitation through the appropriation of surplus-value from unpaid labour time.

To support this argument, Marx presents a pair of tables (1981, ch. 9) showing the redistribution of surplus-value through deviations of price from values proportional to embodied labour times. All variables are measured in hours of labour time, and as a result prices of production are expressed in the same unit. Because Marx's own calculations involve some extraneous complexity (differential turnover rates among sectors), it is more useful to consider the simplified case shown in Table 1. Two industries exist, each of which advances the same capital of 100, but divided in different proportions between the purchase of non-labour inputs (C) and labour inputs (V). All capital is used up during the period, so that the rate of profit is the ratio of surplus-value to total capital advanced, $r=s/(c+v)$. The rate of surplus-value is uniform and equal to 100 per cent. Consequently, surplus-values are equal to variable capitals. Surplus-values and values are computed in each industry. When prices are proportional to values, profit rates differ between the two sectors. Prices of production are determined in Marx's procedure by summing up all surplus-value, a total of 40, and redistributing it in proportion to total capital, that is 20 in each industry, to equalize profit rates on the capitals advanced.

Marx's calculation of prices of production from values

Industry	Constant capitals, C	Variable capitals, V	Total capitals, $K=C+V$	Surplus-values, $S=V$	Values of commodities produced, $\Lambda=K+S$	Profits, Π	'Prices of production' of commodities produced, $P=K+\Pi$
1	70	30	100	30	130	20	120
2	90	10	100	10	110	20	120
Total economy	160	40	200	40	240	40	240

The procedure illustrates a straightforward 'redistribution' of surplus-value. Clearly, the sum of prices, 240, is equal to the sum of values, and total surplus-value is, by construction, conserved in the form of profit. These observations are expressed in two *Marxian equations* concerning the entire economy:

$$\text{Sum of values} = \text{sum of prices of production} \quad \text{Sum of surplus-value} = \text{sum of profits}$$

Note that these compact formulations are not rigorous, since values and surplus-value are measured in labour time and prices and profits in money. Thus, 'Sum of values' should read 'Sum of prices proportional to values'. A simple way out of the problem of units is to use one of these equations to define the general level of prices. For example, the sum of prices of production could be set equal to the number of hours corresponding to the sum of values. Then, Marx's line of argument implies that the surpluses in both sets of prices are equal, as in the second equation. This simple calculation illustrates the idea that profits are 'forms' of surplus-value, that is, unpaid labour.

Approximations

Marx is, however, aware that the type of computation illustrated in Table 1 is not satisfactory, since the evaluations of constant and variable capital have not been modified despite the fact that prices have changed.
First, when natural prices are prices of production, non-labour inputs are purchased on the market at prices of production, not at prices proportional to values. It is, therefore, not correct to conserve the evaluation of constant capital:

> We had originally assumed that the cost-price of a commodity equalled the value of the commodities consumed in its production. But for the buyer the price of production of a specific commodity is its cost-price, and may thus pass as cost-price into the prices of other commodities. Since the price of production may differ from the value of a commodity, it follows that the cost-price of a commodity containing this price of production of another commodity may also stand above or below that portion of its total value derived from the value of the means of production consumed by it. It is necessary to remember this modified significance of the cost-price, and to bear in mind that there is always the possibility of an error if the cost-price of a commodity in any particular sphere is identified with the value of the means of production consumed by it. Our present analysis does not necessitate a closer examination of this point. (Marx, 1981, ch. 9)

Second, there is a similar problem concerning variable capital. When commodities exchange at prices of production, workers will not be able to buy the same bundle of commodities with a wage corresponding to a purchasing power expressed, as in Marx's calculation, as a certain number of hours of labour time, as when prices are proportional to values. Marx is also aware of this problem:

> […] the average daily wage is indeed always equal to the value produced in the number of hours the labourer must work to produce the necessities of life. But this number of hours is in its turn obscured by the deviation of the prices of production of the necessities of life from their values. However, this always resolves itself to one commodity receiving too little of the surplus-value while another receives too much, so that the deviations from the values which are

> embodied in the prices of production compensate one another. Under capitalist production, the general law acts as the prevailing tendency only in a very complicated and approximate manner, as a never ascertainable average of ceaseless fluctuations. (Marx, 1981, ch. 9)

It is not easy to understand Marx's position from these notes (which he never revised for publication). It does seem that the analysis requires a ‘closer analysis’, since the revaluation of constant capital at prices of production will in general make the sum of prices of production deviate from the sum of values, or make the sum of profits deviate from the sum of surplus-values. While it is true that a redistribution of surplus-value through a system of prices of production does not alter the living labour expended in production, so that over the whole economy the deviations from value ‘compensate one another’, the value of labour power will remain constant only if workers consume commodities in the same proportion as they are produced in the whole economy, which is implausible. The phrase ‘average of ceaseless fluctuations’ suggests the averaging out of market prices to prices of production rather than the averaging of surplus-value across sectors.
If Marx's use of the term ‘approximately’ is taken literally, it would appear that the LTV and the theory of exploitation he introduced in Volume 1 of *Capital* are only ‘approximately’ true! Although Marx is conscious of the problem, it is impossible to consider his solution as rigorous. In the formulation of the two equations above, it appears that, when the calculation is done rigorously as in the formal setting below, the second equation *does not hold*! Later critics have judged this a devastating refutation of Marx's theories of value and exploitation, which in turn has led to ongoing controversy.

Earlier approaches

The foundations of the transformation problem can be found in the first analyses of competition and prices in capitalism, beginning with Adam Smith and David Ricardo, on which Marx elaborated. The distinction between values and prices remains somewhat fuzzy in these authors. Smith fails to establish a clear relationship between value and profit rate equalization as the principle determining ‘natural prices’. Thus, one characteristic feature of these approaches, from which Marx was unable to depart completely, is that two sets of prices (the two laws of exchange above) are considered, one proportional to values (embodied labour times), and the other equalizing profit rates (a *dual system*), when only one price system prevails in real-world capitalism (a *single system*):

1. 1. A system of prices proportional to values (embodied labour times) plays a role in the analyses of Smith, Ricardo and Marx. Only Marx, however, clearly distinguishes the two systems from the start.
2. 2. The determination of the ‘surplus’, when such a concept exists (as in Ricardo and Marx), is posed in the first system and imported into the second, instead of being analysed directly within the second system.

This dual system approach lies at the basis of the phrase ‘transformation problem’, which refers to the transformation from one system into the other.

Adam Smith

Smith's point of departure is an ‘early, rude’ state of society, before the establishment of private property in land and means of production. There, Smith contends, products of human labour will exchange in proportion to the labour time required to produce them. Smith offers as an example that, if it requires two days on average to kill a beaver, but one day to kill a deer, a beaver will tend to exchange for two deer. Smith's argument supporting this conclusion rests on the assumption that any hunter can choose to allocate time to hunting deer or beaver, so that, if the exchange ratio were higher or lower than the labour time ratio, hunters would shift from the under- to the over-remunerated productive activity, and force the exchange ratio back toward the labour time ratio. The viewpoint is clearly that of the commodity law of exchange. Smith applies the same type of reasoning to argue that, once means of production have become private property (which he calls ‘stock’, and later economists called ‘capital’), the ability of owners to shift their capital from one line of production to another will tend to equalize the profit rate across different sectors of production. The viewpoint is now that of the capitalist law of exchange.

David Ricardo

Ricardo critiques and corrects Smith's analysis. Ricardo originally based his theories of prices and distribution on Smith's first principle that the labour expended in

producing a commodity determines its price in exchange. But Ricardo, elaborating on the dual system approach, examines the necessary quantitative difference between the two principles that might determine natural prices more carefully than Smith. Ricardo understood that the proportion between capitals invested in non-labour inputs and labour is not uniform across industries, and that this fact implies a discrepancy between the two sets of prices, but he regarded these deviations as quantitatively limited. Prefiguring Marx's investigation, Ricardo was concerned to work out the properties of the first system (values) to derive conclusions concerning distribution, which he supposed were also valid in the second system (prices of production).

First, when natural prices are proportional to values (embodied labour times), it is obvious that there is a trade-off between the shares of output which respectively go to workers and capitalists: workers create all the value added to inputs, and buy a share of output whose production requires less labour time than they expend. In contrast to Smith, Ricardo had a clear view of this mechanism. This division of total output between workers and capitalists was crucial to his analysis, because of its implications in terms of economic policy. (For example, Ricardo was in favour of a low price of corn, which, in his opinion, would increase the profits of capitalists by lowering wages – and encourage capital accumulation.)

Second, Ricardo would have liked to conserve the straightforward distributional properties he derived from the assumption of prices proportional to values, even while acknowledging the quantitative difference between such natural prices proportional to values and natural prices that would equalize profit rates across industries. But Ricardo understood that, in the profit rate-equalizing system, the natural prices of commodities may change with a change in the real wage (due to the distinct compositions of capital) even if the labour required in production remains unaltered, contrary to what happens in the first system, where values remain unchanged with a change in the wage. Thus, with Ricardo's analysis, we are getting closer to Marx's framework and problems.

The rebellious classical legacy in Marx

Marx adopted key elements from Smith and Ricardo's works: (*a*) a dual system approach to natural prices in capitalism (beginning, with Smith, as if labour was the unique input); (*b*) Ricardo's analysis of distribution as a ‘trade-off’ between wages and profit; and (*c*) Smith's analysis of competition that Ricardo had also adopted.

The two classical economists were the mainstream when Marx started his study of economics. Marx seized this opportunity to establish his theory of exploitation, in which surplus-value arises from unpaid labour time, on ‘mainstream’ grounds. Then he devoted hundreds of pages (in the manuscripts known as *The Theories of Surplus-value*) to the inability of these ‘bourgeois’ economists to establish a theory of exploitation, although Ricardo came close. This very smart political move on Marx's part eventually forced mainstream economic theory to abandon these ‘dangerous’ implications of the LTV.

The transformation controversy

A large literature is devoted to the transformation problem, starting with the critical contributions of Eugen Böhm-Bawerk (1890) and Ladislaus von Bortkiewicz (1952) in the late 19th and early 20th centuries. This literature has led to considerable formal advance, though it has failed to resolve the basic controversy over which of Marx's conclusions, if any, are logically valid.

There are fundamentally two points raised by these critiques. First, the critics claim that the value system is useless as a preliminary to the calculation of prices of production. Paul Samuelson puts this point in the following manner: ‘Contemplate two alternative and discordant systems. Write down one. Now transform by taking an eraser and rubbing it out. Then fill in the other one. *Voilá!* You have completed your transformation algorithm’ (1971, p. 400). This point is, however, not really relevant, since Marx's objective was not to show that it is impossible to compute prices of production if values have not been previously determined, but rather to show that the theory of exploitation is consistent with the principle of capitalist competition.

Second, the main focus of this critique is the incompatibility of the two Marxian equations. This literature calculates surplus-value by deducting the value of a given bundle of worker's consumption from the worker's labour time. Profits, on the other hand, are calculated by deducting the price of this same bundle at prices of production from the value added (in prices). When prices of production are not proportional to values, these two quantities are not equal, violating the second Marxian equation. This treatment of the wage of workers, which allocates their purchasing power to particular commodities, departs from Marx's apparent stipulation in his discussion of the transformation problem of the rate of surplus-value.

In face of this quantitative inequality between surplus-value and profit, the Fundamental Marxian Theorem (see Morishima, 1973) argues that the LTV does provide a qualitative foundation for Marx's theory of exploitation, since the rate of profit will be positive if and only if the rate of surplus-value is positive. This interesting observation, however, falls short of fulfilling Marx's ambition to found his theory of exploitation on the LTV through the two Marxian equations.

A crucial moment in the criticism of Marx's transformation was the publication of Piero Sraffa (1960). This book is simultaneously a critique of Marx and of neoclassical economics, but it is, above all, a bold attempt to elaborate Ricardo's analysis. It is the origin of the neo-Ricardian school, represented by, in particular, Ian Steedman (1977) and Pierangelo Garegnani (1984). The central point, in the neo-Ricardian School, is that the LTV is useless, with respect to both the determination of prices of production and exploitation. The dual-system approach of Ricardo is abandoned in favour of the price of production system, as the reference to value is deemed irrelevant. Sraffa calculates prices of production directly from a description of technology and distribution. In this framework, he shows that Ricardo's trade-off between wages and the profit rate can be derived formally as a downward sloping relation (see the mathematical section below).

The price of net product-unallocated purchasing power labour theory of value (PNP-UPP LTV) approach to exploitation

In the late 1970s, Gérard Duménil (1980; 1983; 1984) and Duncan Foley (1982) (independently) proposed new lines of interpretation of Marx's theory of value. In doing so, they followed distinct routes, but the basic principles underlying these reformulations converge to the same basic framework. This interpretation is inappropriately referred to, in the literature, as the ‘New Interpretation’. It is more precise to describe it as the ‘price of net product-unallocated purchasing power labour theory of value’ (PNP-UPP LTV). It was rapidly adopted by Alain Lipietz (1982).

Value and exploitation in the PNP-UPP LTV approach

Beginning with Marx's two equations, as is traditional, there are two basic principles to this interpretation. First, Marx's equation concerning the ‘sum of values’ and ‘sum of prices’ holds for the net product of the period. ‘Net product’ means here, as in Marx's reproduction schemes and national accounting frameworks, output minus non-labour inputs inherited from the previous period. The important idea here is that it is the expenditure of living labour that creates value. Marx regards the value of a commodity as equal to the value transferred by the inputs consumed and the new value created by labour during the period. But the two perspectives are equivalent:

$$\text{Value transferred from inputs} + \text{value created by new labour} = \text{value of output}$$
$$\text{Value created by new labour} = \text{value of output} - \text{value transferred from inputs}$$

The price form of the value created by the total productive labour expended during a period of time is the price of the net product of the period. (As is well known, the price of this net product is equal to total income, wages plus profits.) The PNP-UPP LTV interpretation argues that, when Marx (in the first quotation above) points to the fact that the cost-prices of commodities used as inputs to production must be adjusted to reflect the change to prices of production, the correct formulation would have been to exclude them from the first Marxian equation, which would then read ‘Sum of values of net product=sum of price of net product’. Since values are expressed in labour time, while prices of production are expressed in terms of money, this equation implicitly defines an equivalence between value-creating labour time and money, the *monetary expression of value or labour time* (MELT), which is the ratio of the price of net product (value added measured in money) to the productive labour time expended. If, for example, 250 billion hours of productive labour were expended in an economy to produce a net product worth $10 trillion, the monetary expression of labour time would be $40 per hour. The MELT expresses quantitatively (as a ratio of the price of the net product to the living labour expended) what Marx calls the ‘price form’ of the total value created during the period.

Second, the PNP-UPP LTV views the term ‘surplus-value’ in the second Marxian equation as referring to the monetary equivalent of unpaid labour time. The wage, as in Marx's calculation, is regarded as unallocated purchasing power giving workers the potential to buy a fraction of the net product. (This is the way capitalists look at wage payments, since the individual capitalist has no interest in how workers actually spend their wages.) Individual workers can allocate this purchasing power among the commodities they jointly produced (or even save some of it), in whatever proportions they choose. This can be described as the unallocated purchasing power (UPP) approach to exploitation. With this definition of surplus-value, the Marxian second equation immediately holds as an identity. The PNP-UPP LTV holds the rate of surplus-value rather than the consumption bundle of workers constant.

There is a sharp contrast between the PNP-UPP LTV and the traditional interpretation in the way they conceptualize distribution. Following Marx's procedure in his calculation, represented in the simplified example introduced earlier, it is impossible to assume that workers can buy the same bundle of commodities before and after the

redistribution of surplus-value, since the purchasing power they receive will be spent at different prices. Consequently, the wage must be changed to keep the bundle of workers' consumption unchanged (and the rate of surplus-value must be altered – hence the controversy). The UPP approach to exploitation conserves the rate of exploitation, or, more rigorously, measures the value of labour power as the value whose price form is the price of the commodities workers can buy: an unallocated purchasing power on any commodities. The rate of surplus-value, as in Marx's calculation, is unchanged.

A single-system approach and exploitation in any set of prices

A key aspect of the PNP-UPP LTV interpretation is that value is present in the theory of exploitation, as a social substance extracted in one place in the economy (firm, industry), and realized in another. But there is no logical anteriority in the value system, compared to the price system. This interpretation is a single-system approach to the LTV.

This property has important analytical consequences. There is only one economy, one system, not two. There is no 'underlying', hidden economy, which operates in 'values' where the distributional realities that structure the functioning of capitalism could be determined. The theory of exploitation is not dependent on the prevalence of any particular set of prices. The consideration of prices of production is not central to Marx's argument concerning exploitation, only an example that illustrates a much more general conclusion. Prices of production are just *one case* in which such a demonstration must be made, which Marx focused on because of the importance of this particular set of prices in competitive capitalism, as centres of gravitation of market prices.

The specific property expressed in the equality of the profit rate among industries cannot play any role in the theory of exploitation. Prices may deviate from prices of production because of gravitation; the amounts of surplus-value realized in each industry may also differ from what is implied by the prevalence of uniform profit rates because of the existence of non-reproducible resources and their rents; counteracting factors, such as monopoly, may also prevent equalization of profit rates. These deviations, inherent to capitalism, and also mentioned in Marx's analysis, do not invalidate his theory of value and exploitation.

An ongoing debate

The shift of perspective to single-system interpretations of Marx's labour theory of value has led to further debate in this vein. Fred Moseley (2003) proposes to apply the reasoning of the SS-LTV approach not just to variable capital, but to constant capital as well. Moseley argues for retaining the original form of the Marxian equations by defining the total value of a commodity as the labour-time equivalent of the price of constant capital plus the living labour expended in adding value. Moseley argues that Marx's comments in the quotations above are unnecessary because Marx's tables themselves express his underlying understanding of the labour theory of value.

Alan Freeman, Giugelmo Carchedi, Alan Kliman, and their co-authors (Freeman and Carchedi, 1996) have put forward a 'temporal single-system' (TSS) interpretation of the labour theory of value. This interpretation sets the transformation problem in a temporal context, defining the value of commodities as the sum of the labour time equivalent of constant capital (calculated using a monetary expression of labour time) and the living labour expended in the current period in production. By construction, this interpretation makes the first Marxian equation hold for the total product, while the second Marxian equation holds when the monetary expression of labour time is appropriately defined (as in the SS-LTV). It is, however, clear in Marx's analysis that the value of a commodity is not determined by the actual amount of labour its production required in the past, but by the labour time it requires under present prevailing conditions:

> … the value of commodities is not determined by the labour-time originally expended in their production, but by the labour-time expended in their reproduction, and this decreases continually owing to the development of the social productivity of labour. On a higher level of social productivity, all available capital appears, for this reason, to be the result of a relatively short period of reproduction, instead of a long process of accumulation of capital. (Marx, 1981, ch. 24)

This evaluation at 'replacement costs', however, does not imply that the economy is necessarily in a stationary state as the TSS critique has claimed.

A mathematical setting

The use of numerical examples to work out the quantitative implications of theoretical ideas is now outdated. The most common framework in the contemporary

literature on the transformation problem is a pure circulating-capital model with a single technique in each sector, in which basic properties of solutions and interpretations can be elegantly and compactly expressed. A single homogeneous labour input works with stocks of an arbitrary but finite number of produced commodities available at the beginning of a production period. One unit of each commodity is produced by a single technique of production. This framework is consistent with the example in the first table above but not with Marx's tables since the circulating capital model does not include fixed capital, while Marx's examples do.

1. *Techniques of production*. The number of goods is n, also the number of techniques. A technique of production, indexed by j, is characterized by a column vector, $\boldsymbol{a}_j=(a_{j1},\ldots, a_{ji},\ldots,a_{jn})$, and a scalar l_j, where a_{ji} is interpreted as the quantity of the commodity i required as inputs, and l_j as the quantity of labour required for the production of one unit of commodity j. A technology consisting of the set of all available techniques is described by collecting corresponding inputs into a matrix $\mathbf{A}$, and the labour input scalars into a row vector $\boldsymbol{l}'$. A pattern of economic production is described by a vector of levels of operation of the techniques, $\boldsymbol{x}=(x_1,\ldots,x_j,\ldots,x_n)$. The inputs required with this pattern of production can compactly be written in matrix notation as $\mathbf{A}\boldsymbol{x}$, while the total labour required is $\boldsymbol{l}'\boldsymbol{x}$.

2. *The determination of values*. The value, λ_j, of commodity j is the sum of the direct labour, l_j, expended in its production, and the indirect labour contained in produced inputs required for its production, $\lambda_1 a_{j1}+\cdots+\lambda_n a_{jn}=\lambda'\boldsymbol{a}_j$, that is $\lambda_j=\lambda'\boldsymbol{a}_j+l_j$. The vector of values of commodities, λ', satisfies the equation: $\lambda'=\lambda'\mathbf{A}+\boldsymbol{l}'$. It can be written as:

$$\boldsymbol{\lambda}' = \boldsymbol{l}'(\mathbf{I}-\mathbf{A})^{-1}$$

The value of the net product $\boldsymbol{y}=(\mathbf{I}-\mathbf{A})\boldsymbol{x}$, is equal to the total labour time expended: $\lambda'\boldsymbol{y}=\boldsymbol{l}'\boldsymbol{x}$. It is the sum of variable capital (wages paid), and total surplus-value. We denote τ as the rate of surplus-value, and v, the value of one unit of labour power, or the share of wages in the net product. These two variables are linked by the relationship $v=1/(1+\tau)$.

3. *The example of the table*. Each element in the table (upper-case notation) refers to industries, that is the product of unit variables (lower-case notation) by levels of operation (industries are marked by the subscript j, while vectors have no subscript). Below we will use the notation, P_j, for the price of the output of industry j, p_j for the price of one unit of commodity j, and $\boldsymbol{p}'$ for the vector of unit prices.

Constant capitals:	$C_j=\lambda'\boldsymbol{a}_j x_j$ and $C=\lambda'\mathbf{A}\boldsymbol{x}$.
Variable capitals:	$V_j=v\, l_j x_j$ and $V=v\,\boldsymbol{l}'\boldsymbol{x}$, with $v=1/(1+\tau)$ or $\tau=(1-v)/v$.
Total capitals:	$K_j=C_j+V_j$ and $K=C+V$.
Surplus-values:	$S_j=\tau V_j=(1-v)\, l_j x_j$ and $S=\tau V=(1-v)\,\boldsymbol{l}'\boldsymbol{x}$.
Values of commodities:	$\Lambda_j=K_j+S_j=(\lambda'\boldsymbol{a}_j+l_j)\, x_j=\lambda_j x_j$ and $\Lambda=K+S=(\lambda'\mathbf{A}+\mathbf{l}')\,\mathbf{x}=\lambda'\mathbf{x}$.

Marx determines the total surplus-value, S, and allocates it proportionally to total capital in each industry, so that the profit rates, r_j, in each industry is uniform: $r=S/K$ (or, equivalently, $1+r=\Lambda/K$). Profits in each industry are: $\Pi_j=r\,K_j$. By construction, total profits are equal to total surplus-value. The price of production of the total output of industry j is: $P_j=K_j+\Pi_j=(1+r)K_j$. For the price of one unit of commodity j, one has:

$$p_j = (1+r)(\boldsymbol{\lambda}'\boldsymbol{a}_j + l_j) \text{ and } \boldsymbol{p}' = (1+r)(\boldsymbol{\lambda}'\mathbf{A} + \boldsymbol{l}').$$

As is obvious, the two equations Sum of values ($\Lambda=\lambda'\mathbf{x}$)=Sum of 'prices of production' ($P=\boldsymbol{p}'\mathbf{x}$) and Sum of surplus-value (S)=Sum of profits ($\Pi=r\,K$) are satisfied.

4. *The determination of prices of production*. In the above calculation, Marx simply transfers the values of inputs to the price of production system instead of estimating them at their prices of production. Prices of production are a stationary price system (in which inputs have the same prices as outputs, as would be the case in a long-period equilibrium) at which profit rates in all sectors are equal to a given r, when the wage is paid at the beginning of the production period:

$$\boldsymbol{p}' = (1+r)(\boldsymbol{p}'\mathbf{A} + w\boldsymbol{l}'), \quad \text{which implies } \boldsymbol{p}'[r, w] = w(1+r)\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}.$$

The profit rate equalization conditions are n equations (one for each produced commodity) in n+2 variables, the n prices $\boldsymbol{p}'$, r, and w. Since the accounting units in which prices and the wage are expressed are arbitrary, it is possible without loss of generality to add one further equation normalizing prices, such as $\boldsymbol{p}'\boldsymbol{N}=1$, where $\boldsymbol{N}$ is a nonnegative bundle of commodities chosen as numéraire for the price system, or, alternatively $w=1$, which specifies the unit wage as the numéraire.
In the treatment of the transformation problem the most intuitive normalization is to express prices in labour time units. These prices are often called ‘direct prices’, and the general price level in this metric is determined by: $\boldsymbol{p}'\boldsymbol{y}=\boldsymbol{l}'\boldsymbol{x}$. The price of the net product $\boldsymbol{p}'\boldsymbol{y}$, evaluated at direct prices, is equal to the total labour time expended: $\boldsymbol{l}'\boldsymbol{x}$. This is equivalent to saying that the numéraire is the net product divided by the total number of hours expended: $\boldsymbol{N}=\boldsymbol{y}/\boldsymbol{l}'\boldsymbol{x}$. Using this numéraire one has:

$$\boldsymbol{p}'[r] = \frac{\boldsymbol{l}'x}{\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}y}\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}$$

Using this relationship and the expression of $\boldsymbol{p}'[r,w]$ above, one can determine the negative relation between wages and the profit rate, à la Ricardo and Sraffa:

$$w = \frac{1}{1+r}\frac{\boldsymbol{l}'x}{\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}y}$$

When the profit rate is 0, we have $w=1$, and $\boldsymbol{p}'=\boldsymbol{l}'(\mathbf{I}-\mathbf{A})^{-1}=\lambda'$: direct prices are equal to values.
5. *The historical transformation controversy*. The dual-system critique is based on comparing the aggregates (sum of values to sum of prices, and sum of surplus-values with sum of prices) under the assumption of a given real wage as a bundle, $\boldsymbol{d}$, of commodities. Thus, the value of labour power and surplus-value are respectively: $v=\lambda'\boldsymbol{d}$, and $S=(1-v)\boldsymbol{l}'\boldsymbol{x}$. Workers are assumed to buy the same commodities when prices of production prevail, so that $w=\boldsymbol{p}'\boldsymbol{d}$. Substituting $\boldsymbol{p}'[r,w]$, as above, for $\boldsymbol{p}'$ in this expression, the profit rate is the solution of the following implicit equation:

$$(1+r)\boldsymbol{l}(\mathbf{I} - (1+r)\mathbf{A})^{-1}\boldsymbol{d} = 1.$$

One can then calculate Π, which has no reason to be equal to S: in the general case, the second Marxian equation does not hold.
6. *The PNP-UPP LTV*. In the PNP-UPP LTV interpretation, in contrast, the same situation of distribution means the same rate of surplus-value. In general this means that

workers will not be able to buy the same bundle of commodities at prices of production. The rate of surplus-values is: $\tau p = \Pi / W$. If, in the two systems, the price of production of the net product is set equal to its value, of which it is the price form (or, equivalently, if the monetary expression of value is set to 1), that is $\boldsymbol{p}' \boldsymbol{y} = \lambda' \boldsymbol{y} = \boldsymbol{l}' \boldsymbol{x}$, then the total price of profits is equal to the sum of surplus-value, of which it is the price form. Thus the two Marxian equations (the first interpreted in terms of the net product) hold.

See Also

- absolute and exchangeable value
- classical production theories
- competition, classical
- labour theory of value
- linear models
- market price
- Marxian value analysis
- Marx's analysis of capitalist production
- natural price
- neo-Ricardian economics

Bibliography

Böhm-Bawerk, E. von. 1890. *Capital and Interest*. New York: Kelley and Millman, 1957.

Bortkiewicz, L. von. 1952. Value and price in the Marxian system. *International Economic Papers* 1952(2), 5–60.

Duménil, G. 1980. *De la valeur aux prix de production*. Paris: Economica.

Duménil, G. 1983. Beyond the transformation riddle: a labor theory of value. *Science and Society* 47, 427–50.

Duménil, G. 1984. The so-called ‘transformation problem’ revisited: a brief comment. *Journal of Economic Theory* 33, 340–8.

Duménil, G. and Lévy, D. 1984. The unifying formalism of domination: value, price, distribution and growth in joint production. *Zeitschrift für Nationalökonomie* 44, 349–71.

Foley, D.K. 1982. The value of money, the value of labor power, and the Marxian transformation problem. *Review of Radical Political Economics* 14(2), 37–47.

Foley, D.K. 2000. Recent developments in the labor theory of value. *Review of Radical Political Economy* 32(1), 1–39.

Freeman, A. and Carchedi, G., eds. 1996. *Marx and Non-equilibrium Economics*. Brookfield, Vermont: Edward Elgar.

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 26, 291–325.

Lipietz, A. 1982. The ‘so-called transformation problem’ revisited. *Journal of Economic Theory* 26, 59–88.

Marx, K. 1976; 1978; 1981. *Capital*, vols. 1, 2, and 2. New York: Random House.

Morishima, M. 1973. *Marx's Economics*. Cambridge: Cambridge University Press.

Moseley, F. 2003. Money and totality: Marx's logic in volume 1 of capital. In *The Constitution of Capital: Essays on Volume 1 of Capital*, ed. R. Bellofiore and N. Taylor. Basingstoke: Palgrave Macmillan.

Samuelson, P.A. 1971. Understanding the Marxian notion of exploitation: a summary of the so-called transformation problem between Marxian values and competitive prices. *Journal of Economic Literature* 9, 399–431.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

How to cite this article

Foley, Duncan and Gérard Duménil. "Marxian transformation problem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000400> doi:10.1057/9780230226203.1052

The New Palgrave Dictionary of Economics Online

Marxian value analysis

J.E. Roemer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

For Marx, the labour theory of value was not a theory of price, but a method for measuring the exploitation of labour. The exploitation of labour, in turn, was important for explaining the production of a surplus in a capitalist economy. In a feudal economy, the emergence of a net product, surplus to the consumption of producers and to the inputs consumed in production, was palpable. For the serf reproduced himself on his family plot of land during part of the week, and then worked for the lord, doing demesne or corvée labour during the other part. There was a temporal and physical division between production for subsistence or reproduction, and production which generated an economic surplus and was appropriated by the lord. Under capitalism, with the division of labour, such a demarcation no longer existed. If capitalism is characterized by competitive markets, where each factor is paid its true ‘value’, and no one makes a windfall profit by cheating his partner in exchange, how could a surplus emerge? In what manner could a sequence of equal exchanges transform an initial set of inputs into a larger quantity of outputs, with the surplus being appropriated systematically by one class, the capitalists? Marx's project was to explain the origin of profits in a perfectly competitive model, where each factor, including labour, received its competitive price in exchange.

Marx thought he had discovered the answer to this apparent economic sleight of hand by tracing what happened to labour as it passed from the workers who expended it, to the products in which it became embodied, and eventually to the profits of capitalists who sold these commodities. In some of his writings, notably in *Capital*, Volume I, he simplified the argument by assuming that the prices of goods were equal to the amounts of labour they embodied. The embodied labour in a good is the amount of labour necessary to produce that good, and to reproduce all inputs used up in its production. (Assume the only non-produced input is labour.) In particular, this is true also for the good ‘labour power’; the embodied labour in a week's labour supplied by a worker is the amount of labour necessary to produce the goods which that worker consumes to reproduce himself for work the following week. If all goods exchange at their embodied labour values (the simplifying assumption) then, in particular, the worker receives a wage in consumption commodities (say, corn) which is just necessary to reproduce himself (which includes the reproduction of the working-class family). The secret of accumulation, for Marx, lay in the discovery that the embodied labour value of one week's labour was, let us say, four days of labour. In four days of socially expended labour, given the existing technology and stock of capital, the

consumption commodities necessary to reproduce the worker could be produced. Thus the worker was paid an amount of corn which required four days to produce, his wage for seven days' labour. The surplus labour of three days became embodied in commodities which were the rightful property of the capitalist who hired the worker. Why would the worker agree to such a deal? Because he had no access to the means of production necessary for producing his consumption goods on any better terms. Those means of production were owned by the capitalist class. (Although the simplifying assumption, that equilibrium prices are equal to or proportional to embodied labour values, is rarely true, Marx conjectured that the deviation of prices from labour values was not crucial to understanding the origin of profits. On this point he was correct. Much ink has been spent on the 'transformation problem', which tries to relate embodied labour values to equilibrium prices in general. As will be shown below, prices need not be proportional to embodied labour values for the theory of class and exploitation to be sensible. Hence the study of the transformation problem is a pointless detour.)

Imagine a corn economy, where there are two technologies for producing corn, a Farm and a Factory: – Farm: 3 days' labour produces 1 corn output – Factory: 1 days' labour+1 corn (seed) produces 2 corn output. On the Farm, corn is produced from labour alone, perhaps by cultivating wild corn on marginal land. In the capital-intensive Factory technology, seed corn is used as capital. One unit of seed capital reproduces itself and produces one additional corn output with one day of labour. Suppose both techniques require one week for the corn to grow to maturity. Let there be 1000 agents, ten of whom each own 50 units of seed corn. The other 990 peasants own only their labour power. Suppose a person requires one corn per week to survive; his preferences are to consume that amount, and then to take leisure. Assume that if he owns a stock of seed corn, he is not willing to run it down: he must replenish the inputs which he uses up before consuming. What is an equilibrium for this economy, which is guaranteed to reproduce the stocks with which it begins?

Since there are only 500 bushels of seed corn, the required consumption of 1000 corn cannot be reproduced using only the Factory technology, since the seed capital of 500 must be replaced. Capital is scarce relative to the labour which is available for it to employ. The wage which the 'capitalists', who own the seed corn, will offer at equilibrium to those whom they employ will therefore be bid down to the wage which peasants can earn in the marginal Farm technology: 1/3 corn per day labour. At any higher wage, all peasants will wish to sell their labour to the capitalists, and there is insufficient capital to employ them all. (It is assumed peasants have no preference for life on the Farm over life in the Factory. All they care about is rate at which they can exchange labour for corn.) At the wage of 1/3 corn per day, 500/3 peasants become workers in the Factory, each working for three days, planting three units of seed corn, and earning a wage of one corn. This exhausts the capital stock. The remaining peasants stay on the Farm, and also earn one corn with three days' labour. The ten capitalists each work zero days; altogether, they make a profit of (500–500/3)=333.3 corn, after paying wages and replenishing their seed stock.

In the Factory technology, the embodied labour value of one corn is one day's labour; that amount of labour produces one corn output and reproduces the seed capital used. But the worker, at equilibrium, must work three days to earn one corn. This is so because he does not own the capital stock required for operating the efficient Factory method. His alternative is to eke out a subsistence of one corn by doing three days' labour on the Farm. The worker is said to be *exploited* if the labour embodied in the wage goods he is paid is less than the labour he expends in production. This is the case here, and it is evidently what makes possible the production of a surplus, in an economy where all agents wish only to work long

enough to reproduce themselves (and their capital stock). Note this last statement characterizes, as well, the capitalists: in this story, they get 333 corn profits and expend no labour, a result consistent with their having subsistence preferences, where each desires to work only so long as he must to consume his one corn per week.

Contrast this capitalist economy, where three classes have emerged – capitalists, workers, and peasants – to the following subsistence, peasant economy. Everything is the same as above, except the initial distribution of corn: let each of the 1000 persons own initially 0.5 corn. At equilibrium, each agent will work two days and consume one corn. First, he uses the Factory to turn his 0.5 seed corn into 0.5 corn net output, which costs him 0.5 days of labour; then he must produce another 0.5 corn for consumption, for which he turns to the Farm, where he works for 1.5 days. Each agent consumes one corn with two days' labour, an egalitarian society, which is classless. (There are other ways of arranging the equilibrium in this economy, in which one group of agents hires another group to work up its capital stock, while they, in turn, work on the Farm. But the final allocation of corn and labour is the same as in the equilibrium just described.) There is a fine point here: perhaps one should say, in both economies, that the amount of labour socially embodied in one corn is two days (not one, as written above), for that is what is required to produce society's necessary corn consumption given the capital stock and available techniques. This will not change the verdict that the workers in the capitalist economy are exploited, while no one is exploited in the egalitarian society.

Contrast these two economies, which differ only in the initial distribution of the capital stock. Inequality in the distribution of the means of production gives rise to: (1) the production of a surplus above subsistence needs, or accumulation; (2) exploitation, in the sense that some agents expend more labour than is embodied in the goods they consume and others expend less labour than is embodied in what the consume; and (3) classes of agents, some of whom hire labour, some of whom sell labour, and some of whom work for themselves. The exploitation of labour emerges with the unequal ownership of capital, or the 'separation' of workers from the means of production. The existence of an industrial reserve army (here, the peasantry) who have access to an inferior technology to reproduce themselves explains the equilibration of the wage at a level below that which exhausts the product of labour in the capitalist sector. Moreover, exploitation may be an indicator of an injustice of capitalism. If it does not seem fair that a serf must work three days a week for the lord perhaps it is not fair either that a wage labourer must expend more labour than is embodied in the wage goods he receives. That verdict, however, is not obvious and requires further analysis. Although the story can be made complicated, these simple models demonstrate the main features of the Marxian theory of labour exploitation.

Class, exploitation and wealth

Consider an economy of N agents, with n produced commodities and labour. The input–output matrix which specifies the linear technology is A, and the row vector of direct labour inputs needed to operate the technology is L. Agent i has an initial endowment vector of goods w^i and one unit of labour power. For simplicity, assume as above subsistence preferences: each agent wishes to earn enough income to purchase some fixed consumption vector b, and not run down the value of his initial endowment, valued at equilibrium prices. After working enough to earn that amount, he takes leisure. It is clear that each agent will only operate activities, at a given price vector, which generate the maximum rate of profit. Normalize prices by setting the wage at unity. For all activities to operate at equilibrium, the commodity

price vector *p* must satisfy:

$$p = (1 + \pi)(pA + L)$$
(1)

Prices *p* obeying (1) generate a uniform and hence maximal rate of profit π for all activities. (The only activities we observe are the ones reported in *A* and hence without loss of generality, we may assume the profit rate must be equalized for all sectors of production, since agents only operate maximal profit rate activities.)
The vector of embodied labour values in commodities is Λ:

$$\Lambda = \Lambda(1 - A)^{-1}$$
(2)

A worker, whose initial endowments are none except his labour power, must earn wages sufficient to purchase the subsistence vector *b*, which requries:

$$pb = 1$$
(3)

From these three equations, it can be demonstrated (see Morishima, 1973; Roemer, 1981) that:

$$\pi > 0 \text{ if and only if } \Lambda b < 1$$
(4)

Equivalence (4) was coined by Morishima the ‘fundamental Marxian theorem’, as it shows that profits are positive precisely when labour is exploited (for the second inequality says that the labour embodied in the wage bundle is less than one unit of labour).
An agent in this model minimizes the labour he expends subject to earning revenues sufficient to buying his consumption *b*, and to replace the finance capital he uses. Suppose, for simplicity, there is no borrowing and all production must be financed from initial wealth. In general, an agent will optimize by

hiring some labour, selling some of his own labour, and/or working on his own capital stock. Let x^i be the vector of activity levels which agent *i* operates himself, financed with his wealth; let y^i be the vector of activity levels he hires others to operate, which he finances; let z^i be the amount of labour he sells to other operators. His problem is to choose vectors x^i, y^i, and z^i to:

$$\min Lx^i + z^i$$

subject to

$$\text{(i)}\ pAx^i + pAy^i \leq pw^i \text{(ii)}\ p(q - A)x^i + p(q - A)y^i - Ly^i + z^i \geq pb$$

The first constraint requires him to finance the activities operated out of his endowment, and the second requires that his revenues, net of wages paid and replacement costs, suffice to purchase the consumption bundle *b*. As well as the price vector satisfying (1), equilibrium requires that the markets for production inputs, consumption goods, and labour must clear. It can be proved that at such a ‘reproducible solution’, society is divided into five classes of agents, characterized by their relation to the hiring or selling of labour, as follows. There is a class of *pure capitalists*, who only hire labour (y^i is non-zero, but x^i and z^i are zero vectors); there is a class of *mixed capitalists*, who hire labour and work for themselves as well $(y^i \neq 0 \neq x^i, z^i = 0)$; there is a class of *petty bourgeoisie*, who only work for themselves, and neither hire nor sell labour $(x^i \neq 0; y^i = 0 = z^i)$; there is a class of *mixed proletarians*, who work for themselves part-time, and also sell their labour power on the market $(x^i \neq 0 \neq z^i, y^i = 0)$; and there are *proletarians*, who only sell their labour power $(z^i \neq 0, x^i = 0 = y^i)$. It is clear, from consulting the agent's programme, that this last class comprises those agents who own nothing but their labour power. More generally, the *Class-Wealth Correspondence Theorem* states that the five classes named, in that order, list agents in descending order of wealth. This verifies an intuition of classical Marxism.
There is, as well, a relation of class to exploitation. The *Class–Exploitation Correspondence Principle* states that the agents who hire labour are exploiters and the agents who sell labour are exploited. The exploitation status of agents in the petty bourgeoisie is ambiguous. Exploitation is defined as before: an agent is exploited if he expends more labour than is embodied in the vector *b*, and he is an exploiter if he expends less labour than that. It is important to note that this relationship of class to exploitation is a theorem of the model, not a postulate. Both the class and exploitation status of an agent emerge in the model as a consequence of optimizing behaviour, determined by the initial distribution of endowments, technology and preferences. These aspects of agents which in classical Marxism were taken as given (their class and exploitation status) are here proved to emerge as part of the description of agents in equilibrium, from initial given data of a more fundamental sort (endowments, etc.). For this reason, the

model described provides microfoundations for classical Marxian descriptions. Generalizations and discussion of the model are pursued in Roemer (1982, 1985a). See Wright (1985) and Bardhan (1984, ch. 13) for empirical applications. For a general evaluation of the Marxian theory of exploitation and class, see Elster (1985, ch. 2, 4 and 5).

From the viewpoint of modern capitalism, many criticisms can be levelled against these stories. Foremost among them, perhaps, is the assumption of subsistence preferences. What happens if agents have more general preferences for income and leisure? The Class–Exploitation Correspondence Principle continues to hold, but the correspondence between class and wealth may fail. It fails, however, only for preference orderings which are unusual: the Class–Wealth Correspondence is true if the elasticity of labour supplied by the population viewed cross-sectionally with respect to its wealth is less than or equal to unity. There can, therefore, be no general claim that exploitation corresponds to wealth, in the classical way – that the poor are exploited by the rich. Whether the exploitation–wealth correspondence holds depends on the labour supply behaviour of agents as their wealth changes.

Exploitation as a statistic

Note that the fundamental conclusions of classical Marxian value analysis – the association of exploitation with class, in a certain way, and the association of exploitation with profits and accumulation – hold even when equilibrium prices are not proportional to labour values. For the prices of equation (1) are not, except in a singular case, proportional to the labour values of equation (2). Therefore, the usefulness of exploitation theory need not rest upon the false labour theory of value. It is for this reason that the transformation problem, for so long a central concern in Marxian economics, is unimportant.

That usefulness, instead, depends on how good a statistic exploitation is for the phenomena it purports to represent. Does the exploitation of labour explain accumulation? The ‘fundamental Marxian theorem’ would seem to say so. But, in fact, it can be shown that in an economy capable of producing a surplus, every commodity can be viewed as exploited, not just labour power. If corn is chosen as the value numeraire, then the amount of corn value embodied in a unit of corn is less than one unit of corn, so long as profits are positive. Thus labour power is not unique, as Marx thought, in regard to its potential for being exploited, and it is a false inference that the exploitation of labour ‘explains’ profits any more than the exploitation of corn or steel or land does. (For versions of this ‘generalized commodity exploitation theorem’, see Vegara (1979), Bowles and Gintis (1981), Samuelson (1982), and Roemer (1982).)

Is exploitation a good statistic for the injustice of capitalist appropriation of the surplus? Only if the initial distribution of endowments, which gives rise to such appropriation, is unjust. Marx claimed this was so, by arguing that initial capitalist property was established by plunder and enclosure (*Capital*, Volume 1, Part 8). But suppose there were a clean capitalism, in which initial inequalities in the ownership of capital were generated by differential hard work, skills, risk-taking postures, and perhaps luck of the agents. Would the ensuing class structure, exploitation and differential wealth indicate an injustice, or would it reflect the consequences of persons exercising traits which are rightfully theirs, and from which they deserve differentially to benefit? These topics are pursued in Cohen (1979) and Roemer (1985b).

In sum, the Marxian theory of exploitation is liberated from the labour theory of value. The link between

class and exploitation is robust; but Marx's claim that the exploitation of labour is the unique explanans of accumulation is false. If one's class, defined above as one's relation to hiring or selling of labour, is important sociologically in determining behaviour (such as collective action against another class) and preferences, then the positive theory of class determination described is of use. Exploitation remains a statistic, of some value, for the inequality in the distribution of productive assets. But in this role, exploitation may not correspond to wealth as in the classical story: if the labour supplied by agents responds with excessive enthusiasm to increases in their wealth, then the rich can be ‘exploited’ by the poor. The ethical conclusion from an observation of exploitation is in this case unclear.

Even aside from this peculiar case, exploitation is a circuitous proxy for differential wealth in productive assets, and one's normative evaluation of exploitation depends on one's view of the process that generates that inequality. If agents are the rightful owners of their alienable means of production, because they accumulated them based on the exercise of their rightfully owned talents and preferences then exploitation does not represent unjust expropriation. If agents are not entitled to own alienable productive assets, either because they have no right to their talents and preferences (whose distribution is morally arbitrary), or because they came to possess those assets in some other unjustifiable way, then exploitation represents an expropriation. Inheritance, for example, might be an unjust way of acquiring assets which were originally acquired in an untainted manner. The essential question which lies behind the theory of exploitation concerns the fairness of a system of property allowing private ownership of alienable productive assets. The concept of exploitation based on the calculation of surplus labour accounts is, in this writer's view, a circuitous route towards the discussion of that central issue.

Ethical views concerning what kinds of asset may justifiably be privately appropriated change through history. Property in other persons, as in slavery, or more limited rights over the powers of other persons, as in feudalism, are no longer viewed as legitimate. The Marxian theory of exploitation is associated with a call for the abolition of private property in the productive assets external to persons. (Marx himself did not explicitly base his call for the abolition of such property on grounds of fairness, but on grounds of efficiency, despite the clear ethical tone of his attacks on capitalism. For an evaluation of the debate surrounding this question, see Geras (1985).) The cogency of that call must be established independently of the theory of exploitation.

See Also

- exploitation
- labour theory of value
- market price
- Marx, Karl Heinrich

Bibliography

Bardhan, P. 1984. *Land, Labor and Rural Poverty: Essays in Development Economics*. New York: Columbia University Press.

Bowles, S. and Gintis, H. 1981. Structure and practice in the labor theory of value. *Review of Radical*

Political Economics 12(4), 1–26.

Cohen, G.A. 1979. The labor theory of value and the concept of exploitation. *Philosophy and Public Affairs* 8(4), Summer, 338–60.

Elster, J. 1985. *Making Sense of Marx*. Cambridge: Cambridge University Press.

Geras, N. 1985. The controversy about Marx and justice. *New Left Review* No. 150, March–April, 47–85.

Morishima, M. 1973. *Marx's Economics*. Cambridge: Cambridge University Press.

Roemer, J.E. 1981. *Analytical Foundations of Marxian Economic Theory*. Cambridge: Cambridge University Press.

Roemer, J.E. 1982. *A General Theory of Exploitation and Class*. Cambridge, Mass.: Harvard University Press.

Roemer, J.E. 1985a. *Value, Exploitation, and Class*. London: Harwood Academic Publishers.

Roemer, J.E. 1985b. Should Marxists be interested in exploitation? *Philosophy and Public Affairs* 14(1), Winter, 30–65.

Samuelson, P. 1982. The normative and positivistic inferiority of Marx's values paradigm. *Southern Economic Journal* 49(1), July, 11–18.

Vegara, J.M. 1979. *Economia politica y modelos multisectoriales*. Madrid: Editorial Tecnos.

Wright, E.O. 1985. *Classes*. London: New Left Books.

How to cite this article

Roemer, J. E. "Marxian value analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000098> doi:10.1057/9780230226203.1053

The New Palgrave Dictionary of Economics Online

Marx's analysis of capitalist production

Duncan Foley and Gérard Duménil

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses Marx's analysis of capitalism, including the concepts of historical materialism, class society, exploitation, commodity, value, money, capital, labour-power, value of labour-power, surplus-value, constant and variable capital, commodity law of exchange, capitalist law of exchange, equalization of the profit rate, prices of production, absolute and relative surplus value, the circuit of capital, simple and expanded reproduction, capital accumulation, centralization and concentration of capital, technical change, reserve armies of labour, rent, interest, commercial and bank profit, the falling rate of profit, viable technical change, and cyclical crises.

Keywords

absolute rent; abstract labour time; biased technical change; business cycles; capital accumulation; capitalism; capitalist law of exchange; circulating and fixed capital; circulation of capital; class; commercial, industrial and money-dealing capital; commodity; commodity law of exchange; commodity money; competition; composition of capital; concentration and centralization of capital; constant and variable capital; crises of overproduction; differential rent; division of labour; exchange value and use value; exploitation; fictitious capital; fixed capital; historical materialism; increasing returns; industrial capital; innovation; iron law of wages; labour power; labour theory of value; law of value; Marx's analysis of capitalist production; Marxian transformation problem; money; money-dealing capital; monopoly; organic composition of capital; over-accumulation of capital; population growth; primitive accumulation; private property; productive and unproductive labour; profit rate; rate of exploitation; rate of surplus value; Ricardo, D.; Say's Law; slavery; Smith, A.; socialism; surplus; surplus value; technical change; transformation problem; use value; value; variable capital; velocity of circulation

Article

Karl Marx's analysis of capitalist production is best understood in the context of his broad theory of human societies and their history, namely, *historical materialism*. This theory argues that, after passing through various stages in which societies are divided into classes and the exploitation of a majority of producers by a privileged minority prevails, humanity will finally eliminate classes and class domination by a revolutionary process conducted by the organized *proletariat* in capitalism. This revolutionary stand was based on a ‘scientific’ investigation of history in general and capitalism in particular, with a special emphasis on economics, always with a political perspective. Whether historical materialism has a scientific or ideological character obviously remains controversial between Marxists and non-Marxists: Marxist theory is considered a discredited doctrine of the past by non-Marxists, while Marxists consider mainstream social and economic thinking as a continuing apologetics of capitalism.
After an introductory section devoted to locating the capitalist mode of production as a particular epoch in human history, the main focus below is on Marx's analysis of capitalist production. There are two facets to the theory of capital in the strict sense: *surplus value* (exploitation), and *the circuit of capital* (its ‘circulation’). These are introduced separately, and then gradually combined in the analysis of more complex phenomena. Finally, we consider three broad sets of basic mechanisms directly related to the hold of capital on the functioning of the economy: (1) competition, (2) accumulation, technological and distributional changes, and (3) crises and the business cycle. We do not consider other important aspects of Marx's thinking such as his analysis of class struggle, and his theory of the state. The interpretation of even very fundamental aspects of Marx's thought remains contested among Marxist scholars. The bibliography contains a selective list of works that represent some of these different perspectives.

The capitalist mode of production

The historical materialist point of view starts from the observation that all human societies must produce in order to reproduce both individuals and society itself. Production in this general sense always involves the combination of human labour with previously produced means of production and the natural resources of the earth. With the emergence of settled agriculture a *surplus product* over and above what is necessary for reproduction becomes possible. In societies with a surplus product, class exploitation, an institutionalized form of inequality, arises. Societies divide into a small *exploiting class* which appropriates, controls, and distributes the surplus product created by the labour of a much larger *exploited class* of producers who receive on average only what is necessary for their reproduction. Marx and Engels distinguish two aspects of these *class societies*. The *forces of production* comprise the population, natural resources, and technology which make a surplus product possible; the *social* relations *of production* comprise the institutional framework (such as property relations) through which the exploiting class appropriates the surplus product. The forces and social relations of production together constitute a *mode of production*. For example, in the slave mode of production characteristic of ancient Greek and Roman civilizations, the institution of slavery sustained by military force and political power was the means through which slave-owners appropriated a surplus product created by the labour of slaves, who received a minimum subsistence. In the feudal mode of production, the institutions of serfdom sustained by military force and religious and political power were the means through which the lords of the manor appropriated a fraction of the labour time of serfs, who also laboured in their own fields to feed and reproduce themselves (or the serf had to pay a rent in kind or, later, in money, in addition to various taxes). This is what *exploitation* means in Marx's thought: to live on the product of

the labour of other people.

From the historical materialist point of view, capitalism is a class society in which the institutions of private property in the means of production and free wage labour are the means through which capitalists appropriate the *surplus value* created by workers producing commodities (or services), who receive wages. In feudalism, the exploitation of the serfs was transparent: the serfs worked a certain part of the week on their own plots for their own subsistence, and a certain part of the week on the lord's land to supply his consumption and armies. Marx's theory of capitalism demonstrates that, though the mechanism of capitalist exploitation through the social relation of wage labour based on the formal legal equality of workers and employers is less transparent, capitalists also appropriate the surplus labour time of the workers. Capitalism, therefore, defines a specific stage of the history of class societies.

Capitalism's decentralized, highly competitive organization creates powerful incentives for the rapid development of the forces of production through population growth, technical innovation, and a widening division of labour, but it is unable to control the forces it has itself stimulated.

Marx and Engels expected that the capitalist working class (the *proletariat*), once it had a clear understanding of capitalist exploitation and reached a high degree of organization, would overthrow the social relations of capitalism in a revolution to establish a classless society based on social control of the large surplus product made possible by the forces of production developed by capitalism. A violent transition was required, the *dictatorship of the proletariat*, to attain socialism and finally communism, marking the end of the ‘prehistory’ of humanity. Marx developed this analysis in collaboration with Friedrich Engels in *The Communist Manifesto* (1848).

Marx's main work, *Capital*, is devoted to the analysis of capitalist production. The first volume was published, in 1867, while Marx (1818–83) was still alive. Volumes II and III were published later by Friedrich Engels, from extensive notebooks still in draft form at the time of Marx's death. In what follows, we refer to *Capital* by volumes and chapters; for example, ‘III, 25’ means Chapter 25 of Volume III. References and quotations can be found on Internet, for example in the *Marx/Engels Library*, http://www.marxists.org/archive/marx/works, or in Marx (1976; 1978; 1981). We have put square brackets around our own interpolations in quotes; everything else comes from the source.

The definition of capital (I, 4)

Marx defines capital as *value* (to be defined below) participating in a dynamic process of self-expansion. A capitalist spends money to hire workers and buy means of production, and then sells the resulting output for enough money to cover his initial outlay and secure a profit (the form taken by ‘surplus value’). In this process value appears in various forms: first under the form of money; then as the value of productive inputs (labour power, raw materials, machinery, and buildings); then as the value of the commodities produced; and finally as money value again after the produced commodities have been sold. This process of capital is pointless unless, as is normally the case when capitalists make a profit, the money realized in the sale of commodities is greater than the money initially spent to start the process. Capital is not value as such, but value in movement:

> If we pin down the specific forms of appearance assumed in turn by self-valorising value, in the course of its life, we reach the following elucidation: capital is money, capital is commodities. In truth, however, value is here the subject of a process in which, while

> constantly assuming the form in turn of money and commodities, it changes its own magnitude, throwing off surplus-value from itself considered as original value. (I, 4)

Two aspects of capital are present in this definition: (1) capital is expanding value; and (2) capital value changes its form. These two aspects of capital are also called the *process of self-expansion* (sometimes called *valorization*), and the *process of circulation of capital* (or circuit of capital). Marx means here that: (1) the capitalist invests a certain capital with the intent of making profits (expansion); (2) capital is invested in commodities and money, and constantly passes from one form to the other (for example, when an output is sold, value changes form from commodity to money).

The first two volumes of *Capital* treat the processes of *self-expansion* and *circulation* of capital separately (with a few exceptions); the third volume considers the combination of these two elements. Before entering into the analysis of capital, it is necessary, however, to introduce two other preliminary concepts, *commodity* and *money*, and the related concepts of *value* (at the centre of the definition of capital) and *price*, to which Marx devotes the first three chapters of Volume I, prior to the analysis of capital. In Volumes I and II, the three concepts are considered successively: commodity (including value), money (including price), and capital (valorization and circulation). (This outline is logical, not historical: historically commodities and money reach their full development only with the capitalist mode of production.) We will follow this outline in our exposition here.

Commodities, value, money, and prices

Commodities and value (I, 1)

A *product* is the result of human labour, working with produced means of production and the natural resources of the earth. Useful products become *commodities* when they are regularly exchanged rather than being consumed directly by their producers. ‘Useful’ must be taken in a very broad sense as something desired by someone, for whatever reason. A producer who exchanges his product receives social recognition for his own labour in the form of the other commodities he acquires. Marx denotes the labour time required for the production of a commodity under average conditions, as *socially necessary labour time*. As the outcome of a parcel of social labour time, the commodity has an *exchange value*, or more briefly a *value*. Thus, according to Marx (who here follows Adam Smith), a commodity has a dual character as: (1) *object of utility*, or equivalently a *use value*, and (2) an *exchange value*, or *value*. The value of the commodity is the sum of labour embodied in previously produced inputs, *dead labour*, and newly incorporated labour, *living labour*. Marx sometimes calls this definition the *law of value*, although he rarely uses the expression. Later economists often refer to this framework as the *labour theory of value*.

The dual character of the commodity is reflected on labour itself. The concrete quality of labour (weaving, computer-programming) corresponds to the use-value aspect of the commodity it produces. But all categories of social labour materialized in the production of commodities have in common the ability to produce exchange values and, as such, are defined as *abstract labour*. There is no a priori rule accounting for this process of abstraction. Exchange dissolves the specific character of concrete labours, and the repetition of exchange establishes their quantitative equivalence. If one category of concrete labour is not adequately compensated, its supply will decline, and its wage will rise. In a similar manner,

it is exchange which establishes the normal degree of intensity, skill, and technical efficiency in production.

Abstracting from the capitalist character of production, commodities would ‘normally’ exchange in proportion to their values. For example, if the value of commodity A is twice that of commodity B, one unit of A will exchange for two of B. If the exchange ratio were only one B for one A, producers of A would switch to producing B; a shortage of A would ensue and the exchange value of A would rise. This is the *commodity law of exchange*, sometimes confused with the law of value. The distinction is important because the law of value is a fundamental characteristic of commodity production, whether commodities exchange in proportion to their values or not. (In competitive capitalist economies they typically do not, as we will see.)

Money and prices (I, 3)

We begin with the definition of money, and its first function as *measure of value*, and introduce the other functions of money, and the concept of the *price form* of value.

The value of commodities cannot be expressed on the market directly in abstract labour time (which nobody can observe or measure). In the exchange of two commodities, such as linen for a coat, the value of one commodity is expressed in the body of the other (measured in units such as a length or weight) as its direct *equivalent*. With the repetition of exchange, some specific commodity, such as gold, will emerge as a *socially accepted general equivalent*, that is, as *money*. Thus for Marx the original function of money is as *measure of value*. In addition to its function as measure of value, money comes to serve as *medium of circulation* if purchases and sales are paid for directly, and as *means of payment* if payment is deferred. Value can be accumulated temporarily in money hoards. Another function of money is, therefore, as a *store of value* (though any durable, valuable commodity can serve as a store of value).

Prices are values as expressed in monetary units. They are *forms of value*. When commodities exchange at prices proportional to their values, the price of a commodity expresses the socially necessary (abstract) labour time required for its production of this commodity, qualitatively and quantitatively in a straightforward manner. This is the framework of Volumes I and II. But the prices of commodities may deviate from their values, and we will later return to this issue. The State can establish a *standard of price* by defining a local currency unit such as the franc or dollar as a certain amount of gold or other money commodity. Valueless tokens, ‘symbols or tokens of value’ in Marx's words, such as paper currency, may also be circulated in place of commodity money:

> In the same way as the exchange-value of commodities is crystallised into gold money as a result of exchange, so gold money in circulation is sublimated into its own symbol, first in the shape of worn gold coin, then in the shape of subsidiary metal coin, and finally in the shape of worthless counters, scraps of paper, mere *tokens of value*. (Marx, 1859, 2.B.2. c)

Money also takes the form of a stock of purchasing power in an account in a financial institution. In contemporary capitalism, there is no commodity money.

The monetary expression of value and the quantity of money

Inherent in Marx's theory is the relation between abstract labour time and its price form in money terms. There is a quantitative aspect to this relation. The ratio – for example, dollars per hour of abstract socially necessary labour time – can be called the monetary expression of labour time, or the monetary expression of value.

The determination of this ratio, which is a way of looking at the general price level in an economy, is discussed by Marx in his critique of Ricardo's quantity of money theory of prices, under the assumption of the existence of a commodity money. Marx explains that the quantity of money required to circulate the mass of commodities produced in any period depends on the quantity of the commodities exchanged, their money prices, determined by their costs of production, and the velocity of money, the average number of transactions in which each unit of money participates in the period (an institutional characteristic). Money flows in and out of hoards (reserves) to accommodate the requirements of circulation. He interprets this principle as governing the quantity of money required for purchases and sales, in contrast to Ricardo's quantity of money theory of prices, which sees the prices of commodities adjusting to a given quantity of money. In Marx's theory the general level of prices is determined by the relative costs of production of the money commodity and other commodities when a commodity like gold is used as money. (The critique of Ricardo's theory is developed in Marx, 1859, 2.C.)

The theory of surplus value

The labour theory of value is the foundation of Marx's theory of exploitation, or surplus value. When a produced commodity is purchased or sold no new value is created. If a commodity sells at a price proportional to its value, given the monetary expression of value, the buyer and seller exchange money and commodity representing equal values. If the commodity sells above or below its value, the value gained by one party is just offset by the value lost by the other.

Productive labour-power and surplus-value (I, 7 – 9)

Marx explains surplus value in relation to the purchase of the labour power of waged workers. The capability to work, denoted as *labour power*, is a commodity, with a use-value and a value. The *use value* of labour power is labour itself, because a capitalist buys labour power to obtain the right to use the labour of the worker. The *value* of labour power is the value equivalent of the purchasing power of the wage on the commodities the worker can buy. (We will discuss later Marx's view of the actual purchasing power of workers.)

Only ‘productive workers’, that is, workers involved directly in production within capitalist enterprises, produce new value in Marx's analysis, in contrast to ‘unproductive workers’, whose labour power is employed by capitalists to maximize their profit rate. If the value of the labour power of productive workers is less than the value they produce, capitalist production on average adds more value in the production of commodities than it expends in hiring workers. (One can, equivalently, say that the money wage must be smaller than the monetary expression of the labour time expended by the average worker.) Because capitalist production can produce a surplus over the subsistence of productive workers,

typically the value of labour power is smaller than the value labour produces, and surplus value results. Thus, labour power has a property not shared by other commodities. While the purchase and sale of a produced commodity can only redistribute a given value between buyer and seller, the capitalist's purchase and use of labour-power, in contrast, results in the creation of surplus-value. The capitalist buys labour-power at a wage reflecting the necessary labour time required by the production of the consumption basket of the worker, say, four hours a day, but on average the worker can work longer, say, eight hours. Thus, the capitalist can appropriate *surplus labour*, here four hours, in the form of surplus value. (If the monetary expression of labour time is ten dollars per hour, the surplus value created by an average worker in a day under these assumptions would be 40 dollars.) Under the wage system, once a capitalist has paid a worker the agreed wage, the product of the worker's labour and its value belong to the capitalist. The production of surplus value is thus compatible with transactions at prices proportional to values, including the purchase of labour power at a wage proportional to the value of productive labour power. Marx argues that capitalist exploitation does not violate the commodity law of exchange, that is, it would take place even if all commodities exchanged at prices proportional to their values.

The actual appearance of labour power available for hire historically depends on two preconditions. First, workers must be legally free to sell their labour power. This explains the historic hostility of capitalism to bound forms of labour such as serfdom and slavery. Second, workers cannot have access to their own means of production, such as the feudal commons, so that they have no choice but to sell their labour power to the owner of means of production to live. This explains the historic support of capitalism for the enclosure of common lands and their conversion into private property. Marx devotes the last part of Volume I to *primitive accumulation*, the actual historical process through which the capitalist mode of production came into being. There he shows how, in the first steps of accumulation in England, the availability of labour power was achieved by way of straightforward social violence. The *enclosure* of common lands deprived the rural population of its old conditions of reproduction, and subjected it to the dependency on capital. It is important to keep such mechanisms in mind in the investigation of the historical dynamics of capitalism. Marx emphasizes the crucial historical importance of the transformation of produced means of production and labour, which are universal aspects of human production, into the specific commodity forms of capital, including labour power.

The value of the produced inputs the capitalist purchases to undertake production is recovered in the sales price unchanged, so that Marx calls it *constant capital*, denoted by the symbol c. The value of the labour power the capitalist buys as an input to production, on the other hand, is recovered in the sales price expanded by the addition of the surplus value, so that Marx calls it *variable capital*, denoted by the symbol v. The sum of constant capital, c, variable capital, v, and surplus value, s, is the total value of the product. The sum $c+v$ is the total cost of the commodity. The sum $v+s$ is the *living labour*, as opposed to *dead labour*, c, and measures the *value added* by the production process. The *rate of surplus value*, s/v, is the ratio of *unpaid* to *paid* labour time, so that Marx also calls it the *rate of exploitation*. The ratio c/v, which measures the ratio of dead to living labour in the cost of the commodity, is the *value composition* of capital.

This decomposition of the value of a commodity is parallel to the income statement of a capitalist firm, which exhibits *profit* (Marx's surplus-value, s) as the difference between sales price (Marx's value of the commodity, $c+v+s$), and the cost of the means of production and wages required to produce the commodity (Marx's $c+v$).

Absolute and relative surplus value, manufacture and industry (I, 12 – 16)

Identifying surplus value as surplus labour time does not tell what determines its magnitude and variation. Many natural, social and political conditions are involved, and vary historically. Labour performed by members of the family at home, women in particular, crucially affects the level of exploitation compatible with the reproduction of the workers and their families. In his analysis of surplus value in Volume I, Marx introduces important developments concerning the historical transformation of technology and organization.

Surplus value can be increased in two analytically distinct ways (which can be combined in real production): first, by lengthening the duration of labour time without increasing the value of labour power, *absolute surplus value*; second, by diminishing the value of labour power by cheapening worker's consumption through productivity gains holding the duration of labour time constant, *relative surplus value*. In Marx's view relative surplus value is the origin of the most important developments in the historical transformation of the organization of labour and technology by capitalism.

Marx sees distinct periods in which this transformation of production took different forms. In ‘manufacture’, a large number of individual workers, each processing his or her own means of production, are brought together in one location primarily for the purpose of increasing the capitalist's surveillance and control of production (which Marx describes as the ‘formal subsumption’ of labour to capital). In ‘large-scale industry’, the capitalist takes the further step of imposing a detailed division of labour on the production process, transforming the workers’ relation to the production process (which Marx describes as the ‘real subsumption of labour to capital’). Both technology and organization enter into these transformations. In manufacture, workers originally worked with the same tools they previously used in production at home; in large-scale industry, by contrast, capital has completely transformed technology and the organization of labour.

We will return to Marx's theory of technical change in capitalism in the discussion of the falling rate of profit.

The circulation of capital

As defined earlier, capital is self-expanding value moving through various forms (money, commodity…). We now turn to the analysis of the circulation of capital. The emphasis is on the motion from one form to the other, and the coexistence of the various fractions of capital under the three forms at a given point in time.

The circuit of capital (II, 1 – 4)

A capitalist spends money to buy inputs (means of production and labour power); organizes production; stockpiles and sells the resulting product; and realizes a certain amount of money in sales revenue, normally larger than the original capital outlay. Each atom of capital goes through the various *forms*: money-capital, *M*, commodity capital in the form of inputs to production, *C*, *productive capital*, *P*, the value of partially finished commodities and plant and equipment in the workshop, and again commodity capital in the form of inventories of commodities awaiting sale, C', and finally returning to money

through the sale of the produced commodities, M'. Marx represents this sequence in a diagram of the *circuit of capital*:

$$M - C \ldots P \ldots C' - M'$$

Here M is the money the capitalist uses to buy inputs to production C, P represents the actual production process, and C' are the produced commodities which are sold for money M'. The dashes represent purchase and sale of commodities on the market. The circuit is a chain, which can be viewed as beginning in M, C, or P, the *circuits of money, commodity*, and *productive capital*, three distinct formula of the same circuit.

The speeds at which the values of the various components of capital go through the productive form of capital, P, can be quite different. The value of some components, like raw materials, returns quickly to the money form in the sale of the commodity, while others like the value of buildings and machinery (whose value is only transferred to the product along their service life) returns only after a long period of time. From these differences in turnover time follows the distinction between *circulating* and *fixed capital*.

Capital is also a stock of value at any point in time. All the circuits overlap simultaneously: at the same moment new means of production and labour-power are being purchased while production is going on and finished output is being sold. The capital of a capitalist is the total value, tied up at any moment in these circuits. The total capital, K, is divided into three component stocks: *money capital*, M, *commodity capital*, C, and *productive capital*, P. The sum $K=M+C+P$ parallels the total of the assets on the capitalist's balance sheet.

Industrial, commercial and money-dealing capital (III, 16; III, 19)

Industrial capital undergoes the complete circuit of capital as above, taking on the forms M, C, and P in turn. Some capitals, however, are specialized to limited segments of the circuit. The first is *commercial capital*, which buys finished commodities from industrial capitalists to sell them to final purchasers, in the reduced circuit $M—C—M'$: commercial capitalists buy in order to sell the same commodity. The second category, *money-dealing capital*, refers to the technical activity of banks in handling money payments into and out of accounts (and the exchange of currencies). Since no productive labour is expanded in these circuits, no surplus value is created. How industries engaged in such activities can make profits is part of the theory of competition considered below.

Marx's schemes of reproduction (II, 18 – 21)

Although Volume II of *Capital* is devoted to the circulation of capital, the analysis of the *schemes of reproduction*, combines valorization (c, v, s) and circulation (M, C and P).

Three departments are distinguished which produce the physical commodities to satisfy the demand emanating from c, v, and s: Department I produces means of production, Department II commodities

consumed by workers, and Department III commodities consumed by capitalists. If all of the surplus value is consumed, no accumulation takes place, and the size of the capitalist economy remains unchanged, the case of *simple reproduction*. If a fraction of the surplus value is accumulated, the corresponding purchasing power is spent on additional means of production, and the capitalist economy expands, the case of *expanded reproduction*.

Marx assumes that all capital in the three industries accomplishes exactly one circuit: at the beginning and at the end of the period, all capital is assumed to be under the form C (the stocks of means of production and worker and capitalist consumption goods waiting to be sold). In this setting reproduction requires certain proportionalities to hold: for example, in simple reproduction the value added of Department I must equal the constant capital of Departments II and III.

In this framework, Marx considers two types of issues. The first issue is the definition of output and its relation to income. The *net product* is the value of the final product, C', minus the value of what is now denoted as intermediate inputs, either produced in the previous period, in C, or during the present period but purchased as inputs by firms. Marx shows that the value of this net product is equal to total income or *value added*, as in contemporary national accounting, the sum of wages and surplus value (including rent, interest and profit as we will see): $v+s$. Second, Marx investigates the circulation of money. He attempts to demonstrate how the money thrown into circulation by capitalists returns as sales revenue, taking into account the activities of a sector producing the money commodity if such a money commodity exists.

The functions of the capitalist and their delegation to employees (II, 6)

Being a capitalist is not a sinecure: both the appropriation of surplus value and the circuit of capital require active attention. In contemporary language: enterprises must be managed. Marx refers to these tasks as ‘capitalist functions’, in particular commercial transactions:

> The transformations of the forms of capital from commodities into money and from money into commodities are at the same time transactions of the capitalist, acts of purchase and sale. The time in which these transformations of forms take place constitutes subjectively, from the standpoint of the capitalist, the time of purchase and sale; … the time in which the capitalist buys and sells and scours the market is a necessary part of the time in which he functions as a capitalist, i.e., as personified capital. It is a part of his business hours. (II, 6)

The tasks considered are variegated, from the overseeing of labour in the workshop to the acceleration of the circuit of capital (as in the market activities mentioned above). All these tasks are unproductive, though they are useful. Their purpose is the *maximizing of the profit rate* of the capitalist. (The profit rate is defined below in the treatment of competition.)

The capitalist delegates some of these unproductive tasks to employees. They require means of production as well as labour power, like industrial capitalist production, though they produce no value. The wage and capital costs of these unproductive activities are a deduction from the surplus value. Marx denotes them as ‘costs’, in particular *costs of circulation* (the control and acceleration of the circuit of capital). As a consequence Marx categorizes some wage labour employed in capitalist production as

unproductive, as in, for example, the case of overseers and employees in trade.

The distribution of surplus value as income

In Volume III, surplus value in its relation to both self-expansion and circulation is renamed *profit*. Profit is a form of surplus value. Once extracted, surplus value is at the origin of various categories of incomes, which appear as deductions from profit. The payment of such incomes to agents who employ no labour is thus consistent with the labour theories of value and surplus value. These channels of distribution of surplus value correspond to specific fractions of ruling classes in capitalism, such as active capitalists (entrepreneurs), money capitalists and landowners.

Interest and profit of enterprise. interest-bearing capital (III, 21 – 3)

Some capitalists do not engage directly in capitalist production, but put their capital at the disposal of another functioning industrial capitalist, the *active capitalist* (or *entrepreneur*). This transaction may take the form of a loan in exchange for a share of the surplus value as *interest*, or the purchase of shares of stock in the firm which pays *dividends*. Marx treats both cases as *interest-bearing capital*, and this category of capitalists as *money capitalists* (sometimes referred to as ‘financial capitalists’). Marx explains interest as a portion of the surplus-value realized by active capitalists. The profit remaining after the active capitalist has paid dividends and interest is *profit of enterprise*. The existence of a developed loan market with a uniform rate of interest (for each maturity and risk of the loan) leads active capitalists to regard their own capital as loan capital, and to impute interest on it as an opportunity cost. Thus profit of enterprise appears as a kind of wage to the entrepreneurial activities of the active capitalist.

Rent (III, 38, 45)

Owners of scarce natural resources (‘land’ in the terminology of the classical political economists) also receive incomes in deduction from profits, in the form of rents. Due to their monopoly ownership of specific pieces of land, landowners can bargain with individual capitalists for a share of the surplus-value as rent (or royalties in other instances). How rents are quantitatively determined can only be examined in relation of the theory of competition.

Finance

Banking capital and money capitalists (II, 19; III, 29)

The tasks of money-dealing capital are performed by banks. This represents their first source of income. Banks also concentrate and use available masses of capital. One source of funds for banks is the idle balances of money in the economy, which are deposited in bank accounts. Thus, the money capital of enterprises is pooled within banks together with the balances of money held by other agents, such as households. While individual balances fluctuate, the aggregate pools are much more stable. A second

source of funds is the capital of money capitalists (interest-bearing capital, including stock shares), who, instead of dealing directly with entrepreneurs, use banks as intermediaries. (Marx is aware of the capability of banks to ‘create’ money, but his view of banking mechanisms remains dominated by intermediation.) The theory of banking capital unites these two facets of the theory of capital: money-dealing capital and the handling of the capital of money capitalists.
Besides the management of accounts, the main function of banks is to make these funds available to agents seeking financing. Banks actually become the ‘administrators’ of the capital of money capitalists, and ‘confront’ capital as used by enterprises:

> Borrowing and lending money becomes their [banks’] particular business. They act as middlemen between the actual lender and the borrower of money capital. Generally speaking, this aspect of the banking business consists of concentrating large amounts of the loanable money capital in the bankers’ hands, so that, in place of the individual money-lender, the bankers confront the industrial capitalists and commercial capitalists as representatives of all money-lenders. They become the general managers of money capital. On the other hand by borrowing for the entire world of commerce, they concentrate all the borrowers vis-à-vis all the lenders. A bank represents a centralisation of money capital of the lenders, on the one hand, and, on the other, a centralisation of the borrowers. (III, 25)

It is in these pages of Volume III of *Capital* that Marx analyses the issuance of paper currency by private banks and the Bank of England.

Fictitious capital and financial instability (III, 25)

Marx's original definition of capital, as value in a movement of self-expansion, does not apply to securities like Treasury bills, or even to the stock shares of corporations. To refer to these securities, Marx uses the phrase *fictitious capital*. A public bond is in no way ‘fictitious’ for its holder, but it has no counterpart in the *M*, *C* and *P* of the circuit of capital. Once bonds or equities have been sold by a capitalist firm and are being traded on a secondary market, their values are also fictitious. The emergence of a market interest rate leads to the phenomenon of the *capitalization* of income flows such as the interest on government debt and dividends on equity: the market, where expectations concerning the future of these flows are taken into account, assigns a principal value to any flow of income. Thus, the accumulation of capital is paralleled in capitalism by that of such fictitious capital. Marx sees this capitalization of revenue flows as a source of instability.

The institutional framework of modern capitalism (III, 21 – 3)

As noted earlier, with the development of capitalism, the functions of the active capitalist are gradually delegated to managers and employees. This configuration, in which funding is provided by money capitalists with banks acting as intermediary, and the bulk of capitalist functions is delegated to a salaried personnel is that of modern capitalism:

> But since, on the one hand, the mere owner of capital, the money capitalist, has to face the functioning capitalist, while money capital itself assumes a social character with the advance of credit, being concentrated in banks and loaned out by them instead of its original owners, and since, on the other hand, the mere manager who has no title whatever to the capital, whether through borrowing it or otherwise, performs all the real functions pertaining to the functioning capitalist as such, only the functionary remains and the capitalist disappears as superfluous from the production process. (III, 23)

The trinity formula of capital and classes in capitalism (III, 48; III, 52)

A major objective of *Capital* is to establish surplus value as the source of all incomes in capitalist society except wages. But capitalist practice hides this origin of capitalist incomes in what Marx calls the ‘trinity formula’:

> *Capital—profit (profit of enterprise plus interest), land—ground-rent, labour—wages*, this is the trinity formula which comprises all the secrets of the social production process. (III, 48)

Actually, this configuration is again altered in what we called above the institutions of modern capitalism:

> Furthermore, since as previously demonstrated interest appears as the specific characteristic product of capital and profit of enterprise on the contrary appears as wages independent of capital, the above trinity formula reduces itself more specifically to the following: *Capital—interest, land—ground-rent, labour—wages*, where profit, the specific characteristic form of surplus-value belonging to the capitalist mode of production, is fortunately eliminated. (III, 48)

To Marx, this trinity formula is ‘irrational’, because it confuses the source of incomes in the distribution of surplus-value with the role of necessary inputs in the production of use-values.
Volume III of *Capital* stops on a single-page chapter (obviously incomplete), entitled ‘Classes’. There Marx establishes a straightforward relationship between his analysis of incomes and the fundamental class pattern of capitalism:

> The owners merely of labour-power, owners of capital, and land-owners, whose respective sources of income are wages, profit and ground-rent, in other words, wage-labourers, capitalists and land-owners, constitute the three big classes of modern society based upon the capitalist mode of production. (III, 52)

To this one could add fractions of capitalist classes corresponding to the various circuits of capital and the division of surplus value as above: (1) industrial capitalists, commercial capitalists, bankers, and (2) entrepreneurs (active capitalists) and money capitalists.

The distribution of surplus value through competition

The analysis of capitalist production we have summarized so far, based on the idea that surplus value (and hence capitalist profit) arises from the exploitation of productive labour, runs counter to the apparent linkage of profit to the value of capital invested, regardless of the amount of labour it employs, or indeed whether or not that labour produces commodities at all. Marx offers a systematic account of the way in which competition among capitals gives rise to this linkage of profit with total capital invested by redistributing the surplus value created by productive labour.

Prices and the collective character of exploitation (III, 9)

Because prices are not necessarily proportional to values, surplus value is not necessarily realized by the capitalists who hired the labour-power that created it. Exploitation is thus a ‘collective’ mechanism for the capitalist class. It is as if surplus labour was collected in a single pool, and then distributed among capitalists in proportion to their invested capital (though the division of the surplus value among the individual capitals is actually the result of a fierce competitive struggle):

> Thus, although in selling their commodities the capitalists of the various spheres of production recover the value of the capital consumed in their production, they do not secure the surplus-value, and consequently the profit, created in their own sphere by the production of these commodities. What they secure is only as much surplus-value, and hence profit, as falls, when uniformly distributed, to the share of every aliquot part of the total social capital from the total social surplus-value, or profit, produced in a given time by the social capital in all spheres of production. … So far as profits are concerned, the various capitalists are just so many [100] stockholders in a stock company in which the shares of profit are uniformly divided per 100, so that profits differ in the case of the individual capitalists only in accordance with the amount of capital invested by each in the aggregate enterprise, i.e., according to his investment in social production as a whole, according to the number of his shares. (III, 9)

It is, consequently, necessary to distinguish between the mechanisms which govern the overall *appropriation* of surplus-value and its *realization* by particular capitalists:

1. 1. The total surplus value depends on the value of labour power and the total number of workers capitalists employ.
2. 2. Any system of commodity prices ‘distributes’ this total surplus value to individual producers (and landowners).

Marx describes this process of redistribution of surplus value as a ‘metabolism’ of value. Note that prices remain ‘forms of value’, as stated in the analysis of money and prices, but the hours of social abstract labour are reshuffled. At issue is no longer the labour actually expended to produce each

commodity individually, but value as socially ‘distributed’ by prices (purchasing power as a fraction of social value ‘conveyed’ by the price of each commodity).

The transformation problem (III, 9)

At the beginning of Volume III, Marx pursues two objectives simultaneously. On the one hand, he analyses the basic mechanisms of competition in capitalism, in which the determination of a particular set of prices is implied, with equalized profit rates among industries, and, on the other hand, he uses this particular case to discuss the metabolism of value introduced above. This exposition obscures the fact that the underlying mechanism of exploitation operates whatever the prevailing system of prices; the theory of exploitation does not depend on the particular properties of commodity prices and, in particular, not on the attainment of a market equilibrium at which profit rates are equalized. The failure to separate the two projects, and to appreciate the restricted context of the discussion of the metabolism of value in this particular case, has created much confusion in the history of Marxist economic theory. In the later literature the two problems, those of the metabolism of value and the prevalence of a particular set of prices in capitalist competition, are usually treated jointly as the *transformation problem*. Because of its importance in the history of Marxism, a specific entry is devoted to this controversial issue (see Marxian transformation problem).

The classical Marxian long-period equilibrium. prices of production (III, 10)

The analysis of this process of redistribution of surplus value through competition marks an important break in the present account of Marx's analysis in *Capital*. Beginning with the definition of capital (and the corresponding requirement of the analysis of commodity and money, actually a preliminary to the exposition of capital), we first followed Marx in his investigation of the two components of the theory of capital: the extraction of surplus value and the circuit of capital. These two aspects were then combined in analyses such as the reproduction schemes or capitalist functions. Finally, attention turned to the division of surplus value: (1) its distribution as interest and dividends to money capitalists, and as rents to landowners; (2) its realization by various categories of capitals, such as commercial capital and banking capital, in which no surplus value is produced; and (3), in the present section, its reallocation to capitalists of various industries independently of the extraction by individual capitalists, as in competition. We now enter a new category of developments, in which dynamic processes are involved: the mechanisms of competition, accumulation and employment, technical and distributional changes, and crises and the business cycle.

The basic idea in the analysis of capitalist competition is straightforward. If capital is free to move from one line of production to another in search of profit, the competitive movement of capitals will tend to move prices of specific commodities up or down until the rate of profit is equalized in all sectors. The *equalization of the rate of profit*, clearly stated by Adam Smith and David Ricardo, represents competition at the most fundamental level of analysis. The appropriation and realization of surplus value, as stated above, is thus specified quantitatively: one industry where little labour is used proportionally to total capital, in comparison to another industry, realizes more surplus value as profit than its workers actually contribute to the total surplus value (and conversely).

The *profit rate* is central in this analysis of competition. The profit rate is defined as the ratio of profit, *s*,

to total capital, $K=M+C+P$, that is $r=s/K$. The ratio of the value of the average total capital invested during one unit of time (for example, a year) to the flow of value corresponding to the cost of production engaged during this unit of time, $T=K/(c+v)$, is the *turnover time* of capital measured in units of time such as months or years. In the Marxist literature, the turnover time is often implicitly or explicitly assumed to be unity, in which case the profit rate $r=s/K$ is equal to the *profit margin*, the ratio of profit to costs of production, $s/(c+v)$.

The movement of capital in the pursuit of profit results in a tendency toward the equalization of profit rates among industries. Marx calls commodity prices which are consistent with an equalized profit rate *prices of production*:

> But capital withdraws from a sphere with a low rate of profit and invades others, which yield a higher profit. Through this incessant outflow and influx, or, briefly, through its distribution among the various spheres, which depends on how the rate of profit falls here and rises there, it creates such a ratio of supply to demand that the average profit in the various spheres of production becomes the same, and values are, therefore, converted into prices of production. (III, 10)

Actual *market prices* tend to gravitate around prices of production, and this property defines the *capitalist law of exchange* (which supersedes the commodity law of exchange when production is organized by capital). As stated earlier, Marx calls the substitution of one law of exchange for the other a ‘transformation’, the *transformation of values* (actually prices proportional to individual values) *into prices of production*.

The profit of commercial and money-dealing capital (III, 16; III, 19)

Although commercial and money-dealing capitals do not contribute to the extraction of surplus value, they do participate in its realization, along the lines indicated above, like any other capital. Commercial capital, for example, must secure a profit by buying commodities from industrial capitalists at prices below the prices at which those commodities will be sold to final purchasers. In this way commercial capital appropriates part of the surplus value actually created in the circuit of industrial capital. Similarly, the fees charged by money-dealing capital transfer surplus value created in the circuit of other capitals (abstracting from interest paid by other agents such as households or the state). Thus, the profit of commercial and money-dealing capital is part of the surplus value produced by labour employed by industrial capital.

Differential and absolute rent (III, 38; III, 45)

The level at which rents can be established is directly related to the level of the average and tendentially uniform profit rate in the overall economy. The condition for the cultivation of a land of lesser fertility or for a more intensive investment is that the marginal investment must yield the average profit rate. All capitalists (including capitalist farmers) expect to realize the average profit rate prevailing throughout the economy. This condition is assured if landowners bargain for rents just high enough to assure

capitalists the average rate of profit on their land. This defines *differential rent*. Marx also assumes that landowners as a class may withhold their lands until a minimum rent is paid, which defines *absolute rent*.

The centralization and concentration of capital, monopoly (I, 25)

The Classical–Marxian analysis, which assumes equalized profit rates among industries (not firms, because of differences in their productive efficiency), does not seem to match the features of competition in modern capitalism. Followers of Marx, from Hilferding and Lenin in the early 20th century to contemporary Marxist economics, point to the historical transformation of competition through the emergence of monopolies and oligopolies. The notion of the interplay of large firms is already part of Marx's analysis. In the process of accumulation the size of individual capitalist firms is altered by the *concentration* and *centralization* of capital. In Marx's account, concentration refers to the rise of the size of firms which parallels accumulation, while centralization denotes the outcome of merger or acquisition (and the process of competitive elimination of smaller and less efficient firms in an industry). Monopoly capital is not, however, part of Marx's analysis of capitalism, and Marx does not question the classical analysis of competition on such grounds. Rather than the view that the size of firms could hamper the process of equalization of profit rates among industries, Marx repeatedly asserts that credit mechanisms, including banks, are a crucial factor in the ability of capital to migrate among industries and, therefore, in the formation of prices of production.

Accumulation, and technological and distributional change

The *accumulation of capital* refers to the situation where a fraction of surplus value is saved and devoted to increasing the value of capital. While the analysis of expanded reproduction considers a steady growth path of the economy (on which the key ratios, the rate of surplus value, the organic composition of capital, the value of labour power, and the composition of demand, are assumed to remain constant), Marx's theory of accumulation incorporates the qualitative change in capitalist production that actually accompanies its expansion.

Capital accumulation and employment (I, 25)

For accumulation to succeed, a number of conditions must be met. In particular, an expanded supply of labour power must be made available to permit the expansion of production, an issue which Marx addresses at the end of Volume I. Marx rejects the conclusions of classical economists such as Thomas Malthus, who proposed universal laws governing population growth and a ‘natural’ path of accumulation of capital, and blamed low wages on the fecundity of workers and the limits of natural resources. Marx argues that each mode of production evolves its own characteristic laws of population, and that capitalism in particular gives rise to a number of mechanisms that ensure a rough proportionality between population growth and the accumulation of capital.

How much labour is necessary to meet the demands of capital accumulation? How is the supply of labour roughly adapted to accumulation? Marx explains, in his *law of capitalist accumulation*, that the amount of labour required depends on (1) the pace of accumulation and (2) technical change as

manifested in the variation of the composition of capital – that is, the ratio of capital outlays on means of production (*constant capital*) to capital outlays on wages (*variable capital*). If accumulation is rapid, and the composition of capital unaltered, the demand for labour power grows in proportion to accumulation and real wages tend to increase. This is the most favourable situation for workers. Technical change may moderate this tendency through an increase in the composition of capital, as the same accumulation requires less additional labour, and the demand for labour power grows more slowly than capital as a whole. A priori, any relation between the pace of accumulation and the change in the composition of capital may occur. Marx points, however, to the fact that the composition of capital tends historically to rise and, thus, the pressure on employment is regularly relaxed.

Two mechanisms contribute to remedy any potential lack of available labour power. First, technical change leading to increases in the composition of capital makes some employed labour redundant. Second, recurrent crises periodically restore what Marx calls the *floating reserve army of labour*, with the decline of output. Thus, the process of accumulation is uneven. Accumulation first proceeds during phases of more or less balanced growth; gradually the reserve army of unemployed workers is reabsorbed and wages rises. This is an inducement towards technical change increasing the composition of capital. If, however, the demand for labour grows too rapidly, a crisis occurs, the demand for labour is relaxed. Finally, a new wave of accumulation resumes after the crisis, during which a fraction of capital is devalued or destroyed. We will return below to these episodes in which a rise of wages provokes crises, which Marx calls situations of ‘over-accumulation’.

In addition to this recurring fluctuation of unemployment, capitalism historically has drawn workers from the *latent reserve army*, through the destruction of traditional agricultural modes of production, and the consequent migration of displaced workers to the capitalist labour market. The potential competition of the latent reserve army puts a long-term downward pressure on wages as well.

The overall interaction of these factors is complex, because technical change and the income distribution cannot be treated as independent mechanisms. Marx considers that rising wages, and a correspondingly diminished rate of surplus value, increase the incentives for capitalists to seek labour-saving technical changes. This leads to a rise in the composition of capital, as more machinery is employed, precisely in order to avoid increased wage costs. This analysis must be supplemented by the consideration of fundamental political conditions, in particular, the strength of workers’ class struggle, since Marx believed that, over and above the mechanisms involved in the law of capitalist accumulation, organized labour struggles could influence both wages and the length of the working day.

One of Marx's main goals in presenting his theory of accumulation, at the end of Volume I of *Capital*, is to show that the scarcity of labour power is not an absolute barrier to capital accumulation. The main thesis there is that, in the race between capital accumulation and the supply of labour power that governs the evolution of real wages, employment, and the rate of surplus value, capital has the edge over labour as a result of the capability of capital to substitute fixed capital (machinery) for labour:

> The same causes which develop the expansive power of capital, develop also the labour-power at its disposal. The relative mass of the industrial reserve army increases therefore with the potential energy of wealth. But the greater this reserve army in proportion to the active labour-army, the greater is the mass of a consolidated surplus-population, whose misery is in inverse ratio to its torment of labour. The more extensive, finally, the Lazarus-layers of the working-class, and the industrial reserve army, the greater is official

> pauperism. *This is the absolute general law of capitalist accumulation*. Like all other laws it is modified in its working by many circumstances, the analysis of which does not concern us here. (I, 25-4)

Besides the resistance of organized workers, this capability of capitalism to perpetuate an available reserve army by technical change is limited by the cost of the addition of capital which is required to displace labour, as Marx will contend in his analysis of technical change and the tendency for the profit rate to fall.

Technical change (III, 13 – 15)

The social and technical conditions of production and their historical transformation are central to Marx's analysis of capitalist production. The term ‘technology’ is convenient but somewhat misleading. Marx always describes conditions of production in a perspective which combines technology in the strict sense and organization, that is, the institutional framework in which production is performed; the notion of social relations cannot be neglected in this context. This is the case, for example, in the analysis of relative surplus value, as discussed earlier in reference to manufacture and large-scale industry.
Although Marx often discussed specific historical determinants of technical innovations, his main theory of technical change in capitalism sees it as an endogenous response to pressures from competitors and workers. Each capitalist has a strong motivation to find cost-reducing technical innovations (or profit-increasing product innovations) because the firm which first successfully exploits such innovations is in a position to capture higher-than-average profit rates (‘super-profits’) as a result of its temporary monopoly on the innovation. Innovating capitalists may also use their cost advantage aggressively to increase their market share. (In this respect Marx develops the theory of technical change Ricardo, 1817, presents in his chapter on machinery.) Over time, competitors will find equivalent innovations and the advantage of the innovating capitalist will erode.
Capitalist technical innovation in Marx's framework begins with the discovery of a range of potential new productive techniques and forms of labour organization. The accumulated store of technical knowledge available to capitalist society at any moment is the result of this historical process of innovation: there is no set of predetermined techniques as is assumed in the neoclassical production function. Marx's theory of induced technical change is basically evolutionary. The capitalist evaluates the cost of these alternatives at prevailing prices and wages, and forms expectations concerning profit rates. Only those technologies that promise to reduce costs or increase profits at prevailing prices and wages are *viable* candidates for adoption. The criterion is an increased profit rate.
Marx emphasizes that, because capitalism places both strong incentives for technical change and the power to implement in the hands of competing capitalist firms, it is a *technically progressive* mode of production, in contrast to slavery and feudalism. In this respect Marx resembles Smith, who emphasizes increasing returns inherent in the division of labour, rather than Ricardo, who emphasizes diminishing returns due to limited natural resources (land).

The tendency for the profit rate to fall (III, 13 – 15)

In Volume III of *Capital*, Marx describes trajectories of technical and distributional changes that he denotes as *historical tendencies*. They are unbalanced (nonhomothetic) growth trajectories, which Marx considered typical of the dynamics of capitalism, which we will describe as *trajectories à la Marx*. Along such very long-term paths, the growth rates of capital, output, and employment gradually fall, labour productivity and the composition of capital rise, the share of wages in total income is constant or diminishing, and the profit rate declines. In the speaking of historical tendencies, ‘historical’ refers to a very long-term time frame; ‘tendency’ means that though accumulation in capitalism tends to follow such trajectories, the trajectory does not necessarily prevail due to the action of what Marx labels *counteracting factors*. It is in this framework that Marx defines the *tendential fall in the rate of profit*. This ‘law’ expresses sophisticated insights into the historical dynamics of capitalist economic growth. It is one of the major disputed issues in contemporary Marxist economics (along with the transformation problem).

In Volume III, the profit rate is written as a ratio of two flows or, equivalently, the turnover time of capital is assumed to be unity: $r=s/K=s/(c+v)$. Dividing by v, Marx obtains: $r=(s/v)/(c/v+1)$. The numerator is the rate of surplus value, and the denominator is the *value composition of capital*, *the ratio of constant to variable capital*, plus 1. Marx calls this value composition the *organic composition* of capital. In this simple presentation, the conflicting impacts of the rate of exploitation and the organic composition of capital are clearly evident.

Although *labour productivity* does not appear in this formal setting, it is explicitly a key variable in Marx's analysis. Without altering the basic framework, it is possible to write: $r=(s/(v+s))/((c+v)/(v+s))$. Here, $s/(v+s)$ is the share of profit in total income, and $(c+v)/(v+s)$ is total capital per hour worked, which is another measure of the organic composition of capital. (This ratio can also be read as the ratio of capital to output, since output is equal to total income, or equivalently the inverse of what is frequently loosely called ‘capital productivity’.) The numerator, the share of profit, can be written $1-(v/(v+s))$, that is, 1•minus the share of wages. The share of wages is equal to real wages divided by labour productivity. Thus, the profit rate can be expressed as the ratio of the profit share to the total capital per hour worked, which we call simply the *composition of capital*:

$$\text{profit rate} = \frac{1 - \dfrac{\text{real wage}}{\text{labor productivity}}}{\text{composition of capital}}$$

Marx's fundamental insight can be sketched as follows. To maintain or increase profits (which appear in the numerator of the profit rate), when there is no fall in the real wage, capitalists must increase the productivity of labour, which is the mechanism of relative surplus value. Marx contends, however, that this increase has a considerable cost for capitalists because increases in labour productivity typically require the investment of more capital per hour worked: productivity gains are realized by way of an increased mechanization of production. Thus, the composition of capital rises, and the rate of profit may fall. The actual evolution of the rate of profit also depends on what happens to the real wage and, consequently, to the rate of surplus value as labour productivity increases, which depends on labour

market factors and class struggle, which are beyond the control of any individual capitalist.

Marx considers the case where the rate of surplus value remains constant to refute the argument that the falling profit rate is the result of an excessive growth in the cost of labour to the capitalists. When the productivity of labour rises, a constant rate of surplus value implies a rising real wage. Thus in making this argument, Marx does not assume a constant real wage. His thesis is rather that it is difficult for capitalists to counteract rising wages by technical change, since a more efficient technique in terms of labour productivity typically requires a rising composition of capital. The linchpin of Marx's thesis is, therefore, a hypothesis on the features of available techniques, that is, the profile of innovation: it is comparatively easy to find labour-saving devices if the cost of mechanization is not considered, but opportunities to reduce labour costs without inflating capital costs are rare.

Thus, on trajectories à la Marx the productivity of labour rises, while the productivity of capital (the inverse of the composition of capital) falls, a pattern of technical change sometimes called *Marx-biased*:

> The law of the falling rate of profit, which expresses the same, or even a higher, rate of surplus-value, states, in other words, that any quantity of the average social capital, say, a capital of 100, comprises an ever larger portion of means of production, and an ever smaller portion of living labour. Therefore, since the aggregate mass of living labour operating the means of production decreases in relation to the value of these means of production, it follows that the unpaid labour and the portion of value in which it is expressed must decline as compared to the value of the advanced total capital. … The relative decrease of the variable and increase of the constant capital, however much both parts may grow in absolute magnitude, is, as we have said, but another expression for greater productivity of labour. (III, 13)

Though Marx never articulated the entire framework, this analysis of the biased pattern of technical change supplements the mechanisms at work in the law of capitalist accumulation. Accumulation recurrently pushes employment to the limits of the supply of labour power available and drives real wages upward. Technical change and recurrent crises allow for the partial relaxation of this pressure (as we have seen), but, in typical periods, the new techniques available are such that technical change can only partially offset the rise in real wages, and the profit rate falls. Accumulation is pursued in spite of the diminished profit rate, which will only be apparent after the fact, when a major crisis occurs.

The analysis Engels published from Marx's notes in Volume III of *Capital* is incomplete, and was not intended for publication in the form in which we read it. Consequently, it is not too surprising that Marx's analysis of the tendency for the profit rate to fall remains controversial among Marxists. A central issue is the assumption made concerning real wages, and its relationship to the profitability criterion in the adoption of new techniques. Marx is clear that the innovating capitalist initially makes a surplus profit, while his competitors gradually adopt the new technique and prices fall through competition towards the prices of production corresponding to the new technology. Marx contends that the new uniform average profit rate tends to be lower than the original one. Nobuo Okishio (1972) has demonstrated that if the real wage remains unchanged during this process the new average profit rate can never fall. But along a trajectory à la Marx real wages do increase, as we have explained, although the possibility of a tendency for the rate of profit to fall is consistent with Marx's assumption that the rate of surplus value is constant or even rising.

The problem of the evolution of real wages, the value of labour power, and the rate of surplus value over time as labour productivity rises is controversial among Marxists, due to a change of Marx's view on this subject during his lifetime. Engels explained that Marx originally accepted the so-called *iron law of wages*, which assumes that real wages are constantly driven downward to a minimum compatible with the reproduction of the labour force, but later abandoned it. Marx sometimes refers to a ‘socially and historically determined’ cost of reproduction of labour power, as an external constraint on the evolution of the real wage. But this ‘exogenous’ variable is explicitly subject to a number of economic and social determinations: (i) class struggle impacts on wages and the duration of labour; and (ii) the outcome of struggles crucially depends on the conditions of accumulation and the population available to work (as in the law of accumulation). Marx's understanding of the determination of wages is similar to his view of technical change: the path of real wages is the result of the interaction of extra-economic factors with economic mechanisms such as accumulation and crises.

Crises and the business cycle (III, 15)

There is no systematic treatment of crises and of the business cycle in Marx's work, although the issue plays a prominent role in his analysis of capitalism. In early works, like the *Communist Manifesto*, even prior to Marx's serious study of political economy, the idea that crises will prove more violent with the evolution of capitalism is central. Recurrent crises became a feature of capitalism during the first half of the 19th century. This link between economic mechanisms and class struggle had a considerable impact on Marx's view of the historical dynamics of capitalism. Then, Marx became gradually better aware of the complexity of the phenomenon of crises, in particular the relationship between real and financial mechanisms and crises.

Partial crises and crises of general overproduction

Before capitalism, poor crops and the devastation of war and disease were the major causes of disruptions of production. David Ricardo (1817) observed the existence of recurring crises more directly related to the nature of capitalism, which he called *states of distress*. These crises struck specific industries, like textiles. Consequently, Ricardo interpreted these situations as the effect of *disproportions*, that is, the outcome of the excessive accumulation of capital in one industry. Ricardo did not believe in the possibility of a general glut of the market. Marx devoted much energy to the refutation of Ricardo's interpretation. He contended that the existence of a delay between the sale of a commodity and the spending of its money price on another commodity invalidates ‘Say's Law’, the principle that the sale of a commodity constitutes a direct demand for another commodity. Monetary exchange thus implies the *possibility* of crises, because, by functioning as intermediary in exchanges, money allows for the interruption of the chain of exchanges. Only the ‘possibility’ of crises is, however, implied, not their actual mechanisms in capitalism.
Marx identified a new category of crises, *crises of general overproduction*, where all industries were simultaneously affected. Marx did not deny the existence of crises specific to particular industries, that he called *partial crises*, but contrasted the two types of situations, partial and general, and was specifically concerned with the latter.

The ultimate ground of crisis. profitability and social needs

Marx described general crises of overproduction as typical of capitalism. In capitalism, the purpose of production is not the satisfaction of the needs of the population, but the appropriation of profits. The ‘ultimate ground’ of crisis in capitalism is this disconnection between production and social needs:

> The ultimate reason [*ground*] for all real crises always remains the poverty and restricted consumption of the masses as opposed to the drive of capitalist production to develop the productive forces as though only the absolute consuming power of society constituted their limit. (III, 30)

This quotation is often misunderstood. Marx did not believe that higher wages would solve the problem of crises in capitalism. The cause of crises, proper to capitalism, is the recurrent inability to pursue production *at a certain rate of profit*. Therefore, profitability is always the crucial variable in Marx's explanation of crises:

> Over-production of capital is never anything more than overproduction of means of production – of means of labour and necessities of life – which may serve as capital, *i.e.*, may serve to exploit labour at a given degree of exploitation; … too many means of labour and necessities of life are produced at times to permit of their serving as means for the exploitation of labourers at a certain rate of profit. (III, 15-3)

The business cycle and its determinants

Marx described the fluctuating pattern of production in capitalism as ‘the cycles in which modern industry moves – state of inactivity, mounting revival, prosperity, over-production, crisis, stagnation, state of inactivity, etc.’ (III, 22).

Production is recurrently destabilized by mechanisms which affect the profitability of capital in the short run (a sudden decline rather than a steady downward trend). The first mechanism is *over-accumulation*. Periodically, employment gets closer to the limits of the population available to work (the reserve army is reabsorbed, as in the law of capitalist accumulation). Wages tend to rise, and profitability is diminished. A second mechanism is the rise of interest rates. During the phase of rapid accumulation, the mass of credits increases and, at a certain point, interest rates rise. Again, profitability is affected and the economy destabilized. Marx is well aware of the relationship between real and financial mechanism, and he interprets the direction of causation as reciprocal.

As stated above, Marx did not explain crises by the deficient level of wages (except in his very early work), and refuted this explanation in the manuscripts of Volume II:

> It is sheer tautology to say that crises are caused by the scarcity of effective consumption, or of effective consumers. The capitalist system does not know any other modes of consumption than effective ones, except that of *sub forma pauperis* or of the swindler. That commodities are unsalable means only that no effective purchasers have been found

> for them, i.e., consumers (since commodities are bought in the final analysis for productive or individual consumption). But if one were to attempt to give this tautology the semblance of a profounder justification by saying that the working-class receives too small a portion of its own product and the evil would be remedied as soon as it receives a larger share of it and its wages increase in consequence, one could only remark that crises are always prepared by precisely a period in which wages rise generally [over-accumulation] and the working-class actually gets a larger share of that part of the annual product which is intended for consumption. From the point of view of these advocates of sound and 'simple' (!) common sense, such a period should rather remove the crisis. (II, 20)

Structural crises and the falling profit rate

Since the profitability of capital is central in Marx analysis of crises, there is a link between the tendency for the profit rate to fall and crises. Marx's view is that actual phases of decline of the profit rate make crises more likely, more frequent and deeper. He points to the existence of periods of sustained instability, which, although Marx does not use the term, can be called *structural crises*. A declining and depressed profit rate (both the tendency and levels are at issue) disturbs capitalist accumulation:

> …in view of the fact that the rate at which the total capital is valorised, i.e. the rate of profit, is the spur to capitalist production …, a fall in this rate slows down the formation of new, independent capitals and thus appears as a threat to the development of the capitalist production process; it promotes overproduction, speculation and crises, and leads to the existence of excess capital alongside a surplus population. (III, 15)

This insight concerning the link between the profit rate and the occurrence of periods of historical perturbation in the course of accumulation provides a powerful framework for understanding the real history of capitalist economies.

See Also

- British classical economics
- capitalism
- class
- classical distribution theories
- classical growth model
- commodity fetishism
- commodity money
- exploitation
- labour theory of value
- labour's share of income

- Marx, Karl Heinrich
- Marxian transformation problem
- Marxian value analysis
- ‘political economy’
- profit and profit theory

Bibliography

Arthur, C.J. 2004. *New Dialectic and Marx's Capital*. Leiden, Boston: Brill.

Brewer, A. 1984. *Guide to Marx's Capital*. Cambridge: Cambridge University Press.

Cleaver, H. 2000. *Reading Capital Politically*. Edinburgh: AK Press.

Duménil, G. and Lévy, D. 1993. *Economics of the Profit Rate: Competition, Crises, and Historical Tendencies in Capitalism*. Aldershot: Edward Elgar.

Duménil, G. and Lévy, D. 2003. *Économie Marxiste du Capitalisme*. Paris: La Découverte.

Fine, B. 1975. *Marx's Capital*. Leiden, Boston: Brill.

Fine, B. and Saad-Filho, A. 2004. *Marx's Capital*. New York: Pluto.

Foley, D. 1986. *Understanding Capital: Marx's Economic Theory*. Cambridge, MA: Harvard University Press.

Itoh, M. and Bullock, P. 1987. *The Basic Theory of Capitalism: The Forms and Substance of the Capitalist Economy*. Lanham, MD: Rowman and Littlefield.

Marx, K. 1859. *A Contribution to the Critique of Political Economy*, ed. M. Dobb. New York: International Publishers, 1970.

Marx, K. 1976, 1978, 1981. *Capital, Volumes I, II, III*. New York: Random House.

Marx, K. and Engels, F. 1848. *The Communist Manifesto*. New York: Signet Classics, 1998.

Moseley, F. 1993. *Marx's Method in Capital: A Reexamination*. Leiden, Boston: Brill.

Okishio, N. 1972. On Marx's production prices. *Keizaigaku Kenkyu* 19, 38–63.

Pilling, G. 1980. *Marx's Capital*. London, New York: Routledge.

Ricardo, D. 1817. *On the Principles of Political Economy and Taxation*. Cambridge: Cambridge University Press, 1951.

Rosdolsky, R. 2005. *The Making of Marx's Capital*. London: Pluto Press.

Rubin, I. 1972. *Essays on Marx's Economic Theory*. Detroit: Black and Red.

Sekine, T. 1997. *Outline of the Dialectic of Capital*. London: Palgrave Macmillan.

Shaikh, A. and Tonak, E. 1996. *Measuring the Wealth of Nations: The Political Economy of National Accounts*. Cambridge: Cambridge University Press.

Sweezy, P. 1970. *Theory of Capitalist Development*. New York: Monthly Review Press.

Weeks, J. 1994. *Capital and Exploitation*. Princeton: Princeton University Press.

Wolff, R. and Cohen, M. 1985. *Understanding Marx: A Reconstructive and Critique of Capital*. Princeton: Princeton University Press.

How to cite this article

Foley, Duncan and Gérard Duménil. "Marx's analysis of capitalist production." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000399> doi:10.1057/9780230226203.1054

The New Palgrave Dictionary of Economics Online

Mason, Edward Sagendorph (1899 – 1992)

Gustav F. Papanek

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

development economics; firm, theory of; industrial economics; Mason, E. S.; monopolistic competition.

Article

Edward Mason had a significant impact on the economics profession in four disparate areas: through his influence on the Harvard Economics Department; through his role in two separate sub-fields of the discipline – industrial and development economics; and by exemplifying the dual role of academic and practitioner.

When he came to Harvard in 1919 as a graduate student, he was not in the Harvard mould: he was from the public schools of Kansas and its University. Later, another newcomer and ‘provincial’, J. Kenneth Galbraith, characterized young Mason's presence by saying ‘even when he was an Instructor, where Ed Mason sat there was the head of the table’. He remained a central figure in Harvard Economics for well over 50 years, his only absence a stint in Washington during the Second World War (1941–44). He was one of a handful of senior faculty who dominated the department during its glory days, when it produced about half of all economics Ph.D. degree holders in the United States and was responsible for a substantial fraction of the research produced in the country.

Mason's role at Harvard extended beyond economics. He served for 11 years (1947–58) as Dean of the School of Public Administration and for a short period was second-in-command of the university. While he was its Dean the School of Public Administration became the leading exponent of an emphasis on policy analysis, especially economic analysis, rather than administrative tools and institutions.

Mason taught many of the economists who expanded industrial economics from a preoccupation with regulation and monopoly to an analysis of markets and firms. The ‘Masonic Lodge’ of ex-students, together with its Grand Master, came to dominate the newly developed applied discipline of industrial organization. Mason stimulated the work by his proposition that the performance of a firm was largely explained by the structure of the market in which it operated. The controversy stimulated by this idea and by the concept of monopolistic competition to which he contributed, helped the subfield of industrial economics to flourish intellectually.

It was typical of Mason to come to research from policy concerns; he exemplified the practitioner–academician. Mason's advice was sought by governments and he was privy to their problems and the facts at their disposal. During the Second World War, he headed the economic staff of the OSS, probably the first US intelligence agency to gather economic intelligence systematically and to analyse possibilities for economic warfare. Later he was Deputy Assistant Secretary of State in charge of Economic Affairs and chief economic adviser to the US delegation at the 1947 Moscow Conference.
In the early 1950s, development economists began to apply many of the standard tools of economics to the special problems and institutions of the poor countries. Mason developed an interest in this subject in typical fashion by setting out to deal with a specific set of policy problems. After returning to Harvard at the end of the war, as Dean of Public Administration, he was asked to organize technical assistance to help Pakistan carry out the economic analysis crucial to rational government decision making. Out of that effort emerged another institution that rightfully could be seen as created by Mason; the Development Advisory Service, later the Harvard Institute of International Development. A surprising proportion of those who consider themselves development economists had their initial field experience as an adviser, or as their local counter part, in one of the teams fielded by an organization that existed only because Mason managed to persuade Harvard to undertake the non-traditional university function of advising foreign governments.

Selected works

1926. The doctrine of comparative cost. *Quarterly Journal of Economics* 41, November 63–93.

1932. *The Street Railway in Massachusetts*. Cambridge, MA: Harvard University Press.

1937. Monopoly in law and economics. *Yale Law Journal* 47(1), 34–49. Repr. in *Public Policy and the Modern Corporation: Selected Readings*, ed. D. Grunewald and H.L. Bass. New York: Meredith, 1966.

1938. Price inflexibility. *Review of Economics and Statistics* 20, May, 53–64.

1939. Price and production policies of larger-scale enterprise. *American Economic Review* 29, 61–74. Repr. in American Economic Association, *Readings in Industrial Organization and Public Policy*, ed. R. B. Hefelbower and G.W. Stocking. Homewood, IL: Richard D. Irwin, 1958.

1946. *Controlling World Trade Cartels and Commodity Agreements*. New York: McGraw-Hill.

1949a. The effectiveness of the federal anti-trust laws: a symposium. *American Economic Review* 39, 712–13. Repr. in *Monopoly Power and Economic Performance: The Problems of Industrial Concentration*, ed. E. Mansfield. New York: W.W. Norton, 1964. Also reprinted in *Problems of the Modern Economy*, ed. E.S. Phelps. New York: W.W. Norton, 1966.

1949b. The current status of the monopoly problem in the United States. *Harvard Law Review* 62, 1265–85.

1956. Market power and business conduct: some comments on the Report of the Attorney General's Committee on Antitrust Policy. *American Economic Review, Papers and Proceedings* 46, 471–81.

1957. *Economic Concentration and the Monopoly Problem*. Cambridge, MA: Harvard University Press.

1958. The apologetics of ‘Managerialism’. *Journal of Business* 31, January, 1–11. Repr. in *The Business System: Readings in Ideas and Concepts* (vols 1–3). *In Commemoration of the Fiftieth Anniversary, Graduate School of Business, Columbia University, 1966*, ed. C.C. Walton and R.S. Eells. New York: Arkville Press, 1967.

1959., ed. *The Corporation in Modern Society*. Cambridge. MA: Harvard University Press.

1952. Raw materials, rearmament, and economic development. *Quarterly Journal of Economics* 66, 327–41.

1960. The role of government in economic development. *American Economic Review, Papers and Proceedings* 50, 636–40. Repr. in *Studies in Economic Development*, ed. A.M. Okun and R.W. Richardson. New York: Holt, Rinehart & Winston, 1961.

1962. Some aspects of the strategy of development planning: centralization vs. decentralization. In *Organizations, Planning and Programming for Economic Development*, vol. 8 of Science, Technology and Development; US papers prepared for the U.N. Conference on the Application of Science and Technology for the Benefit of the Less Developed Area, Washington, DC: Government Printing Office.

1963. Interests, ideologies and the problem of stability and growth. *American Economic Review* 53, 1–18.

1964. *Foreign Aid and Foreign Policy*. New York: Harper and Row for the Council on Foreign Relations.

1966. *Economic Development in India and Pakistan*. Cambridge, MA: Center for International Affairs, Harvard University.

1967. Monopolistic competition and the growth process in less developed countries: Chamberlin and the Schumpeterian dimension. In *Monopolistic Competition Theory: Studies in Impact. Essays in Honor of Edward H. Chamberlin*, ed. R.E. Kuenne. New York: John Wiley.

1971. Controlling industry. In D.V. Brown et al., *The Economics of the Recovery Program*. New York: Da Capo.

1973. (With R.E. Asher.) *The World Bank since Bretton Woods*. Washington, DC: Brookings Institution.

1980. (With others.) *The Economic and Social Modernization of the Republic of Korea*. Cambridge, MA: Harvard University Press.

How to cite this article

Papanek, Gustav F. "Mason, Edward Sagendorph (1899–1992)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000101> doi:10.1057/9780230226203.1055

The New Palgrave Dictionary of Economics Online

Mass é , Pierre (1898 – 1987)

Marcel P. Boiteux

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bellman's Optimum Principle; centre and periphery; dynamic programming; electricity markets; forecasting; linear programming; Massé, P.; optimum control; planning

Article

Massé was born on 13 January 1898, the same day that Emile Zola in his ‘J'accuse’ revealed the truth about the Dreyfus Affair, which arouses so much passion in French political circles to this day. His family was quick to take sides – in defence of the innocent – and from them Massé quite probably inherited his deep humanism, in which realistic thought was allied with optimism in action.

In 1916, he passed the competitive entrance examinations to both the Ecole Normale Supérieure, in science, and the Ecole Polytechnique. He was a Second Lieutenant in the Artillery from 1917 to 1918, and then opted for the Polytechnique and a career as an engineer. This choice of career foreshadowed a life in which, in a happy marriage, thought and action were to mingle unceasingly. A further spell of training at the Ecole des Ponts et Chaussées, and the start of his career as a government servant, channelled him towards major civil engineering works, and then, quite soon, towards the business world and hydroelectrical improvement works.

This was a decisive turning-point for both the man of action, the builder, and for the thinking economist. Obliged to deal with the management problems raised by the water stocks accumulated in reservoirs, and also with the need to turn them to account, Massé identified the key role of reserves as the means of regulating systems in order to cope with random factors. In his first work, *Les réserves et la régulation de l'avenir*, published in 1946, whose findings had been published two years previously in a paper submitted to the Société Statistique de Paris, Pierre Massé can be seen to be a forerunner of dynamic programming and of the theory of optimum control. In particular, he set forth in this paper two rules for the optimal management of random processes: (*a*) reservoirs should be managed so as to equalize the marginal utility of the water releases and the marginal expected value of the water held in stock; and (*b*) in order to calculate that expected value a strategy for the future should be defined, that is, a sequence of conditional decisions combining at any time the impact of past decisions, the actual outcome of the

random processes, and the perception of what future natural conditions will probably be. Kenneth Arrow was later to note, in 1956, that this was the earliest formulation of Richard Bellman's Optimum Principle. Massé's work was deeply marked by the recognition that in a random world – and the more so with an uncertain future – one could not confine oneself to just a single forecast, and by the need to adopt strategies and regulate stocks. This was to be borne out 20 years later, when he was in charge of the French Commissariat au Plan (Planning Commission). The consistency of the forecasts carried out as part of the National Accounts exercise certainly went some way towards making the plan a ‘reducer of uncertainties’. Moreover, under his guidance this achievement was crowned by a forecasting approach in which the seeking of a consensus on the type of development that was desirable was combined with the concern to identify ‘factors with potential for the future’, and by the devising of ‘warning lights’, as instruments for marking the future course that were capable of setting corrective actions in motion.

He was Directeur de l'Equipement at Electricité de France in 1946 for the start of the Plan Monnet and became its Directeur Général Adjoint two-and-a-half years later, a post he held until 1959. In those 12 years he developed, and then applied, linear programming techniques for determining the overall volume of electricity generating plant, and furnished justification for using a national discount rate for setting off present and future income and expenditure against each other. He tirelessly argued with the government in favour of using these clear and rigorous tools, already finding support on the Commissariat Général du Plan. In 1957, he published *Le choix des investissements*, a work which was to become authoritative both in France and abroad.

In February 1959, General de Gaulle appointed Massé to head the Commissariat Général du Plan. He took up his duties backed by the sound experience of a microeconomist who was thoroughly at ease with the idea of maximizing the benefit to the community in managing a public service, and who was attached to the pricing system and to its role in providing guidance and regulation. He sought to make the Plan – which had been largely governed by the concern for consistency and accordingly gave pride of place to analysing interlocking strengths and weaknesses – a structure better directed towards achieving competitiveness, both domestically and on foreign markets. His aim was not only to produce more, but also to produce better quality, with consciousness of costs.

With these goals, he strove to lighten the Plan's structure and make it interlocking with, and not a substitute for, the market. Without losing the valuable contribution of a generalized market survey, backed up with the use of an input–output matrix, he endeavoured to better pinpoint future price and income trends: in this way, programming by volume was to be backed up by an early attempt at programming by value. While the market could show what present prices were, it said practically nothing about future prices, since forward markets covered only narrow sectors and near time-horizons. By the light it shed on the future, the Plan was seen to be an indispensable adjunct to a smoothly working market economy. The ‘Centre’ had the task of successfully conveying to the ‘Periphery’ the right price system, and on the basis of this information, the Periphery was able to return to the Centre information on what the intentions were of the decentralized economic agents concerning volumes of goods to be consumed or invested in, and the volumes of factors to be mobilized. In this way, consultation was established between the Centre and the Periphery, converging, after a few successive iterations, towards a dynamic equilibrium. Pierre Massé had already analysed this converging dialogue between the Centre and the Periphery as early as 1952, in ‘Pratique et philosophie de l'investissement’. The Commissariat au Plan in fact organized consultations among the major socio-occupational categories; experts could intervene to put figures to the impacts from selecting the options adopted,

while the representatives of the state were there to ensure observance of the major policy guidelines defined by the government in agreement with Parliament. Such at least was the theory. Certain departures from it were unavoidable in practice.

However, ‘at the same time as it was an act of faith’, the Plan continued to be ‘an affirmation of the will’. Concerned as he was for a ‘less incomplete view of man’ to be taken into consideration, Massé succeeded in convincing the most influential circles in his country that a better balance should be struck between private and collective consumption. Thus, a feature of the early 1960s was a new concern for developing communal infrastructures. At the same time, while investigating various development scenarios, he concluded that it was necessary to raise the discount rate (of profitability) – an indicator of the scarcity of capital for government investors – so that it actually corresponded to the marginal efficiency of capital. He also concerned himself with disseminating the practice of constant-price calculations, so that, while changes in relative prices were not ignored, the profitability of infrastructure projects was not made attributable to inflated profits.

Having stressed future values, Massé necessarily broadened the scope of studies to cover price and income trends, and unavoidably brought discussion round to the knotty point of social tensions. To clarify and persuade, he worked on surplus accounts, establishing a rigorous relationship between the overall productivity gains made from one year to the next, and the sum of benefits available for distribution to customers, suppliers, workers and investors. From this attempt there at least remains a learning approach which the Centre d'Etude des Revenus et des Coûts, set up in 1966 at his instigation, has been engaged in disseminating and extending.

After helping start the Fifth Plan, Massé returned to Electricité de France, of which he was chairman for three years, and secured from the political authorities a sounder channelling of the necessary efforts for investment in the generation of electricity by nuclear power. He thereupon resumed acquaintanceship with business economics, though he did not forsake reflections upon the problems of the national economy, which he was never to abandon thereafter.

In 1977, Massé was elected a member of the Institut de France, which for almost 200 years has gathered together the most eminent French personalities in the humanities, science, history, philosophy and art. He pursued research for the remainder of his life. His body of work attests to a lifelong endeavour to reconcile macro and microeconomists and to ensure the cross-fertilization of their ideas for the social good.

Selected works

1946. *Les résberves et la régulation de l'avenir dans la vie économique*. Paris: Hermann.

1952. Pratique et philosophie de l'investissement. *Economie appliquée* 5, 625–58.

1953. Les investissements électriques. *Revue de statistique appliquée* 1(3–4), 119–29.

1957. (With R. Gibrat.) Application of linear programming to investments in the electric power industry. *Management Science* 3(2), 149–66.

1959. Prévision et prospective. *Cahiers de prospective*. Paris.

1964. *Le choix des investissements*. 2nd edn. Paris: Dunod.

1965. *Le plan ou l'anti-hasard*. Paris: Gallimard.

1969. (With P. Bernard.) *Les dividendes du progrès*. Paris: Le Seuil.

1973. *La crise du développement*. Paris: Gallimard.

1976. Prédation et création. *Etudes*. Paris.

1982. Le chiffré et le vécu, 1960–1980. *Revue des sciences morales et politiques*. Paris.

1984. *Aléas et Progrès, entre Candide et Cassandre*. Paris: Economica.

How to cite this article

Boiteux, Marcel P. "Massé, Pierre (1898–1987)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000102> doi:10.1057/9780230226203.1056

The New Palgrave Dictionary of Economics Online

matching and market design

Muriel Niederle, Alvin E. Roth and Tayfun Sönmez
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Matching is the part of economics concerned with who transacts with whom, and how. Models of matching, starting with the Gale–Shapley deferred acceptance algorithm, have been particularly useful in studying labour markets and in helping design clearing houses to fix market failures. Studying how markets fail also gives us insight into how marketplaces work well. They need to provide a thick, uncongested market in which it is safe to participate. Clearing houses that do this have been designed for many entry-level professional labour markets, for the assignment of children to public schools, and for exchange of live-donor kidneys for transplantation.

Keywords

centralized matching; clearing houses; congestion; kidney exchange; labour markets; market design; marriage markets; matching; medical labour markets; National Resident Matching Program; school choice; one-sided and two-sided markets; priority algorithms; revelation of preferences; strategy-proof allocation mechanisms; two-sided markets

Article

‘Matching’ is the part of economics that focuses on the question of who gets what, particularly when the scarce goods to be allocated are heterogeneous and indivisible; for example, who works at which job, which students go to which school, who receives which transplantable organ, and so on. Studying how particular matching markets succeed at creating efficient matches, or fail to do so, has yielded insights into how markets in general work well or badly.

Because market failures have sometimes been successfully fixed by devising new rules for both centralized and decentralized market organization, matching has been a major focus of the emerging field of market design. Some designs by economists have included labour market clearing houses for doctors and other health-care workers in the United States, both for their first jobs and as they enter

specialties. Clearing houses have also been implemented in less traditional markets, which cannot adjust prices or wages to help clear the market, such as the matching of students to schools in New York City and in Boston. And new clearing houses are being implemented for the organization of live-donor kidney exchanges among patients in need of a kidney transplant who have willing donors with whom they are incompatible.

In the next section we review some studies of matching, including some market failures that have been addressed either by introducing appropriate rules to a decentralized market (as in admissions to graduate programmes in American universities), or by introducing a centralized clearing house (as in the markets for new doctors in the United States, Canada, and Britain). The subsequent two sections consider the simple theory behind some clearinghouse designs. Then we return to some of the successful market design applications, which build on the theoretical models, but handle practical problems that are sometimes not yet fully understood in theory.

We focus on three kinds of market failure that sometimes impede efficient matching.

1. 1. Failure to provide *thickness*; that is, to bring together enough buyers and sellers (or firms and workers, schools and students, and so forth) to transact with each other.
2. 2. Failure to overcome the *congestion* that thickness can bring, that is, that can result when lots of buyers and sellers are trying to transact. That is, failure to provide enough time, or failure to make transactions fast enough so that market participants can consider enough alternative possible transactions to arrive at satisfactory ones.
3. 3. Failure to make it *safe* for market participants to reliably reveal or otherwise act on their information.

Some market failures and their consequences

Unravelling, congestion and centralized clearing houses

A variety of professional labour markets have suffered from the *unravelling* of appointment dates: from year to year, appointments were made earlier and earlier in advance of actual employment. Markets that had once been thick, with many employers and applicants on the market at the same time, became thin, as potential employees faced early offers, dispersed in time, to which they had to respond before they could learn what other offers might be forthcoming. That is, applicants often received ‘exploding’ offers that had to be accepted or rejected without waiting to see whether a more desirable offer might be forthcoming. An applicant who accepts such an offer, in the case that acceptances are binding, will never learn of the more desirable offers that might have become available, but if the offer is reasonably desirable rejecting it might be very risky. And, when applicants are quickly accepting offers in this way, employers, when they make offers, have to start taking into account whether the offer is likely to be accepted, since by the time an offer is rejected other desirable applicants may have already accepted offers elsewhere. This often makes unravelling a dynamic process, with offers being made earlier and shorter in duration from year to year. This kind of unravelling has been described in detail in markets for lawyers (Avery et al., 2001), gastroenterologists (Niederle, Proctor and Roth, 2006) and many others (see Roth and Xing, 1994). A clear example is the market for new doctors (Roth, 1984).

The first job for almost all new doctors in the United States and Canada is as an intern or resident at a hospital. In the early 1900s, medical graduates were hired for such jobs near the end of their fourth year of medical school, just before graduation. By the 1930s, hiring was largely completed half a year before graduation, and by the 1940s it had moved to sometimes as much as two years before graduation. That is, in the early 1940s, students were being hired long before they would begin work, at dispersed times, and without much opportunity to consider alternatives, and long before they had sufficient experience to reveal either to employers or to themselves what kinds of medicine they would most prefer and be best able to practise. There was widespread recognition among the participants that the market was often failing to create the most productive matches of doctors to hospitals, both because there was too little opportunity to consider alternatives and because the matching was being done before important information about students became available.

One way in which many markets tried to address this failure was by attempting to establish rules concerning when offers could be made. In the market for new American doctors, the most concerted attempt at this kind of solution began in 1945 with the help of the medical schools, which agreed not to release any information to hospitals about students until a specified date.

However, the market experienced *congestion* in that hospitals found that they did not always have enough time to make all the offers they would like if their first offer was declined. Over the next few years students were called upon to make increasingly prompt decisions whether to accept offers. In 1945 offers were supposed to remain open for 10 days. By 1949 a deadline of 12 hours had been rejected as too long. Hospitals were finding that, if an offer was rejected after even a brief period of consideration, it was often too late for them to reach their next most preferred candidates before they had accepted other offers. Even when there was a long deadline much of this action was compressed into the last moments, because a student who had been offered a position that wasn't his first choice would be inclined to wait as long as possible before accepting, in the hope of eventually being offered a preferable position. So hospitals felt compelled to pressure students to reply immediately, and offers conveyed by telegram were frequently followed by phone calls requesting an immediate reply.

Congestion can be a problem in any market in which transactions take some time, but it is especially visible in entry-level professional labour markets in which many workers and jobs become available at the same time (for example, after graduation from university, medical school, law school, and so on). In the face of congestion, many markets unravel, as employers try to gain more time to make offers by starting to do so earlier (Roth and Xing, 1994). But the market for new doctors found a solution in the form of a centralized clearing house. Starting in the early 1950s, the various medical groups organized a centralized clearing house, which remains in use today, having undergone some changes over the years. Nowadays, a medical student applies to hospitals and goes on interviews in the winter of the final year of medical school, and then in February submits an ordered preference list of positions to the centralized clearing house, the National Resident Matching Program (NRMP). At the same time, the residency programmes (the employers) submit an ordered preference list of candidates. Once all the preference lists are collected, the clearing house uses an algorithm to produce a match, and residency programmes and applicants are informed to whom they have been matched. Although this clearing house began as an entirely voluntary one, it has been so successful that today it is virtually the only way that most residencies are filled. As we will see below, that success depends critically on the matching algorithm. The NRMP, and clearing houses like it, also make very clear the kinds of issues involved for a marketplace to make it safe for participants to reveal their information. In a clearing house in which you

are asked to state your preferences, the question is simply: is it a good idea to state your true preferences, or would you do better otherwise? For the NRMP we'll see that stating true preferences is indeed both safe and sensible. We'll also discuss clearing houses that failed this test, like the one for placing students into schools in Boston that consequently failed to accomplish their objectives in other ways also, and were redesigned.

Before presenting some formal models that will allow us to start to explain which matching algorithms and clearing houses have been successful and which have failed, it will be helpful to think about several different kinds of matching markets.

Two-sided and one-sided matching markets

Labour markets, like the market for new doctors, are usually modelled as two-sided markets, in which agents on one side of the market (workers) need to be matched with agents on the other side (employers), and each agent has preferences over possible matches. We'll see below that this two-sided structure allows strong conclusions to be drawn about the properties of matchings and matching mechanisms.

In many markets this two-sided structure is absent. One way this occurs is when any participant in the market can be matched with any other. For example, if a group of people want to form pairs to be roommates or bridge partners, any one of them can in principle be matched with any other, although not all matchings would be efficient. We encounter markets of this kind when we speak of kidney exchange. Another way in which markets can be one-sided is if the agents in the market need to be matched to objects, for example when people need to be assigned rooms in a dormitory, or places in a public school that doesn't itself have preferences or take strategic actions (unlike in a two-sided matching market). That is, such a market matches people to places, but only one side, the people, are active participants in the market. Some markets can also be hybrids, with both two and one-sided properties (as when schools aren't strategic players, but still have priorities over students).

Below we consider some static models of two and one-sided matching that have proved useful in the design of clearing houses, and in understanding what they do. In the section on design, we'll also speak about some decentralized design solutions to various market failures, such as unravelling. While there has been some good initial progress on formal models of decentralized markets, and dynamic models in which phenomena like unravelling can play out over time (see for example, Li and Rosen, 1998), these areas are still in need of development, and have not yet received the theoretical attention commensurate with their importance in the study of markets generally (though see Niederle and Yariv, 2007).

Formal models of matching

Two-sided matching models

The workhorse models of two-sided matching come in several varieties. The simplest, presented in detail below, is the ‘marriage model’ in which each firm seeks to hire only a single worker, and wages and other kinds of price adjustment are represented simply in the preferences that workers and firms have for each other (for example, in these models, wages are part of the job description that determines

preferences). However the kinds of results we present here generalize to models in which wage and price formation is explicitly included, some pointers to such models are included in the references.

The marriage model consists of two disjoint sets of agents, men={m_1,…, m_n} and women={w_1,…, w_p}, each of whom has complete and transitive preferences over the agents on the other side (and the possibility of being unmatched, which we model as being 'matched to yourself'). Preferences can be represented as rank order lists, for example, if man m_i's first choice is w_3, his second choice w_2 [w_3>m_i w_2] and so on, until at some point he prefers to remain unmatched, that is P(m_i)=w_3, w_2, … m_i …. If agent k (on either side of the market) prefers to remain single rather than be matched to agent j, that is, if k>k j, then j is said to be *unacceptable* to k. If an agent is not indifferent between any two acceptable mates, or between being matched and unmatched, we'll say he/she has *strict* preferences.

An *outcome* of the game is a *matching*: $\mu : M \cup W \rightarrow M \cup W$ such that w= μ (m) iff μ (w)=m, and for all m and w either μ (w) is in M or μ (w)=w, and either μ (m) is in W or μ (m)=m. That is, a matching matches agents on one side to agents on the other side, or to themselves, and if w is matched to m, then m is matched to w.

A matching μ is *blocked by an individual* k if k prefers being single to being matched with μ (k), that is, k>k μ (k). A matching μ is *blocked by a pair of agents* (m,w) if they each prefer each other to the partner they receive at μ, that is, w>m μ (m) and m>w μ (w).

A matching μ is *stable* if it isn't blocked by any individual or pair of agents.

A stable matching is Pareto efficient, and in the core, and in this simple model the set of (pairwise) stable matchings equals the core.

Theorem 1: (Gale and Shapley, 1962). A stable matching exists for every marriage market.

Gale and Shapley approached this problem from a purely theoretical perspective, but proved this theorem via a constructive algorithm of the kind that has subsequently turned up at the heart of a variety of clearing houses.

Deferred acceptance algorithm, with men proposing

(roughly the Gale and Shapley, 1962 version)

- *Step 0. If some preferences are not strict*, arbitrarily break ties (for example, if some m is indifferent between wi and wj, order them consecutively in alphabetical order. Different agents may break ties differently: that is, tie-breaking can be decentralized by having each agent fill out a strict preference list…).
- *Step 1*(*a*). Each man m proposes to his 1st choice (if he has any acceptable choices).
- *Step1*(*b*). Each woman rejects any unacceptable proposals and, if more than one acceptable proposal is received, 'holds' the most preferred and rejects all others… .
- *Step k*(*a*). Any man who was rejected at step k-1 makes a new proposal to its most preferred acceptable mate who hasn't yet rejected him. (If no acceptable choices remain, he makes no proposal.)
- *Step k*(*b*). Each woman holds her most preferred acceptable offer to date, and rejects the rest. STOP: when no further proposals are made, and match each woman to the man (if any) whose proposal she is holding.

Note that the proof of the theorem now follows from the observation that the matching produced in this way is itself stable. If some man would prefer to be matched to a woman other than his assigned mate, he must, according to the algorithm, have already proposed to her, and she has rejected him, meaning she has a man she strictly prefers, hence they cannot form a blocking pair.

Roth (1984) showed that the algorithm adopted by the medical clearing house in the 1950s was equivalent to the hospital proposing deferred acceptance algorithm. Gale and Shapley observed that which side of the market proposes in a deferred acceptance algorithm has consequences.

Theorem 2: (Gale and Shapley, 1962). When all men and women have strict preferences, there always exists an M-optimal stable matching (that every man likes at least as well as any other stable matching), and a W-optimal stable matching. Furthermore, the matching μM produced by the deferred acceptance algorithm with men proposing is the M-optimal stable matching. The W-optimal stable matching is the matching μW produced by the algorithm when the women propose.

Note that the algorithm has been stated as if people take actions in the course of the algorithm, and we can ask whether those actions would best serve their interests. To put it another way, is it possible to design a clearing house in which a matching is produced from participants' stated rank order lists in such a way that it will never be in someone's interest to submit a rank order list different from their true preferences? The following theorem answers that question in the negative.

Theorem 3: **Impossibility Theorem** (Roth – see Roth and Sotomayor, 1990). No stable matching mechanism exists for which stating the true preferences is a dominant strategy for every agent. However it is possible to design the mechanism so that one side of the market can never do any better than to state their true preferences.

Theorem 4: (Dubins and Freedman, Roth – see Roth and Sotomayor, 1990).

The mechanism that yields the M-optimal stable matching (in terms of the stated preferences) makes it a dominant strategy for each man to state his true preferences.

The conclusions of Theorems 1–3 also hold for a variety of related models (in which firms employ multiple workers, and wages are explicitly allowed to vary; see, for example, Shapley and Shubik, 1971; Kelso and Crawford, 1982, for notable early models of matching with money, and see Roth and Sotomayor, 1990; Hatfield and Milgrom, 2005). However, when we look at many-to-one matching models (in which firms employ multiple workers but workers seek just one job), we have to be careful. It turns out that no procedures exist that give firms a dominant strategy, but that a worker proposing deferred acceptance algorithm still makes it a dominant strategy for workers to state their true preferences (see Roth and Sotomayor, 1990 for more details and further references). (These results are closely connected to related results in auction theory; see in particular Hatfield and Milgrom, 2005; Milgrom, 2004.)

When the market for medical residents was redesigned (Roth and Peranson, 1999), a number of practical complications had to be dealt with, such as the fact that about 1,000 graduates a year go through the match as couples who wish to be matched to nearby jobs, and hence have joint preferences over pairs of residency programmes. While this can cause the set of stable matchings to be empty, in practice this has not proved to be a significant problem (see also Roth, 2002, on engineering aspects of economic design).

One-sided matching models

Shapley and Scarf's ‘ house ’ markets

Another basic model of matching markets was introduced by Shapley and Scarf (1974). They model a simple barter economy in which each one of n agents owns an indivisible good (which they call a house) and has preferences over all houses in the economy. Each agent has use for only one house and trade is only feasible in houses (that is, there is no money in their model). An allocation μ in this context is a matching of houses and agents so that each agent receives one and only one house. An exchange in this market does not need to be bilateral. An allocation μ is in the core if no coalition (including single agent coalitions) of agents can improve upon it (in the sense that all are weakly better off and at least one is strictly better off) by swapping their own houses. Shapley and Scarf attribute to David Gale the following *top trading cycles* algorithm (TTC) which can be used to find a core allocation for any housing market:

- *Step 1*: Each agent points to the owner of her most preferred house (which could possibly be herself). Since there are finite number of agents there is at least one *cycle* (where a cycle is an ordered list (i_1, i_2, …, i_k) of agents with each agent pointing to the next agent in the list and agent *ik* pointing to agent i_1). In each cycle the implied exchange is carried out and the procedure is repeated with the remaining agents.

In general, at

- *Step k*: Each remaining agent points to the owner of her most preferred house among the remaining houses. There is at least one cycle. In each cycle the implied exchange is carried out and the procedure is repeated with the remaining agents.

The algorithm terminates when each agent receives a house.
Theorem 5: (Shapley and Scarf, 1974). The TTC algorithm yields an allocation in the core for each housing market.
The core has some remarkable properties in the context of housing markets. The following propositions summarize the most notable of these results.
While exchange is feasible only in houses, a *competitive allocation* of a housing market can be defined via ‘token money’. There is an important relation between the core and the competitive allocation for this very basic barter economy.
Theorem 6: (Roth and Postlewaite, 1977). There is a unique allocation in the core (which can be obtained with the TTC algorithm) when agents have strict preferences over houses. Moreover the unique core allocation coincides with the unique competitive allocation.
Another remarkable feature of this model is that the top trading cycles mechanism makes it safe for agents to reveal their true preferences.
Theorem 7: (Roth, 1982). The core as a mechanism is *strategy-proof* when agents have strict preferences over houses. That is, truth-telling is a dominant strategy for all agents in the preference revelation game in which TTC is applied to the stated preferences to produce an allocation.

Moreover, it is essentially the only mechanism that is strategy-proof among those that are Pareto efficient and *individually rational* (in the sense that an agent never receives a house inferior to her own).
Theorem 8: (Ma, 1994). The core is the only mechanism that is Pareto efficient, individually rational and strategy-proof.

House allocation problems

Hylland and Zeckhauser (1979) introduced the *house allocation problem* which only differs from housing markets in *property rights*: There are *n* houses to be allocated for *n* agents where each agent has use for only one house and has strict preferences over all houses. Unlike in housing markets, no agent owns a specific house. The mechanism known as *random serial dictatorship* (RSD) is widely used in real-life allocation problems of this sort, such as assigning students to dormitory rooms. Under RSD agents are randomly ordered (from a uniform distribution) in a list and the first agent in the list is assigned her top choice house, the next agent is assigned her top choice among the remaining houses, and so on. In addition to its popularity in practice, RSD has good incentive and efficiency properties.
Theorem 9: RSD is *ex post* Pareto efficient and strategy-proof.
Recall that the only difference between house allocation problems and housing markets is the initial property rights, and the core is very well-behaved in the context of the latter. This observation motivates the mechanism *core from random endowments* (CRE): randomly assign houses to agents with uniform distribution, interpret the resulting matching as the initial allocation of houses, and pick the core of the resulting housing market. It turns out, CRE is equivalent to RSD.
Theorem 10: (Abdulkadiroglu and Sönmez, 1998). For any house allocation problem CRE and RSD yield the same lottery and hence they are equivalent mechanisms.

House allocation with existing tenants

Housing markets and house allocation problems have very different property rights. The former is a pure private ownership economy where each house ‘belongs’ to a specific agent, whereas in the latter no strict subset of the grand coalition has claims on any house. Abdulkadiroglu and Sönmez (1999) introduced the following hybrid *house allocation with existing tenants* model. There are two kinds of agents: *existing tenants* each of whom owns a house, and *newcomers* none of whom has claims on a specific house. In addition to the *occupied houses* owned by existing tenants, there are also *vacant houses*. As in house allocation problems no specific person or group has claims on any vacant house. Suppose that the number of newcomers is equal to the number of vacant houses and hence the number of agents is equal to the number of houses. Agents have strict preferences over all houses and each existing tenant is allowed to keep her current house.
Abdulkadiroglu and Sönmez introduced the following *you request my house – I get your turn* algorithm (YRMH–IGYT) which generalizes TTC as well as RSD. Under YRMH–IGYT, agents are randomly ordered in a line and initially only the vacant houses are *available*. The first agent in the line is assigned her top choice provided that it is either her own house or an available house (in which case her own house becomes available) and the process continues with the next agent in the line. If, however, her top choice is an occupied house, the line is adjusted and the owner of the requested house is moved right in front of the requester. The process continues in a similar way with either the owner of the requested

house getting assigned his own house or an available house (making his own house available), or otherwise his requesting an occupied house and upgrading its owner to the top of the line. When the process continues in a similar way there will either be a *cycle* of existing tenants (as in TTC) who can swap their own houses or a *chain* (i_1, i_2, …, i_k) of agents where agent i_1 is assigned an available house and each of the following agents is assigned the preceding agent's house.
The resulting mechanism inherits the attractive properties of its ‘parents’.
Theorem 11: (Abdulkadiroglu and Sönmez, 1999). The YRMH–IGYT mechanism is strategy-proof, *ex post* Pareto efficient, and individually rational (in the sense that no existing tenant receives a house inferior to her own).

Kidney exchange

Living donors are an important source of kidneys for transplantation. But a patient with a willing living donor may not be able to receive a transplant because of a blood-type or immunologic incompatibility between her and her donor. Recently transplant centres around the world developed the possibility of *pairwise kidney exchange* in which two such pairs can exchange donors in case the donor in each pair is compatible with the patient in the other. Another interesting option is *indirect kidney exchange* in which the patient of an incompatible pair receives priority in the deceased donor waiting list if her incompatible donor donates a kidney to that waiting list. However, prior to 2004 only a very few exchanges had been accomplished, in large part because the market wasn't thick, and no databases were being maintained of incompatible patient–donor pairs. In an effort to organize kidney exchange on a larger scale, Roth, Sönmez and Ünver (2004) introduced the following *kidney exchange* model. There are a number of patients each with a (possibly) incompatible donor. For each patient a subset of donors can feasibly donate a kidney and the patient has strict preferences over these donors and his own donor (who may or may not be compatible with him). In addition to ranking all compatible donors, each patient also ranks a ‘waiting list option’ which represents trading his donor's kidney with a priority in the waitlist. An *allocation* in this context is a matching of patients and donors such that:

- each patient is matched with either a donor or the waiting list option, and
- each donor can be matched with at most one patient while the waiting list option may be matched with multiple patients.

(The donors who remain unmatched are offered to the waitlist in exchange for the equal number of priorities awarded by the allocation). We are only interested in *individually rational* allocations where patients receive neither a donor nor the waiting list option unless it is indicated to be at least as good as his donor's kidney. If the waiting list option is ranked inferior to his donor for a patient, that means the patient is not interested in such an exchange. As in the case of house allocation with existing tenants model, an allocation consists of cycles and chains where

- each patient in a cycle receives a kidney from the donor of the next patient in the cycle, and
- all but the last patient in a chain receive a kidney from the donor of the next patient in the chain whereas the last patient in the chain receives a priority in the waiting list.

If the waiting list option is infeasible, then the resulting problem is formally equivalent to a housing market and therefore has a unique allocation in the core which can be obtained via the TTC algorithm. In this simpler model an allocation (including the one in the core) consists of only cycles. When the waiting list option is feasible an allocation can also have chains (which are indirect exchanges and their more elaborate versions). In this more general model Roth, Sönmez and Ünver (2004) introduce a class of *top trading cycles and chains* (TTCC) algorithms each of which extend the TTC. Among these algorithms Roth, Sönmez and Ünver (2004) identify one that is Pareto efficient and strategy-proof:

Theorem 12: (Roth, Sönmez and Ünver, 2004). There exists a TTCC mechanism that is Pareto efficient and strategy-proof.

In practice, as kidney exchanges have become organized on a larger scale in New England and elsewhere (see Roth, Sönmez and Ünver, 2005a; 2005b; 2007), there has been a focus, for logistical reasons, on cycles and chains that are relatively short, typically only involving exchanges among two or three patient–donor pairs.

The deferred acceptance algorithm (for two-sided markets) also has some uses in one-sided allocation problems in which children are to be allocated to schools, if the schools, although not active strategic players, have priorities over students that need to be treated like preferences (Abdulkadiroglu and Sönmez, 2003).

Design and engineering

Introducing a centralized stable match

Of the several dozen markets and submarkets we know of that established clearing houses in response to unravelling in a (two-sided) labour market, those that produce stable matchings have been most successful. Of particular note in this regard are the markets used in the various regions of the British National Health Service. In the 1960s, these markets suffered from the same kind of unravelling that had afflicted the American medical market in the 1940s. A Royal Commission recommended that each region organize a centralized clearing house (see Roth, 1991), and the various regions each invented their own matching algorithms, some of which were stable and some of which were not (an example of such unstable algorithms will be given later). Those clearing houses that produced stable matches succeeded, while those that did not most often failed and were abandoned. But over a broad range of markets, the correlation between stability and success in halting unravelling isn't perfect; some unstable mechanisms remain in use, and some stable mechanism have occasionally failed, as we will discuss later. And there are other differences between markets than the way their clearing houses are designed. This is why, in order to establish that producing a stable outcome is an important feature for the success of a match, controlled experiments in the laboratory can be informative.

The laboratory experiments reported by Kagel and Roth (2002) help to verify the influence of a stable or unstable matching mechanism. After unravelling had begun in a small laboratory market, a clearing house was introduced using either the stable deferred acceptance algorithm or the unstable algorithms that failed in various regions of the British National Health service (Roth, 1991). In the lab, as in the field, participants learned to wait for and use the stable algorithm, but learned to arrange their matches

early and thus avoid using the unstable algorithm. Note that a laboratory market is quite different from a naturally occurring labour market, but it has the advantage that it allows the effect of the different algorithms to be observed in an environment in which everything else is the same.

Centralized clearing houses that yield stable outcomes have sometimes been introduced to organize markets suffering from failures other than unravelling (and the resulting lack of thickness), but related to congestion or the safety of revealing private information.

Examples of algorithms that produce unstable outcomes, but have been used in a number of market clearing houses, are so called priority algorithms, used for example by some British clearing houses, and also in several school choice problems in the United States. A priority algorithm classifies different matches in terms of priorities, based on the rank orders submitted, and then makes feasible matches in order of priority. In Boston, for example, the centralized system attempted to give as many students as possible their first choice school. The difficulty with the system was that students who did not get assigned to their first choice were much less likely to be assigned to the school they had listed as their second choice than they would have been if they had listed it as their first choice, since those schools often get filled by students who list them as their first choice. This means participants have strong incentives to not report their preferences truthfully, if there is a good chance that they would not be admitted to their true first choice school; it might be wiser to list their second-choice school as their first choice. The newly adopted Boston clearing house fixes this problem using a deferred acceptance algorithm (Abdulkadiroglu et al., 2005; 2006).

Some markets manage to halt unravelling, but still suffer from congestion. The market for clinical psychologists (before it reorganized through a modified deferred acceptance algorithm, see Roth and Xing, 1997) and the match of students to New York City high schools before it was redesigned (Abdulkadiroglu et al., 2005; 2006) are good examples. Clinical psychologists tried to run a deferred acceptance algorithm over the phone in the course of a day, ‘match day,’ from 9:00 a.m. to 4:00 p.m.. All offers had to remain open until 4 p.m., and students were supposed to hold only one offer at a time. Even though turnaround time in this market was very fast (offers took about five minutes, rejections about one minute), simulating a deferred acceptance algorithm in real time, for a market with about 2,000 positions in 500 programmes, takes much longer than the seven hours of match day. (And making the market longer may increase the effects of congestion, if it means that participants can no longer stay by the phone for the whole market, so that the time for an offer to be made and rejected becomes disproportionately longer.) Congestion is an issue whenever a large number of offers have to be made. The system used to assign students to New York City high schools used to be carried out through the mails, and over 30,000 students a year were ‘stranded’ on waiting lists and had to be assigned to a school for which they had expressed no preference. The new New York City clearing house is able to process preferences quickly, and in the four years following its adoption in 2003 fewer than 3,000 students had to be assigned each year to a school for which they had expressed no preference.

What are the effects of a centralized match?

Centralized clearing houses can help make markets thick and uncongested, and avoid unravelling. Studying their effect on various markets can also help us understand how clearing houses and the timing of the market (for example, how far a labour market operates in advance of employment) influence the

outcome of the market in other respects. For example, the market for gastroenterology fellows provides us with a natural case study of the effects of a clearing house not only on hiring practices (namely the timing of the market, and the kinds of offers that are made), but also employment opportunities, job placement and the potential impact on wages.

Gastroenterology fellows are doctors who have completed three years of residency in internal medicine, and are now employed in a fellowship that will result in their becoming board certified sub-specialists in gastroenterology. The market in which gastroenterology fellows are hired operated in a decentralized way for many years, and experienced the problems of congestion, unravelling and exploding offers, as described above in connection with the market for medical residents. In 1986, various internal medicine sub-specialties organized a clearing house called the Medical Specialties Matching Program (MSMP), sponsored by and organized along the same lines as the NRMP (which operates the resident match). But in the mid-1990s, gastroenterology fellowship programmes, and applicants, started to defect from the match, and the gastroenterology market again unravelled. A match was successfully re-established only in 2006 (Niederle, Proctor and Roth, 2006). In those intervening years, as the market unravelled, the national market broke up into more local markets (Niederle and Roth, 2003b). Fellowship programmes, particularly smaller ones, had a larger tendency to hire their own residents than under a centralized match.

A second aspect of the outcome that received prominence in 2002 is the question of whether a match affects wages. An antitrust lawsuit against the NRMP and numerous other defendants was brought in 2002 by 16 law firms on behalf of three former residents seeking to represent the class of all former residents (and naming as defendants a class including all hospitals that employ residents).

Niederle and Roth (2003a) showed empirically that in fact there is no difference in wages between medicine sub-specialties that use a match and those that don't. The suit was dismissed in 2004 following legislation intended to clarify that the medical match is a marketplace and does not violate antitrust laws. The theory of the complaint was that a match holds down wages for residents and fellows. Bulow and Levin (2006) present a very stylized theoretical model providing some logical support for this possibility, by comparing a market with impersonal prices (to represent the NRMP) with perfectly competitive prices at which each worker is paid his or her marginal product. Subsequent theoretical papers have shown that the conclusion about wage suppression doesn't necessarily follow if the model is expanded to include the possibility of firms hiring more than one worker (Kojima, 2007), or when the model incorporates the actual procedures by which the medical match is conducted (Niederle, 2007). Furthermore, decentralized markets may often fail to achieve stable outcomes (Niederle and Yariv, 2007).

Beyond centralized matching. why do some markets work well, while others do not?

We have seen that stability is an important feature for a centralized match to remain in use. However, the history of the gastroenterology market shows that producing a stable outcome is not sufficient to guarantee a successful clearing house. For a centralized match to work well, participants need to have incentives to participate in the match. McKinney, Niederle and Roth (2005) observed that the collapse of the gastroenterology fellowship match seems to have been caused by an unusual shock to the supply of highly qualified gastroenterology fellows, a kind of shock that was not observed in other internal

medicine sub-specialties that continued to use a match. Furthermore, market conditions seemed to have stabilized, so that a centralized match would work well once again, if it could be successfully reinstated. However, many gastroenterology fellowship programmes, when they considered reinstituting a match, were concerned that, while they were willing to refrain from making the early offers that had become customary, and wait for the match, their main competitors would continue to make early exploding offers to promising applicants. Such concerns could effectively prevent a successful restart of a centralized clearing house.

This raises the more general question as to why some markets unravel and experience congestion problems in the first place, while others do not. Empirically, most markets that experience congestion also experience that employers (hospitals, federal judges, colleges…) make short-term offers, with a binding deadline, and in which the acceptance of an offer is often effectively binding (Niederle and Roth, 2007, for descriptions in the markets for law graduates, and for college admissions, see for example, Avery et al., 2001; 2007; Avery, Fairbanks and Zeckhauser, 2003).

On the other hand there are markets that do not unravel, such as the market for graduate school admission. In this market, a policy (adopted by the large majority of universities) states that offers of admission and financial support to graduate students should remain open until 15 April. Furthermore, a student faced with an earlier deadline is explicitly encouraged to accept this offer, and, in case a better one is received before 15 April, to renege on that former acceptance. This of course makes early exploding offers much less attractive to make. Niederle and Roth (2007) explore environments in which either eliminating the possibility of making exploding offers or making early acceptances non-binding helps prevents markets from operating inefficiently early.

These insights were used to help reorganizing the gastroenterology fellowship match. To reduce the concerns of programmes that their competitors would start making exploding offers before the match, a resolution was adopted by the four main professional gastroenterology organizations that stated that acceptances made before the match were not to be considered binding, and such applicants could still change their minds and participate in the match. For an account of the effects of a centralized clearinghouse on the outcomes of a market, and the experience of the gastroenterology fellowship market, see Niederle and Roth (2008).

Directions for future research

As economists' understanding of the matching function of markets increases, and as economists are more often called upon to help design markets, one challenge will be to understand better how decentralized markets work well or badly, and not only in the final transactions.

For example, a common problem in many entry-level labour markets (and in dating and marriage markets) is that participants do not have well formed preferences over potential matching partners, and forming those preferences is often very costly. For example, in the American market for assistant professors, economics departments receive hundreds of applications for any position, but in general interview only about 30 candidates at the annual winter meetings. From among those they interview, they must decide whom to fly out for extended campus visits and seminars, and it is from among this latter set of candidates that they eventually choose to whom to make an offer. Because this is a time-consuming and costly process many departments have to take care to interview applicants who not only

have a good chance of being desirable colleagues, but who also have a good chance of accepting an offer if one is made. This often amounts to a coordination problem: not all departments should interview the same applicants. Allowing applicants to credibly submit information about their interest in particular schools can help alleviate this coordination problem, and in 2007 the American Economic Association implemented a signalling mechanism of this sort in the market for economists.

In general, the study of the matching function of markets has directed attention at the design of rules and procedures of both centralized and decentralized markets. The goal of the growing interest among economists in matching and market design is to understand the operation of markets, both centralized and decentralized, well enough so we can fix them when they're broken.

See Also

- experimental economics
- experimental labour economics
- game theory
- labour market institutions
- matching
- mechanism design experiments
- mechanism design (new developments)

Bibliography

Abdulkadiroglu, A., Pathak, P.A. and Roth, A.E. 2005. The New York City high school match. *American Economic Review* 95, 364–67.

Abdulkadiroglu, A., Pathak, P.A. and Roth, A.E. 2006. Strategy-proofness versus efficiency in matching with indifferences: redesigning the NYC high school match, Working paper, Harvard University.

Abdulkadiroglu, A. and Sönmez, T. 1998. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66, 689–701.

Abdulkadiroglu, A. and Sönmez, T. 1999. House allocation with existing tenants. *Journal of Economic Theory* 88, 233–60.

Abdulkadiroglu, A. and Sönmez, T. 2003. School choice: a mechanism design approach. *American Economic Review* 93, 729–47.

Abdulkadiroglu, A., Pathak, P.A., Roth, A.E. and Sönmez, T. 2005. The Boston public school match. *American Economic Review* 95, 368–71.

Abdulkadiroglu, A., Pathak, P.A., Roth, A.E. and Sönmez, T. 2006. Changing the Boston school choice

mechanism, Working Paper No. 11965. Cambridge, MA: NBER.

Avery, C., Fairbanks, A. and Zeckhauser, R. 2003. *The Early Admissions Game: Joining the Elite*. Cambridge, MA: Harvard University Press.

Avery, C., Jolls, C., Posner, R.A. and Roth, A.E. 2001. The market for federal judicial law clerks. *University of Chicago Law Review* 68, 793–902.

Avery, C., Jolls, C., Posner, R.A. and Roth, A.E. 2007. The new market for federal judicial law clerks. *University of Chicago Law Review* 74, 447–86.

Bulow, J. and Levin, J. 2006. Matching and price competition. *American Economic Review* 96, 652–68.

Gale, D. and Shapley, L.S. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69, 9–15.

Hatfield, J. and Milgrom, P. 2005. Matching with contracts. *American Economic Review* 95, 913–35.

Hylland, A. and Zeckhauser, R. 1979. The efficient allocation of individuals to positions. *Journal of Political Economy* 87, 293–314.

Kagel, J.H. and Roth, A.E. 2000. The dynamics of reorganization in matching markets: a laboratory experiment motivated by a natural experiment. *Quarterly Journal of Economics* 115, 201–35.

Kelso, A.S. and Crawford, V.P. 1982. Job matching, coalition formation, and gross substitutes. *Econometrica* 50, 1483–504.

Kojima, F. 2007. Matching and price competition. *American Economic Review* 97, 1027–31.

Li, H. and Rosen, S. 1998. Unraveling in matching markets. *American Economic Review* 88, 371–87.

Ma, J. 1994. Strategy-proofness and the strict core in a market with indivisibilities. *International Journal of Game Theory* 23, 75–83.

Milgrom, P. 2004. *Putting Auction Theory to Work*. Cambridge: Cambridge University Press.

McKinney, C.N., Niederle, M. and Roth, A.E. 2005. The collapse of a medical labor clearinghouse (and why such failures are rare). *American Economic Review* 95, 878–89.

Niederle, M. 2007. Competitive wages in a match with ordered contracts, *American Economic Review*.

Niederle, M., Proctor, D.D. and Roth, A.E. 2006. What will be needed for the new GI fellowship match to succeed? *Gastroenterology* 130, 218–24.

Niederle, M. and Roth, A.E. 2003a. Relationship between wages and presence of a match in medical fellowships. *Journal of the American Medical Association* 290, 1153–4.

Niederle, M. and Roth, A.E. 2003b. Unraveling reduces mobility in a labor market: gastroenterology with and without a centralized match. *Journal of Political Economy* 111, 1342–52.

Niederle, M. and Roth, A.E. 2007. Making markets thick: how norms governing exploding offers affect market performance. Working paper.

Niederle, M. and Roth, A.E. 2008. The effects of a central clearinghouse on job placement, wages, and hiring practices. In *Labor Market Intermediation*, ed. D. Autor. Chicago: University of Chicago Press.

Niederle, M. and Yariv, L. 2007. Matching through decentralized markets. Working paper.

Roth, A.E. 1982. Incentive compatibility in a market with indivisible goods. *Economics Letters* 9, 127–32.

Roth, A.E. 1984. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of Political Economy* 92, 991–1016.

Roth, A.E. 1991. A natural experiment in the organization of entry level labor markets: regional markets for new physicians and surgeons in the U.K. *American Economic Review* 81, 415–40.

Roth, A.E. 2002. The economist as engineer: game theory, experimental economics and computation as tools of design economics. *Econometrica* 70, 1341–78.

Roth, A.E. and Peranson, E. 1999. The redesign of the matching market for American physicians: some engineering aspects of economic design. *American Economic Review* 89, 748–79.

Roth, A.E. and Postlewaite, A. 1977. Weak versus strong domination in a market with indivisible goods. *Journal of Mathematical Economics* 4, 131–7.

Roth, A.E., Sönmez, T. and Ünver, M.U. 2004. Kidney exchange. *Quarterly Journal of Economics* 119, 457–88.

Roth, A.E., Sönmez, T. and Ünver, M.U. 2005a. A kidney exchange clearinghouse in New England. *American Economic Review* 95, 376–80.

Roth, A.E., Sönmez, T. and Ünver, M.U. 2005b. Pairwise kidney exchange. *Journal of Economic Theory* 125, 151–88.

Roth, A.E., Sönmez, T. and Ünver, M.U. 2007. Efficient kidney exchange: coincidence of wants in markets with compatibility-based preferences. *American Economic Review* 97, 828–51.

Roth, A.E. and Sotomayor, M. 1990. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. New York: Cambridge University Press.

Roth, A.E. and Xing, X. 1994. Jumping the gun: imperfections and institutions related to the timing of market transactions. *American Economic Review* 84, 992–1044.

Roth, A.E. and Xing, X. 1997. Turnaround time and bottlenecks in market clearing: decentralized matching in the market for clinical psychologists. *Journal of Political Economy* 105, 284–329.

Roth, A.E., Sönmez, T., Ünver, M.U., Delmonico, F.L. and Saidman, S.L. 2006. Utilizing list exchange and undirected Good Samaritan donation through ‘chain’ paired kidney donations. *American Journal of Transplantation* 6, 2694–705.

Shapley, L. and Scarf, H. 1974. On cores and indivisibility. *Journal of Mathematical Economics* 1, 23–37.

Shapley, L. and Shubik, M. 1971. The assignment game I: the core. *International Journal of Game Theory* 1, 111–30.

How to cite this article

Niederle, Muriel, Alvin E. Roth and Tayfun Sönmez. "matching and market design." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000356> doi:10.1057/9780230226203.1058

The New Palgrave Dictionary of Economics Online

matching estimators

Petra E. Todd
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Matching methods are a popular method for evaluating the effects of programme or other treatment interventions. This article reviews recent developments in the econometric literature on matching estimators, including the assumptions required to justify their application, different ways of implementing the estimators and some recent empirical applications.

Keywords

bootstrap; curse of dimensionality; kernel estimation; local linear estimation; matching; matching estimators; nearest-neighbour matching; programme effect; propensity score; semiparametric estimation; treatment effect

Article

1 Introduction

Matching is a widely used non-experimental method of evaluation that can be used to estimate the average effect of a treatment or programme intervention. The method compares the outcomes of programme participants with those of matched non-participants, where matches are chosen on the basis of similarity in observed characteristics. One of the main advantages of matching estimators is that they typically do not require specifying the functional form of the outcome equation and are therefore not susceptible to misspecification bias along that dimension. Traditional matching estimators pair each programme participant with a single matched non-participant (see, for example, Rosenbaum and Rubin, 1983), whereas more recently developed estimators pair programme participants with multiple non-participants and use weighted averaging to construct the matched outcomes.

We next define some notation and discuss how matching estimators solve the evaluation problem. Much of the treatment effect literature is built on the potential outcomes framework of Fisher (1935), exposited more recently in Rubin (1974) and Holland (1986). The framework assumes that there are two potential outcomes, denoted (Y_0, Y_1) that represent the states of being without and with treatment. An individual can be in only one state at a time, so only one of the outcomes is observed. The outcome that is not observed is termed a *counterfactual outcome*. The treatment impact for an individual is

$$\Delta = Y_1 - Y_0,$$

which is not directly observable. Assessing the impact of a programme intervention requires making an inference about what outcomes would have been observed in the no-programme state. Let D=1 for persons who participate in the programme and D=0 for persons who do not. The D=1 sample often represents a select group of persons who were deemed eligible for a programme, applied to it, got accepted into it and decided to participate in it. The outcome that is observed is $Y = DY_1 + (1 - D)Y_0$.

Before considering different parameters of interest and their estimation, we first consider what is available directly from the data. The conditional distributions $F(Y_1|X, D = 1)$ and $F(Y_0|X, D = 0)$ can be recovered from the observations on Y_1 and Y_0, but not the joint distributions $F(Y_0, Y_1|X, D = 1)$, $F(Y_0, Y_1|X)$ or the impact distribution, $F(\Delta|X, D = 1)$. Because of this missing data problem, researchers often aim instead on recovering some features of the impact distribution, such as its mean. The parameter that is most commonly the focus of evaluation studies is the *mean impact of treatment on the treated*, $TT = E(Y_1 - Y_0|D = 1)$, which gives the benefit of the programme to programme participants. (If the outcome were earnings and the TT parameter exceeded the average cost of the programme, then the programme might be considered to at least cover its costs.)

Matching estimators typically assume that there exist a set of observed characteristics Z such that outcomes are independent of programme participation conditional on Z. That is, it is assumed that the outcomes (Y_0,Y_1) are independent of participation status D conditional on Z,

$$(Y_0, Y_1) \perp \perp D|Z. \tag{1}$$

The independence condition can be equivalently represented as $\Pr(D = 1|Y_0, Y_1, Z) = \Pr(D = 1|Z)$, or $E(D|Y_0, Y_1, Z) = E(D|Z)$. In the terminology of Rosenbaum and Rubin, 1983, treatment assignment is ‘strictly ignorable’ given Z. It is also assumed that for all Z there is a positive probability of either participating (D=1) or not participating (D=0) in the programme: that is,

$$0 < \Pr(D = 1|Z) < 1. \tag{2}$$

This assumption is required so that matches for D=0 and D=1 observations can be found. If assumptions (1) and (2) are satisfied, then the problem of determining mean programme impacts can be solved by substituting the Y_0 distribution observed for matched on Z non-participants for the missing participant Y_0 distribution.

The above assumptions are overly strong if the parameter of interest is the mean impact of treatment on the

treated (TT), in which case a weaker conditional mean independence assumption on Y_0 suffices (see Heckman, Ichimura and Todd, 1998):

$$E(Y_0|Z, D = 1) = E(Y_0|Z, D = 0) = E(Y_0|Z). \tag{3}$$

Furthermore, when TT is the parameter of interest, the condition $0 < \Pr(D = 1|Z)$ is also not required, because that condition is only needed to guarantee a participant analogue for each non-participant. The TT parameter requires only

$$\Pr(D = 1|Z) < 1. \tag{4}$$

Under these assumptions, the mean impact of the programme on programme participants can be written as

$$\Delta_{TT} = E(Y_1 - Y_0|D = 1) = E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 1, Z)\} = E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 0, Z)\},$$

where the second term can be estimated from the mean outcomes of the matched on Z comparison group. (The notation $E_{Z|D=1}$ denotes that the expectation is taken with respect to the $f(Z|D = 1)$ density.)
Assumption (3) implies that D does not help predict values of Y_0 conditional on Z which rules out selection into the programme directly on values of Y_0. However, there is no similar restriction imposed on Y_1, so the method does allow individuals who expect to experience higher levels of Y_1 to select into the programme on the basis of that information. For estimating the *TT* parameter, matching methods allow selection into treatment to be based on possibly unobserved components of the anticipated programme impact, but only in so far as the programme participation decisions are based on the unobservable determinants of Y_1 and not those of Y_0.
Second, the matching method also requires that the distribution of the matching variables, Z, not be affected by whether the treatment is received. For example, age, gender, and race would generally be valid matching variables, but marital status may not be if it were potentially affected by receipt of the programme. To see why this assumption is necessary, consider the term

$$E_{Z|D=1}\left\{E_Y(Y|D = 0, Z)\right\} = \int_{z\in Z}\int_{y\in Y} y\, f(y|D = 0, z)\, f(z|D = 1)\mathrm{d}z.$$

It uses the $f(z|D = 1)$ conditional density to represent the density that would also have been observed in the no

treatment (D=0) state, which rules out the possibility that receipt of treatment changes the density of Z. Variables that are likely to be affected by the treatment or programme intervention cannot be used in the set of matching variables.

With non-experimental data, there may or may not exist a set of observed conditioning variables for which (1) and (2), or (3) and (4), hold. A finding of Heckman, Ichimura and Todd (1997) and Heckman et al. (1996; 1998) in their application of matching methods to data from the Job Training and Partnership Act (JTPA) programme is that (2) and (4) were not satisfied, because no match could be found for a fraction of the participants. If there are regions where the support of Z does not overlap for the D=1 and D=0 groups, then matching is justified only when performed over the *region of common support*. The estimated treatment effect must then be defined conditionally on the region of overlap. Some methods for empirically determining the overlap region are described below.

Matching estimators can be difficult to implement when the set of conditioning variables Z is large. If Z are discrete, small-cell problems may arise. If Z are continuous and the conditional mean $E(Y_0|D = 0, Z)$ is estimated nonparametrically, then convergence rates will be slow due to the so-called *curse of dimensionality* problem. Rosenbaum and Rubin (1983) provide a theorem that can be used to address this dimensionality problem. They show that for random variables Y and Z and a discrete random variable D

$$E(D|Y, P(D = 1|Z)) = E(E(D|Y, Z)|Y, \Pr(D = 1|Z)),$$

so that

$$E(D|Y, Z) = E(D|Z) \Rightarrow E(D|Y, \Pr(D = 1|Z) = E(D|\Pr(D = 1|Z)).$$

This result implies that, when Y_0 outcomes are independent of programme participation conditional on Z, they are also independent of participation conditional on the probability of participation, $P(Z) = \Pr(D = 1|Z)$. That is, when matching on Z is valid, matching on the summary statistic $\Pr(D = 1|Z)$ (the *propensity score*) is also valid. Provided that $P(Z)$ can be estimated parametrically (or semiparametrically at a rate faster than the nonparametric rate), matching on the propensity score reduces the dimensionality of the matching problem to that of a univariate problem. For this reason, much of the literature on matching focuses on propensity score matching methods. (Heckman, Ichumura and Todd, 1998, and Hahn, 1998, consider whether it is better in terms of efficiency to match on $P(X)$ or on X directly.) With the use of the Rosenbaum and Rubin (1983) theorem, the matching procedure can be broken down into two stages. In the first stage, the propensity score $\Pr(D = 1|Z)$ is estimated, using a binary discrete choice model. (Options for first the stage estimation include, for example, a parametric logit or probit model or a semiparametric estimator, such as semiparametric least squares – Ichimura, 1993 – maximum score – Manski, 1973 – smoothed maximum score – Horowitz, 1992 – or semiparametric maximum likelihood – Klein and Spady, 1993. If $P(Z)$ were estimated using a fully nonparametric method, then the curse of dimensionality problem would reappear.) In the second stage, individuals are matched on the basis of their predicted probabilities of participation.

We next describe a simple model of the programme participation decision to illustrate the kinds of assumptions

needed to justify matching. (This model is similar to an example given in Heckman, Lalonde and Smith, 1999.) Assume that an individual chooses whether to apply to a training programme on the basis of the expected benefits. He or she compares the expected earnings streams with and without participating, taking into account opportunity costs and net of some random training cost ε, which may include a psychic component expressed in monetary terms. The participation decision is made at time t=0 and the training programme lasts for periods 1 through τ, during which time earnings are zero. The information set used to determine expected earnings is denoted by W, which might include, for example, earnings and employment history. The participation model is

$$D = 1 \text{ if } E\left(\sum_{j=\tau}^{T} \frac{Y_{1j}}{(1+r)^j} - \sum_{k=1}^{T} \frac{Y_{0k}}{(1+r)^k} | W\right) > \varepsilon + Y_{00}, \text{ else } D = 0.$$

The terms in the right-hand side of the inequality are assumed to be known to the individual but not to the econometrician.
If $f(Y_{0k}|\varepsilon + Y_{00}, X) = f(Y_{0k}|X)$, then

$$E(Y_{0k}|X, D = 1) = E(Y_{0k}|X, \varepsilon + Y_{00} < \eta(W)) = E(Y_{0k}|X),$$

which would justify application of a matching estimator. This assumption places restrictions on the correlation structure of the earnings residuals. For example, the assumption would not be plausible if X=W and Y_{00}=Y_{0k}, because knowing that a person selected into the programme (D=1) would likely be informative about subsequent earnings. We could assume, however, a model for earnings such as

$$Y_{0k} = \varphi(X) + v_{0k},$$

where v_{0k} follows an MA(q) process with q<k, which would imply that Y_{0k} and Y_{00} are uncorrelated conditional on X. The matching method does not require that everything in the information set be known, but it does assume sufficient information to make the selection on observables assumption plausible.

2 Cross-sectional matching methods

For notational simplicity, let P=P(Z). A prototypical propensity score matching estimator takes the form

$$\hat{\alpha}_M = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} [Y_{1i} - \hat{E}(Y_{0i}|D = 1, P_i)] \quad (5)$$

$$\hat{E}(Y_{0i}|D = 1, P_i) = \sum_{j \in I_0} W(i, j) Y_{0j},$$

where I_1 denotes the set of programme participants, I_0 the set of non-participants, S_P the region of common support (see below for ways of constructing this set). n_1 is the number of persons in the set $I_1 \cap S_P$. The match for each participant $i \in I_1 \cap S_P$ is constructed as a weighted average over the outcomes of non-participants, where the weights $W(i, j)$ depend on the distance between P_i and P_j. Define a neighbourhood $C(P_i)$ for each i in the participant sample. Neighbours for i are non-participants $j \in I_0$ for whom $P_j \in C(P_i)$. The persons matched to i are those people in set A_i where $A_i = \{ j \in I_0 | P_j \in C(P_i) \}$. We describe a number of alternative matching estimators below, that differ in how the neighbourhood is defined and in how the weights $W(i, j)$ are constructed.

2.1 Alternative ways of constructing matched outcomes

2.1.1 Nearest-neighbour matching

Traditional, pairwise matching, also called *nearest-neighbour matching*, sets:

$$C(P_i) = \min_j \|P_i - P_j\|, \quad j \in I_0.$$

That is, the non-participant with the value of P_j that is closest to P_i is selected as the match and A_i is a singleton set. The estimator can be implemented either matching with or without replacement. When matching is performed with replacement, the same comparison group observation can be used repeatedly as a match. A drawback of matching without replacement is that the final estimate will usually depend on the initial ordering of the treated observations for which the matches were selected.

Caliper matching (Cochran and Rubin, 1973) is a variation of nearest neighbour matching that attempts to avoid ‘bad’ matches (those for which P_j is far from P_i) by imposing a tolerance on the maximum distance $\|P_i - P_j\|$ allowed. That is, a match for person i is selected only if $\|P_i - P_j\| < \varepsilon, \quad j \in I_0$, where ε is a pre-specified tolerance. Treated persons for whom no matches can be found within the caliper are excluded from the analysis, which is one way of imposing a common support condition. A drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

2.1.2 Stratification or interval matching

In this variant of matching, the common support of P is partitioned into a set of intervals, and average treatment impacts are calculating through simple averaging within each interval. A weighted average of the interval impact estimates, using the fraction of the D=1 population in each interval for the weights, provides an overall average impact estimate. Implementing this method requires a decision on how wide the intervals should be. Dehejia and Wahba (1999) implement interval matching using intervals that are selected such that the mean values of the estimated P_i and P_j are not statistically different from each other within intervals.

2.1.3 Kernel and local linear matching

More recently developed matching estimators construct a match for each programme participant using a weighted average over multiple persons in the comparison group. Consider, for example, the nonparametric *kernel matching estimator*, given by

$$\widehat{\alpha}_{KM} = \frac{1}{n_1}\sum_{i\in I_1}\left\{Y_{1i} - \frac{\sum_{j\in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k\in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}\right\}$$

where $G(\cdot)$ is a kernel function and a_n is a bandwidth parameter. (See Heckman, Ichimura and Todd, 1997; 1998; and Heckman et al., 1998.) In terms of eq. (5), the weighting function, $W(i,j)$, is equal to

$$\frac{G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k\in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}.$$

For a kernel function bounded between –1 and 1, the neighbourhood is

$$C(P_i) = \left\{\left|\frac{P_i - P_j}{a_n}\right| \le 1\right\}, \quad j\in I_0.$$

Under standard conditions on the bandwidth and kernel,

$$\frac{\Sigma_{j\in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\Sigma_{k\in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$$

is a consistent estimator of $E(Y_0|D = 1, P_i)$. (Specifically, we require that $G(\cdot)$ integrates to one, has mean zero and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$. One example of a kernel function is the quartic kernel, given by $G(s) = \frac{15}{16}(s^2 - 1)^2$ if $|s| < 1$, $G(s) = 0$ otherwise.)

Heckman, Ichimura and Todd (1997) also propose a generalized version of kernel matching, called local linear matching. Recent research by Fan, 1992a; 1992b, demonstrated advantages of local linear estimation over more standard kernel estimation methods. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities; see Fan, 1992a; 1992b.) The local linear weighting function is given by

$$W(i, j) = \frac{G_{ij}\Sigma_{k\in I_0} G_{ik}(P_k - P_i)^2 - [G_{ij}(P_j - P_i)]\left[\Sigma_{k\in I_0} G_{ik}(P_k - P_i)\right]}{\Sigma_{j\in I_0} G_{ij}\Sigma_{k\in I_0} G_{ij}(P_k - P_i)^2 - \left(\Sigma_{k\in I_0} G_{ik}(P_k - P_i)\right)^2}. \quad (6)$$

As demonstrated in research by Fan (1992a; 1992b), local linear estimation has some advantages over standard kernel estimation. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities (see Fan, 1992a; 1992b). Thus, local linear regression would be expected to perform better than kernel estimation in cases where the non-participant observations on *P* fall on one side of the participant observations.

To implement the matching estimator given by eq. (5), the region of common support S_P needs to be determined. The common support region can be estimated by

$$\hat{S}_P = \left\{P: \hat{f}(P|D = 1) > 0 \text{ and } \hat{f}(P|D = 0) > c_q\right\},$$

where $\hat{f}(P|D = d)$, $d \in \{0, 1\}$ are standard nonparametric density estimators. To ensure that the densities are strictly greater than zero, it is required that the densities be strictly positive (that is, exceed zero by a certain amount), determined using a ‘trimming level’ *q*. That is, after excluding any *P* points for which the estimated density is zero, an additional small percentage of the remaining *P* points is excluded for which the estimated density is positive but very low. The set of eligible matches is thus given by

$$\hat{S}_q = \left\{P \in \hat{S}_P : \hat{f}(P|D = 1) > c_q \text{ and } \hat{f}(P|D = 0) > c_q\right\},$$

where c_q is the density cut-off level that satisfies

$$\sup_{c_q} \frac{1}{2J} \sum_{\{i \in I_1 \cap \hat{S}_P\}} \left\{1(\hat{f}(P|D = 1)) < c_q + 1(1(\hat{f}(P|D = 0)) < c_q)\right\} \leq q.$$

Here, J is the cardinality of the set of observed values of P that lie in $I_1 \cap \hat{S}_P$. That is, matches are constructed only for the programme participants for which the propensity scores lie in $\hat{S}_q$.

The above estimators are representations of matching estimators and are commonly used. They can be easily adapted to estimate other parameters of interest, such as the average effect of treatment on the untreated $(\mathrm{UT} = E(Y_1 - Y_0|D = 0, X))$, or the average treatment effect $(\mathrm{ATE} = E(Y_1 - Y_0|X))$, which is just a weighted average of treatment on the treated (TT) and treatment on the untreated (UT).

The recent literature has also developed alternative matching estimators that employ different weighting schemes to increase efficiency. See, for example, Hahn (1998) and Hirano, Imbens and Ridder (2003) for estimators that attain the semiparametric efficiency bound. The methods are not described in detail here, because those studies focus on the ATE and not on the average effect of treatment on the treated (TT) parameter. Heckman, Ichimura and Todd (1998) develop a regression-adjusted version of the matching estimator, which replaces Y_{0j} as the dependent variable with the residual from a regression of Y_{0j} on a vector of exogenous covariates. The estimator uses a Robinson (1988) type estimation approach to incorporate exclusion restrictions: that is, that some of the conditioning variables in an equation for the outcomes do not enter into the participation equation or vice versa. In principle, imposing exclusion restrictions can increase efficiency. In practice, though, researchers have not observed much gain from using the regression-adjusted matching estimator. Some alternatives to propensity score matching are discussed in Diamond and Sekhon (2005).

2.2 When does bias arise in matching?

The success of a matching estimator depends on the availability of observable data to construct the conditioning set Z, such that (1) and (2) are satisfied. Suppose only a subset $Z_0 \subset Z$ of the required variables is observed. The propensity score matching estimator based on Z_0 then converges to

$$\alpha'_M = E_{P(Z_0)|D=1}(E(Y_1|P(Z_0), D = 1) - E(Y_0|P(Z_0), D = 0)). \quad (7)$$

The bias for the parameter of interest, $E(Y_1 - Y_0 | D = 1)$, is

$$\text{bias}_M = E(Y_0 | D = 1) - E_{P(Z_0)|D=1}\{E(Y_0 | P(Z_0), D = 0)\}.$$

There is no way of a priori choosing the set of Z variables to satisfy the matching condition or of testing whether a particular set meets the requirements. In rare cases, where data are available on a randomized social experiment, it is sometimes possible to ascertain the bias (see, for example, Heckman, Ichimura, and Todd, 1997; Dehejia and Wahba, 1999; 2002; Smith and Todd, 2005).

3 Difference-in-difference matching estimators

The estimators described above assume that, after conditioning on a set of observable characteristics, outcomes are conditionally mean independent of programme participation. However, for a variety of reasons there may be systematic differences between participant and non-participant outcomes, even after conditioning on observables, which could lead to a violation of the identification conditions required for matching. Such differences may arise, for example, because of programme selectivity on unmeasured characteristics or because of levels differences in outcomes that might arise when participants and non-participants reside in different local labour markets or if the survey questionnaires used to gather the data differ in some ways across groups. A difference-in-differences (DID) matching strategy, as defined in Heckman, Ichimura and Todd (1997) and Heckman et al. (1998), allows for temporally invariant differences in outcomes between participants and non-participants. This type of estimator matches on the basis of differences in outcomes using the same weighting functions described above. The propensity score DID matching estimator requires that

$$E(Y_{0t} - Y_{0t'} | P, D = 1) = E(Y_{0t} - Y_{0t'} | P, D = 0),$$

where t and t' are time periods after and before the programme enrolment date. This estimator also requires the support condition given above, which must now hold in both periods t and t'. The local linear difference-in-difference estimator is given by

$$\widehat{\alpha}_{DM} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_0 \cap S_P} W(i, j)(Y_{0tj} - Y_{0t'j}) \right\},$$

where the weights correspond to the local linear weights defined above. If repeated cross-section data are available, instead of longitudinal data, the estimator can be implemented as

$$\widehat{\alpha}_{DM} = \frac{1}{n_{1t}} \sum_{i \in I_{1t} \cap S_p} \left\{ \left(Y_{1ti} - \sum_{j \in I_{0t} \cap S_p} W(i, j) Y_{0tj} \right) \right\} - \frac{1}{n_{1t'}} \sum_{i \in I_{1t'} \cap S_p} \left\{ \left(Y_{1t'i} - \sum_{j \in I_{0t'}} W(i, j) Y_{0t'j} \right) \right\},$$

where $I_{1t}, I_{1t'}, I_{0t}, I_{0t'}$ denote the treatment and comparison group data-sets in each time period.
Finally, the DID matching estimator allows selection into the programme to be based on anticipated gains from the programme in the sense that D can help predict the value of Y_1 given P. However, the method assumes that D does not help predict changes $Y_{0t}–Y_{0t'}$ conditional on a set of observables (Z) used in estimating the propensity score. In their analysis of the effectiveness of matching estimators, Smith and Todd (2005) found difference-in-difference matching estimators to perform much better than cross-sectional methods in cases where participants and non-participants were drawn from different regional labour markets and/or were given different survey questionnaires.

4 Matching when the data are choice-based sampled

The samples used in evaluating the impacts of programmes are often choice-based, with programme participants oversampled relative to their frequency in the population of persons eligible for the programme. Under choice-based sampling, weights are generally required to consistently estimate the probabilities of programme participation. (See, for example, Manski and Lerman, 1977, for discussion of weighting for logistic regressions.) When the weights are unknown, Heckman and Todd (1995) show that with a slight modification matching methods can still be applied, because the odds ratio ($P/(1–P)$) estimated using a logistic model with incorrect weights (that is, ignoring the fact that samples are choice-based) is a scalar multiple of the true odds ratio, which is itself a monotonic transformation of the propensity scores. Therefore, matching can proceed on the (misweighted) estimate of the odds ratio (or on the log odds ratio).

5 Using balancing tests to check the specification of the propensity score model

As described earlier, the propensity score matching estimator requires the outcome variable to be mean independent of the treatment indicator conditional on the propensity score, $P(Z)$. An important consideration in implementation is how to choose Z. Unfortunately, there is no theoretical basis for choosing a particular set Z to satisfy the identifying assumptions, and the set is not necessarily the most inclusive one.
To guide in the selection of Z, there is some accumulated empirical evidence on how bias estimates depended on the choice of Z in particular applications. For example, Heckman et al. (1998), Heckman, Ichimura and Todd (1997) and Lechner (2001) show that the choice of variables included in Z can make a substantial difference to the estimator's performance. These papers found that biases tended to be higher when the participation equation was estimated using a cruder set of conditioning variables. One approach adopted is to select the set Z to maximize the percentage of people correctly classified under the model. Another finding in these papers is that the matching estimators performed best when the treatment and control groups were located in the same geographic area and when the same survey instrument was administered to both treatments and controls to ensure comparable measurement of outcomes.
Rosenbaum and Rubin (1983) suggest a method to aid in the specification of the propensity score model. The

method does not provide guidance in choosing which variables to include in *Z*, but can help to determine which interactions and higher-order terms to include in the model for a given *Z* set. They note that for the true propensity score, the following holds:

$$Z \perp \perp D|\Pr(D = 1|Z),$$

or equivalently $E(D|Z, \Pr(D = 1|Z)) = E(D|\Pr(D = 1|Z))$. The basic intuition is that, after conditioning on $\Pr(D = 1|Z)$, additional conditioning on *Z* should not provide new information about *D*. If after conditioning on the estimated values of $P(D=1|Z)$ there is still dependence on *Z*, this suggests misspecification in the model used to estimate $\Pr(D=1|Z)$. The theorem holds for any *Z*, including sets *Z* that do not satisfy the conditional independence condition required to justify matching. As such, the theorem is not informative about what set of variables to include in *Z*.

This result motivates a specification test for $\Pr(D=1|Z)$, that is a test whether or not there are differences in *Z* between the $D=1$ and $D=0$ groups after conditioning on $P(Z)$. The test has been implemented in the literature a number of ways (see, for example Eichler and Lechner, 2002; Dehijia and Wahba, 1999; 2002; Smith and Todd, 2002; Diamond and Sekohn, 2005).

6 Assessing the variability of matching estimators

The distribution theory for the cross-sectional and difference-in-difference kernel and local linear matching estimators described above is derived in Heckman, Ichimura and Todd (1998). However, implementing the asymptotic standard error formulae can be cumbersome, so standard errors for matching estimators are often instead generating using bootstrap resampling methods. (See Efron and Tibshirani, 1993, for an introduction to bootstrap methods, and Horowitz, 2003, for a recent survey of bootstrapping in econometrics.) A recent paper by Abadie and Imbens (2006a) shows that standard bootstrap resampling methods are not valid for assessing the variability of nearest neighbour estimators, but can be applied to assess the variability of kernel or local linear matching estimators for a suitably chosen bandwidth. Abadie and Imbens (2006b) present alternative standard error formulae for assessing the variability of nearest neighbour matching estimators.

7 Applications

There have been numerous evaluations of matching estimators in recent decades. For a survey of many applications in the context of evaluating the effects of labour market programmes (see Heckman, Lalonde and Smith, 1999). More recently, propensity score matching estimators have been used in evaluating the impacts of a variety of programme interventions in developing countries. Jalan and Ravallion (1999) assess the impact of a workfare programme in Argentina (the *Trabajar* programme), and Jalan and Ravallion (2003) study the effects of public investments in piped water on child health outcomes in rural India. Galiani, Gertler and Schargrodsky (2005) use difference-in-difference matching methods to analyse the effects of privatization of water services on child mortality in Argentina. Other applications include Gertler, Levine and Ames (2004) in a study of the effects of parental death on child outcomes, Lavy (2004) in a study of the effects of a teacher incentive programme in Israel on student performance, Angrist and Lavy (2001) in a study of the effects of teacher

training on children's test scores in Israel, and Chen and Ravallion (2003) in a study of a poverty reduction project in China.
Behrman, Cheng and Todd (2004) use a modified version of a propensity score matching estimator to evaluate the effects of a preschool programme in Bolivia on child health and cognitive outcomes. They identify programme effects by comparing children with different lengths of duration in the programme, using matching to control for selectivity into alternative durations. Also, see Imbens (2000) and Hirano and Imbens (2004) for an analysis of the role of the propensity score with continuous treatments. Lechner (2001) extends propensity score analysis for the case of multiple treatments.

See Also

- propensity score
- selection bias and self-selection
- semiparametric estimation
- treatment effect

Bibliography

Abadie, A. and Imbens, G. 2006a. On the failure of the bootstrap for matching estimators. Technical Working Paper No. 325. Cambridge, MA: NBER.

Abadie, A. and Imbens, G. 2006b. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74, 235–67.

Angrist, J. and Lavy, V. 2001. Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics* 19, 343–69.

Behrman, J., Cheng, Y. and Todd, P. 2004. Evaluating preschool programs when length of exposure to the program varies: a nonparametric approach. *Review of Economics and Statistics* 86, 108–32.

Chen, S. and Ravallion, M. 2003. Hidden impact? Ex-post evaluation of an anti-poverty program. Policy Research Working Paper No. 3049. Washington, DC: World Bank.

Cochran, W. and Rubin, D. 1973. Controlling bias in observational studies. *Sankyha* 35, 417–46.

Dehejia, R. and Wahba, S. 1999. Causal effects in non-experimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 1053–62.

Dehejia, R. and Wahba, S. 2002. Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84, 151–61.

Diamond, A. and Sekhon, J.S. 2005. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Working paper, Department of Political

Science, Berkeley.

Efron, B. and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Eichler, M. and Lechner, M. 2002. An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics* 9, 143–86.

Fan, J. 1992a. Design adaptive nonparametric regression. *Journal of the American Statistical Association* 87, 998–1004.

Fan, J. 1992b. Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics* 21, 196–216.

Fisher, R.A. 1935. *Design of Experiments*. New York: Hafner.

Friedlander, D. and Robins, P. 1995. Evaluating program evaluations: new evidence on commonly used nonexperimental methods. *American Economic Review* 85, 923–37.

Galiani, S., Gertler, P. and Schargrodsky, E. 2005. Water for life: the impact of the privatization of water services on child mortality in Argentina. *Journal of Political Economy* 113, 83–120.

Gertler, P., Levine, D. and Ames, M. 2004. Schooling and parental death. *Review of Economics and Statistics* 86, 211–25.

Hahn, J. 1998. On the role of the propensity score in efficient estimation of average treatment effects. *Econometrica* 66, 315–31.

Heckman, J., Ichimura, H. and Todd, P. 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Review of Economic Studies* 64, 605–54.

Heckman, J., Ichimura, H. and Todd, P. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65, 261–94.

Heckman, J., Lalonde, R. and Smith, J. 1999. The economics and econometrics of active labor market programs. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Heckman, J., Smith, J. and Clements, N. 1997. Making the most out of social experiments: accounting for heterogeneity in programme impacts. *Review of Economic Studies* 64, 487–536.

Heckman, J. and Todd, P. 1995. Adapting propensity score matching and selection models to choice-based samples. Manuscript, Department of Economics, University of Chicago.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. 1996. Sources of selection bias in evaluating social

programs: an interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proceedings of the National Academy of Sciences* 93, 13416–20.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. 1998. Characterizing selection bias using experimental data. *Econometrica* 66, 1017–98.

Hirano, K., Imbens, G. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–89.

Hirano, K. and Imbens, G. 2004. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives*, ed. A. Gelman and X.L. Meng. New York: Wiley.

Holland, P.W. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81, 945–70.

Horowitz, J.L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–32.

Horowitz, J.L. 2003. The bootstrap. *Handbook of Econometrics*, vol. 5, ed. J.J. Heckman and E.E. Leamer. Amsterdam: North-Holland.

Ichimura, H. 1993. Semiparametric least squares and weighted SLS estimation of single index models. *Journal of Econometrics* 58, 71–120.

Imbens, G. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 706–10.

Jalan, J. and Ravallion, M. 1999. Efficient estimation of average treatment effects: evidence for Argentina's Trabajar program. Policy Research Working Paper. Washington, DC: World Bank.

Jalan, J. and Ravallion, M. 2003. Does piped water reduce diarrhea for children in rural India. *Journal of Econometrics* 112, 153–73.

Klein, R.W. and Spady, R.H. 1993. An efficient semiparametric estimator for binary response models. *Econometrica* 61, 387–422.

LaLonde, R. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76, 604–20.

Lavy, V. 2002. Evaluating the effects of teachers’ group performance incentives on pupil achievement. *Journal of Political Economics* 110, 1286–387.

Lavy, V. 2004. Performance pay and teachers’ effort, productivity and grading ethics. Working Paper No.

10622. Cambridge, MA: NBER.

Lechner, M. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. M. Lechner and F. Pfeiffer. Heidelberg: Physica.

Manski, C. 1973. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C. and Lerman, S. 1977. The estimation of choice probabilities from choice-based samples. *Econometrica* 45, 1977–88.

Robinson, P. 1988. Root-*N* consistent nonparametric regression. *Econometrica* 56, 931–54.

Rosenbaum, P. and Rubin, D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Rosenbaum, P. and Rubin, D. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 39, 33–8.

Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Smith, J. and Todd, P. 2005. Does matching overcome Lalonde's critique of nonexperimental estimators? *Journal of Econometrics* 125, 305–53.

How to cite this article

Todd, Petra E. "matching estimators." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000365> doi:10.1057/9780230226203.1059

The New Palgrave Dictionary of Economics Online

matching

Giuseppe Moscarini

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Matching (or job-matching) is the process whereby a firm and a worker meet, learn whether their characteristics combine productively and, in light of this information, sequentially contract a wage and decide whether to separate or to continue production. This hypothesis implies that wages rise and the risk of separation declines with seniority, wage changes are unpredictable and have declining variability, and valuable specific human capital is accumulated in the form of knowledge about the quality of the match. These and other observable implications have found strong support in available empirical evidence, and make job-matching a central theory of worker turnover.

Keywords

Bellman equation; job-matching hypothesis; labour-market contracts; matching; matching markets; returns to tenure; Roy model; selection; wage distribution; worker turnover

Article

Matching (or job-matching) is the process whereby a firm and a worker meet, learn whether their characteristics combine productively and, in light of this information, sequentially contract a wage and decide whether to separate or to continue production.

In many respects, a job is like a marriage. Two parties (a firm and a worker) engage in a long-run relationship, whose success depends on a myriad of factors, all quite difficult to describe. Only the actual outcome of the match can reveal the underlying ‘fit’. If the match works, it continues; otherwise it is scrapped and the partners try their luck elsewhere.

Jovanovic (1979a) formalizes the job-matching hypothesis in a dynamic, rational-expectations context. This hypothesis hinges on two pivotal ideas: learning and selection. The emphasis on selection follows the tradition of equilibrium sorting in labour markets going back to the static Roy model (Roy, 1951). Now, dynamics and imperfect information take centre stage. A job is viewed as an ‘inspection’ as well

as an ‘experience’ good. The worker and the firm have to ‘taste’ the match to decide its value, just like two people first date (to ‘inspect’ the match) then possibly get married (to ‘experience’ the match), with varying degrees of success. Unlike in marriage markets, utility is typically transferable through the wage. The fit between firm and worker characteristics is modelled as a match-specific productivity component, a parameter of the output process, summarizing how well the innumerable relevant characteristics of the worker and of the task actually dovetail. Random noise in production creates a signal extraction problem. The firm and/or the worker continuously observe the output performance of the match, incorporate this information in wages, and reassess it against alternative opportunities offered by the market.

A job-matching model

Output y_t is produced at time $t = 1, 2 \ldots$ by a firm and a worker with a 1:1 Leontief technology:

$$y_t = \theta + \varepsilon_t.$$

There is no hours or effort choice. θ is average productivity or ‘match quality’, drawn by nature, unobserved by firm and worker, at the beginning of the match from $\theta : N(m_{-1}, 1/h_{-1})$, which are also parties’ prior beliefs. $\varepsilon_t : N(0, 1/p_\varepsilon)$ is white noise, i.i.d. and independent of θ. Therefore, risk-neutral firm and worker are interested in the permanent component θ. Following the bulk of the literature, assume that firm and worker are symmetrically informed. This is not a crucial assumption: all that matters is that *some* learning drives match selection.
Upon matching at time 0, parties inspect the match and observe a signal

$$x = \theta + \eta$$

where $\eta : N(0, 1/p_\eta)$ independent of θ. By Bayes’ rule, $\theta | x : N(m_0, 1/h_0)$, where $h_0 = h_{-1} + p_\eta$ and $h_0 m_0 = m_{-1} h_{-1} + x p_\eta$. If the match begins and output is produced at $t = 1, 2 \ldots$, posterior beliefs about match quality conditional on the worker's track record are recursively updated as follows:

$$\theta | x, y_1, y_2, \ldots y_t : N(m_t, 1/h_t) \text{where} h_t = h_{-1} + p_\eta + t p_\varepsilon h_t m_t = h_{t-1} m_{t-1} + p_\varepsilon y_t.$$

That is, m_t and h_t are the mean and precision of the normal posterior distribution of θ, conditional on all information available to date t. After solving backward, m_t is an average of the prior expectation m_{-1},

the initial signal *x* and the history of output $\Sigma_{s=1}^{t} y_s$, weighted by their respective precisions. Given the model's parameters, history and beliefs are summarized by expected productivity m_t and by tenure *t*, which jointly measure the specific human capital accumulated in the relationship.

With no uncertainty and perfect information ($h_{-1} = \infty$ and/or $p_n = \infty$), workers and firms would immediately discard unpromising matches and keep drawing better and better outcomes. With imperfect information, equilibrium behaviour is ‘sequential’ and non-trivial. Equilibrium cannot be perfectly competitive, due to the specificity of match quality and consequently of human capital. Nonetheless, with free entry, no mobility and no capital costs, there is a contracting equilibrium where the wage offered by the firm to the worker equals the worker's expected (marginal) productivity m_t, and firms break even. The worker captures the entire option value of learning. By Bayes's rule, the distribution of the future wage m_{t+1}, unconditional on unknown match quality θ but conditional on current beliefs $\{m_t, t\}$, is normal with

$$\mathrm{E}\left[m_{t+1} | m_t, t\right] = m_t \text{ and } Var\left[m_{t+1} | m_t, t\right] = \frac{p_\varepsilon}{\left[h_{-1} + p_n + (t+1)p_\varepsilon\right](h_{-1} + p_n + t p_\varepsilon)} \quad (1)$$

The worker's value of employment solves the Bellman equation

$$V(m_t, t) = \max\left\{\beta \mathrm{E}\left[V(\tilde{m}_0, 0)\right], m_t + \beta \mathrm{E}\left[V(m_{t+1}, t+1) | m_t, t\right]\right\} \quad (2)$$

for some discount factor $\beta \in [0, 1]$. At each point in time, including $t = 0$ right after observing the initial signal *x* and before starting production, the worker decides whether to quit this match at once and to inspect another one next period (expected value $E[V(\tilde{m}_0, 0)]$, independent of $\{m_t, t\}$ because θ is match-specific) or to accept the wage m_t, produce, observe the output realization y_t, update beliefs to $\{m_{t+1}, t+1\}$, and decide again.

The worker's employment value $V(m, t)$ is increasing in expected match quality *m* and decreasing in tenure *t*. The first effect is obvious. Formally, an increase in m_t raises the right-hand side of (2) directly and, by (1), the normal distribution of future wages in a first-order stochastic dominance sense. Standard dynamic programming arguments establish monotonicity of *V*. To see why the value *V* is also decreasing in tenure *t*, consider the following thought experiment. Before deciding whether to quit or to produce y_{t+1}, the worker is provided with a free signal υ which has the same distribution as y_{t+1}, and is then informative about match quality. After observing this signal, the worker cannot do worse, because he or she can always ignore it. So, before observing υ, she must value this additional information:

$$E_v[E[V(m_{t+1}, t+1)|v, m_t, t]] \geq V(m_t, t) = V(E_v[E[m_{t+1}|m_t, t, v]], t)$$

where the equality follows from $v: V_{t+1}$ and then $E_v[E[m_{t+1}|m_t, t, v]] = E[m_{t+1}|m_t, t] = m_t$. The inequality implies that *V* is convex in m. Since tenure t reduces the variance of *m* from (1), it follows from Jensen's inequality that *V*(*m*,*t*) declines in *t* for given *m*. Intuitively, a match of equally expected but more uncertain productivity is more valuable: there is some chance it will turn out to be great, otherwise it can always be scrapped.

Testable implications and empirical evidence

The key implications of the model derive from selection and learning, and those implications that are testable have indeed found strong empirical support.

Selection

Given the properties of *V*, the worker quits as soon as the wage falls short of a reservation wage, which is increasing in tenure. As the option value of learning is consumed, a given expected match quality is no longer sufficient to support the match. Reservation wages are not directly observable, but the resulting selection does have indirect, testable implications. Only promising matches survive, so the average m_t (wage) in continuing jobs increases (cross-sectionally) with tenure *t*. Indeed, seniority has modest but consistently positive wage returns (Altonji and Williams, 2005). As better matches are less likely to end, the hazard rate of separation, after an initial ‘discovery’ phase, declines with tenure, a very robust stylized fact (Farber, 1994). Finally, censoring bad matches skews the distribution of wage residuals, conditional on observable worker and firm exogenous characteristics: a symmetric and thin-tailed Gaussian distribution of output turns into a distribution of ‘unexplained’ wages with a thick Pareto upper tail (Moscarini, 2005), as in a typical empirical wage distribution.

Learning

From (1), unconditional on the unobserved quality of the match, the wage m_t is a martingale, with variance of innovations declining with tenure *t*. Beliefs updated in a Bayesian fashion cannot be expected to drift in any direction, for the same reason that asset prices are a random walk in efficient financial markets. Thus, unconditionally on tenure, within-job wage changes are uncorrelated and, as uncertainty about match quality is resolved, have declining variance (Mortensen, 1988). Wage growth slows down over the course of a career. Indeed, the search for serial correlation in wage changes has been inconclusive, but the slowdown of wage growth is prevalent (Topel and Ward, 1992). The wages of a cohort of workers ‘fan out’, as some workers are luckier than others and find earlier a good match that pays a high wage, and as commonly observed empirically. When a match separates due to an exogenous

layoff (not modelled here, but easy to accommodate), the worker loses the entire match-specific human capital, so she suffers a persistent wage loss. This fully agrees with the available evidence (Jacobson, Lalonde and Sullivan, 1993). More problematic is the prediction (Mortensen, 1988) that, as $V(m,t)$ falls with t, separation rises with tenure given the wage: empirical evidence (Topel and Ward, 1992) suggests the opposite.

Alternative hypotheses about worker turnover

In light of its intuitive appeal and empirical success, job-matching has become the benchmark model of worker turnover. It has in part inspired the canonical search-and-matching model of the labour market (Mortensen and Pissarides, 1994), where *ex post* idiosyncratic uncertainty drives job flows while search frictions account for involuntary unemployment. But, despite its vast influence, the job-matching approach still faces alternative and competing views of worker turnover, which provide conceptually quite different explanations for the same set of stylized facts. The starker contrast is with pure search models, which may dispose of heterogeneity altogether. In the search literature, wage dispersion and dynamics originate from firms' power of monopsony and commitment to contracts, due to purely strategic considerations. Retention concerns and counter-offers (Burdett and Mortensen, 1998; Burdett and Coles, 2003) explain returns to seniority, declining separations rate and so forth. Closer to the job-matching approach is a class of models that retain heterogeneity and selection, but allow for the quality of the job to change physically over time, while in the job-matching model everything is predetermined, and parties only have to learn their fate. Notable examples are firm-specific training (Jovanovic, 1979b) and learning-by-doing, as well as stochastic match-specific productivity shocks (Mortensen and Pissarides, 1994). In these models, general properties of Bayesian learning, like the declining variance of innovations, must be assumed as ad hoc properties of the productivity process. Nonetheless, this lack of identification poses a formidable challenge, and motivates an ongoing research effort.

See Also

- assortative matching
- bandit problems
- learning in macroeconomics
- matching and market design
- Roy model
- search theory
- selection bias and self-selection
- sequential analysis

Bibliography

Altonji, J. and Williams, N. 2005. Do wages rise with job seniority? A reassessment. *Industrial and*

Labor Relations Review 58, 370–97.

Burdett, K. and Coles, M. 2003. Equilibrium wage-tenure contracts. *Econometrica* 71, 1377–404.

Burdett, K. and Mortensen, D. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39, 257–73.

Farber, H. 1994. The analysis of interfirm worker mobility. *Journal of Labor Economics* 12, 554–93.

Jacobson, L., Lalonde, R. and Sullivan, D. 1993, Earnings losses of displaced workers. *American Economic Review* 83, 685–709.

Jovanovic, B. 1979a. Job matching and the theory of turnover. *Journal of Political Economy* 87, 972–90.

Jovanovic, B. 1979b. Firm-specific capital and turnover. *Journal of Political Economy* 87, 1246–60.

Mortensen, D. 1988. Wages, separations, and job tenure: on-the-job specific training or matching? *Journal of Labor Economics* 6, 445–71.

Mortensen, D. and Pissarides, C. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61, 397–415.

Moscarini, G. 2005. Job matching and the wage distribution. *Econometrica* 73, 481–516.

Roy, A. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.

Topel, R. and Ward, M. 1992. Job mobility and the careers of young men. *Quarterly Journal of Economics* 107, 439–79.

How to cite this article

Moscarini, Giuseppe. "matching." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000104> doi:10.1057/9780230226203.1057

The New Palgrave Dictionary of Economics Online

material balances

Gregory Grossman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

command economy; material balances; planning

Article

A material balance is a simple planning device developed (if not originated) early in Soviet planning for the purpose of equating prospective availabilities of a given good and its prospective requirements over the plan period (or at some target date in case of a stock). It occupies a central role in Soviet-type planning. The phrase, a literal rendering of the Russian *material'nyi balans*, is somewhat inexact and possibly confusing inasmuch as each of the two words has a variety of meanings in English. A more exact term would be ‘sources-and-uses account’ for a flow or ‘balance sheet’ for a stock. As such, material balances have counterparts in planning and management the world over.

In Soviet-type planning, a material balance is typically constructed *ex ante*. It can pertain to any good or resource requiring planners’ attention or administrative disposition; thus, ‘balance’ is drawn up not only for material products, but also for labour, capacity, foreign exchange, and so on. While it can be drawn up at any level of the hierarchy of a command economy and by any relevant organizational entity, these alternatives carry important economic, bureaucratic and even political implications in a Soviet-type economy. ‘In the course of preparing the annual plan•…•the USSR State Planning Commission draws up [some] 2,000 single-product balances, the State Commission for Supply – up to 15,000, and the ministries – up to 50,000’ (*EKO*, August, 1983, p. 26). Though there may be some duplication in terms of goods between these figures, they nonetheless do suggest the magnitude of the annual task, especially if one bears in mind the interconnections.

In Soviet-type practice a material balance not only has the passive purpose of checking requirements against availabilities, but forms the operational basis for specific production or import directives to designated organizations and firms, and for specific acquisition permits to designated users of the good. Note that nearly all producer goods are administratively allocated (rationed) to users.

A material balance may take the following form (adapted from Levine, 1959):

Table 1

Material balance for good X for (year)

Sources	*Uses (distribution)*
1. Current production – by major producing organizations, firms	1. For production – by organizations, firms
2. Imports	2. For construction – by organizations, firms
3. Other sources	3. For household sector (‘market fund’)
4. Beginning-year stocks – by organizations	4. For export
5. Total sources	5. To central reserve stocks
	6. End-year stocks at suppliers – by organizations, firms
	7. Total uses (distribution)

Two kinds of questions arise: (*a*) operational – how is the balance initially compiled and ‘balanced’, and later adjusted for outside effects (from other balances) and the extent to which successive iterations are required to converge? and (*b*) policy – the bounds and degree of aggregation of a ‘good’, the organizational locus and level of compilation, and so on?

Little is known about the initial compilation. There must be serious problems of the requisite detailed information in the case of many goods, given that the preparation of the annual plan extends over most of the pre-plan year (and often into the plan year). Thus, the database may anticipate the plan year by one-and-a-half to two years whose projection is obviously subject to uncertainty. A common problem is the uncertainty of going-on-stream of capacity under construction. Also, the data may not be very accurate to start with, given the cat-and-mouse game that firms and other subordinates play with their superiors. What is more, thousands of balances are being drawn up simultaneously, often by different organizations or subdivisions, with the obvious difficulty of mutual coordination.

The ‘balancer’ must take into account – in addition to technical parameters – political and other high-level decisions, existing economic programmes, bureaucratic politics, and the usual pressure to squeeze more out of the economy's resources. Corruption is not unknown. The work is largely done manually and inevitably to some extent subjectively. While computers are beginning to be used, the input–output technique – which in principle is eminently suitable for the purpose – seems to be applied for the grosser computations and checks, not for the drawing up of operational, short-term material balances. The main reasons are that the sectors in even the largest matrices are too aggregative for the material balances, and the data underlying the technical coefficients are not current enough.

Among the balancer's technical parameters, pride of place is occupied by the ‘norm’ – a disaggregated input–output ratio, which assists the compiler in filling in parts of both sides of the account. Much effort goes into computation of the norms, given their crucial role in the preparation of plans and the issuing of specific assignments. They are supposed to be ‘scientific’, that is, representing the best applicable engineering practice (note: for technical rather than economic efficiency), but given their enormous number and informational problems, this remains an ideal. In the event, the balancer must employ short-cuts and resort to optimistic assumptions in order to achieve equality of requirements and availabilities while under pressure to deliver high (‘taut’) production targets. A common and much criticized short-cut is simply to raise output targets of all producers by a uniform percentage, with corresponding adjustments of the norms.

The weakest link in the material balance method is coordination among the many balances to achieve a reasonably internally consistent plan for the whole economy or a sector thereof. (Montias, 1959, discusses this at length.) Even if the implicit inter-industry matrix is close to triangular, every iteration is a major undertaking under the actual conditions. Aggregating the goods would simplify the iteration process, but would not suit well the demands posed by detailed production assignments and allocation orders. So would the holding of ample reserve stocks, which are not always there or accessible. In fact, adjustments and corrections tend ordinarily to be carried to only a few adjoining balances.
The overall annual plan that emerges is typically of low internal consistency (not to say, economic efficiency), causing considerable difficulties to those charged with its implementation and necessitating continual further correction and adjustment during the plan year, with the same effect.

See Also

- command economy
- economic calculation in socialist countries
- planning

Bibliography

Levine, H.S. 1959. The centralized planning of supply in Soviet industry. In U.S. Congress, Joint Economic Committee, *Comparison of the United States and Soviet Economies* I. Washington, DC.

Montias, J.M. 1959. Planning with material balances in Soviet-type economies. *American Economic Review* 49, 963–85.

How to cite this article

Grossman, Gregory. "material balances." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000105> doi:10.1057/9780230226203.1060

The New Palgrave Dictionary of Economics Online

mathematical economics

Gerard Debreu
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A summary of the emergence and triumph of mathematical economics. The modern phase was deeply influenced by John von Neumann's article of 1928 on games and his paper of 1937 on economic growth. His 1944 *Theory of Games and Economic Behavior*, coauthored by Oskar Morgenstern, went beyond differential calculus and linear algebra and paved the way for the axiomatization of economic theory. This has enabled researchers to use precisely stated and flawlessly proved results, in the quest for the most direct link between the assumptions and the conclusions of a theorem. Economic theory is fated for a long mathematical future.

Keywords

activity analysis; Allais, M.; Samuelson, P.; Anderson, R.; Arrow, K.; asymptotic equality; Aumann, R.; axiomatized theory; Brouwer's fixed point theorem; Brown, D.; commodity space; competitive equilibrium; contract curve; convexity; cores; Cournot, A.; Debreu, G.; differential calculus; econometrics; Edgeworth, F.; existence of general equilibrium; first theorem of welfare economics; fixed point theorems; general equilibrium; geometry; Hicks, J.; hyperplanes; implicit prices; Kakutani's fixed point theorem; Koopmans, T.; Lebesgue measure; linear programming; Lyapunov's theorem; mathematical economics; McKenzie, L.; Nash, J.; measure theory; minimax theorem; Morgenstern, O.; non-standard analysis; Pareto, V.; price space; real vector space; Robinson, A.; Sard's theorem; Scarf, H.; set of negligible agents; Shubik, M.; Sonnenschein, H.; uncertainty; uniqueness of equilibrium; Vind, K.; von Neumann, J.; von Neumann's lemma; Wald, A.; Walras, L.; Walras's Law

Article

I. The steady course on which mathematical economics has held for the past four decades sharply contrasts with its progress during the preceding century, which was marked by several major scientific accidents. One of them occurred in 1838, at the beginning of that period, with the publication of Augustin Cournot's *Recherches sur les principes mathématiques de la théorie des richesses*. By its mathematical form and by its economic content, his book stands in splendid isolation in time; and in explaining its data historians of economic analysis in the first half of the 19th century must use a wide confidence interval.

The University of Lausanne was responsible for two other of those accidents. When Léon Walras delivered his first professorial lecture there on 16 December 1870, he had held no previous academic appointment; he had published a novel and a short story but he had not contributed to economic theory before 1870; and he was exactly 36. The risk that his university took was vindicated by the appearance of the *Eléments d'économie politique pure* in 1874–7. For Vilfredo Pareto, who succeeded Walras in his chair in 1893, it was also a first academic appointment; he had not contributed to economic theory before 1892; and he was 45. This second gamble of the University of Lausanne paid off when Pareto's *Cours d'économie politique* appeared in 1896–97, followed by his *Manuel d'économie politique* in 1909, and by the article 'Economie mathématique' in 1911.

In the contemporary period of development of mathematical economics, profoundly influenced by John von Neumann, his article of 1928 on games and his paper of 1937 on economic growth also stand out as major accidents, even in a career with so many facets.

The preceding local views would yield a distorted historical perception, however, if they were not complemented by a global view which sees in the development of mathematical economics a powerful, irresistible current of thought. Deductive reasoning about social phenomena invited the use of mathematics from the first. Among the social sciences,

economics was in a privileged position to respond to that invitation, for two of its central concepts, commodity and price, are quantified in a unique manner, as soon as units of measurement are chosen. Thus for an economy with a finite number of commodities, the action of an economic agent is described by listing his input, or his output, of each commodity. Once a sign convention distinguishing inputs from outputs is made, the action of an agent is represented by a point in the commodity space, a finite-dimensional real vector space. Similarly the prices in the economy are represented by a point in the price space, the real vector space dual of the commodity space. The rich mathematical structure of those two spaces provides an ideal basis for the development of a large part of economic theory.

Finite dimensional commodity and price spaces can be, and usually are, identified and treated as a Euclidean space. The stage is thus set for geometric intuition to take a lead role in economic analysis. That role is manifest in the figures that abound in the economics literature, and some of the great theorists have substituted virtuosity in reasoning on diagrams for the use of mathematical form. As for mathematical economists, geometric insight into the commodity-price space has often provided the key to the solution of problems in economic theory.

The differential calculus and linear algebra were applied to that space at first as a matter of course. By the time John Hicks's *Value and Capital* appeared in 1939, Maurice Allais' *A la recherche d'une discipline économique* in 1943, and Paul Samuelson's *Foundations of Economic Analysis* in 1947, they had both served economic theory well. They would serve it well again, but the publication of the *Theory of Games and Economic Behavior* in 1944 signalled that action was also going to take new directions. In mathematical form, the book of von Neumann and Oskar Morgenstern set a new level of logical rigour for economic reasoning, and it introduced convex analysis in economic theory by its elementary proof of the MiniMax theorem. In the next few years convexity became one of the central mathematical concepts, first in activity analysis and in linear programming, as the *Activity Analysis of Production and Allocation* edited by Tjalling Koopmans attested in 1951, and then in the mainstream of economic theory. In consumption theory as in production theory, in welfare economics as in efficiency analysis, in theory of general economic equilibrium and in the theory of the core, the picture of a convex set supported by a hyperplane kept reappearing, and the supporting hyperplane theorem supplied a standard technique for obtaining implicit prices. The applications of that theorem to economics were a ready consequence of the real vector space structure of the commodity space; yet they were made more than thirty years after Minkowski proved it in 1911.

Algebraic topology entered economic theory in 1937, when von Neumann generalized Brouwer's fixed point theorem in a lemma devised to prove the existence of an optimal growth path in his model. The lag from Brouwer's result of 1911 to its first economic application was shorter than for Minkowski's result. It should, however, have been significantly longer, for von Neumann's lemma was far too powerful a tool for his proof of existence. Several authors later obtained more elementary demonstrations, and David Gale in particular based his in 1956 on the supporting hyperplane theorem. Thus von Neumann's lemma, reformulated in 1941 as Kakutani's fixed point theorem, was an accident within an accidental paper.

But in a global historical view, the perfect fit between the mathematical concept of a fixed point and the social science concept of an equilibrium stands out. A state of a social system is described by listing an action for each one of its agents. Considering such a state, each agent reacts by selecting the action that is optimal for him given the actions of all the others. Listing those reactions yields a new state, and thereby a transformation of the set of states of the social system into itself is defined. A state of the system is an equilibrium if, and only if, it is a fixed point of that transformation. More generally, if the optimal reactions of the agents to a given state are not uniquely determined, one is led to associate a set of new states, instead of a single state, with every state of the system. A point-to-set transformation of the set of states of the social system into itself is thereby defined; and a state of the system is an equilibrium if, and only if, it is a fixed point of that transformation. In this view, fixed point theorems were slated for the prominent part they played in game theory and in the theory of general economic equilibrium after John Nash's one-page note of 1950.

A perfect fit of mathematical form to economic content was also found when the traditional concept of a set of negligible agents was formulated exactly. In 1881, in *Mathematical Psychics*, Francis Edgeworth had studied in his box the asymptotic equality of the 'contract curve' of an economy and of its set of competitive allocations. Basic to his proof of convergence is the fact that in his limiting process every agent tends to become negligible. A long period of neglect of his contribution ended in 1959, when Martin Shubik brought out the connection between the contract curve and the game theoretic concept of the core. After the second impulse given in 1962 by Herbert Scarf's first extension of Edgeworth's result, a new phase of development of the economic theory of the core was under way; and in 1964 Robert Aumann formalized the concept of a set of negligible agents as the unit interval of the real line with its Lebesgue measure. The power of that formulation was demonstrated as Aumann proved that in an exchange economy with that set of agents, the core and the set of competitive allocations coincide. Karl Vind then gave, also in 1964, a different formulation of this remarkable result in the context of a measure space of agents without atoms, and showed that it is a direct consequence of Lyapunov's theorem of 1940 on the range of an atomless vector measure. The convexity of that range explains the convexing effect of large economies. In the important case of a set of negligible agents, it justifies the convexity assumption on aggregate sets to which economic theory frequently appeals. A privileged place was clearly marked for measure theory in mathematical economics.

An alternative formulation of the concept of a set of negligible agents was proposed by Donald Brown and Abraham Robinson in 1972 in terms of Non-standard Analysis, created by Robinson in the early 1960s. Innovations in the mathematical tools of economic theory had not always been immediately and universally adopted in the past. In this case the lag from mathematical discovery to economic application was exceptionally short, and Non-standard Analysis had not been widely accepted by mathematicians themselves. Predictably the intrusion of this strange, sophisticated new tool in economic theory was greeted mostly with indifference or with scepticism. Yet it led to the form given by Robert Anderson to inequalities on the deviation of core allocations from competitive allocations, which are central to the theory of the core. In the article published by Anderson in 1978 those inequalities are stated and proved in an elementary manner, but their expression was found by means of Non-standard Analysis.

The differential calculus, which had been used earlier on too broad a spectrum of economic problems, turned out in the 1970s to supply the proper mathematical machinery for the

study of the set of competitive equilibria of an economy. A partial explanation of the observed state of an economic system had been provided by proofs of existence of equilibrium based on fixed point theorems. A more complete explanation would have followed from persuasive assumptions on a mathematical model of the economy ensuring uniqueness of equilibrium. Unfortunately the assumptions proposed to that end were excessively stringent, and the requirement of global uniqueness had to be relaxed to that of local uniqueness. Even then an economy composed of agents on their best mathematical behaviour (for instance each having a concave utility function and a demand function both indefinitely differentiable) may be ill-behaved and fail to have locally unique equilibria. If one considers the question from the generic viewpoint, however, one sees that the set of those ill-behaved economies is negligible. This time the ideal mathematical tool for the proof of that assertion is Sard's theorem of 1942 on the set of critical values of a differentiable function.

By providing appropriate techniques for the study of the set of equilibria, differential topology and global analysis came to occupy in mathematical economics a place that seemed to have been long reserved for them.

As new fields of mathematics were introduced into economic theory and solved some of its fundamental problems, a growth-generating cycle operated. The mathematical interest of the questions raised by economic theory attracted mathematicians who in turn made the subject mathematically more interesting. The resulting expansion of mathematical economics was unexpectedly rapid. Attempting to quantify it, one can use as an index the total number of pages published yearly by the five main periodicals in the field: *Econometrica* and the *Review of Economic Studies* (which both started publishing in 1933), the *International Economic Review* (1960), the *Journal of Economic Theory* (1969), and the *Journal of Mathematical Economics* (1974). The graph of that index is eloquent. It shows a first phase of decline to 1943, followed by a 33-year period of exuberant, nearly exponential growth. The annual rate of increase that would carry the index exponentially from its 1944 level to its 1977 level is 8.2 per cent, a rate that implies doubling in slightly less than nine years and that cannot easily be sustained. The years 1977–84 have indeed marked a pause that will soon resemble a stagnation phase if it persists. Among its imperfections the index gives equal weights to *Econometrica*, the *Review of Economic Studies*, and the *International Economic Review*, all of which publish articles on econometrics as well as on mathematical economics, and to the *Journal of Economic Theory* and the *Journal of Mathematical Economics*, which do not. But given lower relative weights to the first three yields even higher annual rates of exponential growth of the index for the period 1944–77.

The sweeping movement that took place from 1944 to 1977 suggests an inevitable phase in the evolution of mathematical economics. The graph illustrating that phase hints at the deep transformation of departments of economics during those 33 years. It also hints at the proliferation of discussion papers and at the metamorphosis of professional journals like the *American Economic Review*, which was almost pure of mathematical symbols in 1933 but had lost its innocence by the late 1950s. Figure 1.

Figure 1

Number of pages published yearly by the leading journals in mathematical economics (Econometrica (abbr. Eta), *Review of Economic Studies*, (For the first 29 years the *Review of Economic Studies* was published on an academic rather than on a calendar year basis. As a result, only one issue appeared in 1933, compared with three in 1934; hence the spurious initial increase in the graph.) *International Economic Review, Journal of Economic Theory, Journal of Mathematical Economics*).

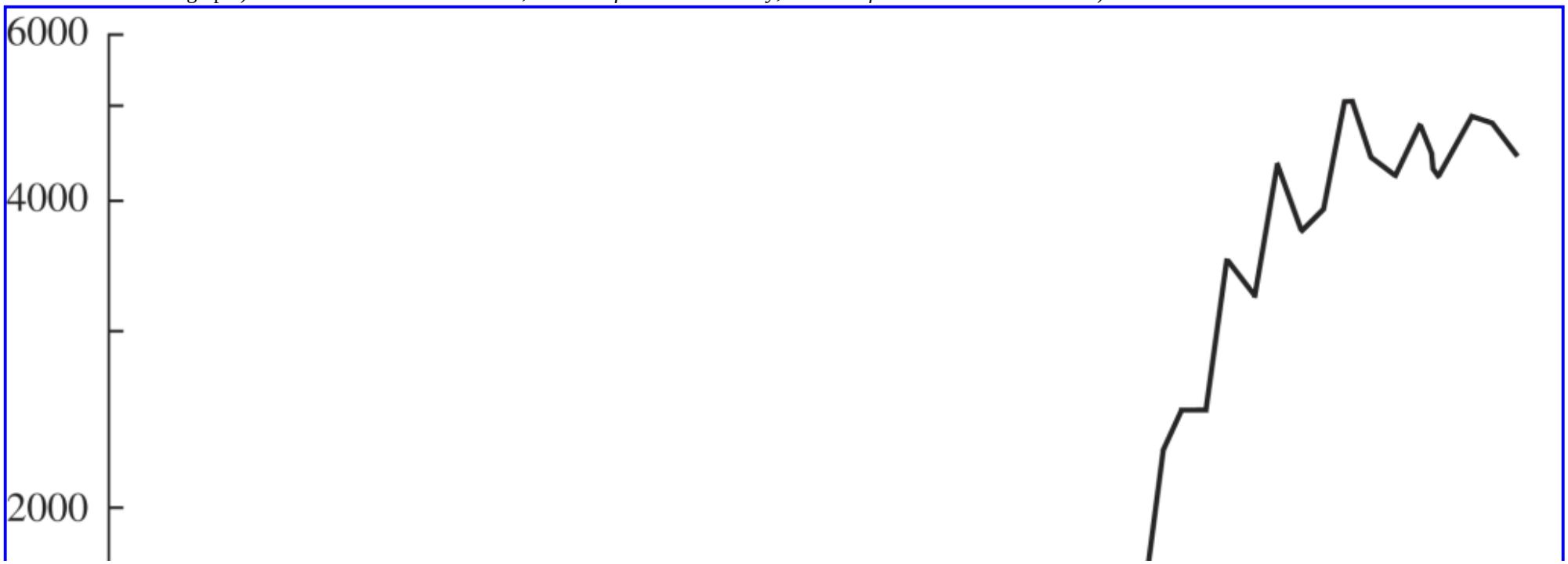

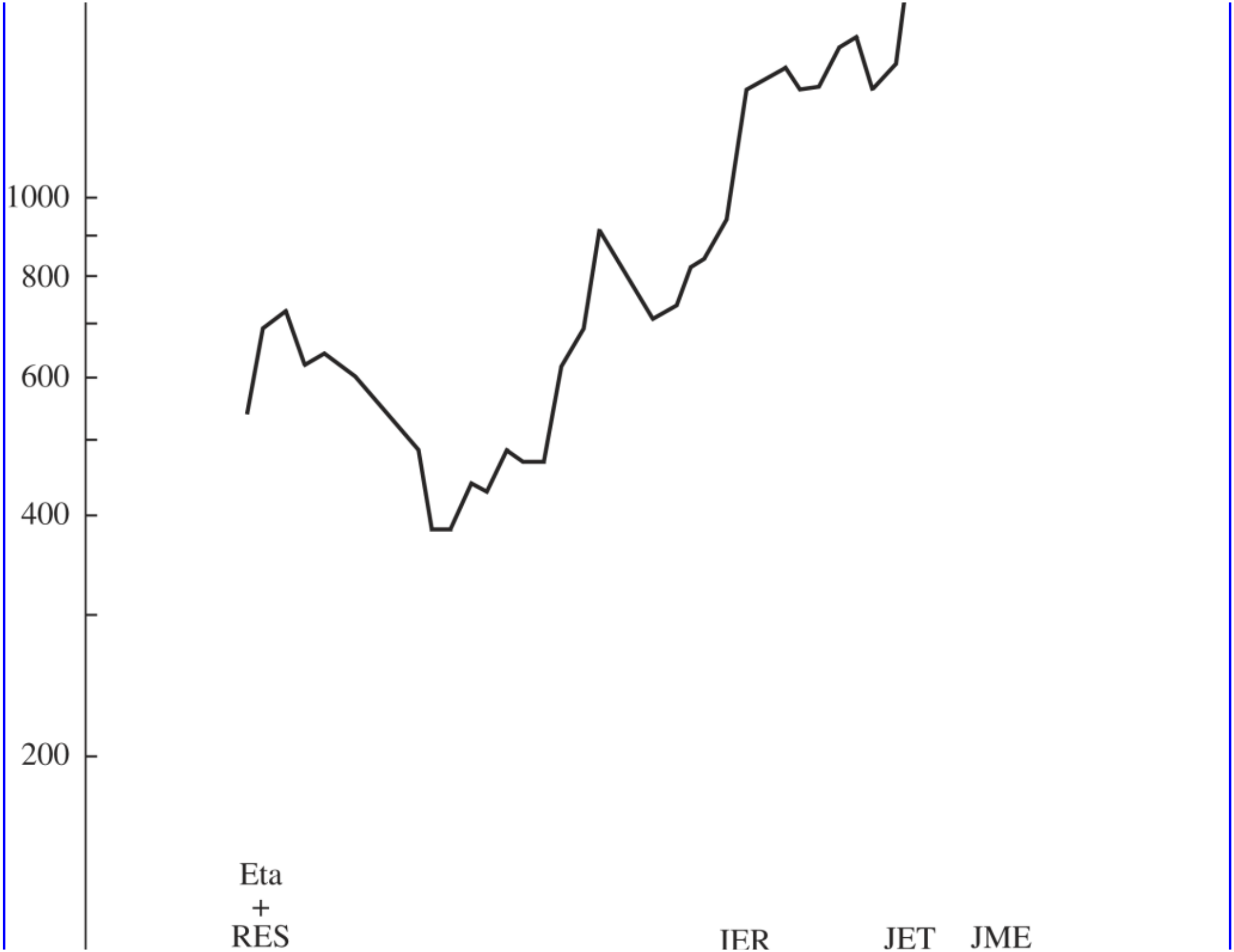
1000
800
600
400
200
Eta
+
RES
JER
JET
JME

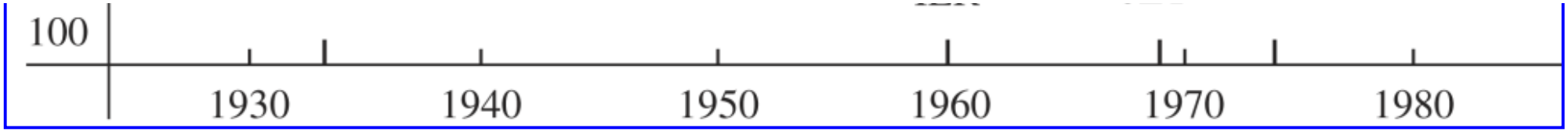

II. As a formal model of an economy acquires a mathematical life of its own, it becomes the object of an inexorable process in which rigour, generality and simplicity are relentlessly pursued.

Before 1944, articles on economic theory only exceptionally met the standards of rigour common in mathematical periodicals. But several of the exceptions were outstanding, among them the two papers of von Neumann of 1928 and of 1937, and the three papers of Abraham Wald of 1935–6 on the existence of a general economic equilibrium. In 1944 the *Theory of Games and Economic Behavior* gained full rights for uncompromising rigour in economic theory and prepared the way for its axiomatization. An axiomatized theory first selects its primitive concepts and represents each one of them by a mathematical object. For instance the consumption of a consumer, his set of possible consumptions and his preferences are represented respectively by a point in the commodity space, a subset of the commodity space and a binary relation in that subset. Next, assumptions on the objects representing the primitive concepts are specified, and consequences are mathematically derived from them. The economic interpretation of the theorems so obtained is the last step of the analysis. According to this schema, an axiomatized theory has a mathematical form that is completely separated from its economic content. If one removes the economic interpretation of the primitive concepts, of the assumptions and of the conclusions of the model, its bare mathematical structure must still stand. This severe test is passed only by a small minority of the papers on economic theory published by *Econometrica* and by the *Review of Economic Studies* during their first decade.

The divorce of form and content immediately yields a new theory whenever a novel interpretation of a primitive concept is discovered. A textbook illustration of this application of the axiomatic method occurred in the economic theory of uncertainty. The traditional characteristics of a commodity were its physical description, its date, and its location when in 1953 Kenneth Arrow proposed adding the state of the world in which it will be available. This reinterpretation of the concept of a commodity led, without any formal change in the model developed for the case of certainty, to a theory of uncertainty which eventually gained broad acceptance, notably among finance theorists.

The pursuit of logical rigour also contributed powerfully to the rapid expansion of mathematical economics after World War II. It made it possible for research workers to use the precisely stated and flawlessly proved results that appeared in the literature without scrutinizing their statements and their proofs in every detail. Another cumulative process could thus gather great momentum.

The exact formulation of assumptions and of conclusions turned out, moreover, to be an effective safeguard against the ever-present temptation to apply an economic theory beyond its domain of validity. And by the exactness of that formulation, economic analysis was sometimes brought closer to its ideology-free ideal. The case of the two main theorems of welfare economics is symptomatic. They respectively give conditions under which an equilibrium relative to a price system is a Pareto optimum, and under which the converse holds. Foes of state intervention read in those two theorems a mathematical demonstration of the unqualified superiority of market economies, while advocates of state intervention welcome the same theorems because the explicitness of their assumptions emphasizes discrepancies between the theoretic model and the economies that they observe.

Still another consequence of the axiomatization of economic theory has been a greater clarity of expression, one of the most significant gains that it has achieved. To that effect, axiomatization does more than making assumptions and conclusions explicit and exposing the deductions linking them. The very definition of an economic concept is usually marred by a substantial margin of ambiguity. An axiomatized theory substitutes for that ambiguous concept a mathematical object that is subjected to definite rules of reasoning. Thus an axiomatic theorist succeeds in communicating the meaning he intends to give to a primitive concept because of the completely specified formal context in which he operates. The more developed this context is, the richer it is in theorems, and in other primitive concepts, the smaller will be the margin of ambiguity in the intended interpretation.

Although an axiomatic theory may flaunt the separation of its mathematical form and its economic content in print, their interaction is sometimes close in the discovery and elaboration phases. As an instance, consider the characterization of aggregate excess demand functions in an l-commodity exchange economy. Such a function maps a positive price vector into an aggregate excess demand vector, and Walras' Law says that those two vectors are orthogonal in the Euclidean commodity-price space. That function is also homogeneous of degree zero. For a mathematician, these are compelling reasons for normalizing the price vector so that it belongs to the unit sphere. Then aggregate excess demand can be represented by a vector tangent to the sphere at the price vector with which it is associated. In other words, the aggregate excess demand function is a vector field on the positive unit sphere. Hugo Sonnenschein conjectured in 1973 that any continuous function satisfying Walras' Law is the aggregate excess demand function of a finite exchange economy. A proof of that conjecture (Debreu, 1974) was suggested by the preceding geometric interpretation since any vector field on the positive unit sphere can be written as a sum of l elementary vector fields, each one obtained by projecting a positive vector on one of the l coordinate axes into the tangent hyperplane. There only remains to note that every continuous elementary vector field is the excess demand function of a mathematically well-behaved consumer. Mathematical form and economic content alternately took the lead in the development of this proof.

The pursuit of generality in a formalized theory is no less imperative than the pursuit of rigour, and the mathematician's compulsive search for ever weaker assumptions is reinforced by the economist's awareness of the limitations of his postulates. It has, for example, expurgated superfluous differentiability assumptions from economic theory, and prompted its extension to general commodity spaces.

Akin in motivation, execution and consequences is the pursuit of simplicity. One of its expressions is the quest for the most direct link between the assumptions and the conclusions

of a theorem. Strongly motivated by aesthetic appeal, this quest is responsible for more transparent proofs in which logical flaws cannot remain hidden, and which are more easily communicated. In extreme cases the proof of an economic proposition becomes so simple that it can dispense with mathematical symbols. The first main theorem of welfare economics, according to which an equilibrium relative to a price system is a Pareto optimum, is such a case.

In the demonstration, we study an economy consisting of a set of agents who have collectively at their disposal positive amounts of a certain number of commodities and who want to allocate these total resources among themselves. By the consumption of an agent, we mean a list of the amounts of each commodity that he consumes. And by an allocation, we mean a specification of the consumption of each agent such that the sum of all those individual consumptions equals the total resources. Following Pareto, we compare two allocations according to a unanimity principle. We say that the second allocation is collectively preferred to the first allocation if every agent prefers the consumption that he receives in the second to the consumption that he receives in the first. According to this definition, an allocation is optimal if no other allocation is collectively preferred to it. Now imagine that the agents use a price system, and consider a certain allocation. We say that each agent is in equilibrium relative to the given price system if he cannot satisfy his preferences better than he does with his allotted consumption unless he spends more than he does for that consumption. We claim that an allocation in which every agent is in equilibrium relative to a price system is optimal. Suppose, by contradiction, that there is a second allocation collectively preferred to the first. Then every agent prefers his consumption in the second allocation to his consumption in the first. Therefore the consumption of every agent in the second allocation is more expensive than his consumption in the first. Consequently the total consumption of all the agents in the second allocation is more expensive than their total consumption in the first. For both allocations, however, the total consumption equals the total resources at the disposal of the economy. Thus we asserted that the value of the total resources relative to the price system is greater than itself. A contradiction has been obtained, and the claim that the first allocation is optimal has been established.

This result, which provides an essential insight into the role of prices in an economy and which requires no assumption within the model, is remarkable in another way. The two concepts that it relates might have been isolated, and its symbol-free proof might have been given early in the history of economic theory and without any help from mathematics. In fact that demonstration is a late by-product of the development of the mathematical theory of welfare economics. But to economists who have even a casual acquaintance with mathematical symbols, the previous exercise is not more than an artificial *tour de force* that has lost the incisive conciseness of a proof imposing no bar against the use of mathematics. That conciseness is one of the most highly prized aspects of the simplicity of expression of a mathematized theory.

In close relationship with its axiomatization, economic theory became concerned with more fundamental questions and also more abstract. The problem of existence of a general economic equilibrium is representative of those trends. The model proposed by Walras in 1874–7 sought to explain the observed state of an economy as an equilibrium resulting from the interaction of a large number of small agents through markets for commodities. Over the century that followed its publication, that model came to be a common intellectual framework for many economists, theorists as well as practitioners. This eventually made it compelling for mathematical economists to specify assumptions that guarantee the existence of the central concept of Walrasian theory. Only through such a specification, in particular, could the explanatory power of the model be fully appraised. The early proofs of existence of Wald in 1935–6 were followed by a pause of nearly two decades, and then by the contemporary phase of development beginning in 1954 with the articles of Arrow and Debreu, and of Lionel McKenzie.

In the reformulation that the theory of general economic equilibrium underwent, it reached a higher level of abstraction. From that new viewpoint a deeper understanding both of the mathematical form and of the economic content of the model was gained. Its role as a benchmark was also perceived more clearly, a role which prompted extensions to incomplete markets for contingent commodities, externalities, indivisibilities, increasing returns, public goods, temporary equilibrium, … .

In an unanticipated, yet not unprecedented, way greater abstraction brought Walrasian theory closer to concrete applications. When different areas of the field of computable general equilibrium were opened to research at the University of Oslo, at the Cowles Foundation, and at the World Bank, the algorithms of Scarf included in their lineage proofs of existence of a general economic equilibrium by means of fixed point theorems. This article has credited the mathematical form of theoretic models with many assets. Their sum is so large as to turn occasionally into a liability, as the seductiveness of that form becomes almost irresistible. In its pursuit, research may be tempted to forget economic content and to shun economic problems that are not readily amenable to mathematization. No attempt will be made here, however, to draw a balance sheet, to the debit side of which justice would not be done. Economic theory is fated for a long mathematical future, and in other editions of Palgrave authors will have the opportunity, and possibly the inclination, to choose as a theme ‘Mathematical Form vs. Economic Content’.

First published in Econometrica, November 1986, with revisions.

See Also

- computation of general equilibria
- cores
- existence of general equilibrium
- game theory
- regular economies

- uncertainty and general equilibrium

Bibliography

Allais, M. 1943. *A la recherche d'une discipline économique*. Paris: Imprimerie Nationale.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46, 1483–7.

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Berkeley: University of California Press, 507–32.

Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie*, Paris: Centre National de la Recherche Scientifique, 41–8.

Arrow, K.J. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Arrow, K.J. and Intriligator, M.D., eds. 1981–5. *Handbook of Mathematical Economics*, vols 1,2 and 3. Amsterdam: North-Holland Publishing Company.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Balasko, Y. 1986. *Foundations of the Theory of General Equilibrium*. New York: Academic Press.

Brouwer, L.E.J. 1912. Über Abbildungen von Mannigfaltigkeiten. *Mathematische Annalen* 71, 97–115.

Brown, D.J. and Robinson, A. 1972. A limit theorem on the cores of large standard exchange economies. *Proceedings of the National Academy of Sciences of the USA* 69, 1258–60.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: L. Hachette.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19, 273–92.

Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences of the USA* 38, 886–93.

Debreu, G. 1959. *Theory of Value: an Axiomatic Analysis of Economic Equilibrium*. New York: Wiley.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–21.

Debreu, G. 1977. The axiomatization of economic theory. Unpublished lecture given at the University of Bonn, 22 April.

Debreu, G. 1982. Existence of competitive equilibrium. Chapter 15 in Arrow and Intriligator (1981–5).

Dierker, E. 1974. *Topological Methods in Walrasian Economics*. Berlin: Springer.

Dierker, E. 1975. Gains and losses at core allocations. *Journal of Mathematical Economics* 2, 119–28.

Dierker, E. 1982. Regular economies. Chapter 17 in Arrow and Intriligator (1981–5).

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Gale, D. 1956. The closed linear model of production. In *Linear Inequalities and Related Systems*, ed. H.W. Kuhn and A.W. Tucker, Princeton: Princeton University Press.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton: Princeton University Press.

Hildenbrand, W. 1982. Core of an economy. Chapter 18 in Arrow and Intriligator (1981–5).

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8, 457–9.

Koopmans, T.C., ed. 1951. *Activity Analysis of Production and Allocation*. New York: Wiley.

Lyapunov, A.A. 1940. Sur les fonctions-vecteurs complètement additives. *Izvestiia Akademii Nauk SSSR* 4, 465–78.

Mantel, R. 1974. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7, 348–53.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: a Differentiable Approach*. Cambridge: Cambridge University Press.

McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22, 147–61.

Minkowski, H. 1911. Theorie der konvexen Körper. *Gesammelte Abhandlungen*, Leipzig and Berlin: Teubner, vol. 3, 131–229.

Nash, J.F. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the USA* 36, 48–9.

Neumann, J.von. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320.

Neumann, J.von. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. In *Ergebnisse eines mathematischen Kolloquiums*, vol. 8, 73–83.

Neumann, J.von. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Pareto, V. 1896–7. *Cours d'économie politique*. Lausanne: Rouge.

Pareto, V. 1909. *Manuel d'économie politique*. Paris: Giard.

Pareto, V. 1911. Economie mathématique. In *Encyclopédie des sciences mathématiques*, Paris: Gauthier-Villars, vol. I(4), 591–640.

Robinson, A. 1966. *Non-standard Analysis*. Amsterdam: North-Holland.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Sard, A. 1942. The measure of the critical points of differentiable maps. *Bulletin of the American Mathematical Society* 48, 883–90.

Scarf, H. 1962. An analysis of markets with a large number of participants. In H. Scarf, *Recent Advances in Game Theory*, Princeton: Princeton University Press.

Scarf, H. 1973. (With the collaboration of T. Hansen) *The Computation of Economic Equilibria*. New Haven: Yale University Press.

Scarf, H. 1982. The computation of equilibrium prices: an exposition. Chapter 21 in Arrow and Intriligator (1981–5).

Scarf, H.E. and Shoven, J.B. 1984. *Applied General Equilibrium Analysis*. Cambridge: Cambridge University Press.

Shubik, M. 1959. Edgeworth market games. In *Contributions to the Theory of Games*, vol. 4, Annals of Mathematical Studies 40, Princeton: Princeton University Press.

Smale, S. 1981. Global analysis and economics. Chapter 8 in Arrow and Intriligator (1981–5).

Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6, 345–54.

Vind, K. 1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5, 165–77.

Wald, A. 1935. Über die eindeutige positive Lösbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines mathematischen Kolloquiums* 6, 12–20.

Wald, A. 1936a. Über die Produktionsgleichungen der ökonomischen Wertlehre. In *Ergebnisse eines mathematischen Kolloquiums*. 7, 1–6.

Wald, A. 1936b. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie* 7, 637–70.

Walras, L. 1874–7. *Eléments d'économie politique pure*. Lausanne: L. Corbaz.

How to cite this article

Debreu, Gerard. "mathematical economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000107> doi:10.1057/9780230226203.1061

The New Palgrave Dictionary of Economics Online

mathematical methods in political economy

F.Y. Edgeworth

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

calculus of variations; mathematical economics; mathematical method in political economy; simultaneous equations; statistics

Article

The idea of applying mathematics to human affairs may appear at first sight an absurdity worthy of Swift's Laputa. Yet there is one department of social science which by general consent has proved amenable to mathematical reasoning – statistics. The operations not only of arithmetic, but also of the higher calculus, are applicable to statistics. What has long been admitted with respect to the average results of human action has within the last half-century been claimed for the general laws of political economy. The latter, indeed, unlike the former, do not usually present numerical constants; but they possess the essential condition for the application of mathematics: constancy of *quantitative* – though not necessarily numerical – relations. Such, for example, is the character of the law of Diminishing Returns: that an increase in the capital and labour applied to land is (tends to be) attended with a less than proportionate increase in produce. The language of Functions is well adapted to express such relations. When, as in the example given, and frequently in economics (see Marshall, *Principles*, 5th edn, Preface, p. xix), the relation is between *increments* of quantities, the differential calculus is appropriate. In the simpler cases the geometrical representations of functions and their differentials may with advantage be employed.

Among the branches of the economic calculus *simultaneous equations* are conspicuous. Given several quantitative – though not in general numerical – relations between several variable quantities, the economist needs to know whether the quantities are to be regarded as *determinate*, or not. A beautiful example of numerous prices determined by numerous conditions of supply and demand is presented by Professor Marshall in his ‘bird's-eye view of the problems of joint demand, composite demand, joint supply, and composite supply’ (*Principles*, Mathematical Appendix, note xxi). ‘However complex the problem may become, we can see that it is theoretically determinate’ (ibid., cf. Preface, p. xx). When we have to do with only *two* conditions, two *curves* may be advantageously employed instead of two

equations.

The mathematical operations which have been mentioned, and others – in particular the integral calculus, are all contained in the calculus of *maxima* and *minima*, or, as it is called, of *variations;* which seems to comprehend all the higher problems of abstract economics. For instance, Prof. Marshall, after writing out a number of equations ‘representing the causes that govern the investment of capital and effort in any undertaking’, adds, ‘they may all be regarded as mathematically contained in the statement that H–V [the net advantages] is to be made a maximum’ (*Principles*, Mathematical Appendix, 2nd and later editions, note xiv). It was profoundly said by Malthus, ‘Many of the questions both in morals and politics seem to be of the nature of the problems *de maximis et minimis* in fluxions.’ The analogy between economics and mechanics in this respect is well indicated by Dr Irving Fisher in his masterly *Mathematical Investigations*.

The property of dealing with quantities not expressible in numbers, which is characteristic of mathematical economics, is not to be regarded as a degrading peculiarity. It is quite familiar and allowed in ordinary mathematics. For instance, if one side of a plane triangle is greater than another, the angle opposite the greater side is greater than the angle opposite the less side (*Euclid*, Book I). Quantitative statements almost as loose as those employed in abstract economics occur in the less perfectly conquered portions of mathematical physics, with respect to the distances of the fixed stars, for instance (see Sir Robert Ball, *Story of the Heavens*, ch. xxi); e.g. before 1853 it was only known that ‘the distance of 61 Cygni could not be *more* than sixty billions of miles’. It is really less than forty billions.

The instance of astronomy suggests a secondary or indirect use of mathematical method in economics, which physical science has outgrown. As the dawn of the Newtonian, or even of the Copernican, theory put to flight the vain shadows of astrology, so the mere statement of an economic problem in a mathematical form may correct fallacies. Attention is directed to the data which would be required for a scientific solution of the problem. Variable quantities expressed in symbols are less liable to be treated as constant. This sort of advantage is obtained by formulating the relation between quantity of precious metal in circulation and the general level of prices, as Sir John Lubbock (senior) has done in his pamphlet *On currency* (anonymous, 1840). Thus the mathematical method contributes to that negative or dialectic use of theory which consists in meeting fallacious arguments on their own ground of abstract reasoning (see some remarks on this use of theory by Prof. Simon Newcomb in the June number of the *Quarterly Journal of Economics*, 1893; and compare Prof. Edgeworth, *Economic Journal*, vol. i, p. 627). The mathematical method is useful in clearing away the rubbish which obstructs the foundation of economic science, as well as in affording a plan for the more regular part of the structure.

The modest claims here made for the mathematical method of political economy may be illustrated by comparing it with the literary or classical method in the treatment of some of the higher problems of the science. The fundamental principle of supply and demand has been stated by J.S. Mill with much precision in ordinary language (*Political Economy*, book iii, ch. 2, §§ 4, 5, and, better, review of Thornton, *Dissertations*, vol. iv). But he is not very happy in indicating the distinction between a rise of price which is due to a diminution of supply – the dispositions of the buyers, the Demand Curves remaining constant – and the rise of price which is due to a displacement of the demand curve. He appears not to perceive that the position of equilibrium between supply and demand is *determinate*, even where it is not *unique* – a conception supplied by equations with multiple roots or curves intersecting in several points. The want of this conception seems to involve even Mill's treatment of the subject in obscurity (*Political Economy*, book iii, ch. 18, § 6).

The use of simultaneous equations or intersecting curves facilitates the comprehension of the ‘fundamental symmetry’ (Marshall) between the forces of demand and supply; the littérateurs lose themselves in wordy disputes as to which of the two factors ‘regulates’ or ‘determines’ value.

The disturbance of the conditions of supply by a tax or bounty, or other impediment or aid, gives rise to problems too complicated for the unaided intellect to deal with. Prof. Marshall, employing the mathematical theory of Consumers’ rent, reaches the conclusion that it might *theoretically* be advantageous to tax commodities obeying the law of decreasing returns in order with the proceeds to give bounty to commodities following the opposite law (*Principles*, book v, ch. xiii, § 7). The want of the theory of consumer's rent renders obscure Mill's treatment of the ‘gain’ which a country may draw to itself by taxing exports or imports (*Political Economy*, book v, ch. 4, § 6; cf. book iii, ch. 18, § 5). This matter is much more clearly expressed by the curves of Messrs. Auspitz and Lieben (*Untersuchungen*, Article 81).

The preceding examples presuppose free competition; the following relate to monopoly. The relation between the rates and the traffic of a railway is shown with remarkable clearness by the aid of a diagram in the appendix to Prof. Hadley's *Railroad Transportation*. By means of elaborate curves Prof. Marshall shows that a government having regard to the interest of the consuming public, as well as to its revenue, may fix a much lower price than a monopolist actuated by mere self-interest. The taxation of monopolies presents problems which require the mathematical method initiated by Cournot. His reasoning convinces of error the following statement made by Mill (book v, chs 4, 6) and others: ‘A tax on rare and high-priced wines will fall only on the owners of the vineyard,’ for ‘when the article is a strict monopoly … the price cannot be further raised to compensate for the tax’. Cournot obtains by mathematical reasoning the remarkable theorem that in cases where there is a joint demand for articles monopolized by different individuals, the purchaser may come off worse than if he had dealt with a single monopolist. This case is more important than at first appears (Marshall, *Principles*, 2nd edn, book v, ch. x, § 4; 5th edn, book v, ch. xi, § 7).

Under the head of monopoly may be placed the case of two individuals or corporate units dealing with each other. The indeterminateness of the bargain in this case is perhaps best contemplated by the aid of diagrams.

These examples, which might be multiplied, seem to prove the usefulness of the mathematical method. But the estimate would be imperfect without taking into account the abuses and defects to which the method is liable. One of these is common to every *organon* – especially new ones – liability to be overrated. As Prof. Marshall says, ‘When the actual conditions of particular problems have not been studied, such [mathematical] knowledge is little better than a derrick for sinking oil-wells where there are no oil-bearing strata.’ Again, the mathematical method is a machinery, the use of which is very liable to be overbalanced by the cost to others than the maker of acquiring it. Not only is mathematics a foreign language ‘to the general’; but even to mathematicians a new notation is an unknown dialect which it may not repay to learn. As Prof. Marshall says, ‘It seems doubtful whether any one spends his time well in reading lengthy translations of economic doctrines into mathematics that have not been made by himself.’

This estimate of the uses and dangers of mathematical method may be confirmed by reference to the works in the subjoined list; which does not pretend to be exhaustive.

Bibliography

Auspitz, R. and Lieben, R. 1889. *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Dupuit, E.T. 1844. De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées*, 2nd series 8, 332–75.

Dupuit, E.T. 1849. *De l'influence des Péages*. Paris: Guillaumin.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Gossen, H.H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs*. 2nd edn, Berlin: Praeger, 1889.

Jevons, W.S. 1871. *Theory of Political Economy*. 3rd edn, London: Macmillan, 1888.

Keynes, J.N. 1891. *The Scope and Method of Political Economy*. 3rd edn, revised, London: Macmillan, 1904.

Launhardt, W. 1885. *Mathematische Begründung der Volkwirthschaftslehre*. Leipzig.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan 2nd edn, 1891; 5th edn, 1907.

Pantaleoni, M. 1889. *Principii di economia pura*. Florence: G. Barb•ra.

Pareto, V. 1896. *Cours d’économie politique*. Lausanne: Rouge.

Walras, L. 1874. *Eléments d’économie politique pure*. 2nd edn, Lausanne: F. Rouge, 1889.

Wicksteed, P.H. 1888. *Alphabet of Economic Science*. London: Macmillan.

How to cite this article

Edgeworth, F. Y. "mathematical methods in political economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000108> doi:10.1057/9780230226203.1062

The New Palgrave Dictionary of Economics Online

mathematics and economics

E. Roy Weintraub

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The interconnection of mathematics and economics reflects changes in both the mathematics and economics communities over time. The respective histories of these disciplines are intertwined, so that both changes in mathematical knowledge and changing ideas about the nature of mathematical knowledge have effected changes in the methods and concerns of economists.

Keywords

Arrow, K.; Axiomatics; Cowles Commission; Debreu, G.; econometrics; Euclid; Koopmans, T.; Marshall, A.; mathematical economics; mathematics and economics; Newton, I.; Whewell, W.

Article

Understanding the connection between mathematics and economics is not the same as understanding the nature and role of mathematical economics. ‘Mathematical economics’ is the employment of mathematics in economics itself. Explaining or justifying mathematical economics often involves essentialist arguments concerning the true nature of economic objects and the true nature of the economy, as well as arguments suggesting that employing mathematics is appropriate since the underlying ‘economy’ is quantitative in nature. Consequently, an historical discussion of mathematical economics will be a narrative of increased sophistication over time in economics, as mathematical tools, techniques and methods move into economic discourse and enrich economic analysis.
Alternatively, one can discuss the relation between mathematics and economics in terms of separate intellectual activities performed in separate intellectual communities, and in that case one will wish to look over time at the interpenetration of the ideas and practices of the two communities across their highly permeable boundaries. The history of mathematics concerns the changing body of mathematical knowledge such as new theorems proved, new research areas opened, and new techniques developed. But the history also involves changing images of mathematical knowledge: changing perspectives and understandings, for example, about the nature of mathematical objects, what constitutes a proof, what

constitutes rigour, what constitutes useful versus not useful mathematics, and so forth (see Corry, 1996, p. 3). Similarly the history of economics involves a history of not only the development of economic knowledge, but the development and changes in images of economic knowledge: what constitutes the economy, what constitutes a good explanation in economics, what constitutes serious empirical work in economics, what a good model is, and so on. Consequently a discussion of the interconnection of mathematics and economics requires not just attention to the interconnection of the bodies of knowledge, as is reflected in the historical discussion of mathematical economics, but a historical discussion of the interconnection of their respective images of knowledge. Put another way, a discussion of the connection of mathematics and economics must reflect economists' changing conceptions of the image of mathematical knowledge and not just their changing understandings of the body of mathematical knowledge.

This distinction between the body of knowledge and the images of knowledge provides a different perspective on the relation between mathematics and economics. The central point for economists to understand is that there were three distinct shifts in the image of mathematics from the beginning of the 19th century to the end the 20th century.

From geometry to mechanics

As a starting point, consider the conditions and perspectives under which mathematics was produced early in the 19th century. Looking closely, we see, particularly in England, the importance of both Euclid's *Elements* and Newton's *Principia*. That is, from relatively early in the 19th century, through the modifications of the Cambridge Tripos in 1849, and on through the middle third of the 19th century at Cambridge, mathematics was understood as flowing, in its purpose and nature, from both Euclid and Newton. From Euclid one understood that geometry was the paradigm of mathematics, and that it was a path to truth. Theorems were derived from assumptions called axioms, where the truth of those assumptions was self-evident from our understanding of the physical world. To learn geometry was to understand how rigorous arguments could lead to truth. One studied mathematics, specifically geometry, as an exemplar of how one deduced truths about the world, and thus mathematics was the paradigm of deductive thought and logical ratiocination. Parallel to this view of how deductive reasoning from true premises could lead to true conclusions, Newton's *Principia* (his mathematical proofs of course were all based on Euclidian geometry – even the calculus derivations were geometrical), suggested how this kind of mathematics could also open up an understanding of the physical world. Students were required to study mathematics because it provided a way of achieving truth.

This image of mathematics is at the root of Ricardo's arithmetical models, and is present in Whewell's papers (1829; 1831; 1850) on economics using mathematics, for Whewell himself was central in reconstructing the Cambridge Tripos around Euclidean geometry and Newton's *Principia* at mid-19th century. Economics was to employ a particular kind of mathematics, Euclidean geometry, to demonstrate its propositions. Just as Newton employed geometrical proofs of his propositions, so too did Marshall. It is an interesting exercise to open Alfred Marshall's *Principles* next to the Newton's *Principia* and see the physical similarity of the proofs or demonstrations of the propositions in each book. Marshall, as Second Wrangler in the Mathematical Tripos of January 1865, had had to master both Euclid and Newton.

The first change in the image of mathematics was developed from a new conception of what mathematical truth might mean. It occurred over the second third of the 19th century and was then well incorporated in the Continental tradition in mathematics. That is, outside Britain there was a change in the image of mathematics between the time of Whewell's defence of mathematics in the educational process, a defence based in the notion that mathematics (*vide* Euclid, Newton) was the paradigm of certain and secure knowledge (the time of Marshall's student days), and Marshall's later time as Professor of Political Economy. The emergence of non-Euclidean geometries had made Whewell's argument about axiomatics, and inevitable truth, ring hollow long before the turn of the 20th century. In the time of the new geometries, the difficulty of linking mathematical truth to a particular (Euclidean) geometry produced a real crisis of confidence for Victorian educational practice (Richards, 1988). This first crisis prepared the late Victorian mind for the new idea that mathematical rigour had to be associated with physical argumentation. And it was this new image of mathematics in science that helps us to understand the concerns of individuals like Edgeworth and Pareto.

An emergent set of themes in mathematics developed from the increased awareness of alternatives to Euclidean geometry, and the recognition that no one set of axioms could be selected for demonstrating the truth of all mathematical propositions. Thus the success of the new rational mechanics (Lagrange's programme of applying techniques of advanced calculus to the study of motions of solids and liquids) in making sense of the world of physical systems encouraged a refinement of the truth-producing view of mathematics. That is, in the last third of the 19th century, in Britain as well as Italy, France and Germany, a rigorous mathematical argument began to be seen as one based on a substrate of physical reasoning. For an argument to be rigorous, and thus believable, the mathematical structure had to be founded generally on the most successful of applied mathematical practices, namely, rational mechanics. *A valid and good and useful mathematical model was a model that had physical interpretations*. The ‘marginal revolution’ in economics was precisely this new understanding. One sees this very clearly in Marshall, who was at the cusp of this changed image of mathematics, for his derivations were offered using Euclidian geometry, but whose mathematical arguments about equilibrium and stability are instantiations of mechanical devices like an egg in a bowl, or a pair of scissors. Put another way, through much of the 19th century in British mathematics, and thus to a degree among insular British economists for whom British mathematics *was* mathematics, rigour in argument was associated with geometric proofs based on assumptions, called axioms, that could be linked to constrained optimization processes associated with particular physical systems. Rational mechanics was taken as a paradigm for what economists came to call the marginal revolution, which, however, was hardly revolutionary but rather the migration of rational mechanical ideas into economic discourse (Mirowski, 1989). Thus, by the last decades of the 19th century one finds economists employing specific mechanical models of economic behaviour. Walras, Pareto, Marshall, Edgeworth and Fisher were producing rigorous mathematical models of economic processes, where rigour was associated with a mathematics tied to physical processes.

From mechanics to axiomatics

But by 1900 the images of, and styles of doing, mathematics were beginning to change again in response to new challenges in mathematics and physics. In mathematics, there were problems associated with the foundations of mathematics. There were apparent inconsistencies in set theory associated with Georg

Cantor's new ideas on ‘infinity’ (that is, transfinite cardinals, and the continuum of real numbers), and apparent inconsistencies in the foundations of arithmetic and logic, associated with work by Frege and Peano. Similarly troubling was the failure of physics, particularly rational mechanics, to solve the new problems associated with black-body radiation, quanta and relativity. If the deterministic mechanical mode of physical argumentation was to be replaced by an alternative physical theory, what constituted a rigorous mathematical argument had to be re-described. In any event, some established areas of mathematics were no longer connected to a canonical physical model (Weintraub, 2002).

Consequently, around the end of the 19th century, just as economists had begun to understand that constructing a mathematical science required basing argumentation of the physical reasoning of rational mechanics, and the measurement of quantities to further ground those reasoning chains, the image of mathematical knowledge was again changing. Modelling the concerns of the new physics appeared to require a new mathematics, based less on deterministic dynamical systems and more on statistical argumentation, algebra, and new beliefs about appropriate axioms for logic and arithmetic.

Just as the objects of the physical world appeared changed – gone were billiard balls, newly present were quanta – the recognition that the paradoxes of set theory and logic were intertwined led mathematicians to seek new foundations for their subject. Analysis of those foundations of set theory, logic and arithmetic, and thus the foundations of sciences based on mathematics, were now to be based on axiomatic thinking. A rigorous argument was to be one built on strong foundations, and axiomatizing the structure of theories, in both physics and mathematics, was a path to the development of those theories (Hilbert, 1918). Thus, following a late 19th century period in which mathematical rigour was to be established by basing the mathematics on physical reasoning, around 1900 – as understanding of the physical world became less secure – *mathematical truth was to be established not relative to physical reasoning but relative to other mathematical theories and objects*. From a physical reductionism mathematics moved to a mathematical reductionism, in the guise of one or another set of ideas about formalism: problems and paradoxes and confusions in turn-of-the-century mathematics were to be resolved by a re-conceptualization of the nature of the fundamental objects of mathematics. The images of mathematical knowledge and ideas of rigour, truth, formalization and proof all changed over this period.

It took a number of decades for this new image of mathematics to become securely established in the mathematical community. From Hilbert's 1918 call for axiomatization as the road to knowledge in mathematics and science, through the interwar years, mathematicians were slow to reframe their working concerns. So too did economists’ use of mathematics in the interwar period reflect the earlier perspectives of modelling economic problems as constrained optimization demonstrations imitating 19th century mechanics. Beginning in the 1930s, however, a group of French mathematicians, collectively called ‘Nicholas Bourbaki’, began rewriting mathematics from the foundationalist perspective (Weintraub and Mirowski, 1994). Mathematics was conceived of, in their project, as growing organically from very basic ideas about sets, which led inexorably to the identification of a small number of ‘mother structures’ (algebraic, order, and topological) from which other structures, other branches of mathematics, could be derived. Rigorous mathematics was not grounded in physical models but rather in mathematics itself. Mathematics was to concern itself with analyses of mathematical structures. Over the next few decades pure mathematics, or mathematics uncontaminated by applications and disengaged from the world of applications, gained sway in the mathematics community. It was in this period that the eminent mathematician Paul Halmos (1981) famously titled an article ‘Applied

mathematics is bad mathematics'. In economics, this concatenation of ideas moved into mainstream theory with the work of Gerard Debreu, Kenneth Arrow and Tjalling Koopmans. The Cowles Commission, in the 1940s at the University of Chicago, became the site for production of this kind of work in mathematical economic theory, particularly general equilibrium theory.

Yet, even as a pure mathematics was taking hold in economics, the exigencies of the Second World War and economists' involvement with scientists, engineers, and other social scientists, moved mathematical economists' concerns back from axiomatization and into what would become operations research. This, of course, was not 'pure' at all, but based on concrete problems of real systems. As Amy Dahan Dalmedico, the historian of mathematics, noted:

> The second World War initiated what I shall call 'image war' or 'representation war' concerning what mathematics was about, what it dealt with, and how. Over the course of the 1950s and 1960s, this 'war' was progressively developed until the balance of power began to shift perceptibly at the end of the 1970s and during the 1980s. This 'war' was focused mainly on the cleavage between pure and applied mathematics, and on the tacit hierarchy – of concepts as much as of values – informing these categories of 'pure' and 'applied'. (Dalmedico, 2001, p. 224)

Thus, Bourbakist images of mathematics were becoming dominant in economics at the same time as the major challenge to those ideas was forming outside 'pure' theory. The image of mathematics as a discipline concerned with understanding the structures of mathematical objects was indeed dominant in the 1950s and 60•s, not only in the United States but in a number of other countries. Yet, from the Second World War on through the cold war, applied mathematics was taking root in disparately profound ways, and was attracting more and more support in the form of grants and contracts and students. New fields of statistics, computer science and operations research flourished. Consequently, economists' ideas about mathematics began to undergo changes, as usual with some time lag, mirroring the changing images of mathematics that were reshaping interests and methods in the mathematics community itself. 'While *structure* was the emblematic term of the 1960s, *model* has now taken its place. In the physical sciences, climatology, engineering science, economics, and the social sciences, the practice of model-building has gradually dominated the terrain. It is today absolutely massive and intrinsically bound up with numerical experimentation and simulation' (Dalmedico, 2001, 249).

If the important lesson from mathematics in the first third of the 19th century was that economics needed to become a deductive science (as geometry was), in the late 19th century the lesson from mathematics was that economics needed to model itself on rational mechanics. Over the first two thirds of the 20th century the lesson was that economics was to become scientific by grounding its models and theories on a modest set of axioms concerning pure economic agents' preferences and choices. But, beginning nearly at mid-century, mathematics was re-imagining itself as a discipline that historically had developed by solving real problems presented to it from other sciences. And in a similar fashion, and partially in response to that changing image of mathematical knowledge, the notion of a serious economic science, connected to data-based reasoning, was reshaping the idea of rigorous argumentation in economics. Econometrics and applied microeconomics were to form the reconstructed core of economic science much as work in algorithmics and applied mathematics were re-commanding attention in the mathematics community. 'At the Berlin International Congress of Mathematicians in August

1998, the old opposition between the pure and the applied – still widely shared in the community – has been formulated in quite different terms: “mathematicians who build models versus those who prove theorems”. [Mumford, 1998]. But the respect enjoyed by the former is now definitely as high as that of the latter’ (Dalmedico, 2001, p. 249). So too in economics, as the prestige accorded ‘good work’ in applied economics now rivals that accorded to work in pure theory.

See Also

- Debreu, Gerard
- existence of general equilibrium
- Fisher, Irving
- Marshall, Alfred
- mathematical economics
- mathematical methods in political economy
- Whewell, William

Bibliography

Corry, L. 1996. *Modern Algebra and the Rise of Mathematical Structures*. Boston: Birkhäuser.

Dalmedico, A. 2001. An image conflict in mathematics after 1945. In *Changing Images in Mathematics: From the French Revolution to the New Millennium*, ed. U. Bottazzini and A. Dalmedico. London and New York: Routledge.

Halmos, P. 1981. Applied mathematics is bad mathematics. In *Mathematics Tomorrow*, ed. L. Steen. New York and Heidelberg: Springer-Verlag.

Hilbert, D. 1918. Axiomatisches Denken. *Mathematische Annalen* 78, 405–15.

Mirowski, P. 1989. *More Heat Than Light*. New York and Cambridge: Cambridge University Press.

Mumford, D. 1998. Trends in the profession of mathematics: choosing our directions. *Berlin Intelligencer*, ICM August 1998, 2–5.

Richards, J. 1988. *Mathematical Visions: the Pursuit of Geometry in Victorian England*. San Diego: Academic Press.

Weintraub, E. 2002. *How Economics Became a Mathematical Science*. Durham, NC: Duke University Press.

Weintraub, E. and Mirowski, P. 1994. The pure and the applied: Bourbakism comes to mathematical economics. *Science in Context* 72, 245–72.

Whewell, W. 1829; 1831; 1850. *Mathematical Exposition of Some Doctrines of Political Economy*. Reprints of Economic Classics. New York: Augustus M. Kelley, 1971.

How to cite this article

Weintraub, E. Roy. "mathematics and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000372> doi:10.1057/9780230226203.1063

The New Palgrave Dictionary of Economics Online

mathematics of networks

M.E.J. Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The patterns of interactions, both economic and otherwise, between individuals, groups or corporations form social networks whose structure can have a substantial effect on economic outcomes. The study of social networks and their implications has a long history in the social sciences and more recently in applied mathematics and related fields. This article reviews the main developments in the area with a focus on practical applications of network mathematics.

Keywords

Bernoulli random graph; centrality measures; graph theory; Milgram, S.; Moreno, J.; network formation; networks, mathematics of; non-Poisson degree distributions; Perron–Frobenius theorem; small worlds; social interactions; social networks

Article

In much of economic theory it is assumed that economic agents interact, directly or indirectly, with all others, or at least that they have the opportunity to do so in order to achieve a desired outcome for themselves. In reality, as common sense tells us, things are quite different. Traders in a market have preferred trading partners, perhaps because of an established history of trust, or simply for convenience. Buyers and sellers have preferred suppliers and customers. Consumers have preferred brands and outlets. And most individuals limit their interactions, economic or otherwise, to a select circle of partners or acquaintances. In many cases partners are chosen not on economic grounds but for social reasons: individuals tend overwhelmingly to deal with others who revolve in the same circles as they do, socially, intellectually or culturally.

The patterns of connections between agents form a social network (Figure 1), and it is intuitively clear that the structure of such networks must affect the pattern of economic transactions, not to mention essentially every other type of social interaction among human beings. Any theory of interaction that ignores these networks is necessarily incomplete. In the last few decades, therefore, researchers have conducted extensive investigations of networks in economics, mathematics, sociology and a number of other fields, in an effort to understand and explain network effects.

Figure 1
A social network of collaborative links. *Note*: The nodes (squares) represent people and the edges (lines) social ties between them.

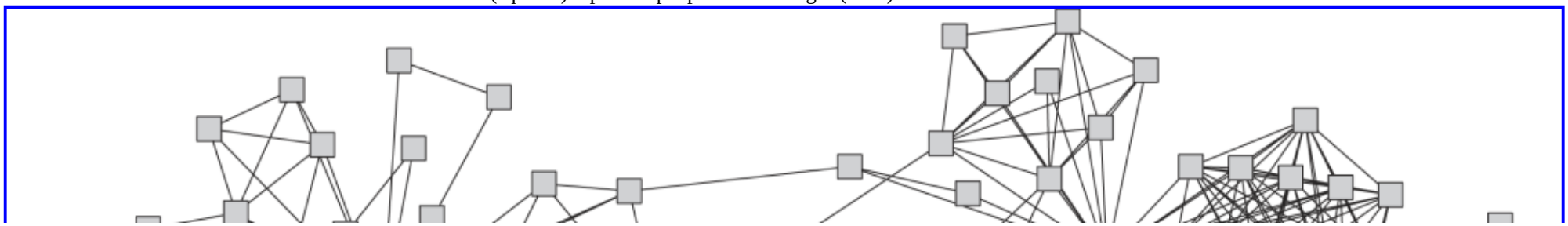

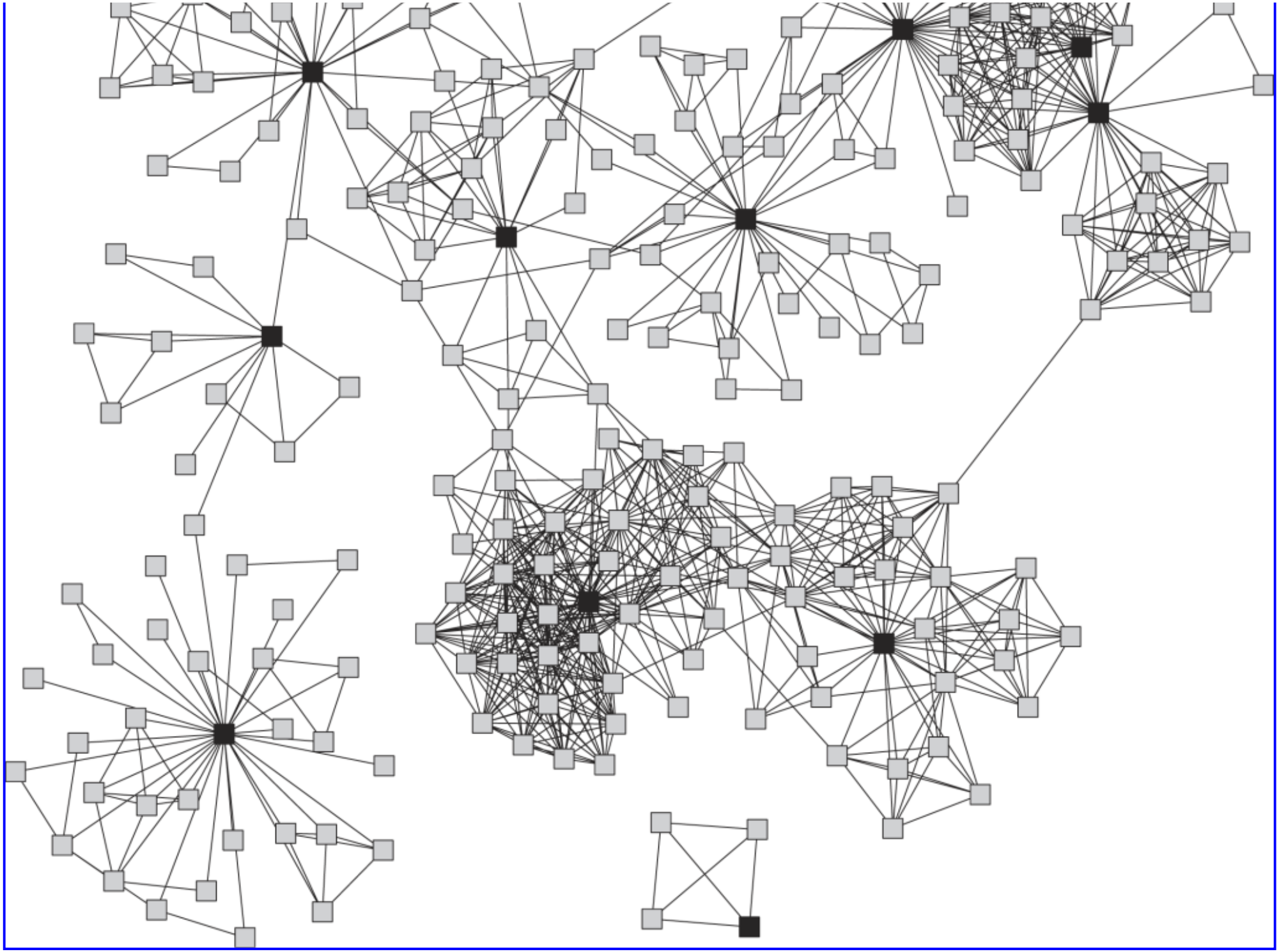

The study of social (and other) networks has three primary components. First, empirical studies of networks probe network structure using a variety of techniques such as interviews,

questionnaires, direct observation of individuals, use of archival records, and specialist tools like ‘snowball sampling’ and ‘ego-centred’ studies. The goal of such studies is to create a picture of the connections between individuals, of the type shown in Figure 1. Since there are many different kinds of possible connections between people – business relationships, personal relationships, and so forth – studies must be designed appropriately to measure the particular connections of interest to the experimenter.

Second, once one has empirical data on a network, one can answer questions about the community the network represents using mathematical or statistical analyses. This is the domain of classical social network analysis, which focuses on issues such as: who are the most central members of a network and who are the most peripheral? Which people have most influence over others? Does the community break down into smaller groups, and if so what are they? Which connections are most crucial to the functioning of a group?

And third, building on the insights obtained from observational data and its quantitative analysis, one can create models, such as mathematical models or computer models, of processes taking place in networked systems – the interactions of traders, for example, or the diffusion of information or innovations through a community. Modelling work of this type allows us to make predictions about the behaviour of a community as a function of the parameters affecting the system.

After a brief historical review, the primary purpose of this article is to describe the mathematical techniques involved in the second and third of these three components: the quantitative analysis of network data and the mathematical modelling of networked systems. Necessarily, this review is short. Much more substantial coverage can be found in the many books and review articles in the field (Wasserman and Faust, 1994; Scott, 2000; West, 1996; Harary, 1995; Ahuja, Magnanti and Orlin, 1993; Dorogovtsev and Mendes, 2003; Albert and Barabási, 2002; Newman, 2003).

History of social network analysis

The study of social networks has roots in the 19th-century beginnings of sociology, especially the ‘gestalt’ tradition of Koehler and others, but is widely regarded as having begun in earnest in the 1930s with the work of psychologist Jacob Moreno, a Romanian immigrant to the United States who had spent a number of years in Vienna and was influenced there by the work of Freud. Moreno advocated an approach to psychoanalysis that involved participants discussing or physically enacting issues that concerned them in front of the analyst. Another approach, which Moreno employed with schoolchildren among others, involved the analyst passively watching participants’ interactions with one another and recording their nature and pattern. In the process of his studies he developed a new tool, the sociogram, which was a map of interactions between individuals drawn on paper as a set of points and lines (Moreno, 1934, p. 38).

In 1933 Moreno presented some of his sociograms during a lecture at a medical conference in New York City, and the work attracted sufficient interest to be featured in the *New York Times*. In everything but name, Moreno's sociograms were what we would now call social networks, and his methods, although strange by today's standards, were the intellectual precursor of social network analysis, which is now a flourishing branch of the social sciences (Wasserman and Faust, 1994). (The term ‘social network’ was not invented until some years later; it is usually credited to John Barnes,1954.)

Apart from a gap during the war years, social network analysis was pursued vigorously following its early popularization. Particularly well-known studies include the ‘southern women’ study of Davis, Davis, and Gardner (1941), Anatol Rapoport's investigations of friendship networks among school children in the 1950s (Rapoport and Horvath, 1961), Pool and Kochen's (1978) mathematical models of social networks that circulated widely in the 1950s and 1960s (although they were not published until much later), and Stanley Milgram's (1967) famous ‘small world’ experiments. Today, social network analysis is one of the standard quantitative tools in the social science toolbox, finding use both in academia and in the business world as a microscope with which to view the details of social interactions.

Mathematics of networks

Turning to the mathematical methods of network analysis, which are the principal focus of this article, let us begin with some simple definitions. A network – also called a *graph* in the mathematics literature – is made up of points, usually called *nodes* or *vertices*, and lines connecting them, usually called *edges*. Mathematically, a network can be represented by a matrix called the *adjacency matrix* **A**, which in the simplest case is an $n \times n$ symmetric matrix, where n is the number of vertices in the network. The adjacency matrix has elements

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } i \text{ and } j, \\ 0 & \text{otherwise}. \end{cases} \tag{1}$$

The matrix is symmetric since if there is an edge between *i* and *j* then clearly there is also an edge between *j* and *i*. Thus $A_{ij} = A_{ji}$.
In some networks the edges are *weighted*, meaning that some edges represent stronger connections than others, in which case the nonzero elements of the adjacency matrix can be generalized to values other than unity to represent stronger and weaker connections. Another variant is the *directed network*, in which edges point in a particular direction between two vertices. For instance, in a network of cash sales between buyers and sellers the directions of edges might represent the direction of the flow of goods (or conversely of money) between individuals. Directed networks can be represented by an asymmetric adjacency matrix in which $A_{ij} = 1$ implies the existence (conventionally) of an edge pointing from *j* to *i* (note the direction), which will in general be independent of the existence of an edge from *i* to *j*.
Networks may also have *multiedges* (repeated edges between the same pair of vertices), *self-edges* (edges connecting a vertex to itself), *hyperedges* (edges that connect more than two vertices together) and many other features. We here concentrate primarily on the simplest networks, having undirected, unweighted single edges between pairs of vertices.

Centrality measures

Now let us consider the analysis of network data. We start by looking at *centrality measures*, which are some of the most fundamental and frequently used measures of network structure. Centrality measures address the question, 'Who is the most important or central person in this network?' There are many answers to this question, depending on what we mean by 'important'. Perhaps the simplest of centrality measures is *degree centrality*, also called simply *degree*. The degree of a vertex in a network is the number of edges attached to it. In mathematical terms, the degree k_i of a vertex *i* is

$$k_i = \sum_{j=1}^{n} A_{ij}. \tag{2}$$

Though simple, degree is often a highly effective measure of the influence or importance of a node: in many social settings people with more connections have more power.
A more sophisticated version of the same idea is the so-called *eigenvector centrality*. Where degree centrality gives a simple count of the number of connections a vertex has, eigenvector centrality acknowledges that not all connections are equal. In general, connections to people who are themselves influential will lend a person more influence than connections to less influential people. If we denote the centrality of vertex *i* by x_i, then we can allow for this effect by making x_i proportional to the average of the centralities of *i*'s network neighbours:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j, \tag{3}$$

where λ is a constant. Defining the vector of centralities $\mathbf{x} = (x_1, x_2, \ldots)$, we can rewrite this equation in matrix form as

$$\lambda \mathbf{x} = \mathbf{A} \cdot \mathbf{x}, \tag{4}$$

and hence we see that **x** is an eigenvector of the adjacency matrix with eigenvalue λ. On the assumption that we wish the centralities to be non-negative, it can be shown (using the

Perron–Frobenius theorem) that λ must be the largest eigenvalue of the adjacency matrix and **x** the corresponding eigenvector.

The eigenvector centrality defined in this way accords each vertex a centrality that depends on both the number and the quality of its connections: having a large number of connections still counts for something, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts. Eigenvector centrality turns out to be a revealing measure in many situations. For example, a variant of eigenvector centrality is employed by the well-known Web search engine Google to rank Web pages, and works well in that context.

Two other useful centrality measures are *closeness centrality* and *betweenness centrality*. Both are based upon on the concept of network paths. A path in a network is a sequence of vertices traversed by following edges from one vertex to another across the network. A *geodesic path* is the shortest path, in terms of number of edges traversed, between a specified pair of vertices. (Geodesic paths need not be unique; two or more paths can tie for the title of shortest.) The closeness centrality of vertex *i* is the mean geodesic distance (that is, the mean length of a geodesic path) from vertex *i* to every other vertex. Closeness centrality is *lower* for vertices that are more central in the sense of having a shorter network distance on average to other vertices. (Some writers define closeness centrality to be the reciprocal of the average so that higher numbers indicate greater centrality. Also, some vertices may not be reachable from vertex *i* – two vertices can lie in separate ‘components’ of a network, with no connection between the components at all. In this case closeness as above is not well defined. The usual solution to this problem is simply to define closeness to be the average geodesic distance to all *reachable* vertices, excluding those to which no path exists.)

The betweenness centrality of vertex *i* is the fraction of geodesic paths between other vertices that *i* lies on. That is, we find the shortest path (or paths) between every pair of vertices, and ask on what fraction of those paths vertex *i* lies. Betweenness is a crude measure of the control *i* exerts over the flow of information (or any other commodity) between others. If we imagine information flowing between all pairs of individuals in the network and always taking the shortest possible path, then betweenness centrality measures the fraction of that information that will flow through *i* on its way to wherever it is going. In many social contexts a vertex with high betweenness will exert substantial influence by virtue not of being in the middle of the network (although it may be) but of lying ‘between’ other vertices in this way. It is in most cases only an approximation to assume that information flows along geodesic paths; normally it will not, and variations of betweenness centrality such as ‘flow betweenness’ and ‘random walk betweenness’ have been proposed to allow for this. In many practical cases, however, the simple (geodesic path) betweenness centrality gives quite informative answers.

Other network properties

The study of shortest paths on networks also leads to another interesting network concept, the *small-world effect*. It is found that in most networks the mean geodesic distance between vertex pairs is small compared with the size of the network as a whole. In a famous experiment conducted in the 1960s, the psychologist Stanley Milgram (1967) asked participants (located in the United States) to get a message to a specified target person elsewhere in the country by passing it from one acquaintance to another, stepwise through the population. Milgram's remarkable finding that the typical message passed though just six people on its journey between (roughly) randomly chosen initial and final individuals has been immortalized in popular culture in the phrase ‘six degrees of separation’, which was the title of a 1990 Broadway play by John Guare in which one of the characters discusses the small-world effect. Since Milgram's experiment, the small-world effect has been confirmed experimentally in many other networks, both social and nonsocial.

Other network properties that have attracted the attention of researchers in recent years include network *transitivity* or *clustering* (the tendency for triangles of connections to appear frequently in networks – in common parlance, ‘the friend of my friend is also my friend’), vertex similarity (the extent to which two given vertices do or do not occupy similar positions in the network), communities or groups within networks and methods for their detection, and, crucially, the distribution of vertex degrees, a topic discussed in more detail below.

Models of networks

Turning to models of networks and of the behaviour of networked systems, we find that perhaps the simplest useful model of a network (and one of the oldest) is the *Bernoulli random graph*, often called just the *random graph* for short (Solomonoff and Rapoport, 1951; Erd•s and Rényi, 1960; Bollobás, 2001). In this model one takes a certain number of vertices *n* and creates edges between them with independent probability *p* for each vertex pair. When *p* is small there are only a few edges in the network, and most vertices exist in isolation or in small groups of connected vertices. Conversely, for large *p* almost every possible edge is present between the $\binom{n}{2}$ possible vertex pairs, and all or almost all of the vertices join together in a single large connected group. One might imagine that for intermediate values of *p* the sizes of groups would just grow smoothly from small to large, but this is not the case. It is found instead that there is a *phase transition* at a special value $p = 1/n$ above which a *giant component* forms, a group of connected vertices occupying a fixed fraction of the whole network, i.e., with size varying as *n*. For values of *p* less than this, only small groups of vertices exist of a typical size that is independent of *n*. Many real-world networks show behaviour reminiscent of this model, with a large component of connected vertices filling a sizable fraction of the entire network, the remaining vertices falling in much smaller components that are unconnected to the rest of the network.

The random graph has a major shortcoming, however: the distribution of the degrees of the vertices is quite unlike that seen in most real-world networks. The fraction p_k of vertices in

a random graph having degree *k* is given by the binomial distribution, which becomes Poisson in the limit of large *n*:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \to \frac{z^k e^{-z}}{k!}, \tag{5}$$

where $z = (n-1)p$ is the mean degree. Empirical observations of real networks, social and otherwise, show that most have highly non-Poisson distributions of degree, often heavily right-skewed with a fat tail of vertices having unusually high degree (Albert and Barabási, 2002; Dorogovtsev and Mendes, 2003). These high-degree nodes or 'hubs' in the tail can, it turns out, have a substantial effect on the behaviour of a networked system.

To allow for non-Poisson degree distributions, one can generalize the random graph, specifying a particular, arbitrary degree distribution p_k and then forming a graph that has that distribution but is otherwise random. A simple algorithm for doing this is to choose the degrees of the *n* vertices from the specified distribution, draw each vertex with the appropriate number of 'stubs' of edges emerging from it, and then pick stubs in pairs uniformly at random and connect them to create complete edges. The resulting model network (or more properly the ensemble of such networks) is called the *configuration model*.

The configuration model also shows a phase transition, similar to that of the Bernoulli random graph, at which a giant component forms. To see this, consider a set of connected vertices and consider the 'boundary vertices' that are immediate neighbours of that set. Let us grow our set by adding the boundary vertices to it one by one. When we add one boundary vertex to our set the number of boundary vertices goes down by 1. However, the number of boundary vertices also increases by the number of new neighbours of the vertex added, which is one less than the degree *k* of that vertex. Thus the total change in the number of boundary vertices is $-1 + (k-1) = k-2$. However, the probability of a particular vertex being a boundary vertex is proportional to *k*, since there are *k* times as many edges by which a vertex of degree *k* could be connected to our set than there are for a vertex of degree 1. Thus the average change in the number of boundary vertices when we add one vertex to our set is a weighted average $\Sigma_i k_i(k_i-2) / \Sigma_j k_j = \Sigma_i k_i(k_i-2) / (nz)$, where *z* is again the mean degree. If this quantity is less than zero, then the number of boundary vertices dwindles as our set grows bigger and will in the end reach zero, so that the set will stop growing. Thus in this regime all connected sets of vertices are of finite size. If on the other hand this number is greater than zero, then the number of boundary vertices will grow without limit, and hence the size of our set of connected vertices is limited only by the size of the network.

Thus, a giant component exists in the network if and only if

$$\langle k^2 \rangle - 2\langle k \rangle > 0, \tag{6}$$

where $\langle k \rangle = z = n^{-1}\Sigma_i k_i$ is the mean degree and $\langle k^2 \rangle = n^{-1}\Sigma_i k_i^2$ is the mean-square degree.

The mean-square degree appears over and over in the mathematics of networks. Another context in which it appears is in the spread of information (or anything else) over a network. Taking a simple model of the spread of an idea (or a rumour or a disease), imagine that each person who has heard an idea communicates it with independent probability *q* to each of his or her friends. If the person's degree is *k*, then there are $k-1$ friends to communicate the idea to, not counting the one from whom he or she heard it in the first place, so the expected number who hear it is $q(k-1)$. Performing the weighted average over vertices again, we find that the average number of people a person passes the idea onto, also called the *basic reproductive number* R_0, is

$$R_0 = q\frac{\Sigma_i k_i(k_i-1)}{\Sigma_i k_i} = q\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}. \tag{7}$$

If R_0 is greater than 1, then the number of people hearing the idea grows as it gets passed around and it will take off exponentially. If R_0 is less than 1 then the idea will die. Again, we have a phase transition, or *tipping point*, for the spread of the idea: it spreads if and only if

$$q > \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}. \tag{8}$$

The simple understanding behind the appearance of the mean-square degree in this expression is the following. If a person with high degree hears this idea he or she can spread it to many others, by virtue of having many friends. However, such a person is also more likely to hear the idea in the first place because of having many friends to hear it from. Thus, the degree enters twice into the process: a person with degree 10 is $10 \times 10 = 100$ times more effective at spreading the idea than a person with degree 1.

The appearance of the mean-square degree in expressions like (6) and (8) can have substantial effects. Of particular interest are networks whose degree distributions have fat tails. It is possible for such networks to have very large values of $\langle k^2 \rangle$ – in the hundreds or thousands – so that, for example, the right-hand side of eq. (8) is very small. This means that the probability of each individual person spreading an idea (or rumour or disease) need not be large for it still to spread through the whole community.

Another important class of network models is the class of generative models, models that posit a quantitative mechanism or mechanisms by which a network forms, usually as a way of explaining how the observed structure of the network arises. The best-known example of such a model is the ‘cumulative advantage’ or ‘preferential attachment’ model (Price, 1976; Barabási and Albert, 1999), which aims to explain the fat-tailed degree distributions observed in some networks. In its simplest form this model envisages a network that grows by the steady addition of vertices, one at a time. Many networks, such as the World Wide Web and citation networks, grow this way; it is a matter of current debate whether the model applies to social networks as well. Each vertex is added with a certain number *m* of edges emerging from it, whose other ends connect to pre-existing vertices with probability proportional to those vertices’ current degree. That is, the higher the current degree of a vertex, the more likely that vertex is to acquire new edges when the graph grows. This kind of rich-get-richer phenomenon is plausible in many network contexts and is known to generate Pareto degree distributions. Using a rate-equation method (Price, 1976; Simon, 1955; Krapivsky, Redner and Leyvraz, 2000), we find that in the limit of large network size the degree distribution obeys:

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}. \tag{9}$$

This distribution has a tail going as $p_k \sim k^{-3}$ in the large-*k* limit, which is strongly reminiscent of the degree distributions seen particularly in citation networks and also in the World Wide Web. Generative models of this type have been a source of considerable interest in recent years and have been much extended by a number of authors (Dorogovtsev and Mendes, 2003; Albert and Barabási, 2002) beyond the simple ideas described here.

Concepts such as those appearing in this article can be developed a great deal further and lead to a variety of useful, and in some cases surprising, results about the function of networked systems. More details can be found in the references.

See Also

- artificial neural networks

- business networks
- graph theory
- interacting agents in finance
- network formation
- Pareto distribution
- Perron–Frobenius theorem
- power laws
- psychology of social networks
- small-world networks
- social networks in labour markets

Bibliography

Ahuja, R., Magnanti, T. and Orlin, J. 1993. *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ: Prentice Hall.

Albert, R. and Barabási, A.-L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 47–97.

Barabási, A.-L. and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509–12.

Barnes, J. 1954. Class and committees in a Norwegian island parish. *Human Relations* 7, 39–58.

Bollobás, B. 2001. *Random Graphs*, 2nd edn. New York: Academic Press.

Davis, A., Gardner, B. and Gardner, M. 1941. *Deep South*. Chicago: University of Chicago Press.

Dorogovtsev, S. and Mendes, J. 2003. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford: Oxford University Press.

Erd•s, P. and Rényi, A. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 17–61.

Harary, F. 1995. *Graph Theory*. Cambridge, MA: Perseus.

Krapivsky, P., Redner, S. and Leyvraz, F. 2000. Connectivity of growing random networks. *Physical Review Letters* 85, 4629–32.

Milgram, S. 1967. The small world problem. *Psychology Today* 2, 60–7.

Moreno, J. 1934. *Who Shall Survive?* Beacon, NY: Beacon House.

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45, 167–256.

Pool, I. and Kochen, M. 1978. Contacts and influence. *Social Networks* 1, 1–48.

Price, D. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of American Society for Information Science* 27, 292–306.

Rapoport, A. and Horvath, W. 1961. A study of a large sociogram. *Behavioral Science* 6, 279–91.

Scott, J. 2000. *Social Network Analysis:* A Handbook, 2nd edn. London: Sage.

Simon, H. 1955. On a class of skew distribution functions. *Biometrika* 42, 425–40.

Solomonoff, R. and Rapoport, A. 1951. Connectivity of random nets. *Bulletin of Mathematical Biophysics* 13, 107–17.

Wasserman, S. and Faust, K. 1994. *Social Network Analysis*. Cambridge: Cambridge University Press.

West, D. 1996. *Introduction to Graph Theory*. Upper Saddle River, NJ: Prentice Hall.

How to cite this article

Newman, M. E. J. "mathematics of networks." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000361> doi:10.1057/9780230226203.1064

The New Palgrave Dictionary of Economics Online

maximum likelihood

Jack R. Porter

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Maximum likelihood is a method of estimation developed for fully specified parametric likelihood settings. In smooth parametric models, maximum likelihood has a number of desirable properties, including consistency, asymptotic normality, and asymptotic efficiency. Maximum likelihood has been usefully extended to various semiparametric and nonparametric settings.

Keywords

asymptotic normality; bootstrap; confidence region; consistency; EM algorithm; empirical likelihood; Fisher, R. A.; generalized method of moments; invariance; law of large numbers; likelihood principle; local likelihood; log likelihood ratio; maximum likelihood; nonparametric regression; semiparametric estimation; statistical inference; sufficiency

Article

Given data from some member of a parametric family of distributions, maximum likelihood provides a general purpose method of estimation frequently accompanied by useful statistical properties.
In a series of papers, R.A. Fisher (1922; 1925; 1934) proposed and argued for a method of estimation he dubbed ‘maximum likelihood’. The intuitive appeal and broad applicability continue to drive its use as a primary tool of statisticians. Suppose data $z = (z_1, \ldots, z_n)$ is drawn from a distribution with density $f_n(z; \theta_0)$, and further suppose that this distribution is a member of a family of parametric distributions with densities $\{f_n(z; \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ (for k finite and $\theta_0 \in \Theta$). The likelihood function is simply defined by the joint density as the function $l_n(\theta, z) = f_n(z; \theta)$ with argument θ and data z held fixed. The maximum likelihood estimator (MLE) is then defined as

$$\hat{\theta}_{ML} = \arg\min_{\theta \in \Theta} l_n(\theta, z).$$

One motivation for the MLE comes from the likelihood principle, which implies that statistical inference on θ given data *z* should be based solely on the likelihood function, $l_n(\theta, z)$ (Berger and Wolpert, 1988). According to this principle the relative evidence on two different values of θ given by the data is fully summarized by their likelihood ratio. In this sense, the MLE is the value of θ most supported by the data. Of course, most econometric work involving maximum likelihood does not take a strict likelihood principle viewpoint, but is typically more concerned with sampling properties from a frequentist viewpoint. It is this perspective that will be our main emphasis in what follows.

It is often convenient to (equivalently) think of the MLE as maximizing the log likelihood ratio, $\mathcal{L}_n(\theta) = \ln \frac{f_n(z;\theta)}{f_n(z;\theta_0)}$. When the data is independent and identically distributed (i.i.d.) with marginal density *f*, the log likelihood ratio can be written $\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \ln \frac{f(z_i;\theta)}{f(z_i;\theta_0)}$. By the law of large numbers, the normalized log likelihood ratio $(\frac{1}{n}\mathcal{L}_n(\theta))$ approaches $\mathcal{L}(\theta) = E_{\theta_0}[\ln \frac{f(z_i;\theta)}{f(z_i;\theta_0)}]$ (where the expectation is taken with respect to the 'population' density $f(z_i; \theta_0)$) asymptotically. Though $-\mathcal{L}(\theta)$ does not satisfy the formal definition of metric, it is often taken as distance measure between the densities $f(z_i;\cdot\theta)$ and $f(z_i;\cdot\theta_0)$. Not surprisingly, this distance is minimized at $\theta = \theta_0$ (when the identification condition that $f(z;\theta) \neq f(z;\theta_0)$ for $\theta \neq \theta_0$ is satisfied). The log likelihood ratio $\mathcal{L}_n(\theta)$ can be interpreted as a sample approximation to this discrepancy measure, which is minimized at the MLE. The likelihood ratio test statistic, for testing the null hypothesis that $\theta = \theta_0$, is also based on this value.

Fisher emphasized the usefulness of the maximized likelihood itself. The density $f_n(z;\hat{\theta}_{ML})$ provides an approximation to the population density. If, for instance, there is interest in some feature of $f_n(z;\cdot\theta_0)$, then an approximation can often be obtained from the corresponding feature of $f_n(z;\hat{\theta}_{ML})$ (as in the parametric bootstrap). More generally, Efron (1982) notes that $f_n(z;\hat{\theta}_{ML})$ acts as a data summary.

Properties

For most commonly used parametric distributional families, the MLE is consistent. Note, for instance, that when $\mathcal{L}(\theta)$ is maximized at θ_0, and the convergence of the log likelihood ratio $\frac{1}{n}\mathcal{L}(\theta)$ mentioned above is uniform on Θ, then $\hat{\theta}_{ML}$ will correspondingly converge to θ_0. More general sufficient conditions for consistency are also available; see Ibragimov and Has'minskii (1981).

Under appropriate regularity conditions (which essentially amount to smoothness of the parametric model), the MLE is asymptotically normal.

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, J(\theta_0)^{-1})$$

where $J(\theta)$ is the Fisher information matrix, with value $E_\theta[\nabla_\theta \ln f(z;\theta)(\nabla_\theta \ln f(z;\theta))']$ when the data is i.i.d. and $\nabla_\theta \ln f(z_i;\cdot\theta)$ is called the 'score function'. Define the Hessian as $H(\theta) = E_\theta[\nabla_{\theta\theta} \ln f(z_i;\theta)]$. By the information matrix equality, $J(\theta) = -H(\theta)$, which adds to the variety of estimators for the information matrix. Frequentist confidence intervals are then immediately available based on the asymptotic normality property and an estimator for $J(\theta_0)$.

Other approximations for the distribution of the MLE are also available for certain statistical models. Barndorff-Nielsen (1983) provides an accurate approximation to the conditional distribution of the MLE given a maximal ancillary statistic. (An ancillary statistic is a statistic whose distribution does not depend on θ, and if every ancillary statistic is a function of a given ancillary statistic then that statistic is called maximal.) When the MLE is sufficient, Barndorff-Nielsen's formula is exact. For non-regular models, asymptotic normality may no longer hold and a general limiting distribution result is then unavailable. Ibragimov and Has'minskii (1981), for instance, characterize the asymptotic behavior of the MLE for certain non-regular classes of models.

The Fisher information matrix is additionally useful as an efficiency bound. Accordingly, the MLE itself enjoys certain optimality properties. For regular models, the MLE is asymptotically efficient under classical criteria. Hirano and Porter (2005) show that a shifted version of the MLE is asymptotically efficient for an even broader class of statistical models (and allow for asymmetric loss). Higher-order efficiency of the MLE has been established in Pfanzagl and Wefelmeyer (1978).

Intuition for the asymptotic normality and efficiency of the MLE can be gained through a consideration of the behaviour of the log likelihood ratio in the i.i.d. case. If we re-parametrize the likelihood in terms of the 'local' parameter $h = \sqrt{n}(\theta - \theta_0)$, then with enough smoothness the log likelihood ratio can be expanded as follows

$$\sum_{i=1}^{n} \ln \frac{f(z_i;\theta_0 + h/\sqrt{n})}{f(z_i;\theta_0)} \approx \frac{h'}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \ln f(z;\theta_0) + \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^{n} h' \nabla_{\theta\theta} \ln f(z;\theta_0) h.$$

Under regularity conditions, the log likelihood ratio converges in distribution (for each h) to $N(-\frac{1}{2}h'J(\theta_0)h, h'J(\theta_0)h)$. (This kind of 'Taylor' expansion actually holds under a mild condition of differentiability in quadratic mean which is weaker than the twice continuous differentiability of the likelihood that appears necessary.) Models with log likelihood ratios obeying this kind of convergence are called 'locally asymptotically normal'. Now, consider the statistical model consisting of a single observation on a random variable $X \sim N(h, J(\theta_0)^{-1})$. Notably, the log likelihood ratio for this simple

statistical model $\{N(h, J(\theta_0)^{-1}) : h \in \mathbb{R}^k\}$ has the same distribution as the asymptotic distribution for the log likelihood ratio of the general model above. Since the log likelihood ratio captures all the statistical information in a given statistical model, there is an equivalence between the asymptotic behaviour of the original model with densities $\prod_{i=1}^{n} f(z_i; \theta)$ and the much simpler model given by a single observation from a normal with unknown mean *h* (and known variance-covariance, $J(\theta_0)^{-1}$). This equivalence is formalized in the limits of experiments theory (Le Cam, 1986). Intuitively, one might expect that the MLE for the local parameter $\hat{h}_{ML} = \sqrt{n}(\hat{\theta}_{ML} - \theta_0)$ in the original model will behave (asymptotically) like the MLE of the 'limit' normal model, which is simply given by *X*. The normality of *X* and the efficiency (minimax) of the mean in a normal model then corresponds to the asymptotic normality and asymptotic efficiency of the MLE in the original model.

Other important properties of the MLE are invariance and sufficiency. The MLE is necessarily a function of all sufficient statistics. The MLE is also invariant to parametrization of the family of distributions. So, if the distributions are re-parametrized in terms of $\lambda = T(\theta)$, then $\hat{\lambda}_{ML} = T(\hat{\theta}_{ML})$. Additionally, the MLE satisfies a group equivariance property (Eaton, 1989). Suppose the family of distributions is invariant under the group of transformations $\mathcal{G}$ defined on both the sample and parameter spaces. If $g \in \mathcal{G}$, then $\hat{\theta}_{ML}(gz) = g\hat{\theta}_{ML}(z)$, where the MLE is written as a function of the observations.

Limitations

Since densities are not uniquely defined, the likelihood criterion on which the MLE is based is not uniquely defined. For a given likelihood, a solution to the maximization problem that defines the MLE need not necessarily exist (or multiple solutions are also possible).

The consistency, asymptotic normality and efficiency properties (discussed above) are all asymptotic, leaving the possibility that the small sample behaviour of the MLE may be quite poor in given applications. Even the asymptotic properties themselves are assured under regularity conditions. Neyman and Scott (1948) describe a famous example where maximum likelihood can be poorly behaved in small samples. The random variables x_{ij} are distributed $N(\mu_i, \sigma^2)$ for $i = 1, \ldots, n, \quad j = 1, \ldots, J$, and all random variables are independent. Consider the case with fixed *n* and $J = 2$. Since $x_{i2} - x_{i1}$ is distributed $N(0, 2\sigma^2)$, $s_n^2 = \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}(x_{i2} - x_{i1})^2$ is a natural and reasonable estimator for σ^2. But $\hat{\sigma}^2_{ML} = \frac{1}{2}s_n^2$, which could be a quite poor estimator with significant bias. This poor small sample performance is particularly notable, since this model consists only of independent normally distributed random variables. Asymptotically, if *n* remains fixed and *J* grows, then the MLE has all the usual favourable large sample properties. If *J* is fixed and *n* grows, then the assumption of a finite dimensional parameter space is violated, and the MLE is not even consistent.

Stein's well-known shrinkage estimator shows that, even in a simple normal model with known variance-covariance and unknown mean, the MLE need not be (mean-squared error) optimal. It is also notable that, outside of regular models, asymptotic efficiency of the MLE can frequently fail. A simple example of such a non-regular model is data drawn from a uniform distribution on $[0, \theta]$. More general, parameter-dependent support models can be found in the auction literature, and the MLE is generally

suboptimal by traditional asymptotic efficiency criteria (Hirano and Porter, 2003). Le Cam (1990) lists a number of additional examples where the deficiencies of maximum likelihood are highlighted.

Extensions

Suppose the parameter is partitioned $\theta' = (\theta_1', \theta_2')$, and we define $\theta_2^*(\theta_1) = \arg\max_{\theta_2} l_n(\theta_1, \theta_2)$. Then, the profiled likelihood, $l_n(\theta_1, \theta_2^*(\theta_1))$ can be maximized to give the MLE for θ_1. Sometimes this is useful for computational purposes. This formulation has also been useful for conceptual purposes, such as developing semiparametric efficiency bounds, where θ_2 contains nuisance parameters.
Maximum likelihood theory also extends immediately to conditional likelihood formulations. Other methods have been developed to ease the computational burden of maximum likelihood in certain problems. The EM algorithm can be especially helpful in missing data cases (MacLachlan and Krishnan, 1997). Simulated maximum likelihood is useful when the likelihood can be expressed as high-dimensional integral without a closed form solution (Hajivassiliou and Ruud, 1994).
A natural concern with maximum likelihood is its reliance on correct specification of the family of distributions. Quasi-likelihood methods suggest parametric families that have robustness properties beyond the family specified. Exponential linear families often play a prominent role in this approach (Gourieroux, Monfort and Trognon, 1984). Typically, efficiency is sacrificed, but consistency and asymptotic normality still hold where the asymptotic variance of the limiting normal distribution is given by the ‘sandwich’ formula, $H(\theta_0)^{-1}J(\theta_0)H(\theta_0)^{-1}$.
Extensions of maximum likelihood have also been usefully applied in semiparametric and nonparametric contexts. Ai (1997) considers semiparametric estimation in a model with unknown conditional density that is assumed only to satisfy an index restriction. The conditional density is estimated nonparametrically, and the corresponding score function is constructed to produce a semiparametric maximum likelihood estimate of a finite-dimensional parameter of the model. Tibshirani and Hastie (1987) introduced the notion of local likelihood estimation where regression functions are fit locally according to a maximum likelihood criterion. This idea has been extended to density estimation and other regression-type settings (Fan, Farmen and Gijbels 1998). Linton and Xiao (2007) develop an adaptive nonparametric regression approach that estimates the unknown density of the disturbance (or its score function) and then uses this estimate for local likelihood estimation of the unknown regression function. Empirical likelihood methods are another offshoot of nonparametric maximum likelihood. The basic insight that the empirical distribution function is the nonparametric MLE for a general cumulative distribution function has led to new approaches to confidence region formation, estimation in regression models, generalized method of moments inference, and bootstrapping (Owen, 2001; Brown and Newey, 2002).

See Also

- classical distribution theories

- econometrics
- efficiency bounds
- empirical likelihood
- Fisher, Ronald Aylmer
- nonparametric structural models
- optimality and efficiency
- semiparametric estimation
- statistical inference

Bibliography

Ai, C. 1997. A semiparametric maximum likelihood estimator. *Econometrica* 65, 933–63.

Barndorff-Nielsen, O.E. 1983. On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343–65.

Berger, J. and Wolpert, R. 1988. *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.

Brown, B. and Newey, W. 2002. Generalized method of moments, efficient bootstrapping, and improved inference. *Journal of Business and Economic Statistics* 20, 507–17.

Eaton, M. 1989. *Group Invariance Applications in Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Efron, B. 1982. Maximum likelihood and decision theory. *Annals of Statistics* 10, 340–56.

Fan, J., Farmen, M. and Gijbels, I. 1998. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society* (Series B) 60, 591–608.

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* A 222, 309–60.

Fisher, R.A. 1925. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* 22, 700–25.

Fisher, R.A. 1934. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London* A 144, 285–307.

Gourieroux, C., Monfort, A. and Trognon, A. 1984. Pseudo maximum likelihood methods: theory.

Econometrica 52, 681–700.

Hajivassiliou, V. and Ruud, P. 1994. Classical estimation methods for LDV models using simulation. *Handbook of Econometrics*, vol. 4, ed. D. McFadden and R. Engle. Amsterdam: North-Holland.

Hirano, K. and Porter, J. 2003. Asymptotic efficiency in parametric structural models with parameter dependent support. *Econometrica* 71, 1307–38.

Hirano, K. and Porter, J. 2005. Efficiency in asymptotic shift models. Working paper, University of Wisconsin.

Ibragimov, I.A. and Has'minskii, R.Z. 1981. *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag.

Le Cam, L. 1986. *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Le Cam, L. 1990. Maximum likelihood: an introduction. *International Statistical Review* 58, 153–71.

Linton, O. and Xiao, Z. 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory* 23(3) (forthcoming).

MacLachlan, G. and Krishnan, T. 1997. *The EM Algorithm and Extensions*. New York: Wiley.

Neyman, J. and Scott, E.L. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.

Owen, A.B. 2001. *Empirical Likelihood*. New York: Chapman-Hall.

Pfanzagl, J. and Wefelmeyer, W. 1978. A Third-order optimum property of the maximum likelihood estimator. *Journal of Multivariate Analysis* 8, 1–29.

Tibshirani, R. and Hastie, T. 1987. Local likelihood estimation. *Journal of the American Statistical Association* 82, 559–67.

How to cite this article

Porter, Jack R. "maximum likelihood." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000112> doi:10.1057/9780230226203.1065

The New Palgrave Dictionary of Economics Online

maximum score methods

Robert P. Sherman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes some aspects of maximum score estimation of parameters of multinomial and, especially, binomial choice models. In the context of binomial choice models, strengths and weaknesses of the estimation procedure are discussed, as well as its relation to classical quantile regression estimation and its nonstandard rate of convergence. The benefits of smoothing the score criterion function are also noted.

Keywords

binary response models; bootstrap; central limit theorems; heteroskedasticity; linear median regression; maximum likelihood; maximum score methods; multinomial choice models; quantile regression; random utility maximization; semiparametric estimation

Article

In a seminal paper, Manski (1975) introduces the maximum score estimator (MSE) of the structural parameters of a multinomial choice model and proves consistency without assuming knowledge of the distribution of the error terms in the model. As such, the MSE is the first instance of a semiparametric estimator of a limited dependent variable model in the econometrics literature.

Maximum score estimation of the parameters of a binary choice model has received the most attention in the literature. Manski (1975) covers this model, but Manski (1985) focuses on it. The key assumption that Manski (1985) makes is that the latent variable underlying the observed binary data satisfies a linear α-quantile regression specification. (He focuses on the linear median regression case, where $\alpha=0.5$.) This is perhaps an under-appreciated fact about maximum score estimation in the binary choice setting. If the latent variable were observed, then classical quantile regression estimation (Koenker and Bassett, 1978), using the latent data, would estimate, albeit more efficiently, the same regression parameters that would be estimated by maximum score estimation using the binary data. In short, the estimands would be the same for these two estimation procedures. Assuming that the underlying latent variable satisfies a linear α-quantile regression specification is equivalent to assuming that the regression parameters in the linear model do not depend on the regressors and that the error term in the model has zero α-quantile conditional on the regressors. Under these assumptions, Manski (1985) proves strong consistency of the MSE. The zero conditional α-quantile assumption does not require the existence of any error moments and allows heteroskedastic errors of an unknown form. This flexibility is in

contrast to many semiparametric estimators of comparable structural parameters for the binary choice model. As discussed in Powell (1994), many of these latter estimators require the existence of error moments and most require more restrictive assumptions governing the relation of errors to regressors.

The weak zero conditional α-quantile assumption comes at a price, however. Extrapolation power is limited: off the observed support of the regressors it is not possible to identify the conditional probability of the choice of interest, but only whether this probability is above or below $1-\alpha$. See Manski (1995, pp. 149–50). There are also disadvantages associated with the estimation procedure. The maximum score criterion function is a sum of indicator functions of sets involving parameters. This lack of smoothness precludes using standard optimization routines to compute the MSE. Moreover, Kim and Pollard (1990) show that this type of discontinuity leads to a convergence rate of $n^{-1/3}$ rather than the $n^{-1/2}$ convergence rate attained by most semiparametric estimators of parameters in this model. In addition, Kim and Pollard (1990) show that the MSE has a nonstandard limiting distribution. The properties of this distribution are largely unknown, making asymptotic inference problematic. Also, Abrevaya and Huang (2005) prove that the bootstrapped MSE is an inconsistent estimator of the parameters of interest, precluding bootstrap inference.

To repair some of these shortcomings, Horowitz (1992) develops a smoothed MSE (SMSE) for the linear median regression case. This estimator retains the attractive flexibility properties of the MSE, but can be computed using standard optimization routines. In addition, the SMSE converges at a faster rate than the MSE and has a normal limit law allowing first order asymptotic inference. Horowitz (2002) proves that bootstrapped SMSE provides asymptotic refinements and in various simulations demonstrates the superiority of bootstrap tests over first-order asymptotic tests. Kordas (2006) generalizes Horowitz's (1992) SMSE to cover all α-quantiles.

In the next section, we present the multinomial choice model under random utility maximization as well as some intuition behind maximum score estimation in this context. We then discuss the relation between maximum score estimation in the binary response model and quantile regression. Next, we present Kim and Pollard's (1990) heuristic argument for the nonstandard rate of convergence of the MSE in the binary model. Finally, we discuss the method of Horowitz (1992) for smoothing the MSE.

The random utility maximization model of choice and the MSE

Manski (1975) developed the MSE for the multinomial choice model in the context of random utility maximization. Suppose the ith individual in a sample of size n from a population of interest must make exactly one of J choices, where $J \geq 2$.

For $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, J\}$, let U_{ij} denote the utility to individual i of making choice j. Assume the structural form $U_{ij} = X_{ij}'\beta + \varepsilon_{ij}$ where X_{ij} is an observable $m \times 1$ vector of explanatory variables, β is a unknown $m \times 1$ parameter vector, and ε_{ij} is an unobservable random disturbance. (A more general set-up can be accommodated. For example, there can be a different parameter vector associated with each choice.)

The utilities associated with the choices an individual faces are latent, or unobservable. However, an individual's choice is observable. Suppose we adopt the maximum utility model of choice: if individual i makes choice j then $U_{ij} > U_{ik}$ for all $k \neq j$. For any event E, define the indicator function $\{E\}=1$ if E occurs and 0 otherwise. Define

$$Y_{ij} = \left\{U_{ij} > U_{ik}, \text{ for all } k \neq j\right\} = \left\{X_{ij}'\beta + \varepsilon_{ij} > X_{ik}'\beta + \varepsilon_{ik}, \text{ for all } k \neq j\right\}. \quad (1)$$

If choice j has maximum utility, then Y_{ij}=1. Otherwise, Y_{ij}=0. Thus, for each individual i, we observe X_{ij}, j=1, 2, …, J and Y_{ij}, j=1, 2,…, J.

The traditional approach to estimating β in the multinomial choice model under the assumption of random utility maximization is the method of maximum likelihood in which the errors are iid with a distribution known up to scale. The likelihood function to be maximized has the form

$$\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij} \log P\{Y_{ij} = 1 | X_{i1}, X_{i2}, \ldots, X_{iJ}, b\}.$$

For example, when ε_{ij} has the Type 1 extreme-value cdf F(t)=exp(–exp(–t)), $t \in R$, McFadden (1974) shows that the likelihood probabilities have the multinomial logit specification $\exp(X_{ij}'b)\left[\Sigma_{k=1}^{J}\exp(X_{ij}'b)\right]^{-1}$. The corresponding likelihood function is analytic and globally concave. Despite the consequent computational advantages, this specification makes very strong assumptions about the distribution of the errors. The MSE is consistent under much weaker assumptions about the errors. Manski (1975) only assumes that the disturbances ε_{ij} are independent and identically distributed (iid) across choices and independent but not necessarily identically distributed across individuals.

Write b for a generic element of the parameter space. It follows trivially from (1) that the infeasible criterion function

$$\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij}\left\{X_{ij}'b + \varepsilon_{ij} > X_{ik}'b + \varepsilon_{ik}, k \neq j\right\}$$

attains its maximum value of n at b= β. Since, for each i, the disturbances ε_{ij} are iid variates, this suggests estimating β with the maximizer of the so-called score function

$$\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij}\left\{X_{ij}'b > X_{ij}'b, k \neq j\right\}.$$

A score for a parameter b is the number of correct predictions made by predicting Y_{ij} to be 1 whenever $X_{ij}'b$ exceeds $X_{ij}'b$ for all $k \neq j$. A maximizer of the score function is an MSE of β. The maximizer need not be unique.

The MSE in the binary choice model and quantile regression

Now consider the binary model where J=2. Define $Y_i=Y_{i1}$ (implying $Y_{i2}=1-Y_i$) and $X_i=X_{i1}-X_{i2}$. Then the score function in (2) reduces to

$$\sum_{i=1}^{n}\left[Y_i\left\{X_i'b>0\right\}+(1-Y_i)\left\{X_i'b<0\right\}\right]. \tag{3}$$

Substitute $1-\left\{X_i'b>0\right\}$ for $\left\{X_i'b<0\right\}$ in (3) and expand each summand to see that maximizing (3) is equivalent to maximizing

$$S_n(b)=n^{-1}\sum_{i=1}^{n}(2Y_i-1)\left\{X_i'b>0\right\}. \tag{4}$$

Note that $Y_i=\left\{Y_i^*>0\right\}$ where $Y_i^*=X_i'\beta+\varepsilon_i$ with $\varepsilon_i=\varepsilon_{i1}-\varepsilon_{i2}$. For ease of exposition, write (Y^*, Y, X, ε) for $(Y_1^*, Y_1, X_1, \varepsilon_1)$ and x for an arbitrary point in the support of X. Thus, $Y=\{Y^*>0\}$ where $Y^*=X'\beta+\varepsilon$.
Before proceeding further, we must consider what interpretation to give to the parameter β in the last paragraph. The interpretation depends on our assumptions. For example, if we assume that β does not depend on x and that for every x, $E[Y^*|x]=x'\beta$, then β is such that the conditional mean of Y^* given $X=x$ is equal to $x'\beta$. However, if we assume that $\mathrm{MED}(Y^*|x)=x'\beta$, then β is such that the conditional median of Y^* given $X=x$ is equal to $x'\beta$. In general, the β satisfying the conditional mean assumption will be different from the β satisfying the conditional median assumption. Similarly, if we assume that for $a\neq 0.5$, the conditional α-quantile of Y^* given x is equal to $x'\beta$, then this β will, in general, be different from the β satisfying the conditional median assumption.
With this in mind, for $\alpha\in(0,1)$, write $Q_\alpha(Y^*|x)$ for the α-quantile of Y^* given $X=x$. Fix an $\alpha\in(0,1)$ and assume the linear α-quantile regression specification. That is, assume that for each x in the support of X, there exists a unique parameter β_α, depending on α but not on x, such that $Q_\alpha(Y^*|x)=x'\beta_\alpha$. This implies a zero conditional α-quantile restriction on ε: $Q_\alpha(\varepsilon|x)=0$ for all x.
For $\alpha\in(0,1)$, define

$$S_n^\alpha(b)=n^{-1}\sum_{i=1}^{n}\left[(2Y_i-1)-(1-2\alpha)\right]\left\{X_i'b>0\right\}. \tag{5}$$

Clearly, $S_n^{0.5}(b) = S_n(b)$ in (4). Assume that the linear α-quantile regression specification holds for some $\alpha \in (0,1)$. To see that it makes sense, under this assumption, to estimate β_α with the maximizer of $S_n^\alpha(b)$, consider $S^\alpha(b) = ES_n^\alpha(b)$. We see that

$$S^\alpha(b) = E^X\left[E[(2Y-1)-(1-2\alpha)]\left\{X'b>0\right\}|X\right] = E^X\left[\left[(2P\left\{-\varepsilon < X'\beta_\alpha|X\right\}-1)-(1-2\alpha)\right]\left\{X'b>0\right\}\right].$$

The linear α-quantile regression specification implies a zero conditional α-quantile restriction on ε: for all *x*, $P\{\varepsilon \le 0|x\} \le \alpha$ and $P\{\varepsilon \ge 0|x\} \ge 1-\alpha$. Thus, $x'\beta_\alpha > 0$ if and only if $P\left\{-\varepsilon \le x'\beta_\alpha|x\right\} \ge P\left\{-\varepsilon \le 0|x\right\} \ge 1-\alpha$. Deduce that for each possible value of *X*, the term in outer brackets in the last expression is maximized at $b=\beta_\alpha$. It follows that $S^\alpha(b)$ is maximized at $b=\beta_\alpha$. The analogy principle (Manski, 1988) prescribes using a maximizer of $S_n^\alpha(b)$ to estimate β_α.

The nonstandard convergence rate

The summands of the criterion function in (5) depend on *b* only through indicator functions of sets. As such, each summand has a 'sharp edge', to use the terminology of Kim and Pollard (1990). These authors provide a beautiful heuristic for why estimators that optimize empirical processes with sharp-edge summands converge at rate $n^{-1/3}$, rather than the usual $n^{-1/2}$ rate. They decompose the sample criterion function into a deterministic trend plus noise. Then, for each possible parameter value, they consider how the trend and the noise compete for dominance. Only a parameter value for which the trend does not overwhelm the standard deviation of the noise has a fighting chance of being an optimizer. Sharp edges produce standard errors with nonstandard sizes leading to the nonstandard $n^{-1/3}$ rate. We now examine how their argument works for the MSE for a very simple model. Assume the median regression specification for the model $Y = \{\beta - X - \varepsilon > 0\}$. Thus, $\beta_{0.5}=(\beta,-1)$ where the slope coefficient is known to equal –1 and the intercept β is the unknown parameter of interest. Assume that ε has median zero and is independent of *X*, so that the conditional median zero restriction is trivially satisfied. Also, assume that the distributions of *X* and ε have everywhere positive Lebesgue densities.

Refer to (4). Define $S(b) = ES_n(b) = E(2Y-1)\left\{X'b>0\right\}$. In the intercept example, $S(b) = E(2\{\varepsilon < \beta - X\}-1)\{X<b\}$. Simple calculations show that

$$S(b) = 2\int_{-\infty}^{b} F_\varepsilon(\beta - t) f_X(t)\,dt - F_X(b)$$

where $F_\varepsilon(\cdot)$ is the cdf of ε, $f_X(\cdot)$ is the pdf of *X*, and $F_X(\cdot)$ is the cdf of *X*. Write $f_\varepsilon(\cdot)$ for the pdf of ε.

Again, simple calculations show that

$$S'(b) = 2F_\varepsilon(\beta - b) f_X(b) - f_X(b) S''(b) = 2F_\varepsilon(\beta - b) f'_X(b) - f_X(b) 2 f_\varepsilon(\beta - b) - f'_X(b).$$

By the median restriction, we see that $S'(\beta) = 0$ and $S''(\beta) = -2 f_X(\beta) f_\varepsilon(0) < 0$. Thus, $S(b)$ is locally maximized at $b = \beta$. In fact, the given assumptions imply that $S(b)$ is globally and uniquely maximized at $b = \beta$. The MSE maximizes $S_n(b) - S_n(\beta)$. For each b, decompose $S_n(b) - S_n(\beta)$ into a sum of a deterministic trend and a random perturbation:

$$S_n(b) - S_n(\beta) = S(b) - S(\beta) + [S_n(b) - S_n(\beta) - [S(b) - S(\beta)]].$$

A Taylor expansion about β shows that for b near β, the trend $S(b) - S(\beta)$ is approximately quadratic with maximum value zero at $b = \beta$:

$$S(b) - S(\beta) \approx S''(\beta)(b - \beta)^2.$$

By a central limit theorem, for large n, the random contribution $S_n(b) - S_n(\beta) - [S(b) - S(\beta)]$ is approximately normally distributed with mean zero and variance σ_b^2 / n where

$$\sigma_b^2 = E[(2Y - 1)[\{X < b\} - \{X < \beta\}]]^2 - [E(2Y - 1)[\{X < b\} - \{X < \beta\}]]^2.$$

For b near β, the second term is much smaller than the first. It is the first term that accounts for the sharp-edge effect. It equals

$$F_X(\beta) + F_X(b) - 2[F_X(\beta)\{b > \beta\} + F_X(b)\{b < \beta\}].$$

A Taylor expansion of both $F_X(b)$ terms about β shows that this term is approximately equal to $|b - \beta| f_X(\beta)$ for b near β. Thus, near β, the criterion function $S_n(b) - S_n(\beta)$ is approximately equal to a quadratic maximized at β, namely, $-c_1(b - \beta)^2$ for $c_1 > 0$, plus a zero-mean random variable with standard deviation equal to $c_2 n^{-1/2} |b - \beta|^{1/2}$ for $c_2 > 0$. Values of b for which $-c_1(b - \beta)^2$ is much bigger in absolute value than

$c_2 n^{-1/2}|b-\beta|^{-1/2}$ have little chance of maximizing $S_n(b)–S_n(\beta)$. Rather, the maximizer is likely to be among those b values for which, for some $c>0$,

$$(b-\beta)^2 \leq cn^{-1/2}|b-\beta|^{1/2}.$$

Rearranging, we see that the maximizer is likely to be among the b values for which

$$|b-\beta| \leq cn^{-1/3}.$$

This is the essence of the heuristic presented by Kim and Pollard (1990) for $n^{-1/3}$ convergence rates. These authors also note that, when criterion functions are smooth, the variance of the random perturbation usually has order $|b–\beta|^2$ (instead of $|b–\beta|$) which, by the same heuristic, leads to the faster $n^{-1/2}$ convergence rate.

Smoothing the MSE

In order to remedy some of the shortcomings of the MSE, Horowitz (1992) develops a smoothed maximum score estimator (SMSE) under a linear median regression specification for the latent variable in the binary model. He replaces the indicator function in (4) with a smooth approximation. His SMSE maximizes a criterion function of the form

$$n^{-1}\sum_{i=1}^{n}(2Y_i-1)K(X_i'b/\sigma_n)$$

where K is essentially a smooth cdf and σ_n approaches zero as the sample size increases. Thus, $K(X_i'b/\sigma_n)$ approaches the indicator function $\{X_i'b>0\}$ as $n\to\infty$. By smoothing out the sharp-edge of the indicator function in (4), Horowitz is able to use Taylor expansion arguments to show that the SMSE, under slightly stronger conditions than those required for consistency of the MSE, converges at rate n^{δ} for $2/5 \leq \delta < 1/2$ and has a normal limit. The exact rate of convergence depends on certain smoothness assumptions and satisfies an optimality property (see Horowitz, 1993). The normality result makes it possible to do standard asymptotic inference with the SMSE. Horowitz (2002) shows that the bootstrapped SMSE provides asymptotic refinements. Kordas (2006) applies the smoothing technique of Horowitz (1992) to the criterion function in (5) and obtains asymptotic results similar to those of Horowitz (1992) for any $\alpha \in (0, 1)$.

See Also

- quantile regression

Bibliography

Abrevaya, J. and Huang, J. 2005. On the bootstrap of the maximum score estimator. *Econometrica* 73, 1175–204.

Horowitz, J.L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–31.

Horowitz, J.L. 1993. Optimal rates of convergence of parameter estimators in the binary response model with weak distributional assumptions. *Econometric Theory* 9, 1–18.

Horowitz, J.L. 2002. Bootstrap critical values for tests based on the smoothed maximum score estimator. *Econometrica* 111, 141–67.

Kim, J. and Pollard, D. 1990. Cube root asymptotics. *Annals of Statistics* 18, 191–219.

Koenker, R. and Bassett, G., Jr. 1978. Regression quantiles. *Econometrica* 46, 33–50.

Kordas, G. 2006. Smoothed binary regression quantiles. *Journal of Applied Econometrics* 21, 387–407.

Manski, C.F. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C.F. 1985. Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27, 313–33.

Manski, C.F. 1988. *Analog Estimation Methods in Econometrics*. New York: Chapman and Hall.

Manski, C.F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

Powell, J.L. 1994. Estimation of semiparametric models. In *Handbook of Econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

How to cite this article

Sherman, Robert P. "maximum score methods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000363> doi:10.1057/9780230226203.1066

The New Palgrave Dictionary of Economics Online

Mazzola, Ugo (1863 – 1899)

F. Caffè

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

free trade; marginalist theory; Mazzola, U.; Pantaleoni, M.; price determination; public finance

Article

Italian economist, born in Naples on 16 September 1863; died in Courmayeur on 14 August 1899. When he was just 20 years old, he graduated from the University of Naples, where his economic course had been largely based on the ideas of Francesco Ferrara. He did postgraduate research in Berlin, where he came in touch with Adolf Wagner and carried out research on behalf of the Italian Ministry of Agriculture into the problems associated with providing insurance for the working classes. At the age of 24 Mazzola was appointed to the Chair of Public Finance at the University of Pavia, where in 1896 he, with other economists, bought up the *Giornale degli Economisti* and transformed it into a centre of liberal thought. It was in this journal that the most eminent economists of the time, Maffeo Pantaleoni, Antonio De Viti De Marco, and Vilfredo Pareto, had their work published.

Mazzola believed fervently in the concept of free trade and he fought the protectionist trends which were threatening the country as it moved towards industrialization. In economic doctrine he was attracted to marginalist theory, as advocated by Jevons and Menger, and to its more comprehensive version in general equilibrium theory. Using these tools of analysis he wrote *I dati scientifici della finanza pubblica*, published in 1890, and thus made ‘a lasting contribution’ (to quote Pantaleoni) to the foundation of the theory of public finance. Mazzola was an expert on German fiscal theories, but he disagreed with the rather ambiguous way in which they differentiated between individual and collective aims. Mazzola stressed that, in his view, individual objectives are conditioned by public aims (defence, security, and so on) which can only be achieved by means of political cooperation. He believed that the provision of public welfare was necessary for the attainment of collective aims. So he analysed the characteristics of public welfare and then examined the process of price determination by means of the principles of maximization of utility – characteristics of the marginalist theory. In this way the phenomenon of fiscal theory was brought within the sphere of general economic analysis.

Selected works

1886. *L'assicurazione degli operai nella scienza e nella legislazione germanica*. Rome: Botta.

1890. *I dati scientifici della finanza pubblica*. Rome: Loescher.

1895. *L'imposta progressiva in economia pura e sociale*. Pavia.

1967. The formation of the prices of public goods. In *Classics in the Theory of Public Finance*, ed. R.A. Musgrave and A.T. Peacock. London: Macmillan.

Bibliography

Pantaleoni, M. 1899. Ugo Mazzola. *Giornale degli Economisti*. 2nd Series, 19(September), 189–98.

How to cite this article

Caffè, F. "Mazzola, Ugo (1863–1899)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000114> doi:10.1057/9780230226203.1067

The New Palgrave Dictionary of Economics Online

McCulloch, John Ramsay (1789 – 1864)

D.P. O'Brien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

absolute advantage; cobweb; Combination Laws; commodity money; convertibility; Corn Laws; cost of production; inflation; invariable measure; labour supply; labour theory of value; McCulloch, J. R.; Poor Law; population theory; public finance; relative value; rent; transfer problem; trade unions; wages fund

Article

MuCulloch was born in Galloway, Scotland on 1 March 1789. After attending Edinburgh University he secured employment as a lawyer's clerk. In 1816 he began his contributions to economics with two essays on the national debt. He was editor of *The Scotsman*, 1817–21, and a contributor to that paper until 1827. In 1818 he began writing for the *Edinburgh Review* and continued doing so until 1837, contributing nearly 80 articles. He also contributed to *Encyclopaedia Britannica* and was a prolific author, his works including editions of the *Wealth of Nations*, a *Commercial Dictionary*, a *Geographical Dictionary*, *Principles of Political Economy*, a *Statistical Account of the British Empire* and a *Treatise on Taxation*. He was a noted bibliophile, and after his death his library was purchased by his friend Lord Overstone and ultimately presented to Reading University.

It was not only as an author that McCulloch was influential; he was also called as an expert witness before the Select Committee on machinery of 1824 and that on Ireland of 1825. He was also one of the first public teachers of economics. He began lecturing in Edinburgh in 1820, and although attempts to establish a chair at Edinburgh University on his behalf in 1825 failed, he had been selected in 1824 to give the Ricardo Memorial lectures in London. In 1828, largely through the agency of James Mill, he was appointed the first professor of political economy at London University. He remained professor until 1837, though largely supporting himself by his pen; but in 1838 he was appointed Comptroller of the Stationery Office, a post he occupied until his death in 1864. This did not prevent him from continuing his literary activities and he remained active as an author, producing new editions of his earlier works and some completely new ones, notably the *Treatise on Taxation* (1845, 1852, 1863) and the *Geographical Dictionary* (1841–2, 1845–6, 1849, 1852, 1854).

As an economist, McCulloch was at one stage very much under the influence of Ricardo, but the

influence was transient. He put forward a simple labour theory of value under the influence of Ricardo, while aware that this was a simplified version (assuming away the problem of capital), for popular exposition. But though the emphasis is on labour quantity, and undoubtedly derived from Ricardo, it was a theory of relative (exchangeable) value, and McCulloch's analysis never borrowed from Ricardo the fundamental (to Ricardo) concept of the invariable measure. Indeed, McCulloch rejected the concept emphatically.

The vicissitudes of McCulloch's exposition over the years included, without doubt, a number of erroneous positions and a strange solution to the origin of profits as being in stored labour continuing to work. He also attempted to make labour cost a real (disutility) cost. But in the second edition (1830) of his *Principles* he advanced an almost complete cost of production theory, where cost was, as in an earlier *Scotsman* article, marginal cost. Labour cost, amortization and profit were now recognized as costs as they had been with Smith; the influences of Ricardo was now largely apparent in treating rent not as a cost but as a price-determined surplus, thus ignoring transfer earnings. Finally, in the 1838 edition of his notes to the *Wealth of Nations*, McCulloch clarified the nature of the profit reward with both waiting (as stressed by Ricardo) and productivity (as stressed by McCulloch) combining to produce a positive return.

McCulloch's treatment of money and banking derived from Smith, Hume and Thornton, though he long followed Ricardo in advancing the idea that the value of commodity money was determined by its cost of production, an idea he finally abandoned in his contributions to the eighth edition of *Encyclopaedia Britannica*. He accepted Hume's theory of the distribution of the precious metals, drawing also on Smith and especially Thornton, though he did accept, at least in relation to external losses of metal, the Ricardian definition of excess. In considering the internal level of activity in relation to money, he accepted Say's identify as an equilibrium proposition; but he also recognized the possibility that excess demand for money in disequilibrium might cause economic dislocation, thus magnifying fluctuations originating in the real side of the economy, such as over-investment. He also recognized, and approved, in contrast to Ricardo, the effects of mild inflation in producing forced saving and economic growth. On the issue of banking control he at first accepted Ricardo's view that convertibility was its own safeguard, but he came to recognize the problems of over-issue of notes much earlier than many writers – he was firmly opposed to laissez-faire in banking – and was one of the earliest writers to put forward the principle that a note issue should fluctuate in amount in response to the balance of payments exactly as an identically circumstanced metallic money would do, though he saw this as providing only a partial solution to monetary control.

McCulloch's analysis of international trade followed Smith rather than Ricardo, in basing trade on absolute advantage assuming international factor mobility. McCulloch may have done this because he saw that the possibility of trade advantage, as explained in comparative cost theory, was incomplete until this was translated into relative costs and prices. At all events he considered Ricardo's treatment of international trade to be faulty. In his view there was a complete parallel between international and inter-regional trade. McCulloch's treatment went well beyond that of Smith in some respects; and in particular, his analysis of the transfer problem, based on the work of Parnell, was an important precursor of modern developments. He discussed not only the effects of a transfer in the form of specie, or of commodities, but also a demand transfer of the kind made famous by Ohlin in his controversy with Keynes after the First World War. On matters of trade policy McCulloch has had the image of a crude free trader: but in fact, though he recognized the harm that protection could do, freedom of trade could,

in his view, involve the imposition of substantial import duties – even as high as 25 per cent – as long as these were balanced by home excise duties so as to avoid distortion of choice.

McCulloch's treatment of public finance used the Smithian framework: the analysis inevitably acquired a number of Ricardian accretions, but many of these were ultimately discarded. Moreover, McCulloch drew on a wide range of earlier writers on taxation and was particularly indebted to Hume and to Robert Hamilton. He presented a broad synthetic treatment which did much to give tax theory practicality after the Ricardian detour into the corn model. His main focus of attention was the use of fiscal policy in such a way as to ensure the maintenance of growth. Heavy taxation could interfere with growth; but if taxation was sensibly used it could be a stimulus to growth, increasing the supply of both effort and savings. A widely based regime of moderate indirect taxes, extending even to postage, was what McCulloch favoured, on ability-to-pay grounds, despite the distortions of the price system and the regressive elements. He opposed Gladstone's taxation policy, not only reliance upon income tax which McCulloch believed to interfere with growth, falling to stimulate effort and, where not proportional, subverting economic motivation.

The basic process of economic development was seen largely in terms of Smith's apparatus involving the accumulation of capital (but including human capital) and division of labour. It also involved the institutional requirements of security of property, internal freedom of trade, and a substantial role for government including education, control of public utilities, and employer liability for accidents. On to this Smithian basis McCulloch grafted specific Ricardian features, notably the Ricardian explanation for the declining rate of profit. However, he finally rejected the idea of inevitably diminishing returns in agriculture, as well as the inverse movement of wages and profits, and the Ricardian stagnation thesis. Writing later than Smith, McCulloch's treatment of development shows a much heavier emphasis on technology. Indeed, this was the basis of a fundamental disagreement with Ricardo over the role of machinery, and led him also to reject the primacy which Smith had afforded to agriculture – he attached key importance to the manufacturing sector though he was worried about the distributional consequences.

On agriculture itself however he wrote a good deal. He believed in large-scale capitalist farming (and supported primogeniture though not entails), and came, after a long period of believing in the (at least ultimate) inevitability of diminishing returns, to the view that under such a system improvements might continuously offset the diminution of returns. Thus his attack on the Corn Laws did not emphasize the Ricardian concept of stagnation but other matters, notably the idea that they encouraged price fluctuation. In this context he employed not only arguments about elasticity of demand and supply but also an agricultural cobweb. On this basis he argued, in contrast to Ricardo, that all classes, including the landlords, lost by the Corn Laws (though he followed Ricardo's argument for a measure of agricultural protection) and he also came to reject Ricardo's argument that improvements were against the interest of the landlords.

Classical economics, unlike modern economics, contained, as an integral part, a theory of population, which served in turn as the basis for various theorems about wages and welfare. McCulloch's population theory initially followed the first two editions of Malthus's celebrated *Essay*. But, probably under the influence of Nassau Senior, he become opposed to the Malthusian argument, believing prudential restraints to predominate. Indeed, amongst mainstream classical economists he was probably the most extreme anti-Malthusian, a development which harmonized with his move away from Ricardo's influence. Having sided with Malthus and Ricardo in opposing the Poor Law, he changed his mind quite

openly after 1826 and became a supporter of the old Poor Law, believing that it did not undermine prudential restraint and that it preserved social stability. He opposed the 1834 Poor Law as harsh and over-centralized. As a measure to raise wages generally he supported emigration and colonization, though he did not favour retention of control of colonies, and objected, in particular, to Wakefield's schemes for ‘scientific’ colonization.

All this raises directly McCulloch's concept of the operation of the labour market. McCulloch is particularly associated with the idea of the wage fund, because of his *Essay on Wages* (1826); and with a given wage fund, reducing supply will raise wages, as implied in McCulloch's treatment of emigration. His analysis of demand for labour thus equated capital with demand for labour, though in his more careful treatments he distinguished total and wage capital (and total population and labour supply). But this only provided half the analysis: on the supply side McCulloch employed four different labour supply functions, including a rising supply schedule as normal; two negatively inclined short-run schedules (the first when wages rose after excessive hours had been required to survive at the previous level of wages, the second where women and children entered the labour force as wages fell, to maintain family income). Fourthly, there was a secular population function. McCulloch favoured high wages, and his writings in defence of trades unions were important in the successful struggle to secure repeal of the Combination Laws.

As an economist, McCulloch was a fairly representative classical writer in that his work involved a synthesis of elements deriving from Smith and Ricardo. Yet McCulloch's case is particularly interesting because his own evolution over they very long period (48 years) of his writings, mirrors the development of classical economics itself. Starting from a basis of Smith and Malthus, he fell under the influence of Ricardo's magnetic personality and remarkable powers of abstraction; but he gradually passed through this phase, the Smithian elements in his work resuming their predominance and leading in turn to an emphasis on empirical work and a methodological position foreign to the tenor of Ricardo's work.

Selected works

1825. *The Principles of Political Economy: with some inquiries respecting their application, and a sketch of the rise and progress of the science*, 3rd edn. Edinburgh: W. Tait, 1843.

1832. *A Dictionary Practical, Theoretical, and Historical, of Commerce and Commercial Navigation*, 10th edn. London: Longmans, 1859.

1845. *A Treatise on the Principles and Practical Influence of Taxation and the Funding System*, ed. D.P. O'Brien. Edinburgh: Scottish Academic Press for The Scottish Economics Society, 1975.

Bibliography

O'Brien, D.P. 1970. *J.R. McCulloch: A Study in Classical Economics*. London: Allen & Unwin.

How to cite this article

O'Brien, D. P. "McCulloch, John Ramsay (1789–1864)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000115> doi:10.1057/9780230226203.1068

The New Palgrave Dictionary of Economics Online

McFadden, Daniel (born 1937)

John Rust

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the contributions of Daniel L. McFadden, 2000 co-recipient of the Nobel Prize in Economics. The article focuses on his seminal contributions to the theoretical and econometric literatures on discrete choice.

Keywords

curse of dimensionality; discrete choice; duality; generalized extreme value; independence of irrelevant alternatives; infinite horizons; logit models of individual choice; mathematical psychology; maximum likelihood; McFadden, D. L.; method of simulated moments; Monte Carlo methods; multinomial logit model; multinomial probit model; neural networks; overlapping generations models; probabilistic choice theory; random utility models; revealed preference; simulation-based estimation

Article

1 Introduction

Daniel L. McFadden, the E. Morris Cox Professor of Economics at the University of California at Berkeley, was the 2000 co-recipient of the Nobel Prize in Economics, awarded ‘for his development of theory and methods of analyzing discrete choice’. (The prize was split with James J. Heckman, awarded ‘for his development of theory and methods for analyzing selective samples’). McFadden was born in North Carolina, USA, in 1937 and received a BS in physics from the University of Minnesota (with highest honors) in 1956, and a Ph.D. in economics from Minnesota in 1962. His academic career began as a postdoctoral fellow at the University of Pittsburgh. In 1963 he was appointed as assistant professor of economics at the University of California at Berkeley, and tenured in 1966. He has also held tenured appointments at Yale University (as Irving Fisher Research Professor in 1977), and at the Massachusetts Institute of Technology (from 1978 to 1991). In 1990 he was awarded the E. Morris Cox Chair at the

University of California at Berkeley, where he has also served as Department Chair and as Director of the Econometrics Laboratory.

2 Research contributions

McFadden is best known for his fundamental contributions to the theory and econometric methods for analysing *discrete choice*. Building on a highly abstract, axiomatic literature on *probabilistic choice theory* due to Thurstone (1927), Block and Marschak (1960), and Luce (1959), McFadden developed the econometric methodology for estimating the utility functions underlying probabilistic choice theory. McFadden's primary contribution was to provide the econometric tools that permitted widespread *practical empirical application* of discrete choice models, in economics and other disciplines. According to his autobiography (McFadden, 2001),

> In 1964, I was working with a graduate student, Phoebe Cottingham, who had data on freeway routing decisions of the California Department of Transportation, and was looking for a way to analyze these data to study institutional decision-making behavior. I worked out for her an econometric model based on an axiomatic theory of choice behavior developed by the psychologist Duncan Luce. Drawing upon the work of Thurstone and Marshak, I was able to show how this model linked to the economic theory of choice behavior. These developments, now called the multinomial logit model and the random utility model for choice behavior, have turned out to be widely useful in economics and other social sciences. They are used, for example, to study travel modes, choice of occupation, brand of automobile purchase, and decisions on marriage and number of children.

Thousands of papers applying his technique have been published since his path-breaking papers, ‘Conditional Logit Analysis of Qualitative Choice Behavior’ (1973) and ‘The Revealed Preferences of a Government Bureaucracy: Empirical Evidence’ (1976). In December 2005, a search of the term ‘discrete choice’ using the Google search engine yielded 10,200,000 entries, and a search on the Google Scholar search engine (which limits search to academic articles) returned 759,000 items.
Besides the discrete choice literature itself, McFadden's work has spawned a number of related literatures in econometrics, theory, and industrial organization that are among the most active and productive parts of the economic literature in the present day. This includes work in game theory and industrial organization (for example, the work on discrete choice and product differentiation of Anderson, De Palma and Thisse (1992), estimation of discrete games of incomplete information (Bajari et al., 2005), and discrete choice modelling in the empirical industrial organization literature (Berry, Levinsohn and Pakes, 1995, and Goldberg, 1995), the econometric literature on semiparametric estimation of discrete choice models (Manski, 1985; McFadden and Train, 2000), the literature on discrete/continuous choice models and its connection to durable goods and energy demand modelling (Dagsvik, 1994; Dubin and McFadden, 1984; Hannemann, 1984), the econometric literature on choice based and stratified sampling (Cosslett, 1981; Manski and McFadden, 1981), the econometric literature on ‘simulation estimation’ (McFadden, 1994; Hajivassiliou and Ruud, 1994; Pakes and Pollard, 1989),

and the work on structural estimation of dynamic discrete choice models and extensions thereof (Dagsvik, 1983; Eckstein and Wolpin, 1989; Heckman, 1981; Rust, 1994).

McFadden has also made significant contributions to other fields, particularly to economic theory and production economics. Due to space limitations, I can only briefly mention several of his best known contributions here. McFadden's earliest published work was in pure theory, including seminal work on duality theory of production functions that was subsequently published in his book on Production Economics edited with Melvyn Fuss in 1978. McFadden made important contributions to growth theory including his 1967 Review of Economic Studies paper that showed how the overtaking criterion could be used to evaluate infinite horizon development programmes, resolving an outstanding paradox raised by Diamond and Koopmans. In a series of papers with Mitra and Majumdar (1976; 1980), McFadden extended the classical competitive equilibrium welfare theorems established by Debreu and others in finite economies, (that is, competitive equilibria are Pareto efficient, and any Pareto efficient allocation can be sustained as a competitive equilibrium after a suitable reallocation of resources), to infinite horizon economies. This work was not a simple technical extension or previous work by Debreu: it resolved serious conceptual problems created by the fact that in an infinite horizon economy (which includes standard overlapping generations models) the commodity space is infinite-dimensional and the number of consumers is infinite. These papers provided sufficient conditions for the existence of these fundamental welfare theorems, resolving paradoxes raised by Paul Samuelson, who showed special cases of infinite horizon overlapping generation economies where competitive equilibria can be strikingly inefficient. Another landmark paper is McFadden's (1974) paper on excess demand functions with Mantel, Mas-Colell and Richter. This paper provided one of the most general proofs of a classic conjecture by Hugo Sonnenschein that the necessary and sufficient properties of any system of aggregate excess demand functions are that it satisfy the following three properties: (1) homogeneity, (2) continuity, and (3) Walras's Law. McFadden has made numerous other contributions to economic theory that I do not have space to cover here. Instead, I now return to a more in depth review of McFadden's contributions to the discrete choice literature, the primary contributions that were cited in his Nobel Prize award.

3 Contributions to discrete choice

McFadden's contributions built on prior work in the literature on *mathematical psychology* (see logit models of individual choice for further details). McFadden's contribution to this literature was to recognize how to operationalize the random utility interpretation in an empirically tractable way. In particular, he provided the first a random utility interpretation of the multinomial logit (MNL) model. His other fundamental contribution was to solve an analogue of the *revealed preference problem:* that is, using data on the actual choices and states of a sample of agents $\{(d_i, x_i)\}_{i=1}^{N}$, he showed how it was possible to ‘reconstruct’ their underlying random utility functions via the method of maximum likelihood, where the likelihood is a product of individuals’ *conditional choice probabilities*. Given the simplicity of the MNL choice probabilities, this worked helped to spawn a huge empirical literature that applied discrete choice models to a wide variety of phenomena. Further, McFadden introduced a new class of multivariate distributions, the *generalized extreme value family* (GEV), and derived tractable formulas for the implied choice probabilities including the *nested multinomial logit model*, and showed

that these models relax some of the empirically implausible restrictions implied by the multinomial logit model, particularly the *independence from irrelevant alternatives* (IIA) property.

3.1 Multivariate extreme value distributions and the multinomial logit model

McFadden assumed that an individual's utility function has the following *additive separable* representation

$$U(x, z, d, \theta) = u(x, d, \theta) + v(z, d). \tag{3.1}$$

Define $\varepsilon(d) \equiv v(z, d\bullet)$. It follows that an assumption on the distribution of the random vector z implies a distribution for the random vector $\varepsilon \equiv \{\varepsilon(d) | d \in D(x)\}$. McFadden's approach was to make assumptions directly about the distribution of ε, rather than making assumptions about the distribution of z and deriving the implied distribution of ε. Standard assumptions for the distribution of ε that have been considered include the *multivariate normal* which yields the *multivariate probit* variant of the discrete choice model. Unfortunately, in problems where there are more than only two alternatives (the case that Thurstone studied), the multinomial probit model becomes intractable in higher dimensional problems. The reason is that, in order to derive the conditional choice probabilities, one must do numerical integrations that have a dimension equal to $|D(x)|$, the number of elements in the choice set. In general this multivariate integration is computationally infeasible when $|D(x)|$ is larger than 5 or 6, using standard quadrature methods on modern computers.
McFadden introduced an alternative assumption for the distribution of ε, namely the *multivariate extreme value distribution* given by

$$F(z|x) = Pr\{\varepsilon_d \le z_d | d \in D(x)\} = \prod_{d \in D(x)} \exp\{-\exp\{-(z_d - \mu_d)/\sigma)\}\}, \tag{3.2}$$

and showed that (when the location parameters μ_d are normalized to) the corresponding random utility model produces choice probabilities given by the multinomial logit formula

$$P(d|x, \theta) = \frac{\exp\{u(x, d, \theta)/\sigma\}}{\Sigma_{d' \in D(x)} \exp\left\{u(x, d', \theta)/\sigma\right\}}.$$

This is McFadden's key result, that is, the *MNL choice probability is implied by a random utility model when the random utilities have extreme value distributions*. It leads to the insight that the *independence from irrelevant alternatives* (IIA) property of the MNL model is a consequence of the statistical independence in the random utilities. In particular, even if the *observed attributes* of two alternatives d and d' are identical (which implies $u(x, d, \theta \bullet)=u(x, d', \theta)$), the statistical independence of unobservable components $\varepsilon(d)$ and $\varepsilon(d')$ implies alternatives d and d' are not perfect substitutes even when their observed characteristics are identical. In many cases this is not problematic: individuals may have different idiosyncratic perceptions and preferences for two different items that have the same observed attributes. However in the case of the ‘red bus/blue bus’ example or the concert ticket example discussed by Debreu (1960), there are cases where it is plausible to believe that the observed attributes provide a sufficiently good description of an agent's perception of the desirability of two alternatives. In such cases, the hypothesis that choices are also affected by additive, *independent* unobservables $\varepsilon(d)$ provides a poor representation of an agent's decisions. What is required in such cases is a random utility model that has the property that the degree of correlation in the unobserved components of utility $\varepsilon(d)$ and $\varepsilon(d')$ for two alternatives $d, d' \in D(x)$ is a function of the degree of closeness in the observed attributes. This type of dependence can be captured by a *random coefficient probit model*. This is a random utility model of the form $U(x, z, d, \theta) = x_d(\theta + z)$ where x_d is a $k\times1$ vector of observed attributes of alternative d, and θ is a $k\times1$ vector of utility weights representing the mean weights individuals assign to the various attributes in x_d in the population and $z\sim N(0,\Omega)$ is a $k\times1$ normally distributed random vector representing agent specific deviations in their weighting of the attributes relative the population average values, θ. Under the random coefficients probit specification of the random utility model, when $x_d = x_{d'}$, alternatives d and d' are in fact perfect substitutes for each other and this model is able to provide the intuitively plausible prediction of the effect of introducing an irrelevant alternative – the red bus – in the red bus/blue bus problem (see, for example, Hausman and Wise, 1978).

3.2 Generalized extreme value distributions and nested logit models

McFadden (1981) introduced the generalized extreme value (GEV) family of distributions. This family relaxes the independence assumption of the extreme value specification while still yielding tractable expressions for choice probabilities. The GEV distribution is given by

$$F(z|x) = Pr\{\varepsilon_d \le z_d | d \in D(x)\} = \exp\{-H(\exp\{-z_1\}, \ldots, \exp\{-z_{|D(x)|}\}, x, D(x))\},$$

for any function $H(z, x, D(x))$ satisfying certain consistency properties. McFadden showed that choice probabilities for the GEV distribution are given by

$$P(d|x,\theta) = \frac{\exp\{u(x,d,\theta)\}H_d(\exp\{u(x,1,\theta)\},\ldots,\exp\{u(x,|D(x)|,\theta)\},x,D(x))}{H(\exp\{u(x,1,\theta)\},\ldots,\exp\{u(x,|D(x)|,\theta)\},x,D(x)),}$$

where $H_d(z, x, D(x))=\partial/\partial_{Zd} H(z, x, D(x))$. A prominent subclass of GEV distributions is given by H functions of the form

$$H(z,\gamma,D(x)) = \sum_{i=1}^{n}\left[\sum_{d\in D_i(x)} z_d^{1/\sigma_i}\right]^{\sigma_i},$$

where $\{D_1(x),\bullet\ldots\bullet,D_n(x)\}$ is a partition of the full choice set $D(x)$. This subclass of GEV distributions yields the *nested multinomial logit* (NMNL) choice probabilities (see logit models of individual choice for further details).

The NMNL model has been applied in numerous empirical studies especially to study demand where there are an extremely large number of alternatives, such as modelling consumer choice of automobiles (for example, Berkovec, 1985; Goldberg, 1995). In many of these consumer choice problems there is a natural partitioning of the choice set in terms of *product classes* (for example, luxury, compact, intermediate, sport-utility, and so on, classes in the case of autos). The nesting avoids the problems with the IIA property and results in more reasonable implied estimates of demand elasticities than those obtained using the MNL model. In fact, Dagsvik (1995) has shown that the class of random utility models with GEV distributed utilities is ‘dense’ in the class of all random utility models, in the sense that choice probabilities implied from any random utility model can be approximated arbitrarily closely by a random utility model in the GEV class. However a limitation of nested logit models is that they imply a highly structured pattern of correlation in the unobservables induced by the econometrician's specification of how the overall choice set $D(x)$ is to be partitioned, and the number of levels in the nested logit ‘tree’. Even though the NMNL model can be nested to arbitrarily many levels to achieve additional flexibility, it is desirable to have a method where patterns of correlation in unobservables can be estimated from the data rather than being imposed by the analyst. Further, even though McFadden and Train (2000) recognize Dagsvik's (1995) finding as a ‘powerful theoretical result’, they conclude that ‘its practical econometric application is limited by the difficulty of specifying, estimating and testing the consistency of relatively abstract generalized Extreme Value RUM’ (McFadden and Train, 2000, p. 452).

3.3 Method of simulated moments and simulation based inference for discrete choice

As noted above, the random coefficients probit model has many attractive features: it allows a flexibly specified covariance matrix representing correlation between unobservable components of utilities that

avoid many of the undesirable features implied by the IIA property of the MNL model, in a somewhat more direct and intuitive fashion than is possible via the GEV family. However as noted above, the multinomial probit model is intractable for applications with more than four or five alternatives due to the ‘curse of dimensionality’ of the numerical integrations required, at least using deterministic numerical integration methods such as Gaussian quadrature. One of McFadden's most important contributions was his (1989) *Econometrica* paper that introduced the *method of simulated moments* (MSM). This was a major breakthrough that introduced a new econometric method that made it feasible to estimate the parameters of multinomial probit models with arbitrarily large numbers of alternatives. The basic idea underlying McFadden's contribution is to use *Monte Carlo integration* to approximate the probit choice probabilities. While this idea had been previously proposed by others, it was never developed into a practical, widespread estimation method because ‘it requires an impractical number of Monte Carlo draws to estimate small choice probabilities and their derivatives with acceptable precision’ (McFadden, 1989, p. 997). However McFadden's insight was that it is not necessary to have extremely accurate (and thus very computationally time-intensive) Monte Carlo estimates of choice probabilities in order to obtain an estimator for the parameters of a multinomial probit model that is consistent and asymptotically normal and performs well in finite samples. McFadden's insight is that the *noise from Monte Carlo simulations can be treated in the same way as random sampling error and will thus ‘average out’ in large samples*. In particular, his MSM estimator has good asymptotic properties even when *only a single Monte Carlo draw is used to estimate each agent's choice probability*. See simulation-based estimation for further details on the MSM estimator.

The idea behind the MSM estimator is quite general and can be applied in many other settings besides the multinomial probit model. McFadden's work helped to spawn a large literature on ‘simulation estimation’ that developed rapidly during the 1990s and resulted in computationally feasible estimators for a large new class of econometric models that were previously considered to be computationally infeasible. However, there are even better simulation estimators for the multinomial probit model, which generally outperform the MSM estimator in terms of having lower asymptotic variance and better finite sample performance, and which are easier to compute. One problem with the simple Monte Carlo estimator $\hat{P}(x_i, \theta)$ underlying the MSM estimator is that it is a discontinuous and ‘locally flat’ function of the parameters θ, and thus the MSM criterion function is difficult to optimize. Hajivassiliou and McFadden (1998) introduced the *method of simulated scores* (MSS) that is based on Monte Carlo methods for simulating the *scores* of the likelihood function for a multinomial probit model and a wide class of other *limited dependent variable models* such as Tobit and other types of censored regression models. (In the case of a discrete choice model, the score for the ith observation is $\partial / \partial \theta \log(P(d_i|x_i, \theta))$.) Because it simulates the score of the likelihood rather than using a method of moments criterion, the MSS estimator is more efficient than the MSM estimator. Also, the MSS is based on a *smooth simulator* (that is, a method of simulation that results in an estimation criterion that is a continuously differentiable function of the parameters θ), so the MSS estimator is much easier to compute than the MSM estimator. Based on numerous Monte Carlo studies and empirical applications, MSS (and a closely related simulated maximum likelihood estimator based on the Geweke–Hajivassiliou–Keane’, GHK, smoother simulator) are now regarded as the estimation methods of choice for a wide class of econometric models with limited dependent variable that are commonly encountered in empirical applications (see simulation-based estimation for further details).

3.4 Mixed logit models

A mixed MNL model has choice probabilities of the form

$$P(d|x,\theta) = \int \left[\frac{\exp\{u(x,d,\alpha)}{\Sigma_{d' \in D(x)} \exp\left\{u(x,d',\alpha)\right\}} \right] G(d\alpha|\theta). \tag{3.3}$$

There are several possible random utility interpretations of the mixed logit model. One interpretation is that the α vector represents ‘unobserved heterogeneity’ in the preference parameters in the population, so the relevant choice probability is marginalized using the population distribution for the α parameters in the population, $G(\alpha|\theta)$. The other interpretation is that α is similar to vector ε, that is, it represents information that agents observe and which affects their choices (similar to ε) but which is unobserved by the econometrician, except that the components of ε, $\varepsilon(d)$ enter the utility function additively separably, whereas the variables α are allowed to enter in a non-additively separable fashion and the random vectors α and ε are statistically independent. It is easy to see that, under either interpretation, the mixed logit model will not satisfy the IIA property, and thus is not subject to its undesirable implications. McFadden and Train proposed several alternative ways to estimate mixed logit models, including maximum simulated likelihood and MSM. In each case, Monte Carlo integration is used to approximate the integral in equation (3.3) with respect to $G(\alpha|\theta\bullet)$. Both of these estimators are smooth functions of the parameters θ, and both benefit from the computational tractability of the MNL while at the same time having the flexibility to approximate virtually any type of random utility model. The intuition behind McFadden and Train's approximation theorem is that a mixed logit model can be regarded as a certain type of *neural network* using the MNL model as the underlying ‘squashing function’. Neural networks are known to have the ability to approximate arbitrary types of functions and enjoy certain optimality properties, that is, the number of parameters (that is, the dimension of the α vector) needed to approximate arbitrary choice probabilities grows only linearly in the number of included covariates x. (Other approximation methods, such as *series estimators* formed as tensor products of bases that are univariate functions of each of the components of x, require a much larger number of coefficients to provide an comparable approximation, and the number of such coefficients grows exponentially fast with the dimension of the x vector.)

4 Conclusion

This brief survey of McFadden's contributions to the discrete choice literature has revealed the immense practical benefits of his ability to link theory and econometrics, innovations that lead to a vast empirical literature and widespread applications of discrete choice models. Beginning with his initial discovery, that is, his demonstration that multinomial logit choice probabilities result from a random utility model

with multivariate extreme value distributed unobservables, McFadden has made a series of fundamental contributions that have enabled researchers to circumvent the problematic implications of the IIA property of the MNL model, providing computationally tractable methods for estimating ever wider and more flexible classes of random utility and limited dependent-variable models in econometrics.

See Also

- logit models of individual choice
- simulation-based estimation

Selected works

1967. The evaluation of development programmes. *Review of Economic Studies* 34, 25–50.

1973. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, ed. P. Zarembka. New York: Academic Press.

1974. The measurement of urban travel demand. *Journal of Public Economics* 3, 303–28.

1974. (With R. Mantel, A. Mas-Colell and M.K. Richter.) A characterization of community excess demand functions. *Journal of Economic Theory* 9, 361–74.

1976. The revealed preferences of a government bureaucracy: empirical evidence. *Bell Journal of Economics and Management Science* 7, 55–72.

1976. (With M. Majumdar and T. Mitra.) On efficiency and Pareto optimality of competitive programs in closed multisector models. *Journal of Economic Theory* 13, 26–46.

1978. Cost, revenue, and profit functions. In *Production Economics: A Dual Approach to Theory and Applications*, vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

1980. (With T. Mitra and M. Majumdar, M.) Pareto optimality and competitive equilibrium in infinite horizon economies. *Journal of Mathematical Economics* 7, 1–26.

1981. Econometric models of probabilistic choice. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

1981. (With C.F. Manski, ed.) *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.

1984. Econometric models of qualitative response models. In *Handbook of Econometrics*, vol. 2, ed. Z.

Griliches and M. Intriligator. Amsterdam: North-Holland.

1989. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57, 995–1026.

1994. (With P. Ruud.) Estimation by simulation. *Review of Economics and Statistics* 76, 591–608.

2000. (With K. Train.) Mixed MNL models of discrete response. *Journal of Applied Econometrics* 15, 447–70.

2001. Autobiography. In *Les Prix Nobel. The Nobel Prizes 2000*, ed. T. Frängsmyr. Stockholm: Nobel Foundation.

Bibliography

Anderson, S.P., De Palma, A. and Thisse, J.F. 1992. *Discrete Choice Theory of Product Differentiation*. Cambridge: MIT Press.

Bajari, P., Hong, H., Krainer, J. and Nekipelov, D. 2005. Estimating static models of strategic interactions. Manuscript, University of Michigan.

Berkovec, J. 1985. New car sales and used car stocks: a model of the automobile market. *RAND Journal of Economics* 16, 195–214.

Berry, S., Levinsohn, J. and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–90.

Block, H. and Marschak, J. 1960. Random orderings and stochastic theories of response. In *Contributions to Probability and Statistics*, ed. I. Olkin. Stanford: Stanford University Press.

Cosslett, S.R. 1981. Efficient estimation of discrete-choice models. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

Dagsvik, J.K. 1983. Discrete dynamic choice: an extension of the choice models of Luce and Thurstone. *Journal of Mathematical Psychology* 27, 1–43.

Dagsvik, J.K. 1994. Discrete and continuous choice, max-stable processes and independence from irrelevant attributes. *Econometrica* 62, 1179–205.

Dagsvik, J.K. 1995. How large is the class of generalized extreme value models? *Journal of Mathematical Psychology* 39, 90–8.

Debreu, G. 1960. Review of R.D. Luce '*Individual Choice Behavior*'. *American Economic Review* 50, 186–8.

Dubin, J. and McFadden, D. 1984. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52, 345–62.

Eckstein, Z. and Wolpin, K. 1989. The specification and estimation of dynamic stochastic discrete choice models. *Journal of Human Resources* 24, 562–98.

Goldberg, P. 1995. Product differentiation and oligopoly in international markets: i the case of the U.S. automobile industry. *Econometrica* 63, 891–951.

Hajivassiliou, V. and McFadden, D.L. 1998. The method of simulated scores for the estimation of LDV models. *Econometrica* 66, 863–96.

Hajivassiliou, V.A. and Ruud, P.A. 1994. Classical estimation methods for LDV models using simulation. In *Handbook of Econometrics*, vol. IV, ed. R.F. Engle and D.L. McFadden. Amsterdam: Elsevier.

Hannemann, M. 1984. Discrete/continuous models of consumer demand. *Econometrica* 52, 541–62.

Hausman, J. and Wise, D. 1978. A conditional probit model of qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46, 403–26.

Heckman, J.J. 1981. Statistical models for discrete panel data. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

Luce, R.D. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Manski, C.F. 1985. Semiparametric estimation of discrete response: asymptotics of the maximum score estimator. *Journal of Econometrics* 27, 303–33.

Manski, C.F. 2001. Dan McFadden and the econometric analysis of discrete choice. *Scandinavian Journal of Economics* 103, 217–29.

Pakes, A. and Pollard, D. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–57.

Rust, J. 1994. Structural estimation of Markov decision processes. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D.L. McFadden. Amsterdam: Elsevier.

Thurstone, L.L. 1927. Psychophysical analysis. *American Journal of Psychology* 38, 368–89.

How to cite this article

Rust, John. "McFadden, Daniel (born 1937)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000239> doi:10.1057/9780230226203.1069

The New Palgrave Dictionary of Economics Online

Meade, James Edward (1907 – 1995)

David Vines

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article explains the part that Meade played in the creation of Keynes's *General Theory*, describes his work with Keynes during the Second World War in the creation of the IMF and the GATT, and summarizes the ideas in *The Theory of International Economic Policy* for which Meade was awarded the Nobel Prize in 1977. It also sets out the role that Meade played in the construction of the inflation-targeting regime which became the centrepiece of British macroeconomic policymaking in the 1990s.

Keywords

absorption approach to the balance of payments; aggregate demand; balance of payments; brain drain; Chamberlin, E. H.; compensation principles; consumer surplus; control theory; demand management; Devaluation; domestic distortions, theory of; double-entry national accounts; Dutch Disease; elasticities approach to the balance of payments; exchange rate mechanism (ERM); factor trade; fiscal policy; fixprice models; Fleming, J. M.; foreign direct investment; Friedman, M.; Harrod, R. F.; Harrod–Domar growth model; Heckscher–Ohlin trade theory; Hicks, J. R.; imperfect competition; imperial preference; income–expenditure theory; incomes policy; inflation; inflation targeting; Institute of Fiscal Studies (UK); interest rates; international policy coordination; international trade theory; IS–LM models; Johnson, H. G.; Kahn, R. F.; Kalecki, M.; Kay, J.; Keynes, J. M.; Keynesian Revolution; King, M.; Marshall, A.; Meade, J. E.; monetarism; monetary policy; monetary theory of the balance of payments; money multiplier; money supply; multinational corporations; multiplier; multivariable control methods; Mundell–Fleming model; national income accounting; new Keynesian macroeconomics; new welfare economics; nominal-income targeting; non-accelerating inflation rate of unemployment (NAIRU); open-economy macroeconomics; optimal savings rate; optimum population; Phillips curve; Phillips, A. W. H.; potential welfare; profit sharing; protection; Ramsey model; regional free-trade areas; Robbins, L. C.; Robertson, D.; Robinson, E. A. G.; Robinson, J. V.; saving and investment; second best; Sidgwick, H.; social security; Sraffa, P.; stabilization bias; sterilization; sticky wages; stocks and flows; Stolper–

Samuelson theorem; Stone, J. R. N.; supply side reform; Taylor rules; Taylor, J.; terms of trade; Tinbergen, J.; Tobin's q; trade creation and trade diversion; unemployment; wage fixing

Article

James Meade was one of the truly great economists of the 20th century. He was profoundly internationalist in his outlook, and was awarded the Nobel Memorial Prize in Economics in 1977, jointly with Bertil Ohlin, for *The Theory of International Economic Policy* (1951–5). But his contributions spanned the whole of the discipline. He made fundamental, and widely influential, contributions to economic theory, in both macroeconomics and microeconomics. More than this, his main concern was always with the part that economic analysis has to play in the solution of practical economic policy problems. As a result, he contributed to the theory of economic policy in a very wide range of subjects, including macroeconomic management, trade policy reform, public finance, economic growth, income distribution, wage determination, and population growth. He served actively in policymaking as an economist for the League of Nations, and in the Economic Section of the UK Cabinet Office during, and immediately after, the Second World War. In all that he did, Meade saw the role of an economist as helping to design a better society – both by the creation of good institutions of economic management and by the provision of appropriate incentives for private individuals.

In this article I concentrate on four main issues. I first explain the large part that Meade played in the creation of Keynes's *General Theory* in the 1930s. After this, I describe his work with Keynes during the Second World War in the creation of the International Monetary Fund and the GATT (which has since become the World Trade Organization, or WTO). I then turn to Meade's work on international economics at the London School of Economics (LSE), immediately after the war, for which he was awarded the Nobel Prize; I spend some time showing how this theoretical work was related to his earlier work on policy with Keynes. Finally, I set out the role that Meade played, along with a group of young economists to which I belonged, in the construction of the inflation-targeting regime that became the centrepiece of British macroeconomic policymaking in the 1990s.

1 Activities before the Second World War: Keynesian macroeconomics

Meade was born on 23 June 1907 in Swanage, Dorset, and brought up in Bath. He went to school at Malvern College, and then won a scholarship in classics to Oriel College in Oxford. But like many others of his generation he was appalled by the problem of mass unemployment, which, as he said, caused ‘poverty in the midst of plenty’. As a result he turned to the study of economics for the last two years of his university education. Meade gained greatly from studying classics, but, as a result of doing so, he had to teach himself the mathematics that he later used extensively.

Immediately upon graduating in 1930, Meade was elected to a fellowship at Hertford College, and appointed to a lectureship in economics at Oxford University. But in October 1930 his college first sent him to Cambridge for the academic year 1930/31, ‘to learn my subject before I started to teach it. I had the greatest good fortune of being taken into Trinity College … by Dennis Robertson, to whose teaching that year I owe a deep debt of gratitude. At an early stage he told me that there was a young man in Kings called Richard Kahn whom I should get to know’ (Meade, 1983b, p. 263).

And so Meade spent a formative and creative year as a member of the ‘Circus’ which was gathered around Keynes. This group of people were debating Keynes’ *Treatise on Money* (Keynes, 1930) which had just been published, and included Joan and Austin Robinson, and Piero Sraffa, as well as Kahn. Meade enjoyed describing the ‘workshop style’ of the Circus meetings. Keynes took no part in the proceedings, but after each meeting Richard Kahn orally recounted to Keynes the subject matter of the discussions and the lines of argument.

> From the point of view of a humble mortal like myself Keynes seemed to play the role of God in a morality play; he dominated the play but rarely appeared on the stage. Kahn was the Messenger Angel who brought messages and problems from Keynes to the ‘Circus’ and went back to Heaven with the result of our deliberations. (Keynes, 1971–88, vol. 13, p. 339; see also p. 338)

The casting of Keynes in this role was first suggested by Meade's wife Margaret in 1934 when they were staying for the weekend with Austin and Joan Robinson in Cambridge. That weekend, too, was dominated by messages from people who had just spoken with Keynes, though Keynes himself never appeared in person.

The Circus was discussing the failure of Keynes's *Treatise*. Keynes had expected that book to become his magnum opus. In it he set out the theoretical work which he had done since the end of the First World War, about the causes of the economic cycle, and about how this cycle should be managed on a national basis and internationally. There is much modern macroeconomics in the *Treatise*, and the international macroeconomics is particularly good: it is possible to find elements of the Swan diagram, of the Fleming–Mundell model, and even of the Dornbusch model (see Vines, 2003). But the *Treatise* contains a fatal flaw. It aims to analyse the problem of the economic cycle, but the discussion rests upon a formal model in which the level of economic activity is exogenous. This mistaken model is nevertheless of interest because it contains the necessary clues about what Keynes needed to do next. In the *Treatise*, an increase in aggregate demand, caused – say – by an exogenous increase in investment, and not prevented by the central bank from having an overall effect on aggregate demand, would cause demand to increase relative to supply, which was assumed to be fixed, and so would cause a rise in the level of prices. This would redistribute income to profits, away from wages, because the level of the money wage is ‘somewhat sticky’ in the model. That would raise the overall level of savings, because the propensity to save is assumed to be higher for capitalists than for workers. A new equilibrium would be regained after the working out of a ‘multiplier’ process: one in which the price level rises by the amount necessary to re-equilibrate leakages (the extra savings) with injections (the increased investment). This all makes sense, except if the model is meant to help in a discussion of booms and slumps in *output*, which it cannot do because output is exogenous. Sorting out this mess required Keynes to write the *General Theory of Employment, Interest and Money*, which took him until 1936.

By the time Meade arrived in Cambridge in October, Kahn had already drafted his famous article on the ‘multiplier’ (Kahn, 1931). In this piece Kahn showed that, if output is endogenous, one can sum an infinite geometric series to show that the overall effect on output of an increase in investment is ‘multiplied’ because of the increases in consumption which happen as output increases. It appears that it was Meade, the young graduate student from Oxford, who showed how Kahn's multiplier analysis could

be connected with Keynes's argument in the *Treatise*. There were two steps in this demonstration. First, Meade showed, by summing the series of the effects of output on savings instead of the series of the effects of output on consumption, that movements in output would cause an increase in savings which would be equal to the original increase in investment. This idea, called ‘Mr Meade's Relation’ in Kahn's article, was written down in a note which was subsequently lost. It has become fundamental to our understanding of the multiplier process, and is explained in all basic macroeconomics textbooks. Meade (1993) explains the way in which his approach was complementary to that of Kahn. This approach was useful to the Circus, in that it showed how Kahn's multiplier process was a flex-output version of the fixed-output argument of the *Treatise* explained above.

Meade once described the second step of his demonstration to me as follows. ‘I said the following to the other members of the Circus. “Haven't any of you read Marshall's *Principles of Economics*? In that book, in the short run, the economy lies on a short-run, upward-sloping, supply curve. But that curve adds an extra equation to the model. This means that – in comparison with the model in the *Treatise* – we can make *both* prices *and* output endogenous at the same time.”‘ This second idea of Meade's is explained, with much less clarity, in Kahn's article. Once it was properly understood, it led the Circus to the view that it is primarily variations in the level of *output* which bring savings into line with investment, and so re-establish the conditions of macroeconomic equilibrium, rather than variations only in the level of *prices*, as had been supposed, unsatisfactorily, by Keynes in the *Treatise*. Kahn himself warmly acknowledged his debt to Meade, both in his article of 1931 and in his fascinating account of the period published in 1984 and called *The Making of Keynes’ General Theory* (Kahn, 1984).

Meade stated (Keynes, 1971–88, vol. 13, p. 342) that when he returned to Oxford in 1931 he took back with him in his head ‘most of the essential ingredients of the subsequent system of the *General Theory*’. In his ‘Simplified model of Mr. Keynes’ system’ (Meade, 1937) Meade set out these ‘essential ingredients’ in a system of eight equations, which included those of the IS–LM model. It is known that Hicks saw a draft of this paper before he prepared his own celebrated article explaining the IS–LM system (Hicks, 1937); indeed Hicks uses Meade's notation in his presentation (see Young, 1987).

Meade's ‘simplified model’ is more general than that of Hicks, because it includes the upward-sloping supply curve discussed above. It therefore enables one to see *much* more of what is going on in the *General Theory*. But Meade's article is very difficult to understand, because it takes so much for granted. One can read it carefully without ever seeing the point that Hicks was concerned to make about the relationship between Keynes and the ‘classics’, and it was Hicks who invented the famous diagram to explain this point. This appears to be the first of a number of occasions during Meade's career on which he would set out a fully specified piece of economic theory, only to find that someone else would extract a simple, essential, idea from what Meade had written, publish it, and become famous as a consequence.

Meade taught in Oxford until 1937. During this period he synthesized with great clarity the ideas of the Keynesian Revolution in *An Introduction to Economic Analysis and Policy* (1936), published almost simultaneously with the *General Theory*. This was the first-ever economics textbook: until then books had been written as a means of expounding new ideas in economics. Many undergraduates in Cambridge at that time were confused by the turbulent debate concerning the Keynesian Revolution which swirled around them, and bemused by the associated misunderstandings, a number of which seemed to be deliberate. Subsequent oral tradition in Cambridge maintained that many of these people found Meade's book exceptionally helpful, since it cut straight through all of these difficulties.

Interestingly, the Keynesian model is expounded in Meade's book using an exogenous rate of interest.

That is, Meade set out the ‘Keynesian-cross’ version of Keynes's model, rather than setting out the full IS–LM system. My own view of the reason for this is that Meade never really believed in the LM curve and so thought, like many of us now do, that the IS–LM system was a distraction. The reason that I say this is that in Meade (1937) he discusses the (realistic) possibility that banks might adjust the quantity of money, in the face of shocks to the economy, so as to keep the interest rate unchanged, unlike what happens in the IS–LM system (except in an extreme case). This view of his was in turn based on the analysis in his first published article (Meade, 1934), in which he invented the money-multiplier – quite independently of the work in the United States begun by Phillips (1920) – and in which he carefully showed what banks would need to do in order to behave in this way. That Meade should have presented the Keynesian system in this way right back at the beginning, even although he fully understood the IS–LM system, has implications for understanding why – as discussed below – he proceeded the same way when he wrote *The Balance of Payments* in the late 1940s.

Economic Analysis and Policy also contains a discussion of longer-run growth, and presents an exposition of the Ramsey model of optimal growth. This was nearly ten years before Harrod and Domar invented what now looks like a very primitive version of growth theory, and long before the famous papers of Solow (1956) and Swan (1956), who invented a simplified version of the Ramsey model, in which the savings rate is *exogenous*. In two key pages, published 20 years before the papers by Swan and Solow, Meade explains how the optimal savings rate could be chosen *endogenously* in order to produce a welfare-maximizing growth process. These pages provide an astonishingly clear verbal exposition of the first order-conditions which must be satisfied if growth is to be optimal.

A second edition of *Economic Analysis and Policy* was published in 1937. This gave an exposition of the new ideas in imperfect competition, invented by Edwin Chamberlin and Joan Robinson, which were challenging Marshallian microeconomics in the 1930s (see Shackle, 1967). These ideas only really bore fruit in mainstream microeconomics, and in macroeconomics, in the 1970s and 1980s, after the rise of game theory. But they had a more or less immediate effect on Meade's work in macroeconomics, as I will show below.

Economic Analysis and Policy also includes a section on problems of international order and disorder, which shows that, already as a young man, Meade (unlike many in Britain and America at this time) was thinking about the macroeconomic problems of the world, as distinct from those of an individual nation. In this part of the book, Meade expands on the ideas on international macroeconomics, which were already to be found in Keynes's *Treatise*, as I have described above. (Notably, Joan Robinson, 1937, and Roy Harrod, 1933, were also busy doing the same thing.) Meade's volume ends with a prescient chapter on the economic causes of war. It is chilling to re-read this chapter, written three or four years before the outbreak of the Second World War. One also realizes that many of the problems connected with internationally ill-coordinated macroeconomic policies, which emerged in the world economy in the 1980s, and which have re-emerged at the beginning of the new millennium, are very like those which Meade wrote about nearly 75 years ago.

Throughout his time at Oxford, Meade was actively involved with the group of Fabian socialist intellectuals who were helping the British Labour Party to recover its sense of purpose, after the disastrous collapse of the Labour Government in 1931. Meade contributed to discussions across the same wide range of macroeconomic, microeconomic and international issues that he had treated in *Economic Analysis and Policy*, and he was most influential in his advocacy of expansionary Keynesian

policies (Durbin, 1985, see especially pp. 136–44, 194–8, 211–12 and 220).

Meade elaborated on this last theme in *Consumers' Credits and Unemployment* (1938). In this work, he proposes Keynesian demand management in the form of *automatic*, countercyclical variations in taxation to stabilize macroeconomic fluctuations. This book foreshadows both the Full Employment White Paper of 1944, and the nominal-income targeting project on which he worked for ten years from 1978, both of which are discussed in more detail below. *Consumers' Credits and Unemployment* is perhaps the earliest official published advocacy of fine-tuned Keynesian policies.

In 1937 Meade went with his wife and young family to Geneva, where he was to stay for three years, as an economist for the League of Nations. Meade often spoke with admiration of the remarkable band that were assembled there, which included Tinbergen, Koopmans, Haberler, Nurkse, and Marcus Fleming. His job was to prepare, more or less single-handed, the *World Economic Survey* (the forerunner of the present IMF *World Economic Outlook*) and he transformed this publication. Keynes was much influenced by Meade's work in Geneva. In the *General Theory*, Keynes had made use of an upward-sloping short-run supply curve, as described above in my account of the discussions of the Circus, which relies on goods being produced in a manner subject to diminishing returns, and being supplied under competitive conditions. Keynes's important article of 1939 makes extensive reference to Meade's work, and then goes on to argue that the quantity produced in an economy is determined by demand, even if there are constant marginal costs. But this was inconsistent with the competitive analysis that Keynes had utilized in the *General Theory*, which requires the assumption of increasing marginal costs. This article by Keynes was exceptionally difficult to understand. It was extensively discussed in Cambridge in the 1970s, when people were comparing Keynes's macroeconomics with that of Kalecki, who had carefully evaded this difficulty by assuming 'markup' pricing. The confusion was resolved only by the arrival of the classic Dixit and Stiglitz (1977) paper. That paper brought Chamberlin's ideas about imperfect competition into macroeconomics, suggesting a setup in which each individual profit-maximizing producer faces a downward-sloping demand curve and sets prices above marginal costs, but in which the existence of free entry prevents the emergence of monopoly profits. I know, from working with Meade from the late 1970s onwards, that the standard macroeconomic model which he used daily, as part of his mental equipment, had this feature, and I believe that, unlike many others, this had been true for him since the mid-1930s, when he wrote the material on imperfect competition in *Economic Analysis and Policy*, which I described above. It is probable that this framework influenced his empirical work in Geneva, and thus influenced Keynes's article of 1939.

The war finally caused the Meades to leave Geneva for London in 1940, with three young children, the smallest a three-week old infant. They set out in a small car for one of the Channel ports, not knowing that at this very time the Germans had broken through at Sedan. After an increasingly desperate journey the family ended up in Nantes as refugees, and finally crossed the Channel in an RAF 'transport ship' – a converted tramp steamer – at the very time of the Dunkirk evacuation.

2 The war years, 1940 – 45: building the post-war world order

On his return to Britain, Meade was brought into the Economic Section of the Cabinet Office. There was a grim feeling of impotence amongst the economists, who wished to do something for the war effort. Keynes's *How to Pay for the War* (Keynes, 1940) had just been published. Meade therefore set to work

on a set of national accounts, so that, at the very least, those making war policy might be able to attach some numbers to Keynes's ideas. Richard Stone who had only recently graduated in economics from Gonville and Caius College, Cambridge, was brought in to help with this work. Together they produced what is probably the first full logical structure of ‘double-entry’ national accounts (Meade and Stone, 1941; 1944). Meade recalled how in the statistical work he quickly became Stone's research assistant – and so began Stone's work on national income accounting that eventually led to another Nobel Prize. During his subsequent time at the Economic Section, Meade worked in three crucial areas. He was to become Director of the Economic Section from 1945 to 1947.

First, Meade was involved in the planning for post-war international monetary arrangements. He participated in the initial excited responses in Whitehall to Keynes's ‘Clearing Union’ plan for a new post-war international monetary system (Keynes, 1971–88, vol. 25, pp. 41–67; see also van Dormael, 1978). He became a member of the British delegation to Washington in September 1943 which discussed these issues with Harry Dexter White and others (Keynes, 1971–88, vol. 25, pp. 338 ff.). And he took part in the subsequent British deliberations, leading up to the Bretton Woods conference in 1944 at which the International Monetary Fund was established.

I have described the analytical content of these negotiations in some detail in Vines (2003), drawing on the wonderful historical account by Skidelsky (2000), and on the papers of Keynes and Meade. Keynes's policy objectives were to create a post-war global system in which full-employment policies could be adopted by the Allied nations, and in which such full-employment policies could be reconciled with the requirement that their trade balances not get too far out of line. Keynes's initial response to this problem was a highly illiberal one: balance-of-payments restrictions should be the mechanism which re-equilibrated exports with imports after any negative external shock to a country (through tariffs, quotas, and ‘managed’ trade). He was persuaded away from this view by an ‘outstandingly able group of economists’ (Williamson, 1983a, p. 91) which included Meade and also Marcus Fleming, Roy Harrod, Lionel Robbins, and Dennis Robertson. This group managed to convince Keynes that exchange rate devaluation should be the adjustment device. Keynes vacillated literally for years on the issue, deliberately suspending judgement, drawing forth from this talented group an extraordinary collection of papers, particularly on the tariffs-versus-devaluation issue (Keynes, 1971–88, vol. 26, ch. 2 and pp. 239–327). James Meade once told me that, one day in a particularly tedious meeting at the Board of Trade in 1944, Keynes scribbled a note to him to the effect that he (Keynes) was, at last, intellectually converted to a regime in which external adjustment would be achieved by exchange-rate change.

Second, when Keynes produced his Clearing Union plan, Meade quickly produced a project for a ‘Commercial Union’ as a companion piece. It was on the basis of this document that the debate in Whitehall on post-war commercial policy (concerning such sensitive issues as imperial preference and the use of import restrictions on balance-of-payments grounds) took place. Meade devoted much time to drafting and redrafting these ideas and, as he said, ‘helping to get them through Whitehall’. And it was to promote these ideas that he was a member of the September 1943 delegation to Washington (mentioned above), and he was subsequently a member of the British delegation to the international conferences in London in 1946 and in Geneva in 1947 which worked on a charter for a proposed International Trade Organization (ITO). Although in the end the ITO proved to be unacceptable to the United States, the Geneva conference resulted in a General Agreement on Tariffs and Trade (GATT) which took on many of the projected functions of the ITO (see Keynes, 1971–88, vol. 26, ch. 2). And the

GATT was eventually turned into the World Trade Organization (WTO) in 1994.

These international discussions laid down, amongst other things, the conditions under which nations should be permitted to form regional free-trade-areas, in which discriminatory regional preference is allowed to overrule the most-favoured-nation rule for international trade which lies at the centre of the WTO, and which lay at the centre of the GATT. Those discussions duly led to Article 24 of the GATT (and to a similar provisions in the international agreements which underpin the WTO). The technical discussions on this Article were particularly difficult, since the relevant theory by Viner and Meade, on 'trade creation' and 'trade diversion', had not yet been invented. (This theory is discussed in Section 3 below.) The discussions also contained much which was non-technical, which dealt more fundamentally with the nature of the international trading system. On one occasion, Meade told me, he could not understand why a senior US official – I believe that it was Dean Acheson – was speaking up so strongly against imperial preference, and yet so much in favour of Britain joining up with European nations, a joining-up which has, in due course, led to the European Common Market, and ultimately to the creation of the European Union. 'I have relatives who are farmers in New Zealand', said Meade, 'who sell their lamb to Britain. They a natural part of the British economic system. Why should we not have an Imperial Free Trade Area which includes them? This would be just like your setup in the US, in which you have a free trade area which includes all 50 of your states?'. 'But there is a lot of water between Britain and New Zealand', replied Acheson. 'There is also quite a lot of water between Britain and France', replied Meade.

If we take these two activities together, it is clear that Meade was one of the architects, on both the monetary side and the trade side, of the liberal world economic regime which sustained the long post-war boom in the Western world from 1945 to 1973. Meade always believed that *both* pieces of this regime stand or fall together. Free trade would – he thought – be resisted if there were severe global macroeconomic imbalances. (This point became clear once again in the mid-1980s, and it is becoming even more clear in the mid-2000s. But conversely, if there is not free trade then macroeconomic order will be difficult to maintain, since devaluation will tend to be much less effective at adjusting trade imbalances. Meade summarized this point clearly, in a diary entry which he made on 31 December 1944 (Meade, 1988–90, vol. 4, p. 22.) He emphasized 'the need for flexible exchange rates to adjust balance of payments [to avoid pushing the burden of adjustment onto] rigid trade controls … in a world in which internal wage levels were not easily reduced. [But such adjustment might be] more easily acceptable if it was preceded by an international agreement to lower trade barriers, since in that case smaller movements in exchange rates would be required'. This belief – that macro management and micro liberalism should go together – had already informed his work in the 1930s. It would form the central organizing principle for the work that Meade did at the LSE on international economics, which I discuss immediately below. As noted in the introduction and conclusion to this article, it recurs again and again throughout his work.

At the meeting at the Board of Trade in 1944, to which I referred above, Keynes followed up his scribble with a sketch, on the back of an envelope, of the desired features of the whole international system that he and his colleagues were trying to build (see Vines, 2003). This sketch went something like the following.

Objective	*Instrument (s)*	*Responsible authority*
Full employment	Demand management (mainly fiscal)	National governments

Balance of payments adjustment	Pegged but adjustable exchange rates	International Monetary Fund
Promotion of international trade	Tariff reductions etc.	International Trade Organisation
Economic development	Official international lending	World Bank

Keynes was aware that this plan would need to work, not just for individual countries, but for the global system as a whole. (It would have been surprising if someone who had invented macroeconomics did not take such an overall, systemic view.) I discuss in some detail in Vines (2003) how Keynes feared that difficulties in the balance-of-payments adjustment process might impose, on deficit countries, an obligation to deflate demand below full employment, something which might not be matched by symmetrical over-expansion by surplus countries, and might thereby create pressures towards global deflation. I also describe how Keynes differed in this view from Harry Dexter White, his US counterpart in the Washington negotiations, who feared an outcome in which the International Monetary Fund would be so expansive with liquidity that there would be a great post-war inflation, worldwide. In that article I claim that, during these discussions, Keynes' negotiating strategy in pursuit of a balanced global outcome was underpinned by a significant theoretical understanding of what was going on. In particular I maintain that (*a*) Keynes took from his *Treatise* something akin to an IS–LM–BP model (without the flaw in the analysis of the *Treatise*, which had been fixed up by the invention of the multiplier and the publication of the *General Theory*), and (*b*) Keynes, as he negotiated, was using something akin to a two-country version of that model to understand what was being discussed. These two claims of mine are vital for a proper understanding of the work that Meade did at the LSE on international economics, which I discuss below.

I will be brief about Meade's third activity while he was in the Economic Section during the war, although it was important. Meade's paper 'Internal Measures for the Prevention of General Unemployment', dated 8 July 1941, reached the Intern-Departmental Committee on Post-War Internal Economic Problems in November, and as Skidelsky (2000, p. 270) says 'never quite lost its place as front-runner in the development of post-war employment policy'. This was, in effect, the first draft of what finally became the Full Employment White Paper, published as an official paper with the title of *Employment Policy* (Minister of Reconstruction, 1944), which laid the basis for a transformed macroeconomic management within the United Kingdom after the war. In the drafting of this document, there were long discussions between Meade and Keynes on the possibility and desirability of automatic fiscal fine-tuning (see Keynes, 1971–88, vol. 27, pp. 207–19 and 308–79; Wilson, 1982). Meade advocated countercyclical variations in social security contributions; this proposal featured in the final White Paper and was endorsed by Keynes. That idea would remain more or less an article of faith for Meade, and underpinned his work in inflation targeting which I describe in Section 5.

3 The LSE, 1947 – 58: international economics

Meade became Professor of Commerce (with special reference to international trade) at the LSE in 1947, where he was to stay for ten years, and where his great work on international economics was done. It had been Meade's intention to begin his time back at a university by rewriting his textbook *Economic Analysis and Policy*. But this was not to happen. Someone once observed to me how different the teaching of our subject might have been if Meade had actually rewritten his book, rather than leaving the

field open for Samuelson's great *Principles* book (Samuelson, 1948), which was not published until 12 years after Meade's book had first appeared. In his Nobel Prize autobiography Meade (1977a) explained why this did not happen.

> I realised that it might be necessary to [rewrite the book] in more than one volume. So, as I was appointed at the LSE to teach international economics, I started on *The Theory of International Economic Policy*. It grew into my two books, *The Balance of Payments*, and *Trade and Welfare*, with their two mathematical appendices….These books took up practically the whole of my ten years at the LSE; but even so they did not cover the whole of the international problem…. My original project was over-ambitious; but the part which I did manage to cover was sufficient, eventually, to gain for me the Nobel award. (Meade, 1977a.)

It is characteristic of Meade's modesty that he should describe the work for which he received the Nobel Prize as an attempt to rewrite a textbook.

The balance of payments

In his introduction the *The Balance of Payments* (Meade, 1951–5), he had, equally modestly, described it as a book which ‘does not claim to make any significant contribution of original work in the fundamentals of pure economic analysis’ (p. vii). This is something which turns out not to be true. Meade also said of his book that it is one which has an ‘indebtedness to the ideas of Lord Keynes [which] is too obvious to need any emphasis’ (p. ix). Many people have said to me that they think that this remark is there because the book contains lots of ‘multiplier Keynesianism’, of a kind derived from the *General Theory*, which was still new and exciting in the 1950s. If that reading is correct, the generous acknowledgement of Keynes's contributions would not be particularly significant. But I believe the remark meant something rather different and rather more interesting. On more than one occasion Meade said to me that all he had done in this book was to write down what he learned from his work with Keynes during the war, about how to understand the international position of the British economy, and about how the world economy should be managed. That is a much more thought-provoking connection to acknowledge. (He did also admit that he had added quite a lot of algebra in the appendix.)
Volume I of the *Theory of International Economic Policy* (Meade, 1951–5) was entitled *The Balance of Payments*. There were three new features of this book. First, at the level of technical analysis, it integrated income effects and price effects so as to study the balance of payments in a general-equilibrium framework. In doing so, it extended the theory of the balance of payments beyond its traditional identification with the current account to so as to consider the overall balance by including international capital movements. Second, it had a policy orientation, focusing on two instruments (exchange rate adjustment and domestic demand management) and two targets (internal balance – that is, full employment – and external balance – that is, a satisfactory overall balance of payments position). Third, Meade carried out his tasks in this book using a two-country model rather than merely developing the analysis for a single open economy.
At the level of technical analysis, investigations of the effects of exchange rate change had previously been separated from investigations of Keynesian income–expenditure theory. The former was normally

based on the assumption of constant incomes, and carried out in terms of Marshallian partial equilibrium concepts, using the elasticities approach. (For a few key exceptions, see Robinson, 1937; Harrod, 1933; Laursen and Metzler, 1950; and Harberger 1950.) The latter was normally carried out using fixprice models, which led to the ‘absorption’ approach to the balance of payments, published by Alexander in the same year as Meade's book (Alexander, 1952). The formal integration of the elasticities approach and the absorption approach to balance-of-payments theory, which Meade achieved, was very important.
At the level of the theory of economic policy, Meade's basic idea – that, if internal and external balance are to be attained simultaneously, then two policy instruments are needed (exchange rate adjustment and the management of domestic demand) – was not a new one to him. He would have been familiar with this idea from his work with Keynes at the beginning of the war on Keynes’ book *How to Pay for the War*, and also from his work with Keynes on Britain's financial crisis at the end of the war. (See Vines, 2003, for a detailed discussion of this claim.) Furthermore, as noted at the beginning of this article, many of the necessary components of this idea are already to be found in the *Treatise*, published more than 20 years earlier; and many of them are also to be found in *Economic Analysis and Policy*, published 15 years earlier, and in the work of Robinson and Harrod referred to above.
Indeed, this idea now seems deeply obvious to all of us. But that is only because we know the Swan diagram, which collapses all of the complex analysis by Meade into just one diagram (Swan, 1963), just like Hicks had done with the IS–LM system. At the time, Meade's idea was absolutely revolutionary. In reminding ourselves of this fact, we should not forget that Tinbergen was awarded the Nobel Prize in 1969 for stating a more general, but equally obvious, idea – that to achieve *n* targets simultaneously one (normally) needs *n* instruments. (Tinbergen's analysis was developed simultaneously to, and independently of, Meade's book.) And it took a *very* long time for Meade's idea to be learned. For many years after the Second World War in the UK, full employment policies appear to have been carried out without sufficient regard for their effects on the balance of payments, and they often needed to be reversed at times of balance-of-payments crisis. Also, to take another example, many policymakers still continue to forget that if a devaluation is to improve a current-account deficit then it must be accompanied by a reduction in domestic absorption relative to domestic output, so as to release the resources needed to improve the trade account.
All of what I have said so far is about open-economy macroeconomics. We should also notice the third important feature of Meade's book which I have mentioned above – that it develops everything for a two-country world, and discusses *global* macroeconomics, not just *open-economy* macroeconomics. You might think that this would be the obvious way to proceed. After all, any treatment of *trade* theory is normally done this way, by analysing trade in a two-country world, and this is what Meade himself would do in Volume II of the *Theory of International Economic Policy*, published a few years later. Furthermore, all of us have now lived through the 1980s, in which we studied the effects of Reaganomics on Europe, something which clearly required a *two*-country model. (We are all at present trying to understand the interrelationships between the United States, East Asia and Europe, which seems to need a *three*-country model.) But nobody had ever done global macroeconomics before Meade wrote his book. As I note in Vines (2003), even Keynes, when writing down of the key components of the necessary theory in the *Treatise* in 1930, wrote about nearly everything for a single open economy rather than for a global system. But in Vines (2003) I also develop the argument, described at the end of Section 2 above, that Keynes worked out, informally, aspects of the needed two-country model when he

was negotiating with Harry Dexter White about the establishment of the IMF. It is my belief that Meade had seen, when working on these negotiations with Keynes, that such a model was necessary for a systemic discussion of global, policy-related, questions. This is my view of why he set out his analysis in this way, even although doing this made his book *much* harder to read.

Harry Johnson made two important criticisms of Meade's book at the level of technical analysis. The first, developed in Johnson's long review of the book (Johnson, 1951), concerned the treatment of saving in the model. Meade assumes that the amount of real saving coming from a given real income is independent of the terms of trade. Laursen and Metzler (1950) and Harberger (1950) had already shown how to avoid this mistake; many practitioners of open economy macroeconomics still forget how hard it is to defend what Meade assumes.

Johnson's second criticism, made in the paper which Johnson published at the time the Meade received the Nobel Prize (Johnson, 1978), leads in a valuable direction. What Johnson said was that Meade did not succeed in fully integrating real and monetary theory in his book. What he meant by this is that Meade assumed a flexible money supply policy designed to maintain a given exogenous rate of interest, with monetary policy changes being expressed in terms of (exogenous) interest rate changes. In the IS–LM–BP model subsequently developed by Fleming (1962) and Mundell (1962), the interest rate instead becomes endogenous, so as to ensure that the economy lies on a given LM curve. Under fixed exchange rates, the LM curve moves around because of the endogeneity of the money supply, unless monetary sterilization is possible, in a way which was analysed in the monetary theory of the balance of payments (which led to fame for Harry Johnson). Under floating exchange rates the money supply is held constant, and the interest rate and the exchange rate together become jointly endogenous, along with output. It seems odd that Meade did not think to make the interest rate endogenous by introducing an LM curve into his model, since this is exactly what he had done, more than 15 years earlier, when he had explained Keynes's *General Theory* to the world. Had Meade done this, surely he would have instantly invented the Fleming–Mundell model. My suggestion of the reason why that didn't happen is related to my view, stated earlier, that Meade never really believed in the LM curve. In his subsequent work, which I discuss below, work that was contemporaneous with that of John Taylor, Meade allowed for the endogeneity of the interest rate without having to make the ridiculous assumption of a fixed money supply – essentially by supposing that the interest rate would follow something like a Taylor rule. One can easily build a Fleming–Mundell-like model with a Taylor rule in it, instead of an LM curve. I believe that, although Meade did not like the way that the interest rate was made endogenous in the LM curve, at the time he was writing *The Balance of Payments* he could not yet see how to replace the LM curve by a policy–behaviour relationship like the Taylor rule. This is why, I think, it was not possible for him to take the next step and construct something akin to the Fleming–Mundell model.

Trade and welfare

The second volume of the *Theory of International Economic Policy* was titled *Trade and Welfare*. In this, Meade presented a systematic analysis of neoclassical trade theory, essentially the theory of Heckscher and Ohlin, with the latter of whom he shared the Nobel Prize. But he combined this with an analysis of trade in factors – both capital and labour. He discusses policy in this book – the issue of protection versus free trade – but in relation to the movement of *both* goods and factors of production.

Meade's inclusion of international factor movements, in the main corpus of his theory of international trade, was innovative, and chimed with growing concerns, at the time, and since, about the ‘brain drain’, foreign direct investment, and the multinational corporation. Surprisingly, very few expositors of trade theory have followed Meade in explaining trade theory in this way, so that these subjects are more normally studied in isolation. Perhaps, again, it is because Meade's integrated analysis makes for such difficult reading.

The book made a number of important innovations at the level of technical analysis, whose influence in economic theory went far beyond the study of international phenomena.

First, Meade introduced a new method for measuring small changes in welfare, which was a generalization of Marshallian consumer surplus, with its attendant limitations. And he then went on to present a whole new approach to welfare economics, defining overall welfare as an appropriately weighted sum of individual welfares. Johnson (1978) describes it as a brilliant feat of imagination for Meade to see that he could take over this approach from Fleming (1951) and then rework it into a powerful general technique for welfare analysis of practical policy problems. Doing this enabled Meade to escape from the nihilism of the new welfare economics, which worked in terms of ‘potential welfare’ and the ‘compensation principles’, but which made it difficult to say anything practical at all about the welfare effects of economic policy changes. Nearly all of us now do welfare economics in the manner pioneered by Meade.

Second, Meade invented the theory of domestic distortions in order to show that a move towards free trade may not be welfare-improving if there are already distortions elsewhere in the economy. This idea was later carried forward by Bhagwati and Ramaswami (1963) and Johnson (1965). Meade invented the theory of the second best in his discussion of these ideas (inventing the technical term ‘second best’ as he did so), and explored many of its implications. As Corden (1996a) says, it is hard to see how something which now seems so obvious needed to be invented. Jacob Viner, in a book on customs unions (Viner, 1950), had already established the distinction between trade creation and trade diversion in the creation of free trade areas. This also seems totally obvious to us now, but it really only became obvious after Meade published the *Theory of Customs Unions* (1955b), which clarified and extended Viner's distinction, and located it within his general theory of the second best.

Finally, it is important to add that the *Trade and Economic Welfare* includes a discussion of the meaning of optimum population and of optimum savings and of the relationship between these two concepts. This discussion, too, broke new ground.

Phillips

While at the LSE, Meade also did something else which was stunningly important: he brought Bill Phillips into economics. Meade once said to me that Phillips was the closest to genius of anyone that he had ever known. Phillips's really important work in economics was on the use of control theory for macroeconomic stabilization purposes, rather than in estimating the ‘Phillips curve’ (for which he is so famous, but which he did in a rush in a few weeks, just before leaving London go on sabbatical leave). Phillips had trained before the war as an electrical engineer (having previously left school without any formal qualifications), and immediately after the war he had graduated from the LSE with a third-class honours degree in sociology. One day, soon after receiving this unremarkable qualification, Phillips

explained to Meade that he wished to build a strange ‘water-machine’ model of a macroeconomic system. Meade listened patiently because ‘the pipes seemed to have the right labels’, and so encouraged Phillips to build the machine, offering Phillips the inducement that he could demonstrate it at Lionel Robbins's prestigious seminar for graduate students. The machine was duly built, and it is described in Phillips (1950). A brilliant performance followed at the Robbins seminar, in front of most of the London economics professoriat, who had got word of what was coming. In the course of that seminar, said Meade, Phillips gave the best exposition that anyone present had ever heard of the Keynes-versus-Robertson debate, about whether the rate of interest was determined by liquidity preference or by the supply of, and demand for, loanable funds. This, said Phillips, was an argument about stocks versus flows; he then illustrated his claim by displaying the effects of water sitting in tanks, on the one hand, and water flowing through pipes, on the other. Phillips was duly instructed to write up his machine in a Ph.D. thesis, and John Hicks, who was by then Drummond Professor of Political Economy in Oxford, was asked to examine the thesis so as to ensure that somebody with a third-class degree in sociology could be given a Ph.D. in economics with a clear conscience. Phillips was then promptly brought on to the staff, and became one of the professors in the department within a few years. In Vines (1996) I give a detailed account of how the Phillips machine works. In particular I describe the stock–flow intuition which it provides, which is almost impossible to obtain any other way than by looking at this machine in action, and which certainly cannot be obtained from modern computer simulation models. As I describe below, Meade was closely involved with the use of the machine.

Work on his machine led Phillips to write his classic article on the use of control theory to help stabilize an economy (Phillips, 1954). This paper argued that a feedback policy can have destabilizing effects if the instrument of policy responds too strongly to a disturbance to the target of policy, and there is a lag in the effect of the instrument on the target. In a subsequent paper, Phillips concluded on a cautious note: ‘the problem of economic stabilisation is, even in principle, a very intricate one, and … a much more thorough investigation of both theoretical principles and empirical relationships would be needed before detailed policy recommendations could be justified’ (Phillips, 1957, p. 275). Meade was involved with the preparation of both of these papers, and he agreed with their conclusions.

Milton Friedman came to hold similar views on the potentially destabilizing effects of macroeconomic policy, at a very similar time (Friedman, 1953). He went on to declare that active macroeconomic policymaking is too difficult to do properly and, worse still, too dangerous. Friedman's response to this problem was to set off in pursuit of his holy grail: a non-interventionist macroeconomic policy.

Meade's and Phillips's response to this problem was rather different. Phillips thought that it would be possible to do good macroeconomic policy, but only if the policy was carefully designed. Indeed he ended his 1957 paper on an optimistic note. He called for the use of multivariable control methods, to regulate multiple objectives in an economy in the face of multiple disturbances, and he noted that methods for doing this were just, in the late 1950s, becoming available. He also called for the econometric estimation of the parameters of the econometric model which would be necessary for the study of such regulation. Meade said to me on more than one occasion that he regarded his own last big project, carried out more than 20 years later, and described in Section 5 below, as a response to Phillips's call to action.

Other activities

At the LSE Meade acquired a further generation of very able young disciples, drawn from many countries, who included Max Corden, Richard Lipsey, Robert Mundell and Harry Johnson, the last of these ‘at one remove’ (Johnson, 1978, p. 66). Meade had persuaded Phillips to build two of his water machines, joined together by an ingenious model of a foreign-exchange market. Peter Kenen (now retired from Princeton) vividly remembers a graduate student seminar in which he was asked to run fiscal and monetary policy for the United States on one of these machines. At the same time on the other machine Richard Cooper (now retired from Harvard) was required to run fiscal and monetary policy for Europe. They made the world develop unstable cycles – and spilt a lot of water (Vines, 1996). By such means did that generation of students learn about the need for an international coordination of macroeconomic policies, 25 years before the subject became fashionable.

During this time Meade also went on sabbatical leave to Australia. With Eric Russell of Adelaide he wrote a short theoretical analysis of the effects of the Korean War boom on the Australian economy, via its effects in raising the world price of wool, which was – at that time – a major export commodity for Australia (Meade and Russell, 1957). This article is one of the most profound pieces ever written about that economy. (Harcourt, 1982, ch. 21, describes how it came to be written: Meade became the expositor of Russell's perceptions, which then existed only in note form.) The authors first explain the ‘Stolper–Samuelson’ theorem concerning the effects of protection on income distribution. Their exposition is different from the one given by Stolper and Samuelson, and much more like that to be found in the original source of that theorem – the Brigden Report of 1929 on the Australia tariff (Brigden et al. 1929). This is because it discusses the effects of protection on income distribution in an Australian ‘dependent-economy’ model, in which there are non-traded goods as well as traded goods. (See Vines, 1994.) Meade and Russell then use this model to examine what has subsequently become called the ‘Dutch Disease’. They show how an export boom can, by raising wages, give rise to cost pressures for the protected sector, which can cause it to contract, even at a time of general boom. Their paper directly influenced the subsequent discussion of this problem, first in Australia in the 1970s (see Gregory, 1976), and then worldwide in the 1980s (see Corden, 1984). This ‘problem’ has returned in a big way in the early 21st century, with the high prices of primary commodities, worldwide.

4 Cambridge, 1957 – 69: growth theory

Meade became Professor of Political Economy at Cambridge in 1957, when he succeeded his teacher, Dennis Robertson. When Meade moved to Cambridge, growth theory was in the air. His useful book, *A Neo-Classical Theory of Economic Growth* (1961b), ‘brings the subject within the range of the undergraduate student, covers a number of aspects (such as the presence of the fixed factor land) usually omitted in more high powered mathematical treatments, and presents in detail the mathematics of a two-sector growth model’ (Johnson, 1978, p. 79). He also made advanced contributions to growth theory (1965, with Frank Hahn, and 1966). But his lasting contribution in this area is his essay *Efficiency, Equality, and the Ownership of Property* (1964). This ‘provides a very suggestive account of the forces underlying the accumulation of capital and the relationship between earned and unearned income’ and ‘stimulated much of the revival of interest in this subject, at least in the United Kingdom’ (Corden and Atkinson, 1979, p. 530). Meade regarded this as, in many ways, his best book, because it puts together

into a single synoptic framework his views on economic growth, on the microeconomic role of the price mechanism, on the size and the genetic composition of the population, and on the distributional implications of property ownership. He analysed further the interplay of these last factors in his Keynes Lecture on ‘The Inheritance of Inequalities: Some Biological, Demographic, Social, and Economic Factors’ (Meade, 1973c).

In 1960 Meade visited Mauritius and contributed to a report to the Governor, applying for the first time his ideas on growth theory and on population policy to the problems of a less developed country (Meade 1961a). His prediction for Mauritius of Malthusian stagnation turned out to be spectacularly wrong, in interesting ways.

In 1973 Meade also began in Cambridge a grand scheme of work entitled *The Principles of Political Economy*. The purpose of this series of books was ‘to bring the best of modern theory within the range of an intelligent and educated adult, the volumes being intended to tackle successively departures from the assumptions of a model of perfect static general equilibrium’ (Johnson, 1978, p. 79).

5 Retirement, 1969 – 95

In 1969 Meade took early retirement, five years before the statutory retiring age. As Atkinson and Weale (2000) say, ‘[a]lways the most gentle and courteous of men, he had found extremely depressing the quarrels between those labelled “post Keynesian” and those in the Faculty who researched the mainstream of Economics’. But he did not stop working; indeed the next quarter century was to be one of his most productive.

Meade initially worked on the *Principles of Political Economy* but subsequently, perhaps sensing that this enterprise did not provide the best outlet for his unflagging energy, he turned to other schemes. The *Intelligent Radical's Guide to Economic Policy* (1975a) had ‘wide influence in Britain, particularly on debates about economic planning’ (Corden, and Atkinson, 1979, p. 530). In it Meade returned to a theme set forth in his *Planning and the Price Mechanism* (1948a) which I summarize in my concluding section below.

An expenditure tax

Meade's first big activity in retirement was to chair a committee which was established by the Institute of Fiscal Studies, to look into the structure of the UK tax system and to advise on how it might be simplified. The report, entitled *The Structure and Reform of Direct Taxation* (1978a), is a monumental study of British personal taxation. As Atkinson and Weale (2000) say,

> The Committee observed that the tax system at the time was a mixture of taxes on income and taxes on expenditure, and concluded that it should be more desirable that tax should be levied on one or the other, all but one of the Committee favouring a shift towards an expenditure tax. In the 20 years since the report was written, exemptions for saving have appeared in the form of TESSAs, PEPs and ISAs, and the shift to indirect taxation has been a move towards a tax on expenditure. In this respect, the Report was influential, but its lasting value lies in the outstandingly high quality of the analysis.

Meade was fortunate in having as assistants for that committee three additional able young scholars, John Flemming, John Kay and Mervyn King, who all subsequently achieved distinction in various aspects of public life.

A return to the theory of macroeconomic policy

In 1977 Meade returned to the great questions of national macroeconomic management, at the age of 70, when most people might have felt ready for a holiday. His work began with his Nobel Prize lecture entitled ‘The Meaning of Internal Balance’ (Meade, 1978b). It has been explained above how, in the *Balance of Payments* (1951) – one of the volumes for which Meade received the prize – he talked about the problems of reconciling internal balance (full employment) with external balance (a satisfactory overall balance of payments position). In his Nobel Prize lecture, Meade returned to question this framework, arguing that the concept of ‘internal balance can now no longer be taken merely to refer to the achievement of full employment, but must also make reference to the achievement of low and stable inflation’ (Meade, 1978b). He argued that it is not sufficient to rely on incomes policy, of the conventional kind which was still fashionable in Britain. The fundamental problem is that a commitment to ‘full employment’ removes the threat of unemployment as a response to over-rapid wage increases, and it is on this threat which wage and price stability in part depends. As a result, Meade argued that Keynesian policies should be ‘stood on their head’. Demand management policy should be responsible for the maintenance of a slow and restrained rate of growth of money incomes, so as to put a ‘lid’ upon inflationary pressures. Incomes policy, or, more generally, the ‘reform of wage-fixing’, should be used – he argued – not to hold down prices but to promote employment.

This lecture contained three striking claims.

First, Meade's assertion that demand management would, inevitably, be excessively expansionary and would thereby promote inflation was essentially the same claim as that made the following year by Kydland and Prescott (1978) – a claim which went on to help them, too, to win the Nobel Prize. Meade's claim was made five years before the macroeconomic implications of the Kydland and Prescott idea were properly worked out by Barro and Gordon (1983).

Second, Meade's claim that macroeconomic policy should be confined to ‘putting a lid on inflation’ implied that employment would no longer be determined by a macroeconomic policy which was promoting full employment. As a result the levels of employment and of unemployment would be determined in some other way. At any point in time, said Meade, the ‘reform of wage fixing’ could be taken as given, and that would determine what we would now call the non-accelerating inflation rate of unemployment, or NAIRU. Meade then said that unemployment would gravitate towards the NAIRU, using the following argument. If unemployment was lower than the NAIRU, then inflation would be rising. But if the rate of growth of money incomes was effectively controlled by policy, then this would mean that policy would need to ensure that output fell, so as to prevent the growth of money incomes from rising above target. That would cause unemployment to rise towards the NAIRU, which would – in turn – stop inflation from rising. Meade used a similar argument to describe what would happen if unemployment was above the NAIRU. This line of reasoning effectively made Meade a follower of Friedman, who had claimed, in his fundamental paper published ten years earlier, that macroeconomic policy could not itself control the level of unemployment (Friedman, 1968). Friedman's idea had been

publicly broadcast in Great Britain, by Prime Minister Callaghan, in a famous speech given two years before Meade's lecture. But at the time this idea was too revolutionary for most macroeconomists in Britain. It was still widely thought that only monetarists believed something like this; Bob Rowthorn had caused uproar amongst the Cambridge Keynesians by claiming something of this kind just a year before Meade gave his lecture (Rowthorn, 1977). Meade's lecture had the effect of detaching such a claim from its monetarist proponents, and began the process of making this claim mainstream in Britain, something which was eventually achieved by Layard, Nickell and Jackman (1991).

Third, Meade discussed how, exactly, demand management policy (that is, fiscal and monetary policy) should be used to achieve the required slow and restrained growth of money incomes. His answer was that this should be done mainly by fine-tuned changes in tax rates, which, as mentioned above, he had discussed in *Consumers' Credit and Employment* (1938) and as he had suggested in his draft of the Full Employment White Paper in 1944. This answer made him very *unlike* Milton Friedman. In a subsequent mischievous talk to the Royal Economic Society, Meade (1981) created a taxonomy, so as to compare his new proposals with orthodox Keynesianism on the one hand and monetarism on the other. His mischief was to make the monetarists end up on the far left of his taxonomy, and to make the 'old-fashioned' Keynesians end up on the far right.

Meade presented a draft of this Nobel Prize lecture to the Marshall Society – Cambridge's student economics society. I was a research student in Cambridge at the time. As I recall, we did not know that Meade had just been awarded the Nobel Prize, or that what we were hearing was a dry run for his lecture in Stockholm. His lecture was a bit too un-Keynesian for me, and I stood up and said so. Meade defended his claims to the (large) audience, using the argument that policy works well when each policymaker is given an objective which he is likely to be able to achieve – and that macroeconomic policymakers would be able to achieve a nominal income target, but would not be able to achieve an excessively optimistic employment target. (This was, again, a very Barro–Gordon-like answer.) But I had, by that time, read the papers by Phillips referred to above. So I stood up again and – rather bravely – said that, although this might be true, I thought that if his system was set up as a set of differential equations it would probably be unstable.

This question was to set in train a large research programme in the Department of Applied Economics in Cambridge. I had never met Meade before this lecture, but within a week he had asked me to work with him, and he then gradually gathered a large team to work with us, which included Andy Blake, Nicos Christodoulakis, Martin Weale and Peter Westaway, and also brought the control engineer Jan Maciejowski into the group. The resulting activity led to four substantial books (Meade, 1982; Meade et al., 1983a; Meade, 1986a; and Meade et al., 1989) and also to a number of tracts and articles in both technical and popular journals. Two main strands emerged in this work; we can describe these as being about inflation targeting and about supply side reforms.

Inflation targeting

The second and fourth of the books just described set out Meade's proposed policy regime, in which there would be a target for nominal GDP, to be controlled primarily by means of changes in taxes. In Meade et al. (1983a) it was shown, using an estimated econometric model of the economy, that fine-tuned feedback rules for taxes really could be found which would keep nominal income close to a target path. The work used the multivariable control methods which Phillips (1957) had predicted would

become available, which were supplied to the group by Jan Maciejowski.

This work culminated in Meade et al. (1989), called *Macroeconomic Policy: Inflation, Wealth and the Exchange Rate*. As a central part of the work for this book, Martin Weale oversaw the construction of an original empirical macroeconomic model, which developed the model being used at the National Institute of Economics Research in London at the time. It contained a number of rational-expectations features, which, at that time, were highly innovative. In particular, the model investment was driven by Tobin's q (that is, by the value of the stock market), which, following earlier work by Blanchard, was forward-looking, and which jumped in response to the expected future level of the interest rate (in exactly the same way as the exchange rate jumps in the Dornbusch model). And the model contained a forward-looking consumption function, with consumption partly depending on expected future income and thus on the expected future level of taxes. This is by now all rather familiar, but was ground-breaking at the time, although some of the underlying ideas had already been explained by Meade himself, nearly 20 years earlier, in *The Growing Economy* (Meade, 1968).

Meade's policies were tried out on this model, using taxes as the policy instrument (and also the interest rate, for reasons explained below). This required the application of feedback control to a forward-looking model. That was necessary, given the rational-expectations features in the model, which made consumption and investment at any point in time depend on the expected level of taxes and the interest rate in the future, as has been explained above. The new ideas necessary for this work were developed jointly by the group in Cambridge, by a group in London led by David Currie and Paul Levine, and by Marcus Miller and Willem Buiter in Warwick and Bristol. The central idea driving that work on control methods was that rule-bound policies are necessary to guide an economy, if the world is forward-looking, since what economic agents do now depends on what they expect policy to do in the future. Such ideas were largely put on one side in the early to mid-1990s, when inflation-target regimes were first analysed theoretically, using simple backward-looking models (see, for example, Bean, 1998.) But many of them have re-emerged in more recent technical work on targeting inflation in forward-looking, dynamic economies. For example, the idea of ‘stabilization bias’, which was understood very clearly by this group of people in England in the mid-1980s, was rediscovered and made popular by Michael Woodford nearly 15 years later, in the late 1990s.

Meade's young colleagues came to experience his skill at running a group of researchers – which I have begun to think he partly inherited from his experience in the Cambridge Circus so many years previously. As he passed 80 years of age, Meade presided over a weekly programme of meetings, at which his research group discussed the rational-expectations developments which I have described above. The day after each meeting, Meade would sit at home, in his village outside Cambridge, and write down an algebraic formulation of what we had all discussed. He would then walk to his local post office and send us a letter containing a photocopy of these handwritten notes. We would all then analyse his algebra and diagrams, in preparation for the next week's meeting.

It is fair to say that the policy proposals, which Meade's group developed, have not withstood the test of time. There are two explanations for this.

First, we now target the rate of inflation, not nominal incomes; Meade's nominal-income target regime was, effectively, only a precursor to the inflation-target regime which is now in place in the UK. Meade proposed a nominal income target, in part, because it was inherently more flexible than a *rigid* inflation target. It did not require that the inflation rate be exactly pinned down to an exactly pre-announced rate, but instead allowed, as explained above, for a (temporary) increase in inflation to be met by a

(temporary) reduction in output, so as to ensure – on balance – that there would be no change in the rate of growth of nominal incomes. We now know that a *flexible* inflation-target regime is better than such a nominal-income target regime. But it took some years of research for us to understand just why this is so, work which is described, for example, in Hall and Mankiw (1993), Leiderman and Svensson (1995), and Woodford, (2003). We have realized that the chief disadvantage of a nominal income target is that it does not ‘let bygones be bygones’: it requires that any overshoot which has occurred in the *level* of prices be clawed back, by means of a recession and lower subsequent inflation. But – at the same time – we have also realized that significant institutional development is required if one is to move from a purely rule-based system, like a nominal-income-target regime, to something like a rule-based but flexible inflation-target regime. To do this requires that the macroeconomic policymaking authorities be shielded from political influences which might force them to use their flexibility in an over-inflationary manner.

Second, we now use changes in interest rates, not changes in tax rates, in order to control inflation. We do this for three well-known reasons. First, it is easier to shield monetary policy from political influence than it is to do this for fiscal policy. Second the interest rate can be changed more regularly than taxes can be changed, and more quickly in response to new information – although fiscal procedures are less inflexible in some countries (such as the UK and New Zealand) than in others (such as the United States). Third, in an open economy monetary policy can have effects beyond those which can be caused by changes in taxes, because it can cause changes in a country's exchange rate which can, in turn, cause movements in exports and imports. This allows a country to externalize some of the costs of controlling shocks. That is a good idea if shocks in the world happen at different times in different countries.

Thus, as to both target and instrument, it appears that the world has moved on from Meade's proposals. Nevertheless Meade's proposals had more general features, which *do* seem to have survived the test of time. Meade came to describe them as ‘New Keynesian’. They were ‘Keynesian’ since, unlike Friedman, Meade continued to see the need for interventionist macroeconomic policies. (On this see Gordon, 1990.) They were ‘new’ because Meade proposed a target for a nominal variable (nominal income) instead of having a target for real output. And, in addition, they were proposals for rule-bound policies. It is hard to remember how unusual, and how original, it was to combine these three features, in the early 1990s.

These three features of Meade's proposals seem to have had a significant influence on the development of macroeconomic policymaking in the UK in the early 1990s. It is also hard to remember just what a mess macroeconomic policymaking was in Britain at that time. Following the country's brief flirtation with monetarism, it had joined the ill-fated ‘exchange-rate mechanism’ of the European Monetary System, which, in retrospect, appears to have been a pretty stupid policy framework. Following the UK's ejection from that system in September 1992, the Bank of England needed to quickly design a new policy regime. There was very little good theoretical guidance on what to do – other than Meade's. I say this, in particular, because the proposals of John Taylor, that monetary policy could follow a ‘Taylor rule’, really emerged only two years later (Taylor, 1994). When the new regime was announced by the Bank, *within days*, it had Meade's three features – it was one in which interest rates would be actively used, in pursuit of a nominal variable (the inflation target), in a rule-bound (if flexible) manner. The outcome was one of the world's earliest inflation-targeting regimes (the British regime was second only to that established in New Zealand), a set-up which has developed into the world's best inflation-targeting system. I believe that *Macroeconomic Policy* (Meade et al., 1989) exerted some influence in

the construction of this valuable new regime in Britain.

Furthermore, there are aspects of Meade's proposals which may yield further benefits in the future. *Macroeconomic Policy* suggests that policy should not just pursue a nominal anchor (taken to be a nominal income target in that book but it could just as well be an inflation target). The book also suggests that policy should pursue a target for the allocation of GDP between consumption and investment, so as to avoid ‘selling off the family silver’ (a phrase much discussed at the time), that is, so as to ensure that the supply side of the economy grows sufficiently rapidly. To do this, the book suggests that there should be rule-bound procedures for *two* policy instruments (both monetary policy – that is, interest-rate policy – and fiscal policy) in the joint pursuit of *two* targets (both the nominal anchor and the consumption–investment split). This was a very Meade-like suggestion in two ways: it synthesized a number of different ideas being discussed at the time, and it was characteristically complex and difficult to investigate (making this aspect of *Macroeconomic Policy* quite hard to understand).

One might argue that, in most circumstances, interest-rate policy can adequately control inflation, in the short to medium run, leaving fiscal policy to be more gradually adjusted so as to being about any desired changes in the consumption–investment mix, in the longer term. This is, for example, how the current British macroeconomic policymaking framework operates. In such cases monetary policy and fiscal policy can be considered separately, and a complex analysis of two instruments in pursuit of two targets is positively unhelpful.

Nevertheless, practical experience – in the United States, Japan, Europe and Australia – has shown that there are circumstances in which fiscal policy may need to assist in the pursuit of the inflation target, particularly when there are large falls, or increases, in demand. (See Garnaut, 2005.) And recent theoretical work has shown that there may be more general advantages if fiscal and monetary policymakers can rely on each other to act in appropriate ways. (See Allsopp and Vines, 2005.) The problems which might arise if the monetary and fiscal authorities cannot do this, and act independently of each other, were examined in Meade's very last published journal article (Meade and Weale, 1995b). These problems have arisen very seriously in the Eurozone, where the European Central Bank and European governments do not cooperate, but not much, if at all, in the United Kingdom, for reasons examined theoretically in Kirsanova, Stehn and Vines (2005).

The reform of wage fixing

The second part of Meade's project considered measures to promote employment through the reform of wage fixing. These were described in Meade (1982; 1984a; 1986a; 1986b). Looking back, one can credit Meade with having helped to create a sea change in the 1980s in British discussion of how wages ought to be fixed. Gone entirely are the ideas of rigid, centralized policies to hold down wages and prices by centralized administration. In their stead are proposals for policies which reinforce market mechanisms and which have their major impact as employment-creating rather than price-controlling devices (Layard, 1986). Meade's own suggestions included proposals for arbitration, a wage inflation tax and profit sharing.

On profit sharing and related topics Meade had already written a number of papers, starting in 1972 with ‘The Theory of Labour Managed Firms and of Profit Sharing’; and his views on this subject also become influential in Britain. He was sympathetic to the ideas about workers’ remuneration espoused in

Weitzman's book *The Share Economy* (1984). But his criticisms of Weitzman were also important. Profit sharing *might* have beneficial effects for macroeconomic stability, through encouraging greater flexibility of workers' remuneration. But it *might also* do the opposite, if workers who concede profit sharing also come to exert an influence on the employment decisions of their firms, and use this influence to restrict employment opportunities and raise their own wages.

Meade went on working in this second area, long after the group of those working on demand management broke up after the publication of *Macroeconomic Policy* in the late 1980s. An important driving force in this work, and something which I have not discussed adequately in this article, was Meade's interest in the reform of the social security system. Such reform might make it possible to reconcile an efficient labour market – which might necessitate a pay-bargaining system that delivered low wages to some people – with a distribution of income which was equitable and just.

The year 1995 saw the production of Meade's last book *Full Employment Regained* (Meade, 1995a), in which he attempted a synthesis of his ideas on demand management and on supply side reforms, arguing that full employment was possible providing that the appropriate reforms were undertaken. This, as Atkinson and Weale (2000) say, brought his career full circle. That career began, and ended, with Meade being concerned about the waste of resources and misery generated by high levels of unemployment. The Institute of Fiscal Studies hosted a seminar at which the ideas in his book were discussed. This was his last public appearance. And it was a gathering of many of the people whom he had influenced throughout his long career.

6 Influence

There can be no doubt that the *Theory of International Economic Policy* had an enormous influence upon our discipline. Corden and Atkinson (1979) and Johnson (1978) pay eloquent tribute to this. What I have said above suggests that there are many other ways in which he has exerted considerable influence. However, it is true that Meade is not as visible as some others of his generation.

This is probably due to his difficult manner of writing. This meant that his books were not as widely read as they might have been. Immediately one must exempt from this blanket statement Meade's popular articles and semi-popular tracts, which were beautifully written, and which displayed Meade's classical training to great effect, at the same time as being very persuasive. However his form of exposition, when he was doing fundamental economic theory, was very different. His 'style of work and presentation consists in the development of a *general* mathematical model of a problem, followed by translation of analysis of the various possible cases into literary English illustrated at most by arithmetical examples or simple diagrams' (Johnson, 1978, p. 65; emphasis added). Johnson went on (p. 66) to complain about his 'taxonomic approach and dependence on rather inelegant personal mathematics'. This means, said Johnson, that 'students find it incredibly tedious to read his books and [find it] difficult to convince themselves that the effort is worthwhile in terms of the knowledge gained' (p. 65).

Corden and Atkinson made similar complaints, specifically about Meade's *Theory of International Economic Policy*, but by implication about his other work as well:

> … the … model of *The Balance of Payments* was very influential and … had a rapid

> impact on key writers and policy makers in the field.… By contrast, the influence of Trade and Welfare was more delayed, and to a great extent many of its original ideas were rediscovered later … Both books … are written in a taxonomic and rather heavy style, with no footnote references to the literature and a failure to highlight the author's original contributions. Although the books are immensely rewarding to serious students, their messages often reach a wider audience only through the intermediation of more succinct, if less original, writers. (Corden and Atkinson, 1979, p. 530)

Elsewhere Corden and Atkinson talk of Meade's ‘distinctive literary–arithmetical style’, which now seems somewhat old fashioned compared with modern concise, simple algebraic expositions. Johnson (1978, p. 74) sums up the complaints, talking about the ‘reader-repellent character of Meade's literary-arithmetical-cum-idiosyncratic-mathematical-appendix style of presentation’.

All this enables one to see why the spread of Meade's ideas had to rely, more than is usual, on his personal influence over his colleagues. It is easy to see how, in such circumstances, his influence could be underrated.

Yet one can easily see, too, why Meade deliberately chose to work in the way just described. It was his prodigious power to generalize – to see competing theories about any subject as part of a yet larger encompassing scheme of things – which caused him to create his vast architectural structures of taxonomy. These put off many readers. But, the structures having been created, a dedicated band of followers managed to climb up onto them, and then – when they came down again – to explain what they had seen to the rest of the profession. It was obviously easier to do this for those disciples and colleagues who had the good fortune to work directly with Meade, for they were able to discuss the insights of his work with him, as they worked through it. As will be clear, there were many such disciples and colleagues throughout Meade's long career. It is mainly through them, and thus mainly indirectly, that his influence spread so far.

Those who worked with Meade shared in his zest for life and in his acute sense of fun, some of which may be apparent to the reader of this account. They saw, too, his respect for careful argument, and his pleasure in a slow and measured conversation, through which such argument can be developed. But, especially, Meade conveyed to them his sense of the underlying moral purpose of our discipline. It is appropriate to end this assessment of Meade's work by discussing his views on that subject.

7 Underlying philosophy

I mentioned in the introduction that Meade took up the study of economics because he wanted to help make the world a better place for ordinary men and women. In this he stood in the great Cambridge tradition of secular moralists, who might in the early Victorian age have become priests, but under the later influences of Darwinism and religious doubt turned instead to social improvement. The first volume of Skidelsky's biography of Keynes (Skidelsky, 1983) links Keynes back to Marshall and Sidgwick in that enterprise. Meade, in turn, emphasized this shared objective as ‘the decisive factor in binding me so closely to [Keynes] … he had … a passionate desire to devise a better domestic and international society’ (Meade, 1983b, p 268).

What is the conception of this better society that Meade strove for? And what is the role of the

economist in helping to create it? The following few paragraphs, taken from a review article which he wrote in the late 1940s (Meade, 1948b, p. 34), summarize some of his key ideas with a deceptive simplicity.

Meade writes that one's overall purpose is that ‘of combining freedom, efficiency and equity in social affairs …’

> Two points should, however, be emphasised. First, this does not beg the question of planning. There may well be occasions … on which the State should rightly prepare general programmes for far-reaching structural changes in the use of the community's resources; and there may be sections of the economy (such as public investment) where the State should on all occasions plan ahead. But where planning takes place, it is still possible to use money and prices as a main, if not the main, instrument for getting the plan carried out.
>
> Secondly, there is no suggestion that on those occasions in which money and prices have been extensively used in the past the arrangements have been satisfactory. Far from it. In order that money and prices may fulfil their purpose three main conditions must be fulfilled. First, the total supply of monetary counters must be neither too great nor too small in relation to the total supply of goods and services to be purchased. Secondly, the total supply of monetary counters must be equitably distributed so that no one obtains more than a fair share of command over resources. Thirdly, no private person or body of persons must be allowed to remain in a sufficiently powerful position to rig the market for his own advantage.
>
> These conditions have not been fulfilled in the past. On the contrary, considerable state planning and much state intervention is required to ensure that these conditions are fulfilled. If, however, we wish to combine freedom, efficiency and equity in our economic life, we should proceed to make arrangements to see that these fundamental conditions are satisfied; and as they are more and more nearly fulfilled we should make a progressively greater use of the monetary and pricing systems….

These ‘fundamental conditions’ have indeed been more nearly satisfied, in OECD countries, in the several decades since Meade wrote these words. But there is still much work to do. We still need to design intelligent monetary and pricing systems to deal with pressing global microeconomic problems (such as the threat of global warming, or the miserable health of the world poorest people), and with pressing global macroeconomic problems (such as large international imbalances, and the risks of financial crises in emerging market economies). It remains Meade's challenge to economists that we should develop policymaking institutions, and pricing systems, to deal with these problems, in ways which combine all of freedom, efficiency and equity, as much as possible.

See Also

- absorption approach to the balance of payments
- elasticities approach to the balance of payments
- Heckscher–Ohlin trade theory
- inflation targeting
- International Monetary Fund
- World Trade Organization

Selected works

This article is a development of my entry about Meade in the first edition of *The New Palgrave: A Dictionary of Economics* (Vines, 1987). That piece made use of two full-scale assessments of his work by Harry Johnson (1978) and by Corden and Atkinson (1979), and also used the short account by Harcourt (1985). In writing this piece I have been helped by a number of full-scale assessments of Meade's work which have appeared since 1987, by Atkinson (1996), Atkinson and Weale (2000), and Howson (2000). I have also referred to shorter accounts in Solow (1987), Greenaway (1990), and Corden (1996a; 1996b). There is a bibliography of Meade's work attached to Johnson (1978), which was complete when it was assembled, except for ‘a number of ephemeral newspaper articles and reviews’. From 1978 until his death in 1995, Meade continued to publish a remarkable amount, and a full bibliography is attached to Howson (2000). Accordingly, in the list below I include only those pieces of Meade's work to which I have referred explicitly, plus a few other particularly important pieces to which I do not refer. The bibliography, of course, includes all the pieces by other authors to which I make reference.

1933. *The Rate of Interest in a Progressive State*. London: Macmillan.

1934. The amount of money and the banking system. *Economic Journal* 4, 98–107.

1936. *An Introduction to Economic Analysis and Policy*. London: Oxford University Press. 2nd edn, 1937. American edition, ed. C.J. Hitch with an Introduction by A.J. Hanson, New York: Oxford University Press, 1938.

1937. A simplified model of Mr Keynes’ system. *Review of Economic Studies* 4, 98–107.

1938. *Consumers’ Credits and Unemployment*. London: Oxford University Press.

1940. *The Economic Basis of a Durable Peace*. London: Oxford University Press.

1941. (With J.R.N. Stone.) The construction of Tables of National Income, Expenditure, Savings and Investment. *Economic Journal* 51, 216–33.

1944. (With J.R.N. Stone.) *National Income and Expenditure*. London: Oxford University Press.

1948a. *Planning and the Price Mechanism: The Liberal Socialist Solution*. London: Allen & Unwin; New York: Macmillan, 1949.

1948b. Planning without prices. *Economica NS* 15, 28–35.

1951–5. *The Theory of International Economic Policy*. Vol. 1: *The Balance of Payments* (1951); Vol. 2: *Trade and Welfare* (1955a); with mathematical supplements. London and New York: Oxford University Press.

1952. *A Geometry of International Trade*. New York: A. Kelley, 1969. London: Allen & Unwin.

1953a. *The Atlantic Community and the Dollar Gap*. London: Friends of the Atlantic Union.

1953b. *Problems of Economic Union*. The Charles R. Walgreen Foundation Lectures. Chicago: University of Chicago Press.

1955b. *The Theory of Customs Unions*. The De Vries Lectures, Vol. 1. Amsterdam: North-Holland.

1955c. The case for variable exchange rates. *Three Banks Review* 27(September), 3–27. Reprinted in *Readings in Money, National Income and Stabilization Policy*, ed. W. Smith and R. Teigen. Chicago: Richard D. Irwin, 1965.

1956a. *The Belgium–Luxembourg Economic Union, 1921–39: Lessons from an Early Experiment*. Essays in International Finance No. 25. Princeton: International Finance Section.

1956b. *Japan and the General Agreement on Tariffs and Trade*. The Joseph Fisher Lectures in Commerce, Adelaide University.Adelaide: Adelaide University Press.

1957. (With E.A. Russell.) Wage rates, the cost of living and the balance of payments. *Economic Record* 33(April), 23–8.

1958. *The Control of Inflation*. Inaugural Lecture, Cambridge University. London: Cambridge University Press.

1961a. (With others.) *The Economic and Social Structure of Mauritius*. Report to the Governor of Mauritius. London: Methuen.

1961b. *A Neo-Classical Theory of Economic Growth*. London: Allen & Unwin. 2nd edn, 1964.

1964. *Efficiency, Equality and the Ownership of Property*. London: AlIen & Unwin.

1965. *Principles of Political Economy: I. The Stationary Economy*. London: Allen & Unwin; Chicago: Aldine Press.

1965. (With F.H. Hahn.) The rate of profit in a growing economy. *Economic Journal* 75, 445–8.

1966. The outcome of the Pasinetti process: a note. *Economic Journal* 76, 161–5.

1968. *Principles of Political Economy: 1I. The Growing Economy*. London: Allen & Unwin; Chicago: University of Chicago Press.

1971a. *Principles of Political Economy: III. The Controlled Economy*. London: Allen & Unwin.

1971b. *Wages and Prices in a Mixed Economy*. Occasional Paper No. 35. London: Institute of Economic Affairs.

1972. The theory of labour managed firms and of profit sharing. *Economic Journal* 82, 402–28.

1973a. *The Theory of Economic Externalities*. Geneva: Institut Universitaire des Hautes Etudes Internationales.

1973b. Economic policy and the threat of doom. In *Resources and Population*, ed. B. Benjamin, P. Cox and J. Peel. London: Academic Press.

1973c. The inheritance of inequalities: some biological, demographic, social and economic factors. *Proceedings of the British Academy* 59, Oxford: Oxford University Press.

1975a. *The Intelligent Radical's Guide to Economic Policy*. London: Allen & Unwin.

1975b. The Keynesian revolution. In *Essays on John Maynard Keynes*, ed. M. Keynes. London: Macmillan.

1976. *Principles of Political Economy: IV. The Just Economy*. London: Allen & Unwin.

1977a. Autobiography.In *Nobel Lectures: Economics 1969–80*, ed. A. Lindbeck. Singapore: World Scientific Publishing Co, 1992. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1977/meade-autobio.html, accessed 27 March 2007.

1977b. Banquet Speech on award of the Nobel Prize. In *Les Prix Nobel. The Nobel prizes 1977*, ed. W. Odelberg. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/

economics/laureates/1977/meade-speech.html, accessed 27 March 2007.

1978a. (With others.) *The Structure and Reform of Direct Taxation*. London: Allen & Unwin.

1978b. The meaning of internal balance. *Economic Journal* 88, 423–35.

1981. Comment on the papers by Professors Laidler and Tobin. *Economic Journal* 91, 49–55.

1982. *Stagflation*, Vol. I: *Wage Fixing*. London: Allen & Unwin.

1983a. (with D. Vines and J. Maciejowski.) *Stagflation*, Vol. II: *Demand Management*. London: Allen & Unwin.

1983b. Impressions of Maynard Keynes. In Worswick and Trevithic (1983).

1984b. A new Keynesian Bretton Woods. *Three Banks Review* (June), 8–25.

1984c. A new Keynesian approach to full employment. *Lloyds Bank Review* (150), 1–18.

1984a. *Wage Fixing Revisited*. Occasional Paper No. 72.London: Institute of Economic Affairs.

1986a. *Alternative Systems of Business Organization and Workers' Remuneration*. London: Allen & Unwin.

1986b. *Different Forms of Share Economy*. London: Public Policy Centre, London.

1988–90. *The Collected Papers of James Meade Volumes I–IV*, (Volume IV jointly edited with D. Moggridge). London: Unwin Hyman.

1989. (With M. Weale, A. Blake, N. Christodoulakis and D. Vines.) *Macroeconomic Policy: Inflation, Wealth and the Exchange Rate*. London: Unwin and Hyman.

1993. The relation of Mr Meade's relation to Kahn's multiplier. *Economic Journal* 103, 464–5.

1995a. *Full Employment Regained*. Occasional Paper No. 61, Department of Applied Economics, Cambridge University.

1995b. (With M. Weale.) Monetary union and the assignment problem. *Scandinavian Journal of Economics* 97, 201–22.

Bibliography

Alexander, S. 1952. The effects of a devaluation on a trade balance. *IMF Staff Papers* 2, 263–78.

Allsopp, C. and Vines, D. 2005. The macroeconomic role of fiscal policy. *Oxford Review of Economic Policy* 21, 485–508.

Atkinson, A. 1996. James Meade's vision: full employment and social justice. *National Institute Economic Review* (July), 90–7.

Atkinson, A. and Weale, M. 2000. James Meade: a memoir. In *1999 Lectures and Memoirs*. London: published for the British Academy by Oxford University Press.

Barro, R. and Gordon, D. 1983. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91, 589–610.

Bean, C. 1998. The new UK monetary arrangements: a view from the literature. *Economic Journal* 108, 1795–809.

Bhagwati, J. and Ramaswami, V.K. 1963. Domestic distortions, tariffs, and the theory of optimum subsidy. *Journal of Political Economy* 71, 44–50.

Brigden, J., Copeland, D., Dyason, E., Giblin, L. and Wickens, C. 1929. *The Australian Tariff: An Economic Enquiry*. Melbourne: Melbourne University Press.

Corden, W.M. 1984. Booming sector and Dutch Disease economics: Survey and consolidation. *Oxford Economic Papers* 36, 359–80.

Corden, W.M. 1996a. Special profile: James Meade, 1907–1995. *Review of International Economics* 4, 382–6.

Corden, W.M. 1996b. James Meade 1907–1995. *Economic Record* 72, 172–4.

Corden, W.M. and Atkinson, A. 1979. Meade James E. In *International Enyclopaedia of the Social Sciences Bibliographical Supplement*, vol. 18, ed. D.L. Sills. New York: Free Press; London: Macmillan.

Durbin, E. 1985. *New Jerusalems*. London: Routledge & Kegan Paul.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 57, 297–308.

Fleming, J.M. 1951. On making the best of balance of payments restrictions on imports. *Economic*

Journal 61, 48–71.

Fleming, J.M. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9, 369–79.

Friedman, M. 1953. The effects of a full employment policy on economic stability: a formal analysis. In *Essays in Positive Economics*. Chicago: Chicago University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Garnaut, R. 2005. Is macroeconomics dead? Monetary and fiscal policy in historical context. *Oxford Review of Economic Policy* 21, 524–31.

Gordon, R. 1990. What is New Keynesian Economics? *Journal of Economic Literature* 28, 1115–71.

Greenaway, D. 1990. The intelligent radical on economic policy: an essay on the work of James Meade. *Scottish Journal of Political Economy* 37, 288–98.

Gregory, R. 1976. Some implications of the growth of the mineral sector. *Australian Journal of Agricultural Economics* 20(2), 71–91.

Harberger, A.C. 1950. Currency depreciation, income and the balance of trade. *Journal of Political Economy* 58, 47–60.

Hall, R. and Mankiw, G. 1993. Nominal income targeting, NBER Working Paper no 4439. Published in *Monetary Policy*, ed. G. Mankiw. Chicago: University of Chicago Press, 1994.

Harcourt, G.C. 1982. The Social Science Imperialists: Selected Essays, ed. P. Kerr. London: Routledge & Kegan Paul.

Harcourt, G.C. 1985. Meade, James (1907–). In *The Social Science Encyclopaedia*, ed. A. Kuper and J. Kuper. London: Routledge & Kegan Paul.

Harrod, R.F. 1933. *International Economics*. London: Nisbet.

Hicks, J. 1937. Mr Keynes and the classics: a suggested simplification. *Econometrica* 5, 147–59.

Howson, S. 2000. James Meade. *Economic Journal* 110, 122–45.

Johnson, H.G. 1951. The taxonomic approach to economic policy. *Economic Journal* 61, 812–32.

Johnson, H.G. 1965. Optimal trade intervention in the presence of domestic distortions. In *Trade Growth and the Balance of Payments*, ed. R. Baldwin. Amsterdam: North-Holland.

Johnson, H.G. 1978. James Meade's contribution to economics. *Scandinavian Journal of Economics* 80, 64–85.

Kahn, R. 1931. The relation of home investment to unemployment. *Economic Journal* 41, 173–98.

Kahn, R. 1984. *The Making of Keynes' General Theory*. Cambridge: Cambridge University Press.

Keynes, J.M. 1930. *A Treatise on Money*. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1939. Relative movements of real wages and output. *Economic Journal* 49, 34–51.

Keynes, J.M. 1940. *How to Pay for the War*. Repr. in Keynes (1971–88, vol. 9). London: Macmillan.

Keynes, J.M. 1971–88. *The Collected Writings of John Maynard Keynes*. London: Macmillan.

Kirsanova, T., Stehn, J. and Vines, D. 2005. The interactions between fiscal policy and monetary policy. *Oxford Review of Economic Policy* 21, 532–64.

Kydland, F.E. and Prescott, E.C. 1978. Rules rather than discretion. *Journal of Political Economy* 84, 473–91.

Laursen, S. and Metzler, L. 1950. Flexible exchange rates and the theory of employment. *Review of Economics and Statistics* 32, 281–99.

Layard, R. 1986. *How to Beat Unemployment*. Oxford: Oxford University Press.

Layard, R., Nickell, S. and Jackman, R. 1991. *Unemployment: Macroeconomic Performance and the Labour Market*. Oxford: Oxford University Press.

Leiderman, L. and Svensson, L., eds. 1995. *Inflation Targets*. London: CEPR.

Minister of Reconstruction. 1944. *Employment Policy*. Cmd. 6527. London: HMSO.

Mundell, R.A. 1962. The appropriate use of fiscal and monetary policy for internal and external stability. *IMF Staff Papers* 9, 70–7.

Phillips, A.W. 1950. Mechanical models in economic dynamics. *Economica* 17, 283–305.

Phillips, A.W. 1954. Stabilization policy in a closed economy. *Economic Journal* 64, 290–323.

Phillips, A.W. 1957. Stabilization policy and the time form of lagged responses. *Economic Journal* 67, 265–77.

Phillips, C. 1920. *Bank Credit*. New York: Macmillan.

Robinson, J. 1937. The foreign exchanges. In *Essays in the Theory of Employment*. London: Macmillan. 2nd edn, 1947.

Rowthorn, R. 1977. Conflict, inflation and money. *Cambridge Journal of Economics* 1, 215–39.

Samuelson, P. 1948. *Economics*. New York: McGraw Hill.

Shackle, G.L.S. 1967. *The Years of High Theory*. Cambridge: Cambridge University Press.

Skidelsky, R. 1983. *John Maynard Keynes: Hopes Betrayed 1883–1920*. London: Macmillan.

Skidelsky, R. 2000. *John Maynard Keynes: Fighting for Britain 1937–1946*. London: Macmillan.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Solow, R.M. 1987. James Meade at eighty. *Economic Journal* 97, 986–8.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Swan, T. 1963. Longer run problems of the balance of payments. In *The Australian Economy*, ed. H.W. Arndt and W.M. Corden. Melbourne: Cheshire. Repr. in *Readings in International Economics*, ed. R. Caves and H. Johnson. Homewood, IL: Irwin, 1968.

Taylor, J. 1994. *Macroeconomic Policy in a World Economy: From Econometric Design to Practical Operation*. New York: Norton.

Van Dormael, A. 1978. *Bretton Woods: Birth of a Monetary System*. London: Macmillan.

Viner, J. 1950. *The Customs Union Issue*. New York: Carnegie Institute for International Peace.

Vines, D. 1987. Meade James Edward. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J.

Eatwell, M. Milgate and P. Newman. London: Macmillan.

Vines, D. 1994. Unfinished business: Australian protectionism, Australian trade liberalisation and APEC. Shann Memorial Lecture, University of Western Australia. In *Australian Macroeconomic Policy Debates: Contributions from the Shann Memorial Lectures, 1991–2000*, ed. P.L. Crompton. Perth: University of Western Australia Press, 2005.

Vines, D. 1996. The Phillips machine. In *A.W. Phillips: Collected Writings in Contemporary Perspective*, ed. R. Leeson. Cambridge: Cambridge University Press.

Vines, D. 2003. John Maynard Keynes, 1937–1946: The creation of international macroeconomics. *Economic Journal* 113, F338–361.

Weitzman, M. 1984. *The Share Economy*. Cambridge, MA: Harvard University Press.

Williamson, J. 1983a. Keynes and the international economic order. In Worswick and Trevithick (1983).

Williamson, J. 1983b. *The Exchange Rate System*. Washington, DC: Institute for International Economics.

Wilson, T. 1982. Planning for the war and for the peace. In *Keynes as a Policy Adviser*, ed. A.P. Thirlwall. London: Macmillan.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

Worswick, D. and Trevithick, D., eds. 1983. *Keynes and the Modern World*. Cambridge: Cambridge University Press.

Young, W. 1987. *Interpreting Mr Keynes: The IS/LM Enigma*. Cambridge: Polity Press.

How to cite this article

Vines, David. "Meade, James Edward (1907–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000301> doi:10.1057/9780230226203.1070

The New Palgrave Dictionary of Economics Online

meaningfulness and invariance

Louis Narens and R. Duncan Luce

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Given a qualitative scientific structure, a structure preserving mapping into a numerical, vectorial, or geometric structure is called a *representation* of it. Within such numerical, vectorial or geometric structures other concepts can always be defined. Some of these correspond to a qualitative property of the underlying system, and they are called ‘meaningful’ concepts. And others do not correspond to a qualitative property; and they are called ‘meaningless’. The article investigates precise meanings of ‘meaningfulness’ and ‘meaninglessness’ and their relation to several notions of invariance, some of which are widely used in science.

Keywords

dimensional analysis; interpersonal comparison of utilities; invariance; meaningfulness; measurement; representations; scientific definability

Article

Few disavow the principle that scientific propositions should be meaningful in the sense of asserting something that is verifiable or falsifiable about the qualitative or empirical situation under discussion. What makes this principle tricky to apply in practice is that much of what is said is formulated not as simple assertions about qualitative or empirical events – such as a certain object sinks when placed in water – but as laws formulated in rather abstract, often mathematical, terms. It is not always apparent exactly what class of qualitative observations corresponds to such (often numerical) laws. Theories of meaningfulness are methods for investigating such matters, and invariance concepts are their primary tools.

The problem of meaningfulness, which has been around since the inception of mathematical science in ancient times, has proved to be difficult and subtle; even today it has not been fully resolved. This article surveys some of the current ideas about it, and illustrates, through examples, some of its uses. The presentation requires some elementary technical concepts of measurement theory (such as

representation, scale type, and so on), which are explained in measurement, theory of.

1 Concepts of meaningfulness

1.1 Some notation and definitions

The operation of functional composition is denoted *. The Cartesian product of $T_1, \ldots, T_n$ is denoted $\Pi_i^n T_i$.

A *scale* $\mathcal{S}$ is a set of functions from a qualitative domain, a set *X* endowed with one or more relations, into the real numbers. Elements of $\mathcal{S}$ are called *representations*. An example is the usual physical scale to measure length. Two of its representations are the *foot representation* and the *centimeter representation*. $\mathcal{S}$ is said to be

- a *ratio scale* if and only if for each ϕ in $\mathcal{S}$,

$$\mathcal{S} = \{r\varphi | r > 0\},$$

- an *interval scale* if and only if for each ϕ in $\mathcal{S}$,

$$\mathcal{S} = \{r\varphi + s | r > 0,\ s \text{ a real}\},$$

- an *ordinal scale* if and only if for each ϕ in $\mathcal{S}$, the range of ϕ is a (possibly infinite) interval of reals and

$$\mathcal{S} = \{f * \varphi | f \text{ is a strictly monotonic function from the range of } \varphi \text{ onto itself}\}.$$

1.2 Intuitive formulation of meaningfulness and some examples

The following example, taken from Suppes and Zinnes (1963), nicely illustrates part of the problem in a very elementary way. Which of the following four sentences are meaningful?

1. (i) Stendhal weighed 150 on 2 September 1839.
2. (ii) The ratio of Stendhal's weight to Jane Austen's on 3 July 1814 was 1.42.

3. (iii) The ratio of the maximum temperature today to the maximum temperature yesterday is 1.10.
4. (iv) The ratio of the difference between today's and yesterday's maximum temperature to the difference between today's and tomorrow's maximum temperature will be 0.95.

Suppose that weight is measured in terms of the ratio scale $\mathcal{W}$ (which includes among its representations the pound and kilogram representations and all those obtained by just a change of unit), and that temperature is measured by the interval scale $\mathcal{T}$, which for this example includes the Fahrenheit and Celsius representations. (The Kelvin representation for temperature, which assumes an absolute zero temperature, is not in $\mathcal{T}$.) Then Statement (*ii*) is meaningful, because with respect to each representation in $\mathcal{W}$ it says the same thing, that is, its truth value is the same no matter which representation in $\mathcal{W}$ is used to measure weight. That is not true for Statement (*i*), because (*i*) is true for exactly one representation in $\mathcal{W}$ and false for all of the rest. Thus we say that (*i*) is ‘meaningless’. Similarly, (*iv*) is meaningful with respect to $\mathcal{T}$ but (*iii*) is not.

The somewhat intuitive concept of meaningfulness suggested by these examples is usually stated as follows. Suppose a qualitative or empirical attribute is measured by a representation from a scale of representations $\mathcal{S}$. Then a numerical statement involving values of the representation is said to be *quantitatively meaningful* if and only if its truth (or falsity) is constant no matter which representation in $\mathcal{S}$ is used to assign numbers to the attribute. There are obvious formal difficulties with this definition, for example the concept of ‘numerical statement’ is not a precise one. More seriously, it is unclear under what conditions this is the ‘right’ definition of meaningfulness, for it does not always lead to correct results in some well-understood and non-controversial situations. (See the discussion involving situations where the measurement scale consists of a single representation for an example.) Nevertheless, it is the concept most frequently employed in the literature, and invoking it often provides insight into the correct way of handling a quantitative situation – as the following still elementary but somewhat less obvious example shows.

Consider a situation where *M* persons rate *N* objects (for example, *M* judges judging *N* contestants in a sporting event). For simplicity, assume that person i rates objects according to the ratio scale of representations $\mathcal{R}_i$. The problem is to find an ordering on the N objects that aggregates the judgements of the judges in a reasonable way. It can be shown that their judgements cannot be coordinated in such a way that, for all R_i in $\mathcal{R}_i$ and R_j in $\mathcal{R}_j$ that for some object *a*, the assertion $R_i(a) = R_j(a)$ is justified philosophically. The difficulties underlying such a coordination are essentially those that arise in attempting to compare individual utility functions. The latter problem – the ‘interpersonal comparison of utilities’ – has been much discussed in the literature, as for example in Narens and Luce (1983). It is generally agreed that there are great, if not insurmountable, difficulties in carrying out such comparisons. Any rule that does not involve coordination among the raters can be formulated as follows. First, let *F* be a function that assigns to an object the value $F(r_1, \ldots, r_M)$ whenever person *i* assigns the number r_i to the object. Second, assume that object *a* is ranked just as high as *b* if and only if the value assigned by *F* to *a* is at least as great as that assigned by *F* to *b*. In practice *F* is often taken to be the arithmetic mean of the ratings $r_1, \ldots, r_M$ (for example, Pickering, Harrison and Cohen, 1973). Observe, however, that arithmetic means for this kind of rating situation, in general, produce a non-quantatively meaningful ranking of objects, as illustrated by the following special case. Suppose *M*=2 and, for $i = 1, 2$, R_i is person's *i* representation that is being used for generating ratings, and

$$R_1(a) = 2, \quad R_1(b) = 3, \quad R_2(a) = 3, \quad \text{and } R_2(b) = 1.$$

Then the arithmetical mean of the ratings for *a*, 2.5, is greater than that for *b*, 2, and thus *a* is ranked above *b*. However, meaningfulness requires the same order if any other representations of persons 1 and 2 rating scales are used, for example, $10R_1$ and $2R_2$. But for this choice of representations, the arithmetic mean of *a*, 13, is less than that of *b*, 16, and thus *b* is ranked higher than *a*.
It is easy to check that the geometrical mean of rankings for an object,

$$F(r_1, \ldots, r_M) = [r_1 \cdots r_M]^{\frac{1}{M}},$$

gives rise to a quantitatively meaningful, non-coordinated rule for ranking objects. It can be shown under plausible conditions that all other meaningful, non-coordinated rules give rise to the same ranking as that given by the geometric mean (Aczél and Roberts, 1989).
Many other applications of quantitative meaningfulness have been given by various researchers. In particular, Roberts (1985) provides a wide range of social science examples. In some contexts, quantitative meaningfulness presents certain technical difficulties that require some modification in its definition (see, for example, Roberts and Franke, 1976; Falmagne and Narens, 1983).

1.3 Meaningfulness and statistics

Another area of importance to social scientists in which invariance notions are thought to be relevant is applying statistics to numerical data. The role of measurement considerations in statistics and of invariance under admissible scale transformations was first emphasized by Stevens (1946; 1951); this view quickly became popularized in numerous textbooks, and it produced extensive debates in the literature. Continued disagreement exists, mainly created by confusion arising from the following two simple facts:

- Measurement scales are characterized by groups of admissible transformations of the real numbers.
- Statistical distributions exhibit certain invariances under appropriate transformation groups, often the same groups (especially the affine transformations), as those that arise from measurement considerations.

Because of these facts, some scientists have concluded that the suitability of a statistical test is determined, in part, by whether or not the measurement and distribution groups are the same. Thus, it is

said that one may be able to apply a test, such as a *t*-test, that rests on the Gaussian distribution to ratio or interval scale data, but surely not to ordinal data, because the Gaussian distribution is invariant under the group of positive affine transformations, $x \to rx + s$, *r*, *s* real, $r > 0$ – which arises in both the ratio and the interval case but not in the ordinal one. Neither half of the assertion is correct. First, a significance test should be applied only when its distributional assumptions are met, and they may very well hold for some particular representation of ordinal data. And second, a specific distributional assumption may well not be met by data arising from a particular scale of measurement. For example, reaction times, being times, are measured on a physical ratio scale, but they are rarely well approximated by a Gaussian distribution.

What is true, however, is that any proposition (hypothesis) that one plans to put to statistical test or to use in estimation had better, itself, be quantitatively meaningful with respect to the scale used for the measurements. In general, it is not quantitatively meaningful to assert that two means are equal when the quantities are measured by an ordinal scale, because equality of means is not invariant under strictly increasing transformations. Thus, no matter what distribution holds and no matter what test is performed, the result may not be quantitatively meaningful, because the hypothesis is not. In particular, if an hypothesis is about the measurement structure itself, for example that the representation is additive over a concatenation operation, then it is essential that the following propositions (*a*) and (*b*) hold, where a *symmetry* of a structure is by definition an isomorphism of the structure onto itself: (*a*) the hypothesis be invariant under the symmetries of the structure and therefore invariant under the scale used to measure the structure. (Because it is assumed that scales of measurement are structure preserving functions from a qualitative structure onto a quantitative one, (*a*) immediately follows). And (*b*) the hypotheses of the statistical test be met without going outside the transformations of the measurement representation. See Luce et al. (1990) for a more detailed discussion of this issue.

2 Concepts of invariance

Measurement laws are quantitative laws based primarily on interrelationships of scales of measurement. They have in common with quantitative meaningfulness that they are derived through considerations of admissible transformations of the measurements of relevant variables. In the view of Falmagne and Narens (1983, p. 298) they arise in an empirical situation ‘that is governed by an empirical law of which we know little of its mathematical form and a little of its invariance properties, but a lot about the structure of the admissible transformations of its variables, and use this information to greatly delimit the possible equations that express the law’. They are generalizations of the kind of laws that have a long-standing tradition in physics, where they are known as laws derived according principles of ‘dimensional analysis’. These principles involve the assertion that laws of nature are in a deep sense invariant under changes of unit, which correspond to invariance under symmetries. Thus, knowledge of the scale type of the relevant variables – a strong presupposition – greatly limits the forms of laws.

2.1 Measurement laws. simplest case

These principles were introduced into the behavioural sciences by Luce (1959), which was concerned with special cases of ‘possible psychophysical laws’. He generalized dimensional analysis, which only

assumed ratio scale transformations of the several variables, to the more general situation of the measurement scale types described by S.S. Stevens. Luce (1964) extended the 1959 formulation to include a few important cases of a single function of many variables.

Luce (1959) considered the case where the independent variable *x* and the dependent variable *y* were related by a law, $y = f(x)$, where *f* was some continuous function. He assumed that this law was invariant under admissible transformations of measurements, that is, for each admissible transformation ϕ of the independent variable, there was an admissible transformation ψ of the dependent variable such that for all *x* and *y*,

$$y = f(x) \text{ iff } \psi(y) = f(\phi(x)). \tag{1}$$

The following is an example of a use of Luce's theory. Suppose *x* is an objective variable measured by a ratio scale, for example, a physical variable such as the intensity of light or the weight of gold, and *y* is the subjective evaluation of *x*, for example, the subjective brightness of light, the subjective value of gold, and *f* is the law linking *x* and *y*. Suppose *x* and *y* are both measured on ratio scales and *f* is continuous. Suppose further that *f* satisfies eq. (1). Under these conditions, Luce shows that there are real numbers *r* and *a*, *a* depending on ψ, such that

$$f(x) = ax^r. \tag{2}$$

His method of proof was to show that eq. (1) implied that *f* satisfied the functional equation $h(s)f(t) = f(st)$ for some continuous function *h* and all positive *s* and *t*, and that this functional equation had eq. (2) as its only solution.

For most applications, such as the above brightness and subjective value examples, the scale for the independent variable is known and continuity is a reasonable idealized approximation. Sometimes theory will specify the measurement scale for the dependent variable. However, often the scale for the dependent variable is unknown, and in many cases, unobservable, as, for example, when it is subjective. In such situations, the measurement scale for the dependent variable has to be hypothesized or derived from theory. It can be hypothesized to be one of several theoretically reasonable types of measurement scales, and then methods similar to the one used to derive Equation (2) can be used to arrive at a measurement law for each type of hypothesized scale. The set of resultant measurement laws provides a clear set of quantitative hypotheses for empirical testing. Quite often such hypotheses turn out to be a good place to begin a scientific investigation.

2.2 Measurement laws. more complex cases

In a number of ways, Falmagne and Narens (1983) greatly generalized Luce's 1959 approach for deriving laws from measurement considerations. In particular:

- Instead of one independent variable and one dependent variable, they assumed *n* independent variables and one dependent variable. (They formulated matters for two independent variables to simplify notation, but their approach easily extends to *n* independent variables.)
- They allowed for a general relationship *R* among the admissible transformations of the independent variables to hold; that is, for the sets T_i of admissible transformations of the independent variables $x_1, \ldots, x_n$, *R* can be any nonempty subset of $\prod_i^n T_i$.
- They allowed for more general kinds of laws by allowing for a family $\mathcal{F}$ of functions to relate the dependent variable with *n* independent variables. They interpret $\mathcal{F}$ as follows. Initially, representations $\phi_1, \ldots, \phi_n$ are used to measure the *n* independent variables, $x_1, \ldots, x_n$. These measurements determine a function $f(\phi_1(x_1), \ldots, \phi_n(x_n))$ that is the value of the dependent variable measured on an unknown scale when $x_1, \ldots, x_n$ are measured by $\phi_1, \ldots, \phi_n$. There are other equally valid ways of measuring *each* independent variable x_i. These are obtained by transforming φ_i by the elements of T_i. However, valid measurements for the *set* of independent variables may be additionally constrained by the empirical law relating the dependent variable to the independent variables. The additional constraint is captured by the relation *R*. Thus each other valid measurement of the independent variables is given by $\tau_1 * \phi_1, \ldots, \tau_n * \phi_n$ for some $\tau_1, \ldots, \tau_n$ such that $R(\tau_1, \ldots, \tau_n)$. The law giving the numerical value of the dependent variable, when the set of independent variables $x_1, \ldots, x_n$ are measured respectively by $\tau_1 * \phi_1, \ldots, \tau_n * \phi_n$, is given by

$$f_{\tau_1, \ldots, \tau_n}(\tau_1 * \phi_1(x_1), \ldots, \tau_n * \phi_n(x_n)).$$

In this way, it is the family of functions,

$$\mathcal{F} = \{f_{\tau_1, \ldots, \tau_n}(\tau_1 * \phi_1(x_1), \ldots, \tau_n * \phi_n(x_n)) | R(\tau_1, \ldots, \tau_n)\},$$

that expresses the empirical law relating the dependent variable to the independent variables $x_1, \ldots, x_n$. Only in very restrictive cases will $\mathcal{F}$ consist of a single function.

2.3 Order meaningfulness

In place of assuming the scale type of the dependent variable, they assume ‘order meaningfulness’, that is, they assume the following. Using the just presented notation, suppose $\mathfrak{F}$ is a family of functions that is a law relating the dependent variable with *n* independent variables and $f_{\sigma_1,\ldots,\sigma_n}$ and $f_{\tau_1,\ldots,\tau_n}$ are in $\mathfrak{F}$. Then for all $x_1, \ldots, x_n$ and $u_1, \ldots, u_n$,
$f_{\sigma_1,\ldots,\sigma_n}(\sigma_1 * \phi_1(x_1), \ldots, \sigma_n * \phi_n(x_n)) \leq f_{\sigma_1,\ldots,\sigma_n}(\sigma_1 * \phi_1(u_1), \ldots, \sigma_n * \phi_n(u_n))$ if and only if $f_{\tau_1,\ldots,\tau_n}(\tau_1 * \phi_1(x_1), \ldots, \tau_n * \phi_n(x_n)) \leq f_{\tau_1,\ldots,\tau_n}(\tau_1 * \phi_1(u_1), \ldots, \tau_n * \phi_n(u_n))$.
By considering families of functions rather than a single function for laws, Falmagne and Narens generalized the notion of ‘dimensional constants’ that appear in many laws. Their generalization allows for the formulation of behavioural laws (Falmagne and Narens, 1983; Falmagne, 1985) and physical laws (Falmagne, 2004) that cannot be obtained by considering only a single function. Of course, Falmagne and Narens’ theory also allows for the case of a single function, by allowing the family of functions to degenerate to a set consisting of a single function.
In many situations order meaningfulness is a testable condition, making it a preferable assumption to assuming a scale type for a dependent variable unless, of course, one already has a well-developed theory for the dependent variable. In the Falmagne–Narens theory, the scale type of the dependent variable is not needed to obtain the law linking the independent and dependent variables.
For the case where the family $\mathfrak{F}$ consists of a single function *f* of *n*-independent variables, Aczél, Roberts and Rosenbaum (1986) provided more general results. Through an insightful mathematical argument, they were able to characterize measurement laws using only measurability assumptions from real analysis about *f* instead of monotonicity or continuity assumptions. Aczél and Roberts (1989) use the general approach of Aczél, Roberts and Rosenbaum (1986) to derive measurement laws of economic interest.

3 Relation between meaningfulness and invariance

Quantitative meaningfulness lacks a serious account as to why it is a good concept of meaningfulness; that is, it lacks a sound theory as to why it should yield correct results. Formulating a serious account for it is difficult. One tack (Krantz et al., 1971; Luce, 1978; Narens, 1981) is to observe that, if meaningfulness expresses valid qualitative relationships, then it must correspond to something purely qualitative, and therefore it should have a purely qualitative description. A long tradition in mathematics for formulating qualitative relationships that belong naturally to some structure or concept goes back to at least 19th-century geometry and was the centrepiece of the famous Erlanger Programme for geometry of Felix Klein. It was based on the idea that associated with each geometry was a set of transformations $\mathcal{T}$, and the relations and concepts belonging to the geometry were exactly those that were left invariant by all the transformations in $\mathcal{T}$. There are strong connections between (*a*) geometric techniques of establishing coordinate systems and measurement techniques for establishing scales, and (*b*) the Erlanger Programme's concept of ‘geometric’ and the measurement-theoretic concept ‘meaningfulness’. To examine these connections, some definitions and conventions are needed.

3.1 Convention

Throughout the remainder of this article, it is assumed that $\mathcal{X}$ is a qualitative structure, which consists of a qualitative set *X* as its domain and relations based on *X* (called the *primitives of* $\mathcal{X}$); $\mathcal{N}$ is a numerically based structure, that is, $\mathcal{N}$ is a structure that has a subset of the real numbers as its domain; and $\mathcal{S}$ is the measurement scale consisting of all isomorphisms from $\mathcal{X}$ onto $\mathcal{N}$. (See measurement, theory of for a more detailed description of this kind of measurement scale.)

3.2 Qualitative meaningfulness

An isomorphism of $\mathcal{X}$ onto itself is called a *symmetry* (or *automorphism*) of $\mathcal{X}$. It easily follows that if α is a symmetry of $\mathcal{X}$ and ϕ and ψ are elements of $\mathcal{S}$, then

- $\phi * \alpha$ is in $\mathcal{S}$,
- $\phi^{-1} * \psi$ is a symmetry of χ,
- $\theta = \phi * \psi^{-1}$ is an admissible transformation of $\mathcal{S}$, that is, $\phi * \eta$ is in $\mathcal{S}$ for each η in $\mathcal{S}$, and all admissible transformations can be obtained in the just mentioned manner by appropriate selections of ϕ and ψ.

An *n*-ary relation *R* on *X* is said to be *qualitatively meaningful* if and only if it is invariant under the symmetries of $\mathcal{X}$, that is, if and only if for each symmetry α of $\mathcal{X}$ and each $x_1, \ldots, x_n$ in *X*,

$$R(x_1, \ldots, x_n) \text{ iff } R(\alpha(x_1), \ldots, \alpha(x_n)).$$

3.3 Quantitative meaningfulness

Although a relation *T* being ‘quantitatively meaningful’ was previously defined, it is defined again here to make explicit the role the scale $\mathcal{S}$ plays in qualitative meaningfulness: an *n*-ary relation *T* on *N* is said to be *quantitatively* $\mathcal{S}$*-meaningful* if and only if for each admissible transformation τ of $\mathcal{S}$ and each $r_1, \ldots, r_n$ in *N*,

$$T(r_1, \ldots, r_n) \text{ iff } T(\tau(r_1), \ldots, \tau(r_n)).$$

$\mathcal{S}$ can be used to interpret *T* as a relation *U* on *X* as follows. The *n*-ary relation *U* on *X* is said to be the $\mathcal{S}$-*inpt* of *T* if and only if for all ϕ in $\mathcal{S}$ and all $r_1, \ldots, r_n$

$$T(r_1, \ldots, r_n) \text{ iff } U(\varphi^{-1}(r_1), \ldots, \varphi^{-1}(r_n)).$$

3.4 Basic result

The above definitions and relationships between symmetries and admissible transformations immediately yield the following theorem relating qualitatively and quantitatively meaningful relations:
Theorem: A relation T is quantitatively $\mathcal{S}$-meaningful if and only if its $\mathcal{S}$-inpt is qualitatively meaningful.
The above theorem shows that each quantitatively meaningful relation has, through measurement, a corresponding qualitatively meaningful relation. Luce (1978) used this idea to provide a qualitative theory for practice of dimensional analysis in physics: Luce produced a qualitative structure $\mathcal{X}$ for measuring physical attributes. He showed that, under measurement, the quantitatively meaningful relationships among the attributes were the ‘dimensionally invariant functions’ of dimensional analysis. It is a principle of dimensional analysis that physical laws are such dimensionally invariant functions. Thus, by the just mentioned theorem, it then follows from the principles of dimensional analysis that each physical law corresponds to a qualitatively meaningful relation of $\mathcal{X}$. (Measurement-theoretic foundations for dimensional analysis can be found in Krantz et al., 1971; Luce et al., 1990; Narens, 2002.)
Qualitative meaningfulness is just the Erlanger concept of ‘geometric’ applied to science. Mathematically, the two concepts are identical. The Erlanger Programme, as formulated by Klein (1872) and as used in mathematics, lacks a serious justification for assuming that the invariance of a relation under the symmetries of a geometry implies that the relation belongs to the geometry.

3.5 Scientific definability

Narens (2002; 2007) sought to find a justification for Klein's assumption. He thought that a reasonable concept of a relation *R* belonging to a structure $\mathcal{X}$ was that *R* should somehow be definable in terms of the primitives of $\mathcal{X}$. But the usual concepts of ‘definable’ used in logic failed to provide a match with the Erlanger's concept of ‘geometric’. Narens developed a new definability concept to capture the Erlanger Programme's concept of ‘geometric’. He called the new concept *scientific definability*.
Scientific definability assumes that the quantitative world is constructed from relationships based on real numbers and is completely separated from the qualitative situation under investigation, $\mathcal{X}$, which is conceptualized as a qualitative structure. Unlike definability concepts from logic, scientific definability allows the free use of concepts from the quantitative world for defining relationships based on the domain *X* of a qualitative structure $\mathcal{X}$. Narens shows that a relation on *X* is qualitatively meaningful if and only if it is scientifically defined in terms of $\mathcal{X}$.
There is one obvious case where the Erlanger Programme appears to produce a rmkably poor concept of ‘geometric’. This is where the geometry $\mathcal{X}$ has the identity function as its only symmetry, yielding that every relation on *X* is ‘geometric’, and for measurement situations where the scale consists of a single representation, making each relation on the domain of the numerical representing structure quantitatively meaningful, and thus, by the above theorem, each relation on *X* qualitatively meaningful. There are many important examples of this case, for example the geometry of physical universe under Einstein's

general theory of relativity.

Narens (2002) provides generalizations of ‘scientific definability’ that appear to yield reasonable and productive concepts of ‘geometric’ (‘qualitatively meaningful’) for situations where the geometry (qualitative structure) has the identity as its only symmetry. The main idea for the generalizations is the following. Instead of formulating meaningfulness in terms of a single qualitative structure, a family $\mathfrak{F}$ of isomorphic qualitative structures is used. It is assumed that all the structures in $\mathfrak{F}$ have the same domain called the *common domain* (of $\mathfrak{F}$). A relation *R* on the common domain is said to be $\mathfrak{F}$*-meaningful* if and only if there exist a structure $\mathcal{X}$ in $\mathfrak{F}$, primitives $R_{j_1}, \ldots, R_{j_n}$ of $\mathcal{X}$, and a formula φ used for scientific definitions such that

1. (i) *R* has a scientific definition in terms $R_{j_1}, \ldots, R_{j_n}$ and φ, and
2. (ii) *R* has the same scientific definition for all $\mathcal{X}' = (X, R'_j)_{j \in J}$ in $\mathfrak{F}$; that is, *R* has the same scientific definition as in (*i*) but with $R_{j_1}, \ldots, R_{j_n}$ replaced by $R'_{j_1}, \ldots, R'_{j_n}$.

For the case where $\mathfrak{F}$ consists of a single structure, $\mathfrak{F}$-meaningfulness coincides with qualitative meaningfulness.

See Also

- measurement, theory of

Bibliography

Aczél, J. and Roberts, F.S. 1989. On the possible merging functions. *Mathematical Social Sciences* 17, 205–43.

Aczél, J., Roberts, F.S. and Rosenbaum, Z. 1986. On scientific laws without dimensional constants. *Journal of Mathematical Analysis and Applications* 119, 389–416.

Falmagne, J.-C. 1985. *Elements of Psychophysical Theory*. New York: Cambridge University Press.

Falmagne, J.-C. 2004. Meaningfulness and order-invariance: two fundamental principles for scientific laws. *Foundations of Physics* 34, 1341–8.

Falmagne, J.-C. and Narens, L. 1983. Scales and meaningfulness of quantitative laws. *Synthese* 55, 287–325.

Klein, F. 1872. *Vergleichende Betrachtungen über neuere geometrische Vorschungen: Programm zu Eintritt in die philosophische Facultät und den Senat der Universität zu Erlangen*. Erlangen: Deichert.

Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A. 1971. *Foundations of Measurement*, vol. 1, New York: Academic Press.

Luce, R.D. 1959. On the possible psychophysical laws. *Psychological Review* 66(2), 81–95.

Luce, R.D. 1964. A generalization of a theorem of dimensional analysis. *Journal of Mathematical Psychology* 1, 278–84.

Luce, R.D. 1978. Dimensionally invariant numerical laws correspond to meaningful qualitative relations. *Philosophy of Science* 45, 1–16.

Luce, R.D., Krantz, D.H., Suppes, P. and Tversky, A. 1990. *Foundations of Measurement*, vol. 3, New York: Academic Press.

Narens, L. 1981. A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision* 13, 1–70.

Narens, L. 1985. *Abstract Measurement Theory*. Cambridge, MA: MIT Press.

Narens, L. 2002. *Theories of Meaningfulness*. Mahwah: Lawrence Erlbaum and Associates.

Narens, L. 2007. *Introduction to the Theories of Measurement and Meaningfulness and the Role of Invariance in Science*. Mahwah: Lawrence Erlbaum and Associates.

Narens, L. and Luce, R.D. 1983. How we may have been misled into believing in the interpersonal comparability of utility. *Theory and Decision* 15, 247–60.

Pfanzagl, J. 1968. *Theory of Measurement*. New York: Wiley. 2nd edn., Vienna: Physica, 1971.

Pickering, J.F., Harrison, J.A. and Cohen, C.D. 1973. Identification and measurement of consumer confidence: methodology and some preliminary results. *Journal of the Royal Statistical Society* Series A 136, 43–63.

Roberts, F.S. 1980. On Luce's theory of meaningfulness. *Philosophy of Science* 47, 424–33.

Roberts, F.S. 1985. Applications of the theory of meaningfulness to psychology. *Journal of Mathematical Psychology* 229, 311–32.

Roberts, F.S. and Franke, C.H. 1976. On the theory of uniqueness in measurement. *Journal of Mathematical Psychology* 14, 211–18.

Stevens, S.S. 1946. On the theory of scales of measurement. *Science* 103, 677–80.

Stevens, S.S. 1951. Mathematics, measurement and psychophysics. In *Handbook of Experimental Psychology*, ed. S.S. Stevens. New York: Wiley.

Suppes, P. and Zinnes, J.L. 1963. Basic measurement theory. In *Handbook of Mathematical Psychology*, vol. 1, ed. R.D. Luce, R.R. Bush and E. Galanter. New York: Wiley.

How to cite this article

Narens, Louis and R. Duncan Luce. "meaningfulness and invariance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000121> doi:10.1057/9780230226203.1071

The New Palgrave Dictionary of Economics Online

Means, Gardiner Coit (1896 – 1988)

Leon H. Keyserling

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Berle, A.A.; concentration; corporations; inflation; inflation–unemployment trade-off; Means, G. C.; New Deal; ownership and control

Article

The Modern Corporation and Private Property appeared in 1932, co-authored by Means and Adolph Berle. This book fused the abilities of a great economist and a great lawyer, and became deservedly famous.

The prevalent economic and legal thinking at that time did not recognize adequately the emergence of corporate giantism. It envisaged a system characterized in the main by small private enterprises. And it assumed that this worked well because the law of supply and demand would determine price levels and thus automatically produce adjustments assuring the greatest good of the greatest number. This laissez-faire approach made private property almost sacrosanct and almost free from public intervention.

Berle and Means *proved* that this was not how the economy actually worked. As they revealed, the monster size of existing corporations and their dominating power negated the attributes of private property as then conceived; their ability and determination to fix or ‘administer’ prices prevented the benign operation of supply and demand. These findings suggested that increased government intervention in the private sector was essential in the public interest; and this was fortified by the book's additional findings *in re* economic concentration and the separation between ownership and control. Thus the book indicated among other things the need for legal change, including judicial reinterpretations of governmental powers under the Constitution.

Substantial parts of the book were used to support New Deal and judicial action between 1933 and 1939, which viewed corporations and private property in a new light. And the New Deal brought Means to Washington. He alone among three economic advisers to the Secretary of Agriculture dealt with the effect of farm conditions upon the overall economy. Next, he was Director of the industrial division of the National Resources Planning Board (NRPB), where he developed techniques for depicting what composition of business activity would maintain full employment. This type of work, continued by him

on the staff of the Committee for Economic Development (after a spell as fiscal analyst in the Bureau of the Budget) was intrinsic to the Committee's portrayal of the post-war markets requisite to full employment.

During subsequent decades, Means poured forth a Niagara of writing and speeches. Insistently, he built upon his original thesis of corporate power, especially through its pricing practices. Refuting the prevalent view among economists that there is a ‘trade-off’ between unemployment and inflation, he showed that the great increases in inflation during recent decades have come mainly, not during a highly used economy near full employment, but rather during periods when the economy moved into stagnation and recession. This squared with his early finding that the modern corporation can and does lift prices to compensate for low volume. It also revealed that his humanistic concern about full employment and economic justice must reject the frequent and unsuccessful efforts to achieve price stability by spawning the misery of vast unemployment. Always, unlike so many economists, Means eschewed outmoded or untested theories, and spent himself in exhaustive empirical studies in aid of his own analysis and policy recommendations.

See Also

- Berle, Adolf Augustus, Jr.

Selected work

1932. (With A.A. Berle.) *The Modern Corporation and Private Property*. New York: Commerce Clearing House.

How to cite this article

Keyserling, Leon H. "Means, Gardiner Coit (1896–1988)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000123> doi:10.1057/9780230226203.1072

The New Palgrave Dictionary of Economics Online

mean-variance analysis

Harry M. Markowitz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mean-variance analysis is concerned with combining risky assets in a way that minimizes the variance of risk at any desired mean return. In the use of mean-variance analysis for actual money management, the issue is how to estimate the large number of required covariances. Many-factor models of covariance are widely used, as are scenario and combined scenario and factor models, and constant correlation models. This simplifies the parameter estimation problem and can accelerate the computation of efficient sets for analyses containing hundreds of securities.

Keywords

capital asset pricing models; constant correlation models; corner portfolio; covariance; dynamic programming; factor models; liquidity preference; Markowitz, H. M.; mean-variance analysis; option prices; quadratic approximation; risk aversion; scenario models

Article

In a mean–variance portfolio analysis (Markowitz, 1959) an *n*-component vector (portfolio) *X* is called feasible if it satisfies

$$AX = b \quad X \geq 0$$

where *A* is an $m \times n$ matrix of constraint coefficients, and *b* an *m*-component constant vector. An *EV* combination is called feasible if

$$E = \mu^T X$$
$$V = X^T C X$$

for some feasible portfolio. Here *E* is the expected return of the portfolio, *V* the variance of the portfolio, μ the vector of expected returns of securities, and *C* a positive semidefinite covariance matrix of returns among securities.

A feasible *EV* combination is called inefficient if some other feasible combination has either less *V* and no less *E*, or else greater *E* and no greater *V*. A feasible *EV* combination is called efficient if it is not inefficient. A feasible portfolio *X* is efficient or inefficient according to whether its *EV* combination meets the one definition or the other. As in linear programming, the constraints $(AX = b,\ X \geq 0)$ can represent inequalities by introducing slack variables, and can incorporate variables which are allowed to be negative, by separating the positive and negative parts of such variables.

Markowitz (1956) shows that if *V* is strictly convex over the set of feasible portfolios – for example when *C* is positive definite – the set of efficient portfolios is piecewise linear, and the set of efficient *EV* combinations is piecewise parabolic. There may or may not be a kink in the efficient *EV* set at a ‘corner portfolio’, where two pieces of the efficient portfolio set meet. Markowitz (1959, Appendix A), shows for arbitrary semidefinite *C* that, while there may be more than one efficient portfolio for given efficient *EV* combination, there is a piecewise linear set of efficient portfolios which contains one and only one efficient portfolio for each efficient *EV* combination. The piecewise linear nature of the efficient set is illustrated graphically, for small *n*, in Markowitz (1952) and (1959).

The fact that the mean–variance analysis selects a portfolio for only one period does not imply that the investor plans to retire at the end of the period. Rather, it assumes that in the dynamic programming (Bellman, 1957) solution to the many period investment problem, current wealth is the only state variable to enter the implied single period utility function (see Markowitz, 1959, Ch. 13; Samuelson, 1969; Ziemba and Vickson, 1975). Mossin (1968) shows conditions under which the optimum solution to the many period problem is ‘myopic’ in that the single period utility function is the same as an end-of-game utility function. This is an example of – but not the only example of – a class of games in which wealth is the only state variable.

The Markowitz (1959) justification for the use of mean-variance analysis further assumes that if one knows the *E* and *V* of a portfolio one can estimate with acceptable accuracy the expected value of the one-period utility function. Samuelson (1970) and Ohlson (1975) present conditions under which mean and variance are asymptotically sufficient as the length of holding periods – that is, the intervals between portfolio revisions – approaches zero. For ‘long’ holding periods, for example for time between revisions as long as a year, Markowitz (1959), Young and Trent (1969), Levy and Markowitz (1979), Pulley (1981) and Kroll, Levy and Markowitz (1984) have each found mean-variance approximations to be quite accurate for a variety of utility functions and historical distributions of portfolio return.

This leads to an apparent anomaly: if you know mean and variance you practically know expected utility; the mean-variance approximation to expected utility is based on a quadratic approximation to the single-period utility function; yet Arrow (1965) and Pratt (1964) show that any quadratic utility function

has the objectionable property that an investor with such a utility function becomes increasingly averse to risks of a given dollar amount as his wealth increases. Levy and Markowitz (1979) show that the anomaly disappears if you distinguish three types of quadratic approximation:

1. (1) Assuming that the investor has a utility-of-wealth function that remains constant through time – so that as the investor's wealth changes he moves along the curve to a new position – fit a quadratic to this curve at some instant of time, and continue to use this same approximation subsequently. (Note that the assumption here, that the investor has a constant utility-of-wealth function is sufficient, but not necessary, for the investor to have a single period utility function at each period.)
2. (2) Fit the quadratic to the investor's current single period utility function. For example, if the investor has an unchanging utility-of-wealth function, choose a quadratic to fit well near current wealth (i.e. near portfolio return equal zero).
3. (3) Allow the quadratic approximation to vary from one portfolio to another, that is, let the approximation depend on the mean, and perhaps the standard deviation, of the probability distribution whose expected value is to be estimated.

The Pratt–Arrow objections apply to an approximation of type (1). The approximations proposed in Markowitz (1959) are of types (2) and (3). Levy and Markowitz (1979) show that, under quite general assumptions, the type 3 mean-variance maximizer has the same risk aversion in the small (in the sense of Pratt) as does the original expected utility maximizer.

Uses of mean-variance analysis

Two areas of use deal with: (a) actual portfolio management using mean–variance analysis, and (b) implications for the economy as a whole of the assumption that all investors act according to the mean–variance criteria. We refer to these, respectively, as ‘normative’ and ‘positive’ uses of mean–variance analysis.

The positive application of mean-variance analysis is dealt with elsewhere in this Dictionary. Seminal works in the field include the Tobin (1958) analysis of liquidity preference; and the Sharpe (1964), Lintner (1965) and Mossin (1966) Capital Asset Pricing Models (CAPMs). As in the Tobin model, these CAPMs assume that the investor can either lend all he has or borrow all he wants at the same ‘risk-free’ rate of interest. From this assumption (plus assumptions that all investors have the same beliefs and seek mean–variance efficiency subject to the same constraint set) they conclude that the excess return on each security (its expected return minus the risk-free rate) is proportional to its ‘beta’, where the latter is the regression of the security's return against the return of the market as a whole. Black (1972) drops the assumption that the investor can borrow at a risk-free rate; assumes instead that the investor can sell short and use the proceeds to buy long; and derives a formula for excess return just like that of Sharpe–Lintner–Mossin except that the expected return on a zero-beta portfolio is substituted for the risk-free rate in the formula for excess return. Merton (1969) has developed mean–variance theory in continuous time. This has been used, for example, in the analysis of option prices by Black and Scholes (1973) from which a vast literature of further implications followed.

As compared with the models used in normative analysis, the models of positive analysis tend to use quite simple constraint sets and other special assumptions (e.g. all investors have the same beliefs). The justification for such assumptions is that they give concrete, therefore testable, implications; and indeed have been the subject of extensive empirical testing.

In the use of mean-variance analysis for actual money management, the question immediately arises as to how to estimate the large number of required covariances. Sharpe (1963) concluded, and Cohen and Pogue (1967) confirmed, that a simple one-factor model of covariance was sufficient. King (1966) showed that, in addition to one pervasive factor, there were ample industry sources of covariance. By the mid-1970s it was clear to many practitioners that the one-factor model was not adequate, since, for example, sometimes ‘the market’, as measured by some broad index, went up while high beta stocks went down, to an extent that could not be explained by chance. Many-factor models such as that of Rosenberg (1974) are now widely used.

Other models of covariance used in practice include scenario and combined scenario and factor models (Markowitz and Perold, 1981), and a model which assumes that all correlation coefficients are the same (Elton and Gruber, 1973). The use of factor, scenario or constant correlation models, in addition to simplifying the parameter estimation problem, can considerably accelerate the computation of efficient sets for analyses containing hundreds of securities. For example, the Perold (1984) code will solve large portfolio selection problems for arbitrary A and C, but is especially efficient in handling upper bounds on variables and sparse (mostly zero) A and C matrices. (The introduction of ‘dummy’ securities into the analysis allows one to ‘sparsify’ the C matrix for factor, scenario or constant correlation models.) Even faster solutions are obtained by Elton, Gruber and Padberg (1976 and 1978) for the one-factor and constant correlation models for certain common constraint sets.

See Also

- capital asset pricing model
- finance

Bibliography

Arrow, K. 1965. *Aspects of the Theory of Risk Bearing*. Helsinki: Yrjö Jahnsson Foundation.

Bellman, R.E. 1957. *Dynamic Programming*. Princeton: Princeton University Press.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45, 444–55.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Cohen, J.K. and Pogue, J.A. 1967. An empirical evaluation of alternative portfolio-selection models. *Journal of Business* 40, April, 166–93.

Elton, E.J. and Gruber, M.J. 1973. Estimating the dependence structure of share prices. *Journal of Finance* 28, 1203–32.

Elton, E.J., Gruber, M.J. and Padberg, M.W. 1976. Simple criteria for optimal portfolio selection. *Journal of Finance* 31, 1341–57.

Elton, E.J., Gruber, M.J. and Padberg, M.W. 1978. Simple criteria for optimal portfolio selection: tracing out the efficient frontier. *Journal of Finance* 33, 296–302.

King, B.F. 1966. Market and industry factors in stock price behavior. *Journal of Business* 39, Supplement, January, 139–90.

Kroll, Y., Levy, H. and Markowitz, H.M. 1984. Mean variance versus direct utility maximization. *Journal of Finance* 39, 47–61.

Levy, H. and Markowitz, H.M. 1979. Approximating expected utility by a function of mean and variance. *American Economic Review* 69, 308–17.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, February, 13–37.

Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7, 77–91.

Markowitz, H.M. 1956. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly* 3, 111–33.

Markowitz, H.M. 1959. *Portfolio Selection: Efficient Diversification of Investments*. New Haven: Yale University Press. Reprinted, New York: John Wiley and Sons, 1970.

Markowitz, H.M. and Perold, A. 1981. Portfolio analysis with factors and scenarios. *Journal of Finance* 36, 871–7.

Merton, R.C. 1969. Lifetime portfolio selection under uncertainty: the continuous-time case. *Review of Economics and Statistics* 51(3), 247–57.

Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34, 768–83.

Mossin, J. 1968. Optimal multiperiod portfolio policies. *Journal of Business* 41, 215–29.

Ohlson, J.A. 1975. The asymptotic validity of quadratic utility as the trading interval approaches zero. In

Stochastic Optimization Models in Finance, ed. W.T. Ziemba and R.G. Vickson. New York: Academic Press.

Perold, A.F. 1984. Large-scale portfolio optimization. *Management Science* 30, 1143–60.

Pratt, J.W. 1964. Risk aversion in the small and in the large. *Econometrica* 32, 122–36.

Pulley, L.B. 1981. A general mean-variance approximation to expected utility for short holding periods. *Journal of Financial and Quantitative Analysis* 16, 361–73.

Rosenberg, B. 1974. Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis* 9(2), 263–74.

Samuelson, P.A. 1969. Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics* 51(3), 239–46.

Samuelson, P.A. 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances and higher moments. *Review of Economic Studies* 37, 537–42.

Sharpe, W.F. 1963. A simplified model for portfolio analysis. *Management Science* 9(2), 277–93.

Sharpe, W.F. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25, 65–86.

Young, W.E. and Trent, R.H. 1969. Geometric mean approximations of individual security and portfolio performance. *Journal of Financial and Quantitative Analysis* 4(2), 179–99.

Ziemba, W.T. and Vickson, R.G., eds. 1975. *Stochastic Optimization Models in Finance*. New York: Academic Press.

How to cite this article

Markowitz, Harry M. "mean-variance analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000124> doi:10.1057/9780230226203.1073

The New Palgrave Dictionary of Economics Online

measure theory

A.P. Kirman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Measure theory is that part of mathematics which is concerned with the attribution of weights of ‘measure’ to the subsets of some given set. Such a measure is required to satisfy a natural condition of additivity, that is that the measure of the union of disjoint sets should be equal to the sum of the measure of those sets. The fundamental problems of measure arise when one has to treat infinite sets or infinite unions of sets. It is perhaps not clear why such a tool should be of use in economics.

Apart from the rather trivial observation that, since measure theory provides the basis for probability theory it underlies all of the economics of uncertainty, there have been direct applications of this theory to several basic problems in economic theory (for a more detailed account, see Kirman, 1982). A first example of such an application is given by the idea of ‘pure’ or ‘perfect’ competition. The fundamental characteristic of a perfectly competitive economic situation is one in which no individual can influence the outcome. Thus, in a competitive market economy, although prices are the result of the collective activity of all the agents, no individual by acting alone can modify them and hence takes them as given. Now strictly speaking in a finite economy this cannot be true and in the work of Torrens, Cournot and Edgeworth can be found lengthy discussions as to whether it is rational for individuals to behave in a perfectly competitive way. Indeed, as Viner once observed, the fact that it is profitable for them to do otherwise is a ‘skeleton in the cupboard of free trade’.

Economists have typically avoided the contradiction involved in analysing economies in which individuals do have positive influence but behave as if they do not, by saying that individuals behave ‘as if’ or ‘believe that’ they have no effect on the outcome. To a mathematician there is no contradiction involved in the idea of individual elements having no weight but sets of such elements having positive weight. If we think of the unit interval, each point has no length but sub-intervals made up of such points do have positive weight. This is, of course, due to the fact that there are infinitely many, indeed a continuum, of such points. Aumann (1964) in his path-breaking article made use of these ideas to define an ‘ideal’ or ‘continuum’ economy which corresponded logically to the idea of perfect competition. If instead of the set of agents A in an economy being finite, we substiute the unit interval [0, 1] a continuum exchange economy can be defined by $e[0, 1] \rightarrow \mathcal{P}_{mo} \times R^{l}_{+}$ where *l* is the number of goods

and $\mathcal{P}mo$ is the set of monotonic continuous preferences on R^l_+ positive orthant of Euclidian *l* space. Thus with each agent or point is associated a preference relation and an initial bundle of goods. Now we have defined an economy which has the right framework for perfect competition. To be able to use this model requires a little more. If we think of an allocation *f* which assigns to each agent a bundle of goods how do we say that what is allocated is equal to the sum of the initial resources *e*(*a*) of the agents. To write

$$\sum_{\alpha \in A} e(a) = \sum_{\alpha \in A} f(a)$$

no longer makes sense. However, in a finite economy with *n* agents, we could also write.

$$\underset{\text{averages resources}}{1/n \sum e(a)} = \underset{\text{average allocation}}{1/n \sum f(a)}$$

without changing anything. In the continuum economy, just such a statement can be made by writing

$$\int e(a) = \int f(a) .$$

When taking an average in this way by integrating we are assigning weights to the various subsets of agents. In other words, we integrate ‘with respect to some measure μ‘. In the case where *A*=[0, 1] there is a natural measure (Lebesgue measure) which corresonds to the length of the intervals which make up a set. This allows us to carry through all the standard analysis in such an economy and indeed allows one to obtain two interesting results which do not hold in finite economies. The first is that in such an economy a competitive equilibrium exists even if preferences are not convex. The second is that the set of Walrasian allocations $W(\mathcal{E})$ is equal to the set of allocations which no coalitition can improve upon, called the core $C(\mathcal{E})$ of the economy (see cores). This last result is the ‘perfect’ or ‘ideal’ version of an old result of Edgeworth. In fact, Aumann's results can be shown to be approximately true for large but finite economies and thus, as one might hope, the ideal case gives us a good idea of what happens in large economies.

Two observations are in order. In fact, the choice of the unit interval and Lebesgue measure is arbitrary. All that one needs is a *measure space* $(A, \mathcal{A}, \mu)$ where *A* is the set of agents $\mathcal{A}$ is the collection of subsets or coalitions of agents and μ is the measure of these subsets. $\mathcal{A}$ can be thought of as the set of all subsets

of *A* although strictly speaking this is not correct. What is required to model perfect competition is that no individual has weight. Thus one must add the condition that the measure space be ‘atomless’ that is for any set *C* with $\mu(C) > 0$ there must be a subset *B* contained in *C* with $\mu(C) > \mu(B) > 0$. This is in contradiction with the standard term ‘atomistic competition’ which is supposed to describe perfect competition. Another aspect of economies which implicitly makes use of the notion of a continuum economy is the discussion of the *distribution of agents’ characteristics*. It is common practice in economics to use a continuous function such as the Pareto distribution to describe the income distribution. For this to be fully appropriate a continuum economy is needed. How may we formally describe distributions? Suppose that we start with a measure space of agents as explained above. Now consider an economy ε i.e. an attribution of preferences and initial resources to each agent. Take a set *B* in the characteristics space and consider the set *C* of those agents who have characteristics in *B* i.e. $C = \varepsilon^{-1}(B)$ Now let the measure $\Psi(B) = \mu(C)$ Thus the measure μ on the set of agents induces another measure Ψ on the set of characteristics. This defines the distribution of characteristics in that economy. Now, one could maintain that a good argument for the distribution approach would be that two economies with the same distribution of characteristics should have the same equilibria, for example. Hildenbrand (1975) gives a detailed discussion of this problem. The general merit of the distribution approach is of course that individualistic descriptions of the characteristics of agents make little economic sense in very large economies. Furthermore, in such economies, putting conditions on the distribution of characteristics may help in restricting the class of outcomes that may be observed. An illustration of the necessity for this is given by the results of Sonnenschein (1973) and Debreu (1974) which show that all the standard individualistic assumptions on individuals put no restrictions on the aggregate excess demand of an economy other than continuity and Walras's Law. This means that there is essentially no *a priori* restriction on the form of aggregate excess demand functions and hence on observable outcomes. Indeed, in finite economies, even specifying the income distribution does not help (see Kirman and Koch, 1986). However, Hildenbrand (1983) has shown that, in a continuum economy, if one puts a condition on the income distribution, then the ‘law of demand’ is satisfied. This ‘normality’ of goods with respect to prices is a fairly strong restriction on excess demand functions and indicates that other results in the same direction might be obtained.

Rather than make assumptions about the specific form of the distribution it is sometimes useful to be able to say something about how ‘dispersed’ agents characteristics are. This involves requiring that the ‘support’ of the measure Ψ representing the distribution of characteristics, that is the smallest set which has full measure, should not be ‘too small’. For example, a bothersome feature of the standard assumption of convex rather than strictly convex preferences is that demand is a ‘correspondence’ rather than a function. This involves considerable technical difficulties. However, it has been shown by various authors (an account may be found in Mas-Colell (1985) for example) that if the support of the distribution of agents characteristics is sufficiently large then aggregate demand will be a function and not a correspondence.

Another use of measure theory is to give precision to the idea that phenomena are ‘unlikely’. Thus one cannot exclude, for example, the possibility that an economy will have an infinite set, even a continuum, of equilibrium allocations. However, as Debreu (1970) has shown, ‘almost no’ economies have this property. To see the idea consider an Edgeworth box representing a two man, two good exchange economy. Each point in the box can be considered as a possible location of the individual endowments.

Naturally, the equilibria vary with initial endowments. What is true is that if we consider the set of endowments which give rise to infinite equilibria, its ‘area’ or ‘measure’ is zero. Thus the probability that an economy drawn at random, in some sense, will have infinite equilibria is zero.

A classic problem which has received considerable attention is that of how to divide some object fairly among n individuals. Suppose that the object is not homogeneous, a cake with different layers for example, then if an individual assigns value 1 to the whole cake he can give a value to any piece of the cake. In other words, each individual i has a measure μ_i on the cake. It has been shown that it is possible to find partitions of U ($U_1 \ldots U_n$) so that each individual i considers that his share U_i is worth more than $1/n$ of the cake. This does not exclude some individuals being jealous of each other. However, Dubins and Spanier (1961) have shown that it is possible to find partitions where each individual considers that all the pieces U_i of the cake are worth $1/n$. Thus:

$$\mu(U_j) = 1/n\text{i}, \; j = 1 \ldots n$$

and everybody believes that the division is perfectly equitable.

Another illustration of the measure theoretic approach is the following. Arrow (1963) discussed the problem of establishing a rule which aggregates individual preferences on a set of social allocations into social preferences. He showed that if all individual preferences are allowed then no rule satisfying certain basic axioms exists. In particular, he showed that his first axioms implied that there must be a ‘dictator’, who has the property that if he prefers state x to state y, then society prefers x to y. Fishburn (1970) showed that this was not true in a society with an infinite number of individuals, thus raising hopes that in large economies Arrow's result loses its importance. In fact, this is not the case, Arrow's axioms impose a very special structure on those sets of individuals who ‘dictate’ society's preferences in the above sense. This structure implies that no matter how large the finite economy there will always be a dictator. Thus the infinite case is exceptional. However, in the infinite society individuals make little sense and one can give a measure theoretic equivalent of Arrow's result. For a society in which the set of individuals is represented by the unit interval then any dictatorial set C contains a dictatorial set B with positive but smaller measure that is

$$\mu(C) > \mu(B) > 0 \text{with} B \subset C.$$

Thus there are dictatorial sets of arbitrarily small measure.

As a last example consider the problem of ‘temporary equilibrium’. In an economy in which one can only transfer wealth to the future by keeping money and in which individuals anticipate future prices, one wishes to find an equilibrium for the goods and money markets today. Each individual forms a distribution over tomorrow's prices p_2 as a function ψ of today's prices p_1. Now if for example,

individuals always expect prices tomorrow to be higher than today there may be no incentive at any prices to hold money. In this case, there can be no equilibrium. However, if we require that prices tomorrow should not be ‘too dependent’ on today's then equilibrium exists. Formally, we require that the family of the price distributions ψ over all prices should be ‘tight’. An explanation of this with results is given by Grandmont (1977). However, intuitively, it is clear that one excludes the ever increasing expectations that lead to hyperinflation.

These examples illustrate the ways in which a formal mathematical tool, measure theory, has been incorporated into economic theory. In particular, its use in characterizing ideal economies, those corresponding to the notion of perfect competition, has been invaluable.

See Also

- cores
- functional analysis
- Lyapunov functions
- non-standard analysis

Bibliography

Any standard text in measure theory such as Halmos (1971) will give the essential mathematical notions, and more specialized references are given in the bibliographies of the articles cited here.

Arrow, K.J. 1963. *Social Choice and Individual Values*. 2nd edn, New York: Wiley.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1(1), 15–21.

Dubins, I.E. and Spanier, E.H. 1961. How to cut a cake fairly. *American Mathematical Monthly* 1, 1–17.

Fishburn, P.C. 1970. Arrow's impossibility theorem, concise proof and infinite voters. *Journal of Economic Theory* 2, 103–6.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45(3), 535–72.

Halmos, P.R. 1961. *Measure Theory*. 7th edn, Princeton: Van Nostrand.

Hildenbrand, W. 1975. Distributions of agents characteristics. *Journal of Mathematical Economics* 2,

129–38.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51(4), 997–1020.

Kirman, A.P. 1982. Measure theory with applications to economics. Chapter 5 in *Handbook of Mathematical Economics*, ed. K.J. Arrow and M. Intriligator, Vol. 1, 159–209.

Kirman, A.P. and Koch, K.J. 1986. Market excess demand in economies with identical preferences and co-linear endowments. *Review of Economic Studies* 53(3), 457–64.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium*. Cambridge; Cambridge University Press.

Sonnenschein, H. 1973. Do Walras's identity and continuity characterise the class of community excess demand functions. *Journal of Economic Theory* 6(4), 345–54.

How to cite this article

Kirman, A. P. "measure theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000126> doi:10.1057/9780230226203.1074

The New Palgrave Dictionary of Economics Online

measurement error models

Han Hong
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Measurement error is important in econometric analysis. Its presence causes inconsistent parameter estimates. Under the classical measurement error assumption, instrumental variable methods can be used to eliminate the bias caused by measurement errors using a second measurement. This technique can be extended to polynomial regression models. For nonlinear models, deconvolution methods have been developed to cope with classical measurement errors. When the classical measurement error assumption is violated, auxiliary data-sets are usually needed to provide additional source of identification. When the true variable takes only discrete values, the mismeasurement problem takes the form of misclassification and requires special techniques.

Keywords

attenuation bias; auxiliary data; deconvolution method; generalized method of moments; instrumental variables; inverse probability weighting estimation; linear models; measurement error models; misclassification; nonlinear models; permanent-income hypothesis; polynomial regression; semiparametric method; sieve estimator

Article

1 Introduction

Many economic data-sets are contaminated by the mismeasured variables. Measurement error is one of the fundamental problems in empirical economics. The presence of measurement errors causes biased and inconsistent parameter estimates, and leads to erroneous conclusions to various degrees in both linear and nonlinear econometric models. Techniques for addressing measurement error problems can be classified along two dimensions. Different techniques are used in linear models and in nonlinear models. Measurement error models that are valid under the classical measurement error assumption often are not applicable when the classical measurement error assumption does not hold.

2 Linear models with classical measurement errors

The classical measurement error assumption maintains that the measurement errors in any of the variables in a data-set are independent of all the true variables that are the objects of interest. The implication of this assumption in the linear least square regression model $y_i^* = x_i^{*\prime}\beta + \varepsilon_i$ is well understood and is usually described in standard econometrics textbooks. Under this assumption, measurement errors in the dependent variable $y_i = y_i^* + v_i$ do not lead to inconsistent estimates of the regression coefficients. Its only consequence is to inflate the standard errors of those regression coefficient estimates. On the other hand, independent errors that are present in the observations of the regressors $x_i = x_i^* + \eta_i$ lead to attenuation bias in simple univariate regression models and to inconsistent regression coefficient estimates in general. The importance of measurement errors in analysing the empirical implications of economic theories is highlighted in Milton Friedman's seminal book on the consumption theory of the permanent income hypothesis (Friedman, 1957). In Friedman's model, both consumption and income consist of a permanent component and a transitory component that can arise from measurement errors or genuine fluctuations. The marginal propensity to consume relates the permanent component of consumption to the permanent income component. Friedman showed that, because of the attenuation bias, the slope coefficient of a regression of observed consumption on observed income would lead to an underestimate of the marginal propensity to consume.
Econometric work on linear models with classical independent additive measurement error dates back to Frish (1934), who derived bounds on the slope and the constant term. Instrumental variables (IV) is a popular method for obtaining consistent point estimators of the parameters of interest in this classical independent additive measurement error model. A valid instrument often comes from the second measurement of the error-prone true variable: $w_i = x_i^* + v_i$ which is subject to another independent measurement error v_i. The second measurement w_i is a valid instrument for the first measurement x_i because it is independent of both ε_i and η_i, but is correlated with the regressor x_i based on the first measurement.
The double-measurement instrumental variable method for linear regression models has been generalized by Hausman et al. (1991) to certain nonlinear regression models in which the regressors are polynomial functions of the error-prone variables. The following is a simplified version of the polynomial regression model that they considered:

$$y = \sum_{j=0}^{K} \beta_j z^j + r'\varphi + \varepsilon.$$

Among the two sets of regressors z and r, r is precisely observed but z is observed only with errors. In particular, two measurements of z, x and w, are observed which satisfy

$$x = z + \eta \text{ and } w = z + v.$$

An i.i.d. sample of observations is assumed to be available. Therefore we focus on identification of population moments. For convenience, assume that ε, η and v are mutually independent and they are independent of all the true regressors in the model.

First assume that $\omega = 0$, then identification of β depends on population moments $\xi_j \equiv E(yz^j)$, $j = 0, \ldots, K$ and $\zeta_m \equiv Ez^m$, $m = 0, \ldots, 2K$, which are the elements of the population normal equations for solving for β. Except for ξ_0 and ζ_0, these moments depend on z which is not observed, but they can be solved from the moments of observable variables Exw^{j-1}, Ew^l for $j = 0, \ldots, 2K$ and Eyw^j, $j = 0, \ldots, K$. Define $v_k = Ev^k$. Then the observable moments satisfy the following relations:

$$Exw^j = E(z+\eta)(z+v)^j = E\sum_{l=0}^{j}\binom{j}{l}(z+\eta)z^l v^{j-l} = \sum_{l=0}^{j}\binom{j}{l}\zeta_{l+1}v_{j-l},\ j = 1, 2K-1, \tag{1}$$

and

$$Ew^j = E(z+v)^j = E\sum_{l=0}^{j}\binom{j}{l}z^l v^{j-l} = \sum_{l=0}^{j}\binom{j}{l}\zeta_l v_{j-l},\ j = 1, \ldots, 2K, \tag{2}$$

and

$$Eyw^j = Ey(z+v)^j = E\sum_{l=0}^{j}\binom{j}{l}yz^l v^{j-l} = \sum_{l=0}^{j}\binom{j}{l}\xi_l v_{j-l},\ j = 1, \ldots, K. \tag{3}$$

Since $v_1 = 0$, we have a total of ($5K - 1$) unknowns in $\zeta_1, \ldots, \zeta_{2K}$, $\xi_1, \ldots, \xi_K$ and $v_2, \ldots, v_{2K}$. Equations (1), (3) and (4) give a total of $5K - 1$ equations that can be used to solve for these $5K - 1$ unknowns. In particular, the $4K - 1$ eqs. in (1) and (3) jointly solve for $\zeta_1, \ldots, \zeta_{2K}$, $v_2, \ldots, v_{2K}$. Subsequently, given knowledge of these ζ's and ν's, ξ's can then be recovered from eq. (4). Finally, we can use these identified quantities of ξ_j, $j = 0, \ldots, K$ and ζ_m, $m = 0, \ldots, 2K$ to recover the parameters β from the normal equations

$$\xi_l = \sum_{j=0}^{K}\beta_j\zeta_{j+l},\ l = 0, \ldots, K.$$

When $\omega \neq 0$, Hausman et al. (1991) noted that the normal equations for the identification of β and ϕ depends on a second set of moments Eyr, Err' and Erz^j, $j = 0, \ldots, K$, in addition to the first set of moments ξ's and ζ's. Since Eyr and Err' can be directly observed from the data, it only remains to identify Erz^j, $j = 0, \ldots, K$. But these can be solved from the following system of equations, for $j = 0, \ldots, K$:

$$Erw^j = Er(z+v)^j = E\sum_{l=0}^{j}\binom{j}{l}rz^l v^{j-l} = \sum_{l=0}^{j}\binom{j}{l}\left(Erz^l\right)v_{j-l},\ j = 0, \ldots, K.$$

In particular, using the previously determined ν coefficients, the jth row of the previous equation can be solved recursively to obtain

$$Eyz^j = Eyw^j - \sum_{l=0}^{j-1} \binom{j}{l} \left(Ez^l r\right) v_{j-l}.$$

Once all these elements of the normal equations are identified, the coefficients β and ϕ can then be solved from the normal equations $\left[Eyz', Eyr\right]' = D\left[\beta', \varphi'\right]$, where $z = (1, z, \ldots, z^K)$ and $D = E\left[\left(z'r'\right), \left(z'r'\right)\right]$.

3 Nonlinear model with classical measurement errors

The deconvolution method is a useful technique to analyse general nonlinear model

$$Em\left(y^*; \beta\right) = 0$$

under the classical measurement error assumption with double measurements. These techniques are developed by Schennach (2004), Li (2002) and Taupin (2001). Suppose one knows the characteristic function $\psi_n(t) = Ee^{it\eta_i}$ of the errors η_i where only $y_i = y_i^* + \eta_i$ is observed and $y^j \in R^k$. Then the characteristic function of y_i^* can be recovered from the ratio of the characteristic functions $\widehat{\varphi}_y(t)$ and $\phi_n(t)$ of y_i and η_i:

$$\widehat{\varphi}_{y^*}(t) = \widehat{\varphi}_y(t) \,/\, \varphi_\eta(t).$$

where $\widehat{\varphi}_y(t)$ can be estimated using a smooth version of $\frac{1}{n}\sum_{i=1}^{n} e^{ity_i}$. Once the characteristic function of y^* is known, its density can be recovered from the inverse Fourier transformations

$$\hat{f}(y^*) = \left(\frac{1}{2\pi}\right)^k \int \widehat{\varphi}_{\mathbf{y}^*}(\mathbf{t})\, e^{-i\mathbf{y}^{*\prime}\mathbf{t}} d\mathbf{t}.$$

For each β, a sample analog of the moment condition can then be estimated by

$$\int m\left(y^*; \beta\right) \hat{f}\left(y^*\right) dy^*.$$

A semiparametric generalized method of moment (GMM) estimator can be formed by minimizing over β a quadratic distance of the above estimated moment condition from zeros. Often, the characteristic function of the measurement errors $\phi_n(t)$ might not be known. However, if two independent measurements of the latent true variable y^* with additive errors are observed and the errors are i.i.d, an estimate of $\widehat{\varphi}_y(t)$ can be obtained using the two independent measurements.

For certain parametric families of the measurement error distribution, $\phi_n(t)$ can be parameterized and its parameters can be estimated jointly with β. Hong and Tamer (2003) assume that the marginal distributions of the measurement errors are Laplace (double exponential) with zero means and unknown variances, and the measurement errors are independent of the latent variables and are independent of each other. Under these assumptions, they derive simple revised moment conditions in terms of the observed variables that lead to a simple estimator for nonlinear method of moment models with measurement error of the classical type when no additional data are available.

When the distributions of η are independent double Laplace, its characteristic function takes the form of

$$\varphi_\eta(t) = \Pi_{j=1}^{k}\left(1 + \frac{1}{2}\sigma_j^2 t_j^2\right)^{-1}.$$

Using this characteristic function, Hong and Tamer (2003) (Theorem 1) show that the moment condition $Em\left(y^*;\beta\right)$ can be translated into observable variable *y* as

$$Em\left(\mathbf{y}^*;\beta\right) = Em(\mathbf{y};\beta) + \sum_{l=1}^{k}\left(-\frac{1}{2}\right)^l \sum_{j_1<\cdots<j_l}\cdots\sum \sigma_{j1}^2\cdots\sigma_{jl}^2 E\frac{\partial^{2l}}{\partial y_{j1}^2\cdots\partial x_{jl}^2} m(\mathbf{y};\beta).$$

For each candidate parameter value β, the right-hand side of the above can be estimated from the sample analog by replacing the expectation with the empirical sum. It can then be used to form a quadratic GMM objective function which can be used to estimate jointly β and the variance parameters $\sigma_j^{'s}$ of the double exponential distributions.

4 Non-classical measurement errors

The recent applied economics literature has raised concerns about the validity of the classical measurement error assumption. For example, in economic data it is often the case that data-sets rely on individual respondents to provide information. It may be hard to tell whether or not respondents are making up their answers and, more crucially, whether the measurement error is correlated with some of the variables. Studies by Bound and Krueger (1991), Bound et al. (1994) and Bollinger (1998) have all documented evidences of non-classical measurement errors. In order to obtain consistent estimates of the parameters β in the moment conditions $m\left(y^*;\beta\right)$, Chen, Hong and Tamer (2005) and Chen, Hong and Tarozzi (2004) make use of an auxiliary data-set to recover the correlation between the measurement errors and the underlying true variables by estimating the conditional distribution of the measurement errors given the observed reported variables or proxy variables. In their model, the auxiliary data-set is a subset of the primary data, indicated by a dummy variable $D = 0$, which contains both the reported variable *Y* and the validated true variable Y^*. Y^* is not observed in the rest of the primary data-set ($D = 1$) which is not validated. The authors assume that the conditional distribution of the true variables given the reported variables can be recovered from the auxiliary data-set:

Assumption 4.1: $Y^* \perp D|Y$.

Under this assumption, an application of the law of iterated expectations gives

$$E[m(Y^*;\beta)] = \int g(Y;\beta)\, f(Y)\, dY \text{ where } g(Y;\beta) = E[m(Y^*;\beta)|Y, D = 0].$$

This suggests a semiparametric GMM estimator for the parameter β. For each value of β in the parameter space, the conditional expectation function $g(Y;\beta)$ can be nonparametrically estimated using the auxiliary data-set where $D = 0$. Chen, Hong, and Tamer (2005) suggest using sieve methods to implement this nonparametric regression. Let *n* denote the size of the entire primary data-set and let n_a denote the size of the auxiliary data-set where $D = 0$. Let $\{q_l(Y), l = 1, 2, \ldots\}$ denote a sequence of known basis functions that can approximate any square-measurable function of *X* arbitrarily well. Also let

$$q^{k(n_a)}(Y) = \left(q_1(Y), \ldots, q_{k(n_a)}(Y)\right)' \text{and}$$
$$Q_a = \left(q^{k(n_a)}(Y_{a1}), \ldots, q^{k(n_a)}(Y_{an_a})\right)'$$

for some integer $k(n_a)$, with $k(n_a) \to \infty$ and $k(n_a)/n \to 0$ when $n \to \infty$. In the above Y_{aj} denotes the *j*th observation in the auxiliary sample. Then for each given β, the first step nonparametric estimation can be defined as,

$$\hat{g}(Y;\beta) = \sum_{j=1}^{n_a} m\left(Y_{aj}^*;\beta\right) q^{k(n_a)}\left(Y_{aj}\right)\left(Q_a' Q_a\right)^{-1} q^{k(n_a)}(Y).$$

A GMM estimator for β_0 can then be defined using a positive definite weighting matrix $\widehat{W}$ as

$$\hat{\beta} = \underset{\beta\in B}{\operatorname{argmin}}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{g}(Y_i;\beta)\right)' \widehat{W}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{g}(Y_i;\beta)\right).$$

Chen, Hong, and Tarozzi (2004) show that a proper choice of $\hat{W}$ achieves the semiparametric efficiency bound for the estimation of β. They called this estimator the 'conditional expectation projection estimator'.
Assumption (4.1) allows the auxiliary data-set to be collected using a *stratified sampling* design where a *non-random response-based subsample* of the primary data is validated. In a typical example of this stratified sampling design, we first oversample a certain subpopulation of the mismeasured variables Y, and then validate the true variables Y^* corresponding to this nonrandom stratified subsample of Y. It is very natural and sensible to oversample a sub-population of the primary data-set where more severe measurement error is suspected to be present. Assumption 3.1 is valid as long as, in this sampling procedure of the auxiliary data-set, the sampling scheme of Y in the auxiliary data is based only on the information available in the distribution of the primary data-set $\{Y\}$. For example, one can choose a subset of the primary data-set $\{Y\}$ and validate the corresponding $\{Y^*\}$, in which case the Y's in the auxiliary data set are a subset of the primary data Y. The stratified sampling procedure can be illustrated as follows. Let U_{pi} be i.i.d $U(0,1)$ random variables independent of both Y_{pi} and Y^*_{pi}, and let $T(Y_{pi}) \in (0,1)$ be a measurable function of the primary data. The stratified sample is obtained by validating every observation for which $U_{pi} < T(Y_{pi})$. In other words, $T(Y_{pi})$ specifies the probability of validating an observation after Y_{pi} is observed.
A special case of assumption 3.1 is when the auxiliary data is generated from the same population as the primary data, where a full independence assumption is satisfied:
Assumption 4.2: $Y, Y^* \perp D$.
This case is often referred to as a validation sample. Semiparametric estimators that make use of a validation sample include Carroll and Wand (1991), Sepanski and Carroll (1993), Lee and Sepanski (1995), and the recent work of Devereux and Tripathi (2005). Interestingly, in the case of a validation sample, Lee and Sepanski (1995) suggest that the nonparametric estimation of the conditional expectation function $g(Y; \beta)$ can be replaced by a finite dimensional linear projection $h(Y; \beta)$ into a fixed set of functions of Y. In other words, instead of requiring that $k(n_a) \to \infty$ and $k(n_a)/n \to 0$, we can hold $k(n_a)$ to be a fixed constant in the above least square regression for $\hat{g}(Y; \beta)$. Lee and Sepenski (1995) show that this will still produce a consistent and asymptotically normal estimator for β as long as the auxiliary sample is also a validation sample that satisfies assumption 4.2. However, if the auxiliary sample satisfies assumption 4.1 but not assumption 4.2, then it is necessary to require $k(n_a) \to \infty$ to obtain consistency. Furthermore, even in the case of a validation sample, requiring $k(n_a) \to \infty$ typically results in a more efficient estimator for β than a constant $k(n_a)$.
An alternative consistent estimator that is valid under assumption 4.1 is based on the *inverse probability weighting* principle which provides an equivalent representation of the moment condition $Em(y^*; \beta)$. Define $p(Y) = p(D = 1|Y)$,

$$Em\left(y^*; \beta\right) = E\left[m(Y^*; \beta_0)\frac{1-p}{1-p(Y)}|D = 0\right].$$

To see this, note that

$$E\left[m(Y^*; \beta_0)\frac{1-p}{1-p(Y)}|D = 0\right] = \int m(Y^*; \beta_0)\frac{1-p}{1-p(Y)}\frac{f(Y)(1-p(Y))f(Y^*|Y, D = 0)}{1-p}dY^*dY = \int m(Y^*; \beta_0)f(Y^*|Y)f(Y)dY^*dY = Em(y^*; \beta),$$

where the third equality follows from assumption 3.1 that $f(Y^*|Y, D = 0) = f(Y^*|Y)$.
This equivalent reformulation of the moment condition $Em(Y^*; \beta)$ suggests a two-step inverse probability weighting estimation procedure. In the first step, one typically obtains a parametric or nonparametric estimate of the so-called propensity score $\hat{p}(Y)$ using, for example, a logistic binary choice model with a flexible functional form. In the second step, a sample analog of the re-weighted moment conditions is computed using the auxiliary data-set:

$$\hat{g}(\beta) = \frac{1}{n_a}\sum_{j=1}^{n_a} m\left(Y_j^*; \beta\right)\frac{1}{1-\hat{p}(Y_j)}.$$

This is then used to form a quadratic norm to provide a GMM estimator:

$$\hat{\beta} = \arg\min_{\beta}\hat{g}(\beta)\, W_n\hat{g}(\beta).$$

Interestingly, an analog of the conditional independence assumption 4.1 is also rooted in the program evaluation literature and is typically referred to as the assumption of un-confoundedness, or selection based on observables. Semiparametric efficiency results for the mean treatment effect parameters to nonlinear GMM models have been developed by, among other, Robins, Mark, and Newey (1992), Hahn (1998), Hirano, Imbens, and Ridder (2003) and Imbens, Newey, and Ridder (2005). Many of the results presented here generalize these results for the mean treatment effect parameters to nonlinear GMM models.

5 Misclassification of binary of discrete variables

Measurement problems on binary or discrete variables usually take the form of *mis-classification*: for example, a unionized worker might be mis-classified as one who is not unionized. When the variable of interest and its measurement are both binary, the measurement error can not be independent of the true binary variable. Typically, mis-classification introduces a negative correlation, or mean reversion, between the errors and the true values. Estimation methods that address

the mis-classification problem have been developed by, among others, Abrevaya, Hausman, and Scott-Morton (1998), Manski and Horowitz (1995), Molinari (2005) and Mahajan (2006).
In particular, the recent work by Mahajan (2006) studies a nonparametric regression model where one of the true regressors is a binary variable:

$$y = g(x^*, z) + \varepsilon, \text{ where } E(\varepsilon | x^*, z) = 0.$$

Instead of observing x^*, the researchers are able only to observe a potentially misreported binary value x of x^*. In the rest of this section we present the identification and estimation results developed in Mahajan (2006).
Mahajan (2006) assumes that, in addition, another random variable v is observed such that the following four assumptions hold.

Assumption 5.1: $E(y|x^*, z, x, v) = g(x^*, z)$.

Assumption (5.1) requires that conditional on the true variable x^*, the measurement error $x - x^*$ does not provide additional information about the outcome variable y. It also requires that v satisfies the following additional assumptions.

Assumption 5.2: $x \perp v | x^*, z$,

and for $\eta_2^*(z, v) = P(x^* = 1 | z, v)$,

Assumption 5.3: $\eta_2^*(z, v)$ is a non-trivial function of v.

Mahajan (2006) calls the variable v an *instrument like variable* that is conditionally independence of the outcome y (assumption 5.1) and of the misreported value x (assumption 5.2), but is correlated with x^* given z (assumption 5.3).
Assumption 5.1 is similar to the exclusion restriction for instrument variables in standard linear models. Assumption 5.3 is analogous to the requirement that an instrument should be correlated with regressors. Because of assumption 5.2, assumption 5.3 implies that $\eta_2(z, v) = P(x = 1 | z, v)$ is also a non-trivial function of v given z.
In addition, Mahajan (2006) also imposes the following monotonicity assumption to restrict the extent of misclassification:

Assumption 5.4: Define $\eta_0(z) = P(x = 1 | x^* = 0, z)$, and $\eta_1(z) = P(x = 0 | x^* = 1, z)$. $\eta_0(z) + \eta_1(z) < 1$.

This assumption is innocuous since it can almost certainly be satisfied by relabelling the binary variables. Under these assumptions, Mahajan (2006) demonstrates that the regression function $g(x^*, z)$ can be nonparametrically identified. To see this, note that $\eta^2(z,v)$ is observable and note the following relations:

$$E(x|z, v) \equiv \eta_2(z, v) = (1 - \eta_1(z))\eta_2^*(z, v) + \eta_0(z)\left(1 - \eta_2^*(z, v)\right) E(y|z, v) = g(1, z)\,\eta_2^*(z, v) + g(0, z)\left(1 - \eta_2^*(z, v)\right) E(yx|z, v) = g(1, z)(1 - \eta_1(z))\,\eta_2^*(z, v) + g(0, z)\,\eta_0(z)\left(1 - \eta_2^*(z, v)\right).$$

Suppose v takes n_v values. For each z, $\eta_0(z)$, $\eta_1(z)$, $g(0, z)$, $g(1, z)$ and $\eta_2^*(z, v)$ are unknown. There are $4+n_v$ parameters. There are $3n_v$ equations. Therefore, as long as $n_v \geq 2$, all the parameters can possibly be identified. Intuitively, if $\eta_2^*(z, v)$ is known, the second moment condition $E(y|z, v)$ identifies $g(1, z)$ and $g(0, z)$. Information from the other moment conditions also allows one to identify both $\eta_1(z)$ and $\eta_0(z)$.
A constructive proof is given in Mahajan (2006) using the above three moment conditions. First of all, using the first moment condition

$$\eta_2^*(z, v) = \frac{\eta_2(z, v) - \eta_0(z)}{1 - \eta_0(z) - \eta_1(z)}.$$

If this is substituted into the next two moment conditions, then one can write

$$E(y|z, v) = g(0, z) + (g(1, z) - g(0, z))\frac{\eta_2(z, v) - \eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} = g(0, z) - \frac{(g(1, z) - g(0, z))\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} + \frac{g(1, z) - g(0, z)}{1 - \eta_0(z) - \eta_1(z)}\eta_2(z, v)$$

$$E(yx|z, v) = g(0, z)\eta_0(z) - [g(1, z)(1 - \eta_1(z)) - g(0, z)\eta_0(z)]\frac{\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} + \frac{[g(1, z)(1 - \eta_1(z)) - g(0, z)\eta_0(z)]}{1 - \eta_0(z) - \eta_1(z)}\eta_2(z, v) = -\frac{(g(1, z) - g(1, z))\eta_0(z)(1 - \eta_1(z))}{1 - \eta_0(z) - \eta_1(z)} + \frac{[g(1, z)(1 - \eta_1(z)) - g(0, z)\eta_0(z)]}{1 - \eta_0(z) - \eta_1(z)}\eta_2(z, v).$$

Mahajan (2006) suggests that, if one runs a regression of $E(y|z, v)$ on $\eta_2(z, v)$ and runs a regression of $E(yx|z, v)$ on $\eta_2(z, v)$, then one can recover the intercepts and the slope coefficients:

$$a = g(0, z) - \frac{(g(1, z) - g(0, z))\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} b = \frac{g(1, z) - g(0, z)}{1 - \eta_0(z) - \eta_1(z)} c = g(0, z)\eta_0(z) - [g(1, z)(1 - \eta_1(z)) - m(0, z)\eta_0(z)] \frac{\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} = - \frac{(g(1, z) - g(1, z))\eta_0(z)(1 - \eta_1(z))}{1 - \eta_0(z) - \eta_1(z)} d = \frac{[g(1, z)(1 - \eta_1(z)) - g(0, z)\eta_0(z)]}{1 - \eta_0(z) - \eta_1(z)}.$$

Therefore, one can write

$$a = m(0, z) - \eta_0(z) b \tag{4}$$

$$c = m(0, z)\eta_0(z) - d\eta_0(z) \tag{5}$$

and

$$c = -b(1 - \eta_1(z))\eta_0(z). \tag{6}$$

Equation (4) can be used to concentrate out $m(0, z)$. One can then substitute it into (5) and make use of (6) to write

$$(a + \eta_0(z) b)\eta_0(z) - d\eta_0(z) = -b(1 - \eta_1(z))\eta_0(z).$$

Then we can factor out $\eta_0(z)$ and rearrange:

$$1 - \eta_1(z) + \eta_0(z) = \frac{d - a}{b}. \tag{7}$$

Now we have two eqs. (6) and (7) in two unknowns $1 - \eta_1(z)$ and $\eta_0(z)$. Obviously the solutions to this quadratic system of equation is unique only up to an exchange between $1 - \eta_1(z)$ and $\eta_0(z)$. However, assumption 5.4 rules out one of these two possibilities and allows for point identification. Hence Mahajan (2006) demonstrates that the model is identified.
Mahajan (2006) further develops his identification strategy into a nonparametric estimator, and also develops a semiparametric estimator for a single index model.

6 Conclusion

Despite numerous articles that have been written on the topic of measurement errors in econometrics and statistics over the years, there are still many unresolved important qsts that are related to models of measurement errors. For example, the implications of measurement errors and data contaminations on complex structural models in labour economics and industrial organization are yet to be understood and studied. Recent empirical studies of precautionary saving and the permanent income hypothesis make use of panel data to address the issue of measurement errors (see, for example, Parker and Preston, 2005). Also, it is often the case that not all variables are validated in auxiliary data-sets. How to make use of partial information in validation studies is also an open qst.

See Also

- econometrics
- efficiency bounds

- linear models
- semiparametric estimation

The author acknowledges generous research support from the NSF (SES-0452143) and the Sloan Foundation.

Bibliography

Abrevaya, J., Hausman, J. and Scott-Morton, F. 1998. Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics* 87, 239–69.

Bollinger, C. 1998. Measurement error in the current population survey: a nonparametric look. *Journal of Labor Economics* 16, 576–94.

Bound, J., Brown, C., Duncan, G. and Rodgers, W. 1994. Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics* 12, 345–68.

Bound, J. and Krueger, A. 1991. The extent of measurement error in longitudinal earnings data: do two wrongs make a right. *Journal of Labor Economics* 12, 1–24.

Carroll, R. and Wand, M. 1991. Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society* 53, 573–85.

Chen, X., Hong, H. and Tamer, E. 2005. Measurement error models with auxiliary data. *Review of Economic Studies* 72, 343–66.

Chen, X., Hong, H. and Tarozzi, A. 2004. Semiparametric efficiency in GMM models nonclassical measurement errors. Working paper, Duke University and New York University.

Devereux, P. and Tripathi, G. 2005. Combining datasets to overcome selection caused by censoring and truncation in moment bases models. Working paper, University of Connecticut and UCLA.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Frish, R. 1934. *Statistical Confluence Study*. Oslo: University Institute of Economics.

Hahn, J. 1998. On the role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66, 315–32.

Hausman, J., Ichimura, H., Newey, W. and Powell, J. 1991. Measurement errors in polynomial regression models. *Journal of Econometrics* 50, 271–95.

Hirano, K., Imbens, G. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–89.

Hong, H. and Tamer, E. 2003. A simple estimator for nonlinear error in variable models. *Journal of Econometrics* 117, 1–19.

Imbens, G., Newey, W. and Ridder, G. 2005. Mean-squared-error calculations for average treatment effects. Working paper, Harvard University, MIT and USC.

Lee, L. and Sepanski, J. 1995. Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association* 90(429), 130–40.

Li, T. 2002. Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics* 110, 1–26.

Mahajan, A. 2006. Identification and estimation of regression models with misclassification. *Econometrica* 74, 631–65.

Manski, C. and Horowitz, J. 1995. Identification and robustness with contaminated and corrupted data. *Econometrica* 63, 281–302.

Molinari, F. 2005. Partial identification of probability distributions with misclassified data. Working paper, Cornell University.

Parker, J. and Preston, B. 2005. Precautionary savings and consumption fluctuations. *American Economic Review* 95, 1119–44.

Robins, J., Mark, S. and Newey, W. 1992. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 48, 479–95.

Schennach, S. 2004. Estimation of nonlinear models with measurement error. *Econometrica* 72, 33–75.

Sepanski, J. and Carroll, R. 1993. Semiparametric quasi-likelihood and variance estimation in measurement error models. *Journal of Econometrics* 58, 223–56.

Taupin, M. 2001. Semiparametric estimation in the nonlinear structural errors-in-variables model. *Annals of Statistics* 29, 66–93.

How to cite this article

Hong, Han. "measurement error models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000412> doi:10.1057/9780230226203.1076

The New Palgrave Dictionary of Economics Online

measurement, theory of

R. Duncan Luce and Louis Narens

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Physical measurement, as embodied in dimensional analysis, consists of interlocked, qualitative, ordered structures. Analogous approaches to behavioural science are outlined for sensory scaling such as loudness, utility of uncertain alternatives, and qualitative foundations of probability. In many cases, they are an ordered structure with a binary operation for combining elements and a Cartesian product where each factor affects the ordering of the attribute. Axioms sufficient for measurement – for the existence of a homomorphism onto the positive real numbers – are mentioned, and their uniqueness (scale type – for example, ratio, interval, ordinal) is formulated qualitatively in terms of the structure's symmetries (or automorphisms).

Keywords

additivity; Allais paradox; averaging; completeness; conditional probability; de Finetti, B.; extended sure-thing principle; independence; invariance; Kolmogorov, A. N.; mass measurement; measurement, theory of; monotonicity; preference reversals; probability; ratio scale; representation; scale of measurement; subjective expected utility; subjective probability; transitivity; unboundedness; unconditional probability

Article

Most mathematical sciences rest upon quantitative models, and the theory of measurement is devoted to making explicit the qualitative assumptions that underlie them. This is accomplished by first stating the qualitative assumptions – empirical laws of the most elementary sort – in axiomatic form and then showing that there are structure preserving mappings, often but not always isomorphisms, from the qualitative structure into a quantitative one. The set of such mappings forms what is called a ‘scale of measurement’.
A theory of the possible numerical scales plays an important role throughout measurement – and

therefore throughout science. Just as the qualitative assumptions of a class of structures narrowly determine the nature of the possible scales, so also the nature of the underlying scales greatly limits the possible qualitative structures that give rise to such scales. Two major themes of this entry reflect research results of the 1970s and 1980s: (*a*) the possible scales that are useful in science are necessarily very limited; (*b*) once a type of scale is selected (or assumed to exist) for a qualitative structure, then a great deal is known about that structure and its quantitative models. A third theme concerns applications of these ideas to the behavioural sciences, especially to utility theory and psychophysics from 1980 onward.

There are several general references to the axiomatic theory. Perhaps the most elementary and the one with the most examples is Roberts (1979). Pfanzagl (1968) and Krantz et al. (1971) are on a par, with the latter more comprehensive. Narens (1985), which is the mathematically most sophisticated, covers much of the basic material mentioned here. Later additions are: Luce et al. (1990), which has much in common with Narens (1985); Suppes et al. (1989), which is focused on geometric representations and probability generalizations; Narens (2007), which is a more narrowly focused introductory book with examples mainly from psychophysics; and Suppes (2002). Mostly, we cite only references not included in wither Krantz et al. (1971) or Narens (1985).

1 Axiomatizability

The qualitative situation is usually conceptualized as a relational structure $\mathcal{X} = (X, S_0, S_1, \ldots)$, where the S_0, S_1,… are finitary relations on X. The number of relations can be either finite or infinite, but in applications almost always finite. X is called the domain of the structure and the S_i its primitive relations. In most applications, S_0 will be some type of ordering relation that is usually written as $\succsim$. The following are some examples of qualitative structures used in measurement situations.

The first goes back to Helmholtz (see Section 4). It has for its domain a set X of objects with the properties like those of mass. There are two primitive relations. The first, $\succsim$, is a binary ordering according to mass (which may be determined, for example, by using an equal-arm pan balance so that $x \succsim y$ means that the pans either remain level or the one containing x drops). The second is a binary operation $\circ$, which formally is a ternary relation. For mass it is empirically defined as follows: if x and y are placed in the same pan and are exactly balanced by z, then we write $x \circ y \sim z$, where ~ means equivalence in the attribute. Other interpretations of the primitives of $(X, \succsim, \circ)$ can be found in the above references. Axiomatic treatments of the structure $(X, \succsim, \circ)$ are discussed in Section 4.

A second example is from economics. Suppose that $C_1,\ldots,C_n$ are sets each consisting of different amounts of a commodity, and $\succsim$ is a preference ordering exhibited by a person or an institution over the set of possible commodity bundles $C = \prod_i C_i$. $(C, \succsim)$ is called a *conjoint structure*, and axioms about it are given that among other things induce an ordering, $\succsim_i$, of an individual's preferences for the commodities associated with each component i.

A third example, due to *B*. de Finetti, has as its domain an algebra of subsets, called ‘events’, of some non-empty set Ω. The primitives of the structure consist of an ordering relation $\succsim$ of ‘at least as likely as’, the events Ω and $\varnothing$ and the set theoretical operations of union $\cup$, intersection $\cap$, and complementation $\neg$.

The relational structure

$$\mathcal{P} = (\mathcal{E}, \succsim, \Omega, \varnothing, \cup, \cap, \neg)$$
(1)

is intended to characterize qualitatively probability-like situations. The primitive $\succsim$ can arise from many different processes, depending upon the situation. In one, which is of considerable importance to Bayesian probability theorists and statisticians, $\succsim$ represents a person's ordering of events according to how likely they seem, using whatever basis he or she wishes in making the judgements.
In such a case, $\mathcal{P}$ is thought of as a *subjective or personal probability structure*. In another, $\succsim$ is an ordering of events based on some probability model for the situation (possibly one coupled with estimated relative frequencies), as in much of classical probability theory.

2 Ordered structure

2.1 Weak order, Dedekind completeness, and unboundedness

Two types of ‘quantitative’ representations have played a major role in science: systems of coordinate geometry and the real number system (the latter being the one-dimensional specialization of the former). Results about the former are in Suppes et al. (1989), but our focus here is the latter. The absolutely simplest case, included in all of the above examples, is the order-preserving representation ϕ of $(X, \succsim)$ into $(\mathbb{R}, \geq)$, where $\mathbb{R}$ denotes the real numbers. An immediate implication is that $\succsim$ must be transitive, reflexive, and connected (for all *x* and *y*, either $x \succsim y$ or $y \succsim x$). Such relations are given many different names including *weak order*. An antisymmetric weak order is called a total or *simple order*. There has been much empirical controversy about the transitivity of $\succsim$, with the most recent Bayesian analyses favouring transitivity of $\succ$ but not of ~ (Myung, Karabatsos and Iverson, 2005). Some doubt has been expressed about completeness. Nevertheless, most of the well-developed measurement-theoretic techniques assume both the completeness and transitivity of $\succsim$ as idealizations.
G. Cantor showed that for $(X, \succsim)$ to be so represented, necessary and sufficient conditions are that $\succsim$ be a weak order and that there be a finite or countable subset *Y* of *X* that is order dense in *X* (that is, for each $x \succ z$ there exists a *y* in *Y* such that $x \succsim y \succsim z$). For many purposes, this subset plays the same role as do the rational numbers within the system of real numbers.

In order for the representation to be onto either $(\mathbb{R}, \geq)$ $(\mathbb{R}^+, \geq)$, where $\mathbb{R}^+$ denotes the positive real numbers, which often happens in physical measurement, two additional conditions are necessary and sufficient: *Dedekind completeness* (each non-empty bounded subset of *X* has a least upper bound in *X*) and *unboundedness* (there is neither a least nor a greatest element).
In measurement axiomatizations, one usually does not postulate a countable, order-dense subset, but derives it from axioms that are intuitively more natural. For example, with a binary operation of combining objects, order density follows from a number of properties including an *Archimedean axiom*

which states in some fashion that no object is either infinitely larger than or infinitesimally close to another object. When the structure is Dedekind complete and the operation is monotonic, it is also Archimedean. Dedekind completeness and Archimedeaness are what logicians call ‘second-order axioms’, and in principle they are incapable of direct empirical verification.

The most fruitful and intensively examined measurement structures are those with a weak ordering $\succsim$ and an associative, positive binary operation $\circ$ that is strictly monotonic ($x \succsim y$ iff $x \circ z \succsim y \circ z$). They have been the basis of much physical measurement. However, for much of the 20th century they played little role in the behavioural and social sciences but, as seen in Sections 6 and 7, since the 1997s such operations have come to be useful. The development of a general non-associative and non-positive ($x \succsim x \circ y$ for some *x* and *y*) theory began in 1976, and it is moderately well understood in certain situations having many symmetries. (Technically, symmetries or automorphisms of the structure are isomorphic transformations of the structure onto itself.) This, and its specialization to associative structures, is the focus of Section 3.

2.2 Representations and scales

A key concept in the theory of measurement is that of a *representation*, which is defined to be a structure preserving map ϕ of the qualitative, weakly ordered relational structure $\mathcal{X}$ into a quantitative one, $\mathcal{R}$, in which the domain is a subset of the real numbers. Representations are either isomorphisms or homomorphisms. The latter are used in cases where equivalences play an important role (for example, conjoint structures where trade-offs between components are the essence of the matter), in which case equivalence classes of equivalent elements are assigned the same number. We say ϕ is a $\mathcal{R}$-*representation for* $\mathcal{X}$.

From 1960 to 1990, measurement theorists were largely focused on certain types of qualitative structures for which numerical representations exist. The questions faced are two. The first, the ‘existence’ problem, is to establish that the set of $\mathcal{R}$-representations is non-empty for $\mathcal{X}$. Cantor's conditions above establish existence of a numerical representation of any weak order. The second, the ‘uniqueness’, problem is to describe compactly the set of all $\mathcal{R}$-representations. Several examples are cited. Since 1990, the focus has been increasingly on applying these insights to behaviour. We cite aspects of utility theory, global psychophysics, and probability.

For the qualitative mass structure $\mathcal{X} = (X, \succsim, \circ)$ described previously, the qualitative representing structure is taken to be $\mathcal{R} = (\mathbb{R}^+, \geq, +)$ where $\geq$ and+have their usual meanings in $\mathbb{R}^+$. The set of $\mathcal{R}$-representations of $\mathcal{X}$ consist of all functions ϕ from *X* into $\mathbb{R}^+$ such that for each *x* and *y* in *X*,

1. (i) $x \succsim y$ iff $\phi(x) \geq \phi(y)$, and
2. (ii) $\phi(x \circ y) = \phi(x) + \phi(y)$.

Such a function is called a *homomorphism for* $\mathcal{X}$, and the set of all of them is called a *scale* (*for* $\mathcal{X}$). In addition to Helmholtz, others – including Hölder, Suppes, Luce and Marley, and Falmagne – have stated axioms about the primitives that are sufficient to show the existence of such homomorphisms and to show the following uniqueness theorem: any two homomorphisms ϕ and ψ are related by positive multiplication, that is, there is some real $r>0$ such that $\psi = r\phi$. In the language introduced by Stevens

(1946), such a form of measurement is said to form a ‘ratio scale’. For cases where is an operation (defined for all pairs), Alimov (1950) and Roberts and Luce (1968) gave necessary and sufficient conditions for such a representation. Such a complete characterization as this one is rather unusual in measurement; sufficient conditions are far more the norm. Often they entail structural assumptions, such as a solvability condition, as well as necessary ones.

Representations of the structure $\mathcal{C} = \langle \Pi_i C_i, \succsim \rangle$ of commodity bundles are usually taken in economics to be n-tuples $\langle \phi_1,\ldots,\bullet \phi_n \rangle$ of functions, where ϕ_i maps C_i into $\mathbb{R}^+$, such that for each x_i and y_i in C_i, i=1, …,•n,

$$(x_1, \ldots, x_n) \succsim (y_1, \ldots, y_n) \text{iff} \sum_i \varphi_i(x_i) \geq \sum_i \varphi_i(y_i). \tag{2}$$

In the measurement literature such a conjoint representation is called ‘additive’.

Debreu, Luce and Tukey, Scott, Tversky, and others gave axioms about $\mathcal{C}$ for which existence of an additive representation can be shown, and such that any two representations $\langle \varphi_1, \ldots, \varphi_n \rangle$ and $\langle \psi_1, \ldots, \psi_n \rangle$ are related by affine transformations of the form $\psi_I = r\phi_I + s_i$, i=1,…,•n, r>0. Note that r is common to all components. In Stevens’ nomenclature, the set of such representations ψ_i for each fixed i are said to form an ‘interval scale’.

In the example of the subjective probability structure, eq. (1), the usual sort of representation is a probability function P from $\mathcal{E}$ into [0, 1], such that, for all A, B in $\mathcal{E}$,

1. (i) $P(\Omega)$=1 and P($\varnothing$)=0,
2. (ii) $A \succsim B$ iff $P(A) \geq P(B)$, and
3. (iii) if $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$.

Unlike the previous two examples, here any two representations are identical, which scales Stevens called ‘absolute’. Such a scale might be appropriate for representing a qualitative structure describing a relative frequency approach to probability. However, for subjective probability, it is better to view P as being a representation of the bounded ratio scale $\{rP|r>0\}$ that is normalized by setting the bound, Ω, to be $1 = rP(\Omega)$.

A number of authors have given sufficient conditions in terms of the primitives for P to exist. Fine (1973) gave the first good, early summary of a variety of approaches to probability. Additional approaches to qualitative and subjective probability can be found in Narens (2008).

2.3 Interlocked measurement structures

A very common, and fundamentally important, feature of measurement is the existence of two or more

ways to manipulate the same attribute. Again, mass measurement is illustrative. The mass order $(X, \succsim)$ is determined as above. Mass can be manipulated in at least two ways by varying volumes and/or substances. Let $(V, \succsim', \circ_V)$ be a structure for combining volumes, where $\circ_V$ is a set of volumes and V is a strictly monotonic, positive, and associative operation over V, and let $X=V\times S$ be a structure of masses, where S is a set of homogeneous substances of various densities. (v, s) is interpreted as an object of volume v filled with substance s and that, therefore, has mass. By definition, $\circ_V$ is the operation on $V\times\{s\}$ such that $(v,s) \circ_V (v',s)=(v \circ_V v',s)$. The first manipulation is to vary $\succsim$ via volume concatenation of a single homogeneous material s, $(V\times\{s\}, \succsim, \circ_V)$. The second is to manipulate the conjoint trade-off between volumes and substances, $(V\times S, \succsim)$. Let m and m^* be the resulting representations of mass which, because they both preserve $\succsim$, must be strictly monotonically related. The ordering interlock alone is insufficient to develop measurement as was done in classical physics and as reflected in the familiar structure of physical units. Comparable developments are now beginning to appear in the behavioural and social sciences. The two structures must be interlocked beyond $\succsim$. Such interlocks are often types of distribution laws. In the mass case, the *distributive interlock* is: For $u, v\in V$ and $r, s\in S$,

$$(u, r) \sim (v, s) \text{ and } (u', r) \sim (v', s) \text{ imply } (u \circ_V u', r) \sim (u \circ_V v', s).$$

For much more detail, see Luce et al. (1990). Such laws are the source of the structure reflected in the units of physical measurement that are used and underlie dimensional analysis (Krantz et al., 1971; Luce et al., 1990; Narens, 2002).

Typically, one is able to use the two separate numerical representations to reduce the interlock to solving a functional equation. (A functional equation resembles a differential one in that its solutions are the unknown functions satisfying the equation. It is unlike a differential equation in that no derivatives are involved; rather, the equation relates the value of the function at several values of the independent variable. See Aczél (1966; 1987) for a general introduction and classical examples of functional equations. Some arising in the behavioural and social sciences were novel and have required the aid of specialists to solve.)

Behavioural examples of interlocked structures are cited in Sections 6 and 7.

2.4 Empirical usefulness of axiomatic treatments

One, seemingly under-appreciated, advantage of a measurement approach to some scientific questions is that it offers an alternative way of testing quantitative models other than attempting to fit the representation to data and to evaluate it by a measure of goodness of fit. Because representations, such as utility and subjective probability, in general have free parameters and often free functions, estimation is necessary. In contrast, the axioms underlying such representations are (usually) parameter free. Testing the axioms often makes clear the source of a problem, thereby giving insight into what must be

altered. Not everyone values the overall axiomatic (as compared with an analytic mathematical) approach to scientific questions; in particular, Anderson (1981, pp. 347–56) has sharply attacked it.

A familiar economic example arose in the theory of subjective expected utility (Fishburn, 1970; Savage, 1954). In its simplest form the domain is gambles of the form $x \circ_A y$, meaning that x is the consequence attached to the occurrence of the chance event A, whereas y is the consequence when the chance outcome is $\neg A$. The x and y may be pure consequences or may be themselves gambles, and the theory postulates a preference ordering $\succsim$ over the pure consequences and gambles constructed from pure consequences and gambles. Classical axiomatizations establish conditions on preferences over gambles so that there exists a probability measure P on the algebra of events, as in a probability structure, and a 'utility function' U over the gambles such that U preserves $\succsim$ and

$$U(x \circ_A y) = P(A)U(x) + [1 - P(A)]U(y). \tag{3}$$

A series of early empirical studies (for summaries see Allais and Hagen, 1979; Kahneman and Tversky, 1979) made clear that this representation, which can be readily defended on grounds of rationality, fails to describe human behaviour. Among its axioms, the one that appears to be the major source of difficulty is the 'extended sure-thing principle'. It may be stated as follows: For events A, B and C, with C disjoint from A and B,

$$x \circ_A y \succsim x \circ_B y \text{ iff } x \circ_{A \cup C} y \succsim x \circ_{B \cup C} y. \tag{4}$$

It is easy to verify that eq. (3) implies equation eq. (4), but people seem unwilling to abide by eq. (4). Any attempt at a descriptive theory must abandon it (see below).

2.5 Non-uniqueness of axiom systems

The isolation of properties in the axiomatic approach has an apparently happenstance quality because the choice of qualitative axioms is by no means uniquely determined by the representation. Any infinite structure has an infinity of equivalent axiom systems, and it is by no means clear why we select the ones that we do. It is entirely possible for a descriptive failure to be easily described in one axiomatization and to be totally obscure in another. Thus, some effort is spent on finding alternative but equivalent axiomatizations.

A related use of axiomatic methods, including the notion of scale (see Sections 2.2 and 3) is to study scientific meaningfulness, which is treated under meaningfulness and invariance.

3 Scale types

3.1 Classification

As was noted in the examples, scale type has to do with the nature of the set of maps from one numerical representation of a structure into all other equally good representations, in a particular numerical structure such as the multiplicative real numbers. For some fixed numerical structure $\mathfrak{R}$, a scale of the structure $\mathcal{A}$ is the collection of all $\mathfrak{R}$-representations of $\mathcal{A}$. Much the simplest case, the one to which we confine most of our attention, occurs when $\mathcal{A}$ is totally ordered, the domain of $\mathfrak{R}$ is either $\mathbb{R}$ or $\mathbb{R}^+$, and the $\mathfrak{R}$-representations are all onto the domain and so are isomorphisms. Such scales are then usually described in terms of the (mathematical) group of real transformations that take one representation into another. As Stevens (1946) noted, four distinct groups of transformations have appeared in physical measurement: any strictly monotonic function, any linear function $rx+s$, $r>0$, any similarity transformation rx, $r>0$, and the identity map. The corresponding scales are called *ordinal*, *interval*, *ratio*, and *absolute*. (Throughout this article, although not in all of the literature, ratio scales are assumed to be onto $\mathbb{R}^+$ thereby ruling out cases where an object maps to zero.)

A property of the first three scale types, called *homogeneity*, is that for each element x in the qualitative structure and each real number r in the domain of $\mathfrak{R}$, some representation maps x into r. Homogeneity, which is typical of physical measurement, plays an important role in formulating many physical laws. Two general questions are: what are the possible groups associated with homogeneous scales, and what are the general classes of structures that can are represented by homogeneous scales?

It is easiest to formulate answers to these questions in terms of automorphisms (=symmetries), that is, isomorphisms of the qualitative structure onto itself. The representations and the automorphisms of the structure are in one-to-one correspondence, because, if ϕ and ψ are two representations and juxtaposition denotes function composition, then $\psi^{-1}\phi$ is an automorphism of the structure, and if ϕ is a representation and α is an automorphism, then $\psi=\phi\alpha$ is a representation.

It is not difficult to see that homogeneity of a scale simply corresponds to there being an automorphism that takes any element of the domain of the structure into any other element. To make this more specific, for M a positive integer, $\mathcal{A}$ is said to be *M-point homogeneous* if and only if each strictly ordered set of M points can be mapped by an automorphism onto any other strictly ordered set of M points. A structure that fails to be homogeneous for $M=1$ is said to be 0-*point homogeneous*; one that is homogeneous for every positive integer M is said to be ∞-*point homogeneous*.

A second important feature of a scale is its degree of redundancy, formulated as follows: a scale is said to be *N-point unique*, where N is a non-negative integer if and only if for every two representations ϕ and ψ in the scale that agree at N distinct points, $\phi=\psi$. By this definition, ratio scales are 1-point unique, interval scales are 2-point unique, and absolute scales 0-point unique. Scales, such as ordinal ones, that take infinitely many points to determine a representation are said to be ∞-*point unique*. Equally, we speak of the structure being *N-point unique* if and only if every two automorphisms that agree at N distinct points are identical.

The abstract concept of scale type can be given in terms of these concepts. The *scale type* of $\mathcal{A}$ is the pair (M, N) such that M is the maximum degree of homogeneity and N is the minimum degree of uniqueness of $\mathcal{A}$. For the types of cases under consideration, it can be shown that $M\leq N$. Ratio scales are of type (1,

1) and interval scales of type (2, 2). Narens (1981a; 1981b) showed that the converses of both statements are true. And Alper (1987) showed that, if $M>0$ and $N<\infty$, then $N=1$ or 2. The group in the (1, 2) case consists of transformations of the form $rx+s$, where s is any real number and r is in some non-trivial, proper subgroup of the multiplicative group $(\mathbb{R}^+, \cdot)$, One example is $r=k^n$, where $k>0$ is fixed and n ranges over the integers. So a structure is homogeneous if and only if it is of type (1, 1), (1, 2), (2, 2), or (M, ∞). The (M, ∞) case is not fully understood. Ordinal scalable (∞, ∞) structures appear frequently in science, and a $(1, \infty)$ structure for threshold measurement appears in psychophysics. We focus here on the (1, 1), (1, 2), (2, 2) cases. For detailed references, see Luce et al. (1990), Narens (1985), or Narens (2007).

3.2 Unit representations of homogeneous concatenation structures

The next question is: which structures have scales of these types? Although the full answer is unknown, it is completely understood for ordered structures with binary operations. This is useful because, as was noted, the associative form of these operations plays a central role in much physical measurement and, as we shall see below, both associative and non-associative forms arise naturally in two distinct ways of interest to behavioural and social scientists.

Consider real concatenation structures of the form $\mathfrak{R} = (\mathbb{R}^+, \succeq, *')$ where $\geq$ has its usual meaning and we have replaced+by a general binary, numerical operation $*'$ that is strictly monotonic in each variable. The major result is that if $\mathfrak{R}$ satisfies $M>0$ and $N<\infty$ (a sufficient condition for finite N is that $*'$ be continuous – Luce and Narens, 1985) then the structure can be mapped canonically into an isomorphic one of the form $(\mathbb{R}^+, \succeq, *)$, with a function f from $\mathbb{R}^+$ onto $\mathbb{R}^+$ such that

1. (i) f is strictly increasing,
2. (ii) $f(x)/x$ is strictly decreasing, and
3. (iii) for all x, y in $\mathbb{R}^+$, $x*y=yf(x/y)$ (Cohen and Narens, 1979)

This type of canonical representation, which is called a *unit representation*, is invariant under the similarities of a ratio scale, that is, for each positive real r,

$$rx * ry = ryf(rx/ry) = ryf(x/y) = r(x * y).$$

The two most familiar examples of unit representations are ordinary additivity, for which $f(z)=1+z$ and so $x*y=x+y$, and bisymmetry, for which $f(z)=z^c$, $c\in(0, 1)$, and so $x*y=x^cy^{1-c}$. Situations where such representations arise are discussed later.

A simple invariance property of the function f corresponds to the three finite scale types (Luce and Narens, 1985). Consider the values of $\rho>0$ for which $f(x^\rho)=f(x)^\rho$ for all $x>0$. The structure is of scale

type (1, 1) if and only if $\rho = 1$; of type (1, 2) if and only if for some fixed $k>0$ and all integers n, $\rho = k^n$; and of type (2, 2) if and only if there are constants c and d in (0, 1) such that

$$f(z) = \begin{cases} z^c, & z \geq 1 \\ z^d, & z < 1. \end{cases}$$

If, as is the usual practice in the social sciences (see subjective expected utility, Section 6), but not in physics, the above representation is transformed by taking logarithms, it becomes a weighted additive form on $\mathbb{R}$:

$$x * y = \begin{cases} cx + (1-c)y, & x \geq y \\ dx + (1-d)y, & x < y. \end{cases}$$

That representation is called *dual bilinear* and the underlying structures are called dual bisymmetric (when $c=d$, the 'dual' is dropped). For references see Luce et al. (1990).

4 Axiomatization of concatenation structures

Given this understanding of the possible representations of homogeneous, finitely unique concatenation structures, it is natural to return to the classical question of axiomatizing the qualitative properties that lead to them. Until the 1970s, the only two cases that were understood axiomatically were those leading to additivity and averaging (see below). We now know more, but our knowledge remains incomplete.

4.1 Additive representations

The key mathematical result underlying extensive measurement, due to O. Hölder, states that when a group operation and a total ordering interlock so that the operation is strictly monotonic and is Archimedean in the sense that sufficient copies of any positive element (that is, any element greater than the identity element) will exceed any fixed element, then the group is isomorphic to an ordered subgroup of the additive real numbers. Basically, the theory of extensive measurement restricts itself to the positive subsemigroup of such a structure. Extensive structures can be shown to be of scale type (1, 1). Various generalizations involving partial operations (defined for only some pairs of objects) have been developed. (For a summary, see Krantz et al., 1971, chs 2, 3, and 5; Luce et al., 1990, ch. 19). Not only are these structures with partial operations more realistic, they are essential to an understanding of the partial additivity that arises in such cases as probability structures. They can be shown to be of scale type

(0, 1). Michell (1999) gives an alternative perspective on measurement in the behavioural sciences and a critique of axiomatic measurement approaches.

The representation theory for extensive structures not only asserts the existence of a numerical representation, but provides a systematic algorithm (involving the Archimedean property) for constructing one to any pre-assigned degree of accuracy. This construction, directly or indirectly, underlies the extensive scales used in practice.

The second classical case, due to J. Pfanzagl, leads to weighted average representations. The conditions are monotonicity of the operation, a form of solvability, an Archimedean condition, and bisymmetry, $(x \circ u) \circ (y \circ v) \sim (x \circ y) \circ (u \circ v)$ which replaces associativity. One method of developing these representations involves two steps: first, the bisymmetric operation is recoded as a conjoint one (see Section 5) as follows: $(u, v) \succsim (x, y)$ iff $u \circ v \succsim x \circ y$; and second, the conjoint structure is recoded as an extensive operation on one of its components. This reduces the proof of the representation theorem to that of extensive measurement, that is to Hölder's theorem, and so it too is constructive.

4.2 Non-additive representations

The most completely understood generalization of extensive structures, called positive concatenation structures or PCSs for short, simply drops the assumption of associativity. Narens and Luce (see Narens, 1985; Luce et al., 1990, ch. 19) showed that this was sufficient to get a numerical representation and that, under a slight restriction which has since been removed, the structure is 1-point unique, but not necessarily 1-point homogeneous. Indeed, Cohen and Narens (1979) showed that the automorphism group is an Archimedean ordered group and so is isomorphic to a subgroup of the additive real numbers; it is homogeneous only when the isomorphism is to the full group. As in the extensive case, one can use the Archimedean axiom to construct representations, but the general case is a good deal more complex than the extensive one and almost certainly requires computer assistance to be practical.

For Dedekind complete PCSs that map onto $\mathbb{R}^+$, a nice criterion for 1-point homogeneity is that, for each positive integer n and every x and y, then $n(x \circ y)=nx \circ ny$, where by definition $1x=x$ and $nx=(n-1)x \circ x$. The form of the representations of all such homogeneous representations was described earlier.

The remaining broad type of concatenation structures consists of those that are idempotent: that is, for all x, $x \circ x=x$. The following conditions have been shown to be sufficient for idempotent structures to have a numerical representation (Luce and Narens, 1985): $\circ$ is an operation that is strictly monotonic and satisfies an Archimedean condition (for differences) and a solvability condition that says for each x and y, there exist u and v such that $u \circ x=y=x \circ v$. If, in addition, such a structure is Dedekind complete, it can be shown that it is N-point unique with $N \leq 2$.

5 Axiomatization of conjoint structures

5.1 Binary structures

A second major class of measurement structures, widely familiar from both physics and the social sciences, comprises those involving two or more independent variables exhibiting a trade-off in the to-

be-measured dependent variable. Their commonness and importance in physics is illustrated by familiar physical relations among three basic attributes, such as kinetic energy=$mv^2/2$, where m is the mass and v the velocity of a moving body. Such conjoint trade-off structures are equally common in the behavioural and social sciences: preference between commodity bundles or between gambles; loudness of pure tones as a function of signal intensity and frequency; trade-off between delay and amount of a reward, and so on. Although there is some theory for more than two independent variables in the additive case, with the general representation given by eq. (2), for present purposes we confine attention to the two-variable case $(X \times S, \succsim)$. Michell (1990) gives detailed analyses of a number of behavioural examples.

As with concatenation structures, the simplest case to understand is the additive one in which the major non-structural properties are:

1. (i) *independence (monotonicity)*: if $(x, s) \succsim (x', s)$ holds for some s, then it holds for all s in S, and the parallel statement for the other component. Note that this property allows us to induce natural orderings, $\succsim_X$ on X and $\succsim_S$ on S;
2. (ii) *Thomsen condition*: if $(x, r)\sim(y, t)$ and $(y, s)\sim(z, r)$, then $(x, s)\sim(z, t)$; and
3. (iii) an *Archimedean condition* which says that if $\{x_i\}$ is a bounded sequence and if for some $r \succ s$ it satisfies $(x_i, r) \sim (x_{i+1}, s)$, then the sequence is finite. A similar statement holds for the other component.

These properties, together with some solvability in the structure, are sufficient to prove the existence of an interval scale, additive representation (for a summary of various results, see Krantz et al., 1971, chs 6, 7, and 9). The result has been generalized to non-additive representations by dropping the Thomsen condition, which leads to the existence of a non-additive numerical representation (Luce et al., 1990, chs 19 and 20). The basic strategy is to define an operation, say $\circ_X$ on component X, that captures the information embodied in the trade-off between components. The induced structure can be shown to consist of two PCSs pieced together at an element that acts like a natural zero of the concatenation structure. The results for PCSs are then used to construct the representation. As might be anticipated, $\circ_X$ is associative if and only if the conjoint structure satisfies the Thomsen condition.

6 Interlocked structures and applications to utility theory

6.1 Interlocked conjoint/extensive structures

The next more complex structure has the form $\mathcal{D} = (X \times S, \succsim, \circ)$, where $\circ$ is an operation on S. Such structures appear in the construction of the dimensional structure of physical units. The key qualitative axioms for physical measurement are that $(X \times S, \succsim)$ is a conjoint structure satisfying independence, $(S, \succsim_S, \circ)$ is an extensive structure, where $\succsim_S$ is the induced ordering on S, and $\mathcal{D}$ is distributive, that is,

$$\text{if } (x, p) \sim (y, q) \text{ and } (x, s) \sim (y, t), \text{ then } (x, p \circ s) \sim (y, q \circ t).$$

These axioms yield the following representation for $\mathcal{D}$: There exists a ratio scale S for the extensive structure $(S, \succeq_S, \circ)$ such that for each $\phi \in S$ there exists ψ from X into the positive reals such that for all x, y in X and all s, t, p, q in S, there exists a representation ϕ on S that is part of a multiplicative representation of the conjoint structure and additive over the concatenation operation: that is,

1. (i) $(x, s) \succeq (y, t)$ iff $\psi(x)\varphi(s) \geq \psi(y)\varphi(t)$, and
2. (ii) $\phi(p \circ q) = \phi(p) + \phi(q)$.

Discussions of how to construct the full algebra of physical dimensions using distributive structures and how to generalize these algebras to situations where there are no primitive associative operations are discussed in Luce et al. (1990) and Narens (2002).

6.2 Rationality assumptions in traditional utility theory

As was noted earlier, an extensive literature exists on preferences among uncertain alternatives, often called ‘gambles’. The first major theoretical development was the axiomatization of subjective expected utility (SEU), which is a representation satisfying, in the binary case, eq. (3). Although such axiomatizations are defensible theories in terms of principles of rationality, they fail as descriptions of human behaviour. The rationality axioms invoked are of three quite distinct types.

First, preference is assumed to be transitive. This assumption has been shown to fail in various empirical contexts (especially multifactor ones), with perhaps the most pervasive and still ill-understood example being the ‘preference reversal phenomenon’, discovered by Slovic and Lichtenstein and investigated extensively by others, most famously by Grether and Plott (1979), and several later references given in Luce (2000, pp. 39–45). Nevertheless, transitivity is the axiom that is least easy to give up. Even subjects who violate it are not inclined to defend their ‘errors’. A few attempts have been made to develop theories without it, but so far they are complex and have not received much empirical scrutiny (Bell, 1982; Fishburn, 1982; 1985; Suppes et al., 1989, chs 16 and 17).

The second type of rationality postulates so-called ‘accounting’ principles in which two gambles are asserted to be equivalent in preference because when analysed into their component outcomes they are seen to be identical. For example, if $x \circ_A y$ is a gamble and $(x \circ_A y) \circ_B y$ means that if the event B occurs first and then, independent of it, A occurs, then on accounting grounds $(x \circ_A y) \circ_B y \sim (x \circ_B y) \circ_A y$ is rational because, on both sides, x is the outcome when A and B both occur (although in opposite orders) and y otherwise. One of the first ‘paradoxes’ of utility theory, that of Allais, is a violation of an accounting equation which assumes that certain probability calculations also take place.

The third type of rationality condition is the extended sure-thing principle, eq. (4). Its failure, which occurs regularly in experiments, is substantially the ‘paradox’ pointed out earlier by Ellsberg. Subjects have insisted on the reasonableness of their violations of this principle (MacCrimmon, 1967).

6.3 Some generalizations of SEU

Kahneman and Tversky (1979) proposed a binary modification of the expected utility representation designed to accommodate the last two types of violations, and Tversky and Kahneman (1992) generalized it to general finite gambles. During the 1980s and 1990s a great deal of attention was devoted to this general class of so-called rank- (and sometimes sign-) dependent representations (RDU or RSDU) (also called cumulative and Choquet, 1953, representations). Summaries of this work, much of it of an axiomatic character for both risky cases, where probabilities are assumed known, and uncertain cases, where a subjective probability function is constructed, can be found in Quiggin (1993) and Luce (2000). These developments rests very heavily on modifying the distribution laws that are assumed. A far more general survey of utility theory, covering many aspects of it from an economic but not primarily an axiomatic measurement-theoretic perspective, is Barberà, Hammond and Seidl (1998; 2004).

To return to an axiomatic approach, suppose in what follows that $x_1 \succsim x_2 \succsim \cdots \succsim x_n$ and their associated event partition is (E_1,•E_2,…,•E_n). Define $E(i) = \cup_{j=1}^{i} E_j$. The class of RDU representations involve proving from the axioms the existence of an order-preserving, utility function U over pure consequences and gambles and, in general, non-additive weighting function S over the chance events such that

$$U(x_1, E_1; x_2, E_2; \ldots; x_n, E_n) = \sum_{i=1}^{n} U(x_i)\,[S(E_i \cup E(i-1)) - S(E(i-1))]. \tag{5}$$

Note that the weighting function is essentially the incremental impact of adding E_i to $E(i-1)$. When S is finitely additive, that is, for disjoint A and B, $S(A \cup B)=S(A)+S(B)$, then eq. (5) reduces to subjective expected utility (SEU).

If there is a unique consequence e, sometimes called a reference level and sometimes taken to be no change from the status quo, then the consequences and gambles can be partitioned into gains, where $x_i \succsim e$, and the remainder, losses. In such cases, usually it follows from the assumptions made that $U(e)=0$ and, usually, the weighting functions are sign dependent (that is, their form depends on whether their consequences are positive with respect to e or negative). Also, the RSDU representation includes cumulative prospect theory (Tversky and Kahneman, 1992) as a special case having added restrictions on both U and S.

Other interesting developments involving different patterns of weighting are cited in Luce (2000).

A great deal of attention has been paid to issues of accounting for empirical phenomena discovered over the years that have discredited SEU and EU as descriptive models of human behaviour. For some summaries see Luce (2000) and Marley and Luce (2005). M.H. Birnbaum (numerous citations of his articles appear in the last reference) has discovered experimental designs that discredit a major feature of eq. (5) called *coalescing* or, equally, *event splitting*: Suppose $x_k=x_k=y$, then

$$(x_1, E_1; \ldots; y, E_k; y, E_{k+1}; \ldots; x_n, E_n) \sim (x_1, E_1; \ldots; y, E_k \cup E_{k+1}; \ldots; x_n, E_n). \quad (6)$$

The left-hand side of eq. (6) is called ‘split’ because y is attached to each of two events, E_k and E_{k+1}. The right-hand side is called ‘coalesced’ because y is attached to the single coalesced event $E_k \cup E_{k+1}$. Birnbaum has vividly demonstrated that experimental subjects often fail to split gambles in ways that help facilitate rational decisions. The other direction, coalescing, is effortless because no choice is involved. Indeed, Birnbaum (2007) has shown that splitting the branch (x_1, E_1), which has the best consequence, x_1, enhances the apparent worth of a gamble, whereas splitting (x_n, E_n), the branch with the poorest consequence, diminishes it. Long ago, he proposed a modified representation, called TAX, because it ‘taxes’ the poorest consequence in favour of the best one, which accommodates many empirical phenomena, including this one, but neither he nor anyone else has offered a measurement axiomatization of TAX. This remains an open problem.

6.4 Joint receipt

Beginning in 1990, Luce and collaborators have investigated an operation $\oplus$ of joint receipt in gambling structures and ways that it may interlock with gambling structures. Its interpretation is suggested by its name, having two goods at once which, because $\oplus$ is assumed to be associative and commutative, can be extended to any finite number of goods. Several possible interlocking laws have been studied, and improved axiomatizations involving them have been given for a number of classical representations (for a summary, see Luce, 2000). The representation that has arisen naturally is called p-additive (so named, at the suggestion of A.J. Marley, because it is the only polynomial form that can be transformed into an additive one), namely, for some real δ,

$$U(x \oplus y) = U(x) + U(y) + \delta U(x) U(y).$$

(By rescaling U there is no loss of generality in assuming that δ is either –1, 0, or 1.)

6.5 Lack of idempotence and the utility of gambling

A feature of very many utility models, in particular, of all RDU or RSDU ones, is *idempotence*:

$$(x, E_1; \ldots; x, E_i; \ldots; x, E_n) \sim x.$$

Among other things, this has been thought to be a way to connect gambles to pure consequences, but that feature is redundant with the certainty principle $(x, E(n)) \sim x$. Further, if there is an inherent utility or disutility to risk or gambling, as widespread behaviour suggests there is – witness Las Vegas and mountain climbing – violations of idempotence assess it. Luce and Marley (2000) proposed partitioning a gamble g into a pure consequence, called a *kernel equivalent of g*, $KE(g)$, with the joint receipt, $\oplus$, of its unrewarded event structure $(e, E_1; \ldots; e, E_i; \ldots; e, E_n)$, which is called an *element of chance*. Although they found properties of such a decomposition based on the assumption that utility is additive over joint receipt, $\oplus$, they did not discover much about the form of the utility of an element of chance. In the case of risk, further work has led to a detailed axiomatic formulation of that leads either to EU plus a Shannon entropy term, or to a linear weighted form plus entropy of some degree different from 1. In the case of uncertainty, the form for elements of risk is much less restrictive (Luce et al., 2008a; 2008b). This risky form was first arrived at by Meginniss (1976) using a non-axiomatic approach. Because of the symmetry of entropy, this representation is unable to account for Birnbaum's differential event splitting data. This approach needs much more work.

7 Other applications of behavioural interest

7.1 A psychophysical one

A modified version of one of the RDU axiomatizations has been reinterpreted as a theory of global psychophysics, meaning that the focus is on the full dynamic range of intensity dimensions (for example, in audition the range 5–130•dB SPL; contemporary IRBs restrict the top of the range to 85•dB), not just local ranges as in discrimination studies. An example of the primitives are sound intensities x and y to the left and right ears, respectively, denoted (x, u), about which the respondent makes loudness judgments. Given two such stimuli, (x, x) and (y, y), $x>y$, and a positive number p, the respondent also can be requested to judge which stimulus (z, z) makes the subjective 'interval' from (y, y) to (z, z) seem to be p times the 'interval' from (y, y) to (x, x). The data are z, which we may denote in operator notation as $(x, x) \square p(y, y) := (z, z)$. Luce (2002; 2004) (for a summary of theory, tests, and references, see Luce and Steingrimsson, 2006) provided testable axioms, it is shown that there is a real valued mapping ψ, called a psychophysical function, and a numerical distortion function W such that

$$\Psi(x, u) = \Psi(x, 0) + \Psi(0, u) + \delta\Psi(x, 0)\Psi(0, u) \; (\delta \geq 0), \tag{7}$$

$$W(p) = \frac{\Psi[(x, x) \circ_p (y, y)] - \Psi(y, y)}{\Psi(x, x) - \Psi(y, y)} (x > y \geq 0). \quad (8)$$

The axioms have been empirically tested by Steingrimsson and Luce in four papers. The 2005a focused on each structure, the conjoint one and the operator; the 2005b focused on the interlocks between them for audition. The results are supportive of the theory. Possible mathematical forms for Ψ and W have been reduced to testable conditions that, with one exception (the cases where $\delta \neq 0$, $\Psi(x, 0)$ and $\Psi(0, x)$ are both power functions but with different exponents), were evaluated with considerable, but not perfect, support, for power functions (2006; 2007). Narens (1996) earlier proposed a closely related theory that included an axiom that forced $W(1)=1$. Empirical data of Ellermeier and Faulhammer (2000) and Zimmer (2005) soundly rejected the joint hypothesis that W is a power function with $W(1)=1$. Subsequent theory and experiments found considerable support for power functions with $W(1) \neq 1$.

7.2 Foundations of probability

Today, the usual approach to probability theory is the classical one due to Kolmogorov (1933). It assumes that probability is a σ-additive (the countable extension of finite additivity) measure function P with a sure event having probability 1. It defines the important concepts of independence and conditional probability in terms of P.

There are many objections to this approach as a foundation for probability. A summary of most of them can be found in Fine (1973) and Narens (2008). In particular, independence and conditional probability appear to be more basic concepts than unconditional probability: for example, one often needs to know the independence of events in order to estimate probabilities. Also, in most empirical situations one cannot exactly pin down the probabilities: that is, there are many probability functions consistent with the data. This suggests that in such situations the underlying probabilistic concept should be a family of probability functions instead of a single probability function. Obviously, with many consistent probability functions explaining the data, the Kolmogorov method of defining independence by $P(A \cap B) = P(A)P(B)$ really does not work. These and other difficulties disappear with measurement-theoretic approaches to probability (for example, see Krantz et al., 1971; Fine, 1973; Narens, 1985; 2008). The qualitative approach provides richer and more flexible methods than Kolmogorov's for formulating and investigating the foundations of probability.

Both the Kolmogorov and the measurement-theoretic approaches assume an event space that is a Boolean algebra of subsets. This assumption works for most applications in science and is routinely assumed in theoretical and empirical studies of subjective probability. A major exception to it is quantum mechanics, where a different event space is needed (von Neumann, 1995).

It is well-known that Boolean algebras of events correspond to the classical propositional calculus of logic. The classical propositional calculus captures deductions for propositions that are either true or false. It is not adequate for capturing various concepts of ‘vagueness’, ‘ambiguity’, or ‘incompleteness based on lack of knowledge’. For these, logicians use nonclassical propositional calculi. In general,

these nonclassical calculi cannot be interpreted as the classical propositional calculus with ‘true’ and ‘false’ replaced with probabilities. It is plausible that some of the just-mentioned concepts are relevant to how individuals make judgements and decisions. Their incorporation into formal descriptions of behaviour requires the event space to be changed from the usual algebra of events used in the Kolmogorov approach to probability to a different kind of event space. This issue and proposals for alternative event spaces are discussed in detail in Narens (2008).

In summary, the Kolmogorov approach to probability is flawed at a foundational level and is too narrow to account for many important scientific phenomena. The measurement-theoretic approach is one alternative for providing a better foundation and generalizations for the kind of probability theory described by Kolmogorov. One should also consider the possibility of developing probabilistic theories for event spaces different from algebras of events, especially for phenomena that fall outside of usual forms of observation, including various phenomena arising from mentation.

See Also

- expected utility hypothesis
- meaningfulness and invariance
- measurement
- non-expected utility theory
- prospect theory
- Savage's subjective expected utility model
- utility

Bibliography

Aczél, J. 1966. *Lectures on Functional Equations and their Applications*. New York/London: Academic Press.

Aczél, J. 1987. *A Short Course on Functional Equations: Based upon Recent Applications to the Social and Behavioral Sciences*. Dordrecht: Reidel.

Alimov, N.G. 1950. On ordered semigroups. *Izvestia Akademii Nauk SSSR, Serija Mat*. 14, 569–76.

Allais, M. and Hagen, O. 1979. *Expected Utility Hypotheses and the Allais Paradox*. Dordrecht: Reidel.

Alper, T.M. 1987. A classification of all order-preserving homeomorphism groups that satisfy uniqueness. *Journal of Mathematical Psychology* 31, 135–54.

Anderson, N.H. 1981. *Foundations of Information Integration Theory*. New York: Academic.

Barberà, S., Hammond, P. and Seidl, C. 1998, 2004. *Handbook of Utility Theory. Vol. I: Principles; Vol.*

II: Extensions. Boston: Kluwer.

Bell, D. 1982. Regret in decision making under uncertainty. *Operations Research* 30, 961–81.

Birnbaum, M.H. 2007. Tests of branch splitting and branch-splitting independence in Allais paradoxes with positive and mixed consequences. *Organizational Behavior and Human Decision Processes* 102, 154–73.

Choquet, G. 1953. Theory of capacities. *Annals Institut Fourier* 5, 131–295.

Cohen, M. and Narens, L. 1979. Fundamental unit structures: a theory of ratio scalability. *Journal of Mathematical Psychology* 20, 193–232.

Ellermeier, W. and Faulhammer, G. 2000. Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception and Psychophysics* 62, 1505–11.

Fine, T. 1973. *Theories of Probability*. New York: Academic.

Fishburn, P.C. 1970. *Utility Theory for Decision Making*. New York: Wiley.

Fishburn, P.C. 1982. Nontransitive measurable utility. *Journal of Mathematical Psychology* 26, 31–67.

Fishburn, P.C. 1985. Nontransitive preference theory and the preference reversal phenomenon. *International Review of Economics and Business* 32, 39–50.

Grether, D.M. and Plott, C.R. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69, 623–38.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Kolmogorov, A. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. New York: Chelsea, 1946.

Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A. 1971. *Foundations of Measurement*, vol. 1. New York: Academic. Repr. New York: Dover, 2007.

Luce, R.D. 2000. *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*. Mahwah, NJ: Erlbaum. Errata: see Luce's web page at http://www.socsci.uci.edu.

Luce, R.D. 2002. A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review* 109, 520–32.

Luce, R.D. 2004. Symmetric and asymmetric matching of joint presentations. *Psychological Review* 111, 446–54.

Luce, R.D., Krantz, D.H., Suppes, P. and Tversky, A. 1990. *Foundations of Measurement*, vol. 3. New York: Academic. Repr. New York: Dover, 2007.

Luce, R.D. and Marley, A.A.J. 2000. On elements of chance. *Theory and Decision* 49, 97–126.

Luce, R.D. and Narens, L. 1985. Classification of concatenation measurement structures according to scale type. *Journal of Mathematical Psychology* 29, 1–72.

Luce, R.D., Ng, C.T., Aczél, J. and Marley, A.A.J. 2008a. Utility of gambling I: entropy-modified linear weighted utility. *Economic Theory*.

Luce, R.D., Ng, C.T., Aczél, J. and Marley, A.A.J. 2008b. Utility of gambling II:. risk, paradoxes, data. *Economic Theory*.

Luce, R.D. and Steingrimsson, R. 2006. Global psychophysical judgments of intensity: summary of a theory and experiments. In *Measurement and Representations of Sensations*, ed. H. Colonious and E. Dzharfov. Mahwah, NJ: Lawrence Erlbaum Associates.

MacCrimmon, K.R. 1967. Descriptive and normative implications of the decision theory postulates. In *Risk and Uncertainty*, ed. K. Borch and J. Mossin. New York: Macmillan.

Marley, A.A.J. and Luce, R.D. 2005. Independence properties vis-à-vis several utility representations. *Theory and Decision* 58, 77–143.

Meginniss, J.R. 1976. A new class of symmetric utility rules for gambles, subjective marginal probability functions, and a generalized Bayes' rule. *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 471–6.

Michell, J. 1990. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Michell, J. 1999. *Measurement in Psychology: Critical History of a Methodological Concept*. Cambridge: Cambridge University Press.

Myung, J.I., Karabatsos, G. and Iverson, G.J. 2005. A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology* 49, 205–25.

Narens, L. 1981a. A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision* 13, 1–70.

Narens, L. 1981b. On the scales of measurement. *Journal of Mathematical Psychology* 24, 249–75.

Narens, L. 1985. *Abstract Measurement Theory*. Cambridge, MA: MIT Press.

Narens, L. 1996. A theory of ratio magnitude estimation. *Journal of Mathematical Psychology* 40, 109–29.

Narens, L. 2002. *Theories of Meaningfulness*. Mahwah, NJ: Lawrence Erlbaum Associates.

Narens, L. 2007. *Introduction to the Theories of Measurement and Meaningfulness and the Use of Invariance in Science*. Mahwah, NJ: Lawrence Erlbaum Associates.

Narens, L. 2008. *Theories of Probability: An Examination of Logical and Qualitative Foundations*. London: World Scientific.

Pfanzagl, J. 1968. *Theory of Measurement*. New York: Wiley. 2nd edn, Vienna: Physica, 1971.

Quiggin, J. 1993. *Generalized Expected Utility Theory: The Rank-Dependent Model*. Boston: Kluwer.

Roberts, F.S. 1979. *Measurement Theory*. Reading, MA: Addison-Wesley.

Roberts, F.S. and Luce, R.D. 1968. Axiomatic thermodynamics and extensive measurement. *Synthese* 18, 311–26.

Savage, L.J. 1954. *The Foundations of Probability*. New York: Wiley.

Steingrimsson, R. and Luce, R.D. 2005a. Evaluating a model of global psychophysical judgments I: behavioral properties of summations and productions. *Journal of Mathematical Psychology* 49, 290–306.

Steingrimsson, R. and Luce, R.D. 2005b. Evaluating a model of global psychophysical judgments II: behavioral properties linking summations and productions. *Journal of Mathematical Psychology* 49, 308–19.

Steingrimsson, R. and Luce, R.D. 2006. Empirical Evaluation of a model of global psychophysical judgments III: a form for the psychophysical and perceptual filtering. *Journal of Mathematical Psychology* 50, 15–29.

Steingrimsson, R. and Luce, R.D. 2007. Empirical Evaluation of a model of global psychophysical

judgments IV: Forms for the weighting function. *Journal of Mathematical Psychology* 51, 29–44.

Stevens, S.S. 1946. On the theory of scales of measurement. *Science* 103, 677–80.

Suppes, P. 2002. *Representation and Invariance of Scientific Structures*. Stanford, CA: CSLI publications.

Suppes, P., Krantz, D.H., Luce, R.D. and Tversky, A. 1989. *Foundations of Measurement*, vol. 2. New York: Academic. Repr. New York: Dover, 2007.

Tversky, A. and Kahneman, D. 1992. Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5, 297–323.

von Neumann, J. 1995. *Mathematical Foundations of Quantum Mechanics*. Princeton, NJ: Princeton University Press.

Zimmer, K. 2005. Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics* 67, 569–79.

How to cite this article

Luce, R. Duncan and Louis Narens. "measurement, theory of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000128> doi:10.1057/9780230226203.1077

The New Palgrave Dictionary of Economics Online

measurement

Marcel Boumans
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Measurement theory takes measurement as the assignment of numbers to properties of an empirical system so that a homomorphism between the system and a numerical system is established. To avoid operationalism, two approaches can be distinguished. In the axiomatic approach it is asserted that if the empirical system satisfies a certain set of axioms such a homomorphism can be constructed. In the empirical approach, empirical adequacy is established by aiming at accuracy, precision and standardization. Precision is achieved by least-squares-errors methods, accuracy by calibration and standardization by the involvement of independent theoretical and empirical studies.

Keywords

axiomatic index theory; axiomatic theory; calibration; ceteris paribus; Fisher, I.; functional equation theory; index numbers; measurement error models; measurement theory; model theory of measurement; operationalism; passive observations; price indexes; representation theorems; representational theory of measurement; structural parameters; uniqueness theorems

Article

The dominant measurement theory is the representational theory of measurement (RTM), which takes measurement as a process of assigning numbers to attributes of the empirical world in such a way that the relevant qualitative empirical relations among these attributes are reflected in the numbers themselves as well as in important properties of the number system.
The RTM defines measurement set-theoretically. Given a set of empirical relations $\mathbf{R} = \{R_1, \ldots, R_m\}$ on a set of extra-mathematical entities **X** and a set of numerical relations $\mathbf{P} = \{P_1, \ldots, P_m\}$ on the set of numbers **N** (in general a subset of the set of real numbers), a function ϕ from **X** into **N** takes each R_i into P_i, $i = 1, \ldots, m$, provided that the elements $x, y, \ldots,$ in **X** stand in relation R_i if and only if the

corresponding numbers $\varphi(x), \varphi(y), \ldots,$ stand in relation P_i. In other words, measurement is conceived of as establishing homomorphisms (also called scales) from empirical relational structures $\{\mathbf{X}, \mathbf{R}\}$ into numerical relational structures $\{\mathbf{N}, \mathbf{P}\}$. A numerical relational structure representing an empirical relational structure is also called a model, therefore the RTM is sometimes called the model theory of measurement.

The problem is that when the requirements for choosing a representation or model are not further qualified, it can easily lead to an operationalist position, which is most explicitly expressed by Stevens (1959, p. 19): ‘Measurement is the assignment of numerals to objects or events according to rule – any rule.’ A model should meet certain criteria to be considered homomorphic to an empirical relational structure. In economics, there are two different foundational approaches, an axiomatic and an empirical approach (Boumans, 2007).

Axiomatic theory

The axiomatic theory is most comprehensively presented in Krantz et al. (1971–90). According to this literature the foundations of measurement are established by axiomatization. The analysis into the foundations of measurement involves, for any particular empirical relation structure, the formulation of a set of axioms that is sufficient to establish two types of theorems, a representation theorem and a uniqueness theorem.

A representation theorem asserts that if a given relational structure satisfies certain axioms, then a homomorphism into a certain numerical relational structure can be constructed. A uniqueness theorem sets forth the permissible transformations $\varphi \to \varphi'$. A transformation $\varphi \to \varphi'$ is permissible if and only if ϕ and ϕ' are both homomorphisms of $\{\mathbf{X}, \mathbf{R}\}$ into the same numerical structure $\{\mathbf{N}, \mathbf{P}\}$.

Probably the first example of the axiomatic approach in economics is Frisch (1926), in which three axioms define utility as a quantity. The work more often referred to as the one that introduced the axiomatic approach to economics, however, is Von Neumann and Morgenstern (1944). They required the transformation $\varphi: \mathbf{X} \to \mathbf{N}$ to be order-preserving: $x > y$ implies $\varphi(x) > \varphi(y)$, and linear:

$$\varphi(\alpha x + (1-\alpha)y) = \alpha\varphi(x) + (1-\alpha)\varphi(y), \text{ where } \alpha \in (0, 1).$$

Another field in economics in which the axiomatic approach has been influential is the axiomatic index theory. This theory originates from Fisher's work on index numbers (1911; 1922). Fisher evaluated in a systematic manner a very large number of indices with respect to a number of criteria. These criteria were called ‘tests’. Fisher himself did not expect that it would be possible to devise an index number that would satisfy all these tests. Moreover, Frisch (1930) proved the impossibility of maintaining a certain set of Fisher's tests simultaneously. It is, however, Eichhorn and Voeller (1976) who provide a definite evaluation of Fisher's tests by their axiomatic approach.

Eichhorn and Voeller (1976) look systematically at the inconsistencies between various tests (and how

to prove such inconsistencies) by means of the functional equation theory. Functional equation theory is transferred into index theory if the price index is defined as a positive function $P(\mathbf{p}_s, \mathbf{x}_s, \mathbf{p}_t, \mathbf{x}_t)$ that satisfies a number of axioms, where **p** is a price vector and **x** a commodity vector, and the subscripts are time indices. These axioms do not, however, determine a unique form of the price index function. Several additional tests are needed for assessing the quality of a potential price index. Both axioms and tests are formalized as functional equations. When the axioms are formalized as functional equations, inconsistency theorems can then be proven by showing that for the relevant combinations of functional equations, the solution space is empty.

In current axiomatic index theory, axioms specify mathematical properties that are essential or desirable for an index formula. One of the problems of axiomatic index theory is the impossibility of simultaneously satisfying all axioms. In practice, however, a universally applicable solution to this problem is not necessary. The specifics of the problem at hand, including the purpose of the index and the characteristics of the data, determine the relative merits of the possible attributes of the index formula.

Empirical approach

Relation-rich structures, in contrast to object-rich structures, do not lend themselves to axiomatization. This does not mean, however, that measurement is impossible, but that a representation should, beside theoretical requirements, also satisfy empirical criteria. Moreover, economic measurements are often developed for purposes of economic policy; so, representations should also satisfy criteria of applicability. For example, a national account system should be a consistent structure of interdependent definitions, enabling uniform analysis and comparison of various economic phenomena.

To understand empirical measurement approaches, let us consider the problem of measuring a property *x* of an economic phenomenon. $y_i (i = 1, \ldots, n)$ are repeated observations to be used to determine value *x*. Each observation involves an observational error, ε_i. This error term, representing noise, reflects the operation of many different, sometimes unknown, background conditions, indicated by *B*:

$$y_i = f(x, B_i) = f(x, 0) + \varepsilon_i (i = 1, \ldots n) \quad (1)$$

Now, accuracy is obtained by reducing noise as much as possible. One way of obtaining accuracy is by taking care that the background conditions *B* are held constant, in other words, that *ceteris paribus* conditions are imposed. To show this, eq. (1) is rewritten to expresses how *x* and possible other conditions (*B*) influence the observations:

$$\Delta y = f_x \Delta x + f_B \Delta B = f_x \Delta x + \Delta \varepsilon \quad (2)$$

Then, imposing *ceteris paribus* conditions: $\Delta B \approx 0$ reduces noise.
However, *ceteris paribus* conditions imply full control of the circumstances and complete knowledge of all potential influence quantities. However, in economics we have often to deal with open systems, in which full control is not feasible. As a result, accuracy has to be obtained by modelling in a specific way. To measure *x* a model *M* has to be specified of which the values of the observations y_i functions as input and the output estimate $\hat{x}$ as measurement result: $\hat{x} = M[y_i; \alpha]$, where α denotes the parameter set of the model. If one substitute eq. (1) into model *M*, one can derive that, assuming that *M* is a linear operator (usually the case):

$$\hat{x} = M[f(x) + \varepsilon; \alpha] = M_x[x; \alpha] + M_\varepsilon[\varepsilon; \alpha]. \quad (3)$$

A necessary condition for the measurement of *x* is that a model *M* must entail a representation of the measurand, M_x, and a representation of the environment of the measurand, M_ε.
The performance of a model built for measuring purposes is described by the terms accuracy and precision. In metrology, accuracy is defined as the statement about the closeness of the model's outcome to a value declared as the standard. Precision is a statement about the spread of the estimated measurement errors. The usual procedure to attain precision is by minimizing the variance of errors. The procedure to obtain accuracy is calibration, which is the establishment of the relationship between values indicated by a model and the corresponding values realized by standards. So, we can split the measurement error in three parts:

$$\hat{\varepsilon} = \hat{x} - x = M_\varepsilon + (M_x - S) + (S - x) \quad (4)$$

where *S* represents a standard value. The error term M_ε is reduced as much as possible by aiming at precision. $(M_x - S)$ is the part of the error term that is reduced by calibration. The reduction of the last term $(S - x)$ is called standardization and is dealt with by finding an invariant structure representing the measurement system.
Attempting to find these invariant structures, we have to deal with the so-called problem of passive observation: it is not possible to identify the reason for a disturbing influence, say *z*, being negligible, $f_z \Delta z \approx 0$. We cannot distinguish whether its potential influence is very small, $f_z \approx 0$, or whether the factual variation of this quantity over the period under consideration is too small, $\Delta z \approx 0$. The variation of *z* is determined by other relationships within the economic system. In some cases, a virtually dormant quantity may become active because of changes in the economic system elsewhere. Each found

empirical relationship is a representation of a specific data-set. So, for each data-set it is not clear whether potential influences are negligible or only dormant. This is what Haavelmo (1944) called the problem of autonomy. Some of the empirical found relations have very little ‘autonomy’ because their existence depends upon the simultaneous fulfilment of a great many other relations. Autonomous relations are those relations that could be expected to have a great degree of invariance with respect to various changes in the economic system.

This problem of autonomy is dealt with by the following modelling strategy: when a relationship appears to be inaccurate, this is an indication that a potential factor is omitted. As long as the resulting relationship is inaccurate, potential relevant factors should be added. The expectation is that this strategy will result in the fulfilment of two requirements: (*a*) the resulting model captures a complete list of factors that exert large and systematic influences and (*b*) all remaining influences can be treated as a small noise component. The problem of passive observations is solved by accumulation of data-sets: the expectation is that we converge bit by bit to a closer approximation to the complete model, as all the most important factors reveal their influence. This strategy, however, is not applicable in cases when there are influences that we cannot measure, proxy or control for, but which exert a large and systematic influence on the outcomes.

A very influential paper in macroeconometrics (Lucas, 1976) showed that the estimated so-called structural parameters (α) achieved by the above strategy are not invariant under changes of policy rules. Policy-invariant parameters should be obtained in an alternative way. Either they could be supplied from independent microeconometric studies, accounting identities or institutional facts, or they are chosen to secure a good match between a selected set of characteristics of the actual observed time series and those of the simulated model output. These alternative ways of obtaining parameter values are all covered by the label calibration. It is important that, whatever the source, the facts being used for calibration should be as stable as possible. An important result of this calibration strategy is that for accurate measurement it is no longer required for representations to be homomorphic to an empirical relational structure.

See Also

- calibration
- ceteris paribus
- econometrics
- meaningfulness and invariance
- measurement error models
- measurement, theory of

Bibliography

Boumans, M. 2007. *Measurement in Economics: A Handbook*. Elsevier.

Eichhorn, W. and Voeller, J. 1976. *Theory of the Price Index*. Berlin: Springer.

Fisher, I. 1911. *The Purchasing Power of Money*. New York: Kelley, 1963.

Fisher, I. 1922. *The Making of Index Number*. New York: Kelley, 1967.

Frisch, R. 1930. Necessary and sufficient conditions regarding the form of an index number which shall meet certain of Fisher's tests. *Journal of the American Statistical Association* 25, 397–406.

Frisch, R. 1926. On a problem in pure economics. In *Preferences, Utility, and Demand*, ed. J. S. Chipman, L. Hurwicz, M.K. Richter and H.F. Sonnenschein. New York: Harcourt Brace Jovanovich, 1971.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12, 1–118.

Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A. 1971–90. *Foundations of Measurement*, 3 vols. New York: Academic Press.

Lucas, R. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A.H. Meltzer. Amsterdam: North-Holland.

Stevens, S.S. 1959. Measurement, psychophysics, and utility. In *Measurement. Definitions and Theories*, ed. C.W. Churchman and P. Ratoosh. New York: Wiley.

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1956.

How to cite this article

Boumans, Marcel. "measurement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000371> doi:10.1057/9780230226203.1075

The New Palgrave Dictionary of Economics Online

mechanism design (new developments)

Sandeep Baliga and Tomas Sjöström

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mechanism design concerns the question: given some desirable outcome, can we design a game which produces it? This theory provides a foundation for many important fields, such as auction theory and contract theory. We survey the recent literature dealing with topics such as robustness of mechanisms, renegotiation and collusion. An important issue is whether simple and intuitively appealing mechanisms can be optimal. Finally, we discuss what can be learned from recent experiments.

Keywords

adverse selection; asymmetric information; auctions; centralization; cheap talk; collusion; commitment; consequentialism; credibility; decentralization; delegation; dominant strategy mechanisms; free riding; hold-up; incentive compatibility; incomplete contracts; information rent; limited liability; mechanism design; moral hazard; Nash equilibrium; Paretian liberal; pooling equilibrium; principal and agent; prospect theory; refinements of Nash equilibrium; renegotiation; repeated games; revelation mechanism; revelation principle; side contracts; side transfers; signalling; social choice functions; social choice rules; soft-budget constraint; surplus; truthful implementation

Article

1 Possibility results and robustness

Game theory provides methods to predict the outcome of a given game. Mechanism design concerns the *reverse* question: given some desirable outcome, can we design a game which produces it? Formally, the *environment* is $\langle A, N, \Theta \rangle$, where *A* is a set of feasible and verifiable alternatives or outcomes, $N = \{1, \ldots, n\}$ is a set of agents, and Θ is a set of possible *states of the world*. Except where indicated, we consider *private values* environments, where a state is $\theta = (\theta_1, \ldots, \theta_n) \in \times_i \Theta_i = \Theta$, each agent *i*

knows his own 'type' $\theta_i \in \Theta_i$, and his payoff $u_i(a, \theta_i)$ depends only on the chosen alternative and his own type. (This does not rule out the possibility that the agents know something about each others' types.) If values are not private, then they are said to be *interdependent*. A *mechanism* or *contract* $\Gamma = (S, h)$ specifies a set of feasible actions S_i for each agent *i*, and an outcome function $h: S \equiv \times_{i=1}^{n} S_i \to A$. An outside party (a principal or social planner), or the agents themselves, want to design a mechanism which produces optimal outcomes. These are often represented by a *social choice rule* (SCR) $F: \Theta \to A$. A *social choice function* (SCF) is a single-valued SCR. Implicitly, it is assumed that the mechanism designer does not know the true θ, and this lack of information makes it impossible for her to directly choose an outcome in $F(\theta)$. Instead, she uses the more roundabout method of designing a mechanism which produces an outcome in $F(\theta)$, *whatever the true* θ *may be*.

In a *revelation mechanism*, each agent simply reports what he knows (so if agent *i* only knows θ_i then $S_i = \Theta_i$). By definition, an *incentive compatible* revelation mechanism has a *truthful* Bayesian–Nash equilibrium, that is, it achieves *truthful implementation*. Truthful implementation plays an important role in the theory because of the revelation principle (see the dictionary entry on mechanism design, which surveys the early literature on truthful implementation). The early literature produced powerful results on optimal mechanisms for auction design, bargaining problems, and other applications. However, it generally made quite strong assumptions, for example, that the agents and the principal share a common prior over Θ, that the principal can *commit* to a mechanism, that the agents cannot side-contract and always use equilibrium strategies, and so on. We survey the recent literature which deals with these issues. In addition, we note that the notion of truthful implementation has a drawback: it does not rule out the possibility that non-truthful equilibria also exist, and these may produce suboptimal outcomes. (A non-truthful equilibrium may even Pareto dominate the truthful equilibrium for the agents, and hence provide a natural focal point for coordinating their actions.) To rule out the possibility of suboptimal equilibria, we may require *full implementation*: for all $\theta \in \Theta$, the set of equilibrium outcomes should precisely equal $F(\theta)$.

Maskin (1999) assumed *complete information*: each agent knows the true θ. If $n \geq 3$ agents know θ, then any SCF can be truthfully implemented: let the agents report θ, and if at least $n - 1$ agents announce the same θ then implement the outcome $F(\theta)$. Unilateral deviations from a consensus are disregarded, so truth-telling is a Nash equilibrium. Of course, this revelation mechanism will also have non-truthful equilibria. For full implementation, more complex mechanisms are required. (Even if $n = 2$, any SCF can be truthfully implemented if the principal can credibly threaten to 'punish' both agents if they report different states; in an economic environment, this might be achieved by making each agent pay a fine.)

A necessary condition for full Nash implementation is (*Maskin*) *monotonicity* (Maskin, 1999). Intuitively, monotonicity requires that moving an alternative *up* in the agents' preference rankings should not make it *less* likely to be optimal. This condition can be surprisingly difficult to satisfy. For example, if the agents can have any complete and transitive preference relation on *A*, then any Maskin monotonic SCF must be a constant function (Saijo, 1987). The situation is quite different if we consider *refinements* of Nash equilibrium. For example, there is a sense in which almost any (ordinal) SCR can be fully implemented in *undominated Nash equilibrium* when the agents have complete information (Palfrey and Srivastava, 1991; Jackson, Palfrey and Srivastava, 1994; Sjöström, 1994). Chung and Ely

(2003) showed that this possibility result is not robust to small perturbations of the information structure that violate private values (there is a small chance that agent *i* knows more about agent *j*'s preferences than agent *j* does). The violation of private values is key. For example, in Sjöström's (1994) mechanism, an agent who knows his own preferences can eliminate his dominated strategies, and a second round of elimination of *strictly* dominated strategies generates the optimal outcome. This construction is robust to small perturbations that respect private values.

A different kind of robustness was studied by McLean and Postlewaite (2002). Consider an economic environment where each agent *i* observes an independently drawn signal t_i which is correlated with the state θ. The complete information structure is approximated by letting each agent's signal be very accurate. With complete information, any SCF can be truthfully implemented. McLean and Postlewaite (2002) show robustness to perturbations of the information structure: any outcome can be approximated by an incentive-compatible allocation, if the agents' signals are accurate enough. There is no need to assume private values.

The literature on *Bayesian mechanism design* typically assumes each agent *i* knows only his own type $\theta_i \in \Theta_i$, the agents share a common prior *p* over $\Theta \equiv \times_{i=1}^{n} \Theta_i$, and the principal knows *p*. In fact, for truthful implementation with $n \geq 3$, the assumption that the principal knows *p* is redundant. Suppose for any common prior *p* on Θ, there is an incentive-compatible revelation mechanism $\Gamma_p = (\times_{i=1}^{n} \Theta_i, h_p)$. By definition, Γ_p truthfully implements the SCF $F_p \equiv h_p$. The mechanism Γ_p is 'parametric', that is, it depends on *p*. To be specific, consider a quasi-linear public goods environment with independent types, and suppose Γ_p is the well-known mechanism of d'Aspremont and Gérard-Varet (1979). Now consider a nonparametric mechanism Γ, where each agent *i* announces *p* and θ_i. If at least $n-1$ agents report the same *p*, the outcome is $h_p(\theta_1,\ldots,\theta_n)$. Now, if agent *i* thinks everyone will announce *p* truthfully, he may as well do so. If in addition he thinks the other agents report θ_{-i} truthfully, then he should announce θ_i truthfully by incentive compatibility of Γ_p. Therefore, for any common prior *p*, the nonparametric mechanism Γ truthfully implements F_p. In this sense, the principal can use Γ to extract the agents' shared information about *p*. Of course, this particular mechanism also has non-truthful equilibria. Choi and Kim (1999) *fully* implemented the d'Aspremont and Gérard-Varet (1979) outcome in *undominated* Bayesian–Nash equilibrium, using a nonparametric mechanism. Naturally, their mechanism is quite complex. Suppose we restrict attention to mechanisms where each agent *i* only reports θ_i, truthfully in equilibrium. Then the necessary and sufficient condition for full nonparametric Bayesian–Nash implementation for any common prior *p* is (dominant strategy) incentive compatibility plus the *rectangular property* (Cason et al., 2006).

The d'Aspremont and Gérard-Varet (1979) mechanism is budget balanced and surplus maximizing. The above argument shows that such outcomes can be truthfully implemented by a nonparametric mechanism in quasi-linear environments with independent types. As is well known, this cannot be achieved by any dominant strategy mechanism. Thus, in general, nonparametric truthful implementation is easier than dominant strategy implementation. However, there are circumstances where the two concepts coincide. Bergemann and Morris (2005a) consider a model where each agent *i* has a *payoff type* $\theta_i \in \Theta_i$ and a *belief type* π_i. The payoff type determines the payoff function $u_i(a, \theta_i)$, while the belief type determines beliefs over other agents' types. The set of socially optimal outcomes $F(\theta)$ depends on

payoff types, but not on beliefs. Bergemann and Morris (2005a) show that in quasi-linear environments with no restrictions on side payments (hence no budget-balance requirement), truthful implementation for all possible type spaces with a common prior implies dominant strategy implementation. (For related results, see Section 4.)

Bergemann and Morris (2005b) consider *full* implementation of SCFs in a similar framework. The SCF $F: \Theta \to \mathcal{A}$ is *fully robustly implemented* if there exists a mechanism which fully implements *F* on all possible type spaces. They make no common prior assumption. Full robust implementation turns out to be equivalent to implementation using iterated elimination of strictly dominated strategies. Although a demanding concept, there are situations where full robust implementation is possible. For example, a Vickrey-Clarke-Groves (VCG) mechanism in a public goods economy with private values and strictly concave valuation functions achieves implementation in strictly dominant strategies. However, Bergemann and Morris (2005b) show the impossibility of full robust implementation when values are sufficiently interdependent.

A generalization of Maskin monotonicity called *Bayesian monotonicity* is necessary for (‘parametric’) full Bayesian-Nash implementation (Postlewaite and Schmeidler, 1986; Palfrey and Srivastava, 1989a; Jackson, 1991). Again, refinements lead to possibility results (Palfrey and Srivastava, 1989b). Another way to expand the set of implementable SCRs is *virtual* implementation (Abreu and Sen, 1991; Duggan, 1997). Serrano and Vohra (2001) argue that the sufficient conditions for virtual implementation are in fact quite strong.

The work discussed so far is *consequentialist*: only the final outcome matters. The mechanisms are clearly not meant to be descriptive of real-world institutions. For example, they typically require the agents to report ‘all they know’ before any decision is reached, an extreme form of centralized decision making hardly ever encountered in the real world. (The question of how much information must be transmitted in order to implement a given SCR is addressed by Hurwicz and Reiter, 2006, and Segal, 2004.) Delegating the power to make (verifiable) decisions to the agents would only create additional ‘moral hazard’ constraints, as discussed in the entry on mechanism design. Since centralization eliminates these moral hazard constraints, it typically strictly dominates decentralization in the basic model. However, as discussed below, by introducing additional aspects such as renegotiation and collusion, we can frequently prove the optimality of more realistic decentralized mechanisms. The implicit assumption is that decentralized decision making is in itself a good thing, which is a mild form of non-consequentialism. (Other non-consequentialist arguments are discussed in Section 4.) We might add that there is, of course, no way to eliminate the moral hazard constraints if the agents take *unverifiable* decisions that cannot be contracted upon. In this case, the issue of centralization versus decentralization of decisions is moot.

2 Renegotiation and credibility

Suppose $n = 2$ and both agents know the true θ. If a revelation mechanism is used and the agents announce different states, then we cannot identify a deviator from a ‘consensus’, so it may be necessary to punish *both* agents in order to support a truth-telling equilibrium. But this threat is not credible if the agents can avoid punishment by renegotiating the outcome. Maskin and Moore (1999) capture the

renegotiation process by an exogenously given function $r: A \times \Theta \to \mathcal{A}$ which maps outcome *a* in state θ into an efficient outcome $r(a, \theta)$. They derive an incentive-compatibility condition which is necessary for truth-telling when $n = 2$, and show that *renegotiation monotonicity* is necessary for full Nash implementation (see also Segal and Whinston, 2002).

The idea that renegotiation may preclude the implementation of the first-best outcome, even when information is complete, has received attention in models of bilateral trade with relationship-specific investments (the hold-up problem). It is possible to implement the first-best outcome if trade is one-dimensional and investments are ‘selfish’, in the sense that each agent's investment does not directly influence the other agent's payoff (Nöldeke and Schmidt, 1995; Edlin and Reichelstein, 1996). If investments are not selfish, then the first-best cannot always be achieved, while the second-best can often be implemented without any explicit contract (Che and Hausch, 1999). Segal (1999) found a similar result in a model with *k* goods and selfish investments, for *k* large (see also Maskin and Tirole, 1999; Hart and Moore, 1999). It should be noted that the case $n = 2$ is quite special, and adding a third party often alleviates the problem of renegotiation (Baliga and Sjöström, 2006).

Credibility and renegotiation also impact trading with asymmetric information. Suppose the seller can produce goods of different quality, but the buyer's valuation is his private information. It is typically second-best optimal for the seller to offer a contract such that low-valuation buyers consume less than first-best quality (‘underproduction’), while high-valuation buyers enjoy ‘information rents’. Incentive compatibility guarantees that the buyer reveals his true valuation. Now suppose trading takes place twice, and the buyer's valuation does not change. Suppose the seller cannot credibly commit to a long-run (two-period) contract. If the buyer reveals his true valuation in the first period, then in the second period the seller will leave him no rent. This is typically not the second-best outcome. The seller may prefer a ‘pooling’ contract which does not fully reveal valuations in the first period, a commitment device which limits his ability to extract second period rents. This idea has important applications. When a regulator cannot commit to a long-run contract, a regulated firm may hide information or exert less effort to cut costs, the *ratchet effect* (Freixas, Guesnerie and Tirole, 1985). A borrower may not exert effort to improve a project knowing that a lender with deep pockets will bail him out, the *soft budget constraint* (Dewatripont and Maskin, 1995a). These problems are exacerbated if the principal is well informed and cannot commit not to use his information. Institutional or organizational design can alleviate the problems. By committing to acquire less information via ‘incomplete contracts’, or by maintaining an ‘arm's-length relationship’, the principal can improve efficiency (Dewatripont and Maskin, 1995b; Crémer, 1995). Less frequent regulatory reviews offset the ratchet effect, and a decentralized credit market helps to cut off borrowers from future funding. Long-run contracts can help, but they may be vulnerable to renegotiation (Dewatripont, 1989). In particular, the second-period outcome may be renegotiated if quality levels are known to be different from the first-best. Again, some degree of pooling may be optimal.

If the principal cannot commit even to short-run contracts, then, after receiving the agents’ messages, she always chooses an outcome that is optimal given her beliefs. She cannot credibly threaten punishments that she would not want to carry out. Refinements proposed in the cheap-talk literature suggest that a putative pooling equilibrium may be destroyed if an agent can reveal information by ‘objecting’ in a credible way. This leads to a necessary condition for full implementation with complete information which is reminiscent of Maskin monotonicity, but which involves the principal's preferences

(Baliga, Corchón and Sjöström, 1997).

3 Collusion

A large literature on collusion was inspired by Tirole (1986). A key contribution was made by Laffont and Martimort (1997), who assumed an uninformed third party proposes side contracts. This circumvents the signalling problems that might arise if a privately informed agent makes collusive proposals. A side contract for a group of colluding agents is a *collusive mechanism* which must respect incentive compatibility, individual rationality and feasibility constraints. The original mechanism Γ, designed by the principal, is called the *grand mechanism*. The objective is to design an optimal grand mechanism when collusion is possible. Typically, collusion imposes severe limits on what can be achieved.

Baliga and Sjöström (1998) study a model with moral hazard and limited liability. Two agents share information not known to the principal: agent 1's effort is observed by both agents. Agent 2's effort is known only to himself. In the absence of collusion, the optimal grand mechanism specifies a ‘message game’: agent 2 reports agent 1's effort to the principal. Now suppose the agents can side contract on agent 1's effort, but not on agent 2's effort (which is unobserved). Side contracts can specify side transfers as a function of realized output, but must respect limited liability. This collusion may destroy centralized ‘message games’, and we obtain a theory of optimal delegation of decision making. For some parameters, it is optimal for the principal to contract only with agent 2, and let agent 2 subcontract with agent 1. This is intuitive, since agent 2 observes agent 1's effort and can contract directly on it. More surprisingly, there are parameter values where it is better for the principal to contract only with agent 1.

Mookherjee and Tsumagari (2004) study a similar model, but with adverse selection: the agents privately observe their own production costs. In this model, delegating to a ‘prime supplier’ creates ‘double marginalization of rents’: the prime supplier uses underproduction to minimize the other agent's information rent. A centralized contract avoids this problem. Hence, in this model delegation is always strictly dominated by centralization, even though the agents can collude.

Mookherjee and Tsumagari (2004) assume the agents can side contract before deciding to participate in the grand contract. Che and Kim (2006) assume side contracting occurs only after the decision to participate in the grand mechanism has been made. In this case, collusion does not limit what the principal can achieve. Hence, the timing of side contracting is important. In a complete information environment with $n \geq 3$, Sjöström (1999) showed that neither renegotiation nor collusion limit the possibility of undominated Nash implementation.

4 Other theoretical issues

In quasi-linear environments with uncorrelated types, there exist incentive-compatible mechanisms which maximize the social surplus (for example, d'Aspremont and Gérard-Varet, 1979). But the principal cannot extract all the surplus: the agents must get informational rents. However, Crémer and McLean (1988) showed that the principal can extract all the surplus in auctions with *correlated types*.

McAfee and Reny (1992) extended this result to general quasi-linear environments.

Jehiel and Moldovanu (2001) considered a quasi-linear environment with multidimensional (uncorrelated) types and interdependent values. Generically, a standard revelation mechanism cannot be designed to extract information about multidimensional types, and no incentive-compatible and surplus-maximizing mechanism exists. Mezzetti (2004) presents an ingenious *two-stage* mechanism which maximizes the surplus in interdependent values environments, even when types are independent and multidimensional. In the first stage, the mechanism specifies an outcome decision but not transfers. Transfers are determined in the second stage by reports on payoffs realized by the outcome decision. Mezzetti (2007) shows that the principal can sometimes extract all the surplus by this method, even if types are independent. For optimal mechanisms for a profit-maximizing monopolist when consumers have multidimensional types and private values, see Armstrong (1996).

Incentive compatibility does not require that each agent has a dominant strategy. Nevertheless, incentive-compatible outcomes can often be replicated by dominant strategy mechanisms (Mookherjee and Reichelstein, 1992). In quasi-linear environments, all incentive-compatible mechanisms that maximize the social surplus are *payoff-equivalent* to dominant strategy (VCG) mechanisms (Krishna and Perry, 1997; Williams, 1999). However, as pointed out above, dominant strategies (but not incentive compatibility) rules out budget balance.

Bergemann and Välimäki (2002) assume agents can update a common prior by costly information acquisition. Suppose a single-unit auction has two bidders *i* and *j* who observe statistically independent private signals θ_i and θ_j. Bidder *i*'s valuation of the good is $u_i(\theta_i, \theta_j) = \alpha\theta_i + \beta\theta_j$, where $\alpha > \beta > 0$. Thus, values are interdependent. Efficiency requires that bidder *i* gets the good if and only if $\theta_i \geq \theta_j$. Suppose bidders report their signals, the good is allocated efficiently given their reports, and the winning bidder *i* pays the price $(\alpha + \beta)\theta_j$. This VCG mechanism is incentive compatible (Maskin, 1992). If bidder *i* acquires negative information which causes him to lose the auction, then he imposes a negative externality on the other bidder (as $\beta > 0$). This implies the bidders have an incentive to collect too much information. Conversely, there is an incentive to collect too little information when $\beta < 0$. Bergemann and Välimäki (2002) provide a general analysis of these externalities. Similar externalities occur when members of a committee must collect information before voting. If the committee is large, each vote is unlikely to be pivotal, and free riding occurs. Persico (2004) shows how the optimal committee is designed to encourage the members to collect information.

Some authors reject consequentialism and instead emphasize agents' *rights*. For example, suppose a mechanism implements *envy-free outcomes*. An agent might still feel unfairly treated if his own bundle is worse than a bundle which another agent *had the right to choose* (but did not). Such agents may demand 'equal rights' (Gaspart, 1995). Unfortunately, once we leave the classical exchange economy, Sen's 'Paretian liberal' paradox (Sen, 1970) suggests that rights are incompatible with efficiency (Deb, Pattanaik and Razzolini, 1997). Sen originally considered rights embodied in SCRs rather than mechanisms. Peleg and Winter (2002) study *constitutional implementation* where the mechanism embodies the same rights as the SCR it implements.

5 Learning from experiments

Cabrales, Charness and Corchón (2003) tested the so-called canonical mechanism for Nash implementation. A Nash equilibrium was played only 13 per cent of the time (20 per cent when monetary fines were used). Remarkably, the optimal outcome was implemented 68 per cent of the time (80 per cent with ‘fines’), because deviations from equilibrium strategies frequently did not affect the outcome. This suggests that a desirable property of a mechanism is *fault-solerance*: it should produce optimal outcomes even if some ‘faulty’ players deviate from the theoretical predictions. Eliaz (2002) showed that, if at most $k < \frac{1}{2} n - 1$ players are ‘faulty’ (that is, unpredictable), then full Nash implementation is possible if *no-veto-power* and $(k + 1)$*-monotonicity hold*.

Equilibrium play can be justified by epistemic or dynamic theories. According to epistemic theories, common knowledge about various aspects of the game implies equilibrium play even in one-shot games. Experiments provide little support for this. However, there is evidence that players can reach equilibrium through a dynamic adjustment process. If a game is played repeatedly, with no player knowing any other player's payoff function, the outcome frequently converges to a Nash equilibrium of the one-shot complete information game (Smith, 1979). Dynamic theories have been applied to the mechanism design problem (for example, Cabrales and Ponti, 2000). Chen and Tang (1998) and Chen and Gazzale (2004) argue that mechanisms which induce supermodular games produce good long-run outcomes. Unfortunately, these convergence results are irrelevant for decisions that are taken infrequently, or if the principal is too impatient to care only about the long-run outcome.

The idea of dominant strategies is less controversial than Nash equilibrium, and should be more relevant for decisions that are taken infrequently. Unfortunately, experiments on dominant-strategy mechanisms have yielded negative results. Attiyeh, Franciosi and Isaac (2000, p. 112) conclude pessimistically, ‘we do not believe that the pivot mechanism warrants further practical consideration … . This is due to the fundamental failure of the mechanism, in our laboratory experiments, to induce truthful value revelation.’ However, VCG mechanisms (such as the pivotal mechanism) frequently have a multiplicity of Nash equilibria, some of which produce suboptimal outcomes. Cason et al. (2006) did experiments with *secure* mechanisms, which fully implement an SCR both in dominant strategies and in Nash equilibria. The players were much more likely to use their dominant strategies in secure than in non-secure mechanisms. In the non-secure mechanisms, deviations from dominant strategies tended to correspond to Nash equilibria. However, these deviations typically did not lead to suboptimal *outcomes*. In this sense, the non-secure mechanisms were fault-tolerant. Kawagoe and Mori (2001) report experiments where deviations from dominant strategies typically corresponded to suboptimal Nash equilibria.

In experiments, subjects often violate standard axioms of rational decision making. Alternative theories, such as prospect theory, fit the experimental evidence better. But, if we modify the axioms of individual behaviour, the optimal mechanisms will change. Esteban and Miyagawa (2005) assume the agents have Gul–Pesendorfer preferences (Gul and Pesendorfer, 2001). They suffer from ‘temptation’, and may prefer a smaller menu (choice set) to a larger one. Suppose each agent first chooses a menu, and then chooses an alternative from this menu. Optimal menus may contain ‘tempting’ alternatives which are never chosen in equilibrium, because this relaxes the incentive-compatibility constraints pertaining to the choice of menu. Eliaz and Spiegler (2006) assume some agents are ‘sophisticated’ and some are

‘naive’. Sophisticated agents know that they are dynamically inconsistent, and would like to commit to a future decision. Naive agents are unaware that they are dynamically inconsistent. The optimal mechanism screens the agents by providing commitment devices that are chosen only by sophisticated agents.

Experiments reveal the importance of human emotions such as spite or kindness (Andreoni, 1995; Saijo, 2003). In many mechanisms in the theoretical literature, by changing his strategy an agent can have a big impact on another agent's payoff without materially changing his own. Such mechanisms may have little hope of practical success if agents are inclined to manipulate each others’ payoffs due to feelings of spite or kindness.

See Also

- auctions (experiments)
- auctions (theory)
- contract theory
- hold-up problem
- incentive compatibility
- mechanism design
- revelation principle

Bibliography

Abreu, D. and Sen, A. 1991. Virtual implementation in Nash equilibria. *Econometrica* 59, 997–1022.

Andreoni, J. 1995. Cooperation in public goods experiments: kindness or confusion? *American Economic Review* 85, 891–904.

Armstrong, M. 1996. Multiproduct nonlinear pricing. *Econometrica* 64, 51–75.

Attiyeh, G., Franciosi, R. and Isaac, R.M. 2000. Experiments with the pivotal process for providing public goods. *Public Choice* 102, 95–114.

Baliga, S., Corchón, L. and Sjöström, T. 1997. The theory of implementation when the planner is a player. *Journal of Economic Theory* 77, 15–33.

Baliga, S. and Sjöström, T. 1998. Decentralization and collusion. *Journal of Economic Theory* 83, 196–232.

Baliga, S. and Sjöström, T. 2006. Contracting with third parties. Working paper No. 75, CSIO, Northwestern University.

Bergemann, D. and Morris, S. 2005a. Robust mechanism design. *Econometrica* 73, 1771–813.

Bergemann, D. and Morris, S. 2005b. Robust implementation: the role of large type spaces. Discussion Paper No. 1519, Cowles Foundation, Yale University.

Bergemann, D. and Välimäki, J. 2002. Information acquisition and efficient mechanism design. *Econometrica* 70, 1007–34.

Cabrales, A., Charness, G. and Corchón, L. 2003. An experiment on Nash implementation. *Journal of Economic Behavior and Organization* 51, 161–93.

Cabrales, A. and Ponti, G. 2000. Implementation, elimination of weakly dominated strategies and evolutionary dynamics. *Review of Economic Dynamics* 3, 247–82.

Cason, T., Saijo, T., Sjöström, T. and Yamato, T. 2006. Secure implementation experiments: do strategy-proof mechanisms really work? *Games and Economic Behavior* 57, 206–35.

Che, Y.K. and Hausch, D. 1999. Cooperative investments and the value of contracting. *American Economic Review* 89, 125–47.

Che, Y.K. and Kim, J. 2006. Robustly collusion-proof implementation. *Econometrica* 74, 1063–107.

Chen, Y. and Gazzale, R. 2004. When does learning in games generate convergence to Nash equilibria? *American Economic Review* 94, 1505–35.

Chen, Y. and Tang, F.F. 1998. Learning and incentive compatible mechanisms for public goods provision: an experimental study. *Journal of Political Economy* 106, 633–62.

Choi, J. and Kim, T. 1999. A nonparametric, efficient public decision mechanism: undominated Bayesian Nash implementation. *Games and Economic Behavior* 27, 64–85.

Chung, K. and Ely, J. 2003. Implementation with near-complete information. *Econometrica* 71, 857–71.

Crémer, J. 1995. Arm's length relationships. *Quarterly Journal of Economics* 110, 275–95.

Crémer, J. and McLean, R. 1988. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica* 56, 1247–57.

d'Aspremont, C. and Gérard-Varet, L.A. 1979. Incentives and incomplete information. *Journal of Public Economics* 11, 25–45.

Deb, R., Pattanaik, P. and Razzolini, L. 1997. Game forms, rights and the efficiency of social outcomes. *Journal of Economic Theory* 72, 74–95.

Dewatripont, M. 1989. Renegotiation and information revelation over time: the case of optimal labor contracts, *Quarterly Journal of Economics* 104, 589–619.

Dewatripont, M. and Maskin, E. 1995a. Credit and efficiency in centralized and decentralized economies. *Review of Economic Studies* 62, 541–55.

Dewatripont, M. and Maskin, E. 1995b. Contractual contingencies and renegotiation. *RAND Journal of Economics* 26, 704–19.

Duggan, J. 1997. Virtual Bayesian implementation. *Econometrica* 67, 1175–99.

Edlin, A. and Reichelstein, S. 1996. Hold-ups, standard breach remedies and optimal investment. *American Economic Review* 86, 478–501.

Eliaz, K. 2002. Fault tolerant implementation. *Review of Economic Studies* 69, 589–610.

Eliaz, K. and Spiegler, R. 2006. Contracting with diversely naive agents. *Review of Economic Studies* 73, 689–714.

Esteban, S. and Miyagawa, E. 2005. Optimal menu of menus with self-control preferences. Unpublished paper, Penn State University.

Freixas, X., Guesnerie, R. and Tirole, J. 1985. Planning under incomplete information and the ratchet effect. *Review of Economic Studies* 52, 173–92.

Gaspart, F. 1995. Fair implementation in the cooperative production problem: two properties of normal form mechanisms. Unpublished manuscript, FUNDP Namur.

Gul, F. and Pesendorfer, W. 2001. Temptation and self-control. *Econometrica* 69, 1403–35.

Hart, O. and Moore, J. 1999. Foundations of incomplete contracts. *Review of Economic Studies* 66, 115–38.

Hurwicz, L. and Reiter, S. 2006. *Designing Economic Mechanisms*. New York: Cambridge University Press.

Jackson, M. 1991. Bayesian implementation. *Econometrica* 59, 461–77.

Jackson, M., Palfrey, T. and Srivastava, S. 1994. Undominated Nash implementation in bounded mechanisms. *Games and Economic Behavior* 6, 474–501.

Jehiel, P. and Moldovanu, B. 2001. Efficient design with interdependent valuations. *Econometrica* 69, 1237–59.

Kawagoe, T. and Mori, T. 2001. Can the pivotal mechanism induce truthtelling? *Public Choice* 108, 331–54.

Krishna, V. and Perry, M. 1997. Efficient mechanism design. Unpublished manuscript, Penn State University.

Laffont, J.J. and Martimort, D. 1997. Collusion under asymmetric information. *Econometrica* 65, 875–911.

McAfee, P. and Reny, P. 1992. Correlated information and mechanism design. *Econometrica* 60, 395–421.

McLean, R. and Postlewaite, A. 2002. Informational size and incentive compatibility. *Econometrica* 70, 2421–53.

Maskin, E. 1992. Auctions and privatization. In *Privatization*, ed. H. Siebert. Tübingen: J.C.B. Mohr.

Maskin, E. 1999. Nash equilibrium and welfare optimality. *Review of Economic Studies* 66, 23–38.

Maskin, E. and Moore, J. 1999. Implementation and renegotiation. *Review of Economic Studies* 66, 39–56.

Maskin, E. and Tirole, J. 1999. Two remarks on the property rights literature. *Review of Economic Studies* 66, 139–50.

Mezzetti, C. 2004. Mechanism design with interdependent valuations: efficiency. *Econometrica* 72, 1617–26.

Mezzetti, C. 2007. Mechanism design with interdependent valuations: surplus extraction. *Economic Theory* 3, 473–99.

Mookherjee, D. and Reichelstein, S. 1992. Dominant strategy implementation of Bayesian incentive compatible allocation rules. *Journal of Economic Theory* 56, 378–99.

Mookherjee, D. and Tsumagari, M. 2004. The organization of supplier networks: effects of delegation

and intermediation. *Econometrica* 72, 1179–219.

Nöldeke, G. and Schmidt, K. 1995. Option contracts and renegotiation: a solution to the hold–up problem. *RAND Journal of Economics* 26, 163–79.

Palfrey, T. and Srivastava, S. 1989a. Implementation with incomplete information in exchange economies. *Econometrica* 57, 115–34.

Palfrey, T. and Srivastava, S. 1989b. Mechanism design with incomplete information: a solution to the implementation problem. *Journal of Political Economy* 97, 668–91.

Palfrey, T. and Srivastava, S. 1991. Nash implementation using undominated strategies. *Econometrica* 59, 479–501.

Peleg, B. and Winter, E. 2002. Constitutional implementation. *Review of Economic Design* 7, 187–204.

Persico, N. 2004. Committee design with endogenous information. *Review of Economic Studies* 71, 165–91.

Postlewaite, A. and Schmeidler, D. 1986. Implementation in differential information economies. *Journal of Economic Theory* 39, 14–33.

Saijo, T. 1987. On constant Maskin monotonic social choice functions. *Journal of Economic Theory* 42, 382–86.

Saijo, T. 2003. Spiteful behavior in voluntary contribution mechanism experiments. In *Handbook of Experimental Economics Results*, ed. C. Plott and V. Smith, Elsevier Science.

Segal, I. 1999. Complexity and renegotiation: a foundation for incomplete contracts. *Review of Economic Studies* 66, 57–82.

Segal, I. 2004. The communication requirements of social choice rules and supporting budget sets. Working Paper No. 39, School of Social Science, Institute for Advanced Study, Princeton.

Segal, I. and Whinston, M. 2002. The Mirrlees approach to mechanism design with renegotiation. *Econometrica* 70, 1–46.

Sen, A. 1970. The impossibility of a Paretian liberal. *Journal of Political Economy* 78, 152–7.

Serrano, R. and Vohra, R. 2001. Some limitations of virtual Bayesian implementation. *Econometrica* 69, 785–92.

Sjöström, T. 1994. Implementation in undominated Nash equilibria without using integer games. *Games and Economic Behavior* 6, 502–11.

Sjöström, T. 1999. Undominated Nash implementation with collusion and renegotiation. *Games and Economic Behavior* 26, 337–52.

Smith, V. 1979. Incentive compatible experimental processes for the provision of public goods. In *Research in Experimental Economics*, vol. 1, ed. V. Smith. Greenwich, CT: JAI Press.

Tirole, J. 1986. Hierarchies and bureaucracies. *Journal of Law, Economics and Organization* 2, 181–214.

Williams, S. 1999. A characterization of efficient, Bayesian incentive compatible mechanisms. *Economic Theory* 14, 155–80.

How to cite this article

Baliga, Sandeep and Tomas Sjöström. "mechanism design (new developments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000395> doi:10.1057/9780230226203.1079

The New Palgrave Dictionary of Economics Online

mechanism design experiments

Yan Chen and John O. Ledyard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mechanism design experiments bridge the gap between a theoretical mechanism and an actual economic process. In the domain of public goods, matching and combinatorial auctions and laboratory experiments identify features of mechanisms that lead to good performance when implemented among boundedly rational agents. These features include dynamic stability and security in public goods mechanisms, transparency in matching mechanisms, package bidding, simultaneity and iteration in combinatorial auctions.

Keywords

bounded rationality; combinatorial auction; compensation mechanism; convergence; cooperation; dynamic stability; English auction; externalities; matching; mechanism design; mechanism design experiments; misrevelation; public goods; public goods production functions; sealed bid auction; social identity theory; social loafing; subgame perfection; supermodularity; value complementarities

Article

Mechanism design is the art of designing institutions that align individual incentives with overall social goals. Mechanism design theory was initiated by Hurwicz (1972) and is surveyed in Groves and Ledyard (1987). To bridge the gap between a theoretical mechanism and an actual economic process that solves fundamental social problems, it is important to observe and evaluate the performance of the mechanism in the context of actual decision problems faced by real people with real incentives. These situations can be created and carefully controlled in a laboratory. A mechanism design experiment takes a theoretical mechanism, recreates it in a simple environment in a laboratory with human subjects as economic agents, observes the behaviour of human subjects under the mechanism, and assesses its performance in relation to what it was created to do and to the theory upon which its creation rests. The laboratory serves as a wind tunnel for new mechanisms, providing evidence which one can use to eliminate fragile ones, and to identify the characteristics of successful ones.
When a mechanism is put to test in a laboratory, behavioural assumptions made in theory are seriously challenged. Theory assumes perfectly rational agents who can compute the equilibrium strategies via introspection. When a mechanism is implemented among boundedly rational agents, however, characteristics peripheral to theoretical implementations, such as transparency, complexity and dynamic stability, become important, or even central, to the success of a mechanism in a laboratory, and we suspect, ultimately in the real world. Mechanism design experiments cover several major domains, including public goods and externalities, matching, contract theory, auctions, market design and information markets. In what follows, we will review the experimental results of some of these topics.

Public goods and externalities

With the presence of public goods and externalities, competitive equilibria are not Pareto optimal. This is often referred to as market failure, since competitive markets on their own either result in underprovision of public goods (that is, the free-rider problem) or overprovision of negative externalities, such as pollution. To solve the free-rider problem in public goods economies, incentive-compatible mechanisms use innovative tax-subsidy schemes that utilize agents' own messages to achieve the Pareto optimal levels of public goods

provision. A series of experiments test these mechanisms in the laboratory (see Chen, 2008, for a comprehensive survey).

When preferences are quasi-linear, the Vickrey–Clarke–Groves (VCG) mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973) is strategy-proof, in the sense that reporting one's preferences truthfully is always a dominant strategy. It has also been shown that any strategy-proof mechanism selecting an efficient public decision at every profile must be of this type (Green and Laffont, 1977). Two forms of the VCG mechanism have been tested in the field and laboratory by various groups of researchers. The pivot mechanism refers to the VCG mechanism when the public project choice is binary, while the cVCG mechanism refers to the VCG mechanism when the level of the public good is selected from a continuum. Under the pivot mechanism, misrevelation can be prevalent. Attiyeh, Franciosi and Isaac (2000) show that about ten per cent of the bids were truthfully revealing their values. Furthermore, there was no convergence tendency towards value revelation. In a follow-up study, Kawagoe and Mori (2001) show that more information about the payoff structure helps reduce the degree of misrevelation. More recently, Cason et al. (2006) provide a novel explanation for the problem of misrevelation in strategy-proof mechanisms. As Saijo et al. (2005) point out, the standard strategy-proofness concept in implementation theory has serious drawbacks, that is, almost all strategy-proof mechanisms have a continuum of Nash equilibria. They propose a new implementation concept, secure implementation, which requires the set of dominant strategy equilibria and the set of Nash equilibria to coincide. Cason et al. (2006) compare the performance of two strategy-proof mechanisms in the laboratory: the Pivot mechanism where implementation is not secure and truthful preference revelation is a weakly dominant strategy, and the cVCG mechanism with single-peaked preferences where implementation is secure. Results indicate that subjects play dominant strategies significantly more often in the secure cVCG mechanism (81 per cent) than in the non-secure Pivot mechanism (50 per cent). The importance of secure implementation in dominant strategy implementation is replicated in Healy (2006), where he compares five public goods mechanisms, voluntary contributions, proportional taxation, Groves–Ledyard, Walker and cVCG. The cVCG is found to be the most efficient of all mechanisms.

Although the VCG mechanism admits dominant strategies, the allocation is not fully Pareto-efficient. In fact, it is impossible to design a mechanism for making collective allocation decisions, which is informationally decentralized, non-manipulable and Pareto optimal. This impossibility has been demonstrated in the work of Hurwicz (1975), Green and Laffont (1977), Roberts (1979), Walker (1980) and Mailath and Postlewaite (1990) in the context of resource allocation with public goods.

Many ‘next-best’ mechanisms preserve Pareto optimality at the cost of non-manipulability, some of which preserve ‘some degree’ of non-manipulability. Some mechanisms have been discovered which have the property that Nash equilibria are Pareto optimal. These can be found in the work of Groves and Ledyard (1977), Hurwicz (1979), Walker (1981), Tian (1989), Kim (1993), Peleg (1996), Falkinger (1996) and Chen (2002). Other implementation concepts include perfect Nash equilibrium (Bagnoli and Lipman, 1989), undominated Nash equilibrium (Jackson and Moulin, 1992), subgame perfect equilibrium (Varian, 1994), strong equilibrium (Corchon and Wilkie, 1996), and the core (Kaneko, 1977), and so forth. Apart from the above non-Bayesian mechanisms, Ledyard and Palfrey (1994) propose a class of Bayesian Nash mechanisms for public goods provision.

Experiments on Nash-efficient public goods mechanisms underscore the importance of dynamic stability, that is, whether a mechanism converges under various learning dynamics. Most of the experimental studies of Nash-efficient mechanisms focus on the Groves–Ledyard mechanism (Smith, 1979a; 1979b; Harstad and Marrese, 1981; 1982; Mori, 1989; Chen and Plott, 1996; Arifovic and Ledyard, 2006). Chen and Tang (1998) also compare the Walker mechanism with the Groves–Ledyard mechanism. Falkinger et al. (2000) study the Falkinger mechanism. Healy (2006) compares Nash-efficient mechanisms to cVCG and other benchmarks.

Among the series of experiments exploring dynamic stability, Chen and Plott (1996) first assessed the performance of the Groves–Ledyard mechanism under different punishment parameters. They found that by varying the punishment parameter the dynamics and stability changed dramatically. For a large enough parameter, the system converged very quickly to its stage game Nash equilibrium and remained stable; while under a small parameter, the system did not converge to its stage game Nash equilibrium. This finding was replicated by Chen and Tang (1998) with more independent sessions and a longer time series in an experiment designed to study the learning dynamics.

Figure 1 presents the time series data from Chen and Tang (1998) for two out of five types of players. Each graph presents the mean (the black dots) and standard deviation (the error bars) for each of the two different types averaged over seven independent sessions for each mechanism – the Walker mechanism, the Groves–Ledyard mechanism under a low punishment parameter (GL1), and the Groves–Ledyard mechanism under a high punishment parameter (GL100). From these graphs, it is apparent that GL100 converged very quickly to its stage game Nash equilibrium and remained stable, while the same mechanism did not converge under a low punishment parameter; the Walker mechanism did not converge to its stage game Nash equilibrium either.

Figure 1

Mean contribution and standard deviation in Chen and Tang (1998)

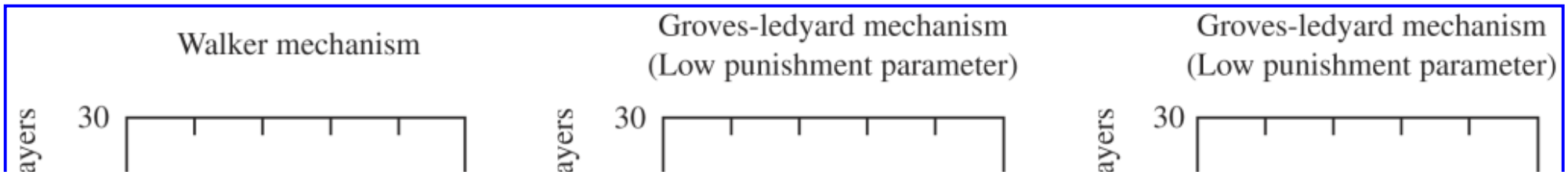

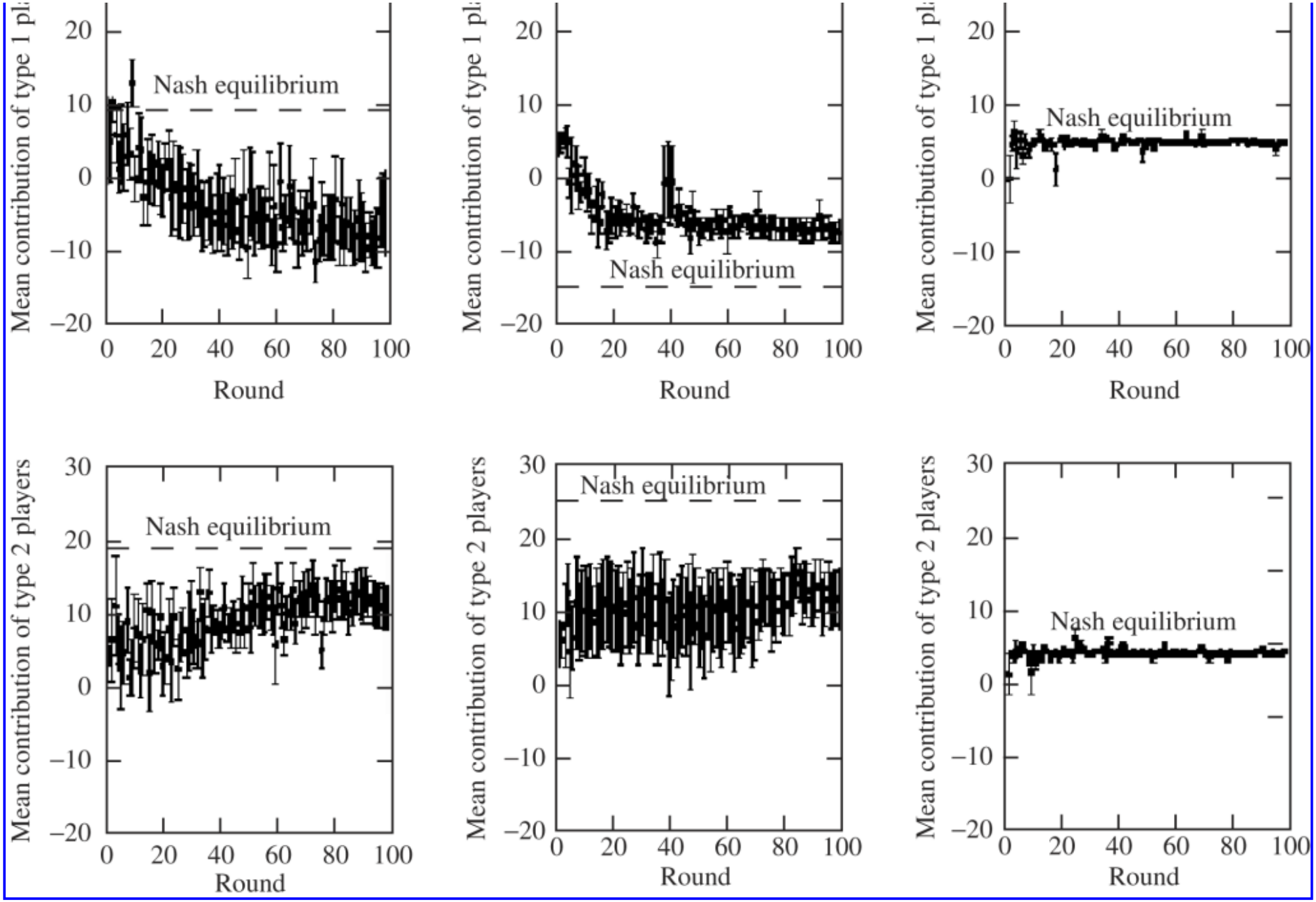

Because of its good dynamic properties, GL100 had significantly better performance than GL1 and Walker, evaluated in terms of system efficiency, close to Pareto optimal level of public goods provision, fewer violations of individual rationality constraints and convergence to its stage game equilibrium.

These past experiments serendipitously studied supermodular mechanisms. Two recent studies systematically vary the parameters from below, close to, at and above the supermodularity threshold to assess the effects of supermodularity on learning dynamics.

Arifovic and Ledyard (2006) conduct computer simulations of an individual learning model in the context of a class of the Groves–Ledyard mechanisms. They vary the punishment

parameter systematically, from extremely small to extremely high. They find that their model converges to Nash equilibrium for all values of the punishment parameter. However, the speed of convergence depends on the value of the parameter. As shown in Figure 2, the speed of convergence is U-shaped: very low and very high values of the punishment parameter require long periods for convergence, while a range of intermediate values requires the minimum time. In fact, the optimal punishment parameter identified in the simulation is much lower than the supermudularity threshold. Predictions of the computation model are validated by experimental data with human subjects.

Figure 2

Convergence speed in Groves–Ledyard in Arifovic and Ledyard (2006)

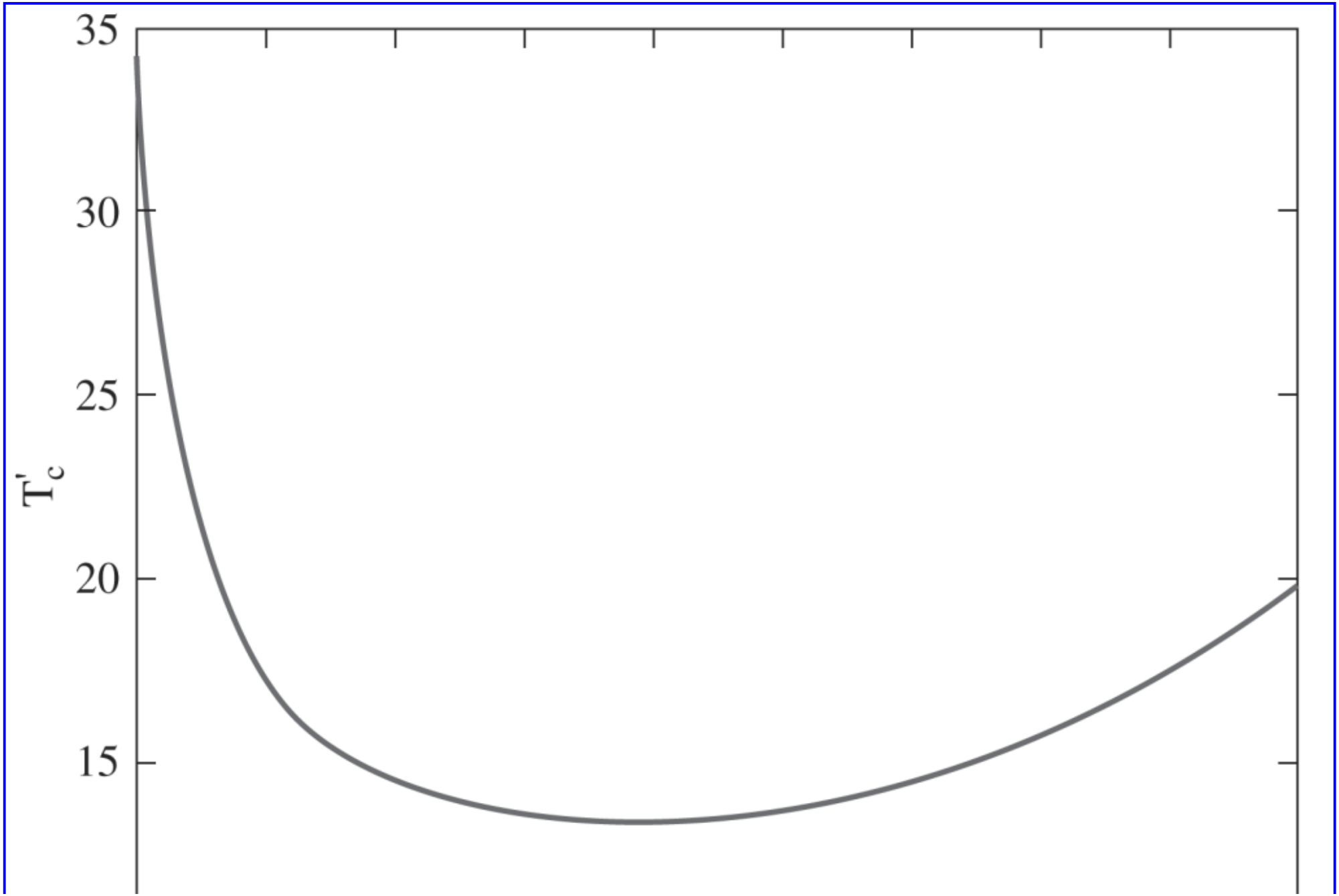

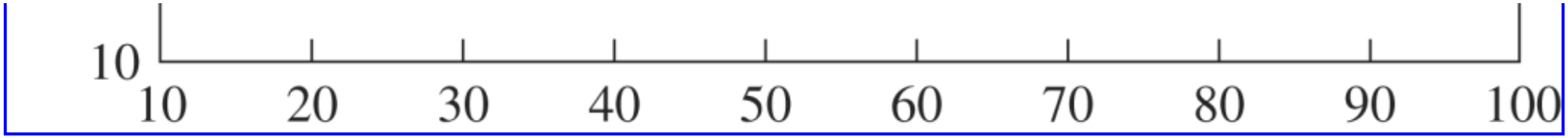

In a parallel research project on the role of supermodularity on convergence, Chen and Gazzale (2004) experimentally study the generalized version of the compensation mechanism (Varian, 1994), which implements efficient allocations as subgame-perfect equilibria for economic environments involving externalities and public goods. The basic idea is that each player offers to compensate the other for the 'costs' incurred by making the efficient choice. They systematically vary the free parameter from below, close to, at and beyond the threshold of supermodularity to assess the effects of supermodularity on the performance of the mechanism. They have three main findings. First, in terms of proportion of equilibrium play and efficiency, they find that supermodular and 'near supermodular' mechanisms perform significantly better than those far below the threshold. This finding is consistent with previous experimental findings. Second, they find that from a little below the threshold to the threshold, the improvement in performance is statistically insignificant. This implies that the performance of 'near supermodular' mechanisms, such as the Falkinger mechanism, ought to be comparable to supermodular mechanisms. Therefore, the mechanism designer need not be overly concerned with setting parameters that are firmly above the supermodular threshold – close is just as good. This enlarges the set of robustly stable mechanisms. The third finding concerns the selection of mechanisms within the class of supermodular mechanisms. Again, theory is silent on this issue. Chen and Gazzale find that within the class of supermodular mechanisms, increasing the parameter far beyond the threshold does not significantly improve the performance of the mechanism. Furthermore, increasing another free parameter, which is not related to whether or not the mechanism is supermodular, does improve convergence.

In contrast to the previous stream of work which identifies supermodularity as a robust sufficient condition for convergence, Healy (2006) develops a *k*-period average best response learning model and calibrates this new learning model on the data-set to study the learning dynamics. He shows that subject behaviour is well approximated by a model in which agents best respond to the average strategy choices over the last five periods under all mechanisms. Healy's work bridges the behavioural hypotheses that have existed separately in dominant strategy and Nash-efficient mechanism experiments.

In summary, experiments testing public goods mechanisms show that dominant strategy mechanisms should also be secure, while Nash implementation mechanisms should satisfy dynamic stability, if any mechanism is to be considered for application in the real world in a repeated interaction setting.

While experimental research demonstrates that incentive-compatible public goods mechanisms can be effective in inducing efficient levels of public goods provision, almost all the mechanisms rely on monetary transfers, which limit the scope of implementation of these mechanisms in the real world. In many interesting real world settings, such as open source software development and online communities, sizable contributions to public goods are made without the use of monetary incentives. We next review a related social psychology literature, which studies contribution to public goods without the use of monetary incentives.

Social loafing

Analogous to free riding, social loafing refers to the phenomenon whereby individuals exert less effort on a collective task than they do on a comparable individual task. To determine conditions under which individuals do or do not engage in social loafing, social psychologists have developed and tested various theoretical accounts. Karau and Williams (1993) present a review of this literature and develop a collective effort model, which integrates elements of expectancy value, social identity and self-validation theories, to explain social loafing. A meta-analysis of 78 studies shows that social loafing is robust across studies. Consistent with the prediction of the model, several variables are found to moderate social loafing. The following factors are of particular interests to a mechanism designer.

1. 1. *Evaluation potential*: Harkins (1987) and others show that social loafing can be reduced or sometimes eliminated when a participant's contribution is identifiable and evaluable. In a related public goods experiment, Andreoni and Petrie (2004) find a substantial increase (59 per cent) in contribution to public goods compared to the baseline of a typical VCM experiment, when both the amount of individual contribution and the (photo) identification of donors are revealed.
2. 2. *Task valence*: the collective effort model predicts that the individual tendency to engage in social loafing decreases as task valence (or perceived meaningfulness) increases.
3. 3. *Group valence and group-level comparison standards*: Social identity theory (Tajfel and Turner, 1986) suggests that 'individuals gain positive self-identity through the accomplishments of the groups to which they belong' (Karau and Williams, 1993, p. 686). Therefore, enhancing group cohesiveness or group identity might reduce or eliminate social loafing. In a closely related economics experiment, Eckel and Grossman (2005) use induced group identity to study the effects of varying strength of identity on cooperative behaviour in a repeated public goods game. They find that while cooperation is unaffected by simple and artificial group identity, actions designed to enhance group identity contribute to higher levels of cooperation. This stream of research suggests that high degrees of group identification may limit individual shirking and free riding in environments with a public good.

4. 4. *Expectation of co-worker performance influences individual effort*. This set of theories might be sensitive to individual valuations for the public good as well as the public goods production functions. The meta-analysis indicates that individuals loafed when they expected their co-workers to perform well, but did not loaf otherwise.
5. 5. *Uniqueness of individual inputs*: individuals loafed when they believed that their inputs were redundant, but did not loaf when they believe that their individual inputs to the collective product were unique. In an interesting application, Beenen et al. (2004) conducted a field experiment in an online community called MovieLens. They found that users who were reminded of the uniqueness of their contributions rated significantly more movies than the control group.
6. 6. *Task complexity*: individuals were more likely to loaf on simple tasks, but less likely on complex tasks. This finding might be related to increased interests when solving complex tasks.

Exploring non-monetary incentives to increase contribution to public goods is an important and promising direction for future research. Mathematical models of social psychology theories are likely to shed insights on the necessary and sufficient conditions for a reduction or even elimination of social loafing.

Matching

Matching theory has been credited as ‘one of the outstanding successful stories of the theory of games’ (Aumann, 1992). It has been used to understand existing markets and to guide the design of new markets or allocation mechanisms in a variety of real world contexts. Matching experiments serve two purposes: to test new matching algorithms in the laboratory before implementing them in the real world, and to understanding how existing institutions evolved. We focus on one-sided matching experiments, and refer the reader to matching and market design for a summary of the two-sided matching experiments.

One-sided matching is the assignment of indivisible items to agents without a medium of exchange, such as money. Examples include the assignment of college students to dormitory rooms and public housing units, the assignment of offices and tasks to individuals, the assignment of students to public schools, the allocation of course seats to students (mostly in business schools and law schools), and timeshare exchange. The key mechanisms in this class of problems are the top trading cycles (TTC) mechanism (Shapley and Scarf, 1974), the Gale–Shapley deferred acceptance mechanism (Gale and Shapley, 1962), and variants of the serial dictatorship mechanism (Abdulkadiroglu and Sonmez, 1998). Matching experiments explore several issues. For strategy-proof mechanisms, they explore the extent to which subjects recognize and use their dominant strategies without prompting. For mechanisms which are not strategy-proof, they explore the extent of preference manipulation and the resulting efficiency loss. As a result, they examine the robustness of theoretical efficiency comparisons when the mechanisms are implemented among boundedly rational subjects and across different environments.

For the class of house allocation problems, two mechanisms have been compared and tested in the laboratory. The random serial dictatorship with squatting rights (RSD) is used by many US universities for on-campus housing allocation, while the TTC mechanism is theoretically superior. Chen and Sonmez (2002) report the first experimental study of these two mechanisms. They find that TTC is significantly more efficient than RSD because it induces significantly higher participation rate of existing tenants.

Another application of one-sided matching is the time-share problem. Wang and Krishna (2006) study the top trading cycles chains and spacebank mechanism (TTCCS), and two status quo mechanisms in the time-share industry, that is, the deposit first mechanism and the request first mechanism, neither of which is efficient. In the experiment, the observed efficiency of TTCCS is significantly higher than that of the deposit first mechanism, which in turn, is more efficient than the request first mechanism. In fact, efficiency under TTCCS converged to 100 per cent quickly, while the other two mechanisms do not show any increase in efficiency over time.

More recently, the school choice problem has received much attention. We review two experimental studies. Chen and Sonmez (2006) present an experimental study of three school choice mechanisms. The Boston mechanism is influential in practice, while the Gale–Shapley and TTC mechanisms have superior theoretical properties. Consistent with theory, this study indicates a high preference manipulation rate under Boston. As a result, efficiency under Boston is significantly lower than that of the two competing mechanisms in the designed environment. However, contrary to theory, Gale–Shapley outperforms TTC and generates the highest efficiency. The main reason is that a much higher proportion of subjects did not realize that truth-telling was a dominant strategy under TTC, and thus manipulated their preferences and ended up worse off. While Chen and Sonmez (2006) examine these mechanisms under partial information, where an agent only knows his own preference ranking, and not those of other agents, a follow-up study by Pais and Pinter (2006), investigates the same three mechanisms under different information conditions, ranging from complete ignorance about the other participants’ preferences and school priorities to complete information on all elements of the game. They show that information condition has a significant effect on the rate of truthful preference revelation. In particular, having no information results in a significantly higher proportion of truth-telling than under any treatment with additional information. Interestingly, there is no significant difference in the efficiency between partial and full information treatments. Unlike Chen and Sonmez (2006), in this experiment, TTC outperforms in terms of efficiency. Furthermore, TTC is also less sensitive to the amount of information in the environment.

Owing to their important applications in the real world, one-sided matching experiments provide insights on the actual manipulability of the matching mechanisms which are valuable in their real world implementations. Some issues, such as the role of information on the performance of the mechanisms, remain open questions.

Combinatorial auctions

In many applications of mechanism design, theory is not yet up to the task of identifying the optimal design or even comparing alternative designs. One case in which this has been true is in the design of auctions to sell collections of heterogeneous items with value complementarities, which occur when the value of a combination of items can be higher than the sum of the values for separate items. Value complementarities arise naturally in many contexts, such as broadcast spectrum rights auctioned by the Federal Communications Commission, pollution emissions allowances for consecutive years bought and sold under the RECLAIM programme of the South Coast Air Quality Management District in Los Angeles, aircraft take-off and landing slots, logistics services, and advertising time slots. Because individuals may want to express bids for combinations of the items for sale, requiring up to 2^N bids per person when there are N items, these auctions have come to be known as combinatorial auctions.

As was discussed earlier under public goods mechanisms, theory has identified the VCG mechanism as the unique auction design that would implement an efficient allocation assuming bidders use dominant strategies. Theory has not yet identified the revenue-maximizing combinatorial auction, although Ledyard (2007) shows that it is not the VCG mechanism. Theory has also been of little use in comparing the expected revenue collection between different auction designs. This has opened the way for many significantly different auction designs to be proposed, and sometimes even deployed, with little evidence to back up various claims of superiority.

To give some idea of the complexity of the problem we describe just some of the various design choices one can make. Should the auctions be run as a sealed bid or should some kind of iterative procedure be used? And, if the latter, should iteration be synchronous or asynchronous? What kinds of bids should be allowed? Proposals for allowable bids include only bids for a single item, bids for any package, and some which allow only a limited list of packages to be bid on. What stopping rule should be used? Proposals have included fixed stopping times, stop after an iteration in which revenue does not increase by more than x per cent, stop if demand is less than or equal to supply, and an imaginative but complex system of eligibility and activity rules created for the Federal Communications Commission (FCC) auctions. Should winners pay what they bid or something else? Alternatives to pay what you bid include VCG prices and second-best prices based on the dual variables to the programme that picks the provisional winners. What should bidders be told during the auction? Some designs provide information on all bids and provisional winners and the full identity of the bidders involved in them. Some designs provide minimal information such as only the winning bids without even the information as to who made them. The permutations and combinations are many. Because theory has not developed enough to sort out what is best, experiments have been used to provide some evidence.

The very first experimental analysis of a combinatoric auction can be found in Rassenti, Smith and Bulfin (1982), where they compared a sealed bid auction (RSB) allowing package bids to a uniform price sealed bid auction (GIP), proposed by Grether, Isaac and Plott (1981), that did not allow package bids. Both designs included a double auction market for re-trading after the auction results were known. The RSB design yielded higher efficiencies than the GIP design. Banks, Ledyard and Porter (1989) compared a continuous, asynchronous design (AUSM), a generalization of the English auction with package bidding, to a synchronous iterative design with myopic VCG pricing and found AUSM to yield higher efficiencies and revenues on average. Ledyard, Porter and Rangel (1997) compare the continuous AUSM to a synchronous iterative design (SMR) developed by Millgrom (2000) for the FCC auctions, which only allowed simultaneous single item bids. The testing found that ASUM yielded significantly higher efficiencies and revenues. Kwasnica et al. (2005) compare an iterative design (RAD) with package bidding and price feedback to both AUSM and SMR. RAD and SMR use the same stopping rule. Efficiencies observed with RAD and AUSM are similar and higher than those for SMR, but revenue is higher in SMR since many bidders lose money due to a phenomenon known as the exposure problem, which is identified in Bykowsky, Cull and Ledyard (2000). If it is assumed that bidders default on bids on which they make losses and thus set the prices of such bids to zero, revenues are in fact higher under AUSM and RAD than under SMR. At the behest of the FCC, Banks et al. (2003) ran an experiment to compare an iterative, package bidding design (CRA) from Charles River Associates and Market Design (1998) with the FCC SMR auction format. They also found that the package bidding design provides more efficient allocations but less revenue, due to bidder losses in the SMR.

Parkes and Unger (2000) proposed an ascending price, generalized VCG auction (iBEA) that maintains nonlinear and non-anonymous prices on packages, and charges VCG prices to the winners. The design would theoretically produce efficient allocations as long as bidders bid in a straightforward manner. Straightforward bidding is myopic and non-strategic and involves bidding on packages that yield the locally highest payoff in utility. There is no evidence that actual bidders will behave this way. Chen and Takeuchi (2005) have experimentally tested iBEA against the VCG sealed bid auction and found that VCG was superior in both revenue generation and efficiency attained. Takeuchi et al. (2006) tested RAD against VCG and found that RAD generated higher efficiencies, especially in the earlier auctions. They were using experiments to test combinatoric auctions as a potential alternative to scheduling processes in situations with valuation complementarities. In many cases current procedures request orderings from users and then employ a knapsack algorithm of some kind to choose good allocations without any concern for incentive compatibility. Takeuchi et al. (2006) find that both RAD and VCG yield higher efficiencies than the knapsack approach. Ledyard, Porter and Noussair (1996) found similar results when comparing a more vanilla combinatoric auction to an administrative approach. These findings suggest there are significant improvements in organization performance being overlooked by management.

Porter et al. (2003) proposed and tested a combinatorial clock (CC) auction. After bids are submitted, a simple algorithm determines the demand for each item by each bidder and for those items that have more than one bidder demanding more units than are available the clock price is raised. They test their design against the SMR and CRA. They do not report revenue but in their tests the CC design attained an almost perfect average efficiency of 99.9 per cent. CRA attained an average of 93 per cent, while SMR attained only 88 per cent.

Brunner et al. (2006) have carried out a systematic comparison of SMR and three alternatives, CC, RAD and a new FCC design called SMRPB, which takes the basic RAD design and changes two things. SMRPB allows bidders to win at most one package and the pricing feedback rule includes some inertia that RAD does not. They find that in terms of efficiency RAD is better than CC which is equivalent to SMRPB which is better than SMR. In terms of revenue, they find CC is better than RAD which is better than SMRPB which is better than SMR.

Most of these papers compare only two or three auction designs at a time and the environments used as the basis for comparison is often different in different papers. Further, environments can often be chosen that favour one auction over another. To deal with this, many research teams stress test their results by looking at boundary environments' collections of payoff parameters that give each auction under examination its best or worst chance of yielding high revenue or efficiency. But it is still unusual for a research team to report on a comparative test of several auctions in which their own design ends up being out-performed by another. Nevertheless, there are some tentative conclusions one can draw from this research.

The easiest and most obvious conclusion is that allowing package bidding improves both efficiency and revenue. In all the studies listed, anything that limits bidders' ability to express the full extent of their willingness to pay for all packages does interfere with efficiency and revenue. Less obvious but also easy to see is that simultaneity and iteration are also good design features. Bidding in situations in which value complementarities exist can be difficult since bidders need to discover where their willingness to pay is more than others but also where they fit with others interests. Getting this right improves both efficiency and revenue. Iteration and relevant price feedback both help here. Stopping rules also matter. Although this is an area that could benefit from more research, it is clear that in many cases complicated stopping rules that allow auctions to proceed for very long periods of time provide little gain in revenue or efficiency.

Summary

Mechanism design experiments identify features of mechanisms that lead to good performance when they are implemented among real people. Experiments testing public goods mechanisms show that dominant strategy mechanisms should also be secure, while Nash-efficient mechanisms should satisfy dynamic stability if it is to be considered for application in the real world in a repeated interaction setting. For matching mechanisms, transparency of the dominant strategy leads to better performance in the laboratory. Lastly, in combinatorial auctions, package bidding, simultaneity and iteration are shown to be good design features. In addition to the three domains covered in this article, there has been a growing experimental literature on market design, information markets and contract theory. We do not cover them in this article, due to lack of robust empirical regularities. However, they are excellent areas in which to make a new contribution.

See Also

- computing in mechanism design
- matching and market design
- mechanism design
- public goods

Bibliography

Abdulkadiroglu, A. and Sönmez, T. 1998. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66, 689–701.

Andreoni, J. and Petrie, R. 2004. Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of Public Economics* 88, 1605–23.

Arifovic, J. and Ledyard, J. 2006. Computer testbeds and mechanism design: application to the class of Groves-Ledyard mechanisms for provision of public goods. Caltech working paper. Pasadena, CA.

Attiyeh, G., Franciosi, R. and Isaac, M. 2000. Experiments with the pivot process for providing public goods. *Public Choice* 102, 95–114.

Aumann, R. 1992. Foreword. In *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, ed. E. Alvin, M.A. Roth, O. Sotomayor. Cambridge: Cambridge University Press.

Bagnoli, M. and Lipman, B. 1989. Provision of public goods: fully implementing the core through private contributions. *Review of Economic Studies* 56, 583–602.

Banks, J.S., Ledyard, J.O. and Porter, D.P. 1989. Allocating uncertain and unresponsive resources: an experimental approach. *Rand Journal of Economics* 20, 1–25.

Banks, J., Olson, M., Porter, D., Rassenti, S. and Smith, V. 2003. Theory, experiment and the federal communications commission spectrum auctions. *Journal of Economic Behavior and Organization* 51, 303–50.

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P. and Kraut, R. 2004. In *Proceedings of ACM Computer Supported Cooperative Work 2004*, Conference on Computer Supported Cooperative Work. Chicago, IL: ACM.

Brunner, C., Goeree, J., Holt, C. and Ledyard, J. 2006. Combinatorial auctioneering, Caltech working paper. Pasadena, CA.

Bykowsky, M., Cull, R. and Ledyard, J. 2000. Mutually destructive bidding: the FCC auction design problem. *Journal of Regulatory Economics* 17, 205–28.

Cason, T., Saijo, T., Sjöström, T. and Yamato, T. 2006. Secure implementation experiments: do strategy-proof mechanisms really work? *Games and Economic Behavior* 57, 206–35.

Charles River Associates Inc. and Market Design Inc. 1998. Report No. 1351–00.

Chen, Y. 2002. A family of supermodular Nash mechanisms implementing Lindahl allocations. *Economic Theory* 19, 773–90.

Chen, Y. 2008. Incentive-compatible mechanisms for pure public goods: a survey of experimental literature. In *The Handbook of Experimental Economics Results*, ed. D. Plott and V. Smith. Amsterdam: Elsevier.

Chen, Y. and Gazzale, R. 2004. Supermodularity and convergence: an experimental study of the compensation mechanism. *American Economic Review* 94, 1505–35.

Chen, Y. and Plott, C.R. 1996. The Groves–Ledyard mechanism: an experimental study of institutional design. *Journal of Public Economics* 59, 335–64.

Chen, Y. and Sonmez, T. 2002. Improving efficiency of on-campus housing: an experimental study. *American Economic Review* 92, 1669–86.

Chen, Y. and Sonmez, T. 2006. School choice: an experimental study. *Journal of Economic Theory* 127, 202–31.

Chen, Y. and Tang, F.-F. 1998. Learning and incentive compatible mechanisms for public goods provision: an experimental study. *Journal of Political Economy* 106, 633–62.

Chen, Y. and Takeuchi, K. 2005. Multi-object auctions with package bidding: an experimental comparison of Vickrey and iBEA. Working paper.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11, 17–33.

Corchon, L. and Wilkie, S. 1996. Double implementation of the ratio correspondence by a market mechanism. *Review of Economic Design* 2, 325–37.

Eckel, C. and Grossman, P. 2005. Managing diversity by creating team identity. *Journal of Economic Behavior & Organization* 58, 371–92.

Falkinger, J. 1996. Efficient private provision of public goods by rewarding deviations from average. *Journal of Public Economics* 62, 413–22.

Falkinger, J., Fehr, E., Gächter, S. and Winter-Ebmer, R. 2000. A simple mechanism for the efficient provision of public goods: experimental evidence. *American Economic Review* 90, 247–64.

Gale, D. and Shapley, L. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69, 9–15.

Green, J. and Laffont, J.-J. 1977. Characterization of satisfactory mechanisms for the revelation of the preferences for public goods. *Econometrica* 45, 427–38.

Grether, D., Isaac, M. and Plott, C. 1981. The allocation of landing rights by unanimity among competitiors. *American Economic Review* 71, 166–71.

Groves, T. 1973. Incentives in Teams. *Econometrica* 41, 617–31.

Groves, T. and Ledyard, J. 1977. Optimal allocation of public goods: a solution to the ‘free rider’ problem. *Econometrica* 45, 783–809.

Groves, T. and Ledyard, J. 1987. Incentive compatibility since 1972. In *Essays in Honor of Leonid Hurwicz*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Harkins, S.G. 1987. Social loafing and social facilitation. *Journal of Experimental Social Psychology* 23, 1–18.

Harstad, R.M. and Marrese, M. 1981. Implementation of mechanism by processes: public good allocation experiments. *Journal of Economic Behavior & Organization* 2, 129–51.

Harstad, R.M. and Marrese, M. 1982. Behavioral explanations of efficient public good allocations. *Journal of Public Economics* 19, 367–83.

Healy, P.J. 2006. Learning dynamics for mechanism design: an experimental comparison of public goods mechanisms. *Journal of Economic Theory* 129, 114–49.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and Organization*, ed. C. McGuire and R. Radner. Amsterdam: North-Holland.

Hurwicz, L. 1975. On the existence of allocation systems whose manipulative Nash equilibria are Pareto-optimal. Paper presented at Third World Congress of the Econometric Society, Toronto.

Hurwicz, L. 1979. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *Review of Economic Studies* 46, 217–25.

Isaac, R. and James, D. 2000. Robustness of the incentive compatible combinatorial auction. *Experimental Economics* 3, 31–53.

Jackson, M. and Moulin, H. 1992. Implementing a public project and distributing its cost. *Journal of Economic Theory* 57, 125–40.

Kaneko, M. 1977. The ratio equilibria and the core of the voting game in a public goods economy. *Econometrica* 45, 1589–94.

Kauru, S.J. and Williams, K.D. 1993. Social loafing: a meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology* 65, 681–706.

Kawagoe, T. and Mori, T. 2001. Can pivotal mechanism induce truth-telling? An experimental study. *Public Choice* 108, 331–54.

Kim, T. 1986. On the nonexistence of a stable Nash mechanism implementing Lindahl allocations. Mimeo. University of Minnesota.

Kim, T. 1993. A stable Nash mechanism implementing Lindahl allocations for quasi-linear environments. *Journal of Mathematical Economics* 22, 359–71.

Kwasnica, A.M., Ledyard, J.O., Porter, D. and DeMartini, C. 2005. A new and improved design for multi-object iterative auctions. *Management Science* 51, 419–34.

Ledyard, J. 2007. Optimal combinatoric auctions with single-minded bidders. *Proceedings of the 8th ACM Conference on Electronic Commerce*. San Diego, CA: ACM.

Ledyard, J., Olson, M., Porter, D., Swanson, J. and Torma, D. 2002. The first use of a combined value auction for transportation services. *Interfaces* 32, 4–12.

Ledyard, J. and Palfrey, T. 1994. Voting and lottery drafts as efficient public goods mechanisms. *Review of Economic Studies* 61, 327–55.

Ledyard, J., Porter, D. and Noussair, C. 1996. The allocation of a shared resource within an organization. *Economic Design* 2, 163–92.

Ledyard, J., Porter, D. and Rangel, A. 1997. Experiments testing multiobject allocation mechanisms. *Journal of Economics and Management Strategy* 6, 639–75.

Mailath, G. and Postlewaite, A. 1990. Asymmetric information bargaining problems with many agents. *Review of Economic Studies* 57, 351–67.

McAfee, P.R. and McMillan, J. 1996. Analyzing the airwaves auction. *Journal of Economic Perspectives* 10(1), 159–75.

McCabe, K., Rassenti, S. and Smith, V. 1989. Designing ‘Smart’ computer assisted markets. *European Journal of Political Economy* 5, 259–83.

Millgrom, P. 2000. Putting auction theory to work: the simultaneous ascending auction. *Journal of Political Economy* 108, 245–272a.

Milgrom, P. and Roberts, J. 1990. Rationalizability, learning and equilibrium in games with strategic complementarities. *Econometrica* 58, 1255–77.

Milgrom, P. and Shannon, C. 1994. Monotone comparative statics. *Econometrica* 62, 157–80.

Mori, T. 1989. Effectiveness of mechanisms for public goods provision: an experimental study. *Economic Studies* 40, 234–46.

Pais, J. and Pintér, A. 2006. School choice and information: an experimental study on matching mechanisms. Working Paper, Institute for Economics and Business Administration (ISEG). Lisbon: Technical University.

Parkes, D. and Unger, L. 2000. Iterative combinatorial auctions: theory and practice. In *Proceedings of the 17th National Conference on Artificial Intelligence* (AAAI-00).

Peleg, B. 1996. Double implementation of the Lindahl equilibrium by a continuous mechanism. *Economic Design* 2, 311–24.

Porter, D.P. 1999. The effect of bid withdrawal in a multi-object auction. *Review of Economic Design* 4, 73–97.

Porter, D., Rassenti, S., Roopnarine, A. and Smith, V. 2003. Combinatorial auction design. *Proceedings of the National Academy of Sciences* 100, 11153–7.

Rassenti, S., Smith, V. and Bulfin, R. 1982. A combinatorial auction mechanism for airport time slot allocation. *Bell Journal of Economics* 13, 402–17.

Roberts, J. 1979. Incentives and planning procedures for the provision of public goods. *Review of Economic Studies* 46, 283–92.

Saijo, T., Sjöström, T. and Yamato, T. 2005. Secure Implementation. Working Paper, No. 567-0047. Osaka: Institute of Social and Economic Research, Osaka University.

Shapley, L.S. and Scarf, H. 1974. On cores and indivisibilities. *Journal of Mathematical Economics* 1, 23–37.

Smith, V. 1979a. Incentive compatible experimental processes for the provision of public goods. In *Experimental Economics*, vol. 1, ed. R. Smith. Greenwich, CT: JAI Press.

Smith, V. 1979b. An experimental comparison of three public goods decision mechanisms. *Scandinavian Journal of Economics* 81, 198–251.

Tajfel, H. and Turner, J.C. 1986. The social identity theory of intergroup behaviour. In *Psychology of Intergroup Relations*, ed. S. Worchel and W. Austin. Chicago: Nelson-Hall.

Takeuchi, K., Lin, J., Chen, Y. and Finholt, T. 2006. Shake it up baby: scheduling with package auctions. Working paper. School of Information, University of Michigan.

Tian, G. 1989. Implementation of the Lindahl correspondence by a single-valued, feasible, and continuous mechanism. *Review of Economic Studies* 56, 613–21.

Varian, H. 1994. A solution to the problems of externalities when agents are well-informed. *American Economic Review* 84, 1278–93.

Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16, 8–37.

Walker, M. 1980. On the impossibility of a dominant strategy mechanism to optimally decide public questions. *Econometrica* 48, 1521–40.

Walker, M. 1981. A simple incentive compatible scheme for attaining Lindahl allocations. *Econometrica* 49, 65–71.

Wang, Y. and Krishna, A. 2006. Timeshare exchange mechanisms. *Management Science* 52, 1123–37.

How to cite this article

Chen, Yan and John O. Ledyard. "mechanism design experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000404> doi:10.1057/9780230226203.1080

close print preview

The New Palgrave Dictionary of Economics Online

mechanism design

Roger B. Myerson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A mechanism is a specification of how economic decisions are determined as a function of the information that is known by the individuals in the economy. Mechanism theory shows that *incentive constraints* should be considered coequally with *resource constraints* in the formulation of the economic problem. Where individuals' private information and actions are difficult to monitor, the need to give people an incentive to share information and exert efforts may impose constraints on economic systems just as much as the limited availability of raw materials. Mechanism design is the fundamental mathematical methodology for analysing economic efficiency subject to incentive constraints.

Keywords

adverse selection; Bayesian collective-choice problems; Bayesian equilibrium; bilateral bargaining; Coase Theorem; cooperative game theory; correlated equilibrium; direct-revelation mechanisms; first-price auctions; incentive constraints; incentive efficiency; incentive-compatible mechanisms; incomplete information; inscrutability principle; mechanism design; Pareto efficiency; private information; revelation principle; revenue equivalence theorems; sealed-bid auctions

Article

Overview

A mechanism is a specification of how economic decisions are determined as a function of the information that is known by the individuals in the economy. In this sense, almost any kind of market institution or economic organization can be viewed, in principle, as a mechanism. Thus mechanism theory can offer a unifying conceptual structure in which a wide range of institutions can be compared, and optimal institutions can be identified.
The basic insight of mechanism theory is that *incentive constraints* should be considered coequally with *resource constraints* in the formulation of the economic problem. In situations where individuals' private information and actions are difficult to monitor, the need to give people an incentive to share information and exert efforts may impose constraints on economic systems just as much as the limited availability of raw materials. The theory of mechanism design is the fundamental mathematical methodology for analysing these constraints.

The study of mechanisms begins with a special class of mechanisms called *direct-revelation* mechanisms, which operate as follows. There is assumed to be a mediator who can communicate separately and confidentially with every individual in the economy. This mediator may be thought of as a trustworthy person, or as a computer tied into a telephone network. At each stage of the economic process, each individual is asked to report all of his private information (that is, everything that he knows that other individuals in the economy might not know) to the mediator. After receiving these reports confidentially from every individual, the mediator may then confidentially recommend some action or move to each individual. A direct-revelation mechanism is any rule for specifying how the mediator's recommendations are determined, as a function of the reports received.

A direct-revelation mechanism is said to be *incentive compatible* if, when each individual expects that the others will be honest and obedient to the mediator, then no individual could ever expect to do better (given the information available to him) by reporting dishonestly to the mediator or by disobeying the mediator's recommendations. That is, if honesty and obedience is an equilibrium (in the game-theoretic sense), then the mechanism is incentive compatible.

The analysis of such incentive-compatible direct-revelation mechanisms might at first seem to be of rather narrow interest, because such fully centralized mediation of economic systems is rare, and incentives for dishonesty and disobedience are commonly observed in real economic institutions. The importance of studying such mechanisms is derived from two key insights: (i) for any equilibrium of any general mechanism, there is an incentive-compatible direct-revelation mechanism that is essentially equivalent; and (ii) the set of incentive-compatible direct-revelation mechanisms has simple mathematical properties that often make it easy to characterize, because it can be defined by a set of linear inequalities. Thus, by analysing incentive-compatible direct-revelation mechanisms, we can characterize what can be accomplished in all possible equilibria of all possible mechanisms, for a given economic situation.

Insight (i) above is known as the *revelation principle*. It was first recognized by Gibbard (1973), but for a somewhat narrower solution concept (dominant strategies, instead of Bayesian equilibrium) and for the case where only informational honesty is problematic (no moral hazard). The formulation of the revelation principle for the broader solution concept of Bayesian equilibrium, but still in the case of purely informational problems, was recognized independently by many authors around 1978 (see Dasgupta, Hammond and Maskin, 1979; Harris and Townsend, 1981; Holmstrom, 1977; Myerson, 1979; Rosenthal, 1978). Aumann's (1974; 1987) concept of *correlated equilibrium* gave the first expression to the revelation principle in the case where only obedient choice of actions is problematic (pure moral hazard, no adverse selection). The synthesis of the revelation principle for general Bayesian games with incomplete information, where both honesty and obedience are problematic, was given by Myerson (1982). A generalization of the revelation principle to multistage games was stated by Myerson (1986).

The intuition behind the revelation principle is as follows. First, a central mediator who has collected all relevant information known by all individuals in the economy could issue recommendations to the individuals so as to simulate the outcome of any organizational or market system, centralized or decentralized. After the individuals have revealed all of their information to the mediator, he can simply tell them to do whatever they would have done in the other system. Second, the more information that an individual has, the harder it may be to prevent him from finding ways to gain by disobeying the mediator. So the incentive constraints will be least binding when the mediator reveals to each individual only the minimal information needed to identify his own recommended action, and nothing else about the reports or recommendations of other individuals. So, if we assume that the mediator is a discrete and trustworthy information-processing device, with no costs of processing information, then there is no loss of generality in assuming that each individual will confidentially reveal all of his information to the mediator (maximal revelation to the trustworthy mediator), and the mediator in return will reveal to each individual only his own recommended action (minimal revelation to the individuals whose behaviour is subject to incentive constraints).

The formal proof of the revelation principle is difficult only because it is cumbersome to develop the notation for defining, in full generality, the set of all general mechanisms, and for defining equilibrium behaviour by the individuals in any given mechanism. Once all of this notation is in place, the construction of the equivalent incentive-compatible direct-revelation mechanism is straightforward. Given any mechanism and any equilibrium of the mechanism, we simply specify that the mediator's recommended actions are those that would result in the given mechanism if everyone behaved as

specified in the given equilibrium when his actual private information was as reported to the mediator. To check that this constructed direct-revelation mechanism is incentive compatible, notice that any player who could gain by disobeying the mediator could also gain by similarly disobeying his own strategy in the given equilibrium of the given mechanism, which is impossible (by definition of equilibrium).

Mathematical formulations

Let us offer a precise general formulation of the proof of the revelation principle in the case where individuals have private information about which they could lie, but there is no question of disobedience of recommended actions or choices. For a general model, suppose that there are n individuals, numbered 1 to n. Let C denote the set of all possible combinations of actions or resource allocations that the individuals may choose in the economy. Each individual in the economy may have some private information about his preferences and endowments, and about his beliefs about other individuals' private information. Following Harsanyi (1967), we may refer to the state of an individual's private information as his *type*. Let T_i denote the set of possible types for any individual i, and let $T=T_i\times\ldots\times T_n$ denote the set of all possible combinations of types for all individuals.

The preferences of each individual i may be generally described by some *payoff function* $u_i: C\times T\to\mathbb{R}$, where $u_i(c,(t_i,\ldots,t_n))$ denotes the payoff, measured in some von Neumann–Morgenstern utility scale, that individual i would get if c was the realized resource allocation in C when $(t_i,\ldots,t_n)$ denotes the actual types of the individuals $1,\ldots,n$ respectively. For short, we may write $t=(t_i,\ldots,t_n)$ to describe a combination of types for all individuals.

The beliefs of each individual i, as a function of his type, may be generally described by some function $p_i(\cdot|\cdot)$, where $p_i,(t_1,\ldots,t_{i-1}, t_{i+1},\ldots,t_n|t_i)$ denotes the probability that individual i would assign to the event that the other individuals have types as in $(t_1,\ldots,t_{i-1}, t_{i+1},\ldots,t_n)$, when i knows that his own type is t_i. For short, we may write $t_{-i}=(t_1,\ldots,t_{i-1}, t_{i+1},\ldots,t_n)$, to describe a combination of types for all individuals other than i. We may let $T_{-i}=T_1\times\ldots\times T_{i-1}\times T_{i+1}\times\ldots\times T_n$ denote the set of all possible combinations of types for the individuals other than i.

The general model of an economy defined by these structures $(C, T_1,\ldots,T_n, u_1,\ldots,u_n, p_1,\ldots,p_n)$ is called a Bayesian collective-choice problem.

Given a Bayesian collective-choice problem, a general mechanism would be any function of the form $\gamma: S_1\times\ldots\times S_n,\to C$, where, for each i, S_i is a nonempty set that denotes the set of strategies that are available for individual i in this mechanism. That is, a general mechanism specifies the strategic options that each individual may choose among, and the social choice or allocation of resources that would result from any combination of strategies that the individuals might choose. Given a mechanism, an equilibrium is any specification of how each individual may choose his strategy in the mechanism as a function of his type, so that no individual, given only his own information, could expect to do better by unilaterally deviating from the equilibrium. That is, $\sigma=(\sigma_1,\ldots,\sigma_n)$ is an equilibrium of the mechanism γ if, for each individual i, σ_i is a function from T_i to S_i, and, for every t_i in T_i and every s_i in S_i,

$$\Sigma_{t_{-i}\in T_{-i}}\ p_i(t_{-i}|t_i)u_i(\gamma(\sigma(t)),\ t)\ \geq \Sigma_{t_{-i}\in T_{-i}}\ p_i(t_{-i}|t_i)u_i(\gamma(\sigma_{-i}(t_{-i}),\ s_i),\ t).$$

(Here $\sigma(t)=(\sigma_1(t_1),\ldots,\sigma_n(t_n))$ and $(\sigma_{-i}(t_{-i}), s_i)=(\sigma_1(t_1),\ldots,\sigma_{i-1}(t_{i-1}), s_i, \sigma_{i+1}(t_{i+1}),\ldots,\sigma_n(t_n))$.) Thus, in an equilibrium σ, no individual i, knowing only his own type t_i, could increase his expected payoff by changing his strategy from $\sigma_i(t_i)$ to some other strategy s_i, when he expects all other individuals to behave as specified by the equilibrium σ. (This concept of equilibrium is sometimes often called *Bayesian equilibrium* because it respects the assumption that each player knows only his own type when he chooses his strategy in S_i. For a comparison with other concepts of equilibrium, see

Dasgupta, Hammond and Maskin, 1979, and Palfrey and Srivastava, 1987.)

In this context, a direct-revelation mechanism is any mechanism such that the set S_i of possible strategies for each player i is the same as his set of possible types T_i. A direct-revelation mechanism is (Bayesian) incentive-compatible iff it is an equilibrium (in the Bayesian sense defined above) for every individual always to report his true type. Thus, $\mu : T_1 \times \ldots \times T_n \rightarrow C$ is an incentive-compatible direct-revelation mechanism if, for each individual i and every pair of types t_i and r_i in T_i,

$$\Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t), t) \geq \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t_{-i}, r_i), t).$$

(Here $(t_{-i}, r_i)=(t_1,\ldots,t_{i-1}, r_i, t_{i+1},\ldots,t_n)$.) We may refer to these constraints as the *informational incentive constraints* on the direct-revelation mechanism μ. These informational incentive constraints are the formal representation of the economic problem of *adverse selection*, so they may also be called adverse-selection constraints (or self-selection constraints).

Now, to prove the revelation principle, given any general mechanism γ and any Bayesian equilibrium σ of the mechanism γ, let μ be the direct-revelation mechanism μ defined so that, for every t in T,

$$\mu(t) = \gamma(\sigma(t)).$$

Then this mechanism μ always leads to the same social choice as γ does, when the individuals behave as in the equilibrium σ. Furthermore, μ is incentive compatible because, for any individual i and any two types t_i and r_i in T_i,

$$\Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t), t) = \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\gamma(\sigma(t)), t) \geq \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\gamma(\sigma_{-i}(t_{-i}), \sigma_i(r_i)), t) = \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t_{-i}, r_i), t).$$

Thus, μ is an incentive-compatible direct-revelation mechanism that is equivalent to the given mechanism γ with its equilibrium σ.

Notice that the revelation principle asserts that any pair consisting of a mechanism *and* an equilibrium is equivalent to an incentive-compatible direct-revelation mechanism. Thus, a general mechanism that has several equilibria may correspond to several different incentive-compatible mechanisms, depending on which equilibrium is considered.

Furthermore, the same general mechanism will generally have different equilibria in the context of different Bayesian collective-choice problems, where the structure of individuals' beliefs and payoffs are different. For example, consider a first-price sealed-bid auction where there are five potential bidders who are risk-neutral with independent private values drawn from the same distribution over $0 to $10. If the bidders' values are drawn from a uniform distribution over this interval, then there is an equilibrium in which each bidder bids 4/5 of his value. On the other hand, if the bidders' values are drawn instead from a distribution with a probability density that is proportional to the square of the value, then there is an equilibrium in which each bidder bids

8/9 of his value. So in one situation the first-price sealed-bid auction (a general mechanism) corresponds to an incentive-compatible mechanism in which the bidder who reports the highest value gets the object for 4/5 of his reported value; but in the other situation it corresponds to an incentive-compatible mechanism in which the bidder who reports the highest value gets the object for 8/9 of his reported value. There is no incentive-compatible direct-revelation mechanism that is equivalent to the first-price sealed-bid auction in all situations, independently of the bidders' beliefs about each others' values. Thus, if we want to design a mechanism that has good properties in the context of many different Bayesian collective-choice problems, we cannot necessarily restrict our attention to incentive-compatible direct-revelation mechanisms, and so our task is correspondingly more difficult. (See Wilson, 1985, for a remarkable effort at this kind of difficult question.)

Even an incentive-compatible mechanism itself may have other dishonest equilibria that correspond to different incentive-compatible mechanisms. Thus, when we talk about selecting an incentive-compatible mechanism and assume that it will then be played according to its honest equilibrium, we are implicitly making an assumption about the selection of an equilibrium as well as of a mechanism or communication structure. Thus, for example, when we say that a particular incentive-compatible mechanism maximizes a given individual's expected utility, we mean that, if you could choose any general mechanism for coordinating the individuals in the economy and if you could also (by some public statement, as a focal arbitrator, using Schelling's, 1960, *focal-point effect*) designate the equilibrium that the individuals would play in your mechanism, then you could not give this given individual a higher expected utility than by choosing this incentive-compatible mechanism and its honest equilibrium.

In many situations, an individual may have a right to refuse to participate in an economic system or organization. For example, a consumer generally has the right to refuse to participate in any trading scheme and instead just consume his initial endowment. If we let $w_i(t_i)$ denote the utility payoff that individual i would get if he refused to participate when his type is t_i, and if we assume that an individual can make the choice not to participate after learning his type, then an incentive-compatible mechanism μ must also satisfy the following constraint, for every individual i and every possible type t_i:

$$\Sigma_{t_{-i}\in T_{-i}}\ p_i(t_{-i}|t_i)\,u_i(\mu(t),\,t)\ \geq\ w_i(t_i).$$

These constraints are called *participational incentive constraints*, or *individual-rationality* constraints.

In the analysis of Bayesian collective-choice problems, we have supposed that the only incentive problem was to get people to share their information, and to agree to participate in the mechanism in the first place. More generally, a social choice may be privately controlled by one or more individuals who cannot be trusted to follow some pre-specified plan when it is not in their best interests. For example, suppose now that the choice in C is privately controlled by some individual (call him 'individual 0') whose choice of an action in C cannot be regulated. To simplify matters here, let us suppose that this individual 0 has no private information. Let $p_0(t)$ denote the probability that this individual would assign to the event that $t=(t_1,\ldots,t_n)$ is the profile of types for the other n individuals, and let $u_0(c, t)$ denote the utility payoff that this individual receives if he chooses action c when t is the actual profile of types. Then, to give this active individual an incentive to obey the recommendations of a mediator who is implementing the direct-revelation mechanism μ, μ must satisfy

$$\Sigma_{t\in T}\ p_0(t)\,u_0(\mu(t),\,t)\ \geq\ \Sigma_{t\in T}\ p_0(t)\,u_0(\delta(\mu(t)),\,t)$$

for every function $\delta : C \rightarrow C$. These constraints assert that obeying the actions recommended by the mediator is better for this individual than any disobedient strategy δ under which he would choose $\delta(c)$ if the mediator recommended c. Such constraints are called *strategic incentive constraints* or *moral-hazard constraints*, because they are the formal representation of the economic problem of moral hazard.

For a formulation of general incentive constraints that apply when individuals both have private information and control private actions, see Myerson (1982) or (1985).

Applications

In general, the mechanism-theoretic approach to economic problems is to list the constraints that an incentive-compatible mechanism must satisfy, and to try to characterize the incentive-compatible mechanisms that have properties of interest.

For example, one early contribution of mechanism theory was the derivation of general *revenue equivalence* theorems in auction theory. Ortega-Reichert (1968) found that, when bidders are risk-neutral and have private values for the object being sold that are independent and drawn from the same distribution, then a remarkably diverse collection of different auction mechanisms all generate the same expected revenue to the seller, when bidders use equilibrium strategies. In all of these different mechanisms and equilibria, it turned out that the bidder whose value for the object was highest would always end up getting the object, while a bidder whose value for the object was zero would never pay anything. By analysing the incentive constraints, Harris and Raviv (1981), Myerson (1981) and Riley and Samuelson (1981) showed that all incentive-compatible mechanisms with these properties would necessarily generate the same expected revenue, in such economic situations.

Using methods of constrained optimization, the problem of finding the incentive-compatible mechanism that maximizes some given objective (one individual's expected utility, or some social welfare function) can be solved for many examples. The resulting optimal mechanisms often have remarkable qualitative properties.

For example, suppose a seller, with a single indivisible object to sell, faces five potential buyers or bidders, whose private values for the object are independently drawn from a uniform distribution over the interval from $0 to $10. If the objective is to maximize the sellers' expected revenue, optimal auction mechanisms exist and all have the property that the object is sold to the bidder with the highest value for it, except that the seller keeps the object in the event that the bidders' values are all less than $5. Such a result may seem surprising, because this event could occur with positive probability (1/32) and in this event the seller is getting no revenue in an 'optimal' auction, even though any bidder would almost surely be willing to pay him a positive price for the object. Nevertheless, no incentive-compatible mechanism (satisfying the participational and informational incentive constraints) can offer the seller higher expected utility than these optimal auctions, and thus no equilibrium of any general auction mechanism can offer higher expected revenue either. Maximizing expected revenue requires a positive probability of seemingly wasteful allocation.

The threat of keeping the object, when all bidders report values below $5, increases the seller's expected revenue because it gives the bidders an incentive to bid higher and pay more when their values are above $5. In many other economic environments, we can similarly prove the optimality of mechanisms in which seemingly wasteful threats are carried out with positive probability. People have intuitively understood that costly threats are often made to give some individual an incentive to reveal some information or choose some action, and the analysis of incentive constraints allows us to formalize this understanding rigorously.

In some situations, incentive constraints imply that such seemingly wasteful allocations may have to occur with positive probability in all incentive-compatible mechanisms, and so also in all equilibria of all general mechanisms. For example, Myerson and Satterthwaite (1983) considered bilateral bargaining problems between a seller of some object and a potential buyer, both of whom are risk-neutral and have independent private values for the object that are drawn out of distributions that have continuous positive probability densities over some pair of intervals that have an intersection of positive length. Under these technical (but apparently quite weak) assumptions, it is impossible to satisfy the participational and informational incentive constraints with any mechanism in which the buyer gets the object whenever it is worth more to him than to the seller. Thus, we cannot hope to guarantee

the attainment of full *ex post* efficiency of resource allocations in bilateral bargaining problems where the buyer and seller are uncertain about each other's reservation prices. If we are concerned with welfare and efficiency questions, it may be more productive to try to characterize the incentive-compatible mechanisms that maximize the expected total gains from trade, or that maximize the probability that a mutually beneficial trade will occur. For example, in the bilateral bargaining problem where the seller's and buyer's private values for the object are independent random variables drawn from a uniform distribution over the interval from $0 to $10, both of these objectives are maximized subject to incentive constraints by mechanisms in which the buyer gets the object if and only if his value is greater than the seller's value by $2.50 or more. Under such a mechanism, the event that the seller will keep the object when it is actually worth more to the buyer has probability 7/32, but no equilibrium of any general mechanism can generate a lower probability of this event.

The theory of mechanism design has fundamental implications about the domain of applicability of Coase's (1960) theorem, which asserts the irrelevance of initial property rights to efficiency of final allocations. The unavoidable possibility of failure to realize mutually beneficial trades, in such bilateral trading problems with two-sided uncertainty, can be interpreted as one of the ‘transaction costs’ that limits the validity of Coase's theorem. Indeed, as Samuelson (1985) has emphasized, reassignment of property rights generally changes the payoffs that individuals can guarantee themselves without selling anything, which changes the right-hand sides of the participational incentive constraints, which in turn can change the maximal social welfare achievable by an optimal incentive-compatible mechanism.

For example, consider again the case where there is one object and two individuals who have private values for the object that are independent random variables drawn from a uniform distribution over the interval from $0 to $10. When we assumed above that one was the ‘seller’, we meant that he had the right to keep the object and pay nothing to anyone, until he agreed to some other arrangement. Now, let us suppose instead that the rights to the object are distributed equally between the two individuals. Suppose that the object is a divisible good and each individual has a right to take half of the good and pay nothing, unless he agrees to some other arrangement. (Assume that, if an individual's value for the whole good is t_i, then his value for half would be $t_i/2$.) With this symmetric assignment of property rights, we can design incentive-compatible mechanisms in which the object always ends up being owned entirely by the individual who has the higher value for it, as Cramton, Gibbons and Klemperer (1987) have shown.

For example, consider the game in which each individual independently puts money in an envelope, and then the individual who put more money in his envelope gets the object, while the other individual takes the money in both envelopes. This game has an equilibrium in which each individual puts into his envelope an amount equal to one-third of his value for the whole good. This equilibrium of this game is equivalent to an incentive-compatible direct-revelation mechanism in which the individual who reports the higher value pays one-third of his value to buy out the other individual's half-share. This mechanism would violate the participational incentive constraints if one individual had a right to the whole good (in which case, for example, if his value were $10 then he would be paying $3.33 under this mechanism for a good that he already owned). But with rights to only half of the good, no type of either individual could expect to do better (at the beginning of the game, when he knows his own value but not the other's) by keeping his half and refusing to participate in this mechanism.

More generally, redistribution of property rights tends to reduce the welfare losses caused by incentive constraints when it creates what Lewis and Sappington (1989) have called *countervailing incentives*. In games where one individual is the seller and the other is the buyer, if either individual has an incentive to lie, it is usually because the seller wants to overstate his value or the buyer wants to understate his value. In the case where either individual may buy the other's half-share, neither individual can be sure at first whether he will be the buyer or the seller (unless he has the highest or lowest possible value). Thus, a buyer-like incentive to understate values, in the event where the other's value is lower, may help to cancel out a seller-like incentive to overstate values, in the event where the other's value is higher.

The theory of mechanism design can also help us to appreciate the importance of mediation in economic relationships and transactions. There are situations in which, if the individuals were required to communicate with each other only through perfect noiseless communication channels (for example, in face-to-face dialogue), then the set of all possible equilibria would be much smaller than the set of incentive-compatible mechanisms that are achievable with a mediator. (Of course, the revelation principle asserts that the former set cannot be larger than the latter.)

For example, consider the following ‘sender–receiver game’ due to J. Farrell. Player 1 has a privately known type that may be α or β, but he has no payoff-relevant action to choose. Player 2 has no private information, but he must choose an action from the set $\{x, y, z\}$. The payoffs to players 1 and 2 respectively depend on 1's type and 2's action as follows.

	x	y	z
α	2, 3	1, 2	0, 0
β	4, −3	8, −1	0, 0

At the beginning of the game, player 2 believes that each of 1's two possible types has probability 1/2.

Suppose that, knowing his type, player 1 is allowed to choose a message in some arbitrarily rich language, and player 2 will hear player 1's message (with no noise or distortion) before choosing his action. In every equilibrium of this game, including the randomized equilibria, player 2 must choose y with probability 1, after every message that player 1 may choose in equilibrium (see Farrell, 1993; Myerson, 1988). If there were some message that player 1 could use to increase the probability of player 2 choosing x (for example, ‘I am α, so choosing x would be best for us both!’), then he would always send such a message when his type was α. (It can be shown that no message could ever induce player 2 to randomize between x and z.) So not receiving such a message would lead 2 to infer that 1's type was β, which implies that 2 would rationally choose z whenever such a message was not sent, so that both types of 1 should always send the message (any randomization between x and y is better than z for both types of 1). But a message that is always sent by player 1, no matter what his type is, would convey no information to player 2, so that 2 would rationally choose his *ex ante* optimal action y.

If we now allow the players to communicate through a mediator who uses a randomized mechanism, then we can apply the revelation principle to characterize the surprisingly large set of possible incentive-compatible mechanisms. Among all direct-revelation mechanisms that satisfy the relevant informational incentive constraints for player 1 and strategic incentive constraints for player 2, the best for player 2 is as follows: if player 1 reports to the mediator that his type is α then with probability 2/3 the mediator recommends x to player 2, and with probability 1/3 the mediator recommends y to player 2; if player 1 reports to the mediator that his type is β then with probability 2/3 the mediator recommends y to player 2, and with probability 1/3 the mediator recommends z to player 2. Notice that this mechanism is also better for player 1 than the unmediated equilibria when 1's type is α, although it is worse for 1 when his type is β.

Other mechanisms that player 2 might prefer would violate the strategic incentive constraint that player 2 should not expect to gain by choosing z instead of y when y is recommended. If player 2 could pre-commit himself always to obey the mediator's recommendations, then better mechanisms could be designed.

Efficiency

The concept of efficiency becomes more difficult to define in economic situations where individuals have different private information at the time when the basic decisions about production and allocation are made. A welfare economist or social planner who analyses the Pareto efficiency of an economic system must use the perspective of an outsider, so he cannot base his analysis on the individuals’ private information. Otherwise, public testimony as to whether an economic mechanism or its outcome would be ‘efficient’ could implicitly reveal some individuals’ private information to other individuals, which could in turn alter their rational behaviour and change the outcome of the mechanism! Thus, Holmstrom and Myerson (1983) argued that efficiency should be considered as a property of mechanisms, rather than of the outcome or allocation ultimately realized by the mechanism (which will depend on

the individuals' private information).

Thus, a definition of Pareto efficiency in a Bayesian collective-choice problem must look something like this: 'a mechanism is efficient if there is no other feasible mechanism that may make some other individuals better off and will certainly not make other individuals worse off.' However, this definition is ambiguous in at least two ways.

First, we must specify whether the concept of feasibility takes incentive constraints into account or not. The concept of feasibility that ignores incentive constraints may be called *classical feasibility*. In these terms, the fundamental insight of mechanism theory is that incentive constraints are just as real as resource constraints, so that incentive compatibility may be a more fruitful concept than classical feasibility for welfare economics.

Second, we must specify what information is to be considered in determining whether an individual is 'better off' or 'worse off'. One possibility is to say that an individual is made worse off by a change that decreases his expected utility payoff as would be computed before his own type or any other individuals' types are specified. This is called the *ex ante* welfare criterion. A second possibility is to say that an individual is made worse off by a change that decreases his conditionally expected utility, given his own type (but not given the types of any other individuals). An outside observer, who does not know any individual's type, would then say that an individual may be made worse off, in this sense, if this conditionally expected utility were decreased for at least one possible type of the individual. This is called the *interim* welfare criterion. A third possibility is to say that an individual is made worse off by a change that decreases his conditionally expected utility given the types of all individuals. An outside observer would then say that an individual may be worse off in this sense if his conditionally expected utility were decreased for at least one possible combination of types for all the individuals. This is called the *ex post* welfare criterion.

If each individual knows his own type at the time when economic plans and decisions are made, then the interim welfare criterion should be most relevant to a social planner. Thus, Holmstrom and Myerson (1983) argue that, for welfare analysis in a Bayesian collective-choice problem, the most appropriate concept of efficiency is that which combines the interim welfare criterion and the incentive-compatible definition of feasibility. This concept is called *incentive efficiency*, or *interim incentive efficiency*. That is, a mechanism $\mu : T \rightarrow C$ is incentive efficient if it is an incentive-compatible mechanism and there does not exist any other incentive-compatible mechanism $\gamma : T \rightarrow C$ such that for every individual i and every type t_i in T_i,

$$\Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i)\, u_i(\gamma(t), t) \geq \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i)\, u_i(\mu(t), t),$$

and there is at least one type of at least one individual for which this inequality is strict. If a mechanism is incentive efficient, then it cannot be common knowledge among the individuals, at the stage when each knows only his own type, that there is some other incentive-compatible mechanism that no one would consider worse (given his own information) and some might consider strictly better.

For comparison, another important concept is classical *ex post* efficiency, defined using the *ex post* welfare criterion and the classical feasibility concept. That is, a mechanism $\mu : T \rightarrow C$ is (*classically*) *ex post efficient* iff there does not exist any other mechanism $\gamma : T \rightarrow C$ (not necessarily incentive compatible) such that, for every individual i and every combination of individuals' types t in $T = T_1 \times \ldots \times T_n$,

$$u_i(\gamma(t), t) \geq u_i(\mu(t), t),$$

with strict inequality for at least one individual and at least one combination of individuals' types.

The appeal of *ex post* efficiency is that there may seem to be something unstable about a mechanism that sometimes leads to outcomes such that, if everyone could share their information, they could identify another outcome that would make them all better off. However, we have seen that bargaining situations exist where no incentive-compatible mechanisms are *ex post* efficient. In such situations, the incentive constraints imply that rational individuals would be unable to share their information to achieve these gains, because if everyone were expected to do so then at least one type of one individual would have an incentive to lie.

Thus, a benevolent outside social planner who is persuaded by the usual Paretian arguments should choose some incentive-efficient mechanism. To determine more specifically an 'optimal' mechanism within this set, a social welfare function is needed that defines tradeoffs, not only between the expected payoffs of different individuals but also between the expected payoffs of different types of each individual. That is, given any positive utility-weights $\lambda_i(t_i)$ for each type t_i of each individual i, one can generate an incentive-efficient mechanism by maximizing

$$\Sigma_{i=1}^{n} \Sigma_{t_i \in T_i} \lambda_i(t_i) \Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i) u_i(\mu(t), t)$$

over all $\mu : T \rightarrow C$ that satisfy the incentive constraints; but different vectors of utility weights may generate different incentive-efficient mechanisms.

Bargaining over mechanisms

A positive economic theory must go beyond welfare economics and try to predict the economic institutions that may actually be chosen by the individuals in an economy. Having established that a social planner can restrict his attention to incentive-compatible direct-revelation mechanisms, which is a mathematically simple set, it is natural to assume that rational economic agents who are themselves negotiating the structure of their economic institutions should be able to bargain over the set of incentive-compatible direct-revelation mechanisms. But if we assume that individuals know their types already at the time when fundamental economic plans and decisions are made, then we need a theory of mechanism selection by individuals who have private information.

When we consider bargaining games in which individuals can bargain over mechanisms, there should be no loss of generality in restricting our attention to equilibria in which there is one incentive-compatible mechanism that is selected with probability 1 independently of anyone's type. This proposition, called the *inscrutability principle*, can be justified by viewing the mechanism-selection process as itself part of a more broadly defined general mechanism and applying the revelation principle. For example, suppose that there is an equilibrium of the mechanism-selection game in which some mechanism μ would be chosen if individual 1's type were α and some other mechanism v would be chosen if 1's type were β. Then there should exist an equivalent equilibrium of the mechanism-selection game in which the individuals always select a direct-revelation mechanism that coincides with mechanism μ when individual 1 confidentially reports type α to the mediator (in the implementation of the mechanism, after it has been selected), and that coincides with mechanism v when 1 reports type β to the mediator.

However, the inscrutability principle does not imply that the possibility of revealing information during a mechanism-selection process is irrelevant. There may be some mechanisms that we should expect not to be selected by the individuals in such a process, precisely because some individuals would choose to reveal information about their types rather than let these mechanisms be selected. For example, consider the following Bayesian collective-choice problem, due to Holmstrom and Myerson (1983). There are two individuals, 1 and 2, each of whom has two possible types, α and β, which are independent and equally likely. There are three social choice options, called x, y and z. Each individual's utility for these options depends on his type according to the following table.

Option	1, α	1, β	2, α	2, β
x	2	0	2	2
y	1	4	1	1
z	0	9	0	−8

The incentive-efficient mechanism that maximizes the *ex ante* expected sum of the two individuals' utilities is as follows: if 1 reports type α and 2 reports α then choose *x*, if 1 reports type β and 2 reports α then choose *z*, and if 2 reports β then choose *y* (regardless of 1's report). However, Holmstrom and Myerson argue that such a mechanism would not be chosen in a mechanism-selection game that is played when 1 already knows his type, because, when 1 knows that his type is α, he could do better by proposing to select the mechanism that always chooses *x*, and 2 would always want to accept this proposal. That is, because 1 would have no incentive to conceal his type from 2 in a mechanism-selection game if his type were α (when his interests would then have no conflict with 2's), we should not expect the individuals in a mechanism-selection game to agree inscrutably to an incentive-efficient mechanism that implicitly puts as much weight on 1's type-β payoff as the mechanism described above.

For another example, consider again the sender–receiver game due to Farrell. Recall that *y* would be the only possible equilibrium outcome if the individuals could communicate only face-to-face, with no mediation or other noise in their communication channel. Suppose that the mechanism-selection process is as follows: first 2 proposes a mediator who is committed to implement some incentive-compatible mechanism; then 1 can either accept this mediator and communicate with 2 thereafter only through him, or 1 can reject this mediator and thereafter communicate with 2 only face-to-face. Suppose now that 2 proposes that they should use a mediator who will implement the incentive-compatible mediation plan that is best for 2 (recommending *x* with probability 2/3 and *y* with probability 1/3 if 1 reports α, recommending *y* with probability 2/3 and *z* with probability 1/3 if 1 reports β). We have seen that this mechanism is worse than *y* for 1 if his type is β. Furthermore, this mechanism would be worse than *y* for player 1 under the *ex ante* welfare criterion, when his expected payoffs for type α and type β are averaged, each with weight 1/2. However, it is an equilibrium of this mechanism-selection game for player 1 always to accept this proposal, no matter what his type is. If 1 rejected 2's proposed mediator, then 2 might reasonably infer that l's type was β, in which case 2's rational choice would be *z* instead of *y*, and *z* is the worse possible outcome for both of 1's types.

Now consider a different mechanism-selection process for this example, in which the informed player 1 can select any incentive-compatible mechanism himself, with only the restriction that 2 must know what mechanism has been selected by 1. For any incentive-compatible mechanism μ, there is an equilibrium in which 1 chooses μ for sure, no matter what his type is, and they thereafter play the honest and obedient equilibrium of this mechanism. To support such an equilibrium, it suffices to suppose that, if any mechanism other than μ were selected, then 2 would infer that 1's type was β and therefore choose *z*. Thus, concepts like sequential equilibrium from non-cooperative game theory cannot determine the outcome of this mechanism-selection game, beyond what we already know from the revelation principle; we cannot even say that 1's selected mechanism will be incentive-efficient. To get incentive efficiency as a result of mechanism-selection games, we need some further assumptions, like those of cooperative game theory.

An attempt to extend traditional solution concepts from cooperative game theory to the problem of bargaining over mechanisms has been proposed by Myerson (1983; 1984a; 1984b). In making such an extension, one must consider not only the traditional problem of how to define reasonable compromises between the conflicting interests of different individuals, but also the problem of how to define reasonable compromises between the conflicting interests of different types of the same individual. That is, to conceal his type in the mechanism-selection process, an individual should bargain for some inscrutable compromise between what he really wants and what he would have wanted if his type had been different; and we need some formal theory to predict what a reasonable inscrutable compromise might be. In the above sender–receiver game, where only type β of player 1 should feel any incentive to conceal his type, we might expect an inscrutable compromise to be resolved in favor of type α. That is, in the mechanism-selection game

where 1 selects the mechanism, we might expect both types of 1 to select the incentive-compatible mechanism that is best for type α. (In this mechanism, the mediator recommends *x* with probability 0.8 and *y* with probability 0.2 if 1 reports α; and the mediator recommends *x* with probability 0.4, *y* with probability 0.4, and *z* with probability 0.2 if 1 reports β.) This mechanism is the *neutral optimum* for player 1, in the sense of Myerson (1983).

See Also

- incentive compatibility
- mechanism design experiments
- mechanism design (new developments)
- revelation principle

Bibliography

Aumann, R.J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96.

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55, 1–18.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Cramton, P., Gibbons, R. and Klemperer, P. 1987. Dissolving a partnership efficiently. *Econometrica* 55, 615–32.

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules: some general results on incentive compatibility. *Review of Economic Studies* 46, 185–216.

Farrell, J. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5, 514–31. Repr. in *Mathematical Models in Economics*, ed. M. Bacharach and M. Dempster. Oxford: Oxford University Press, 1997.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–602.

Harris, M. and Raviv, A. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49, 1477–99.

Harris, M. and Townsend, R.M. 1981. Resource allocation under asymmetric information. *Econometrica* 49, 33–64.

Harsanyi, J.C. 1967. Games with incomplete information played by Bayesian players. *Management Science* 14, 159–82, 320–34, 481–502.

Holmstrom, B. 1977. On incentives and control in organizations. Ph.D. thesis, Graduate School of Business, Stanford University.

Holmstrom, B. and Myerson, R.B. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51, 1799–19.

Lewis, T.R. and Sappington, D.E.M. 1989. Countervailing incentives in agency problems. *Journal of Economic Theory* 49, 294–313.

Myerson, R.B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47, 61–74.

Myerson, R.B. 1981. Optimal auction design. *Mathematics of Operation Research* 6, 58–73.

Myerson, R.B. 1982. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of Mathematical Economics* 10, 67–81.

Myerson, R.B. 1983. Mechanism design by an informed principal. *Econometrica* 51, 1767–97.

Myerson, R.B. 1984a. Two-person bargaining problems with incomplete information. *Econometrica* 52, 461–87.

Myerson, R.B. 1984b. Cooperative games with incomplete information. *International Journal of Game Theory* 13, 69–86.

Myerson, R.B. 1985. Bayesian equilibrium and incentive compatibility. In *Social Goals and Social Organization*, ed. L. Hurwicz, D. Schmeidler and H. Sonnenschein. Cambridge: Cambridge University Press.

Myerson, R.B. 1986. Multistage games with communication. *Econometrica* 54, 323–58.

Myerson, R.B. 1988. Incentive constraints and optimal communication systems. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, ed. M.Y. Vardi. Los Altos: Morgan Kaufmann.

Myerson, R.B. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29, 265–81.

Ortega-Reichert, A. 1968. Models for competitive bidding under uncertainty. Ph.D. thesis, Department of Operations Research, Stanford University.

Palfrey, T. and Srivastava, S. 1987. On Bayesian implementable allocations. *Review of Economic Studies* 54, 193–208.

Riley, J.G. and Samuelson, W.F. 1981. Optimal auctions. *American Economic Review* 71, 381–92.

Rosenthal, R.W. 1978. Arbitration of two-party disputes under uncertainty. *Review of Economic Studies* 45, 595–604.

Samuelson, W. 1985. A comment on the Coase Theorem. In *Game-Theoretic Models of Bargaining*, ed. A.E. Roth. Cambridge: Cambridge University Press.

Schelling, T.C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Wilson, R. 1985. Incentive efficiency of double auctions. *Econometrica* 53, 1101–15.

How to cite this article

Myerson, Roger B. "mechanism design." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000132> doi:10.1057/9780230226203.1078

The New Palgrave Dictionary of Economics Online

medieval guilds

Gary Richardson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Guilds operated throughout Europe during the Middle Ages, and in many places into the early modern era. Merchant guilds were organizations of merchants involved in long-distance commerce and local wholesale trade, and may also have been retail sellers of commodities in their home cities and distant venues where they possessed rights to trade. Craft guilds were organized along lines of particular trades, their members typically owning and operating small family businesses. After the Black Death guilds grew rapidly in number, becoming the central economic and social institutions in medieval towns.

Keywords

Black Death; collective liability; craft guilds; credit; fertility; free rider problem; Hanseatic League; insurance; medieval guilds; merchant guilds; monopoly; reputation; urbanization

Article

Guilds operated throughout Europe during the Middle Ages, and in many places, lasted into the early modern era. Guilds were groups of individuals with common goals whose activities, characteristics, and composition varied greatly across centuries, regions, and industries.

Guilds filled many niches in medieval economy and society. Typical taxonomies divide urban occupational guilds into two types: merchant and craft.

Merchant guilds were organizations of merchants who were involved in long-distance commerce and local wholesale trade, and may also have been retail sellers of commodities in their home cities and distant venues where they possessed rights to set up shop. The largest and most influential merchant guilds participated in international commerce and politics and established colonies in foreign cities. In many cases, they evolved into or became inextricably intertwined with the governments of their home towns.

Merchant guilds enforced contracts among members and between members and outsiders. Guilds policed members' behaviour because medieval commerce operated according to the community

responsibility system. If a merchant from a particular town failed to fulfil his part of a bargain or pay his debts, all members of his guild could be held liable. When they were in a foreign port, their goods could be seized and sold to alleviate the bad debt. They would then return to their hometown, where they would seek compensation from the original defaulter.

Merchant guilds also protected members against predation by rulers. Rulers seeking revenue had an incentive to seize money and merchandise from foreign merchants. Guilds threatened to boycott the realms of rulers who did this, a practice known as *withernam* in medieval England. Since boycotts impoverished both kingdoms which depended on commerce and governments for whom tariffs were the principal source of revenue, the threat of retaliation deterred medieval potentates from excessive expropriations.

Craft guilds were organized along lines of particular trades. Members of these guilds typically owned and operated small businesses or family workshops. Craft guilds operated in many sectors of the economy. Guilds of victuallers bought agricultural commodities, converted them to consumables, and sold finished foodstuffs. Examples included bakers, brewers, and butchers. Guilds of manufacturers made durable goods and, when profitable, exported them from their towns to consumers in distant markets. Examples include makers of textiles, military equipment, and metalware. Guilds of a third type sold skills and services. Examples include clerks, teamsters, and entertainers.

These occupational organizations engaged in a wide array of economic activities. Some manipulated input and output markets to their own advantage. Others established reputations for quality, fostering the expansion of anonymous exchange and making everyone better off. Because of the underlying economic realities, victualling guilds tended towards the former. Manufacturing guilds tended towards the latter. Guilds of service providers fell somewhere in between. All three types of guilds managed labour markets, lowered wages, and advanced their own interests at their subordinates' expense. These undertakings had a common theme. Merchant and craft guilds acted to increase and stabilize members' incomes.

Non-occupational guilds also operated in medieval towns and cities. These organizations had both secular and religious functions. Historians refer to these organizations as social, religious, or parish guilds as well as fraternities and confraternities. The secular activities of these organizations included providing members with mutual insurance, extending credit to members in times of need, aiding members in courts of law, and helping the children of members afford apprenticeships and dowries.

The principal pious objective was the salvation of the soul and escape from Purgatory. Guilds served as mechanisms for organizing, managing, and financing members' collective quest for eternal salvation. Efforts centered on three types of tasks. The first were routine and participatory religious services such as prayers, processions, the singing of psalms, the illumination of holy symbols, and the distribution of alms to the poor. The second category consisted of actions performed on members' behalf after their deaths and for the benefit of their souls. Post-mortem services began with funerals and continued perpetually as guilds prayed (or hired priests to pray) for the salvation of the souls of all deceased members. The third category involved indoctrination and monitoring to maintain the piety of members. Righteous living was important because members' fates were linked together. The more pious one's brethren, the more helpful their prayers, and the more quickly one escaped from purgatory. So, in hopes of minimizing purgatorial pain and maximizing eternal happiness, guilds beseeched members to restrain physical desires and forgo worldly pleasures.

Guilds also operated in villages and the countryside. Rural guilds performed the same tasks as social and

religious guilds in towns and cities. Recent research on medieval England indicates that guilds operated in most, if not all, villages. Villages often possessed multiple guilds. Most rural residents belonged to a guild. Some may have joined more than one organization.

Guilds often spanned multiple dimensions of this taxonomy. Members of craft guilds participated in wholesale commerce. Members of merchant guilds opened retail shops. Social and religious guilds evolved into occupational associations. All merchant and craft guilds possessed religious and fraternal features.

In sum, guild members sought prosperity in this life and providence in the next. Members wanted high and stable incomes, quick passage through Purgatory, and eternity in heaven. Guilds helped them coordinate their collective efforts to attain these goals.

To attain their collective goals, guilds had to persuade members to contribute to the common good and deter free riding. Guilds that wished to develop respected reputations had to get all members to sell superior merchandise. Guilds that wished to lower the costs of labour had to get all masters to reduce wages. Guilds that wished to raise the prices of products had to get all masters to restrict output. Guilds whose members wished to enter heaven had to get all members to live piously, abstaining both from the pleasures of the flesh and the material temptations of secular society.

To persuade members to cooperate and advance their common interests, guilds formed stable, self-enforcing associations that possessed structures for making and implementing collective decisions. A guild's members met periodically to elect officers, audit accounts, induct new members, debate policies, and amend ordinances. Officers administered a nexus of agreements among a guild's members. Details of these agreements varied greatly from guild to guild, but the issues addressed were similar in all cases. Members agreed to contribute certain resources or take certain actions that furthered the guild's occupational and spiritual endeavors.

Members who failed to fulfil their obligations faced punishments. Punishments varied across transgressions, guilds, time and place, but a pattern existed. First-time offenders were punished lightly, perhaps suffering public scolding and paying small monetary fines, and repeat offenders punished harshly. The ultimate threat was expulsion.

Within large guilds, a hierarchy existed. Masters were full members who usually owned their own workshops, retail outlets, or trading vessels. Masters employed journeymen, who were labourers who worked for wages on short-term contracts or a daily basis (hence the term journeyman, from *jour*, the French word for ‘day’). Journeymen hoped to one day advance to the level of master. To do this, journeymen usually had to save enough money to open a workshop and pay for admittance, or, if they were lucky, receive a workshop through marriage or inheritance.

Masters also supervised apprentices, who were usually boys in their teens who worked for room, board and perhaps a small stipend in exchange for a vocational education. Both guilds and government regulated apprenticeships, usually to ensure that masters fulfilled their part of the apprenticeship agreement. Terms of apprenticeships varied, usually lasting from five to nine years.

Relationships between guilds and governments varied over centuries and around Europe. Guilds typically began as voluntary associations with little legal standing. Most guilds operated without formal recognition or authorization from the government. Successful occupational guilds aspired to attain recognition as a self-governing association with the right to possess property and other legal privileges. Merchant and craft guilds often purchased these rights from municipal and royal authorities.

The history of guilds stretches back to times with few written records. In the late Roman Empire,

organizations resembling guilds existed in most towns and cities. These voluntary associations of artisans, known as *collegia*, were organized along trade lines. Members shared religious observances and fraternal dinners. Most of these organizations disappeared during the Dark Ages, when the Western Roman Empire disintegrated and urban life collapsed. In the Eastern Roman Empire, some *collegia* may have survived from late antiquity and evolved into medieval guilds, but it is unlikely that even the most resilient *collegia* survived in Western Europe.

In the centuries following the collapse of the Roman Empire, evidence indicates that guild-like associations operated in most towns and many rural areas. These organizations functioned as modern burial and benefit societies, whose objectives included prayers for the souls of deceased members, payments of *weregilds* in cases of justifiable homicide, and supporting members involved in legal disputes. These rural guilds were descendents of Germanic social organizations known as *gilda* which the Roman historian Tacitus referred to as *convivium*.

During the 11th through 13th centuries, considerable economic development occurred. The revival of long-distance trade coincided with the expansion of urban areas. Merchant guilds formed an institutional foundation for this commercial revolution. Merchant guilds sprung up in towns throughout Europe, and in many places rose to prominence in urban political structures. Merchant guilds' principal accomplishment was establishing the institutional foundations for long-distance commerce.

Merchant guilds first flourished in Italian cities in the 12th century. Craft guilds became ubiquitous in Italy during the succeeding century. In northern Europe, merchant guilds rose to prominence a century later, when local merchant guilds in trading cities such as Lubeck and Bremen formed alliances with merchants throughout the Baltic region. The alliance system grew into the Hanseatic League which dominated trade around the Baltic and North Seas and in northern Germany.

As economic expansion continued in the 13th and 14 centuries, the influence of the Catholic Church grew, and the doctrine of Purgatory developed. The doctrine inspired the creation of countless religious guilds, since the doctrine provided members with strong incentives to want to belong to a group whose prayers would help one enter heaven and it provided guilds with mechanisms to induce members to exert effort on behalf of the organization.

The number of guilds grew rapidly after the Black Death, for several reasons. The decline in population raised per capita incomes, which encouraged the expansion of consumption and commerce, which in turn necessitated the formation of institutions to satisfy this demand. Repeated epidemics decreased family sizes, particularly in cities, where the typical adult had on average perhaps 1.5 surviving children, few surviving siblings, and only a small extended family, if any. Guilds replaced extended families in a form of fictive kinship. The decline in family size and impoverishment of the Church also forced individuals to rely on their guild more in times of trouble, since they no longer could rely on relatives and priests to sustain them through periods of crisis. All of these changes bound individuals more closely to guilds, discouraged free riding, and encouraged the expansion of collective institutions.

For nearly two centuries after the Black Death, guilds dominated life in medieval towns. Any town resident of consequence belonged to a guild. Most urban residents thought guild membership to be indispensable. Guilds dominated manufacturing, marketing, and commerce. Guilds dominated local politics and influenced national and international affairs. Guilds were the centre of social and spiritual life.

The heyday of guilds lasted into the 16th century. The Reformation weakened guilds. Afterwards, in Protestant nations the influence of guilds waned. Guilds often asked governments for assistance. Guilds

requested monopolies on manufacturing and commerce and asked courts to force members to live up to their obligations. Guilds lingered where governments provided such assistance. Guilds faded where governments did not. Guilds retained strength in nations which remained Catholic until they were swept away by the reforms following the French Revolution and the Napoleonic Wars.

Bibliography

Basing, P. 1990. *Trades and Crafts in Medieval Manuscripts*. London: British Library.

Cooper, R.C.H. 1985. *The Archives of the City of London Livery Companies and Related Organizations*. London: Guildhall Library.

Davidson, C. 1996. *Technology, Guilds, and Early English Drama*. Early Drama, Art, and Music Monograph Series 23. Kalamazoo, MI: Medieval Institute Publications, Western Michigan University.

Epstein, S.R. 1998. Craft guilds, apprenticeships, and technological change in pre-industrial Europe. *Journal of Economic History* 58, 684–713.

Epstein, S. 1991. *Wage and Labor Guilds in Medieval Europe*. Chapel Hill, NC: University of North Carolina Press.

Gross, C. 1890. *The Guild Merchant: A Contribution to British Municipal History*. Oxford: Clarendon Press.

Gustafsson, B. 1987. The rise and economic behavior of medieval craft guilds: an economic-theoretical interpretation. *Scandinavian Journal of Economics* 35, 1–40.

Hanawalt, B. 1984. Keepers of the lights: late medieval English parish guilds. *Journal of Medieval and Renaissance Studies* 14, 21–37.

Hatcher, J. and Miller, E. 1995. *Medieval England: Towns, Commerce and Crafts, 1086–1348*. London: Longman.

Hickson, C.R. and Thompson, E.A. 1991. A new theory of guilds and European economic development. *Explorations in Economic History* 28, 127–68.

Lopez, R. 1971. *The Commercial Revolution of the Middle Ages, 950–1350*. Englewood Cliffs, NJ: Prentice-Hall.

Mokyr, J. 1990. *The Lever of Riches: Technological Creativity and Economic Progress*. Oxford: Oxford University Press.

Pirenne, H. 1952. *Medieval Cities: Their Origins and the Revival of Trade*, trans. F. Halsey. Princeton: Princeton University Press.

Richardson, G. 2001. A tale of two theories: monopolies and craft guilds in medieval England and modern imagination. *Journal of the History of Economic Thought* 23, 217–42.

Richardson, G. 2000. Brand Names Before the Industrial Revolution. Working paper, UC Irvine.

Richardson, G. 2004. Guilds, laws, and markets for manufactured merchandise in late-medieval England. *Explorations in Economic History* 41, 1–25.

Richardson, G. 2005. Christianity and craft guilds in late medieval England: a rational choice analysis. *Rationality and Society* 17, 139–89.

Richardson, G. 2005. The prudent village: risk pooling institutions in medieval English agriculture. *Journal of Economic History* 65, 386–413.

Smith, T. 1870. *English Guilds*. London: N. Trübner & Co.

Swanson, H. 1983. *Building Craftsmen in Late Medieval York*. York: University of York Press.

Thrupp, S. 1989. *The Merchant Class of Medieval London 1300–1500*. Chicago: University of Chicago Press.

Unwin, G. 1904. *The Guilds and Companies of London*. London: Methuen & Company.

Ward, J. 1997. *Metropolitan Communities: Trade Guilds, Identity, and Change in Early Modern London*. Palo Alto: Stanford University Press.

Westlake, H.F. 1919. *The Parish Guilds of Mediaeval England*. London: Society for Promotion of Christian Knowledge.

How to cite this article

Richardson, Gary. "medieval guilds." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000537> doi:10.1057/9780230226203.1081

The New Palgrave Dictionary of Economics Online

Menger, Carl (1840 – 1921)

Karen I. Vaughn
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian School; bilateral competition; Böhm-Bawerk, E. von; capital accumulation; classical economics; commodity money; double coincidence of wants; economic development; economic good; economic man; exchange of equivalent values; exchange theory; fiat money; German Historical School; Hayek, F. A. von; Imputation; isolated exchange; labour theory of value; limits of trade; marginal productivity theory; marginal utility analysis; market intermediaries; Menger, C.; Methodenstreit; methodological individualism; Mises, L. E. von; money; opportunity cost; period of production; precautionary balances; price formation; Schmoller, G. von; spontaneous order; subjective theory of value; time preference; value; Wieser, F. F. von

Article

Carl Menger is known as one of the co-founders, along with W.S. Jevons and Leon Walras, of marginal utility analysis. As such, he can be counted as one of the originators of modern neoclassical economics. He is also recognized as the founder of the Austrian School of economics which developed a distinct tradition of economic thought over the century following his writing.

Menger was born in Neu-Sandez, Galizieu, a part of Austria that later became Poland. His family were mostly civil servants and army officers. Menger's father was a lawyer, and Carl studied law and political science first at the University of Vienna (1859–60) and then at Prague (1860–63). He took a doctorate at the University of Cracow and soon after, began a career in journalism. He worked in Lemberg and later Vienna where his main interests were in economic and fiscal problems of Austria. In 1871, Menger entered the Austrian civil service. However, 1871 was also the year in which his first book, *Grundsätze der Volkswirtschaftslehre*, later translated as *Principles of Economics*, was published. He presented this work for his habilitation for the faculty of law and political science at the University of Vienna. As a consequence, he became a ‘privatdozent’ and quit his position in the civil service. In 1873, he was appointed extraordinary professor and began his very long and very successful academic career. In 1876, Menger was appointed tutor to Crown Prince Rudolf of Austria and for two years accompanied him on

travels through Germany, France, Switzerland and the British Isles. Upon his return, he resumed his teaching responsibilities, and he received a chair in political economy in 1879. He continued to teach until 1903 when, at the comparatively early age of 63, he retired to devote himself exclusively to completing the treatise he had begun with the *Principles*. He died within three days of his 81st birthday in 1921, his project still incomplete. He was survived by his one son, Karl.

The Principles of Economics

When Menger published the *Principles*, he was 31 years old and a journalist who had recently been appointed to the prestigious ‘Ministerratsprasidium’ in the Austrian civil service. Several biographers report that during his years as a journalist Menger became interested in economics because he observed that current economic theories did not seem to explain current economic events. He therefore wanted to work out the laws of economics for himself. It is apparent from the internal textual evidence of the *Principles*, however, that Menger must have had a more than cursory interest in the subject of economics during his years as a student. He must have read deeply and widely in the history of economics, since his first major work cites a wide range of earlier thinkers on economic problems including Aristotle, the medieval scholastics, Turgot, Smith, Ricardo, the German historicists and the contemporary socialists. Menger's knowledge of the history of economic thought is also evidenced by the outstanding library he accumulated during his lifetime, and by the fact that most of the major works in economic thought bear the marks of his close study.

Menger's clear purpose was to show how his theory of value could solve satisfactorily and in a unified manner all the problems of economic theory posed by earlier thinkers. The major target of his work was the labour theory of value which he believed was not only incorrect as an explanation of value and prices, but also failed to provide a unified explanation for factor prices on its own terms. However, Menger also took as his task to explain away the paradox of value, the erroneous view of Aristotle that exchange was an exchange of equivalent values, the mistaken view that capital as such was productive and the notion that money had to be explained according to different principles from other goods. In fact, every chapter of the *Principles* contains a refutation of some earlier doctrine or other that required a correct theory of value to elucidate.

Menger was writing partly against the background of classical economics. He was also, however, writing to an audience of German scholars who, in their rejection of classical economics were also rejecting the whole notion that one could develop a scientific theory of economic phenomena. Nothing, however, could be further from Menger's approach. Part of his aim, then, was to explain to the German historical economists that scientific economic theory was possible and compatible with empirical reality. To that end, Menger dedicated the book ‘with respectful esteem to D. Wilhelm Roscher’, a major figure in the older German Historical School.

To Menger, the central unifying principle of economics was the phenomenon of value. One had to explain the source of value before any of its particular manifestations could be understood. However, to develop adequately a theory of value, Menger had to prepare the ground upon which the theory rests. For Menger, that meant spending the first two chapters of his book (62 pages and over one-fourth the main text) on an exhaustive discussion of the meaning of a good and of an economic good in particular. While to the modern reader this might seem excessively thorough, Menger, the innovator, wished to take nothing for granted in establishing the firm basis of his theory. One had to move from the notion of

useful things to the notion of a good to the concept of an economic good before one could understand the real meaning of economic value. Since all of economic theory hangs on this concept, he must have believed it imperative to be sure the reader understands each step of the argument.

Right from the beginning we see Menger's distinctive approach to economic theory. ‘All things are subject to the law of cause and effect’ (*Principles*, p. 51). Economic theory is an exercise in discovering and explaining the causal relationship between things and human values. He thus begins by pointing out that there are many useful things in the world, but for a useful thing to have ‘goods-character’, men must (*a*) recognize a causal connection between the good and its ability to satisfy a need and (*b*) have the power to make use of the thing for need satisfaction. (Goods, Menger points out, can be concrete things or they can be intangible relationships such as firms, copyrights or good will, an observation that is distinctly modern.) This is a pattern repeated again and again in Menger's writing: men must have knowledge and power. Economic life is built around gaining knowledge and power; knowledge of causal relationship between things and satisfaction, knowledge of technical production relationships, knowledge of trading opportunities, knowledge of ‘economic’ prices, knowledge of the qualities of goods, and the power to make the best use of man's knowledge.

Knowledge of causal connections among goods permits men to rank goods in accordance with their relationship to want satisfaction. Goods that have the ability to satisfy needs directly (consumer goods), Menger called ‘first order goods’. Goods that can only indirectly satisfy needs by being transformed with complementary goods into first order goods, Menger called ‘goods of a higher order’ (inputs). Furthermore, higher order goods are not valued in themselves, but derive their ‘goods character’ from first order goods, an observation that will later allow Menger to develop his refutation of the labour theory of value.

Having established the concept of a good in general in Chapter 1, Menger goes on in Chapter 2 to explain the concept of an economic good. Menger's definition of an economic good is completely familiar to modern readers; the way in which Menger develops his argument is not. Menger sees men's strictly economizing activities as taking place within the context of an overall plan through time. He argues that men must estimate both their needs for various goods and the quantities of goods that will be available for fulfilling their consumption plans for specific periods of time. Their estimated needs, he calls their ‘requirements’ (*bedarfs*) a concept for which we have no modern equivalent although Stigler (1941, p. 140) has argued that requirements are the quantities of goods sufficient to make marginal utility go to zero (all the economic goods men could rationally consume). An economic good, then, is one where available quantities fall short of men's requirements.

The notion of requirements is important to Menger's argument because it allows him to discuss how men get information about requirements and quantities of goods and how they plan for their consumption in the face of uncertainty. It is obvious in this discussion that Menger does not hypothesize given utility functions which are maximized subject to fixed constraints. While he eventually gets to a verbal explanation of economizing behaviour that is consistent with the standard model, to him the interesting questions involve how men go about estimating their requirements over time and how they plan to satisfy them. Their planning activities require not only that they estimate future needs based on present tastes and preferences, but that they take into account the fact that their needs may change in unexpected ways. Further, their planning activity also encompasses plans to change the quantities of goods available. Hence, the plan is one of production as well as consumption. Only after Menger establishes the importance of human planning through time does he go on to discuss economizing behaviour in the

modern sense of maximizing satisfaction within the known resource constraints.

Economizing to Menger, then, is a two-step process that involves first formulating a general plan for meeting one's requirements by assessing probable needs given an uncertain future and gathering information about the probable availability of goods, and then actually economizing based on the actual needs and quantities available at a moment in time.

Menger's discussion of an economic good is rich with associated insights. In this chapter, he gives an account of how non-economic goods become economic (through growth of population, growth of human needs and advances in knowledge as civilization progresses), a description of public goods (goods that are economic goods in general but are provided in such a way that people treat them as non-economic goods), an account of the origin and function of private property (to protect economic goods owned by the haves from the predation of the have-nots). Property is the ‘only practically possible solution of the problem that is, in the nature of things, imposed upon us by the disparity between requirements for, and available quantities of, all economic goods’ (*Principles*, p. 97), and a discussion of the economic implications of differing qualities of goods. He devotes part of his discussion of economic goods to discussing the nature of individual wealth – the entire sum of economic goods at an individual's command – and of national wealth – a slippery concept that can only be accurately described as ‘a complex of wealths linked together by intercourse and trade’ (*Principles*, p. 112).

Finally, after this detailed groundwork, Menger gets to his theory of value in Chapter 3. Menger has been called a member of the ‘psychological school’ because of his thoroughly subjective notion of value. However, it is not a Jevons-like utilitarian subjectivism. Goods are valued not because they provide various quantities of utils to individuals, but because they serve various uses that have different levels of importance to individuals. The difference may seem small to the reader, but it makes for subtle but important differences in understanding the valuation process. ‘Value is … the importance that individual goods or quantities of goods attain for us because we are conscious of being dependent on command of them for the satisfaction of our needs’ (*Principles*, p. 115). Value is a judgement men make about the importance of goods; it adheres in concrete units of goods and not in abstract utility. The problem of a theory of value is to explain the differences in value among different goods.

Menger develops his theory of value in two stages. First he shows, with the use of a numerical table, how the importance men attach to the acquisition of additional units of a good that satisfies a particular need declines as more of the good is acquired, and by comparing the declining satisfactions associated with the acquisition of increasing amounts of various goods, he explains why a man might satisfy some of his desire for tobacco, for example, before he has completely satisfied his desire for food. In fact, Menger's tables are vivid examples of Gossen's first and second laws. Menger's use of numbers may give the impression that he is explaining utility as a cardinally measurable quantity. However, the impression is immediately dispelled when he points out that his chart is merely illustrative of a general psychological principle and is not meant to be taken literally. Furthermore, his chart, he explains, describes only a special case of valuation – the case where a single good serves for a single satisfaction. The more important case – where a single good has multiple uses – is more complex and requires more discussion. Interestingly, it is only in the context of the following more complex case that he states clearly his principle of diminishing marginal valuation.

When a single good, such as sacks of grain or pails of water, can serve many different uses, the first units will be used to serve the most important uses for the good while successive units of the good will be put to less and less important uses. Menger concludes, then, that the value of any one sack of grain is

equal to the satisfaction associated with the least important use that would go unsatisfied if one sack of grain is removed, a statement of diminishing marginal utility that is completely free of mathematical metaphor.

Menger drew two immediate implications from his value theory: (1) the diamonds–water paradox was easily solved because given their respective quantities, the marginal unit of water in most cases served no use while the marginal unit of scarce diamonds had very important desires to satisfy, and (2) the labour theory of value was obviously incorrect.

> The determining factor in the value of a good, then, is neither the quantity of labor or other goods necessary for its production not the quantity necessary for its reproduction, but rather the magnitude of importance of those satisfactions with respect to which we are conscious of being dependent on command of the good. This principle of value determination is universally valid, and no exception to it can be found in human economy. (*Principles*, p. 145)

This leads Menger to one of the most important theoretical implications of his theory – that the value of goods of a higher order depends on the prospective value of corresponding goods of lower order. In fact, the value of an input is equal to the satisfaction that would be forgone if the input were not available for use. Note that this is not so much a marginal productivity theory of factor value as it is a ‘marginal utility product’ theory completely consistent with his subjective theory of value.

Despite his comments on the value of goods of a higher order, Menger did not develop a theory of production in the modern sense. He did observe, however, that all production takes place in time, and that the higher the order of goods employed, the more distant in time will be the final satisfaction obtained. The only way men can increase output is ‘to lengthen the period of time over which their provident activity is to extend in the same degree that they progress to goods of higher order’ (*Principles*, p. 153). This suggestion was the basis upon which Böhm-Bawerk constructed his theory of the period of production that led to so much controversy by the end of the 19th century.

Menger also points out that the limit to economic progress is the degree to which men value the same satisfaction more highly in the present rather than the future. Later called ‘time-preference’ by Austrian economists, Menger believed it was a consequence of men's continuous and finite life span. Without time preference, one would have to expect infinite capital accumulation. Notice that time preference is an explanation for why there is a limit to capital accumulation rather than an explanation for why capital is accumulated at all.

Menger is best known for his theory of value and its implication for goods of a higher order. His theory of exchange and price is neither so well-known nor so highly regarded. This is a pity since the chapters following the theory of value are equally rich with economic insights and deserve close attention by modern readers. Predictably, Menger's theory of exchange is derived from his theory of value. His starting point is Adam Smith's statement that men are possessed of a ‘propensity to truck, barter and exchange one good for another’, a statement Menger finds objectionable since it provides no explanation for the particular kinds of trade men make or for the limits of their trading activity. Men do not trade because of a propensity to do so, but because of a rational desire to improve their well-being. Men seek out trade opportunities in order to exchange something less valuable for something more valuable and hence trade is productive of value for both trading partners. The problem for the economist, then, is to

determine the limit of trade, limits that will be reached when neither party any longer stands to gain. While Menger's theory of trade is fairly standard, less standard is his very modern discussion of the importance of transactions costs in limiting trade. These ‘economic sacrifices of exchange’ (*Principles*, p. 189) arise because men and their possessions are separate in space and time and must be brought together for trade to take place. Sometimes these economic sacrifices are so great that a potentially productive trade does not take place at all. It is the role of market intermediaries (including entrepreneurs) to reduce the economic sacrifices of trade through improved knowledge and improved market organizations. Entrepreneurs bring together potential traders, and the source of the intermediary's income is the gain in satisfaction permitted by his activities. The idea of transactions costs and the role of market ‘intermediaries’ in reducing transactions costs was rediscovered in the 1950s.

Menger's theory of exchange leads him to develop his theory of price. This chapter eventually arrives at propositions that are now standard in price theory, but it does so in a peculiar way. Menger states in the very beginning of the chapter that contrary to the beliefs of some earlier thinkers, price is not the fundamental feature of exchange. While price is directly observable, it is derivative of the real fundamental feature of exchange: the utility gain from trade. Price is merely a ‘symptom of an economic equilibrium between the economies of individuals’ (*Principles*, p. 191). There should be no misunderstanding then about exchange involving an exchange of equivalent values. If such were the case, men would be willing to reverse their trades since there would be no gain or loss involved. But we do not observe such ‘reversible’ trades in the real world because trades are not of equivalent values but of subjective values that differ for each party to the trade. Price theory, then, is not a theory of establishing equivalents for exchange but rather a theory that seeks to explain why men give specific quantities of goods for specific quantities of other goods.

Menger approaches this problem in a way that was to become common in neoclassical economics – according to the number of traders in the market. However, instead of taking the case of many buyers and sellers as the norm and examining various monopolies as deviations, he begins with the simplest case of two party exchange (‘isolated exchange’) and progresses through various monopoly models finally to reach the case of ‘bilateral competition’. The reason for this progression is not simply analytic simplicity; he believed that this was the way trade actually developed in history with monopolies giving way to more and more competitive conditions, and he gives several historical examples to support his case.

Under isolated exchange, price will fall within a range set by the marginal utilities of the two traders. The actual price is indeterminate from the point of view of theory, but in most cases, Menger argued, neither party will have any special bargaining power and they will agree to a price that gives them a more or less equal utility gain.

From there Menger progresses to the case where a monopolist provides a single good to several competing buyers. In this case, the limits within which price will fall are narrowed by the intensity of demand of the most eager buyer and the one next most eager to acquire the good.

The case of monopoly provision of several units of a good to competing buyers is even more interesting. There, assuming a uniform price is established:

> price formation takes place between the limits that are set by the equivalent of one unit of the monopolized good to the individual least eager and least able to compete who still participates in the exchange and the equivalent of one unit of the monopolized good to the

> individual most eager and best able to compete of the competitors who are economically excluded from the exchange. (*Principles*, p. 207)

One important implication is that the larger the quantity offered for sale by the monopolist, the ‘lower in terms of purchasing power and eagerness to trade will he have to descend among the classes of competitors for the monopolized good in order to sell the whole quantity, and hence the lower also will be the price of one unit of the monopolized good’ (*Principles*, p. 207). In this way, Menger established the inverse price–quantity relationship that had been assumed by economists prior to the introduction of marginalism into economic thought.

What is interesting in Menger's approach is that he emphasizes that the process of price formation is the same regardless of the market conditions. Monopolists are subject to the limits placed on their actions by the utilities of the buyers just as competitors are so limited. What does vary according to market conditions are the policies of sellers. Monopolists may well follow a policy of restricting supply in order to sell few units at higher prices, or they may follow a policy of selling different units at different prices depending on the buyers. Competitors in supply of a product, however, will never find those policies to their advantage and hence under bilateral competition, one would expect prices to be lower and quantities supplied to be greater.

There is some debate as to whether Menger was offering an equilibrium theory of price determination in the *Principles* (Streissler, 1972; Jaffé, 1976). Certainly, his method of reasoning implies some underlying equilibrium price within any given market, and he even states that from time to time equilibrium prices will be observed. Equilibrium prices are ‘economic’ prices in that transactions at these prices are the result of economizing behaviour where no one could have been better off at another price. Further, he describes prices that reflect the full ‘economic situation’, a phrase that seems to indicate a more widespread economic equilibrium. However, it is also true that Menger did not describe economies settling down to a strict general equilibrium in the manner of Walras. Indeed, given the barriers to strict economic behaviour, especially barriers of incomplete knowledge, that are inherent in real life, Menger would find a Walrasian general equilibrium in principle unattainable. Men did the best they could, and with economic progress their best got better, but the very conception of a Walrasian general equilibrium is foreign to Menger's method of reasoning. This will become clearer below when Menger's methodology is discussed.

The next two chapters, ‘Use Value and Exchange Value’ and ‘The Theory of the Commodity’, while containing several interesting discussions about market organization, are really prelude to the very important last chapter on ‘The Theory of Money’. In the ‘Commodity’ chapter, Menger defines a commodity as a good intended for sale and then discusses the varying degrees of saleability of commodities based on their characteristics and market organization. The point he is leading to is to define money as the most marketable of all commodities, his starting point for the last chapter.

Menger does not develop a theory of the value of money in the *Principles*. While he does stress the importance of holding precautionary balances, to Menger the most important questions are how does money come to exist and what functions does it serve. These are questions he addresses both in his *Principles*, in the later work on methodology and in his two articles on money written in 1892. From a modern perspective, there are two particularly interesting features of Menger's discussion that should be noted. First, Menger's account of the origin of money is developed in a way reminiscent of the reasoning of the Scottish Enlightenment and Adam Smith's ‘invisible hand’ in particular (although it is doubtful

that the writers of the Scottish Enlightenment were the direct sources for his reasoning. In fact, at one point, he criticizes Adam Smith for a too mechanistic and rationalistic view of economic and social institutions! [*Investigations*, p. 177]). Money, according to Menger, arises out of the self-interested actions of individuals aimed at attaining their own ends through trade, but not specifically aimed at developing a money commodity as such. Second, because money arises as an unintended by-product of human action, it is not a creation of government.

The process Menger describes for the origin of money is a straightforward extension of his theory of economizing behaviour through trade. Following Aristotle, Menger points out the difficulties men face under barter in finding trading partners, (the problem of the ‘double coincidence of wants’). Rational men soon come to realize that goods have different degrees of marketability. A cow, for instance, is far more marketable than custom made shoes. Hence, men learn that if they exchange their less marketable goods for goods that may not directly satisfy their needs but that are more marketable, they will be more successful at bartering for what they really need. Eventually, Menger reasons, some one commodity will emerge as the most marketable commodity and men will be willing always to accept it in exchange for other goods because they know they will have no trouble trading it for what they really want. This most marketable commodity then becomes money. While specific money commodities have differed from one society to another, in the most developed countries, precious metals become the money commodity because of their suitable characteristics: their portability, divisibility, scarcity, and so on.

Obviously, in such a theory money cannot be a creation of government because it is a naturally evolved social institution. Government can enhance the acceptability of a money commodity by declaring it legal tender, but government cannot create money. In this way, Menger's theory is meant to solve several long-standing controversies in the theory of money. The nominalist–realist debate is resolved by acknowledging that the value of money commodity is equal to the value of the money (except for small coins where it would be uneconomic to spend the resources to make full-weight coins) but by also pointing out that the actual commodity can be anything consistent with the accepted standards and level of development of the community. The commodity–fiat debate is resolved by showing a role for government in enhancing the acceptability of money even though it originates first through a natural process of human choices.

The last chapter is not the only place in which Menger discusses the origin of an economic phenomenon. All through the *Principles*, Menger is interested in establishing the origin and meaning of phenomena where the meaning is often elucidated through a description of their evolution through time. Erich Streissler (1972, p. 430) has gone so far as to credit Menger with presenting foremost a theory of economic development in the *Principles*. There is much to recommend that position. One of Menger's main themes is that economic development is a process of increasing knowledge and the consequent improvement in the variety and quality of goods available. Economic development is characterized by better communication among traders, more complex trading institutions, more and better commodities, and a greater ability of men to establish ‘economic’ prices.

We can perhaps understand Menger's vision better if we remember how he thought of the human predicament. Man in his original state is ignorant of his environment and uncertain about his (finite) future. He must plan for the satisfaction of his wants in this difficult world, and his primary aid is his ability to learn. The progress of civilization is nothing so much as a process of reducing ignorance and developing institutions that make dealing with the uncertain future more manageable. Smith emphasized the division of labour and capital accumulation as the causes of the wealth of nations. Menger

emphasized the priority of improved knowledge to the improvement of wealth. Indeed, progress is evidenced by ‘the increasing understanding of the causal connections between things and human welfare’ (*Principles*, p. 74).

Methodology

While Menger's *Principles* was well received and eventually became very influential in his native Austria, his theories came in for criticism – or, more to the point, apathy and neglect – in the one audience Menger had most hoped to convince, the German Historical School. While the older members of the historical school, Knies, Roscher and Hildebrand, understood classical economic theory and wanted to overcome its shortcomings with detailed historical investigations which would have the purpose of allowing them to infer them empirical regularities in economic events, the younger Germans, led by Gustav Schmoller, rejected the theory entirely. They believed there could be no such thing as scientific economic theory, and they insisted on viewing an economy as an organic whole at one with politics, law and custom. Menger's new theory, then, was considered not only incorrect, but useless. To Menger, who was convinced that he had discovered the key to unlocking the mysteries of all economic phenomena, such cavalier dismissal must have been particularly galling.

Having failed to make headway with his new theory in Germany on what appeared to be methodological grounds, Menger began work in 1875 on his second book, *Untersuchungen über die Methode der Socialwissenschaften und der politischen Oekonomie insbesondere* (*Investigations into the Method of the Social Sciences with special reference to Economics*). This book, essentially a defence of economic theory and an account of its relationship to historical methods, was published in 1883. Menger's ambition was this time to attract the attention of German academics. This time he succeeded, but unfortunately, the attention he attracted was negative. Gustav Schmoller's review of the *Investigations* was particularly unsympathetic and incited Menger to respond with an impassioned pamphlet entitled *The Errors of Historicism* in 1884. In this pamphlet, Menger dropped all attempts at cordial conciliation and, in Hayek's words, ‘ruthlessly demolished Schmoller's position’ (Hayek, 1981, p. 24). If so, Schmoller never discovered the demolition since he returned the book to Menger unread and wrote a final scathing attack on Menger in his journal.

This exchange has been referred to as the ‘Methodenstreit’ or war of methods, a war that at the time seemed to have no clear winners and certainly led to no resolution of the opposing views. Ultimately, of course, Menger's position was far closer to the methodological turn economics took in the subsequent century, although in Germany, Menger's approach and the school that formed around it remained excluded from the university curriculum well into the 20th century.

The vehemence and hostility with which the Germans greeted Menger's *Investigations* is to some degree surprising. Far from an attempt to displace the approach of the Historical School, Menger's *Investigations* is a conscious plan for incorporating many of the features of the historical–empirical approach into a more comprehensive general methodology. (Although it must be admitted that Menger's tone is not always cordial when discussing the mistaken views of the Historical School.) Menger divides economics into three parts: the historical–statistical which investigates the individual nature and individual connection of economic phenomena, the theoretical which investigates the general nature and general connections of phenomena, and the ‘practical sciences of national economy’, the basic principles for suitable action in the field of national economy, or in modern terminology, economic policy

(*Investigations*, pp. 38–9). Menger defends the idea that science requires knowledge *both* of individual (or concrete) aspects of phenomena and of the general (formal) aspects. Presumably, the methods of the Historical School are appropriate to the investigation of concrete aspects of economic phenomena while economic theory is necessary to understand the general aspects. The general form of things, Menger calls *types* and the general form of relationships, Menger calls *typical*.

Menger defends the scientific quality of economic theory despite the fact that its laws are not as strict as some other sciences may be. All sciences, Menger argues, show varying degrees of strictness, and ‘the number of natural sciences which absolutely comprise strict laws of nature is also small, and the value of those which only show empirical laws is nevertheless beyond question’ (*Investigations*, p. 52). Economic science develops exact laws, but the observation of these laws in reality is hindered by the complexity of the events in which they are manifested and by the impingement of non-economic goals on the actions of observable human beings. Hence, one can never refute the exact laws of economics by pointing to contrary empirical cases. Such a procedure would be analogous to testing the laws of geometry by measuring triangular shapes. In any case, the fact that economic laws are not as strict as some other sciences is irrelevant to its scientific character.

The problem of economic science is to find the causal laws of typical events even though they are manifested in complex reality. Hence it is necessary to ‘ascertain the simplest elements of everything real, elements which must be thought of as strictly typical just because they are the simplest’ (*Investigations*, p. 60). The appropriate procedure, then, is to start with the simplest elements of economic phenomena and from there investigate the laws by which more complicated human phenomena are formed from simplest elements. Menger called this the ‘causal–genetic’ approach. Obviously, the simplest elements of economic theory are human valuations and from this can be derived the more complicated economic relationships that are observable in the real world. While Menger does not call this approach ‘methodological individualism’, it is clear from his discussion of the exact approach and his later criticisms of the excesses of the organic approach that he is a methodological individualist where that means explaining economic phenomena in terms of the choices and consequences of individual human valuation.

Menger's example that he uses to contrast the exact approach with the ‘realistic–empirical’ is particularly interesting since it clarifies a point of debate about his use of equilibrium constructs. He claims that the exact method can be used to predict ‘economic prices’ even though one rarely observes true economic prices in the real world. The four criteria for prices to be ‘economic’ are that (*a*) individuals protect their economic interests completely; (*b*) people have complete knowledge about their goals and their means to achieving them; (*c*) they know the full economic situation (complete knowledge about quantities offered for sale and what prices are being charged) and (*d*) they have the freedom to act in their own interests according to their knowledge. It does not take much imagination to see in these requirements a form of perfect competition where complete knowledge and freedom of entry and exit allow economic man full scope to arrive at equilibrium prices. However, while the laws which predict economic prices are true and exact, the empirical manifestation of them will vary due to circumstances. Indeed, Menger argues that it would be surprising indeed if any of the circumstances required for the establishment of ‘economic’ prices were ever met completely in the real world. Real prices will deviate from economic prices, and the role of the realistic–empirical approach, then, is to discover to what degree real prices deviate from economic prices. The realistic–empirical approach, however, must take the exact theory of economic prices as the point of departure.

While Menger insists on the necessity of an exact theory of economics to understand economic phenomena, it is clear that he does not believe economics is an all-purpose science. Economics provides exact laws, but only of a subset of human action. In answer to the charge that his vision of human experience is too limited, he emphasizes that a full understanding of social phenomena requires the aid of the totality of exact sciences of man as well as the historical context of the actions. He is also careful to point out that his assumption of economic man – man guided exclusively by self-interest – is a fiction that does not capture real action. The theory of political economy ‘teaches us to follow and understand in an exact way the manifestations of human self-interest in the efforts of economic humans aimed at the provision of their material needs’ (*Investigations*, p. 87) but this provides understanding of a special side, by no means the only side, of human life.

One of the criticisms of economic theory that Menger attempted to answer was the charge that pure theory ignored the reality of development and change in economic life and failed to take account of the organic nature of real economic phenomena. While Menger in principle acknowledged the importance of change brought about in time both to empirical forms and to strict types, he believed that the way to explain such change was always with reference to exact theory. In fact, those who discussed organic development missed one of the most important sources of institutional change in social organization. In the *Principles*, Menger had developed a theory of the origin of money as an unintended social order. In the *Investigations*, Menger generalized his theory to encompass many different social forms. The Historical School's emphasis on historical development required a theory of development, a theory that explained how institutions arise from the unintended consequences of human attempts to improve their own well-being.

Menger saw the problem of exact research to be to discover ‘how institutions which serve the common welfare and are extremely significant for its development come into being without a *common will* directed toward establishing them?’ (*Investigations*, p. 146). His answer, developed using examples of such social institutions as money, law, language, markets, the origin of communities and of the state itself, was that individuals following their own economic interests provide spillovers to others in the form of increased knowledge of potential advantages or increased ability to pursue their interests. Money, as we have already learned, arises as individuals attempt to overcome the difficulties of barter by acquiring more saleable commodities for the purposes of trade. New localities develop as individuals of different abilities and different professions settle in new areas because they believe they have a better market for their skills. States mostly came into being as families living in close proximity to each other decided it was to their advantage to unite. Most such social organization, Menger argued were not the consequences of conscious planning, but the unconscious result of human will directed toward other, more personal ends. This is the nature of organic development in social science.

What makes Menger's discussion of ‘organic’ orders (or ‘spontaneous orders’ as Hayek was later to call them) particularly interesting, is the fact that he not only describes them, but he also provides a brief theoretical analysis of how they can develop. He mentions in his theory of the origin of money that some individuals will be quicker than others to recognize the advantages of acquiring more marketable commodities because it helps them to come closer to their own ends. Not everyone will discover the advantages of indirect exchange at once, but they will soon learn because ‘there is no better means to enlighten people about their economic interests than their perceiving the economic successes of those who put the right means to work for attaining them’ (*Investigations*, p. 155). It does not take much of a leap to interpret Menger's theory as describing the development of an organic order as a process of

discovery and transmission of new information through imitation, motivated by the interests of economic persons. Menger's theory of unintended organic institutions is thus an attempt to reconcile the organic and developmental approach to economics with the exact laws of economic science. Compared to the frenetic publishing activity of a 20th-century economist, Menger published relatively little during his long career. Nevertheless, he had a major influence on the history of economic thought primarily because he attracted a number of bright and ambitious students. Although his two major disciples, Friedrich Wieser and Eugen Böhm-Bawerk, were never technically his students (both had studied at the University of Vienna before Menger began teaching there), they were clearly his students in the most important sense: they absorbed and finally extended major aspects of the work of the master. Wieser worked specifically on the problem of imputation which led him to be the first to use the term ‘opportunity cost’, the utility of the forgone alternative. Wieser also extended Menger's notion of national economy in ways that brought him closer to the to the general equilibrium school. Böhm-Bawerk is best known for his development of Menger's suggestions about the importance of time in production and the implication of goods of higher order for a theory of the structure of production. While Wieser and Böhm-Bawerk were the best known of Menger's students, there were many others who gathered around him and formed a school. Those who published works in the Austrian tradition included Emil Sax, Johann von Komorzynski, Robert Zuckerkandl, and H. von Schullern-Schrattenhofen. Although not directly his student, Ludwig von Mises (who actually studied under Böhm-Bawerk) made his first major contribution to economics by extending Menger's notion of marginal utility combined with Menger's process analysis to develop a theory of the value of money. Friedrich Hayek, a student of von Mises, later developed Menger's ideas of spontaneous orders and the problem of knowledge into a comprehensive social theory. Both Mises and Hayek, in turn, have inspired a number of contemporary economists to work in the tradition of Menger to reformulate modern economics in a more ‘Austrian’ form.

See Also

- Austrian economics

Selected works

1871. *Grundsätze der Volkswirtschaftslehre*. Trans. J. Dingwall and B.F. Hoselitz as *Principles of Economics* by with an Introduction F.A. Hayek. New York and London: New York University Press, 1981.

1883. *Untersuchungen über die Methode der Sozialwissenschaften und der politischen Ökonomie insbesondere*. Trans. F.J. Nock as *Problems of Economics and Sociology*, edited and with an Introduction by L. Schneider Urbana: University of Illinois Press, 1963. Reprinted as *Investigations into the Method of the Social Sciences with Special Reference to Economics* with a new Introduction by L.H. White, New York and London: New York University Press, 1985.

1884. *Irrthümer des Historismus in der deutschen Nationalökonomie*. Vienna: Hölder.

1887. *Zur Kritik der politischen Ökonomie*. Vienna.

1888. Zur Theorie des Kapitals. *Conrad's Jahrbücher für Nationalökonomie und Statistik* 17, 1–49.

1889. Grundzüge einer Klassification der Wirtschaftswissenschaften. *Conrad's Jahrbücher für Nationalökonomie und Statistik* 14.

1892. Geld. In *Handworterbuch der Staatswissenschaften*, vol. 3. Vienna.

1892a. Die Valutaregulierung in Österreich-Ungarn. *Conrad's Jahrbücher fur Nationalökonomie und Statistik* 3.

1892b. *Der Übergang zur Goldwahrung. In Untersuchungen über die Wertprobleme der österreichisch-ungarischen Valutareform*. Vienna.

1892c. La monnaie mesure de la valeur. *Revue d'économie politique* 6.

1892d. On the origin of money. *Economic Journal* 2(6), 239–55.

Bibliography

Alter, M., 1982. Carl Menger and Homo Oeconomicus: some thoughts on Austrian theory and methodology. *Journal of Economic Issues* 16(1), 149–60.

Bloch, H.-S., 1940. Carl Menger: the founder of the Austrian school. *Journal of Political Economy* 48, 428–33.

Hayek, F.A. 1981. Carl Menger. Introduction to Carl Menger, *Principles of Economics*. New York and London: New York University Press.

Hicks, J.R. and Weber, W., eds. 1973. *Carl Menger and the Austrian School of Economics*. Oxford: Clarendon Press.

Jaffé, W. 1976. Menger, Jevons and Walras de-homogenized. *Economic Inquiry* 14, 511–24.

Kauder, E. 1957. Intellectual and political roots of the older Austrian school. *Zeitschrift für Nationalökonomie* 17, 411–25.

Kauder, E. 1959. Menger and his library. *Economic Review* 10 (Hitotsubashi University).

Kauder, E. 1965. *A History of Marginal Utility Theory*. Princeton: Princeton University Press.

Kirzner, I.M. 1979. The entrepreneurial role in Menger's system. In *Perception, Opportunity and Profit: Studies in the Theory of Entrepreneurship*, ed. I.M. Kirzner. Chicago: University of Chicago Press.

Martin, D.T. 1979. Alternative views of Mengerian entrepreneurship. *History of Political Economy* 11, 271–85.

Mises, L. von. 1978. Carl Menger and the Austrian school of economics. In *The Clash of Group Interests and Other Essays*. ed. L. von. Mises. New York: Center for Libertarian Studies.

Schumpeter, J.A. 1951. Carl Menger, 1840–1921. In *Ten Great Economists, From Marx to Keynes*. ed. J. A. Schumpeter. New York: Oxford University Press.

Stigler, G.J. 1941. Carl Menger. In *Production and Distribution Theories*, ed. G.J. Stigler. New York: Macmillan.

Streissler, E. 1972. To what extent was the Austrian School marginalist? *History of Political Economy* 4, 426–41.

Wagner, R.E. et al. 1978. Carl Menger and Austrian economics. *Atlantic Economic Journal* 6(3), Special Issue. Contributions by R.E. Wagner, S. Bostaph, L.S. Moss, I. Kirzner and H. Nelson Gram and V.C. Walsh, L.M. Lachmann and K.I. Vaughn.

Bibliographic addendum

See also M. Alter, *Carl Menger and the Origins of Austrian Economics*, Boulder, CO: Westview Press, 1990. Different facets of Menger's work are discussed in B.

Caldwell, ed., *Carl Menger and His Legacy in Economics*, Durham, NC: Duke University Press, 1991. M. Latzer and S. Schmitz, eds, *Carl Menger and the Evolution of Payments Systems*, Cheltenham: Edward Elgar, 2002, provides the first English translation of Menger's ‘Geld’ as well as essays both evaluating Menger's views on monetary systems and applying them to contemporaneous issues.

How to cite this article

Vaughn, Karen I. "Menger, Carl (1840–1921)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000141> doi:10.1057/9780230226203.1082

The New Palgrave Dictionary of Economics Online

mercantilism

Laura LaHaye

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mercantilism is economic nationalism that seeks to limit the competition faced by domestic producers. It refers to the economic thought and policies that were characteristic of the dominant Western European trading nations during the transition from feudalism to modern capitalism from the 16th to the late 18th century. It is often depicted as the school of thought that confused money with wealth, promoting a favourable balance of trade as the best method to increase the wealth of a nation that did not possess gold or silver mines.

Keywords

Colbert, Jean-Baptiste; East India Company; German Historical School; Heckscher, E. F.; mercantilism; money; Monopoly; Mun, T.; Navigation Acts; protection; specie; tariffs

Article

Mercantilism is economic nationalism that seeks to limit the competition faced by domestic producers. The tools of mercantilist policies include the granting of monopoly privileges, regulation of prices and business practices and especially prohibitions, tariffs, subsidies and other regulations regarding the conduct of international trade. The goals of mercantilism are supposedly to contribute to the development of a rich and powerful state; however, the principal beneficiaries are the merchants and producers who are protected or encouraged under a mercantile system. Although mercantilism was frequently promoted as means of obtaining long-term development objectives, it is significant that such promotion typically increased in fervour following periods of trade crisis, such as that in England in the 1620s.

Mercantilism refers to the economic thought and policies that were characteristic of the dominant western European trading nations during the transition from medieval feudalism to modern capitalism from the 16th to the late 18th century. Adam Smith (1776, p. 399) characterized the ‘principle of the

commercial or mercantile system' – that a 'favourable' balance of trade would bring gold or silver into the country – which could be used to 'carry on foreign wars, and to maintain fleets and armies in distant countries'. Import restraints and encouragement to exportation were the mercantile policies that would enrich and empower the newly emerging nation-states. At the end of the 19th century authors of the German Historical School popularized the term 'mercantilism' while rationalizing the mercantile policies as necessary for the unification of feudal power centres by large competitive states.

The mercantile era emerged following the discovery of the New World and the East Indies by European explorers at the close of the 15th century. Shipping and trading grew in importance during this period as did the frequency of military battles at sea and in the colonies. Anglo-French rivalry remained intense, and Henry VIII invested heavily in shipping while fortifying the coastline against possible attack. Meanwhile, the Spanish Habsburgs were at war all over Europe. Mercantile economic warfare complemented the military objectives of the antagonistic nations and served to unify each nation against an external threat.

As a concept of society, mercantilism reflects the medieval view that wise government intervention is necessary to delicately balance the tendencies of unbridled competition to produce unjust wages or income below a subsistence level, when too many workers or businesses operate in a particular activity, or to result in an unregulated monopoly that would reap unjust profits charging prices that are too high. The market could certainly not be left to itself to find a 'just price' or wage. The 1563 Statute of Artificers marked one of the first efforts by Queen Elizabeth of England to extend the restrictive and regulatory policies of medieval towns to the nation as a whole. A century later, Louis XIV of France, with the assistance of his powerful mercantilist finance minister Jean-Baptiste Colbert, undertook similar national regulation of industry and simplification of the internal tolls of France which Heckscher (1935, vol. 1, p. 103) 'ranks with Elizabeth's Statute of Artificers as one of the two unquestionable triumphs of mercantilism in the sphere of economic unification'.

The granting of monopoly privileges was a relatively more important form of state protection during the earlier part of the mercantile era. The British East India Company was granted a monopoly charter by Queen Elizabeth in 1600 which encouraged the United Provinces to consolidate the independent Dutch traders into the Dutch East India Company in 1604. A number of short-lived East Indies trading monopolies were chartered by the French Crown throughout the first half of the 17th century, culminating in the 1664 charter of, and royal participation in, the French East Indies Company. These privileges were intended to benefit the developing shipping and long-distance trading industries themselves as well as to provide revenues to the state either directly, in the case of state monopolies, or indirectly through modest duties on imports of the private monopolies.

When, however, the successful conclusion of the Dutch Revolt in 1648 exposed the English to an increased level of competition in intra-European shipping and trading, Cromwell eventually responded with the first Navigation Act of 1651. This Act stipulated that all goods imported into England or her territories had to be carried in English ships, unless they were carried directly from a European country of origin on ships owned and crewed by citizens of that country of origin, and that no foreign vessels could engage in the coastal trade among English ports. Furthermore, no type of salted fish or fishing by-product of the type usually caught and processed by English people could be imported unless it was caught and processed by an English ship. Additional navigation laws further protected English fishing, shipping and trading industries from competition, especially from the Dutch, who largely dominated maritime activity at the time.

More general industrial protection followed the navigation laws, although several early examples of discriminatory protective policies were already in existence. The 1667 anti-Dutch tariff imposed by Colbert in France, and the subsequent quadrupling of import duties in England during the 15 years following the 1688 accession to the throne of William III and Mary marked the major shift from moderate revenue-generating customs duties on imports and exports to the more protective import tariffs as well as bounties and drawbacks on exports that constituted the mercantile system in Smith's view. English export duties on woollens were abolished in 1700 and export duties were abolished in general by the Walpole customs reform of 1722. Protection was further extended throughout the 18th century until ‘the building up of the protective system showed signs of becoming a general and recognized policy … in the decade in which Adam Smith was collecting material and writing his great blast against commercial regulation, *The Wealth of Nations*’ (Davis, 1966, p. 314).

Following Smith's (1776, p. 418) lengthy examination of the ‘popular notion that wealth consists in money’, mercantilism has often been depicted as the school of thought that confused money with wealth. Although this interpretation has been thoroughly debated, there is certainly much evidence to suggest that mercantile pamphleteers did believe an inflow of precious metals would increase the wealth of the nation and that foreign but not domestic or internal trade was the only way to increase the wealth of a nation that did not possess gold or silver mines. Exportation of bullion or coin had generally been regulated or prohibited since medieval times, and it was in an effort to get those restrictions relaxed that mercantilist authors such as Mun (1664, p. 5), a director of the British East Indies Trading Company, argued that the ‘means therefore to increase our wealth and treasure is by *Foreign Trade*, wherein we must ever observe this rule; to sell more to strangers yearly than we consume of theirs in value’. That the wealth of the nation was not perceived to be primarily related to its ability to provide goods and services to its consumers is revealed when reading Mun's (1664, p. 7) recommendations for reducing imports such as using waste grounds ‘to supply our selves and prevent the importations of Hemp, Flax, Cordage, Tobacco and divers other things which we now fetch from strangers to our great impoverishing’.

In all fairness, the proponents of an export surplus did not generally advocate the accumulation of specie for the simple purpose of hoarding it, although they did like to make the analogy between the kingdom and an individual that would grow poor if its purchases exceeded its income. Of course, neither the individual nor the kingdom will grow poor if the purchases include investment expenditures that yield a rate of return in excess of the borrowing cost. As a store of value, money is only a component of wealth to the extent that one intends to spend it one day, and there is a limit to this precautionary motive for accumulating specie. It is sensible to accumulate specie following a period of declining reserves (excessive expenditure) or in response to increased uncertainty, which requires a larger precautionary balance, or in response to increased hostility, which requires a larger defence balance, but not ad infinitum, except perhaps to maintain a desired ratio of specie to growing royal expenditures over time.

For the merchant adventurers engaged in long-distance trading, specie was a valuable factor of production as a medium of exchange, and they recognized the relationship between the quantity of money in circulation and the amount of trading activity that could be financed. Mun (1664, p. 68) was careful to recommend that the royal treasure should not be augmented by more than the favourable balance of trade, ‘for if he should mass up more money than is gained by the over-balance of his foreign trade, he shall not *Fleece*, but *Flea* his Subjects, … whereby the life of lands and arts must fail and fall to the ruin both of the public and private wealth’. This indicates that he perceived a relationship between the quantity of money and the level of national economic activity, although his immediate concern was

probably the economic activity of his own British East India Company, which imported exotic goods that could not be produced at home.

More important, perhaps, than enabling the royal treasure to be augmented, an export surplus is generally perceived to stimulate domestic employment directly or indirectly by reducing interest rates. According to Heckscher (1935, vol. 2, p. 121), the '"fear of goods" was nourished … by the idea of creating work at home and of taking measures against unemployment'. References to the unemployment argument date back to the early 15th century, and in English legislation in 1455, 'foreign competition was blamed for having caused the unemployment in the silk industry' (Heckscher, 1935, vol. 2, p. 122). The preference for encouraging exportation of manufactured consumer goods, as opposed to raw materials or productive equipment, and allowing the importation of raw materials are consistent with this employment concern. An export surplus – an excess of domestic saving over investment – naturally arises when productivity growth outpaces the growth of profitable domestic investment opportunities, and this may include an accumulation of international reserves to finance the growth of monetized transactions; but to try to engineer such a surplus with protective trade policies would be futile at best. In addition to competitively induced innovation and increased specialization, limited by the extent of the market, productive investment is the true source of a sustainable increase in the wealth of a nation, and there is no reason to suppose, a priori, that domestic investment is inferior to foreign investment.

Most of the vestiges of the mercantile era were removed during the laissez-faire era of the 19th and early 20th centuries, especially in England, where monarchical power was weaker and property rights were clearer than in France and Spain. Yet mercantilism has remained a topic of considerable debate, especially since Heckscher's broad treatment of the subject and the emergence of global depression in the 1930s (Heckscher, 1935; Viner, 1937; Minchinton, 1969; Coleman, 1969; Magnusson, 1993). Whether mercantilist policies re-emerge in the 21st century will depend on the institutional framework within which the special interests seeking protection must function (Ekelund and Tollison, 1997), as there exists no coherent economic doctrine to support such policies.

See Also

- cameralism
- Colbert, Jean-Baptiste
- Hume, David
- Misselden, Edward
- Mun, Thomas
- Schmoller, Gustav von

Bibliography

Coleman, D. 1969. *Revisions in Mercantilism*. London: Methuen.

Davis, R. 1966. The rise of protection in England, 1689–1786. *Economic History Review* 19, 306–17.

Ekelund, R. and Tollison, R. 1997. *Politicized Economies: Monarch, Monopoly and Mercantilism*. College Station: Texas A & M University Press.

Heckscher, E. 1935. *Mercantilism*, 2 vols. London: Allen & Unwin.

LaHaye, L. 2007. Mercantilism. In *The Concise Encyclopedia of Economics*, ed. D. Henderson. Indianapolis: Liberty Fund.

Magnusson, L. 1993. *Mercantilist Economics*. Boston: Kluwer.

Magnusson, L. 1994. *Mercantilism, the Shaping of an Economic Language*. London: Routledge.

Minchinton, W. 1969. *Mercantilism, System or Expediency?* Lexington: Raytheon.

Mun, T. 1664. *England's Treasure by Foreign Trade or The Balance of our Foreign Trade is the Rule of our Treasure*. London. Reprints of Economic Classics, New York: Kelley, 1968.

Smith, A. 1776. *The Wealth of Nations*. Edwin Cannan edn. New York: Random House, 1937.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

How to cite this article

LaHaye, Laura. "mercantilism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000144> doi:10.1057/9780230226203.1083

The New Palgrave Dictionary of Economics Online

Mercier De La Rivi è re, Pierre-Paul (Mercier or Lemercier) (1720 – 1793/4)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balance of trade; Mercier de la Rivière, P. P.; money; Physiocracy; private property; taxation theory; value theory, Physiocratic

Article

Lawyer, administrator and economist, born into a financier's family in 1720. From 1749 to 1759, he was Councillor of the Paris Parlement; from 1759 to 1764, Governor of Martinique. Although Garnier (1854, p. 188) claims that Mercier became acquainted with Quesnay and Mirabeau while Governor of Martinique, this is doubtful. However, after 1765 he became a prominent Physiocrat and published what many (for example, Smith, 1776, p. 679; Mill, 1824, p. 712) considered to be the most comprehensive exposition of Physiocratic doctrine in his *L'ordre naturel et essentiel des sociétés politiques* (1767). This gained him both Catherine the Great's invitation to advise her on a new legal code and the enmity of Voltaire (1768), who devastatingly satirized his cumbrous prose. Du Pont (1768) wrote a summary of Mercier's work for *Ephémérides*, confirming thereby its enormous importance for the Physiocrats. Subsequently, Mercier published a reply to Galiani's dialogues attacking the Physiocratic position on the grain trade (1770) and an essay on the importance of public education dedicated to the King of Sweden (1775). He died in Paris in either 1793 or 1794.

Mercier's *L'ordre naturel* (1767) is therefore the major general treatise of Physiocratic doctrine both political and economic. The work divides into three parts with a concluding summary chapter. Part I develops the theory and necessity of the social order based on the duties and rights inherent in private property, without which a society cannot be sustained. ‘The greatest possible happiness comes from the greatest possible abundance of means of enjoyment and the greatest possible freedom to profit from [the ownership of property]’ (1767, I, pp. 42–3). Hence the sanctity of private property and complete freedom for its owners to use it are the first principles of the theory of natural order (pp. 45, 50–51). These principles need to be inculcated in society through a system of public instruction (pp. 91–2). Part II discusses the manner in which social order is achieved in practice through the establishment of three

fundamental institutions: law and magistrature, the sovereign as bearer of authority, and institutions of public instruction for spreading knowledge of the social order among all members of society. In his lengthy elaboration on these institutions (chs 11–24) Mercier presents his famous defence of legal despotism.

Part III (the greater part of Volume 2 in the original edition) further discusses the practical promotion of the social order by examining the political economy of wealth creation. After reviewing the essential association between the king and his subjects (ch. 25) the theory of taxation is presented as the way in which kings share the net product of their common property with the landlords (chs 28–34). The dogmatic presentation of Physiocratic tax theory was the special target of Voltaire (1768). These chapters also contain interesting economic contributions. In them Mercier emphasizes the role of consumption and effective demand in stimulating reproduction (vol. 2, pp. 138–9); presents an argument showing the possibility of a downward spiral in economic activity ‘in geometrical progression’ if taxation reduces the advances of agriculture (pp. 150–1), an analysis having both real and value aspects (pp. 160–4). The second half of Part III examines commerce and industry and their function in the Physiocratic social order (chs 35–43), the last chapter being a particularly dogmatic demonstration of these activities’ unproductive nature. However, they likewise contain interesting analytical contributions on the role of money and its circulation (vol. 2, pp. 262–3, 297–9, 334), the impact of trade on wealth via the profits of agriculture and hence accumulation when it provides a wider market for agricultural produce (p. 273) and a critique of the balance of trade doctrine based on the logical impossibility for all nations to enjoy a favourable balance (p. 349) and a type of specie mechanism argument (pp. 360–7) from which Mercier concludes that nations can have too much as well as too little money (pp. 368–9). His discussion of commerce and industry highlights, in particular, the richness of Physiocratic value theory and its importance for their theory of distribution and economic development. As Vaggi (1987) has demonstrated, recognition of this importance is indispensable for a proper understanding of Physiocracy, as is the full social and political framework in which their policy recommendations are framed and for which Mercier was particularly noted by his contemporaries.

Selected works

1767. *L'ordre naturel et essentiel des sociétés politiques*. London and Paris.

1770. *L'intérêt général de l’état, ou la liberté du commerce des blés, avec une réfutation d'un nouveau système publié par l'abbé Galiani en forme de dialogues sur le commerce des blés*. Amsterdam and Paris.

1775. *De l'instruction publique, ou considérations morales et politiques sur la nécessité, la nature et la source de cette institution*. Stockholm and Paris.

Bibliography

Du Pont de Nemours, P.S. 1768. De l'origine et des progrès d'une science nouvelle. Reprinted in *Physiocrates*, vol. 1, ed. E. Daire, Paris, 1846.

Garnier, J. 1854. Mirabeau. In *Dictionnaire de l'économie politique*, vol. 2, ed. Ch. Coquelin and Guillaumin. Paris.

Mill, J. 1824. Economists. Supplement to *Encyclopaedia Britannica*, vol. 3. Edinburgh.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. R.H. Campbell, A. S. Skinner and W.B. Todd. Oxford: Clarendon Press, 1976.

Vaggi, G. 1987. *The Economics of François Quesnay*. London: Macmillan.

Voltaire, F.M.A. de. 1768. *The Man of Forty Crowns*. London, trans. from the French.

How to cite this article

Groenewegen, Peter. "Mercier De La Rivière, Pierre-Paul (Mercier or Lemercier) (1720–1793/4)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000146> doi:10.1057/9780230226203.1084

The New Palgrave Dictionary of Economics Online

Mercosur

Marcelo Olarreaga
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mercosur is an ambitious economic integration project, launched in 1991, which includes Argentina, Brazil, Paraguay and Uruguay. The early and quasi-complete liberalization of intra-regional trade and the adoption of a common external tariff by 1996 were accompanied by significant increases in intra-regional trade. However, the most difficult and challenging steps towards a common market (its original objective) has been slow since then, in part due to the absence of strong regional institutions.

Keywords

Andean Community; central-bank independence; common external tariffs; comparative advantage; customs unions; foreign direct investment; free-trade agreements; intra-regional trade; Mercosur; monetary union; non-tariff barriers; tariffs; trade diversion; World Trade Organization

Article

Mercado Común del Sur (Mercosur, Southern Common Market) is an ambitious economic integration project which includes Argentina, Brazil, Paraguay and Uruguay. It represents 70 per cent of the gross domestic product (GDP) of South America and 60 per cent of its population. In terms of geographic size, Mercosur is four times larger than the European Union, which would rank Mercosur as the largest customs union in the world. Its economic size, however, is similar to that of the Netherlands.
Mercosur was launched in March 1991 with the signing of the Asunción Treaty. Aiming at creating a common market, Article I calls for full internal mobility of goods, services and factors of production, the implementation of common external policies in these areas, as well as the coordination of macroeconomic policies and cooperation in education, health and transport policies.
It is an agreement that is open to accession by all members of the Latin American Integration Association (which regulates partial bilateral trade agreements among members). By 1996 Bolivia and Chile were associate members of Mercosur; later, a free-trade agreement (FTA) was signed with the

Andean Community. At the time of writing other Latin American countries are in different stages of association with Mercosur. Negotiations for trade agreements are ongoing with China, the European Union, Mexico, India, South Africa, Egypt, and Morocco.

Mercosur members had agreed in the Asunción Treaty to create the common market within four years. However, this proved politically impossible and little progress was made a part from very rapid reductions in internal tariffs (with some negotiated exceptions). Very quickly it became clear that the ambitious objectives of the Asunción Treaty had to be scaled back. An imperfect customs union became a more realistic objective, and the Protocol of Ouro Preto signed in December 1994 called for the implementation of a common external tariff (CET) by early 1995. It was an imperfect ‘common’ external tariff as each member was allowed some deviations from the negotiated CET; and more than ten years later the CET is still to be defined in some politically entrenched sectors (such as sugar). Nevertheless, by 1996 internal tariffs were applied on less than three per cent of tariff lines, and the CET was implemented in 80 per cent of tariff lines.

In all other areas progress has been slow or non-existent. For example, non-tariff barriers (NTBs) are not only not subject to common external policies but are routinely used as an impediment to intra-regional trade, contrary to what is explicitly required in Article V of the Asunción Treaty. For example, non-automatic import licensing, sanitary measures and other technical regulations (such as labelling) on Brazilian imports of powdered milk impose an equivalent tax of 54 per cent on Argentina's exporters (Berlinski, 2004). Internal trade in the automobile sector is managed with bilateral trade quotas at the firm level (for those firms with a presence in several Mercosur members) and a trade balance constraint on global automobile trade, which if removed could double bilateral trade (Brambilla, 2005).

Negotiations on services trade and factor mobility were still at a very early stage 15 years after the treaty was signed. The Services Trade Protocol signed in 1997 merely states the multilateral commitments of Mercosur members at the World Trade Organization (WTO). The dispute settlement mechanism (DSM) remained unused until 1997; an appeal court was created only in 2002. Steps have been taken for the mutual recognition of standards, but enforcement has been largely absent (for example, in the area of education, mutual recognition stops at the high-school level). Macroeconomic coordination is limited to routine exchange of (public) information.

Internal tariffs and the FTA

In spite of the slow progress in the ‘non-tariff’ areas (NTBs, services, factor mobility, macroeconomic coordination), by the late 1990s Mercosur was considered one of the most successful attempts at regional integration between developing countries. This was partly due to the unprecedented rapid elimination of internal tariffs, a sixfold increase in intra-regional trade, a twentyfold increase in flows of foreign direct investment (FDI) (mainly from the United States and Spain), and the longevity that the agreement was achieving in spite of several financial, economic and institutional crises.

A more careful analysis, however, reveals a more subtle picture. Let me start with the rapid increase in trade. Yeats (1998) argued that intra-regional trade appeared to be concentrated in products in which Mercosur did not have a clear comparative advantage (capital goods), and that these were the goods with the most rapid growth after the creation of Mercosur. He concluded that this provided evidence of trade diversion and should raise questions regarding the (static) welfare impacts of such rapid growth in intra-

regional trade. Olarreaga and Soloaga (1998) showed that fast-growing intra-regional trade was concentrated on products with trade-diverting potential partly because deviations from zero internal tariffs occurred in products with substantial trade-creation potential, as predicted by the theoretical political economy literature on regional agreements.

External tariffs and the CET

It has also been argued that a significant part of the increase in intra-regional trade need not be attributed to the creation of Mercosur, but rather to the tremendous trade liberalization vis-à-vis the rest of the world that Mercosur members were independently undertaking after the mid-1980s. For example, Brazil's external tariff declined from an average of 80 per cent in the mid-1980s to an average of 15 per cent by the mid-1990s. This can explain a large share of the rapid growth in imports, including those from other Mercosur members. On the other hand, it has been suggested that the important external liberalization undertaken by Mercosur members needs to be partly attributed to the creation of Mercosur. Without the significant competitive pressure imposed by the increase in intra-regional flows, the move towards lower external tariffs would have been more difficult. Bohara, Gawande and Sanguinetti (2004) showed that the lobbying for high external tariffs was eroded by the increase in intra-regional trade due to internal tariff preferences. Also, it has been shown that a significant force for lower CET was the prospect of the elimination of duty drawbacks for intra-regional exports (a by-product of the creation of Mercosur) as agreed in Ouro Preto. Indeed, the elimination of duty drawbacks on intra-regional exports increased counter-lobbying by regional exporters for lower tariffs on their imports of intermediate inputs from the rest of the world. This led to a 25 per cent reduction in the negotiated CET (Cadot, de Melo and Olarreaga, 2003).

An additional trade-related benefit for Mercosur members is that rest-of-the-world exporters to the regional market started pricing their products more competitively due to the more intense competition in the internal market brought by tariff preferences granted to other Mercosur members. This led to significant welfare gains for Mercosur consumers of imported products at the expense of foreign firms exporting to the region (Chang and Winters, 2002). Schiff and Chang (2003) further showed that the pro-competitive forces that led rest-of-the-world exporters to price more competitively after the creation of Mercosur were also present even when Mercosur partners did not export to each other, as long as they had the potential to do so (that is, markets were contestable). Thus, Mercosur created trade-related gains to its members even in the absence of any intra-regional trade flow or external tariff reduction.

Beyond tariffs

Moreover, regardless of whether Mercosur led to trade diversion, it has been shown that most households and in particular poor households within the region benefited from the agreement. Porto (2006) provided evidence of a pro-poor bias of Mercosur in Argentina: on average, poor households gain more from the reform than middle-income households, whereas the effects on rich families are positive but not statistically significant. Prior to Mercosur, Argentine trade policy protected the rich over the poor. As relative pre-Mercosur tariffs are higher on relatively skill-intensive goods, the tariff removals tend to benefit the poor over the rich. Thus, Mercosur not only helps reduce poverty in

Argentina, but it improves the distribution of income.

Regarding the rapid increase in FDI flows, it seems that the creation of Mercosur was not the main cause. Most statistical analysis shows no direct causality between the creation of Mercosur and the rapid growth in FDI (Castilho and Zignago, 2002). The main forces were the simultaneous privatization processes in Argentina and Brazil, the macroeconomic stabilization and the external tariff reduction independently undertaken by Mercosur members, which provided foreign firms investing in the region access to imported inputs (Chudnovski, 2001). The creation of a larger regional market only marginally contributed to foreign firms' decisions to invest in Mercosur.

The longevity of Mercosur has come at the cost of achievements in the area of internal free trade, the implementation of the CET, and an (implicit) consensus to move slowly in other areas. For example, at the end of 1992 Argentina increased its statistical import tax surcharge (applied to intra-Mercosur imports) from three to ten per cent as its trade deficit with Brazil widened. An optional increase in external tariffs of up to three percentage points was authorized in 1997. In June 2001, on the eve of a major financial and fiscal crisis, the Argentine government unilaterally altered its tariff rates on capital goods and consumer goods. A waiver was granted by the Common Market Council. In 2006, duty drawbacks and temporary admission regimes which were to be eliminated by 2000 were still in place; the customs code drafted in 1994 had not been adopted by any of the members' parliaments; and no common safeguard mechanism had been put in place to deal with unforeseen changes in competitive pressures, leading to the adoption of unilateral ad hoc measures and private sector marketing agreements (for example, dairy, paper and steel) after the devaluation of the real in January 1999. In sum, flexibility rather than consistency has been the norm, and time-inconsistent policies have often been reversed with the associated cost for the credibility of Mercosur institutions (Bouzas, 2002).

Regional institutions

From the very beginning Mercosur decisions were driven by national private-sector interests, and weak and relatively politicized regional institutions emerged, partly because Brazil (the largest member) wanted to preserve its hegemony. Mercosur is ruled by a Consejo del Mercado Común (CMC, Common Market Council), which is responsible for the political decisions of the integration process. Sitting members are the four national presidents and their cabinets, who regularly meet twice a year. The Grupo Mercado Común (GMC, Common Market Group) is directly answerable to the CMC and is the executive organ, which includes the ministers of foreign affairs and economics, the chairmen of the central banks, and the permanent coordinators from each member country. The GMC enforces resolutions. The GMC branches out into the Trade Commission of Mercosur, which is responsible for counselling and enforcing trade policy instruments as well as setting directives; the Joint Parliamentary Commission in representation of the four parliaments; the Economic and Social Consultation Forum, which has representatives from the different economic and social groups; and finally a weak Administrative and Technical Secretariat, which supports the whole operation from Montevideo. With such a structure, any decision is likely to be highly politicized (Vaillant, 2005).

The absence of strong regional institutions has been particularly felt in the area of macroeconomic coordination. Throughout the 1990s the variability of nominal exchange rates within the region was twice as great as in other comparable countries, leading to the strong backlashes against regional

integration discussed above. This led some regional leaders, including the former Argentine President Carlos Menem, to call for the creation of a monetary union. As argued by Eichengreen (1998), this is the optimal instrument to avoid wide fluctuations in intra-regional exchange rates while keeping some flexibility with respect to bilateral exchange rates with the rest of the world. However, a monetary union will not be an option as long as the other institutions of Mercosur remained politicized and weak. As the experience of the European Union shows, a monetary union not only requires a strong and politically independent central bank but should also be part of an interlocking web of strong economic and political agreements, all of which could be jeopardized if a country abandoned the single currency. The latter acts as a significant barrier to exit for members, reinforcing credibility and stabilizing markets. Mercosur members have all taken significant steps towards central-bank independence. But in terms of barriers to exit there is not much apart from a relatively well-functioning customs union. Very little has been achieved in terms of common trade, economic, social or security policies. If Mercosur does not engage in a deeper integration project, a monetary union cannot be successful.

To conclude, Mercosur is an unprecedented example of successful and enduring regional integration among developing countries. It has proven its resilience by emerging relatively unscathed from acute financial and fiscal crises in the region. However, the most difficult and challenging steps towards the economic integration envisaged in the Treaty of Asunción remain to be taken.

See Also

- currency unions
- foreign direct investment
- regional and preferential trade agreements

Bibliography

Berlinski, J. 2004. *Los impactos de la política comercial: Argentina y Brazil (1988–1997)*. Buenos Aires: Siglo XXI.

Bohara, A., Gawande, K. and Sanguinetti, P. 2004. Trade diversion and declining tariffs: evidence from Mercosur. *Journal of International Economics* 64, 65–88.

Bouzas, R. 2002. MERCOSUR after ten years: learning process or déjà-vu? In *Paths to Regional Integration: The Case of Mercosur*, ed. J. Tulchin and R. Espach. Washington, DC: Woodrow Wilson International Center.

Brambilla, I. 2005. A customs union with multinational firms: the automobile market in Argentina and Brazil. Working Paper No. 11745. Cambridge, MA: NBER.

Cadot, O., de Melo, J. and Olarreaga, M. 2003. The protectionist bias of duty drawbacks: evidence from Mercosur. *Journal of International Economics* 59, 161–82.

Castilho, M. and Zignago, S. 2002. Trade effects of FDI in Mercosur: a disaggregated Analysis. In *An Integrated Approach to the EU-Mercosur Association*, ed. P. Giordano. Paris: Sciences Po.

Chang, W. and Winters, L. 2002. How regional blocs affect excluded countries: the price effects of Mercosur. *American Economic Review* 92, 889–904.

Chudnovski, D. 2001. *El boom de la inversión extranjera directa en el MERCOSUR*. Buenos Aires: Siglo XXI.

Eichengreen, B. 1998. Does Mercosur need a single currency? Working Paper No. 1018, Center for International and Development Economic Research, UC Berkeley.

Olarreaga, M. and Soloaga, I. 1998. Endogenous tariff formation: the case of Mercosur. *World Bank Economic Review* 12(1), 297–320.

Porto, G. 2006. Using survey data to assess the distributional effects of trade policy. *Journal of International Economics* 70(1), 140–60.

Schiff, M. and Chang, W. 2003. Market presence, contestability and the terms-of-trade effects of regional integration. *Journal of International Economics* 60, 161–75.

Vaillant, M. 2005. Mercosur: southern integration under construction. *Internationale Politik und Gesellschaft* 2, 52–71.

Yeats, A. 1998. Does Mercosur's trade performance raise concerns about the effects of regional trade arrangements? *World Bank Economic Review* 12(3), 1–28.

How to cite this article

Olarreaga, Marcelo. "Mercosur." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000369> doi:10.1057/9780230226203.1085

The New Palgrave Dictionary of Economics Online

merger analysis (United States)

Dennis W. Carlton and Jeffrey M. Perloff
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There are three different types of mergers: horizontal, vertical, and conglomerate. We discuss all three and explain why mergers can be a desirable way to expand a firm. Then we turn to the evidence on the amount of merger activity. Finally, we address one of the important questions surrounding mergers: whether they are motivated by the desire to improve efficiency or by the desire to acquire market power. Although the evidence is sometimes ambiguous, the overwhelming consensus is that most merger activity in the United States is motivated by efficiency considerations.

Keywords

antitrust enforcement; conglomerate mergers; economies of scale; economies of scope; efficiency; horizontal mergers; litigation; market power; mergers; reputation; takeovers; taxation of corporate profits; transaction costs; vertical mergers

Article

In the economics literature, a merger is the combination of the assets of two or more firms. Economists usually distinguish three different types of merger: horizontal, vertical, and conglomerate. Horizontal mergers are between rivals; vertical mergers involve firms one of which supplies inputs to the other(s); conglomerate mergers are between firms in unrelated businesses. Mergers represent one way for a firm to acquire assets as an already assembled package.

We first discuss why a merger is sometimes a desirable way to expand a firm. Then we turn to the evidence on the amount of merger activity. Finally, we address one of the important questions surrounding mergers: whether they are motivated by the desire to improve efficiency or by the desire to acquire market power. Although the evidence is sometimes ambiguous, the overwhelming consensus is that most merger activity in the United States is motivated by efficiency considerations.

Reasons for mergers

The most important obvious reason for mergers is to increase efficiency. There is a variety of ways in which mergers can enhance efficiency. By increasing its size, a firm may be able to achieve economies of scale in production, distribution, management, or other aspects of the firm's operation, such as research and development. By eliminating duplication of certain management functions, firms may be able to cut their total costs. Certain scale efficiencies may arise naturally when firms are regulated or have reporting requirements. For example, a merged firm may have to submit tax and other government forms only once as a result of the merger.

By increasing the number of its activities, the merged firm may achieve economies of scope, efficiencies that result from engaging in related activities done together in one firm. For example, the ability of one firm to provide a wide range of products may make distribution easier. Alternatively, the ability to use in one activity knowledge gained in another can make it more efficient to have one firm perform both activities rather than having each activity performed by a different firm.

A common reason for vertical mergers is to eliminate transaction costs associated with using the marketplace to obtain supplies. (Of course, there is the offsetting cost of running a larger firm.) An example of a transaction cost is opportunism. In marketplace transactions, a buyer may (unexpectedly) be able to exploit the seller (or vice versa). For example, the

seller may have no other possible buyer in the short run, and the buyer could demand a lower price than the once originally agreed upon. A vertical merger is an alternative to other mechanisms such as reputations or contract litigation to deal with this problem.

A vertical merger may eliminate the distortion of an upstream (input) monopoly. Prior to merger, the downstream (output) firm decides how to produce and how to price its output based on this distorted input price. If the output is produced with variable input proportions, there is a loss of efficiency to the economy that a vertical merger can fix. There is a private incentive to vertically integrate, but the effect on social welfare is ambiguous.

If both the upstream and the downstream firms are non-competitive, a vertical merger eliminates ‘double marginalization’. An upstream firm with market power raises its price above its marginal cost. Then the downstream firm adds an additional markup so that the final consumers pay a double markup. If the firms merge, they set only a single markup, causing output price to fall, output to expand and social welfare to rise.

So far, these explanations do not answer the question why one firm merges with another rather than buying the underlying assets and assembling them itself. Aside from competitive effects, one answer is that another firm is a package of already assembled assets and it may be cheaper to buy an existing firm than to create one.

Mergers can also be used to transfer assets from the control of bad managers (or investors) to good ones. Suppose that Firm X has very smart managers, while Firm Y has either incompetent managers or managers who are not performing well because no one is monitoring their actions. Here, a transfer of assets to X should allow Y's assets to be more productive. X should be able to pay more for Y's assets than they are worth based on the market's valuation of Y's cash flows under its current incompetent management. This disparity in value creates an incentive for X to purchase Y. To avoid being taken over, Y's managers can improve their performance (that is, the takeover threat disciplines them) or engage in defensive tactics designed to thwart such a takeover in order to save their jobs. If these defensive tactics induce the acquiring firm to raise its price for Y, the tactics can benefit Y's shareholders. There is a large literature on defensive tactics as well as their sometimes ambiguous efficiency consequences. In a hostile takeover, X buys Y despite the desire of Y (or its managers) to remain independent. The use of hostile takeovers in the 1980s coincided with the ability of acquiring firms to obtain financing through junk bonds (bonds below investment grade).

Aside from efficiency motivations, another rationale for a merger is to eliminate competition between the merging firms. The antitrust laws of the United States forbid mergers that result in a lessening of competition with a consequent increase in price. Although antitrust concerns about mergers mainly arise in the context of horizontal mergers, such concerns can also arise with vertical mergers. One concern is that a vertical merger could eliminate a key supplier for a rival firm. Typically, theories of vertical harm are much less certain in their predictions than theories of competitive harm arising from horizontal mergers.

In addition to efficiency and market power explanations, there is a variety of other reasons for mergers. Tax considerations can sometimes make it advantageous for one firm to merge with another. For example, if one firm has a loss and another a profit, a merger can lower their total tax liability. The merged firm may be able to report no profit and therefore owe no corporate tax. Separately, one of the firms (the profitable one) would have to pay a tax. Mergers can also allow managers to engage in empire building, or allow a firm to have an ‘excuse’ (‘I'm no longer in charge’) to renege on certain informal promises made to workers or other firms.

Evidence: merger activity

Mergers come in waves, being common in certain industries at certain times. Because no single data series on merger activity goes back to 1900, we must splice together sometimes inconsistent data sources to study mergers over the 20th century. Figure 1 presents data on the amount of US merger activity relative to the size of the economy back to 1900. By controlling for the economy's size (a larger economy is likely to generate more merger activity), we can compare the intensity of merger activity at different times.

Figure 1

Annual number of mergers and acquisitions per billion US dollars of real GNP (United States, 1895–2003). *Note*: 1982 dollars. *Sources*: Nelson Series, Federal Trade Commission (FTC) ‘Board’ series and Mergerstat. In 2003 the Bureau of Economic Analysis made comprehensive revisions to its National Income and Product Accounts. These revised figures were used to calculate deflators for the FTC series and Mergerstat. The Mergerstat series has broader converage (for example, more industries, lower thresholds for reporting) than the FTC series. Adapted from Golbe and White (1998, Figure 9.7).

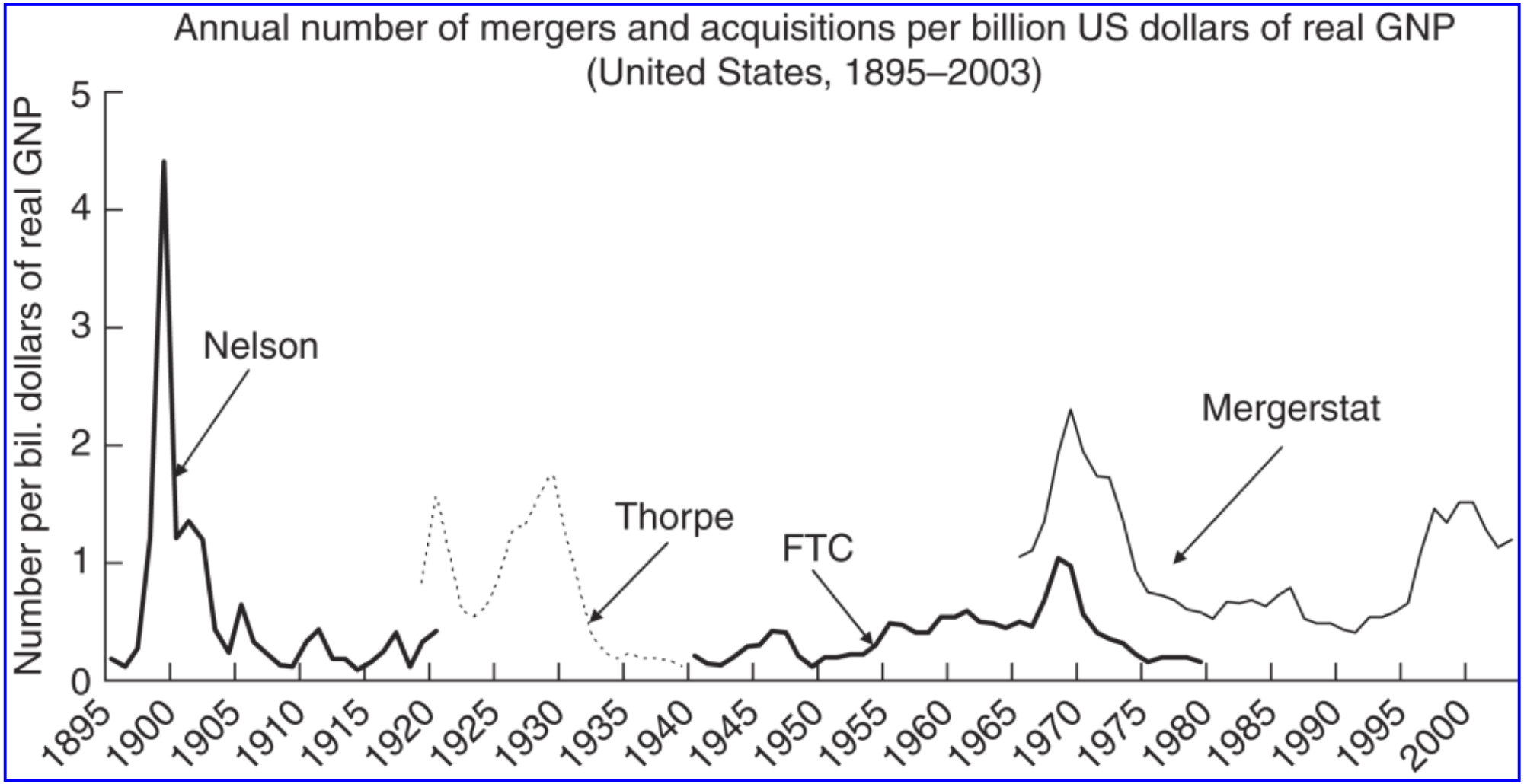

Figure 1 indicates that there have been several waves of merger activity. The first, around 1900, was (relatively) the largest and represents the creation of some of the best-known firms in the United States, such as General Electric and U.S. Steel. This was a time of great change with significant developments in transportation and communications. The second wave was in the 1920s and helped to create several oligopolies. The third, in the 1960s, involved conglomerate mergers. In the 1980s, the fourth wave (which would be more evident in the figure if we had dollar value of merger activity instead of the number of mergers) arose as hostile takeovers became popular in the United States. The fifth was in the late 1990s and disproportionately involved airlines, telecommunications, banking and other industries that had previously been heavily regulated.

The timing of merger waves seems to coincide with stock market booms for reasons no one has completely explained. One recent explanation by Shleifer and Vishny (2003) maintains that during stock booms stocks are overvalued (a fact known by managers but not outside investors) and managers use the (overvalued) stock to purchase other firms. In stock market booms, the use of stock rather than cash to buy other firms' assets does increase, consistent with the hypothesis.

Empirical evidence: rationales for mergers

A central question is whether mergers improve efficiency or, especially in the case of horizontal mergers, reduce competition and harm consumers. Despite an enormous number of studies of this question, the answer is still somewhat controversial. Our conclusion is that, although there is no doubt that some mergers are poorly motivated and turn out badly for the firms, and that some horizontal mergers reduce competition and harm consumers, most are expected to be profitable, to enhance efficiency and not to reduce competition.

Researchers have used three types of data: stock market data, accounting data, and price or output data. Because of their availability, stock market data have been used most often.

Stock market studies rely on the premise that stock market prices are a good indication of a firm's expected future profitability (and make subtle assumptions about when information gets reflected in prices). These stock market studies can capture the effect of a merger on the acquiring firm, the acquired (target) firm, and rivals. Accounting data may have certain biases that can be hard to correct, and can be difficult to obtain. The same is true of data on price and output. In contrast to stock market studies, studies using either accounting or performance data are *ex post* studies of mergers (what happened after the mergers), while studies using stock market data are *ex ante* studies (what is expected to happen). We present a brief summary of the major findings (see Carlton and Perloff, 2005, and the references cited there, especially Andrade and Stafford, 2001, and Pautler, 2003).

Shareholders of an acquired firm earn a premium of between 16 and 25 per cent above the price prevailing prior to the merger. This premium is now higher than it used to be before the Williams Act of 1968 was passed. The Williams Act requires a firm to reveal publicly its intentions to acquire another firm.

Shareholders of the acquiring firm do not do very well. Although they earned slightly positive returns in the 1960s (plus four per cent), their returns became slightly negative (minus three per cent) in the 1980s and 1990s. Interestingly, the form of the acquisition (whether cash or stock) influences the return, with acquirers doing better when more cash is used, though it is unclear why this should occur. (The use of stock to finance mergers has increased over time, with about 60 per cent of transactions in the 1990s financed entirely by stock.)

Overall, the total return (which is what matters for efficiency) to the combined acquiring and acquired firms is positive. That is, the total value of the merged firm is about 2–7.5 per cent higher after a merger than the sum of each firm's value pre-merger.

Researchers using accounting or other performance data have had more difficulty documenting gains from mergers. Using data from the 1960s and 1970s, Scherer (1988) and Ravenscreft and Scherer (1987) do not find increased profits post-merger. Andrade and Stafford (2001) use Scherer's data and show that the data support the efficiency hypothesis if one controls for industry benchmarks. Lichtenberg and Siegal (1987) find significant positive effects of mergers on productivity.

Studies of stock markets and of individual industries have been used to investigate whether horizontal mergers generally create market power. The stock market studies exploit the idea that a merger that creates efficiency will cause the stock price of the (to-be-merged) firm to rise and that of its rivals to fall. In contrast, a horizontal merger that eliminates a rival should be expected to also benefit other rivals (since industry price will rise if competition is eliminated). Banerjee and Eckard (1998) show that, even for the massive merger wave around 1900 (prior to strict enforcement of antitrust laws forbidding mergers that eliminated competition), rival firms suffered as a result of a horizontal merger, supporting the efficiency hypothesis. There are of course exceptions, and some studies of recent mergers (for example, some airline mergers) show that horizontal mergers can harm competition and raise price. However, most of the literature (though certainly not all) supports the view that mergers generally should be expected to help consumers.

See Also

- firm boundaries (empirical studies)
- mergers, endogenous
- merger simulations

This article draws heavily on Carlton and Perloff (2005, ch. 2). The reader interested in more detailed discussion should consult that work together with the references cited therein.

Bibliography

Andrade, G. and Stafford, E. 2001. New evidence and perspectives on mergers. *Journal of Economic Perspectives* 15(2), 103–20.

Banerjee, A. and Eckard, E. 1998. Are mega-mergers anti-competitive? Evidence from the first great merger wave. *RAND Journal of Economics* 29, 803–27.

Carlton, D. and Perloff, J., 2005. *Modern Industrial Organization*. New York: Pearson Addison Wesley.

Golbe, D. and White, L. 1988. A time series analysis of mergers and acquisitions in the US Economy. In *Corporate Takeovers: Causes and Consequences*, ed. A. Auerbach. Chicago: University of Chicago Press.

Lichtenberg, F. and Siegal, D. 1987. Productivity and changes in ownership of manufacturing plants. *Brookings Papers on Economic Activity* 1987(3), 63–83.

Pautler, P. 2003. Evidence on mergers and acquisitions. *Antitrust Bulletin* 48, 119–221.

Ravenscraft, D. and Scherer, F. 1987. *Mergers, Sell-offs and Economic Efficiency*. Washington, DC: Brookings Institution.

Scherer, F. 1988. Corporate takeovers: the efficiency arguments. *Journal of Economic Perspectives* 2(1), 69–82.

Shleifer, A. and Vishny, R. 2003. Stock market driven acquisitions. *Journal of Financial Economics* 70, 295–311.

How to cite this article

Carlton, Dennis W. and Jeffrey M. Perloff. "merger analysis (United States)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000148> doi:10.1057/9780230226203.1086

The New Palgrave Dictionary of Economics Online

merger simulations

Aviv Nevo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

airline industry; merger simulations; mergers

Article

The key in an evaluation of a proposed merger is to determine whether the reduction of competition it would cause is outweighed by potential cost reductions. Traditional analysis of mergers is primarily based on industry-concentration measures. A market is defined and market shares of the relevant firms are used to compute a pre-merger concentration measure as well as a change in this measure due to the merger. Both the pre-merger level and the change in concentration are then compared with preset levels. The intuition is that, if the industry is concentrated, or if the change in concentration is large, then the anti-competitive effect will dominate. Using this approach to evaluate mergers in some industries is problematic for at least two reasons. In many cases the product offerings make the definition of the relevant product (or geographic) market difficult. Even if the relevant market can be defined, the computed concentration index provides a reasonable standard by which to judge the competitive effects of the merger only under strong assumptions.

Merger simulation attempts to deal with these challenges. The basic idea consists of ‘front-end’ estimation, in which the structural primitives of the model are estimated, and a ‘back-end’ analysis, in which the estimates are used to simulate the post-merger equilibrium. The approach proceeds as follows. First, demand parameters are recovered by econometric estimation, if the data are rich enough, or, if data (with enough variation) are not available, then marketing and other anecdotal evidence can be used to approximate the effects of prices on demand (Werden and Froeb, 1994). Estimation has to deal with two main challenges: a flexible functional form, especially with a large number of products, and reasonable identifying assumptions. The most commonly used approaches, to deal with the large number of products, are multi-level budgeting (Hausman, Leonard and Zona, 1994) and the discrete-choice, characteristics, approach (Berry, Levinsohn and Pakes, 1995; Nevo, 2000). Prices are set endogenously and typically respond to demand shocks that are unobserved by the researcher, and therefore instrumental variables are needed. Two common instrumental variables are observed characteristics of

other products (Bresnahan, 1987; Berry, Levinsohn and Pakes, 1995) and out-of-market prices (Hausman, Leonard and Zona, 1994; Nevo 2000).

Second, pre-merger cost parameters are recovered. One approach is to assume a model of pricing (Bertrand, say) and to use it jointly with the estimated demand parameters to recover implied marginal costs. If needed, the implied marginal costs can be regressed on characteristics in order to recover cost functions. Alternatively, the pricing equation, and the cost functions, can be estimated jointly with demand. Either way, the model of pricing can, and should, be tested (Porter, 1983; Bresnahan, 1987; Nevo, 2001). Finally, marginal cost can be approximated from accounting data, but these tend to be unreliable.

Third, the recovered marginal costs and estimated demand parameters are used jointly to simulate the new equilibria that would result from a merger. Usually, the analysis focuses on ‘unilateral effects’, with the likelihood of (tacit) collusion fixed. In principle, however, the simulation can use a different model of competition post-merger from the one used to recover the parameters. In order to address potential cost reductions, the simulation can be performed with marginal cost fixed, by changing marginal costs or by asking what cost saving is required to keep consumer welfare, or any other measure, at a certain level (Nevo, 2000). Finally, the model can be used to assess the likelihood of entry and/or the change in incentive to collude.

The end result is a prediction of post-merger prices and quantities under several scenarios. With the use of the estimated demand and supply functions, these equilibrium quantities can be converted into consumer welfare and (variable) profits. The change in welfare and profits can be used as the basis for evaluating the merger instead of the change in concentration. This has the advantage of being linked to economic theory and the underlying trade-off between reduction in competition and improved efficiency. It also allows the parties to assess the accuracy of the prediction due to the assumptions by simulating under different assumptions, or due to the data by computing standard errors.

There are several potential pitfalls in using merger simulation. The simulation is only as good as the model it is based on and the parameter estimates that go into the simulation. Therefore, one should take extra care in choosing a model suitable for the industry. Furthermore, in some cases data and time constraints might limit the ability to consistently estimate the parameters required for the simulation. Despite the fact that merger simulation has been used extensively in practice, there is little work testing its accuracy with the use of post-merger data. One exception is a study of mergers in the airline industry (Peters, 2003) that finds that simulation methods do a reasonable job at predicting the price effects of mergers. Peters also finds that a large fraction of the unexplained change in prices comes from changes in marginal costs or firm conduct (his analysis cannot separate the two). Retrospective analysis of this sort is useful not just in evaluating the quality of predictions but also in pointing to directions in which the modelling and analysis can be improved.

For further readings and details see Whinston (2005, ch. 3).

Bibliography

Berry, S., Levinsohn, J. and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–90.

Bresnahan, T. 1987. Competition and collusion in the American automobile oligopoly: the 1955 price war. *Journal of Industrial Economics* 35, 457–82.

Hausman, J., Leonard, G. and Zona, J. 1994. Competitive analysis with differentiated products. *Annales d'Economie et de Statistique* 34, 159–80.

Nevo, A. 2000. Mergers with differentiated products: the case of the ready-to-eat cereal industry. *RAND Journal of Economics* 31, 395–421. Reprinted in *Empirical Industrial Organization*, ed. P. Joskow and M. Waterson. Cheltenham: Edward Elgar, 2004.

Nevo, A. 2001. Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69, 307–42.

Peters, C. 2003. Evaluating the performance of merger simulation: evidence from the US airline industry. Working Paper No. 32. Center for the Study of Industrial Organization, Northwestern University.

Porter, R. 1983. A study of cartel stability: The Joint Executive Committee, 1880–1886. *Bell Journal of Economics* 14, 301–14.

Werden, G. and Froeb, L. 1994. The effects of mergers in differentiated products industries: logit demand and merger policy. *Journal of Law, Economics and Organization* 10, 407–26.

Whinston, M. 2005. *Lectures on Antitrust Economics*. Cambridge, MA: MIT Press.

How to cite this article

Nevo, Aviv. "merger simulations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000357> doi:10.1057/9780230226203.1087

The New Palgrave Dictionary of Economics Online

mergers, endogenous

Volker Nocke
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bidding games; antitrust; coalition formation games; collusion; concentration; Cournot model; endogenous mergers; exogenous mergers; horizontal mergers; market power; mergers; monopolization; oligopoly; vertical mergers

Article

The term ‘endogenous mergers’ reflects the view in economic theory that mergers are equilibrium outcomes. The literature on endogenous mergers explicitly analyses firms’ incentives to merge and makes predictions on the volume and type of mergers that are likely to occur. In this literature, merger formation is modelled as a bidding game or non-cooperative coalition formation game (Kamien and Zang, 1990; Gowrisankaran, 1999; Nocke, 2000; Pesendorfer, 2005), or as an anonymous merger market where firms can buy or sell corporate assets (Jovanovic and Rousseau, 2002; Nocke and Yeaple, 2007). The literature on endogenous mergers is conceptually distinct from the literature on *exogenous* mergers, which considers the positive and normative effects of a merger between a given (‘exogenous’) set of firms.

To analyse the endogenous merger process, one first needs to understand why firms may want to merge. Several motives for mergers have been identified in the literature.

First, firms may want to merge to realize efficiency gains or ‘synergies’. Mergers may allow firms to exploit complementarities in their capabilities (Nocke and Yeaple, 2007), or they may be an efficient way to reallocate used capital from less productive firms to more productive firms (Jovanovic and Rousseau, 2002).

Second, firms may want to merge to increase their market power. However, as Salant, Switzer and Reynolds (1983) have shown for the Cournot model, a merger solely aimed at increasing market power may not be profitable: to the extent that merging firms want to reduce joint output to raise price, non-participating outsiders will increase their output in response, imposing a negative externality on the merging firms. (This point relies heavily on the Cournot assumption; see Deneckere and Davidson,

1985.) While it has generally been acknowledged that horizontal mergers (between firms competing in the same market) may lead to higher prices and lower welfare, the Chicago School of antitrust has long held the view that vertical mergers (between upstream suppliers and their downstream customers) are efficiency-enhancing. By showing that vertical mergers may allow foreclosure of upstream suppliers or downstream buyers, this view has recently been refuted in a series of articles (see Rey and Tirole, 2005, for a survey).

Third, firms may want to merge to facilitate collusion. A horizontal merger may facilitate collusion by reducing the number of players in the industry, or by reallocating industry capacity in a way that equalizes firms' incentives to cheat (Compte, Jenny and Rey, 2002). A vertical merger may facilitate upstream collusion by reducing the number of downstream outlets through which an upstream firm can profitably deviate. Furthermore, to the extent that collusion is sustainable only if the vertically integrated firm receives a larger market share than an unintegrated firm, firms may have an incentive to merge so as to demand and obtain a larger share of the collusive pie (Nocke and White, 2003).

Finally, a variety of other motives for merger have been proposed, some of which are based on the view that firms do not necessarily maximize profits. For example, it has been argued that managers may have an incentive to engage in empire building.

Focusing on the market power motive, much of the recent literature on endogenous mergers has been concerned with studying the limits to monopolization through mergers and acquisitions, and making predictions on the relationship between concentration levels and industry characteristics (Kamien and Zang, 1990; Nocke, 2000; Gowrisankaran and Holmes, 2004). The starting point of this literature is the observation by Stigler (1950, pp. 25–6) that 'the promoter of a merger is likely to receive much encouragement from each firm – almost every encouragement, in fact, except participation'.

To understand Stigler's point that a merger to monopoly may not obtain even when feasible, consider an industry with N firms, each running a single plant to produce a homogeneous or differentiated good. If a subset of these firms merge, they will internalize any externality in the price/output decisions they impose on each other. Unless efficiency gains from merging are large, a merged entity would thus produce a smaller output per plant than a single-plant firm: a firm participating in the merger ('insider') would be better off than a firm not participating ('outsider'). Let $\Pi(N; 0)$ denote monopoly profits, and $\Pi(1; N-1)$ the profit of a single-plant firm competing with a larger firm owning $N-1$ plants. Assume the merger would take place even when only $N-1$ firms agreed to merge. Then, firm i will agree to merge with its $N-1$ rivals only if $\Pi(1; N-1) \leq s_i \Pi(N; 0)$, where s_i is firm i's equity share in the merged entity. Since this must hold for any firm i, merger to monopoly will occur only if $\Pi(1; N-1) \leq \Pi(N; 0) / N$. In standard oligopoly models, this inequality is often violated if efficiency gains from merging are small, the number of firms is large, and competition is not too 'tough'. Merger to monopoly may thus fail to occur, even though it would maximize joint profits, as some firm(s) may be better off staying outside and taking a free ride on the merged entity's effort to restrict output.

There may also be limits to monopolization through mergers and acquisitions because of entry. To the extent that a merger makes the industry *less* competitive, a merger between incumbents may induce more entry in the future, reducing the incumbents' profits. By not merging with their rivals, incumbent firms may thus credibly commit to compete vigorously and deter further entry.

See Also

- cartels
- merger analysis (United States)
- merger simulations

Bibliography

Compte, O., Jenny, F. and Rey, P. 2002. Capacity constraints, mergers and collusion. *European Economic Review* 46, 1–29.

Deneckere, R. and Davidson, C. 1985. Incentives to form coalitions with Bertrand competition. *RAND Journal of Economics* 16, 473–86.

Gowrisankaran, G. 1999. A dynamic model of endogenous horizontal mergers. *RAND Journal of Economics* 30, 56–83.

Gowrisankaran, G. and Holmes, T.J. 2004. Mergers and the evolution of industry concentration: results from the dominant firm model. *RAND Journal of Economics* 35, 561–82.

Jovanovic, B. and Rousseau, P.L. 2002. The Q-theory of mergers. *American Economic, Review Papers and Proceedings* 92, 198–204.

Kamien, M.I. and Zang, I. 1990. The limits of monopolization through acquisition. *Quarterly Journal of Economics* 105, 465–99.

Nocke, V. 2000. Monopolisation and industry structure. Economics Working Paper No. 2000-W27, Nuffield College, Oxford.

Nocke, V. and White, L. 2003. Do vertical mergers facilitate upstream collusion? Working Paper No. 03-033, PIER, University of Pennsylvania.

Nocke, V. and Yeaple, S. 2007. Cross-border mergers and acquisitions versus greenfield foreign direct investment: the role of firm heterogeneity. *Journal of International Economics* 72(2), 336–65.

Pesendorfer, M. 2005. Mergers under entry. *RAND Journal of Economics* 36, 661–79.

Rey, P. and Tirole, J. 2005. A primer on foreclosure. In *Handbook of Industrial Organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: North-Holland.

Salant, S.W., Switzer, S. and Reynolds, R.J. 1983. Losses from horizontal merger: the effects of an exogenous change in industry structure on Cournot–Nash equilibrium. *Quarterly Journal of Economics*

98, 185–99.

Stigler, G.J. 1950. Monopoly and oligopoly by merger. *American Economic Review, Papers and Proceedings* 40, 23–34.

How to cite this article

Nocke, Volker. "mergers, endogenous." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000362> doi:10.1057/9780230226203.1088

The New Palgrave Dictionary of Economics Online

merit goods

R.A. Musgrave
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘merit goods’ has no generally agreed application. It is best applied where individual choice is restrained by community values. It may apply also where charity or political redistribution imposes the donors’ preferences on recipients; in primary redistribution, society may define fair shares in cash or kind, the latter chosen with regard to what are considered meritorious items for the recipient. However, the concept of merit goods remains within the realm of consumer sovereignty when individuals’ ‘higher’ preferences are imposed on their ‘lower’ ones.

Keywords

Aristotle; asymmetric information; charity; commitment; community preferences; consumer sovereignty; fair shares; fiscal theory; Harsanyi, J. C.; individual choice; Kant, I.; majority rule; merit goods; Musgrave, R. A.; myopia; paternalism; Pigou, A. C.; preferences; primary goods; private vs. public goods; subjective vs. ethical preferences; Rawls, J.; redistribution; Smith, A.; Vickrey, W. S.

Article

The concept of merit goods, since its introduction thirty years ago (Musgrave, 1957, 1959), has been widely discussed and given divergent interpretations (for surveys, see Head, 1966; Andel, 1984). Since no patent attaches to the term, it is thus difficult to provide a unique definition. However, most interpretations relate to situations where evaluation of a good (its merit or demerit) derives not simply from the norm of consumer sovereignty but involves an alternative norm. In the following, various situations and their bearing on the concept will be considered.

1 Merit goods, private goods and public goods

While the concept of merit goods was raised in the context of fiscal theory, the term has broader

application and should not be confused with that of public (Musgrave, 1957, 1959) goods. The distinction between private and public or social goods arises from the mode in which benefits become available, i.e., rival in the one and non-rival in the other case (see Public Goods). As a result, conditions of Pareto optimality differ, as do the appropriate mechanisms of choice. But whether met through a market or political process, both choices and the normative evaluation of outcomes squarely rest on the premise of individual preference. Consumer sovereignty is taken to apply to both cases. The concept of merit (or, for that matter, of demerit) goods questions that premise. It thus cuts across the traditional distinction between private and public goods. A more fundamental set of issues is raised, issues which do not readily fit into the conventional framework of micro theory as based on a clearly designed concept of free consumer choice.

2 Pathological cases

Next, we consider various settings where the norm of consumer sovereignty remains the preferred solution, but where difficulties in implementation have to be met. The most extreme case arises with regard to the mentally deficient or children. In both cases, some guidance is needed and custodial choices have to be made. These, however, may be viewed as exceptional circumstances and not part of the essential merit good problem. It is also evident that rational choice requires correct information, and that the quality of choice is impeded where information is imperfect or misleading. Situations may arise, as in the design of educational programmes, where the quality of choice as eventually valued by the beneficiary's own preference is improved by initial delegation of choice to others whose prior information is superior. Once more, the implementation of individual preferences is affected, but without questioning their dominance at the normative level.

Other instances arise where rational choice is impeded by oversight or myopia. Individuals, though informed and generally competent to choose, may be inclined to depart from rational choice on certain issues. Thus future consumption tends to be undervalued relative to present consumption (Pigou, 1928), while public services may be overvalued because they seem free or undervalued due to dislike of taxation. Rational choice may be impeded in the context of risk-taking, and so forth. Certain goods may thus come to be under or oversupplied for such reasons of misjudgment and their promotion or restriction may be called for. Such situations again pose some departure from the premise of rational choice, but they deal with defects in the implementation of consumer sovereignty, rather than its rejection as a norm.

3 Rule of fashion

By assuming individuals to have a well-defined preference structure which may then be interfered with, it is tempting to bypass the fact that individual preferences are not fixed in isolation but are affected by the societal setting in which individuals operate. Taking an extreme view of this dependence (Galbraith, 1958), the existence of independent preferences may be denied. Individual preferences become mirror images of fashions in what society approves or holds desirable. But this is too extreme a position. While societal influences enter, they are nevertheless met by individual responses, leaving effective preferences to differ across individuals. Though the preferences of individuals are conditioned by their

social environment, own-preferences enter in shaping the individual's responses thereto. It thus seems inappropriate to equate the concept of merit goods with that of fashion.

4 Community preferences

As distinct from the rule of fashion, consider a setting where individuals, as members of the community, accept certain community values or preferences, even though their personal preferences might differ. Concern for maintenance of historical sites, respect for national holidays, regard for environment or for learning and the arts are cases in point. Such acceptance in turn may affect one's choice of private goods or lead to budgetary support of public goods even though own preferences speak otherwise. By the same token, society may come to reject or penalize certain activities or products which are regarded as demerit goods. Restriction of drug use or of prostitution as offences to human dignity (quite apart from potentially costly externalities) may be seen to fit this pattern. Community values are thus taken to give rise to merit or demerit goods. The hard-bitten reader regards this as merely another instance of fashion which may be disposed of accordingly. But such is not the case. Without resorting to the notion of an ‘organic community’, common values may be taken to reflect the outcome of a historical process of interaction among individuals, leading to the formation of common values or preferences which are transmitted thereafter (Colm, 1965). As this author sees it, this is the setting in which the concept of merit or demerit goods is most clearly appropriate, and where consumer sovereignty is replaced by an alternative norm.

5 Paternalism in distribution

In viewing the problem of individual choice and preferences, we so far have assumed that the individual's endowment from which to choose is given. It remains to consider a set of problems which arise in the context of distribution.

We begin with the case of voluntary giving (Hochman and Rogers, 1969). Donor D may derive utility from giving to recipient R, but more so if the grant is specified in kind (e.g., milk) than given in cash (and used for beer). Such paternalistic giving interferes with R's preferences. While R cannot be damaged (the grant can be refused) his or her gain is less than it would be from a cash grant. Charity by way of paternalistic giving thus involves imposition of D's preferences, of what goods *he* considers of merit for R. At the same time, giving in kind is in line with consumer sovereignty at the donor level, as D's satisfaction depends on what R consumes. Moreover, R cannot suffer a loss, since the grant may be rejected.

A similar problem arises in the context of redistribution through the political process of majority rule. Here, taking as well as giving is involved. While the Rs would prefer to take cash, they may do better by setting for in-kind programmes which appeal to the Ds. Redistribution by majority vote may thus take in-kind form. Once more the Ds may impose their preferences on the Rs, but subject to the terms of the social contract which now permits such intervention via majority rule. Many budget programmes rendering services to the poor (such as health, welfare, and low-cost housing) are of this type, and have indeed come to be classified as merit goods (OECD, 1985).

Having considered merit goods in relation to redistribution, it remains to note their bearing on the more

basic issue of *primary* distribution. Models of distributive justice have taken a variety of forms, including entitlement to earnings in the Lockean tradition, utilitarian criteria, and entitlement to ‘fair shares’ (Vickrey, 1960; Harsanyi, 1955; Rawls, 1971). The latter may be viewed in terms of fair shares in income and wealth, while leaving its use to individual choice; or, it may be viewed in terms of a fair share in particular goods or bundles thereof. The role of merit goods arises in the latter context, and indeed bears some relation to the philosopher's concept of ‘primary goods’. Moreover, both approaches may be combined in various ways. Thus, society may view it as fair to modify the distribution of income via a tax-transfer scheme, while also arranging the distribution of certain goods (e.g., scarce medical treatment) outside the market rule (Tobin, 1970), or society may wish to assure an adequate minimum provision, but do so by providing for a bundle of necessities rather than an equivalent minimum income to be spent at the recipient's choice. Goods separated out for non-market distribution might then be viewed as merit goods.

6 Multiple preferences or ‘ higher values ’

The reader will note that up to this point we have dealt with settings which, in one way or another, involve some form of departure from the rule of consumer sovereignty. It remains to consider a further perspective, which views the problem within the sovereignty context. This approach postulates that preferences may derive from conflicting sets. This has been noted over the ages, from Artstotle's concept of *‘atrasia,’* over the Kantian imperative and Faust's ‘two souls’ to Adam Smith's impartial observer (Smith, 1759). Later the same thought appears in Harsanyi's distinction between subjective and ethical preferences (Harsanyi, 1955). A recent illustration follows in Rawls's concept of disinterested choice (Rawls, 1971) and Sen's usage of commitment (Sen, 1977). The term merit goods has then been applied to goods chosen under the latter (‘ethically superior’) set of preferences. Such choice may involve private as well as public goods, although they may be more likely to enter in the latter context where they may prove less costly due to the sharing of tax burdens (Brennan and Lomasky, 1983).

Conclusion

As the preceding discussion shows, the term merit goods has been applied to a variety of situations. In (1) we have noted that the merit good concept should not be confused with that of public goods. In section (2) we noted that a variety of situations may arise where interference with individual choice is needed but without questioning its validity as the basic norm. In (3) we have granted that individual preferences are influenced by social environment, but not to the point of excluding individual-preference based responses. None of these cases offered an appropriate setting in which to apply the merit or demerit concept. The case considered in section (4), offering community values as a restraint on individual choice, did, however, fit the pattern and, as I see it, goes to the heart of the merit concept. Section (5) posed related issues in the context of distribution. Voluntary giving was shown to permit the donor to impose his or her preferences on the donee, and this remains the case, if with lesser force, for political redistribution. Redistribution will tend to be in goods which the donor consider meritorious for the donee. Turning to primary distribution, we noted that society may define fair shares in cash or kind, the latter chosen with regard to what are considered meritorious items for the recipient. Only in section

(6) did use of the merit goods concept remain within the context of the sovereignty norm, dealing now with preferences (merit or demerit wants) of a higher or lower kind. In all, it seems difficult to assign a unique meaning to the term. This writer's preference, as noted before, would reserve its use for the setting dealt with under (4), but that of (5) and (6) may also have a claim.

See Also

- public finance
- public goods

Bibliography

Andel, N. 1984. Zum Konzept der meritorischen Güter. *Finanz Archiv*, New Series 42(3), where extensive literature references are given.

Brennan, G. and Lomasky, L. 1983. Institutional aspects of merit goods analysis. *Finanz Archiv*, New Series 41, 183–206.

Colm, G. 1965. National goals analysis and marginal utility economics. *Finanz Archiv*, New Series 24, 209–24.

Galbraith, K. 1958. *The Affluent Society*. Boston: Houghton Mifflin.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Head, J.C. 1966. On merit goods. *Finanz Archiv*, New Series 25(1), 1–29.

Hochman, H.H. and Rogers, J.D. 1969. Pareto-optimal redistribution. *American Economic Review* 59, 542–57.

Musgrave, R.A. 1957. A multiple theory of budget determination. *Finanz Archiv*, New Series 17, 333–43.

Musgrave, R.A. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

OECD. 1985. *The Role of the Public Sector*. Paris: OECD.

Pigou, A.C. 1928. *A Study in Public Finance*. London: Macmillan.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Sen, A. 1977. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6, 317–44.

Smith, A. 1749. *The Theory of Moral Sentiments*. Reprinted, New York: Liberty, 1969.

Tobin, J. 1970. On limiting the domain of inequality. *Journal of Law and Economics* 13, 263–77.

Veblen, T. 1899. *The Theory of the Leisure Class*. New York: New American Library.

Vickrey, W. 1960. Utility, strategy, and social decision rules. *Quarterly Journal of Economics* 74, 507–35.

How to cite this article

Musgrave, R. A. "merit goods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000152> doi:10.1057/9780230226203.1089

The New Palgrave Dictionary of Economics Online

Merton, Robert C. (born 1944)

Darrell Duffie
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robert C. Merton, who developed the theory of option pricing with Myron Scholes and the Fischer Black, is responsible for a new approach to investments and asset pricing, based in part on stochastic calculus. Awarded the Nobel prize in 1997, Merton's other contributions to financial economics include the intertemporal capital asset pricing model (ICAPM). He has written extensively on pension planning, social security, and bank deposit insurance.

Keywords

American Finance Association; arbitrage; Black, F.; Black–Scholes option pricing model; derivative securities; intertemporal capital asset pricing model; Long-Term Capital Management (LTCM); Markowitz, H.; mean-variance investment theory; Merton, R.; Miller, M.; Modigliani, F.; rational option pricing theory; Scholes, M.; Sharpe, W.; stochastic calculus

Article

Robert C. Merton, awarded the 1997 Nobel Memorial Prize in Economics, was born in New York City on 31 July 1944. His father, Robert K. Merton, was a noted sociologist, to say the least. This biographical sketch of Robert C. Merton and his contributions to financial economics may seem brief, given the gigantic impact that he had on economics and financial-market practice.

Merton's university education veered from applied mathematics at Columbia University (BS, 1966) and the California Institute of Technology (MS, 1967) to economics at MIT (Ph.D., 1970), where he quickly joined Paul Samuelson as student, then research assistant, faculty colleague, and collaborator. Their paper on warrant pricing (1969a) hinted at Merton's later massive contributions to ‘the option-pricing formula’ and to dynamic investment theory, which followed almost immediately. Within a few years of his arrival at MIT in 1967, it is no exaggeration to say that Merton had transformed his newly chosen field of financial economics and, more broadly, dynamic modelling in economics.

Only a decade before Merton framed his revolutionary new approach to financial modelling, Modigliani and Miller (1958) had used arbitrage reasoning to discover the irrelevance of corporate capital structure and dividend policy in perfect capital markets. About five years before Merton came on to the scene, William Sharpe (1964) had adapted Markowitz's mean-variance investment theory to establish the relationship between risk and expected return in market equilibrium. These pre-Merton breakthroughs were based on static reasoning. Merton exploited stochastic calculus – a completely new approach to dynamic modelling under uncertainty – in order to extend these insights and to open entirely new paths of discovery. The crucial tool of stochastic calculus that Merton brought into financial modelling is the formula of Kiyoshi Itô (1951), whereby, under suitable technical regularity, the rate of change of the conditional expectation of $f(X(t), t)$, for an Itô process X and a smooth function $f(\cdot\ \cdot)$, is

$$f_t(X(t), t) + f_x(X(t), t)m(t) + \frac{1}{2} f_{xx}(X(t), t)v(t), \tag{1}$$

where $m(t)$ is the rate of change of the conditional expectation of $X(t)$, and $v(t)$ is the rate of change of the conditional variance of $X(t)$. (Subscripts indicate partial derivatives.)
Consider, for example, Merton's approach (1969b; 1971) to investment, in which $X(t)$ is the wealth of a risk-averse investor whose current optimal conditional expected utility for final wealth is $f(X(t), t)$. (The conjectured dependence of indirect utility on wealth and time only is tantamount to the independence over time of asset returns, which Merton relaxed in 1973b.) The current portfolio p of investments determines the ‘local mean’ $m(t, p)$ and ‘local variance’ $v(t, p)$ of changes in wealth. At anything other than an optimal portfolio strategy, Bellman's principle of optimality implies that the conditional expected change of $f(X(t), t)$ is negative, so (1) suggests that

$$\begin{aligned} &\mathrm{Max}_p\, f_t(X(t), t) + f_x(X(t), t) \\ &m(t, p) + \frac{1}{2} f_{xx}(X(t), t)v(t, p) = 0\,. \end{aligned} \tag{2}$$

Because the mean $m(t, p)$ and variance $v(t, p)$ of the ‘local return’ on wealth are linear and quadratic, respectively, with respect to the portfolio choice p, the first-order optimality conditions for (2) provide an explicit solution for p in terms of the derivatives of $f(\cdot\ \cdot)$. Substitution of this solution for p into the same equation (2) leaves a partial differential equation to solve for $f(x, t)$. Merton was able to give explicit solutions in certain cases. For example, with expected power utility for final wealth, the indirect utility must inherit the same degree of homogeneity with respect to wealth. Merton's problem is still the classic textbook example of stochastic control to which graduate students in finance and other fields,

even beyond economics, are first exposed. The associated insights into lifetime investment planning are striking, and have led to an immense literature of extensions.

Although this is no place to derive it, the Black–Scholes formula $f(x, t)$ for the price at time *t* of an option on an asset whose current market value is *x* is similarly obtained by the ‘risk-neutral’ valuation equation

$$f_t(x, t) + f_x(x, t)rx + \frac{1}{2} f_{xx}(x, t)\sigma^2 x^2 = rf(x, t), \tag{3}$$

where *r* is the continuously compounding risk-free borrowing rate and σ is the volatility (the standard deviation of annualized continuously compounding returns) of the underlying asset. The boundary condition for (3), in the case of a call option with an exercise date *T* and exercise price *K*, is $f(x, T) = max(x - K, 0)$, because this is the market value of the right, but not the obligation, to buy the stock for *K* when it trades in the market for *x*. Black and Scholes (1973) solved this equation with the famous formula named for them. For the market value of a general contingent claim paying $g(X(T))$ at *T*, the same differential equation (3) applies under technical conditions, with the boundary condition $f(x, T) = g(x)$.

By virtue of Itô's formula (1), one can view (3) as a statement that the option's expected rate of return may be treated as the risk-free rate of return, provided that we replace the actual mean rate of return on the underlying asset with the risk-free rate. Indeed, this is roughly how Black and Scholes (1973) interpreted their original derivation of (3), which was based on a particular general-equilibrium model. Merton, however, noted that changes in the market value of the option over time could actually be replicated by trading the underlying asset, financing any cash needs with risk-free borrowing. This ‘arbitrage’ strategy leads to (3) without reference to a particular general-equilibrium model, since arbitrage is ruled out in any equilibrium. Black and Scholes acknowledged Merton for this alternative approach, which was the genesis of both an enormous academic literature on contingent claims pricing and the professional practice of ‘financial engineering’, a field that includes a vast array of financial pricing and risk-management methods.

Among the most influential applications that Merton developed on the basis of his approach to derivative asset pricing was his insight (1974) that the equity and debt of a corporation may be viewed as derivative securities written on the assets of the firm, and priced accordingly. This idea was developed independently in Black and Scholes (1973). In any case, this widely known ‘Merton model of corporate debt’ is the basis of much modern fundamental market analysis of corporate debt and credit derivatives, in practice and academic research, including both pricing and default prediction.

Rounding out the series of major results that Merton produced within a stunningly short period of time were his intertemporal capital asset pricing model (ICAPM) (1973b) and his theory of rational option pricing (1973a). Merton's ICAPM extended Sharpe's CAPM to a dynamic framework, relieving it of its dependence on mean-variance utility because, from (2), we see that only the (‘instantaneous’) mean and variance of returns matter for conditional mean rates of change of utility, under technical conditions.

More importantly, the ICAPM showed how the expected returns of assets in a multi-period setting compensate not only for exposure to the risk associated with the return of the market portfolio but also for exposure to the risks associated with changes in state variables determining future conditional distributions of asset returns. These latter risks introduce hedging motives not present in a static model.
Merton's Theory of Rational Option Pricing (1973a) shored up the foundations of the Black–Scholes option pricing model and treated a variety of related issues, in particular a rational approach to exercising and pricing American options. A few years later Merton (1977) provided deeper foundations for the basic arbitrage reasoning underlying the pricing of derivatives by replacing his earlier ‘instantaneous return’ arbitrage argument, the original basis of the Black–Scholes formula, with the construction of dynamic portfolio trading strategies that specified, at each state and date, the actual quantities of each type of security that an investor would hold in order to replicate the final payoff of the target contingent claim.
After 1978 Merton shifted his attention from foundational theories of investment and asset pricing to applications of those theories, paying special attention to the institutional features of financial markets and to related issues of public policy. For example, a series of papers addressed pension planning, social security, and bank deposit insurance. He also worked on corporate capital budgeting, labour contracts, financial intermediation, and the risk management of financial institutions, among many other applications. Merton even turned his hand to some empirical research on investments. His 1987 presidential address to the American Finance Association raised some influential new ideas regarding the impact of market imperfections and incomplete information on equilibrium asset prices.
In 1988 Merton moved from MIT, of whose faculty he had been a member since 1970, to Harvard University. While he has maintained direct involvement in financial markets in various capacities throughout his professional career, for example as a consultant, in 1993 Merton took a more significant step in this direction by becoming one of the first principals of the now notorious hedge fund, Long-Term Capital Management (LTCM). In its first years, the great financial successes of LTCM were attributed in large measure to the unusually deep team of talented financial minds, notably including both Merton and Myron Scholes, which had been assembled by John Merriwether, LTCM's founder. When LTCM failed spectacularly in 1998, some pundits ironically blamed undue reliance on sophisticated financial modelling, in some cases singling out Merton and Scholes. The record, however, seems to point to initial successes based on high leverage, attractive financing, and good trading, and then failure caused by high leverage coupled with the results of some unwise or unlucky trading, exacerbated by a ‘rush to the exits’ by other investors who held large positions similar to those of LTCM. In 2002, Merton co-founded Integrated Finance, a financial advisory firm.
As of this writing, Merton continues to publish and speak influentially, and remains on Harvard's faculty. In addition to the Nobel Prize, Merton is the recipient of numerous awards and honorary degrees, and is widely viewed as one of the all-time most respected leaders and researchers of his profession.

See Also

- contingent valuation
- Miller, Merton

- Modigliani, Franco
- Scholes, Myron
- Sharpe, William F.
- social insurance
- Social Security in the United States

Selected works

1969a. (With P. Samuelson.) A complete model of warrant pricing that maximizes utility. *Industrial Management Review* 10 (Winter), 17–46. Chapter 7 in *Continuous-Time Finance*.

1969b. Lifetime portfolio selection under uncertainty: the continuous-time case. *Review of Economics and Statistics* 51, 247–57. Chapter 4 in *Continuous-Time Finance*. 1992.

1971. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* 3, 373–413. Chapter 5 in *Continuous-Time Finance*, 1992.

1973a. Theory of rational option pricing. Bell Journal of Economics and Management Science 4, 141–83. Chapter 8 in *Continuous-Time Finance*, 1992.

1973b. An intertemporal capital asset pricing model. *Econometrica* 41, 867–87. Chapter 15 in *Continuous-Time Finance*, 1992.

1974. On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* 29, 449–70. Chapter 12 in *Continuous-Time Finance*, 1992.

1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–44. Chapter 9 in *Continuous Time Finance*, 1992.

1977. On the pricing of contingent claims and the Modigliani–Miller theorem. *Journal of Financial Economics* 5, 241–9. Chapter 13 in *Continuous Time Finance*, 1992.

1978. On the cost of deposit insurance when there are surveillance costs. *Journal of Business* 51, 439–52. Chapter 20 in *Continuous-Time Finance*, 1992.

1980. On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics* 8, 323–61.

1983a. On the role of social security as a means for efficient risk-bearing in an economy where human capital is not tradeable. In *Financial Aspects of the US Pension System*, ed. Z. Bodie and J. Shoven. Chicago: University of Chicago Press.

1983b. On consumption-indexed public pension plans. In *Financial Aspects of the US Pension System*, ed. Z. Bodie and J. Shoven. Chicago: University of Chicago Press 1983. Chapter 18 in *Continuous-Time Finance*, 1992.

1985. Implicit labor contracts viewed as options: a discussion of ‘Insurance Aspects of Pensions’. In *Pensions, Labor, and Individual Choice*, ed. D. Wise. Chicago: University of Chicago Press.

1987a. (With Z. Bodie and A. Marcus.) Pension plan integration as insurance against social security risk. In *Issues in Pension Economics*, ed. Z. Bodie, J. Shoven and D. Wise. Chicago: University of Chicago Press.

1987b. A simple model of capital market equilibrium with incomplete information. *Journal of Finance* 42, 483–510.

1987c. (With Z. Bodie and A. Marcus.) Defined benefit versus defined contribution pension plans: what are the real tradeoffs? In *Pensions in the U.S. Economy*, ed. J. Shoven and D. Wise. Chicago: University of Chicago Press.

1990. The financial system and economic performance. *Journal of Financial Services Research* 4, 263–300.

1992. *Continuous-Time Finance*, Rev. edn. Cambridge, MA: Basil Blackwell.

1993a. (With Z. Bodie.) Deposit insurance reform: a functional approach. *Carnegie–Rochester Conference Series on Public Policy* 38, 1–34.

1993b. (With A. Perold.) Theory of risk capital in financial firms. *Journal of Applied Corporate Finance* (Fall), 16–32.

1993c. (With Z. Bodie.) Pension benefit guarantees in the United States: a functional analysis. In *The Future of Pensions in the United States*, ed. R. Schmitt. Philadelphia: University of Pennsylvania Press.

1995. Financial innovation and the management and regulation of financial institutions. *Journal of Banking and Finance* 19, 461–81.

1997. A model of contract guarantees for credit-sensitive, opaque financial intermediaries. *European Finance Review* 1, 1–13.

2000. (With Z. Bodie.) *Finance*. New Jersey: Prentice-Hall.

2005. (With Z. Bodie.) The design of financial systems: towards a synthesis of function and structure. *Journal of Investment Management* 3, 1–23.

Bibliography

Bernstein, P. 1992. *Capital Ideas*. New York: Free Press.

Black, F. 1989. How we came up with the option formula. *Journal of Portfolio Management* 15, 4–8.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Duffie, D. 1998. Black, Merton and Scholes – their central contributions to economics. *Scandinavian Journal of Economics* 100, 411–24.

Greenspan, A. 1998. private-sector refinancing of the large hedge fund, Long-Term Capital Management. Testimony before the Committee on Banking and Financial Services, US House of Representatives, 1 October. Online. Available at http://www.federalreserve.gov/boarddocs/testimony/19981001.htm, accessed 26 July 2005.

Itô, K. 1951. On a formula concerning stochastic differentials. *Nagoya Mathematics Journal* 3, 55–65.

Jarrow, R. and Protter, P. 2004. A short history of stochastic integration and mathematical finance: the early years, 1880–1970. In *The Herman Rubin Festschrift, IMS Lecture Notes* 45, 75–91. Institute of Mathematical Statistics: Bethesda, MD.

Markowitz, H. 1959. *Portfolio Selection: Efficient Diversification of Investment*. New York: Wiley.

Modigliani, F. and Miller, M. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48, 261–97.

Royal Swedish Academy of Sciences. 1998. The Nobel Memorial Prize in Economics 1997. *Scandinavian Journal of Economics* 100, 405–9.

Schaefer, S. 1998. Robert Merton, Myron Scholes and the development of derivative pricing. *Scandinavian Journal of Economics* 100, 425–45.

Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

How to cite this article

Duffie, Darrell. "Merton, Robert C. (born 1944)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000343> doi:10.1057/9780230226203.1090

The New Palgrave Dictionary of Economics Online

Methodenstreit

Daniel R. Fusfeld

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Hobson, J. A.; institutional economics; Menger, C.; Methodenstreit; methodology of economics; Schmoller, G. von; Veblen, T.

Article

The ‘battle of methods’ between Carl Menger (1840–1921) and Gustav Schmoller (1838–1917) is one of the most important methodological debates in the history of economics. It began with the publication of Menger's book on method (1883), which made the case for pure theory based on assumptions about behaviour and antecedent conditions. Schmoller responded with a strongly worded review (1883) that argued for principles of economics based on empirical historical data and the inductive method. Menger answered with an equally vehement statement of *The Errors of Historicism* (1884). The infuriated Schmoller refused even to read it (Schmoller, 1884). A torrent of books and papers by others followed over the next several decades. The best summary of the entire controversy is Ritzel (1951).

Like most disputes over method in economics, the opposing views were related to more complex disagreements over the nature and scope of economics and its policy implications. Menger's assumptions about behaviour implied a social system composed of selfishly motivated individuals; Schmoller assumed the existence of individuals grouped into nations, with group as well as individual goals. More important, Menger's conclusions emphasized the primacy of laissez-faire policies designed to allow as large a scope as possible to market adjustment processes. Schmoller's conclusions supported the interventionist and state-building policies of the newly unified German nation. In addition, the Ministry of Education in Berlin gave almost exclusive preference to the Schmoller school in appointing university professors. Menger was attacking the ‘official’ economics prevailing in Germany and its almost monopolistic control over university appointments. In addition to economic method, academic freedom and the role of the state were at issue.

On the basic issue of the place of theory and empirical studies in economics, Menger and Schmoller agreed that both were necessary. They disagreed, however, on the emphasis to be placed on each and their role in the development of conclusions. Menger argued that ‘pure’ economics based on

assumptions of wide and perhaps universal generality, could be developed through correct logical analysis to arrive at conclusions of equally broad applicability and usefulness. Propositions based on empirical data, however, would be correct only for the limited data on which they were based. Since empirical data were always partial, as well as bounded by time and space, the conclusions drawn from them must be both problematic and of limited generality. Correct and general propositions could be derived through rigorous logic from assumptions not bounded by time, space or special circumstance, however.

Empirical studies entered Menger's method in two ways. First, they could be used to verify or illustrate the results of theoretic inquiry. Second, they were necessary when theoretic principles were applied to specific instances or policy problems. Empirical studies were required to define the situation to which theoretic principles were applied, and to delimit the applicability of the conclusions. Data acted as a bridge between the principles of pure economics and the policy problems of applied economics. Indeed, Menger warned against application of pure theory to applied problems without thorough empirical studies.

Schmoller also advocated use of both empirical studies and theory, but in a different combination. He rejected Menger's logical deductive method for three chief reasons: its assumptions were unrealistic, its high degree of abstraction made it largely irrelevant to the real-world economy, and it was devoid of empirical content. The theory was therefore useless in studying the chief questions of importance to economists: how have the economic institutions of the modern world developed to their present state, and what are the laws and regularities that govern them? The proper method was induction of general principles from historical–empirical studies (Schmoller, 1883). In the Hegelian tradition of 19th-century German scholarship, Schmoller conceived of the economy as a dynamic and evolving set of interrelated institutions whose laws of development could not be understood in terms of an abstract theory of constrained choice. One reason for the polarized arguments of the Methodenstreit was that the disputants were talking about different things.

How were the historical laws of economic development and change to be determined? Schmoller was not clear on that point, although he devoted five chapters of the introductory section of his *Grundriss* to a survey of the history and method of economics (Schmoller, 1900–4). The starting point of his method was empirical research rather than assumptions. The second step was to organize the data in a logical fashion, to bring out the essential nature of economic phenomena. The third step was to identify the relationship between phenomena in the context of their continually changing interaction and development. At all stages of the inquiry, empirical research was to be used to obtain the propositions of steps two and three. The connecting link between data and generalizations was not spelled out, although in retrospect we can interpret the procedure as an early version of the gestalt method and the use of pattern models.

The Methodenstreit had a significant impact on the development of economics. Schmoller's attack on the logical deductive method as inherently devoid of empirical content coincided with similar critiques by the British economic historians, John A. Hobson and the American institutionalists led by Thorstein Veblen. These critics forced the adherents of neoclassical economics to bring empirical studies more fully into the mainstream of economic thought and practice. After the Methodenstreit a combination of theory and empirical studies was almost universally accepted by economists as necessary.

Menger's method of combining them was adopted, however. In the 20th century economics became increasingly a theoretic discipline based on ‘as if’ assumptions, which are developed by rigorous logical

methods to derive general propositions. Hypotheses about reality, derived from the general propositions, are then tested against empirical studies. Schmoller's vision of an empirical discipline based on factual studies, in which generalizations are both derived from and tested against data as they are developed, remains only among critics of the mainstream in a new battle of methods that has erupted a hundred years later.

See Also

- Historical School, German
- institutional economics
- Menger, Carl
- methodology of economics
- Schmoller, Gustav von

Bibliography

Menger, K. 1883. *Untersuchungen über die Methode der Sozialwissenschaften, und der politischen Oekonomie insbesondere*. Berlin: Duncker & Humblot. Ed. L. Schneider and trans. F.J. Nock as *Problems of Economics and Sociology*, Urbana: University of Illinois Press, 1963.

Menger, K. 1884. *Die Irrthümer des Historismus in der deutschen Nationalökonomie*. Vienna: Alfred Hölder.

Ritzel, G. 1951. *Schmoller Versus Menger*. Offenbach: Bollwerk-Verlag.

Schmoller, G. von. 1883. Zur Methodologie der Staats- und Sozialwissenschaften. *Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft im deutschen Reich* 8, 974–94.

Schmoller, G. von. 1884. *Jahrbuch für Gesetzgebung, Verwaltung und Volkwirtschaft in deutschen Reich*, 677.

Schmoller, G. von. 1900–4. *Grundriss der allgemeinen Volkswirtschaftslehre*. Leipzig: Duncker & Humblot.

How to cite this article

Fusfeld, Daniel R. "Methodenstreit." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000155> doi:10.1057/9780230226203.1091

The New Palgrave Dictionary of Economics Online

methodological individualism

Kaushik Basu
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Methodological individualism holds that a proper explanation of a social regularity or phenomenon is grounded in individual motivations and behaviour. Although many economists claim to be methodological individualists, economics has always used social concepts and categories. As Schumpeter pointed out, nearly a century ago, price in a competitive model is an irreducibly social concept. Each individual takes the price as given but the price that comes to prevail is an outcome of the choices made by all individuals. Since Veblen, economists have increasingly recognized that individual preferences are endogenous and may be responsive to what happens in society at large.

Keywords

Arrow, K.; atomism; Bertrand oligopoly; Buchanan, J.; class interest; Cournot oligopoly; Dworkin, R.; endogenous preferences; German school; Hayek, F. A.; homo economicus vs. homo sociologicus; Menger, C.; methodological holism; methodological individualism; normative individualism; Schumpeter, J.; self-interest; Smith, A.; Veblen, T.; Weber, M.

Article

Methodological individualism is a doctrine in the social sciences according to which a proper explanation of a social regularity or phenomenon is one that is grounded in individual motivations and behaviour. In other words, according to this methodology, individual human beings are the basic units from which we must build *up* in order to understand the functioning of society, economy and polity. We may not in all our research succeed in doing so but to committed methodological individualists such research must be viewed as provisional and ideally be accompanied by a slight feeling of inadequacy on the part of the researcher.
The social scientists who have been the focus of much of the debate on methodological individualism and, paradoxically, also the ones least touched by the debate are economists. Economists are typically

held up as examples of the most unbending methodological individualists; and, on the rare occasions when economists have joined this debate, they have tended to agree with this. The difference is that most non-economists mean this as criticism, whereas most economists take it as praise.

At first sight this characterization of economics seems right. Textbooks of microeconomics almost invariably begin by specifying individual utility functions or preference relations and asserting that human beings are rational in the sense that they behave so as to maximize their own utilities. They then build up from this to explain market phenomena, make claims about social welfare and discuss prospects of national economic growth. In some macroeconomic models economists are unable to build all the way up from individual behaviour and use aggregate behaviour descriptions as the starting point. But these models are almost always accompanied by an effort to ‘complete’ them with proper micro-foundations; and the profession regards these models as somewhat incomplete and awaiting the definitive work.

That economics may not actually be quite as methodologically individualistic as often presumed by both the discipline's admirers and its critics is a matter to which I return below. What is interesting to note here is that the debate on methodological individualism has been a surprisingly cantankerous one that has spawned enemies and intrigues. Some social scientists have sworn by it: no other method is worth its salt. Others have castigated it as an instrument of exploitation and maintenance of the status quo. Concepts and categorizations have multiplied over the years. We have come across methodological holism, methodological solipsism, atomism, ‘MIs’ (that is, methodological individualisms) of different types – 1, 2, 3… – creating the impression that the British intelligence had somehow got involved in the quest to understand this elusive concept.

One cause of the controversy is the confounding of positive and normative social science. To some commentators, methodological individualism implies that it is fine to leave it all to individuals, and by implication it amounts to an argument against government intervention. Friedrich von Hayek (1942) and James Buchanan (1989), for instance, have taken this line, as have some sociologists, who felt that the conservatism of traditional economics is founded in its adherence to methodological individualism. But this happens because of a possibly logical error, a failure to appreciate Hume's law, namely, that a normative proposition cannot be derived from a purely positive analysis. Kenneth Arrow (1994) has rightly criticized the tendency of some writers to treat methodological individualism and ‘normative individualism’ as inextricably linked. Similarly, Marxists often link methodological individualism automatically to certain ethical implications. Roemer (1981) and Elster (1982) argue that this is not a valid link. In what follows I treat the two as separate and assume that methodological individualism has no automatic normative implications.

Origins

The term ‘methodological individualism’ was probably used for the first time in the English language in 1909 by Joseph Schumpeter. Even if that is not so, Schumpeter certainly thought so, and he pointed out in his paper in the *Quarterly Journal of Economics* that year that he had actually coined the term in German the previous year. But methodological individualism had been *practised* from much earlier, at least as early as Adam Smith (1776), and was described as a deliberate methodology, though without the term itself being used, by Carl Menger in 1883 (Menger, 1883). Max Weber's later exposition of it was

published posthumously in 1922 (Weber, 1922).

From the perspective of economics it seems reasonable to treat Menger as the first proponent of methodological individualism. He did so vociferously, dismissing the German historical school of economists and their methods as outdated and flawed. He advanced the idea of ‘spontaneous order’ in society, which sprang from atomistic individual behaviour, reminiscent of Adam Smith's ‘invisible hand’ and the efficiency of markets that was an outcome of rational, self-interested behaviour on the part of individuals. Menger not only failed to acknowledge that some of his ideas on spontaneous social order were already there in Adam Smith but wrote in a tone almost suggesting that Smith had taken those ideas from him.

A distinction is often drawn in philosophy between methodological individualism and ‘atomism’. The latter is treated as a more extreme version of individualism, in which it is possible to characterize each individual fully without reference to society and then explain social behaviour by simply imagining such individuals being brought together in one society. Since the proponents of these ideas did not really define terms with that much care – and when they did, they went on to write in a way that disregarded their own definitions – I shall refrain from drawing fine distinctions and treat these neighbouring terms as all representing the broad *idea* of methodological individualism. Moreover, concepts like these are probably innately indefinable. They are understood through a combination of approximate definitions and repeated use.

It is useful in an exposition like this to think of the polar opposite of the term under consideration. This is captured in the concept of ‘methodological holism’, developed (without endorsement) by the philosopher John Watkins. Methodological holism is the belief that there are ‘macroscopic laws that are *sui generis* and apply to the system as an organic whole’ (Watkins, 1952, p. 187), and the behaviour of its components had to be deduced from it. In economics, this would imply beginning our analysis by stating the laws of an aggregate economy and, perhaps, the behaviour of prices and industries and, from that, deducing how individuals behaved and what motivated them. Stated in these terms, it immediately becomes clear from a perusal of almost any microeconomics textbook that economics belongs essentially to the methodological individualist end of the spectrum defined by methodological holism at one end and individualism at the other.

After these writings, interest in the subject flagged. Social scientists, especially economists, continued to do research without trying to explicitly articulate the method that they were in fact using. The feeling developed among economists that the issue of methodological individualism was either trivial or had been resolved in their favour.

In the early 1990s the economists’ gathering insouciance was challenged by Rajeev Bhargava (1993) and Kenneth Arrow (1994). Bhargava summarizes various points of view on the subject and then challenges the orthodoxy, especially within economics. But he also expresses well the philosopher's inevitable anxiety in a debate like this, which stems from not knowing whether what one is grappling with is something profound or trivial. As he writes, ‘On reading the literature one is swung between exuberance and despair, from feeling that all problems have been resolved to one that none has … Gradually an intense frustration overwhelms the reader: perhaps there was nothing worth discussing in the first place. What on earth was all the fuss about?’ (1993, p. 5).

What he settles for as the best face of methodological individualism is ‘intentionalism’. The intentional man is somewhere between the imaginary *homo economicus* and equally rare *homo sociologicus*. He can choose and decide individually but he is not a relentless, maximizing agent. He has psychology and a

sense of social norms, which get in the way of selfish maximization. Bhargava then develops the idea of ‘contextualism’ as a challenge to methodological individualism, including intentionalism. The challenge consists of arguing that a variety of beliefs and practices in everyday life make sense only in the *context* of the society where they occur. Hence, in describing a society or an economy we are compelled to use concepts which are *irreducibly social*.

The reason why the assertion that certain beliefs and concepts are inextricably social is unlikely to stir a hornet's nest is that, although many economists claim to be rigid adherents of methodological individualism, they do use and have always used social concepts and categories. This is convincingly argued by Arrow. He points out how a variable such as price in a competitive model is an irreducibly social concept. Each individual takes the price to be given but the price that comes to prevail is an outcome of the choices made by all the individuals. So economists constructing equilibrium models, who claim to be hardened methodological individualists, are actually not so, at least in the sense that they use some concepts that are irreducibly social. Knowingly or unknowingly they follow a method which uses social categories. In fact, this was explicitly recognized by Schumpeter in his classic essay on methodological individualism, where he noted ‘prices are obviously social phenomena’ (1909, p. 217).

Preferences and groups

There are more contentious claims that one can make about the role of social concepts in economics. One of these relates to the *permissibility* of a certain class of propositions in social science, such as: ‘The landlord will undertake action A, *because* it is in the landlord's class interest to do so.’ (Action A could, for instance, be: ‘refuse to hire a servant who has fled another landlord's employment and offers to work for this landlord for a low wage’). Let me call this proposition P.

The bone of contention between neoclassical and traditional Marxist economists is frequently whether such propositions are permissible. Many neoclassical economists and some political scientists (especially those belonging to the positive political economy school) believe that P is not permissible – a person's class interest must be not be treated as an innate characteristic in the same way that his self-interest may be. A small group of writers maintain that even Marxism is compatible with methodological individualism and that class and other aggregate behaviour should, ideally, be built from individual motivations and preferences (Roemer, 1981; Elster, 1982).

In any case, whether or not proposition P is wrong, mainstream economics certainly considers it so. If an economist were to use an axiom like proposition P, she would usually want to first satisfy herself why it may be in the landlord's *self*-interest to behave in a way which is in his *class* interest. However, this does not negate the use of beliefs and other concepts and variables which are irreducibly social. It is not clear whether a researcher who does both (that is, resists explaining individual behaviour solely in terms of its ability to serve group or class interests but uses concepts and beliefs which are inherently social) is a methodological individualist. But this is a purely definitional matter and of no great consequence. The important and contestable question is whether assumptions like proposition P should or should not be used. I take the view that it is best to avoid such assumptions as far as possible, without making that into a dogma.

There are some fundamental ways in which modern economics has moved further away from methodological individualism than merely by using irreducibly social concepts, like prices, and even

without using propositions, like P. I here mention two. First, most models of economics make use of the idea of ‘rules of the game’. In Cournot oligopoly, firms choose quantities and then wait for prices to form. In Bertrand oligopoly, firms set prices and then wait to sell what the market demands from them. In most real-life situations, these rules evolve over time through intrinsically social processes. We may not fully understand what these social processes are, but few individuals will deny their existence. Arrow (1994) has emphasized this and also the importance of ‘*social* knowledge.’

Second, there is increasing recognition in economics that individual preferences are endogenous. They evolve over time and may be responsive to what happens in society at large. As Thorstein Veblen (1899) recognized, around the time when neoclassical economics was taking shape, human preferences for certain objects often depend on who else is consuming those objects. If a film star wears a brand-name shirt, you may be willing to pay more for that same shirt. If the elite likes a particular wine, then some people will acquire a taste for that wine; moreover, such people will be viewed by others as belonging to the elite because of their taste in wine. In other words, people often use goods to associate themselves with other people who use those goods (Basu, 1989). These are obvious matters (though they were sidelined during the time of Veblen) and any economist whose ability to think is not damaged by excessive textbook education will recognize that these kinds of preference endogeneity exist. What is remarkable about Schumpeter's (1909) essay is that he understood (admittedly in a somewhat inchoate way) that this recognition may cut into the methodological individualism of economics. He observed how, given the human tendency to conform to society, ‘there will be a tendency to give [each individual's] utility curves shapes similar to those of other members of the community’ (1909, p. 219). To see how this can ruffle methodological individualism, suppose that each person likes to wear jeans if more than 60 per cent of society wears jeans; more precisely, suppose that, if over 60 per cent of society wears jeans, each person is willing to pay for jeans more than the marginal cost of producing them; otherwise they are willing to pay less. This society will have two possible equilibria: one in which no one wears jeans and another (however revolting it may be to visualize this) in which everyone wears jeans. In models of this kind there is an interdependence between society's behaviour and each individual's preference. Once we recognize this, there is no reason to start our analysis by characterizing the individual. We may still do so through force of habit. But we could equally begin by considering a social behaviour postulate – for instance, that 50 per cent of people wear jeans. Then we work out how much each individual prefers to wear jeans (and so how much each is willing to pay for his or her jeans) and check whether the initial social postulate is sustainable. If it is, then we have found an equilibrium. If not (as in the above example), then the behaviour is not one that will prevail in equilibrium. This method is one of neither methodological individualism nor methodological holism. It is therefore evident that, as economics becomes more sophisticated, it is moving away from pure individualism towards this kind of a hybrid methodology.

Normative statements

An interesting and unexpected area where methodological individualism is violated is in some of our normative statements. We often pass moral judgement on groups of people which cannot be reduced to the individuals in the group. Normative propositions of the following kind are common: ‘It is a shame that no one in your university does research on poverty.’ If you asked the person making this

observation whether he was blaming you for not doing research on poverty, he would typically claim that he was not; in fact, he would deny that he was casting aspersions on *any* individual but criticizing the collectivity of individuals in the university. This amounts to an implicit rejection of individualism.

Methodological individualism in the context of normative statements like the above one has not been much analysed, but Ronald Dworkin has provided an interesting analysis. He argues that in situations of group responsibility it may be reasonable to *personify* the group. Thus, when a corporation produces a dangerously defective good but it is not possible to pin down the responsibility on any particular individual, we may need to treat the corporation as a moral agent and apply ‘facsimiles of our principles about individual fault and responsibility to it’ (Dworkin, 1986, p. 170). And *then*, by virtue of the *corporation's* responsibility, we may proceed to hold some or all of the individual *members* of the corporation responsible. This is interesting because it comes as close to Watkins's ‘methodological holism’ as we are likely to encounter anywhere. Individuals are still essential units in Dworkin's analysis but, unlike in standard methodological individualism, judgement of the group *precedes* judgement of the individual.

Dworkin argues that we unwittingly often use this method. This happens when we talk of the state's responsibility for certain kinds of individual rights. Thus we talk of the state's obligation to ensure that no one is assaulted by others. Moreover, we do this even before agreeing on how this responsibility is to be apportioned across various units and agents of the state, such as the police and the bureaucracy. Dworkin (1986, p. 171) points out how we discuss the community's responsibility and ‘leave for *separate* consideration the different issue of which arrangement of official duties would best acquit the communal responsibility’ (emphasis added).

It is possible to criticize Dworkin's line (see Basu, 2000) by arguing that the personification of the corporation or the community has to be an interim construct. It will be sustained if we *can* then apportion the blame among the members of the corporation. If, however, we find that we are not able to spread the blame among the individuals in some reasonable way, then we may have to forgo our initial stand, which held the corporation responsible, or at least maintain that there is no way to take the next step of tracing the fault to individuals.

Interestingly, this brings us back to the kind of analysis defended in the case of endogenous preferences. And this suggests, once again, that what is needed for modern social science is neither holism not individualism but a hybrid methodology that, at least for now, lacks a name.

See Also

- collective rationality
- economic man
- individualism versus holism
- social norms

Bibliography

Arrow, K. 1994. Methodological individualism and social knowledge. *American Economic Review* 84,

1–10.

Basu, K. 1989. A theory of association: social status, prices and markets. *Oxford Economic Papers* 41, 653–71.

Basu, K. 2000. *Prelude to Political Economy: A Study of the Social and Political Foundations of Economics*. Oxford: Oxford University Press.

Bhargava, R. 1993. *Individualism in Social Science: Forms and Limits of Methodology*. Oxford: Oxford University Press.

Buchanan, J. 1989. The State of Economic Science. In *The State of Economic Science*, ed. W. Sichel. Kalamazoo, MI: Upjohn Institute for Employment Research.

Dworkin, R. 1986. *Law's Empire*. Cambridge, MA: Harvard University Press.

Elster, J. 1982. Marxism, functionalism and game theory. *Theory and Society* 11, 453–82.

Hayek, F. von. 1942. Scientism and the study of science. *Economica* 9, 267–91.

Menger, C. 1883. *Investigations into the Method of the Social Sciences with Special Reference to Economics*. English translation. New York: New York University Press, 1986.

Roemer, J. 1981. *Analytical Foundations of Marxian Economic Theory*. Cambridge: Cambridge University Press.

Schumpeter, J. 1909. On the concept of social value. *Quarterly Journal of Economics* 23, 213–32.

Smith, A. 1776. *Inquiry into the Nature and Causes of the Wealth of Nations*. Indianapolis: Liberty Classics, 1981.

Veblen, T. 1899. *The Theory of the Leisure Class*. London: Macmillan.

Watkins, J. 1952. The principle of methodological individualism. *British Journal for the Philosophy of Science* 3, 186–89.

Weber, M. 1922. *Economy and Society*, vol. 1. New York: Bedminster Press, 1968.

How to cite this article

Basu, Kaushik. "methodological individualism." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000352> doi:10.1057/9780230226203.1092

The New Palgrave Dictionary of Economics Online

methodology of economics

Roger E. Backhouse
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The methodology of economics concerns the principles underlying economic argumentation. Though systematic reflections on the subject go back at least to the 19th century, the methodology of economics emerged as a recognizable field, at the boundaries of economics, philosophy and science studies, only in the 1980s. This article outlines the way the field developed and its current position.

Keywords

assumptions controversy; behavioural economics; Cairnes, J. E.; causality; conventionalism; data mining; economics, definition of; experimental economics; explanation; falsificationism; Friedman, M.; German Historical School; heterodox economics; history of economic thought; ideology; individualism versus holism; instrumentalism; Jevons, W. S.; Keynes, J. N.; Koopmans, T. C.; Kuhn, T.; Lakatos, I.; Leslie, T. E. C.; measurement theory; Menger, C.; Mill, J. S.; models; operationalism; paradigms; philosophy and economics; pluralism in economics; Popper, K.; positive economics; positivism; postmodernism; Robbins, L, C.; stylized facts; theory appraisal

Article

Since the 1970s, the methodology of economics has developed from a series of reflections by practising economists on the methods employed in their field, to a field at the boundaries of economics and philosophy (and to a lesser extent sociology). After an initial focus on falsificationism, the range of issues pursued has considerably broadened.

In the social sciences, which include economics, the term ‘methodology’ is used with two different meanings. When an article or thesis contains a section called ‘methodology’ in which the author explains how a piece of research was conducted the word is used as a synonym for ‘method’. In the literature on ‘the methodology of economics’, on the other hand, it is used as a label for enquiries into the principles underlying economic reasoning. This article is concerned only with the second of these

two meanings.

The methodology of economics is inevitably an interdisciplinary activity. Economists may analyse their own reasoning using ideas drawn from philosophy, sociology, linguistics or discourse analysis, or they can draw simply on their own experience as practising economists. For philosophers, enquiries into economic methodology are part of philosophy – a branch of the philosophy of science or, if the word science is thought inappropriate, of knowledge. Economic methodology is thus largely covered by the article on philosophy and economics. The two are, however not synonymous, for the latter covers decision theory, rational choice and ethics, fields not traditionally thought of as economic methodology. That article traces the interrelations between these disciplines from the 19th century to the present. However, though this overlaps with the story of economic methodology, the two are not synonymous.

The explicit study of methodology has always had a mixed reputation within economics. Most of the time, economists simply get on with their work, reflecting on specific methodological problems as and when they arise, refraining from more general speculation. They are suspicious of general theories about how to practise economics (or any other subject for that matter), especially when such theories are written by those who do not themselves engage in the research they are analysing. Against this there are those who believe that methodological reflection by those who are more distant from practice, whether they are trained as economists or philosophers, even if it does not tell economists how to do their work better, can provide a valuable perspective on what economists do that would be otherwise be missed. When it comes to publication, some, even if they find methodological argument valuable, hold that it should not have a place within economics journals as it is not economics, but writing about economics. A further reason for scepticism is that methodological arguments are frequently used by non-economists and heterodox economists to show that certain economic theories cannot possibly be right: it is held that, rather than speculate on methodology, those who believe this would do better if they showed by example, how things could be done better. This attitude has a parallel in divisions within the field of economic methodology between those who believe that the task of methodology is primarily to understand what economists do (a stance that does not imply an absence of criticism, even if the methodologist deliberately refrains from telling economists what to do) and those who use methodological arguments to argue for heterodox positions within economics. These two categories overlap significantly, but this divide nevertheless reflects important tensions within the field.

Historical background

The 19th century was an age when disciplinary boundaries began to be established. Given the extremely high regard in which John Stuart Mill was held by contemporaries, both as a philosopher and as an economist, it would be rash to classify him according to modern disciplinary categories. His *Logic* (1843) was a standard textbook in the philosophy of science and his *Essays on Some Unsettled Questions of Political Economy* (1844) were an influential statement of economic methodology. Methodological arguments among economists were frequent (see, for example, the work collected in Smyth, 1962; Backhouse, 1997) and were primarily by economists using methodological arguments to criticize positions with which they disagreed. Cliffe Leslie and John Elliott Cairnes are good examples. Both established reputations for their work in economics itself, but wrote extensively on methodology, Leslie in a series of essays (1879) and Cairnes in *The Character and Logical Method of Political*

Economy (1857). William Stanley Jevons made a methodological case for a particular way of practising economics in his *Theory of Political Economy* (1871) but in addition to being a leading economist was also the author of *The Principles of Science* (1873), a major textbook in the philosophy of science. If anyone should be classified as a professional methodologist in this period, it is John Neville Keynes, author of a textbook in formal logic, but whose *The Scope and Method of Political Economy* (1890) was his major work. Even if some considered it a worthy book, but one that students did not in practice need to bother with, it played a role in establishing the Marshallian consensus within British economics and preventing the methodological dispute between the Carl Menger (1883) and the German Historical School from dividing the British profession in the same way as it divided German-speaking economists. There were methodological disputes over historicism in British economics, but they were nowhere nearly as divisive as the German.

The tradition of economists writing article and occasionally book-length reflections on methodology continued through the 20th century, and was linked to disputes over the direction in which economics should be moving. In the first half of that century, the most influential such work was undoubtedly Lionel Robbins's *An Essay on the Nature and Significance of Economic science* (1932), which helped define modern welfare economics and, more broadly to redefine the subject, though this took much longer than is commonly believed (Backhouse and Medema, 2007). The 1930s saw a profusion of articles and books on methodology, many of which discussed Robbins's *Essay*. However, what we find is a literature that, though containing much that was perceptive, can be seen as a series of comparatively isolated works in which trends are hard to identify.

After the Second World War, this pattern continued, though the literature became more focused, due to the way economic theory was developing. There was also, in the background, the emergence of what came to be known as the ‘received view’ in the philosophy of science and the work of Karl Popper, though lags in translation meant that this permeated the economic literature only gradually. Several of the leading economists wrote on methodology, their work being given focus, despite their varied perspectives, by their concern with models and the role of assumptions. The most influential work was Milton Friedman's ‘The Methodology of Positive Economics’ (1953) with its provocative thesis that it was actually desirable for the assumptions of a theory to be unrealistic. Good theories are ones that pick out the relevant features of reality, using a sparse set of assumptions to explain a wide range of phenomena, which means that they must be descriptively unrealistic. Tjalling Koopmans included an essay defending unrealistic models on quite different grounds in his *Three Essays on the State of Economic Science* (1957): unrealistic models should be seen not in isolation but as part of a series of models – they are prototypes of subsequent models that will be more realistic. These, and above all Friedman's essay, provoked a significant literature. The question of testability linked this with issues being discussed in the philosophy of science in a way that was not generally true of the period before the Second World War (an exception was Hutchison, 1938).

Methodology of economics as a field

The emergence of the methodology of economics as a recognizable field within economics, into which this earlier literature could (retrospectively) be incorporated, is best dated to the appearance of Mark Blaug's *The Methodology of Economics* (1980). This was not the first textbook on economic

methodology but it served to define a field in a way that previous textbooks (for example, Stewart, 1979) had not. It offered a survey of what economists needed to know about the philosophy of science and a series of case studies in economics. The theme of the book was the importance of falsificationism, as found in the work of Karl Popper and Imre Lakatos. He offered a typically robust conclusion:

> [T]he ultimate question we can and must pose about any research program is the one made familiar by Popper: what events, if they materialized, would lead us to reject that program? A program that cannot meet that question has fallen short of the highest standards that scientific knowledge can attain. (Blaug, 1980, p. 264)

The common practice among economists was what he called ‘innocuous falsificationism’: to preach falsificationism but not to practise it. The theme of his case studies was that the subject had made progress when economists had sought to test theories, even when such testing had, as in the examples of human capital theory or monetarism, been inconclusive.

Blaug's book was important. His thesis offered a challenge, both to those who felt that there must be reasons why economists approached their subject in ways that, according to Blaug, were fundamentally flawed, and to those who were concerned about the philosophical coherence of falsificationism. This defined a research agenda. His approach to methodology also pointed to ways in which it could be combined with the history of economic thought. Popperian methodology, especially in its Lakatosian variant, with its focus on progress, provided a criterion that could be used to assess the history of economics, an approach already explored in Latsis (1976). For the first time, economic methodology came to be linked with the history of economic thought.

A further stimulus to economic methodology came from heterodox economics. Movements such as radical economics, Post Keynesian economics and Austrian economics, though their proponents might construct longer histories, originated in the late 1960s and early 1970s, and were characterized by methodological critiques of the way they saw economic enquires being undertaken. Their interest in both methodology and the history of economics brought tensions: their interest in the subject was welcomed but this was associated with concerns that their ideological commitments might cause a problem for the field (in his article on history of economic thought, Goodwin hints at similar concerns).

The result of this activity was the emergence of an identifiable field of economic methodology, defined not simply by textbooks but by a community of specialists engaging with each other as well as economists and philosophers who chose to explore the subject. This blend of economics and philosophy was reflected in the specialist journals, those publishing articles in English including *Economics and Philosophy* (established 1985), *Journal of Economic Methodology* (1994), *Research in the History of Economic Thought and Methodology* (1983), and in anthologies such as Caldwell (1993) or Hausman (1994) (the former contains a useful list of previously anthologized articles). There are also foreign-language journals, of which *Revue de philosophie économique* has begun publication in English. The difference between the *Journal of Economic Methodology* and *Economics and Philosophy* illustrates the point made earlier that, though there is substantial overlap, economic methodology is not synonymous with philosophy and economics.

The emergence of the field of economic methodology was centred on the philosophy of Popper and Lakatos. By 1990, these approaches had ceased to be dominant. Though some (for example, Blaug,

Lawrence Boland and Terence Hutchison) continued to defend it, falsificationism was generally seen as too restrictive a criterion against which to appraise economic theories: the methodologies of Popper and Lakatos had technical problems and there were good reasons why economists behaved differently. The most significant problem arises from the fact that theories are almost never testable on their own, creating problems with what it means for a theory to be falsifiable (see falsificationism). Lakatosian methodology raised questions concerning the definition of a research programme and the meaning of novelty (see paradigms). Corroboration seemed more important than either Popper or Lakatos admitted. The most influential alternative was articulated by Daniel Hausman's *The Inexact and Separate Science of Economics* (1992). Dismissive of Popper and Lakatos, this book opened up other themes, such as ways of thinking about economic models, but its significance was engaging in what Hausman elsewhere summarized as ‘empirical philosophy of science’. This involved exploring in detail economists’ practices, his main example being the way economists had responded to the phenomenon of preference reversals. Hausman defended the view that economics was, as Mill had expressed it in the mid-19th century, an inexact science, but he was more critical of the motivation, implicit in economists’ responses to experimental evidence, to keep their science separate from any dependence on philosophy. This conclusion may, at least in part, have been rendered out of date by the rise of behavioural economics, but its significance lay in the method of starting from economics, drawing out methodological conclusions (as opposed to the method) characteristic of the Popperian era, and of applying the methodology developed in the context of natural science to economics.

The problems involved in defining Lakatosian research programmes were widely considered to render the concept a rather blunt tool for analysing economics. Instead, the trend was towards analysing problems that arose in particular fields of economics. The rise of experimental economics raised new methodological questions about experimental procedures and the transferability of experimental results to behaviour out of the laboratory. Econometric practices raised issues not covered in traditional methodology such as the significance of data mining, the meaning of causality and how measurement of economic quantities should be understood. Disagreements over the relation between macro and microeconomics raised questions about individualism and the meaning of aggregate analysis, many of which were familiar to the philosophy of social science. Economists had come, almost universally, to argue in terms of models, raising the question of what was going on in the process of economic modelling. Postmodernism failed to have anything like the effect that it had in some other social sciences, but some postmodern ideas were explored. Older questions, such as conventionalism, instrumentalism, positivism and falsification, all concerned with questions of theory appraisal, remained, but they received proportionately much less attention.

The methodology of economics today

The methodology of economics has remained a field with very elastic boundaries. There is a substantial literature in which specialists in the field engage with each other, but perhaps more than most fields in economics, it is one where outsiders, who engage in varying degrees with this literature, have much to say. These outsiders include philosophers, economists specializing in other fields and other social scientists. This results in a variety of perspectives, one of the main divides concerning the extent to which writers see methodology as aimed at understanding economists’ practices and the extent to which

they seek to criticize those practices. Sometimes these aims overlap, but sometimes they do not, those concerned with explication considering that others do not take economics sufficiently seriously, and those concerned with criticism considering that the others are too defensive about what they find within the discipline.

The result is that it has become increasingly difficult to find a framework within which to offer a coherent survey of the field. The most comprehensive recent attempt is Wade Hands's *Reflection without Rules* (2001), which he describes as an ‘interpretive survey’ of the economic methodology. Part of the difficulty with such a task is, as Hands (2001, p. ix) recognized, that any such survey is aiming at a moving target. That would be true in any field; however, a further difficulty is that much work on methodology cuts across the philosophical categories he employs to structure his survey and hence provide the basis for his interpretation. His starting point (as for Blaug, 1980; Caldwell, 1982), was the breakdown of the received view within the philosophy of science, after which he identified a series of turns – the naturalistic, the sociological, the pragmatic and the economic – as well as attempts to develop the Popperian, Millian and other traditions. From the point of view of elucidating the philosophical foundations of economic methodology, this is a successful strategy. However, this framework does not shed much light on some practice-based methodological work. For example Hoover's (2001a; 2001b) work on macroeconomics and causality could be fitted into its framework but, despite its deep philosophical engagement, it is not clear that it is helpful to do so. The problems with work by reflective practitioners, such as Reder (1999) or Goldfarb (1997), simply does not fit at all.

Economic methodology has become a very active field that tackles a range of questions that is much broader than was the case in, say, 1950. In part this reflects the broadening that has taken place within economics, which has meant that methodologists have faced the challenge of broadening their focus to encompass developments as varied as experimental economics and time-series econometric methods. There has also been a shift away from abstract questions of theory appraisal towards understanding the variety of practices found in economics, the aim being to work out a philosophical framework that is appropriate to economics rather than simply applying one derived from consideration of natural science. Finally, economic methodology has turned not only to philosophy of science, as that term has traditionally been understood, but also to disciplines such as sociology, linguistics and science studies. This plethora of new developments suggests that it may change as much in the next quarter century as it has done in the past.

See Also

- assumptions controversy
- causality in economics and econometrics
- conventionalism
- data mining
- economics, definition of
- experimental methods in economics
- explanation
- falsificationism

- Historical School, German
- history of economic thought
- individualism versus holism
- instrumentalism and operationalism
- measurement
- models
- paradigms
- philosophy and economics
- pluralism in economics
- positive economics
- positivism
- postmodernism
- stylized facts
- theory appraisal

Bibliography

Backhouse, R.E. 1997. *The Methodology of Economics: Nineteenth-Century British Contributions*, 6 vols. London and Bristol: Routledge & Thoemmes Press.

Backhouse, R.E. and Medema, S.G. 2007. Defining economics: Robbins's essay in theory and practice. Working paper. Abstract online. Available at http://ssrn.com/abstract=969994, accessed 19 May 2007.

Blaug, M. 1980. *The Methodology of Economics: How Economists Explain*, 2nd edn. Cambridge: Cambridge University Press, 1992.

Cairnes, J.E. 1857. *The Character and Logical Method of Political Economy*. London: Macmillan, 1888.

Caldwell, B.J. 1982. *Beyond Positivism: Economic Methodology in the Twentieth Century*. London: Allen & Unwin.

Caldwell, B.J. 1993. *The Philosophy and Methodology of Economics*, 3 vols. Cheltenham: Edward Elgar.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*, ed. M. Friedman. Chicago: University of Chicago Press.

Goldfarb, R. 1997. Now you see it, now you don't: emerging contrary results in economics. *Journal of Economic Methodology* 4, 221–44.

Hands, D.W. 2001. *Reflection without Rules: Economic Methodology and Contemporary Science*

Theory. Cambridge: Cambridge University Press.

Hausman, D.M. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Hausman, D.M. 1994. *The Philosophy of Economics: An Anthology*. Cambridge: Cambridge University Press.

Hoover, K.D. 2001a. *The Methodology of Empirical Macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K.D. 2001b. *Causality in Economics*. Cambridge: Cambridge University Press.

Hutchison, T.W. 1938. *The Significance and Basic Postulates of Economic Theory*. London: Macmillan.

Jevons, W.S. 1873. *The Principles of Science: A Treatise on Logic and Scientific Method*, 2nd edn. London: Macmillan, 1877.

Jevons, W.S. 1871. *The Theory of Political Economy*, 2nd edn. London: Macmillan, 1879.

Keynes, J.N. 1890. *The Scope and Method of Political Economy*. London: Macmillan, 1917.

Koopmans, T.C. 1957. *Three Essays on the State of Economic Science*. New York: McGraw Hill.

Latsis, S.J. 1976. *Method and Appraisal in Economics*. Cambridge: Cambridge University Press.

Leslie, T.E.C. 1879. *Essays on Political and Moral Philosophy*. Dublin: Hodges, Foster & Figgis.

Menger, C. 1883. *Investigations into the Method of the Social Sciences with Special Reference to Economics*, tr. F.J Nock. New York: New York University Press, 1985.

Mill, J.S. 1843. *A System of Logic. The Collected Works of John Stuart Mill*, vols 7–8, ed. J.M. Robson. Toronto: University of Toronto Press, 1974.

Mill, J.S. 1844. Essays on Some Unsettled Questions of Political Economy. In *The Collected Works of John Stuart Mill*, vol. 4, ed. J.M. Robson. Toronto: University of Toronto Press, 1967.

Reder, M.W. 1999. *Economics: The Culture of a Controversial Science*. Chicago: University of Chicago Press.

Robbins, L.C. 1932. *An Essay on the Nature and Significance of Economic Science*. London:

Macmillan, 1935.

Smyth, R.L. 1962. *Essays in Economic Method: Selected Papers read to Section F of the British Association for the Advancement of Science, 1860–1913*. London: Duckworth.

Stewart, I. 1979. *Reasoning and Method in Economics*. New York: McGraw Hill.

How to cite this article

Backhouse, Roger E. "methodology of economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000393> doi:10.1057/9780230226203.1093

The New Palgrave Dictionary of Economics Online

Metzler, Lloyd Appleton (1913 – 1980)

George Horwich, John Pomery

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

dynamic stability; flexible exchange rates; gross substitutes; income determination; inventory cycles; inventory investment; Laursen–Metzler effect; matrix theory; Metzler matrix; Metzler paradox; Metzler, L. A.; optimal tariffs; perfect stability; propensity to consume; purchasing power parity; real balances; stability; tariffs; terms of trade; transfer problem; wealth effect

Article

Metzler was born in Lost Springs, Kansas and took AB and MBA degrees at the University of Kansas. He was one of a long line of brilliant students that John Ise sent to Harvard, where Metzler arrived in 1937. He served as an instructor and tutor, receiving his Ph.D. in 1942. From 1943 to 1946 he held a number of positions with government agencies in wartime Washington, including the Office of Strategic Services, several economic policy and planning commissions, and, from 1944 to 1946, with the research staff of the Board of Governors of the Federal Reserve System. From 1946 to 1947 he was a member of the economics department of Yale University. In 1947 he joined the department at the University of Chicago, where he remained until his death. His health declined in the early 1950s; removal of a brain tumour in 1952 left him with a markedly reduced energy level. He continued to teach and produced an occasional paper for the next 20 years.

Metzler's *Collected Papers*, most of them written between 1941 and 1951, were published by Harvard University Press in 1973. A Festschrift, *Trade, Stability, and Macroeconomics*, co-edited by his fellow student Paul Samuelson and one of his own students, George Horwich, was published by Academic Press in 1974.

I. Metzler's contribution to the business cycle literature centred on his integration of inventories into a dynamic model of income determination. Employing the income/expenditure multiplier/accelerator framework pioneered by Robertson, Keynes, Hicks, Lundberg, Samuelson and others, Metzler added a rigorous formulation of inventory behaviour against a backdrop of production lags and endogenously determined anticipation of sales.

His initial assumption, supported by his later empirical investigation (1948), was that the response of output to sales receipts was the longest of the three basic lags in the circular flow of income, of which the other two were the lag between the receipt of income and consumption spending and the lag between output and the distribution of earnings to factors of production. Featuring the output/sales lag, his classic study, ‘The Nature and Stability of Inventory Cycles’ (1941), demonstrated that any disturbance, such as an autonomous increase in investment, tended to produce cycles about the new level of income provided that the marginal propensity to consume is less than unity. The cycles are damped if businesses demand a constant level of inventories and expect sales in the current period to be unchanged at the level of the preceding period. Explosive cycles may occur if anticipated sales change when actual sales change and if firms try to maintain a constant ratio of inventories to anticipated sales.

If inventory demand varies with anticipated sales, a reduction in inventories below desired levels during an expansionary phase of the business cycle raises the demand for inventories and hence reinforces the rise in income and expenditures. The increases gradually taper off, causing a fall in planned inventory investment. Income and sales therefore fall and inventories rise about desired levels. The demand for inventories drops further and reinforces the decline in income. The model thus gives rise to predictions about the relative timing of movements in income, sales, and inventories, and of response coefficients for which the cyclical process will converge to a new equilibrium.

Investigations in the 30 years following the 1941 paper, including several of his own (1946; 1947b; 1973e), tended to support Metzler's initial formulation. The basic model was also enriched by the addition of monetary factors and the rate of interest, the price level, and a disaggregation of inventories into finished goods and goods in process (Zarnowitz, 1985, pp. 541–2).

II. Metzler's contribution to macro-monetary theory came from a single influential paper, ‘Wealth, Saving, and the Rate of Interest’ (1951b). Metzler wrote in the wake of the great debates between Keynes and his critics (Haberler, 1941; Pigou, 1943; 1947; Scitovszky, 1941), who had invoked a positive relation between real cash balances and expenditures on goods and services as the basis of achieving a stable macro equilibrium. Metzler, taking a broader view of wealth as including both real balances and financial claims to the capital stock, argued that the implied inverse wealth/saving relation introduced a monetary element in the determination of the interest rate. Whereas money was traditionally without any lasting influence on the real side of the ‘classical’ model, the presence of the wealth/saving relation meant that the exchange of money and securities (the prototype of which is an interest-bearing equity claim) between the central bank and the private sector altered the latter's perceived wealth total and hence its rate of saving and the equilibrium rate of interest.

In Metzler's analysis, an open-market purchase, for example, raises cash balances and removes securities from private portfolios. In the resulting adjustment process, real balances are reduced by the rise of the general price level, but securities are not restored. The consequent net wealth loss stimulates a greater rate of saving and a lower rate of interest in the post-purchase equilibrium.

Most commentators disputed Metzler's claim that in the classical tradition, monetary changes tended not to affect the equilibrium rate of interest (Haberler, 1952, p. 245; Patinkin, 1956, pp. 260–1). On the other hand, the early critics would probably have agreed that Metzler was the first to articulate a specific influence on the interest rate that sprang from open-market operations in a model containing the wealth/saving relation.

Later critics, notably Robert Mundell (1960), questioned the *direction* of that influence on the interest rate. Metzler had been careful to account for the disposition of the earnings on the securities acquired by

the central bank in the open-market purchase. In order to prevent private disposable income from falling continuously as these earnings are received by the bank in the future, the fiscal authority is assumed to reduce taxes by an equal amount (Metzler, 1951b, p. 109n). Mundell pointed out that if the taxes reduced are those on capitalizable income, the securities sold to the central bank are exactly replaced by an upward revaluation of remaining privately held securities. The wealth effect of the operations, taking account of the inflation-induced loss of real balances, is a ‘wash’. A reduction in taxes on capital, however, also raises the net return to capital and hence the investment demand schedule. In the new equilibrium, the rate of interest is higher.

Metzler (1973d, pp. 354–62) replied that capitalizable federal taxes are at most 30 per cent of total federal taxes. Proportionately, 70 per cent of a tax cut falls on non-capitalizable personal income. Only a small part of the value of securities sold to the bank is thus recovered by any likely tax cut, and the operation's wealth effects remain predominantly as Metzler described them.

David McCord Wright (1952) pointed out that the lower equilibrium interest rate generated by Metzler's open-market purchase will itself promote a more rapid investment rate, offsetting thereby the community's loss of securities and real wealth due to the purchase. Metzler (1952) objected that such an offset fell outside the limited time frame that his analysis and macrotheory generally were properly concerned with.

George Horwich (1962; 1964) argued that offsetting wealth changes due to forced saving tend to characterize the very process by which the new equilibrium interest rate is reached. While Wright saw a long-run wealth offset spurred by the wealth/saving-induced lower rate of interest, Horwich questioned whether the short-run adjustment creating a reduced equilibrium interest rate was viable. His characterization of the adjustment process (influenced by Metzler's account of the securities markets underlying the model) emphasized the equilibrating role of new or flow security supply and demand originating in new investment and saving, respectively. The excess of investment over saving created by the operation is thus, in its financial counterpart, an excess supply of new securities that directly moves the interest rate towards its new equilibrium and funds additional investment spending. If, through forced saving, the excess investment is realized in additional real capital, the excess new securities tend to replace those sold to the bank. The process of reaching any post-operation equilibrium is thus one in which additional security issues and increments of capital stock are necessarily involved and tend, more or less, to maintain the pre-operation level of wealth.

Niehans (1978) questioned Metzler's specification of the capital stock as a determinant of saving, while real balances, which are at desired levels in equilibrium, are not so specified. Both wealth components, Niehans argued (1978, pp. 91–2), should be at desired (optimum) levels in equilibrium and should not appear in individual demand functions. He saw the main contribution of the ‘wealth’ article in its elegant formulation of the neoclassical synthesis (see also Haberler, 1952, p. 246 for this viewpoint), in the distinction it made in the differing impacts of the various types of monetary change, and in its broad influence on the methodology employed by the major monetary writers of the next quarter century.

III. At least half of Metzler's papers are related to the field of international trade. He is probably best known for papers on tariff theory, international macroeconomics, and the transfer problem, but other work includes a lucid survey (1949a), a discussion of difficulties of applying purchasing power parity in post-Second World War exchange rate realignments (1947a), and a discussion of the views of Frank Graham (1950a).

In tariff theory, Metzler's contribution has largely been summarized in the statement that, in a two-

country, two-good world, a tariff can fail to be protective in the sense that it can lead to a reduction in the domestic relative price of the import-competing good. This is the so-called ‘Metzler paradox’. It is ironic that Metzler himself described this result as well known, and viewed his own contribution as analyzing the implications for income distribution of such a non-protective tariff (1949b, p. 10). Metzler's papers on this topic were apparently motivated by pronouncements of Australian economists, and the two papers (1949b; 1949c) include discussion of alternative assumptions about expenditure of tariff revenue – by government (non-dutiable) or the private sector (dutiable), with various marginal propensities to consume, and from a zero or non-zero initial tariff. However the cleanest and best-known result comes with zero initial tariff and with tariff revenue implicitly going to the private sector as an increase in disposable income. Metzler shows that the domestic relative price will not change if the elasticity of import demand in the foreign country is equal to 1•minus the marginal propensity to import in the home (tariff-levying) country.

This foregoing result is easily understood by considering the world market for the home importable. The imposition of a trade tax, at constant domestic relative price, will imply a vertical shift of the foreign export supply curve (as in an elementary tax-incidence problem), since foreign export supply is a function of world relative price. If the foreign offer curve is elastic, this implies decreased export supply by the foreign country at the fixed domestic price, but if the foreign offer curve is inelastic (implying that the export supply curve is backward-bending), then the result is increased export supply at a given domestic price. On the demand side, a fixed domestic price implies a lower world relative price of the home importable. Thus the home country is better off, via the improvement in its terms of trade. If the importable is a normal good, the improved real income in the home country implies a rightward shift of the home import demand curve. In a Walrasian-stable market, the domestic price falls if and only if the shifts in home import demand and foreign export supply combine to yield excess supply at the initial domestic price. Metzler's condition is a requirement that the shifts in the two curves exactly offset each other. Thus, if the importable good is normal at home, an inelastic foreign offer curve is necessary, but not sufficient, for the Metzler paradox.

The subsequent literature has enshrined this simplest version of the non-protective tariff (despite at least one attempt to refute the result). It is now understood that, given the normality assumption, the Metzler paradox is inconsistent with the home country levying the so-called optimal tariff; thus the ‘paradox’ is just one of many possible consequences of a second-best situation (from a myopic national viewpoint).

Another result to bear Metzler's name is the Laursen–Metzler effect, in honour of the joint authors (1950). Laursen and Metzler were concerned with integrating models of flexible exchange rates which focused either on income–expenditure effects or on terms of trade effects, but not on both simultaneously. They posited a channel from devaluation through the terms of trade and onto expenditure; specifically, a deterioration in the terms of trade was assumed to lead to increased expenditure with given nominal income. This was then applied to discussion of the extent of insulation via flexible exchange rates, and of the ‘acceptability’ of exchange rate changes in certain policy scenarios. The Laursen–Metzler effect has been integrated into the literature, although it was eclipsed in periods where there was extreme emphasis on flexible product prices (thus weakening the link from exchange rate changes to changes in the terms of trade) and although there is some question as to the sign of the effect. In a period of current-account and government-budget imbalances, there has been some emphasis on real intertemporal models of trade. With more sophisticated models of simultaneous intertemporal and intratemporal optimization than were available to Laursen and Metzler, a deeper

understanding of the link between intratemporal terms of trade and current expenditure is now possible. Many of Metzler's international papers involved the transfer problem; see Metzler (1942; 1951a; 1951c; 1973b; 1973c). Metzler's focus was on endogenous income and expenditure effects, holding prices and interest rates fixed. In later papers, the analysis is tied to Metzler's contributions to the applications of matrix theory to economics. In the transfer problem, since the initial transfer is a pure redistribution, the analytic question concerns what changes in endogenous variables are required to re-establish equilibrium. The pure trade literature has emphasized the endogenous adjustment of the terms of trade in real, flexible-price models – including a somewhat incestuous literature, mainly involving Samuelson and Jones, on the likelihood of orthodox or anti-orthodox bias. More recently, the possibility of ‘paradox’ in a multi-country setting has been the central topic. Metzler's assumption of constant prices led to analysis of the impact on trade balances and income at home and abroad, with discussion of the role of stability conditions nationally and globally and of the relevant roles for alternative income concepts in the presence of imported inputs for production. Chapter 4 of the *Collected Papers* (1973c) comes closest to linking up to the orthodox theory. While Metzler was one of many contributors to the transfer literature, his strong Keynesian perspective may have limited the long-term importance of his contribution more on this topic than on those mentioned earlier.

IV. In the field of mathematical economics, Metzler has been honoured by having a matrix named after him. The central paper is perhaps Metzler (1945), but see also Metzler (1950b; 1951a; 1951c). The Metzler matrix is a square matrix with positive diagonal elements, negative off-diagonal elements, positive principal minors and determinant, and a positive inverse matrix. Metzler investigated this class of matrices in the context of market stability (1945) and comparative statics (1950b). The stability analysis linked the Hicksian concept of market stability, which can be interpreted as essentially static, and Samuelson's explicitly dynamic approach to stability.

Metzler showed that if multiple markets are stable for any (relative) speeds of adjustment, then they must satisfy Hicks's concept of perfect stability. Perfect stability says that a fall in price in any single market creates excess demand in that market – after any subset of other prices is adjusted to clear the ‘own’ markets – and all other prices are held fixed. Metzler's proof revolved about the alternating sign of principal minors of a matrix of partial derivatives of excess demands with respect to prices, the negative of this matrix leading to a Metzler matrix. Metzler also showed, by counterexample, that Hicksian perfect stability does not imply Samuelsonian dynamic stability. Another Metzler result showed that in the presence of gross substitutability, dynamic stability and perfect stability are equivalent. Gross substitutability guarantees that the matrix of partial derivatives has negative diagonal terms and positive off-diagonal terms, so that its negative has the sign pattern of a Metzler matrix. The intuition of the gross substitute case is that the impact of a change in ‘own price’ on excess demand for a good exceeds the aggregate impact of all ‘other price’ changes; thus, in a sense, the system generalizes the intuition of single-market stability analysis. While cross-effects can exist, the own effects dominate in each market. Metzler applied this theory to the comparative statics of fixed-coefficient regional models, multicountry income transfers, and taxes and subsidies in fixed-coefficient models. As is better understood after integrative work on matrices with dominant diagonals (McKenzie, 1960) and on P-matrices (Gale and Nikaido, 1965), strong results in ‘square’ – that is, n x n systems – usually require strong assumptions closely related to the existence of appropriate Metzler matrices. While there were many other contributors in this area, for example Hawkins and Simon, and while the majority of key results were already known to mathematicians, Metzler's work provided a crucial synthesis of stability literature and

an important step in the evolution of matrix theory as applied in economics.

Selected works

1941. The nature and stability of inventory cycles. *Review of Economics and Statistics* 23(3), 113–29.

1942. The transfer problem reconsidered. *Journal of Political Economy* 50, 397–414.

1945. Stability of multiple markets: the Hicks conditions. *Econometrica* 13(4), 277–92.

1946. Business cycles and the theory of employment. *American Economic Review* 36, 278–91.

1947a. Exchange rates and the International Monetary Fund. In *International Monetary Policies*, Postwar Economic Studies No. 7, ed. L.A. Metzler, R. Triffin and G. Haberler. Washington, DC: Board of Governors of the Federal Reserve System.

1947b. Factors governing the length of inventory cycles. *Review of Economics and Statistics* 29(1), 1–15.

1948. Three lags in the circular flow of income. In *Income, Employment, and Public Policy: Essays in Honor of Alvin H. Hansen*, ed. L.A. Metzler. New York: W.W. Norton.

1949a. The theory of international trade. In *A Survey of Contemporary Economics*, ed. H.S. Ellis. Philadelphia: Blakiston.

1949b. Tariffs, the terms of trade, and the distribution of national income. *Journal of Political Economy* 57, 1–29.

1949c. Tariffs, international demand, and domestic prices. *Journal of Political Economy* 57, 345–51.

1950a. Graham's theory of international values. *American Economic Review* 40, 301–22.

1950b. A multiple-region theory of income and trade. *Econometrica*18(4), 329–54.

1950. (With S. Laursen.) Flexible exchange rates and the theory of employment. *Review of Economics and Statistics* 32(4), 281–99.

1951a. A multiple-country theory of income transfers. *Journal of Political Economy* 59, 14–29.

1951b. Wealth, saving, and the rate of interest. *Journal of Political Economy* 59, 93–116.

1951c. Taxes and subsidies in Leontief's input–output model. *Quarterly Journal of Economics* 65, 433–

38.

1952. A reply. *Journal of Political Economy* 60, 249–52.

1973a. *Collected Papers*. Cambridge, MA: Harvard University Press.

1973b. Imported raw materials, the transfer problem, and the concepts of income. In Metzler (1973a).

1973c. Flexible exchange rates, the transfer problem, and the balance-budget theorem. In Metzler (1973a).

1973d. The structure of taxes, open-market operations, and the rate of interest. In Metzler (1973a).

1973e. Partial adjustment and the stability of inventory cycles. In Metzler (1973a).

Bibliography

Gale, D. and Nikaido, H. 1965. The Jacobian matrix and the global univalence of mappings. *Mathematische Annalen* 159(2), 81–93.

Haberler, G. 1941. *Prosperity and Depression*. 3rd edn. Geneva: League of Nations.

Haberler, G. 1952. The Pigou effect once more. *Journal of Political Economy* 60, 240–6.

Horwich, G. 1962. Real assets and the theory of interest. *Journal of Political Economy* 70, 157–69.

Horwich, G. 1964. *Money, Capital, and Prices*. Homewood, IL: R.D. Irwin.

Horwich, G. and Samuelson, P.A., eds. 1974. *Trade, Stability, and Macroeconomics: Essays in Honor of Lloyd A. Metzler*. New York: Academic Press.

McKenzie, L. 1960. Matrices with dominant diagonals and economic theory. In *Mathematical Methods in the Social Sciences 1959*, ed. K.J. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

Mundell, R.A. 1960. The public debt, corporate income taxes, and the rate of interest. *Journal of Political Economy* 68, 622–26.

Niehans, J. 1978. Metzler, wealth, and macroeconomics: a review. *Journal of Economic Literature* 16, 84–95.

Patinkin, D. 1956. *Money, Interest and Prices: An Integration of Monetary and Value Theory*. Evanston: Row, Peterson.

Pigou, A.C. 1943. The classical stationary state. *Economic Journal* 53, 343–51.

Pigou, A.C. 1947. Economic progress in a stable environment. *Economica*, NS 14(55), 180–8.

Scitovszky, T. 1941. Capital accumulation, employment, and price rigidity. *Review of Economic Studies* 8(2), 69–88.

Wright, D.M. 1952. Professor Metzler and the rate of interest. *Journal of Political Economy* 60, 247–9.

Zarnowitz, V. 1985. Recent work on business cycles in historical perspective: a review of theories and evidence. *Journal of Economic Literature* 23, 523–80.

How to cite this article

Horwich, George and John Pomery. "Metzler, Lloyd Appleton (1913–1980)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000157> doi:10.1057/9780230226203.1094

The New Palgrave Dictionary of Economics Online

micro-credit

Jonathan Morduch
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Most providers of micro-credit lend without requiring collateral. In doing so, they can provide poor households with access to small-scale loans to expand household businesses and meet consumption needs. Micro-credit institutions demonstrate that a combination of mechanisms can overcome the market imperfections created when banks lack good information about borrowers and when borrowers lack collateral. Micro-credit innovations are of both theoretical interest and practical importance. Proponents argue that micro-credit can be a tool to reduce poverty and, in the best cases, can operate profitably and on a large scale, free of public subsidy.

Keywords

adverse selection; behavioural economics; capital access; collateral; collusion; credit rationing; Grameen Bank; group lending; information revelation; joint liability; micro-credit; micro-finance; moral hazard; Muhammad Yunus; poverty alleviation

Article

Micro-credit encompasses a broad movement to supply professional banking services to poor households; micro-credit innovations offer new insights into the economics of information and new mechanisms for reducing poverty.

Karl Marx (1867) famously tied inequality in access to capital to broader social and economic inequalities driven by markets; micro-credit presents the promise that market mechanisms may instead help to broaden capital access.

The economics of information shows why poor customers are usually shunned by commercial lenders. Customers typically lack the assets and ownership documents that banks require as collateral, and banks lack cost-effective ways to monitor and enforce contracts. Theorists demonstrate how credit rationing can emerge in these contexts, with adverse selection and moral hazard as culprits. The challenge for

banks is exacerbated by the small size of transactions. The *MicroBanking Bulletin*'s (2006) survey of 302 leading micro-credit institutions, for example, found that the average loan balance was 436 dollars for the median ‘micro-bank’. For the median micro-bank focusing on poorer customers, the average loan balance was just 109 dollars. These amounts tend to fall below the threshold of interest for large commercial banks, even in low-income economies. Hence the poor lose twice: they begin with less income and fewer assets than others, and, as a result, they have worse access to the financial institutions that might offer a route away from poverty. To this extent, poverty reinforces poverty.

Micro-credit is part of an approach that aims to undo this equation. Despite the challenges, providers of micro-credit aim to deliver reliable and reasonably priced financial services to the under-served, and most institutions aim to do so without ongoing subsidies. While loans are relatively small, advocates argue that the funds are sufficient to finance small businesses and cover emergency consumption needs – and thus to contribute meaningfully to poverty reduction.

Early micro-credit successes were realized in Bangladesh, Indonesia and Bolivia, gaining global attention in the 1980s. By the end of 2005, one annual survey counted over 3,000 institutions, collectively serving 113 million customers worldwide (Daley-Harris, 2006). Of these, 82 million customers were classified as being among the ‘poorest’, and 84 per cent of those were women. Rough estimates place unmet demand at over one billion people. The 2006 Nobel Peace Prize to Muhammad Yunus and the Grameen Bank of Bangladesh recognized the potential of micro-credit to reduce global poverty, though Yunus's boldest claims about the potential scale and impact of micro-credit remain untested with reliable data.

The *MicroBanking Bulletin* (2006) survey shows both the promise and the challenges of micro-credit. The survey, which is skewed towards institutions with strong commitments to financial self-sufficiency, finds that 69 per cent of the 302 institutions were earning profits in 2004, and just two per cent of loan portfolios were deemed ‘at risk’ as a result of loan payments left unpaid beyond 30 days. Real interest rates on loans average about 25–35 per cent per year in the survey, though some top 90 per cent per year. The encouragement of profit and the tolerance of relatively high interest rates are central to the logic of micro-credit policy. Escaping reliance on subsidy, it has been argued, allows institutions to expand beyond the constraints imposed by donors’ purses, creating the prospect of a truly global market-based industry. Despite innovations, though, institutions focused on the poorest customers face stiff challenges in generating profits. The *MicroBanking Bulletin* (2006) survey shows that the median micro-bank serving the poorest customers faces almost twice the cost of lending (per unit of assets) compared with the median micro-bank serving better-off (but still low-income) customers. The extent of trade-offs between meeting profit targets and achieving social objectives remains largely unexplored, as does the nature of productivity-enhancing roles for subsidy (Armendáriz de Aghion and Morduch, 2005, ch. 9).

Group lending

The high rates of loan repayment are attributed to innovative loan contracts, most notably the ‘group lending’ contract. The group approach is associated with the Grameen Bank, although it has been employed more faithfully by others. The Grameen approach begins with the bank entering a village and inviting villagers to form themselves into five-person groups. A cluster of groups is then formed into a centre that meets once a week in the village, where all business is transacted by a loan officer from the

bank. Loans are given to individuals, but the group is deployed to improve incentives and provide a support network. As long as all loans are repaid on time, loans continue to be made to group members, but if any group member cannot repay his or her loan (and the four others cannot fix the problem themselves), the entire group is excluded from future borrowing. This element of the contract is often referred to as creating ‘joint liability’ – even though, in the Grameen model at least, individuals are not explicitly liable for the repayment of fellow group members.

The contract addresses moral hazard by giving borrowers incentives to monitor each other and to sanction members whose lack of effort jeopardizes loan repayments. The customers often have advantages in these activities, which stem from living and working alongside each other and from being able to employ ‘social’ sanctions that the bank cannot use. The contracts may also foster mutual support mechanisms that provide insurance and other assistance, a point stressed by Muhammad Yunus, Grameen's founder. Early theoretical analyses on moral hazard in group lending include Stiglitz (1990), while Besley and Coate (1995) raise the possibility of collusive behaviour by borrowers.

The contracts, in principle, can also address adverse selection (and the inefficiencies created by the withdrawal of safer borrowers in markets with asymmetric information; Stiglitz and Weiss, 1981). Adverse selection in credit markets arises because banks cannot distinguish between potential customers who are likely to reliably repay loans and those that will not. Without such information, the bank must charge all customers the same interest rate, and the safer borrowers implicitly subsidize the riskier ones. In principle, the process of group formation can improve outcomes by screening risky borrowers and matching safer individuals with other relatively safe individuals. Because the effective cost of the loan depends in part on the probability that one's fellow group members will default, safer individuals will then face lower effective borrowing costs than riskier individuals – even when all individuals face identical nominal contracts; the contract combined with the sorting process reduces the extent of cross-subsidization and thus adverse selection (Ghatak, 1999; see also references in Armendáriz de Aghion and Morduch, 2005, ch. 3). This mechanism has received little empirical verification, though, and in practice lenders devote substantial resources to information acquisition.

Beyond group lending

The use of groups has clear attractions but has proved cumbersome when customers have diverse needs and growth prospects. It also relies on the willingness and ability of customers to carry out monitoring and enforcement tasks that are usually the responsibility of bankers. Rai and Sjöström (2004) point to inefficiencies in group lending that can be mitigated through simple information revelation mechanisms, and, as noted above, collusion remains a theoretical possibility. A push to move beyond group lending with joint liability reached an important milestone in practice when two early pioneers, BancoSol of Bolivia and the Grameen Bank, independently abandoned group lending with joint liability as the basis of their operations.

The move beyond group lending highlights other contract innovations that have been overshadowed. Among the most important is the repayment schedule. In Grameen Bank loan contracts, for example, loans are repaid in small increments weekly over the course of several months to a year. It is an odd structure for loans that are ostensibly for business investments that may take time to bear fruit. The schedule, though, allows households to repay loans from other income sources in small, manageable

increments. In this way, loans can often be repaid even if businesses fail. Perhaps more important, the structure allows households to easily use loans to finance consumption, strengthening the ability to cope with health crises, pay school fees and keep food on the table. The extent to which such ‘diversion’ occurs, and its costs and benefits, has yet to receive much research attention, but it may hold a key to new directions for micro-credit.

A second important mechanism is the use of long-term lending relationships. Lenders gain information and instill incentives for loan repayment by repeatedly interacting with customers, allowing borrowers to start with small loans and become eligible for steadily larger loans with each successful cycle.

From micro-credit to ‘micro-finance’

Most of the evidence in favour of micro-credit is anecdotal, though rigorous empirical studies are accumulating (Armendáriz de Aghion and Morduch, 2005, ch. 8). In data from Mexico, for example, McKenzie and Woodruff (2006) find returns to capital of above 20 per cent per month for small-scale businesses with capital stocks below 200 dollars. As capital stocks rise above 400 dollars, estimated returns to capital fall to around five per cent per month. These returns are still substantial and help to explain the ability to pay relatively high interest rates.

The returns pose a puzzle, though. There are no signs of poverty traps, and if returns are so high, why have households been unable to save more on their own, overcoming credit gaps through self-finance? With the realization that customers indeed seek better ways to save and insure (and seek credit for a wide range of uses), micro-banks have started expanding their services. The next wave of innovations focuses there, and draws in part on insights from behavioural economics (for example, Ashraf, Karlan, and Yin, 2006). The focus will thus continue to shift from ‘micro-credit’ to ‘micro-finance’ more broadly.

See Also

- adverse selection
- credit rationing
- development economics
- moral hazard
- poverty alleviation programmes
- poverty traps

Bibliography

Armendáriz de Aghion, B. and Morduch, J. 2005. *The Economics of Microfinance*. Cambridge, MA: MIT Press.

Ashraf, N., Karlan, D. and Yin, W. 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. *Quarterly Journal of Economics* 121, 635–72.

Beck, T., Demirguc-Kunt, A. and Martinez Peria, M.S. 2006. Banking services for everyone? Barriers to bank access and use around the world. World Bank Policy Research Working Paper 4079. Washington, DC: World Bank.

Besley, T. and Coate, S. 1995. Group lending, repayment incentives, and social collateral. *Journal of Development Economics* 46, 1–18.

Daley-Harris, S. 2006. *State of the Microcredit Summit Campaign Report 2006*. Washington, DC: Microcredit Summit.

Ghatak, M. 1999. Group lending, local information and peer selection. *Journal of Development Economics* 60, 27–50.

Marx, K. 1867. *Capital*, vol. 1, tr. Ben Fowkes. Harmondsworth: Penguin, 1990.

McKenzie, D. and Woodruff, C. 2006. Do entry costs provide an empirical basis for poverty traps? Evidence from Mexican microenterprises. *Economic Development and Cultural Change* 55, 3–42.

Microbanking Bulletin 12. 2006. Washington, DC: Microfinance Information Exchange.

Rai, A. and Sjöström, T. 2004. Is Grameen lending effcient? Repayment incentives and insurance in village economies. *Review of Economic Studies* 71, 217–34.

Stiglitz, J. 1990. Peer monitoring and credit markets. *World Bank Economic Review* 4, 351–66.

Stiglitz, J. and Weiss, A. 1981. Credit markets with imperfect information. *American Economic Review* 71, 393–410.

How to cite this article

Morduch, Jonathan. "micro-credit." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000348> doi:10.1057/9780230226203.1095

The New Palgrave Dictionary of Economics Online

microfoundations

Maarten C.W. Janssen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The microfoundations literature has attempted to bridge the gap between microeconomic and macroeconomic models. Many models in this literature have used the theoretical construct of a representative agent. Economy-wide outcomes are thereby presented as if they were the result of the optimizing behaviour of one individual. Emergent properties at the macro level are by construction precluded from the analysis. Other literatures exist where emergent properties are taken to be at the heart of the quest for microfoundations.

Keywords

aggregation (theory); Austrian economics; behavioural macroeconomics; bounded rationality; classical economics; coordination failure; effective demand; efficiency wages; evolution; expectations; fairness; fixprice models; frictions; game theory; general equilibrium; Hayek, F. von; imperfect competition; involuntary unemployment; IS–LM; Keynesianism; labour supply; learning; Lucas, R.; market clearing; menu cost; methodological individualism; microfoundations; multiplier; neoclassical synthesis; New Classical economics; New Keynesian macroeconomics; other-regarding preferences; overlapping generations model of general equilibrium; Phillips curve; post-Keynesian economics; residential segregation; rational behaviour; rational expectations; real business cycles; reciprocity; representative agents; spontaneous order; sticky prices; strategic complementarities; technology shocks; Vienna Circle

Article

The quest to understand microfoundations is an effort to understand aggregate economic phenomena in terms of the behaviour of individual economic entities and their interactions. These interactions can involve both market and non-market interactions. The quest for microfoundations grew out of the widely felt, but rarely explicitly stated, desire to stick to the position of methodological individualism (see Agassi, 1960; 1975; Brodbeck, 1958), and also out of the growing uneasiness among economists in the

late 1950s and 1960s with the coexistence of two sub-disciplines – namely, microeconomics and macroeconomics – both aiming to explain features of the economy as a whole. Methodological individualism is the view that proper explanations in the social sciences are those that are grounded in individual motivations and their behaviour. The urge to make microeconomics and macroeconomics compatible can be understood from the perspective of the unity-of-science discussion initiated by the Vienna Circle in the philosophy of science in the beginning of the 20th century (see Nelson, 1984).

Efforts to understand microfoundations go far beyond the questions that lie at the heart of formal aggregation theory, that is, the analysis of how to map aggregate economic variables and relationships back to similar individual variables and relationships that underlie them. One crucial issue in the microfoundations literature is the extent to which aggregate economic variables and/or relationships exhibit features that are similar to the features of individual variables and/or relationships, and in particular whether certain features are emergent properties at the macro level that do not have a natural counterpart at the individual level. An important early example of emergence is Schelling's analysis (1978) of segregation. He shows that segregation in neighbourhoods may be an emergent property at the micro level that can be viewed as an unintended consequence of the individual decisions concerning where to live.

The discussion on emergence shows that there is no reason to assume or expect macro behaviour to be in any way similar or analogous to the behaviour of individual units. In order to have ‘proper’ microfoundations in line with methodological individualism, it is thus by no means required that aggregate outcomes are represented as if they were the outcome of a single agent's decision problem. On the contrary, the restriction to single individual decision problems found in modern macroeconomics is self-imposed and not implied by the methodological position of methodological individualism (see Kirman, 1989). In fact, one may argue that the interaction between different, and possibly heterogeneous, individual units should be at the core of macroeconomic analysis.

As the quest for ‘proper’ microfoundations has arisen in the debate concerning the microfoundations for macroeconomics, this article's main focus is on this debate. The article starts with a historical perspective on this debate and continues to discuss New Classical and New Keynesian approaches to macroeconomics that emerged out of the microfoundations debate. The role of equilibrium notions and expectations is discussed in a separate section. The article argues that the microfoundations for macroeconomics literature is best understood from the perspective of attempting to make microeconomics and macroeconomics compatible with each other. The article closes with a discussion of non-mainstream approaches to microfoundations and more recent approaches to microfoundations using the perspective of evolutionary forces and boundedly rational behaviour.

Historical background to the microfoundations for macroeconomics debate

Around the mid-1950s two more or less separate approaches existed to studying economy-wide phenomena: general equilibrium theory and (Keynesian) macroeconomics. Some of the more important theoretical issues within each of these approaches were settled. Existence of a general equilibrium point was proved by Arrow and Debreu (1954) and the macroeconomic IS–LM framework was well established (following the seminal paper by Hicks, 1937). Of course, some other issues were still to be tackled, such as questions related to how to deal with imperfect competition, incomplete markets and/or

overlapping generations.

Both approaches explained economy-wide phenomena, but there were important differences between the perspectives from which they started. Flexible prices and market-clearing were at the core of general equilibrium theory; involuntary unemployment and effective demand were important concepts in macroeconomics. The neoclassical synthesis reconciled general equilibrium theory and (Keynesian) macroeconomics by giving each of them its own domain of applicability: macroeconomics (with its assumption of sticky money wages) gives an accurate description of the economy in the short run, while long-run developments of the economy were considered to be adequately described by the general equilibrium approach.

From a theoretical point of view this state of affairs was unsatisfactory. One cannot simply attribute unemployment to sticky money wages while leaving the theoretical structure of general equilibrium theory intact: the imposition of a fixed money wage (or, more generally, fixed prices) deeply affects the theory of supply and demand. It was natural, then, to inquire into the relationship between the two approaches, especially given that they study the same phenomena. In addition, the generally accepted view was that it is the market interaction between many individual agents from which economy-wide phenomena result, implying that general equilibrium theory is the more fundamental theory of the two. The quest for microfoundations was born.

The rise of interest in microfoundations can also be at least party conceived as being driven by the perceived failings of important elements of empirical macroeconomics and in particular the fact that the Phillips curve turned out to be not a stable relationship that can be used for economic policy purposes (see, for example, Friedman, 1968). Several essays in Phelps (1970) are written to reconcile microeconomic theory with the apparent temporary trade-off between wages and unemployment embodied in the new interpretation of the Phillips curve.

New Classical and New Keynesian economics

One key controversy in the quest for microfoundations is how to explain the widely observed phenomenon of unemployment. From a market-clearing perspective, unemployment simply means that at the current (real) wage rate people do not want to supply more labour to the market. If there is registered unemployment, it is thus either of a ‘voluntary’ nature or a short-run phenomenon that quickly disappears. In this vein, Lucas (1978, p. 354) argued that involuntary unemployment is not a fact that needs to be explained, but rather a theoretical construct Keynes introduced in the hope it would be helpful in explaining fluctuations in measured unemployment.

In line with these ideas, New Classical economists have attempted to reconcile macroeconomic phenomena such as inflation and unemployment, and the empirical observed trade-off between the two measured by the Phillips curve, with a Walrasian notion of market clearing. Early models, such as Lucas and Rapping (1969) and Lucas (1972), stressed the idea that incomplete information about the money supply may cause business fluctuations. Later real business cycle models (such as that of Kydland and Prescott, 1982) looked at technology shocks to explain cyclical behaviour. Thus, an important difference between the Lucas–Rapping approach and early real business cycle models is that the former, but not the latter, introduces frictions to explain business cycles. With these New Classical models, the concept of the representative agent (consumer, firm or producer/consumer agent) became widely used in modern

macroeconomics. In its most extreme form, the economy as a whole is represented as if it were the outcome of a single individual's decision problem. The possible differences between individual and aggregate economic behaviour are thereby assumed away.

Economists who were oriented towards Keynesian ideas thought that there is an involuntary, non-transient component in observed unemployment figures. Many New Keynesian contributions therefore try to reconcile the notion of involuntary unemployment with a notion of market equilibrium.

A first approach considers the question how to incorporate the notion of price stickiness, especially concerning money wages, with the traditional theory of demand and supply. This issue was first studied by Clower (1965). He emphasized that, because of the interdependence of markets, demand and supply curves on all markets are affected if money wages are fixed. If prices are restrained from bringing about market clearing allocations, then other variables have to bring about some kind of fixed-price equilibrium. Clower (1965) and Leijonhufvud (1968) set out a research programme studying the existence of fixed-price equilibria and their properties. The resulting equilibrium notion and the properties of such fixprice equilibria were formulated by Barro and Grossman (1971), Drèze (1975) and Benassy (1975), among others. The idea of this literature is that agents express their demands on the basis of market prices and perceived quantity constraints. These models have microfoundations in the sense that they are based on decision-making individuals and a notion of equilibrium. Moreover, it turned out that the fixprice models capture quite a number of ideas associated with Keynesian economics. By means of these alternative equilibrium notions, involuntary unemployment could be regarded as an equilibrium phenomenon in which optimizing households face a quantity constraint on the amount of labour they can supply. Also, the Keynesian notions of effective demand and the multiplier were reformulated within the new models. Finally, the models provided arguments for demand policies by the government. Of course, from a market-clearing perspective, these fixprice models are unsatisfactory as they do not explain why (rational) individuals do not propose changes to the terms of trade at which they exchange. Clearly, if prices are fixed at no market clearing levels, some agents in the economy can mutually benefit by exchanging at different prices, and therefore have an incentive to propose changes in prices. A literature on small menu cost appeared arguing that introducing a very small cost for economic agents to change prices may result in large fluctuations in aggregate output (see Mankiw, 1985).

Another approach New Keynesian economists followed is to incorporate the literature on imperfect competition into macroeconomic models. Hart (1982), Blanchard and Kiyotaki (1987), Kiyotaki (1988) and d'Aspremont, Dos Santos Ferreira and Gérard-Varet (1990) are among the pioneering articles in this area. These models can explain why aggregate output is below the optimal full employment output level. Unemployment can be involuntary when there is imperfect competition in the labour market.

A third approach to explaining non-competitive wages is to introduce some type of informational problem, as in the literature on efficiency wages. The basic idea of this literature is that the average labour productivity is positively related to the wage a firm offers. Firms may set wages above the competitive level in order to induce employees to work harder, and therefore may be unwilling to lower their wage offers (see Yellen, 1984; Lindbeck and Snower, 1987).

Yet another approach relies on coordination failures formally analysed in terms of multiple equilibria (see Bryant, 1983; Roberts, 1987). Cooper and John (1988) point out that many New Keynesian models are based on strategic complementarities between agents’ actions, that is, these models do not rely on an

assumption that prices cannot adjust to their market equilibrium values. When strategic complementarity exists, there may be multiple equilibria that can be Pareto-ranked. Agents may then find themselves in a ‘bad’ equilibrium, but individually they cannot benefit by deviating to another choice. The authors call this a ‘coordination failure’.

There is a parallel between the coordination failures literature and the overlapping generations general equilibrium literature (see, for example, Geanakoplos and Polemarchakis, 1986). The latter literature views the economy as a process without definite end, such that what happens today is underdetermined as it depends on what people expect to happen tomorrow, which in turn depends on what people expect to happen the day after tomorrow, and so on. In such a world there is a continuum of equilibria. Geanakoplos and Polemarchakis (1986) show that, depending on how this indeterminacy is solved, that is, which variables are chosen to be exogenously determined, classical or Keynesian-oriented conclusions may be derived.

Work on all these different models has resulted in a shared methodology of how to go about building macroeconomic models. The traditional distinction in macroeconomics between Keynesian and classical economists is disappearing and a common methodology is surfacing. Economists share the understanding that the ultimate question that matters is how well markets function. The differences in importance attached to various market frictions are more a matter of degree than of fundamental divergence between different methodologies. The nature of what used to be macroeconomic theory has undergone dramatic changes alongside these developments. Traditional macroeconomic issues such as how to explain the business cycle or how to account for inflation are now studied with the same tools and techniques as those that are used in microeconomics. Along these lines, and by using the assumption of the representative agent, modern macroeconomics has assumed away the heterogeneity that may exist at the individual level. Lucas's prediction that we may soon simply speak of economic theory instead of separate microeconomic and macroeconomic theories has turned out to be fairly accurate (see Lucas, 1987, pp. 107–8). Somewhat paradoxically, one may say that the modern economist who still is a ‘hard line microeconomist’ is now called a macroeconomist.

Rationality, equilibrium and expectations

The efforts to create microfoundations for macroeconomics have resulted in a more unified approach to doing economic theory. The approaches discussed so far (also Keynesian-oriented models) all postulate rational behaviour on the part of economic agents and some notion of equilibrium. If expectations are important, it is postulated that agents’ expectations concerning important variables coincide with the model's predicted values concerning these same variables. This assumption concerning agents’ expectations have been termed ‘rational expectations’ (see Muth, 1961).

Parallel to the microfoundations literature, a literature questioning the eductive justifications for the notions of equilibrium and rational expectations emerged. This literature on the foundations of game theory basically argued that, if we assume that agents (players) are rational and that their rationality and the model (game) in which they operate are common knowledge, then it is not implied that these agents will play according to an equilibrium of the game. Fundamental papers in this respect are Bernheim (1984) and Pearce (1984), among others. These and other papers show that a much weaker notion, named (correlated) rationalizability, can be derived from assumptions regarding common knowledge of

the rationality of players.

On the basis of this literature, Guesnerie (1992) argues that rational expectations should be regarded as an equilibrium notion that is also not solely based on postulates regarding the rational behaviour of individual players. It is rational for individual players to have ‘rational expectations’ if other players have these very same ‘rational expectations’, but not necessarily otherwise. As the notion of rational expectations is essentially an equilibrium or consistency notion, it suffers from the same drawbacks in that it is not implied by the individual rationality assumptions that players will form rational expectations.

Another literature (see, for example, Bray and Savin, 1986, and several essays in Frydman and Phelps, 1983) studies the question whether in a decentralized economy economic agents may learn over time to have expectations that are consistent with those that are assumed by the rational expectations hypothesis. The general conclusion of this literature is that, due to the feedback from expectations to economic behaviour, the outcomes of an economic model with learning agents do not converge to the rational expectations solution.

It then follows that the microfoundations literature mentioned so far has not really succeeded in deriving all macroeconomic propositions from fundamental hypotheses on the behaviour of individual agents. The requirements of methodological individualism have thus not been satisfied by the microfoundations literature, which has predominantly presumed that individuals behave rationally (see Janssen, 1993).

Non-mainstream approaches to the microfoundations of macroeconomics

Apart from a long-lasting debate in the mainstream literature, the term ‘microfoundations’ has also stimulated work by other economists, and they have publicized their views on the relation between microeconomics and macroeconomics. Horwitz (2000) provides an overview of the Austrian perspective where individual knowledge, prices as conveyers of information, and subjective evaluations play important roles. The essays in Hayek (1948) and his views on spontaneous order are especially important in this respect. It may seem, then, that macroeconomics is not an important term in the Austrian vocabulary. However, this is only partly true. From an Austrian perspective an important question is what kind of monetary system will most likely preserve the communicative function of prices. Austrian economists have, as Horwitz shows, addressed such issues in a way that is compatible with methodological individualism.

A post-Keynesian view of the economy holds that long-term expectations are largely determined by non-economic processes such as those determined by mass psychology. These expectations therefore should be regarded as exogenous to the economic model, rather than as endogenously determined as in the case of rational expectations. Interestingly, this post-Keynesian view comes close to the result that is established by Geanakoplos and Polemarchakis (1986) in their overlapping generations general equilibrium model, where they show that indeterminacy of equilibria implies that expectations concerning future market outcomes may be chosen exogenously. Important investment decisions are, according to post-Keynesian economists, by their nature long-term decisions, and these decisions are thus largely determined by the state of these long-term expectations. This fundamental uncertainty requires a different decision-theoretic approach from what is typically used by mainstream economics. Informally, some post-Keynesians have argued for the irreducibility of macroeconomic issues to purely

microeconomic considerations where individuals' actions are based on expected utility calculations (see Weintraub, 1979).

Alternative types of microfoundations

Most of the literature up to the 1990s discussing the microfoundations of macroeconomics has focused on rationally behaving self-interested economic agents. More recently, attention has shifted to other forms of behaviour. Using evolutionary mechanisms or learning, economists have studied the evolutionary foundations of equilibrium notions (see Kandori, Mailath and Rob, 1993; Young, 1993). Allowing agents to imitate best practices they observe around them, or choosing best replies to some adaptively formed expectations of what others will do, the literature shows that under some conditions concerning the dynamic process the economy will converge to equilibrium play. Early work in this direction by Schelling (1978) shows, as noted in the introduction to this article, that macro phenomena such as racial segregation may be regarded as the unintended long-run outcome of the interactive effects of decisions of individual households to move into other neighbourhoods.

Alternatively, economists such as Fehr and Falk (1999) have looked at the consequences of non-selfish preferences for macroeconomic outcomes. They consider preferences for fairness and reciprocity to be important in explaining why managers do not consider cutting employees' wages. Wage cuts may be perceived as unfair and hostile, and managers fear that they will be followed by hostile actions on the part of employees. This literature provides an alternative foundation for the downward rigidity of monetary wages, and may start a literature on behavioural macroeconomics.

Conclusions

The microfoundations literature has brought about many changes in economic theory. Macroeconomic theory in the form of studies of the interplay of a few aggregate relationships is almost non-existent nowadays. Instead, an extreme form of 'microfoundations' is sometimes used in which the economy as a whole is represented in terms of a single agent decision problem. In this way, emergent properties appearing at the macro level that do not exist at the individual level are precluded from the analysis as the micro and macro level simply coincide!

Along with the many other models in the microfoundations literature reviewed in this article, we now see a wide spectrum of partly overlapping models dealing with different types of market frictions and market imperfections. Most of the literature before the 1990s adopts fairly traditional assumptions concerning individual behaviour. More recent contributions in the area of behavioural economics and evolutionary models with (adaptively) learning individuals are starting to explore the implications of different behavioural assumptions at the individual level and to consider the macro implications. These models have the potential to analyse how macro phenomena may emerge from the interactions among a heterogeneous set of individuals. Thereby, they may provide economic theory with a more plausible empirical underpinning, while sticking to the requirements of methodological individualism.

See Also

- involuntary unemployment
- methodological individualism
- social interactions (theory)

Bibliography

Agassi, J. 1960. Methodological individualism. *British Journal of Sociology* 11, 144–70.

Agassi, J. 1975. Institutional individualism, *British Journal of Sociology* 26, 144–55.

Arrow, K. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

d'Aspremont, C., Dos Santos Ferreira, R. and Gérard-Varet, L. 1990. On monopolistic competition and involuntary unemployment. *Quarterly Journal of Economics* 105, 895–919.

Barro, R. and Grossman, H. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

Benassy, J.-P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42, 502–23.

Bernheim, D. 1984. Rationalizable strategic behavior. *Econometrica* 52, 1007–28.

Blanchard, O. and Kiyotaki, N. 1987. Monopolistic competition and the effects on aggregate demand. *American Economic Review* 77, 647–66.

Bray, M. and Savin, N. 1986. Rational expectations equilibria, learning and model specification. *Econometrica* 54, 1129–60.

Brodbeck, M. 1958. Methodological individualisms: definition and reduction. *Philosophy of Science* 25, 1–22.

Bryant, J. 1983. A rational expectations Keynes type model. *Quarterly Journal of Economics* 98, 525–9.

Clower, R. 1965. The Keynesian counterrevolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cooper, R. and John, A. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103, 441–63.

Drèze, J. 1975. Existence of an equilibrium with price rigidity and quantity rationing. *International Economic Review* 16, 301–20.

Fehr, E. and Falk, A. 1999. Wage rigidity in a competitive incomplete contract labour market. *Journal of Political Economy* 107, 106–34.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Frydman, R. and Phelps, E., eds. 1983. *Individual Forecasting and Aggregate Outcomes*. Cambridge: Cambridge University Press.

Geanakoplos, J. and Polemarchakis, H. 1986. Walrasian indeterminacy and Keynesian macroeconomics. *Review of Economic Studies* 53, 755–79.

Guesnerie, R. 1992. An exploration of the eductive justifications of the rational expectations hypothesis. *American Economic Review* 82, 1254–78.

Hart, O. 1982. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97, 109–38.

Hayek, F. 1948. *Individualism and Economic Order*. Chicago: Chicago University Press.

Hicks, J. 1937. Mr. Keynes and the classics: a suggested interpretation. *Econometrica* 5, 147–59.

Horwitz, S. 2000. *Microfoundations and Macroeconomics: An Austrian Perspective*. London: Routledge.

Janssen, M. 1993. *Microfoundations: A Critical Inquiry*. London: Routledge.

Kandori, M., Mailath, G. and Rob, R. 1993. Learning, mutation and long-run equilibria in games. *Econometrica* 61, 29–56.

Kirman, A. 1989. The intrinsic limits of modern economic theory: the emperor has no clothes. *Economic Journal* 99(supplement), 126–39.

Kirman, A. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6(2), 117–36.

Kiyotaki, N. 1988. Multiple expectational equilibria under monopolistic competition. *Quarterly Journal of Economics* 103, 695–713.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press.

Lindbeck, A. and Snower, D. 1987. Efficiency wages versus insiders and outsiders. *European Economic Review* 31, 407–16.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R. 1978. Unemployment policy. *American Economic Review* 68, 353–7.

Lucas, R. 1987. *Models of Business Cycles*. Oxford: Basil Blackwell.

Lucas, R. and Rapping, L. 1969. Real wages, employment and inflation. *Journal of Political Economy* 77, 721–54.

Mankiw, N. 1985. Small menu cost and large business cycles: a macroeconomic model of monopoly. *Quarterly Journal of Economics* 100, 529–37.

Muth, J. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Nelson, A. 1984. Some issues surrounding the reduction from macroeconomics to microeconomics. *Philosophy of Science* 51, 573–94.

Pearce, D. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52, 1029–50.

Phelps, E., ed. 1970. *Microeconomic Foundations of Unemployment and Inflation Theory*. London: Macmillan.

Roberts, J. 1987. An equilibrium model with involuntary unemployment at flexible competitive prices and wages. *American Economic Review* 57, 856–74.

Schelling, T. 1978. *Micromotives and Macrobehavior*. New York: Norton.

Weintraub, E. 1979. *Microfoundations*. Cambridge: Cambridge University Press.

Yellen, J. 1984. Efficiency-wage models of unemployment. *American Economic Review* 74, 200–5.

Young, H. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

How to cite this article

Janssen, Maarten C. W. "microfoundations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000380> doi:10.1057/9780230226203.1096

The New Palgrave Dictionary of Economics Online

Mill, James (1773 – 1836)

Donald Winch

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

James Mill, Indian civil servant, Benthamite, and father and mentor of John Stuart Mill, introduced Jean-Baptiste Say's law of markets into British economic discourse. In addition to important works on the history of India, political and legal reform, and associationist psychology, he was the author of a textbook of Ricardian economics and played a major part in convincing Ricardo that he should write his *Principles of Political Economy* (1817). Through his son he was also responsible for giving prominence to proposals for taxing the ‘unearned increment’ in rental incomes that were influential in forming radical and socialist thinking in Britain.

Keywords

associationist psychology; Bentham, J.; betterment charge; birth control; Corn Laws; deductive method; effective demand; Fabian economics; labour theory of value; land nationalization; land tax; law of rent; Macauley, T. B.; Malthus, T. R.; methodology of economics; Mill, J.; Mill, J. S.; philosophic radicalism; productive and unproductive labour; Ricardo, D.; Say, J.-B.; Say's Law; Stewart, D.; underconsumptionism

Article

Mill was born in a village near Montrose in Scotland, the son of a cobbler-cum-smallholder. With the support of a local laird, Sir John Stuart, he was able to attend Montrose Academy and then, in 1790, Edinburgh University, where his original goal was to become a minister in the Scottish Kirk. During the seven years he spent in Edinburgh, he appears to have virtually become a member of the Stuart family, acting as tutor to the daughter of the house. Mill attended Dugald Stewart's lectures on moral philosophy and may have attended his class on political economy as well. Mill obtained his MA in 1794 and acquired a licence to preach in 1798. After an unsuccessful spell as an itinerant preacher and tutor, he moved to London in 1802, where he became part of an expatriate community of young Scots attempting

to make their way in the world through journalism. In addition to various freelance jobs, Mill edited the *Literary Journal* from 1803 to 1806, writing most of the articles dealing with political and economic topics. This enabled him to marry in 1805 and begin a family of nine children that was to prove a strain on his finances and temperament. He also began work on what was to be an 11-year enterprise, the research for and writing of his *History of British India* (1817). In addition to his income from journalism, Mill obtained assistance from Jeremy Bentham, whose disciple and intermediary with the world of affairs he became after 1808. In this way, and especially through his articles for the Supplement to the 4th, 5th and 6th editions of the Encyclopaedia *Britannica* (1815–24), Mill became the leading light of the movement known as philosophic radicalism, an intellectual grouping dedicated to the reform of parliament and other legal and political institutions according to Benthamite criteria for ‘good government’. In contradistinction to Bentham, however, Mill was a mature devotee of associationist psychology, as can be judged from his *Analysis of the Phenomena of the Human Mind* (1829). Mill also provided his eldest son, John Stuart, with an education which became part of the father's claim, both positive and negative, to have formed his son's mind and character. In 1819, partly as a result of the reception given to his *History*, Mill was appointed to the post of Assistant Examiner with the East India Company, rising to the post of Chief Examiner in 1830, a position he held until his death in 1836.
Mill's early economic writings consist of a large body of articles and two pamphlets, the first of which was *An Essay on the Impolicy of a Bounty on the Exportation of Grain* (1804), constructed along Smithian lines, the second entitled *Commerce Defended: An Argument by which Mr. Spence, Mr. Cobbett, and others have attempted to prove that Commerce is not a source of National Wealth* (1808). The latter is of interest to the history of economics, for two main reasons. The work contains the first enunciation in English of what was originally known as the Say–Mill law of markets; and it was through this work that Mill made the acquaintance of David Ricardo. The pamphlet was an attack on the views of those neo-physiocratic authors who argued during the period of Napoleon's economic blockade that agriculture rather than manufacturing and commerce was the true source of Britain's wealth. Mill agreed that claims on behalf of commerce had frequently been pitched too high, but he defended the Smithian view that manufacturing and other profits were a legitimate form of net surplus. He also upheld a pre-comparative cost interpretation of the gains from trade judged by the difference between the real costs incurred in producing goods for export and the putative domestic cost of producing imported goods. In countering Spence's underconsumptionist arguments on the relationship between capital accumulation, consumption and public expenditure, Mill defended Smith's distinction between productive and unproductive labour, translating it into the goods consumed by each category in order to show the importance of accumulation and productive consumption to economic growth. In refuting the idea of excessive accumulation, or general overproduction, Mill invoked Say's principle: ‘The production of commodities creates, and is the one and universal cause which creates a market for the commodities produced’ (1808, p. 135). Since the argument was conducted in barter terms, however, it amounts to little more than a statement of Say's identity, though the implication was that the conclusions applied equally to a money economy. Hence Mill's conclusion: the claims of commerce could be exaggerated whenever it was suggested that the extension of foreign markets was necessary to guarantee full employment. Here then was the English origin of the idea, expressed in characteristically unqualified terms, that was to lie at the heart of the controversy between Ricardo and Malthus over general gluts, and was later to be taken up by Keynes as the distinguishing mark of orthodox classical (and neoclassical) macroeconomics – an intellectual obstacle that had to be removed by a new theory of

effective demand in order to open the way for an explanation of involuntary unemployment in the *General Theory*.

It was largely as a result of Mill's encouragement that Ricardo overcame his doubts as to his capacity to move from being an economic pamphleteer to writing his *Principles of Political Economy*, which embodied those new doctrines that were necessary in order to replace those of Smith and other predecessors. Mill became Ricardo's impresario, coach and disciple; he was responsible for completing Ricardo's education and inducing him to enter parliament as spokesman for the ‘true’ principles of political economy and the reform programme of philosophic radicalism. Mill wrote one of the first ‘schoolbook’ accounts of Ricardo's doctrines in his *Elements of Political Economy* (1821), a record of what his son was taught at the tender age of 13. Ricardian doctrines appear in their most simplified and abstract form, but arranged according to the model provided by Jean Baptiste Say's *Traité d’économie politique* (1814), and with some embellishments that were not always acceptable to Ricardo himself. Thus in attempting to defend Ricardo's labour theory of value from attack by Robert Torrens, Mill bowdlerized the theory.

On policy matters, however, Mill struck out more boldly than Ricardo on two main issues: the advocacy of birth control as a solution to the problem of low wages, and a proposal for taxing the increment in rents accruing to landowners as a result of any legislative action which increased the demand for land. (Mill chiefly had the Corn Laws in mind.) He was sympathetic to land nationalization (if only as a way of frightening the landed aristocracy) and to the view that taxes on rent were one of the best means of raising government revenue; but he recognized that such proposals could not be introduced into a country where property had already exchanged hands at prices reflecting rental expectations. Nevertheless, since this only gave a legitimate expectation to present rents, plus an allowance for improvements undertaken by the landowner, Mill was in favour of levying what would later be called a ‘betterment’ charge on increments in rent beyond this. In adopting this position he believed he was merely carrying Ricardo's conclusions on the special nature of rent, as compared with wages and profits, to their logical policy conclusion.

Mill then, rather than Ricardo, is the source of that strand of radical thinking on the ‘law of rent’ that was to be passed on via his son to the Fabians. More significantly, when judged by results, the official positions occupied by Mill and his son in India House ensured that their views on taxation and land revenue were influential in practice. It was primarily through his efforts that a determined attempt was made in the Bengal provinces to replace a landowner-based (*zemindari*) system of land tenure with one based on the view that the government should retain the ultimate rights in land and deal directly with the peasant cultivator or *ryot*, basing the tax assessment on Ricardian or pure rent.

Mill is also of some importance for his views on the methodology of political economy and other moral sciences, as can be best illustrated – negatively at least – by the attack mounted by Macaulay on Mill's essay on ‘Government’ for the *Encyclopaedia Britannica*. Mill was an extreme upholder of the virtues of the deductive method, and a critic of practical men who professed to be ‘all for fact and nothing for theory’. In this respect Mill is sometimes credited with an influence on, certainly as encouraging, Ricardo's adoption of the a priori method of working from unqualified assumptions to ‘strong cases’, and from there to policy conclusions. Since there is little firm evidence to establish this proposition, those who are either critical or defensive of the Ricardian method should probably dispense with Mill rather than attempt to draw attention to similarities or differences between the practice of both men. We do know, however, that Mill produced a son who believed that his education had peculiarly fitted him to

engage in ‘the science of science itself, the science of investigation – of method’. We also know that in the aftermath of the Macaulay attack the son wrote an essay ‘On the Definition of Political Economy; and on the Method of Investigation Proper to it’, later to be the basis for Book VI of his *System of Logic* (1843), in which he criticized both his father and Macaulay in the course of expounding an interpretation of the role of deductive methods in political economy which remained canonical for much of the 19th century.

See Also

- classical distribution theories
- Enlightenment, Scottish
- land tax
- Mill, John Stuart
- Ricardo, David
- Say, Jean-Baptiste

Selected works

1804. *An Essay on the Impolicy of a Bounty on the Exportation of Grain*. London. Repr. in Winch (1966).

1808. *Commerce Defended: An Argument by which Mr. Spence, Mr. Cobbett, and others have attempted to prove that Commerce is not a source of National Wealth*. Repr. in Winch (1966).

1817. *History of British India*. London: Baldwin & Cradock.

1821. *Elements of Political Economy*. London: Baldwin Cradock & Joy; 2nd edn, 1824; 3rd edn, 1826. Repr. in Winch (1966).

1829. *Analysis of the Phenomena of the Human Mind*. London: Baldwin & Cradock.

Bibliography

Bain, A. 1882. *James Mill: A Biography*. London: Longmans.

Ball, T. 1992. *James Mill: Political Writings*. Cambridge: Cambridge University Press.

De Marchi, N.B. 1983. The case for James Mill. In *Methodological Controversy in Economics: Historical Essays in Honor of T.W. Hutchison*, ed. A.W. Coats. Greenwich, CT: JAI Press.

Keynes, J.M. 1936. *General Theory of Employment Interest and Money*. London: Macmillan.

Macaulay, T.B. 1829. Mill's essay on government: utilitarian logic and politics. *Edinburgh Review* 97 (March).

Mill, J.S. 1844. On the definition of political economy; and on the method of investigation proper to it. In *Essays on Some Unsettled Questions in Political Economy*. London: Parker. [Written 1829; first published in the *London and Westminster Review*, 1836.]

Mill, J.S., 1843. *A System of Logic*. London: Parker.

Ricardo, D. 1817. *Principles of Political Economy and Taxation*. London: John Murray.

Say, J.-B. 1814. *Traité d'économie politique*, 2nd edn. Paris: A.-A. Renouard.

Winch, D. 1966. *James Mill: Selected Economic Writings*. Edinburgh: Oliver & Boyd. Reissue, New Brunswick: Transaction, 2006.

How to cite this article

Winch, Donald. "Mill, James (1773–1836)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000306> doi:10.1057/9780230226203.1097

The New Palgrave Dictionary of Economics Online

Mill, John Stuart (1806 – 1873)

Neil De Marchi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mill approached economic theory using conceptual and verbal analysis. This worked well for settled truths applied to circumscribed situations, such as a rise in the ratio of food prices to manufactured goods prices under growth subject to decreasing returns. He needed, but did not develop, a different method for multi-causal problems. Mill insisted that value and production were settled areas of political economy but was open to societal reforms that would result in altered shares of income and wealth. This distracted from the coherence of his *Principles of Political Economy* and from his reputation as a theorist, while ensuring that he will be remembered for challenging readers to entertain breathtaking prospects for human improvement.

Keywords

absolute and exchangeable value; Bentham, J.; birth control; communal ownership; Corn Laws; education; falling rate of profit; happiness; inheritance and bequests; labour as a measure of value; laissez-faire; Mill, J.; Mill, J.S.; rent; Ricardo, D.; role of government; Smith, A.; social preferences; speculation; stationary state; Taylor, H.; Value

Article

John Stuart Mill was the pre-eminent British economist of the mid-19th century. But he was much more besides, commanding a hearing in public debates on subjects from logic to liberty, the position of women to the problem of Ireland. Yet, though his *Principles of Economics, with Some of Their Applications to Social Philosophy* (1848) dominated economic discourse for 40 years, there is little in it of technical, or even conceptual, advance that would justify placing him in a pantheon of great economists, if one judges by what is understood as economics today. Mill should be known and honoured more for his vision of an improved condition for humankind and for the novel economic views that formed part of that vision than for his economic analysis as such.

Approaching Mill's economic ideas from this perspective necessitates attending to his passage from early Benthamite propagandist and defender of Ricardian doctrine to more pragmatic reform strategist, with a greatly expanded notion of happiness. The transformation was traumatic in that it involved a lapse, and relapses, into depression, and it meant modifying some old convictions. Positively, however, Mill also discovered the possibility of cultivating feelings – ‘of inward joy, of sympathetic and imaginative pleasure’ – and began to see for others the prospect of ‘perennial sources of happiness, when all the greater evils of life shall have been removed’ (1873, pp. 147, 151). Certain elements in this prospect seemed to him to require, ultimately, the replacement of competition with cooperation, and there were various other novel economic aspects to this notion. But the inspiration was quite different from the motivations reflected in the economics Mill had learned from James Mill (his father) and Ricardo. This makes it hard to find the strong logical link between economic doctrine and social philosophy implied by the word ‘Applications’ in the subtitle of his *Principles*. In fact there is a switch in mode between the doctrines and the applications, from the demonstrative to the conditional – from result to possibility – which seriously weakens that link. There is also a difference of tone: Mill wrote with great immediacy and verve about possibilities for the improvement of humankind, but defensively on the economic doctrines he had inherited. He embraced new social thinkers, borrowing freely from them even if, as he often allowed, their views were incomplete, not always coherent and even downright misleading in certain respects. But he chose not to keep up with new developments in economics, more especially those that employed mathematics. Instead he stuck to the method that was his forte, clearing up terminological and logical confusion and thus ‘perplexities’.

1 The constraints of a Benthamite education

Mill was the eldest of eight children born to James Mill and Harriet (née Barrow). He was home-schooled by his demanding father, whom he eventually succeeded as Examiner in the East India Company. The elder Mill was a Scots literary émigré in London, disciple of Bentham, and a leading protagonist of utilitarian reform. His writing took precedence, and, besides giving John basic instruction, he largely turned over to him the education of the younger children. John's mother, worn down, developed no intellectual interests, and became for him a model of what women should not be, in sharp contrast to Harriet Taylor, with whom he fell in love in the 1830s and married in 1851. Harriet Taylor shared Mill's reformist ideas and emboldened him in expressing his notions concerning autonomy, not least for women.

John's spectacular childhood achievements are well known: beginning to learn Greek words at the age of three, and starting Latin at eight, acquiring the language by dint of having to instruct his siblings. Studies in Logic began at 12, and Political Economy at 13. Instruction in the latter took the form of lectures from his father, which he was to summarize and repeat the following day, on their daily walk. James Mill's *Elements of Political Economy* (1821), which the daily lectures became, was essentially a set of logical propositions. John always regarded logical analysis as the most valuable of all mental trainings.

At the age of 20, however, Mill discovered that something was lacking. In describing what he would later call a crisis in his mental history, he recalls imagining the accomplishment of all the Benthamite reforms for which he was agitating, but finding himself without satisfaction at the prospect. Recovery was effected, he tells us, through reading new authors and modifying his circle of friends; central to the process, however, was the realization that the Benthamite views he had imbibed were entirely too

narrow.

During Mill's childhood the family spent their summers close to Bentham, and in 1821 he began reading him – in fact Dumont's edited version of notes, published as the *Traité de Législation*. This work gave him ‘a vista of improvement’ for human life based on coherent laws and opinions founded on the principle of utility (1873, pp. 69, 71). Somewhat later Mill was given the task of editing Bentham's manuscript of *The Rationale of Judicial Evidence* (1827); so, by the time of his depression, Mill must have been as well equipped as anyone in Britain to convey accurately Bentham's thinking on government and to comment on matters of English law, which he did, frequently, in the daily and periodical press.

Mill's post-depression reappraisals were cautious, even after Bentham and his own father had died. In 1838, however, he was able to present a lengthy and balanced account. He praised Bentham for having accomplished the first scientific investigation of the large and messy body of precepts that comprised English law, using as his tool what Mill called ‘the method of detail’– separating wholes into their parts, resolving abstractions into concrete things – in short, ‘breaking every question into pieces before attempting to solve it’ (1838, pp. 83, 100). The method was Baconian; Bentham's originality lay not in having invented it, but in having applied it to the law. He had not yet had the impact that he deserved, partly because of his obsessive verbal partitioning of every topic. This resulted in tedious intricacies for which few readers cared. But the method had the great merit of bringing into question even commonly accepted truths and constituted a tool for identifying the rationale, or lack of it, in every existing or proposed law.

Take murder, for example. According to common sense and religion it is a crime. But why? A rational examination would ask whether the benefits to the perpetrator were outweighed by costs, in terms of the suffering inflicted on the victim; the feelings of insecurity aroused in others; and the discouragement to certain sorts of industry and useful pursuits through fear, as well as any diversion of resources to warding off the perceived danger. If the costs dominate, then murder must count as a crime and the infliction of punishment is warranted (1838, p. 83). Mill judged it useful to challenge even basic truths in this way, both because they support many subsidiary truths, and for the mental discipline involved, which we need in order to guard us against too readily following moralists who invoke, unexamined, phrases such as ‘law of nature’ or ‘right reason’, and politicians who call for ‘liberty’ and ‘social order’ (1838, p. 84; 1873, p. 67).

On the negative side, Mill found Bentham's approach cripplingly narrow. By focusing on pain and pleasure exclusively Bentham implied that human beings are governed solely by their own immediate interest and their sympathy or antipathy towards others.

Among things ignored are a feeling of moral approbation or disapprobation (conscience); standards of excellence or the desire of perfection as an end in itself; a sense of honour or personal dignity; a love of beauty; the passion of the artist; love of the congruency or consistency of things, or of their conforming to their intended ends (1838, pp. 95–6). This philosophy, devoid of morality and spiritual interests, did not sit well with the new Mill, who had now ‘learnt by experience that the passive susceptibilities needed to be cultivated as well as the active capacities’ (1873, p. 147). But neither was Bentham's philosophy able to cope well with even the purely business aspects of life, since in practice every action influences our own and others’ affections and desires (1838, p. 98).

Mill's own revised aspiration was to give due place to the *moral*, *aesthetic*, and *sympathetic* aspect of every human action. We must ask of each, is it right or wrong? Is it beautiful – inspiring, estimable?

And is it ‘loveable?’ (1838, p. 112). These additions could have come from Smith's *Theory of Moral Sentiments* (1759), though it is not clear that Mill knew that work.

Returning to Bentham's philosophy, Mill concluded that, since it ignored feelings and the moral, the inspiring and the lovable, it simply offered a crude guide to desirable outward circumstances and regulations to effect them. But circumstances and punishments cannot instil the sympathy that binds us. Mill drew from the aftermath of the French Revolution the lesson that social feelings are only shallow-rooted in human nature, that, once a society has torn down old institutions that have grown corrupt, conflicting interests are likely to produce anarchy (1838, p. 99). For sympathy to prevail, then, there must be education directed towards making it second nature to care for others as we care for ourselves. On the political level, Bentham, it was true, had urged that government be delegated to those whose interests are identical with the interests of the population at large. But Mill feared giving even such a group control over the whole; without a serious opposition its members are apt to become tyrannical (1838, pp. 106–8). Mill's fears in this regard presage those expressed by Hayek in his *Road to Serfdom* (1944).

What did it mean to be a Benthamite propagandist, as Mill was before 1826? Two examples will illustrate. He distributed pamphlets on methods of birth control, convinced that the average condition of the working classes could be permanently improved only by voluntary reduction in their numbers relative to the means available for their support. And he opposed the Corn Laws because, in restricting imports, they kept the price of grain higher than it need be, making the most basic means of sustenance less accessible, which was a clear net loss of aggregate happiness. By contrast with these cut-and-dried policy choices, Mill's views in the decade or so after 1826 were largely an outworking of the enlarged basis for personal and social happiness that he had adopted.

He also became more practical; he saw that radicals must co-opt conservatives to command a parliamentary majority. The new Mill judged that there is no simple and direct connection between first principles, such as the principle of utility, and actions that will increase happiness. For individuals differ in their primary beliefs, making happiness ‘too complex and indefinite an end’ to pursue in the Benthamite manner (1838, p. 110). Fortunately, division on ultimate standards does not preclude agreement on intermediate ends. During the 1830s, therefore, Mill strove to engage erstwhile opponents on such intermediate goals, arguing, for example, that the landed interest's support for the Corn Laws would be weakened if it could be shown that those laws actually increased wage costs, harming landowners both as employers and consumers.

2 Early political economy

Philosophically and in terms of political practice Mill's new views had far-reaching implications for his life and writing. Much of his political economy, however, underwent relatively little change. Not only prior to 1826, but even in his *Principles* he retained and defended the core doctrines of his father's *Elements* and Ricardo's *Principles of Political Economy and Taxation* (1817). At times, especially early on, his defence was conducted with a fierceness that blinded him to any merit in alternative views. In the 1840s, by which time he wanted to make place, alongside the old core doctrines, for many additional topics in economic analysis as well as for his favourite ideas for improving society, the combination lent to his *Principles* the appearance of a patchwork. An illustration of his early treatment of critics can be

given here; the patchwork aspect of the *Principles* will be touched upon in Section 5.

The illustration concerns the central subject of value. Ricardo and James Mill chose to discuss value strictly in terms of exchange value, for both wanted to show that, as population grows, the value of food *relative* to manufactures rises because land of lesser productivity has to be brought under cultivation. On the assumption that returns in manufacturing are constant but unit labour costs in agriculture are rising, it is obvious what causes an observed rising trend in the relative price of food.

Smith, however, in addition to allowing value to be relative, stressed the pain cost of labour and insisted that the true cost of goods is how much labour they command. Behind this emphasis was a concern that the sacrifice of ease involves a loss of happiness, since ease for Smith was linked with tranquillity of mind, and the latter with happiness. Mill understood this (see 1848, pp. 580–1). Nevertheless, as a young defender of his father and Ricardo, he dismissed Smith's alternative measure of value lest readers be deflected from focusing on relative labour input, so central to the case against the Corn Laws. Hence, when Malthus, in his *Measure of Value* (1823), opted for Smith's sacrifice measure, Mill, aged 17, portrayed him as logically incompetent: to make labour command a measure of value, Malthus had in fact to assume what he needed to show, that the value of wages is always the same (1823, p. 57). Mill was correct but also quite one-eyed. Malthus showed in his *Definitions in Political Economy* (1827) that he too grasped the difference between an invariable measure of value and exchange value, yet preferred to measure even exchange value by how much labour commodities can command because that is appropriate if one's purpose is to ascertain ‘the sacrifice which people are willing to make in order to obtain [commodities]’ (1827, p. 211).

Against this crabbed performance, it is refreshing to find Mill, a very few years later, writing comparatively wide-ranging and subtle analyses of current events. The best of these was an essay, ‘Paper Currency and Commercial Distress’, in the short-lived radical *Parliamentary Review* of 1826, on the recent ‘commercial revulsion’.

Mill insisted that the proximate cause of recession in this case was a prior speculation, not in new ventures, but in existing activities. The dominant group of parliamentarians instead blamed an over-issue of small notes by country bankers – an attribution of causation, Mill suggested, that betrayed a deeper scepticism about paper currency. Drawing on Tooke's recently revised *Considerations on the State of the Currency* (1826), Mill showed that the speculation had begun after trade papers reported below normal stocks in a few key goods, including grains. In the usual way this had induced dealers to increase their purchases, causing an immediate price increase, a pattern that then extended itself, though for purely speculative reasons, to a wider range of goods. Mill agreed that there had been an increase in credit associated with speculative buying, but observed that this did not require small notes: trade credit and bills of exchange would have sufficed. He also showed that the observed movements in the currency were not what one would expect from an expansion of the circulation. What had happened was merely a redistribution of, rather than an overall expansion in, the circulation. When grain prices first began to rise, means of circulation shifted from London to the country, sustaining the rise in agricultural prices but lowering the prices of manufactures in the city. Manufactured exports therefore rose, and, because grain imports were restricted under the Corn Laws, the exports occasioned an influx of gold. This would have happened whether the medium of circulation was metallic or paper, and no net expansion of their notes by country bankers need have been involved. It followed that the ultimate culprit was the Corn Laws, which prevented imports from offsetting the speculative purchases occasioned by the initial shortfalls in stocks of grain.

This was a tour de force in applied economic analysis. Mill contrasted his analysis with an account of why the parliamentarians had got it wrong. At root they lacked general principles. Inevitably, then, the views of ‘practical men’ – men who observed a few facts near at hand and generalized on that inadequate basis – prevailed with them.

Practical men as nemesis was a theme in Mill's famous essay ‘On the Definition of Political Economy; and on the Method of Investigation Proper to It’, which was published in 1836 and again, along with other youthful exercises in clarification of the principles of the new political economy, in *Essays on Some Unsettled Questions of Political Economy* (1844). From an economist's point of view perhaps the most useful portion of Mill's *System of Logic, Ratiocinative and Inductive* (1843) was his extended analysis of the social scientist's equivalent of experimentation: the various methods of ascertaining causes (Book III). The earlier methodological essay started him down that road.

The methodological essay is by far the most sophisticated of the set published in 1844, the others, with two other exceptions, suffering from being of the crabbed, defensive sort. In an essay on ‘The Influence of Consumption on Production’ Mill allowed that a general glut could occur, temporarily, if there were a sudden general preference for liquidity. The other exception was an essay on ‘The Laws of Interchange between Nations’, in which Mill elaborated on his father's suggestion that the division of the gains from trade would depend on the relative strengths of demand of the participating countries. This was one of Mill's few lasting contributions to economic analysis. Marshall utilized it in his essay *The Pure Theory of Foreign Trade* (1879), and his demonstration, in the context of the 1903 tariff debate that whether the foreigner bears the cost of a tariff will depend on the shape of his offer curve.

3 Espousing selective conservatism, and incorporating social evolution

By the late 1830s, as we have seen, Mill had begun to make explicit what was required to make good on Bentham's omissions. But how exactly were these to be supplied? Here Mill had recourse to German views, conveyed in language more palatable to English minds by the poet and essayist Samuel Taylor Coleridge. He also drew on the writings of the Saint Simonians, particularly the early work of Auguste Comte.

Mill took from Coleridge the idea that education should assist in forming national character. The young need to be imbued with an ‘active principle of cohesion’, of sympathy, not hostility; union, not separation. This might require heroes, or at least common beliefs; either way the goal must be to make caring for others second nature. By implication, there was a very active role here for government, a role more positive than either the pre-revolutionary French philosophers had allowed, or than their English counterparts had felt to be necessary. The French had wanted to tear down corrupt and spent institutions, after which government should basically leave people be (laissez-faire). On the English side, the national discomfort with conflict and a preference for compromise had asserted itself in the 18th century; after the strife of the 17th century the English had settled for living with whatever institutions there were, provided they were reduced to practical nullities (1840, pp. 142–4, 146). There was no sense in England that education should be reformed to build national character and supply an active force for social cohesion.

Mill picked up on three intriguing ways in which government might contribute to or reflect cohesion; and each had an economic aspect. First, the state ‘ought to be considered as a great benefit society, or

mutual insurance company, for helping (under the necessary regulations for preventing abuse) that large proportion of its members who cannot help themselves' (1840, p. 156). The details of this idea were not filled in, and it does not reappear in Mill's later work, but it sounds not unlike the Social Security system of the United States or the mandatory contributions towards retirement now applied in many countries.

Second, the land must be considered a trust. Mill distinguished this notion from calls for the state to reclaim private property, though he noted that the law of real property originally applied only to movables. It was his view that, if an owner possesses more land than is necessary for him to sustain himself and his family by his own labour, the excess confers on him power over others and the state may require that this power not be abused. This meant that even the system of cultivation is a proper concern of society (1840, pp. 157–8). The notion reappears in the *Principles*, though as one among several possibilities for limiting bequest and tenure (1848, p. 227).

Third, Mill insisted that education, being of almost boundless power, should be used by the state to foster public opinion in favour of the attracting forces within society. These forces derive from our love of praise, favour, admiration and respect, and our dread of shame and ill repute – again, ideas central to Smith's *Moral Sentiments*, though Smith was not acknowledged by Mill. Mill held that, once the basic means of living has been obtained, almost the whole of our remaining effort is directed to acquiring the favourable regard of others. In fact this is the driving force behind the industrial and commercial activity that advances civilization. Love of praise, however, is also the source of the selfish thirst for aggrandizement; hence the state must tip the balance in favour of social sympathy (1840, pp. 410–1).

A possible explanation of Mill's slighting of Smith is available in this instance. Mill might easily have seen Smith as insufficiently positive about the role of government. Smith, for example, advocated basic education for the poor in the hope that, for those condemned by excessive specialization to repetitive, trivial tasks, it might mitigate the risk of moral deformity (1776, p. 788). But for Mill that was too feeble a response, too restrained an expectation. For him education was the key to all future social and personal improvement.

The expectation of improvement also impelled Mill farther in a related direction. There is an implication for distribution in the notion that mutuality of interests makes it easier to cultivate and fix social feelings (1861, p. 231). Mill took from Comte the conviction that there had been considerable social progress towards cooperation, a trend likely to continue. The cooperative spirit, in turn, ought to make it possible for individuals to regard working for the benefit of others as a good it itself, requiring no compensation. Ideally, what we get for ourselves should not be viewed as a quid pro quo for our cooperation but in terms of 'how much the circumstances of society permit to be assigned' to us, 'consistent with the just claims of others'. The market method of settling a worker's share of the produce may be a temporary practical necessity, but morally is not ideal. Society, Mill understood, was not yet ready to relinquish the market, so he judged it better to let competition decide rather than to impose any artificial mode of distribution as yet untried – save in the army, where it was the de facto norm (1865, pp. 340–1). The idea reappears in the discussion, in the *Principles*, of cooperative arrangements in industry, though Mill's emphasis there was strongly on shared ownership for harnessing 'productive energies', and cooperation as still for the future, while competition is not only dominant but also has its positive side emphasized (1848, pp. 216, 337, 356).

4 Happiness: an enlarged view

Feelings aside, morals, aesthetics and sympathy – the other three missing elements in Bentham's philosophy – put happiness firmly in the social sphere. Mill continued to hold, with Bentham, that the general end should be the multiplication of happiness, but increasingly happiness had to involve the desire to care for others. Even if by nature we have only a small germ of this feeling it is one which can and should be ‘laid hold of and nourished by the contagion of sympathy and the influences of education’ and supported by external sanctions (1861, p. 233). Social ends would thus be rendered part of our inmost motivation.

On the one hand, then, Mill naturally came to think of happiness as linked to the growth of the cooperative spirit. On the other, he also saw it, crucially, as involving the development of the inward man, which is where the three added dimensions really have their purchase on our emotions and motives. He would eventually redefine individual happiness as a satisfied life, one with a balance between tranquillity and excitement. A person who finds this balance can be content with little pleasure, and can even be reconciled to much pain (1861, p. 215). Mill saw no reason why the mass of humankind could not unite tranquillity with excitement, since, even without great improvement in outward circumstances, the inward balance could be struck.

Notice, however, that inward happiness, since it does not depend essentially on a person's material resources or situation, removes the end – happiness – from the status of positional good. It may be that this realization predisposed Mill to accept Comte's ideas on cooperation – that cooperation itself is made easier if the overall end in view does not involve rivalry – though there is no collaborating evidence for this.

5 Mill's mature political economy

Mill's *Principles* was an uneasy amalgam of Smith, Ricardo, Mill's own refined insights on various discrete topics, and new social ideas.

The treatise can be dissected for its insights on a wide range of topics, as Hollander has done in his full-length study of Mill (1985). Hollander shows Mill to have had an unusually clear grasp of mechanisms: the determination of (long-run or cost) prices by variation of supply; of the rate of return by the proportion of the work day required to produce wage goods; of the alternative to wage reduction that exists in population control, as a way of equating the growth rates of population and capital accumulation; of the feedback between speculation engendered by a declining rate of profit and the loss of capital due to business failures, which in itself will raise the rate of profit; of a general desire to hold money as a cause of depressions; and so on.

These mechanisms summarize clearly and appropriately Mill's analytical contribution, which he even recorded on occasion as a list of propositions established (for example 1848, pp. 497–9). Since, however, there are various possibilities implicit in the application of such propositions, circumstances matter, as Mill himself stressed in his essay on method. This distinction between demonstrated truth and institutional possibilities inevitably loosened the logical connection between Mill's economic analysis and his social views. Thus he could analyse in a Malthusian-cum-Ricardian way the growth tendencies that issue in stationariness: given diminishing returns in agriculture, constant population growth and a fixed state of the productive arts, growth will eventually cease. Yet he could also freely explore possibilities for human nature, society and the ‘Art of Living’ in the stationary state, unconstrained by those economic tendency laws.

This is partly responsible for the patchwork appearance of the *Principles* noted earlier; yet it probably owes as much or more to Mill's having retained key doctrines from Ricardo and his father while accepting that they had not always elaborated the general case. In acknowledging this, on rent for example, Mill ended up incorporating qualifications that the reader must locate here and there – in the case of rent, in four separate chapters, spread across three books.

Marshall chose an alternative way of addressing rent. He noted that rent of land is ‘no isolated economic doctrine’ but ‘simply the chief species of a large genus of economic phenomena’ (1890, I, p. 629). Mill, sticking to the Ricardian view that rent of land is ‘differential and peculiar’ (1848, p. 495), concluded that rent only enters into the cost of production if there is a scarcity element involved – cases ‘rather conceivable than actually existing’ (1848, p. 498). Marshall, however, constructed a continuum of cases in which, at one extreme, a productive resource is in strictly fixed supply and its return therefore a surplus or ‘rent in the strictest sense of the term’, while at the other end the resource is quickly reproducible and its return no more than the interest on the money cost of obtaining more of it. There are multiple combinations in between where revenue might temporarily diverge from interest, for reasons originating either on the supply or the demand side. Specifying the exact circumstances may have consequences, as when a choice must be made whether to impose a tax on producers rather than consumers. Marshall's point was that interest and quasi-rent ‘shade into one another gradually’, making such choices very difficult (1890, I, pp. 412–21). Hence, as to ‘rent not entering into cost’, he concluded that the phrase cannot be rescued by verbal analysis but ‘only by experience’. At the same time, it is a ‘denial of subtle truths’ to generalize either in the direction chosen by Mill or its opposite (1890, II, p. 439). Mill's fierce defence of Ricardian doctrine in this instance, as in some others, did not advance the cause of clarity nor did it allow experience the crucial role his own method suggested it should have.

As noted, Mill incorporated analytical developments in economics selectively; he left aside those that involved mathematics – not the strongest component in his early education (1873, p. 15, though see also p. 59) and a mode of reasoning he later came to suspect of strengthening the false claim that moral, political and ‘supersensual’ truth may be had without self-observation and common experience (1832, p. 331; 1873, p. 233). Not to speak of French works, he failed to mention even contemporary English analyses of profit-maximizing equilibrium, and the gains and losses from the supposition of various changes (in technique – hence machinery – or in taxes), such as those due to Tozer (1838) and Lardner (1850). Much later he responded to Jevons, though probably not from having studied the *Theory of Political Economy* (1871) at first hand. And from reviews of the *Theory* Mill *mis*judged that Jevons just offered ‘a notation implying the existence of greater precision in the data than the questions admit of’ (Mill to Cairnes, 5 December 1871, in Mill, 1963–91, vol. 17, p. 1862).

There remain, as the freshest contributions of the *Principles*, those of Mill's notions on future social possibilities that have some economic content.

1. In the context of reflecting on possible distributions of property (Book II, Ch. 1), Mill posited that a society might be in the position of having to choose between communal ownership and private. He supplied arguments why the communal arrangement ought not to be rejected out of hand. Shirking, he allowed, would be a serious problem; moreover, the experiment should not be tried without universal education first being implemented and numbers (population) controlled, so that none would lack for subsistence. Under such circumstances one might assume more public spirit than we are used to seeing. Nevertheless, and difficulties with the alternative notwithstanding, he thought existing production

arrangements far from ideal: in nine-tenths of cases there are principal–agent problems (not his terminology). All said, he suggested, the choice should turn on the most important issue of all: which system ‘is consistent with the greatest amount of human liberty and spontaneity’? (1848, Book II, Ch. 1, p. 208).

2. In the very next chapter Mill argued for a distinction between the right of private ownership and the right to bequeath and inherit. On the one hand, the power to bequeath might be inconsistent with the permanent interest of the race; on the other, the essential principle of property – to assure to all what they themselves have produced – cannot apply to the raw materials of the earth. After universally agreed exceptions, Mill observed, where doubt is present the presumption should be against the owner (1848, Book II, Ch. 2).

3. In Book IV, Chapter 4, Mill adumbrated his own non-Smithian tendency for the rate of profits to fall. He accepted the tendency, but argued that it reflects not only the natural (Ricardian) consequences of the extension of cultivation, but also the progress of civilization. As people become more rational they also become more self-controlled, and find lower rates of interest and profits acceptable. Not only are rational people less apt to discount the future; they also save against contingencies even in the absence of any immediate need. In a more civilized world, moreover, risks are lower because the strong social spirit renders capital and wealth generally more secure.

4. Building on the arguments just listed, Mill was also able to contemplate a future with zero growth (the stationary state: Book IV, Ch. 6). Here he reiterated the theme that ‘a population might be too crowded’ for that solitariness and tranquillity so essential to depth of character. Quite apart from that, zero growth of course does not preclude ‘improving the Art of Living’. And in any case, the social ideal cannot be the elbowing, crushing competition all around us. We should be able to get beyond the struggle for (relative) riches, so as to realize a state in which ‘while no one is poor, no one desires to be richer, nor has any reason to fear being thrust back’ (1848, p. 754).

5. A third chapter in Book IV, ‘On the Probably Futurity of the Labouring Classes’, expands on all this, but stresses the importance of making people more ‘rational’ by increasing their independence, this by reversing the hiring–service relationship and replacing it with employer–employee associations (1848, p. 763). As so often, Mill qualified this sweepingly optimistic view with a pragmatic caution: competition need not be dispensed with; after all, cheaper goods come of it and labour must therefore benefit (1848, p. 794).

6. Finally, in Book V, especially Chapter 11, there is an exploration of laissez-faire, the general rule, and the ‘large exceptions’ to it that Mill also deemed necessary. The positive role of government should extend to education; the care of minors (from which category he was careful to exclude women); and a long list of cases where private initiative would be preferable if only it were not generally lacking for one reason or another. The list reads quite like the one Smith provided, of desirable projects for which no individual or small group can find the necessary financing; only Mill extended it beyond roads, harbours, canals, and so on, to hospitals, schools, colleges and printing presses (1848, pp. 944, 947, 950, 970).

See Also

- competition

- cooperation
- property rights

Selected works

1823. Malthus's measure of value. In *Collected Works*, vol. 23.

1826. Paper currency and commercial distress. In *Collected Works*, vol. 4.

1836. On the definition of political economy; and on the method of investigation proper to it. In *Collected Works*, vol. 4.

1838. Bentham. In *Collected Works*, vol. 10.

1840. Coleridge. In *Collected Works*, vol. 10.

1843. *A System of Logic, Ratiocinative and Inductive*. In *Collected Works*, vols 7 and 8.

1844. Of the laws of interchange between nations. In *Essays on Some Unsettled Questions of Political Economy*. In *Collected Works*, vol. 4.

1844. On the influence of consumption on production. In *Essays on Some Unsettled Questions of Political Economy*. In *Collected Works*, vol. 4.

1848. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*. In *Collected Works*, vols 2 and 3.

1861. Utilitarianism. In *Collected Works*, vol. 10.

1865. Auguste Comte and positivism. In *Part I*, *Collected Works*, vol. 10.

1873. Autobiography. In *Collected Works*, vol. 1.

1963–91. *Collected Works of John Stuart Mill*, 33 vols. ed. J.M. Robson. Toronto and London: University of Toronto Press and Routledge & Kegan Paul.

Bibliography

Bentham, J. 1827. *Rationale of Judicial Evidence, Specially Applied to English Practice*, 5 vols. ed. J.S. Mill. London: Hunt & Clarke.

De Marchi, N. 2002. Putting evidence in its place: John Mill's early struggles with ‘facts in the concrete’. In *Fact and Fiction in Economics. Models, Realism, and Social Construction*, ed. U. Maki. Cambridge: Cambridge University Press.

Dumont, P.É.L. 1802. *Traité de législation civile et pénale*, vols. 3. Paris: Boussange, Masson, Besson.

Hayek, F.A. 1944. *The Road to Serfdom*. London: Routledge.

Hollander, S. 1985. *The Economics of John Stuart Mill*. Oxford: Basil Blackwell.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan.

Lardner, D. 1850. *Railway Economy*. New York: A.M. Kelley, 1968.

Maas, H. 2005. *William Stanley Jevons and the Making of Modern Economics*. Cambridge: Cambridge University Press.

Malthus, T.R. 1823. *The Measure of Value Stated and Illustrated*. London: John Murray.

Malthus, T.R. 1827. *Definitions in Political Economy*. London: John Murray.

Marshall, A. 1879. *The Pure Theory of Foreign Trade. The Pure Theory of Domestic Values*. Circulated privately; repr. London: London School of Economics and Political Science, 1930.

Marshall, A. 1890. *The Principles of Economics*, 9th (variorum) edn. vol. 2, ed. C. Guillebaud. London: Macmillan for the Royal Economic Society, 1961.

Mill, J. 1821. *The Elements of Political Economy*. London: Baldwin, Cradock & Joy.

Ricardo, D. 1817. *The Principles of Political Economy and Taxation*. London: John Murray.

Robson, J.M. 1968. *The Improvement of Mankind*. London: Routledge & Kegan Paul.

Smith, A. 1759. In *The Theory of Moral Sentiments*, ed. A.L. Macfie and D.D. Raphael. Oxford: Oxford University Press, 1976.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, vol. 2, ed. R.H. Campbell and A.S. Skinner. Oxford: Oxford University Press, 1976.

Tooke, T. 1826. *Considerations on the State of the Currency*. London: John Murray.

Tozer, J.E. 1838. *Mathematical Investigation of the effect of Machinery on the Wealth of a Community in which it is Employed and on the Fund for the Payment of Wages*. Cambridge: Cambridge Philosophical Society, Transaction 6.

How to cite this article

De Marchi, Neil. "Mill, John Stuart (1806–1873)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000172> doi:10.1057/9780230226203.1098

The New Palgrave Dictionary of Economics Online

Millar, John (1735 – 1801)

Nicholas Phillipson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

government; Millar, John; profit; Scottish Enlightenment; Smith, Adam

Article

Born in Lanarkshire, the son of a Presbyterian minister, Millar was educated at Glasgow University for the Scottish Bar. He became a protégé of Adam Smith and Lord Kames, both of whom were instrumental in securing his appointment as Professor of Civil Law at Glasgow, a post he held until his death. He was a charismatic teacher who transformed the civil law curriculum by placing it on the jurisprudential foundations Smith had created for his moral philosophy lectures. He was a radical Foxite Whig and a member of the Society of the Friends of the People.

Millar is now much admired by historians of social thought for his *Origin of the Distinction of Ranks* (1771) in which he appeared to develop the sociological implications of Smith's account of the progress of civilization in a history of different systems of social authority. Unfortunately this view will not stand. The publication of the text of Smith's *Lectures on Jurisprudence* in 1978 showed that Millar's apparently original analysis, for all its closeness of texture and acuity, was intellectually entirely dependent on Smith's earlier work.

He never gave a separate course of lectures on political economy, and dealt with that subject in unpublished lectures on government whose character can be inferred from a series of essays first published in the posthumous edition of his *Historical View of the English Government* (1803). His Smithian interest in the natural history of property led him to formulate a distinctive theory of profit as the wage of the manufacturer plus the saving he derived from investing in the division of labour, a subject which also interested his pupil, the Earl of Lauderdale. The attraction of this theory lay in its radical political implications. It allowed Millar to show that the attributes of property ownership and personal independence which lay at the heart of contemporary ideas of political rights extended to all of those who participated in the productive relationships of a commercial society. It led him to campaign for the radical reform of parliament in order to adjust the old Whig constitution to the social and economic changes of the past century.

See Also

- Scottish Enlightenment

Selected works

1771. *The Origin of the Distinction of Ranks*, 4th edn. Edinburgh: William Blackwood; London: Longman, Hurst, Rees & Orme, 1806.

1803. *An Historical View of the English Government*. 4 vols. London: J. Mawman.

Bibliography

Igantieff, M. 1983. John Millar and individualism. In *Wealth and Virtue: The Shaping of Political Economy in the Scottish Enlightenment*, ed. I. Hont and M. Ignatieff. Cambridge: Cambridge University Press.

Lehmann, W.C. 1960. *John Millar of Glasgow. 1735–1801: His Life and Thought and His Contributions to Sociological Analysis*. Cambridge: Cambridge University Press.

How to cite this article

Phillipson, Nicholas. "Millar, John (1735–1801)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000174> doi:10.1057/9780230226203.1860

The New Palgrave Dictionary of Economics Online

Miller, Merton (1923 – 2000)

René M. Stulz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Merton Miller was at the centre of the transformation of academic finance from a descriptive field to a science. His principal contribution to this transformation was the introduction of arbitrage arguments which underlie most theoretical contributions in finance and remain central to the way financial economists analyse finance problems to this day. These arbitrage arguments underlie his and Franco Modigliani's famous irrelevance propositions.

Keywords

American Finance Association; arbitrage; bankruptcy costs; capital gains taxation; capital structure; contracting costs; corporate bonds; corporate debt; corporate finance; dividend policy; financial markets; interest rates; irrelevance propositions; Miller, M.; MM irrelevance propositions; Modigliani, F.; risk; tax subsidy; tax shield; taxation of corporate profits

Article

From the late 1950s to the early 1970s the field of finance changed fundamentally. A reader of the *Journal of Finance* in the early 1950s would find a field that was mostly descriptive. After the early 1970s the field had become a science. Merton Miller was at the centre of that transformation. His work started it in 1958. For the rest of his life he was at the heart of modern finance. (Grundy, 2001, provides a complete list of Merton Miller's publications.)
After obtaining a Ph.D. in economics from Johns Hopkins University in 1952 and a brief stay at the London School of Economics, he joined the Graduate School of Industrial Administration at what was then known as Carnegie Tech. As an assistant and associate professor there, he made the contributions to the theory of corporate finance with Franco Modigliani, another faculty member, that made him famous. He joined the University of Chicago in 1961. From Chicago he exerted a huge influence on finance which lasted until he died in 2000. Merton Miller's research had a prodigious impact – he made major

contributions in monetary economics, operations research, derivatives pricing, and asset pricing, as well as his seminal contributions in corporate finance – but his influence went far beyond the contributions of his papers. He mentored countless Chicago graduate students and faculty members from Chicago and throughout the profession. At times he played the role of the nurturing patriarch, while at other times he used his wit and intellect to keep people on the straight and narrow path of solid economic thinking. From ‘his’ seat on the left of the speaker in the Rosenwald seminar room, often in a worn-out sweater, he changed the course of numerous papers. Sometimes his intervention went further – for example, he was instrumental in persuading the *Journal of Political Economy* to publish the paper by Black and Scholes that is the foundation of option pricing theory. When he ventured outside of the University of Chicago, he often did so to be ‘an activist supporter of free-market solutions to economic problems’, as he stated in a brief Nobel autobiography (1991a). He knew how to make his case – he was not the son of an attorney, a Harvard graduate also, for nothing – and had a well-deserved reputation for unparalleled eloquence in the finance profession.

The irrelevance propositions and the role of arbitrage

Merton Miller earned a Nobel Prize in economics in 1990 for his ‘fundamental contributions to the theory of corporate finance’ (Franco Modigliani already had a Nobel Prize by then for his life cycle theory of saving). Just about every MBA in the world has learned the famous MM irrelevance propositions he developed with Franco Modigliani. (One paper had Modigliani's name first and the other had Miller's name first, so I will proceed using the moniker MM to represent the team.) The two key MM irrelevance propositions are developed in a world with perfect markets, so that there are no frictions. In particular, there are no transactions costs or taxes, and no costs are incurred to induce managers to maximize the value of the firm.

The first irrelevance proposition, Proposition I in the paper titled ‘The Cost of Capital, Corporation Finance and the Theory of Investment’ published in the *American Economic Review* (1958, p. 268) states that ‘the market value of any firm is independent of its capital structure and is given by capitalizing its expected return at the rate.•…•appropriate to its class’. The second irrelevance proposition concludes that ‘given a firm's investment policy, the dividend payout it chooses to follow will affect neither the current price of its shares nor the total return to its shareholders’ (1961, p. 414). In other words, in perfect markets neither capital structure choices nor dividend policy decisions matter. Since then, corporate finance has refined these results and built theories based on the existence of market imperfections.

If we had to remember one thing about Merton Miller's contributions to finance, what should it be? It would not be the irrelevance propositions themselves. Rather, it would be the way the irrelevance propositions were proved (for a more complete analysis, see Stulz, 2000). The approach used to prove these propositions is central to the thinking of practitioners of modern finance. It has spawned many seminal contributions to the field. The method used to prove Proposition I is the method of arbitrage. MM did not invent arbitrage, but made it the foundation of modern finance. MM assume that financial markets are perfect and then show that

> if Proposition I did not hold, an investor could buy and sell stocks and bonds in such a way as to exchange one income stream for another stream, identical in all relevant

> respects but selling at a lower price. The exchange would therefore be advantageous to the investor quite independently of his attitudes toward risk. As investors exploit these arbitrage opportunities, the value of the overpriced shares will fall and that of the underpriced shares will rise, thereby tending to eliminate the discrepancy between the market values of the firms. (1958, p. 269)

The arbitrage mechanism is how Merton Miller thought about finance phenomena. Results that would lead to arbitrage opportunities could not possibly be important because market forces would step in to make prices right. However, in his thinking arbitrage was never limited to existing financial instruments and institutions. For him, arbitrage opportunities that exist in the real world will eventually disappear because, when needed, financial innovations will occur that will prevent these opportunities from persisting.

Though arbitrage arguments are now pervasive throughout finance and, more generally, economics, the more immediate and direct impact of the arbitrage proof of Proposition I was to provide the foundation for modern corporate finance because it specifies sufficient conditions for leverage not to matter. Because of the proof, we know that, if financial markets are perfect, the value of a firm does not depend on its leverage. As a result, practitioners and academics alike know that, if leverage affects value, it must be that one or more of the assumptions required by the arbitrage proof do not hold.

In their papers MM eliminated once and for all the argument that leverage is costly simply because it increases the interest rate the corporation pays for its debt. As leverage increases in a world of perfect markets, the coupon paid on debt increases, but that is because bondholders bear more risk and must be compensated for this additional risk. This will happen even though the firm's cash flows are unaffected by the additional leverage. Hence, as Merton Miller pointed out in his Nobel lecture (Miller, 1991c), the increase in the risk of debt has no social costs because the firm's total risk is unaffected by the change in leverage.

Beyond the irrelevance propositions

With corporate income taxes, the cost of debt for the firm is the cost after taxes since interest paid on debt is tax deductible at the corporate level. If the only departure from the assumptions leading to Proposition I were a tax subsidy to corporate debt, one would expect firms to maximize the value of that subsidy and therefore have extremely high leverage. Empirically, however, leverage is not extreme. To make sense of the limited levels of leverage in the presence of what appeared to be a large tax subsidy for debt, finance had either to relax other assumptions leading to Proposition I or to conclude that the subsidy was illusory. Initially, the route chosen by finance was to take into account bankruptcy costs. Bankruptcy costs occur because contracting is costly – firms that default on their debt contracts cannot be costlessly reorganized. In the presence of bankruptcy costs and tax subsidy to debt, each firm has an optimal debt level such that the increase in the present value of expected bankruptcy costs resulting from an additional dollar of debt equals the present value of the expected tax subsidy from that additional dollar of debt.

Merton Miller always doubted that expected bankruptcy costs could be large enough to explain why firms did not take greater advantage of the tax subsidy of debt. His assessment of the evidence on bankruptcy and financial distress costs was that ‘neither empirical research nor simple common sense

could convincingly sustain these presumed costs of bankruptcy as a sufficient, or even as a major, reason for the failure of so many large, well-managed US corporations to pick up what seemed to be billions upon billions of dollars of potential tax subsidies' (1991b, p. 274). This assessment led him to one of his most memorable statements, namely, that 'the supposed trade-off between tax gains and bankruptcy costs looks suspiciously like the recipe for the fabled horse-and-rabbit stew – one horse and one rabbit' (1976, p. 264).

Since direct bankruptcy costs could not explain why firms were not taking advantage of the apparent tax subsidy of debt, the field of finance turned to other explanations for low leverage based on contracting costs. Jensen and Meckling (1976) showed that, as leverage increases, shareholders have incentives to take advantage of bondholders by undertaking highly risky projects with high payoffs to shareholders in some states even though such projects have a negative net present value. The bondholder–shareholder conflict identified by Jensen and Meckling makes debt more costly because firms either behave inefficiently as a result of leverage or spend real resources to convince bondholders that they will not take advantage of them. A large literature emphasizing contracting costs has developed over time.

Merton Miller always had doubts that the bondholder–shareholder conflict could explain why firms did not take greater advantage of the tax shield of debt. Not surprisingly, his scepticism stemmed from the role of arbitrage in his thinking. If the tax shield of debt was so large, why was it that investment bankers would not devise solutions that would enable firms to take advantage of this tax shield and overcome the agency costs of debt through clever contracting? As always, he viewed no finance problem as solved unless he could find a solution that would not provide clever arbitrageurs with profit opportunities.

In 1976, in his address as President of the American Finance Association, Merton Miller revisited the issue of the impact of corporate taxation on the MM irrelevance propositions in a classic paper titled 'Debt and Taxes'. This paper shows perhaps better than any of his other papers how he could use arbitrage arguments to change the way finance academics and practitioners understood how the world works. In that paper he pointed out that the tax advantage of corporate debt might be mostly if not completely illusory. Because interest on corporate debt is taxed as income to the bondholder, the interest paid must be sufficiently high to ensure that the after-tax income from holding corporate bonds is attractive relative to the income from equity which, when it accrues as capital gains, is taxed at a lower effective rate. While corporate interest payments generate tax deductions, personal taxes on interest income are higher than on capital gains, and so the before-tax cost of capital on debt must be higher than on equity to induce investors to hold debt. In his paper Merton Miller showed that under specific conditions the only feasible equilibrium is the one in which the after-tax cost of debt equals the after-tax cost of equity. When this equilibrium obtains, Proposition I holds in the presence of taxes, and no firm has a financial incentive to alter its mix of debt and equity even though interest payments on debt are tax deductible. 'Debt and Taxes' demonstrated that the perfect-markets assumptions are sufficient, but not necessary, conditions for leverage to be irrelevant. Showing that the assumptions required for Proposition I do not hold is not enough to conclude that leverage matters; rather it must also be the case that clever arbitrageurs cannot profit from the situation.

The legacy

With the contributions to the field of finance that I have described, Merton Miller provided a way to

think about financial phenomena that remains at the core of all major theoretical developments in the field. Throughout his life, Merton Miller used arbitrage reasoning to organize his thoughts about important phenomena. His first publication appeared in the 1948 *American Economic Review*. In 1990, he published a paper in the *Journal of Finance* (co-authored with David Hsieh) that analysed the impact on stock prices of changes in margin requirements. That paper was awarded a Smith–Breeden prize for best paper in the *Journal of Finance*. At the time, Merton Miller was thrilled because he had published refereed papers in top journals in five different decades. He never stopped wanting to write papers that merited publication in top journals. Three days before his death he was preparing a paper for submission. Throughout his life he was first, last, and foremost a scholar.

See Also

- arbitrage
- Modigliani, Franco
- Modigliani–Miller theorem

I am grateful for comments from Harry DeAngelo, Linda DeAngelo, Steven Durlauf and Andrew Karolyi.

Selected works

1948. (With R. Musgrave.) Built-in flexibility. *American Economic Review* 38, 122–8.

1958. (With F. Modigliani.) The cost of capital, corporation finance and the theory of investment. *American Economic Review* 48, 261–97.

1961. (With F. Modigliani.) Dividend policy, growth, and the valuation of shares. *Journal of Business* 34, 411–33.

1976. Debt and taxes. *Journal of Finance* 32, 261–75.

1990. (With D. Hsieh.) Margin regulation and stock market volatility. *Journal of Finance* 45, 3–29.

1991a. Autobiography. In *Les Prix Nobel. The Nobel Prizes 1990*, ed. Tore Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1990/miller-autobio.html, accessed 28 June 2006.

1991b. *Financial Innovations and Market Volatility*. Cambridge, MA and Oxford, UK: Blackwell.

1991c. Leverage. *Journal of Finance* 46, 479–88.

Bibliography

Grundy, B. 2001. M. H. Miller: his contributions to financial economics. *Journal of Finance* 56, 1183–206.

Jensen, M. and Meckling, W. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–60.

Stulz, R. 2000. Merton Miller and modern finance. *Financial Management* 29, 119–31.

How to cite this article

Stulz, René M. "Miller, Merton (1923–2000)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000345> doi:10.1057/9780230226203.1099

The New Palgrave Dictionary of Economics Online

Minard, Charles Joseph (1781 – 1870)

R.F. Hébert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost–benefit analysis; demand theory; Dupuit, A.-J.; income distribution; Jevons, W.; Minard, C.; Navier, L.; public works; Say, J.-B.; subjective measures of utility; substitution effect; value of time

Article

A French engineer and economist, Charles Joseph Minard was widely recognized as the creator of graphical statistics, a means of figuratively portraying railway traffic routes on illustrated maps. Minard served as professor at the École Nationale des Ponts et Chaussées (ENPC) in the 1830s, where he taught the course on interior navigation, which included roads, rivers, canals and railways. In 1831 Minard wrote a lengthy monograph designed to establish a course in economics that he proposed for ENPC students. Although Minard viewed this work as a manual for practising engineers, J. B. Say immediately recognized the manuscript as a systematic treatise on the economics of public works, and urged Minard to publish it for the benefit of economists as well as engineers. For reasons that are not entirely clear, Minard shelved his manuscript instead – probably owing to the delay by ENPC in establishing an economics chair until 1847. In 1850, a year before his retirement from public service, Minard published his ‘Notions élémentaires d’économie politique appliqué aux travaux publics’ in the *Annales des Ponts et Chaussées*.

In this monograph Minard explored such fundamental notions as utility, demand, opportunity costs, the value of time and services, the effects of taxes on income distribution, and the use of compound interest in calculating the value of capital expenditures – a treatment lauded by W. S. Jevons in his *Theory of Political Economy* (1871). Despite its unfortunate delay in publication, the ideas in Minard's monograph were clearly part of the oral tradition in economics at ENPC in the first half of the 19th century. Thus, Minard served as an important link between Navier and Dupuit in the development of demand theory and cost–benefit analysis. This claim is based on four major aspects of his work: he introduced subjective elements, such as the value of time, into the operational measure of utility; he insisted that the magnitude of social utility depends on the distribution of income; he recognized that price increases cause substitution effects among existing consumers and that price decreases draw new consumers into

the market; and he developed subjective notions of cost associated with public works.

See Also

- cost–benefit analysis
- demand theory
- Dupuit, Arsene-Jules-Emile Juvenal
- Jevons, William Stanley
- Navier, Claude Louis Marie Henri
- public works
- Say, Jean-Baptiste

Selected works

1850. Notions élémentaires d'économie politique appliqué aux travaux publics. *Annales des Ponts et Chaussées: Mémoires et Documents*, 2d ser. 19(1), 1–125.

1851. Motifs pour préférer dans les travaux publics des ouvrages moins coûteux, quoique moins durables. *Journal des Economistes* 21, 65–7.

Bibliography

Coronio, G. 1997. *250 ans de L'École des Ponts en cent portraits*. Paris: Presses de l'école des Ponts et Chaussées, 84–5.

Ekelund, R., Jr. and Hébert, R. 1978. French engineers, welfare economics, and public finance in the nineteenth century. *History of Political Economy* 10, 636–68.

Ekelund, R., Jr. and Hébert, R. 1999. *Secret Origins of Modern Microeconomics: Dupuit and the Engineers*. Chicago: University of Chicago Press.

Etner, F. 1987. *Histoire du calcul économique en France*. Paris: Economica.

Hébert, R. 1994. Fondements et développements de l'économie publique. *Revue Du Dix-Huitième Siècle* 26, 37–49.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan.

How to cite this article

Hébert, R. F. "Minard, Charles Joseph (1781–1870)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000385> doi:10.1057/9780230226203.1100

The New Palgrave Dictionary of Economics Online

Mincer, Jacob (1922 – 2006)

Finis Welch

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Jacob Mincer was one of the founding fathers of modern labour economics. Along with Gary Becker and T.W. Schultz, Mincer's ideas led to the evolution of labour economics as perhaps the premier applied field in economics. His work on personal income distributions and the associated wage–age profiles has dominated empirical research on these topics since the mid-1960s. The work extended to many related areas, most importantly the labour force participation of married women, the wage–age profiles associated with interrupted work careers, and migration decisions of two-career families.

Keywords

age–wage profiles; Becker, G.; labour economics; Mincer, J.; on the job training; returns to schooling; Schultz, T. W.; women's work and wages

Article

Born in Poland, Jacob Mincer was a college freshman in Czechoslovakia when the Germans invaded in early 1939. He spent most of the Second World War in prisons and concentration camps, but survived to enter Emory University in 1948 on an Hillel Foundation scholarship. After completing his first degree in two years, Mincer began his graduate studies in economics at the University of Chicago. He then transferred to Columbia University, having followed the lady who would be his wife from Chicago to New York for her residency in radiation oncology. Later, Flora Kaplan Mincer, MD, took six years from her practice to bring up their three children. That interruption in her career would be the basis for Mincer's subsequent paper, with his student Solomon Polachek (1974), which was the first to empirically tackle the complications of women's careers in earnings determination.
Jacob Mincer received his Ph.D. from Columbia in 1957, taught for two years at the City College of New York, and then returned to Columbia, where he remained until his 1991 retirement. In the interim, there were visiting appointments at the University of Chicago, the Stockholm School of Economics and the Hebrew University of Jerusalem.

Mincer was one of the very best 20th-century economists. He is one of the four or five who led the way into modern labour economics. The ideas of investing in man per se that were circulating in the early to mid-1950s had become environmental at Chicago and Columbia by the end of the decade. Given publication lags, it is impossible to know who came first, but the three papers that introduced the economics world to human capital were Theodore W. Schultz's ‘Capital Formation by Education’ (1960), Jacob Mincer's ‘Investment in Human Capital and Personal Income Distribution’ (1958) and Gary S. Becker's ‘Underinvestment in College Education?’ (1960).

Schultz argued simply that skills are malleable, that they are durable and acquired at a cost. As such they fit the capital formation rubric nicely. He also demonstrated that the opportunity costs of students who forgo work to remain in school in the aggregate are roughly equal to the costs of all purchased resources of schools and colleges. Soon afterwards he suggested that an extraordinarily large part of US per capita income growth in the first half of the 20th century was due to growth in education of the citizenry (1961). Mincer's 1958 *Journal of Political Economy* (*JPE*) paper was an extension of his thesis, which relied on the 1940, and 1950 decennial censuses. In this paper he challenged the traditional literature regarding income distributions that had focused only on the aggregate shape, with differences among individuals presumably owing only to luck and ability. After presenting a simple theory showing that with the discounted value of lifetime incomes constant, there would nonetheless be differences in income attributable to the time spent in both formal training and informal on-the-job training. The empirical support followed. In addition to laying the groundwork for what would become possibly *the* major area of empirical study in all of economics, he made the fundamental argument that the distribution of lifetime incomes is more, much more, equal than the point in time distributions. Becker simply showed that as a pure and simple investment for subsequent income alone, rates of return to a college education compare favourably to investments in physical capital. The traditional assumptions regarding the consumption and external benefits of education – for example, the ability to enjoy the arts, and so on, improved choices regarding health and life style and the externality of an informed citizenry – were not to be ignored: education is undoubtedly consumption in part, but there is also real productive value.

The introduction of the ideas of human capital to economics cumulated in T.W. Schultz's 1960 Presidential Address to the American Economics Association, ‘Investment in Human Capital’ (1961), followed by the collection of articles in the October 1962 *JPE* supplement headlined with Becker's, ‘Investment in Human Capital: A Theoretical Analysis' and Mincer's, ‘On the Job Training: Costs, Returns, and Some Implications’. The ideas presented in these and a few of the others in the supplement, for example, Stigler's article on search, triggered an intellectual excitement and enthusiasm that coloured almost all of labour economics for the two succeeding decades. During that period, labour economics became one of the foremost applied fields of economics.

Although Mincer matched his contemporaries in insight and imagination, his work is distinguished by his insistence on empirical applicability. An elegant case in point is the piece in the 1962 *JPE* supplement. Noting that age–wage profiles have a consistent tendency to rise rapidly early in a career, less rapidly thereafter, and then stabilize or decline slightly, he characterized the shape of the profile as the result of investment in learning on the job. He began by assuming that the individual has as an option a relatively flat profile equal in discounted value to the value of the observed profile. It follows that immediately after leaving school the difference between the flat profile and the lower actual one is an investment in higher subsequent higher wages. As such, the second period opportunity wage exceeds the

flat alternative by the return on the first period's investment, and so on. Assuming, further, that investment declines linearly during the early career, Mincer observed that if the rate of return on the investment in training is approximately equal to the rate used for calculating the flat alternative, then the two will intersect a number of years after leaving school, that is, approximately equal to the inverse of the rate of return. This is his famous ‘overtaking point’. Some invest heavily and have a steeply inclined wage profile, while others invest less and have a profile that increases less rapidly. Even so, if the rates of return are independent of the intensity of investment, the alternative paths will intersect at the overtaking point. These are simple descriptives that are easy for graduate students to follow. Maybe that is part of the reason Mincer is so revered. The ultimate pedagogic piece came in his 1974 book *Schooling, Experience and Earnings*.

In that most influential work, Mincer specified the details of ‘the’ human capital earnings equation. In it the left-hand variable is the logarithm of a rate of wages or earnings. The right-hand side has years of schooling (linearly) and a quadratic is years of work experience approximated by the number of years since leaving school. In the mid-1980s Kevin Murphy and I prepared a paper on empirical age earnings profiles. By way of introduction, we wanted an approximate count of the number of articles in economics journals that had used the Mincer specification. Once we saw it was well over 1,000, we gave up and simply noted that fact.

Mincer's early work on human capital and earnings was interspersed with work on the labour force participation of married women (1960a; 1960b; 1962a). As for his work on wages and experience, it set the stage for a voluminous literature to follow. As noted, he introduced work on earnings profiles of women that included interrupted work careers (1974a; 1974b; 1978a; 1979; 1980). Although he subsequently added excellent work on wage growth and job mobility, the nemesis of economists – the minimum wage, and economic growth – I believe that the most outstanding is the 1978 *JPE* paper, ‘Family Migration Decisions’, where he made the *ex post* obvious point regarding tried husband–wife movers, namely, in two-career families it is more difficult to find superior alternatives than it is for individuals or for one-career families. Moreover, the difficulty of finding superior alternatives increases as the specialization of the careers increases. As is true of most of his work, whether he pioneered or joined an existing literature, he greatly influenced what was to follow.

Mincer thought about and worked on important problems. He was original. You expected to learn any time you read a Mincer paper. Further, he always looked for applications: the theory had an empirical counterpart. Equally important, he was simply very good at what he did. He was an excellent colleague, teacher, and mentor to his doctoral students. He was also a great man. I am honoured to have known him.

Jacob Mincer retired from Columbia University in 1991. In 2002, the Institute for the Study of Labor (IZA) in Bonn awarded him the inaugural IZA Prize in Labor Economics. The prize was announced at his 80th birthday celebration hosted by Columbia University. In 2003 Mincer and Gary Becker were the inaugural recipients of the Society of Labor Economists (SOLE) Career Achievement Award. That award was then renamed the Jacob Mincer Award, in honour of the great man.

See Also

- returns to schooling
- women's work and wages

Selected works

1958. Investment in human capital and personal income distribution. *Journal of Political Economy* 66, 281–302.

1960a. Employment and consumption. *Review of Economics and Statistics* 42, 20–6.

1960b. Labor supply, family income and consumption. *Papers and Proceedings of the American Economic Association* 50, 574–83.

1962a. Labor force participation of married women. In *Aspects of Labor Economics*, ed. H.G. Lewis. Princeton: Princeton University Press.

1962b. On the job training: costs, returns, and some implications. In *Investment in Human Beings*, 2, U-NB Conference, printed in *Journal of Political Economy* 70(Suppl), 50–79.

1974a. Family investment in human capital: earnings of women, with S. Polacheck. *Journal of Political Economy* 82, 76–108.

1974b. *Schooling, Experience and Earnings*. New York: NBER.

1978. Family migration decisions. *Journal of Political Economy* 86, 749–73.

1978. (With S. Polachek.) Women's earnings reexamined. *Journal of Human Resources* 13, 118–34.

1979. (With H. Ofek.) Lifetime distribution of labor supply of married women. *Journal of Political Economy* 87, 197–202.

1980. Research in earnings and labor supply of women. In *Conference on Women in the Labor Market*, ed. C. Lloyd, E. Andrews and C. Gilroy. New York: Columbia University Press.

Bibliography

Becker, G.S. 1960. Underinvestment in college education? *American Economic Review* 50, 346–54. Repr. in *Problems of Economic Growth* ed. E. Phelps. New York: Norton, 1962.

Becker, G.S. 1962. Investment in human capital: a theoretical analysis. *Journal of Political Economy* 5, 9–49.

Schultz, T.W. 1960. Capital formation by education. *Journal of Political Economy* 6, 571–83.

Schultz, T.W. 1961. Education and economic growth. In *Social Forces Influencing American Education*, ed. N.B. Henry. Chicago: University of Chicago Press.

Schultz, T.W. 1961. Investment in human capital. *American Economic Review* 51, 1–17.

How to cite this article

Welch, Finis. "Mincer, Jacob (1922–2006)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000414> doi:10.1057/9780230226203.1101

The New Palgrave Dictionary of Economics Online

minimax regret

Jörg Stoye
From The New Palgrave Dictionary of Economics Online, , 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Minimax regret (Savage, 1951) is the principle of optimizing worst-case loss relative to some measure of unavoidable risk. In statistical decision theory, it provides a non-Bayesian alternative to minimax. It differs from minimax by fulfilling von Neumann–Morgenstern independence but exhibiting menu dependence. Minimax regret has seen occasional use in statistics, and implausible implications of minimax in certain economic problems recently led to its reconsideration by economists.

Keywords

decision theory; econometrics; estimation; maxmin; minimax; minimax regret; model uncertainty

Article

Minimax regret is the principle in statistical decision theory of optimizing worst-case efficiency loss relative to an ex post optimal decision. It was originally proposed in Savage's (1951) review of Wald (1950). In fact, Savage misinterpreted Wald (1950) and took it that he had proposed minimax regret rather than minimax; this was clarified in Savage (1954). The principle saw occasional use in statistics and machine learning (Das Gupta and Studden, 1991; Droge, 1998; Foster and Vohra, 1999) and recently enjoyed some revival in economics and econometrics, especially with regard to treatment choice (see below for references).

Definition and foundations

This article uses the same notation as the entries on minimax and on econometrics and decision theory; see either for elaborations. A minimax regret statistical decision rule minimizes (over D, the set of feasible decision rules)

$$\sup_{\theta\in\Theta}\left\{R(\delta,\theta)-\inf_{\delta^*\in D}R(\delta^*,\theta)\right\},$$

where *R* denotes a risk function. Thus, $R(\delta, \theta)$ is the expected loss incurred by decision rule δ as function of some unknown parameter value θ, and $\inf_{\delta^*\in D}R(\delta^*,\theta)$ indicates the lowest expected loss achievable given θ. Minimax regret differs from standard minimax by considering not loss in and of itself, but excess loss relative to this unavoidable risk. Intuitively, it thereby optimizes not against parameter values that are unfavourable to any decision rule, but against ones where a decision rule can cause great damage. Unlike with minimax, *D* enters the criterion function, hence minimax regret is menu dependent. Expanding *D* can affect the preferred decision rule even if the newly available rules are themselves unattractive.

Important variations are as follows. First, the preceding criterion is prior-free: the supremum is taken over all possible parameter values, without any concern for prior probabilities. The Γ-minimax regret criterion (Berger, 1985) takes the supremum with respect to a set of priors over Θ and thereby allows for compromises with Bayesianism. Second, the benchmark $\inf_{\delta^*\in D}R(\delta^*,\theta)$ could be defined via the ex post best among a subset D' of decision rules, a well-known example being Hannan regret (Cesa-Bianchi and Lugosi, 2006; Hannan, 1957). A close relative of minimax regret which enjoys some popularity in computer science is the competitive ratio, defined by taking the ratio rather than the difference to unavoidable risk (Borodin and El-Yaniv, 1998).

The prior-less minimax regret preference ordering was axiomatized by Milnor (1954) and Stoye (2006). However, menu dependence implies that preferences over decision rules that are not in fact chosen lack a behavioural or ‘revealed preference’ interpretation. Hayashi (2008) provides a revealed preference characterization of the Γ-minimax regret choice correspondence. Stoye (2007b) subsequently unifies the literature and considers prior-less minimax, Γ-minimax and Hannan regret. In either framework, the core message is that the trade-off between minimax and minimax regret can be cast as choice among two well-known axioms. Minimax avoids the aforementioned menu dependence but violates von Neumann–Morgenstern independence (for which see the entry on minimax); minimax regret fulfils independence but is menu dependent.

Mathematically, minimax regret is minimax with a respecified risk function, so remarks on finding minimax regret rules mirror the relevant remarks for minimax. Recent applications of game theory to identify finite sample minimax regret rules include Schlag (2007) and Stoye (2007c, 2007d). Asymptotic minimax regret efficiency is treated by Hirano and Porter (2006).

Applications

Minimax regret coincides with minimax loss if $\inf_{\delta^*\in D}R(\delta^*,\theta)$ is constant on Θ, as in the estimation example given in the entry on minimax. The criteria differ in the following, simple application. A

decision maker must assign one of two treatments $t \in \{0, 1\}$ to some population; treatments induce random outcomes (Y_0, Y_1) supported on {*success, failure*}. Treatment choice can condition on observations from a simple random sample of N subjects, half of whom were assigned to treatment $t=0$. Then the no-data rule that always assigns treatment 0 is minimax, as is any other decision rule. This is because any decision rule's risk is maximized – and all these maxima are identical – if both treatments induce only failures. The more natural rule that assigns everybody to the treatment that scores more successes in the sample (with even tie-breaking) is asymptotically (Hirano and Porter, 2008) as well as finite sample (Canner, 1970) minimax regret efficient, and essentially uniquely so (Stoye, 2007d).
The example illustrates a classic criticism (Savage, 1951) of minimax, namely that its 'ultra-pessimism' can lead to complete ignorance of data, as well as how minimax regret may avoid the problem. Extensions of this application are analyzed in Brock (2006), Manski (2004, 2007), Stoye (2007a, 2007c, 2007d), and Schlag (2007). Empirical applications of minimax regret to treatment choice are found in Eozenou, Rivas and Schlag (2006), Manski (2008) and Stoye (2009). In other applications to economics, Schlag (2003) brings minimax regret to bandit problems; Hansen (2005) evaluates kernel density estimators in terms of minimax regret; Bergemann and Schlag use minimax regret (2008) and Γ-minimax regret (2007) to analyse monopoly pricing; Chamberlain (2000) applies Γ-minimax regret to portfolio choice problems; and Hart and Mas-Colell (2001) use Hannan regret to evaluate learning rules.

Criticisms

The menu dependence of minimax regret has attracted criticism at least since Chernoff (1954). Other criticisms mirror those of the minimax principle, namely that minimax regret may implicitly optimize against unreasonable priors. It is worth noting that while minimax regret avoids no-data rules in natural examples, examples that go the other way can be constructed (Parmigiani, 1992). A natural example in which both principles inform no-data rules occurs if one modifies the above application by conditioning on a continuous covariate (Stoye, 2007d).

See Also

- decision theory in econometrics
- minimax

Bibliography

Bergemann, D. and Schlag, K.H. 2007. Robust monopoly pricing. Cowles Foundation Discussion Paper 1527R. New Haven, Conn: Yale University.

Bergemann, D. and Schlag, K.H. 2008. Pricing without priors. *Journal of the European Economic Association (Papers and Proceedings)* 6, 560–9.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.

Borodin, A. and El-Yaniv, R. 1998. *Online Computation and Competitive Analysis*. Cambridge and New York: Cambridge University Press.

Brock, W.A. 2006. Profiling problems with partially identified structure. *Economic Journal* 92, F427–F440.

Canner, P.L. 1970. Selecting one of two treatments when the responses are dichotomous. *Journal of the American Statistical Association* 65, 293–306.

Cesa-Bianchi, N. and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge: Cambridge University Press.

Chamberlain, G. 2000. Econometrics and decision theory. *Journal of Econometrics* 95, 255–83.

Chernoff, H. 1954. Rational selection of decision functions. *Econometrica* 22, 422–43.

Das Gupta, A. and Studden, W. 1991. Robust Bayesian experimental designs in normal linear models. *Annals of Statistics* 19, 1244–56.

Droge, B. 1998. Minimax regret analysis of orthogonal series regression estimation: selection versus shrinkage. *Biometrika* 85, 631–43.

Eozenou, P., Rivas, J. and Schlag, K.H. 2006. Minimax regret in practice: four examples on treatment choice. Discussion paper. Florence: European University Institute.

Foster, D. and Vohra, R. 1999. Regret in the on-line decision problem. *Games and Economic Behavior* 29, 7–36.

Hannan, J. 1957. Approximation of Bayes risk in repeated play. In *Contributions to the Theory of Games*, Vol. III, ed. M. Dresher, A.W. Tucker and P. Wolfe. Princeton, N.J.: Princeton University Press.

Hansen, B.E. 2005. Exact mean integrated squared error of higher-order kernels. *Econometric Theory* 21, 1031–57.

Hart, S. and Mas-Colell, A. 2001. A general class of adaptive strategies. *Journal of Economic Theory* 98, 26–54.

Hayashi, T. 2008. Regret aversion and opportunity-dependence. *Journal of Economic Theory* 139, 242–68.

Hirano, K. and Porter, J. 2008. Asymptotics for statistical treatment rules. Discussion paper. Tucson: University of Arizona.

Manski, C.F. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72, 1221–46.

Manski, C.F. 2007. Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics* 139, 105–15.

Manski, C.F. 2008. *Identification for Prediction and Decision*. Cambridge, Mass.: Harvard University Press.

Milnor, J. 1954. Games against nature. In *Decision Processes*, ed. R.M. Thrall, C.H. Coombs and R.L. Davis. New York: Wiley.

Parmigiani, G. 1992. Minimax, information and ultrapessimism. *Theory and Decision* 33, 241–52.

Savage, L.J. 1951. The theory of statistical decision. *Journal of the American Statistical Association* 46, 55–67.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Schlag, K.H. 2003. How to minimize maximum regret in repeated decisions. Discussion paper. Florence: European University Institute.

Schlag, K.H. 2007. Eleven: designing randomized experiments under minimax regret. Discussion paper. Florence: European University Institute.

Stoye, J. 2006. Statistical decisions under ambiguity. Discussion paper. New York: New York University.

Stoye, J. 2007a. Minimax regret treatment choice with incomplete data and many treatments. *Econometric Theory* 23, 190–9.

Stoye, J. 2007b. Axioms for minimax regret choice correspondences. Discussion paper. New York: New York University.

Stoye, J. 2007c. Minimax regret treatment choice with finite samples and missing outcome data. In *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, ed. G. de Cooman, J. Veinarová and M. Zaffalon. Prague.

Stoye, J. 2007d. Minimax regret treatment choice with finite samples. Discussion paper. New York:

New York University.

Stoye, J. 2009. Minimax regret treatment choice with missing data: an application to young offenders, *Journal of Statistical Theory and Practice*, forthcoming.

Wald, A. 1950. *Statistical Decision Functions*. New York: Wiley.

How to cite this article

Stoye, Jörg. "minimax regret." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_M000423> doi:10.1057/9780230226203.1896

The New Palgrave Dictionary of Economics Online

minimax

Jörg Stoye
From The New Palgrave Dictionary of Economics Online, , 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Minimax (Wald, 1950) is the principle in statistical decision theory of minimizing worst-case risk. It is the subject of a rich literature in statistics and saw occasional normative application in economics. Minimax is related to the maximin expected utility model (Gilboa and Schmeidler, 1989) in economics, an model of ambiguity aversion that was recently used to analyse model uncertainty.

Keywords

ambiguity; decision theory; econometrics; estimation; maxmin; minimax; minimax regret; model uncertainty

Article

Minimax is the principle in statistical decision theory of optimizing worst-case outcomes. The minimax principle was first formalized by Wald in a sequence of papers culminating in Wald (1950). In statistics, minimax estimators or decision rules have since become the objects of a rich literature. Minimax is related to maxmin expected utility, a leading model of ambiguity aversion in economic theory that recently became prominent as a way to approach model uncertainty. While not the focus of this article, Rawls' (1999, first edition 1971) use of minimax as component of a normative theory of justice deserves mention.

Minimax in statistics and economics

This entry uses the same notation (essentially due to Wald, 1950) as the one on econometrics and decision theory. A statistical experiment consists of a family of distributions $\{P_\theta : \theta \in \Theta\}$ over an outcome set Z. (Θ may be infinite dimensional, so the model underlying the experiment need not be

parametric in the usual sense.) The decision maker must pick an act *a* from some feasible set *A*, possibly at random, after observing a draw *z* from P_θ. A complete contingent plan for this decision maker can be summarized by a decision function $\delta: Z\times[0, 1]\rightarrow A$, where $\delta(z, u)$ assigns treatment conditional on observation *z* and randomization *u* (*u* is normalized to be drawn from a uniform (0, 1) distribution). The decision maker incurs loss $L(a, \theta)\geq 0$; a decision rule's risk function $R(\delta, \theta) = \int\int L(a, \theta)\, dP_\theta du$ maps possible parameter values onto expected losses.

The question is which decision rule to pick, given that there are typically many undominated or admissible ones. Under minimax the answer is to minimize $\sup_{\theta\in\Theta} \bullet R(\delta, \theta)$: that is, worst-case risk. The best known alternative is Bayesianism, that is, to minimize $\int R(\delta, \theta)\, d\pi$, where π is a prior over Θ. A compromise between the two is Γ-minimax (Berger 1985), which imposes a set of priors Γ over Θ and then minimizes $\sup_{\pi\in\Gamma}\int R(\delta, \theta)\, d\pi$, the maximal expected risk over Γ. This nests standard minimax as the extremal case where Γ contains all possible priors over Θ.

It is instructive to compare this with the ‘maxmin expected utility’ model (Gilboa and Schmeidler, 1989), in which a decision maker ranks acts according to $\min_{\pi\in\Gamma}\int u\circ f(s)\, d\pi(s)$. Here, *f* denotes an act and maps states of the world *s* into lotteries over ultimate outcomes *x*; Γ is a set of priors π; and *u* is a von Neumann–Morgenstern utility, thus $u\circ f(s) = \int U(x)\, df(s)$ for some utility function *U*. The notation is due to Anscombe and Aumann (1963) and is introduced in detail in this dictionary's entry on ambiguity.

These formalisms are related as follows. States of the world *s* correspond to parameter values θ. Losses $L(a, \theta)$ correspond to (negative) utility evaluations of outcomes $U(x)$: that is, they are already expressed in utility terms. Because of this, outcomes themselves, as well as acts, do not have a direct analogue in Wald's setting. Conversely, risk functions correspond not to any of Anscombe and Aumann's primitives, but to so-called utility acts $u\bullet\bullet\bullet f(s)$ that map states of the world into expected utilities and that play important roles in many axiomatic developments. Finally and most importantly, maxmin expected utility corresponds to Γ-minimax. The criterion function of classical minimax translates into the decision theoretic notation as $\min_{s\in S}\bullet u\bullet\bullet\bullet f(s)$.

Foundations of minimax

Foundations for minimax can be found in the axiomatic literature on decision theory. A natural starting point is Gilboa and Schmeidler's (1989) characterization of maxmin expected utility (and hence, Γ-minimax). The core insight behind this characterization concerns the following axioms for a preference ordering • over acts.

Independence: $f\bullet g$ iff $\alpha f+(1-\alpha)h\bullet \alpha g+(1-\alpha)h$ for all scalars $\alpha\in(0, 1)$ and acts *f*, *g*, *h*; here, $\alpha f+(1-\alpha)h$ denotes a statewise probabilistic mixture of acts.

C-independence: Like independence, but imposed only if *h* is constant: that is, *h* yields the same lottery in every state.

Uncertainty aversion: $f\sim g$ implies $\alpha f+(1-\alpha)g\bullet f$ for all $\alpha\in(0, 1)$.

The first of these axioms, von Neumann–Morgenstern (1947) independence, is crucial for characterizations of Bayesianism. Gilboa and Schmeidler (1989) replace it with the next two, weaker ones. Uncertainty aversion states that decision makers exhibit weak preference for mixtures, intuitively

because these constitute a hedging of bets across states. Any such strict preference would violate independence. C-independence limits the potential for such violations by reinstating independence whenever the mixing act is constant, intuitively because mixing with constant acts cannot generate a hedge.

The resulting characterization leaves Γ unspecified. This is appropriate from a 'revealed preference' point of view because sets of beliefs are not directly observable, but users of the statistical minimax criterion might desire axiomatizations that imply the according – that is, the maximal – specification of Γ. These were originated by Milnor (1954) and modernized and made comparable to Gilboa and Schmeidler (1989) by Stoye (2006). Specifically, Γ can be made maximal by adding a symmetry axiom (Arrow and Hurwicz, 1972; Cohen and Jaffray, 1980) that excludes any prior weighting of states and thereby eliminates any vestige of Bayesianism.

Finding minimax rules

No universal method for finding minimax decision rules exists, but a number of helpful ones are detailed in any statistics textbook. See for example Berger (1985) for an overview and Ferguson (1967) for an advanced treatment. A technique of special interest to economists is direct application of game theory (Wald, 1945).

Let π^* be a prior and δ^* a decision rule such that (i) $\int R(\delta^*, \theta)\, d\pi^* \geq \int R(\delta^*, \theta)\, d\pi$ for any prior π over Θ, thus π^* maximizes risk given δ^*; and (ii) δ^* is the Bayes rule relative to π^*. Then δ^* achieves minimax risk. π^* is also called a least favorable prior, and it can be instructive to think of (δ^*, π^*) as Nash equilibrium of a fictitious zero-sum game between the decision maker and a malicious Nature. The minimax theorem (von Neumann, 1928) gives conditions under which π^* exists; subsequent existence results for Nash equilibria imply other such conditions. The technique can also be extended to cases where π^* fails to exist; specifically, δ^* is minimax if there exists a sequence (δ_n, π_n) such that δ_n is Bayes relative to π_n and $\sup_{\theta \in \Theta} R(\delta^*, \theta) \leq \lim_{n \to \infty} \int R(\delta_n, \theta)\, d\pi_n < \infty$. Some other techniques for finding minimax rules – for example, minimaxity of a constant-risk Bayes rule – are corollaries.

As an example, let Z be binomial with parameters (θ, n) and let $L(a, \theta)=(a-\theta)^2$. Then $\delta(z) = (z + \sqrt{n}/2) / (n + \sqrt{n})$ can be shown to have constant risk and to be Bayes if π^* is a Beta $(\sqrt{n}/2, \sqrt{n}/2)$-distribution, hence it is a minimax estimator. Note that the sample analogue of θ, z/n, might appear a more natural estimator; δ shrinks it toward 1/2.

In more involved problems, finding exact minimax rules may not be feasible, and one may have to resort to asymptotic analysis (Le Cam, 1986). A classic result is that under certain conditions, Bayes as well as maximum likelihood estimators are locally asymptotically minimax.

Applications

A famous early application of minimax to estimation is Hodges and Lehmann (1950). A sizeable

literature developed from this and is surveyed in the textbooks mentioned above. Chamberlain (2008) applies minimax analysis to an instrumental variables model; a core result is that under normality and other conditions, maximum likelihood is (finite sample) minimax for some parameters. Chamberlain (2000) applies minimax to portfolio choice problems. Robust control in macroeconomics has a maxmin expected utility interpretation; see Hansen et al. (2006) and this dictionary's entry on ‘model uncertainty’. In economic theory, maxmin expected utility is an early benchmark in the large literature on ambiguity aversion; see this dictionary's entry on ‘ambiguity and ambiguity aversion’.

Criticisms of minimax

Criticisms of minimax centre on the facts that it may be perceived as extremely conservative and that it may optimize against an implausible prior. For example, π^* in the above example is much concentrated near 1/2. (Intuitively, values of θ close to 1/2 are unfavourable because they imply a large variance of the signal.) The sample analogue of θ accordingly underperforms against the minimax estimator if θ is indeed close to 1/2, but outperforms it by a much greater margin for θ near 0 or 1 and is generally considered more attractive. Also, minimax estimators need not be admissible; while admissible minimax rules exist under regularity conditions, the techniques described above might not identify them. Furthermore, it is easy to construct decision problems in which minimax decision rules ignore available data (Savage 1954), and in economics, recent work on treatment choice uncovered natural examples of this (Manski 2004). Users who are comfortable with priors can avoid some of these criticisms by using the Bayesian or Γ-minimax criteria, taking care to specify reasonable priors. Economists looking for non-Bayesian approaches recently explored minimax regret as an alternative (Manski 2004 and other references in the ‘minimax regret’ entry). Finally, it is not obvious how to adapt the minimax principle to dynamic decision problems; see the entry on ambiguity aversion.

See Also

- ambiguity and ambiguity aversion
- decision theory in econometrics
- minimax regret
- model uncertainty

Bibliography

Anscombe, F.J. and Aumann, R.J. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34, 199–205.

Arrow, K.J. and Hurwicz, L. 1972. An optimality criterion for decision-making under ignorance. In *Uncertainty and Expectations in Economics: Essays in Honour of G.L.S. Shackle*, ed. C.F. Carter and J. L. Ford. London: Blackwell.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.

Chamberlain, G. 2000. Econometric applications of maxmin expected utility. *Journal of Applied Econometrics* 15, 625–44.

Chamberlain, G. 2008. Decision theory applied to an instrumental variables model. *Econometrica* 75, 609–52.

Cohen, M. and Jaffray, J.-Y. 1980. Rational behavior under complete ignorance. *Econometrica* 48, 1281–99.

Ferguson, T. 1967. *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Hansen, L.P., Sargent, T.J., Turmuhambetova, G. and Williams, N. 2006. Robust control and model misspecification. *Journal of Economic Theory* 128, 45–90.

Hodges, J.L. and Lehmann, E.L. 1950. Some problems in minimax point estimation. *Annals of Mathematical Statistics* 21, 182–97.

Le Cam, L. 1986. *Asymptotic Methods in Statistical Decision Theory*. New York: Springer.

Manski, C.F. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72, 1221–46.

Milnor, J. 1954. Games against nature. In *Decision Processes*, ed. R.M. Thrall, C.H. Coombs and R.L. Davis. New York: Wiley.

Rawls, J. 1999. *A Theory of Justice*, rev. edn. Cambridge, Mass.: Harvard University Press.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Stoye, J. 2006. Statistical decisions under ambiguity. Discussion paper. New York: New York University.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320.

von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd edn. Princeton, N.J.: Princeton University Press.

Wald, A. 1945. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics* 46, 265–80.

Wald, A. 1950. *Statistical Decision Functions*. New York: Wiley.

How to cite this article

Stoye, Jörg. "minimax." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_M000417> doi:10.1057/9780230226203.1895

The New Palgrave Dictionary of Economics Online

minimum wages

Original by Donald O. Parsons; revised by Bruce Weinberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The minimum wage, the lowest wage rate legally payable by employers to workers, derives support from concern about the equity of market processes. Because employment may fall in response to an increase in the minimum wage and because the majority of low-wage workers do not come from families in poverty, the minimum wage may have modest benefits as a poverty reduction tool. While there are variations across studies, evidence from the United States suggests that the economy-wide employment effects of wage minimums at the levels at which they have been implemented in the United States are negative but not large.

Keywords

envelope theorem; hours worked; labour market participation; labour supply; minimum wages; poverty alleviation; unemployment

Article

The term minimum wages refers to various legal restrictions on the lowest wage rate payable by employers to workers. Until relatively recently, wage floors usually had a very specific focus; in Great Britain and the United States, for example, minimum wages were initially limited to women and children. Only following the Great Depression were such laws extended systematically to the general work force in many industrial and industrializing economies. The minimum wage restrictions were often industry specific, in France for example, extensions of trade union legislation (Rosa, 1981). In the United States, industry-specific wage restrictions were held to be unconstitutional; in 1938 a uniform national minimum wage rate was established for non-farm, non-supervisory personnel under the Fair Labor Standards Act. Subsequently, coverage was extended to the bulk of the labour force.
The social appeal of minimum wage legislation appears to be strong, its intuitive base rooted in concern about the equity of market processes. Dissatisfaction with the share of production allocated to the least

able members of the work force is prevalent even among individuals impressed with the enormous capacity of the market system to organize productive activity. An obvious solution to this problem, and one that can be implemented with a modest government budget commitment for statute enforcement, is to redefine the wage structure politically to achieve a socially preferable distribution of income. Although the political interests that have formed the most prominent support for minimum wage legislation may have had less socially oriented goals, for example, Colberg (1960) and Silberman and Durden (1978), broad public support for such legislation is, I believe, based on this equity issue and it is usually against the social criterion of poverty reduction that minimum wage legislation has been judged. Stigler (1946) provides the classic discussion of the potential deficiencies of minimum wage legislation as an antipoverty device; employment may fall more than in proportion to the wage increase from the minimum, thereby reducing earnings: wage rates in uncovered sectors may decrease more than those in the covered sector rise as the uncovered sector is forced to absorb the workers released by the covered sector: the impact of the legislation on family income distribution may be perverse unless the fewer but better jobs are allocated to members of needy families rather than to low-wage workers, most obviously teenagers, from wealthier families. A crucial insight by economists is that minimum wage legislation alters the opportunity set of the least able but does not unambiguously expand it. The legal restriction that employers cannot pay less than a specified wage is equivalent to the legal stipulation that workers cannot work at all in the protected sector unless they find employers willing to hire them at that wage. Much of the progress in the analysis of minimum wage effects in the last several decades has focused on the theoretical and empirical modelling needed to assess the welfare implications of this altered opportunity set.

As the theoretical modelling of the low-wage labour market has become more complete, theoretical predictions of minimum wage law effects have, unfortunately, become qualitatively ambiguous. Most models have been designed to capture the major features of minimum wage legislation in the United States, a uniform wage minimum covering a portion of a competitive economy. The principal implication of such models is that employment in the covered sector will fall with the establishment of an effective minimum wage. If labour supply is inelastic, these disemployed workers will seek and presumably find employment in the uncovered sector. The wages and well-being of workers in the uncovered sector might be expected to fall as that sector is forced to absorb additional workers (Stigler, 1946; Welch, 1974; 1978). Johnson (1969) demonstrates, however, that in a general equilibrium framework with two factors (labour and capital) the well-being of uncovered workers could in fact rise. If the covered sector is sufficiently capital intense and faces a sufficiently high demand elasticity, the quantity of capital released as the covered sector contracts could potentially increase the well-being of workers in the uncovered sector. The introduction of an elastic labour supply function (and implicitly or explicitly some valuable non-market activity) suggests additional parameters that must be estimated before theoretical considerations can be brought to bear on the assessment of the minimum wage (Welch, 1974).

The modelling of minimum wage effects on unemployment and labour force participation is more complex than on employment, requiring careful specification of the search process (Mincer, 1976). The effect of a minimum wage on unemployment, for instance, depends critically on the queuing method required to secure high paying jobs and on the optimal search strategy induced by this hiring regime. If, for example, workers must wait in a union hall to secure jobs in the covered sector, the extent of

unemployment will be quite different than if they can maintain their places in the queue for covered employment while in an uncovered sector job or while out of the labour force entirely.

Before turning to the empirical evidence on minimum wage effects, a brief comment on compliance is warranted. Legal compliance with the minimum wage laws in the United States appears to be surprisingly high (Ashenfelter and Smith, 1979). Effective evasion of small minimum wage restrictions, however, is probably quite high since wages are only a portion of the employment compensation package (Wessels, 1980). Non-wage benefits such as paid vacations are almost completely fungible. Indeed, the envelope theorem would suggest that modest adjustments among components in the total compensation package could be made without affecting employer costs or, equally important, worker welfare. Larger minimum wage restrictions would presumably raise covered worker welfare and employer costs, but not at the rate suggested by the wage-only compensation models.

Among other adjustments employers could make to an increase in the wage minimum would be an increase in effort demands or a reduction in the convenience (or number) of scheduled work hours. Perhaps of greater concern to economists is the potential for a reduction in the provision of on-the-job training to the young. The adverse training effects of legal minimum wages appear to be significant (Leighton and Mincer, 1981; Hashimoto, 1982), although perhaps partly offset by increased schooling in a broader picture (Mattila, 1981).

Clearly the effect of minimum wage laws on the wages and well-being of the labour force must be resolved empirically, either by estimation of the parameters in the theoretical models or by direct estimation of labour market effects. The latter approach has been the most common. Unfortunately, the evidence for the United States labour market (for which such estimation is most prevalent) is not as useful as one might hope. The political equilibrium in the United States has apparently kept the legal wage minimum relatively low. Only in a few circumstances has the minimum been so large as to induce major industrial contractions, for example in the South in the early years of the legislation (Colberg, 1960), and in Puerto Rico, most dramatically in that same period (Reynolds and Gregory, 1965). For most of the more recent period, the wage minimum has been primarily limited in impact to teenagers of both sexes and to adult females (Kneisner, 1981), both of which groups have significant non-market alternatives subject to their own exogenous forces.

The empirical literature on employment effects of the legal minimum wage in the United States suggests that the economywide employment effects of wage minimums at recent levels are negative but not large (Eccles and Freeman, 1982; Brown, Gilroy and Kohen, 1982). Most estimates are bounded by employment elasticities of minus 1 (a reduction in employment equiproportional to the increase in the wage minimum) and zero. Brown, Gilroy and Kohen argue for an estimate towards the zero portion of that range. The effects may, however, not be constant over a wider range of minimum wage restrictions; as the potential for substitution within the total compensation package is reduced, the employment effects will almost surely increase. Certainly minimum wage restrictions that are ‘large’ relative to customary wages appear to have very large effects, whether considered regionally (again Colberg, 1960; Reynolds and Gregory, 1965), or by economic sector (Fleisher, 1981).

Highly visible work by Card and Krueger (1994; 1995) has focused on ‘natural experiments’ generated by changes in the minimum wage. In 1992 the minimum wage was increased in New Jersey. Card and Krueger estimated the effect of the minimum wage on employment in fast-food restaurants in New Jersey compared with neighbouring Pennsylvania, where there was no increase in the minimum wage,

and found that employment increased in New Jersey relative to Pennsylvania. Kennan (1995) discussed this work and potential explanations and subsequent research by Neumark and Wascher (2000) questioned their results.

For most individuals not directly involved in buying or selling low skilled labour, the critical empirical question is not the magnitude of the employment effects of minimum wages but rather the effect on income poverty. Obviously large negative employment effects would suggest that the antipoverty effects of the minimum wage are small or possibly even perverse. Direct empirical studies of antipoverty effects (Gramlich, 1976; Parsons, 1980) indicate, however, that the antipoverty effects in the United States would be quite modest even if employment effects were zero. The great majority of low-wage workers do not come from families in poverty. Moreover, the groups primarily affected, teenagers and low-skilled adult females, are predominantly part-time workers and any wage-rate effect on earnings and income is strictly proportional to hours worked. Even a fully effective wage minimum with no offsetting employment adjustment would provide little relief to poverty-level families (Parsons, 1980). Negative employment effects simply enhance other fundamental limitations of minimum wage legislation as a poverty programme.

Wage rate restrictions alone appear to be an unsatisfactory solution to social concerns about labour market outcomes. Politically manipulating the price system seems like a direct and inexpensive method of assisting the disadvantaged. Almost surely it is not. Employment opportunities and the factors that limit labour market participation must be considered as well as wage rates if market outcomes are to be supplanted in a socially satisfactory way for low-skilled workers.

See Also

- labour economics

Bibliography

Ashenfelter, O. and Smith, R.S. 1979. Compliance with the minimum wage law. *Journal of Political Economy* 87, 333–50.

Brown, C., Gilroy, C. and Kohen, A. 1982. The effect of the minimum wage on employment and unemployment. *Journal of Economic Literature* 20, 487–528.

Card, D. and Krueger, A.B. 1994. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84, 772–93.

Card, D. and Krueger, A.B. 1995. *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton, NJ: Princeton University Press.

Colberg, M.R. 1960. Minimum wage effects on Florida's economic development. *Journal of Law and Economics* 3, 106–17.

Eccles, M. and Freeman, R.B. 1982. What! Another minimum wage study? *American Economic Review* 94, 226–32.

Fleisher, B.M. 1981. *Minimum Wage Regulation in Retail Trade*. Washington, DC: American Enterprise Institute.

Gramlich, E.M. 1976. Impact of minimum wages on other wages, employment, and family incomes. *Brookings Papers on Economic Activity* 1976(2), 409–51.

Hashimoto, M. 1982. Minimum wage effects on training on the job. *American Economic Review* 72, 1070–87.

Johnson, H.J. 1969. Minimum wage laws: a general equilibrium analysis. *Canadian Journal of Economics* 2, 599–604.

Kennan, J. 1995. The elusive effects of minimum wages. *Journal of Economic Literature* 33, 1950–65.

Kneisner, T.J. 1981. The low-wage workers: who are they? In Rottenberg (1981).

Leighton, L. and Mincer, J. 1981. The effects of minimum wages on human capital formation. In Rottenberg (1981).

Mattila, J. 1981. The impact of minimum wages on teenage schooling and part-time full-time employment of youths. In Rottenberg (1981).

Mincer, J. 1976. Unemployment effects of minimum wages. *Journal of Political Economy* 84, S87–S104.

Neumark, D. and Wascher, W. 2000. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: comment. *American Economic Review* 90, 1362–96.

Parsons, D.O. 1980. *Poverty and the Minimum Wage*. Washington, DC: American Enterprise Institute.

Reynolds, L.G. and Gregory, P. 1965. *Wages, Productivity, and Industrialization in Puerto Rico*. Homewood, IL: Richard D. Irwin.

Rosa, J.J. 1981. The effects of minimum wage regulation in France. In Rottenberg (1981).

Rottenberg, S. (ed.) 1981. *The Economics of Legal Minimum Wages*. Washington, DC: American Enterprise Institute.

Silberman, J. and Durden, G.C. 1978. Determining legislative preferences on the minimum wage: an economic approach. *Journal of Political Economy* 84, 317–29.

Stigler, G.J. 1946. The economics of minimum wage legislation. *American Economic Review* 36, 358–65.

Welch, F. 1974. Minimum wage legislation in the United States. *Economic Inquiry* 12, 285–318.

Welch, F. 1978. *Minimum Wages: Issues and Evidence*. Washington, DC: American Enterprise Institute.

Wessels, W.J. 1980. The effect of minimum wages on fringe benefits: an expanded model. *Economic Inquiry* 18, 293–13.

How to cite this article

Parsons, Donald O. and Bruce A. Weinberg. "minimum wages." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000176> doi:10.1057/9780230226203.1102

The New Palgrave Dictionary of Economics Online

Mirrlees, James (born 1936)

Gareth D. Myles

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Sir James Mirrlees has been influential in several areas. With Little he contributed to development economics through ‘the manual’, a practical guide to the use of cost–benefit analysis. He developed the theory of optimal income taxation and in so doing introduced the concept of incentive compatibility. With Diamond he revolutionized the theory of commodity taxation. His work on the principal–agent problem characterized contracts designed to counter moral hazard. The analytical tools he pioneered have benefited every area of economics to which they have been applied.

Keywords

asymmetric information; border prices; commodity taxation; cost–benefit analysis; endogenous growth theory; incentive compatibility; income taxation and optimal policies; Mirrlees, J.; monotone likelihood ratio condition; optimal growth; optimum income tax; principal–agent problem; production efficiency lemma; revelation principle; shadow pricing; single-crossing condition; technical change; value-added taxation; World Bank

Article

Professor Sir James Mirrlees was born in Scotland in 1936 and educated at Edinburgh University and Trinity College, Cambridge. He held academic posts at Trinity College and at Nuffield College, Oxford, and was awarded the Nobel Prize in economics in 1996 for his work on optimal income taxation and its extension to information and incentive problems in general. Mirrlees also made important contributions to growth theory, development economics and public economics.

Growth and development

The initial work of Mirrlees focused upon technical progress in models of economic growth. Kaldor and

Mirrlees (1962) assumed technical progress was embodied in new investment with the growth rate of productivity per worker operating on new machines an increasing concave function of the growth rate of investment per worker. The incorporation of externalities between different firms' investment decisions made this paper a precursor of the literature on endogenous growth theory. The problem of optimal growth in an economy subject to deterministic technical change was discussed in Mirrlees (1967). An extension to stochastic diffusion in continuous time showing that increased uncertainty would often lead to more saving rather than less was given in Mirrlees's Ph.D. dissertation and circulated in unpublished work (Mirrlees, 1965). These themes were also addressed in Mirrlees (1974a).

Mirrlees contributed to development economics via the influential Little and Mirrlees (1968; 1974) handbook of project appraisal ('the manual'). The manual was a practical guide to the use of cost–benefit analysis designed to contribute to improvements in the economic conditions of developing countries. It took as its starting point the use of shadow prices to value all inputs and outputs, regardless of whether they were marketable or non-marketable, and showed how shadow prices should be determined. In particular, it emphasized the use of border prices to value inputs and outputs when the project was located in a small country. When goods were not traded, it provided methods for valuing them based on the prices of traded goods. The manual emphasized that investment finance was scarce because of the government's budget constraint, so social profits should be discounted at the internal rate of return for the marginal investment project. The manual also studied constraints upon policy choices and how these affected shadow prices.

The recommendations of the manual provided a simple but powerful methodology. Since its publication they have been subjected to much theoretical scrutiny that has generally confirmed their validity. The practical impact of the Little–Mirrlees approach can be judged from the number of donor agencies that adopted it to guide their decisions. Foremost among these was the World Bank, where cost–benefit analysis was the dominant decision-making method throughout the 1970s. Its use has steadily declined since, which is attributed by Little and Mirrlees (1994) to the changing nature of lending and the internal institutional structure of the World Bank.

Income taxation

In a seminal paper Mirrlees (1971, p. 175) studied 'what principles should govern an optimum income tax; what such a tax schedule would look like; and what degree of inequality would remain once it was established'. Addressing this question required a model that included a motive for the redistribution of income because of endogenously generated inequality, incentive effects in labour supply and a justification for using an income tax rather than lump-sum taxation. The success of Mirlees's model was that it managed to capture all this but remained tractable and allowed the optimum tax to be characterized. No better model of income taxation has yet been proposed, although new results are still being discovered within the original framework and its specializations – see, for example, Diamond (1998), Saez (2001) and Hashimzade and Myles (2007).

The paper demonstrated that an optimum income tax leads to an allocation in which pre-tax income is increasing with ability, and that the marginal tax rate is between zero and one. Furthermore, unemployment is possible at the optimum and, when it occurs, is of the lowest-ability workers. The numerical analysis provides some of the most surprising findings. The optimal marginal rate of tax is low, at least compared with the rates applied in many countries at the time the paper was written.

Furthermore, the marginal tax rate is fairly constant, so the tax function is close to being linear. These results motivated Mirlees's observation (1971, p. 207) that ‘I had expected the rigorous analysis of income-taxation in the utilitarian manner to provide an argument for high tax rates. It has not done so.’ The analysis of the model required Mirrlees to formulate and solve a series of novel theoretical problems. In doing so he developed a series of techniques that have since become standard tools of economic analysis. In the income tax problem the government must offer the workers a budget constraint along which each chooses an optimal location through utility maximization. Since the budget constraint can be nonlinear it is possible for there to be multiple optimal choices for a worker, so choice cannot be represented by a demand function. The fundamental contribution of the paper was to show how this problem could be circumvented by viewing the government as selecting an allocation (an income–consumption pair) for each worker. If every worker prefers his allocation to that of any other, then each will willingly select the allocation intended for him. This is the notion of incentive compatibility: a worker of ability level *s* must find that the allocation designed for someone of this ability gives at least as much utility as the allocation designed for any other ability *s’*. The government then conducts its optimization over the set of incentive-compatible allocations. The imposition of incentive compatibility reduces the set of feasible allocations and is responsible for the second-best nature of the optimum tax.

The paper also showed that the problem can be reduced further if workers’ preferences over allocations are consistently related to ability. The restriction upon preferences introduced in Mirrlees (1971) has since become known as the single-crossing condition and implies that at every point in income–consumption space the indifference curve of a high-ability worker is flatter than that of a low-ability worker. Under single crossing, incentive compatibility requires high-ability workers to earn higher incomes and enjoy higher levels of consumption. The single-crossing condition has since found countless applications in problems involving the design of contracts for populations with agents of differing characteristics. With a continuum of consumers, it is not practical to state the incentive compatibility constraints directly. Mirlees (1971) surmounted this problem in a simple but ingenious way by showing that incentive compatibility is equivalent to utility being maximized at a worker's true skill level. The first-order condition for this optimization generates a differential equation that determines the evolution of utility as a function of ability. The differential equation can be used as a constraint on the optimization. This technique has since become known as the first-order approach to ‘maximization subject to maximization’. The first-order condition is necessary but not sufficient, so there exists the possibility that the tax function arising from the optimization analysis may violate the monotonicity requirement. A direct solution to this problem is to incorporate the second-order condition into the optimization (see Ebert, 1992). The 1971 income tax paper appreciated this issue, and the limitations of the first-order approach remained an issue that was addressed further in Mirlees's later work on the principal–agent problem.

That the optimum involved monotonicity implied an important observation: those with higher skills earn and consume more, so, although the government cannot directly observe skill, in equilibrium it can infer skill from income. Hence, given the optimum tax function, the announcement by a consumer of an income level is just a proxy for the direct announcement of a level of skill. This observation was later formalized in the revelation principle (Dasgupta, Hammond and Maskin, 1979; Myerson, 1979) that shows it is possible to replace the income tax with an equivalent direct mechanism in which each consumer announces a skill level and, furthermore, announcing the true skill level is a dominant

strategy. The revelation principle is now applied routinely in the analysis of incentive problems.

Commodity taxation

Diamond and Mirrlees (1971a; 1971b) revolutionized the theory of commodity taxation. The papers clarify the separation between consumer and producer prices and show that the choice of untaxed commodity is just a normalization that plays no role in determining the optimum allocation. They were among the first to employ the emerging duality methods and used the indirect utility function to phrase the problem in terms of the after-tax consumer prices that were the natural choice variables. As well as these innovations, the commodity taxation papers contain two fundamental results. The first is the simple rule of thumb that the imposition of an optimum commodity tax system requires an equal proportionate reduction in compensated demand for all commodities. This conclusion emphasizes that the real effect of a tax system is on consumers' demands and that the effect on prices is of secondary importance. The second result, now known as the production efficiency lemma, is more surprising and of significant practical value for policy.

The production efficiency lemma states that the optimum commodity tax system results in an equilibrium that is on the frontier of the production set. There are some limitations to this result, most notably non-constant returns to scale, which imply that achieving efficiency may require some firms to be shut down, thus adversely affecting their owners' incomes. Such restrictions are clarified in Mirrlees (1972). The policy value of the lemma follows from observing that efficiency is only possible if there are no distortions in the input prices faced by producers. Input taxes should not therefore be a feature of the optimum set of commodity taxes, implying that intermediate goods should not be taxed. This observation justifies the use of value-added taxation with tax rebates available for producers who purchase intermediate goods. It also suggests that capital held by firms should not be subject to taxation, though dividends paid to consumers can and probably should be, along with their realized capital gains.

Theoretically, the production efficiency lemma is especially surprising when contrasted with the conclusions of Lipsey and Lancaster's (1956) second-best theory. The central message of Lipsey and Lancaster was that a distortion in any sector of the economy should generally be offset by introducing distortions in all other sectors. This finding had achieved great prominence at the time the Diamond–Mirrlees article was published. In contrast, the lemma states that, even when distortionary taxes and subsidies are being introduced into consumer decisions in order to redistribute real income or to finance public goods, there is no reason to distort producer decisions. This special case runs counter to the general message of Lipsey–Lancaster.

Principal – agent

The third area to which Mirrlees made a fundamental contribution is the principal–agent problem that arises when one party wishes another to undertake an act on his or her behalf. If the act undertaken cannot be observed directly and its consequences observed only with some random error, then moral hazard can occur: the agent can attempt to hide behind the randomness to take an action which is less costly to the agent but which yields a lower expected return to the principal. Such a problem can arise in any economic relationship based on contingent contracts, for example between the owner and the

manager of a firm. Mirrlees (1974b; 1975) analysed the problem facing the principal in designing a contract that provides an incentive to the agent to take the action that yields the highest expected payoff to the principal. There are considerable analytical similarities between the design of this contract and the choice of an optimum income tax. These similarities arise because the principal is choosing the contract to maximize expected payoff subject to the agent choosing an action to maximize his or her payoff. This leads again to a situation of maximization subject to maximization and its analysis via incentive compatibility.

When the agent must choose from a finite set of actions the incentive compatibility constraints can be employed directly. This is impractical for the continuous case where there would be an uncountable infinity of constraints. Consequently, it again becomes necessary to use the first-order conditions for the agent's choice problem as a constraint on the optimization of the principal. Although this had been used prior to Mirrlees's analysis of the principal–agent problem (Zeckhauser, 1970), it had not been noticed that the approach might fail to generate the optimum. This possibility was made very clear in Mirrlees (1975), which provided an example where the first-order approach failed to generate the optimum and proceeded to discuss how the problem could be overcome. The method proposed identified the possible maxima and incorporated them as constraints into the optimization. This method works but has proved unwieldy in practice, so most analyses rely on the first-order approach despite its known weaknesses. These issues were explored even further in Mirrlees (1986) and in Mirrlees and Roberts (1980).

A further issue that arises in principal–agent relationships is the conditions that guarantee the reward from the contract is monotonic: that is, the payment to the agent increases as observed output increases. If there are only two possible output levels, monotonicity arises naturally. With three possible output levels, monotonicity can easily fail (Grossman and Hart, 1983). Mirrlees (1976) introduced the monotone likelihood ratio condition that is sufficient for monotonicity. This condition requires that actions that are more costly for the agent to undertake make more profitable outcomes relatively more likely. Although weaker conditions are available (Jewitt, 1988), the monotone likelihood ratio condition has become another essential component of the economic theorist's toolkit. It is, of course, closely related to the single-crossing property that plays such an important role in the income tax paper.

The work of Mirrlees has contributed to the understanding of economic policy via the manual and the papers on tax policy. His work also laid the foundation for the analysis of incentive problems in the presence of asymmetric information. Taken together, incentive compatibility, the extension of the first-order approach, the single-crossing property and the monotone likelihood ratio condition provide the basic tools that no economic theorist can be without. There has not been a single area of economics in which they have not been used to great advantage.

See Also

- incentive compatibility
- income taxation and optimal policies
- neoclassical growth theory
- principal and agent (i)
- principal and agent (ii)
- revelation principle

Selected works

1962. (With N. Kaldor.) A new model of economic growth. *Review of Economic Studies* 29, 174–90.

1965. Optimal capital accumulation under uncertainty. Unpublished manuscript.

1967. Optimum growth when technology is changing. *Review of Economic Studies* 34, 95–124.

1968. (With I. Little.) *Manual of Industrial Project Analysis in Developing Countries, Vol. II: Social Cost–Benefit Analysis*. Paris: OECD.

1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

1971a. (With P. Diamond.) Optimal taxation and public production I: Production efficiency. *American Economic Review* 61, 8–27.

1971b. (With P. Diamond.) Optimal taxation and public production II: Tax rules. *American Economic Review* 61, 261–78.

1972. On producer taxation. *Review of Economic Studies* 39, 105–11.

1974a. Optimal allocation under uncertainty. In *Allocation Under Uncertainty*, ed. J. Drèze. London: Macmillan.

1974b. Notes on welfare economics, information and uncertainty. In *Essays in Equilibrium Behavior under Uncertainty*, ed. M. Balch, D. McFadden and S. Wu. Amsterdam: North-Holland.

1974. (With I. Little.) *Project Appraisal and Planning for Developing Countries*. London: Heinemann.

1975. The theory of moral hazard and unobservable behaviour, Part I. Mimeo. Oxford: Nuffield College. Published in *Review of Economic Studies* 66 (1999), 3–21.

1976. The optimal structure of incentives and authority within an organization. *Bell Journal of Economics* 7, 105–31.

1980. (With K. Roberts.) Functions with multiple maxima. Mimeo. Oxford: Nuffield College.

1986. The theory of optimal taxation. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M. Intriligator. Amsterdam: North-Holland.

1994. (With I. Little.) The costs and benefits of analysis: project appraisal and planning twenty years on. In *Cost–Benefit Analysis*, ed. R. Layard and S. Glaister. Cambridge: Cambridge University Press.

Bibliography

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules: some general results in incentive compatibility. *Review of Economic Studies* 46, 185–216.

Diamond, P. 1998. Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates. *American Economic Review* 88, 83–95.

Ebert, U. 1992. A reexamination of the optimal nonlinear income tax. *Journal of Public Economics* 49, 47–73.

Grossman, S. and Hart, O. 1983. An analysis of the principal–agent problem. *Econometrica* 51, 7–45.

Hashimzade, N. and Myles, G. 2007. The structure of the optimal income tax in the quasi-linear model. *International Journal of Economic Theory* 3, 5–33.

Jewitt, I. 1988. Justifying the first-order approach to principal–agent problems. *Econometrica* 56, 1177–90.

Lipsey, R. and Lancaster, K. 1956. The general theory of second best. *Review of Economic Studies* 24, 11–32.

Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47, 61–73.

Saez, E. 2001. Using elasticities to derive optimal income tax rules. *Review of Economic Studies* 68, 205–29.

Zeckhauser, R. 1970. Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* 2, 10–26.

How to cite this article

Myles, Gareth D. "Mirrlees, James (born 1936)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000346> doi:10.1057/9780230226203.1103

The New Palgrave Dictionary of Economics Online

Mises, Ludwig Edler von (1881 – 1973)

Murray N. Rothbard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Böhm-Bawerk, E. von; business cycles; cash-balance analysis; central banking; Cuhel, F.; Currency School; equation of exchange; fractional reserve banking; German Historical School; Great Depression; Hayek, F. A. von; inflation; laissez-faire; marginal utility of money; Mises, L. E. von; money; money supply; neutrality of money; praxeology; price level; purchasing power parity; socialism; socialist calculation debate

Article

Mises was born in Lemberg, Austria-Hungary, on 29 September 1881 and died in New York City on 18 October 1973. The son of a Viennese construction engineer for the Austrian railroads, Mises enrolled in the University of Vienna in 1900. He earned his doctorate in law and economics in 1906, after which he became a leading member of Böhm-Bawerk's famous seminar at the university. From 1913 to 1934, Mises taught as an unpaid *Privatdozent* at the University of Vienna, conducting a seminar on economic theory. From 1909 to 1934, he was an economist for the Vienna Chamber of Commerce, serving as the principal economic adviser to the Austrian government.

Disturbed at encroaching Nazi influence in Austria, Mises accepted a professorship at the Graduate Institute of International Studies in Geneva, where he taught from 1934 to 1940, after which he emigrated to New York City. Mises became a visiting professor at New York University in 1948, where he continued to teach a seminar on economic theory until he retired in 1969, spry and energetic at the age of 87.

Mises' multifaceted achievements in economic theory built upon the insights and methodology of the Menger–Böhm-Bawerk Austrian School of economics. In contrast to the Jevons and Walras branches of marginal utility theory, the Austrians engaged in a logical analysis of the action of individuals, their major focus on a step-by-step process analysis rather than on the necessarily unreal world of static general equilibrium. Furthermore, 'cause', for the Austrians, was a unilinear 'causal-genetic' flow from individual utilities and actions to price, rather than the familiar neoclassical mutual determination of mathematical functions.

Mises' first pioneering accomplishment was to extend Austrian analysis to money. In his *Theory of Money and Credit* (1912) he succeeded in integrating money into micro-theory, demonstrating how the marginal utility of money interacts with utilities of other goods and with the supply of money to determine money prices. In doing so, Mises solved 'the problem of the Austrian circle', a formidable obstacle for any causal-genetic theorist. Since money, unlike other goods, is demanded not for its own sake but to purchase other goods in exchange, a demand to purchase and hold money must assume a pre-existing purchasing power in terms of other goods. How, then, can one explain the existence of that purchasing power; that is, of money prices? In his 'regression theorem', Mises, building on Menger's insights into the origin of money, demonstrated that the demand for money can be pushed back logically to the 'day' before the money-commodity became money, when it had purchasing power only as a commodity valuable in barter. Hence, every money must originate on the market as a valuable non-monetary commodity and cannot begin by being imposed by the state, or in an ad hoc social contract. There were many other notable contributions in *Money and Credit*. Though superficially similar to the quantity theory of money, Mises' process analysis demonstrated the inevitable non-neutral impact of money on relative prices and incomes. Indeed, he levelled a devastating critique of such neutral-money concepts as Fisher's equation of exchange and the idea of stabilizing 'the price level'. Moreover, Mises developed a cash-balance analysis, independently of the Cambridge School and on an individualistic rather than an aggregative and holistic basis. And before Gustav Cassel, Mises set forth a purchasing-power parity theory of exchange rates under fiat money, based on a Ricardian array of goods rather than on Cassel's price-level approach (Wu, 1939, pp. 115–16, 126–7, 232–5).

Money and Credit also revived the Ricardian–Currency School insight that no quantity of the money supply can be more optimal than any other. Since money's sole function is to exchange, an increase in its quantity can only dilute the purchasing power of each money unit and can confer no social benefit. Mises concluded that fractional reserve banking, or 'circulation credit', is inflationary and distorts prices and production. He showed the ideal banking system to be 100 per cent reserves of bank notes and demand deposits to standard gold or silver. On the other hand, eight years before C.A. Phillips (Phillips, 1920), Mises showed that any individual bank is necessarily severely restricted in expanding credit, so that the abolition of central banking would go far to eliminate the problem of inflationary banking. Finally, in analysing marginal utility, Mises incorporated the insights of the Czech Franz Cuhel (1907), a fellow member of Böhm-Bawerk's seminar, to demonstrate that marginal utility can in no sense be a measurable, mathematical quantity. Instead, it can only be a strictly ordinal subjective preference ranking; hence there can be no 'total utility' as an integral of marginal utilities. There can only be varying marginal utilities depending on the size of the 'margin', the actual unit of human choice. Although two semesters of Böhm-Bawerk's seminar were devoted to discussing *Money and Credit*, the older Austrians resisted this new development (Mises, 1978, pp. 59–60). Mises proceeded to found his own 'neo-Austrian' school, centred in his renowned biweekly private seminar at the Chamber of Commerce. Leading participants and followers included F.A. Hayek, Fritz Machlup, Gottfried von Haberler, Oskar Morgenstern, Wilhelm Ropke, Richard von Strigl, Alfred Schutz, Felix Kaufmann, Erich Voegelin, Georg Halm, Paul Rosenstein-Rodan and Lionel Robbins.

During the 1920s Mises developed (from its beginnings in *Money and Credit*) his notable theory of the business cycle, one of the few to be integrated with general micro-theory (Mises, 1923–31). Formed out of the Currency School, Böhm-Bawerk's theory of capital and Wicksell's distinction between natural and loan rates of interest, Mises' 'monetary malinvestment' theory sees the boom–bust cycle as the

inevitable product of inflationary credit expansion. This expansion artificially lowers interest rates and induces unsound overinvestments in higher-order capital goods, as well as underinvestment in consumer goods. Any cessation of credit expansion reveals the malinvestments and the lack of sufficient savings, and the ensuing recession liquidates the distortions of the boom and restores a healthy economy.

Mises founded the Austrian Institute for Business Cycle Research in 1926, and his cycle theory later won attention as an explanation of the Great Depression. His most important student and follower, F.A. Hayek, who had elaborated on the theory, emigrated to the London School of Economics in 1931 and strongly influenced a rising generation of English economists. Unfortunately, most of this influence was swept away in the flush of enthusiasm for the Keynesian Revolution.

When socialism emerged after the First World War, Mises wrote a classic article (Mises, 1920; 1922), demonstrating that a socialist government could not calculate economically and therefore could not organize a complex industrial economy. For two decades, socialists in Europe tried to rebut Mises' contentions, but not only had he anticipated their objections, he explicitly refuted them in the late 1940s (Mises, 1949; Hoff, 1949). If socialism could not calculate, and state interventionism only creates problems in the name of solving them (Mises, 1929), then the only viable and truly prosperous economy is laissez-faire. In a century marked by accelerating statism and collectivism, Mises stood out among scholars as an uncompromising stalwart of laissez-faire (Mises, 1927).

Austrian economists had virtually begun with defence of economic theory against the German Historical School (Menger, 1883). Amidst a rising tide of logical positivism, Mises now set forth and elaborated 'praxeology', the methodology of Nassau Senior and of the Austrians (Bowley, 1937). In contrast to the physical sciences, economic laws are discovered by logical deduction from self-evident axioms, such as that human beings exist and pursue goals. Praxeology develops the logical implications of the fact of individual human action (Mises, 1933; 1949). Historical events are the complex resultants of many causal factors; they are not simple, homogeneous events that can, as in the positivist schema, be used to 'test' theory. Instead, prior theory must be used to explain and understand history (Mises, 1957; Robbins, 1932; Kirzner, 1960).

In the culmination of his life-work, Mises put his methodological precepts into practice by constructing a systematic edifice of economic theory, completing the neo-Austrian integration of micro- and macroeconomics. First published in German in 1940, this monumental treatise was refined and expanded in his English-language *Human Action* (Mises, 1949). Some notable features were a resurrection of Fetter's pure time-preference theory of interest; a theory of subjective costs; and a dynamic emphasis on profit-and-loss as the motive power of the economy, and on profit as a reward for successful entrepreneurial forecasting.

Even though an exile late in life, the trend of the world and of academia against him, and remaining only a visiting professor, Mises maintained his good cheer and productivity and gradually built up a new group of followers in the United States. Since his death there has been a veritable renaissance of interest in his thought and works, including the establishment of an institute in his name at Auburn University (Moss, 1976; Andrews, 1981; Kirzner, 1982; Rothbard, 1973).

Selected works

1912. *The Theory of Money and Credit*, 3rd English edn. Indianapolis: Liberty Classics, 1981.

1920. Economic calculation in the socialist commonwealth. In *Collectivist Economic Planning: Critical Studies on the Possibilities of Socialism*, ed. F. von Hayek. London: Routledge & Sons, 1935.

1922. *Socialism: An Economic and Sociological Analysis*, 3rd English edn. Indianapolis: Liberty Classics, 1981.

1923–31. *On the Manipulation of Money and Credit*. Ed. P.E. Greaves, Dobbs Ferry, NY: Free Market Books, 1978.

1927. *Liberalism*, 3rd English edn. Irvington-on-Hudson, NY: Foundation for Economic Education, 1985.

1929. *A Critique of Interventionism*. New Rochelle, NY: Arlington House, 1977.

1933. *Epistemological Problems of Economics*. New York: New York University Press, 1981.

1940. *Nationalökonomie: Theorie des Handelns und Wirtschaftens*. Geneva: Editions Union.

1944a. *Bureaucracy*. New Haven: Yale University Press.

1944b. *Omnipotent Government: The Rise of the Total State and Total War*. Springs Mill, PA: Libertarian Press, 1985.

1949. *Human Action: A Treatise on Economics*, 3rd edn. Chicago: Regnery, 1966.

1957. *Theory and History: An Interpretation of Social and Economic Evolution*, 2nd edn. Auburn, AL: Ludwig von Mises Institute, 1985.

1978. *Notes and Recollections*. South Holland, IL: Libertarian Press.

Bibliography

Andrews, J.K., ed. 1981. *Homage to Mises: The First Hundred Years*. Hillsdale, MO: Hillsdale College Press.

Bien, B., ed. 1969. *The Works of Ludwig von Mises*. Irvington-on-Hudson, NY: Foundation for Economic Education.

Bowley, M. 1937. *Nassau Senior and Classical Economics*. New York: Kelley, 1949.

Cuhel, F. 1907. *Zur Lehre von den Bedurfrissen. Theoretische Untersuchungen über das Grenzgebiet der Ökonomik und Psychologie*. Innsbruck: Wagner.

Hoff, T.J.B. 1949. *Economic Calculation in the Socialist Society*. London: William Hodge.

Kirzner, I.M. 1960. *The Economic Point of View: An Essay in the History of Economic Thought*. Princeton, NJ: Van Nostrand.

Kirzner, I.M., ed. 1982. *Method, Process, and Austrian Economics: Essays in Honor of Ludwig von Mises*. Lexington, MA: Lexington Books.

Menger, C. 1883. *Problems of Economics and Sociology*. Ed. L. Schneider, Urbana, IL: University of Illinois Press, 1963.

Mises, M. von. 1976. *My Years with Ludwig von Mises*, 2nd enlarged edn. Cedar Falls, IA: Center for Futures Education, 1984.

Moss, L., ed. 1976. *The Economics of Ludwig von Mises: Toward a Critical Reappraisal*. Kansas City: Sheed & Ward.

Phillips, C.A. 1920. *Bank Credit*. New York: Macmillan.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Rothbard, M. 1973. *The Essential von Mises*, 3rd edn. Washington, DC: Ludwig von Mises Institute, 1983.

Wu, C.-Y. 1939. *An Outline of International Trade Theories*. London: Routledge.

How to cite this article

Rothbard, Murray N. "Mises, Ludwig Edler von (1881–1973)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000178> doi:10.1057/9780230226203.1104

The New Palgrave Dictionary of Economics Online

Misselden, Edward (fl. 1608 – 1654)

Douglas Vickers
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balance of trade; free trade; market price; Misselden, E.; rate of interest

Article

Edward Misselden, merchant-economist, held a number of appointments, including that of Deputy Governor of the Merchant Adventurers' Company at Delft, 1623–33, and representative of the Merchant Adventurers and the East India Company in various trade negotiations. His economic writings stem from his testimony before the Standing Commission on Trade appointed in 1622. His *Free Trade, or the means to make trade flourish* (1622) attributes the 'decay of trade' to the 'undervaluation of His Majesty's coin', 'the want of money', 'the excess of … consuming the commodities of foreign countries', particularly luxury goods (against which he proposed sumptuary laws), the export of bullion by the East India Company, and the decay and inadequate enforcement of regulation of the cloth trades. His proposal to remedy the shortage of money by increasing the denomination of the coin would, he acknowledged, raise general commodity prices, but this would be offset by the 'quickening of trade in every man's hand' that would result from the 'plenty of money'. Landlords and creditors could be protected by requiring that 'contracts made before the raising of monies shall be paid at the value the money went at when the contracts were made'.

Misselden's *Circle of Commerce, or the Balance of Trade* (1623), a long rejoinder to Malynes's attack on his earlier work, effectively sorts out the relation between commodity trade and the international exchange rate. 'It is not the rate of exchange, whether it be higher or lower, that maketh the price of commodities dear or cheap … but it is the plenty or scarcity of commodities, their use or non-use, that maketh them rise or fall in price.' He recognized that actual market prices may deviate from what might be thought to be intrinsic or par values.

Misselden's impressive work on the definition and computation of the balance of trade (explicitly including the earnings from re-exports, profits on fisheries, and freight income) contained an estimate for the year 1621–22, made by multiplying the five per cent customs revenue by 20 to obtain trade volume data. He pointed to the idea of the self-balancing international mechanism in his claim that

‘there is a fluxus and refluxus, a flood and ebbe of the monies of Christendom traded within itself: for sometimes there is more in one part … less in another, as one country wanteth and another aboundeth’. While Misselden advocated a high degree of governmental intervention in the economy, particularly in the granting of exclusive international trading privileges and the regulation of quality standards in domestic trade, he generally opposed the encouragement of monopolies. The free market analogy, which reappears frequently in his arguments, pointed also to the theory of the rate of interest: ‘As it is the scarcity of money that maketh the high rates of interest, so the plenty of money will make interest low better than any statute for that purpose.’ Misselden's various proposals to correct the ‘decay of trade’ shared the widespread concern for the state of the ‘idle poor’. What was given away as charity, he proposed, should be ‘orderly collected and prudently ordered for the employment of the poor’.

See Also

- mercantilism

Selected works

1622. *Free Trade, or the means to make trade flourish*. London.

1623. *The Circle of Commerce, or the Balance of Trade*. London.

Bibliography

Letiche, J.M. 1959. *Balance of Payments and Economic Growth*. New York: Harper.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

How to cite this article

Vickers, Douglas. "Misselden, Edward (fl. 1608–1654)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000179> doi:10.1057/9780230226203.1105

The New Palgrave Dictionary of Economics Online

Mitchell, Wesley Clair (1874 – 1948)

G. Moore

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; history of economic thought; Mitchell, W.C.; National Bureau of Economic research

Article

Wesley C. Mitchell was born in Rushville, Illinois, on 5 August 1874 and died on 29 October 1948. Most of his professional life was spent at Columbia University (1913–19, 1922–44) and as Director of Research at the National Bureau of Economic Research in New York (1920–45).
Mitchell's principal contribution to economic theory was indirect – through the emphasis he placed throughout his working life upon the need for close interaction between the development of hypotheses and testing their conformity to fact. This emphasis was explicit both in his own work on business cycles and in the research that he promoted and guided in many fields at the National Bureau. In his final report as Director (1945) he said:

> We like to think of ourselves as helping to lay the foundations of an economics that will consist of statements warranted by evidence a competent reader may judge for himself.… Speculative systems can be quickly excogitated precisely because they do not require the economist to collect and analyze masses of data, to test hypotheses for conformity to fact, to discard those which do not fit, to invent new ones and test them until, at long last, he has established a factually valid theory.

One of the hypotheses that Mitchell formulated has generated one of the longest continued and most widely applied scientific experiments in the field of economics. The hypothesis is that in free enterprise economies business cycles are generated by the continuous interaction of economic activities which lead or lag one another by varying intervals and which differ in amplitude of fluctuation by varying amounts. The processes have been identified and their leads and amplitudes measured on an *ex ante* as well as *ex post* basis. Historical records covering a century or more have been used to test the hypothesis. Private research institutes, governmental and international agencies in more than 30 countries have set up the

statistical apparatus required to test the hypothesis, keep the information up-to-date, and derive economic forecasts from it. The information is generally summarized in the form of leading, coincident and lagging indexes.

The types of economic processes that Mitchell considered crucial to his hypothesis were largely identified in his first major work on the subject, *Business Cycles* (1913). The variables included costs, prices and profits; investment decisions and investment expenditures; employment, income and consumption expenditures; interest rates, the volume of money, credit, and bank reserves; inventories and sales. In the original and subsequent treatises he observed how economic agents reacted to changes in economic conditions and how these reactions in turn affected others. To measure leads and lags he defined business cycles in such a way that their peaks and troughs could be dated. Measures of timing, amplitude and rates of change in successive cycles were devised, and summarized across cycles to find out what patterns were typical. Patterns of change within the cycle enabled Mitchell to test whether what happened during one phase had a bearing on what happened in the next, and whether the repetitive sequences corresponded with his expectations based upon economic practices and institutions.

Among the economic processes in business cycles that Mitchell stressed was the imbalance that develops between costs and prices. As an expansion in business activity proceeds, costs of production begin to rise faster than prices. This reduces profit margins and dims the outlook for future profits. This in turn prompts cutbacks in decisions to invest, and leads to reductions in sales, output, and employment. As a recession develops, costs as well as prices increase less rapidly or are reduced, but cost reduction soon begins to exceed price reduction, enhancing prospects for profits and incentives to invest. This in turn helps to bring the recession to an end and get recovery started. Mitchell and others have looked into such questions as what kinds of costs and prices behave in this manner, why they do so, and whether or not the phenomenon is widespread. Comprehensive data bearing on the matter have only become available in recent years, but they show that the pattern has continued to emerge in market-oriented economies some 70 years after Mitchell gave it a central position in his hypothesis about the self-generating character of business cycles.

One of Mitchell's great objectives was to construct a general theory of business cycles consistent with the facts of cyclical experience that he took such pains to observe and record. His respect for the value of economic theory was demonstrated not only by this objective, but also by his long concern with the history of economic thought. For many years he taught a famous course at Columbia University called Types of Economic Theory, and lecture notes taken stenographically by students were subsequently published under that title (1949). The lectures traced the historical origins of economic theories and related their development to particular legal, political, social, and economic institutions and events. Such was his interest in theory that in 1941 Mitchell allowed the theoretical portion (Part III) of his 1913 volume, *Business Cycles*, to be re-published under the title *Business Cycles and Their Causes* (1941). This early effort to construct a dynamic theory, indeed, serves well as an interpretation of Mitchell's last work, *What Happens During Business Cycles* (1951). But the self-generating theory of business cycles, which was the hallmark of Mitchell's ideas on the subject, remained and still remains to be written.

Selected works

1903. *A History of the Greenbacks, with Special Reference to the Economic Consequences of their Issue: 1862–65*. Chicago: University of Chicago Press.

1913. *Business Cycles*. Berkeley: University of California Press.

1915. The making and using of index numbers. *Bulletin of the US Bureau of Labor Statistics* No. 173 (July), 5–114.

1923. (With others) *Business Cycles and Unemployment*. New York: McGraw-Hill.

1927. *Business Cycles: The Problem and its Setting*. New York: National Bureau of Economic Research.

1937. *The Backward Art of Spending Money, and Other Essays*. New York: McGraw-Hill.

1938. (With A.F. Burns.) *Statistical Indicators of Cyclical Revivals*. Bulletin No. 69. New York: NBER.

1941. *Business Cycles and Their Causes*. Berkeley: University of California Press.

1945. *The National Bureau's First Quarter-Century*. 25th Annual Report of the National Bureau of Economic Research. New York: NBER.

1946. (With A.F. Burns.) *Measuring Business Cycles*. New York: NBER.

1949. *Lecture Notes on Types of Economic Theory*. New York: Kelley.

1951. *What Happens during Business Cycles: A Progress Report*. New York: NBER.

Bibliography

Burns, A.F., ed. 1952. *Wesley Clair Mitchell: The Economic Scientist*. New York: NBER.

Mitchell, L.S. 1953. *Two Lives: The Story of Wesley Clair Mitchell and Myself*. New York: Simon & Schuster.

How to cite this article

Moore, G. "Mitchell, Wesley Clair (1874–1948)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000182> doi:10.1057/9780230226203.1106

The New Palgrave Dictionary of Economics Online

mixed strategy equilibrium

Mark Walker and John Wooders
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A mixed strategy is a probability distribution one uses to randomly choose among available actions in order to avoid being predictable. In a *mixed strategy equilibrium* each player in a game is using a mixed strategy, one that is best for him against the strategies the other players are using. In laboratory experiments the behaviour of inexperienced subjects has generally been inconsistent with the theory in important respects; data obtained from contests in professional sports conforms much more closely with the theory.

Keywords

competition; equilibrium; game theory; minimax strategy; mixed strategy equilibrium; price dispersion; pure strategy; quantal response equilibrium; reinforcement learning

Article

In many strategic situations a player's success depends upon his actions being unpredictable. Competitive sports are replete with examples. One of the simplest occurs repeatedly in soccer (football): if a kicker knows which side of the goal the goalkeeper has chosen to defend, he will kick to the opposite side; and if the goalkeeper knows to which side the kicker will direct his kick, he will choose that side to defend. In the language of game theory, this is a simple 2×2 game which has no pure strategy equilibrium.

John von Neumann's (1928) theoretical formulation and analysis of such strategic situations is generally regarded as the birth of game theory. Von Neumann introduced the concept of a *mixed strategy*: each player in our soccer example should choose his Left or Right action randomly, but according to some particular binomial process. Every *zero sum* two-person game in which each player's set of available strategies is finite must have a *value* (or *security level*) for each player, and each player must have at least one *minimax* strategy – a strategy that assures him that, no matter how his opponent plays, he will achieve at least his security level for the game, in expected value terms. In many such games the minimax strategies are pure strategies, requiring no mixing; in others, they are mixed strategies.

John Nash (1950) introduced the powerful notion of *equilibrium* in games (including non-zero-sum games and games with an arbitrary number of players): an equilibrium is a combination of strategies (one for each player) in which each player's strategy is a *best* strategy for him against the strategies all the other players are using. An equilibrium is thus a sustainable combination of strategies, in the sense that no player has an incentive to change unilaterally to a different strategy. A *mixed-strategy equilibrium* (MSE) is one in which each player is using a mixed strategy; if a game's only equilibria are mixed, we say it is an MSE game. In two-person zero-sum games there is an equivalence between minimax and equilibrium: it is an equilibrium for each player to use a minimax strategy, and an equilibrium can consist only of minimax strategies.

An example or two will be helpful. First consider the game tic-tac-toe. There are three possible outcomes: Player A wins, Player B wins, or the game ends in a draw. Fully defining the players' possible strategies is somewhat complex, but anyone who has played the game more than a few times knows that each player has a strategy that guarantees him no worse than a draw. These are the players' respective minimax strategies and they constitute an equilibrium. Since they are *pure strategies* (requiring no mixing), tic-tac-toe is not an MSE game.

A second example is the game called 'matching pennies'. Each player places a penny either heads up or tails up; the players reveal their choices to one another simultaneously; if their choices match, Player A gives his penny to Player B, otherwise Player B gives his

penny to Player A. This game has only two possible outcomes and it is obviously zero-sum. Neither of a player's pure strategies (heads or tails) ensures that he won't lose. But by choosing heads or tails randomly, each with probability one-half (for example, by ‘flipping’ the coin), he ensures that in expected value his payoff will be zero *no matter how his opponent plays*. This 50–50 mixture of heads and tails is thus a minimax strategy for each player, and it is an MSE of the game for each player to choose his minimax strategy.

Figure 1 provides a matrix representation of matching pennies. Player A, when choosing heads or tails, is effectively choosing one of the matrix's two rows; Player B chooses one of the columns; the cell at the resulting row-and-column intersection indicates Player A's *payoff*. Player B's payoff need not be shown, since it is the negative of Player A's (as always in a zero-sum game). Matching pennies is an example of a 2×2 game: each player has two pure strategies, and the game's matrix is therefore 2×2.

Figure 1

		Player B	
		H	T
Player A	H	–1	1
	T	1	–1

Figure 2 depicts our soccer example, another 2×2 MSE game. The kicker and the goalie simultaneously choose either Left or Right; the number in the resulting cell (at the row-and-column intersection) is the probability a goal will be scored, given the players’ choices. The probabilities capture the fact that for each combination of choices by kicker and goalie the outcome is still random – a goal is less likely (but not impossible) when their choices match and is more likely (while not certain) when they don't. The specific probabilities will depend upon the abilities of the specific kicker and goalie: the probabilities in Figure 2 might represent, for example, a situation in which the kicker is more effective kicking to the left half of the goal than to the right half. For the specific game in Figure 2 it can be shown that the kicker's minimax strategy is a 50–50 mix between Left and Right and the goalie's minimax strategy is to defend Left 3/5 of the time and Right 2/5. The reader can easily see that the value of the game is therefore 3/5, that is, in the MSE

the kicker will succeed in scoring a goal 60 per cent of the time.

Figure 2

		Goalkeeper	
		L	R
Kicker	L	0.4	0.9
	R	0.8	0.3

Non-zero-sum games and games with more than two players often have mixed strategy equilibria as well. Important examples are decisions whether to enter a competition (such as an industry, a tournament, or an auction), ‘wars of attrition’ (decisions about whether and when to exit a competition), and models of price dispersion (which explain how the same good may sell at different prices), as well as many others.

How do people actually behave in strategic situations that have mixed strategy equilibria? Does the MSE provide an accurate description of people's behaviour? Virtually from the moment Nash's 1950 paper was distributed in preprint, researchers began to devise experiments in which human subjects play games that have mixed strategy equilibria. The theory has not fared well in these experiments. The behaviour observed in experiments typically departs from the MSE in two ways: participants do not generally play their strategies in the proportions dictated by the game's particular MSE probability distribution; and their choices typically exhibit negative serial correlation – a player's mixed strategy in an MSE requires that his choices be independent across multiple plays, but experimental subjects tend instead to switch from one action to another more often than chance would dictate. Experimental psychologists have reported similar ‘switching too often’ in many experiments designed to determine people's ability to intentionally behave randomly. The evidence suggests that humans are not very good at behaving randomly.

The results from experiments were so consistently at variance with the theory that empirical analysis of the concept of MSE became all but moribund for nearly two decades, until interest was revived by Barry O'Neill's (1987) seminal paper. O'Neill pointed out that

there were features of previous experiments that subtly invalidated them as tests of the theory of mixed strategy equilibrium, and he devised a clever but simple experiment that avoided these flaws. Although James Brown and Robert Rosenthal (1990) subsequently demonstrated that the behaviour of O'Neill's subjects was still inconsistent with the theory, the correspondence between theory and observation was nevertheless closer in his experiment than in prior experiments.

Mark Walker and John Wooders (2001) were the first to use field data instead of experiments to evaluate the theory of mixed strategy equilibrium. They contended that, while the rules and mechanics of a simple MSE game may be easy to learn quickly, as required in a laboratory experiment, substantial experience is nevertheless required in order to develop an understanding of the strategic subtleties of playing even simple MSE games. In short, an MSE game may be easy to play but not easy to play *well*. This fact alone may account for much of the theory's failure in laboratory experiments.

Instead of using experiments, Walker and Wooders applied the MSE theory to data from professional tennis matches. The ‘serve’ in tennis can be described as a 2×2 MSE game exactly like the soccer example in Figure 2: the server chooses which direction to serve, the receiver chooses which direction to defend, and the resulting payoff is the probability the server wins the point. Walker and Wooders obtained data from matches between the best players in the world, players who have devoted their lives to the sport and should therefore be expert in the strategic subtleties of this MSE game. Play by these world-class tennis players was found to correspond quite closely to the MSE predictions. Subsequent research by others, with data from professional tennis and soccer matches, has shown a similar correspondence between theory and observed behaviour.

Thus, the empirical evidence to date indicates that MSE is effective for explaining and predicting behaviour in strategic situations at which the competitors are experts and that it is less effective when the competitors are novices, as experimental subjects typically are. This leaves several obvious open questions. In view of the enormous disparity in expertise between world-class athletes and novice experimental subjects, how can we determine, for specific players, whether the MSE yields an appropriate prediction or explanation of their play? And when MSE is not appropriate, what *is* a good theory of play? We clearly need a generalization of current theory, one that includes MSE, that tells us in addition when MSE is ‘correct’, and that explains behavior when MSE is not correct. Moreover, the need for such a theory extends beyond MSE games to the theory of games more generally.

A more general theory will likely comprise either an alternative, more general notion of equilibrium or a theory of out-of-equilibrium behaviour in which some players may, with enough experience, come to play as the equilibrium theory predicts. Recent years have seen research along both lines. Among the most promising developments are the notion of quantal response equilibrium introduced by Richard McKelvey and Thomas Palfrey (1995), the theory of level-*n* thinking introduced by Dale Stahl and Paul Wilson (1994), and the idea of reinforcement learning developed by Ido Erev and Alvin Roth (1998).

See Also

- game theory
- game theory in economics, origins of
- Nash, John Forbes
- purification
- quantal response equilibria
- von Neumann, John

Bibliography

Brown, J. and Rosenthal, R. 1990. Testing the minimax hypothesis: a re-examination of O'Neill's game experiment. *Econometrica* 58, 1065–81.

Erev, I. and Roth, A.E. 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88, 848–81.

McKelvey, R. and Palfrey, T. 1995. Quantal response equilibria for normal-form games. *Games and Economic Behavior* 10, 6–38.

Nash, J.F. 1950. Equilibrium points in N person games. *Proceedings of the National Academy of Sciences* 36, 48–9.

O'Neill, B. 1987. Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the National Academy of Sciences* 84, 2106–9.

Stahl, D. and Wilson, P. 1994. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization* 25, 309–27.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320.

Walker, M. and Wooders, J. 2001. Minimax play at Wimbledon. *American Economic Review* 91, 1521–38.

How to cite this article

Walker, Mark and John Wooders. "mixed strategy equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000405> doi:10.1057/9780230226203.1107

The New Palgrave Dictionary of Economics Online

mixture models

Bruce G. Lindsay and Michael Stewart
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses statistical models involving mixture distributions. As well as being useful in identifying and describing sub-populations within a mixed population, mixture models are useful data-analytic tools, providing flexible families of distributions to fit to unusually shaped data. Theoretical advances since the mid-1970s, as well as advances in computing technology, have led to the widespread use of mixture models in ecology, machine learning, genetics, medical research, psychology, reliability and survival analysis. In particular, recent advances in non-linear time series involving mixtures have helped explain various features of financial and econometric data that more traditional models cannot capture.

Keywords

ARCH models; autoregressive models; bootstrap; convex models; density estimation; estimation; expectation-minimization (EM) algorithm; finite mixture distributions; hypothesis testing; maximum likelihood; mixture autogressive (MAR) models; mixture models; nonlinear time series; random variables; time series analysis

Article

Suppose that $\mathcal{F} = \{F_\theta : \theta \in S\}$ is a parametric family of distributions on a sample space *X*, and let *Q* denote a probability distribution defined on the parameter space *S*. The distribution

$$F_Q = \int F_\theta \, dQ(\theta)$$

is a mixture distribution. An observation *X* drawn from F_Q can be thought of as being obtained in a two-step procedure: first, a random Θ is drawn from the distribution *Q* and then, conditional on $\Theta = \theta$, *X* is drawn from the distribution F_θ. Suppose we have a random sample $X_1, \ldots, X_n$ from F_Q. We can view this as a missing data problem in that the 'full data' consists of pairs $(X_1, \Theta_1), \ldots, (X_n, \Theta_n)$, with $\Theta_i \sim Q$ and $X_i | \Theta_i = \theta \sim F_\theta$, but then only the first member X_i of each pair is observed; the labels Θ_i are hidden.
If the distribution *Q* is discrete with a finite number *k* of mass points $\theta_1, \ldots, \theta_k$ then we can write

$$F_Q = \sum_{j=1}^{k} q_j F_{\theta_j},$$

where $q_j = Q\{\theta_j\}$. The distribution F_Q is called a finite mixture distribution, the distributions F_θ are the component distributions and the q_j are the component weights.

There are several reasons why mixture distributions, and in particular finite mixture distributions, are of interest. First, there are many applications where the mechanism generating the data is truly of a mixture form; we sample from a population which we know or suspect is made up of several relatively homogeneous sub-populations in each of which the data of interest have the component distributions. We may wish to draw inferences, based on such a sample, relating to certain characteristics of the component sub-populations (parameters θ_j) or the relative proportions (parameters q_j) of the population in each sub-population, or both. Even the precise number of sub-populations may be unknown to us. An example is a population of fish, where the sub-populations are the yearly spawnings. Interest may focus on the relative abundances of each spawning, an unusually low proportion possibly corresponding to unfavourable conditions one year.

Second, even when there is no a priori reason to anticipate a mixture distribution, families of mixture distributions, in particular finite mixtures, provide us with particularly flexible families of probability distributions and densities which can be used to fit to unusually (skewed, long-tailed, multimodal) shaped data which would otherwise be difficult to describe with a more conventional parametric family of densities. Also, such a fit is often comparable in flexibility to a fully nonparametric estimate but structurally simpler, and often requires less subjective input, for example in terms of choosing smoothing parameters. For example, it has been shown that the very skewed log-normal density can often by well approximated by a two- or three-component mixture of normals, each with possibly different means and variances.

Third, many problems can be recast as mixture problems. An example is the problem of estimating a decreasing density function on the positive half-line. Such a density can be expressed as a mixture of uniform distributions, and, in the nonparametric maximum likelihood estimation of mixing distributions discussed below, we see that the solution to this density estimation problem follows from the solution to the general mixture problem.

Formal interest in finite mixtures dates back to at least Karl Pearson's laborious method-of-moments fitting of a two-component normal mixture to data on physical dimensions of crabs in the late 19th century. The mathematical difficulties inherent in fitting mixtures in that time have been greatly eased with the advent of the expectation-minimization (EM) algorithm in the 1970s. This algorithm yields an iterative method for computing maximum likelihood estimates (or very accurate approximations thereof) in a general missing-data situation. As mentioned above, mixtures have a natural missing-data interpretation and so the EM algorithm, together with improved computing technology, has made the task of fitting mixtures models to data much easier, leading to a renewal of interest in them.

Fitting finite mixtures using maximum likelihood

The EM-algorithm generates a sequence of parameter estimates each of which is guaranteed to give a larger likelihood than its predecessor. It can be used whenever the original log-likelihood $\log \cdot f_X(x;\ \theta)$ is difficult to maximize over θ for given x, but $f_X(x;\ \theta)$ can be expressed as the marginal distribution of X in a pair (X, J) whose corresponding log-likelihood $\log \cdot f_{XJ}(x, j;\ \theta)$ is easier to maximize over θ for given x and j. Given a 'current estimate' θ_0, the next in the sequence θ_1 is defined as the maximizer of the EM-log-likelihood $\cdot_{EM}(\theta; x)$ which is

defined as the conditional expectation of log•$f_{XJ}(x, J; \theta)$ over the 'missing data' J given $X = x$ computed under θ_0, that is

$$\ell_{EM}(\theta; x) = E\log f(x, J; \theta) \text{ where } J \text{ has density } f_{J|X}(j|x; \theta_0) = f_{XJ}(x, j; \theta_0) / f_X(x; \theta_0).$$

It is guaranteed that $\log f_X(x; \theta_1) \geq \log f_X(x; \theta_0)$.
If we wish to fit a finite mixture

$$f(x; Q) = \sum_{j=1}^{k} q_j f(x; \theta_j)$$

where the number of components k is known, the EM-algorithm works in almost the same way for either one or both of the q_j's or θ_j's unknown. We regard the x_i's as the observed first members of random pairs $(X_1, J_1), \ldots, (X_n, J_n)$, but the J_i's are unobserved. We can write the full data log-likelihood as

$$\sum_{i=1}^{n} \sum_{j=1}^{k} 1\{J_i = j\}\{\log q_j + \log f(x_i; \theta_j)\}$$

(here $q_j = P\{J_i = j\}$). We now outline how to go from an initial set of estimates $q_{01}, \ldots, q_{0k}, \theta_{01}, \ldots, \theta_{0k}$ to the next in the EM-sequence $q_{11}, \ldots, q_{1k}, \theta_{11}, \ldots, \theta_{1k}$. If some of these values are known, then they of course remain unchanged. The first step is to compute the posterior probabilities

$$\pi_{j|i} = P\left\{J_i = j | X_i = x_i\right\} \text{ computed under the } q_{0j}\text{s and } \theta_{0j}\text{s} = \frac{q_{0j} f(x_i; \theta_{0j})}{\sum_{j=1}^{k} q_{0j} f(x_i; \theta_{0j})}.$$

The EM-log-likelihood is then obtained by replacing the $1\{J_i = j\}$'s in the full data log-likelihood with the $\pi_{j|i}$'s; note that the EM-log-likelihood thus obtained separates into a term involving the q_j's only and one involving the θ_j's only.
If the q_j's are unknown, we maximize

$$\sum_{j=1}^{k} \log q_j \left\{ \sum_{i=1}^{n} \pi_{j|i} \right\}$$

with respect to the q_j's; this is maximized at

$$q_{1j} = n^{-1} \sum_{i=1}^{n} \pi_{j|i},$$

simply the averages of the posterior probabilities over the data:
If the θ_j's are unknown, we maximize

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \pi_{j|i} \log f(x_i; \theta_j)$$

with respect to the θ_j's. Differentiating with respect to each θ_j and setting to zero yields *k* weighted score equations:

$$\sum_{i=1}^{n} \pi_{j|i} \frac{\partial \log f(x_i; \theta_j)}{\partial \theta_j} = 0.$$

In many common models these are easily solved. For example, in one-parameter exponential families of the form $f(x; \theta) = e^{\theta x - K(\theta)} f_0(x)$, (for example, normal with known variance, Poisson, and so on) let $\hat{\theta}(t)$ be that value of θ that solves $K'(\theta) = t$. Then for each *j* one can explicitly find the EM update as

$$\theta_{j1} = \hat{\theta}\left(\frac{\sum_{i=1}^{n} \pi_{j|i} x_i}{\sum_{i=1}^{n} \pi_{j|i}} \right),$$

a known function of a $\pi_{j|i}$-weighted average of the x_is.

Further inferences

Once the model has been fitted, further inferences may consist of confidence intervals for, or hypothesis tests

concerning, the component parameters θ_j and/or the mixing proportions q_j. When the model is correctly specified (that is, there really are k components and all the q_j's are positive), the parameter estimates behave more or less in a standard fashion: they are asymptotically normal with an estimable covariance matrix, subject to the component densities $f(x;\ \theta_j)$ being suitably regular. Hence confidence regions can be computed in a standard fashion, bearing in mind the restrictions on the q_j's: they are non-negative and add to 1. In addition, one should be aware that, when the weights q_j are small or the parameters θ_j for two or more groups are similar, there is a sharp loss of estimating efficiency as well as good reason to be doubtful of the accuracy of asymptotic approximations. This occurs because of the near loss of identifiability of the parameters near the boundaries of the parameter space.

Hypothesis tests are perhaps not so standard, at least not for tests concerning the q_js. If one wishes to test whether an estimate $\hat{q}_j$ is significantly different from zero, the non-negativity constraints have a significant impact, at least when it comes to using large-sample χ^2 approximations to the p-values. Since such a hypothesis constrains a parameter to be on the boundary of the parameter space, the asymptotic distribution of twice the log-likelihood ratio will be a mixture of χ^2 distributions rather than a pure χ^2, on the assumption that the model is otherwise suitably regular. In such a case, a parametric bootstrap approach can be used to obtain an approximate p-value.

An unknown number of components, or completely unknown Q

If the number of components of a putatively finite mixture is unknown, we are essentially on the same footing as knowing absolutely nothing about Q, for reasons we now explain.

For any given data-set $x_1, \ldots, x_n$ with $d \leq n$ distinct x_i's and any pre-specified Q, no matter it be discrete or continuous, so long as the likelihoods $f(x_i;\ \theta)$ are bounded in θ we can find a discrete $\ddot{Q}$ with $m \leq d$ support points such that Q and $\ddot{Q}$ provide exactly the same density values at the observed data. That is, for any mixing distribution Q there is a possibly different $\ddot{Q}$ yielding a finite mixture such that Q and $\ddot{Q}$ cannot be distinguished, at least in terms of the data $x_1, \ldots, x_n$. So it suffices to restrict attention to such $\ddot{Q}$s.

An implication of this, when the likelihoods are bounded in θ, is that the maximum likelihood estimate of Q over all distributions, which we denote by $\hat{Q}$, exists and is finite with at most d (the number of distinct x_is) support points. So we never need leave the realm of finite mixtures in this setting.

This is not to say, however, that an estimate of an unknown k is readily available. The number of components in $\hat{Q}$ may be an overestimate in that some support points (respectively mixing proportions) may be so close together (small) that combining them into a single point (removing them) hardly decreases the likelihood. This and other issues related to trying to infer something about the number of components in a mixture, like hypothesis tests concerning k, are difficult problems. Some problems are still open, others have solutions that are possibly too complex to be useful.

The nonparametric estimate of Q

When the estimate $\hat{Q}$ discussed above exists, it is discrete with at most d support points. Hence a strategy for computing it is to try to fit a finite mixture with d components using the EM-algorithm. In many situations this yields a sensible result. More sophisticated algorithms exist however which are related to the following gradient function characterization.

The gradient function

$$D_Q(\theta) = \sum_{i=1}^{n} \left[\frac{f(x_i, \theta)}{f(x_i, Q)} - 1 \right]$$

measures the rate of increase in the log-likelihood if we remove a small amount of weight from the mixing distribution *Q* and put it at the point θ. Hence, for a candidate estimate *Q*, if for some θ we have $D_Q(\theta) > 0$, we know that we can increase the log-likelihood by putting some weight at θ.

In light of this the following result is not surprising: if the nonparametric maximum likelihood estimate $\hat{Q}$ exists, then $D_{\hat{Q}}(\theta) \leq 0$ for all θ, and the support points of $\hat{Q}$ are included in the set of values θ where $D_{\hat{Q}}(\theta) = 0$. The fact that $D_{\hat{Q}}(\theta) > 0$ for no θ makes sense; moving mass around from $\hat{Q}$ to any other θ cannot increase the likelihood.

The nonparametric version of the mixture model falls into the class of convex models, a subject with its own independent literature. Often convex models can be written as mixture models. For example, a distribution function that is concave on the positive half-line can also be written as a nonparametric mixture of the form $\int f(x, \theta)\, dQ(\theta)$ with component density $f(x, \theta) = 1\{0 < x < \theta\} / \theta$. One can deduce that the nonparametric likelihood estimator is the least concave majorant of the empirical distribution function using the above gradient characterization. See McLachlan and Peel (2000), Titterington, Smith and Makov (1985) or Lindsay (1995) for further examples and other references.

Mixtures and nonlinear time series

Methods related to mixtures of distributions have in recent times enjoyed a surge in popularity in finance and econometrics, in particular in the area of time series analysis. Traditional (linear) time series models, while intuitive and tractable, are well-known to be unable to capture certain features of much financial or econometric data, including variability that changes over time and marginal distributions that can be multimodal or long-tailed. Traditional linear time series models with Gaussian innovations have marginal and conditional distributions which are Gaussian. However, in many applications both marginal and conditional distributions can be multimodal, skewed, and fat-tailed, and exhibit other non-Gaussian features. Also, series can exhibit bursts of volatility, where the variability changes in strange ways, sometimes with some dependence on past and current values of the observable series or an unobserved underlying process of ‘shocks’. In several different settings, ideas of mixtures have led to new types of models that have been quite successful at capturing many of these problematic features. One example is the mixture of autoregressive (AR) models idea. The standard autoregressive model, where the observation at time *t*, Y_t, has a conditional distribution, given the past $Y_{t-1}, Y_{t-2}, \cdots$ of the form

$$Y_t = \theta_0 + \sum_{\ell=1}^{L} \theta_\ell Y_{t-\ell} + sZ_t,$$

where the $\theta_\bullet$'s are fixed constants and the Z_t's are independent (often standard Gaussian) random variables. Assuming $\theta_k \neq 0$ here, the model is said to be autoregressive of order *L* (we abbreviate this to AR(*L*)). The mixture version can be represented by replacing the parameter vector $\theta = (\theta_0, \theta_1, \ldots, \theta_L, s)^T$ above at each time point *t* with

a random version $\Theta_t = (\Theta_{0t}, \Theta_{1t}, \ldots, \Theta_{Lt}, S_t)^T$, yielding

$$Y_t = \Theta_{0t} + \sum_{\ell=1}^{L} \Theta_{\ell t} Y_{t-\ell} + S_t Z_t$$

where $P\left(\Theta_t = \theta^{(j)}\right) = q_j$, each $q_j \geq 0$ and $\Sigma_{j=1}^{k} q_j = 1$. For each *j*, we have a different AR *regime* with corresponding parameter vector $\theta^{(j)} = (\theta_0^{(j)}, \ldots, \theta_L^{(j)}, s_j)^T$ which is chosen randomly at each time point according to the probability distribution given by the q_j's, independently of Z_t and past values of the series. All regimes need not be of the same order; an $AR(L')$ regime with $L' < L$ can be obtained by just setting $\theta_{L'+1}^{(j)} = \cdots = \theta_L^{(j)} = 0$.
This so-called mixture autoregressive (MAR) model has several appealing features. Its mathematical form means that it is relatively straightforward to derive its autocorrelation function, and indeed its stationarity properties are similarly easy to derive. An interesting point here is that it is possible to have some of the component regimes non-stationary, but, so long as their mixing proportions q_j are small enough, the overall series can still have a second-order stationarity property (see time series analysis for more details). In looser terms, we can have occasional explosive behaviour but still have a series that is well-behaved in the long run. For example, when the stock market becomes volatile we can have short bursts of heightened activity which eventually settle down. Such features cannot be captured by a single AR model.
Another feature of the MAR model is that the marginal as well as conditional distributions can change with time and be multimodal. Again, during a period of stock market volatility we might expect some sharp increases and/or decreases during these periods which may result in bi- or multimodal conditional distributions. Consider the following example (simplified version of fit to IBM data from Wong and Li, 2000):

$$Y_t = \begin{cases} 0.7Y_{t-1} + 0.3Y_{t-2} + 5Z_t + & \text{with prob. } 0.55; \\ 1.7Y_{t-1} - 0.7Y_{t-2} + 5Z_t + & \text{with prob. } 0.4; \\ Y_{t-1} + 20Z_t & \text{with prob. } 0.5. \end{cases}$$

If the series has been quite volatile and Y_{t-1} and Y_{t-2} are very different, say $Y_{t-1} = 200$ and $Y_{t-2} = 300$, then the conditional distribution of Y_t would be a mixture of the form

$$Y_t \sim \begin{cases} N(230, 25) & \text{with prob. } 0.55; \\ N(130, 25) & \text{with prob. } 0.4; \\ N(200, 400) & \text{with prob. } 0.05. \end{cases}$$

However, if the series had been quite stable, say with $Y_{t-1} = 200$ and $Y_{t-2} = 201$ say, then the conditional

distribution would be

$$Y_t \sim \begin{cases} N(200.3, 25) & \text{with prob. } 0.55; \\ N(199.3, 25) & \text{with prob. } 0.4; \\ N(200, 400) & \text{with prob. } 0.05. \end{cases}$$

So we still have a component for increases, a component for decreases and the same component for outliers. However, the first two components are so similar that the mixture density is markedly unimodal. This example illustrates that the MAR can capture volatility as well as a changing, possibly multimodal conditional distribution.

Estimation

The mixture structure also enables maximum-likelihood estimation of unknown parameters via the EM algorithm. We briefly outline how this would work when fitting a mixture of k AR(L) regimes, although the basic steps are the same in cases where the order of each regime can differ from component to component. As in the i.i.d. case though, the question of choosing k, the number of components of the mixture, is a difficult open problem.
We can represent the mixture in terms of an unobserved label J_t at each time point which indicates which regime applies; it is equal to j with probability q_j, $j = 1, \ldots, k$. If these were known, then the full log-likelihood of observed $(y_{L+1}, \ldots, y_n)^T$ (conditional on $y_1, \ldots, y_L$) would be

$$\ell_{\text{full}}\left(\boldsymbol{q}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}\right) = \sum_{t=L+1}^{n} \left\{ \sum_{j=1}^{k} 1\{J_t = j\} \left[\log q_j + \log f\left(y_t, \ldots, y_{t-L}; \boldsymbol{\theta}^{(j)}\right) \right] \right\},$$

where $f(\cdot;\ \theta)$ is the conditional density of Y_t given $Y_{t-1}, \ldots, Y_{t-L}$ under a single AR(L) regime. We now show how a current set of estimates $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{\theta}}^{(1)}, \ldots, \tilde{\boldsymbol{\theta}}^{(k)}$ would be updated. There are two steps, an E-step and an M-step. At the E-step the missing data is set equal to its conditional expectation, given current parameter estimates and data, which here reduce to the posterior probabilities:

$$\pi_{j|t} = P\left\{ J_t = j \,|\, Y_t = y_t, \ldots, Y_{t-L} = y_{t-L} \right\} \text{ computed under current estimates} = \frac{\tilde{q}_j f\left(y_t, \ldots, y_{t-L}; \tilde{\boldsymbol{\theta}}^{(j)}\right)}{\sum_{t=L+1}^{n} \tilde{q}_j f\left(y_t, \ldots, y_{t-L}; \tilde{\boldsymbol{\theta}}^{(j)}\right)}$$

The M-step consists of firstly defining the EM-log-likelihood $\ell_{\text{EM}}\left(\boldsymbol{q}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}\right)$ obtained by replacing $1\{J_j = t\}$ with $\pi_{j|t}$, and then maximizing over the remaining parameters. As in the i.i.d. case, the EM-log-likelihood separates

into two pieces, one involving just the q_j's, which is maximized at

$$\hat{q}_j = \frac{\sum_{t=L+1}^{n} \pi_{jt}}{n-L}$$

and another involving the other parameters of the form

$$\sum_{j=1}^{k} \sum_{t=L+1}^{n} \pi_{jt} \log f\left(y_t, \ldots, y_{t-L}; \boldsymbol{\theta}^{(j)}\right)$$

which when differentiated partially with respect to each $\theta^{(j)}$ yields a separate set of *weighted likelihood* equations just as in the i.i.d. case, for example,

$$\frac{\partial}{\partial \theta_0^{(j)}} \ell_{\mathrm{EM}}\left(\boldsymbol{q}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}\right) = \sum_{t=L+1}^{n} \pi_{jt} \frac{\partial}{\partial \theta_0} \log f\left(y_t, \ldots, y_{t-L}; \boldsymbol{\theta}\right) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{(j)}}.$$

Thus, if one has a computational method to obtain the maximum likelihood estimate for a straight AR(L) model, it is possible to use the same computations on this weighted form in the M-step for the more general mixture case. Note that this method is not restricted to the Gaussian-Z_t case or a linear autoregression function.

As mentioned earlier, the autocorrelation structure of the MAR model is quite straightforward to analyse; in fact, it inherits much of the simplicity of the standard AR model. One thing that one cannot obtain using an AR or MAR model is a first-order stationary series whose square exhibits some autocorrelation, which is a key feature of certain time series models designed to capture time-varying volatility. The main breakthrough in this area was the introduction of the autoregressive conditional heteroscedastic (ARCH) model for time series errors in the early 1980s by Engle, where S_t^2, the variance of the error at time t, is allowed to depend on squares of earlier errors: if Z_t's are i.i.d. mean-zero-unit-variance errors then the series $\{\varepsilon_t\}$ given by

$$\varepsilon_t = S_t Z_t; S_t = \left(\beta_0 + \sum_{\ell=1}^{M} \beta_\ell \varepsilon_{t-\ell}^2\right)$$

is an ARCH(M)-series. One can incorporate this into a mixture setting by using the same specification for the conditional mean as in the MAR case, but allowing the errors to be generated *within each regime* by a different ARCH mechanism. Hence the full specification is

$$Y_t = \Theta_{0t} + \sum_{\ell=1}^{L} \Theta_{\ell t} Y_{t-\ell} + \varepsilon_t, \varepsilon_t = S_t Z_t; S_t = \left(B_0 + \sum_{\ell=1}^{M} B_\ell \varepsilon_{t-\ell}^2 \right),$$

where now (Θ_t, B_t) takes the value ($\theta^{(j)}$, $\beta^{(j)}$) with probability q_j.
The resulting MAR–ARCH model combines the extra flexibility of the MAR model with the superior modelling of volatility enjoyed by ARCH series. In addition, the ability to fit several different AR–ARCH regimes provides an aid to interpretation; as in the MAR case, we can have a different regime for each of several possible reactions at each time point, and furthermore the choices (that is, conditional distributions) can change with time. The EM-algorithm can be employed in essentially the same way as the MAR model, so long as weighted maximum likelihood estimation can be performed in the M-step for each AR-ARCH regime (allowing the possibility of non-normal errors).

Connection to threshold models

There is some connection between MAR and MAR-ARCH models and another class of non-linear time series known as (self-exciting) threshold autoregressive (SETAR) models. An elementary version is

$$Y_t = \begin{cases} \theta_0^{(1)} + \theta_1^{(1)} Y_{t-1} + s_1 Z_t \text{if } Y_{t-1} < c, \\ \theta_0^{(2)} + \theta_1^{(2)} Y_{t-1} + s_2 Z_t \text{if } Y_{t-1} \geq c, \end{cases}$$

That is, follows one of two possible AR(1) regimes, the choice depending on whether the previous value Y_{t-1} exceeds a threshold c, in contrast to the MAR model where the choice is made independently of the earlier values of the series.

It can be shown that if the Z_t's are Gaussian then the marginal distribution of the zeroth order (where $\theta_1^{(j)} \equiv 0$) is a mixture of Gaussians, permitting multimodality.
A class of models intermediate between the SETAR models and MAR involves having several AR regimes, but the choice at each time point is partly influenced by earlier values of the series, but not in a completely deterministic way. A simple version involves replacing the thresholding rule $Y_{t-1} < c$ with $Y_{t-1} + \eta_t < c$ for an independent random variable η_t. In this case, we have a mixture of AR regimes where the mixing proportions $q_j = q_j(Y_{t-1}, c)$ depend on earlier values of the series and the threshold.
These models (MAR, MAR-ARCH, SETAR and intermediate versions) are still being fully developed, however an excellent introduction is provided in Tong (1990).

Summary

Mixture distributions, particularly finite mixtures, in general permit a great increase in flexibility of modelling without an overwhelming increase in computation difficulty, while also helping in interpretation by modelling heterogeneity in a natural way. In particular, if distributions within a certain model can be fitted by maximum

likelihood, then finite mixtures of distributions from the same model can in general also by fitted by maximum likelihood using the EM-algorithm. Such finite mixtures can capture heterogeneity or other complex behaviour that single components (that is, when there is no mixture) cannot capture.

See Also

- statistical inference
- testing
- time series analysis

Bibliography

Lindsay, B.G. 1995. *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, 5. Hayward, CA: Institute of Mathematical Statistics and American Statistical Association.

McLachlan, G. and Peel, D. 2000. *Finite Mixture Models*. New York: Wiley-Interscience.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. 1985. *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley.

Tong, H. 1990. *Nonlinear Time Series: A Dynamical System Approach*. Oxford: Clarendon Press.

Wong, C.S. and Li, W.K. 2000. On a mixture autoregressive model. *Journal of the Royal Statistical Society*, Series B 62, 95–115.

How to cite this article

Lindsay, Bruce G. and Michael Stewart. "mixture models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000382> doi:10.1057/9780230226203.1108

The New Palgrave Dictionary of Economics Online

model averaging

Gernot Doppelhofer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Model averaging estimates the distribution of quantities of interest across models. Model averaging can be used for inference, prediction and policy analysis to address model uncertainty. Three main approaches are discussed: Bayesian model averaging (BMA), empirical Bayes (EB) methods, and frequentist model averaging (FMA). Differences in prior specifications are contrasted using the example of normal, linear regression models. Finally, the article discusses implementation issues such as numerical simulation techniques and software for model averaging.

Keywords

Bayes' rule; Bayesian estimation; Bayesian model averaging; empirical Bayes methods; exchangeability; frequentist model averaging; homoskedasticity; likelihood; Markov chain Monte Carlo methods; Metropolis–Hastings algorithm; model averaging; model selection criteria; model uncertainty; posterior model probabilities; sensitivity analysis; statistical decision theory; stochastic search variable selection

Article

Model averaging allows the estimation of the distribution of unknown parameters and related quantities of interest across different models. The basic principle of model averaging is to treat models and associated parameters as unobservable and estimate their distributions based on observable data. Model averaging can be employed for inference, prediction and policy analysis in the face of model uncertainty. Many areas of economics give rise to model uncertainty, including uncertainty about theory, specification and data issues. A naive approach that ignores model uncertainty generally results in biased parameter estimates, overconfident (too narrow) standard errors and misleading inference and predictions (see Draper, 1995). Taking model uncertainty seriously implies a departure from conditioning on a particular model and calculating quantities of interest by averaging across different models instead.

Model averaging is conceptually straightforward. The sample information contained in the likelihood function for a particular model is combined with relative model weights or posterior model probabilities to estimate the distribution of unknown parameters across models. Three main approaches – Bayesian, empirical Bayes, and frequentist – have been developed, and they differ in their underlying statistical foundations and practical implementation.

Bayesian model averaging (*BMA*) was developed first to systematically deal with model uncertainty. The idea of combining evidence from different models is readily integrated into a Bayesian framework. Jeffreys (1961) laid the foundation for BMA, further developed by Leamer (1978). Hoeting et al. (1999), Wasserman (2000) and Koop (2003) give excellent introductions to BMA. A drawback of the Bayesian approach is that it requires assumptions about prior information about distribution of unknown parameters. In response, *empirical Bayes* (*EB*) approaches have been developed to estimate elements of the prior using observable data. Chipman, George and

McCulloch (2001) argue for a pragmatic approach that introduces objective or frequentist considerations into model averaging. In contrast to Bayesian approaches, *frequentist model averaging* (*FMA*) methods were developed only relatively recently. Recent contributions include Yang (2001), Hjort and Claeskens (2003) and Hansen (2007).

Model averaging was not widely used until advances in statistical techniques and computing power facilitated its practical use (see Chib, 2001; Geweke and Whiteman, 2006). Economic applications of model averaging include economic growth (Fernandez, Ley and Steel, 2001a; Sala-i-Martin, Doppelhofer and Miller, 2004), finance (Avramov, 2002), policy evaluation (Brock, Durlauf and West, 2003; Levin and Williams, 2003), macroeconomic forecasting (Garratt et al., 2003).

This article is organized as follows. The statistical model averaging framework is introduced in the next section. Different model averaging approaches are illustrated with applications to linear regressions. Finally, implementation issues, including model priors, numerical methods, and software are discussed.

Statistical framework

Suppose a decision maker observes data *Y* and wishes to learn about quantities of interest related to an unknown parameter (vector) θ, such as the effect of an economic variable (say $\theta > 0$ or $\theta \leq 0$) or predictions of future observations Y^f. The utility (or loss) function of the decision maker describes the relation between parameter of interest θ and action *a*. For example, the decision maker could maximize expected utility

$$\max_a E[u(a, \theta|Y)] = \int u(a, \theta|Y)\, p(\theta|Y)\, d\theta. \tag{1}$$

In general, the preferred action depends on the preferences of the decision-maker and the unconditional distribution of parameters. Alternative preference structure can have important consequences for optimal estimators and implied policy conclusions. Bernardo and Smith (1994) give an accessible introduction to statistical decision theory. In the context of economic policy, Brock, Durlauf and West (2003) present an interesting discussion of alternative preferences and implied policies.

A key ingredient in decision making is the *posterior* distribution of the parameter θ, which can be calculated using Bayes's rule:

$$p(\theta|Y) = \frac{L(Y|\theta)\, p(\theta)}{p(Y)} \propto L(Y|\theta)\, p(\theta). \tag{2}$$

The posterior distribution is therefore proportional to the *likelihood* function $L(Y|\theta)$, which summarizes all information about θ contained in the observed data, and the *prior* distribution $p(\theta)$. In contrast, the classical approach assumes that the parameter θ is fixed (non-random) and does not have a meaningful distribution. The estimator $\hat{\theta}$ on the other hand is viewed as a random variable.

In many economic and more generally non-experimental applications, a decision maker might face considerable model uncertainty given potentially overlapping, economic theories. Brock and Durlauf (2001) refer to this as ‘open-endedness’ of economic theories. Also, there might be alternative empirical specifications of these theoretical channels. In sum, the number of observations may be smaller than the number of suggested explanations, and the problem may be compounded by data problems, such as missing data or outliers.

Formally, there may be many candidate models $M_1, \ldots, M_K$ to explain the observed data. A model M_j can be described by a probability distribution $p(Y|\theta_j, M_j)$ with

model-specific parameter (vector) θ_j. In a situation of model uncertainty, the decision-maker evaluates the utility function (1) using the posterior distribution of θ. The posterior distribution is unconditional with respect to the set of models and is calculated by averaging conditional or model-specific distributions across all models

$$p(\theta|Y) = \sum_{j=1}^{K} w_j \cdot p(\theta_j|M_j, Y), \tag{3}$$

where the model weights w_j are proportional to the fit in explaining the observable data. In a Bayesian context, the weights are the *posterior model probabilities*, $w_j = p(M_j|Y)$. Using Bayes's rule,

$$p(M_j|Y) = \frac{L(Y|M_j)\,p(M_j)}{\sum_{j=1}^{K} L(Y|M_j)\,p(M_j)} \propto L(Y|M_j)\,p(M_j). \tag{4}$$

The posterior model weights are proportional to the product of prior model probability $p(M_j)$ and model-specific marginal likelihood $L(Y|M_j)$. The marginal likelihood is obtained by integrating a model-specific version of equation (2) with respect to θ_j

$$L(Y|M_j) = \int_{\theta} L(Y|\theta_j, M_j)\,p(\theta_j|M_j)\,d\theta_j \tag{5}$$

using the fact that $\int p(\theta_j|M_j, Y)\,d\theta_j = 1$.
When comparing two models, M_i and M_j say, the posterior model probabilities or *posterior odds ratio* equals the ratio of integrated likelihoods times the prior odds

$$\frac{p(M_i|Y)}{p(M_j|Y)} = \frac{L(Y|M_i)}{L(Y|M_j)}\,\frac{p(M_i)}{p(M_j)}. \tag{6}$$

Similarly, the weight for model M_i relative to K models under consideration is given by (4), where the normalizing factor $\sum_{j=1}^{K} L(Y|M_j)\,p(M_j)$ ensures consistency of

model weights.

The decision maker may be interested in particular aspects of the unconditional distribution (3), such as posterior mean or variance. Leamer (1978) derives the following expressions for unconditional mean or variance of the parameter θ

$$E(\theta|Y) = \sum_{j=1}^{K} p(M_j|Y)E(\theta_j|Y, M_j). \tag{7}$$

$$Var(\theta|Y) = E(\theta^2|Y) - [E(\theta|Y)]^2 = \sum_{j=1}^{K} p(M_j|Y)\left\{Var(\theta_j|Y, M_j) + [E(\theta_j|Y, M_j)]^2\right\} - [E(\theta|Y)]^2 = \sum_{j=1}^{K} p(M_j|Y)Var(\theta_j|Y, M_j) + \sum_{j=1}^{K} p(M_j|Y)[E(\theta_j|Y, M_j) - E(\theta|Y)]^2. \tag{8}$$

The expression for the unconditional mean of θ in (7) is simply the model-weighted sum of conditional means. Notice that the unconditional variance of θ in (8) exceeds the sum of model-weighted conditional variances by an additional term, reflecting the distance between the estimated conditional mean in each model $E(\theta_j|Y, M_j)$ and the unconditional mean $E(\theta|Y)$. Ignoring this last term overestimates the precision of estimated effects and underestimates parameter uncertainty (see Draper, 1995).

The advantage of the Bayesian approach to model averaging is its generality and the explicit treatment of model uncertainty and decision theory. The decision maker simply combines prior information about the distribution of parameters and models with sample information to calculate the unconditional posterior distribution of θ in eq. (3).

However, there are several problems that can make implementation of BMA difficult in practice (see Hoeting et al., 1999; Chipman, George and McCulloch, 2001):

1. 1. The specification of prior distribution of parameters θ requires assumptions about functional forms and unknown hyper-parameters which will in general affect the marginal likelihood (5) and hence posterior model weights (4).
2. 2. The specification of prior probabilities over the model space $p(M_j)$ might have important effects on posterior model weights (4).
3. 3. The number of models K in eq. (3) can be too large for a complete summation across models, implying the use simulation techniques to approximate the unconditional distribution $p(\theta|Y)$ in equation (3).
4. 4. Choices of utility function (1) and class of models are other important issues.

These issues are discussed in turn, contrasting the fully Bayesian, empirical Bayes and frequentist approaches.

Linear regression example

Many of the implementation problems of model averaging and approaches suggested in the literature can be illustrated using the linear regression example (see Koop,

2003). Raftery, Madigan and Hoeting (1997) and Fernandez, Ley and Steel (2001b) discuss BMA for linear regression models.

Consider linear regression models of the form

$$y = x_1\beta_1 + \cdots + x_k\beta_k + \varepsilon = X\beta + \varepsilon, \tag{9}$$

where *y* is the vector of *N* observations of the dependent variable and $X = [x_1, \ldots, x_k]$ is a set of *k* regressors (including a constant) with associated coefficient vector β. Each model M_j is characterized by a subset of explanatory variables X_j with coefficient vector β_j. With *k* regressors, the set of linear models equals $K = 2^k$. The residuals are drawn from a multivariate normal distribution and are assumed to be conditionally homoskedastic, $\varepsilon_j \sim N(0, \sigma^2 I)$. Notice that this implies that the residuals are also conditionally exchangeable (see Bernardo and Smith, 1994; Brock and Durlauf, 2001).

Suppose the decision maker is interested in the effect of different explanatory variables, represented by slope parameters β with posterior distribution of $p(\beta|Y)$. As shown in eq. (3), the posterior distribution is estimated by weighting conditional distributions of parameters by posterior model probabilities. The relative posterior model weights in eqs (6) and (4) are proportional to the marginal likelihood and prior model weights.

For the normal regression model, the likelihood function can be written as

$$L(y|\beta_j, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}}\left\{\exp\left[-\frac{1}{2\sigma^2}(y - X\beta_j)'(y - X\beta_j)\right]\right\} \propto \left\{\exp\left[-\frac{1}{2\sigma^2}(\beta_j - \hat{\beta}_j)'X_j'X_j(\beta_j - \hat{\beta}_j)\right]\right\}\left\{\sigma^{-(v_j+1)}\exp\left[-\frac{v_j s_j^2}{2\sigma^2}\right]\right\}. \tag{10}$$

The second line of the likelihood substitutes the ordinary least squares (OLS) estimates for the slope and variance

$$\hat{\beta}_j = (X_j'X_j)^{-1}X_j'y, \tag{11}$$

$$s_j^2 = \frac{(y - X_j\hat{\beta}_j)'(y_j - X_j\hat{\beta}_j)}{v_j}, \tag{12}$$

with degrees of freedom $v_j = N - k_j - 1$. The implementation of model averaging – Bayesian, empirical Bayes, or frequentist – requires the specification of prior distributions $p(\theta_j)$ for the model parameters $\theta_j = (\beta_j, \sigma^2)$.

Bayesian conjugate priors

A standard way to specify priors in Bayesian estimation is to assume a prior structure that is analytically and computationally convenient. A *conjugate prior* distribution leads to a posterior distribution of the same class of distributions when combined with the likelihood. The likelihood (10) is part of the Normal-Gamma family of distributions, proportional to the product of a normal distribution for the slope β_j, conditional on the variance σ^2, and an inverse-Gamma distribution for the variance σ^2. The conjugate prior therefore takes the form

$$p(\beta_j|\sigma^2, M_j) \sim N\left(\beta_{0j}, \sigma^2 V_{0j}\right) p(\sigma^2|M_j) = p(\sigma^2) \sim IG(s_0^2, v_0) \tag{13}$$

where the prior hyper-parameters for slope and variance are denoted by subscript 0. Notice that the error variance is assumed to be drawn from the same distribution across all regression models, reflecting the assumption of conditional homoskedasticity and exchangeability of the residuals.
A drawback of the Bayesian approach is that marginal likelihood and posterior model weights depend on unknown hyper-parameters (β_0, V_0, s_0, v_0). Different subjective priors therefore affect the posterior model weights and distribution of parameters, and hence also the decision maker's action. The standard Bayesian approach to check for robustness with respect to the choice of prior parameters is sensitivity analysis. An alternative strategy is to limit the use of subjective prior information and use objective methods based on observed data.

Empirical Bayes priors

Empirical Bayes (EB) approaches make use of sample information to specify prior parameters. Different versions of empirical Bayes methods have been proposed in the literature (see Hoeting et al., 1999; George and Foster, 2000; Chipman, George and McCulloch, 2001). To limit the importance of prior information, EB methods often use non-informative or diffuse priors that are dominated by the sample information (see Leamer, 1978). Jeffreys (1961) proposes non-informative priors to represent lack of prior knowledge and derives a formal relationship to the expected information in the sample.
A drawback of non-informative priors is that they are usually not proper distributions, which can lead to undesirable properties when comparing models with different parameters. In this case, relative model weights can depend on arbitrary constants. However, this problem is not present when comparing models with common parameters, since normalizing constants drop out from *relative* model weights (see Kass and Raftery, 1995). Koop (2003) argues that informative or proper priors should be used for all other (non-common) parameters.
Fernandez, Ley and Steel (2001b) propose *benchmark priors* for BMA that limit the subjective prior information to a minimum while maintaining the Bayesian natural conjugate framework. They suggest the following non-informative priors for the error variance, assumed to be the same in all *k* models:

$$p(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{14}$$

The slope parameter β_j is drawn from a normal prior distribution as in eq. (13) with prior mean $\beta_{0j} = 0$ and prior covariance matrix V_{oj} equal to the so-called *g*-prior suggested by Zellner (1986):

$$V_{oj} = \left(g_0 X_j' X_j\right)^{-1}. \tag{15}$$

Intuitively, the prior covariance matrix is assumed to be proportional to the sample covariance with a factor of proportionality g_0. The *g*-prior simplifies the specification of prior covariances to choosing a single parameter g_0. For example, $g_0 = 0$ corresponds to completely non-informative priors, and $g_0 = 1$ implies a very informative prior receiving equal weight to the sample information. Based on extensive simulations, Fernandez, Ley and Steel (2001b) recommend the following benchmark values:

$$g_0 = \begin{cases} 1/k^2, & \text{if } N \le k^2 \\ 1/N, & \text{if } N > k^2 \end{cases}. \tag{16}$$

Note that the ratio of prior to sample variance g_0 decreases with the sample size or with the square of estimated parameters. If the number of parameters is relatively large $k^2 \ge N$, the variance is assumed to be relatively more diffuse.
Using this prior structure, the posterior weights for model M_j can be written as

$$p(M_j|Y) \propto p(M_j) \cdot \left(\frac{1+g_0}{g_0}\right)^{-k_j/2} \cdot SSE_j^{-(N-1)/2}. \tag{17}$$

The weight for model $p(M_j|Y)$ depends on three terms: (i) the prior model weight $p(M_j)$, (ii) a penalty term for the number of regressors $((1+g_0)/g_0)^{-k_j/2}$ implying a preference for parsimonious models, and (iii) a term involving the sum of squared errors of the regression $SSE_j \equiv (y - X_j\beta_j)'(y - X_j\beta_j)$, corresponding to the kernel of the normal likelihood.

Frequentist sample dominated priors

A potential problem of using non-informative *g*-priors for the error covariance matrix is that the limit of posterior weights may be very sensitive to specification of the prior (see Leamer, 1978). Alternatively, Leamer (1978) assumes that a proper, conjugate Normal-Gamma prior (13) is 'dominated' by the sample information as the number of observations *N* grows. For stationary regressors with $\lim_{N \to \infty} (X_j' X_j) / N$ converging to a constant, the implied model weight is approximately equal to the (exponentiated) Schwarz (1978) model selection criterion (BIC)

$$p(M_j|Y) \propto p(M_j) \cdot N^{-k_j/2} \cdot SSE_j^{-N/2}. \tag{18}$$

On closer inspection, the relative model weights using non-informative *g*-priors (17) or sample-dominated prior (18) are essentially the same, using $g_0 = 1/N$ in eq. (16). This is very reassuring for a decision maker, since the relative model weights are very similar under an empirical Bayesian or frequentist interpretation.

The BIC weights can also be derived from a unit information prior, where the information introduced by the prior corresponds to *one* datapoint from the sample (see Kass and Wasserman, 1995; Raftery, 1995). Klein and Brown (1984) give an alternative derivation of the BIC model weights (18) by minimizing the so-called Shannon information in the prior distribution; this approach also lends support for using the BIC model weights in small samples.

The underlying model space and its interpretation are important issues in the model uncertainty literature. Bernardo and Smith (1994) distinguish between *M*-closed and *M*-open environments, where the former includes the true model and the latter does not necessarily. A set of Akaike (AIC) model weights can be derived in the *M*-open environment as the best approximation to the true distribution (see Burnham and Anderson, 2002). The AIC weights have the disadvantage that they will not be consistent in *M*-closed environments.

Prior over model space

An important ingredient to model averaging is the choice of prior model probability. A popular choice is to impose a uniform prior over the space of models

$$p(M_j) = 1/K. \tag{19}$$

This prior might represent diffuse information about the set of models, but does have important implications for the size of models.

There are different approaches to modelling the inclusion of explanatory variables in the linear regression models (9). Mitchell and Beauchamp (1988) assign a discrete prior probability mass $p(\beta_i = 0|M_j)$ to excluding regressors x_i from the regression model M_j, that is a 'spike' at zero. A more Bayesian approach assigns a mixture of a relatively informative prior at zero (corresponding to a spike at zero) and a more diffuse prior if the variable is included (see George and McCulloch, 1993).

An alternative to specifying prior model probabilities is to think about prior model size and the implied probability of including individual variables. Sala-i-Martin, Doppelhofer and Miller (2004) argue that in the context of economic growth regressions a prior model size $\bar{k}$ smaller than the one implied by uniform priors *k*/2 might be preferable. Notice that this translates into a prior probability $\pi = p(\beta_i \neq 0|M_j) = \bar{k}/k$ of including a regressor x_i in model M_j. The implied model probability can then be written as

$$p(M_j) = \pi^{k_j} \cdot (1-\pi)^{k-k_j}. \tag{20}$$

Notice that the prior inclusion probabilities π_i and implied prior model weights can also differ across variables, which is used in the 'stratified' sampler of the BACE approach by Sala-i-Martin, Doppelhofer and Miller (2004) to speed up numerical convergence.

George (1999) observes that, when allowing for a large number of explanatory variables which could be correlated with each other, posterior model probabilities can be spread across models with 'similar' regressors. To address this problem, George (1999) proposes *dilution priors*, which reduce the prior weight on models that include explanatory variables measuring similar underlying theories. Alternatively, one can impose a hierarchical structure on the set of models and variables and partition the model space accordingly (see Chipman, George and McCulloch, 2001; Brock, Durlauf and West, 2003). Doppelhofer and Weeks (2007) propose to estimate the degree of dependence or jointness among regressors over the model space. If we are only interested in prediction, the orthogonalization of regressors greatly reduces the computational burden of model averaging (see Clyde, Desimone and Parmigiani, 1996). The costs are the loss of interpretation of associated coefficient estimates and the need to recalculate orthogonal factors with changing sample information.

Numerical simulation techniques

A major challenge for the practical implementation of model averaging is the computational burden of calculating posterior quantities of interest when the model space is potentially very large. In the linear regressions example, an exhaustive integration over all 2^k models becomes impractical for a relatively moderate number of 30 regressors.

Recent advances in computing power and development of statistical methods have made numerical approximations of posterior distributions feasible. Chib (2001) gives an overview of computationally intensive methods. Such methods include Markov chain Monte Carlo techniques (Madigan and York, 1995), stochastic search variable selection (George and McCulloch, 1993), the Metropolis–Hastings algorithm (Chib and Greenberg, 1995), and the Gibbs sampler (Casella and George, 1992). Chipman, George and McCulloch (2001) contrast different approaches in the context of Bayesian model selection.

The main idea of Monte Carlo simulation techniques is to estimate the empirical distribution of the parameter θ or related functions of interest $g(\theta)$ by sampling from the posterior distribution

$$E[g(\theta)|Y] = \int g(\theta)\, p(\theta|Y)\, d\theta, \tag{21}$$

where $g(\theta)$ could be any function, such as variance of θ or predicted values of the dependent variable y. Consider the sample counterpart

$$\widehat{g}_S = \frac{1}{S}\sum_{s=1}^{S} g\left(\theta^{(s)}\right), \tag{22}$$

where $\theta^{(s)}$ is a random *i.i.d.* sample drawn from $p(\theta|Y)$ and *S* is the number of draws. Provided that $E[g(\theta)|Y] < 0$ exists, a weak law of large numbers implies

$$\widehat{g}_S \xrightarrow{p} E[g(\theta)|Y]. \tag{23}$$

A central limit theorem implies that

$$\sqrt{S}\{\widehat{g}_S - E[g(\theta)|Y]\} \xrightarrow{d} N(0, \Sigma_g) \tag{24}$$

where Σ_g is the estimated covariance matrix of $g(\theta)|Y$.

Markov chain Monte Carlo (MCMC) techniques strengthen these results by constructing a Markov chain moving through the model space $\{M(s),\ s = 1, \ldots, S\}$ that simulates from a transition kernel $p(\theta^{(s)}|\theta^{(s-1)})$, starting from an initial value $\theta^{(0)}$. There are various approaches to constructing a Markov chain that converges to the posterior distribution $p(\theta|Y)$. This limiting distribution can be estimated from simulated values of $\theta^{(s)}$.

Simulation methods differ with respect to the choice of sampling procedure and transition kernels. A sampling algorithm that uses the underlying structure of the model can greatly improve the efficiency of the simulation. For example, the Gibbs sampler uses the structure of the statistical model to partition parameters and their distribution into blocks, which breaks up the simulation into smaller steps. In the linear regression example, the Gibbs sampler can draw from the conditional distributions for slope and variance parameters (13) separately. A disadvantage of numerical methods can be the technical challenges in their implementation (for an excellent introduction, see Gilks, Richardson and Spiegelhalter, 1996). Links to software packages and codes that facilitate implementation, such as BACC, BACE, BUGS and the BMA project website, are listed at the end of this article.

An alternative approach is to limit the set of models and rule out dominated models by Occam's razor, see Hoeting et al. (1999). This can speed up computation of posterior distributions and can be useful tool for model selection. Evidence by Raftery, Madigan and Volinsky (1996) suggests that model averaging leads to important improvements in predictive performance over any single model, and gives a small predictive advantage relative to the restricted set of models. The relative performance of different model averaging techniques and associated model weights depends on sample size and stability of estimated model (see Yuan and Yang, 2005; Hansen, 2007).

See Also

- Bayesian econometrics
- Bayesian methods in macroeconometrics
- Bayesian statistics
- decision theory in econometrics

- econometrics
- extreme bounds analysis
- hierarchical Bayes models
- model uncertainty
- shrinkage-biased estimation in econometrics
- testing

Bibliography

Avramov, D. 2002. Stock return predictability and model uncertainty. *Journal of Financial Economics* 64, 423–58.

Bernardo, J.M. and Smith, A.F.M. 1994. *Bayesian Theory*. New York: Wiley.

Brock, W.A. and Durlauf, S.N. 2001. Growth empirics and reality. *World Bank Economic Review* 15, 229–72.

Brock, W.A., Durlauf, S.N. and West, K. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 2003(1), 235–322.

Burnham, K.P. and Anderson, D.R. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. New York: Springer.

Carlin, B.P. and Louis, T.A. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. New York: Chapman & Hall.

Casella, G. and George, E.I. 1992. Explaining the Gibbs sampler. *The American Statistician* 46, 167–74.

Chib, S. 2001. Markov chain Monte Carlo methods: computation and inference. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North- Holland.

Chib, S. and Greenberg, E. 1995. Understanding the Metropolis–Hastings algorithm. *The American Statistician* 49, 327–35.

Chipman, H., George, E.I. and McCulloch, R.E. 2001. The practical implementation of Bayesian model selection. In *Model Selection. IMS Lecture Notes: Monograph Series*, ed. P. Lahiri. Beachwood, OH: Institute of Mathematical Statistics.

Clyde, M., Desimone, H. and Parmigiani, G. 1996. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* 91, 1197–208.

Doppelhofer, G. and Weeks, M. 2007. Jointness of growth determinants. *Journal of Applied Econometrics*.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* 57, 45–97.

Fernandez, C., Ley, E. and Steel, M.F.J. 2001a. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16, 563–76.

Fernandez, C., Ley, E. and Steel, M.F.J. 2001b. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100, 381–427.

Garratt, A., Lee, K., Pesaran, M.H. and Shin, Y. 2003. Forecast uncertainties in macroeconomic modelling: an application to the U.K. economy. *Journal of the American Statistical Association* 98, 829–38.

George, E.I. 1999. Discussion of Bayesian model averaging and model search strategies by M.A. Clyde. *Bayesian Statistics* 6, 175–77.

George, E.I. and Foster, D.P. 2000. Calibration and empirical Bayes variable selection. *Biometrika* 87, 731–47.

George, E. and McCulloch, R.E. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88, 881–9.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57, 1317–39.

Geweke, J. and Whiteman, C. 2006. Bayesian forecasting. In *Handbook of Economic Forecasting*, vol. 1, ed. G. Elliott, C.W.J. Granger and A. Timmermann. Amsterdam: North-Holland.

Gilks, W., Richardson, S. and Spiegelhalter, D. 1996. *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall.

Hansen, B.E. 2007. Least squares model averaging. *Econometrica* 75, 1175–89.

Hjort, N.L. and Claeskens, G. 2003. Frequentist model averaging. *Journal of the American Statistical Association* 98, 879–99.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. 1999. Bayesian model averaging: a tutorial. *Statistical Science* 14, 382–417.

Jeffreys, H. 1961. *Theory of Probability*, 3rd edn. Oxford: Clarendon Press.

Kass, R.E. and Raftery, A.E. 1995. Bayes factors. *Journal of the American Statistical Association* 90, 773–95.

Kass, R.E. and Wasserman, L. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90, 928–34.

Klein, R.W. and Brown, S.J. 1984. Model selection when there is ‘minimal’ prior information. *Econometrica* 52, 1291–312.

Koop, G. 2003. *Bayesian Econometrics*. Chichester: Wiley.

Leamer, E. 1978. *Specification Searches*. New York: Wiley.

Levin, A.T. and Williams, J.C. 2003. Robust monetary policy with competing reference models. *Journal of Monetary Economics* 50, 945–75.

Madigan, D. and York, J. 1995. Bayesian graphical models for discrete data. *International Statistical Review* 63, 215–32.

Mitchell, T.J. and Beauchamp, J.J. 1988. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83, 1023–32.

Raftery, A.E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25, 111–63.

Raftery, A.E., Madigan, D. and Hoeting, J.A. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92, 179–91.

Raftery, A.E., Madigan, D. and Volinsky, C.T. 1996. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* 5, 323–49.

Sala-i-Martin, X., Doppelhofer, G. and Miller, R.M. 2004. Determinants of economic growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94, 813–35.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.

Wasserman, L. 2000. Bayesian model selection and model averaging. *Journal of Mathematical Psychology* 44, 92–107.

Yang, Y. 2001. Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–88.

Yuan, Z. and Yang, Y. 2005. Combining linear regression models: when and how? *Journal of the American Statistical Association* 100, 1202–14.

Zellner, A. 1986. On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, ed. P.K. Goel and A. Zellner. Amsterdam: North-Holland.

Model averaging software and codes

BACC package: http://www2.cirano.qc.ca/~bacc

BACE website: http://www.nhh.no/sam/bace

BMA homepage: http://www.research.att.com/~volinsky/bma.html

BUGS project: http://www.mrc-bsu.cam.ac.uk/bugs

LeSage's Econometrics Toolbox: http://www.spatial-econometrics.com

How to cite this article

Doppelhofer, Gernot. "model averaging." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000350> doi:10.1057/9780230226203.1109

The New Palgrave Dictionary of Economics Online

model selection

Jean-Marie Dufour
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The problem of statistical model selection in econometrics and statistics is reviewed. Model selection is interpreted as a decision problem through which a statistical model is selected in order to perform statistical analysis, such as estimation, testing, confidence set construction, forecasting, simulation, policy analysis, and so on. Broad approaches to model selection are described: (1) hypothesis testing procedures, including specification and diagnostic tests; (2) penalized goodness-of-fit methods, such as information criteria; (3) Bayesian approaches; (4) forecast evaluation methods. The effect of model selection on subsequent statistical inference is also discussed.

Keywords

ARMA models; autocorrelation; Bayesian statistics; deterministic models; econometrics; endogeneity; forecasting; forecast evaluation; heteroskedasticity; linear models; model selection; models; parsimony; probability models; serial correlation; specification problems in econometrics; statistical decision theory; statistical inference; stochastic models; structural change; testing; time series analysis

Article

The purpose of econometric analysis is to develop mathematical representations of observable phenomena, which we call *models* or *hypotheses* (models subject to restrictions). Such models are then used to perform parameter estimation, test hypotheses, build confidence sets, make forecasts, conduct simulations, analyse policies, and so on. A central feature of modelling activity is the fact that models are usually interpreted as stylized (or simplified) representations that can perform certain tasks – such as prediction – but (eventually) not others, and they are treated *as if they were true* for certain purposes. Indeed, summarizing and stylizing observed phenomena can be viewed as essential components of modelling activity, which make it useful. This feature is not specific to economics and is shared by other sciences (see Cartwright, 1983).

Models can be classified as either *deterministic* or *stochastic*. *Deterministic models*, which often claim to make arbitrarily precise predictions, can be useful in theoretical activity. However, such models are rarely viewed as appropriate representations of observed data; for example, unless they are highly complex or

indeterminate, they are typically logically inconsistent with data. For this reason, models used for econometric analysis are usually *stochastic* (or *statistical*).

Formally, a statistical model is a family of probability distributions (or measures) which are proposed to represent observed data. *Model selection*, in this context, is the task of selecting a family of proposed probability distributions, which will then be used to analyse data and perform other statistical inference operations (such as parameter estimation, hypothesis testing, and so on).

A basic feature of probability models is that they are typically *unverifiable*: as for any theory that makes an indefinite number of predictions, we can never be sure that the model will not be at odds with new data. Moreover, they are *logically unfalsifiable*: in contrast with deterministic models, a probabilistic model is usually logically compatible with all the possible observation sets. Consequently, model selection can depend on a wide array of elements, such as the objectives of the model, (economic) theory, the data themselves, and various conventions.

Features which are often viewed as desirable include: (*a*) simplicity or *parsimony* (Zellner, Keuzenkamp and McAleer, 2001); (*b*) the ability to deduce testable (or falsifiable) hypotheses (Popper, 1968); (*c*) the possibility of interpreting model parameters in terms of economic theory, if not consistency with economic theory; (*d*) the ability to satisfactorily perform the tasks for which the model is built (prediction, for example); (*e*) consistency with observed data. It is important to note that these characteristics depend (at least, partially) on conventional elements, such as the objectives of the model, criteria upon which a model will be deemed ‘satisfactory’, and so on. For further discussions of these general issues, the reader may consult Poirier (1994), Morgan and Morrison (1999), Keuzenkamp (2000), Zellner, Keuzenkamp and McAleer (2001) and Dufour (2003).

In this article, we focus on statistical methods for selecting a model on the basis of the available data. Methods for that purpose can be classified in four broad (not mutually exclusive) categories:

1. 1. hypothesis testing procedures, including specification and diagnostic tests;
2. 2. penalized goodness-of-fit methods, such as information criteria;
3. 3. Bayesian approaches;
4. 4. forecast evaluation methods.

The three first approaches are meant to be applicable ‘in-sample’, while the last approach *stricto sensu* requires observations that are not available when the model is selected, but may lead to model revision. (For general reviews of the topic of statistical model selection in econometrics and statistics, see Hocking, 1976; Leamer, 1978; 1983; Draper and Smith, 1981; Judge et al., 1985, chs 7 and 21; Sakamoto, Ishiguro and Kitagawa, 1985; Grasa, 1989; Choi, 1992; Gouriéroux and Monfort, 1995, ch. 22; Charemza and Deadman, 1997; McQuarrie and Tsai, 1998; Burnham and Anderson, 2002; Clements and Hendry, 2002; Miller, 2002; Bhatti, Al-Shanfari and Hossain, 2006). It is also interesting to note that classification techniques in statistics contain results that may be relevant to model selection. This topic, however, goes beyond the scope of the present article (for further discussion, see Krishnaiah and Kanal, 1982).

Model selection and specification errors

Most model selection methods deal in different ways with a trade-off between model *realism* – which usually suggests considering relatively general, hence complex models – and *parsimony*. From the viewpoint of

estimation, for example, a model which is too simple (or parsimonious) involves *specification errors* and *biases* in parameter estimation, while too complex a model leads to parameter estimates with large variances. If the objective is forecasting, it is usually unclear which effect dominates.

For example, let us consider a linear regression model of the form

$$y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \cdots + x_{tk}\beta_k + u_t,\ t = 1, \ldots, T, \tag{1}$$

where y_t is a dependent variable and $x_{t1},\ldots,x_{tk}$ are explanatory variables, and u_t is a random disturbance which is typically assumed to be independent of (or uncorrelated with) the explanatory variables. In the classical linear model, it is assumed that the regressors can taken as fixed and that the disturbances $u_1,\ldots,u_T$ are independent and identically distributed (i.i.d.) according to a $N(0, \sigma^2)$ distribution. In this context, model selection typically involves selecting the regressors to be included as well as various distributional assumptions to be made upon the disturbances.

An especially important version of (1) is the autoregressive model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t,\ t = 1, \ldots, T. \tag{2}$$

Then a central model selection issue consists in setting the order p of the process. In such models, there is typically little theoretical guidance on the order, so data-based order selection rules can be quite useful. A related set-up where model selection is usually based on statistical methods is the class of autoregressive-moving-average (ARMA) models

$$y_t = \beta_0 + \phi y_{t-1} + \cdots + \phi_p y_{t-p} + u_t - \theta_1 u_{t-1} + \cdots + \theta_1 u_{t-q}, \tag{3}$$

where the orders p and q must be specified.

By considering the simple linear regression model, it is easy to see that excluding irrelevant variables can lead to biases in parameter estimates (Theil, 1957). On the other hand, including irrelevant regressors raises the variances of the estimators. The overall effect on the mean square error (MSE) of the estimator and, more generally, how closely it will tend to approach the parameter value may be ambiguous. It is well known that a biased estimator may have lower MSE than an unbiased estimator. This may be particularly important in forecasting, where a simple ‘false’ model may easily provide better forecasts than a complicated ‘true’ model, because the latter may be affected by imprecise parameter estimates (Allen, 1971).

Hypothesis testing approaches

Since hypothesis tests are based on a wide body of statistical theory (see Lehmann, 1986; Gouriéroux and Monfort, 1995), such procedures are widely used for assessing, comparing and selecting models.
Furthermore, econometric models are also based on economic theory which suggests basic elements that can be used for specifying models. This entails a form of asymmetry, in which restrictions suggested by economic theory will be abandoned only if ‘sufficient evidence’ becomes available. Although significance tests are meant to decide whether a given hypothesis (which usually takes the form of a restricted model) is compatible with the data, such procedures can also be used for model selection. It is interesting to note that the methodology originally proposed by Box and Jenkins (1976) for specifying ARMA models was almost exclusively based on significance tests (essentially, autocorrelation tests).
There are two basic ways of using hypothesis tests for that purpose. The first one is *forward* or *specific-to-general* approach, in which one starts from a relatively simple model and then checks whether the model can be deemed ‘satisfactory’. This typically involves various specification tests, such as:

1. 1. residual-based tests, including tests for heteroskedasticity, autocorrelation, outliers, distributional assumptions (for example, normality), and so on;
2. 2. tests for unit roots and/or stationarity, to decide whether corrections for integrated variables may be needed;
3. 3. tests for the presence of structural change;
4. 4. exogeneity tests, to decide whether corrections for endogeneity – such as instrumental variable (IV) methods – are required;
5. 5. tests for the addition of explanatory variables;
6. 6. tests of the functional form used (for example, linearity vs. nonlinearity).

There is a considerable literature on specification tests in econometrics (see Godfrey, 1988; MacKinnon, 1992; Davidson and MacKinnon, 1993). Systematic procedures for adding variables are also know in statistics as *forward selection* or *stepwise regression* procedures (Draper and Smith, 1981).
The second way is the *backward* or *general-to-specific* approach, in which one starts from a relatively comprehensive model which includes all the relevant variables. This model is then simplified by checking which variables are significant. *Backward selection* procedures in statistics (Draper and Smith, 1981) and the general-to-specific approach in econometrics (Davidson et al., 1978; Charemza and Deadman, 1997) can be viewed as illustrations of this approach.
In practical work, the backward and forward approaches are typically combined. Both involve a search for a model which is both parsimonious and consistent with the data. However, the results may differ. Specifying a model through significance tests involves many judgements and depends on idiosyncratic decisions. Further, standard hypothesis tests involve the use of typically conventional levels (such as the commonly used five per cent level). The powers of the tests can also have a strong influence on the results.

Penalized goodness-of-fit criteria

As pointed out by Akaike (1974), it is not clear that hypothesis testing is a good basis for model selection. Instead, the problem of model selection may be better interpreted as an estimation problem involving a well-

defined loss function. This leads to the topic of goodness-of-fit criteria.

A common way of assessing the performance of a regression model, such as (1), consists in computing the coefficient of determination, that is, the proportion of the dependent variable variance which is ‘explained’ by the model:

$$R^2 = 1 - \frac{\hat{V}(u)}{\hat{V}(y)} \tag{4}$$

where $\hat{V}(u) = \Sigma_{t=1}^{T} \hat{u}_t^2 / T$, $\hat{V}(y) = \Sigma_{t=1}^{T} (y_t - \bar{y})^2 / T$, $\bar{y} = \Sigma_{t=1}^{T} y_t / T$ and $\hat{u}_1, \ldots, \hat{u}_T$ are least squares residuals. This measure, however, has the inconvenient feature that it always increases when a variable is added to the model, even if it is completely irrelevant, and it can be made equal to its maximal value of one by including a sufficient number of regressors (for example, using any set of *T* linearly independent regressors).

An early way of avoiding this problem was proposed by Theil (1961, p. 213) who suggested that $\hat{V}(u)$ and $\hat{V}(y)$ be replaced by the corresponding unbiased estimators $s^2 = \Sigma_{t=1}^{T} \hat{u}_t^2 / (T - k)$ and $s_y^2 = \Sigma_{t=1}^{T} (y_t - \bar{y})^2 / (T - 1)$. This yields the adjusted coefficient of determination:

$$\bar{R}^2 = 1 - \frac{s^2}{s_y^2} = 1 - \frac{T-1}{T-k}(1 - R^2) = R^2 - \frac{k-1}{T-k}(1 - R^2).$$

It is easy to see that $\bar{R}^2$ may increase when the number of regressors increases. Note that maximizing $\bar{R}^2$ is equivalent to minimizing the ‘unbiased estimator’ s^2 of the disturbance variance. Further, if two regression models (which satisfy the assumptions of the classical linear model) are compared, and if one of these is the ‘true’ model, then the value of s^2 associated with the true model is smaller on average than the one of the other model (see Theil, 1961, p. 543). On the other hand, in large samples, the rule which consists in maximizing $\bar{R}^2$ does not select the true model with a probability converging to one: that is, it is not consistent (see Gouriéroux and Monfort, 1995, section 2.3).

Another approach consists in evaluating the ‘distance’ between the selected model and the true (unknown) model. Let $f(y)$ the density associated with the postulated model and $f_o(y)$ the density of the true model, where $Y=(y_1,\ldots, y_T)'$. One such distance is the *Kullback distance*:

$$I(f, f_o) = \int \log[f_o(y) / f(y)] f_o(y)\, dy = \underset{f_o}{E}\{\log[f_o(Y) / f(Y)]\} = \underset{f_o}{E}\{\log[f_o(Y)]\} - \underset{f_o}{E}\{\log[f(Y)]\}.$$

Minimizing $I(f, f_o)$ with respect to f is equivalent to minimizing $-E_{f_o}\{\log[f(Y)]\}$. We obtain an information criterion by selecting an 'estimator' of $E_{f_o}\{\log[f(Y)]\}$.
For the case where the model is estimated by maximizing a likelihood function $L_T(\theta)$ over a $K\times 1$ parameter vector θ, Akaike (1973) suggests that $L(\hat{\theta})$ can be viewed as a natural estimator of $E_{f_o}\{\log[f(Y)]\}$.
However, the fact that θ has been estimated introduces a bias. This bias is (partially) corrected – using an expansion argument – by subtracting the number K from $L(\hat{\theta})$. This suggests the following information criterion:

$$AIC_L(\hat{\theta}_T) = -2L_T(\hat{\theta}_T) + 2K \tag{5}$$

where K is the dimension of θ (the number of estimated parameters) and multiplication by 2 is introduced to simplify the algebra. Among a given set of models, the one with the lowest AIC is selected.
The above criterion has also been generalized by various authors leading the following general class of criteria:

$$IC_L(\hat{\theta}_T) = -2L_T(\hat{\theta}_T) + c(T, K)K \tag{6}$$

where $c(T, K)$ is a function of T and K. In the case of Gaussian models, such as (1) or (2) with i.i.d. $N(0, \sigma^2)$ disturbances, we have $L_T(\hat{\theta}_T) = -(T/2)\ln(\hat{\sigma}_T^2) + d_T$, where d_T is a constant which only depends on T, so that minimizing $IC_L(\hat{\theta}_T)$ is equivalent to minimizing

$$IC(\hat{\theta}_T) = \ln(\hat{\sigma}_T^2) + c(T, K)\frac{K}{T}. \tag{7}$$

Alternative values of $c(T, K)$ which have been proposed include:

1. 1. $c(T, K)=2$ (Akaike, 1969), which yields what is usually called the AIC criterion;

2. 2. $c(T, K)=\ln(T)$ (Schwarz, 1978);
3. 3. $c(T, K)=2\delta_T \ln(\ln T)$ where $\limsup_{T\to\infty} \delta_T > 1$ (Hannan and Quinn, 1979);
4. 4. $c(T, K) = 2 + \frac{2K(K+1)}{T-K-1}$ (Hurvitch and Tsai, 1989), which leads to the AIC_c criterion.

An especially convenient feature of such information criteria is the fact that they can be applied to both regression models (through (7)) as well as to various nonlinear models (using (6)).
Other related rules include: (*a*) criteria based on an estimate of the *final prediction error*, which try to estimate the mean square prediction error taking into account estimation uncertainty (Akaike, 1969; 1970; Mallows, 1973; Amemiya, 1980); (*b*) the *criterion autoregressive transfer* (CAT) function proposed by Parzen (1977) for selecting the order of an autoregressive process; (*c*) Sawa's (1978) Bayesian information criterion (BIC).
By far, the information criteria are the most widely used in practice. Some theoretical (non-)optimality properties have been established. In particular, when one of the models compared is the ‘true’ one, it was observed by Shibata (1976) that Akaike's criterion is not consistent, in the sense that it does not select the most parsimonious true model with probability converging to one (as the sample size goes to infinity). Instead, even in large samples it has a high probability of picking a model with ‘too many parameters’. By contrast, the criterion proposed by Hannan and Quinn (1979) is consistent under fairly general conditions, which also entails that Schwarz's (1978) criterion also leads to consistent model selection. On the other hand, the AIC criterion has a different optimality property, in the sense that it tends to minimize the one-step expected quadratic forecast error (Shibata, 1980).
On consistency, it is also interesting to observe that consistent model selection rules can be obtained provided each model is tested through a consistent test procedure (against all the other models considered) and the level of the test declines with the sample size at an appropriate rate (which depends on the asymptotic behaviour of the test statistic) (see Pötscher, 1983).
Model selection criteria of the information have the advantage of being fairly mechanical. On the other hand, they can be become quite costly to apply in practice when the number of models considered is large.

Bayesian model selection

Bayesian model selection involves comparing models through their ‘posterior probabilities’ giving observed data. Suppose we have two models M_1 and M_2 each of which postulates that the observation vector **y** follows a probability density which depends on a parameter vector: $p_y(y|\theta_1, M_1)$ under M_1, and $p_y(y|\theta_2, M_2)$ under M_2, where θ_1 and θ_2 are unknown parameter vectors (which may have different dimensions). Further, each one of the parameter vectors is assigned a ‘prior distribution’ ($p(\theta_1|M_1)$ and $p(\theta_2|M_2)$), and each model a ‘prior probability’ ($p(M_1)$ and $p(M_2)$). Then one may compute the ‘posterior probability’ of each model given the data

$$p(M_i|y) = p(M_i)\int p_y(y|\theta_i, M_1)\, p(\theta_1|M_i)\, d\theta_i,\ i = 1, 2\ . \tag{8}$$

This posterior probability of each model provides a direct measure of the ‘plausibility’ of each model. In such contexts, the ratio

$$K_{12} = \frac{p(M_1|y)}{p(M_2|y)} \tag{9}$$

is called the ‘posterior odds ratio’ of M_1 relative to M_2.
A rational decision rule for selecting between M_1 and M_2 then emerges if we can specify a loss function such as

$$L(i, j) = \text{cost of choosing } M_j \text{ when } M_i \text{ is true.} \tag{10}$$

If $L(i, i)=0$ for $i=1, 2$, expected loss is minimized by choosing M_1 when

$$K_{12} \geq \frac{L(2, 1)}{L(1, 2)}, \tag{11}$$

and M_2 when otherwise. In particular, if $L(1, 2)=L(2, 1)$, expected loss is minimized by choosing the model with the highest posterior probability. Such rules can be extended to problems where more than two models are compared.
The Bayesian approach automatically introduces a penalty for non-parsimony and easily allows the use of decision-theoretic considerations. The main difficulty consists in assigning prior distributions on model parameters and prior probabilities to competing models. For further discussion, see Zellner (1971, ch. 10), Leamer (1978; 1983), Gelman et al. (2003) and Lancaster (2004).

Forecast evaluation

In view of the fact that forecasting is one of the most common objectives for building econometric models, alternative models are often assessed by studying *post-sample* forecasts. Three types of assessments are typically considered in such contexts: (*a*) tests of predictive failure; (*b*) descriptive measures of forecast

performance, which can be compared across models; (*c*) tests of predictive ability.

A test of predictive failure involves testing whether the prediction errors associated with a model are consistent with the model. This suggests testing whether forecasts are ‘unbiased’ or ‘too large’ to be consistent with the model. The well-known predictive test for structural change proposed by Chow (1960) is an early example of such an approach. (For further discussion and extensions, see Box and Tiao, 1976; Dufour, 1980; Pesaran, Smith and Yeo, 1985; Dufour, Ghysels and Hall, 1994; Dufour and Ghysels, 1996; Clements and Hendry, 1998.)

Common measures of forecast performance involve mean errors, mean square errors, mean absolute errors, and so on (see Theil, 1961; Diebold, 2004). Although commonly used, such measures are mainly descriptive. They can usefully be complemented by tests of predictive ability. Such procedures test whether the difference between expected measures of forecast performance is zero (or less than zero) against an alternative where it is different from zero (or larger than zero). Tests of this type were proposed, among others, by Meese and Rogoff (1988), Diebold and Mariano (1995), Harvey, Leybourne and Newbold (1997), West (1996), West and McCracken (1998) and White (2000) (for reviews, see also Mariano, 2002; McCracken and West, 2002).

It is important to note that predictive performance and predictive accuracy depend on two features: first, whether the theoretical model used is close to the unknown data distribution and, second, the ability to estimate accurately model parameters (hence on sample size available for estimating these). For a given sample size, a false but parsimonious model may well have better predictive ability than the ‘true’ model.

Post-model selection inference

An important issue often raised in relation with model selection is its effect on the validity of inference – such as estimation, tests and confidence sets – obtained after a process of model selection (or pretesting). This issue is subtle and complex. Not surprisingly, both positive and negative assessments can be found.

On the positive side, it has been observed that pretesting (or model selection) does allow one to produce so-called ‘super-efficient’ (or Hodges) estimators, whose asymptotic variance can be at least as low as the Cramér–Rao efficiency bound and lower at certain points (see Le Cam, 1953). This may be viewed as a motivation for using consistent pretesting.

Furthermore, consistent model selection does not affect the asymptotic distributions of various estimators and test statistics, so the asymptotic validity of inferences based on a model selected according to such a rule is maintained (see Pötscher, 1991; Dufour, Ghysels and Hall, 1994).

On the negative side, it is important to note that these are only asymptotic results. In particular, these are pointwise convergence results, not uniform convergence results, so they may be quite misleading concerning what happens in finite samples (for some examples, see Dufour, 1997; Pötscher, 2002). For estimation, there is a considerable literature on the finite-sample distribution of pretest estimators, which can be quite different of their limit distributions (Judge and Bock, 1978; Danilov and Magnus, 2004). For a critical discussion of the effect of model selection on tests and confidence sets, see Leeb and Pötscher (2005).

Conclusion

The problem of model selection is one of the most basic and challenging problems of statistical analysis in econometrics. Much progress has been done in recent years in developing better model selection procedures

and for understanding the consequences of model selection.
But model building remains largely an art in which subjective judgements play a central role. Developing procedures applicable to complex models, which may involve a large number of candidate variables, and allowing for valid statistical inference in the presence of model selection remain difficult issues to which much further research should be devoted.

See Also

- Bayesian statistics
- econometrics
- endogeneity and exogeneity
- forecasting
- heteroskedasticity and autocorrelation corrections
- linear models
- models
- serial correlation and serial dependence
- specification problems in econometrics
- statistical decision theory
- statistical inference
- structural change
- testing
- time series analysis

Bibliography

Akaike, H. 1969. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21, 243–7.

Akaike, H. 1970. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22, 203–17.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, ed. B.N. Petrov and F. Csaki. Budapest: Akademiai Kiado.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–23.

Allen, D.M. 1971. Mean square error prediction as a criterion for selecting variables. *Technometrics* 13, 469–75.

Amemiya, T. 1980. Selection of regressors. *International Economic Review* 21, 331–54.

Bhatti, M.I., Al-Shanfari, H. and Hossain, M.Z. 2006. *Econometric Analysis of Model Selection and Model Testing*. Aldershot: Ashgate.

Box, G.E.P. and Jenkins, G.M. 1976. *Time Series Analysis: Forecasting and Control*, 2nd edn. San Francisco: Holden-Day.

Box, G.E.P. and Tiao, G.C. 1976. Comparison of forecast and actuality. *Applied Statistics* 64, 195–200.

Burnham, K.P. and Anderson, D.R. 2002. *Model Selection and Multi-model Inference: A Practical Information Theoretic Approach*. New York: Springer.

Cartwright, N. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Charemza, W.W. and Deadman, D.F. 1997. *New Directions in Econometric Practice: General to Specific Modelling, Cointegration and Vector Autoregression*, 2nd edn. Aldershot: Edward Elgar.

Choi, B. 1992. *ARMA Model Identification*. New York: Springer.

Chow, G.C. 1960. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28, 591–605.

Clements, M.P. and Hendry, D.F. 1998. *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, M.P. and Hendry, D.F., eds. 2002. *A Companion to Economic Forecasting*. Oxford: Blackwell.

Danilov, D. and Magnus, J.R. 2004. On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122, 27–46.

Davidson, J.E.H., Hendry, D.F., Srba, F. and Yeo, S. 1978. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88, 661–92.

Davidson, R. and MacKinnon, J.G. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Diebold, F.X. 2004. *Elements of Forecasting*, 3rd edn. Mason, OH: Thomson South-Western.

Diebold, F.X. and Mariano, R.S. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253–63.

Draper, N.R. and Smith, H., eds. 1981. *Applied Regression Analysis*, rev. edn. New York: Wiley.

Dufour, J.-M. 1980. Dummy variables and predictive tests for structural change. *Economics Letters* 6, 241–7.

Dufour, J.-M. 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. *Econometrica* 65, 1365–89.

Dufour, J.-M. 2003. Identification, weak instruments and statistical inference in econometrics. *Canadian Journal of Economics* 36, 767–808.

Dufour, J.-M. and Ghysels, E. 1996. Recent developments in the econometrics of structural change. *Journal of Econometrics* 70(1).

Dufour, J.-M., Ghysels, E. and Hall, A. 1994. Generalized predictive tests and structural change analysis in econometrics. *International Economic Review* 35, 199–229.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 2003. *Bayesian Data Analysis*, 2nd edn. London: Chapman and Hall/CRC.

Godfrey, L.G. 1988. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge: Cambridge University Press.

Gouriéroux, C. and Monfort, A. 1995. *Statistics and Econometric Models*, vols. 1 and 2. Cambridge: Cambridge University Press.

Grasa, A.A. 1989. *Econometric Model Selection: A New Approach*. Dordrecht: Kluwer.

Hannan, E.J. and Quinn, B. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41, 190–1.

Harvey, D.I., Leybourne, S.J. and Newbold, P. 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13, 281–91.

Hocking, R.R. 1976. The analysis and selection of variables in linear regression. *Biometrika* 32, 1–49.

Hurvitch, C.M. and Tsai, C.-L. 1989. Regression and time series model selection in small samples. *Biometrika* 76, 297–307.

Judge, G.G. and Bock, M.E. 1978. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.

Judge, G.G., Griffiths, W.E., Carter Hill, R., Lütkepohl, H. and Lee, T.-C. 1985. *The Theory and Practice of Econometrics*, 2nd edn. New York: Wiley.

Keuzenkamp, H.A. 2000. *Probability, Econometrics and Truth: The Methodology of Econometrics*.

Cambridge: Cambridge University Press.

Krishnaiah, P.R. and Kanal, L.N. 1982. *Handbook of Statistics 2: Classification, Pattern Recognition and Reduction of Dimensionality*. Amsterdam: North-Holland.

Lancaster, T. 2004. *An Introduction to Modern Bayesian Econometrics*. Oxford: Blackwell.

Le Cam, L. 1953. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics* 1, 277–330.

Leamer, E. 1978. *Specification Searches: Ad Hoc Inferences with Nonexperimental Data*. New York: Wiley.

Leamer, E.E. 1983. Model choice. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Leeb, H. and Pötscher, B. 2005. Model selection and inference: facts and fiction. *Econometric Theory* 21, 29–59.

Lehmann, E.L. 1986. *Testing Statistical Hypotheses*, 2nd edn. New York: Wiley.

MacKinnon, J.G. 1992. Model specification tests and artificial regressions. *Journal of Economic Literature* 30, 102–46.

Mallows, C.L. 1973. Some comments on *Cp*. *Technometrics* 15, 661–75.

Mariano, R.S. 2002. Testing forecast accuracy. In Clements and Hendry (2002).

McCracken, M.W. and West, K.D. 2002. Inference about predictive ability. In Clements and Hendry (2002).

McQuarrie, A.D.R. and Tsai, C.-L. 1998. *Regression and Time Series Model Selection*. Singapore: World Scientific.

Meese, R.A. and Rogoff, K. 1988. Was it real? The exchange rate-interest differential relation over the modern floating-rate period. *Journal of Finance* 43, 933–48.

Miller, A. 2002. *Subset Selection in Regression*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC.

Morgan, M.S. and Morrison, M. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Parzen, E. 1977. Multiple time series: determining the order of approximating autoregressive schemes. In *Multivariate Analysis IV: Proceedings of the Fourth International Symposium on Multivariate Analysis*, ed. P.R. Krishnaiah. Amsterdam: North-Holland/Elsevier.

Pesaran, M.H., Smith, R.P. and Yeo, J.S. 1985. Testing for structural stability and predictive failure: a review. *Manchester School* 3, 280–95.

Poirier, D.J. 1994. *The Methodology of Econometrics*, 2 vols. Aldershot: Edward Elgar.

Popper, K. 1968. *The Logic of Scientific Discovery*, rev. edn. New York: Harper Torchbooks.

Pötscher, B.M. 1983. Order estimation in ARMA-models by Lagrange multiplier tests. *Annals of Statistics* 11, 872–85.

Pötscher, B. 1991. Effects of model selection on inference. *Econometric Theory* 7, 163–85.

Pötscher, B. 2002. Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots and estimation of long memory parameters. *Econometrica* 70, 1035–65.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. 1985. *Akaike Information Criterion Statistics*. Dordrecht: Reidel.

Sawa, T. 1978. Information criteria for discriminating among alternative regression models. *Econometrica* 46, 1273–91.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.

Shibata, R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 71, 117–26.

Shibata, R. 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8, 147–64.

Theil, H. 1957. Specification errors and the estimation of economic relationships. *Review of the International Statistical Institute* 25, 41–51.

Theil, H. 1961. *Economic Forecasts and Policy*, 2nd edn. Amsterdam: North-Holland.

West, K.D. 1996. Asymptotic inference about predictive ability. *Econometrica* 64, 1067–84.

West, K.D. and McCracken, M.D. 1998. Regression-based tests of predictive ability. *International Economic Review* 39, 817–40.

White, H. 2000. A reality check for data snooping. *Econometrica* 68, 1097–126.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Zellner, A., Keuzenkamp, H.A. and McAleer, M. 2001. *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*. Cambridge: Cambridge University Press.

How to cite this article

Dufour, Jean-Marie. "model selection." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000379> doi:10.1057/9780230226203.1110

The New Palgrave Dictionary of Economics Online

model uncertainty

Alexei Onatski

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Model uncertainty is a condition of analysis when the specification of the model of analysed process is open to doubt. A failure to account for model uncertainty may result in poor decisions. This article reviews various approaches to representing model uncertainty. The approaches depend on the research context, differ in their degree of generality, and may be classified as deterministic versus stochastic, Bayesian versus frequentist, and treating model uncertainty as static versus viewing model uncertainty as evolving over time.

Keywords

ambiguity aversion; approximating model; axioms of choice; Bayesian statistics; Bootstrap; detection error probability; dynamic stochastic general equilibrium models; Ellsberg paradox; ε -contaminated priors; heterogeneity uncertainty; identification scheme; identified vector autoregressions; incommensurability; incomplete information; judgement; Laplace transform; learning; linear quadratic Gaussian problem; Markov processes; model averaging; model expansion; model uncertainty; model uncertainty set; models; monetary policy; overconfidence; probability; rational expectations; relative entropy; risk-sensitive control; specification uncertainty; theory uncertainty; uncertainty; vector autoregressions

Article

Model uncertainty is a condition of analysis when the specification of the model of analysed process is open to doubt. One of the fundamental sources of model uncertainty is the tradition of critical reasoning that, in words of Karl Popper (1962, pp. 151–3), ‘admits a plurality of doctrines which all try to approach the truth by means of critical discussion’. Popper traces the critical tradition back to ancient Greek philosophy. He cites Xenophanes, 570–480 bc, who wrote (see Diels, 1951, vol. 1, pp. 133 and 137):

The gods did not reveal, from the beginning,
All things to us; but in the course of time,
Through seeking, men find that which is the better …

But as for certain truth, no man has known it,
Nor will he know it; neither of the gods,
Nor yet of all the things of which I speak.
And even if by chance he were to utter
The final truth, he would himself not know it:
For all is but a woven web of guesses.

Another fundamental source of model uncertainty is the necessity for models to be simple enough to provide an efficient link between theory and reality (see Morgan and Morrison, 1999, for a book-length discussion of the nature of models). Complicated models may be less useful than simple ones even though the accuracy of the simple models' description of the modelled process may be more doubtful. The understanding of what constitutes a model and how to model model uncertainty itself depends on the research context. For example, for engineers a prototypical model is represented by a system of differential equations:

$$\dot{x}(t) = Ax(t) + Bu(t),$$
$$y(t) = Cx(t) + Du(t),$$

where $x(t)$, $u(t)$, and $y(t)$ are square-integrable functions of $t \in [0,\infty)$, interpreted, respectively, as internal states, inputs and outputs of the modelled mechanism. By means of the Laplace transform, we get an alternative representation of the above model (with zero initial states):

$$\hat{y}(s) = M(s)\hat{u}(s), M(s) = D + C(sI - A)^{-1}B,$$

where $\hat{u}(s)$ and $\hat{y}(s)$ are the Laplace transforms of $u(t)$ and $y(t)$ (see Kwakernaak and Sivan, 1972, p. 33). For engineers, the interest often lies in checking whether particular inputs into the modelled mechanism are such that the corresponding internal states and outputs satisfy an admissibility criterion. Since the model is only an approximation of the mechanism, the check of the admissibility criterion should take into account possible deviations of the model from the truth. These possible deviations constitute model uncertainty, which can be represented by a set of models built around the above reference model.
A model of the uncertainty set which is very flexible and well suited for the purpose of the admissibility

checks is the so-called linear fractional model (see Zhou, Doyle and Glover, 1996, chs 10 and 11). It replaces the reference model represented by $M(s)$, by a set of models:

$$M(\Delta) = M(s) + L(s)\Delta(I - \Delta G(s))^{-1}R(s), \Delta \in \Lambda,$$

where Λ is a set of block-diagonal matrices with the largest singular value bounded by unity. The number and the structure of the blocks, and the form of matrix functions $L(s)$,$R(s)$, and $G(s)$ must be chosen so that the resulting model uncertainty set accurately represents the engineer's understanding of possible deviations of the reference model from the modelled mechanism.

To take another example, for researchers in statistics a model is often defined as a family of the joint probability distributions of the data. Draper (1995, pp. 45–6) notes that statistical models can be expressed in two parts, the first representing structural assumptions, such as distributional choices for residuals, or a particular functional form of the regression function and so on, and the second representing parameters, whose meaning is specific to the assumed structure. He points out that ‘even in controlled experiments and randomized sample surveys key aspects of … [the structure] will usually be uncertain, and this is even more true with observational studies’. Statistical model uncertainty can be interpreted as the structural uncertainty that Draper is concerned about in the above citation.

A failure to account for statistical model uncertainty often leads to overconfidence in the results of a statistical study. For example, the forecast intervals, which are computed ignoring possible model uncertainty, may be too narrow, p-values of a test of significance of coefficients in a linear regression too small, and so forth.

Typically, statistical models analyse reality, which is much more complicated than man-made mechanisms. Therefore, building a crisp set of models that represent model uncertainty is more problematic in statistics than in engineering. Most of the statistical approaches to modelling model uncertainty are Bayesian. They represent model uncertainty by a prior distribution defined in the model space and propagate this uncertainty to the statistical decisions by integrating models out from the posterior distribution. Such a technique of model uncertainty propagation is called Bayesian model averaging (see Hoeting et al., 1999 for a tutorial).

There are no standardized ways of specifying a prior that would represent model uncertainty. One approach is to expand a given model to a more general class and to formulate a subjective prior over this class. An early example of this approach can be found in Box and Tiao (1962), who re-examine Darwin's paired data on the heights of self- and cross-fertilized plants earlier analysed by Fisher (1935). To take into account a possible misspecification of Fisher's model for differences in the heights of the i-th pair of plants, y_i:

$$y_i = \theta + \sigma e_i, \quad e_i \propto i.i.d.\ N(0, 1),$$

Box and Tiao expand it to a more general model:

$$y_i = \theta + \sigma e_i, \quad e_i \text{ are i.i.d. with density proportional to } \exp\left(-\frac{1}{2}|e|^{2/(1+\beta)}\right),$$

and formulate a beta-type prior distribution for the ‘extra’ parameter β.

As emphasized by Draper et al. (1987, p. 12), for the model expansion approach to be successful it is important to ‘stake out the corners in model space’, that is, to find ‘the plausible variations on the model… that strongly influence what actions would be taken’. Of course, such an exercise would necessarily be subjective and context specific.

An interesting frequentist alternative to specifying a prior in the model space is to bootstrap the modelling process. Efron and Gong (1983) consider a databased process of explanatory variable selection for a logistic model of the probability of death from a certain disease. They apply the selection process to bootstrap replications of the data, obtaining, thus, a distribution of logistic models, which represents uncertainty about the model whose explanatory variables were chosen on the basis of the original data-set.

If we turn to a discussion of model uncertainty in economics, we first note that the economic model uncertainty is much broader in scope than engineering or statistical model uncertainty. The economic reality is so complex that it may be impossible in principle to approximate it by any model. Different research communities may disagree on what should be understood by economic reality in the first place. For example, Frankel and Rockett (1988, p. 318), in their study of potential gains from cooperation of different countries’ monetary authorities, write: ‘The assumption that policy makers agree on the true model has little, if any, empirical basis. Different governments subscribe to different economic philosophies.’

The idea of incommensurability of different views of economic reality is a focus of Dow's (2004) methodological study of model uncertainty. Dow puts the incommensurability idea in the context of uncertainty research originating in Keynes's (1921) *Treatise on Probability*, and discusses the role of judgement in a situation when it is impossible in principle to compare models on the basis of their closeness to ‘the truth’.

Further, in views of Keynes (1921) and Knight (1921), economic uncertainty may be conceptually different from the uncertainty modelled by randomness. So even if the incommensurability issue does not arise, an economic decision-maker may be hesitant in assigning probabilities to different economic models and comparing them on the basis of these probabilities. Such a view is supported by a range of experimental studies initiated by the Ellsberg paradox.

The Ellsberg paradox shows that people prefer to bet on 50–50 lotteries rather than on lotteries with completely unknown odds. Such behaviour is a variant of a more general phenomenon called ambiguity aversion. It reveals that people fail to assign prior probabilities to events that happen in incomplete information environments.

As Gilboa and Schmeidler (1989) show, failing to assign prior probabilities to events is perfectly

rational because it is consistent with axioms of choice as reasonable as those used by Savage (1954). Gilboa and Schmeidler's axioms imply that a rational decision-maker acts as if he or she contemplates a set of probability distributions over the possible events. The decision is then made so as to minimize the expected loss under the worst possible distribution from the set.

Much of modern research on economic model uncertainty concerns monetary policy formulation and evaluation when policymakers do not have a single reliable model of the economy. In the rest of this article we will therefore focus on the model uncertainty arising in monetary policy research.

Global model uncertainty

Different approaches to macroeconomic model uncertainty can roughly be separated in two broad categories, which Brock, Durlauf and West (2003) call global and local approaches. The global approach assumes that a set of possible models consists of the substantially different economic theories. An early example of the global approach is posited by McCallum (1988), who uses a real business cycle model, a monetary misperception model and a Keynesian model to represent model uncertainty confronted by a monetary policymaker. In contrast, the local approach builds the model uncertainty set by continuously expanding a single reference model.

Brock, Durlauf and West (2003) distinguish three different components of model uncertainty in the global approach. The first component is ‘theory uncertainty’, which represents economists’ ‘disagreement over fundamental aspects of the economy’. The second component is ‘specification uncertainty’, which includes uncertainty about lag length specification, functional form, and the choice of proxy variables representing particular theoretical concepts. The last component is ‘heterogeneity uncertainty’ which ‘concerns the extent to which different observations are assumed to obey a common model’.

A model for the global model uncertainty itself is based on a set of models which represent different theories, have different specifications given a particular theory, and may include dummy variables or other devices that capture possible data heterogeneity. Brock, Durlauf and West (2003) propose to complete the model of model uncertainty by specifying a prior distribution over the models in the set. They propose three principles that should guide the formulation of the prior. First, it ‘should assign relatively high probability to those areas of the likelihood that are relatively large’ (see, however, Chris Sims's critique of this principle in the discussion published with Brock, Durlauf and West, 2003). Second, ‘a prior should be robust in the sense that a small change in the prior should not induce a large change in the posterior’. Finally, ‘priors should be flexible enough to allow for their use across similar studies’.

To accommodate the possibility of the ambiguity aversion on behalf of policymakers, Brock, Durlauf and West (2003) suggest that the chosen prior, π, be extended to a class of ε-contaminated priors $\{(1-\varepsilon)\pi+\varepsilon P, P\in P(M)\}$, where $0\leq\varepsilon\leq 1$ and $P(M)$ is the set of all possible probability measures on the model uncertainty set. The policy which takes into account the model uncertainty can then be chosen by minimizing the expected posterior loss under the worst possible prior from the ε-contaminated class. Classes of ε-contaminated priors are often used in robust Bayesian analysis to model uncertainty in the prior distribution (see Berger and Berliner, 1986). Such classes are easy to work with, and it is not difficult to show that the policy described above differs from the policy which minimizes expected

posterior loss under the original prior by putting an extra weight on the worst possible model from the model uncertainty set. The higher the ε, the larger the extra weight.

In the extreme case when $\varepsilon = 1$, specifying prior probabilities of the models from the model uncertainty set is not necessary. The very set completely describes model uncertainty. The policy is then formulated as if the worst possible model were true. The policy choice under the extreme model uncertainty is often visualized as a zero-sum game between a policymaker and malevolent 'nature' who chooses adversary models from the model uncertainty set. Early advocates of this useful visualization were Brunner and Meltzer (1969) and von zur Muehlen (2001).

Describing model uncertainty by an un-weighted set of models was advocated by John Tukey. He says in his comment on Draper (1995) (see Draper, 1995, p. 78):

> The most acceptable pattern, as far as I am concerned, for the development of a bouquet of models begins with a predata choice of a collection of models likely to be relevant in the field in question, followed by an examination of the reasonability of the data in the light of each model. For those models for which the data seem unreasonable, we have a choice:
>
> 1. (a) drop them from consideration or
> 2. (b) move them sufficiently close to a smoothed version of the data to make the data reasonable.
>
> Here reasonability is a yes-no decision, not a probability reduction, and the models are thought as challenges, trying to mark the boundaries of reasonability, not to represent likely outcomes. Taking the worst of what remains is a conservative but, in my judgment, reasonable step.

Local model uncertainty

An unweighted set description of model uncertainty is also preferred by Lars Hansen and Thomas Sargent, who initiated a broad research programme addressing model uncertainty in macroeconomics (see Hansen and Sargent, 2006, for a book-length development of their research plan). In contrast to Tukey, their choice of the unweighted representation is primarily motivated by the difficulty of formulating a sensible prior over a large set of models.

Hansen and Sargent's approach to model uncertainty is an example of the local approach. They assume that by an unspecified search process a policymaker comes to a single approximating model of the economy. Then, the model uncertainty set is built around this model. The set includes all models that are statistically difficult to distinguish from the original approximating model.

More formally, Hansen and Sargent (2006, p. 8) consider a policymaker whose approximating model can be formalized as a Markov process characterized by transition density $f(y_t|y_{t-1})$, where y_t is a state vector at time t. The policymaker's model uncertainty set consists of the Markov processes with transition densities $g(y_t|y_{t-1})$, which are difficult to statistically distinguish from the approximating model in the sense that the expected discounted sum of conditional relative entropies of models g with

respect to model *f* is reasonably small:

$$E_g \sum_{t=0}^{\infty} \beta^t \int \log\left(\frac{g(z|Y_t)}{f(z|Y_t)}\right) g(z|Y_t)\, dz \leq \eta.$$

The conditional relative entropy measures the mean information for discrimination between g and f on the basis of a new observation of the state vector, which comes from *g* (see Kullback and Leibler, 1951). What ‘reasonably small’ means depends on how uncertain the policymaker is about her approximating model. When η is large, the amount of uncertainty may be very large. Hence, the classification of the Hansen–Sargent approach as ‘local’ does not mean that the uncertainty they address is insignificant. Anderson, Hansen and Sargent (2003) relate η to a more transparent concept of detection error probability, which can be used to calibrate η.

Using the relative entropy concept for the formulation of the model uncertainty set is very convenient for design of macroeconomic policy that works well across all models from the set. In an engineering context, Petersen, James and Dupuis (2000) show how to construct a risk-sensitive control problem which has the same solution as the problem of finding a controller that maximizes the worst possible performance over the set of models subject to the relative entropy constraint. The risk-sensitive control problem is extensively studied in Whittle (1990). It has a very simple solution, which is a modification of a standard solution of the linear quadratic Gaussian problem. Hansen and Sargent (2006) substantially modify and extend the control methods so that they are applicable to economic problems.

A very important economic setting that calls for an extensive modification of the engineering ideas is that with multiple decision-makers. The rational expectations literature assumes that the economic agents living inside the model and the policymaker who uses the model to formulate her policy agree on the model. The possibility that the agents and the policymaker have doubts about the model calls for a revision of the rational expectations paradigm. Giannoni (2002) is an early example of a study that assumes model uncertainty on behalf of the policymaker but requires the modelled economic agents to know the true model. Hansen and Sargent (2003) consider a situation when the policymaker and the economic agents are uncertain about a common approximating model.

Another example of the local approach to model uncertainty is provided by Schorfheide and Del Negro (2005). In contrast to Hansen and Sargent, they represent model uncertainty about an approximating model by a prior distribution in the model space, centred at the approximating model. To cope with the difficulty of specifying sensible and manageable priors over a vast set of models, they restrict attention to the alternative models that have form of identified vector autoregressions (VARs). The prior density over the alternative models is taken to be proportional to the relative entropy distance between the alternative and the approximating model, which is chosen to be a state-of-the-art dynamic stochastic general equilibrium model.

Using identified VARs for construction of the model uncertainty set potentially raises an extra model uncertainty issue: which identification scheme to use to identify structural shocks in VARs? Different identification schemes cannot be evaluated on the basis of data because the implied identified VARs are

observationally equivalent. To take into account uncertainty about the identification schemes, Faust (1998) proposes forming a set of identified VARs so that the corresponding impulse responses look reasonable in the sense that they are consistent with some particularly strong prior beliefs about the effects of structural shocks.

Evolving model uncertainty

Schorfheide and Del Negro's (2005) analysis of policy choice under model uncertainty is one of few studies that allows for changes in the model uncertainty depending on the prospective policy choice. Such flexibility comes from their obtaining the joint posterior distribution for the set of possible models and policy parameters. As long as policy parameters are set in the historically observed region, the policymaker can take as his or her model of model uncertainty the posterior distribution over the set of models conditional on the particular parameter values.

Policymakers' perceptions of model uncertainty may depend on many factors beyond particular policy choices. As new data emerge, policymakers learn and adjust their model uncertainty sets. Even more importantly, unforeseen events may substantially change the set of possible models. This fact is at the heart of Keynes's (1939, p. 567) critique of Tinbergen's econometric method: '[The] main prima facie objection to the application of the method of multiple correlation to complex economic problems lies in the apparent lack of any adequate degree of uniformity in the environment.'

An interesting theory of learning under the condition of model uncertainty in a non-stationary environment is Epstein and Schneider (2006). These authors consider a decision-maker who receives a sequence of signals generated by an uncertain model. Some features of the model are constant over time. Those features are represented by a parameter θ, which the decision-maker hopes to learn about, although it is ambiguous initially. Other features may 'vary over time in a way … [the decision-maker] does not understand well enough even to theorize about and therefore she does not try to learn about them' (see Epstein and Schneider, 2006, p. 3). These features are captured by an assumption that the decision-maker considers a non-singleton set of data distributions, which are all parameterized by θ but have different structure. Which structure is used to deliver observations may erratically change over time. Epstein and Schneider show how the set of priors for θ changes over time, and prove that, under certain regulatory conditions, it converges to a distribution assigning probability 1 to a single vector θ^* so that the ambiguity about θ is asymptotically resolved. On the contrary, by assumption, the uncertainty associated with the multiplicity of the structures representing the poorly understood factors influencing the dynamics is never resolved or learned about.

Can we form any idea about the nature and the strength of the poorly understood factors? After all, we have to somehow specify a set of distributions representing these factors to analyse model uncertainty. Tetlow (2006) may be a first step in answering this question. He studies the real-time evolution of the principal macroeconomic model of the Federal Reserve Board in the 1996–2003 period. He finds a surprisingly large amount of variation in the model over the period, and shows how the changes in the model were driven by the data and 'the economic issues of the day'.

The literature on model uncertainty is large and rapidly growing. In engineering, the entire field of automatic control is motivated to a large extent by issues of robustness to model uncertainty. Although above we have given an example of engineers' approach to model uncertainty, we have not even

scratched upon the surface of the literature. Similarly, many important approaches to model uncertainty in statistics and economics have been left aside. We hope, however, that the reader has gained a general idea about the topic and will find a further discussion in the references provided below.

See Also

- ambiguity and ambiguity aversion
- model averaging
- models
- robust control
- specification problems in econometrics
- uncertainty

Bibliography

Anderson, E., Hansen, L.P. and Sargent, T.J. 2003. A quartet of semigroups for model specification, robustness, prices of risk, and model detection. *Journal of the European Economic Association* 1, 68–123.

Berger, J. and Berliner, L.M. 1986. Robust Bayes and empirical Bayes analysis with ε -contaminated priors. *Annals of Statistics* 14, 461–86.

Box, G.E.P. and Tiao, G.C. 1962. A further look at robustness via Bayes's theorem. *Biometrika* 49, 419–32.

Brock, W.A., Durlauf, S.N. and West, K.D. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 2003(1), 235–322.

Brunner, K. and Meltzer, A. 1969. The nature of policy problem. In *Targets and Indicators of Monetary Policy*, ed. K. Brunner and A. Meltzer. San Francisco: Chandler Publishing Company.

Diels, H. 1951. *Die Fragmente der Vorsokratiker*. Berlin: Weidmannsche Verlagsbuchhandlung.

Dow, S.C. 2004. Uncertainty and monetary policy. *Oxford Economic Papers* 56, 539–61.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society* B 57, 45–97.

Draper, D., Hodges, J.S., Leamer, E.E., Morris, C.N. and Rubin, D.B. 1987. A research agenda for assessment and propagation of model uncertainty. Report No. 2683-RC. Santa Monica: Rand

Corporation.

Efron, B. and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistican* 37, 36–49.

Epstein, L.G. and Schneider, M. 2006. Learning under ambiguity. Working Paper No. 527, Rochester Center for Economic Research.

Faust, J. 1998. The robustness of identified VAR conclusions about money. *Carnegie-Rochester Conference Series on Public Policy* 49, 207–44.

Fisher, R.A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Frankel, J.A. and Rockett, K.E. 1988. International macroeconomic policy coordination when policymakers do not agree on the true model. *American Economic Review* 78, 318–40.

Giannoni, M. 2002. Does model uncertainty justify caution? Robust optimal monetary policy in a forward-looking model. *Macroeconomic Dynamics* 6, 111–44.

Gilboa, I. and Schmeidler, D. 1989. Maximin expected utility with non-unique priors. *Journal of Mathematical Economics* 18, 141–53.

Hansen, L.P. and Sargent, T.J. 2003. Robust control of forward-looking models. *Journal of Monetary Economics* 50, 581–604.

Hansen, L.P. and Sargent, T.J. 2006. Robustness. Working paper, New York University.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. 1999. Bayesian model averaging: a tutorial. *Statistical Science* 14, 382–417.

Keynes, J.M. 1921. *A Treatise on Probability*. London: Macmillan. Repr. for the Royal Economic Society as *Collected Writings*, vol. 8, 1973.

Keynes, J.M. 1939. Professor Tinbergen's method. *Economic Journal* 49, 558–68.

Knight, F.H. 1921. *Risk, Uncertainty and Profit*. New York: Houghton Mifflin.

Kullback, S. and Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.

Kwakernaak, H. and Sivan, R. 1972. *Linear Optimal Control Systems*. New York: John Wiley & Sons.

McCallum, B.T. 1988. Robustness properties of a rule for monetary policy. *Carnegie-Rochester Conference Series on Public Policy* 29, 173–203.

Morgan, M.S. and Morrison, M. 1999. *Models as Mediators*. Cambridge: United Kingdom: Cambridge University Press.

Petersen, I.R., James, M.R. and Dupuis, P. 2000. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control* 45, 398–412.

Popper, K.R. 1962. *Conjectures and Refutations*. New York, London: Basic Books.

Savage, L.G. 1954. *The Foundations of Statistics*. New York: Dover Publications.

Schorfheide, F. and Del Negro, M. 2005. Monetary policy analysis with potentially misspecified models. Mimeo, University of Pennsylvania.

Tetlow, R.J. 2006. Real-time model uncertainty in the United States: ‘robust’ policies put to the test. Manuscript, Division of Research and Statistics, Federal Reserve Board.

von zur Muehlen, P. 2001. Activist vs. non-activist monetary policy: optimal rules under extreme uncertainty. Finance and Economics Discussion Series Working Paper No. 2, Federal Reserve Board.

Whittle, P. 1990. *Risk-Sensitive Optimal Control*. New York: Wiley.

Zhou, K., Doyle, J.C. and Glover, K. 1996. *Robust and Optimal Control*. Upper Saddle River, NJ: Prentice Hall.

How to cite this article

Onatski, Alexei. "model uncertainty." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000407> doi:10.1057/9780230226203.1111

The New Palgrave Dictionary of Economics Online

models

Mary S. Morgan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Philosophical analysis of the historical development of modelling, as well as the programmatic statements of the founders of modelling, support three different functions for modelling: for fitting theories to the world; for theorizing; and as instruments of investigation. Rather than versions of data or of theories, models can be understood as complex objects constructed out of many resources that defy simple description. These accounts also suggest a kinship between the ways models work in economics and various kinds of experiment, found most obviously in simulation but equally salient in older traditions of mathematical and statistical modelling.

Keywords

business cycles; caricatures; correspondence rules; Cowles Commission; design of experiments; econometrics; economic man; Edgeworth, F.Y; experiments and econometrics; Fisher, I.; Friedman, M.; Frisch, R.A.K.; Haavelmo, T.; ideal type; instrumentalism; idealization; inference; Koopmans. T.C.; laboratory experiments in economics; Lucas, R.; macroeconometric models; matching; mathematical economics; mathematics and economics; metaphor; methodology of economics; Mill, J.S.; Marshall, A.; model design; models; model construction; model functions; model experiments; National Bureau of Economic Research; Pigou, A.C.; prediction; probability; Quesnay, F.; random shock models; Shubik, M.; simulation; Slutsky, E.; Sutton, J.; statistical inference; statistics and economics; Tableau économique; tendency laws; testing; Tinbergen, J.; Weber, M.

Article

Modelling became the dominant methodology of economics during the 20th century.
Yet, despite its ubiquitous usage in modern economics, the term ‘model’ was introduced relatively recently. In the late 19th century, ‘models’ were not even a recognized category in discussions about methodology (as for example in Palgrave's *Dictionary of Political Economy* of the 1890s), although a

few existed as practical working objects. The effective usage of the term ‘model’ in economics is associated with the econometrics movement of the interwar period, a movement whose aim was both to develop and to meld together mathematical and statistical approaches to economics. From this original broad notion, in the 1950s grew separate fields of mathematical economists and econometricians, and both maintained modelling as a central tool of their scientific practice. It became conventional then to think of models in modern economics as either mathematical objects used in economic theory or as econometric objects (involving both statistical and mathematical properties) in empirical work. Historical accounts of models in modern economics may conveniently begin then with this division. Philosophical commentaries, too, have mostly tended to follow this division, treating the models of economic theory as different kinds of creatures, with different roles, from those which are applied to data. The latter role of models, that of ‘fitting theories to the world’, is exemplified in the empirical modelling, econometric work and methodological statements of Jan Tinbergen in the 1930s. By contrast, the mathematical models of modern economics are primarily viewed as a way in which economic theory building goes on. This ‘modelling as theorizing’ view is exemplified in the programmatic pronouncements of Tjalling Koopmans in 1957. A third methodological framework presents ‘models as investigative instruments’: tools to learn about economic theory or the economic world, a position typified in the late 19th century and early 20th century work of Irving Fisher, who might be seen as another of the founders of modelling in economics.

This article covers the historical emergence and roles of models in economics according to these three different methodological accounts, and discusses how these approaches fit into the modern science of economics.

Modelling as fitting theories to the world

Although a vibrant econometrics community developed in the two decades up to the 1920s, its products (regressions of demand, statistical accounts of business cycles and so forth) were presented as direct descriptions of the underlying economic relations, rather than as models put forward tentatively to represent them (see Morgan, 1990). The difference is a subtle one, but illuminated by Philippe Le Gall's (2007) use of the term ‘natural econometrician’ for those 19th century economists who believed, in parallel to the natural sciences, that the laws that governed the economy were written in mathematics, and clever manipulation of statistical data (without, it must be said, much in the way of analytical techniques) would reveal these laws.

Into this descriptive statistical framework, Jan Tinbergen not only introduced the term ‘model’ in 1935 (see Boumans, 1993) but he was also responsible – along with Ragnar Frisch – for the development of such joint mathematical–statistical objects in the econometrics of the 1930s. (Prior to this, the rare use of the term ‘model’ typically referred to physical object models as Boltzmann defined them in 1911. Paul Ehrenfest is the probable source of a broadening in scope of the term to include mathematical models, and Tinbergen was his assistant during the mid-1920s; see Boumans, 2005, ch. 2.) Frisch in 1933 had developed – in the context of business cycle research – the notion of a ‘macro-dynamic scheme’: a three-equation model with random errors. He even simulated it to show that it could reproduce the generic characteristics of time-series data of his time. But it was Tinbergen who developed Frisch's design into an econometric model – a model that could be fitted to real data from the economy. As is well known,

he built the first generation of macroeconometric models (see Tinbergen, 1937; 1939; and Bodkin, Klein and Marwah, 1991), and in doing so he made explicit the notion of a model as a vehicle for bridging the gap between theories of the business cycle and specific (time and place) statistical data of the cycle, as Morgan (1990, ch. 4) argues. To appreciate the task, it needs to be remembered that most existing theories of the cycle were expressed verbally, and the nascent mathematical theories of the cycle were too small and simplified to represent the characteristics of real cycles, so even building up a system of equations from these theories was a considerable task. The data played a role, too, in deciding the time sequence of the relations and which variables should be included or omitted, for both these elements were determined in the statistical work. In other words, Tinbergen created a set of usable mathematical–statistical relations which both incorporated theoretical ideas about how the economy worked and represented empirically the different parts of the economy. Having fitted theories and data together in the format of the econometric model, he then used the model to test the viability of various theories of the cycle, to explain events in the economy, and to run the model forward with different policy options relevant for the Great Depression years – all this in the pre-computer age using hand calculators! This ‘new practice’ of models, as Boumans (2005) terms it, involved a creative building of mathematical economic theory in relation to the statistical data of the economic world and of craft skill in using those models. For both Frisch and Tinbergen, modelling was a project to explain how the economic world worked.

The next stage in the history may be marked by Trygve Haavelmo's famous blueprint for econometrics of 1944 which brought another subtle change of focus to the task of econometric modelling. He suggested that econometrics ought to be concerned, not with a process of matching theory and data in an iterative process, but with finding the correct model for the observed data using probability reasoning (see Morgan, 1990, ch. 8). He effectively introduced into econometrics not only the notion of the theoretical model (the mathematical model derived from a priori theory) but also that of the ‘true’ (but unknown) model: ‘the ‘true’ mechanism under which the data considered are being produced’ (Haavelmo, 1944, p. 49). Yet he was by no means a ‘natural econometrician’ (in Le Gall's sense for the 19th century), arguing of models of economic behaviour that ‘whatever be the ‘explanations’ [of economic phenomena] we prefer, it is not to be forgotten that they are all our own artificial inventions in a search for an understanding of real life; they are not hidden truths to be “discovered‘” (Haavelmo, 1944, p. 3). Though he urged that a well-fitting econometric model (a theory which fits the data well) might not be the ‘true’ model, nevertheless, his blueprint probability approach was destined to alter the accepted task of econometrics. The Cowles Commission approach that followed (whose contributions are analysed by Qin, 1993, and Epstein, 1987) stressed the use of the correct methods of identifying and estimating the theoretically derived complete structural model as the means to discover that true model. The ‘strong apriorism’ of their approach to econometrics, in which theory proposes the model and the data dispose (or not) of these hypotheses, sparked the famous ‘measurement without theory’ debate with the more empiricist branch of the field over how to do econometrics in the late 1940s.

It is tempting to see Haavelmo's provision of a philosophical basis for econometrics as paving the way for a post-1950 division of labour in the use of models – namely, the economists provide mathematical models from economic theory, and the job of the econometrician is to use statistics for model estimation and theory testing. To some extent this division of labour is borne out, for it is in this period that a much

clearer distinction emerges between theoretical and applied economics (as seen in Backhouse, 1998). However, despite the rhetoric of post-1950 econometrics which talks of ‘confronting theory with data’, or ‘applying theory to data’, from the point of view of econometric modelling the practical division is not nearly so clear-cut. There are several reasons. First, it remains a prosaic but generally valid comment that theory rarely provides all the resources needed to make models that can be immediately applied to the data from the world. This is precisely why econometric models have featured as a necessary intermediary, a matching device, between them. Second, this matching process of fitting theories to the world is done with many different purposes – to test theories, to measure relations, to explain events, and so on – each needing different resources from theory and with different criteria. Third, there are no general scientific rules for modelling. There have been fierce arguments within the econometrics community in recent decades over various scientific principles for modelling (and associated criteria): whether models should be theory driven or data driven; whether the modelling process should be simple to general or general to specific; and so forth (see Pagan, 1987; Heckman, 2000). Regardless of which principles are followed, the creative element is still very much evident wherever applied econometric modelling occurs, whether such modelling is at the pattern-seeking end of the spectrum or theory-led modelling, and whether the field is macro- or micro-econometrics.

A more recent shift in focus, particularly in the macroeconometric field and associated with Robert Lucas, entails giving up on the aim of using theory to make models that represent the true general structure as a way to uncover that structure. As he wrote:

> A ‘theory’ is not a collection of assertions about the behavior of the actual economy but rather an explicit set of instructions for building a parallel or analogue system – a mechanical, imitation economy. A ‘good’ model, from this point of view, will not be exactly more ‘real’ than a poor one, but will provide better imitations. (Lucas, 1980, p. 697)

This move changes the relation between models and theory, for now the task of theory is to produce models as analogues of the world, rather than to use them to explain the behaviour of the world (see Boumans, 1997). At the same time, it shifts the focus of ‘fitting’: the aim is no longer to fit theory to the world but to fit the model to the world in the particular sense of being able to imitate certain sorts of data characteristics.

Another recent account, developed this time in micro-econometrics by John Sutton (2000), validates itself in relation to the earlier econometric agenda held by Frisch and Tinbergen, for, like those early pioneers, he thinks of models not as devices for the discovery of the true general model as in the Cowles Commission interpretation of Haavelmo's project, nor as mathematical machines that imitate the world as in Lucas's account, but as investigative devices for finding out about the world. In Sutton's view, the economic world produces reasonably stable regularities or variability only within a class of cases, not across all cases; thus, looking for a general model is too ambitious. The aim of modelling is to describe the economic mechanisms that produce the data characteristics that are shared within a subset of all cases and so explain the regularities observed within that subclass. Sutton describes this as a ‘class of models’ approach. Once again, models appear as an intermediary device between theory and data, but this time function to sort out like cases in the world and so offer explanations for their characteristic

behaviour.

Models apparently play a critical epistemological role in econometrics – but there are different ways of characterising this. Econometrics can be seen as fulfilling the function of laboratory experiments in some other sciences – a claim that lies implicit in Haavelmo's discussion of the data of economics as being the result of passive observation of nature's experiments and explicit in his discussion of econometric modelling as designing experiments (see Haavelmo, 1944, chs. 1 and 2). His conceptualization of econometrics appeals to the importance of probability and statistical reasoning as the bases for both model design and statistical inference: models have to be designed to match data that could be observed, and be framed in probability terms. The ‘design of experiments’ notion requires the econometrician to think about the fitting problem, while the probability set-up gives rules for inferences from the model experiment, ones that are in fact much better specified than those for laboratory experiments in most sciences. Thus Haavelmo's blueprint explicitly buys into a tradition of statistical thinking as a valid mode of scientific reasoning, but reinterprets it as a form of experimental work.

A more recent characterisation of the epistemological function of models in econometrics is to understand them as instruments of observation and measurement that enable economists to identify stable phenomena in the world of economic activity. Kevin Hoover's account of ‘econometrics as observation’ describes ‘econometric calculations’ as ‘the economist's telescope’ (1994, p. 74) where rules for focusing the telescope come from statistical theory and where economic theory, and the purpose engaged in, guide the observation process. Marcel Boumans (2005) understands models as the primary instrument in this process, without which the economists could not ‘model the world to number’. Rather than a means of observation, he portrays models as complex scientific instruments that generate the numbers for those economic objects, concepts and relations that cannot be observed directly and that are not yet measured. Like Haavelmo, Boumans's account of model work invokes a careful design of experiments, but he provides a more concrete discussion of how econometric modelling provides measurement structures to deal with *ceteris paribus* clauses; how statistical and other criteria provide ways of assessing the reliability of model instruments (via calibration, filtering and so forth); and how precision and rigour are obtained in the measurement process.

Neither Boumans nor Hoover is instrumentalist about models in the sense that has come to be associated with Milton Friedman's 1953 argument that models need be efficient only for prediction, not for explanation. (Friedman's essay has been much argued over, and interpretations of this particular point vary; see particularly Hirsch and De Marchi, 1990; instrumentalism and operationalism; and Mäki, 2007.) Nor are they operationists in the Bridgmanian sense (that informed, for example, Paul Samuelson's early work in economics; see Bridgeman, 1927), namely, that a concept is defined by its measuring process (such as an econometric model). Both Hoover and Boumans might be termed ‘sophisticated instrumentalists’ for they regard econometric calculations or models as cleverly designed instruments for observing and measuring the relations of economics, and so understanding and explaining, the world.

Modelling as theorizing

The term ‘model’ had rarely been used in economics before the 1930s, even though things we would now label ‘models’ had been developed and used for theorizing before then. We can certainly recognize

some earlier examples of modelling in the late 19th century; for example, we can happily denote the Edgeworth–Bowley box diagram, and Alfred Marshall's trade diagrams and supply–demand scissor diagrams as models. These examples signal that modelling was an unrecognized element in the mathematizing process of that earlier period (see Morgan, 2008). Yet it was only after the 1950s that modelling became a widely recognized way of using mathematics in economics and became one of the dominant forms of economic theorizing. Whereas the establishment of the statistical–econometric notion is associated with Tinbergen, the mathematical–theorizing one may be associated with another Dutch econometrician, Tjalling Koopmans, whose account, given in a set of three essays in 1957, is widely understood as a paradigmatic statement of the modelling approach of modern mathematical economics. Koopmans had developed Tinbergen's earlier ideas about modelling to fit with contemporary discussions of the role of mathematics in economics in the 1940s and 1950s and with the formal mathematical idea of a model at that time. As such, his statement fits into a broader history of mathematics and economics treated particularly in Weintraub (2002) and Ingrao and Israel (1987).

Koopmans defined an economic theory as a set of postulates with which we reason in order to work out and make explicit the otherwise implicit effects of the set of postulates taken together: a reasoning practice that apparently involves models. For Koopmans, this reasoning was an important part of theorizing since these implications are not self-evident, nor is any particular set of postulates necessarily fruitful. His portrayal of ‘Economic Theory as a Sequence of Models’ (to quote his 1957, p. 142, section title) is presented as his answer to the ongoing argument of his day about the status of the assumptions and the predictions of economics in which he explicitly defined the role of models almost as an aside:

> neither are the postulates of economic theory entirely self-evident [as Robbins had argued in 1932], nor are the implications of various sets of postulates readily tested by observation [as Friedman had argued in 1953]. In this situation, it is desirable that we arrange and record our logical deductions in such a manner that any particular conclusion or observationally refutable implication can be traced to the postulates on which it rests … Considerations of this order suggest that we look upon economic theory as a sequence of conceptional *models* that seek to express in simplified form different aspects of an always more complicated reality. At first these aspects are formalized as much as feasible in isolation, then in combinations of increasing realism. (Koopmans, 1957, p. 142)

Koopmans suggests, then, that models are an essential element in theorizing, and that their role comes in their sequenced ability to express different and combined aspects of a simplified reality. But his projection that such a sequence of models would represent ‘combinations of increasing realism’ seems not to have been borne out. While tractability suggests that increasing realisticness in some aspects will have to be traded off against simplification in others, the history of modelling suggests that model sequences are more often driven by changes in problems, in questions, and in the mathematical tools available. This last was a possibility that Koopmans himself discusses in the context of the move from arithmetical to diagrammatic to algebraic forms of theorizing. And, as just noted with Lucas, some modern modelling no longer aims to represent the world as it is, but to develop artificial systems that mimic outputs from the world.

There are various ways of characterizing the use of mathematical models in economic theory. For Daniel

Hausman, the connection of models with concept formation is both more explicit and more important than Koopmans suggests, for economic modelling is where theory development goes on:

> A theory must identify regularities in the world. But science does not proceed primarily by spotting correlations among various known properties of things. An absolutely crucial step is constructing new concepts – new ways of classifying and describing phenomena. Much of scientific theorizing consists of developing and thinking about such new concepts, relating them to other concepts and exploring their implications.
> This kind of endeavor is particularly prominent in economics, where theorists devote a great deal of effort to exploring the implications of perfect rationality, perfect information, and perfect competition. These explorations, which are separate from questions of application and assessment, are, I believe, what economists (but *not* econometricians) call 'models'. (Hausman, 1984, p. 13)

Nowadays, the explorations would be into bounded rationality, imperfect information and imperfect competition: the agenda has moved on, but the mode of theorizing via modelling remains the same.
Hausman's attention to the role of models in conceptual innovation is given credence and depth in his own analysis of Samuelson's 'overlapping generations' model, a story about creative exploration in the theoretical realm. The Edgeworth Box history (see Humphrey, 1996, and Morgan, 2004a) provides another good example of the way modelling is associated with new concepts and descriptions – it is after all where indifference curves, contract curves and so forth were first introduced.
The development of the supply and demand diagram we find in Marshall's *Principles* (1890) exemplifies Hausman's claims. It is not just that Marshall's diagrams describe in new ways some older ideas about the phenomena of supply and demand that go way back in the purely verbal literatures of economics, but that in his hands these curves are fashioned to represent various kinds of markets and relations, resulting in new concepts and classification of types of supply or demand at a level that sits between any general theory and one-off cases (see Morgan, 2002). It is this function of modelling as a classification device that Sutton (2000) reprises in a different form in his 'class of models' work on industrial competition (discussed above). And, historically between these two economists, we can situate, as just one example, the work by Martin Shubik (1959) who used game theoretic models to classify kinds of competition and industry structure according to the kind of game that most matches the economic situation involved.
Hausman is keen to make his account of the methodology of economics not only fit to the practice of modern economics, but philosophically sensible, so he separates the activity of modelling from the more general assertions and truth claims of theories. At first sight this strict separation may look curious to economists who often talk of 'testing models' rather than theories, and do not bother to pull apart the categories of theories and models in their everyday scientific work. This conflation may occur because, as Hausman suggests, 'Models are not themselves empirical applications, but they have the same structure' (Hausman, 1992, p. 80). Having the same structure might enable empirical application by econometricians, though this is not how economists mostly use mathematical models in arguing about the world: rather, they are more often linked to the world in a much more casual fashion.
Indeed, 'casual application' is exactly the term used by Alan Gibbard and Hal Varian to describe how mathematical models are applied 'to explain aspects of the world that can be noticed or conjectured

without explicit techniques of measurement' (1978, p. 672). In their view, mathematical models are designed only to *approximate* the world, and, unlike econometric models which go through a serious process of fitting to the world, they are casually connected to the world by 'stories' which interpret the terms in the model to elements in the world. But they stress that such applications of models do not pertain to particular situations or things in the world. In contrast, Hausman (1990) argues that economists do often use their models in this way to discuss particular real world events, and they use narratives to fill in the descriptions given in the model in order to provide explanations of those events in the world. Morgan (2001; 2007) takes a stronger position with regard to these stories, suggesting that they form an integral part of the application of models to the world – both in general and for particular cases – and equally form an essential part of the identity of the model. Steven Rappaport (1998), like Hausman, finds mathematical models to be quite stretchable in function: in conceptual work, in normative work (for example in discussions of policy), and in heuristic explanatory work. However, in other respects Rappaport's account of models and their function contrast with Hausman's and with Morgan's, for he portrays models as 'mini-theories' within a research programme that function in counterfactual format: that is, their function is to provide accounts of what might happen if the model were a true description of the world.

These accounts of how mathematical models connect to the world all suggest a dependence on cognitive, intuitive or informal elements of economists' theorizing with respect to the world, in strong contrast to the statistical and economic criteria that attend the way econometricians use models to fit theories to the world. On the other hand, mathematical models appear to fulfil a wider variety of functions ranging from devices for new concept formation and classificatory work in theorizing to inference devices that purport to give explanations of general or particular events. Policy usage often involves mathematical models for analysis of policy interventions and for mechanism design purposes – as, for example, in the design of auctions. So far there is little historical or reflective philosophical literature on this side of model work (though see Guala, 2001). By contrast, there is a considerable reflective literature on the policy activities associated with empirical or econometric models (see examples and references in Den Butter and Morgan, 2000).

Models as investigative instruments

We have already seen various ways in which models are understood as investigative devices. In the commentaries on econometrics, we found models portrayed as tools or instruments of observation and measurement, and in the early econometric work models were also understood as tools to help explain the world. The idea of models as instruments is also present in the mathematical modelling literature, but is associated with a more active sense of investigation. Irving Fisher, for his thesis, physically built a three good, three consumer, hydraulic analogue general equilibrium model:

> The mechanism just described is the physical analogue of the ideal economic market. The elements which contribute to the determination of prices are represented each with its appropriate role and open to the scrutiny of the eye. We are thus enabled not only to obtain a clear and analytical *picture* of the interdependence of the many elements in the causation of prices, but also to employ the mechanism as an instrument of investigation

> and by it, study some complicated variations which could scarcely be successfully followed without its aid. (Fisher, 1892, p. 44)

This chimes well with the commentary from Scott Gordon, who, from his historical and philosophical analysis of economics, claims that ‘the purpose of any model is to serve as a tool or instrument of scientific investigation’ (1991, p. 108).

The notion of tools in economics has not been well-developed. Arthur Pigou (1929) introduced the distinction between ‘tool makers’ and ‘tool users’, labelling Francis Edgeworth as a maker of tools, and Marshall as both a maker and user. For Pigou, the term ‘tools’ referred not to processes of induction as opposed to deduction, or even to the mathematical as opposed to the literary method, but to something he referred to as a ‘wider’ analytical movement involving specific statistical and mathematical techniques or ‘machinery’ (such as the method of analysis of demand and supply). It was in following him that Joan Robinson (1933), in oft-quoted comments, wrote about the ‘tool-box of economics’ which she presented as consisting of ‘assumptions’ (theory) and ‘geometry’ (methods) though we might more naturally think of these combining to form models. Koopmans (1957), too, wrote about tools, referring not only to numerical examples and diagrammatic representations, but also to formal mathematics, computing techniques, input–output analysis and so forth, thus (for our time) mixing up methods or modes of analysis (ones we associate now with modelling) and kinds of models. Yet there is a striking similarity between the way Fisher referred to and used his physical hydraulic model and the way modern economists use their equivalent mathematical models of modern economics as tools of investigation. Both seem to be well covered by the notions of tool using that Pigou introduced.

Indeed, attention to the functions of models has emphasized that much of the classifying and conceptual development work of theorizing discussed in the previous section occurs not so much in building mathematical models as in using them. For example, the models developed by Hicks, Samuelson, Meade and others in the late 1930s based on Keynes's *General Theory* were used to explore, develop and understand that theory in ways that involved substantive conceptual and classifying work of their own (see Darity and Young, 1995). In deriving solutions to theoretical problems, or in exploring the limits of behaviour implied by the theoretical relations represented in the models, and in applying their models to think about problems of the economic world represented in the model, those economists used their models as instruments of investigation. These investigations appear as glorified thought experiments, too complicated to do in the mind and so requiring a representation of the case or system in the form of the model and associated mathematical modes of reasoning about it. In Fisher's case, he had a material object to experiment with. Mathematical models in economics also typically provide such internal resources for experimental manipulation. Morgan (2002) argues the case for regarding mathematical modelling activity as experimental work on mathematical models in parallel with statistical experiments practised on econometric models. But whereas we have well-grounded statistical rules for making inferences from econometric experiments, the application of mathematical models to the world (or inferences from such model experiments) is more casual or approximate, as we have already seen.

This notion that mathematical modelling work is a form of experimental activity is most evident in the founding literature on simulation in economics around 1960 (surveyed at the time by Shubik, 1960a; 1960b). In some other fields of science, simulation has been introduced primarily as a method of numerical, rather than analytical, solution. But in economics, simulation has been more usually

presented and used as a process of experiment on models, a process that effectively investigates in a systematic manner the full range of behaviours of the system or the actors portrayed in the model. There were isolated examples of simulation earlier in the history of economics – most particularly Tinbergen's 1936 simulation of his macroeconometric model, Paul Samuelson's (1939) simulation of a little Keynesian mathematical system and Eugen Slutsky's (1927) famous random shock models that mimicked business cycles. The possibilities of simulation were then explored more effectively during 1950s and 1960s Cold War activities that brought the social sciences and mathematics together.
The birth of simulation in economics has usually been attributed to Herbert Simon, but equally important were concurrent developments connected with other pioneers, particularly Frank and Irma Adelman, Martin Shubik and Guy Orcutt (see Morgan, 2004b). Simon's simulation projects in economics involved, for example, programming computers to imitate decisions and choices in the same way that investment bankers made those decisions and choices, that is, on the same information and by the same processes of comparison and assessment (see Clarkson and Simon, 1960). The Adelmans's work was particularly important in the development of simulation methods in econometrics following the lead of Tinbergen's earlier work (see also Duesenberry, Eckstein and Fromm, 1960), while in economics at that time simulations involved both ‘game playing’, meaning experiments in which people role-played making economic decisions where the model simulated the environment and all the interest was in the behaviour of the people (for example, managers making decisions), and mathematical model simulations in which the behaviour was taken as given (for example, rational economic behaviour) and the environment varied to see how that altered the outcomes projected by the model. (This broad category of simulations around 1960 thus included some things we would now label experiments.) Shubik was involved in many of these different types of simulations ranging from game-playing experiments, to business games, to model experiments. Orcutt (1960) meanwhile pioneered the method of microsimulation, in which he constructed a representative virtual sample of the population, endowed the sample individuals with characteristics of the real population, and then simulated their behaviour through time to explore the characteristics of the aggregate system as well as the individual parts. This is complicated model-experimental work that was possible only with the new-found computing power of that day. All these economists significantly extended the ways in which models worked as instruments of investigation via different forms of experimental activity in which each ‘run’ of the model provided a slightly different experiment with the model. Simulation, since its introduction into economics, has been characterized as a form of experiment with models that aims at mimicking a variety of different economic behaviours, at different levels and in different ways.

Model construction

Model making (as opposed to formal or informal definitions of models) has been a fertile ground for philosophical commentators on economics who have presented it as a process of ‘idealization’, a term that covers a range of things including abstraction, simplification and isolation (see Hamminga and De Marchi, 1994). This general idea goes back to the ‘ideal type’ concept defined by Max Weber (1904; 1913) for the social sciences. His discussion included notions of the ideal type of individual economic behaviour and the ideal type notion of a market. Certainly it is easy to see the late 19th century portrait of economic man as ideal type, divorced from all but his pure economic motivations without any deeper

psychology. The term ‘idealization’ suggests that models are arrived at by processes of *abstracting* to the level of ideas or concepts; of *simplifying* the case or system treated by omitting irrelevant or negligible influences; of *isolating* the elements that are really thought to be important by *ceteris paribus* clauses; and so forth (see Morgan, 2006). These processes can be understood as working on theories (for example, moving from a full equilibrium account down to a single particular market) or as starting with the complicated world and isolating a small part of it for model representation. Leszek Nowak (for example, 1994) presents a rather general analysis in which ‘idealization’ takes one from the world to theory and ‘concretization’ from theory to the world in two rather seamless parallel processes. This account known as the ‘Pozna• approach’ (named after the University that hosted its development: see Hamminga, 1998), was formulated for Marxian economics, but might well be applied more generally.

Two other commentators particularly associated with questions of idealization in economic modelling are Nancy Cartwright and Uskali Mäki. Cartwright (1989) is interested in what has been called ‘causal’ idealization, that is, in isolating the causal capacities that actually work in the world. She associates this aim both with how econometric modelling works and with Millian tendencies (the account of tendency laws in economics provided by John Stuart Mill in the mid-19th century). Mäki (1992) is more interested in ‘construct’ idealization, that is, in how economic theorizing goes on by constructing versions of theory with more or less scope along different dimensions of isolation. (The distinction between construct and causal idealization used here is due to McMullin, 1985.) We can find both these kinds of process going on in the history of model making. Von Thünen's (1826) construction of his diagrammatic model of an ‘isolated state’ provides a clear example of model-making by isolating the factors that determine farm profitability. His isolations can be interpreted as creating a theoretical model (that is, he constructed an idealized model) but he was also interested in getting at real causes for he fitted this model to his own farm's statistical data (that is, he isolated the causes, using informal econometric procedures).

Idealization itself may involve not just simplifications or isolations but the addition of false elements. Max Weber (1904) discusses how ideal types present certain features in an exaggerated form, not just by accentuating those features left by the omission of others but as a strategy to present the most ideal form of the type. This notion of exaggeration comes up again in Gibbard and Varian's (1978) notion of caricature modelling in economics, where the exaggeration is designed to enable the economist to investigate the robustness of the model (the virtue that Friedman had, of course, earlier associated with the use of unrealistic assumptions). But if we interpret this caricaturing process to involve not just an extreme degree of exaggeration but the addition of features, then we have an idealization of a qualitatively different kind from those that come from methods of isolation or simplification. For example, Frank Knight's (1921) assumption of perfect information involves adding a feature to the portrait of economic man; the assumption can be specified in different ways, each creates a different model. Caricature models are not to be confused with the artificial constructions of Lucas's models, which are not derived by idealization from either theory or the world. Idealizations, even in caricaturing form, are still understood as representations of the system or man's behaviour (however unrealistic or positively false these might be) whereas the artificial world models do not seek to represent the system or agent's behaviour – rather, the aim is to mimic the output of such systems or behaviour. In imitating the system outputs, one might of course argue that representational power is sought at a different point.

In economics itself, as opposed to in the analyses of commentators, these processes of model making

may all be going on together at the same time. That is, models may be constructed to represent the idealized versions of grander theories, be abstracted from the particularities of economic life, and provide simplifications of the more complicated world. These features are all at play in François Quesnay's famous 18th century *Tableau économique*, a construction that may be regarded as the general ancestor of models in economics. But that model makes a telling example, for as a construction it is only in part a derivation or isolation from a general set of ideas or theory, only in part a simplification of the relations in the world or abstraction into a more conceptual framework. It does not seem to be derived entirely from theory, nor does it appear as a description of his contemporary data. Yet while it does embody elements of all these things, it is also a construction of its own (see Charles, 2004). Quesnay moulded these elements together to create a wonderful table-cum-picture that represents the French economy of his day, one that few later economists can understand easily (at least without translating it into a different form, which of course changes its meaning and working).

This interpretation of Quesnay's modelling assumes that models are neither just derived from theory nor solely built up from data, for they typically involve bits of both and oftentimes other things as well, such as metaphors, imported mathematical forms, and so on. The notion that econometric models are constructed from both theoretical relations and statistical elements is probably not that contentious. The mixture of elements is also obvious in a case like the Phillips–Newlyn model, a real hydraulic machine in which red water, representing the various aggregate stocks and flows of the economy, circulated around the machine and sometime spilt into the lecture room (see Leeson, 2000; Boumans and Morgan, 2004). But these mixtures are equally characteristic in mathematical models, according to the case work account of model building by Boumans (1999), who argues that we should think of model making as like cooking new recipes, in which mathematics provides the means of integrating such several, sometimes disparate, elements into new models. This account of model construction goes against much traditional philosophizing, even by economists, about model making. Yet more recently economists have begun to write about their modelling work as a much more ad hoc activity in which past practices, new intuitions and even speculations guide their model making (see, for example, Krugman, 1993; Sugden, 2000).

Understanding model making according to Boumans's recipe-making account suggests that models – by construction – are partially independent of both theory and the world (or its data), and this accounts for their apparently autonomous existence as working objects in modern economics. This construction account is part of the ‘models as mediators’ view of the role of models, which analyses their use as investigative instruments (see Morrison and Morgan, 1999). According to this account, models can function in this autonomous in-between way because of their construction. However, the possibility of learning from using models depends on another element in their construction, namely, that models are devices made to represent in some way or form something in our economic theories or in the economic world or both at once. It is this representing quality, built in at the construction stage, which makes it possible to use a model not just as an instrument of prediction but as an investigative instrument to learn something about the world or the theory which it represents. This account can apply even to the artificial world models proposed by Lucas which are constructed not to represent the workings of the system but the outputs of the system, though here the modellers’ ambitions to learn from the modelling in order to understand the economic system and explain the outcome phenomena that they mimic seems somewhat reduced.

This recent recipe account of model-making stands in marked contrast to accounts of how model making goes on according to those mid-20th century commentators discussed earlier. Recall that Koopmans had labelled mathematical models as ‘defined by a set of postulates’ where the full set of postulates form the theory – a definition consistent with the then current axiomatic approach to theories. In econometrics, the Cowles Commission presented econometric models as being derived – directly given in some sense – from a priori theory. Indeed, it was the basis of their position in the ‘measurement without theory’ debate that econometrics needed models that were clearly versions of theories to get anywhere at all, against the data-derived models of the National Bureau of Economic Research (NBER) that they decried as unscientific. Another description that fits the philosophical inclinations of the mid-20th century, but is more model-oriented, was given by Friedman, who defined a theory as consisting of two parts: ‘a conceptual world or abstract model simpler than the “real world” and containing only the forces that the hypothesis [theory] asserts to be important’ and a second part defining the ‘class of phenomena for which the “model” can be taken to be an adequate representation of the “real world”’ along with the correspondence rules linking the model terms and the phenomena (Friedman, 1953, p. 24). Friedman here neatly depicts the model as both a version of theory and at the same time a representation of the real world, yet the correspondence rules are by no means unproblematic. While one could argue that the main work of econometrics has been to develop both the theory and practices of such correspondence rules for models, for mathematical models, in contrast, methodological accounts have often foundered on how such correspondence criteria might be formulated. Despite the long shadow of these rather formal mid-20th century definitions, it is in keeping with our observations about how models are used in modern economic science that they may now be understood as autonomous working objects, rather than as either proto-theories or versions of data.

Conclusion

There is more that might be said, and that remains to be researched, about the philosophy of modelling, for example about the nature of reasoning with mathematical models; about the role of mathematical models within the design of classroom/laboratory experiments in economics; about the use of models in policy advice and intervention; and about the absence of formal criteria for working with mathematical models that are equivalent to the statistical criteria associated with econometric model work. There is also much to be done in filling in the skeletal history of modelling offered here: in separating the history of modelling from both the history of mathematical economics and the history of econometrics; in demarcating the historical range of scope of modelling; and in discerning why and how the method took hold. Nevertheless, the basic trajectory of the history is clear: modelling becoming defined as a mode of reasoning and working for economics in the 1930s, it was developed and used in various ways in the 1940s and 1950s, setting the scene for modelling to become a dominant methodology in the latter part of the century. And once defined, we can look back and recognize earlier prototypes for such a method going back to Quesnay in the 18th century. When we so look back, and consider the scientific world view that we have lost in economics by adopting modelling as one of our favoured methods of doing economics, what stands out is that the science is a radically different one. No longer do economists believe and enquire into a few grand governing laws, nor even propose wide-ranging general theories – rather, economics has become a science of many different and particular models.

See Also

- econometrics
- Edgeworth, Francis Ysidro
- Fisher, Irving
- instrumentalism and operationalism
- Koopmans, Tjalling Charles
- mathematics and economics
- methodology of economics
- Tinbergen, Jan

Bibliography

Adelman, I. and Adelman, F.L. 1959. The dynamic properties of the Klein–Goldberger model. *Econometrica* 27, 596–625.

Backhouse, R.E. 1998. The transformation of U.S. economics, 1920–1960. In *From Interwar Pluralism to Postwar Neoclassicism*, ed. M.S. Morgan and M. Rutherford. Annual Supplement to *History of Political Economy*, vol. 30, Durham, NC: Duke University Press.

Bodkin, R.G., Klein, L.R. and Marwah, K. 1991. *A History of Macroeconometric Model-Building*. Aldershot: Elgar.

Boltzmann, L. 1911. Models. In *Encyclopaedia Britannica*, 11th edn. Cambridge: Cambridge University Press.

Boumans, M. 1993. Paul Ehrenfest and Jan Tinbergen: a case of limited physics transfer. In *Non-natural Social Science: Reflecting on the Enterprise of More Heat Than Light*, ed. N. De Marchi. Durham: Duke University Press.

Boumans, M. 1997. Lucas and artificial worlds. In *New Economics and its History*, ed. J.B. Davis. Durham: Duke University Press.

Boumans, M. 1999. Built-in justification. In *Models as Mediators*, ed. M.S. Morgan and M. Morrison. Cambridge: Cambridge University Press.

Boumans, M. 2005. *How Economists Model the World to Numbers*. London: Routledge.

Boumans, M. and Morgan, M.S. 2001. *Ceteris paribus* conditions: materiality and the application of economic theories. *Journal of Economic Methodology* 8, 11–26.

Boumans, M. and Morgan, M.S. 2004. Secrets hidden by two-dimensionality: the economy as a hydraulic machine. In *Models: The Third Dimension of Science*, ed. S. de Chadarevian and N. Hopwood. Stanford: Stanford University Press.

Bridgeman, P. 1927. *The Logic of Modern Physics*. New York: Macmillan.

Cartwright, N. 1989. *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

Charles, L. 2004. The Tableau économique as rational recreation. *History of Political Economy* 36, 445–74.

Clarkson, G.P.E. and Simon, H.A. 1960. Simulation of individual and group behaviour. *American Economic Review* 50, 920–32.

Darity, W. and Young, W. 1995. IS–LM: an inquest. *History of Political Economy* 27, 1–41.

Den Butter, F. and Morgan, M.S. 2000. *Empirical Models and Policy Making: Interaction and Institutions*. London: Routledge.

Duesenberry, J.S., Eckstein, O. and Fromm, G. 1960. A simulation of the United States economy in recession. *Econometrica* 28, 749–809.

Epstein, R.J. 1987. *A History of Econometrics*. Amsterdam: North-Holland.

Fisher, I. 1892. *Mathematical Investigations in the Theory of Value and Prices*. Yale University thesis, repr. New Haven: Yale University Press, 1925.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Frisch, R. 1933. Propagation and impulse problems in dynamic economics. In *Economic Essays in Honour of Gustav Cassel*. London: Allen & Unwin.

Gibbard, A. and Varian, H.R. 1978. Economic models. *Journal of Philosophy* 75, 664–77.

Gordon, S. 1991. *The History and Philosophy of Social Science*. New York: Routledge.

Guala, F. 2001. Building economic machines: the FCC auctions. *Studies in History and Philosophy of Science* 32, 453–77.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement), iii–iv, 1–115.

Hamminga, B. 1998. Pozna• approach. In *Handbook of Economic Methodology*, ed. J.B. Davis, D. Wade Hands and U. Mäki. Cheltenham: Edward Elgar.

Hamminga, B. and De Marchi, N., eds. 1994. *Idealization in Economics*. Amsterdam: Rodopi.

Hausman, D.M., ed. 1984. *The Philosophy of Economics: An Anthology*. Cambridge: Cambridge University Press.

Hausman, D.M. 1990. Supply and demand explanations and their *ceteris paribus* clauses. *Review of Political Economy* 2, 168–87.

Hausman, D.M. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Heckman, J. 2000. Causal parameters and policy analysis in economics: a twentieth century retrospective. *Quarterly Journal of Economics* 115, 45–97.

Hirsch, A. and De Marchi, N. 1990. *Milton Friedman: Economics in Theory and Practice*. New York: Harvester Wheatsheaf.

Hoover, K.D. 1994. Econometrics as observation: the Lucas critique and the nature of econometric inference. *Journal of Economic Methodology* 1, 65–80.

Humphrey, T.M. 1996. The early history of the box diagram. *Federal Reserve Board of Richmond Economic Review* 82(1), 37–75.

Ingrao, B. and Israel, G. 1987. *The Invisible Hand: Economic Equilibrium in the History of Science*. Trans. I. McGilvray, Cambridge, MA: MIT Press, 1990.

Knight, F.H. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

Koopmans, T. 1957. *Three Essays on the State of Economic Science*. New York: McGraw Hill.

Krugman, P. 1993. How I work. *American Economist* 37(2), 25–31.

Le Gall, P. 2007. *A History of Econometrics in France: From Nature to Models*. Routledge: London.

Leeson, R. 2000. *A.W.H. Phillips: Collected Works in Contemporary Perspective*. Cambridge: Cambridge University Press.

Lucas, R.E. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12, 696–715.

Mäki, U. 1992. On the method of isolation ineconomics. In *Idealization IV: Intelligibility in Science*, ed. C. Dilworth. Amsterdam: Rodopi.

Mäki, U., ed. 2002. *Fact and Fiction in Economics*. Cambridge: Cambridge University Press.

Mäki, M., ed. 2007. *The Methodology of Positive Economics: Milton Friedman's Essay Fifty Years Later*. Cambridge: Cambridge University Press.

Marshall, A.W. 1890. *Principles of Economics*. 8th edn, London: Macmillan, 1930.

McMullin, E. 1985. Galilean idealization. *Studies in History and Philosophy of Science* 16, 247–73.

Morgan, M.S. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Morgan, M.S. 2001. Models, stories and the economic world. *Journal of Economic Methodology* 8, 361–84 Repr. in Mäki (2002).

Morgan, M.S. 2002. Model experiments and models in experiments. In *Model-Based Reasoning: Science, Technology, Values*, ed. L. Magnani and N. Nersessian. New York: Kluwer Academic/Plenum Press.

Morgan, M.S. 2004a. Imagination and imaging in economic model-building. *Philosophy of Science* 71, 753–66.

Morgan, M.S. 2004b. Simulation: the birth of a technology to create ‘evidence’ in economics. *Revue d'Histoire des Sciences* 57, 341–77.

Morgan, M.S. 2006. Economic man as model man: ideal types, idealization and caricatures. *Journal of the History of Economic Thought* 28, 1–27.

Morgan, M.S. 2007. The curious case of the Prisoner's Dilemma: model situation? Exemplary narrative? In *Science Without Laws: Model Systems, Cases, and Exemplary Narratives*, ed. A. Creager, E. Lunbeck and N. Wise. Durham, NC: Duke University Press.

Morgan, M.S. 2008. *The World in the Model*.

Morgan, M.S. and Morrison, M. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Morrison, M. and Morgan, M.S. 1999. Models as mediating instruments. In Morgan and Morrison (1999).

Nowak, L. 1994. The idealization methodology and econometrics. In Hamminga and De Marchi (1994).

Orcutt, G.H. 1960. Simulation of economic systems. *American Economic Review* 50, 893–907.

Pigou, A.C. 1929. The function of economic analysis. The Sidney Ball Lecture, University of Oxford, May. In *Economic Essays and Addresses* with D.H. Robertson. London: P.S. King, 1931.

Qin, D. 1993. *The Formation of Econometrics*. Oxford: Clarendon Press.

Rappaport, S. 1998. *Models and Reality in Economics*. Cheltenham: Edward Elgar.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics* 21, 75–8.

Shubik, M. 1959. *Strategy and Market Structure*. New York: Wiley.

Shubik, M. 1960a. Bibliography on simulation, gaming, artificial intelligence and allied topics. *Journal of the American Statistical Association* 55, 736–51.

Shubik, M. 1960b. Simulation of the industry and the firm. *American Economic Review* 50, 908–19.

Slutsky, E.E. 1927. The summation of random causes as the source of cycle processes. *Econometrica* 5 (1937), 105–46.

Sugden, R. 2000. Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology* 7, 1–31.

Sutton, J. 2000. *Marshall's Tendencies: What Can Economists Know*? Cambridge, MA: MIT Press.

Tinbergen, J. 1937. *An Econometric Approach to Business Cycle Problems*. Paris: Hermann.

Tinbergen, J. 1939. *Statistical Testing of Business Cycle Theories*. Geneva: League of Nations.

Von Thünen, J.H. 1826. *Der Isolierte Staat*. Hamburg: Perthes. Trans. C. Wartenberg as *Von Thünen's Isolated State*, Oxford: Pergamon, 1966.

Weber, M. 1904. ‘Objectivity’ in social science and social policy. In *The Methodology of the Social Sciences*, trans. and ed. E.A. Shils and H.A. Finch. New York: Free Press, 1949.

Weber, M. 1913. *The Theory of Social and Economic Organisations*, trans. A.M. Henderson and T. Parsons as Part I of *Wirtschaft und Gesellschaft*. New York: Free Press, 1947.

Weintraub, E.R. 2002. *How Economics Became a Mathematical Science*. Durham, NC: Duke University Press.

How to cite this article

Morgan, Mary S. "models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000391> doi:10.1057/9780230226203.1112

The New Palgrave Dictionary of Economics Online

Modigliani, Franco (1918 – 2003)

Richard Sutch

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article focuses on the scholarly contributions of Franco Modigliani, 1985 Nobel laureate in economics. Particular attention is given to his formulation of the determinants of equilibrium in Keynesian macroeconomics, the life-cycle hypothesis of saving, his contributions to the theory of expectations, and the Modigliani–Miller theorems of corporate finance. The objective is to demonstrate Modigliani's importance in the history of economics.

Keywords

aging and retirement; American Economic Association; American Finance Association; Ando, A.; Arbitrage; Barro, R.; bequest motive; bonds; budget deficits; buffer stocks; business cycle; classical economists; consumption function; corporate finance; cost of capital; dividend policy; Duesenberry–Modigliani hypothesis; Econometric Society; expectations; Federal Reserve System; finance theory; fiscal policy; FMP model; Friedman, M.; Great Depression; Greenspan, A.; Hicks, J.; life cycle hypothesis; life-cycle theory of saving; infinite horizons; inflation; information cost; inventories; investment function; IS–LM model; Keynes, J. M.; Keynesian revolution; Koopmans, T. C.; Lange, O. R.; Lerner, A. P.; life-cycle hypothesis; liquidity trap; long-term contracts; marginal utility analysis; Marschak, J.; microfoundations; Miller, M.; MIT model; Modigliani, Franco; Modigliani–Miller theorem; monetarism; monetary policy; money demand; money income vs real income; money supply; Muth, J. F.; national debt; neoclassical synthesis; new classical macroeconomics; permanent income hypothesis; precautionary savings; preferred maturity habitat; price control; quantity theory of money; rational expectations; real money; relative income hypothesis; Ricardian equivalence theorem; rules vs discretion in policy; saving and growth; saving–income ratio; Shiller, R.; simultaneous equations; Social Security (USA); stabilization policy; sticky wages; stock price volatility; Sutch, R.; term structure of interest rates; uncertainty; underemployment equilibrium

Article

Franco Modigliani was awarded the Sveriges Riksbank (Bank of Sweden) Prize in Economic Sciences in Memory of Alfred Nobel in 1985 for ‘pioneering studies of saving and of financial markets’. A life-long Keynesian, his contributions to macroeconomics and finance transformed both fields. The life-cycle approach to consumption and saving pioneered a microfoundations approach to macroeconomic theory and remains the standard model of consumption in macroeconomics. The Modigliani–Miller theorems on the cost of capital had a profound influence on subsequent research in finance. He was also a pioneer in modelling expectations in macroeconomic models. Modigliani was an influential and critical voice on macroeconomic policy in the United States, in his native country of Italy, and in the European community.

Biography and intellectual development

Modigliani was born in Rome, Italy on 10 June 1918. His father, who died when Modigliani was only 14, was a pediatrician. He entered the University of Rome to study law at 17. In his second year he won a national competition in economics with an essay on the price controls imposed in Italy during the annexation of Abyssinia (now Ethiopia). He records in his autobiography (2001) that, following the receipt of this award, he began a self-study of economics reading the classics, an approach he deemed more satisfactory then taking courses during the fascist regime.

At about the same time he became a committed anti-fascist. After the Italian government promulgated anti-Semitic laws in 1938, he and his fiancée, Serena Calabi, fled to Paris, where they were married in 1939. He and Serena applied for an immigration visa to the United States and arrived in New York in August 1939, a few days before the beginning of the Second World War. Modigliani was immediately taken on as a postgraduate scholar by the New School for Social Research, which had been newly created as a haven for social scientists fleeing Europe. He was mentored there in economic theory and econometrics by Jacob Marschak. Modigliani always took care to acknowledge the powerful influence that Marschak had on his development as an economist. During 1941–3 Modigliani taught as an instructor at the New Jersey College for Women (now Douglass College and at the time part of Rutgers University) and at Bard College of Columbia University (now independent). During these years he continued to work on his doctoral dissertation in social sciences for the New School, and received a Ph. D. in 1944. This work was reported in the same year in his first published article, ‘Liquidity Preference and the Theory of Interest and Money’. He then returned to the New School as a lecturer.

Modigliani taught briefly at the University of Illinois (1949–52), where he was promoted from associate professor to full professor in 1950 at the age of 32. There he found a friend and collaborator, Richard Brumberg, a graduate student. Together they developed the life-cycle theory of saving, which became Modigliani's most important contribution and one of the two cited by the Nobel judges. He next taught at the Carnegie Institute of Technology (now Carnegie–Mellon) as a Professor of Economics and Industrial Administration (1952–60). Most of the other projects now associated with his name were begun there including his collaboration with Merton Miller on the founding theorems of corporate finance. These theorems were the second contribution cited in 1985 by the Nobel committee.

Modigliani visited Harvard University (1957–8) and the Massachusetts Institute of Technology (1960–1). He was appointed to the faculty at Northwestern University (1960–2) and taught there one year

before returning to MIT in 1962 as a Professor of Economics and Finance. He remained at MIT for the balance of his career. By the mid-1960s MIT was regarded as the premier graduate school in the world for the study of economics.

Modigliani's scientific output is impressive for its breadth of coverage, the depth to which each topic was pursued, and the sheer volume of brilliant, highly original papers. In six volumes of collected papers (1980–2005), Modigliani assembled 87 published papers from a corpus of nearly 200 (and a famous previously unpublished paper with Richard Brumberg, 1980). Modigliani also wrote or coauthored ten books and edited several more. This huge output is all the more remarkable when one considers that throughout his academic career Modigliani always subjected his economic theory to rigorous empirical verification, often employing sophisticated statistical technique with ingeniously (and laboriously) derived data.

In 1970 Modigliani was named an Institute Professor, an honorific title that MIT reserves for scholars of great distinction. He was elected President of the American Economic Association (1975–6). He also served as President of the Econometric Society and the American Finance Association. He became Professor Emeritus in 1988.

Franco Modigliani died on 25 September 2003 at the age of 85 in Cambridge, Massachusetts. MIT Institute Professor Paul Samuelson, a colleague and friend, said, ‘Franco Modigliani could have been a multiple Nobel winner. When he died he was the greatest living macroeconomist. He revised Keynesian economics from its Model-T, Neanderthal, Great Depression model to its modern-day form’ (MIT, 2003).

The Keynesian revolution and the debate over stabilization policy

When he arrived in New York in 1939, Modigliani began several years of study of macroeconomics (and mathematics and statistics as well) under the tutelage of Jacob Marschak, Abba Lerner, Oskar Lange and Tjalling Koopmans. The hot topic, of course, was *The General Theory of Employment, Interest, and Money* by John Maynard Keynes. Published in 1936, Keynes's analysis was truly revolutionary. Keynes pioneered modern macroeconomics by proposing a novel and compelling explanation of the gyrations of the economic system. Those fluctuations had had a devastating impact on the US economy during the Great Depression of the 1930s, a catastrophe whose lingering effects were still evident in 1939. The Keynesian model also suggested a set of active policy prescriptions that might be used, first, to lift an economy out of depression and, second, to prevent recessions and depressions from occurring in the first place. Furthermore, Keynes suggested that if the curative policies were not applied, the economy might languish with mass unemployment for a long time.

Neither the theoretical formulation nor the policy prescriptions of *The General Theory* were easy to accept in the early 1940s. Keynes's argument was complicated and subtle, the book's prose was at points cumbersome and inelegant, and the concepts that Keynes introduced were unfamiliar to economists and sometimes counter-intuitive. The policy implications seemed almost impossibly unorthodox. Government spending should not be based on the need for public services. Taxes should not be based on the need for revenue to pay for the government services. Instead government spending and taxation should be directed to restoring and then maintaining full employment which might lead to levels of spending far in excess of the perceived need for public services and to a level of taxation that might

produce substantial deficits.

Working in 1942 and 1943, Modigliani sought to reduce the confusion generated by the debate over what Keynes was saying and to articulate the common sense of the Keynesian policy message. In the process he made an important clarification of the Keynesian argument. The result was his now famous *Econometrica* paper of 1944, ‘Liquidity Preference and the Theory of Interest and Money’. The paper did three things. First, Modigliani reduced the 384 pages of Keynes's complex argument to a mathematical system of nine simultaneous equations. The virtue of a mathematical representation is that it served to insure that the variables considered important by Keynes were consistently and precisely defined and that the relationships among them were made rigorously explicit. Modigliani was not the first to attempt a mathematical reduction to clarify the logical structure of *The General Theory*. One of his mentors, Oskar Lange, the noted Polish economist, had preceded him. But Modigliani's version became the standard, taught to graduate students for decades (who generally left *The General Theory* unread), until it was replaced by a revised (and more complex) presentation produced by Modigliani in 1963, ‘The Monetary Mechanism and its Interaction with Real Phenomena.’ Second, Modigliani clarified the role played in the model by Keynes's assumption that money wages were inflexible. We return to this point below. Third, Modigliani argued that fiscal policy was not the only weapon available for fighting recessions. Monetary policy could be effective in many, if not all cases. In this third effort, Modigliani was taking issue with another of his mentors, Abba Lerner, who was suggesting at that time that fiscal policy, and only fiscal policy, would work.

The mathematical formulation of the determinants of macroeconomic equilibrium did much to make Keynes acceptable to economists, though it must be said that the mathematics required was most easily mastered by young economists still in graduate school or only recently accepted into the professorship. Many of the ‘old guard’ seemed unable or unwilling to shed their pre-Keynesian conceptions. By salvaging and later defending a role for monetary policy, Modigliani had a major influence on the conduct of anti-recession policy, particularly in the 1950s and 1960s. But it was the clarification of the role of ‘sticky wages’ that helped to transform Keynesian economics into its modern form.

Modigliani established that both the classics and Keynes shared a conception of the macroeconomic demand for money derived from basic microeconomic principles. Any such model would necessarily connect money to real variables such as output and employment only when money entered the formulation as a ratio to the price level. This ratio is known as *real* money and was defined by Keynes and Modigliani in terms of ‘wage units’. The ‘classical’ quantity theory of money, for example, made real money proportional to real output. If changes in the nominal money supply are to influence real output, there must be some reason why those changes are not immediately followed by an equi-proportionate change in wages. In the pre-Keynesian, ‘classical’ model wages would adjust rapidly, unemployment would thus be briefly transitory, and monetary policy would be both ineffective and unnecessary to increase output and employment.

Modigliani's equations revealed that idle resources and price flexibility could simultaneously exist only in the extreme case when the demand for money became infinite. Modigliani considered this an unlikely situation which he called the ‘Keynesian case’. It later became better known as the ‘liquidity trap’. In the *General Theory* Keynes had been critical of the flexible wage assumption and introduced what Modigliani considered the more realistic assumption that, in the short run at least, money wages would not adjust in the downward direction. With this specification added to the system of equations, underemployment equilibrium was possible even when liquidity trap conditions were not present.

Modigliani argued on this basis that the hypothesis of wage rigidity was a necessary part of the Keynesian system if monetary policy was to play a role in influencing real variables.

As a corollary of the argument, Modigliani pointed out that the economy could not be ‘dichotomized’ into real and monetary sectors that operated independently of each other. In his demonstration Modigliani was following John Hicks who made the same point with the IS–LM apparatus made famous by introductory textbooks. In Hicks's (1937) diagram the interest rate and the level of output are jointly determined by the intersection of an LM curve reflecting an equilibrium of the demands and supplies that characterize the monetary sector (L for ‘liquidity preference’ and M for the money supply) with the IS curve reflecting the equilibrium of real forces (I for the demand for investment and S for the supply of saving). It might be noted, however, that Hicks expressed the IS–LM relationship in terms of the rate of interest and *money* income. It was Modigliani who gave it the appropriate interpretation in terms of the interest rate and *real* income.

Modigliani considered the 1944 paper one of his most significant contributions. It set the stage for the ‘neoclassical synthesis’ of the Keynesian and the classical traditions. This synthesis came to dominate the economics profession for the next three or four decades. That approach accepted that labour and capital would be underutilized over the course of the business cycle, that unemployment was not a transitory problem but a variable that helped clear the money market, and that activist monetary and fiscal policies can be welfare improving. Indeed, avoiding unemployment would take close management of the money supply and interest rates. In the United States these views became most influential during the 1960s when the administration of John F. Kennedy put them into practice in a serious way. But these academic and political developments pulled Modigliani into an extended debate with the ‘monetarists’. Led by Milton Friedman, the monetarists held that the quantity of money is the key factor in determining economic change and that the fiscal variables advocated by Modigliani and other Keynesians are not important. Modigliani was particularly disturbed by Friedman's proposal that neither discretionary fiscal nor monetary policy should be employed; rather, the money supply should be strictly regulated to grow at a constant rate (say three per cent per year). Modigliani ridiculed this prescription as a ‘blind rule’ and consistently argued that wise discretionary control of the money supply was essential.

In a pair of empirical papers, one with Albert Ando, his student at Carnegie, Modigliani, went on the attack (1964; 1965). In his presidential address to the American Economic Association, Modigliani rejected the idea that Keynesians did not think that money mattered, and he cited his 1944 and 1963 papers as proof (1977). After winning the rhetorical and empirical debate with Friedman, he sought to ‘make peace’ with the monetarists by declaring ‘We are all monetarists’. And yet he went on to defend the case for policy discretion in a fashion he later described as ‘a full, passionate, and polemical’. In an interview conducted in 1999, he declared victory. ‘There is not a country in the world today that uses a mechanical rule’ (2000, p. 236). He might have added that the highly praised success of Alan Greenspan as the Chairman of the U.S. Federal Reserve Board was based on the careful discretionary management of money and interest rates. In a series of lectures, later published as *The Debate over Stabilization Policy* (1986c), Modigliani traced the history of these disputes.

The monetarist debates of the 1960s, and the empirical success of the work testing Keynesian propositions, led to another important direction for Modigliani's research. He was asked to construct an econometric model of the US economy by the Federal Reserve. The model would be an empirically estimated system of simultaneous equations that would be used by the Federal Reserve to make and guide policy and forecast future developments. He asked Albert Ando to join him on the project and

they created what was first known as the ‘MIT model’ and, after Ando moved to the University of Pennsylvania, as the ‘Federal Reserve-MIT-University of Pennsylvania Model’ (FMP) (1975a). The result embodied many of Modigliani's ideas about the structure of the economy, the consumption function, the structure of interest rates, and the workings of other financial markets. Emblematic of Modigliani's willingness to learn from the data were the many modifications he made to his early formulations in the process of constructing the FMP model. In particular he explicitly extended the theories to include the causes and consequences of inflation which had only begun to become a noticeable problem for the American economy in the 1970s (for his major contributions on inflation see Part III of *Collected Papers*, vol. 5). The model proved sufficiently valuable that the Federal Reserve continued to use it into the 1980s.

The life-cycle model of saving and consumption

In his 1944 paper on the Keynesian model, Modigliani presented an equation for the national flow of saving that described saving as a positive function of aggregate income in a manner consistent with the ‘consumption function’ famously introduced by John Maynard Keynes in the *General Theory*. Keynes had postulated a ‘fundamental psychological law’ whereby an individual's consumption would increase as his or her income increased but not as much as the increase in income. Thus saving, defined as income less consumption, should increase when income grows and the aggregate saving rate, defined as the national saving–income ratio, should increase with aggregate income. According to this part of the *General Theory*, rich people saved, poor people did not; rich countries saved, poor countries did not. Despite his acceptance of this simple Keynesian formulation in 1944, Modigliani reports that he was not convinced that the saving–income ratio should rise with aggregate income, and began to systematically reconsider the Keynesian law in 1946. He was particularly unhappy with the notion that saving should be regarded as a luxury good that would be ‘purchased’ in greater quantities by the rich than poor in order to ‘bequeath a fortune’. ‘This explanation satisfied me not a jot’ (2001, p. 52).

In the late 1940s Modigliani's alternative suggestion was that the saving–income ratio should fluctuate around a constant (or slowly moving) trend and that these fluctuations would be driven by the relationship of actual income to the normal income that the household could expect. In other words, the household's saving rate was explained not by its absolute level of income (as Keynes would have it) but by its income *relative* to the aggregate mean income in the economy. Modigliani formulated his hypothesis in an elegant linear model in which the saving–income ratio was related negatively to the ratio of income at its previous peak to the current level of income (1949). When the economy was in recession (and current income was below its previous peak), saving and the saving–income ratio would both fall. This movement reflected the cyclical movement of consumption emphasized by Keynes. But, when the economy was growing and incomes were pushed above their previous peak, saving would rise and the saving–income ratio would return to its previous level. Thus the aggregate consumption function would shift upward in a ratcheting movement as aggregate income set new records.

Modigliani tested this formulation and estimated the parameters of the model using aggregate data for 1921 to 1940. James Duesenberry independently hit upon a very similar formulation. Duesenberry's ‘relative income hypothesis’ reconciled the time series and cross-sectional data by suggesting that the higher consumption of the poor was an attempt to keep up with those better situated economically. Both contributions were published in 1949. The differences in their theoretical justifications were generally

glossed over by subsequent commentators and the empirical model became known as the Duesenberry–Modigliani hypothesis.

The success of the Duesenberry-Modigliani empirical work (and the growing sophistication of econometric technique) produced a flurry of follow-up empirical studies. Modigliani and his collaborator Richard Brumberg described the state of affairs, in a passage that reflects Modigliani's scientific philosophy (2001, p. 129). Empirical work should test theory; theory should be inspired by empirical observation; and progress would be made only through the constant interplay between the two:

> It may be said that, at the date of this writing (1952), the analysis of the consumption function has degenerated into a morass of seemingly contradictory, or at least disconnected, results, with each new empirical finding adding less to our understanding than to the existing confusion. Further empirical analysis is not likely to advance us very far until the economic theorist has been able to provide a conceptual framework to give coherence to past findings and guidance for the collection of more ‘facts.’

Shortly after arriving at the University of Illinois, Modigliani began working with Brumberg to provide the missing conceptual foundation for the macroeconomic theory of consumption based on microeconomic marginal utility analysis. They produced two papers in 1952. The first, ‘Utility Analysis and the Consumption Function,’ was published in 1954. The other, ‘Utility Analysis and Aggregate Consumption Functions,’ was unpublished at the time of Blumberg's sudden and tragic death from a cerebral embolism in 1955. Modigliani was devastated by his friend's death and ‘lost all interest in revising the manuscript’ for publication (2001, p. 66). It remained unpublished for a quarter of century. It finally appeared in Modigliani's *Collected Papers* (1980) exactly as it had been left at the time of Blumberg's death. Together the two papers describe the life-cycle hypothesis (LCH).

The microeconomic model of consumption and saving proposed by Modigliani and Brumberg took the perspective of a forward-looking individual (or a couple) with a finite lifespan and no desire to bequeath a fortune to heirs (Keynes's proposed motive for aggregate saving was thus explicitly rejected). The model recognized that income will vary over the lifetime, rising at first as the individual's career advances and he or she gains experience and skill, but income will ultimately fall with age and may even disappear during retirement. With this view saving behaviour would vary over a person's lifetime. When young, the individual would save very little (when income is low relative to what can be expected in middle age). During the period of peak earnings in middle age, the individual's saving will be high as assets are accumulated to finance late life consumption and to afford retirement. When retired, the individual dissaves (saving is negative) as the accumulated assets are sold to support a planned retirement lifestyle.

The most familiar (and most simplified) exposition of the microeconomic model is that published by Modigliani in *Social Research* in 1966. That version pictured the expected income profile as flat and constant until retirement when it fell to zero and the desired consumption profile as flat throughout life. Over the lifespan the total of consumption would exactly exhaust the total income earned, but, since consumption must be maintained during the retirement years, consumption is less than income during the earning years. The 1966 article first introduced the diagram of the ‘Modigliani pyramid’, made famous by macroeconomic textbooks, that was reproduced in Modigliani's Nobel lecture, and which he came to view as his ‘trademark’ (2001, p. 60). In this diagram, the lifetime profile of wealth rises

linearly with age until it reaches a maximum on the day of retirement and then declines linearly with age until death, thus tracing out the pyramid shape. In the more general case, the wealth profile would be hump-shaped.

In the elementary formulation of the model it was assumed for convenience (and also to make a sharp contrast to the common view) that individuals had no desire to make a net bequest to heirs; that is, they had no reason to make net accumulations in order to bequeath a greater inheritance then they had received. Modigliani had always argued, however, that a bequest motive could be added to the LCH without disturbing its implications. Yet he maintained that empirically a bequest motive would be ‘relevant for the very rich (and especially for the *nouveaux riches*)’. At the same time Modigliani argued that in the absence of bequest motives there would still be substantial bequests left at death. If individuals knew the date of their death in advance, as assumed in the simplified exposition, then each individual over his lifetime would consume one hundred per cent of his or her lifetime income. The ‘life-time propensity to consume’ would be 1. Since people do not generally foresee the timing of their death, they must plan their saving to be sufficient to support them to a very old age. Since, alas, many die at a younger age then this, inheritance bequests are commonplace, but are for the most part unintended.

The importance of Modigliani–Brumberg *micro*economic model of saving lies in the *macro*economic implications of life-cycle behaviour. Aggregate saving in the LCH does not depend upon current income but on life-cycle income. Thus the age structure of the population matters. In a population that is growing rapidly because of natural increase or immigration, there will be more young and middle-aged savers than older retired dissavers. Aggregate saving will be higher. Likewise, in an economy that is experiencing rapid economic growth, perhaps produced by new technologies and strong investment in new capital, the young and middle-aged savers will look forward to higher lifetime earnings while the older dissavers are consuming at a level commensurate with their assets accumulated over a lifetime when productivity was lower. Thus growth is good for saving. Moreover, the higher aggregate rates of saving generated by either population growth or by economic growth can help sustain the forward progress by financing investment at continuing high levels. Saving is good for growth.

Another important long-run implication of the life-cycle hypothesis is that sustained government deficits will be a drag on economic growth. Modigliani called attention to the burden of the national debt in a famous paper published in the *Economic Journal* (1961). The government finances its deficit spending by issuing government bonds which are a form of net worth for those who purchase them. When members of the public hold some of their wealth in the form of bonds, the bonds substitute for the physical capital (machines, structures, and other productive capital) that would otherwise be created to satisfy the demand for life-cycle assets. The burden of the national debt is the reduced rate of growth attributable to the reduced rate of capital formation. This burden can be said to fall on future generations by reducing their income below what it would be otherwise.

Modigliani's analysis of the burden of the national debt was criticized by Robert Barro (1974). Barro's approach to the issue is also known as the ‘Ricardian equivalence theorem’ because it echoed a suggestion of David Ricardo. Barro rejected Modigliani's view that an individual's planning horizon is constrained to his expected lifetime and took the extreme opposite position that the planning horizon is infinite. Government deficits today, Barro argued, should lead to an increase in saving as taxpayers reasoned that taxes would have to be raised in some indefinite future to pay off the debt. To have the assets needed to meet this forecast tax increase, taxpayers would temporarily increase saving to set the required sum aside. Modigliani viewed Barro's assumption of an infinite horizon as ‘incredible’ and the

equivalence theorem ‘untenable’ (2000, p. 235). In characteristic fashion, however, he responded with carefully designed empirical tests rather than theoretical debate. In his presentation of the data, the LCH and the burden of the debt were supported and Ricardian equivalence rejected (1983a; 1986b).

The simple version of the LCH made no allowance for the Social Security pension system as an alternative to private saving. Modigliani argued that incorporating a mandatory government retirement plan into the model was straightforward and, more importantly, that treating Social Security consistently would clear up several important misunderstandings. As he would model it, Social Security's payroll taxes should be considered a form of forced or ‘compulsory’ saving that builds up ‘Social Security wealth’. The benefits received in old age should then be seen as drawing down those assets. When Social Security is included as a form of wealth, the empirical wealth profile has the hump shape predicted by the LCH (1983 with Arlie Sterling; 1987 and 2005 with Tullio Jappelli). This answered those critics who failed to find much dissaving in old age when using a conventional definition of saving. The critics had simply defined wealth too narrowly.

The introduction of Social Security into an economy that previously relied exclusively on private saving, according to Modigliani, would have two effects on the private saving rate. One is the replacement effect. Because the Social Security tax is a form of forced saving, individuals who count on the promised benefits can save less and on this account the society's wealth–income ratio will be reduced. On the other hand, there might be an offsetting ‘retirement effect’. A Social Security system will encourage earlier retirement both directly and through a social emulation effect. Longer retirement periods require greater wealth accumulation and thus increased saving rates. Empirical work reported by Modigliani and his coauthor Arlie Sterling suggests that the two effects roughly cancel each other out (1983).

The long-run implications of the LCH that saving is increased by economic growth, that the national debt produces a burden, and that there is little reason to think that the introduction of Social Security significantly reduced the saving rate challenged conventional views at the time. Not surprisingly, there were many critics. Modigliani's persistent defence of the logic of the theory and his continuous production (with the help of many coauthors) of ingeniously designed and carefully executed empirical verifications and rejoinders kept the model in the forefront of academic analysis and policy debate. It remains the accepted view.

Yet it was the short-run or cyclical implications – not the long-run consequences – that received the more immediate attention. In a pair of papers coauthored with Albert Ando, Modigliani directed attention to the short-run considerations and the implications for the aggregate time-series consumption function (1963; 1965). The underlying theory had been formulated in the still unpublished second paper with Richard Brumberg but the work with Ando brought the cyclical implications to the attention of the profession. The short-term consumption function proposed by Ando and Modigliani made consumption a linear function of aggregate disposable *labour* income (that is, income excluding the return to asset holdings and less the amount of personal taxes) and aggregate net worth. The coefficients of the two variables could be taken as empirically constant in the short run determined by the length of life, the length of retirement, and the rate of growth. It was not until estimates of the aggregate stock of net worth became available that the model could be verified empirically. When Raymond Goldsmith published his wealth estimates (1962), the life-cycle consumption function passed the battery of tests designed by Ando and Modigliani with the highest marks (Ando and Modigliani, 1963; Modigliani, 1966).

The cyclical properties of the LCH equation were not in themselves particularly novel. The LCH behaved in the short run not unlike the Duesenberry–Modigliani model or the roughly contemporaneous

theory of consumption put forward by Milton Friedman, the permanent income hypothesis (1957). The saving–income ratio would fall during recessions and rise during upturns, but would fluctuate about a fairly stable long-run average. And, like the simpler Keynesian model, the Ando–Modigliani formulation implied that tax cuts could stimulate consumption and thus help counteract recessionary tendencies. There were, however two novel implications of the cyclical formulation of the LCH with important policy implications. The short-run life-cycle consumption function postulates that consumption would be responsive to the value of assets; thus a stock market crash, like that of 1929, would tend to reduce consumption as individuals sought to restore their lost wealth. This was not an implication of the alternative models. As another contrast, Friedman suggested that consumption each period should depend upon the current rate of interest since consumers would be willing to save more (and consume less) when the reward to asset holding is high. Modigliani conjectured that the saving rate would be ‘largely independent’ of the interest rate. While accepting Friedman's point that an increase in the reward for saving (higher interest rates) would induce an increase in saving, Modigliani pointed out another consequence of high rates. High interest rates would allow the stock of assets to accumulate more rapidly, thus requiring less saving to reach the target level of assets needed for retirement. Modigliani suggested the two effects would largely cancel out. If Modigliani is correct, short-term policy strategies to increase saving by manipulating the rate of interest would be ruled out.

Expectations and fluctuations

One of Keynes's foremost contributions, according to his own view, was his emphasis on the importance of expectations. The central conclusion of his *General Theory*, announced in the Preface, was that a ‘monetary economy … is essentially one in which changing views about the future are capable of influencing the quantity of employment and not merely its direction’ (1936). It is somewhat ironic, then, that Modigliani's 1944 reformulation of the *General Theory* took expectations as given. We are told that this was a simplification for ‘convenience’ since the paper was concerned with ‘the determinants of equilibrium, and not with the explanation of business cycles’ (1944, p. 46). Most of the equations of his equilibrium model, to be sure, contain variables that represent the expectations of economic agents. But, to take account of any relevant *change* of views about the future, the analyst would have to shift one or more of the relationships expressed in the system of equations.

A few years after formulating the equilibrium model with static expectations, Modigliani began a far-ranging investigation of the role of anticipations and uncertainty in the explanation of business cycles. In 1949 Modigliani began work on a project he called ‘Economic Expectations and Fluctuations’. It would occupy him for more than ten years. Modigliani moved the project to Carnegie Tech in 1952. There he collaborated with Herbert Simon, Charles Holt, and John Muth (1960) and Kalman Cohen (1961) on two books concerned with anticipations, forecasting, and the use of inventories to smooth production. While the work on the life cycle and production smoothing explicitly recognized the importance of expectations in microeconomic models, a breakthrough came when Modigliani turned his attention to modelling the formation of expectations in a macroeconomic context. Modigliani collaborated with Emile Grunberg, a colleague at Carnegie, on the ‘Predictability of Social Events’ (1954), a famous paper that is widely recognized as introducing the concept of ‘rational expectations’ into economic theory. The concept itself is simple. Rational expectations are forecasts of the future that are consistent with the way the economy is believed to work. To adopt any other expectation would be to ignore whatever

knowledge one had about the workings of the economy. The macroeconomic implications of rational expectations, however, proved to be profound. In the hands of others, rational expectation formation was used to question the effectiveness of Keynesian monetary and fiscal stabilization policy, and thus Modigliani's ‘invention’ reappeared later as a challenge to the legitimacy of Keynesian economics.

The problem that Grunberg and Modigliani set out to explore was whether a widely believed public prediction of a future event might change individuals’ behaviour in such a way as to invalidate the prediction. Their answer was that a correct private prediction would be a wrong public prediction. Nevertheless, accurate public prediction was possible because the reaction of the public to the announcement can be taken into account by the social scientist. Accurate public predictions are predictions that are ‘internally consistent’ in the sense that they recognize and incorporate any change in public expectations induced by the prediction itself that would influence the course of events.

It was left to Modigliani's student at Carnegie, John Muth, to extend the concept of internally consistent expectations to become ‘rational expectations’, an exercise Muth (1961) carried out in a microeconomic context. Ten years later Robert Lucas (1972) returned the concept to a macroeconomic setting (in the context of a market-clearing model) and suggested that stabilization policies could not change real output in a predictable way if those policies were fully anticipated. The macroeconomic rational expectations model became the foundation of the ‘new classical macroeconomics’, so called because money had no real effects in this model. These developments took place in the 1970s and were led by others; meanwhile, Modigliani's thinking about expectations had been developing in another direction during the 1950s and 1960s. In his paper with Brumberg, Modigliani argued that, while anticipations about the future life course would be relevant to the individual's decision about current consumption, it was not necessary to take explicit account of uncertainty about the future. Uncertainty would simply give rise to an additional precautionary motive for saving, but the assets accumulated to satisfy the life-cycle motive would do double duty as a buffer stock to insure against emergencies. In making this argument Modigliani was echoing arguments that he had advanced in the books on business planning. In this work Modigliani observed firm behaviour and offered a description of how businesses form expectations about the future and how they make use of those anticipations in current decision making. The upshot was that, while knowledge about the future would be important, firms need not (and therefore did not) attempt to acquire all possible information. Much information about the future would be beyond the relevant planning horizon and therefore irrelevant, other information would not need to be precise, and some information might not be worth the effort to acquire. Businesses in the real world attempt to make the best possible forecasts of the variables deemed important, but sometimes best practice would be a rule of thumb or a simple extrapolation of the past. Any uncertainty that remained would be adequately hedged since inventory could do double duty and serve as a buffer stock against inadequately foreseen contingencies as well as smooth production.

This is a clearly pragmatic approach to expectations. Modigliani suggested that a pragmatic formulation was realistic. The ‘expectation function will, at best, appear in the form of broad statistical generalizations’ since expectations about the future range in practice ‘from the elaborate scientific forecast of the large business enterprise to primitive guesses and dark hunches’ (Grunberg and Modigliani, 1954, p. 471). It was this realistic approach to expectations that Modigliani later carried over to macroeconomic models. It is important to note, however, that Modigliani was not opposed to the idea of rational expectations in principle. He declared the concept ‘a good starting point’ for analysis and thought the assumption would be ‘sensible’ in some circumstances, for example, in financial markets

(1983b, pp. 123–4). He considered Muth's contribution ‘fundamental’ and an improvement over ‘naïve or *ad hoc* assumptions’ regarding the formation of expectations (1986b, p. 25). But when rational expectations were used to support the new classical economics and its startling proposition that stabilization policy would be ineffective, he thought that this was ‘pushing the idea of rationality well beyond the range where it is useful’ (1983b, p. 123). ‘It is a ‘wonderful theory … [but] it is *not* a description of the world’ (2000, p. 235).

Characteristically, Modigliani was not content to simply debate the logical merits of a model or the realism of its assumptions. The new classical macroeconomics could be rejected because its conclusions were inconsistent with the empirical evidence. The model implied that fluctuations in unemployment should be mild, short-lived, and random, contrary to all experience. The new classical model was also inconsistent with the existence of long-term contracts. If such contracts are rational, then wages are rigid, contrary to a postulate of the new classical view. If they are not rational, then they ‘should have long ago disappeared’.

Modigliani's pragmatic approach to modelling expectations was put to work in a series of papers on the term structure of interest rates (Modigliani and Sutch, 1966; 1967; Modigliani and Shiller, 1973). The ‘term structure’ refers to the relationship between interest rates on assets with different terms to maturity. In his collaboration with Sutch the long-term rate of interest was linked to the short-term rate through financial arbitrage. Since the investor could obtain a return over the long term by investing either in a long-term bond or alternatively in a sequence of short-term bills, the choice between the two would be influenced by the investor's expectations about the future course of short-term interest rates. Because the expectations would be subject to uncertainty, each investor would have a natural preference for assets with a maturity that matched their needs. But they could be tempted out of this ‘preferred maturity habitat’ if the advantage with shorter or longer maturities were forecast to be large enough.

An empirical characterization of expectation formation was required to complete the model of the relationship between short- and long-term rates. Here again Modigliani looked to how investors actually behaved. Modigliani and Sutch suggested that future expected rates were formulated by extrapolating past movements. They proposed that the recent trend in the rate would be anticipated to continue for a while, but that the best guess for the long run was that rates would return to their long historical average (as Keynes had suggested). Modigliani and Sutch considered this formulation a ‘plausible’ representation of how investors actually thought about the problem. Modigliani and Shiller went on to demonstrate that the Modigliani–Sutch model of expectations was also rational in the sense that it represented the best forecast possible on the basis of all information available.

Corporate finance

A key component of the Keynesian macroeconomic structure is the investment function, which held that the aggregate volume of investment would be responsive to the cost of capital. In the 1950s Modigliani turned his attention to this topic as well. The result was spectacular. In citing Franco Modigliani for the Nobel Prize in 1986, the Nobel Foundation's judges singled out both the life-cycle hypothesis and the path-breaking Modigliani–Miller theorems on corporate dividends, leverage, and the cost of capital (Modigliani and Miller, 1958; 1963; Miller and Modigliani, 1961; Modigliani. 1982). The two MM theorems, as they are called, not only overturned the existing thinking about the cost of capital but launched modern finance theory. Indeed, this line of research was deemed so important that Merton

Miller later received his own Nobel Prize in 1990 for his contribution to the joint work. In 1956 Merton Miller was an assistant professor auditing Modigliani's course at the Carnegie Institute of Technology. He became excited when Modigliani introduced the topic in class and agreed to join him in working out the proof.

The first Modigliani–Miller theorem establishes, when a firm's investment policy is fixed, that the market evaluation of a firm would be unaffected by its volume of debt in a simplified world with well-functioning financial markets, rational investors, and neutral taxes (1958). The second theorem, an extension of the first, states under the same assumptions that the value of the firm is independent of its dividend policy (1961). Taken together they suggest that ‘financial policy does not matter!’ (1982, p. 255).

The contribution to the scientific analysis of finance was profound. First, the MM papers introduced the application of microeconomic theory – and in particular the notion of arbitrage – to problems in corporate finance. Rigorous mathematical modelling has been the hallmark of the field ever since. Second, the two theorems taken together allow the separation for analytical and management purposes of investment decisions from financial decisions. The implication for the structuring of corporate management has led over time to the division of managerial responsibilities between the CEO and the CFO. Third, the MM theorems were established in the context of a highly stylized model, so a good deal of subsequent theoretical and empirical work has been devoted to understanding the impact of relaxing the simplifying assumptions and extending the application of the model. This research agenda has enriched the field immeasurably. The MM theorems, for example, have led directly to subsequent developments in the evaluations of options.

At the time the first MM theorem was published it was held to be self-evident that borrowing and taking on debt would lower the cost of capital to the firm because the rate of interest on the loans was below the cost of raising capital through the sale of equity. Modigliani and Miller elegantly demonstrated that the old theory was seriously flawed. As is typical of Modigliani's work, the result rests on the clear application of a microeconomic principle, in this case, the role of arbitrage. The intuition behind the two theorems is simple. No matter what the debt–equity structure of the firm (or its dividend–retained earnings policy) the investor can always undo the impact on his or her own stock portfolio by adding or subtracting other equities or forms of debt to the mix. The resulting arbitrage will mean that the market value of the firm will depend only on the income stream generated by its assets.

Despite the enormous literature that the MM theorems generated and despite the transformation of the field of corporate finance as a consequence, Modigliani was fond of trivializing the idea behind MM as ‘obvious’ and said that the theorems were written with ‘tongue-in-cheek’ as a way of chastising the ‘old school’ of finance for its reliance on anecdotes and rules of thumb relayed through case studies and the reminiscences of managers and accountants (2000, pp. 233–4). Yet the papers from 1958 and 1963 are among the top three most-cited papers by Modigliani and also among those he listed as his ‘personal favorites’ (Merton, 1987).

Other topics

In 1997 Modigliani published with Leah Modigliani, his granddaughter and a financial analyst, a paper entitled ‘Risk-Adjusted Performance: How to Measure It and Why’. Together they proposed a measure of the rate of return on an investment portfolio that was adjusted for risk so that the performance of

different fund portfolios could be compared with the same measuring rod. Their technique of risk adjustment is now widely used on Wall Street and has become known as M^2 – 'M-squared' – for the two Modiglianis. It applies the same concept of arbitrage introduced by Modigliani and Miller to neutralize the risk. If the historical volatility of a portfolio has been high relative to a benchmark (say the S&P 500), its risk could hypothetically be reduced to match that of the benchmark by adding treasury bills in sufficient quantity to the mix. If the investment portfolio under study has a volatility below that of the benchmark, it could hypothetically be levered up to match the risk standard by borrowing on margin and investing additional sums in the fund. The rates of return can then be calculated and compared for these blended portfolios.

A longer review of Modigliani's work would have to find space to discuss his writings on the Italian economy with La Malfa (1967), Tarantelli (1975), Padoa-Schioppa and Rossi (1986), and Jappelli (1987) and on the European economy and international finance (Part III of *Collected Papers*, vol. 3), to mention just a few of the papers that were published in English. For Modigliani's recounting of this work, much of which is in Italian, see chapters 2 and 3 of his autobiography, *Adventures of an Economist* (2001). A review of those contributions will suggest that there are serious omissions from the present survey.

Retrospect

Modigliani was a brilliant economist who took the real problems of the real world seriously and then developed powerful theories to explain what he saw. Yet he was uncomfortable with his theories until he had rigorously tested them against the data and against alternative explanations. These econometric explorations invariably stimulated him to re-examine his thinking. For him, it was a process without end. Modigliani rarely let a topic go, he worked continuously to refine, improve, and, when necessary, to defend each of his signature contributions. He was driven by a strong faith in the power of an economic theory so derived to inform policy, to solve problems, and to right social wrongs. He acted on his beliefs by becoming an advisor to – or, when they would not listen, a public opponent of – those who made economic policy. He changed his mind when logic or facts dictated it, yet he remained steadfast in his belief that economics was a science with the potential to make the world a better place. This combination of dedication, intellectual honesty, and liberal values made him one of the most influential macroeconomists of the 20th century.

His personal characteristics were an important ingredient of his success. He was warm and caring, intense and excitable, enthusiastic and full of an infectious joy for life. He was a charismatic teacher, a tenacious debater, and a seminal thinker with the rare ability to stimulate others to think and imagine beyond their usual capacity.

See Also

- expectations
- Keynesian revolution
- Miller, Merton
- Modigliani–Miller theorem

- rational expectations
- term structure of interest rates

Partial support for this research was provided by the National Science Foundation and the Center for Social and Economic Policy at the University of California, Riverside.

Selected works

The most important scientific contributions of Modigliani have been collected and reprinted in *The Collected Papers of Franco Modigliani*, 6 vols., ed. A. Abel, S. Johnson and F. Franco (1980–2005), Cambridge, MA: MIT Press.

1944. Liquidity preference and the theory of interest and money. *Econometrica* 12, 45–88.

1949. Fluctuations in the saving–income ratio: a problem in economic forecasting. In *Studies in Income and Wealth*, vol. 11. New York: NBER.

1954. (With R. Brumberg.) Utility analysis and the consumption function: an interpretation of cross-section Data. In *Post-Keynesian Economics*, ed. K. Kurihara. New Brunswick, NJ: Rutgers University Press.

1954. (With E. Grunberg.) The predictability of social events. *Journal of Political Economy* 62, 465–78.

1958. (With M. Miller.) The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48, 261–97.

1960. (With C. Holt, J. Muth and H Simon.) *Planning Production, Inventories, and Work Force*. Englewood Cliffs, NJ: Prentice-Hall.

1961. Long-run implications of alternative fiscal policies and the burden of the national debt. *Economic Journal* 71, 730–55.

1961. (With K. J. Cohen.) *The Role of Anticipations and Plans in Economic Behavior and their Use in Economic Analysis and Forecasting*. Urbana: University of Illinois.

1961. (With M. Miller.) Dividend policy, growth, and the valuation of shares. *Journal of Business* 34, 411–33.

1963. The monetary mechanism and its interaction with real phenomena. *Review of Economics and Statistics* 45(1) Part 2, Supplement, 79–107.

1963. (With A. Ando.) The ‘life-cycle’ hypothesis of saving: aggregate implications and tests. *American Economic Review* 53, 55–84.

1963. (With M. Miller.) Corporate income tax and the cost of capital: a correction. *American Economic Review* 53, 433–43.

1964. Some empirical tests of monetary management and of rules versus discretion. *Journal of Political Economy* 72, 211–45.

1965. (With A. Ando.) The relative stability of monetary velocity and the investment multiplier. *American Economic Review* 55, 693–728.

1966. The life-cycle hypothesis of saving: the demand for wealth and the supply of capital. *Social Research* 33, 160–217.

1966. (With R. Sutch.) Innovations in interest rate policy. *American Economic Review* 56, 178–97.

1967. (With G. La Malfa.) Inflation, balance of payments deficit, and their cure through monetary policy: the Italian example. *Banca Nazionale del Lavoro Quarterly Review* 80, 3–47.

1967. (With R. Sutch.) Debt management and the term structure of interest rates: an empirical analysis. *Journal of Political Economy* 75, 569–89.

1969. (With A. Ando.) Econometric analysis of stabilization policies. *American Economic Review* 59, 296–314.

1973. (With R. Shiller.) Inflation, rational expectations and the term structure of interest rates. *Economica* 40, 12–43.

1975a. Channels of monetary policy in the Federal Reserve-MIT-University of Pennsylvania Econometric Model of the United States. In *Modelling the Economy*, ed. G. Renton. London: Heinemann Educational Books.

1975b. The life-cycle hypothesis of saving twenty years later. In *Contemporary Issues in Economics*, ed. M. Parkin and A. Nobay. Manchester: Manchester University Press.

1975. (With E. Tarantelli.) The consumption function in a developing economy and the Italian experience. *American Economic Review* 65, 825–42.

1977. The monetarist controversy or, Should we forsake stabilization policies? Presidential address delivered at the American Economic Association. *American Economic Review* 67, 1–19.

1980. (With R. Brumberg.) Utility analysis and aggregate consumption functions: an attempt at integration. In *The Collected Papers of Franco Modigliani*, vol. 2, ed. A. Abel. Cambridge, MA: MIT Press.

1982. Debt, dividend policy, taxes, inflation, and market valuation. Presidential address delivered at the American Finance Association. *Journal of Finance* 37, 255–73.

1983a. Government deficits, inflation, and future generations. In *Deficits: How Big and How Bad?*, ed. D. Conklin and T. Courchene. Ontario: Ontario Economic Council.

1983b. Interview with Franco Modigliani. In *Conversations with Economists: New Classical Economists and Their Opponents Speak Out on the Current Controversy in Macroeconomics*, ed. A. Klamer. Totowa, NJ: Rowman and Littlefield.

1983. (With A. Sterling.) Determinants of private saving with special reference to the role of Social Security – cross-country tests. In *The Determinants of National Saving and Wealth*, ed. F. Modigliani and R. Hemming. London: Macmillan.

1986a. Autobiography. *Les Prix Nobel. The Nobel Prizes 1985*, ed. W. Odelberg. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1985/modigliani-autobio.html, accessed 9 September 2006.

1986b. Life cycle, individual thrift, and the wealth of nations. Nobel Prize Lecture, 1985. Reprinted in *American Economic Review* 76, 297–313.

1986c. *The Debate over Stabilization Policy*. Cambridge: Cambridge University Press.

1986. (With F. Padoa-Schioppa and N. Rossi.) Aggregate unemployment in Italy, 1960–1983. *Economica* 53, S245–S273, S347–S352.

1986. (With A. Sterling.) Government debt, government spending, and private sector behavior: comment. *American Economic Review* 76, 1168–79.

1987. (With T. Jappelli.) Fiscal policy and saving in Italy since 1860. In *Private Saving and Public Debt*, ed. M. Boskin, J. Flemming and S. Gorini. Oxford: Basil Blackwell.

1988a. The role of intergenerational transfers and life cycle saving in the accumulation of wealth. *Journal of Economic Perspectives* 2(2), 15–40.

1988b. MM – past, present, future. *Journal of Economic Perspectives* 2(4), 149–58.

1997. (With L. Modigliani.) Risk-adjusted performance: how to measure it and why. *Journal of Portfolio Management* 23(2), 45–54.

2000. An interview with Franco Modigliani. Interviewed by W. Barnett and R. Solow, 5–6 November 1999. *Macroeconomic Dynamics* 4, 222–56.

2001. *Adventures of an Economist*. New York: Texere.

2005. (With T. Jappelli.) The age–saving profile and the life–cycle hypothesis. In *The Collected Papers of Franco Modigliani*, vol. 6, ed. F. Franco. Cambridge, MA: MIT Press.

Bibliography

Barro, R. 1974. Are government bonds net worth? *Journal of Political Economy* 82, 1095–117.

Duesenberry, J. 1949. *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton, NJ: Princeton University Press.

Goldsmith, R. 1962. *The National Wealth of the United States in the Postwar Period*. Princeton: Princeton University Press.

Hicks, J. 1937. Mr Keynes and the ‘Classics’. *Econometrica* 5, 147–59.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. New York: Harcourt, Brace.

Lucas, R., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Merton, R. 1987. In honor of Nobel Laureate, Franco Modigliani. *Journal of Economic Perspectives* 1 (2), 145–55.

MIT (Massachusetts Institute of Technology). 2003. ‘Nobel laureate Franco Modigliani dies at 85’. News Office, MIT. Online. Available at http://web.mit.edu/newsoffice/2003/modigliani.html, accessed 11 November 2003.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

How to cite this article

Sutch, Richard. "Modigliani, Franco (1918–2003)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000341> doi:10.1057/9780230226203.1113

The New Palgrave Dictionary of Economics Online

Modigliani – Miller theorem

Anne P. Villamil

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Modigliani–Miller theorem provides conditions under which a firm's financial decisions do not affect its value. The theorem is one of the first formal uses of a no arbitrage argument and it focused the debate about firm capital structure around the theorem's assumptions, which set the conditions for effective arbitrage. The search for the source of the ‘failure of irrelevance’ has led to important advances in the nature of financial structure, and more fundamentally to the types of frictions that would cause agents to have different market opportunities, information sets or commitment frictions.

Keywords

agency problems; arbitrage; bankruptcy; capital gains; control rights; debt versus equity; debt–equity ratio; dividend policy; finance; imperfect information; incomplete contacts; information revelation; insider trading; law of one price; leasing vs. buying; Miller, M.; Modigliani, F.; Modigliani–Miller theorem; monitoring; optimal capital structure; optimal contracts; separating equilibrium; separation of ownership and control; shareholder voting rights; signalling models of equity; taxation of capital income; verification costs

Article

The Modigliani–Miller theorem is a cornerstone of modern corporate finance. At its heart, the theorem is an irrelevance proposition: it provides conditions under which a firm's financial decisions do not affect its value. Modigliani explains the theorem as follows:

> … with well-functioning markets (and neutral taxes) and rational investors, who can ‘undo’ the corporate financial structure by holding positive or negative amounts of debt, the market value of the firm – debt plus equity – depends *only* on the income stream generated by its assets. It follows, in particular, that the value of the firm should not be

> affected by the share of debt in its financial structure or by what will be done with the returns – paid out as dividends or reinvested (profitably). (Modigliani, 1980, p. xiii)

In fact, what is currently understood as the Modigliani–Miller theorem comprises four distinct results from a series of papers (1958; 1961; 1963). The first proposition establishes that under certain conditions, a firm's debt–equity ratio does not affect its market value. The second proposition establishes that a firm's leverage has no effect on its weighted average cost of capital (that is, the cost of equity capital is a linear function of the debt–equity ratio). The third proposition establishes that firm market value is independent of its dividend policy. The fourth proposition establishes that equity-holders are indifferent about the firm's financial policy.

Miller (1991, p. 5) explains the intuition for the theorem with a simple analogy. ‘Think of the firm as a gigantic tub of whole milk. The farmer can sell the whole milk as it is. Or he can separate out the cream, and sell it at a considerably higher price than the whole milk would bring.’ He continues: ‘The Modigliani–Miller proposition says that if there were no costs of separation (and, of course, no government dairy support programme), the cream plus the skimmed milk would bring the same price as the whole milk.’ The essence of the argument is that increasing the amount of debt (cream) lowers the value of outstanding equity (skimmed milk) – selling off safe cash flows to debt-holders leaves the firm with more lower-valued equity, keeping the total value of the firm unchanged. Put differently, any gain from using more of what might seem to be cheaper debt is offset by the higher cost of now riskier equity. Hence, given a fixed amount of total capital, the allocation of capital between debt and equity is irrelevant because the weighted average of the two costs of capital to the firm is the same for all possible combinations of the two.

The theorem makes two fundamental contributions. In the context of the modern theory of finance, it represents one of the first formal uses of a no arbitrage argument (though the ‘law of one price’ is long-standing). More fundamentally, it structured the debate on why irrelevance fails around the theorem's assumptions: (i) neutral taxes; (ii) no capital market frictions (that is, no transaction costs, asset trade restrictions or bankruptcy costs); (iii) symmetric access to credit markets (that is, firms and investors can borrow or lend at the same rate); and (iv) firm financial policy reveals no information. Modigliani and Miller (1958) also assumed that each firm belonged to a ‘risk class’, a set of firms with common earnings across states of the world, but Stiglitz (1969) showed that this assumption is not essential. The relevant assumptions are important because they set conditions for effective arbitrage: When a financial market is not distorted by taxes, transaction or bankruptcy costs, imperfect information or any other friction which limits access to credit, then investors can costlessly replicate a firm's financial actions. This gives investors the ability to ‘undo’ firm decisions, if they so desire. Attempts to overturn the theorem's controversial irrelevance result were a fortiori arguments about which of the assumptions to reject or amend. The systematic analysis of these assumptions led to an expansion of the frontiers of economics and finance.

The importance of taxes for the irrelevance of debt versus equity in the firm's capital structure was considered in Modigliani and Miller's original paper (1958). Modigliani and Miller (1963) and Miller (1977) addressed the issue more specifically, showing that under some conditions, the optimal capital structure can be complete debt finance due to the preferential treatment of debt relative to equity in a tax code. For example, in the United States, interest payments on debt are excluded from corporate taxes. As

a consequence, substituting debt for equity generates a surplus by reducing firm tax payments to the government. Firms can then pass this surplus on to investors in the form of higher returns. This raised the further provocative question – were firms that issued equity leaving stockholder money on the table in the form of unnecessary corporate income tax payments? Miller (1977) resolved this problem by showing that a firm could generate higher after-tax income by increasing the debt–equity ratio, and this additional income would result in a higher payout to stockholders and bondholders, but the value of the firm need not increase. The crux of the argument is that as debt is substituted for equity, the proportion of firm payouts in the form of interest on debt rises relative to payouts in the form of dividends and capital gains on equity. Taxes that are higher on interest payments than on equity returns reduce or eliminate the advantage of debt finance to the firm.

The remaining Modigliani–Miller assumptions deal with various types of capital market frictions (for example, transaction costs or imperfect information) that are at the heart of arbitrage. The driving force in a perfect market for a homogeneous good is the ‘law of one price’. If debt and equity are merely different packages of an underlying homogeneous good – capital – and there are no market imperfections, then it follows immediately that the law of one price holds due to arbitrage. Investors simply engage in arbitrage until any deviation in the price of the two forms of capital is eliminated. Thus, the remaining discussion is organized around the implications of the theorem for firm capital structure, dividend policy, and the method of capital finance (lease versus buy).

With regard to firm capital structure, the theorem opened a literature on the fundamental nature of debt versus equity. Are debt and equity distinct forms of capital? Why and in what specific ways? In order to answer these questions about the nature of capital, the optimal contract literature examines debt and equity as financial contracts that arise optimally in response to particular market frictions, when contracting possibilities are complete or incomplete. Complete contracts can be written on all states if this is optimal; incomplete contacts cannot depend on some states of nature.

In one of the earliest contributions, Townsend (1979) combines elements of imperfect information and bankruptcy costs to examine the nature of debt in a complete contracting environment. In his costly state verification model, debt is an optimal response to costly monitoring and differential information: all agents know *ex ante* the distribution of firm returns, but only the firm privately and costlessly observes the return *ex post*. The lender can acquire this information, but must irrevocably commit to pay a deadweight verification cost. Townsend shows that debt is optimal because it minimizes this cost. When the firm makes the required fixed debt repayment, no cost is incurred. Only when the firm is insolvent, and hence cannot repay its debt fully, does verification occur. Townsend interprets this as costly bankruptcy (liquidation): the firm is shut down; firm assets are seized by a ‘court’, which verifies their magnitude and transfers the residual to the lender, net of the verification cost. Lacker and Weinberg (1989) extend the approach by specifying conditions under which equity is optimal in an analogue of the model, costly state falsification. Neither debt nor equity is *ex post* efficient in this class of models because no agent wishes to request costly intervention and incur the deadweight cost, *ex post*. Agents know that bankruptcy occurs only when the firm is truly unable to repay due to a low realization, but they are implicitly assumed to be committed to the decisions they made *ex ante*. Otherwise, debt is no longer optimal.

Krasa and Villamil (2000) show that a firm–lender investment problem with multiple stages, costly enforcement, limited commitment and an explicit enforcement decision, can illuminate debt's distinct properties. The analysis also solves the *ex post* inefficiency problem in the costly state verification

model. Agents write a contract in the initial period, knowing only the distribution of project returns. The contract specifies payments and when enforcement will occur, and can be altered if agents receive new information. In the next period, the borrower privately observes the return and can make the unenforceable payment specified in the original contract or propose an alternative payment (that is, renegotiate). In the final stage the investor can seek costly enforcement of the contractually specified payment or renegotiate enforcement. The opportunity to renegotiate is important because it introduces a new source of information: any positive renegotiation payment by the firm would reveal information to the investor about the firm's state. Debt is optimal because it minimizes information revelation. Renegotiation, which imposes a constraint on the contract problem, is only relevant when an agent acquires new information and can use the information to alter the initial contract. Debt weakens agents' incentive to renegotiate by minimizing information revelation (a fixed face value reveals no information about the firm). The contract is *ex post* efficient because all decisions are chosen optimally as part of a perfect Bayesian Nash equilibrium. This minimal information revelation of debt stands in sharp contrast to the active information revelation in signalling models of equity. For example, in Leland and Pyle (1977) retained equity by a firm signals a profit increase sufficient to offset the owner's forgone diversification. In Myers and Majluf (1984), issuing equity signals bad news – owners with inside information sell shares when markets overvalue them. These signalling models leave open why a firm would use financial decisions to reveal information, a problem that does not arise in Krasa and Villamil.
In incomplete contracting models, control rights are an alternative justification for debt and equity contracts. Aghion and Bolton (1992) view debt as a particular assignment of control rights with important incentive properties. They show that when contracting possibilities are exogenously incomplete and control rights are assigned entirely to the investor or the firm, the first-best contract cannot be implemented. If the investor has sole control, the investor may force the firm to expand to a suboptimal level. Alternatively, if the firm has sole control it may not liquidate optimally. Aghion and Bolton show that, under some conditions, debt is the optimal contract because it assigns control to the firm in good states but to the investor in bad states. This ensures that optimal decisions are made in solvency and default states. Zender (1991) extends the model to include both debt and equity contracts. Grossman and Hart (1988) and Harris and Raviv (1988) examine control in the context of voting rights. They focus on the 'one vote per share' property of equity and majority voting, showing circumstances under which equity is optimal and when other 'extreme securities' are optimal.
Instead of focusing on the properties of debt and equity per se, Allen and Gale (1988; 1991) examine the properties of optimal securities more broadly, especially financial innovation. They study the problem of a firm that can issue securities in a market where the transaction cost of issuing securities makes the market incomplete. Market structure is endogenous in the sense that firms choose the securities they issue, which determines the transaction costs they incur. Allen and Gale (1988) prohibit short sales and show that neither debt nor equity is optimal. In contrast, Allen and Gale (1991) permit unlimited short sales, and show by example that debt and equity can be optimal. They note that the example is a special case; in general their model predicts that optimal securities are much more complex than those typically observed. The debt–equity puzzle unleashed by Modigliani and Miller continues to be an active area of research. The common theme of both the complete and incomplete contracting literatures is that debt, equity, and hybrid securities arise endogenously to overcome frictions in capital markets. Debt and equity have unique properties that resolve these frictions.

The Miller and Modigliani (1961) and Miller (1977) result that firm value is independent of dividend policy has also been examined extensively. Bhattacharya (1979) and others show that firm dividend policy can be a costly device to signal a firm's state, and hence relevant, in a class of models with: (i) asymmetric information about stochastic firm earnings; (ii) shareholder liquidity (a need to sell makes firm valuation relevant); and (iii) deadweight costs (to pay dividends, refinance cash flow shocks or cover underinvestment). In a separating equilibrium, only firms with high anticipated earnings pay high dividends, thus signalling their prospects to the stock market. As in other costly signalling models, the question as to why a firm would use financial decisions to reveal information, rather than direct disclosure, must be addressed. As noted previously, taxes are another important friction that effect dividend policy (for example, see Allen, Bernardo and Welch, 2000).

Finally, Miller and Upton (1976) show that firms are indifferent between leasing and buying capital, except when they face different tax rates. Myers, Dill and Bautista (1976) develop a formula to evaluate the lease versus buy decision, where different tax rates across firms create different discount rates. They show it is optimal for low tax rate, and hence high discount rate firms, to lease. Alchian and Demsetz (1972) show that leasing involves agency costs due to the separation of ownership and control of capital; a lessee may not have the same incentive as an owner to properly use or maintain the capital. Coase (1972) and Bulow (1986) argue that a durable goods monopolist may lease in order to avoid time inconsistency, and Hendel and Lizzari (1999; 2002) show that it may lease to reduce competition or adverse selection in secondary (used goods) markets. Eisfeldt and Rampini (2007) show that leasing has a repossession advantage relative to buying via secured lending. The trade-off involves the benefit of the enforcement advantage for leased capital, relative to the cost of the ownership, versus a standard control agency problem which arises because ownership and control are separated.

In addition to these specific advances in financial structure, an essential part of Modigliani and Miller's innovation was to put agents on an equal footing. They, and others, then asked what types of friction would cause agents to have different market opportunities, information sets or commitment frictions? This perspective, which was novel at the time, has been used productively to analyse problems in monetary economics, public finance, international economics, and a number of other applications. In summary, the most profound and lasting impacts of the Modigliani–Miller theorem have been this notion of ‘even footedness’ and the systematic investigation of the theorem's assumptions. The approach has motivated decades of research in economics and finance in a search for what *is* relevant in a host of economic problems (between borrowers and lenders, governments and citizens, and countries). As Miller (1988, p. 100) said: ‘Showing what doesn't matter can also show, by implication, what does.’

See Also

- arbitrage
- finance
- Miller, Merton
- Modigliani, Franco

Bibliography

Aghion, P. and Bolton, P. 1992. An incomplete contracts approach to financial contracting. *Review of Economic Studies* 59, 473–94.

Alchian, A. and Demsetz, H. 1972. Production, information costs and economic organization. *American Economic Review* 62, 777–95.

Allen, F., Bernardo, A. and Welch, I. 2000. A theory of dividends based on tax clienteles. *Journal of Finance* 55, 2499–536.

Allen, F. and Gale, D. 1988. Optimal security design. *Review of Financial Studies* 1, 229–63.

Allen, F. and Gale, D. 1991. Arbitrage, short sales and financial innovation. *Econometrica* 59, 1041–68.

Bhattacharya, S. 1979. Imperfect information, dividend policy, and the ‘bird in the hand’ fallacy. *Bell Journal of Economics* 10, 259–70.

Bulow, J. 1986. An economic theory of planned obsolescence. *Quarterly Journal of Economics* 101, 729–49.

Coase, R. 1972. Durability and monopoly. *Journal of Law and Economics* 15, 142–9.

Eisfeldt, A. and Rampini, A. 2007. Leasing, ability to repossess and debt capacity. *Review of Financial Studies*, forthcoming.

Grossman, S. and Hart, O. 1988. One share one vote and the market for corporate control. *Journal of Financial Economics* 20, 175–202.

Harris, M. and Raviv, A. 1988. Corporate governance: voting rights and majority rule. *Journal of Financial Economics* 20, 203–35.

Hendel, I. and Lizzari, A. 1999. Interfering with secondary markets. *RAND Journal of Economics* 30, 1–21.

Hendel, I. and Lizzari, A. 2002. The role of leasing under adverse selection. *Journal of Political Economy* 110, 113–43.

Krasa, S. and Villamil, A.P. 2000. Optimal contracts when enforcement is a decision variable. *Econometrica* 68, 119–34.

Lacker, J. and Weinberg, J. 1989. Optimal contracts under costly state falsification. *Journal of Political*

Economy 97, 1345–63.

Leland, H. and Pyle, D. 1977. Informational asymmetries, financial structure and financial intermediation. *Journal of Finance* 32, 371–87.

Miller, M.H. 1977. Debt and taxes. *Journal of Finance* 32, 261–75.

Miller, M.H. 1988. The Modigliani-Miller proposition after thirty years. *Journal of Economic Perspectives* 2(4), 99–120.

Miller, M.H. 1991. *Financial Innovations and Market Volatility*. Cambridge, MA: Blackwell.

Miller, M.H. and Modigliani, F. 1961. Dividend policy, growth and the valuation of shares. *Journal of Business* 34, 411–33.

Miller, M.H. and Upton, C. 1976. Leasing, buying and the cost of capital services. *Journal of Finance* 31, 761–86.

Modigliani, F. 1980. Introduction. In *The Collected Papers of Franco Modigliani*, vol. 3, ed. A. Abel. Cambridge, MA: MIT Press.

Modigliani, F. and Miller, M.H. 1958. The cost of capital, corporate finance and the theory of investment. *American Economic Review* 48, 261–97.

Modigliani, F. and Miller, M.H. 1963. Corporate income taxes and the cost of capital: a correction. *American Economic Review* 53, 433–43.

Myers, S., Dill, D. and Bautista, A. 1976. Valuation of financial lease contracts. *Journal of Finance* 31, 799–819.

Myers, S. and Majluf, N. 1984. Corporate financing and investment when firms have information that investors do not have. *Journal of Financial Economics* 11, 187–221.

Stiglitz, J. 1969. A re-examination of the Modigliani–Miller theorem. *American Economic Review* 59, 784–93.

Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 22, 265–93.

Zender, J. 1991. Optimal financial instruments. *Journal of Finance* 46, 1645–63.

How to cite this article

Villamil, Anne P. "Modigliani–Miller theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000187> doi:10.1057/9780230226203.1114

The New Palgrave Dictionary of Economics Online

monetarism

Phillip Cagan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Brunner, K.; budget deficits; cash balance approach; central banks; Chicago School; consumer expenditure; consumption multiplier; crowding out; Friedman, M.; gold standard; inflation; Keynesianism; liquidity trap; Meltzer, A.; Mints, L.; monetarism; monetary approach to the balance of payments; monetary base; monetary targeting; monetary transmission mechanism; money supply; natural rate of unemployment; Phillips curve; quantity theory of money; rational expectations; Simons, H.; stabilization; unemployment; velocity of circulation

Article

Monetarism is the view that the quantity of money has a major influence on economic activity and the price level and that the objectives of monetary policy are best achieved by targeting the rate of growth of the money supply.

Background and initial development

Monetarism is most closely associated with the writings of Milton Friedman who advocated control of the money supply as superior to Keynesian fiscal measures for stabilizing aggregate demand. Friedman (1948) had proposed that the government finance budget deficits by issuing new money and use budget surpluses to retire money. The resulting countercyclical variations in the money stock would stabilize the economy, provided that the government set its expenditures and tax rates to balance the budget at full employment. In his *A Program for Monetary Stability* (1960), however, Friedman proposed that constant growth of the money stock, divorced from the government budget, would be simpler and equally effective for stabilizing the economy.
In their emphasis on the importance of money, these proposals followed a tradition of the Chicago School of economics. Preceding Friedman at the University of Chicago, Henry Simons (1936) had advocated control of the money stock to achieve a stable price level, and Lloyd Mints (1950) laid out a

specific monetary programme for stabilizing an index of the price level. These writers rejected reliance on the gold standard because it had failed in practice to stabilize the price level or economic activity. Such views were not confined to the University of Chicago. In the 1930s James Angell of Columbia University (1933) advocated constant monetary growth, and in the post-Second World War period Karl Brunner and Allan Meltzer were influential proponents of monetarism. The term ‘monetarism’ was first used by Brunner (1968). He and Meltzer founded the ‘Shadow Open Market Committee’ in the 1970s to publicize monetarist views on how the Federal Reserve should conduct monetary policy. Monetarism gradually gained adherents not only in the United States but also in Britain (Laidler, 1978) and other Western European countries, and subsequently around the world. The growing prominence of monetarism led to intense controversy among economists over the desirability of a policy of targeting monetary growth.

The roots of monetarism lie in the quantity theory of money which formed the basis of classical monetary economics from at least the 18th century. The quantity theory explains changes in nominal aggregate expenditures – reflecting changes in both the physical volume of output and the price level – in terms of changes in the money stock and in the velocity of circulation of money (the ratio of aggregate expenditures to the money stock). Over the long run changes in velocity are usually smaller than those in the money stock and in part are a result of prior changes in the money stock, so that aggregate expenditures are determined largely by the latter. Moreover, over the long run growth in the physical volume of output is determined mainly by real (that is, non-monetary) factors, so that monetary changes mainly influence the price level. The observed long-run association between money and prices confirms that inflation results from monetary overexpansion and can be prevented by proper control of the money supply. This is the basis for Friedman's oft-repeated statement that inflation is always and everywhere a monetary phenomenon.

The importance of monetary effects on price movements had been supported in empirical studies by classical and neoclassical economists such as Cairnes, Jevons and Cassel. But these studies suffered from limited data, and the widespread misinterpretation of monetary influences in the Great Depression of the 1930s fostered doubts about their importance in business cycles. As Keynesian theory revolutionized thinking in the late 1930s and 1940s, it offered an influential alternative to monetary interpretations of business cycles.

The first solid empirical support for a monetary interpretation of business cycles came in a series of studies of the United States by Clark Warburton (for example, 1946). Subsequently Friedman and Anna J. Schwartz compiled new data at the National Bureau of Economic Research in an extension of Warburton's work. In 1962 they demonstrated that fluctuations in monetary growth preceded peaks and troughs of all US business cycles since the Civil War. Their dates for significant steps to higher or lower rates of monetary growth showed a lead over corresponding business cycle turns on the average by about a half year at peaks and by about a quarter year at troughs, but the lags varied considerably. Other studies have found that monetary changes take one to two years or more to affect the price level.

In *A Monetary History of the United States, 1867–1960* (1963b) Friedman and Schwartz detailed the role of money in business cycles and argued in particular that severe business contractions like that of 1929–33 were directly attributable to unusually large monetary contractions. Their monetary studies were continued in *Monetary Statistics of the United States* (1970) and *Monetary Trends in the United States and the United Kingdom* (1982). A companion National Bureau study *Determinants and Effects of*

Changes in the Stock of Money (1965) by Phillip Cagan presented evidence that the reverse effect of economic activity and prices on money did not account for the major part of their observed correlation, which therefore pointed to an important causal role of money.

The monetarist proposition that monetary changes are responsible for business cycles was widely contested, but by the end of the 1960s the view that monetary policy had important effects on aggregate activity was generally accepted. The obvious importance of monetary growth in the inflation of the 1970s restored money to the centre of macroeconomics.

Monetarism versus Keynesianism

Monetarism and Keynesianism differ sharply in their research strategies and theories of aggregate expenditures. The Keynesian theory focuses on the determinants of the components of aggregate expenditures and assigns a minor role to money holdings. In monetarist theory money demand and supply are paramount in explaining aggregate expenditures.

To contrast the Keynesian and monetarist theories, Friedman and David Meiselman (1963) focused on the basic hypothesis about economic behaviour underlying each theory: for the Keynesian theory the consumption multiplier posits a stable relationship between consumption and income, and for the monetarist theory the velocity of circulation of money posits a stable demand function for money. Friedman and Meiselman tested the two theories empirically using US data for various periods by relating consumption expenditures in one regression to investment expenditures, assuming a constant consumption multiplier, and in a second regression to the money stock, assuming a constant velocity. They reported that the monetarist regression generally fitted the data much better. These dramatic results were not accepted by Keynesians, who argued that the Keynesian theory was not adequately represented by a one-equation regression and that econometric models of the entire economy, based on Keynesian theory, were superior to small-scale models based solely on monetary changes.

The alleged superiority of Keynesian models was contested by economists at the Federal Reserve Bank of St Louis (see Andersen and Jordan, 1968). They tested a ‘St Louis equation’ in which changes in nominal GNP depended on current and lagged changes in the money stock, current and lagged changes in government expenditures, and a constant term reflecting the trend in monetary velocity. When fitted to historical US data, the equation showed a strong permanent effect of money on GNP and a weak transitory (and in later work, non-existent) effect of the fiscal variables, contradicting the Keynesian claim of the greater importance of fiscal than monetary policies. Although the St Louis equation was widely criticized on econometric issues, it was fairly accurate when first used in the late 1960s to forecast GNP, which influenced academic opinion and helped bring monetarism to the attention of the business world.

Although budget deficits and surpluses change interest rates and thus can affect the demand for money, monetarists believe that fiscal effects on aggregate demand are small because of the low interest elasticity of money demand. Government borrowing crowds out private borrowing and associated spending, and so deficits have little net effect on aggregate demand. The empirical results of the St Louis equation are taken as confirmation of weak transitory effects. The debate over the effectiveness of fiscal policy as a stabilization tool has produced a large literature.

In their analysis of the transmission of monetary changes through the economy, Brunner and Meltzer

(1976) compare the effects of government issues of money and bonds. If the government finances increased expenditures in a way that raises the money supply, aggregate expenditures increase and nominal income rises. Moreover, the increased supply of money adds to the public's wealth, and greater wealth increases the demand for goods and services. This too raises nominal income. The rise in nominal income is at first mainly a rise in real income and later a rise in prices. They compare this result with one in which the government finances its increased expenditures by issuing bonds rather than money. Again wealth increases, and this raises aggregate expenditures. As long as the government issues either money or bonds to finance a deficit, nominal income must rise due to the increase in wealth. Brunner and Meltzer therefore agree with Keynesians that in principle a deficit financed by bonds as well as by new money is expansionary. However, they show that the empirical magnitudes of the economy are such that national income rises more from issuing a dollar of money than a dollar of bonds.

Policy implications of monetarism

Because monetary effects have variable lags of one to several quarters or more, countercyclical monetary policy actions are difficult to time properly. Friedman as well as Brunner and Meltzer argued that an active monetary policy, in the absence of an impossibly ideal foresight, tends to exacerbate, rather than smooth, economic fluctuations. In their view a stable monetary growth rate would avoid monetary sources of economic disturbances, and could be set to produce an approximately constant price level over the long run. Remaining instabilities in economic activity would be minor and, in any event, were beyond the capabilities of policy to prevent. A commitment by the monetary authorities to stable monetary growth would also help deflect constant political pressures for short-run monetary stimulus and would remove the uncertainty for investors of the unexpected effects of discretionary monetary policies.

A constant monetary growth policy can be contrasted with central bank practices that impart pro-cyclical variations to the money supply. It is common for central banks to lend freely to banks at times of rising credit demand in order to avoid increases in interest rates. Although such interest-rate targeting helps to stabilize financial markets, the targeting often fails to allow rates to change sufficiently to counter fluctuations in credit demands. By preventing interest rates from rising when credit demands increase, for example, the policy leads to monetary expansion that generates higher expenditures and inflationary pressures. Such mistakes of interest-rate targeting were clearly demonstrated in the 1970s, when for some time increases in nominal interest rates did not match increases in the inflation rate, and the resulting low rates of interest in real terms (that is, adjusted for inflation) overstimulated investment and aggregate demand.

The same accommodation of market demands for bank credit results from the common practice of targeting the volume of borrowing from the central bank. Attempts to keep this volume at some designated level require the central bank to supply reserves through open market operations as an alternative to borrowing by banks when rising market credit demands tighten bank reserve positions, and to withdraw reserves in the opposite situation. The resulting procyclical behaviour of the money supply could be avoided by operations designed to maintain a constant growth rate of money.

Brunner and Meltzer (1964a) developed an analytic framework describing how monetary policy should aim at certain intermediate targets as a way of influencing aggregate expenditures. The intermediate targets are such variables as the money supply or interest rates. (Since the Federal Reserve does not

control long-term interest rates or the money stock directly, it operates through instrumental variables, such as bank reserves or the federal funds rate, which it can affect directly.) The question of the appropriate intermediate targets of monetary policy soon became the most widely discussed issue in monetary policy.

In recognition of the deficiencies of interest-rate targeting, some countries turned during the 1970s to a modified monetary targeting in which annual growth ranges were announced and adhered to, though with frequent exceptions to allow for departures deemed appropriate because of disturbances from foreign trade and other sources. Major countries adopting some form of monetary targeting included the Federal Republic of Germany, Japan, and Switzerland, all of which kept inflation rates low and thus advertised by example the anti-inflationary virtues of monetarism. In the United States the Federal Reserve also began to set monetary target ranges during the 1970s but generally did not meet them and continued to target interest rates. In October 1979, when inflation was escalating sharply, the Federal Reserve announced a more stringent targeting procedure for reducing monetary growth. Although the average growth rate was reduced, the large short-run fluctuations in monetary growth were criticized by monetarists. In late 1982 the Federal Reserve relaxed its pursuit of monetary targets.

By the mid-1980s the US and numerous other countries were following a partial form of monetary targeting, in which relatively broad bands of annual growth rates are pursued but still subject to major departures when deemed appropriate. These policies are monetarist only in the sense that one or more monetary aggregates are an important indicator of policy objectives; they fall short of a firm commitment to a steady, let alone a non-inflationary, monetary growth rate.

Monetarist theory

Monetarist theory of aggregate expenditures is based on a demand function for monetary assets that is claimed to be stable in the sense that successive residual errors are generally offsetting and do not accumulate. Given the present inconvertible-money systems, the stock of money is treated as under the control of the government. Although a distinction is made in theory between the determinants of household and business holdings of money, money demand is usually formulated for households and applied to the total. In these formulations the demand for money depends on the volume of transactions, the fractions of income and of wealth the public wishes to hold in the form of money balances, and the opportunity costs of holding money rather than other income-producing assets (that is, the difference between yields on money and on alternative assets). The alternative assets are viewed broadly to include not only financial instruments but also such physical assets as durable consumer goods, real property, and business plant and equipment. The public is presumed to respond to changes in the amount of money supplied by undertaking transactions to bring actual holdings of both money and other assets into equilibrium with desired holdings. As a result of substitutions between money and assets, starting with close substitutes, yields change on a broad range of assets, including consumer durables and capital goods, in widening ripples that affect borrowing, investment, consumption, and production throughout the economy.

The end result is reflected in *aggregate* expenditures and the average level of prices. Independently of this monetary influence on aggregate expenditures and the price level, developments specific to particular sectors determine the distribution of expenditures among goods and services and relative prices. Thus monetarist theory rejects the common technique for forecasting aggregate output by adding

up the forecasts for individual industries or the common practice of explaining changes in the price level in terms of price changes for particular goods and services.

Monetarists were early critics of the once influential Keynesian theory of a highly elastic demand for money with respect to short-run changes in the interest rate on liquid short-term assets, which in extreme form became a ‘liquidity trap’. Empirical studies have found instead that interest rates on savings deposits and on short-term market securities have elasticities smaller even than the $-\frac{1}{2}$ implied by the simple Baumol–Tobin cash balance theory (Baumol, 1952; Tobin, 1956).

In empirical work a common form of the demand function for money includes one or two interest rates and real GNP as a proxy for real income. A gradual adjustment of actual to desired money balances is allowed for, implying that a full adjustment to a change in the stock is spread over several quarters. The lagged adjustment is subject to an alternative interpretation in which money demand reflects ‘permanent’ instead of current levels of income and interest rates. This interpretation de-emphasizes the volume of transactions as the major determinant of money demand in favour of the monetarist view of money as a capital asset yielding a stream of particular services and dependent on ‘permanent’ values of wealth, income, and interest rates (in most studies captured empirically by a lagged adjustment). Treatment of the demand for money as similar to demands for other assets stocks is now standard practice.

The monetarist view of money as a capital asset suggests that the demand for it depends on a variety of characteristics, and not uniquely on its transactions services. The definition of money for policy purposes depends on two considerations: the ability of the monetary authorities to control its quantity, and the empirical stability of a function describing the demand for it. In their study of the United States Friedman and Schwartz used an early version of M2, which included time and savings deposits at commercial banks, but they argued that minor changes in coverage would not greatly affect their findings. Subsequently the quantity of transaction balances M1 has become the most widely used definition of money for most countries, though many central banks claim to pay attention also to broader aggregates in conducting monetary policy.

In view of the wide range of assets into which the public may shift any excess money balances, the transmission of monetary changes through the economy to affect aggregate expenditures and other variables can follow a variety of paths. Monetarists doubt that these effects can be adequately captured by a detailed econometric model which prescribes a fixed transmission path. Instead they prefer models that dispense with detailed transmission paths and focus on a stable overall relationship between changes in money and in aggregate expenditures.

In both the monetarist model and large-scale econometric models, changes in the money stock are usually treated as exogenous (that is, as determined outside the model). It is clear that money approaches a strict exogeneity only in the long run. The US studies by Friedman and Schwartz and by Cagan established that the money supply not only influences economic activity but also is influenced by it in turn. This creates difficulties in testing empirically for the monetary effects on activity because allowance must be made for the feedback effect of economic activity on the money supply. Econometric models of the money supply can allow for feedback through the banking system (Brunner and Meltzer, 1964b). Under modern systems of inconvertible money, however, the feedback is dominated by monetary policies of the central banks, and attempts to model central bank behaviour have been less than satisfactory. Statistical tests of the exogeneity of the money supply using the Granger–Sims

methodology have given mixed results. Although the concurrent mutual interaction between money and economic activity remains difficult to disentangle, the longer the lag in monetary effects the less likely that the feedback from activity to money can account for the observed association. In the St Louis equation, for example, while the correlation between changes in GNP and in money concurrently could largely reflect feedback from GNP to money, the correlation between changes in GNP and lagged changes in money are less likely to be dominated by such feedback.

Opposition to monetary targeting

While monetarism has refocused attention on money and monetary policy, there is widespread doubt that velocity is sufficiently stable to make targeting of monetary growth desirable. Movements in velocity when monetary growth is held constant produce expansionary and contractionary effects on the economy. In the United States the trend of velocity was fairly stable and predictable from the early 1950s to the mid-1970s, but money demand equations based on that period showed large overpredictions after the mid-1970s (Judd and Scadding, 1982). Financial innovations providing new ways of making payments and close substitutes for holding money were changing the appropriate definition of money and the parameters of the demand function. In the United States the gradual removal of ceilings on interest rates banks could pay on deposits played a major role in these developments by increasing competition in banking. In Great Britain the removal of domestic controls over international financial transactions led to unusual movements in money holdings in 1979–80. Germany and Switzerland also found growing international capital inflows at certain times a disruptive influence on their monetary policies.

The ‘monetary theory of the balance of payments’ (Frenkel and Johnson, 1976) is an extension of monetarism to open economies where money supply and demand are interrelated among countries through international payments. A debated issue is whether individual countries, even under flexible exchange rates, can pursue largely independent monetary policies. The growing internationalization of capital markets is often cited as an argument against the monetarist presumption that velocity and the domestic money supply under flexible foreign exchange rates are largely independent of foreign influences.

Uncertainties over the proper definition of money and instability in the velocity of money as variously defined led to monetarist proposals to target the monetary liabilities of the central bank, that is, the ‘monetary base’ consisting of currency outstanding and bank reserves. The monetary base has the advantage of not being directly affected by market innovations and so of not needing redefinitions when innovations occur. Monetarists have proposed maintaining a constant growth rate of the base also because it would simplify – indirectly virtually eliminate – the monetary policy function of central banks and governments. Some of the European central banks have found targeting the monetary base preferable to targeting the money supply, though not without important discretionary departures from the target.

Yet financial market developments can also produce instabilities in the relationship between the monetary base and aggregate expenditures. Economists opposed to monetarism propose instead that stable growth of aggregate expenditures be the target of monetary policy and that it be pursued by making discretionary changes as deemed appropriate in growth of the base. This contrasts sharply with the monetarist opposition to discretion in the conduct of policy.

The Phillips curve trade-off

The inflationary outcome of discretionary monetary policy since the Second World War can be explained in terms of the Phillips curve trade-off between inflation and unemployment. Along the Phillips curve lower and lower unemployment levels are associated with higher and higher inflation rates. Such a relationship, first found in historical British data, was shown to fit US data for the 1950s and 1960s and earlier. The trade-off depends on sticky wages and prices. As aggregate demand increases, the rise in wages and prices trails behind, inducing an expansion of output to absorb part of the increase in demand. US experience initially suggested that any desired position on the Phillips curve could be maintained by the management of aggregate demand. Thus a lower rate of unemployment could be achieved and maintained by tolerating an associated higher rate of inflation. Given this presumed trade-off, policymakers tended to favour lower unemployment at the cost of higher inflation. In the 1970s, however, the Phillips curve shifted towards higher rates of inflation for given levels of unemployment. Friedman (1968) argued that the economy gravitates toward a ‘natural rate of unemployment’ which in the long run is largely independent of the inflation rate and cannot be changed by monetary policy. Wages and prices adjust sluggishly to unanticipated changes in aggregate demand but adjust more rapidly to maintained increases in demand and prices that are anticipated. Consequently, the only way to hold unemployment below the natural rate is to keep aggregate demand rising faster than the anticipated rate of inflation. Since the anticipated rate tends to follow the actual rate upward, this leads to faster and faster inflation. This ‘acceleration principle’ implies that there is no permanent trade-off between inflation and unemployment. The existence of a natural rate of unemployment also implies that price stability does not lead to higher unemployment in the long run.

Monetarist thought puts primary emphasis on the long-run consequences of policy actions and procedures. It rejects attempts to reduce short-run fluctuations in interest rates and economic activity as usually beyond the capabilities of monetary policy and as generally inimical to the otherwise achievable goals of long-run price stability and maximum economic growth. Monetarists believe that economic activity, apart from monetary disturbances, is inherently stable. Much of their disagreement with Keynesians can be traced to this issue.

Rational expectations

One version of the rational expectations theory goes beyond monetarism by contending that there is little or no Phillips curve trade-off between inflation and unemployment even in the short run, since markets are allegedly able to anticipate any systematic countercyclical policy pursued to stabilize the economy. Only unanticipated departures from such stabilization policies affect output; all anticipated monetary changes are fully absorbed by price changes. Since unsystematic policies would have little countercyclical effectiveness or purpose, the best policy is to minimize uncertainty with a predictable monetary growth.

This theory shares the monetarist view that unpredictable fluctuations in monetary growth are an undesirable source of uncertainty with little benefit. But the two views disagree on the speed of price adjustments to predictable monetary measures and on the associated effects on economic activity.

Monetarists do not claim that countercyclical policies have no real effects, but they are sceptical of our ability to use them effectively. It is the ill-timing of countercyclical policies as a result of variable lags in monetary effects that underlies the monetarist preference for constant monetary growth to avoid uncertainty and inflation bias.

Interest in private money supplies

Monetarism is the fountainhead of a renewed interest in a subject neglected during the Keynesian Revolution: the design of monetary systems that maintain price-level stability. Scepticism that price-level stability can be achieved even by a constant growth rate of money however defined or of the monetary base has led to proposals for a strict gold standard or for a monetary system in which money is supplied by the private sector under competitive pressures to maintain a stable value. While monetarists are sympathetic to proposals to eliminate discretionary monetary policies, they view such alternative systems as impractical and believe that a non-discretionary government policy of constant monetary growth is the best policy.

Associated views of the Monetarist School

Monetarism is associated with various related attitudes towards government (see Mayer, 1978). Monetarism shares with laissez-faire a belief in the long-run benefits of a competitive economic system and of limited government intervention in the economy. It opposes constraints on the free flow of credit and on movements of interest rates, such as the US ceilings on deposit interest rates (removed by the mid-1980s except on demand deposits). The disruptive potential of such ceilings became evident in the 1970s when financial innovations, partly undertaken to circumvent the ceilings, produced the transitional shifts in the traditional money-demand functions that created difficulties for the conduct of monetary policy. Government control over the quantity of money is viewed as a justifiable exception to laissez-faire, however, in order to ensure the stability of the value of money.

See Also

- Friedman, Milton
- Keynesianism
- monetary policy, history of
- new classical macroeconomics
- quantity theory of money
- rational expectations

Bibliography

Andersen, L.C. and Jordan, J.L. 1968. Monetary and fiscal actions: a test of their relative importance in economic stabilization. *Federal Reserve Bank of St Louis Review* 50(November), 11–24.

Angell, J. 1933. Monetary control and general business stabilization. In *Economic Essays in Honour of Gustav Cassel*. London: Allen and Unwin.

Baumol, W.J. 1952. The transactions demand for cash: an inventory theoretic approach. *Quarterly Journal of Economics* 66, 545–56.

Brunner, K. 1968. The role of money and monetary policy. *Federal Reserve Bank of St Louis Review* 50 (July), 8–24.

Brunner, K. and Meltzer, A. 1964a. The Federal Reserve's attachment to the free reserve concept. U.S. Congress House Committee on Banking and Currency, Subcommittee on Domestic Finance, April.

Brunner, K. and Meltzer, A. 1964b. Some further investigations of demand and supply functions for money. *Journal of Finance* 19, 240–83.

Brunner, K. and Meltzer, A. 1976. An aggregative theory for a closed economy. In *Studies in Monetarism*, ed. J. Stein. Amsterdam: North-Holland.

Cagan, P. 1965. *Determinants and Effects of Changes in the Stock of Money 1875–1960*. New York: Columbia University Press for the NBER.

Frenkel, J.A. and Johnson, H.G., eds. 1976. *The Monetary Approach to the Balance of Payments*. Toronto: University of Toronto Press.

Friedman, M. 1948. A monetary and fiscal framework for economic stability. *American Economic Review* 38, 256–64.

Friedman, M. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Friedman, M. and Meiselman, D. 1963. The relative stability of monetary velocity and the investment multiplier in the United States, 1897–1958. In Commission on Money and Credit, *Stabilization Policies*. Englewood Cliffs, NJ: Prentice–Hall.

Friedman, M. and Schwartz, A.J. 1963a. Money and business cycles. *Review of Economics and Statistics* 45(1), Part II, Supplement, 32–64.

Friedman, M. and Schwartz, A. 1963b. *A Monetary History of the United States 1867–1960*. Princeton: Princeton University Press for the NBER.

Friedman, M. and Schwartz, A. 1970. *Monetary Statistics of the United States Estimates, Sources, Methods*. New York: NBER.

Friedman, M. and Schwartz, A. 1982. *Monetary Trends in the United States and the United Kingdom Their Relation to Income, Prices and Interest Rates, 1867–1975*. Chicago: University of Chicago Press.

Judd, J.P. and Scadding, J.L. 1982. The search for a stable money demand function: a survey of the post-1973 literature. *Journal of Economic Literature* 20, 993–1023.

Laidler, D. 1978. Mayer on monetarism: comments from a British point of view. In *The Structure of Monetarism*, ed. T. Mayer. New York: Norton.

Mayer, T., ed. 1978. *The Structure of Monetarism*. New York: Norton.

Mints, L.W. 1950. *Monetary Policy for a Competitive Society*. New York: McGraw-Hill.

Simons, H. 1936. Rules versus authorities in monetary policy. *Journal of Political Economy* 44 (February), 1–30.

Tobin, J. 1956. The interest elasticity of transactions demand for cash. *Review of Economics and Statistics* 38(August), 241–7.

Warburton, C. 1946. The misplaced emphasis in contemporary business-fluctuation theory. *Journal of Business* 19, 199–220.

How to cite this article

Cagan, Phillip. "monetarism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000191> doi:10.1057/9780230226203.1115

The New Palgrave Dictionary of Economics Online

monetary aggregation

William A. Barnett

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Aggregation theory and index-number theory provide the foundations for official governmental data. However, the monetary quantity aggregates and interest rate aggregates supplied by many central banks are not based on index-number or aggregation theory, but rather are the simple unweighted sums of the component quantities and the quantity-weighted or unweighted arithmetic averages of interest rates. The result has been instability of estimated money demand and supply functions, and a series of ‘puzzles’ in the related applied literature. In contrast, the Divisia monetary aggregates are derived directly from economic index-number theory.

Keywords

aggregation theory; Barnett critique; data construction; equity premium puzzle; European Central Bank; Divisia index; index number theory; inflation targeting; interest rate targeting; intermediate targets; monetary aggregation; monetary policy; monetary quantity targeting; monetary targeting; monetary velocity; money demand; money supply

Article

Aggregation theory and index-number theory have been used to generate official governmental data since the 1920s. One exception still exists. The monetary quantity aggregates and interest rate aggregates supplied by many central banks are not based on index-number or aggregation theory, but rather are the simple unweighted sums of the component quantities and quantity-weighted or arithmetic averages of interest rates. The predictable consequence has been induced instability of money demand and supply functions, and a series of ‘puzzles’ in the resulting applied literature. In contrast, the Divisia monetary aggregates, originated by Barnett (1980), are derived directly from economic index-number theory. Financial aggregation and index number theory was first rigorously connected with the literature on microeconomic aggregation and index number theory by Barnett (1980; 1987). A collection of many

of his contributions to that field is available in Barnett and Serletis (2000).

Data construction and measurement procedures imply the theory that can rationalize the procedure. The assumptions implicit in the data construction procedures must be consistent with the assumptions made in producing the models within which the data are to be used. Unless the theory is internally consistent, the data and its applications are incoherent. Without that coherence between aggregator function structure and the econometric models within which aggregates are embedded, stable structure can appear to be unstable. This phenomenon has been called the 'Barnett critique' by Chrystal and MacDonald (1994).

Aggregation theory versus index number theory

The exact aggregates of microeconomic aggregation theory depend on unknown aggregator functions, which typically are utility, production, cost, or distance functions. Such functions must first be econometrically estimated. Hence the resulting exact quantity and price indexes become estimator- and specification-dependent. This dependency is troublesome to governmental agencies, which therefore view aggregation theory as a research tool rather than a data construction procedure.

Statistical index-number theory, on the other hand, provides indexes which are computable directly from quantity and price data, without estimation of unknown parameters. Such index numbers depend jointly on prices and quantities, but not on unknown parameters. In a sense, index number theory trades joint dependency on prices and quantities for dependence on unknown parameters. Examples of such statistical index numbers are the Laspeyres, Paasche, Divisia, Fisher ideal, and Törnqvist indexes.

The loose link between index number theory and aggregation theory was tightened, when Diewert (1976) defined the class of second-order 'superlative' index numbers. Statistical index number theory became part of microeconomic theory, as economic aggregation theory had been for decades, with statistical index numbers judged by their nonparametric tracking ability to the aggregator functions of aggregation theory.

For decades, the link between statistical index number theory and microeconomic aggregation theory was weaker for aggregating over monetary quantities than for aggregating over other goods and asset quantities. Once monetary assets began yielding interest, monetary assets became imperfect substitutes for each other, and the 'price' of monetary-asset services was no longer clearly defined. That problem was solved by Barnett (1978; 1980), who derived the formula for the user cost of demanded monetary services. Subsequently Barnett (1987) derived the formula for the user cost of supplied monetary services. A regulatory wedge can exist between the demand and supply-side user costs if non-payment of interest on required reserves imposes an implicit tax on banks.

Barnett's results on the user cost of the services of monetary assets set the stage for introducing index number theory into monetary economics.

The economic decision

Consider a decision problem over monetary assets that illustrates the capability of monetary aggregation theory. The decision problem will be defined so that the relevant literature on economic aggregation over goods is immediately applicable. Initially we shall assume perfect certainty.

Let $\mathbf{m}_t' = (m_{1t}, m_{2t}, \ldots m_{nt})$ be the vector of real balances of monetary assets during period *t*, let $\boldsymbol{r}_t$ be the vector of nominal holding-period yields for monetary assets during period *t*, and let R_t be the one-period holding yield on the benchmark asset during period *t*. The benchmark asset is defined to be a pure investment that provides no services other than its yield, R_t, so that the asset is held solely to accumulate wealth. Thus, R_t is the maximum holding period yield in the economy in period *t*.

Let y_t be the real value of total budgeted expenditure on monetary services during period *t*. Under simplifying assumptions for data within one country, the conversion between nominal and real expenditure on the monetary services of one or more assets is accomplished using the true cost of living index on consumer goods. But for multi-country data or data aggregated across heterogeneous regions, the correct deflator can be found in Barnett (2003; 2007). The optimal portfolio allocation decision is:

$$\text{maximize } u(\mathbf{m}_t) \tag{1}$$

$$\text{subject to } \pi_t' \mathbf{m}_t = y_t,$$

where $\pi_t' = (\pi_{1t}, \ldots, \pi_{nt})$ is the vector of monetary-asset real user costs, with

$$\pi_{it} = \frac{R_t - r_{it}}{1 + R_t}. \tag{2}$$

This function *u* is the decision maker's utility function, assumed to be monotonically increasing and strictly concave. The user cost formula (2), derived by Barnett (1978; 1980), measures the forgone interest or opportunity cost of holding monetary asset *i*, when the higher yielding benchmark asset could have been held.

To be an admissible quantity aggregator function, the function *u* must be weakly separable within the consumer's complete utility function over all goods and services. Producing a reliable test for weak separability is the subject of much intensive research by an international group of econometricians (see, for example, Jones, Dutkowsky and Elger, 2005; Fleissig and Whitney, 2003; De Peretti, 2005). Two

approaches exist. One approach uses stochastic extensions of nonparametric revealed preference tests, while the other uses parametric econometric models.

Let $\mathbf{m}_t^*$ be derived by solving decision (1). Under the assumption of linearly homogeneous utility, the exact monetary aggregate of economic theory is the utility level associated with holding the portfolio, and hence is the optimized value of the decision's objective function:

$$M_t = u(\mathbf{m}_t^*). \tag{3}$$

The Divisia index

Although equation (3) is exactly correct, it depends upon the unknown function, *u*. Nevertheless, statistical index-number theory enables us to track M_t exactly without estimating the unknown function, *u*. In continuous time, the exact monetary aggregate, $M_t = u(\mathbf{m}_t^*)$, can be tracked exactly by the Divisia index, which solves the differential equation

$$\frac{d\log M_t}{dt} = \sum_i s_{it}\frac{d\log m_{it}^*}{dt} \tag{4}$$

for M_t, where

$$s_{it} = \frac{\pi_{it} m_{it}^*}{Y_t}$$

is the *i*'th asset's share in expenditure on the total portfolio's service flow. In equation (4), it is understood that the result is in continuous time, so the time subscripts are a shorthand for functions of time. We use *t* to be the time period in discrete time, but the instant of time in continuous time. The dual user cost price aggregate $\Pi_t = \Pi(\pi_t)$, can be tracked exactly by the Divisia price index, which solves the differential equation

$$\frac{d\log\Pi_t}{dt} = \sum_i s_{it}\frac{d\log\pi_{it}}{dt}.$$
(5)

The user cost dual satisfies Fisher's factor reversal in continuous time:

$$\Pi_t M_t = \pi_t' \mathbf{m}_t.$$
(6)

As a formula for aggregating over quantities of perishable consumer goods, that index was first proposed by François Divisia (1925) with market prices of those goods inserted in place of the user costs in equation (4). In continuous time, the Divisia index, under conventional neoclassical assumptions, is exact. In discrete time, the Törnqvist approximation is:

$$\log M_t - \log M_{t-1} = \sum_i \bar{s}_{it}(\log m_{it}^* - \log m_{i,t-1}^*),$$
(7)

where

$$\bar{s}_{it} = \frac{1}{2}(s_{it} + s_{i,t-1}).$$

In discrete time, we often call equation (7) simply the Divisia quantity index. After the quantity index is computed from (7), the user cost aggregate most commonly is computed directly from equation (6). Diewert (1976) defines a ‘superlative index number’ to be one that is exactly correct for a quadratic approximation to the aggregator function. The discretization (7) to the Divisia index is in the superlative class, since it is exact for the quadratic translog specification to an aggregator function. With weekly or monthly monetary data, Barnett (1980) has shown that the Divisia index growth rates, (7), are accurate to within three decimal places. In addition, the difference between the Fisher ideal index and the discrete Divisia index growth rates are third order and comparably small. That third-order differential error

typically is smaller than the round-off error in the component data.

Prior applications

Divisia monetary aggregates were first constructed for the United States by Barnett (1980), when he was on the staff of the Special Studies Section of the Board of Governors of the Federal Reserve System, and are now maintained by the Federal Reserve Bank of Saint Louis in its data base, called FRED (see Anderson, Jones and Nesmith, 1997, who produced the Divisia data for FRED). A Divisia monetary-aggregates data base also has been produced for the United Kingdom by the Bank of England. An overview of Divisia data maintained by many central banks throughout the world can be found in Belongia and Binner (2000; 2005) and in Barnett, Fisher and Serletis (1992), along with a survey of empirical results with that data. The most extensive collection of relevant applied and theoretical research in that area is in Barnett and Serletis (2000) and Barnett and Binner (2004).

The state of the art

The European Central Bank is implementing a multilateral extension of the Divisia monetary aggregates for monetary quantity and interest rate aggregation within the euro area. This aggregation is multilateral in the recursive sense that it permits aggregation of monetary service flows first within countries, then over countries. The resulting aggregation will be in a strictly nested, internally consistent manner. The multilateral extension of the theory was produced by Barnett (2003; 2007). This extension was produced under three increasingly strong sets of assumptions: (*a*) with the weakest being produced from heterogeneous agents theory, (*b*) followed by the somewhat stronger assumption of existence of a multilateral representative agent, and (*c*) finally with the strongest being the assumption of the existence of a unilateral representative agent. The intent is to move from the weakest towards the strongest assumptions, as progress is made within the European Monetary Union towards its harmonization and economic convergence goals. Since Barnett's three assumption structures are nested, construction of the data under the most general heterogeneous countries approach would continue to be valid, as the stronger assumptions become more reasonable and are attained within the euro area.

Extension of index number theory to the case of risk was introduced by Barnett, Liu and Jensen (2000), who derived the extended theory from Euler equations rather than from the perfect-certainty first-order conditions used in the earlier index number-theory literature. Since that extension is based upon the consumption capital-asset-pricing model (CCAPM), the extension is subject to the ‘equity premium puzzle’ of smaller than necessary adjustment for risk. We believe that the under-correction produced by CCAPM results from its assumption of intertemporal blockwise strong separability of goods and services within preferences. Barnett and Wu (2005) have extended Barnett, Liu, and Jensen's result to the case of risk aversion with intertemporally non-separable tastes.

The extension to risk is likely to be especially important to countries whose residents hold significant deposits in foreign denominated assets, since exchange-rate risk can cause rates of return on monetary assets to be subject to non-negligible risk. With the recent trend towards financial integration in many parts of the world, exchange-rate risk is likely to grow in importance in monetary aggregation. In many countries, the largest holder of foreign-denominated deposits is the central bank itself. Within the United

States, the extension to risk is highly relevant to the so called ‘missing M2’ episode of the early 1990s, when substitutability among small time deposits, stock funds, and bond funds produced ‘puzzles’.
User cost aggregates are duals to monetary quantity aggregates. Either implies the other uniquely. In addition, user-cost aggregates imply the corresponding interest-rate aggregates uniquely. The interest-rate aggregate r_t implied by the user-cost aggregate Π_t is the solution for r_t to the equation:

$$\frac{R_t - r_t}{1 + R_t} = \Pi_t.$$

Accordingly, any monetary policy that operates through the opportunity cost of money (that is, interest rates) has a dual policy operating through the monetary quantity aggregate, and vice versa. Aggregation theory implies no preference for either of the two dual policy procedures or for any other approach to policy, so long as the policy does not violate principles of aggregation theory.

Conclusion

Aggregation theory is about measurement, and has little, if anything, to say about the choice of policy instrument, such as the funds rate or the base. But accurate measurement, through proper application of aggregation theory, has much to say about the transmission of policy, modelling of structure, and the measurement of intermediate targets (if any) and final targets.
Policies that violate aggregation theoretic principles include the following oversimplified approaches: (*a)* inflation targeting that targets one arbitrary consumer-good price as a final target, while ignoring all other consumer goods prices, rather than targeting the true cost-of-living index over all consumer goods prices; (*b*) interest rate targeting that analogously targets one arbitrary interest rate as an intermediate target while ignoring all other interest rates, rather than targeting the aggregation-theoretic interest-rate or user-cost aggregate over a weakly separable collection of monetary assets; (*c*) monetary quantity targeting that targets a simple-sum monetary aggregate as an intermediate target rather than the aggregator function over a weakly separable collection of monetary assets; and (*d*) policy simulations using money-demand or money-supply functions containing simple-sum monetary aggregates or quantity-weighted interest-rate aggregates. The measurement defects in the above four cases are unrelated to the choice of the funds rate or monetary base as an instrument of policy. Unlike intermediate targets, final targets, and variables in models, the chosen instruments of policy tend to be highly controllable, disaggregated variables, presenting few serious measurement problems.
The objective of the Divisia monetary aggregates is measurement of the economy's monetary service flow and its dual opportunity cost (user cost) and implied interest rate aggregate, not advocacy of any particular policy use of the correctly measured variables. But all uses of data are adversely affected by improper measurement, and a long series of ‘puzzles’ in monetary economics have been shown to have been produced by improper measurement (see, for example, Barnett and Serletis, 2000, ch. 24).

See Also

- European Central Bank
- Federal Reserve System
- foreign direct investment
- inflation targeting
- measurement
- monetary economics, history of
- monetary and fiscal policy overview
- statistics and economics

Bibliography

Anderson, R., Jones, B. and Nesmith, T. 1997. Building new monetary services indexes: concepts, data and methods. *Federal Reserve Bank of St Louis Review* 79, 53–82.

Barnett, W. 1978. The user cost of money. *Economics Letters* 1, 145–49. Repinted in Barnett and Serletis (2000, ch. 1).

Barnett, W. 1980. Economic monetary aggregates: an application of aggregation and index number theory. *Journal of Econometrics* 14, 11–48. Reprinted in Barnett and Serletis (2000, ch. 2).

Barnett, W. 1987. The microeconomic theory of monetary aggregation. In *New Approaches in Monetary Economics*, ed. W. Barnett and K. Singleton. Cambridge: Cambridge University Press. Reprinted in Barnett and Serletis (2000, ch. 3).

Barnett, W. 2003. Aggregation-theoretic monetary aggregation over the euro area, when countries are heterogeneous. Working Paper No. 260. Frankfurt: European Central Bank.

Barnett, W. 2007. Multilateral aggregation-theoretic monetary aggregation over heterogeneous countries. *Journal of Econometrics* 136, 457–82.

Barnett, W. and Binner, J. 2004. *Functional Structure and Approximation in Econometrics*. Amsterdam: North-Holland.

Barnett, W., Fisher, D. and Serletis, A. 1992. Consumer theory and the demand for money. *Journal of Economic Literature* 30, 2086–119. Reprinted in Barnett and Serletis (2000, ch. 18).

Barnett, W., Liu, Y. and Jensen, M. 2000. CAPM risk adjustment for exact aggregation over financial assets. *Macroeconomic Dynamics* 1, 485–512.

Barnett, W. and Serletis, A., eds. 2000. *The Theory of Monetary Aggregation*. Amsterdam: North-Holland.

Barnett, W. and Wu, S. 2005. On user costs of risky monetary assets. *Annals of Finance* 1, 35–50.

Belongia, M. and Binner, J. 2000. *Divisia Monetary Aggregates: Theory and Practice*. Basingstoke: Palgrave.

Belongia, M. and Binner, J. 2005. *Money, Measurement, and Computation*. Basingstoke: Palgrave.

Chrystal, A. and MacDonald, R. 1994. Empirical evidence on the recent behaviour and usefulness of simple-sum and weighted measures of the money stock. *Federal Reserve Bank of St. Louis Review* 76, 73–109.

De Peretti, P. 2005. Testing the significance of the departures from utility maximization. *Macroeconomic Dynamics* 9(3), 373–97.

Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4, 115–45.

Divisia, F. 1925. L'Indice monétaire et la théorie de la monnaie. *Revue d'Economie Politique* 39, 980–1008.

Fleissig, A. and Whitney, G. 2003. A new PC-based test for Varian's weak separability conditions. *Journal of Business and Economic Statistics* 21, 133–44.

Jones, B., Dutkowsky, D. and Elger, T. 2005. Sweep programs and optimal monetary aggregation. *Journal of Banking and Finance* 29, 483–508.

How to cite this article

Barnett, William A. "monetary aggregation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000351> doi:10.1057/9780230226203.1116

The New Palgrave Dictionary of Economics Online

monetary and fiscal policy overview

Narayana R. Kocherlakota
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides an overview of economic thinking about monetary and fiscal policy. I discuss the methodology of answering policy questions in macroeconomics. I then explain what is known from using this methodology for positive and normative analyses of monetary and fiscal policy. Finally, I describe the challenges associated with endogenizing monetary and fiscal policy.

Keywords

adjustment costs; commitment; computational experimentation; deflation; endogenous versus exogenous variables; game theory; government budget constraint; implementability; incentive compatibility; inflationary expectations; lump-sum taxes; Markov-perfect equilibria; methodology of economics; Mirrlees approach to optimal taxation; monetary versus fiscal policy; multiple equilibria; optimal taxation; Ramsey approach to optimal taxation; rational expectations; recursive equilibria; Ricardian equivalence; social insurance; sticky prices; sunspot equilibrium; Taylor rule; wedges

Article

In this article I provide an overview of economic thinking about monetary and fiscal policy. There are three terms that need to be defined in this sentence: policy, monetary, and fiscal. I begin by defining each in turn.

A government's *policy* is akin to a strategy in game theory. It specifies a function at each date that maps the government's information at that date into the government's actions. This information typically takes two forms. First, it includes *endogenous* variables such as past prices, past quantities or past actions of the government. For example, under the famous Taylor rule, a government's choice of current short-term interest rates is based on past observations of the consumer price index and gross domestic product. Second, the government's information includes *exogenous* variables, like the realizations of shocks to productivity or to money demand.

These sources of information may be public or they may be known only to the government. Thus, in the United States the Federal Reserve collects information about the state of the economy that it uses for making decisions but is kept confidential from households in the economy. Note, too, that the government's actions themselves may be private information to the government; for example, until recently, the Federal Open Market Committee publicly announced its decisions only with a lag.

In the popular press, the term ‘policy’ is commonly used in a different way, to refer only to the *current* choice of the government. However, as long as *some* economic actors (firms, households or the government itself) are forward-looking, such a specification of policy is intrinsically incomplete. Forward-looking decision-makers need to know not just the government's choice of policy today but also how the government will respond to new information in the future. (This is true even if these forward-looking actors have expectations that are far from rational.) Thus, if the government raises taxes today, my response to that increase depends crucially on whether I believe it will persist for a long time. To make that judgement, I need to know not just the government's choices today but also how its choices in the future depend on new information that the government receives.

Whether a policy is monetary or fiscal or neither depends on the nature of the actions specified by that policy. A policy is said to be *monetary* if the relevant actions are those generally undertaken by a central bank. These may include the size of monetary injections, reserve requirements, the discount rate, or the scale of interventions in bond or foreign exchange markets. A policy is said to be *fiscal* if the relevant actions are tax rates and/or expenditures on various commodities. Of course, many government policies (should Iran be invaded or not?) are neither fiscal nor monetary.

In the body of this article, I discuss several lessons from the study of monetary and fiscal policy. Before doing so, though, it is useful to understand the methodology that was used to learn those lessons (see Lucas, 1980, and Prescott, 2005, for a fuller discussion of this methodology). Any analysis of policy starts with the following question: on the assumption that no other exogenous variables change, how does the economy respond to a change in policy? This kind of question is really asking about the outcome of a controlled *experiment*. It would be best answered by constructing giant national or super-national laboratories in order to conduct these experiments. But it is clearly impossible to perform controlled experiments of this kind. How then do macroeconomists proceed?

The approach taken by macroeconomists is closely related to the methods used by other non-experimental sciences. Consider for example the issue of global warming. There have been no prior episodes in world history in which man has been able to generate such a large amount of CO_2 in such a short period of time. Hence, there is no way to use prior data to understand the impact of this build-up on climatic variables like temperature. Instead, climatologists rely on computer simulations of abstract models to understand the impact of greenhouse gases on the world's climate.

Similarly, macroeconomists build abstract computational models to answer questions about the impact of monetary and fiscal policy. It is well-understood from many years of computational experimentation that useful models must have certain elements to provide reliable answers to policy questions. The models need to be both dynamic and stochastic in nature. The models need to be explicit about aggregate resource constraints: the amount of goods consumed by governments and households cannot exceed the amount of goods produced. The models should feature households with well-defined objectives and budget constraints. The households and firms in the models should be forward-looking (although they may or may not be fully informed about the state of the economy).

To provide a quantitative answer about the impact of a particular policy, macroeconomists need to be

specific about many other elements of the computational model (preferences of households, shocks hitting the economy, and so on). Again, it is useful to refer to the natural sciences as a way to understand how macroeconomists proceed. Consider a biologist that wants to understand the impact of a new drug on human beings. At least initially, she experiments on animals. For some kinds of drugs, she may use mice. For others, she may use more expensive animals like monkeys or dogs. Her decision about which proxy to use is a complex one, grounded in theory, collective prior experience about other drugs and these animals, and individual judgement.

In the same fashion, macroeconomists do not use the same model for all policy questions. Instead, they choose the model based on the question at hand. Thus, for questions concerning the short-run impact of monetary policies, they may include adjustment costs in physical capital and/or prices. For other questions concerning the long-run impact of monetary policy, they may neglect these elements. Like the biologist, their decisions are based on theory, collective prior experience and judgement.

One aspect of this decision-making that receives particular attention in macroeconomics is how to quantify the various elements of the model. How risk-averse are the households in the model economy? What is the elasticity of substitution between capital and labour in the model economy? Fortunately, for many of these parameter choices, there is a profession-wide consensus, informed by many years of experience and discussion. For other parameters, new choices have to be made. Generally, macroeconomists use a mix of information from both microeconomic and macroeconomic sources to make these choices. There may well be a range of plausible choices for a given parameter, and then the answer to the policy question under consideration is really a set, not a single point.

In the remainder of this article, I discuss some of the conclusions about monetary and fiscal policy that macroeconomists have reached from using this methodology. I focus on results that are highly robust, in the sense that they occur across a wide class of models. I begin by looking at lessons from the *positive* approach to policy, which studies the response of the private sector to different specifications of policy. I then look at lessons from the *normative* approach, which looks at properties of *ex ante* optimal policies. Finally, I discuss some difficulties associated with modelling policy choices as being an endogenous response to economic conditions.

The positive approach to policy

There is a large amount of macroeconomic research that treats monetary and fiscal policy as wholly exogenous to the economy. It asks questions of the sort: how does some aspect of private sector economic behaviour respond to a given specification of monetary and fiscal policy? Macroeconomists have described the outcomes to many specific experiments of this kind. There is no useful way to summarize this knowledge. However, there are several general lessons that one can draw from this research. In what follows, I discuss three of these.

Lesson 1. fiscal vs. monetary policy

I have drawn a distinction between fiscal and monetary policy. However, this distinction is more than a little artificial for two reasons. First, in macroeconomic models households face budget constraints and aggregate resource constraints are satisfied. Together, these imply that the government itself must satisfy a budget constraint in equilibrium: the present value of the government's revenues must equal the present

value of its expenditures. (There are overlapping-generations model economies in which this restriction need not be satisfied. However, these models are typically not thought to be empirically relevant; Abel et al., 1989.) This constraint implies a sharp linkage between fiscal and monetary policy. Changes in monetary policies affect the government's revenue from money creation. Hence, the two types of policies are inextricably linked, because they cannot be changed separately. (This fundamental linkage between fiscal and monetary policy was made especially clear by Sargent and Wallace, 1981.)

The second reason is that, in terms of its impact on the economy, monetary policy is merely fiscal policy by another name. People and firms who hold money are forgoing the interest that they could receive by holding bonds instead. They hold that money because it helps them buy goods and services that are difficult to purchase using bonds. Higher interest rates makes money more costly to hold, and makes those goods and services more costly to buy. The interest rate acts like a sales tax on those goods and services.

Monetary policy has still other distorting effects on the economy when some prices are more flexible than others. For example, suppose nominal wages do not respond rapidly to changes in inflation, but gas prices do. Then, the relative price of labour and gasoline may vary in response to variations in monetary instruments. Again, though, a particular kind of fiscal policy – variations in the gasoline tax – can affect the economy in exactly the same way. (This equivalence between fiscal and monetary policy is stressed by Correia, Nicolini and Teles, 2004.)

Lesson 2. Ricardian equivalence

I pointed out above that the present value of government expenditures must equal the present value of government revenues. This simple fact has surprising consequences. Consider two policies with the same government purchases. Suppose one policy generates lower tax revenue in the next ten years than the other policy. Obviously, under the first policy, the government must borrow more. This extra demand in loans puts upward pressure on interest rates.

However, the government's intertemporal budget constraint also implies that the first policy must necessarily generate *higher* tax revenue in the future. Forward-looking households anticipate this increase in their future tax burden. They respond by saving more to meet this tax burden. In a classic paper, Barro (1974) shows that, if households are sufficiently forward-looking, and markets are frictionless, then the households’ extra demand for savings under the first policy is exactly equal to the government's extra demand for loans. Hence, even though the government is borrowing more, there is no extra pressure on interest rates; they should be the same under the two policies. This result is generally termed *Ricardian equivalence* (because of some antecedents in the work of David Ricardo).

The exact Ricardian equivalence result is not robust to adding plausible frictions like borrowing constraints on households. Nonetheless, there is a qualitative lesson that holds much more generally and is often forgotten in policy discussions: economics does not predict a stable relationship between current government debt or deficits and interest rates.

Lesson 3. Expectations matter

I have emphasized above that households’ expectations about future government actions matter for

current outcomes. However, in many macroeconomic models a given household's behaviour depends also on its expectations of other households' current and future actions. This feedback generates the possibility of multiple equilibrium outcomes for a given government policy.

Here's a simple example of this phenomenon. Suppose both government investment and household labour are necessary inputs into production – that is, either zero government investment or zero labour input leads to zero output. Suppose as well that the government collects resources to fund its investment by taxing output. In such a world, regardless of the government's policy, there is always an equilibrium in which households do not work at all. In this equilibrium, because other households are not working, a given household realizes that the government cannot fund any investment. Hence, it is individually optimal for that household not to provide any labour input.

This kind of multiplicity leads to the possibility of what are called *sunspot* fluctuations in macroeconomic variables. The idea here is that households use some arbitrary random variable to coordinate their behavior. Thus, if they all see rain in Peoria, they decide not to work. If they see sun in Peoria, they decide to work. Whether it is sunny or not in Peoria, of course, is irrelevant for economic fundamentals – but in this economy, this variable can still affect equilibrium outcomes. (For early expositions of the concept of sunspot equilibria, see Azariadis, 1981, and Cass and Shell, 1983.)

Note that this example is only an illustration of a much more general phenomenon. It is especially prevalent in monetary economies. In these settings, a household's decision about how many real balances to hold today depends crucially on the household's expectations about future inflation rates. Obstfeld and Rogoff (1983) demonstrate how this intertemporal feedback can generate a continuum of welfare-indexed possible inflation paths as equilibria, even if the money supply is fixed. Sargent and Wallace (1975) demonstrate how this intertemporal feedback can generate a continuum of welfare-indexed possible inflation paths as equilibria even if interest rates are fixed. (Pareto-ranked equilibria do not occur in all economies. In many economies – especially non-monetary ones – it may be possible to prove that any equilibrium allocation solves a maximization problem in which the objective is a weighted average of households' utilities. In such settings, equilibrium allocations are necessarily Pareto non-comparable. Without such a proof in hand, though, one has to be aware that there is the potential for sunspot fluctuations between Pareto-ranked outcomes. Many macroeconomists restrict attention to so-called recursive equilibria or Markov-perfect equilibria. Under these notions of equilibrium, outcomes have the property that they depend on the past only through a small number of state variables. This restriction is undoubtedly useful for simplifying computational or econometric work. However, the restriction may inadvertently rule out important sources of potential multiplicity. See, for example, Woodford's, 1994, analysis of Lucas and Stokey's, 1987, model economy.)

The normative approach to policy

I now turn to the second approach to studying macroeconomic policy. This approach posits a government that chooses a policy at the beginning of time; its objective is to maximize some weighted average of household utilities. Crucially, the government is able to commit to never change the policy. This kind of commitment power is clearly artificial; the goal of the second approach is to tell us what kinds of policies maximize *ex ante* social welfare, not what policies are actually adopted by governments. By construction, there is no requirement that the optimal policies be realistic: normative

analyses tell us what the government should do, not what they actually do. Thus, economists use normative analyses to argue strongly in favour of free trade, which is a policy that has never been followed by any country at any time.

Everything in this approach hinges on what is assumed about the set of instruments available to the government. It is well-known that *lump-sum* taxes are a highly desirable taxation instrument. A lump-sum tax is a tax on a household or firm which is independent of their actions. Such a tax is desirable because it does not distort the choices of the household or the firm.

But lump-sum taxes are typically not used by governments. Once one notices this fact, there are at least two ways to proceed in thinking about optimal taxes. One can assume that the governments can only use a limited set of tax instruments that does not include lump-sum taxes. This approach is generally called the *Ramsey* approach. Alternatively, one can build model economies in which governments have access to all possible tax instruments, but *choose*, because of a particular private information friction, not to use lump-sum taxes. This approach is generally called the Mirrlees approach.

The Ramsey approach and its lessons

Suppose the government can impose a linear tax on capital income, a linear tax on labour income, and can print money. It must optimally choose these instruments so as to finance an optimally chosen process for government purchases. What are the properties of the optimal taxes? An enormous amount of work has been done on this question; see Chari and Kehoe (1999) for a survey. I first briefly describe the mathematical approach, and then turn to the properties of the optimal taxes.

One way to proceed here would be to solve for the households' and firms' response to all possible tax policies. Then, given this response function, we could solve the government's optimization problem. This problem turns out to be difficult in most circumstances.

Fortunately, there is a way to substitute out the tax schedules; we can instead think of the government directly choosing quantities subject to two types of restrictions. The first is the usual physical feasibility constraints. The other is a set of constraints called *implementability* constraints. These look like household budget constraints, except that we substitute the household marginal rates of substitution in for all prices; the constraints then contain only physical quantities. Somewhat remarkably, these simple implementability constraints turn out to capture exactly the seemingly complicated restriction that the government can use only linear taxes.

Of course, because it is couched only in terms of quantities, the solution to this problem does not contain direct information about optimal taxes. Once one solves the optimization problem, one sees that there are differences (commonly termed *wedges)* between marginal rates of substitution and marginal rates of transformation in the solution. The optimal taxes in equilibrium are equal to these wedges from the solution of the optimization problem. Note these wedges exist only because of the implementability constraints; without them, all wedges would be zero, and it would be optimal to set all taxes to zero.

What then are the properties of optimal taxes when we apply this kind of analysis? In general, the quantitative properties of the optimal taxes depend on many precise details of the specification of the environment. However, there are (at least) two remarkably robust properties of the optimal taxes. The first is that if the government can accumulate assets, the long-run optimal capital income tax rate is zero. (This result was originally derived by Chamley, 1986, and Judd, 1985.) Intuitively, suppose the long-run capital income tax is positive. This tax rate affects the rate of return in every period, and its impact

cumulates as the horizon of the investment grows. Hence, the tax rate on accumulating capital between period t and period $t+s$ gets arbitrarily large as t, s get large. This arbitrarily large tax rate creates too much social waste, given that it is raising only a finite amount of revenue. The second property of optimal taxes is that, under very general conditions, the optimal nominal interest rate is zero (in all periods, not just in the long run) (see Chari, Christiano and Kehoe, 1996; Correia and Teles, 1999; Correia, Nicolini and Teles, 2004).

Here, the basic intuition is that any positive nominal interest rate is a tax on money holdings (as discussed above). But money is not a final good; it is only an intermediate input into production and consumption. A tax on intermediate inputs creates two distortions: people are deterred both from using the intermediate input and from consuming any final goods that use the intermediate input. It is generally optimal to eliminate this double distortion by simply taxing final goods and not taxing any intermediate inputs, including money.

Even though the nominal interest rate is zero, the real interest rate can still be positive as long as the price index is falling over time. If prices are fully flexible, then this consistent deflation has no real effects. However, if prices are sticky, this steady deflation may create inefficiencies in a world with sticky prices. In particular, if some prices are adjusted downward more frequently than others, then any consistent deflation creates distortions in relative prices.

Correia, Nicolini and Teles (2004) demonstrate that this kind of systematic distortion can be fixed by using *sales* taxes. Their key observation is that the nominal interest rate can be zero and the real interest rate can be positive as long as the *after-tax* price level is falling over time. Hence, if the government sets the sales tax to fall at the correct rate, firms will find it optimal to never change their prices even though the nominal interest rate is zero.

The Mirrlees approach and its lessons

The Ramsey approach simply assumes that governments cannot use lump-sum taxes. But why do governments not use lump-sum taxes? One problem is that, if the government imposes a tax of, say, $10,000 per head, then some people will have the ability to generate this income and others will not. This is not a difficulty if the government can tell who is in which group – it can just exempt those who cannot pay.

Unfortunately, people can *pretend* to be unable to generate this level of income by pretending to have back pain, mental illness or other sources of disability. The government cannot figure out whom to exempt from the head tax.

This observation suggests that governments are deterred from using lump-sum taxes because people are privately informed about their abilities or skills. The Mirrlees approach starts with this informational restriction. The government is allowed to use any form of taxes that it wishes (linear, nonlinear, and so on) on any private sector choice. Because it is not restricted to linear taxes, the implementability constraint discussed above vanishes. Instead, the government faces an *incentive-compatibility constraint* that reflects the ability of people to pretend to be less able than they truly are.

Given this difference in constraints, one can proceed much as in the Ramsey approach. The first step is to set up a maximization problem in which the government maximizes *ex ante* welfare subject to feasibility constraints and incentive-compatibility constraints. This type of maximization problem is

roughly equivalent to the kind of dynamic contracting problems originally considered by Green (1987). One considerable complication is that abilities may change over time due to health shocks. Dynamic contracting models with persistent shocks are highly challenging to solve even with a computer (see Fernandes and Phelan, 2000).

The next step is to design a tax system such that the optimal allocations emerge as equilibrium outcomes. These tax systems are complicated objects when abilities evolve over time. Nonetheless, we can draw remarkably strong conclusions about the structure of optimal capital income taxes. If preferences are additively separable between consumption and leisure, then one can show that there exists an optimal tax system which is *linear* in capital income. (Remember that the government is free to use an arbitrarily nonlinear system.) The optimal tax system *subsidizes* the capital income of surprisingly highly skilled people and *taxes* the capital income of surprisingly low-skilled people. While seemingly regressive, this tax system actually provides better social insurance. Intuitively, the tax system provides better incentives because it deters people from accumulating lots of wealth and then pretending to be low-skilled. These better incentives expand the scope for social insurance.

The heterogeneity in tax rates across people means that the Mirrlees prescription for optimal capital taxes differs from the Ramsey prescription for optimal capital tax rates. However, the two approaches do coincide in their recommendations for total and average capital income taxes. The Mirrlees approach recommends subsidies on some people and taxes on others. However, one can prove that, in the optimal tax system, both the average tax rate (across people) and the total tax revenue from capital income taxes are zero at every date. (See Kocherlakota, 2006, for a survey article on the Mirrlees approach.)

Making government endogenous

In both the positive approach and the normative approach, the government is a pre-programmed robot during the life of the economy. It would be useful to develop models in which the government is another economic actor (or, even more realistically, a collection of economic actors) that makes choices at intervals based on its information. Such models would allow us to understand what forces lead to the kinds of policy choices that we see in reality. (See Persson and Tabellini, 2000, for a much more complete discussion of these issues.)

These models need to capture at least two types of conflict. One source of conflict is heterogeneity. Households differ in their attributes and so in their preferences over policies. Old people have shorter horizons and typically prefer to set public investment to lower levels than young people. People with lots of capital prefer lower capital tax rates than do people with little capital. People with lots of nominal debt would like to raise nominal interest rates; their lenders prefer the opposite.

There is a great deal of research studying these kinds of conflicts. Unfortunately, it has been hard to generate the kind of robust answers that macroeconomists have obtained from the positive and normative approaches. There is no real consensus about how to model the games that get played by the different groups. Some researchers use voting games, while others use bargaining games. Some researchers treat conflicts in isolation, while others model conflicts as being resolved in bundles. These different modelling choices generate substantially different predictions about policy formation.

In a classic article, Kydland and Prescott (1977) set forth a second source of conflict. Suppose the world lasts two periods, and a government wants to raise taxes to finance purchases using capital income taxes

and labour income taxes. Assume that all households are identical – so that the first type of conflict is removed – and that the government cares only about maximizing household welfare. It would seem that all sources of conflict have been removed in this situation.

But this is not true. The period 1 government's preferences over period 2 capital taxes are fundamentally different from the period 2 government's preferences. In period 2, the amount of capital in the economy is fixed – there is no way to get any more. The period 2 government would like to set a high tax rate on this fixed tax base to raise as much revenue as possible.

In period 1, though, the amount of capital in period 2 has yet to be determined. The period 1 government has to consider how the tax rate in period 2 affects the size of the period 2 tax base. Its preferred period 2 tax rate is much smaller than the tax rate that the period 1 government likes.

Thus, even if governments at different dates are all benevolent, there is a dynamic conflict between them. How this conflict gets resolved is, again, a non-trivial matter. The dynamic game does have a unique equilibrium in a finite horizon. Unfortunately, this unique equilibrium is unrealistic in most countries: capital tax rates are set very high in every period. On the other hand, if the game has an infinite horizon, then there are an infinite number of equilibrium outcomes, including ones with high capital tax rates, low capital tax rates, and paths that vary between the two (see Chari and Kehoe, 1990). The predictive power of the model is then quite limited.

Conclusions

There is an old joke to the effect that if you ask 10 macroeconomists about a policy question, you'll get 11 different answers. This joke provided a disturbingly accurate picture of the state of the field in the 1970s and 1980s. To a remarkable extent, it was no longer applicable as of 2005. There is a profession-wide consensus on methods that simply did not exist in the early 1980s. This consensus has led to a set of results about monetary and fiscal policy that are sharp, robust and surprising.

See Also

- optimal taxation
- ‘political economy’
- Ricardian equivalence theorem
- social insurance

I thank Barbara McCutcheon for her comments. The opinions expressed herein are mine and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

Bibliography

Abel, A., Mankiw, N.G., Summers, L. and Zeckhauser, R. 1989. Assessing dynamic efficiency: theory and evidence. *Review of Economic Studies* 56, 1–19.

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25, 380–96.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1095–117.

Cass, D. and Shell, K. 1983. Do sunspots matter? *Journal of Political Economy* 91, 193–227.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54, 607–22.

Chari, V.V., Christiano, L. and Kehoe, P. 1996. Optimality of the Friedman rule in economies with distorting taxes. *Journal of Monetary Economics* 37, 203–23.

Chari, V.V. and Kehoe, P. 1990. Sustainable plans. *Journal of Political Economy* 98, 783–802.

Chari, V.V. and Kehoe, P. 1999. Optimal fiscal and monetary policy. In *Handbook of Macroeconomics*, vol. 1C, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Correia, I. and Teles, P. 1999. The optimal inflation tax. *Review of Economic Dynamics* 2, 325–46.

Correia, I., Nicolini, J.-P. and Teles, P. 2004. Optimal fiscal and monetary policy: equivalence results. Working paper, Centre for Economic Performance, London School of Economic.

Fernandes, A. and Phelan, C. 2000. A recursive formulation for repeated agency with history dependence. *Journal of Economic Theory* 91, 223–47.

Green, E. 1987. Lending and smoothing of uninsurable income. In *Contractual Agreements for Intertemporal Trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

Judd, K. 1985. Redistributive taxation in a perfect foresight model. *Journal of Public Economics* 28, 59–84.

Kocherlakota, N. 2006. Advances in dynamic optimal taxation. *Advances in Economics and Econometrics: Theory and Applications: Ninth World Congress of the Econometric Society*, vol. 1, ed. R. Blundell, W. K. Newey and T. Persson. Cambridge: Cambridge University Press.

Kydland, F. and Prescott, E. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

Lucas, R.E., Jr. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit, and Banking* 12, 696–715.

Lucas, R.E., Jr. and Stokey, N. 1987. Money and interest in a cash-in-advance economy. *Econometrica* 55, 491–513.

Obstfeld, M. and Rogoff, K. 1983. Speculative hyperinflations in maximizing models: can we rule them out? *Journal of Political Economy* 91, 675–87.

Persson, T. and Tabellini, G. 2000. *Political Economics, Explaining Economic Policy*. Cambridge, MA: MIT Press.

Prescott, E. 2005. The transformation of macroeconomic policy and research. In *Les Prix Nobel. The Nobel Prizes 2004*, ed. T. Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/2004/prescott-lecture.pdf, accessed 18 October 2006.

Sargent, T. and Wallace, N. 1975. Rational expectations, the optimal monetary instrument and the optimal money supply rule. *Journal of Political Economy* 83, 241–54.

Sargent, T. and Wallace, N. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5(3), 1–18.

Woodford, M. 1994. Monetary policy and price level determinacy in a cash-in-advance economy. *Economic Theory* 4, 345–80.

How to cite this article

Kocherlakota, Narayana R. "monetary and fiscal policy overview." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000383> doi:10.1057/9780230226203.1122

The New Palgrave Dictionary of Economics Online

monetary approach to the balance of payments

Mario I. Blejer and Jacob A. Frenkel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The monetary approach to the balance of payments is an analytical formulation which emphasizes the interaction between the supply and the demand for money in determining the country's overall balance of payments position. It could be seen as an extension, to the case of an open economy, of traditional closed-economy monetary theory, which stresses the stability of the money demand function and considers the various channels through which changes in the money supply affect the economy. If changes in the money supply are not matched by equivalent changes in demand, then a stock disequilibrium arises. In responding to the stock disequilibrium, individuals alter their spending patterns. These adjustments are subject to the budget constraints which link the excess flow supply of money to the corresponding excess flow demand for goods and services. In a closed economy nominal income rises and interest rates may change so as to eliminate the disequilibrium in the money market; the increase in prices, and possibly output, in conjunction with the change in interest rates, raises the nominal demand for money to a level equivalent to the rise in the nominal money stock.

In contrast to the closed economy, the open economy has additional channels through which monetary imbalances are resolved. In the open economy changes in the money stock can arise from domestic credit creation as well as from the foreign exchange operations of the monetary authorities. As a result, the monetary approach to the balance of payments stresses that money market disequilibria are reflected not only in changes in nominal income but also in the country's overall balance of payments, as represented by changes in foreign exchange reserves. Thus the monetary approach to the balance of payments focuses on the relation among prices, output, interest rates, *and* the balance of payments.

In developing the simplest version of the monetary approach to the balance of payments, it is assumed that the country is small, fully employed, that it has a fixed exchange rate, and that there is perfect international mobility of goods and financial assets. These assumptions mean that domestic prices and interest rates equal their respective (exogenously given) world values, and that output is determined exogenously. Under such circumstances, any disequilibrium emerging from the money market is fully reflected in the balance of payments. For example, an excess supply of money arising from domestic credit expansion results in a loss of international reserves. This loss reduces the outstanding money stock to its equilibrium level consistent with the given demand. By concentrating on the direct connection between the money market and the balance of payments, rather than working through the implied

changes in the goods or financial assets markets, the monetary approach distinguishes itself from other analytical approaches to balance of payments theory.

The development of the approach

The monetary approach to the balance of payments has a long intellectual history originating with the 18th-century writings of David Hume. The continuity of its development, however, was reversed for upwards of a quarter of a century by the events of the 1930s. This included the international monetary collapse of 1931 and after, and the ‘Keynesian revolution’.

The modern revival of the monetary approach originated with the writings of James Meade in the early 1950s followed by Harry G. Johnson and Robert A. Mundell in the 1960s. At the same time, important contributions to the formal development of the approach were carried out, under the leadership of Jacques J. Polak at the International Monetary Fund, thereby yielding analytical foundations to the Fund's operational practices.

By the late 1960s a long series of articles, subsequently collected in Mundell (1968, 1971), Frenkel and Johnson (1976) and International Monetary Fund (1977), gave an increasing stimulus to the rapid development of theoretical and empirical work on the monetary approach. Many of the contributions are surveyed in Kreinin and Officer (1978) and in Frenkel and Mussa (1985).

Theoretical Underpinnings

In order to assess the major implications of the monetary approach to the balance of payments, it is useful to present a simplified model which embodies the central characteristics of the analytical approach. The stripdown basic model considers a small, fully employed country operating under a fixed exchange rate system and assumes full integration of domestic and foreign goods and capital markets. Perfect arbitrage determines the prices of domestic commodities and of financial assets.

Because of its concentration on the money market, the monetary approach to the balance of payments involves the explicit specification of the money supply process and of a demand for money function. The supply of money (M^s) is the product of the stock of high-powered money (H) and the money supply multiplier (m) where the latter reflects the behaviour of asset-holders and the banking system:

$$M^s = mH. \tag{1}$$

By definition, the stock of high-powered money (the liabilities of the monetary authorities) is equal to the domestic currency value of the stock of international reserves, eR (where e is the exchange rate, defined as the domestic-currency price of foreign exchange, and R is the foreign currency value of international reserves), and the domestic asset (net of liabilities) holdings of the monetary authorities (D):

$$H = eR + D. \quad (2)$$

The demand for real money balances is specified as a positive function of real income and a negative function of the opportunity cost of holding money. This opportunity cost is measured by the yield on alternative financial assets, usually represented by the rate of interest. The demand for money in nominal terms (M^d) can be written as:

$$M^d = Pf(Y, i) \quad (3)$$

where P denotes the domestic price level, Y is the level of domestic real income, and i stands for the domestic nominal interest rate.
Money market equilibrium implies that $M^s=M^d$. Under the assumptions of the simplified model, the mechanism responsible for maintaining equilibrium operates through changes in international reserves. Accordingly, using equations (1), (2), and (3) the (endogeneously determined) stock of international reserves can be specified as:

$$R = g(P, Y, i, m, D). \quad (4)$$

Equation (4) represents the key relationship implied by the monetary approach to the balance of payments under a fixed exchange rate system. The assumed specifications of M^s and M^d imply that an increase in real income and in (world) prices raises the stock of international reserves while an increase in the rate of interest, in the money multiplier, and in the net domestic assets of the central bank reduces the stock of international reserves. These changes in the stock of international reserves are reflected in balance of payments surpluses or deficits. The size of the income and interest rate effects depends on the elasticities of the money demand function. In this simple model a rise in the money supply brought about by an open-market purchase (an increase in D) is completely offset by a corresponding fall in R. An important implication of the analysis is that under a fixed exchange rate regime the nominal money supply is no longer within the direct control of the monetary authorities and becomes an endogenous variable of the system. The monetary authorities, however, do retain control over the volume of domestic credit, which is one of the sources of money creation. The distinction between high-powered money and its domestic credit component becomes crucial: the central bank controls the latter but not the former. Given a rate of growth in the demand for money, an equivalent growth in the supply can be

obtained by an appropriate increase in domestic credit. However, if the rate of domestic credit expansion differs from the growth in demand, then the difference between the two is made up by changes in net foreign assets, brought about through a balance of payments surplus or deficit.

Extensions and analytical applications

The simplified model presented in the previous section may be seen as a prototype of the monetary approach to the balance of payments and can also be regarded as a representation of its long-run equilibrium characteristics, when all adjustments have taken place. In these circumstances, monetary imbalances tend to affect primarily the balance of payments. However, if the degree of international capital mobility is not high and if the share of non-tradeable goods in GNP is relatively high, then the speed of adjustment to monetary disturbances is reduced. In the short run, therefore, monetary imbalances also affect prices, output, and interest rates, and the relative importance of these effects depends on various factors such as the nature of exchange rate management, the degree of openness of the economy in both the goods and the capital markets, the proportion of tradeable and non-tradeable goods, the degree of resource utilization, the degree of nominal and real wage rigidities, and so forth. Many of those elements have been specifically modelled within the framework of the monetary approach, and the effects of considering different sets of alternative assumptions have been carefully analysed. A central feature of most of the short-term extensions of the basic model is that the excess demand in the commodity market, caused by excess supply in the money market, results in a combination of price increases (which reduce the real value of the outstanding nominal money stock) and balance of payments deficits (which, by depleting the level of international reserves, reduce the level of the nominal money stock). These changes take place in addition to income changes, which in the short run depend on the degree of resource utilization and on the degree to which the public has anticipated the monetary expansion. The effects of monetary disequilibrium fall more heavily on the domestic price level and on the domestic interest rate, and less on the balance of payments, the lower the degree to which the economy is integrated into the world markets for goods and capital. Therefore, the effects of monetary imbalances on the domestic price level and interest rate are stronger the larger are the relative shares of nontraded goods and financial assets, and the more prohibitive are import tariffs, quantitative restrictions and exchange controls.

Further extensions of the basic framework have considered the effects of exchange rate changes on prices and on the balance of payments. In contrast with other approaches to balance of payments analysis (notably the elasticities approach), the monetary approach stresses that the effects of a once-and-for-all exchange rate adjustment in a small economy are transitory. A devaluation (a rise in e) raises the price of internationally tradeable goods. This increase in price reduces the real value of the nominal money stock and, in order to restore money market equilibrium, a balance of payments surplus is generated as foreign exchange reserves flow into the country. As monetary equilibrium is restored, the flow of reserves stops.

The negative relationship between the rate of expansion of domestic credit and the rate of change of foreign exchange reserves implied by the monetary approach does not necessarily imply a unidirectional causality. In fact, it is possible that central banks manipulate their domestic assests in order to sterilize the impact of exogenous changes in foreign reserves on the domestic supply of money. Assume, for example, a reserve-gaining country which desires to avoid an increase in its money supply. The central

bank will tend to counteract the inflow of reserves by reducing its credit to commercial banks or its lending to the government. The required volume and the effects of these sterilization operations could easily be analysed within the framework of the monetary approach, if the parameters underlying the money demand function and the money supply process were known.

The effects of income growth and of external shocks can also be examined within the same set-up. As shown by equation (4), changes in the level of income have a direct impact on the balance of payments through their effect on the demand for money. Therefore, an acceleration in a country's rate of growth, by increasing the demand for liquidity, tends to improve the balance of payments provided that domestic credit policy does not expand accordingly. Similarly, external shocks, such as terms of trade changes, which affect domestic activity, also affect the balance of payments through the same mechanism. In particular, a negative external shock which reduces real income results in a once-and-for-all reduction in the demand for money and (in the absence of domestic credit policy) results in foreign exchange reserves.

The basic model can also be used to determine the effects of commercial policies such as an import tariff. A tariff affects the balance of payments by raising the domestic price level and thereby, by lowering the real value of the outstanding money stock. These changes are likely to induce an excess demand for money which, other things being equal, results in an inflow of international reserves. Similar principles can be used to analyse the effects of other forms of taxation and commercial policies.

Finally, the model could be generalized to the ‘large-country’ case. When the country is not small relative to the rest of the world, one needs to take account of the impacts of its policy and economic behaviour on the world price of tradeable goods and on the world rate of interest. While the monetary mechanism of balance of payments adjustment is more complicated for the large-country case, the basic elements of this mechanism are essentially the same. Starting from a situation in which the domestic nominal money supply is below its long-run equilibrium level and, correspondingly, the foreign money supply is above its long-run equilibrium level, reserve flows associated with trade imbalances gradually move the economic system to long-run equilibrium by raising the domestic money supply and reducing the foreign money supply to their respective long-run equilibrium levels. As in the case of the small country, the essential ingredient underlying this adjustment process is the relationship through which a deficiency in a country's money supply relative to its long-run equilibrium level leads to an excess of domestic income over domestic expenditure which implies a trade surplus which brings an inflow of foreign exchange reserves and a gradual restoration of money balances to their long-run equilibrium level.

In the two-country world, it remains true that a given initial divergence of a country's money supply will ultimately lead to a cumulative payments surplus and change in reserves just equal to this initial divergence, assuming there is no change in the non-reserve assets of central banks.

Overview

In general, a proper analysis of the balance of payments emphasizes the budget constraint imposed on the country's international spending and views the various accounts of the balance of payments as the ‘windows’ to the outside world, through which the excesses of domestic flow demands over domestic flow supplies, and of excess domestic flow supplies over domestic flow demands, are cleared. Accordingly, surpluses in the trade account and the capital account, respectively, represent excess flow

supplies of goods and securities, and a surplus in the money account reflects an excess domestic flow demand for money. Consequently, in analysing the money account, or more familiarly the rate of increase or decrease in the country's international reserves, the monetary approach focuses on the determinants of the excess domestic flow demand for or supply of money.

Although it concentrates on the money account of the balance of payments, the monetary approach should, in principle, give an answer not different from that provided by a correct analysis in terms of the other balance of payments accounts. The surplus or deficit in the goods account (more generally the current account) measures the extent to which the economy's income is greater than consumption (‘absorption’) and the economy is therefore accumulating claims on future income (assets) from abroad or vice versa. By virtue of the budget constraint, the sum of the deficit on the capital account (net purchase of foreign securities) and the surplus on the money account equally represents the accumulation of foreign assets (decumulation if negative). The so-called ‘absorption approach’ to the balance of payments, associated with Sidney Alexander (1952), emphasizes the rate of accumulation or decumulation of foreign assets (securities plus money). In so doing, it differs from the ‘elasticity approach’, which emphasizes relative-price mechanisms.

The monetary approach selects for emphasis a subset of the spectrum of foreign assets whose accumulation or decumulation is emphasized by the absorption approach. The main reasons for this are, firstly, that the accumulation of foreign assets does not necessarily imply the accumulation of money through the balance of payments – it may mean the opposite, as for example when a monetary policy of lowering interest rates leads domestic asset-holders to move their funds from domestic to foreign securities. Secondly, the monetary authorities, in their role as stabilizers of the exchange rate in a fixed rate system, are concerned with what causes the stock of international reserves to change and how to prevent such changes. Thirdly, the monetary authorities, as the ultimate source of domestic money, control the rate of change of the domestic credit component of the monetary base – the other component being international reserves. The assumption that the residents of the country have a demand for money which depends on variables at least in part different from those that determine the quantity of domestic credit extended by the banking system, or alternatively, that the rate of change of money demanded (the rate of hoarding) is independent of the rate of change of the domestic credit source component of the monetary base, implies that the money account of the balance of payments is influenced directly by monetary policy.

See Also

- absorption approach to the balance of payments
- elasticities approach to the balance of payments
- international finance
- purchasing power parity
- specie-flow mechanism

Bibliography

Alexander, S.S. 1952. Effects of a devaluation on a trade balance. *IMF Staff Papers* 2, June, 263–78.

Frenkel, J.A. and Johnson, H.G., eds. 1976. *The Monetary Approach to the Balance of Payments*. London: Allen & Unwin; Toronto: University of Toronto Press.

Frenkel, J.A. and Mussa, M.L. 1985. Asset markets, exchange rates and the balance of payments. In *Handbook of International Economics*, Vol. II, ed. R.W. Jones and P.B. Kenen. New York: Elsevier.

International Monetary Fund. 1977. *The Monetary Approach to the Balance of Payments*. Washington, DC: International Monetary Fund.

Kreinin, M. and Officer, L. 1978. *The Monetary Approach to the Balance of Payments: A Survey*. Princeton Studies in International Finance No. 43, Princeton: Princeton University Press.

Meade, J.E. 1951. *The Theory of International Economic Policy*, vol. I,: The Balance of Payments, Oxford: Oxford University Press.

Mundell, R.A. 1968. *International Economics*. New York: Macmillan.

Mundell, R.A. 1971. *Monetary Theory*. Pacific Palisades, Cal.: Goodyear Publishing Company.

How to cite this article

Blejer, Mario I. and Jacob A. Frenkel. "monetary approach to the balance of payments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000193> doi:10.1057/9780230226203.1117

The New Palgrave Dictionary of Economics Online

monetary business cycle models (sticky prices and wages)

Christopher J. Erceg

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Monetary business cycle (MBC) models are general equilibrium models designed to analyse how monetary shocks affect output, prices, and interest rates. This article describes the analytic framework underlying sticky prices and wages in modern MBC models, and highlights the prominent role that these rigidities play in the transmission of nominal and real shocks.

Keywords

cost-push inflation; Friedman, M.; inflation; inflationary expectations; intertemporal optimization; Keynes, J. M.; Lucas, R.; microfoundations; monetary business cycle models; monetary policy rules; monetary shocks; monetary transmission mechanism; New Keynesian Phillips curve; nominal rigidities; nominal shocks; nominal wage inflation; output gap; Phelps, E.; Phillips curve; price dynamics; rational expectations; real business cycles; real rigidities; real shocks; staggered contracts model; sticky prices; sticky wages; technology shocks; unemployment

Article

Since the earliest analysis of the monetary transmission mechanism by pre-eminent classical economists of the 18th and early 19th century, sticky prices and wages have been identified as playing a central role (Humphrey, 2004). The classical economists believed that prices adjusted gradually to a change in the nominal money stock, so that monetary changes could exert substantial short-run effects on output. Nominal wages were regarded as particularly slow to change, and thus helped account for gradual price adjustment by mitigating short-run pressures on factor costs.

The classical economists and their successors used this framework both to guide recommendations about policy and to evaluate alternative monetary regimes. For example, the belief that prices would respond slowly to a monetary contraction led Thornton and Ricardo to recommend a gradualist approach to deflation.

Early Keynesian models, and some critiques

A major contribution of Keynes (1936) and prominent successors such as Hicks to understanding the monetary transmission mechanism consisted in developing an explicit theoretical framework expressed in terms of equilibrium conditions in goods and asset markets. This IS–LM framework was of great value in illuminating the channels through which monetary shocks affected interest rates and output. However, the assumption of fixed prices and wages was a major shortcoming. It was eventually supplanted by the famous ‘Phillips curve’ relation linking nominal wage inflation to the unemployment rate, or variants relating price inflation to the output gap:

$$p(t) - p(t-1) = b^{*}(y(t) - y(t)^{*}) b > 0 \quad (1)$$

where $p(t)$ is (the log of) the price level, $y(t)$ output, $y(t)^*$ potential output, and b is a parameter. The Phillips curve filled a missing link in earlier ‘fixed price’ IS–LM analysis by making it feasible to trace the dynamic effects of a monetary shock on prices and output. Thus, an initial rise in output following a monetary expansion boosts prices via (1), which in turn causes real balances and output to revert gradually to pre-shock levels. However, the Phillips curve had weak theoretical underpinnings, so that there was little economic rationale for what determined the sensitivity of prices to the output gap (that is, ‘*b*’ in (1)), for the activity variable(s) driving price dynamics, and for how inflation might be influenced by expectations.

A series of remarkable critiques beginning with the analysis of Friedman (1968) and Phelps (1968) provided impetus for developing more theoretically coherent models of price and wage dynamics. These authors argued that the Phillips curve should be augmented so that actual inflation depended directly on inflation expectations in addition to real activity. In this framework, output could be pushed above potential only through surprising private agents by keeping inflation above the level that they had forecast in previous periods. Since such surprises could not continue indefinitely, there could be no long-run trade-off between inflation and output: expansionary monetary policy would eventually raise expected inflation, resulting in higher inflation with no output stimulus.

Shortly thereafter, Lucas (1972) derived an ‘expectations-augmented’ Phillips curve in a clearly specified rational expectations model. Lucas adopted a signal extraction framework in which agents partly misinterpreted aggregate nominal shocks as shocks to the relative price of their own output good (due to limited information), and responded by adjusting their supply. Consistent with Friedman and Phelps, Lucas's model implied that aggregate output varied positively with the unanticipated component of inflation (with anticipated inflation exerting no real effects). But because unanticipated inflation was linked explicitly to a ‘rational expectations’ forecast error in Lucas's model – which would be expected to die away quickly as agents learned about the nature of underlying shocks – monetary shocks could exert only transient effects on output. This posed a serious challenge to traditional Keynesian models by suggesting that their ability to derive persistent effects in response to a monetary injection relied on ad

hoc assumptions about price dynamics or expectations formation. Moreover, because only unanticipated changes in inflation affected output, Lucas's supply relation implied that any predictable policy was as good as any other (the ‘policy ineffectiveness’ proposition). This point, emphasized by Sargent and Wallace (1975), contrasted sharply with the activist policy stance that emerged from typical Keynesian models.

Monetary transmission in optimization-based MBC models

Since the mid-1990s a new generation of optimization-based MBC models has emerged that can generate ‘traditional’ Keynesian implications, but in a framework consistent with rational expectations and rigorous microfoundations. Roughly speaking, these new MBC models graft features that can induce sluggish price and/or wage adjustment onto an underlying real business cycle (RBC) model. (Blanchard, 2000, and Taylor, 1999, provide comprehensive surveys of the foundations of modern optimization-based MBC models, which were laid in a series of important contributions spanning several decades.)
To highlight salient features of the modern approach, it is helpful to examine a specific characterization of price-setting that has been utilized extensively in the literature. This relation, often called the ‘New Keynesian Phillips curve (NKPC)’, takes the form

$$p(t) - p(t-1) = B * E(t)\,[\,p(t+1) - p(t)\,] + b * (y(t) - y(t)^{*}) \qquad (2)$$

where $E(t)$ is the conditional expectation operator, and B is the discount factor.
Following Calvo (1983) and Yun (1996), the NKPC can be derived in a framework consistent with intertemporal optimization. Firms are assumed to behave as monopolistic competitors in the output market, and face downward-sloping demand curves for their distinctive products. Firms face a dynamic decision problem, because they are constrained to set a price that remains fixed in nominal terms over some random duration of time (referred to as the ‘contract period’, since firms are assumed to meet all demand at this fixed price until allowed to adjust). When a firm receives a signal enabling it to adjust its price, the firm resets it based on estimates of current and future marginal costs expected to prevail over the contract period. Because not all firms can change their price in a given period, price-setting is staggered – similar to the decentralized price-setting in actual economies. (See sticky wages and staggered wage setting for a discussion of the staggered contracts model.)
From a qualitative perspective, an MBC model in which prices are determined by the NKPC provides a conventional Keynesian account of the monetary transmission mechanism. Thus, a monetary shock increases nominal spending and, since the price level adjusts gradually, real output exhibits a persistent increase (in contrast to the transient real effects in Lucas's model). But as time passes, a larger proportion of firms receive a signal that allows them to raise their price in response to higher projected marginal costs. At an aggregate level, these relative price adjustments translate into a higher price level, which eventually restores real balances and output to pre-shock levels.

A major virtue of the microfounded approach is that it illuminates how the monetary policy rule and various structural features of the economy affect the transmission of nominal (and real) shocks. First, given that price adjustment is influenced directly by inflation expectations (as in (2)), monetary surprises have smaller effects on current inflation to the extent that the policy rule is expected to keep future inflation near target (that is, ‘anchors’ inflation expectations). Second, while the sensitivity of price inflation to the output gap (‘*b*’) clearly plays a key role in determining how quickly prices and output adjust to a monetary injection, this parameter is itself determined by features of the microeconomic environment. Quite intuitively, the parameter ‘*b*’ varies inversely with the mean duration of price contracts, so that longer contracts imply slower price adjustment and more persistent effects on output. But ‘*b*’ also depends on the responsiveness of firm-level marginal costs to the aggregate output gap, which in turn hinges on features of the specific microeconomic environment, including assumptions about factor mobility, capital utilization, and preferences. While some assumptions constrain ‘*b*’ to be large, a considerable literature has emerged showing how various ‘real rigidities’ such as firm-specific capital and labour can account for a low ‘*b*’ (even with fairly short-lived contracts); an insightful overview is provided in Woodford (2003). Such real rigidities appear important in allowing macro models to account for persistent output effects, while remaining consistent with disaggregate price data suggesting that firms change prices frequently (Bils and Klenow, 2004).

The NKPC in (2), in which the output gap enters as the activity variable, is derived under the assumption that wages are fully flexible. But, as noted above, there is a long precedent in macroeconomics suggesting that sticky wages play an important role in the transmission process. As shown by Erceg, Henderson and Levin (2000), wage rigidity may be modelled in a framework isomorphic to that rationalizing price rigidity, with households acting as monopolistic suppliers of differentiated labour services. Christiano, Eichenbaum and Evans (2005) have shown that a model that incorporates both wage and price rigidity can account remarkably well for the estimated dynamic effects of a monetary shock on output, prices, and interest rates. The presence of wage rigidity damps the rise in marginal cost due to a positive monetary injection, helping account for estimated persistence in the response of output. Moreover, a model including both types of rigidities can help account for the observed acyclicality of the real wage. By contrast, sticky prices alone imply too much procyclicality in the real wage, while sticky wages alone (in the spirit of the classical economists and Keynes) imply too much counter-cyclicality.

Real shocks and alternative policies in MBC models

Given that monetary policy is widely perceived to have been much more stable since the mid-1980s, the literature has focused greater attention on how policy should respond to real shocks. Modern optimization-based MBC models are useful in this regard, because they provide a coherent framework for examining the transmission of real shocks in the presence of sticky wages and prices, and for assessing the role of monetary policy in affecting the economy's responses.

The presence of nominal rigidities can markedly affect the economy's responses to real shocks. Following Gali (1999), this can be illustrated by contrasting the effects of a persistent rise in technology in an RBC model (in which prices and wages are flexible) with the effects in an MBC model in which prices adjust according to eq. (2). For simplicity, it is assumed that money demand takes the interest-

inelastic form $M = P^{*}Y$, and that the monetary authority holds the nominal money stock constant. In either model, money market equilibrium implies that output can expand only if prices fall proportionally. But as prices can drop instantaneously in the RBC model, the money supply rule is irrelevant in determining the real effects of the shock. Thus, the technology shock immediately boosts employment (as the substitution effect dominates the income effect), and the (percentage) jump in output exceeds the magnitude of the shock. By contrast, prices fall gradually in the MBC model, so that output is constrained to rise slowly given the fixed money stock. With prices determined by the NKPC, negative output gaps are required to induce prices to fall, consistent with employment remaining persistently below its pre-shock level.

As in the case of nominal shocks, the effects of real shocks may be highly sensitive to underlying features of the microeconomic framework, including those that determine the speed of price or wage adjustment. Thus, features that affect '*b*' in the NKPC can markedly change how real shocks impact the economy. In the case of the technology shock, additional price sluggishness would translate into a smaller short-run expansion in output and greater employment contraction. Similarly, the inclusion of wage stickiness can markedly affect the responses to technology shocks. For example, while the NKPC derived under the assumption of flexible wages (eq. (2)) implies that price inflation stabilization also keeps output at potential, the same policy could generate large output gap fluctuations if wages were sticky as well as prices.

Modern MBC models have also been applied fruitfully to normative issues. Optimal policy is derived by maximizing an objective function subject to the model's behavioural equations. Importantly, the objective function used in ranking alternative policies is typically derived from the utility functions of the economy's households (Woodford provides an extensive treatment).

A compelling message of this normative literature is that a well-designed policy must take account of its ability to influence inflation through an expectations channel. Thus, a policymaker acting 'under discretion' in an environment where inflation was determined by (2) would act as if the only margin on which to trade in devising a policy involved current inflation and output. However, such a 'discretionary' policy is suboptimal, because it fails to take account of its influence on the expected inflation term in (2). The analysis of Clarida, Gali and Gertler (1999) and Woodford shows that rules that are devised to take account of their influence on future expected inflation can perform much better in maximizing social welfare than discretionary policies that take future inflation as outside the central bank's control. For example, these authors show that well-designed policies can reduce substantially the impact of an adverse cost-push shock on current inflation (relative to the effects under discretion) by creating the perception that future policy will bring inflation back quickly to baseline.

Woodford emphasizes that the optimal monetary policy rule in an environment with forward-looking price-setting exhibits history dependence, so that current monetary policy actions depend on past inflation and activity. This inertial character reflects that the optimal policy rule is derived in a framework in which future policy is expected to take full account of its influence on inflation expectations at earlier dates, much as optimal tax rules recognize their impact on previous investment decisions. Consistent with this history dependence, Woodford shows that it is generally optimal for monetary policy to reverse spikes in inflation above its target value, rather than follow the conventional wisdom of allowing 'bygones to be bygones'. Interestingly, this analysis provides strong support for some form of price level targeting – as recommended by Fisher and Keynes nearly a century ago – with the twist that the modern justification highlights the role it can play in optimally anchoring inflation

expectations.

See Also

- IS–LM in modern macro
- monetary transmission mechanism
- Phillips curve (new views)
- real rigidities
- sticky wages and staggered wage setting

Bibliography

Bils, M. and Klenow, P. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112, 947–85.

Blanchard, O. 2000. What do we know about macroeconomics that Fisher and Wicksell did not? *Quarterly Journal of Economics* 115, 1375–409.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal rigidities and the dynamic effects of shocks to monetary policy. *Journal of Political Economy* 113, 1–45.

Clarida, R., Gali, J. and Gertler, M. 1999. The science of monetary policy: a new Keynesian perspective. *Journal of Economic Literature* 37, 1661–707.

Erceg, C., Henderson, D. and Levin, A. 2000. Optimal monetary policy with staggered wage and price contracts. *Journal of Monetary Economics* 46, 281–313.

Fisher, I. 1920. *Stabilizing the Dollar*. Norwood, MA: Norwood Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1), 1–17.

Gali, J. 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations? *American Economic Review* 89, 249–71.

Humphrey, T. 2004. Classical deflation theory. *Federal Bank of Richmond Economic Quarterly* 90, 11–32.

Keynes, J. 1936. *The General Theory of Interest, Employment, and Money*. London: Macmillan.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Phelps, E. 1968. Money–wage dynamics and labor market equilibrium. *Journal of Political Economy* 76, 678–711.

Sargent, T. and Wallace, N. 1975. Rational expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83, 169–83.

Taylor, J. 1999. Staggered wage and price setting in macroeconomics. In *Handbook of Macroeconomics*, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37, 345–70.

How to cite this article

Erceg, Christopher J. "monetary business cycle models (sticky prices and wages)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000403> doi:10.1057/9780230226203.1118

The New Palgrave Dictionary of Economics Online

monetary business cycles (imperfect information)

Christian Hellwig
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Business cycle theories based on incomplete information start from the premise that key economic decisions on pricing, investment or production are often made on the basis of incomplete knowledge of constantly changing aggregate economic conditions. As a result, decisions tend to respond slowly to changes in economic fundamentals, and small or temporary economic shocks may have large and long-lasting effects on macroeconomic aggregates. This article provides an introductory overview of incomplete information-based theories of business cycles, from their origins to the most recent theoretical developments.

Keywords

common information; forecasting the forecasts of others; heterogeneity in beliefs; higher-order expectations; higher-order uncertainty; imperfect common knowledge; imperfect information; information aggregation; informational frictions; law of iterated expectations; local markets; menu cost; monetary business cycle models; monetary shocks; monopolistic competition; New Keynesian economics; pricing complementarities; rational expectations; rational inattention; real rigidities; sticky prices

Article

Business cycle theories based on incomplete information start from the premise that key economic decisions on pricing, investment or production are often made on the basis of incomplete knowledge of constantly changing aggregate economic conditions. As a result, decisions tend to respond slowly to changes in economic fundamentals, and small or temporary economic shocks may have large and long-lasting effects on macroeconomic aggregates.
Incomplete information theories have been popular in particular for explaining sluggish price or wage adjustment in response to monetary shocks. At the heart of this theory lies the assumption that firms or

households only pay attention to a relatively small number of indicators regarding conditions in markets relevant to their own activities, but they may not acquire information more broadly about aggregate economic activity. With imprecise information about these aggregate conditions, it takes the firms some time to sort out temporary from permanent changes, or nominal from real disturbances. Prices then respond with a delay to changes in nominal spending, and monetary shocks may have significant effects on real economic activity in the intervening periods – despite the fact that firms have the opportunity to constantly readjust their decisions.

This basic idea was proposed first by Phelps (1970) and formalized by Lucas (1972). In Lucas (1972), economic agents produce in localized markets, in which they observe the market-clearing price at which they can sell their output. This price is affected both by aggregate spending shocks and by market-specific supply shocks. Under perfect information, quantities adjust in response to local supply shocks, but not prices, and prices respond to aggregate spending shocks, but not quantities. With imperfect information, agents are unable to filter out the magnitudes of the aggregate and market-specific shocks from the observed prices in the short run. Output then responds positively to price changes and spending shocks in the short run, but not in the long run, once agents have been able to sort out the spending shocks from the market-specific supply shocks.

Lucas (1972) formulated this idea in a rational expectations market equilibrium model, in which agents' expectations are fully Bayesian, and the resulting output responses are optimal. His model also includes stark assumptions about the nature of local versus aggregate market interactions, as well as the nature of shocks (monetary versus real, demand versus supply, aggregate versus market-specific) and the information to which firms have access.

Importantly, the model lacks a natural internal amplification mechanism: the extent of incomplete nominal adjustment depends almost entirely on the degree of informational incompleteness. Subsequent work has tried to address these issues, for example by introducing richer information structures.

Townsend (1983) considers an investment model in which firms get to observe how much some of the other firms invest. Therefore, they need to form forecasts about each others' beliefs – forecasting the forecasts of others. This leads to a complicated infinite regress problem, whereby a firm's current investment level depends on its observation of other firms' past investment, which in turn depended on observations about past investment… Townsend showed that this type of problem does not admit a simple finite-dimensional recursive structure. As a result, firms must draw inference about all past realizations of shocks simultaneously, leading to an infinite-dimensional filtering and fixed point problem, with no easily characterized solution.

These and other important technical and computational hurdles effectively imposed limitations on the complexity and economic realism of the early incomplete information models. Moreover, the model is open to the criticism that if incomplete information is a major source of business cycle fluctuations, then there seems to be an important societal benefit to making the relevant information publicly available to everyone. In part because of these difficulties, economists have, from the mid-1980s, turned their attention to New Keynesian sticky price theories that emphasize the role of adjustment and coordination frictions in price-setting. (Among others, see Calvo, 1983; Blanchard and Kiyotaki, 1987.)

Recently, the incomplete information theories have made a comeback, which can be traced to two factors. First, technological progress has made models such as Townsend (1983) computationally tractable. Second, new game-theoretic results regarding equilibrium analysis with a lack of common knowledge and heterogeneity in beliefs, as well as insights borrowed from the sticky price literature

regarding the role of real rigidities and pricing complementarities (Ball and Romer, 1990) have enabled us to paint a much richer picture of the adjustment dynamics resulting from incomplete information models. The empirical performance of these new incomplete information models, however, still remains to be seen.

In the remainder of this article I provide a unified exposition of the main ideas behind the incomplete information theories, from the original contributions to the more recent renewal. I also attempt to chart out some of the challenges that lie ahead. This is a lively and active area of research, with many open questions and few definite answers.

A canonical framework

Consider the following model, which is based on the New Keynesian models of monopolistic competition. There is a large number of firms, indexed by $i \in [0,1]$. In each period, each firm sets its (log-)price $p_t(i)$ equal to its expectation of a target price p_t^*, $p_t(i) = E(p_t^* | \Phi_t^i)$, where Φ_t^i denotes the information set of firm i at date t, that is all signals on which it can condition its pricing decision. p_t^* is characterized as

$$p_t^* = ky_t + p_t, \tag{1}$$

where $p_t = \int p_t(i)\,di$ denotes the average of the firms' pricing decisions, y_t denotes the aggregate real output in period t, relative to its trend level that would prevail with complete information, and $k>0$ measures the response of optimal pricing decisions to real output. A firm's ideal relative price $p_t^* - p_t$ is determined by real output deviations from trend.

We augment this pricing rule by a quantity equation, $y_t+p_t=m_t$, where m_t denotes nominal spending. Combining the two, we find

$$p_t^* = km_t + (1-k)\,p_t. \tag{2}$$

Nominal spending m_t is driven by exogenous shocks; for simplicity, assume that $m_t=m_{t-1}+\varepsilon_t$, where $\{\varepsilon_t\}$ is i.i.d. white noise.

Each firm's target price is therefore a linear combination of the exogenous shocks and the prices set by the other firms. If $k\in(0,1)$, prices are complementary, that is, an increase in the average price level

implies that each firm has an incentive to raise its own price. The parameter value of k depends on the substitution elasticity between the firms' products, the firms' returns to scale parameter in the technology, and the Frisch elasticity of labour supply.

To complete the model description, we need to specify each firm's information set $\mathcal{I}_t^i$ – this is where different incomplete information theories vary. An equilibrium of this model requires that prices satisfy the optimality condition $p_t(i) = E(p_t^* | \mathcal{I}_t^i)$, taking into account that p_t^* itself depends on the aggregate price level.

Common information

Suppose first that all firms have identical information sets, $\mathcal{I}_t^i = \mathcal{I}_t$. Then, they will set identical prices, equal to $p_t(i) = p_t = E(m_t | \mathcal{I}_t)$. This reflects the implications of the original Lucas model that prices adjust to the common expectation of the underlying shocks. When information is incomplete, firms will only learn gradually about m_t, prices adjust slowly, and monetary surprises have real effects: y_t is determined directly by the discrepancy between the realized and the expected value of m_t. However, if the available information on which these expectations are based is sufficiently precise, then $E(m_t | \mathcal{I}_t)$ cannot be far from the true value of m_t. As discussed above, the real effects of monetary shocks are bounded by the degree of informational incompleteness – as firms have better information, their prices track m_t more closely, and monetary shocks have smaller real effects.

Heterogeneous beliefs, but independent strategies

A similar conclusion emerges when firms have different information sets, but their target prices do not respond to the other firms' decisions (k=1). Each firm's price is set equal to its expectation of the spending shock $p_t(i) = E(m_t | \mathcal{I}_t^i)$, and the average price adjusts according to the average expectation $p_t = \bar{E}(m_t) = \int E(m_t | \mathcal{I}_t^i)\, di$ of the spending shock. Once again, if firms are sufficiently well informed, their pricing decisions will on average not be far from the nominal spending shock, which implies little delay in price adjustment and only small real output effects.

Heterogeneous beliefs and complementary strategies

Suppose now that instead $k \in (0,1)$, so that there are complementarities in pricing decisions. Averaging the pricing equation, and substituting forward, firm i's equilibrium price is given by

$$p_t(i) = k \sum_{s=0}^{\infty} (1-k)^s E\left[\bar{E}^{(s)}(m_t) | \mathcal{I}_t^i\right] \tag{3}$$

where $\bar{E}^{(s)}(m_t)$ denotes the *s*-order average expectation of m_t, or the average expectation of the average expectation of … (repeat *s* times) … of m_t. A firm's optimal price is therefore given as a geometrically weighted average of higher-order expectations – a firm needs to forecast not only the realized shock but also the other firms' expectations of the shock, the other firms' expectations of the other firms' expectations of the shock, and so on.

If the firms all had identical information, the law of iterated expectations would simply collapse the right-hand side above into the common first-order expectation of m_t. The model thus derives its interest from the fact that with heterogeneous information, higher-order expectations respond differently to new information than first-order expectations about m_t.

The following example illustrates this point and serves also to derive the main results of this model. Suppose that all firms observe m_{t-1} exactly, but only a fraction λ (the *informed*) gets to observe m_t. Then, $\bar{E}(m_t) = \lambda m_t + (1-\lambda)m_{t-1}$, but the second order average expectation is $\bar{E}^{(2)}(m_t) = \lambda[\lambda m_t + (1-\lambda)m_{t-1}] + (1-\lambda)m_{t-1} = \lambda^2 m_t + (1-\lambda^2)m_{t-1}$. By iteration, the *s*-order average expectation of m_t is $\bar{E}^{(s)}(m_t) = \lambda^s m_t + (1-\lambda^s)m_{t-1}$. The average price is

$$p_t = k\sum_{s=0}^{\infty}(1-k)^s\bar{E}^{(s+1)}(m_t) = m_{t-1} + \frac{k\lambda}{1-(1-k)\lambda}(m_t - m_{t-1}). \tag{4}$$

Two important conclusions emerge. First, note that $\frac{k\lambda}{1-(1-k)\lambda} < \lambda$. The informed firms whose prices may react to m_t take into account that the uninformed firms won't respond, which in turn reduces their incentives to adjust prices. Therefore, while incomplete information serves as the initial source of sluggish price adjustment, the complementarity and the heterogeneity in beliefs dampen the response of prices far beyond what the initial degree of informational incompleteness would suggest. To illustrate the strength of this amplification effect, consider the following numerical example: suppose that k=0.15 (as in standard parametrizations of New Keynesian sticky price models), and that half the firms are informed. Then, the contemporaneous response of average prices is $\frac{k\lambda}{1-(1-k)\lambda} \approx 0.13$, that is a one per cent increase in nominal spending leads to only a 0.13 per cent increase in prices, and a 0.87 per cent increase in real output – despite the fact that half of the firms actually observe the increase in nominal spending and are hence able to respond to it!

Second, this amplification can be large, even if the degree of informational incompleteness is small. If λ is close to 1, almost all firms exactly observe the current realization m_t. Nevertheless, if k is close to 0, that is if there is a strong pricing complementarity, they still won't respond to the monetary shock. The presence of only a few uninformed firms is therefore enough to radically overturn the conclusions of the

complete information model.

These two observations apply quite generally, once firms have heterogeneous beliefs. They form the central insight of the new incomplete information theories. In Mankiw and Reis (2002), heterogeneous beliefs result because, in any given period, only a fraction of firms observe new information. This generalizes the above example to allow for richer adjustment dynamics. In Woodford (2002), all firms observe a conditionally independent idiosyncratic signal x_t^i of the current realization of m_t in each period. The resulting inference problem is more complicated but can be solved numerically. Again, the response of prices to monetary shocks is significantly dampened by the fact that firms do not share in common information, yet their pricing decisions are complementary.

The role of public information

Hellwig (2002) provides a simplified version of Woodford (2002), providing closed-form solutions to a general class of information structures. This simplified model also accommodates the presence of additional public sources of information such as central bank announcements. Besides dampening the response to idiosyncratic private signals, the complementarity in prices generates overreaction to public news. Public announcements thus speed up price adjustment and reduce the real effects of monetary shocks, but the noise in public news creates an additional source of volatility, which in some cases may increase rather than decrease real output fluctuations. (Similar results are derived by Amato and Shin, 2003, for Woodford's model, and by Ui, 2003, in the original Lucas island model.)

Looking ahead

These new contributions have provided promising insights into the amplification and propagation mechanisms of incomplete information models. But they also abstract from important modelling issues that need to be addressed before a comprehensive quantitative evaluation becomes possible.

So far, much of the analysis is based on a stylized price-setting model that captures the essence of pricing complementarities as described above, without deriving them within a fully specified dynamic general equilibrium model. This short-cut is not without problems. First, the lack of a proper context of markets makes it difficult to interpret these propagation results. Presumably in a market firms obtain some information about price and quantity variables – so far, this is not formally modelled.

Second, the assumption that firms are heterogeneously informed implies that other frictions must be present – in particular, the extent to which information about fundamental shocks can be inferred from publicly observable prices must be limited, implying that the asset market must be incomplete. But then, one faces the problem of isolating the effects of informational heterogeneity from the effects of other market imperfections. In Lorenzoni, (2006) for instance, a precautionary savings motive generates a multiplier effect in household spending, which is further amplified by the presence of heterogeneous information.

Third, there is an issue of interpretation. At this point, there exist several different interpretations regarding the source of the differences in beliefs across firms, and they may lead to radically different model conclusions. In Mankiw and Reis (2002), firms update their information only infrequently, and in

the intervening periods set prices on the basis of outdated information; Reis (2006) further develops this idea on the basis of menu costs in updating decisions. Woodford (2002) instead bases his model on the notion of ‘rational inattention’, developed by Sims (2003; 2006a). Sims argues that decision makers only have a finite capacity to process new information, which constrains the quality of the signals they observe in any given period. Heterogeneity in beliefs then arises naturally through the idiosyncratic noise in each individual's information processing channel (see Sims, 2006b, for further discussion of the resulting conceptual and modelling issues). A third interpretation suggests that individuals are Bayesian, but access to information is limited – for example, firms observe the demand for their own products, but not the demand for competitors’ products. If each firm is subject to idiosyncratic, as well as common shocks, then an information structure much like the above with idiosyncratic private signals emerges. On the other hand, firms also observe market prices, which generates a source of common information.
Finally, all these models treat the information structure as an exogenous primitive. In reality, firms and households have access to overwhelming amounts of information, and information processing becomes a matter of choice, given the existing constraints and trade-offs. By and large, the effects of information costs and choices and the strategic interaction that results from these choices remains unexplored. Preliminary developments in this direction include Mackowiak and Wiederholt (2005) and Hellwig and Veldkamp (2005). In Mackowiak and Wiederholt, firms need to allocate a fixed information processing capacity between firm-specific and aggregate variables. Hellwig and Veldkamp explore how the pricing complementarities that are relevant for business cycle implications also shape incentives for information acquisition.
In summary, the most important issue that remains to be resolved is the grounding of new incomplete information theories within a fully specified model of goods and asset markets, with special emphasis on the origins of the informational frictions. Beyond that, the new incomplete information theories raise many intriguing questions, which merit further attention, or have already been addressed to some extent: for example, Ball, Mankiw and Reis (2005) reconsider the role of monetary policy, and Morris and Shin (2002), Hellwig (2005) and Angeletos and Pavan (2004; 2007) discuss the welfare effects of information disclosures. Finally, the combination of new evidence on the cross-sectional and business cycle properties of expectations (Mankiw, Reis and Wolfers, 2004) and new micro-level data on price adjustments (Bils and Klenow, 2004) promises to provide an interesting avenue for evaluating the empirical performance of the model's cross-sectional and business cycle implications.

See Also

- information aggregation and prices
- Lucas, Robert
- monetary business cycle models (sticky prices and wages)
- monetary transmission mechanism
- Phelps, Edmund

Bibliography

Amato, J. and Shin, H.S. 2003. Public and private information in monetary policy models. Working Paper No. 138, Bank of International Settlements.

Angeletos, G.-M. and Pavan, A. 2004. Transparency of information and coordination in economies with investment complementarities. *American Economic Review* 94, 91–8.

Angeletos, G.-M. and Pavan, A. 2007. Efficient use of information and social value of information. *Econometrica* 75, 1103–42.

Ball, L., Mankiw, G. and Reis, R. 2005. Monetary policy for inattentive economies. *Journal of Monetary Economics* 52, 703–25.

Ball, L. and Romer, D. 1990. Real rigidities and the non-neutrality of money. *Review of Economic Studies* 57, 183–203.

Bils, M. and Klenow, P. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112, 947–85.

Blanchard, O. and Kiyotaki, N. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77, 647–66.

Calvo, G. 1983. Staggered prices in a utility maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Hellwig, C. 2002. Public announcements, adjustment delays and the business cycle. Discussion paper, University of California, Los Angeles.

Hellwig, C. 2005. Heterogeneous information and the welfare effects of public information disclosures. Discussion paper, University of California, Los Angeles.

Hellwig, C. and Veldkamp, L. 2005. Knowing what others know: coordination motives in information acquisition. Discussion paper, University of California, Los Angeles and New York University.

Lorenzoni, G. 2006. A theory of demand shocks. Discussion paper, Massachusetts Institute of Technology.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Mackowiak, B. and Wiederholt, M. 2005. Optimal sticky prices under rational inattention. Discussion paper, Humboldt University Berlin.

Mankiw, G. and Reis, R. 2002. Sticky information versus sticky prices: a proposal to replace the new Kaynesian Phillips curve. *Quarterly Journal of Economics* 117, 1295–328.

Mankiw, G., Reis, R. and Wolfers, J. 2004. Disagreement about inflation expectations. In *NBER Macroeconomics Annual 2003*. Cambridge, MA: MIT Press.

Morris, S. and Shin, H.S. 2002. The social value of public information. *American Economic Review* 92, 1521–34.

Phelps, E. 1970. Introduction: the new microeconomics in employment and inflation theory. In *Microeconomic Foundations of Employment and Inflation Theory*. New York: Norton.

Reis, R. 2006. Inattentive producers. *Review of Economic Studies* 73, 1–29.

Sims, C. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50, 665–90.

Sims, C. 2006a. Rational inattention: beyond the linear-quadratic case. *American Economic Review* 96, 158–63.

Sims, C. 2006b. Rational inattention: a research agenda. Discussion paper, Princeton University.

Townsend, R. 1983. Forecasting the forecasts of others. *Journal of Political Economy* 91, 546–88.

Ui, T. 2003. A note on the Lucas model: iterated expectations and the non-neutrality of money. Discussion paper, Yokohama National University.

Woodford, M. 2002. Imperfect common knowledge and the effects of monetary policy. In *Knowledge, Information and Expectations in Modern Macroeconomics*, ed. P. Aghion. et al. Princeton: Princeton University Press.

How to cite this article

Hellwig, Christian. "monetary business cycles (imperfect information)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000375> doi:10.1057/9780230226203.1119

The New Palgrave Dictionary of Economics Online

monetary cranks

David Clark

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Douglas, C. H.; gold standard; monetary cranks; money supply; Social Credit; underconsumptionism

Article

The history of ideas tends to concentrate on the successful ideas – ideas which appear to have been precursors of the orthodoxy of the day. As a result, ideas which had large followings but which are later considered ‘cranky’ tend to be ignored. This is especially true of the ideas of those who we can loosely call the monetary cranks.

These persons have placed money at the centre of their economic analysis, have usually placed major blame for society's evils on alleged financial conspiracies and bankers’ ramps – on the ‘Money Power’ – and have advocated a variety of monetary experiments. Over the past century particularly, such concerns can be found in all Western countries, on both the Left and the Right of politics. This article can only provide the broadest of overviews of the voluminous literature in this field.

Opposition to financial oligarchies has a long history. The Medicis of 15th-century Florence aroused suspicion and hostility. In *Lombard Street* (1873), Walter Bagehot described the streets around the Bank of England in London as ‘by far the greatest combination of power and economic oligarchy that the world has ever seen’. But it was the fiery late-19th-century American populist, William Jennings Bryan, who popularized the term ‘Money Power’ (cited in Douglas, 1924, Preface):

> The Money Power preys upon the nation in times of peace and conspires against it in times of adversity. It is more despotic than monarchy, more insolent than autocracy, more bureaucratic than bureaucracy. It denounces, as public enemies, all who question its methods, or throw light upon its crimes. It can only be overthrown by the awakened conscience of the nation.

Monetary parables have a long history, ranging from David Hume's 1752 hope that ‘by miracle, every man in Great Britain should have five pounds slipped into his pocket in one night’, through to Milton

Friedman's 1969 postulated helicopter miracle, whereby dollars would be dropped from the heavens. (These are discussed in Clayton, Gilbert and Sedgwick, 1971, p. 6.) Over the past three centuries, however, actual monetary experiments have taken two main forms: attempts to overcome economic fluctuations by means of adjusting note issue; and attempts to achieve a more stable price level through the formulation and adoption of a new or different monetary standard.

Such experiments were first undertaken in the North American colonies. The first paper money issued by any government in Europe or the Americas was printed by Massachusetts to pay the wages of its soldiers engaged in conflict with the French in Canada at the end of the 17th century. Other New England colonies followed suit and a competitive depreciation of the individual currencies followed. The French Canadians even used playing cards as a form of money.

In 1721, a Mr Wise of Chebacco, Massachusetts, concerned at the depreciation of the notes admonished his fellow colonists (cited in Lester, 1939):

> Gentleman! You must do by your Bills, as all Wise Men do by their Wives; Make the Best of them… Wise Men Love their Wives; and what ill-conveniences they find in them they bury; and what Vertues they are inrich't with they Admire and Magnifie. And thus you must do by your Bills for there is not doing without them; if you Divorce or Disseize yourselves of them you are undone.

Hence the American colonies developed the practice of adjusting note issue to stimulate business or countervail a recession. They believed that there is a very close relationship between money, prices and business conditions and that the appropriate note issue would greatly stimulate business. Their efforts were made easier by the fact that there was no bank-issued money.

In England, after the Napoleonic Wars, the first great debate about monetary reform occurred, with persons such as Joseph Lowe, John Rooke and Poulett Scrope, proposing a ‘managed currency’, the volume of which was to be controlled according to changing prices in such a way as to keep the price level steady. Similarly, Henry Thornton's *Paper Credit* (1802) argued that contraction or expansion of the money supply had real effects on the level of economic activity. In the 1840s, Thomas Attwood claimed that if Britain's coinage ‘were accommodated to man and man to our coinage then world would be capable of multiplying its production to an unlimited extent’. However, David Ricardo's and John Stuart Mill's failure to appreciate that credit expansion might stimulate the level of economic activity, rather than just increase prices, dominated economic thinking for the rest of the 19th century (see Viner, 1937).

This opened the door for the monetary cranks, who argued that money did matter. Their main inspiration came from the underconsumptionist tradition. A number of authorities have emphasized that underconsumptionist literature is difficult to categorize (for example, Schumpeter, 1954, p. 740; Haberler, 1937, ch. 5; Bleaney, 1976, ch. 1). Still, the argument that there is a permanent deficiency of purchasing power produced all kinds of suggestions as to how such a deficiency could be remedied.

In the interwar period, underconsumptionist ideas fell on particularly receptive ears. Many persons, particularly those concerned with high unemployment, were prepared to believe that the schemes of the monetary cranks would increase demand and hence create jobs. The quantity of pamphlet literature on monetary reform over this era is thus enormous. A common argument was that because the First World

War was financed by printing money, the same method could be used to eliminate unemployment. Opposition to the gold standard usually accompanied this argument.

Academic discussion of monetary matters was disparate and disputatious (see, for example the famous debate between F.A. von Hayek and P. Sraffa in the *Economic Journal*, March–June 1932) and this was seized upon by the monetary reformers, who sought to penetrate what they claimed were the obfuscations of the academics. They also pointed to the fact that discussion of money and banking tended to be confined to tendentious tomes written for bank employees, while economic theory textbooks devoted little space to arguments against Say's Law.

Major C.H. Douglas was probably the best-known reformer in English-speaking countries in this era (see Douglas, 1924) but there were many, many others who wrote on monetary reform. These included: A.H. Abbati, who attracted the interest of John Maynard Keynes and D.H. Robertson; Sir Normal Angel, whose set of cards *The Money Game* was widely used in high schools in Britain and the US; W.T. Foster and W. Catchings, who were probably the best known US reformers; and Frederick Soddy of Oxford University, who, after being awarded the Nobel Prize for chemistry, set out to solve the money problem inspired by John Ruskin's *Unto this Last* (1862) and an Australian invention. Soddy argued that the gold standard could be replaced with a machine based on the automatic totalizator at Sydney's Randwick Racecourse (Soddy, 1931). Cole (1933) discusses some of this literature.

Strangely, Schumpeter (1954) contains no reference to Douglas but he does mention (pp. 1090–91) G.F. Knapp's *The State Theory of Money* (1924), which promoted similar ideas and had considerable impact in interwar Germany. For example, in the dying days of the Weimar Republic, at the suggestion of H.J. Rustow and W. Lavtenbach of the Ministry of Economics, interest-bearing tax certificates were issued in lieu of treasury bills and exchequer bonds. Employers were given these certificates if they employed additional employees and reduced the wages of existing employees (see Rustow, 1978).

With the Keynesian revolution and the increased emphasis given to monetary theory by academic economists in recent decades, the monetary cranks have largely disappeared from public debate, although underconsumptionist ideas will probably have supporters while ever there is unemployment. Any explanation of the appeal of these ideas over generations would have to invoke sociology and psychology. Such ideas found strong support because they enabled persons to impress their peers with their apparent understanding of economics, even though they had no formal training in the discipline. They offered the false hope that there were simple solutions to the complexities of modern economic life. They also transcended party political allegiances – similar passages about ‘credit slavery’ and ‘Shylocks’ can be found in Hitler's *Mein Kampf* and left-wing pamphlets of the same era. A very wide range of individuals can be opposed to private banks and the ‘Money Power’ without their opposition leading to more sophisticated political analysis. In fact, as the history of populism shows, ‘Funny Money’ beliefs provided a kind of ideological release valve.

The history of ideas contains numerous examples of the power of the phrase-monger. The simpler the panacea, the greater the chance the agitator will have of attracting a following. As the Chartist agitator Ernest Jones once advised (cited in Martin and Rubinstein, 1979, p. 43): ‘We say to the great minds of the day, come among the people, write for the people and your fame will live forever.’

Bibliography

Angell, N. 1936. *The Money Mystery: An Explanation for Beginners*. London: Dent. (*The Money Game*, a set of cards for teaching purposes, was sold in conjunction with this book.)

Bleaney, M. 1976. *Underconsumption Theories: A History and Critical Analysis*. London: Lawrence & Wishart.

Clayton, G., Gilbert, J.C. and Sedgwick, R., eds. 1971. *Monetary Theory and Policy in the 1970s*. London: Oxford University Press.

Cole, G.D.H. 1933. *What Everybody Wants to Know about Money: A Planned Outline of Monetary Problems*. London: Victor Gollancz.

Douglas, C.H. 1924. *Social Credit*. London: Eyre & Spottiswoode.

Durbin, E.F.M. 1934. *Purchasing Power and Trade Depression*. London: Chapman & Hall.

Haberler, G. 1937. *Prosperity and Depression*. Cambridge, MA: Harvard University Press.

Lester, R.A. 1939. *Monetary Experiments: Early American and Recent Scandinavian*. Princeton: Princeton University Press.

Martin, D. and Rubinstein, D., eds. 1979. *Ideology and the Labour Movement*. London: Croom Helm.

Rustow, H.J. 1978. The economic crisis of the Weimar Republic and how it was overcome. *Cambridge Journal of Economics* 2, 409–21.

Schumpeter, J.A. 1954. *A History of Economic Analysis*. London: George Allen & Unwin.

Soddy, F. 1931. *Money versus Man*. London: Elkin Mathews & Marrot.

Viner, J. 1937. *Studies in the Theory of International Trade*. London: George Allen & Unwin.

How to cite this article

Clark, David. "monetary cranks." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000198> doi:10.1057/9780230226203.1120

The New Palgrave Dictionary of Economics Online

monetary economics, history of

Robert W. Dimand

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Throughout its long history, monetary economics has been concerned with the role of money in exchange, with what determines the purchasing power of money, and with the effects of changes in the purchasing power of money on interest rates and real economic activity.

Keywords

aggregate demand; Allais, M.; Aristotle; Bagehot, W.; banking crises; Banking School; barter; Bentham, J.; bimetallism; Bullionist Controversy; Cantillon, R.; central banking; convertibility; credit cycles; Currency School; debasement of currency; distributed lags; effective demand; equation of exchange; exchange; fiat money; Fisher, I.; forced saving; free banking; Friedman, M.; Fullarton, J.; general equilibrium approach to monetary theory; gold standard; Hawtrey, R. G.; Hicks, J. R.; Hume, D.; index numbers; inflation; inflation; interest rate limits; involuntary unemployment; Jevons, W. S.; Keynes, J. M.; Malthus, T. R.; Markowitz, H. M.; Marshall, A.; medium of exchange; Mercantilism; Mill. J. S.; monetarism; monetary approach to the balance of payments; monetary economics; monetary policy rules; money; money illusion; money supply; Mun, T.; natural rate and market rate of interest; neutrality of money; nominal interest rates; open market operations; overlapping generations models; Pigou, A. C.; Plato; portfolio choice; price revolution; quantity theory of money; real balances; real bills doctrine; Ricardo, D.; Robertson, D. H.; Say's equality; Say's identity; Say's Law; Sismondi, J. C. L. S. de; Smith, A.; specie-flow mechanism; stabilization; Stockholm School; symmetallism; Taylor rule; term structure of interest rates; Thornton, H.; time preference; Tobin, J.; Tooke, T.; Turgot, A. R. J.; uncovered interest parity; unemployment; veil of money; velocity of circulation]; Walras, L; Walras's, Law; Wicksell, J. G. K.

Article

Origins of monetary economics

As with so much else in the Western tradition, theorizing about the role of money can be traced back to Plato and Aristotle in the fourth century bce, although they may have drawn on pre-Socratic philosophers whose works survive, if at all, only in fragments. In his *Republic* (1974), Plato remarked that money was a symbol devised to make exchange easier. He disapproved of gold and silver as money, preferring a currency that would have value only internally, not in external commerce. The analysis in Aristotle's *Nicomachean Ethics* (1996) and *Politics* (1984) of what constitutes just exchange led Aristotle to a more systematic discussion of a medium of exchange. His account of the functions of money, and of the properties that suit a commodity such as gold or silver to be the medium of exchange, as well as his use of the myth of Midas to distinguish between gold and wealth, influenced comparable presentations by Nicolas Oresme in about 1360 (Oresme, de Sassoferrato and Buridan, 1989), Adam Smith (1776), and, through Smith, any number of 19th-century textbooks (see Menger, 1892; Monroe, 1923). Barter might be the most basic form of exchange, but it involves accepting goods one does not wish to consume in order to make a further exchange for what is desired. Aristotle noted the convenience of a generally accepted medium of exchange in reducing the number of transactions required. He saw the convenience of stating prices in terms of the medium of exchange, and that, if a commodity is to serve as a medium of exchange, it must also be a store of value, retaining purchasing power between being received and being spent (but he did not mention the function of money as a standard of deferred payment). Precious metals provided a suitable medium of exchange because of being homogenous, divisible, portable, and sufficiently scarce to a have a high value relative to their weight, although that value could change. Unlike Plato, Aristotle viewed the weight and purity of the precious metals as the source of the purchasing power of money, with coinage just saving the inconvenience of having to weigh and assay the metals at every transaction.

The quantity theory of money, described by David Laidler (1991b) as ‘always and everywhere controversial’ and by Mark Blaug as ‘the oldest surviving theory in economics’ (Blaug et al., 1995), holds that the price level (the inverse of the purchasing power of money) depends on how large the stock of money is compared with the demand for real money balances, with the direction of causation running from money to prices (Hegeland, 1951). The quantity theory originated in the 16th century, when Martin de Azpilcueta Navarro in Salamanca in 1556 and Jean Bodin in France in 1568 identified the inflow of silver from the Spanish colonies of Mexico and Upper Peru as the cause of the rise in prices and depreciation of silver throughout Europe, a phenomenon now known as the ‘price revolution’ (Grice-Hutchinson, 1952; O'Brien, 2000). In contrast to the recognition by Navarro and Bodin of the inverse relationship between the quantity of the precious metals and their purchasing power, contemporaries such as the Seigneur de Malestroit had attributed rising prices of commodities to the debasement of various national coinages. The astronomer Copernicus had remarked earlier that money usually depreciates when it is too abundant (Grice-Hutchinson, 1952, p. 34), but Navarro and Bodin went beyond such passing insights to formulate a theory they could use to explain the observed trend of commodity prices. Later research has shown that the 16th century quadrupling of prices was also due in part to the growing output of central European silver mines and to an increase in the velocity of circulation of money as systems of payment and communication evolved, notably the use of bills of exchange.

Mercantilists also took note of the inflow of precious metals from Spain's conquest in the New World, viewing this gold and silver as the ‘sinews of war’ with which Spain could pay armies in Europe. Although both alchemy and seizure of the Spanish treasure fleet were attempted (the physicist Isaac Newton was both Master of the Royal Mint and an avid alchemist), mercantilists such as Thomas Mun advocated interventionist government policies to achieve a surplus of exports over imports as the way to bring gold and silver into a country that lacked its own mines. Mercantilists held that increased circulation of gold and silver in a country would both increase national power and stimulate real economic activity (Viner, 1937; Vickers, 1959). Isaac Gervaise (1720), Richard Cantillon (2001, written c. 1730 and published posthumously in 1755), and, most fully and forcefully, David Hume (1752) used the quantity theory of money to develop the specie-flow mechanism of international payments adjustment that rendered such mercantilist schemes futile. An increase in gold and silver circulating in a country, whether due to colonial conquests, discovery of new mines, or a trade surplus engineered by tariffs on imports and bounties on exports, would increase spending. Although Hume recognized that one immediate, temporary effect of such increased spending would be to stimulate production (see Humphrey, 1993) in due course prices and wages would rise, making domestic goods more expensive in relation to foreign goods. This would reduce exports and increase imports, eliminating the trade surplus, so that the only lasting result would be the misallocation of resources caused by tariffs, bounties and quotas. For Adam Smith (1776), a small open economy such as that of Scotland took prices under the gold standard as given by the world market, so the balance of payments adjustment would take place without any change in the relative price of foreign and domestic goods. An excess supply of money in a country would directly cause more imports and more exportable goods to be purchased domestically (and the contrary in a country with an excess demand for money) unless the world's supply of monetary metal was distributed across countries in proportion to their demand for money. Humphrey (1993) and Laidler (2003, ch. 1) show that Smith's analysis bore a closer resemblance than that of Hume to the modern monetary approach to the balance of payments.

From Aristotle and the Bible onwards, payment of interest on loans had been condemned as usury on the grounds that it was unnatural for gold (‘barren metal’) to breed and that interest violated justice (exchange of equal values), as the amount of money repaid exceeded the initial loan. Cantillon, Hume, A. R.J. Turgot, and Jeremy Bentham argued for the legitimacy of an interest rate set by market forces of supply and demand, with Turgot invoking time preference to point out that the amount of money lent and the larger amount of money repaid represented the same present value. Contrary to his general stand against government intervention, Adam Smith (1776) endorsed legal limits on interest to prevent high-risk lending for speculation and reckless consumption, and was rebuked for inconsistency by the young Bentham (West, 1997).

Monetary controversies in classical economics

Monetary theory was advanced by two British debates, the Bullionist Controversy, which surrounded the suspension of the convertibility of Bank of England notes into gold from 1797 to 1821, and the clash between the Banking School and the Currency School in the 1840s leading up to and following the Bank Act revision that separated and regulated the Bank of England's Issue Department (whose liabilities were bank notes, with gold held in reserve) and Banking Department (whose liabilities were deposits, with

Bank of England notes held in reserve). During the suspension of convertibility during the Napoleonic Wars, Henry Thornton (1802) and, from 1809 onwards, David Ricardo (1810) argued the high price of bullion and foreign exchange showed that the Bank of England had engaged in over-issue of bank notes, raising commodity prices and depreciating the pound sterling (Fetter, 1965; Marcuzzo and Rosselli, 1991). Christiernin (1761) had made a similar argument in Sweden, but appears not to have been known in Britain. Thornton was the leading figure on a House of Commons Select Committee on the High Price of Gold Bullion in 1810 that adopted this view in the Bullion Report, but the directors of the Bank of England persuaded the full House not to act on the committee's report. The directors, invoking the authority of Adam Smith, held that they could not have been guilty of any inflationary overissue of notes beyond what the needs of trade required as long as they issued notes only by discounting bills of exchange created by genuine commercial transactions, rather than financial speculation. This version of the real bills doctrine ignored Smith's assumption that bank notes were convertible into gold upon demand, so that any increase in the quantity of notes sufficient to depress their value below their gold par would cause the excess notes to be redeemed. Without convertibility as a constraint on overissue, the demand for bills would be unbounded as long as the discount rate was less than the prevailing rate of profit. The distinction between real and fictitious bills also failed to recognize that the length of time a bill was discounted need not correspond to the length of time goods were in process (Mints, 1945; Laidler, 2003; Davis, 2005).

The depression that accompanied the end of the Napoleonic Wars and Britain's subsequent return to the gold standard stimulated a debate over the possibility of a general glut of commodities. Thomas Robert Malthus and J.C.L. Simonde de Sismondi attributed the depression to an insufficiency of effective demand. Malthus's argument was acclaimed by John Maynard Keynes a century later, although, unlike Keynes, Malthus did not distinguish between a decision to save and a decision to invest (see Keynes's 1933 essay on Malthus). Ricardo and Jean-Baptiste Say upheld Say's (or James Mill's) Law of Markets, denying the possibility of a general glut of commodities or an insufficiency of aggregate effective demand, since a commodity was offered for sale only with the intention of acquiring the means to purchase some other commodity, not with intent to hoard money, which is only a medium of exchange (Say was not quite as unambiguous as James Mill). Ricardo and Say recognized that unemployment would occur during the adjustment to a major change in the mix of commodities demanded, as the end of the Napoleonic Wars curtailed military and naval spending and as the purchasing power of money changed: Ricardo was prepared to accept restoration of gold convertibility at the depreciated parity, to avoid the price deflation associated with going back to the pre-war parity, and Say endorsed public works to employ those who would otherwise be jobless during the transition period. But, according to Ricardo, Say and James Mill, such distress resulted from a temporary mismatch between the mix of commodities produced and those demanded, with excess supply in some markets and excess demand in others, not from generalized excess supply.

Throughout the 19th century, classical economists such as John Stuart Mill struggled to formulate an acceptable version of the law of markets that would be stronger than what Oskar Lange later labelled Say's Equality but weaker than what Lange called Say's Identity (Corry, 1962; Sowell, 1972; Baumol, 1977; 1999; Davis, 2005). Say's equality, which held that at equilibrium prices the value of excess demand sums to zero across all markets except that for money, is a trivial implication of the market-clearing equilibrium condition that at market-clearing prices supply equals demand in each market. Say's

identity, which held that at any prices the value of excess demand always sums to zero across all markets except money, implies (when combined with the summation of individual budget constraints) that money demand always equals the money supply at any prices, which leaves the absolute price level (the inverse of the purchasing power of money) indeterminate. In the 1870s, Leon Walras reformulated Say's Law as what Lange termed Walras's Law: the value of aggregate excess demand summed over all markets (including money) is identically zero, from the summation of individual budget constraints (the net value of each individual's transactions is at most zero, since people must pay for their purchases) plus local non-satiation (so that no one is willing to throw away purchasing power). Robert Clower (1984), seeking to understand Keynes's rejection of Say's Law of Markets, argued that Walras's Law only applies to notional demands, not to quantity-constrained effective demands when markets do not clear (in Keynes's case, the labour market): if workers cannot sell all the labour they wish at the prevailing wage rate, then the quantity of labour they cannot sell multiplied by the wage rate that they would have received should not be included in their budget constraint for demanding goods. Currency School adherents (for example, J.R. McCulloch, G.W. Norman and Lord Overstone), whose ideas shaped Sir Robert Peel's Bank Act of 1844, urged that, beyond maintaining convertibility, the Bank of England should conduct its operations so that a mixed metallic and paper currency would fluctuate in the same way that a purely metallic currency would. Building on Ricardo's presentation of the quantity theory of money and the price specie-flow mechanism, the Currency School wished the central bank to follow a stabilizing policy that would prevent gold outflows, rather than waiting for such international cash drains to bring about adjustment. The Currency School attributed the banking crises of 1825, 1832 and 1836–37 to monetary mismanagement by the Bank of England, which could have regulated the volume of coin and notes in circulation so as to stabilize prices. In contrast, Banking School writers such as Thomas Tooke and John Fullarton, drawing on Thornton, emphasized the endogeneity of the total volume of credit (financial instruments convertible into gold), of which bank notes were only a small part (Fullarton, 1836; 1845; Fetter, 1965; Arnon, 1991; Cassidy, 1998; Skaggs, 1999). Karl Marx also held that the volume of money adjusted to satisfy the equation of exchange (de Brunhoff, 1976). Elements of both Currency School and Banking School positions appeared in the writings of John Stuart Mill. The Banking School thought that the volume of credit was as likely to respond to changes in prices as to cause them, and so did not share the Currency School view of the banking system as the initiator of credit cycles. The Banking School prescription was for the Bank of England to hold a bullion reserve large enough to ride out temporary disturbances in credit and international payments. While the Currency and Banking Schools differed on the appropriate policy for a central bank, another group of writers, including Henry Dunning Macleod (1855), James Wilson of *The Economist* and Jean-Gustave Courcelle-Seneuil, opposed having a central bank with a legally protected dominant position and special privileges. Instead, they advocated a system of free banking, with the market valuing the notes of competing banks, a proposal revived by Vera Smith (1936) and later by Friedrich Hayek (1976), who had been her dissertation adviser. Walter Bagehot's *Lombard Street* (1873) established the monetary orthodoxy, emerging from the Currency School–Banking School debates, on how the central bank should manage the discount rate to maintain convertibility and its role as a lender of last resort to preserve the liquidity of the banking system, rather than simply acting in the interests of its shareholders.

The golden age of the quantity theory

In studies collected posthumously in Jevons (1884), William Stanley Jevons used index numbers, with equal weights on different commodities, to show the rise in prices following the gold rushes in California in 1849 and Australia in 1851, as did John Elliot Cairnes. Commodity prices tended downwards from 1873 to 1896 as the world's demand for real money balances grew faster than its money supply, a decline halted by the introduction of the cyanide process for extracting gold from low-grade ores and by gold discoveries in South Africa and the Klondike. Together with the return of the United States to gold convertibility of the dollar in 1873 after the issue of inconvertible greenbacks during the Civil War, this deflation contributed to bimetallist agitation that reached its peak in William Jennings Bryan's presidential campaign in 1896, in which Bryan spoke against ‘crucifying mankind on a cross of gold’. The bimetallists argued that monetizing silver as well as gold would raise the price by increasing the quantity of money, and this would have lasting real benefits. This led hard-money, classical economists such as J. Laurence Laughlin of the University of Chicago to associate the quantity theory of money with claims of long-run non-neutrality (Skaggs, 1995). In place of the quantity theory, Laughlin (1903) derived the value of money from the convertibility into gold, whose value depended on its cost of production, a view which David Glasner (1985; 2000) shows had figured alongside the quantity theory in classical political economy. The quantity theorists David Kinley (1904), Edwin Kemmerer (1907) and Irving Fisher (with Harry G. Brown, *The Purchasing Power of Money*, 1911, in Fisher, 1997, vol. 4) responded by seeking to show, contrary to Laughlin and his Chicago associates, that exogenous changes in the quantity of money explained the behaviour of prices (given the trend in money demand), and, contrary to the bimetallists, that money is neutral in the long run. These quantity theorists extended earlier statements of the equation of exchange by Simon Newcomb (to whom Fisher dedicated his 1911 book) and Sir John Lubbock. Fisher allowed currency (M) and bank deposits (M') to have different velocities of circulation, restating the equation of exchange as $MV+M'V'=PT$, where T is an index of the volume of transactions and P is the price level. To use the equation of exchange to make the case that the changing money supply explained the observed movements of US prices (rather than just having the equation as a tautology defining the velocity of circulation) required independent measures of the velocity of circulation. To estimate V, Fisher persuaded 116 people at Yale (including 113 male undergraduates) to keep daily records of their spending and cash balances. For V', the velocity of circulation of bank deposits, Fisher used linear interpolation between the estimates from two empirical studies by David Kinley counting all bank clearings in the United States for a day in 1896 (for the Comptroller of the Currency) and a day in 1910 (for the National Monetary Commission). From an Austrian perspective, Ludwig von Mises (1935) objected to the aggregative reasoning of the quantity theorists, arguing that an index number of the price level gives a distorted picture of how agents respond to prices.

Systematically developing earlier remarks by John Stuart Mill and Alfred Marshall and an article by Jacob de Haas, Irving Fisher argued in *Appreciation and Interest* (1896, in Fisher, 1997, vol. 1) that that nominal interest is the sum of real interest and the expected rate of inflation, so that only unanticipated changes in the purchasing power of money change the real interest rate and redistribute wealth. Contrary to bimetallist claims, expected inflation or deflation would have no real effects. Fisher's 1896 analysis included uncovered interest arbitrage parity (the difference between nominal interest rates in two

currencies is the expected rate of change of the exchange rate) and the expectations theory of the term structure of interest rates (variations in nominal interest on loans of different duration reflects expectations of the time-path of prices). But from *The Purchasing Power of Money* onwards, while continuing to insist on the long-run neutrality of money, Fisher argued that money was not neutral during transition periods (of up to ten years), as nominal interest adjusted only slowly to monetary shocks, and that the ‘so-called “business cycle”’ was really a ‘dance of the dollar’. While Ralph Hawtrey (1919) and Fisher advanced monetary theories of economic fluctuations, many economists in the late 19th and early 20th centuries, from Jevons on sunspot cycles to Joseph Schumpeter on clusters of innovations, emphasized real shocks and truly periodic cycles of varying lengths such as Juglar, Kondratiev and Kitchin cycles. Fisher's article, ‘A Statistical Relationship between Unemployment and Price Level Changes’ (1926) correlated unemployment with a distributed lag of past price level changes and was reprinted in the *Journal of Political Economy* in 1973 as ‘Lost and Found: I Discovered the Phillips Curve’. Fisher correlated nominal interest with a distributed lag of price changes (a version of adaptive expectations) to show the slow adjustment of nominal interest and inflation expectations (*The Theory of Interest*, 1930), resulting from what he termed *The Money Illusion* (the title of his 1928 book), the widespread tendency to think in nominal rather than real terms.

Bimetallism foundered on its insistence on fixing the relative price of gold and silver, at 15 or 16 ounces of silver per ounce of gold. As the relative market valuation changed, due to changing marginal costs of production or shifts in non-monetary demand for precious metals, one of the two metals would disappear from circulation and its coins be melted down. Alfred Marshall's (1887) suggestion of symmetallism, a unit of value consisting of a quantity of gold plus a quantity of silver (reprinted in Pigou, 1925), was more practical, but did not seem so to bimetallists or the general public. Marshall's tentative proposal to peg the monetary value of a basket of two commodities instead of just one (gold) marked a step towards a monetary policy of targeting the price level (or its rate of change) rather than the exchange rate with gold. Like Jevons (1884), Marshall suggested voluntary indexation, with contracts made in terms of a ‘standard unit of purchasing power’, which Marshall argued would reduce cyclical fluctuations (Laidler, 1991a, pp. 172–8). Irving Fisher and Senator Robert Owen attempted unsuccessfully to get such a price level target into the Federal Reserve Act of 1913. The Federal Reserve Act, influenced by J.L. Laughlin and his student H. Parker Willis, instead adopted a fixed price of gold and, inconsistent with that goal, a version of the real bills doctrine that the volume of currency and bank credit should vary pro-cyclically with the needs of trade. As Knut Wicksell (1915) and others objected, Fisher compromised his compensated dollar plan by disguising it as a version of the gold standard, with the gold weight of the dollar changed periodically to peg the dollar price of a basket of commodities, a system vulnerable to speculative attacks. By 1935, when Fisher endorsed open market operations under a floating exchange rate to achieve a price-level target, he had lost his audience.

While Fisher distinguished nominal and real interest rates, Knut Wicksell (1898; 1915) stressed the distinction between the market rate of interest, set by the banking system, and the natural rate of interest that would equilibrate desired investment and saving (Laidler, 1991a; Humphrey, 1993). As long as the market rate is less than the natural rate, entrepreneurs can profit by borrowing and investing, causing total spending to increase and prices to rise. Such a cumulative inflation would continue until the growth of loans and deposits and a drain of cash out of the banking system reduced the ratio of reserves to bank deposits, forcing banks to raise the market rate to restore their liquidity. Wicksell pointed out that in a

cashless economy, with only bank money used for transactions and no reserves held by banks, there would be no such force to automatically halt a cumulative inflation or deflation, and stability would depend on deliberate action by the monetary authority to match the market rate to the changing natural rate. To explain observed price movements, Wicksell emphasized real shocks that changed the natural rate as initiating fluctuations. Wicksell's two-rate model greatly influenced the Stockholm School (Karin Kock, Erik Lindahl, Erik Lundberg, Gunnar Myrdal, Bertil Ohlin) and John Maynard Keynes's *Treatise on Money* (1930). Recent financial innovations, diminishing the role of money as a means of payment and as an asset, have renewed attention to Wicksell's analysis of a cashless economy in which the monetary authority pursues stabilization by setting the interest rate rather than the quantity of money. The title of Michael Woodford's (2003) *Interest and Prices* deliberately echoes the title of Wicksell's (1898) *Interest and Prices* and a change of emphasis from Don Patinkin's (1965) *Money, Interest and Prices*. The ‘Taylor rule’, the influential monetary policy rule proposed by John Taylor, amounts to an attempt to set the market rate of interest equal to a Wicksellian natural rate that changes over time and is not directly observable.

Cambridge monetary theory and the Keynesian revolution

In his lectures at Cambridge, evidence to official inquiries (collected by Keynes after Marshall's death as Marshall, 1926), and manuscripts from the 1870s that half a century later formed the basis of Marshall (1923), Alfred Marshall expounded the quantity theory of money in a version that emphasized that desired cash balances are proportional to nominal income, M=kPY (see Robertson, 1922; Marget, 1938–42; Eshag, 1963; Bridel, 1987; Laidler, 1999 on Cambridge monetary economics). The Cambridge coefficient k is the reciprocal of V, the income velocity of circulation of money in the equation of exchange, so that the two versions of the quantity theory are formally equivalent, although Marshall's disciples A.C. Pigou and J.M. Keynes claimed that Cambridge discussions of the determinants of k were more choice-theoretic and less mechanical than Fisher's discussion of the determinants of velocity. Related contributions emerged from both traditions: Fisher was the first to correctly state the marginal opportunity cost of holding real money balances (1930), Keynes the first to explicitly write money demand as a function of income and nominal interest (*General Theory*, 1936). Writing in a time of floating exchange rates and Continental European hyperinflations after the First World War, the young Keynes, in *A Tract on Monetary Reform* (1923), extended Marshall's monetary economics to analyse inflation as a tax on holding money and government bonds, the social costs of inflation (both distortions from incorrectly anticipated inflation and higher transactions costs as expected inflation reduces the demand for real money balances), and covered interest arbitrage parity (the spread between spot and forward exchange rates is the difference between nominal interest in two currencies). Keynes opposed Britain's return to the gold standard at the pre-war parity in 1925 as entailing domestic deflation and, until wages declined, unemployment. Keynes's position recalled Ricardo's preference for restoring convertibility as a depreciated parity after the Napoleonic Wars. D.H. Robertson (1926), deeply Marshallian although a student of Keynes and Pigou rather than directly of Marshall, examined the effect of price level changes on saving and investment, notably how an increase in the price level causes forced saving (‘induced lacking’) to restore real money balances (Laidler, 1999).

Reflecting on Britain's stagnation after the return to gold and on the worldwide Great Depression of the

1930s, Keynes's *General Theory of Employment, Interest and Money* (1936) denied the automatic restoration of full employment in a monetary economy after a negative demand shock. Keynes lumped together economists from Ricardo to Marshall and Pigou as ‘classical’ economists who accepted Say's Law (summarized by Keynes as ‘supply creates its own demand’). Keynes subsequently clarified that he did not regard Fisher, Hawtrey, Robertson or Wicksell's Swedish followers as classical (but he did think that Wicksell himself was trying to be classical), and, as Ellis (1934) showed, German monetary theorists such as Joseph Schumpeter and L. Albert Hahn were far from classical about the real effects of an expansion of the banking system. In contrast to von Mises (1935) and Hayek (1931), who viewed depressions as necessary corrections of earlier overinvestment, Keynes held that depressions were calamities that the government and monetary authority could overcome by increasing aggregate demand, rather than relying on wage and price deflation to restore full employment. Keynes considered it crucial that wage bargains are made in money terms, so that workers concerned about relative wages might accept a price level increase to clear the labour market while quite rationally opposing money wage cuts as staggered contracts came up for renegotiation (1936, ch. 2). Wage cuts, and the associated deflation of prices, would increase demand for real money balances, exerting a contractionary effect on aggregate demand (1936, ch. 19). Keynes identified volatile private investment, resulting from fundamental uncertainty about future profitability, as the source of economic fluctuations, and, like the generations of Keynesian, New Keynesian and Post Keynesian economists after him, saw a need for management of aggregate demand to stabilize the economy.

The revival of the quantity theory of money

While Keynes was arguing the case for stabilization policy, Henry Simons of the University of Chicago made the case for rules rather than discretion in monetary policy (Simons, 1936). Keynes saw a role for government to counteract the instability resulting from volatile private spending, but Chicago quantity theorists (later called monetarists) such as Simons (1936) and Milton Friedman and his students (Friedman, 1956) blamed volatile, unpredictable monetary policy for economic instability. Keynesians invoked the Great Depression of the 1930s as demonstrating the need for government stabilization of an unstable private sector in a monetary economy, but Friedman and Anna J. Schwartz (1963) blamed the depression on a misguided Federal Reserve system that permitted a ‘great contraction’ of the money supply. Misled by the real bills doctrine, the Federal Reserve Board had not paid sufficient attention to the quantity of money. Where Keynes had emphasized the fundamental uncertainty underlying long-period expectations of profitability, Friedman (like Fisher) stressed the endogeneity of expectations of inflation: people cannot be fooled indefinitely by inflation into working more for a lower real wage that they think they are getting, because they will learn from experience (see Friedman and his critics in Gordon, 1974). Keynes worried about involuntary unemployment – an excess supply of labour because the labour market did not clear – while Friedman held that at any correctly anticipated inflation rate unemployment would be at its natural rate, reflecting voluntary investment in search and consumption of leisure. Friedman claimed in 1956 to be following a Chicago oral tradition of monetary theory taught by Frank Knight, Jacob Viner, Henry Simons and Lloyd Mints that had replaced J. Laurence Laughlin's opposition to the quantity theory. Don Patinkin (1981) and David Laidler (2003), who both held Chicago Ph.D.s, argued that Friedman overstated the purely Chicago sources of his monetarism: Friedman's

teachers had taught the works of non-Chicago quantity theorists such as Fisher as well as Keynes's earlier Marshallian *Tract on Monetary Reform* (1923) and his Wicksell-influenced *Treatise on Money* (1930). Friedman took a course in which the main textbook was Keynes's *Treatise*, which Keynes's detailed and extensive contribution to monetary analysis. Fisher had advocated a monetary policy rule (a price level target, rather than the constant of money growth proposed by Friedman), while Keynes's *Tract* was as attentive as any Chicago monetarist to the social costs of inflation. A key element of Friedman's monetarism, money demand as a function of a small list of variables, had first appeared in Keynes's *General Theory*. There were also parallel, independent revivals of the quantity theory of money far from Chicago, such as that associated with Marius Holtrop, longtime president of the Netherlands central bank (De Jong, 1973).

Integrating the theory of money into general economic theory

Rationalizing the use of money has been a problem in the development of general equilibrium theory: if markets are complete, or all debts will be repaid with certainty, there is no need for a particular asset to be singled out as a generally accepted means of payment. Irving Fisher's 1892 dissertation introduced general equilibrium analysis in North America, but he did not integrate his later monetary economics into a general equilibrium framework. Leon Walras, the founder of general equilibrium theory, wrote on the theory of money (for example, Walras, 1886), starting with the equation of exchange and later discussing desired cash balances, *encaisse désirée*, but simply assumed that monetary exchange is superior to barter, rather than demonstrating that the use of money reduces transactions costs: ‘In Walras's economy, agents hold money not out of choice but of a technological necessity’ (Bridel, 1997, p. 119; see also Patinkin, 1965, pp. 531–72). In Walras's analysis, prices were stated in terms of a particular commodity, the *numéraire*, but it was not clear why transactions should use that commodity. The idea that money is only a veil over the real side of the economy long predates the introduction of the term ‘veil of money’ in English by Dennis Robertson (1922) and of ‘neutrality of money’ by Hayek (1931): (see Pigou, 1949; Patinkin and Steiger, 1989). Don Patinkin (1965) argued that a long list of classical and neoclassical economists postulated, at least implicitly, an invalid dichotomy between the real and nominal sides of the economy, in which an equi-proportional change in all money prices (so that no relative prices changed) would not affect the excess demands for commodities. Such a dichotomy would exclude the real balance effect that would bring the general price level to equilibrium. The valid dichotomy would hold that an equi-proportional change in all money prices, the quantity of money, and any exogenous nominal variables (such as quantities of government bonds) would have no real effects.

John Hicks (1935) set the agenda for much later work integrating the theory of money into the more general theory of value, seeking choice-theoretic explanations of why fiat money, not backed by convertibility into a commodity such as gold or silver, has a positive purchasing power, and why people choose to hold part of their wealth in money (either non-interest-bearing high-powered money or highly liquid close substitutes paying low rates of interest) rather than in alternative assets that pay a higher rate of return. Following Hicks's argument for treating the decision to hold money as part of the allocation of wealth across a portfolio of assets, James Tobin (1958) introduced money as a riskless asset (at least in nominal terms) into Harry Markowitz's theory of portfolio choice. Risk-averse individuals would divide

their wealth between money (zero return, zero risk) and a portfolio of risky assets with positive expected return. Each investor would combine risky assets in the same proportions, differing from other investors only in the fraction of wealth held in the riskless asset. If returns were normally distributed or investors had quadratic loss functions, this portfolio choice could be conveniently captured by a two-dimensional diagram (the mean and standard distribution of portfolio returns), and if investors had constant relative risk aversion, the share of wealth held in each asset (including money) would be independent of the level of wealth (see Tobin, 1958; 1969; Tobin and Golub, 1998). However, money is a risky asset in real terms, as its purchasing power may be eroded by inflation, and is dominated in rate of return by such short-term, highly liquid assets as Treasury bills, which, like money, have no default risk. While Treasury bills have some nominal risk, since a rise in nominal interest would lower their market price, this risk is limited by the short maturity of the bills. Tobin (1969) extended his portfolio approach to a ‘general equilibrium approach to monetary theory’ that treated money as one of a range of imperfectly substitutable assets whose rates of return are determined simultaneously, with an adding-up constraint that asset demands sum to total wealth, but without assuming continuous clearing of non-financial markets (Tobin, 1971; Tobin and Golub, 1998).

Another approach to a choice-theoretic explanation of demand for fiat money assumes that money must be used as a means of payment and that it is costly to trade between money and interest-bearing assets, so that individuals trade off the interest forgone by holding money against the transaction costs (including the value of one's time spent going to the bank) incurred by having to liquidate interest-bearing assets when having to make payments. Maurice Allais in 1947, William Baumol in 1952, and James Tobin in 1956 independently derived the square-root rule for this inventory approach to the transactions demand for money by minimizing the total cost of cash management, forgone interest plus transactions costs (see Allais, 1947, pp. 238–41; Tobin and Golub, 1998), unaware that Francis Ysidro Edgeworth (1888), followed by Wicksell (1898, pp. 57–8), had derived a similar square-root rule for the demand for reserves by banks given randomness in withdrawals of deposits.

Another explanation for a positive value of fiat money is provided by overlapping generations (OLG) models, pioneered independently by Allais (1947) and Paul Samuelson (1958). In OLG models, agents live for two periods, but produce consumption goods only when young. The young trade goods to the old in return for money in anticipation of being able to exchange that money for goods in the next period when they themselves are old. Such models explain the existence of positive-valued fiat money on the assumption that no other assets exist. Other efforts to provide microeconomic foundations for fiat money emphasize monitoring costs and default risks, so that liabilities of a single, more easily monitored monetary authority are less risky than private promissory notes and therefore more acceptable as means of payments.

The long history of monetary economics reveals several recurring issues: why fiat money has value, how the real and monetary sides of the economy are related, whether a central bank should follow a rule (and if so which rule) or have discretion (or whether a central bank should even exist), is the lender of last resort function consistent with a policy rule, whether money has a special role or is just one of many assets and forms of credit, how should monetary exchange be incorporated in the general theory of value. Monetary analysis has also been focused and stimulated by external events and current policy issues: the ‘price revolution’ of the 16th century, the high price of bullion while the convertibility of Bank of England notes was suspended during the Napoleonic Wars, the Bank of England's charter

coming up for renewal in 1844 after several banking crises, the decline in the purchasing power of gold following the California and Australian gold rushes and its appreciation from 1873 to 1896, the Continental European hyperinflations after the First World War, Britain's return to the gold exchange standard at the pre-war parity in 1925, and the Great Depression.

See Also

- Banking School, Currency School, Free Banking School
- bullionist controversies (empirical evidence)
- equation of exchange
- natural rate and market rate of interest
- quantity theory of money
- real bills doctrine versus the quantity theory

Bibliography

Allais, M. 1947. •*conomie et intérêt*. Paris: Librairie des Publications Officieles.

Aristotle. 1984. *The Politics*, tr. Carnes Lord. Chicago: University of Chicago Press.

Aristotle. 1996. *The Nicomachean Ethics*, tr. Harris Rackham Ware, Herts., UK: Wordsworth Editions.

Arnon, A. 1991. *Thomas Tooke, Pioneer of Monetary Theory*. Aldershot, UK, and Brookfield, VT: Edward Elgar.

Bagehot, W. 1873. *Lombard Street*. In Bagehot (1974–86).

Bagehot, W. 1974–86. *The Collected Works of Walter Bagehot*. London: The Economist.

Baumol, W.J. 1977. Say's (at least) eight laws, or what Say and James Mill may really have meant. *Economica* NS 44, 145–62.

Baumol, W.J. 1999. Retrospectives: Say's law. *Journal of Economic Perspectives* 13, 195–204.

Blaug, M., Eltis, W., O'Brien, D., Patinkin, D., Skidelsky, R. and Wood, G.E. 1995. *The Quantity Theory of Money from Locke to Keynes and Friedman*. Aldershot, UK, and Brookfield, VT: Edward Elgar.

Boyer, J.de. 2003. *La pensée monétaire: Histoire et analyse*. Paris: •ditions Les Solos.

Bridel, P. 1987. *Cambridge Monetary Thought: The Development of Saving-Investment Analysis*. Basingstoke: Macmillan.

Bridel, P. 1997. *Money and General Equilibrium Theory: From Walras to Pareto (1870–1923)*. Cheltenham: Edward Elgar.

Brunhoff, S. de 1976. *Marx on Money*, tr. M. Goldbloom. New York: Urizen.

Cantillon, R. 2001. *Essay on the Nature of Commerce in General*, tr. H. Higgs. New Brunswick, NJ: Transaction.

Cassidy, M. 1998. The development of John Fullarton's monetary thought. *European Journal of the History of Economic Thought* 5, 509–36.

Christiernin, P.N. 1761. *Lectures on the High Price of Foreign Exchange in Sweden*, tr. in R.V. Eagly, ed., *The Swedish Bullionist Controversy*. Philadelphia: American Philosophical Society, 1967.

Clower, R.W. 1984. *Money and Markets: Essays by Robert W. Clower*, ed. Donald A. Walker. Cambridge: Cambridge University Press.

Corry, B. 1962. *Money, Saving and Investment in English Economics 1800–1850*. London: Macmillan.

Davis, T. 2005. *Ricardo's Macroeconomics: Money, Trade Cycles and Growth*. Cambridge: Cambridge University Press.

De Jong, F.J. 1973. *Developments of Monetary Theory in the Netherlands*. Rotterdam: Rotterdam University Press.

Edgeworth, F.Y. 1888. Mathematical theory of banking. *Journal of the Royal Statistical Society* 51, 113–27.

Ellis, H.S. 1934. *German Monetary Theory 1905–1933*. Cambridge, MA: Harvard University Press.

Eshag, E. 1963. *From Marshall to Keynes: An Essay on the Monetary Theory of the Cambridge School*. Oxford: Basil Blackwell.

Fetter, F.W. 1965. *The Development of British Monetary Orthodoxy 1797–1875*. Cambridge, MA: Harvard University Press.

Fisher, I. 1892. *Mathematical Investigations in the Theory and Value of Prices*. New York: Macmillan In Fisher (1997), vol. 1.

Fisher, I. 1896. *Appreciation and Interest*. In Fisher (1997), vol. 1.

Fisher, I. and Brown, H. 1911. *The Purchasing Power of Money*. In Fisher (1997), vol. 4.

Fisher, I. 1926. ‘A statistical relationship between unemployment and price level changes’. In *International Labour Review* 13, 785–92. Repr. 1973 as ‘Lost and found: I discovered the Phillips curve’, *Journal of Political Economy* 81, 496–502. Also in Fisher (1997), vol. 8.

Fisher, I. 1928. *The Money Illusion*. In Fisher (1997), vol. 8.

Fisher, I. 1930. *The Theory of Interest*. In Fisher (1997), vol. 9.

Fisher, I. 1997. *The Works of Irving Fisher*, 14 vols. ed. W.J. Barber, assisted R.W. Dimand and K. Foster. London: Pickering & Chatto.

Friedman, M. 1956. *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States 1867–1960*. Princeton, NJ: Princeton University Press for the NBER.

Fullarton, J. 1836. Response to a proposal for a bank of India. Repr. 1998 in *European Journal of the History of Economic Thought* 5, 480–508.

Fullarton, J. 1845. *Regulation of Currencies of the Bank of England*. New York: Augustus M. Kelley, 1969.

Gervaise, I. 1720. *The System or Theory of Trade of the World*. London: J. Roberts; repr. Baltimore: Johns Hopkins Press, 1954.

Glasner, D. 1985. A reinterpretation of classical monetary theory. *Southern Economic Journal* 52, 46–68.

Glasner, D. 2000. Classical monetary theory and the quantity theory. *History of Political Economy* 32, 39–59.

Gonnard, R. 1936. *Histoire des doctrines monétaires, dans ses rapports avec l'histoire des monnaies*, 2 vols. Paris: Sirey.

Gordon, R.J. 1974. *Milton Friedman's Monetary Framework: A Debate with his Critics*. Chicago: University of Chicago Press.

Grice-Hutchinson, M. 1952. *The Salamanca School: Readings in Spanish Monetary Theory 1544–1605*.

Oxford: Clarendon Press.

Guggenheim, T. 1989. *Preclassical Monetary Theories*. London and New York: Pinter.

Hawtrey, R.G. 1919. *Currency and Credit*. London: Longmans, Green, 3rd edn. 1934.

Hayek, F.A. 1931. *Prices and Production*. London: Routledge.

Hayek, F.A. 1976. *The Denationalisation of Money*. London: Institute of Economic Affairs.

Hegeland, H. 1951. *The Quantity Theory of Money*. Göteborg: Elanders Boktryckeri; New York: Augustus M. Kelley, 1969.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* NS 2, 1–19.

Hume, D. 1752. *Writings on Economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.

Humphrey, T.M. 1993. *Money, Banking, and Inflation: Essays in the History of Economic Thought*. Aldershot, UK, and Brookfield, VT: Edward Elgar.

Jevons, W.S. 1875. *Money and the Mechanism of Exchange*. New York: D. Appleton, 1897.

Jevons, W.S. 1884. *Investigations in Credit and Prices*, ed. H.S. Foxwell. London: Macmillan.

Kemmerer, E.W. 1907. *Money and Credit Instruments in their Relation to General Prices*. New York: Henry Holt.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. In Keynes (1971–89), vol. 4.

Keynes, J.M. 1930. *Treatise on Money*. In Keynes (1971–89), vols 5 and 6.

Keynes, J.M. 1933. Robert Malthus: the first of the Cambridge economists. In J.M. Keynes, *Essays in Biography*, London: Macmillan. Repr. in Keynes (1971–89), vol. 9.

Keynes, J.M. 1936. *General Theory of Employment, Interest and Money*. In Keynes (1971–89), vol. 7.

Keynes, J.M. 1971–89. *Collected Writings of John Maynard Keynes*, 30 vols. ed. D.E. Moggridge and E. A.G. Robinson. London: Macmillan, and New York: Cambridge University Press, for the Royal Economic Society.

Kinley, D. 1904. *Money, a Study of the Theory of the Medium of Exchange*. New York: Macmillan.

Laidler, D. 1991a. *The Golden Age of the Quantity Theory*. Princeton, NJ: Princeton University Press.

Laidler, D. 1991b. The quantity is always and everywhere controversial – why? *Economic Record* 67, 289–306.

Laidler, D. 1999. *Fabricating the Keynesian Revolution*. Cambridge: Cambridge University Press.

Laidler, D. 2003. *Macroeconomics in Retrospect: Selected Essays*. Cheltenham: Edward Elgar.

Laughlin, J.L. 1903. *The Principles of Money*. New York: Scribner.

Lowry, S. Todd. 1987. *The Archaeology of Economic Ideas: The Classical Greek Tradition*. Durham, NC: Duke University Press.

Macleod, H.D. 1855. *The Theory and Practice of Banking*, 5th edn. London: Longmans, Green, 1893.

Marcuzzo, M.C. and Rosselli, A. 1991. *Ricardo and the Gold Standard: The Foundations of the International Monetary Order*. London: Macmillan.

Marget, A.W. 1938–42. *The Theory of Prices: A Re-examination of the Central Problems of a Monetary Theory*, 2 vols. New York: Augustus M. Kelley, 1966.

Marshall, A. 1887. Remedies for fluctuations in general prices. *Contemporary Review*, reprinted in *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan, 1925.

Marshall, A. 1923. *Money, Credit and Commerce*. London: Macmillan.

Marshall, A. 1926. *Official Papers*, ed. J.M. Keynes. London: Macmillan.

Menger, C. 1892. On the origin of money. *Economic Journal* 2, 239–55.

Mints, L. 1945. *A History of Banking Theory in Great Britain and the United States*, 5th edn. Chicago: University of Chicago Press, 1970.

Mises, L. Von. 1935. *The Theory of Money and Credit*, tr. H. Batson. London: Cape.

Monroe, A.E. 1923. *Monetary Theory before Adam Smith*. New York: Augustus M. Kelley, 1969.

O'Brien, D.P. 2000. Bodin's analysis of inflation. *History of Political Economy* 32, 267–92.

Oresme, N., de Sassoferrato, B. and Buridan, J. 1989. *Traité des monnaies et autres écrits monétaires du XIVe siècle*, ed. C. Dupuy, tr. F. Chartrain. Lyon: La Manufacture.

Patinkin, D. 1965. *Money, Interest and Prices*, 2nd edn. New York: Harper & Row.

Patinkin, D. 1981. *Essays on and in the Chicago Tradition*. Durham, NC: Duke University Press.

Patinkin, D. and Steiger, O. 1989. In search of the ‘Veil of Money’ and the ‘Neutrality of Money’: a note on the origin of terms. *Scandinavian Journal of Economics* 91, 131–46.

Pigou, A.C. 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Pigou, A.C. 1949. *The Veil of Money*. London: Macmillan.

Plato. 1974. *The Republic*, tr. G.M.A. Grube. Indianapolis: Hackett Publishing.

Ricardo, D. 1810. *The High Price of Bullion, a Proof of the Depreciation of Bank Notes*. London: John Murray, 4th edn 1811. Repr. in Ricardo (1951–73), vol. 3.

Ricardo, D. 1951–73. *Works and Correspondence of David Ricardo*, Cambridge: Cambridge University Press 11 vols, ed. P. Sraffa and M.H. Dobb.

Rist, C. 1938. *Histoire des doctrines relative au crédit et à la monnaie depuis John Law jusqu’à nos jours*. Paris: Sirey. Tr. as *History of Monetary and Credit Theory from John Law to the Present Day*, New York: Augustus M. Kelley, 1966.

Robertson, D.H. 1922. *Money*. Cambridge: Cambridge Economic Handbooks.

Robertson, D.H. 1926. *Banking Policy and the Price Level*. London: P.S. King.

Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66, 467–82.

Simons, H.C. 1936. Rules versus authorities in monetary policy. *Journal of Political Economy* 44, 1–30.

Skaggs, N.T. 1995. The methodological roots of J. Laurence Laughlin's anti-quantity theory of money and prices. *Journal of the History of Economic Thought* 17, 1–20.

Skaggs, N.T. 1999. Changing views: twentieth-century opinion on the banking school-currency school controversy. *History of Political Economy* 31, 361–91.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. E. Cannan. New York: Random House, 1937.

Smith, V.C. 1936. *The Rationale of Central Banking*. London: P.S. King.

Sowell, T. 1972. *Say's Law: An Historical Analysis*. Princeton, NJ: Princeton University Press.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*, with an introduction by F.A. Hayek. London: George Allen & Unwin, 1939. New York: Augustus M. Kelley, 1965.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, 65–86.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1, 15–29.

Tobin, J. 1971. *Essays in Economics*, vol. 1, *Macroeconomics*. Chicago: Markham.

Tobin, J. and Golub, S. 1998. *Money, Credit, and Capital*. New York: McGraw-Hill.

Vickers, D. 1959. *Studies in the Theory of Money 1690–1776*. New York: Chilton.

Viner, J. 1937. *Studies in the Theory of International Trade*. London: George Allen & Unwin.

Walker, D.A. 1984. *Money and Markets: Selected Essays of Robert Clower*. Cambridge: Cambridge University Press.

Walras, L. 1886. *Théorie de la monnaie*. Paris: •ditions Larose et Forcel.

West, E.G. 1997. Adam Smith's support for money and banking regulation: a case of inconsistency. *Journal of Money, Credit and Banking* 29, 127–35.

Wicksell, K.G. 1898. *Interest and Prices*, tr. R.F. Kahn. London: Macmillan, 1936.

Wicksell, K.G. 1915. *Lectures on Political Economy*, vol. 2, *Money*, tr. E. Claasen. London: Routledge, 1935.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, NJ: Princeton University Press.

How to cite this article

Dimand, Robert W. "monetary economics, history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000394> doi:10.1057/9780230226203.1121

The New Palgrave Dictionary of Economics Online

monetary overhang

Holger C. Wolf

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A monetary overhang emerges when individuals jointly hold more money than they wish and all adjustment processes are rendered unavailable through price and quantity controls. While monetary overhangs can in principle be eliminated through increased real money demand, their magnitude in practice typically implies a resolution through a reduction in real money supply through a cut in the nominal money supply or through higher prices. The former is impeded by the difficulty of estimating the appropriate reduction, the latter risks triggering sustained inflation in the presence of distorted relative wage and price structures.

Keywords

forced saving; inflation; monetary overhang; money supply; price control; price liberalization; repressed inflation; velocity of circulation

Article

In functioning market economies, an excess of nominal money supply over nominal money demand is resolved through a combination of price, interest rate and real income changes. If these adjustment mechanisms are effectively blocked, a *monetary overhang* may emerge. Periods of pervasive monetary overhangs occurred in 1940s Europe (Gurley, 1953; Ames, 1954; Dornbusch and Wolf, 2001) and in the final period of some centrally planned economies, though for the latter episodes the magnitude of monetary overhangs – and thus the degree to which they contributed to rapid inflation in the aftermath of liberalization – has been debated (Nuti, 1989; Cochrane and Ickes, 1991; Chawluk and Cross, 1997).

A pure monetary overhang requires three conditions. Individuals (*a*) face a binding upper limit on nominal expenditures on goods and services (typically reflecting rationing of goods at controlled prices), (*b*) face binding limits on the purchase of (non-monetary) assets, and (*c*) are holding monetary balances that exceed the levels they would choose to hold in the absence of restrictions on goods and asset

purchases. In practice, for a number of reasons discussed below, these constraints are unlikely to bind absolutely for all individuals; the term monetary overhang is hence also used more loosely to describe situations of extensive constraints on monetary spending.

First, access to unofficial markets may allow consumers a choice between converting monetary balances into goods at the higher unofficial price (hidden inflation) and holding cash balances (possibly in expectation of greater availability of rationed products at the lower official price in the future). As access to black markets is often limited and subject to penalties, the aggregate situation may still be described as a monetary overhang. Second, individuals may be able to convert cash into savings accounts. If interest rates are controlled, a situation may arise in which individuals prefer buying more goods at controlled prices to holding either cash or deposits, but, unable to buy goods, prefer the interest-bearing asset to cash. In this setting, the overhang situation persists, but now becomes a broader financial asset overhang (forced savings).

A monetary overhang – which might be alternatively characterized as a situation of excess nominal money supply, of below equilibrium prices (repressed inflation) and of below equilibrium velocity – can be eliminated by a combination of (*a*) a cut in the nominal money supply, (*b*) an increase in prices, (*c*) a decrease in equilibrium velocity, and (*d*) an increase in output.

In practice, the degree of disequilibrium is typically such that an increase in money demand through the third and fourth channel does not provide more than a partial solution. In episodes of often substantial uncertainty, higher nominal interest rates on demand deposits are unlikely to elicit pronounced increases in desired holdings and may, moreover, adversely affect stability in financial sectors often characterized by significant non-performing loans accrued during the period of price and interest rate controls. Rapid output growth following a return to free prices has at times acted as an anti-inflation force in a post-monetary-reform period, but rarely suffices to raise money demand sufficiently.

Severe monetary overhangs consequently tend to be cured by a reduction in the real money supply, either through an increase in the price level measured from the controlled price baseline (some black market prices may well fall after price liberalization) or through a cut in the nominal money supply (typically accompanied by the removal of price controls).

A cut in money supply (often embedded in a more comprehensive reform package) may be voluntary, for instance through the issue of bonds (with fiscal implications), or involuntary, either through a straight cancellation of part of the outstanding monetary balances or a forced conversion into public assets (again with associated fiscal implications). In principle, the cut in the nominal money supply can be set so that the post-reform equilibrium price level coincides with the pre-reform controlled price level. Determining the necessary cut requires estimates of the reform-induced change in velocity and output levels. The combination of extensive economic distortions in the pre-reform period, possible responses to anticipated monetary reform and the endogeneity of the post-reform developments to the success of the reform renders this estimation highly challenging. In economies with a recent market experience, historical velocity provides a useful baseline. Alternatively, velocity estimates can be based on comparable market economies. On the implementation side, the difficulty can be overcome by a two-stage approach combining an outright cancellation of part of the nominal money supply with a freeze on a further part, with an option to either cancel or release the frozen balances at a future point depending on the post-reform evolution of output and velocity.

Price liberalization relies on market forces to restore monetary equilibrium and avoids the need to estimate the extent of the overhang. If price controls kept prices for all goods below their equilibrium by

the same proportion, the monetary overhang can in principle be resolved with a one-time proportionate jump in all prices. In practice, the disequilibrium price level typically combines with a disequilibrium relative price structure. Price liberalization may then lead to a period of inflation depending on the wage and price setting structures, possibly reinforced by an adverse fiscal impact of inflation.

See Also

- command economy
- forced saving
- inflation
- inflation dynamics
- rationing

Bibliography

Ames, E. 1954. Soviet bloc currency conversions. *American Economic Review* 44, 339–53.

Chawluk, A. and Cross, R. 1997. Measures of shortage and monetary overhang in the Polish economy. *Review of Economics and Statistics* 79, 105–15.

Cochrane, J. and Ickes, B. 1991. Inflation stabilization in reforming socialist economies. *Comparative Economic Studies* 33, 97–122.

Dornbusch, R. and Wolf, H. 2001. Curing a monetary overhang. In *Money, Capital Mobility, and Trade: Essays in Honor of Robert A. Mundell*, ed. G.A. Calvo, M. Obstfeld and R. Dornbusch. Cambridge, MA: MIT Press.

Gurley, J. 1953. Excess liquidity and European monetary reforms. *American Economic Review* 43, 76–100.

Nuti, D. 1989. Hidden and repressed inflation in soviet type economies. *Contributions to Political Economy* 5, 37–82.

How to cite this article

Wolf, Holger C. "monetary overhang." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000409> doi:10.1057/9780230226203.1123

The New Palgrave Dictionary of Economics Online

monetary policy, history of

Michael D. Bordo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Monetary policy has evolved over the centuries, with the development of the money economy. To implement monetary policy the monetary authority uses its policy instruments (short-term interest rates or the monetary base) to achieve its goals of low inflation and real output close to potential. This article surveys the origins of monetary policy from the classical gold standard to the evolution of central banks and their quest for goal independence, documenting the evolution of the goals, instruments and intermediate targets of monetary policy, and surveying the development of theories of monetary policy, including the debate over rules versus discretion.

Keywords

assignats ; Bank of England; bank rate; Banking School; Bretton Woods system; Bullionist Controversy; central bank independence; commitment; convertibility; credit rationing; Currency School; debasement; discount rate; Federal Reserve System; fiat money; fiduciary money; fiscal policy; gold standard; Great Depression; inflation; inflation targeting; interest rate; Law, J.; lender of last resort; monetarism; monetary base central bank; monetary policy targets; monetary policy, history of; money; money supply; natural rate of interest; open market operations; Phillips curve; price stability; quantity theory of money; real bills doctrine; real rate of return on capital; rediscounting; reserve requirements; rules versus discretion; seigniorage; sterilization; Taylor rule; time consistency; velocity of circulation

Article

Today monetary policy is the principle way in which governments influence the macroeconomy. To implement monetary policy the monetary authority uses its policy instruments (short-term interest rates or the monetary base) to achieve its desired goals of low inflation and real output close to potential. Monetary policy has evolved over the centuries, along with the development of the money economy.

The origins

Debate swirls between historians, economists, anthropologists and numismatists over the origins of money. In the West it is commonly believed that coins first appeared in ancient Lydia in the eighth century bc. Some date the origins to ancient China.

Money evolved as a medium of exchange, a store of value and unit of account. According to one authority – Hicks (1969), following Menger (1892) – its rise was associated with the growth of commerce. Traders would hold stocks of another good, in addition to the goods they traded in, which was easily stored, widely recognized, and divisible, with precious metals evolving as the best example. This good would serve as the unit of account and then as a medium of exchange. According to this story money first emerged from market activity.

Governments became involved when the monarch realized that it was easier to pay his soldiers in terms of generalized purchasing power than with particular goods. This led to the origin of seigniorage or the government's prerogative in the coining of money. Seigniorage originally represented the fee that the royal mint collected from the public to convert their holdings of bullion into coin. Governments generally since ancient times had a monopoly over the issue of coins (either licensing their production or producing them themselves).

The earliest predecessors to monetary policy seem to be those of debasement, where the government would call in the coins, melt them down and mix them with cheaper metals. They would alter either the weight or the quality of the coins (fineness). An alternative method used was to alter the unit of account (see Redish, 2000; Sussman, 1993; Sargent and Velde, 2002). The practice of debasement was widespread in the later years of the Roman Empire (Schwartz, 1973), but reached its perfection in western Europe in the late Middle Ages. Sussman (1993) describes how the French monarchs of the 15th century, unable to collect more normal forms of taxes, used debasement as a form of inflation tax to finance the ongoing Hundred Years War with the English. Debasement was really a form of fiscal rather than monetary policy, but it set the stage for the later development of monetary policy using fiduciary money.

Fiduciary or paper money evolved from the operations of early commercial banks in Italy (Cipolla, 1967) to economize on the precious metals used in coins (although there is evidence that paper money was issued by imperial decree in China centuries earlier: see Chown, 1994). This development has its origins in the practice of goldsmiths who would issue warehouse receipts as evidence of their storing gold coins and bullion for their clients. Eventually these certificates circulated as media of exchange. Once the goldsmiths learned that not all the claims were redeemed at the same time, they were able to circulate claims of value greater than their specie reserves. Thus was borne fiduciary money (money not fully backed by specie) and fractional reserve banking. The goldsmiths and early commercial bankers learned by experience to hold a precautionary reserve sufficient to meet the demands for redemption in the normal course of business.

Governments began issuing paper money in Europe only in the 18th century. An early example was Sweden's note issue, initiated to finance its participation in the Seven Years War (Eagly, 1969). Fiat money reached its maturity during the American Revolutionary Wars when the Congress issued continentals to finance military expenditures. These were promissory notes to be convertible into specie; but the promise was not kept. They were issued in massive quantities. However, the rate of issue and the

average inflation rate of 65 per cent per annum (Rockoff, 1984) was not far removed from the revenue-maximizing rate of issue by a monopoly fiat money issuing central bank of the 20th century (Bailey, 1956). During the French Revolution the overissue of paper money, the *assignats*, which were based initially on the value of seized Church lands, led to hyperinflation (White, 1995).

An early predecessor of monetary policy was John Law's system. In 1719 Law persuaded the Regent of France to convert the French national debt into stock in his Compagnie des Indes. He then used the stock as backing for the issue of notes in his Banque Royale. Note issue could then support and finance the issue of further shares. Law then conducted a proto typical form of monetary policy in 1720 to save his system when he attempted both to peg the exchange rate of notes in terms of specie and provide a support price to stem the collapse in the price of shares (Bordo, 1987; Velde, 2007).

Central banks

Monetary policy is conducted by the monetary authority. It is the issuer of national currency and the source of the monetary base. Usually we think of central banks as fulfilling these functions, but in many countries, until well into the 20th century, in the absence of a central bank, these were performed by the Treasury or in some cases (Australia, Canada, New Zealand) by a large commercial bank entrusted with the government's tax revenues (Goodhart, 1989). The earliest central banks were established in the 17th century (the Swedish Riksbank founded in 1664, the Bank of England founded in 1694, the Banque de France, founded in 1800, and the Netherlands Bank in 1814) to aid the fisc of the newly emerging nation states.

In the case of the Bank of England a group of private investors was granted a royal charter to set up a bank to purchase and help market government debt. The establishment of the bank helped ensure the creation of a deep and liquid government debt market which served as the base of growing financial system (Dickson, 1969; Rousseau and Sylla, 2003. The bank eventually evolved into a bankers’ bank by taking deposits from other nascent commercial banks. Its large gold reserves and monopoly privilege eventually allowed it to become a lender of last resort, that is, to provide liquidity to its correspondents in the face of a banking panic – a scramble by the public for liquidity.

Monetary policy as we know it today began by the bank discounting the paper of other financial institutions, both government debt and commercial paper. The interest rate at which the bank would lend, based on this collateral became known as bank rate (in other countries as the discount rate). By altering this rate the bank could influence credit conditions in the British economy. It could also influence credit conditions in the rest of the world by attracting or repelling short-term funds (Sayers, 1957).

A second wave of central banks was initiated at the end of the 19th century. This was not based explicitly on the fiscal revenue motive as had been the case with the first wave, but on following the rules of the gold standard and ironing out swings in interest rates induced by seasonal forces and by the business cycle. Included in this group are the Swiss National Bank founded in 1907 (Bordo and James, 2007) and the Federal Reserve founded in 1913 (Meltzer, 2003). Subsequent waves of new central banks followed in the interwar period as countries in the British Empire, the new states of central Europe and Latin America attempted to emulate the experiences of the advanced countries (Capie et al., 1994).

Central bank independence

Although the early central banks had public charters, they were privately owned and they had policy independence. A problem that plagued the Bank of England in its early years was that it placed primary weight on its commercial activities and on several occasions of financial distress was criticized for neglecting the public good. Walter Bagehot formulated the responsibility doctrine in 1873 according to which the bank was to place primary importance on its public role as lender of last resort (Bagehot, 1873).

From the First World War onwards central banks focused entirely on public objectives, and many fell under public control. Their objectives also changed from emphasis on maintaining specie convertibility towards shielding the domestic economy from external shocks and stabilizing real output and prices. This trend continued in the 1930s and after the Second World War. Moreover, the Great Depression led to a major reaction against central banks, which were accused of creating and exacerbating the depression. In virtually every country monetary policy was placed under the control of the Treasury and fiscal policy became dominant. In every country central banks followed a low interest peg to both stimulate the economy and aid the Treasury in marketing its debt.

Monetary policy was restored to the central banks in the 1950s (for example, in the United States, after the Treasury–Federal Reserve Accord of 1951), and there followed a brief period of price stability until the mid-1960s. This was followed by a significant run up in inflation worldwide. The inflation was broken in the early 1980s by concerted tight monetary policies in the United States, the United Kingdom and other countries and a new emphasis placed on the importance of low inflation based on credible monetary policies. Central banks in many countries were granted goal independence and were given a mandate to keep inflation low.

Classical monetary policy

The true origin of modern monetary policy occurred under the classical gold standard, which prevailed from 1880 to 1914. The gold standard evolved from the earlier bimetallic regime. Under the gold standard all countries would define their currencies in terms of a fixed weight of gold and then all fiduciary money would be convertible into gold. The key role of central banks was to maintain gold convertibility. Central banks were also supposed to use their discount rates to speed up the adjustment to external shocks to the balance of payments, that is, they were supposed to follow the ‘rules of the game’ (Keynes, 1930). In the case of a balance of payments deficit, gold would tend to flow abroad and reduce a central bank's gold reserves. According to the rules, the central bank would raise its discount rate. This would serve to depress aggregate demand and offset the deficit. At the same time the rise in rates would stimulate a capital inflow. The opposite set of policies was to be followed in the case of a surplus.

There is considerable debate on whether the rules were actually followed (Bordo and MacDonald, 2005). There is evidence that central banks sterilized gold flows and prevented the adjustment mechanism from working (Bloomfield, 1959). Others paid attention to the domestic objectives of price stability or stable interest rates or stabilizing output (Goodfriend, 1988). There is also evidence that because the major central banks were credibly committed to maintaining gold convertibility they had

some policy independence to let their interest rates depart from interest rate parity and to pursue domestic objectives (Bordo and MacDonald, 2005).

After the First World War the gold standard was restored, but in the face of a changing political economy – the extension of suffrage and organized labour (Eichengreen, 1992) – greater emphasis was placed by central banks on the domestic objectives of price stability and stable output and employment than on external convertibility. Thus for example the newly created Federal Reserve sterilized gold flows and followed countercyclical policies to offset two recessions in the 1920s (Meltzer, 2003).

The depression beginning in 1929 was probably caused by inappropriate monetary policy. The Federal Reserve followed the flawed real bills doctrine, which exacerbated the downturn, and the gold sterilization policies followed by the Fed and the Banque de France greatly weakened the adjustment mechanism of the gold standard. As mentioned above, the central banks were blamed for the depression and monetary policy was downgraded until the mid-1950s.

The goals of monetary policy

The goals of monetary policy have changed across monetary regimes. Until 1914, the dominant monetary regime was the gold standard. Since then the world has gradually shifted to a fiat money regime. Under the classical gold standard the key goal was gold convertibility with limited focus on the domestic economy. By the interwar period gold convertibility was being overshadowed by emphasis on domestic price level and output stability, and the regime shifted towards fiat money. This continued after the Second World War. Under the 1944 Bretton Woods Articles of Agreement, member countries were to maintain pegged exchange rates and central banks were to intervene in the foreign exchange market to do this, but the goal of domestic full employment was also given predominance. The Bretton Woods system evolved into a dollar gold exchange standard in which member currencies were convertible on a current account basis into dollars and the dollar was convertible into gold (Bordo, 1993). A continued conflict between the dictates of internal and external balance was a dominant theme from 1959 to 1971 as was the concern over global imbalance because the United States, as centre country of the system, would provide through its balance of payments deficits and its role as a financial intermediary more dollars than could be safely backed by its gold reserves (Triffin, 1960).

The collapse of Bretton Woods between 1971 and 1973 was brought about largely because the United States followed an inflationary policy to finance both the Vietnam War and expanded social welfare programmes like Medicare under President Johnson's Great Society, thus ending any connection of the monetary regime to gold and propelling the world to a pure fiat regime. In this new environment the balance was largely tipped in favour of domestic stability and was coupled with the now dominant belief by central bankers in the Phillips curve trade-off between unemployment and inflation (Phillips, 1958): this led to a focus on maintaining full employment at the expense of inflation.

The resulting ‘great inflation’ of the 1970s finally came to an end in the early 1980s by central banks following tight monetary policies. Since then the pendulum has again swung towards the goal of low inflation and the belief that central banks should eschew control of real variables (Friedman, 1968; Phelps, 1968).

The instruments of monetary policy

The original policy instrument was the use of the discount rate and rediscounting. Open market operations (the buying and selling of government securities) was first developed in the 1870s and 1880s by the Bank of England in order to make bank rate effective, that is to force financial institutions to borrow (Sayers, 1957). Other countries with less developed money markets than those of Britain used credit rationing (France) and gold policy operations to alter the gold points and impede the normal flow of gold (Sayers, 1936) .

In the interwar period the newly established Federal Reserve initially used the discount rate as its principal tool, but after heavy criticism for its use in rolling back the post-First World War inflation and thereby creating one of the worst recessions of the 20th century in 1920–1 (Meltzer, 2003), the Fed shifted to open market policy, its principal tool ever since. In the 1930s it also began changing reserve requirements. Its policy of doubling reserve requirements in 1936 was later blamed as the cause for the recession of 1937–8 (Friedman and Schwartz, 1963). In the 1930s and 1940s, along with the downgrading of monetary policy, came an increased use of various types of controls and regulations such as margin requirements on stock purchases, selective credit controls on consumer durables and interest rate ceilings. Similar policies were adopted elsewhere. The return to traditional monetary policy in the 1950s restored open market operations to the position of predominance.

Intermediate targets

Traditionally, central banks altered interest rates as the mechanism to influence aggregate spending, prices and output. In the 1950s, the monetarists revived the quantity theory of money and posited the case for using money supply as the intermediate target (Friedman, 1956; Brunner and Meltzer, 1993). The case for money was based on evidence of a stable relationship between the growth of money supply, on the one hand, and nominal income and the price level, on the other hand, and the evidence that, by focusing on interest rates, the Fed and other central banks aggravated the business cycle, and then – in part because of their inability to distinguish between real and nominal rates – generated the great inflation of the 1970s (Brunner and Meltzer, 1993).

By the 1970s most central banks had monetary aggregate targets. However, the rise in inflation in the 1970s (which was followed by disinflation) as well as continuous financial innovation (which was in turn exacerbated by inflation uncertainty) made the demand for money function less predictable (Laidler, 1980; Judd and Scadding, 1982). This meant that central banks had difficulty in meeting their money growth targets. In addition, the issue was raised as to which monetary aggregate to target (Goodhart, 1984). By the late 1980s most countries had abandoned monetary aggregates and returned to interest rates. But since the early 1990s monetary policy in many countries has been based on pursuing an inflation target (implicit or explicit) with the policy rate set to allow inflation to hit the target, a policy which seems to be successful.

Theories of monetary policy

The development of the practice of monetary policy described above was embedded in major advances

in monetary theory that began in the first quarter of the 19th century. A major controversy in England, the Currency Banking School debate, has shaped subsequent thinking on monetary policy ever since. That debate evolved out of the Bullionist debate during the Napoleonic wars over whether inflation in Britain was caused by monetary or real forces (Viner, 1937). In a later debate, Currency School advocates emphasized the importance for the Bank of England to change its monetary liabilities in accordance with changes in its gold reserves – that is, according to the currency principle, which advocated a rule tying money supply to the balance of payments. The opposing Banking School emphasized the importance of disturbances in the domestic economy and the domestic financial system as the key variables the Bank of England should react to. They advocated that the bank directors should use their discretion rather than being constrained by a rigid rule. The controversy still rages.

Later in the 19th century, the two principles became embedded in central banking lore (Meltzer, 2003, ch. 2). The Federal Reserve and other central banks (including the Swiss National Bank) were founded on two pillars that evolved from this debate – the gold standard and the real bills doctrine.

The latter evolved from 19th-century practice and the Banking School theory. The basic premise of real bills is that as long as commercial banks lend on the basis of self-liquidating short-term real bills they will be sound. Moreover, as long as central banks discount only eligible real bills the economy will always have the correct amount of money and credit. Adherence to real bills sometimes clashed with the first pillar, gold adherence, for example when the economy was expanding and real bills dictated ease while the balance of payments was deteriorating, which dictated tightening. This conflict erupted in the United States on a number of occasions in the 1920s (Friedman and Schwartz, 1963).

Adherence to the two pillars led to disaster in the 1930s. The Fed made a serious policy error by following real bills. A corollary of that theory urged the Fed to defuse the stock market boom because it was believed that speculation would lead to inflation, which would ultimately lead to deflation (Meltzer, 2003). According to Friedman and Schwartz, Meltzer and others, the Fed's tight policy triggered a recession in 1929 and its inability to stem the banking panics that followed in the early 1930s led to the Great Depression. The depression was spread globally by the fixed exchange rate gold standard. In addition, the gold standard served as ‘golden fetters’ for most countries because, lacking the credibility they had before 1914, they could not use monetary policy to allay banking panics or stimulate the economy lest it trigger a speculative attack (Eichengreen, 1992).

The Great Depression gave rise to the Keynesian view that monetary policy was impotent. This led to the dominance of fiscal policy over monetary policy for the next two decades. The return to traditional monetary policy in the 1950s was influenced by Keynesian monetary theory. According to this approach monetary policy should influence short-term rates and then by a substitution process across the financial portfolio would affect the real rate of return on capital. This money market approach dominated policy until the 1960s.

The monetarists criticized the Fed for failing to stabilize the business cycle, for still adhering to vestiges of real bills (for example, free reserves: Calomiris and Wheelock, 1998), and for its belief in a stable Phillips curve – that unemployment could be permanently reduced at the expense of inflation. This, they argued, led to an acceleration of inflation as market agents’ expectations adjusted to the higher inflation rate, which produced the great inflation of the 1970s. As mentioned above, the subsequent adoption of monetary aggregate targeting was short lived because of unpredictable shifts in velocity.

The approach to monetary policy followed since the early 1990s has learned the basic lesson from the

monetarists of the primacy of price stability. It also learned about the distinction between nominal and real interest rates (Fisher, 1922). Moreover, it has adopted a principle from the earlier gold standard literature, Wicksell's (1898) distinction between the natural rate of interest and the bank rate (Woodford, 2003). In Wicksell's theory, central banks should gear their lending rate to the natural rate (the real rate of return on capital). If it keeps bank rate too low, inflation will ensue, which under the gold standard will lead to gold outflows and upward market pressure on the bank rate. Today's central banks, dedicated to low inflation, can be viewed as following the Taylor rule, according to which they set the nominal policy interest rate relative to the natural interest rate as a function of the deviation of inflation forecasts from their targets and real output from its potential (Taylor, 1999).

Rules versus discretion

A key theme in the monetary policy debate is the issue of rules versus discretion. The question that followed the Currency Banking School debate was whether monetary policy should be entrusted to well meaning authorities with limited knowledge or to a rule that cannot be designed to deal with unknown shocks (Simons, 1936; Friedman, 1960).

A more recent approach focuses on the role of time inconsistency. According to this approach a rule is a credible commitment mechanism that ties the hands of policymakers and prevents them from following time-inconsistent policies – policies that take past policy commitments as given and react to the present circumstances by changing policy (Kydland and Prescott, 1977; Barro and Gordon, 1983). In this vein, today's central bankers place great emphasis on accountability and transparency to support the credibility of their commitments to maintain interest rates geared towards low inflation (Svensson, 1999).

Conclusion

Monetary policy has evolved since the early 19th century. It played a relatively minor role before 1914, although it was then that many of its tools and principles were developed. The role of monetary policy in stabilizing prices and output came to fruition in the 1920s, but for the Federal Reserve, which used a flawed model – the real bills doctrine – and adhered to a less than credible gold standard, the policy was a recipe for disaster and led to the great contraction of 1929–33. When monetary policy was restored in the 1950s in the United States, it still was influenced by real bills (Calomiris and Wheelock, 1998), which may have led to the policy mistakes that created the great inflation. The rest of the world was tied to the United States by the pegged exchange rates of Bretton Woods. Since the early 1990s monetary policy in many countries has returned back to a key principle of the gold standard era – price stability based on a credible nominal anchor (Bordo and Schwartz, 1999) and to Wicksell's distinction between real and nominal interest rates. Yet it is based on a fiat regime and the commitment of central banks to follow credible and predictable policies.

See Also

- Bank of England

- central bank independence
- fiat money
- gold standard
- inflation targeting
- monetary and fiscal policy overview

Bibliography

Barro, R.I. and Gordon, D.B. 1983. Rules, discretion and reputation in a model of monetary policy. *Journal of Monetary Economics* 12, 101–21.

Bloomfield, A.I. 1959. *Monetary Policy under the International Gold Standard*. New York: Federal Reserve Bank of New York.

Bagehot, W. 1873. *Lombard Street: A Description of the Money Market*. Reprint edn. London: John Murray, 1917.

Bailey, M.J. 1956. The welfare costs of inflationary finance. *Journal of Political Economy* 64, 93–110.

Bordo, M.D. 1987. John Law. In *The New Palgrave: A Dictionary of Economic Theory and Doctrine*, ed. J. Eatwell and M. Milgate. London: Macmillan.

Bordo, M.D. 1993. The Bretton Woods international monetary system: a historical overview. In *A Retrospective on the Bretton Woods System: Lessons for International Monetary Reform*, ed. M.D. Bordo and B. Eichengreen. Chicago: University of Chicago Press.

Bordo, M.D. and James, H. 2007. The SNB 1907–1946: a happy childhood or a troubled adolescence? In Swiss National Bank. Centenary Conference volume, Zurich.

Bordo, M.D. and MacDonald, R. 2005. Interest rate interactions in the classical gold standard: 1880–1914: was there monetary independence? *Journal of Monetary Economics* 52, 307–27.

Bordo, M.D. and Schwartz, A.J. 1999. Monetary policy regimes and economic performance: the historical record. In *Handbook of Macroeconomics*, ed. J.B. Taylor and M. Woolford. New York: North-Holland.

Brunner, K. and Meltzer, A.H. 1993. *Money and the Economy: Issues in Monetary Analysis*. Cambridge: Cambridge University Press.

Calomiris, C.W. and Wheelock, D.C. 1998. Was the great depression a watershed for American monetary policy? In *The Defining Moment: The Great Depression and the American Economy in the*

Twentieth Century, ed. M.D. Bordo, C. Goldin and E.N. White. Chicago: University of Chicago Press.

Capie, F., Goodhart, C., Fischer, S. and Schnadt, N. 1994. *The Future of Central Banking*. Cambridge: Cambridge University Press.

Chown, J.F. 1994. *The History of Money from AD 800*. London: Routledge.

Cipolla, C.M. 1967. *Money, Prices, and Civilization in the Mediterranean World, Fifth to Seventeenth Century*. New York: Gordian Press.

Dickson, P.M. 1969. *The Financial Revolution in England: A Study in the Development of Public Credit, 1688–1756*. London: Macmillan.

Eagly, R.U. 1969. Monetary policy and politics in mid-eighteenth century Sweden. *Journal of Economic History* 29, 739–57.

Eichengreen, B. 1992. *Golden Fetters*. New York: Oxford University Press.

Fisher, I. 1922. *The Purchasing Power of Money*. New York: Augustus M. Kelley, 1965.

Friedman, M. 1956. Quantity theory of money: a restatement. In *Studies in the Quality Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press.

Goodfriend, M. 1988. Central banking under the gold standard. *Carnegie Rochester Conference Series on Public Policy* 19, 85–124.

Goodhart, C.A.E. 1984. Chapter 3 problems of monetary management. In *Monetary Theory and Practice. The UK Experience*. London: Macmillan.

Goodhart, C. 1989. *The Evolution of Central Banks*. Cambridge, MA: MIT Press.

Hicks, J.R. 1969. *A Theory of Economic History*. Oxford: Clarendon Press.

Judd, J.P. and Scadding, J.L. 1982. The search for a stable money demand function: a survey of the post-

1973 literature. *Journal of Economic Literature* 20, 993–1023.

Keynes, J.M. 1930. *A Treatise on Money*, vol. 2: *The Applied Theory of Money*. Repr. in *The Collected Writings of John Maynard Keynes*. 30 vols, ed. A. Robinson and D. Moggridge, vol. 6. London: Macmillan for the Royal Economic Society, 1971.

Kydland, F.E. and Prescott, E.C. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–92.

Laidler, D. 1980. The demand for money in the United States – yet again. In *The State of Macro-Economics, Carnegie-Rochester Conference Series on Public Policy*, vol. 12, ed. K. Brunner and A.H. Meltzer. New York: North-Holland.

Meltzer, A.H. 2003. *A History of the Federal Reserve*, vol. 1. Chicago: University of Chicago Press.

Menger, K. 1892. On the origins of money. *Economic Journal* 2, 238–58.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom 1861–1957. *Economica* 25, 283–99.

Phelps, E.S. 1968. Money-wage dynamics and labor market equilibrium. *Journal of Political Economy* 76, 678–711.

Redish, A. 2000. *Bimetallism: An Economic and Historical Analysis*. Cambridge: Cambridge University Press.

Ricardo, D. 1811. High price of bullion: a proof of the depreciation of bank notes. In *The Works and Correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.

Rockoff, H. 1984. *Drastic Measures: A History of Wage and Price Controls in the United States*. New York: Cambridge University Press.

Rousseau, P. and Sylla, R. 2003. Financial systems ,economic growth and globalization. In *Globalization in Historical Perspective*, ed. M.D. Bordo, A. Taylor and J. Williamson. Chicago: University of Chicago Press.

Sargent, T. and Velde, F. 2002. *The Big Problem of Small Change*. Princeton: Princeton University Press.

Sayers, R.S. 1936. *Bank of England Operations, 1890–1914*. London: P.S. King & Son.

Sayers, R.S. 1957. *Central Banking after Bagehot*. Oxford: Oxford University Press.

Schwartz, A.J. 1973. Secular price change in historical perspective. *Journal of Money, Credit and Banking* 5, 243–69.

Simons, H.C. 1936. Rule versus authorities in monetary policy. *Journal of Political Economy* 44, 1–30.

Sussman, N. 1993. Debasement, royal reviews and inflation in France during the second stage of the Hundred Years War. *Journal of Economic History* 56, 789–808.

Svensson, L.E.O. 1999. Inflation targeting as a monetary policy rule. *Journal of Monetary Economics* 43, 607–54.

Taylor, J.B. 1999. A historical analysis of monetary policy rules. In *Monetary Policy Rule*, ed. J.B. Taylor. Chicago: University of Chicago Press.

Thornton, H. 1802. *An Inquiry into the National Effects of the Paper Credit of Great Britain*. Fairfield, NJ: Augustus M. Kelley, 1978.

Triffin, R. 1960. *Gold and the Dollar Crisis*. New Haven: Yale University Press.

Velde, F. 2007. *Government Equity and Money: John Laws System in 1720 France*. Princeton: Princeton University Press.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Augustus M. Kelley, 1975.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

Wicksell, K. 1898. *Interest and Prices*. New York: Augustus M. Kelley, 1965.

White, E.N. 1995. The French Revolution and the politics of government finance, 1770–1815. *Journal of Economic History* 55, 227–55.

How to cite this article

Bordo, Michael D. "monetary policy, history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_H000180> doi:10.1057/9780230226203.1124

The New Palgrave Dictionary of Economics Online

monetary transmission mechanism

Peter N. Ireland

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The monetary transmission mechanism describes how policy-induced changes in the nominal money stock or the short-term nominal interest rate impact on real variables such as aggregate output and employment. Specific channels of monetary transmission operate through the effects that monetary policy has on interest rates, exchange rates, equity and real estate prices, bank lending, and firm balance sheets. Recent research on the transmission mechanism seeks to understand how these channels work in the context of dynamic, stochastic, general equilibrium models.

Keywords

asset price channels of monetary transmission; balance sheet channel of monetary transmission; balance sheet credit channel; bank lending credit channel; bank reserves; bonds; central banks; currency; exchange rate channel of monetary transmission; inflation targeting; interest rate channel of monetary transmission; interest rates; IS–LM model; Keynesianism; life-cycle theory of consumption; liquidity trap; loanable funds theory; monetarism; monetary base; monetary policy; monetary transmission mechanism; New Keynesian economics; New Keynesian Phillips curve; open market operations; Phillips curve; rational expectations models; real business cycles; Taylor rule

Article

The monetary transmission mechanism describes how policy-induced changes in the nominal money stock or the short-term nominal interest rate impact on real variables such as aggregate output and employment.

Key assumptions

Central bank liabilities include both components of the monetary base: currency and bank reserves.

Hence, the central bank controls the monetary base. Indeed, monetary policy actions typically begin when the central bank changes the monetary base through an open market operation, purchasing other securities – most frequently, government bonds – to increase the monetary base or selling securities to decrease the monetary base.

If these policy-induced movements in the monetary base are to have any impact beyond their immediate effects on the central bank's balance sheet, other agents must lack the ability to offset them exactly by changing the quantity or composition of their own liabilities. Thus, any theory or model of the monetary transmission mechanism must assume that there exist no privately issued securities that substitute perfectly for the components of the monetary base. This assumption holds if, for instance, legal restrictions prevent private agents from issuing liabilities having one or more characteristics of currency and bank reserves.

Both currency and bank reserves are nominally denominated, their quantities measured in terms of the economy's unit of account. Hence, if policy-induced movements in the nominal monetary base are to have real effects, nominal prices must not be able to respond immediately to those movements in a way that leaves the real value of the monetary base unchanged. Thus, any theory or model of the monetary transmission mechanism must also assume that some friction in the economy works to prevent nominal prices from adjusting immediately and proportionally to at least some changes in the monetary base.

The monetary base and the short-term nominal interest rate

If, as in the US economy today, neither component of the monetary base pays interest or if, more generally, the components of the monetary base pay interest at a rate that is below the market rate on other highly liquid assets such as short-term government bonds, then private agents' demand for real base money M/P can be described as a decreasing function of the short-term nominal interest rate i: $M/P = L(i)$. This function L summarizes how, as the nominal interest rate rises, other highly liquid assets become more attractive as short-term stores of value, providing stronger incentives for households and firms to economize on their holdings of currency and banks to economize on their holdings of reserves. Thus, when the price level P cannot adjust fully in the short run, the central bank's monopolistic control over the nominal quantity of base money M also allows it to influence the short-term nominal interest rate i, with a policy-induced increase in M leading to whatever decline in i is necessary to make private agents willing to hold the additional volume of real base money and, conversely, a policy-induced decrease in M leading to a rise in i. In the simplest model where changes in M represent the only source of uncertainty, the deterministic relationship that links M and i implies that monetary policy actions can be described equivalently in terms of their effects on either the monetary base or the short-term nominal interest rate.

Poole's (1970) analysis shows, however, that the economy's response to random shocks of other kinds can depend importantly on whether the central bank operates by setting the nominal quantity of base money and then allowing the market to determine the short-term nominal interest rate or by setting the short-term nominal interest rate and then supplying whatever quantity of nominal base money is demanded at that interest rate. More specifically, Poole's analysis reveals that central bank policy insulates output and prices from the effects of large and unpredictable disturbances to the money demand relationship by setting a target for i rather than M. Perhaps reflecting the widespread belief that money demand shocks are large and unpredictable, most central banks around the world today –

including the Federal Reserve in the United States – choose to conduct monetary policy with reference to a target for the short-term nominal interest rate as opposed to any measure of the money supply. Hence, in practice, monetary policy actions are almost always described in terms of their impact on a short-term nominal interest rate – such as the federal funds rate in the United States – even though, strictly speaking, those actions still begin with open market operations that change the monetary base.

The channels of monetary transmission

Mishkin (1995) usefully describes the various channels through which monetary policy actions, as summarized by changes in either the nominal money stock or the short-term nominal interest rate, impact on real variables such as aggregate output and employment.

According to the traditional Keynesian *interest rate channel*, a policy-induced increase in the short-term nominal interest rate leads first to an increase in longer-term nominal interest rates, as investors act to arbitrage away differences in risk-adjusted expected returns on debt instruments of various maturities as described by the expectations hypothesis of the term structure. When nominal prices are slow to adjust, these movements in nominal interest rates translate into movements in real interest rates as well. Firms, finding that their real cost of borrowing over all horizons has increased, cut back on their investment expenditures. Likewise, households facing higher real borrowing costs scale back on their purchases of homes, automobiles and other durable goods. Aggregate output and employment fall. This interest rate channel lies at the heart of the traditional Keynesian textbook IS–LM model, due originally to Hicks (1937), and also appears in the more recent New Keynesian models described below.

In open economies, additional real effects of a policy-induced increase in the short-term interest rate come about through the *exchange rate channel*. When the domestic nominal interest rate rises above its foreign counterpart, equilibrium in the foreign exchange market requires that the domestic currency gradually depreciate at a rate that, again, serves to equate the risk-adjusted returns on various debt instruments, in this case debt instruments denominated in each of the two currencies – this is the condition of uncovered interest parity. Both in traditional Keynesian models that build on Fleming (1962), Mundell (1963), and Dornbusch (1976) and in the New Keynesian models described below, this expected future depreciation requires an initial appreciation of the domestic currency that, when prices are slow to adjust, makes domestically produced goods more expensive than foreign-produced goods. Net exports fall; domestic output and employment fall as well.

Additional *asset price channels* are highlighted by Tobin's (1969) *q*-theory of investment and Ando and Modigliani's (1963) life-cycle theory of consumption. Tobin's *q* measures the ratio of the stock market value of a firm to the replacement cost of the physical capital that is owned by that firm. All else equal, a policy-induced increase in the short-term nominal interest rate makes debt instruments more attractive than equities in the eyes of investors; hence, following a monetary tightening, equilibrium across securities markets must be re-established in part through a fall in equity prices. Facing a lower value of *q*, each firm must issue more new shares of stock in order to finance any new investment project; in this sense, investment becomes more costly for the firm. In the aggregate across all firms, therefore, investment projects that were only marginally profitable before the monetary tightening go unfunded after the fall in *q*, leading output and employment to decline as well. Meanwhile, Ando and Modigliani's life-cycle theory of consumption assigns a role to wealth as well as income as key determinants of

consumer spending. Hence, this theory also identifies a channel of monetary transmission: if stock prices fall after a monetary tightening, household financial wealth declines, leading to a fall in consumption, output and employment.

According to Meltzer (1995), asset price movements beyond those reflected in interest rates alone also play a central role in *monetarist* descriptions of the transmission mechanism. Indeed, monetarist critiques of the traditional Keynesian model often start by questioning the view that the full thrust of monetary policy actions is completely summarized by movements in the short-term nominal interest rate. Monetarists argue instead that monetary policy actions impact on prices simultaneously across a wide variety of markets for financial assets and durable goods, but especially in the markets for equities and real estate, and that those asset price movements are all capable of generating important wealth effects that impact, through spending, on output and employment.

Two distinct *credit channels*, the *bank lending channel* and the *balance sheet channel*, also allow the effects of monetary policy actions to propagate through the real economy. Kashyap and Stein (1994) trace the origins of thought on the bank lending channel back to Roosa (1951) and also highlight Blinder and Stiglitz's (1983) resurrection of the loanable funds theory and Bernanke and Blinder's (1988) extension of the IS–LM model as two approaches that account for this additional source of monetary non-neutrality. According to this lending view, banks play a special role in the economy not just by issuing liabilities – bank deposits – that contribute to the broad monetary aggregates but also by holding assets – bank loans – for which few close substitutes exist. More specifically, theories and models of the bank lending channel emphasize that for many banks, particularly small banks, deposits represent the principal source of funds for lending and that for many firms, particularly small firms, bank loans represent the principal source of funds for investment. Hence, an open market operation that leads first to a contraction in the supply of bank reserves and then to a contraction in bank deposits requires banks that are especially dependent on deposits to cut back on their lending, and firms that are especially dependent on bank loans to cut back on their investment spending. Financial market imperfections confronting individual banks and firms thereby contribute, in the aggregate, to the decline in output and employment that follows a monetary tightening.

Bernanke and Gertler (1995) describe a broader credit channel, the balance sheet channel, where financial market imperfections also play a key role. Bernanke and Gertler emphasize that, in the presence of financial market imperfections, a firm's cost of credit, whether from banks or any other external source, rises when the strength of its balance sheet deteriorates. A direct effect of monetary policy on the firm's balance sheet comes about when an increase in interest rates works to increase the payments that the firm must make to service its floating rate debt. An indirect effect arises, too, when the same increase in interest rates works to reduce the capitalized value of the firm's long-lived assets. Hence, a policy-induced increase in the short-term interest rate not only acts immediately to depress spending through the traditional interest rate channel, it also acts, possibly with a lag, to raise each firm's cost of capital through the balance sheet channel, deepening and extending the initial decline in output and employment.

Recent developments

Recent theoretical work on the monetary transmission mechanism seeks to understand how the

traditional Keynesian interest rate channel operates within the context of dynamic, stochastic, general equilibrium models. This recent work builds on early attempts by Fischer (1977) and Phelps and Taylor (1977) to combine the key assumption of nominal price or wage rigidity with the assumption that all agents have rational expectations so as to overturn the policy ineffectiveness result that McCallum (1979) associates with Lucas (1972) and Sargent and Wallace (1975). This recent work builds on those earlier studies by deriving the key behavioural equations of the New Keynesian model from more detailed descriptions of the objectives and constraints faced by optimizing households and firms. More specifically, the basic New Keynesian model consists of three equations involving three variables: output y_t, inflation π_t, and the short-term nominal interest rate i_t. The first equation, which Kerr and King (1996) and McCallum and Nelson (1999) dub the expectational IS curve, links output today to its expected future value and to the *ex ante* real interest rate, computed in the usual way by subtracting the expected rate of inflation from the nominal interest rate:

$$y_t = E_t y_{t+1} - \sigma(i_t - E_t \pi_{t+1}),$$

where σ, like all of the other parameters to be introduced below, is strictly positive. This equation corresponds to a log-linearized version of the Euler equation linking an optimizing household's intertemporal marginal rate of substitution to the inflation-adjusted return on bonds, that is, to the real interest rate. The second equation, the New Keynesian Phillips curve, takes the form

$$\pi_t = \beta E_t \pi_{t+1} + \gamma y_t$$

and corresponds to a log-linearized version of the first-order condition describing the optimal behavior of monopolistically competitive firms that either face explicit costs of nominal price adjustment, as suggested by Rotemberg (1982), or set their nominal prices in randomly staggered fashion, as suggested by Calvo (1983). The third and final equation is an interest rate rule for monetary policy of the type proposed by Taylor (1993),

$$i_t = \alpha \pi_t + \psi y_t,$$

according to which the central bank systematically adjusts the short-term nominal interest in response to movements in inflation and output. This description of monetary policy in terms of interest rates reflects the observation, noted above, that most central banks today conduct monetary policy using targets for

the interest rate as opposed to any of the monetary aggregates. A money demand equation could be appended to this three-equation model, but that additional equation would serve only to determine the amount of money that the central bank and the banking system would need to supply to clear markets, given the setting for the central bank's interest rate target (see Ireland, 2004, for a detailed discussion of this last point).

In this benchmark New Keynesian model, monetary policy operates through the traditional Keynesian interest rate channel. A monetary tightening in the form of a shock to the Taylor rule that increases the short-term nominal interest rate translates into an increase in the real interest rate as well when nominal prices move sluggishly due to costly or staggered price setting. This rise in the real interest rate then causes households to cut back on their spending, as summarized by the IS curve. Finally, through the Phillips curve, the decline in output puts downward pressure on inflation, which adjusts only gradually after the shock.

Importantly, however, the expectational terms that enter into the IS and Phillips curves displayed above imply that policy actions will differ in their quantitative effects depending on whether these actions are anticipated or unanticipated; hence, this New Keynesian model follows the earlier rational expectations models of Lucas and Sargent and Wallace by stressing the role of expectations in the monetary transmission mechanism. And, as emphasized by Kimball (1995), by deriving these expectational forms for the IS and Phillips curves from completely spelled-out descriptions of the optimizing behaviour of households and firms, the New Keynesian model takes advantage of the powerful microeconomic foundations introduced into macroeconomics through Kydland and Prescott's (1982) real business cycle model while also drawing on insights from earlier work in New Keynesian economics as exemplified, for instance, by the articles collected in Mankiw and Romer's (1991) two-volume set.

Clarida, Gali and Gertler (1999) and Woodford (2003) trace out the New Keynesian model's policy implications in much greater detail. Obstfeld and Rogoff (1995) develop an open-economy extension in which the exchange rate channel operates together with the interest rate channel of monetary transmission. Andres, Lopez-Salido and Nelson (2004) enrich the New Keynesian specification to open up a broader range of asset price channels and, similarly, Bernanke, Gertler and Gilchrist (1999) extend the basic model to account for the balance sheet channel of monetary transmission. Hence, all of these papers contribute to a large and still growing body of literature that examines the workings of various channels of monetary transmission within dynamic, stochastic, general equilibrium models.

Other recent research on the monetary transmission mechanism focuses on the problem of the zero lower bound on nominal interest rates – a problem that appears most starkly in the basic New Keynesian model sketched out above, in which monetary policy affects the economy exclusively through the Keynesian interest rate channel. Private agents always have the option of using currency as a store of value; hence, equilibrium in the bond market requires a non-negative nominal interest rate. In a low-inflation environment where nominal interest rates are also low on average, the central bank may bump up against this zero lower bound and find itself unable to provide further monetary stimulus after the economy is hit by a series of adverse shocks. Interest in the zero lower bound grew during the late 1990s and early 2000s when, in fact, nominal interest rates approached zero in Japan, the United States and a number of other countries. Among recent studies, Summers (1991) and Fuhrer and Madigan (1997) rank among the first to call for renewed attention to the problem of the zero lower bound; Krugman (1998) draws parallels between the zero lower bound and the traditional Keynesian liquidity trap; and Eggertsson and

Woodford (2003), Svensson (2003), and Bernanke, Reinhart and Sack (2004) propose and evaluate alternative monetary policy strategies for coping with the zero lower bound.
Finally, on the empirical front, quite a bit of recent work looks for evidence of quantitatively important credit channels of monetary transmission. Kashyap and Stein (1994) and Bernanke, Gertler and Gilchrist (1996) survey this branch of the literature. Also, the striking rise in equity and real estate prices that began in the mid-1990s in the United States, the United Kingdom, and elsewhere has sparked renewed interest in quantifying the importance of the asset price channels described above. Noteworthy contributions along these lines include Lettau and Ludvigson (2004) and Case, Quigley and Shiller (2005).

See Also

- inflation dynamics
- liquidity trap
- monetary business cycles (imperfect information)
- monetary business cycle models (sticky prices and wages)
- money supply
- Phillips curve (new views)
- Taylor rules

I would like to thank Steven Durlauf and Jeffrey Fuhrer for extremely helpful comments and suggestions.

Bibliography

Ando, A. and Modigliani, F. 1963. The ‘life cycle’ hypothesis of saving: aggregate implications and tests. *American Economic Review* 53, 55–84.

Andres, J., Lopez-Salido, J. and Nelson, E. 2004. Tobin's imperfect asset substitution in optimizing general equilibrium. *Journal of Money, Credit, and Banking* 36, 665–90.

Bernanke, B. and Blinder, A. 1988. Credit, money, and aggregate demand. *American Economic Review* 78, 435–39.

Bernanke, B. and Gertler, M. 1995. Inside the black box: the credit channel of monetary policy transmission. *Journal of Economic Perspectives* 9(4), 27–48.

Bernanke, B., Gertler, M. and Gilchrist, S. 1996. The financial accelerator and the flight to quality. *Review of Economics and Statistics* 78, 1–15.

Bernanke, B., Gertler, M. and Gilchrist, S. 1999. The financial accelerator in a quantitative business

cycle framework. In *Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Bernanke, B., Reinhart, V. and Sack, B. 2004. Monetary policy alternatives at the zero bound: an empirical assessment. *Brookings Papers on Economic Activity* 2004(2), 1–78.

Blinder, A. and Stiglitz, J. 1983. Money, credit constraints, and economic activity. *American Economic Review* 73, 297–302.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Case, K., Quigley, J. and Shiller, R. 2005. Comparing wealth effects: the stock market versus the housing market. *Advances in Macroeconomics* 5 . http://www.bepress.com/bejm/advances/vol5/iss1/art1/.

Clarida, R., Gali, J. and Gertler, M. 1999. The science of monetary policy: a New Keynesian perspective. *Journal of Economic Literature* 37, 1661–707.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84, 1161–76.

Eggertsson, G. and Woodford, M. 2003. The zero bound on interest rates and optimal monetary policy. *Brookings Papers on Economic Activity* 2003(1), 139–211.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85, 191–205.

Fleming, J. 1962. Domestic financial polices under fixed and under floating exchange rates. *International Monetary Fund Staff Papers* 9, 369–79.

Fuhrer, J. and Madigan, B. 1997. Monetary policy when interest rates are bounded at zero. *Review of Economics and Statistics* 79, 573–85.

Hicks, J. 1937. Mr. Keynes and the ‘Classics’: a suggested interpretation. *Econometrica* 5, 147–59.

Ireland, P. 2004. Money's role in the monetary business cycle. *Journal of Money, Credit, and Banking* 36, 969–83.

Kashyap, A. and Stein, J. 1994. Monetary policy and bank lending. In *Monetary Policy*, ed. N. Mankiw. Chicago: University of Chicago Press.

Kerr, W. and King, R. 1996. Limits on interest rate rules in the IS model. *Federal Reserve Bank of Richmond Economic Quarterly* 82, 47–75.

Kimball, M. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27, 1241–77.

Krugman, P. 1998. It's baaack: Japan's slump and the return of the liquidity trap. *Brookings Papers on Economic Activity* 1998(2), 137–87.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lettau, M. and Ludvigson, S. 2004. Understanding trend and cycle in asset values: reevaluating the wealth effect on consumption. *American Economic Review* 94, 276–99.

Lucas Jr., R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Mankiw, N. and Romer, D. 1991. *New Keynesian Economics. Volume 1: Imperfect Competition and Sticky Prices. Volume 2: Coordination Failures and Real Rigidities*. Cambridge, MA: MIT Press.

McCallum, B. 1979. The current state of the policy-ineffectiveness debate. *American Economic Review* 69, 240–5.

McCallum, B. and Nelson, E. 1999. An optimizing IS–LM specification for monetary policy and business cycle analysis. *Journal of Money, Credit, and Banking* 31, 296–316.

Meltzer, A. 1995. Monetary, credit and (other) transmission processes: a monetarist perspective. *Journal of Economic Perspectives* 9, 49–72.

Mishkin, F. 1995. Symposium on the monetary transmission mechanism. *Journal of Economic Perspectives* 9(4), 3–10.

Mundell, R. 1963. Capital mobility and stabilization policy under fixed and flexible exchange rates. *Canadian Journal of Economics and Political Science* 29, 475–85.

Obstfeld, M. and Rogoff, K. 1995. Exchange rate dynamics redux. *Journal of Political Economy* 103, 624–60.

Phelps, E. and Taylor, J. 1977. Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85, 163–90.

Poole, W. 1970. Optimal choice of monetary policy instruments in a simple stochastic macro model.

Quarterly Journal of Economics 84, 197–216.

Roosa, R. 1951. Interest rates and the central bank. In *Money, Trade, and Economic Growth: Essays in Honor of John Henry Williams*. New York: Macmillan.

Rotemberg, J. 1982. Sticky prices in the United States. *Journal of Political Economy* 90, 1187–211.

Sargent, T. and Wallace, N. 1975. ‘Rational’ expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83, 241–54.

Summers, L. 1991. How should long-term monetary policy be determined? *Journal of Money, Credit, and Banking* 23, 625–31.

Svensson, L. 2003. Escaping from a liquidity trap and deflation: the foolproof way and others. *Journal of Economic Perspectives* 17(4), 145–66.

Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39, 195–214.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1, 15–29.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

Ireland, Peter N. "monetary transmission mechanism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000214> doi:10.1057/9780230226203.1125

The New Palgrave Dictionary of Economics Online

money and general equilibrium

Douglas Gale
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The study of money and general equilibrium deals with the integration of monetary theory and the classical theory of value. It includes such topics as the role of money in exchange, the determination of the price level, and the ‘real’ effects of money on the allocation of goods and services.

Keywords

cash-in-advance constraint; classical dichotomy; complete markets; excess demand; financial securities; incomplete markets; intertemporal substitution effect; money; money demand; money in general equilibrium; money supply; nominal prices; optimum quantity of money; Pigou effect; real balance effect; Say's Law; stationary states; temporary equilibrium; uniform tightness property; value theory; Walras's Law; wealth effect

Article

The general equilibrium theory of value, as developed by Walras (1874–77) and his followers, determines the relative prices of goods in terms of non-monetary factors such as technology, preferences, and endowments. Monetary factors are used to determine the nominal price level once relative prices have been determined. Relative prices are determined by the market-clearing conditions for goods whereas the general price level is determined by the market-clearing condition for money. Given a vector of nominal prices $p = (p_1, \ldots, p_\ell)$, the market excess demand functions can be denoted by $f(p) = (f_1(p), \ldots, f_\ell(p))$, where p_h denotes the nominal price of good h and $f_h(p)$ denotes the market excess demand for good h. The functions $f(p)$ are assumed to be homogeneous of degree zero in nominal prices:

$$f(p) = f(tp),$$

for any positive scalar $t>0$. The market-clearing conditions for goods require that the excess demand for each good vanishes at the equilibrium price vector p^*, that is $f(p^*)=0$. These conditions can at most determine relative prices, because if p^* is an equilibrium price vector, then so is tp^*, for any positive scalar $t>0$.

To determine the nominal price level, a demand function for money is introduced. The aggregate demand for money is assumed to be a function of prices $M(p)$. Money demand is homogeneous of degree one in prices:

$$M(tp) = tM(p),$$

for any price vector p and any scalar $t>0$. For any vector of nominal prices p^* satisfying the goods market-clearing condition $f(p^*)=0$, there is a unique value of $t>0$ such that

$$M(tp^*) = \overline{M},$$

where $\overline{M} > 0$ is the exogenous money supply. Thus, once relative prices have been determined by the real factors, the level of nominal prices is determined by monetary factors. This doctrine, which became known as the classical dichotomy, characterized the classical (pre-Keynesian) thinking about monetary economics (see Fisher, 1963, for example).

The integration of monetary theory and the theory of value was stimulated by the appearance of Keynes's General Theory (Keynes, 1936). Pigou (1943) argued that the demand for goods could not be homogeneous of degree zero in prices, because a general fall in prices would increase the real value of money and the wealth effect would in turn increase demand for goods. The Pigou effect (the effect of a general fall in prices on the aggregate demand for goods) is a special case of the real balance effect: that is, the effect of any change in real balances on the aggregate demand for goods. In an attempt to make sense of Keynes's short period analysis, Hicks (1946) introduced the concept of temporary equilibrium, in which prices adjust to clear markets in a particular time period, taking as given expectations about prices in future periods. Building on the work of Hicks and Pigou, Patinkin (1965) argued that the real balance effect is essential for the existence and stability of equilibrium. The classical writers assumed that the market excess demand functions satisfy Say's Law, that is, the value of excess demands for goods sum to zero or

$$p \cdot f(p) = 0,$$

for any price vector *p*. However, Patinkin pointed out that Walras's Law should also be satisfied: that is, the value of the excess demands for goods and money should sum to zero, or

$$p \cdot f(p) + M(p) - \overline{M} = 0,$$

for any price vector *p*. Say's Law and Walras's Law together imply that

$$M(p) = \overline{M},$$

for any price vector *p*. Then homogeneity of the excess demand function *f*(*p*) once again implies that, if p^* is a market-clearing price vector, so is tp^* for any $t>0$ and the price level is once again undetermined. To avoid this indeterminacy, Patinkin argued that there must be a real balance effect: a change in the general price level implies a change in real balances, and hence a change in wealth which must change the demand for commodities. Thus, in a monetary economy the excess demand for goods $f(p, \overline{M})$ is a (homogeneous of degree zero) function of nominal prices and the money supply.

Hahn (1965) pointed out another problem in the theory of monetary equilibrium, viewed from the Walrasian perspective. The problem was the lack of a proof that money has positive value in equilibrium. Hahn observed that the uses of money that might be expected to give rise to a positive demand for money all require money to have positive value in exchange. If the value of money were zero, the economy would be identical to a barter economy. Under the usual assumptions on the excess demand functions, such a non-monetary economy would possess an equilibrium, but it would not be a monetary equilibrium, because money would have no role in exchange.

Grandmont (1983) provided an elegant solution to the problem posed by Hahn (1965). He showed that, while the real balance effect might be necessary, it was not sufficient for the existence of an equilibrium in which the value of money is positive. A strong intertemporal substitution effect is needed as well. Consider an economy in which there are two periods (the present and the future). In the first period, agents buy and sell goods for immediate consumption. They also demand money as a store of value, which they hold until the following period. The value of money is given by an indirect utility function $v(m, p')$, where $m>0$ is the amount of money held until the future and p' is the vector of future nominal prices. An agent's expectations are represented by a probability measure μ on the space of price vectors. Expectations of future prices depend on current prices *p* via the expectation function $\mu = \psi(p)$. Then the expected utility associated with the cash balance *m* is simply the expected value of $v(m, p')$, conditional on the current price vector *p*:

$$v(m, p) = \int \tilde{v}(m, p')\,d\psi(p) .$$

Let $u(x)$ denote the utility associated with the consumption of a vector of current goods x. Then the agent seeks to maximize

$$u(x) + v(m, p)$$

subject to the budget constraint

$$p \cdot x + m \leq p \cdot e + \overline{m},$$

where e is the agent's endowment of goods and $\overline{m}$ his endowment of money. The crucial assumption (sufficient condition) for the existence of an equilibrium in which money has a positive value is that the expectation function $\psi(p)$ satisfies the uniform tightness property: for any number $\varepsilon > 0$ and for every current price vector p, there is a compact set K in the space of positive prices such that $\psi(p)$ assigns probability at least $1-\varepsilon$ to the event that the future price vector p' belongs to K.
While the classical dichotomy cannot hold in the short run, Archibald and Lipsey (1958) argued that it would hold in the long run because the allocation of money balances is endogenous in the long run. This gave rise to the study of stationary states (see Grandmont, 1983).

The cash-in-advance constraint

Introduced by Clower (1967), the cash-in-advance constraint provides a simple motivation for the use of money as a medium of exchange. Lucas (1980) derives the cash-in-advance constraint as follows. Every household is assumed to consist of two agents, one of whom is responsible for selling the household's endowment of goods (for example, supplying labour) and the other is responsible for purchasing goods. At the beginning of each day, the seller sets off for the market with a bundle of goods to sell, while the buyer sets of for a different set of markets to buy the goods they need. Following Clower's dictum that ‘money buys goods and goods buy money but goods do not buy goods’, the buyer needs to have a stock of money at the beginning of the day. The money earned by the seller is not available until the end of the day, so the buyer's purchases are constrained by the amount of money she has at the beginning of the day. The money brought home by the seller must be held until the next day. If $\overline{m}$ is the amount of money held initially and m is the amount carried forward to the next day, the budget constraint can be written as

$$p \cdot x + m \leq p \cdot e + \overline{m}$$

and the cash-in-advance constraint can be written as

$$p \cdot (x - e)^{+} \leq \overline{m},$$

where ξ^{+} denotes the vector consisting of the non-negative part of the vector ξ.
Grandmont and Younes (1973) used a cash-in-advance constraint to study the efficiency of monetary equilibrium. They considered stationary equilibria of an infinite-horizon, pure-exchange economy in which a finite number of individuals i=1,…,I maximize the discounted sum of utilities $\sum_{s=t}^{\infty} \delta^{s-t} u_i(x_i(s))$ subject to a sequence of budget constraints and a cash-in-advance constraint in the form

$$p(t) \cdot (x_i(t) - e_i)^{+} + kp \cdot (x_i(t) - e)^{-} \leq m(t-1),$$

where $0 \leq k \leq 1$. For k=0 this constraint reduces to the Clower–Lucas version. Grandmont and Younes established Friedman's optimum quantity of money result: any laissez-faire, stationary equilibrium of this economy is Pareto inefficient but, if the rate of price deflation equals the subjective rate as time preference, this is sufficient to guarantee that equilibrium is efficient. Grandmont and Laroque (1975) also showed that the payment of interest on money has no effect on efficiency. More precisely, it is the gap between the inflation rate and the interest rate which has an effect, and this is attributable to the lump-sum taxes rather than the interest payments.
The cash-in-advance constraint has played an important role in macroeconomics, particularly in the study of the effect of fiscal and monetary policy (see, for example, Lucas and Stokey, 1983; 1987; Sargent, 1987).

Financial securities

The classical model of general competitive equilibrium assumes that markets are complete. Hart (1975) showed that, with incomplete markets, the existence of equilibrium is no longer guaranteed and the fundamental theorems of welfare economics no longer hold. In Hart's model, incomplete markets are represented by trade in real securities, which are promises to deliver bundles of commodities at some

future date and event. Cass (2006) and Werner (1985) introduced financial securities, whose payoffs are denominated in units of money, and showed that this resolved the existence problem. However, as Balasko and Cass (1989) and Geanakoplos and Mas-Colell (1989) showed, financial securities also introduced indeterminacy of equilibrium. The problem is that a change in the price level in some state changes the real purchasing power of money and hence changes the real payoffs of the financial securities. Magill and Quinzii (1992) pointed out that the indeterminacy arises from the fact that 'money' serves only as a unit of account in the Cass–Werner model. Money has no role in exchange or savings and investment, and hence there is no well defined demand for money.

To address this problem, Magill and Quinzii introduce a cash-in-advance constraint in the spirit of Clower (1967). There are two dates, t=0,1, and S states of nature, s=1,…, S. The state is unknown at date 0; the true state is revealed at date 1. It is convenient to treat the situation at date 0 as another state, denoted s=0. Then each period s is divided into three sub-periods, denoted s_1, s_2, and s_3. In sub-period s_1, agents sell their entire endowment of money to a central exchange and receive money instead. In sub-period s_2, they invest in financial securities (at date 0) and receive dividends (at date 1). In sub-period s_3, they use money to purchase goods from the central exchange. The separation of the sale and purchase of goods between sub-periods s_1 and s_3 forces agents to hold money in equilibrium. Money can also be used to store wealth between periods 0 and 1, but agents will do this only if they anticipate deflation. The supply of money is determined exogenously by the government.

Three main results were established by Magill and Quinzii. First, they showed that, *generically in endowments and money supply, an economy has a finite number of locally unique monetary equilibria*. This means that equilibrium is locally determinate: the well-defined demand for money has eliminated the indeterminacy of the price level. Second, *if money is used as a medium of exchange only, local changes in the money supply have no real effects if the asset markets are complete* – changes in the money supply will change the price level but this will have no effect on the real allocation as long as markets are complete – *whereas, if markets are incomplete, local changes in money supply translate into an S*–1 *dimensional submanifold of real allocations*. When markets are incomplete, any change in the price level implies a change in the real payoffs of the securities, and this translates into a real change in the allocation. Finally, *if money is used as a store of value, local changes in the money supply translate into an S-dimensional submanifold of real allocations in the case of both complete and incomplete markets*. This follows because the use of money as a store of value to transfer wealth between periods implies that the real allocation is directly impacted by changes in the real payoffs from holding money. A related study by Geanakoplos and Dubey (1992) addresses a similar set of questions, but does so in the context of a model with a banking system.

Market games

To provide microeconomic foundations for monetary equilibrium, Shubik (1972) introduced a game that integrates the use of money as a medium of exchange with a generalized Nash–Cournot model of markets. The generalization by Shapley and Shubik (1977) can be summarized as follows. There is an exchange economy with • commodities, indexed by $h = 1, \ldots, \ell$, and I traders, indexed by I=1,…,I. Each

trader is characterized by a consumption set $\mathbf{R}^{\ell}_{+}$, an endowment $e_i \in \mathbf{R}^{\ell}_{+}$, and a utility function $u_i: \mathbf{R}^{\ell}_{+} \to \mathbf{R}$. The utility functions are assumed to be C^1, non-decreasing and concave. We assume that each commodity has a positive aggregate endowment $e_h>0$ and that each individual has a non-zero endowment $e_i>0$.

For simplicity, we assume that traders offer their entire endowment of assets for sale and then bid for the assets they want to hold using fiat money as a means of payment. Each trader i has an endowment of fiat money $m_i>0$. The amount of money he bids for asset h is denoted by $b_{ih}\geq 0$ and the vector consisting of his bids is denoted by $b_i \in \mathbf{R}^{\ell}_{+}$.

A trader cannot bid more money than he holds, so the bid vector chosen by trader i must satisfy the cash-in-advance constraint

$$\sum_{h=1}^{\ell} b_{ih} \leq m_i.$$

The set of bid vectors satisfying the cash-in-advance constraint for trader i is denoted by B_i, where it is understood that the initial balance m_i is exogenously given.

For any strategy profile $b=(b_1,\ldots, b_I)$, define an attainable allocation of commodities as follows. Let the price of commodity h be denoted by $p_h(b)$ and defined by

$$p_h(b) = \frac{b_h}{e_h},$$

where $b_h \equiv \Sigma_{i=1}^{I} b_{ih}$ and $e_h = \Sigma_{i=1}^{I} e_{ih}$. Then let the quantity of commodity h received by trader i be denoted by $\xi_{ih}(b)$ and defined by

$$\xi_{ih}(b) = \begin{cases} b_{ih} / p_h & \text{if } p_h > 0 \\ 0 & \text{if } p_h = 0 . \end{cases}$$

Then the commodity bundle achieved by i for any strategy profile b is denoted by $\xi_i(b)$. It is easy to see that the I-tuple $\{\xi_i(b)\}$ is an attainable allocation for any $b\in B$.

The traders must return their initial balances of fiat money to the government at the end of the game. This means that trader *i* must end the trading period with at least m_i units of money. We assume that any choice of b_i resulting in end-of-period money balances that are lower than m_i will yield a payoff of $-\infty$. The terminal balance for trader *i* equals his initial balance m_i minus the sum of his bids $\sum_{h=1}^{\ell} b_{ih}$ plus the revenue from the sale of his initial portfolio $p(b)\cdot e_i$. It is easy to show that the terminal balance satisfies

$$m_i - \sum_{h=1}^{\ell} b_{ih} + p(b) \cdot e_i = m_i - p(b) \cdot (\xi_i(b) - e_i),$$

so the terminal constraint is satisfied if and only if $p(b) \cdot (\xi_i(b) - e_i) \leq 0$. For any strategy profile *b*, let trader *i*'s payoff be denoted by $\pi_i(b)$ and defined by

$$\pi_i(b) = \begin{cases} u_i(\xi_i(b)) & \text{if } p(b) \cdot (\xi_i(b) - e_i) \leq 0, \\ -\infty & \text{if } p(b) \cdot (\xi_i(b) - e_i) > 0. \end{cases}$$

Shapley and Shubik (1977) demonstrate the existence of a Nash equilibrium for this game under the additional assumption that for each commodity *h* there are at least two individuals whose utility is increasing in that commodity. They also provide conditions under which the equilibrium allocation converges to a competitive equilibrium as the number of traders increases without bound.

Concluding remarks

As Joseph Ostroy wrote in the first edition of *The New Palgrave* (1987, p. 515),

> We shall argue that the incorporation of monetary exchange tests the limits of general equilibrium theory, exposing its implicitly centralized conception of trade and calling for more decentralized models of exchange.

That comment is just as true today as it was then, and remains the great challenge for economists who want to develop more satisfactory models of the process of monetary exchange at the level of the economy as a whole.

See Also

- monetary approach to the balance of payments
- monetary cranks
- monetary policy, history of
- money illusion
- money supply

Bibliography

Archibald, G. and Lipsey, R. 1958. Monetary and value theory: a critique of Patinkin and Lange. *Review of Economic Studies* 26, 1–22.

Balasko, Y. and Cass, D. 1989. The structure of financial equilibrium: I. exogenous yields and unrestricted participation. *Econometrica* 57, 135–62.

Cass, D. 2006. Competitive equilibrium with incomplete financial markets. *Journal of Mathematical Economics* 42, 384–405.

Clower, R. 1967. A reconsideration of the microfoundations of monetary theory. *Economic Inquiry* 6, 1–8.

Fisher, I. 1963. *The Purchasing Power of Money*, rev. edn. New York: Kelley.

Geanakoplos, J. and Mas-Colell, A. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47, 22–38.

Geanakoplos, J. and Dubey, P. 1992. The value of money in a finite-horizon economy: a role for banks. In *Economic Analysis of Markets and Games*, ed. P. Dasgupta et al. Cambridge, MA: MIT Press.

Grandmont, J.-M. 1983. *Money and Value: A Reconsideration of Classical and Neoclassical Monetary Theories*. Cambridge: Cambridge University Press; Paris: Maison des Sciences de l'Homme.

Grandmont, J.-M. and Laroque, G. 1975. On money and banking. *Review of Economic Studies* 42, 207–36.

Grandmont, J.-M. and Younes, Y. 1973. On the efficiency of monetary equilibrium. *Review of Economic Studies* 40, 149–65.

Hahn, F. 1965. On some problems of proving the existence of an equilibrium in a monetary economy. In *The Theory of Interest Rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11, 418–43.

Hicks, J. 1946. *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory*, 2nd edn. Oxford: Clarendon.

Keynes, J. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Lucas, R. 1980. Equilibrium in a pure currency economy. *Economic Enquiry* 18, 203–20.

Lucas, R. and Stokey, N. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12, 55–93.

Lucas, R. and Stokey, N. 1987. Money and interest in a cash-in-advance economy. *Econometrica* 55, 491–513.

Magill, M. and Quinzii, M. 1992. Real effects of money in general equilibrium. *Journal of Mathematical Economics* 21, 301–42.

Ostroy, J.M. 1987. Money and general equilibrium. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Patinkin, D. 1965. *Money, Interest, and Prices: An Integration of Monetary and Value Theory*, 2nd edn. New York: Harper and Row.

Pigou, A.C. 1943. The classical stationary state. *Economic Journal* 53, 343–51.

Sargent, T. 1987. *Dynamic Macroeconomic Theory*. Cambridge, MA: Harvard University Press.

Shapley, L. and Shubik, M. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 85, 937–68.

Shubik, M. 1972. Commodity money, oligopoly, credit and bankruptcy in a general equilibrium model. *Western Economic Journal* 10, 24–38.

Werner, J. 1985. Equilibrium in economies with incomplete financial markets. *Journal of Economic Theory* 36, 110–19.

How to cite this article

Gale, Douglas. "money and general equilibrium." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000219> doi:10.1057/9780230226203.1128

The New Palgrave Dictionary of Economics Online

money illusion

Peter Howitt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The term money illusion is commonly used to describe any failure to distinguish monetary from real magnitudes. It seems to have been coined by Irving Fisher, who defined it as ‘failure to perceive that the dollar, or any other unit of money, expands or shrinks in value’ (1928, p. 4). To Fisher, money illusion was an important factor in business-cycle fluctuations. Rising prices during the upswing would stimulate investment demand and induce business firms to increase their borrowing, thus causing a rise in the nominal rate of interest. Lenders would accommodate them by increasing their savings in response to the rise in the nominal rate, not taking into account that, because of the rise in inflation, the real rate of interest had not risen but had actually fallen (Fisher, 1922, esp. ch. 4).

Beginning with Haberler (1941, p. 460, fn. 1) other writers have used the term money illusion as synonomous with a violation of what Leontief (1936) called the ‘homogeneity postulate’, the postulate that demand and supply functions be homogeneous of degree zero in all nominal prices; that is, that they depend upon relative prices but not upon the absolute price level. This usage differs from Fisher's in two senses. It refers to people's reactions to a change in the level of prices rather than to a change in the rate of change of prices, and it is cast in operational terms, as a property of potentially observable supply and demand functions rather than as a property of people's perceptions or lack thereof.

Patinkin (1949) objected to the latter use of the term money illusion on the grounds that it failed to take into account the real balance effect. A doubling of all money prices should affect household demand functions even if people are perfectly rational and suffer from no illusions, because it reduces at least one component of the real wealth that constrains their demands – the real value of their initial money holdings. Accordingly he defined the absence of money illusion as the zero-degree homogeneity of net demand functions in all money prices and the money values of initial holdings of assets.

In a fiat money economy in Hicksian temporary equilibrium, under the assumption of static expectations, the absence of money illusion in Patinkin's sense is operationally equivalent to the assumption of rational behaviour, in the following sense. Let each agent's demand functions $\hat{x}_i(p_1, \ldots, p_n, W)$ for goods i=1,•…,•n, together with his demand-for-money function $M(p_1, \ldots, p_n, W)$ be defined as the maximizers of the utility function $U(x_1, \ldots, x_n, M, p_1, \ldots, p_n)$ subject to the budget constraint: $p_1x_1 + \ldots + p_nx_n + M = W$, where W is initial nominal wealth. The utility function includes M and

the money prices p_i because M is assumed to yield unspecified services whose value depends upon the vector of prices expected to prevail next period, and those expected prices are proportional to today's prices.

A rational agent would realize that a proportional change in M and all prices would leave unaffected the purchasing power of M, and thus also the services rendered by M. Accordingly U is said to be illusion-free if it is homogeneous of degree zero in $(M, p_1, \ldots, p_n)$. This homogeneity property was first assumed explicitly in the context of demand theory by Samuelson (1947, p. 118) although it was implicit in the earlier analysis of Leser (1943), who used the equivalent formulation: $U(x_1, \ldots, x_n; M/p_1, \ldots, M/P_n)$. It is easily verified that the $\hat{x}'s$ are illusion-free in Patinkin's sense if and only if they can be derived from an illusion-free U (see Howitt and Patinkin, 1980).

The assumption of static expectations is crucial to this equivalence. If expected future prices were not proportional to current prices then a proportional change in $p_1, \ldots, p_n$, W would alter intertemporal relative prices and it would not be irrational for the agent to respond by changing his demands.

Patinkin's original definition can be generalized to take this possibility into account and to allow for the presence of productive non-money assets by requiring demand functions for real goods to be unaffected by a proportional change in W, all current prices, and all expected future prices, holding constant the rates of return on all non-money assets. If future prices p_i' were uncertain then current demands would depend upon the probability distribution $F(p_i', \ldots, p_n')$, and the proportional change in future expected prices in the above statement would have to be replaced by a change from $F(p_i', \ldots, p_n')$ to $F_\lambda(p_i', \ldots, p_n') \equiv F(p_i'/\lambda, \ldots, p_n'/\lambda)$ where λ is the factor of proportionality.

The absence of money illusion is the main assumption underlying the long-run neutrality proposition of the quantity theory of money. But the presence of money illusion has also frequently been invoked to account for the short-run non-neutrality of money, sometimes by quantity theorists themselves, as in the case of Fisher. On the other hand, many monetary economists have reacted adversely to explanations based on such illusions, partly because illusions contradict the maximizing paradigm of microeconomic theory and partly because invoking money illusion is often too simplistic an explanation of phenomena that do not fit well into the standard equilibrium mould of economics. Behaviour that seems irrational in a general equilibrium framework may actually be a rational response to systemic coordination problems that are assumed away in that framework.

Thus, for example, Leontief (1936) attributed Keynes's denial of the quantity theory to an assumption of money illusion. He interpreted Keynes as saying that the supply of labour depended upon the nominal wage rate whereas the demand depended upon the real wage. A rise in the price level would thus raise the equilibrium quantity of employment. Leijonhufvud (1968, ch. 2) questioned this interpretation and argued that Keynes was dealing with information problems that don't exist in Leontief's general equilibrium analysis. Specifically, Leijonhufvud argued that workers might continue supplying the same amount of labour services in the event of a rise in the general price level, not because they irrationally identified nominal with real wages but because in a world of less than perfect information it would take time for them to learn of the changed value of money.

Likewise, Friedman (1968) objected to the then standard formulation of the Phillips-relation between unemployment and the rate of wage-inflation. Friedman argued that the rate at which firms raised their

wage offers and households raised their reservations wages, given any existing amount of unemployment, should depend upon these agents' expectations of the future value of money. To assume otherwise would be to assume money illusion. Friedman's argument implied that an expected-inflation term should be added to the usual specification of the Phillips curve. His analysis of the expectations-augmented Phillips curve was similar to Leijonhufvud's imperfect-information argument.

More recently, Barro (1977) has argued against the assumption of nominal wage stickiness in the work of Fischer (1977) and others, on the grounds that microeconomic theories of wage contracts imply that these contracts should be signed in real, not nominal terms, unless people suffer from money illusion. Although monetary economists have thus been reluctant to attribute money illusion to private agents they have not hesitated to attribute it to governments. Indeed, as Patinkin (1961) demonstrated, money illusion on the part of the monetary authority is necessary for an economy to possess a determinate equilibrium price level. More recently, several writers have attributed real effects of inflation to money-illusion in tax laws (e.g., Feldstein, 1983). Specifically, in many countries interest income and expenses are taxed at the same rate regardless of the rate of inflation, and historical money costs rather than current replacement costs are used for evaluating inventories and calculating depreciation allowances. Because of these effects inflation can distort the after-tax cost of capital.

In short, the attitude of economists to the assumption of money illusion can best be described as equivocal. The assumption is frequently invoked and frequently resisted. The persistence of a concept so alien to economists' pervasive belief in rationality indicates a deeper failure to understand the importance of money and of nominal magnitudes in economic life. This failure is evident, for example, in the lack of any convincing explanation for why people persist in signing non-indexed debt contracts, or why the objective of reducing the rate of inflation, even at the cost of a major recession, should have such wide popular support in times of high inflation.

See Also

- neutrality of money
- real balances

Bibliography

Barro, R.J. 1977. Long-term contracting, sticky prices, and monetary policy. *Journal of Monetary Economics* 3(3), July, 305–16.

Feldstein, M. 1983. *Inflation, Tax Rules, and Capital Formation*. Chicago: University of Chicago Press.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85(1), February, 191–205.

Fisher, I. 1922. *The Purchasing Power of Money*. 2nd edn, New York: Macmillan.

Fisher, I. 1928. *The Money Illusion*. New York: Adelphi.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1), March, 1–17.

Haberler, G. 1941. *Prosperity and Depression*. 3rd edn, Geneva: League of Nations.

Howitt, P. and Patinkin, D. 1980. Utility function transformations and money illusion: comment. *American Economic Review* 70(3), September, 819–22.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press.

Leontief, W. 1936. The fundamental assumptions of Mr Keynes' monetary theory of unemployment. *Quarterly Journal of Economics* 5(November), 192–7.

Leser, C.E.V. 1943. The consumer's demand for money. *Econometrica* 11(2), April, 123–40.

Patinkin, D. 1949. The indeterminacy of absolute prices in classical economic theory. *Econometrica* 17 (1), January, 1–27.

Patinkin, D. 1961. Financial intermediaries and the logical structure of monetary theory. *American Economic Review* 51(1), March, 95–116.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

How to cite this article

Howitt, Peter. "money illusion." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000225> doi:10.1057/9780230226203.1129

The New Palgrave Dictionary of Economics Online

money supply in the American colonies

Farley Grubb
From The New Palgrave Dictionary of Economics Online, , 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The British colonies in North America experimented with legislature-issued paper monies to supplement their specie monies which were in chronic short supply. These experiments were designed to produce inside monies that, unlike specie monies, could not profitably be exported. The nature of British regulation, while leaving room for some variation, constrained the colonies to issuing fiat currencies that were typically tied to paying the future taxes and other dues levied by the issuing colonies. After some early failures, most of these experiments performed well over the quarter century before the Revolution as revealed by the presence of long-run price stability.

Keywords

backing or asset theory of money; barter; commodity money; fiat currency; Benjamin Franklin; inside money; John Law; purchasing power parity; quantity theory of money; real bills doctrine

Article

The money supply in the British North American colonies was a complex mixture of colonial legislature-issued inside fiat paper monies and outside specie monies. Gold and silver coins (specie) were the principal local and international monies of exchange for Europeans. In British North America such monies could only be acquired through trade or government transfers as gold and silver were not yet mined there. Specie was acquired through trade surpluses with Spanish America and the Caribbean. In addition, British military spending, especially during King George's War (1744–48) and the Seven Years' War (1756–63), injected specie into colonial economies. This specie, however, quickly flowed out to cover the colonies' trade deficits with Britain. The British government used mercantilist policies to prevent specie outflows from Britain and encourage specie inflows by holding their colonies in a state of chronic trade deficit with the mother country. As specie passed through the colonies it could only serve in a limited capacity as a local medium of exchange. Given the frequent disruptions to trade flows caused by wars, weather shocks and, in the decade before the Revolution, political boycotts, periods of specie dearth and glut in the colonies were not uncommon. Colonists often complained of a lack of specie. As a result, extensive barter systems using merchant book credit and non-specie commodity monies, such as tobacco, developed to support local exchange. These barter systems were never completely displaced by monetized exchange during the colonial era (Brock, 1975; Grubb, 2004; Grubb, 2008; McCusker, 1978; Mossman, 1992; Rabushka, 2008).

Ironically, the relative efficiency of colonial barter was likely responsible for the chronic scarcity of specie. Individual colonists could not capture the positive externalities resulting from the lowered transaction costs in all subsequent local trades that used their specie (as 19th-century banks could by loaning banknotes fractionally backed by specie reserves), and they gained more by quickly exporting their specie to buy British goods. Exchanging specie for an entire British good was more welfare-enhancing than the gain from lowered transaction costs in using that specie to acquire a good in local exchange, relative to using barter to obtain that good. As such, the colonies could only gain the lowered transaction costs that a monetary medium of exchange could bring if they could create a medium that could not to any great advantage be exported – a pure inside money. Colonial experiments with paper money, which were also called bills of credit, can be understood in this light.

During the 18th century the British North American colonies became the first western economies to rely on legislature-issued fiat paper monies as their principal internal media of exchange. These monetary experiments were neither uniform nor coordinated across the colonies. They were instituted piecemeal – at different times with different motivations and goals. Their institutional structures and relative performances varied as well. These experiments, while wide-ranging, were constrained by British regulations that effectively limited the colonies' rights and abilities to mint their own coins, institute capital controls to limit specie exports, and form private corporations that could effectively function as banks using specie as reserves to support private paper money emissions (which dominated 19th-century US paper money creation) (Brock, 1975; Grubb, 2006; Grubb, 2008; Mossman, 1992; Rabushka, 2008).

The Massachusetts legislature in 1690 was the first to issue paper money, followed by South Carolina in 1703, New York, New Jersey, and New Hampshire in 1709, Rhode Island and Connecticut in 1710, North Carolina in 1712, Pennsylvania and Delaware in 1723, Maryland in 1733, and finally Virginia and Georgia in 1755. In the first eight cases, paper money was created as a solution to the short-run fiscal crises caused by emergency military spending during King William's (1689–97) and Queen Anne's (1702–13) wars. Emergency spending during the Seven Years' War led Georgia and Virginia to create their paper monies. For paper money to become a permanent medium of exchange in the peacetime economies of these colonies was not necessarily the motive behind these experiments, although for many colonies it evolved in that direction (Brock, 1975; Newman, 1997; Rabushka, 2008).

In 1723, Pennsylvania and Delaware became the first colonies to initiate paper money systems that were not motivated by wartime fiscal crises. Their goal was to ameliorate internal economic crises caused by temporary depressions in their overseas trade balances. The paper money was to be removed from circulation by the end of the decade – presumably once the trade depression had passed. In 1729, paper-money advocates in Pennsylvania, such as Benjamin Franklin, won the debate to renew and expand the paper money experiment and turn it into a more or less permanent feature of the peacetime economy of these colonies (Grubb, 2006). In 1733, Maryland became the third colony to initiate a paper money system that was not motivated by a wartime fiscal crisis. From the beginning, Maryland's paper money experiment was intended to be a permanent restructuring of the medium of exchange within the colony. Its goal was tied to transforming the transatlantic tobacco trade, which in turn required demonetarization of tobacco within the colony – Maryland's principal non-specie commodity money (Grubb, 2008). The only colony issuing paper money subsequently to return to a specie standard for the remainder of the colonial period was Massachusetts in 1750, largely because of the rapid inflation that accompanied its paper emissions used to support military operations in King George's War (Officer, 2005).

Following the English system, colonial paper currencies were denominated in pounds, shillings and pence (except for Maryland's money after 1766), but with the unit-of-account or proclamation exchange rate (par rate) to pounds sterling typically set higher than one-to-one. This par rate differed among the colonies, so for example there were 1.33 Maryland pounds and 1.67 Pennsylvania pounds to one pound sterling. Exchange rates to pounds sterling fluctuated considerably around, and sometimes departed from, these par rates. Some colonial legislatures made their paper money a legal tender for all transactions within their jurisdiction, some only for a subset of transactions, and some only for public debts (taxes). Some colonial currencies stated a specie exchange rate on their face, as did some early paper monies issued by New England colonies, Georgia and New Jersey, and Maryland paper money after 1766, but most did not (McCusker, 1978; Newman, 1997; Rabushka, 2008). Colonial legislatures did not otherwise fix or defend an exchange rate between their paper money and specie coins.

No colonial government succeeded in consistently exchanging its paper money on demand for specie coins in its colony, and most did not even try. Colonial treasuries did not typically keep specie reserves and so could not effectively act like banks. Colonial paper currencies seldom circulated far beyond the issuing colony's borders in any substantial quantities for any considerable period of time. Cross-colony and cross-oceanic trade was consummated using specie coins – the outside money – or credit in specie, often through bills of exchange.

As such, colonial paper monies are best thought of as inside monies on floating exchange rates to outside (specie) monies and to each other. They were true fiat currencies: that is, backed by nothing other than the promise that nominal taxes and mortgage payments owed to a government could be paid in the paper money issued by that government. Colonial legislatures passed taxes when issuing their paper monies that could be paid with these paper monies, or held mortgages on their subjects' lands in exchange for loaning them these monies. Often paper money emissions were redeemed via these tax and mortgage payments within a few years, with the money burnt upon redemption. This emission-redemption structure gave an immediate contemporaneous use and nominal anchor to colonial paper monies which supported their face value in current local exchange (Brock, 1975; Rabushka, 2008).

Several colonies, such as Rhode Island, New Jersey, Pennsylvania and Maryland, issued and redeemed portions of their paper money through land banks. Subjects borrowed paper money from their governments, pledging their lands as collateral. They could pay their mortgage principal and interest either in specie or in the paper money of their government, with the interest earned being an important source of income for some colonial governments. The amount any subject could borrow relative to the total sums available was typically restricted so that borrowings would be widespread (Rabushka, 2008; Thayer, 1953). In 1729, Benjamin Franklin argued that this land-bank method created a flexible money supply that passively expanded and contracted with the economy's money demand. This in turn produced a market-clearing monetary equilibrium that prevented excess quantities from being in circulation relative to demand and so prevented the paper money from depreciating. In essence, Franklin's argument was a primitive statement of the real bills doctrine, and may have been the first, and possibly only, statement of this idea by an American writer in the colonial era (Grubb, 2006). Franklin may have taken the idea from John Law's 1705 Scottish land-bank pamphlet which was reprinted in London in 1720. Franklin visited London in the early 1720s. Whether colonial land banks actually functioned effectively as Franklin argued is yet to be conclusively determined.

The quantity of paper money in the initial authorization across colonies averaged 0.6 sterling-equivalents per capita, and ranged between 0.1 and 2.0. Thereafter, it stayed within this range, but averaged closer to 1.0. It only systematically exceeded 2.0 sterling-equivalents per capita in Rhode Island (1714–46) and only briefly spiked above 2.0 in New Hampshire and New York during King George's War and the Seven Years' War, respectively. By contrast, the US money stock from 1795 to 1830 in sterling-equivalents per capita hovered between 1.4 and 2.2, with an average around 1.8 (Rousseau, 2006).

The early experiments were often less than successful. The South Carolina pound in the late 1720s, the Maryland pound between 1736 and 1760, the Virginia pound between 1756 and 1765, and the Massachusetts pound in the 1740s suffered substantial depreciations. The British Parliament's response to the Massachusetts crisis was the Currency Act of 1751, which allowed colonies to issue paper money as long as it met two conditions: (1) that it not be a legal tender, and (2) that ample provisions (taxes) be put in place to redeem each issue within a reasonable time. While this Act applied only to New England, the Virginia crisis in the early 1760s led Parliament in 1764 to extend a version of the 1751 Act to all the colonies (Brock, 1975; Ernst, 1973; Rabushka, 2008).

These early struggles were caused by excessive emissions relative to expected redemptions, which in turn were caused by perceived mismanagement in some cases and the overwhelming burden of war in other cases. The structure and backing of a paper emission could also affect its performance. For example, unlike other colonies, the 1733 Maryland paper pound was to be redeemed at par in specie by the Maryland government at designated future dates via a sinking fund. Most of the emission was handed out to its subjects in exchange for destroying trash tobaccos. Use of the money to pay contemporaneous local taxes was thwarted. Thus its value rested principally on the promised payoff in specie, one-third in 1748 and two-thirds in 1764. The colony taxed tobacco exports and invested the money in Bank of England stock at a rate that would generate the sums needed to meet the promised payoff, which the colony successfully executed. This structure, however, meant that the contemporaneous value of the Maryland pound relative to its par (face) value would track its present discounted value relative to its redemption date. In effect the Maryland pound was a zero-interest bearer bond (Grubb, 2008; Rabushka, 2008). Figure 1 shows this outcome, albeit with a lot of volatility around the discount trend early on (McCusker, 1978). The Continental paper dollar issued by Congress during the American Revolution, of all colonial paper monies, most closely resembled the 1733–64 Maryland pound. Before the Revolution, Benjamin Franklin had noted the performance of the Maryland pound, as depicted in Figure 1, with disapproval. While a general supporter of colonial paper money, his unexplained objection to the Continental dollar might be because its structure closely resembled that of the 1733–64 Maryland pound (Grubb, 2006). The idea, and occasionally practice, of having colonial (and later Continental) paper monies pay interest came from an effort to counterbalance the discount off their face value that resulted when they were backed by a bond-like redemption structure.

Figure 1
The value of Maryland's paper money, 1734–1765

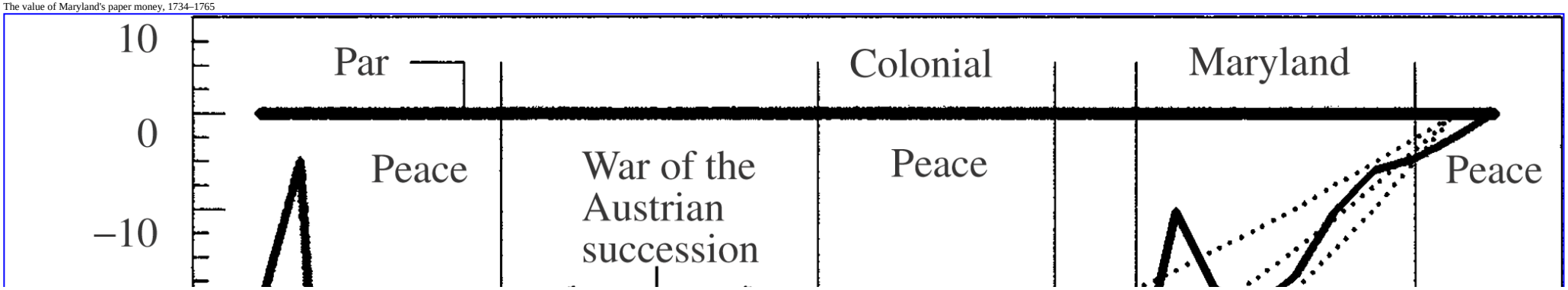

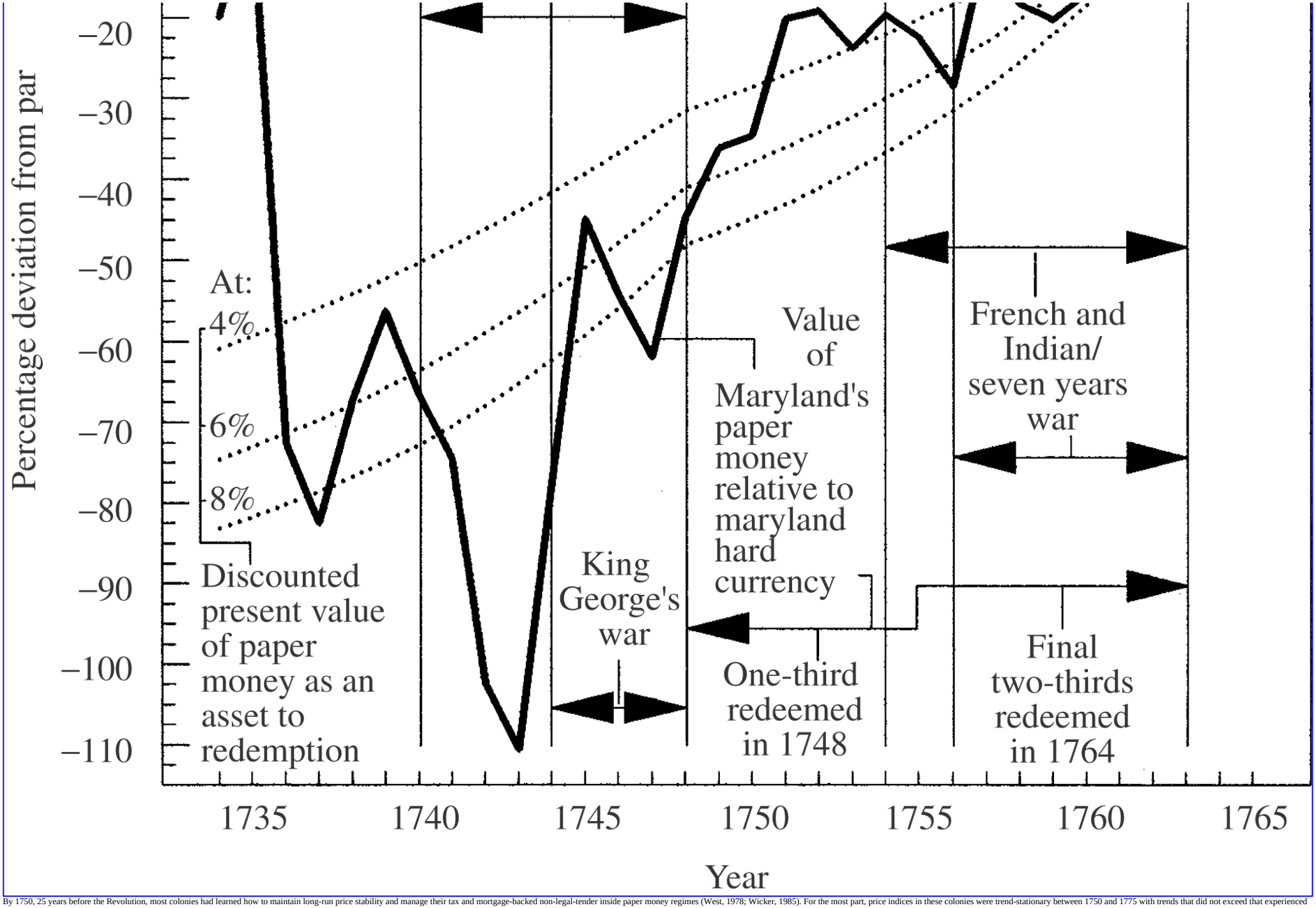

By 1750, 25 years before the Revolution, most colonies had learned how to maintain long-run price stability and manage their tax and mortgage-backed non-legal-tender inside paper money regimes (West, 1978; Wicker, 1985). For the most part, price indices in these colonies were trend-stationary between 1750 and 1775 with trends that did not exceed that experienced by colonies on a specie standard only (Grubb, 2003). In addition, for the most part, each colony's paper money exchange rate to pounds sterling was stationary, and purchasing power parity cannot be rejected between each colony and between each colony and England. This performance is consistent with colonial legislatures successfully using their emission-redemption backing structures within a long-run quantity theory of money framework to manage their macro-economies (MV=PY where M=money supply, V=velocity of money circulation, P=price level, Y=real output, and where the growth in V and Y are long-run constants). Substantial short-run volatility in velocity and real output per capita (which equals real money balances per capita) was still present (Rousseau, 2007). For example, Figure 2 shows the movement in real money balances per capita (ln(M/P*Pop) where Pop=population) for paper plus specie (total) money in Pennsylvania from 1729 to 1775. This series exhibits no trend and is stationary with a three-year half-life to shocks (Grubb, 2004).

Figure 2

The movement in per capita real money balances in Pennsylvania, 1929–1775

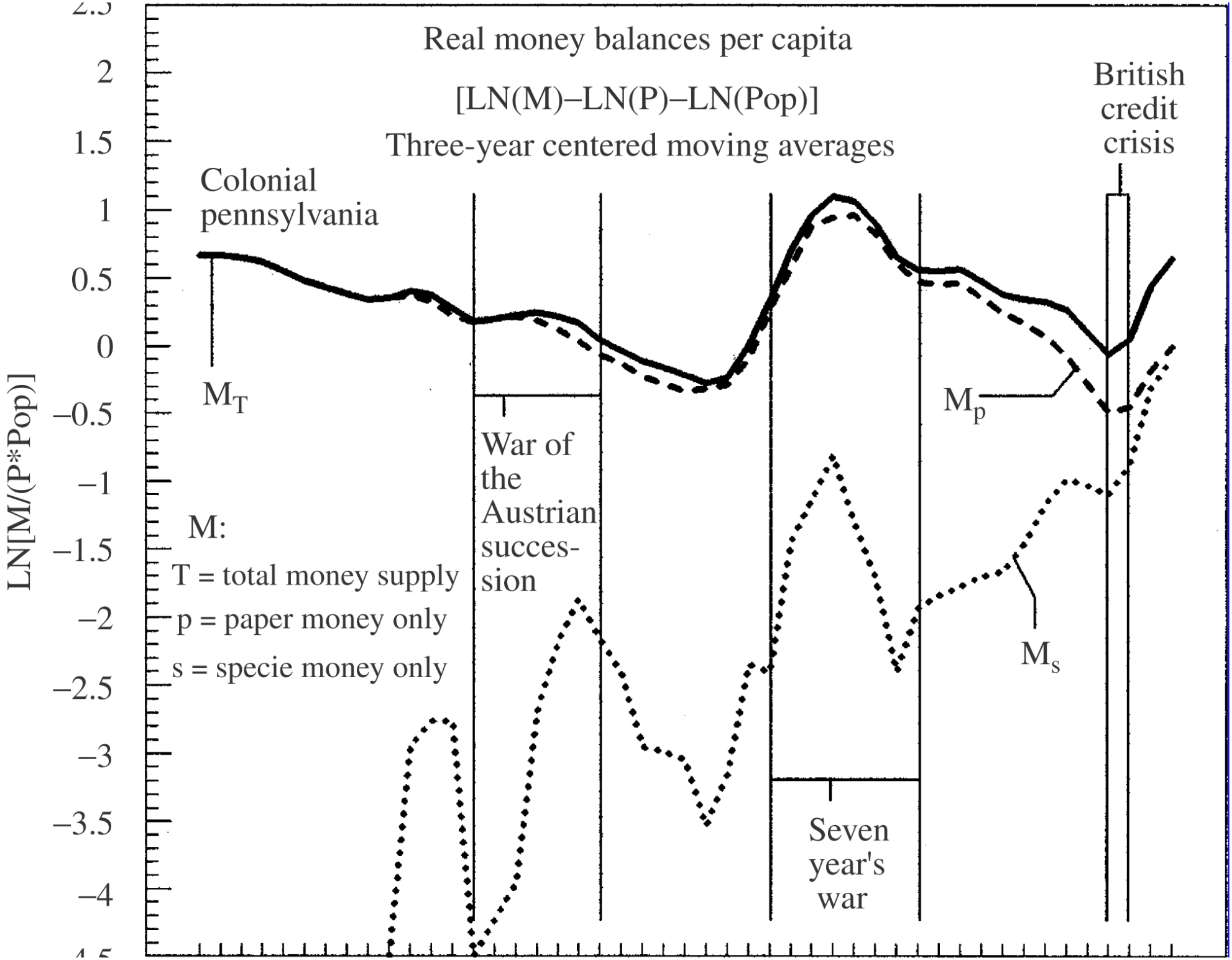
Real money balances per capita
[LN(M)–LN(P)–LN(Pop)]
Three-year centered moving averages
British credit crisis
Colonial pennsylvania
LN[M/(P*Pop)]
2.5
2
1.5
1
0.5
0
–0.5
–1
–1.5
–2
–2.5
–3
–3.5
–4
4.5
M_T
War of the Austrian succes-sion
M_p
M:
T = total money supply
p = paper money only
s = specie money only
M_s
Seven year's war

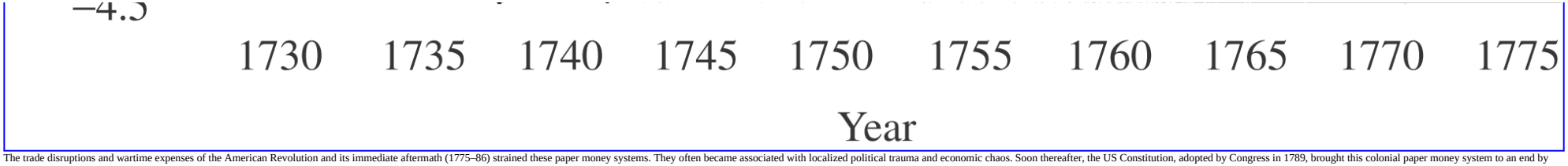

The trade disruptions and wartime expenses of the American Revolution and its immediate aftermath (1775–86) strained these paper money systems. They often became associated with localized political trauma and economic chaos. Soon thereafter, the US Constitution, adopted by Congress in 1789, brought this colonial paper money system to an end by constitutionally barring national and state legislatures from issuing paper monies.

See Also

- barter
- commodity money
- fiat money
- inside and outside money
- Law, John
- money
- quantity theory of money
- purchasing power parity
- real bills doctrine

Bibliography

Brock, L.V. 1975. *The Currency of the American Colonies*, 1700–1764. New York: Arno.

Ernst, J.E. 1973. *Money and Politics in America*, 1755–1775. Chapel Hill, NC: University of North Carolina Press.

Grubb, F. 2003. Creating the U.S.-dollar currency union, 1748–1811: a quest for monetary stability or a usurpation of state sovereignty for personal gain? *American Economic Review* 93, 1778–98.

Grubb, F. 2004. The circulating medium of exchange in colonial Pennsylvania, 1729–1775: new estimates of monetary composition, performance, and economic growth. *Explorations in Economic History* 41, 329–60.

Grubb, F. 2006. *Benjamin Franklin and the Birth of a Paper Money Economy*. Philadelphia: Federal Reserve Bank of Philadelphia. Available at: <http://www.philadelphiafed.org/education/index.html#publications, accessed 8 November 2008.

Grubb, F. 2008. Creating Maryland's paper money economy, 1720–1739: the role of power, print, and markets. NBER working paper #13974. Available at: <http://0-www.nber.org.library.lemoyne.edu/papers/w13974, accessed 8 November 2008.

McCusker, J.J. 1978. *Money and Exchange in Europe and America*, 1600–1775. Chapel Hill, NC: University of North Carolina Press.

Mossman, P.L. 1992. *Money of the American Colonies and Confederation: A Numismatic, Economic and Historical Correlation.* New York: American Numismatic Society.

Newman, E.P. 1997. *The Early Paper Money of America*, 4th edn. Iola, WI: Krause.

Officer, L.H. 2005. The quantity theory in New England, 1703–1749: new data to analyze an old question. *Explorations in Economic History* 42, 101–21.

Rabushka, A. 2008. *Taxation in Colonial America*. Princeton, NJ: Princeton University Press.

Rousseau, P.L. 2006. A common currency: early US monetary policy and the transition to the dollar. *Financial History Review* 13, 97–122.

Rousseau, P.L. 2007. Backing, the quantity theory, and the transition to the US dollar, 1723–1850. *American Economic Review–Papers and Proceedings* 97, 266–70.

Thayer, T. 1953. The land-bank system in the American colonies. *Journal of Economic History* 13, 145–59.

West, R.C. 1978. Money in the colonial American economy. *Economic Inquiry* 16, 1–15.

Wicker, E. 1985. Colonial monetary standards contrasted: evidence from the Seven Years' War. *Journal of Economic History* 45, 869–84.

How to cite this article

Grubb, Farley. "money supply in the American colonies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_M000418> doi:10.1057/9780230226203.1891

The New Palgrave Dictionary of Economics Online

money supply

Benjamin M. Friedman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Governments supply money not only for use in everyday transactions but also, in the modern era, in order to influence their economies. In most advanced industrialized economies the demand for money is sufficiently unstable to make the quantity of money supplied, or its growth rate, an unreliable guide to how monetary policy influences either prices or real economic activity. Most central banks therefore set a designated interest rate, not the quantity or growth of money supplied. But because money supply and money demand help determine market interest rates, the money supply process remains essential to analysing how monetary policy operates.

Keywords

aggregate demand; central bank independence; central bank reserves; central banking; central banks; Friedman, M.; Goodhart's Law; hyperinflation; inflation; interest rates; monetary base; monetary policy; monetary policy rules; money; money demand; money multiplier; money supply; open market operations; optimal monetary policy; reserve-to-deposit ratio.

Article

Supplying money for use in everyday transactions, so as to obviate the need for cumbersome barter, has been a function of governments for more than 2,000 years. Not surprisingly, government-issued money, once in existence, rapidly became a store of value as well. As an aspect of the history of human society and institutions, the process by which governments supply money has naturally attracted substantial attention. But the primary interest in money supply within the discipline of economics has stemmed from the proposition that movements in money are an important – according to some views, the most important – determinant of movements in prices, in output and employment, and in other economic phenomena of well-established interest on their own account.
Two analytical frameworks that rose to prominence in the latter half of the 20th century – indeed, that

dominated macroeconomic thinking during much of that period – attached just this importance to money: quantity-theory monetarism, and IS–LM Keynesianism. Both these frameworks, however, took for granted that governments conduct their affairs (specifically in this context, that central banks conduct monetary policy) in such a way as to create independent movements in the supply of money, as opposed to merely passive movements in response to changes in money demand that therefore could not plausibly be the cause of movements in either prices or real economic activity. As of the outset of the 21st century, however, the number of central banks that in fact carry out their responsibilities in such a way is small and shrinking. Instead, most central banks implement monetary policy by setting some designated short-term interest rate.

As a result, interest in how money is supplied has sharply diminished among economists, and the details of the money supply process are now often omitted from the standard economics curriculum. (Examples at the graduate level are the instructional text by David Romer, 2006, and the theoretical treatise by Michael Woodford, 2003.) In the absence of some substantive knowledge of how money is supplied, however, just how a central bank can set ‘the interest rate’ would remain mysterious. Even if the number of central banks that actively seek to influence money supply as an element of the conduct of monetary policy shrinks to zero, therefore, money supply is unlikely to disappear from the purview of economics altogether.

The analytical basics

The first recognized monies supplied by governments for ordinary economic use mostly consisted of precious metals. The authorities’ role was to provide standardized units, together with what amounted to stamped certification that the amount of metal in the coin or other object conformed. Apart from the certification, therefore, anyone who had an adequate quantity of the chosen metal could supply money along with the government.

In the more modern conception of money supply, relevant only since the 19th century, money is a form of debt. Most government-issued money consists of currency, which represents the liability of a partly or wholly government-owned central bank. Currency is typically not interest-bearing, and so the motives for holding it do not stem from its role as an earning asset. And although it is the government's (the central bank's) liability, in modern times it usually does not represent an obligation on the government's part to pay the bearer in some other form. Instead, both private citizens and businesses hold these government liabilities for their convenient use in everyday transactions, normally enforced by their statutory status as legal tender.

The fact that government-issued money is supplied as the liability of the central bank, and the presumption that the central bank has control over its balance sheet, together create the conceptual foundation for viewing the supply of money as a tool of economic policy. Indeed, much of the initial interest in this subject in the modern era arose from the experience of countries where the central bank had lost control of its balance sheet for some period of time, often in the aftermath of war or under other circumstances that prevented the government from raising ordinary revenues to cover its ongoing expenditures. The observation that such episodes often led to spiralling hyperinflation, with rising prices requiring the government to issue more money (in the absence of other revenues) and the larger supply of money leading to further increases in prices, immediately suggested a connection between money supply and prices, if not real economic activity as well.

Apart from situations of runaway money supply and hyperinflation, however, the issuance of currency is usually not the focus of economists' interest in how the supply of money relates to economic activity. While the great majority of government-issued money in the economically advanced countries now consists of currency held by the public (as of 2006, 69 per cent in the United Kingdom, and 95 per cent in the United States), currency is nonetheless only a small part of the money that individuals and firms use for savings and to execute everyday economic transactions. The money that individuals and firms use mostly consists of deposits issued by banks and other financial institutions. In the United Kingdom, deposit money outweighs currency by more than 30 to 1. Even in the United States, where the country's currency is also commonly used in both legal and illegal transactions around the world, the ratio is more than 8 to 1. Moreover, although in principle a central bank could seek to influence the economy by manipulating how much currency it supplies, in practice most central banks supply currency passively to accommodate whatever demands the public may have. (The role of currency issuance as a source of government finance – the heart of most examples of hyperinflation – is likewise limited in most economically advanced countries. Even in the United States, with demand for the currency enlarged by the use of US dollars in other countries, issuance of currency in a typical year amounts to only one to two per cent of the federal government's spending.) The simple construct of an economy in which the public depends entirely on government-issued currency to execute economic transactions, and the central bank exerts its economic influence by expanding or contracting the supply of that currency, is a textbook instructional device with limited relevance to most actual economies.

From the perspective of any active connection to either nonfinancial economic activity or the pricing of assets in the financial markets, therefore, what matters is the larger money supply issued by banks and other depository institutions (hereafter simply 'banks' for short). And in most modern banking systems, what gives the central bank the ability to influence the volume of deposits that banks in the aggregate create is its control over the amount of its own liabilities that it supplies for banks to hold. While most of the central bank's liabilities consist of currency held by the public, the remainder (31 per cent in the United Kingdom, and only five per cent in the United States, as of 2006) are held as assets – normally called 'reserves' – by the banks. The link between the banks' creation of deposits for the public to hold and their own holdings of reserves at the central bank constitutes the heart of the money supply process for purposes of a connection to most matters of concern to monetary policy.

Banks hold central bank reserves – and, importantly, hold more reserves as they have more deposits outstanding (all other things equal) – for several reasons. First, in traditional 'fractional reserve' banking systems, banks are required by law to hold such reserves in amounts equal to at least some fixed percentage of their outstanding deposits. Hence a larger supply of reserves makes it possible for the banks to do more lending (or buy more securities) and therefore create more money. Conversely, contracting the supply of reserves requires banks to shrink the amount of deposits they have outstanding, normally by not extending new loans to replace existing credits that mature or are otherwise repaid, or by selling securities.

Second, banks need a supply of currency to satisfy customers who draw on their accounts or present checks or other negotiable instruments for payment. In some banking systems, currency held by banks (as opposed to currency held by the public) is counted as part of banks' reserves. When a customer cashes a check, therefore, bank reserves fall and there is a corresponding increase in currency held by the public. (Because the central bank is not a party to the transaction, the total amount of central bank liabilities remains unchanged.) But banks cannot satisfy such demands unless they are holding an

adequate amount of currency to begin with. And the greater the bank's volume of business, including in particular the amount of deposits it has outstanding against which its customers may want to draw, the more currency – hence the more reserves, if bank-held currency counts as reserves – the bank will ordinarily hold.

Third, banks also need to settle transactions with one another. If a customer of one bank deposits a check written against an account at another bank, the two banks must transfer some asset from one to the other. The same is true if one bank sells a security to another. Although banks in most countries have various mechanisms, like private clearing houses, for effecting such transfers without involving the central bank, some inter-bank transactions do normally settle by transferring reserves at the central bank from the paying bank to the receiving bank. In order to participate in that process, banks therefore need to hold at least some amount of reserves; and the more deposits the bank has outstanding, the more inter-bank transactions it may have to settle on a given day, and so the more reserves it will ordinarily hold. Moreover, in some banking systems the central bank reinforces the demand for its reserves by requiring banks to settle certain classes of inter-bank transactions in this way. Especially in systems where there are no reserve requirements in the traditional form of a stated minimum percentage of outstanding deposits, requiring the banks to settle inter-bank transactions in this way reinforces the banks' need to hold central bank reserves.

Banks' demand for reserves, therefore, is in many ways analogous to the public's demand for money. Reserves provide banks with an ability to do business, just as the money that individuals and nonfinancial firms hold enables them carry out their everyday economic affairs. That ability has value, but not infinite value. Hence the more expensive it is for banks to hold reserves, in terms of interest forgone by holding reserves instead of some other asset, the more banks will seek to economize on their reserve holdings in relation to their outstanding volume of deposits. For a given amount of deposits, therefore, banks' demand for reserves is negatively elastic with respect to the interest rate on alternative assets (typically loans or securities), just as the public's demand for money is negatively interest elastic for a given amount of income being earned or transacting being done. If reserves at the central bank bear an interest rate that varies in close step with what banks can get from holding other earning assets, this negative interest elasticity is likely to be small, or even trivial. But if the interest rate that the central bank pays on reserves is fixed (in the United States, for example, it is fixed at zero), or even if it varies together with market returns but only imperfectly, the negative interest elasticity in banks' reserve demand is likely to be significant. (The classic paper making this point is Dewald, 1963.)

The analytical mirror image of banks' negatively elastic demand for reserves, for a given volume of deposits outstanding, is their positively elastic willingness to create deposits for a given amount of reserves that they hold. The higher are market interest rates on earning assets, compared to whatever rate the central bank pays on reserves, the greater is the incentive for banks to stretch their reserves further by making more loans and buying more securities – and in the process creating more deposits – rather than leaving an increasingly expensive cushion of reserves that may provide benefits (less risk of having to take abrupt action in the event of a shortfall, for example) but are costly nonetheless.

The result is a positively interest-elastic supply of money, representing the behaviour of banks, to go along with the usual negatively interest-elastic demand for money representing the behaviour of the households and firms that hold bank deposits, together with currency, as the money that they use for economic purposes. In the absence of some pathology, the intersection of this positively interest-elastic money supply and negatively interest-elastic money demand determines the equilibrium quantity of

money created and held, for a given supply of reserves and a given level of income, together with the interest rate at which the market clears. (And, because the positively interest-elastic supply of money is simply the mirror image of the negatively interest-elastic demand for reserves – both represent the same aspect of banks' behaviour – the market for reserves is likewise in equilibrium, with demand equal to whatever quantity of reserves the central bank is supplying, at the same interest rate.) Integrating this partial equilibrium of the money market (and the reserves market) with the demand for goods and services then completes a simple representation of the economy's aggregate demand. Further integrating that aggregate demand representation with aggregate supply, importantly including the labour market, in turn completes the economy's short-run general equilibrium (short-run in that such dynamic elements as the stocks of capital, technology, and other relevant factors are still unaccounted for).

In some treatments of money supply within the economics literature, this explicit supply–demand equilibrium in the markets for money and reserves is, instead, implicitly represented by a simple 'money multiplier' stating the relationship between the total liabilities supplied by the central bank – often called the 'monetary base' – and the resulting amount of money, including bank deposits as well as currency. Purely as a matter of arithmetic, specifying the ratio of reserves to deposits that the banks choose to hold (influenced in part by whatever reserve requirements and other institutional strictures banks face), and the ratio of currency to deposits that the public chooses within its holdings of money, is sufficient to determine the quantity of money that goes along with any given monetary base set by the central bank. But the banks' reserve-to-deposit ratio depends in part on interest rates as well, and the public's demand for currency often varies with a host of factors (confidence in the banking system, use of currency abroad or for purposes of illegal transactions, and so on), so that the 'money multiplier' representation is really just a short-hand simplification that works well or badly depending on the strength of the relevant interest elasticities and the extent of variation in interest rates and the many other factors involved. (See, for example, Cagan, 1965. A brief statement of the central ideas appeared in Friedman and Schwartz, 1963, ch. 2, sec. 4.) Underneath, the supply–demand equilibrium established by the central bank's supply of reserves, banks' behaviour in demanding reserves and supplying deposits, and the public's behaviour in demanding both deposits and currency, is what establishes an economy's money supply. (For a fully articulated treatment, see Modigliani, Rasche, and Cooper, 1970.)

The link to monetary policy

The logical starting point in this process is the central bank's supply of its own liabilities, and it is the central bank's control over the liabilities it issues that gives the supply of money its place in economic policy. Until fairly recently – well into the 19th century – governments issued either coins or paper currency mostly as a means of payment for goods and services they purchased. Such actions were, in effect, a combination of what have come to be known as fiscal and monetary policies. In the modern era, however, especially with the advent of central banks as distinct and often quasi-independent governmental institutions, economists have thought of fiscal and monetary policies as likewise distinct. In the absence of a securities market, or some similar set of financial institutions, it is difficult to conceive of how monetary policy would operate independently of fiscal policy: how could the government, in such a setting, increase the amount of money outstanding without simultaneously making either a purchase or at least a transfer payment? One metaphor sometimes used in the theoretical economics literature to represent such an action – and which only serves to indicate how far-fetched

such a situation is – is to picture the government dropping money from a helicopter. While monetary and fiscal policies are distinguishable in most modern economies, central banks, of course, do not drop money from helicopters. The reason is that the economies in which they operate in fact have securities markets.

The primary means by which central banks in most modern economies change the amount of their liabilities outstanding is to purchase, or sell, securities – actions typically called ‘open market operations’. When the central bank buys a security, it makes payment by increasing the amount of reserves credited to the seller's bank. (In systems in which bank-held currency is counted as part of reserves, the consequence is the same even if the central bank makes payment by delivering currency to the seller's bank.) When the central bank sells a security, it correspondingly receives payment by reducing the amount of reserves credited to the buyer's bank. In either case, the central bank's assets, consisting mostly of the securities it holds, and its liabilities, consisting partly of the reserves credited to banks, rise or fall in lockstep. But because of the ways in which banks’ ability to create deposits depends on their holdings of reserves, the change is not economically irrelevant. Changes in the supply of reserves, effected via open market operations, shift a key underpinning of the equilibrium in the reserves market and the money market, thereby changing not only the resulting quantity of money but the yields and prices of non-money assets and ultimately the equilibrium of the nonfinancial economy as well.

Not all open market operations carried out by central banks change the quantity of reserves. Most importantly, the central bank also needs to accommodate the public's changing demand for currency. In a growing economy with rising prices, the demand for currency is usually increasing. When individuals and businesses go to their banks to get more currency, their doing so increases the amount of currency in public circulation but reduces the amount of the banks’ reserves (as long as bank-held currency is counted as reserves). As a part of their normal ongoing procedures, therefore, most central banks routinely purchase securities – that is, carry out open market operations – in order to offset such reductions in reserves due to increasing public demand for currency. Central banks also regularly carry out open market purchases or sales in order to prevent short-run fluctuation in other technical factors, such as international transactions and variations in the amount of checks currently in the clearing process, from affecting the supply of reserves.

Central banks can also create reserves by lending to banks, rather than buying earning assets from them, and in some countries’ systems the lending of reserves is more important for purposes of carrying out monetary policy than open market operations. Whether banks distinguish between reserves that they have borrowed from the central bank and reserves that they simply own outright (often called ‘nonborrowed reserves’ to distinguish the two) depends on the specifics of the individual system's institutions. Most obviously, borrowed reserves are a liability of the bank, on which it presumably has to pay interest, while its nonborrowed reserves are an asset on which it may or may not earn interest. In addition, in some systems (the United States, for example), borrowing reserves from the central bank exposes a bank to regulatory oversight with implicit costs well beyond what the interest rate paid would suggest.

Whether reserves are borrowed or nonborrowed, however, the essence of monetary policy is the central bank's provision of reserves to the banking system. The recognition of the way in which that role played by the central bank potentially affects an economy's money supply, interest rates, asset prices, nonfinancial activity, and prices and wages, in turn sets the stage for both normative and positive consideration of monetary policy. The ensuing economics literature has become vast. In most countries

the corresponding public discussion is likewise active and intense.

The modern economist most identified with emphasizing the role of money supply in the conduct of monetary policy – as opposed to focusing on interest rates, or measures of reserves in the banking system, or other relevant indicators of what a central bank is doing in this respect – is Milton Friedman. At the most fundamental normative level, Friedman advocated a long-run policy of shrinking the supply of money (by which he meant government-issued money) at a rate adequate to render nominal interest rates on assets closely substitutable for money equal to zero on average over time. The basic logic was that, since the government could create such money at essentially no cost, it should be costless for the public to hold; the public's effort to economize on holdings of money balances, when market interest rates on money substitutes are positive, represents a deadweight loss to the economy (see Friedman, 1969). Given the demonstrated dangers of deflation, however – with a positive real rate of interest, negative inflation would be necessary to achieve a zero average nominal interest rate – this recommendation had little impact on actual monetary policy.

At a more practical level, however, over short- and medium-run horizons Friedman advocated keeping the supply of money (by which he meant the deposits and currency held by the public) growing at a constant rate. Here the argument was that the influence of monetary policy on both prices and real economic activity operates with lengthy delays, subject to unpredictable variation, and that active attempts by the central bank to use monetary policy to offset nonmonetary influences on the economy were likely to be destabilizing (see Friedman, 1953; 1956). Many other economists, more optimistic about the prospects for using active variation in monetary policy to blunt the influence on the economy of factors that the central bank could either foresee or at least recognize quickly once they had occurred, followed Friedman in advocating the use of growth in the money supply as the way to gauge whether the central bank was exerting a stimulative or a contractionary force on economic activity. Beginning in the 1960s, but more so in the 1970s, many central banks around the world implemented these recommendations by adopting one or another form of explicit target for the growth of its money supply.

The role of empirical evidence

The crucial empirical underpinning of such policy frameworks, whether they involved constant money growth or attempts at active stabilization nonetheless benchmarked by money growth, was the observation that movements in money bore a reliable relationship to movements in income and prices. Early in the post-Second World War period, Philip Cagan documented such a relationship between money growth and price inflation in several well-known episodes of hyperinflation in Europe that had followed each of the two world wars (Cagan, 1956). But hyperinflation in the context of post-war chaos (especially for the war's losers) bore only limited implications for the conduct of monetary policy under more normal circumstances. In a massive historical study, Milton Friedman and Anna J. Schwartz documented the relationships between money and prices, and also money and income, for the United States during the period 1867–1960 – including the Great Depression of the 1930s but also many more ordinary business fluctuations as well – and following their work many other empirical researchers attempted similar (though mostly smaller-scale) studies for other countries and other time periods (Friedman and Schwartz, 1963).

At the conceptual level, the central idea linking this empirical research to the implied role of money

supply in conducting monetary policy was that, if fluctuations in money growth and fluctuations in income and/or prices are systematically related, and if the observed fluctuations in money growth within those relationships represent independent movements of money supply, then the central bank can exploit those relationships by purposefully steering the money supply along an optimally chosen course (which may or may not be a simply constant-growth path). Following the work of Friedman and Schwartz, and the many other researchers who applied ever more sophisticated empirical methodologies to the same line of enquiry, questions about each of these two underlying issues – how strong the observed relationships are, and whether they result from independent movements of money supply – generated a similarly large literature.

One immediate difficulty, recognized early on, is that, since money supply necessarily equals money demand, inferences about the money–income or money–price relationship on the basis of observed movements in money are subject to the usual problem of statistical identification. (An early paper making this point was Teigen, 1964. Another, addressed more explicitly to the work of Friedman and Schwartz, was Tobin, 1970.) Hence what may look like a relationship between movements of prices and income induced by movements in money supply may in reality be movements in money demand induced by movements in prices and income. Further, unless the central bank takes its decisions affecting money supply with no regard for the behaviour of prices and income, the observed relationships may also represent the reactive behaviour of the central bank itself. Indeed, under some plausible accounts of how central banks make monetary policy, relationships of the kind observed in the data would spuriously emerge. (An early paper making this point was Goldfeld and Blinder, 1972.) Still more fundamentally, even if the relationships observed between money and either income or prices actually did represent exactly the kind of causal influence of money supply that was claimed, the attempt by the central bank to exploit such a relationship for policy purposes, once widely recognized, could cause the relationship to change or even break down altogether. (The classic statement of this proposition in a general context is Lucas, 1976. For a formulation in the specific context of monetary policy, see Goodhart, 1984; the original formulation of ‘Goodhart's Law’ dates to 1975 when this paper was first presented.)

Starting in the mid-1970s, however, and then increasingly so over the next two decades, these questions became moot. Fluctuations in money growth no longer appeared to bear much observed relation to fluctuations in either income or prices over time horizons that were useful for conducting monetary policy, especially after controlling for other obvious information like past movements of income and prices themselves. In parallel, the evidence indicated that money demand was unstable. The presumption of a stable functional relationship between money demand and income or prices had always been central to the claim that money supply was a useful tool for purposes of monetary policy. But now evidence for a stable money demand gave way, in one country after another, to evidence of instability.

The reasons for the disappearance of stable money demand were many, and, at a qualitative level, straightforward to understand. (The empirical money demand literature is a separate subject; for a survey, see Goldfeld and Sichel, 1990. For an earlier survey, written before the instability became so widespread or so evident, see Laidler, 1977, ch. 7.) One reason was changing regulation (in the United States, for example, the removal of the prohibition against banks’ paying interest on checkable deposits, and also of the ceilings limiting the interest that banks could pay on interest-bearing savings deposits). Another, in part prompted by regulatory changes, was innovation in the kinds of deposits and deposit-

like instruments that banks and other financial institutions offered their customers (for example, money market mutual funds). A third was the electronic revolution, which made various forms of financial transactions ever easier and less costly (for example, shifting funds between checkable and noncheckable accounts). A fourth was rapid globalization, which made businesses in particular, but many individuals as well, increasingly willing to hold assets, and to borrow, in multiple currencies, and to substitute readily among them. But regardless of the precise reasons, which presumably varied from one country to another, money demand no longer appeared to be stable. Nor, in parallel, did the relationships of a simpler form between money and either income or prices that had spurred policy interest in money supply in the first place.

The decline of money supply as a tool of monetary policy

In the absence of empirical evidence of stable money demand, the rationale for the role of money supply as a tool of monetary policy collapsed as well. If money demand is unstable, then even perfectly stable money supply introduces into income and prices the influence of whatever disturbances to the public's money-holding behaviour occur. Under those circumstances, the central bank can do a better job of stabilizing either prices or income, over the short or medium run, by fixing some interest rate and thereby allowing fluctuations in money supply to accommodate fluctuations in money demand that occur for reasons unrelated to movements of income and prices. (The classic paper making this point is Poole, 1970; for a survey of the optimal monetary policy literature along these lines, including the role of money supply behaviour along with money demand, see Friedman, 1990. In the long run, however, there must be at least some absolute nominal element in the policy mechanism to anchor the price level; the interest rate is a relative price, not an absolute price.) Following the increasing evidence of money demand instability, and the collapse of money–price and money–income relationships, that is precisely what an increasing number of central banks have done.

The experience in the Unites States is illustrative. The Federal Reserve System, the US central bank, first began to take explicit note of money supply movements in formulating its monetary policy in 1970. In 1975 the US Congress adopted a resolution requiring the Federal Reserve to announce, in advance, quantitative targets for the growth of key money (and credit) aggregates and, after the fact, to report to the relevant Congressional oversight committees on its success or failure in meeting these targets. In 1979 the Federal Reserve publicly declared an intensified dedication to controlling money growth, with the main focus on the narrow M1 aggregate (consisting primarily of currency and checkable deposits), and adopted new day-to-day operating procedures, centred on the supply of nonborrowed reserves, designed to enhance its ability to achieve control of M1.

The movement towards ever greater emphasis on money supply in US monetary policy took less than a decade; unwinding it took only a little longer. In 1982, the Federal Reserve recognized the increasing instability of demand for M1 and shifted its focus to the broader M2 (including not only currency and demand deposits but also most forms of time and savings deposits). Soon thereafter, it abandoned its operating system based on nonborrowed reserves, in favour of simply setting the federal funds rate (the overnight interest rate on bank reserves) at the level most likely to achieve the desired M2 growth. After 1986 the Federal Reserve stopped setting a target for M1 growth, but continued to do so for M2 and M3 (a still broader aggregate). In the late 1980s evidence based on how the Federal Reserve changed the federal funds rate in response to observed movements of money suggested that the M2 growth target still

bore significant influence on US monetary policy. (See, for example, Friedman, 1997; but the empirical literature on this issue is voluminous.)

That influence had mostly dissipated by 1990, and in 1993 the Federal Reserve publicly ‘downgraded’ the role of its M2 target. Thereafter it continued to set ‘ranges’ for M2 and M3 growth, but it made clear that these were not actual money growth targets; they were merely ‘intended to communicate its expectation as to the growth of these monetary aggregates that would result’ under specified assumed conditions. In 1998 the Federal Reserve further confirmed that these ranges were not ‘guides to policy’. In 2001 it stopped setting such ranges altogether.

The pattern in most other countries was roughly parallel. By 1980 the use of money supply targets for monetary policy was an idea whose time had come. Most of the major central banks had put such targets at the core of their policymaking process. By 1990 money growth targets were already largely a thing of the past. By the mid-1990s most central banks had either de-emphasized such targets or dropped them altogether. By 2000 it had become standard that central banks carry out monetary policy by setting some short-term interest rate. Money supply mostly disappeared from public discussion, and the professional economics literature largely dispensed with the now-unnecessary apparatus of money demand, money supply, and likewise demand and supply in the market for reserves. (See, for example, Clarida, Gali and Gertler, 1999.)

Implicitly, however, that conceptual apparatus nonetheless stands behind the ability of central banks to set the designated interest rate in the first place. In principle, a central bank – or anyone else with large enough resources, for that matter – could fix the price or yield on any asset simply by buying or selling that asset in sufficient volume to shift the entire market equilibrium, ultimately including the real returns established by the fundamental economic forces of thrift and productivity. (Given the lags with which monetary policy influences price inflation, in the short run the interest rate the central bank is setting is a real interest rate.) But in fact most central banks normally move the interest rate they use for monetary policy purposes by executing only very small transactions, and in an increasing number of cases they do so without executing any transactions at all; often the mere announcement of what the central bank would like the designated rate to be is sufficient.

What gives a central bank the ability to do so is, presumably, market participants’ knowledge that the interest rate being set is closely tied to that on the central bank's own liabilities (in systems like that in the United States, it is exactly that rate), and that the central bank can make the supply of those liabilities whatever it chooses. But market equilibrium requires that the demand for those liabilities equal the supply, and the demand for central bank liabilities in turn is an aspect of the same behavioural process that determines the supply of money. Hence money supply remains a part of the story, even if now mostly a hidden one.

See Also

- Friedman, Milton
- inside and outside money
- monetary and fiscal policy overview
- monetary policy, history of
- monetary transmission mechanism

- money

Bibliography

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Cagan, P. 1965. *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. New York: NBER.

Clarida, R., Gali, J. and Gertler, M. 1999. The science of monetary policy: a new Keynesian perspective. *Journal of Economic Literature* 37, 1661–1707.

Dewald, W.G. 1963. Free reserves, total reserves and monetary control. *Journal of Political Economy* 71, 141–53.

Friedman, B.M. 1990. Targets and instruments of monetary policy. In *Handbook of Monetary Economics*, vol. 2, ed. B.M. Friedman and F. Hahn. Amsterdam: North-Holland.

Friedman, B.M. 1997. The rise and fall of money growth targets as guidelines for U.S. monetary policy. In *Towards More Effective Monetary Policy*, ed. I. Kuroda. London: Macmillan.

Friedman, M. 1953. The effects of a full-employment policy on economic stability: a formal analysis. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Friedman, M. 1956. The quantity theory of money: a restatement. In *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Friedman, M. 1969. The optimum quantity of money. In *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press.

Goldfeld, S.M. and Blinder, A.S. 1972. Some implications of endogenous stabilization policy. *Brookings Papers on Economic Activity* 1972(3), 585–644.

Goldfeld, S.M. and Sichel, D.E. 1990. The demand for money. In *Handbook of Monetary Economics*, vol. 2, ed. B.M. Friedman and F. Hahn. Amsterdam: North-Holland.

Goodhart, C. 1984. Problems of monetary management: the U.K. experience. In *Monetary Theory and*

Practice: The U.K. Experience. London: Macmillan.

Laidler, D.E. 1977. *The Demand for Money: Theories and Evidence*, 2nd edn. New York: Harper & Row.

Lucas, R.E., Jr. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A.H. Meltzer. Amsterdam: North-Holland.

Modigliani, F., Rasche, R. and Cooper, J.P. 1970. Central bank policy, the money supply and the short-term rate of interest. *Journal of Money, Credit and Banking* 2, 166–218.

Poole, W. 1970. Optimal choice of monetary policy instruments in a simple stochastic macro model. *Quarterly Journal of Economics* 84, 197–216.

Romer, D. 2006. *Advanced Macroeconomics*, 3rd edn. Boston: McGraw-Hill/Irwin.

Teigen, R.L. 1964. Demand and supply functions for money in the Unites States: some structural estimates. *Econometrica* 32, 467–509.

Tobin, J. 1970. Money and income: post hoc ergo propter hoc? *Quarterly Journal of Economics* 84, 301–17.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

Friedman, Benjamin M. "money supply." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000236> doi:10.1057/9780230226203.1130

The New Palgrave Dictionary of Economics Online

money, classical theory of

Roy Green
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An integral part of the classical theory of value and distribution, the classical theory of money emerged largely in response to the issue of the relationship between changes in the money supply and the price level. This issue was central to the Price Revolution of the 16th and 17th centuries, the Napoleonic war inflation and the industrial crises of the mid-19th century. It was not the existence of an empirical correlation that was in dispute, but the direction of causation. A solution would therefore require a *theoretical* approach as well as knowledge of the facts.

Keywords

Bank Charter Act 1844 (UK); Bank Restriction period; Banking School; Bullion Report (1810) (UK); bullionist controversies; Cantillon, R.; classical dichotomy; classical law of circulation; classical theories of distribution; classical theory of money; convertibility; credit; Currency School; equation of exchange; fiduciary money; Hume, D.; inflation; law of reflux; Law, J.; metallic money; monetarism; money, classical theory of; money supply; paper money; Price Revolution; quantity theory of money; real bills doctrine; Say's Law; specie-flow mechanism; Steuart, J.; Tooke, T.; Torrens, R.; value theory

Article

The classical theory of money is an integral part of the classical theory of value and distribution; and its conceptual categories have real counterparts in historical experience. These categories begin with metallic money and progress to the more complex forms of fiduciary money and credit.

Classical framework

The equation of exchange forms a common point of reference for all approaches to monetary theory, since the relationships it expresses simply constitute a truism and do not in themselves imply causality:

$MV = PT$, where *M* denotes the money supply, *V* the velocity of circulation, *P* an index of prices and *T* the number of commodity transactions. This equation may also be written: $MV = PY$, where *Y* denotes total output, the index *P* is correspondingly adjusted and *V* no longer reflects the circulation of a stock of commodities but the rate of expenditure of a flow of income (corresponding to the flow of output). We use this alternative formulation to specify the classical approach to monetary theory. The only difference of substance is the replacement of the sum of commodity transactions with a measure of net output over a given period, hence excluding non-produced assets (such as land) from the exchange process.

The classical theory of money was developed largely as a response to the practical issue of the relationship between changes in the money supply and the price level. This issue was central to three historical episodes which form the background to our discussion: the Price Revolution of the 16th and 17th centuries, the Napoleonic war inflation and the industrial crises of the mid-19th century. It was not the existence of an empirical correlation that was in dispute, but the direction of causation. A solution would therefore require a *theoretical* approach as well as knowledge of the facts.

The basic structure of the solution arose from discussion of the Price Revolution. Instead of augmenting wealth in the manner suggested by mercantilist doctrine, the influx of gold and silver from the newly discovered American mines seemed only to devalue the unit of account. An immediate interpretation was offered by the quantity theory of money, which attributed the increase in the price level throughout Europe entirely to monetary expansion. According to David Hume, money had no intrinsic value and was simply a means of circulation, in which capacity it served simultaneously as money of account (1752, p. 33). This approach ‘essentially amounted to treating money not as a commodity but as a voucher for buying goods’ (Schumpeter, 1954, p. 313). Once in circulation, money acquired merely a ‘fictitious value’, whose magnitude was established by demand and supply (Hume, 1752, p. 48; also Montesquieu, 1748, pp. 50–1; Vanderlint, 1734, pp. 2–3; Locke, 1691, p. 233).

Classical economists, by contrast, treated money as a *real commodity*, whose value was determined like other commodities by the labour time socially necessary for its production (Petty, 1963, vol. 1, pp. 43–4; Smith, 1776, p. 24; Ricardo, 1821, pp. 85–6). They traced the cause of the Price Revolution not to monetary phenomena but to lowered production costs at the mines (Nef, 1941; Outhwaite, 1969, esp. p. 29; Vilar, 1976, esp. p. 343). It followed that, in the long run, when economic activity is regulated by permanent forces, the magnitude of *P* in the equation of exchange is determined on the basis of value theory and both *Y* and *V* are fixed due to Say's Law and institutional factors respectively. Hence *P* is the independent variable in the equation and *M* the dependent variable. Any movement in *P* as a result of changes in the production costs of commodities (or money) has a commensurate effect on *M*. This determination of aggregate monetary requirements in the ‘real’ sector of the economy became known as the ‘classical dichotomy’ and constitutes the basic classical law of circulation (Petty, 1963, vol. 1, p. 36; Smith, 1776, pp. 332–3; Ricardo, 1923, p. 158; Marx, 1867, pp. 123–4). In other words, causation runs from prices to money in classical economics and not the reverse as we find in both traditional quantity theory and neoclassical monetarism (Eatwell, 1983; Green, 1982). All things being equal, ‘The quantity of money that can be employed in a country must depend on its value’ (Ricardo, 1821, p. 352). The type of money employed in the circulation process has no bearing on this conclusion, since *V* will be determinate whatever *its* numerical value.

Had the scope of classical economics extended no further than the study of permanent economic forces, the question of whether it possessed a ‘quantity theory of money’ would not have arisen. But the

limitations of a long-run approach in explaining concrete developments and formulating relevant policies convinced most classical writers to take into account the role of temporary factors. In particular, the effect of exogenous changes in the money supply needed to be explained. Now the problem became complicated by the definition of money and the nature of financial organization. If Say's Law kept *Y* constant, only two possibilities remained open: a price adjustment, that is, a change in *P*, or a quantity adjustment, that is, a change in *V* (by hoarding or dishoarding). This was the essence of the division among the classical economists. One group was led by Ricardo and included the bullionists (that is, supporters of the 1810 Bullion Report), and later, the Currency School. The other group comprised the anti-bullionists and the Banking School and was given qualified approval by Marx.

The dominant Ricardian group held consistently that both *Y* and *V* were always fixed. The quantity ‘theory’ of money was therefore no theory at all in this view, but simply a logical outcome of assuming Say's Law. The inflationary process was seen as the transitional mechanism by which monetary deviations were corrected: ‘That commodities would rise or fall in price, in proportion to the increase of diminution of money, *I assume as a fact which is incontrovertible*’ (Ricardo, 1923, p. 93 fn., emphasis added).

The opponents of quantity theory, on the other hand, were prepared to sacrifice logical consistency in an attempt to interpret the real events with which they were confronted. Their often pioneering expositions generally placed the weight of adjustment on *V*, although the extent was seen as contingent upon the composition of *M* – whether the money supply was metallic, fiduciary or credit. The flaw in their approach was their failure to overthrow Say's Law and develop an analysis of the saving–investment process, that is, a theory of output. Had they done so, their challenge to the incorporation of quantity theory into classical economics may have been more successful.

Currency and credit

By the time the Bank of England suspended cash payments in 1797, a body of principles on the role and behaviour of paper money had already been formed. The collapse of Law's system led to considerable discussion which culminated in Smith's authoritative exposition of banking in the *Wealth of Nations*. There Cantillon's view was accepted – as against Law and Steuart – that banking could not increase the quantity of capital but only its turnover (Smith, 1776, p. 246). This accorded with the given output assumption of Say's Law. It was also established that paper money would not depreciate provided its total amount did not exceed the value of gold and silver that would otherwise have circulated at any given level of economic activity (1776, p. 227).

More contentiously, Smith argued that the economic convertibility of paper and metallic money could be maintained not only by enforcing legal convertibility but also by having banks adopt the practice of discounting ‘real bills’, that is, securities backed by real assets (1776, p. 239 and *passim*). This became known as the ‘real bills doctrine’. It was repudiated first by Thornton and then by Ricardo and the Currency School, but rehabilitated as the ‘law of reflux’ by the Banking School.

The Bank Restriction period was marked by high inflation accompanied by a rise in the market price of bullion over its mint price. This indicated a depreciation of paper currency in terms of the monetary standard, a phenomenon which could not have existed when convertibility was enforced by law. The central problem was to explain the appearance of this premium on bullion, and to find a principle whose practical implementation would restore and maintain economic convertibility, thus ensuring that the

bank notes conformed to the behaviour of metallic currency. The explanation which gained widest acceptance was based upon the quantity theory of money. It was presented officially in the Bullion Report and then developed by Ricardo. The remedy for inflation implied by this approach was control over the money supply by the authorities.

Ricardo began his analysis by recognizing the need to replace gold and silver in the sphere of circulation by paper – provided only that it was issued in the same amount, that is, the amount prescribed by the value of the metal which served as the monetary standard: ‘A currency is in its most perfect state when it consists wholly of paper money, but of paper money with an equal value with the gold which it professes to represent’ (Ricardo, 1821, p. 355). Ricardo's discussion of legally convertible bank notes followed Smith, with some of Thornton's modifications. Since their equivalence with gold was guaranteed, they could not be issued in a greater quantity than the value of the coin which would otherwise have circulated. Any attempt to exceed this sum would precipitate a return of notes for specie, a depreciation of both paper and metallic currency, and the subsequent export of superfluous bullion (Ricardo, 1923, pp. 7–13). Overextension of inconvertible notes in a ‘mixed currency’ of notes and coins had the same effect so long as the degree of excess was no greater than the amount of coin in circulation (1923, p. 13, n., pp. 108–12).

In 1809, however, when Ricardo entered the bullion controversy, the currency was composed almost entirely of inconvertible paper. He therefore ascribed the rise in commodity prices, in so far as it corresponded with the premium on bullion, wholly to monetary overissue. Such an overissue would have no other effect than to ‘raise the *money* price of bullion without lowering its *value*, in the same manner, and in the same proportion, as it will raise the prices of other commodities’. In other words, although paper money was depreciated, the ‘bullion price’ of commodities was unaltered. Hence the deterioration of the foreign exchanges ‘will only be a *nominal*, not a *real* fall, and will not occasion the exportation of bullion’ (1923, p. 13 n. and p. 109).

Ricardo was criticized for ignoring the real reasons for the inflation, which had more to do with harvest failures, war subsidies and the Napoleonic blockade (Morgan, 1965, pp. 46–7). Moreover, he left himself open to the charge of superimposing a theory of *fiduciary* money on a *credit* system. Had bank notes been issued at will by the state, Ricardo would have been correct in his characterization of their relationship to the price level. Fiduciary money only *represents* gold in the circulation process, and is depreciated to the extent of its overissue. The depreciation persists until the quantity is reduced, for there are no self-correcting tendencies as in the case of convertible paper. However, the fact that the notes of the Bank Restriction period were not forced currency but credit responding to the demand of the non-bank public was excluded from Ricardo's consideration by Say's Law. He treated the notes as though they were fiduciary because output and velocity were independently given. The possibility of disintermediation when the authorities tried to contract the note issue was also excluded. The fixed velocity assumption implied that the rest of the spectrum of credit would shrink commensurately with the notes. In fact, as the Banking School was to demonstrate, credit instruments simply expanded in their place.

The resumption of specie payments in 1819 on the advice of Ricardo and the bullionist spokesmen did nothing to eliminate price instability from Britain's developing industrial economy. In 1825 and 1836, phases of vigorous expansion ended with an adverse balance of payments, a gold drain from the Bank of England and an inflationary collapse into recession. The Currency School – a new orthodoxy which Morgan describes as the ‘heirs of the Bullion Report’ – attributed the recurrent dislocation to excessive

monetary growth. The convertibility of bank notes was no longer seen as a sufficient safeguard against overissue and consequent depreciation. The Currency School argued that rules would have to be devised to make the paper currency fluctuate as though it were metallic, in other words to replicate the ‘automatic’ operation of Ricardo's international specie-flow mechanism. This implied regulation of the note issue by the monetary authorities in strict conformity with the foreign exchanges; the export and import of bullion was treated as an index of monetary excess or deficiency, and thus of the value of the notes.

The currency principle was given practical effect by the Bank Charter Act of 1844, which set the pattern of the UK financial system for almost a century. It was challenged by the Banking School, which Morgan calls ‘the heirs to the opposition to the Bullion Report, but the opposition as it might have been rather than as it was’.

The long-run determination of aggregate monetary requirements by nominal output – the ‘supply side’ of the equation of exchange – was common ground in the debate. The real point at issue was again the *short-run* behaviour of the variables. Whereas the Currency School adopted Ricardian quantity theory and applied it to a credit system made up of convertible bank notes, the Banking School took the alternative view of metallic circulation and tried to develop a theory specific to credit. Both sides recognized the importance of theorizing the laws of metallic circulation as a precondition for the analysis of paper currency. The entire Currency School case for monetary control rested upon the assertion that the note issue would not by itself emulate the behaviour of a metallic system. Despite legal convertibility, it might depart at least temporarily from the amount and value of the metallic money which would otherwise have circulated. In practice, therefore, economic convertibility could be ensured only by quantitative intervention on the part of the authorities (Torrens to Lord Melbourne, *cit*. Tooke, 1844, p. 7).

Banking School criticism took three main lines. First, starting from the assumption that legal convertibility necessarily implied economic convertibility, they pointed out that any discrepancy between the note issue and a purely metallic system arose from the Currency School's erroneous theory of metallic circulation rather than from the supposed autonomy of the notes. Second, any effect of prices attributed to bank notes could not be denied to a range of financial assets excluded by the Currency School from their definition of money. Third, bank notes were in any case not money but credit, and therefore never could be overissued, through the credit structure as a whole might be extended beyond the limits of real accumulation by ‘speculation and overtrading’.

The Banking School emphasized that the volume of notes in circulation could not be increased at will by the authorities, but only in response to the demand of the non-bank public. This crucial difference between fiduciary money and bank notes was explained by Tooke as consisting, ‘not only in the limit prescribed by their convertibility to the amount of them, but in the *mode of issue*’ (Tooke, 1844, pp. 70–1, emphasis added; see also Fullarton, 1845, ch. 3, and Wilson, 1859, pp. 48, 51–2, 57–8). The currency principle, by contrast, ‘completely identifies *monetary turnover* with *credit*, which is economically wrong’ (Marx, 1973, p. 123). An advance of bank notes did not *add* to the money supply, but merely changed its *composition*, allowing the substitution of one financial asset for another in the hands of the public. Excess notes returned automatically to the bank ‘in the shape of deposits or of a demand for bullion’ (Tooke, 1844, p. 60; see also Wilson, 1859, p. 58; Marx, 1867, III, pt. 5). This was the basis of the law of reflux, which Fullarton called ‘the great regulating principle of the internal currency’. It held

that economic convertibility could be ensured not only by a legal right to exchange notes for specie but also by maintaining a balance between the notes advanced on loan and those returned to the bank at maturity. Provided lending took place on commercial paper which represented a real or (within a given timescale) potential sum of values, ‘the reflux and the issue will, in the long run, always balance each other’ (1845, pp. 64–7; also p. 207; Marx, 1973, p. 131).

The Banking School did not imagine that the economic cycle could be eliminated by monetary measures. Instead, they evolved a new set of criteria by which the authorities could operate on the ‘state of credit’ through interest rate and reserve management (Tooke, 1844, p. 124; Fullarton, 1845, p. 164; Marx, 1867, III, p. 447). In practice, all that lay between the Currency and Banking Schools was ultimately a matter of timing, but this reflected profound theoretical differences. Within the framework of classical analysis, it was the Banking School which came closer to constructing a modern philosophy of monetary regulation.

See Also

- Banking School, Currency School, Free Banking School
- bullionist controversies (empirical evidence)
- Hume, David
- quantity theory of money
- Ricardo, David
- Thornton, Henry

Bibliography

Cantillon, R. 1775. *Essai sur la nature du commerce en général*. London: Macmillan, 1931.

Eatwell, J. 1983. The analytical foundations of monetarism. In *Keynes's Economics and the Theory of Value and Distribution*, ed. J. Eatwell and M. Milgate. London: Duckworth.

Fullarton, J. 1845. *On the Regulation of the Currency*. London: John Murray.

Green, R. 1982. Money, output and inflation in classical economics. *Contributions to Political Economy* 1, 59–85.

Hume, D. 1752. *Writings on Economics*. Oxford: Oxford University Press, 1955.

Law, J. 1705. *Money and Trade Considered*. Edinburgh: Anderson.

Locke, J. 1691. Consequences of the lowering of interest and raising the value of money. In *Principles of Political Economy*, ed. J.R. McCulloch. London: Ward, Lock & Co., 1825.

Marx, K. 1867. *Capital*. Moscow: Progress Publishers, 1971.

Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin.

Montesquieu, C. 1748. *The Spirit of Laws*. London: George Bell & Sons, 1900.

Morgan, E.V. 1965. *The Theory and Practice of Central Banking, 1797–1913*. London: Frank Cass.

Nef, J.U. 1941. Silver production in central Europe: 1450–1618. *Journal of Political Economy* 49, 575–91.

Outhwaite, R.B. 1969. *Inflation in Tudor and Early Stuart England*. London: Macmillan.

Petty, W. 1963. *The Economic Writings of Sir William Petty*. New York: Kelley.

Ricardo, D. 1821. *Principles of Political Economy and Taxation*. Ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1923. *Economic Essays*. London: Frank Cass.

Schumpeter, J. 1954. *A History of Economic Analysis*. London: Allen & Unwin.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Routledge, 1890.

Steuart, J. 1767. *An Inquiry into the Principles of Political Economy*. Edinburgh: Oliver and Boyd, 1966.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. London: LSE reprint series, 1939.

Tooke, T. 1844. *An Inquiry into the Currency Principle*. London: LSE reprint series, 1959.

Vanderlint, J. 1734. *Money Answers All Things*. London: T. Cox.

Vilar, P. 1976. *A History of Gold and Money: 1450–1920*. London: New Left Books.

Viner, J. 1937. *Studies in the Theory of International Trade*. London: Allen & Unwin.

Wilson, J. 1859. *Capital, Currency and Banking*. 2nd edn, London: The Economist.

How to cite this article

Green, Roy. "money, classical theory of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_C000608> doi:10.1057/9780230226203.1127

The New Palgrave Dictionary of Economics Online

money

James Tobin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

silver standard; numeraire; free banking

Article

Money as a social institution and public good

Among the conventions of almost every human society of historical record has been the use of *money*, that is, particular commodities or tokens as measures of value and media of exchange in economic transactions. Somehow the members of a society agree on what will be acceptable tender in making payments and settling debts among themselves. General agreement to the convention, not the particular media agreed upon, is the source of money's immense value to the society. In this respect money is similar to language, standard time, or the convention designating the side of the road for passing.
The reason for the universality of money as a social institution is that it facilitates trade. Trade among individuals enables them to achieve much higher standards of living than if each person or family were restricted to autarchic subsistence. Because of economies of scale, division of labour among specialists yields enormous gains. Of course, trades have always taken place by barter, and even in modern economies many exchanges occur without money. Barter is usually bilateral, thus in Jevons's famous phrase it requires ‘a double coincidence [of wants], which will rarely happen’ (1875: 3). Multilateral trade is much more efficient, permitting each trader bilateral imbalances provided her trade in aggregate is balanced. Imagine, for example, that for lack of double coincidences no bilateral trades are possible among A, B and C because A wants C's goods, B wants A's and C wants B's. Obviously three-way exchange would benefit everyone.
Multilateral barter is conceivable. It could be arranged by putting participants in simultaneous communication with each other – in person as at a village market or a commodity or stock exchange, or by modern telecommunications. But any multi-participant multi-commodity market would need a clearing mechanism. A trader would not have to be balanced with every other trader. But in the absence of a money each trader would have to be balanced in every commodity. This would be awkward and

inefficient. Participants would need to come to market with inventories of many goods. A natural conclusion of any one market session would be intertemporal deals, commodities acquired today in exchange for promised future deliveries of the same or other commodities. Without money, this too would be awkward: a typical trader would end up with debts to or claims on other traders in many specific commodities.

One could imagine using intrinsically valueless tokens during a market session to lubricate barter – like poker chips for scorekeeping in a stakeless poker game. The tokens would make it possible to price each commodity in a common *numéraire* rather than in each of numerous other commodities. But if the tokens became worthless at the end of the session, each participant would have to be required to return as many tokens as he or she started with. Otherwise no one would sell useful goods for tokens, for fear of leaving the market with them rather than with commodities of value. If instead the tokens will be acceptable tenders in this and other markets in future – well, then they are money (on these issues see Hawtrey, 1927, ch. 1; Starr, 1972; Shubik, 1984; Kareken and Wallace, 1980).

The social convention makes a society's money generally acceptable within it, and the practice of general acceptability reinforces the convention. Y accepts money from X in exchange for goods and services and other things of value because Y is confident that Z, A, B,…, and indeed X will in turn accept that same money. Moreover, money is accepted from the bearer immediately and impersonally – without delay, without identification. Since an economic agent's purchases and sales, outlays and receipts, are not perfectly synchronous, each agent's inventory of money fluctuates in size as money circulates throughout the economy. These fluctuations in individual money holdings enable essential intertemporal exchanges to take place. Workers are paid for their labour today, and next week they buy the food and clothing that are the truly desired proceeds of their work. The farmer and the tailor accumulate money from those sales; on payday they pay it out to their hired hands.

The moneys chosen by societies have varied tremendously over human history. So have their languages. In each case, what is universal and important is that something is chosen, not what is chosen. The variety of choice defies generalizations about the intrinsic properties of moneys. Livestock, salt, glass beads and seashells have served as money. Major grain crops were natural media for payments of wages and rents, and therefore in other transactions and accounts. Cigarettes were money in prisoner-of-war camps. On the island of Yap debts were settled by changing the ownership of large immovable stone wheels. The practice continued after the sea flooded their site and the stones were invisible at the bottom of a lagoon. (Similarly when gold was international money in the twentieth-century title to it often changed while the gold itself, safe in underground vaults, never moved.)

Some moneys have been commodities valued independently of their monetary role, intrinsically useful in production or consumption. Others have been tokens of no intrinsic utility and negligible cost of production, coins or pieces of paper. Commodity moneys derive their value partly, and token moneys wholly, from the social convention that designates them as money.

In modern nation-states the sovereign government can generally determine the society's money. For example, the United States constitution assigns to the federal government (thus, not to the states) the power ‘to coin money, regulate the value thereof, and of foreign coin’. The central government defines the monetary unit, decides in what media taxes and other debts to the government itself may be paid, and defines what media are legal tender in the settlement of other debts and contracts (Starr, 1974).

Precious metals as money

Gold and silver have histories going back many centuries as the moneys of choice of many societies and as international media of exchange. Copper coinage antedates them, but copper became too abundant and was relegated to subsidiary coins. The precious metals are durable. They are divisible into convenient denominations. They can be made into ingots, bars and coins of standard weights. When used as moneys, they have been sufficiently scarce – relative to the non-monetary demands for them – as to pack considerable value into convenient portable forms. They glitter. They have long been prized for ornament and display. Gold and silver, one or the other or both, were the basic moneys of Europe and of European dominions and settlements throughout the world from the 17th century, or before, until recently. In modern times gold, in particular, acquired awesome mystique (Keynes, 1930).

Sovereigns minted these precious metals on demand into coins of their own realms, with their own names. In addition to minting *full-bodied* coins for public circulation, sovereigns commonly provided *token* coins made of metals, convenient for retail transactions, negligible in intrinsic value but convertible into the basic money of the realm. Many full-bodied coins circulated across national boundaries with values equivalent to their weight. For example, the original monetary unit of the United States was the silver dollar of Spanish America.

Until the late nineteenth century silver was more prevalent than gold as a monetary commodity. From medieval times silver was the English money of account; the pound sterling was initially a weight of silver. England and many other countries coined both silver and gold, but there were frequent periods when bimetallism degenerated *de facto* into one standard or the other. This happened when their prices at the mint diverged enough from their relative values in other countries or in commerce to offset the costs of arbitrage. Then ‘Gresham's law’ would take over, and the metal undervalued at the mint, the ‘good money’, would disappear from monetary circulation, ‘driven out’ by the ‘bad money’ overpriced at the mint (Hawtrey, 1927: 202–4, 283).

In England in 1717 Isaac Newton, Master of the Mint, unintentionally overvalued gold, pushing silver out of circulation and in effect putting England on a gold standard. The switch was formalized in 1816. During the nineteenth century other European countries and the United States likewise gravitated from bimetallism to gold. Alexander Hamilton, America's first Secretary of the Treasury, complemented the silver dollar with gold coins. But it was not until the late nineteenth century that gold overtook silver as the basic money of the United States. The values of sterling and dollars in gold set by Newton and Hamilton, implying an exchange rate of $4.86 per pound, lasted until 1931, with several wartime interruptions.

The heyday of the international gold standard was 1880–1914, when all major national currencies were convertible into gold at fixed rates. Silver, like copper before it, was eventually demoted to token coin status (Hawtrey, 1927, chs 16–20).

Functions of money

A triad long familiar to students of introductory economics lists the functions of money: (1) unit of account, or *numéraire*, (2) means of payment, or medium of exchange, and (3) store of value.

The US dollar, for example, is the unit of account in the United States. Prices of everything are quoted in dollars, and accounts are kept in dollars. The various media that change hands in transactions – coins,

paper currency, deposits – are denominated in dollars. That does not prevent anyone who cares to do so from quoting prices in a foreign currency or in bushels of wheat, or from finding sellers who will accept them in payment for other things. It just would not be very efficient as a general practice.

To be sure, some societies have used, and kept accounts in, more than one money – in both gold and silver or, for example, in Japan two centuries ago, both in coins and in standard weights of rice. Today some national currencies may be acceptable means of payment in other jurisdictions – dollars in Russia, Israel and Canada, yen in Hawaii, Deutschemarks in Eastern Europe. The reason may be the frequency of cross-border tourism and trade. Or it may be that as a consequence of hyperinflation people turn to a ‘hard’ foreign currency as unit of account. For still a different reason, a new European currency, the ecu, may become a *numéraire* parallel to national currencies like pounds, francs and Deutschemarks during the period of transition to a common currency.

A society's money is necessarily a store of value. Otherwise it could not be an acceptable means of payment. (New York subway tokens cannot be generally acceptable money; they can become valueless any day, even for use as subway fare. US food stamps, intended to be in-kind welfare benefits, are exchanged with cash at par, while grocery brands’ discount coupons are disqualified by their expiration dates.)

Money is the principal means of payment of a society, but it is only one of many stores of value – and quantitatively a minor one at that. Through most of human history land has been the major form of wealth, increasingly augmented by livestock and reproducible capital – buildings, tools, machines and durable goods of all kinds. Claims to much of this wealth today take the form of bonds and shares and other securities. In the United States, basic money is only 6 per cent of total privately owned wealth.

Even though a particular commodity or token is established as the generally acceptable medium for discharging debts denominated in the unit of account, it need not be and generally is not the sole means of payment in use. *Derivative* media, often termed *representative* money, arise and circulate as media of exchange. They are promises to pay the *basic*, sometimes called *definitive*, money on demand. In the commercial city states of northern Italy, merchants left gold with goldsmiths for safekeeping. They then found it convenient to circulate the ‘warehouse’ receipts in place of the gold. Those payable to bearers were precursors of paper currency and banknotes. Those payable to named persons, and on their order to third parties, were precursors of cheques. Indeed, once the goldsmiths realized that they need not keep 100 percent gold reserves against the outstanding claims upon them, and that they could lend their certificates to merchants promising to deliver gold later, they became banks.

Besides providing token coins, states issued paper currency redeemable in gold or silver, or delegated the privilege to a private bank chartered to serve the state, like the Bank of England, founded in 1694. In addition, ordinary private banks issued their own notes, backed only by their own promises to pay basic money, gold or silver. In the nineteenth and twentieth centuries, governments and their central banks came to monopolize the issue of paper currency. This was not a catastrophe for banks. In modern economies, demand deposits in banks, transferable to third parties by cheque or wire or other order, have become the most important derivative media of exchange.

Whether derivative moneys were officially or privately issued, the ability of the issuers to carry out their promises to redeem them in basic money, gold or silver, was a recurrent problem. In wars and other emergencies governments often suspended these promises and issued irredeemable paper money. The trend in the twentieth century was to dispense with commodity money and to replace it with fiat money of no intrinsic value. Within each nation, the official derivative money, government currency, became

the basic money. In 1933 United States paper dollars became inconvertible into gold except by foreign governments or central banks.

Internationally, gold was dethroned in 1971 as the medium for settlement of imbalances of payments between countries. Governments are no longer prepared to buy or sell gold at prices fixed in their own currencies. Gold is traded freely in private markets all over the world. Its price fluctuates as people speculate about its future. In the United States there is still an official weight of gold that theoretically corresponds to the dollar – 0.0231 oz, that is a gold price of $43.22, about one eighth of the free market price. But the US government is not prepared to sell any gold for dollars at the official price – or at the free market price, for that matter.

The US monetary base (M0) is the amount of fiat currency the government, mainly its central bank, the Federal Reserve System, has issued. It is a ‘debt’ to the public on which the government pays no interest and against which the government holds virtually no assets (other than its remaining gold stock, $11 billion at the official price, and its drawing rights at the International Monetary Fund, $19 billion). Derivative promises to pay dollars are now, directly or indirectly, commitments to pay this fiat money. Those promises include bank deposits and all other debts, private and public, denominated in dollars and payable at specified future times, tomorrow or 30 years hence.

In the United States in the fourth quarter of 1991 the stock of *transactions money* (M1) held by economic agents other than the federal government and banks averaged $890 billion, $265 of currency (paper and coin) and $617 of chequable deposits available on demand. The banks held reserves of $53 billion in currency in their vaults or on deposit in the 12 Federal Reserve Banks, collectively the American central bank. The sum of the currency in public circulation and the currency or equivalent held as bank reserves is the *monetary base* (M0), $318 billion. It is often called *high-powered* money: every dollar of M0 was supporting $2.80 of Ml, and GNP transactions of $18.20 a year.

Sovereigns have long profited from their money monopolies. Their mints charged ‘seigniorage’ fees – and sometimes they cheated. Likewise, issue of currency bearing zero interest is a way for a government to pay its bills, easier than taxation and cheaper than interest-bearing debt. By regularly issuing base money to keep up with economic growth and inflation, the sovereign collects seigniorage year after year. In the United States today seigniorage is a minor source of revenue. Since base money is only 6 per cent of GNP, growth of dollar GNP at 7 per cent a year means new issue of base money of only 0.42 per cent of GNP, 1.68 per cent of the federal budget. But for many less developed countries printing money is a major way of financing public expenditures; seigniorage is a major source of revenue, because implicit taxation by inflation is politically easier than explicit taxation.

Commodity money vs fiat money

The age of fiat money, first in one nation after another and finally internationally as well, has been more inflationary than the century of silver and gold standards between the Napoleonic wars and the First World War. During and following the 1914–18 war the gold standard broke down, and attempts to re-establish it during the Great Depression did not succeed. The Bretton Woods regime established in 1945 linked the world's currencies to gold via their fixed parities with the US dollar, because foreign governments could convert dollars into gold at a fixed price. But this system differed radically from the pre-1914 gold standard in that currency exchange rates could be and were frequently changed. The discipline imposed on a government and economy by an exchange parity fixed for a long time was

diluted. In 1971, when this discipline became too much for the US itself, the gold–dollar parity gave way, and the international monetary system was wholly a regime of fiat money.

Discontent with inflation since the Second World War, and with the volatility of currency exchange rates since 1971, has led to agitation for return to the gold standard or some other commodity money. A commodity standard, if adhered to, provides a real anchor for nominal prices; its discipline prevents hyperinflation.

However, although the long-run trend of prices during the gold standard period was flat, there were violent inflationary and deflationary fluctuations around it. More important, real economic activity was highly volatile, to a degree that would be politically unacceptable nowadays (Cooper, 1982, 1991).

Irving Fisher, writing during the gold standard era, was greatly concerned by the instability of prices. He was complaining, in effect, about the volatility of the relative price of gold. Ideally, he would define the dollar in terms of a representative package of goods and services, the bundle priced in a comprehensive index number. Thus he revived the idea of a ‘tabular standard’, proposed by several early-nineteenth-century writers, and described with approval by Jevons (1875, ch. 25). But exchange between paper currency and such bundles is impractical. Fisher proposed instead to make periodic adjustments of the gold content of the dollar, raising or lowering it in proportion to the rise or fall in the price index since the previous adjustment. In effect, the Treasury would be selling gold for dollars to fight inflation and buying gold for dollars to fight deflation (Fisher, 1920).

A recent proposal by Robert Hall (1982) would tie the dollar to a composite commodity ‘ANCAP’ of ammonium nitrate, copper, aluminium and plywood. Because ANCAP's prices have historically mirrored general indices, it is meant to be a feasible proxy for the economy's aggregate market basket (other proposals for commodity standards are described in Cooper, 1991).

The Fisher strategy could be followed, even imposed as a nondiscretionary rule on the central bank, in a regime of fiat money. The market operations to implement it would be carried out in securities rather than in gold. The fundamental issue is not the monetary standard but whether stabilizing a price index should be the exclusive objective of monetary policy, to the exclusion of stabilization of real output growth and employment.

Free market money?

Would it be possible to privatize money? Certainly it is possible to privatize derivative issues of money, promises to pay fixed amounts of base money on demand. But United States experience suggests that the supply of money, even derivative ‘low-powered’ money, cannot safely be left to free market competition.

Before the establishment of the national banking system in 1864, private banknotes were the only paper currency of the United States. The several states freely chartered banks, and those banks freely issued their own banknotes. These were promises to pay silver dollars, but so-called ‘wildcat’ banks contrived to make it tough for noteholders to find them. There was no central bank to control the aggregate issue of banknotes. The notes circulated at varying discounts from par and often became worthless, stranding innocent holders.

As a result, Congress established a system of nationally chartered banks in 1864, and taxed state banknotes out of existence. Only nationally chartered banks could issue notes, and these had to be fully backed by US Treasury debt securities. In effect, they were Treasury currency, supplementing various

direct issues of Treasury currency (including the inconvertible ‘greenbacks’ the union government issued during the 1861–5 Civil War, which were made convertible into specie in 1879). Central banking did not begin in the United States until the Federal Reserve Act of 1914, which confined the issue of banknotes to Federal Reserve Banks.

Although private banks, state and national, were out of the business of issuing demand notes, they were still in the business of accepting demand deposits, the increasingly prevalent form of derivative money. Banks’ balance sheets were regulated, but depositors were at risk. Their banks might not be able to pay in gold or equivalent on demand. After the epidemic bank failures of the 1920s and 1930s, Congress initiated a system of federal deposit insurance. Deposits in banks and other financial institutions became governmentally guaranteed, like banknotes after 1864. In the 1980s, these deposit guarantees became an expensive burden on federal taxpayers.

Could government get out of the money business altogether? It seems barely possible with commodity money and not possible with fiat money. If the government defined the *dollar* as a certain weight of gold or ANCAP or some other commodity or bundle, then private entrepreneurs could issue ‘dollars’, either chequable deposits or paper notes. They would be promises to pay the bearer the equivalent in the chosen commodities. The commodities themselves would not necessarily circulate on their own; indeed ANCAP and other composites could not.

The money entrepreneurs would have to keep inventories of the commodity as reserves. If one hundred per cent reserves were required, the currency would be like goldsmiths’ warehouse receipts, and the private issuers would earn just a small fee for ‘minting’ the commodity into paper. Left to themselves, they would become banks, acquiring risky and illiquid assets while incurring demand liabilities. *Caveat emptor* would reign. The rates various banks would have to pay to attract funds would reflect depositors’ appraisals of the risks. Notes and cheques of risky banks would not be honoured at par. In short, the very problems that resulted in consensus that issue of money cannot safely be left to unregulated free markets would recur.

Could the government's role be confined to defining the unit of account, the commodity equivalent of a dollar, in the same way that the government – through the Bureau of Standards in the United States – defines weights and measures? Could the system operate without any government-owned or government-issued base money? In its absence, clearings among private banks would require awkward transfers of ownership of the commodities kept as reserves against their liabilities. Very probably some one bank or consortium would arise as an unofficial central bank, and its liabilities would play the role of base money, the medium in which clearing imbalances among other banks are settled. The central bank, official or unofficial, would have to hold inventories of the standard commodity, gold or ANCAP or whatever, and be prepared to convert currency into the commodity and vice versa. That institution, history also suggests, would eventually be nationalized.

A fortiori, if there is neither an official definition of the ‘dollar’ nor any issue of dollars by the government or a quasi-governmental institution, there would be no standard commodity for private banks to compete in supplying to the public. Barter trading would be the rule, and the public-good advantages of social agreement on money would be lost. Since the institution of money is a public good, it is not surprising that its advantages cannot be realized by private market competition unassisted and uncontrolled.

How can money have positive value in exchange?

Economists have long regarded the theory of value as the central question of their discipline. What determines the prices at which goods and services are traded for each other? The prices in question include the wages of labour in terms of consumer goods, the rent of land in terms of its produce, and many other relative prices. They encompass interest rates and asset prices, thus the terms of trade of commodities to be delivered in future for commodities available today. They cover interregional and international trade, where the prices of concern are the terms on which imports can be obtained by exports.

Money, however, is an embarrassment to value theory. According to standard theory, something can have positive value only if it generates positive marginal utility in individuals' consumption or positive marginal productivity in the making of goods and services that do generate marginal utility. The embarrassing puzzle is sharpest for fiat money. All of its value comes from the fiat that makes it money. Fiat money has no intrinsic non-monetary source of value. It cannot be eaten or worn or be used in any other way that generates utility for consumers, except a few numismatists. Nor can it contribute to the production of things that consumers do value. It can be produced at zero social cost. Yet it is a scarce commodity for any individual agent. Why is it worth anything at all? That the institution of money is of value to the society as a whole as a public good does not automatically give it value to individuals in market exchanges.

The uphill struggle of modern economic theorists to cope with these challenges is exhibited in the proceedings of a recent conference (Kareken and Wallace, 1980). Their solutions relied principally on the overlapping generations model, which unrealistically assigns to money the function of being the sole or the principal store of value that links one generation to the next. The most careful, thoughtful and perceptive formal models of the roles of credit and money in transactions and strategies, in partial equilibrium and general equilibrium systems, are those of the game theorist Martin Shubik (1984).

It was argued at the beginning that a condition for fiat money to be held and valued today is that it will be acceptable in exchange for intrinsically useful commodities tomorrow. But this bootstrap story may not work. Suppose the world itself is known to be finite; its end will come at a definite future time. In the last period, one minute before midnight so to speak, you may need money to buy whatever consumer goods might generate utility, at least solace. Otherwise you will be confined to your own resources. But who will sell you anything, knowing that the money will be worthless while the goods might be a source of some utility? Thus money is worthless one minute before midnight, and by iterations of the same argument, it is worthless today. Even if the institution of money had public-good value between now and the end of the world, the money itself would have no market value to individuals.

The escape from this logical impasse is that we do not all and will not all expect with certainty the end of the world at any definite time. We always do, always will, assign some probability to its continuation. Since there are many other paradoxes involved in thinking about human behaviour in a world with no chance of a future beyond a definite time, it is best not to take that prospect seriously in economic modelling.

Formal general equilibrium theory, which describes the imaginary world of frictionless barter, does of course express the prices of goods and services in a *numéraire*. It is tempting to identify *numéraire* prices as money prices. But the *numéraire* is just a mathematical normalization convenient for handling the fact that the supply equals-demand equations for N commodities determine only the $N - 1$ relative prices. Those relative prices are, by construction, independent of the scalar arbitrarily attached to the

numéraire.

Standard value theory does, of course, have something to say about the value of commodity money in terms of other goods and services. In a gold standard regime, the relative prices of gold in other commodities have to be the same at the mint and in the market; they cannot depend on whether the gold is circulating in coins or being used in jewellery, dentistry or rocketry. That is simply a condition of the absence of arbitrage profits. It definitely does not say that under the gold standard the relative price of gold is the same as it would be if gold were not money. As argued above, gold's role as money must increase the demand for it, and that must affect its price unless it is supplied perfectly elastically. The same will be true of any other commodity or bundle of commodities chosen as the monetary standard. A substantial part of the value of any commodity used as money arises from the convention or the fiat that makes it money. The distinction between commodity money and fiat money is not absolute.

The neutrality of money

Although business managers, financiers, politicians and workers worry a great deal about monetary institutions and policies and their consequences for economic activity and well-being, pure economic theory minimizes these consequences. Theory puts the burden of proof on anyone who contends that money and monetary inflations or deflations do much good or much ill.

Classical economists liked to insist that money is a veil, obscuring but not altering the real economic scenario (Robertson, [1922] 1959:7). Their modern descendants expound ‘real business cycle theory’, premised on the view that economic developments that matter to societies and individuals are independent of monetary events and policies (Prescott, 1986). It is true that economic fluctuations and trends are frequently misinterpreted by stressing superficial monetary phenomena to the neglect of resources, technologies and tastes. But money does matter, really.

Does an economy arrive at the same *real* outcomes (in variables like volumes of production, consumption and employment, and in relative prices such as the purchasing power of wages and the price of oil relative to that of bread) as it would without the institution of money? Clearly not. Without money, confined to barter, the economy would produce a different menu of products, less of most things. People would spend more time searching for trades and less in actual production, consumption and leisure.

That is not the comparison the classical economists, old and new, intend by the ‘veil’ metaphor. Their fantasy is a frictionless, costless system of multilateral barter, in which relative prices and the allocations of labour and capital among various productive activities are determined in competitive markets. Their proposition is that the outcomes of an economy with money are the same as those that would arise from their ideal barter model. The corollary is that real economic outcomes are independent of the particular nature of the monetary institutions (Dillard, 1988).

These propositions cannot be true of commodity money. Real economic outcomes with commodity money will differ from those with fiat money, and will also depend on what commodity is selected as money. Inventories of the chosen commodity have to be held for exchange purposes and for governmental and bank reserves, beyond the stocks held in connection with the commodity's non-monetary uses in production and consumption. In growing economies demands for monetary inventories will be steadily increasing. The relative demands for monetary and non-monetary inventories are bound

to change with economic and technological developments that alter the incentives to produce the commodity and change its prices in terms of other goods and services. Examples are discoveries or exhaustions of gold and silver deposits and innovations in mining and processing technologies. Since the monetary commodity's price is fixed in money, its output will decline when there is general inflation and rise when there is deflation. Intertemporal choices involving the monetary commodity, as well as contemporaneous choices, will be significantly affected by its monetary use.

The availability of moneys, whether commodity or fiat, whether basic or derivative, as stores of value necessarily brings about significant deviations in real outcomes from the hypothetical regime of frictionless barter. This is true even though that regime is postulated to include markets in state-contingent commodity futures, ‘Arrow–Debreu’ contracts (Arrow and Debreu, 1954). Holding monetary assets gives agents more flexibility: they can convert them into consumption of any kind at any time in any ‘state of nature’, though not at predictable prices. The flexibility is a convenience to individual agents. But, as Keynes saw, it opens the door to ‘coordination failures’ which are the essence of macroeconomics – demand for goods and services may at times diverge seriously from supplies (Keynes, 1936, chs 16, 17).

The classical dichotomy

It is possible to recognize that an economy with monetary institutions is different in real outcomes from a barter economy, even from an ideal frictionless barter economy, and still to argue that its real outcomes are independent of the purely nominal parameters of those institutions. It would be terribly convenient if the determination of the absolute price level, the reciprocal of the value of the monetary unit in a representative bundle of consumer goods, could be split off from the determination of relative prices and the associated real quantities.

Don Patinkin (1956) called this separation the *classical dichotomy*. Only monetary shocks would affect the general price level, and those shocks would raise or lower the nominal prices of all commodities in the same proportions. Only real shocks – to tastes, technologies and resource supplies – would affect relative prices and real quantities. This proposition would not exclude the fact that the monetary institutions themselves matter. The choice between commodity money and fiat money, the choice among possible commodity standards, and the arrangements for derivative moneys might well affect the social efficiency of markets and trade.

What are the nominal parameters whose settings, according to the classical dichotomy, would make no real difference? For a commodity money, such a parameter is the definition of the monetary unit in terms of the standard commodity, for example the weight in gold of a dollar. For fiat money, the key nominal parameter is the quantity of money – base money, all transactions money, or some even more inclusive aggregate.

Why should cutting the gold content of the dollar from 0.0484 ounces to 0.0286 ounces, raising the dollar price of gold from $20.67 to $35.00 (as Franklin Roosevelt did in 1933), make any real difference? The dollar values of existing public and private stocks of gold, and of monetary claims to gold would rise in the same proportion. Will not all other commodity prices do likewise? Then all relative prices and real quantities, including those of gold, will be the same as before.

For fiat money systems, and for commodity standards where issues of derivative moneys have become

essentially independent of the commodity, the *quantity theory of money* achieved similar dichotomization. According to the theory, which might more accurately be called the quantity-of-money theory of prices, an increase in the nominal quantity of money would raise all nominal commodity prices in the same proportion, leaving relative prices and real quantities unchanged. Quantity theorists argue that an increase in the quantity of money is equivalent to a change in the monetary unit. A hundred-fold increase in the stock of French francs would be – would it not? – the same as De Gaulle's decree changing the unit of account to a new franc equivalent to 100 old francs. Since the units change could make no real difference, the other way of multiplying the money stock could not either.

These analogies fail, for several related reasons. In most economies money is by no means the only asset denominated in the monetary unit. There are many promises to pay base money on demand or at specified dates. If there is a thorough units change, like De Gaulle's, all these assets are automatically converted to the new unit of account. Roosevelt's devaluation of the dollar relative to gold was not a pure units change. He did not scale up the dollar values of outstanding currency or even of Treasury bonds with provisions for such revaluation. Naturally private assets and debts expressed in dollars were not scaled up either. Likewise, when the quantity of money is changed by normal operations of governments or central banks or by other events, the outstanding amounts of other nominally denominated assets are not scaled up or down in the same proportion. They may remain constant, as when money is printed to finance government expenditures. They may move in the opposite direction, as when central banks engage in open-market operations, which typically increase the amount of base money outstanding by buying bills or bonds, thus reducing the quantities of them in the hands of the public.

The quantity theory

The quantity theory goes back to David Hume, probably farther, but its major and most effective protagonists have been Irving Fisher (1911) and Milton Friedman (1956).

In its crudest form, the quantity theory is a mechanistic proposition strangely alien to the assumptions of rational maximizing behaviour on which classical and neoclassical economic theories generally rely, as J. R. Hicks eloquently pointed out in a famous article (1935). Specifically, it ignores the effects of the returns to holding money on the amounts economic agents choose to hold. The technology of monetary circulation fixes the annual turnover of a unit of money. Suppose that every dollar ‘sitting’ supports just *V* dollars per year ‘on the wing’, to use D.H. Robertson's famous terms ([1922] 1959: 30). Suppose, further, that the economy is assumed to be in real equilibrium and the supply of money is doubled. The public will not wish to hold the additional money until the dollar value of transactions is doubled, and this requires prices to double.

Surely the demand for money to hold is not so mechanical. The velocity of money can be speeded up if people put up with more inconvenience and risk more illiquidity in managing their transactions. Money holdings depend, therefore, on the opportunity costs, the expected changes in the value of money and the real yields of other assets into which the same funds could be placed. Fisher and Friedman would agree.

The quantity theory can still be rationalized, as a proposition in comparative statics. Compare, for example, two stationary situations of a given economy, in each of which the money supply and price level are constant over time. Let the money supply in the second situation be twice that in the first. Then

an equilibrium in the second situation will be the equilibrium of the first with a nominal price level twice as high. This will be true even if the demand for money is modelled as behavioural, not mechanical, and is allowed to depend on interest rates, expected inflation and other variables.

However, it is not sufficient to double solely the quantity of money, narrowly defined. All exogenous nominal quantities, including outstanding stocks of debts and assets, must also be doubled. Or the second equilibrium must be interpreted as a stationary state that will be reached only when all these other nominal stocks have had time to adjust endogenously to the new quantity of money. This quantity theory does not apply to short-run changes in monetary quantities engineered by central banks, for the same reasons that render the ‘units change’ metaphor inapplicable.

In its interpretation as a proposition in long-run comparative statics, the quantity theory supports ‘neutrality’ as asserted in the classical dichotomy. Neutrality has come to have two meanings in monetary economics. Simple *neutrality* means that real economic outcomes are independent of the levels of nominal prices. *Superneutrality* means that those outcomes are also independent of the rates of change of nominal prices.

The case for superneutrality appeals to, and depends upon, the ‘Fisher equation’. Early on, Fisher (1896) saw the importance of distinguishing between nominal and real rates of interest on assets and debts denominated in monetary units. *Ex post*, the algebraic difference between them is by definition the rate of inflation or deflation. This is a tautology. But Fisher (1911) is also credited with a meaningful proposition: anticipation of inflation (deflation) raises (lowers) nominal rates of interest but does not alter real rates of interest. The corollary is that whatever is the time path of money stocks that determines the path of prices, the paths of real economic variables are the same. Fisher himself was enough of a classical economist to believe this as a long-run theoretical truth, but enough of a pragmatic empiricist to find that nominal rates were very slow to incorporate adjustments for ongoing inflations and deflations.

The price of money

A 1975 conference on monetarism at Brown University is remembered for a pithy observation by Milton Friedman, offered only half in jest:

> For the monetarist/non-monetarist dichotomy, I suspect that the simplest litmus test would be the conditioned reflex to the question, ‘What is the price of money?’ The monetarist will answer, ‘The inverse of the price level’; the non-monetarist (Keynesian or central banker) will answer, ‘the interest rate’. The key difference is whether the stress is on money viewed as an asset with special characteristics, or on credit and credit markets, which leads to the analysis of monetary policy and monetary change operating through organized ‘money’, i.e. ‘credit’, markets, rather than through actual and desired cash balances. Though not so obvious, the answer given also affects attitudes toward prices: whether their adjustment is regarded as an integral part of the economic process analyzed, or as an institutional datum to which the rest of the system will adjust (Stein, 1976: 316).

‘What am I’, asked the chairman of the session, George Borts, ‘if I answer “one”?’

Any durable good has at least two ‘prices’, the price at which it can be bought or sold, and the price of

the services it renders per unit time. The price of the good itself is the present value of the expected, though uncertain, values of the services it will render in future. For money, the first price is its purchasing power. Its services come in two forms: as a store of value, the capital gain or loss from changes in its purchasing power, and, as a medium of exchange, the benefits it yields in convenience, effort-saving and risk reduction. Without cash on hand, an economic agent may find it costly to make desirable transactions, or to forgo them. The marginal productivity of holding money is the value of an additional dollar in reducing those costs.

What is the marginal opportunity cost to which agents will equate the marginal productivity of holding money? It depends on what alternatives are available. If money proper were the only store of value in the economy, the opportunity cost of holding money would be the marginal utility of immediate consumption relative to future consumption. Although this set-up is all too common in the literature, it confuses theories of money and of saving. Acknowledging the availability of other stores of value makes the cost of holding money the difference between the real capital gain or loss on money and the real rate of return on the non-money assets in which a marginal dollar could be invested.

If money proper were the only store of value in the monetary unit of account, though not the only one in the economy at large, the relevant opportunity cost would be the return on real capital, that is storable or durable commodities. In modern economies, however, the immediate substitutes for money are promises to pay money in future. Since money and these substitutes are affected equally by price level changes, the opportunity cost is simply the nominal interest rate on those non-money substitutes. (This assumes zero nominal interest on money itself.)

Friedman's Keynesian is careless if he calls any of these opportunity cost concepts the price of money. These are prices of the services of money. Friedman's monetarist is right, therefore, to say that the price of money is the reciprocal of the commodity price level – the real price, that is, for Borts was right about money's nominal price. Of course, there are as many relative prices as there are non-monetary commodities, and any average value of money requires using an arbitrary commodity price index.

To implement Friedman's asset valuation approach to the price of money, suppose that the nominal supply of money per capita, real per capita output and the real interest rate all follow arbitrary variable paths, anticipated in advance. Assume, at least for illustrative purposes, the Allais–Baumol–Tobin model of the demand for money (Baumol and Tobin, 1989). The marginal productivity of nominal cash holdings for a representative agent is the reduction in the frequency and cost of exchanges back and forth between money and dollar-denominated interest-bearing substitutes. It is, by the usual approximation equal to $a(t)y(t)/(2 \cdot m(t)^2 v(t))$, where a is the real cost of one of those exchanges, y is the agent's real income per period, m is the agent's average nominal cash holding, and v is the value of money, the reciprocal of the price level. Of these, a, y and m are arbitrary exogenous functions of time, while the valuation v is a function of time to be determined. Let $r(t)$ be the exogenous path of the real interest rate. The value of money at any time T is the discounted value of its future marginal productivities:

$$v(T) = a(T)\int_T^{\infty} \exp\left(-\int_T^t r(s)\,ds\right) y(t) \,/\, (2m(t)^2 v(t))\,dt, \tag{1}$$

$$v'(T) = r(T)v(T) - a(T)v(T) / (2m(T)^2 v(T)), \tag{2}$$

$$r(T) - v'(T) / v(T) = a(t)v(T) / (2m(T)^2 v(T)^2). \tag{3}$$

Equation (3), with the nominal interest rate on the left, is the familiar equation for optimal *real* cash holdings. It involves the stronger Fisher equation, because the real rate has been taken as exogenous. Interpreted as the price dynamics of the economy, these equations describe the time path of the ‘price of money’. The level of prices at each time converts the autonomous nominal money supply into the real quantity on which its marginal productivity depends. The price path itself generates the rates of price change which, added to the autonomous real interest rates, give the nominal rates. The marginal productivity of money at each point in time is equated to the nominal interest rate. Future as well as current values of money supplies, as well as other variables, affect current prices. An expected increase in future money supply raises prices today, and so does an expected future increase in real rates of interest. The Fisher equation is essential to maintain the assumed dichotomy between the paths of real and nominal variables (for a calculation in this same spirit, see Sargent and Wallace 1981).

Money and macroeconomics

In the above scenario, a key institutional fact is that the nominal interest rate on money proper is fixed, at zero. Expected inflation makes money's real interest rate negative and reduces the attraction of holding money compared to assets bearing the economy's real interest rate. For the same reason, an increase in that real interest rate is a disincentive to hold money.
However, the same institution – the fixed nominal interest rate on money – threatens the classical dichotomy. It calls into question the Fisher equation, which is central to the independence from monetary influence of the real rate of interest and related real variables. It calls it into question in principle, in long runs and short, in equilibrium and in disequilibrium. If expected inflation diminishes demand for money, it by the same token increases demands for other assets, both interest-bearing promises to pay money and real capital. These substitutions will reduce the real interest rates on those assets; their nominal interest rates will rise less than the full inflation premium. This effect – associated

in the literature with the names of Mundell (1963) and Tobin (1965, 1969) – refutes superneutrality, which is essential to neutrality in any general dynamic meaning. That is to say, it is not possible to determine the real interest rate and related real variables independently of the money equation, or to determine the value of money from the demand=supply equation for money by itself.

This is true whether the economy is assumed to be classical, with full employment assured by flexibility of nominal interest rates and prices, or Keynesian, with aggregate demand short of full employment. However, the real effects of expected price inflation and deflation are a reason for doubting the efficacy of price flexibility in sustaining or restoring full employment equilibrium in the face of aggregate demand shocks (Fisher, 1933; Keynes, 1936, ch. 19; Tobin, 1975).

Irving Fisher, Alfred Marshall and other monetary economists of the early twentieth century regarded neutrality in any sense as properties of long-run static equilibrium, not of the dynamic transitions that dominate empirical observations of monetary and real variables. According to them, people are slow in translating experience of inflation into their expectations of the future. This is how Fisher interpreted the strong positive correlations he found between inflation rates and real output (Fisher, 1911). However, the Mundell–Tobin effect suggests a still stronger conclusion, since it calls into question the Fisher equation even when inflation expectations are correct and people are not victims of ‘money illusion’.

In Friedman's litmus test there is much more at stake than meets the eye. The issue is how the price level, whose reciprocal is the ‘price of money’, is determined. The monetarist's trained instinct is to think of it as determined by the demand=supply equation for money ‘as an asset with special characteristics’. With the absolute price level thus determined, the function of markets for goods and services is to generate real, relative prices, just as in Walrasian general equilibrium theory. Those real variables, in turn, are exogenous to the path of the ‘price of money’.

The Keynesian's trained instinct, on the other hand, is to think of the price level as an index of nominal prices of goods and services. As Keynes (1936, Book I) emphasized – for labour markets especially – markets in our monetary economies determine in the first instance nominal prices, not real prices. The pricel ‘level’ is a synthetic aggregate of multitudes of individual prices determined in diverse imperfect markets, often decided by administrative decisions or by negotiations. For price determination the most relevant equations of a macroeconomic model are price and wage equations, often members of the Phillips curve family. These specify inertia of varying degrees in nominal prices and relate their changes to measures of real excess demand or supply. As a result, price indices move smoothly and sluggishly over time, not ‘jumping’ like the price of a financial asset sensitive to market views of the future.

With the price level determined in goods markets, the function of the money demand=supply equation is to generate interest rates. That explains the Keynesian's instinctive response to the test question. Of course, the Keynesian recognizes that the endogenous variables of a simultaneous equations system are determined jointly, not equation by equation. That real variables are among those endogenous variables can be attributed to the fact that there is usually a non-zero discrepancy between the price path determined by the full system and the path that would be generated by the monetarist's asset price of money. The non-monetarist view does not take prices ‘as an institutional datum to which the rest of the system will adjust’, but it does rely on variables besides prices to equate ‘actual and desired cash balances’.

The equation of money demand and supply is just one of many relations in a theoretical or econometric macroeconomic model. The small tail cannot wag the big dog. That was too much to expect. The price level is a factor common to the valuation of many assets denominated in the monetary unit, many of

them close substitutes for transactions money. Their quantities now and in future must make a difference. Of course monetary policies and supplies, current and prospective, are important determinants of the price level, and so are credit markets. But the channels of these influences run through demands and supplies in markets for goods and services. Understanding the process belongs to the messy subject of macroeconomics. Finance theory, however elegant, cannot provide a shortcut. Monetary events and policies are not a sideshow to the main performance. The real variables of a monetary economy are hopelessly entangled with monetary phenomena. They do not behave as if an economy enjoying the societal advantages of money were a frictionless multilateral barter economy seen through a veil. That barter economy would never have business cycles characterized by economy-wide excess demands and supplies of labour and other goods and services. The public-good advantages of the institution of money do not come so cheap. Among their costs are fluctuations in business activity and in the value of money itself. Pragmatic monetary economics is a central part of macroeconomics in general.

See Also

- fiat money
- financial intermediation
- gold standard
- international finance
- monetarism
- new classical macroeconomics
- quantity theory of money
- rational expectations

Bibliography

Arrow, K.J. and Debreu, G. 1954. Existence of equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Baumol, W.J. and Tobin, J. 1989. The optimal cash balance proposition: Maurice Allais's priority. *Journal of Economic Literature* 27 , 1160–62.

Cooper, R.N. 1982. The gold standard: historical facts and future prospects. *Brookings Papers on Economic Activity* 1982(1), 1–45.

Cooper, R.N. 1991. Toward an international commodity standard? In *Money, Macroeconomics, and Economic Policy*. ed. W.C. Brainard, W.D. Nordhaus and H.W. Watts. Cambridge, MA: MIT Press.

Dillard, D. 1988. The barter illusion in classical and neoclassical economics. *Eastern Economic Journal* 14 , 299–318.

Fisher, I. 1896. *Appreciation and Interest*. Publications of the American Economic Association, 3rd series 11(4); reprinted, Fairfield, NJ: A.M. Kelley, 1991.

Fisher, I. 1906. *The Nature of Capital and Income*. New York: Macmillan.

Fisher, I. 1911. *The Purchasing Power of Money*. New York: Macmillan.

Fisher, I. 1920. *Stabilizing the Dollar*. New York: Macmillan.

Fisher, I. 1933. The debt-deflation theory of great depressions. *Econometrica* 1, 337–57.

Friedman, M. 1956. *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Hall, R.E. 1982. Explorations in the gold standard and related policies for stabilizing the dollar. In *Inflation: Causes and Effects*, ed. R.E. Hall. Chicago: University of Chicago Press.

Hawtrey, R.G. 1927. *Currency and Credit*. 3rd edn. London: Longmans, Green & Co.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica*, NS 2(1), 1–19.

Jevons, W.S. 1875. *Money and the Mechanism of Exchange*. London: King.

Kareken, J.H. and Wallace, N., eds. 1980. *Models of Monetary Economics*. Minneapolis, MN: Federal Reserve Bank.

Keynes, J.M. 1930. Auri sacra fames. In *Essays in Persuasion*, reprinted in *The Collected Writings of John Maynard Keynes*, vol. 9, London: Macmillan, 1972; New York: Harcourt Brace.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. Reprinted in *The Collected Writings of John Maynard Keynes*, vol. 7. London: Macmillan, 1973; New York: Harcourt Brace.

Mundell, R.A. 1963. Inflation and real interest. *Journal of Political Economy* 71, 280–83.

Patinkin, D. 1956. *Money, Interest, and Prices*. New York: Harper and Row; 2nd edn, 1965.

Prescott, E. 1986. Theory ahead of business cycle measurement. *Federal Reserve Bank of Minneapolis Quarterly Review* 10(4), 9–22.

Robertson, D.H. 1922. *Money*. Cambridge Economic Handbook, 4th edn, Chicago: University of Chicago Press, 1959.

Sargent, T.J. and Wallace, N. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5(3), 1–17.

Shubik, M. 1984. *A Game-Theoretic Approach to Political Economy*. Cambridge, MA: MIT Press.

Starr, R.M. 1972. The structure of exchange in barter and monetary economies. *Quarterly Journal of Economics* 86, 290–302.

Starr, R.M. 1974. The price of money in a pure exchange economy with taxation. *Econometrica* 42, 45–54.

Stein, J.L., ed. 1976. *Monetarism*. Amsterdam: North-Holland.

Tobin, J. 1965. Money and economic growth. *Econometrica* 33, 671–84.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1(1), 15–29.

Tobin, J. 1975. Keynesian models of recession and depression. *American Economic Review* 65, 195–202.

How to cite this article

Tobin, James. "money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000217> doi:10.1057/9780230226203.1126

The New Palgrave Dictionary of Economics Online

moneylenders in developing countries

Hanan G. Jacoby

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Moneylenders are a principal source of credit in developing countries. They thrive where collateral is scarce and legal enforcement of debt contracts is difficult. Their advantages over banks include better knowledge about creditworthiness of their clientele and greater ability to enforce repayment. Landlords lend to their share-tenants because they can capture a larger share of the tenants' surplus than can outside lenders. Other credit by moneylenders is in kind, such as in the form of input advances or deferred rent. The effects of government credit policies depend importantly on the relationship between moneylenders and banks.

Keywords

agricultural finance; asymmetric information; credit markets in developing countries; informal sector; moneylenders in developing countries

Article

Moneylenders are a principal source of credit in developing countries, especially in rural areas, but are notoriously difficult to classify. They may be shopkeepers, millers, traders, landlords, or professional financiers. Moneylenders operate within a broad spectrum of lending 'formality' bounded above by the activities of commercial or agricultural banks and below by credit from friends, relatives, and fellow clan members. Banks normally take deposits, ask lenders for collateral, have formal procedures for loan applications with written contracts, and operate within the legal system; moneylenders may do none of the above. Friends, relatives and clan-members, on the other hand, do not require their loans to be secured, make verbal agreements, generally do not charge interest, and often allow state-contingent repayment (Udry, 1994). Reciprocity and social pressure, rather than legal sanctions, enforce such kin- or clan-based credit (La Ferrara, 2003). Moneylenders, by contrast, are less flexible about the terms of repayment, more likely to charge interest, and less able to mobilize social opprobrium to punish default.

Formal sector lending is limited by the value of collateral, which in agricultural areas is usually in the form of land. Land is useful as collateral only to the extent that it can be legally repossessed upon default of the loan. This, in turn, requires that land be titled, or that ownership be otherwise documented, and that foreclosure be enforceable in court. Moneylenders thrive in settings where collateral is scarce or legal enforcement of debt contracts is weak or non-existent. But such conditions are not sufficient for the presence of moneylenders, who ultimately face the same problem as do banks; earning profit in the face of potential default. One way to do so is by setting a low interest rate and rationing credit, as in Stiglitz and Weiss (1981). This presumes, however, that moneylenders have no particular informational advantage over banks.

What, then, is the comparative advantage of the moneylender? There are three, not mutually exclusive, answers to this question, all related to the fact that the moneylender either resides in the same village or locality as his clientele, and is thus likely to have much more personal knowledge of and contact with them than would a bank, or is simultaneously dealing with his borrowers in another market. By virtue of proximity, a moneylender may, first of all, have a better idea as to whether a borrower can successfully implement a given project and thus repay the loan. In other words, it is mainly the bank, not the moneylender, that faces asymmetric information about the creditworthiness of the borrower.

A second advantage the moneylender may have over a bank is in enforcing repayment. Traders or millers often advance credit against the forthcoming harvest. By acquiring the right to market his debtor's harvest as a condition of the loan, and to deduct principal and interest at the time of sale, the trader–lender effectively guarantees debt seniority. Indeed, the trade–credit linkage may serve the dual purpose of enforcement and screening. The frequent exclusivity of such marketing agreements insures that the lender can observe the *entire* output of his borrowers, so as to monitor their ability to repay, as well as that of his prospective borrowers, so as to assess their future creditworthiness; at the same time, no other lender can have access to this information and thereby compete away borrowers (Siamwalla et al., 1990; Aleem, 1990). Moneylenders may also have more effective means of preventing their clients from absconding with the loan principal or diverting it to non-productive uses (Giné, 2005). While banks cannot legally prevent such strategic default beyond confiscating what collateral they hold, moneylenders may be able to exert various kinds of physical and psychological pressures to ensure repayment.

Lastly, moneylenders may more readily exchange information about borrowers’ repayment histories than banks in developing countries. An informal borrower with a reputation for default will not only be unable to obtain future loans from the same moneylender but may lose access to all local creditors. Kletzer and Wright (2000) show that, when credit histories are public information, punishing default by a debt moratorium until such time as the lender is repaid is a credible strategy. If any competing moneylender fails to respect this punishment by subsequently lending to the delinquent borrower, the other moneylenders can induce the borrower to default on this loan by offering him a better deal, thus ‘cheating the cheater’. When there is a high enough probability that credit histories are ‘forgotten’ or hidden, however, this type of equilibrium breaks down (Hoff and Stiglitz, 1997). Reputation equilibria are thus sensitive to the extent of village information networks, about which little is known.

These arguments aside, collecting on a past debt may not be an unalloyed benefit to the moneylender. When the borrower's output or investment depends importantly on his unverifiable effort, debt creates an incentive problem. The higher the debt burden, the more the borrower is working merely to pay off the

loan, the less willing he is to work, the lower his output, and, consequently, the more likely he is to default. Given ‘debt overhang’, the moneylender has to trade off higher debt collection against higher probability of default and collecting nothing. The resolution may involve forgiving debt. Evidence on the extent of debt forgiveness in informal credit markets is lacking (Fafchamps and Gubert, 2004, is a notable exception), but there are at least two reasons to believe that it is not widespread. First, in a long-term credit relationship, the moneylender has the option of rescheduling debt in the hopes that the borrower's fortunes will improve, a less drastic step than forgiveness. Second, the impact of forgiveness on incentives is diluted when the borrower and lender are not in an exclusive credit relationship. Since other creditors can free ride on the lowering of total debt, there may be too little forgiveness in equilibrium.

Landlord–moneylenders have motivated a considerable literature on ‘interlinked’ tenancy-credit contracts. Because the landlord must always verify the harvest of a share-tenant, he is in a better position to enforce debt repayment than an outside moneylender. Perhaps more importantly, however, the landlord has a stronger incentive to provide credit to his share-tenant than any other moneylender. This is because the landlord, in general, captures a larger share of incremental surplus due to an increase in the tenant's working capital than does an otherwise equivalent outside moneylender (see, for example, Basu, Bell and Bose, 2000). Even if the landlord himself faces relatively high credit costs, given his enforcement advantage, he may still prefer to on-lend funds to his tenant from a moneylender under a so-called ‘credit-layering’ arrangement (Mansuri, 2007).

The boundaries of moneylending are further obscured by the multifarious nature of credit. Traders, for example, often advance inputs in kind rather than cash, with interest collected through a markup on the price. Burkhart and Ellingsen (2004) rationalize this form of trade credit on the grounds that inputs are less easily diverted to non-productive uses than cash; in-kind loans thus alleviate a monitoring problem. Another form of in-kind lending occurs when landlords defer rental payments until after the harvest. Besides the possible monitoring advantage, such debt contracts have better incentive properties than share-contracts when the tenant's liability is limited (Innes, 1990) or when tenant risk aversion and yield variability are not too high (Arimoto, 2005). Since land is far and away the most important factor of agricultural production, the value of deferred rent may dwarf that of other seasonal borrowing.

Interest in moneylenders has centred around their role in modulating the impact of government policies, such as interest rate subsidies or controls, that can be effectively implemented only in the formal sector. The effects of such policies depend critically on the relationship between moneylenders and banks. The literature has taken two approaches to the formal–informal sector interaction. The first assumes a vertical structure whereby moneylenders act as middlemen, borrowing from the formal sector and on-lending to uncollateralized peasants (Hoff and Stiglitz, 1997; Floro and Ray, 1998). In the second approach, moneylenders and bankers compete with one another, with the residual demand for credit in the formal sector spilling over into the informal sector. Bell, Srinivasan and Udry (1997) and Kochar (1997) have moneylenders coexisting with banks by virtue of exogenous ceilings on formal sector credit, whereas Giné (2005) and Jain (1999) explicitly model moneylenders’ informational advantage over banks to obtain coexistence in equilibrium without imposing formal sector credit rationing.

See Also

- adverse selection
- agricultural finance
- micro-credit
- sharecropping

Bibliography

Aleem, I. 1990. Imperfect information, screening and the costs of informal lending: a study of rural credit markets in Pakistan. *World Bank Economic Review* 4, 329–49.

Arimoto, Y. 2005. State-contingent rent reduction and tenancy contract choice. *Journal of Development Economics* 76, 355–75.

Basu, K., Bell, C. and Bose, P. 2000. Interlinkage, limited liability and strategic interaction. *Journal of Economic Behavior and Organization* 42, 445–62.

Bell, C., Srinivasan, T. and Udry, C. 1997. Rationing, spillover and interlinking in credit markets: the case of rural Punjab. *Oxford Economic Papers* 49, 557–87.

Burkhart, M. and Ellingsen, T. 2004. In-kind finance: a theory of trade credit. *American Economic Review* 94, 569–90.

Fafchamps, M. and Gubert, F. 2004. Contingent loan repayment in the Philippines. Discussion Paper No. 215. Department of Economics, Oxford University. Online.

Floro, M. and Ray, D. 1998. Vertical links between formal and informal financial institutions. *Review of Development Economics* 1, 34–56.

Giné, X. 2005. Access to capital in rural Thailand: an estimated model of formal vs. informal credit. Policy Research Working Paper No. 3502. Washington, DC: World Bank.

Hoff, K. and Stiglitz, J. 1997. Moneylenders and bankers, price–increasing subsidies in a monopolistically competitive market. *Journal of Development Economics* 52, 429–62.

Innes, R. 1990. Limited liability and incentive contracting with ex-ante action choices. *Journal of Economic Theory* 52, 45–67.

Jain, S. 1999. Symbiosis vs. crowding-out, the interaction of formal vs. informal credit markets in developing countries. *Journal of Development Economics* 59, 419–44.

Kletzer, K. and Wright, B. 2000. Sovereign debt as intertemporal barter. *American Economic Review* 90,

621–39.

Kochar, A. 1997. An empirical investigation of rationing constraints in rural credit markets in India. *Journal of Development Economics* 53, 339–71.

La Ferrara, E. 2003. Kin groups and reciprocity: a model of credit transactions in Ghana. *American Economic Review* 93, 1730–50.

Mansuri, G. 2007. Credit Layering in informal financial markets. *Journal of Development Economics* 84, 715–30.

Siamwalla, A., Pithong, C., Poapongsakorn, N., Satsanguan, P., Nettayarak, P., Mingmaneenakin, W. and Tubpun, Y. 1990. The Thai Rural credit system: public subsidies, private information, and segmented markets. *World Bank Economic Review* 4, 271–96.

Stiglitz, J. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 383–410.

Udry, C. 1994. Risk and insurance in a rural credit market: an empirical investigation in Northern Nigeria. *Review of Economic Studies* 61, 495–526.

How to cite this article

Jacoby, Hanan G. "moneylenders in developing countries." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000246> doi:10.1057/9780230226203.1131

The New Palgrave Dictionary of Economics Online

monocentric versus polycentric models in urban economics

Tomoya Mori

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article overviews the development of the formal modelling framework for the urban spatial structure which started in 1960s and grew dramatically thereafter. Modelling in the 1970s focused on the endogenous formation of the central business district within a city. Then richer polycentric city models were developed in 1980s, where the number, location and spatial extent of the business districts are determined endogenously. The emergence of the new economic geography in 1990s provided a framework capable of explaining the spatial distribution of cities (rather than the business districts within a city) and their industrial structure in a general location-equilibrium model.

Keywords

urban economics, monocentric versus polycentric models in; spatial impossibility theorem; urban agglomeration; monopolistic competition; new economic geography; central business district

Article

The formal modelling of urban spatial structure originated in the monocentric city model by Alonso (1964). The model was extended to include production, transportation and housing by Mills (1967; 1972) and Muth (1969), and was eventually integrated into a unified framework by Fujita (1989). In these traditional models, the city is a priori assumed to be monocentric, that is, all production activities within a city are supposed to take place in a point representing the *central business district* (CBD), and all workers living in the surrounding area are supposed to commute to the CBD. The success of this model is primarily due to its compatibility with the competitive paradigm, since the existence of the CBD is a priori assumed. In order to explain the urban morphology, however, it is essential to endogenize the CBD formation. For this purpose, Fujita (1986) provided a very useful insight based on the spatial impossibility theorem of Starrett (1978): in order to have endogenous formation of economic agglomeration, the model must have at least one of the following three elements: (*a*) heterogeneous

space, (*b*) non-market externalities in production and/or consumption, and (*c*) imperfectly competitive markets.

The approach based on (*a*) explains the formation of the CBD by *comparative advantage* among locations, while otherwise retaining the competitive paradigm. One of the earliest such attempts was made by Schweizer, Varaiya and Hartwick (1976).

Most models of type (*b*) are based on *externalities from non-market interactions*. The earliest attempt was made by Solow and Vickrey (1971). In the one-dimensional location space, they considered the optimal allocation of urban land between business areas and roads when each unit of business area is assumed to generate a given number of trips to every other unit. But the first model of residential land use of this type is by Beckmann (1976), where the utility of each individual directly depends on the average distance to all other individuals and the amount of her land consumption. This preference leads to a *bell-shaped spatial population distribution as well as land rent curves*, where the CBD is represented by a densely inhabited area around the central location.

While Beckmann, Solow and Vickrey considered only a single type of agents (firms or consumers), Ogawa and Fujita (1980) and Imai (1982) developed two-sector monocentric models of a one-dimensional city. The dispersion force in this case is generated through land and labour markets. That is, the agglomeration of firms increases the commuting distance for their workers on average, which in turn pushes up the wage rate and land rent around the agglomeration, and this higher cost of labour and land discourages further agglomeration of firms. The most recent contribution along this line is by Lucas and Rossi-Hansberg (2002), who formally demonstrate the existence of an equilibrium and the endogenous formation of the CBD.

In the endogenous monocentric models discussed so far, the optimal distribution of firms requires greater concentration near the centre than does the equilibrium distribution. The reason is the locational externality generated by individuals: while the location of each individual directly affects the travelling cost for others to make contact with this individual, it is not taken into account when each individual makes a location decision.

Building on Ogawa and Fujita (1980), the first model of a *polycentric city* was developed by Fujita and Ogawa (1982). Their key assumption is that the benefit from interactions between two firms is a negative exponential function of the distance between them, unlike the linear dependence in previous models. When commuting costs are relatively high, this assumption leads to the formation of *multiple business districts* and the possibility of *multiple equilibria*.

The first urban economic model based on (*c*) is by Fujita (1988). His model demonstrated that pure market interactions alone can explain the agglomeration of economic activities with the use of the Chamberlinian monopolistic competition model. The agglomeration force is generated from the interaction among preference for product variety, transport costs, and increasing returns at the level of individual producers. In this model, the city may be monocentric or polycentric. Also it is possible that business and residential districts are mixed. These works were critical for the emergence of *the new economic geography* (NEG) in the 1990s (Krugman, 1991a; 1991b; Fujita, 1993).

In the application of the NEG to urban economics initiated by Fujita and Krugman (1995), there are two key features. The first is the *general equilibrium modelling of an entire spatial economy* unlike all the models presented so far. The second is its focus on *the spatial distribution of cities*, while abstracting from the intra-city spatial structure. In particular, it is assumed that mobile firms and workers do not

occupy land, so that an agglomeration of firms and population, that is, a city, forms at a point on the continuous location space. The second feature dramatically increases the tractability of the model. The agglomeration force in this model is essentially the same as in Fujita (1988), while the dispersion force is generated from the presence of immobile resources through transport costs between cities and non-city locations. The key to this approach is the recognition that the profitability of any given location for a firm can be represented by an *index of market potential*. The market potential at a given location reflects the trade-off among the proximity to consumers, the degree of competition, and the production cost at that location. In particular, the market potential of a given industry sharply decreases when it moves away from a city in which this industry is agglomerated, and then starts increasing again after a certain distance, exhibiting the presence of an *agglomeration shadow*. Differences in the degree of product differentiation and/or transport costs among industries lead to differences in the size of the agglomeration shadow, which in turn result in variations in the (roughly constant) spacing of agglomerations among industries (Fujita and Mori, 1997). In the presence of multiple industries, the *inter-industry demand externalities* lead to a formation of *hierarchical city systems* (Fujita, Krugman and Mori, 1999). This is reminiscent of Christaller (1933): the set of industries found in a smaller city is a subset of those found in a larger city. Furthermore, the relative decrease in transport costs for urban sectors may eventually lead to the formation of a *megalopolis* consisting of large core cities that are connected by an *industrial belt*, that is, *a continuum of small cities* (Mori, 1997). NEG remains the only general location-equilibrium framework which can investigate the spatial distribution of cities and their industrial structure in a unified manner.

There is also a large literature of spatial oligopoly (hence, type *c*) aiming to explain the spatial concentration of stores through *statistical economies of scale*. These models assume that consumers have imperfect information regarding the types (and the prices) of commodities sold by stores before they visit them. The greater the agglomeration of stores, the more likely it is that consumers will find their favourite commodities. The concentration of stores is explained by the market-size effect due to taste uncertainty and/or lower price expectation (see, for example, Konishi, 2005).

Finally, in all the models introduced thus far, all agents are assumed to be atomistic. Hence, land and labour markets are perfectly competitive. In contrast, Henderson and Mitra (1996) offer a model of *suburbanization* in which new *edge cities* are formed by *large land-developers* in the suburbs of the old CBD, formalizing Garreau's observation (1991) on the recent development of edge cities within large US metro areas. Given an existing CBD, the developer of a new edge city chooses the location and capacity of its business district strategically to maximize profits. The developer exercises monopsony power in the labour market in the edge city though her control over aggregate employment there. The proximity to the old CBD increases production efficiency through easier communication of firms between the CBD and the edge city, while it also increases residential land rents and wages of workers in the edge city. This model thus incorporates elements (*b*) and (*c*).

See Also

- location theory
- spatial economics

- urban agglomeration
- urban economics
- urban growth
- urban production externalities
- urbanization

Bibliography

Alonso, W. 1964. *Location and Land Use*. Cambridge, MA: Harvard University Press.

Beckmann, M. 1976. Spatial equilibrium in the dispersed city. In *Mathematical Land Use Theory*, ed. Y. Papageorgiou. Lexington MA: Lexington Books.

Christaller, W. 1933. *Central Places in Southern Germany*, trans. C. Baskin. London: Prentice Hall, 1966.

Fujita, M. 1986. Urban land use theory. In *Location Theory*, ed. R. Arnott. London: Harwood Academic Publishers.

Fujita, M. 1988. A monopolistic competition model of spatial agglomeration: a differentiated product approach. *Regional Science and Urban Economics* 18, 87–124.

Fujita, M. 1989. *Urban Economic Theory: Land Use and City Size*. Cambridge: Cambridge University Press.

Fujita, M. 1993. Monopolistic competition and urban systems. *European Economic Review* 37, 308–15.

Fujita, M. and Krugman, P. 1995. When is the economy monocentric?: von Thünen and Chamberlin unified. *Regional Science and Urban Economics* 25, 505–28.

Fujita, M., Krugman, P. and Mori, T. 1999. On the evolution of hierarchical urban systems. *European Economic Review* 43, 209–51.

Fujita, M. and Mori, T. 1997. Structural stability and evolution of urban systems. *Regional Science and Urban Economics* 27, 399–442.

Fujita, M. and Ogawa, H. 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics* 12, 161–96.

Fujita, M. and Thisse, J.-F. 2002. *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge: Cambridge University Press.

Garreau, J. 1991. *Edge City: Life on the New Frontier*. New York: Doubleday.

Henderson, J. and Mitra, A. 1996. The new urban landscape: developers and edge cities. *Regional Science and Urban Economics* 26, 613–43.

Imai, H. 1982. CBD hypothesis and economies of agglomeration. *Journal of Economic Theory* 28, 275–99.

Krugman, P. 1991a. Increasing returns and economic geography. *Journal of Political Economy* 99, 483–99.

Krugman, P. 1991b. *Geography and Trade*. Cambridge, MA: MIT Press.

Konishi, H. 2005. Concentration of competing retail stores. *Journal of Urban Economics* 58, 488–512.

Lucas, R. and Rossi-Hansberg, E. 2002. On the internal structure of cities. *Econometrica* 70, 1445–76.

Mills, E. 1967. An aggregative model of resource allocation in a metropolitan area. *American Economic Review* 57, 197–210.

Mills, E. 1972. *Studies in the Structure of the Urban Economy*. Baltimore, MD: Johns Hopkins University Press.

Mori, T. 1997. A modeling of megalopolis formation: the maturing of city systems. *Journal of Urban Economics* 42, 133–57.

Muth, R. 1969. *Cities and Housing*. Chicago: University of Chicago Press.

Ogawa, H. and Fujita, M. 1980. Equilibrium land use patterns in a non-monocentric city. *Journal of Regional Science* 20, 455–75.

Schweizer, U., Varaiya, P. and Hartwick, J. 1976. General equilibrium and location theory. *Journal of Urban Economics* 3, 285–303.

Solow, R. and Vickrey, W. 1971. Land use in a long narrow city. *Journal of Economic Theory* 3, 1468–88.

Starrett, D. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 9, 418–48.

How to cite this article

Mori, Tomoya. "monocentric versus polycentric models in urban economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000368>
doi:10.1057/9780230226203.1132

The New Palgrave Dictionary of Economics Online

monopolistic competition

G.C. Archibald
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advertising; Chamberlin, E. H.; characteristics; discriminating monopoly; efficient allocation; entry; excess capacity; goods approach versus characteristics approach; Hotelling, H.; imperfect competition; increasing returns; labour market discrimination; marginal productivity theory; market failure; monopolistic competition; monopsony; non-convexithy; ologopoly; optimality; planning; price discrimination; quasi-rent; Robinson, J. V.; side-payments; spatial competition; tangency solution

Article

There is at least an oral tradition that the origin of theories of monopolistic competition is Sraffa's (1926). In the case of Joan Robinson (1933) this may well be true. In the case of Edward Chamberlin (1933) it cannot be: the book was developed from a Ph.D. thesis supervised by Allyn A. Young submitted on 1 April 1927. Indeed, Chamberlin (1933, p. 5 n.) refers to Sraffa's paper as appearing ‘since the above was written’.

It is, none the less, convenient to take Sraffa's implicit criticism of Marshall (1890) as a starting point. The increasing-marginal-cost condition, necessary for a competitive equilibrium, was, he asserted, not satisfied in many firms that could not possibly be described as ‘Marshallian monopolies’. Thus there existed no appropriate model for an apparently common class of firms (or markets – Sraffa was quite aware of the problems of product heterogeneity). The works of Chamberlin and Mrs Robinson, however diversely prompted, may be seen as attempts to fill what became known as the gap between Marshall's polar cases of monopoly and perfect competition. The gap they had in mind was not filled by oligopoly models, which were already well known. Chamberlin certainly had a ‘more competitive’ model in mind (free entry). Mrs Robinson was so vague about the construction of the demand curve that it is hard to be sure where ‘imperfect competition’ leaves off and oligopoly begins, but I read her as in the same spirit as Chamberlin. Whether we can in fact reasonably construct a model of imperfect or monopolistic competition which is not an oligopoly model is still an open question.

The work of Edward Chamberlin and Joan Robinson

It would not, I think, be a wise use of space to review here the old dispute between Chamberlin (persistent and vociferous) and Mrs Robinson (reluctant and *dégagée*) about whether or not their models were ‘the same’. Nor do I wish to dismiss the question as merely ‘braces versus suspenders’. Instead, I shall note briefly what it seems to me they had in common and what not. I start with what they had in common.

Both had downward-sloping demand curves (although their construction differed somewhat; see below), but tried to distinguish their models from that of simple or Marshallian monopoly.

This they were able to do because they assumed that the competitive mechanism worked not only through prices but, most importantly, through entry of firms (products). Thus both made an important generalization and extension of Marshall's proposition that competition would ensure that pure profits were only quasi-rents. Indeed, both thought that free entry is a sufficient condition for the elimination of all pure profit in full equilibrium, and thus both exhibited the famous tangency solution.

Thanks to the downward-sloping demand curve, both were able to exhibit profit-maximizing equilibria consistent with non-convexities in the technology, that is to answer Sraffa. (One consequent result, the familiar excess-capacity theorem, is discussed below.)

Both should, in my judgement, be credited with a major extension of the marginal productivity theory of distribution.

There are, none the less, some differences, and they may explain why, in spite of the many elegant features of Mrs Robinson's analysis, Chamberlin's ‘monopolistic competition’ seems to have been the more enduring model (or, at least, title).

First, there are the famous Chamberlinian ‘groups’ or industries, groups of similar but not identical products, ill-defined as they may have been. The lack of identity justified the downward slope of the individual demand curve; the assumptions of large numbers and symmetry were carefully stated to justify the assumption of Cournot–Nash behaviour instead of the recognition of oligopolistic interdependence. (The famous construction of the ‘perceived’ demand curve, DD' and the ‘share-of-the-market’ demand curve, dd' was designed to explain disequilibrium adjustment behaviour. It has little to do with the properties of full equilibrium which, as in Mrs Robinson's version, is characterized by the elimination of super-normal profit.)

By contrast, Mrs Robinson's treatment of the demand curve seems cavalier. She simply asserted (1933, p. 21) that it shows what the firm will sell at each price when all other adjustments are completed. Whether she had in mind the Cournot–Nash assumption of Chamberlin, or intended to encompass in her model some types of oligopolistic behaviour, is obscure. No adjustment mechanism was suggested. The existence of a full-adjustment demand curve, on which the firm's profit-maximizing decisions are based, was simply postulated as a primitive of the model.

Chamberlin was much more ambitious than Mrs Robinson: he attempted to include product-choice and advertising in the model. I say ‘attempted’ because it was here that his technique let him down most seriously. Two-dimensional geometry only allowed him to illustrate equilibrium conditions pairwise, and he was never able to exhibit the full set of simultaneous equilibrium conditions, to consider second-order conditions, or to carry out any comparative static analysis. Mrs Robinson confined her attention to what her two-dimensional geometry could handle, omitting advertising and quality from the model, and gave us her elegant analysis of discriminating monopoly and monopsony (with its arresting application

to the theory of labour market discrimination).

Criticisms

It would be impossible to review the whole debate over monopolistic competition in limited space. I shall concentrate on those criticisms which seem to be still with us, and lead us to recent advances. There is no doubt that ‘groups’ were ill-defined. A common definition, still employed, is that we have a group if we can isolate a set of products such that (i) cross-elasticities of demand between them are ‘large’ and (ii) cross-elasticities of demand between all members of the set and its complement are ‘small’. Triffin (1940) pointed out that there is no analytical cut-off between small and large, and concluded that there was no valid analytical construct between the individual firm and the whole economy. We may take this a little further. We may say that a satisfactory taxonomy induces the discrete metric. A continuous function, such as a cross-elasticity, cannot induce the discrete metric and, accordingly, cannot generate a satisfactory taxonomy. I shall argue below that there now exists an analytically satisfactory way of defining groups, that is, one that induces the discrete metric.

Kaldor (1934; 1935) suggested very early in the discussion that chains of overlapping oligopolies might be empirically more likely than competitive groups operating in virtual isolation from other groups. This raises sharply a question which is still with us: what are the necessary and sufficient conditions for competition to be general, or ‘diffuse’, that is for the assumption of Cournot–Nash behaviour to be plausible, as opposed to localized or oligopolistic so that the possibility of strategic behaviour has to be admitted.

Several writers on spatial competition have shown recently that free entry cannot be relied upon as a sufficient condition to eliminate super-normal profit, that is to generate the tangency solution (see, for example, Eaton, 1976; Eaton and Lipsey, 1978). This follows basically from the idea that capital is product- (location-) specific, and long-lasting, and has accordingly to be *committed*. Hotelling (1929) and Chamberlin (1957) thought that monopolistic and spatial competition were, in some sense, the ‘same’ subject. Given the spatial results, the ‘sameness’ of the subjects, or models, becomes an urgent question.

Application of Samuelson's (1947) programme, the ‘qualitative calculus’, to Chamberlin's model, even when ‘making the best of it’ (to make the criticism more effective) by, for example ignoring the fact that groups were ill-defined, unfortunately showed it to be qualitatively almost empty in the sense of generating few qualitative comparative-static predictions (Archibald, 1961). For the individual firm, the reason is the now familiar one: in the multivariate case, the assumption that sufficient extremum conditions are satisfied is not enough to sign the cofactors of off-diagonal elements in the matrix of second-order coefficients. For the group, the reason is essentially the non-convexity of the technology. If, for example demand falls (due, say, to an excise tax), firms exit. When full equilibrium is restored, surviving firms may be producing more or less, that is incurring lower or higher average costs. It also turned out that even the excess capacity theorem did not survive the explicit introduction of advertising in the model (excess capacity remains a possibility but not an entailment). Demsetz (1964) made the interesting suggestion that, by the processes of spin-off, merger, and subcontracting, firms would become so structured that the quantity that minimized average production costs would also minimize average selling costs, in which case equilibrium could not entail excess capacity. It unfortunately turned

out that, analytically, this model was inadequately specified (Archibald, 1967), but the idea might still be worth pursuing.

Some reactions to the welfare implications of Chamberlin's model were strange. The reaction of several writers to the excess capacity theorem seems to have been ‘It can't be true, but, if it is, it is wicked’. Chamberlin replied (1957), quite reasonably, that optimality conditions for an economy with homogeneous product-groups would not necessarily serve as benchmarks for an economy with some increasingness in returns and differentiated product-groups. Little was in fact known about the welfare economics of an economy with non-convexities in its technology.

Some recent advances and unsolved problems

After some years in which the theory of monopolistic competition was relatively neglected, or at least not much advanced, there has been a recent revival of interest, and a new approach to the subject has emerged. The standard approach, which I shall call the ‘goods approach’, is in the traditional Walrasian (or Hicksian) style: see the papers by Dixit and Stiglitz (1977), Hart (1979) and Spence (1976). The new, or ‘characteristics approach’, follows the work of Lancaster: see his (1966), (1971), (1975) and (1979), also Gorman (1980). This approach to monopolistic competition was advocated in Archibald, Eaton and Lipsey (1986). I note briefly the main features of these two quite distinct approaches.

The goods approach is familiar and traditional, but some features deserve emphasis in the present context. Goods themselves are, of course, the primitives of analysis. There is a fixed vector of possible goods, usually either finite or countably infinite. The utility function is defined on the goods, and there is usually a ‘representative consumer’ (in some sense that requires definition) who consumes some of each of the goods actually produced. If groups are to be identified, the cross-elasticity taxonomy is employed. If individual firm behaviour is considered, the Cournot–Nash assumption is commonly employed. Full equilibrium is characterized by normal profit.

There are some points to notice here. In some models, the assumption of a fixed vector of goods implies that the technology is not continuous: a firm may choose to produce a good (or quality) x_0 or x_1, say, but cannot produce a good arbitrarily close to either of them (in some space of attributes). Now, if these attributes (characteristics) of goods are continuous (for example, the fuel consumption of automobilies, the alcohol content of beer), this is a restrictive, and somewhat strange, assumption. Furthermore, it induces an immediate, and perhaps unwelcome, answer to the question, ‘are models of monopolistic and spatial competition in some sense the same?’, as Hotelling and Chamberlin thought. The space in most spatial models is a continuum, whether in one dimension or two, whence any analogy between the models breaks down at the first step in their construction.

The assumption of a representative consumer who, necessarily, consumes some of each good produced prevents us from taking into account that diversity of tastes which is an obvious feature of the real world. In a characteristics model, the consumer buys no more goods than there are characteristics that he wishes to consume, and if the number of goods produced exceeds the number of ‘relevant’ characteristics, he buys none of many (perhaps most) goods. This seems to capture an important feature of reality; but it must immediately be admitted that tractable methods of modelling the diversity of preferences have yet to be developed.

The characteristics model is doubtless now familiar too, and only a few points need to be made. The

characteristics of goods, rather than the goods themselves, are the primitives of analysis. The technology is assumed to be continuous, in the sense that, if y_1 and y_2 are two goods embodying different mixes of two characteristics, z_1 and z_2 say, then it is possible to produce any good y_i embodying a convex combination of the quantities of z_1 and z_2 embodied in y_1 and y_2. It is thus always possible to produce a good $\in$-close to any other good in the characteristics space. As in spatial models, possible locations form a continuum. Some increasingness of returns is necessarily assumed: with everywhere constant returns, we might expect a ‘production point’ at every ‘consumption point’, whether in physical or characteristics space. Thus out of the continuum of possibilities, only a finite number of goods is produced at any time. None the less, it is assumed that, at least in developed economies, the number of goods produced exceeds the number of characteristics desired by consumers. This can only be the consequence of diversity of tastes: if all consumers wanted the same characteristics mix, the number of goods produced would be *less* than the number of characteristics.

An immediate advantage of the characteristics approach is that it allows us to give an analytic definition of a ‘group’ or industry. It is assumed that the consumption technology is linear, that is, characteristics are ‘produced’ by goods according to $z=Ay$ where z is the $1\times m$ vector of characteristics, A is $m\times n$, and y is the $n\times 1$ vector of produced goods. Suppose now that we can partition z, and correspondingly y, so that the corresponding arrangement of the elements of A is block diagonal. Consider one such block, and the corresponding subsets of z and y. We may call this a group: the elements of the subset of y produce only elements of the corresponding subset of z, and no elements of the complement of this subset in y produce any elements of this subset in z.

This taxonomy induces the discrete metric: two goods either do or do not unambiguously belong to the same group. Whether or not there exist, empirically, any groups corresponding to this definition is yet to be discovered.

To complete this sketch of the characteristics approach, let us consider a group of possible goods embodying only two characteristics, z_1 and z_2, say. Then any produced good, y_i say, can be described by θ_i where $\tan \theta_i = z_2 / z_1$. The good is completely described by the pair of numbers (p_i, θ_i) where p_i is the price (reciprocal of the length of the vector θ_i to the z_1, z_2 point that can be bought for some fixed amount). The firm's problem is to choose θ_i as well as p_i. The economist's first problem is to characterize the equilibrium vector of θ's as well as p's. His second problem is to characterize the optimal vector of θ's as well as p's. For the first problem, he needs, of course, to know whether competition is oligopolistic or diffuse. (The prior problem of existence has, of course, been thoroughly investigated for the competitive general equilibrium model, and for some partial equilibrium spatial and small-group models. Little seems to have been done on existence in a Chamberlinian model, at least in the characteristics approach.)

The problem of the socially optimal product choice is of great practical as well as theoretical interest. There is evidence, mostly anecdotal, that the planned economies frequently produce the ‘wrong’ goods. In the planning literature a fixed vector of homogeneous goods is commonly assumed, and the problem of product choice is not addressed. We similarly lack welfare criteria to tell us if a capitalist economy makes a good job of product selection.

The problem is in fact most difficult. Lancaster (1979) showed that, given some increasingness in returns, considerations of efficiency cannot be successfully divorced from distributional considerations.

The problem appears even more starkly in a series of papers by Brown and Heal (1979), (1980), (1981), since they stay with the conventional goods approach. Consider an ‘economy’ as described by an endowment of resources, a given but non-convex technology, and tastes. They show that, for an arbitrary distribution of ownership, there may exist no efficient allocation of resources. They are also able to show that, for the given ‘economy’, there always exists a share-ownership distribution (in particular, the equal distribution) such that an efficient allocation does exist.

If this is true for an economy with a fixed product vector, we might conjecture that it is true for an economy in which the product vector is yet to be chosen. We can at least see why efficiency and distribution are entangled in a spatial model (whether the space be geographical or of characteristics). Let there be a given distribution of consumers, whether by location or preferred characteristics mix, and a distribution of stores or products. Assume that the capital specific to one product or location (store) wears out, and is due for replacement. Assume further that some of the mass of demand has shifted (arbitrarily, to the left in the appropriate space). ‘Common sense’ suggests that the new capital be installed to the left of the old. But this is not a Pareto-efficient move: those consumers remaining to the right will unambiguously lose.

Spence (1976) investigates optimality in monopolistic competition. He adopts the conventional goods model, assuming away income effects, and takes the sum of consumer and producer surplus as his welfare criterion. He is able to show (i) that if sellers can price-discriminate, their profit function will coincide with the welfare maximand, and the optimal product vector (from the possible set) will be produced; and (ii) if not, not: there may be too many or too few products marketed. The reason, roughly speaking, is that, with some increasingness of returns, and products to be chosen, price is not a sufficient signal: we have a species of market failure. Thus there is no market in which you and I and the producers may arrange side-payments such that, by agreeing on the same good(s), we all benefit from the increasingness of returns.

I conjecture that Spence has given us all the pure efficiency results that are to be had. If we follow Lancaster, and Brown and Heal, and do not ignore distributional considerations, it is not obvious what results we may hope for.

We urgently need to know the necessary and sufficient conditions for competition to be diffuse (Chamberlinian) as opposed to local or oligopolistic. The next question follows: on what conditions does the small-group model become asymptotically competitive? So far, we have only a scattering of results. Consider the set of products in a space of two characteristics, or of stores along a line. It is obvious that no product, or store, can have more than two neighbours: we appear to have Kaldor's chain of overlapping oligopolies. What happens if we increase the number of consumers and products, or stores, without limit is less obvious. While each outlet still has no more than two neighbours, its scope for price setting is evidently diminished. We might conjecture that the asymptotic results in this case will be approximately competitive. Now let the dimensions of the tangency solution space increase. It has been shown, by Archibald and Rosenbluth (1975), that when the number of tangency solution is four, the number of neighbours (immediate competitors) each product may have approaches half the number of products in the space. This is a necessary condition for competition among diverse products to be Chamberlinian. Sufficient conditions have not been established. These authors, and others, also considered the possibility of ‘pre-emptive entry’: an incumbent firm in a growing market occupies a point (in physical or characteristics space) before it is normally profitable to do so in order to deter new competition.

Hart (1979) gets asymptotically competitive results in a goods model. He assumes, however that the output of each firm is bounded from above so that the output of each firm can be made as small as we like relative to the whole economy. Further, replication involves increasing the number of consumers each of whom has one of a finite set of preferences, that is, cloning them. What is not yet known is what happens asymptotically in an economy in which (i) the output of the individual firm is not bounded, (ii) the 'address' of products, in the sense of Archibald, Eaton and Lipsey matters, and (iii) as the number of consumers increases, so does the diversity of preferences.

See Also

- advertising
- Chamberlin, Edward Hastings
- competition
- market structure
- oligopoly
- product differentiation
- Robinson, Joan Violet

Bibliography

Archibald, G.C. 1961. Chamberlin versus Chicago. *Review of Economic Studies* 24, 9–28.

Archibald, G.C. 1967. Monopolistic competition and returns to scale. *Economic Journal* 77, 405–12.

Archibald, G.C. and Rosenbluth, G. 1975. The 'new' theory of consumer demand and monopolistic competition. *Quarterly Journal of Economics* 80, 569–90.

Archibald, G.C., Eaton, B.C. and Lipsey, R.G. 1986. Address models of value theory. In *New Developments in the Analysis of Market Structure*, ed. J.E. Stiglitz. Cambridge, MA: MIT Press.

Brown, D.J. and Heal, G. 1979. Equity, efficiency and increasing returns. *Review of Economic Studies* 57, 571–85.

Brown, D.J. and Heal, G.M. 1980. Two-part tariffs, marginal cost pricing and increasing returns in a general equilibrium model. *Journal of Public Economics* 13, 25–49.

Brown, D.J. and Heal, G. 1981. Welfare theorems for economies with increasing returns. Essex Economic Papers No. 179.

Chamberlin, E.H. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University

Press; London: Oxford University Press.

Chamberlin, E.H. 1957. *Towards a More General Theory of Value*. New York: Oxford University Press.

Demsetz, H. 1964. The welfare and empirical implications of monopolistic competition. *Economic Journal* 74, 623–41.

Dixit, A.K. and Stiglitz, J.E. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Eaton, B.C. 1976. Free entry in one dimensional models. *Journal of Regional Science* 16, 21–33.

Eaton, B.C. and Lipsey, R.G. 1978. Freedom of entry and the existence of pure profit. *Economic Journal* 88, 455–69.

Gorman, W.M. 1980. A possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies* 47, 843–57.

Hart, O. 1979. Monopolistic competition in a large economy with differentiated commodities. *Review of Economic Studies* 46, 1–30.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39, 41–57.

Kaldor, N. 1934. Mrs. Robinson's ‘Economics of imperfect competition’. *Economica* NS 1, 335–41.

Kaldor, N. 1935. Market imperfections and excess capacity. *Economica* NS 2, 33–50.

Lancaster, K.J. 1966. A new approach to consumer theory. *Journal of Political Economy* 74, 132–57.

Lancaster, K. 1971. *Consumer Demand: A New Approach*. New York: Columbia University Press.

Lancaster, K. 1975. Socially optimal product differentiation. *American Economic Review* 65, 567–85.

Lancaster, K.J. 1979. *Variety, Equity, and Efficiency*. New York: Columbia University Press.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Spence, A.M. 1976. Product selection, fixed costs, and monopolistic competition. *Review of Economic Studies* 43, 217–35.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36, 535–50.

Triffin, R. 1940. *Monopolistic Competition and General Equilibrium Theory*. Cambridge, MA: Harvard University Press.

How to cite this article

Archibald, G. C. "monopolistic competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000250> doi:10.1057/9780230226203.1133

The New Palgrave Dictionary of Economics Online

monopoly capitalism

Paul M. Sweezy

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

To Marxian economists, ‘monopoly capitalism’ denotes the stage of capitalism beginning in the last quarter of the 19th century and maturing after the Second World War. While Marx and Engels wrongly thought it heralded the demise of capitalism, later thinkers, like Sweezy and Baran, have tried to identify its main features and ‘laws of motion’. They argue that, by increasing the savings potential of the economy and reducing opportunities for productive investment, monopoly capitalism suppresses levels of income and employment. No other approach, whether mainstream or traditional Marxian, has satisfactorily explained capitalism's growing tendency towards stagnation in the 20th century.

Keywords

accumulation of capital; Baran, P.; Burns, A.; capitalism; cartellization; concentration of capital; Engels, F.; Hilferding, R.; Kalecki, M.; Keynesianism; laws of motion of capitalism; Lenin, V.; Marx, K.; monopoly capitalism; stagnation; Sweezy, P.; Veblen, T.

Article

Among Marxian economists ‘monopoly capitalism’ is the term widely used to denote the stage of capitalism which dates from approximately the last quarter of the 19th century and reaches full maturity in the period after World War II. Marx's *Capital*, like classical political economy from Adam Smith to John Stuart Mill, was based on the assumption that all commodities are produced by industries consisting of many firms, or capitals in Marx's terminology, each accounting for a negligible fraction of total output and all responding to the price and profit signals generated by impersonal market forces. Unlike the classical economists, however, Marx recognized that such an economy was inherently unstable and impermanent. The way to succeed in a competitive market is to cut costs and expand production, a process which requires incessant accumulation of capital in ever new technological and organizational forms. In Marx's words: ‘The battle of competition is fought by cheapening of

commodities. The cheapness of commodities depends, *ceteris paribus*, on the productiveness of labour, and this again on the scale of production. Therefore the larger capitals beat the smaller.' Further, the credit system which 'begins as a modest helper of accumulation' soon 'becomes a new and formidable weapon in the competitive struggle, and finally it transforms itself into an immense social mechanism for the centralization of capitals' (Marx, 1867, ch. 25, sect. 2). Marx, and even more clearly Engels when preparing the second and third volumes of *Capital* for the printer two decades later, concluded, in the latter's words, that 'the long cherished freedom of competition has reached the end of its tether and is compelled to announce its own palpable bankruptcy' (Marx, 1894, ch. 27).

There is thus no doubt that Marx and Engels believed capitalism had reached a turning point. In their view, however, the end of the competitive era marked not the beginning of a new stage of capitalism but rather the beginning of a transition to the new mode of production that would take the place of capitalism. It was only somewhat later, when it became clear that capitalism was far from on its last legs that Marx's followers, recognizing that a new stage had actually arrived, undertook to analyse its main features and what might be implied for capitalism's 'laws of motion'.

The pioneer in this endeavour was the Austrian Marxist Rudolf Hilferding whose magnum opus *Das Finanzkapital* appeared in 1910. A forerunner was the American economist Thorstein Veblen, whose book *The Theory of Business Enterprise* (1904) dealt with many of the same problems as Hilferding's: corporation finance, the role of banks in the concentration of capital, etc. Veblen's work, however, was apparently unknown to Hilderding, and neither author had a significant impact on mainstream economic thought in the English-speaking world, where the emergence of corporations and related new forms of business activity and organization, though the subject of a vast descriptive literature, was almost entirely ignored in the dominant neoclassical orthodoxy.

In Marxist circles, however, Hilferding's work was hailed as a breakthrough, and its pre-eminent place in the Marxist tradition was assured when Lenin strongly endorsed it at the beginning of his Imperialism, the Highest Stage of Capitalism. 'In 1910,' Lenin wrote, 'there appeared in Vienna the work of the Austrian Marxist, Rudolf Hilferding, Finance Capital … . This work gives a very valuable theoretical analysis of "the latest phase of capitalist development", the subtitle of the book.'

As far as economic theory in the narrow sense is concerned, Lenin added little to Finance Capital, and in retrospect it is evident that Hilferding himself was not successful in integrating the new phenomena of capitalist development into the core of Marx's theoretical structure (value, surplus value, and above all the process of capital accumulation). In chapter 15 of his book ('Price Determination in the Capitalist Monopoly. Historical Tendency of Finance Capital') Hilferding, in seeking to deal with some of these problems, came up with a very striking conclusion which has been associated with his name ever since. Prices under conditions of monopoly, he thought, are indeterminate and hence unstable. Wherever concentration enables capitalists to achieve higher than average profits, suppliers and customers are put under pressure to create counter combinations which will enable them to appropriate part of the extra profits for themselves. Thus monopoly spreads in all directions from every point of origin. The question then arises as to the limits of 'cartellization' (the term is used synonymously with monopolization). Hilferding answers:

> The answer to this question must be that there is no absolute limit to cartellization. What exists rather is a tendency to the continuous spread of cartellization. Independent industries, as we have seen, fall more and more under the sway of the cartellized ones,

> ending up finally by being annexed by the cartellized ones. The result of this process is then a *general cartel*. The entire capitalist production is consciously controlled from one center which determines the amount of production in all its spheres … . It is the consciously controlled society in antagonistic form.

There is more about this vision of a future totally monopolized society, but it need not detain us. Three quarters of a century of monopoly capitalist history has shown that while the tendency to concentration is strong and persistent, it is by no means as ubiquitous and overwhelming as Hilferding imagined. There are powerful counter-tendencies – the breakup of existing firms and the founding of new ones – which have been strong enough to prevent the formation of anything even remotely approaching Hilferding's general cartel.

The first signs of important new departures in Marxist economic thinking began to appear toward the end of the interwar years, i.e., the 1920s and 1930s; but on the whole this was a period in which Lenin's *Imperialism* was accepted as the last word on monopoly capitalism, and the rigid orthodoxy of Stalinism discouraged attempts to explore changing developments in the structure and functioning of contemporary capitalist economies. Meanwhile, academic economists in the West finally got around to analysing monopolistic and imperfectly competitive markets (especially Edward Chamberlin and Joan Robinson), but for a long time these efforts were confined to the level of individual firms and industries. The so-called Keynesian revolution which transformed macroeconomic theory in the 1930s was largely untouched by these advances in the theory of markets, continuing to rely on the time-honoured assumption of atomistic competition.

The 1940s and 1950s witnessed the emergence of new trends of thought within the general framework of Marxian economics. These had their roots on the one hand in Marx's theory of concentration and centralization which, as we have seen, was further developed by Hilferding and Lenin; and on the other hand in Marx's famous Reproduction Schemes presented and analysed in Volume II of *Capital*, which were the focal point of a prolonged debate on the nature of capitalist crises involving many of the leading Marxist theorists of the period between Engels' death (1895) and World War I. Credit for the first attempt to knot these two strands of thought into an elaborated version of Marxian accumulation theory goes to Michal Kalecki, whose published works in Polish in the early 1930s articulated, according to Joan Robinson and others, the main tenets of the contemporaneous Keynesian 'revolution' in the West. Kalecki had been introduced to economics through the works of Marx and the great Polish Marxist Rosa Luxemburg, and he was consequently free of the inhibitions and preconceptions that went with a training in neoclassical economics. He moved to England in the mid-1930s, entering into the intense discussions and debates of the period and making his own distinctive contributions along the lines of his previous work and that of Keynes and his followers in Cambridge, Oxford and the London School of Economics. In April 1938 Kalecki published an article in *Econometrica* ('The Distribution of the National Income') which highlighted differences between his approach and that of Keynes, especially with respect to two crucially important and closely related subjects, namely, the class distribution of income and the role of monopoly. With respect to monopoly, Kalecki stated at the end of the article a position which had deep roots in his thinking and would henceforth be central to his theoretical work:

> The results arrived at in this essay have a more general aspect. A world in which the

> degree of monopoly determines the distribution of the national income is a world far removed from the pattern of free competition. Monopoly appears to be deeply rooted in the nature of the capitalist system: free competition, as an assumption, may be useful in the first stage of certain investigations, but as a description of the normal state of capitalist economy it is merely a myth.

A further step in the direction of integrating the two strands of Marx's thought – concentration and centralization on the one hand and crisis theory on the other – was marked by the publication in 1942 of *The Theory of Capitalist Development* by Paul M. Sweezy, which contained a fairly comprehensive review of the pre-war history of Marxist economics and at the same time made explanatory use of concepts introduced into mainstream monopoly and oligopoly theory during the preceding decade. This book, soon translated into several foreign languages, had a significant effect in systematizing the study and interpretation of Marxian economic theories.

It should not be supposed, however, that these new departures were altogether a matter of theoretical speculation. Of equal if not greater importance were the changes in the structure and functioning of capitalism which had emerged during the 1920s and 1930s. On the one hand the decline in competition which began in the late 19th century proceeded at an accelerated pace – as chronicled in the classic study by Arthur R. Burns, *The Decline of Competition: A Study of the Evolution of American Industry* (1936) – and on the other hand the unprecedented severity of the depression of the 1930s provided dramatic proof of the inadequacy of conventional business cycle theories. The Keynesian revolution was a partial answer to this challenge, but the renewed upsurge of the advanced capitalist economies during and after the war cut short further development of critical analysis among mainstream economists, and it was left to the Marxists to carry on along the lines that had been pioneered by Kalecki before the war. Kalecki spent the war years at the Oxford Institute of Statistics whose Director, A. L. Bowley, had brought together a distinguished group of scholars, most of them emigrés from occupied Europe. Among the latter was Josef Steindl, a young Austrian economist who came under the influence of Kalecki and followed in his footsteps. Later on, Steindl recounted the following:

> On one occasion I talked with Kalecki about the crisis of capitalism. We both, as well as most socialists, took it for granted that capitalism was threatened by a crisis of existence, and we regarded the stagnation of the 1930s as a symptom of such a major crisis. But Kalecki found the reasons, given by Marx, why such a crisis should develop, unconvincing; at the same time he did not have an explanation of his own. I still do not know, he said, why there should be a crisis of capitalism, and he added: Could it have anything to do with monopoly? He subsequently suggested to me and to the Institute, before he left England, that I should work on this problem. It was a very Marxian problem, but my methods of dealing with it were Kaleckian (Steindl, 1985).

Steindl's work on this subject was completed in 1949 and published in 1952 under the title *Maturity and Stagnation in American Capitalism*. While little noticed by the economics profession at the time of its publication, this book nevertheless provided a crucial link between the experiences, empirical as well as theoretical, of the 1930s, and the development of a relatively rounded theory of monopoly capitalism in

the 1950s and 1960s, a process which received renewed impetus from the return of stagnation to American (and global) capitalism during the 1970s and 1980s.

The next major work in the direct line from Marx through Kalecki and Steindl was Paul Baran's book, *The Political Economy of Growth*, which presented a theory of the dynamics of monopoly capitalism and opened up a new perspective on the nature of the interaction between developed and underdeveloped capitalist societies. This was followed by the joint work of Baran and Sweezy, *Monopoly Capital: An Essay on the American Economic and Social Order*, incorporating ideas from both of their earlier works and attempting to elucidate, in the words of their Introduction, the ‘mechanism linking the foundation of society (under monopoly capitalism) with what Marxists call its political, cultural, and ideological superstructure’. Their effort, however, still fell short of a comprehensive theory of monopoly capitalism since it neglected ‘a subject which occupies a central place in Marx's study of capitalism’, that is, a systematic inquiry into ‘the consequences which the particular kinds of technological change characteristic of the monopoly capitalist period have had for the nature of work, the composition (and differentiation) of the working class, the psychology of workers, the forms of working-class organization and struggle, and so on.’ A pioneering effort to fill this gap in the theory of monopoly capitalism was taken by Harry Braverman a few years later (Braverman, 1974) which in turn did much to stimulate renewed research into changing trends in work processes and labour relations in the late 20th century.

Marx wrote in the Preface to the first edition of Volume 1 of *Capital* that ‘it is the ultimate aim of this work to lay bare the economic law of motion of modern society’. What emerged, running like a red thread through the whole work, could perhaps better be called a theory of the accumulation of capital. In what respect, if at all, can it be said that latter-day theories of monopoly capitalism modify or add to Marx's analysis of the accumulation process?

As far as form is concerned, the theory remains basically unchanged, and modifications in content are in the direction of putting even greater emphasis on certain tendencies already demonstrated by Marx to be inherent in the accumulation process. This is true of concentration and centralization, and even more spectacularly so of the role of what Marx called the credit system, now grown to monstrous proportions compared to the small beginnings of his day. In addition, and perhaps most important, the new theories seek to demonstrate that monopoly capitalism is more prone than its competitive predecessor to generating unsustainable rates of accumulation, leading to crises, depressions, and prolonged periods of stagnation.

The reasoning here follows a line of thought which recurs in Marx's writings, especially in the unfinished later volumes of *Capital* (including *Theories of Surplus Value*): individual capitalists always strive to increase their accumulation to the maximum extent possible and without regard for the ultimate overall effect on the demand for the increasing output of the economy's expanding capacity to produce. Marx summed this up in the well-known formula that ‘the real barrier of capitalist production is capital itself’. The upshot of the new theories is that the widespread introduction of monopoly raises this barrier still higher. It does this in three ways.

1. (1) Monopolistic organization gives capital an advantage in its struggle with labour, hence tends to raise the rate of surplus value and to make possible a higher rate of accumulation.
2. (2) With monopoly (or oligopoly) prices replacing competitive prices, a uniform rate of profit gives way to a hierarchy of profit rates – highest in the most concentrated industries, lowest in the

most competitive. This means that the distribution of surplus value is skewed in favour of the larger units of capital which characteristically accumulate a greater proportion of their profits than smaller units of capital, once again making possible a higher rate of accumulation.

3. (3) On the demand side of the accumulation equation, monopolistic industries adopt a policy of slowing down and carefully regulating the expansion of productive capacity in order to maintain their higher rates of profit.

Translated into the language of Keynesian macro theory, these consequences of monopoly mean that the savings potential of the system is increased, while the opportunities for profitable investment are reduced. Other things being equal, therefore, the level of income and employment under monopoly capitalism is lower than it would be in a more competitive environment.

To convert this insight into a dynamic theory, it is necessary to see monopolization (the concentration and centralization of capital) as an ongoing historical process. At the beginning of the transition from the competitive to the monopolistic stage, the accumulation process is only minimally affected. But with the passage of time the impact grows and tends sooner or later to become a crucial factor in the functioning of the system. This, according to monopoly capitalist theory, accounts for the prolonged stagnation of the 1930s as well as for the return of stagnation in the 1970s and 1980s following the exhaustion of the long boom caused by World War II and its multifaceted aftermath effects.

Neither mainstream economics nor traditional Marxian theory had been able to offer a satisfactory explanation of the stagnation phenomenon which has loomed increasingly large in the history of the capitalist world during the 20th century. It is thus the distinctive contribution of monopoly capitalist theory to have tackled this problem head on and in the process to have generated a rich body of literature which draws on and adds to the work of the great economic thinkers of the last 150 years. A representative sampling of this literature, together with editorial introductions and interpretations, is contained in Foster and Szlajfer (1984).

See Also

- Baran, Paul Alexander
- capitalism
- finance
- foreign direct investment
- Marx, Karl Heinrich

Bibliography

Baran, P.A. 1957. *The Political Economy of Growth*. New York: Monthly Review Press.

Baran, P.A. and Sweezy, P.M. 1966. *Monopoly Capital: An Essay on the American Economic and Social Order*. New York: Monthly Review Press.

Braverman, H. 1974. *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. New York: Monthly Review Press.

Burns, A.R. 1936. *The Decline of Competition: A Study of the Evolution of American Industry*. New York: McGraw-Hill.

Foster, J.B. and Szlajfer, H., eds. 1984. *The Faltering Economy: The Problem of Accumulation Under Monopoly Capitalism*. New York: Monthly Review Press.

Hilferding, R. 1910. *Das Finanzkapital*. Trans. M. Watnick and S. Gordon as *Finance Capital*, ed. T. Bottomore, London: Routledge & Kegan Paul, 1981.

Kalecki, M. 1938. The distribution of the national income. *Econometrica*, April.

Lenin, V.I. 1917. *Imperialism, The Highest Stage of Capitalism*.

Marx, K. 1867. *Capital*, Vol. 1. Moscow: Progress Publishers.

Marx, K. 1885. *Capital*, Vol. 2. Moscow: Progress Publishers.

Marx, K. 1894. *Capital*, Vol. 3. Moscow: Progress Publishers.

Steindl, J. 1952. *Maturity and Stagnation in American Capitalism*. Oxford: Blackwell.

Steindl, J. 1985. The present state of economics. *Monthly Review*, February.

Sweezy, P.M. 1942. *The Theory of Capitalist Development*. New York: Monthly Review Press.

Sweezy, P.M. 1966. See Baran and Sweezy (1966).

Veblen, T. 1904. *The Theory of Business Enterprise*. New York: Charles Scribner's Sons.

How to cite this article

Sweezy, Paul M. "monopoly capitalism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000253> doi:10.1057/9780230226203.1135

The New Palgrave Dictionary of Economics Online

monopoly

Edwin G. West
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Irving Fisher (1923), once defined monopoly simply as an ‘absence of competition’. From this point of view various attitudes to, or criticisms of, monopoly are connected with the particular vision of competition that each writer has in mind. To the neoclassical economist monopoly is the polar opposite to the now familiar ‘perfect competition’ of the textbooks. Modern writers in the classical tradition, on the other hand, complain that perfect competition neglects the *process* of competitive activity, overlooks the importance of time to competitive processes and assumes away transaction or information costs.

In effect, ‘perfect competition’ to the neoclassical implies perfect decentralization wherein exchange costs happen to be zero. But the modern critics insist that exchange is not costless. And for this reason competition can be consistent with a wide variety of institutions that are employed to accommodate time, uncertainty and the costs of transacting (Demsetz, 1982). Such arrangements include, for example, tie-in sales, vertical integration and manufacturer-sponsored resale price maintenance. Such price-making behaviour means that in the real world decentralization is imperfect. And it is imperfect decentralization that is embodied in the classical paradigm of laissez faire. Consequently many phenomena that are automatically treated by the neoclassical as the absence of perfect competition or the presence of behaviour that *looks* monopolistic, are often viewed approvingly by those in the classical tradition.

It is widely believed that, historically, Adam Smith's *Wealth of Nations* provided the most sustained and devastating attack on monopoly. It is true that he speaks of ‘monopoly’ quite frequently, but typically he uses the term in a wide 18th-century sense to include all kinds of political restrictions. Monopoly under the modern meaning of a single uncontested firm was not Smith's usual target. He employed the term most often to refer to multi-firm industries enjoying statutory protection. Thus, ‘the law gave a monopoly to our boot-makers and shoe-makers, not only against our graziers, but against our tanners’ (Smith [1776], 1960, vol. 2, p. 153). Again, the whole system of mercantilism was condemned as monopolistic: ‘Monopoly of one kind or another, indeed, seems to be the sole engine of the mercantile system’ (ibid., p. 129).

The Ricardians too were more concerned with general restrictions, and especially with the fixed supply of land. Ricardo's *Principles of Political Economy and Taxation* in fact has only five pages out of 292 that discuss monopoly, while John Stuart Mill's *Principles of Political Economy* has only two out of 1,004. Following the Ricardians, the development of Darwinian philosophy in the mid-19th century only served to reinforce the classical emphasis on the necessity, if not inevitability, of competition. It is true that the ‘modern’ and more rigorous theory of monopoly, showing equilibrium to be determined by the equality of marginal revenue with marginal cost, was introduced by Cournot in 1838. But it received very little attention until much later.

In America the classical laissez-faire view of competition and imperfect decentralization prevailed at least to the end of the 19th century. When the Sherman Antitrust Act was passed in 1890, economists were almost unanimously opposed to it. Thus, despite his general disposition for widespread government intervention, the founder of the American Economic Association, Richard T. Ely (1900), firmly rejected the politically popular policy of ‘trust busting’. In the late 1880s John Bates Clark similarly feared that antitrust laws would involve a loss of the efficiency advantages of combinations or trusts. Combination itself was often necessary to generate adequate capital and to insure against adversity during the depressing period of the business cycle. Other contemporary economists, including Simon N. Patten, David A. Wells and George Gunton, had similar views. The last argued that the concentration of capital does not drive small producers out of business, ‘but simply integrates them into a larger and more complex system of production, in which they are enabled to produce wealth more cheaply for the community and obtain a larger income for themselves’. Instead of the concentration of capital tending to destroy competition, the reverse was true: ‘By the use of large capital, improved machinery and better facilities, the trust can and does undersell the corporation’ (Gunton, 1888, p. 385).

Consider now, and in contrast, the subsequent neoclassical approach which eventually involved the comparison of monopoly with what is said to be its polar opposite market structure of perfect competition. The method was gradually developed from the last part of the 19th century and ultimately, in the 1950s, reached the stage of empirical measurement of what was described as the social cost of monopoly. The most influential study has been that of Harberger (1954), whose basic argument can be summarized in terms of Figure 1.

Figure 1

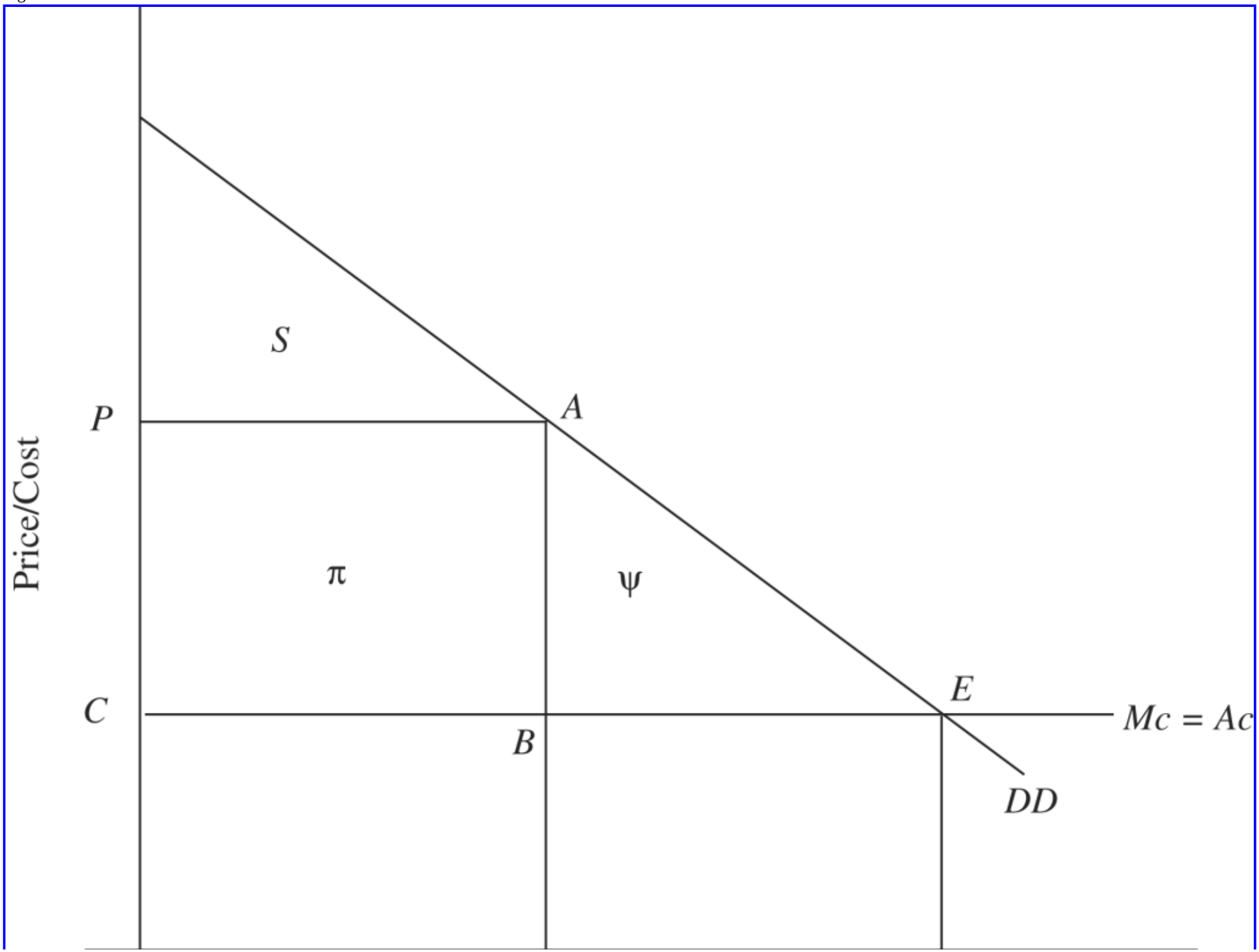

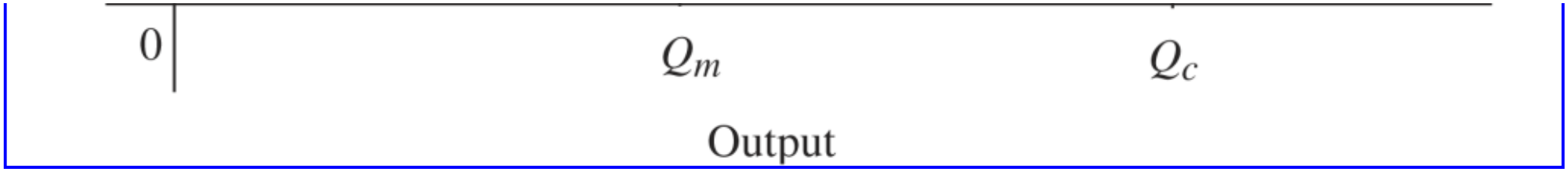

Assume that long-run average costs are constant for both firm and industry and are represented by the line M_c=A_c. The perfectly competitive output would be at Q_c where M_c intersects the demand curve *DD*. If a monopolist were substituted, he could maximize profits by producing Q_m at price *P*. His monopoly profit, π, would be represented by the rectangle *ABCP*. The loss of consumers' surplus is measured by the trapezoid *AECP*. The part of this area represented by *ABCP*, however, is not destroyed welfare but simply a transfer of wealth from consumers to the monopolist. The net loss to society as a whole from the monopoly is given by the 'welfare triangle' *ABE*, denoted in Figure 1 by ω. After making some heroic assumptions, in particular that marginal cost (M_c) was constant for all industries and that the price elasticity of demand was unity everywhere, Harberger estimated an annual welfare loss of $59 million for the US manufacturing sector in the 1920s. This figure was surprisingly small since it represented only one-tenth of 1 per cent of the US national income for that period.

Subsequent writers have argued that Harberger's measure was a serious underestimate for statistical and other reasons. George Stigler (1956) objected that (1) monopolists normally produce in the range where elasticity is greater than unity; (2) some monopoly advantages become embodied in the accounted costs of assets, so leading to an underestimate in reported profits. Subsequent studies that allowed for Stigler's objections reported social costs of monopoly much higher than Harberger's. Thus D.R. Kamerschen (1966) reported an annual welfare loss due to monopoly in the 1956–61 period amounting to around 6 per cent of national income. D.A. Worcester, Jr. (1973), on the other hand, using *firm* rather than *industry* data, and assuming an elasticity of (minus) 2, reported a maximum estimate of welfare loss in the range of 0.5 per cent of national income for the period 1965–9. Focusing on the complaint that Harberger assumed the normal competitive profit rate to be represented by the actual average profit rate earned, whereas the latter itself contains a monopoly profit element, Cowling and Mueller (1978), reported that 734 large firms in the US generated welfare losses totalling $15 billion annually over the period 1963–6, and this amounted to 13 per cent of Gross Corporate Product. All such criticisms have obviously been of a technical nature and implicitly accept Harberger's basic methodology.

Consider next another type of qualification that also accepts the same central methodology. In the frictionless world of the neoclassical model, where all exchange costs are zero, it would be profitable for the monopolist to produce more than Q_m in Figure 1. This would be the case, for example, with the institution of a two-part tariff where a second price is charged for all purchases in excess of Q_m. If this price were located exactly halfway between *P* and *C*, it could be shown that the triangle of welfare loss would shrink to one-quarter of the existing size of ω. An extension of such multi-part pricing, of course, would reduce the welfare triangle of loss still further. With the presence of zero exchange costs, which pertains to the neoclassical world, perfect price discrimination is possible. In this case the whole of the trapezoid *CPAE* would consist of transferred wealth from consumers to producers. Deadweight welfare loss from monopoly would be zero.

If the neoclassical analyst objects that perfect price discrimination does not exist in the real world, he has to offer reasons. It is difficult, meanwhile, to conceive of any practical explanation that could be couched in terms of anything else but significant costs of exchange, such as positive information costs and risk. But such explanation undermines the 'purity' of the neoclassical model and points us back in the direction of the classical world of imperfect decentralization featuring real-world limitations on knowledge, and the existence of dynamic change under uncertainty.

It will be helpful now to describe classical analysis in terms of Figure 1. But first recall that, instead of the notion of perfect competition as a static long-term equilibrium, we start with the view of competition, espoused by Adam Smith and his successors, as a process of rivalry within a time dimension. In Schumpeter, for instance, competition is seen as 'a perennial gale of creative destruction'. It is the possibility of profit, of course, that drives the innovating entrepreneur. Without it the laissez-faire model of decentralization would collapse. But once profits are obtained by a successful pioneer his operation is immediately copied by others, so that there is a constant tendency for entrepreneurial profit to be competed away. It is this focus on a continual series of short runs that distinguishes the analysis from that of 'perfect competition', which is always expressed in terms of the very long run.

Assume then the discovery of a new product, product *X*, by an entrepreneur who proceeds to offer Q_m of it at price *P* (see Figure 1). It is only academically true that he is restricting output compared with what potential rivals would produce if they possessed his knowledge and business acumen. But since, in reality, they do not, the only alternative to Q_m supply of product *X* is some positive quantity of conventional products that the factors were previously producing (the supply of *X* being zero). The result of his activity in producing *X*, therefore, is pure social gain, and this is measured in Figure 1 by the profit plus the consumer surplus *S*. The welfare triangle of social loss (ω) does not exist. It can be expected that the entrepreneur's action will lead to the eventual entry of rivals. At this stage competition will lead to a lowering of price towards cost. This process will then involve a transfer of wealth from the original entrepreneur to consumers. But the latter's original and temporary profit is necessary to induce him to introduce the product at an earlier time than otherwise. It is this earlier introduction indeed that produces the social gains. So while such temporary profit may be described as proceeding from the market structure of 'imperfect

competition', nevertheless, according to the Smithian/Schumpeterian analysis, the monopolies so described are necessary institutions, since economic growth would be much weaker without them. Indeed, society recognizes such logic when it grants temporary legal monopolies in the form of patents.

It is necessary now to examine the special place that is usually accorded to the phenomenon of what is called 'natural monopoly'. This is said to exist when it is technically more efficient to have a single producer or enterprise. The ultimate survival of such a single firm is usually the natural outcome of initial rivalry between several competitors. J.S. Mill ([1848], 1965, p. 962) appears to have been the first to use the adjective 'natural' and to use it interchangeably with 'practical'. Examples quoted by Mill included gas supply, water supply, roads, canals and railways.

In his *Social Economics* (1914) Friedrich von Wieser was probably the first to distinguish the modern from the classical doctrine of monopoly. The classical (Marxian?) attribution to monopoly of the 'favoured' market position of capital over labour was incorrect. So was Ricardo's reference to the 'monopoly' of agricultural soil. The price of urban rents was a competitive price. A typical real monopoly for von Wieser consisted of what he called the 'single-unit enterprise', that was identical to the organization that Mill had previously identified as a 'natural monopoly'. The postal service was an excellent illustration:

> In the face of [such] single-unit administration, the principle of competition becomes utterly abortive. The parallel network of another postal organization, beside the one already functioning, would be economically absurd; enormous amounts of money for plant and management would have to be expended for no purpose whatever. [von Wieser [1914], 1967, pp. 216–17]

The conclusion was that some kind of government control such as price regulation was required.

One must conjecture that von Wieser would have been astonished by the application (in the 1980s) to natural monopolies of the new theory of 'the contestable market'. According to its promulgators, this is a situation in which 'entry is absolutely free, and exit is absolutely costless' (Baumol, 1982). To such economists, even von Wieser's postal service is, at least conceptually, open to such market contestability (although the main example quoted by the new analysts has been that of airlines). The essence of a contestable market is that it is vulnerable to hit-and-run entry: 'Even a very transient profit opportunity need not be neglected by a potential entrant, for he can go in, and before prices change, collect his gains and then depart without cost, should the climate grow hostile' (Baumol, 1982, p. 4).

In effect, such new analysis is a theoretical development of the neoclassical concern with perfect competition and especially with its condition of free entry. Indeed, one writer prefers the term 'ultra-free entry' to 'perfect contestability' (Shepherd, 1984). What is involved is not only the possibility of a new firm gaining a foothold (which is conventional 'free entry') but the ability to duplicate immediately and entirely replace the existing monopolist. The entrant can, moreover, establish itself before the existing firm makes any price response (the Bertrand–Nash assumption). Finally, exit is perfectly free and without cost. Sunk cost, in other words, is zero. Given these conditions, even the threat of entry (potential competition) may hold price down to cost. A government scheme of regulated prices might therefore be socially detrimental.

Although such theoretical innovation is challenging, it has given rise to considerable controversy concerning both the internal consistency of the theory and empirical support for it. The assumption of zero sunk costs has been the one that has come under most attack. It has been observed for instance that in most markets sunk costs are more obvious in the short run than in the long run; and this is by definition. With *any* element of sunk cost the existing firm has a proportionate potential pricing advantage over an entrant. But it is in the very short period that the pure contestability theory stipulates a zero-price response from the incumbent. Meanwhile, with respect to the question of the empirical basis for the theory, Baumol et al. concede that very little is available so far.

Doubts about the efficiency of government price regulation of natural monopolies have also been raised by Demsetz (1968). He has proposed that formal regulation is unnecessary where governments can allow 'rivalrous competitors' to bid for the exclusive rights to supply a good or service over a given 'contract period'. The appearance of a single firm may not imply monopoly pricing, because competition could have previously asserted itself at the franchise bidding stage. Monopoly *structure* therefore does not inevitably predict monopoly *behaviour*, although some element of the latter could appear if conditions, say of production, change during the period of the contract.

An ostensibly similar line of argument to that of Demsetz was offered by Bentham and Chadwick. Chadwick's investigation into water supply in London in the 1850s revealed circumstances of natural monopoly. But he argued that inefficiency was prevalent because the field was divided among 'seven separate companies and establishments of which six were originally competing within the field of supply, with two and three sets of pipes down many of the same streets' (quoted in Crain and Ekelund, 1976). Following Chadwick's recommendation, rivalry was channelled into what he called competition *for* the field and away from (costly) competition '*within* the field'. The same reasoning applied to the railways. Public ownership was advocated while management (operation) of the services was to be contracted out via a competitive franchise bidding process from among potential private enterprises.

It must next be recognized that very many monopolies, if not most, are *unnatural;* that is, they arise not from inexorable economic conditions but from man-made arrangements, usually through the exercise of political power. In these cases the monopoly is typically awarded by government but not usually with the intention of encouraging the introduction of a new product (as with patents). Instead, one supplier is granted the sole right of trading an existing product or service to the exclusion of all other suppliers. A natural state of competition is thus converted by fiat into one of (statutory) monopoly. In this case the classical analyst might see more potential relevance in Harberger's model of welfare loss from monopoly.

Where the monopoly right is granted by the government, and assuming that price discrimination is prohibitively costly, it would seem, again at first sight, that the monopoly rent or 'prize' to the successful producer could indeed be represented by a rectangle such as *ABCP* in Figure 1. But since the seminal writing of Tullock (1967), economists have come to recognize that the pursuit of such monopoly rents is itself a competitive activity, and one that consumes resources. Since Krueger (1974) this process has become known as 'rent seeking' and it frequently takes the form of lobbying, offering campaign contributions, bribery, and other ways of influencing the authorities to grant exclusive rights to production, rights that are then policed by the coercive powers of government.

Recent work has modified the conclusion that the value of resources used in pursuit of the rents would exactly equal the value of the rents. Some writers have urged that lobbying by consumers might to some extent offset that of potential monopolists such that a regulated price at a magnitude lower than *P* (but higher than *C*) in Figure 1 would result. In this case, of course, the expected rectangle of monopoly rent would be reduced and the producers collectively would not spend more than this in rent seeking.

Jadlow (1985) has reduced still further the expected magnitude of such monopoly rent rectangles by introducing a multi-period model wherein other rent seekers continue to compete for the valued monopoly prize while consumers, regulators and antitrusters continue their endeavours to eliminate the rents over a protracted period into the future. Since, therefore, instead of a one-time prize, the monopoly rent is viewed as the expected present value of a stream of rents over a series of future time periods in which uncertainty is present, there is likely to be a significant reduction of resources invested in rent-seeking activities.

It is usually implied by economists that the task of public policy with regard to monopoly is to eliminate monopoly profit by one means or another. The above analysis reveals, however, that the conventional measures of social losses via the welfare triangles, plus the rectangles of potential transfers that are partially 'eaten up' by resources devoted to rent-seeking, are predominantly applicable to monopolies that are politically bestowed. We are thus left with the conclusion that appropriate public policy (according to usual economic reasoning) involves government 'correcting for' something it has created itself. The direct way of solving such a problem, at least to the innocent, would be for the government simply to abstain from granting statutory monopoly privileges in the first place. The newer 'economics of politics', however, has produced reasons why the legislative activity of monopoly rent creation is inherent in the very structure of majority voting democracies. Indeed, some writers (Brennan and Buchanan, 1980) argue that the very institution of government is usually a monopoly. In so far as this is true, we face the paradoxical situation that the public policy prescribed in economics textbooks is one whereby monopoly in general is policed or controlled by an institution that is itself a monopoly.

The problem of government sponsored monopolies is currently receiving considerable attention. Indeed, it constitutes one of the most profound issues of the day. For hopeful developments we must look again, presumably, to still further research in the modern economics of politics.

See Also

- competition
- contestable markets
- market structure
- monopoly

Bibliography

Baumol, W.J. 1982. Contestable markets: an uprising in the theory of industry structure. *American Economic Review* 72(1), March, 1–15.

Brennan, G. and Buchanan, J. 1980. *The Power to Tax: Analytical Foundations of the Fiscal Constitution*. Cambridge: Cambridge University Press.

Cowling, K. and Mueller, D.C. 1978. The social costs of monopoly power. *Economic Journal* 88, December, 727–48.

Crain, W.M. and Ekelund, R.E., Jr. 1976. Chadwick and Demsetz on competition and regulation. *Journal of Law and Economics* 19(1), April, 149–62.

Demsetz, H. 1968. Why regulate utilities? *Journal of Law and Economics* 11, April, 55–65.

Demsetz, H. 1982. *Economic, Legal, and Political Dimensions of Competition, Professor Dr. F. de Vries Lectures in Economics*. Vol. 4, Amsterdam: North-Holland.

Ely, R.T. 1900. *Monopolies and Trusts*. New York: Macmillan.

Fisher, I. 1923. *Elementary Principles of Economics*. New York: Macmillan.

Gunton, G. 1888. The economic and social aspects of trusts. *Political Science Quarterly* 3(3), September, 385–408.

Harberger, A.C. 1954. Monopoly and resource allocation. *American Economic Association, Papers and Proceedings* 44, May, 77–87.

Jadlow, J.M. 1985. Monopoly rent-seeking under conditions of uncertainty. *Public Choice* 45(1), 73–87.

Kamerschen, D.R. 1966. An estimation of the ‘welfare losses’ from monopoly in the American economy. *Western Economic Journal* 4, Summer, 221–36.

Krueger, A.O. 1974. The political economy of the rent-seeking society. *American Economic Review* 64, June, 291–303.

Mill, J.S. 1848. *Principles of Political Economy*. Ed. W.J. Ashley. Reprinted, New York: A.M. Kelley, 1965.

Shepherd, W.G. 1984. ‘Contestability’ vs. competition. *American Economic Review* 74(2), September, 572–87.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 2 vols, ed. E. Cannan, London: Methuen, 1960.

Stigler, G. 1956. The statistics of monopoly and merger. *Journal of Political Economy* 64, February, 33–40.

Tullock, G. 1967. The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal* 5, June, 224–32.

von Wieser, F. 1914. *Social Economics*. Trans. A. Ford Hinrichs, New York: A.M. Kelley, 1967.

Worcester, D.A., Jr. 1973. New estimates of the welfare loss to monopoly, United States: 1956–69. *Southern Economic Journal* 40(2), October, 234–45.

How to cite this article

West, Edwin G. "monopoly." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000252>
doi:10.1057/9780230226203.1134

The New Palgrave Dictionary of Economics Online

monopsony

Alan Manning
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Monopsony refers to the situation where a firm has some market power over the price it pays for its inputs, so that a higher price must be paid the more input is used. Monopsony could exist in any input market but is usually discussed in the context of the labour market. Employers will have monopsony power over their workers because of frictions in the labour market. Employers will use this monopsony power to pay workers less than their marginal product. This gap between marginal product and wage offers policy an opportunity to raise the wage of workers without necessarily jeopardizing their employment.

Keywords

collusion; frictions; human capital; job mobility; labour markets; labour supply; law of one wage; minimum wage; monopsony; oligopsony; partial equilibrium; Robinson, J.; search capital; search models; Smith, A.; training; wage discrimination; wage dispersion; wage heterogeneity, sources of

Article

The definition of a monopsony in the *Oxford English Dictionary* (OED) is ‘a market situation in which there is only one buyer’. Joan Robinson (1933) is credited with inventing the term (but see Thornton, 2004, for a discussion of the origins of the term) as a counterpart to the more commonly used and understood term ‘monopoly’.

Taken literally, it is very likely that a pure monopsony has never existed in any market, but the term is more generally used to denote a situation in which the supply curve to an individual firm has an input price elasticity that is finite, that is, is increasing in the input price, and this article follows that usage. If one is pedantic, one might think that ‘oligopsony’ is a more accurate term to use (defined by the OED as ‘a state of the market in which only a small number of buyers exists for a product’), or ‘oligopsonistic competition’ if one believes that free entry of firms will bid away any monopsony rents.

The market for any type of good or service could, in principle, be monopsonistic. To give some examples from the economic literature, Schroeter (1988) considers the meat-packing industry as an oligopsonistic buyer of cattle, Just and Chern (1980) consider the tomato-canning industry as an oligopsonistic buyer of tomatoes, and Murray (1995) considers saw-mills as oligopsonistic buyers of logs. But the idea of monopsony is most commonly applied to the labour market, and this article focuses on that application. Employers are often felt to have monopsony power only in a few specific labour markets – those for professional athletes in the United States, nurses and teachers (for whom outside cities there may only be one potential employer), and miners and mill workers in company towns in the early days of the Industrial Revolution are some of the more common examples. But, in recent years, some labour economists have argued that monopsony is pervasive in all labour markets.

The plan of this article is the following. We first review the simple partial equilibrium of monopsony, discussing the differences from and similarities to the more conventional perfectly competitive model. We then discuss why it is plausible to believe that employers have some monopsony power over their workers, after which we discuss how the monopsony perspective can help us to a better understanding of the workings of labour markets. The monopsonistic approach is more in line with the way that workers and employers experience the labour market, and can explain a wide range of what are puzzles and anomalies from the perspective of labour markets as perfectly competitive. Many of these puzzles and anomalies have other potential explanations but monopsony offers a simple unified account of their existence.

The simple textbook model of monopsony

In a perfectly competitive labour market, an employer can hire as many workers of a particular type as it wants at the market wage for that type of workers (and none at all if it tries to pay below the market wage). But, if an employer has some Monopsony power, the labour supply to an individual employer depends positively on the wage paid. The wage elasticity of the labour supply curve facing the firm is therefore finite not infinite. Figure 1 represents such a labour supply curve.

Figure 1

The textbook model of monopsony

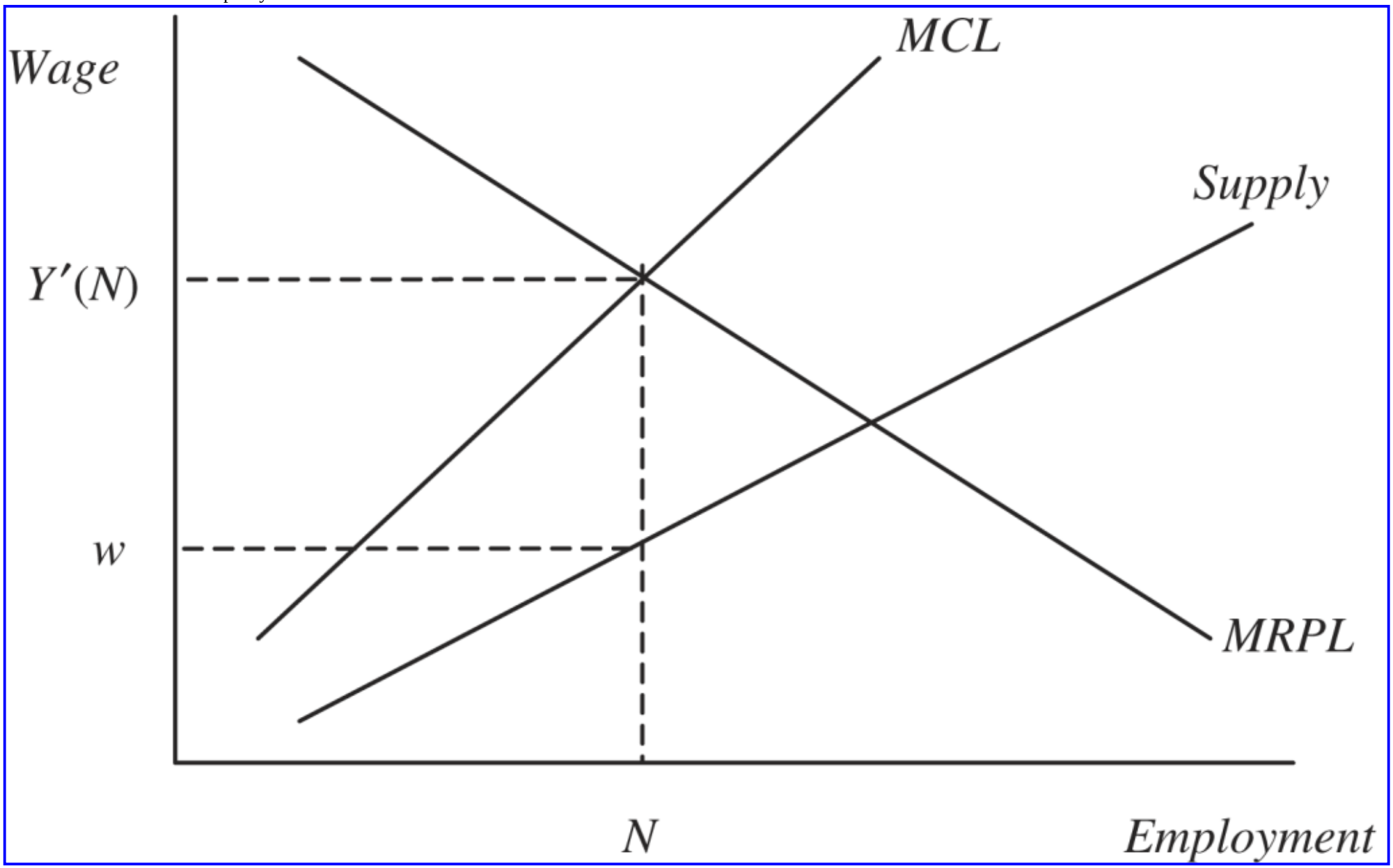

How does this affect the decisions of employers? Denote the supply of labour to the firm if it pays w by $N(w)$ and the inverse of this relationship by $w(N)$. Total labour costs are given by $w(N)N$. Assume that the firm has a revenue function $Y(N)$ and is a simple monopsonist who has to pay a single wage to all its workers. It wants to choose N to maximize profits

which are given by:

$$\pi = Y(N) - w(N)N. \quad (1)$$

This leads to the first-order condition:

$$Y'(N) = w(N) + w'(N)N. \quad (2)$$

The left-hand side of (2) is the marginal revenue product of labour. The right-hand side is the marginal cost of labour, the increase in total labour costs when an extra worker is hired. The marginal cost of labour (MCL) has two parts: the wage, *w*, that must be paid to the new worker hired and the increase in wages that must be paid to all existing workers. The MCL is always above the labour supply curve to the firm and is also drawn on Figure 1. The profit maximizing employer will choose the level of employment where MRPL = MCL and the wage necessary to supply this amount of labour – the solution is represented graphically in Figure 1.

In equilibrium, the wage paid to workers is less than their marginal revenue product. Although the employer is making positive profit on the marginal worker they have no incentive to increase employment because doing so would require increasing the wage (to attract the extra worker) and this higher wage must be paid not just to the new worker but also to all the existing workers. One particularly useful way of representing the choice of the firm is that marginal cost of labour is a mark-up on the wage, the mark-up being given by the elasticity of the labour supply curve facing the firm. Write the elasticity of the labour supply curve facing the firm as $\varepsilon_{Nw} = \frac{wN'(w)}{N(w)}$ and let ε be the inverse of this elasticity. Then (2) can be written as:

$$\frac{Y' - w}{w} = \frac{1}{\varepsilon_{Nw}} = \varepsilon \quad (3)$$

so that the proportional gap between the wage and the marginal revenue product is a function of the elasticity of the labour supply curve facing the firm. Perfect competition can be thought of as a special case of this model where $\varepsilon_{Nw} = \infty$ and $\varepsilon = 0$, in which case (3) says that the wage will be equal to the marginal revenue product.

Some of the comparative statics of the monopsony model are the same as the perfectly competitive model and some are different. For example, consider an increase in the marginal revenue product of labour for a single firm – this will lead to an increase in employment and a rise in wages in a monopsony model. The former would occur in a competitive model but the latter would not, as a competitive firm would simply continue to pay the market wage (which would not change if the change in the MRPL affected only a single firm). The impact of shifts in the labour supply curve to the firm is more complicated as the impact depends on how the change affects the marginal cost of labour and not just the average cost of labour. An increase in the supply of labour to the firm that keeps the elasticity the same will result in a rise in employment and a fall in wages, just as in the competitive model. But matters are more complicated if the elasticity of the labour supply curve can also change as the average and marginal cost of labour can move in opposite directions, the most familiar example of which is the impact of a minimum wage. The minimum wage raises the average cost of labour but (if it is binding) reduces $w'(N)$ so its effect on the marginal cost of labour (see (2)) is ambiguous. In fact, one can show that a minimum wage that just binds must raise employment (a demonstration of this can be found in most labour economics textbooks).

Although the model described here captures the fundamentals of a monopsonistic labour market, there are a number of ways in which it is too simplistic, and it is important to be aware of its limitations. First, we have assumed that the employer is a simple monopsonist who must pay the same wage to all workers – that is, wage discrimination is not allowed by

assumption.

Second, the simple model assumes that the only way an employer can raise employment is by raising the wage paid, something that is quite implausible. Manning (2006) considers the case where employers can also increase their employment by spending resources on recruitment activities. He shows that monopsony can be thought of as the case where the marginal cost of recruiting an extra worker is increasing in the number of workers recruited.

Third, the simple model is a model of partial equilibrium – it ignores the interactions with other employers that are very important in reality. One would expect the actions of other employers to affect the labour supply curve facing an individual firm; for example, if other firms pay higher wages we would expect the labour supply to this firm to fall for a given wage. Taking account of these interactions is particularly important when considering the impact of policies like the minimum wage that will affect all employers in a market. Manning (2003a, ch. 12) shows that, while in the simple monopsony model a just-binding minimum wage always raises employment, this is not necessarily the case in general equilibrium models of oligopsony, where there is more than one employer.

The sources of monopsony power

Labour economists have often doubted whether many employers have significant monopsony power over their employees (though this scepticism has diminished in recent years – see Boal and Ransom, 1997, for a generally sympathetic survey). So it is important to think about why employers are likely to have monopsony power over their workers.

Traditionally, employers are thought to have monopsony power only in labour markets in which there is a small number of employers. A typical example would be a mill town or mine village in the early days of industrialization, where the employer dominated the local labour market. Most economists are rightly sceptical of the view that the number of employers in many labour markets is small. Classical monopsony could also occur when there are many employers but they collude in wage-setting so that there are only a few effective employers in the labour market. But most economists do not think employer collusion is important in labour markets. (Yet Adam Smith, 1776, p. 169, strongly believed that employer collusion was a frequent outcome in labour markets:

> … we rarely hear, it has been said, of the combinations of masters, though frequently of those of workmen. But, whoever imagines, upon this account, that masters rarely combine, is as ignorant of the world as of the subject. Masters are always and everywhere in a sort of tacit, but constant and uniform combination, not to raise the wages of labour above their actual rate. To violate this combination is everywhere a most unpopular action, and a sort of reproach to a master among his neighbours and equals. We seldom, indeed hear of this combination, because it is the usual, and one may say, the natural state of things.)

However, modern theories of monopsony do not generally argue that employer market power over their workers derives from there being a small number of employers. They tend to emphasize the role of frictions in the labour market. The perfectly competitive model implies that an employer who cuts wages by one cent will find all their existing workers quit immediately. While it is likely that cutting wages will increase the quit rate and make it harder to recruit replacements, these effects are not as strong as the perfectly competitive model would have us believe.

To illustrate how this can lead to a model from the perspective of firms that looks something like Figure 1, suppose that the quit rate of workers is a negative function of the wage, $q(w)$ and the flow of recruits to the firm is a positive function of the wage, $R(w)$. Then, in steady state employment in the firm is:

$$N = \frac{R(w)}{q(w)} \tag{4}$$

which will be a positive function of the wage – that is, the employer will face an upward-sloping labour supply curve as represented in Figure 1.

What are the sources of these frictions in labour markets? In *The Economics of Imperfect Competition* (1933, p. 296), Joan Robinson argued that ignorance (about what all employers are offering), heterogeneous preferences and mobility costs are the most plausible sources of frictions in the labour market. The formal models of recent years are built on these ideas. Models based on worker ignorance are typically search models (the canonical versions of which are probably Albrecht and Axell, 1984, and Burdett and Mortensen, 1998) in which it takes time and/or money for workers to change jobs. One the other hand, there are the models that assume workers have full information and no mobility costs but that jobs are differentiated in some way (a canonical model of this sort is Bhaskar and To, 1999, though all such models have roots in the model of product differentiation by Salop, 1979). In these models, jobs might be differentiated by physical location or skill or any other plausible characteristic. This product differentiation gives employers some monopsony power over their

workers because employers are not perfect substitutes from the perspective of workers, so a cut in the wage does not cause all workers to leave for other firms.

These theories of 'modern monopsony' might appear to be very different to classical models of monopsony, but Manning (2003b) argues that they are more similar than one might have thought as they all use different mechanisms to argue that the choice of employers of a particular worker is limited at a particular moment in time.

It is plausible to think that labour markets have frictions; but is this any more than a complication? The next three sections argues that it does matter, emphasizing how our analysis of labour markets from the perspective of workers, employers and public policy is affected in important ways by the recognition that employers have monopsony power over their workers.

Monopsony from an employer perspective

Here the key idea of the monopsony model is that the labour supply curve facing an individual employer is not perfectly elastic. It is helpful to think about the decisions employers must make about pay, the structure of pay and non-wage aspects of jobs.

First, monopsonistic employers who want, for whatever reason, to be large will have to pay higher wages as they need to be further up their labour supply curve. Hence monopsony offers a simple explanation for the very robust empirical correlation between employer size and wages (see Brown and Medoff, 1989). It can also explain why wages seem to be positively correlated with measures of how 'good' an employer is like productivity and profitability (for example, Blanchflower, Oswald and Sanfey, 1996). As noted in the previous section, 'good' firms that have a higher MRPL curve will choose to pay higher wages, something that should not happen in a perfectly competitive labour market.

We also have robust evidence that low-wage employers find it harder to recruit and retain workers, as predicted by monopsony. Low-wage employers have higher vacancy rates, take longer to fill vacancies and have higher quit rates among their workers.

As already mentioned, employers have an incentive to wage discriminate, to pay different wages to workers who might have the same level of productivity. In particular, we would expect them to pay wages that rise with seniority, since pushing the rewards of employment into the future helps to deter quits as workers get the high wages only if they remain with the firm (see Stevens, 2004). This is consistent with the empirical evidence (admittedly a bit patchy) that pay varies more strongly than productivity with seniority, though there are also incentive theories that make similar predictions.

Monopsony also offers a simple explanation of why employers often seem to pay for general training of their workers. In a perfectly competitive market this is something of a puzzle; since workers should receive all the returns to general training, employers should not be prepared to pay for it. But in a monopsonistic labour market, where wages are below marginal products, some of the returns to general training are likely to accrue to employers, giving them an incentive to provide training.

Monopsony from a worker perspective

From the perspective of workers, a monopsonistic labour market will appear to be one in which there is heterogeneity in the jobs available (definitely in the wage but quite likely in other dimensions as well) and jobs are hard to find, so getting and losing jobs are occasions for joy and sadness. If one wants a formal model to capture these ideas, a search model is the right conceptual framework to use. Of course, one can use search models to think about workers' choices whenever they face a distribution of wages even if the origin of that distribution is not the monopsony power of employers, so this area of research is not distinctive to monopsony.

First, it can explain the existence of wage dispersion even in very tightly defined labour markets. This violation of the 'law of one wage' was first documented by the so-called neo-realist labour economists (see Kaufman, 1988) in the United States in the 1940s, but most subsequent studies have confirmed it (for example, Groshen, 1991). This wage dispersion is exactly what we would expect to see in a monopsonistic labour market in which different employers will choose different wages even if faced with the same labour supply curve. This can then help to explain why high-wage workers are, other things equal, less likely to quit and less likely to be looking for another job as these workers have been lucky enough to find themselves in one of the good jobs in their segment of the labour market.

Second, monopsony can explain part of the rapid growth in earnings over the early stages of the life cycle (as first identified by Mincer, 1974). The human capital explanation of this is that workers are accumulating skills but monopsony/search suggests that workers are working themselves into the best jobs in the market (what might be called the accumulation of search capital, the knowledge of which employers pay higher wages). Consistent with this, Topel and Ward (1992) find that one-third of the wage growth of young men in the US labour market is the result of job mobility.

Third, monopsony can explain the earnings losses suffered by displaced workers. It is well-documented that workers who lose their jobs through no fault of their own (for example, through plant closure) tend to suffer losses in earnings (see Kletzer, 1998, for a review) and the losses do not completely fit the pattern suggested by human capital theory – in particular, older workers suffer greater losses, even when we control for job tenure.

Monopsony can also explain systematic wage differentials between workers, even if they do not differ in their productivity. For example, if women are less attached to market employment or their decisions on which jobs to take are less motivated by money (Manning, 2003a, provides evidence on both these points), then women will earn less than men even

if the wage offer distribution they face is the same. The reason is that women will find it harder to accumulate search capital. There may also be incentives for employers to then pay lower wages to women, giving a further twist to their earnings disadvantage. Ransom and Oaxaca (2005) provides some evidence that the quit rate for women is less sensitive to the wage than is the quit rate for men.

Monopsony also has implications for the incentives to acquire human capital. Because the wage is below the marginal product, it is quite likely that some of the returns to investments in human capital accrue to future employers of the worker, though the interests of these employers are not internalized in the education decision. Hence, the social return to education is likely to exceed the private return, leading, in a free market, to underinvestment.

Monopsony from a public policy perspective

Thinking of labour markets as pervasively monopsonistic rather than perfectly competitive has implications for how one thinks about the likely effects of interventions in the labour market. In a perfectly competitive labour market one tends to think of the free market outcome as efficient, of any intervention as causing some inefficiency and justifiable only on equity grounds, especially if the equity effect is large and/or the efficiency cost is small. In contrast, if the labour market is monopsonistic, then there is no presumption that the free market is efficient and interventions might be justifiable on efficiency grounds alone. Based on the simple textbook model of a monopsonist presented earlier, one might be tempted to go further and argue that, because wages are below marginal products, interventions to raise wages must, over some range, improve efficiency. However, in more sophisticated models of monopsony or models of oligopsonistic competition, such a simple conclusion is not necessarily valid. So the monopsonistic approach does suggest approaching the analysis of the impact of interventions with a more open mind than a true believer in perfect competition might be inclined to do.

A good example is the employment impact of the minimum wage. If the labour market is perfectly competitive, one can prove with nothing more than pencil and paper that the minimum wage must reduce employment, and the only purpose of empirical analysis is to decide on how large the reduction is. However, a monopsony approach suggests going to the data with a less certain view about the ‘right’ answer. The intuition is that, while a rise in the minimum wage reduces the profitability of employing workers for firms, it increases the incentives for workers to work, and the net effect on employment depends on whether the ‘demand’ or ‘supply’ effect is dominant. Hence monopsony can explain why the empirical literature often fails to find evidence that it reduces employment (Card and Krueger, 1995; Dickens, Machin and Manning, 1999).

Another good example of apparently ‘perverse’ employment effects can be found in the impact of equal pay legislation. In the UK, this legislation led to a big increase in the pay of women relative to that of men but did not, as the perfectly competitive model would predict, lead to big falls in the relative employment of women (Manning, 1996).

Conclusions

There are good reasons to believe that employers have some monopsony power over their workers. Assuming labour markets are monopsonistic also brings the thinking of labour economists in line with the way in which agents perceive the workings of labour markets. Workers do not perceive labour markets as frictionless and changing; getting and losing jobs are routinely reported as major life events. And employers perceive they have discretion over the wages paid, as a reading of any human resource management textbook can confirm. And, as demonstrated in this article, a whole range of puzzles and anomalies melt away once one adopts the monopsony perspective. However, the impact of regulations is more ambiguous than in perfectly competitive markets, and the theoretical perspective should go hand-in-hand with an open-minded empirical approach.

There is much work yet to be done. For example, the size of the wage elasticity of the labour supply curve to an individual firm is very much unknown. The literature on the subject is small and not entirely convincing. The best estimates we do have (probably from Staiger, Spetz and Phibbs, 1999; Falch, 2001; Clotfelter et al., 2006) suggest quite a low wage elasticity, with the implication that employers do have significant monopsony power.

See Also

- labour market search
- minimum wages
- Robinson, Joan Violet
- wage inequality, changes in

Bibliography

Albrecht, J. and Axell, B. 1984. An equilibrium model of search unemployment. *Journal of Political Economy* 92, 824–40.

Bhaskar, V. and To, T. 1999. Minimum wages for Ronald McDonald monopsonies: a theory of monopsonistic competition. *Economic Journal* 109, 190–203.

Blanchflower, D., Oswald, A. and Sanfey, P. 1996. Wages, profits, and rent-sharing. *Quarterly Journal of Economics* 111, 227–51.

Boal, W. and Ransom, M. 1997. Monopsony in the labor market. *Journal of Economic Literature* 35, 86–112.

Brown, C. and Medoff, J. 1989. The employer size–wage effect. *Journal of Political Economy* 97, 1027–59.

Burdett, K. and Mortensen, D. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39, 257–73.

Card, D. and Krueger, A. 1995. *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton: Princeton University Press.

Clotfelter, C., Glennie, E., Ladd, H. and Vigdor, J. 2006. Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. Working Paper No. 12285. Cambridge, MA: NBER.

Dickens, R., Machin, S. and Manning, A. 1999. The effects of minimum wages on employment: theory and evidence from Britain. *Journal of Labor Economics* 17, 1–22.

Falch, T. 2001. Estimating the elasticity of labor supply utilizing a quasi–natural experiment. Working Paper, Norwegian University of Science and Technology, Trondheim.

Groshen, E. 1991. Sources of intra-industry wage dispersion: how much do employers matter? *Quarterly Journal of Economics* 106, 869–84.

Just, R. and Chern, W. 1980. Tomatoes, technology, and oligopsony. *Bell Journal of Economics* 11, 584–602.

Kaufman, B. 1988. *How Labor Markets Work*. Lexington, MA: Lexington Books.

Kletzer, L. 1998. Job displacement. *Journal of Economic Perspectives* 12(1), 115–36.

Manning, A. 1996. The Equal Pay Act as an experiment to test theories of the labour market. *Economica* 63, 191–212.

Manning, A. 2003a. *Monopsony in Motion: Imperfect Competition in Labor Markets*. Princeton: Princeton University Press.

Manning, A. 2003b. The real thin theory: monopsony in modern labour markets. *Labour Economics* 10(2), 105–31.

Manning, A. 2006. A generalised model of monopsony. *Economic Journal* 116, 84–100.

Mincer, J. 1974. *Schooling, Experience and Earnings*. New York: NBER.

Murray, B. 1995. Measuring oligopsony power with shadow Prices: U.S. markets for pulpwood and sawlogs. *Review of Economics and Statistics* 77, 486–98.

Ransom, M. and Oaxaca, R. 2005. Sex Differences in Pay in a ‘New Monopsony’ Model of the Labor Market. Discussion Paper No. 1870. Bonn: IZA.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Salop, S. 1979. A model of the natural rate of unemployment. *American Economic Review* 69, 117–25.

Schroeter, J. 1988. Estimating the degree of market power in the beef packing industry. *Review of Economics and Statistics* 70, 158–62.

Smith, A. 1776. *The Wealth of Nations*. London: Penguin, 1986.

Staiger, D., Spetz, J. and Phibbs, C. 1999. Is there monopsony in the labor market? Evidence from a natural experiment. Working Paper No. 7258. Cambridge, MA: NBER.

Stevens, M. 2004. Wage-tenure contracts in a frictional labour market: strategies for recruitment and retention. *Review of Economic Studies* 71, 535–51.

Thornton, R. 2004. Retrospectives: how Joan Robinson and B. L. Hallward named Monopsony. *Journal of Economic Perspectives* 18(2), 257–61.

Topel, R. and Ward, M. 1992. Job mobility and the careers of young men. *Quarterly Journal of Economics* 107, 439–79.

How to cite this article

Manning, Alan. "monopsony." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000373>
doi:10.1057/9780230226203.1136

The New Palgrave Dictionary of Economics Online

Moore, Henry Ludwell (1869 – 1958)

A.W. Coats

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agricultural economics; Moore, H. L.; Schultz, H.; statistics and economics

Article

An outstanding pioneer econometrician, Moore was a retiring, highly sensitive, intensely dedicated man, who devoted his whole life to the construction of ‘a statistical complement to economics’, as he termed it. He was born at Moore's Rest, Maryland, on 21 November 1869. After graduating from Randolph Macon College in 1892, he studied under Carl Menger in Vienna, and Simon Newcomb and John Bates Clark at Johns Hopkins, where in 1896 he completed his Ph.D. dissertation on von Thünen's theory of the natural wage. Following a year's instructorship at Hopkins, and five years at Smith College, Moore taught at Columbia, mainly mathematical economics and statistics, from 1902 to 1929. Essentially a researcher rather than a pedagogue, he attended Karl Pearson's courses on mathematical statistics and correlation in London, in 1909 and 1913, and for several years took a voluntary salary reduction in order to avoid undergraduate teaching. Ill health forced his early retirement.

In a series of powerful and highly original volumes Moore endeavoured, among other things, to verify the marginal productivity of wages, render the Walrasian system statistically operational, and reveal the fundamental law and cause of cycles – wherein he concluded that ‘the law of the cycles of rainfall is the law of the cycles of the crops and the law of Economic Cycles’ (1914, p. 135). Needless to say, this immensely ambitious undertaking was often severely attacked by contemporaries and subsequent commentators who exposed the data deficiencies, lax hypotheses, unavoidably heroic oversimplifications, and other shortcomings (cf. Stigler, 1965; 1968). Nevertheless, the strength and purity of Moore's scientific vision, and the careful and sophisticated statistical methods he employed, commanded respect and admiration.

Not surprisingly, Moore founded no school. Yet his principal disciple, Henry Schultz, was only one among the many economists who produced the 20th-century ‘avalanche of statistical demand curves’ (Schumpeter, 1954, p. 213) inspired by Moore, whose researches exerted a major impact on agricultural economics. Thus Moore may be credited in part with the high scientific standing American

agricultural economics now enjoys (Leontief, 1971). However, despite his seminal efforts to develop empirical estimates of theoretical economic relationships, Moore's achievements have been insufficiently acknowledged, partly, no doubt, because he was unwilling to propagandize his methods among his fellow professionals.

Selected works

1908. The statistical complement of pure economics. *Quarterly Journal of Economics* 23, November, 1–33.

1911. *Laws of Wages: An Essay in Statistical Economics*: New York: Macmillan Co.

1914. *Economic Cycles: Their Law and Cause*. New York: Macmillan. (Japanese translation, Tokyo, 1926).

1917. *Forecasting the Yield and Price of Cotton*. New York: Macmillan.

1923. *Generating Economic Cycles*. New York: Macmillan.

1929. *Synthetic Economics*. New York: Macmillan.

Bibliography

Leontief, W. 1971. Theoretical assumptions and non-observed facts. *American Economic Review* 61, 1–7.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Stigler, G.J. 1965. Henry L. Moore and statistical analysis. In *Essays in the History of Economics*, ed. G. J. Stigler. Chicago: University of Chicago Press.

Stigler, G.J. 1968. Moore, H.L. In *International Encyclopedia of the Social Sciences*, vol. 10, ed. David, L. Sills. New York: Macmillan and Free Press.

How to cite this article

Coats, A. W. "Moore, Henry Ludwell (1869–1958)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000258> doi:10.1057/9780230226203.1137

The New Palgrave Dictionary of Economics Online

moral hazard

Y. Kotowitz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adverse selection; agency theory; assignment problems; asymmetric information; contingent contracts; contract enforcement; exclusive contracts; externalities; hidden actions; incomplete contracts; incomplete information; insurance; joint production; law of large numbers; licensing; monitoring; moral hazard; noise; non-cooperative game theory; non-market institutions; opportunistic behaviour; optimal contracts; rationing; reputation; risk aversion; risk sharing; spot markets; transaction price; vertical integration

Article

The problem of moral hazard is pervasive in economic activities. Economists have been well aware of its existence as the following quote from the *Wealth of Nations* will testify:

> The directors of such companies, however, being the managers rather of other peoples' money than of their own, it cannot well be expected, that they should watch over it with the same anxious vigilance with which the partners in a private copartnery frequently watch over their own … Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company. (Smith, 1776, p. 700)

However, theoretical developments and their application to specific problems have only proceeded since the 1960s and are still the subject of vigorous research. While we have a considerable understanding of the problem, we do not as yet understand fully market and social responses to it. In the following I shall attempt to explain the nature of the problem and selectively illustrate the flavour of current theoretical developments.

Moral hazard may be defined as actions of economic agents in maximizing their own utility to the detriment of others, in situations where they do not bear the full consequences or, equivalently, do not enjoy the full benefits of their actions *due to uncertainty and incomplete information or restricted*

contracts which prevent the assignment of *full* damages (benefits) to the agent responsible. It is immediately apparent that this definition includes a wide variety of externalities, and thus may lead to nonexistence of equilibria or to inefficiencies of equilibria when they exist.

It is a special form of incompleteness of contracts which creates the conflict between the agent's utility and that of others. Such incompleteness may arise due to several reasons: the coexistence of unequal information and risk aversion or joint production, costs and legal barriers to contracting and costs of contract enforcement. We shall analyse each in turn.

Unequal information

Agents may possess exclusive information. Arrow (1985) classifies such informational advantages as ‘hidden action’ and ‘hidden information’. The first involves actions which cannot be accurately observed or inferred by others. It is therefore impossible to condition contracts on these actions. The second involves states of nature about which the agent has some, possibly incomplete information, information which determines the appropriateness of the agent's actions, but which are imperfectly observable by others. Thus, even if agents’ actions are costlessly observable by others, they do not know with certainty whether the actions were in their interest.

Commonly analysed examples of hidden actions are: workers’ effort, which cannot be costlessly monitored by employers, precautions taken by insured to reduce the probability of accidents or damages due to them, which cannot be costlessly monitored by insurers. Criminal activity clearly belongs in this category as well.

Examples of hidden information are expert services – such as physicians, lawyers, repairmen, managers and politicians.

Where consequences of specific agents’ actions can be separated from those of others, even though the consequences may be affected by random, unobservable states of nature, the problem may be easily solved if agents are risk neutral, by simply assigning the full consequences to the agent, in exchange for a fixed fee. This is in effect a complete contract. The problem of contract incompleteness arises when agents are risk averse or where assignment of responsibility to one agent cannot be made.

When agents are risk averse, assigning full damages (benefits) to them assigns them all risk due to random states of nature. Risk-averse agents would like to purchase insurance against such risks. However, it is impossible for others to separate the consequences of agents’ actions from random elements which cannot be controlled by the agent. Insurance against the latter will inevitably insulate agents from the consequences of their own actions. The agent may, of course, offer to supply information about the unobserved actions or states – but such information cannot be credible.

Optimal contracts generally involve some degree of insurance and hence lead to a conflict between incentives and risk sharing. Most of the literature on moral hazard has concentrated on this case. We shall come back to it.

When precise assignment of responsibility to individual agents is impossible, full assignment of consequences to individual agents cannot be achieved. By definition, this is the case for crime, where the identity of the perpetrator is generally not known with certainty. The design of punishments and the interaction with enforcement activities to apprehend and convict criminals is treated extensively in the literature (see for example Becker, 1968).

Group production is another area where assignment may be impossible. Some forms of collective

punishment on the group as a whole when output falls short of a specified quota, with some allocation rule when output meets or exceeds the quota may serve to elicit the desired output (Holmstrom, 1982). However, the conditions under which this is possible are quite restricted.

Similar problems arise where quality of products is difficult to ascertain because they must be used jointly with another service or product, because their performance is affected by conditions and nature of use. For example drugs must be used in conjunction with physicians' services. Failure of the drug may result from its poor quality, from misdiagnosis by the physician (who may prescribe the wrong drug) or from failure to follow instructions by the patient. In the absence of these complications, it would be optimal for the manufacturer, who knows the quality of his product, to supply a guarantee of performance, in order to remove the incentive to supply lower quality. As well, the guarantee serves at least partly to insure risk averse consumers against random variations in the performance of the drug. Even if the manufacturer is risk-averse, his risk is mitigated by the 'law of large numbers', so it is optimal for him to act as insurer.

However, under the circumstances above such insurance creates a moral hazard problem for the physician and the patient, who may use insufficient care in diagnosis and use. Any risk sharing among the relevant parties therefore induces a moral hazard problem which cannot be avoided in the presence of private information, even if all parties are risk neutral.

Barriers to contracting

Incomplete contracts may also arise in the absence of private information due to costs of writing detailed contingent contracts. This problem is particularly severe in contracts involving complex transactions and long periods. When uncertainty about the future is great, the number and nature of eventualities to be considered is clearly very large. The cost of anticipating them and writing a contract which specifies or elicits desired actions may be very large. The cost of reaching agreement on the proper actions in each eventuality may well be prohibitive. If the probability of any event is small, and the cost of agreement high, it may pay to leave the contract vague and wait for the resolution of uncertainty before reaching agreement. Of course, this is precisely the case in spot market transactions. However, frequently decisions must be made prior to the resolution of uncertainty. For example, specialized investments in physical or human capital must be made by the parties before production and trading begin (Becker, 1964). The nature of the investment may well depend on the transaction price, which may in turn depend on information revealed after the investment is made. A limited agreement on investment and trading may be optimal, leaving transaction price to future negotiation. This, however, may lead to a moral hazard problem. Opportunistic behaviour in subsequent periods by one of the parties may lead to termination of trading or unfavourable contract terms, for the party which invested in specialized capital. Knowing that this may occur, the incentive to invest is reduced. The resulting inefficiency may well fall short of the costs of complete contracts. Williamson (1985) argues that such problems may give rise to vertical integration.

Contracts are too costly to write when transactions are infrequent and small. Most spot market transactions between retailers and consumers falls in this category. Blanket contracts offered by sellers in the form of 'money back guarantees' or exchange privileges may be substituted for explicit contingent contracts – but they are subject to moral hazard on the consumers' side. Alternatively the state legislates

fair trading laws which serve as generalized contracts.

Contracts are lacking altogether when transactions are random or involuntary. Accidental damages inflicted on one party by another as in a traffic accident are good examples. Here again, the law must form a generalized contract. It is obvious that such a law cannot possibly allow for all contingencies, so that it constitutes an incomplete contract, giving rise to moral hazard problems. The question of the design of liability rules has been extensively analysed in the law and economics literature (Posner, 1977).

Finally, contracts may be restricted by law or by limited financial resources of agents. For example, even if managers are risk neutral, their financial resources may be insufficient to become sole proprietors, without relying on outside capital. Shareholders and bondholders must then share in the risk – raising a moral hazard problem due to the informational advantages of managers. For an extensive analysis of these problems, see Jensen and Meckling (1976).

Similarly, when punishments are limited by law, moral hazard may not be resolved even where actions can be costlessly observed *ex post*. Thus, for example, bankruptcy and limited liability provisions insure borrowers against extremely unfavourable states of nature without limiting the gains from extremely favourable ones. This creates a moral hazard problem, inducing borrowers to undertake riskier projects. Stiglitz and Weiss (1981) show that lenders will sometimes require collateral and ration loans in attempting to overcome these difficulties.

Problems of enforcement

A related barrier to complete contracting arises from costs and other limitation on enforcement. When enforcement is costly, it may be more efficient to live with the inefficiencies generated by the moral hazard, than to try to enforce the optimal contingent contract. A common way to overcome such difficulties is by way of posting a bond, which is forfeit in the event of non-performance. However, restricted financial resources generally prevent bonding.

Under conditions where enforcement is not economical, contracts must be *self-enforcing*. It is unimportant whether contracts are explicit or implicit, as they frequently are in labour markets. To be viable contracts must make subsequent actions by contracting parties consistent with their self-interest, that is, they must allow for the potential exercise of moral hazard. This problem is at the heart of non-cooperative game theory, which defines moral hazard as opportunistic behaviour.

So far we have surveyed the conditions under which a moral hazard problem cannot be trivially resolved. This raises three questions which theorists have begun tackling in the past two decades: (*a*) the nature of optimal contracts in the presence of moral hazard; (*b*) market and institutional/legal response to mitigate these problems; and (*c*) welfare consequences.

Optimal contracts

The problem has mainly been tackled by agency theory. Following seminal work by Wilson (1969) and Ross (1973) the optimal (typically second best) reward structure for an agent is derived on the basis of observed variables, usually under ‘hidden action’ assumptions. Some of the main results for risk-averse agents are: (*a*) Optimal contracts require risk sharing between principal and agent which creates a moral hazard problem in the form of insufficient incentives. (*b*) Efficient contracts should utilize all the

information available, that is they should be constructed on the basis of statistical inference from the information available on the hidden action of the agent (Holmstrom, 1979). Thus monitoring, which reduces inference errors, is productive. (*c*) The nature of the reward schedule is sensitive to the nature of the information available, the residual uncertainty and the degree of risk aversion of the agent and principal. This observation is troubling because incentive contracts observed in reality are generally simple and uniform across a variety of agents and information sets. Long-term contracts, explicit or implicit (client relations), tend to mitigate moral hazard problems, by introducing a reward for not exploiting short-term informational advantages, and because cumulative information reduces uncertainty. Hence, for example, experience rating in insurance contracts.

Market and institutional responses

Market responses may invalidate or reinforce the special features of contracts to mitigate the moral hazard problem. These responses depend on the nature of competition. Free entry and the existence of unobserved differences among agents create the additional problem of adverse selection. We shall therefore reflect only on market responses which are mainly a consequence of moral hazard.

As indicated above, contracts typically require some risk sharing (coinsurance) between the parties when agents are risk averse. Therefore, agents generally bear more risk than they desire. If they are able to purchase additional insurance from third parties, the moral hazard problem is aggravated, making the original contract inefficient. This requires exclusivity in contracting. Thus for example, insurance companies do not allow insurance claims for damage due to fire, health or accident insurance from more than one company. It is obvious that any restriction on coinsurance can be circumvented if such claims are allowed. At the extreme, agents might have more than full coverage, inducing intentional damages, such as arson.

This tendency for exclusivity is reinforced by the advantages of long-term contracting. In the presence of risk aversion or limits on agents’ capital which prevent effective bonding, it may be necessary to promise future rewards to mitigate short term opportunistic behaviour. Termination of the agreement will deny these rewards and thus operates as a threat. This requires that contracts yield some rents to agents, so that their removal may constitute a punishment. Thus for example, the utility of being employed must exceed the utility of being unemployed (Shapiro and Stiglitz, 1984).

This requires rationing, which is not undone by competition. If being fired by one's employer leads to immediate employment elsewhere at the same wage, rather than to a significant period of unemployment, the threat of firing is ineffective. An equilibrium must be supported by transaction costs of finding new employment or by a collective use of the information contained in the firing. Such information is indeed relevant for hiring decisions by other firms. Its use depends on the costs of obtaining such information. Markets develop to supply such information, thereby increasing the effectiveness of such agreements. Credit information bureaus and employment agencies are some examples. Fama (1980) argues that such ‘reputation’ mechanisms eliminate moral hazard problems in executive markets. However, as the information is subject to noise, it is clear that moral hazard problems cannot be entirely resolved.

Non-market institutions may develop to mitigate some of these problems. Professional licensing and certification limit the number of physicians, lawyers and many other professionals. Aside from issues of

assurance of minimum quality and monopoly, these arrangements insure rents to the professions involved and, hence, make license removal a significant penalty (Arrow, 1963).

The consequences of moral hazard in political processes have largely been neglected by economists. Exceptions are Stigler (1971) and Peltzman (1976), who analysed the motivations of regulators, and Buchanan and Tullock (1962). The theoretical tools of agency, contract and game theory have yet to be fruitfully employed in this area. Given the expanding role of government and the evidence of widespread abuses in the political process, such application promises to yield significant dividends.

General equilibrium and welfare effects

There has been little research on the welfare implications of moral hazard. An exception is Stiglitz (see for example Arnott and Stiglitz, 1985), who noted that the existence of moral hazard creates second best contracts. In an economy characterized by such contracts, changes in contracts between any two parties have significant first order effects on social welfare, in contrast to the Arrow–Debreu economy, where first order effects of individual actions are zero at an optimum. As we have seen, moral hazard may lead to rationing and queues, suboptimal expenditure of hidden actions and imperfections in capital markets. This is not surprising because moral hazard is basically a form of externality. It is well known that uninternalized externalities lead to non-concavities, possible non-existence of equilibria and inefficiencies. The existence of such inefficiencies signals a possible role for government. However, government intervention may well cause more problems than it solves. For example, attempts to supplement deficient insurance markets in the form of universal income (social security, income taxation) insurance have run into serious moral hazard problems of work incentives, tax avoidance and evasion, and so on. It is at least partly because of these moral hazard problems that such markets failed to develop. It is therefore unclear whether government supply of these services enhances welfare.

In contrast, government policies which enhance complete contracts and improve their enforcement, can be welfare enhancing. Examples are contract law, liability rules and trade regulations.

See Also

- adverse selection
- health economics
- incomplete contracts
- principal and agent (i)
- principal and agent (ii)

Bibliography

Arnott, R. and Stiglitz, J. 1985. Labor turnover, wage structures, and moral hazard: the inefficiency of competitive markets. *Journal of Labor Economics* 3, 434–62.

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review*

53, 541–67.

Arrow, K. 1985. The economics of agency. In *Principals and Agents: The Structure of Business*, ed. J. Pratt and R. Zeckhauser. Boston: Harvard Business School Press.

Becker, G. 1964. *Human Capital*. New York: Columbia University Press.

Becker, G. 1968. Crime and punishment – an economic approach. *Journal of Political Economy* 76, 169–217.

Becker, G. and Stigler, G. 1974. Law enforcement, malfeasance and compensation of enforcers. *Journal of Legal Studies* 3, 1–18.

Buchanan, J.M. and Tullock, G. 1962. *The Calculus of Consent*. Ann Arbor: University of Michigan Press.

Fama, E. 1980. Agency problems and theory of the firm. *Journal of Political Economy* 88, 288–307.

Green, J. 1985. Differential information, the market and incentive compatibility. In *Frontiers of Economics*, ed. K.J. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

Harris, M. and Raviv, A. 1979. Optimal incentive contracts with imperfect information. *Journal of Economic Theory* 20, 231–59.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13, 314–40.

Jensen, M. and Meckling, W. 1976. Theory of the firm: managerial behavior, agency costs, and capital structure. *Journal of Financial Economics* 3, 305–60.

Peltzman, S. 1976. Towards a more general theory of regulation. *Journal of Law and Economics* 19, 211–40.

Posner, R. 1977. *The Economic Analysis of Law*. 2nd edn. Boston: Little, Brown.

Ross, S. 1973. The economic theory of agency: the principal's problem. *American Economic Review* 63, 134–9.

Shapiro, C. and Stiglitz, J. 1984. Equilibrium unemployment as a worker incentive device. *American Economic Review* 74, 433–44.

Shavell, S. 1979. Risk sharing and incentives in the principal and agent relationship. *Bell Journal of Economics* 10, 55–73.

Smith, A. 1776. *An Inquiry into the Nature and the Causes of the Wealth of Nations*. Ed. E. Cannan, New York: Modern Library, 1937.

Stigler, G. 1971. The theory of economic regulation. *Bell Journal of Economics and Management Science* 2, 3–21.

Stiglitz, J. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

Williamson, O. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

Wilson, R. 1969. The structure of incentives for decentralisation under uncertainty. In *La Décision*. Paris: Editions du CNRS.

How to cite this article

Kotowitz, Y. "moral hazard." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000259> doi:10.1057/9780230226203.1138

The New Palgrave Dictionary of Economics Online

Morgenstern, Oskar (1902 – 1977)

Martin Shubik

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

game theory; Morgenstern, O.; perfect foresight; prediction; rational behaviour; von Neumann, J.

Article

Morgenstern was born in Goerlitz, Silesia, on 24 January 1902. He died on 26 July 1977 at his home in Princeton, New Jersey. The two main intellectual centres of his life were Vienna and Princeton. In each case the source of his intellectual stimulation was not primarily the university but institutions such as the Wienerkreis of Moritz Schlick in Vienna, where he counted among his friends Karl Popper, Kurt Gödel and Karl Schlesinger, and the Institute for Advanced Study at Princeton. He obtained his doctorate in 1925 from the University of Vienna, where he was greatly influenced by Karl Menger and the writings of Eugen Böhm-Bawerk.

Morgenstern's first major work, *Wirtschaftsprognose* (1928), which was published in Vienna, served as his Habilitation thesis leading to his appointment as a *privatdozent* at the University of Vienna in 1929. In this book he began to consider the difficulties and paradoxes inherent in economic prediction, being particularly concerned with prediction where the action of a few powerful individuals could influence the outcome. He illustrated some of these difficulties with the example of Sherlock Holmes's pursuit of Professor Moriarty (an example repeated in the *Theory of Games*, 1944).

He became a professor at the University of Vienna in 1935, and in the same year published in the *Zeitschrift für Nationalökonomie* (of which he was managing editor) an article on fundamental difficulties with the assumption of perfect foresight in the study of economic equilibrium. It was then that the mathematician Edward •ech noted that the problems raised by Morgenstern were related to those treated by von Neumann in his article ‘Zur Theorie der Gesellschaftspiele’, published in 1928. Morgenstern did not have the opportunity to meet von Neumann until somewhat later. They both recalled meeting at the Nassau Inn in Princeton on 1 February 1939, although each believed that they had met once before. They became close friends and remained so until von Neumann's death on 8 February 1957.

In Vienna, Morgenstern was also director of the Austrian Institute for Business Cycle Research (1931–8)

where he employed Abraham Wald, whom he later helped go to the United States. In 1938, due to his opposition to the Nazis, Morgenstern was dismissed from the University of Vienna as ‘politically unbearable’ and he accepted an offer from Princeton, to some extent because of the presence of von Neumann at the Institute for Advanced Study. Their close collaboration resulted in the publication in 1944 of their book, *The Theory of Games and Economic Behavior*. This major work contained a radical reconceptualization of the basic problems of competition and collaboration as a game of strategy among several agents, as well as an important novel approach to utility theory (presented in detail in the second edition, 1947).

Both Morgenstern and von Neumann were well aware of the limitations of their great work. They stressed that they were beginning by offering a sound basis for a static theory of conscious individually rational economic behaviour and that the history of science indicated that a dynamic theory might be considerably different. They warned against premature generalization.

In his years at Princeton from 1938 until his retirement in 1970, Morgenstern encouraged the work of a distinguished roster of younger scholars in game theory and combinatoric methods. This was feasible primarily through the strength of the Mathematics Department and its connections with the Institute. There was little interest in the subject in the Department of Economics at the time. The ideas of the *Theory of Games* were so radical that they have taken many years to permeate the social sciences. Even at the time of his death many in the economics profession were sceptical of or indifferent to its contributions.

Although his work on the theory of games was undoubtedly Morgenstern's greatest contribution and collaboration, his interests were wide-ranging. His two books, *On the Accuracy of Economic Observations* (1950), and *Predictability of Stock Market Prices* (1970), written jointly with Clive W. Granger, indicate these interests. He was also concerned with matters of national defence and in 1959 published *The Question of National Defense*.

In 1959 he was one of the founders of Mathematica, a highly successful and sophisticated consulting firm, and served as Chairman of the Board. After retiring from Princeton he was Distinguished Professor at New York University until his death.

Selected works

1928. *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten*. Vienna: Julius Springer.

1935. Vollkommene Voraussicht und wirtschaftliches Gleichgewicht. *Zeitschrift für Nationalökonomie* 6 (3), 337–57.

1944. (With J. von Neumann.) *Theory of Games and Economic Behavior*. Princeton: Princeton University Press. 2nd edn, 1947.

1950. *On the Accuracy of Economic Observations*. Princeton: Princeton University Press.

1959. *The Question of National Defense*. New York: Random House.

1970. (With C.W.J. Granger.) *Predictability of Stock Market Prices*. Lexington, MA: Heath Lexington Books.

How to cite this article

Shubik, Martin. "Morgenstern, Oskar (1902–1977)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000262> doi:10.1057/9780230226203.1139

The New Palgrave Dictionary of Economics Online

Morishima, Michio (1923 – 2004)

Meghnad Desai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Morishima's contribution to economic theory was in tackling questions of equilibrium and dynamics with and without money, with heterogenous capital and in a multisectoral framework. He tried to synthesize and answer questions raised by Ricardo, Marx, Walras, Wicksell, Keynes and Schumpeter. His work was influenced by von Neumann's model and Hicks's style of theorizing.

Keywords

capital accumulation; capital controversy; credit creation; deflation; Econometric Society; effective demand; excess demand and supply; exploitation; false trading; growth paths; heterogeneous capital; Hicks, J. R.; homogeneity postulate; inflation; innovations; investment functions; involuntary unemployment; IS–LM models; Keynes, J. M.; Marx, K. H.; Marx–von Neumann model; microfoundations; Morishima, M.; natural rate of interest; nominal demand; nominal rate of interest; quantity theory of money; Ricardo, D.; saving and investment; Say's Law; Schumpeter, J. A.; stability of equilibrium; tâtonnement; turnpikes; von Neumann, J.; Walras, L.; Walras–Leontieff model; Wicksell, J. G. K.

Article

Michio Morishima was one of the most distinguished economic theorists of his generation. He taught in Japan at Kyoto and Osaka Universities, and in the UK he was the Keynes Visiting Professor at the University of Essex 1969–70 and Professor of Economics, later the John Hicks Professor of Economics, at the London School of Economics 1970–84 and Emeritus Professor for the rest of his life. He was awarded the Order of Culture [Bunka Kunsho] of Japan by the Emperor in 1976, a Fellowship of the British Academy in 1981 and an Honorary Fellowship of the LSE upon his retirement. Morishima became the first Japanese to be the President of the Econometric Society in 1965. He died aged 80 on 13 July 2004, leaving behind his wife Yoko and two sons and a daughter.

Morishima's work encompasses general equilibrium theory with heterogeneous capital, growth and money, as part of a coherent attempt to tackle one of the most intractable problems in economic theory, namely, the construction of an adequate theory of a dynamic growing economy with heterogeneous capital and money as well as credit or, to put it another way, a theory of how the capitalist system works.

Morishima's Ph.D. thesis at Kyoto University (published in Japanese in 1950 and in English in 1996 under the title *Dynamic Economic Theory*) dealt with stability of equilibrium. The standard (Hicksian) theory says that if the market starts out at a price away from the equilibrium given by the intersection of the demand and supply curves, then the price must change until the equilibrium point is reached. But how? Walrasians posit an auctioneer who would call out prices and register demands and supplies at each price. No trades are made until the auctioneer is satisfied that demands and supplies balance, that is, no false trading. The corollary of a no false trading equilibrium is that there can be never be involuntary unemployment, raising the issue of the consistency of micro and macro theories with each other.

Morishima prefers the case in which trading takes place at each price, but the price changes if, at that price, after transactions are closed, there is excess supply or demand. This would be a non-tâtonnement process, where some traders may buy (sell) at a price higher (lower) than the equilibrium price. He does not, however, develop this any further in the thesis but asks: are we exploring the path of convergence of the ‘groping’ prices, that is, virtual prices at which no trades are carried out and hence within ‘the market day’, or are we talking of the path of equilibrium prices arrived at, at the end of the tâtonnement in each market day from one day to the next?

Within the Hicksian week, the groping process traces out a path of virtual prices which converge to equilibrium under certain well-known conditions. But what of the sequence over several weeks of the equilibrium price? What are the dynamics of the path itself? It is this question that Morishima poses in *Dynamic Economic Theory* and pursues over his entire career. It is obviously connected to the stability of a growth path, since the path of income is analogous to the path of equilibrium prices. Morishima's discussion of growth paths was therefore always concerned not only with the quantity variables such as income and the stock of capital but also prices and interest rates.

Morishima's first book in English, *Equilibrium, Stability and Growth* (1964), tried to integrate Walras into the growth story, which had not hitherto been attempted, and also gave prominence to Marx's work on accumulation at the same time. Morishima constructed Walras–Leontieff and Marx–von Neumann models, which are pioneer efforts. *Equilibrium, Stability and Growth* is growth-oriented with an emphasis on linear technology and balanced maximal growth paths with fixed coefficients. But there is also a chapter on a spectrum of techniques. This is Morishima's response to the then ongoing capital controversy between Cambridge England and Cambridge Massachusetts.

Very soon after *Equilibrium, Stability and Growth* was published, Morishima came out with his most ambitious work to date, *Theory of Economic Growth* (1969). Here Morishima sets out a rigorous multisectoral framework – the von Neumann model – and integrates Walras as well as Hicks into this framework. Prices are solved out along with quantities throughout. Turnpikes are discussed under various assumptions. But Morishima also deals with the issue of the optimality of the maximal growth paths.

Morishima was not happy with *Theory of Economic Growth*. Thus started his long detour via Marx, Walras and Ricardo, until he could come back to his major concern. Morishima's book *Marx's Economics* (1973) deals with the statics and dynamics of Marx's growth and exploitation theory and

tackled joint production with innovative insights. It shows that labour values can be used to tackle the aggregation problem for heterogeneous capital.

The crucial next step is provided by Walras. Most economists think that Walras provided consistent microfoundations for a full employment–all markets clearing theory of the macroeconomy. Morishima had a different Walras in his 1977 book with the intriguing title *Walras' Economics: A Pure Theory of Capital and Money*. Morishima's purpose in the book is to see whether he can exploit Walras's work to provide the microfoundations of Keynesian macroeconomics. He focuses on the contrast between nominal demands (neoclassical) and effective demands (Keynesian) as well as the alternative hypotheses that investments adjust to savings (neoclassical) and that investments are prior and saving adjust (Keynesian). Walras's entrepreneurs have no income; they work on altruistic principles. Morishima adjusts Walras's investment function as well as giving entrepreneurs an income (profits) which makes the model closer to real capitalism. But he also shows why one needs a theory of accumulation and growth, that is, a story with time and future in it, in order to have a rationale for holding money in a Walrasian world. In a static general equilibrium, money can, and does, play no role.

The heart of Morishima's book *Ricardo's Economics* (1989) is in the final section entitled 'Three Paradigms Compared'. Say's Law is at issue. Ricardo established Say's Law as a dominant mode of theorizing. Usual departures from Say's Law involve a non-trivial role for money and/or a growth process via an active investment function. Ricardo had neither and so could subscribe to Say's Law. Marx had both but his investment function was very restrictive and made no use of money or credit. Walras had money towards the end of *Elements* but his growth theory lacked an investment function which led the way for savings to adjust to investment. Keynes of course had money and investment functions, but he did not spell out the microfoundations. Growth is not sufficient to justify a violation of Say's Laws; money or an investment function which has a role for entrepreneurs to respond to uncertainty is required.

In *Ricardo's Economics* a model is set up in which excess demand and supply for labour and capital are modelled in a simple diagram (1989, fig. 6, p. 218). Here, around an equilibrium point, zones of excess supply and demand for the two factors are mapped out. Morishima's axes are the real wage and the output capital ratio. Within the same general model all the three paradigms are embedded. Again, the investment function turns out to be the crucial relationship for the Anti-Say's Law result that Keynes established.

Capital and Credit: A New Formulation of General Equilibrium Theory (1992) brings together all the major themes of money, heterogeneous capital, underemployment equilibria and growth. The major innovation in *Capital and Credit* is that banks play a crucial role in financing production. This is Schumpeter rather than Keynes. While in Keynes's scheme entrepreneurs may underinvest because of expectations or a low marginal efficiency of capital relative to the rate of interest, Schumpeter allows for overshooting of credit creation by bankers. Thus, inflation as well as underemployment is possible. *Capital and Credit* is therefore concerned with innovations and their financing and monetary disequilibrium. The economy is split into Say's Law and Anti-Say's Law activities. There is a scope for Anti-Say's Law if production is financed by credit, and this of course requires that it is not instantaneous but has an input–output lag. With instantaneous production and investment adjusting to savings, Say's Law is confirmed. But in any realistic capitalist economy it breaks down due to the presence of credit. The amount of credit determines activity in the Anti-Say's Law sector (manufacturing industry, in other words), and this, via the multiplier, determines the overall levels of activity and employment. This need

not be full employment.

The separation of the economy as between relative prices determined by demand/supply and absolute prices as determined by money –‘the classical dichotomy’ is no longer valid. It is only by omitting banks and the financial requirements for production that the dichotomy is sustained.

In the last chapter, on ‘Monetary Disequilibrium’, Wicksell's cumulative process is examined from the point of view of von Neumann. The real system establishes the rate of profits (= rate of growth), but it leaves the price level indeterminate. Credit creation by bankers determines the nominal level of interest with the natural rate given by the real system. Then the monetary side determines the price level by the intersection of the money demand function and the real growth rate. But it is not a stable equilibrium. It is a kind of IS–LM model, but with its axes as interest rate and price level rather than income.

So we now enter a new development in monetary and growth theory. If the economy is growing and/or if the natural rate is a variable, then we need to extend Wicksell's analysis, which assumed a constant natural interest rate. But the natural rate may be above or below the von Neumann rate, and if the natural rate is also variable then the gap between the natural and the money rate is variable over the cycle. Thus, if the natural rate is above the money rate and the von Neumann rate, then inflation follows, but that may reduce the natural rate. If it then crosses over to being below the money rate, deflation follows and the natural rate may approach the von Neumann rate from above. Prices keep falling, and the economy may converge to the von Neumann rate.

In the converse case, the economy starts off with the natural rate below the money rate and below the von Neumann rate, and then deflation comes first as the natural rate approaches the von Neumann rate from below. Once it crosses over the constant money rate, then inflation follows and the economy approaches the von Neumann rate in an explosive inflationary situation.

This is the most sophisticated discussion of money and growth in the classical Wicksell framework. A variable natural rate is seldom modelled, and the deflation–inflation cycles enrich the Wicksell model greatly. But we are still in the world of Say's Law. What happens if we break away from it? The shortage of credit will restrict the economy below full employment, as Keynes envisaged, and abundance of credit will start off an inflationary growth process, as Schumpeter said. This then is the climax of the entire edifice of Morishima's work. He can now combine Anti-Say's Law with credit and disequilibrium. Credit creation determines the natural rate via the Anti-Say's Law sector, which is often the most innovative and dynamic. Morishima can then tackle the classical dichotomy.

This is the homogeneity postulate whereby nominal variables cannot have real effects and so money must be a veil. But the homogeneity postulate requires that a monetary shock be evenly spread across all agents. It also requires that the elasticity of demand with respect to money balances be identical across all agents. Morishima shows in the final pages of *Capital and Credit* that neither of these assumptions is likely to be fulfilled in a monetary economy. Agents including households and firms and the Anti-Say's Law firms are much more credit-sensitive than other firms, for one thing. And if the homogeneity postulate falls, so does the quantity theory.

The challenge of integrating money and growth with general equilibrium but without Say's Law has been accomplished. There is much more to be gained from a careful study of these writings and one can only hope that future scholars will mine the rich source of theoretical insights in the decades to come.

See Also

- capitalism
- dynamic models with non-clearing markets
- growth models, multisector
- Say's Law

Selected works

1950.*Degaakuteki Keizai Riron*. Tokyo: Kobundo.

1964. *Equilibrium, Stability and Growth*. Oxford: Clarendon Press.

1969. *Theory of Economic Growth*. Oxford: Clarendon Press.

1973. *Marx's Economics: A Dual Theory of Value and Growth*. Cambridge: Cambridge University Press.

1973. (ed.) *Theory of Demand: Real and Monetary*. New York: McGraw Hill.

1976. *Economic Theory of Modern Society*. Cambridge: Cambridge University Press.

1976. *Walras' Economics: A Pure Theory of Capital and Money*. Cambridge: Cambridge University Press.

1978. (With G. Catephores.) *Value, Exploitation and Growth*. London: McGraw Hill.

1986. *The Economics of Industrial Society*. Cambridge: Cambridge University Press.

1989. *Ricardo's Economics: A General Equilibrium Theory of Distribution and Growth*. Cambridge: Cambridge University Press.

1992. *Capital and Credit: A New Formulation of General Equilibrium Theory*. Cambridge: Cambridge University Press.

1996. *Dynamic Economic Theory* [an English translation of *Degaakuteki Keizai Riron*, 1950, with additional articles]. Cambridge: Cambridge University Press.

How to cite this article

Desai, Meghnad. "Morishima, Michio (1923–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_M000353> doi:10.1057/9780230226203.1140

The New Palgrave Dictionary of Economics Online

mortality

James W. Vaupel, Kristín G. von Kistowski and Roland Rau
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mortality is a demographic component that contributes to shaping the size, structure, and dynamics of populations. Life expectancy has been rising remarkably in the more developed countries since the 19th century and the process of rising life expectancy also has begun in most of the less developed countries. Increases in adult life expectancy and declines in birth rates result in aging societies. Survival is increasing as a result of progress in economic development, social improvements, and advances in medicine. However, death rates vary significantly in different parts of the world and are particularly high in sub-Saharan Africa.

Keywords

biodemography; central death rate; demography; fertility; Gompertz law of mortality; health care; innovations; life expectancy; marriage and divorce; migration; mortality; pensions; population aging; standard of living; time-series analysis

Article

Mortality is one of the three demographic components that shape the size, structure, and dynamics of populations; the other two are fertility and migration. Death rates have declined remarkably in modern times. The populations of the more developed countries have been aging for more than 100 years and the process of rising life expectancy also has begun in most of the less developed countries. Survival is increasing as a result of progress in economic development, social improvements, and advances in medicine. Mortality has been falling steadily especially in wealthier, economically advanced countries and has continued to do so during the second half of the 20th century and after, particularly at higher ages. We are getting older and the number of the elderly is increasing in most countries.
While the reduction in human mortality can be considered one of the greatest achievements of modern civilization, rising longevity and the increasing number of elderly will pose major challenges to health

care and social security systems. Declines in birth rates and increases in adult life-expectancy result in aging societies. These demographic changes will impact the life-course decisions of individuals, social interaction, economic development, and policy reforms in the countries involved.

Rising life expectancy has globally been a widespread phenomenon, but mortality differentials remain. Death rates vary significantly in different parts of the world and are particularly high in sub-Saharan Africa by global standards. Mortality conditions have changed throughout history and vary among and within populations. Death rates differ according to the country of origin, place of residence, sex, socio-economic status, level of education and marital status.

Mortality and life expectancy

Various indicators exist to measure mortality. Two of the indicators most often used and cited are the central death rate and life expectancy. The former, which is age-specific and time-specific, is defined as the number of deaths occurring at a given age during a given year, divided by the mean population of that age and year.

Life expectancy is an estimate of average age at death under current death rates. It is calculated by imposing the age-specific death rates of the respective year on a hypothetical cohort of newborns. In 2004, Japan reached the highest female life expectancy (85.59 years) ever obtained by a country. Lowest life expectancy is generally recorded in sub-Saharan Africa. An example is Zimbabwe, a country that in 2004 suffered the world's lowest life expectancy, 34 years for men and 37 years for women, according to WHO (2006). The United Nations estimated worldwide life expectancy for 2000–5 at 67.7 and 63.2 for women and men, respectively (United Nations, 2005).

Remaining life expectancy at age x is usually denoted as e_x and e^0_x. A value of x=0 leads to the most often published indicator, ‘life expectancy at birth’. Note that ‘life expectancy’ for a given year is based on a hypothetical cohort. Only if death rates are not changing can the average newborn be expected to live the number of years indicated by life expectancy. If age-specific mortality continues to decrease – as was the case in many developed countries during recent decades – then the actual average age at death of a birth cohort would be higher than the one estimated for the hypothetical cohort.

Age trajectories of human mortality and the Gompertz law of mortality

As individuals age, they tend to suffer an increasing loss of physical function and greater susceptibility to disease and injury. Benjamin Gompertz, a British actuary, described in 1825 the gradual increase in mortality rates with age, using an exponential curve, today known as the ‘Gompertz law of mortality’. The model implies that there is a constant rate of increase in the age-specific mortality of adult populations; for many populations this rate of increase is about ten per cent per year. The Gompertz model fits human mortality rates well for adults aged 30 to 85 in most modern populations with high life expectancies.

The overall age trajectory of human mortality is roughly U-shaped. Mortality is high immediately after birth. During infancy it decreases rapidly with age to reach a minimum between the ages of 10 and 15. Thereafter, the risk of dying rises more or less exponentially according to the Gompertz law of mortality, with some excess mortality among young adults. A rise in mortality during early adulthood is

often referred to as ‘accident hump’, as it is mainly caused by accidents in many modern populations (Heligman and Pollard, 1980). Especially in industrialized countries, this hump is more pronounced for men than for women. The hazards associated with being a woman of childbearing age have been greatly reduced in developed countries, but those connected with the transition to manhood are still substantial. Maternal mortality, in contrast, is confined almost exclusively to developing countries. Among women worldwide, those in sub-Saharan Africa are at highest risk of dying during pregnancy and at childbirth: The lifetime risk was estimated by WHO at 1 in 16 in 2002; this compares to a risk of 1 in 2,500 in the United States (WHO, 2004).

Most deaths in developed countries today are concentrated at older ages. Death rates at older ages have, however, declined markedly during the second half of the 20th century. Furthermore, after age 80 death rates rise more slowly than predicted by the Gompertz exponential formula, and may roughly level off around age 110, albeit at the high level of about 50 per cent mortality per year (Thatcher, Kannisto and Vaupel, 1998; Robine and Vaupel, 2002).

Rising life expectancy in industrialized countries

The rise in life expectancy is one of the great achievements of modern times. In the countries with the highest levels, female life expectancy has been rising for 160 years at a steady pace of almost three months per year (Oeppen and Vaupel, 2002). The four-decade increase in best-practice life expectancy is so extraordinarily linear that it may be the most remarkable regularity of mass endeavour observed. On average, women live longer than men, but record life expectancy has also risen linearly for men since 1840, albeit a little more slowly than for women. The improvements in survival leading to the linear climb in record life expectancy result from the intricate interplay of advances in income, salubrity, nutrition, education, sanitation and, in recent decades, medicine (Riley, 2001).

When we look at individual countries, gains in life expectancy have not progressed as linearly. The gap between the record level and the national level can be regarded as a measure of how much better a country might do. Neither the trend in record life expectancy nor the life expectancy trajectories in different countries suggest that a limit to life expectancy is in sight. Although rapid progress in catch-up periods is typically followed by slower increases, none of the curves appear to be approaching a maximum value (Oeppen and Vaupel, 2002).

The rising numbers of centenarians in developed countries is another striking piece of evidence for the continuing increase in longevity. Lifespans exceeding 100 years, which seemed almost impossible to achieve in the past, despite spectacular reports, are increasingly becoming part of our reality today.

It is unlikely that any person living in Sweden before 1800 attained the age of 100 (Jeune, 1995) and throughout the world centenarians must have been very rare (Wilmoth, 1995). Data on the pre-18th century period have to be interpreted with caution. Few reliable statistics are available on mortality levels among the very old living under conditions of low life expectancy. The lower life expectancy is, the greater is the tendency to exaggerate age at older ages (Kannisto, 1994). Today, the number of centenarians in developed countries is increasing at an exceptionally rapid rate of six to nine per cent per year in many countries. While 265 centenarians were counted in England and Wales in 1950, there were 5,895 of them 50 years later, that is, more than 20 times the 1950 figure (Kannisto–Thatcher Database). In developed countries, the number of people celebrating their 100th birthday doubled each decade

between 1950 and 1980; by the end of the 20th century it was multiplying by a factor of 2.4 per decade.

The history of mortality decline

How can the transition from high to low mortality be explained? Over most of the course of human existence, life expectancy hovered between 20 and 30 years. Infant mortality was high, people fell victim to infectious and parasitic diseases or simply to the harshness of everyday living conditions. Even in western Europe life expectancy did not reach age 40 until after 1800, and it stayed below age 50 until after 1900 (Vaupel and Jeune, 1995). Over the course of the 20th century, life expectancy rose dramatically by more than 30 years in many industrialized countries. Rising life expectancy in industrialized countries since the 19th century is related to a fundamental epidemiological transition. There was a shift from the predominance of high mortality from infectious disease to conditions in which non-communicable and degenerative diseases among the elderly became more important. By the beginning of the 19th century in European areas of the world epidemics had been reduced, food supply became more stable, and fluctuations in mortality decreased. Over the course of the 19th century, the standard of living and hygiene improved and some public health services were established in a number of countries (Bongaarts and Bulatao, 2000). Infectious disease was the greatest scourge of mankind until the first half of the 20th century, that is, until vaccination, antibiotics, and other medical advances finally began to combat successfully many of the life-threatening diseases in industrialized countries. By the same token, they lowered the rates of infant and child mortality and limited the devastating effects of the largest epidemics, although some outbreaks of influenza and the HIV/AIDS epidemic are exceptions. Parallel to these changes, there was a shift from high to low fertility. Mortality associated with pregnancy and birth decreased considerably.

The second half of the 20th century saw a dramatic reduction in death rates at advanced ages (Vaupel and Jeune, 1995; Kannisto, 1994; Kannisto et al., 1994; Vaupel, 1997). The time around 1950 marks a distinct change in mortality conditions among the ‘oldest old’ (85 or more years of age) in developed countries: While improvements in survival were slow in the years preceding 1950, progress made after 1950 and especially after 1970 has been impressive. Data from England, Wales, France, Iceland, Japan, and the United States show clearly that old-age survival has been increasing since 1950 (Vaupel, 1997; Vaupel et al., 1998). The population of centenarians and even super-centenarians (persons older than 110 years) is growing rapidly. The increase in the number of births about a century ago coupled with a sharp decline in mortality from childhood to age 80 contributed to the rising numbers. Demographic analyses, however, demonstrate that the most important factor behind the explosion of the centenarian population has been the decline in the mortality rate after age 80, a factor that has been two to three times more important than the other factors combined (Vaupel and Jeune, 1995). The ongoing increase in life expectancy is largely attributable to continuous improvements in survival at advanced ages (Vaupel and Jeune, 1995; Vaupel, 1997).

In developed countries, the decline in mortality caused by infectious diseases and the postponement of degenerative diseases has delayed deaths to increasingly older ages. Today, cardiovascular disease and cancer are the major causes of death in industrialized countries. In 2002, heart disease and stroke accounted for more than half of all deaths, and cancers were responsible for around 20 per cent of all deaths in Europe (WHO, 2004).

The human survival curve, which depicts the proportion of an initial (hypothetical) cohort still alive, has changed its shape as a consequence. The survival curve is becoming more rectangular due to the concentration of deaths at higher ages. To provide an example, the 2002 life table for Japanese women shows that more than 95 per cent of the initial hypothetical cohort would be still alive under current mortality rates at age 60. Mortality decline is neither a regular process in industrialized countries nor is it a process confined to these nations. Life expectancy has risen in most developing countries, too, especially in many Asian states and in Latin America. The mortality transition is driven by the same factors as in the developed countries – combating infectious disease plays a major role here. However, the transition proceeds much faster than it did in industrialized nations and there are considerable differences in the degree of progress (Bongaarts and Bulatao, 2000).

The plateau in late-life mortality

Human death rates increase slowly after age 80. Data analyses of very large cohorts reveal that death rates reach a plateau at advanced ages and may level off around age 110 (Thatcher, Kannisto and Vaupel, 1998; Robine and Vaupel, 2002). This observation is not unique to humans, however. Late-life mortality deceleration has been noticed in and confirmed for a number of model organisms as diverse as yeast, nematodes, or fruit flies. For all species for which large cohorts have been followed to extinction, age-specific mortality decelerates and, for the largest populations studied, even declines at older ages (Vaupel et al., 1998).

Some concepts contributing to an understanding of the astonishing improvement in survival at late ages come from biodemography, a subject that has emerged at the confluence of demography and biology. One biodemographic explanation builds on heterogeneity in frailty. All populations are heterogeneous, and even genetically identical populations display phenotypic differences. Frailer individuals have a lower probability of survival to late ages; robust individuals have a higher one. The frail tend to suffer high mortality, leaving a select subset of robust survivors. This results in compositional change in the surviving, aging population and in slower increases in age-specific death rates (Vaupel, Manton and Stallard, 1979; Curtsinger et al., 1992; Vaupel and Carey, 1993; Yashin, Vaupel and Iachine, 1994).

Another biodemographic explanation refers to changes in survival capacities at the individual level. Generally, the longevity of individual organisms is influenced by the living conditions to which they are exposed. Studies with different species have shown that several environmental factors of non-lethal stress, for example dietary restriction or heat shock, can induce increases in both resistance and longevity (Lithgow et al., 1995; Murakami and Johnson, 1996; Masoro, 2000). Hormesis, a biologically favourable response to low exposure to stress or toxins, is a well-known physiological phenomenon. Caloric restriction has proven to be an effective way to extend life span in a wide range of species, from yeast to mammals (Masoro, 2000). It is not clear, however, whether fasting is a way of prolonging life in humans.

The influence of current conditions on age-specific death rates

Studies involving model organisms have provided valuable insights into the biological processes of aging. An example can be drawn from a study on the Drosophila fruit fly. When flies fed a restricted diet

were switched to a full diet, mortality soared to the level suffered by flies that had been fully fed all their lives. Conversely, when the diet of fully fed Drosophila was restricted, mortality plunged within 48 hours to the level enjoyed by flies that had experienced a lifelong restricted diet (Mair et al., 2003). The results support the repeated finding that age-specific death rates for humans (and other species) are strongly influenced by current conditions and behaviour (Kannisto, 1994; Vaupel et al., 1998).

Placed in a broader context, the conclusion drawn from the fruit fly study also applies to humans. This can be illustrated neatly by an unplanned ‘natural experiment’ in Germany's recent history. Before reunification, both East and West Germany saw a radical decline in old-age mortality, as is characteristic for most developed countries. In the former GDR, however, mortality was considerably higher than in West Germany. Following unification (1989–1990), old-age mortality in East Germany declined to reach the levels prevailing in the West (Gjonca, Brockmann and Maier, 2000), a development largely attributed to improved health care for the elderly after unification. Thus, interventions even late in life can switch death rates to a lower, healthier trajectory. It's never too late to start prolonging your life (Vaupel, Carey and Christensen, 2003).

Longevity in humans has a relatively low heritability. Studies of twins indicate that a modest 25 per cent of the variation in life spans is attributable to genetic differences among people (McGue et al., 1993; Herskind et al., 1996; Finch and Tanzi, 1997). The discoveries of genetic and environmental factors that contribute to extensions of the lifespan do not fully explain the malleability of aging. Nevertheless, the findings show that there are means and ways of delaying aging.

The plasticity of aging

The rise in life expectancy has provoked discussion of the question whether we are approaching a limit to life expectancy, a biologically determined maximum lifespan that inevitably halts further improvements of old-age survival.

A common assumption still widely held is that lifespan cannot be extended beyond a biologically determined limit. The notion of an inevitable maximum lifespan also influences scientific studies of longevity (Fries, 1980; Olshansky, Carnes and Cassel, 1990). Ever since research into longevity began, attempts have been made to determine the maximum life expectancy that humans could reach. The ceilings proposed by various authors differ but all have been exceeded, apart from those proposed most recently (Oeppen and Vaupel, 2002). The assumption of a finite, biological limit to life can be traced back to Aristotle (350 bc). In his treatise ‘On Youth and Old Age, On Life and Death’, Aristotle contrasted two types of death: premature death caused by disease or accident, and senescent death due to old age. He believed that nothing could be done about old age and thus about the end to life. More than 2,300 years later, James Fries quantified Aristotle's distinction in a widely cited article published in the *New England Journal of Medicine*. If life is not cut short by accident or illness, then the lifespan of man will inevitably approach a potential maximum limit that is fixed for every human but differs from individual to individual (Fries, 1980). According to Fries, the fixed value of the maximum lifespan is normally distributed with a mean of 85 years and a standard deviation of seven years. Fries emphasizes that nothing can be done to alter a person's maximum lifespan as the latter is beyond the influence of environmental, behavioural, or medical intervention currently conceivable. Accordingly, death rates at older ages are intractable. The notion of unavoidable senescent death has been reinforced by

evolutionary biologists who hypothesize that mortality must rise with age as the force of selection against deleterious, late-acting mutations declines (Hamilton, 1966).

The notion of an upper biological limit to lifespan may be commonly accepted, yet there is no empirical evidence of a proximate limit to human longevity. The steady rise in human life expectancy shows no signs of levelling off. Experts repeatedly asserting that life expectancy is approaching a ceiling have repeatedly been proven wrong. If life expectancy were approaching an unavoidable biologically maximum, then the increase in life expectancy should be slowing, especially in countries such as Japan or France, both of which enjoy exceptionally low death rates. This, however, is not the case (Oeppen and Vaupel, 2002; Vaupel, 1997). Mortality is plastic even at advanced ages.

The prevailing causes of rising life expectancy have undergone changes and are complex. Combined, they have nonetheless led to a stable and linear increase in life expectancy since 1840. This will probably also apply to the future. Just as medical breakthroughs – for example, the discovery of antibiotics or advances in organ transplantation – were not foreseen, we do not know what major technological innovations the future will bring to promote long and healthy lives. There is no reason, however, to assume that progress in technological knowledge and its exploitation will come to a halt. It would not make sense to take the standards of today to estimate the conditions influencing life expectancy tomorrow. Future advances in life expectancy will be made as we progress in the prevention, diagnosis, and treatment of deadly age-related diseases (Barbi and Vaupel, 2005).

Future prospects of longevity

Because best-practice life expectancy has been increasing by 2.5 years per decade for the past 160 years, one reasonable scenario is that this trend will continue in the coming decades. To date, there is no indication that a change in the trend is in sight. If the trend continues, there may be a country in about six decades' time with life expectancy beyond the threshold of 100 years (Oeppen and Vaupel, 2002).

An application of this extrapolation in conjunction with methods from time-series analysis to project the gap between best-practice and national life expectancy results in national forecasts that are considerably higher than many official projections. From the use of this method, female life expectancy for Germany, for example, is expected to rise significantly above 90 years by 2050. Official projections, however, do not exceed 87 years (medium scenario). In many countries, official projections assume a deceleration in reductions of death rates. Such projections made in the past have resulted in underestimates of actual increases in life expectancy. These errors distort planning for future pensions, health care, and other social needs as well as the decision-making of individuals drawing up saving plans or planning for retirement. Increases in life expectancy of a few years can produce large changes in the numbers of old and oldest old who will need support and care. In developed countries, centenarians may well become commonplace during the lifetime of people alive today.

Mortality divergences

Although health trends have been generally positive throughout the world and remarkable improvements in survival have been achieved in developed and many developing countries (Tuljapurkar, Li and Boe, 2000; Vallin and Meslé, 2005), death rates still vary among countries and even within countries. In the

1970s and early 1980s, many demographers expected a convergence in life expectancies worldwide by assuming gains would be higher for the countries with lower life expectancies (McMichael et al., 2004). A quarter of a century later, however, it is clear that this assumption did not hold. On the one hand, increases in life expectancy of some of the best-performing countries, such as Japan or France, did not show any levelling off at all and life expectancy climbed higher than expected. On the other hand, there have been exceptions to the widespread phenomenon of general mortality decline in the second half of the 20th century. Mortality reversals were observed in the 1980s and 1990s in as many as 42 countries (McMichael et al., 2004; Caselli, Meslé and Vallin, 2002; Vallin and Meslé, 2005) as life expectancy fell. Most of these countries are situated in sub-Saharan Africa or in eastern Europe. Life expectancy in several sub-Saharan countries was more than ten years lower in 2004 than predicted by the UN Population Division about 20 years earlier (United Nations, 1981). Other countries that experienced reversals in life expectancy at the end of the 20th century are North Korea, Haiti, Fiji, the Bahamas, and Iraq. Setbacks apart from those caused by war and famine were not taken into account by early demographers, with the result that future setbacks in national mortality were considered unlikely (McMichael et al., 2004).

In sub-Saharan Africa, HIV/AIDS and other infectious diseases, such as tuberculosis and malaria, caused death rates to rise, and many of the countries involved were additionally faced with economic hardships, political conflicts, and violence between groups or individuals. Russia, like other countries of the former USSR or of eastern Europe, experienced increased mortality among working-age adults, especially among men aged between 20 and 65 (Shkolnikov et al., 1998; Meslé et al., 2003). Adults are normally less vulnerable to mortality increase than are children or the elderly. The drastic political and socio-economic transition increased unemployment rates and income inequalities, and led to weakened safety nets and to psycho-social stress among those most affected, particularly the less educated population groups (Shapiro, 1995; Shkolnikov et al., 1998; Bobak et al., 2000). Adverse male behaviours, such as alcohol abuse, crime, and violence, contributed to male excess adult mortality. In addition, rates of cardiovascular disease and cancer mortality are high in Russia.

Some industrialized countries perform less well than others. Since the mid-1980s in the United States, for example, death rates have declined more slowly than in most other developed countries. Until about 1980, the United States enjoyed relatively low death rates for both women and men after aged 65. Since then, however, death rates at older ages have fallen less rapidly than in Japan, France and other countries. The reasons for the slow increase in life expectancy in the United States are not yet well understood.

Mortality differentials

The U-shape of the mortality risk trajectory applies to all humans. Nevertheless, remarkable differentials exist by geographical region and along other dimensions. The best-known differential is between females and males. In most developed countries, the difference between female and male life expectancy is between four and seven years. The gap between women and men is typically smaller in less developed countries. It is not clear how much of the gap is biological as opposed to social, in part because biological factors interact with social ones. While men take more health risks (such as smoking), women are more careful about their health (for example, visits to the doctor).

Socio-economic status (SES) and mortality have an inverse relationship: individuals with higher SES usually enjoy lower mortality, regardless of how SES is measured (Goldman, 2001). Although measures of SES are correlated with each other, they address different dimensions: education is related to health behaviour and knowledge of healthy lifestyle, occupation to health hazards of the job, and income to access to health care as well as to the ability to provide a healthy living environment (such as housing conditions).

Marital status is another important mortality determinant. Married individuals usually have lower death rates than do never-married women and men, the widowed, or the divorced. Two different hypotheses have been discussed in the literature to explain this differential. On the one hand, marriage is expected to have a protective effect via pooled financial resources, higher social support, the adoption of healthier lifestyles, and other factors. On the other hand, it is argued that there is a selection effect into marriage: healthy women and men have higher chances of finding a spouse than less healthy individuals (Goldman, 1993).

See Also

- fertility in developing countries
- fertility in developed countries
- retirement

Bibliography

Aristotle. 350 bc. On youth and old age, on life and death, on breathing. Trans. G. Ross. Online. Available at http://classics.mit.edu/Aristotle/youth_old.html, accessed 20 August 2006.

Barbi, E. and Vaupel, J. 2005. Comment on ‘Inflammatory exposure and historical change in human life-spans’. *Science* 308, 1743.

Bobak, M., Pikhart, H., Hertzman, C., Rose, R. and Marmot, M. 2000. Socioeconomic factors, material inequalities, and perceived control in self-rate health: cross-sectional data from seven post-communist countries. *Social Science and Medicine* 51, 1343–50.

Bongaarts, J. and Bulatao, R., eds. 2000. *Beyond the Billion: Forecasting the World's Population*. Washington, DC: National Academy Press.

Caselli, G., Meslé, F. and Vallin, J. 2002. Epidemiologic transition theory exceptions. *Genus* 58(1), 9–51.

Curtsinger, J., Fukui, H., Townsend, D. and Vaupel, J. 1992. Demography of genotypes: failure of the limited life-span paradigm in *Drosophila melanogaster*. *Science* 258, 461–63.

Finch, C. and Tanzi, R. 1997. Genetics of aging. *Science* 278, 407–11.

Fries, J. 1980. Aging, natural death, and the compression of morbidity. *New England Journal of Medicine* 303, 130–5.

Gjonca, A., Brockmann, H. and Maier, H. 2000. Old-age mortality in Germany prior to and after reunification. *Demographic Research* 3(1). Online. Available at http://www.demographic-research.org/volumes/vol3/1/3-1.pdf, accessed 20 August 2006.

Goldman, N. 1993. Marriage selection and mortality patterns: inferences and fallacies. *Demography* 30, 189–208.

Goldman, N. 2001. Mortality differentials: selection and causation. In *International Encyclopedia of the Social and Behavioral Sciences*, ed. N. Smelser and P. Baltes. Oxford: Elsevier Science.

Gompertz, B. 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London* 11, 513–83.

Hamilton, W. 1966. The moulding of senescence by natural selection. *Journal of Theoretical Biology* 12, 12–45.

Heligman, M. and Pollard, J. 1980. The age pattern of mortality. *Journal of the Institute of Actuaries* 107, 49–80.

Herskind, A., McGue, M., Soerensen, T. and Vaupel, J. 1996. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. *Human Genetics* 97, 319–23.

Jeune, B. 1995. In search for the first centenarians. In *Exceptional Longevity: From Prehistory to the Present*, ed. B. Jeune and J. Vaupel. Odense: Odense University Press.

Kannisto, V. 1994. *Development of the Oldest-Old mortality, 1950–1990*. Odense: Odense University Press.

Kannisto, V., Lauritsen, J., Thatcher, A. and Vaupel, J. 1994. Reductions in mortality at advanced ages: several decades of evidence from 27 countries. *Population and Development Review* 20, 793–30.

Kannisto–Thatcher Database on old age mortality. Online: available at http://www.demogr.mpg.de/databases/ktdb/, accessed 20 August 2006.

Lithgow, G., White, T., Melov, S. and Johnson, T. 1995. Thermotolerance and extended life-span

conferred by single-gene mutations and induced by thermal stress. *Proceedings of the National Academy of Sciences USA* 92, 7540–4.

Mair, W., Goymer, P., Pletcher, S. and Partridge, L. 2003. Demography of dietary restriction and death in *Drosophila*. *Science* 301, 1731–3.

Masoro, E.J. 2000. Caloric restriction and aging: an update. *Experimental Gerontology* 35, 299–305.

McGue, M., Vaupel, J., Holm, N. and Harvald, B. 1993. Longevity is moderately heritable in a sample of Danish twins born 1870–1880. J*ournals of Gerontology: Series A, Biological Sciences and Medical Sciences* 48, B237–B244.

McMichael, A., McKee, M., Shkolnikov, V. and Valkonen, T. 2004. Mortality trends and setbacks: global convergence or divergence? *The Lancet* 262, 1155–9.

Meslé, F., Vallin, J., Hertrich, V., Andreev, E. and Shkolnikov, V. 2003. Causes of death in Russia: assessing trends since the 1950s. In *Population of Central and Eastern Europe. Challenges and Opportunities*, ed. I. Kotowska and J. Joswiak. Warsaw: Statistical Publishing Establishment.

Murakami, S. and Johnson, T. 1996. A genetic pathway conferring life extension and resistance to UV stress in *Caenorhabditis elegans*. *Genetics* 143, 1207–18.

Oeppen, J. and Vaupel, J. 2002. Broken limits to life expectancy. *Science* 296, 1029–31.

Olshansky, S., Carnes, B. and Cassel, C. 1990. In search of Methuselah: estimating the upper limits of human longevity. *Science* 250, 634–40.

Riley, J. 2001. *Rising Life Expectancy: A Global History*. Cambridge: Cambridge University Press.

Robine, J. and Vaupel, J. 2002. Emergence of supercentenarians in low mortality countries. *North American Actuarial Journal* 6, 54–63.

Shapiro, J. 1995. The Russian mortality crisis and its causes. In *Russian Economic Reform at Risk*, ed. A. Aslund. London: Pinter.

Shkolnikov, V., Cornia, G., Leon, D. and Meslé, F. 1998. Causes of the Russian mortality crisis: evidence and interpretations. *World Development* 26, 1995–2011.

Thatcher, A., Kannisto, V. and Vaupel, J. 1998. *The Trajectory of Mortality from Age 80 to 120*. Odense: Odense University Press.

Tuljapurkar, S., Li, N. and Boe, C. 2000. A universal pattern of mortality decline in the G7 countries. *Nature* 405, 789–92.

United Nations. 1981. *World Population Prospects As Assessed in 1980*. New York, United Nations.

United Nations. 2005. *World Population Prospects, The 2004 Revision, Highlights*. New York: Department of Economic and Social Affairs, United Nations.

Vallin, J. and Meslé, F. 2005. Convergences and divergences: an analytical framework of national and sub-national trends in life expectancy. *Genus* 61(1), 83–123.

Vaupel, J. 1997. The remarkable improvements in survival at old ages. *Philosophical Transactions of the Royal Society of London, Series B* 352, 1799–804.

Vaupel, J. and Carey, J. 1993. Compositional interpretations of medfly mortality. *Science* 260, 1666–7.

Vaupel, J.W., Carey, J. and Christensen, K. 2003. It's never too late. *Science* 301, 1679–81.

Vaupel, J., Carey, J., Christensen, K., Johnson, T., Yashin, A., Holm, N., Iachine, I., Kannisto, V., Khazaeli, A., Liedo, P., Longo, V., Zeng, Y., Manton, K. and Curtsinger, J. 1998. Biodemographic trajectories of longevity. *Science* 280, 855–60.

Vaupel, J. and Jeune, B. 1995. The emergence and proliferation of centenarians. In *Exceptional Longevity: From Prehistory to the Present*, ed. B. Jeune and J. Vaupel. Odense: Odense University Press.

Vaupel, J.W., Manton, K.G. and Stallard, E. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16, 439–54

WHO (World Health Organization). 2004. *Maternal Mortality in 2000: Estimates Developed by WHO, UNICEF, and UNFPA*. Geneva: WHO.

WHO. 2006. *World Health Report 2006: Working Together for Health*. Geneva: WHO.

Wilmoth, J. 1995. The earliest centenarians: a statistical analysis. In *Exceptional Longevity: From Prehistory to the Present*, ed. B. Jeune and J. Vaupel. Odense: Odense University Press.

Yashin, A., Vaupel, J. and Iachine, I. 1994. A duality in aging: the equivalence of mortality models based on radically different concepts. *Mechanisms of Ageing and Development* 74, 1–14.

Free internet sources on mortality data

The Human Mortality Database contains calculations of the death rates and life tables of almost 30 countries. Access to the data requires registration, which is free at http://www.mortality.org, accessed 20 August 2006.

The Kannisto–Thatcher Database on Old Age Mortality is tailored to analyse mortality at ages 80 and over. It has death and population counts by sex, age, birth year, and calendar year for more than 30 countries for ages 80 and above. Available at http://www.demogr.mpg.de/databases/ktdb., accessed 20 August 2006.

The Human Life-Table Database is a collection of life tables for more than 30 countries. Available at http://www.lifetable.de, accessed 20 August 2006.

How to cite this article

Vaupel, James W., Kristín G. von Kistowski and Roland Rau. "mortality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000263> doi:10.1057/9780230226203.1141

The New Palgrave Dictionary of Economics Online

motion pictures, economics of

W. David Walls

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Film-goers discover the films they like by consuming them, and through the exchange of information the demand for motion pictures evolves dynamically. The supply of screens adjusts in response to demand through flexible state-contingent exhibition contracts. This article presents an overview of the economics of motion pictures that focuses on how the demand process affects the distribution of outcomes, how the distribution of outcomes can be quantified with the use of statistical models, and how the industry's organization and business practices can be understood in light of the behavioural and statistical models.

Keywords

asymmetric information; Bose–Einstein distribution; contagion; motion pictures, economics of; optimal contracts; Pareto distribution; power laws; superstars, economics of

Article

The business of motion pictures is a fascinating laboratory for applied researchers in the social sciences. The glamorous subject matter makes the industry inherently interesting, but more important for empirical research is the availability of project-level data on investment and financial returns. Most studies of investment decisions are conducted at the industry or firm level, so that the researcher observes the return only on a portfolio of projects. In the movie business, the unit of observation is the individual project, and data are collected and reported in fine detail by many industry sources.
Early research on the movie business applied microeconomic theory to the industry and made little use of its detailed data and rich institutions. This early literature is important in providing the historical context in which many of the movie industry's business practices emerged. Kindem (1982) has collected in his volume many papers that provide organizational and institutional analyses of the motion-picture industry from its origins through the modern era. More recent papers in this line of applied research provide revisionist analyses of the industry's history and development (Chisholm, 1993; 1997; De Vany

and Eckert, 1991; De Vany and McMillan, 2004; Sedgwick, 2000).

The market for motion pictures is difficult to understand quantitatively, though the intuition is transparent. Film-goers discover the films they like by consuming them, and through the exchange of information the demand for motion pictures evolves over time. Supply adjusts as the available screens respond to demand through flexible state-contingent exhibition contracts. The present article provides an overview of the economics of motion pictures. The focus is on the how the demand process affects the distribution of outcomes, how the distribution of outcomes can be quantified, and how this relates to the industry's organization and business practices.

Movie-goer choices and outcome uncertainty

Understanding demand is essential if one is to make sense of the movie industry's contracts and business practices. Early viewers of a movie affect the choices of potential viewers – behaviour that goes under the names of herding, contagion, network effects, bandwagons, path-dependence, momentum, and information cascades. The particular models differ in their details, but they are dynamic in that demand depends on revealed demand, or more generally on how group behaviour arises from the interaction of individual decision-makers (Epstein and Axtell, 1996). Initial advantages in movie attendance can lead to extreme differences in outcomes when demand has recursive feedback. De Vany and Walls (1996) showed that box-office revenues have a contagion-like property where the week-to-week change in demand is stochastically dependent on previous demand. A big opening of a bad movie can kill it but a big opening of a good movie can lead to an avalanche of attendance and large revenues. Let's examine the demand for movies more closely to see the origins of extreme success and failure.

Assume for simplicity initially that there was only one movie that could be viewed by one consumer at any one time. Consumers choose in random sequence whether or not to go to the movie. If we further assume that the consumers have a common prior belief about the film's quality, then there is a common probability p that a randomly chosen person will choose to see the film. If we let X be the number attending the film, then X is a binomial random variable; it follows that when consumers share a common prior the film's revenues would follow a binomial distribution. When quality is unknown and priors over quality differ among viewers, p is a random variable. By conditioning on p and integrating over the binomial distribution, we see that each of the $n+1$ possible outcomes is equally likely; adding uncertainty to the priors transforms the distribution of revenue from the binomial to the uniform distribution.

Now consider information sharing, as has been modelled by Jovanovic (1987), where potential consumers can use information revealed during a film's run to refine their prior on its quality; this sort of information includes the opinions of other viewers, such as expert reviewers, advertising, and information from box office reports and queuing at cinemas. De Vany and Walls (1996) let the distribution of customers over screens be multinomial uniform, so the movie search problem – a search for quality with an unknown distribution – is similar to the search for price with an unknown distribution. Viewers who do not know the distribution begin with a uniform prior and adapt from there. The result of this process is the Bose–Einstein distribution which has the property that all of the possible outcome vectors are equally likely! This means the vector in which the attendance at every theatre is equal to zero is as likely as one in which all n trials go to only one theatre and every other vector is

equally likely (Feller, 1957). The Bose–Einstein distribution has uniform mass over a space of s+1-vectors; the s-vectors correspond to the revenues of the s theatres and one bin collects those who go to no film.

What is important about the evolution of choice probabilities under the Bose–Einstein choice logic is the way past successes are leveraged into future successes: as soon as individual differences emerge among the films, they are compounded by information feedback into very large differences over the course of a film's theatrical lifetime. A broad opening at many theatres can produce large and rapidly growing audiences, but it also can lead to early failure if the large crowd relays negative information. Movie customers sequentially select movies, and the probability that a given customer selects a particular movie is proportional to the fraction of previous customers who selected that movie. This result obtains because the probabilities are not known and sampling reveals information that causes previous selections to attract new ones.

Quantifying the distribution of movie outcomes

Box-office revenue is asymptotically power law or Pareto distributed (De Vany and Walls, 1999; 2002). One of the attractions of the power law distributions in explaining the movie business is that they allow for the heavy tails and skewness that are characteristic of box-office outcomes. Power laws emerge in many other systems with feedback of the type discussed above (Brock, 1999).

The stable Paretian model

Mandelbrot (1963) proposed the stable Paretian distribution as a general model for natural and social systems; it is applied in economics, finance, biology, geology, physiology, and other sciences (McCulloch, 1996; Uchaikin and Zolotarev, 1999; Mantegna and Stanley, 1995; Levy and Soloman, 1997). The stable distribution is the limiting distribution of all stable processes so that it contains the other well-known stable distributions (Cauchy, Lévy, Gaussian) as special cases. Motion picture profit is well fit by a stable distribution with infinite variance and positive skew (Walls, 2000; De Vany and Walls, 2004). The stable distribution's ability to capture the empirical regularities found in motion picture data and the distribution's statistical foundation on the most general form of central limit theorem make it a natural model of motion picture outcomes. The theoretical reason for thinking that a stable distribution might apply to motion pictures is that Mandelbrot (1963) showed that a dynamic process that is stable under choice, mixture, and aggregation converges in distribution to the stable distribution. If motion picture revenues and costs are discrete time processes with stable increments, then profit will converge to a stable distribution.

Conditional stable distribution

In empirical studies it is possible to model the stable Paretian distribution of movie outcomes conditional on a vector of explanatory variables with the use of McCulloch's (1998) stable regression model in which the index of stability α and the regression coefficients are estimated jointly. The stable

regression model has the familiar form of a linear regression $y_i = \beta_0 + \Sigma_{j=1}^{k} \beta_{ij} x_i + \varepsilon_i$ where the β's are the coefficients to be estimated and the *x*'s are the regressors, but the random disturbance term is assumed to follow a stable distribution with median zero. Estimation of the stable regression model results in an estimate of the regression coefficient β's as well as an estimate of the characteristic exponent α. The regression coefficients in this model represent what is known about the correlates of film success while at the same time permitting the variance of film success at the box office to be infinite. Estimates of this model show that the distribution of returns conditional on a movie's attributes has infinite variance and that returns to production budgets are substantially larger and returns to stars substantially lower than one would estimate using an improperly specified least-squares model (Walls, 2005b).

Stretched exponentials

Concavity in log-log plots of size against rank, also known as a parabolic power law, are interpreted as evidence of increasing returns to information in the demand for motion pictures (De Vany and Walls, 1996; Walls, 1997; Hand, 2001). Frisch and Sornette (1997) propose a multiplicative stochastic process that can explain the deviation of the data relative to a power law distribution, and Sornette (1998) provides rigorous technical details on multiplicative processes leading to power laws and stretched exponentials. Walls (2005a) finds that the stretched exponential distribution fits motion-picture revenue data remarkably well. The stretched exponential distribution does not truncate the upper tail in its estimates of the probability of a movie earning a larger amount than previous movies. The distribution also accounts for the deviation from the strict Pareto power law in a way that does not place artificial restrictions on the possibility that a movie can earn far more than our experience suggests.

Understanding the movie business

We now discuss how the behavioural and statistical models help us to understand the way the motion-picture industry operates and how contracts and business practices adjust the supply of theatrical engagements to capture the increasing returns inherent in the demand process.

The opening

Stars, large production budgets and national advertising campaigns can place a film on many exhibitor screens when it opens. This can generate high initial revenues and, if viewers like the film and spread the word, it will earn high revenues in the following weeks. But a wide release is vulnerable to negative feedback – if viewers do not like the film, the large opening audience transmits a large flow of negative information, and revenue may decline at a rapid rate. A wide release lowers the gross revenue per theatre, and this may cause exhibitors to drop the film sooner than they would otherwise. The willingness of exhibitors and downstream sources of revenue like cable television, videocassette distributors, pay per view and network television as well as foreign distributors to pay advance guarantees for motion pictures before their theatrical run is a major inducement for distributors to

produce big budget films and promote them heavily. The theatrical market can be less important than other sources of revenues (Rusco and Walls, 2004).

Decentralization

Each film's run through the market is sequential in order to exploit information dynamics. The run is self-organized because it decentralizes the decision to extend the run to each theatre and uses only local information to extend or close the run at each location. The initial release is modified over time through this process, and new engagements can be added subject to prior contractual obligations. These contractual features interact to adaptively capture revenue and generate strongly increasing returns from highly successful films. When demand has positive feedback, supply responds flexibly to allow some films to become blockbusters.

Admission pricing

Fixed admission prices (across films but within a given customer class or time of day) are a common industry practice. As a result, demand is accommodated by lengthening a film's run. A relatively stationary admission price combined with a count of admissions gives a reliable signal of demand, and this signal is transmitted throughout the industry by real time reporting of box office revenues. This reporting is required in the exhibition contract and encouraged by other means as well. If the admission price were increased to ration excess demand, the number of people who would see the film in the opening weeks would fall and this would reduce the flow of information from this source to potential viewers. This lower rate of information transfer would lead to a shorter run and a lower total level of demand. The ability to extend the run makes an almost perfectly elastic supply response possible, so there is no need for price to rise to ration excess demand. Fixed admission prices lead to a pure quantity signal and an adaptive supply response to accommodate demand discovery.

Contracting

Optimal contract theory does not fit the environment of motion pictures where expected values are dominated by the rare and unpredictable events that are so large. The incentive clauses of optimal contract theory are designed to alter the probabilities of favourable outcomes and raise expected values, but the asymmetric information often emphasized by optimal contract theory is not a factor because both principal and agent are in a state of symmetric ignorance about the prospects of a movie owing to the ‘nobody knows’ property (Caves, 2000).

A difficult problem to solve contractually is how to keep a film on screens long enough for it to build an audience. If an exhibitor takes such a film, it is with the risk that it may build so slowly during his or her run that only exhibitors who show it later will benefit from information feedback. Because the Paramount decrees bar long-term, exclusive showings, it is difficult to guarantee that the exhibitor who takes the risk of introducing the film will benefit if the film later becomes a success (De Vany and Eckert, 1991). When the audience grows recursively, the Paramount contracting restrictions may prevent risk-taking exhibitors from capturing the demand externality which they create.

Extreme events drive the business, so contracts condition pay on rare events with compensation related to the outcome of the movie. Many Hollywood superstar movie contracts contain some form of profit participation (Weinstein, 1998). Many contracts are contingent on theatrical box office revenues, which are readily monitored. In this case, the share of gross revenue paid often is nonlinear, with the share rising at higher outcomes, to reflect the nonlinear dependence of profit on revenue. In a complex contract, there may be several breakpoints where the star's percentage share increases, this nonlinearity reflecting the nonlinearity of profit in revenue.

Film rentals

The exhibition contracts are rich in contingencies that make them highly adaptive: they rely on locally generated information; they set the rental fee in a precise and nonlinear way in response to demand; they share risk between exhibitors and distributors; and they create incentives for exhibitors to show films by granting a measure of exclusivity. The rental price adapts to the state of demand and the rental schedule is nonlinear. Events in the tail are the high-revenue weeks during a movie's theatrical run, and these weeks can occur at any time during the run. During these high-revenue weeks, the rental clause allows the exhibitor to retain his or her (negotiated) cost per week of operation plus ten per cent while allocating the remaining 90 per cent to the distributor (De Vany and Eckert, 1991).

Star power

Movies with superstars have a different distribution of profit from other movies (De Vany and Walls, 2004). The profit distribution for superstar movies is an asymmetric stable distribution with infinite variance. Stars place much more mass in the upper tail of the profit distribution. The probability of extreme catastrophes – losses in excess of $95 million, say – is higher for movies *without* stars than for movies with stars. This is not at all obvious and may not be observed in a given sample. Putting a star in a movie places more mass on the upper tail and less on the lower tail. Expected profit is positive for star movies and negative for non-star movies. These values are consistent with the fact that probability is skewed to the positive tail in superstar movies and to the negative tail for others. Superstar movies are more profitable and less risky than other movies.

Success breeds success

An interesting property of the stable Paretian distribution discussed above is that conditional expectation does not converge. The tails of stable distributions are Paretian and the conditional probability that $x \geq x_0$ is $P[x > x_0] = (x_0 / x)^{\alpha}$. The conditional mean, given that $x > x_0$ equals $\bar{x}_{x_0} = x_0 \alpha / (\alpha - 1)$. Since α is a constant, the conditional expected value of profit depends linearly on x_0. Conditional on having earned a profit, the expected profit continues to rise with current profit, and this does not end as the movie earns more profit. This is not paradoxical because movies that make it into the upper tail of the profit distribution have been selected from among their competitors. The heavy tails of the stable distribution imply that probability does not decline rapidly enough for the conditional expectation to converge. For the Gaussian or log Gaussian distributions, the conditional expectation converges to a

constant as the conditioning event increases. The linear conditional expectation of the Paretian distribution means that blockbuster movies that have already attained high profit have an expectation of even higher profit, and this prospect does not diminish as profit grows. This captures the idea of demand momentum.

Conclusion

When movie audiences see a movie they like, they make a discovery and they tell their friends about it. This and other information is transmitted to other consumers, and demand develops dynamically as the audience sequentially discovers which movies it likes. Supply adapts to revealed demand through flexible exhibition contracts and other business practices that permit the increasing returns in film demand to be realized.

See Also

- information cascades
- Pareto distribution
- path dependence
- power laws
- superstars, economics of

Bibliography

Brock, W. 1999. Scaling in economics: a reader's guide. *Industrial and Corporate Change* 8, 403–46.

Caves, R. 2000. *Creative Industries: Contracts between Art and Commerce*. Cambridge, MA: Harvard University Press.

Chisholm, D. 1993. Asset specificity and long-term contracts: the case of the motion-pictures industry. *Eastern Economic Journal* 19, 143–55.

Chisholm, D. 1997. Profit-sharing versus fixed-payment contracts: evidence from the motion pictures industry. *Journal of Law, Economics, and Organization* 13, 169–201.

De Vany, A. and Eckert, R. 1991. Motion picture antitrust: the Paramount cases revisited. *Research in Law and Economics* 14, 51–112.

De Vany, A. and McMillan, H. 2004. Was the antitrust action that broke up the movie studios good for the movies? Evidence from the stock market. *American Law and Economics Review* 6, 135–53.

De Vany, A. and Walls, W.D. 1996. Bose–Einstein dynamics and adaptive contracting in the motion

picture industry. *Economic Journal* 106, 1493–514.

De Vany, A. and Walls, W.D. 1999. Uncertainty in the movie industry: does star power reduce the terror of the box office? *Journal of Cultural Economics* 23, 285–318.

De Vany, A. and Walls, W.D. 2002. Does Hollywood make too many R-rated movies?: Risk, stochastic dominance, and the illusion of expectation. *Journal of Business* 75, 425–51.

De Vany, A. and Walls, W.D. 2004. Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. *Journal of Economic Dynamics and Control* 28, 1035–57.

Epstein, J. and Axtell, R. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, MA: Brookings Institution and MIT Press.

Feller, W. 1957. *An Introduction to Probability Theory and Its Applications*. New York: Wiley.

Frisch, U. and Sornette, D. 1997. Extreme deviations and applications. *Journal de Physique* 1, 1155–71.

Hand, C. 2001. Increasing returns to information: further evidence from the UK film market. *Applied Economics Letters* 8, 419–21.

Jovanovic, B. 1987. Micro shocks and aggregate risk. *Quarterly Journal of Economics* 17, 395–409.

Kindem, G., ed. 1982. *The American Movie Industry: The Business of Motion Pictures*. Carbondale: Southern Illinois University Press.

Levy, M. and Soloman, S. 1997. New evidence for the power law distribution of wealth. *Physica A* 242, 90–4.

Mandelbrot, B. 1963. New methods in statistical economics. *Journal of Political Economy* 71, 421–40.

Mantegna, R. and Stanley, H. 1995. Scaling in financial markets. *Nature* 376, 46–9.

McCulloch, J. 1996. Financial applications of stable distributions. In *Statistical Methods in Finance*, vol. 14 of *Handbook of Statistics*, ed. G. Maddala and C. Rao. New York: North-Holland.

McCulloch, J. 1998. Numerical approximation of the symmetric stable distribution and density. In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, ed. R. Adler, R. Feldman and M. Taqqu. Berlin: Birkhäuser.

Rusco, F. and Walls, W.D. 2004. Independent film finance, pre-sale agreements, and the distribution of

film earnings. In *The Economics of Art and Culture*. Contributions to Economic Analysis No. 260, ed. V. Ginsburgh. Amsterdam: Elsevier.

Sedgwick, J. 2000. *Popular Filmgoing in 1930s Britain: A Choice of Pleasures*. Exeter: University of Exeter Press.

Sornette, D. 1998. Multiplicative processes and power laws. *Physical Review E* 57, 4811–13.

Uchaikin, V. and Zolotarev, V. 1999. *Chance and Stability: Stable Distributions and their Applications*. Utrecht: VSP.

Walls, W.D. 1997. Increasing returns to information: evidence from the Hong Kong movie market. *Applied Economics Letters* 4, 187–90.

Walls, W.D. 2000. Measuring and managing uncertainty with an application to the Hong Kong movie business. *International Journal of Management* 17, 118–27.

Walls, W.D. 2005a. Demand stochastics, supply adaptation, and the distribution of film earnings. *Applied Economics Letters* 12, 619–23.

Walls, W.D. 2005b. Modeling movie success when ‘nobody knows anything’: conditional stable-distribution analysis of film returns. *Journal of Cultural Economics* 29(3), 177–90.

Weinstein, M. 1998. Profit-sharing contracts in Hollywood: evolution and analysis. *Journal of Legal Studies* 27, 67–112.

How to cite this article

Walls, W. David. "motion pictures, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000223> doi:10.1057/9780230226203.1142

The New Palgrave Dictionary of Economics Online

multilingualism

Victor Ginsburgh and Shlomo Weber

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Multilingualism or *linguistic diversity* in a heterogeneous society provides extraordinary challenges and room for policies which may have important economic implications in shaping the flows of interregional or international trade, investment and migrations. Given the often uncompromising nature of linguistic conflicts, linguistic policies and, especially, the choice of official languages should take into account the preferences of those groups of individuals whose cultural, societal, historical values and sensibilities could be affected. In evaluating linguistic policies an important role is played by the dynamic nature of language environments driven by individual choices of learning other languages.

Keywords

communicative benefits; linguistic disenfranchisement; linguistic standardization; multilingualism; official languages

Article

Multilingualism or linguistic diversity is an important societal phenomenon that can generate gains or losses resulting from the economic interactions between individuals, regions or countries. The effects of multilingualism have recently come to the forefront of public policy debates. Linguistic issues and, in particular, the treatment of minority languages are almost unparalleled in terms of their explosiveness and emotional appeal, much more so than any other question of resource allocation or responsibility sharing within a polity. As noted by Bretton (1976, p. 447), ‘language may be the most explosive issue universally and over time. This mainly because language alone, unlike all other concerns associated with nationalism and ethnocentrism is so closely tied to the individual self. Fear of being deprived of communicating skills seems to raise political passion to fever pitch.’
Language policies in multilingual societies are beset by the trade-off between standardization and disenfranchisement. Linguistic *standardization* comprises any set of policies that promote the dominant

use of a unique or several languages while limiting the usage of languages spoken by other population groups. Indeed, linguistic standardization may deliver important benefits in terms of greater ease of communication, reducing costs of translation, increased trade, improved economic performance and administrative efficiency. However, excessive standardization may exacerbate the alienation of large minorities and widen the existing chasm between linguistic communities (Laponce, 2003). A restriction of basic linguistic rights may create *disenfranchisement* of groups of individuals and cause citizens to lose their ability to communicate in the language of their choice. Standardization, which is often represented by a selection of official languages and allocation of linguistic rights, may alienate those groups of individuals whose cultural, societal and historical values and sensibilities are not represented by the official languages (Laitin, 1989). As Pool (1991) points out, non-official languages may suffer from their 'minority status' and limit employment and advancement possibilities of their native speakers. Since in many cases it is not feasible to include all the languages in the set of official ones, a multilingual society must design some language standardization policies (for example, the 'three-language formula' in India; Baldridge, 1996) and the implementation of certain standardization measures (De Swaan, 2001; Grin, 2004). However, the explosive and uncompromising nature of linguistic conflicts, the reluctance of linguistic majorities to concede rights to minorities, makes the choice of official languages a challenging and daunting task. Thus, the choice of the set of official languages has to take into account the sensitivity of a society towards possible disenfranchisement of large groups of its citizens (Ginsburgh, Ortuño-Ortín and Weber, 2005) and has to rely on a delicate resolution of the interplay between administrative and cost efficiency, on the one hand, and the rights and desires of various linguistic groups, on the other (Van Parijs, 2005).

To illustrate the individual and aggregate cost and benefits of standardization and disenfranchisement, we consider a society M and the set of languages L spoken in this society. We assume that every citizen i is endowed with a unique native language $n(i) \in L$ and a set of languages $L(i) \subset L$ that, to simplify, she commands with identical ease. A linguistic profile of each individual i is the pair $(n(i), L(i))$, and society's linguistic profile is given by $P = (n(i), L(i))_{i \in M}$. A linguistic policy is represented by a set of official languages $K \subset L$ that is chosen for administrative, educational, and official communication functions in the society (Pool, 1991; 1996, and the extensive list of references therein; Ginsburgh, Ortuño-Ortín and Weber, 2005.) The choice of the set K represents a linguistic *standardization policy*. If the set of official languages K is non-empty and smaller than L, those members of the society whose native language is not included in K will be *disenfranchised* and some of their linguistic rights will be denied.

In order to evaluate the costs of disenfranchisement, we assume that every citizen i has utility function u_i defined over all subsets of L. We will denote $u_i(K)$ for $i \in M$ and $K \subset L$, where citizens with the same linguistic profiles have identical utility functions. It is important to stress that the functions u_i are defined over the set of languages as a whole, rather than being dissected into preferences over single languages. Though citizens may have preferences over single languages, their evaluation of the set of official languages could be crucially affected by inclusion or exclusion of their native language. The aggregate utility (welfare) function for the entire society is given by $W(u, P, K)$, where u is the vector of u_i's. Our description indicates the special role played by the native languages of citizens in M, which can be viewed as the union of linguistic clusters M_l, where, for each $l \in L$, M_l consists of citizens whose native

language is *l*. Assuming additivity of the aggregate utility, we have $W(u, P, K) = \sum_{l \in L} \sum_{i \in M_l} u_i(K)$. As a simple example, consider the *dichotomous* function based on the citizens' native languages (Ginsburgh and Weber, 2005), for which the value of $u_i(K)$ is 1 if *i* 's native language, $n(i)$, is included in *K*, and zero if it is not. The latter group contains individuals who are *disenfranchised* by the imposed standardized measures. The value taken by the function *W* is the number of citizens whose native language belongs to the set *K*, $W^1(u, P, K) = \sum_{\{i \in N | n(i) \in K\}} 1$. One generalization of the dichotomous approach is to take into account the entire language profile of every citizen rather than her native language only. Then, the value of her utility function is 1 if at least one of the languages spoken by her is included in *K* and zero otherwise. Here, the notion of disenfranchisement is limited to those who speak no official language: $W^2(u, P, K) = \sum_{\{i \in N | L(i) \cap K \neq \emptyset\}} 1$.

In evaluating citizens' preferences over subsets of languages one may take into account the similarity or the proximity between languages (see, for example, Dyen, Kruskal and Black, 1992, for a matrix of distances between 95 Indo-European languages). Let $\delta(l, l')$ be the linguistic distance between two languages *l* and l'. Denote the linguistic distance between any two subsets T, T' of *L* as the minimal distance between a language from *T* and a language from T': $\delta(T, T') = \min_{l \in T, l' \in T'} \delta(l, l')$. Then, the 'linguistic welfare' of the society is function of the distances between citizens' native languages and the set of official languages K: $W^3(u, P, K) = w(\delta(n(1), K), \delta(n(2), K), \ldots, \delta(n(M), K))$, where $w: \Re_+^M \to \Re$ is decreasing in each of its *M* arguments. Again, a modified utility function could be defined over the distances between the sets $L(i)$ and *K* instead:

$W^4(u, P, K) = w(\delta(L(1), K), \delta(L(2), K), \ldots, \delta(L(M), K))$.

Note that enlarging the set of official language is welfare improving in all four specifications above. Thus, if the only goal of the society is to maximize aggregate utility, it should set *K*=*L*. However, there are also other considerations to take into account. Difficulties of communication, costs incurred by translation and interpretation, possible errors causing delays and sometimes paralysing multilateral discussions and negotiations impose a non-negligible burden on societies with a large number of official languages (in 2007, the European Union had to manage 23 official languages at a cost over $1.5 billion). Denote then by $C(K)$ the cost of maintaining the set *K* of official languages. Obviously, C is increasing, but its specific form depends on the intensity of the linguistic regime. There could be various requirements, including a 'full' regime that every official document needs to exist in all official languages.

There is thus a trade-off between language standardization (and disenfranchisement of some citizens) and the translation, interpretation and communication costs generated by every additional official language. Formally, the society's objective is to find a set of languages *K* that maximizes the difference between aggregate utility and costs: $\max_{K \subseteq L} W(u, P, K) - C(K)$. A solution to this problem is discussed by Grin (2004, p. 201), who argues that there must be an optimum, since 'it is reasonable to assume that the benefits of diversity increase at a decreasing rate, while its costs increase at an increasing rate', and is addressed in Ginsburgh, Ortuño-Ortín and Weber (2005).

Language profiles considered so far are assumed given. In fact, they can be remarkably dynamic and change over time as individuals may decide to learn other languages. The reasons that induce citizens to do so can be analysed by examining the benefits and the costs that such learning generates. Benefits are often linked with the increased earning potential, especially in the case of immigrants who acquire the native language of the country in which they live (see, for example, MacManus, Gould and Welsch, 1978; Grenier, 1985; Lang, 1986; Chiswick, 1998; and references in Grin and Vaillancourt, 1997). We consider the Selten and Pool (1991, p. 66) ‘communicative benefits’ approach that frees itself from the restriction that ‘earnings [are] a mechanism and firms a milieu of the incentive to learn languages’. For every language l consider the set M_l of its native speakers, whose number is denoted by m_l. Assume for simplicity that $L = j, k$ and that all citizens speak only their native language, so that the linguistic profile $L(i)$ consists of $n(i)$ for every $i \in M$. Citizens may learn the other language. Denote by $m_{j,k}$,($m_{j,k}$) the number of citizens in M_j(M_k) who do so. A citizen $i \in M_j$ who learns language k incurs a cost $C(\delta(j, k))$, where C is an increasing function of linguistic distance. Let $u_j(m_j,\cdot)$ be the utility of $i \in M_j$, where the second argument indicates the number of individuals i can communicate with. We assume that the utility functions are increasing and, moreover, identical for all individuals with the same native language. If i learns k, it costs her $C_{j,k}$, but she will be able to communicate with all citizens in M_k. Her gross benefit will be given by $u_j(m_j, m_k)$. If i does not learn k, she will be able to communicate with those in M_k who learn language j, and her gross (and net) benefit will be $u_j(m_j, m_k)$. This formulation leads to the following equilibrium condition that makes individuals in M_k indifferent between learning the other language and deciding not to do so: $u_j(m_j, m_k) - C_{j,k} = u_j(m_j, m_{k,j})$. This equation allows us to determine the number of citizens in group M_k who learn j, and in a similar manner the number of those in group M_j who learn k (see Selten and Pool, 1991; Church and King, 1993; Shy, 2001; Gabszewicz, Ginsburgh and Weber. 2005; Ginsburgh, Ortuño-Ortín and Weber, 2007). By imposing some additional conditions, such as continuity, concavity and super-modularity of the utility functions one can derive some comparative statics results. In particular, one can show that the number of learners of the foreign language j in country k is positively correlated with the number of j-speakers in other countries and negatively correlated with the population size of their own country k (Lazear, 1999; Ginsburgh, Ortuño-Ortín and Weber, 2007). These results also show that public policies may be useful in stimulating learning (for a cost–benefit analysis of linguistic policies in Quebec, see, for example, Breton and Mieskowski, 1975, Vaillancourt, 1987; see also Fidrmuc and Ginsburgh, 2007, for policy suggestions in the EU).

In short, the questions raised by multilingualism offer serious challenges and the main reason is that linguistic policies are concerned not only with difficult trade-offs and resource allocation issues, but enter also the area of public policies that touch so closely personal values, beliefs and traditions.

See Also

- culture and economics
- social welfare function

We should like to thank Yuval Weber for his help in preparing this manuscript.

Bibliography

Baldridge, J. 1996. Reconciling linguistic diversity: the history and future of linguistic policies in India. Discussion paper, University of Pennsylvania.

Breton, A. and Mieskowski, P. 1975. The returns to investment in language: The economics of bilingualism. Working Paper No. 7512, Toronto Institute for the Quantitative Analysis of Social and Economic Policy, University of Toronto.

Bretton, H. 1976. Political science, language, and politics. In *Language and Politics*, ed. W.M. O'Barr and J.F. O'Barr. The Hague: Mouton.

Chiswick, B. 1998. Hebrew language usage: determinants and effects on earnings among immigrants in Israel. *Journal of Population Economics* 15, 253–71.

Church, J. and King, I. 1993. Bilingualism and network externalities. *Canadian Journal of Economics* 26, 337–45.

De Swaan, A. 2001. *Words of the World*. Cambridge: Polity Press.

Dyen, I., Kruskal, J.B. and Black, P. 1992. An Indo-European classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5).

Fidrmuc, J. and Ginsburgh, V. 2007. Languages in the European Union: the quest for equality and its cost. *European Economic Review* 51, 1351–69.

Gabszewicz, J., Ginsburgh, V. and Weber, S. 2005. Bilingualism and communicative benefits. Discussion paper, CORE, Catholic University of Louvain.

Ginsburgh, V., Ortuño-Ortín, I. and Weber, S. 2005. Language disenfranchisement in linguistically diverse societies: the case of European Union. *Journal of the European Economic Association* 3, 946–65.

Ginsburgh, V., Ortuño-Ortín, I. and Weber, S. 2007. Learning foreign languages: theoretical and empirical implications of the Selten and Pool model. *Journal of Economic Behavior and Organization* 64, 337–47.

Ginsburgh, V. and Weber, S. 2005. Language disenfranchisement in the European Union. *Journal of Common Market Studies* 43, 273–86.

Grenier, G. 1985. Bilinguisme, transferts linguistiques et revenus du travail au Québec, quelques éléments d'interaction. In *Economie et Langue*, ed. F. Vaillancourt. Québec: Editeur officiel.

Grin, F. 2004. On the costs of cultural diversity. In *Cultural Diversity versus Economic Solidarity*, ed. F. Van Parijs. Brussels: De Boeck Université.

Grin, F. and Vaillancourt, F. 1997. The economics of multilingualism: overview of the literature and analytical framework. In *Multilingualism and Multilingual communities*, ed. W. Grabbe. Cambridge, MA: Cambridge University Press.

Laitin, D. 1989. Language policy and political strategy in India. *Policy Sciences* 21, 415–36.

Lang, K. 1986. A language theory of discrimination. *Quarterly Journal of Economics* 100, 363–81.

Laponce, J.A. 2003. Minority languages and globalization. *Nationalism and Ethnic Policies* 10, 15–24.

Lazear, E. 1999. Culture and language. *Journal of Political Economy* 107, 95–126.

MacManus, W., Gould, W. and Welsch, F. 1978. Earnings of Hispanic men: the role of English language proficiency. *Journal of Labor Economics* 1, 101–30.

Pool, J. 1991. The official language problem. *American Political Science Review* 85, 495–514.

Pool, J. 1996. Optimal language regimes for the European Union. *International Journal of Sociology of Language* 121, 159–79.

Selten, R. and Pool, J. 1991. The distribution of foreign language skills as a game equilibrium. In *Game Equilibrium Models*, vol. 4, ed. R. Selten. Berlin: Springer-Verlag.

Shy, O. 2001. *The Economics of Network Industries*. Cambridge: Cambridge University Press.

Van Parijs, P. 2005. Europe's three language problems. In *The Challenge of Multilingualism in Law and Politics*, ed. D. Castiglione and C. Longman. Oxford: Hart Publishing.

Vaillancourt, F. 1987. The benefits and costs of language policies in Quebec, 1974–1984: some partial estimates. In *The Economics of Language Use*, ed. H. Tonkin and K. Johnson-Weiner. New York: Center for Research and Documentation on World Language Problems.

How to cite this article

Ginsburgh, Victor and Shlomo Weber. "multilingualism." The New Palgrave Dictionary of Economics.

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000411> doi:10.1057/9780230226203.1143

The New Palgrave Dictionary of Economics Online

multiple equilibria in macroeconomics

Costas Azariadis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The multiple equilibrium literature seeks explanations for excessive economic volatility, persistent poverty, market fads and fashions, and related macroeconomic phenomena that appear to be anomalies in standard models of rational economic behaviour. Terms like *animal spirits, sunspots, irrational exuberance, indeterminacy*, and *bubbles* describe situations of multiple equilibrium. All such ideas assert that future values of macroeconomic states cannot be predicted accurately from knowledge of economic fundamentals. This article describes four types of multiple equilibria common in macroeconomics (missing initial conditions, multiple laws of motion, multiple attractors, and non-fundamental state variables), discusses their causes and reviews what they teach us about economic policy.

Keywords

animal spirits; bubbles; extraneous random variables; imperfect asset markets; income effect; increasing returns; indeterminacy; jump variables; missing initial conditions; multiple equilibria in macroeconomics; multiple laws of motion; overlapping generations models; public debt; stable manifolds

Article

The multiple equilibrium literature seeks explanations for excessive economic volatility, persistent poverty, market fads and fashions, and related macroeconomic phenomena that appear to be anomalies in standard models of rational economic behaviour. Terms like *animal spirits, sunspots, irrational exuberance, indeterminacy*, and *bubbles* describe situations of multiple equilibrium. All of these ideas assert that future values of macroeconomic states cannot be predicted accurately from current values of these states or from knowledge of economic fundamentals, even if households and firms behave with complete rationality.

Most of the economics research community has been sceptical of multiple equilibrium (cf. McCallum, 1990), believing that it undermines the comparative statics and comparative dynamics exercises that are essential for policy evaluation and econometric prediction. Is it unreasonable, ask the sceptics, to know how the economy selects one equilibrium when many are possible, and how the expectations of economic actors settle on that particular outcome?

Economists have to weigh these legitimate reservations against direct evidence from laboratory experiments that beliefs do matter (Duffy and Fisher, 2005) as well as against the continuing difficulties of unique equilibrium models to come to grips with an expanding array of empirical anomalies in many sub-fields of macroeconomics, from excessively volatile asset prices and exchange rates to persistent underdevelopment. This article describes briefly four types of multiple equilibria common in macroeconomics, discusses what causes them, and reviews briefly what they teach us about economic policy.

Typology and examples

Multiple equilibria occur in dynamic economies whenever the laws of motion that describe macroeconomic states over time admit more than one solution sequence or, more broadly, several asymptotic states. The simplest mathematical example is a set valued, piecewise linear, deterministic law of motion for a scalar state variable $x(t)$, expressed in terms of a vector $v = (A, B, m, a, b)$ of fundamental parameters:

$$x(t+1) = f(x(t), v) = mx(t) + a \text{ if } 0 < x(t) < A = g(x(t), v) = mx(t) + b \text{ if } B < x(t) \quad (1)$$

for all $t = 0, 1, \ldots,$ *with* $0 < m < 1, 0 < A, 0 < B, 0 < a < b$, *and possibly some initial condition* $x(0) > 0$ *fixed by history*.

For different values of the parameter vector v, eq. (1) illustrates explicitly three major types of multiple equilibria: *indeterminacy from missing initial conditions*, *indeterminacy from multiple laws of motion*, and *multiple attractors*. A fourth type, *non-fundamental state variables* or *sunspots*, occurs when we randomly combine the two laws of motion f and g. All four types are associated with excessively volatile behaviour, that is, with macroeconomic states exhibiting abnormal sensitivity to small changes in fundamentals.

Missing initial conditions is the simplest and best-known type of indeterminacy. Suppose, for example, that there is a unique law of motion f, that is, the parameters A and B are infinitely large. If $x(0)$ is an initial price or, more generally, a *jump variable* that is not predetermined by history but emerges instead from forward-looking markets, then there is a one-dimensional continuum of solutions $x(t,a)$ to eq. (1) indexed on the indeterminate initial condition $x(0)$:

$$\log(x(t, a) - a/(1-m)) = t \log m + \log(x(0) - a/(1-m)) \quad (2)$$

More generally, an indeterminacy with $S - I$ degrees of freedom appears in any dynamic economy when: (*a*) history predetermines I initial conditions; (*b*) the law of motion has S stable eigenvalues; and (*c*) $I < S$. Equation (2) illustrates the case $(S, I) = (1, 0)$. A major set of economic examples for this kind of multiplicity comes from overlapping generations models. Fiat money in a dynamically inefficient exchange economy (Wallace, 1980) has an indeterminate steady state with worthless money at which $(S, I) = (1, 0)$ because history does not fix the initial price of money. Public debt in a dynamically inefficient production economy (Diamond, 1965) leads to an indeterminate steady state, with worthless public debt and $(S, I) = (2, 1)$ because the price of debt is also a jump variable. Finally, two-sector growth environments (Galor, 1992), in which the distribution of capital between sectors is again a jump variable, exhibit indeterminacy with $(S, I) = (2, 1)$ whenever the consumption good is more capital-intensive than the investment good.

Multiple laws of motion describe a less understood but more pernicious kind of indeterminacy that arises even if there are no jump variables. Examples of this phenomenon are growth models with private information or limited enforcement (Azariadis and Smith, 1998; Azariadis and Kaas, 2008) as well as Markov switching models in time-series econometrics and empirical finance (Hamilton, 1994). To illustrate, let us choose the parameter vector v in eq. (1) so that

$$(1-m)B < a < b < (1-m)A$$

(3)

Then the two laws of motion, *f* and *g*, overlap in the interval (*B*, *A*); each of them has a steady state, $a / (1 - m)$ and $b / (1 - m)$ respectively, which is a suitable initial condition for the *other* law. If *x*(*t*,*a*) and *x*(*t*,*b*) are dynamic equilibria for the two laws in eq. (2), then for any initial condition *x*(0) in the interval (*B*, *A*), we can write down a deterministic general solution *z*(*t*) that combines regimes *f* and *g* in *any arbitrary time sequence*, that is,

$$z(t) = x(t, a) \quad \text{for some } t = x(t, b) \quad \text{for all other } t$$

(4)

For each *x*(0), we may freely select either regime in each time period. In particular, choosing the same regime every period leads to the steady state of that regime; switching regimes periodically leads to deterministic periodic cycles, as in Grandmont (1985), and so on.

Sunspot equilibria are mixtures of multiple deterministic equilibria – static ones as in Cass and Shell (1983) or dynamic ones as in Azariadis (1981) – connected by a *non-fundamental* or *extraneous* random variable. Market sentiment, investor beliefs, and consensus forecasts are three examples of extraneous random variables which often take on more colourful names like ‘animal spirits’, ‘sunspots’ or ‘self-fulfilling prophecies’. A simple illustration of a non-fundamental state variable is a lottery *s*(*t*) played each period over the intercept, *a* or *b*, of the two laws of motion in eq. (1). For instance, if *s*(*t*) is a two-state Markov process, then $s(t) = s(t - 1)$, with probability *p*(*a*) if $s(t - 1) = a$, and with probability *p*(*b*) if $s(t - 1) = b$. The general stochastic solution *Z*(*t*, *s*(*t*)) to eq. (1) shows how outcomes depend on the non-fundamental macroeconomic state *s*(*t*). Specifically,

$$\text{If } s(t - 1) = a, \text{ then } z(t, s(t)) = x(t, a) \text{ w. p. } p(a) = x(t, b) \text{ w. p. } 1 - p(a) \text{If } s(t - 1) = b, \text{ then } z(t, s(t)) = x(t, a) \text{ w. p. } 1 - p(b) = x(t, b) \text{ w. p. } p(b)$$

(5)

The last type of non-uniqueness, *multiple attractors*, describes environments with several asymptotic states. Here long-run values of macroeconomic states depend on the corresponding initial values, as in Murphy, Shleifer and Vishny (1989), Azariadis and Drazen (1990), and Matsuyama (1991). We call these environments ‘non-ergodic’ or ones in which ‘history matters’. For example, suppose we pick the parameter vector *v* in eq. (1) to eliminate the overlap between regimes *f* and *g*, and obtain one piecewise linear law of motion. Specifically, we replace (3) by

$$a < (1 - m)A < (1 - m)B < b$$

(6)

Then, for each initial *x*(0), the general deterministic solution *z*(*t*) to eq. (1) is a unique step function, which traces the law *f* up to $x = A$, and jumps to the other law

g at that point. Mathematically,

$$z(t) = x(t, a) \text{ if } z(t-1) < A = x(t, b) \text{ if } z(t-1) > A \tag{7}$$

Equilibrium here is completely determinate and utterly predictable if history fixes $x(0)$, but the asymptotic state is $a/(1-m)$ if $x(0) < A$ and $b/(1-m)$ if $x(0) > A$. History *matters* in this situation because small or temporary shocks to the macroeconomic state $z(t)$ can have substantial and long-lasting consequences if that state is anywhere near the critical value A.

Causes

Dynamic inefficiency and dynamic complementarities are the two most common proximate causes of multiple equilibrium in macroeconomic models. Dynamic inefficiency is a property of economies with very patient consumers who are energetic savers at low interest rates. For example, holders of short-term US Treasury bills in the last 50 years seem content with an average real pre-tax annual yield of about one per cent. Very patient savers are willing to invest in *bubbles*, paying top dollar for assets with low dividends. Bubbles themselves (Tirole, 1985; Shiller, 1989) are notoriously indeterminate objects in their initial conditions and laws of motion; they may deflate now, later or not at all, depending on investor sentiment.

Economies with externalities, increasing returns and, most notably, imperfect asset markets often exhibit complementarities in production or consumption which cause excess demands for consumption goods and productive factors to bend backward instead of sloping downward. The typical outcome is several steady states and several laws of motion or *stable manifolds*, each one leading to a distinct asymptotic state. In particular, multiple equilibria occur when externalities or increasing returns link the payoffs of each agent with the actions of others, both in strategic environments (Cooper and John, 1988) and in competitive ones (Benhabib and Farmer, 1994). Producers, for example, find it advantageous to raise, hold steady, or lower output in tandem with their industry or the whole national economy.

Imperfect asset markets, especially restrictions on debt and short sales (Bewley, 1986; Kehoe and Levine, 1993; Kiyotaki and Moore, 1997) are an intellectually bountiful and empirically compelling source of complementarities in consumption. This literature motivates restrictions on short sales by the collateral requirements of creditors and, more generally, as a deterrent to debtor default. Short-sales constraints depend on the excess payoff of solvency (which guarantees unfettered participation in future asset markets) over default (which restricts trading in future asset markets). Constraints on short sales are tighter the smaller this excess payoff is because smaller excess payoffs strengthen the temptation to default.

Debt constraints cause two dynamic complementarities in consumption, one through prices and the other through quantities (Azariadis and Kaas, 2007). Either one may be sufficient to overcome the intertemporal substitution effect embedded in the consumer's utility function. Specifically, price changes create a dynamic complementarity when the ordinary income effect is amplified by a relaxation of binding short-sale restrictions. The same outcome is achieved by quantity changes when an anticipated relaxation of future constraints increases the current payoff to solvency, and to continued market participation, thus slackening today's constraints.

Lessons for policy

What is the function of economic policy in a deterministic world of many steady states like the one described in eq. (7)? What should policy do in the stochastic world of eq. (5) where non-fundamental variables like beliefs, forecasts, consumer sentiment, ‘sunspots’, or ‘animal spirits’ could be every bit as important as fundamentals? Dynamic economies with several asymptotic states have two special properties: long-run performance depends on the starting state $x(0)$; and

temporary shocks may have permanent consequences. Any economy that is headed towards an inferior or undesirable steady state may be shocked temporarily until it finds a path leading to a more desirable state. In growth models with many asymptotic states, these shocks are easy to achieve in principle via short-lasting gifts of physical or human capital, by forgiving international debt, and so on. The US-supported Marshall Plan for Europe did exactly that in the 1940s and 1950s. Africa seems in need of a similar plan now but the internal situation in that continent is more problematic than Europe's was at the end of the Second World War. A bigger conceptual, as distinct from political, challenge is to formulate policies appropriate for environments swayed by non-fundamental variables and vulnerable to spurious volatility. If equilibria were well described by the stochastic process of eq. (5), could we find an economic policy to eliminate the unnecessary randomness, and bolster among consumers the belief that the economy is headed toward the more desirable of the two steady states, say, $b/(1 - m)$? Viewing economic policy as *equilibrium selection* is fairly widespread in the monetary policy literature (Woodford, 2003), and broadly consistent with monetary neutrality. On this view, credible monetary policy may be unable to influence the set of possible long-run equilibria, but it does bear on which one the economy selects. In eq. (5), for example, reactive policy rules may be unable to change the laws of motion f and g but they can still deliver the long-run state $b/(1 - m)$ if they influence the public's beliefs about the long-run likelihood of each state. All it takes to achieve the high state is nudging the two mixing probabilities, $p(a)$ towards zero and $p(b)$ towards 1.

See Also

- animal spirits
- bubbles

Bibliography

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25, 380–96.

Azariadis, C. and Drazen, A. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105, 501–26.

Azariadis, C. and Kaas, L. 2008. Credit and growth under limited commitment. *Macroeconomic Dynamics* 12(Supp. 1).

Azariadis, C. and Smith, B. 1998. Financial intermediation and regime switching in business cycles. *American Economic Review* 88, 516–36.

Benhabib, J. and Farmer, R. 1994. Indeterminacy and increasing returns. *Journal of Economic Theory* 63, 19–41.

Bewley, T. 1986. Dynamic implications of the form of the budget constraint. In *Models of Economic Dynamics*, ed. H. Sonnenschein. New York: Springer Verlag.

Cass, D. and Shell, K. 1983. Do sunspots matter? *Journal of Political Economy* 91, 193–227.

Cooper, R. and John, A. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103, 441–64.

Diamond, P. 1965. National debt in a neoclassical growth model. *American Economic Review* 55, 1126–50.

Duffy, J. and Fisher, E. 2005. Sunspots in the laboratory. *American Economic Review* 95, 510–29.

Hamilton, J. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Galor, O. 1992. A two-sector overlapping generations model. *Econometrica* 60, 351–86.

Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53, 995–1045.

Kehoe, T. and Levine, D. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60, 865–88.

Kiyotaki, N. and Moore, J. 1997. Credit cycles. *Journal of Political Economy* 105, 221–48.

Matsuyama, K. 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. *Quarterly Journal of Economics* 106, 617–50.

McCallum, B. 1990. New classical macroeconomics: a sympathetic account. In *The State of Macroeconomics*, ed. S. Honkapohja. Oxford: Basil Blackwell.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Industrialization and the big push. *Journal of Political Economy* 97, 1003–26.

Shiller, R. 1989. *Market Volatility*. Cambridge, MA: MIT Press.

Tirole, J. 1985. Asset bubbles and overlapping generations. *Econometrica* 53, 1499–528.

Wallace, N. 1980. The overlapping generations model of fiat money. In *Models of Monetary Economies*, ed. J. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Azariadis, Costas. "multiple equilibria in macroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000374> doi:10.1057/9780230226203.1144

The New Palgrave Dictionary of Economics Online

multiplier analysis

Edward J. Amadeo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

equilibrium; excess demand; expectations; Hicks, J. R.; Kalecki, M.; Keynes, J. M.; investment; investment multiplier; multiplier analysis; real wage rate; saving; saving equals investment; shifting equilibrium; Wicksell, J. G. K.

Article

What is the effect of a change in the level of investment? Wicksell (1935) was the first economist to pose this question explicitly in the context of his ‘pure credit economy’. Voluntary or anticipated saving is not a requirement if the banking system is willing to supply the necessary credit to finance an increase of investment demand. The effect of this increase of investment demand is an increase in the level of prices (if the level of output is fixed or given), or output if there is idle capacity and unemployed labour. In his *Treatise on Money* (1930) Keynes analyses the same question. Just as in Wicksell's model, in the *Treatise*, investment is independent of current saving. The effects of a change of investment are studied through the *Treatise*'s ‘Fundamental Equations’ according to which a difference between current (or voluntary) saving and investment will give rise to a change in the price level. It is a pure excess demand effect. Changes in the price level will lead to unforeseen (or windfall) profits or losses which, in turn, will affect producers’ next period decision to produce and employ. Windfall profits will have the effect of inducing producers to increase the level of output; losses will have the opposite effect. The effect may not be as mechanical as described here if new informations (concerning, for example, changes in economic policies) come into the picture.
Book IV of the *Treatise* studies the ‘credit cycle’, that is, the effects of changes in monetary or banking policies on the rate of interest which may have an effect on the decisions to save and invest, and therefore, on the price and output levels. Changes in both the price and output levels are seen as deviations from their long-period or equilibrium counterparts; they are short-period or disequilibrium levels of price and output which, so to speak, oscillate around the equilibrium as defined by the equality between voluntary saving and investment. However, just as in Wicksell's analysis, once the system deviates from the equilibrium position, very little is said in terms of the path towards a new equilibrium;

indeed, the latter is not really determined.

Multiplier analysis is very much related to the adjustment process described above. The real *differentia* is that it focuses predominantly on the notions of stability and equilibrium of the process. The most important contributors for the development of the multiplier analysis were Kahn (1931), Keynes (1936) and Kalecki (1971).

The multiplier as an exercise in comparative statics

Let us consider the effects of a change in the level of investment which is known to all the relevant agents of the economy. Also let us temporarily assume that producers of consumption goods fully anticipate the effects of this change in investment on the demand for their products. An increase in the level of investment demand implies a greater level of production of capital goods. The degrees of capacity utilization and employment in the capital goods sector increase, thus leading to higher profits and a greater wage bill. Part of the extra profits and wages earned will be spent in consumption goods; the rest will be saved. The share of profits and wages spent in consumption goods are determined respectively by the propensities to consume out of profits and wages. These, according to Keynes (1936, chs 8 and 9), depend on objective factors (other than income) such as the money wage rate and agents' rates of time-discounting, and subjective factors such as precaution and avarice.

Thus the main effect of an increase in investment is that it induces an increase in consumption, saving, and income. The final effect on the level of income will depend essentially on the propensity to consume of the economy. The greater the propensity to consume, the greater will be the increase in the demand for consumption goods resulting from an initial increase in the income generated in the capital goods sector. The immediate effect on the demand for consumption goods will be given by $C=cI$ where C and I are respectively the levels of consumption and investment, and c is the weighted average of the propensities to consume out of profits and wages. The immediate effect on the level of income will be given by $\Delta Y = \Delta I + c\Delta I$. Note that a second round of the multiplier process will lead to an increase in the level of income given by $\Delta Y = \Delta I + c\Delta I + c^2\Delta I$. In the limit the effect will be given by $\Delta Y = \Delta I + c\Delta I + c^2\Delta I + \ldots = [1/(1-c)]\Delta I$. The term $1/(1-c)$ is called the investment multiplier. According to Keynes, the multiplier 'tells us that, when there is an increment of aggregate investment, income will increase by an amount which is $[1/(1-c)]$ times the increment in investment' (Keynes, 1936, p. 115).

Note that the change in the level of saving (ΔS) is given by the propensity to save ($s=1-c$) times the level of income, that is, $\Delta S = s\Delta Y$, which, according to the above analysis, is also equal to the initial change in the level of investment. Thus, through the multiplier mechanism, a change in the level of investment gives rise to an equal level of saving. The multiplier is essentially an equilibrating mechanism. It refers to the adjustment of the economy given an exogenous change, and it determines the equilibrium levels of income and saving associated with different levels of investment demand. It describes the changes in the level of consumption which eventually makes the latter compatible to each level of investment given the propensity to consume of the economy.

The essential difference between the multiplier mechanism and the description of credit cycles found in Keynes's *Treatise on Money* as well as in the analyses of Wicksell and the Swedish economists (Ohlin

and Lindhal for example), is that it emphasizes the notion of equilibrium. It determines the new equilibrium configuration associated with any change in the level of investment demand rather than only its immediate effects. Because it is an equilibrating mechanism it must also take into account the stability conditions of the process. In terms of the simple static version discussed above, the only stability condition is that the propensity to consume must be smaller than one. If it were greater than one the system would always explode either to a situation of full employment or zero-employment of the labour force and capacity utilization. As noted by Keynes, ‘if the [community] seek to consume the whole of any increment in income, there will be no point of stability and prices with rise without limit’ (Keynes, 1936, p. 117). However, since the propensity to consume is always positive, the multiplier is always greater than one which implies that fluctuations in investment will lead to fluctuations of income of greater magnitude. Thus, the workings of the multiplier mechanism itself may be regarded as a source of instability.

The multiplier as an exercise in dynamics

What makes the analysis of the above section static is the fact that it emphasizes the equilibrium configuration associated with a given (and known) level of investment, and a given propensity to consume. The decision to consumer is rather passive and taking it into account does not really make the analysis dynamic. What is most important, however, is that the decisions to produce are not considered. Production takes time, and therefore decisions to produce involve expectations over a period of time. A dynamic approach to the analysis of the multiplier should emphasize the role of time and expectations associated with the decisions to produce.

What is the appropriate time unit for the analysis of the multiplier process if decisions to produce are to be considered explicitly? Following Keynes we shall take the short period as the appropriate time unit. The short period is associated with ‘daily’ decisions, and daily here stands ‘for the shortest interval of time after which the firm is free to revise its decisions as to how much employment to offer’ (Keynes, 1936, p. 47). Producers make their decisions as to how much to produce based on their short-period expectations.

On the demand side the object of such expectations are either the expected sale-proceeds or the expected price, that is, the price which the producer expects to get for his product at the end of the period of production. Let us take the expected price as the relevant variable, and assume that the producer knows the remuneration rates of the variable inputs and the shape of his cost curve. Given this information we may assume that the producer goes through the following optimization exercise in order to determine the levels of output and employment: max $E[p]X-wN$ st. $X=F(N, K)$ where $E[p]$ is the expected price, X and N are the levels of output and employment respectively, w is the money-wage rate, K is the stock of capital (assumed to be given), and F is a production function. The level of employment associated with the expected price must satisfy the following condition: $w/E[p] = F'(N^*)$. The level of output is obviously $X^*=F(N^*)$.

Let us assume that the level of investment has been stable for a rather long period of time. Producers of consumption goods know not only the level of investment but also the demand for their products associated with this level. Therefore they are able to form correct expectations concerning the demand for their products, and their price. In short, in each and every period the expected price corresponds to

the market price, that is, $E[p]=p$. We now let the level of investment increase but assume that the producers of consumption goods either do not know that the change has taken place or the effect of the change on the demand for their products. If the latter is the case, assume that they underestimate the effect on demand. In either case the actual price will be greater than the expected price associated with the predetermined level of output (X^*), that is, $p>E[p]$ where p is the market price. In this example producers will experience a windfall profit given by $Q=(p-E[p])X^*$. The same exercise could be carried on taking stocks rather than the price as the adjustment variable (see Hicks, 1974, ch. 1).

The process initiated with a change in investment demand could go on for a long period. Producers would continue to get their expectations wrong, profits or losses would appear, new decisions would be taken and so on. Will producers ever get their prices (and production decisions) right? If we assume that the level of investment will not be affected by changes in short-period expectations, and depending on the way producers form their expectations, they will eventually converge to an equilibrium position. If, for example, producers form their short-period expectations in an adaptive fashion, for certain values of the parameters of the expectation function, the system will converge to a position of rest. For other values of the parameters the system will not converge. This only implies that the way producers form their expectations may affect the stability of the multiplier process and the trajectory of the relevant variables.

Does the way producers form their expectations affect the equilibrium configuration? The answer here is no. If the level of investment is assumed to be given and the process is assumed to be stable (which, again, depends on the parameters of the expectation function), the equilibrium configuration will be exactly the same as the one associated with a process in which producers form their expectations in a rational fashion. By ‘rational’ here we mean that expectations are recurrently correct. Keynes was aware of this result: in his lecture notes written in 1937 he argued that his principle of effective demand is substantially the same independently of the way expectations are formed (see Keynes, 1973, pp. 180–1).

The multiplier and the notion of ‘shifting equilibrium’

So far we have examined the multiplier mechanism assuming that either the level or the expected level of investment is given. In both the static and dynamic analyses the multiplier tells us the levels of income and saving compatible with a given level or expected level of investment. The advantage of these approaches to the multiplier is that they emphasize the notion of equilibrium, that is, they provide a definite result to the effect of a change in investment.

However, once the notion of equilibrium has become clear, we should turn our attention to the interactive relation between the level of investment and the workings of the multiplier. The level of investment is quite a volatile variable. Long-period expectations (which play a central role in the determination of the level of investment) change for various reasons. They change due to changes in the political or international environments; due to changes in economic policies; or due to objective problems of individual industries which tend to affect the expected performance of other industries of the economy. To different states of long-period expectations there corresponds different levels of investment and, therefore, different ‘levels of long-period employment’ (Keynes, 1936, p. 48). The extent to which short-period expectations are fulfilled may also affect the level of investment. If the actual demand is persistently greater than the expected demand, producers will tend to revise their long-

period expectations and investment decisions.
We may associate the notion of ‘shifting equilibrium’ with the evolution of the economic system as determined by different states of long-period expectations, and therefore, characterized by a sequence of equilibrium configurations of income and saving. By shifting equilibrium Keynes meant ‘the theory of a system in which changing views about the future are capable of influencing the present situation’ (1936, p. 293).

Distribution and the multiplier

The relationship between the distribution of income (or the real wage) and the multiplier depends on assumptions about the exogeneity or endogeneity of the real wage. In the *General Theory* Keynes assumed perfect competition *cum* profit maximization and decreasing marginal returns which, for a given money-wage rate, implies that the real wage is endogenously determined. It also implies that the greater the levels of employment and output, the smaller the real wage. This result has an important implication for the workings of the Keynesian multiplier. If we assume – as Keynes and Kalecki usually do – that the propensity to consume out of wages is greater than the propensity to consume out of other types of incomes (profits, interests, and so on), as the level of income increases and the real wage falls, the value of the multiplier decreases. Keynes pointed out to this result in the *General Theory*:

> the increase of employment will tend, owing to the effect of diminishing returns, … to increase the proportion of aggregate income which accrues to the entrepreneurs, whose … propensity to consume is probably less than the average for the community as a whole. (1936, p. 121)

Kalecki (1971) assumed constant marginal returns and gave up profit maximization. Instead he assumed that firms determine their prices through a markup over variable costs which, in a closed economy, also determines the real wage. Therefore, according to Kalecki, the real wage is exogenously determined, and does not change as the levels of output and employment change. This means that the multiplier does not change either as the level of output changes; it depends on the propensity to consume out of wages and profits and the level of the markup, both assumed to be constant over the cycle.

See Also

- multiplier–accelerator interaction

Bibliography

Hicks, J.R. 1974. *The Crisis in Keynesian Economics*. Oxford: Blackwell.

Kahn, R. 1931. The relation of home investment to unemployment. *Economic Journal* 41(June), 173–98. Reprinted in R. Kahn, *Selected Essays on Employment and Growth*. Cambridge: Cambridge University

Press, 1972.

Kalecki, M. 1971. *Selected Essays on the Dynamics of the Capitalist Economy*. Cambridge: Cambridge University Press.

Keynes, J.M. 1930. *A Treatise on Money. Vol. 1: The Pure Theory of Money*. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1973. *The Collected Writings of John Maynard Keynes*, vol. 14, ed. D.E. Moggridge and E. Johnson. London: Macmillan for the Royal Economic Society.

Wicksell, K. 1935. *Lectures in Political Economy*, vol. 2. London: Routledge & Kegan Paul.

How to cite this article

Amadeo, Edward J. "multiplier analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000274> doi:10.1057/9780230226203.1146

The New Palgrave Dictionary of Economics Online

multiplier – accelerator interaction

A. Medio

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The phrase ‘multiplier–accelerator’ refers to a combination of a theory of income as determined by investment and a theory of investment as determined by the rate of change of income.

The concept of multiplier is usually attributed to Richard Kahn (1931), from whom it was adopted by Keynes and used as a building block for his General Theory. The idea was probably shared by a number of European economists in the Thirties and was certainly known to Michael Kalecki, independently of Keynesian influence.

The theory of multiplier in its pure (and static) form can be described thus. In a capitalist economy, investment can always be realized in real terms. The necessary saving will be made available by means of corresponding variations of the level of income, given the propensity to save. With generally underutilized capacity and labour and fixed prices – the most common hypothesis – *real* income will take whatever value generates a flow of saving equal to planned investment. Alternatively, in the presence of supply constraints, the level of prices will adjust and deflate consumption expenditure so as to make available the real resources required for investment.

The former, ‘fixprice’ version of this simple relation can be stated in the form of algebraic equations, as follows

$$Y = C + I, \tag{1}$$

$$C = cY, \quad c = 1 - s \tag{2}$$

$$I = \bar{I} \tag{3}$$

where *Y*, *C*, *I* indicate, respectively, actual income, consumption and investment; $\bar{I}$ is desired investment; *c* and *s* are the propensities to consume and to save, respectively.
Elementary manipulation yields:

$$Y = (1 / s)\bar{I} \tag{4}$$

where (1/*s*) measures the multiplier and the causal relation runs from right to left.
The concept of accelerator appeared in the economic literature much earlier than the *General Theory* and was perhaps first developed by Aftalion (1909) and J.M. Clark (1917). It is based on the idea that the relation between productive capacity (somewhat measured by a scalar quantity, the capital stock) and production can vary only within narrow limits and, in a first approximation, may be taken as a constant.
The constancy of the capital–output ratio may be defended on the basis of two main arguments:
(i) Technical coefficients are fixed (or change little) even though the interest rate may vary: in economists' parlance, the isoquants are L-shaped. Whatever the plausibility of this hypothesis may be from an engineering point of view, it is difficult to accept it on economic grounds. Indeed, when 'capital' is a vectorial quantity (i.e. a list of different capital goods), the capital–output ratio depends both on technical coefficients and on relative prices and the rate of interest.
(ii) Technical coefficients vary, within a certain technology, as functions of the rate of interest. If the latter is constant so are the former.
The assumption on (ii) may be accepted or rejected for lack of realism but is formally correct. On the other hand, it is also consistent with the fix-price approach to income determination. In its starkest form, the accelerator (Harrod called it 'the Relation') can be described by the equation

$$K = vY, \tag{5}$$

(where *K* indicates the capital stock and *v* the desired capital–output ratio) or, in its incremental form

$$I = v\dot{Y} \tag{6}$$

where an overdot indicates the derivative with respect to time.
The idea came naturally to combine multiplier and accelerator and derive a model ‘complete’ in the sense that, given initial conditions, it determines the time evolution of capital stock and income. This was first attempted in the late 1930s by Harrod (1936, 1939) and, in a more mathematical manner, by Samuelson (1939). In the subsequent years, a substantial part of the literature on cycle and growth was also based on the interaction between multiplier and accelerator.
In order to discuss this idea formally, let us couple equations (4) and (6). We shall obtain

$$sY = v\dot{Y} \tag{7}$$

and

$$(\dot{Y} / Y) = (s / v) \tag{8}$$

Equation (8) represents the proportional rate of growth of income as a function of the propensity to save and the acceleration coefficient and was first investigated by Harrod and Domar, after whom it has been named ever since.
The model described by equations (1)–(8) implicitly assumes that equality always holds between demand (=consumption+investment) and supply (=income), as well as between actual and desired consumptions, the results may become drastically different. This line of research was pursued early by Samuelson, Hicks and, in an apparently very different context, Kalecki, and provided the basis for a theory of the trade cycle which prevailed in the economic profession in the early post-World War II years (the best reference is perhaps Phillips, 1954).
Suppose that, while desired and actual consumption are still equal, discrepancies are permitted to exist between demand and supply and between actual and desired investment. We therefore need to replace the relevant equilibrium conditions

$$\dot{Y} = C + I \text{(or equivalently, } sY = I)$$

and

$$I = v\dot{Y}$$

by adjustment mechanisms which reflect economic agents' reactions to undesired situations.
The most commonly used such adjustments are those of a tâtonnement type, according to which the relevant variables change at a rate proportional to the differences between their desired and actual values. In terms of our model, we have

$$\dot{Y} = \tau_Y[(C + I) - Y] = \tau_Y[I - sY], \tag{9}$$

$$\dot{I} = \tau_Y\left[v\dot{Y} - I\right], \tag{10}$$

where τ_Y and τ_i are the (positive) speeds of adjustment of income and investment. The equation (9) can be interpreted as a (typically Keynesian) situation in which, prices being fixed and potential supply unlimited, producers are constrained only by demand and adjust their production in relation to (positive or negative) excess demand.
The system (9)–(10) can be easily transformed into a single second-order differential equation in *Y*. By choosing the arbitrary unit of measure of time such that τ_i=1 we have

$$\ddot{Y} + [1 + \tau_Y s = \tau_Y v]\dot{Y} + \tau_Y sY = 0 \tag{11}$$

System (11) has a unique position of stationary equilibrium at Y=0. (‘Zero’ must be taken here to indicate a level of income determined by factors not considered in the present discussion, such as government expenditure.) Its dynamic behaviour depends on the structural coefficients and may induce decline or growth, in either case with or without fluctuations.

Generally speaking, we may say that the accelerator is an explosive factor in so far as, for given s and τ_Y the greater v the more likely it is for the system to grow in time. Moreover, the relative size of the accelerator affects the oscillatory behaviour of the system: if the motion is damped, a large v tends to make the system fluctuate; vice versa, if the motion is explosive, a strong acceleration leads to sustained growth without fluctuations. In agreement with intuitive considerations, large speeds of adjustment tend to produce explosive behaviour, whereas the saving ratio acts as a damper. The effect of these factors on oscillations is more complicated and cannot be ascertained in any obvious way.

A very special and unlikely case arises when we have

$$1 + \tau_Y(s - v) = 0 \tag{12}$$

and the time path followed by the system is a pure sinusoid describing a persistent and perfectly regular cycle, neither damped nor explosive. This of course is a watershed situation which would be destroyed by any small perturbation of the model and is therefore not a suitable idealization of economic cycles.

The multiplier–accelerator model constitutes a rough but effective idealization of certain basic mechanisms deemed to determine or influence cycles and growth in a capitalist economy under certain specific circumstances.

Two major extensions of the model, which have made it theoretically more robust (and complicated), should be mentioned in concluding this entry.

First of all, the assumption that the structural coefficients are constant may be dropped and they may instead be treated as functions of the level (or the rate of change) of income, thus making the model nonlinear. Formal investigation of nonlinear multiplier–accelerator models was initiated in the 1950s by Richard Goodwin (1951a and 1951b) and is still a very active area of research. Nonlinear models have two distinct advantages over the linear ones. For one thing, they better correspond to empirical observation of economic facts. Secondly, and most importantly, they can reproduce a far richer (and economically more interesting) diversity of dynamic behaviours. In particular, only they can represent sustained fluctuations of income, i.e. cycles that neither expire nor explode, without requiring very special configurations of parameters for which no economic justification could be found.

A second important extension of the model has been the generalization of some basic results to the multidimensional case. In an economy with an indefinitely large number of sectors the Harrod–Domar equation (7) can be rewritten as

$$[I - A]\,x = B\dot{x}, \tag{13}$$

where $A \in R^{nxn}$ is the flow input–output matrix, i.e. the generalized propensity to consume; $B \in R^{nxn}$ is the stock input–output matrix, i.e. the generalized accelerator; $x \in R^{n}$ is the vector of production levels; *I* is of course the identity matrix.
In analogy to the one-dimensional case we can introduce error-adjustment mechanisms for production and investment, obtaining the system of differential equations

$$\ddot{x} + \{T_i + T_y[I - A] - T_i T_y B\}\dot{x} + T_i T_y [I - A]x = 0 \quad (14)$$

where T_y and T_i are diagonal matrices whose (positive) elements are the speeds of adjustment of production and investment, respectively, in the various sectors.
The analysis of system (13) is obviously more complex than that of (11), even in the linear case, as now the coefficients are of order n^2. However, it is possible to define multidimensional equivalents of the main explosive and damping factors, and to indicate the conditions for oscillatory behaviour. It is also possible to show that – in perfect analogy to the one-dimensional case – the loss of stability which takes place when the explosive forces (the accelerators) become too strong vis-à-vis the damping forces (saving and lags), leads to cyclical behaviour of the system.

See Also

- acceleration principle
- aggregate demand theory
- growth and cycles
- multiplier analysis
- trade cycle

Bibliography

Aftalion, A. 1909. La réalité des surproductions générales. *Revue d'économie politique* 23(March), 81–117, 201–29 and 241–59.

Clark, J.M. 1917. Business acceleration and the law of demand: a technical factor in economic cycles. *Journal of Political Economy* 25(March), 217–35.

Domar, E.D. 1946. Capital expansion, rate of growth and employment. *Econometrica* 14(2), April, 137–

47.

Goodwin, R. 1951a. The nonlinear accelerator and the persistence of business cycles. *Econometrica* 19 (1), January, 1–17.

Goodwin, R. 1951b. Econometrics in business cycle analysis. In *Business Cycles and National Income*, ed. A.H. Hansen, New York: W.W. Norton.

Harrod, R.F. 1936. *The Trade Cycle*. Oxford: Clarendon Press.

Harrod, R.F. 1939. An essay in dynamic theory, *Economic Journal* 49(March), 14–33.

Hicks, J.R. 1950. *A Contribution to the Theory of the Trade Cycle*. Oxford: Clarendon Press.

Kahn, R.F. 1931. The relation of home investment to employment. *Economic Journal* 41(June), 173–98.

Kalecki, M. 1971. *Selected Essays on the Dynamics of the Capitalist Economy 1933–1970*. Cambridge: Cambridge University Press.

Phillips, A.W. 1954. Stabilization policy in a closed economy. *Economic Journal* 64(June), 290–323.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics* 21(2), May, 75–8.

How to cite this article

Medio, A. "multiplier–accelerator interaction." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000275> doi:10.1057/9780230226203.1145

The New Palgrave Dictionary of Economics Online

Mummery, Albert Frederick (1855 – 1895)

Michael Bleaney
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Hobson, J.A.; Mummery, Albert Frederick; underconsumption

Article

Co-author with J.A. Hobson of the *Physiology of Industry* (1889), Mummery was also a famous mountaineer who wrote a book on climbing in the Alps and the Caucasus and died in the Himalayas in 1895. According to Hobson's own account (*Confessions of an Economic Heretic*, pp. 29–30), it was Mummery who set him on the path to intellectual heresy; considering that Hobson's later economic writings may in many ways be regarded as a development of the theme established in the *Physiology of Industry*, this is a considerable achievement.
Mummery was a businessman who seems to have become acquainted with Hobson by chance while the latter was teaching in Exeter. He managed to convince Hobson, after considerable argument, that the economy contained a serious tendency to over-saving, and that depressions were the expression of this tendency. Unfortunately we do not know how far this idea had developed in Mummery's mind before he met Hobson, or how much each contributed to the published version of the argument. Since Hobson subsequently became a prolific writer on economic matters, one suspects that the meat of the book was his work. It is not certain that Mummery had received much training in economics, and he may have contributed little more than the germ of the idea.

See Also

- Hobson, John Atkinson
- underconsumptionism

Selected works

1889. (With J.A. Hobson) *The Physiology of Industry*. London: Murray.

Bibliography

Hobson, J.A. 1938. *Confessions of an Economic Heretic*. Brighton: Harvester Press, 1976.

How to cite this article

Bleaney, Michael. "Mummery, Albert Frederick (1855–1895)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000280> doi:10.1057/9780230226203.1861

The New Palgrave Dictionary of Economics Online

Mun, Thomas (1571 – 1641)

Walter Eltis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balance of payments; balance of trade; East India Company; mercantilism; Mun, T.; specie-flow mechanism

Article

Thomas Mun, the distinguished mercantilist, was born in London in June 1571 and died in July 1641. He was the third son of John Mun, a mercer, whose father, also John Mun, held the office of provost of the moneyers in the Royal Mint and received a grant of arms in 1562.
Thomas Mun became an extremely wealthy merchant, and a Director (Member of the Committee) of the East India Company in 1615. In 1624 he had the opportunity to serve as Deputy Governor which he declined, but he remained a director until he died.
The East India Company was much criticized because its trade involved exports of bullion (in order to purchase spices). In 1621 Mun was author of a pamphlet, *A Discourse of Trade, from England unto the East-Indies*, in which he set out the benefits that England derived from this trade. His argument was that the same spices (and he details the amounts) would otherwise have been imported from Turkey at three times the sterling cost, and that the purchase of spices in the Indies thus produced satisfactory results for British consumers, while merchants also benefited, and so ultimately did the balance of trade. On Mun's figures the East India Company exported £100,000 of silver yearly to import silk and spices which sold in England for £500,000 (out of which customs duties took a substantial fraction). But only £120,000 of these goods were actually consumed in England, and the remaining £380,000 were re-exported with the consequence that England gained back considerably more bullion than the original outflow of £100,000.
In 1622 he was the leading member of a committee of merchants which submitted evidence to a Commission set up by James I to investigate the causes of the fall in the exchange rate and the loss of specie from which Britain was suffering. Mun was principal author of their first memorandum in 1622, and sole author of later memoranda submitted in 1623. He strongly opposed Malynes' view that the fall in the exchange rate was attributable to conspiratorial behaviour by foreign merchants, and argued that the balance of *trade* was the principal determinant of specie flows and the exchange rate. His

memoranda resurfaced in 1664, as chapters in his posthumously published magnum opus, *England's Treasure by Forraign Trade, or the Ballance of our Forraign Trade is the Rule of our Treasure*, which Schumpeter has referred to as ‘the classic of English mercantilism’. This was published by his son, John Mun, with the imprimatur and personal approval of Charles II's Secretary of State, Sir Henry Bennet. *England's Treasure* demolished the previous mercantilist literature which advocated detailed interventionist policies to sustain the English money supply and the exchange rate, such as banning gold exports, currency appreciation, lowering the metallic content of the currency, and encouraging the domestic circulation of foreign coin. Mun reiterated the fundamental balance of payments equation that specie flows must be determined primarily by the excess of exports over imports, and therefore insisted that there could not be a sustained loss of gold and silver while there was a trade surplus, while none of the above expedients could prevent a monetary outflow in the face of a sustained deficit.

His book hammered home the significance of the balance of payments equation, with numerous examples to demonstrate the impotence of detailed interventionist policies to hold or attract bullion while trade was in deficit. At the same time, he developed examples like those in his earlier *Discourse of Trade*, to show how it was ultimately the domestic consumption of imports and not imports as such that needed to be compared with exports to determine the net balance of trade. Imports by English merchants which were not destined for consumption in England were bound to result in equivalent exports, plus of course merchants’ profits and duties for the King.

Mun went on to explain the relationship between the balance of trade and the excess of home production over consumption, and to distinguish carefully between the financial interests and impact on the trade balance of Merchants, the Commonwealth (the whole population) and the King. Merchants were solely concerned with profit. The Commonwealth determined the trade balance via the relationship between the aggregate expenditures and incomes of the whole population, while the Sovereign's interest in trade depended considerably upon customs and excise, ‘the King by his Customs and Imposts may get notoriously, even when the Merchant notwithstanding shall lose grievously’ (p. 147).

Mun may well have been the first to state the celebrated proposition (which Lord Kaldor made much of in the 1970s) that the current account trade surplus must correspond to the sum of the financial surpluses of the public and private sectors. He set out an example where a King enjoys revenues of £900,000, spends £400,000 and accumulates the resulting budget surplus of £500,000. Then if the trade surplus is merely £200,000, the King will

> lay up £300,000 more in his Coffers than the whole Kingdom gains from strangers by forraign trade: who sees not then that all the money in such a State, would suddenly be drawn into the Princes treasure, whereby the life of lands and arts must fail and fall to the ruin both of the publick and private wealth? So that a King who desires to lay up much money must endeavour by all good means to maintain and encrease his forraign trade. (pp. 188–9)

Mun believed that the achievement of a trade surplus on which monetary inflows depended would be best achieved where the population moderated consumption, and merchants enjoyed maximum freedom to exploit opportunities for trade. He has been much praised in the secondary literature for his perception that it was the trade balance that determined specie flows. This has been universally judged vastly superior to the previous literature which recommended piecemeal interventionism in financial markets.

According to McCulloch's 1847 *Edinburgh Review* article ‘Mun's book was received as the gospel of finance and commercial policy; and his principles ruled for above a century the policy of England, and much longer that of the rest of Europe’ (p. 450).

Mun's analysis was superseded in 18th-century England because he failed to go a vital stage further and appreciate the potentially self-correcting nature of the balance of payments. This led Hume and his followers to cease to regard the trade balance as a primary policy objective in comparison with the achievement of a growing capital stock, and increasing levels of output and employment, about which Mun was also deeply concerned.

But those who have been satisfied that the trade balance is self-correcting have sometimes failed to appreciate that a continuing deficit is inevitable where consumption (modern writers would say, domestic absorption) exceeds production. They also lost Mun's perception that in a protectionist world, winning trade away from other countries may permit increases in domestic capital and employment with would not otherwise occur.

Selected works

1621. *A Discourse of Trade, from England unto the East-Indies*. London.

1664. *England's Treasure by Forraign Trade. Or, The Ballance of our Forraign Trade is the Rule of our Treasure*. London. Repr. in the Economic History Society Reprints of Economic Classics, Oxford, 1928.

These works by Mun are both reprinted in J.R. McCulloch, ed., *Early English Tracts on Commerce*, Cambridge: Cambridge University Press, 1952, which is itself a reprint of the London *Political Economy Club’s* 1856 publication, and page references are to this edition.

Bibliography

Appleby, J.O. 1978. *Economic Thought and Ideology in Seventeenth-Century England*. Princeton: Princeton University Press.

Heckscher, E.F. 1955. *Mercantilism*, 2 vols. London: Allen & Unwin.

McCulloch, J.R. 1847. Primitive political economy of England. *Edinburgh Review* 172, 426–52.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Supple, B.E. 1970. *Commercial Crisis and Change in England 1600–42*. Cambridge: Cambridge University Press.

Wilson, C. 1957. *Profit and Power: A Study of England and the Dutch Wars*. London: Longmans.

How to cite this article

Eltis, Walter. "Mun, Thomas (1571–1641)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000281> doi:10.1057/9780230226203.1147

The New Palgrave Dictionary of Economics Online

Mundell, Robert (born 1932)

Charles Engel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mundell is best known for his work creating an open-economy version of the IS–LM model. Its special interest lies in the analysis of monetary and fiscal policy. He emphasized the importance of the speed of adjustment in capital markets and the role of fixed versus flexible exchange rates in determining the impact of policy changes and the determination of a desirable monetary–fiscal policy mix. Mundell has also been influential on optimum currency areas, the effect of inflation on portfolio investment, and trade theory.

Keywords

assignment problem; balance of payments; central banking; expectations; external balance; factor mobility; fiscal policy; fixed exchange rates; flexible exchange rates; Heckscher–Ohlin–Samuelson model; interest rates; international capital flows; IS–LM analysis; monetary approach to the balance of payments; monetary equilibrium; monetary policy; Mundell, R.; open-economy models; optimum currency areas; sterilization; sticky prices

Article

Robert Mundell is one of the key figures in the development of thought in international monetary economics. His work on the IS–LM model in open economies, equilibrium in a world of perfect capital mobility, monetary dynamics in open economies, and optimal currency areas constitutes the core of the research for which Mundell is best known. His work continues to this day to be influential in the analysis of policy decisions in open economies, but an equally important legacy of Mundell's is the role his work played in determining the direction of research in open-economy macroeconomics in the 1960s, 1970s and through to the present. Mundell's work had such a great impact in part because it combined theoretical rigor with elegant presentation. Mundell was awarded the Nobel Prize in 1999 for ‘his analysis of monetary and fiscal policy under different exchange rate regimes and his analysis of

optimum currency areas’.

Mundell was born in Kingston, Ontario, in 1932. His undergraduate education was undertaken at the University of British Columbia and the University of Washington. He engaged in postgraduate studies at the London School of Economics and received his Ph.D. from MIT in 1956. He taught at Stanford University and the Bologna Center of the School of Advanced International Studies of the Johns Hopkins University, and joined the staff of the International Monetary Fund in 1961. He was a Professor of Economics at the University of Chicago from 1966 to 1971. In 1974 he joined the faculty at Columbia University, where he has spent the remainder of his career.

Mundell is perhaps best known for his work creating an open-economy version of the IS–LM model. Mundell's (1960; 1961; 1962; 1963a) model is still the workhorse model of most undergraduate texts in international macroeconomics. Mundell, like Meade, Metzler, and a few others whose work preceded Mundell's, recognized that the analysis of exchange rates and balance of payments flows must proceed in a monetary general equilibrium framework. Under Mundell's initial formulation, the equilibrium conditions in money markets and goods markets were augmented by an external balance condition. Mundell's concept of external balance was a balance of payments equilibrium, in which the net flow demand for foreign exchange is zero. Demand for foreign exchange comes from importers of goods and from importers of foreign assets. In his initial work, Mundell modelled the demand for foreign assets as a flow that depended on the difference between home and foreign interest rates. As long as there was a positive spread between home and foreign interest rates, capital inflows would persist at a steady rate.

Mundell's special interest was in the analysis of monetary and fiscal policy. He emphasized the importance of the speed of adjustment in capital markets and the role of fixed versus flexible exchange rates in determining the impact of policy changes and the determination of a desirable monetary–fiscal policy mix. His framework was extended and used to consider policy issues by academics and central bankers for many years.

One basic insight of these models concerns the difference in the impact of fiscal and monetary policy under fixed and floating exchange rates. Consider a monetary expansion. Under a floating exchange rate, external balance requires a depreciation of the domestic currency. The monetary expansion lowers interest rates, leading to a capital outflow and a decline in demand for the domestic currency. With sticky nominal goods prices (the hallmark of the Keynesian IS–LM analysis), the depreciation makes imported goods more expensive, so expenditure switches to home goods. This expenditure switching effect would not be present if exchange rates were fixed. Indeed, Mundell (1961) makes the point that in the absence of sterilization (see below) the monetary expansion would be reversed over time. That is, under fixed exchange rates, the monetary expansion leads to a balance of payments deficit. Under a balance of payments deficit, as the central bank's foreign reserves decline, the money supply falls.

In contrast, an expansionary fiscal policy might have greater impact under fixed exchange rates, when capital mobility is high. In the IS–LM framework, an increase in aggregate demand raises interest rates. This should lead to an inflow of capital and an appreciation of the home currency under flexible exchange rates. But the appreciation switches demand away from home goods, thereby dampening the effect of the fiscal expansion. Under fixed exchange rates, the expenditure switching does not occur. Moreover, in the absence of sterilization operations the balance of payments surplus that ensues from the fiscal expansion will lead to a domestic monetary expansion as the central bank acquires foreign reserves.

Note how the analysis of the effects of fiscal expansions depends on the assumption that capital flows

respond significantly to changes in the interest rate. If capital flows were not significant, the analysis would be reversed. A fiscal expansion leads to an increase in domestic income. Some of that increased income is spent on imports. There may be increased capital inflows because the interest rate has risen domestically, but if these flows are slight then the decline in the trade balance dominates, so the country's balance of payments deteriorates. Under floating exchange rates, then, there will be a currency depreciation that further boosts aggregate demand. That effect is not present under fixed exchange rates, and indeed there could be a contractionary effect of the balance of payments deficit in the absence of sterilization.

Of special note is Mundell's (1963a) version of his model under the assumption of perfect capital mobility, so that the rates of return on home and foreign nominal bonds are equalized. At one level this paper is a simple extension of his earlier work to consider the extreme case in which capital flows infinitely quickly to equalize rates of return. But at another level the model is fundamentally different. In essence this case turns the external balance condition from a flow equilibrium (analogous to the IS curve) into an asset-market equilibrium condition (analogous to the LM curve.) In this model, for asset markets to be in equilibrium households must be satisfied not only with their holdings of money relative to interest-earning assets (LM) but also with their holdings of domestic bonds relative to foreign bonds. This model laid the foundation for virtually all later work in the field that understands the market for foreign exchange to be an asset market.

The key distinction analytically is that the flow of assets plays no role per se in determining equilibrium in this formulation. For example, the trade balance plays no direct role in establishing the equilibrium in the foreign exchange market. In contrast to many models of the 1950s in which the exchange rate adjusted to set the trade balance to zero, here the trade balance plays a role only in its contribution to the net demand for domestic output. The balance of payments simply reflects the central bank's net accumulation of foreign assets. As Obstfeld (2001) points out, the balance of payments is no longer a relevant indicator of external balance in this setting. By modelling the external balance condition as an asset-market equilibrium, Mundell opened the door for subsequent models that considered the role of expectations in determining exchange rates and laid the foundation for models of balance of payments crises under fixed exchange rates in which speculative attacks play a key role.

Subsequent developments in the field have replaced Mundell's ad hoc formulations of behaviour with optimizing models, and have explicitly modelled expectations formation. But Mundell's work was a cornerstone of the development of more sophisticated models, and open-economy macroeconomic models are still often evaluated by comparing their implications with those of the models of Mundell.

Dynamics was a key concern of Mundell's. Even within the IS–LM framework, Mundell examined the evolution of output, interest rates, exchange rates and prices. Mundell paid special attention to the dynamic effects of balance of payments ‘disequilibrium’ under fixed exchange rates. When the net private flow demand for foreign exchange is not zero (that is, the sum of the current account and the private component of the capital account is not zero), then, in Mundell's terms, there is balance of payments disequilibrium. Mundell made explicit the distinction between balance of payments flows that were sterilized – so that the monetary base did not change – and policies that allowed the money supply to change automatically when there was balance of payments disequilibrium. Mundell (1961) especially was a precursor of the literature that became known as the ‘monetary approach to the balance of payments’. That literature emphasized the automatic adjustment mechanism when there is no sterilization. Most of that analysis was undertaken in classical-style models in which nominal goods

prices were assumed to be flexible. Indeed, Mundell (1967) was a contributor in that tradition. But what Mundell's (1961) piece makes clear is that it is the assumption of non-sterilization that is key to understanding the dynamics of adjustment. Even in a world of sticky nominal prices, automatic adjustment to balance of payments disequilibrium can occur through adjustment in the money supply.
Dynamics were central in Mundell's development of what became known as ‘the assignment problem’. The question was whether the central bank should be responsible for external balance and fiscal authorities for internal balance, or vice versa. Mundell's answer was that each policy tool should be assigned to the market in which it has the greater effect, which depends on the speed of adjustment of goods markets relative to capital markets. Mundell modelled policymaking in a realistic world in which policymakers have an imperfect understanding of the state of the economy, and in which macroeconomic adjustment to policy changes is slow. These concerns have all but disappeared from more recent research in macroeconomic policymaking, but Mundell's focus still seems relevant. Moreover, Mundell's work recognizes that policymaking at the national level is not in the hands of a single policymaker, but instead involves the interaction of decisions by central banks and fiscal authorities whose actions and goals may not be perfectly coordinated.
Mundell's (1961) paper on optimum currency areas also is still very influential. This paper determines some conditions under which it is optimal for countries to share a common currency. Mundell's view was that there may be some advantage to sharing currencies in terms of reduced transactions costs. But the adoption of a common currency means, of course, that each country is not free to pursue its own independent monetary policy. That loss may not be so large when factors of production can flow freely between the countries in a currency area. If there is a downturn in one country, adjustment can occur through factor flows towards the country with the stronger economy. But if factor mobility is weak, then there is a case for each country to have its own independent money. In general, in Mundell's framework the optimum currency area is determined by a trade-off between these considerations about factor mobility and considerations involving the transactions costs of having many separate currencies. Mundell's work in this area spawned a large literature that considered other factors that determine whether a set of countries were good candidates for adoption of a single currency.
Mundell is also known for his short paper (1963b) that develops what became known as the ‘Mundell–Tobin effect’. Mundell argued that inflation reduced the demand for real money balances. That led to a portfolio shift that could induce greater investment in real capital.
Mundell (1957) also made a lasting contribution in pure trade theory. This paper examined the effects of factor mobility in the Heckscher–Ohlin–Samuelson model. Factor mobility could be a substitute for goods trade, just as goods trade could substitute for factor mobility (as in the well-known factor-price equalization theorem.)
The Nobel Prize citation notes that ‘Mundell chose his problems with uncommon – almost prophetic – accuracy in terms of predicting the future development of international monetary arrangements and capital markets.’ When Mundell wrote much of his influential work in the early 1960s, much of the world was on a fixed-exchange rate system – although his native Canada had a freely floating exchange rate. Moreover, there were still significant barriers to international flows of capital that had been erected in the 1930s and 1940s, even among advanced industrialized countries. Nonetheless, Mundell focused in much of his work on the contrast between the fixed and floating exchange rate systems, with an emphasis on the role of capital mobility. Only in the early 1970s did most of the advanced world move to floating exchange rates, and obstacles to capital flows were gradually eliminated in the decades

following Mundell's early writings. His work on optimum currency areas was frequently cited in the economic analysis that preceded the introduction of the euro.

Many of Mundell's contributions are collected in *International Economics* (1968). Excellent brief surveys of Mundell's work can be found in Royal Swedish Academy of Sciences (1999) and Obstfeld (2001). Mundell (2001) provides an interesting history of the development of some of Mundell's work.

See Also

- international capital flows
- international financial institutions (IFIs)
- international monetary institutions

Selected works

1957. International trade and factor mobility. *American Economic Review* 47, 321–55.

1960. The monetary dynamics of international adjustment under fixed and flexible exchange rates. *Quarterly Journal of Economics* 74, 227–57.

1961. A theory of optimum currency areas. *American Economic Review* 51, 657–65.

1962. The appropriate use of monetary and fiscal policy for internal and external stability. *IMF Staff Papers* 9, 70–9.

1963a. Capital mobility and stabilization policy under fixed and flexible exchange rates. *Canadian Journal of Economics* 29, 475–85.

1963b. Inflation and real interest. *Journal of Political Economy* 71, 280–3.

1967. Barter theory and the monetary mechanism of adjustment. In *Capital Movements and Economic Development*, ed. J. Adler. London: Macmillan.

1968. *International Economics*. New York: Macmillan.

2001. On the history of the Mundell–Fleming model. Keynote speech. *IMF Staff Papers* 47(special issue), 215–27.

Bibliography

Obstfeld, M. 2001. International macroeconomics: beyond the Mundell–Fleming model. *IMF Staff*

Papers 47(special issue), 1–39.

Royal Swedish Academy of Sciences. 1999. *Bank of Sweden Prize in Economic Sciences in Memory of Alfred A. Nobel, 1999*. Online. Available at http://nobelprize.org/economics/laureates/1999/ecoback99.pdf, accessed 9 August 2005.

How to cite this article

Engel, Charles. "Mundell, Robert (born 1932)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000344> doi:10.1057/9780230226203.1148

The New Palgrave Dictionary of Economics Online

municipal bonds

James M. Poterba

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Municipal bonds are issued by state and local governments in many nations. In the United States the interest on these bonds is usually exempt from federal income tax, which provides an incentive for taxable investors to hold municipal bonds even if their before-tax yield falls below that of other taxable bonds. This article describes the various types of municipal bonds, the yield spread between taxable bonds and municipal bonds, and the factors that determine the efficiency of the federal income tax exemption as a means of subsidizing capital outlays by state and local governments.

Keywords

alternative minimum tax (USA); bonds; implicit tax rate; municipal bonds; Social Security in the United States

Article

Municipal bonds differ from most other securities because of their special tax status. Interest payments on bonds issued by state and local governments in the United States are exempt from federal income tax. Most states with income taxes also exempt their own interest payments from tax. The federal income tax exemption for municipal bond interest is usually justified on the grounds that it reduces borrowing costs for states and localities, thereby facilitating their investment in public infrastructure.

When the federal income tax was enacted in 1913, there was some question as to the constitutionality of such a federal tax on interest paid by states and localities. In 1988, the Supreme Court affirmed the federal prerogative to tax such interest in the case of *South Carolina v. Baker*. The tax exemption for municipal bond interest should therefore be viewed in the same way as any other tax expenditure, namely as a political decision about the structure of income taxation.

There are three types of municipal bonds: *general obligation bonds*, which are backed by the ‘full faith and credit’ of the borrowing jurisdiction; *revenue bonds*, which are backed by the stream of income from a particular project such as a highway or publicly operated power plant; and *private purpose bonds*,

which are tax-exempt bonds issued by private borrowers with the authorization of a state or local government. Only general obligation or 'GO' bonds have a potential claim on the tax revenues of a state and local government. The interest payments on revenue bonds are dependent on the revenues associated with the project that issued the bonds. Private purpose bonds are typically used to finance private sector projects that are deemed beneficial to the state or local economy or community; in practice these bonds finance a wide range of activities. The market value of outstanding tax-exempt bonds in 2006 was 2.3 trillion dollars according to estimates from the Federal Reserve Board Flow of Funds Accounts. GO bonds account for roughly 40 per cent of outstanding tax-exempt debt.

While municipal bond interest is generally exempt from federal income taxation, the relevant tax rules are complicated in some situations. For example, retirees who receive Social Security benefits must include tax-exempt bond interest in the income concept that is used to determine how much of their Social Security income is included in taxable income. In addition, the interest paid on many private purpose bonds is taxable under the federal alternative minimum tax (AMT). While the AMT affected only 3.5 million taxpayers in 2006, projections suggest that provided there are no changes in the basic structure of the tax, it will apply to more than 20 million taxpayers by 2010. Bonds that are not exempt from the AMT typically offer investors a higher yield than bonds that pay interest that is completely tax exempt.

In part as a result of changes in the tax law, there have been changes over time in the ownership patterns for municipal bonds. Prior to 1986, commercial banks were the primary holders of short-term municipal bonds while households and insurance companies were the primary holders of long-term municipals. The Tax Reform Act of 1986 sharply limited the incentives for banks to hold tax-exempt bonds, and since then the ownership mix has shifted towards households. According to Flow of Funds data for the third quarter of 2006, households were the direct owners of 37 per cent of outstanding municipal bonds. Mutual funds, which are largely owned by households, accounted for another 33 per cent. Commercial banks hold seven per cent, while property and casualty insurance companies hold 14 per cent.

Investors who hold municipal bonds avoid paying income taxes on their interest income, but they pay an 'implicit tax' when the pre-tax interest rate on municipal bonds is lower than that on an equally risky taxable bond. The yield spread between taxable and municipal bonds is often summarized by the *implicit tax rate*. This is the value of θ for which $(1-\theta)R_T=R_M$ where R_T is the yield on newly issued Treasury bonds and R_M is the yield on prime grade municipal bonds of comparable maturity. This relationship is only satisfied by newly issued taxable and municipal bonds under the assumption that investors plan to hold these bonds to maturity. Poterba (1986) shows that with forward-looking investors, the implicit tax rate measured from current bond yields reflects not just current marginal tax rates on taxable interest but future marginal tax rates as well. For seasoned bonds, the tax treatment of differences between the purchase price of the bond and the par value complicates the calculation of the implicit tax rate. More generally, when investors sell their bonds before maturity, changes in bond prices may result in taxable capital gains or losses. The definition of the implicit tax rate also assumes that Treasury bonds and prime grade municipals are equally risky, an assumption that some might question.

The implicit tax rate on municipal bonds varies across bond maturities at a given point in time, and it varies over time in part as a result of changes in tax rates and tax rules. During the first week of 2007, the interest rate on 30-year GO bonds with an AAA rating was 4.14 per cent, while the yield on a 30-year Treasury bond was 4.59 per cent. The implicit tax rate based on these values is 9.8 per cent, well below the top statutory marginal tax rate on individual investors, 35 per cent. The yield spread between

AAA-rated municipal bonds and AAA-rated corporate bonds is larger, but this comparison raises the challenge of risk adjustment. For the same week, the yield on one-year AAA-rated municipals was 3.53 per cent, while that on one-year Treasury bonds was 4.92 per cent. The implicit tax rate at the one year maturity was therefore 28.3 per cent.

One of the challenges in analysing the municipal bond market is explaining why implicit tax rates are substantially below top statutory rates. Chalmers (1998) discusses various potential explanations and rejects the possibility that differential default risk explains this long-standing pattern. The yield curve puzzle has motivated research on the relative pricing of taxable and tax-exempt bonds. Green (1993) argues for moving beyond yield-to-maturity analysis, such as that underlying the foregoing implicit tax rate computations, and developing a more subtle analysis of the tax-exempt bond market.

The key insight in Green (1993) and several subsequent studies is that fully taxable individual investors are unlikely to regard newly issued tax-exempt bonds and newly issued taxable bonds as competitive investment alternatives. If such investors chose to hold taxable bonds, they should do so by holding bonds that generate income in a way that generates less tax liability than a newly issued bond. The opportunities to earn bond returns that face a lighter tax burden are greater at longer than at shorter maturities, because divergences between the purchase price of a bond and its par value are potentially greater at long maturities. This role of tax-wise investing appears to receive empirical support in yield curve comparisons at different maturities. It may help to explain why implied tax rates in the municipal bond market are often lower for longer-maturity than for shorter-maturity bonds.

Whether the policy of exempting interest on state and local government bonds from federal taxation is an efficient method of encouraging capital formation by states and localities is a long-standing subject of debate. The answer turns on the difference between the implicit marginal tax rate on municipal bonds, which determines the interest saving of state and local government borrowers, and the weighted-average marginal tax rate of municipal bond investors, with weights equal to the tax-exempt interest receipts of each investor. The latter determines the federal government's revenue cost from exempting interest on state and local government obligations from tax. If the revenue cost exceeds the interest saving, it would cost less for the federal government to provide cash transfers to state and local governments equal to the amount of their current interest saving, while taxing interest on their bonds, than to pursue the current policy of tax exemption. In 2002, the weighted-average marginal tax rate for individual investors who received tax-exempt interest was 30.2 per cent. Feenberg and Poterba (1991) describe the calculation of such marginal tax rates. Since the implicit tax rate on 20-year municipal bonds and Treasuries varied between ten and 20 per cent during calendar 2002, the revenue cost of the exemption for households appears to exceed the interest saving for state and local government borrowers.

The progressivity of the federal income tax schedule is a key determinant of the efficiency of policies that exempt interest from taxation. When the yield spread between taxable and municipal bonds is determined by the marginal tax rate of the *lowest* tax rate investor who holds those bonds, but the revenue cost is determined by the weighted average marginal tax rate of the investors who hold municipal bonds, then the efficiency cost of the tax exemption will be greater when the top marginal tax rates affect many but not all municipal bond investors, and when the top rates are substantially higher than the rates on lower-income households.

When investors have access to taxable and tax-exempt bonds of equal risk, market equilibrium should involve investor clienteles in which investors segment themselves according to their tax rates. High tax

rate investors should hold tax-exempt bonds, while low tax rate investors should hold taxable bonds. In practice, this separation does not occur. Poterba and Samwick (2003) show that among households that hold tax-exempt bonds, 55 per cent also hold taxable bonds. In contrast, only 15 per cent of the households that hold taxable bonds also hold tax-exempt bonds. There are risks inherent to holding municipal bonds, such as the risk of tax change, that are difficult to hedge and may incline investors to diversify their portfolios. This may explain why most investors who hold municipal bonds also hold taxable bonds.

There are many innovative products in the municipal bond market, including variable rate municipals, insured municipal bonds, and zero coupon tax-exempt bonds. The bonds issued by several large issuers, particularly large states and revenue authorities, trade in active after-markets, but the markets for many smaller municipal bond issues are not very liquid.

See Also

- bonds
- fiscal federalism
- local public finance
- tax expenditures
- taxation of income

Bibliography

Chalmers, J.M.R. 1998. Default risk cannot explain the muni puzzle: evidence from municipal bonds that are secured by U.S. Treasury obligations. *Review of Financial Studies* 11, 281–308.

Feenberg, D.R. and Poterba, J.M. 1991. Which households own municipal bonds? Evidence from tax returns. *National Tax Journal* 44, 93–104.

Green, R.C. 1993. A simple model of the taxable and tax-exempt yield curves. *Review of Financial Studies* 6, 233–64.

Poterba, J.M. 1986. Explaining the yield spread between taxable and tax-exempt bonds. In *Studies in State and Local Public Finance*, ed. H. Rosen. Chicago: University of Chicago Press.

Poterba, J.M. and Samwick, A. 2003. Taxation and household portfolio composition: U.S. evidence from the 1980s and 1990s. *Journal of Public Economics* 87, 5–38.

How to cite this article

Poterba, James M. "municipal bonds." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000282> doi:10.1057/9780230226203.1149

The New Palgrave Dictionary of Economics Online

Musgrave, Richard Abel (1910 – 2007)

Original by Peter Mieszkowski, updated by the editors
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Richard Musgrave is best known for his treatise *The Theory of Public Finance* (1959). His most original and lasting contributions are in taxation theory and public goods theory. His work on tax incidence has been the starting point for all subsequent studies on tax burdens by income classes, and he broke new ground by introducing the concept of equal options as the basis for horizontal equity. His separation of budgetary functions into allocation and distribution branches has acquired increased practical significance as much of the expansion of the public sector has consisted of increased transfer payments.

Keywords

allocative versus distributive budgetary functions; charitable giving; choice under uncertainty; direct and indirect taxation; excise taxes; fiscal stabilization; horizontal equity; indirect taxation; merit goods; Musgrave, R. A.; Nozick, R.; primary and secondary redistribution; public finance; public goods; Rawls, J.; revealed preference; risk taking; social justice; social rights; tax incidence; taxation of capital income; taxation of corporate profits

Article

Born in Koenigstein, Germany, Musgrave was educated at Heidelberg (where he obtained a Diplom Volkswirt in 1933) and at Harvard University (where he obtained his Ph.D. in 1937). After serving at the Federal Reserve System in Washington, he held appointments at a number of leading North American universities and ended his formal teaching career at Harvard, where he was Professor Emeritus. He was an economic adviser to a number of governments, headed foreign tax commissions, and served as editor of the *Quarterly Journal of Economics*.
Richard Musgrave is best known for his outstanding treatise *The Theory of Public Finance*, published in 1959 at a time when social expenditures were growing rapidly throughout the industrial world, and when poverty and social justice had become primary policy concerns. This book, which is comprehensive, has

served as a fundamental source for scholars and as a teaching reference. In it Musgrave summarizes and extends his original contributions to expenditure theory and the theory of taxation, provides an extensive review of the classical literature in public finance, and includes a thorough discussion of fiscal and monetary policy developed from a Keynesian perspective. One of the great strengths of the book is Musgrave's broad knowledge of the early European masters of public finance, notably Wicksell and Lindahl. By reviewing the classical writers and relating his theory of the public household to their work, Musgrave built an essential bridge between earlier ideas and the development of modern public goods theory.

Musgrave made significant contributions to virtually all areas of public finance. He wrote on the theory of fiscal federalism and revenue sharing, international aspects of taxation, alternative measures of income tax progressivity, land value taxation, the theory of fiscal sociology, and the effects of tax policy on private capital formation, as well as on various aspects of debt and monetary policy. His most original and lasting contributions can be grouped into two categories: taxation theory, which includes three major contributions, and public goods theory, in particular his theory of the public household.

One of Musgrave's most distinguished contributions to taxation theory is his joint paper with E.D. Domar on the effects of taxes on risk taking (1944). The authors show that taxes on capital income will not necessarily decrease investment in relatively risky ventures once the loss offset provisions of the tax and its income effects are accounted for. In fact, it is quite likely that risk taking will be encouraged by an interest income tax. This article ranks with the half-dozen most influential articles on taxation written since the mid-1950s, and it represents the first application of the theory of choice under uncertainty to taxation. Its conclusions have proved to be quite robust to more general formulations of the theory of risk taking.

Musgrave's second contribution to taxation theory is his theoretical and empirical work on tax incidence. He has developed most of the general concepts currently used in incidence analysis, and, in one of the first general equilibrium analyses, established the fundamental equivalences between direct and indirect taxes and between general and specific factor taxes. These contributions clarify a much confused issue: whether general excise taxes are shifted forwards to purchasers of taxed commodities or backwards to providers of factor services. They also established the importance of both uses and sources aspects of incidence theory.

Musgrave's work on the allocation of tax burden (1951) to different income groups is a basic contribution to applied analysis and has been the starting point for all subsequent studies on tax burdens by income classes. More recently (1974) he refined this earlier work and covered the distributive aspects of expenditures as well as taxes. In another important study, *The Shifting of the Corporation Tax* (1963), with M. Krzyzaniak, Musgrave developed the first econometric estimates of incidence and concluded that the corporate profits tax is shifted forwards, a finding which gave rise to a large literature.

Musgrave extended and refined the normative theory of equitable taxation and its implications for income taxation and the concept of horizontal equity (1959, ch. 8). Later, in 1976, he broke new ground by introducing the concept of equal options as the basis for horizontal equity. Within this framework, two persons are considered to be in equal positions and should be treated equally if they face the same options. Thus, two persons with the same present value of lifetime earnings would be considered equal. One of the important insights of this concept is that under certain assumptions a consumption-based tax system is more equitable than an income tax system: the first treats equals equally while the second discriminates against persons who save relatively more.

The theory of the public household, Musgrave's unifying perspective on public goods, has provided the basis for many of his insights into that fundamental topic. This theory distinguishes between three branches of government – the allocative branch, which provides for social goods and deals with related questions of efficiency, the distribution branch, which modifies the distribution of income as determined by market forces and inheritance, and the stabilization branch, which is concerned with unemployment and overall economic stability.

He stresses that the failure to distinguish between the three different objectives of budget policy will involve unnecessary conflict and inefficient policy design. For instance, different voters may agree on the objective of fiscal stabilization but may fail to enact a proportional cut in taxes in recession if the proposals to combat recession will increase expenditures or change the distribution of income. Hence, one of the practical principles to emerge from the three-budget classification is that expenditure levels and the distribution of income, or tax shares of individual groups, should be determined independently of stabilization objectives. Similarly, the distinction between allocation and distribution leads to the principle that redistribution should be implemented primarily through a tax-transfer process. This will avoid inefficient increases of public expenditures in the name of progressive objectives.

The distinction between allocation and distribution has acquired increased practical significance as much of the expansion of the public sector has consisted of increased transfer payments: Social Security and publicly financed medical care. Also, in a wide variety of policy areas, from the regulation of the prices of energy resources to efficient congestion-pricing of urban highways, the conflict between allocation and distribution has led to poor policy design, as he predicted. Compensation systems are needed to offset the redistributive effects of efficient allocation policies.

The value of the distinction between allocation and distribution has been enhanced by the work of Robert Nozick and John Rawls on social justice. Nozick has restated and extended John Locke's doctrine that one is fully entitled to the fruit of one's labour. Rawls developed a very different theory based on a communal claim to the output of high-ability persons. However, the claim structure is voluntarily agreed upon through a social compact, as risk-averse individuals, not knowing their future position, agree behind a veil of ignorance to share their income. This contractual approach to distribution is fully consistent with Musgrave's separation between the allocation and distribution branches.

Musgrave distinguishes between primary and secondary redistribution. Primary redistribution is determined by the social rights that entitle the individual to some share of the social product based on membership in the community, rather than on property ownership or labour supplied. Secondary redistribution is voluntary giving that occurs either through private charities or collective provision. Secondary redistribution is Pareto optimal in that the donor derives more satisfaction from providing the gift to the poor than from additional personal consumption.

The mix between primary and secondary redistribution will vary across societies, according to differences in social values. Also, some social rights, or primary redistribution, may be provided in part in the form of goods and services, such as education, training programmes and medical care. This possibility blurs the separation of allocation and distribution functions.

The primary shortcoming of the distinction between allocation and distribution, however, is not the existence of transfers in kind and the subsidization of certain goods, which Musgrave has classified as merit goods (1957; 1959). As stressed by Samuelson, the fundamental issue is that numerous allocations between social and private goods are Pareto efficient, and the choice of an efficient allocation, a task for the distribution branch, has allocative consequences. In a planning solution, then, allocation and

distribution are decided simultaneously, not separately.

Musgrave agrees to the formal correctness of this argument but he argues that this approach implicitly assumes that the planner knows individual preferences, and that the question of distribution is dealt with *de novo*. If, however, the distribution of income is determined primarily by market forces and preferences are not known, a pricing rule or voting rule that induces preference revelation must be designed. The determination of the pricing rule is the allocative function of government. The determination of money income, in conjunction with the pricing rule, is the distributive function.

When considered from a broader perspective the separation of budgetary functions into allocation and distribution branches has been invaluable, both as a normative theory and as a description of the way public agencies operate. Experience shows that it is very important to develop coordination between branches of government. Also, Musgrave's three-branch theory clarifies many positive issues, such as the causes of large foreign trade deficits and the demise of central cities in metropolitan areas, as well as the design of policies to deal with these trends.

The establishment of a framework for the systematic solution of fiscal problems is Musgrave's most significant contribution. His work combined theory, institutional and historical information, a deep understanding of prior work and empirical testing. Like a number of other outstanding economists educated during the turbulent 1930s, he emphasized the practical and concrete applications of academic research in the belief that ‘intelligent conduct of government is at the heart of democracy’, and until the end of his life was an active commentator on policy issues. A lovely delineation of his views, along with a contrasting perspective, is found in Buchanan and Musgrave (1999); see also his review of the evolution of ideas on fiscal policy (1987).

See Also

- horizontal and vertical equity
- merit goods
- public finance

Selected works

1939. The voluntary exchange theory of public economy. *Quarterly Journal of Economics* 53, 213–37.

1944. (With E.D. Domar.) Proportional income taxation and risk taking. *Quarterly Journal of Economics* 58, 388–422.

1948. (With T. Thin.) Income tax progression. *Journal of Political Economy* 56, 498–514.

1951. (With J.J. Carrol, L.D. Cook and L. Frane.) Distribution of tax payments by income groups: a case study for 1948. *National Tax Journal* 4, 1–53.

1953. On incidence. *Journal of Political Economy* 61, 306–23.

1957. A multiple theory of budget determination. *Finanz Archiv* NS 17, 333–43.

1959. *The Theory of Public Finance: A Study in Public Economy*. New York: McGraw-Hill.

1963. (With M. Krzyzaniak.) *The Shifting of the Corporation Tax*. Baltimore: Johns Hopkins University Press.

1974. (With K.E. Case and H. Leonard.) The distribution of fiscal burdens and benefits. *Public Finance Quarterly* 2, 259–31.

1976. ET, OT, and SBT. *Journal of Public Economics* 6, 3–16.

1987. A brief history of fiscal doctrine. In *Handbook of Public Economics*, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

1999. (With J. Buchanan) *Public Finance and Public Choice: Two Contrasting Visions of the State*. Cambridge, MA: MIT Press.

How to cite this article

Mieszkowski, Peter. "Musgrave, Richard Abel (1910–2007)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000284> doi:10.1057/9780230226203.1150

The New Palgrave Dictionary of Economics Online

music markets, economics of

Frederic M. Scherer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

With the growth of economic prosperity, the demise of feudalism, and the weakening of Western religious institutions, markets for music have been transformed radically. By the 19th century, freelance composition and performance endeavours outweighed the employment of musicians by churches and noble courts. Further changes came from the invention of electrical and then electronic means of recording and disseminating music. The emergence of copyright for musical works strengthened economic incentives for the composition of music.

Keywords

Baumol's cost disease; copyright; music markets; superstars

Article

On 15 January 1787, Wolfgang Amadeus Mozart wrote from Prague to a friend in Vienna that ‘here [in Prague] nothing is talked about but *Figaro*; nothing is played, tootled, sung or whistled but [Mozart's *Marriage of*] *Figaro*.’ Music was ubiquitous, and Mozart was at the time Prague's favorite composer. More than two centuries later, music is played and listened to incessantly, usually through some electronic medium, in homes, shops, automobiles, trains, and on the streets. But the diversity of composers and forms is much greater. And in the means by which music is created and reaches the ears of its countless appreciators, the market institutions have changed radically.

Early history

The history of musical performance is as old as the history of humanity. A seven-hole Chinese flute has been carbon-dated to the year 7,000 bc. Prehistoric tribes celebrated military events and other special occasions with music from drums, horns, flutes, and a variety of stringed instruments. By the Middle Ages in Europe, the professional performance of music was concentrated in the churches, following traditions inherited from the Judaic temple music of King David, and in the residences of the wealthy, especially the nobility. The Roman Schola Cantorum was founded in the seventh century ad to perform what came to be called Gregorian chant. The first Holy Roman Emperor, Charlemagne (d. 814), imported from Rome a delegation of 12 specialists to propagate the correct use of Gregorian chant in northern Europe. Chapels established in residences of the nobility maintained their own cadre of instrumentalists and singers. Competition between Protestant and Roman Catholic denominations during the 16th century led to innovations in the richness of church music, ranging from the eminently singable hymns of Martin Luther to the polyphonic masses of Giovanni da Palestrina. The musicians employed in noble chapels also provided entertainment at dinners and celebrations, and during the Renaissance period wealthier nobles initiated further specialization, maintaining one group of musicians for chapel and another for secular entertainment. During the second half of the 17th century, a kind of ‘cultural arms race’ emerged among the hundreds of noble courts in Germany, Bohemia, and Austria. Each court competed for prestige through the quality of the musicians and composers it employed to entertain visitors (see Elias, 1969; Baumol and Baumol, 1994).

As a golden age of classical music dawned in the 17th century, much of Europe was organized along feudal lines. There was an active market for the hiring of promising musicians, who travelled far and wide in search of the best employment opportunities. But once a musician was retained by a feudal lord, at least throughout much of the European continent, he

(seldom she) was often bound to continued servitude at the noble's whim and on the noble's terms. Claudio Monteverdi was able to leave his badly paid, demoralizing position with Duke Vincenzo I of Mantua only after his employer's death in 1612. Johann Sebastian Bach was imprisoned for nearly a month in 1717 when he sought to leave the service of the Duke of Weimar. His contemporary Georg Friedrich Händel was advised by friends to reject an employment offer from the King of Prussia (Scherer, 2004, p. 94):

> For they well knew, that if he once engag'd in the King's service, he must remain in it, whether he liked it, or not; that if he continued to please, it would be reason for not parting with him; and that if he happened to displease, his ruin would be the certain consequence.

When he was discharged in an economy move during 1769, Niccolò Jommelli was denied permission to take with him copies of the music he had written for the Duke of Württemberg.

Gradually, however, a new set of opportunities materialized for musicians to earn a living as freelance artists. Opera was the forerunner of this new tradition (see Bianconi and Pestelli, 1998). Having pioneered the first modern opera *Orpheo* under ducal auspices at Mantua, Monteverdi migrated to the free city of Venice, where operas were financed by a consortium of wealthy patricians, organized by a hired impresario, and written and performed under contracts individually negotiated with composers, librettists, and soloists. The paradigm spread to other parts of Italy, then to England and parts of Germany, and eventually to other European nations and the United States. Opportunities for the performance of instrumental music at private locales also began to emerge. One predecessor appeared in mercantile London, where King Charles II, embarrassed over his perennial money problems and his inability to pay his court musicians adequately, allowed Henry Purcell and others to perform their music privately in local theaters, taverns, and music halls. In 1697 Thomas Hickford opened a ‘Great Dancing Room’ in London, perfecting the emerging model for private music halls. In 1735 Vauxhall Gardens, south-east across the Thames from today's Victoria Station, began offering open-air summer concerts at admission prices sufficiently modest to draw Londoners of nearly all economic classes (see McVeigh, 1993). These innovations spread to other locations in London and then to many parts of the European continent. By the third decade of the 19th century, private ballrooms had proliferated in Vienna to the point at which they could accommodate some 50,000 music lovers simultaneously, with entertainment provided, inter alia, by 300 musicians under contract to Johann Strauss, sen., and deployed by Strauss in groups of 25. His son Johann was paid $100,000 to conduct his own and others’ compositions at the Boston, Massachusetts, Peace Festival in 1872, performed in a huge wooden shed by an orchestra of 2,000 and chorus of 20,000 before audiences of approximately 100,000 persons.

The transition from church and court employment to freelance music composition is depicted in Figure 1. (Scherer, 2004, pp. 69–71). It summarizes by 50-year birth cohort intervals the principal occupational choices of 646 musical composers of enduring fame born between 1650 and 1849. Strong downward trends are evident for court and church employment along with an upward trend for freelance activity. With double-counting allowed to reflect multiple career phases, we see that the fraction employed in noble courts or regularly subsidized by them fell from 62.4 per cent for composers born between 1650 and 1699 to 19.0 per cent for those born in the first half of the 19th century, by which time the Napoleonic wars had undermined much of the feudal system. For church employment the sharpest decline occurs for composers born in the second half of the 18th century. The fraction earning a significant component of their living through freelance composition and performance activities increased from 35.5 per cent for composers born in 1650–99 to 81 per cent for those born in 1800–49.

Figure 1

Trends in composers’ principal modes of employment, 1650–1849

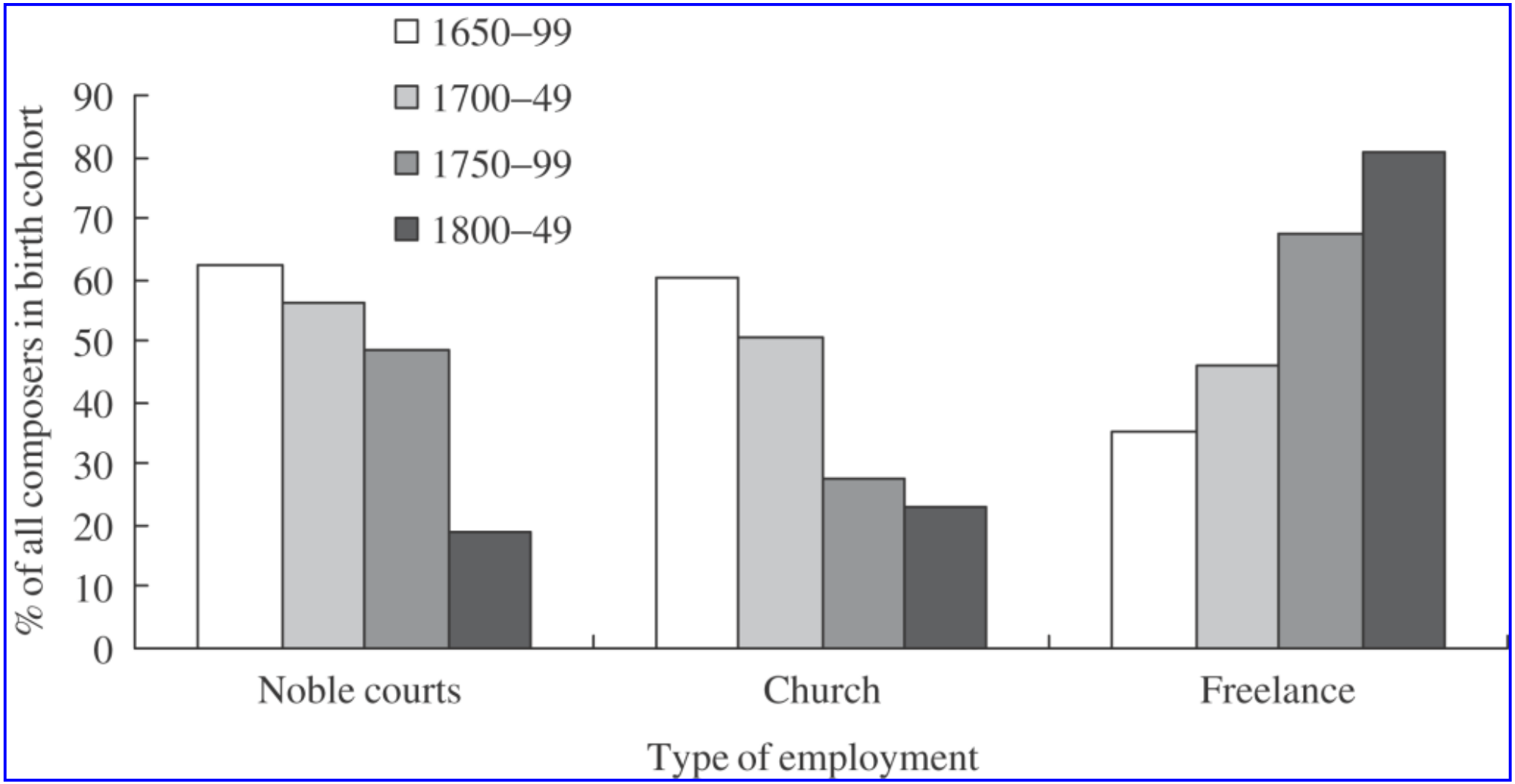

Music market organization

Markets for music are both vertically and horizontally complex. Final demand exists for hearing music performed or for performing it oneself. From that demand are derived a host of other demands: for new musical compositions, for the sheet music through which compositions are disseminated to performers, for training (for example, at conservatories and local schools) in performance, for the concerts and other venues at which music is performed, for the instruments with which it is performed, and for recorded means by which performed music is propagated more widely. The composition, instrument-making, and dissemination stages have for many centuries experienced particularly vigorous innovation. In some subsets, however, such as organ building and violin-making, the technology attained a remarkable degree of perfection as early as the 17th century.

Although data permitting a direct statistical test have not been available, the growth of concert-going during the 18th and 19th centuries in tandem with the commercial and industrial revolutions in Europe implies a substantial income elasticity of demand for music consumption. Indirect evidence is presented in Figure 2 (from Scherer, 2004, p. 35), showing trends in the production of pianos in the United States between 1850 and 1939. Values for years other than those on which specific data were available (points) are interpolated. The implied income elasticity of demand is in the range of 2.4 to 4.3, depending upon what other variables are included in multiple regressions. There is a sharp and lasting break in the series during the mid-1920s, when an economic boom was in full swing. The 1909 and 1923 production peaks were not surpassed over the next 60 years, after which imports began to outweigh domestic production. Two coincident events are responsible for the mid-1920s slump: the introduction of electrical phonographs (with fidelity superior to acoustic phonographs marketed successfully in the 1890s) and the advent of radio broadcasting, including the transmission of classical and popular music. Up to that time, the principal

alternative to expensive concert-going (or free summer concerts in urban parks) was the active performance of music within one's home. After the mid-1920s, music could be enjoyed passively at home by listening to radios and phonographs. An era of participatory family musicales began to fade, and a new era dawned.

Figure 2

Trends in US piano production, 1850–1939

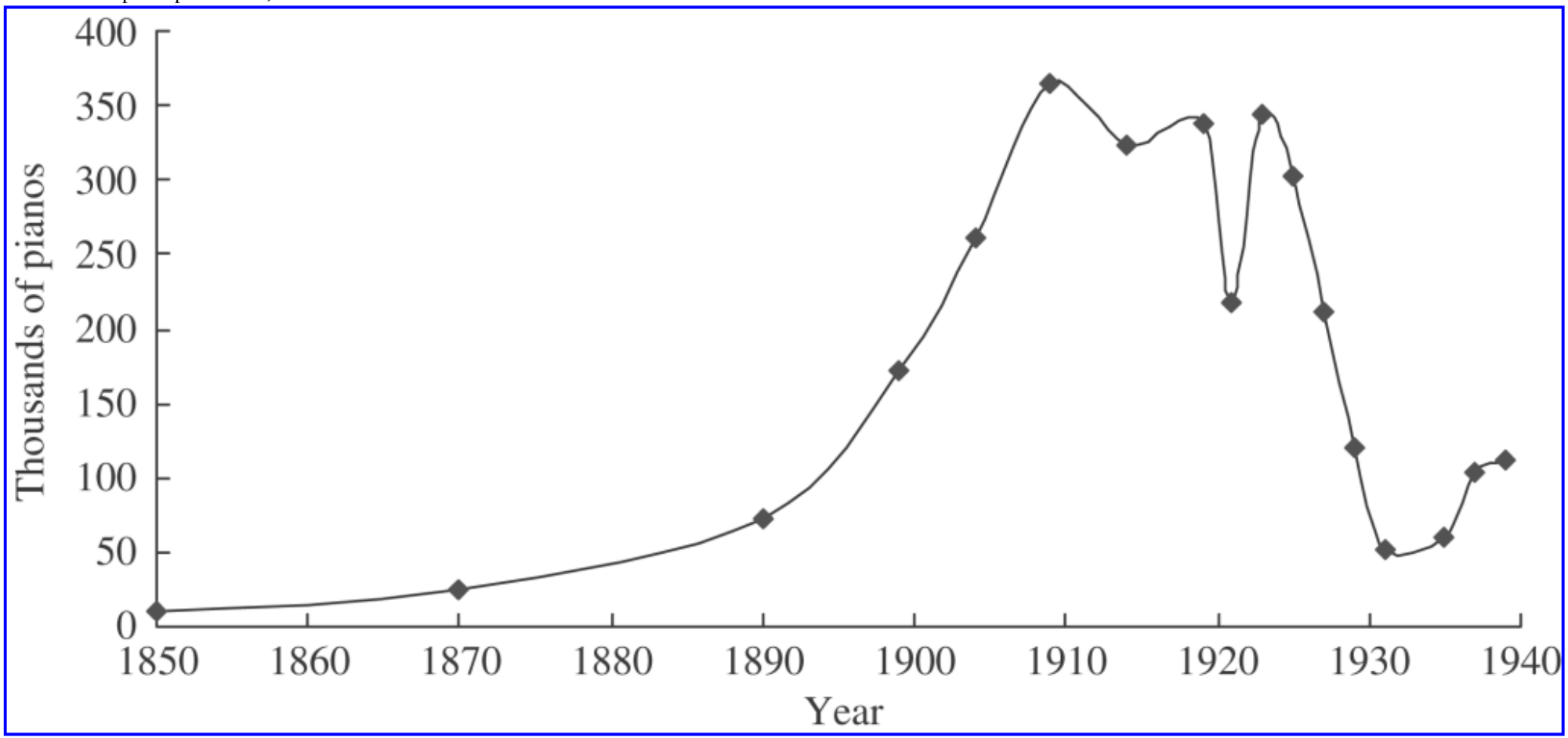

The marriage of electronics with music wrought further radical changes in markets for music. Through records, radio, television, and still later, the internet, audiences for musical performance were no longer limited to those who could be assembled to hear a specific concert. The whole world was a stage, with four noteworthy consequences. First, for the world as a whole, musical record sales in 1998 (if we count only those sold legally, consistent with applicable copyrights) amounted to more than four billion units. Second, through amplification live performances could be heard by unprecedented numbers of concert-goers. The Woodstock Festival of August 1969 attracted an estimated 300,000 to 500,000 participants. Third, the expansion of potential audiences enhanced incentives for product differentiation. New musical styles proliferated during the second half of the 20th century at an accelerating pace. Fourth, the prerequisite for success as a vocal performer was no longer a beautiful voice that could carry through the expanse of an opera house. Electronics made popular acclaim attainable for faint voices, and even for performers whose histrionics, dancing ability, costuming, and sex appeal outweighed their vocal talent.

The expansion of markets also intensified a phenomenon already in evidence at the start of the 18th century: superstardom. The received theory (see, for example, Rosen, 1981) asserts that the broader the market for talent is, the higher the income differential tends to be between performers with the greatest ability to please and performers of inferior talent. In 1998, for example, the Three Tenors (Luciano Pavarotti, Placedo Domingo, and Jose Carreras) along with their agent received an advance of $18 million for a single performance accompanying the World Cup football finals in Paris, including broadcast and recording rights. Michael Jackson's '*Thriller*' album, introduced in 1992, achieved an all-time world high of 46 million unit sales, and in 2002, before he plunged into legal and financial difficulties, Jackson's net financial worth was estimated to be in the order of $300 million. But superstardom was not entirely new. In the early 18th century, the leading opera singers made arduous journeys throughout Europe in quest of the most remunerative engagements. The most famous of them all, the castrato Farinelli (Carlo Broschi), is said to have earned £5,000 during the 1735–36 opera season in London at a time when an English building

craftsman averaged £30 a year.

For live musical performances that are neither amplified nor broadcast, another economic law operates, known as ‘Baumol's cost disease’ (Baumol and Bowen, 1965). Many musical works require a more or less fixed complement of musicians expending a nearly fixed amount of rehearsal and performance time. Thus, labour productivity hardly grows from one century to the next. Meanwhile, most other goods and services experience appreciable rates of productivity growth, permitting those who supply them to earn increasing real incomes over time. For musical performers to stay abreast economically with alternative high productivity growth vocational opportunities, musicians’ hourly pay levels must rise apace commensurately, which means that the costs of live musical performances increase relative to the prices of all other goods and services, threatening possibly severe adverse substitution effects. To maintain a thriving supply of live musical performances, subsidization becomes increasingly necessary – not by noble patrons, as in the 18th century, but by governments (preponderantly in Europe and Asia) or affluent concert-goers and private philanthropists (the United States pattern).

For 42 leading US symphony orchestras, all unionized, admissions receipts during the 2002–03 season defrayed on average only 43 per cent of annual budgets. Balancing budgets (which was often not achieved) required voluntary contributions and drawing upon endowments (the latter varying from virtually nothing to $248 million, with a mean of $48 million and median of $19 million). A regression analysis spanning 1980 to 2002 revealed that those orchestras’ budgets were higher, if local population is also taken into account, the greater the concentration of manufacturing, mining, and service corporation headquarters assets was in the relevant metropolitan area. A local corporate headquarters presence subsidized symphony orchestra performance both directly through endowment contributions and through the annual donations of well-paid company officials (Scherer, 2005).

Music publication and copyright

For at least three centuries the composers and publishers of new music have complained about the unauthorized use, or ‘piracy’, of their works. The copyright system – having governments confer upon composers and publishers (including record producers) exclusive rights to their productions, which can then be licensed to others upon payment of royalties and/or performance fees – has been the standard means of compromising the maintenance of economic incentives for creative contributions against widespread public dissemination. The first formal copyright law was enacted in England in 1709, but it was interpreted initially not to cover musical works. Extension to musical works came first in 1777 through a lawsuit brought in England by Johann Christian Bach, the son of Johann Sebastian Bach. Musical works were then included under copyright laws passed in the United States, France, various German states, and then, thanks to an initiative led by Johann Nepomuk Hummel and Ludwig van Beethoven culminating in 1837, the German, Austrian, Italian, Czech, and Hungarian territories that previously comprised the Holy Roman Empire.

Prior to the enactment of copyright laws, some protection against unauthorized use was provided by ‘privileges’ – ad hoc grants of exclusivity conferred upon composers or publishers by royal sovereigns. Securing such grants required access to the relevant sovereign and, in the politically fragmented territories of the old Holy Roman Empire, the grants were mostly localized and prone to being undermined by competitors producing and selling from another territory. Composers protected their works by contracting with publishers having a reputation for respecting their contracts and keeping manuscripts secret for as long as possible before published versions reached the market. Hand-copying posed a particular threat, for in the early days of the music publishing industry's rapid growth – for example, around 1800 – a copyist could turn out copies by hand at a unit cost lower than the average front-end set-up costs plus variable costs incurred with mechanical printing for production runs of fewer than 25–40 copies (Scherer, 2004, p. 162). Like his contemporaries, Mozart attempted to prevent hand copyists from pirating his works by keeping the copyists he hired under constant supervision and dividing work on any given manuscript among multiple copyists. Publishers combatted piracy through secrecy, announcing fixed prices lower than copyists’ minimum costs for publications expected to secure a considerable volume (which would now be called ‘limit pricing’), keeping composers’ honoraria low for works of limited or uncertain appeal, and entering into collusive anti-piracy agreements with fellow publishers.

Giuseppe Verdi and his publisher Giovanni Ricordi were the first to make aggressive use of the copyright laws enacted in German-speaking and Austrian-controlled regions (for example, northern Italy). Previously, local opera houses had purchased or leased manuscripts at cut-rate prices from copyists. With copyright and a network of local enforcement employees, Verdi and Ricordi were able to extract fees from each house, graduating them in a discriminatory fashion to extract more revenue from those serving large, affluent audiences than those located in small provincial towns. They were also particularly energetic in publishing ‘reductions’ of each separate overture, aria, and chorus, along with bundles covering a full opera, for a diversity of instruments – for example, voice, piano, violin, flute, clarinet, and various ensembles – played by middle-class citizens in their homes. In this way they were able to create a mass market for their works, and as a result Ricordi could pay unprecedentedly large sums for the rights to publish Verdi's works. Verdi became quite rich, accumulating an estate equivalent to nearly £40,000 at the time of his death in 1901 (when English building craftsmen's annual income averaged £100) and beginning semi-retirement at his Busetto villa in the fifth decade of his nearly nine-decade life.

Verdi's extensive written correspondence leaves little doubt that, as his fortune grew, he consciously reduced his work effort along a backward-bending supply curve. Few 18th and 19th century composers achieved as much prosperity as Verdi did; the terminal wealth distribution is highly skew. (Gioachino Rossini became even wealthier and, after reaching the age of 37, spent the remaining four decades of his life in retirement.) It is unlikely that the majority of composers found themselves on the backward-bending portion of a labour supply curve. It is not unreasonable to suppose that the spectacular financial successes achieved by a relatively few composers under the copyright laws inspired many others to try their luck at musical composition. An attempt to test this hypothesis quantitatively (Scherer, 2004) was inconclusive, largely because of the difficulty of holding other relevant

variables constant. What can be said, however, is that the lack of copyright laws did not prevent classical music from experiencing its golden age of creativity before copyright protection was available in the most musically productive parts of Europe, that is, before the death of Beethoven in 1827 and Schubert in 1828. Despite this limping recommendation, advocates for copyright have been successful in extending greatly both the length of time for which creative individuals and publishers can be protected and, given a continuing stream of new technological challenges, in the range of media over which copyright applies (see Lessig, 2004).

See Also

- art, economics of
- intellectual property
- superstars, economics of

Bibliography

Baumol, W. and Bowen, W. 1965. On the performing arts: the anatomy of their economic problems. *American Economic Review* 55, 495–502.

Baumol, W. and Baumol, H. 1994. On the economics of musical composition in Mozart's Vienna. *Journal of Cultural Economics* 18, 171–98.

Bianconi, L. and Pestelli, G., eds. 1998. *Opera Production and Its Resources*. Chicago: University of Chicago Press.

Elias, N. 1969. *Die höfische Gesellschaft* [The Courtly Society]. Frankfurt: Suhrkamp Verlag.

Lessig, L. 2004. *Free Culture*. New York: Penguin.

McVeigh, S. 1993. *Concert Life in London from Mozart to Haydn*. Cambridge: Cambridge University Press.

Raynor, H. 1972. *A Social History of Music from the Middle Ages to Beethoven*. New York: Schocken.

Rosen, S. 1981. The economics of superstars. *Journal of Political Economy* 71, 845–57.

Scherer, F. 2004. *Quarter Notes and Bank Notes: The Economics of Music Composition in the 18th and 19th Centuries*. Princeton, NJ: Princeton University Press.

Scherer, F. 2005. Corporate structure and the financial support of U.S. symphony orchestras. Working paper.

How to cite this article

Scherer, Frederic M. "music markets, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000216> doi:10.1057/9780230226203.1151

The New Palgrave Dictionary of Economics Online

Muth, John F. (1930 – 2005)

F. Robert Jacobs

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

John F. Muth, Professor of Operations Management, is known for his seminal work in rational expectations, aggregate planning and production scheduling. He received his Ph.D. from Carnegie Tech and spent most of his academic career at Indiana University. A colleague for over 20 years, in this article we give insight into his eclectic interests and intellectual motivation.

Keywords

aggregate planning; artificial intelligence; innovation cycles; linear decision rule; Muth, J. F.; rational expectations

Article

John F. (Jack) Muth was a brilliant individual, though somewhat awkward socially, and little understood by most people. He was born in Chicago, where his father worked as an accountant at a national accounting firm. Eventually, Jack moved with his parents and two brothers to St. Louis, Missouri. He was very weak as a youngster, suffering from severe asthma and allergies. An avid reader, Jack loved playing the cello and studying mathematics. Jack's cello-playing days continued through the 1980s, and he was a member of the Bloomington symphony orchestra for many years. He studied industrial engineering at Washington University in St. Louis, and continued with graduate work in mathematical economics at Carnegie Tech in Pittsburgh, Pennsylvania. His thesis advisor was Franco Modigliani, with Herb Simon and Merton Miller serving on his committee. All three individuals would later become Nobel laureates in economics.

While a doctoral student, Muth was the first recipient of the Alexander Henderson Award in 1954 (for his work in economics). While finishing his doctorate, he spent the 1957–8 academic year as visiting lecturer at the University of Chicago, returned to Carnegie Tech as an assistant professor during 1959 to 1961, spent the 1961–2 academic year at the Cowles Foundation at Yale University, and finally returned to Carnegie Tech as an associate professor without tenure from 1962 to 1964. It is said that it took him

very long to graduate because he did not see the need to take a foreign language examination which would have completed requirements for the Ph.D. degree. Eventually, a colleague whose wife was a French instructor joined the faculty. She tutored Muth in French, and he was finally allowed to graduate. He went on to Michigan State as a professor in 1964, and moved to Indiana University in 1969. He stayed at Indiana University until he retired in 1994.

Throughout his entire academic career John Muth loved to challenge conventional thought. He would explore alternative explanations mathematically, his most famous work being three papers that develop the rational expectations hypothesis (1960; 1961; 1981). Later work by Robert Lucas, the economist, popularized the idea of rational expectations, and Lucas received the Nobel Prize for his efforts. Esther-Mirjam Sent has written a comprehensive paper that describes Muth's work on rational expectations from a historical perspective (Sent, 2002).

Many have asked why Muth did not himself further develop his ideas. If you knew him, this is easy to explain. He knew that there were alternative ways to explain the macroeconomic relationships that were the hot topic of the day. All he wanted to do was show an alternative; in essence, to create an academic debate that challenged conventional wisdom. His colleagues at Carnegie Tech were heavily involved in related research, and he wanted to have some fun and add his thoughts at the same time. Whenever he saw an opening to challenge an idea, he enjoyed developing his elegant ideas, often running concise computer simulations to accompany his mathematical models, subsequently writing up his results. He started doing this very early in his academic career.

A true intellectual, Muth had little interest in promoting his ideas through workshops, presentations or other academic portals. He felt his papers would be interpreted and stand the test of time. Being a good friend, I remember on many occasions Jack talking about invitations to speak at international conferences and at schools. These invitations were usually declined, I am sure not because he was uninterested but rather because he felt these activities would take a significant amount of time, he would probably have trouble with his allergies, and he was more interested in working on his current ideas. He always had something that he was actively working on and would talk about these ideas often over a few beers in the late afternoon at Nick's in Bloomington, Indiana, with his friends.

The late 1960s and early 1970s were spent developing industrial scheduling theory in the field of operations management. He wrote about the importance of the ‘aggregate planning’ problem and established it in the literature in 1960 with his colleagues at Carnegie Tech (Holt, Modigliani, Muth and Simon, 1960). It was Muth who established the proof of the linear decision rule in aggregate planning. This effort developed into a series of books published with Gerry Thompson and Gene Groff (Muth and Thompson, 1963; Groff and Muth, 1969; 1972).

He spent the late 1970s through the early 1980s studying artificial intelligence. His main interests were in inference engines and inductive and deductive logic. To my knowledge he published only one paper on the topic (Jacobs, Hancock, Mathieson and Muth, 1991). I often heard him refer to his work on artificial intelligence as his ‘ten-year sink hole’.

Later in the 1980s he began studying innovation cycles. He would often muse on the fact that many of the most innovative ideas were developed by individuals working at home, and how corporations that spent gigantic sums to develop new ideas so often produced only minor incremental innovations. He wrote simple simulation programs that simulated a random progress function, and matched these results with what was documented in the literature, often musing on the fit. He published an important paper in this area in 1989.

During the late 1980s until his retirement in 1994, Muth spend his time teaching undergraduate courses in process design and scheduling. As one might imagine, he was an awkward teacher and often had difficulty coming down to the level of doctoral students, much less undergraduate students. When he realized during this time that he was going to have to teach undergraduate students to see out his career, it was interesting to observe how he worked to improve. He worked with the teaching resource group at the university, who videotaped his lectures and helped him develop a better teaching style. His colleges in the department were amazed when he was listed as a recommended instructor in the student newspaper in the early 1990s, an event that gave him great personal satisfaction.

He loved sailing in the Florida Keys and had a 30-foot Auburn sailboat that was docked in Marathon until he moved it to Cudjoe Key around 1989. The boat was well suited for sailing around the Florida Keys having a shallow keel. He retired in 1994 and initially split his time between Bloomington and the Keys. For a time, he worked as a consultant to the business school to develop the integrative cases used in the undergraduate core curriculum, taking Indiana's integrative core to yet another level, as he had done in economics and in almost any area in which he became involved. From around 2000, he remained permanently in the Keys.

This article has possibly not emphasized enough the impact of Muth's work. He was truly a brilliant intellectual and influenced numerous doctoral students throughout his career at Michigan State and Indiana University. As an aside, he was also an amazing Trivial Pursuits player. You always wanted Muth on your team since he seldom missed a question! Muth also had a private side as an aggressively loyal and caring person to his close friends. He was always willing to spend time to talk through important career decisions, and always willing to comment quickly and brilliantly on a manuscript (although it might cost you a beer).

Finally, a funny story, I can still remember being in the Keys with my wife and children, and visiting Jack when he first bought the Cudjoe Key house. Late one afternoon we were all driving up from Key West with Jack. We stopped at a store to pick up some food for dinner, my wife and I leaving the kids with Jack in the car. When we returned to the car, there we saw Jack teaching our two young daughters how to make ‘unusual noises’ by putting their hands over their armpits and pumping their arms up and down. We all still laugh when we think about that time and the other wonderful times we enjoyed with that nervous little genius who was such a great friend.

Selected works

1960. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55, 299–306.

1960. (With C.C. Holt, F. Modigliani and H.A. Simon.) *Planning Production, Inventories, and Work Force*. Englewood Cliffs, NJ: Prentice-Hall.

1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

1963. (With G.L. Thompson.) *Industrial Scheduling*. New York: Prentice-Hall.

1969. (With G.K. Groff.) *Operations Management: Selected Readings*. Homewood, IL: Irwin.

1972. (With G.K. Groff.) *Operations Management: Analysis for Decisions*. Homewood, IL: Irwin.

1981. Estimation of economic relationships containing latent expectations variables. In *Rational Expectations and Econometric Practice*, ed. R.E. Lucas and T.J. Sargent. Minneapolis: University of Minnesota Press.

1986. Search theory and the manufacturing progress function. *Management Science* 32, 948–62.

1989. A stochastic theory of the generalized Cobb–Douglas production function. In *Cost Analysis Applications of Economics and Operations Research*, ed. T.R. Gulledge, Jr. and L.A. Litteral. New York: Springer-Verlag.

1991. (With F.R. Jacobs, T. Hancock and K. Mathieson.) A rule–based system to generate NC programs from CAD exchange files. *Computers & Industrial Engineering* 20, 167–76.

Bibliography

Sent, E.-M. 2002. How (not) to influence people: the contrary tale of John F. Muth. *History of Political Economy* 34, 291–319.

How to cite this article

Jacobs, F. Robert. "Muth, John F. (1930–2005)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000408> doi:10.1057/9780230226203.1152

The New Palgrave Dictionary of Economics Online

Myrdal, Gunnar (1898 – 1987)

Paul Streeten

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cumulative causation; economic development; ex ante and ex post; expectations; Myrdal, G.; non-economic variables; planning; population policy; racial disadvantage; stagflation; underdevelopment

Article

Gunnar Myrdal was born in the province of Dalarná in Sweden. He attributed his faith in Puritan ethics and his egalitarianism to his sturdy farming background.

He was a student of the giant figures Knut Wicksell, David Davidson, Eli Heckscher, Gösta Bagge and above all Gustav Cassel. His personal friendship was warmest with Cassel, to whose chair in Political Economy at Stockholm University he succeeded (1933–9).

At first a pure theorist, Myrdal's year in the United States as a Rockefeller Fellow, following the crash of 1929, turned his interests to political issues. On his return to Sweden from America he, with his wife Alva, became active in politics. In 1935 he became a Member of Parliament. Together, they pioneered modern population policy. His involvements in Swedish politics between 1931 and 1938 turned him from a theoretical economist into a political economist and what he himself describes as an institutionalist. In 1938 the Carnegie Corporation selected him for a major investigation of the Negro problem in America, a project which resulted in *An American Dilemma* (1944a).

He returned to Sweden in 1942 and for five years was again involved in political activities. He headed the committee that drafted the social democratic post-war programme. He returned to Parliament and became a member of the board of directors of the Swedish Bank, chairman of the Swedish Planning Commission, and Minister for Trade and Commerce (1945–7). As Minister he arranged for a highly controversial treaty with the Soviet Union and was also involved in controversy over the dismantling of wartime controls. In 1947 he became Executive Secretary of the United Nations Economic Commission for Europe, to which he recruited an outstandingly able team. After ten years with the Commission in Geneva he embarked on a ten-year study of development in Asia, the result of which was the monumental *Asian Drama* (1968). In 1973 he was awarded the Nobel Prize in economics jointly with Friedrich von Hayek.

Methodological questions occupied Myrdal's thought throughout his life. They were already present in the young Myrdal's *Political Element in the Development of Economic Theory* (1930, English edition 1953). It was under the influence of the remarkable Uppsala University philosopher Axel Hägerström that he had begun to question the wisdom of the economic establishment.

Myrdal's doctoral dissertation on price formation and economic change (1927) introduced expectations systematically into the analysis of prices, profits and changes in capital values. The microeconomic analysis focused on planning by the firm. Many of these ideas were used in his later macroeconomic work, including *Monetary Equilibrium* (1931, English expanded translation 1939).

Much confusion had been caused by the lack of distinction between anticipations and results. The concepts *ex ante* and *ex post* that Myrdal developed greatly clarified the discussion of savings, investment and income, and their effects on prices. In anticipation, intention and planning, savings can diverge from investment; after the event they must be identical, because the community can save only by accumulating real assets. It is the process by which anticipations *ex ante* are adjusted so as to bring about the bookkeeping identity *ex post* that explains unexpected gains and losses as well as fluctuations in prices. Only in equilibrium are *ex ante* savings equal to *ex ante* investment, so that there is no tendency for prices to change. By introducing expectations into the analysis of economic processes he made a major contribution to liberalizing economics from static theory, in which the future is like the past, and to paving the way for dynamics, in which time, uncertainty and expectations enter in an essential way.

What is common to his three important later books, *The Political Element* (1930), *American Dilemma* (1944a) and *Asian Drama* (1968) is the emphasis on realistic and relevant research, whether on economic problems, race relations, or world poverty, and with it the effort to purge economic thinking of systematic biases.

Starting on the study of Blacks in the United States, he soon discovered that he had to study ‘American civilization in its entirety, though viewed in its implications for the most disadvantaged population group’ (Introduction to *An American Dilemma*, Section 4). The way to reach objectivity was to state explicitly the value premisses of the study. These premisses were not chosen arbitrarily, but were what Myrdal called the ‘American Creed’ of justice, liberty and equality of opportunity. But while these value premisses were chosen for their relevance to American society, they corresponded to Myrdal's own valuations. The major contribution of the book, which Myrdal regarded as his war service, is the analysis of six decades after Reconstruction as a ‘temporary interregnum’ not a ‘stable equilibrium’, and of the incipient changes, on which the prediction of the Black revolt in the South was based.

Apart from his work on expectations and on racial problems, Myrdal is best known for his critique of conventional economic theory applied to underdeveloped countries.

Through his whole work run five lines of criticism of mainstream economic and social theory. First, his appeal for realism is not a critique of abstraction. His criticism is that irrelevant features are selected and relevant ones ignored (‘opportunistic ignorance’). A second line of criticism has been the narrow or abstract definitions of development, economic growth, or welfare. The actual needs and valuations of people, not the abstractions of statisticians or the empty concepts of metaphysicians, should be the basis for formulating aims. His third line of criticism is directed at the narrow definitions and the limits of disciplines. The essence of the institutional approach, advocated by Myrdal, is to bring to bear all relevant knowledge and techniques on the analysis of a problem. In an interdependent social system there are no economic, political or social problems, there are only problems. His fourth line of criticism is directed at spurious objectivity which, under the pretence of scientific analysis, conceals political

valuations and interests. Myrdal argues that this pseudoscience should be replaced by explicit valuations. He is, of course, aware of the complex nexus between valuations and facts but, ever since his youthful *Political Element*, has constantly fought the inheritance of natural law and utilitarianism, according to which we can derive recommendations from pure analysis. A fifth line of criticism is directed against biases and twisted terminology. He lays bare the opportunistic interests and the ‘diplomacy’ underlying the use of such concepts as ‘United Nations’, ‘international’, ‘values’, ‘welfare’, ‘developing countries’, ‘unemployment’, ‘the free world’. The features against which these lines of criticism are advanced are combined in the technocrat. He isolates economic (or other technical) relations from their social context; he neglects social and political variables and thereby ministers to the vested interests that might otherwise be hurt; he pretends to scientific objectivity and is socially and culturally insensitive. Since the majority of experts, academics and planners are of this type, he has ruffled many feathers.

The question may be asked whether the narrow technocrat cannot be replaced by an approach that introduces social variables openly into the formal model? Myrdal's answer would be, yes and no. In certain areas, a widening or redefinition of concepts can be successful. The productive effects of better nutrition can be studied and the line between investment and consumption be redrawn. The influence of climate, of attitudes, and of institutions can be introduced as constraints or as variables. An agricultural production function can be constructed in which health, education, distance from town, and so on figure as ‘inputs’. ‘Capital’ can be redefined so as to cover anything on which expenditure of resources now raises the flow of output later.

But there are limits to such revisionism. These limits apply both to the analysis of facts and to recommendations of policies. On the factual side, the reformulation runs into difficulties if the connection between expenditure now and ‘yield’ later is only tenuous, as in the initiation of a birth control programme or a land reform.

In the analysis of values, the construction of a social welfare function is not, in Myrdal's view, a logical task. The unity of a social programme of a party is unlike that of a computer program or a logically consistent system, and more like the unity of a personality. It is discovered not only by deductive reasoning but by empathy, imagination, and even artistic and intuitive understanding. Means and ends, targets and instruments, are misleading ways of grasping the valuations of a class, an interest group or a whole society, for their unity is not logical but psychological.

In *Asian Drama* the explicitly formulated valuations are the ‘Modernization Ideals’. A list would include rationality, planning for the future, raising productivity, raising levels of living, social and economic equalization, improved institutions and attitudes, national consolidation, national independence, political democracy, social discipline.

An important idea in Myrdal's arsenal of ideas is that of circular or cumulative causation (or the vicious – or virtuous – circle), first fully developed in *An American Dilemma*. It postulates increasing returns through specialization and economies of scale and shows how small advantages are magnified.

The principle goes back to Wicksell who, in *Interest and Prices* (1898), had analysed divergences between the natural and the market rates of interest in terms of upward or downward cumulative processes, until the divergence was eliminated. Wicksell pointed out that, if banks keep their loan rate of interest below the real rate of return on capital, they will encourage expansion of production and investment in plant and equipment. As a result, prices will rise and will continue to rise cumulatively as long as the lending rate is kept below the real rate.

The principle of cumulative causation can be used to show movements away from an equilibrium

position as a result of the interaction of several variables. Myrdal has not always been entirely clear in the formulation of this important principle, and there has been the suggestion that any form of circular or mutual causation or interaction is cumulative and hence disequilibrating. This would be false, for a series of mutually caused events can, after a disturbance, rapidly converge either on the initial or on some other point of stable equilibrium. In order to get instability, a cumulative movement away from the initial situation, the numerical values of the coefficients of interdependence have to be above a critical minimum size. For example, an increase in consumption will raise incomes which in turn will raise consumption, and so on, ad infinitum. But as long as the marginal propensity to consume is less than unity, the infinite series will converge on a finite value.

The notion of cumulative causation was applied by Myrdal most illuminatingly to price expectations (*Monetary Equilibrium*) and to the relations between regions (*Economic Theory and Underdeveloped Regions*, 1957; American title: *Rich Lands and Poor*). He showed how the advantages of growth poles can become cumulative, while the backward region may be relatively or even absolutely impoverished. Myrdal applied the notion of sociological variables, such as the prejudices against Negroes and their level of performance (low skills, crime, disease, and so on); to economic variables; and, above all, to the interaction of so-called ‘economic’ and ‘non-economic’ variables. Thus, the relation between better nutrition, better health and better education, higher productivity and hence ability further to improve health, education and nutrition shows that the inclusion of non-economic variables in the analysis opens up the possibility of numerous cumulative processes to which conventional economic analysis is blind. It also guards against uni-causal explanations and panaceas.

The revolutionary character of the concept of cumulative causation is brought out by the fact that interaction takes place not only within a social system in which the various elements interact, but also in time, so that memory and expectations are of crucial importance. The responses to any given variable, say a price, are different according to what the history of this variable has been. It is this dynamic feature of analysis and its implications for policy that distinguishes Myrdal's approach from that of economists who think in terms of general equilibrium.

In *Economic Theory and Underdeveloped Regions* (1957), and later in *Asian Drama* (1968), he used the concepts ‘backwash’ and ‘spread’ effects to analyse the movement of regions or whole countries at different stages of development and the effects of unification. It is a highly suggestive, realistic and fruitful alternative explanation to that of stable equilibrium analysis, usually based on competitive conditions and diminishing returns, and concluding that gains are widely and evenly distributed.

Like the Marxists, Myrdal emphasizes the unequal distribution of power and property as an obstacle not only to equity but also to efficiency and growth. But his conclusion is not Marxist. He regards a direct planning of institutions and shaping of attitudes (what Marx regarded as part of the superstructure) as necessary, though very difficult, partly because he believes that attitudes and institutions are inert, and partly because the policies which aim at reforming attitudes and institutions are themselves part of the social system, part of the power and property structure. There are clearly also logical difficulties in operating on variables that are thought to be fully determined within the system.

In *Asian Drama* Myrdal criticizes the kind of government he calls the ‘soft state’. This critique has sometimes been misunderstood. It is plain that ‘softness’ in Myrdal's sense is quite compatible with a high degree of coercion, violence and cruelty. The Tamils in Sri Lanka, the Indians in Burma, the Chinese in Indonesia, the Hindus in Pakistan, the Moslems in India, the Biharis in Bangladesh – to take six states he calls ‘soft’ – could not claim excessively soft treatment. ‘Soft states’ also go in for military

violence, both internal and external. Their ‘softness’ lies in their unwillingness to coerce in order to implement declared policy goals. It is not the result of gentleness or weakness but reflects the power structure and a gap between real intentions and professions.

Myrdal applied his method to the analysis of inflation combined with widespread unemployment in the developed countries of the West in the 1970s, and either coined or was one of the first to use the term ‘stagflation’. He attributes the situation to the organization of producers as pressure groups, and the dispersion and comparative weakness of consumers, to the tax system which encourages speculative expenditures, to the structure of markets and to the methods of oligopoly administrative pricing, and he condemns inflation as a socially highly divisive force.

The approach favoured by Myrdal is one of neither Soviet authority and force nor of capitalist laissez-faire but of a third way: that of using prices for planning purposes and of attacking attitudes and institutions directly to make them the instruments of reform. His approach has more affinity with those socialists who were dismissed by Marx as utopian. The difficulty is that any instrument, even if used with the intention to reform, within a given power structure may serve the powerful and re-establish the old equilibrium. Even well-intentioned allocations, rationing, licensing and controls may reinforce monopoly and big business. How does one break out of this lock? Myrdal does not draw revolutionary conclusions but relies on the, admittedly difficult, possibility of self-reform that arises, in both the American Creed and in the Modernization Ideals, from the tensions between preferred and proclaimed beliefs and actions.

Both *An American Dilemma* and *Asian Drama* are books about the interaction and the conflict between ideals and reality, and about how, when the two conflict, one of them must give way. Much of conventional economic theory is a rationalization whose purpose it is to conceal that conflict. But it is bound to reassert itself sooner or later. When this happens, either the ideals will be scaled down to conform to the reality or the reality will be shaped by the ideals.

See Also

- ex ante and ex post
- institutional economics

Selected works

1927. *Prisbildningsproblemet och föränderligheten*. Uppsala and Stockholm: Almqvist & Wiksell.

1930. *Vetenskap och politik i nationalokonomien*. Stockholm: P.A. Norstedt.

1931. Om penningteoretisk jämvikt. *Ekonomisk Tidskrift* 33, 191–302.

1932. *Das Politische Element in der nationalökonomischen Doktrinbildung*. Berlin: Junker and Dünnhaupt.

1933a. *The Cost of Living in Sweden 1830–1930*. London: P.S. King & Son.

1933b. Der Gleichgewichtsbegriff als Instrument der geldtheoretischen Analyse. In *Beiträge zur Geldtheorie*, ed. P. Friedrich von Hayek. Vienna: Springer-Verlag (expanded version of Myrdal, 1931).

1933c. Konjunktur och offentlig hushållning. Bihang til riksdagens protokoll, 1st collection, Appendix III.

1933d. Das Zweck-Mittel-Denken in der Nationalökonomie. *Zeitschrift für Nationalökonomie* 4, 305–29.

1934a. *Finanspolitikens ekonomiska verkningar*. Stockholm: P.A. Norstedt.

1934b. (With A. Myrdal.) *Kris i befolkningsfrågan*. Stockholm: A. Bonnier.

1939. *Monetary Equilibrium*. London: Hodge.

1940. *Population: A Problem for Democracy*. Cambridge, MA: Harvard University Press.

1944a. *An American Dilemma: The Negro Problem and Modern Democracy*. New York: Harper. Paperback edn, New York: McGraw-Hill, 1964.

1944b. *Varning för fredsoptimism*. Stockholm: Bonniers.

1945. *Warnung gegen Friedensoptimismus*. Zurich: Europa Verlag.

1953. *The Political Element in the Development of Economic Theory*. Trans. P. Streeten London: Routledge & Kegan Paul; Cambridge, MA: Harvard University Press. (Originally published in German.)

1955. *Realities and Illusions in Regard to Intergovernmental Organisations*. Oxford: Oxford University Press.

1956a. *Development and Underdevelopment: A Note on the Mechanism of National and International Inequality*. Cairo: National Bank of Egypt.

1956b. *An International Economy: Problems and Prospects*. London: Routledge & Kegan Paul.

1957. *Economic Theory and Underdeveloped Regions*. London: Duckworth; New York: Harper.

1958. *Value in Social Theory: A Selection of Essays on Methodology*, ed. P. Streeten. London: Routledge & Kegan Paul.

1960. *Beyond the Welfare State*. New Haven: Yale University Press.

1961. Value-loaded concepts. In *Money, Growth and Methodology and Other Essays in Honor of Johan Åkerman*, ed. H. Hegeland. Lund: Gleerup.

1963. *Challenge to Affluence*. New York: Pantheon.

1968. *Asian Drama: An Inquiry Into the Poverty of Nations*. New York: Twentieth Century Fund.

1969. *Objectivity in Social Research*. New York: Pantheon.

1970a. *The Challenge of World Poverty: A World Anti-poverty Program in Outline*. New York: Pantheon.

1970b. The ‘soft state’ in underdeveloped countries. In *Unfashionable Economics*, ed. P. Streeten. London: Weidenfeld & Nicolson.

1972. *Against the Stream: Critical Essays on Economics*. New York: Pantheon.

Bibliographic addendum

Myrdal's intellectual development is described in J. Angresano, *The Political Economy of Gunnar Myrdal*, Cheltenham: Edward Elgar, 1997. The influence of *An American Dilemma* on public policy has been sufficiently great to generate scholarly studies on its genesis and impact, including W. Jackson, *Gunnar Myrdal and America's Conscience: Social Engineering and Racial Liberalism, 1938–1987*, Durham: University of North Carolina Press, 1990, and D. Southern, *Gunnar Myrdal and Black–White Relations: The Use and Abuse of an American Dilemma, 1944–1969*, Baton Rouge: Louisiana State University Press, 1987.

How to cite this article

Streeten, Paul. "Myrdal, Gunnar (1898–1987)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000288> doi:10.1057/9780230226203.1153

The New Palgrave Dictionary of Economics Online

Nash equilibrium, refinements of

Srihari Govindan and Robert B. Wilson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes ways that the definition of an equilibrium among players' strategies in a game can be sharpened by invoking additional criteria derived from decision theory. Refinements of John Nash's 1950 definition aim primarily to distinguish equilibria in which implicit commitments are credible due to incentives. One group of refinements requires sequential rationality as the game progresses. Another ensures credibility by considering perturbed games in which every contingency occurs with positive probability, which has the further advantage of excluding weakly dominated strategies.

Keywords

backward induction; Bayes' rule; dynamic games; game theory; Harsanyi, J.; imperfect information; Nash equilibrium; Nash, J.; non-cooperative games; perfect information; refinements of Nash equilibrium; sequential rationality; small-worlds criterion; subgame perfection; symmetric games

Article

Game theory studies decisions by several persons in situations with significant interactions. Two features distinguish it from other theories of multi-person decisions. One is explicit consideration of each person's available strategies and the outcomes resulting from combinations of their choices; that is, a complete and detailed specification of the ‘game’. Here a person's strategy is a complete plan specifying his action in each contingency that might arise. In non-cooperative contexts, the other is a focus on optimal choices by each person separately. John Nash (1950; 1951) proposed that a combination of mutually optimal strategies can be characterized mathematically as an *equilibrium*. According to Nash's definition, a combination is an equilibrium if each person's choice is an optimal response to others' choices. His definition assumes that a choice is optimal if it maximizes the person's expected utility of outcomes, conditional on knowing or correctly anticipating the choices of others. In some applications, knowledge of others' choices might stem from prior agreement or communication, or accurate prediction of others' choices might derive from ‘common knowledge’ of strategies and outcomes and of optimizing behaviour. Because many games have multiple equilibria, the predictions obtained are incomplete. However, equilibrium is a weak criterion in some respects, and therefore one can refine the criterion to obtain sharper predictions (Harsanyi and Selten, 1988; Hillas and Kohlberg, 2002; Kohlberg, 1990; Kreps, 1990).

Here we describe the main refinements of Nash equilibrium used in the social sciences. Refinements were developed incrementally, often relying on ad hoc criteria, which makes it difficult for a non-specialist to appreciate what has been accomplished. Many refinements have been proposed but we describe only the most prominent ones. First we describe briefly those refinements that select equilibria with simple features, and then we focus mainly on those that invoke basic principles adapted from single-person decision theory.

Equilibria with simple features

Nash's construction allows each person to choose randomly among his strategies. But randomization is not always plausible, so in practice there is a natural focus on equilibria in ‘pure’ strategies, those that do not use randomization. There is a similar focus on strict equilibria, those for which each person has a unique optimal strategy in response to others' strategies. In games with some symmetries among the players, the symmetric equilibria are those that reflect these symmetries. In applications to dynamic interactions the most useful

equilibria are those that, at each stage, depend only on that portion of prior history that is relevant for outcomes in the future. In particular, when the dynamics of the game are stationary one selects equilibria that are stationary or that are Markovian in that they depend only on state variables that summarize the history relevant for the future. Applications to computer science select equilibria or, more often, approximate equilibria, using strategies that can be implemented by simple algorithms. Particularly useful are equilibria that rely only on limited recall of past events and actions and thus economize on memory or computation.

Refinements that require strategies to be admissible

One strategy is strictly dominated by another if it yields strictly inferior outcomes for that person regardless of others' choices. Because an equilibrium never uses a strictly dominated strategy, the same equilibria persist when strictly dominated strategies are deleted, but after deletion it can be that some remaining strategies become strictly dominated. A refinement that exploits this feature deletes strictly dominated strategies until none remain, and then selects those equilibria that remain in the reduced game. If a single equilibrium survives then the game is called ‘dominance solvable’. An equilibrium can, however, use a strategy that is weakly dominated in that it would be strictly dominated were it not for ties – in decision theory such a strategy is said to be inadmissible. A prominent criterion selects equilibria that use only admissible strategies, and sometimes this is strengthened by iterative deletion of strictly dominated strategies after deleting the inadmissible strategies. A stronger refinement uses *iterative deletion of* (both strictly and weakly) *dominated strategies* until none remain; however, this procedure is ambiguous because the end result can depend on the order in which weakly dominated strategies are deleted.

A particular order is used for dynamic games that decompose into a succession of subgames as time progresses. In this case, those strategies that are weakly dominated because they are strictly dominated in final subgames are deleted first, then those in penultimate subgames, and so on. In games with ‘perfect information’ as defined below this procedure implements the criterion called ‘backward induction’ and the equilibria that survive are among those that are ‘subgame-perfect’ (Selten, 1965). In general a subgame-perfect equilibrium is one that induces an equilibrium in each subgame. Figure 1 depicts an example in which there are two Nash equilibria, one in which A moves down because she anticipates that B will move down, and a second that is subgame-perfect because in the subgame after A moves across, B also moves across, which yields him a higher payoff than down.

Figure 1

Player A moves down or across, in which case player B moves down or across. Payoffs for A and B are shown at the end of each sequence of moves

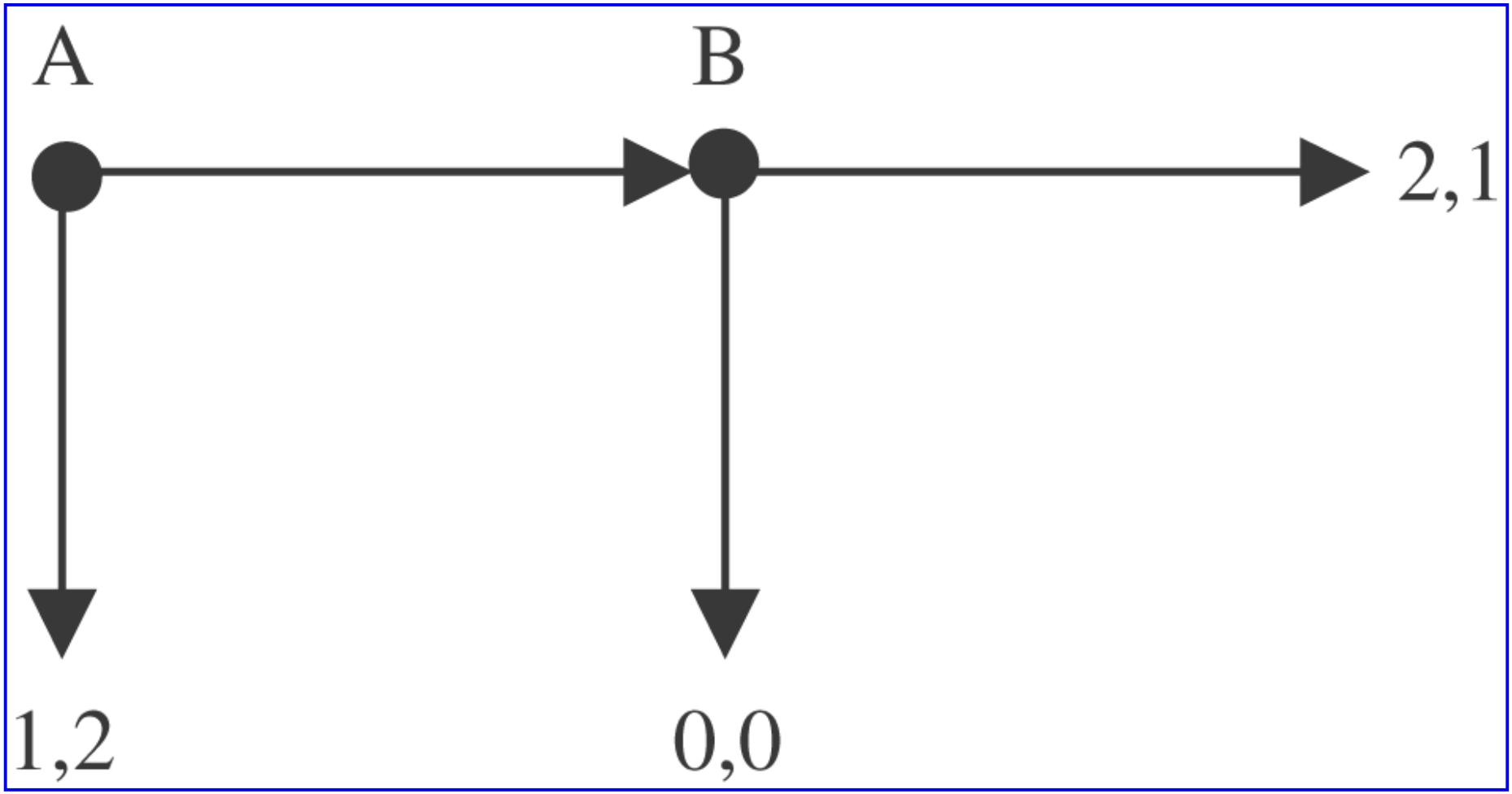

The informal criterion of ‘forward induction’ has several formulations. Kohlberg and Mertens (1986) require that a refined set of equilibria contains a subset that survives deletion of strategies that are not optimal responses at any equilibrium in the set. Van Damme (1989; 1991) requires that if player A rejects a choice X in favour of Y or Z then another player who knows only that Y or Z was chosen should consider Z unlikely if it is chosen only in equilibria that yield player A outcomes worse than choosing X, whereas Y is chosen in an equilibrium whose outcome is better. A typical application mimics backward induction but in reverse – if a person previously rejected a choice with an outcome that would have been superior to the outcomes from all but one equilibrium of the ensuing subgame, then presumably the person is anticipating that favourable equilibrium and intends to use his strategy in that equilibrium of the subgame. In Figure 2, if A rejects the payoff 5 from Down then B can infer that A intends to play Top in the ensuing subgame, yielding payoff 6 for both players.

Figure 2

First A and then B can avoid playing the subgame in which simultaneously each chooses between two options

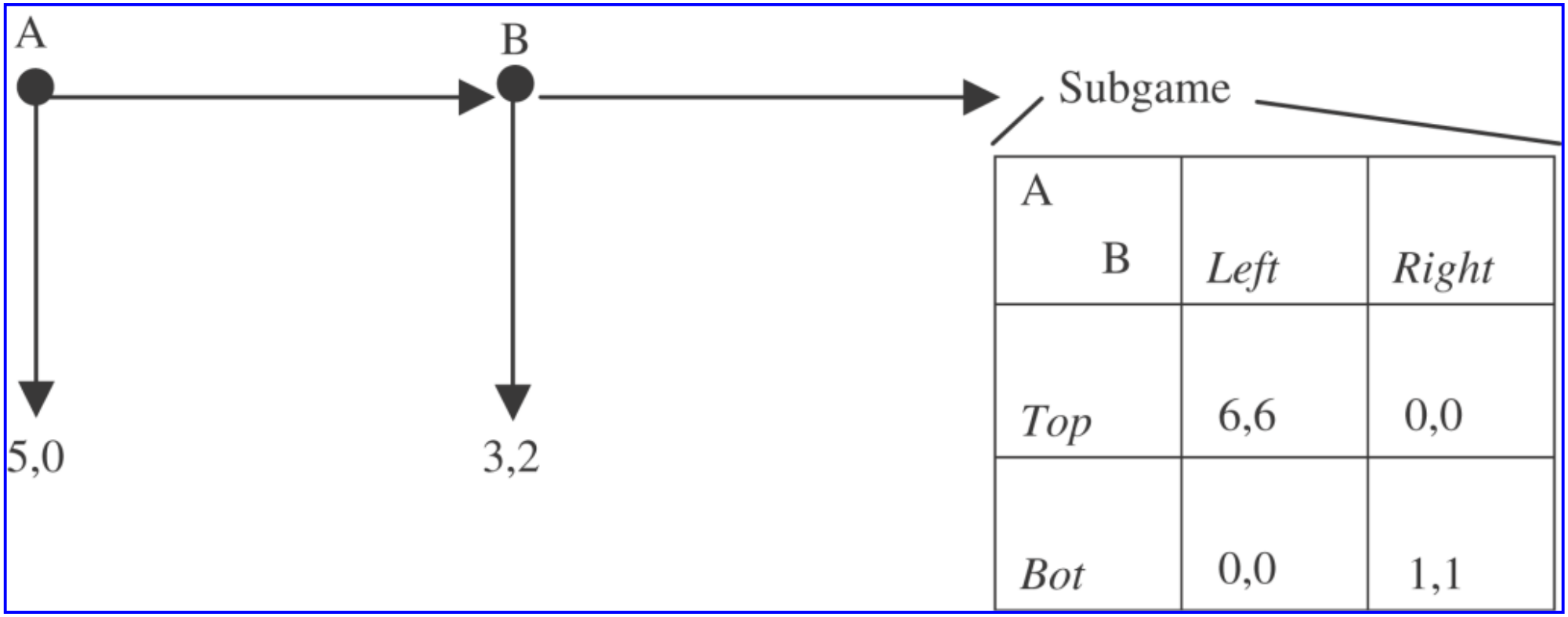

Dynamic games

Before proceeding further we describe briefly some relevant features of dynamic games, that is, games in which a player acts repeatedly, and can draw inferences about others' strategies, preferences, or private information as the game progresses. A dynamic game is said to have ‘perfect information’ if each person knows initially all the data of the game, and the prior history of his and others' actions whenever he acts, and they do not act simultaneously. In such a game each action initiates a subgame; hence backward induction yields a unique subgame-perfect equilibrium if there are no ties. But in many dynamic games there are no subgames. This is so whenever some person acts without knowing all data of the game relevant for the future. In Figure 3 player C acts without knowing whether player A or B chose down.

Figure 3

Player A moves down or across, in which case player B moves down or across. Player C does not observe whether it was A or B who moved down when she chooses to move left or right

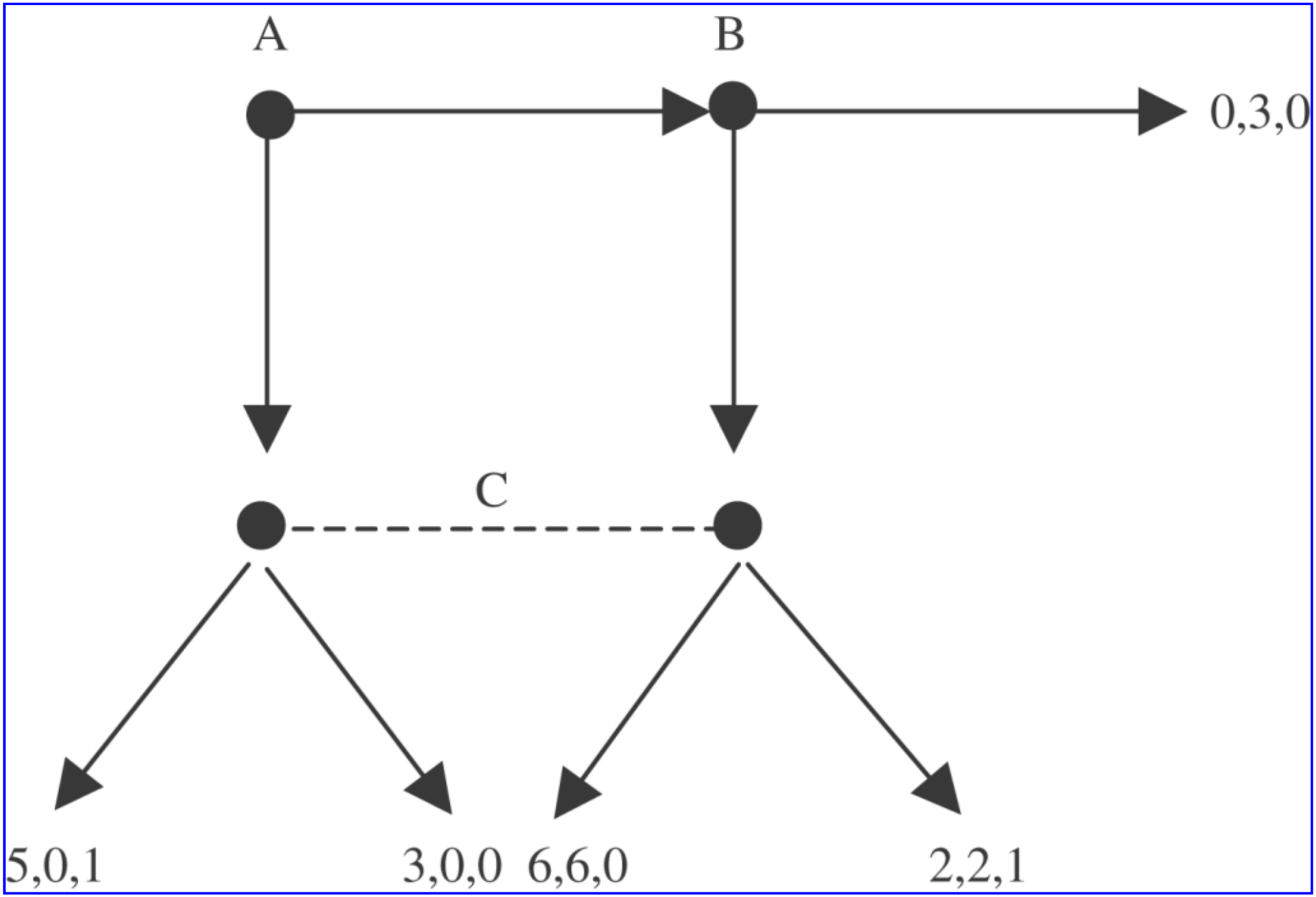

The source of this deficiency is typically that some participant has private information – for example, about his own preferences or about outcomes – or because his actions are observed imperfectly by some others. Among parlour games, chess is a game with perfect information (if players remember whether each king has been castled). Bridge and poker are games with imperfect information because the cards in one player's hand are not known to others when they bet. In practical settings, auctions and negotiations resemble poker because each party acts (bids, offers, and so on) without knowing others' valuations of the transaction. Analyses of practical economic games usually assume (as we do here) ‘perfect recall’ in the sense that each player always remembers what he knew and did previously. If bridge is treated as a two-player game between teams, then it has imperfect recall because

each team alternately remembers and forgets the cards in one member's hand as the bidding goes round the table, but bridge has perfect recall if it is treated as a four-player game. In card games like bridge and poker each player can derive the probability distribution of others' cards from the assumption that the deck of cards was thoroughly shuffled. Models of economic games impose analogous assumptions; for example, a model of an auction assumes that each bidder initially assesses a probability distribution of others' valuations of the item for sale, and then updates this assessment as he observes their bids. More realism is obtained from more complicated scenarios; for example, it could be that player A is uncertain about player B's assessment of player A's valuation. In principle the model could allow a hierarchy of beliefs – A's probability assessment of B's assessment of A's assessment of … . To adopt a proposal by John Harsanyi (1967–1968) developed by Mertens and Zamir (1985), such situations are modelled by assuming that each player is one of several types. The initial joint distribution of types is commonly known among the players, but each player knows his own type, which includes a specification of his available strategies, his preferences over outcomes, and, most importantly, his assessment of the conditional probabilities of others' types given his own type. In poker, for instance, a player's type includes the hand of cards he is dealt, and his hand affects his beliefs about others' hands.

Refinements of Nash equilibrium are especially useful in dynamic games. Nash equilibria do not distinguish between the case in which each player commits initially and irrevocably to his strategy throughout the game, and the case in which a player continually re-optimizes as the game progresses. The distinction is lost because the definition of Nash equilibrium presumes that players will surely adhere to their strategies chosen initially. Most refinements of Nash equilibrium are intended to resurrect this important distinction. Ideally one would like each Nash equilibrium to bear a label telling whether it assumes implicit commitment or relies on incredible threats or promises. Such features are usually evident in the equilibria of trivially simple games, but in more complicated games they must be identified by augmenting the definition of Nash equilibrium with additional criteria.

In the sequel we describe two classes of refinements in detail, but first we summarize their main features, identify the main selection criteria they use, and mention the names of some specific refinements. Both classes are generalizations of backward induction and subgame perfection, and they obtain similar results, but their motivation and implementation differ.

1. The criterion of sequential rationality

The presumption that commitment is irrevocable is flawed if other participants in the game do not view commitment to a strategy as credible. Commitment can be advantageous, of course, but if commitment is possible (for example, via enforceable contractual arrangements) then it should properly be treated as a distinct strategy. Absent commitment, some Nash equilibria are suspect because they rely implicitly on promises or threats that are not credible. For example, one Nash equilibrium might enable an incumbent firm to deter another firm from entering its market by threatening a price war. If such a threat succeeds in deterring entry then it is costless to the incumbent because it is never challenged; indeed, it can be that this equilibrium is sustained only by the presumption that the incumbent will never need to carry out the threat. But this threat is not credible if, after entry occurs, the incumbent would recognize that accommodation is more profitable than a price war. In such contexts, the purpose of a refinement is to select an alternative Nash equilibrium that anticipates correctly that entry will be followed by accommodation. For instance, the subgame-perfect equilibrium in Figure 1 satisfies this criterion.

Refinements in the first class exclude strategies that are not credible by requiring explicitly that a strategy is optimal in each contingency, even if it comes as a surprise. (We use the term ‘contingency’ rather than the technical term ‘information set’ used in game theory – it refers to any situation in which the player chooses an action.) These generally require that a player's strategy is optimal initially (as in the case of commitment), *and* that in each subsequent contingency in which the player might act his strategy remains optimal for the remainder of the game, even if the equilibrium predicts that the contingency should not occur. This criterion is called ‘sequential rationality’. As described later, three such refinements are *perfect Bayes*, *sequential*, and *lexicographic* equilibria, each of which can be strengthened further by imposing additional criteria such as *invariance*, the *intuitive criterion* and *divinity*.

2. The criterion of perfection or stability

The presumption that commitment is irrevocable is also flawed if there is some chance of deviations. If a player might ‘tremble’ or err in carrying out his intended strategy, or his valuation of outcomes might be slightly different from others anticipated, then other players can be surprised to find themselves in unexpected situations. Refinements that exploit this feature are implemented in two stages. In the first stage one identifies the Nash equilibria of a perturbation of the original game, usually obtained by restricting each player to randomized strategies that assign positive probabilities to all his original pure strategies. In the second stage one identifies those equilibria of the original game that are limits of equilibria of the perturbed game as this restriction is relaxed to allow inferior strategies to have zero probabilities.

Refinements in the second class also exclude strategies that are not credible, but refinements in this class implement sequential rationality indirectly. The general criterion that is invoked is called ‘perfection’ or ‘stability’, depending on the context. In each case a refinement is obtained from analyses of perturbed games. This second class of refinements is typically more restrictive than the first class due to the stronger effects of perturbations. As described later, two such refinements are *perfect* and *proper* equilibria. These are equilibria that are perturbed slightly by *some* perturbation of the players' strategies. A more stringent refinement selects a subset of equilibria that is *truly perfect* or *stable* in the sense that it is perturbed only slightly by *every* perturbation of players' strategies. This refinement selects a subset of equilibria rather than a single equilibrium because there need not exist a single equilibrium that is *essential* in that it is perturbed slightly by every perturbation of strategies. A stringent refinement selects a subset that is *hyperstable* in that it is stable

against perturbations of both players' strategies and their valuations of outcomes, or against perturbations of their optimal responses; and further, it is *invariant* in that it is unaffected by addition or deletion of redundant strategies.

The crucial role of perturbations in the second class of refinements makes them more difficult for non-specialists to understand and appreciate, but they have a prominent role in game theory because of their desirable properties. For example, in a two-player game a perfect equilibrium is equivalent to an equilibrium that uses only admissible strategies. In general, refinements in the second class have the advantage that they satisfy several selection criteria simultaneously.

After this overview, we now turn to detailed descriptions of the various refinements.

Refinements that require sequential rationality

In dynamic games with perfect information, the implementation of backward induction is unambiguous because in each contingency the player taking an action there knows exactly the subgame that follows. In chess, for example, the current positions of the pieces determine how the game can evolve subsequently. Moreover, if he anticipates his opponent's strategy then he can predict how the opponent will respond to each possible continuation of his own strategy. Using this prediction he can choose an optimal strategy for the remainder of the game by applying the *principle of optimality* – his optimal strategy in the current subgame consists of his initial action that, when followed by his optimal strategies in subsequent subgames, yields his best outcome. Thus, in principle (although not in practice, since chess is too complicated) his optimal strategy can be found by working backward from final positions through all possible positions in the game.

In contrast, in a game with imperfect information a player's current information may be insufficient to identify the prior history that led to this situation, and therefore insufficient to identify how others will respond in the future, even if he anticipates their strategies. In poker, for example, knowledge of his own cards and anticipation of others' strategies are insufficient to predict how they will respond to his bets. Their strategies specify how they will respond conditional on their cards but, since he does not know their cards, he remains uncertain what bets they will make in response to his bets. In this case, it is his assessment of the probability distribution of their cards that enables construction of his optimal strategy. That is, this probability distribution can be combined with their strategies to provide him with a probabilistic prediction of how they will bet in response to each bet he might make. Using this prediction he can again apply the principle of optimality to construct an optimal strategy by working backward from the various possible conclusions of the game.

Those refinements that select equilibria satisfying sequential rationality use an analogous procedure. The analogue of the probability distribution of others' cards is a system of ‘beliefs’, one for each contingency in which the player might find himself. Each belief is a conditional probability distribution on the prior history of the game given the contingency at which he has arrived. Thus, to whatever extent he is currently uncertain about others' preferences over final outcomes or their prior actions, his current belief provides him with a probability distribution over the various possibilities. As in poker, this probability distribution can be combined with his anticipation of their strategies to provide him with a probabilistic prediction of how they will act in response to each action he might take – and again, using this prediction he can apply the principle of optimality to construct an optimal strategy by working backward from the various possible conclusions of the game.

There is an important proviso, however. These refinements require that, whenever one contingency follows another with positive probability, the belief at the later one must be obtained from the belief at the earlier one by Bayes' rule. This ensures consistency with the rules of conditional probability. But, importantly, it does not restrict a player's belief at a contingency that was unexpected, that is, had zero probability according to his previous belief and the other players' strategies.

In Figure 3, in one Nash equilibrium A chooses down, B chooses across, and C chooses left. This is evidently not sequential because if A were to deviate then B could gain by choosing down. In a sequential equilibrium B chooses down and each of A and C randomizes equally between his two strategies. The strategies of A and B imply that C places equal probabilities on which of A and B chose down.

The weakest refinement selects a *perfect-Bayes* equilibrium (Fudenberg and Tirole, 1991). This requires that each player's strategy is consistent with some system of beliefs such that (*a*) his strategy is optimal given his beliefs and others' strategies, and (*b*) his beliefs satisfy Bayes' rule (wherever it applies) given others' strategies. A stronger refinement selects *sequential* equilibria (Kreps and Wilson, 1982). A sequential equilibrium requires that each player's system of beliefs is consistent with the structure of the game. Consistency is defined formally as the requirement that each player's system of beliefs is the limit of the conditional probabilities induced by players' strategies in some perturbed game, as described previously. A further refinement selects *quasi-perfect* equilibria (van Damme, 1984), which requires admissibility of a player's strategy in continuation from each contingency, excluding any chance that he himself might deviate from his intended strategy. And even stronger are *proper* equilibria (Myerson, 1978), described later. This sequence of progressively stronger refinements is typical. Because proper implies quasi-perfect implies sequential implies perfect-Bayes, one might think that it is sufficient to always use properness as the refinement. However, the prevailing practice in the social sciences is to invoke the weakest refinement that suffices for the game being studied. This reflects a conservative attitude about using unnecessarily restrictive refinements. If, say, there is a unique sequential equilibrium that uses only admissible strategies, then one refrains from imposing stronger criteria.

Additional criteria can be invoked to select among sequential equilibria. In Figure 4 there is a sequential equilibrium in which both types of A move left and B randomizes equally between middle and bottom, and another in which both types of A move right and B chooses middle. An alternative justification for the second, due to Hillas (1998), is shown in Figure 5, where the game is restructured so that A either commits initially to left or they play the subgame with simultaneous choices of strategies. The criterion of subgame

perfection selects the second equilibrium in Figure 4 because in Figure 5 the subgame has a unique equilibrium with payoff 6 for A that is superior to his payoff 4 from committing to left.

Figure 4

Nature chooses whether player A's type is A1 or A2 with equal probabilities. Then A chooses Left or Right, in which case player B, without knowing A's type, chooses one of three options

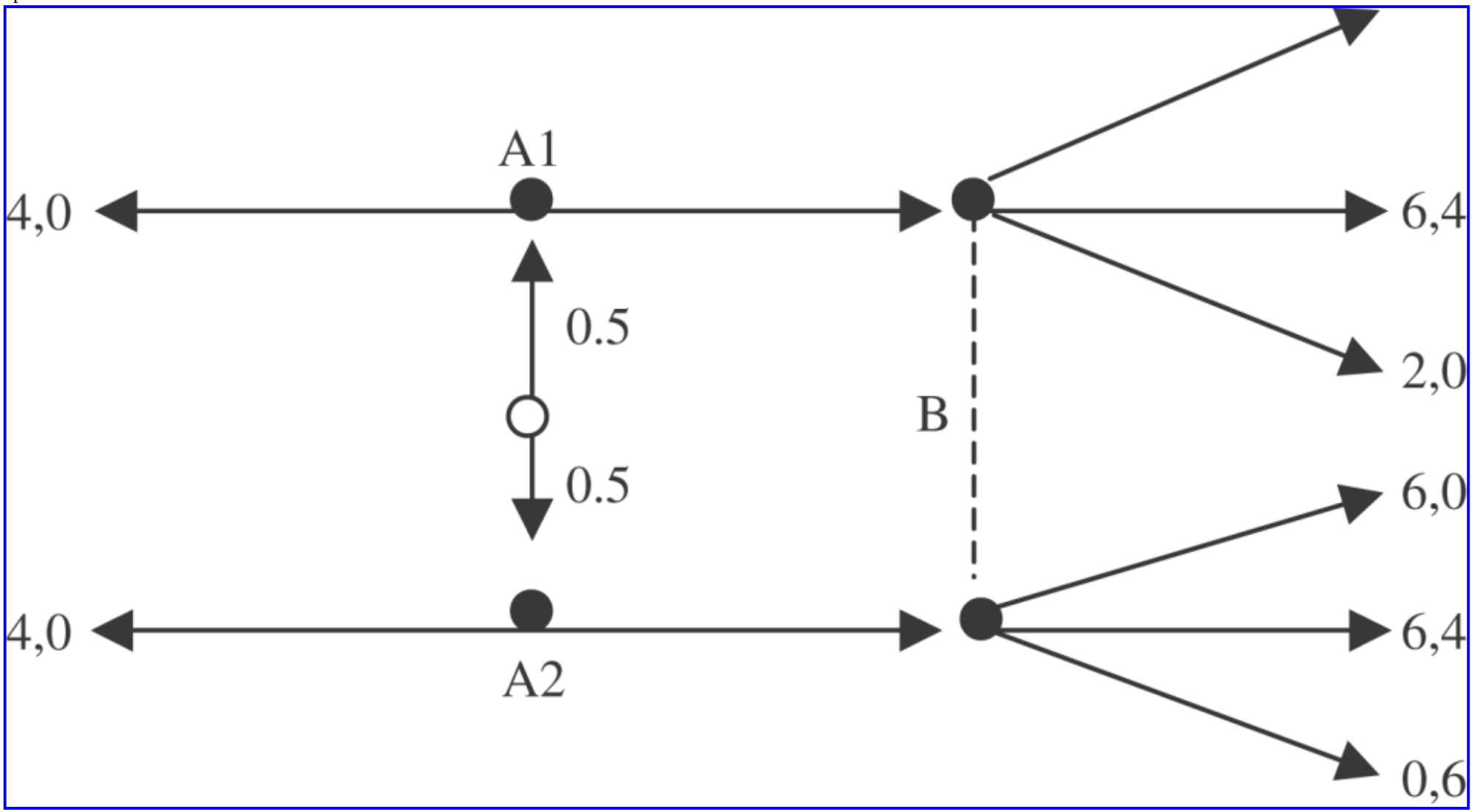

Figure 5

The game in Figure 4 restructured so that either A commits to Left regardless of his type, or plays a subgame with simultaneous moves in which he chooses one of his other three type-contingent strategies. The payoffs 6,4 to A and B from the unique Nash equilibrium of the subgame are shown with an asterisk

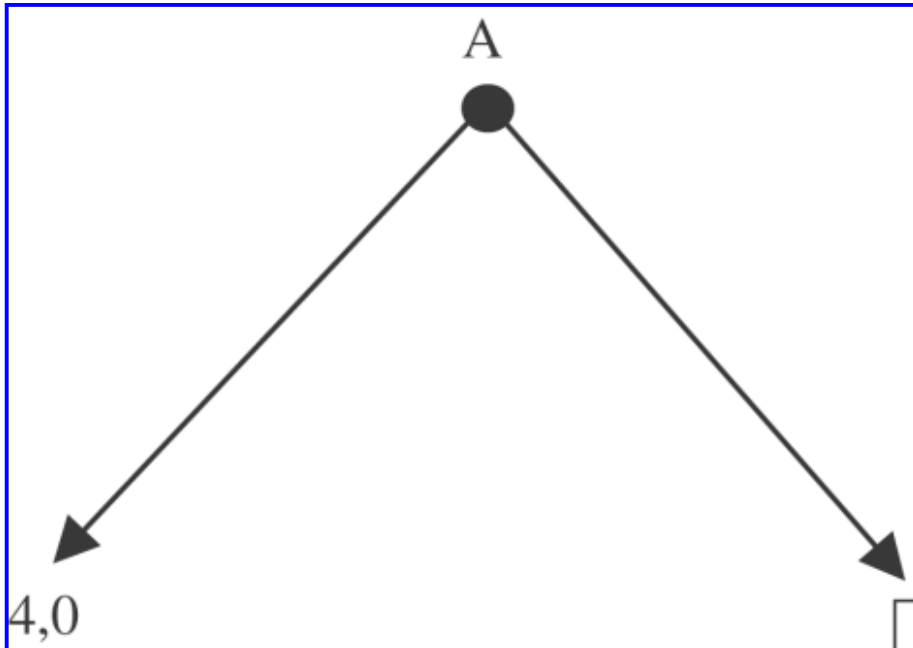

A / B	*Top*	*Middle*	*Bottom*
1.*Left* 2.*Right*	5,0	5,2	2,3
1.*Right* 2.*Left*	3,2	5,2	3,0
1.*Right* 2.*Right*	4,3	**6,4** *	1,3

These refinements can be supplemented with additional criteria that restrict a player's beliefs in unexpected contingencies. The most widely used criteria apply to contexts in which one player B could interpret the action of another player A as revealing private information; that is, A's action might signal something about A's type. These criteria restrict B's belief (after B observes A deviating from the equilibrium) to one that assigns positive probability only to A's types that might possibly gain from the deviation, provided it were interpreted by B as a credible signal about A's type. The purpose of these criteria is to exclude beliefs that are blind to A's attempts to signal what his type is when it would be to A's advantage for B to recognize the signal. In effect, these criteria reject equilibria that commit a player to unrealistic beliefs. Another interpretation is that these criteria reject equilibria in which A

is ‘threatened by B's beliefs’ because B stubbornly retains these beliefs in spite of plausible evidence to the contrary.

The simplest version requires that B's belief assigns zero probability to those types of A that cannot possibly gain by deviating, regardless of how B responds. The *intuitive* criterion (Cho and Kreps, 1987) requires that there cannot be some type of A that surely gains from deviating in every continuation for which B responds with a strategy that is optimal based on a belief that assigns zero probability to those types of A that cannot gain from the deviation. That is, an equilibrium fails the intuitive criterion if B's belief fails to recognize that A's deviation is a credible signal about his type. They apply this criterion to the game in Figure 6, which has two sequential equilibria. In one both types of A choose left and B chooses down or up contingent on left or right. In another both types choose right and B chooses up or down contingent on left or right. In both equilibria B's belief in the unexpected event (right or left respectively) assigns probability greater than 0.5 to A's type A1. The intuitive criterion rejects the second equilibrium because if A2 were to deviate by choosing left, and then B recognizes that this deviation credibly signals A's type A2 (because type A1 cannot gain by deviating regardless of B's response) and therefore B chooses down, then type A2 obtains payoff 3 rather than his equilibrium payoff 2.

Figure 6

A signalling game in which Nature chooses A's type A1 or A2, then A chooses left or right, and then B, without knowing A's type, chooses up or down

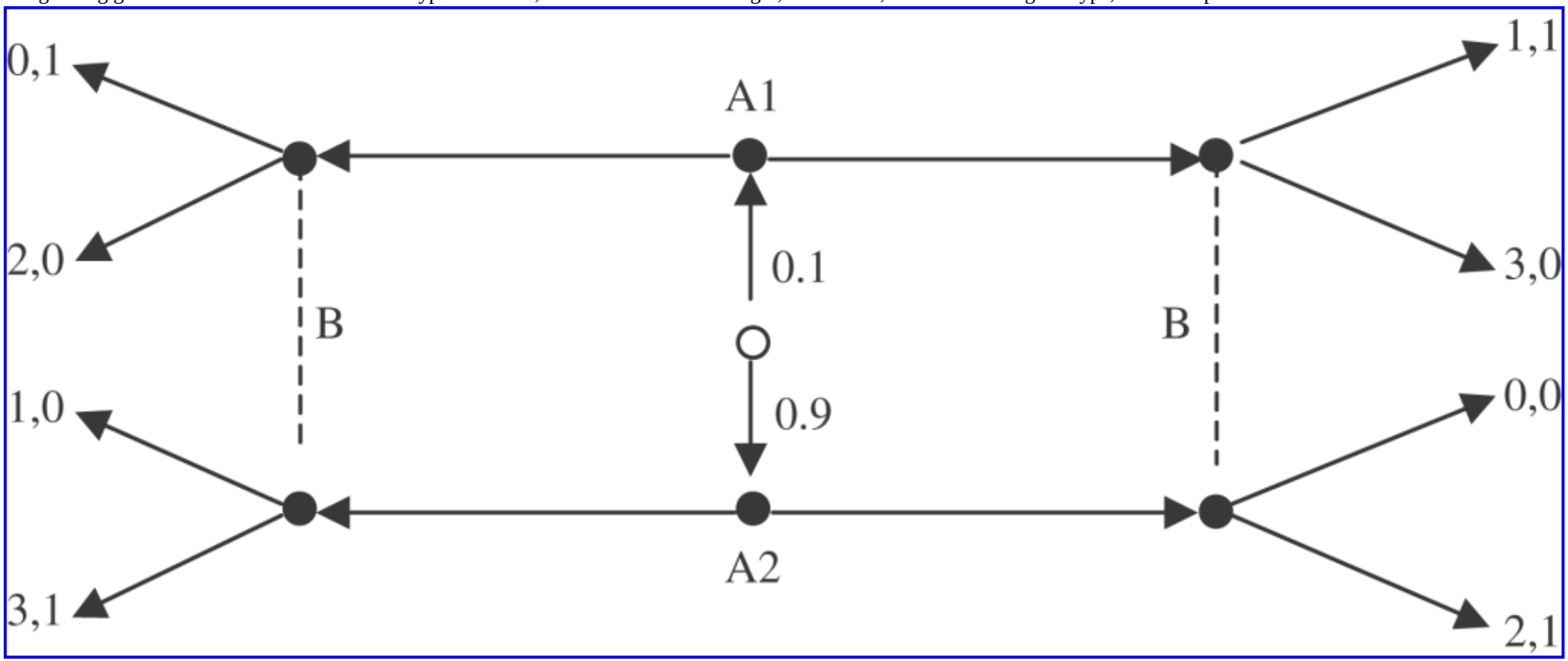

Cho and Kreps also define an alternative version, called the ‘equilibrium domination’ criterion. This criterion requires that, for each continuation in which B responds with a strategy that is optimal based on a belief that assigns zero probability to those types of A that cannot gain from deviating, there cannot be some type of A that gains from deviating. More restrictive is the criterion **D1** (Banks and Sobel, 1987), also called ‘divinity’ when it is applied iteratively, which requires that, if the set of B's responses for which one type of A gains from deviating is larger than the set for which a second type gains, then B's beliefs must assign zero probability to the second type. The criterion **D2** is similar except that some (rather than just one) types of A gain. All these criteria are weaker than the *never weak best reply* criterion that requires an equilibrium to survive deletion of a player's strategy that is not an optimal reply to any equilibrium with the same outcome. In Figure 6 this criterion is applied by observing that the second equilibrium does not survive deletion of those strategies of A in which type A2 chooses left.

A *lexicographic* equilibrium (Blume, Brandenburger and Dekel, 1991a; 1991b) uses a different construction. Each player is supposed to rely on a sequence of ‘theories’ about others' strategies. He starts the game by assuming that his first theory of others' strategies is true, and uses his optimal strategy according to that theory. He continues doing so until he finds himself in a situation that cannot be explained by his first theory. In this case, he abandons the first theory and assumes instead that the second theory is true – or if it too cannot

explain what has happened then he proceeds to the next theory in the sequence. This provides a refinement of Nash equilibrium because each player anticipates that deviation from his optimal strategy for any theory will provoke others to abandon their current theories and strategies and thus respond with their optimal strategies for their next theories consistent with his deviant action. Lexicographic equilibria can be used to represent nearly any refinement. The hierarchy of a player's theories serves basically the same role as his system of beliefs, but the focus is on predictions of other players' strategies in the future rather than probabilities of what they know or have done in the past. The lexicographic specification has the same effect as considering small perturbations of strategies; for example, the sequence of strategies approximating a perfect or proper equilibrium can be used to construct the hierarchy of theories.

Refinements derived from perturbed games

The other major class of refinements relies on perturbations to select among the Nash equilibria. The motive for this approach stems from a basic principle of decision theory –the *equivalence* of alternative methods of deriving optimal strategies. This principle posits that constructing a player's optimal strategy in a dynamic game by invoking auxiliary systems of beliefs and the iterative application of the principle of optimality (as in perfect-Bayes and sequential equilibria) is a useful computational procedure, but the same result should be obtainable from an initial choice of a strategy, that is, an optimal plan for the entire game of actions taken in each contingency. Indeed, the definition of Nash equilibrium embodies this principle. Proponents therefore argue that whatever improvements come from dynamic analysis can and should be replicated by static analysis of initial choices among strategies, supplemented by additional criteria. (We use the terms ‘static’ and ‘dynamic’ analysis rather than the technical terms ‘normal-form’ and ‘extensive-form’ analysis used in game theory.) The validity of this argument is evident in the case of subgame-perfect equilibria of games with perfect information, which can be derived either from the principle of optimality using backward induction, or by iterative elimination of weakly dominated strategies in a prescribed order. The argument is reinforced by major deficiencies of dynamic analysis; for example, we mentioned above that a sequential equilibrium can use inadmissible strategies. Another deficiency is failure to satisfy the criterion of *invariance*, namely, the set of sequential equilibria can depend on which of many equivalent descriptions of the dynamics of the game is used (in particular, on the addition or deletion of redundant strategies).

On this view one should address directly the basic motive for refinement, which is to exclude equilibria that assume implicitly that each player commits initially to his strategy – since Nash equilibria do not distinguish between cases with and without commitment. Thus one considers explicitly that during the game any player might deviate from his equilibrium strategy for some exogenous reason that was not represented in the initial description of the game. Recognition of the possibility of deviations, however improbable they might be, then ensures that a player's strategy includes a specification of his optimal response to others' deviations from the equilibrium. The objective is therefore to characterize those equilibria that are affected only slightly by small probabilities of deviant behaviours or variations in preferences. This programme is implemented by considering perturbations of the game. These can be perturbations of strategies or payoffs, but actually the net effect of a perturbation of others' strategies is to perturb a player's payoffs.

In the following we focus on the perturbations of the static (that is, the normal form) of the game but similar perturbations can also be applied to the dynamic version (that is, the extensive form) by applying them to each contingency separately. This is done by invoking the principle that a dynamic game can also be analysed in a static framework by treating the player acting in each contingency as a new player (interpreted as the player's agent who acts solely in that contingency) in the ‘agent-normal-form’ of the game, where the new player's payoffs agree with those of the original player.

The construction of a *perfect* equilibrium (Selten, 1975) illustrates the basic method, which uses two steps.

1. 1. For each small positive number ε one finds an ε -*perfect* equilibrium, defined by the requirement that each player's strategy has the following property: every one of his pure strategies is used with positive probability, but any pure strategy that is an inferior response to the others' strategies has probability no more than ε . Thus an ε -perfect equilibrium supposes that every strategy, and therefore every action during the game, might occur, even if it is suboptimal.
2. 2. One then obtains a perfect equilibrium as the limit of a convergent subsequence of ε -perfect equilibria.

One method of constructing an ε -perfect equilibrium starts by specifying for each player *i* a small probability $\delta_i < \varepsilon$ and a randomized strategy σ_i that uses every pure strategy with positive probability – that is, the strategy combination σ is ‘completely mixed’. One then finds an ordinary Nash equilibrium of the perturbed game in which each player's payoffs are as follows: his payoff from each combination of all players' pure strategies is replaced by his expected payoff when each player *i*'s pure strategy is implemented only with probability $1-\delta_i$ and with probability δ_i that player uses his randomized strategy σ_i instead. In this context one says that the game is perturbed by less than ε toward σ – we use this phrase again later when we describe stable sets of equilibria. An equilibrium of this perturbed game induces an ε -perfect equilibrium of the original game.

An alternative definition of perfect equilibrium requires that each player's strategy is an optimal response to a convergent sequence of others' strategies for which all their pure strategies have positive probability – this reveals explicitly that optimality against small probabilities of deviations is achieved, and that a perfect equilibrium uses only admissible strategies. In fact, a perfect equilibrium of the agent-normal-form induces a sequential equilibrium of the dynamic version of the game. Moreover, if the payoffs of the dynamic game

are generic (that is, not related to each other by polynomial equations) then every sequential equilibrium is also perfect.

A stronger refinement selects *proper* equilibria (Myerson, 1978). This refinement supposes that the more inferior the expected payoff from a strategy is, the less likely it is to be used. The construction differs only in step 1: if one pure strategy S is inferior to another T in response to the others' strategies then S has probability no more than ε times the probability of T. A proper equilibrium induces a sequential equilibrium in every one of the equivalent descriptions of the dynamic game.

A perfect or proper equilibrium depends on the particular perturbation used to construct an ε-perfect or ε-proper equilibrium. Sometimes a game has an equilibrium that is *essential* or *truly perfect* in that any σ can be used when perturbing the game by less than ε toward σ, as above. This is usual for a static game with generic payoffs because in this case its equilibria are isolated and vary continuously with perturbations. However, such equilibria rarely exist in the important case that the static game represents a dynamic game, since in this case some strategies have the same equilibrium payoffs. This occurs because there is usually considerable freedom about how a player acts in contingencies off the predicted path of the equilibrium; in effect, the same outcome results whether the player ‘punishes’ others only barely enough to deter deviations, or more than enough. Indeed, for a dynamic game with generic payoffs, all the equilibria in a connected set yield the same equilibrium outcome because they differ only off the predicted path of equilibrium play. One must therefore consider sets of equilibria when invoking stringent refinements like truly perfect. One applies a somewhat different test to sets of equilibria. When considering a set of equilibria one requires that every sufficiently small perturbation (within a specified class) of the game has an equilibrium near some equilibrium in the set. Some refinements insist on a minimal closed set of equilibria with this property, but here we ignore minimality.

The chief refinement of this kind uses strategy perturbations to generate perturbed games. Kohlberg and Mertens (1986) say that a set of equilibria is *stable* if for each neighbourhood of the set there exists a positive probability ε such that, for every completely mixed strategy combination σ, each perturbation of the game by less than ε toward σ has an equilibrium within the neighbourhood. Stability can be interpreted as truly perfect applied to sets of equilibria and using the class of payoff perturbations generated by strategy perturbations. Besides the fact that a stable set always exists, it satisfies several criteria: it uses only *admissible* strategies, it contains a stable set of the reduced game after deleting a strategy that is weakly dominated or an inferior response to all equilibria in the set (these assure *iterative elimination of weakly dominated strategies* and a version of *forward induction*), and it is *invariant* to addition or deletion of redundant strategies. However, examples are known in which a stable set of a static game does not include a sequential equilibrium of the dynamic game it represents. This failure to satisfy the backward induction criterion can be remedied in various ways that we describe next.

One approach considers the larger class of all payoff perturbations. In this case, invariance to redundant strategies is not assured so it is imposed explicitly. For this, say that two games are equivalent if deletion of all redundant strategies results in the same reduced game. Similarly, randomized strategies in these two games are equivalent if they yield the same randomization over pure strategies of the reduced game. Informally, a set of equilibria is hyperstable if, for every payoff perturbation of every equivalent game, there is an equilibrium equivalent to one near the set. Two formal versions are the following. Kohlberg and Mertens (1986) say that a set S of equilibria is *hyperstable* if, for each neighbourhood N of those strategies in an equivalent game that are equivalent to ones in S, there is a sufficiently small neighbourhood P of payoff perturbations for the equivalent game such that every game in P has an equilibrium in N. A somewhat stronger version is the following. A set S of equilibria of a game G is *uniformly hyperstable* if, for each neighbourhood N of S, there is a $\delta>0$ such that every game in the δ-neighbourhood of any game equivalent to G has an equilibrium equivalent to one in N. This version emphasizes that uniform hyperstability is closely akin to a kind of continuity with respect to payoff perturbations of equivalent games. Unfortunately, both of these definitions are complex, but the second actually allows a succinct statement in the case that the set S is a ‘component’ of equilibria, namely, a maximal connected set of the Nash equilibria. In this case the component is uniformly hyperstable if and only if its topological index is non-zero, and thus *essential* in the sense used in algebraic topology to characterize a set of fixed points of a function that is slightly affected by every perturbation of the function. This provides a simply computed test of whether a component is uniformly hyperstable.

Hyperstable sets tend to be larger than stable sets of equilibria because they must be robust against a larger class of perturbations, but for this same reason the criterion is actually stronger. Within a hyperstable component there is always a stable set satisfying the criteria listed previously. There is also a proper equilibrium that induces a sequential equilibrium in every dynamic game with the same static representation – thus, the criterion of *backward induction* is also satisfied. Selecting a stable subset or a proper equilibrium inside a hyperstable component may be necessary because there can be other equilibria within a hyperstable component that use inadmissible strategies. Nevertheless, for a dynamic game with generic payoffs, all the equilibria within a single component yield the same outcome, since they differ only off the path of equilibrium play, so for the purpose of predicting the outcome rather than players' strategies it is immaterial which equilibrium is considered. However, examples are known in which an inessential hyperstable component contains two stable sets with opposite indices with respect to perturbations of strategies.

The most restrictive refinement is the revised definition of stability proposed by Mertens (1989). Although this definition is highly technical, it can be summarized briefly as follows for the mathematically expert reader. Roughly, a closed set of equilibria is (Mertens-) *stable* if the projection map (from its neighbourhood in the graph of the Nash equilibria into the space of games with perturbed strategies) is essential. Such a set satisfies all the criteria listed previously, and several more. For instance, it satisfies the *small-worlds* criterion (Mertens, 1992), which requires that adding other players whose strategies have no effect on the payoffs for the original players has no effect on the selected strategies of the original players. The persistent mystery in the study of refinements is why such sophisticated constructions seem to be necessary if a single definition is to satisfy all the criteria simultaneously. The clue seems to be that, because Nash equilibria are the solutions of a fixed-point problem, a fully adequate refinement must ensure that fixed points exist for every perturbation of this problem.

The state of the art of refinements

The development of increasingly stronger refinements by imposing ad hoc criteria incrementally was a preliminary to more systematic development. Eventually, one wants to identify decision-theoretic criteria that suffice as axioms to characterize refinements. The two groups of refinements described above approach this problem differently. Those that consider perturbations seek to verify whether there exist refinements that satisfy many or (in the case of Mertens-stability) most criteria. From its beginning in the work of Selten (1975), Myerson (1978), and Kohlberg and Mertens (1986), this has been a productive exercise, showing that refinements can enforce more stringent criteria than Nash (1950; 1951) requires. However, the results obtained depend ultimately on the class of perturbations considered, since Fudenberg, Kreps and Levine (1988) show that each Nash equilibrium of a game is the limit of strict equilibria of perturbed games in a very general class. Perturbations are mathematical artefacts used to identify refinements with desirable properties, but they are not intrinsic to a fundamental theory of rational decision making in multi-person situations. Those in the other group directly impose decision-theoretic criteria – admissibility, iterative elimination of dominated or inferior strategies, backward induction, invariance, small worlds, and so on. Their ultimate aim is to characterize refinements axiomatically. But so far none has obtained an ideal refinement of the Nash equilibria.

See Also

- behavioural game theory
- epistemic game theory: incomplete information
- game theory
- Harsanyi, John C.
- Markov equilibria in macroeconomics
- Nash, John Forbes
- Nash program
- Selten, Reinhard
- signalling and screening.

Bibliography

Banks, J. and Sobel, J. 1987. Equilibrium selection in signaling games. *Econometrica* 55, 647–61.

Blume, L., Brandenburger, A. and Dekel, E. 1991a. Lexicographic probabilities and choice under uncertainty. *Econometrica* 59, 61–79.

Blume, L., Brandenburger, A. and Dekel, E. 1991b. Lexicographic probabilities and equilibrium refinements. *Econometrica* 59, 81–98.

Cho, I. and Kreps, D. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102, 179–221.

Fudenberg, D., Kreps, D. and Levine, D. 1988. On the robustness of equilibrium refinements. *Journal of Economic Theory* 44, 351–80.

Fudenberg, D. and Tirole, J. 1991. Perfect Bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory* 53, 236–60.

Harsanyi, J. 1967–1968. Games with incomplete information played by ‘Bayesian’ players, I–III. *Management Science* 14, 159–82, 320–34, 486–502.

Harsanyi, J. and Selten, R. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Hillas, J. 1998. How much of ‘forward induction’ is implied by ‘backward induction’ and ‘ordinality’? Mimeo. Department of Economics, University of Auckland.

Hillas, J. and Kohlberg, E. 2002. The foundations of strategic equilibrium. In *Handbook of Game Theory*, vol. 3, ed. R. Aumann and S. Hart. Amsterdam: North-Holland/Elsevier Science Publishers.

Kohlberg, E. 1990. Refinement of Nash equilibrium: the main ideas. In *Game Theory and Applications*, ed. T. Ichiishi, A. Neyman and Y. Tauman. San Diego: Academic Press.

Kohlberg, E. and Mertens, J.-F. 1986. On the strategic stability of equilibria. *Econometrica* 54, 1003–38.

Kreps, D. 1990. *Game Theory and Economic Modeling*. New York: Oxford University Press.

Kreps, D. and Wilson, R. 1982. Sequential equilibria. *Econometrica* 50, 863–94.

Mertens, J.-F. 1989. Stable equilibria – a reformulation, Part I: definition and basic properties. *Mathematics of Operations Research* 14, 575–624.

Mertens, J.-F. 1992. The small worlds axiom for stable equilibria. *Games and Economic Behavior* 4, 553–64.

Mertens, J.-F. and Zamir, S. 1985. Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory* 14, 1–29.

Myerson, R. 1978. Refinement of the Nash equilibrium concept. *International Journal of Game Theory* 7, 73–80.

Nash, J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences USA* 36, 48–9.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. *Zeitschrift fur die gesamte Staatswissenschaft* 121, 301–24, 667–89.

Selten, R. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4, 25–55.

van Damme, E. 1984. A relation between perfect equilibria in extensive form games and proper equilibria in normal form games. *International Journal of Game Theory* 13, 1–13.

van Damme, E. 1989. Stable equilibria and forward induction. *Journal of Economic Theory* 48, 476–96.

van Damme, E. 1991. *Stability and Perfection of Nash Equilibria*. Berlin: Springer-Verlag.

How to cite this article

Govindan, Srihari and Robert B. Wilson. "Nash equilibrium, refinements of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000242> doi:10.1057/9780230226203.1155

The New Palgrave Dictionary of Economics Online

Nash program

Roberto Serrano
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article is a brief survey on the Nash program for coalitional games. Results of non-cooperative implementation of the Nash solution, the Shapley value and the core are discussed.

Keywords

cooperative games; core; Edgeworth, F.; Nash program; Nash solution; non-cooperative games; Shapley value; subgame perfect equilibrium; Walrasian outcome

Article

In game theory, ‘Nash program’ is the name given to a research agenda, initiated in Nash (1953), intended to bridge the gap between the cooperative and non-cooperative approaches to the discipline. Many authors have contributed to the program since its beginnings (see Serrano, 2005, for a comprehensive survey). The current article concentrates on a few salient contributions. One should begin by introducing some preliminaries and providing definitions of some basic concepts.

Preliminaries

The non-cooperative approach to game theory provides a rich language and develops useful tools to analyse strategic situations. One clear advantage of the approach is that it is able to model how specific details of the interaction may affect the final outcome. One limitation, however, is that its predictions may be highly sensitive to those details. For this reason it is worth also analysing more abstract approaches that attempt to obtain conclusions that are independent of such details. The cooperative approach is one such attempt.
Here are the primitives of the basic model in cooperative game theory. Let $N = \{1, \ldots, n\}$ be a finite set of players. For each *S*, a non-empty subset of *N*, we shall specify a set *V*(*S*) containing $|S|$-dimensional

payoff vectors that are feasible for coalition *S*. Thus, a reduced form approach is taken because one does not explain what strategic choices are behind each of the payoff vectors in *V*(*S*). In addition, in this formulation, referred to as the characteristic function, it is implicitly assumed that the actions taken by the complement coalition (those players not in *S*) cannot prevent *S* from achieving each of the payoff vectors in *V*(*S*). There are more general models in which these sorts of externalities are considered, but for the most part the contributions to the Nash program have been confined to the characteristic function model. Given a collection of sets *V*(*S*), one for each *S*, the theory formulates its predictions on the basis of solution concepts.

A solution is a mapping that assigns a set of payoff vectors in *V*(*N*) to each characteristic function $(V(S))_{S \subseteq N}$. Thus, a solution in general prescribes a set, although it can be single-valued (when it assigns a unique payoff vector as a function of the fundamentals of the problem). The leading set-valued cooperative solution concept is the core, while the most used single-valued ones are the Nash bargaining solution and the Shapley value.

There are several criteria to evaluate the reasonableness or appeal of a cooperative solution. One could start by defending it on the basis of its definition alone. In the case of the core, this will be especially relevant: in a context in which players can freely get together in groups, the prediction should be payoff vectors that cannot be improved upon by any coalition. Alternatively, one can propose axioms, abstract principles, that one would like the solution to have, and the next step is to pursue their logical consequences. Historically, this was the first argument to justify the Nash solution and the Shapley value. However, some may think that the definition may be somewhat arbitrary, or one may object that the axiomatic approach is ‘too abstract’. By proposing non-cooperative games that specify the details of negotiation, the Nash program may help to counter these criticisms. First, the procedure will tell a story about how coalitions form and what sort of interaction among players is happening. In that process, because the tools of non-cooperative game theory are used for the analysis, the cooperative solution will be understood as the outcome of a series of strategic problems facing individual players. Second, novel connections and differences among solutions may now be uncovered from the distinct negotiation procedures that lead to each of them. Therefore, a result in the Nash program, referred to as a ‘non-cooperative foundation’ or ‘non-cooperative implementation’ of a cooperative solution, enhances its significance, being looked at now from a new perspective. Focusing on the features of the rules of negotiation that lead to different cooperative solutions takes one a long way in opening the ‘black box’ of how a coalition came about, and contributes to a deeper understanding of the circumstances under which one solution versus another may be more appropriate to use.

The Nash bargaining solution

A particular case of a characteristic function is a two-player bargaining problem. In it, $N = \{1, 2\}$ is the set of players. The set $V(\{1, 2\})$, a compact and convex subset of $\mathbb{R}$, is the set of feasible payoffs if the two players reach an agreement. Compactness may follow from the existence of a bounded physical pie that the parties are dividing, and convexity is a consequence of expected utility and the potential use of lotteries. The sets $(V(\{i\}))_{i \in N}$ are subsets of $\mathbb{R}$, and let $d_i = \max V(\{i\})$ be the disagreement payoff for player *i*, that is, the payoff that *i* will receive if the parties fail to reach an agreement. It is assumed that $V(\{1, 2\})$ contains payoff vectors that Pareto dominate the disagreement payoffs. A solution assigns a feasible payoff pair to each bargaining problem.

This is the framework introduced in Nash (1950), where he proposes four axioms that a solution to bargaining problems should have. First, expected utility implies that, if payoff functions are rescaled via positive affine transformations, so must be the solution (scale invariance). Second, the solution must prescribe a Pareto efficient payoff pair (efficiency). Third, if the set $V(\{1, 2\})$ is symmetric with respect to the 45 degree line and $d_1 = d_2$, the solution must lie on that line (symmetry). Fourth, the solution must be independent of 'irrelevant' alternatives, that is, it must pick the same point if it is still feasible after one eliminates other points from the feasible set (IIA). Because of scale invariance, there is no loss of generality in normalizing the disagreement payoff to 0. We call the resulting problem a normalized problem.

Nash (1950) shows that there exists a unique solution satisfying scale invariance, efficiency, symmetry and IIA, and it is the one that assigns to each normalized bargaining problem the point (u_1, u_2) that maximizes the product v_1v_2 over all $(v_1, v_2) \in V(\{1, 2\})$. Today we refer to this as the 'Nash solution'. The use of the Nash solution is pervasive in applications and, following the axioms in Nash (1950), it is usually viewed as a normatively appealing resolution to bargaining problems.

In the first paper of the Nash program, Nash (1953) provides a non-cooperative approach to his axiomatically derived solution. This is done by means of a simple demand game. The two players are asked to demand simultaneously a payoff: player 1 demands v_1 and player 2 demands v_2. If the pair (v_1, v_2) is feasible, so that $(v_1, v_2) \in V(\{1, 2\})$, the corresponding agreement and split of the pie takes place to implement these payoffs. Otherwise, there is disagreement and payoffs are 0. To fix ideas, let us think of the existence of a physical pie of size 1 that is created if agreement is reached, while no pie is produced otherwise. Thus, player *i*'s demand v_i corresponds to demanding a share x_i of the pie, $0 \leq x_i \leq 1$, such that player *i*'s utility or payoff from receiving x_i is v_i.

The Nash demand game admits a continuum of Nash equilibria. Indeed, every point on the Pareto frontier of $V(\{1, 2\})$ is a Nash equilibrium outcome, as is the disagreement payoff point if each player demands the payoff corresponding to having the entire pie. However, Nash (1953) introduces uncertainty concerning the exact size of the pie. Now players, when formulating their demands, must have to take into account the fact that with some probability the pair of demands may lead to disagreement, even if they add up to less than 1. Then, it can be shown that the optimal choice of demands at a Nash equilibrium of the demand game with uncertain pie converges to the Nash solution payoffs as uncertainty becomes negligible. Hence, the Nash solution arises as the rule that equates marginal gain (through the increase in one's demanded share) and marginal loss (via the increase in the probability of disagreement) for each player when the problem is subject to a small degree of noise and demands/commitments are made simultaneously.

Rubinstein (1982) proposes a different non-cooperative procedure. In it, time preferences – impatience – and credibility of threats are the main forces that drive the equilibrium. The game is a potentially infinite sequence of alternating offers. In period 0, player 1 begins by making the first proposal. If player 2 accepts it, the game ends; otherwise, one period elapses and the rejector will make a counter-proposal in period 1, and so on. Let $\delta \in [0, 1)$ be the common per period discount factor, and let $v_i(\cdot)$ be player *i*'s utility function over shares of the pie, assumed to be concave and strictly monotone. Thus, if player *i* receives a share x_i in an agreement reached in period *t*, his payoff is $\delta^{t-1}v_i(x_i)$. Perpetual disagreement has a payoff of 0.

Using subgame perfect equilibrium as the solution concept (the standard tool to rule out non-credible threats in dynamic games of complete information), Rubinstein (1982) shows that there exists a unique prediction in his game. Specifically, the unique subgame perfect equilibrium prescribes an immediate agreement on the splits (x,1–x) – offered by player 1 – and $(y, 1-y)$ – by player 2 – which are described by the following equations:

$$v_1(y) = \delta v_1(x)$$
$$v_2(1-x) = \delta v_2(1-y)\,.$$

That is, at the unique equilibrium, the player acting as a responder in a period is offered a share that makes him exactly indifferent between accepting and rejecting it to play the continuation: the bulk of the proof is to show that any other behaviour relies on non-credible threats.
As demonstrated in Binmore, Rubinstein and Wolinsky (1986), the unique equilibrium payoffs of the Rubinstein game, regardless of who is the first proposer, converge to the Nash solution payoffs as $\delta \to 1$. First, note that the above equations imply that, for any value of δ, the product of payoffs $v_1(x)v_2(1-x)$ is the same as the product $v_1(y)v_2(1-y)$. Thus, both points, $(v_1(x),v_2(1-x))$ and $(v_1(y),v_2(1-y))$, lie on the same hyperbola of equation $v_1 v_2 = K$ and, in addition, since they correspond to efficient agreements, both points also lie on the Pareto frontier of $V(\{1, 2\})$. Finally, as $\delta \to 1$, one has that $x \to y$ so that the two proposals (the one made by player 1 and the other by player 2) converge to one and the same, the one that yields the Nash solution payoffs. Thus, credible threats in dynamic negotiations in which both players are equally and almost completely patient also lead to the Nash solution.

The Shapley value

Now consider an n-player coalitional game where payoffs are transferable in a one-to-one rate among different players (for instance, because utility is money for all of them). This means that $V(S)$, the feasible set for coalition S, is the set of payoffs $(x_i)_{i \in S}$ satisfying the inequality $\sum_{i \in S} x_i \leq v(S)$ for some real number $v(S)$. This is called a transferable utility or TU game in characteristic function form. The number $v(S)$ is referred to as the ‘worth of S’, and it expresses S's initial position (for example, the maximum total utility that the group S of agents can achieve in an exchange economy by redistributing their endowments when utility is quasi-linear).
Therefore, without loss of generality, we can describe a TU game as a collection of real numbers $(v(S))_{S \subseteq N}$. A solution is then a mapping that assigns to each TU game a set of payoffs in the set $V(N)$, that is, vectors $(x_1, \ldots, x_n)$ such that $\sum_{i \in N} x_i \leq v(N)$. In this section, as in the previous one, we shall require that the solution be single-valued. Shapley (1953) is interested in solving in a fair way the problem of distribution of surplus among the players, when taking into account the worth of each coalition. To do this, he resorts to the axiomatic method. First, the payoffs must add up to $v(N)$,which means that the entire surplus is allocated (efficiency). Second, if two players are substitutes because they contribute the same to each coalition, the solution should treat them equally (symmetry). Third, the

solution to the sum of two TU games must be the sum of what it awards to each of the two games (additivity). Fourth, if a player contributes nothing to every coalition, the solution should pay him nothing (dummy).

The result in Shapley (1953) is that there is a unique single-valued solution to TU games satisfying efficiency, symmetry, additivity and dummy. It is what today we call the Shapley value, the function that assigns to each player *i* the payoff

$$\mathrm{Sh}_i(N, v) = \sum_{S, i \in S} \frac{(|S| - 1)!(|N| - |S|)!}{|N|!} \times [v(S) - v(S \setminus \{i\})].$$

That is, the Shapley value awards to each player the average of his marginal contributions to each coalition. In taking this average, all orders of the players are considered to be equally likely. Let us assume, also without loss of generality, that $v(\{i\}) = 0$ for each player *i*.

Hart and Mas-Colell (1996) propose the following non-cooperative procedure. With equal probability, each player $i \in N$ is chosen to publicly make a feasible proposal to the others: $(x_1, \ldots, x_n)$ is such that the sum of its components cannot exceed *v*(*N*). The other players get to respond to it in sequence, following a pre-specified order. If all accept, the proposal is implemented; otherwise, a random device is triggered. With probability $0 \leq \delta < 1$, the same game continues being played among the same *n* players (thus, a new proposer will be chosen again at random among them), but with probability $1-\delta$ the proposer leaves the game. He is paid 0 and his resources are removed so that, in the next period, proposals to the remaining *n*–1 players cannot add up to more than $v(N \setminus \{i\})$. A new proposer is chosen at random among the set $N \setminus \{i\}$, and so on.

As shown in Hart and Mas-Colell (1996), there exists a unique stationary subgame perfect equilibrium payoff profile of this procedure, and it actually coincides with the Shapley value payoffs for any value of δ. (Stationarity means that strategies cannot be history dependent.) As $\delta \to 1$, the Shapley value payoffs are also obtained not only in expectation but independently of who the proposer is. One way to understand this result, as done in Hart and Mas-Colell (1996), is to check that the rules of the procedure and stationary behaviour in it are in agreement with Shapley's axioms. That is, the equilibrium relies on immediate acceptances of proposals, stationary strategies treat substitute players similarly, the equations describing the equilibrium have an additive structure, and dummy players will have to receive 0 because no resources are destroyed if they are asked to leave. It is also worth stressing the important role in the procedure of players' marginal contributions to coalitions: following a rejection, a proposer incurs the risk of being thrown out and the others of losing his resources, which seem to suggest a ‘price’ for them.

The core

The idea of agreements that are immune to coalitional deviations was first introduced to economic theory in Edgeworth (1881), who defined the set of coalitionally stable allocations of an economy under

the name ‘final settlements’. Edgeworth envisioned this concept as an alternative to Walrasian equilibrium (Walras, 1874), and was also the first to investigate the connections between the two concepts. Edgeworth's notion, which today we refer to as ‘the core’, was rediscovered and introduced to game theory in Gillies (1959). Therefore, the origins of the core were not axiomatic. Rather, its simple definition appropriately describes stable outcomes in a context of unfettered coalitional interaction. (The axiomatizations of the core came much later: see, for example, Peleg, 1985; 1986; Serrano and Volij, 1998).

For simplicity, let us continue to assume that we are studying a TU game. In this context, the core is the set of payoff vectors $x = (x_1, \ldots, x_n)$ that are feasible, that is, $\sum_{i \in N} x_i \leq v(N)$, and such that there does not exist any coalition $S \subseteq N$ for which $\sum_{i \in S} x_i < v(S)$. If such a coalition *S* exists, we shall say that *S* can improve upon or block *x*, and *x* is deemed unstable. The core usually prescribes a set of payoffs instead of a single one, and it can also prescribe the empty set in some games.

To obtain a non-cooperative implementation of the core, the procedure must embody some feature of anonymity, since the core is usually a large set and it contains payoffs where different players are treated very differently. Perry and Reny (1994) build in this anonymity by assuming that negotiations take place in continuous time, so that anyone can speak at the beginning of the game instead of having a fixed order. The player that gets to speak first makes a proposal consisting of naming a coalition that contains him and a feasible payoff for that coalition. Next, the players in that coalition get to respond. If they all accept the proposal, the coalition leaves and the game continues among the other players. Otherwise, a new proposal may come from any player in *N*. It is shown that, if the TU game has a non-empty core (as well as any of its subgames), the stationary subgame perfect equilibrium outcomes of this procedure coincide with the core. If a core payoff is proposed to the grand coalition, there are no incentives for individual players to reject it. Conversely, a non-core payoff cannot be sustained because any player in a blocking coalition has an incentive to make a proposal to that coalition, who will accept it (knowing that the alternative, given stationarity, would be to go back to the non-core status quo). Moldovanu and Winter (1995) offer a discrete-time version of the mechanism: in their work, the anonymity required is imposed on the solution concept by looking at order-independent equilibria.

Serrano (1995) sets up a market to implement the core. The anonymity of the procedure stems from the random choice of broker. The broker announces a vector $(x_1, \ldots, x_n)$, where the components add up to $v(N)$. One can interpret x_i as the price for the productive asset held by player *i*. Following an arbitrary order, the remaining players either accept or reject these prices. If player *i* accepts, he sells his asset to the broker for the price x_i and leaves the game. Those who reject get to buy from the broker, at the called out prices, the portfolio of assets of their choice if the broker still has them. If a player rejects but does not get to buy the portfolio of assets he would like because someone else took them before, he can always leave the market with his own asset. The broker's payoff is the worth of the final portfolio of assets that he holds, plus the net monetary transfers that he has received. Serrano (1995) shows that the prices announced by the broker will always be his top-ranked vectors in the core. If the TU game is such that gains from cooperation increase with the size of coalitions, the set of all subgame perfect equilibrium payoffs of this procedure will coincide with the core. Core payoffs are here understood as those price vectors where all arbitrage opportunities in the market have been wiped out. Finally, yet another way to build anonymity in the procedure is by allowing the proposal to be made by brokers

outside of the set *N*, as done in Pérez-Castrillo (1994).

See Also

- bargaining
- non-cooperative games (equilibrium existence)
- Shapley value

Bibliography

Binmore, K., Rubinstein, A. and Wolinsky, A. 1986. The Nash bargaining solution in economic modelling. *RAND Journal of Economics* 17, 176–88.

Edgeworth, F. 1881. Mathematical psychics. In *F. Y. Edgeworth's Mathematical Psychics and Further Papers on Political Economy*, ed. P. Newman. Oxford: Oxford University Press, 2003.

Gillies, D. 1959. Solutions to general non-zero-sum games. In *Contributions to the Theory of Games IV*, ed. A. Tucker and R. Luce. Princeton, NJ: Princeton University Press.

Hart, S. and Mas-Colell, A. 1996. Bargaining and value. *Econometrica* 64, 357–80.

Moldovanu, B. and Winter, E. 1995. Order independent equilibria. *Games and Economic Behavior* 9, 21–34.

Nash, J. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Nash, J. 1953. Two person cooperative games. *Econometrica* 21, 128–40.

Peleg, B. 1985. An axiomatizationof the core of cooperative games without side payments. *Journal of Mathematical Economics* 14, 203–14.

Peleg, B. 1986. On the reduced game property and its converse. *International Journal of Game Theory* 15, 187–200.

Pérez-Castrillo, D. 1994. Cooperative outcomes through non-cooperative games. *Games and Economic Behavior* 7, 428–40.

Perry, M. and Reny, P. 1994. A non-cooperative view of coalition formation and the core. *Econometrica* 62, 795–817.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50, 97–109.

Serrano, R. 1995. A market to implement the core. *Journal of Economic Theory* 67, 285–94.

Serrano, R. 2005. Fifty years of the Nash program, 1953–2003. *Investigaciones Económicas* 29, 219–58.

Serrano, R. and Volij, O. 1998. Axiomatizations of neoclassical concepts for economies. *Journal of Mathematical Economics* 30, 87–108.

Shapley, L. 1953. A value for *n*-person games. In *Contributions to the Theory of Games II*, ed. A. Tucker and R. Luce. Princeton, NJ: Princeton University Press.

Walras, L. 1874. *Elements of Pure Economics, or the Theory of Social Wealth*, trans. W. Jaffé. Philadelphia: Orion Editions, 1984.

How to cite this article

Serrano, Roberto. "Nash program." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000137> doi:10.1057/9780230226203.1156

The New Palgrave Dictionary of Economics Online

Nash, John Forbes (born 1928)

Joel Watson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Nash originated general non-cooperative game theory in seminal articles in the early 1950s by formally distinguishing between non-cooperative and cooperative models and by developing the concept of equilibrium for non-cooperative games. Nash developed the first bargaining solution characterized by axioms, pioneered methods and criteria for relating cooperative-theory solution concepts and non-cooperative games, and also made fundamental contributions in mathematics. Nash was the 1994 recipient of the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, jointly with John C. Harsanyi and Reinhard Selten.

Keywords

coalitions; commitment; contract curve; cooperative games; Cournot, A. A.; dominance; equilibrium; equilibrium refinements; evolutionary stability; expected utility; fixed-point methods; game theory; maximin strategy; Morgenstern, O.; multiple equilibria; Nash bargaining solution; Nash demand game; Nash equilibrium; Nash program; Nash. J. R., Jr.; non-cooperative games; prisoner's dilemma; rational behaviour; strategic and extensive-form games; strategic independence; von Neumann, J.

Article

The context of for Nash's work: von Neumann and Morgenstern

Nash's contributions to the theory of games were fundamental to the development of the discipline and its interface with applied fields of study. This section provides is a short account of the state of affairs before Nash's work. For a more detailed account, see the suggestions for further reading at the end of this article.

The first significant step in mathematical modelling of strategic situations was Augustin Cournot's (1838) book on oligopoly, where Cournot presented models of firm interaction that were analysed using what we now call Nash equilibrium. But Cournot did not attempt, or perhaps even recognize, how the analysis might generalize. Further, in the ensuing years confusion persisted regarding whether it would be appropriate for a firm to incorporate a response by its rivals when considering whether to change its own action. The concept of *strategic independence* – that the players' strategies can be considered to be

chosen simultaneously and independently – began to be clarified by Emile Borel's (1921) description of a *method of play*.

Game theory became a discipline with the work of John von Neumann (1928), which was incorporated into the path-breaking book by von Neumann and Oscar Morgenstern (1944; 1947). In the book, von Neumann and Morgenstern formally defined both the extensive form (tree-based) and normal form (strategy-based) representations of games, related by the notion of a strategy; they studied for the first time a general class of games, defining solutions and proving existence using fixed-point methods; they introduced the idea of analysing how coalitions of players can take advantage of binding agreements; and they provided a theory of utility and decision-making under risk (the expected utility criterion). With one book, game theory was created and put on solid footing.

Von Neumann and Morgenstern were interested in developing a positive theory of behaviour in games – for any given game, a ‘solution’. In a nutshell, their analysis progresses as follows:

1. 1. Formulate a solution concept for two-player *zero-sum games*, which have the defining property that, for each *strategy profile* (one strategy for each player), the players' payoffs sum to zero. Such a game is special because the only economic concern is distributional; in other words, the game models a situation of pure conflict between the players, where one player's winnings come at the other's expense.
2. 2. Analyse *n*-player zero-sum games by assuming that coalitions of players could bind together and play as a team against the other players. This requires assuming that coalitions can communicate before the game and make binding agreements on how to play. The value of forming a coalition is calculated in reference to the implied zero-sum game that the coalitions play against one another, which ultimately is a two-player game to which the solution from Part 1 above is applied.
3. 3. To evaluate a non-zero-sum, *n*-player game, imagine the existence of a fictitious player *n*+1 whose payoff is defined as negative of the sum of the other players' payoffs. This creates a zero-sum game to which the preceding applies.

For an illustration of von Neumann and Morgenstern's analysis of two-player zero-sum games (Part 1 above), consider a simple example. Suppose that players 1 and 2 interact in the normal form game depicted in the following table.

1\2	X	Y	Z
A	4, – 4	0, 0	– 2, 2
B	3, – 3	1, – 1	1, – 1
C	2, – 2	1, – 1	1, – 1

Player 1 selects between strategies A, B, and C. Simultaneously, player 2 chooses between X, Y, and Z. The players' payoffs, which might as well be in

monetary terms, are shown in the cells of the table, with player 1's payoff written first. Note that this is a zero-sum game in that, in each cell of the table, the players' payoffs sum to zero.

Von Neumann and Morgenstern motivated their solution concept by considering sequential variations of games in which one player would move first and then the other player, having seen what the first selected, would respond. Their key concept is what is generally known as a ‘maximin strategy’, also called a ‘security strategy’. A security strategy for a given player is a strategy that gives the highest guaranteed payoff level; that is, it maximizes the minimum that the player could get, where the minimum is calculated over all of the strategies of the other player.

In the example, B and C are both security strategies for player 1 because, regardless of what player 2 does, player 1 gets a payoff of at least 1 when using either of these strategies, whereas it is feasible for player 1 to obtain a lower payoff (0 or –2, in particular) by selecting strategy A. For player 2, Y and Z are security strategies and they guarantee a payoff of at least –1.

Von Neumann and Morgenstern's general analysis focuses on mixed strategies (probability distributions over pure strategies) in finite two-player games, to which the maximin definition extends. They prove that the players' security levels (the amounts that the security strategies guarantee) sum to zero. Thus, when each player selects his security strategy, each player obtains exactly his security level payoff. Further, when one player selects his security strategy, the other player can do no better than select her own security strategy; that is, the two players' security strategies are optimal responses to each other. Security strategies also describe optimal play in zero-sum games that are played sequentially. For example, if player 1 had the privilege of selecting among A, B, and C *after* observing player 2's choice, both players would still select security strategies. Finally, security strategies are interchangeable in that the preceding conclusions hold equally well for any combination of security strategies, for instance (B, Y) as well as (B, Z).

Although von Neumann and Morgenstern had developed a theory that applied to all finite games, their theory is essentially empty for non-zero-sum games. For example, in converting a two-player game into a three-player game by adding the fictitious player 3, von Neumann and Morgenstern basically change the rules of the game for the original two players, who now can make binding agreements. The resulting prediction is that the two players will bind themselves to a strategy profile that maximizes the sum of their payoffs, with each player getting at least his security level. Von Neumann and Morgenstern's theory is therefore incomplete and unsatisfying on two fronts. First, for non-zero-sum games, it offers no treatment of rationality in the absence of binding commitments. Second, it offers no way of predicting the outcome of a two-player bargaining problem beyond Francis Ysidro Edgeworth's (1881) contract curve and it relies on transferable utility. Nearly all interesting economic examples involve efficiency concerns and hence are not zero-sum in nature, so economics had little to benefit from game theory until another significant step could be made in the modelling of rational behaviour.

Nash's contributions

Nash's contributions to the emerging discipline of game theory were equally as bold as were von Neumann and Morgenstern's and, in terms of applicability, even more significant. Nash's main contributions were made in a series of four papers published between 1950 and 1953 and summarized in this section.

In his articles in the *Proceedings of the National Academy of Sciences* in 1950 and the *Annals of Mathematics* in 1951, which reported his dissertation research, Nash (*a*) introduced and made clear the distinction between cooperative and non-cooperative games – the latter being games in which players act independently (that is, without the assumption about coalitions that von Neumann and Morgenstern adopted) – and (*b*) defined a solution concept for non-cooperative games. The first four paragraphs from Nash's *Annals of Mathematics* article describe the context and the contribution succinctly:

> Von Neumann and Morgenstern have developed a very fruitful theory of two-person zero-sum games in their book *Theory of Games and Economic Behavior*. This book also contains a theory of *n*-person games of a type which we would call cooperative. This theory is based on an analysis of the interrelationships of the various coalitions which can be formed by the players of the game.
>
> Our Theory, in contradistinction, is based on the *absence* of coalitions in that it is assumed that each participant acts independently, without collaboration or communication with any of the others.
>
> The notion of an *equilibrium point* is the basic ingredient in our theory. This notion yields a generalization of the concept of the solution of a

> two-person zero-sum game. It turns out that the set of equilibrium points of a two-person zero-sum game is the set of all pairs of opposing 'good strategies.'
> In the immediately following sections we shall define equilibrium points and prove that a finite non-cooperative game always has at least one equilibrium point. We shall also introduce the notions of solvability and strong solvability of a non-cooperative game and prove a theorem on the geometrical structure of the set of equilibrium points of a solvable game. (1951, p. 286)

Nash's equilibrium concept became known as 'Nash equilibrium'. It and the cooperative/non-cooperative distinction were cited by the Royal Swedish Academy of Sciences in awarding Nash the Nobel Prize.

In more mathematical and modern language, here are the definitions of *best response* (in Nash's words, a 'good strategy') and Nash equilibrium. Consider any game defined by a number n of players; a strategy set S_i for each player $i = 1, 2, \ldots, n$; and, for each player i, a payoff function $u_i: S \to \mathbf{R}$, where S is the set of strategy profiles. The strategy sets may be defined as mixed strategies for some underlying set of pure strategies, in which case the payoff functions, as expectations, are linear in the mixed strategies. For a player i, we write '$-i$' to refer to the other players. Given a strategy vector s_{-i} for the other players, player i's strategy s_i is called a best response if player i can do no better than to select s_i; that is, we have $u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i})$ for every strategy s_i' of player i. Then strategy profile $s^* = (s_1^*, s_2^*, \ldots, s_n^*)$ is called a Nash equilibrium if every player is best responding to the others—that is, if for each player i, it is the case that s_i^* is a best response to s_{-i}^*.

For an illustration of Nash equilibrium and its relation to security strategies, consider the game depicted in the following table.

1\2	X	Y	Z
A	2,3	1,2	6,5
B	1,0	0,2	4,0
C	3,4	2,2	2,0

Observe that, in this game, C and Y are the players' security strategies, so a naive application of von Neumann and Morgenstern's maximin theory (absent

binding agreements) would predict that strategy profile (C, Y) be played. However, this strategy profile is plainly inconsistent with the idea that players are rational in responding to each other. In particular, if player 1 is expected to select C then player 2 behaves quite irrationally by choosing Y. In fact, strategy Y is *not even rationalizable* for player 2; it does not survive iterated removal of dominated strategies (see below). Thus, the notion of a security strategy is not a good theory of behaviour for non-zero-sum games, demonstrating the limits of von Neumann and Morgenstern's analysis.

Next, observe that the game has two Nash equilibria in pure strategies, (C, X) and (A, Z). Both of these are reasonable predictions in the sense that, in both cases, the players are best responding to one another. For example, if player 1 is sure that player 2 will select X, then it is best for player 1 to select C; likewise, if player 2 is convinced that player 1 will select C, then it is optimal for player 2 to choose X. There is also a mixed-strategy Nash equilibrium in which player 1 randomizes between A and C, and player 2 randomizes between X and Z. That the game has multiple Nash equilibria demonstrates the general economic problem of coordination, in particular the possibility that the players will coordinate on the less efficient Nash equilibrium. Other games, such as the *Prisoner's Dilemma*, have only inefficient equilibria and thus reveal a fundamental tension between individual and joint incentives.

Nash's intuitive concept of equilibrium facilitated the analysis of *all* non-cooperative games, opening the door to widespread application of game theory. Indeed, Nash equilibrium has become the dominant solution concept for the analysis of games. Through an ingenious fixed-point argument, Nash also proved the existence of an equilibrium point in every finite game. Further, in his dissertation (1950) Nash offered two interpretations of the concept, one based on rational reasoning by individual players and the other describing stability of the distribution of strategies chosen by a population of individuals who interact over time. The latter is a precursor to the methodology of the literature on learning in games and to the modern theories of *evolutionary stability* in biology (John Maynard Smith, 1984). Nash's 1951 *Annals of Mathematics* article also contains a section that defines ‘dominance’ (meaning one strategy yields a strictly higher payoff than another, regardless of what the other players do) and explains how an iterated dominance procedure can be used to rule out strategies that are not equilibria. Thus, Nash also made observations that would resurface in the concept of ‘rationalizable strategic behaviour’ (B. Douglas Bernheim, 1984; David Pearce, 1984), the main non-equilibrium notion of rationality. Nash even was among the first to perform game experiments, as his co-authored article in the volume *Decision Processes* (Kalisch et al., 1954) attests.

In his 1950 *Econometrica* article, Nash tackled the two-person bargaining problem with the objective of determining a unique solution (a precise ‘value’ that eluded von Neumann and Morgenstern) from the underlying set of alternatives and the players' preferences. Nash took a cooperate-theory approach by positing a system of four axioms that reasonably characterize properties one might expect the outcome of a bargaining process to exhibit: (*a*) a notion of equal bargaining power, (*b*) invariance to inessential utility transformations, (*c*) efficiency, and (*d*) independence of the solution to the removal of so-called irrelevant alternatives. Nash proved that a particular function of parameters (which maximizes the product of surpluses) is exactly characterized by the axioms. The analysis showed that it is possible to reasonably identify a precise outcome of a bargaining problem. It also initiated the axiomatic method for the analysis of bargaining (where theorists explore how different axioms characterize various functional solutions), starting a literature that thrived for several decades. The *Nash bargaining solution* is still the dominant solution in applied economic models.

Nash's second paper on bargaining (the 1953 *Econometrica* article) took another major step by connecting the non-cooperative and cooperative approaches to strategic analysis. At the heart of this theoretical exercise is an underlying non-cooperative game, which gives a set of feasible payoffs, and a technology for the players to make binding commitments about the mixed strategies that they will play in the underlying game. In the model, players first simultaneously make threats, which are mixed strategies they are bound to play if they do not reach an agreement. Then the players interact in a non-cooperative bargaining game in which they simultaneously make payoff demands – this stage is now called the ‘Nash demand game’. If their payoff demands are feasible in the underlying game, then the players obtain their demanded payoffs; otherwise, the players get what their threats imply.

Nash observed that the demand game has generally an infinite number of equilibria, revealing a coordination aspect to the bargaining problem. But Nash went further in developing a brilliant method to ‘escape from this troublesome non-uniqueness’ by looking at the limit of ‘smooth’ approximations of the demand game. Amazingly, Nash showed that the limit is unique and coincides with the prediction of his axiomatic model; that is, the limit is the Nash bargaining solution. Nash's limit argument was the forerunner to the enormous literature on *equilibrium refinements*, an area of research that thrived decades later and was the primary subject of Nash's Nobel co-recipients. More significantly, Nash argued that the relation between the cooperative solution concept and the equilibrium in the non-cooperative model justifies wide use of the cooperative solution as a reasonable shorthand for the actual non-cooperative setting.

Nash's argument, and fascinating theoretical result, established the profession's understanding of the connection between cooperative and non-cooperative

models and initiated the literature on what is now called the ‘Nash program’.

After completing the work in game theory just described, Nash made fundamental contributions in pure mathematics – contributions that, in terms of mathematical depth and originality, were of an even higher order of sophistication and importance. According to leading mathematician John Milnor, Nash's

> subsequent mathematical work is far more rich and important [in this mathematical sense]. During the following years he proved that every smooth compact manifold can be realized as a sheet of a real algebraic variety, proved the highly anti-intuitive C1-isometric embedding theorem, introduced powerful and radically new tools to prove the far more difficult C1-isometric embedding theorem in high dimensions, and made a strong start on fundamental existence, uniqueness, and continuity theorems for partial differential equations. (Milnor, 1998, p. 1330)

It is not appropriate to provide here details on Nash's pure mathematics work (nor is it possible, due to the limitations of the author's fields of expertise).

Nash's personal life

Nash's character became legendary with the publication of a biography by Sylvia Nasar (1998) and a 2001 feature film produced by Brian Grazer and Ron Howard. Nash's remarkable personal journey began in Bluefield, West Virginia, where he was born and raised. He explored mathematics and conducted science experiments as a child, and attended Carnegie Institute of Technology, where the mathematics department discovered in him a budding genius. Nash's ideas on bargaining that were published as ‘The Bargaining Problem’ (1950) were developed while he was an undergraduate student at Carnegie, during the only economics course he took, on international trade.

Nash studied mathematics in the graduate program at Princeton University, where, as his biography describes, he was boorish, cocky, and a renowned adversary in strategic contests. At Princeton, Nash added to his prodigious achievements, finishing his dissertation – the work on non-cooperative games and equilibrium that would bring him the Nobel Prize – in his second year. (Nash also invented the board game *Hex*, a game independently created by Danish mathematician Piet Hein.) Nash taught at Princeton for one year and then took a position at Massachusetts Institute of Technology, where he was on the faculty until 1959. There he conducted the research that won him great acclaim in the mathematics community.

Nash's genius in advancing game theory and mathematics was paired with deep personal challenges. In 1959 Nash began experiencing the severe mental disturbances of paranoid schizophrenia. He resigned from MIT and began a phase of life marked by delusional thinking, an escape to Europe, repeated hospitalizations, unsuccessful medical treatments, and then a long, disengaged presence at Princeton. In the mid-1980s Nash miraculously began to emerge from the delusional haze in what he describes as a gradual rejection of psychotic thinking on intellectual grounds (Nash, 1995). After a quarter century of detachment, Nash's life regained a measure of normality.

Nash's legacy in game theory and economics

There is no simple way of quantifying the enormous reach of Nash's ideas. The notions of Nash equilibrium, the Nash bargaining solution, the Nash demand game, and the Nash program have found such widespread acceptance and application that it has become customary, and perhaps even appropriate, for researchers to forgo formally citing Nash's articles when utilizing these concepts. Nash ideas helped to propel game theory from a mathematical sub-field into a full discipline, with major use and application in not only economics, where it is the main and worthy alternative to the competitive-market framework, but also in theoretical biology, political science, international relations and law.

Beyond its theoretical content, Nash's work also made a stylistic departure from that of von Neumann and Morgenstern, whose book methodically records definitions, examples, and analysis for numerous special cases in the process of developing general theory. Nash, in contrast, used the terse style of the mathematician, presenting his ideas with minimal obscuring features. His 1950 *Proceedings of the National Academy of Sciences* entry, for instance, is generously allotted two pages and could have been typeset on one. The benefit of focusing on the basic mathematical concepts is that it allows for a broad

range of interpretations and extensions. For example, there are several motivations for Nash equilibrium, including as a condition for self-enforcement of a contract (which is an important topic in the current literature). A hallmark of excellent theoretical modelling is precise and straightforward expression of assumptions and conclusions, with their relation shown in the most simple and elegant way possible.

Mathematician Milnor, after offering the assessment of Nash's work in pure mathematics that is quoted above, continues with by saying: ‘However, when mathematics is applied to other branches of human knowledge, we must really ask a quite different question: To what extent does the new work increase our understanding of the real world? On this basis, Nash's thesis was nothing short of revolutionary’ (1998, p. 1330). Two leading game theorists of today say ‘Nash's theory of non-cooperative games should now be recognized as one of the outstanding intellectual advances of the twentieth century’ (Myerson, 1999, p. 1067) and ‘His work lay the foundation of non-cooperative game theory, now the predominant mode of analysis of strategic interactions in economics, political science, and biology’ (Crawford, 2002, p. 380).

When viewed from the perspective of five short decades, game theory has caused a revolution in economics and other fields of study. It was with the work of John Nash that the flame so exquisitely ignited by von Neumann and Morgenstern became the torch that would eventually set the social sciences ablaze.

See Also

- bargaining
- game theory
- Morgenstern, Oskar
- Nash program
- non-cooperative games (equilibrium existence)
- von Neumann, John

The author thanks Vincent Crawford, Joel Sobel and Martin Dufwenberg for comments on a preliminary draft.

Bibliography

Items indicated with an asterisk provide good further background reading on John F. Nash, Jr. Also, the *Scandinavian Journal of Economics*, vol. 97, issue 1 (1995), contains articles on John Nash and his co-Nobel Prize recipients, John C. Harsanyi and Reinhard Selten. For a complete list of Nash's publications, including his papers in pure mathematics, see Milnor (1998).

Bernheim, B.D. 1984. Rationalizable strategic behavior. *Econometrica* 52, 1007–28.

Borel, E. 1921. La théorie du jeu et les équations intégrales à noyau symétrique gauche. *Comptes Rendus de l'Académie des Sciences* 173, 1304–08. English translation by L.J. Savage, *Econometrica* 21 (1953), 97–100.

Cournot, A. 1838. *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. Paris: Hatchette. English translation by N.T. Bacon, Researches into the Mathematical Principles of the Theory of Wealth. New York: Macmillan, 1927.

Crawford, V.P. 2002. John Nash and the analysis of strategic behavior. *Economics Letters* 75, 377–82.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Hammerstein, P. et al. 1996. The work of John Nash in game theory: Nobel seminar, December 8, 1994. *Journal of Economic Theory* 69, 153–85.

Kalisch, C., Milnor, J., Nash, J. and Nering, E. 1954. Some experimental n-person games. *Decision Processes*, ed. R.M. Thrall, C.H. Coombs and R.L. Davis. New York: Wiley.

Mayberry, J.P., Nash, J.F. and Shubik, M. 1953. A comparison of treatments of a duopoly situation. *Econometrica* 21, 141–54.

Maynard Smith, J. 1984. *Evolution and the Theory of Games*. New York: Cambridge University Press.

*Milnor, J. 1995. A Nobel Prize for John Nash. *The Mathematical Intelligencer* 17, 11–7.

*Milnor, J. 1998. John Nash and ‘A Beautiful Mind’. *Notices of the American Mathematical Society* 45, 1329–32.

Myerson, R.B. 1999. ‘Nash equilibrium and the history of economic theory. *Journal of Economic Literature* 37, 1067–82.

*Nasar, S. 1998. *A Beautiful Mind*. New York: Simon and Schuster.

Nash, J.F., Jr. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences, USA* 36, 48–9.

Nash, J.F., Jr. 1950. Non-cooperative games. Doctoral dissertation, Princeton University.

Nash, J.F., Jr. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Nash, J.F., Jr. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

Nash, J.F., Jr. 1953. Two-person cooperative games. *Econometrica* 21, 128–40.

*Nash, J.F., Jr. 1995. Autobiography. *Les Prix Nobel. The Nobel Prizes 1994*, ed. T. Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1994/nash-autobio.html, accessed 29 November 2006.

Pearce, D. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52, 1029–50.

von Neumann, J. 1928. Zur theories der gesellschaftsspiele. *Mathematische Annalen* 100, 295–320. English translation by S. Bergmann in *Contributions to the Theory of Games IV*, ed. R.D. Luce and A.W. Tucker. Princeton: Princeton University Press, 1959.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press (2nd edn 1947).

How to cite this article

Watson, Joel. "Nash, John Forbes (born 1928)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000155> doi:10.1057/9780230226203.1154

The New Palgrave Dictionary of Economics Online

national accounting, history of

André Vanoli

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

With antecedents as far back as the late 17th century, national accounting is a product of the Great Depression, the Second World War and the subsequent period of recovery and economic growth. Soon after the war, country experiences and international harmonization processes interacted, eventually leading to a complete accounting framework with the 1993 SNA/ESA 1995. Until the mid-1970s, national accounting experienced a kind of golden age, after which greater difficulties arose, in terms of the increased complexity of economic life, widened social concerns and theoretical challenges. In that context, impressive achievements and a sense of frustration have coexisted.

Keywords

aggregates; Aukrust, O.; balance sheets; basic prices; Boisguilbert, P.; Brown, A.; Clark, C. G.; computable general equilibrium models; Copeland, M.A.; Denison, E.; domestic product; Economic and Monetary Union; European System of Accounts; factor cost valuation; fair value valuation; Fisher, I.; Frisch, R. A. K.; Gilbert, M.; Goldsmith R.; Gordon, R.; Great Depression; Grüning, F.; Gruson, C.; Hicks, J. R.; historical cost valuation; income; indicative planning; input–output analysis; International Monetary Fund; Kampmann, V.; Keynes, J. M.; King, G.; Kravis, I.; Kuznets, S.; Leontief, W.; Lindahl, E. R.; Maastricht criteria; market price valuation; Marx, K.; Material Product System; mesoeconomic level; national accounting; national accounts statistics; national income; Ohlsson, I.; Petty, W.; Popov, P. I.; purchasing power parities; Quesnay, F.; Ruggles, N.; Ruggles, R.; satellite accounts; Smith, A.; social accounting matrices; social accounting,; Stone, J. R. N.; structural change; sustainability; System of Balances of the National Economy; System of National Accounts; *tableau économique* ; Tinbergen, J.; value added; Van Cleeff, E.; Vauban, S.; Vincent, A.; Walras, L.; Warburton, C.; wealth; welfare indicators

Article

National accounting is a product of the 20th century, more precisely of the Great Depression, the Second World War and the subsequent period of recovery and economic growth. However, two and a half centuries earlier, estimates of national income had started with William Petty and Gregory King in England, and Vauban and Boisguilbert in France. This innovation in England, by the end of the 17th century, has been attributed to ‘the spirit of the age’ (Phyllis Deane, 1955), ‘an age of great intellectual vigour, scientific curiosity and inventiveness’ (Richard Stone, 1986). This early work had two main purposes: on the one hand, taxation and fiscal reforms, and on the other the assessment of the nations' comparative economic strength in an age when England, France and the Netherlands were frequently at war. Exceptionally, King, an outstanding pioneer, made consistent estimates of various economic magnitudes (income, expenses, increase or decrease in wealth, and so on) for a series of years. However, as a rule, national income was estimated as an isolated magnitude using various methods. Estimates were intermittent and extended slowly (according to Studenski, 1958, national income had been estimated at least once for only eight countries by the end of the 19th century, and for some 20 by 1929. From 1850, earlier in England, evaluations of fortune or wealth, more numerous, were disconnected from national income estimates.

From national income estimate to national accounting

The influence of the First World War was limited, with some exceptions (for example, an NBER 1909–19 series in current and constant dollars published by Wesley Mitchell et al. in 1921–22). The 1929 crisis was a turning point. Official demand appeared (US Senate, 1932; Carson, 1975, p. 156) leading to a 1934 report prepared by Simon Kuznets and his assistants (*National Income 1929–1932*, in current prices, by type of economic activity and distributed income). Estimates were then extended to expenditures (final consumption and capital formation) by Clark Warburton. In a number of countries – the Netherlands (Jan Tinbergen), Sweden, Denmark (Viggo Kampmann) – large programs were developed, such as the one resulting in *National Income in Sweden 1861–1930* published in 1937 by Erik Lindahl, Einar Dahlgren and Karin Koch. Working on his own, Colin Clark in the United Kingdom extended his previous 1932 estimates to a quite comprehensive coverage (*National Income and Outlay*, 1937).

The 1930s were a period of maturation in economics, apart from the conceptual and methodological deepening directly involved in this stream of quantitative estimates. The stimulus to quantitative macroeconomics given by Keynes's *General Theory* (1936) provided the theoretical basis for the estimation of interdependent economic aggregates, for the relationships between income and expenditure and between saving and investment were central to his argument. Such interrelationships had not previously been absent from economic theories (think of Quesnay's *Tableau économique*, Marx's reproduction schemes or Walras's general equilibrium analysis). However, after the Great Depression, such concepts and their statistical representations became central to macroeconomic concerns and policies. Keynes's works were focused on macroeconomic relations, but others sought representations of the economic system as a whole in different ways. Ferdinand Grüning in Germany (1933) analysed the economic circuit at a level later called ‘mesoeconomic’, half-way between the macro and micro levels. Wassily Leontief's research (1941) introduced input–output analysis at the level of homogeneous industrial groups, with a much broader view, in terms of general equilibrium, than the descriptive

detailed balances of relations between branches (industries) prepared by P.I. Popov (1926) in the Soviet Union. The idea of an accounting approach for the economy as a whole, similar to the business accounting approach, was introduced either as a tool for improving national income estimates (as by Morris A. Copeland, following an intuition of Irving Fisher) or as part of a new proposed economic organization (André Vincent in France, Ed Van Cleeff in the Netherlands). The idea of micro/macro relationships was present in much of this work. Coming from a very different perspective, Ragnar Frisch developed an axiomatic, bottom-up representation of economic circulation.

The Second World War was the second, decisive, turning point. National accounting, often called at the beginning social accounting, crystallized in a direct response to the problem of war finance in the UK, as explicitly stated in the April 1941 White Paper (UK Treasury, *An Analysis of the Sources of War Finance and Estimate of the National Income and Expenditure in 1938 and 1940*). This was backed up by a technical paper by James Meade and Richard Stone in 1941. A more elaborated ‘social accounting’ system was soon proposed by Stone in an appendix to *Measurement of National Income and The Construction of Social Accounts* (published by the United Nations in 1947). Inspired by business accounting, it included sector accounts grouping accounting entities and their transactions organized according to a sequence of sub-accounts, with a set of detailed definitions and the discussion of many unsettled issues. Although it covered neither balance sheets nor a detailed analysis of the productive system, this accounting system was well ahead of its time. Actually, before and during the war, the United States was in advance in both national income and related aggregates estimates and their use, as for instance in the 1942 feasibility study of the Victory Program led by Kuznets or the analysis of the inflationary gap (Carson, 1975, p. 174–7). However, the National Income Division of the Commerce Department, with Milton Gilbert, evolved towards a simple accounting framework rather than a developed accounting system.

Though they encountered many difficulties and though it was a very uneven development, mostly due to deficiencies in statistical information and staffing, national accounting experienced a kind of golden age in the three decades following the war. Economic reconstruction and growth policies, the large increase in the economic role of government and the welfare state, the extension of international cooperation (for example, the Marshall Plan and, later, the Common Market in Europe), with the consequent emphasis on measuring of the rate of growth, led to a great demand for national accounts. This comprised the requirements of Keynesian macroeconomic demand management for short-term economic budget forecasts and longer-term projections needed for various types of indicative planning (the latter being particularly important in France). The development of econometric techniques and national accounts estimates reinforced each other. This trend towards greater use of national accounting data was general, even though the economies involved ranged from basically liberal economies such as the United States to more controlled economies such as France, the Netherlands and Norway.

International harmonization and extensions

Country experiences interacted with the process of international harmonization very early. Discussion between Canada, the UK and the USA took place in September 1944. There was a meeting of a League of Nations Committee, for which Stone prepares a memorandum, in December 1945. Stone played a prominent role in the first generation of standardized systems (OEEC, 1950; 1952; United Nations,

1952). This first attempt at standardization across the Western world as a whole, however, was too limited in scope, and was very far from the ambitions of the 1945 accounting scheme. Conceived as a simplified model for countries that were only beginning to develop their national accounts, it could not meet the needs of countries that were already more advanced, such as Scandinavian countries (Odd Aukrust in Norway, Ingvar Ohlsson in Sweden) or even a country like France. Under the impulse of Claude Gruson, France was, in the 1950s, in order to implement far-reaching economic policies, beginning the process of building a comprehensive and ambitious system of its own, integrating accounts for economic agents, input–output tables and financial transactions in a way that was more integrated than the Copeland's money-flows accounts in the United States.

Until the end of the 1960s the Western stage was characterized by the existence of a variety of national systems that were difficult to reconcile, even among those countries that adopted, in principle, the same comprehensive concept of production, including non-market government services. The new French system adopted a narrower concept of production, limited to market goods and services. The Soviet Union and its satellites used the even more restricted concept of material production, limited to goods and the so-called material services (mostly the transport of goods), following the old tradition of Smith and Marx. However, during the 1960s intense international discussions took place, on the basis of the wide range of national experiences in Europe and North America and the demands of international organizations. The result was the adoption of a second generation of standardized systems, the 1968 System of National Accounts (SNA) and the new European System of Accounts (ESA 1970), prepared on the basis of a report by Stone for the UN (the OECD deleting its system) and a French expert for the European Community. The European Community, thinking the 1952 system was too narrow and unsuited to harmonizing the accounts of its original six members and to meeting the needs of Community policies, had decided in 1964 to establish its own system.

The new system (they can be described as a single system, for SNA and ESA were very close) was closer to Stone's 1945 inspiration and to the French, Scandinavian and British systems than to the 1952 standardized system, in terms of coverage (in particular of input–output tables and financial accounts), integration and institutional orientation. The main weakness remained the absence of balance sheets, despite the pioneering work of Raymond Goldsmith in the United States at the beginning of the 1960s. Fixed capital formation was limited to tangible assets and the relation between income and changes in wealth was not fully shown.

The System of Balances of the National Economy, built around the material product concept, was also standardized, though little innovation was involved, through the framework of the Council of Mutual Economic Assistance, and then published by the United Nations (1971). Careful comparisons between the SNA and the Material Product System (MPS) were carried out in the UN European Economic Commission in Geneva.

France decided to leave its own peculiar system and join, via ESA 1970, the international system, this being achieved by 1976. The USA was not actively involved in the elaboration of the 1968 SNA, keeping its National Income and Product Accounts, whose accounting and conceptual framework had evolved little since 1947.

A quarter of a century later, a third generation of normalized systems has taken the trend towards a universal system a step further. The 1993 SNA/ESA 1995 closed the accounting framework by including balance sheets and completing the accumulation accounts with the introduction of a revaluation account (holding gains and losses) and an account for other types of capital gains and losses. Intangible capital

formation was partly accounted for. In the current accounts, the analysis of income distribution was deepened (primary income distribution, secondary distribution, and redistribution in kind), actual final consumption was differentiated from final consumption expenditures, via the re-routing of social transfers in kind from government to households. This clarification of the accounting relation between income and changes in wealth (net worth) has deep implications (see below).

Nearly full integration was achieved between the SNA and the International Monetary Fund manuals (Balance of Payments, Government Finance Statistics, Monetary and Financial Statistics). The MPS disappeared at the beginning of the 1990s with the collapse of the Soviet Union and the fast transition of China towards a market economy. Paradoxically, the USA followed a slower path towards adopting the SNA framework.

During this long process of extension and harmonization of the accounting framework, the substance of the accounts changed dramatically in comparison with what was involved when the focus was on estimating national income. The product aggregate soon became the most important one, on a par with the expenditure aggregate. The income aggregate not only lost its position of being the single aggregate, but was often given a secondary position. From that, a series of consequences resulted.

The factor cost method of valuation, when still in use, was reduced to a lower rank than the market price valuation (in spite of the recurrent objection of ‘double counting’). The latter was much more convenient for the valuation of expenditure and the analysis of consumer behaviour. In an integrated framework, the market price valuation was then applied also to the product aggregate (domestic product takes progressively the first place) and much later on to the income aggregate. In the 1993 SNA, full recognition was given to the concept of national income at market prices, which is in fact the new name given to the earlier concept of national product (which was not actually a product but an income concept).

Partly for similar reasons, gross concepts have generally come to be preferred in practice, even though net concepts, that is, after deduction of consumption of fixed capital (depreciation in the usual business terminology), were considered closer to what was generally understood by the idea of national income. Both gross and net concepts of product, income and expenditure are finally considered part of the SNA/ESA.

The analysis and measurement of production and flows of products (goods and services), both in current value and in volume, have been given an increasing importance in relation to the integration of supply and use or input–output tables (a characteristic feature of the 1968 SNA/ESA 1970). This is increasingly done using the framework of annual tables. The integration with income estimates is less clear in practice, though the concept of value added, a significant improvement, and not only in words, on the old expression ‘net output’ or ‘net product’, provides the necessary link.

In this context, thanks to Stone's contribution, significant improvements in valuation concepts were made in the 1968 SNA. This widens and differentiates the usual notion of market prices. Basic prices, excluding net taxes on products, were introduced on the output side, resulting in the measurement of value added at basic prices. All taxes, minus subsidies, on products are then introduced. On the use side, acquisition prices are defined as purchasers' prices including only non-deductible taxes.

Measures in constant prices (described as volume measures), combining quantity and quality changes, also changed significantly. The trend was from globally deflating national income using a single price index in the 1930s, to deflating each of the main items in the balance of products (output, final consumption, and so on) using specific indices, and finally to an integrated system of price and volume

measures, at a detailed level, using an input–output framework when annual tables were available (with Denmark, France, the Netherlands and Norway leading here). Double deflation, of output and inputs respectively, was used for value added in this context. International manuals by Stone (1956; 1968 SNA, ch. 4) and Peter Hill (1972; United Nations, 1979) recommended such an approach. Later on the 1993 SNA/ESA 1995 recommended replacing the traditional fixed base indices with chain indices, preferably Fisher volume and price indices or acceptable alternatives.

Much more complex, both conceptually and practically, international comparisons of volume levels of aggregates were the object of an International Comparison Project (ICP), launched in 1968, after the pioneering research of Colin Clark (1940) and Gilbert and Irving Kravis (1954) at the OEEC. Purchasing power parities, more significant than exchange rates, were calculated. The results of the ICP, however, were not as widely implemented or as widely accepted as national volume measures, something that is unfortunate in a globalized world.

Beyond the progressive completion of its integrated framework, attempts were made to broaden the scope of national accounting by developing semi-integrated additional constructs, such as the satellite accounts whose idea was introduced (by Vanoli) by the end of the 1960 (for example, accounts for social protection, health, education, and environmental protection). In such an approach, the fully integrated system itself becomes the central framework (the expression often used, ‘core accounts’, is ambiguous).

Social accounting matrices (SAMs) were designed by Stone and Alan Brown in 1962, in order to achieve more flexibility than was possible using the usual account presentation. Though the word ‘social’ here means only ‘for the whole economy’, it gave rise to a certain ambiguity. SAMs are sometimes presented as a kind of alternative framework.

In the late 1980s, the Dutch proposed an ambitious ‘system of economy-related statistics’ as a way of organizing a vast array of statistics. A ‘core system’, narrower than the SNA central framework, was linked with ‘system modules’, such as social and environmental modules. This proposal had some similarity with the unsuccessful attempt by Stone, in the first half of the 1970s, to design for the United Nations a system of social and demographic statistics. It echoes the growing importance given to the micro–macro linkages (for example, Richard and Nancy Ruggles, 1986), in parallel with the increased availability of micro-databases.

Concern for statistical coordination had, of course, been present in national accounting from the very beginning.

New challenges since the mid-1970s

The achievements of national accounting, in the face of an enormous development of statistics, have been impressive. However, many countries are still far from fully implementing the international system (for example, few countries prepare integrated balance sheets), and economic and social conditions have changed drastically, especially since the mid-1970s. As a result national accounting, often questioned, sometimes radically, has had to face new challenges.

Since around 1980, after the supply shocks of the 1970s and the decreasing role played by macroeconometric models, national accounting has no longer been supported by the Keynesian paradigm. Some people even think it is obsolete. However, the demand for national accounts continues

to grow, even if it also changes. Predominantly short-term concerns have led to a pressing demand for quarterly accounts, and even sometimes for a monthly GDP, resulting in conflicts between timeliness (early estimates are required) and accuracy. Though more accurate, through successive revisions, annual accounts seem less used and their results are less commented upon.

In the opposite direction, computable general equilibrium models have multiplied since the mid-1970s as a means of studying policies aimed at structural change. Without any concern for the setting up of time series, they are based on the accounts of a single year supplemented, as required, by other data dictated by the models' specificities and purposes. Although they use the somewhat misleading SAM terminology, they actually need national accounts bases.

It remains true, however, that for the study of structural and social policies economists and social researchers, since the last two decades of the 20th century, have generally preferred to make use of micro-simulation models. The role of national accounts data is relatively reduced in this context.

In contrast, a considerable extension of the institutional and political role of national accounting took place during the 1990s, mostly in Europe. Certain aggregates (GDP or GNP) had been used fairly early for administrative purposes such as country contributions to international organizations, eligibility thresholds to preferential World Bank loans, regional allocation of European structural funds, and the ‘Fourth own budgetary resource’ of the Community budget. However, the debate over accession criteria to the European Economic and Monetary Union (the creation of the euro) marked a qualitative jump in the consideration of national accounting by policymakers and public opinion. Most Maastricht criteria were defined in reference to the ESA (ratios of public deficit and public debt to GDP). The ESA became compulsory for member states of the European Union. This marked the culmination of the European statistical strategy adopted in the 1960s. Closely related to the international statistical systems, like the SNA, European statistical tools are in effect very often legally based.

The policy uses of the ESA necessitate effective harmonization of the content of the accounts. A procedure of verification and evaluation of the comparability and representativeness of GDP is established. Full harmonization is, however, difficult. Because conceptual and statistical issues and political considerations intervene, especially in the procedure for identifying excessive deficits, specific cases have to be studied, sometimes through a rather difficult process. Here, and in issues such as the ratio between compulsory levies and GDP, national accounts appear at the forefront of sensitive political concerns. While it clearly shows their importance, this situation may also have less positive aspects for the national accounts. There is the possibility of political pressures, though this is rare; there may be lack of flexibility; official obligations and procedures can be very time-consuming and, as a result of limited human resources, European national accountants may become insufficiently involved in research work.

No similar policy-led process is taking place at the world level. However the need for regulation on a global scale is increasingly felt. Monitoring and intervention aimed at remedying local and regional crises and at preventing systemic crises falls to the International Monetary Fund, in agreement with the principal economic powers. Hence the growing role of the IMF in the supply, by member states, of timely and well-documented harmonized information. In the last decade of the 20th century, the Fund set up a system of standards to guide countries in data dissemination, including meta-information concerning various characteristics of the data. The structuring role of national accounts has been particularly highlighted. The Fund has conducted assessment missions in order to evaluate the quality of countries' national accounts and data systems.

The impressive increase in the demand for and use of national accounts statistics has taken place against

the background of economies which have become much more complex, and hence more difficult to describe and measure, than was the case in the three decades following the Second World War. The number and sophistication of available products have grown; changes in product quality have become more rapid; the share of services, generally more difficult to measure, especially in volume, has increased. The effects of technical change, opening the global economy, the transformation of enterprises and groups, refinements of price policies and consumer behaviour, continuing financial innovations, frequent extension of informal activities, and so on have caused a tendency for economic information systems to maladjust. Hence many controversies arise, notably on price and volume measurements of capital goods – quality change based on performances (Robert Gordon) or on resource cost (the traditional solution championed by Edward Denison) – or measurement of consumption goods and services, where the Boskin Report in the United States (Boskin, Dulberger and Griliches, 1996) argued that the price increase was overestimated.

Significant methodological progress has been in areas such as the measurement of quality change of durable goods based on the change in their performance, the US having taken the lead. However the field is huge, and research is mostly concentrated on information and communication technology products. The measurement of financial and insurance services is in progress. Intangible assets are increasingly investigated. For non-market services, the necessary focusing on direct output–volume measurement instead of the traditional input–volume approach opens, at the start of the 21st century, another wide field of research. It soon appears that the relationship between the concepts of output and outcome must be clarified. On the other hand, some very important issues, like interest and inflation, the treatment of R&D expenditures and the extraction of subsoil resources, have remained outstanding for a long time, defying consensus, though relevant solutions do exist.

After a long emphasis on the relationship between production, income and expenditure, national accounting concerns have in recent decades been extended to the full set of relations between production, income, accumulation and wealth. This raises complex issues concerning the analysis and measurement of capital, particularly intangible assets, and consequently income. By the end of the 20th century business accountants faced similar difficulties with the emerging international accounting standards, moving from historical cost, which national accounting always rejected, to fair value valuation of assets.

Thus, national accounting is fighting for a better coverage of its traditional object at the same time that, at least since the early 1970s, new social concerns have given rise to requests for aggregate monetary indicators synthesizing broader sets of phenomena. There remain things that national accountants cannot do. One is the provision of a welfare indicator, a function that Kuznets assigned to national income, and which gave rise, in the 1940s, to an intense debate involving John Hicks and Paul Samuelson that reached negative conclusions. (William Nordhaus and James Tobin, 1973, later tried to provide such a measure with their ‘measure of economic welfare’.) Another is the measurement of an environmentally adjusted domestic product. The suggestions in this direction included in the 1993 United Nations Handbook, *Integrated Environmental and Economic Accounting*, do not reach a consensus and are not implemented. There was then a move towards wanting a sustainable product or income measure, but this does not make any answer easier, though Hicks's concept of income (the maximum amount that can be consumed in a period while expecting total wealth to be unchanged at the end of it) has increasingly been advocated in recent decades.

Most difficulties relate to the observation and measurement of non-market non-monetary flows and

stocks. Economists propose at least partial measurement solutions, within the framework of standard economic theory, using, for instance, contingent valuation methods (which raises problems of combination with actual exchange values, transfer of results and aggregation), or theoretical constructs with idealized conditions, seeking to justify a possible interpretation of net domestic product in terms of both welfare and sustainability. Other approaches, however, lean towards synthetic indicators combining both monetary and non-monetary variables.

Tensions between social concerns, theoretical issues and observation constraints of actual economies are increasingly at stake.

See Also

- green national accounting
- Kuznets, Simon
- national accounting, history of
- national income
- Stone, John Richard Nicholas
- *Tableau économique*

Bibliography

Aukrust, O. 1994. The Scandinavian contribution to national accounting. In *The Accounts of Nations*, ed. Z. Kenessey. Amsterdam: IOS Press.

Boskin, M.J., Dulberger, E.R. and Griliches, Z. 1996. *Toward a More Accurate Measure of the Cost of Living*. Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index. Washington, DC: Government Printing Office.

Carson, C.S. 1975. The history of the United States National Income and Product Accounts: the development of an analytical tool. *Review of Income and Wealth* 21, 153–81.

Clark, C. 1937. *National Income and Outlay*. London: Macmillan.

Clark, C. 1940. *The Conditions of Economic Progress*. London: Macmillan.

Commission of the European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, World Bank. 1993. *System of National Accounts* 1993. Brussels/Luxembourg, New York, Paris, Washington, DC.

Deane, P. 1955. The implications of early national income estimates for the measurement of long term economic growth in the United Kingdom. *Economic Development and Cultural Change* 4, 3–38.

Eurostat. 1996. *European System of Accounts ESA 1995*. Luxembourg: Eurostat.

Gilbert, M. and Kravis, I.B. 1954. *An International Comparison of National Products and the Purchasing Power of Currencies*. Paris: OEEC.

Grüning, F. 1933. *Der Wirtschaftskreislauf*. München: Beck.

Hill, T.P. 1972. *A System of Integrated Price and Volume Measures (Indices)*. Luxembourg: Statistical Office of the European Communities.

Kenessey, Z., ed. 1994. *The Accounts of Nations*. Amsterdam: IOS Press.

Kuznets, S. 1934. *National Income 1929–1932*. US Senate Document No. 124, 73rd Congress, 2nd session. Washington, DC: Government Printing Office.

Kuznets, S. 1942. U.S. War Production Board, Planning Committee Document No. 151. A memorandum to the Planning Committee from Simon Kuznets on ‘Analysis of the Production program’, dated 12 August.

Leontief, W. 1941. *The Structure of the American Economy 1919–1929: An Empirical Application of Equilibrium Analysis*. Cambridge, MA: Harvard University Press.

Lindahl, E., Dahlgren, E. and Koch, K. 1937. *National Income in Sweden 1861–1930*. London: P.S. King.

Meade, J. and Stone, R. 1941. The construction of tables of national income, expenditure, savings and investment. *Economic Journal* 51, 216–233.

Mitchell, W.C., King, W.I., Macaulay, F.R. and Knauth, O.W. 1921; 1922. *Income in the United States: Its Amount and Distribution, 1909–1919*, Parts I and II. New York: NBER.

Nordhaus, W. and Tobin, J. 1973. Is growth obsolete? In *The Measurement of Economic and Social Performance*, ed. M. Moss. New-York: Columbia University Press for NBER.

OEEC (Organisation for European Economic Co-operation). 1950. *A Simplified System of National Accounts*. Paris: OEEC.

OEEC. 1952. *A Standardised System of National Accounts*. Paris: OEEC.

Popov, P.I., ed. 1926. *Balans narodnogo khoziaistva Soyuza SSSR 1923–1924 goda*. Moskva: Trudi Tsentralnogo Statisticheskogo Upravlenia, Tom XXIX.

Ruggles, R. and Ruggles, N.D. 1986. The integration of macro and micro data for the household sector. *Review of Income and Wealth* 32, 245–76.

Statistical Office of the European Communities. 1970. *European System of Integrated Economic Accounts (ESA)*. Luxembourg: OSCE.

Stone, R. 1947. Definition and measurement of the national income and related totals. Appendix to *Measurement of National Income and the Construction of Social Accounts*. Geneva: United Nations.

Stone, R. 1956. *Quantity and Price Indexes in National Accounts*. Paris: OEEC.

Stone, R. 1986. Nobel Memorial Lecture 1984: the accounts of society. *Journal of Applied Econometrics* 1, 5–28.

Studenski, P. 1958. *The Income of Nations*. New York: New York University Press.

UK Treasury. 1941. *An Analysis of the Sources of War Finance and an Estimate of the National Income and Expenditure in 1938 and 1940*. Cmd. 6261. London: HMSO.

United Nations. 1952. *A System of National Accounts and Supporting Tables*. New York: United Nations.

United Nations. 1968. *A System of National Accounts*. Studies in Methods Serie F n° 2 Rev. 3. New-York: United Nations.

United Nations. 1971. *Basic Principles of the System of Balances of the National Economy*. New York: United Nations.

United Nations. 1979. *Manual on National Accounts at Constant Prices*. New York: United Nations.

United Nations. 1993. *Integrated Environmental and Economic Accounting*. Interim version. New York: United Nations.

US Senate. 1932. S. Res. 220, 72nd Cong., 1st sess., *Congressional Record* 75, 12285.

Vanoli, A. 2005. *A History of National Accounting*. Amsterdam: IOS Press.

How to cite this article

Vanoli, André. "national accounting, history of." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000160> doi:10.1057/9780230226203.1157

The New Palgrave Dictionary of Economics Online

National Bureau of Economic Research

Malcolm Rutherford

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The National Bureau of Economic Research was founded in 1920 and has been regarded as one of the leading research organizations in economics ever since. This entry deals briefly with the founding of the NBER, its early research on national income and business cycles, its later research directions and contributions, and some of the more important changes in organization and direction that have occurred up to 2007.

Keywords

Burns, A. F.; business cycle measurement; Commons, J. R.; Cowles Commission; Fabricant, S.; Feldstein, M.; Friedman, M.; Gay, E.; Kuznets, S.; Meyer, J.; Mitchell, W. C.; Moore, G.; National Bureau of Economic Research; national income; national income accounting; Rorty, M.; Schwartz, A.; Stigler, G.; Stone, N. I.

Article

The National Bureau of Economic Research (NBER) was founded in January 1920, and from the moment of its founding was seen as one of the leading independent research organizations in economics in the world (Fabricant, 1984).

The NBER was established as an independent, non-partisan, research organization focused on empirical investigation. The original research orientation was towards ‘basic’ knowledge of the economy, but was, nevertheless, clearly intended to inform and improve the policymaking process. More recently the research focus has shifted to become more explicitly applied and policy orientated, but empirical work is still central to the bureau's mission. From the first, its Board included a large number of directors from various universities, scientific associations and other organizations. This and the system of manuscript review were designed to ensure the scientific impartiality of its work. These aspects of bureau organization still exist today.

The idea for an independent research bureau in economics sprang from discussions between Malcolm Rorty and N.I. Stone in 1916. Rorty was a statistician with AT&T, Stone an economist working as an arbitrator and economic advisor. Their policy views clashed but they could agree on the need for more reliable information. They involved Wesley Mitchell (Columbia), Edwin Gay (Harvard), and John R. Commons (Wisconsin, and then President of the American Economic Association). The First World War interrupted progress, but the experience of the war made the lack of quantitative information concerning the economy even more apparent, and by the AEA meeting of December 1919 all the necessary elements were in place.

The NBER began with a research agenda directed at the measurement of the size and distribution of national income, and the problem of business cycles. Wesley Mitchell was the first director of research, Edwin Gay the first president, while Rorty and Stone were members of the Board of Directors. Funding was obtained for a small research staff, originally consisting of Mitchell, Willford King, Frederick Macaulay and Oswald Knauth. The major financial contributors were the Commonwealth Fund, followed by the Carnegie Corporation, and, after 1923, the Laura Spelman Rockefeller Memorial Foundation (and its successor organization, the Social Science Division of the Rockefeller Foundation). The NBER also sold subscriptions and engaged in research commissioned by the President's Conference on Unemployment. In 1921 and 1922 the NBER published its first national income estimates: *Income in the United States: Its Amount and Distribution*. This was followed in 1923 by *Business Cycles and Unemployment*, produced by a special staff of the NBER for the President's Conference on Unemployment.

The NBER grew and prospered during the 1920s and early 1930s. The senior research staff were paid a modest stipend by the bureau, but generally held university appointments in the New York area. The bureau also employed research assistants and received funding for research fellowships and for statistical laboratory and library facilities. The research staff came to include Leo Wolman, F.C. Mills, Simon Kuznets, Arthur Burns and Solomon Fabricant. The bureau's research expanded to include Wolman's work on trade union membership, a substantial project on the topic of labour migration (undertaken by Harry Jerome, who was ‘borrowed’ from Wisconsin), F.C. Mills's extensive series of price studies, as well as further work on national income and business cycles. Mitchell produced the first of his projected volumes on business cycles, *Business Cycles: The Problem and its Setting*, in 1927. The bureau also continued its association with the President's Conference on Unemployment by contributing the research for *Recent Economic Changes in the United States* (1929). Kuznets took over the work on national income from King in 1931, and from 1933 he was ‘loaned’ to the Department of Commerce to work on the construction of official national income estimates. The first result of Kuznets's efforts was his report *National Income, 1929–32*, published in 1934.

A financial crisis in 1932 resulted in significant retrenchment at the bureau, which had suffered loss of income due to the Depression and faced uncertainty over the future of Carnegie support. The crisis was overcome thanks to the flexibility shown by Edmund Day of the Social Science Division of the Rockefeller Foundation, but Day expressed concerns with the bureau – its dependence on Rockefeller funding, its domination by a staff drawn heavily from Columbia University, and its lack of interaction with the broader academic community (Rutherford, 2005).

Rockefeller continued to fund the NBER core programmes on national income, business cycles, price and price relationships, the labour market, and savings and capital formation. The bureau also took on a programme of financial research funded by the Association of Reserve City Bankers and headed by

Ralph Young. Mitchell and Burns developed what became known as the ‘NBER method’ of specific and reference cycles to deal with the variations they found between cycles, but the project became ever larger. By the late 1930s the bureau's financial position had recovered and staff numbers again grew substantially, with Milton Friedman joining as an assistant to Kuznets in 1937 (he took over Kuznets's work on *Incomes from Independent Professional Practice*), Moses Abramovitz and Julius Shiskin arriving in 1938, and Geoffery Moore, among numerous others, in 1939.

Day's concerns were not without results. A Universities National Bureau Committee was established in 1935 to examine the potential of NBER–university cooperation. Out of this came the Conference on Income and Wealth (headed by Kuznets) and the Conference on Prices (headed by Mills). The first of these was particularly successful, producing the series Studies in Income and Wealth from 1938 onwards. In addition, Joseph Willits joined the bureau in 1936 as executive director, to deal with administration and fund raising. In 1939, Willits was appointed as Director of the Division of Social Science of the Rockefeller Foundation, and the NBER enjoyed strong support from Rockefeller until Willits left that position in 1954 Rutherford 2005.

Mitchell retired as Director of Research and was succeeded by Arthur Burns in 1945. Kuznets and Burns disagreed over the future direction of the bureau. Kuznets wished to shift the research emphasis to long-run growth, while Burns wished to maintain the focus on business cycles. Kuznets was to pursue his interests through the Conference on Income and Wealth with the financial support of the Social Science Research Council. Burns stayed as Director of Research until appointed to the Council of Economic Advisers in 1953. He was succeeded by Solomon Fabricant. Burns, however, returned to the bureau as President in 1957 and regained much of his previous authority within the organization.

In 1946, Burns and Mitchell published *Measuring Business Cycles*, the result of almost 20 years of effort on the business-cycle project, and the much delayed second volume of the three that were planned. The final, theoretical, volume was never completed. *Measuring Business Cycles* drew sharp criticism from Tjalling Koopmans of the Cowles Commission for its failure to utilize a formal model. Although Koopman's 1947 characterization of the work as ‘measurement without theory’ is a misrepresentation of the Mitchell–Burns programme, there can be no doubt that Burns and others at the bureau were sceptical of what might be achieved by the econometric methods being pioneered at Cowles. Also at this time Burns was engaged in a criticism of Keynesian economics as represented by Alvin Hansen. For Burns, Keynesian theorizing was too speculative and not sufficiently well grounded empirically (Burns, 1946).

The period from the late 1940s through to the mid-1960s was a mixed time for the bureau. Some excellent projects were undertaken. Milton Friedman and Anna Schwartz began their work on US monetary history in 1948, a project that took until 1963 to publish. Friedman did other important work, particularly on consumption theory. Abramovitz worked on inventories and business cycles. George Stigler, who had joined the bureau staff in1943, worked on output and employment trends. Geoffrey Moore refined the system of leading indicators for business cycles, and Morris Copeland developed the analysis of money flows, later to become flow of funds accounts. All the same, the focus of the bureau's efforts had become less sharp; it was conducting much work of lesser value, and running into considerable financial difficulty. Once Willits left Rockefeller, those at Rockefeller were not so sympathetic to the bureau's plight. With the exception of a programme on international economic relations, Rockefeller declined to continue funding the NBER, and in 1958 the bureau turned to the Ford Foundation. Ford established a review committee of Gardiner Ackley, Richard Ruggles, and George Stocking. They criticized the bureau, but recommended that Ford provide funding, which they did. This

allowed the bureau to continue, with relatively few changes until 1965. The research conducted over this period covered a wide range of projects that were loosely grouped into the categories of economic fluctuations, economic growth, wages and other incomes, the economic impact of government and international economic relations.

In 1965, Solomon Fabricant retired as Director of Research and was replaced by Geoffrey Moore, which was seen by many as a decision by the bureau to stay pretty much on its existing track. At the same time, Ford embarked on a major review of the bureau, again with a committee, but this time consisting of Emile Despres, R.A. Gordon, Lawrence Klein, Lloyd Reynolds, Theodore Schultz, George Shultz and James Tobin. This committee was sharply critical of the bureau, its leadership, project selection and research methods. Burns resigned as President and was replaced by John Meyer of Harvard. Meyer took over many of the functions previously held by the Director of Research, created two Vice Presidents of Research, and reorganized the bureau's efforts into specific programmes under their own Directors. Meyer also shifted the focus of the bureau's research into a number of new areas of social policy importance such as urban economics, health, human resources, education, environmental standards, the economics of the family, and crime and punishment. A number of important NBER studies were published during Meyer's term on subjects such as these by Theodore Schultz, Gary Becker, William Landes, Jacob Mincer and Victor Fuchs. Work on cycles was carried on, but no longer using the older NBER methods (Rutherford, 2005).

Meyer left the bureau in 1977 and was replaced as President by Martin Feldstein, also of Harvard. Feldstein has remained as President except for a few years when he was with the Council of Economic Advisors (1982–4), and Eli Shapiro took over. Feldstein brought about further changes at the bureau, doing away with the senior research staff employed directly by the bureau, and changing the bureau into an organization designed to promote and coordinate research being conducted by university-based ‘research associates’ funded largely by National Science Foundation and other research grants. This rearrangement vastly increased the bureau's involvement with the larger academic community.

The focus has remained on empirical and policy-related research. Feldstein added programmes on issues such as aging, and asset pricing, and reinvigorated the NBER programmes on macroeconomics and on taxation. As of 2007, the NBER lists 17 major research programmes each involving 20 or more NBER research associates and each with its own director(s). These include aging, asset pricing, children, corporate finance, education, economic fluctuations and growth, health, industrial organization, international finance, labour, law and economics, monetary economics, productivity and public economics. In addition are smaller working groups working on another 16 topics from behavioural finance to the Chinese economy. The Conference on Income and Wealth also continues. Details of these programmes, those involved, and their publications can be found on the NBER website. The NBER's Research Associates now number about 600, and the NBER working paper series is a major research outlet. Links to the original NBER emphasis on measurement and business cycles are still to be found, however, notably in the NBER's data collection and in the Business Cycle Dating Committee.

See Also

- Burns, Arthur Frank
- business cycle measurement

- Kuznets, Simon
- Mitchell, Wesley Clair
- national accounting, history of

Bibliography

Committee of the President's Conference on Unemployment. 1923. *Business Cycles and Unemployment*. New York: McGraw Hill.

Committee on Recent Economic Changes of the President's Conference on Unemployment. 1929. *Recent Economic Changes in the United States*. New York: McGraw Hill.

Burns, A.F. 1946. Economic research and the Keynesian thinking of our times. *Twenty Sixth Annual Report of the National Bureau of Economic Research*. New York: NBER.

Burns, A.F. and Mitchell, W.C. 1946. *Measuring Business Cycles*. New York: NBER.

Fabricant, S. 1984. Toward a firmer basis of economic policy: the founding of the National Bureau of Economic Research. www.nber.org/nberhistory/sfabricantrev.pdf.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press.

Koopmans, T.C. 1947. Measurement without theory. *Review of Economic Statistics* 29, 161–72.

Kuznets, S. 1934. *National Income, 1929–1932*. New York: NBER.

Kuznets, S. and Friedman, M. 1939. *Incomes from Independent Professional Practice, 1919–1936*. New York: NBER.

Mitchell, W.C. 1927. *Business Cycles: The Problem and its Setting*. New York: NBER.

Mitchell, W.C., King, W.I., Macaulay, F.R. and Knauth, O.W. 1921. *Income in the United States: Its Amount and Distribution, 1909–1919, Part 1, Summary*. New York: NBER.

Mitchell, W.C., King, W.I., Macaulay, F.R. and Knauth, O.W. 1922. *Income in the United States: Its Amount and Distribution, 1909–1919, Part 2, Detailed Report*. New York: NBER.

National Bureau of Economic Research.Online. Available at http://www.nber.org, accessed 4 May 2007.

Rutherford, M. 2005. ‘Who's afraid of Arthur Burns?’ The NBER and the foundations. *Journal of the History of Economic Thought* 27, 109–39.

How to cite this article

Rutherford, Malcolm. "National Bureau of Economic Research." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000147> doi:10.1057/9780230226203.1158

The New Palgrave Dictionary of Economics Online

national income

Thomas K. Rymes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article emphasizes how classical, neoclassical and real Keynesian economic theories are related to accounts of national income and its distribution. The more traditional parts of the analysis focus on rates of growth, capital accumulation and real net rates of return to capital, factoral distributions of income, and capital=theoretic problems in constructing matching national income accounts. More modern neo=Keynesian and monetary approaches are examined to account for theoretical roles played by money and banking in determining output, national income and technical progress. The effects of measures of banking output on modern national income accounts are stressed.

Keywords

capital gains and losses; capital theory; central banking; classical economists; depreciation; distribution of income and wealth; factoral distribution of income; Friedman, M.; Hicks, J. R.; Hulten, C.; human capital; Keynesian revolution; Kuznets, S.; Leontief, W.; Meade, J. E.; measurement of capital; monetary theory; national accounting; national income; national income accounting; neo=Ricardianism; net rates of return; obsolescence; output of banks; quantity theory of money; returns to human capital; Ricardian equivalence theorem; stationary state; Stone, J. R. N.; sustainable consumption; System of National Accounts; technical change; unemployment

Article

Comprehensive systems of national accounts consist today of traditional national income, expenditure and product accounts, input output or production accounts, financial transactions and revaluation accounts (Rymes, 1992) and national balance sheets. While many parts of this modern system are expressed in current and constant prices, national income, its factor and individual income distributions are meaningfully expressed only in current prices. Constant price, or ‘quantity’, indexes are used to measure ‘real’ expenditures over time and across nations, in productivity studies both partial and for all factors again over time and across industries and countries (see Erwin W. Diewert's contributions in ILO, 2004, and IMF, 2004). Indeed, much of modern economic history can now be written in terms of the nominal and real economic accounts over time.

Yet, to date, no one has put together a comprehensive examination of the whole accounting system seen from a particular set or sets of economic theory. Theorists, such as J.R. Hicks, Richard Stone, Wassily Leontief and James Meade, and quantitative economic historians such as Simon Kuznets have made notable contributions to national accounting and have been so recognized with Nobel Prizes. The general lack of emphasis on the connection with economic theory, however, causes the poor student of economics to find the structure of the official accounts a bewildering maze of ‘uses and resources’, which seem more the product of much worthwhile international compromise than the development of the accounts from basic principles of economic theory. Anyone who has tried to teach economics students with the assistance of the 1993 System of National Accounts (SNA 1993, Washington, DC.; Commission of the European Communities; International Monetary Fund; OECD; United Nations; and the World Bank (*sic*)) will not find in all the bureaucratic compromises of admittedly needed reconciliation and international comparisons those flashes of illumination which economic theories can give. A recent OEDC publication (Blades and Lequiller, 2006) further illustrates dangers of the lack of economic theory. It never adequately explains the economic meaning behind consumers ‘real’ expenditures and producers ‘real’ outputs making up GDP, though such knowledge must be held if the reader is to understand the very useful warnings about ‘real shares’ and additivity problems associated with index numbers. Thus it is sad to read one of the best practitioners of national accounting today asserting ‘… the conceptual foundations of the present model of the national accounts are

being progressively undermined by the shifting quicksand of economic theory…' (Ward, 2006, p. 327). Of course, Ward describes other eroding forces, but to give economic theory priority of place in conceptually undermining the accounts seems to me an error resulting from a despairing denigration of economic theory.

I concentrate here on how economic theory contributed to and conditioned national income accounting developments and to some extent how problems in constructing national accounts condition good economic theory. The central theme of this article, then, is the interplay between economic theory and national income accounting. Modern readers, especially students, once they see the interconnection between the accounts and economic theory, should find the national accounts as fascinating and exciting as I do and will each become, I hope, a '… passionate accountant' (Lathen, 1974, p. 183).

Classical and neoclassical national income theories

David Ricardo argued the principal problem of political economy was the determination of the laws governing the distribution of national income among the classes of society (Ricardo, 1971, vol. 1, p. 5). His question was a major concern of classical economic theorists and it has returned to some pre=eminence among economists today (Milanovic, 2005). Consider the following set of extremely simple national income and expenditure accounts set out for a market economy to examine classical economic theory.

Incomes	*Expenditures*
WL	$P_C C$
$RP_K K$	$P_K \Delta K$
$RP_N N$	
$D_N P_N N$	
$D_K P_K K$	

$$D_K P_K K$$

$$Y \equiv E$$

where National Income (Y) is shown as identically equal to National Final Expenditures (E) or Product.

Examining the accounts for one country among many, one must distinguish between National Income and Domestic Product whereas, of course, World Income (WI) and World Expenditure or Product (WP) will be the same. Some economists regard the Domestic Product concept as more useful since it extracts from effects of the international redistribution of returns to capital. (For a contrary opinion, see Beckerman, 1987.) More technical but telling objections can be raised against the Domestic Product concept when it is expressed in constant price terms in a world experiencing technical change in which international trade takes place in intermediate inputs of production.

Why however, does Y identically equal E? If we imagine the accounts were for an even simpler world where there was no capital, then the equality among the circular flows would be clear. Owners of labour would sell their time to producers and the value of their expenditures for the goods produced would cover the cost of the producers. For an extensive discussion of circular flows and the crucial capital=theoretic problems in national accounting, see Hulten (2006).

The notation involves the income of workers (WL), with W the set of money wage rates and L the corresponding set of the working times (hours, days, and so on) offered and demanded by the suppliers and demanders of labour; $RP_N N$ is the net rents earned by the natural agents of production, which, for illustrative purposes, we shall take mainly to be the inalienable and inexhaustible powers of the soil, where R is net rates of return, P_N is prices of the stocks of land so that RP_N is the net rents on the stocks of land (N); and RP_K is rentals earned by the stocks (K) of reproducible capital goods like machines, inventories and buildings. Inanimate things like land and capital goods earn nothing by themselves, and clearly what the classical economists had in mind when then they wrote of the factoral distribution of income was that the net rents on land were garnered by landowners for their husbandry, and the net rents being earned by capital were the net flow of income being earned by the owners of the capital goods, capitalists playing their rentier roles as savers and holders of the stock of capital in the economy. By the ‘factoral distribution of income’ classical economists meant the distribution of income among people, aggregated as the classes of society: labourers, landlords and capitalists. When it is borne in mind that the classical economists also saw labour, land and capital as factors of production, it can be clearly seen that classical theoretical economics was an immensely great scientific undertaking, one which still echoes throughout economics today.

The notation $D_N P_N N$ and $D_K P_K K$ refers to the rates of depletion or exhaustion of natural agents of production, such as the using up of pools of oil, which do not apply to our simple theoretical case of N being Ricardian land. Nor is there any discussion here of the rate of degradation of the environment capital (see Rymes, 1991). Very importantly, $D_K P_K K$ refers to the rates of depreciation or using up of capital in production.

On the Expenditure side of the accounts, $P_C C$ is the values of the final consumption of the society, which, to many economists, is the be all and end all of economics. $P_K \Delta K$ represents the values of the gross capital formation taking place in the society. It is gross in that no allowance is taken of the fact that the new capital goods being produced may or may not be sufficient to replace the wear and tear on existing capital goods. Y and E refer then to Gross National Income and Expenditure respectively.

One of the major theoretical problems in contemporary theory and classical and contemporary national income accounting is the meaning of capital and the conception and measurement of ‘maintaining capital intact’. Even today, despite advances in accounting and economic theory, it is difficult if not almost impossible empirically to measure well the ‘wear and tear’ on capital in modern economic systems. Where depreciation arises from obsolescence, so severe are the problems of measurement that almost all economists today use Gross Domestic Income (Product) or Expenditure as the principal aggregate for economic analysis. National income analysis, then, is greatly hampered by the fact that good estimates of capital consumption and the depletion of natural agents of production, again to say nothing of the degradation of the environment, are generally not available.

If we did have such estimates, the National Accounts just set out could be revised further to appear as

Incomes		*Expenditures*
WL		PcC
$RP_K K$		$P_K(G - D)K = P_K nK$
$RP_N N$		
$Y*N$	$\equiv$	$E*N$

where P_K (G – D)K=P_K nK is net capital formation, with n being the rate of growth so that one would be able to see how important net returns were to capital in net national income, which also in this case is said to measure ‘sustainable’ consumption.

The importance of the capital problem extends to the measurement of labour income as well. Today, wages are paid not so much for the application of pure labour time but for the services of the human capital accumulated by the individuals through expenditures on education, health and even the raising of families. On such capital expenditures, though there is a direct link between the forgoing of present consumption and the accumulation of capital by the individuals, the difficulties of measuring the depreciation on intangible human capital in the so=called knowledge economies are as bad as, if not worse than, those for physical capital. Yet the problem of measuring the returns to human capital gripped the classical economists as well.

One could argue that the consumption of the workers was not final at all, but was perhaps just sufficient to maintain the labour force either at a particular level or at a certain growth rate. Suppose we could extend all of the capital measurement thinking previously outlined to the classical and modern neoclassical treatment of labour. We could write off the consumption of the workers as required inputs into the maintenance of the labour force. Much of PC would vanish along with WL. The above accounts could be then even further dramatically reduced to

Incomes		*Expenditures*
$RP_K K$		PcC^*
$RP_N N$		$P_K(G - D)K = P_K nK$
$Y^{**}N$	$\equiv$	$E^{**}N$

where PcC* is the consumption of the capitalists (and landholders). The extreme classical Ricardian stationary state comes into focus, where the economy is said to have converged to a position where savings and accumulation have been pushed to the point where R, the net rates of return, *are* positive but so low that net savings and the rate of growth of net capital stock and national income, n, would be zero.

Though classical economists were aware that capital accumulation was unlikely to occur in given states of technology, the modern treatment of technical progress is to assume that it serendipitously occurs or, more interestingly, is an endogenous function of the rate of capital accumulation. If, however, technical progress were steadily occurring, then the long=period equilibrium of modern classical analysis and *theory* comes into view. If we ignore land and landholders, and if the consumption of the capitalists were some function of their income and the rate of return so that PC*=c((R),RPK), then national income for steady growth, the modern variant of the Ricardian stationary state, becomes

	Incomes		*Expenditures*
	$RPK - c((R)RPK)$	$\equiv$	$Pn'K$
or	$(1 - c((R)RPK))$	$\equiv$	$Pn'K$
	$s(R)R$	$\equiv$	n'

that is, the economy may be said to have converged to an equilibrium where rates of return to capital exceeds the rate of growth of the income of the economy arising from technical progress, n ′ , if the fraction of returns to capital saved, s, is less than 1. If one assumes that the rate of technical progress is a function of R, then the whole structure of the classical and neoclassical national income accounts can be boiled down to reflect basic theories

$$S(R)R = n'(R(R^{*}(R))$$

where the net rates of return to capital, the intertemporal prices in modern economies, are seen by the simplest accounts to be a function of the rates of saving, or intertemporal choice, and rates of technical change, itself the product of investing and expected rates of return, R*, themselves seen as some function of R. Thus, we see that, when asking questions about the distribution of national income, the national accounts can be set out to illuminate the forces of growth which play vital roles in determining national income. It can also be seen that Ricardo's question about the determinants of the factoral distribution of national income lies at the very heart of modern economic analysis, of both the neoclassical and neo=Ricardian growth varieties (see Barro and Sala=i=Martin, 1995, in particular the chapter on growth accounting; and Pasinetti, 1995). While economic theories may be said to generate the accounts designed to illuminate them, we have seen that they also illuminate the great theoretical difficulties and aggregation problems associated with Professor Hulten's questions about capital theory.

Readers should please note that I am largely by=passing the *severe* capital=theoretic difficulties alluded to by him. One of Hulten's observations that ‘… all aspects of capital ultimately are derived from the decision to defer current consumption in order to enhance or maintain expected future consumption’ (2006, p. 195) means that capital is not a factor of production independently of the ‘willingness to wait’ and that multifactor productivity advance should be conceived as the improvement in the efficiency of working and waiting, n ′ , rather than an improvement in the efficiency of labour and capital. The deep theoretical questions involved in measuring capital, the growth of nations and the aggregation questions may be resolved to some extent by the application of Leontief's disaggregated production and capital accumulation accounts (see Cas and Rymes, 1991; Rymes, 1997).

Keynesian theory

The Keynesian revolution clashed with classical and neoclassical theories and led to some of the modern ‘advances’ in national income accounting. Indeed, some national accountants argue that, partly as a result of Keynes and other theorists such as Jan Tinbergern, modern national accounting started in the 1930s (Bos, 2003; 2006). At the same time economic theory started paying increased attention to institutional forms such as corporations and governments. Under these influences, our simplified national accounts now appear as

Incomes		*Expenditures*
WL		$P_C C$
Ω		$P_K \Delta K$
Y	$\equiv$	E

where the net returns to capital and net rents on natural agents of production are largely replaced by corporate profits, Ω, which generally have measures of depreciation of limited economic meaning, and may or may not well reflect the distribution of interest to bondholders and dividends to shareholders with almost certainly no account being taken of capital gains and losses, and where the switch away from national income to gross national product reflects concern with unemployment rather than the level and the distribution of national income. When the revaluation accounts are added to the standard income accounts, theory again comes to the forefront.

Suppose that modern corporations distribute none of the profits or returns to capital they earn as dividends to their shareholders, ignoring for simplicity the payment of interest to bondholders, but reinvest their profits in the acquisition of capital goods for their firms. The value of the shares held by shareholders (and bought and sold among them) rise along with increases in the corporate stock of capital. It would appear from the national accounts as if the corporations did the saving whereas they may be used to test theories which have the corporations as mere intermediaries, whose investment decisions reflect the wishes of their shareholders.

The neo=Ricardian and Keynesian theories can be put together for the determination of not just the level but also the distribution of national income. If good estimates of the wear and tear on capital are available, one can revert from gross to net income and develop arguments addressed to the question of whether corporate firms and governments can affect the level and the distribution of national income. Here the national accounts can contribute to our knowledge of the extent to which individual households can be said to ‘see through’ corporate firms and governments in such matters as the Ricardian equivalence theorem (see Gillespie, 1980; 1991). To do this, the accounts must be prepared with the various theories of institutional forms in mind; otherwise they may be dismissed with some derision by contemporary theorists (Prescott, 2006).

When the personal distribution of national income is considered, national income accounts must be supplemented by longitudinal surveys of the distribution of income and wealth

among individuals and families, the latter of which can be taken as representing constellations of individuals through time. Here again the theory of why certain families have such time preferences as to permit them to form dynasties requires much work if national income is to be so disaggregated so that those forces playing upon it may be extended to portray and explain individual and dynastic distributions of income and wealth.

Controversies among modern monetary theories and national accounting

Recent developments in monetary theory present great challenges to national accounting. Some monetary theories, those based fundamentally on the quantity theory of money, assert that once=over changes in ‘costless’ fiat money cannot have effects on such real phenomenon as national income, whereas continuous changes in such monies, affecting continuous changes in price indexes, may have rather dramatic effects. Yet, as national balance sheets and wealth accounts show, outside fiat monies are becoming increasingly marginal. How is national income affected by these matters?

National income reflects differences in the underlying classical, neoclassical and Keynesian theories. Keynesian models of unemployment rest upon the empirical and theoretical unimportance of outside or fiat money. Friedman argues, against the Keynesian position, that with real capital gains (losses) accruing to holders of money because of Keynesian disequilibria, real national income will tend to equilibrate at classical economic levels. Thus, if money wage rates and prices are falling because of unemployment, then, according to Friedman, the real income of people, holding given amounts of outside fiat money, will be positive, and will rise faster and faster and become bigger and bigger the more quickly prices fall, thus causing the unemployment to vanish even if there were some adverse effects on expenditures while the deflations were going on (Friedman, 1976, pp. 319–21). As monetary economies are characterized by less and less outside or fiat money, the less and less important is the Friedman counter to Keynes. The question which must be asked is this: is it meaningful to introduce capital gains and losses associated with deflations and inflations and the holding of fiat money into the revaluation accounts associated with national income estimates when, under modern monetary and central banking theory, such holdings, at least in the form of reserves with central banks, are vanishing?

The basic problem with the current national accounts is that we do not have meaningful measures of the output of private banks nor, even more importantly, of the output of central banks. If we applied the current method of imputation for the output of banks to modern central banks, their output would be seen to be zero (Rymes, 2004). Since the banks are the principal producers of transactions services and affect monetary production technologies, it follows that the inability of the national accounts to arrive at satisfactory measures of the output of banks in general means that they cannot measure satisfactorily production in monetary economies (see Fixler and Reinsdorf, 2006). Thus, though one of the central questions dividing Keynesian and neoclassical analyses and the effects of monetary developments on the concepts and measures of national income cannot be currently understood using the current national income accounts, even deeper questions emerge. Does the growth of banks and central bank policies affect capital accumulation, technical progress and national income? We simply do not know now!

Conclusion

The national income accounts have played central roles in the development of economic theory and analysis. Concepts and measures must be improved and developed to reflect better the fact that we live in monetary economies where we do not understand and do not accordingly measure well the outputs of banks and central banks, capital inputs, accumulation and technical progress, all which affect the distribution of national income. Ricardo's question still needs answers. Our current theories and measures of national income need work. Readers and students should therefore realize that there is much exciting and profitable theoretical and empirical study remaining to be done in national income accounting.

Acknowledgements

I am indebted to Dr Harry Postner, with whom I collaborated in preparing a paper given at a session we organized on ‘National accounts and the teaching of economics’ at the 1998 meetings of the International Association for Research in Income and Wealth (IARIW), for his continued criticisms of my work. The most dispiriting finding of that session was how few departments around the world maintained courses in national accounting as part of their curricula. I think the reason is that official national accounting places too little emphasis on connecting with modern economic theories and accordingly lose too much academic contact. I alone am responsible for this basic theme in this article. I also record with much gratitude the extremely critical and penetrating comments on the initial draft made by Professor Duncan McDowall of the history department at Carleton. Professor McDowall took valuable time away from writing a history of the Canadian national accounts to comment on my manuscript. I am also much obliged to Professor Mark Bils whose suggestion for tightening the monetary part of the argument was useful indeed. Finally, I thank Michael James for making my article far more readable than I could.

See Also

- national accounting, history of

Bibliography

Barro, R. and Sala=i=Martin, X. 1995. *Economic Growth*. Toronto: McGraw Hill.

Beckerman, W. 1987. National income. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. Toronto: Macmillan.

Blades, D. and Lequiller, F. 2006. *Understanding National Accounts*. Paris: OECD.

Bos, F. 2003. The national accounts as a tool for analysis and policy; past, present and future. Ph.D. thesis, Twente University.

Bos, F. 2006. The development of the Dutch national accounts as a tool for analysis and policy. *Statistica Neerlandica* 60, 225–58.

Cas, A. and Rymes, T.K. 1991. *On Concepts and Measures of Multifactor Productivity in Canada, 1961–81*. Cambridge: Cambridge University Press, 2006.

Fixler, D. and Reinsdorf, M. 2006. Computing real bank services. A paper prepared for the NBER/CRIW Workshop, 18 July.

Friedman, M. 1976. *Price Theory*. Chicago: Aldine.

Gillespie, I. 1980. *The Redistribution of Income in Canada*. Ottawa: Carleton Library.

Gillespie, I. 1991. *Tax, Borrow and Spend: Financing Federal Spending in Canada 1867–1990*. Ottawa: Carleton University Press.

Hulten, C. 2006. The ‘architecture’ of capital accounting: basic design principles. In *A New Architecture for the U.S. National Accounts*, ed. D.W. Jorgenson, J.S. Landefeld and W.D. Nordhaus. Chicago: University of Chicago Press.

ILO (International Labour Organization). 2004. *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO.

IMF (International Monetary Fund). 2004. *Producer Price Index Manual: Theory and Practice*. Washington, DC: IMF.

Lathen, E. 1974. *Accounting for Murder*. Richmond Hill, ON: Pocket Books.

Milanovic, B. 2005. *Worlds Apart: Measuring International and Global Inequality*. Princeton: Princeton University Press.

Pasinetti, L. 1995. *Structural Change and Economic Growth*. Cambridge: Cambridge University Press.

Prescott, E.C. 2006. The transformation of macroeconomic policy and research. *Journal of Political Economy* 114, 203–35.

Ricardo, D. 1971. On the principles of political economy and taxation. In *The Works and Correspondence of David Ricardo*, vol. 1, ed. P. Sraffa. Cambridge: Cambridge University Press.

Rymes, T.K. 1991. Some theoretical problems in accounting for sustainable consumption. In *Approaches to Environmental Accounting*, ed. A. Franz and C. Stahmer. Heidelberg: Physica.

Rymes, T.K. 1992. National accounting and financial flows. In *The New Palgrave Dictionary of Money and Finance*, vol. 3, ed. P. Newman, M. Milgate and J. Eatwell. London: Macmillan.

Rymes, T.K. 1997. The productivity of working and waiting. In *Capital Controversy: Post=Keynesian Economics And The History Of Economic Thought*, ed. P. Arestis, G. Palma and M. Sawyer. London: Routledge.

Rymes, T.K. 2004. Modern central banks only have *real* effects. In *Central Banking in the Modern World: Alternative Perspectives*, ed. M. Lavoie and M. Seccareccia. Cheltenham: Edward Elgar.

Ward, M. 2006. An intellectual history of national accounting: a review of André Vanoli. *A History of National Accounting. Review of Income and Wealth* 52, 327–40.

How to cite this article

Rymes, Thomas K. "national income." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000007> doi:10.1057/9780230226203.1159

The New Palgrave Dictionary of Economics Online

national system

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balanced growth; Bowen, F.; Carey, H. C.; Eucken, W.; full employment; Hamilton, A.; infant-industry protection; List, F.; national system; protection; social overhead; Sombart, W.; tariffs

Article

The term ‘national system of political economy’ stems from a filiation of American and German ideas that arose in opposition to the universalist character of classical economics and were designed to promote public policies serving the economic development of the nation. The development was visualized as one that would yield a balance of agriculture and industry and make the most of a country's potential economic strength. The term ‘American system’ occurs as early as 1787 in No. 11 of *The Federalist*, where Alexander Hamilton launches this appeal to his readers: ‘Let the thirteen states, bound together in a strict and indissoluble Union, concur in erecting one great American system, superior to the control of all transatlantic force or influence and able to dictate the terms of the connection between the old and the new world.’

Hamilton's more detailed proposals regarding the ways and means to construct the American system can be found in his great state papers, written when he served as Secretary of the Treasury in President Washington's cabinet, and dealing with manufactures, a national bank, and the public debt. With the help of these three instruments he wished to emancipate the new nation from the rural economy of its forefathers, one that Thomas Jefferson, Hamilton's great antagonist, attempted to preserve. Among Hamilton's specific devices to promote industrial development, bounties, or subsidies, stood out. Later writers emphasized protective tariffs rather than bounties.

These writers included Daniel Raymond, a Baltimore attorney, whose *Thoughts on Political Economy* of 1820, while not elaborating the notion of a national system in so many words, made a substantial contribution to the later interpretation of the term by introducing the concept of ‘capacity’ to produce goods, identified by him with national wealth. Raymond placed on government the duty of utilizing and enlarging this capacity by a policy of protection. His plea for protective tariffs was supported both by the infant-industry argument and the employment argument, in conjunction with which Raymond wrote

explicitly of ‘full employment.’

The next step in elaborating the concept of a national system was taken by Frederick List, the German writer and promoter, who in 1827 during his residence in the United States published *Outlines of American Political Economy*. Like Hamilton, List writes of the ‘American system’, which was to realize its potential with the help of tariff protection. This work was written and distributed at the behest of a Pennsylvania manufacturers association whose members clamoured for tariff protection. Composed ostensibly in the form of letters addressed to a leading protectionist, the work appeared serially in the *National Gazette* of Philadelphia and was reprinted by more than 50 other newspapers. When published in pamphlet form, it was distributed in ‘many thousand’ copies, as List later reported. It was sent to the members of Congress and was apparently helpful in securing the adoption of the Tariff Act of 1828.

In an abortive attempt to win a prize, List wrote in French in 1837 an essay on *The Natural System of Political Economy*, which remained, however, unpublished until 1927, when it was printed in French and German. An English translation appeared only in 1983. This work anticipates in a number of respects List's principal work, *National System of Political Economy*, in which the national-system doctrine reached its full flowering. This work was published in German in 1841; an English translation, sponsored by protectionist interests in the United States, appeared in 1856, and another one, published in England, in 1885. The work, while substantial enough in itself, was intended to be the first part of a larger project, which, however, was never completed. Of the English translations, the earlier one omits the preface, while the later one contains extracts from the preface but omits the introductory chapter that provides a summary of the work.

In the *National System*, List finds fault with the classics for a variety of reasons. He takes them to task for having constructed a system of thought that is permeated by individualism and cosmopolitanism but neglects the nation. According to List, the community of nations is not a homogenous group but made up of members that find themselves at different stages of their development. List then goes on to construct a stage theory which visualizes progress from the agricultural stage to one in which agriculture is combined with industry, and to still another one in which agriculture, industry, and trade are joined together. List tends to equate agriculture with poverty and low level of culture, whereas industry and urbanization bring wealth and cultural achievement. The classics, with their homogenized picture of the world which neglected national differences, would tend to perpetuate the underdeveloped status of the United States and continental Europe vis-à-vis the highly developed Britain. According to List, each stage, or each nation at its respective stage, requires a different set of economic doctrines, whereas the classics claimed universal validity for their doctrines.

At heart, List wanted to improve on Providence by turning all people into Englishmen. To allow the underdeveloped countries of his time to participate in the march toward higher stages, attention would have to be paid to their productive capacities. The development and utilization of these was a task that List placed squarely on the national governments. In this connection List called for liberal political institutions, for the construction of what is now known as social overhead, especially in the form of transportation facilities, for balanced growth and for tariff protection for infant industries (not for agricultural products). The free-trade orientation of the classics List was willing to endorse as valid for the future, when all nations had utilized their potential and attained the most progressive stage. Then free trade would be combined with universal peace and a world federation.

There are a number of questions that List left unanswered. To begin with the most often heard objection to the infant-industry argument for protection, what tests are there to identify infant industries and to

mark their eventual attainment of maturity, when protection presumably is to terminate? Moreover, List did not explain how the type of economic warfare that he envisaged would prepare the ground for universal peace. Nor did he show awareness of the likelihood that, once all nations had progressed to what he called the normal state one nation would again get ahead of the others, perhaps for reasons of technological advances, a matter treated with so much insight by Hume in his analysis of the migration of economic opportunities.

List had been a protectionist of sorts already in his young years in his native Germany. His protectionist leanings came to the fore in the United States, where he encountered an even richer potential for economic development and where changing economic conditions were more rapid and conspicuous. Here List's strictures on the classics fell on fertile ground because so many features of their dismal science did not seem to fit into the American environment, especially Malthus's population doctrine and Ricardo's theories of subsistence wages, diminishing returns, and free trade. Thus List's work coalesced with the works of native American critics of the classics, especially of Henry Carey, who developed theories of increasing rather than diminishing returns and of rising wages and profits and declared that each successive addition to the population brings a consumer and a producer. According to Samuelson, Carey's ‘logic was often bad and his prolix style atrocious. But his fundamental empirical inferences seem correct for his time and place’ (p. 1,732). Beginning in 1848, Carey became an ardent exponent of protectionism. By this time List was dead and it is uncertain to what extent, if any, Carey was indebted to List's thought. Neither of the two developed his proposal for tariff protection in isolation but as parts of a wider system of thought, of a theory of economic development in the case of List and of a theory of a harmoniously ordered society in the case of Carey.

Among political leaders in the United States Henry Clay is often mentioned as an architect of the American system, in which the industrial east and the agrarian west were allied in a powerful union. He pleaded for such a system in a famous speech in 1824, in which he supported protective tariffs as instruments of industrial development. Later still, in 1870, Francis Bowen, an early teacher of economics at Harvard, would publish *American Political Economy*, in which he supported tariff protection and which caused him to lose his teaching job in economics, the president easing him into the presumably less controversial field of history.

In Germany, List's ideas had a profound and lasting influence. He promoted the customs union, which by 1844 covered almost all of Germany, and agitated for railroad construction and tariff protection. The very name of economics in Germany, Nationalökonomie, conveys associations with List. Some German interpreters of the history of economics have compared List with Marx. Both had utopian visions of a society to come in the fullness of time. Both made much of a fusion of theory and practice and of economics and politics. Both are linked by their reputation as rebels who opposed the established order. It is an interesting trivium that in 1841 List turned down an offer to serve as the editor of a newspaper that was to be published under the name of *Rheinische Zeitung*, a post that Marx filled the following year.

List's thought has an affinity with the historical schools and institutional economists, who had ideas of their own about the possibility of universally valid economic doctrines. The word ‘system’, cleansed of its protectionist implications, continued to play a key role in the writings of such 20th-century German economists as Walter Eucken and Werner Sombart. An equally faint echo of the Hamiltonian idea can be discerned in the current usage of the word in conjunction with the study of comparative economic *systems*.

See Also

- comparative advantage
- Corn Laws, free trade and protectionism
- growth and international trade
- infant-industry protection

Bibliography

Bowen, F. 1870. *American Political Economy*. New York: Scribner.

Carey, H. 1858–59. *Principles of Social Science*. 3 vols, Philadelphia: Lipincott.

Conkin, P.K. 1980. *Prophets of Prosperity: America's First Political Economists*. Bloomington: Indiana University Press.

Dorfman, J. 1946. *The Economic Mind in American Civilization*, Vols 1–2. New York: Viking Press.

Hamilton, A. 1934. *Papers on Public Credit, Commerce and Finance*. Ed. S. McKee, New York: Columbia University Press.

Henderson, W.O. 1983. *Friedrich List: Economist and Visionary 1789–1846*. London: Cass.

Hirst, M.E. 1909. *Life of Friedrich List and Selections from His Writings*. London: Smith, Elder.

List, F. 1827. *Outlines of American Political Economy*. Reprinted in *The Life of Friedrich List and Selection from his Writings*, ed. M.E. Hirst. London: Smith, Elder & Co., 1909.

List, F. 1837. *The Natural System of Political Economy*. Trans. and ed. W.O. Henderson, London: Cass, 1983.

List, F. 1956. *National System of Political Economy*. Philadelphia: Lippincott.

Samuelson, P.A. 1960. American economics. In *Postwar Economic Trends in the United States*, ed. R.E. Freeman, New York: Harper. Reprinted in *P.A. Samuelson, Collected Scientific Papers*, Vol. 2, ed. J.E. Stiglitz, Cambridge, Mass.: MIT Press, 1966, 1732–47.

Spiegel, H.W. 1960. *The Rise of American Economic Thought*. Philadelphia: Chilton.

Spiegel, H.W. 1983. *The Growth of Economic Thought*. Revised and expanded edn,Durham, North

Carolina: Duke University Press.

How to cite this article

Spiegel, Henry W. "national system." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000010> doi:10.1057/9780230226203.1160

The New Palgrave Dictionary of Economics Online

natural and warranted rates of growth

J.A. Kregel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The concepts of the natural and warranted rates of growth of national income, associated with the work of R.F. Harrod and E.D. Domar, were first developed in the 1930s and 1940s as part of the rethinking of the theory of economic fluctuations generated by Keynes's *General Theory*. Somewhat paradoxically, they formed an initial impetus for the theories for long-run steady growth elaborated in the 1950s and 1960s.

In the early 1930s Harrod criticized the static nature of economic analysis, suggesting that it be supplemented by a ‘dynamic’ theory: static theory determined the levels of variables, dynamic theory should explain the ‘rates of change’ of the variables taken at a point in time. Harrod's first attempt at dynamic theory, *The Trade Cycle* (1936), appeared almost simultaneously with Keynes's book, which Harrod considered limited to statics, even though it argued that the system could achieve equilibrium at less than full employment, because it dealt with the equilibrium *levels* of output and employment. After a lengthy correspondence with Keynes (cf. Keynes, 1973, pp. 151ff), Harrod published a new version of his theory, ‘An Essay on Dynamic Theory’, (1939) in which he formulated a ‘dynamic equilibrium’ for income, Y, defined as the ‘warranted rate of growth’ $g_w=(\mathrm{d}Y/\mathrm{d}t)/Y$, to complement Keynes's static equilibrium. Due to the outbreak of war the theory did not attract attention until he presented it in a series of popular lectures (Harrod, 1948) after the war.

In Keynes's theory any level of output and employment, including full employment as a special case, was a potential equilibrium; the actual equilibrium was determined by the point of effective demand given the general state of expectations expressed in the propensity to consume, the marginal efficiency of capital and liquidity preference. Harrod was thus led to analyse a ‘dynamised version of Keynes' … effective demand’ (Harrod, 1959), defined as the rate of growth produced by the rate of investment chosen by entrepreneurs which is warranted in the sense of maintaining a rate of expansion of effective demand which is consistent with entrepreneurial expectation and with individuals' autonomous decisions to save. The level of income, Y_0, prevailing at any point in time in the actual development of the economy will be determined by the entrepreneurs' expectations of the rate of growth of income $(\mathrm{d}Y/\mathrm{d}t)/Y_0$. On the basis of the expected $\mathrm{d}Y/\mathrm{d}t$ they will decide the investment necessary to satisfy this expected

expansion in demand. This decision is made on the basis of the ‘capital coefficient’ (which Harrod called C, but is now generally written as v), ($I_0=v(\mathrm{d}Y/\mathrm{d}t)$, defined as the total money expenditure that must be made on new investment projects to create an additional £ of output. The public's decisions to spend and save expressed as $S=sY_0$ will then determine the actual increase in income via the multiplier $(\mathrm{d}I/\mathrm{d}t)/s=\mathrm{d}Y/\mathrm{d}t$. Entrepreneurs' expectations will only be confirmed if $v(\mathrm{d}Y/\mathrm{d}t)=sY_0$ which when rearranged produces Harrod's famous growth equation $g_w=(\mathrm{d}Y/\mathrm{d}t)/Y_0=s/v$, with $S/Y_0=I/Y_0$ which is Keynes's equilibrium. The rate of expansion of income is thus warranted and since entrepreneurs' expectations have been confirmed they are preseumed to expect income to continue to expand at that rate. Thus, given Y_0 and s there is a set of expectations which produces a dynamic equilibrium rate which will describe an expansion of income through time of $Y=Y_t \exp(g_w t)$.

For Harrod, the analytical importance of his dynamics was to be found in the proposition that while in static analysis any departure from equilibrium produced centripetal forces driving the variable back to its equilibrium value, in dynamic analysis any movement away from equilibrium (in this case the warranted rate of growth of income) would set up centrifugal forces which would move the system further away from its equilibrium position. For example, if income were growing at the warranted rate and investment rose above the warranted rate, $I_t>I_0 \exp(g_w t)$, income would expand at a higher rate, inventories would be drawn down and additional investment would be required to restore them to normal; the expectations which produced the warranted rate would be revised upwards as investment would appear insufficient relative to the expansion in sales, leading to further increases which would eventually surpass available labour and resources. Thus, instead of returning to the equilibrium rate, g_w, an inflationary boom in which expectations would eventually be disappointed by shortages of supply, leads to a collapse of investment and expectations. Since the dynamic equilibrium is unstable, Harrod thus concludes that the warranted rate of growth is inherently unstable.

Just as in Keynes's theory, there is no reason for the warranted rate to be associated with full employment, nor is there any reason for a disturbance of the system from a dynamic equilibrium to lead to a full employment rate. Disturbances will in general lead to a series of erratic booms and slumps of variable duration with respect to the warranted rate. The full employment rate of growth does however play a role in this cyclical process by setting a limit beyond which it is impossible for the economy permanently to grow, either in equilibrium or disequilibrium. If the rate of growth of potentially employable labour, given by the rate of population growth, is $n=(\mathrm{d}N/\mathrm{d}t)/N$, the full employment rate of growth representing the maximum sustainable growth rate would be $g=n=s/v$ unless technical progress expanded output per man employed. When available technical progress is used to increase labour productivity by $\tau = (\mathrm{d}(Y/N)\mathrm{d}t)/(Y/N)$ the maximum sustainable rate, which Harrod called the ‘natural’ rate, would be $g = n + \tau$ The natural rate will only be an equilibrium position, i.e. a warranted rate, if households save the required proportion of income s_r which given the optimal introduction of new production techniques producing v_r, is required to produce $g_n = s_r / v_r = n + \tau$. Since there is no economic mechanism that links s and v to n and τ the natural rate is unstable, but for different reasons than if it happened by chance to be a warranted rate.

Thus, for any actual state of the economy there will be a value for $g_w \leq g_n$ which is given by the values of Y_0, v and s determined by the past history of the economy. There can, of course, be only one value for g_w since there cannot be more than one value of Y_0, s or v for any given point in time. If the economy

grows at some other rate, say g_a, then Y will not expand along the warranted path $Y_t=Y_0\exp(g_w t)$, so that the rate which would be required to produce warranted growth from any subsequent point in time, t, would depend on the actual values of Y_t, s and v.

For example, if $g_a = s_a / v > g_w = s / v$, then $s_a > s$ and investment will continue to increase g_a until the upper limit of g_n is surpassed. This may be conceivable, for example in the period after a deep slump, but physical bottlenecks and increases in money wages due to labour market shortages will eventually lead to inflationary boom and a subsequent collapse back into a slump which will cause incomes and investment to fall, causing s_a to fall. At any time in this process it would be possible to calculate on the basis of the level of income, Y_t, and associated s and v, the rate of investment which, if adopted would produce warranted equilibrium growth from that time onwards. Although it is highly unlikely that the economy would adopt this rate, it serves as a benchmark with which to compare the actual behaviour of the economy and thus to predict the direction of its subsequent cyclical movements.

There will thus be a different, but unique, value of g_w for every actual position of the system as it develops through time. Only if g_w is in fact attained will the economy exhibit stable, non-cyclical growth, while departures from the rate will not set up self-correcting movements to instantly restore it.

These two aspects of Harrod's theory have caused much misunderstanding. The fact that there is only one ‘unique’ or ‘knife-edge’ equilibrium growth path for any given t and condition of the system has led some economists to consider this as the main cause of instability. Yet Harrod himself considered ‘instability’ to be an inherent property of the general concept of dynamic equilibrium as represented by the warranted growth rate. Since there would be only one warranted rate for any given condition of the economy it could be used to explain the cyclical behaviour of the economy if g_a diverged from g_w. But in Harrod's theory there would be a new warranted rate for every new combination of Y, s and v thrown up by the actual growth of the economy; g_w was only unique because each point in time was characterized by unique conditions. The role of the instability property of the warranted rate, given the natural rate, was to explain how the system would move when it was not growing at its dynamic equilibrium rate.

Domar (1946, 1947), writing after the publication of the *General Theory*, reacted to a specific problem in Keynes's theory, pointing out that the very investment expenditure that provides the demand for the output of existing productive capacity implies increased productive capacity in future periods. Investment as a means of increasing aggregate demand is thus a ‘mixed blessing’, for if the investment sufficient to prevent unemployment today creates excess capacity tomorrow then even more investment will be required tomorrow. Long-run unemployment could be avoided only by increasing investment at an increasing rate. To analyse this problem it was inevitable that Domar recast Keynes's analysis in terms of rates of change.

Domar approached the problem by separating the influence of investment on aggregate demand and on productive capacity or supply. Keynes had already provided the analysis of demand in terms of the multiplier (k=1/s) giving the expansion in demand resulting from increasing investment as $\mathrm{d}Y_d/\mathrm{d}t=k(\mathrm{d}I/\mathrm{d}t)$. On the supply side, however, since all of net investment, and not only the increase, expands productive capacity Domar amends Keynes's approach and considers the fraction of the labour force employed as a function of the ratio of income to *potential* productive capacity rather than as a simple function of income. Defining α as the net value added produced by a £ of net investment, potential

productive capacity will then increase by αI where I is the aggregate cost of new investment projects. On the micro level, however, some new capacity will be competing with older capacity, and since some investment projects will be carried out on the basis of expectations which will not be realized, Domar defines σ as the ‘potential social average productivity of investment’ for the economy. The divergence between α and σ (as well as the assumption that $\alpha < \sigma$) thus represents errors in investment decisions, investment outpacing the growth in the labour force or investments incorporating inappropriate technology. The supply-side effect is thus $dY_s / dt = \sigma I$. The answer to Domar's question of whether there is a constant rate of growth of investment at which the demand will rise sufficiently rapidly to offset the effect of investment on supply is thus found where $dY_d / dt = dY_s / dt$ or where $k(dI / dt) = \sigma I$. This equality can be rewritten as $(dI / dt) / I = \sigma / k$ which Domar calls the ‘required’ rate of growth of investment.

Domar's assumption that unemployment is determined by the relation of income to potential capacity means that the ‘required’ rate implies full capacity utilization and thus full employment. The failure of the economy to grow at this rate implies excess capacity. If productive potential arising from net investment σI is defined as P, $\sigma = (dP / dt) / I$, then a coefficient of utilization determined by the relative expansion of demand and capacity can be defined as, $\theta = (dY_d / dt) / (dP / dt)$ Since $dY_d / dt = k(dI / dt)$ and $dP / dt = \sigma I$, θ can be written as $(dI / dt) / I \cdot k / \sigma$ assuming that $\alpha = \sigma$ If investment is expanding at the required rate $(dI / dt) / I = \sigma / k$, $dY_d / dt = dP / dt$ and $\theta = 100$ per cent capacity utilization. Domar's required rate is thus equivalent to Harrod's natural rate of growth (s_r/v_r) when $\alpha = \sigma$ since k=1/s and $\sigma = (dY / dt) / I = 1 / v_r$, $\sigma / k = s / v_r$.

Domar's analysis of divergence of the actual growth rate from the ‘required’ rate also produces an analysis of instability, for when (dI/dt)/I is below k/σ the required rate, dY_d/dt is less than dP/dt, so part $(1 - \theta)$ of new productive potential is unused. This excess capacity thus implies the existence of unemployment. A higher rate of growth of investment would be required to eliminate the excess capacity and unemployment, but since current productive capacity is already excess to needs, entrepreneurs are more likely to try to reduce than to increase their desired capacity by lowering (dI/dt)/I, which will increase rather than decrease both unemployment and excess capacity, producing a slump. Thus, in difference from Harrod's analysis, the natural rate is a unique equilibrium or ‘knife edge’ rate as well as being unstable. For Domar instability is not linked to the conceptual definition of dynamic equilibrium by means of a warranted rate, but rather to the ‘paradox’ that is the dynamic equivalent to the Keynesian paradox of saving: given s, the elimination of excess capacity, whether it is caused by the effects of investment on the expansion of demand or productive capacity, requires more capital to be built, while a shortage of productive capacity requires a reduction in the rate of growth of investment. This result is parallel to Harrod's statement to the effect that a general glut of commodities is due to entrepreneurs producing too little rather than too much.

While both Harrod and Domar sought to use the concepts of warranted and natural or required rates as an aid to understanding the cyclical implications of Keynes's analysis, and despite the differences in their approach, their work served to form the basis of what came to be known as the ‘Harrod–Domar’ theory of steady growth. By interpreting the variables s and v as being given exogenously the theory produced what Kaldor (1951) called ‘Harrod's problem’, or as Joan Robinson (1965, p. 52) put it:

> Given s,… and v,… g is determined. There is only one value of g which (provided it does

not exceed *n*) is not impossible. The uniqueness of *g*, not any question about the stability of the corresponding growth path, created the problem of the ‘knife edge’.

This ‘problem’ was ‘resolved’ by introducing differential savings propensities from wages and profits to make *s* a variable determined by the distribution of income, which would allow multiple long-period unemployment growth equilibria, as in the post-Keynesian theories of growth and distribution. Alternatively (cf. e.g. Solow, 1970, ch. 2), if movements in relative prices of capital and labour services are allowed to produce substitution of capital for labour, as in an aggregate production function, then *v* would become variable over time and lead to the full employment of both factors, despite Domar's (1952, pp. 23–6) explicit warning that the introduction of a Cobb–Douglas production function to solve this problem would lead directly to this traditional pre-Keynesian result.
These two conflicting interpretations of the applicability of Keynes's unemployment equilibrium in the long period, soon enlarged to include the wider question of capital theory, created a debate in which steady state theories overwhelmed the interests of both Harrod and Domar in the implications of Keynes's theory for the problem of economic fluctuations and dynamics.

See Also

- aggregate demand theory

Bibliography

Domar, E.D. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14(April), 137–47. Reprinted in Domar (1957), 70–82.

Domar, E.D. 1947. Expansion and employment. *American Economic Review* 37(March), 34–55. Reprinted in Domar (1957).

Domar, E.D. 1952. Economic growth: an econometric approach. *American Economic Review, Papers and Proceedings* 42(May), 479–95. Reprinted as ‘A theoretical analysis of economic growth’, in Domar (1957).

Domar, E.D. 1957. *Essays in the Theory of Economic Growth*. New York and Oxford: Oxford University Press.

Harrod, R.F. 1936. *The Trade Cycle*. Oxford: Clarendon Press.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49(March), 14–33.

Harrod, R.F. 1948. *Towards a Dynamic Economics*. London: Macmillan.

Harrod, R.F. 1959. Domar and dynamic economics. *Economic Journal* 69(September), 451–64.

Kaldor, N. 1951. Mr Hicks on the trade cycle. *Economic Journal* 61(December), 833–47.

Keynes, J.M. 1973. *The Collected Writings of J.M. Keynes*, Vol. 14: *The General Theory and After* – Part II, *Defence and Development*. Ed. D. Moggridge, London: Macmillan.

Robinson, J. 1965. Harrod's knife edge. In J. Robinson, *Collected Economic Papers*, Vol. 3, Oxford: Basil Blackwell.

Solow, R.M. 1970. *Growth Theory: an Exposition*. Oxford: Clarendon Press.

How to cite this article

Kregel, J. A. "natural and warranted rates of growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000015> doi:10.1057/9780230226203.1161

The New Palgrave Dictionary of Economics Online

natural experiments and quasi-natural experiments

J. DiNardo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Natural experiments or quasi-natural experiments in economics are serendipitous situations in which persons are assigned randomly to a treatment (or multiple treatments) and a control group, and outcomes are analysed for the purposes of putting a hypothesis to a severe test; they are also serendipitous situations where assignment to treatment ‘approximates’ randomized design or a well-controlled experiment.

Keywords

experiment; natural experiments; quasi-natural experiments; randomization; regression discontinuity design; returns to schooling; social experiments; treatment effect

Article

The term ‘natural experiment’ has been used in many, often, contradictory, ways. It is not unfair to say that the term is frequently employed to describe situations that are neither ‘natural’ nor ‘experiments’ or situations which are ‘natural, but not experiments’ or vice versa.

It will serve the interests of clarity to initially direct most of our attention to the second term – experiment. A useful, albeit philosophically charged definition of an experiment ‘is a set of actions and observations, performed in the context of solving a particular problem or question, to support or falsify a hypothesis or research concerning phenomena’ (Wikipedia, 2006).

With such a broad definition in hand, it may not be surprising to observe a wide range of views among economists about whether or not they perform experiments. Vernon Smith, for example, in experimental methods in economics, begins with the premise that ‘historically, the method and subject matter of economics have *presupposed* that it was a *non–experimental … science more like astronomy or meteorology than physics or chemistry*’ (emphasis added). As he makes clear, his observation implies that *today*, economics is an experimental science. Bastable's article on the same subject in the first edition of *The New Palgrave* overlaps only superficially with Smith's and divides experiments along the lines suggested by Bacon: *experimenta lucifera*, in which ‘theoretical’ concerns dominate, and *experimenta fructifera*, which concern themselves with ‘practical’ matters. In sharp contrast to Smith, Bastable concludes that *experimenta lucifera* are ‘a very slight resource’ (1987, p. 240) in economics.

These two views of experiment, however, do not seem helpful in understanding the controversy regarding natural experiments. ‘Experiment’ in our context is merely the notion of putting one's view to the most ‘severe’ test possible. A good summary of the the spirit of experiment (natural or otherwise) comes from the American philosopher Charles Sanders Peirce (and see Mayo, 1996 for a nice exposition of this and related points):

> [After posing a question or theory], the next business in order is to commence deducing from it whatever experimental predictions are extremest and most unlikely … in order to subject them to the *test of experiment*.
> The process of testing it will consist, not in examining the facts, in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis as would be capable of direct verification, especially those consequences which would be very unlikely or surprising in case the hypothesis were not true.
> When the hypothesis has sustained a testing as severe as the present state of our knowledge … renders imperative, it

> will be admitted provisionally … subject of course to reconsideration. (Peirce, 1958, 7.182 (emphasis added) and 7.231 as cited in Mayo, 1996)

The philosophy of experimentation in natural science

In the emergence of modern natural science during the 16th century, experiments represented an important break with a long historical tradition in which observation of phenomenon was used *in* theories as a way to justify or support a priori reasoning. In Drake's (1981) view: ‘The Aristotelian principle of appealing to experience had degenerated among philosophers into dependence on reasoning supported by casual examples among philosophers and the refutation of opponents by pointing to apparent exceptions not carefully examined.’ In the useful historical account provided by Shadish, Cook, and Campbell (2002) it is suggested that this ‘break’ was twofold: first, experiments were frequently employed to correct or refute theories. This naturally led to conflict with political and religious authorities: Galileo Galilei's conflict with the Church and his fate at the hands of the Inquisition is among the best-known examples of this conflict. Second, experiments increasingly involved ‘manipulation’ to learn about ‘causes’. Passive observation was not sufficient. As Hacking (1983, p. 149) says of early experimenter Sir Francis Bacon: ‘He taught that not only must we observe nature in the raw, but that we must also “twist the lion's tale”, that is, manipulate our world in order to learn its secrets.’

Indeed, at some level in the natural sciences there has been comparatively little debate about the centrality of experiment – ironically, it has typically been only philosophers of science who have downplayed the importance of experiment. Hacking (1983) makes a strong case that philosophers typically have exhibited a remarkably high degree of bias in minimizing their importance in favour of ‘theory’. Until the 19th century, the term experiment was typically reserved for studies in the natural sciences.

In the low sciences such as economics and medicine, the role of experiment is been the subject of extensive debate, much tied up with the debate on whether all the types of experiments possible in real science are possible in economics as well as with debates about the many meanings of the word ‘cause’.

A key distinction between much real science and economics involves the centrality of ‘randomization’. No randomization is required, for example, to study whether certain actions will produce nuclear fission, since ‘control’ is possible: if a set of procedures applied to a piece of plutonium – under certain pre-specified experimental conditions – regularly produces nuclear fission, as long as agreement exists on the pre-specified conditions and on what constitutes plutonium, and so on, it is possible to put the implied propositions to the type of severe test that would gain widespread assent – all without randomization. Put in a different way, randomization is required only when it is difficult to put a proposition to a severe test without it.

A related issue is whether a study of ‘causes’ requires some notion of ‘manipulation’. Most definitions of ‘cause’ in social science involve some notion of ‘manipulation’ (Heckman, 2005) – Bacon's ‘twisting of the tail’, so to speak. In physics, by way of contrast, some important ‘causes’ do not involve manipulation per se. One might argue that Newton's law of gravitation was an example of a mere empirical regularity that became a ‘cause’. Indeed, when proposed by Newton, Leibnitz objected to this new ‘law’: in the prevailing intellectual and scientific climate where the world was understood in terms of ‘mechanical pushes and pulls’, this new law seemed to require the invocation of ‘occult powers’ (Hacking, 1983). (There is an element of irony in Leibnitz's objection. Leibnitz is believed by some to be the object of Voltaire's satire as the character Dr Pangloss in *Candide* of whom it is said that he ‘proved admirably that there is no effect without a cause … in this the best of all possible worlds’ – a very different notion of causation! Voltaire, 1759, ch. 1.)

In this article, we take the view that, even if manipulation were not necessary to *define* causality, manipulation is central to whether it is possible to discuss the idea intelligibly in social sciences and whether some kind of ‘severe test’ is possible (DiNardo, 2007). Some philosophers have sought to *define* science around issues related to ‘control’, arguing that the phenomena economists try to investigate are impossible to study scientifically at all. Philosophers have articulated numerous reasons for the difference between social and natural science. A few examples may be helpful: Nelson (1990, pp. 102–6) argues, for example, that the objects of enquiry by the economist do not constitute ‘a natural kind’. Put very crudely, the issue is the extent to which all the phenomena that we lump into the category ‘commodity’, for example, can be refined to some essence that is sufficiently ‘similar’ so that a scientific theory about commodities is possible in the same way as a ‘body’ is in Newtonian mechanics. This is often discussed as the issue of whether the relevant taxonomy results in ‘carving nature at the joints’. Hacking (2000) introduces the notions of ‘indifferent kinds’ – the objects in the physical science – atoms, quarks, and so on with ‘interactive’ kinds – the objects of study in medicine or the social sciences. We might interact with plutonium or bacteria, but neither the plutonium nor the bacteria are aware of how we are classifying them or what we are doing to them. This can be contrasted with ‘interactive kinds’ that are aware and for which ‘looping’ is possible. For example, mental retardation might lead to segregation of those so designated. This segregation might lead to new behaviours which then might not fall under the old label, and so on. Consequently, investigation of such phenomena might be likened to ‘trying to hit a moving target’. Searle (1995) on the other hand, notes that the objects of interest in social science while epistemologically objective, are ontologically subjective. While the loss of 100 dollars may be very ‘real’ to someone, the notion of money requires

groups of individual to accept money as a medium of exchange. Again the existence of atoms does not require us to recognize their existence.

Randomization: an attempt to evade the problems of imperfect ‘control’

If one accepts the centrality of manipulation (or something like it), it will not be surprising that the application of principles of experimentation to humans who have free will, make choices, and so on entails a host of issues that, inter alia, sharply constrain what might be reasonable to expect of experiments, natural, or otherwise.

If it is not possible, desirable, or ethical to ‘control’ humans or their ‘environment’ as it sometimes is in the natural sciences, is it possible to learn anything at all from experiment broadly construed? *Randomization* in experiments developed in part to try to evade the usual problems of isolating the role of the single phenomenon in situations. In the 19th century, it was discovered that by the use of ‘artificial randomizers’ (such as a coin toss) it was possible, in principle, to create two groups of individuals which were the same ‘on average’ apart from a single ‘treatment’ (cause) which was under (at least partial) control of the experimenter. Hacking (1988, p. 427) has observed that their use began primarily in contexts ‘marked by complete ignorance’: the economist F. Y. Edgeworth was early to apply the mathematical logic of both Bayesian and ‘classical’ statistics to a randomized trial of the existence of ‘telepathy’. Although economists played an important role in the development of randomization, economists as a whole were quite slow to embrace the new tools. In an echo of debates that faced natural sciences in the 1600s, this was due in part ‘because the theory [of economics] was not in doubt, applied workers sought neither to verify nor to disprove’ (Morgan, 1987, pp. 171–2).

Over time, the term ‘experiment’ evolved to include both experiments of the ‘hard sciences’ where a measure of control was possible as well as situations in which artificial randomizers were used to assign individuals (or plots of land, and so on) to different ‘treatments’. A key role was played by R. A. Fisher (1935) and his seminal *Design of Experiments* as well subsequent publications which discussed the theory and practice of using artificial randomizers to learn about causes.

There are at least two key limitations of randomized experiments relative to experiments where ‘scientific’ control is possible:

- Without real control, one only has a weak understanding of the ‘cause’ in question. For instance, one can do a randomized controlled trial of the effect of aspirin on heart failure while understanding nothing of the mechanism by which aspirin affects the outcome. Moreover, it is clear that the experiment is ‘context specific’. One's generalization about atoms in a laboratory often extends to atoms in other contexts in a way not possible in social science.
- Any single experiment – even under the ideal situation – does not always reveal the true answer. In the logic of randomized design, the usual inference procedure is merely one that *would* give the right answer on average *if* the experiment were repeated. At best, the true answer is just a ‘long-run tendency’ in repeated identical experiments.

Social experiments: why not do a ‘real’ randomized trial?

Even without these limitations, there is a long list of reasons why economists frequently have little interest in randomized trials. The most important reason is that many of the real randomized experiments (often called ‘Social experiments’) of which one could conceive (or have been implemented), are immoral or unethical. At a most basic level, the decision as to who ‘performs an experiment’ and who ‘decides’ or is recruited to be experimented upon often reflects deep-seated social injustice. Even Brandeisian (see below) experiments can take on a sinister cast – state governments surely do not consider the interests of all their citizens equally. Indeed, historically the conduct of experiments on persons has told us as much or more about the structure of society than anything else: one well-known example is the series of ‘experiments’ conducted by the US Public Health Service from 1932 to 1972 on about 400 poor black men who had advanced syphilis. One aim of the experiment was to determine the effect of untreated syphilis. To this end, the medical doctors misrepresented themselves to the subjects (the sons and grandsons of slaves), claiming to provide free medical care. For example, when penicillin became the standard of care, the subjects were deliberately not provided with the medication: rather, the doctors were content to observe the horrific progress of the disease as some went blind or insane.

Another set of reasons is practical – experiments are costly to administer. Another reason is attrition: often people drop out of such experiments (often in non-random ways), greatly complicating the problem of inference. A distinct, although sometimes related, reason is that the results of social experiments involving randomization are sometimes difficult to interpret. One often cited reason is that those recruited to participate in such experiments may be different from those for whom the policy is ultimately intended. In even the simplest experiments, ‘compliance’ is imperfect. Not everyone assigned to a treatment takes it up – indeed, it is often the case that analysis is made on an ‘intent to treat’ basis. That is, those ‘assigned’ to treatment are compared to those assigned to the control whether or not those assigned to treatment actually ‘took’ the treatment. Another often cited reason is that what is likely when a social experiment is conducted with a small number of persons might be very different when applied to much larger numbers of persons. Persons, unlike atomic particles, enjoy free will. In the world of persons, the ‘experiment’ does not necessarily stop after the

experimenters have made their observations. For example, even in the context of a true randomized experiment, those denied treatment often have the opportunity to find it elsewhere (see Heckman and Smith, 1995, with references, for one discussion of the merits of randomized trials in the social science).

Types of natural experiments

Thus far we have seen that the word ‘experiment’ can be used in two very different senses: one to denote situations where real ‘control’ is possible and second involving artificial randomizers. As a consequence, the term ‘natural experiment’ has been used in very different senses. I now turn to the origins of the term and the different ways the term has been used, although we focus on natural experiments most frequently arising in economics.

Natural experiments in natural science

An early use of the term ‘natural experiment’ in English describes an investigation into the functioning of ‘nature’. The term comes from a translation *Saggi di naturali esperienze fatte nell'Accademia del Cimento* published in Italian in 1667 which appeared in an English translation by Richard Waller in 1684 as *Essayes of natural experiments made in the Academie del Cimento* (Waller, 1684). The short-lived Accademia del Cimento was founded in Florence in 1657 by the Medici brothers, Prince Leopold and Grand Duke Ferdinand II, and the *Saggi* record a small subset of the large number of experiments by the Cimento that involved such issues as ‘smells do not traverse Glass’, and ‘the failure to confirm Existence of Atoms of cold’ (1684, p. xx). Although the experiments of the Academy included trials involving humans, they did not involve randomization. Indeed, the legacy of these investigations into humans is more relevant to the study of 16th-century culture and authority relations than 16th-century science. (Tribby, 1994, for example, discusses an investigation into a ‘gentler’ laxative that could ‘satisfy’ the needs of Grand Duke Ferdinand II as well as those of the many ‘delicate persons’ who visited or had dealings with the court that involved experimentation on individuals described variously as ‘a mercenary’, ‘a vagrant’, ‘the Little Moor’, and so on.)

Over time, in the hard sciences, the term natural experiment has also come to describe both cases where ‘nature’ provides an experiment that resembles the controlled situation that scientists would like observe but are unable to create themselves. An unsuccessful experiment may help make the point clear: in a famous quote by Albert Einstein to Erwin Findlay Freundlich (who was attempting to assess the whether path of a ray of light was affected by gravity), Einstein wrote: ‘If only we had a considerably larger planet than Jupiter! But nature has not made it a priority to make it easy for us to discover its laws.’ (‘Wenn wir nur einen ordentlich grösseren Planeten als Jupiter hätten! Aber die Natur hat es sich nicht angelegen sein lassen, uns die Auffindung ihrer Gesetze bequem zu machen’, (as cited in Ashtekar et al., 2003; translation from the *New York Times*, 24 March 1992).

Natural experiments as serendipitous randomized trials

In contrast to the natural experiment of the hard sciences, the term natural experiment is often used by economists to denote a situation where real randomization was employed, without the intent of providing a randomized experiment. For example, between 1970 and 1972 men from specific birth cohorts were conscripted into the US military by way of a draft lottery. Each day of the year was randomly assigned a number which (in part) determined whether or not one was at risk of being inducted into the military service to fight in the US war on Indochina. As a consequence, men of specific birth cohorts born only a day apart, for example, had very different risks of serving in the military. In Hearst, Newman and Hulley (1986), the authors asked whether the war continued to kill after the warrior returned home. The authors compared, among other things, the suicide rates among individuals who on average were *ex ante* similar, but who had very different probabilities of having completed military service.

The example is sufficiently simple to make a number of points about the limitations of natural experiments. *If* one can assume that the mere fact of having such a birth date put one at high risk of military duty, and that having a birth date raised (or did not lower) any person's risk of serving in the military, then it is possible to use something akin to two stage least squares (2SLS) to estimate an ‘average’ effect of military service for those who were induced to serve in the military by the draft lottery. However, Hearst, Newman and Hulley (1986) are quick to observe that *whether or not* one actually served in the military, the mere fact of having been put at risk of the lottery might have had an effect on delayed mortality. In econometric terms, this would be a violation of the ‘exclusion restriction’ of 2SLS. If such is the case, it is apparent that a comparison of men with high-risk birthdays to those with low-risk birthdays will be an admixture of the effect of the military service on later mortality *and* any direct effect of the lottery itself. An additional problem is the possibility of non-random selection induced by men dying while at war. This was judged to be small due since the fraction of US soldiers who died while serving in action was a small fraction of the total.

Returning to how one might go from an estimate generated in this way to more general inference, one has a number of other obstacles. For example, the delayed mortality effects of military service on those *induced* to serve by an unlucky birth date might be

different from the effect on those who *volunteered* to fight in the war. If the effects are very different, it would obviously be incorrect to use estimates generated by those induced to serve to extrapolate to the broader population of interest.

More generally, our ability to generalize the valid results of an experiment is much more limited when we can only manipulate the cause indirectly (as in the example above) than when we can manipulate the cause directly: there is often the possibility of important differences between persons who take up the treatment as a result of having been encouraged to participate and those who were similarly encouraged but did not take up the treatment.

The regression discontinuity design as a natural experiment

One research design that involves the ‘serendipitous’ randomization of individuals into a treatment is called the regression discontinuity design. Since it is a relatively ‘clean’ example of something that approaches a truly randomized experiment without involving explicit randomization, it provides a good illustration of the strengths and weaknesses of natural experiments. (For an analysis of the relationship between the regression discontinuity design and randomized controlled trials see Lee, 2007.) For illustration, let us consider DiNardo and Lee's (2004) analysis of the causal effect of ‘unionization’ on firms in the United States. The naive approach would be to compare unionized firms to non-unionized firms.

The basis of the regression discontinuity design is the existence of a ‘score’ or a ‘vote’ which assigns persons to one treatment or another. In the US context, workers at a firm can win the right to form a labour union by means of a secret ballot election. If 50 per cent plus one of the workers votes in favour of the union, the workers win the right to be represented by a union; less than that, and they are denied such rights.

To understand how this works, consider elections at two different sets of work sites that employ large numbers of workers. In one set, $0.5+\Delta$ of the workers vote in favour of the union and win the right to bargain collectively where Δ is some small number. In another set, slightly less than 50 per cent vote in favour of the union, and are denied the right to bargain collectively. The vote share in these sites is $0.5-\Delta$. Suppose we have large amounts of data on such elections and can accurately estimate the average outcome (say the fraction of firms that continue to exist 15 years after the vote).

Using almost exactly the same set-up as before, we compare those places where the union wins with those where the union loses:

$$E[\bar{y}_{\text{Union}} - \bar{y}_{\text{No Union}}] = E[y|\text{vote} = 0.5 + \Delta] - E[y|\text{vote} = 0.5 - \Delta]$$

If firm survival is described by the same ‘model’ as in å above, where now $T = 1$ denotes winning the right to bargain collectively, we get:

$$E[\bar{y}_{\text{Union}} - \bar{y}_{\text{No Union}}] = \beta + \left(\underbrace{E[f(X)|\text{vote} = 0.5 + \Delta] - E[f(X)|\text{vote} = 0.5 - \Delta]}_{\text{Observable Differences}}\right) + \left(\underbrace{E[\varepsilon|\text{vote} = 0.5 + \Delta] - E[\varepsilon|\text{vote} = 0.5 - \Delta]}_{\text{Unobservable Differences}}\right)$$

The ‘trick’ is that if we choose Δ to be small enough (that is, close to zero), then

$$E[f(X)|\text{vote} = 0.5 + \Delta] \approx E[f(X)|\text{vote} = 0.5 - \Delta] \text{ and } E[\varepsilon|\text{vote} = 0.5 + \Delta] \approx E[\varepsilon|\text{vote} = 0.5 - \Delta]$$

and we get a ‘good’ estimate of the ‘effect of unions’ in the same sense that we get a good estimate of the effect of a treatment in a randomized controlled trial. That is, if we focus our attention on the difference in outcomes between ‘near winners’ and ‘near losers’ such a contrast is formally equivalent to a randomized controlled trial if there is at least some ‘random’ component to the vote share. For example, sometimes people take ill on the day of the vote – if that happens randomly in some sites, two sites that would have had the same final vote tally had everyone shown up are now different. When such differences are the difference between recognition or not, one has the practical equivalent of a randomized controlled trial. The mere existence of a ‘score’ that discontinuously exposes one to a treatment is not enough. This design would not be appropriate, for example, to analyse the causal effects of US Congressional votes on various issues. Substantial ‘manipulation’ – that is, through negotiation, and so on – of the final vote tally is common and

suggests that individuals near but on opposite sides of the threshold are not otherwise similar (see regression-discontinuity analysis).

A few moments' reflection will make clear both the appeal of such experiments and their limits. Advocates of a natural experiment approach point to the fact that the implicit randomization involved in this design means that we can be more confident with such a comparison than a naive comparison that merely compares unionized to non-unionized firms. This would almost certainly confound the true ‘effect’ with pre-existing differences in unionized and non-unionized firms with ‘unionization’. Advocates will also point to the fact that the experiment is relevant to a potential policy – say lowering the threshold required to win representation rights by a small amount.

Detractors will observe many limitations. Is the effect of a union that is set into a place by a 51 per cent vote the same as the effect of a union where the workers vote unanimously? Possibly not. Stipulating the validity of the estimate, is it reasonable to suggest that the effect of unionization would be the same if all workplaces were allowed to vote on a union? Probably not. Is it possible that a union at one work site affects other work sites? What about the effect on the firm's competitors? Indeed, it is even possible to question the premise that a union is a ‘treatment’ at all. Does it make sense to talk of a single effect of a labour union when there is such heterogeneity in what the notion ‘labour union’ represents? While the anarcho-syndicalist Industrial Workers of the World (IWW) of Joe Hill (a famous militant IWW member and subject of a well-known folksong) and the American Federation of Labor and Congress of Industrial Organizations (AFL-CIO) of George Meany (a conservative ‘anti-communist’ who was its president for many years) were both labour unions, they had virtually contrary aims and wildly different political structures.

More generally, ‘causes’, ‘treatments’, and so on are much more fragile objects for the types of things usually interesting to economists than the types of things interesting to natural science. The concepts of natural science are often capable of quite substantial refinement in a way that concepts in the human sciences rarely are.

‘Natural natural experiments’?

As I have already mentioned, the term ‘natural experiment’ has been used in several different ways inconsistent with our definition. It seems pointless, however, to claim that our definition is the ‘true’ or correct one. We shall therefore consider some cases that use the term which do not obviously involve randomization of a treatment or something that approximates such randomization.

Rosenzweig and Wolpin (2000) for instance, have coined the expression ‘natural natural experiments’ to denote a wide range of studies involving the use of twins. The emphasis on the word ‘natural’ is intended to highlight the role of nature in providing the variation. Twins have been of inordinate interest to the social scientists since they seem to offer the possibility of ‘controlling’ for ‘genetics’. Consider one case of interest to economists, ‘returns to schooling’. Does acquiring an additional year of school result in higher wages in the labour market? How much higher? To fix ideas consider a simple model of the sort:

$$y_{ij} = \beta S_{ij} + a_j + \varepsilon_{ij}.$$

We are interested in some outcome, say hourly wages, and the causal effect of years of schooling S. It will greatly simplify the discussion if we assume that all persons ‘treated’ with ‘schooling’ experience the same increase in their wages – that is, the treatment effect is a constant across individuals. We have gathered a random sample of j=1, …, J ‘identical’ (monozygotic) twins (i=1, 2). The term a_j is not directly observable but includes everything that the twins have in common – genetics, environment, and so on. The error term ε_{ij} includes everything that the twins do not have in common and cannot be observed as well as the effects of misspecification, and so on. Though this simple set-up can be greatly elaborated (see Ashenfelter and Krueger, 1994, for a clear exposition) the essential idea is that the *difference* between the twins purges the outcome of the a_j term so that an ordinary least squares regression of the difference in wages Δy_{ij} on ΔS_{ij} yields a good estimate of

$$\hat{\beta} \text{ is a good estimate of } \beta + \frac{\mathrm{Var}(\Delta\varepsilon, \Delta S)}{\mathrm{Var}(\Delta S)}.$$

The first term is the goal of such studies. The second term points to the possibility that there are other influences which might be correlated both with schooling and that affect the outcome. The second term can be interpreted as the slope coefficient from the following hypothetical ordinary least squares (OLS) regression, where δ is the slope of the ‘best-fitting’ line in this expression:

$$\varepsilon = \text{constant} + S\delta + \text{error}.$$

When will $\hat{\beta}$ to be a good estimate of the returns to schooling β? The conditions are essentially the same as for the randomized controlled trial: if we can treat the assignment of schooling to the two twins as if it were determined by a random coin toss then differences in the level of schooling between the two twins – ΔS_{ij} – will be independent of differences between the two twins in unobserved influences on wages – $\Delta \varepsilon_{ij}$. Detractors of this approach doubt that such an assumption is plausible. In simple language, if the twins are so 'identical' why do they have different levels of schooling? Perhaps the parents noticed that one twin was more interested or had more 'aptitude' for schoolwork than another. If that were the case, estimates of the returns to schooling would be confounded with differences in the aptitude for schooling despite the fact that we had 'controlled' for a large number of other factors. The key difference between this case and what I have identified as a natural experiment is the lack of an obvious approximation to randomization. Bound and Solon (1999) discuss, inter alia, a host of difficulties in treating twin differences as experimental variation. I do not discuss twins studies that utilize twins as a 'surprise' to family size which have some element of randomization.

Other research designs. quasi-experiments

Finally, I should make note of the fact that some authors use the term natural experiment more broadly than I have construed it here. Meyer (1995, p. 151) for instance, considers natural experiments the broad class of research designs 'patterned after randomized experiments' but not (generally) involving actual randomization. One term often used for such situations is 'quasi-experiment'. The relationship between these quasi-experiments and the natural experiments I have been describing is quite varied and ranges from those whose difference from the standard of randomized assignment is merely a matter of 'degree' to those in which assignment to treatment differs so much from the standard of randomization that it is really a difference in 'kind'.

Most of these quasi-experiments are variants of a 'before and after' where an observation is made before and after a treatment. Often a before–after comparison for one set of observations (the treatment – *T*) is compared to another set (the control – *C*). A typical set-up might compute a treatment effect by taking the difference in two differences:

$$\text{Treatment Effect} = \{\bar{Y}_{T,\text{after}} - \bar{Y}_{T,\text{before}}\} - \{\bar{Y}_{C,\text{after}} - \bar{Y}_{C,\text{before}}\}.$$

For this reason, such quasi-experiments are described as using 'difference-in-differences' approach to identifying a causal relationship. In the United States, the fact that the state (or city) governments have some liberty to enact laws independently of the federal government, for example, has led to a great deal of research using 'Brandeisian' experiments. The term comes by way of US Supreme Court Justice Louis Brandeis, in the case *New State Ice* v. *Liebmann*:

> There must be power in the States and the Nation to remould, through experimentation, our economic practices and institutions to meet changing social and economic needs. … It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country. (U.S. Supreme Court *New State Ice Co.* v. *Liebmann*, 285 U.S. 262 (1932))

To give one such example, consider DiNardo and Lemieux's (2001) evaluation of the effect of changing the age at which it is legal to purchase alcohol or the consumption of marijuana. At the beginning of the 1980s states generally enforced two types of legal regimes. In one set, alcohol could not be legally sold to those under the age of 21. In another, the legal minimum drinking age (LMDA) was 18. In the mid-1980s, the federal government put a great deal of pressure on those states with LMDA of 18 to raise them to 21 and by the end of the 1980s, in all states drinking age was 21.

The assignment of drinking age statutes to the states at the beginning of the 1980s could not be considered 'approximately' random. Utah, for example, which is home to a large number of adherents to the Mormon religion – which proscribes alcohol use – had a 21-year drinking age at the beginning of the 1980s. However, due to a federal policy implemented in the mid-1980s of eventually denying federal highway funds to states with legal minimums less than 21 years old, something perhaps approximating an 'experiment' can be arrived at by comparing *changes* in alcohol or marijuana consumption during the 1980s in those states which

were forced to change (and changed early) with those who were forced to but raised their drinking age later.
Let Δy_t denote the change in the fraction of 18–21 year olds who reported smoking marijuana in the previous 30 days from 1980 to 1990 in states that had 18-year-old drinking ages that were increased, and Δy_c denote the similar change in states whose drinking age was always 21. Then an estimate of the effect of the drinking age might be:

$$\Delta y_t - \Delta y_c = \text{Effect of LMDA.}$$

Although randomization is not employed per se, the credibility of these exercises can be at least partially evaluated. For instance, if the outcome of interest has been approximately constant in both the treatment and control groups for a long time preceding the change in legal regime, the estimate is generally more credible. Less credible is the case in which the outcomes in the control group and the treatment group are quite variable over time, the control group and the treatment group do not follow similar patterns *before* the proposed experiment, or when both are true.

Controversies: concluding remarks

Natural experiments and their like have been at the heart of much work in economics. Nonetheless, they are the subject of considerable debate. One of the most cited limitations of natural experiments – by both supporters and detractors – is that such experiments are context specific. Indeed, one frequently encountered ‘strength’ of natural experiments is that it often concerns the evaluation of an actual policy. There are limitations, however. If we assume that the experiment is ‘internally valid’ we still have to ask: how do we generalize from one experiment to the broader questions of policy? The foregoing has suggested that it is difficult. There are at least three broad classes of reasons:

1. 1. While a natural experiment might provide a credible estimate of some particular serendipitous ‘intervention’, this may have only a weak relation to the type of interventions being contemplated as policies. Many of the potential reasons for a weak relationship are similar to those encountered in social experiments (among other things, for example, the effect of a treatment in a demonstration programme might be quite different from the outcome that would obtain if the treatment were applied more broadly or to different persons).
2. 2. Some interesting questions are unanswerable with such an approach because serendipitous randomized experiments are few and far between. The extent to which this criticism is warranted, of course, depends on the availability of alternative ways of putting our views to a severe test.
3. 3. More generally, without a ‘theory’, estimates from natural experiments are uninterpretable.

I am sympathetic with all three criticisms although (3) deserves some qualification. While it has been argued that even in the natural sciences it is impossible to have ‘pre-theoretical’ observations or experiments, Hacking (1983) makes a strong case that experimentation has a life of its own, sometimes suggesting ideas in advance of theory, other times the consequence of theory, and sometimes testing theories. Much of this debate in the natural sciences revolves around the notion of what constitutes a ‘theory’. Whatever the validity of the view that one cannot experiment in advance of ‘theory’ in the natural sciences, in the social sciences, it is clear that no theory has the same standing as, say, general relativity in physics. This is the sense in which Noam Chomsky observes that ‘as soon as questions of will or decision or reason or choice of action arise, human science is pretty much at a loss’ (Magee, 2001, p. 184). Indeed, the standing of randomized experiments – in some fields of enquiry regarded as ‘the gold standard’ of evidence – is a great deal lower than the best experiments of natural science; they are most often useful in situations otherwise marked by ‘complete ignorance’ (Hacking, 1988). In short, while the human sciences might have the same ambition as natural science, the status of what we know will almost surely be quite limited.
Nonetheless, one does not need a ‘correct’ theory to hand, nor an understanding as rich as that found in some of the natural sciences to find an experiment useful. At the risk of over-using such metaphors, the fact that the Michelson–Morley experiments were in part about testing for the existence of ‘ether’ did not make them uninteresting. Experiments are just ways to use things we (think we) understand to learn about something we do not. And while the sorts of ‘natural’ experiments ‘serendipitously’ provided by society may be very limited and are often the product of unhappy social realities, they can sometimes perhaps serve a small role in enhancing our understanding.
Any assessment of the usefulness of natural experiments depends on how one judges the power of other methods of enquiry. Such a discussion is well beyond the scope of this article. Nonetheless, not discounting their many limitations, one benefit of natural

experiments I have tried to highlight is that for some they might open up the possibility of revising their beliefs in light of evidence or suggest new ways to think about old problems, however limited. A key aspect of experiments (natural or otherwise) is the willingness to put one's ideas ‘to the test’. Often, careful study of a natural experiment, however limited, may also make one aware of how complicated and difficult are the problems we call ‘economics’. Even if the success we might have in generalizing natural experiments more broadly may be quite limited, if they bring nothing but humility to the claims social scientists make about much we actually understand, that alone would justify an interest in natural experiments.

See Also

- difference-in-difference estimators
- experimental economics
- experimental economics, history of
- experimental labour economics
- experimental methods in economics
- experiments and econometrics
- Fisher, Ronald Aylmer
- regression-discontinuity analysis

Bibliography

Ashenfelter, O. and Krueger, A.B. 1994. Estimates of the economic returns to schooling from a new sample of identical twins. *American Economic Review* 84, 1157–73.

Ashtekar, A., Cohen, R.S., Howard, D. Renn, J. Sarkear, S. and Shimony, A. 2003. *Revisiting the Foundations of Relativistic Physics: Festschrift in Honor of John Stachel*. Boston Studies in the Philosophy of Science, vol. 234. Dordrecht: Kluwer Academic.

Bastable, C.F. 1987. Experimental methods in economics (i). In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Bound, J. and Solon, G. 1999. Double trouble: on the value of twins-based estimation of the return to schooling. *Economics of Education Review* 18, 169–82.

DiNardo, J. 2007. Interesting questions in freakonomics. *Journal of Economic Literature*.

DiNardo, J. and Lee, D.S. 2002. The impact of unionization on establishment closure: a regression discontinuity analysis of representation elections. Working Paper No. 8993. Cambridge, MA: NBER.

DiNardo, J. and Lee, D.S. 2004. Economic impacts of new unionization on private sector employers: 1984–2001. *Quarterly Journal of Economics* 119, 1383–441.

DiNardo, J. and Lemieux, T. 2001. Alcohol, marijuana, and American youth: the unintended consequences of government regulation. *Journal of Health Economics* 20, 991–1010.

Drake, S. 1981. *Cause, Experiment, and Science: A Galilean Dialogue, Incorporating a New English Translation of Galileo's Bodies that Stay atop Water, or Move in It*. Chicago: University of Chicago Press.

Fisher, R.A. 1935. *Design of Experiments*. Edinburgh, London: Oliver & Boyd.

Hacking, I. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hacking, I. 1988. Telepathy: origins of randomization in experimental design. *Isis* 79, 427–51.

Hacking, I. 2000. *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Hearst, N., Newman, T.B. and Hulley, S.B. 1986. Delayed effects of the military draft on mortality: a randomized natural experiment. *New England Journal of Medicine* 314, 620–4.

Heckman, J.J. 2005. The scientific model of causality. *Sociological Methodology* 35, 1–97.

Heckman, J.J. and Smith, J.A. 1995. Assessing the case for social experiments. *Journal of Economic Perspectives* 9(2), 85–110.

Lee, D.S. 2008. Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*.

Magee, B. 2001. *Talking Philosophy: Dialogues with Fifteen Leading Philosphers*. Oxford: Oxford University Press.

Mayo, D.G. 1996. *Error and the Growth of Experimental Knowledge Science and Its Conceptual Foundations*. Chicago: University of Chicago Press.

Meyer, B. 1995. Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics* 13, 151–61.

Morgan, M.S. 1987. Statistics without probability and Haavelmo's revolution in econometrics. In *The Probabilistic Revolution: Ideas in the Sciences*, vol. 2, ed. L. Krüger, G. Gigerenzer and M.S. Morgan. Cambridge, MA: MIT Press.

Nelson, A. 1990. Are economic kinds natural? In *Scientific Theories of Minnesota Studies in the Philosophy of Science*, vol. 14, ed. C. Wade Savage. Minneapolis: University of Minnesota Press.

Peirce, C.S. 1958. In *Collected Papers*, vols. 7–8, ed. A. Burks. Cambridge, MA: Harvard University Press.

Rosenzweig, M.R. and Wolpin, K.I. 2000. Natural ‘natural experiments’ in economics. *Journal of Economic Literature* 38, 827–74.

Searle, J. 1995. *The Construction of Social Reality*. New York: Free Press.

Shadish, W.R., Cook, T.D. and Campbell, D.T. 2002. *Experimental and Quasi–Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Tribby, J. 1994. Club Medici: natural experiment and the imagineering of ‘Tuscany’. *Configurations* 2, 215–35.

Voltaire. 1759. *The History of Candide; or All for the Best*, ed. C. Cooke. London, 1796.

Waller, R. 1684. *Essayes of natural experiments made in the academie del cimento, under the protection of the most serene Prince Leopold of Tuscany*. Facsimile edn, ed. R. Hall, trans. R. Waller. New York and London, 1964.

Wikipedia. 2006. Experiment. http://en.wikipedia.org, accessed 28 September 2006.

How to cite this article

DiNardo, J. "natural experiments and quasi-natural experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000142>
doi:10.1057/9780230226203.1162

The New Palgrave Dictionary of Economics Online

natural price

G. Vaggi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bon prix; Cantillon, R.; competition; cost of production; effectual demand; exchange value; fundamental price; Hutcheson, F.; Locke, J.; market price; Marshall, A.; Marx, K. H.; natural price; net product; normal price; North, D.; Petty, W.; Physiocracy; price theory; Pufendorf, S.; Quesnay, F.; rent; Smith, A.; Sraffa, P.; Steuart, Sir J.; Turgot, A. R. J.; uniform rate of profit; value theory

Article

In the *Wealth of Nations* Smith says that

> when the price of any commodity is neither more nor less than what is sufficient to pay the rent of the land, the wages of labour, and the profits of the stock employed in the raising, preparing and bringing it to market, according to their natural rates, the commodity is then sold for what may be called its natural price. (Smith, 1776, p. 72)

In the same chapter he explains that in economic theory this particular price level is important because it is a sort of benchmark for the actual price of the commodity, its market price (p. 73). The market price is different from the natural price but tends to move towards it all the time because of competition between producers. ‘The natural price, therefore, is, as it were, the central price, to which the prices of all commodities are continuously gravitating’ (p. 73). Smith's concept of natural price and his description of the competitive mechanism which guarantees that the market prices tend to move towards it became an important element in classical political economy. Smith's analysis was entirely subscribed to by Ricardo (Ricardo, 1821, pp. 88–91), and was a central point in the classical theory of value and in the price theories of some neoclassical economists.

Smith's notion of natural price is part of a more general analysis of the normal and regular causes which determine the value of commodities. Smith's theory can be divided into three main aspects. First of all, there is the definition of natural price, which is made up of three component parts, wages, profits, and

rent. In Chapter 6 of the *Wealth of Nations*, Smith explains that the price of all commodities resolves itself into wages, profits and rent, as soon as we abandon the ‘early and rude state of society which precedes both the accumulation of stock and the appropriation of land’ (Smith, 1776, p. 65). The price must also repay the raw materials and the capital equipment consumed in production, but the prices of these commodities are also made up of the wages, profits and rent required in their own production (p. 68). Thus ultimately the price of each product is entirely made up of those three parts, which include the incomes of workers, landlords and capitalists who take part in the final production of the good and also the incomes of all those who have indirectly contributed to produce it in previous years. The techniques of production of a commodity have an important influence on its natural price, because they determine the relative shares of profits, rent, and wages. But the natural price also depends on the distribution of income, that is to say, on the level of the natural rates at which wages, rent and profits must be paid. According to Smith, each rate is determined on a different market and this depends on several circumstances. Therefore the natural price of each commodity is determined by the methods of production and by the exogenously given values of the rates which remunerate the three classes which take part in production. It is worth noticing that for Smith, society is made up of different classes, labourers, landlords and capitalist entrepreneurs, whose economic functions are clearly separated. When all the commodities that make up the output of society are assessed according to their natural prices, the part of this value given by wages is the capital stock of society (p. 110), while rent and profits make up the net product, or surplus.

The second feature of Smith's price theory is the description of the reasons why the natural price is the price level which prevails in the long run, and around which market prices gravitate. This price mechanism is an important element in the notion of natural price because it guarantees that the permanent causes of value are those which influence the natural price, while market price deviations are due to temporary circumstances. The market price fluctuates and may differ from the natural price, but there are forces which compel it towards the natural price.

The factors affecting natural prices must be regarded as the permanent and fundamental forces that determine the value of produced commodities, quite independently from the day-to-day changes in their market prices. This second part of Smith's analysis of natural prices contains several concepts. First, there is the notion of effectual demand which is used to explain the differences between natural and market prices. Effectual demand is the ‘demand of those who are willing to pay the natural price of a commodity’ (p. 73). Of course a change in this price affects the effectual demand. The quantity produced and brought to the market may be lower (or higher) than the effectual demand, in which case the market price of the commodity will be higher (or lower) than the natural one. This mechanism explains why there are differences between natural and market prices.

The second step in Smith's analysis of the gravitation of market prices around natural prices consists in the competitive mechanism itself. Here, too, several logical stages may be distinguished. (a) For Smith the fact that the market price is higher than the natural one implies that at least one of the three parts which make up the price of a product is higher than it would have been if its contribution to production was remunerated according to its natural rate; it seems reasonable to assume that profits are the share which takes advantage of the favourable market conditions (but the process works in the same way if wages and rent are higher than their natural rates). (b) Entrepreneurs are aware of the existence of these different rates of profit in the different sectors of the economy. (c) There are no barriers to the free circulation of capital, thus entrepreneurs move towards the most remunerative sectors; this is the crucial

aspect of Smith's analysis of competition (see Sylos-Labini, 1976). (d) These capital movements lead to an increase in the output of the products which yield the highest rates of profit. (e) Since the quantity produced and brought to the market of these products increases while the effectual demand in unchanged, the market price falls. This does not mean that there is a downward-sloping demand schedule. In Smith's price theory there is no continuous differentiable inverse relationship between quantities and prices, as is found in neoclassical economics (Garegnani, 1983).

Free competition tends to bring about a uniform rate of profit throughout the economy. Hence the concept of natural price is related to the existence of a single rate of profit on the capital invested in all sectors, and is regarded by Smith as ‘a centre of repose and continuance’ for the actual market price (Smith, 1776, p. 75).

The view that it is possible and useful to separate the day-to-day fluctuations in market prices from the stable and permanent causes of the value of commodities can be traced back to the 16th century. It was part of Scholastic tradition to believe that there was a logical distinction between the actual price of a product and its *true* value. The former price can vary quite a lot according to the state of trade, while the value is always the same. Von Pufendorf believed that the value, or just price, of a commodity depended mostly on the difficulty of acquiring and producing it (Pufendorf, 1688, pp. 684–9). Theoreticians of the just price regarded it as the level to which actual prices ought to conform. They gave no indication of any spontaneous mechanism which should guarantee that market values would adapt to these just levels. As a student, Adam Smith read the works of von Pufendorf, and his teacher, Francis Hutcheson, wrote a book entitled *A System of Moral Philosophy* in which the distinction between value and price was restated along very similar lines (Hutcheson, 1754–5, pp. 53–5). At the end of the 17th century, Dudley North and John Locke maintained that regulations and government interventions could not affect the price of commodities, which depended on market conditions (North, 1691, Preface; Locke, 1691, pp. 4, 11, 13).

Some years before the publication of the works of Locke and North, Sir William Petty regarded the cost of production of commodities as the main cause determining their true value. Ultimately all commodities are produced by two common denominators, land and labour, and their exchange values are in proportion to the quantities of these non-produced goods which have been employed in their production (Petty, 1662, p. 44). The value of goods is regulated by the physical cost of production, which is regarded as the true measure of the difficulty of acquiring them. For Petty, the natural price depends upon the amount of labour required to produce a commodity with the best available technique (pp. 50–1). Richard Cantillon developed Petty's analysis of land and labour as the original components of the value of each commodity. He transformed the amount of labour employed in production into an equivalent quantity of land. Thus, the value of each commodity is given by the quantity of land which has been directly and indirectly used in its production (Cantillon, 1755, p. 29). This is the intrinsic value of the products, and their market price fluctuates around it (pp. 28–30). Moreover, Cantillon presented the well-known theory of the ‘three rents’; the farmer receives two thirds of the products of land, one third is required to pay workers' wages and other expenses, the second third is the profit from his enterprise; the final third accrues to landlords as rent (p. 43).

Quesnay and the Physiocrats also distinguish the permanent value of commodities from their market price. For Quesnay, the fundamental price is the lowest level of the selling price for the producer. This value is the minimum level of the market price: it is the sum of all the expenses incurred by the

cultivator in the production of a commodity, and there is a loss when the market price is lower than this value (Quesnay, 1757, p. 555). The fundamental value of commodities is stable and varies quite slowly, on the other hand market prices change rapidly. Quesnay concentrated his attention on the fundamental price of primary commodities, which included the technical costs of production plus the annual rent paid to the landlords (1757, p. 555; Quesnay, 1756, p. 443). Quesnay believed that two elements contribute to determining the fundamental value of agricultural products: farming techniques, which determine the physical cost of production, and the rule which fixes the distribution of income, at least in the form of rent. The inclusion of an element, rent (which is part of the country's surplus), in the fundamental value of a commodity is an important step towards Smith's concept of natural price. Now the permanent value of commodities is not only the result of technical conditions but also of the social rules and customs which determine the distribution of the net product.

Quesnay used the term ‘natural price’ to indicate the state of prices when free and unobstructed competition in all the markets regulates the exchanges between buyers and sellers (Quesnay, 1766, pp. 829–30). In this case the actual exchange value of the products of land is a *bon prix*, it exceeds the fundamental price and leaves the farmer with a profit (Quesnay, 1757, p. 529). Quesnay provided a good explanation of the reasons why the market price cannot be lower than the fundamental one, but there is no indication of the existence of market forces which lead the actual price towards the *bon prix*. In Quesnay's value theory the notion of fundamental price is only a sort of threshold which fixes the lowest market price, but profits are still not part of the fundamental price.

In 1767 Sir James Steuart published *An Inquiry into the Principles of Political Oeconomy* in which he made at least two important contributions to the classical theory of value. The first was the notion of the real, or intrinsic, value of the goods. He says that two things make up the price of a product, ‘the real value of a commodity and the profit upon alienation’ (Steuart, 1767, p. 159). The real value is the cost of production, which depends upon the average techniques which have been adopted and which establishes the amount of time needed to produce a commodity. The ‘profit upon alienation’ is the positive difference between the actual price and the real value (1767, p. 159). Thus profits are not part of the value of commodities, but according to Steuart ‘such profits subsisting for a long time, they insensibly become *consolidated*, or as it were, transformed into the intrinsic value of the goods’ (1767, p. 193, Steuart's italics). Thus, in the normal condition of the market, the value of commodities must also include entrepreneurs' profits, which are a permanent feature of the exchange value of goods. Steuart's second contribution to price theory is the concept of effectual demand; this notion indicates the demand of consumers who can actually pay for a product and is clearly distinguished from wants and desires (1767, pp. 151–3). Steuart's analysis does not provide a theory of profit capable of explaining the level which becomes consolidated in the intrinsic value of commodities. The normal value is not yet defined in a way which explains the existence of a regular element of profit in the exchange value of commodities.

In the *Obsérvations sur le mémoire de Saint Péravy*, Turgot distinguished the fundamental and market price of commodities. The first concept is defined as the cost of production, which includes wages, raw materials and interests on the capital advanced. The fundamental value is fairly stable, while the exchange value is ruled by supply and demand and ‘it has a tendency to approach it (the fundamental price) continually, and can never move away from it permanently’ (Turgot, 1767, p. 120, n. 16). There is an important difference between Quesnay's and Turgot's use of the term ‘fundamental price’. Turgot's

notion does not simply indicate the lowest level of the market price, but is the value to which this price must tend. Turgot included a regular profit among the necessary expenses of production (Meek, 1973, p. 17). Turgot's interest on the capital advanced is not only a depreciation allowance but includes profit for the entrepreneur. In *Réflexions sur la formation et la distribution des richesses* (1766) Turgot clearly says that the return to the capitalist entrepreneur must be divided into three main categories: ‘depreciation of the capital’, ‘wages of superintendence and direction as well as the risk premium’ and ‘pure return on his capital which he could have earned if he had not employed it in industry’ (Groenewegen, 1971, p. 333; see Turgot, 1766, pp. 152, 154). Now profits are an essential part of the permanent value of commodities, but above all Turgot's notion of profit is different from those of Steuart and Quesnay. Profit is defined as a rate on the capital invested. This definition of profits is quite different from that of profit upon alienation, according to which profits are influenced by market conditions where the products are sold. For Turgot, on the contrary, the rate of profit depends mainly on competition between capitalist producers who act with a view to obtaining the highest possible rate of profit. This mechanism explains the existence of a continuous tendency towards the equalization of rates of return in all of the capital.

In the *Lectures on Jurisprudence* which Adam Smith gave at Glasgow in the academic year 1762–3, we already find the distinction between natural and market price, together with the description of the mechanism by which the latter price gravitates around the natural value (Smith, 1762–3, pp. 353 ff.). Smith's analysis of competition among producers explains that natural prices are bound with the existence of a uniform rate of profit in all the sectors of the economy. The existence of this uniform rate has been traditionally adopted to describe the prices which prevail in the long run, when it is possible to abstract from all the accidental causes which influence market prices. In Smith's economics, technology and income distribution are the permanent forces which determine the value of natural prices.

In classical economics, the notion of natural price is necessary to build up an abstract analysis of the main features of the economy. This notion helps to single out the main characteristics of the capitalistic process of development and their relationships to changes in the distribution of income. Thus the concept of natural price is part of the study of the long-term changes in economic systems, which derive from capital accumulation. Natural price is an essential element of the classical method of analysis, which investigates the features of the long-term positions of the economy, when demand does not affect prices and income distribution (Garegnani, 1976, section 1).

In Chapter 4 of *On the Principles of Political Economy and Taxation*, Ricardo subscribes to Smith's theory of natural prices (1821, pp. 88–92). He was interested in the analysis of the permanent changes in income distribution, and was not interested in the temporary deviation of market prices from their natural value.

However, there is a major difference between Smith's and Ricardo's theories of profit. Smith says that profits and wages are determined on separate markets and that the natural price is the sum of these shares plus rent, while Ricardo says that the rate of profit and the real wage are inversely related.

Marx's notion of prices of production shares many of the features of Smith's natural price; both concepts are associated with the existence of a uniform rate of profit in all sectors of the economy (see Marx, 1894, pp. 153–8). Moreover, Marx accepted Ricardo's analysis of the reasons why market prices fluctuate around natural ones (1894, p. 179). Like Ricardo, he believed that real wages and the rate of profit vary in opposite directions. In his 1951 Introduction to *The Works and Correspondence of David*

Ricardo, Sraffa clearly singled out the implications of Ricardo's theory of profit determining commodities natural value. Sraffa explicitly mentioned the concepts of natural price and prices of production in presenting his theory of price determination and retained the notion of a uniform rate of profit throughout the economy (Sraffa, 1960, pp. 9, 6).

In the *Principles of Economics* (1920), Alfred Marshall referred to Smith's natural price, for which he substituted the notion of normal price (Marshall, 1920, p. 289). In his discussion of the causes which influence the value of commodities he said that in general, market values are deeply affected by demand, while normal prices depend on the cost of production of commodities. The former price prevails in the short run, but ‘the longer the period, the more important becomes the influence of cost of production on value’ (1920, p. 291). Normal prices are determined by the persistent causes of value, and are not influenced by fitful and irregular events (1920, pp. 304–5). It should be pointed out that Marshall's notion of cost of production is not the same as the notion put forward by Ricardo and Marx. Moreover, he was sceptical about the existence of a tendency towards the equalization of the rates of profit in all economic activities (1920, pp. 506–7, 512). Nevertheless inside each branch of trade there can be a fair rate of profit which must be reckoned as a component element of the normal price (1920, pp. 513–14).

See Also

- British classical economics
- market price

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Ed. H. Higgs, London: Cass, 1959.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in Modern Capital Theory*, ed. M. Brown, K. Sato and P. Zarembka. Amsterdam: North-Holland.

Garegnani, P. 1983. The classical theory of wages and the role of demand schedules in the determination of relative prices. *American Economic Review* 73, 309–13.

Groenewegen, P.D. 1971. A reinterpretation of Turgot's theory of capital and interest. *Economic Journal* 81, 327–40.

Hutcheson, F. 1754–5. *A System of Moral Philosophy*. Glasgow: Robert and Andrew Foulis.

Locke, J. 1691. Some considerations of the consequences of the lowering of interest and raising the value of money. In *The Works of John Locke*, vol. 5. London, 1823.

Marshall, A. 1920. *Principles of Economics*, 8th edn. Reprinted, London: Macmillan, 1972.

Marx, K. 1894. *Capital*, vol. 3. London: Lawrence & Wishart, 1977.

Meek, R.L. 1962. *The Economics of Physiocracy*. London: George Allen & Unwin.

Meek, R.L., ed. 1973. *Turgot on Progress, Sociology and Economics*. Cambridge: Cambridge University Press.

North, D. 1691. Discourses upon trade. In *Early English Tracts on Commerce*, London: The Political Economy Club. Reprinted, ed. J.R. McCulloch, Cambridge: Cambridge University Press, 1954.

Petty, W. 1662. A treatise of taxes and contributions. In *The Economic Writings of Sir William Petty*, ed. C.H. Hull. Cambridge: Cambridge University Press, 1899.

Pufendorf, S. von. 1688. *De jure naturae et gentium – libri octo*. Oxford: Clarendon Press, 1934.

Quesnay, F. 1756. Fermiers. In *François Quesnay et la Physiocratie*, Paris: INED, 1958.

Quesnay, F. 1757. Hommes. In *François Quesnay et la Physiocratie*, Paris: INED, 1958.

Quesnay, F. 1766. Du commerce. In *François Quesnay et la Physiocratie*, Paris: INED, 1958.

Ricardo, D. 1821. On the principles of political economy and taxation. In *The Works and Correspondence of David Ricardo*, ed. P. Sraffa with the collaboration of M.H. Dobb. Cambridge: Cambridge University Press, 1951.

Smith, A. 1762–3. *Lectures on Jurisprudence*. Ed. R.L. Meek, D.D. Raphael and P.G. Stein, Oxford: Oxford University Press, 1978.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Oxford University Press, 1976.

Sraffa, P. 1951. *Introduction to The Works and Correspondence of David Ricardo*, vol. 1. Cambridge: Cambridge University Press.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Steuart, J. 1767. *An Inquiry into the Principles of Political Oeconomy*. Ed. A. Skinner, Edinburgh and London: Oliver & Boyd, 1966.

Sylos-Labini, P. 1976. Competition: the product market. In *The Market and the State*, ed. T. Wilson and A. Skinner. Oxford: Clarendon Press.

Turgot, A.R.J. 1766. Réflexions sur la formation et la distribution des richesses. In Meek (1973).

Turgot, A.R.J. 1767. Observations sur le mémoire de Saint-Péravy. In *The Economics of A.R.J. Turgot*, ed. P.D. Groenewegen. The Hague: Martinus Nijhoff, 1977.

How to cite this article

Vaggi, G. "natural price." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000020> doi:10.1057/9780230226203.1163

The New Palgrave Dictionary of Economics Online

natural rate and market rate of interest

Axel Leijonhufvud
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The terms ‘natural rate’ and ‘market rate’ of interest were introduced by Wicksell (1898; 1906) to denote an equilibrium value and the actual value of the real rate of interest. Wicksell applied these concepts to explain the inter-equilibrium movement of money and prices using the hypothesis of maladjustments in the interest rate. Wicksell's work made the nexus between money creation, intertemporal resource allocation disequilibrium and movements in money income the dominant theme in macroeconomics for three decades. However, Keynes's conclusions over the saving–investment problem in the *General Theory* led to the abandonment of the concept of ‘natural’ rate of interest.

Keywords

credit; cumulative processes; fiat money; forced saving; Hayek, F. A. von; inflation; inflation targeting; inside money; intertemporal resource allocation; IS–LM model; Keynes, J. M.; liquidity preference; loanable funds; monetarism; money supply; national income; natural rate and market rate of interest; neutrality of money; new classical macroeconomics; overinvestment; quantity theory of money; saving–investment coordination; Stockholm School; unemployment equilibrium; wage rigidity; Wicksell, J. G. K.

Article

The main analytical elements of Knut Wicksell's *Interest and Prices* can be found in the works of earlier writers. Wicksell was familiar with Ricardo's distinction between the direct and indirect transmission of monetary impulses. Although unknown to Wicksell in 1898, Henry Thornton had provided a clear account of the cumulative process in 1802, as had Thomas Joplin of the saving–investment analysis somewhat later (cf. Humphrey, 1986).

Yet Wicksell did not just coin the terms ‘natural rate’ and ‘market rate of interest’. His development (1898; 1906) of these ideas made the nexus between money creation, intertemporal resource allocation

disequilibrium and movements in money income the dominant theme in macroeconomics for three decades until it was submerged in Keynesian economics. His starting point was the quantity theory, understood as the proposition that in the long run the price level will tend to be proportional to the money stock. His objective was to explain how both money and prices come to move from one equilibrium level to another. This inter-equilibrium movement became his famous ‘cumulative process’. The maladjustment of the interest rate was the key hypothesis in Wicksell's explanation.

The ‘market rate’ denotes the actual value of the real rate of interest while the ‘natural rate’ refers to an equilibrium value of the same variable. The latter term by itself divulges Wicksell's engagement in the ancient quest for a ‘neutral’ monetary system, that is, a system neutral in the original sense that all relative prices develop as they would in a hypothetical world without paper money. Wicksell asserted three equilibrium conditions that the interest rate should satisfy; the first of these was that the market rate should equal the rate that would prevail if capital goods were lent and borrowed in kind (*in natura*). This criterion was later shown by Myrdal, Sraffa and others not to have an unambiguous meaning outside the single input–single output world of Wicksell's example. The further development of Wicksellian theory, therefore, centred around the two remaining criteria: saving–investment coordination and price level stability.

The interest rate has two jobs to do. It should coordinate household saving decisions with entrepreneurial investment decisions and it should balance the supply and demand for credit. If the supply of credit were always to equal saving and the demand for credit investment, the two conditions could always be met simultaneously. But there is no such necessary relationship between saving and investment on the one hand and credit supply and demand on the other. In Wicksell's system the banks make the market for credit; they may, for instance, go beyond the mere intermediation of saving and finance additional investment by creating money; the injection of money drives a wedge between saving and investment; this could only be so if the banks set the market rate below the ‘natural’ value required for the intertemporal coordination of real activities. The resulting inflation and endogenous growth of the money supply would continue as long as the banking system maintained the market rate below the natural rate. Wicksell analysed the case of a ‘pure credit’ economy in which the cumulative process could go on indefinitely, but he also pointed out that, in a gold standard world, the banks would eventually be checked by the need to maintain precautionary balances of reserve media in some proportion to their demand obligations.

Wicksell used the model to explain long-term trends in the price level and was critical of those who, like Gustav Cassel, used it to explain the business cycle. Nonetheless, subsequent developments of his ideas went altogether in the direction of shorter-run macroeconomic theory. In Sweden, Erik Lindahl (1939) and Gunnar Myrdal (1939) refined the conceptual apparatus, in particular by introducing the distinction between *ex ante* plans and *ex post* realizations and thereby clarifying the relationship between Wicksellian theory and national income analysis. The attempts by the Stockholm School to improve on Wicksell's treatment of expectations were less successful, however, producing a brand of generalized process-analysis in which almost ‘everything could happen’.

In Austria, Ludwig von Mises and Friedrich von Hayek focused on the allocational consequences of the Wicksellian inflation story. The Austrian overinvestment theory of the business cycle became known to English-speaking economists primarily through Hayek's *Prices and Production* (1931). In expanding the money supply, the banks hold market rate below natural rate. At this disequilibrium interest rate, the business sector will plan to accumulate capital at a rate higher than the planned saving of the household

sector. If the banks lend only to business, the entrepreneurs are able to realize their investment plans whereas households will be unable to realize their consumption plans (‘forced saving’). The too rapid accumulation of capital (which also has the wrong temporal structure) cannot be sustained indefinitely. The eventual collapse of the boom may then be exacerbated by a credit crisis as some entrepreneurs are unable to repay their bank loans.

The Austrian ‘monetary’ theory of the cycle has been overshadowed first by Keynesian ‘real’ macrotheory and later by monetarist theory. One problem with it is the firm association of inflation with overinvestment. The US stagflation in the 1970s, for example, will not fit. The reasons lie largely in the changes that the monetary system has undergone. Most obviously, commercial banks now lend to all sectors and not only to business. More importantly, however, inflation in a pure fiat regime does not tend to distort intertemporal values in any particular direction (although it may destroy the system's capacity for coordinating activities over time): it simply blows up the nominal scale of real magnitudes at a more or less steady or predictable rate. In contrast, the Austrian situation that preoccupied Mises and Hayek in the late 1920s was one of credit expansion by a small open economy on the gold standard. Given the inelastic nominal expectations appropriate to this regime, the growth of inside money would be associated with the distortion of relative prices and misallocation effects predicted by the Austrian theory.

In England, Dennis Robertson and J. Maynard Keynes both worked along Wicksellian lines in the 1920s. The novel and complicated terminology of Robertson's *Banking Policy and the Price Level* (1926) may have made the work less influential than it deserved. Keynes's *Treatise on Money* (1930), although also remembered as a flawed work, nonetheless remains important as a link in the development of macroeconomics from Wicksell to the *General Theory*.

In the *Treatise*, Keynes, like Wicksell, assumes that the process starts with a real impulse, that is, a change in investment expectations. Unlike Wicksell, he focuses on deflation rather than inflation. For Keynes with his City experience, *the* interest rate was determined on the Exchange rather than set by the banks. Consequently, a deflationary situation with the market rate exceeding the natural rate can only arise when bearish speculation keeps the rate from declining. When saving exceeds investment, therefore, money leaks out of the circular spending flow into the idle balances of bear-speculators. Thus the analysis stresses declining velocity rather than endogenously declining money stock. At this stage of the development of Keynesian economics, the banks are already edging out of the theoretical field of vision and the original connection of natural rate theorizing with criteria for neutral money is by and large severed.

The model of the *Treatise* still assumes that, when market rate exceeds the natural rate, the resulting excess supply of present goods will cause falling spot prices but not unemployment of present resources. Although the focus is on a disequilibrium process, at a deeper level the theory is still comfortingly classical. As long as the economy remains at full employment, the bear-speculators who are maintaining the disequilibrium are forced, period after period, to sell income-earning securities and accumulate cash at a rate corresponding to the difference between household saving and business sector investment. Automatic market forces, therefore, are seen to put those responsible for the undervaluation of physical capital under inexorably mounting pressure to allow correction of the market rate. And the longer those agents acting on incorrect expectations persist in obstructing the intertemporal coordination of activities, the larger the losses that they will eventually suffer.

In the *General Theory*, Keynes starts the story in the same way: investment expectations take a turn for

the worse – ‘the marginal efficiency of capital declines’; the speculative demand for money prevents the interest rate from falling sufficiently to equate *ex ante* saving with investment. But at this point the *General Theory* takes a different tack: the excess supply of present resources, which is the immediate result of the failure of intertemporal price adjustments to bring intertemporal coordination, is eliminated through falling output and employment. Real income falls until saving has been reduced to the new lower investment level.

This change in the lag-structure of Keynes's theory (‘quantities reacting before prices’) is not necessarily revolutionary by itself. But Keynes combines it with the assumption that the subsequent price adjustments will be governed, in Clower's terminology, not by ‘notional’ but by ‘effective’ excess demands. For the economy to reach a new general equilibrium, on a lower growth path, interest rates should fall but money wages stay what they are. Following the real income response, however, saving no longer exceeds investment so there is no accumulating pressure on the interest rate from this quarter; at the same time, unemployment does put effective pressure on wage rates. Interest rates, which should fall, do not; wages, which should not, do. From this point, Keynes went on to argue that nominal wage reductions would not eliminate unemployment unless, in the process, they happened to produce a correction of relative prices (an eventuality that he considered unlikely). This argument was the basis for his ‘revolutionary’ claim that a failure of saving–investment coordination could end with the economy in ‘unemployment equilibrium’.

Prior to the *General Theory*, writers in the Wicksellian tradition had generally treated ‘saving exceeds investment’ and ‘market rate exceeds natural rate’ as interchangeable characterizations of the same intertemporal disequilibrium. The basic proposition could be couched equally well in terms of quantities as in terms of prices. In the *General Theory*, Keynes moved away from this language. Constructing a model with output and employment variable in the short run was a novel task and Keynes, as the pioneer, was unsure in his handling of expected, intended and realized magnitudes. Thus his preoccupation with the ‘necessary equality’ of saving and investment (*ex post*) was to produce endless confusion over interest theory. If saving and investment are always equal, the interest rate cannot be governed by the difference between them; nor can the interest rate mechanism possibly coordinate saving and investment decisions. To Keynes, two things seemed to follow. One was the substitution of the liquidity preference theory of the interest rate for the loanable funds theory; the other was the abandonment of the concept of a ‘natural’ rate of interest (Leijonhufvud, 1981, pp. 169 ff.)

These were not innocent terminological adjustments. The brand of Keynesian economics that developed on the basis of the IS–LM model had only a shaky grasp at the best of times of the intertemporal coordination problem originally at the heart of Keynes's theory. The Keynesian position shifted already at an early stage back to the pre-Keynesian hypothesis of money wage ‘rigidity’ as the cause of unemployment. This switched the focus of analytical attention away from the role of intertemporal relative prices (the market rate) in the coordination of saving and investment to the relationship between aggregate money expenditures and money wages. This brand of ‘Keynesian’ theory which excludes the saving–investment problem (that is, excludes the market-natural rate problem) could hardly be distinguished from Monetarism in any theoretically significant way.

Monetarism gained enormously in influence during the inflationary 1970s. But its period of dominance was brief. This was so in part because, in its New Classical form, it was both theoretically implausible and empirically weak. In part, however, it was swept aside by a wave of innovations in payments technology and in forms of short-term credit that undermined the stability of the relationship between

the money stock and income which had been the very linchpin of monetarist doctrine.
Most recently, this has led to a return to a basically Wicksellian doctrine of what monetary policy should aim to accomplish and how it should be conducted. Leading central banks are now committed to targeting the inflation rate (rather than the price level) and use the interest rate as their primary instrument for pursuing that goal. This policy doctrine has been elaborated in the book by Woodford (2003) which borrows its title from Wicksell.

See Also

- Stockholm School
- Wicksell, Johan Gustav Knut

Bibliography

Cassel, G. 1928. The rate of interest, the bank rate, and the stabilization of prices. *Quarterly Journal of Economics* 42(August), 511–29.

Hayek, F.A. 1931. *Prices and Production*. London: Routledge & Kegan Paul.

Humphrey, T.M. 1986. Cumulative process models from Thornton to Wicksell. *Federal Reserve Bank of Richmond Economic Review*, May–June, 18–25.

Keynes, J.M. 1930. *A Treatise on Money*, 2 vols. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Leijonhufvud, A. 1981. The Wicksell connection. In A. Leijonhufvud, *Information and Coordination*. New York: Oxford University Press.

Lindahl, E. 1939. *Studies in the Theory of Money and Capital*. New York: Holt, Rinehart & Winston.

Myrdal, G. 1939. *Monetary Equilibrium*. Edinburgh: William Hodge.

Palander, T. 1941. On the concepts and methods of the Stockholm School. *International Economic Papers* No. 3, London: Macmillan, 1953.

Robertson, D.H. 1926. *Banking Policy and the Price Level*. New York: Augustus M. Kelley, 1949.

Taylor, J.B., ed. 1999. *Monetary Policy Rules*. Chicago: University of Chicago Press.

Wicksell, K. 1898. *Interest and Prices*. New York: Augustus M. Kelley, 1962.

Wicksell, K. 1906. *Lectures on Political Economy*, vol. 2. London: Routledge & Kegan Paul, 1934.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

Leijonhufvud, Axel. "natural rate and market rate of interest." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000021> doi:10.1057/9780230226203.1164

The New Palgrave Dictionary of Economics Online

natural rate of unemployment

Michael J. Pries

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Milton Friedman defined the natural rate of unemployment as the level of unemployment that resulted from real economic forces, the long-run level of which could not be altered by monetary policy. Macroeconomic policymakers continue to view the natural rate as a key benchmark due to the belief that monetary policy can counter short-run deviations of the unemployment rate from the natural rate. It is important, however, that policymakers focus as much attention on understanding the real determinants of the natural rate, and the policies that can affect it, as they do on trying to identify and counteract deviations from it.

Keywords

American Economics Association; demography; Friedman, M.; inflationary expectations; labour supply; natural rate of unemployment; Phillips curve; rational expectations; real business cycles; search models of unemployment; Taylor rule; unemployment insurance; unemployment–inflation tradeoff; wage rigidity

Article

In his 1968 presidential address to the American Economics Association, Milton Friedman famously defined the natural rate of unemployment as

> … the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on. (1968, p. 8)

This definition is incomplete, however, because it conspicuously lacks any mention of inflation. A more complete definition emerges from the remainder of Friedman's presidential address, in which he extensively examined the relationship between the unemployment rate and inflation. He argued that, whereas the natural rate of unemployment is determined by the real factors described in the passage quoted above, deviations from the natural rate are monetary phenomena: ‘I use the term “natural” for the same reason Wicksell did – to try to separate real forces from monetary forces’ (Friedman, 1968, p. 9).

The unemployment – inflation trade-off

Friedman's ‘natural rate hypothesis’ maintained that ‘… there is a ‘natural rate of unemployment’ which is consistent with the real forces and with accurate perceptions; unemployment can be kept below that level only by an accelerating inflation; or above it only by accelerating deflation’ (Friedman, 1976, p. 458). This view of the relationship between the unemployment rate and inflation grew out of the experiences of the previous decades. In 1958, Phillips had observed a negative empirical relationship between the unemployment rate and the growth rate of wages (Phillips, 1958). Understanding that high wage growth would ultimately translate into inflation, policymakers believed that there was a stable trade-off between unemployment and inflation that they could exploit. In other words, monetary and fiscal policy could be used to drive down unemployment at the cost of a certain degree of inflation. Experience showed, however, that the relationship was not stable. As individuals started to anticipate the inflation that resulted from attempts to exploit the trade-off, stimulative policy ceased to lower unemployment. Consequently, the Phillips curve appeared to have shifted outward, with higher inflation accompanying higher unemployment.

Friedman provided an explanation for this apparent shift. Over the long run, there is an unemployment rate determined by real factors that cannot be affected by monetary policy: the natural rate. In the short run, unanticipated inflation can temporarily push the unemployment rate below its natural rate. If workers do not perceive the higher inflation, then they will respond to higher nominal wages by increasing labour supply; similarly, employers who do not immediately perceive the higher inflation will respond to a higher price for their product by demanding more labour. This temporarily lowers unemployment, but the unemployment rate returns to its natural level when workers and employers begin to perceive the inflation. As emphasized in the literature on rational expectations (for example, Lucas, 1973) that followed Friedman, inflation has no impact on real variables like the unemployment rate once individuals have already built the level of inflation into their expectations. In other words, as expectations about inflation change, the Phillips curve shifts.

Although the absence of any long-run trade-off between inflation and unemployment has gained wide acceptance, the possibility of a short-run trade-off has kept the natural rate of unemployment at the centre of policymaking. In particular, policy rules such as the Taylor rule (see Taylor, 1999) maintain that central banks can stabilize the inflation rate by assessing where the economy stands relative to economic benchmarks such as the natural rate of unemployment, ‘potential output’, or the ‘natural rate of interest’. When unemployment is high relative to the natural rate, and when output is below potential output, the policy rules call for stimulative monetary policy.

However, several important questions arise when one contemplates the usefulness of the natural rate of unemployment as a policy benchmark. First, although the natural rate clearly cannot be observed

directly, can it be estimated with enough accuracy to be useful for policy? Or do movements in the natural rate itself make it too difficult to distinguish the natural rate and deviations from the natural rate in a sufficiently timely manner to be useful for policymakers? Second, rather than focusing so much on deviations from the natural rate, should policymakers also focus on policies that would alter the natural rate, either at low frequencies or perhaps even at business cycle frequencies? What would those policies be?

Identifying the natural rate

Although the natural rate is often simplistically described as the long-run average unemployment rate, economists widely recognize that this rate varies over time. Friedman (1968, p. 9) was clear on this point:

> To avoid misunderstanding, let me emphasize that by using the term ‘natural’ rate of unemployment, I do not mean to suggest that it is immutable and unchangeable. On the contrary, many of the market characteristics that determine its level are man-made and policy-made.… Improvements in employment exchanges, in availability of information about job vacancies and labor supply, and so on, would tend to lower the natural rate of unemployment.

Friedman (1968, p. 10) further argued that the mutability of the natural rate of unemployment significantly reduces its policy usefulness:

> What if the monetary authority chose the ‘natural’ rate – either of interest or unemployment – as its target? One problem is that it cannot know what the ‘natural’ rate is. Unfortunately, we have as yet devised no method to estimate accurately and readily the natural rate of either interest or unemployment. And the ‘natural’ rate will itself change from time to time.

Since Friedman's work, however, economists have achieved additional understanding of some of the factors that contribute to low-frequency fluctuations in the natural rate of unemployment. It is now generally understood that demographic changes can have a significant impact on the natural rate of unemployment (see Shimer, 1998). For instance, young workers experience substantially more job turnover than more experienced workers, with the spells between jobs often spent in unemployment. Accordingly, when younger workers make up a larger fraction of the workforce (as they did in the 1970s when the baby boom generation entered the workforce in significant numbers), unemployment will be higher on average. Nevertheless, it is not clear whether this greater understanding of the factors that affect the natural rate can be translated into an estimate of the natural rate that is accurate enough to be useful for policy. Often changes in the natural rate can only be detected with a significant lag, after which time a policy response may actually increase volatility by causing the economy to overshoot its target.

Further complicating the question of the natural rate's usefulness as a policy benchmark is the question

of whether even higher-frequency (that is, business cycle) fluctuations in the unemployment rate could in fact represent movements in the natural rate. For example, modern search theory views unemployment fluctuations at business cycle frequencies as movements in the natural rate, in the sense that they result from real rather than monetary forces. Evidence from data on job flows shows that jobs are constantly being reallocated across firms, industries, geographical regions, and so on (see Davis, Haltiwanger and Schuh, 1996). Moreover, periods of above-average unemployment rates tend to coincide with an increased level of this reallocative activity. In this sense, unemployment rate fluctuations at business cycle frequencies can be viewed as the outcome of real phenomena of the type described in Friedman's famous quote – that is, as cyclical movements in the natural rate.

This emphasis on the real determinants of movements in the unemployment rate is part of the broader view that a significant portion of economic fluctuations reflects real factors as opposed to monetary phenomena. The vast real business cycle literature has explored this proposition since the seminal paper by Kydland and Prescott (1982). Hall (2005b) argues that real fluctuations, and the difficulty of distinguishing them from monetary phenomena, render useless the various benchmark concepts such as the natural rate of unemployment, potential output, and the equilibrium real interest rate.

Optimality of the natural rate and policies to alter it

If real sources of unemployment fluctuations are in fact as important as monetary sources, then the proper response by monetary policymakers to the fluctuations is much less clear. However, even if unemployment fluctuations are primarily driven by real factors, it would be incorrect to conclude that either the level or fluctuations of the natural rate are optimal. Accordingly, there may be a role for policy to improve welfare by affecting the natural rate (either at low frequencies or perhaps even at high frequencies). This suggests that research on the optimality of the natural rate, and on policies that can affect it, is as important as research aimed at detecting and proposing policies to counteract deviations from it.

The idea that the natural rate can be either too high or too low has been a primary focus of modern search and matching models of the labour market. In those models, the process whereby workers and firms meet may be subject to various externalities. When a worker chooses to search for a job, it has a positive externality on the probability that employers will find a suitable worker and a negative externality on the probability that other workers will find a job. Employers' search decisions cause similar externalities.

Hosios (1990) analyses the conditions under which, in a broad class of search and matching models, the various externalities result in an unemployment rate that is either too high or too low. He finds that in general there is no economic force that draws the unemployment rate towards its optimal level. One suspects that the wage might play that role. When employers decide whether to open job vacancies (the number of which ultimately determines the unemployment rate), they anticipate the wages that they will have to pay and the profits that they will earn when they form an employment relationship. However, the level of those wages and the resulting profits are determined after the fact by bargaining between workers and firms who have been matched, and who are not contemplating the impact that their bargain has on firms posting new vacancies. If the wages that result from bargaining are too low (high), firms anticipate this and create many (few) vacancies, and the unemployment rate is inefficiently low (high).

As a complement to this more theoretical examination of the optimal level of the natural rate, there is a more applied literature that tries to understand cross-country differences (particularly between continental Europe and the United States) in the average unemployment rate and how those differences relate to various policies. For example, Hopenhayn and Rogerson (1992) examine the impact of firing costs on unemployment and on productivity. They find that, in addition to increasing average unemployment, firing costs reduce productivity by impeding the reallocation of workers towards more productive employers. Ljungqvist and Sargent (1998) argue that the interaction between generous unemployment insurance in many western European countries and an increased turbulence in labour markets can explain the secular rise in European unemployment rates relative to the US rate over the last several decades.

In addition to this work on the determinants of average unemployment rates in the long run, recent work has also focused on trying to better understand the sources of non-monetary movements in the unemployment rate over the business cycle, and whether they are efficient. What real factors contribute to spikes in unemployment, and why is the subsequent recovery so slow? Pries (2004) argues that the slow recovery occurs because workers who lose their job in the initial spike may pass through several short-lived jobs, and several intervening unemployment spells, before ultimately settling into more stable employment. In this environment, policies that try to accelerate a recovery may be counterproductive if they encourage worker–firm pairs to hang on to low-quality matches.

Shimer (2005), on the other hand, argues that the slow recovery of the unemployment rate during economic downturns results from a significant reduction in posted vacancies and, consequently, a decline in workers' job-finding rates. More research is needed to understand the causes of the decline in posted vacancies. The canonical Mortensen–Pissarides (1994) matching model, in which wages are flexibly renegotiated as part of a Nash bargaining solution, struggles to produce a sizeable decline in vacancies during recessions. In the model, wages fall considerably during economic downturns, and the lower wages mean that firms still find it quite profitable to post vacancies. This model's failure to deliver the observed cyclicality in vacancies leads Hall (2005a) to suggest that in fact wages are much less flexible than assumed in Mortensen–Pissarides (1994). If so, then should the fluctuations be seen as monetary in nature, and is stimulative monetary policy the correct policy response? Or are tax incentives for investment, which may spur the creation of new jobs, a better policy response? As with countercyclical monetary policy, tax incentives may take effect with a lag and exacerbate fluctuations.

Milton Friedman's assertion in 1968 that there is a natural rate of unemployment that is determined by real economic forces and is impervious to monetary policy has become relatively uncontroversial. Nevertheless, important unresolved questions about the natural rate remain. What is the optimal natural rate? To what extent do unemployment rate fluctuations reflect movements in the natural rate as opposed to deviations from it? What policies, if any, are appropriate for counteracting movements in the natural rate or deviations from it?

See Also

- Friedman, Milton
- Phillips curve
- real business cycles

- search models of unemployment
- Taylor rules

Bibliography

Davis, S., Haltiwanger, J. and Schuh, S. 1996. *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Friedman, M. 1976. Nobel lecture: Inflation and unemployment. *Journal of Political Economy* 85, 451–72.

Hall, R. 2005a. Employment fluctuations with equilibrium wage stickiness. *American Economic Review* 95, 50–65.

Hall, R. 2005b. Separating the business cycle from other economic fluctuations. In *The Greenspan Era: Lessons for the Future* Proceedings of the Federal Reserve Bank of Kansas City Symposium, August.

Hopenhayn, H. and Rogerson, R. 1992. Job turnover and policy evaluation: a general equilibrium analysis. *Journal of Political Economy* 101, 915–38.

Hosios, A. 1990. On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies* 57, 279–98.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–71.

Ljungqvist, L. and Sargent, T. 1998. The European unemployment dilemma. *Journal of Political Economy* 106, 514–50.

Lucas, R. 1973. Some international evidence on output-inflation tradeoffs. *American Economic Review* 63, 326–34.

Mortensen, D. and Pissarides, C. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61, 397–415.

Phillips, A. 1958. The relationship between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 58, 283–99.

Pries, M. 2004. Persistence of employment fluctuations: a model of recurring job loss. *Review of Economic Studies* 71, 193–215.

Shimer, R. 1998. Why is the U.S. unemployment rate so much lower? In *NBER Macroeconomics Annual*, vol. 13, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Shimer, R. 2005. The cyclical behavior of equilibrium unemployment and vacancies. *American Economic Review* 95, 25–49.

Taylor, J. 1999. *Monetary Policy Rules*. NBER Conference Report series. Chicago and London: University of Chicago Press.

How to cite this article

Pries, Michael J. "natural rate of unemployment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000024> doi:10.1057/9780230226203.1165

The New Palgrave Dictionary of Economics Online

Navier, Louis Marie Henri (1785 – 1836)

R.F. Hébert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

consumption externalities; cost–benefit analysis; demand theory; Dupuit, A.-J.; jointness of consumption; Navier, L.; Pigou, A.; public goods; public works; Samuelson, P, on public goods; subjective utility; utility measurement

Article

A French engineer and economist, Louis Marie Henri Navier was a pioneer in the construction of suspension bridges, and is also known as the creator of that branch of mechanics known as structural analysis. In his economic inquiries, he sought a practical measure of public utility that provided the springboard for Dupuit's pioneer contributions to demand theory. Orphaned at the age of nine, Navier was adopted by his great-uncle, the celebrated architect–engineer, Émiland-Marie Gauthey (1732–1806), who likely inspired his adopted son to follow in his illustrious footsteps. Navier died prematurely at the age of 51, thus cutting short a distinguished career of public service.

Navier was one of the earliest formulators of a cost–benefit rule to guide the construction of public works. His rule advocates expenditures on public works if the total benefit derived – in the form of before–after cost savings – exceeds the total recurring costs of the new construction. In choosing recurring costs over total costs as the element to be covered by tolls, Navier was showing a greater appreciation of consumption externalities than Pigou (1947, p. 3n.), who wrote more than a century later. In fact, Navier's rule is a somewhat less sophisticated version of Stephen Marglin's (1967, pp. 22–4) ‘myopic rule’ of public investment.

Navier's rule was the proximate cause of Dupuit's innovative attempt to establish demand based on subjective utility. Dupuit (1844) objected to Navier's attempt to measure utility on two grounds: (*a*) in competitive markets the proper measure of utility of the quantity of goods and services consumed is not the reduction of transport costs but rather the reduction of production costs; (*b*) increases in the quantity taken at lower prices do not all have the same utility, but rather take on smaller values as more is consumed. Thus, Dupuit's rule overcame the limitations of Navier's rule, and, in addition, launched the

neoclassical theory of demand. Kölm (1968) argues that, in the context of public finance, Dupuit's rule moves us closer to Samuelson's (1954, pp. 387–9) decision rule regarding public goods. However, a valid comparison of Dupuit's performance with Samuelson's must recognize that Samuelson employed a highly restrictive definition of a public good and the assumption of true consumption jointness – aspects missing from Dupuit's analysis or from Navier's.

See Also

- consumption externalities
- cost-benefit analysis
- Dupuit, Arsene-Jules-Emile Juvenal
- Pigou, Arthur Cecil
- public goods
- public works

Selected works

1832. De l'exécution des travaux publics, et particulièrement des concessions. *Annales des Ponts et Chaussées: Mémoires et Documents*, 1 ser. 3, 1–31.

1835. Note sur la comparison des avantages respectifs de diverse lignes de chemins de fer, et sur l'emploi des machines locomotives. *Annales des Ponts et Chaussées: Mémoires et Documents*, 1 ser. 9, 129–79.

Bibliography

Coronio, G. 1997. *250 ans de L'École des Ponts en cent portraits*. Paris: Presses de l'école des Ponts et Chaussées, 86–7.

Dupuit, J. 1844. On the measurement of the utility of public works. *Annales des ponts et chaussées*, 2d ser. 8, 332–75. Trans. R. Barback, *International Economic Papers* 2 (1952), 83–110.

Ekelund, R., Jr. and Hébert, R. 1978. French engineers, welfare economics, and public finance in the nineteenth century. *History of Political Economy* 10, 636–68.

Ekelund, R., Jr. and Hébert, R. 1999. *Secret Origins of Modern Microeconomics: Dupuit and the Engineers*. Chicago: University of Chicago Press.

Etner, F. 1987. *Histoire du calcul économique en France*. Paris: Economica.

Hébert, R. 1994. Fondements et développements de l’économie publique. *Revue Du Dix-Huitième Siècle* 26, 37–49.

Kölm, S.-C. 1968. Léon Walras' correspondence and related papers: the birth of mathematical economics. *American Economic Review* 58, 1330–41.

Marglin, S. 1967. *Public Investment Criteria: Studies in the Economic Development of India*. Cambridge, MA: MIT Press.

Pigou, A. 1947. *A Study in Public Finance*, 3rd edn. London: Macmillan.

Samuelson, P. 1954. The pure theory of public expenditures. *Review of Economics and Statistics* 36, 387–9.

How to cite this article

Hébert, R. F. "Navier, Louis Marie Henri (1785–1836)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000158> doi:10.1057/9780230226203.1166

The New Palgrave Dictionary of Economics Online

neighbours and neighbourhoods

Ingrid Gould Ellen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

census data; Chicago School of Sociology; employment; externalities; hedonic regression analysis; house prices; human capital; neighbourhood; networks; residential segregation; social capital; tipping; zoning

Article

The concept of neighbourhood has long been a topic of popular discourse and a subject of academic interest. Despite this attention, there is little agreement on what the term ‘neighbourhood’ means. The *American Heritage Dictionary* (Pickett, 2000) simply defines a neighbourhood as ‘a district or an area with distinctive characteristics’.

‘A district or an area’ is not very specific, and social scientists (outside of economics) have struggled for decades to define more precisely the geographic boundaries of neighbourhoods (Keller, 1968). Beyond the fact that neighbourhoods are sub-jurisdictional units, characterized by some degree of social cohesion, there is no accepted standard. The report prepared by the National Commission on Neighborhoods (1979, p. 7) stated that ‘each neighborhood is what the inhabitants think it is’. Yet the evidence suggests that such subjective perceptions vary greatly (Keller, 1968).

For economists, who generally focus on externalities when considering neighbourhoods, an individual's neighbourhood should theoretically extend as far as the individuals or facilities that affect her satisfaction with the community (Segal, 1979; Galster, 1986). In practice, economists and other social scientists studying neighbourhoods in the United States typically use census tracts to proxy for neighbourhoods. Including between 2,500 and 8,000 people on average, census tracts are close in size to what most envision as a neighbourhood and have the practical advantage of supplying demographic and economic data from the decennial census. In Australia and Europe, census data are typically available at sub-jurisdictional levels, defined by electoral wards or postcodes, and in some cases, smaller enumeration or collection districts (Overman, 2002; Bolster et al., 2004; Drever, 2004). Increasingly, researchers in the United States and Europe are able to link individual census data and other national

household surveys to geographic identifiers, and they are experimenting with smaller and more flexible neighbourhood definitions (Bolster et al., 2004; Ioannides, 2004; Bayer, Ross and Topa, 2005).

As for the term ‘distinctive characteristics’, economists identify several types of goods or services delivered by neighbourhoods. First, neighbourhoods offer distinct physical amenities, ranging from the style and condition of local housing to the number and quality of local parks. Second, neighbourhoods embody a particular set of ‘neighbours’, who have a distribution of income, human capital, and racial characteristics. Third, neighbourhoods often approximate local public service delivery areas such as attendance zones for public elementary schools, which often vary significantly in performance, even within the same jurisdictions. Fourth, neighbourhoods provide accessibility to shopping and employment opportunities. Finally, economists increasingly view neighbourhoods as possessing a stock of social capital, or norms and networks that facilitate interaction and can help residents work together to address problems like crime (Glaeser, 2000).

Social scientists have been preoccupied with the evolution and nature of neighbourhoods for decades. Modern academic discourse on neighbourhoods has its roots in the Chicago School of the 1920s. These University of Chicago sociologists hypothesized that cities naturally grow outward in a series of concentric rings. Through this growth, a neighbourhood life cycle emerges, from richer residents to poorer, as more affluent residents opt for newer, less dense and quieter areas (Park, Burgess and McKenzie, 1925).

Economists came later to the study of neighbourhoods, also initially drawn by an interest in the transition of neighbourhoods from high to low income and from predominantly white to predominantly minority residents. Muth (1972) and Sweeney (1974) propose variations of the filtering model, which, similar to the Chicago School theory, posits that neighbourhoods decline because, as their housing ages and deteriorates, higher-income residents exit, opting for newer neighbourhoods with newer housing. Other economists focused instead on the role of racial or class preferences in driving neighbourhood change (Bailey, 1959). In his simple, elegant model, Schelling (1971) shows that, if households care about the composition of their neighbours, then small changes in demographic make-up can lead to the rapid tipping of a neighbourhood from one group to another.

Another strand of economic literature examines the relationship between various neighbourhood attributes and housing prices, typically using hedonic regression analysis (Kain and Quigley, 1970; Bartik and Smith, 1987). Mills and Hamilton (1994) argue that economists have historically failed to identify the external effects of housing quality and neighbourhood conditions. But more recent research finds strong evidence that housing prices are lower in areas with higher crime, lower-quality schools, dilapidated housing and vacant lots, and fewer homeowners (Grieson and White, 1989; Black, 1999; Coulson, Hwang and Imai, 2003; Schwartz, Susin and Voicu, 2003; Schwartz et al., 2005). As for the impacts of racial composition, more recent papers find that a neighbourhood's housing prices are negatively correlated with the percentage of black residents (Yinger, 1976; Kiel and Zabel, 1996; Myers, 2004).

Finally, following Wilson (1987), economists have more recently turned to the study of how neighbourhoods and social interactions in them influence resident behaviour and outcomes.

See Also

- ghettoes
- residential segregation
- spatial mismatch hypothesis
- urban housing demand

Bibliography

Bailey, M. 1959. Note on the economics of residential zoning and urban renewal. *Land Economics* 35, 288–92.

Bartik, T. and Smith, K. 1987. Urban amenities and public policy. In *Handbook of Regional and Urban Economics*, vol. 2, ed. E. Mills. Amsterdam: North-Holland.

Bayer, P., Ross, S. and Topa, G. 2005. Place of work and place of residence: informal hiring networks and labor market outcomes. Working Paper 11019. Cambridge, MA: NBER.

Black, S. 1999. Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics* 114, 579–99.

Bolster, A., Burgess, S., Johnston, R., Jones, K., Propper, C. and Sarker, R. 2004. Neighbourhoods, households, and income dynamics: a semi-parametric investigation of neighbourhood effects. Research Discussion Paper 4611. London: Centre for Economic Policy.

Coulson, N., Hwang, S.-J. and Imai, S. 2003. The value of owner-occupation in neighborhoods. *Journal of Housing Research* 13(2), 153–74.

Drever, A. 2004. Separate spaced, separate outcomes? Neighbourhood impacts on minorities in Germany. *Urban Studies* 41, 1423–39.

Galster, G. 1986. What is a neighborhood? *International Journal of Urban and Regional Research* 10, 243–63.

Glaeser, E. 2000. The future of urban research: non-market interactions. *Brookings-Wharton Papers on Urban Affairs* 2000, 101–38.

Grieson, R. and White, J. 1989. The existence and capitalization of neighborhood externalities: a reassessment. *Journal of Urban Economics* 25, 68–76.

Ioannides, Y. 2004. Neighborhood income distributions. *Journal of Urban Economics* 56, 435–57.

Kain, J. 1968. Housing segregation, negro unemployment, and metropolitan decentralization. *Quarterly Journal of Economics* 82, 175–97.

Kain, J. and Quigley, J. 1970. Measuring the value of housing quality. *Journal of the American Statistical Association* 5, 532–48.

Katz, L., Kling, J. and Liebman, J. 2001. Moving to opportunity in Boston: early results of a randomized mobility experiment. *Quarterly Journal of Economics* 116, 607–54.

Keller, S. 1968. *The Urban Neighborhood: A Sociological Perspective*. New York: Random House.

Kiel, K. and Zabel, J. 1996. House price differentials in U.S. cities: household and neighborhood racial effects. *Journal of Housing Economics* 5, 143–65.

Mills, E. and Hamilton, B. 1994. *Urban Economics*. 5th edn. New York: HarperCollins College Publishers.

Muth, R. 1972. A vintage model of the housing stock. *Regional Science Association Papers Proceedings* 30(2), 141–56.

Myers, C. 2004. Discrimination and neighborhood effects: understanding racial differences in U.S. house prices. *Journal of Urban Economics* 56, 279–302.

National Commission on Neighborhoods. 1979. *People, Building Neighborhoods*. Final Report to the President and the Congress of the United States. Washington, DC: Government Printing Office.

Overman, H. 2002. Neighbourhood effects in large and smallneighbourhoods. *Urban Studies* 39, 117–30.

Park, R. Burgess, E. and McKenzie, R., eds. 1925. *The City*. Chicago: University of Chicago Press.

Pickett, J., ed. 2000. *The American Heritage® Dictionary of the English Language*, 4th edn. Boston: Houghton Mifflin Company.

Schelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1, 143–86.

Schwartz, A., Ellen, I., Voicu, I. and Schill, M. 2005. The external effects of place-based, subsidized housing. Working paper. New York: Furman Center for Real Estate and Urban Policy, New York University.

Schwartz, A., Susin, S. and Voicu, I. 2003. Has falling crime driven New York City's real estate boom? *Journal of Housing Research* 14, 101–36.

Segal, D. 1979. Introduction. In *The Economics of Neighborhood*, ed. D. Segal. New York: Academic Press.

Sweeney, J. 1974. A commodity hierarchy model of the rental housing market. *Journal of Urban Economics* 1, 288–323.

Wilson, W. 1987. *The Truly Disadvantaged: The Inner-City, the Underclass and Public Policy*. Chicago: University of Chicago Press.

Yinger, J. 1976. Racial prejudice and racial residential segregation in an urban model. *Journal of Urban Economics* 3, 383–96.

How to cite this article

Ellen, Ingrid Gould. "neighbours and neighbourhoods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000139> doi:10.1057/9780230226203.1167

The New Palgrave Dictionary of Economics Online

Nemchinov, Vasily Sergeevich (1894 – 1964)

M.C. Kaser

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

linear programming; mathemetical economics; Nemchinov, V. S.

Article

Born the son of a State Bank messenger in Grabovo, Russia, on 2 January 1894; died in Moscow on 5 November 1964. Nemchinov graduated from the Moscow Commercial Institute between the February and October Revolutions of 1917, but joined the Communist Party only in 1940 on appointment as Director of the K.A. Timiryazev Agricultural Institute, the Statistics Faculty of which he had headed since 1928. He showed courage in prohibiting from his Institute the pseudo-genetics (‘Michurinism’) of T.D. Lysenko, but when at Stalin's instigation mainstream genetics were condemned in 1948 he was forced from the directorship. The Academy of Sciences (to which he had been elected in 1946) then made him chairman of its Council for the Study of Productive Resources, a post retained (with a chair at the party's Academy of Social Sciences) until his fatal illness. In 1958 he established the first group in the USSR to study mathematical economics (from 1963 the Central Economic Mathematical Institute) and was posthumously awarded a Lenin Prize for elaborating linear programming and economic modelling for the USSR.

The research embodied in Nemchinov (1926; 1928) was distorted to justify Stalin's coercion of the peasantry: his data on rural social stratification gave cover to ‘liquidation of the kulaks as a class’ (though Nemchinov had avoided the term ‘kulak’); his measurement of absolute gross harvest (Nemchinov, 1932) was used to extort deliveries from collective farms. As soon as Stalin died, Nemchinov campaigned for the publication of official statistics and for more sophisticated techniques to utilize them – cybernetics had been damned as a pseudo-science serving capitalist interests. His organization of experimental national and regional input–output tables led him to question the meaningfulness of administered pricing, and his last book (1962) sought, as his widow put it (Nemchinova, 1985, pp. 202–21), ‘a broad-based system of social valuations … as a single, internally consistent set of values’.

Selected works

1926. O statisticheskom izuchenii klassovogo rassloenniya derevni [On the statistical study of rural class stratification]. *Bulleten' Ural'skogo oblastnogo statisticheskogo upravleniya* [Bulletin of the Urals Regional Statistical Administration] 1. Reprinted in *Selected Works*, vol. 1.

1928. Opyt kalssifikatsii krest'yanskikh khozyaistv [Experience from the classification of peasant households]. *Vestnik statistiki* [Statistical bulletin] 1. Reprinted in *Selected Works*, vol. 1.

1932. Vyborochnye izmereniya urozhainosti [Sampling measurement of yields]. *Narodnoe khozyaistvo SSSR* [National economy of the USSR] 5–6. Reprinted in *Selected Works*, vol. 1.

1962. *Ekonomiko-matematicheskie metody i modeli* [Methods and models of mathematical economics]. Moscow: Sotsegiz. 2nd (posthumous) edn, 1965. Reprinted in *Selected Works*, vol. 3.

1967–9. *Izbrannye proizvedeniya* [Selected Works]. 6 vols. Moscow: Izdatel'stvo Nauka.

Bibliography

Nemchinova, M.B. 1985. The scientific work of Vasily Sergeevich Nemchinov (on the 90th anniversary of his birth). *Matekon. Translations of Russian and East European Mathematical Economics* 21(2) (1984–5), 3–25; translation of an article in *Ekonomika i matematicheskie metody* [Economics and mathematical methods] 20(1) (1984).

How to cite this article

Kaser, M. C. "Nemchinov, Vasily Sergeevich (1894–1964)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000037> doi:10.1057/9780230226203.1168

The New Palgrave Dictionary of Economics Online

neoclassical growth theory (new perspectives)

Rodolfo E. Manuelli
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The neoclassical growth model captures the basic trade-off between saving and investment. It has proven to be a useful tool to study development paths, and the interactions of technology shocks, money and fertility choices with growth.

Keywords

competitive equilibrium; convexity; endogenous growth; fertility; human capital accumulation; infinite horizons; innovation; marginal rate of transformation; neoclassical growth theory; optimal development paths; optimal quantity of money; optimal taxation; population growth; recursive equilibrium; representative agent; Solow–Swan growth model; technical change; technology shocks; transversality condition; turnpike property

Article

This article complements neoclassical growth theory. It discusses some developments of the neoclassical growth theory that endogenize the saving rates.

Infinite horizons

The planning problem

The standard neoclassical growth model assumes that the planning horizon is infinite. One justification is that forward-looking parents act ‘as if’ they were to live forever. To see this, assume that each individual lives for one period and has exactly one descendant. The utility of a member of generation 0 is given by

$$U_0 = u(c_0) + \beta U_1, \tag{1}$$

where u is an increasing, continuous and concave function of consumption at time t, c_t. Iterating on this expression yields

$$U_0 = \sum_{t=0}^{\infty} \beta^t u(c_t),\ \beta = \frac{1}{1+\rho},\ \rho > 0, \tag{2}$$

which shows that altruism implies that the effective planning horizon for each individual is infinite.
In the simplest one-sector version of the model, the technology is summarized by

$$c_t + x_t \leq zf(k_t),\ t = 0, 1, \ldots \tag{3a}$$

$$k_{t+1} \leq (1 - \delta_k)k_t + x_t,\ t = 0, 1, \ldots \tag{3b}$$

$$k_0 > 0,\ given, \tag{3c}$$

where k_t is the stock of capital per person available at the beginning of period t, x_t is gross investment, z is a measure of productivity, and δ_k is the depreciation rate of capital. The function f is assumed to be increasing, continuous and strictly concave.
The planning problem corresponds to the maximization of the utility criterion (2), subject to the feasibility constraints (3). The analysis of this problem was initially carried out by Ramsey (1928), Cass (1965) and Koopmans (1965). A thorough analysis of the model can be found in Stokey and Lucas (1989).
The model has sharp predictions for the properties of an optimal development path. The relevant first-order conditions (in the interior case) require that the marginal rate of substitution between consumption at time t and $t+1$ equal the marginal rate of transformation,

$$\frac{u(c_t)}{\beta u(c_{t+1})} = 1 - \delta_k + zf'(k_{t+1}),\ t = 0, 1, \ldots,$$
(4)

and a transversality condition which is naturally interpreted as requiring that the value, at time 0, of the stock of capital at time T+1 converge to 0 as $T \to \infty$. Formally, the condition is

$$\lim_{T \to \infty} \beta^T u'(c_T) k_{T+1} = 0.$$

Some properties of the solution are as follows:

1. 1. There exists a unique steady state; that is, there are constant sequences of consumption, investment and capital that satisfy (3) (except at time 0) and (4). From (4) it follows that, in the steady state, the marginal product of capital equals the sum of the discount rate, ρ, and the depreciation factor, δ_k,

$$\rho + \delta_k = zf'(k^*),$$
(5)

which determines capital per worker. The steady state level of consumption is given by

$$c^* = zf(k^*) - \delta_k k^*.$$
(6)

2. 2. For any k_0>0, the solution to the problem converges to the steady state. Convergence is monotone.
3. 3. In general, the savings rate – defined as $1 - c_t / zf(k_t)$ – is not constant, or even monotone. This distinguishes the optimal neoclassical growth model from the Solow–Swan version that assumes exogenous (and generally constant) saving rates.

The steady state is the model's prediction about the long-run levels of capital, consumption and investment. From the point of view of a theory of growth there are some interesting results:

1. 1. The steady state level of output per worker is independent of the form of the utility function.
2. 2. If a fixed level of government consumption, g, is introduced in the model, the steady state condition (5) remains unchanged. The new steady state level of consumption is $c^* = zf(k^*) - \delta_k k^* - g$. Thus the

model predicts that, in the long run, permanent increases in government spending have no impact on output per worker, and they crowd out private consumption one for one, with no effect on investment.

The basic model has been extended in many dimensions. In the case of multiple sectors, existence of optimal paths has been established very generally. Burmeister (1980) provides conditions for the existence and uniqueness of steady states with many capital goods.

The properties of optimal paths depend on the specification of the economic environment. In the case of a discounted twice differentiable utility and dominance diagonal of a matrix of first-order conditions, it is possible to show that the turnpike property holds (see the excellent survey in McKenzie, 1986). Formally, McKenzie shows that if $\{k_t\}$ is an optimal path starting from k_0, then, for every capital stock k_0' near k_0 the associated unique optimal path converges exponentially to $\{k_t\}$.

The monotonicity properties of optimal paths do not extend to the multicapital or multisector case. In general, optimal paths can display cycles (see Burmeister, 1980) and even more complex behaviour.

To illustrate this let the feasible technology set be described as

$$c_t \leq T(k_t, k_{t+1}),$$

and let the (indirect) utility function over capital stocks be

$$v(k_t, k_{t+1}) \equiv u(T(k_t, k_{t+1})).$$

With this notation, the planning problem reduces to

$$\max_{\{k_{t+1}\}} \sum_{t=0}^{\infty} \beta^t v(k_t, k_{t+1}).$$

Let's denote a candidate solution by a function g where

$$k_{t+1} = g(k_t).$$

Boldrin and Montrucchio (1986) showed that – under standard conditions – given any twice differentiable function g, there exists a pair (v, β) so that the associated planner's problem has g as its optimal policy function. Since g can exhibit arbitrary complex dynamics, the result shows that in order to endow the theory with predictive power it is necessary to ‘force’ the chosen specification to quantitatively match moments of the (actual) economy

under study. Most recent research using the neoclassical growth model disciplines the choices of functional forms and parameters by requiring that they predict behaviour consistent with the empirical evidence.

Equilibrium growth

Even though the analysis of the growth model was motivated by normative considerations, under the stated assumptions the planner's solution of the growth model coincides with the competitive equilibrium of the economy. The argument – using the traditional definition of a competitive equilibrium – follows from Debreu (1954). In macro applications – the field in which the model has proved to be most useful – it is more natural to define a competitive equilibrium using the notion of recursive equilibrium first introduced by Prescott and Mehra (1980).

In order to account for wages, let the production function be given by

$$y \leq zF(k, n),$$

where F is concave and homogeneous of degree one, and it satisfies

$$f(k) \equiv F(k, 1).$$

Even though there are many alternative ways of defining an equilibrium, it is easiest to consider the case in which there are rental spot markets for capital and labour, and the households trade consumption, labour and capital services and one-period bonds. The problem solved by the representative household is

$$\max \sum_{t=0}^{\infty} \beta^t u(c_t)$$

subject to

$$b_{t+1} + c_t + x_t \leq w_t n_t + q_t k_t + (1 + r_t) b_t,\ t = 0, 1, \ldots k_{t+1} \leq (1 - \delta_k) k_t + x_t,\ t = 0, 1, \ldots 0 \leq n_t \leq 1,\ t = 0, 1, \ldots$$

and the initial conditions, $[(1+r_0)b_0, k_0]$, given. As stated, this problem has no solution since the budget set is unbounded. Different alternative assumptions on how to deal with debt at infinity have been used to guarantee that the problem is well defined. The most general specification is to rule out Ponzi games by imposing that the present value of debt be nonnegative. Formally, any solution must satisfy

$$\lim_{T \to \infty} \prod_{j=0}^{T} \frac{1}{1 + r_j} b_{T+1} \geq 0.$$

which is the analogue – in the market setting – of the transversality condition in the planning problem.
Firms solve a static problem

$$\max_{k_t, n_t} zF(k_t, n_t) - q_t k_t - w_t n_t.$$

A competitive equilibrium is an allocation $[\{c_t\}, \{n_t\}, \{x_t\}, \{k_{t+1}\}]_{t=0}^{\infty}$, a price system $[\{q_t\}\{w_t\}, \{r_{t+1}\}]_{t=0}^{\infty}$ and a sequence of bond holdings $\{b_{t+1}\}_{t=0}^{\infty}$, such that:

1. 1. Given the price system, the allocation solves the maximization problems of households and firms.
2. 2. Markets clear.

Given that Debreu (1954) shows that the solution to the planner's problem can be decentralized as a competitive equilibrium, the first-order conditions (on the assumption of interiority and differentiability) corresponding to the maximization of utility and profits imply that equilibrium prices (as a function of the planner's allocation) are given by

$$q_t = zf'(k_t),$$
(7a)

$$w_t = zf(k_t) - k_t zf'(k_t),$$
(7b)

$$r_{t+1} = q_{t+1} - \delta_k.$$
(7c)

It is possible to state the implications of the neoclassical growth model more intuitively using equilibrium prices. The consumer's optimal choice between consumption and saving requires that

$$\frac{u(c_t)}{\beta u(c_{t+1})} = 1 + r_{t+1},$$

that is, that the marginal rate of substitution between present and future consumption equal to (gross) interest rate. Optimality on the part of firms requires that the marginal product of labour be equal to the wage rate and that the marginal product of capital equal the cost of capital, $r_t + \delta_k$.

The basic neoclassical growth model (and some of the extensions mentioned) has had a significant impact on how economists view the process of development and the role of markets supporting optimal development paths. It is clear that there is nothing special about dynamic problems that make it more (or less) likely for competitive markets to fail to deliver optimal allocations. In the basic model of this note, Theorems I and II of welfare economics apply.

Applications

Some of the most notable extensions are as follows.

Technology shocks

Brock and Mirman (1972) studied a version of the neoclassical growth model in which the representative agent maximizes the expected value of the discounted flow of utility, and the technology is as in the deterministic growth model except that the technology level, z, is replaced by a stochastic process $\{z_t\}$. Brock and Mirman assumed that the process $\{z_t\}$ is i.i.d. They established the existence of a solution and they showed that, under standard concavity assumptions, the resulting stochastic process of the capital stock has a unique invariant measure, which is the stochastic analogue of the steady state in the deterministic version of the problem. They also showed that the optimal policy function which determines k_{t+1} as a function of k_t and z_t is monotone. The results were extended to the case of serially correlated shocks by Donaldson and Mehra (1983).

This research has provided the theoretical foundations for a large literature that analyses the impact of economic fluctuations on savings and growth. When the model is extended to include an elastic labour supply, this is a natural setting in which to study cyclical movements of employment. For an introduction to this literature see Cooley (1995).

Human capital and development

The neoclassical growth model, extended to allow for human capital accumulation, is a natural candidate to understand the role that technological differences play in accounting for differences in output per worker. In the standard specification – using a Cobb–Douglas specification for f – it follows that output per worker is given by

$$y = z^{1/(1-\alpha)}\bar{y}_0$$

where α corresponds to capital share, and $\bar{y}_0$ (and all the $\bar{y}_j$ in this section) is a constant. This version of the theory implies that the elasticity of output per worker with respect to z is $1/(1-\alpha)$. Since accepted estimates of α cluster around 0.33 – which, approximately, correspond to the share of national income that accrues to capital – the elasticity is estimated to be approximately 1.5. If this model is to explain the differences in output per worker between the richest and poorest countries (which are of the order of 15–20 to 1), it must assume fairly large differences in productivity that exceed the best available estimates.

Klenow and Rodríguez-Claire (1997) (see also, Bils and Klenow, 2000) consider a production function of the form

$$y = zk^{\alpha}(h^{e})^{1-\alpha},$$

and they use the specification $h^e=e^{\psi s}$, where s corresponds to years of schooling to estimate the role of human capital. In this case, the equilibrium level of output per worker is given by

$$y = z^{1/(1-\alpha)}e^{\psi s}\bar{y}_1$$

Klenow and Rodríguez-Claire use data to determine s and ψ. To highlight the role of productivity differences, let $e^{\psi s}=z^{\nu}$. Output per worker is

$$y = z^{1/(1-\alpha)+\nu}\bar{y}_1.$$

Klenow and Rodríguez-Claire find that the implied ν is not large. They conclude that productivity differences account for much of the differences in output.

Manuelli and Seshadri (2007a) endogenize the human capital decision. They adopt Ben Porath's (1967) specification. In discrete time, their model assumes that human capital evolves according to

$$h_{t+1} = z_h(n_t h_t)^{\gamma_1} x_{ht}^{\gamma_2} + (1-\delta_h)h_t,$$

where $n_t h_t$ is the fraction of the available time allocated to producing human capital, and x_{ht} denotes market goods used in the production of human capital. In this setting, $h^e=(1-n)h$. It is possible to show that, in the steady state,

output per worker is given by

$$y = z^{\gamma_2/[(1-\alpha)(1-\gamma_1-\gamma_2)]}\bar{y}_2.$$

This version of the model implies that the elasticity of output with respect to the productivity parameter *z* is $\gamma_2 / [(1-\alpha)(1-\gamma_1-\gamma_2)]$. Manuelli and Seshadri use life age–earnings profile evidence to estimate that γ_1=0.63 and γ_2=0.30. This results in an elasticity of output per worker with respect to productivity of 6.5. This high elasticity implies that productivity differences have a large impact on (endogenously chosen) human capital. As a result, even small productivity differences are consistent with large variations in output per worker. The relative importance of human capital and productivity is an active area of research. More work is needed before the roles of technology and education in accounting for differences in output can be accurately estimated.

The role of taxation

The neoclassical growth model has been widely used to analyse the effect of specific tax policies and to derive properties of optimal tax systems.
Consider a version of the model in which labour is elastically supplied. Let the period utility function be given by *u* (*c*, •), where • is interpreted as leisure. In an economy in which consumption, capital income and labour income are taxed (at constant rates) it follows that the steady state is characterized by

$$\rho = (1-\tau^k)(F_k(k, n) - \delta_k) \tag{8a}$$

$$u_\ell(c, 1-n) = u_c(c, 1-n)F_n(k, n)\frac{1-\tau^n}{1+\tau^c} \tag{8b}$$

$$F(k, n) = c + \delta_k k \tag{8c}$$

$$\rho = (1 - \tau^b) r^b. \tag{8d}$$

From a formal point of view the system of eqs (8) contains four equations in four unknowns. Let $\Phi(c, \ell) = u_\ell(c, 1 - n) / u_c(c, 1 - n)$, and assume that $\Phi(c, \bullet)$ is increasing in c and decreasing in •. In this case, it is possible to show that:

1. 1. An increase in the tax rate of capital income, τ^k, decreases the amount of capital, but has ambiguous effects on employment.
2. 2. An increase in tax rate on labour income (consumption) decreases both k and n.

The effect of taxes on employment and growth is a subject that continues to receive substantial attention.
In the mid-1980s Chamley (1986) and Judd (1985) asked the following question: If a government has to finance a given (say, constant) stream of consumption, and if the only available taxes are distortionary taxes (for example, in the previous example, set τ^c=0 and add government spending to (8c)), how should those taxes be chosen? Chamley and Judd showed that the optimal tax system is such that, in the steady state, capital income taxes are zero while labour income taxes are positive.
This result is delicate in the sense that it does not hold if some of the assumptions are slightly modified. For example, if the function F is strictly concave, and pure profits cannot be taxed away, then the optimal long-run tax rate on capital income need not be zero. Similarly, if there are different types of labour (for example, high and low skill) and it is possible for the planner to distinguish between them, then the zero taxation result is overturned. For other examples see Correia (1996) and Jones, Manuelli and Rossi (1997).

Money and growth

Since the neoclassical growth model satisfies the assumptions of the convex economy studied by Debreu (1959), it is impossible to find an equilibrium in which a non-interest earning asset (for example, money) has positive value in equilibrium. In order to introduce money, the neoclassical growth model has been modified in a variety of ways. One of the first attempts corresponds to Sidrauski's (1967) analysis of a monetary model. Sidrauski studied the case in which money enters the utility function, as a reduced form that captures the services provided by money balances.
In Sidrauski's formulation (adapted to discrete time), the consumer problem is

$$\max \sum_{t=0}^{\infty} \beta^t u(c_t, m_{t+1} / p_t)$$

subject to

$$c_t + \frac{m_{t+1}}{p_t} + x_t + \frac{B_{t+1}}{p_t} \leq w_t + q_t k_t + \frac{m_t}{p_t} + \frac{(1+i_t)B_t}{p_t} + \frac{M_{t+1} - M_t}{p},$$

where m_t is nominal money balances chosen by the household, M_t is the economy-wide per capita money supply (that the individual takes as given), p_t is the price level, B_t is the nominal value of one period bonds purchased at time t–1, and $(1+i_t)$ is the gross nominal interest rate. The specification of the budget constraint reflects the assumption that the government exogenously increases the stock of money through lump-sum transfers.
The first order conditions for this problem are (imposing the standard equilibrium conditions)

$$u_1(c_t, m_{t+1} / p_t) = \lambda_t,$$
(9a)

$$u_2(c_t, m_{t+1} / p_t) = \lambda_t \frac{i_{t+1}}{1 + i_{t+1}},$$
(9b)

$$\lambda_t = \beta \lambda_t [1 - \delta_k + z f'(k_{t+1})],$$
(9c)

and feasibility. In this version of the model, money is superneutral in the steady state. In the steady state eq. (9c) reduces to eq. (5a) and, hence, the rate of money growth has no impact on the long-run level of output. This result is not robust. If labour is supplied elastically, inflation has (in general) real effects through its impact on the marginal rate of substitution between real money balances and leisure. The one case in which money is still neutral is when the utility function is separable in real money balances (see Fischer, 1979).
In an economy in which nominal money balances grow at the (gross) rate $1+\pi$, the nominal interest rate is given by

$$1 + i = (1 + \rho)(1 + \pi),$$

and satisfies the Fisher equation. Friedman (1969) argued that since money is costless to produce, its optimal level should be such that individuals are satiated. This corresponds to $u_2(c_t, m_{t+1} / p_t) = 0$. Inspection of eq. (9b) shows that the optimal quantity of money requires that the nominal interest rate be 0. This can be implemented by engineering a deflation (that is, setting $1+\pi=(1+\rho)^{-1}$) or by keeping the price level constant and paying interest on money holdings.

In general, in the non-separable case, the Friedman rule needs to be modified (see Turnovsky and Brock, 1980).

Fertility and growth

The neoclassical growth model can be easily extended to the case of exogenous population growth and exogenous technical change. It has also been used to understand the interplay between economic forces and fertility decisions (see Barro and Becker, 1989; Becker and Barro, 1988).

To illustrate the relationship between growth and fertility, assume that individuals live for just one period and that each agent gives birth to η offspring. The utility function of a member of generation *t* is given by

$$U_t = u(c_t) + \beta \eta_t^{(1-\varphi)} U_{t+1},\ 0 \le \varphi \le 1,$$

where η_t is the number of children. When $\phi>0$, these preferences display imperfect altruism as increases in the number of children result in lower marginal contribution of the last child to utility.

It is assumed that each child costs υ units of labour, and the per capita labour endowment is normalized to 1. The planner's problem for this economy can be expressed as

$$\max \sum_{t=0}^{\infty} \beta^t N_t u(c_t),$$

subject to

$$c_t + \eta_t(a + k_{t+1}) \le zF(k_t, 1 - \eta_t \upsilon) + (1 - \delta_k)k_t,\ k_0 > 0, N_{t+1} \le N_t \eta_t^{(1-\varphi)},\ N_0 = 1$$

Thus, from a formal point of view, endogenous fertility plays the role of another good, N_t, which is ‘produced’ with a linear technology with current fertility as its only input. This is a special case of a two-sector model. Barro and Becker showed that if the utility function is of the form $u(c)=c^{\sigma}$ – a standard specification – the model can have multiple steady states, with some stable and some unstable.

The model has been used to study the effect of changes in child mortality on fertility (see Doepke, 2005), the impact of introducing social security (see Boldrin and Jones, 2005), and the relationship between fertility, growth

and human capital (see Manuelli and Seshadri, 2007b). In general, the ability of the model to match the evidence depends on the specific parameterization used, and finding the appropriate specification is an active area of research.

Finite lifetimes

What are the properties of the neoclassical growth model if economic agents have short – relative to the economy – horizons? The simplest case is study an economy in which individuals live for two periods, and have preferences defined over first- and second-period consumption. This model was originally analysed by Diamond (1965), and an excellent textbook treatment can be found in Azariadis (1993).

Each agent inelastically offers one unit of labour in his first period, and $e \leq 1$ units in his second period. The representative agent problem is

$$\max U(c_t^t, c_{t+1}^t)$$

subject to

$$c_t^t + (1 + r_{t+1})^{-1} c_{t+1}^t \leq w_t + (1 + r_{t+1})^{-1} w_{t+1} e,$$

where c_t^j denotes consumption at time t of an individual born in period j, and w_t is the wage rate. Feasible allocations satisfy

$$c_t^t + c_t^{t-1} + x_t \leq zF(k_t, 1 + e), k_{t+1} \leq (1 - \delta_k)k_t + x_t, t = 0, 1, \ldots.$$

where, as before, we assume that F is homogeneous of degree 1.

Since the solution to an individual optimization problem is completely summarized (in the two period setting) by its saving function, let

$$s_t = s(w_t, w_{t+1}, r_{t+1}) \tag{10}$$

denote saving by a member of generation t. Firms, as in the case of infinite horizons, are assumed to solve static problems. Equilibrium input prices, satisfy the appropriate version of (7).

An equilibrium in this economy consists of sequences of capital stocks and prices such that individuals and firms optimize and markets clear. A simple (and intuitive) condition that characterizes all the equilibria is the requirement that saving by the young at time t equal the capital stock at the beginning of period t+1. Formally, this corresponds to

$$k_{t+1} = s(\bar{w}(k_t), \bar{w}(k_{t+1}), \bar{r}(k_{t+1})), \tag{11}$$

where,

$$\bar{w}(k) \equiv zF_2(k, 1 + e), \bar{r}(k) = zF_1(k, 1 + e) - \delta_k.$$

For a given k_0, any sequence that satisfies (11) and that does not violate other feasibility conditions (for example, $k_t \geq 0$) is an equilibrium sequence of capital stocks. The other components of an equilibrium (for example, consumption and prices) can be readily obtained from the household and firm optimization problems.
Even though this set-up (with only one type of consumer) appears very close to the infinite horizon model, its implications are quite different. An (incomplete) list of the most interesting properties includes the following:

1. 1. Even if e=0 (young individuals are net savers), and if both consumption goods are normal, the equilibrium need not be unique. A sufficient condition for uniqueness is that the two goods be gross substitutes. This corresponds to the saving function being an increasing function of the interest rate.
2. 2. If e=0 and saving is increasing in the interest rate, eq. (11) can be solved for k_{t+1}. Let the solution be denoted $k_{t+1}=G(k_t)$. Then, if $G'(0)>1$, then this map can have and odd number ($2j$+1) of nontrivial steady states, of which j+1 are asymptotically stable and j are unstable. If $G'(0)<1$ there may be an even number of nontrivial steady states.
3. 3. If e=0 and saving is not increasing in the interest rate, eq. (11) can be solved for k_{t+1} only locally. The major impact of this is that stable steady states need not be separated by unstable steady states.
4. 4. Equilibrium paths of capital may display cycles and, depending on the specification, chaotic dynamics.
5. 5. Equilibria – even stationary equilibria – need not be optimal.

This last result shows that when the individual horizon differs from the economy's horizon, then optimal saving at the individual level need not imply optimality in the aggregate, even in the absence of the standard arguments (for example, externalities) for market failure.
To illustrate what can go wrong, consider an economy in which U is strictly quasi-concave and that, in a stationary equilibrium, the stock of capital is such that $\bar{r}(\bar{k}) = zF_1(\bar{k}, 1) - \delta_k < 0$. Let the levels of consumption in young and old age be denoted $(\bar{c}_1, \bar{c}_2)$. The key condition is that the gross interest rate be less that the gross rate of population growth, which is assumed to be 1 in this example. Consider next the problem of maximizing the utility of a given generation subject to the constraint that allocations be constant and the stock of capital also remains constant. Let k^* be the solution to

$$\max U(c_1, c_2)$$

subject to

$$c_1 + c_2 \leq zF(k, 1) - \delta_k k.$$

Let the solution of this problem be (c_1^*, c_2^*, k^*). Given that k^* is such that $zF_1(k^*,1) - \delta_k = 0$, it follows that $k^* < \bar{k}$. Since $(\bar{c}_1, \bar{c}_2, \bar{k})$ is feasible, it must be the case that $U(c_1^*, c_2^*) > U(\bar{c}_1, \bar{c}_2)$. Thus all generations, starting with generation 1, are better off under this alternative allocation. What about the initial old? Since they only care about consumption they are also better off as fewer resources are allocated to investment.

To summarize, when individual horizons are shorter than the economy's horizon, even the simplest specification of the neoclassical growth model can result in very complicated equilibrium paths.

Concluding comments

For many years, the neoclassical growth model has been the workhorse of researchers interested in fluctuations and growth. The model is not without weaknesses. Perhaps the most important is its inability to explain long-run growth: in the steady state the growth rate is exogenous. Endogenous growth models – versions of which are very close to the neoclassical growth model – can be used to understand the effects of policies and shocks on long-run growth. Currently, there are isolated attempts to integrate both views. This has been done for versions of the models that assume convex technologies. For example, endogenous growth models have been used to eliminate the need for arbitrary detrending in the study of business fluctuations (see, for example, Jones, Manuelli and Siu, 2005). The versions of the models that have been studied so far are, of necessity, the simplest ones. It is too early to tell whether the integration of the two strands will succeed.

A large literature on endogenous growth departs from the assumption of convex technologies and no external effects. This body of research views innovation as a form of public good, and emphasizes the role of institutions (for example, how property rights are protected) in determining growth. Since these assumptions amount to departures from the convexity assumptions of the neoclassical model, competitive equilibria are no longer optimal, and this alternative view suggests that a variety of interventions are needed to attain optimality. Thus, the major difference relies on the presence (or absence) of departures from the assumption that technologies form a convex cone.

If the neoclassical growth model is narrowly interpreted (as in this article) as assuming that government policies are exogenous (and markets are competitive), then it follows that the fundamental cause of cross-country differences in output are differences in policies. More recently, the analysis of the determinants of development has emphasized the role of (endogenous) institutions and geography. Endogenizing the institutional structure seems like a natural next step in the development of the theory. However, serious theoretical limitations of our understanding of social choice theory in dynamic settings has limited progress so far. The direct role of geography is easily incorporated into the framework. However, to the extent that the geographic dimension is viewed as influencing (or determining) institutions and or policies, the same limitations apply.

In summary, the neoclassical growth model is still the basic framework to study questions that require

understanding differences across countries, regions or individuals, in the *level* of some economic variable. The main challenge for future research is to develop a theory of social choices (policy choices) that is consistent with the dynamic framework.

See Also

- neoclassical growth theory

Bibliography

Azariadis, C. 1993. *Intertemporal Macroeconomics*. Cambridge: Blackwell Publishers.

Barro, R.J. and Becker, G.S. 1989. Fertility choice in a model of economic growth. *Econometrica* 57, 481–501.

Becker, G.S. and Barro, R.J. 1988. A reformulation of the economic theory of fertility. *Quarterly Journal of Economics* 103, 1–25.

Ben Porath, Y. 1967. The production of human capital and the life cycle of earnings. *Journal of Political Economy* 75, 352–65.

Bils, M. and Klenow, P. 2000. Does schooling cause growth? *American Economic Review* 90, 1160–83.

Boldrin, M. and Jones, L.E. 2005. Fertility and social security. Staff Report No. 359, Federal Reserve Bank of Minneapolis.

Boldrin, M. and Montrucchio, L. 1986. On the indeterminacy of capital accumulation paths. *Journal of Economic Theory* 40, 26–39.

Brock, W.A. and Mirman, L.J. 1972. Optimal economic growth and uncertainty. *Journal of Economic Theory* 4, 479–513.

Burmeister, E. 1980. *Capital Theory and Dynamics*. Cambridge: Cambridge University Press.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32, 233–40.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lifetimes. *Econometrica* 54, 607–22.

Cooley, T.F. 1995. *Frontiers of Business Cycle Research*. Princeton: Princeton University Press.

Correia, I. 1996. Should capital be taxed in the steady state? *Journal of Public Economics* 60, 147–51.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences*

40, 588–92.

Debreu, G. 1959. *The Theory of Value*. New Haven and London: Yale University Press.

Diamond, P.A. 1965. National debt in a neoclassical growth model. *American Economic Review* 55, 1126–50.

Doepke, M. 2005. Child mortality and fertility decline: does the Barro–Becker model fit the facts? *Journal of Population Economics* 18, 337–66.

Donaldson, J.B. and Mehra, R. 1983. Stochastic growth with correlated production shocks. *Journal of Economic Theory* 29, 282–312.

Fischer, S. 1979. Capital accumulation on the transition path in a monetary optimizing model. *Econometrica* 47, 1433–39.

Friedman, M. 1969. The optimum supply of money. In *The Optimum Supply of Money and Other Essays*, ed. M. Friedman. Chicago: Aldine.

Jones, L.E., Manuelli, R.E. and Rossi, P.E. 1997. On the optimal taxation of capital income. *Journal of Economic Theory* 73, 93–117.

Jones, L.E., Manuelli, R.E. and Siu, H. 2005. Fluctuations in convex models of endogenous growth II: business cycle properties. *Review of Economic Dynamics* 8, 805–28.

Judd, K.J. 1985. Redistributive taxation in a perfect foresight model. *Journal of Public Economics* 28, 59–84.

Klenow, P. and Rodríguez-Clare, A. 1997. The neoclassical revival in growth economics: has it gone too far? In *Macroeconomics Annual 1997*, ed. B. Bernanke and J. Rotenberg. Cambridge, MA: MIT Press.

Koopmans, T.J. 1965. On the concept of optimal economic growth. In *The Econometric Approach to Development Planning*, Chicago: Rand McNally.

Manuelli, R.E. and Seshadri, A. 2007a. Human capital and the wealth of nations. Working paper, University of Wisconsin.

Manuelli, R.E. and Seshadri, A. 2007b. Explaining international fertility differences. Working paper, University of Wisconsin.

McKenzie, L.W. 1986. Optimal economic growth, Turnpike theorems and comparative dynamics. In *Handbook of Mathematical Economics*, vol. 3, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Prescott, E.J. and Mehra, R. 1980. Recursive competitive equilibrium: the case of homogeneous households. *Econometrica* 48, 1365–79.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 28, 543–59.

Sidrauski, M. 1967. Inflation and economic growth. *Journal of Political Economy* 75, 796–810.

Stokey, N.L. and Lucas, R.E. (with E.C. Prescott). 1989. *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

Turnovsky, S.J. and Brock, W.A. 1980. Time consistency and optimal government policies in perfect Foresight equilibrium. *Journal of Public Economics* 13, 183–212.

How to cite this article

Manuelli, Rodolfo E. "neoclassical growth theory (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000167> doi:10.1057/9780230226203.1171

The New Palgrave Dictionary of Economics Online

neoclassical growth theory

F.H. Hahn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Neoclassical growth theory is mostly that of the equilibrium of a competitive economy through time. It stresses capital accumulation, population growth and technical progress. It distinguishes momentary equilibrium (when the capital stock, the working population and technical know-how are fixed) from long-run equilibrium (when none of these elements is given). Long-run equilibrium is not a sequence of momentary equilibria, since it embodies the rational expectations of agents. The theory has little to say about the ‘animal spirits’ that may determine an economy's potential growth rate, but provides a good base camp for sallies into the study of particular economies.

Keywords

accumulation of capital; animal spirits; Arrow, K. J.; capital–labour ratio; classical saving function; Cobb–Douglas functions; convergence; duality; elasticity of substitution; expectations; factor–price frontier; Hahn, F. H.; Harrod, R. F.; investment behaviour; Kaldor, N.; Keynes, J. M.; knife-edge problem; liquidity trap; long-run equilibrium; Lucas, R.; Meade, J. E.; Modigliani, F.; momentary equilibrium; natural and warranted rates of growth; neoclassical economics; neoclassical growth theory; new macroeconomics; overlapping generations; population growth; proportional savings assumption; rational expectations equilibrium; Robinson, J. V.; Samuelson, P. A.; savings; Solow, R.; steady-state equilibrium; technical progress; technical progress function; unemployment; von Neumann, J.; warranted path; Wicksell effect

Article

Neoclassical growth theory is not a theory of history. In a sense it is not even a theory of growth. Its aim is to supply an element in an eventual understanding of certain important elements in growth and to provide a way of organizing one's thoughts on these matters. For instance, the question of whether technical progress is bound to be associated with unemployment cannot be decisively answered by the

theory but it goes a long way in pinpointing those considerations on which an answer depends.

Most of the theory is that of the equilibrium of a competitive economy through time. In particular, attention is paid to the accumulation of capital goods, growth in population and technical progress. Two kinds of equilibria are distinguished. One is the short period or *momentary* equilibrium of the economy when the stock of capital goods, the working population and technical know how can be taken as fixed. The other is the *long-run* equilibrium when none of these three elements are taken as given. It is important to understand that while long-run equilibrium implies momentary equilibrium for all dates it is not the case that a sequence of momentary equilibria constitutes a long-run equilibrium. For the latter has the property that the actions of agents taken at a given date in the light of their expectations of events at subsequent dates are not regretted when these dates arrive. In other words, it is what we would now call a *rational expectations equilibrium*. Harrod (1939) called a path of an economy with this property the *warranted path*.

In principle a warranted path (say of output or output per man) could be quite irregular. Indeed it could be cyclical (Lucas, 1975). But except in very simple models such generality is intractable and most of the attention has been devoted to long-run equilibria which are *steady-state* or *quasi-stationary*. (If a variable $x(t)$ obeys the dynamic equation $x(t) = e^{gt}x(0)$ then $\hat{x}(t) = x(t)e^{-gt} = x(0)$ is a constant, that is x is stationary.) This is one of the reasons why the theory is not really a theory of growth. It is also unwise to identify the steady state – say, the steady state rate or growth in output per head – with historical trends in the variable. That would require a good deal more argument than the theories provide. A steady state equilibrium is simply an extension of stationary equilibrium (an equilibrium in which the stock of capital goods, the population and technical knowledge are all constant). But it allows this now to include accumulation and technical change.

It is of interest to ask whether a steady state equilibrium is possible and if it is, whether a sequence of short period equilibria guides the economy to it. There is also another qst: do all warranted paths eventually become steady states? (See Hahn, 1987.) However the literature on these matters is sometimes confused and confusing. Short period equilibrium plainly depends on agents' expectations and so if they are not postulated to be always correct there are many possible evolutions of such equilibria. In fact except for Harrod's (1939) pioneering discussion of *actual* growth paths and one or two others, little attention has been paid to the expectational problem. Instead the path of the economy has been studied on the hypothesis that what is saved is also invested without explicit attention to what this implies for expectations concerning prices and interest rates. When that is made explicit it turns out that only warranted paths have been examined and not a sequence of short period equilibria. This procedure has been also adopted by the ‘new macroeconomics’ (e.g. Lucas, 1975).

Connected with this is the treatment of investment and savings. The latter are usually taken to be either proportional to income or to come only from profits. Savings are not explained by the optimizing choices of households. This, however, is against the spirit of neoclassical economics. In order to improve on conventional savings theory one either takes a world which one can study ‘as if’ agents were infinitely long lived or one considers an economy of *overlapping generations* first studied by Samuelson (1958). Neither of these moves is discussed in what follows. But I re-emphasize that until savings behaviour has been explained the theories are not fully neoclassical.

Investment behaviour is a more difficult matter. Since the bulk of the theory is one of the warranted path, the marginal return to any investor is always equal to the marginal cost of investment. Thus

investment is never regretted and is simply explained by it not being profitable to undertake more or less investment than is thus warranted. But difficulties arise if the warranted path and particularly the steady state is not unique, and also if investment is in some sense the carrier of technical progress. ‘Animal spirits’, as Keynes called entrepreneurial investment propensities, may be determinants of the rate of growth which the economy is capable of. Equally important is the circumstance that investment behaviour will be of prime importance in the evolution of a sequence of short run equilibria. neoclassical theory has little to offer on these matters and is open to criticism on these grounds.

This brings me back to the beginning. As will be seen from what follows neoclassical theory states quite precisely what kind of economy in what kind of state is being considered. This economy and this state may be considered to be of low descriptive power. That, however, needs empirical argument and neither proponents nor opponents have produced any clinching ones. But an equally interesting question is whether the theory provides a good base camp for sallies into the study of particular economies. For instance, does it allow us to find just that feature of such an economy which is at variance with the postulates of the theory and thence to a modification of the latter, step by step? To this question at the moment the answer must be yes.

There is one last matter. The theories here discussed have provided the arena for much controversy concerning the *logical* coherence of neoclassical theory in general (Robinson, 1965; Harcourt, 1969). This controversy is not here discussed. For what it is worth it is this writer's view that neoclassical theory has survived this controversy unscathed. But the emphasis here is on ‘logical’. There is little to be said for those economists who have taken the question of the descriptive merit of the theory as having been decisively settled in its favour.

1 The simple model

1.1 The single good economy. no technical progress

Consider an economy in which a single good is produced by means of itself and labour. The good can also be consumed. The stock of it devoted to production is denoted by K and called capital. The stock does not depreciate either through use or the passage of time. Further notation is as follows: Y is output, L is the amount of labour used in production, L^0 is the labour force, $y = Y / L$, $k = K / L$, $e = L / L^0$.

Assumption 1.1: The production possibilities of the economy can be represented by a C^2 production function.

$$Y = F(K, L)$$

with the following properties:

1. (a) For all $h > 0$: $hY = F(hK, hL)$. (Constant Returns to Scale)
2. (b) $f'(k) > 0$, $f''(k) < 0$ for $k \in [0, \infty]$. Also $f'(0) = \infty$, $f'(\infty) = 0$

(The ‘Inada Conditions’; see Inada, 1963).
From these assumptions it follows that we may represent the production possibilities by

$$y = f(k). \tag{1.1}$$

Assumption 1.2: The working population L^0 grows at a constant geometric rate λ [i.e. $L^0(t) = L^0(0)e^{\lambda t}$].
Assumption 1.3: A constant fraction s of output is not consumed.
It will thus be a condition of equilibrium that output which is not consumed is invested:

$$sf[k(t)] = sy(t) = \frac{\dot{K}(t)}{L(t)} = \dot{k}(t) + k(t)\frac{\dot{L}(t)}{L(t)}. \tag{1.2}$$

Definition 1.1: The economy is said to be in steady state equilibrium if $k(t)$ and $e(t)$ are constants, profits are maximized and (1.2) holds.
If $e(t)$ is constant then

$$\frac{\dot{L}(t)}{L(t)} = \frac{\dot{L}^0(t)}{L^0(t)} = \lambda.$$

Using this and the condition $\dot{k}(t) = 0$ in (1.2) yields

$$\lambda = \frac{sf(k)}{k} \tag{1.3}$$

as a condition for steady state equilibrium. Harrod (1939) called $sf(k)/k$ the *warranted rate of growth* and we shall abbreviate by writing

$$\frac{sf(k)}{k} \equiv w(k).$$

Clearly $w(k)$ gives us the rate of growth of output required to keep investment and savings equal to each other in steady state. On the other hand, λ is the rate of growth of employment which is needed to keep the proportion employed (possibly = unity) constant. Harrod called it the *natural rate of growth* of output for it tells us the rate at which output grows at a constant e.

Now by A.1.1.(b) one has $w(0) > \lambda$ and $w(\infty) < \lambda$ so there exists k^* satisfying (1.3). Since $w'(k) < 0$ everywhere, k^* is the only value of the capital labour ratio satisfying 1.3. But then for profit maximization, the real wage w^* and the real interest rate, ρ in steady state equilibrium are:

$$w^* = f(k^*) - k^* f'(k^*) \text{ and } \rho^* = f'(k^*). \tag{1.4}$$

So the steady state equilibrium exists and is uniquely characterized by (1.4) and

$$\lambda = w(k^*) \tag{1.5'}$$

Now return to (1.2) and consider the path $k(t)$ out of steady state but with $e(t)$ constant at e. In our new notation we find

$$\frac{\dot{k}}{k} = [w(k) - \lambda] \tag{1.5}$$

by dividing (1.2) by k and rearranging. Now let

$$V(k) = \frac{1}{2}[w(k) - \lambda]^2$$

so that $V(k)$ is a measure of the deviation of the warranted from the natural rate of growth. One has:

$$V(k) \geq 0 \text{all } k \text{ and } V(k^*) = 0. \tag{1.6}$$

Also using (1.5):

$$\dot{V}(k) = [w(k) - \lambda]w'(k)\dot{k} = [w(k) - \lambda]^2 k w'(k) < 0 \text{all } k > 0 \text{ and } k^* \neq k. \tag{1.7}$$

These two results together with the Inada conditions suffice for the conclusion:

$$\text{For all } k(0) \geq 0, \lim_{t \to \infty} k(t) = k^*.$$

We sum up:
Proposition P.1.1: An economy satisfying A.1.1 – A.1.3 has the following properties:

1. (a) There exists a unique steady state equilibrium
2. (b) The path of the economy along which savings are always equal to investment and the proportion of the workforce employed is constant (e is constant) approaches the steady state equilibrium as $t \to \infty$.

1.2 Discussion of the model

There are many lacunae in the theory just presented and we shall be able to fill in some of these below. But first I discuss what can be learned from it.
Harrod (1939) writing in a Keynesian spirit held the view that a steady state equilibrium might not exist. He was particularly interested in the possibility that the warranted growth rate was always above the natural rate. In that case output would have to grow faster than is physically possible in order for

investment to take up the savings generated and that is not possible. There would be a permanent tendency to depression. For many commentators this view of Harrod's rested implicitly on an assumed production function of the form:

$$Y = \min[aK, bL] \tag{1.8}$$

that is on fixed coefficients of production (see e.g. Solow, 1956). However, a careful reading of Harrod suggests that he rather based his argument on the Keynesian liquidity trap. That is he thought that monetary forces set a positive lower bound on the rate of interest which thus on neoclassical theory set an upper bound on *k* and so, given *s*, a lower bound on *w*(*k*).

This argument, however, is suspect. It is the real and not the nominal interest rate which governs (together with the real wage) the choice of *k*. Liquidity preference may set a lower bound on the nominal interest rate (the cost of holding money) but not on the real rate. Thus suppose *r* is the nominal interest rate. Then

$$\rho = r - \frac{\dot{p}}{p}$$

where *p* is the price of the good. Then if *r* is at its minimum level $\underline{r}$ we have from (1.4)

$$\left(\frac{\dot{p}}{p}\right)^* = \underline{r} - f'(k^*) \tag{1.9}$$

as a condition of steady state equilibrium. By assumption $f'(k^*) < \underline{r}$ so for such an equilibrium one requires a constant inflation rate:

$$\left(\frac{\dot{p}}{p}\right)^* > 0.$$

So provided we can graft a monetary sector onto the simple model it would seem that the liquidity trap is not an obstacle to the existence of steady state equilibrium.

But this argument reveals a central weakness in the reasoning which supports P.1.1(b). For suppose at a historically given *k* one has $w(k) > \lambda$. If we *impose* the condition that savings are equal to investment, then indeed there would be pressure on resources and one could tell a story to explain the generation of the required inflation rate of (1.8). But we have no good reason for imposing that condition. By doing so we are not really asking: what actually happens?, that is, what is the actual growth rate?, but rather we are implicitly postulating that the inflation rate is always such that excess savings for *k* constant are taken up by capital deepening $(\dot{k} > 0)$. But why should this be so? If, for instance, the economy grew at λ then there would be excess supply of the good and normal arguments would lead us to suspect falling prices. But these would raise the real rate of interest and raise $w(k)$ above λ even further. The steady state equilibrium even if it exists is an unstable ‘knife-edge’ (Harrod, 1939).

(b) Solow's celebrated paper (1956) established P.1.1. But Solow was mistaken in his belief that it disposed of Harrod's knife-edge. The latter does not deal with paths on which the condition: savings = investment at a constant *e* has been imposed. That is Harrod did not postulate that the actual path was an equilibrium path. In this he was right since there is no good explanation of the Solow condition.

(c) An alternative procedure leading to P.1.1(a) even if 1.8 is the form of the production function is to drop A.1.3 (Hahn, 1951; Kaldor, 1955; Robinson, 1965). This is done by supposing that the saving ratio out of profits is higher than that out of wages. Now if there are fixed coefficients of production (1.8) the equilibrium conditions (1.4) have no meaning since marginal products are not defined. This leaves it open to determine the real wage and interest rate by the requirement that they should generate that distribution of income between wages and profits which makes the warranted growth rate equal to the natural rate. From (1.8) one finds

$$\frac{Y}{K} = a, \frac{Y}{L} = b \text{ and } k = \frac{b}{a} \equiv \beta \text{ say}.$$

Let s_0 be the saving propensity out of wages and s_1 the saving propensity out of profits, with $s_0 < s_1$. Then the aggregate saving propensity, *s*, of the economy is given by

$$\frac{s_1 \rho}{a} + s_0 \frac{w}{b} = s.$$

Imposing the condition $sa = \lambda$ (the warranted rate = natural rate) yields

$$s_1\rho + s_0\frac{w}{a} = \lambda.$$
(1.10)

But also

$$\frac{\rho}{a} + \frac{w}{a} = 1$$
(1.11)

so that we have two equations to determine what w^* and ρ^* must be in steady state equilibrium. A special case arises when $s_0 = 0$ (no saving out of wages) and $s_1 = 1$ (no consumption out of profits). Then

$$\rho^* = \lambda$$
(1.12)

is the condition of equilibrium. The reader should avoid interpreting (1.12) as saying that λ 'determines' the rate of profit. Equation (1.12) tells us what ρ must be if there is to be steady state equilibrium.

Once again a version of P.1.1(a) survives. Also stability fares slightly better than in (a). For if the actual growth rate is less than the warranted rate (because *w* and ρ have the 'wrong' values), and the latter is greater than λ then investment will be less than savings and competition between firms may lead to lower prices, higher real wages and so a fall in *s*. This will lower the warranted rate and bring it closer to λ as well as reducing the investment-savings gap. This *may* be so but what has just been said is not a proof. Indeed, as for instance Meade (1966) has shown, falling profitability may reduce the willingness to invest and so lead the system away from steady state equilibrium.

(d) Of course, (1.8) is not a plausible production function. Suppose we combine the savings assumption of (c) with a neoclassical production function satisfying A.1.1. Then certainly (1.4) must hold in equilibrium. But (1.3) will now read

$$(s_1 + s_0)f'(k) + s_0\frac{f(k)}{k} = \lambda$$
(1.13)

from which we can find k^*. (Since

$$s_1\rho\frac{K}{Y} + s_0\frac{wL}{Y} = s.$$

So

$$s_1\rho + s_0\frac{w}{k} = s\frac{Y}{K} = \lambda.$$

Then substitute from (1.4) for ρ and w. So while the saving hypothesis will be reflected in the steady state value of k it will leave the equality between marginal productivity and factor rewards as an equilibrium condition. Indeed without this, the steady state values of w and ρ would be unknown. This is so even under the 'classical' savings assumption that $s_0 = 0$. The equation derived from (1.13) is then

$$s_1 f'(k) = \lambda$$

and it tells us what k must be in order to generate a profit rate which, given the savings hypothesis, generates just the right amount of savings required for a growth in the capital stock at the rate λ. Thus the savings hypothesis has no direct bearing on the neoclassical equilibrium condition that the rate of profit must equal the marginal product of k.

(e) If workers save and invest their savings at the current rate of return on capital then the foregoing arithmetic needs to be changed. This was first noticed by Pasinetti (1962) whos paper gave rise to a number of others (Meade and Hahn, 1965; Modigliani and Samuelson, 1966).

Let $\sigma = s_1 - s_0 > 0$ Let μ be the fraction of k owned by capitalists – that is by agents who have no income from work. Then savings per employed worker are given by

$$s_0 f(k) + \sigma f'(k)\mu k.$$

So in steady state equilibrium one requires

$$\frac{s_0 f(k)}{k} + \sigma f'(k)\mu = \lambda. \tag{1.14}$$

From which

$$\frac{\mu f'(k)k}{f(k)} = \frac{1}{\sigma}\left[\frac{\lambda k}{f(k)} - s_0\right]. \tag{1.15}$$

The left-hand side measures the capitalists' share in income which cannot be negative. But there is nothing which guarantees a solution to (1.15) with $\lambda k \geq s_0 f(k)$. Pasinetti (1962) simply made the latter (with strict inequality) a condition of the model. But God may have made the world otherwise.
In fact there are two possibilities. Suppose (1.15) has an admissible solution. One notes that in steady state one must have

$$1 - \mu = \frac{s_0\left[f(k) - \mu k f'(k)\right]}{\lambda k}. \tag{1.16}$$

That is the ratio of workers' capital to total capital must equal the ratio of their savings to total savings which in steady state equilibrium is equal to λk. Solving (1.16) for μ yields.

$$\frac{\lambda k - s_0 f(k)}{k[\lambda - s_0 f'(k)]} = \mu. \tag{1.17}$$

Solving (1.14) for μ yields

$$\left[\frac{\lambda k - s_0 f(k)}{k}\right]\frac{1}{\sigma f'(k)} = \mu.$$
(1.18)

Equating (1.17) to (1.18) then yields

$$s_1 f'(k) = \lambda.$$
(1.19)

So even though workers save, the long run equilibrium rate of profit bears the same relation to λ as it does under the classical savings hypothesis. Note that $\lambda k > s_0 f(k)$ is here required as before. In particular write (1.18) as

$$\max\left[0, \frac{\lambda k - s_0 f(k)}{k}\right]\frac{1}{\sigma f'(k)} = \mu.$$
(1.18′)

Then this always has an admissible solution. If that gives $\mu = 0$ then from (1.14)

$$\frac{s_0 f(k)}{k} = \lambda$$
(1.20)

Harrod solution. It should now be emphasized that $\mu = 0$ does *not* mean that capitalists own no capital. All it means is that their share in total capital is zero.

Modigliani and Samuelson (1966) have shown how a warranted growth path may converge to k^* given by (1.12) or to k^{**} given by (1.20) depending on the technology and savings propensities.

(f) It will have been noticed that the whole of the above discussion has been conducted for L/L^0 constant and not $L/L^0 = 1$; that is the steady state is consistent with permanent unemployment. This should cause no surprise since the assumption of constant returns to scale and of constant savings propensities makes all equilibrium conditions independent of scale. if there is unemployment in a steady state

equilibrium it can be argued with equal lack of real sense that either the capital stock is too low or that the real wage is too high. The present model is not suited to a discussion of whether falling interest rates and or money wages as long as there is unemployment would lead the economy to a steady state with full employment.

1.3 The single good economy with technical progress

Growth theory without technical progress seems pretty useless. Yet no really satisfactory account exists of the determinants of technical progress, at least no such account based solely on considerations of economic theory exists. (Schumpeter (1934) is probably still the most interesting attempt but it excludes the possibility of steady state equilibrium.) What follows is therefore rather ad hoc and mechanical. Technical progress shifts the production function through time and so in its most general form when technical progress is *disembodied*, one writes

$$Y(t) = F[K(t), L(t), t] \tag{1.21}$$

and retains the assumption of constant returns to scale for each *t*. Progress is disembodied if it can be taken full advantage of by the stock of the good (capital) accumulated in the past and by the same kind of labour. Even with this strong assumption we need more structure to build a model and accordingly postulate that all technical progress is *factor-augmenting*, that is (1.21) can be written as

$$Y(t) = F[\alpha(t)K(t), \beta(t)L(t)] \quad \text{with } \alpha(t) \geq 0, \beta(t) \geq 0 \text{ all } t.$$

Let

$$\hat{K}(t) = \alpha(t)K(t), \hat{L}(t) = \beta(t)L(t)$$

and

$$\hat{k}(t) = \frac{\hat{K}(t)}{\hat{L}(t)}, \hat{y}(t) = \frac{Y(t)}{\hat{L}(t)}.$$

Then the equilibrium real interest rate is given by $\alpha f'[\hat{k}(t)]$ when $\hat{y}(t) = f[\hat{k}(t)]$.
In steady state equilibrium the real interest rate is constant. Let the operator *E* applied to a function *g*(*x*) denote its elasticity

$$\left[Eg(x) = \frac{g'(x)}{g(x)}x\right].$$

Then for the real interest rate to be constant one requires:

$$\frac{\dot{\alpha}}{\alpha} + \left\{Ef'[\hat{k}(t)]\right\}\left[\frac{\dot{\alpha}}{\alpha} - \frac{\dot{\beta}}{\beta} + \frac{\dot{k}}{k}\right] = 0.$$
(1.22)

Suppose first that $\alpha(0) = \beta(0) = 1$ and that $\dot{\alpha}(t) = 0$ all *t*, $\dot{\beta}(t) = b\beta(t)$ all *t*. Technical progress is purely labour augmenting (at a constant rate) or *Harrod-Neutral*. Clearly $\beta(t) = e^{bt}$. Hence (1.22) will be satisfied if

$$\frac{\dot{k}(t)}{k(t)} - b = 0 \text{ or } \frac{\dot{K}}{K} = b + \lambda.$$
(1.23)

Let $n = b + \lambda$ and call it the *natural rate of growth*. If savings are proportional to income, equilibrium requires

$$\frac{\alpha(0)sf[\hat{k}(t)]}{\hat{k}(t)} = n$$
(1.24)

which can be uniquely solved for $\hat{k}^*$ when the production function is concave and satisfies the Inada conditions. By (1.23), $\dot{k}(t) = 0$ and so we conclude that (i) the capital output ratio and the real interest rate are both constant and (ii) the real wage and the capital labour ratio (*k*) are rising at the rate *b*. But the wage per efficiency unit of labour and capital per efficiency unit of labour are both constant. Hence we are essentially in the same situation as that discussed for the absence of technical progress.

Next suppose that $\dot{\alpha}(t) = a\alpha(t)$ and $a = b$. Technical progress is said to be *Hicks-neutral*. Then (1.22) becomes

$$a + \left\{Ef'[k(t)]\right\}\frac{\dot{k}}{k} = 0. \tag{1.22$'$}$$

Suppose that the production function is characterized by an elasticity of substitution equal to minus one. Then since with Hicks-neutrality one can write: $Y = e^{bt}F[K(t), L(t)]$ one has that KF_k/F is constant when *K* is changed but *F* is constant (if one is moving along an isoquant). This implies

$$Ef'[\hat{k}(t)] = -1$$

and so once again (using (1.22$'$)) one obtains (1.23). A constant rate of profit and a constant share of profits then implies a constant capital output ratio. In other words, Harrod-neutrality is equivalent to Hicks-neutrality with a unit elasticity of substitution (Robinson, 1938). Uzawa (1961) has shown that only a Cobb–Douglas production function will give this equivalence.

If $a \neq b$ technical progress is ‘biased’ in favour of the higher of *a* and *b*. However, there is no fundamental reason why technical progress should be of the factor-augmenting type nor, if it is, why it should proceed at a steady rate. Hence technical progress makes the idea of steady state equilibrium somewhat unconvincing.

However, there have been attempts to formulate a theory which focuses on endogenous economic forces that may cause technical progress to be of a certain kind (Kennedy, 1964; Samuelson, 1965). These attempts are not notably successful or convincing and will only be sketched.

Given a factor-augmenting production function which exhibits constant returns to scale, one can write the minimum unit cost function as

$$c = c[q(t)/\alpha(t), w(t)/\beta(t)]$$

where $q(t)$ is the rental of capital of $w(t)$ the wage. Let s_K and s_L respectively be the shares in unit cost of capital and labour. Then from elementary Duality Theory (e.g. Varian, 1978), if $\dot{w}(t) = \dot{q}(t) = 0$:

$$\frac{\dot{c}}{c} = -[s_K a(t) + s_L b(t)] \tag{1.25}$$

where $b(t) = \dot{\beta}(t) / \beta(t)$, $a(t) = \dot{\alpha}(t) / \alpha(t)$. The idea now is as follows. Firms can choose to 'produce' $a(t)$ and $b(t)$ according to a 'production possibility' function.

$$T[a(t), b(t)] = g[b(t)] - a(t) \geq 0 \tag{1.26}$$

and the pairs (a, b) satisfying (1.26) form a convex compact set with a differentiable boundary. Also $g'(b) < 0$. If the firm's objective is to minimize $\dot{c} / c$ subject to (1.26) it will choose $b(t)$ so as to satisfy

$$-g'[b(t)] = \frac{s_L}{s_K}. \tag{1.27}$$

As Samuelson (1965) has noted, (1.27) is *not* some novel theory of income distribution unrelated to the Neo-classical one. The latter was needed in the definition of c and the derivation of (1.25).
Now s_L/s_K will depend on the relative prices of efficiency units. Since $g(\cdot)$ is monotone (1.27) can be inverted:

$$b(t) = (g')^{-1}(s_L / s_K)$$

and so we write

$$b(t) = h\left[\frac{w(t)}{q(t)}\frac{\alpha(t)}{\beta(t)}\right].$$
(1.28)

The equations (1.26) and (1.28) are two differential equations in $\alpha(t)$, $\beta(t)$ and relative factor prices. It is easy to show that

$$h'(1-\sigma) \geq 0$$

where σ is the elasticity of substitution.
If one can take *w*/*q* constant then one proceeds as follows.

$$b(t) - a(t) = \frac{\mathrm{d}\log[\beta(t)/\alpha(t)]}{\mathrm{d}t} = b(t) - g[b(t)] = v[b(t)] \text{ say}.$$

Substituting from (1.28) one obtains the differential equation

$$\frac{\mathrm{d}\log[\beta(t)/\alpha(t)]}{\mathrm{d}t} = v\left\{h\left[\frac{w}{q}\frac{\alpha(t)}{\beta(t)}\right]\right\}.$$
(1.29)

This equation gives the evolution of relative factor augmentation. If for some $[\alpha/\beta]^*$ one has a critical point of *v* and (1.29) is convergent then there will be a constant relative rate of labour augmentation so $b(t) - a(t) \to 0$. (This does not necessarily imply that *b*(*t*) and *a*(*t*) become constant.) In that situation innovations are derived to be Hicks-Neutral. Even if the rate of innovation is then constant we know that this will not be consistent with steady state unless the elasticity of substitution is unity. But Samuelson (1965) has shown that the stipulated convergence of (1.29) requires an elasticity of substitution which is less than one in absolute value.
All of this is on the assumption *w*/*q* = constant. In fact we know from our earlier discussion that *w*/*q* will depend on $\hat{k}(t)$ so we can replace the r.h.s. of (1.29) by:

$$v^*\left[k(t)\frac{\alpha(t)}{\beta(t)}\right].$$

We then need a differential equation for the evolution of $k(t)$ which we can obtain from the appropriate warranted growth path.

Samuelson (1965) has studied the case: $\dot{k}(t) = 0$. The literature can be consulted for further detail. At this level of aggregation the story is hardly persuasive nor can much be said in favour of the objective function which has been stipulated. On the other hand, all of this is a considerable advance on meaningless claims like: ‘high wages induce labour-saving innovation’ first exposed by Fellner (1961). After all, the marginal return per unit cost of the factor is the same for all factors in equilibrium. None the less one must conclude that the theory of induced innovations and their relations to growth have a long way to go yet.

1.4 The one sector model with embodied technical progress

In this section two related ideas are considered. The first is that capital and labour are substitutable *ex ante* (‘putty’) before investment has been congealed in concrete machines but it is not substitutable *ex post* (‘clay’) once the investment has been made. The second is that technical progress does not benefit old machines; it is embodied in the latest machines. These two ideas are related but can be combined in various ways. Thus one can have embodied technical progress with (traditional) putty–putty (Solow, 1970) or with clay–clay (Solow et al., 1967). One can also have disembodied technical progress as in the previous section with putty–clay. The main lessons are perhaps best learned by combining embodied technical progress with putty–clay. The classic reference here is Bliss (1968).

Some of the technicalities of the analysis now called for are somewhat involved and what follows is more in the nature of a summary of the economic implications.

An investment undertaken at date θ gives rise to machines of vintage θ. If at that date the investment is $I(\theta)$ and employment is $L(\theta, \theta)$, output per man is $y(\theta, \theta)$ and given by

$$y(\theta, \theta) = e^{\bar{a}\theta} f(k(\theta)) \text{ where } k(\theta) = I(\theta) / L(\theta, \theta)e^{\bar{a}\theta}.$$

Let $f(\cdot)$ satisfy Assumption 1.1. The output per man on vintage θ at date $t \geq \theta$ is written as $y(t, \theta)$. It is assumed that as long as output is produced on vintage θ that

$$y(t, \theta) = y(\theta, \theta)$$

(1.30)

This departs somewhat from the ‘clay’ assumption. It will be noticed that Harrod-neutral technical progress has been assumed. It can be shown (Bliss, 1968) that this is necessary for a steady state equilibrium to exist.

Any firm in this technological environment will make its investment and employment decisions in the light of long term expectations. For once machines have been installed they no longer share in technical progress yet the latter will raise real wages and reduce quasi-rents on old machines. These will be scrapped when quasi-rents have fallen to zero so that the economic life of the machines is endogenous to the economic process. The economic life is relevant to the investment decision and hence expectations of the course of real wages are relevant. In the theory it is assumed that all expectations are always correct. None of these considerations apply to the case of disembodied technical progress with putty-putty.

If $w(t)$ is the real wage at t then if $y(t, \theta) - w(t) > 0$ it will pay the firm to set $L(t, \theta) = L(\theta, \theta)$ because of (1.30). It will set $L(t, \theta) = 0$ when $y(t, \theta) - w(t) = 0$. These conditions determine the economic life of a machine. It is easy to show that if T is the economic life of a machine that it must be constant in steady state equilibrium. The value of T is determined by the condition $w(t) = y(t - T, t - T)$, that is, the wage equals its average product on the last vintage in use. When that is the case the firm is indifferent whether it employs labour on that vintage or not. If it does employ some then if the economy had a little more or less labour it would be employment on the last vintage in use which is varied and so $w(t)$ would measure labour's marginal social product. If no labour is employed of the last vintage then a small reduction in labour would mean reducing employment on the next oldest vintage. If there is a continuum of vintages then the economy would still lose just $y(t - T, t - T)$.

Now let $n = a + \lambda$ as in (1.2). We are looking for a steady state equilibrium as before in which output and investment grow at the rate n because gross savings are proportional to income. As before also the ratio of capital to labour measured in efficiency units of the latest vintage (i.e. $k(\theta)$) should be constant. So if $Y(t)$ is aggregate output at t and $Y(\theta, \theta)$ total output with capital of vintage θ we have

$$Y(t) = \int_{t-T}^{t} y(\theta, \theta) L(\theta, \theta) \mathrm{d}\theta = \int_{t-T}^{t} Y(\theta, \theta) \mathrm{d}\theta = \frac{e^{nt} Y(t - T, t - T)(1 - e^{-nT})}{n}. \tag{1.31}$$

If $I(\theta)$ is investment at θ then $I(t) = e^{nt} I(t - T)$ and that must equal $sY(t)$. So using (1.31) and writing $v = Y(t - T) / I(t - T)$ we obtain

$$sv = \frac{n}{1 - e^{-nT}}. \tag{1.32}$$

The left-hand side of (1.32) is again Harrod's warranted growth rate. But the rate at which the economy is capable of expanding indefinitely now depends on *T*, the economic life of equipment and that is an economic variable and *not* a parameter like *n*. One must, of course, show that (1.32) has a solution. If as in Solow et al. (1967) the technology is clay–clay then *v* is given as fixed. Profit maximization together with the condition that the present value of quasi-rents equals the cost of the investment which gives rise to them at the scrapping, fix the equilibrium value of *T*. It is then possible that Harrod's view that (1.32) has no solution is valid. This is a fortiori true if the solution of (1.32) requires $s > 1$.

One can show that the real interest rate (= profit rate) must be constant in steady state equilibrium (see Bliss, 1968). However, the relation between the latter and the equilibrium value of *T* is not straightforward and depends on the elasticity of substitution. That is because in steady state the scrapping condition is $t = 1 / a\log$ (inverse of share of wages in vintage $(t - T)$) and the share will depend on the elasticity of substitution. One can also show that if a steady state exists that the warranted growth path of the economy will approach the steady state. This is even the case with clay–clay.

All in all the simple neoclassical model survives ‘the bolting down’ of concrete machines and embodied technical progress rather well. That does not mean that the resulting model is satisfactorily ‘realistic’. What it does mean is that the theory is a good deal more robust than critics once thought it to be. This is also illustrated by the following episode in the related theory of technical progress.

Kaldor took the view that it was not possible to distinguish between finding another ‘page in the book of blueprints’ (Robinson, 1965), i.e. movements along the production function and finding a new page, i.e. innovations. He proposed that all that could be observed was a relation between the rate of growth in labour productivity and investment per man. This relation he called the ‘technical progress function’ and justified by the view that every act of investment led to learning. He and Mirrlees (1962) constructed a model on this basis. However, except for the assumption that firms required investment ‘to pay for itself’ in a predetermined period, the results of the model were not notably different from the ones already discussed. (A linear technical progress function can be integrated into a Cobb–Douglas production function. A non-linear one of the right shape has the advantage of making steady state equilibrium investment be at the rate at which the capital output ratio is constant, i.e. Harrod-neutrality is a consequence and not a hypothesis of the model.)

Arrow (1962) kept the production function (he uses clay–clay) but made technical improvement depend on the total investment undertaken over the past. This was again justified by learning. The steady state again is one of Harrod-neutral progress which is explained endogenously. There are now obvious external benefits from investment but otherwise the ‘learning by doing’ steady state equilibrium is of the kind we have already discussed.

2 Two sector growth models

One considers an economy with a consumption good and an investment good sector. This was first proposed by Uzawa (1961) and then gave rise to a very large literature (e.g. Solow, 1962; Inada, 1963; Takayama, 1963). We shall discuss only the case where both sectors have ‘well behaved’ constant returns to scale production functions, capital does not depreciate and there is no technical progress. For the latter see Diamond (1965).

2.1 Steady state

It is well known (e.g. Samuelson, 1957; Mirrlees, 1969) that given these assumptions, the equilibrium relative prices of the two goods are determined once ρ (the real interest rate) is determined. So with a classical saving hypothesis we know that steady state requires:

$$\rho = \lambda$$

and so q the price of the investment good in terms of the consumption good can be written as $q(\lambda)$. If w is the wage in terms of consumption good, y_c is output per man employed in the consumption good sector and $\mu = L_c / L$ is the proportion of the labour force employed in that sector, the classical savings assumption yields the equilibrium condition

$$w = y_c \mu \text{ or } \mu = w / y_c. \tag{2.1}$$

(Demand for consumption good equals supply.) But w/y_c is a unique function of ρ. For by profit maximization the marginal product of capital in the consumption sector must equal $\rho q = \lambda q(\lambda)$. So λ determines a unique capital/labour ratio and so a unique share of wages in the consumption sector. Hence we can write $\mu = \mu(\lambda)$. If k is the overall capital labour ratio, k_c and k_I the capital/labour ratios in the consumption and investment sectors respectively then $k = \mu k_c + (1 - \mu) k_I$ It is plain that k is uniquely determined by λ.

Matters are somewhat more complicated with a proportional saving function and we shall not derive all the results in full. Let v be the capital output ratio *in value terms*. In steady state, as usual, we require $s = v\lambda$. The question now is whether putting $v = s / \lambda$ uniquely determines k, k_c k_I and hence the rate of profit and real wage. The answer is: no.

Let ψ be the wage rental ratio. A rise in that ratio will lower q if the consumption goods sector is more labour intensive than the investment goods sector. Hence k_c and k_I will be raised and v will be lowered. But the value of investment output is a constant fraction s of the value of output and q is lower so that output of investment good must rise relatively to that of consumption good and so μ must be lower (

$1 - u$ is higher). Hence *k* will be higher (since $k_I > k_C$) and this will tend to increase *v*. It follows that *v* can have the same value at different *k*'s and ψ's. This is really the story of what Professor Robinson (1965) called the Wicksell effect. To get uniqueness one needs the not very persuasive assumption: $k_C > k_I$ always, or some assumption on the elasticities of substitution (Takayama, 1963).

2.2 Stability

The question may be asked whether a sequence of short period equilibria of the economy starting with an arbitrary *k*(0) at time $t = 0$ lead the economy to steady state equilibrium.

At any moment of time *k* is given from the past. A short period equilibrium is a division of the capital stock and of the labour between the two sectors such that at the resulting prices all markets clear and profits are maximised. The resulting investment good output will augment the capital stock. At the next moment there will also be more labour so we know the new value of *k*. So given *k*(0) it looks as if we could deduce *k*(*t*) for all $t > 0$ and so study the convergence to steady state.

But this is only true if momentary equilibrium is unique. If it is not then there will be a variety of paths the system can follow and we do not know which it will be. More seriously in this case we may have, say, there equilibria for some *k* and only one for another k'. In that case at the point at which we 'lose' equilibria there is a 'catastrophe' (in the technical sense). For this see Inada (1963).

Now consider the proportional savings assumption. It says that consumption and investment are proportional to *aggregate income*, that is, the distribution of income has no effect on the demand for either good. But this is just the case for which non-intersecting community indifference maps exist (see Gorman, 1953) and in that case momentary equilibrium must be unique: it is given by the tangency of the transformation curve between investment and consumption good and the indifference curve. So in this case momentary equilibrium is unique.

But this is not true for the classical saving function where it is clear that demand does depend on the distribution of income so that in general no community indifference maps exist and there may be multiple momentary equilibria. Once again more detailed assumptions concerning elasticities of substitution or $k_C > k_I$ can rescue the situation. They really amount to the postulate of a certain kind of gross-substitutability (Hahn, 1965).

Once uniqueness of momentary equilibrium is assured it is not hard to show that the sequence of momentary equilibria approach the steady state (see Hahn and Matthews, 1964, for an intuitive account). For instance, for a classical saving postulate, *k*(0) must be inversely related to $\psi(k(0))$, the wage rental ratio. So if k^* is the steady state capital labour ratio, $\rho(k(0)) < \rho(k^*)$ whenever $k(0) > k^*$. But $\rho[k(0)] = K/K$ while $\rho(k^*) = \lambda$ hence

$$\frac{\dot{k}}{k} = \rho[k(0)] - \rho(k^*) < 0$$

and $k(0)$ in declining at $t = 0$. In fact the reader can check that $[k(t) - k^*]^2$ is always declining with t as long as $k(t) \neq k^*$ which suffices here to establish convergence to the steady state value k^*.
On the other hand, it should be noted that this argument is very much at risk when there is a variety of capital goods (see Hagemann, 1987).

2.3 Technical progress

With two sectors the nature of technological change in the economy as a whole will clearly depend on what kind of progress occurs in each of the sectors and on the composition of output. For instance, if by Harrod neutrality we mean that the capital/output ratio in value terms is constant when the rate of profit is constant we need to know how the capital/output ratio in each of the sectors is changing as well as what is happening to the relative outputs of the two sectors.
The case of disembodied technical progress is fully analysed in Diamond (1965) while there seems to be no literature on two-sector embodied technical progress.
As an example consider steady state with a proportional savings function. The value share of investment in output must remain constant. Technical progress in the investment sector will have to be Harrod-neutral because the rate of profit equality with the marginal product of capital is there independent of relative prices (input and output are the same). So in steady state the marginal product of capital should remain constant. If the capital labour ratio in both sectors remains constant then technical progress in the consumption goods sector must also be Harrod-neutral. Differences in the rate of technical progress in the two sectors will be reflected in a changing price of consumption good in terms of investment good. However, there could be steady state equilibrium with the labour allocation between the two sectors changing. In that case in general technical progress in the consumption good sector will not be Harrod-neutral.
It is not profitable to go into greater detail.

3 Many sectors

As long as one is only concerned with steady state equilibrium there is no difficulty for neoclassical theory when there are many sectors. Although it was somewhat special the foundations for the study of this case were laid by von Neumann (1945). (He assumed labour to be in infinitely elastic supply (in fact producible) at a given vector input of consumption goods. He also considered a ‘spectrum’ of techniques.) More recent formulations are best studied in Morishima (1964). For a survey see Hahn and Matthews (1964).
The essentials of this case can be illustrated for a classical savings function with only intermediate goods used in production (i.e. no long lived inputs) and no joint production.
Suppose there are N produced goods and one non-produced good (e.g. labour). Production takes time. Let q be the price vector of the N produced goods in terms of the non-produced good. Let all inputs be paid for when purchased and let $c(q)$ be the minimum unit cost function in terms of labour. That is $c(q)$ is the unit cost of production when inputs have been chosen to minimise costs. We can write it in this way because constant returns prevail everywhere. If that were not so there would be no hope of finding a

steady state equilibrium.
In such an equilibrium if all goods are produced and relative prices are constant it must be that

$$q = (1 + \rho)c(q). \tag{3.1}$$

If the economy is productive and indecomposable and every good needs labour in its production then one can solve (3.1) uniquely for $q(\rho) \gg 0$ provided ρ lies in some bounded interval. The function $q(\rho)$ is the *factor-price frontier*.
It is easy to prove that

$$\frac{\partial q_j}{\partial \rho} > 0. \tag{3.2}$$

Provided that the ratio in which wage earners consume goods depends only on q and not on their level of income one can now complete the story. The solution $q(\rho)$ is plainly independent of the scale or composition of output. So one can always make demand equal to supply in each sector provided there is enough labour in the economy. Suppose that labour is inelastically supplied. Then the scale of output can be anything. But if the ratio of employed to unemployed is to remain constant then output must grow at the rate λ hence so must investment and we get $\rho = \lambda$ as a further equilibrium condition. Relative prices will then be given by $q(\lambda)$. In equilibrium the present value of an input's marginal product will equal its price. Moreover ρ can be shown to measure the increase, at constant prices, in consumption made possible tomorrow if there is a little less consumption today and resources saved thereby are allocated efficiently.
An alternative scenario is to suppose that labour can always be had at a constant real wage w^* where the real wage is written as some function of q, say, $w(q)$. Then $w^* = w(q)$ together with (3.1) determine both q^* and ρ^* for steady state equilibrium. Given that there are classical savings the economy will grow at the rate ρ^* which will in fact be the highest (balanced) rate of growth the economy is capable of.
Perhaps a more general insight into these models can be gained as follows. Let Y and X be two n-vectors where the latter is the input of goods at one date and Y the output resulting at the subsequent date. Let L be the labour input. Then

$$T(Y, X, L) \geq 0 \tag{3.3}$$

is the economy's transformation locus which is homogeneous of degree one in its argument. Now a perfectly competitive economy is production efficient. So if all goods are produced in the steady state (Y^*/L^*, X^*/L^*) there must be prices q^* and profit rate p^* such that

$$q^*Y^* - (1+p^*)[q^*X^* + L^*] = 0 \tag{3.4}$$

is a supporting hyperplane of the set of (Y, X, L) satisfying (3.3) at (λ^*, X^*, L^*) Net output is $q^*(Y^* - X^*)$. If there are proportional savings at the rate s then one requires

$$sq^*(Y^* - X^*) = \lambda(q^*X^*) \tag{3.5}$$

if employment is to grow at the rate λ and Y/L and X/L are constant. But that is just the Harrod equation. Now

$$q^*Y^* - (1+p^*)[q^*X^* + L^*] \geq q^*Y - (1+p^*)[q^*X + L] \tag{3.6}$$

for all (Y, X, L) satisfying (3.3). Hence (3.4) is the maximum value of the r.h.s. of (3.6) subject to (3.3). Hence if T is differentiable:

$$q_i^* = \frac{T_{X_i}}{T_L} = -(1+p^*)\frac{T_{Y_i}}{T_L} \tag{3.7}$$

as can be verified by carrying out the maximization. Write (3.3) as

$$T(Y, kX, L) \geq 0$$
(3.8)

take $k = 1$ and differentiate with respect to k at (Y^*, X^*, L^*) to get

$$\left[\sum T_{Y_i}\frac{dY_i}{dk} + \sum T_{X_i}X_i\right]dk = 0.$$
(3.9)

Substitute from (3.7) into (3.9) writing

$$\Delta y_i = \frac{dY_i}{dk}dk, \Delta x_i = X_i dk,$$

to obtain

$$\sum q_i^* \Delta y_i = (1 + \rho^*)\sum q_i^* \Delta x_i$$

or

$$\frac{\Sigma q_i^* \Delta y_i - \Sigma q_i^* \Delta x_i}{\Sigma q_i^* \Delta x_i} = \rho^*$$
(3.10)

Hence the equilibrium rate of profit measures the increase in the value of net output at equilibrium prices as a fraction of the increase in the value of inputs at equilibrium prices. Or the rate of substitution between present and future consumption bundles of constant composition, evaluated at q^*. Of course,

there is no sense to the claim that (3.10) 'determines' ρ^*.

The literature on growth theory is vast and this essay can usefully be supplemented by other accounts such as Meade (1962), Hahn and Matthews (1964) and Solow (1970).

See Also

- classical growth model
- neoclassical growth theory (new perspectives)
- Ramsey model
- two-sector models
- von Neumann, John

Bibliography

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 28(3), 155–73.

Bliss, C.J. 1968. On putty-clay. *Review of Economic Studies* 35(2), 105–32.

Diamond, P. 1965. Disembodied technical change in a two-sector model. *Review of Economic Studies* 32 (2), 161–8.

Fellner, W. 1961. Two propositions in the theory of induced innovations. *Economic Journal* 71, 305–8.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21(1), 63–80.

Hagemann, H. 1987. Capital goods. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Hahn, F.H. 1951. The share of wages in national income. *Oxford Economic Papers* 3(2), 149–57.

Hahn, F.H. 1965. On two sector growth models. *Review of Economic Studies* 32(4), 339–46.

Hahn, F.H. 1987. 'Hahn problem'. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Hahn, F.H. and Matthews, R.C.O. 1964. The theory of economic growth: a survey. *Economic Journal* 74, 779–902. Reprinted in *Surveys of Economic Theory*, vol. 2. London: Macmillan 1965.

Harcourt, G.C. 1969. Some Cambridge controversies in the theory of capital. *Journal of Economic*

Literature 7(2), 369–405.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49, 14–33.

Inada, K. 1963. On a two-sector model of economic growth: comments and a generalisation. *Review of Economic Studies* 30, 119–27.

Inada, K. 1964. On the stability of growth equilibrium in two-sector models. *Review of Economic Studies* 31(2), 127–42.

Kaldor, N. 1955. Alternative theories of distribution. *Economic Journal* 23(2), 83–100.

Kaldor, N. and Mirrlees, J. 1962. A new model of economic growth. *Review of Economic Studies* 29(3), 174–92.

Kennedy, C. 1964. Induced bias in innovation and the theory of distribution. *Economic Journal* 74, 841–7.

Lucas, R. 1975. An equilibrium model of the trade cycle. *Journal of Political Economy* 83, 1113–44.

Meade, J.E. 1962. *A Neoclassical Theory of Economic Growth*. London: Allen & Unwin.

Meade, J.E. 1966. The outcome of the Pasinetti process: a note. *Economic Journal* 76, 161–5.

Meade, J.E. and Hahn, F.H. 1965. The rate of profit in a growing economy. *Economic Journal* 75, 445–48.

Mirrlees, J.A. 1969. The dynamic non-substitution th. *Review of Economic Studies* 36(1), 67–76.

Modigliani, F. and Samuelson, P.A. 1966. The Pasinetti Paradox in neo-classical and more general models. *Review of Economic Studies* 33, 269–301.

Morishima, M. 1964. *Equilibrium, Stability and Growth*. Oxford: Clarendon Press.

Neumann, J. von. 1945. A model of general economic equilibrium. *Review of Economic Studies* 13, 1–9.

Pasinetti, L.L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29(4), 267–79.

Robinson, J.V. 1938. The classification of inventions. *Review of Economic Studies* 5(2), 139–42.

Robinson, J.V. 1965. *The Accumulation of Capital*. 2nd edn. London: Macmillan.

Samuelson, P.A. 1957. Wages and interest: a modern dissection of Marxian economic models. *American Economic Review* 47, 884–912.

Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66, 467–82.

Samuelson, P.A. 1965. A theory of induced innovations along Kennedy–Weizsacker lines. *Review of Economics and Statistics* 47, 343–56.

Schumpeter, J.A. 1934. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70 (1), 65–94.

Solow, R.M. 1962. Comment (on Uzawa 1961). *Review of Economic Studies* 29(3), 255–7.

Solow, R.M. 1970. *Growth Theory: An Exposition*. Oxford: Clarendon Press.

Solow, R.M., Tobin, J., Weizsacker, C.C. von. and Yaari, M. 1967. Neo-classical growth with fixed proportions. *Review of Economic Studies* 33(2), 79–115.

Takayama, A. 1963. On a two-sector model of economic growth: a comparative statics analysis. *Review of Economic Studies* 36, 95–104.

Uzawa, H. 1961. On a two-sector model of economic growth. *Review of Economic Studies* 29(1), 40–47.

Varian, H. 1978. *Micro-economic Analysis*. New York: W.W. Norton.

How to cite this article

Hahn, F. "neoclassical growth theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000040> doi:10.1057/9780230226203.1170

The New Palgrave Dictionary of Economics Online

neoclassical synthesis

Olivier J. Blanchard

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘neoclassical synthesis’ appears to have been coined by Paul Samuelson to denote the consensus view of macroeconomics which emerged in the mid-1950s in the United States. This synthesis remained the dominant paradigm for another 20 years, in which most of the important contributions, by Hicks, Modigliani, Solow, Tobin and others, fit quite naturally. The synthesis had, however, suffered from the start from schizophrenia in its relation to microeconomics, which eventually led to a serious crisis from which it is only now re-emerging. I describe the initial synthesis, the mature synthesis, the crisis and the new emerging synthesis.

Keywords

aggregate demand; animal spirits; consumption function; countercyclical fiscal policy; dynamic stochastic general equilibrium (DSGE) models; econometrics; excess demand and supply; Hicks, J.; imperfect competition; imperfect financial markets; interest rates; intertemporal utility maximization; investment function; IS–LM model; Keynesianism; life-cycle hypothesis; liquidity constraints; Lucas, R.; market clearing; microfoundations; Modigliani, F.; monetary policy; MPS–FMP model; neoclassical synthesis; neoclassical theory of investment; new classical synthesis; new Keynesian synthesis; new neoclassical synthesis; nominal rigidities; permanent-income hypothesis; Phillips curve; Pigou effect; price and wage adjustment; public debt; rational behaviour; real business cycles; Samuelson, P.; stagflation; tâtonnement; wage rigidity; wealth effects

Article

The term ‘neoclassical synthesis’ appears to have been coined by Paul Samuelson to denote the consensus view of macroeconomics which emerged in the mid-1950s in the United States. In the third edition of *Economics* (1955, p. 212), he wrote:

> In recent years 90 per cent of American Economists have stopped being ‘Keynesian economists’ or ‘anti-Keynesian economists’. Instead they have worked toward a synthesis of whatever is valuable in older economics and in modern theories of income determination. The result might be called neo-classical economics and is accepted in its broad outlines by all but about 5 per cent of extreme left wing and right wing writers.

Unlike the old neoclassical economics, the new synthesis did not expect full employment to occur under laissez-faire; it believed, however, that, by proper use of monetary and fiscal policy, the old classical truths would come back into relevance.

This synthesis was to remain the dominant paradigm for another 20 years, in which most of the important contributions, by Hicks, Modigliani, Solow, Tobin and others, were to fit quite naturally. Its apotheosis was probably the large econometric models, in particular the MPS model developed by Modigliani and his collaborators, which incorporated most of these contributions in an empirically based and mathematically coherent model of the US economy. The synthesis had, however, suffered from the start from schizophrenia in its relation to microeconomics. This schizophrenia was eventually to lead to a serious crisis from which it is only now re-emerging. I describe in turn the initial synthesis, the mature synthesis, the crisis and the new emerging synthesis.

The initial synthesis

The post-war consensus was a consensus about two main beliefs. The first was that the decisions of firms and of individuals were largely rational, and as such amenable to study using standard methods from microeconomics. Modigliani, in the introduction to his collected papers, stated it strongly:

> [One of the] basic themes that has dominated my scientific concern [has been to integrate] the main building blocks of the General Theory with the more established methodology of economics, which rests on the basic postulate of rational maximizing behavior on the part of economic agents…’ (1980, p. xi)

The faith in rationality was far from blind: animal spirits were perceived as the main source of movements in aggregate demand through investment. For example, the possibility that corporate saving was too high and not offset by personal saving was considered a serious issue, and discussed on empirical rather than theoretical grounds.

This faith in rationality did not, however, extend to a belief in the efficient functioning of markets. The second main belief was indeed that prices and wages did not adjust very quickly to clear markets. There was broad agreement that markets could not be seen as competitive. But, somewhat surprisingly given the popularity of imperfect competition theories at the time, there was no attempt to think in terms of theories of price and wage setting, with explicit agents setting prices and wages. Instead, the prevailing mode of thinking was in terms of tâtonnement, with prices adjusting to excess supply or demand, along the lines of the dynamic processes of adjustment studied by Samuelson in his *Foundations of Economic Analysis* (1947). The Phillips curve, imported to the United States by Samuelson and Solow in 1960, was in that context both a blessing and a curse. It gave strong empirical support to a tâtonnement-like

relation between the rate of change of nominal wages and the level of unemployment, but it also made less urgent the need for better microeconomic underpinnings of market adjustment. Given the existence of a reliable empirical relation and the perceived difficulty of the theoretical task, it made good sense to work on other and more urgent topics, where the marginal return was higher.

These twin beliefs had strong implications for the research agenda as well as for policy. Because prices and wages eventually adjusted to clear markets, and because policy could avoid prolonged disequilibrium anyway, macroeconomic research could progress along two separate lines. One could study long-run movements in output, employment and capital, ignoring business cycle fluctuations as epiphenomena along the path and using the standard tools of equilibrium analysis: ‘Solving the vital problems of monetary and fiscal policy by the tools of income analysis will validate and bring back into relevance the classical verities’ (Samuelson, 1955, p. 360). Or one could instead study short-run fluctuations around that trend, ignoring the trend itself. This is indeed where most of the breakthroughs had been made by the mid-1950s. Work by Hicks (1937) and Hansen (1949), attempting to formalize the major elements of Keynes's informal model, had led to the IS–LM model. Modigliani (1944) had made clear the role played by nominal wage rigidity in the Keynesian model. Metzler (1951) had shown the importance of wealth effects, and the role of government debt. Patinkin (1956) had clarified the structure of the macroeconomic model, and the relation between the demands for goods, money and bonds, in the case of flexible prices and wages. There was general agreement that, except in unlikely and exotic cases, the IS curve was downward sloping and the LM curve upward sloping. Post-war interest rates were high enough – compared with pre-war rates – to make the liquidity trap less of an issue. There was still, however, considerable uncertainty about the effect of interest rates on investment, and thus about the slope of the IS relation. The assumption of fixed nominal wages made by Keynes and early Keynesian models had been relaxed in favour of slow adjustment of prices and wages to market conditions. This was not seen, however, as modifying substantially earlier conclusions. The ‘Pigou effect’ (so dubbed by Patinkin in 1948), according to which low enough prices would increase real money and wealth, was not considered to be of much practical significance. Only activist policy could avoid large fluctuations in economic activity.

Refinements of the model were not taken as implying that the case for policy activism was any less strong than Keynes had suggested. Because prices and wages did not adjust fast enough, active countercyclical policy was needed to keep the economy close to full employment. Because prices and wages, or policies themselves, eventually got the economy to remain not far from its growth path, standard microeconomic principles of fiscal policy should be used to choose the exact mix of fiscal measures at any point in time. The potential conflict between their relative efficacy in terms of demand management, and their effect on the efficiency of economic allocation, were considered an issue but not a major problem. Nor was the fact that the market failure which led to short-run fluctuations in the first place was not fully understood or even identified.

The ground rules for cyclical fiscal policy were laid in particular by Samuelson in a series of contributions (1951, for example). Countercyclical fiscal policy was to use both taxes and spending; in a depression, the best way to increase demand was to increase both public investment and private investment through tax breaks, so as to equalize social marginal rates of return on both. Where the synthesis stood on monetary policy is less clear. While the potential of monetary policy to smooth fluctuations was generally acknowledged, one feels that fiscal policy was still the instrument of

predilection, that policy was thought of as fiscal policy in the lead with accommodating monetary policy in tow.

The mature synthesis

For the next 20 years the initial synthesis was to supply a framework in which most macroeconomists felt at home and in which contributions fitted naturally. As Lucas remarks in his critique of the synthesis, ‘those economists, like Milton Friedman, who made no use of the framework, were treated with some impatience by its proponents’ (1980, p. 702). The research programme was largely implied by the initial synthesis, the emphasis on the behavioural components of IS–LM and its agnostic approach to price and wage adjustment; to quote Modigliani, ‘the Keynesian system rests on four basic blocks: the consumption function, the investment function, the demand and the supply of money, and the mechanisms determining prices and wages’ (1980, p. xii). Progress on many of these fronts was extraordinary; I summarize it briefly as these developments are reviewed in more depth elsewhere in this dictionary.

The failure of the widely predicted post-war over-saving to materialize had led to a reassessment of consumption theory. The theory of intertemporal utility maximization progressively emerged as the main contender. It was developed independently by Friedman (1957) as the ‘permanent income hypothesis’ and Modigliani and collaborators (1954 in particular) as the ‘life cycle hypothesis’. The life-cycle formulation, modified to allow for imperfect financial markets and liquidity constraints, was, however, to dominate most of empirical research. Part of the reason was that it emphasized more explicitly the role of wealth in consumption, and, through wealth, the role of interest rates. Neither wealth effects nor interest rate effects on consumption had figured prominently in the initial synthesis.

Research on the investment function was less successful. Part of the difficulty arose from the complexity of the empirical task, the heterogeneity of capital, and the possibility of substituting factors *ex ante* but not *ex post*. Many of the conceptual issues were clarified by work on growth, but empirical implementation was harder. Part of the difficulty, however, came from the ambiguity of neoclassical theory about price behaviour, about whether firms could be thought of as setting prices or whether the slow adjustment of prices implied that firms were in fact output constrained. The ‘neoclassical theory of investment’ developed by Jorgenson and collaborators (for example, Hall and Jorgenson, 1967) was ambiguous in this respect, assuming implicitly that price is equal to marginal cost, but estimating empirical functions with output rather than real wages.

Research on the demand for and supply of money was extended to include all assets. Solid foundations for the demand for money were given by Tobin (1956) and Baumol (1952), and the theory of finance provided a theory of the demand for all assets (Tobin, 1958). The expectations hypothesis, which alleviated the need to estimate full demand and supply models of financial markets, was thoroughly tested and widely accepted as an approximation to reality.

In keeping with the initial synthesis, work on prices and wages was much less grounded in theory than work on the other components of the Keynesian model. While research on the microeconomic foundations of wage and price behaviour was proceeding (Phelps, 1972 in particular), it was poorly integrated in empirical wage and price equations. To a large extent, this block of the Keynesian synthesis remained throughout the period the ad hoc but empirically successful Phillips curve,

respecified through time to allow for a progressively larger effect of past inflation on current wage inflation.

All these blocks, together with work on growth theory, were largely developed in relation with and then combined in macroeconometric models, starting with the models estimated by Klein (for example, Goldberger and Klein, 1955). The most important model was probably the MPS–FMP model developed by Modigliani and collaborators. This model, while maintaining the initial IS–LM Phillips curve structure of its ancestors, showed the richness of the channels through which shocks and policy could affect the economy. It could be used to derive optimal policy, show the effects of structural changes in financial markets, and so on. By the early 1970s the synthesis appeared to have been highly successful and the research programme laid down after the war to have been mostly completed. Only a few years later, however, the synthesis was in crisis and fighting for survival.

The crisis and the reconstruction

The initial trigger for the crisis was the failure of the synthesis to explain events. The scientific success of the synthesis had been largely due to its empirical success, especially during the Kennedy and the first phase of the Johnson administrations in the United States. As inflation increased in the late 1960s, the empirical success and, in turn, the theoretical foundations of the synthesis were more and more widely questioned. The more serious blow was, however, the stagflation of the mid-1970s in response to the increases in the price of oil: it was clear that policy was not able to maintain steady growth and low inflation. In a clarion call against the neoclassical synthesis, Lucas and Sargent (1978) judged its predictions to have been an ‘econometric failure on a grand scale’.

One cannot, however, condemn a theory for failing to anticipate the shape and the effects of shocks which have not been observed before; few theories would pass such a test and, as long as the events can be explained after the fact, there is no particular cause for concern. In fact, soon thereafter models were expanded to allow for supply shocks such as changes in the price of oil. It became clear, however, that while the models could indeed be adjusted *ex post*, there was a more serious problem behind the failure to predict the events of the 1970s. To quote again from the polemical article by Lucas and Sargent, ‘That the doctrine on which [these predictions] were made is fundamentally flawed is simply a matter of fact’ (1978, p. 49). The ‘fundamental flaw’ was the asymmetric treatment of agents as being highly rational and of markets as being inefficient in adjusting wages and prices to their appropriate levels. The tension between the treatment of rational agents and that of myopic impersonal markets had been made more obvious by the developments of the 1960s, and the representation of consumers and firms as highly rational intertemporal decision makers. It was further highlighted by the research on fixed price equilibria, which went to the extreme of taking prices as unexplained and solving for macroeconomic equilibrium under non-market clearing. That research made clear, in a negative way, that progress could be made only if one understood why markets did not clear, why prices and wages did not adjust.

The solution proposed by Lucas and others in the ‘new classical synthesis’ was thoroughly unappealing to economists trained in the neoclassical synthesis. It was to formalize the economy as if markets were competitive and clearing instantaneously. The ‘as if’ assumption seemed objectionable on a priori grounds, in that direct evidence on labour and goods markets suggested important departure from competition; it also appeared to many to be an unpromising approach if the goal was to explain

economic fluctuations and unemployment. Soon papers by Fischer (1977) and Taylor (1980) showed that one could replace the Phillips curve by a model of explicit nominal price and wage setting and still retain most of the traditional results of the neoclassical synthesis. These papers led the way to a major overhaul and reconstruction, and by the mid-1990s a new synthesis had emerged, a synthesis now dubbed the ‘new neoclassical synthesis’ (Goodfriend and King, 1997) or the ‘new Keynesian synthesis’ (for example, Clarida, Gali and Gertler, 1999). This new synthesis is described in more detail elsewhere in this dictionary, and I shall limit myself to a few remarks and comparisons between the old and the new. Like the old synthesis, the new synthesis has two major features: on the one hand, optimizing behaviour by firms, consumers and workers; on the other, the presence of distortions, most importantly nominal rigidities. In contrast to the old synthesis, however, the distortions are introduced explicitly, and price and wage behaviour is derived from optimizing behaviour by price and wage setters. These distortions imply that, as in the old synthesis, monetary policy and fiscal policy have a major role to play.

Like the old synthesis, the new synthesis is derived from microfoundations, utility maximization by consumers, and profit maximization by firms. But, while models in the old synthesis used theory as a loose guide to empirical specifications and allowed the data to determine the ultimate specification, models in the new synthesis remain much closer to their microfoundations. Dynamics are derived from the model itself, and the implied behavioural equations, rather than being estimated, are typically derived from assumptions about underlying technological and utility parameters. These more explicit microfoundations allow for a more careful welfare analysis of the implications of policy than was possible with the old models.

The models in the new synthesis are referred to as ‘dynamic stochastic general equilibrium’, or DSGE, models. Because they are typically difficult to solve, even the larger models are smaller than the models of the old synthesis, and their formalization of markets such as those for goods and labour remains primitive compared with the spirit of the formalizations in the old models. Improvements both in the formalization of these markets and in numerical techniques are, however, allowing for steadily richer and larger models.

To parallel the quotation from Samuelson given at the beginning, it is fair to say that the new neoclassical synthesis is attracting wide support, although less so than the old one. Some researchers, particularly those in the ‘real business cycle’ tradition, are sceptical about the importance of nominal rigidities in fluctuations. Others find the rationality assumptions embodied in the new synthesis to be too strong, and the methodology too constraining to capture the complexity present in the data. Nevertheless, DSGE models are increasingly used to guide policy. Many challenges remain, for example in capturing the relevant distortions in goods, labour, financial, and credit markets, or in using econometrics to assess the fit of both the specific components and the overall model to reality. Progress is rapid, however. When I wrote the first version of this contribution in 1991, the emergence of a new synthesis appeared uncertain, and at best far in the future. In updating this contribution, I am struck by the progress that has taken place since then, and by the speed at which progress continues to be made today.

See Also

- Friedman, Milton
- Hicks, John Richard
- Klein, Lawrence R.
- Lucas, Robert
- microfoundations
- Modigliani, Franco
- Patinkin, Don
- Phillips curve (new views)
- Samuelson, Paul Anthony
- Tobin, James

Bibliography

Baumol, W. 1952. The transactions demand for cash. *Quarterly Journal of Economics* 66, 545–6.

Clarida, R., Gali, J. and Gertler, M. 1999. The science of monetary policy: a New Keynesian perspective. *Journal of Economic Literature* 37, 1661–707.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85, 191–205.

Friedman, M. 1957. *A Theory of the Consumption Function*. New York: NBER.

Goldberger, A. and Klein, L. 1955. *An Econometric Model of the United States, 1929–1952*. Amsterdam: North-Holland.

Goodfriend, M. and King, R. 1997. The new neoclassical synthesis and the role of monetary policy. In *NBER Macroeconomics Annual 1997*, ed. B. Bernanke and J. Rotemberg. Cambridge: MIT Press.

Hall, R. and Jorgenson, D. 1967. Tax policy and investment behavior. *American Economic Review* 57, 391–414.

Hansen, A. 1949. *Monetary Theory and Fiscal Policy*. New York: McGraw-Hill.

Hicks, J. 1937. Mr Keynes and the ‘classics’: a suggested interpretation. *Econometrica* 5, 147–59.

Lucas, R. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12, 696–715.

Lucas, R. and Sargent, T. 1978. After Keynesian macroeconomics. In *After the Phillips Curve:*

Persistence of High Inflation and High Unemployment. Boston: Federal Reserve of Boston.

Metzler, L. 1951. Wealth, saving and the rate of interest. *Journal of Political Economy* 59, 93–116.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12, 45–88.

Modigliani, F. 1980. *Collected Papers*. Vol. 1: Essays in Macroeconomics. Cambridge, MA: MIT Press.

Modigliani, F. and Brumberg, R. 1954. Utility analysis and the consumption function: an interpretation of cross section data. In *Post-Keynesian Economics*, ed. K. Kurihara. New Brunswick, NJ: Rutgers University Press.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38, 543–64.

Patinkin, D. 1956. *Money, Interest and Prices*. New York: Harper and Row.

Phelps, E. 1972. *Inflation Policy and Unemployment Theory*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P. 1951. Principles and rules in modern fiscal policy: a neoclassical reformulation. In *Money, Trade and Economic Growth: Essays in Honor of John Henry Williams*, ed. H. Waitzman. New York: Macmillan.

Samuelson, P. 1955. *Economics*, 3rd edn. New York: McGraw-Hill.

Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38, 241–7.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, 65–86.

How to cite this article

Blanchard, Olivier J. "neoclassical synthesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000041> doi:10.1057/9780230226203.1172

The New Palgrave Dictionary of Economics Online

‘ neoclassical ’

Tony Aspromourgos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cambridge School; classical economics; Dobb, M. H.; hedonistic psychology; Hobson, J. A.; marginal productivity theory of distribution; marginalist theory; Marshall, A.; methodological individualism; Mitchell, W. C.; neoclassical; neoclassical economics; neoclassical synthesis; Roll, E.; subjective theory of value; utilitarianism; Veblen, T.

Article

The term ‘neoclassical’ was first used by Veblen (1900, pp. 242, 260–2, 265–8), in order to characterize Marshall and Marshallian economics. Veblen did not appeal to any similarity in theoretical structure between the economics of Marshall and classical economics in order to defend this novel designation. Rather, he perceived Marshall's Cambridge School to have a continuity with classical economics on the alleged basis of a common utilitarian approach and the common assumption of a hedonistic psychology. Derivative from Veblen's use, this meaning of the term subsequently gained some currency, particularly in the 1920s and 1930s; for example, in the writings of Wesley Mitchell, J.A. Hobson, Maurice Dobb and Eric Roll. It is evident that the emergence of this notion of Marshallian economics as a ‘neoclassical’ project also involved, at least in part, an acquiescence to Marshall's portrayal of his own economics as a continuation of the classical tradition, though Marshall's sense of the continuity is not really that perceived by Veblen. Keynes (1936, pp. 177–8) also employed the term, though in an idiosyncratic matter, derivative from his equally idiosyncratic notion of classical economics.
The use of the term with the meaning which became the accepted convention after the Second World War, extending it to embrace marginalist theory in general, can be traced to Hicks (1932, p. 84) and Stigler (1941, pp. 8, 13, 297). From what source they derived the term is not certain. It is highly unlikely that either of them coined it independently. Perhaps the likeliest source of Hicks's use is Dobb's article, published as it was in the London School of Economics' ‘house journal’, *Economica*. Following Hamilton (1923), Dobb (1924, p. 68) writes that ‘neo-classical’ is not an entirely inappropriate term to describe Marshallian economics, ‘for what the Cambridge School has done is to divest Classical

Political Economy of its more obvious crudities, to sever its connection with the philosophy of natural law, and to restate it in terms of the differential calculus. The line of descent is fairly direct from Smith, Malthus, and Ricardo’. Hicks's article, or Veblen, is the most likely source of Stigler's use. He refers to both of them. Hicks and Stigler were certainly more correct than Veblen in perceiving the unifying core of the marginalist theories to be, on the one hand, methodological individualism and on the other, the marginal productivity theory of distribution developed in connection with the subjective theory of value. However, neither of them offered any significant defence for their (then) implicit view that the writings of the classical economists also can be characterized in terms of this theoretical approach. Subsequently this characterization – and the nomenclature for marginalism associated with it – has given way to a recognition of the sharp theoretical disjuncture between classical and marginalist economics. Stigler's use, albeit hesitant, was probably as influential as his book. The term first gained wide currency in the debates on capital and growth in the 1950s and 1960s. It was no doubt also popularized by the extensive use made of it in Samuelson's textbook. From the third edition, Samuelson (1955, p. vi) presents the book as setting forth a ‘grand neoclassical synthesis’. (For a fuller account, see Aspromourgos, 1986.) The question may be raised whether the depiction of ‘neoclassical economics’ in the mid-20th century, understood as a characterization of the mainstream of the discipline, continues to represent an accurate picture of dominant beliefs within economics. Colander (2000), for example, has questioned this. But, even though the term was never sensible, the majority of the profession remains committed to the fundamental convictions which were at issue in those earlier capital and growth debates – in particular, the notion that competition brings about a tendency to full employment of resources (especially labour) and the marginal productivity theory of functional income distribution.

See Also

- Robinson Crusoe
- supply and demand

Bibliography

Aspromourgos, T. 1986. On the origins of the term ‘neoclassical’. *Cambridge Journal of Economics* 10, 265–70.

Colander, D. 2000. The death of neoclassical economics. *Journal of the History of Economic Thought* 22, 127–43.

Dobb, M. 1924. The entrepreneur myth. *Economica* 4, 66–81.

Hamilton, W.H. 1923. Vestigial economics. *New Republic*, 4 April.

Hicks, J.R. 1932. Marginal productivity and the principle of variation. *Economica* 12, 79–88.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Samuelson, P.A. 1955. *Economics: An Introductory Analysis*. 3rd edn, New York: McGraw-Hill.

Stigler, G.J. 1941. *Production and Distribution Theories*. New York: Macmillan.

Veblen, T.B. 1900. The preconceptions of economic science III. *Quarterly Journal of Economics* 14, 240–69.

How to cite this article

Aspromourgos, Tony. "‘neoclassical’." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000038> doi:10.1057/9780230226203.1169

The New Palgrave Dictionary of Economics Online

neo-Ricardian economics

Heinz D. Kurz and Neri Salvadori

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article deals with the revival of the classical theory of value and distribution, championed by Piero Sraffa. The general rate of profits and relative prices are shown to be determined exclusively in terms of the given system of production and real wages (or the share of wages). Prices generally depend on income distribution. So does the cost-minimizing technique. The ‘quantity of capital’ cannot be ascertained independently of prices and thus the rate of profits. Techniques cannot generally be ordered monotonically with the rate of profits. Marginalist ideas regarding input proportions and input prices therefore cannot generally be sustained.

Keywords

actual vs. normal values; Austrian economics; Cantillon, W.; capital accumulation; capital theory; circular flow of production; classical distribution theories; classical economics; cost-minimizing behaviour; division of labour; economic growth; endogenous growth; intertemporal equilibrium theory; labour theory of value; labour's share of income; laws of capitalism; long-period positions; marginalist theory of value and distribution; Marx, K.; methodological individualism; Mill, J.; neo-Ricardian economics; new growth theory; Petty, W.; physical real cost; profits; Quesnay, F.; reswitching; Ricardo, D.; Robinson, J.; Romer, P.; Say's Law; simultaneous equations; Smith, A.; social surplus; Sraffa, P.; stratification; technical change; Torrens, R.; uniform rate of profits; Wicksell effects

Article

The term ‘neo-Ricardian economics’, as it is understood today, can mean several things. It was coined in the aftermath of the publication of *The Works and Correspondence of David Ricardo*, edited by Piero Sraffa with the collaboration of Maurice H. Dobb (Ricardo, 1951–73), and the publication of Sraffa's *Production of Commodities by Means of Commodities* (Sraffa, 1960). One meaning of the term simply refers to these facts and interprets Sraffa's work in the way Sraffa himself saw it: as a return to the

‘standpoint of the old classical economists from Adam Smith to Ricardo, [which] has been submerged and forgotten since the advent of the “marginal’’ method’ (Sraffa, 1960, p. v; see Smith, 1776, and Ricardo, 1951–73). However, the term was first used by Marxist economists to distinguish Sraffa's approach to the theory of value and distribution, which explained relative prices and income distribution strictly in material terms (that is, quantities of commodities and labour), from the Marxist one, which starts from labour values (see Rowthorn, 1974). In some contributions Sraffa's analysis is described in a derogatory manner as a ‘peanut theory of profits’ and rejected together with marginalist (or ‘neoclassical’) theory as a variant of ‘vulgar economics’, dealing with ‘appearances’ only, whereas Marxist theory is taken to investigate ‘the real relations of production in bourgeois society’ (Marx, 1867, p. 85n). Neoclassical economists in turn occasionally (see, for example, Hahn, 1982) applied the term to the analysis of those critics who, in the so-called Cambridge controversies on the theory of capital, had attacked marginalism, especially its long-period version, showing it to be logically flawed (see Kurz and Salvadori, 1995, ch. 14). Because of the nationalities of the critics – especially Joan Robinson, Nicholas Kaldor, Piero Sraffa, Pierangelo Garegnani and Luigi Pasinetti – they also spoke of an ‘Anglo-Italian school’.

Such an unfortunate diversity of meanings may reflect a misunderstanding both of Sraffa's achievement and of the relation of his analysis to that of Marxist and marginalist economics respectively. What Sraffa in fact provides is a reformulation of the *classical* approach to the problem of value and distribution that sheds the weaknesses of its earlier formulations and builds upon their strengths. Put briefly, profits and all property incomes (such as interest and land rents) are explained in terms of the *social surplus* left over after the necessary means of production and the wages in the support of workers have been deducted from the gross outputs produced during a year. As Ricardo had stressed: ‘Profits come out of the surplus produce’ (*Works*, vol. 2, pp. 130–1; cf. vol. 1, p. 95). Therefore, instead of ‘neo-Ricardian economics’ it would be more appropriate to speak of that part of classical economics that deals with value and distribution. As is well known, this part was designed to constitute the foundation of all other economic analysis, including the investigation of capital accumulation and technical progress, of development and growth, of social transformation and structural change, and of taxation and public debt. The pivotal role of the theory of value and distribution in the classical authors can be inferred from the fact that it is typically developed at the beginning of their major works. By rectifying this part, Sraffa revived interest in classical economics. In addition to this constructive task Sraffa also pursued a critical task: the propositions of his book were explicitly ‘designed to serve as the basis for a critique of [the marginal theory of value and distribution]’ (1960, p. vi).

In the following we first summarize the achievements of Sraffa and his followers with respect to the constructive task. We then turn to the criticism of marginalist theory. In conclusion, we point out some of the problems that are currently being tackled by scholars working in the classical tradition.

Reformulating the classical theory of value and distribution

The concern of the classical economists, especially Smith and Ricardo, was the laws governing the emerging capitalist economy, characterized by the stratification of society into three classes: workers, landowners, and the rising class of capitalists; wage labour as the dominant form of the appropriation of other people's capacity to work; an increasingly sophisticated division of labour within and between

firms; the coordination of economic activity through a system of interdependent markets in which transactions were mediated through money; and significant technical, organizational and institutional change. In short, they were concerned with an economic system incessantly in motion. How to analyse such a system? The ingenious device of the classical authors to see through the complexities of the modern economy consisted in distinguishing between the 'actual' values of the relevant variables – the distributive rates and prices – and their 'normal' values. The former were taken to reflect all kinds of influences, many of an accidental or temporary nature, about which no general propositions were possible, whereas the latter were conceived of as expressing the persistent, non-accidental and non-temporary factors governing the economic system, which could be systematically studied.

The method of analysis adopted by the classical economists is known as the method of 'long-period positions' of the economy. Any such position is the situation towards which the system is taken to gravitate as the result of the self-seeking actions of agents, thereby putting into sharp relief the fundamental forces at work. In conditions of free competition the resulting long-period position is characterized by a *uniform rate of profits* (subject perhaps to persistent inter-industry differentials reflecting different levels of risk and of agreeableness of the business; see Kurz and Salvadori, 1995, ch. 11) and uniform rates of remuneration for each particular kind of primary input. Competitive conditions were taken to engender *cost-minimizing behaviour* of profit-seeking producers.

Alfred Marshall (1920) had interpreted the classical economists as essentially early and somewhat crude demand and supply theorists, with the demand side in its infancy. It was this interpretation and the underlying continuity thesis in economics that Sraffa challenged. As he showed, the classical economists' approach to the theory of value and distribution was fundamentally different from the later marginalist one, and explained profits in terms of basically two data: (*a*) the system of production in use and (*b*) a given real wage rate (or, alternatively, a given share of wages). Profits (and rents) were thus conceived of as a *residual* income. Whereas in marginalist theory wages and profits are treated symmetrically, in classical theory they are treated *asymmetrically*. On a still deeper methodological level the divide between the classical and the later marginalist authors could hardly be more pronounced. While the classical authors took the economic system to exist independently of the single agent and actually exert a considerable influence upon the latter depending upon the role ascribed to him as worker, capitalist or landowner, the marginalist authors advocated one version or another of 'methodological individualism', which takes a set of assumedly optimizing agents who exist independently of the system as a whole and who shape the system rather than the other way round.

Let us now examine more closely the scope, content and analytical structure of classical theory. The classical economists proceeded essentially in two steps. In the first step they isolated the kinds of factors that were seen to determine income distribution and the prices supporting that distribution in specified conditions, that is, *in a given place and time*. The theory of value and distribution was designed to identify *in abstracto* the dominant factors at work and to analyse their interaction. In the second step they turned to an investigation of the causes which *over time* affected systematically the factors at work from within the economic system. This was the realm of the classical analysis of capital accumulation, technical change, economic growth and socio-economic development.

It is another characteristic feature of the classical approach to profits, rents and relative prices that these are explained essentially in terms of magnitudes that can, in principle, be observed, measured or calculated. The *objectivist* orientation of classical economics has received its perhaps strongest expression in a famous proclamation by William Petty, who was arguably its founding father. Keen to

assume what he called the '"physician's" outlook', Petty in his *Political Arithmetick*, published in 1690, stressed that he was to express himself exclusively 'in Terms of *Number*, *Weight* or *Measure*' (Petty, 1986, p. 244). And James Mill noted significantly that '*The agents of production are the commodities themselves* …. They are the food of the labourer, the tools and the machinery with which he works, and the raw materials which he works upon' (Mill, 1826, p. 165, emphasis added). According to Sraffa the classical authors advocated essentially a concept of *physical real cost*. Man cannot create matter, man can only change its form and move it. Production involves destruction, and the real cost of a commodity consists in the commodities destroyed in the course of its production. This concept differs markedly from the later marginalist concepts, with their emphasis on 'psychic cost', reflected in such notions as 'utility' and 'disutility'.

In line with what may be called their 'thermodynamic' view, the classical authors saw production as a *circular flow*. This idea can be traced back to William Petty and Richard Cantillon, and was most effectively expressed by François Quesnay (1759) in the *Tableau économique*: commodities are produced by means of commodities. This is in stark contrast with the view of production as a one-way avenue leading from the services of original factors of production via some intermediate products to consumption goods, as was entertained by the 'Austrian' economists.

Why then did the classical economists fail to elaborate a consistent theory of value and distribution on the basis of the twin concepts of (*a*) physical real costs and (*b*) a circular flow of production? According to Sraffa (see Kurz and Salvadori, 2005) a main, if not *the* main, reason consisted in a mismatch between highly sophisticated analytical concepts on the one hand and inadequate tools available to the classical authors to deal with them on the other. More specifically, the tool needed in order to bring to fruition an analysis based on these twin concepts was simultaneous equations: knowledge of how to solve them and how to discover what their properties are. This indispensable tool (alas!) was not at their disposal. They therefore tried to solve the problems they encountered in a roundabout way, typically by first identifying an 'ultimate standard of value' by means of which *heterogeneous* commodities could be rendered *homogeneous*. Several authors, including Smith, Ricardo and Marx, had then reached the conclusion that 'labour' was the standard they sought and had therefore arrived in one way or another at some version of the labour theory of value. This preserved the objectivist character of the theory by taking as data, or known quantities, only measurable things, such as amounts of commodities actually produced and amounts actually used up, including the means of subsistence in the support of workers. This was understandable in view of the unresolved tension between concepts and tools. However, with production as a circular flow, even labour values cannot be known independently of solving a system of simultaneous equations. Hence the route via labour values was not really a way out of the impasse in which the classical authors found themselves: it rather landed them right in that impasse again. Commodities were produced by means of commodities and there was no way to circumnavigate the simultaneous equations approach.

What made it so difficult, if not impossible, for the classical authors to see that the theory of value and distribution could be firmly grounded in the concept of physical real cost? Given their primitive tools of analysis, they did not see that the information about the system of production in use and the quantities of the means of subsistence in support of workers was all that was needed in order to determine *directly* the system of necessary prices and the rate of profits. Sraffa understood this as early as November 1927, as we can see from his hitherto unpublished papers kept at Trinity College Library, Cambridge (UK), with

respect to what he called his ‘first’ (without a surplus) and ‘second’ (with a surplus) ‘equations’.
We may start with James Mill's aforementioned case with three kinds of commodities, tools (*t*), raw materials (*m*), and the food of the labourer (•*f*•). Production in the three industries may then be depicted by the following system of quantities

$$T_t \oplus M_t \oplus F_t \rightarrow T$$
$$T_m \oplus M_m \oplus F_m \rightarrow M$$
$$T_f \oplus M_f \oplus F_f \rightarrow F$$
(1)

where T_i, M_i and F_i designate the inputs of the three commodities (employed as means of production *and* means of subsistence) in industry $i\,(i = t, m, f)$, and *T*, *M* and *F* total outputs in the three industries; the symbol $\oplus$ indicates that all inputs on the LHS of $\rightarrow$, representing production are required to generate the output on its RHS. Invoking classical concepts, Sraffa called these relations ‘the methods of production and productive consumption’ (1960, p. 3). In the hypothetical case in which the economy is just viable, that is, able to reproduce itself without any surplus (or deficiency), we have $T = \Sigma_i T_i$, $M = \Sigma_i M_i$, and $F = \Sigma_i F_i$.
From this schema of reproduction and reproductive consumption we may directly derive the corresponding system of ‘absolute’ or ‘natural’ values, which expresses the idea of physical real cost-based values in an unadulterated way. Denoting the value of one unit of commodity *i* by $p_i\,(i = t, m, f)$, we have

$$T_t p_t + M_t p_m + F_t p_f = T p_t$$
$$T_m p_t + M_m p_m + F_m p_f = M p_m$$
$$T_f p_t + M_f p_m + F_f p_f = F p_f$$
(2)

These linear equations are homogeneous and therefore only relative prices can be determined. Further, only two of the three equations are independent of one another. This is enough to determine the two relative prices. Alternatively, it is possible to fix a standard of value whose price is *ex definitione* equal to unity. This provides an additional (non-homogeneous) equation without adding a further unknown, and allows one to solve for the remaining dependent variables.
A numerical example illustrates the important finding that the given socio-technical relations rigidly fix relative values:

$$\begin{array}{ll} & \textit{Values} \\ 2p_t + 15p_m + 20p_f = 17p_t & p_t = 3p_m \\ 5p_t + 7p_m + 4p_f = 28p_m & p_m = \frac{2}{3}p_f \\ 10p_t + 6p_m + 11p_f = 35p_f & p_f = \frac{1}{2}p_t \end{array}$$

These values depend exclusively on necessities of production. They are the only ones that allow the initial distribution of resources to be restored. Apparently, the value of one commodity may be ‘reduced’ to a certain amount of another commodity needed directly or indirectly in the production of the former. For example, one might reduce one unit of commodity *t* to an amount needed of commodity *m*. Hence one might say that each of the three commodities could serve as a ‘common measure’ and that, for example, commodities *t* and *f* exchange for one another in the proportion 1:2 because commodity *t* ‘contains’ or ‘embodies’ twice as much of commodity *m* as commodity *f*.

There is no need even to talk about labour values at this stage of the argument. The same applies to the next stage, which refers to a system with a surplus and given commodity (or real) wages advanced at the beginning of the production period. In conditions of free competition the surplus will be distributed in terms of a *uniform* rate of profits on the ‘capitals’ advanced in the different industries.

We start again from the system of quantities consumed productively and produced (1), but now we assume that $T \geq \Sigma_i T_i$, $M \geq \Sigma_i M_i$, and $F \geq \Sigma_i F_i$, where at least with regard to one commodity the strict inequality sign holds. In conditions of free competition ‘normal’ prices, or ‘prices of production’, have to satisfy the following system of price equations:

$$\begin{aligned} (T_t p_t + M_t p_m + F_t p_f)(1 + r) &= T p_t \\ (T_m p_t + M_m p_m + F_m p_f)(1 + r) &= M p_m \\ (T_f p_t + M_f p_m + F_f p_f)(1 + r) &= F p_f \end{aligned} \tag{3}$$

The case of a uniform rate of physical surplus across all commodities contemplated by David Ricardo and Robert Torrens

$$\frac{T - \Sigma_i T_i}{\Sigma_i T_i} = \frac{M - \Sigma_i M_i}{\Sigma_i M_i} = \frac{F - \Sigma_i F_i}{\Sigma_i F_i} = r \tag{4}$$

denotes a very special constellation: in it the general rate of profits, *r*, equals the uniform material rate of produce. *Here we see the rate of profits in the commodities themselves, as having nothing to do with their values.* In this case only two of the eqs. (3) are linearly independent so that eq. (4) determines the rate of profits, and eqs. (3), following the same procedure used for eqs. (2), determine relative prices. In general, the rates of physical surplus will be different for different commodities. Unequal rates of commodity surplus do not, however, by themselves imply unequal rates of profit across industries.

In this case there are three numbers, each of which substituted for *r* in eqs. (3) makes them linearly dependent on one another with respect to prices. It is possible to show that, when the highest real number among such numbers is substituted for *r*, the corresponding relative prices are positive, whereas when any of the other numbers is substituted for *r* some relative prices are negative. Since a negative relative price has no economic meaning in the present context, we can assert that there is a single solution which is relevant from an economic point of view. Fixing a standard of value provides a fourth equation and no extra unknown, so that the system of equations can be solved.

The important point to note here is the following. With the real wage rate given and paid at the beginning of the periodical production cycle, the problem of the determination of the rate of profits consists in distributing the surplus product in proportion to the capital advanced in each industry. Obviously,

> such a proportion between two aggregates of heterogeneous goods (in other words, the rate of profits) cannot be determined before we know the prices of the goods. On the other hand, we cannot defer the allotment of the surplus till after the prices are known, for … the prices cannot be determined before knowing the rate of profits. *The result is that the distribution of the surplus must be determined through the same mechanism and at the same time as are the prices of commodities*. (Sraffa, 1960, p. 6; emphasis added)

This passage shows that the idea which underlies Marx's so-called ‘transformation’ of labour values into prices of production (see Marx, 1894, part 2) cannot generally be sustained. Marx had proceeded in two steps; Ladislaus von Bortkiewicz (1906–7, essay 2, p. 38) aptly dubbed his approach ‘successivist’ (as opposed to ‘simultaneous’). In a first step Marx had assumed that the general rate of profits is determined independently of, and prior to, the determination of prices as the ratio between the labour value of the social surplus and that of social capital, consisting of ‘constant capital’ (means of production) and ‘variable capital’ (wages or means of subsistence). In a second step he had then used this rate to calculate prices.

So far we have assumed that real wages are given in kind at some level of subsistence. The classical economists, however, saw clearly that wages may rise above mere sustenance of labourers, which makes necessary a new wage concept. This case had made Ricardo adopt a *share* concept of wages and establish the inverse relationship between the share of wages in the product and the rate of profits: ‘The greater the *portion of the result of labour* that is given to the labourer, the smaller must be the *rate* of profits, and vice versa’ (*Works*, vol. 8, p. 194; emphasis added). The concept of ‘proportional wages’, as Sraffa called it, was then adopted by Marx in terms of a given rate of surplus value. Sraffa also adopted the concept, albeit with two important changes. First, when workers participate in the sharing out of the

surplus product, the original classical idea of wages being entirely paid out of social capital can no longer be sustained. After some deliberation Sraffa decided to treat wages as a whole as paid out of the product. Second, he did not express the share of wages in terms of labour but as the ratio of total wages to the net product expressed in terms of normal prices, *w*. These changes necessitated reformulating the price equations by taking explicitly into account the amounts of labour expended in the different industries, $L_i\,(i = t, m, f)$, because wages are taken to be paid in proportion to these amounts, and by defining these amounts as fractions of the total annual labour of society, that is, $L_t + L_m + L_f = 1$. In addition, it is assumed, following the classical economists, that differences in the quality of labour have been previously reduced to equivalent differences in quantity, so that each unit of labour receives the same wage rate (see Kurz and Salvadori, 1995, ch. 11). We may now formulate the corresponding system of production equations again for the case of the three kinds of commodities mentioned by Mill, where now the quantities represented by T_i, M_i and F_i refer exclusively to the inputs of the three commodities employed as means of production. We get (on the assumption that wages are paid *post factum*)

$$\begin{aligned}(T_t p_t + M_t p_m + F_t p_f)(1 + r) + L_t w &= T p_t\\ (T_m p_t + M_m p_m + F_m p_f)(1 + r) + L_m w &= M p_m\\ (T_f p_t + M_f p_m + F_f p_f)(1 + r) + L_f w &= F p_f\end{aligned} \tag{5.1}$$

With the net product taken as standard of value, we have in addition that

$$(T - \Sigma_i T_i)\, p_t + (M - \Sigma_i M_i)\, p_m + (F - \Sigma_i F_i)\, p_f = 1.$$

Taking one of the distributive variables, the share of wages *w* (or the rate of profits *r*) as given, allows one to determine the remaining variables: *r* (or *w*) and the prices of commodities.
Using this approach, Sraffa was able to show that, whereas the wage rate as a function of the rate of profits is necessarily decreasing (but does not need to be so if commodities are produced jointly), any relative price as a function of the rate of profits typically does not follow a simple rule: the function can alternately be increasing or decreasing, and can pass through unity a number of times (but such a number is constrained by the overall number of commodities involved). This fact is important also because the problem of the choice of technique from among several alternatives can be studied by following substantially the same argument. Suppose, for instance, that commodity *t* can be produced also with process

$$T_t' \oplus M_t' \oplus F_t' \oplus L_t' \rightarrow T'$$

Then we can add to system (5.1) the equation

$$(T_t' p_t + M_t' p_m + F_t' p_f)(1 + r) + L_t' w = T' p_t'$$
(5.2)

with the further unknown p_t'. The study of the ratio p_t' / p_t allows one to say when it is profitable to use the old process and when the new one: if p_t' / p_t is smaller than 1, the new process will be chosen by cost-minimizing producers; if it is larger than 1, the old process will be retained, whereas the two processes can coexist in case $p_t' / p_t = 1$. Obviously, if the new process is chosen and has replaced the old one, and if it is assumed that the rate of profits is unchanged, then eqs. (5.1) give way to the following equations, serving as the new system

$$(T_t' p_t' + M_t' p_m' + F_t' p_f')(1 + r) + L_t' w' = T' p_t'$$
$$(T_m p_t' + M_m p_m' + F_m p_f')(1 + r) + L_m w' = M p_m'$$
$$(T_f p_t' + M_f p_m' + F_f p_f')(1 + r) + L_f w' = F p_f'$$
(6.1)

In this new system prices and the wage are different ($p_j' \neq p_j$ and $w' \neq w$), but they are not so when $p_j' / p_t = 1$ in system (5). If we now evaluate the old process in terms of the prices and wage of the new system by combining system (6.1) and the equation

$$(T_t p_t' + M_t p_m' + F_t p_f')(1 + r) + L_t w' = T p_t$$
(6.2)

we can calculate again the ratio p_t' / p_t, and the property that prices and the wage in the two systems coincide when $p_t' / p_t = 1$ is enough to prove that p_t' / p_t is larger (lower) than 1 for a given *r* in system (6) if and only if it is so in system (5). Hence the comparison between the new process and the old one can be indifferently done at the prices of either the old system or the new system.

In the following a system involving a number of processes equal to the number of commodities involved, each producing a different commodity, is called a *technique*, and a technique which is chosen at a given income distribution is called a *cost-minimizing technique* at that income distribution. The fact that a relative price can pass through unity at several income distributions implies that a technique can be cost-minimizing at different values of the rate of profits, with other techniques being cost minimizing in the interval in between. This fact has been called *reswitching*; it played an important role in the criticism of neoclassical theory.

In the above it has for simplicity been assumed that there is only single production, that is, only circulating capital. While the circulating part of the capital goods advanced in production contributes entirely and exclusively to the output generated, that is, ‘disappears’ from the scene, so to speak, the fixed part of it contributes to a sequence of outputs over time, that is, after a single round of production its items are still there – older but still useful. For a discussion of joint production, fixed capital and scarce natural resources, see Kurz and Salvadori (1995).

Critique of marginalist theory

The passage quoted above from Sraffa (1960, p. 6) contains the key to his critique of the long-period marginalist concept of capital. This concept hinges crucially on the possibility of defining the ‘quantity of capital’, whose relative scarcity and thus marginal productivity was taken to determine the rate of profits, independently of the rate of profits. However, according to the logic of Sraffa's above argument the rate of profits and the quantity (that is, value) of social capital $(\Sigma_i T_i p_t + \Sigma_i M_i p_m + \Sigma_i F_i p_f)$ can only be determined simultaneously.

We may approach the issues under consideration by first discussing what are known as ‘Wicksell effects’. The term was introduced by Joan Robinson (1953, p. 95) during a debate in the theory of capital (see Kurz and Salvadori, 1995, ch. 14). We distinguish between *price Wicksell effects* and *real Wicksell effects* (henceforth PWE and RWE). A PWE relates to a change in relative prices corresponding to a change in income distribution, given the system of production in use. A RWE relates to a change in technique, with the fact taken into account that at the income distribution at which two techniques are both cost-minimizing (one being so at higher, the other at lower levels of the rate of profits) both techniques have the same prices. The ‘changes’ under consideration refer to comparisons of long-period equilibria.

Marginalist theory contends that both effects are invariably positive. A *positive* PWE means that with a rise (fall) in the rate of interest prices of consumption goods will tend to rise (fall) relative to those of capital goods. The reason given is that consumption goods are said to be produced more capital intensively than capital goods: consumption goods emerge at the end of the production process, whereas capital goods are intermediate products that gradually ‘mature’ towards the final product. The higher (lower) is the rate of interest the less (more) expensive are the intermediate products in terms of a

standard consisting of a (basket of) consumption good(s). At the macro level of a stationary economy (in which the net product contains only consumption goods) this implies that with a rise in the rate of interest the value of the net social product rises relatively to the value of the aggregate of capital goods employed. Clearly, seen from the marginalist perspective, a positive PWE with regard to the relative price of the two aggregates under consideration involves a negative relationship between the aggregate capital-to-net output ratio on the one hand and the interest rate on the other. Let $K / Y = \mathbf{x}\mathbf{p}(r) / \mathbf{y}\mathbf{p}(r)$ (**x** is the row vector of capital goods, **y** the row vector of net outputs, and **p**(*r*) the column vector of prices (in terms of the consumption vector) which depends on *r*) designate the capital-output ratio, then the marginalist message is:

$$\frac{\partial (K / Y)}{\partial r} \leq 0$$

Since for a given system of production the amount of labour is constant irrespective of the level of the rate of interest, also the ratio of the value of the capital goods and the amount of labour employed, or capital–labour ratio, *K/L*, would tend to fall (rise) with a rise (fall) in the rate of interest,

$$\frac{\partial (K / L)}{\partial r} \leq 0 \qquad (7)$$

This is the first claim marginalist authors put forward. The second is that RWEs are also positive. A *positive* RWE means that with a rise (fall) in the rate of interest cost-minimizing producers switch to methods of production that generally exhibit higher (lower) labour intensities, ‘substituting’ for the ‘factor of production’ that has become more expensive – ‘capital’ (labour) – the one that has become less expensive – labour (‘capital’). Hence (7) is said to apply also in this case. The assumed positivity of the RWE underlies the marginalist concept of a demand function for labour (capital) that is inversely related to the real wage rate (rate of interest).

Careful scrutiny of the marginalist argument has shown that it cannot generally be sustained: there is no presumption that PWEs and RWEs are invariably positive. In fact there is no presumption that techniques can be ordered monotonically with the rate of interest (Sraffa, 1960). Reswitching implies that, even if PWEs happen to be positive, RWEs cannot always be positive. As Mas-Colell (1989) stressed, the relationship between *K/L* and *r* can have almost any shape whatsoever. In the intervals in which *K/L* is an increasing function of *r* we say that there is *capital reversal*. It implies that, if the neoclassical approach to value and distribution is followed, the ‘demand for capital’ is not decreasing, and therefore the resulting equilibrium, provided there is one, is not stable. Hence the finding that PWEs and RWEs need not be positive challenges the received doctrine of the working of the economic system, as it is portrayed by conventional economic theory with its reference to the ‘forces’ of demand and

supply (see Pasinetti, 1966; Garegnani, 1970; see also Harcourt, 1972; Kurz and Salvadori, 1995, ch. 14; 1998c).

Current work in the classical tradition

In more recent times authors working in the classical tradition, as it was revived by Sraffa, have focused attention on a large number of problems. First, there has been a lively interest in generalizing the results provided by Sraffa on joint production, fixed capital, and land. Then the approach was extended to cover renewable and exhaustible resources and to allow for the more realistic case of costly disposal, which leads to the concept of negative prices of products that have to be disposed of. There is also a renewed interest in the problem of economic growth and development. Freed from the straightjacket of Say's Law, which can be said to be an implication of the finding that conventional equilibrium analysis cannot be sustained, there is no presumption that the economy will consistently follow a full-capacity path of economic expansion. Hence the problem of different degrees and modes of utilization of productive capacity and the role of effectual demand (Adam Smith) have to be analysed. This avenue has opened up avenues for cross-fertilization between classical economics on the one hand, and Keynesian economics, based on the principle of effective demand, and evolutionary economics, concerned with complex dynamics, on the other (see Coase, 1976; Nelson, 2005). This fact is also highlighted in comparisons with the so-called new growth theory, and allows one to better understand the latter's merits and demerits (see Kurz and Salvadori, 1998a, ch. 4; 1999).

In the 1960s and 1970s the long-period versions of marginalist theory revolving around the concept of a uniform rate of return on capital were called into question on logical grounds. While many marginalist authors accepted this criticism, some of them contended that intertemporal equilibrium theory, the ‘highbrow version’ of neoclassicism, was not affected by it (see especially Bliss, 1975; Hahn, 1982). This claim has more recently been subjected to close scrutiny (see Garegnani, 2000, Schefold, 2000, and the special issue of *Metroeconomica*, vol. 56(4), 2006). While the criticism of the long-period versions of marginalist theory is irrefutable, as authors from Paul Samuelson to Andreu Mas-Colell have admitted, surprisingly this has not prevented the economics profession at large from still using this theory. This is perhaps so because in more recent years the way of theorizing in large parts of mainstream economics has fundamentally changed. Whether this change is a response to the criticism need not concern us here. It suffices to draw the reader's attention to a statement by Paul Romer in one of his papers on endogenous growth in which he self-critically pointed out a slip in his earlier argument. The error he had committed, he wrote, ‘may seem a trifling matter in an area of theory that depends on so many other short cuts. After all, if one is going to do violence to the complexity of economic activity by assuming that there is an aggregate production function, how much more harm can it do to be sloppy about the difference between rival and nonrival goods?’ (Romer, 1994, pp. 15–16) Once economic theory has taken the road indicated, criticism becomes a barren instrument. Indeed, why should someone who seeks to provide ‘microfoundations’ in terms of a representative agent with an infinite time horizon find fault with the counter-factual but attractive assumption that there is only a single (capital) good?

See Also

- capital theory
- capital theory (paradoxes)
- classical distribution theories
- classical growth model
- classical production theories
- reswitching of technique
- Ricardo, David
- Smith, Adam
- Sraffa, Piero
- Sraffian economics
- Sraffian economics (new developments)

Bibliography

Bliss, C. 1975. *Capital Theory and the Distribution of Income*. Amsterdam: North-Holland.

Bortkiewicz, L. von. 1906–7. Wertrechnung und Preisrechnung im Marxschen System. *Archiv für Sozialwissenschaft und Sozialpolitik* 23 (1906), 1–50 (essay 1), 25 (1907), 10–51 (essay 2) and 445–88 (essay 3).

Coase, R. 1976. Adam Smith's view of man. *Journal of Law and Economics* 19, 529–46.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37, 407–36.

Garegnani, P. 1987. Surplus approach to value and distribution. In *The New Palgrave: A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Garegnani, P. 2000. Savings, investment and the quantity of capital in general intertemporal equilibrium. In Kurz (2000).

Hahn, F. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6, 353–74.

Harcourt, G. 1972. *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.

Kurz, H., ed. 2000. *Critical Essays on Piero Sraffa's Legacy in Economics*. Cambridge: Cambridge University Press.

Kurz, H. and Salvadori, N. 1995. *Theory of Productio: A Long-period Analysis*. Cambridge: Cambridge

University Press.

Kurz, H. and Salvadori, N. 1998a. *Understanding 'Classical' Economics: Studies in Long-Period Theory*. London: Routledge.

Kurz, H. and Salvadori, N., eds. 1998b. *The Elgar Companion to Classical Economics*, 2 vols. Cheltenham and Northhampton, MA: Edward Elgar.

Kurz, H. and Salvadori, N. 1998c. Reverse capital deepening and the numeraire: a note. *Review of Political Economy* 10, 415–26.

Kurz, H. and Salvadori, N. 1999. Theories of 'endogenous' growth in historical perspective. In *Contemporary Economic Issues. Proceedings of the Eleventh World Congress of the International Economic Association, volume 4: Economic Behaviour and Design*, ed. M. Sertel. London: Macmillan.

Kurz, H. and Salvadori, N. 2005. Representing the production and circulation of commodities in material terms: on Sraffa's objectivism. *Review of Political Economy* 17, 414–41.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Marx, K. 1867. *Capital*, vol. 1. Moscow: Progress Publishers, 1954.

Marx, K. 1894. 1959. *Capital*, vol. 3. Moscow: Progress Publishers, 1959.

Mas-Colell, A. 1989. Capital theory paradoxes: anything goes. In *Joan Robinson and Modern Economic Theory*, ed. R. Feiwel. London: Macmillan.

Mill, J. 1826. *Elements of Political Economy*, 3rd edn, reprinted 1844. London: Baldwin, Cradock, and Joy.

Nelson, R. 2005. *Technology, Institutions, and Economic Growth*. Cambridge, MA, and London: Harvard University Press.

Petty, W. 1986. *The Economic Writings of Sir William Petty*. New York: Kelley.

Quesnay, F. 1759. *Quesnay's Tableau Economique*, ed. M. Kuczynski and R. Meek. London: Macmillan, 1972.

Pasinetti, L. 1966. Changes in the rate of profit and switches of techniques. *Quarterly Journal of Economics* 80, 503–17.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*, 11 vols, ed. P. Sraffa with the collaboration of M. Dobb. Cambridge: Cambridge University Press. (In the text referred to as *Works*, volume number.)

Robinson, J. 1953. The production function and the theory of capital. *Review of Economic Studies* 21, 81–106.

Romer, P. 1994. The origins of endogenous growth. *Journal of Economic Perspectives* 8(1), 3–22.

Rowthorn, R. 1974. Neo-classicism, neo-Ricardianism and Marxism. *New Left Review* 86, 63–87.

Schefold, B. 2000. Paradoxes of capital and counter-intuitive changes of distribution in an intertemporal equilibrium model. In Kurz (2000).

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Oxford University Press, 1976.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

How to cite this article

Kurz, Heinz D. and Neri Salvadori. "neo-Ricardian economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000150> doi:10.1057/9780230226203.1173

The New Palgrave Dictionary of Economics Online

network formation

Matthew O. Jackson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A brief introduction and overview of models of the formation of networks is given, with a focus on two types of model. The first views networks as arising stochastically, and uses random graph theory, while the second views the links in a network as social or economic relationships chosen by the involved parties, and uses game theoretic reasoning.

Keywords

clustering; degree distributions; graph theory; Myerson value; network formation; pairwise stability; random graphs; small worlds

Article

A growing literature in economics examines the formation of networks and complements a rich literature in sociology and recently emerging literatures in computer science and statistical physics. Research on network formation is generally motivated by the observation that social structure is important in a wide range of interactions, including the buying and selling of many goods and services, the transmission of job information, decisions on whether to undertake criminal activity, and informal insurance networks.

Networks are often modelled using tools and terminology from graph theory. Most models of networks view a network as either a non-directed or a directed graph; which type of graph is more appropriate depends on the context. For instance, if a network is a social network of people and links represent friendships or acquaintances, then it would tend to be non-directed. Here the people would be modelled as the nodes of the network and the relationships would be the links. (In terms of a graph, the people would be vertices and the relationships would be edges.) If, instead, the network represents citations from one article to another, then each article would be a node and the links would be directed, as one article could cite another. While many social and economic relationships are reciprocal or require the consent of both parties, there are also enough applications that take a directed form, so that both non-

directed and directed graphs are useful as modelling tools.

Models of how networks form can be roughly divided into two classes. One derives from random graph theory, and views an economic or social relationship as a random variable. The other views the people (or firms or other actors involved) as exercising discretion in forming their relationships, and uses game theoretic tools to model formation. Each of these techniques is discussed in turn.

Models of random networks

Bernoulli random graphs

Some of the earliest formal models used to understand the formation of networks are random graphs: the canonical example is that of a pure Bernoulli process of link formation (for example, see the seminal study of Erdös and Rényi, 1960). For instance, consider a network where the (non-directed) link between any two nodes is formed with some probability p (where $1>p>0$), and this process occurs independently across pairs of nodes. While such a random method of forming links allows any network to potentially emerge, some networks are much more likely to do so than others. Moreover, as the number of nodes becomes large, there is much that can be deduced about the structure the network is likely to take, as a function of p. For instance, one can examine the probability that the resulting network will be connected in the sense that one can find a path (sequence of links) leading from any given node to any other node. We can also ask what the average distance will be in terms of path length between different nodes, among other things. As Erdös and Rényi showed, such a random graph exhibits a number of ‘phase’ transitions as the probability of forming links, p, is varied in relation to the number of nodes, n; that is, resulting networks exhibit different characteristics depending on the relative sizes of p and n.

Whether or not such a uniformly random graph model is a good fit as a model of network formation, it is of interest because it indicates that networks with different densities of links might tend to have very different structures and also provides some comparisons for network formation processes more generally. Some of the basic properties that such a random graph exhibits can be summarized as follows. When p is small in relation to n, so that $p<1/n$ (that is, the average number of links per node is less than one), then with a probability approaching 1 as n grows the resulting graph consists of a number of disjointed and relatively small components, each of which has a tree-like structure. (A component of a network is a subgraph, so that each node in the subgraph can be reached from any other node in the subgraph via a path that lies entirely in the subgraph, and there are no links between any nodes in the subgraph and any nodes outside the subgraph.) Once p is large enough in relation to n, so that $p>1/n$, then a single ‘giant component’ emerges; that is, with a probability approaching 1 the graph consists of one large component, which contains a nontrivial fraction of the nodes, and all other components are vanishingly small in comparison. Why there is just one giant component and all other components are of a much smaller order is fairly intuitive. In order to have two ‘large’ components each having a nontrivial fraction of n nodes, there would have to be no links between any node in one of the components and any node in the other. For large n, it becomes increasingly unlikely to have two large components with absolutely no links between them. Thus, nontrivial components mesh into a giant component, and any other components must be of a much smaller order. As p is increased further, there is another phase transition when p is proportional to $log(n)/n$. This is the threshold at which the network becomes

‘connected’ so that all nodes are path-connected to each other and the network consists of a single component. Once we hit the threshold at which the network becomes connected, we also see further changes in the diameter of the network as we continue to increase p relative to n. (The diameter is the maximal distance between two nodes, where distance is the minimal number of links that are needed to pass from one node to another.) Below the threshold, the diameter of a giant component is of the order of $log(n)$, then at the threshold of connectedness it hits $log(n)/loglog(n)$, and it continues to shrink as p increases.

Similar properties and phase transitions have been studied in the context of other models of random graphs. For example, Molloy and Reed (1995), among others (see Newman, 2003), have studied component size and connectedness in a ‘configuration model’. There, a set of nodes is given together with the number of links that each node should have, and then links are randomly formed to leave each node with the pre-specified number of links.

Clustering and Markov graphs

Although the random graphs of Erdös and Rényi are a useful starting point for modelling network formation, they lack many characteristics observed in most social and economic networks. This has led to a series of richer random graph-based models of networks. The most basic property that is absent from such random networks is that the presence of links tends to be correlated. For instance, social networks tend to exhibit significant clustering. Clustering refers to the following property of a network. If we examine triples of nodes so that two of them are each connected to the third, what is the frequency with which those two nodes are linked to each other? This tends to be much larger in real social networks than one would see in a Bernoulli random graph. On an intuitive level, models of network formation where links are formed independently tend to look too much like ‘trees’, while observed social and economic networks tend to exhibit substantial clustering, with many more cycles than would be generated at random (see Watts, 1999, for discussion and evidence).

Frank and Strauss (1986) identified a class of random graphs that generalize Bernoulli random graphs, which they called ‘Markov graphs’ (also referred to as p^* networks). Their idea was to allow the chance that a given link forms to be dependent on whether or not neighbouring links are formed. Specific interdependencies require special structures, because, for instance, making one link dependent on a second, and the second on the third, can imply some interdependencies between the first and third. These sorts of dependencies are difficult to analyse in a tractable manner, but nevertheless some special versions of such models have been useful in statistical estimation of networks.

Small worlds

Another variation on a Bernoulli network was explored by Watts and Strogatz (1998) in order to generate networks that exhibit both relatively low distances (in terms of minimum path length) between nodes and relatively high clustering – two features that are present in many observed networks but not in the Bernoulli random graphs unless the number of links per node (p(n−1)) is extremely high. They started with a very structured network that exhibits a high degree of clustering. Then, by randomly rewiring enough (but not too many) links, one ends up with a network that has a small average distance

between links but still has substantial clustering. While such a rewiring process results in networks that exhibit some of the features of social networks, it leads to networks that miss out on other basic characteristics that are present in many social networks. For example, the nodes of such a network tend to be too similar in terms of the number of links that they each have.

Degree distributions

One fundamental characteristic of a social network is a network's degree distribution. The degree of a node is the number of links it has, and the degree distribution keeps track of how varied the degree is across the nodes of the network. That is, the degree distribution is simply the frequency distribution of degrees across nodes. For instance, in a friendship network some individuals might have only a few friends while other individuals might have many, and then the degree distribution quantifies this information.

Price (1965) examined a network of citations (between scientific articles), and found that the degree distribution exhibited ‘fat tails’ compared with what one would observe in a Bernoulli random graph; that is, there was a higher frequency of articles that had many citations and a higher frequency of articles that had no citations than should be observed if citations were generated independently. In fact, many social networks exhibit such fat tails, and some have even been thought to exhibit what is known as a ‘scale-free’ degree distribution or said to ‘follow a power law’. A scale-free distribution is one where the frequency of degrees can be written in the form $f(d) = ad^{-b}$, for some parameters *a* and *b*, where *d* is the degree and *f*(*d*) is the relative frequency of nodes with degree *d*. Such distributions date to Pareto (1896), and have been observed in a variety of other contexts ranging from the distribution of wealth in a society to the relative use of words in a language. Price (1976) adapted ideas from Simon (1955) to develop a random link formation process that produces networks with such degree distributions. A similar model was later studied by Barabási and Albert (2001), who called the process of link formation ‘preferential attachment’. The idea is that nodes gain new links with probabilities that are proportional to the number of links they already have (which is closely related to a lognormal growth process). In a system where new nodes are born over time, this process generates scale-free degree distributions.

A simple preferential attachment model also has its limitations. One is that most social networks do not in fact have degree distributions that are scale-free. Observed degree distributions tend to lie somewhere between the extremes of a scale-free distribution and that corresponding to an independent Bernoulli random graph (sometimes known as a Poisson random graph for its approximate degree distribution). Second, the preferential attachment model fails to produce the type of clustering observed in many social networks, just as Bernoulli random graphs do. This has led to the construction of hybrid models that allow for richer sets of degree distributions, as well as clustering and correlation in degrees, and allows for the structural fitting of random graph based network formation models to data (for example, see Jackson and Rogers, 2007, and the discussion there).

Strategic models of network formation

Strategic models of network formation have emerged from the economics literature, and offer a very different perspective from that seen in random graph models, and a complementary set of insights (see

Jackson, 2006, for comparison and discussion). The starting point for a game theoretic approach is to assume that the nodes are active discretionary agents or players who get payoffs that depend on the social network that emerges. For example, if nodes are countries and links are political alliances, or nodes are firms and links are trading or collaboration agreements, then the relationships are entered into with some care and thought. Even in modelling something like a friendship network, while individuals might not be directly calculating costs and benefits from the relationship, they do react to how enjoyable or worthwhile the relationship is and might tend to spend more effort or time in relationships that are more beneficial and avoid ones that are less so. Different social networks lead to different outcomes for the involved agents (for example, different trades, different access to information or favours, and so on). Links are then formed at the discretion of the agents, and various equilibrium notions are used to predict which networks will form. This differs from the random models not only in that links result as a function of decisions rather than at random, but also in that there are natural costs and benefits associated with networks which then allow a welfare analysis.

Some of the first models to bring explicit utilities and choice to the formation of social links were in the context of modelling the trade-offs between ‘strong’ and ‘weak’ ties (links) in labour contact networks. Such models by Boorman (1975) and Montgomery (1991) explored a theory, due to Granovetter (1973), about different strengths of social relationships and their role in finding employment. Granovetter observed that when individuals obtained jobs through their social contacts, while they sometimes did so through strong ties (people whom they knew well and interacted with on a frequent basis), they also quite often obtained jobs through weak ties (acquaintances whom they knew less well and/or interacted with relatively infrequently). This led Granovetter to coin the phrase ‘the strength of weak ties’. Boorman's article and Montgomery's articles provided explicit models where costs and benefits could be assigned to strong and weak ties, and trade-offs between them could be explored.

In a very different setting, another use of utility functions involving networks emerged in the work of Myerson (1977). Myerson analysed a class of cooperative games that were augmented with a graph structure. In these games the only coalitions that could produce value are those that are pathwise connected by the graph, and so such graphs indicate the possible cooperation or communication structures. This approach led Myerson to characterize a variation on the Shapley value, now called the Myerson value, which was a cooperative game solution concept for the class of cooperative games where constraints on coalitions were imposed by a graph structure. Although the graphs in Myerson's analysis are tools to define a special class of cooperative games, they allow the graph structure to influence the allocation of societal value among a set of players. Aumann and Myerson (1988), recognizing that different graph structures led to different allocations of value, used this to study a game where the graph structure was endogenous. They studied an extensive form game where links are considered one by one according to some exogenous order, and formed if both agents involved agree. While that game turns out to be hard to analyse even in three-person examples, it was an important precursor to the more recent economic literature on network formation.

In contrast to the cooperative game setting, Jackson and Wolinsky (1996) explicitly considered networks, rather than coalitions, as the primitive. Thus, rather than deducing utilities indirectly through a cooperative game on a graph, they posited that networks were the primitive structure and agents derived utilities based on the network structure in place. So, once a social network structure is in place, one can then deduce what the agent's payoffs will be. Using such a formulation where players' payoffs are determined as a function of the social network in place, it is easy to model network formation using

game theoretic techniques.

Pairwise stability

In modelling network formation from a game theoretic perspective, one needs to have some notion of equilibrium or stable networks. Since it is natural to require mutual consent in many applications, standard Nash equilibrium based ideas are not very useful. For instance, consider a game where each agent simultaneously announces which other agents he or she is willing to link to. It is always a Nash equilibrium for each agent to say that he or she does not want to form any links, anticipating that the others will do the same. Generally, this allows for a multiplicity of equilibria, many of which make little sense from a social network perspective. Even equilibrium refinements (such as undominated Nash or perfect equilibrium) do not avoid this problem. Given that it is natural in a network setting for the agents prospectively forming a link to be able to communicate with each other, they should also be able to coordinate with each other on the forming of a link. An approach taken by Jackson and Wolinsky (1996) is to define a stability notion that directly incorporates the mutual consent needed to form links. Jackson and Wolinsky (1996) defined the following notion of ‘pairwise stability’: a network is pairwise stable if (i) no player would be better off if he or she severed one of his or her links, and (ii) no pair of players would both benefit (with at least one of the pair seeing a strict benefit) from adding a link that is not in the network. The requirement that no player wishes to delete a link that he or she is involved in implies that a player has the discretion to unilaterally terminate relationships that he or she is involved in. The second part of the definition captures the idea that if we are at a network where the creation of a new link would benefit both players involved, then the network g is not stable, as it will be in the players' interests to add the link.

Pairwise stability is a fairly permissive stability concept – for instance, it does not consider deviations where players delete some links and add others at the same time. While pairwise stability is easy to work with and often makes fairly pointed predictions, the consideration of further refinements can make a difference. A variety of refinements and alternative notions have been introduced, including allowing agents to form and sever links at the same time, allowing coalitions of agents to add and sever links in a coordinated fashion, or behaviour where agents anticipate how the formation of one link might influence others to form further links (see Jackson, 2004, for discussion and references). There are also dynamic models (for example, Watts, 2001) in which the possibility of forming links arises (repeatedly) over time, and agents might ‘tremble’ when they form links (see Jackson, 2004, for references). These various equilibrium/stability concepts have different properties and are appropriate in different contexts.

With pairwise stability, or some other solution in hand, one can address a series of questions. One fundamental question is whether, from society's point of view, efficient or optimal networks will be stable when agents form links with their selfish interests in mind. Given that transfers are being considered here, one natural definition of an ‘efficient’ or ‘optimal’ network is one that maximizes the total value or the sum of utilities of all agents in the society. Another basic question is to ask whether in situations where no efficient network is pairwise stable, is it possible for some sort of intervention (for example, in the form of taxing or subsidizing links), to lead efficient networks to form.

A connections model of social networks

One stylized example from Jackson and Wolinsky (1996) gives some feeling for the issues involved in the above questions and is useful for illustrating the relationship between efficient and pairwise stable networks. Jackson and Wolinsky called this example the ‘symmetric connections model’, in which the links represent social relationships between players such as friendships. These relationships offer benefits in terms of favours, information, and so on, and also involve some costs. Moreover, players benefit from having indirect relationships. A ‘friend of a friend’ produces benefits or utility for a player, although of a lesser value than the direct benefits that come from a ‘friend’. The same is true of ‘friends of a friend of a friend’, and so forth. Benefit deteriorates in the ‘distance’ of the relationship, as represented by a factor δ between 0 and 1, which indicates the benefit from a direct relationship between two agents and is raised to higher powers for more distant relationships. For instance, in the network where player 1 is linked to 2, 2 is linked to 3, and 3 is linked to 4; player 1 gets a benefit of δ from the direct connection with player 2, an indirect benefit of δ^2 from the indirect connection with player 3, and an indirect benefit of δ^3 from the indirect connection with player 4. For $\delta<1$ this leads to a lower benefit from an indirect connection than a direct one. Players also pay some cost c for maintaining each of their direct relationships (but not for indirect ones). Once the benefit parameter, δ, and the cost parameter, $c>0$ are specified, it is possible to determine each agent's payoff from every possible network, allowing a characterization of the pairwise stable networks as well as the efficient networks. The efficient network structures are the complete network if $c<\delta-\delta^2$, a ‘star’ (a network where one agent is connected to each other agent and there are no other connections) encompassing all nodes if $\delta-\delta^2<c<\delta+\frac{(n-2)}{2}\delta^2$, and the empty network if $\delta+\frac{(n-2)}{2}\delta^2<c$. The idea is that if costs are very low it will be efficient to include all links in the network, because shortening any path leads to higher payoffs. When the link cost is at an intermediate level, then the unique efficient network structure is to have all players arranged in a star network, since such a structure has the minimal number of links (n–1) needed to connect all individuals, and yet still has all nodes within at most two links from one another. Once links become so costly that a star results in more cost than benefit, then the empty network is efficient. One can also examine a directed version of such a model, as in Bala and Goyal (2000), who find related results, but with some differences that depend on whether both agents or just one of the agents enjoys the benefits from a directed link.

Inefficiency of stable networks

The set of pairwise stable networks does not always coincide with the efficient ones, and sometimes do not even intersect with the set of efficient networks. For instance, if the cost of a link is greater than the direct benefit ($c>\delta$), then relationships are only valuable to a given agent if they generate indirect benefits as well as direct ones. In such a situation a star is not pairwise stable since the centre player gets benefit of the direct value from each of his or her links, which is less than the cost of each of those links. This model of social networks makes it obvious that there will be situations where individual incentives are not aligned with overall societal benefits.

As it will generally be the case that in economic and social networks there are some sort of externalities present, since two agents' decisions of whether or not to form a relationship can affect the well-being of other agents, one should expect that there will be situations where the networks formed through the

selfish decisions of the agents do not coincide with those that are efficient from society's perspective. In such situations, it is natural to ask whether intervention in the form of transfers among agents might help align individual and overall societal incentives to form the right network. For instance, in the connections model, it would make sense to have the peripheral agents in a star pay the centre of the star in order to maintain their links. The peripheral agents benefit much more from the relationship with the centre agent than vice versa, as the centre agent provides access to many indirect agents. Although a simple set of transfers can align individual and overall incentives in the connections model, it is impossible to always correct this tension between individual incentives and overall efficiency by taxing and subsidizing agents for the links they form (even in a complete information setting). The fact that there are very simple, natural network settings where no ‘reasonable’ set of transfers can help rectify the disparity stability and efficiency was shown in Jackson and Wolinsky (1996). Without providing details, the impossibility of reconciling stability and efficiency stems from the following considerations: from any given network, there are many other networks that can be reached. In fact, if there are n nodes, then there are $n(n-1)/2$ possible links that can be added to or deleted from any given network. In order to ensure that a given efficient network is pairwise stable, payoffs to all neighbouring networks have to be configured so that no agent finds it in his or her interest to delete a link and no two agents find it in their interests to add a link. It is impossible to assign all the necessary taxes and subsidies in such a way that (i) the transfers are feasible (and are not given to unattached agents), (ii) identical agents are treated identically, and (iii) it is always the case that at least one efficient network is pairwise stable.

Much more has been learned about the relationship between stable and efficient networks and possible transfers to ensure that efficient networks form. For instance, one can characterize some classes of settings where the efficient networks and the stable ones coincide (see Jackson and Wolinsky, 1996). One can also design transfers that ensure that some efficient network is stable by treating agents unequally (for example, taxing or subsidizing them differently even though the agents are identical in the problem as shown by Dutta and Mutuswami, 1997). Another important point was made by Currarini and Morelli (2000), who showed that if agents bargain over the division of payoffs generated by network relationships at the time when they form link, then in a nontrivial class of settings equilibrium networks are efficient. While the conclusions hinge on the structure of the link-formation-bargaining game, and in particular on an asymmetry in bargaining power across the agents, such a result tells us that it can be important to model the formation of the links of a network together with any potential bargaining over payoffs or transfers. Further study in this area shows how the types of transfers needed to reach efficient networks relate to the types of network externalities that are present in the setting.

Small worlds and strategic network formation

Beyond understanding the relationship between stable and efficient networks, strategic models of network formation have also shed light on some empirical regularities and helped predict which networks will arise in settings of particular interest. For instance, strategic models of network formation provide substantial insight into the ‘small-worlds’ properties of social networks: the simultaneous presence of high clustering (a high density of links on a local level) and short average path length between nodes (see Jackson, 2006, for references). The reasoning is based on a premise that different nodes have different distances from each other, either geographically or according to some other

characteristic, such as profession, tastes, and so on. The low cost of forming links to other nodes that are nearby then naturally explains high clustering. High benefits from forming links that bridge disparate parts of the network, due to the access and indirect connections that they bring, naturally explain low average path length.

Networks and markets

There is a rich set of studies of markets and networks from an economics perspective, including models that explicitly examine whether or not buyers and sellers have incentives to form an efficient network of relationships (for example, Kranton and Minehart, 2001). The incentives to form efficient networks depend on the setting and which agents bear the cost of forming relationships. In some settings competitive forces lead to the right configuration of links, and in others buyers and sellers over-connect in order to improve their relative bargaining positions. Other studies focus on the context of specific markets, such as labour markets, where people benefit from connections with neighbours who provide information about job opportunities (see Ioannides and Loury, 2004, for an overview and references).

In addition to studies of networks of relationships between buyers and sellers, firms also form relationships amongst themselves that affect their costs and the sets of products they offer. Such oligopoly settings where network formation is important (see Bloch, 2004, for a recent survey), again provide a rich set of results regarding the structure of networks that emerge, and contrasts between settings where efficient networks naturally emerge and others where only inefficient networks are formed.

Network formation has also been studied in the context of many other applications, including risk-sharing in developing countries, social mobility, criminal activity, international trade and banking deposits.

Finally, there have been a number of experiments on network formation, using human subjects. These examine a variety of questions, ranging from how forward-looking agents are when they form social ties, to whether or not agents overcome coordination problems when forming links, to whether there are pronounced differences between network formation when links can be formed unilaterally as opposed to when they require mutual consent, to whether efficient networks will tend to result and how that depends on symmetries or asymmetries in the efficient network structure (see Falk and Kosfeld, 2003, for some discussion and references).

See Also

- business networks
- learning and information aggregation in networks
- mathematics of networks
- power laws
- psychology of social networks
- social networks in labour markets

Bibliography

Aumann, R. and Myerson, R. 1988. Endogenous formation of links between players and coalitions: an application of the Shapley value. In *The Shapley Value*, ed. A. Roth. Cambridge: Cambridge University Press.

Bala, V. and Goyal, S. 2000. A non-cooperative model of network formation. *Econometrica* 68, 1181–230.

Barabási, A. and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509–12.

Bloch, F. 2004. Group and network formation in industrial organization: a survey. In *Group Formation in Economics; Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Bollobás, B. 2001. *Random Graphs*, 2nd edn. Cambridge: Cambridge University Press.

Boorman, S. 1975. A combinatorial optimization model for transmission of job information through contact networks. *Bell Journal of Economics* 6, 216–49.

Currarini, S. and Morelli, M. 2000. Network formation with sequential demands. *Review of Economic Design* 5, 229–50.

Dutta, B. and Mutuswami, S. 1997. Stable networks. *Journal of Economic Theory* 76, 322–44.

Erdös, P. and Rényi, A. 1960. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 17–61.

Falk, A. and Kosfeld, M. 2003. It's all about connections: evidence on network formation. Mimeo, University of Zurich.

Frank, O. and Strauss, D. 1986. Markov graphs. *Journal of the American Statistical Association* 81, 832–42.

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology* 78, 1360–80.

Ioannides, Y.M. and Loury, L.D. 2004. Job information networks, neighborhood effects and inequality. *Journal of Economic Literature* 42, 1056–93.

Jackson, M.O. 2004. A survey of models of network formation: stability and efficiency. In *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders.

Cambridge: Cambridge University Press.

Jackson, M.O. 2006. The economics of social networks. In *Chapter 1, Volume 1 in Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, ed. R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press.

Jackson, M.O. and Rogers, B.W. 2007. Meeting strangers and friends of friends: how random are socially generated networks? *American Economic Review* 97, 890–915.

Jackson, M.O. and Wolinsky, A. 1996. A strategic model of social and economic networks. *Journal of Economic Theory* 71, 44–74.

Kranton, R. and Minehart, D. 2001. A theory of buyer–seller networks. *American Economic Review* 91, 485–508.

Molloy, M. and Reed, B. 1995. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* 6, 161–79.

Montgomery, J. 1991. Social networks and labor market outcomes. *American Economic Review* 81, 1408–18.

Myerson, R. 1977. Graphs and cooperation in games. *Math. Operations Research* 2, 225–9.

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45, 167–256.

Page, F., Wooders, M. and Kamat, S. 2005. Networks and farsighted stability. *Journal of Economic Theory* 120, 257–69.

Price, D.J.S. 1965. Networks of scientific papers. *Science* 149, 510–5.

Price, D.J.S. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27, 292–306.

Simon, H. 1955. On a class of skew distribution functions. *Biometrika* 42, 425–40.

Watts, A. 2001. A dynamic model of network formation. *Games and Economic Behavior* 34, 331–41.

Watts, D.J. 1999. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton: Princeton University Press.

Watts, D.J. and Strogatz, S. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–2.

How to cite this article

Jackson, Matthew O. "network formation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000135> doi:10.1057/9780230226203.1174

The New Palgrave Dictionary of Economics Online

network goods (empirical studies)

Neil Gandal

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A network effect exists if the consumption benefits of a good or service increase with the total number of consumers who purchase compatible products. A growing empirical literature examines technological adoption of products with network effects. The early literature mainly addressed the question of whether network effects are indeed significant; this work typically employed reduced form models. Later literature employed structural methodology, which can address aspects of firm strategy, such as incentives to provide compatible products. Key issues in the empirical work on network industries are examined.

Keywords

hedonic prices; logit models of individual choice; network effect (empirical studies); product differentiation

Article

A network effect exists if the consumption benefits of a good or service increase with the total number of consumers who purchase compatible products. The literature distinguishes between direct and indirect network effects.

In the case of a direct (or physical) network effect, an increase in the number of consumers on the same network raises the consumption benefits for everyone on the network. Communication networks such as telephone and e-mail networks are examples of goods with direct network effects.

A network effect can also arise in a setting with a ‘hardware/software’ system. Here, the benefits of the hardware good increase when the variety of compatible software increases. An indirect (or virtual) network effect arises endogenously in this case because an increase in the number of users of compatible hardware increases the demand for compatible software. Since software goods are typically characterized by economies of scale, the increase in demand leads to increases in the supply of software varieties. Examples of settings where virtual network effects arise include consumer electronics such as

CD players and compact discs, computer operating systems and applications programs, and television sets and programming.

Given the dramatic growth of the internet and information technology industries, and the importance of interconnection in these networks, it is not surprising that there is a large theoretical literature on competition in industries with network goods. Important questions in this literature include

- the examination of the private and social incentives to attain compatibility;
- the trade-off between standardization and variety;
- modelling the dynamics of competition between competing networks; and
- how the private and social choice among competing incompatible networks differs when there are both early and late adopters.

See Farrell and Klemperer (2007) for further discussion.

Although relatively small, a growing empirical literature has developed to examine technological adoption of products with network effects. In this short article, I briefly discuss this literature. The empirical work can be organized by the issues addressed and the methodology employed. The primary issue addressed by the early literature is whether network effects are indeed significant; this work typically employed reduced form models. The article first surveys early work in this genre, then examines papers that employed structural methodology. The main advantage of this methodology is that it can address aspects of firm strategy, such as incentives to provide compatible products. The article closes by examining key issues in empirical work on network industries.

Early work: indirect evidence of network effects

Greenstein (1993), Gandal (1994; 1995), and Saloner and Shepard (1995) provide early evidence that the value of the ‘hardware’ good depends on the variety of compatible complementary software. (Shy, 2001, surveys many of the empirical papers discussed in this article in greater detail than space permits here.)

Software for the IBM 1400 mainframe could not run on succeeding generations of IBM mainframes while software for the IBM 360 could run on succeeding models. Greenstein (1993) finds that, other things being equal, a firm with an IBM 1400 was no more likely than any other firm to purchase an IBM mainframe when making a future purchase. On the other hand, a firm with an IBM 360 was more likely to purchase an IBM mainframe than a firm that did not own an IBM 360. This result can be interpreted as a demand for compatible software.

Gandal (1994) estimates hedonic (quality-adjusted) price equations for spreadsheets to examine whether spreadsheet programs that were compatible with Lotus – the de facto standard – command a premium. The results – that consumers place a positive value on compatibility – suggest (*a*) direct network effects because people want to share files and (*b*) indirect network effects because compatible software enables the transfer of data among a variety of software programs. Gandal (1995) extends the analysis to database management software (DMS) and multiple standards and finds that only the Lotus file compatibility standard is significant in explaining price variations, suggesting that indirect network effects are important in the DMS market.

Saloner and Shepard (1995) test for network effects in the automated teller machine (ATM) industry. In particular, they test whether banks with a larger expected number of ATM locations will adopt the ATM technology sooner. Since expected network size is not an observable variable, they use the number of branches as a proxy. The results suggest that banks with more branches will adopt earlier, which is consistent with virtual network effects.

Structural models: explicitly modelling the complementary goods market

Because hedonic price equations are a reduced form, rather than a structural model, parameter estimates associated with compatibility in Gandal (1994; 1995) may be capturing demand effects or supply effects or some combination of both. In other words, are consumers really willing to pay a premium for compatibility or is the marginal cost of compatibility relatively high? In the case of software, fixed costs of providing characteristics are quite significant, while marginal production costs associated with the characteristics are typically very small; they primarily include duplication of digital material. Hence, in these papers the estimated hedonic price coefficients on compatibility indeed measure consumer willingness to pay for compatibility.

Nevertheless, reduced form models are not suitable for examining business strategies or conducting counterfactuals. Gandal, Kende and Rob (2000) develop a dynamic structural model of consumer adoption and software entry, and use the model to estimate the feedback from hardware to software and vice versa in the CD industry. The advantage of the structural methodology is that it enables researchers to assess business strategies as well as examine conduct counterfactuals. In the case of business strategies, Gandal, Kende and Rob (2000) show that a five per cent reduction in price would have had the same effect as a ten per cent increase in CD variety in terms of increasing sales of CD players. They also show that, if it had been possible to make CD players compatible with LPs, compatibility could have accelerated the adoption process by more than a year. This is just a ‘thought experiment’ for CD players, but it has policy relevance for other systems like HDTV.

Rysman (2004) develops a structural model to examine the importance of network effects in the market for Yellow Pages. The model includes a consumer adoption equation, advertiser demand for space, and a firm's profit maximizing behaviour. He finds that consumers value advertising and advertisers value consumer adoption, suggesting virtual network effects.

In several recent papers, advances in the estimation of discrete choice models of product differentiation – see Berry (1994) and Berry, Levinsohn and Pakes (1995) – have also been employed when testing for indirect network effects in differentiated product markets. Ohashi and Clements (2005), for example, use a logit model to test for indirect network effects in the US video game market.

Key issues in empirical work

As in most fields, empirical work is typically limited by the available data. A key problem exists when one tries to estimate network effects in homogeneous product industries using time series data. For many network industries, technological progress drives down prices and costs. Hence an increase in the number of users on a network might be due to a network effect or to falling prices (see Gowrisankaran

and Stavins, 2004, for further discussion). In order to estimate these effects, one must have additional data.

Gandal, Kende and Rob (2000), for example, have data on the number of available compact disc titles at each point in time. Hence, in their model the two main effects that lead to greater adoption of CD players – lower prices of the hardware good and network effects due to increases in the number of titles – are measured separately. Nevertheless, that is only a start, since both of these variables are typically endogenous. Identification in Gandal, Kende, and Rob (2000) was possible only because there were data on the fixed costs of entering the CD production industry over time. These data were used as an instrument for CD (title) availability. Additionally, case studies indicated that the CD player industry was quite competitive, leading the authors to assume that the price of CD players was exogenous. Without both of these assumptions, it would not have been possible to identify the model.

Additionally, there is the thorny issue of pricing in dynamic models of competition in network industries. Since hardware firms may want to subsidize early adopters in order to build up a network advantage and then (perhaps) charge a higher price when the installed base grows, pricing issues are dynamic; firms will take into account (current and expected future) network size when choosing their prices. Park (2004) develops a dynamic structural model of competition in an oligopolistic market with network effects that addresses the dynamic pricing issues; he then estimates the model for VCRs. To the best of my knowledge, this is the only empirical paper that deals explicitly with dynamic pricing issues.

A similar issue arises in dynamic models of competition in network industries when firms make investment in quality over time. Markovich (2001) examines the trade-off between standardization and variety in a dynamic setting using numerical methods. With suitable data one might be able to use her framework to empirically examine investment incentives and pricing decisions in a dynamic setting with network effects.

Finally, there is a budding empirical literature on standardization via committees. Papers include Simcoe (2006), who examines the standardization process in various committees of the Internet Engineering Task Force, and Gandal, Gantman and Genesove (2006), who examine firms' incentives to participate in Telecommunication Industry Association standardization meetings.

See Also

- hedonic prices
- network goods (theory)

Bibliography

Berry, S. 1994. Estimating discrete-choice models of product differentiation. *RAND Journal of Economics* 25, 334–47.

Berry, S., Levinsohn, J. and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–90.

Farrell, J. and Klemperer, P. 2007. Coordination and lock-in: competition with switching costs and network effects. In *Handbook of Industrial Organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: North-Holland.

Gandal, N. 1994. Hedonic price indexes for spreadsheets and an empirical test for network externalities. *RAND Journal of Economics* 25, 160–70.

Gandal, N. 1995. A Selective survey of the literature on indirect network externalities. *Research in Law and Economics* 17, 23–31.

Gandal, N., Gantman, N. and Genesove, D. 2006. Intellectual property and standardization committee participation in the U.S. modem industry. In *Standards and Public Policy*, ed. S. Greenstein and V. Stango. Cambridge: Cambridge University Press.

Gandal, N., Kende, M. and Rob, R. 2000. The dynamics of technological adoption in hardware/software systems: the case of compact disc players. *RAND Journal of Economics* 31, 43–61.

Gowrisankaran, G. and Stavins, J. 2004. Network externalities and technology adoption: lessons from electronic payments. *RAND Journal of Economics* 35, 260–76.

Greenstein, S. 1993. Did installed base give an incumbent any (measurable) advantages in federal computer procurement? *RAND Journal of Economics* 24, 19–39.

Markovich, S. 2001. Snowball: the evolution of dynamic markets with network externalities. Mimeo. Tel Aviv University.

Ohashi, H. and Clements, M. 2005. Indirect network effects and the product cycle: video games in the U. S., 1994–2002. *Journal of Industrial Economics* 53.

Park, S. 2004. Quantitative analysis of network externalities in competing technologies: the VCR case. *Review of Economics and Statistics* 86, 937–45.

Rysman, M. 2004. competition between networks: a study of the market for Yellow Pages. *Review of Economic Studies* 71, 483–512.

Saloner, G. and Shepard, A. 1995. Adoption of technologies with network externalities: an empirical examination of the adoption of automated teller machines. *RAND Journal of Economics* 26, 479–501.

Shy, O. 2001. *The Economics of Network Industries*. Cambridge: Cambridge University Press.

Simcoe, T. 2006. Committees and the creation of technical standards. In *Standards and Public Policy*,

ed. S. Greenstein and V. Stango. Cambridge: Cambridge University Press.

How to cite this article

Gandal, Neil. "network goods (empirical studies)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000140> doi:10.1057/9780230226203.1175

The New Palgrave Dictionary of Economics Online

network goods (theory)

Paul Klemperer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Network effects arise where current users of a good gain when additional users adopt it (classic examples are telephones and faxes). The effects create multiple equilibria and fierce competition between incompatible networks; users' expectations are crucial in determining which network succeeds. Early choices, such as the QWERTY typewriter keyboard, lock in the market; new entry, especially against established networks with proprietary technology, is often nearly impossible. Incompatible networks can induce efficient ‘competition for the market’, but more often create biases and inefficiencies. Policymakers should scrutinize markets where firms deliberately choose incompatibility.

Keywords

compatible products; competition for the market; competition policy; coordination; entry; excess early power; excess inertia; excess momentum; herding; indirect network effects; intellectual property; lock-in; market share; Microsoft; multiple equilibria; network effects; network externality; penetration pricing; pre-announcements; product variety; proprietary technology; QWERTY; standards; switching costs; tipping

Article

Direct network effects arise if each user's payoff from the adoption of a good, and his incentive to adopt it, increase as more others adopt it; that is, if adoption by different users is complementary. For example, telecommunications users gain directly from more widespread adoption, and telecommunications networks with more users are also more attractive to non-users contemplating adoption.
Indirect network effects arise if adoption is complementary because of its effect on a related market. For example, users of hardware may gain when other users join them, not because of any direct benefit, but because it encourages the provision of more and better software.
Extensive case studies and more formal econometric evidence document significant network effects in

many areas including, for example, telecommunications, radio and television, computer hardware and software, applications software and operating systems (including Microsoft's), securities markets and exchanges (including Ebay), and credit cards (see, for example, Gabel, 1991; Rohlfs, 2001; Shy, 2001; and the article on network goods (empirical studies) in this dictionary).

Usually adoption prices do not fully internalize the network effects, so there is a positive externality from adoption. A single network product therefore tends to be under-adopted at the margin – this issue was the main focus of the early literature (see, for example, Leibenstein, 1950; Rohlfs, 1974). However, if two networks compete, then adopting one network means not adopting the other, which dilutes or reverses the externality.

More interestingly – and what is the starting point for the more recent literature – network effects create incentives to ‘herd’ with others. In a static (simultaneous-adoption) game there are often multiple equilibria, so expectations are crucial, and self-fulfilling. Likewise, a dynamic (sequential-adoption) game exhibits positive feedback or ‘tipping’ – a network that looks like succeeding will *as a result* do so (see, for example, David, 1985; Arthur, 1989; Arthur and Rusczcynski, 1992).

How well competition among incompatible networks works depends dramatically on how adopters form expectations and coordinate their choices. If adopters smoothly coordinate on the best deals, vendors face strong pressure to offer them. Competition may then be unusually fierce because all-or-nothing competition neutralizes horizontal differentiation – since adopters focus not on matching a product to their own tastes but on joining the expected winner.

However, coordination is not easy. With simultaneous adoption, adopters may fail to coordinate at all and ‘splinter’ among different networks, or may coordinate on a different equilibrium from the one that is best for them – for example, each adopter may expect others to choose a low-quality product because it is produced by a firm that was successful in the past. Furthermore, consensus standard-setting (informally or through standards organizations) can be painfully slow when different adopters prefer different coordinated outcomes (see Bulow and Klemperer, 1999). Coordination through contingent contracts is possible in theory (see, for example, Dybvig and Spatt, 1983; Segal, 1999), but seems uncommon in practice.

When adoption is sequential, we see *early instability and later lock-in* (see, for example, Arthur, 1989) – this corresponds to the multiple equilibria that arise with simultaneous adoption. Because early adoptions influence later ones, long-term behaviour is determined largely by early events, whether accidental or strategic. In theory, at least, fully sequential adoption achieves the efficient outcome if it is best for all adopters, but more generally early adopters' preferences count for more than later adopters': this is ‘excess early power’. Note that ‘excess early power’ does not depend on ‘excess inertia’, that is, on incompatible transitions being too hard *given ex post* incompatibility. (Both ‘excess inertia’, and its opposite, ‘excess momentum’, are theoretically possible; see Farrell and Saloner, 1985.)

Firms promoting incompatible networks compete to win the pivotal early adopters, and so achieve *ex post* dominance and monopoly rents. Strategies such as penetration pricing and pre-announcements (see, for example, Farrell and Saloner, 1986) are common. History, and especially market share, matter because an installed base both directly means a firm offers more network benefits and boosts expectations about its future sales. Such ‘Schumpeterian’ competition ‘for the market’ can neutralize (or even overturn) excess early power if promoters of networks that will be more efficient later on set low penetration prices in anticipation of this (see Katz and Shapiro, 1986a). More commonly, though, late

developers struggle while networks that are preferred by early pivotal customers thrive.

So early preferences and early information are likely to be excessively important in determining long-term outcomes. For example, whether or not the Dvorak typewriter keyboard is really much better than QWERTY (as David, 1985, contends), there clearly was a chance in the 1800s that a keyboard superior to QWERTY would later be developed, and it is not clear what could have persuaded early generations of typists to wait, or to adopt diverse keyboards, *if* that was socially desirable. So it seems unlikely that the market gave a very good test of whether or not waiting was efficient. (Liebowitz and Margolis, 1990, and Liebowitz, 2002, contest both the details of the QWERTY example and the claim that network effects are significant more generally, but at least the second view is probably a minority one.)

Despite the possibility of competition for the market passing *ex post* rents through to earlier buyers, incompatibility often reduces efficiency and harms consumers in several ways.

Incompatibility means that consumers are faced with either a segmented market with low network benefits, or – if the market does ‘tip’ all the way to one network – with reduced product variety and without the option value from the possibility that a currently inferior technology might later become superior. Product variety is more sustainable if niche products are compatible with the mainstream, and so don't force users to sacrifice network effects.

These direct costs of poor coordination by adopters may be exacerbated by weaker incentives for vendors to offer good deals. For example, if a firm like Microsoft is widely believed to have the ability to offer the highest quality, it may never bother to do so: the fact that everyone expects Microsoft to recapture the market if it ever lost any one cohort of customers (or lost any one cohort of providers of complementary products) means everyone rationally chooses Microsoft even if it never actually produces high quality or offers a low price (see Katz and Shapiro, 1992).

Ex post rents are often not fully dissipated by *ex ante* competition, especially if expectations fail to track relative surplus. Worse, the rent dissipation that does occur may be wasteful, such as socially inefficient marketing. At best, *ex ante* competition induces ‘bargain-then-rip-off’ pricing (low to attract business, high to extract surplus) but this distorts buyers' quantity choices and gives them artificial incentives to be or appear pivotal.

Furthermore, outcomes are biased in favour of a proprietary technology (for example, Microsoft's) whose single owner has the incentive to market it strategically over ‘open’ unsponsored alternatives (for example, Linux) – see, for example, Katz and Shapiro (1986b). As discussed above, outcomes are also often biased in favour of networks that are more efficient early on, and are generally biased in favour of established firms on whom expectations focus. The last bias implies entry with proprietary network effects is often nearly impossible (and frequently much too hard from the social viewpoint even *given* incompatibility). And this in turn makes it easier to recoup profits after predatory behaviour that eliminates a rival, and so encourages such predation.

So while incompatibility does not necessarily damage competition, it often does, and firms may therefore also dissipate further resources creating and defending incompatibility.

If firms offer compatible products, then consumers don't need to buy from the same firm to enjoy full network benefits, and (differentiated) products will be better matched with customers. Consumers will be willing to pay more for these benefits, and this may encourage firms to choose compatibility. But compatibility often intensifies competition and nullifies the competitive advantage of a large installed base, whereas proprietary networks tend to make competition all-or-nothing, with the advantage going to large firms, and may completely shut out weaker firms. So large firms and those who are good at

steering adopters' expectations may prefer their products to be incompatible with rivals' (see, for example, Katz and Shapiro, 1985; Bresnahan, 2001), and may be able to use their intellectual property to enforce this.

Competition with incompatible network effects is closely related to other forms of competition when market share is important, especially competition when consumers have switching costs (see, for example, Klemperer, 1995; Farrell and Klemperer, 2007; and the companion-piece to this article, switching costs), and has similar broader implications (for example, for international trade, see Froot and Klemperer, 1989).

Because competition ‘for the market’ differs greatly from conventional competition ‘in the market’, and especially because capturing consumers' and complementors' expectations can be so profitable, competition policy needs to be vigilant against predatory or exclusionary tactics by advantaged firms, including deliberately creating incompatibility by misusing intellectual property protection. Thus, for example, the network effect by which more popular operating systems attract more applications software took centre stage in both the US and European Microsoft cases (see, for example, Bresnahan, 2001).

And because coordination is often important and difficult, institutions such as standards organizations matter, and government procurement policy takes on more significance than usual.

In summary, network effects *can* involve efficient competition for larger units of business – ‘competition for the market’ – but very often make competition, especially entry, less effective. So I, and others, recommend that public policymakers should have a cautious presumption in favour of compatibility, and should look particularly carefully at markets where incompatibility is strategically chosen rather than inevitable.

Farrell and Klemperer (2007) contains a recent and comprehensive survey of network effects.

See Also

- network goods (empirical studies)
- switching costs

The views expressed here are personal and should not be attributed to the UK Competition Commission or to any of its individual Members other than myself. Furthermore, although some observers thought some of the behaviour discussed warranted regulatory investigation, I do not intend to suggest that any of it violates any applicable rules or laws.

Bibliography

Arthur, W.B. 1989. Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal* 99, 116–31.

Arthur, W.B. and Rusczcynski, A. 1992. Dynamic equilibria in markets with a conformity effect. *Archives of Control Sciences* 37, 7–31.

Bresnahan, T. 2001. Network effects in the Microsoft case. Discussion Paper No. 00-51, Stanford Institute for Economic Policy Research, Stanford University.

Bulow, J. and Klemperer, P.D. 1999. The generalized war of attrition. *American Economic Review* 89, 175–89.

David, P. 1985. Clio and the economics of QWERTY. *American Economic Review* 75, 332–7.

Dybvig, P.H. and Spatt, C.S. 1983. Adoption externalities as public goods. *Journal of Public Economics* 20, 231–47.

Farrell, J. and Klemperer, P.D. 2007. Coordination and lock-in: competition with switching costs and network effects. In *Handbook of Industrial Organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: North-Holland.

Farrell, J. and Saloner, G. 1985. Standardization, compatibility and innovation. *RAND Journal of Economics* 16, 70–83.

Farrell, J. and Saloner, G. 1986. Installed base and compatibility: innovation, product preannouncements, and predation. *American Economic Review* 76, 940–55.

Froot, K.A. and Klemperer, P.D. 1989. Exchange rate pass-through when market share matters. *American Economic Review* 79, 637–54.

Gabel, H.L. 1991. *Competitive Strategies for Product Standards*. New York: McGraw-Hill.

Katz, M.L. and Shapiro, C. 1985. Network externalities, competition and compatibility. *American Economic Review* 75, 424–40.

Katz, M.L. and Shapiro, C. 1986a. Product compatibility choice in a market with technological progress. *Oxford Economic Papers* 38, 146–65.

Katz, M.L. and Shapiro, C. 1986b. Technology adoption in the presence of network externalities. *Journal of Political Economy* 94, 822–41.

Katz, M.L. and Shapiro, C. 1992. Product introduction with network externalities. *Journal of Industrial Economics* 40, 55–83.

Klemperer, P.D. 1995. Competition when consumers have switching costs. *Review of Economic Studies* 62, 515–39.

Leibenstein, H. 1950. Bandwagon, Snob and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64, 183–207.

Liebowitz, S. 2002. *Re-Thinking the Network Economy: The True Forces that Drive the Digital Marketplace*. New York: American Management Association.

Liebowitz, S.J. and Margolis, S.E. 1990. The fable of the keys. *Journal of Law and Economics* 33, 1–25.

Rohlfs, J. 1974. A theory of interdependent demand for a communications service. *Bell Journal of Economics* 5, 16–37.

Rohlfs, J. 2001. *Bandwagon Effects in High Technology Industries*. Cambridge, MA: MIT Press.

Segal, I. 1999. Contracting with externalities. *Quarterly Journal of Economics* 114, 337–88.

Shy, O. 2001. *The Economics of Network Industries*. Cambridge: Cambridge University Press.

How to cite this article

Klemperer, Paul. "network goods (theory)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000138> doi:10.1057/9780230226203.1176

The New Palgrave Dictionary of Economics Online

neuroeconomics

John Dickhaut and Aldo Rustichini
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Neuroeconomics aims at improving the science of major economic phenomena such as the formation of prices and the design and performance of institutions. A revised model of choice is expected, based on the behaviour of the neuronal structures of the brain. Researchers are tackling issues such as determining how fundamental constructs like probabilities and payoffs are reflected in neuronal activity; disentangling the processing of inputs to choice from the act of choice; isolating learning, impulsive and analytic components of neuronal behaviour; and distinguishing how context affects the processing of the brain and subsequent levels of trust and cooperation in exchange.

Keywords

Allais paradox; choice; Ellsberg paradox; experimental economics; learning; mixed strategy equilibrium; neuroeconomics; preference reversals; Prisoner's Dilemma; probability; regret; reputation; trust; ultimatum game

Article

The fundamental unit of activity of the brain is the neuron. It ingests nutrients, receives chemical signals from other neurons, and fires (produces electro-chemical action potentials), which results in sending chemical signals (that is, neurotransmitters) to other neurons. Human brains are estimated to have as many as 100 billion neurons. A first task of neuroeconomics is to accumulate information about the behaviour of collections of neurons and how they interact to produce economic choices.

Research methods

Research methods employed include single neuron recordings of non-human primates, often macaque monkeys, brain scans (such as functional magnetic resonance imaging, fMRI) of humans and comparative studies of lesioned and normal patients.

Single cell recording

Only in rare instances is it possible to target specific neurons of living human beings (for example, when someone is having open brain surgery). Because many brain structures of non-humans correspond to human brain structures, it is possible to use results from non-human studies to postulate neuronal structures that function in human brains making economic choices. The method for making observations of a neuron's behaviour using monkeys is single cell recording. In this approach specific groups of neurons are targeted. Electrodes are implanted in individual neurons in the group. When a neuron fires, an electrical impulse is sent to a recording device.
Figure 1 shows a typical result for a specific neuron in a targeted group of neurons. The distance along the horizontal axis represents the number of seconds into the experimental trial. In this picture an experimental event such as the receipt of reward occurred roughly one fifth of the way through the experimental trial. The vertical axis represents the sum of activations for this neuron at each particular time over a set of experimental trials; here there is much activation immediately after the experimental event when looking across trials.

Figure 1

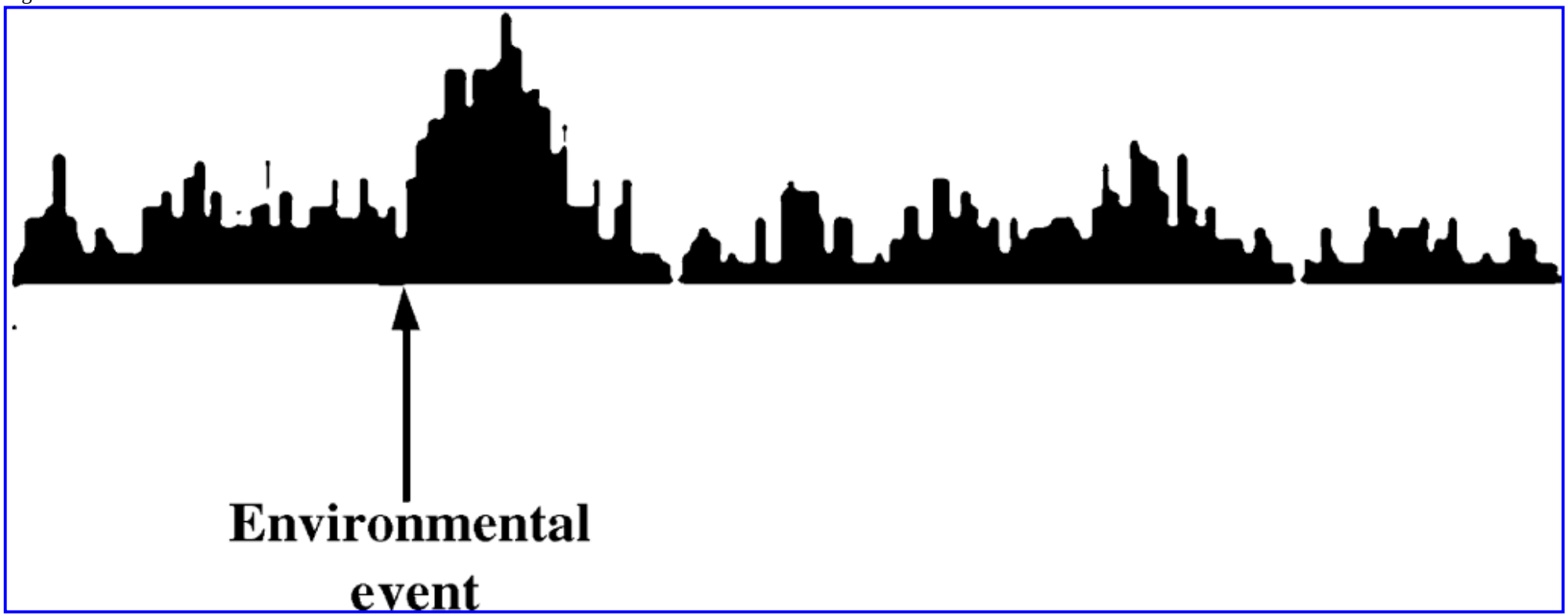

Imaging

In studying the human brain researchers employ scanning, for example, functional magnetic resonance imaging (fMRI). fMRI surrounds the economic agent with a strong magnetic field. When specific neurons are engaged in a task, capillaries near those neurons carry more oxygenated blood than capillaries surrounding neurons not engaged in the task. fMRI assesses where such oxygenated blood is. These assessments can be represented in an image indicating areas of the brain that activate differentially. A typical scan produces an image like that in Figure 2. The image shows the implicit activation in the superior parietal lobe (upper-left darkened spot of image) when a subject performs certain numerical operations. The whitened area surrounding the darkened spot suggests the increasing activation around the location.

Figure 2

Source: Dehaene et al. (2003).

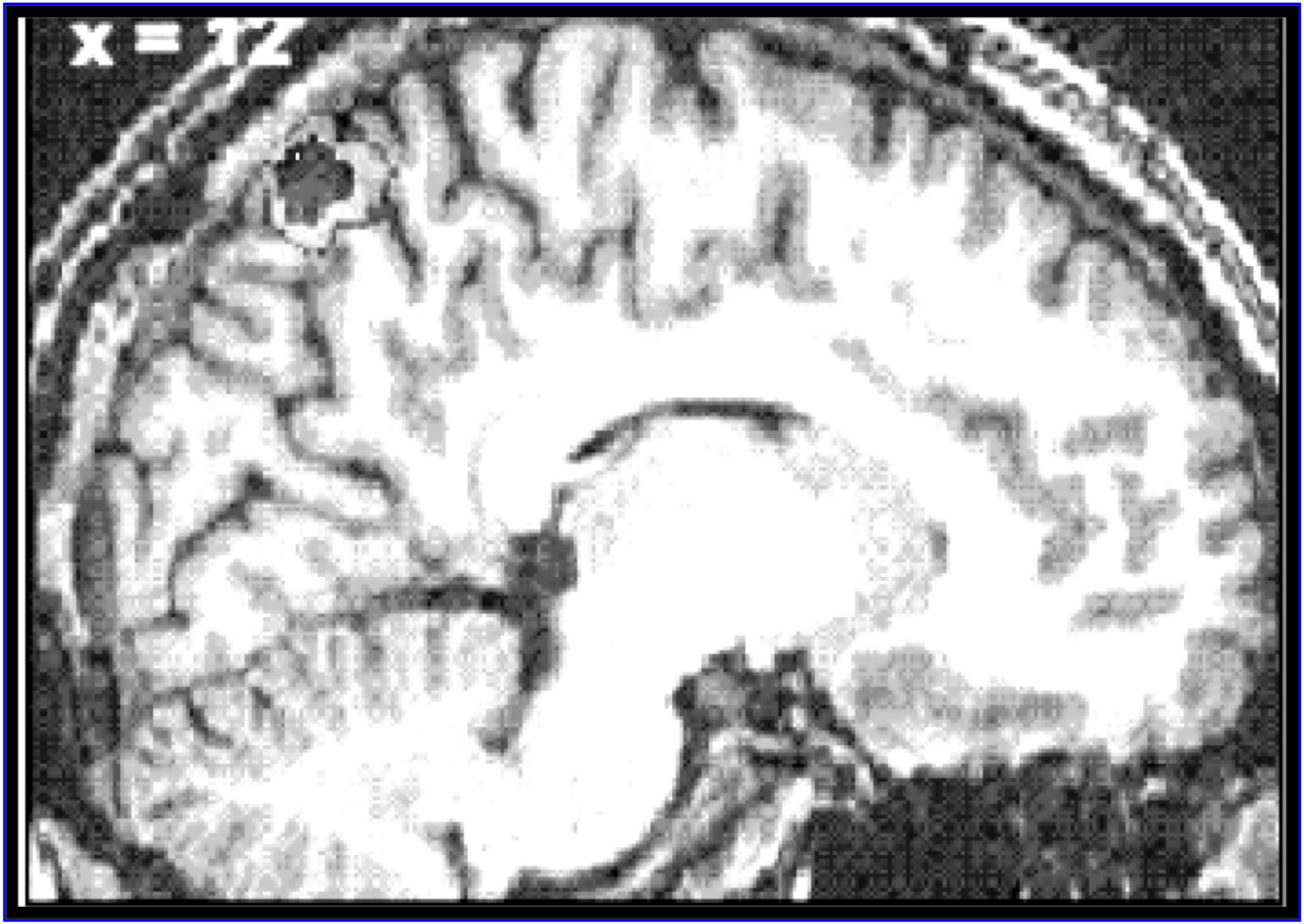

An fMRI captures brain activity at a much coarser level than single unit recording; it cannot isolate some brain structures in humans to the same degree as single unit recording can isolate neuronal activation in monkeys. fMRI allows investigators time resolution in milliseconds.

A related type of scanning is positron emission tomography (PET). In PET studies subjects are injected with radioactive isotopes. Activated neurons in the brain recruit more blood than other neurons and thus brain areas with more positron emissions indicate where more blood is flowing. These areas are then highlighted to produce an image similar to that in

Figure 2.

Using lesioned and normal subjects

Another type of study involves using lesioned (subjects with damaged brain areas) and normal subjects. When normal subjects perform differently on tasks from lesioned subjects, it is evidence consistent with the hypothesis that the area in question is responsible for the differential performance.

Skin conductance

Skin conductance (SCR) measures the ability of skin to conduct electricity (conductance increases with sweat secretion). Generally measures such as SCR and heart rate (HR) have been used to proxy behaviour in the emotional part of the brain. Brain structures associated with emotion send signals to both the heart and the sweat glands.
Figure 3 is intended to assist the reader in identifying brain areas mentioned in the discussion. The image depicts a cross-section (a sagittal view) of the brain taken at the midline of the brain. Approximate locations of brain structures are provided. Where the word 'To' appears in the figure it means the brain part is behind the cross-section at that location.
Figure 3

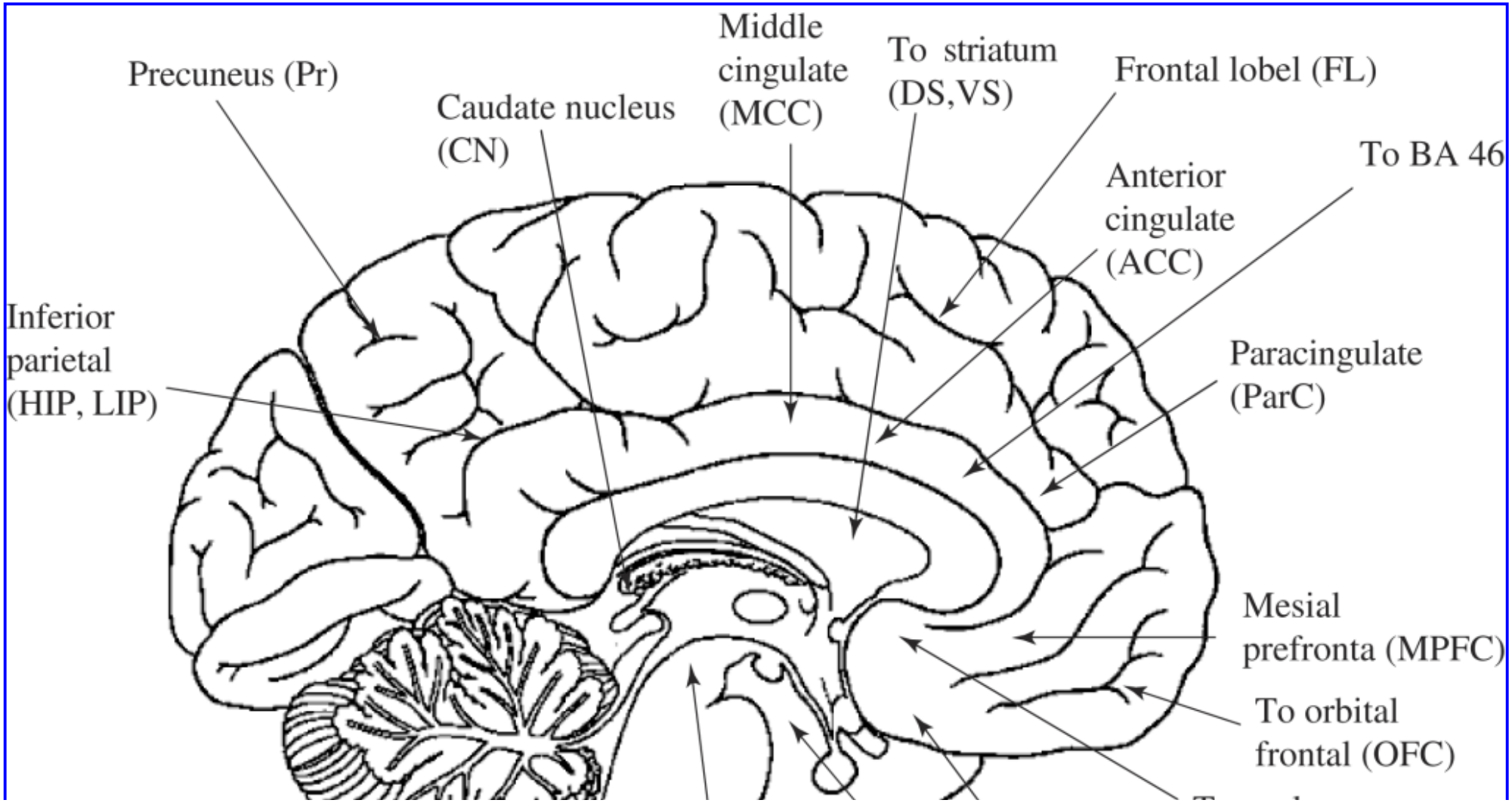

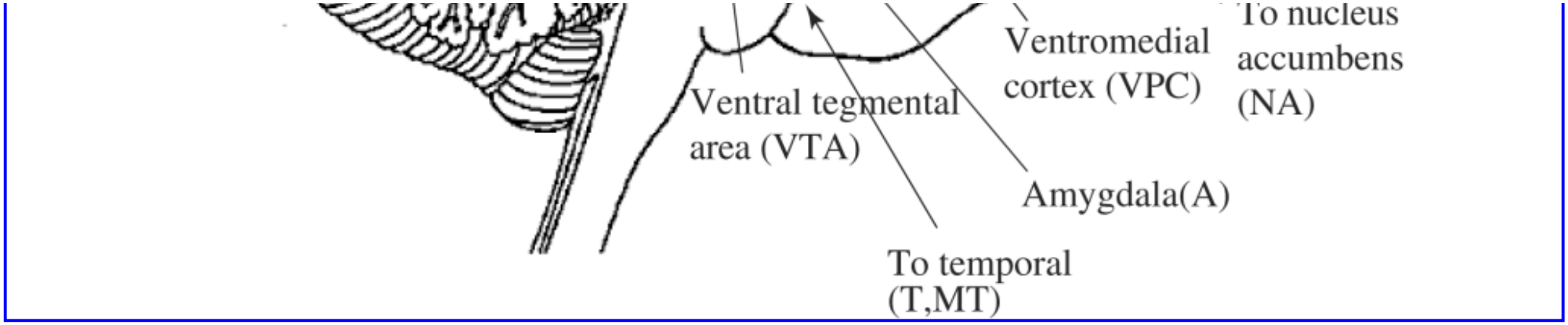

A critical question about such research concerns what we have learned so far about the economic behaviours of humans (in relation to monkeys) using these methods. The remainder of this article suggests several answers to this question.

Results related to games against nature

1. The monkey brain has mechanisms that are sensitive to environmental differences in probabilities (relative frequencies) and payoffs. Typically, neuroeconomists with neuroscience backgrounds use a reinforcement perspective. For example, no representation of a probabilistic process is made. Rather, a subject learns probabilities through repeated exposure to outcome feedback. One important set of findings using this paradigm reveals a collection of neurons responsible for detecting differences in economic information in the environment. Tremblay and Schultz (1999) used single cell recording to demonstrate that a region of the brain, the ventromedial prefrontal cortex (VPC), has some very specialized neurons. These VPC neurons are differentially activated for different reward expectations in macaque monkeys. The experimenters established that monkeys reveal a preference for different food and liquid items. For example, they were able to establish that a raisin was stochastically preferred to a piece of apple, which was stochastically preferred to cereal. Then they established that, when the raisin and pieces of apple were alternated as rewards, the VPC neurons activated more for raisins than for apples; on the other hand when apples and cereal were the rewards, the same neurons activated more for the pieces of apple.
Fiorillo, Tobler and Schultz (2003) showed monkeys were differentially sensitive to differences in probabilities of stimuli. The researchers employed five different visual cues, each of which yielded a reward with different probabilities, 0, .25, .5, .75 and 1.00. Neurons in the ventral tegmental area (VTA) showed higher activation immediately after cues the more likely the cue was to yield a reward. At the actual time of reward the same neurons activated more the less likely it was that the reward would follow.
2. *The findings regarding how monkeys come to know probabilities and payoffs have implications for how humans come to know probabilities and payoffs*. Brain areas such as the VTA are so small that it is not easy to detect them in humans using fMRI. Knutson et al. (2003) exploited neuroanatomy to show that VTA neurons send neural information to the nucleus accumbens (NA) and mesial prefrontal cortex (MPFC) The results from using fMRI indicate that the NA is sensitive to differential gains and that the MPFC encodes differences in probabilities. Thus, Knutson et al., without directly assessing the behaviour of human VTA neurons, were able to look downstream to infer an informational role for these neurons.
3. *Researchers have begun to incorporate results in experiments with feedback into a testable dynamic theory of choice*. The diagnostic role of VTA neurons in relating expectation to outcome serves as a basis for a particular dynamic model of choice, the actor–critic model (Schultz, Dayan and Montague, 1997). In the model the critic assesses the difference between expectation and outcome, the difference forms the basis for evaluating the stimuli in the experiment and for revising the probability for the next choice. Berns et al. (2001) showed that parts of this model are appropriate to human behaviour when they looked specifically at how predictable sequences of squirts of water and juice activate brains of human subjects as compared with unpredictable ones. Areas more activated for unpredicted areas than predicted areas included the NA and the orbital frontal cortex (OFC), clusters of neurons also downstream from the VTA. O'Doherty et al. (2004) pinpointed differential activation associated with the actor, dorsal striatum (DS), and the critic, ventral striatum (VS).
4. *Emotions can play a beneficial role in choice*. Bechara and Damasio (2005) invented the Iowa gambling task (IGT) to assess the role of emotions in choice. In earlier studies, emotions had been shown to be associated with activation in the OFC and the amygdala (A). In the IGT subjects sampled 100 times from four decks of cards and subjects received the reward that showed up on the face of the card drawn. Two of the decks were bad decks, resulting in occasional high losses as well as a low long-run payoff. Two were good decks, which produced moderate gains and an occasional moderate loss, but yielded long-run gains. To show that emotions aided choice, Bechara and Damasio report using three sets of subjects – subjects with damage to the VPC area of the brain, subjects with damage to the A, and normal subjects. None of the subjects knew the composition of the decks, but as they performed the task they received feedback; hence the potential for learning the composition of the decks. Neuronal firing was implicitly detected using skin conductance and heart

rate (SCR, HR).

Normal and VPC damaged subjects showed SCR and HR increases when the card was observed, but A damaged subjects showed no response. Furthermore, while learning the task normal subjects developed ‘anticipatory’ SCRs, that is, their SCR measurement increased as their hand neared the choice of a bad deck even though supplemental evidence showed no awareness of the bad deck. This anticipatory response was not detected in either of the groups with brain damage. The final piece of evidence corroborating that emotions play a positive role in choice is that subjects with brain damage made poorer choices in the task.

5. *A decision itself consists of more than just a choice. There is a neuronal modification of sensory inputs, a choice, and various neuronal communications to muscular structures that reveal the choice*. Shadlen and Newsome (2001) used a task in which a monkey sees moving dots presented on a screen. A portion of the dots had direction determined randomly and a portion had a fixed direction right or left. The monkey's choice involved making an eye movement, a saccade, to the right or left signifying the net direction of movements in the dots. The monkey was rewarded if correct.

Suppose there is a small net movement of dots to the right. When the monkey first sees the dots, they are registered on the retinas of the monkey's eyes. These signals are transferred through the optic chasm back to the occipital lobe and then to secondary areas of the visual cortex (MT). This processing takes place encoding and partially preserving various aspects of the stimuli, including colour, size and background, but most importantly the direction of movement of the dots on the screen. MT enervates (sends signals to) the lateral interior parietal cortex (LIP); however the LIP does not just preserve the signals in MT but summarizes the net activation between groups of neurons in the MT, in particular the difference in activation in neurons representing movement of dots from right to left. The LIP then sends signals which direct the muscle movements of the eye.

Such a structure seems somewhat removed from probability and value as they might be expected to be seen in economic choice. Platt and Glimcher (1999) provided the work that helps make the linkage clear. Using single unit recording, they placed electrodes in the LIP. A monkey indicated choices by making eye movements to the left or right. When appropriate a movement to the left yielded a juice squirt of .01•ml while a movement to the right yielded .03•ml. The monkey was signalled the appropriate direction of eye movement by different coloured fixation points in the middle of the monkey's computer screen. The fixation point signalled left and right with .5 probability. This set-up allowed the investigators to compute the expected payoff at different levels of information (before and after showing the fixation point) to the monkeys. Results revealed a collection of neurons in the LIP that responded monotonically to increases in expected payoff. Thus, in a task with computable expected payoffs, the LIP registers how differences in expected payoff enter into the decision process.

6. *The implicit processes of traditional choice theory tend to be evoked when subjects deal with numerical representations of outcomes*. In results 1–5 a reinforcement paradigm is involved, and many findings are the result of repeated trials with subjects bringing no knowledge of the stimuli to the task. For example, in Fiorillo, Tobler and Schultz (2003), monkeys experienced one signal at a time and seconds later learned whether a reward occurred. On the other hand economic theory often assumes there can be a structured and often numerical representation of the choice problem. In experiments conducted by economists, physical objects such as dice, urns filled with different-coloured marbles, and wheels of fortune with different-coloured segments have been used to convey probabilistic information. At times subjects have been simply told numbers that represent probabilities that the experimenters would like them to believe were the true probabilities.

Furthermore, because decision theory describes the relationship between choices that are available to the decision maker given no changes in subject's endowments, studies done by experimental economists have often provided no feedback after every choice; but rather, a randomly selected choice is played only after a set of choices have been made. In this sense experimental economics has traditionally been concerned with choice, while experiments conducted by neuroscientists are often concerned with learning. Such traditional types of experimental economics studies have unearthed a large number of regularities including the Allais and Ellsberg paradoxes and preference reversals, and in some studies expected utility is supported.

Dickhaut et al. (2003) had subjects make binary choices between gambles. For example, the subject could choose between a certainty gamble and a risky gamble (or two risky gambles). Probabilities were represented to subjects as the number of balls of particular colours that could be drawn from an urn, and after a set of choices was made one or more of the subject's designated choices was played. In the study the balls were drawn from a real urn. The study showed that context plays a role in how the brain functions during choice. For risky gambles comparison brain areas such as the frontal lobe (FL) and parietal (P) are relatively more activated than the OFC and nearby areas. Thus, context alters how parts of the brain come into play in choice and simultaneously how analytical functions of the brain are recruited.

Employing this paradigm, Rustichini et al. (2005) added ambiguous and partially ambiguous gambles. They uncovered key aspects of the choice process that are involved when subjects work with explicit probabilistic representations and payoffs. Subjects behaved as if they were employing cut-offs to distinguish between numerical magnitudes; it was also shown that the closer the gamble evaluated was to the cut-off the more difficult the judgment (that is, the longer was the reaction time.) Areas of major activation found by Rustichini et al. included P, precuneus (Pr) and Brodman area 6. Rustichini et al. raised the possibility that such cut-off rules operate as approximate calculations like those found by Dehaine et al. when subjects compare numbers to a criterion. In monkeys Dehaine et al. isolated the horizontal inferior parietal (HIP) area as an area capable of making relational comparisons.

Within the classical paradigm Hsu et al. (2005) studied ways in which the brain processed information differently under ambiguity and risk. Using three different approaches to approximating ambiguous and risky tasks, they identified the A and OFC as areas in which ambiguity and risk are differentially processed. In supplemental materials the authors reported inferior parietal activation, which is consistent with giving the subjects both verbal and numerical representations of the choice.

Leland and Grafman (2005) also studied the traditional type of economic tasks. Their study was constructed along the lines of the Bechara and Damasio (2005) studies since they

used normal subjects and subjects with brain damage to the VPC. There was no difference between the performance of these groups on these traditional types of tasks, which is consistent with the proposition that people recruit areas other than orbital frontal cortex in performing these tasks.

Another study that examined economic behaviour in a more traditional choice context is McClure et al. (2004), who studied whether agents have a propensity to discount hyperbolically. They found that the evaluation of immediate payoffs produced relatively more VPC activation, but for all decisions (those involving immediate and non-immediate payoffs) a broader set of areas including the Pr and the P areas was activated.

Camille et al. (2004) examined the degree to which normal and subjects with VPC lesions incorporate regret into their choices. In this study regret is the maximum difference in payoffs that exists between two choices. Camille et al. reported that normal subjects are much more likely to incorporate regret into their choices. The authors found that normal and lesioned subjects both incorporated expected value into their choices. In this study subjects saw gambles represented explicitly in terms of payoffs and probabilities. Feedback was provided after every choice. The results of this study and Hsu et al.'s results imply that some of the more analytic processes implied by Dickhaut et al. and Rustichini et al. can be at work in these studies, but that there is emerging a potentially delicate interplay between the reward areas and the analytical areas of the brain.

Results related to game theory

7. *Monkeys' neuronal activity encodes mixed strategies*. Dorris and Glimcher (2004) extended the examination of the behaviour of monkeys to consider how a monkey plays against different strategies of the computer in a game with a mixed strategy equilibrium. Results reveal that monkeys are capable of adjusting their mixed strategies approximately optimally to the mixed strategies played by the computer. Dorris and Glimcher examined the behaviour of LIP neurons and found that they reflected the mixed strategy of the monkeys.

8. *Games with other agents are consistent with a theory of mind*. In typical game theory experiments it is customary to attempt to give players a complete description of the game, from which strategic behaviour ensues. Then it is assumed that individual players generate beliefs contingent on their beliefs about others' strategies. Given this perspective of how choice proceeds, technically it becomes useful to have an experimental design that attempts to ensure that every player has the chance to fully anticipate the other players' actions prior to any moves made by any of the players. Often this common knowledge approach is approximated by representation of a game matrix in a simultaneous-play game or a game tree in a sequential game.

Neuroscientists have isolated the paracingulate cortex (ParC) as a location associated with the ability to understand another person's deception. Utilizing this perspective, McCabe et al. (2001) investigated whether this area was implicated in cooperative games such as the trust game. They uncovered increased ParC activity when subjects knew they were playing against a person as opposed to a computer, and also found increased ParC activity for cooperative as opposed to non-cooperative players. Sanfey et al. (2003) further examined the McCabe results by employing the ultimatum game and Prisoner's Dilemma games. They preprogrammed a set of outcomes for the subjects to play against. The experimenters attempted to lead subjects to believe they were playing against computers for one set of outcomes and against real people for the other set. These differences in procedure yielded some differences in the brain areas activated. McCabe et al. (2001) found P activation that is not reported by Sanfey et al. However, Sanfey et al. found temporal (T), FL and Pr activation in addition to ParC activation.

9. *Economic reputation building is identifiable at a neuronal level*. King-Casas et al. (2005) used fMRI to scan pairs of subjects in a trust game repeated ten periods. The researchers were able to show that activations in the middle cingulate cortex (MCC) of a sender (when an amount is invested) were coterminous with activations of the anterior cingulate cortex (ACC) of the receiver in the game when the receiver saw the money sent. The receiver's intent to reciprocate was reflected in activation of the receiver's caudate nucleus (CN). Initially this activation lagged the receipt of the investment by approximately eight seconds, but with repeated play the activation precedes receiver's knowledge of the investment by approximately eight seconds. In this way the authors implicitly measured the way economic reputation is built by the sender in the receiver's brain in the trust game.

10. *The brain has mechanisms that reveal individuals enjoy punishing norm violators*. De Quervain et al. (2003) examined the neuronal basis of costly punishment. They allowed the sender to penalize the receiver when the receiver did not reciprocate, but at a cost to the sender. They found evidence consistent with the assumption that the sender was comparing the costs of punishment with a derived benefit (satisfaction) from punishing. The locus of the derived benefit from punishing was reflected in behaviour of the CN and the VPC, the area in which the authors argued the evaluations take place.

Conclusion

Neuroeconomics has moved the economics from the discussion of useful fictions regarding choice to the direct examination of the structures in the human brain that are making the choices. Evidence to date suggests that the underpinnings of modern-day *homo economicus* are reflected in brain structures that exist in both monkeys and humans and in both Robinson Crusoe and multi-agent settings, and findings are emerging on which a more informed model of choice and exchange can be formulated using brain function as the underpinning.

See Also

- altruism in experiments
- behavioural game theory
- evolutionary economics
- experimental methods in economics
- market institutions
- trust in experiments
- uncertainty

This material is based upon work supported by the National Science Foundation under Grant No. 0452477.

Bibliography

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: critique des postulates et axiomes de l'école Américane. *Econometrica* 21, 503–46.

Bechara, A. and Damasio, A.R. 2005. The somatic marker hypothesis: a neural theory of economic decisions. *Games and Economic Behavior* 52, 336–72.

Berns, G.S., McClure, S.M., Pagnoni, G. and Montague, P.R. 2001. Predictability modulates human brain response to reward. *Journal of Neuroscience* 21, 2793–8.

Camille, N., Coricelli, G., Sallet, J., Pradat, P., Duhamel, J.R. and Sugrue, A. 2004. The involvement of the orbitofrontal cortex in the experience of regret. *Science* 304, 1167–70.

De Quervain, J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A. and Fehr, E. 2003. The neural basis of altruistic punishment. *Science* 425, 785–91.

Dehaene, S., Piazza, M., Pinel, P. and Cohen, L. 2003. Three parietal circuits for number processing. *Cognitive Neuropsychology* 20, 487–506.

Dickhaut, J., McCabe, K., Nagode, J., Rustichini, A. and Pardo, J. 2003. The impact of the certainty context on the process of choice. *Proceedings of the National Academy of Science* 6, 3536–41.

Dorris, M.C. and Glimcher, P.W. 2004. Activity in posterior parietal cortex is correlated with relative subjective desirability of action. *Neuron* 44, 365–78.

Ellsberg, D. 1961. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Fiorillo, C.D., Tobler, P. and Schultz, W. 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299, 1898–902.

Hsu, M., Bhat, M., Adolphs, R., Tranel, D. and Camerer, C.F. 2005. Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680–3.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R. and Montague, P.R. 2005. Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83.

Knutson, B., Fong, G.W., Bennett, S.M., Adams, C.M. and Hommer, D. 2003. A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *NeuroImage* 18, 263–72.

Leland, J.W. and Grafman, J. 2005. Experimental tests of the somatic marker hypothesis. *Games and Economic Behavior* 52, 386–409.

McCabe, K., Houser, D., Ryan, L., Smith, V. and Trouard, T. 2001. A functional imaging study of cooperation in two person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America* 98, 11832–5.

McClure, S.M., Laibson, D.I., Lowenstein, G. and Cohen, J. 2004. Separate neural systems value immediate and delayed monetary rewards. *Science* 306, 503–7.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. and Dolan, R. 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–4.

Platt, M.L. and Glimcher, P.W. 1999. Neural correlates of decision variables in parietal cortex. *Nature* 400, 233–8.

Rustichini, A., Dickhaut, J., Ghirardato, P., Smith, K. and Pardo, J.V. 2005. A brain imaging study of the choice procedure. *Games and Economic Behavior* 52, 257–82.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E. and Cohen, J.D. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300, 1755–8.

Schultz, W., Dayan, P. and Montague, P.R. 1997. A neural substrate of prediction and reward. *Science* 275, 1593–7.

Shadlen, M.N. and Newsome, W.T. 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology* 86, 1916–36.

Tremblay, L. and Schultz, W. 1999. Relative reward preference in primate orbitofrontal cortex. *Nature* 398, 704–8.

How to cite this article

Dickhaut, John and Aldo Rustichini. "neuroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000157> doi:10.1057/9780230226203.1177

The New Palgrave Dictionary of Economics Online

neutral taxation

Arnold C. Harberger

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Formally, neutral taxation is taxation falling on something that is in completely inelastic supply, with the tax being so designed as not to affect resource allocation either within or among the affected categories or between them and the other activities not subject to the tax. To minimize deadweight loss, the Ramsey rule says that, the more demand-elastic a good is, the less it should be taxed. But in practice, given ignorance about demand elasticities, uniform low-rate, broad-based taxation reliably reduces deadweight loss and implies neutrality on the part of the state between citizens' preferred actions within the rule of law.

Keywords

deadweight loss; efficiency vs. equity; elasticity; Harberger, A. C.; land tax; neutral taxation; optimal taxation; Ramsey rule taxation; uniform taxation; value-added tax

Article

One can detect in the literature of economics two important lines of thinking on the subject of neutral taxation. One emphasizes economic efficiency (i.e. the elimination of deadweight loss) as the objective in terms of which the neutrality of taxation is defined. The other emphasizes the generality of a tax as itself imparting the quality of neutrality. Two examples, each with a long history in economic thinking, illustrate the main lines of the distinction.

On the one hand we have the taxation of land rents or land values. It builds on the notion (not precisely true in fact) that each piece or plot of land is totally fixed in supply, with the consequence that any tax levied upon it will ultimately be paid out of its pure economic rent.

On the other hand we have the relatively modern idea of a general tax on value added, the tax being applied at a uniform rate on all activities in the economy. Here there is no thought that the underlying resources are fixed in each activity; quite to the contrary, mobility among the various taxed activities is taken for granted for most of the resources on whose product the tax will fall.

It is easy enough by making artful assumptions to bring these two notions very close together. For example we can *assume* that no manmade improvements to the soil are possible, or alternatively that the tax assessors can always distinguish between ‘the intrinsic and immutable qualities of the soil’, on which tax is then duly assessed, and the manmade improvements thereon or accretions thereto, on which (under our convenient assumption) no tax is either assessed or paid. Similarly, we can *assume* for the value added tax that there are just three basic resources in the economy – land, labour and capital – and that each of them is fixed in supply. Therefore a uniform tax on the marginal product of any one of them will be neutral, striking the factor equally regardless of the end use to which it is applied, and leaving the factor (because of the assumed zero-elasticity of its supply) no untaxed haven (not even leisure) to which it might choose to escape.

The above assumptions make it easy to define neutral taxation for a Dictionary. (Neutral taxation is taxation falling on something that is in completely inelastic supply, with the tax being so designed as not to affect resource allocation either within or among the affected categories or between them and the other activities not subject to the tax.) But it would probably not add much to the usefulness of the Dictionary.

To be truly useful, I believe, a definition of neutral taxation should be able to throw away such artificial crutches as the two assumptions presented above. It should be able to live in the real world, where we know that the relevant supply elasticities are rarely zero, but where we do not feel at all sure about their magnitudes nor how they vary as between the short, middle and long run. It should be able to cope with reality that, for tax policy at least, the objects of tax do not have an independent essence as commodities; rather, a commodity subject to tax is whatever the tax law (including the regulations and practices followed in enforcing that law) defines it to be. And finally it should come to grips with the serious claims that can be made for considering equality (among the affected activities) in the applicable tax rate to be an attribute whose presence connotes neutrality and whose absence creates a presumption of non-neutrality.

Economics has come the farthest in responding to the first of the desiderata expressed above. Deadweight loss is a concept completely familiar to the discipline, as is the idea of minimizing the deadweight loss of raising a certain amount of tax revenue subject to given constraints. A clear line of thinking runs from Ramsey in the 1920s through Hotelling in the 1930s, Meade in the 1940s, Corlett and Hague and Lipsey and Lancaster in the 1950s, Harberger in the 1960s, to the modern writers on optimal taxation of whom Atkinson, Diamond, Dixit, Mirrlees, and Stiglitz are a representative few. Flowing through this strand of thought are the related ideas (a) that uniform taxation is not always neutral; (b) that the special condition under which uniform taxation of a subset of commodities or activities minimizes the deadweight loss of raising a given amount of revenue from that subset is met when the equilibrium quantity (or activity level) of *each* member of the taxed subset would respond in the same proportion to a (hypothetical) uniform tax on all goods or activities that are *not* in the taxed subset; and (c) that whenever the condition stated in (b) is *not* met then instead of uniform taxation the minimization of deadweight loss requires higher-than-average taxation on goods whose quantities would fall as a result of a (hypothetical) uniform tax on the uncovered group and lower-than-average taxation on those whose equilibrium quantities would rise most sharply.

The analysis underlying the above statements is straightforward, and one can even call economic intuition into play to explain the conclusion. If the tax authorities are denied the possibility of taxing certain goods or activities, then it can to some degree ‘get around’ the ban by putting higher taxes on

those items within the taxable subset which are complements of those that cannot be taxed. In a similar vein, since one way of thinking of the resource misallocation that occurs when only a subset of activities is allowed to be taxed is that resources are ‘artificially’ shunted from the taxed to the untaxed subset, it seems quite plausible that the optimal patterning of tax rates within the taxed subset should entail taxing at somewhat lower-than-average rates those particular activities in which a percentage point increment of tax would lead to notably greater-than-average ‘shunting’ of resources to untaxed activities.

The line of reasoning just presented is persuasive – sufficiently so that some economists have been tempted to write off uniformity altogether as a plausible objective of tax policy. There remain many, however, who adhere to uniformity as a goal. Given the ease with which propositions (a) through (c) above can be derived, one should hope that most of those who hold to uniformity base their adherence on considerations extraneous to the derivation, say, of the Ramsey rule and other similar propositions in the literature on optimal taxation. The discussion that follows assumes so.

To build a case for uniformity in taxation in the face of the foregoing logic, one should (appropriately, I think) postulate that one is not dealing with two quite arbitrary categories of goods and/or activities, viz., the taxed subset and the untaxed subset. Instead, one should assume that the taxed subset, rather than being ‘any arbitrary bundle’, is so selected as to contain all the goods and activities that can plausibly and without unusual administrative or regulatory effort be brought into the tax net. One then proceeds to view the problem not as a simple analytical puzzle but as one of guiding or governing the interaction between the society's fiscal authorities and its members.

With this objective in mind, an advocate of uniform taxation might set up a quite different problem from that posed earlier. He might consider the ‘disturbance’ with which he is dealing to be a consumer changing his mind about how to spend his money or a worker changing his preference about where or for whom to work. A uniform-tax advocate would likely place a considerable value on the authorities' simply not caring about these various changes of mind.

When one solves the Ramsey problem one takes as given the tastes and preferences of economic agents and maximizes government revenue for a given aggregate level of the agents' welfare. Under the differentiated set of tax rates that emerges from this exercise, the maximizer is not indifferent to changes in tastes of the agents. The maximizer likes it when agents shift their tastes from low-taxed to high-taxed activities, and is disappointed by shifts in the other direction.

Something of the same thing occurs when uniform taxation is implemented. Here the ‘good’ event would be a shift in tastes that caused untaxed activities to contract and taxed activities to expand; the ‘bad’ event would be the opposite. But there would be a wide range of changes of tastes that would be neutral – these would cover shifts among commodities or activities within the sector subject to the uniform tax, and also shifts among activities in the untaxed sector. To the degree that the authorities are successful in extending the tax net over quite a wide range, it may turn out to be true that most changes in tastes simply lead to shifts in the composition of goods within the taxed group. This is the sort of scenario that would best fit the vision of an advocate of broad-based, uniform taxation and at the same time would (at least if changes in tastes within the taxed sector were frequent and important) create problems for proponents of Ramsey rule taxation.

Subtle overtones of a less technical nature also arise when Ramsey-rule taxation is compared to a broad-based, uniform levy. In Ramsey-rule taxation individuals are genuinely presented with incentives to shift their demand from high-taxed to low-taxed products, and workers are likewise motivated to shift their labour efforts from high-taxed to low-taxed activities. Both these incentives are counterproductive from

the social point of view. Subtly hidden in the way the problem is framed is the assumption that people's tastes are given. The reality of the world is that tax laws change only rarely; once enacted, they stay in effect for long periods of time, over which economists can be certain that there will be important changes in the parameters of tastes and technology. The goal of having a tax system that is *robust* against these unknown future shifts in demand and supply is not capricious; it deserves to be taken seriously.

In a quite different vein, there arises the question of to what degree we want our choice of tax patterns to depend on parameters like elasticities of supply and demand about which our knowledge is very spotty and imperfect. Proponents of uniform taxation can fairly argue that their choice of such a form does not depend seriously on knowledge about the parameters of demand and supply. Economic theory assures us that the dominant force is substitution (in the sense that a tax on an activity will, other things equal, cause that activity to contract). There is thus a very strong presumption that broadening the coverage and lowering the rate of a uniform tax will reduce the deadweight loss associated with it (for given revenue yield). One can build policy on this basis without having any detailed knowledge of the parameters of supply and demand, without any particular hope of gaining anything more than a very patchy knowledge about them in the future, and indeed *with* an almost absolute assurance that whatever the relevant parameters might be now, they will undergo substantial changes in the future. If one believes that these conditions come close to describing our present and likely future state of knowledge about the relevant parameters, he will likely be predisposed toward uniform as against Ramsey-rule taxation.

The last line of argument favouring uniform taxation has to do with the interplay between equity and efficiency considerations in governing tax policy. The motivations that fall under the umbrella of ‘equity’ are too numerous and too varied to try to recount here. But nowhere among them can one find that it is fairer to tax more heavily factors of production that cannot flee to other activities or that it is more just to tax heavily those items whose demand happens to be less elastic. To tax salt more heavily than sugar simply and solely because it has a lower elasticity of demand is at least as capricious (from the standpoint of equity) as taxing people differently according to the colour of their eyes.

Ultimately, I believe, the issue of uniform versus Ramsey-rule taxation may turn out to be just one facet of much broader philosophical differences. Consider the philosophy of government that assigns to government the role of creating a framework of laws and regulations within which the private sector then is encouraged to operate freely. Under this philosophy a positive value is placed on the authorities' not caring about what private agents do (so long as they abide by the rules). It is a position desideratum to create a tax system that is robust against changes in tastes and technology.

On the other side of the coin we have a philosophy of social engineering, in which the detailed tastes and technology of the society enter as data into a process by which the policy makers choose parameters such as tax rates and coverages so as to maximize some measure of social net benefit.

Each of these philosophies has had its own long trajectory within the profession of economics. Each has its representatives today. Each will surely be reflected in the literature of future decades. In my opinion, the future debate as to how the concept of neutrality in taxation should be reflected in real-world policy decisions will swirl around the subtle differences between the ways in which holders of these two philosophies view the world, between the roles they envision for government, and between the ways they see the science of economics interacting with government in the formation of policy.

See Also

- optimal taxation
- public finance
- Ramsey pricing

Bibliography

Atkinson, A.B. 1977. Optimal taxation and the direct versus indirect tax controversy. *Canadian Journal of Economics* 10, 590–606.

Atkinson, A.B. and Stiglitz, J.E. 1980. *Lectures on Public Economics*. New York and Maidenhead: McGraw-Hill (esp. Lectures 12–14).

Corlett, W.J. and Hague, D.C. 1953. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21, 21–30.

Diamond, P.A. and Mirrlees, J.A. 1971. Optimal taxation and public production. I: Production efficiency; II: Tax rules. *American Economic Review* 61, 8–27; 261–78.

Dixit, A.K. 1970. On the optimum structure of commodity taxes. *American Economic Review* 60, 295–301.

Harberger, A.C. 1964. Taxation, resource allocation and welfare. In *The Role of Direct and Indirect Taxes in the Federal Revenue System*, ed. J.F. Due. Princeton: Princeton University Press.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6, 242–69.

Lipsey, R.G. and Lancaster, K. 1956–7. The general theory of second best. *Review of Economic Studies* 24, 11–32.

Meade, J.E. 1955. *Trade and Welfare*. Vol. 2: *Mathematical Supplement*. Oxford: Oxford University Press.

Mirrlees, J.A. 1976. Optimal tax theory: a synthesis. *Journal of Public Economics* 6, 327–58.

Mirrlees, J.A. 1979. The theory of optimal taxation. In *Handbook of Mathematical Economics*, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

How to cite this article

Harberger, Arnold C. "neutral taxation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000053> doi:10.1057/9780230226203.1178

The New Palgrave Dictionary of Economics Online

neutrality of money

Don Patinkin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

homogeneity postulate; homogeneity property; Hume, D.; inflation; inside and outside money; intermediate products; IS-LM model; liquidity preference; money illusion; money supply; neutrality of money; nominal interest rate; perfect foresight; Phillips curve; quantity theory of money; rational expectations; real interest rate; superneutrality; time preference

Article

‘Neutrality of money’ is a shorthand expression for the basic quantity-theory proposition that it is only the level of prices in an economy, and not the level of its real outputs, that is affected by the quantity of money which circulates in it. Thus the notion – though not the term – goes back to early statements of the quantity theory, such as the classic one by David Hume in his 1752 essays ‘Of Money’, ‘Of Interest’ and ‘Of the Balance of Trade’. At that time the notion also served as one of the arguments against the mercantilist doctrine that the wealth of a nation was to be measured by the quantity of gold (which in 18th-century England constituted a – if not the – major form of metallic money: Feaveryear, 1963, p. 158) that it possessed. The term itself is much more recent. Though attributed by Hayek (1935, pp. 129–31) to Wicksell, it is actually due to continental economists in the late 1920s and early 1930s to whom Hayek also refers (see 1935, pp. 129–31; see also Patinkin and Steiger, 1988).

1. The rigorous demonstration of, the neutrality of money is based on the critical assumption that individuals are free of ‘money illusion’. An individual is said to suffer from such an illusion if he changes his economic behaviour when a currency conversion takes place: when, for example (as in Israel in 1985), a new monetary unit – the ‘new shekel’ – is introduced in circulation and declared to be equivalent to 1,000 old shekels.

It can be shown (Patinkin, 1965) that an illusion-free individual in an economy with borrowing who maximizes utility subject to his budget constraint will have demand functions which depend on relative prices, the rate of interest, and the real value of his initial wealth – which consists of physical capital, bond holdings, and money balances. That is, the demand of this representative individual for the *j*th good, d_j, is described by the function

$$d_j = f_j(p_1/p, \ldots, p_{n-2}/p, r, K_0 + B_0/p + M_0/p)(j = 1, \ldots, n-2),$$

where the p_j are the respective money (or absolute) prices of the $n - 2$ goods; p is the average price level as defined by $p = \Sigma_j w_j p_j$ where the w_j are fixed weights; r is the rate of interest; K_0 is physical capital, B_0 is the initial nominal value of bond holdings (which, for a debtor, is negative), and M_0 is the initial quantity of money. Thus when the new shekel is introduced in circulation, the price of each good in terms of this shekel (and hence the general price level), the terms of indebtedness, and the nominal quantity of initial money holdings are respectively reduced to 1/1,000th of what they were before; hence relative prices and the real value of initial wealth are unaffected; hence so are the amounts demanded of each good.

Mathematically, the foregoing property of the demand functions is described by the statement that these functions are homogeneous of degree zero in the money prices *and* in the

initial quantity of financial assets, including money. Accordingly, the absence of money illusion is sometimes referred to as the homogeneity property of the demand functions. (For the necessary and sufficient conditions that must be satisfied by the utility function in order to generate such illusion-free demand functions, see Howitt and Patinkin, 1980.) This homogeneity property is to be sharply distinguished from what the earlier literature denoted as the 'homogeneity postulate', by which it meant the invariance of demand functions with respect to an equiproportionate change in money prices alone, and which invariance it erroneously regarded as the condition for the absence of money illusion and hence for the neutrality of money (Leontief, 1936, p. 192; Modigliani, 1944, pp. 214–15): for even in the case of an individual who is neither debtor nor creditor, such a change affects the real value of his initial money balances, hence is not analogous to a change in the monetary unit, and hence – by virtue of the real-balance effect – will generally lead him to change the amounts he demands of the various goods.

For a closed economy, the aggregate value of B_0 is obviously zero, for to each creditor there corresponds a debtor. For simplicity, we can also consider the amount of physical capital, K_0, to remain constant. Disregarding distribution effects, the demand functions of the economy as a whole for the n–2 goods can then be represented by

$$D_j = F_j(p_1/p, \ldots, p_{n-2}/p, r, M_0/p) \; (j = 1, \ldots, n-2)$$

and the corresponding supply functions by

$$S_j = G_j(p_1/p, \ldots, p_{n-2}/p, r).$$

The general-equilibrium system of the economy is then

$$\left| \begin{array}{l} F_1(p_1/p, \ldots, p_{n-2}/p, r, M_0/p) = G_1(p_1/p, \ldots, p_{n-2}/p, r) \\ \vdots \quad \vdots \quad \vdots \\ F_{n-2}(p_1/p, \ldots, p_{n-2}/p, r, M_0/p) = G_{n-2}(p_1/p, \ldots, p_{n-2}/p, r) \\ F_{n-1}(p_1/p, \ldots, p_{n-2}/p, r, M_0/p) = 0 \\ F_n(p_1/p, \ldots, p_{n-2}/p, r, M_0/p) = M_0/p. \end{array} \right|$$

The (n – 1)st equation is for real bond holdings, whose aggregate net value is (as already noted) zero; and the nth equation is for real money balances. Assume that this system has a unique equilibrium solution with money prices $p_1^0, \ldots, p_{n-2}^0, p^0$ and the rate of interest r^0, and that the economy is initially at this position. Let the quantity of money now be changed to kM_0, where k is some positive constant. From the preceding system of equations we can immediately see that (on the further assumption that the system is stable) the economy will reach a new equilibrium position with money prices $kp_1^0, \ldots, kp_{n-2}^0, kp^0$ and an unchanged rate of interest r^0. (Clearly, this conclusion would continue to hold if the supply functions $G_j(\ )$ were also dependent on M_0/p.) Thus the increased quantity of money does not affect any of the real variables of the system, namely, relative prices, the rate of interest, the real value of money balances, and hence the respective outputs of the n–2 goods. In brief, money is neutral: or in the picturesque phrase which Robertson (1922, p. 1) apparently coined, money is a veil. (For empirical studies, see Lucas, 1980, and Lothian, 1985.)

Furthermore, Archibald and Lipsey (1958) have shown that if the initial equilibrium exists not only with respect to the economy as a whole, but also with respect to each and every individual in it (which, inter alia, means that each individual was initially holding his optimum quantity of money), then this neutrality will obtain in the long run even if one does take account of distribution effects. That is, even if one takes account of differences in tastes, endowments, and hence individual demand functions, an increase in the quantity of money,

no matter how distributed among individuals, will in the long run cause an equiproportionate increase in prices and leave the rate of interest invariant. This conclusion in turn follows from the fact that the sequence of short-run equilibria generated by the increase in the quantity of money will in the long run redistribute this quantity in a way that results in an equiproportionate increase in the money holdings of each individual, relative to his holdings in the initial equilibrium position (see also Patinkin, 1965, pp. 50–9).

It should also be noted that the preceding analysis has implicitly assumed a unitary elasticity of expectations with respect to future prices, so that neutrality is not disturbed by substitution between present and future commodities.

2. The conclusions of the foregoing analysis are clearly those of long-run comparative-statics analysis. It was this fact that led Keynes – even in his quantity-theory period as represented by his *Tract on Monetary Reform* (1923) – to disparage their policy implications with the famous remark that ‘*in the long run* we are all dead’ (1923, p. 80, italics in original). It should therefore be emphasized that at the same time they demonstrated the long-run neutrality of money, quantity theorists (including Keynes of the *Tract*) also emphasized its non-neutrality in the short run (Patinkin, 1972a). Thus Hume emphasized that prices do not immediately rise proportionately to the increased quantity of money and that in the intervening period this stimulates production. In Hume's words:

> it is of no manner of consequence, with regard to the domestic happiness of a state, whether money be in a greater or less quantity. The good policy of the magistrate consists only in keeping it, if possible, still increasing; because, by that means, he keeps alive a spirit of industry in the nation … (1752, pp. 39–40)

Hume's emphasis on the irrelevance of the absolute level of the money supply (and hence of money prices) in contrast with the significance of the rate of change of this level was also made by later quantity-theorists. Some of them stressed the stimulating effects of rising prices on ‘business confidence’ and hence economic activity. A more frequent explanation of the short-run non-neutrality of money was in terms of the shift in the distribution of real income as between creditors and debtors generated by a changing price level. Of particular importance was the danger that a sharply declining price level would increase the number of bankruptcies among debtors, with all its adverse repercussions on the economy. Another source of non-neutrality was the fact that individual prices do not change at the same rate in response to a monetary change. Thus if after a monetary decrease, wage rigidities cause the decline in wages to lag behind that of product prices, the resulting increase in the real wage rate would generate unemployment; conversely, the lag of wages in the case of an inflation would increase profits and hence stimulate production. This consideration led some quantity-theorists to deny even the long-run neutrality of money on the grounds that profit-recipients had a higher tendency to save than wage-earners, so that the shift in income in favour of profits would increase savings, and that these would lead to an increase in the real stock of physical capital in the economy, and hence to a decline in the long-run rate of interest.

For Irving Fisher, the important lag was that of the nominal rate of interest behind the rate of (say) inflation generated by a monetary increase. In particular, because of the lack of perfect foresight on the part of savers (who are the lenders), the nominal rate does not rise sufficiently to offset this inflation; and the resulting decline in the real rate of interest causes entrepreneurs to increase their borrowings, hence investments and economic activity in general. Conversely, when prices decline, corresponding misperceptions cause an increase in the real rate of interest and hence a decline in economic activity. Indeed, Fisher (1913, ch. 4) based his whole theory of the business cycle on this process: the cycle was for him ‘the dance of the dollar’ (Fisher, 1923).

The greatly increased importance of income and capital-gains taxation since Fisher's time is the background of the present-day view – much stressed by Feldstein (1982, and references there cited) – that inflation would have real effects on the economy even if there were perfect foresight, so that the nominal rate fully adjusted itself to the rate of inflation, leaving the real rate of interest unchanged. This is particularly true for the taxation of income from capital, with the simplest example being the increased tax burden on corporations generated by the calculation of depreciation expenses on the basis of historical (as distinct from replacement) costs in an inflationary economy (see also Birati and Cukierman, 1979). This is a specific instance of the short-run non-neutrality of money generated by the existence of a tax structure formulated in nominal terms (as is the case with, for example, specific taxes and income-tax brackets) which are generally adjusted to the rate of inflation only after a lag.

Short-run non-neutrality is a basic feature of Keynesian monetary theory and stems from the contention that in a situation of unemployment, prices will not rise proportionately to the increased quantity of money, and that the resulting increase in the real quantity of money will cause a decline in the rate of interest and hence an increase in the volume of investment and the level of national income. The short-run non-neutrality of money is, however, also a basic tenet of today's monetarists, who contend that though the long-run effect of a change in the quantity of money is primarily on prices, its short-run effect is primarily on output. In Friedman's words: ‘In the short run, which may be as much as five or ten years, monetary changes affect primarily output. Over decades, on the other hand, the rate of monetary growth affects primarily prices’ (Friedman, 1970, pp. 23–4).

This non-neutrality has been rationalized by Lucas (1972) in terms of the individual's inability to determine whether a change in the price of a good with which he is particularly concerned (for example labour, in the case of a wage-earner) is a change only in the price of that good (in which case it represents a change in its relative price, which calls for a quantity adjustment) or is part of a general change in prices which does not affect relative prices. In accordance with this approach, and under the assumption that markets always clear, it has also been claimed that only an unanticipated change in the quantity of money will have real effects; for an anticipated one will be expected by the individual to affect all prices proportionately (Lucas, 1975; Barro, 1976). A far-reaching corollary of this claim is that if, in accordance with the assumption of rational expectations, the public anticipates the actions that government will carry out within the framework of its proclaimed monetary policy, then this policy too will be neutral: that is, the systematic component of monetary

policy will not affect any of the real variables of the system (cf. McCallum, 1980 and references there cited). Thus under these circumstances even the short-run Phillips curve is – from the viewpoint of systematic monetary policy – vertical.

Empirical support for the claim that only unanticipated monetary changes will have real effects was at first provided by Sargent (1976) and Barro (1978). Contrary conclusions were, however, reached in subsequent empirical studies by Fischer (1980), Boschen and Grossman (1982), Gordon (1982), Mishkin (1982; 1983) and Cecchetti (1986). These differing conclusions stem from different views about the respective ways to estimate (1) that part of a monetary change that is anticipated and/or (2) the extent of the time lags that must be taken account of in measuring the effects of a monetary change on output. In any event, the weight of opinion today is that both anticipated and unanticipated changes in the money supply have short-term real effects. To the extent that anticipated changes have such effects, this can be interpreted either as reflecting the influence of nominally formulated elements (for example the aforementioned tax structure, or long-term wage contracts – Fischer, 1977) in an economy functioning in accordance with the hypothesis of rational expectations cum market-clearing; or, alternatively, it can be interpreted as a refutation of this hypothesis in part or in whole. Thus once again we are confronted with *la condition scientifique* of our discipline: its inability in all too many cases to reach definitive conclusions about theoretical questions on the basis of empirical studies, an inability which increases directly with the political significance of the question at issue.

3. Neoclassical quantity-theorists contended that a shift in the demand curve for money would also have a long-run neutral effect on the economy. Thus consider the Cambridge cash-balance equation, M=*KPY*, where *Y* is the real volume of expenditures and *K* is that proportion of his planned money expenditures, *PY*, which the individual wishes to hold in the form of money. Assume that the economy is in equilibrium with a fixed quantity of money M_0 and price level P_0. Let there now take place a positive shift in the demand for money – that is, an increase in *K*. Because of the budget constraint, this must be accompanied by a negative shift in the demand for goods. Consequently, the price level *P* will decline until equilibrium is reestablished with the same nominal quantity of money, M_0, but at a lower price level, $P_1 < P_0$. Thus the automatic functioning of the market will in the long run generate the additional quantity of real balances that individuals wish to hold, without affecting the output of goods.

This neutrality can also be demonstrated in terms of the general-equilibrium system presented above. In particular, if we assume that the increased demand for money is accompanied by a symmetric decrease in the demand for all other goods and for bonds, then a new equilibrium will be established with all money prices reduced in the same proportion, and with an unchanged rate of interest; correspondingly, the respective outputs of goods are also unchanged. In Keynesian monetary theory, however, the increased demand for money is assumed to be solely at the expense of bond holdings: this, after all, is an implication of Keynes's theory of liquidity preference. Such a shift in liquidity preference will accordingly not be neutral in its effects; instead, it will cause an increase in the rate of interest with consequent effects on investment and other real variables of the system (Patinkin, 1965, chs VIII:5 and X:4).

In an analogous manner, a change in the proportions between inside and outside money generated by a change in the currency/deposit ratio and/or the bank-reserve/deposit ratio will not be neutral in its effects (Gurley and Shaw, 1960, pp. 231–6). It should, however, be emphasized that if the demand and supply functions of the financial sector are also characterized by absence of money illusion, then an increase in outside money will leave these ratios unchanged and hence be neutral (Patinkin, 1965, ch. XII: 5–6).

So far, our concern has implicitly been an increase in the quantity of money generated by a one-time government deficit, after which the government returns to a balanced budget. This results in an initial net increase in the total of financial assets in the economy and is thus the real-world analytical counterpart of an increase in the quantity of money generated by the proverbial helicopter dropping down money from the skies. If, however, the monetary increase is generated by an open-market purchase of government bonds (so that initially there is no change in total financial assets), and if there is a real-balance effect in the commodity market, then, as Metzler (1951) showed in a classic article, the equilibrium rate of interest will decline, so that money will not be neutral in its effects. If, however, individuals fully anticipate and discount the future stream of tax payments needed to service the government bonds (in which case these bonds are not part of net wealth), neutrality will obtain in this case too (Patinkin, 1965, ch. XII:4).

4. The discussion until this point has dealt almost entirely with the neutrality of a once-and-for-all increase in the quantity of money in a stationary economy. An analogous question arises with reference to the long-run neutrality of a change in the rate of growth of the money supply in a growing economy – in which context the notion is referred to as ‘superneutrality’. Thus consider an economy in steady-state equilibrium whose population is growing at the rate *n*. Assume that the nominal quantity of money is growing at a faster rate, $\mu = \dot{M} / M$ so that (in order to maintain the constant level of per-capita real money balances that is one of the characteristics of such a steady state) prices rise at the constant rate $\pi = \mu - n$. Money is said to be superneutral if (say) an increase in the steady-state rate of its expansion, and hence in the corresponding rate of inflation, will not affect any of the steady-state real variables in the system, with the exception of per-capita real-balances: that is, per-capita capital, *k*; per-capita output, *y*; and the real rate of interest, *r*, equal to the marginal productivity of capital. On the other hand, because of the higher costs of holding real balances – in terms of loss of purchasing power, or, alternatively, in terms of the forgone higher nominal rate of interest, *i*, generated by the increased rate of inflation – the steady-state per capita real value of these balances, *m*, should generally be expected to decrease.

As already indicated, for Irving Fisher (1907, ch. 5; 1913, pp. 59–60; 1930, pp. 43–4) it was only the absence of perfect foresight which prevented such superneutrality from obtaining: for were such foresight to exist, the nominal rate of interest would simply increase so as to compensate for the inflation and thus leave the real rate of interest (which, under the assumption of continuous compounding, equals $i - \pi$) unchanged. Fisher, however, did not take account of the possible effects of the way the increased amount of money is injected into the economy and/or the possible effects of the resulting decrease in real balances on other markets. Thus by assuming that the government increases the quantity of

money in the economy by distributing it to households and thereby increasing their disposable income, Tobin (1965; 1967) – in a generalization of the Solow (1956) growth model to a money economy – showed that a higher rate of inflation will generally cause individuals to change the composition of their asset portfolios by shifting out of real money balances and into physical capital, thus increasing the steady-state values of k and y – and hence (by the law of diminishing returns) decreasing that of r – so that superneutrality does not obtain.

Tobin's analysis assumes a constant savings ratio. In a critique of this analysis, Levhari and Patinkin (1968) showed inter alia that if instead this ratio is assumed to depend positively on the respective rates of return on capital and on real money balances – that is, on the real rate of interest and on the rate of deflation – then an increase in the rate of inflation might decrease steady-state savings and hence k, thus causing an increase in the real rate of interest. Similarly, if real money balances were explicitly introduced into the production function, an increase in the rate of inflation might so decrease these balances as to decrease steady-state per-capita output and hence savings sufficiently to offset the positive substitution effect on k, thus generating a decrease in the latter.

Patinkin (1972b) analysed superneutrality by means of an IS–LM model generalized to a full employment economy with a real-balance effect in the commodity market (the following largely reproduces the relevant material in this reference). As in Solow (1956), the economy is assumed to have a linearly homogeneous production function, $Y=F(K, L)$, where Y is output, K capital, and L labour, with the labour force assumed to be growing at the exogenous rate n. The intensive form of this function is then $y=f(k)$ and its derivative, $f'(k)$ is accordingly the marginal productivity of capital, so that the equilibrium real rate of interest is $r = f'(k)$ Following Mundell (1963; 1965), the crucial assumption of this model is that whereas investment and saving (and hence consumption) decisions depend upon the real rate of interest, $r=i-\pi$, the decision with respect to the amount of real money balances to hold depends on the nominal rate of interest, i– for the alternative cost of holding money instead of a bond is precisely this rate. The same is true if we measure this cost in terms of the alternative of holding physical capital: for the total yield on this capital is its marginal product (equal in equilibrium to the real rate of interest) *plus* the capital gain generated by the price change (π): that is, it is $r+\pi=i$. Alternatively, if we measure rates of return in real terms, the rate of return on money balances is $-\pi$ and that on physical capital r; hence the alternative cost of holding money is the difference between these two rates, or $r-(-\pi)=i$.

Consider now the commodity market. Let E represent the aggregate real demand for consumption and investment commodities combined. For simplicity, assume that this demand is a certain proportion, α, of total real income, Y. Assume further that this proportion depends inversely on the real rate of interest and directly on the ratio of real money balances, M/p, to physical capital, K. The second dependence is a type of real-balance effect, reflecting the assumption that the greater the ratio of real money balances to physical capital in the portfolios of individuals, the more they will tend (for any given level of income) to shift out of money and into commodities. The equilibrium condition in the commodity market is then represented by

$$\alpha(i-\pi, (M/p)/K)\cdot Y = Y. \tag{1}$$

By assumption, $\alpha_1(\cdot)$ is negative and $\alpha_2(\cdot)$ positive, where $\alpha_1(\alpha_2)$ is the partial derivative of $\alpha(\cdot)$ with respect to its first (second) argument.

Consider now the money market. Following Tobin (1965, p. 679), assume that the demand in this market depends on the volume of physical capital and the nominal rate of interest. More specifically, assume that the demand for money is a certain proportion, λ of physical capital. Thus the larger K, the greater (other things equal) the total portfolio of the individuals, hence the greater the demand for money: this can be designated as the scale or wealth effect of the portfolio. Assume further that the proportion λ depends inversely on the nominal rate of interest. That is, the higher this rate, the smaller the proportion of money relative to physical capital which individuals wish to hold in their portfolios: this can be designated as the composition or substitution effect. The equilibrium condition in the money market is then

$$\lambda(i)\cdot K = M/p \tag{2}$$

where by assumption the derivative $\lambda'(\cdot)$ is negative.

Dividing equations (1) and (2) through by Y and K, respectively – and transforming them into per capita form – we then obtain the equations

$$\alpha(i - \pi, m / k) = 1$$
(3)

$$\lambda(i) = m / k.$$
(4)

In the steady state,

$$\mu = \pi + n.$$
(5)

Since μ and n are both assumed to be exogenously determined, the same can be said for the steady-state value of π. Thus in steady states, equations (3) and (4) can be considered as a system of two equations in the two endogenous variables i and m/k, and in the exogenous variable π. On the assumption of the solubility of these equations, the specific value of k (and hence m) can then be determined by making use of the additional equilibrium condition that the marginal productivity of capital equals the real rate of interest, or,

$$f'(k) = i - \pi.$$
(6)

In accordance with the usual assumption of diminishing marginal productivity, we also have

$$f''(k) < 0.$$
(7)

The solution of system (3)–(4) can be presented diagrammatically in terms of Figure 1. The curve *CC* represents the locus of points of equilibrium in the commodity market for a given value of π. Its positive slope reflects the assumption made above about the respective influences of the real rate of interest ($i-\pi$) and of the real-balance effect (as represented by m/k) on α. Namely, a (say) increase in i increases the real rate of interest and thus tends to decrease α: hence the ratio m/k must increase in order to generate a compensating increase in α and thus restore equilibrium to the commodity market. On the other hand, *LL* – the locus of points of equilibriums in the money market – must be negatively sloped: an increase in the supply of money and hence in m/k must be offset by a corresponding increase in the demand for money, which means that i must decline. The intersection of the two curves at *W* thus determines the steady-state position of the economy.

Figure 1

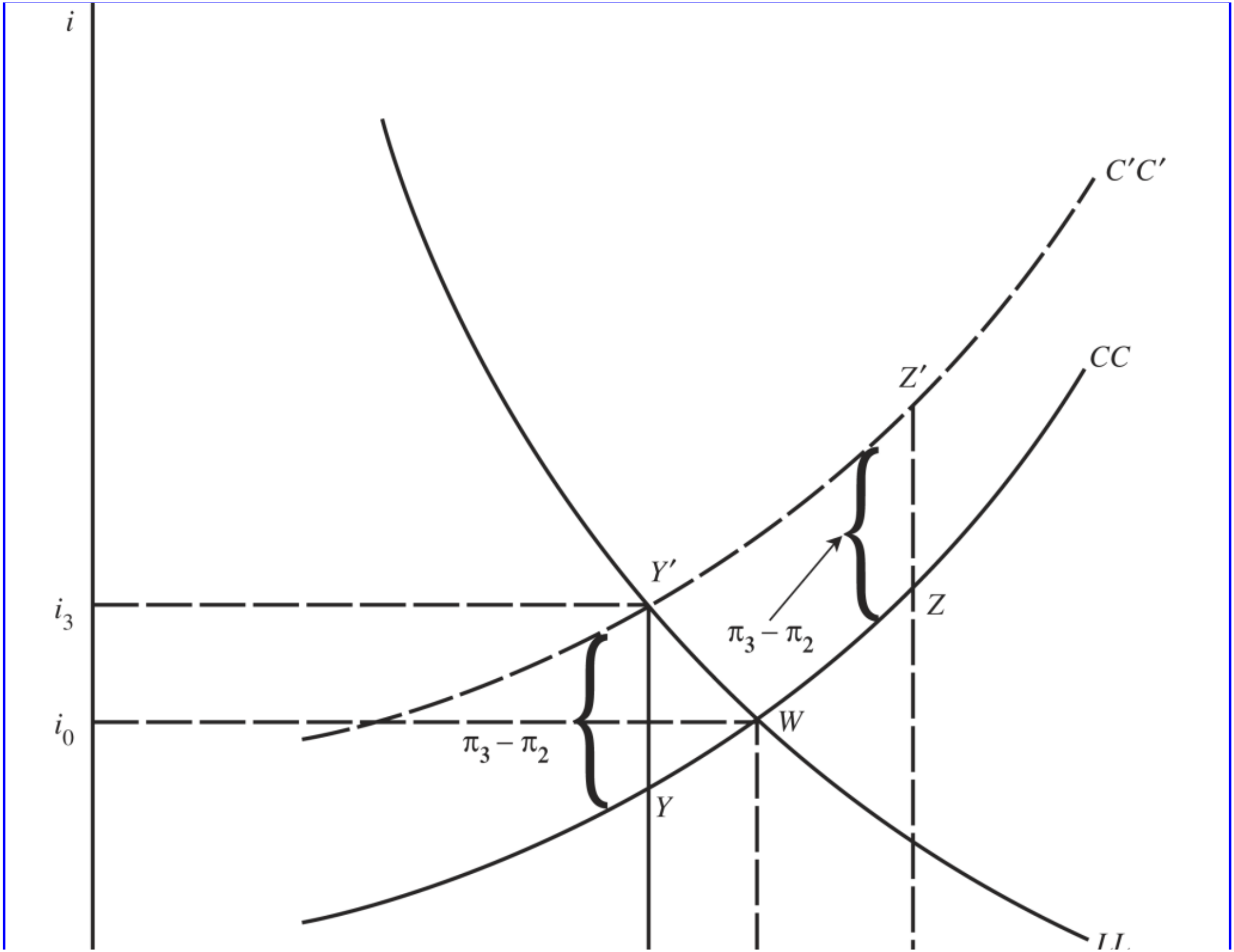
i
C′C′
CC
Z′
Y′
Z
i_3
$\pi_3 - \pi_2$
i_0
W
$\pi_3 - \pi_2$
Y
LL

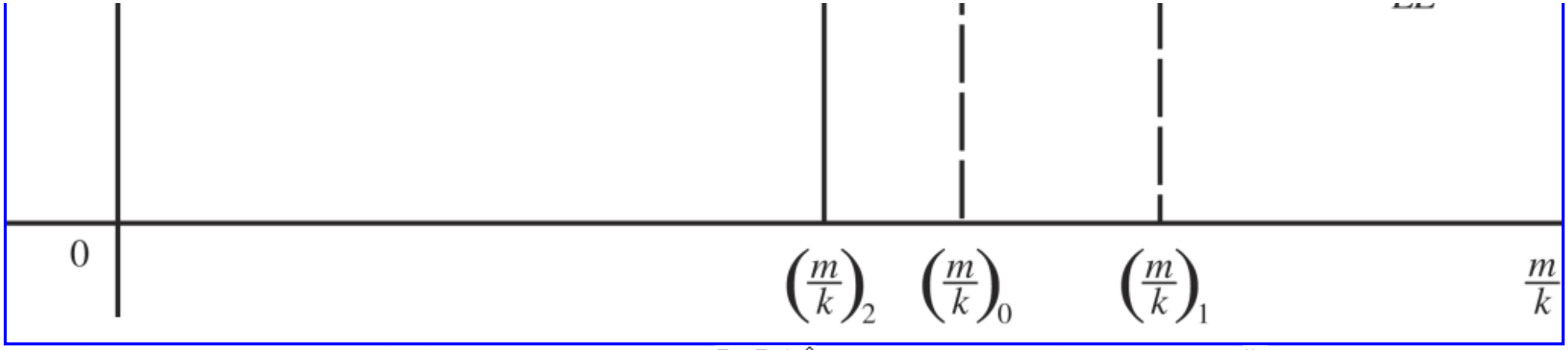

Assume for simplicity that the given value of π for which *CC* and *LL* are drawn is $\pi = \pi_2 > 0$, corresponding to the rate of monetary expansion μ_2. Assume now that this rate is exogenously increased to $\mu = \mu_3$, so that (by (5)) the steady-state value of π is increased accordingly to $\pi_3 = \mu_3 - n > \pi_2$. From the fact that π does not appear in (4), it is clear that *LL* remains invariant under this change. On the other hand, the curve *CC* must shift upwards in a parallel fashion by the distance $\pi_3 - \pi_2$: for at (say) the point Z' on the curve $C'C'$ so constructed, the money/capital ratio *m*/*k* and the real rate of interest $i - \pi$ are the same as they were at point *Z* on the original curve *CC;* hence *Z'* too must be a position of equilibrium in the commodity market.

We can therefore conclude from Figure 1 that the increase in the rate of monetary expansion (and hence rate of inflation) shifts the steady-state position of the economy from *W* to Y'. From the construction of $C'C'$ it is also clear that the real rate of interest at Y' is $r_3 = i_3 - \pi_3$ which is less than the real rate at *W*, namely, $r_0 = i_0 - \pi_2$. Thus the policy of increasing the rate of inflation decreases the steady-state value of the real rate of interest, and also the money/capital ratio.

Because of the diminishing marginal productivity of capital, the decline in *r* implies that *k* has increased. Thus the fact that *m*/*k* has declined does not necessarily imply that *m* has declined. This indeterminacy reflects the two opposing influences operating on *m* reflected in eq. (2), rewritten here in the per capita form as

$$\lambda(i) \cdot k = m. \tag{8}$$

To use the terminology indicated above, the increased inflation increases the steady-state stock of physical capital, and thus exerts a positive wealth effect on the quantity of real-money balances demanded. At the same time, the increased inflation means that the alternative cost of holding money balances (for a given level of *k* and hence *r*) has increased, and this exerts a negative substitution effect on the demand for these balances; that is, individuals will tend to shift out of money and into capital. Thus the final effect on *m* depends on the relative strength of these two forces. As is, however, generally assumed in economic theory, we shall assume that the substitution effect dominates, so that an increase in π decreases *m*.

We now note that the only exogenous variable which appears in system (3)–(5) is the rate of change of the money supply, as represented by its steady-state surrogate, $\pi = \mu - n$. In contrast, the absolute quantity of money, *M*, does not appear. It follows that once-and-for-all changes in *M* (after which the money supply continues to grow at the same rate) will not affect the steady-state values of *m*, *k*, and *i* as determined by the foregoing system for a given value of π. In brief, system (3)–(5) continues to reflect the neutrality of money. On the other hand, because of the Keynesian-like interdependence between the commodity and money markets, the system is not superneutral.

Note that in the absence of this interdependence, the system would also be superneutral. This would be the case either if the demand for commodities depended only on the real rate of interest, and not on *m*/*k* (that is if there were no real-balance effect); or if the demand for money depended only on *k*, and not on the nominal rate of interest – an unrealistic assumption, particularly in inflationary situations which cause this rate to increase greatly.

The first of these cases is analogous to the dichotimized case of stationary macroeconomic models (cf. Patinkin, 1965, pp. 242, 251 (n.19), and 297–8). It would be represented in Figure 1 by a *CC* curve which was horizontal to the abscissa. Correspondingly, the upward shift generated by the rate of inflation would cause the new *CC* curve to intersect the unchanged *LL* curve at a money rate of interest which was $\pi_3 - \pi_2$ greater than the original one, and hence at a real rate of interest (and hence value of *k*) which was unchanged; the value of *m*, however, would unequivocally decline. The second of these cases would be represented by a vertical *LL* curve. Hence the upward parallel shift in the *CC* curve generated

by inflation would once again shift the intersection point to one which represented an unchanged real rate of interest. In this case (which, as already noted, is an unrealistic one) the value of *m* also remains unchanged.

5. A common characteristic of the foregoing money-and-growth models is that their respective savings functions are postulated and not derived from utility maximization. An analysis which does derive consumption (and hence savings) behaviour from such maximization was presented by Sidrauski (1967) in an influential article. As before, consider an economy growing at the constant rate *n* with a linearly homogeneous production function having the intensive form $y=f(k)$. Assume now that the representative individual of this economy is infinitely lived with a utility function which depends on consumption and real balances, and that he maximizes the discounted value of this function over infinite time, using the constant subjective rate of time preference, *q*. Under these assumptions, Sidrauski shows that money is superneutral.

As Sidrauski is fully aware, this conclusion follows from the form of his production function together with his assumption of a constant rate of time preference; for this fixes the steady-state real rate of interest at $r = q + n = f'(k)$, which determines the steady-state value of *k* and hence of *r*. If, however, the production function depends also on real balances – say, $y=g(k, m)$ – then this superneutrality no longer obtains. For the necessary equality between the marginal productivity of capital and $q+n$ in this case is expressed by the equation $g_k(k, m)=q+n$ (where $g_k(k, m)$ is the partial derivative with respect to *k*), which no longer fixes the value of *k* (Levhari and Patinkin, 1968, p. 234). In an analogous argument, Brock (1974) showed that if the individual's utility function depends also on leisure, then an increase in the rate of inflation will affect his demand for leisure, which means that it will affect his supply of labour (that is, labour *per capita*). Hence even though (in accordance with Sidrauski's argument) the increased rate of inflation will not affect the steady-state values of *r*, *k* (that is, capital per *labour-input*), and *y* (that is output per *labour-input*), it will affect the respective amounts of labour and capital *per capita* and hence output *per capita* – so that it will not be superneutral. Needless to say, Sidrauski's results will also not obtain if the rate of time preference is not constant.

6. The conclusion that can be drawn from this discussion is that whereas there is a firm theoretical basis for attributing long-run neutrality to money (but see Gale, 1982, pp. 7–58, and Grandmont, 1983, pp. 38–45, 91–5), there is no such basis for long-run superneutrality: for changes in the rate of growth of the nominal money supply and hence in the rate of inflation generally cause changes in the long-run equilibrium level of real balances; and if there are enough avenues of substitution between these balances and other real variables in the system (viz., commodities, physical capital, leisure), then the long-run equilibrium levels of these variables will also be affected. An exception to this generalization would obtain if money were to earn a rate of interest which varied one-to-one with the rate of inflation, so that the alternative cost of holding money balances would not be affected by changes in the latter rate; but though it is generally true that interest (though not necessarily at the foregoing rate) will eventually be paid on the inside money (that is bank deposits) of economies characterized by significant long-run inflation, this is not the case for the outside money which is a necessary (though in modern times quantitatively relatively small) component of any monetary system.

The discussion to this point has treated the economy's output as a single homogeneous quantity. A more detailed analysis which considers the sectoral composition of this output yields another manifestation of the absence of superneutrality. In particular, it is a commonplace that the higher the rate of inflation, the higher the so-called ‘shoe-leather costs’ of running to and from the banks and other financial institutions in order to carry out economic activity with smaller real money balances. In the case of households, the resulting loss of leisure is denoted as the ‘welfare costs of inflation’ as measured by the loss of consumers' surplus: that is, by the reduction in the triangular area under the demand curve for real money balances (cf. Bailey, 1956). In the case of businesses, the costs of inflation take the concrete form of the costs of the additional time and efforts devoted to managing the cash flow. What must now be emphasized is that the obverse side of the additional efforts of both households and businesses is the additional resources that must be diverted to the financial sector of the economy in order to enable it to meet the increased demand for its services. Thus the higher the rate of inflation, the higher (say) the proportion of the labour force of an economy employed in its financial sector as opposed to its ‘real’ sectors, and hence the smaller its ‘real’ output. This is a phenomenon that has been observed in economies with two- and especially three-digit inflation (cf. Kleiman, 1984 on the Israeli experience). Viewing the phenomenon in this way implicitly assumes that the services of the financial sector are not final products (which are a component of net national product) but ‘intermediate products’, whose function it is ‘to eliminate friction in the productive system’ and which accordingly are ‘not net contributions to ultimate consumption’ (Kuznets, 1951, p. 162; see also Kuznets, 1941, pp. 34–45).

See Also

- general equilibrium
- money illusion
- quantity theory of money
- real balances

Bibliography

Archibald, G.C. and Lipsey, R.G. 1958. Monetary and value theory: a critique of Lange and Patinkin. *Review of Economic Studies* 28, 50–6.

Bailey, M.J. 1956. The welfare cost of inflationary finance. *Journal of Political Economy* 64, 93–110.

Barro, R.J. 1976. Rational expectations and the role of monetary policy. *Journal of Monetary Economics* 2, 1–32.

Barro, R.J. 1978. Unanticipated money, output, and the price level in the United States. *Journal of Political Economy* 86, 549–80.

Birati, A. and Cukierman, A. 1979. The redistributive effects of inflation and of the introduction of a real tax system in the US bond market. *Journal of Public Economics* 12, 125–39.

Boschen, J.F. and Grossman, H.I. 1982. Tests of equilibrium macroeconomics using contemporaneous monetary data. *Journal of Monetary Economics* 10, 309–33.

Brock, W.A. 1974. Money and growth: the case of long run perfect foresight. *International Economic Review* 15, 750–77.

Cecchetti, S.G. 1986. Testing short-run neutrality. *Journal of Monetary Economics* 17, 409–23.

Feaveryear, A. 1963. *The Pound Sterling: A History of English Money*. 2nd edn, revised by E.V. Morgan. Oxford: Clarendon Press.

Feldstein, M. 1982. Inflation, capital taxation, and monetary policy. In *Inflation: Causes and Effects*, ed. R.E. Hall. Chicago: University of Chicago Press.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85, 191–205.

Fischer, S. 1980. On activist monetary policy with rational expectations. In *Rational Expectations and Economic Policy*, ed. S. Fischer. Chicago: University of Chicago Press.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Fisher, I. 1913. *The Purchasing Power of Money: Its Determination and Relation to Credit Interest and Crises*. Rev. edn. New York: Macmillan. Reprinted, New York: Augustus M. Kelley, 1963.

Fisher, I. 1923. The business cycle largely a ‘Dance of the Dollar’. *Journal of the American Statistical Association* 18, 1024–8.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan. Reprinted, New York: Kelley and Millman, 1954.

Friedman, M. 1970. *The Counter-Revolution in Monetary Theory*. London: Institute of Economic Affairs.

Gale, D. 1982. *Money: in Equilibrium*. Cambridge: Cambridge University Press.

Gordon, R.J. 1982. Price inertia and policy ineffectiveness in the United States, 1890–1980. *Journal of Political Economy* 90, 1087–117.

Grandmont, J.-M. 1983. *Money and Value: a Reconsideration of Classical and Neoclassical Monetary Theories*. New York: Cambridge University Press.

Gurley, J.G. and Shaw, E.S. 1960. *Money in a Theory of Finance*. Washington, DC: Brookings Institution.

Hayek, F.A. 1935. *Prices and Production*. 2nd edn. London: Routledge and Kegan Paul.

Howitt, P. and Patinkin, D. 1980. Utility function transformations and money illusion: comments. *American Economic Review* 70, 819–22, 826–8.

Hume, D. 1752. ‘Of money’, ‘Of interest’ and ‘Of the balance of trade’. As reprinted in D. Hume, *Writings on Economics*, ed. E. Rotwein, Wisconsin: University of Wisconsin Press, 1970.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London: Macmillan.

Kleiman, E. 1984. Alut ha-inflatzya [The costs of inflation]. *Rivon Le-kalkalah* [Economic Quarterly] 30, 859–64.

Kuznets, S. 1941. *National Income and its Composition, 1919–1938*. New York: National Bureau of Economic Research.

Kuznets, S. 1951. National income and industrial structure. *Proceedings of the International Statistical Conferences 1947* 5, 205–39. As reprinted in S. Kuznets, *Economic Change*, London: William Heinemann, 1954.

Leontief, W. 1936. The fundamental assumption of Mr Keynes' monetary theory of unemployment. *Quarterly Journal of Economics* 51, 192–7.

Levhari, D. and Patinkin, D. 1968. The role of money in a simple growth model. *American Economic Review* 58, 713–53. As reprinted in Patinkin (1972c), 205–42.

Lothian, J.R. 1985. Equilibrium relationships between money and other economic variables. *American Economic Review* 75, 828–35.

Lucas, R.E., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24. As reprinted in Lucas (1981), 66–89.

Lucas, R.E., Jr. 1975. An equilibrium model of the business cycle. *Journal of Political Economy* 83, 1113–44. As reprinted in Lucas (1981), 179–214.

Lucas, R.E., Jr. 1980. Two illustrations of the quantity theory of money. *American Economic Review* 70, 1005–14.

Lucas, R.E., Jr. 1981. *Studies in Business Cycle Theory*. Cambridge, MA: MIT Press.

McCallum, B.T. 1980. Rational expectations and macroeconomic stabilization policy: an overview. *Journal of Money, Credit, and Banking* 12, 716–46.

Metzler, L.A. 1951. Wealth, saving and the rate of interest. *Journal of Political Economy* 59, 93–116.

Mishkin, F.S. 1982. Does anticipated monetary policy matter? An econometric investigation. *Journal of Political Economy* 90, 22–51.

Mishkin, F.S. 1983. *A Rational Expectations Approach to Macroeconometrics*. Chicago: University of Chicago Press.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12, 45–88. As reprinted in American Economic Association, *Readings in Monetary Theory*, Philadelphia: Blakiston for the American Economic Association, 1951.

Mundell, R.A. 1963. Inflation and real interest. *Journal of Political Economy* 71, 280–3.

Mundell, R.A. 1965. A fallacy in the interpretation of macroeconomic equlibrium. *Journal of Political Economy* 73, 61–6.

Patinkin, D. 1965. *Money, Interest, and Prices*. 2nd edn. New York: Harper & Row.

Patinkin, D. 1972a. On the short-run non-neutrality of money in the quantity theory. *Banca Nazionale del Lavoro Quarterly Review* 100, 3–22.

Patinkin, D. 1972b. Money and growth in a Keynesian full-employment model. In Patinkin (1972c).

Patinkin, D. 1972c. *Studies in Monetary Economics*. New York: Harper & Row.

Patinkin, D. and Steiger, O. 1988. On the terms ‘neutrality of money’ and ‘veil of money’. *Scandinavian Journal of Economics* 90.

Robertson, D.H. 1922. *Money*. Cambridge: Cambridge University Press.

Sargent, T.J. 1976. A classical macroeconometric model for the United States. *Journal of Political Economy* 84, 207–37.

Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review* 57, 534–44.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Tobin, J. 1965. Money and economic growth. *Econometrica* 33, 671–84.

Tobin, J. 1967. The neutrality of money in growth models: a comment. *Economica* 34, 69–72.

How to cite this article

Patinkin, Don. "neutrality of money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000054>
doi:10.1057/9780230226203.1179

The New Palgrave Dictionary of Economics Online

new classical macroeconomics

Stanley Fischer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The new classical macroeconomics (NCM) attempts to build macroeconomics entirely on the foundations of market clearing and optimization by economic agents. It is also known as the rational expectations–equilibrium approach to macroeconomics. The leading figures are Robert Lucas of the University of Chicago and Thomas Sargent of the University of Minnesota, whose 1981 volume contains many of the formative contributions. Lucas (1977) and Sargent (1982) provide non-technical accounts of the approach. Other leading figures include Edward Prescott and Neil Wallace of the University of Minnesota and Robert Barro of the University of Rochester.

IThe monetary approach: the Lucas supply function

The new classical macroeconomics can be dated from work by Robert Lucas in the early 1970s. The article with greatest popular impact is Lucas's (1973) ‘Some International Evidence on Output–Inflation Tradeoffs’. This is a market-clearing model from which the Phillips curve emerges as a result of imperfect information about the aggregate price level. (Lucas (1972) is a more difficult article that produces a similar result.) The nature of the approach is clarified by outlining the Lucas model and by contrasting it with other models of the Phillips curve.

Markets are physically separated. There are two types of disturbance in the economy, aggregate disturbances that move the aggregate price level and relative disturbances that affect price in each market, but by definition average zero across all markets. Knowledge about past events and the probability distributions of disturbances is complete, but suppliers and demanders within each market observe only the nominal price in that market in the current period in which they have to make their output and purchase decisions.

In a full information set-up, supply and demand in an individual market would depend on relative price. Participants in the market know the price in that market, but cannot calculate relative price without an estimate of the aggregate price level. The optimal estimate of the aggregate price level, conditioned on the observed price in the market, is a weighted average of the expected aggregate price level and the absolute price observed in the market.

Estimated relative price in each market thus increases with the absolute price in that market *relative to the expected aggregate price level*. Aggregating across all markets, aggregate output is an increasing function of the absolute price level relative to the expected price level. This is the famous Lucas supply function

$$Y_t = \alpha(P_t - {}_{t-1}P_t)$$

where Y is aggregate output or its logarithm, P is the logarithm of the aggregate price level, and ${}_{t-1}P_t$ is the expectation of P_t based on information available at the end of period $(t - 1)$. The model is closed by assuming that aggregate demand is determined by the quantity equation.

The Lucas model contains a Phillips curve in the sense that output and the price level (relative to the expected price level) are positively correlated. If the price level followed a random walk, the standard Phillips curve relationship between output and the inflation rate would be observed in the data.

What NCM is not

The Lucas supply function illustrates the difference between NCM and alternative approaches. The original Phillips–Lipsey approach views the Phillips curve as a reflection of *dis*equilibrium in the labour market, with the wage adjusting to the excess demand for labour according to ‘the law of supply and demand’. Such an assumption is regarded as unsatisfactory by NCM because the existence of labour market disequilibrium (or disequilibrium anywhere) implies a failure to exploit mutually beneficial trades. NCM would rule out models with that feature – such as Keynesian models with unemployment – unless the failure to trade is explained within the model.

Despite many shared policy positions, the new approach also differs radically from monetarism. While the Lucas supply function is closely related to the Phillips curve model in Friedman's Presidential Address (1968), Friedman assumed that expectations were adaptive and that the monetary authority by accelerating inflation could keep the unemployment rate below the natural rate. The rational expectations assumption distinguishes NCM from monetarism. It is clear from a reading of Friedman and Schwartz (1963) that monetarists are more willing than the NCM to entertain the possibility of disequilibrium and slow adjustment of expectations. Indeed, from the perspective of NCM, monetarism and Keynesianism are of a piece – and equally unsatisfactory – in their willingness to use rules of thumb and crude empirical relationships to model economic behaviour, and in their willingness to proceed on macroeconomic issues in models without firm microfoundations.

Rational expectations is necessary but not sufficient for NCM. Many economists who do not assume that markets clear do assume that expectations are rational.

Policy ineffectiveness

The Lucas supply function has two important implications that are central to the new classical

macroeconomics: the *policy ineffectiveness result*, to be taken up now, and the *econometric policy evaluation critique*, examined later.

The policy ineffectiveness result is that any *anticipated* monetary policy action will not affect output. Rather, such actions are reflected in both the expected and the actual price levels, leading to no effect on output. The result, contained in Lucas (1973) but made most explicit in Sargent and Wallace (1975), is that monetary policy actions affect output only if they are unanticipated – meaning not reflected in pricing decisions. The result has been misinterpreted as applying to all macroeconomic policy, but would not apply to any *real* policy action: for instance an anticipated increase in the investment tax credit would certainly affect investment and typically also aggregate output. The ineffectiveness result relates only to monetary policy, in a model in which money is neutral except for its Phillips curve effects. That is, the Lucas supply curve produces a tradeoff between inflation and unemployment that is not systematically exploitable by policy makers.

The monetary policy ineffectiveness result has been the subject of much controversy. Models in which the monetary policy makers can respond to events after prices have been set leave open the possibility that systematic monetary policy can have real effects. Long-term labour contracts (as in Fischer, 1977, or Taylor, 1980) may be a source of effective monetary policy. Barro (1977) pointed out that the assumed form of contracts in Fischer was not optimal in that output decisions were left to the firm rather than being set as part of the contract. In practice, output decisions are made by firms; subsequent microeconomic research has shown that asymmetric information may generate that feature of contracts (Hart and Holmstrom, 1987) though it remains difficult to account for the failure of contracts to index for nominal disturbances.

Much of the controversy over the effectiveness of monetary policy derives from an implicit view that the aims of the government and the private sector differ. Stabilizing monetary policy may have a useful role to play if contracts cannot fully describe future contingencies, and if there are costs of frequent renegotiation. By creating a stable macro-economic environment, active monetary policy can encourage long-term contracting even when not all states of nature can be described – but it thereby also increases the damage that can be done by inappropriate policy (Fischer, 1980).

Early success

The NCM derived early success from empirical work by Barro (1978) that appeared to support the implication of the Lucas supply function that only unanticipated changes in the money stock had real effects. However, this implication of the NCM approach is shared by sticky wage theories, such as Fischer (1977), and turns out not to distinguish the NCM from other approaches. Further, empirical work by Mishkin (1983) shows that the result that only unanticipated money matters is not robust to lag length.

Within the NCM school, three sets of empirical results led to a loss of confidence in the Lucas supply function approach and the view that monetary shocks affect output. First, Barro (1978) found that although output was closely related to unanticipated changes in the money stock, the aggregate price level was not. This raised doubts about the Lucas supply function, in which prices are the transmission mechanism through which unanticipated money induces suppliers to increase output. Second, Barro and Hercowitz (1980) and Boschen and Grossman (1982) find that currently perceived changes in the money

stock, as reflected in preliminary money stock data, do affect output. Since the theory is built on the assumption that money has real effects only because it is not known, this result was a serious blow to the view that the Phillips curve is a result of imperfect information about current nominal variables. Third, Sims (1980) found in a vector autoregressive system including output, money and interest rates that interest rate shocks accounted for a far larger share of variations in output than money shocks.

IIEconometric implications

The rational expectations assumption used by NCM has led to the development of major new econometric methods for the treatment of expectations. Much of the econometric development is contained in Lucas and Sargent (1981). One focus has been on methods of testing the typical rational expectations *cross equation constraints*. These are restrictions on relations between parameters in different equations that follow from the assumption that expectations are optimal predictors of variables accounted for elsewhere in the model. A second focus is the econometric policy evaluation critique.

Econometric policy evaluation

In deriving the supply function, Lucas shows that the parameter α, the slope of the Phillips curve, is a decreasing function of the variance of the absolute price level. That is because it is a mixture of the structural supply elasticity in an individual market and the signal extraction problem solved by the supplier in deciding how much to respond to any observed nominal price in her market.

The implication is that parameters of macroeconomic models that appear structural, such as α, the slope of the Phillips curve, may not be invariant to changes in policy. In this case a reduction in the variance of the money supply, which is a policy parameter, will make the Phillips curve steeper.

The implication that parameters may not be invariant to changes in policy is the central point of Lucas's influential *econometric policy evaluation critique*, which has had a profound effect on both policy modelling and econometric practice in general (Lucas, 1976). On policy modelling, the argument is that existing econometric models, almost all of which are large-scale versions of textbook IS–LM models with an aggregate supply sector appended, cannot be used for analysing changes in policy, since the parameters in those models would likely change as policy changes. Lucas (1976) concedes that existing econometric models, some of which are commercially successful, may do a good job of forecasting. Nor does he argue that econometric models cannot ever be used for policy evaluation, since the true structural parameters (in the Phillips curve example the micro supply elasticity in an individual market) could in principle sometimes be identified. However in practice identification would be almost impossible for many parameters unless there had been frequent changes in policy ‘regimes’, or policy rules, that would produce variation in parameters such as the variance of the aggregate price level that affect responses to price signals.

The effect of the Lucas critique on econometric practice arises from a pervasive fear that parameters that had previously been thought structural and that were routinely estimated in empirical macroeconomics, such as the propensity to consume out of wealth, or the interest elasticity of money demand, are not invariant to economic policy. Few practising macroeconomists estimate a demand function for money or consumption function without making a pro forma bow in the direction of the Lucas critique – and those

who do not are reminded of the protocol by their discussants.
The influence of the Lucas critique is remarkable in that parameter instability induced by policy changes has not been shown to have been empirically important in whatever failures macroeconometric models have suffered. Nonetheless, the critique has led to a new empirical research agenda in macroeconomics.

Deep structural parameters

The argument is that the only truly structural parameters in the economy are tastes and technology, utility and production functions. Technology is to be widely interpreted as including the transactions technology and mechanisms for intertemporal trade. Once these primitives are known, it becomes possible to deduce how consumers and producers will respond to policy actions, whose only significance is in how they modify the constraints facing economic agents. Sargent (1982) presents an eloquent account of the research agenda.
The new approach has been to estimate parameters of utility and production functions from first order conditions rather than to attempt to estimate structural relations. In intertemporal optimization first order conditions are Euler equations. For instance in the life cycle consumption model with one consumption good and intertemporally and contemporaneously separable utility function, the discrete time Euler equation is:

$$U'(C_1) = \beta E_t\left[(1 + r_{t+1})U'(C_{t+1})\right]$$

where $\beta < 1$ is the discount factor, r is the (perhaps stochastic) rate of return on any asset, and E_t is the expectation conditional on information available in period t.
Aggregate and cross section data can be used to estimate such equations. Hall and Mishkin (1982) on panel data and Hansen and Singleton (1983) are examples. The purpose may be both to estimate utility function parameters and to test restrictions imposed by the underlying model of consumer optimization. Hall and Mishkin for instance conclude that 20 per cent of consumption is accounted for by consumers who are not satisfying the first order condition with equality, and that such consumers may be liquidity constrained. Mankiw, Rotemberg and Summers (1985) attempt using aggregate time series data to estimate parameters of utility functions defined over consumption and leisure. Examples of estimates of technological relations include Sargent (1978) on the demand for labour and Blanchard (1983) on inventory demand. Garber and King (1983) have severely criticized the Euler equation approach on the grounds that the identification problem has not been faced squarely.

IIIReal business cycles

The apparent failure of the Lucas supply function to account for the correlation between inflation and output as a result of imperfect information has led to the alternative real business cycle approach. In this

view, business cycles are equilibrium real phenomena, driven largely by productivity shocks. Endogeneity of the money stock accounts for the inflation- or money-output link.

The most fully worked out real business cycle model is that of Kydland and Prescott (1982). There is a representative agent, an infinite horizon intertemporal maximizer. Production inputs are labour, capital and inventories. The economy is hit by imperfectly observed productivity shocks, which are a mixture of permanent and transitory components. Slow acquisition of information about past shocks is one source of lags in the economy; another is lags in the process by which investment turns into capital. Kydland and Prescott can find parameter values, including the variance of the productivity shocks, that enable them to broadly match the stochastic processes that characterize United States business cycles.

The Kydland–Prescott paper has to deal with a basic problem in the NCM approach, that of the cyclical patterns of wages and leisure.

Intertemporal substitution of leisure

All theories of the business cycle have to account for relatively large movements in labour input accompanied by only small changes in real wages. If disequilibrium is disallowed, then the problem is to explain labour's willingness to supply, say, five per cent more labour in booms than in slumps for real wages that may be only one per cent higher. The obvious explanation, if the real wage is in fact procyclical, is that labour supply is very responsive to the wage. If this hypothesis explains business cycle correlations, it remains to reconcile short- and long-run labour supply behaviour, for in the long run labour supply curves may be backward bending.

The theoretical explanation comes from the distinction between responses to transitory and permanent increases in the real wage (Lucas, 1977). Workers may respond significantly to a transitory increase in the real wage, choosing to work harder now and substitute future for current leisure when the cost of leisure returns to normal. The intertemporal substitution of leisure mechanism plays an extremely significant role in NCM, for at a deeper level it is the rationale for the Lucas supply function.

Direct evidence in support of this hypothesis has been difficult to find (Altonji, 1982). Indeed there is some evidence that the real wage follows a random walk, which means that real wage changes are permanent. Unless transitory wage changes are identifiable at a local level, this result rules out the intertemporal substitution of leisure explanation of large movements of labour input over the cycle.

Alternative explanations may be available in which the observed wage does not measure the marginal utility of leisure because long-term arrangements between firms produce efficient allocations of resources without using the wage for short-term allocative purposes. Hart and Holmstrom (1987) present several models of contracts in which the wage is not equal to the marginal utility of leisure.

Leisure and consumption over the cycle

It is well known that an intertemporally separable utility function in which both consumption and leisure are normal goods implies that consumption and leisure should be positively correlated unless their relative price (the real wage) changes. In fact, measured consumption and leisure move in opposite directions over the cycle. The correlation cannot be explained in the typical model without significant movements in the real wage, which do not occur. Mankiw, Rotemberg and Summers' (1985) empirical

work documents this difficulty.
Kydland and Prescott account for cyclical patterns of leisure and goods consumption by, first, making productivity shocks the driving force in the cycle, and second, by assuming that past levels of leisure affect the current marginal utility of leisure.

Endogenous money

The real business cycle approach accounts for the Phillips curve by assuming that the money stock accommodates itself to the level of economic activity (King and Plosser, 1984). This view derives some support from the fact that the correlation with output is closer for inside than for outside money. Ironically the real business cycle and early Keynesian views of the unimportance of money are close, despite the dissimilarities of the analytic approaches.

IVPolicy analysis

The game-theoretic view of the operation of economic policy implicit in the policy ineffectiveness result has become extremely influential in the wake of the important paper on dynamic inconsistency by Kydland and Prescott (1977). Dynamic inconsistency occurs when a future policy decision that forms part of an optimal plan formulated at an initial date is no longer optimal from the viewpoint of a later date, even though no new information has appeared in the meantime.
The problem is likely to arise when expectations of future policy affect current decisions. For instance, to produce low rates of wage change, policymakers would like it believed that future policy will not accommodate wage increases. However, if wage increases occur, policy may well accommodate them rather than cause unemployment.
Kydland and Prescott view dynamic inconsistency as a major argument for the use of policy rules rather than discretion. Dynamic inconsistency will not occur if policy rules are set out and adhered to. Subsequent developments have analysed the tradeoff between the gains from flexibility produced by discretion and the losses due to dynamic inconsistency (e.g. Rogoff, 1985). It is also possible that a rational concern for reputation by policy makers will produce consistent behaviour (Barro and Gordon, 1983).
The game theory approach implies a stress on the credibility of policy makers, leading for instance to the view that a credible change in monetary policy could lead to a costless disinflation. This view was expressed in the United States before the disinflation of the early Eighties; the subsequent recessionary disinflation helped reduce support for the NCM. Although the game theory approach is not inherently related to NCM, in that expectations of future policy may matter in models without market clearing, it has in practice been pursued largely in an NCM context.

VSummary

The promise of the original Lucas NCM model that an imperfect information market clearing approach to macroeconomics could satisfactorily account for most business cycle phenomena including the Phillips curve has not been fulfilled. Beyond its difficulty in accounting for the apparent real effects of

monetary policy, the theory is not good at explaining unemployment in a market-clearing context. The NCM approach builds on the joint assumptions of market-clearing and optimizing behaviour. The market-clearing hypothesis is unlikely to persist as an analytic axiom, unless it is redefined to the point of being meaningless. But the assumption of maximizing behaviour within a specified environment is the microeconomic ideal to which economists aspire. That component of NCM will surely remain as a major impulse in macroeconomics. So too will the rational expectations assumption and the econometrics associated with that approach.

See Also

- business cycle measurement
- IS–LM
- natural rate of unemployment
- neoclassical synthesis
- rational expectations

Bibliography

Altonji, J.G. 1982. The intertemporal substitution model of labour market fluctuations: an empirical analysis. *Review of Economic Studies* 49(5), Special Issue, 783–824.

Barro, R.J. 1977. Long-term contracting, sticky prices, and monetary policy. *Journal of Monetary Economics* 3(3), July, 305–16.

Barro, R.J. 1978. Unanticipated money, output, and the price level in the United States. *Journal of Political Economy* 86(4), August, 549–80.

Barro, R.J. and Gordon, D. 1983. Rules, discretion, and reputation. *Journal of Monetary Economics* 12 (1), July, 101–21.

Barro, R.J. and Hercowitz, Z. 1980. Money stock revisions and unanticipated money growth. *Journal of Monetary Economics* 6(2), April, 257–67.

Blanchard, O.J. 1983. The production and inventory behaviour of the American automobile industry. *Journal of Political Economy* 91(3), June, 365–400.

Boschen, J.F. and Grossman, H.I. 1982. Tests of equilibrium macroeconomics using contemporaneous monetary data. *Journal of Monetary Economics* 10(3), November, 309–33.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85(1), February, 191–205.

Fischer, S. 1980. On activist monetary policy with rational expectations. In *Rational Expectations and Economic Policy*, ed. S. Fischer. Chicago: University of Chicago Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1), March, 1–17.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States*. Princeton: Princeton University Press.

Garber, P.M. and King, R.G. 1983. Deep structural excavation. University of Rochester, Department of Economics, Working Paper 83–14, September.

Hall, R.E. and Mishkin, F.S. 1982. The sensitivity of consumption to transitory income – estimates from panel data on households. *Econometrica* 50(2), March, 461–81.

Hansen, L.P. and Singleton, K. 1983. Stochastic consumption, risk aversion, and the temporal behaviour of asset returns. *Journal of Political Economy* 91(2), April, 249–65.

Hart, O. and Holmstrom, B. 1987. The theory of contracts. In *Advances in Economic Theory, 5th World Congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

King, R.G. and Plosser, C.I. 1984. Money, credit, and prices in a real business cycle model. *American Economic Review* 74(3), June, 363–80.

Kydland, F.E. and Prescott, E.C. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85(3), June, 473–93.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50(6), November, 1345–70.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4(2), April, 103–24.

Lucas, R.E. 1973. Some international evidence on output–inflation tradeoffs. *American Economic Review* 63(3), September, 326–34.

Lucas, R.E. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A.H. Meltzer, Carnegie-Rochester Conference Series on Public Policy, Vol. 1, Amsterdam: North-Holland.

Lucas, R.E. 1977. Understanding business cycles. In *Stabilization of the Domestic and International*

Economy, ed. K. Brunner and A.H. Meltzer, Carnegie-Rochester Conference Series on Public Policy, Vol. 5, Amsterdam: North-Holland.

Lucas, R.E. and Sargent, T.J. 1981. *Rational Expectations and Econometric Practice*. Minnesota: University of Minnesota Press.

Mankiw, N.G., Rotemberg, J.J. and Summers, J.H. 1985. Intertemporal substitution in macroeconomics. *Quarterly Journal of Economics* 100(1), February, 225–51.

Mishkin, F.S. 1983. *A Rational Expectations Approach to Macroeconometrics*. Chicago: University of Chicago Press.

Rogoff, K. 1985. The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100(4), November, 1169–90.

Sargent, T.J. 1978. Estimation of dynamic labour demand schedules under rational expectations. *Journal of Political Economy* 86(6), December, 1009–44.

Sargent, T.J. 1982. Beyond demand and supply curves in macroeconomics. *American Economic Review, Papers and Proceedings* 72(2), May, 382–9.

Sargent, T.J. and Wallace, N. 1975. ‘Rational’ expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83(2), April, 241–54.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48(1), January, 1–48.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88(1), February, 1–23.

How to cite this article

Fischer, Stanley. "new classical macroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000056> doi:10.1057/9780230226203.1180

The New Palgrave Dictionary of Economics Online

New Deal

Price V. Fishback

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

US President Franklin Roosevelt's New Deal created the most dramatic peacetime expansion of government in American economic history. It established the basic structures for modern federal/state social welfare programmes, farm programmes, labour policies, regulations of many industries, and government insurance of deposits and mortgages. Roosevelt experimented with a cartel-like industrial policy that was declared unconstitutional by the Supreme Court. The emergency public works and relief programmes built a large number of roads, dams, and other public works, and employed millions of labourers. Recent studies suggest that the impact of the New Deal varied greatly by programme.

Keywords

banking crises; cartels; child labour; Civilian Conservation Corps (USA); farm programmes; Federal Reserve System; Friedman, M.; gold standard; Great Depression; Hoover, H.; internal migration; Keynes, J. M.; minimum wages; monetary policy; National Labor Relations Boards (USA); National Recovery Administration (USA); New Deal; pensions; protection; public works; real business cycles; Reconstruction Finance Corporation (USA); Roosevelt, F.D.; Schwartz, A.; Smoot–Hawley Tariff Act of 1930; Social Security in the United States; trade unions; unemployment insurance; Works Progress Administration (WPA)

Article

Franklin Roosevelt's New Deal created the most dramatic peacetime expansion of government in American economic history.

When Franklin D. Roosevelt became president in March 1933, real output had fallen 30 per cent from its 1929 peak and the unemployment rate exceeded 25 per cent. Within his first hundred days in office Roosevelt and the Democratic Congress established an incredible array of programmes, a virtual ‘alphabet soup’ of acronyms. More programmes were added under the First New Deal until 1935, when

the Supreme Court declared the National Recovery Administration's (NRA) codes of ‘fair’ competition for industry and the Agricultural Adjustment Administration (AAA) farm programme unconstitutional. A Second New Deal re-established the farm programme in the name of soil conservation, strengthened the role of unions in collective bargaining, and established the basic structure of most of America's current social insurance and public assistance programmes.

After Roosevelt took office, the federal government, often in conjunction with state and local governments, built a huge number of roads, dams, sanitation facilities, schools, public housing projects, and other public works. The federal government expanded regulation of banking, finance, labour, and a host of other markets, insured and refinanced housing loans, and made extensive loans to numerous private and public entities. In the decades following the 1930s, several waves of historians have provided narratives and interpretations of the New Deal and introductions to their work can be found in collections edited by Dubofksy (1992), Braeman, Bremner and Brody (1975), and Hamby (1969). The recent trends in New Deal studies include a series of studies by economists and economic historians (Fishback et al., 2007; Bordo, Goldin and White, 1998).

Searching for an overarching theme for the programmes is a daunting task. The doubling of annual federal spending between the Hoover (1929–32) and Roosevelt years tempts many to describe the New Deal as Keynesian expansionary policy. But the Roosevelt administration ran relatively small budget deficits, as federal tax collections also more than doubled. In a brief meeting and a letter to the *New York Times* Keynes had encouraged Roosevelt to follow an expansionary policy, but the levels of government spending and the small budget deficits pale in comparison with the fall in output to be counteracted (Barber, 1996; Brown, 1956; Peppers, 1973; Romer, 1992).

One goal appeared to have been to raise prices and wages, as the establishment of the NRA allowed each industry to establish cartel-like codes that stifled price and quality competition, labour policies promoted unionization and high wages, and farm policies offered price guarantees while cutting output. Ultimately, Roosevelt and his advisors were pragmatists faced with terrible economic problems of nearly every kind. They established agencies and programmes meant to try to solve nearly each and every one. At times the programmes operated at cross-purposes. Higher farm and industry prices worsened the plight of the unemployed and other consumers. The pressure to raise wages exacerbated the unemployment problem, and the NRA codes limited output growth. The administration made constant adjustments in policies, creating a climate of uncertainty about the regulatory environment that left businesses wary of making new investments (Higgs, 1997).

New Deal monetary, banking, and international policy

Building on the seminal work by Friedman and Schwartz (1963), many economists argue that monetary policy contributed significantly to the harsh decline in the economy between 1929 and 1933. The Federal Reserve took seriously its international responsibilities in maintaining the gold standard and thus failed to respond sufficiently to three major waves of bank failures in a timely fashion. Many states had begun declaring ‘holidays’ that closed state banks to stave off bank runs. Roosevelt took office in the midst of the third wave of failures and declared a Bank Holiday that closed all national banks. Two-thirds of the banks were declared sound and reopened within the week. The troubled banks were reorganized and the Reconstruction Finance Corporation (RFC) subscribed to their new stock issues,

reassuring the public about the solvency of the banking system (Smiley, 2002; Mason, 2001).

In 1933 Roosevelt also announced that the United States was leaving the gold standard, prohibited gold exports, and devalued the dollar to $35 per ounce of gold. In response, the United States received a substantial flow of gold that stimulated the money supply, and economic growth resumed. Japan, Britain, France and several other leading nations experienced similar resumptions of economic growth when they broke free of their ‘golden fetters’ (Eichengreen, 1992; Temin, 1989; Temin and Wigmore, 1990). Gold inflows continued for the rest of the 1930s as Europe moved towards war. By choosing not to offset the gold inflows, Roosevelt and the Federal Reserve allowed the money supply to expand (Romer, 1992). The Federal Reserve took a misstep, however, when it used its newly awarded control over reserve requirements to double them in three steps between 1935 and 1937. The goal was to prevent a potentially inflationary rise in lending by soaking up the substantial excess reserves that banks were holding at the time. The banks responded by increasing their reserves and keeping the same cushion because they did not trust the Federal Reserve to provide adequate liquidity if a bank run occurred. The money supply fell and contributed to a sharp rise in unemployment and drop in real GDP in 1937–8 (Friedman and Schwartz, 1963; Romer, 1992). There is some disagreement about the impact of the monetary policies. Real business cycle economists argue that monetary and investment changes played much smaller roles than productivity shocks and high-wage labour policies in accounting for the fluctuations during the 1930s (Chari, Kehoe and McGratton, 2005).

The decision to leave the gold standard was accompanied by efforts to expand world trade beginning in 1934 with the Reciprocal Trade Agreement Act (RTA). The Smoot–Hawley Tariff Act of 1930 had helped touch off a series of protectionist responses by other countries that had caused total imports for a group of 75 countries to fall to one-third of their 1929 level. The RTA freed the Roosevelt administration to sign a series of tariff reduction agreements with Canada, several South American countries, Britain and key European trading partners. Consequently, American imports rose from a 20-year low in 1932–3 to an all-time high by 1940 (Irwin, 1998; Kindleberger, 1986).

Meanwhile, the Banking (Glass–Steagall) Act of June 1933 enacted an additional set of banking policies. Despite the checkered history experienced by state deposit insurance programmes (Calomiris and White, 2000), the act created the Federal Deposit Insurance Corporation (FDIC) to insure commercial bank deposits of up to $10,000. Insurance for savings and loans followed within the year. The Banking Act also established regulations, eliminated in the late 1970s, that prevented commercial banks from investing more than ten per cent of their assets in stocks and paying interest on deposits (Regulation Q). To increase the capital available for housing loans, the Home Owners' Loan Corporation (HOLC) provided funds to refinance troubled mortgages between 1933 and 1936, and the Federal Housing Administration (FHA) began offering insurance of mortgages and home improvement loans. Both agencies aided in the spread of the modern long-term, amortized mortgage loan that replaced short-term loans in which repayment of only interest over the course of the loan was followed by a balloon payment of the principal when it fell due.

The Reconstruction Finance Corporation (RFC): New Deal lender

Established by President Herbert Hoover in 1932, the RFC was an off-budget government corporation that maintained control of the funds repaid on its earlier loans. The RFC offered the Roosevelt

administration flexibility because they could start funding programmes without constantly seeking new appropriations from Congress. In consequence, the RFC became the lender during the starting phase of nearly every major New Deal grant and lending programme. In addition, the RFC provided loans to large numbers of financial institutions of all types, railroads, farmers and local governments (Olson, 1998). The RFC loans to private business met with mixed success. The liquidity loans to failing banks in 1932 had not prevented many bankruptcies because the RFC loans were given first priority over depositors and other lenders in case of failure; therefore, banks were prevented from selling their most liquid assets to meet depositor demands for cash. The RFC's purchases of preferred stock in banks reorganized after the Bank Holiday of 1933 exposed the RFC funds to more risk but led to more success at preventing failures (Mason, 2001). RFC lending to railroads succeeded in preventing several railroad bankruptcies. However, the spared railroads continued to underinvest in maintenance and capital improvements. In contrast, railroads forced into bankruptcy had to make such investments to attract enough capital to reopen for business (Mason and Schiffman, 2004).

Emergency relief and public works programmes

Unprecedented unemployment rates ranging from 10 to 25 per cent through the 1930s were the New Deal's greatest challenge. Prior to the New Deal, aid to the poor and labour policies had been the purview of state and local governments. Claiming unemployment to be a national emergency, Roosevelt and Congress raised the federal share of relief spending as high as 79 per cent while nearly quadrupling relief spending even as unemployment rates fell by the mid-1930s. The Federal Emergency Relief Administration (FERA, 1933–5), the Civil Works Administration (CWA, winter of 1933–4), and the Works Progress Administration (WPA, 1935–42) offered work relief jobs to households whose incomes fell below a target budget for necessities. The Civilian Conservation Corps (CCC) offered conservation jobs in the nation's hinterlands to youths whose earnings were shared with their parents. The FERA also handed out direct relief until 1935, when the responsibility for ‘unemployables’ was returned to state and local governments, and the federal government began offering matching grants for public assistance for children, the blind, and the elderly.

Harry Hopkins, who headed the FERA, CWA and the WPA, preferred work relief because it ‘provided a man with something to do, put money in his pocket, and kept his self-respect’ (Adams, 1977, p. 53). To give people incentive to leave work relief for private jobs, WPA monthly earnings averaged 40 to 50 per cent of full-time private earnings, and the WPA assured people that they would be reaccepted should the private job end. Even so, a significant percentage of workers stayed on work relief jobs for periods as long as a year and in some cases several years (Margo, 1993).

Roughly one-fourth of New Deal grant spending went to the Public Works Administration (PWA), Public Buildings Administration (PBA), the Public Roads Administration (PRA), and the Tennessee Valley Authority (TVA). The planning stages on these large-scale projects were longer, the wages were higher, and there was more freedom to hire already employed workers. The relief and public works programmes grants were designed to provide employment, build public projects, and stimulate the economy.

At one level the relief and public works programmes were very successful. Millions of Americans obtained work relief jobs to tide them over, and most of the original public works, many renovated

since, are still in place today. To understand the true impact of the New Deal, areas with different amounts of spending need to be compared to get a sense of how their economies would have performed without the New Deal. Since the mid-1990s economists have been using the substantial variation in spending across local areas to make such comparisons while working to control for the feedbacks caused by administrators using New Deal programmes to respond to economic problems. At the local level the benefits of the projects were likely to be stronger when the general share of goods produced in the area for local consumption was higher, the projects hired the unemployed without crowding out private or state and local government employment, and expansions did not raise incomes enough to generate federal income tax payments.

Although cross-sectional studies show little effect of relief jobs on private employment, analysis of panel data can control for unmeasured factors using the information across time for a cross section of areas. The panel studies suggest that an additional relief job reduced private employment by up to half a job (Wallis and Benjamin, 1981; 1989; Fleck, 1999a). A new relief job also raised ‘measured’ unemployment by one person because many discouraged workers, who had been out of the labour force and thus not counted as unemployed, were defined as re-entering the labour force as unemployed workers when they accepted relief jobs (Darby, 1976; Fleck 1999a).

The impact of public works and relief programmes had more clearly beneficial effects on other measures of socio-economic welfare. Cross-sectional studies of US counties suggest that an added dollar of public works and relief spending per person raised per capita income by roughly 85 cents and stimulated in-migration (Fishback, Horrace and Kantor, 2005; 2006). Panel studies of more than 100 major cities between 1929 and 1940 show that increased relief spending stimulated birth rates, reduced property crime, and reduced infant deaths and deaths from suicide and several diseases. The relief costs per death prevented in today's dollars are within the range of modern market values of life, and the costs are lower than the costs per death prevented of many modern safety programmes (Fishback, Haines and Kantor, 2007; Johnson, Kantor and Fishback, 2006).

Farm programmes

To raise the incomes of farmers, who had struggled through over a decade of hard times, the New Deal established the structure of the modern US farm programmes. The Agricultural Adjustment Administration (AAA) paid farmers to take land out of production. In 1935 in *United States v. Butler* the Supreme Court struck down the output processing tax that had originally funded the payments. The AAA payments were quickly reinstituted (minus the processing tax) under the Soil Conservation and Domestic Allotment Act (1935). The Commodity Credit Corporation (CCC) insured that farmers were paid higher prices by making loans that could be repaid with the crop itself if market prices fell below a target price. The Farm Credit Administration (FCA) reorganized and expanded farm lending, ultimately becoming involved in more than half of all farm mortgages and a large share of production loans. Meanwhile, the Rural Electrification Administration (REA) provided subsidized loans to give farmers access to electricity, while the Farm Security Administration (FSA) developed programmes to aid low-income farmers.

Efforts to determine the AAA's impact on limiting farm output have been confounded because a series of major climatic disasters in the 1930s served to cut output anyway. There is evidence that farmers

stopped planting their least productive land and raised the inputs used on the remaining land. The AAA clearly aided large farmers but possibly at the expense of farm workers and tenants (Alston and Ferrie, 1999; Whatley, 1983). Cross-county studies show that increases in AAA payments in counties led to no increases in retail sales, were associated with higher infant mortality in the South, and stimulated net outmigration (Fishback, Horrace and Kantor, 2005; 2006; Fishback, Haines and Kantor, 2001; Alston and Ferrie, 1999; Whatley, 1983). On the positive side, the AAA soil conservation programmes encouraged a move to larger farms and practices that cut soil erosion, so that the Great Plains avoided a recurrence of the Dust Bowl when the same drought and wind conditions arose later (Hansen and Libecap, 2004).

The political economic geography of New Deal spending

New Deal grant spending across states and counties varied enormously, as some western states received several times more per head than some southern states. Roosevelt in a radio ‘fireside chat’ vowed that the New Deal would promote ‘Relief, Recovery, and Reform’. Critics argued that Roosevelt used the monies primarily to aid his re-election efforts. The distribution process for many programmes was opaque, so New Deal scholars have turned to econometric analysis that simultaneously tests the importance of the stated motives and presidential politicking. Politicking was clearly part of the process in the distribution of total funds and at the programme level. Nearly every study finds that more grants went to swing states and areas with higher political turnout, while some find rewards for loyal Democratic areas as well as districts represented by powerful congressmen. The Roosevelt administration was innovative in targeting radio owners in their push to win elections (Wright, 1974; Wallis, 1998; Fleck, 1999b; Stromberg, 2004; Couch and Shughart, 1998).

Winning elections required more than just manipulation of spending to hit specific political targets. The Roosevelt administration also enhanced its future re-election prospects by following its stated aims. Many studies find evidence that the Roosevelt administration promoted recovery and relief by spending more in areas with higher unemployment and larger declines in income from 1929 to 1933. Few find signs that the total spending was reform-oriented, but specific relief programmes did target areas with long-term poverty. State governments influenced the distribution by the intensity of their lobbying and their spending in matching grant programmes, while the presence of federal land in a state also drew substantial public works grants. Specific programmes typically followed stated goals. There were so many programmes that nearly everybody could find one that benefited them, ranging from relief for the unemployed and poor to loans and AAA grants for large farmers. The HOLC and FHA housing programmes benefited carefully vetted home owners who were perceived as having lower risk of default (Fishback, Wallis and Kantor, 2003). There were constant charges of corruption, but the WPA actively battled corruption at the state and local levels by establishing an internal investigative agency. When the federal government increased its control of the distribution of funds within states in the switch from the FERA to the WPA, the distribution of funds within states more closely mirrored the relief, recovery and reform goals (Wallis, Fishback and Kantor, 2006).

Industrial and labour policies

To combat ‘destructive competition’, low prices and low wages, the National Recovery Administration (NRA) was created to allow industries to establish their own codes for minimum prices, quality standards, trade practices, and labour relations (Bellush, 1975). The NRA appeared to be sponsoring a series of industry cartels, as large firms tended to dominate the code-writing process in most industries. Wholesale prices jumped 23 per cent in two years, although consumer prices were much slower to rise. Simulations of the economy with and without the NRA imply that it served to slow economic recovery (Cole and Ohanian, 2004). The internal problems of cartels were also present, as industries with diverse firms had trouble coming to agreement and a number of firms routinely violated the codes (Alexander and Libecap, 2000). The NRA ended in 1935 when the Supreme Court declared it unconstitutional in the Schechter Poultry case, and few mourned its passing.

The National Labor Relations (Wagner) Act of 1935 expanded the right of workers to collective bargaining through their own representatives beyond the protections originally offered in the 1933 act that created the NRA. Employers were required to bargain with unions when a majority of workers voted for union representation, and employer-sponsored unions were banned. The National Labor Relations Board (NLRB) was established to oversee union elections and the collective bargaining process. As a result, unionization expanded rapidly through a mixture of strikes and elections. In the long run the NLRB policies regularized the union recognition and bargaining process, and the incidence of violent strikes has diminished sharply since (Freeman, 1998).

The emphasis on raising wages continued when the Fair Labor Standards Act (FSLA) of 1938 set a national minimum wage, overtime requirements, and child labour restrictions. Workers in agriculture or not employed in interstate commerce were exempted. Congressional support for the act was centred in states outside the South with high-wage industries, more unionization, and more advocates for teenage workers. As a result, the first minimum wage was binding only for low-wage industries in the South, where employers in some southern industries responded by reducing employment, and others switched to labour-saving technologies or limited their business to intra-state commerce to avoid federal regulation (Seltzer, 1995, 1997; Fleck, 2004).

The Social Security Act of 1935

The legislative centerpiece of the Second New Deal was the Social Security Act (SSA) of 1935, which established the modern structure of public assistance and social insurance programmes. The public assistance grants set some federal guidelines and offered matching grants that gave the states latitude in setting benefits. The new Aid to Dependent Children (ADC), Aid to the Blind (AB), and Old-Age Assistance (OAA) programmes replaced similar state programmes in more than half of the states, and provided coverage for the first time in the remaining states.

State unemployment insurance programmes funded by employer contributions with administrative costs paid by the federal government were established as a long-term alternative to providing emergency work relief. The states retained control over benefits offered. Each designed its own experience-rating system that required employers who laid off more workers to pay higher premiums, a feature not commonly found in other countries' unemployment insurance systems. The experience rating helped reduce seasonal unemployment fluctuations (Baicker, Goldin and Katz, 1998).

Social security is most associated with the federal old-age retirement system. In the debates over social

security, Roosevelt pressed for an actuarially sound system where the individual's retirement benefits were based purely on his and his employer's own contributions. He was not convinced the old-age pensions were necessary and sought to ensure that future generations would not be saddled with the costs. Others pressed for a subsidized system that provided adequate payments to all who contributed. The plan adopted in 1935 was a hybrid, but the inadequacies of the hybrid system had become apparent by 1939, and the current pay-as-you-go structure was created. A worker and his employer pay taxes into an administrative trust fund that pays benefits to current retirees and serves as a commitment by the federal government to collect enough taxes to pay the worker his own social security pension when he reaches retirement age. The initial taxes were one per cent of wages each for workers and employers, and the initial benefits paid in 1940 were roughly 25 per cent of the average earnings of workers contributing to the system. Average pension payments are now roughly 40 per cent of the contributing workers' average earnings, and the increase in average lifespans has caused rapid increases in the ratio of retirees to workers. In consequence, the tax rates had risen to over 5.3 per cent each for worker and employer by 2000, with expectations that relative benefits will have to be cut or taxes raised in the future to sustain the system (Schieber and Shoven, 1999).

Conclusion

The New Deal was a response to the Great Depression, a major peacetime crisis sandwiched between two world wars. All three crises contributed to short-run rapid expansions of the federal government. When each ended, the government's role retracted somewhat but never to the level that would likely have occurred without the crisis (Higgs, 1987). In the span of six years the Roosevelt administration built an incredible array of public works and established a series of regulations, government insurance, and public assistance programmes that are still in place today. The New Deal arguably did more to expand the role of government in the United States than the more evolutionary changes that have occurred since the end of the Second World War.

See Also

- Great Depression

Bibliography

Adams, H.H. 1977. *Harry Hopkins: A Biography*. New York: G.P. Putnam Sons.

Alexander, B. and Libecap, G. 2000. The effect of cost heterogeneity in the success and failure of the New Deal's agricultural and industrial programs. *Explorations in Economic History* 37, 370–400.

Alston, L. and Ferrie, J. 1999. *Southern Paternalism and the American Welfare State*. New York: Cambridge University Press.

Baicker, K., Goldin, C. and Katz, L. 1998. A distinctive system: origins and impact of U.S. unemployment compensation. In *The Defining Moment: The Great Depression and the American Economy in the Twentieth Century*, ed. M. Bordo, C. Goldin, and E.N. White. Chicago: Chicago University Press.

Barber, W.J. 1996. *Designs within Disorder: Franklin D. Roosevelt, the Economists, and the Shaping of American Economic Policy, 1933–1945*. New York: Cambridge University Press.

Bellush, B. 1975. *The Failure of the NRA*. New York: Norton.

Bordo, M., Goldin, C. and White, E.N., eds. 1998. *The Defining Moment: The Great Depression and the American Economy in the Twentieth Century*. Chicago: University of Chicago Press.

Braeman, J., Bremner, R.H. and Brody, D., eds. 1975. *The New Deal*. Columbus: Ohio State University Press.

Brown, E.C. 1956. Fiscal policy in the ‘thirties: a reappraisal. *American Economic Review* 46, 857–79.

Calomiris, C. and White, E.N. 2000. The origins of federal deposit insurance. In *U.S. Bank Deregulation in Historical Perspective*, ed. C. Calomiris. New York: Cambridge University Press.

Chari, V.V., Kehoe, P. and McGrattan, E.R. 2005. Spectral methods in business cycle accounting. *Revista de Economia* 12, 5–18.

Cole, H. and Ohanian, L. 2004. New Deal policies and the persistence of the Great Depression: a general equilibrium analysis. *Journal of Political Economy* 112, 779–816.

Couch, J. and Shughart, W., III 1998. *The Political Economy of the New Deal*. New York: Edward Elgar.

Darby, M.R. 1976. Three and a half million US employees have been mislaid: or, an explanation of unemployment, 1934–1941. *Journal of Political Economy* 84, 1–16.

Dubofksy, M., ed. 1992. *The New Deal: Conflicting Interpretations and Shifting Perspectives*. New York: Garland Publishers.

Eichengreen, B. 1992. *Golden fetters: the gold standard and the depression, 1919–1939*. New York: Oxford University Press.

Fishback, P., Haines, H. and Kantor, S. 2001. The impact of the New Deal on black and white infant mortality in the South. *Explorations in Economic History* 38, 93–122.

Fishback, P., Haines, H. and Kantor, S. 2007. Births, deaths, and New Deal relief during the Great Depression. *Review of Economics and Statistics* 89, 1–14.

Fishback, P., Higgs, R., Libecap, G., Wallis, J.W., Engerman, S., Hummel, J., LaCroix, S., Margo, R., McGuire, R., Sylla, R., Alston, L., Ferrie, J., Guglielmo, M., Pasour, E.C., Rucker, R. and Troesken, W. 2007. *Government and the American Economy: A New History*. Chicago: University of Chicago Press.

Fishback, P., Horrace, W. and Kantor, S. 2005. The impact of New Deal expenditures on local economic activity: an examination of retail sales, 1929–1939. *Journal of Economic History* 65, 36–71.

Fishback, P., Horrace, W. and Kantor, S. 2006. Do federal programs affect internal migration? The impact of New Deal expenditures on mobility during the Great Depression. *Explorations in Economic History* 43, 179–222.

Fishback, P., Wallis, J.J. and Kantor, S. 2003. Can the New Deal's three R's be rehabilitated? A program-by-program, county-by-county analysis. *Explorations in Economic History* 40, 278–307.

Fleck, R. 1999a. The marginal effect of New Deal relief work on county-level unemployment statistics. *Journal of Economic History* 59, 659–87.

Fleck, R. 1999b. The value of the vote: a model and test of the effects of turnout on distributive policy. *Economic Inquiry* 37, 609–23.

Fleck, R. 2004. Democratic opposition to the Fair Labor Standards Act of 1938: reply to Seltzer. *Journal of Economic History* 62, 231–5.

Freeman, R. 1998. Spurts in union growth: defining moments and social processes. In *The Defining Moment: The Great Depression and the American Economy in the Twentieth Century*, ed. M. Bordo, C. Goldin and E.N. White. Chicago: University of Chicago Press.

Friedman, M. and Schwartz, A. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press.

Hamby, A., ed. 1969. *The New Deal: Analysis and Interpretation*. New York: Weybright and Talley.

Hansen, Z. and Libecap, G. 2004. Small farms, externalities, and the dust bowl of the 1930s. J*ournal of Political Economy* 112, 665–94.

Higgs, R. 1987. *Crisis and Leviathan: Critical Episodes in the Growth of American Government*. New York: Oxford University Press.

Higgs, R. 1997. Regime uncertainty: why the Great Depression lasted so long and why prosperity resumed after the war. *Independent Review* 1, 561–90.

Irwin, D. 1998. Changes in U.S. tariffs: the role of import prices and commercial policies. *American Economic Review* 88, 1015–26.

Johnson, R., Kantor, S. and Fishback, P. 2006. Striking at the roots of crime: the impact of social welfare spending on crime during the Great Depression. Working Paper No. 12825, Cambridge, MA: NBER.

Kindleberger, C. 1986. *The World in Depression, 1929–1939*. Rev. edn. Berkeley: University of California Press.

Margo, R. 1993. Employment and unemployment in the 1930s. *Journal of Economic Perspectives* 7(2), 41–59.

Mason, J. 2001. Do lenders of last resort policies matter? The effects of the reconstruction finance corporation assistance to banks during the Great Depression. *Journal of Financial Services Research* 20, 77–95.

Mason, J. and Schiffman, D. 2004. Too-big-to-fail, government bailouts, and managerial incentives: the case of Reconstruction Finance Corporation assistance to the railroad industry during the Great Depression. In *Too-Big-To Fail: Policies and Practices in Government Bailouts*, ed. B.E. Gup. Westport, CT: Greenwood Press.

Olson, J.S. 1998. *Saving Capitalism: The Reconstruction Finance Corporation and the New Deal*. Princeton: Princeton University Press.

Peppers, L. 1973. Full employment surplus analysis and structural change: the 1930s. *Explorations in Economic History* 10, 197–210.

Romer, C.D. 1992. What ended the Great Depression? *Journal of Economic History* 52, 757–84.

Schieber, S.J. and Shoven, J.B. 1999. *The Real Deal: The History and Future of Social Security*. New Haven: Yale University Press.

Seltzer, A.J. 1995. The political economy of the Fair Labor Standards Act. *Journal of Political Economy* 103, 1302–42.

Seltzer, A.J. 1997. The effects of the Fair Labor Standards Act of 1938 on the southern seamless hosiery and lumber industries. *Journal of Economic History* 57, 396–415.

Smiley, G. 2002. *Rethinking the Great Depression: A New View of Its Causes and Consequences*. Chicago: Ivan R. Dee.

Stromberg, D. 2004. Radio's impact on public spending. *Quarterly Journal of Economics* 119, 189–221.

Temin, P. 1989. *Lessons from the Great Depression*. Cambridge, MA: MIT Press.

Temin, P. and Wigmore, B. 1990. The end of one big deflation. *Explorations in Economic History* 27, 483–502.

Wallis, J.J. 1998. The political economy of New Deal spending revisited, again: with and without Nevada. *Explorations in Economic History* 35, 140–70.

Wallis, J.J. and Benjamin, D.K. 1981. Public relief and private employment in the Great Depression. *Journal of Economic History* 41, 97–102.

Wallis, J.J. and Benjamin, D.K. 1989. Private employment and public relief during the Great Depression. Working paper, Department of Economics, University of Maryland.

Wallis, J.J., Fishback, P. and Kantor, S. 2006. Politics, relief, and reform: Roosevelt's efforts to control corruption and manipulation during the New Deal. In *Corruption and Reform*, ed. E. Glaeser and C. Goldin. Chicago: University of Chicago Press.

Whatley, W.C. 1983. Labor for the picking: the New Deal in the South. *Journal of Economic History* 43, 905–29.

Wright, G. 1974. The political economy of New Deal spending: an econometric analysis. *Review of Economics and Statistics* 56, 30–8.

How to cite this article

Fishback, Price V. "New Deal." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000168> doi:10.1057/9780230226203.1181

The New Palgrave Dictionary of Economics Online

new economic geography

Anthony J. Venables
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

New economic geography provides an integrated and micro-founded approach to spatial economics. It emphasizes the role of clustering forces in generating an uneven distribution of economic activity and income across space. The approach has been applied to the economics of cities, the emergence of regional disparities, and the origins of international inequalities.

Keywords

clustering; comparative advantage; congestion; core–periphery mode; dispersion; factor price equalization; foreign direct investment; gravity modelling; imperfect markets; increasing returns to scale; industrial organization; international portfolio investment; international trade (theory); knowledge spillovers; labour mobility; linkages; location of economic activity; market access; Marshall, A.; monopolistic competition; new economic geography; productivity; regional development; spatial economics; urban agglomeration; urban economics

Article

Why is economic activity distributed unevenly across space, with centres of concentrated activity surrounded by ‘peripheral’ regions of lower density? What economic interactions are there between different geographical areas, and how do these shape income levels in the areas? How does the spatial organization of economic activity respond to exogenous shocks, such as technological change or policy measures? The contribution of ‘new economic geography’ (NEG) is to address these questions in a manner that is based on rigorous microeconomic foundations. It shows how the spatial structure of an economy is determined by the interplay between costs of transactions across space and various types of increasing returns to scale. The questions posed above can be addressed at different spatial levels – international, regional and urban. NEG provides a unified framework for analysis at these different levels.

Clustering versus dispersion

The NEG approach has several key analytical ingredients. The first is the recognition that spatial interactions are costly. These costs are shaped by geography and depend on the nature of the interaction. Thus, trade in goods incurs shipping costs and costs of time in transit, depending on distance shipped, on transport infrastructure and on geography. Communications and coordination costs mean that workers may be less effective if they are not in close proximity with co-workers. Factor mobility may be impeded by distance and geography. This approach contrasts with that of international trade theory, in which spatial units are identified solely with countries – jurisdictions rather than geography – and where goods and factors are typically assumed to either be traded freely or to be completely non-tradable. The NEG approach shows how outcomes depend on the extent to which different goods and activities are mobile between locations.
The second key ingredient is the possibility that there are clustering forces, inducing activity to concentrate in space. Clustering arises because of spatially concentrated increasing returns to scale which can derive from a number of different underlying forces. (The classic discussion is Marshall, 1890; for a recent survey see Duranton and Puga, 2004.) One

possibility is that there are public goods, the enjoyment of which depends on geographical access, such as a town centre. Another possibility is that there are positive technological externalities such as knowledge spillovers; firms produce ideas that can be observed and copied by other firms, depending on their proximity. These approaches have been prominent in much of the urban economics literature (for example, Henderson, 1988), but writers in the NEG literature have generally sought to derive clustering forces from spatial interactions in imperfect markets rather than to simply assume them through public goods or technological externalities.

One way to derive clustering forces is through thick market effects, particularly in the labour market. Dense labour markets may allow for better matching of the skills of workers and the requirements of firms (Helsley and Strange, 1990). Incentives to acquire skills may be greater where workers face more prospective employers (Matouschek and Robert-Nicoud, 2005). Another way in which to derive clustering is to use industrial organization models of imperfect competition. The route followed in much of the NEG literature is to suppose that an industry (we will call it ‘manufacturing’) contains a number of firms, each of which has increasing returns to scale. The presence of internal economies of scale means that firms are faced with a location choice (if they had constant or diminishing returns then, given transport costs and dispersed consumers, they would choose to produce a very small amount in all locations – ‘backyard capitalism’, Starrett, 1978). The questions are, then, where do firms choose to locate, and under what circumstances will they cluster together?

The model often used to analyse the choice is the Dixit and Stiglitz (1977) model of monopolistic competition and its international trade extensions (Krugman, 1980). In this model each firm has a distinct variety of product which it produces in a single location and exports to other locations, and entry and exit occur until profits are bid down to zero. It turns out that, as firms take location decisions in order to maximize profits, so their location pattern tends to amplify any underlying differences between locations, and from this it is possible to generate an outcome in which clustering occurs.

To understand the argument, suppose that there are two regions A and B, and that A has demand $k > 1$ times larger than B (we ignore factor supply considerations for the moment). Could there be an equilibrium in which firms are located in proportion to the size of the regions, so A has *k* times more manufacturing firms than B? If trade costs are prohibitively high the answer is ‘yes’; only local firms supply each market, and the number of firms is proportional to the size of the market. (Notice that this argument uses the Dixit–Stiglitz property that all firms are the same size in equilibrium.) But as trade costs are reduced and firms start to export, two things happen. First, the region B market comes to be supplied by *k* times as many importing firms as does the country A market, thus reducing the profitability of producers in B. Second, each firm in B will pay transport costs on a large part of their output (sales to the large country A market) while firms in A will pay transport costs only on a smaller fraction of their output (sales to the smaller region B market). Both arguments suggest that firms in A become relatively more profitable, implying that in equilibrium with free entry the number of firms in A must exceed the number in B by a factor greater than *k*. The large region therefore has a disproportionately large share of manufacturing production, and is a net exporter of manufactures and importer of agriculture. More generally, a region with good ‘market access’ will attract a high share of firms.

This argument holds only if transport costs lie strictly between zero and a prohibitive level. If transport costs are prohibitive no firms ship any exports; each region is self-sufficient, and the location of industry is in proportion to the size of the regions. Conversely, if transport costs are zero, then the argument collapses, as firms in all regions have equally good access to all markets. The argument shows that it is at intermediate levels of transport costs that market access matters, and manufacturing is pulled disproportionately into the large region.

While this argument creates an incentive for clustering of firms, it is balanced by dispersion forces. These could be due to negative externalities, such as congestion, or arise as a consequence of immobility of some factors of production. Which factors are immobile depend on context, but typically include land (as in the tradition of urban economic modelling) and some or all types of labour. Thus, if labour were immobile, any benefit that firms derived from locating in one region rather than another would create a regional wage differential, until profits (more generally, the return to mobile activities) were equalized across regions.

Labour mobility is central to the Krugman (1991) ‘core–periphery’ model. This analyses two regions and two sectors, a constant returns to scale agriculture and manufacturing modelled as outlined above. Each sector uses a sector-specific type of labour (‘peasants’ and manufacturing workers respectively), and the regions’ endowments of these factors are, *ex ante*, identical. Crucially, manufacturing workers are mobile between the locations, whereas peasants are immobile. What is the division of manufacturing workers and firms between the two locations? Outcomes, as a function of trade costs, are illustrated on Figure 1. When trade costs are high manufacturing is equally divided between regions. However, when trade costs are low enough, manufacturing (and all manufacturing workers) concentrate entirely in one region or the other. There are two mutually reinforcing arguments supporting this clustering. The concentration of manufacturing workers creates a large market, so making the location profitable for firms. And the entry of firms bids up wages, so making the location attractive for workers (this effect reinforced by the fact that workers also benefit from not having to pay trade costs on their consumption of manufactures). It is not profitable for any single firm to leave the cluster, because the benefit of lower wages is outweighed by the loss of market access. As Figure 1 makes clear, the switch from dispersed manufacturing to agglomeration arises discontinuously. There is a critical value of trade costs, t^*, above which dispersed production is the stable equilibrium, and below which dispersed activity is unstable, while clustering of activity, in either of the regions, is a stable equilibrium.

Figure 1

Location of manufacturing in two regions

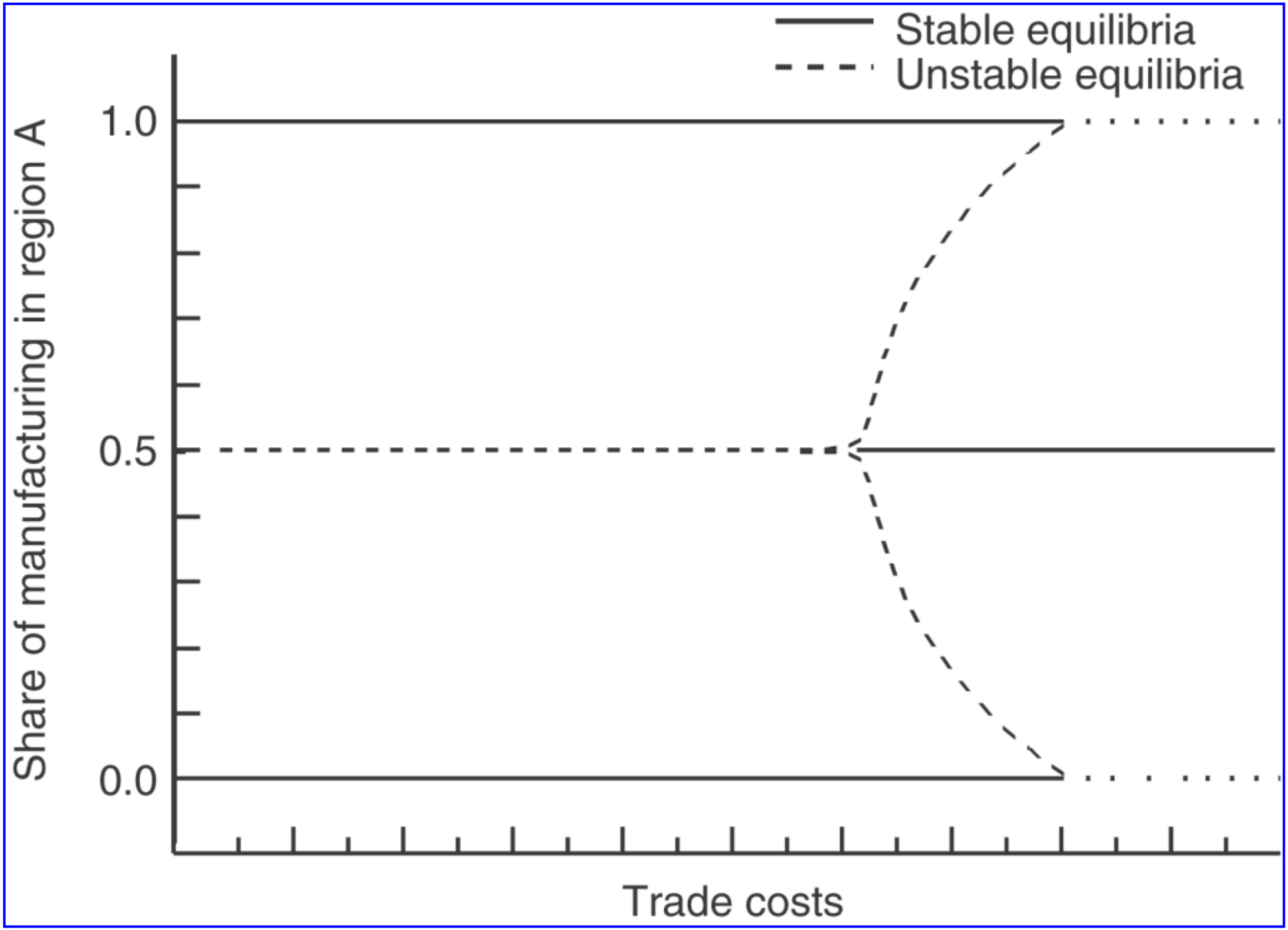

Krugman's ‘core–periphery’ model is perhaps the seminal paper, and brings the insight that agglomeration forces can be derived from a standard model of trade and monopolistic competition (see Fujita, Krugman and Venables, 1999, for further development these ideas). These micro-foundations mean that outcomes (clustering or dispersion) can be linked to

parameters such as trade costs, as in Figure 1. The model also makes it clear that *ex ante* identical locations can be different *ex post*, and that there are multiple equilibria – we have to look outside the model, or rely on chance, to determine which of the regions has the manufacturing cluster.

The model was constructed with just two locations. How do these insights extend when there are many locations? With many locations the number of equilibria increases dramatically, and there is a danger that little can be said about outcomes. There are several ways through this problem. One is to investigate how the size and number of manufacturing centres on a given geographical space depends on underlying parameters such as trade costs and population levels. The approach of Fujita, Krugman and Venables (1999) is to hypothesize a circular economy (with population on the circumference) and to show that an initial random allocation of manufacturing grows into a determinate number of centres, the size of which is greater (and number of which is smaller) the lower trade costs are. Given some number of centres, reducing trade costs will have no effect until some critical point is reached, at which the economy will reorganize itself to a new economic geography with fewer and larger centres. The approach of Fujita and Mori (1997) is to suppose that initially there is a small populated region. Population growth causes this to expand, at first with the spread of agricultural production into the hinterland. However, these agriculture workers demand manufactures, and this will cause new manufacturing centres to develop. The expanding economy therefore grows its urban structure, and cities will tend to be larger (and further apart) the greater increasing returns to scale are and the lower trade costs are. Both of these approaches work with underlying geographies that are undifferentiated. Adding structure to these underlying geographies simplifies the problem in fairly natural ways. A transport node – such as a port or river crossing – will attract manufacturing, as firms in such a location have better access to a larger number of consumers.

Intermediate goods and industrial clusters

The clustering mechanisms described in the preceding section turn on the mobility of labour. Clustering occurs because, as firms and workers move, so do both supply *and* demand for manufactures. What if labour is immobile? An analogous mechanism can work between firms when we take into account intermediate goods, that is, goods that are both supplied and demanded by the manufacturing sector. This mechanism is similar to the idea of ‘linkages’ common in the development economics literature of the 1950s and 1960s. This studied the roles of backward linkages (demands from downstream firms to their suppliers) and of forward linkages (supply from intermediate producers to downstream activities) in developing industrial activity. However, as we saw above, rigorous treatment requires that the concepts are placed in an environment with increasing returns to scale, in order to force firms to make a location choice. This can be done in a model isomorphic to that outlined above, but in which firms in the manufacturing sector produce and use intermediate as well as final goods. Clustering can occur as it is profitable for firms producing intermediate and final goods to co-locate. Depending on the strength of linkages within and between industrial sectors, clustering might occur through a wide part of the economy or within narrowly defined sectors.

In this model clustering arises purely from the mobility of firms, even if there is little or no labour mobility. It is applicable to a number of different situations. For example, within a country there might be inelastic supply of land or housing in each city which places a limit on labour mobility. Clustering of particular sectors can nevertheless occur, and might be associated with different levels of employment and different house prices across cities.

The model has also been applied in the international context, with labour immobile across national boundaries. Manufacturing may then concentrate in a single country or group of countries, and this clustering may lead to international wage differences. This idea is developed by Krugman and Venables (1995) in a model with two countries, N and S, assumed to be *ex ante* identical. Firms produce final and intermediate goods, and use labour and intermediates as inputs. Equilibrium outcomes are summarized in Figure 2, which has trade costs on the horizontal axis and real wages on the vertical axis. At very high trade costs there is no clustering, so the two economies are identical; this is because firms operate in each country to supply local consumers. As trade costs fall (moving left on the figure) so the possibility of supplying consumers through trade rather than local production develops, and clustering forces become relatively more important. Below some level of trade costs, t^*, clustering forces come to dominate, and one of the countries (N) gains most of manufacturing, and consequently has a high real wage. This clustering ‘deindustrializes’ the other country (S), which experiences a fall in its real wage. For the case illustrated in Figure 2, there is a range of trade costs in which the world necessarily has a dichotomous structure. Wages are lower in S than in N, but it does not pay any firm to move to S as to do so would be to forgo the clustering benefits of large markets and proximity to suppliers that are found in N. However, as trade costs fall it becomes cheaper to ship intermediate goods, so the location of manufacturing becomes more sensitive to factor price differences. This is the era of globalization, in which manufacturing starts to move to S and the equilibrium wage gap narrows. In this model factor price equalization is attained when trade is perfectly free – the ‘death of distance’.

Figure 2

Real wages in a two-country model

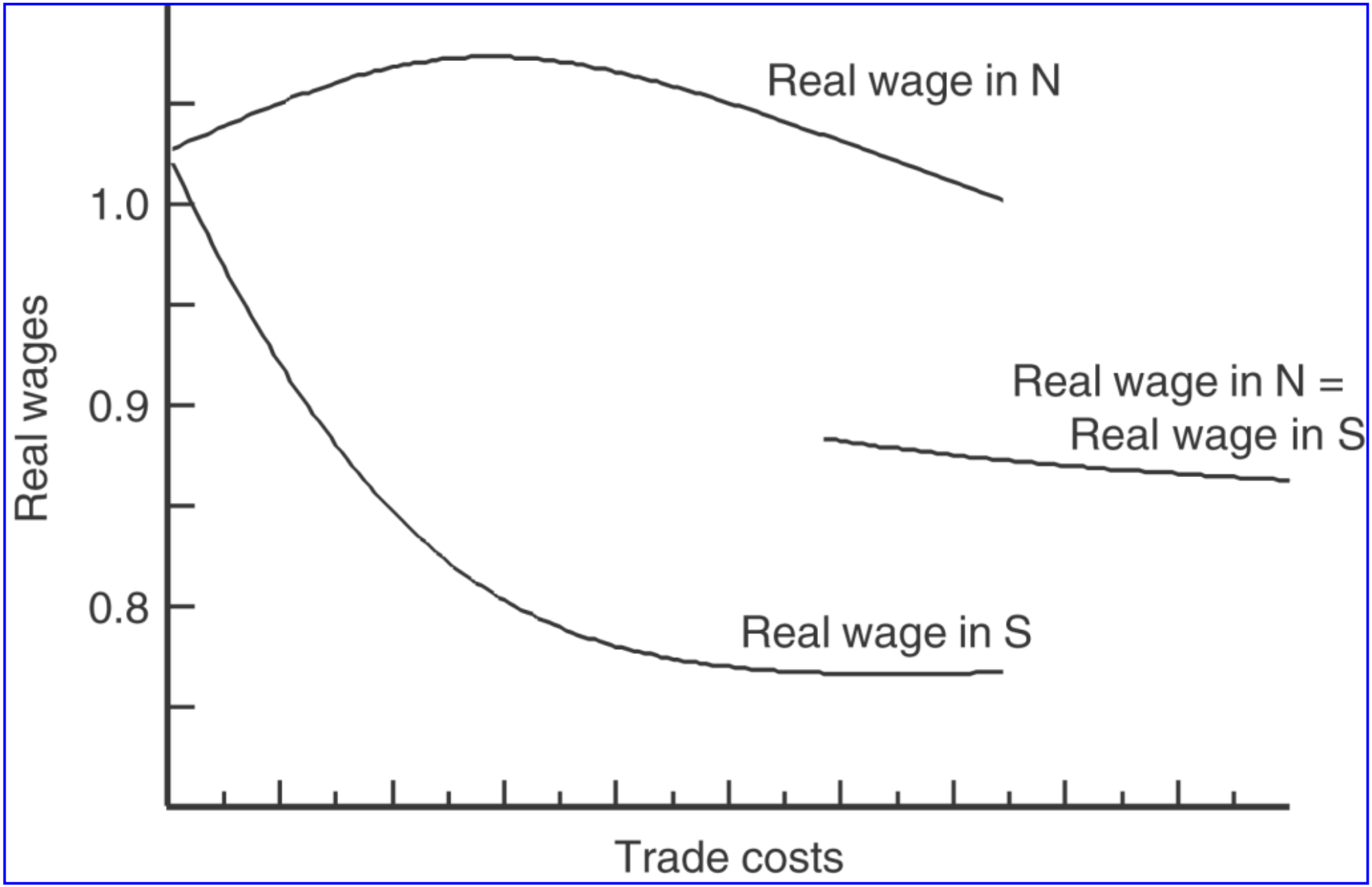

This model offers quite a general theory of location, in which four forces are at work, two of which are dispersion forces, and two favour clustering. The dispersion forces are *factor supply* and *product market competition*: moving a firm from S to N reduces the profitability of firms in N both by bidding up wages and by driving down product prices. Against this there are two agglomeration forces, *demand linkages* and *cost linkages*: moving a firm from S to N raises the profitability of firms in N by increasing the size of the market and by increasing the supply of intermediate goods. The balance between these four forces depends on parameters, including trade costs, giving the outcomes illustrated on Figure 2. It is worth comparing the four forces present in this model with the conventional model of free international trade, in which factor supply alone determines the location of economic activities.

Extensions of this approach provide a number of further insights concerning international inequalities. It suggests that the world may tend to organize into a rich club of countries and

a poor club. Economic development takes the form of countries growing from the poor club to the rich club in sequence rather than in parallel. Parallel growth is unstable because of the tendency of developing manufacturing sectors to cluster in a few countries.

Empirical findings

The new economic geography literature offers explanations of a number of phenomena that are empirically well documented – even obvious – such as the existence of cities and the presence of regional and international inequalities. Its insights range across different spatial scales, from the urban to the international. Empirical work is correspondingly diverse, and we refer to just four elements of it.

First, there is strong evidence of the importance of geography in shaping economic interactions. Trade costs are high (Anderson and van Wincoop, 2004), and 'gravity modelling' points to the fact that bilateral trade flows approximately halve with each doubling of distance between country pairs. Similar results hold for other cross-border interactions such as foreign direct investment flows, telephone calls, and international portfolio investments.

To turn to outcomes, a number of researchers have investigated the extent to which individual sectors are prone to clustering. There is a long business school tradition of work in this area, for example Porter (1990), who studies a number of industrial clusters. Econometric work has established that sectors are more prone to cluster than would be explained by chance or by comparative advantage (Ellison and Glaeser, 1997). A further prediction of NEG is that prices of immobile factors will be high in locations with good market access. As we have seen, in the national context this will show up in the price of land and housing and hence nominal wages differences, a prediction confirmed for US counties by Hanson (2005). In the international context this may show up as real wage differences. Gallup and Sachs (1999) find that 70 per cent of cross-country variation in per capita income can be accounted for by just four measures of physical and economic geography (malaria, hydrocarbon endowment, coastal access and transport costs). A structural approach to identifying the importance of market access in explaining cross-country income differentials is adopted by Redding and Venables (2004), who use gravity modelling to calculate measures of market access for each country. With other factors (such as institutional quality) controlled for, these measures of market access are important determinants of international wage gaps.

Finally, there is considerable evidence of the productivity benefits derived from being located in dense centres of economic activity. A recent survey of the literature on cities (Rosenthal and Strange, 2004) reports a consensus view that doubling city size is associated with a productivity increase of some three to eight per cent. However, a good deal of uncertainty surrounds the extent to which this is driven by the different clustering mechanisms – knowledge spillovers, thick labour markets, market access benefits, or inter-firm linkages – that we described above. Identifying the importance of each of these underlying mechanisms remains an active area of current research.

See Also

- city and economic development
- growth and international trade
- international trade theory
- regional development, geography of
- spatial economics
- symmetry breaking
- systems of cities
- urban agglomeration
- urban economics
- urban production externalities

Bibliography

Anderson, J. and van Wincoop, E. 2004. Trade costs. *Journal of Economic Literature* 42, 691–751.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Duranton, G. and Puga, D. 2004. Micro-foundations of urban agglomeration economies. In *Handbook of Urban and Regional Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Ellison, G. and Glaeser, E. 1997. Geographic concentration in US manufacturing industries: a dartboard approach. *Journal of Political Economy* 105, 889–927.

Fujita, M., Krugman, P. and Venables, A. 1999. *The Spatial Economy: Cities, Regions and International Trade*. Cambridge, MA: MIT Press.

Fujita, M. and Mori, T. 1997. Structural stability and the evolution of urban systems. *Regional Science and Urban Economics* 27, 399–442.

Gallup, J. and Sachs, J. 1999. Geography and economic development. In *Annual World Bank Conference on Development Economics 1998*, ed. B. Pleskovic and J. Stiglitz. Washington, DC: World Bank.

Hanson, G. 2005. Market potential, increasing returns and geographic concentration. *Journal of International Economics* 67, 1–24.

Helsley, R. and Strange, W. 1990. Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* 20, 189–212.

Henderson, J. 1988. *Urban Development: Theory, Fact and Illusion*. New York: Oxford University Press.

Krugman, P. 1980. Scale economies, product differentiation and the pattern of trade. *American Economic Review* 70, 950–9.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 49, 137–50.

Krugman, P. and Venables, A. 1995. Globalization and the inequality of nations. *Quarterly Journal of Economics* 110, 857–80.

Marshall, A. 1890. *Principles of Economics*, 8th edn. London: Macmillan, 1920.

Matouschek, N. and Robert-Nicoud, F. 2005. The role of human capital investments in the location decisions of firms. *Regional Science and Urban Economics* 35, 570–83.

Porter, M. 1990. *The Competitive Advantage of Nations*. New York: Macmillan.

Redding, S. and Venables, A. 2004. Economic geography and international inequality. *Journal of International Economics* 62, 53–82.

Rosenthal, S. and Strange, W. 2004. Evidence on the nature and sources of agglomeration economies. In *Handbook of Urban and Regional Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Starrett, D. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 17, 21–37.

How to cite this article

Venables, Anthony J. "new economic geography." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000247> doi:10.1057/9780230226203.1182

The New Palgrave Dictionary of Economics Online

new institutional economics

L.J. Alston
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The new institutional economics (NIE) consists of a set of analytical tools or concepts from a variety of disciplines in the social sciences, business and law. The NIE addresses two overarching issues: what are the determinants of institutions – the formal and informal rules shaping social, economic and political behaviour? And what impact do institutions have on economic performance? It is the impact of institutions via property rights and transaction costs that ultimately affect the ability of individuals and societies (at a macro level) to extract the gains from trade which in turn can lead to enhanced economic well-being.

Keywords

agenda control; collective action; contract enforcement; corruption; creative destruction; credible commitment; firm, theory of the; free-rider problem; institutional path dependence; monitoring costs; New Deal; new institutional economics; political entrepreneurship; principal and agent; Prisoner's Dilemma; property rights; rational ignorance; rule of law; side payments; social capital; social norms; technology; transaction costs of exchange; transaction costs of production; transformation costs

Article

What is the new institutional economics (NIE)? The NIE adds to the neoclassical framework insights and concepts from a variety of social sciences as well as business organization, history and law. Unlike past interdisciplinary forays by economists into other disciplines, proponents of the NIE have been less imperialist and instead have been importers of various concepts. This does not mean that the NIE is internally inconsistent. Indeed, the NIE is a set of analytical spokes that when put together properly form a wheel of analysis capable of addressing a broad variety of issues. The NIE consists of analytical spokes from a variety of disciplines: anthropology, business organization, economics, history, law, political science, psychology, and sociology. My purpose in this article is to identify the spokes and try to form the wheel in order to give a better understanding of the NIE.

A framework for understanding the new institutional economics

The alpha and the omega of the NIE are institutions and economic performance (Alston and Ferrie, 1999; Eggertsson, 1996; North, 1990). Institutions determine economic performance and economic performance determines institutions. This is nothing new. What is new are the conceptual spokes such as transaction costs, property rights, credible commitment, and agenda control that determine the simultaneous causal links between institutions and economic performance. It is important to emphasize that the NIE does not abandon neoclassical theory. As Figure 1 illustrates, the conceptual arrows beginning with technology to transformation costs (production isoquants, along with relative prices) are still the backbone of the theory of the firm that determine the costs of production and in the neoclassical world led to discussions of how far inside and/or where on the production possibilities frontier a country would be. Because of the limited ability of this stark depiction of the theory of the firm to explain many of the ‘big’ questions facing economists – for example, the lack of convergence in standards of living across countries – many economists added various concepts. Let us begin with the role of institutions.
Figure 1
Institutions and economic performance

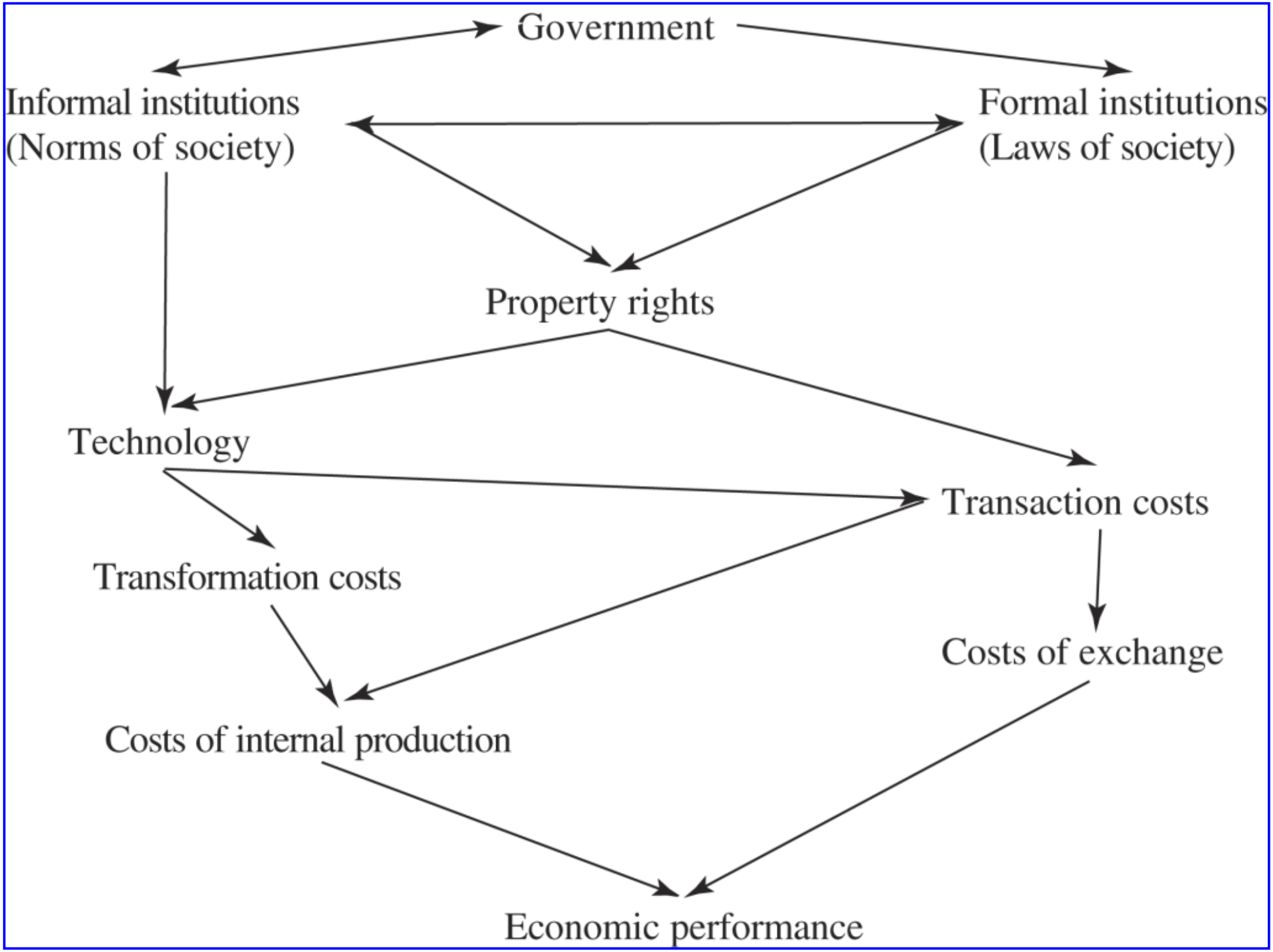

Institutions are the informal norms and formal laws of societies that constrain and shape decision-making or, as North (1990) defined them, ‘the rules of the game’. For a good

treatment of the interaction of norms and laws see Greif (2006). For the importance of social capital or norms see Keefer and Knack (2005). Informal norms do not rely on the coercive power of the state for enforcement whereas formal laws do, in part. The enforcement of formal laws does not rely entirely on the coercive power of the state because some of their force is derived from the beliefs of its citizens For example, if more people believe that littering is morally wrong, the costs that governments incur to police littering are lower. Similarly, if more people believe that recycling is morally right then they will incur their own costs to recycle even though to do so would not be in their self-interest strictly speaking. The existence of certain laws may simply be the codification of the norms of the majority. But, at times, and particularly during crises, some political leaders can influence the norms of citizens (Higgs, 1987). To the extent that political leaders can sway public opinion, the passage of laws may affect the beliefs of the constituents.

As Figure 1 shows, the norms and laws of society determine the property rights that individuals possess. Here I am concerned with rights that individuals have in regard to goods and services: (1) the right to sell an asset; (2) the right to use and derive income from an asset; and (3) the right to bequeath an asset. Property rights are enforced in three ways. Individuals themselves enforce their assigned rights; for example, we put locks on our doors to protect our property. Societal sanctions such as ostracism can deter individuals from violating the assigned rights of others. And the coercive power of the state can be used to enforce property rights; for example, the police will evict trespassers.

Technology, which the standard neoclassical model took as exogenous, is shaped by the property rights, and the norms and endowments of citizens. Property rights along with technology determine the transaction costs and transformation costs associated with exchange and production. Robertson and Alston (1992) present a schematic framework for analysing the impact of technology on the transaction costs of production. Transformation costs are the physical costs (in an engineering sense but based also on relative prices) of combining inputs to produce output. The transformation costs of production depend on the technology in society. The transaction costs of production are the invisible costs of production and initially discussed by Coase (1937) in his seminal article for the NIE, ‘The Nature of the Firm’. Transaction costs include: (1) search and negotiation costs; (2) monitoring labour effort; (3) coordinating the physical factors of production; (4) monitoring the use of the physical and financial capital employed in the production process; and (5) enforcing the terms of the contract. It is the transaction costs within a firm – along with transformation costs – relative to the transaction costs of using the market that Coase first identified as being decisive in determining the firm/market boundary. Others within the tradition of the NIE have extended this considerably, most notably Yoram Barzel (1989) and Oliver Williamson (1985). The extensions have provided answers to issues associated with long-term contracting, for example, Goldberg and Erickson (1987); Joskow (1985); hybrid contracts of various sorts (Menard, 2005) and various forms of business organization, for example, franchises (Lafontaine, 1992).

Both technology and property rights can affect the transaction costs of production in a variety of ways. Technology generally reduces both the direct costs of monitoring, through better surveillance, and reduces the need to monitor, that is, capital standardizes the marginal productivity of labour, holding constant monitoring. As an historical example, in agriculture, when workers cut down weeds by hand, monitoring costs were higher than when workers drove through the fields with a mechanical cultivator that cut down the weeds. Whether on the farm or in the factory, machines by their very nature reduce the discretion of labour. They standardize the production process and thereby reduce the variation in the marginal product of labour. In addition, technology influences the transaction costs of coordinating production; for example the computer is partially responsible for the observed increase in horizontal integration in commercial banking in the United States in the 1990s. The huge merger wave in the banking industry in the 1990s was partially the result of legal changes that in turn could have been prompted by the lobbying efforts from the financial industry in recognition of the cost savings associated with the advent of computer technology. Norms and property rights can also affect the transaction costs of production. For example, if people believe in working hard in some cultures (perhaps because of past incentives), providing ‘an honest day's work for an honest day's pay’, then the monitoring costs borne by the residual claimant are lower. Similarly, if the property rights in a society make it easy to dismiss workers for shirking, then monitoring costs would also decrease.

The transaction costs of exchange include the costs associated with negotiating and enforcing contracts. For some exchanges, the transaction costs of exchange are low because informal norms suffice to uphold bargains. Most local communities have well-established customs that limit opportunistic behaviour. Similarly, repeat transactions often give a sufficient incentive to deal fairly. Though local or repeat exchanges may have low transaction costs, the gains from such trade are limited because the extent of the market limits the number of individuals with whom one can deal locally or repeatedly. Formal institutions are necessary if the full gains from specialization in an extended market are to be captured. I use the term ‘full gains’ because some trade can be accomplished through self-generated reputation and the prospect of repeat business without relying on outside formal government institutions (Telser, 1981). This is particularly evident in the case of international transactions where the participants do not share a common body of law. For example, the extension of the market may require that more trades occur among anonymous parties or that more trades occur where payment and delivery are not simultaneous. Institutions can reduce the potential for unscrupulous behaviour inherent in such arrangements.

The presence of ‘honest’ courts and a body of law that upholds contracts and safeguards exchanges is a formal institution that determines the property rights of individuals which in turn affect the transaction costs of exchange. The shorthand concept used to describe this system is ‘the rule of law’ (Arrunada and Adonova, 2005; Beck and Levine, 2005; Hadfield, 2005). This does not imply that the courts are used frequently, only that they form a backdrop for exchange. The availability of recourse to law and the courts provides a safeguard for market participants engaged in anonymous or non-simultaneous exchanges. In the absence of honest courts, negotiation and enforcement costs will be higher. As a consequence, contracts will be written in ways that will safeguard the exchange should one party desire to act opportunistically. Williamson (1985) describes how contractors shield themselves from the potential opportunistic behaviour of others. Levy and Spiller (1994) illustrate the role of institutions in providing commitment in the context of safeguarding investments in the regulation of telecommunications. Firms (and legislative and executive bodies) also use the courts strategically but here I treat firms as responding exogenously to their

expectation of decisions by courts.

At times there may be insufficient safeguards so that the result is not an exchange. For example, large investments are generally required to reap economies of scale. A part of that investment may not be readily transferable to other uses (that is the investments are asset specific – see Williamson, 1985, for an expansive treatment of specific assets). Before the investment is made, if there is a fear that some of the value of the investment will be expropriated, either through nationalization, taxes, regulations, or opportunistic behaviour by one of the contractors, firms will not invest as much as they would in the absence of such fears (Spiller and Tommasi, 2005). Expropriation could occur either through actions taken by the state (such as regulation or nationalization) or through actions taken by one of the parties (such as refusing to execute the exchange without a renegotiation of terms).

Given the set of institutions in a society, residual claimants will construct contracts with the suppliers of inputs to minimize the sum of transformation and transaction costs within a firm, and between firms and firms and consumers. The results are a variety of contracts with differing transaction cost and production cost components, and different total costs of production. The varying contracts in turn influence economic performance. As an example there is a voluminous literature associated with principal agent problems ranging from tenancy in agriculture (Alston, 2003) to corporate governance (Fama, 1980).

The conceptual framework presented in Figure 1 and discussed thus far is basically static; it illustrates the ultimate importance of institutions for economic performance but it does not address the determinants of institutions and institutional change (Alston, 1996; North, 2005). To understand the process of institutional change, it is useful to think about economic performance or economic growth as a process of creative destruction (Schumpeter, 1942). Creative destruction means that there are winners and losers associated with economic performance (see Figure 2). The losers have an incentive to lobby government for institutional change to protect them from the ravages of the market, while the winners have an incentive to lobby for the status quo or an even better outcome. Consumers have an interest in the outcome, but given the existence of rational ignorance and free-rider problems consumers tend not to be as effective as special interests in the political marketplace. By rational ignorance, we mean that it does not pay the consumer to be as informed about legislation as special interest groups (Olson, 1965; Buchanan and Tullock, 1962). The free-rider problem arises because of the large numbers of consumers have difficulties in organizing collectively to prevent policy changes. Political entrepreneurs may attenuate both these problems because the interests of consumers are represented somewhat through competition amongst politicians who bring issues to the attention of consumers, and thus limit the power of special interests (Denzau and Munger, 1986).

Figure 2

The determinants of formal institutions

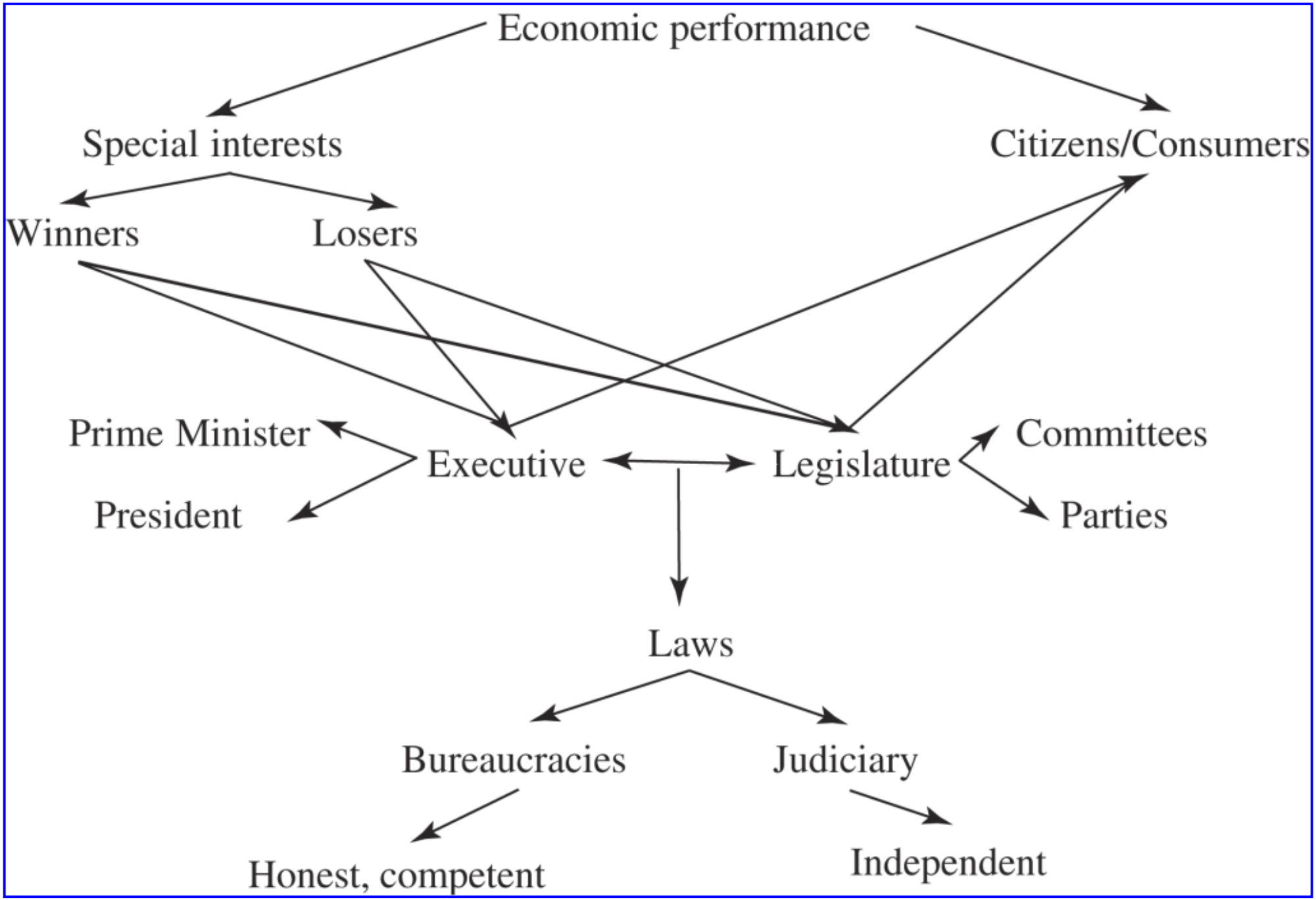

We can think of those who lobby for changes in institutions or for the status quo as the demand side of legislation. But special interest groups do not enact legislation. Their demands get filtered through a political process of government institutions – what I call the supply side of legislation. By using the terms ‘demand’ and ‘supply’ I do not mean that there is necessarily a unique outcome; the term ‘bargaining’ may be more appropriate. Curiously, until recently, economists have paid little attention to the supply side of government, leaving the modelling of the political process to political scientists; ‘curiously’ because the concepts of demand and supply are the two most important components of neoclassical economics. The supply of legislation can be initially decomposed into the executive, legislative and judicial branches. In parliamentary systems, the executive, prime minister, and the legislature

are more interconnected than in presidential systems, so that the same demands may end up with a different result depending on whether a country has a presidential or parliamentarian system (Carey, 2005). Within legislatures there are a myriad of coordinating devices; historically, in the United States, political parties and the committee structure in legislatures have played major roles in shaping political outcomes (Cox and McCubbins, 1993; 2005; Shepsle, 1978; Shepsle and Weingast, 1984). Political parties and committees have a certain amount of agenda control. For example, the party leadership makes appointments to committees, and committees in turn have the power to veto bills simply by refusing to report the bill out of committee. In addition they can amend bills to better suit their preferences. In parliamentary systems, particularly two-party dominant parliamentary systems, the majority power has significant agenda control. In other countries, most notably those with strong executive powers, such as in Brazil or Chile, the demand for legislation is filtered through the preferences of the president who negotiates with members of Congress using his powers to sway votes (Alston and Mueller, 2006). Changes in either demand or supply side forces will result in institutional change. Legislation can be either specific or vague in content (Spiller, 1996). In either case the law is administered through bureaucracies, giving rise to another set of principal–agent problems between the legislature and the agency to which the law is delegated (Ferejohn and Shipan, 1990; Weingast and Moran, 1983; McCubbins, Noll and Weingast, 1987). In the United States, the Environmental Protection Agency (EPA) is frequently cited as an example of a bureaucracy with large discretion because of the vagueness of its mandate from Congress.

The outcomes of this demand and supply side bargaining are the formal laws and regulations of a society, subject to the explicit or implicit sanction of the courts. It matters a great deal whether the courts that interpret the constitutionality of legislation are independent of the executive and legislative branches. If the courts are truly independent the executive and legislative branches will enact legislation ‘in the shadow of the court’, knowing that the court could overturn legislation. The dismal political and economic history of Argentina since 1945 is a good example of the impact on economic performance from a Supreme Court that has not been independent (Alston and Gallo, 2007; Iaryczower, Spiller and Tommasi, 2002).

Where do we go from here?

Before discussing institutional lock-in, the topic to which I believe we should devote more of our intellectual resources, it is worth considering which parts of the framework of the new institutional economics we know best. The hands-down winner is the area of contracting. We have much empirical evidence on how contracts change in response to different transaction costs, which in turn result from the formal laws and informal norms in societies. We also know a good deal about why governments pass the laws and regulations that they do. Here there has been an outpouring by both economists and political scientists, with economists tending to specialize in demand-side explanations – for example, the role of special interests – and political scientists specializing in supply-side explanations – for example, the role of committees and the importance of agenda control. So if we know why we get the laws, and we know how laws affect contracting, what is missing? What is missing is a better understanding of the transaction costs associated with getting laws and regulations that are more conducive to better economic performance, especially when it becomes obvious that the existing laws and regulations are not fostering economic growth (Shirley, 2005). In many scenarios special interests are in a position to either enact legislation or block legislation so that they reap the gains. Yet society is worse off by such activity. The question is: why cannot ‘we’, the citizens or consumers, buy out the special interests? For many societies, poor economic performance is explained by corrupt governments, who are more or less stealing from their own citizens. Here we focus on issues beyond corruption, though corruption is clearly in the domain of the NIE. There are several possible explanations for institutional lock-in:

1. 1. Informational problems abound such that citizens are unaware of possible policy moves that would improve on the status quo (North, 2005 and citations therein).
2. 2. Though citizens do not like the outcome, they approve of the process that produced the outcome.
3. 3. Even when aware, there are serious collective action problems.
4. 4. Insecurity in political property rights prevents transactions from occurring, that is, you cannot buy what someone else does not own.

Let us explore each of these in turn.

Given rational ignorance it may be that many citizens are simply unaware of property rights arrangements that would improve societal welfare. For example, under the Homestead Act in the United States settlers could acquire property rights to 160 acres of unoccupied federal land by residing and ‘improving’ the land. These homestead plots turned out to be economically too small and promoted externalities associated with wind erosion. Even after the great dust bowl of the 1930s, plots remained small because subsidies by the federal government enabled farmers to remain on the land. Why did the federal government not move to reallocate land or at least not interfere with consolidation through markets? It appears that the answer rests with the information available to citizens and their beliefs in the virtues of small landholdings. This is coupled with the efforts of local politicians to maintain a population base (Hansen and Libecap, 2004a; 2004b). Ironically, in the latter part of the 19th century Major John Wesley Powell recognized the potential problems of settlements in the arid or sub-humid regions of the country, but his reports to Congress were ignored in favour of boosterism (Stegner, 1954).

Another example concerns consumers who may simply be unaware of policy moves that would improve their welfare. For example, the United States and many other countries have

allowed their ocean shippers to participate in cartels that set prices on ocean routes around the globe (Sicotte, 1997). When I mention this to scholars, most are unaware of this price fixing. What needs to be done is to determine how many redistributive programmes exist where a policy move would be wealth-enhancing yet does not occur either because of insufficient information or an inability of citizens to process the cause and effects of potential policy moves in the face of risk aversion, that is, we know the effect of the status and do not fully comprehend a counterfactual policy world. Many institutions are bundled in ways that makes decoupling difficult. It is partially a coordination problem and it is partially a case of risk aversion – once you open Pandora's Box you are uncertain as to the final outcome.

Another reason for institutional path dependence is circumstances where citizens have a deep belief in the process that produces laws and regulations even though they may disapprove of some legislative outcomes. The majority may opt to support the status quo legislation because changing the law would entail changing a higher order institution concerning overall institutional development. From US history an example of public disapproval of changing the system of checks and balances was the attempt by President Roosevelt to add Justices to the Supreme Court. Roosevelt wanted to stack the Court because the Court was ruling that some major legislative acts were unconstitutional, for example, the Agricultural Adjustment Act and the National Industrial Recovery Act. By adding Justices, Roosevelt believed that his New Deal legislation would pass the constitutional test. Even though most people supported the New Deal legislation, there was a public outcry against Roosevelt's attempt to change the rules affecting checks and balances so as to achieve his legislative goals.

Alternatively, people may be aware of the dissipation associated with the status quo arrangement of property rights, but it is in no one's self-interest to mount an organizational campaign to change the existing regulations. This is the classic collective action problem developed independently but almost simultaneously by Buchanan and Tullock (1962) and Olson (1965) – one could also model this as a multi-player Prisoner's Dilemma game. The collective action problems are particularly acute in situations entailing multiple governments across international boundaries, for example, overfishing in international waters or global warming. The difficulties for international property rights are twofold: specification and enforcement. Specification is difficult because of knowledge or beliefs about the state of the world differ (for example, global warming) but even if beliefs are the same, preferences can vary across countries because of incomes (for example, the United States versus Mexico) or simply preferences (for example, the United States versus Germany on green issues). Collective action problems occur in representative democracies as well as dictatorial regimes. We have instances of both types of regimes not specifying and enforcing property rights at what would appear to be optimal times. For example, the United States squandered considerable oil reserves in the early 20th century and Indonesia mowed through a large stock of their tropical hardwoods in the latter part of the 20th century.

A fourth possibility for the lack of policy reform is insecure political property rights. It may be that individuals are aware and willing to organize but there is no ‘market’ for the emergence of property rights. Suppose that the winners from a status quo policy have the political power to veto or allow policy changes. Given their power, they would be foolish to acquiesce to policy moves that made them worse off, even if they were wealth enhancing. But, they would allow such a policy move if they were compensated. The actions of the landless peasants' movement (MST) in Brazil are consistent with this argument. The MST is very effective at swaying public opinion and thereby prompting politicians to expropriate land and transfer it to peasants; but they do not support deeding the land to peasants. The MST prefers to keep the peasants dependent on the MST as a collective because it is easier for them to extract payments from the group than individual farmers (Alston, Libecap and Mueller, 2005).

Why is it that we generally do not allow such side payments? One answer is that transparent side payments would undermine the legitimacy of the organization, whether the organization is the MST, a union or a government. If the current property rights arrangement is viewed as inferior to an alternative, people ‘believe’ that they should not pay to move to a better property rights arrangement. The result is institutional lock-in. Yet there have been examples of improving the status quo for all parties involved. A case involving the sale of water in the 1990s illustrates the difficulties in changing the status quo. The Imperial Valley Irrigation District, a governmental unit that has jurisdiction over water, entered into a contract to sell some of its water to the city of San Diego. The Imperial Valley Water District has property rights to water that are subsidized by US taxpayers. As such it can sell water at prices higher than it pays. Interestingly, members of the Imperial Water District decided that they would only sell water that they have conserved through better irrigation technologies. The interesting question is: why didn't they fallow all their land and sell their entire water allocation? I speculate that they were concerned about the political fallout that could have resulted in the district losing its current subsidy. In short, it appears as if they have secure property rights to the rental stream of water but not the clear ‘political’ property right to the stock. The establishment of ‘water banks’ throughout the West – whereby farmers could sell their flow of water to urban users or resort users – have failed primarily because farmers are afraid of losing their property right when it becomes transparent that farming is not the highest-valued use of water in the West.

Another factor promoting the insecurity of political property rights falls under the rubric of credible commitment (North and Weingast, 1989). In representative democracies politicians face the demands of constituents who may be harmed or obtain benefits from a rearrangement of property rights. The demands of the majority of voters may not coincide with the optimal arrangements of property rights, and politicians cannot commit to making side payments over time to compensate the losers. Authoritarian regimes are subject to similar problems associated with catering to populist demands. A good example of this was the infringement in property rights by Peron in Argentina in the late 1940s. Peron imposed rent and price controls in the Pampas, the most fertile and productive agricultural producing area in Argentina. The punitive arrangement in property rights lead to a decline in investment which, along with political instability, affected growth in the long run (Alston and Gallo, 2007; Spiller and Tommasi, 2003; 2007).

A more cynical view of political behaviour suggests that we do not want to encourage paying for changes in property rights because to do so would promote the creation and maintenance of non-optimal property rights in order to be paid to move to a more optimal situation. Campaign finance and corruption around the globe may be testimony to special interests trying to ‘bribe’ politicians to maintain or change property rights. In some instances politicians may use part of the contributions to make side payments (Norlin, 2003).

Explaining institutional rigidities in the face of poor economic performance is a difficult research agenda. To understand the lock-in requires insights from the disciplines that comprise the NIE – anthropology, business organization, economics, history, law, political science, psychology and sociology. Yet the potential reward from an understanding of the forces that account for poor economic performance is huge. The research agenda includes both international cross-sectional studies and case studies of successful and unsuccessful institutional change. The international cross-sections allow us to quickly determine the correlates of successful economic performance, for example secure property rights, while the case studies allow us to stack the building blocks that will ultimately allow us to produce a more general framework for the determinants of institutional change (Alston, 2007).

See Also

- firm, theory of the
- growth and institutions
- ‘political economy’
- political institutions, economic approaches to
- property rights
- transaction costs, history of

Bibliography

Alchian, A. and Demsetz, H. 1972. Production, information costs, and economic organization. *American Economic Review* 62, 777–95.

Alston, L.J. 1996. Empirical work in institutional economics. In *Empirical Studies in Institutional Change*, ed. L.J. Alston, T. Eggertsson and D. North. New York: Cambridge University Press.

Alston, L.J. 2003. Tenant farming. In *The Oxford Encyclopedia of Economic History*, vol. 5, ed. J. Mokyr. New York: Oxford University Press.

Alston, L.J. 2007. The ‘case’ for case studies zin the new institutional economics. In *New Institutional Economics: A Guidebook*, ed. J. Glachant and E. Brousseau. Cambridge: Cambridge University Press.

Alston, L.J., Eggertsson, T. and North, D.C. 1996. *Empirical Studies in Institutional Change*. New York: Cambridge University Press.

Alston, L.J. and Ferrie, J.P. 1993. Paternalism in agricultural labor contracts in the U.S. South: implications for the growth of the welfare state. *American Economic Review* 83, 852–75.

Alston, L.J. and Ferrie, J.P. 1999. *Southern Paternalism and the Rise of the Welfare State: Economics, Politics, and Institutions in the U.S. South, 1865–1965*. New York: Cambridge University Press.

Alston, L.J. and Gallo, A. 2007. Electoral fraud, the rise of Peron, and demise of checks and balances in Argentina. Working Paper No. ES2007-0001, Institute of Behavioral Science, University of Colorado.

Alston, L.J., Libecap, G.D. and Mueller, B. 1999. *Titles, Conflict and Land Use: The Development of Property Rights and Land Reform in the Brazilian Amazon Frontier*. Ann Arbor: University of Michigan Press.

Alston, L.J., Libecap, G.D. and Mueller, B. 2000. Land reform policies: the sources of violent conflict, and implications for deforestation in the Brazilian Amazon. *Journal of Environmental and Economics and Management* 39, 162–88.

Alston, L.J., Libecap, G. and Mueller, B. 2005. How interest groups can influence political outcomes indirectly through information manipulation: the landless peasants movement in Brazil. Working Paper No. EB2005-0005, Institute of Behavioral Science, University of Colorado.

Alston, L.J., Libecap, G.D. and Schneider, R. 1996. The determinants and impact of property rights: land titles on the Brazilian frontier. *Journal of Law, Economics and Organization* 12, 25–61.

Alston, L.J. and Mueller, B. 2005. Property rights and the state. In Menard and Shirley (2005).

Alston, L.J. and Mueller, B. 2006. Pork for policy: executive and legislative exchange in Brazil. *Journal of Law Economics and Organization* 22, 87–114.

Anderson, T.L. and Hill, P.J. 2004. *The Not So Wild, Wild West: Property Rights on the Frontier*. Stanford: Stanford University Press.

Anderson, T.L. and McChesney, F.S. 2003. *Property Rights: Cooperation, Conflict and Law*. Princeton: Princeton University Press.

Arrunada, B. and Andonova, V. 2005. Market institutions and judicial rule-making. In Menard and Shirley (2005).

Barzel, Y. 1989. *Economic Analysis of Property Rights*. New York: Cambridge University Press.

Barzel, Y. 2002. *A Theory of the State: Economic Rights, Legal Rights and the Scope of the State*. New York: Cambridge University Press.

Beck, T. and Levine, R. 2005. Legal institutions and financial development. In Menard and Shirley (2005).

Buchanan, J.M. and Tullock, G. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.

Carey, J.M. 2005. Presidential versus parliamentary government. In Menard and Shirley (2005).

Coase, R.H. 1937. The nature of the firm. *Economica* 4, 386–405.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Coase, R.H. 1992. The institutional structure of production. *American Economic Review* 82, 713–9.

Cox, G.W. and McCubbins, M.D. 1993. *Legislative Leviathan: Party Government in the House*. Berkeley: University of California Press.

Cox, G.W. and McCubbins, M.D. 2005. *Setting the Agenda: Responsible Government in the U.S. House of Representatives*. Cambridge: Cambridge University Press.

Crocker, K.J. and Masten, S.E. 1991. Pretia ex machina? Prices and process in long-term contracts. *Journal of Law and Economics* 34, 69–99.

Denzau, A.T. and Munger, M. 1986. Legislators and interest groups: how unorganized interests get represented. *American Political Science Review* 80, 84–106.

Eggertsson, T. 1990. *Economic Behavior of Institutions*. New York: Cambridge University Press.

Eggertsson, T. 1996. A note on the economics of institutions. In *Empirical Studies in Institutional Change*, ed. L.J. Alston, T. Eggertsson and D. North. New York: Cambridge University Press.

Fama, E.F. 1980. Agency problems and the theory of the firm. *Journal of Political Economy* 88, 283–307.

Ferejohn, J. and Shipan, C. 1990. Congressional influence on bureaucracy. *Journal of Law, Economics, and Organization* 6, 1–27.

Furubotn, E.G. and Richter, R. 2005. *Institutions and Economic Theory: The Contribution of the New Institutional Economics*. Ann Arbor: University of Michigan Press.

Goldberg, V.P. and Erickson, J.R. 1987. Quantity and price adjustment in long-term contracts: a case study in petroleum coke. *Journal of Law and Economics* 31, 369–98.

Greif, A. 2006. *Institutions and the Path to the Modern Economy*. New York: Cambridge University Press.

Hadfield, G.K. 2005. The many legal institutions that support contractual commitments. In Menard and Shirley (2005).

Hansen, Z.K. and Libecap, G.D. 2004a. The allocation of property rights to land: US land policy and farm failure in the northern great plains. *Explorations in Economic History* 41, 103–29.

Hansen, Z.K. and Libecap, G.D. 2004b. Small farms, externalities, and the dust bowl of the 1930s. *Journal of Political Economy* 112, 665–94.

Higgs, R. 1987. *Crisis and Leviathan: Critical Episodes in the Growth of American Government*. New York: Oxford University Press.

Iaryczower, M., Spiller, P. and Tommasi, M. 2002. Judicial independence in unstable environments, Argentina 1935–1998. *American Journal of Political Science* 46, 699–716.

Joskow, P.L. 1985. Vertical integration and long-term contracts: the case of coal-burning electric generating plants. *Journal of Law, Economics and Organization* 1, 33–80.

Keefer, P. and Knack, S. 2005. Social capital, social norms and the new institutional economics. In Menard and Shirley (2005).

Lafontaine, F. 1992. Agency theory and franchising: some empirical results. *Rand Journal of Economics* 23, 263–83.

Levy, B. and Spiller, P.T. 1994. The institutional foundations of regulatory commitment: a comparative analysis of telecommunications regulation. *Journal of Law, Economics and Organization* 10, 201–46.

Libecap, G.D. 1989. *Contracting for Property Rights*. New York: Cambridge University Press.

Libecap, G.D. and Hansen, Z.K. 2003. Small farms, externalities, and the dust bowl of the 1930s. Working Paper No. 04-01, Department of Economics, University of Arizona.

McCubbins, M., Noll, R. and Weingast, B. 1987. Administrative procedures as instruments of political control. *Journal of Law, Economics, and Organization* 3, 243–77.

Menard, C. 2005. A new institutional approach to organization. In Menard and Shirley (2005).

Menard, C. and Shirley, M.M. 2005. *Handbook of New Institutional Economics*. Dordrecht: Springer.

Norlin, K. 2003. Political corruption: theory and evidence from the Brazilian experience. Ph.D. thesis, University of Illinois.

North, D.C. 1990. *Institutions, Institutional Change, and Economic Performance*. New York: Cambridge University Press.

North, D.C. 2005. *Understanding the Process of Economic Development*. Princeton: Princeton Economic Press.

North, D. and Weingast, B. 1989. Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. *Journal of Economic History* 49, 803–32.

Olson, M. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Robertson, P.L. and Alston, L.J. 1992. Technological change and the organization of work in capitalist firms. *Economic History Review* 45, 330–50.

Schumpeter, J. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper.

Shepsle, K. 1978. *The Giant Jigsaw Puzzle: Democratic Committee Assignments in the Modern House*. Chicago: University of Chicago Press.

Shepsle, K.A. and Weingast, B.R. 1984. Legislative politics and budget outcomes. In *Federal Budget Policy in the 1980s*, ed. G.B. Mills and J.L. Palmer. Washington, DC: Urban Institute Press.

Shirley, M.M. 2005. Institutions and development. In Menard and Shirley (2005).

Sicotte, R. 1997. The organization and regulation of international shipping cartels. Unpublished Ph.D. thesis, University of Illinois.

Smith, H.E. 2000. Semicommon property rights and scattering in the open fields. *Journal of Legal Studies* 29, 131–70.

Spiller, P.T. 1996. A positive theory of regulatory instruments: contracts, administrative law or regulatory specificity. *USC Law Review* 69, 477–515.

Spiller, P.T. and Tommasi, M. 2003. The institutional foundations of public policy: a transactions approach with application to Argentina. *Journal of Law Economics and Organization* 19, 281–306.

Spiller, P.T. and Tommasi, M. 2005. The institutions for regulation: an application to public utilities. In Menard and Shirley (2005).

Spiller, P.T. and Tommasi, M. 2007. *The Institutional Foundation of Public Policy in Argentina*. New York: Cambridge University Press.

Stegner, W.E. 1954. *Beyond the Hundredth Meridian: John Wesley Powell and the Second Opening of the West*. New York: Penguin.

Telser, L. 1981. A theory of self-enforcing agreements. *Journal of Business* 53, 27–44.

Weingast, B.R. and Marshall, W. 1988. The industrial organization of congress: or why legislatures like firms are not organized as markets. *Journal of Political Economy* 96, 132–63.

Weingast, B.R. and Moran, M.J. 1983. Bureaucratic discretion or congressional control? Regulatory policymaking by the FTC. *Journal of Political Economy* 91, 765–800.

Williamson, O. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

Williamson, O. 2000. The new institutional economics: taking stock, looking ahead. *Journal of Economic Literature* 38, 595–613.

How to cite this article

Alston, L. J. "new institutional economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000170> doi:10.1057/9780230226203.1183

The New Palgrave Dictionary of Economics Online

new Keynesian macroeconomics

Huw David Dixon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘new Keynesian economics’ refers to a body of work done by macroeconomists in the late 1970s and 1980s in which the notion of imperfect competition was introduced into macroeconomics in order to provide a micro-foundation for nominal rigidities and also to provide an alternative to supply-equals-demand equilibrium. This led in the 1990s to the new-neoclassical-synthesis approach to monetary economics in which dynamic pricing models have become central to our understanding how monetary policy influences output and inflation. Other themes in the new Keynesian approach include the effect of imperfect competition on the fiscal multiplier, and coordination failures.

Keywords

adaptive expectations; aggregate supply and demand models; balanced budget multiplier; competitive equilibrium; constant hazard rate models; consumer surplus; coordination failure; efficiency wages; envelope theorem; extrinisic and intrinsic uncertainty; hybrid Phillips curve; imperfect competition; inflation; invisible auctioneer; IS–LM model; Keynesian Revolution; menu costs; microfoundations; monetary transmission mechanism; monopolistic competition; monopoly; neoclassical synthesis; neutrality of money; new classical macroeconomics; new Keynesian macroeconomics; new Keynesian Phillips curve; new neoclassical synthesis; nominal rigidities; perfect competition; Phillips curve; price indexation; price making; price rigidity; price taking; rational expectations; real business cycles; real rigidities; representative agent; s-S models; staggered price-setting; staggered wage-setting; state-dependent pricing; strategic complementarity; Taylor principle; time-dependent models; wage indexation; wage rigidity

Article

The term ‘new Keynesian economics’ came into popular usage in the 1980s. The origins of the term are

fairly easy to understand in broad historical terms. In the classical approach of the pre-Keynes world (prior to 1936), wages and prices were seen as perfectly flexible and markets competitive (or at least ideally so). The Keynesian Revolution argued that prices, and more importantly wages, were rigid, and in order to understand phenomena like prolonged mass unemployment it was necessary to see how the economy operated when not in competitive equilibrium. In the post-Second World War period there emerged the neoclassical synthesis model that dominated macroeconomics from the 1950s to the mid-1970s. The essence was that in the *long run* all prices are perfectly flexible and the competitive or 'Walrasian' equilibrium will hold. However, in the *short run* prices and/or wages were treated as given. Thus there were the IS–LM and aggregate supply and demand (AS–AD) models, which were the workhorses of macroeconomic research until the mid-1970s and have remained established in many textbooks to the present day.

This approach was in the process of being overtaken at the level of research by the 'new classical' or rational expectations revolution of the 1970s. One aspect of the neoclassical synthesis was that not only prices but also expectations were treated as fixed in the short run, or subject to ad hoc adjustment, as under the adaptive expectations hypothesis. The new classical approach was based on the idea that wages and prices are perfectly flexible, but that agents did not have full information: even though agents used the information they had optimally (rational expectations), markets could deviate from the full information equilibrium. For example, agents might not know about the values of certain current variables such as aggregate price or the money supply when deciding how much output to produce or labour to supply.

The new Keynesian economics was to incorporate the rational expectations framework. However, it was to focus on the key issue of nominal rigidity: how do we understand the short-term rigidity of wages and/or prices in terms of providing a microfoundation that will explain why prices might not be perfectly flexible? Now, this required a 'revolution' of the order of magnitude of the rational expectations revolution. That revolution consisted in one idea: in order to understand nominal rigidity, it was necessary to abandon the approach of perfect competition with price-taking agents, and replace it with an approach where there are wage and price-setting agents. This is self-evident in hindsight: if you want to understand why wages and prices are rigid in the short run, you have to have agents who set the price, so that you can understand the microeconomics of price adjustment. If all agents (firms, households) are price-takers, prices can only be explained by some notion of 'demand equals supply' and a shadowy Walrasian auctioneer acting like an invisible puppet master-cum-market maker, adjusting prices gradually in response to excess demand or supply. This is hardly the basis for a rigorous theory of why prices and wages are not always at their market clearing levels: maybe the auctioneer called in sick or went on holiday!

Just to complete the historical setting, alongside the new Keynesian ideas there was the real business cycle (RBC) research programme which put forward the radical idea that nominal wage and price behaviour were irrelevant for understanding macroeconomic dynamics. Changes in output and employment were seen to be driven by real things such as productivity shocks, and the savings and investment decisions of agents as inherently dynamic. This was a radical agenda, which also pushed macroeconomics into trying to provide a quantitative explanation of economic fluctuations based on a competitive equilibrium model. However, despite many successes, the methodological idea of ignoring nominal things was an unsustainable self-limitation. For one thing, governments and central bankers are interested in the nominal side of the economy – inflation, the transmission mechanism of monetary

policy, to name a few.
So, in the mid-1990s there emerged the ‘new’ neoclassical synthesis (NNS). This combined the dynamic framework of the RBC approach with dynamic pricing models developed by the new Keynesian approach. The key idea is that in the long run money is neutral, but in the short run there is some nominal rigidity resulting from the price-setting behaviour of firms (and wage-setting behaviour of unions). This approach to modelling has certainly become the dominant school of thought, at least in central banks of Europe and the United States. It differs from the old neoclassical synthesis in that the model is fully dynamic and microfounded and the equilibrium imperfectly competitive.

The microfoundations of wage and price rigidity

So the problem in the late 1970s and early 1980s was clear. Most of economics was based on models of perfect competition, where all agents are price-takers. An agent is a ‘price-taker’ if it believes that it can trade any quantity at the market price which it treats as given, or exogenous. Price-taking makes sense only when markets clear, and supply equals demand. If supply does not equal demand, then something has got to give because the chosen trades do not add up to zero. An alternative was needed. Up until then, various ad hoc assumptions had been made: the simplest was that wages and/or prices were simply assumed to be fixed (this was justified by the notion that the model was a short-run model). Another ad hoc fix was that the market was competitive but that the price cleared the market *ex ante*: the invisible auctioneer sets the price which he or she expects to clear the market before it opens. The basic and fundamental new Keynesian insight was that the assumption of price-taking behaviour had to be abandoned. Real agents such as firms, households or unions needed to be price-makers. But this meant that the notion of perfectly competitive equilibrium had to be abandoned: the alternative was going to be an imperfectly competitive equilibrium where (some) agents have market power. The classic imperfectly competitive equilibrium is pure monopoly: a monopolist can set any price he pleases, and will maximize profits. The monopolist equates marginal revenue with marginal cost: if he faces a downward sloping demand curve, this means that the monopolist will set a price above the competitive price and output will be lower than in the competitive equilibrium. While the firm increases its profits there is also a decline in consumer surplus and the total surplus (consumer plus producer) declines.
In the absence of market failure, the perfectly competitive equilibrium is Pareto optimal. If we are adopting a representative agent framework (as has most often been the case in macroeconomics since the neoclassical synthesis), Pareto optimality means that the equilibrium outcome maximizes the utility of the representative agent. Hence, if we look at small deviations from equilibrium (in terms of output, employment and so on), they will not have a first-order effect on welfare. This is an envelope theorem: the first-order conditions for optimality state that the first-order effect is zero at the optimum. With imperfect competition, by contrast, we start away from the optimum. Hence there are first-order effects of changes in output and employment: since the monopolist restricts output, an increase is good and a decrease bad. To many macroeconomists, this seems more plausible and common sense than the implication of the first welfare theorem that holds that, if one starts from the competitive equilibrium, increases and decreases in output and employment are both (slightly) bad.
The introduction of imperfect competition into a tractable general equilibrium framework (albeit a static one) was achieved by Oliver Hart (1982), who stressed the ‘Keynesian features’ of the model. However, Hart's was a real model without money: what was needed was to link this idea to nominal rigidity. It was

a few years later that the concept was taken up simultaneously in three papers: Akerlof and Yellen (1985), Mankiw (1985) and Parkin (1986). The new idea was that of ‘menu costs’, whereby there might be ‘costs’ to changing a price, which might be interpreted broadly as decision and implementation costs (the line taken by Akerlof and Yellen and interpreted as a sort of bounded rationality) or as literally the cost of implementing a price change (having new menus printed). This idea was not new: it was used by the (S,s) models of pricing with inflation developed in the 1970s by Sheshinski and Weiss (1977), and in some other papers in the non-macroeconomic literature.

The insight is that if a monopolist sets its price optimally, a small deviation from the optimum will have no first-order effect on *profits*. If there is a small but lump-sum cost of changing a price, then the effect of a price-setting monopolist to an increase in demand (or cost) might be to leave the price where it is, not to change it. Thus, even small menu costs can give rise to some nominal rigidity: because at the optimum there is no first-order effect on profits, the menu costs only have to overcome the smaller higher order effects. Thus began a theory of nominal rigidity based on monopolistic competition and menu costs. The nice feature of the model was that, although the menu costs could be small, the nominal rigidity they created would give rise to first-order welfare effects (since we start from a level of output and employment below equilibrium). Whilst the idea is very simple and powerful, it did alas run into a problem. In static models it is easy to use the menu-cost approach. However, macroeconomists in the 1980s were interested in dynamic models, and menu-cost models have proven very difficult to solve except under very special cases. For example, Caplin and Spulber (1987) looked at steady-state inflation and found that although the menu costs caused individual firms to have prices that remained fixed for a time, in aggregate prices they drifted up, with the aggregate money supply yielding the same aggregate output and inflation as with flexible prices. It has only been much later, since the late 1990s, that these models are beginning to be solved for interesting dynamic cases (under the new name ‘state-dependent pricing’ models).

However, the menu-cost idea spawned a large literature that looked into how certain features of the economy might allow even smaller menu costs to give rise to nominal rigidity. For example, Ball and Romer (1990) argued that if there were some real rigidity in the economy, it would interact with the nominal rigidity of prices, reducing the size of menu costs required to induce nominal rigidity. The real rigidity might take the form of an efficiency wage model, for example, where the equilibrium determined the real wage which was not sensitive to the level of economic activity. On the empirical level, Ball, Mankiw and Romer (1988) argued that the menu-cost theory had a clear prediction for the relation between inflation and the inflation–output trade-off. If steady-state inflation was higher, this would mean that for a given level of menu costs, firms would change prices more frequently (there is less nominal rigidity). This in turn would mean that changes in nominal demand would have less effect on output when inflation is higher. Thus the non-neutrality of money in the short run was higher in low-inflation economies than in high-inflation economies, which was confirmed in the data.

Whilst there has been until recently quite some difficulty in making state-dependent or menu-cost models tractable enough to model wage and price dynamics out of steady state, another class of models proved well suited to a dynamic setting. These were the *time-dependent* models of pricing, which focused on the notion of staggered wage- and price-setting: Taylor (1979) and Calvo (1983). Indeed, these two models have become the workhorses of the NNS framework. John B. Taylor's model focused on wage-setting: the empirical evidence suggests that many wage contracts take the form of a nominal

wage being set for a period of four quarters. However, wages in different sectors are negotiated at different times. It is usually assumed that there are four equally sized cohorts, one cohort resetting the wage each quarter. Whilst this framework does not explain why wage contracts last for a particular period, it does start out from firm empirical observation and works out the implications of this for the resultant process. What we find is that wages gradually adjust to their new steady state values. The reason for this is that when setting wages the current cohort is facing an aggregate price level partly determined by cohorts that have moved previously. At any one time, with four cohorts, three cohorts will not reset the wage: they reset their wages in the previous three quarters. When the union sets its wage, it looks at what the aggregate price level and demand will be over the period of the contract: in this sense the wage-setting rule is dynamic and forward looking. However, it is also looking back at the previous wages insofar as they are reflected in the current price. This results in a gradual adjustment of wages and prices in response to a nominal shock. Taylor (1999) provides a good survey of this approach.

Calvo's model of nominal rigidity is based on a constant hazard rate model: each period, the firm or union faces a given probability of resetting its price or wage. The expected duration of the price or wage when it is set is the reciprocal of the reset probability. When the firm sets its price it looks into the infinite horizon, and takes into account the future price with the probability that the current price being set will still be in force. Thus, if the reset probability is 0.25 per quarter, we will observe 25 per cent of firms resetting price in any one quarter. In setting the price, each firm expects that the price will last for four quarters, but there is en ever diminishing probability that it might last ever longer. If we look across all firms, the average contract length will be about twice the life expectation at birth (twice the life expectation at birth minus 1). Thus a reset probability of 0.25 implies an average lifetime of prices set by all firms across the economy of seven quarters (see Dixon and Kara, 2006). The firms choose an optimal price in a dynamic setting, but the setting itself leaves the fundamental probability of resetting the price unexplained. However, the model is highly tractable and has since become very popular.

Other new Keynesian themes

Whilst the theoretical microfoundation of nominal rigidity was the main theme of the new Keynesian economics, other themes aimed to establish the implications of imperfect competition and other market imperfections as an alternative equilibrium concept to perfect competition.

One theme that ran through the new Keynesian literature that did not involve nominal rigidity was the effect of imperfect competition on the government expenditure multiplier. Papers by Dixon (1987) and Mankiw (1988) found that in simple general equilibrium models an increase in the degree of imperfect competition reflected in a bigger markup of price over marginal cost meant that the balanced budget government multiplier was bigger. The intuition behind this result was that there was a profit feedback effect: as output increased, so did firms' profits, which were paid back to households in dividends, part of which were spent again, and so on. This feedback effect was bigger than the markup. In a constant returns to scale world, there were no profits in a perfectly competitive equilibrium, so the effect was completely absent. In a follow-up paper, Startz (1989) argued that whilst the Dixon–Mankiw result held in the short run with a fixed number of firms, in the long run free entry would eliminate profits and the relationship between profits and the multiplier would disappear. This argument turned out to be true in general only in the case of constant returns to scale. The point is that when you allow for a concave

production function with diminishing marginal product of labour, a second mechanism comes into effect: as employment rises, the real wage falls, which tends to reduce consumption. In the Walrasian case of perfect competition, the real wage effect always dominates the profit effect: the long-run multiplier with free entry is always greater than the short-run multiplier. It follows that if there is only a little imperfect competition, this will still be true, as shown in Dixon and Lawler (1996). Startz's result holds because with a constant marginal product of labour the real wage mechanism is absent and only the profit feedback is present.

It should be noted that the fiscal multiplier is still always less than unity. What is happening is that in equilibrium imperfect competition leads to lower real wages (the markup in the product market leads to real wages being below the marginal product). Households react to this by choosing more leisure and less consumption for any given utility level (the level of economic activity is below the perfectly competitive level). Now, an increase in government expenditure financed by a lump-sum tax makes the household worse off; so the household reacts by reducing its consumption and leisure (less leisure means working harder). The reason the short-run multiplier tends to be larger when there is more imperfect competition is that the equilibrium ratio of leisure to consumption is larger, so the effect of the tax on labour supply is larger, resulting in a bigger overall increase in labour supply and hence less crowding out of consumption. The mechanism underlying this is essentially a supply side effect, which is not exactly what some people might think of as ‘Keynesian’.

The notion of ‘coordination failure’ was also important in the new Keynesian thought. The idea arose out of the concept of strategic complementarity. Strategic complementarity occurs when the marginal benefit from the action of one agent is increasing in the level of activity chosen by other agents. Effectively, the reaction functions are upward sloping. Cooper and John (1988) applied this idea to several macroeconomic applications, including search models and demand spillovers in multi-sector economies, and the subsequent literature has applied this concept to almost any model with positive externalities. One interesting feature of the coordination failure approach is that there may be multiple equilibria: if this is so and the equilibria are symmetric the equilibria will be Pareto ranked. With positive externalities the high activity equilibria will Pareto dominate the low-level equilibria. The existence of multiple equilibria is not easy to establish: it requires as a necessary condition that the slope of the reaction function must be greater than 1 for some values in between the two symmetric equilibria.

In the labour market, there were several developments in the new Keynesian literature. Perhaps the most important was the development of efficiency wage models. Whilst the model of efficiency wages had a long pedigree, it was seen as a way of modelling how firms might set wages at a level different from the competitive level. In Shapiro and Stiglitz (1984), the internal monitoring problem faced by the firm is influenced by the level of unemployment, since the higher the level of unemployment the costlier it is for an employee to lose his or her job. Unemployment can therefore act as a disciplining device. This model predicts that firms will be forced to pay workers a higher wage when unemployment is lower, leading to a theoretical explanation of pro-cyclical wages.

The new neoclassical synthesis (NNS)

In the 1990s, the new Keynesian ideas become part of the NNS approach, which is a combination of the dynamic structures developed by the RBC theory with a nominal side to the economy, which is based on imperfect competition and nominal rigidity. One of the main contributions has been the new Keynesian

Phillips curve: this can be derived from both the Calvo and Taylor models of dynamic pricing (see Roberts, 1995). The equation relates current inflation to current output and expected inflation next period

$$\pi_t = \beta E_t \pi_{t+1} + \kappa y_t.$$

where inflation is π_t, the discount rate is β and output (deviation from capacity) is y_t. This differs from the traditional Phillips curve in which the expectation of current inflation appears on the right-hand side. The coefficient on the output gap is related to the probability the firm can reset its price, the discount rate and a parameter capturing the sensitivity of marginal cost to output. Empirically, the new Keynesian Phillips curve has not done very well. The evidence seems to support the idea that lagged inflation needs to be included as well (resulting in the so-called ‘hybrid Phillips curve’). This has led to the idea that indexation might be important: in the periods when firms cannot set prices or wages explicitly, they are updated by a ‘rule of thumb’ using last periods inflation rate (see Christiano, Eichenbaum and Evans, 2005) which results in a hybrid Phillips curve.

The Keynesian notion of demand management is very much at the centre of the analysis of monetary policy: the central bank is seen as using interest rate policy to stabilize the economy in two senses. The overall policy design should be to stabilize expectations and rule out explosive or indeterminate solutions: the possibility of economic turbulence caused by sunspot equilibria is seen as welfare reducing and is to be avoided (this is called *extrinisic uncertainty*). Thus policy should give rise to a unique rational expectations equilibrium path. In most models, a necessary condition for a unique equilibrium path is that the interest rate policy satisfies the *Taylor principle*, which states that if nominal inflation rises the central bank should raise the nominal interest rate by more, so that the real interest rate rises. Monetary policy should also be designed to stabilize the economy in response to real shocks, the *intrinsic* uncertainty facing the economy. This has been dubbed by some the ‘science of monetary policy’ (see Clarida, Gali and Gertler, 1999). Of course, the new Keynesian science is different from the old Keynesian art in that the interest rate is the only instrument and fiscal policy is reduced to providing a prudent and sustainable regime of expenditure and taxation. But the view is still Keynesian in that the economy needs and benefits from having an active monetary policy.

An evaluation

The most lasting legacy of the new Keynesian economics was to put imperfect competition and non-competitive models at the heart of macroeconomics. For a long time many economists had been impatient with the assumption of market clearing/demand equals supply as a basis for macroeconomics. However, a quest for a rigorous and consistent alternative was in place since Keynes's *General Theory* in 1936 raised more questions than it had answered. Whilst the book had given rise to the notion of using fiscal and monetary policy to stabilize the economy, this remained a practical art without a proper theoretical framework to underpin it. The macroeconomic theory developed was not consistent with standard microeconomics and was in this sense unsatisfactory. The real achievement of the new

Keynesian literature was to provide the theoretical alternative to demand and supply economics that was rigorous and microfounded. Economics has always been ideological as well as scientific. There are those free market ideologues who believe that the free market is almost always the best and that the state should intervene as little as possible in the market. There are also those who believe that although markets are pretty good at many things, they can also malfunction and so maybe there is a role for some sort of public policy. In macroeconomics this polarity was at its most obvious in the 1980s and 1990s. The real business cycle theorists used models with perfect markets and were largely of the ‘free-market’ variety of economists. The new Keynesian economics provided a rigorous alternative to the free-market perspective and as such has left a lasting legacy which we can see is firmly embedded in the way nominal rigidity is understood and monetary policy is practiced.

Further reading

Insofar as there is a defining book of the new Keynesian macroeconomics, it is Mankiw and Romer's two-volume collection (1991). Some good surveys were made in the early 1990s: Gordon (1990) is one; Silvestre (1993) focuses on the issue of imperfect competition; Dixon and Rankin (1994) focused more on the implications for macroeconomic policy issues. There was also a *Journal of Economic Perspectives* symposium on ‘Keynesian Economics Today’ in 1993 (volume 7, number 1) which takes a broader view of new and old Keynesian macroeconomics. On the NNS approach, the monetary policy aspects are well surveyed by Clarida, Gali and Gertler (1999), and for text book treatment of the modelling foundations turn to Walsh (2003, ch. 5) and Woodford (2003, ch. 3). There is also an excellent survey of several NNS models of nominal rigidity in Ascari (2003).

See Also

- microfoundations
- real business cycles
- real rigidities

Bibliography

Akerlof, G. and Yellen, J. 1985. A near-rational model of the business cycle with wage and price inertia. *Quarterly Journal of Economics* 100, 823–38.

Ascari, G. 2003. Price and wage staggering: a unifying framework. *Journal of Economic Surveys* 17, 511–40.

Ball, L., Mankiw, N.G. and Romer, D. 1988. The new Keynesian economics and the output-inflation trade-off. *Brookings Papers on Economic Activity* 1, 1–82.

Ball, L. and Romer, D. 1990. Real rigidities and the non-neutrality of money. *Review of Economic*

Studies 57, 179–98.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Caplin, A. and Spulber, D. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102, 703–26.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal rigidity and the dynamics effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–45.

Clarida, R., Gali, J. and Gertler, M. 1999. The science of monetary policy. *Journal of Economic Literature* 37, 1661–707.

Cooper, R. and John, A. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103, 441–63.

Dixon, H.D. 1987. A simple model of imperfect competition with Walrasian features. *Oxford Economic Papers* 39, 134–60.

Dixon, H.D. and Kara, E. 2006. How to compare Taylor and Calvo contracts: a comment on Michael Kiley. *Journal of Money, Credit and Banking* 38, 1119–26.

Dixon, H.D. and Lawler, P. 1996. Imperfect competition and the fiscal multiplier. *Scandinavian Journal of Economics* 98, 219–31.

Dixon, H.D. and Rankin, N. 1994. Imperfect competition and macroeconomics: a survey. *Oxford Economic Papers* 46, 171–99.

Gordon, R.J. 1990. What is new Keynesian economics. *Journal of Economic Literature* 28, 1115–71.

Hart, O. 1982. A simple model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97, 109–38.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. London: Macmillan.

Mankiw, N.G. 1985. Small menu costs and large business cycles: a macroeconomic model of monopoly. *Quarterly Journal of Economics* 100, 529–39.

Mankiw, N.G. 1988. Imperfect competition and the Keynesian cross. *Economics Letters* 26, 7–13.

Mankiw, N.G. and Romer, D. 1991. *New Keynesian Economics*. Cambridge, MA: MIT Press.

Parkin, M. 1986. The output-inflation trade-off when prices are costly to change. *Journal of Political Economy* 94, 200–24.

Roberts, J.M. 1995. New Keynesian economics and the Phillips curve. *Journal of Money, Credit and Banking* 27, 975–84.

Silvestre, J. 1993. The market power foundations of macroeconomic policy. *Journal of Economic Literature* 31, 105–41.

Shapiro, C. and Stiglitz, J. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Sheshinski, E. and Weiss, Y. 1977. Inflation and costs of price adjustment. *Review of Economic Studies* 44, 287–303.

Startz, R. 1989. Monopolistic competition as a foundation for Keynesian macroeconomic models. *Quarterly Journal of Economics* 104, 737–52.

Taylor, J. 1979. Staggered wage setting in a micro model. *American Economic Review* 69, 108–13.

Taylor, J. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of Macroeconomics*, vol. 1B, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Walsh, C.E. 2003. *Monetary Theory and Policy*, 2nd edn. Cambridge MA: MIT press.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Dixon, Huw David. "new Keynesian macroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000166> doi:10.1057/9780230226203.1184

The New Palgrave Dictionary of Economics Online

new open economy macroeconomics

Giancarlo Corsetti
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘New open economy macroeconomics’ (NOEM) refers to a body of literature embracing a new theoretical framework for policy analysis in open economy, aiming to overcome the limitations of the Mundell–Fleming model while preserving the empirical wisdom and policy friendliness of traditional analysis. NOEM contributions have developed general equilibrium models with imperfect competition and nominal rigidities, to reconsider conventional views on the transmission of monetary and exchange rate shocks; they have contributed to the design of optimal stabilization policies, identifying international dimensions of optimal monetary policy; and they have raised issues about the desirability of international policy coordination.

Keywords

disconnect puzzle; elasticity of substitution; exchange rate pass-through; expenditure switching effects; imperfect competition; inflationary bias; international policy coordination; international portfolio diversification; international transmission mechanism; Jensen's inequality; law of one price; local currency pricing; Mundell–Fleming model; neo-Wicksellian’ monetary economics; new Keynesian macroeconomics; new neoclassical synthesis; new open economy macroeconomics; nominal rigidities; optimal monetary policy; optimal tariffs; output gap; producer currency pricing; stabilization; staggered price setting; sticky prices; terms of trade; trade costs; uncertainty

Article

The new open economy macroeconomics (NOEM) is a leading development in international economics that began in the early 1990s. Its objective is to provide a new theoretical framework for open economy analysis and policy design, overcoming the limitations of the Mundell–Fleming model, while preserving the empirical wisdom and the close connection to policy debates of the traditional literature. The new framework consists of choice-theoretic, general-equilibrium models featuring nominal rigidities and imperfect competition in the markets for goods or labour. In this respect, the NOEM has close links with related agendas pursued in closed-economy macro, such as the ‘new neoclassical synthesis’ and the ‘neo-Wicksellian’ monetary economics. The assumption of imperfect competition is logically consistent with the maintained hypothesis that firms and workers optimally choose prices and wages subject to nominal frictions, as well as with the idea that output is demand-determined over some range in which firms (workers) can meet demand at non-negative profits (surplus).

NOEM models differ from the Mundell–Fleming approach in at least two notable dimensions. First, all agents are optimizing, that is, households maximize expected utility and managers maximize firms' value. The expected utility of the national representative consumer thus provides a natural welfare criterion for policy evaluation and design. Second, general-equilibrium analysis paves the way towards further integration of international economics as a unified field, bridging the traditional gap between open macroeconomic and trade theory.

From a historical perspective, NOEM was launched by Obstfeld and Rogoff (1995), although Svensson and van Wijnbergen (1989) had also worked out a model with NOEM features as an open economy development of Blanchard and Kiyotaki (1987).

A specific goal of the NOEM agenda is to achieve the standards of tractability which made traditional models so popular and long-lived among academics and policymakers. For instance, many contributions have adopted the model specification by Corsetti and Pesenti (2001), which admits a closed-form solution by virtue of some educated restrictions on

preferences (Tille, 2001, explains the relation of this model to Obstfeld and Rogoff, 1995). At the same time, the NOEM literature has promoted the construction of a new generation of large, multi-country quantitative models by international institutions and national monetary authorities. A leading example is the Global Economic Model (GEM) of the International Monetary Fund (see, for example, Laxton and Pesenti, 2003).

This article first introduces a stylized NOEM model. Based on this model, it then provides a short selective survey of the NOEM literature, and its main advances in the analysis of the international transmission mechanism and policy design in open economies.

1 A stylized NOEM model

To illustrate the basic features of NOEM models, highlighting similarities and differences with the Mundell–Fleming model, it is useful to refer to the model by Corsetti and Pesenti (2001; 2005a; 2005b) and Obstfeld and Rogoff (2000) (henceforth CP–OR). The economy consists of two countries, Home and Foreign, specialized in the production of one type of tradable goods, denoted H and F, respectively. Home consumption falls on both local goods and imports, that is, $C=C(C_H, C_F)$; the price level P includes both local goods and imports prices in Home currency, that is, $P=P(P_H, P_F)$. Preferences over local and imported goods are Cobb–Douglas with identical weights across countries: as the elasticity of substitution is equal to 1, any increase in domestic output is matched by a proportional fall in its price, so that terms-of-trade movements ensure efficient risk sharing. Furthermore, utility from consumption is assumed to be logarithmic, while disutility from labour $\mathscr{L}$ is linear.

Let μ index the Home monetary stance. Specifically, μ is the nominal value of the inverse of consumption marginal utility – for example, with log utility, $\mu = PC$. Whatever the instruments used by monetary authorities, μ indexes its ultimate effect on current spending. With competitive labour markets, the Households' optimality conditions imply that the nominal wage moves proportionally to μ, that is, $W = \mu$. Furthermore, abstracting from investment and government spending, μ indexes nominal aggregate demand. Similar definitions and conditions hold for the Foreign country, whose variables are denoted with a star, that is, $\mu^* = W^*$.

Let ε denote the nominal exchange rate, measured in units of Home currency per unit of Foreign currency. With perfect risk sharing, it is well known that the real exchange rate $\varepsilon P/P^*$ is equal to the ratio between the two countries' consumption marginal utilities (see Backus and Smith, 1993). Rearranging this condition, the nominal exchange rate is equal to the ratio of Home to Foreign monetary stance, that is, $\varepsilon = \mu/\mu^*$. A Home expansion depreciates ε.

Goods are supplied by a continuum of firms, each being the only producer of a differentiated variety of the national good. For simplicity, production is linear in labour. With nominal rigidities, managers optimally set prices as to maximize the market value of the firm. (Since households are assumed to own firms, the discount factor used in calculating the present value is the growth in the marginal utility of consumption.) In the CP–OR model, prices are preset for one period and marginal costs coincide with unit labour costs $W/Z = \mu/Z$. In this model, optimal pricing actually takes a form that is very similar to textbook monopoly pricing: Home firms selling in the domestic market set P_H by charging the optimal markup over *expected* marginal costs, that is:

$$P_H = \text{markup} \cdot E \overbrace{\left(\frac{\mu}{Z}\right)}^{\text{marginal cost}}$$

where E denotes conditional expectations. If prices were flexible, the above would hold with current instead of expected costs.

When modelling nominal rigidities in the exports market, however, the following issue arises: are export prices sticky in the currency of the producers or in the currency of the destination market? In the NOEM literature, this issue has fed an extensive debate on the international transmission mechanism and the design of optimal stabilization policies, discussed in detail in the next sections.

The equilibrium allocation can be characterized in terms of three equilibrium relationships, labelled AD, TT and NR. In Figure 1, these are drawn in the space ‘consumption’ vs. ‘labour’, C vs. $\mathscr{L}$. The horizontal AD locus represents the Home aggregate demand in real terms, given by the ratio of the monetary stance to the price level: $C = \mu/P$. The upward-sloping TT locus shows the level of consumption that Home agents obtain (at market prices) in exchange for $\mathscr{L}$ units of labour. The slope of the TT locus depends on the (exogenous) productivity level Z, and the (endogenous) price of domestic GDP ($Y=Z\mathscr{L}$), in terms of domestic consumption τ, that is, $C = \tau \cdot Z \cdot \mathscr{L}$. Since agents consume both local goods and imports, τ rises with an improvement in the terms of trade of the Home country, conventionally defined as the price of imports in terms of exports. The vertical NR locus marks the equilibrium employment in the flexible prices (or natural rate) allocation, $\mathscr{L}^{flex}$. Because of firms' monopoly power, $\mathscr{L}^{flex}$ is inefficiently low. To stress this point, Figure 1 includes the indifference curve passing through the equilibrium point E, where it crosses the TT locus from above: with monopolistic distortions, the marginal rate of substitution between labour

and consumption differs from the marginal rate of transformation.

Figure 1

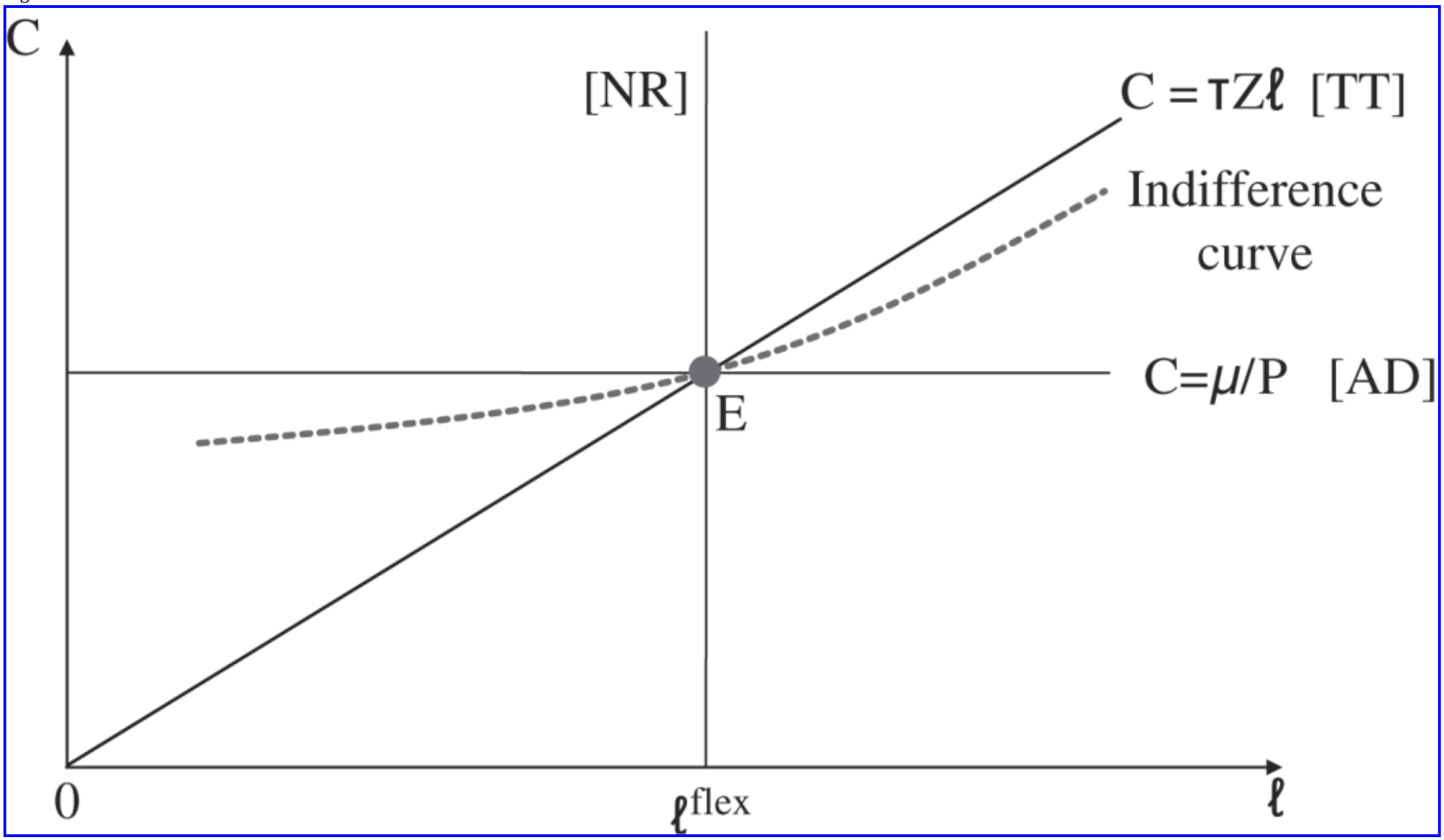

With flexible prices, the macroeconomic equilibrium is determined by the NR locus and the TT locus. For a given μ, nominal prices adjustment ensures that demand is in equilibrium. With nominal rigidities, the equilibrium is instead determined by the AD locus and the TT locus. Depending on the level of demand, employment may fall short of or exceed the natural rate, opening employment and output gaps proportional to ($\ell^{flex} - \ell$).

2 The international transmission mechanism and the allocative properties of the exchange rate

According to traditional open macroeconomic models, exchange rate movements play the stabilizing role of adjusting international relative prices in response to shocks, when frictions prevent or slow down price adjustment in the local currency. At the heart of this view is the idea that nominal depreciation transpires into real depreciation, making domestic goods cheaper in the world markets, hence redirecting world demand towards them: exchange rate movements therefore have ‘expenditure switching effects’.

Consistent with this view, NOEM contributions after Obstfeld and Rogoff (1995) draws on the Mundell–Fleming and Keynesian tradition, and posits that export prices are sticky in the currency of the producers. Thus the nominal import prices in local currency move one-to-one with the exchange rate. This hypothesis is commonly dubbed ‘producer currency pricing’ (PCP).

Under PCP firms preset P_H and P_F^*, so the Home country's terms of trade $\varepsilon P_F^*/P_H$ deteriorate with unexpected depreciation. Moreover, as long as demand elasticities are identical in all markets, firms have no incentive to price discriminate: the price of exports obeys the law of one price, that is, $P_H^* = P_H/\varepsilon$ and $P_F = \varepsilon P_F^*$.

Monetary shocks have two distinct effects on the Home allocation and welfare. Expansions raise demand and output: because of monopolistic distortions in production, positive nominal shocks benefit domestic consumers by raising output towards its efficient (competitive) level. However, currency depreciation also raises the relative price of Foreign goods, reducing the real income of domestic consumers. In terms of Figure 1, monetary expansions shift the AD locus upward and, due to currency depreciation, cause the TT locus to rotate clockwise. The new equilibrium may lie either above or below the indifference curve passing through E, the initial equilibrium. In other words, Home welfare may rise or fall, depending on the relative magnitude of monopoly power in production, vis-à-vis the terms-of-trade externality, in turn related to openness and the degree of substitutability between Home and Foreign tradables. (The size of the monetary shock also matters: by the same argument, by the theory of optimal tariffs a country never gains from monetary shocks which are large enough to raise output up to its competitive – Pareto-efficient – level.)

A noteworthy implication for policy analysis is that, in relatively open economies where terms-of-trade distortions are strong, benevolent policymakers may derive short-run benefits by implementing surprise monetary contractions, which appreciate the Home currency and boost the purchasing power of Home consumers. In these economies, monetary policy can have a deflationary bias.

In the Foreign country, welfare spillovers of a Home monetary expansion are unambiguously positive. Foreign consumers benefit from the terms-of-trade movement, which raises their income in real terms: the Foreign TT rotates counterclockwise. In addition, cheaper imports reduce inflation, raising aggregate demand for a given monetary stance μ^*: the Foreign AD shifts upward.

The high elasticity of import prices to the exchange rate underlying the above analysis is, however, at odds with a large body of empirical studies showing that the exchange rate pass-through on import prices is far from complete in the short run, and deviations from the law of one price are large and persistent (see, for example, Engel and Rogers, 1996; Goldberg and Knetter, 1997; Campa and Goldberg, 2005). This evidence has motivated a thorough critique of the received wisdom on the expenditure switching effects of the exchange rate. Specifically, Betts and Devereux (2000) and Devereux and Engel (2003), among others, posit that firms preset prices in the currency of the markets where they sell their goods. This assumption, commonly dubbed ‘local currency pricing’ (LCP), attributes local currency price stability of imports mainly to nominal frictions, with far-reaching implications for the role of the exchange rate in the international transmission mechanism (see Engel, 2003).

To the extent that import prices are sticky in the local currency, a Home depreciation does not affect the price of Home goods in the world markets; hence, it has no expenditure switching effects. Instead, it raises *ex post* markups on Home exports: at given marginal costs, revenues in domestic currency from selling goods abroad rise. In contrast with the received wisdom, nominal depreciation strengthens a country's terms of trade: if P_F and P_H^* are preset during the period, the Home terms of trade $P_F/\varepsilon P_H^*$ improve when the Home currency weakens. In Figure 1, with LCP, a Home monetary expansion shifts aggregate demand AD upward and rotates the TT counterclockwise.

It follows that monetary authorities cannot derive short-run welfare benefits from surprise contraction. As currency depreciation improves the terms of trade, the inflationary bias in policymaking is even stronger than in a closed economy.

International spillovers from Home monetary expansions are detrimental to Foreign welfare. If prices in local currency remain constant, a Home expansion does not at all affect the aggregate demand in the Foreign country. Yet the adverse terms-of-trade movement forces Foreign agents to work more to sustain an unchanged level of consumption: for a given AD, the TT locus rotates clockwise.

An interesting case with asymmetric transmission is one in which the prices of exports are all preset in one currency, so that Home firms adopt PCP while Foreign firms adopt LCP (see, for example, Devereux, Engel and Tille, 2003).

While the NOEM literature has encompassed additional real and financial aspects in the analysis of the transmission mechanism, the PCP versus LCP debate identifies essential building blocks of optimal stabilization policy.

3 International dimensions of optimal monetary policies

A defining question of open economy macroeconomics is whether monetary and fiscal policy should react to international variables, such as the exchange rate or the terms of trade, beyond the influence that these variables have on the domestic output gap (for example, via external demand) and domestic inflation (for example, via import prices). This is a research area where choice-theoretic NOEM models have comparative advantages over the traditional literature. Indeed early NOEM contributions have established a set of original and provocative results, setting benchmarks for further analytical and quantitative studies.

To account for these results, consider the stabilization problem in a CP–OR economy with country-specific productivity uncertainty. In a flexible price environment (corresponding to the long run of the CP–OR model), a positive productivity shock in the Home country causes the world price of Home goods to fall. This raises both domestic and foreign demand for Home output, and worsens the Home terms of trade. With sticky prices, by contrast, unexpected gains in productivity simply translate into lower employment: given μ and μ^* (hence given the exchange rate), current demand is satisfied with a lower labour input. (In Figure 1, a higher Z rotates the TT locus counterclockwise. With the AD and the TT loci held fixed, the equilibrium employment is below the natural rate. A fall in domestic prices would shift the AD locus up, while offsetting part of the rotation of the TT locus. The flexible price equilibrium always lies on the NR locus.)

However, under the hypothesis of PCP, it is easy to see that monetary policy in a sticky-price environment can support the flexible price allocation. Posit that monetary rules satisfy $\mu = \Gamma Z$, where Γ denotes a (possibly time-varying) variable indexing the level of nominal variables in the Home country. When such rules are implemented, any gain in productivity is matched by a proportional expansion of the monetary stance, which raises Home demand and depreciates the Home currency. Marginal costs remain constant in nominal terms (since $\mu/Z = \Gamma$): hence product prices in domestic currency would remain fixed even if there were no nominal rigidities. At the same time, however, exchange rate movements adjust international relative prices, as monetary policy moves ε in proportion to productivity changes.

A first benchmark result is that, in economies with the CP–OR features, monetary policy rules supporting the flexible price allocation are optimal: no rule welfare-dominates complete marginal cost and output gap stabilization. This is true under different assumptions regarding nominal rigidities, including staggered price setting and partial adjustment (see, for example, Clarida, Galí and Gertler, 2002). Optimal monetary rules are completely ‘inward-looking’: welfare-maximizing central banks stabilize the GDP deflator while letting the consumer price index (CPI) fluctuate with movements in the relative price of imports. There is no need for monetary policies to react to international variables.

The result that monetary rules supporting a flexible price allocation are optimal, however, does not hold in general. In the presence of multiple distortions monetary authorities are generally able to exploit nominal rigidities and improve welfare relative to such allocation (Benigno and Benigno, 2003; Corsetti and Dedola, 2005). Yet, holding PCP, it is unclear whether and under which conditions deviating from full domestic stabilization could yield significant welfare gains.

A second result concerns the costs of inefficient stabilization. The New Keynesian theory has emphasized welfare costs from relative price dispersion when private pricing decisions are not synchronized (see, for example, Galí and Monacelli, 2003). Early NOEM contributions have instead pioneered the analysis of the effect of uncertainty on the level of prices and economic activity. A simple example illustrates this point. Suppose that monetary policy responds to productivity shocks according to rule: $\mu = \Gamma Z^{\gamma}$. When $\gamma < 1$, marginal cost uncertainty due to insufficient stabilization implies $E(\mu/Z) = \Gamma E(1/Z^{1-\gamma}) > \Gamma$: by a straightforward application of Jensen's inequality, expected marginal costs are higher than under complete stabilization. Higher costs transpire into higher prices both in nominal terms and relative to wages, reducing the average supply of domestic goods, thus exacerbating monopolistic distortions in the economy (see, for example, Sutherland, 2005, and Kollmann, 2002, for a quantitative assessment).

Similar effects, with potentially stronger welfare implications, are caused by a noisy conduct of monetary policy and exchange rate variability (Obstfeld and Rogoff, 1998). Notably, Broda (2006) provides evidence consistent with the (NOEM) prediction that incomplete stabilization *and* monetary/exchange rate noise transpire into higher price levels and real appreciation.

A third result, derived on the assumption of LCP, defines a clear-cut argument in favour of policies with an international dimension. To the extent that exporters' revenues and markups are exposed to exchange rate uncertainty, firms' optimal pricing strategies internalize the monetary policy of the importing country. In the CP–OR model, for instance, Foreign firms optimally preset the price of their goods in the Home market P_F by charging the equilibrium markup over expected marginal costs *evaluated in Home currency*, that is,

$$P_F = \text{markupg} \cdot E\left(\epsilon \frac{\mu^*}{Z^*}\right) = \text{markupg} \cdot E\left(\frac{\mu}{Z^*}\right).$$

Clearly, the price of Home imports depends on the joint distribution of Home monetary policy and Foreign productivity shocks.

Suppose that Home monetary authorities ignore the influence of their decisions on the price of Home imports. For the reason discussed above, import prices will tend to be inefficiently high. On the other hand, if Home monetary authorities want to stabilize Foreign firms' marginal costs, they can only do so at the cost of raising costs and markup uncertainty for Home producers, resulting in higher Home good prices. It follows that, to maximize Home welfare, Home policymakers should optimally trade off the stabilization of marginal costs of all producers (domestic and foreign) selling in the Home markets.

When foreign firms' profits are exposed to exchange rate uncertainty, optimal monetary rules are no longer inward-looking. The importance of Foreign shocks in the conduct of monetary policy depends on the degree of openness of the economy, measured by the overall share of imports in the CPI (see Corsetti and Pesenti, 2005a, and Sutherland, 2005, for a discussion of intermediate degrees of pass-through, and Smets and Wouters, 2002, and Monacelli, 2005, for models with staggered price setting).

Notably, the case for an international dimension in monetary policy described above transpires into limited exchange rate variability. Since with LCP optimal monetary policies respond to both domestic and foreign shocks, national monetary stances tend to be more correlated than in the case of inward-looking stabilization of output gaps. This implies lower exchange rate volatility. In the baseline CP–OR model, the optimal policy rules actually prevent *any* short-run fluctuations of the exchange rate, a point stressed by Devereux and Engel (2003). But this exact result holds only when the weights of Home and Foreign goods in final expenditure are assumed to be identical across countries: Home and Foreign monetary authorities de facto stabilize the same weighted average of marginal costs. The presence of non-traded goods or some Home bias in consumption would obviously imply asymmetries in the optimal monetary stances, which would be incompatible with a fixed exchange rate (Duarte and Obstfeld, 2007; Corsetti, 2006). Even if, with LCP, exchange rate variability does not perform any role in adjusting international prices, a fixed rate regime would impose unwarranted constraints on the efficient conduct of monetary policy.

A fourth result concerns the desirability of international policy coordination. Leading NOEM contributions have fed considerable scepticism on this issue. At the core of this scepticism is the disappointing quantitative assessment of welfare gains from coordination. By using the CP–OR model, for instance, it is possible to build economies with either PCP or LCP behaviour, where optimal monetary rules are identical whether national policymakers act independently or cooperatively (maximizing an equally weighted sum of national welfare functions). When this exact result breaks down (depending on the elasticity of substitution between Home and Foreign tradables, and/or sector-specific shocks in the presence of non-tradables), gains from coordination usually remain quite small (see, for example, Pappa, 2004; Benigno and Benigno, 2006).

The lesson from the NOEM literature, stressed by Obstfeld and Rogoff (2002), is a new welfare-based argument against coordination: once policymakers independently pursue efficient stabilization policies in their own country (that is, they ‘keep their house in order’), the room for improving welfare through cooperation is quite limited (see Canzoneri, Cumby and Diba, 2005, for a discussion).

The results reviewed above were first derived in highly stylized economies. A critical question directing current NOEM research is whether they would still hold in richer models with good quantitative performance.

4 Challenges to the NOEM literature

The above debate on the role of exchange rate in the international transmission has motivated further empirical and theoretical work on market segmentation along national borders and on its implications for international macroeconomic adjustment. As stressed by Obstfeld and Rogoff (2001), despite the ongoing process of real and financial globalization, frictions and imperfections appear to keep national economies ‘insular’.

An important issue is the extent to which the evidence of local currency price stability of imports can be explained by nominal rigidities. It is well understood that the low elasticity of import prices with respect to the exchange rate is in large part due to the incidence of distribution (Burstein, Eichenbaum and Rebelo, 2007). Several macro and micro contributions have emphasized the role of optimal destination-specific markup adjustment by monopolistic firms depending on market structure (Dornbusch, 1987; Goldberg and Verboven, 2001), or vertical interactions between producers and retailers (Corsetti and Dedola, 2005).

The main point is that low pass-through is not necessarily incompatible with expenditure switching effects (see, for example, Obstfeld, 2002). In this respect, Obstfeld and Rogoff (2000) emphasizes that, in the data (and consistent with the received wisdom), nominal depreciation does tend to be associated with deteriorating terms of trade. This piece of evidence clearly sets an empirical hurdle for LCP models, if we assume a high degree of price stickiness in local currency (see Corsetti, Dedola and Leduc, 2005, for a quantitative assessment). Interestingly, estimates of LCP models downplaying price discrimination, distribution and other real determinants of incomplete pass-through predict that the degree of price stickiness is implausibly higher for imports than for domestic goods, a result suggesting model misspecification (see, for example, Lubik and Schorfheide, 2006).

Moreover, the currency denomination of exports prices should be treated as an endogenous choice by profit maximizing firms (see, for example, Bacchetta and Van Wincoop, 2005; Devereux, Engel and Storgaard, 2004). To appreciate the contribution by the NOEM literature on this issue, recall that, in the CP–OR model above, expansionary monetary shocks unrelated to productivity raise nominal wages and marginal costs while depreciating the currency. For a firm located in a country with noisy monetary policy, pricing its exports in foreign currency (that is, choosing LCP) is therefore quite attractive: it ensures that revenues from exports in domestic currency will tend to rise in parallel with nominal marginal costs, with stabilizing effects on the markup. This may help explain why exporters from emerging markets with relatively unstable domestic monetary policies prefer to price their exports to advanced countries in the importers' currency. The same argument, however, suggests that LCP is not necessarily optimal for exporters producing in countries where monetary policy systematically stabilizes marginal costs (see Goldberg and Tille, 2005, for empirical evidence).

New waves of studies are building models with trade costs where goods tradability is endogenous, and/or new varieties are created at business cycle frequencies. Trade and transaction costs are also at the heart of recent attempts to integrate current account and macroeconomic dynamics with international portfolio diversification in a unified analytical framework (see, for example, Devereux and Sutherland, 2007).

The discussion above is far from exhausting the range of topics and issues analysed by the NOEM literature, which has marked a radical change of paradigm in international macroeconomics. Many authors have undertaken a systematic reconsideration of classical themes in the new framework. A partial list of themes includes overshooting (for example,

Hau, 2000); current account, debt and exchange rate dynamics (Cavallo and Ghironi, 2002; Ganelli, 2005; Ghironi, 2006); exchange rate uncertainty and trade (Bacchetta and Van Wincoop, 2000); and fiscal policy (Adao, Horta Correia and Teles, 2006). An important set of papers delves into empirical analysis of NOEM models (for example, Bergin, 2003; Lubik and Schorfheide, 2006).

Yet most NOEM contributions so far specify models which predict a counterfactually high degree of consumption risk sharing: even when financial markets are incomplete, intertemporal trade and terms-of-trade spillovers ensure that the consumption risk of productivity shocks is contained, and the market allocation is not too distant from the efficient one (see, for example, Chari, Kehoe and McGrattan, 2002). Not only this is inconsistent with a large body of evidence (see Backus and Smith, 1993); most crucially, a counterfactually high degree of risk sharing built in NOEM models may limit their capacity to comprehend significant cross-border spillovers and policy trade-offs. Similarly, in most models the exchange rate is tightly related to fundamentals, at odds with a large body of evidence showing that the relation between the exchange rate and virtually any macroeconomic aggregate is exceedingly weak – the so-called disconnect puzzle.

Further progress in these areas is crucial towards the fulfilment of the NOEM research agenda.

See Also

- international finance
- international policy coordination
- international real business cycles
- nominal exchange rates
- price discrimination (empirical studies)

Bibliography

Adao, B., Horta Correia, M.I. and Teles, P. 2006. On the relevance of exchange rate regimes for stabilization policy. Discussion Paper No. 5797, CEPR.

Bacchetta, P. and Van Wincoop, E. 2000. Does exchange rate stability increase trade and welfare? *American Economic Review* 50, 1039–109.

Bacchetta, P. and Van Wincoop, E. 2005. A theory of the currency denomination of international trade. *Journal of International Economics* 67, 295–319.

Backus, D.K. and Smith, G.W. 1993. Consumption and real exchange rates in dynamic economies with non–traded goods. *Journal of International Economics* 35, 297–316.

Benigno, G. and Benigno, P. 2003. Price stability in open economy. *Review of Economic Studies* 70, 743–64.

Benigno, G. and Benigno, P. 2006. Designing targeting rules for international monetary policy cooperation. *Journal of Monetary Economics* 53, 473–506.

Bergin, P. 2003. Putting the new open economy macroeconomics to a test. *Journal of International Economics* 60, 3–34.

Betts, C. and Devereux, M. 2000. Exchange rate dynamics in a model of pricing to market. *Journal of International Economics* 50, 215–44.

Blanchard, O. and Kiyotaki, N. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77, 647–66.

Broda, C. 2006. Exchange rate regimes and national price levels. *Journal of International Economics* 70, 52–81.

Burstein, A., Eichenbaum, M. and Rebelo, S. 2007. Modeling exchange rate pass-through after large devaluations. *Journal of Monetary Economics* 54, 346–68.

Campa, J. and Goldberg, L. 2005. Exchange rate pass through into import prices. *Review of Economics and Statistics* 87, 679–90.

Canzoneri, M.B., Cumby, R. and Diba, B. 2005. The need for international policy coordination: what's old, what's new, what's yet to come? *Journal of International Economics* 66, 363–84.

Cavallo, M. and Ghironi, F. 2002. Net foreign assets and the exchange rates: redux revived. *Journal of Monetary Economics* 49, 1057–97.

Chari, V.V., Kehoe, P.J. and McGrattan, E. 2002. Can sticky prices generate volatile and persistent real exchange rates? *Review of Economic Studies* 69, 633–63.

Clarida, R., Galí, J. and Gertler, M. 2002. A simple framework for international policy analysis. *Journal of Monetary Economics* 49, 879–904.

Corsetti, G. 2006. Openness and the case for flexible exchange rates. *Research in Economics* 60, 1–21.

Corsetti, G. and Dedola, L. 2005. A macroeconomic model of international price discrimination. *Journal of International Economics* 67, 129–56.

Corsetti, G., Dedola, L. and Leduc, S. 2005. DSGE models of high exchange rate volatility and low pass through. Discussion Paper No. 5377, CEPR.

Corsetti, G. and Pesenti, P. 2001. Welfare and macroeconomic interdependence. *Quarterly Journal of Economics* 116, 421–46.

Corsetti, G. and Pesenti, P. 2005a. International dimension of optimal monetary policy. *Journal of Monetary Economics* 52, 281–305.

Corsetti, G. and Pesenti, P. 2005b. The simple geometry of transmission and stabilization in closed and open economies. Working Paper No. 11341. Cambridge, MA: NBER.

Devereux, M. and Engel, C. 2003. Monetary policy in open economy revisited: price setting and exchange rate flexibility. *Review of Economic Studies* 70, 765–83.

Devereux, M., Engel, C. and Storgaard, P. 2004. Endogenous exchange rate pass-through when nominal prices are set in advance. *Journal of International Economics* 63, 263–91.

Devereux, M., Engel, C. and Tille, C. 2003. Exchange rate pass-through and the welfare effects of the euro. *International Economic Review* 44, 223–42.

Devereux, M. and Sutherland, A. 2007. Monetary policy and portfolio choice in an open economy macro model. *Journal of the European Economics Association* 5, 491–499.

Dornbusch, R. 1987. Exchange rates and prices. *American Economic Review* 77, 93–106.

Duarte, M. and Obstfeld, M. 2007. Monetary policy in the open economy revisited: the case for exchange-rate flexibility restored. Working paper. Online. Available at http://elsa.berkeley.edu/~obstfeld/DO_JIMF-2.pdf, accessed 29 November 2007.

Engel, C. 2003. Expenditure switching and exchange rate policy. *NBER Macroeconomics Annual 2002* 17, 231–72.

Engel, C. and Rogers, J. 1996. How wide is the border? *American Economic Review* 86, 1112–25.

Galí, J. and Monacelli, T. 2003. Monetary policy and exchange rate volatility in a small open economy. *Review of Economic Studies* 72, 707–34.

Ganelli, G. 2005. The new open economy macroeconomics of government debt. *Journal of International Economics* 65, 167–84.

Ghironi, F. 2006. Macroeconomic interdependence under incomplete markets. *Journal of International Economics* 76, 428–50.

Goldberg, L. and Tille, C. 2005. Vehicle currency use in international trade. Staff Report No. 200, Federal Reserve Bank of New York.

Goldberg, P.K. and Knetter, M.M. 1997. Goods prices and exchange rates: what have we learned? *Journal of Economic Literature* 35, 1243–72.

Goldberg, P.K. and Verboven, F. 2001. The evolution of price dispersion in the European car market. *Review of Economic Studies* 68, 811–48.

Hau, H. 2000. Exchange rate determination: the role of factor price rigidities and nontradeables. *Journal of International Economics* 50, 421–47.

Kollman, R. 2002. Monetary policy rules in the open economy: effects on welfare and business cycles. *Journal of Monetary Economics* 49, 989–1015.

Laxton, D. and Pesenti, P. 2003. Monetary rules for small, open, emerging economies. *Journal of Monetary Economics* 50, 1109–46.

Lubik, T. and Schorfheide, F. 2006. A Bayesian look at new open economy macroeconomics. *NBER Macroeconomics Annual 2005* 20, 313–66.

Monacelli, T. 2005. Monetary policy in a low pass-through environment. *Journal of Money Credit and Banking* 6, 1047–66.

Obstfeld, M. 2002. Inflation-targeting, exchange rate pass-through, and volatility. *American Economic Review* 92, 102–7.

Obstfeld, M. and Rogoff, K. 1995. Exchange rate dynamics redux. *Journal of Political Economics* 102, 624–60.

Obstfeld, M. and Rogoff, K. 1998. Risk and exchange rates. In *Contemporary Economic Policy: Essays in Honor of Assaf Razin*, ed. H. Helpman and E. Sadka. Cambridge and New York: Cambridge University Press.

Obstfeld, M. and Rogoff, K. 2000. New directions for stochastic open economy models. *Journal of International Economics* 50, 117–53.

Obstfeld, M. and Rogoff, K. 2001. The six major puzzles in international finance: is there a common cause? *NBER Macroeconomics Annual 2000* 15, 339–90.

Obstfeld, M. and Rogoff, K. 2002. Global implications of self-oriented national monetary rules. *Quarterly Journal of Economics* 117, 503–36.

Pappa, E. 2004. Do the ECB and the Fed really need to cooperate? Optimal monetary policy in a two-country world. *Journal of Monetary Economics* 51, 753–79.

Smets, F. and Wouters, R. 2002. Openness, imperfect exchange rate pass-through and monetary policy. *Journal of Monetary Economics* 49, 947–81.

Sutherland, A. 2005. Incomplete pass-through and the welfare effects of exchange rate variability. *Journal of International Economics* 65, 375–400.

Svensson, L.E.O. and van Wijnbergen, S. 1989. Excess capacity, monopolistic competition and international transmission of monetary disturbances. *Economic Journal* 99, 785–805.

Tille, C. 2001. The role of consumption substitutability in the international transmission of shocks. *Journal of International Economics* 53, 421–44.

How to cite this article

Corsetti, Giancarlo. "new open economy macroeconomics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave

Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000060> doi:10.1057/9780230226203.1185

The New Palgrave Dictionary of Economics Online

Newcomb, Simon (1835 – 1909)

Milton Friedman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

charitable giving; equation of exchange; Newcomb, S.; quantity theory of money; wages fund

Article

Newcomb entitled his autobiography *Reminiscences of an Astronomer* (1903), devoted only 10 pages out of 416 to his activities in economics, and remarked: ‘Being sometimes looked upon as an economist, I deem it not improper to disclaim any part in the economic research of today’ (p. 408). The 1913 *Encyclopaedia Britannica* in a lengthy article describes him as ‘one of the most distinguished astronomers of his time’ and includes one sentence, ‘He also wrote on questions of finance and economics.’ The 1970 edition of the *Encyclopaedia* describes him as ‘the greatest American astronomer of the 19th century’ and repeats the remark that ‘he wrote on finance and economics’.

Those may well be correct evaluations of the *relative* importance of Newcomb's work in astronomy and economics. Yet they give a wholly misleading impression of the *absolute* importance of his contribution to economics. He wrote two classics of economic science: *A Critical Examination of Our Financial Policy during the Southern Rebellion* (1865) and *Principles of Political Economy* (1885). The first ‘contains the most sophisticated, original, and profound analysis of the theoretical issues involved in Civil War finance that we have encountered, regardless of date of publication’ (Friedman and Schwartz, 1963, p. 18). The second contains what Irving Fisher, in his obituary note on Newcomb, regarded as ‘his chief and most fruitful contribution to economic science’, namely.

> the distinction he applied in particular to what he called ‘societary circulation’, or the equation of exchange between money and goods. So far as I am aware, he was the first definitely to enunciate this equation, expressing the fact that the quantity of money multiplied by its velocity of circulation is equal to price-level multiplied by volume of business transactions. This equation, with due amplifications, represents the so-called ‘quantity theory of money’ in its highest form. He also employed this same distinction … to expose the fallacy of ‘the wage-fund’. (Fisher, 1909, p. 642)

Another notable item in the *Principles* is its final chapter, ‘Of Charitable Effort’, an economic analysis of charity that is highly relevant to modern problems of the welfare state, in part because Newcomb writes about currently sensitive issues with a frankness and plainness that is absent from contemporary literature.

In addition to these two books, Newcomb published well over 50 popular magazine articles on economic issues, some of which formed the basis for two popular books: *The ABC of Finance* (1877) and *A Plain Man's Talk on the Labor Question* (1886). The latter, which was still in print in the 1980s, remains today an extraordinarily persuasive and effective exposition of the basic principles of a market economy and the effects of labour union activity on the interests of the worker.

Had these items constituted the whole of Newcomb's canon, instead of only a surprising few out of a total of well over 500 items, including not only major works in astronomy but also textbooks in mathematics, important contributions to statistics, and even a science fiction novel, he would have come to be regarded as one of the leading American economists of the 19th century. Irving Fisher noted that one reason ‘his economic writings did not attract the attention among economists which they deserved … is that … once a man's name becomes associated with a particular department of knowledge like astronomy, any attempts to contribute to other departments encounter a prejudice which it is difficult to overcome’ (Fisher, 1909, p. 641). Perhaps also it was not irrelevant that he was completely self-taught in economics, as in much else.

Newcomb was born on 12 March 1835 in a small town in Nova Scotia, son of an impecunious country school teacher. He died on 11 July 1909, and was buried with all the military pomp due to his congressionally conferred rank of Rear Admiral, a remarkable transition due entirely to Newcomb's own talents, character and persistence. Though his only formal schooling consisted of occasional attendance at his father's schools, he early displayed unusual intellectual interests and capacities. As what seemed in that remote region and time the only avenue of further instruction, he was apprenticed at the age of 16 to a herb doctor for a five-year period. The doctor turned out to be a quack who treated Simon as a slave, and provided no training whatsoever.

After two years, Simon finally summoned up the courage to run away, hiding in the woods as his erstwhile master sought to track him down. He joined his father, who had gone to New England after the death of Simon's mother, and father and son made their way to the Eastern Shore of Maryland, where both found employment as country teachers. Despite being entirely self-taught, Simon started to write articles on mathematical and astronomical subjects, one of which he sent to Professor Henry, Secretary of the Smithsonian Institute. This led to Professor Henry's becoming interested in Newcomb and ultimately recommending him for a job as a ‘computer’ at the Nautical Almanac in Cambridge, Massachusetts – an event which Newcomb described as ‘an epoch – an entrance into a new world’. Employment at the Nautical Almanac enabled him to take courses at the Lowell Scientific School of Harvard University, where he received a degree in 1857. In 1861 he received a commission as professor of mathematics at the US Naval Observatory; in 1877 he was appointed superintendent of the Nautical Almanac, and in 1884 professor of mathematics and astronomy at the Johns Hopkins University, a position he held concurrently with his posts at the Naval Observatory and the Nautical Almanac.

In addition to his prodigious written output, Newcomb served for many years as editor of the *American Journal of Mathematics* and was active in the National Academy of Sciences, and the American Association for the Advancement of Science, of which he was president in 1877. Truly a Renaissance

man.

See Also

- equation of exchange
- quantity theory of money

Bibliography

Archibald, R.C. 1924. Simon Newcomb 1835–1909. Bibliography of his life and work. *Memoirs of the National Academy of Sciences* 17, 19–69.

Campbell, W.W. 1924. Biographical memoir: Simon Newcomb. *Memoirs of the National Academy of Sciences* 17, 1–18.

Fisher, I. 1909. Obituary. Simon Newcomb. *Economic Journal* 19, 641–4.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States*. Princeton: Princeton University Press.

Stigler, S.M. 1973. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association* 68, 872–9.

How to cite this article

Friedman, Milton. "Newcomb, Simon (1835–1909)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000065> doi:10.1057/9780230226203.1186

The New Palgrave Dictionary of Economics Online

Newmarch, William (1820 – 1882)

D.P. O'Brien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

monetary theory; prices; Newmarch, William; Tooke, Thomas

Article

Newmarch was born in Thirsk, Yorkshire, and died in Torquay. He had little formal education, but rose from the position of bank clerk to be a force in the City of London, being manager of Glyn Mills from 1862 to 1881. An excitable but effective speaker, he was a member of the Political Economy Club from 1852 (Treasurer 1855–82), and a considerable force in the (Royal) Statistical Society of which he was Secretary 1854–62 and President 1869–71. He was also elected a Fellow of the Royal Society and wrote for the *Morning Chronicle* and the *Economist.*

Newmarch was important as the principal author of the last two volumes of Tooke's monumental *History of Prices* (though, oddly, this publication, unlike Newmarch's own work in the *Economist*, did not employ index numbers), and as an economist in his own right for exploring the effects of the gold discoveries, public debt, and questions of monetary control. He was one of the leading opponents of the Currency School and the Bank Act of 1844, arguing that causality ran from prices to note issue, so long as the notes were convertible. He believed that monetary base control, as embodied in the 1844 Act, was not only ineffective – following the work of William Leatham he showed that the actual number of bills of exchange increased in times of monetary contraction – but that it produced, at times, both unnecessary stringency and harmful fluctuations in the rate of interest. Though his position was analytically underdeveloped his work is still of considerable interest.

See Also

- Tooke, Thomas

Selected works

1851. An attempt to ascertain the magnitude and fluctuations of the amount of bills of exchange (Inland and Foreign) in circulation at one time in Great Britain, in England, in Scotland, in Lancashire, and in Cheshire, respectively, during each of the twenty years 1828–1847, both inclusive; and also embracing in the inquiry bills drawn upon foreign countries. *Journal of the Royal Statistical Society* 14, 143–83.

1857. (With T. Tooke) A History of Prices and of the State of the Circulation, during the Nine Years 1848–1856. In two volumes; forming the fifth and sixth volumes of the History of Prices from 1792 to the Present Time. London: Longmans.

Bibliography

Ashbee, R.A. 1979. William Newmarch and Glyns. *The Three Banks Review* 122, 49–60.

Bonar, J. and Macrosty, H.W. 1934. *Annals of the Royal Statistical Society 1834–1934*. London: Royal Statistical Society.

Gregory, T.E. 1928. *An Introduction to Tooke and Newmarch's ‘A History of Prices and of the State of the Circulation’*. London: P.S. King.

O'Brien, D.P. (ed.) 1971 . *The Correspondence of Lord Overstone*. 3 vols, Cambridge: Cambridge University Press.

Wood, E. 1939. *English Theories of Central Banking Control 1819–1858*. Cambridge, Mass.: Harvard University Press.

How to cite this article

O'Brien, D.P. "Newmarch, William (1820–1882)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000066> doi:10.1057/9780230226203.1862

The New Palgrave Dictionary of Economics Online

Nikaido, Hukukane (1923 – 2001)

Kazuo Nishimura
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

existence of general equilibria; factor substitution; Gale–Nikaido lemma; Harrod–Domar model; imperfect competition; knife-edge property; minimax theorem; monopolistic competition; Morishima, M.; Nikaido, H.; objective demand functions; Solow, R.; turnpike theorems; von Neumann growth model

Article

Hukukane Nikaido graduated from the Department of Mathematics, University of Tokyo, in 1949. While he was a university student he became interested in economics and studied Marx's *Das Kapital*, Hicks's *Value and Capital* and Samuelson's *Foundations of Economic Analysis*. After graduating from the University of Tokyo and becoming an Associate Professor of Mathematics at the Tokyo College of Science, he wrote papers concerning the von Neumann growth model and the minimax theorem (1954a; 1954b; 1955). He also worked on the existence of equilibria in the general equilibrium model with many firms and many consumers. His paper had been completed independently of Arrow and Debreu (1954) and McKenzie (1954), and was published in *Metroeconomica* (1956a). One of his results in the existence proof is now known as Gale–Nikaido lemma. These achievements led him to visit Stanford in 1955–6 at the invitation of Kenneth Arrow. At Stanford, Nikaido started to work on the existence of general equilibria for an economy with infinitely many commodities (1956b; 1957), and then published in the *Journal of the Mathematical Society of Japan* (1959a). His contributions on the existence of general equilibria in the infinite dimensional space had long remained unknown.
After he returned from Stanford he was invited by Michio Morishima to join the Institute of Socioeconomic Research at Osaka University. There he began to work on the stability of general equilibria (1959b; 1960; 1964a). Osaka was very active in research in those days, and many well-known economists from abroad visited Osaka University. One was John Hicks, who in those days was interested in the turnpike theorem of multi-sector economic growth. Turnpike theorems had been proved by Morishima (1961) and Radner (1961). Radner's result was improved by Nikaido (1964b). David Gale also visited Osaka in 1961. He and Nikaido wrote a joint paper (1965) on the uniqueness of solutions of

nonlinear simultaneous equations; the condition used in the paper has been called the Gale–Nikaido condition. In 1969 Nikaido moved from Osaka University to Hitotsubashi University in Tokyo. His previous research was published in a book, *Convex Structures and Economic Theory* (1968), which has been read by many graduate students and researchers worldwide.

After moving to Hitotsubashi University, he began to work on general equilibrium combined with monopolistic competition. Nikaido recognized, however, that the demand function, a partial equilibrium theoretic construction, involves inconsistencies in a general equilibrium situation. By introducing the concept of ‘objective demand functions’ Nikaido explored the existence of monopolistically competitive equilibria (1974). His research was published in *Monopolistic Competition and Effective Demand* (1975). Thereafter Nikaido developed his previous work on imperfect competition into a dynamic model (1978; 1979; 1980a). His main concern was in the theory of out-of-equilibrium adjustments. Nikaido also re-examined the knife-edge property in the Harrod–Domar model and the stability property in Solow's neoclassical growth model. He showed that the stability of Solow's model depends on the assumption that an investment is equal to a saving, rather than the smooth factor substitution as had been generally believed, and that, if an intended investment is not the same as a realized investment, the steady state solution is not necessarily stable and the imbalance is not solved even with flexible factor substitution (1975; 1980b).

In 1983 Nikaido joined Tsukuba University and later Tokyo International University. From then on he spent most of his time working on Marxian economics and then Keynesian models, using dynamic analysis developed in his earlier research. They were problems that he had been concerned with when he was a young university student.

See Also

- existence of general equilibrium

Selected works

1954a. Note on the general economic equilibrium for nonlinear production functions. *Econometrica* 22, 49–53.

1954b. On von Neumann's minimax theorem. *Pacific Journal of Mathematics* 4(1), 65–72.

1955. New aspects of von Neumann's model with special regard to computational problems. *Annals of the Institute of Statistical Mathematics* 6, 223–30.

1956a. On the classical multilateral exchange problem. *Metroeconomica* 8(2), 135–45.

1956b. On the existence of competitive equilibrium for infinitely many commodities. Technical Report No. 34, Department of Economics, Stanford University.

1957. Existence of equilibrium based on Walras' Law. ISER Discussion Paper No.2, Institute of Social and Economic Research, Osaka University.

1959a. Coincidence and some systems of inequalities. *Journal of the Mathematical Society of Japan* 11, 354–73.

1959b. Stability of equilibrium by the Brown-von Neumann differential equation. *Econometrica* 27, 654–71.

1960. (With H. Uzawa.) Stability and non-negativity in a Walrasian tâtonnement process. *International Economic Review* 1(1), 50–9.

1964a. Generalized gross substitutability and extremization. In *Advances in Game Theory*, ed. M. Dresher, L. Shapley and A. Tucker. Princeton: Princeton University Press.

1964b. Persistence of continual growth near the von Neumann Ray: a strong version of the Radner turnpike theorem. *Econometrica* 32, 151–62.

1965. (With D. Gale.) The Jacobian matrix and global univalence of mappings. *Mathematische Annalen* 159(2), 81–93.

1968. *Convex Structures and Economic Theory*. New York: Academic Press.

1974. What is an objective demand function? *Zeitschrift für Nationalökonomie* 34, 291–307.

1975. *Monopolistic Competition and Effective Demand*. Princeton: Princeton University Press.

1975. Factor substitution and Harrod's knife-edge. *Zeitschrift für Nationalökonomie* 35, 149–54.

1978. (With S. Kobayashi.) Dynamics of wage-price spirals and stagflation in the Leontief–Sraffa system. *International Economic Review* 19, 83–102.

1979. Wage-price spiral under monopoly: ‘What is an objective demand function?’ set in motion. *Zeitschrift für Nationalökonomie* 39, 299–313.

1980a. Prices and income distribution in a Leontief economy. *Journal of Economic Behavior and Organization* 1, 61–79.

1980b. Harrodian pathology of neoclassical growth. *Zeitschrift für Nationalökonomie* 40, 111–34.

Bibliography

Arrow, K. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

McKenzie, L. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22, 147–61.

Morishima, M. 1961. Proof of a turnpike theorem: the ‘no joint production case’. *Review of Economic Studies* 28, 89–97.

Radner, R. 1961. Paths of economic growth that are optimal states: a turnpike theorem. *Review of Economic Studies* 28, 98–104.

How to cite this article

Nishimura, Kazuo. "Nikaido, Hukukane (1923–2001)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000159> doi:10.1057/9780230226203.1187

The New Palgrave Dictionary of Economics Online

noise traders

James Dow and Gary Gorton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Noise traders are agents whose theoretical existence has been hypothesized as a way of solving certain fundamental problems in financial economics. We briefly review the literature on noise traders.

Keywords

‘no trade’ theorem; arbitrage; Grossman, S.; hedging; imperfect information revelation; information aggregation and prices; information costs; insurance motive; liquidity traders; market microstructure; market selection hypothesis; noise; noise traders; private information; rational expectations equilibrium

Article

‘Noise traders’ are economic agents who trade in security markets for non-information-based reasons. The existence of noise traders was theoretically posited as a solution to the ‘no trade’ or ‘no speculation’ results of Grossman and Stiglitz (1980) and Milgrom and Stokey (1982). These authors showed that it is impossible under most circumstances for an agent with superior information to profit from that information by trading. The intuition for the ‘no trade’ result is as follows. A buyer of an asset is prepared to pay a seller a price *p* only if the buyer believes that, conditional on the seller agreeing to sell the asset, the value of the asset exceeds *p*. But then the seller, knowing this, is at least as well off keeping the asset. So no one trades.

But we do observe trade in the world. Moreover, no trade is difficult to reconcile with the notion of asset market efficiency, in which prices allegedly contain all available information. If some agents produce costly private information and then trade on their private information, security prices will reflect some or all of the information and hence become more informationally efficient. To explain how informed traders can cover the costs of information production when they trade in securities markets, someone in the market must lose money trading against them. ‘Noise traders’ or ‘liquidity traders’ are the names given to the traders who lose money, on average, when they trade. Their trade then provides the subsidy

to cover the informed traders' cost of information production.
The idea that there are traders who systematically lose money trading securities leads to obvious questions. Do noise traders really exist? Who exactly are noise traders in reality? How do noise traders survive and persist when they are losing money trading?

Rational expectations and efficient security markets

In security markets, prices are alleged to reflect ‘all available information’. But how does this come about? What is the information, and how is it aggregated into the price? The concept of a rational expectations equilibrium (REE) gave formal content to the notion of ‘market efficiency’, which has been a central concept in financial economics since the 1960s. The idea is that, if agents understand the economy and understand how markets work, they know that current prices reflect the information which is known to some agents but maybe not to others. The uninformed agents understand the link between current prices and the information of the informed agents, and so can infer something about the information in prices. When the prices that prevail in equilibrium coincide with what the uninformed agents can learn from the prices and with the actions taken by the informed agents, who trade on their information knowing that the uninformed agents will infer (some or all) of the information, then the equilibrium is said to be a rational expectations equilibrium. The idea that prices can convey information, in the sense of REE, is due to Lucas (1972). (See also Green, 1977, and Radner, 1979. Grossman, 1981, provides a brief intellectual history of REE; see also Allen and Jordan, 1998.)
But, when all the information of the informed agents is revealed in a fully revealing REE, there is a problem if information acquisition is costly. Grossman (1976) considers a model of the stock market in which there are two types of traders: ‘informed’ and ‘uninformed’. Informed traders take positions in the market based on their information. Uninformed traders have no information but know that prices will reflect the information of the informed traders. Grossman shows that the equilibrium prices aggregate and reveal the information perfectly, ‘but in doing this the price system eliminates the private incentive for collecting the information’ (1976, p. 574). Grossman is quite clear in identifying the paradox, but he also proposes a solution:

> When a price system is a perfect aggregator of information it removes private incentives to collect information. If information is costly, there must be *noise* in the price system so that traders can earn a return on information gathering. If there is no *noise* and information collection is costly, then a perfect competitive market will break down because no equilibrium exists where one collects information. (1976, p. 574; emphasis added)

Beja (1976) also argues that REE and costly endogenous information acquisition are not compatible when agents are strategic and that consequently asset prices cannot be efficient.
So ‘noise’ is required if agents are to acquire and trade on their costly information. But what is this ‘noise’? The example of ‘noise’ that Grossman points to is ‘an uncertain total stock of the risky asset’ (1976, p. 574). He describes ‘noise’ simply as ‘many other factors’ (1977, p. 431). The device of adding a random noise term to the aggregate supply of the asset is used in Grossman and Stiglitz (1980). They show that, when information production is costly and there is noise in the asset supply, then some

traders will acquire information and trade, but rational expectations prices will not be fully revealing. If there is uncertainty about the supply of the asset in the market, or about the level of demand, or about the risk aversion of other traders, then uninformed traders cannot be sure that prices reflect the information of the informed traders. The basic idea is that the uninformed traders confuse the private information with uncertainty about the other unknown variables. It is this additional uncertainty, or noise, which makes it possible for the informed traders to trade without perfectly revealing their information, and hence profit from its production.

The device of adding a noise term to aggregate supply does result in REE that are only partially revealing. Unfortunately, there were two problems with this approach as a general matter. First, the partially revealing REE models require somewhat special assumptions. Second, it was not clear what the proposed noise shock to aggregate supply really corresponds to in reality. (There are other problems as well. Hellwig, 1980, pointed out that REE requires traders to act rationally with respect to information, yet they ignore the effect of their transactions on the price. This was deemed the ‘Schizophrenia problem’: ‘…Grossman's agents are slightly schizophrenic: (Hellwig’, 1980, p. 478’. The model in Kyle, 1985, avoided this problem.) On the first point, Green's (1977) non-existence example uses a noise term on the traders' endowments, and suggests that this will not be a suitable basis for a general approach. The general equilibrium literature did develop a number of generalizations, including, for example, the difference between the dimensions of the signals and the dimension of the prices (see, for example, Jordan, 1983, and Ausubel, 1990). Others have provided slightly different models that have partially revealing equilibria, but still there seems to be no general approach (see, for example, Allen, 1981, and see Allen and Jordan, 1998, for a discussion).

Noise traders

REE models assume that traders maximize expected utility with *rational* beliefs, where rational beliefs are defined to be consistent with the model itself. There may be ‘noise’, but this was not viewed as emanating from incorrect beliefs. (There is the issue of how traders come to understand the model, that is, how they learn. On that question see, for example, Blume, Bray and Easley, 1982, and Blume and Easley, 2004.) In general, the notion of ‘noise’ in the REE literature was somewhat vague and corresponded to a random error term added to the aggregate excess demand function. Understanding the role of ‘noise’ appeared to require leaving the REE world and explicitly detailing the origin of noise. This was done by Kyle (1985).

Kyle posited the existence of ‘uninformed noise traders who trade randomly’ (1985, p. 1315). (In private correspondence, Kyle said that he did not coin the term ‘noise trading’ but attributes it to Sanford Grossman.) Kyle identified certain people as trading in a way which made noise in the sense that their trade was not based on information. That is, he explicitly posited the existence of a class of agents – people – who traded in a certain way so as to fulfil the role of ‘noise’. By explicitly introducing noise traders, Kyle focused attention on the details of the trading process. This became the foundation for the study of market microstructure. (Garman, 1976, appears to have been the first to use the term ‘market microstructure’. See Easley and O'Hara, 2003, for a survey of the microstructure literature.) Around the same time Kyle, Glosten and Milgrom introduced a similar class of agents: ‘…we assume that there are

informed investors and purely “liquidity” traders’ (1985, p. 76). Earlier, Treynor (1971, under the pseudonym W. Bagehot) talked about ‘liquidity-motivated’ traders.

In REE models agents do not act strategically; the process of learning from prices occurs in equilibrium (as opposed to happening in real time), and the details of trading are treated in reduced form (agents submit demand functions to an auctioneer). Kyle and Glosten and Milgrom changed this by specifying the trading process in a way that was not possible in REE models. In both papers there is a competitive market-maker who receives orders from traders, at least one of whom has superior information. The market-maker must infer the information of the informed trader from the order flow. The market-maker knows that some traders are privately informed, and that others are not trading based on any superior information (the noise traders). Inference about information occurs as the market-maker learns by watching the order flow. Gradually, the market-maker changes his price to reflect the information.

Still, the noise traders in this new type of model were not well-motivated. In fact, their motives are not explained. They earn a lower-than-average return than the informed agents, who earn an above-average return. If the uninformed noise traders could at least buy the market portfolio, then they could earn the average return on the market. But in fact they are not allowed to buy the market portfolio. That is their root problem (see Dow and Gorton, 1995).

Diamond and Verrecchia (1981) suggest adding a noise term to agents' risk exposures (their endowments). Risk-averse agents will then have an insurance motive for trading. DeMarzo and Duffie (1999) propose a model where different traders have different discount rates. Shocks to their discount rates provide an incentive to trade that other traders cannot distinguish from speculative trading intended to profit from information about the liquidation value of the asset. These papers solve the theoretical problem of finding a logically consistent model that can be used as a basis for economic analysis, including welfare statements, of markets with imperfect information revelation. Papers that have applied these models in various settings include Biais and Mariotti (2005) (for the DeMarzo and Duffie model) and Dow and Rahi (2000; 2003) (for the Diamond and Verrecchia approach).

But is it really plausible to believe that there is a significant demand for individual stocks or bonds based on an insurance motive? Stock indexes, exposure to the yield curve, or foreign currency could experience demand variations due to insurance motives, but there are close substitutes for individual stocks and bonds from a risk point of view. Also, if investors do start off with different discount factors, one would expect them to trade these differences away.

In other words, plausibly the demand curve for an individual asset should be almost perfectly elastic. The price at which it becomes elastic (given the prices of all other assets) should be almost identical for all agents. Hence, we revert to the situation where the asset has a unique fundamental value that all agents will agree on if they have the same information about the asset's cash flows. So the question of who noise traders actually are remains open.

Who are the noise traders?

The details of the identity of noise traders or liquidity traders were initially left vague. For example, Glosten and Milgrom write of exogenous events motivating their trade, like ‘job promotions or unemployment, deaths or disabilities…’ (1985, p. 77). These shocks were not well identified. Notably, noise traders were modelled as equally likely to be buying or selling securities, which, while making

models technically tractable, is counter-intuitive. Exogenous reasons for needing money and hence having to sell securities seems more natural than exogenous reasons for having to buy securities.

The details of the identity of noise traders are important, because if noise traders are simply irrational there is clearly an incentive for ‘smart money’ to take advantage of them, and eventually eliminate them from the market. The ‘market selection hypothesis’ holds that irrational traders will eventually be driven out of the market. Noise traders should not survive, and so cannot play the role envisioned for them. In fact, it has long been argued that rational traders will eliminate irrational traders from the market by taking their money when they trade at incorrect prices. This process is what causes prices to be driven to (or close to) fundamental values (see, for example, Friedman, 1953).

Noise traders can survive only if there are some frictions or barriers preventing them from being eliminated by the smart money. That is, there must be some limits to arbitrage. One possibility is that the smart money has a limited horizon over which trade can occur. With a limited horizon, the noise traders could cause losses to the smart money by moving prices further away from fundamentals. This is the idea in DeLong et al. (1990), Dow and Gorton (1994), and Shleifer and Vishny (1997). These papers argue that there are ‘limits to arbitrage’, providing an explanation for the persistence of noise trade.

Still, the question remains: who are the noise traders? On one view, noise traders are simply individuals who are less than rational; they are subject to behavioural biases and fads. For example, Shiller (1984) argued that some investors rely on ‘popular models’ which are wrong, and also that they can be subject to fads. Along the same lines, Shleifer and Summers (1990, p. 19) wrote: ‘their demand for assets is affected by their beliefs or sentiments that are not fully justified by fundamental news.’ A large literature argues that individual investor trading is subject to a myriad of psychological biases, and that such individuals may use various heuristics, ‘popular models’, as the basis for their investment decisions. This literature is surveyed in Barberis and Thaler (2003).

A second rationale for noise trading focuses not on individual investors but on professional traders and money managers (‘funds’) hired by principals/investors. Funds do not invest and trade their own money; they work for others. This creates a potential conflict of interest or agency problem. This notion is developed by Dow and Gorton (1997). They argue that churning by funds, which occurs when they do not become informed and want to pretend that they have, is ‘noise’ in a setting where all market participants are rational. Among the other agents in the market are hedgers. Noise trading, being a manifestation of agency problems, reduces the profitability of traders to the employers of the traders and money managers. But it benefits hedgers who earn more when they hedge. Consequently, they hedge more, which in turn can support more informed fund trading. Dow and Gorton (1997) show that a ‘small’ amount of hedging demand can result in a ‘large’ noise. Irrationality is not needed to explain significant amounts of noise.

Summary

Noise traders play an essential role in modern finance theory, but their identities, motivations, and ability to persist remain topics of research.

We thank Pete Kyle for comments.

Bibliography

Allen, B. 1981. A class of monotone economies in which rational expectations equilibria exist but prices do not reveal all information. *Economics Letters* 7, 227–32.

Allen, B. and Jordan, J. 1998. The existence of rational expectations equilibrium: a retrospective. Research Department Staff Report No. 252, Federal Reserve Bank of Minneapolis.

Ausubel, L. 1990. Partially-revealing rational expectations equilibrium in a competitive economy. *Journal of Economic Theory* 50, 93–126.

Bagehot, W. (pseudonym for Jack Treynor). 1971. The only game in town. *Financial Analysts Journal* 22, 12–14.

Barberis, N. and Thaler, R. 2003. A survey of behavioral finance. In *Handbook of the Economics of Finance: Financial Markets and Asset Pricing*, vol. 1B, ed. G. Constantinides, M. Harris and R. Stulz. Amsterdam: North-Holland.

Beja, A. 1976. The limited information efficiency of market processes. Working Paper No. 43, Research Program in Finance, University of California, Berkeley.

Biais, B. and Mariotti, T. 2005. Strategic liquidity supply and security design. *Review of Economic Studies* 72, 615–49.

Blume, L., Bray, M. and Easley, D. 1982. Introduction to the stability of rational expectations equilibrium. *Journal of Economic Theory* 26, 313–17.

Blume, L. and Easley, D. 2004. If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets. *Econometrica* 74, 929–66.

DeLong, J., Shleifer, A., Summers, L. and Waldman, R. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98, 703–38.

DeMarzo, P. and Duffie, D. 1999. A liquidity-based model of security design. *Econometrica* 67, 65–99.

Diamond, D. and Verrecchia, R. 1981. Information aggregation in a noisy rational expectations economy. *Journal of Financial Economics* 9, 221–35.

Dow, J. and Gorton, G. 1994. Arbitrage chains. *Journal of Finance* 49, 819–50.

Dow, J. and Gorton, G. 1995. Profitable informed trading in a simple general equilibrium model of asset pricing. *Journal of Economic Theory* 67, 327–69.

Dow, J. and Gorton, G. 1997. Noise trading, delegated portfolio management, and economic welfare. *Journal of Political Economy* 105, 1024–50.

Dow, J. and Rahi, R. 2000. Should speculators be taxed? *Journal of Business* 73, 89–107.

Dow, J. and Rahi, R. 2003. Informed trading, investment, and economic welfare. *Journal of Business* 76, 430–54.

Easley, D. and O'Hara, M. 2003. Microstructure and asset pricing. In *Handbook of the Economics of Finance: Financial Markets and Asset Pricing*, ed. G. Constantinides, M. Harris and R. Stulz. Amsterdam: North-Holland.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Garman, M. 1976. Market microstructure. *Journal of Financial Economics* 3, 257–75.

Glosten, L. and Milgrom, P. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14, 71–100.

Green, J. 1977. The non-existence of informational equilibria. *Review of Economic Studies* 44, 451–63.

Grossman, S. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 32, 573–85.

Grossman, S. 1977. The existence of futures markets, noisy rational expectations and informational externalities. *Review of Economic Studies* 44, 431–49.

Grossman, S. 1978. Further results on the informational efficiency of competitive stock markets. *Journal of Economic Theory* 18, 81–101.

Grossman, S. 1981. An Introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48, 541–59.

Grossman, S. and Stiglitz, J. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70, 393–408.

Hellwig, M. 1980. On the aggregation of information in competitive markets. *Journal of Economic Theory* 22, 477–98.

Jordan, J. 1983. On the efficient market hypothesis. *Econometrica* 51, 1325–43.

Kyle, A. 1985. Continuous auctions and insider trading. *Econometrica* 53, 1315–35.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Milgrom, P. and Stokey, N. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26, 17–27.

Radner, R. 1979. Rational expectations equilibrium: existence and information revealed by prices. *Econometrica* 47, 370–91.

Shiller, R. 1984. Stock prices and social dynamics. *Brookings Papers on Economic Activity* 1984(2), 457–98.

Shleifer, A. and Summers, L. 1990. The noise trader approach to finance. *Journal of Economic Perspectives* 4, 19–33.

Shleifer, A. and Vishny, R. 1997. The limits of arbitrage. *Journal of Finance* 51, 35–55.

How to cite this article

Dow, James and Gary Gorton. "noise traders." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000146> doi:10.1057/9780230226203.1188

The New Palgrave Dictionary of Economics Online

nominal exchange rates

Richard T. Baillie

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The nominal exchange rate is the rate at which the currency of one country can be exchanged for that of another. The overall value of a currency can be summarized through the ‘effective nominal exchange rate’, which is a weighted average of a country's nominal bilateral exchange rates. Following the advent of freely floating exchange rates in 1973 there has been intense research on understanding the mechanisms of nominal exchange rate determination and the search for an adequate model, but no model has so far withstood rigorous empirical tests.

Keywords

Bretton Woods system; bubbles; cointegration; currency crises; euro; European Monetary System; exchange rate determination; exchange rate target zones; exchange rate volatility; fixed exchange rates; floating exchange rates; forward premium anomaly; GARCH models; law of one price; monetary transmission mechanism; money supply; nominal exchange rate determination; nominal exchange rates; nominal interest rates; open market operations; overshooting; peso problem; purchasing power parity; rational expectations; real balances; real exchange rates; real interest differential; realized volatility; risk premium; specification problems in econometrics; spot exchange rates; sterilized intervention; transversality condition; uncovered interest parity

Article

The nominal exchange rate is the price at which the money of one country can be exchanged for another. Usually, nominal exchange rates are bilateral, which means they denote the number of units of one country in terms of one unit of another; for example, two US dollars to one UK pound; or 0.50 UK pounds to one US dollar. Bilateral exchange rates can be expressed either in terms of spot rates, which are prices for immediate delivery, or in terms of forward contracts for delivery in the future. Some foreign exchange (FX) markets also trade currency options and futures. The worldwide FX market

transacted approximately $1,700bn a day in 2006, making it by far the largest financial market. On most weeks it operates for 156 of the 168 hours available, with New York, London and Tokyo being considered as the most important and heavily traded markets. Surveys of FX market participants generally suggests that 98 per cent or more of currency transactions are motivated by speculation, arbitrage and international capital movements, rather than for the purposes of importing or exporting goods.

The overall value of a particular currency can be summarized through the ‘effective nominal exchange rate’, which is a weighted average of a country's nominal bilateral exchange rates. A number of international financial institutions regularly report these effective rates, with different weights being used dependent on which criteria – for example, the patterns of trade – are being emphasized.

A real exchange rate is the nominal bilateral exchange rate divided by the ratio of the price indices for the two countries. Usually consumer price indices (CPIs) are used for this purpose, although trade weighted price indices are also sometimes used.

Historical perspective

Following the end of the Second World War in 1945, a conference in Bretton Woods, New Hampshire established a system of fixed exchange rates based on the US dollar, with the US dollar in turn being convertible to gold at a fixed gold standard. However, continuing trade imbalances and apparent exchange rate misalignments led to a collapse of the Bretton Woods fixed exchange rate system in March 1973. Since then the international monetary system has generally followed what is best characterized as a managed or ‘dirty floating’ regime, with governments and/or central banks occasionally intervening to attempt to influence the value of the currencies and volatility of the market. Until the early 1990s the nominal rates between the three major regions of North America, Western Europe and Japan were formally freely floating. However, many bilateral rates were pegged under various arrangements. In particular, the European Monetary System (EMS) allowed individual countries currencies to move in a narrow band, named the ‘snake’ around par rates for each member country's bilateral rate vis-à-vis the German Deutschmark. After several periods of apparent instability, such as the autumn of 1992 when the UK pound exited the EMS, and also the autumn of 1993 when the bands were widened to plus and minus 15 per cent of par rate; the new euro currency was introduced in 1999. Originally ten member countries of the EMS surrendered their sovereign currencies to form the euro area.

The other major development, converse to the formation of the euro, has been the collapse of communism in the late 1980s and the early 1990s, which has led many of the previously fixed exchange rates of eastern Europe and Asia to become floating rates.

As of 2007 the currencies of the US dollar, Japanese yen, euro, British pound, Swiss franc and Canadian dollar are the most actively traded, freely floating currencies.

Empirical behaviour

To a large extent nominal exchange values and returns behave in similar manner to other asset prices. On denoting the spot exchange rate at time t as S_t, then $\Delta s_t = \Delta \ln(S_t)$ is the approximately continuously

compounded rate of return. Many empirical studies have found that the hypothesis of a unit root in $\ln(S_t)$ cannot be rejected, so that returns appear to be stationary. Furthermore, returns generally appear to be approximately serially uncorrelated, so that the returns appear to be close to a martingale difference sequence, which is consistent with the theory of weak form efficiency. This has led to the one of the most striking empirical properties of high frequency, daily, weekly, or even monthly nominal exchange rate returns, concerning their apparent lack of predictability in their conditional mean. Numerous studies such as Meese and Rogoff (1983), using forward rates, surveys of market participant's expectations, and nonlinear time series models have been unable in the MSE sense to improve on random walk predictions of the nominal exchange rate.

However, the unconditional distribution of short-term nominal spot exchange rate returns is non-Gaussian and has substantial excess kurtosis; that is, they are leptokurtic. Also, returns generally exhibit time-dependent volatility, which can be well represented by various types of generalized autoregressive conditionally heteroskedastic (GARCH) models. These models represent the autocorrelated nature of volatility, which is generally considered to be due to arrival of news and to the patterns of trading volume. See Baillie and Bollerslev (1989), who estimate and discuss these models for different levels of temporal aggregation. The degree of non-Gaussianity and the level of persistence of the volatility in GARCH models are particularly high for daily returns and decreases for lower frequencies of returns. Andersen and Bollerslev (1997a; 1997b) have used high-frequency data to examine returns and the volatility process of nominal spot exchange rate returns. They find particular stylized patterns of worldwide FX market volatility which characterizes the volatility process for each spot returns series. Andersen et al. (2003) consider the concept of realized volatility, which is an observable measure of (daily) volatility obtained from aggregating information on high-frequency returns within the day. For example, the sum of squared high-frequency returns is often used to measure daily realized volatility. The daily realized volatility is generally found to be almost pure fractional white noise, with the long-memory parameter generally being in the range of 0.30–0.40.

Purchasing power parity

The theory of purchasing power parity (PPP) is sometimes known as the law of one price and is to be found in the work of Ricardo in the 18th century and by Cassel in the 1920s. If S_t denotes the spot exchange rate, measured in terms of the dollar–yen rate at time t, P_t is the domestic US price level and P_t^* is the foreign country (Japan's) price level, then continuous PPP requires $P_t = S_t P_t^*$. The real exchange rate is defined as Q_t where $Q_t = (S_t P_t^*) / P_t$ and, if PPP held continuously, the real exchange rate would be constant over time. In general, empirical real exchange rates since 1973 have been found to exhibit highly persistent autocorrelation and may possibly be non-stationary (see Abauf and Jorion, 1990). An important area of research in international finance has been to understand the duration of the effect of shocks to the real exchange rate, and the evidence for whether the real exchange rate returns to equilibrium in ‘finite’ time and restores PPP. More empirical work has re-established PPP holding in the long run, but with significant deviations (see Frankel and Rose, 1996).

Uncovered interest rate parity

On denoting domestic interest rates as i_t and foreign rates as i_t^* it is known that covered interest rate parity holds exactly apart from very small transaction costs and brokerage fees so that $(1 + i_t)S_t = (1 + i_t^*)F_t$, where F_t is the forward exchange rate. This relationship implies that the forward premium is equivalent to the interest rate differential. An important extension is the theory of uncovered interest rate parity (UIP), where $(1 + i_t) = (1 + i_t^*)E_t(S_{t+1} / S_t)$, which implies that the interest rate differential is approximately the expected rate of appreciation (depreciation) of a currency. Hence,

$$E_t \Delta S_{t+1} \approx (i_t - i_t^*),$$

where E_t represents the expectation operator conditioned on a sigma field of information available at time t. Hence the country with the higher rate of interest is expected to have the currency depreciation. The UIP hypothesis requires the joint assumptions of rational expectations, risk neutrality, free capital mobility and the absence of taxes on capital transfers. The theory can be derived from the solution of an Euler equation where expected real returns in the forward market are hypothesized to be zero.

Models of exchange rate determination

Following the advent of freely floating exchange rates in 1973 there has been intense research on understanding the mechanisms of nominal exchange rate determination and the search for an adequate model.

Earlier work by Mundell (1963) and Fleming (1962) emphasized a Keynesian approach and considered the relative advantages of fixed versus floating nominal exchange rates. In particular, monetary policy was shown to be ineffective as a policy tool under a fixed exchange rate, while fiscal policy is effective. Conversely, monetary policy was shown to be effective under a flexible exchange rate, and fiscal policy to be ineffective under flexible exchange rates. The dominant modern paradigm is the asset market approach, which implies that the nominal exchange rate is the value of one country's money supply against another. The simplest version of the monetary model assumes PPP to hold continuously and for the existence of stable and static demand for real balances for one and possibly both countries. If the demand for real balances in the United States is $m_t - p_t = \varphi y_t - \alpha i_t$ where the lower case letters m_t, p_t and y_t represent the natural logarithms of money, prices and income respectively, and where i_t is the level of nominal interest rates; where ϕ is the elasticity of the demand for real balances with respect to income and α is the semi-elasticity with respect to the nominal rate of interest. The combination of PPP, uncovered interest rate parity and the demand for real balances equation is sufficient to generate a first-order rational expectations equation of the form

$$s_t = \left(\frac{1}{1+\alpha}\right) z_t + \left(\frac{\alpha}{1+\alpha}\right) E_t s_{t+1}$$

where z_t are the fundamentals, and in this case are $z_t = (m_t - \varphi y_t - p_t^* + \alpha i_t^*)$, and asterisks denote foreign equivalents. On assuming the transversality condition which eliminates bubbles, the forward looking solution is

$$s_t = \left(\frac{1}{1+\alpha}\right) \sum_{j=0}^{\infty} \left(\frac{\alpha}{1+\alpha}\right)^j E_t (m_{t+j} - \varphi y_{t+j} - p_{t+j}^* + \alpha i_{t+j}^*).$$

A more intuitive solution is to further assume the same demand for real balances equation for the foreign country, in which case the solution is

$$s_t = \left(\frac{1}{1+\alpha}\right) \sum_{j=0}^{\infty} \left(\frac{\alpha}{1+\alpha}\right)^j E_t [(m_{t+j} - m_{t+j}^*) - \varphi (y_{t+j} - y_{t+j}^*)].$$

Similarly to the Keynesian approach, the monetary model implies an equivalent depreciation of the exchange rate with respect to an increase in US money supply and prices. However, the model also implies dollar appreciation following an increase in US incomes, and a dollar depreciation following an increase in US nominal interest rates; both implications are contrary to the Keynesian approach.

The empirical realization that nominal exchange rates did not move in perfect synchronization with relative prices and money supplies generated attempts to loosen the constraints of the model. Frankel (1979) introduced the real interest differential (RID) model, while the celebrated concept of overshooting was due to Dornbusch (1976). Under rational expectations, the overshooting model is very similar to an alternative model of Woo (1985), which assumes flexible prices but dynamic adjustments in the demand for real balances. The solution paths for both models are obtained from the forward solution of a second-order forward-looking rational expectations equation.

It has been hard to find rigorous empirical support for any of the models. While the log of the exchange rate and most of the macroeconomic fundamentals appear to be well approximated by integrated processes, there has been an absence of cointegration. This rejects the long-run properties of the basic monetary model as well as the Dornbusch overshooting formulation. In fact, the macro fundamentals are found to add little explanatory power to the model. This again is consistent with the findings of Meese

and Rogoff (1983), who found that most models and forecasting methods were inferior, in the sense of *ex ante* MSE forecasting comparisons, to a simple random walk model. Mark (1995) and Mark and Choi (1997) have found some evidence that fundamentals have increased explanatory power when predicting exchange rates a year or more ahead. An alternative approach considering the possibility of nonlinear adjustment to equilibrium has been advocated by Engel and Hamilton (1990) and Taylor and Peel (2000), and is likely to remain an active area of research for the forseeable future.

High-frequency analyses by Anderson and Bollerslev (1998) and Andersen et al. (2003) have examined the role of macro news announcements on exchange rate returns. Some explanatory power has been detected, but not as much as would be suggested by the macro models. These findings tend to support perceived wisdom in the FX market concerning the fact that traders react less to macroeconomic news than previously expected.

Forward premium anomaly

The forward premium or forward discount anomaly refers to the widespread result that the returns on freely floating exchange rates are invariably negatively correlated with the lagged forward premium. One of the most widespread tests of uncovered interest rate parity is based on the regression of future spot returns on the lagged forward premium, or equivalently the lagged interest rate differential,

$$\Delta s_{t+1} = \alpha + \beta(f_t - s_t) + \varepsilon_{t+1},$$

where ε_{t+1} is the regression disturbance. While the theory of uncovered interest rate parity would suggest that $\alpha = 0$, $\beta = 1$ and ε_{t+1} uncorrelated, a substantial body of empirical work has found the estimate of the slope coefficient β to be negative. Interestingly, this result is found for different currencies, different numeraire currencies and over different sample periods, including the 1920s. As discussed by Baillie and Bollerslev (2000), the estimated β coefficient is time varying and can be as low as –13 for periods within the 1980s. Possible explanations of the forward premium anomaly have included ‘peso problem’ effects, the role of learning and heterogeneous beliefs on the part of agents; while the most dominant explanation has been in terms of the presence of a time-dependent risk premium, ρ_{t+1} which is defined as

$$E_t\Delta s_{t+1} = (f_t - s_t) - \rho_{t+1}.$$

Fama (1984) has shown that a $\hat{\beta} < 0$ implies that $Cov(E_t\Delta s_{t+1}\rho_{t+1}) < 0$, so that the expected rate of appreciation is negatively correlated with the risk premium, and also $Var(\rho_{t+1}) > Var(E_t\Delta s_{t+1})$, so

that the variability of the risk premium must exceed that of the expected rate of appreciation. Models of the time-dependent risk premium are generally motivated by versions of the Lucas–Breeden asset pricing approach (see Lucas, 1978). Following Domowitz and Hakkio (1985) many parametric models for the risk premium have been formulated from micro theoretic models. See Hodrick (1989) for one of the most detailed formulations, which is discussed in detail by Engel (1996). These models generally represent the risk premium in terms of the second conditional moments of fundamentals, and there has been little definitive empirical support for these models. However, Baillie and Kilic (2006) have found evidence for nonlinear smooth transition regime adjustment to uncovered interest rate parity with threshold variables, such as the conditional variability of US money growth, and the interest rate differential, which are variables derived for risk premium from theoretical models.
Other authors have noted that the problem of econometric specification with uncorrelated returns being regressed on the forward premium or interest rate differential appears to have very persistent, or ‘long-memory’ autocorrelation. Baillie and Bollerslev (2000) and Maynard and Phillips (2001) discuss some of the specification issues that result.

Target regimes and intervention

There has been considerable research on the implementation of target zones for nominal exchange rates. In particular, Krugman (1991) has considered the differential equations behind monetary policy-style intervention at the bands of the target zone; while Neely (1999) documents some of the statistical properties of such returns. Complications due to intra-marginal intervention have also been considered, and the empirical success of the models is discussed by Bekaert and Gray (1998). Perhaps most work in this area has been done on trying to understand the transmission mechanism of sterilized intervention, where open market operations by a central bank are designed to maintain levels of money supply following their purchase (sale) of domestic currency. Such intervention is generally officially motivated as an attempt to either move a nominal exchange rate closer to a target level, and/or to reduce FX market volatility. The empirical results are controversial with relatively small effects being detected, although Baillie and Osterberg (1997) use an extension of Hodrick (1989) to motivate intervention affecting the risk premium, and find quite strong supportive econometric evidence. The reasons for currency crises and the possibility of early warning corrective actions that may be taken to avoid crises have also attracted attention (see Kaminsky, Lizondo and Reinhart, 1998; Kaminsky and Schumaker, 2000; Rose and Svensson, 1995).

See Also

- exchange rate target zones
- purchasing power parity
- real exchange rates
- uncovered interest parity

Bibliography

Abauf, N. and Jorion, P. 1990. Purchasing power parity in the long run. *Journal of Finance* 45, 157–74.

Andersen, T.G. and Bollerslev, T. 1997a. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* 4, 115–58.

Andersen, T.G. and Bollerslev, T. 1997b. Heterogeneous information arrivals and return volatility dynamics: uncovering the long-run in high frequency returns. *Journal of Finance* 52, 975–1005.

Andersen, T.G. and Bollerslev, T. 1998. Deutsche mark–dollar volatility: intraday activity, patterns, macroeconomic announcements and longer run dependence. *Journal of Finance* 53, 219–65.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, F. 2003. Modeling and forecasting realized volatility. *Econometrica* 71, 579–625.

Baillie, R.T. and Bollerslev, T. 1989. The message in daily exchange rates: a conditional variance tale. *Journal of Business and Economic Statistics* 7, 297–305.

Baillie, R.T. and Bollerslev, T. 2000. The forward premium anomaly is not as bad as you think. *Journal of International Money and Finance* 19, 471–88.

Baillie, R.T. and Osterberg, W.P. 1997. Central bank intervention and risk in the forward premium. *Journal of International Economics* 43, 483–97.

Baillie, R.T. and Kilic, R. 2006. Do asymmetric and nonlinear adjustments explain the forward premium anomaly? *Journal of International Money and Finance* 25, 22–47.

Bekaert, G. and Gray, S.F. 1998. Target zones and exchange rates: an empirical investigation. *Journal of International Economics* 45, 1–35.

Domowitz, I. and Hakkio, C.S. 1985. Conditional variance and the risk premium in the foreign exchange market. *Journal of International Economics* 19, 47–66.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84, 1161–76.

Engel, C. 1996. The forward discount anomaly and the risk premium: a survey of recent evidence. *Journal of Empirical Finance* 3, 123–92.

Engel, C.M. and Hamilton, J.D. 1990. Long swings in the dollar: are they in the data and do the markets

know it? *American Economic Review* 80, 689–713.

Fama, E.F. 1984. Spot and forward exchange rates. *Journal of Monetary Economics* 14, 319–38.

Fleming, M.J. 1962. Domestic financial policies under fixed and floating exchange rates. *IMF Staff Papers* 9, 369–79.

Frankel, J.A. 1979. On the mark: theory of floating exchange rates based on real interest rate differentials. *American Economic Review* 69, 610–22.

Frankel, J.A. and Rose, A.K. 1996. A panel data project on purchasing power parity: mean reversion within and between bands. *Journal of International Economics* 40, 209–24.

Hodrick, R.J. 1989. Risk, uncertainty and exchange rates. *Journal of Monetary Economics* 23, 433–59.

Kaminsky, G., Lizondo, S. and Reinhart, C.M. 1998. Leading indicators of currency crisis. *IMF Staff Papers* 45, 1–48.

Kaminsky, G.L. and Schumaker, S.L. 2000. What triggers market jitters? A chronicle of the Asian crisis. *Journal of International Money and Finance* 18, 537–60.

Krugman, P. 1991. Target zones and exchange rate dynamics. *Quarterly Journal of Economics* 106, 669–82.

Lucas, R.E. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

Mark, N.C. 1995. Exchange rates and fundamentals: evidence on long-horizon predictability. *American Economic Review* 85, 201–18.

Mark, N.C. and Choi, D.-Y. 1997. Real exchange rate prediction over long horizons. *Journal of International Economics* 43, 29–60.

Maynard, A. and Phillips, P.C.B. 2001. Rethinking an old empirical puzzle: econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics* 16, 671–708.

Meese, R.A. and Rogoff, K.R. 1983. Exchange rate models of the seventies: do they fit out of sample? *Journal of International Economics* 14, 3–24.

Mundell, R.A. 1963. Capital mobility and stabilization policy under fixed and flexible exchange rates. *Canadian Journal of Economics and Political Science* 29, 475–85.

Neely, C.J. 1999. Target zones and conditional volatility: the role of realignments. *Journal of Empirical Finance* 6, 177–92.

Rose, A.K. and Svensson, L.E.O. 1995. Expected and predicted realignments: the FF/DM exchange rate during the European Monetary System 1979–1993. *Scandinavian Journal of Economics* 97, 175–200.

Taylor, M.P. and Peel, D.A. 2000. Nonlinear adjustments and long run equilibrium and exchange rate fundamentals. *Journal of International Money and Finance* 19, 33–53.

Woo, W.T. 1985. The monetary approach to exchange rate determination under rational expectations. *Journal of International Economics* 18, 1–16.

How to cite this article

Baillie, Richard T. "nominal exchange rates." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000071> doi:10.1057/9780230226203.1189

The New Palgrave Dictionary of Economics Online

non-clearing markets in general equilibrium

Jean-Pascal Bénassy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In this article we study models with non-clearing markets in a full general equilibrium framework. The theories we describe synthesize three major schools of thought, Walrasian, Keynesian and imperfect competition. This synthesis is notably achieved by introducing quantity signals in addition to price signals into the traditional general equilibrium model. This considerably enlarges the scope of traditional general equilibrium, allowing us not only to construct equilibria with various price rigidities but also to endogenize prices in a decentralized imperfect competition framework.

Keywords

active demand; budget constraint; dual decision method; fixprice equilibria; fixprice-fixwage macroeconomic model; frictions; general equilibrium; imperfect competition; Keynesianism; market efficiency; monopolistic competition; non-clearing markets in general equilibrium; non-tâtonnement processes; objective demand curve; overbidding; price making; price signals; price taking; quantity constraints; quantity signals; quasi-equilibrium; rational expectations; rationing; subjective demand curves; Walrasian equilibrium

Article

In this article we study how to model situations of non-clearing markets in a full general equilibrium framework. As we shall see from the historical discussion at the end, the theories we obtain synthesize three major schools of thought: (*a*) the Walrasian school, as Walras was the first to study a fully fledged general equilibrium system; (*b*) the Keynesian school, as Keynes emphasized the importance of quantity adjustments in reaching a macroeconomic equilibrium with at least one non-clearing market (that is, the labour market); and (*c*) the imperfect competition school, which endogenized prices through explicit price making by agents internal to the system.
This synthesis is notably achieved by introducing quantity signals into the traditional general equilibrium model. These quantity signals are quantity constraints which tell each agent the maximum quantity he can trade in each market. As we shall see, the introduction of these quantity signals in addition to price signals considerably enlarges the scope of traditional general equilibrium since they allow us not only to treat equilibria with various price rigidities, but also to endogenize prices in a decentralized imperfect competition framework.
The plan of the entry is the following. In the next three sections we describe the general concepts. The fourth section gives a brief historical outline of this line of thought.

Non-clearing markets and quantity signals

In this section and the next two we describe various concepts in the framework of a monetary exchange economy where one good, money, serves as numéraire, medium of exchange and reserve of value (similar concepts have been developed for barter economies – see Bénassy, 1975b; 1982 – but the formalization gets quite clumsy). There are • markets in the period considered, where non-monetary goods indexed by $h = 1, \ldots, \ell$ are exchanged against money at the price p_h. We call p the vector of these prices.
Agents are indexed by $i = 1, \ldots, n$. In market h agent i may make a purchase $d_{ih} \geq 0$ or a sale $s_{ih} \geq 0$. Define his net transaction of good h, $z_{ih} = d_{ih} - s_{ih}$, and z_i the •-dimensional vector of these net transactions.

At the beginning of the period agent *i* holds quantities $\overline{m}_i$ of money, and ω_{ih} of good *h*. Call ω_i the vector of the ω_{ih}. As a result of his trades z_i, agent *i* ends up with final holdings of non-monetary goods and money, x_i and m_i, given respectively by:

$$x_i = \omega_i + z_i m_i = \overline{m}_i - pz_i$$

We assume that agent *i* has a utility function on these final holdings $U_i(x_i, m_i) = U_i(w_i + z_i, m_i)$, which we assume throughout strictly concave in its arguments.

Walrasian equilibrium

In order to contrast it with the non-Walrasian equilibrium concepts that will follow, let us describe briefly the Walrasian equilibrium of this economy (Arrow and Debreu, 1954; Debreu, 1959). Each agent *i* receives (from the implicit auctioneer) a price signal *p*. As a response he expresses a Walrasian net demand given by the function $z_i(p)$, solution in z_i of the following program:

$$\text{Maximize } U_i(\omega_i + z_i, m_i) \text{ s.t.}$$
$$pz_i + m_i = \overline{m}_i$$

A Walrasian equilibrium price vector p^* is defined by the condition that all markets clear, that is:

$$\sum_{i=1}^{n} z_i(p^*) = 0$$

The vector of transactions realized by each agent *i* is $z_i(p^*)$.

Demands and transactions

As we will be studying non-clearing markets, we must now make an important distinction, that between demands and supplies on the one hand, and the resulting transactions on the other.

Transactions, that is, purchases or sales of goods, denoted d^*_{ih} and s^*_{ih}, are exchanges actually made, and must thus identically balance on each market, that is:

$$D^*_h = \sum_{i=1}^{n} d^*_{ih} = \sum_{i=1}^{n} s^*_{ih} = S^*_h \text{ for all } h \quad (1)$$

On the other hand, demands and supplies, denoted $\tilde{d}_{ih}$ and $\tilde{s}_{ih}$, are signals transmitted to the market (that is, to the other agents) before exchange takes place. They represent as a first approximation the exchanges the agents wish to make on each market. So they do not necessarily match in a specific market, and no identity like (1) applies to them:

$$\tilde{D}_h = \sum_{i=1}^{n} \tilde{d}_{ih} \neq \sum_{i=1}^{n} \tilde{s}_{ih} = \tilde{S}_h$$

In order to shorten notation, we often work in what follows with net demands and net transactions defined respectively by:

$$\tilde{z}_{ih} = \tilde{d}_{ih} - \tilde{s}_{ih} z_{ih}^{*} = d_{ih}^{*} - s_{ih}^{*}$$

The equality of aggregate purchases and sales (eq. (1)) is then rewritten:

$$\sum_{i=1}^{n} z_{ih}^{*} = 0 \text{for all } h \qquad (2)$$

Rationing schemes

In each market *h* the exchange process must generate consistent transactions (that is, transactions satisfying eqs. (1) or (2)) from any set of possibly inconsistent demands and supplies. Some rationing will necessarily occur, which may take various forms, such as uniform rationing, queuing, priority systems, proportional rationing, and so forth … depending on the particular organization of each market. We call *rationing scheme* the mathematical representation of each specific organization. To be more precise, the rationing scheme in market *h* is defined by a set of *n* functions:

$$z_{ih}^{*} = F_{ih}(\tilde{z}_{1h}, \ldots, \tilde{z}_{nh}) \; i = 1, \ldots, n \qquad (3)$$

such that:

$$\sum_{i=1}^{n} F_{ih}(\tilde{z}_{1h}, \ldots, \tilde{z}_{nh}) = 0 \text{for all } \tilde{z}_{1h}, \ldots, \tilde{z}_{nh}$$

We assume that F_{ih} is continuous, non-decreasing in $\tilde{z}_{ih}$ and non-increasing in the other arguments. Before examining the possible properties of these rationing schemes, let us take a most simple example with two agents. Agent 1 emits a demand $\tilde{d}_{1h}$, agent 2 a supply $\tilde{s}_{2h}$. Then a natural rationing scheme, implicit in most macroeconomic models, is to take the level of transactions as equal to the minimum of demand and supply, that is:

$$d^*_{1h} = s^*_{2h} = \min(\tilde{d}_{1h}, \tilde{s}_{2h}) \tag{4}$$

Properties of rationing schemes

We first study two possible properties that a rationing scheme may satisfy: voluntary exchange and market efficiency.
The first property is actually an extremely natural one in a free market economy: We shall say that there is *voluntary exchange* in market *h* if no agent can be forced to purchase more than he demands, or to sell more than he supplies, which is expressed by:

$$d^*_{ih} \le \tilde{d}_{ih} s^*_{ih} \le \tilde{s}_{ih} \text{for all } i$$

or equivalently in algebraic terms:

$$\left|z^*_{ih}\right| < |\tilde{z}_{ih}| z^*_{ih} \cdot \tilde{z}_{ih} \ge 0 \text{for all } i.$$

Most markets in reality meet this condition, and we henceforth assume that it always holds. This allows us to classify agents in a market *h* in two categories: unrationed agents for which $z^*_{ih} = \tilde{z}_{ih}$, and rationed agents who trade less than they wanted.
The second property we study here is that of market efficiency, or absence of frictions, which corresponds to the idea of exhaustion of all mutually advantageous exchanges: a rationing scheme on a market *h* is *efficient*, or *frictionless*, if one cannot find simultaneously a rationed demander and a rationed supplier in market *h*. The intuitive idea behind this is that in an efficiently organized market a rationed buyer and a rationed seller would meet and exchange until one of the two is not rationed. Together with voluntary exchange, it implies the ‘short-side rule’, according to which agents on the ‘short side’ of the market can realize their desired transactions:

$$\tilde{D}_h \ge \tilde{S}_h \Rightarrow s^*_{ih} = \tilde{s}_{ih} \text{for all } i \tilde{S}_h \ge \tilde{D}_h \Rightarrow d^*_{ih} = \tilde{d}_{ih} \text{for all } i$$

This rule also implies that the global level of transactions on a market *h* will be equal to the minimum of aggregate demand and supply:

$$D_h^* = S_h^* = \min\ (\tilde{D}_h, \tilde{S}_h)$$

We should note that the market efficiency assumption may not always hold, notably if one considers a fairly wide and decentralized market, because some demanders and suppliers might not meet pairwise. In particular, the market efficiency property is usually lost by aggregation of sub-markets, whereas the voluntary exchange property remains intact in the aggregation process. So we must keep in mind that it does not always hold. Fortunately, this hypothesis is not necessary for most of the microeconomic concepts presented in the next sections.

Quantity signals

Now it is clear that at least rationed agents must perceive a quantity constraint in addition to the price signal. As it turns out, these quantity signals appear quite naturally in the formulation of a number of rationing schemes called *non-manipulable*, which can be written under the form:

$$d_{ih}^* = \min(\tilde{d}_{ih}, \bar{d}_{ih})\ s_{ih}^* = \min(\tilde{s}_{ih}, \bar{s}_{ih}) \quad (5)$$

where the quantity signals $\bar{d}_{ih}$ and $\bar{s}_{ih}$ are functions only of the demands and supplies of the other agents. As an example, we can note that the rationing scheme corresponding to eq. (4) above is of this type with:

$$\bar{d}_{1h} = \tilde{s}_{2h}\bar{s}_{2h} = \tilde{d}_{1h}$$

For non-manipulable schemes the relation between z_{ih}^* and $\tilde{z}_{ih}$ looks as in Figure 1, in which we see where the term ‘non-manipulable’ comes from: once rationed, the agent cannot increase, or ‘manipulate’, the level of his transactions by increasing his demand and supply.

Figure 1

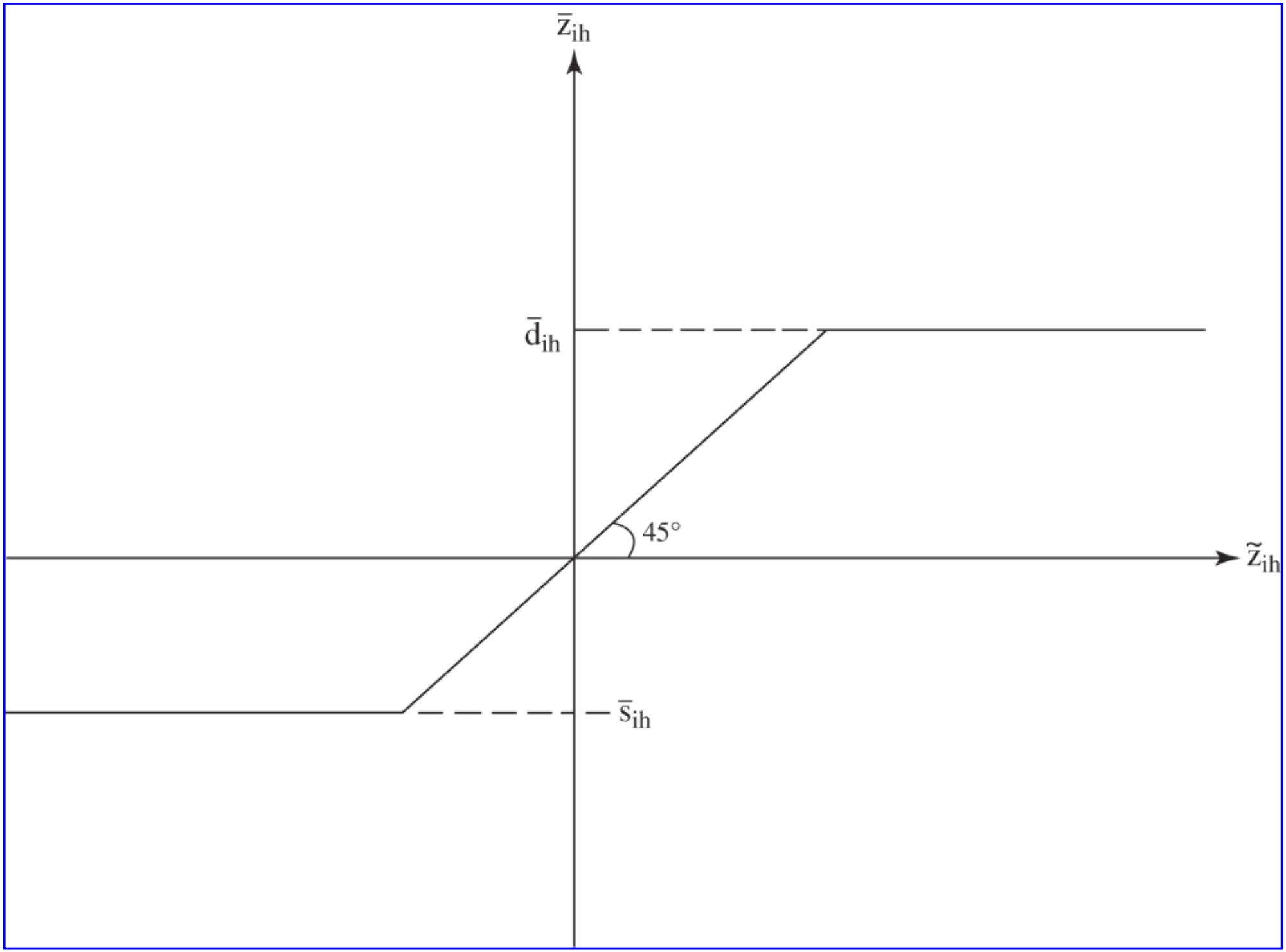

To make things a little more precise, let us rewrite the rationing scheme in market *h* (eq. (3)) under the form:

$$z^*_{ih} = F_{ih}(z_{ih}, z_{-ih})$$
(6)

where z_{-ih} is the set of all net demands on market *h*, except that of agent *i*, that is, $z_{-ih} = \{z_{jh} | j \neq i\}$. The rationing scheme is non-manipulable if it can be rewritten as in eqs. (5), or algebraically:

$$F_{ih}(z_{ih}, z_{-ih}) = \begin{cases} \min(z_{ih}, \bar{d}_{ih}) & z_{ih} \geq 0 \\ \max(z_{ih}, -\bar{s}_{ih}) & z_{ih} \geq 0 \end{cases}$$

where $\bar{d}_{ih}$ and $\bar{s}_{ih}$ are functions of all demands and supplies in market *h*, except that of agent *i*, which we shall write as:

$$\bar{d}_{ih} = G^d_{ih}(z_{-ih}) \geq 0 \quad \bar{s}_{ih} = G^s_{ih}(z_{-ih}) \geq 0$$
(7)

Note that the functions $G^d_{ih}(z_{-ih})$ and $G^s_{ih}(z_{-ih})$ are not arbitrary, but are related to the rationing scheme F_{ih} through:

$$G^d_{ih}(z_{-ih}) = \max\{z_{ih} | F_{ih}(z_{ih}, z_{-ih}) = z_{ih}\}$$
(8)

$$G^s_{ih}(z_{-ih}) = -\min\{z_{ih} | F_{ih}(z_{ih}, z_{-ih}) = z_{ih}\}$$
(9)

where it appears clearly that these quantity constraints are indeed the maximum purchase and sale that agent *i* can make in market *h*.

We may note that some rationing schemes, called *manipulable*, such as the proportional rationing scheme, cannot be written under this form. The phenomenon of manipulation through demand and supply leads then to a perverse phenomenon of overbidding, and to the non-existence of an equilibrium unless additional constraints are put on demands and supplies (Bénassy, 1977b; 1982).

Most rationing schemes in the real world are actually non-manipulable through demand and supply, and we thus from now on study only such rationing schemes as can be

characterized by eqs. (5) or (7). The variables $\bar{d}_{ih}$ and $\bar{s}_{ih}$ in (5) and (7) are *quantity constraints*. These are the quantity signals that each agent receives, and they play a fundamental role in both quantity and price determination, as we see in the next two sections. Before moving to the study of these problems and to the definition of non Walrasian equilibria, it is useful to rewrite eqs. (6) and (7) pertaining to an agent *i* under vector form:

$$z_i^* = F_i(\tilde{z}_i, \tilde{z}_{-i})\,\bar{d}_i = G_i^d(\tilde{z}_{-i})\,\bar{s}_i = G_i^s(\tilde{z}_{-i}) \tag{10}$$

where $\tilde{z}_i$ is the vector of $\tilde{z}_{ih}$, $h = 1, \ldots, \ell$, and $\tilde{z}_{-i}$ is the set of all such vectors, except that of agent *i* himself, i.e. $\tilde{z}_{-i} = \{\tilde{z}_j \mid j \neq i\}$.

Fixprice equilibria

We now study a first concept of non-Walrasian equilibrium, that of fixprice equilibrium. This concept is of interest for several reasons. First, it gives us a very large class of consistent market allocations, since we shall find that under very standard conditions a fixprice equilibrium exists for every positive price system and every set of rationing schemes (we may note that Walrasian allocations are particular fixprice allocations, specifically those corresponding to a Walrasian price vector). Second, as we see in the next section, fixprice equilibria are a very important building block in constructing other non-Walrasian equilibrium concepts with flexible prices.
We thus assume that the price system *p* is given. As indicated, we assume that the rationing schemes in all markets are non-manipulable. Accordingly, transactions and quantity signals are generated in all markets according to the formulas seen above (eqs. (10)). We immediately see that all that remains to be done in order to obtain a fixprice equilibrium concept is to determine how demands themselves are formed, a task to which we now turn.

Effective demands and supplies

Demands and supplies are signals that agents send to the ‘market’ (that is, to the other agents) in order to obtain the best transactions. Consider thus an agent *i* faced with a price vector *p* and vectors of quantity constraints, $\bar{d}_i$ and $\bar{s}_i$. He knows that his transactions will be related to his demands and supplies by formulas (5) seen above, namely,

$$d_{ih}^* = \min(\tilde{d}_{ih}, \bar{d}_{ih})\; s_{ih}^* = \min(\tilde{s}_{ih}, \bar{s}_{ih})$$

Now the problem is to choose a vector of net effective demands $\tilde{z}_i$ which will lead him to the best possible transactions. As it turns out, there exists a simple and workable definition which generalizes Clower's (1965) original ‘dual decision’ method: the effective demand of agent *i* on market *h* is the trade which maximizes his utility subject to the budget constraints and to the quantity constraints on the *other* markets. Formally the effective demand $\tilde{z}_{ih}$ is solution in z_{ih} of the following programme:

$$\text{Maximize } U_i(\omega_i + z_i, m_i) \text{ s.t.}$$
$$\begin{cases} pz_i + m_i = \bar{m}_i \\ -\bar{s}_{ik} \leq z_{ik} \leq \bar{d}_{ik} \; k \neq h \end{cases}$$

Because of the strict concavity of U_i, we obtain a function, denoted $\tilde{\xi}_{ih}(p, \bar{d}_i, \bar{s}_i)$. Repeating the operation for all markets $h = 1, \ldots \ell$, we obtain a vector function of effective demands $\tilde{\xi}_i(p, \bar{d}_i, \bar{s}_i)$. This vector of effective demands has two good properties. First, it leads to the best transactions that it is possible to attain given the price vector *p* and the quantity constraints $\bar{d}_i$ and $\bar{s}_i$. Second, whenever a constraint is binding on a market *h*, the corresponding demand or supply is greater than the quantity constraint, which thus 'signals' to the market that the agent trades less than he would want. Such signals are useful to avoid trivial equilibria where no one would trade because nobody else signals that he wants to trade.

Fixprice equilibrium

With the above definition of effective demand, we are now ready to give a first definition of a fixprice equilibrium, found in Bénassy (1975a; 1982).

Definition 1: A fixprice equilibrium associated with a price system p and rationing schemes represented by functions F_i, $i = 1, \ldots, n$, is a set of effective demands $\tilde{z}_i$, transactions z_i^* and quantity constraints $\bar{d}_i$ and $\bar{s}_i$ such that:

1. (a) $\tilde{z}_i = \tilde{\xi}_i(p, \bar{d}_i, \bar{s}_i)\ i = 1, \ldots, n$
2. (b) $z_i^* = F_i(\tilde{z}_i, \tilde{z}_{-i})\ i = 1, \ldots, n$
3. (c) $\bar{d}_i = G_i^d(\tilde{z}_{-i})\ \bar{s}_i = G_i^s(\tilde{z}_{-i})\ i = 1, \ldots, n$

Equilibria defined in this way exist for all positive prices and all rationing schemes satisfying voluntary exchange and non-manipulability (Bénassy, 1975a; 1982). The 'exogenous' data are the price system *p* and the rationing schemes F_i, $i = 1, \ldots, n$. One may wonder whether for given such exogenous data the equilibrium is likely to be unique. A positive answer has been given by Schulz (1983), who showed that the equilibrium is globally unique, provided the 'spillover' effects (there is a spillover effect when a binding constraint in one market modifies the effective demand in another market) are less than 100 per cent in value terms. For example in the simplest Keynesian model this would amount to a propensity to consume strictly smaller than 1.

In what follows we assume that the Schulz conditions hold, and denote by $\tilde{z}_i(p)$, $z_i^*(p)$, $\bar{D}_i(p)$ and $\bar{S}_i(p)$ the functions giving the values of $\tilde{z}_i$, z_i^*, $\bar{d}_i$ and $\bar{s}_i$ at a fixprice equilibrium corresponding to *p* (the market organization, and thus the rationing schemes, being assumed invariant).

An alternative concept

We shall now present an alternative concept of fixprice equilibrium, due to Drèze (1975) (who actually dealt with the more general case of prices variable between fixed limits), and which we shall recast using our notations. That concept does not separate demands from transactions, and thus considers directly the vectors of transactions z_i^* and quantity constraints $\bar{d}_i$ and $\bar{s}_i$. The original concept actually assumed uniform rationing, so that the vectors $\bar{d}_i$ and $\bar{s}_i$ were the same for all agents.

Definition 2: A fixprice equilibrium for a given set of prices p is defined as a set of transactions z_i^* and quantity constraints $\bar{d}_i$ and $\bar{s}_i$ such that:

1. (a) $\sum_{i=1}^{n} z_{ih}^* = 0\ \forall h$
2. (b) *The vector z_i^* is solution in z_i of:*

$$\text{Maximize } U_i(\omega_i + z_i, m_i) \text{ s.t.} \begin{cases} pz_i + m_i = \bar{m}_i \\ -\bar{s}_{ih} \le z_{ih} \le \bar{d}_{ih}\ \forall\ h \end{cases}$$

3. (c) $\forall h\ z^*_{ih} = \bar{d}_{ih}$ *for some i implies* $z^*_{jh} > -\bar{s}_{jh}\ \forall j\ z^*_{ih} = -\bar{s}_{ih}$ *for some i implies* $z^*_{jh} < -d_{jh}\ \forall j$

Let us now interpret these conditions. Condition (*a*) is the natural requirement that transactions should balance in each market. Condition (*b*) says that transactions must be individually rational, that is, they must maximize utility subject to the budget constraint and the quantity constraints on all markets. We may note at this stage that using quantity constraints under the form of upper and lower bounds on trades implicitly assumes rationing schemes which exhibit voluntary exchange and non-manipulability, as we saw when studying rationing schemes. Condition (*c*) says that rationing may affect either supply or demand, but not both simultaneously. We recognize here with a different formalization the condition of market efficiency which is thus built into this definition of equilibrium, whereas it is not in the previous definition.

Drèze (1975) proved that an equilibrium according to def 2 exists for all positive price systems and for uniform rationing schemes under the standard concavity assumptions for the utility functions. The concept is easily extended to non-uniform bounds (Grandmont and Laroque, 1976; Greenberg and Muller, 1979), but in this last case it is not specified in the concept how shortages are allocated. Because of this there will be usually an infinity of equilibria corresponding to a given price vector, as soon as there are two rationed agents on one side of a market.

As we noted above, the two concepts of fixprice equilibrium we described in this section are based, implicitly or explicitly, on a representation of markets under the form of rationing schemes satisfying voluntary exchange and non-manipulability. This suggests that, if in the first definition we further assume that all rationing schemes are efficient or frictionless, the two definitions should yield similar sets of equilibrium allocations for a given price system. This was indeed proved by Silvestre (1982; 1983) for both exchange and production economies. The relation between the two concepts has been further explored by D'Autume (1985).

Price making and equilibrium

As this stage we still need a description of price making by agents internal to the system. We describe in this section a concept dealing with that problem and we shall see that, just as in demand and supply theory, quantity signals play a prominent role. It is indeed quite intuitive that quantity constraints must be a fundamental part of the competitive process in a decentralized economy: it is the inability to sell as much that they want which leads suppliers to propose, or accept from other agents, a lower price, and conversely it is the inability to purchase as much as they want that leads demanders to propose, or accept, a higher price.

Various modes of price making integrating these aspects can be envisioned. We deal here with a realistic organization of the pricing process where agents on one side of the market (most often the suppliers) quote prices and agents on the other side act as price takers. The general idea relating the concepts in this section to those of the previous one is that price makers change their prices so as to ‘manipulate’ the quantity constraints they face (that is, so as to increase or decrease their possible sales or purchases). As we shall see, this model of price making is quite reminiscent of the imperfect competition line (Chamberlin, 1933; Robinson, 1933; Triffin, 1940; Bushaw and Clower, 1957; Arrow, 1959), and more particularly of the theories of general equilibrium with monopolistic competition, as developed notably by Negishi (1961).

The framework

We thus now assume that agent *i* controls the prices of a (possibly empty) subset H_i of goods. Goods are distinguished both by their physical characteristics and by the agent who sets their price. We thus consider two goods sold by different sellers as different goods, a fairly natural assumption since these goods differ at least by location, quality, and so on, so that:

$$H_i \cap H_j = \{\varnothing\}\ i \neq j$$

We denote by p_i the set of prices controlled by agent *i* and p_{-i} the rest of prices, that is:

$$p_i = \{p_h | h \in H_i\}\ p_{-i} = \{p_h | h \notin H_i\}$$

Each agent chooses his price vector p_i taking the other prices p_{-i} as given. The equilibrium structure is thus that of a Nash equilibrium in prices, corresponding to an idea close to that of monopolistic competition. The basic idea behind the modelling of price making itself in such models is, as we indicated above, that each price maker uses the prices he controls to 'manipulate' the quantity constraints he faces. Consider the markets whose price are determined by agent *i*, and subdivide further H_i into H_i^d (goods demanded by *i*) and H_i^s (goods supplied by *i*). We may note in passing that, although agent *i* appears formally as a monopolist in markets $h \in H_i^s$ or a monopsonist in markets $h \in H_i^d$, his actual 'monopoly power' may be very low due to the fact that other agents sell or buy products which are extremely close substitutes to those he controls. Because the price makers are alone on their side of the markets where they set prices, their quantity constraints on these markets have the simple form:

$$\bar{s}_{ih} = \sum_{j \neq i} \tilde{d}_{jh} h \in H_i^s \bar{d}_{ih} = \sum_{j \neq i} \tilde{s}_{jh} h \in H_i^d$$

that is, the maximum quantity that price setter *i* can sell is the total demand of the others, and conversely if he is a buyer. All we need to know, in order to pose the problem of price setting as a standard decision problem, is the relation, as perceived by the price maker, between the quantity constraints he faces and the prices he sets. Several approaches allow us to treat this problem and to link it with the concepts seen previously. The first, based on Negishi's (1961) subjective demand curve approach, was developed in Bénassy (1976; 1982). The second is an objective demand curve approach, developed in Bénassy (1987; 1988), and which we shall now briefly describe.

Objective demand curves

The implicit idea behind the objective demand curve approach (Gabszewicz and Vial, 1972; Marschak and Selten, 1976; Nikaido, 1975) is that each price maker knows the economy well enough to be able to compute under all circumstances the actual quantity constraints he will face. Since we are considering a Nash equilibrium, he must be able to perform this computation for any set p_i of prices he chooses as well as for any set p_{-i} of the other prices; that is, he must be able to compute his constraints for any vector of prices, once all feedback effects have been accounted for.

But we know from the previous section that, for a given organization of the economy (that is, notably for given rationing schemes), and for a given set of prices *p*, the quantity constraints agent *i* faces are given by the functions $\bar{D}_i(p)$ and $\bar{S}_i(p)$. If the agent has full knowledge of the parameters of the economy (a strong assumption, of course, but which is embedded in the notion of an objective demand curve), then he knows this and the objective demand and supply curves will be respectively given by the functions $\bar{S}_i(p)$ and $\bar{D}_i(p)$. We may note that the objective demand curve $\bar{S}_i(p)$ is denoted as a constraint on agent *i's* supply, which is natural since the sum of all other agents' demands acts as a constraint on the sales of agent *i*, and symmetrically with the objective supply curve $\bar{D}_i(p)$.

Price making and equilibrium

If agent *i* knows the two vector functions $\bar{D}_i(p)$ and $\bar{S}_i(p)$, the programme giving his optimal price p_i is the following:

$$\text{Maximize } U_i(\omega_i + z_i, m_i) \text{ s.t.}$$
$$\begin{cases} pz_i + m_i = \bar{m}_i \\ -\bar{S}_i(p) \leq z_i \leq \bar{D}_i(p) \end{cases}$$

which yields the optimum price p_i chosen by agent i as a function of the other prices p_{-i}.

$$p_i = \psi_i(p_{-i})$$

This naturally leads us to the definition of an equilibrium with price makers:

Definition 3: An equilibrium with price makers is characterized by a set of prices p_i^*, *net demands* $\tilde{z}_i$, *transactions* z_i^* *and quantity constraints* $\bar{d}_i$ *and* $\bar{s}_i$ *such that:*

1. (a) $p_i^* = \psi_i(p_{-i}^*)$
2. (b) $\tilde{z}_i$, z_i^*, $\bar{d}_i$, $\bar{s}_i$ are equal respectively to $\tilde{z}_i(p^*)$, $z_i^*(p^*)$, $\bar{D}_i(p^*)$, $\bar{S}_i(p^*)$.

Condition (*a*) indicates that we have a Nash equilibrium in prices, given each agent's optimal price responses. Condition (*b*) says that the various quantities form a fixprice equilibrium (according to def 1) for the price vector p^*. Further discussion and conditions for existence can be found in Bénassy (1988; 1990).

Bibliographical references

So far we have concentrated in this entry on the microeconomic concepts allowing us to deal with non-clearing markets at a general equilibrium level. We now indicate further bibliographical references both on the early history of the domain and on macroeconomic applications.

History

The field we described in this entry has a triple ancestry. On one hand Walras (1874) developed a model of general equilibrium with interdependent markets where adjustment was made through prices. This model, in its modern reformulation (Arrow and Debreu, 1954; Arrow, 1963; Debreu, 1959) has become the basic benchmark concept in microeconomics. On the other hand Keynes (1936) and Hicks (1937) built, at the macroeconomic level, a concept of equilibrium where adjustment was made by quantities (the level of national income) as well as by prices. Finally, following the contributions by Chamberlin (1933) and Robinson (1933), progress was made on the treatment of imperfect competition. Notably, Negishi (1961) formalized imperfect competition with subjective demand curves in a general equilibrium framework.

A few isolated contributions in the post-war period made some steps towards modern theories of non-clearing markets. Bent Hansen (1951) introduced the ideas of active demand, close in spirit to that of effective demand, and of quasi-equilibrium where persistent disequilibrium created steady inflation. Patinkin (1956, ch. 13) considered the situation where firms might not be able to sell all their Walrasian output. Hahn and Negishi (1962) studied non-tâtonnement processes where trade could take place before a general equilibrium price system was reached.

A stimulating impetus came from the contributions of Clower (1965) and Leijonhufvud (1968), who reinterpreted Keynesian analysis in terms of market rationing and quantity adjustments. These insights were included in the first fixprice-fixwage macroeconomic model by Barro and Grossman (1971; 1976).

The main subsequent development was the construction of rigorous microeconomic concepts allowing us to deal with non-clearing markets and imperfect competition in a full multi-market general equilibrium setting, as described above. Notably, Drèze (1975) and Bénassy (1975a; 1977b; 1982) bridged the gap between the Walrasian and Keynesian lines of thought by generalizing the Walrasian equilibrium concept to integrate non-clearing markets and quantity signals. The link between this new line of work and the imperfect competition equilibrium concepts in the Negishi (1961) line was made in Bénassy (1976; 1977a; 1988). These contributions led to the unified framework we set out in the previous sections. Of course, since one of the main goals of this line of research was to bridge the gap between microeconomics and macroeconomics, there were a number of macroeconomic applications of the above concepts, which we now briefly describe.

Macroeconomic applications

As indicated above, the first fully worked out fixprice-fixwage macroeconomic model embedding the notions set out above is that of Barro and Grossman (1971; 1976). Early attempts are found in Glustoff (1968) and Solow and Stiglitz (1968). Further developments of the model were made in Bénassy (1977a; 1982; 1986), Malinvaud (1977), Hildenbrand and Hildenbrand (1978), Muellbauer and Portes (1978), Honkapohja (1979), Neary and Stiglitz (1983), and Persson and Svensson (1983). Most of these models concentrated on the problem of employment and policy. Other problems have been treated with this methodology, including notably foreign trade (Dixit, 1978; Neary, 1980; Cuddington, Johansson and Löfgren, 1984), growth (Ito, 1980; Picard, 1983; D'Autume, 1985), business cycles (Bénassy, 1984), as well as the specific problems of planned socialist economies (Portes, 1981).
An important part of this line of macroeconomic modelling is that concerned with the explicit introduction of price making and imperfect competition in the macro-setting. Models of that type can be found notably in Bénassy (1977a; 1982; 1987; 1990; 1991), Negishi (1977; 1979), Hart (1982), Snower (1983) Weitzman (1985), Svensson (1986), Blanchard and Kiyotaki (1987), Dixon (1987), Sneessens (1987), Silvestre (1988) and Jacobsen and Schultz (1990).
Now the concepts described in this entry are full general equilibrium models in the tradition of, say, Arrow and Debreu (1954) and Debreu (1959). Contemporaneously to these developments, other authors developed, under the initial name of real business cycles, dynamic stochastic models based on the hypothesis of rational expectations. At some point these two lines of work were synthesized, and the result of this synthesis is described in the dictionary article ‘dynamic models with non-clearing markets’.

See Also

- dynamic models with non-clearing markets
- fixprice models

Bibliography

Arrow, K. 1959. Towards a theory of price adjustment. In *The Allocation of Economic Resources*, ed. M. Abramowitz. Stanford: Stanford University Press.

Arrow, K. 1963. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31, 91–6.

Arrow, K. and Debreu, G. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–90.

Barro, R. and Grossman, H. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

Barro, R. and Grossman, H. 1976. *Money, Employment and Inflation*. Cambridge: Cambridge University Press.

Bénassy, J.-P. 1975a. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42, 503–23.

Bénassy, J.-P. 1975b. Disequilibrium exchange in barter and monetary economies. *Economic Inquiry* 13, 131–56.

Bénassy, J.-P. 1976. The disequilibrium approach to monopolistic price setting and general monopolistic equilibrium. *Review of Economic Studies* 43, 69–81.

Bénassy, J.-P. 1977a. A Neo-Keynesian model of price and quantity determination in disequilibrium. In *Equilibrium and Disequilibrium in Economic Theory*, ed. G. Schwödiauer. Boston: Reidel Publishing Company.

Bénassy, J.-P. 1977b. On quantity signals and the foundations of effective demand theory. *Scandinavian Journal of Economics* 79, 147–68.

Bénassy, J.-P. 1982. *The Economics of Market Disequilibrium*. New York: Academic Press.

Bénassy, J.-P. 1984. A non-Walrasian model of the business cycle. *Journal of Economic Behavior and Organization* 5, 77–89.

Bénassy, J.-P. 1986. *Macroeconomics: An Introduction to the Non-Walrasian Approach*. Orlando, FL: Academic Press.

Bénassy, J.-P. 1987. Imperfect competition, unemployment and policy. *European Economic Review* 31, 417–26.

Bénassy, J.-P. 1988. The objective demand curve in general equilibrium with price makers. *Economic Journal* 98, 37–49.

Bénassy, J.-P. 1990. Non Walrasian equilibria, money and macroeconomics. In *Handbook of Monetary Economics*, ed. B. Friedman and F. Hahn. Amsterdam: North-Holland.

Bénassy, J.-P. 1991. Microeconomic foundations and properties of a macroeconomic model with imperfect competition. In *Issues in Contemporary Economics, Volume 1: Markets and Welfare*, ed. K. Arrow. London: Macmillan.

Blanchard, O. and Kiyotaki, N. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77, 647–66.

Bushaw, D. and Clower, R. 1957. *Introduction to Mathematical Economics*. Homewood, IL: Richard D. Irwin.

Chamberlin, E. 1933. *The Theory of Monopolistic Competition*, 7th edn. Cambridge, MA: Harvard University Press, 1956.

Clower, R. 1965. The Keynesian counterrevolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cuddington, J., Johansson, P.-O. and Löfgren, K. 1984. *Disequilibrium Macroeconomics in Open Economies*. Oxford: Basil Blackwell.

D'Autume, A. 1985. *Monnaie, Croissance et Déséquilibre*. Paris: Economica.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Dixit, A. 1978. The balance of trade in a model of temporary equilibrium with rationing. *Review of Economic Studies* 45, 393–404.

Dixon, H. 1987. A simple model of imperfect competition with Walrasian features. *Oxford Economic Papers* 39, 134–60.

Drèze, J. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16, 301–20.

Gabszewicz, J. and Vial, J.-P. 1972. Oligopoly ‘A la Cournot’ in a general equilibrium analysis. *Journal of Economic Theory* 42, 381–400.

Glustoff, E. 1968. On the existence of a Keynesian equilibrium. *Review of Economic Studies* 35, 327–34.

Grandmont, J.-M. and Laroque, G. 1976. On Keynesian temporary equilibria. *Review of Economic Studies* 43, 53–67.

Greenberg, J. and Muller, H. 1979. Equilibria under price rigidities and externalities. In *Game Theory and Related Topics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.

Hahn, F. and Negishi, T. 1962. A theorem on non tatonnement stability. *Econometrica* 30, 463–69.

Hansen, B. 1951. *A Study in the Theory of Inflation*. London: Allen and Unwin.

Hart, O. 1982. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97, 109–38.

Hicks, J. 1937. Mr. Keynes and the classics: a suggested inpt. *Econometrica* 5, 147–59.

Hildenbrand, K. and Hildenbrand, W. 1978. On Keynesian equilibria with unemployment and quantity rationing. *Journal of Economic Theory* 18, 255–77.

Honkapohja, S. 1979. On the dynamics of disequilibria in a macro model with flexible wages and prices. In *New Trends in Dynamic System Theory and Economics*, ed. M. Aoki and A. Marzollo. New York: Academic Press.

Ito, T. 1980. Disequilibrium growth theory. *Journal of Economic Theory* 23, 380–409.

Jacobsen, H.-J. and Schultz, C. 1990. A general equilibrium macro model with wage bargaining. *Scandinavian Journal of Economics* 92, 379–98.

Keynes, J.M. 1936. *The General Theory of Money, Interest and Employment*. New York: Harcourt Brace.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. Oxford: Oxford University Press.

Malinvaud, E. 1977. *The Theory of Unemployment Reconsidered*. Oxford: Basil Blackwell.

Marschak, T. and Selten, R. 1974. *General Equilibrium with Price-Making Firms*. Berlin: Springer-Verlag.

Muellbauer, J. and Portes, R. 1978. Macroeconomic models with quantity rationing. *Economic Journal* 88, 788–821.

Neary, P. 1980. Nontraded goods and the balance of trade in a neo-Keynesian temporary equilibrium. *Quarterly Journal of Economics* 95, 403–30.

Neary, P. and Stiglitz, J. 1983. Towards a reconstruction of Keynesian economics: expectations and constrained equilibria. *Quarterly Journal of Economics* 98, 199–228.

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28, 196–201.

Negishi, T. 1977. Existence of an underemployment equilibrium. In *Equilibrium and Disequilibrium in Economic Theory*, ed. G. Schwödiauer. Boston: D. Reidel Publishing Company.

Negishi, T. 1979. *Microeconomic Foundations of Keynesian Macroeconomics*. Amsterdam: North-Holland.

Nikaido, H. 1975. *Monopolistic Competition and Effective Demand*. Princeton: Princeton University Press.

Patinkin, D. 1956. *Money, Interest and Prices*, 2nd edn. New York: Harper and Row, 1965.

Persson, T. and Svensson, L. 1983. Is optimism good in a Keynesian economy? *Economica* 50, 291–300.

Picard, P. 1983. Inflation and growth in a disequilibrium macroeconomic model. *Journal of Economic Theory* 30, 266–95.

Portes, R. 1981. Macroeconomic equilibrium and disequilibrium in centrally planned economies. *Economic Inquiry* 19, 559–78.

Robinson, J. 1933. *The Economics of Imperfect Competition*, 2nd edn. London: Macmillan, 1969.

Schulz, N. 1983. On the global uniqueness of fixprice equilibria. *Econometrica* 51, 47–68.

Silvestre, J. 1982. Fixprice analysis of exchange economies. *Journal of Economic Theory* 26, 28–58.

Silvestre, J. 1983. Fixprice analysis in productive economies. *Journal of Economic Theory* 30, 401–9.

Silvestre, J. 1988. Undominated prices in the three good model. *European Economic Review* 32, 161–78.

Sneessens, H. 1987. Investment and the inflation–unemployment trade off in a macroeconomic rationing model with monopolistic competition. *European Economic Review* 31, 781–815.

Snower, D. 1983. Imperfect competition, unemployment and crowding out. *Oxford Economic Papers* 35, 569–84.

Solow, R. and Stiglitz, J. 1968. Output, employment and wages in the short run. *Quarterly Journal of Economics* 82, 537–60.

Svensson, L. 1986. Sticky goods prices, flexible asset prices, monopolistic competition and monetary policy. *Review of Economic Studies* 53, 385–405.

Triffin, R. 1940. *Monopolistic Competition and General Equilibrium Theory*. Cambridge, MA: Harvard University Press.

Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz. Trans. W. Jaffe, *Elements of Pure Economics*. London: Allen and Unwin, 1954.

Weitzman, M. 1985. The simple macroeconomics of profit sharing. *American Economic Review* 75, 937–53.

How to cite this article

Bénassy, Jean-Pascal. "non-clearing markets in general equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000153> doi:10.1057/9780230226203.1190

The New Palgrave Dictionary of Economics Online

non-cooperative games (equilibrium existence)

Philip J. Reny
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides a brief overview of equilibrium existence results for continuous and discontinuous non-cooperative games.

Keywords

auctions; Bertrand price-competition models; convexity; Cournot oligopoly models; discontinuous games; endogenous sharing rules; equilibrium; equilibrium existence; finite-action games; fixed point theorems; infinite-action games; mixed strategy Nash equilibria; Nash equilibrium; non-cooperative games; pure strategy Nash equilibria; quasi-concavity; quasi-convexity; spatial economics; strategic form games

Article

1 Introduction

Nash equilibrium is *the* central notion of rational behaviour in non-cooperative game theory (see Osborne and Rubinstein, 1994, for a discussion of Nash equilibrium, including motivation and inpt). Our purpose here is to discuss various conditions under which a strategic form game possesses at least one Nash equilibrium.

Strategic settings arising in economics are often naturally modelled as games with infinite strategy spaces. For example, models of price and spatial competition (Bertrand, 1883; Hotelling, 1929), quantity competition (Cournot, 1838), auctions (Milgrom and Weber, 1982), patent races (Fudenberg et al., 1983), and so on, typically allow players to choose any one of a continuum of actions. The analytic convenience of the continuum from both an equilibrium characterization and a comparative statics point of view is perhaps the central reason for the prevalence and usefulness of infinite-action games. Because of this, our treatment will permit both finite-action and infinite-action games.

Games with possibly infinite strategy spaces can be divided into two categories: those with continuous payoffs and those with discontinuous payoffs. Cournot oligopoly models and Bertrand price-competition models with differentiated products, as well as all finite-action games, are important examples of continuous games, while Bertrand price-competition with homogeneous products, auctions, and Hotelling spatial competition are important examples in which payoffs are discontinuous. Equilibrium existence results for both continuous and discontinuous games will be reviewed here. We begin with some notation.

A strategic form game, $G = (S_i, u_i)_{i=1}^N$, consists of a positive finite number, *N*, of players, and for each player $i \in \{1, \ldots, N\}$, a non-empty set of pure strategies, S_i, and a payoff function $u_i : S \to \mathbb{R}$, where $S = \times_{i=1}^N S_i$. The notations s_{-i} and S_{-i} have their conventional meanings: $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_N)$ and $S_{-i} = \times_{j \neq i} S_j$. Throughout, we assume that each S_i is a subset of some metric space and that, if any finite number of sets are each endowed with a topology, then the product of those sets is endowed with the product topology.

2 Continuous games

2.1 Pure strategy Nash equilibria

Pure strategy equilibria are more basic than their mixed strategy counterparts for at least two reasons. First, pure strategies do not require the players to possess preferences over lotteries. Second, mixed strategy equilibrium existence results often follow as corollaries of the pure strategy results. It is therefore natural to consider first the case of pure strategies.

Definition: . $s^* \in S$ is a pure strategy Nash equilibrium of $G = (S_i, u_i)_{i=1}^N$ if for every player i, $u_i(s^*) \geq u_i(s_i, s^*_{-i})$ for every $s_i \in S_i$.

An important and very useful result is the following.

Theorem 1: If each S_i is a non-empty, compact, convex subset of a metric space, and each $u_i(s_1, \ldots, s_N)$ is continuous in $(s_1, \ldots, s_N)$ and quasi-concave in s_i, then $G = (S_i, u_i)_{i=1}^N$ possesses at least one pure strategy Nash equilibrium.

Proof: . For each player *i*, and each $s_{-i} \in S_{-i}$, let $B_i(s_{-i})$ denote the set of maximizers in S_i of $u_i(\cdot, s_{-i})$. The continuity of u_i and the compactness of S_i ensure that $B_i(s_{-i})$ is non-empty and also ensure, given the compactness of S_{-i}, that the correspondence, $B_i : S_{-i} \twoheadrightarrow S_i$ is upper hemi-continuous. The quasi-concavity of u_i in s_i implies that $B_i(s_{-i})$ is convex. Consequently, each B_i is upper hemi-continuous, non empty-valued and convex-valued. All three of these properties are therefore inherited by the correspondence $B : S \twoheadrightarrow S$ defined by $B(s) = \times_{i=1}^N B_i(s_{-i})$ for each $s \in S$. Consequently, we may apply Glicksberg's (1952) fixed point theorem to *B* and conclude that there exists $\hat{s} \in S$ such that $\hat{s} \in B(\hat{s})$. This $\hat{s}$ is therefore a pure strategy Nash equilibrium. Q.E.D.

Remark: . Theorem 1 remains valid when ‘metric space’ is replaced by ‘locally convex Hausdorff topological vector space’. See Glicksberg (1952).

Remark: . The convexity property of strategy sets and the quasi-concavity of payoffs in own action cannot be dispensed with. For example, strategy sets are not convex in matching pennies, and, even though the continuity and compactness assumptions hold there, no pure strategy equilibrium exists. On the other hand, in the two-person zero-sum game in which both players' compact convex pure strategy set is [−1,1] and player 1's payoff function is $u_1(s_1, s_2) = |s_1 + s_2|$, all of the assumptions of Theorem 1 hold except the quasi-concavity of u_1 in s_1. But this is enough to preclude the existence of a pure strategy equilibrium because in any such equilibrium player 2's payoff would have to be zero (given s_1, 2 can choose $s_2 = -s_1$) and 1's payoff would have to be positive (given s_2, 1 can choose $s_1 \neq -s_2$).
Remark: . More general results for continuous games can be found in Debreu (1952) and Schafer and Sonnenschein (1975). Existence results for games with strategic complements on lattices can be found in Milgrom and Roberts (1990) and Vives (1990).

2.2 Mixed strategy Nash equilibria

A *mixed strategy* for player *i* is a probability measure, m_i, over S_i. If S_i is finite, then $m_i(s_i)$ denotes the probability assigned to $s_i \in S_i$ by the mixed strategy m_i, and *i*'s set of mixed strategies is the compact convex subset of Euclidean space $M_i = \{m_i \in [0, 1]^{\# S_i} : \Sigma_{s_i \in S_i} m_i(s_i) = 1\}$.
In general, we shall not require S_i to be finite. Rather, we shall suppose only that it is a subset of some metric space. In this more general case, a mixed strategy for player *i* is a (regular, countably additive) probability measure, m_i, over the Borel subsets of S_i; for any Borel subset *A* of S_i, $m_i(A)$ denotes the probability assigned to *A* by the mixed strategy m_i. Player *i*'s set of such mixed strategies, M_i, is then convex. Further, if S_i is compact, the convex set M_i is compact in the weak-* topology (see, for example, Billingsley, 1968).
Extend $u_i : S \to \mathbb{R}$ to $M = \times_{i=1}^{N} M_i$ by an expected utility calculation (hence, the $u_i(s)$ are assumed to be von Neumann–Morgenstern utilities). That is, define
$u_i(m_1, \ldots, m_N) = \int_{S_1} \cdots \int_{S_N} u_i(s_1, \ldots, s_N) dm_1 \ldots dm_N$ for all $m = (m_1, \ldots, m_N) \in M$. (This is an extension because we view *S* as a subset of *M*; each $s \in S$ is identified with the $m \in M$ that assigns probability one to *s*.) Finally, let $\bar{G} = (M_i, u_i)_{i=1}^{N}$ denote the *mixed extension* of $G = (S_i, u_i)_{i=1}^{N}$.
Definition: . $m^* \in M$ is a mixed strategy Nash equilibrium of $G = (S_i, u_i)_{i=1}^{N}$ if *m** is a pure strategy Nash equilibrium of the mixed extension, $\bar{G}$, of G. That is, if for every player i, $u_i(m^*) \geq u_i(m_i, m^*_{-i})$ for every $m_i \in M_i$.
Because $u_i(m_i, m_{-i})$ is linear and therefore quasi-concave, in $m_i \in M_i$ for each $m_{-i} \in M_{-i}$, and because continuity of $u_i(\cdot)$ on *S* implies continuity of $u_i(\cdot)$ on *M* (in the weak-* topology), Theorem 1 applied to the mixed extension of *G* yields the following basic mixed strategy Nash equilibrium existence result:
Corollary 1: . If each S_i is a non-empty compact subset of a metric space, and each $u_i(s)$ is continuous in $s \in S$, then $G = (S_i, u_i)_{i=1}^{N}$ possesses at least one mixed strategy Nash equilibrium, $m^* \in M$.
Remark: . Note that Corollary 1 does not require $u_i(s_i, s_{-i})$ to be quasi-concave in S_i, nor does it require

the S_i to be convex.

Remark: . Corollary 1 yields von Neumann's (1928) classic result for two-person zero-sum games as well as Nash's (1950; 1951) seminal result for Enite games as special cases. To obtain Nash's result, note that if each S_i is Enite, then each u_i is continuous on S in the discrete metric. Hence, the corollary applies and we conclude that every finite game possesses at least one mixed strategy Nash equilibrium.

Remark: . To see how Theorem 1 can be applied to obtain the existence of mixed strategy equilibria in Bayesian games, see Milgrom and Weber (1985).

Remark: . See Glicksberg (1952) for a generalization to non-metrizable strategy spaces.

3 Discontinuous games

The basic challenge one must overcome in extending equilibrium existence results from continuous games to discontinuous games is the failure of the best reply correspondence to satisfy the properties required for application of a fixed point theorem. For example, in auction or Bertrand price-competition settings, discontinuities in payoffs sometimes preclude the existence of best replies. The best reply correspondence then fails to be non-empty valued, and Glicksberg's theorem, for example, cannot be applied.

A natural technique for overcoming such difficulties is to approximate the infinite strategy spaces by a sequence of finer and finer *finite* approximations. Each of the approximating finite games is guaranteed to possess a mixed strategy equilibrium (by Corollary 1) and the resulting sequence of equilibria is guaranteed, by compactness, to possess at least one limit point. Under appropriate assumptions, the limit point is a Nash equilibrium of the original game. This technique has been cleverly employed in Dasgupta and Maskin's (1986) pioneering work, and also by Simon (1987). However, while this finite approximation technique can yield results on the existence of *mixed* strategy Nash equilibria, it is unable to produce equally general existence results for pure strategy Nash equilibria. The reason, of course, is that the approximating games, being finite, are guaranteed to possess mixed strategy, but not necessarily pure strategy, Nash equilibria. Consequently, the sequence of equilibria, and so also the limit point, cannot be guaranteed to be pure.

One might be tempted to conclude that, unlike the continuous game case where the mixed strategy result is a special case of the pure strategy result, discontinuous games require a separate treatment of pure and mixed strategy equilibria. But such a conclusion would be premature. A connection between pure and mixed strategy equilibrium existence results similar to that for continuous games can be obtained for discontinuous games by considering a different kind of approximation. Rather than approximating the infinite strategy spaces by a sequence of finite approximations, one can instead approximate the discontinuous payoff functions by a sequence of continuous payoff functions. This payoff-approximation technique is employed in Reny (1999), whose main result we now proceed to describe.

All of the definitions, notation, and conventions of the previous sections remain in effect. In particular, each S_i is a subset of some metric space. (This is for simplicity of presentation only. The results to follow hold in non metrizable settings as well. See Reny, 1999.)

3.1 Better-reply security

Definition: . Player *i* can *secure* a payoff of $\alpha \in \mathbb{R}$ at $s \in S$ if there exists $\bar{s}_i \in S_i$, such that $u_i(\bar{s}_i, s'_{-i}) \geq \alpha$ a for all s'_{-i} close enough to s_{-i}.
Thus, a payoff can be secured by *i* at *s* if *i* has a strategy that guarantees at least that payoff even if the other players deviate slightly from *s*.
A pair $(s, u) \in S \times \mathbb{R}^N$ is in the closure of the graph of the vector payoff function if $u \in \mathbb{R}^N$ is the limit of the vector of player payoffs for some sequence of strategies converging to *s*. That is, if $u = \lim_n (u_1(s^n), \ldots, u_N(s^n))$ for some $s^n \to s$.
Definition: . A game $G = (S_i, u_i)_{i=1}^N$ is *better-reply secure* if whenever (*s**,*u**) is in the closure of the graph of its vector payoff function and *s** is not a Nash equilibrium, some player *i* can secure a payoff strictly above u_i^* at *s**.
All games with continuous payoff functions are better-reply secure. This is because if (*s**,*u**) is in the closure of the graph of the vector payoff function of a continuous game, we must have $u^* = (u_1(s^*), \ldots, u_N(s^*))$. Also, if *s** is not a Nash equilibrium then some player *i* has a strategy $\bar{s}_i$ such that $u_i(\bar{s}_i, s^*_{-i}) > u_i(s^*)$, and continuity ensures that this inequality is maintained even if the others deviate slightly from *s**. Consequently, player *i* can secure a payoff strictly above $u_i^* = u_i(s^*)$.
The import of better-reply security is that it is also satisfied in many discontinuous games. For example, Bertrand's price-competition game, many auction games, and many games of timing are better-reply secure.

3.2 Pure strategy Nash equilibria

The following theorem provides a pure strategy Nash equilibrium existence result for discontinuous games.
Theorem 2: (Reny, 1999). If each S_i is a non-empty, compact, convex subset of a metric space, and each $u_i(s_1, \ldots, s_N)$ is quasi-concave in s_i, then $G = (S_i, u_i)_{i=1}^N$ possesses at least one pure strategy Nash equilibrium if in addition *G* is better-reply secure.
Remark: . Theorem 1 is a special case of Theorem 2 because every continuous game is better-reply secure.
Remark: . A classic result due to Sion (1958) states that every two-person zero-sum game with compact strategy spaces in which player 1's payoff is upper-semi-continuous and quasi-concave in his own strategy, and lower-semi-continuous and quasi-convex in the opponent's strategy, has a value and each player has an optimal pure strategy. (Sion does not actually prove the existence of optimal strategies, but this follows rather easily from his compactness assumptions and his result that the game has a value, that is, that infsup=supinf.) It is not difficult to show that Sion's result is a special case of Theorem 2.
Remark: . A related result that weakens quasi-concavity but adds conditions to the sum of the players' payoffs can be found in Baye, Tian and Zhou (1993). Dasgupta and Maskin (1986) provide two interesting pure strategy equilibrium existence results, both of which require each player's payoff

function to upper semi-continuous in the vector of all players' strategies.

3.3 Mixed strategy Nash equilibria

One easily obtains from Theorem 2 a mixed strategy equilibrium existence result (the analogue of Corollary 1) by treating each M_i as if it were player *i*'s pure strategy set and by applying the definition of better-reply security to the mixed extension $\bar{G} = (M_i, u_i)$ of *G*. This observation yields the following result.

Corollary 2: (Reny, 1999). If each S_i is a non-empty, compact, convex subset of a metric space, then $G = (S_i, u_i)_{i=1}^{N}$ possesses at least one mixed strategy Nash equilibrium if in addition its mixed extension, $\bar{G} = (M_i, u_i)$, is better-reply secure.

Remark: . Better-reply security of $\bar{G}$ neither implies nor is implied by better-reply security of *G*. (See Reny, 1999, for sufficient conditions for better-reply security.)

Remark: . Corollary 1 is a special case of Corollary 2 because continuity of each $u_i(s)$ in $s \in S$ implies (weak-*) continuity of $u_i(m)$ in $m \in M$, which implies that the mixed extension, $\bar{G}$, is better-reply secure.

Remark: . Corollary 2 has as special cases the mixed strategy equilibrium existence results of Dasgupta and Maskin (1986), Simon (1987) and Robson (1994).

Remark: . Theorem 2 can similarly be used to obtain a result on the existence of mixed strategy equilibria in discontinuous Bayesian games by following Milgrom and Weber's (1985) seminal distributional strategy approach. One simply replaces Milgrom and Weber's payoff continuity assumption with the assumption that the Bayesian game is better-reply secure in distributional strategies. An example of this technique is provided in the next subsection.

3.4 An application to auctions

Auctions are an important class of economic games in which payoffs are discontinuous. Furthermore, when bidders are asymmetric, in general one cannot prove existence of equilibrium by construction, as in the symmetric case. Consequently, an existence theorem applicable to discontinuous games is called for. Let us very briefly sketch how Theorem 2 can be applied in this case.

Consider a first-price single-object auction with *N* bidders. Each bidder *i* receives a private value $v_i \in [0, 1]$ prior to submitting a sealed bid, $b_i \geq 0$. Bidder *i*'s value is drawn independently according to the continuous and positive density f_i. The highest bidder wins the object and pays his bid. Ties are broken randomly and equiprobably. Losers pay nothing.

Because payoffs are not quasi-concave in own bids, one cannot appeal directly to Theorem 2 to establish the existence of an equilibrium in pure strategy bidding functions. On the other hand, it is not difficult to show that all mixed strategy equilibria are pure and non-decreasing. Hence, to obtain an existence result for pure strategies, it suffices to show that there is an equilibrium in mixed, or equivalently in distributional, strategies. (In this context, a distributional strategy for bidder *i* is a joint probability distribution over his values and bids with the property that the marginal density over his values is f_i ; see Milgrom and Weber, 1985.)

Because the set of distributional strategies for each bidder is a non-empty compact convex metric space and each bidder's payoff is linear in his own distributional strategy, Theorem 2 can be applied so long as a first-price auction game in distributional strategies is better-reply secure. Better-reply security can be shown to hold by using the facts that payoff discontinuities occur only when there are ties in bids and that bidders can always break a tie in their favour by increasing their bid slightly. Consequently, a Nash equilibrium in distributional strategies exists and, as mentioned above, this equilibrium is pure and non-decreasing.

3.5 Endogenous sharing rules

Discontinuities in payoffs sometimes arise endogenously. For example, consider a political game in which candidates first choose a policy from the interval [0,1] and each voter among a continuum then decides for whom to vote. Voters vote for the candidate whose policy they most prefer, and if there is more than one such candidate it is conventional to assume that voters randomize equiprobably over them. The behaviour of voters in the second stage can induce discontinuities in the payoffs of the candidates in the first stage since a candidate can discontinuously gain or lose a positive fraction of votes by choosing a policy that, instead of being identical to another candidate's policy, is just slightly different from it.

Simon and Zame (1990) suggest an elegant way to handle such discontinuities. In particular, for the political game example above, they would not insist that voters, when indifferent, randomize equiprobably. Indeed, applying subgame perfection to the two-stage game would permit voters to randomize in any manner whatsoever over those candidates whose policies they most prefer. With this in mind, if s is a joint pure strategy for the N candidates specifying a location for each, let us denote by $U(s)$ the resulting *set* of payoff vectors for the N candidates when all best replies of the voters are considered. If no voter is indifferent, then $U(s)$ contains a single payoff vector. On the other hand, if some voters are indifferent (as would be the case if two or more candidates chose the same location) and $U(s)$ is not a singleton, then distinct payoff vectors in $U(s)$ correspond to different ways the indifferent voters can randomize between the candidates among whom they are indifferent.

The significance of the correspondence $U(\cdot)$ is this. Suppose that we are able to select, for each s, a payoff vector $u(s) \in U(s)$ in such a way that some joint mixed strategy m^* for the N candidates is a Nash equilibrium of the induced policy-choice game between them when their vector payoff function is $u(\cdot)$. Then m^* together with the voter behaviour that is implicit in the definition of $u(s)$ for each s, constitutes a subgame perfect equilibrium of the original two-stage game. Thus, solving the original problem with potentially endogenous discontinuities boils down to obtaining an appropriate selection from $U(\cdot)$. Simon and Zame (1990) provide a general result concerning the existence of such selections, which they refer to as ‘endogenous sharing rules’. This method therefore provides an additional tool for obtaining equilibrium existence when discontinuities are present. Simon and Zame's main result is as follows.

Theorem 3: (Simon and Zame, 1990). Suppose that each S_i is a compact subset of a metric space and that $U: S \twoheadrightarrow \mathbb{R}^N$ is a bounded, upper hemi-continuous, non-empty-valued, convex-valued correspondence. Then for each player i, there is a measurable payoff function, $u_i: S \to \mathbb{R}$, *such that*

$(u_1(s), \ldots, u_N(s)) \in U(s)$ for every $s \in S$ and such that the game $(S_i, u_i)_{i=1}^{N}$ possesses at least one mixed strategy Nash equilibrium.

Remark: . Theorem 3 applies to the political game example above because for any policy choice s of the N candidates, the resulting set of payoff vectors $U(s)$ is convex, a fact that follows from the presence of a continuum of voters. It can also be shown that, as a correspondence, $U(\cdot)$ is upper hemi-continuous.

Remark: . In the context of Bayesian games, an even more subtle endogenous-sharing rule result can be found in Jackson et al. (2002). This result, too, can be very helpful in dealing with discontinuous games. Indeed, Jackson and Swinkels (2005) have shown how it can be used to obtain equilibrium existence results in a variety of auction settings, including double auctions.

See Also

- auctions (theory)
- epistemic game theory: incomplete information
- fixed point theorems
- mathematical methods in political economy
- spatial economics
- strategic and extensive form games

Bibliography

Baye, M., Tian, G. and Zhou, J. 1993. Characterizations of the existence of equilibria in games with discontinuous and non-quasiconcave payoffs. *Review of Economic Studies* 60, 935–48.

Bertrand, J. 1883. Théorie mathématique de la richesse sociale. *Journal des Savants* 67, 499–508.

Billingsley, P. 1968. *Convergence of Probability Measures*. New York: John Wiley and Sons.

Cournot, A. 1838. *Researches into the Mathematical Principles of the Theory of Wealth*, ed. N. Bacon. New York: Macmillan, 1897.

Dasgupta, P. and Maskin, E. 1986. The existence of equilibrium in discontinuous economic games, I: theory. *Review of Economic Studies* 53, 1–26.

Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences* 38, 386–93.

Fudenberg, D., Gilbert, R., Stiglitz, J. and Tirole, J. 1983. Preemption, leapfrogging, and competition in patent races. *European Economic Review* 22, 3–31.

Glicksberg, I. 1952. A further generalization of the Kakutani fixed point th. *Proceedings of the American Mathematical Society* 3, 170–4.

Hotelling, H. 1929. The stability of competition. *Economic Journal* 39, 41–57.

Jackson, M. and Swinkels, J. 2005. Existence of equilibrium in single and double private value auctions. *Econometrica* 73, 93–139.

Jackson, M., Simon, L., Swinkels, J. and Zame, W. 2002. Communication and equilibrium in discontinuous games of incomplete information. *Econometrica* 70, 1711–40.

Milgrom, P. and Roberts, J. 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica* 58, 1255–77.

Milgrom, P. and Weber, R. 1982. A theory of auctions and competitive bidding. *Econometrica* 50, 1089–122.

Milgrom, P. and Weber, R. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10, 619–32.

Nash, J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* 36, 48–9.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

Osborne, M. and Rubinstein, A. 1994. *A Course in Game Theory*. Cambridge, MA: MIT Press.

Reny, P. 1999. On the existence of pure and mixed strategy Nash equilibria in discontinuous games. *Econometrica* 67, 1029–56.

Robson, A. 1994. An ‘informationally robust’ equilibrium in two-person nonzero-sum games. *Games and Economic Behavior* 2, 233–45.

Schafer, W. and Sonnenschein, H. 1975. Equilibrium in Abstract economies without ordered preferences. *Journal of Mathematical Economics* 2, 345–8.

Simon, L. 1987. Games with discontinuous payoffs. *Review of Economic Studies* 54, 569–97.

Simon, L. and Zame, W. 1990. Discontinuous games and endogenous sharing rules. *Econometrica* 58, 861–72.

Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics* 8, 171–6.

Vives, X. 1990. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics* 19, 305–21.

von Neumann, J. 1928. Zur Theorie der Gesellshaftspiele. *Mathematische Annalen* 100, 295–320. Trans. S. Bargmann [On the theory of games of strategy] in *Contributions to the Theory of Games*, vol. 4, ed. R. Luce and A. Tucker. Princeton: Princeton University Press, 1959.

How to cite this article

Reny, Philip J. "non-cooperative games (equilibrium existence)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000136> doi:10.1057/9780230226203.1191

The New Palgrave Dictionary of Economics Online

non-expected utility theory

Mark J. Machina
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Beginning with the work of Allais and Edwards in the early 1950s and continuing through the present, psychologists and economists have uncovered a growing body of evidence that individuals do not necessarily conform to many of the key assumptions or predictions of the expected utility model of choice under uncertainty, and seem to depart from this model in systematic and predictable ways. This has led to the development of alternative models of preferences over objectively or subjectively uncertain prospects, which seek to accommodate these systematic departures from the expected utility model while retaining as much of its analytical power as possible.

Keywords

Allais Paradox; Allais, M.; ambiguity aversion; asset demand theory; Choquet expected utility model; common consequence effect; common ratio effect; comparative statics; Ellsberg Paradox; expected utility hypothesis; first-order stochastic dominance preference; Independence Axiom; insurance; non-expected utility theory; objective vs. subjective uncertainty; regret theory; risk; risk aversion; stochastic dominance; transitivity; uncertainty; von Neumann–Morgenstern utility function

Article

Although the expected utility model has long been the standard theory of individual choice under objective and subjective uncertainty, experimental work by both psychologists and economists has uncovered systematic departures from the expected utility hypothesis, which has led to the development of alternative models of preferences over uncertain prospects.

The expected utility model

In one of the simplest settings of choice under economic uncertainty, the objects of choice consist of finite-outcome *objective lotteries* of the form $\mathbf{P} = (x_1, p_1; \ldots; x_n, p_n)$, yielding a monetary payoff of x_i with probability p_i, where $p_1 + \ldots + p_n = 1$. In such a case, the expected utility model of risk preferences assumes (or posits axioms sufficient to imply) that the individual ranks these prospects on the basis of an *expected utility preference function* of the form

$$V_{EU}(\mathbf{P}) \equiv V_{EU}(x_1, p_1; \ldots; x_n, p_n) \equiv U(x_1) \cdot p_1 + \ldots + U(x_n) \cdot p_n$$

in the standard economic sense that the individual prefers lottery $\mathbf{P}^* = (x_1^*, p_1^*; \ldots; x_{n^*}^*, p_{n^*}^*)$ over lottery $\mathbf{P} = (x_1, p_1; \ldots; x_n, p_n)$ if and only if $V_{EU}(\mathbf{P}^*) > V_{EU}(\mathbf{P})$, and is indifferent between them if and only if $V_{EU}(\mathbf{P}^*) = V_{EU}(\mathbf{P})$. $U(\cdot)$ is termed the individual's *von Neumann–Morgenstern utility function* (von Neumann and Morgenstern, 1944; 1947;

1953) and its various mathematical properties serve to characterize various features of the individual's attitudes toward risk, for example:

- $V_{EU}(\cdot)$ exhibits *first-order stochastic dominance preference* (a preference for shifting probability from lower to higher outcome values) if and only if $U(x)$ is an increasing function of x.
- $V_{EU}(\cdot)$ exhibits *risk aversion* (an aversion to all mean-preserving increases in risk) if and only if $U(x)$ is a concave function of x.
- $V^*_{EU}(\cdot)$ is *at least as risk averse as* $V_{EU}(\cdot)$ (in several equivalent senses) if and only if its utility function $U^*(\cdot)$ is a concave transformation of $U(\cdot)$ (that is, if and only if $U^*(x) \equiv \rho(U(x))$ for some increasing concave function $\rho(\cdot)$).

As shown by Bernoulli (1738), Arrow (1965), Pratt (1964), Friedman and Savage (1948), Markowitz (1952) and others, this model admits of a tremendous flexibility in representing attitudes towards risk, and can be applied to many types of economic decisions and markets.
But in spite of its flexibility, the expected utility model has testable implications which hold regardless of the shape of the utility function $U(\cdot)$, since they follow from the *linearity in the probabilities* property of the preference function $V_{EU}(\cdot)$. These implications can be best expressed by the concept of an $\alpha : (1-\alpha)$ *probability mixture* of two lotteries $\mathbf{P} = (x_1, p_1; \ldots; x_n, p_n)$ and $\mathbf{P}^* = (x^*_1, p^*_1; \ldots; x^*_{n^*}, p^*_{n^*})$, which is defined as the single-stage lottery $\alpha \cdot \mathbf{P} + (1-\alpha) \cdot \mathbf{P}^* = (x_1, \alpha \cdot p_1; \ldots; x_n, \alpha \cdot p_n; x^*_1, (1-\alpha) \cdot p^*_1; \ldots; x^*_{n^*}, (1-\alpha) \cdot p^*_{n^*})$. The mixture $\alpha \cdot \mathbf{P} + (1-\alpha) \cdot \mathbf{P}^*$ can be thought of as a coin flip yielding lotteries $\mathbf{P}$ and $\mathbf{P}^*$ with probabilities $\alpha : (1-\alpha)$, where the uncertainty in the coin and in the subsequent lottery is resolved simultaneously. Linearity in the probabilities is equivalent to the following property, which serves as the key foundational axiom of the expected utility model (Marschak, 1950):

Independence Axiom: If lottery $\mathbf{P}^*$ is preferred (indifferent) to lottery $\mathbf{P}$, then the probability mixture $\alpha \cdot \mathbf{P}^* + (1-\alpha) \cdot \mathbf{P}^{**}$ is preferred (indifferent) to $\alpha \cdot \mathbf{P} + (1-\alpha) \cdot \mathbf{P}^{**}$ for every lottery $\mathbf{P}^{**}$ and every mixture probability $\alpha \in (0, 1]$.
This axiom can be interpreted as saying 'given an $\alpha : (1-\alpha)$ coin, the individual's preferences for receiving $\mathbf{P}^*$ versus $\mathbf{P}$ in the event of a head should not depend upon the prize $\mathbf{P}^{**}$ that would be received in the event of a tail, nor upon the probability α of landing heads (so long as this probability is positive)'. The strong normative appeal of this axiom has contributed to the widespread adoption of the expected utility model.
The property of linearity in the probabilities, as well as the senses in which it has been found to be empirically violated, can be illustrated in the special case of preferences over all lotteries $\mathbf{P} = (\bar{x}_1, p_1; \bar{x}_2, p_2; \bar{x}_3, p_3)$ over a fixed set of outcome values $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$. Since we must have $p_2 = 1 - p_1 - p_3$, each such lottery can be completely summarized by its pair of probabilities (p_1, p_3), as plotted in the 'probability triangle' of Figure 1. Since upward movements in the diagram (increasing p_3 for fixed p_1) represent shifting probability from outcome $\bar{x}_2$ up to $\bar{x}_3$, and leftward movements represent shifting probability from $\bar{x}_1$ up to $\bar{x}_2$, such movements constitute first-order stochastically dominating shifts and will thus always be preferred. Expected utility indifference curves (loci of constant expected utility) are given by the formula

$$U(\bar{x}_1) \cdot p_1 + U(\bar{x}_2) \cdot [1 - p_1 - p_3] + U(\bar{x}_3) \cdot p_3 = \text{constant}$$

and are thus seen to be parallel straight lines of slope $[U(\bar{x}_2) - U(\bar{x}_1)] / [U(\bar{x}_3) - U(\bar{x}_2)]$, as indicated by the solid lines in the figure. The dotted lines in Figure 1 are loci of constant *expected value*, given by the formula $\bar{x}_1 \cdot p_1 + \bar{x}_2 \cdot [1 - p_1 - p_3] + \bar{x}_3 \cdot p_3 = \text{constant}$, with slope $[\bar{x}_2 - \bar{x}_1] / [\bar{x}_3 - \bar{x}_2]$. Since north-east movements along the constant expected value lines shift probability from $\bar{x}_2$ down to $\bar{x}_1$ and up to $\bar{x}_3$ in a manner that preserves the mean of the distribution, they represent simple increases in risk (Rothschild and Stiglitz, 1970; 1971). When $U(\cdot)$ is concave (that is, risk averse), its indifference curves will have a steeper slope than these constant expected value lines, and such increases in risk move the individual from more to less preferred indifference curves, as illustrated in the figure. It is straightforward to show that the indifference curves of any expected utility maximizer with a more risk-averse (that is, more concave) utility function $U^*(\cdot)$ will be steeper than those generated by $U(\cdot)$.
Figure 1
Expected utility indifference curves in the probability triangle

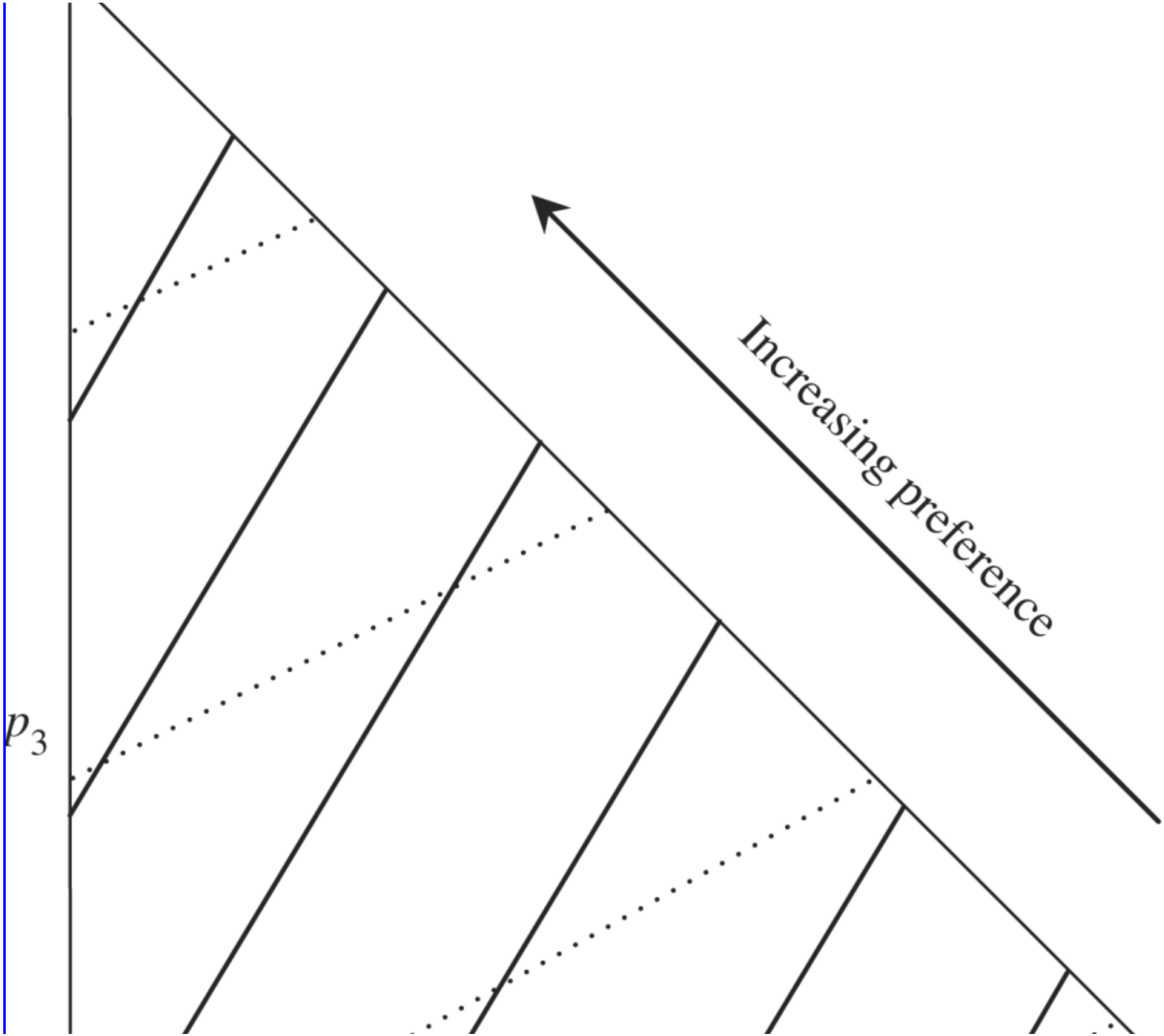
Increasing preference
p_3

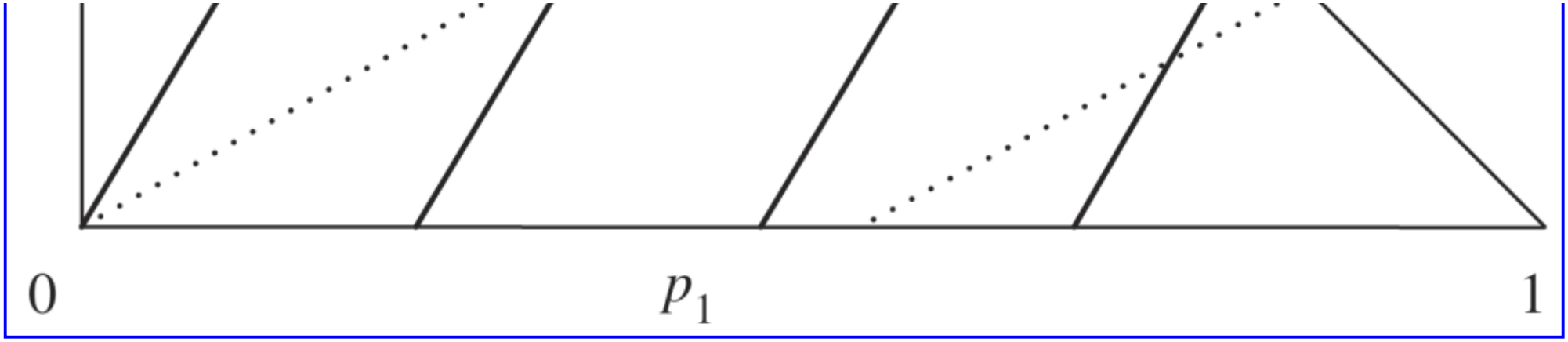

Systematic violations of the expected utility hypothesis

In spite of its normative appeal, researchers have uncovered several types of widespread systematic violations of the expected utility model and its underlying assumptions. These can be categorized into (*a*) violations of the Independence Axiom (such as the common consequence and common ratio effects), (*b*) violations of the hypothesis of probabilistic beliefs (such as the Ellsberg Paradox) and (*c*) violations of the model's underlying assumptions of descriptive and procedural invariance (such as reference-point and response-mode effects).

Violations of the Independence Axiom

The best-known violation of the Independence Axiom is the so-called *Allais Paradox*, in which individuals are asked to rank the lotteries in each of the following pairs, where $\$1\text{M} = \$1,000,000$:

$$a_1: \{1.00 \text{ chance of } \$1\text{M} \quad \text{versus} \quad a_2: \begin{cases} .10 \text{ chance of } \$5\text{M} \\ .89 \text{ chance of } \$1\text{M} \\ .01 \text{ chance of } \$0 \end{cases}$$

$$a_3: \begin{cases} .10 \text{ chance of } \$5\text{M} \\ .90 \text{ chance of } \$0 \end{cases} \quad \text{versus} \quad a_4: \begin{cases} .11 \text{ chance of } \$1\text{M} \\ .89 \text{ chance of } \$0 \end{cases}$$

Researchers such as Allais (1953), Morrison (1967), Raiffa (1968), Slovic and Tversky (1974) and others have found that the modal if not majority preference of subjects is for a_1 over a_2 in the first pair of choices and for a_3 over a_4 in the second pair. However, such preferences violate expected utility, since the first ranking implies the inequality $U(\$1\text{M}) > .10 \cdot U(\$5\text{M}) + .89 \cdot U(\$1\text{M}) + .01 \cdot U(\$0)$ whereas the second implies the inconsistent inequality $.10 \cdot U(\$5\text{M}) + .90 \cdot U(\$0) > .11 \cdot U(\$1\text{M}) + .89 \cdot U(\$0)$. By setting $\bar{x}_1 = \$0$, $\bar{x}_2 = \$1\text{M}$ and $\bar{x}_3 = \$5\text{M}$, the lotteries a_1, a_2, a_3 and a_4 are seen to form a parallelogram when plotted in the probability triangle (Figure 2), which explains why the parallel straight line indifference curves of an expected utility maximizer must either prefer a_1 and a_4 (as illustrated for the relatively steep indifference curves of the figure) or else prefer a_2 and a_3 (for relatively flat indifference curves). Figure 3 illustrates *non-expected utility indifference curves* which *fan out*, and are seen to exhibit the typical Allais Paradox rankings of a_1 over a_2 and a_3 over a_4.

Figure 2
Expected utility indifference curves and the Allais Paradox choices

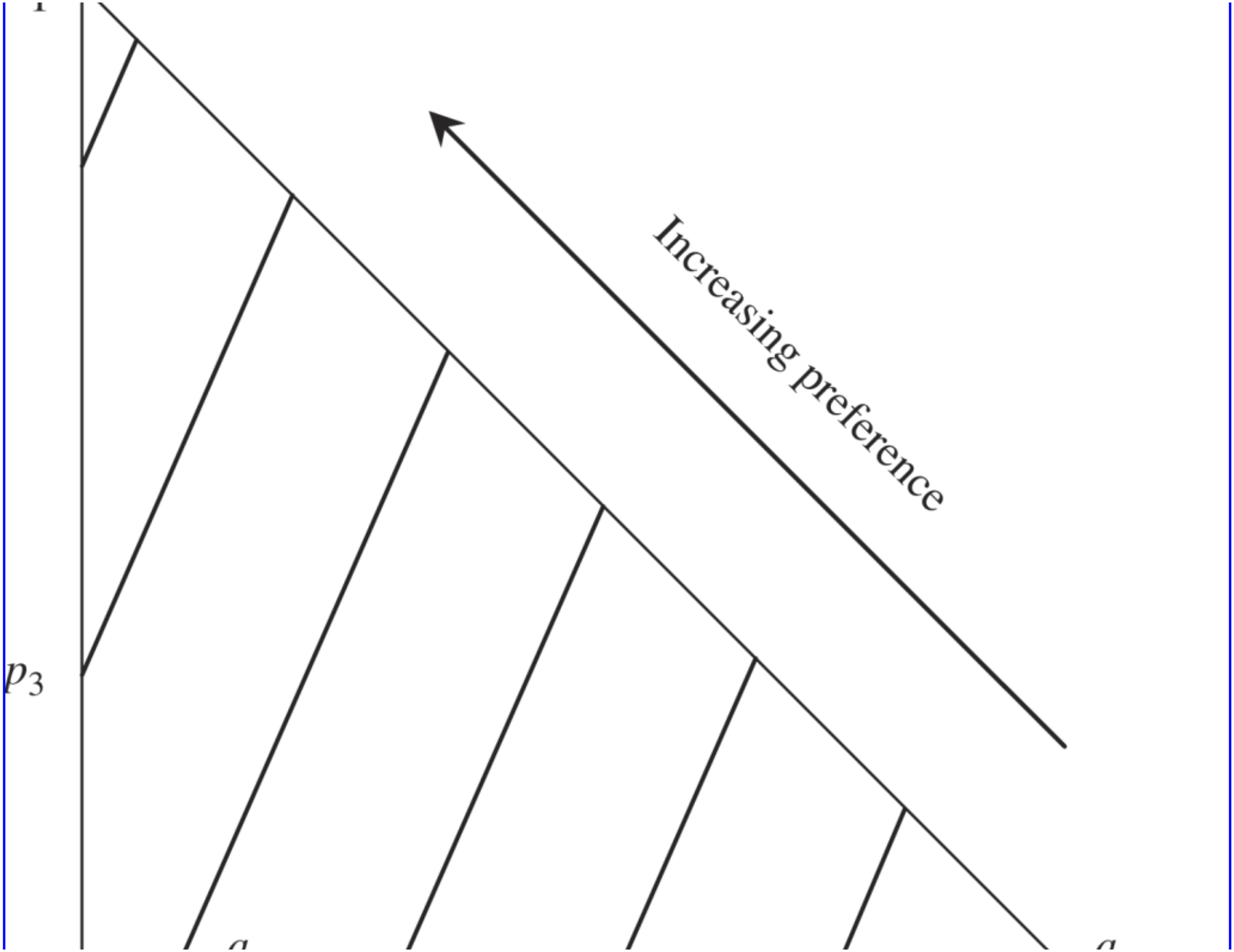
Increasing preference
p_3

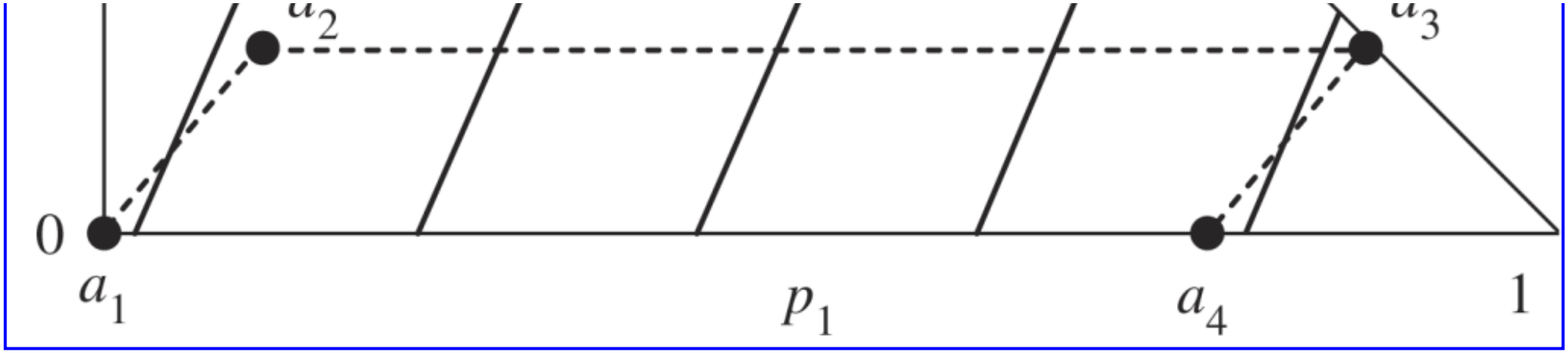

Figure 3
Allais Paradox choices and indifference curves which ‘fan out’

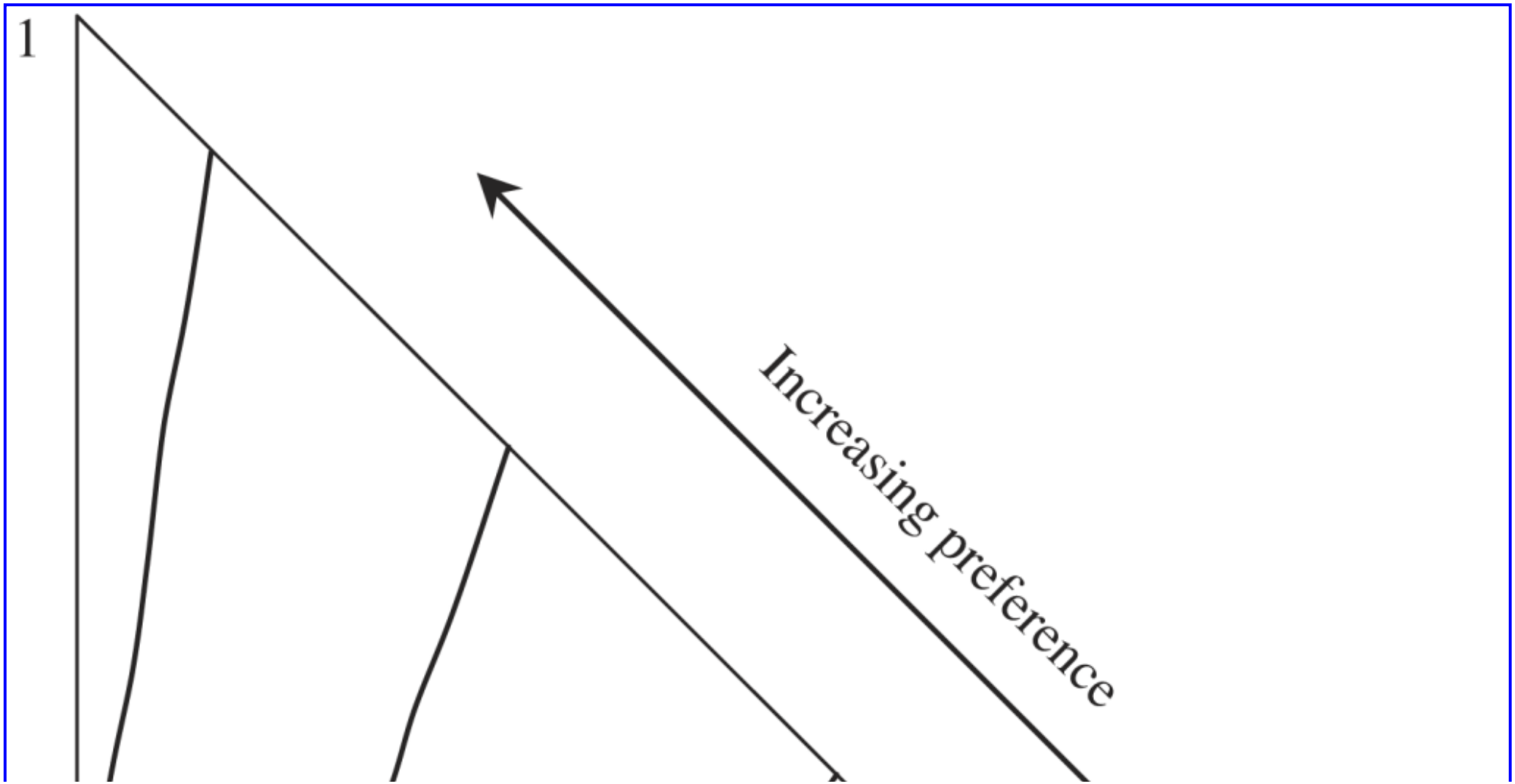

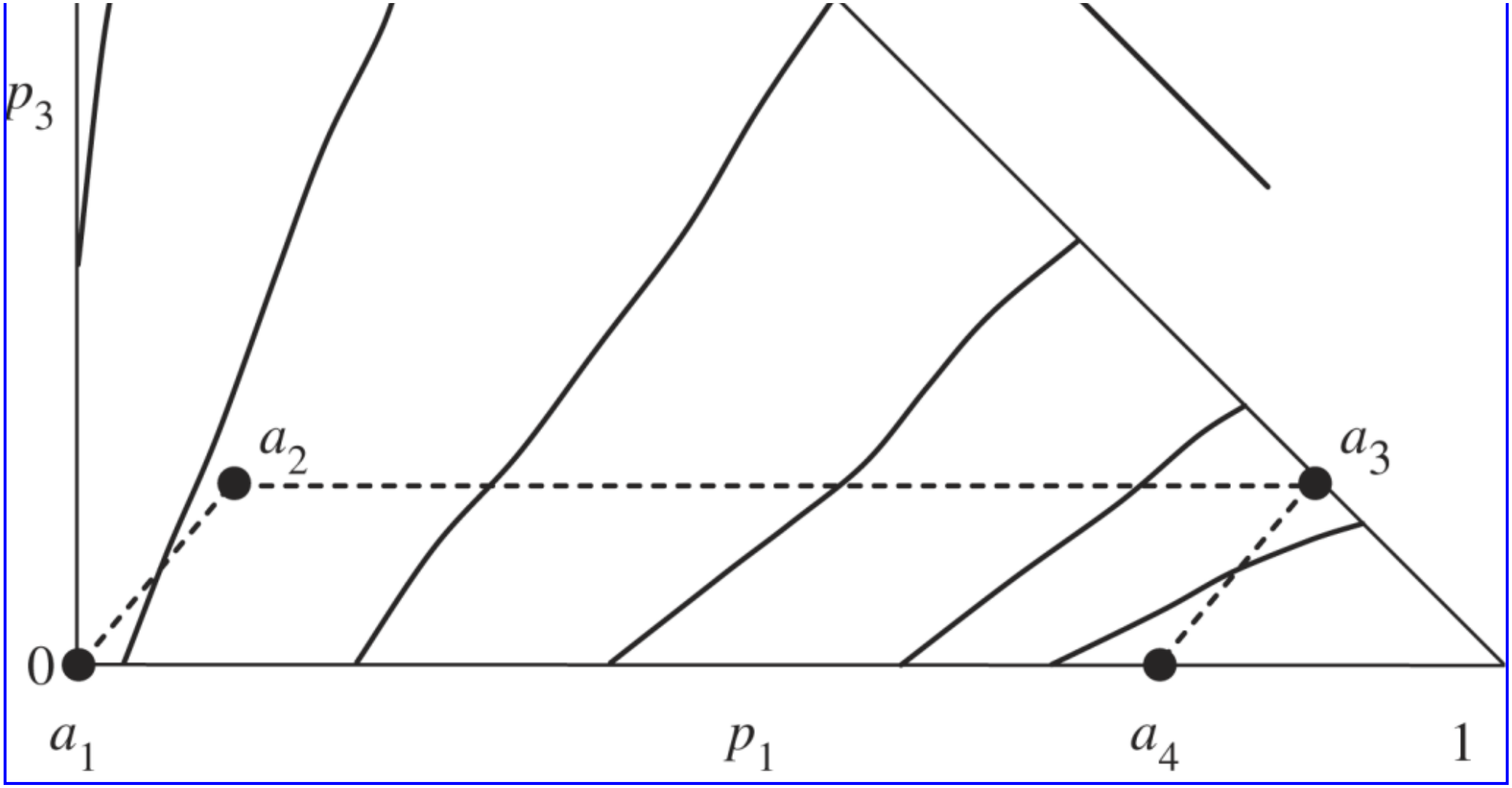

Although the Allais Paradox was originally dismissed as an isolated example, subsequent experimental work by psychologists, economists and others have uncovered a similar pattern of violations over a range of probability and payoff values, and the Allais Paradox is now seen to be a special case of a widely observed phenomenon known as the *common consequence effect.* This effect involves pairs of prospects (probability mixtures) of the form:

$$b_1: \begin{cases} \alpha & \text{chance of } x \\ 1-\alpha & \text{chance of } \mathbf{P}^{**} \end{cases} \quad \text{versus} \quad b_2: \begin{cases} \alpha & \text{chance of } \mathbf{P} \\ 1-\alpha & \text{chance of } \mathbf{P}^{**} \end{cases}$$

$$b_3: \begin{cases} \alpha & \text{chance of } x \\ 1-\alpha & \text{chance of } \mathbf{P}^{*} \end{cases} \quad \text{versus} \quad b_4: \begin{cases} \alpha & \text{chance of } \mathbf{P} \\ 1-\alpha & \text{chance of } \mathbf{P}^{*} \end{cases}$$

where the lottery **P** involves outcomes both greater and less than the amount x, and $\mathbf{P}^{**}$ first order stochastically dominates $\mathbf{P}^{*}$ (in Allais's example, $x = \$1\text{M}$,

$\mathbf{P} = (\$5M, 10/11; \$0, 1/11)$, $\mathbf{P}^* = (\$0, 1)$, $\mathbf{P}^{**} = (\$1M, 1)$ and $\alpha = .11$). Although the Independence Axiom clearly implies choices of either b_1 and b_3 (if x is preferred to **P**) or else b_2 and b_4 (if **P** is preferred to x), researchers have found a tendency for subjects to choose b_1 in the first pair and b_4 in the second. When the distributions **P**, $\mathbf{P}^*$ and $\mathbf{P}^{**}$ are each over a common outcome set $\{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$ with $\bar{x}_2 = x$, the prospects $\{b_1, b_2, b_3, b_4\}$ again form a parallelogram in the (p_1, p_3) triangle, and a choice of b_1 and b_4 again implies indifference curves which fan out.

The intuition behind this phenomenon can be described in terms of the above ‘coin-flip’ scenario. According to the Independence Axiom, one's preferences over what would occur in the event of a head ought not depend upon what would occur in the event of a tail. However, they *may well* depend upon what would otherwise happen (as Bell, 1985, p. 1, notes, ‘winning the top prize of $10,000 in a lottery may leave one much happier than receiving $10,000 as the lowest prize in a lottery’). The common consequence effect states that the *better off* individuals would be in the event of a tail (in the sense of stochastic dominance), the *more risk averse* their preferences over what they would receive in the event of a head. That is, if the distribution $\mathbf{P}^{**}$ in the pair $\{b_1, b_2\}$ involves very high outcomes, one may prefer not to bear further risk in the unlucky event that one doesn't receive it, and hence opt for the sure outcome x over the risky distribution **P** (that is, choose b_1 over b_2). But, if $\mathbf{P}^*$ in $\{b_3, b_4\}$ involves very low outcomes, one might be more willing to bear risk in the lucky event that one doesn't receive it, and prefer going for the lottery **P** rather than the sure outcome x (choose b_4 over b_3).

A second type of systematic violation of linearity in the probabilities, also noted by Allais and subsequently termed the *common ratio effect*, involves prospects of the form:

$$c_1: \begin{cases} p & \text{chance of } \$X \\ 1-p & \text{chance of } \$0 \end{cases} \quad \text{versus} \quad c_2: \begin{cases} q & \text{chance of } \$Y \\ 1-q & \text{chance of } \$0 \end{cases}$$

$$c_3: \begin{cases} \alpha \cdot p & \text{chance of } \$X \\ 1-\alpha \cdot p & \text{chance of } \$0 \end{cases} \quad \text{versus} \quad c_4: \begin{cases} \alpha \cdot q & \text{chance of } \$Y \\ 1-\alpha \cdot q & \text{chance of } \$0 \end{cases}$$

where $p > q$, $0 < X < Y$ and $\alpha \in (0, 1)$. (The term ‘common ratio effect’ comes from the common value of prob($\$X$)/prob($\$Y$) in the upper and lower pairs.) Setting $\{\bar{x}_1, \bar{x}_2, \bar{x}_3\} = \{\$0, \$X, \$Y\}$ and plotting these prospects in the probability triangle as in Figure 4, the line segments $\overline{c_1c_2}$ and $\overline{c_3c_4}$ are seen to be parallel, so that the expected utility model again predicts choices of c_1 and c_3 (if the indifference curves are relatively steep) or else c_2 and c_4 (if they are flat). However, experimental studies by MacCrimmon (1968), Tversky (1975), MacCrimmon and Larsson (1979), Kahneman and Tversky (1979), Hagen (1979), Chew and Waller (1986) and others have found a systematic tendency for choices to depart from these predictions in the direction of preferring c_1 over c_2 and c_4 over c_3, which again suggests that indifference curves fan out, as in the figure. For example, Kahneman and Tversky (1979) found that, while 86 per cent of their subjects preferred a .90 chance of winning $3,000 to a .45 chance of $6,000, 73 per cent preferred a .001 chance of $6,000 to a .002 chance of $3,000. Kahneman and Tversky (1979) observed that, when the positive outcomes $3000 and $6000 in the above gambles are replaced by *losses* of these magnitudes, to obtain the lotteries c_1', c_2', c_3' and c_4', preferences typically ‘reflect,’ to prefer c_2' over c_1' and c_3' over c_4'. Setting $\bar{x}_1 = -\$6000$, $\bar{x}_2 = -\$3000$ and $\bar{x}_3 = -\$0$ (to preserve the ordering $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$) and plotting as in Figure 5, such preferences again suggest that indifference curves in the probability triangle fan out. Battalio, Kagel and MacDonald (1985) found that laboratory rats choosing among gambles involving substantial variations in their daily food intake also exhibited this pattern of choices.

Figure 4

Common ratio effect and fanning out indifference curves

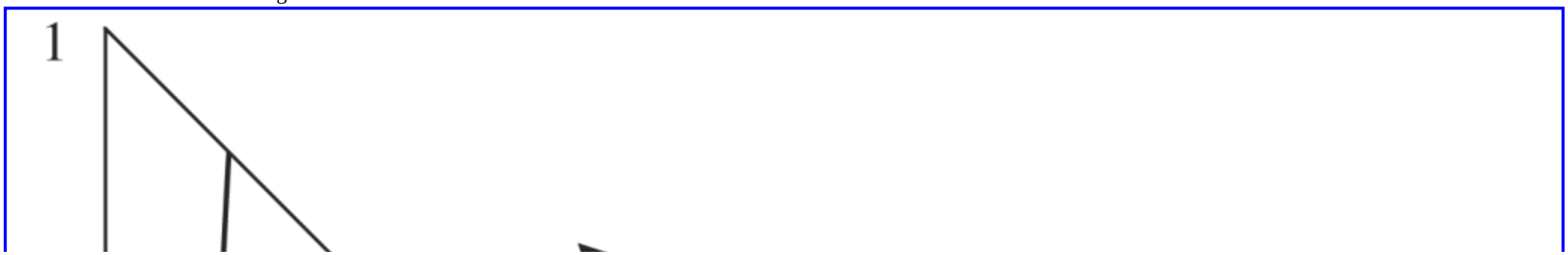

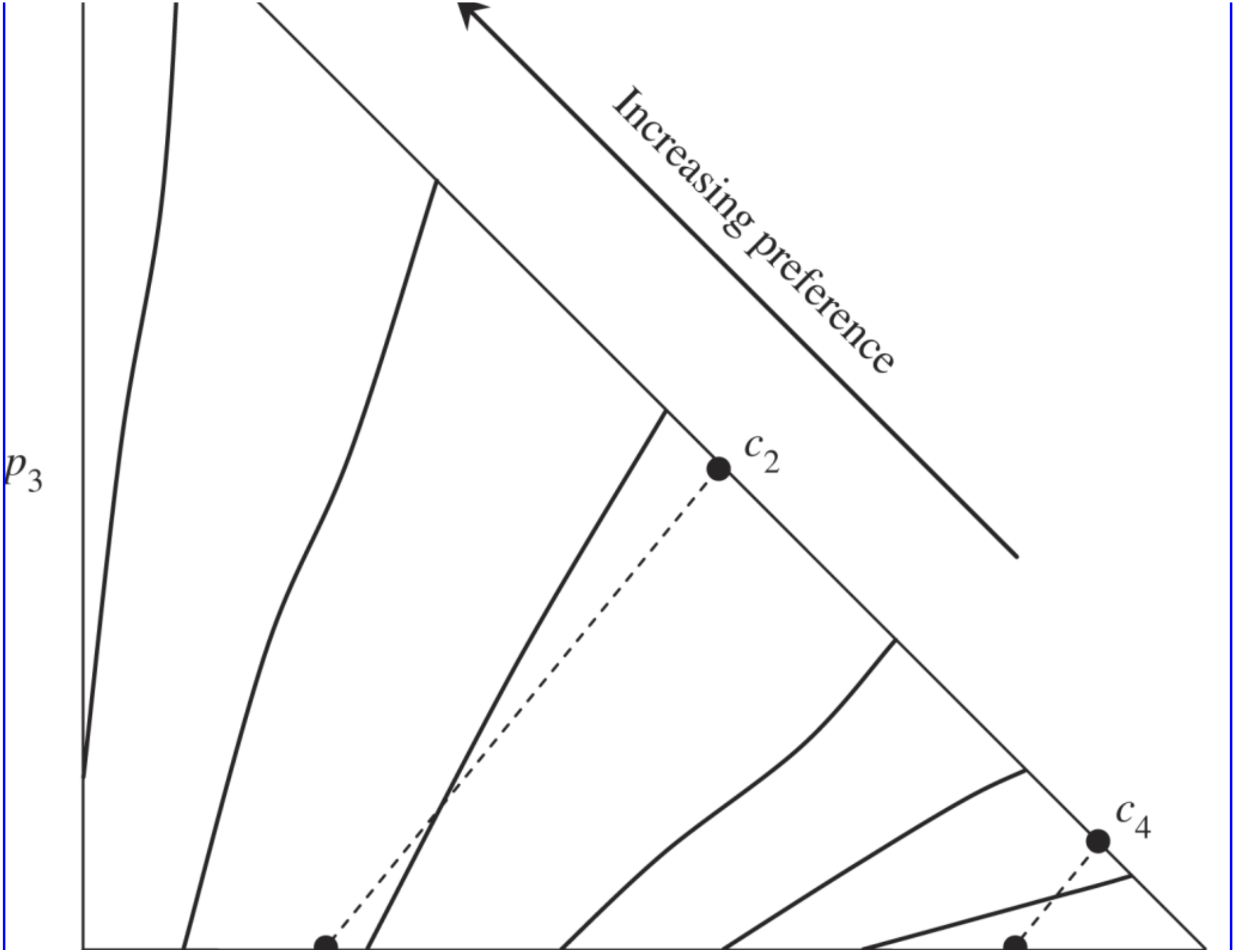
Increasing preference
p_3
c_2
c_4

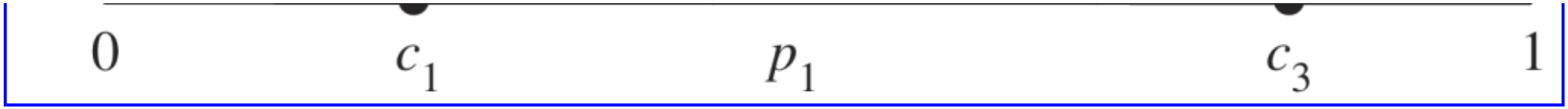

Figure 5
Common ratio effect for losses and fanning out indifference curves

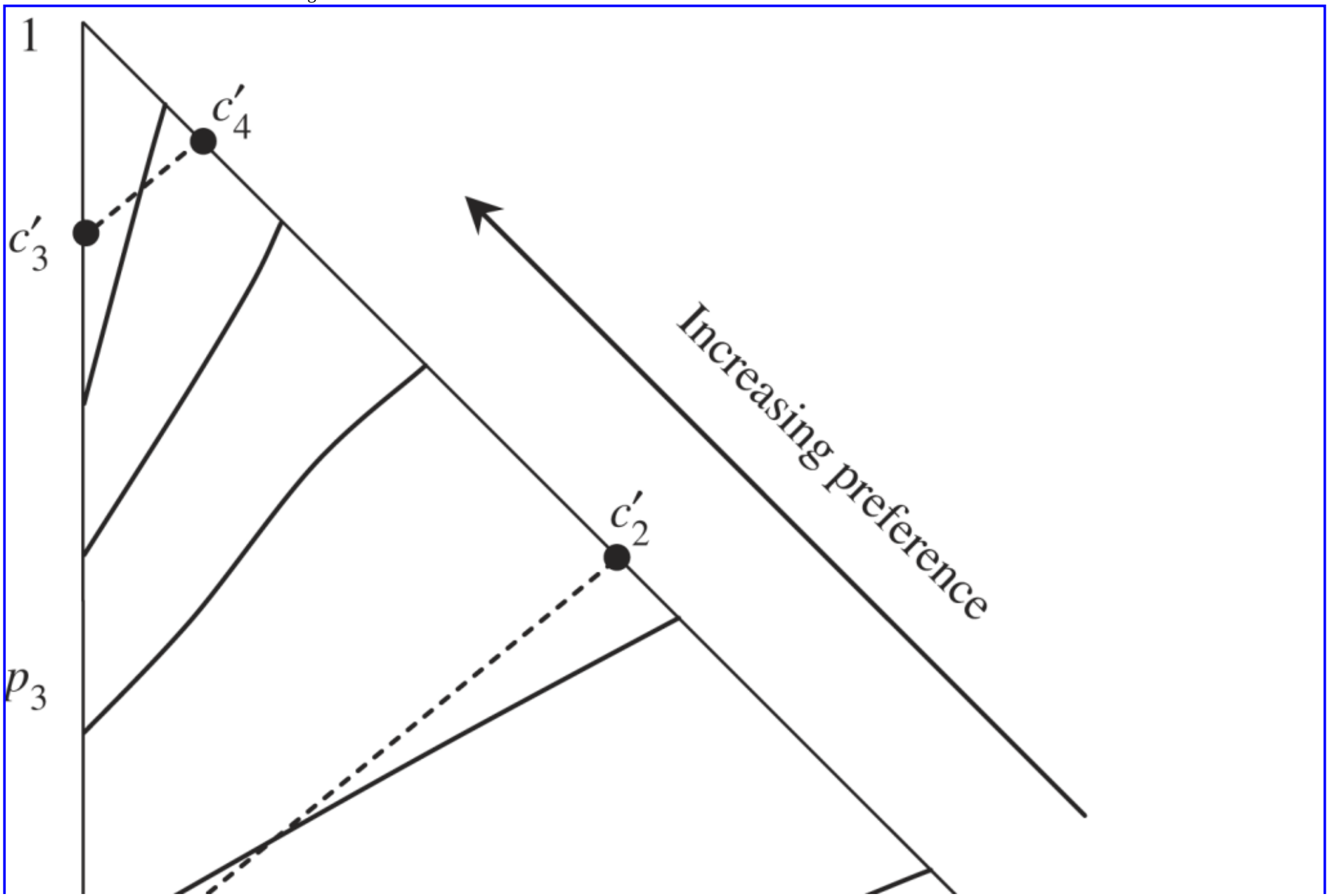

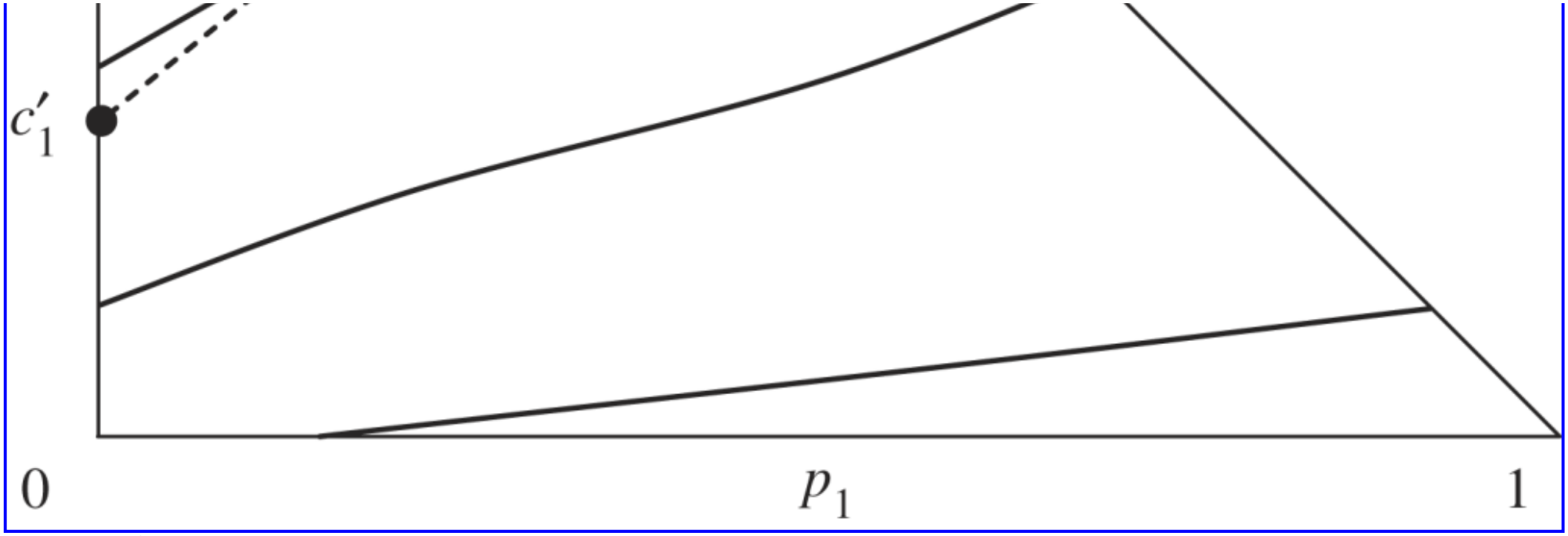

One criticism of this evidence has been that individuals whose initial choices violated the Independence Axiom in the above manners would typically 'correct' themselves once the nature of their violations was revealed by an application of the above type of coin-flip argument. Thus, while even Leonard Savage chose a_1 and a_3 when first presented with such choices by Allais, he concluded upon reflection that these preferences were in error (Savage, 1954, pp. 101–3). Although Moskowitz found that allowing subjects to discuss opposing written arguments led to a decrease in the proportion of violations, 73 per cent of the initial fanning-out type choices remained unchanged after the discussions (1974, pp. 232–7, Table 6). When written arguments were presented but no discussion was allowed, there was a 93 per cent persistency rate of such choices (1974, p. 234, Tables 4, 6). In experiments where subjects who responded to Allais-type problems were then presented with written arguments both for *and against* the expected utility position, neither MacCrimmon (1968), Moskowitz (1974) nor Slovic and Tversky (1974) found predominant net swings toward the expected utility choices.

Further descriptions of these and other violations of the Independence Axiom can be found in Camerer (1989), Machina (1983; 1987), Starmer (2000), Sugden (1986) and Weber and Camerer (1987).

Non-existence of probabilistic beliefs

Although the expected utility model was first formulated in terms of preferences over *objective lotteries* $\mathbf{P} = (x_1, p_1; \ldots; x_n, p_n)$ with pre-specified probabilities, it has also been applied to preferences over *subjective acts* $f(\cdot) = [x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n]$, where the uncertainty is represented by a set $\{E_1, \ldots, E_n\}$ of mutually exclusive and exhaustive *events* (such as the alternative outcomes of a horse race) (Savage, 1954). As long as an individual possesses well-defined *subjective probabilities* $\mu(E_1), \ldots, \mu(E_n)$ over these events, their *subjective expected utility* preference function takes the form

$$W_{SEU}(f(\cdot)) \equiv W_{SEU}(x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n) \equiv U(x_1) \cdot \mu(E_1) + \ldots + U(x_n) \cdot \mu(E_n) .$$

However, researchers have found that individuals may not possess such well-defined subjective probabilities, in even the simplest of cases. The best-known example of this is the *Ellsberg Paradox* (Ellsberg, 1961), in which the individual must draw a ball from an urn that contains 30 red balls, and 60 black or yellow balls in an unknown proportion, and is offered the following bets based on the colour of the drawn ball:

	30 balls	60 balls	
	red	black	yellow
$f_1(\cdot)$	\$100	\$0	\$0
$f_2(\cdot)$	\$0	\$100	\$0
$f_3(\cdot)$	\$100	\$0	\$100
$f_4(\cdot)$	\$0	\$100	\$100

Most individuals exhibit a preference for $f_1(\cdot)$ over $f_2(\cdot)$ and $f_4(\cdot)$ over $f_3(\cdot)$. When asked, they explain that the chance of winning under $f_2(\cdot)$ could be anywhere from 0 to 2/3 whereas under $f_1(\cdot)$ it is known to be exactly 1/3, and they prefer the bet that offers the known probability. Similarly, the chance of winning under $f_3(\cdot)$ could be anywhere from 1/3 to 1 whereas under $f_4(\cdot)$ it is known to be exactly 2/3, so the latter is preferred. However, such preferences are inconsistent with any assignment of subjective probabilities μ (red), μ (black), μ (yellow) to the three events. If the individual were to be choosing on the basis of such probabilistic beliefs, the choice of $f_1(\cdot)$ over $f_2(\cdot)$ would 'reveal' that μ (red)> μ (black), but the choice of $f_4(\cdot)$ over $f_3(\cdot)$ would reveal that μ (red)< μ (black). A preference for gambles based on probabilistic partitions such as {red, black $\cup$ yellow} over gambles based on subjective partitions such as {black, red $\cup$ yellow} is termed *ambiguity aversion*.

In an even more basic example, Ellsberg presented subjects with a pair of urns, the first containing 50 red balls and 50 black balls, and the second with 100 red and black balls in an unknown proportion. When asked, a majority of subjects strictly preferred to stake a prize on drawing red from the first urn over drawing red from the second urn, and strictly preferred staking the prize on drawing black from the first urn over drawing black from the second. It is clear that there can exist no subjective probabilities p:(1–p) of red:black in the

second urn, including 1/2:1/2, which can simultaneously generate both of these strict preferences. Similar behaviour in this and related problems has been observed by Raiffa (1961), Becker and Brownson (1964), MacCrimmon (1965), Slovic and Tversky (1974) and MacCrimmon and Larsson (1979).

Violations of descriptive and procedural invariance

Researchers have also uncovered several systematic violations of the standard economic assumptions of stability of preferences and invariance with respect to problem description in choices over risky prospects. In particular, psychologists have found that alternative means of representing or *framing* probabilistically equivalent choice problems lead to systematic differences in choice. Early examples of this were reported by Slovic (1969), who found that offering a gain or loss contingent on the joint occurrence of four independent events with probability p elicited different responses than offering it on the occurrence of a single event with probability p^4 (all probabilities were stated explicitly). In comparison with the single-event case, making a gain contingent on the joint occurrence of events was found to make it more attractive, and making a loss contingent on the joint occurrence of events made it more unattractive.

One class of framing effects exploits the phenomenon of a *reference point*. According to economic theory, the variable which enters an individual's von Neumann–Morgenstern utility function should be total (that is, final) wealth, and gambles phrased in terms of gains and losses should be combined with current wealth and re-expressed as distributions over final wealth levels before being evaluated. However, risk attitudes towards gains and losses tend to be more stable than can be explained by a fixed utility function over final wealth, and utility functions might be best defined in terms of changes from the *reference point* of current wealth. In his discussion of this phenomenon, Markowitz (1952, p. 155) suggested that certain circumstances may cause the individual's reference point to temporarily deviate from current wealth. If these circumstances include the manner in which a problem is verbally described, then differing risk attitudes towards gains and losses from the reference point can lead to different choices, depending upon the exact description of an otherwise identical problem. A simple example of this, from Kahneman and Tversky (1979, p. 273), involves the following two choices:

> In addition to whatever you own, you have been given 1,000 (Israeli pounds). You are now asked to choose between a 1/2:1/2 chance of a gain of 1,000 or 0 or a sure chance of a gain of 500.

and

> In addition to whatever you own, you have been given 2,000. You are now asked to choose between a 1/2:1/2 chance of a loss of 1,000 or 0 or a sure loss of 500.

These two problems involve identical distributions over final wealth. But, when put to two different groups of subjects, 84 per cent chose the sure gain in the first problem but 69 per cent chose the 1/2:1/2 gamble in the second.

In another class of examples, not based on reference point effects, Moskowitz (1974), Keller (1985) and Carlin (1990) found that the proportion of subjects choosing in conformity with the Independence Axiom in examples like the Allais Paradox was significantly affected by whether the problems were described in the standard matrix form, decision tree form, roulette wheels, or as minimally structured written statements. Interestingly, the form judged the ‘clearest representation’ by the majority of Moskowitz's subjects (the tree form) led to the lowest degree of consistency with the Independence Axiom, the highest proportion of Allais-type (fanning out) choices, and the highest persistency rate of these choices Moskowitz (1974, pp. 234, 237–8).

In other studies, Schoemaker and Kunreuther (1979), Hershey and Schoemaker (1980), Kahneman and Tversky (1982; 1984), and Slovic, Fischhoff and Lichtenstein (1977) found that subjects' choices in otherwise identical problems depended upon whether they were phrased as decisions whether or not to gamble as opposed to whether or not to insure, whether statistical information for different therapies was presented in terms of cumulative survival probabilities or cumulative mortality probabilities, and so on (see the references in Tversky and Kahneman, 1981).

Whereas framing effects involve alternative *descriptions* of an otherwise identical choice problem, alternative *response formats* have also been found to lead to different choices, leading to what have been termed *response-mode effects*. For example, under expected utility, an individual's von Neumann–Morgenstern utility function can be assessed or elicited in a number of different manners, which typically involve a sequence of pre-specified lotteries $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$, …, and ask for (a) the individual's certainty equivalent $CE(\mathbf{P}_i)$ of each lottery $\mathbf{P}_i$, (b) the *gain equivalent* G_i that would make the gamble (G_i,1/2; \$0,1/2) indifferent to $\mathbf{P}_i$, or (c) the *probability equivalent* $\wp_i$ that would make the gamble (\$1000, $\wp_i$; \$0,1−$\wp_i$) indifferent to $\mathbf{P}_i$. Although such procedures should generate equivalent assessed utility functions, they have been found to yield systematically different ones (for example, Hershey, Kunreuther and Schoemaker, 1982; Hershey and Schoemaker, 1985).

In a separate finding now known as the *preference reversal phenomenon*, subjects were first presented with a number of pairs of lotteries and asked to make one choice out of each

pair. Each pair of lotteries took the following form:

$$p\text{-bet}: \begin{cases} p & \text{chance of } \$X \\ 1-p & \text{chance of } \$0 \end{cases} \quad \text{versus} \quad \$\text{-bet}: \begin{cases} q & \text{chance of } \$Y \\ 1-q & \text{chance of } \$0 \end{cases}$$

where $0 < X < Y$ and $p > q$. The terms '*p*-bet' and '$-bet' derive from the greater probability of winning in the first bet, and greater possible gain in the second bet. Subjects were next asked for their certainty equivalents of each of these bets, via a number of standard elicitation techniques. Standard theory predicts that, for each such pair, the prospect selected in the direct choice problem would also be assigned the higher certainty equivalent. However, subjects exhibit a systematic departure from this prediction in the direction of choosing the *p*-bet in a direct choice, but assigning a higher certainty equivalent to the $-bet (Lichtenstein and Slovic, 1971). Although this finding initially generated widespread scepticism, it has been replicated by both psychologists and economists in a variety of settings involving real-money gambles, patrons in a Las Vegas casino, group decisions and experimental market trading. By expressing the implied preferences as '$-bet ~ *CE*($-bet) • *CE*(*p*-bet) ~ *p*-bet • $-bet', some economists have categorized this phenomenon as a violation of transitivity and tried to model it as such (see the 'regret theory' model below). However, most psychologists and economists now view it as a response-mode effect: specifically, that the psychological processes of valuation (which generates certainty equivalents) and direct choice are differentially influenced by the probabilities and payoffs involved in a lottery, and that under certain conditions this can lead to choices and valuations which 'reveal' opposite preference rankings over a pair of gambles.

Non-expected utility models of risk preferences

Non-expected utility functional forms

Researchers have responded to departures from linearity in the probabilities in two manners. The first consists of replacing the expected utility form $V_{EU}(\mathbf{P}) = U(x_1) \cdot p_1 + \ldots + U(x_n) \cdot p_n$ by some more general form for the preference function $V(\mathbf{P}) = V(x_1, p_1; \ldots; x_n, p_n)$. Several such forms have been proposed (for the Rank Dependent, Dual and Ordinal Independence forms, the payoffs must be labelled so that $x_1 \leq \ldots \leq x_n$, and $G(\cdot)$ must satisfy $G(0) = 0$ and $G(1) = 1$):

Prospect theory	$\Sigma_{i=1}^{n} v(x_i) \cdot \pi(p_i)$	Edwards (1955; 1962), Kahneman–Tversky (1979)
Subjectively weighted utility	$\Sigma_{i=1}^{n} v(x_i) \cdot \pi(p_i) \,/\, \Sigma_{i=1}^{n} \pi(p_i)$	Karmarkar (1978; 1979)
Rank-dependent expected utility	$\Sigma_{i=1}^{n} v(x_i) \cdot \left[G\left(\Sigma_{j=1}^{i} p_j\right) - G\left(\Sigma_{j=1}^{i-1} p_j\right)\right]$	Quiggin (1982)
Dual expected utility	$\Sigma_{i=1}^{n} x_i \cdot \left[G\left(\Sigma_{j=1}^{i} p_j\right) - G\left(\Sigma_{j=1}^{i-1} p_j\right)\right]$	Yaari (1987)
Ordinal independence	$\Sigma_{i=1}^{n} h\left(x_i, \Sigma_{j=1}^{i} p_j\right) \cdot \left[G\left(\Sigma_{j=1}^{i} p_j\right) - G\left(\Sigma_{j=1}^{i-1} p_j\right)\right]$	Segal (1984), Green–Jullien (1988)
Moments of utility	$M\left(\Sigma_{i=1}^{n} v(x_i) \cdot p_i, \Sigma_{i=1}^{n} v(x_i)^2 \cdot p_i, \ldots\right)$	Múnera–de Neufville (1983), Hagen (1979)
Weighted utility	$\Sigma_{i=1}^{n} v(x_i) \cdot p_i \,/\, \Sigma_{i=1}^{n} \tau(x_i) \cdot p_i$	Chew (1983)
Optimism–pessimism	$\Sigma_{i=1}^{n} v(x_i) \cdot g(p_i, x_1, \ldots, x_n)$	Hey (1984)
Quadratic in the probabilities	$\Sigma_{i=1}^{n} \Sigma_{j=1}^{n} K(x_i, x_j) \cdot p_i \cdot p_j$	Chew, Epstein and Segal (1991)
Regret theory	$\Sigma_{i=1}^{n} \Sigma_{j=1}^{n^*} R(x_i, x_j^*) \cdot p_i \cdot p_j^*$	Loomes–Sugden (1982)

Most of these forms have been formally axiomatized, and, under the appropriate monotonicity and/or curvature assumptions on their constituent functions $\upsilon(\cdot)$, $G(\cdot)$, and so on, most are capable of exhibiting first-order stochastic dominance preference, risk aversion, and the above types of systematic violations of the Independence Axiom. Researchers such as Konrad and Skaperdas (1993), Schlesinger (1997) and Gollier (2000) have used these forms to revisit many of the applications previously modelled by expected utility theory, such as asset and insurance demand, in order to determine which expected-utility-based results are, and which are not, robust to departures from linearity in the probabilities, and which additional properties of risk-taking behaviour can be modelled.

Although the form $\Sigma_{i=1}^{n} v(x_i) \cdot \pi(p_i)$ was the earliest non-expected utility model to be proposed, it was largely abandoned when it was realized that, whenever the weighting function $\pi(\cdot)$ was nonlinear, the generic inequalities $\pi(p_i) + \pi(p_j) \neq \pi(p_i + p_j)$ and $\pi(p_1) + \ldots + \pi(p_n) \neq 1$ implied discontinuities in the payoffs and inconsistency with first-order stochastic dominance preference. Both problems were corrected by adopting weights $[G(\Sigma_{j=1}^{i} p_j) - G(\Sigma_{j=1}^{i-1} p_j)]$ based on the *cumulative* probability values $p_1, p_1 + p_2, p_1 + p_2 + p_3, \ldots$, to obtain the Rank-Dependent form. Under the above-mentioned restrictions on this form, these weights necessarily sum to unity, and the Rank-Dependent form has emerged as the most widely adopted model in both theoretical and applied analyses. The Dual Expected Utility and Ordinal Independence forms are based on similar weighting formulas.

Unlike the other models, the regret theory form dispenses with the assumption of a preference function over lotteries, and instead derives choice from the psychological notions of *rejoice* and *regret* – that is, the reaction to receiving outcome x when an alternative decision would have led to outcome x^*. The primitive of this model is a *regret:rejoice function* $R(x, x^*)$ which is positive if x is preferred to x^*, negative if x^* is preferred to x, zero if they are indifferent, and satisfies the skew-symmetry condition $R(x, x^*) \equiv -R(x^*, x)$. In a choice between lotteries $\mathbf{P} = (x_1, p_1; \ldots; x_n, p_n)$ and $\mathbf{P}^* = (x_1^*, p_1^*; \ldots; x_{n^*}^*, p_{n^*}^*)$ which are realized independently, the individual's *expected rejoice* from choosing **P** over $\mathbf{P}^*$ is given by $\Sigma_{i=1}^{n} \Sigma_{j=1}^{n^*} R(x_i, x_j^*) \cdot p_i \cdot p_j^*$, and the individual is predicted to choose **P** if this value is positive, $\mathbf{P}^*$ if it is negative, and be indifferent if it is zero (various proposals for extending this approach beyond pairwise choice have been offered). Since this model specifies choice in pairwise comparisons rather than preference levels of individual lotteries, it allows choice to be intransitive, so the individual might select **P** over $\mathbf{P}^*$, $\mathbf{P}^*$ over $\mathbf{P}^{**}$, and $\mathbf{P}^{**}$ over **P**. Though some have argued that such cycles allow for the phenomenon of 'money pumps', it has allowed the model to serve as a proposed solution to the Preference Reversal Phenomenon.

Generalized expected utility analysis

An alternative approach to non-expected utility preferences does not rely upon any specific functional form, but links properties of attitudes toward risk directly to the probability derivatives of a general 'smooth' preference function $V(\mathbf{P}) = V(x_1, p_1; \ldots; x_n, p_n)$. Such analysis reveals that the basic analytics of the expected utility model are in fact quite robust to general smooth departures from linearity in the probabilities. This approach is based on the observations that for the expected utility function $V_{EU}(x_1, p_1; \ldots; x_n, p_n) \equiv U(x_1) \cdot p_1 + \ldots + U(x_n) \cdot p_n$, the value $U(x_i)$ can be interpreted as the coefficient of p_i, and that many theorems involving a linear function's *coefficients* continue to hold when generalized to a nonlinear function's *derivatives*. By adopting the notation $U(x; \mathbf{P}) \equiv \partial V(\mathbf{P}) / \partial \mathrm{prob}(x)$ and the term 'local utility function' for the function $U(\cdot; \mathbf{P})$, standard expected utility characterizations such as those listed at the beginning of this article can be generalized to any smooth non-expected utility preference function $V(\mathbf{P})$ in the following manners (Machina, 1982):

- $V(\cdot)$ exhibits global *first order stochastic dominance preference* if and only if, at each lottery **P**, its local utility function $U(x; \mathbf{P})$ is an increasing function of x.
- $V(\cdot)$ exhibits global *risk aversion* (aversion to small or large mean-preserving increases in risk) if and only if, at each lottery **P**, its local utility function $U(x; \mathbf{P})$ is a concave function of x.
- $V^*(\cdot)$ is globally *at least as risk averse as* $V(\cdot)$ if and only if, at each lottery **P**, $V^*(\cdot)$'s local utility function $U^*(x; \mathbf{P})$ is a concave transformation of $V(\cdot)$'s local utility function $U(x; \mathbf{P})$.

Similar generalizations of expected utility results and characterizations can be obtained for general comparative statics analysis, the theory of asset demand, and the demand for insurance. With regard to the Allais Paradox and other observed violations of the Independence Axiom, it can be shown that the indifference curves of a smooth preference function $V(\cdot)$ will fan out in the probability triangle if and only if $U(x; \mathbf{P}^*)$ is a concave transformation of $U(x; \mathbf{P})$ whenever $\mathbf{P}^*$ first-order stochastically dominates **P**. This analytical approach has been extended to larger classes of preference functionals and distributions by Chew, Karni and Safra (1987), Karni (1987; 1989) and Wang (1993), formally axiomatized by Allen (1987), and applied to the analysis of choices under uncertainty by Chew, Epstein and Zilcha (1988), Chew and Nishimura (1992), Dekel (1989), Green and Jullien (1988), Machina (1984; 1989; 1995) and others.

Non-expected utility preferences under subjective uncertainty

Recent years have seen a growing interest in models of choice under subjective uncertainty, with efforts to represent and analyse departures from both expected utility risk preferences and probabilistic beliefs. A non-expected utility preference function $W(f(\cdot)) \equiv W(x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n)$ over subjective acts $f(\cdot) = [x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n]$ is said to be

probabilistically sophisticated if it takes the form $W(f(\cdot)) \equiv V(x_1, \mu(E_1); \ldots; x_n, \mu(E_n))$ for some subjective probability measure $\mu(\cdot)$ over the space of events and some non-expected utility preference function $V(\mathbf{P}) = V(x_1, p_1; \ldots; x_n, p_n)$. Such preferences have been axiomatized in a manner similar to Savage's (1954) axiomatization of the subjective expected utility form $W_{SEU}(f(\cdot)) \equiv U(x_1) \cdot \mu(E_1) + \ldots + U(x_n) \cdot \mu(E_n)$ (Machina and Schmeidler, 1992). Although such preferences can be consistent with Allais-type departures from linearity in (subjective) probabilities, they are not consistent with Ellsberg-type departures from probabilistic beliefs.

Efforts to accommodate the Ellsberg Paradox and the general phenomenon of ambiguity aversion have led to the development of several non-probabilistically sophisticated models of preferences over subjective acts (see the analysis of Epstein, 1999, as well as the surveys of Camerer and Weber, 1992, and Kelsey and Quiggin, 1992). One such model, the *maximin expected utility* form, replaces the unique probability measure $\mu(\cdot)$ of the subjective expected utility model by a finite or infinite family $\mathcal{M}$ of such measures, to obtain the preference function

$$W_{maximin}(x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n) \equiv \min_{\mu(\cdot) \in \mathcal{M}} [U(x_1) \cdot \mu(E_1) + \ldots + U(x_n) \cdot \mu(E_n)]$$

When applied to the Ellsberg Paradox, the family of subjective probability measures $\mathcal{M} = \{(\mu(\text{red}), \mu(\text{black}), \mu(\text{yellow})) = (1/3, \gamma, 2/3 - \gamma) | \gamma \in [0, 2/3]\}$ will yield the typical Ellsberg-type choices of $f_1(\cdot)$ over $f_2(\cdot)$ and $f_4(\cdot)$ over $f_3(\cdot)$ (Gilboa and Schmeidler, 1989).

Another important model for the representation and analysis of ambiguity averse preferences, based on the Rank Dependent form under objective uncertainty, is the *Choquet expected utility* form:

$$W_{Choquet}(x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n) \equiv \sum_{i=1}^{n} U(x_i) \cdot \left[C\left(\cup_{j=1}^{i} E_j\right) - C\left(\cup_{j=1}^{i-1} E_j\right) \right]$$

where for each act $f(\cdot) = [x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n]$, the payoffs must be labelled so that $x_1 \leq \ldots \leq x_n$, and $C(\cdot)$ is a *nonadditive* measure over the space of events which satisfies $C(\varnothing) = 0$ and $C(\cup_{i=1}^{n} E_i) = 1$ (Gilboa, 1987; Schmeidler, 1989). This model has been axiomatized in a manner similar to the subjective expected utility model, and with proper assumptions on the shape of the utility function $U(\cdot)$ and the nonadditive measure $C(\cdot)$ it is capable of demonstrating ambiguity aversion as well as a wide variety of observed properties of risk preferences.

The technique of generalized expected utility analysis under objective uncertainty has also been adopted to the analysis of general non-expected utility/non-probabilistically sophisticated preference functions $W(f(\cdot)) \equiv W(x_1 \text{ on } E_1; \ldots; x_n \text{ on } E_n)$ over subjective acts. So long as such a function is ‘smooth in the events’ it will possess a ‘local expected utility function’ (which may be state-dependent) and a ‘local probability measure’ at each act $f(\cdot)$, and classical results involving expected utility risk preferences and probabilistic beliefs can typically be generalized in the manner described above (Machina, 2005).

See Also

- Allais paradox
- expected utility hypothesis
- preference reversals
- prospect theory
- risk
- risk aversion
- Savage's subjective expected utility model
- uncertainty

Bibliography

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'Ecole Américaine. *Econometrica* 21, 503–46.

Allen, B. 1987. Smooth preferences and the local expected utility hypothesis. *Journal of Economic Theory* 41, 340–55.

Arrow, K. 1965. *Aspects of the Theory of Risk Bearing*. Helsinki: Yrjö Jahnsson Säätiö.

Battalio, R., Kagel, J. and Macdonald, D. 1985. Animals' choices over uncertain outcomes. *American Economic Review* 75, 597–613.

Becker, S. and Brownson, F. 1964. What price ambiguity? Or the role of ambiguity in decision-making. *Journal of Political Economy* 72, 62–73.

Bell, D. 1985. Disappointment in decision making under uncertainty. *Operations Research* 33, 1–27.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*. Trans. as Exposition of a new theory on the measurement of risk. *Econometrica* 22(1954), 23–36.

Camerer, C. 1989. An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty* 2, 61–104.

Camerer, C. and Weber, M. 1992. Recent developments in modeling preferences: uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5, 325–70.

Carlin, F. 1990. Is the Allais Paradox robust to a seemingly trivial change of frame? *Economics Letters* 34, 241–4.

Chew, S. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais Paradox. *Econometrica* 51, 1065–92.

Chew, S., Epstein, L. and Segal, U. 1991. Mixture symmetry and quadratic utility. *Econometrica* 59, 139–63.

Chew, S., Epstein, L. and Zilcha, I. 1988. A correspondence theorem between expected utility and smooth utility. *Journal of Economic Theory* 46, 186–93.

Chew, S., Karni, E. and Safra, Z. 1987. Risk aversion in the theory of expected utility with rank dependent probabilities. *Journal of Economic Theory* 42, 370–81.

Chew, S. and Nishimura, N. 1992. Differentiability, comparative statics, and non-expected utility preferences. *Journal of Economic Theory* 56, 294–312.

Chew, S. and Waller, W. 1986. Empirical tests of weighted utility theory. *Journal of Mathematical Psychology* 30, 55–72.

Dekel, E. 1989. Asset demands without the independence axiom. *Econometrica* 57, 163–9.

Edwards, W. 1955. The prediction of decisions among bets. *Journal of Experimental Psychology* 50, 201–14.

Edwards, W. 1962. Subjective probabilities inferred from decisions. *Psychological Review* 69, 109–35.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Epstein, L. 1999. A definition of uncertainty aversion. *Review of Economic Studies* 66, 579–608.

Friedman, M. and Savage, L. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56, 279–304.

Gilboa, I. 1987. Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics* 16, 65–88.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Gollier, C. 2000. Optimal insurance design: what can we do without expected utility. In *Handbook of Insurance*, ed. G. Dionne. Boston: Kluwer Academic Publishers.

Green, J. and Jullien, B. 1988. Ordinal independence in non-linear utility theory. *Journal of Risk and Uncertainty* 1, 355–87.

Hagen, O. 1979. Towards a positive theory of preferences under risk. In *Expected Utility Hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen. Dordrecht: Reidel.

Hershey, J., Kunreuther, H. and Schoemaker, P. 1982. Sources of Bias in assessment procedures for utility functions. *Management Science* 28, 936–54.

Hershey, J. and Schoemaker, P. 1980. Risk-taking and problem context in the domain of losses – an expected utility analysis. *Journal of Risk and Insurance* 47, 111–32.

Hershey, J. and Schoemaker, P. 1985. Probability versus certainty equivalence methods in utility measurement: are they equivalent? *Management Science* 31, 1213–31.

Hey, J. 1984. The economics of optimism and pessimism: a definition and some applications. *Kyklos* 37, 181–205.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Kahneman, D. and Tversky, A. 1982. The psychology of preferences. *Scientific American* 246, 160–73.

Kahneman, D. and Tversky, A. 1984. Choices, values and frames. *American Psychologist* 39, 341–50.

Karmarkar, U. 1978. Subjectively weighted utility: a descriptive extension of the expected utility model. *Organizational Behavior and Human Performance* 21, 61–72.

Karmarkar, U. 1979. Subjectively weighted utility and the Allais Paradox. *Organizational Behavior and Human Performance* 24, 67–72.

Karni, E. 1987. Generalized expected utility analysis of risk aversion with state-dependent preferences. *International Economic Review* 28, 229–40.

Karni, E. 1989. Generalized expected utility analysis of multivariate risk aversion. *International Economic Review* 30, 297–305.

Keller, L. 1985. The effects of problem representation on the sure-thing and substitution principles. *Management Science* 31, 738–51.

Kelsey, D. and Quiggin, J. 1992. Theories of choice under ignorance and uncertainty. *Journal of Economic Surveys* 6, 133–53.

Konrad, K. and Skaperdas, S. 1993. Self-insurance and self-protection: a nonexpected utility analysis. *Geneva Papers on Risk and Insurance Theory* 18, 131–46.

Lichtenstein, S. and Slovic, P. 1971. Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89, 46–55.

Loomes, G. and Sugden, R. 1982. Regret Theory: an alternative theory of rational choice under uncertainty. *Economic Journal* 92, 805–24.

MacCrimmon, K. 1965. An experimental study of the decision making behavior of business executives. Doctoral dissertation, University of California, Los Angeles.

MacCrimmon, K. 1968. Descriptive and normative implications of the decision-theory postulates. In *Risk and Uncertainty: Proceedings of a Conference Held by the International Economic Association*, ed. K. Borch and J. Mossin. London: Macmillan.

MacCrimmon, K. and Larsson, S. 1979. Utility theory: axioms versus ‘paradoxes’. In *Expected Utility Hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen. Dordrecht: Reidel.

Machina, M. 1982. ‘Expected utility’ analysis without the independence axiom. *Econometrica* 50, 277–323.

Machina, M. 1983. Generalized expected utility analysis and the nature of observed violations of the independence axiom. In *Foundations of Utility and Risk Theory with Applications*, ed. B. Stigum and F. Wenstøp. Dordrecht: Reidel.

Machina, M. 1984. Temporal risk and the nature of induced preferences. *Journal of Economic Theory* 33, 199–231.

Machina, M. 1987. Choice under uncertainty: problems solved and unsolved. *Journal of Economic Perspectives* 1(1), 121–54.

Machina, M. 1989. Comparative statics and non-expected utility preferences. *Journal of Economic Theory* 47, 393–405.

Machina, M. 1995. Non-expected utility and the robustness of the classical insurance paradigm. *Geneva Papers on Risk and Insurance Theory* 20, 9–50.

Machina, M. 2005. ‘Expected utility/subjective probability’ analysis without the sure-thing principle or probabilistic sophistication. *Economic Theory* 26, 1–62.

Machina, M. and Schmeidler, D. 1992. A more robust definition of subjective probability. *Econometrica* 60, 745–80.

Markowitz, H. 1952. The utility of wealth. *Journal of Political Economy* 60, 151–8.

Marschak, J. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18, 111–41.

Morrison, D. 1967. On the consistency of preferences in Allais' paradox. *Behavioral Science* 12, 373–83.

Moskowitz, H. 1974. Effects of problem representation and feedback on rational behavior in Allais and Morlat-type problems. *Decision Sciences* 5, 225–42.

Múnera, H. and de Neufville, R. 1983. A decision analysis model when the substitution principle is not acceptable. In *Foundations of Utility and Risk Theory with Applications*, ed. B. Stigum and F. Wenstøp. Dordrecht: Reidel.

Pratt, J. 1964. Risk aversion in the small and in the large. *Econometrica* 32, 122–36.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3, 323–43.

Raiffa, H. 1961. Risk, ambiguity, and the Savage axioms: comment. *Quarterly Journal of Economics* 75, 690–4.

Raiffa, H. 1968. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Reading, MA: Addison-Wesley.

Rothschild, M. and Stiglitz, J. 1970. Increasing risk: I. A definition. *Journal of Economic Theory* 2, 225–43.

Rothschild, M. and Stiglitz, J. 1971. Increasing risk: II. Its economic consequences. *Journal of Economic Theory* 3, 66–84.

Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley and Sons. Revised edn, New York: Dover Publications, 1972.

Schlesinger, H. 1997. Insurance demand without the expected utility paradigm. *Journal of Risk and Insurance* 64, 19–39.

Schmeidler, D. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57, 571–87.

Schoemaker, P. and Kunreuther, H. 1979. An experimental study of insurance decisions. *Journal of Risk and Insurance* 46, 603–18.

Segal, U. 1984. Nonlinear decision weights with the independence axiom. Working Paper No. 353, Department of Economics, University of California, Los Angeles.

Slovic, P. 1969. Manipulating the attractiveness of a gamble without changing its expected value. *Journal of Experimental Psychology* 79, 139–45.

Slovic, P., Fischhoff, B. and Lichtenstein, S. 1977. Behavioral decision theory. *Annual Review of Psychology* 28, 1–39.

Slovic, P. and Tversky, A. 1974. Who accepts Savage's axiom? *Behavioral Science* 19, 368–73.

Starmer, C. 2000. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38, 332–82.

Stigum, B. and Wenstøp, F., eds. *Foundations of Utility and Risk Theory with Applications*. Dordrecht: Reidel.

Sugden, R. 1986. New developments in the theory of choice under uncertainty. *Bulletin of Economic Research* 38, 1–24.

Tversky, A. 1975. A critique of expected utility theory: descriptive and normative considerations. *Erkenntnis* 9, 163–73.

Tversky, A. and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science* 211, 453–8.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press. 2nd edn, 1947; 3rd edn, 1953.

Wang, T. 1993. Lp-Fréchet differentiable preference and ‘local utility’ analysis. *Journal of Economic Theory* 61, 139–59.

Weber, M. and Camerer, C. 1987. Recent developments in modeling preferences under risk. *OR Spektrum* 9, 129–51.

Yaari, M. 1987. The dual theory of choice under risk. *Econometrica* 55, 95–115.

How to cite this article

Machina, Mark J. "non-expected utility theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000081>

doi:10.1057/9780230226203.1192

The New Palgrave Dictionary of Economics Online

non-governmental organizations

Caren Grown
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article defines the term ‘non-governmental organizations’ (NGOs) and describes how they operate. It reviews the growth of the NGO sector since the 1980s, examines the reasons why NGOs have proliferated, reviews evidence on NGO impact, and summarizes how economists have modelled and tested hypotheses about the role of NGOs in development assistance.

Keywords

charitable giving; contract theory; development assistance; foreign aid; Grameen Bank; humanitarian and relief work; imperfect information; incomplete contracts; international development; international donors; non-governmental organizations; non-profit organizations; participation of the poor; poverty alleviation; principal agent; private voluntary organizations; public goods; volunteerism

Article

The term ‘non-governmental organization’ came into currency in 1945 when the United Nations Charter distinguished between participation rights for intergovernmental specialized agencies and international organizations (Willetts, 2002). Non-governmental organizations (NGOs) form that subset of non-profit organizations working in development assistance, international disaster relief, poverty alleviation, and human rights in developing countries (see non-profit organizations). In the literature and in practice, the term ‘NGO’ is often used interchangeably with ‘private voluntary organization’, a term used to refer to organizations based in the United States engaged in overseas provision of services (Anheier and Salamon, 1998).

As the NGO sector has grown, so too has the number of definitions, classifications, and taxonomies (Vakil, 1997). According to Bebbington (2004, p. 729), ‘discussions of NGOs continue to be plagued by the vexed and ultimately unanswerable question of “what is an NGO” and haunted by endless typologies. While some of these clarify functional differences, they are less helpful in an explanatory sense – why NGOs emerge, why they do what they do and where, and why certain ideas underlie their actions.’ Despite the lack of a uniform definition, most commentators agree that NGOs can be characterized as private, autonomously managed, value-based organizations that depend, in whole or in part, on charitable donations and voluntary service. Although the sector has become increasingly professionalized since the mid-1980s, principles of altruism and volunteerism remain key defining characteristics.

The lack of a uniform definition reflects the heterogeneity of NGOs around the world. They can be structured as large global federated entities, small community-based organizations, local or national cooperatives, or large national or international membership organizations. They can carry out a range of functions, from advocacy on behalf of vulnerable or other groups, to direct service (such as providing credit, education and health), research, organizing and public education, humanitarian and relief operations, and peacekeeping operations. Their geographic reach may be in a local community or an entire country, or they may operate across many countries. They are not part of the public sector nor are they dependent on the political process, but in various countries some may seek to influence the formal political process. In many countries, they are exempt from taxes on corporate income. Some NGOs receive funding in the form of grants and contracts from governments and private foundations, others from membership dues and individual contributions, and still others from fees for goods or services. Some receive funding from all these sources.

Examples of organizations that fall in the broad category of NGOs include:

- Centro Mujeres, a small community health organization with 12 staff in 2004 dedicated to fostering the empowerment and well-being of women and adolescents in La Paz, Mexico;
- the Bangladesh Rural Advancement Committee (BRAC), a nationwide organization dedicated to poverty alleviation in Bangladesh with branches in 65,000 villages and more than 97,000 employees in 2006;
- Amnesty International, a global organization with over 1.8 million members in over 150 countries in 2006 and local chapters that undertake research and campaigns to protect human rights and prevent abuses.

Given their heterogeneity of purpose, form and function, NGOs are a multidisciplinary topic, and studies on this topic tend to be published in multidisciplinary journals such as *World Development*, the *Journal of International Development* and *Third World Quarterly*. By contrast, the non-profit sector has its own specialized journals. Research on NGOs is far more common in disciplines other than economics; it is an active field in international relations and development studies. Much of the literature is descriptive, relying on historical analysis or contemporary case studies of single countries, single sectors, or single organizations (Bebbington, 2004; Edwards and Hulme, 1996). There is surprisingly little survey-based research on NGOs in developing countries, especially Africa (Barr, Fafchamps and Owens, 2005). The broad literature explores the growth, evolution, and impact of NGOs in development and relief work in different contexts, NGO relationships with states and donors (and firms in a few instances), and community-based action and social change (Lewis and Opoku-Mensah, 2006). NGOs are frequently cast in a favourable light. It is quite common to read articles about the potential of NGOs to transform the development process as opposed to articles about corruption or project failure.

By contrast, the economics literature has tended to develop a narrow range of theoretical models and to take a more critical view of NGOs. Theoretical models explore imperfect information, contracting problems, and accountability in developing countries, using the broad descriptive literature to provide support. With a few exceptions, empirical work by economists has focused mostly on NGOs that provide micro-finance services (Morduch, 1999; Pitt and Khandker, 1998). The exceptions include Barr, Fafchamps and Owens (2005), who conducted a survey to document the funding sources and examine monitoring and oversight procedures of NGOs in Uganda; Gauri and Galef (2005), who analysed data from a nationally representative survey of NGOs in Bangladesh; Gauri and Fruttero (2003), who used the Bangladesh Household Income and Expenditure Survey to examine location decisions of NGO programs; and Leonard (1998), who analysed data on health care providers in Cameroon.

Growth of the NGO sector

Although statistics are hard to come by and what is covered in the numbers can be unclear, Figure 1 shows spectacular growth of the NGO sector since the 1980s.

Figure 1

Total number of NGOs worldwide by year, 1909–1999. *Source*: Agg (2006).

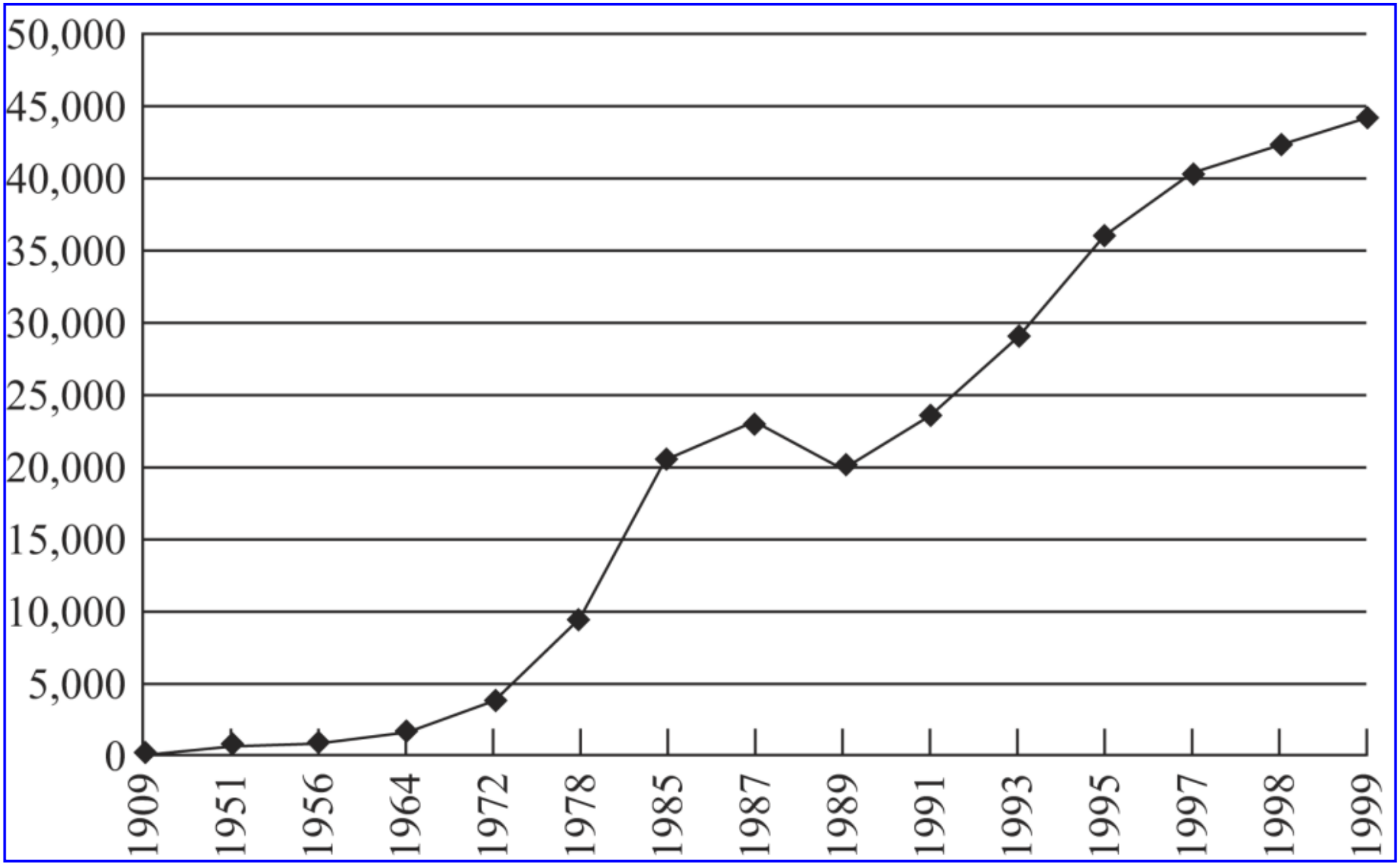

The NGO sector has also proliferated in various countries. A recent survey by Gauri and Galef (2005) shows that Bangladesh has one of the largest and most sophisticated NGO sectors in the developing world: over 90 per cent of villages in the country had at least one NGO in 2000 (Gauri and Fruterro, 2003) and foreign assistance channelled through NGOs has been above ten per cent since 1993 (Gauri and Galef, 2005). As Phinney (2002) notes, ‘In some villages in Bangladesh, you can send your child to an NGO school, have a vasectomy arranged by an NGO health worker, sell your milk to an NGO dairy, and talk on an NGO phone. And, there's usually a choice of NGO banks.’ International NGOs were responsible for the creation of the NGO community in Bangladesh, although they have withdrawn in recent years and now play a secondary role to local NGOs (Stiles, 2002). Price (1999) has noted a significant concentration of NGOs in Latin America, although they are unevenly distributed across countries. In Uganda, Barr, Fafchamps and Owens (2005) identified 3,500 NGOs registered with the government. Ghanaian NGOs provide 40 per cent of clinical care needs, 27 per cent of hospital beds, and 35 per cent of outpatient services, and in Tanzania NGOs provide half of all hospitals and beds and receive half of all curative visits (Leonard and Leonard, 2004). Few national surveys have been undertaken to identify NGO prevalence and incidence in other African countries.

What explains the rise of NGOs?

Development studies scholars (geographers, political scientists, anthropologists) argue that NGO involvement in public projects in developing countries has grown in response to budgetary stringency and public sector cutbacks often imposed by macroeconomic stabilization policies (Bebbington and Farrington, 1993; Edwards and Hulme, 1996). Economists take a less political position, arguing that NGOs are a response to the undersupply of public goods. Bebbington (1997) provides empirical support, noting that Latin American states shifted away from direct implementation of development initiatives in the 1980s and increasingly subcontracted or financed programmes implemented by non-state institutions. In Bolivia, NGOs manage national parks, reserves, and protected areas. In Chile, since the mid-1980s, governments have subcontracted extension services to the private sector; beginning in the 1990s, NGOs and farmers' organizations can also bid for these contracts.

Political scientists and others also highlight the changing preferences of international funders to direct money through private channels due to increasing donor frustration with the public sector because of corruption, inefficiency, and poor results in reducing poverty (Clark, 1991; Edwards and Hulme, 1996). Empirical evidence suggests that donors have played a key role in the proliferation of international and national NGOs. According to Woods (2003, p. 9), ‘resources channeled through NGOs in all OECD member countries rose from 0.2 percent of the total bilateral ODA [official development assistance] of members of the Development Assistance Committee in 1970 to 17.0 percent in 1996, to reach, in absolute terms an amount equal to twice the total 1996 ODA of the United Kingdom, the DAC's sixth largest donor by volume.’ OECD Development Assistance Committee (DAC) figures show that net grants by NGOs rose from five per cent of total net flows in 2000 to eight per cent of flows in 2004 (OECD, 2005, Table 2).

The data have many limitations, and these numbers are likely to be an underestimate. There are complex reporting requirements that are interpreted differently by different governments. For example, donors must choose between designating a disbursement as ‘emergency and distress relief’ or a grant to an NGO (Agg, 2006). Nor do the data include US funds channeled through NGOs. Nonetheless, the OECD data are the only aid data collected over time and from all donor governments.

Meyer (1995) concurs that NGOs arise in part because of donor dissatisfaction with the level of public goods in developing countries, so donors turn to NGOs, which are seen to have some comparative advantages over governments. A number of contributions to the *World Development* special issue on NGOs in 1987 claim that NGOs have better information on the needs of poor people than do governments; have lower transaction costs; are more flexible than governments and better able to respond to crises such as drought or floods. Because they are part of dense networks with close ties to the community, they are also better at fostering community participation and responding to local needs (Bebbington, 2004). Bebbington (2004), for instance, documents how NGOs use methodologies and actions that strengthen capacity and involve poor people in project activities in Latin America. Finally, NGOs are seen to promote new ideas and practices (Scott and Hopkins, 1999; Meyer, 1995).

What is the evidence on NGO impact?

There is little systematic empirical evidence to either support or refute the notion that NGOs are more cost-effective than governments. Some country case studies find that large NGOs working in some sectors do provide some services more cost-effectively than governments (Hasan, 1993; AFK/NOVIB 1993; Riddell and Robinson, 1992), while others find little difference between governments and NGOs (Tendler, 1982; 1989).

Similarly, the evidence on whether NGOs are better at reaching the poorest is also mixed (Fowler, 2000; Edwards and Hulme, 1996; Arellano-Lopez and Petras, 1994; Riddell and Robinson, 1992; Tendler, 1982). Most NGOs reach the poor, but not necessarily the poorest (UNRISD, 2000). An analysis of NGO activity in Bangladesh found that NGO assistance reached those in the second wealth quintile but not those in the poorest (Gauri and Galef, 2005). Even the most well-known NGO in Bangladesh, the Grameen Bank, was found to reach less than 20 per cent of landless households in the country (Farrington and Lewis, 1993).

There is greater empirical support for the notion that NGOs have pioneered and used instruments that emphasize the participation of the poor in poverty and development projects (Clark, 1995; Bratton, 1990). Kilby (2006) finds that formal participation measures and ‘downward’ accountability practices (for example, to members, clients, other beneficiaries) are correlated with empowerment outcomes in India. Bebbington and Farrington (1993) observe that NGOs that emphasized project methodologies and actions that promote participation have increased the impacts of agricultural development projects.

A number of studies document NGO innovations in various sectors of service delivery, for example in financial services for the poor (Hulme and Mosely, 1996), in the creation of debt-for-nature swaps (Meyer, 1995), in agriculture technology development (Bebbington and Farrington, 1993), and in oral rehydration therapy (Howes and Sattar, 1992).

The literature also highlights concerns about the effects of donor financing on NGOs. Edwards and Hulme (1996) argue that increasing reliance on donor funding weakens key attributes that make NGOs attractive to donors in the first place. It can reduce advocacy efforts on behalf of poor and vulnerable groups, negatively affect NGO institutional development, weaken their legitimacy as independent actors, distort their accountability away from internal constituencies to donors and patrons, and lead to an overemphasis on short-term outputs. Fyvie and Ager (1999) argue that donor requirements constrain NGO capacity for innovation. Bebbington (1997; 2005) shows that donor funding of three poverty-

oriented rural development NGOs in Peru has over time had several of these effects.

Concerns have also been raised about the nature of NGO, government and donor relations. Scholars have uncovered a range of relationships between NGOs and government, from strongly adversarial to tight partnerships, and between NGOs and donors, from dependent recipient to co-financers/implementers of projects. They have also identified the institutional, economic and political factors that condition which types of relationship emerge and are sustained in different contexts (Bebbington, 1997; Nelson, 2006; Atack, 1999; Anheier and Salamon, 1998; Ahmad, 2006). By contrast, the economics literature has focused largely on the relationship between NGOs and donors and the conditions for different types of partnerships.

NGO role in development assistance: theoretical models

The theoretical economics literature has yet to reflect the diversity of NGO types, the varied impacts of NGO projects, and the multiplicity of NGO, donor, and government relations. Most of this literature focuses on how NGOs compensate for the undersupply of government-provided public goods or are a device to overcome imperfect information and incomplete contracts. Economists have applied principal–agent models with NGOs to the African health-care sector and foreign assistance chains.

Scott and Hopkins (1999) identify the organizational comparative advantage of NGOs and develop a model that explains the circumstances under which they emerge and dominate other types of firms/entities. NGOs predominate in environments where public goods are undersupplied to citizens whose demand for that good exceeds demand of the median voter. The authors argue that the potential superiority of NGOs derives from an institutional environment that selectively attracts altruists who have a lower reservation wage than egotists, and who have the ability to develop efficient technologies for converting the effort of their staffs into local outputs highly valued by the target group of beneficiaries.

The technical superiority of NGOs stems from the way NGOs operate – their interaction with local communities, which enables them to articulate and aggregate local demands. Additionally, NGOs recruit field staff from among beneficiaries and target groups, which facilitates communication and assists in creation of trust between beneficiaries and the target agency. As donors get to know the field and seek the most efficient organizations, NGOs would generally dominate when they have the same or better development technologies than public agencies and wages are similar in both sectors. They may dominate even when wages paid by public agencies are higher, if NGO technology is superior and warm-glow effects are strong enough to outweigh the wage differential.

Besley and Ghatak (2001) develop a model of NGO involvement in public goods provision and enumerate several propositions based on observations from the case study literature. Pure NGO involvement will be more prevalent in projects where the marginal cost of public funds is high and/or the public sector is relatively less efficient in input provision. In activities where performance is hard to measure, NGOs are perceived to be committed to high quality or serve some groups better than others due to their religious or ideological orientation. NGO involvement in supplying services is less dominant in types of projects that are infrastructure-intensive and in countries where the government manages infrastructure well. Decentralization initiatives have often resulted in increased NGO involvement, in part because resource constraints are more severe. NGO provision will also be more prevalent in projects where the NGO cares more about the beneficiaries. Support for this proposition is provided by the *World Development Report 1997* (World Bank, 1997), which described how governments typically prefer NGOs for delivery of social services while preferring for-profit contractors for the management of infrastructure, such as road maintenance in Brazil.

The models of Leonard (2002) and Leonard and Leonard (2004) address imperfect information and incomplete contracts in the health-care sector in Africa. In sectors where goods or services are characterized by asymmetric information, such as in health care, mechanisms other than prices are needed for the market to function well. In Africa, NGOs are one mechanism to solve the asymmetric information problem. Leonard (2002) and Leonard and Leonard (2004) show they have a stock of attributes which, when combined with the institutional environment in Africa, make them more successful than governments with similar values in providing quality services and reducing the transaction costs of asymmetric information. (There are few private providers in Africa so the relevant comparison is between government and non-governmental services.)

Finally, Azam and Laffont (2003) apply contract theory to shed light on the aid relationship between a donor and recipient country, where consumption of the poor is assumed to be an international public good. The authors model the intricacies of coordinating the efforts of government and NGOs in the fight against poverty. When aid is introduced, several possibilities emerge. Most importantly, free riding problems arise in the provision of aid to the poor when there are several providers. Yontcheva (2003) models a dynamic game between a principal (donor) and agent (NGO), where the model's objective is to identify the long-term determinants of the principal's choice of whether to delegate a project to an NGO and to verify the impact of the allocation on the principal's payoff and the agent's effort.

Conclusion

NGOs are a burgeoning field of cross-disciplinary study. Economists can learn from this voluminous literature both to enrich their models and to contribute theoretical and empirical rigour to a rather messy descriptive literature. They can develop richer theorizations of NGOs roles, relationships and power vis-à-vis governments and donors. They can also work with other social scientists to gather better data on the range of NGO motivations, roles and impacts in various country contexts. This information can help fill an important gap in

understanding the dynamic and growing NGO sector in developing countries.

See Also

- non-profit organizations
- poverty alleviation programmes

Bibliography

AKFC/NOVIB (Aga Khan Foundation Canada/Netherlands Organization for International Development Cooperation). 1993. *Going To Scale: The BRAC Experience 1972–1992 and Beyond*. The Hague: AKFC/NOVIB.

Agg, C. 2006. Trends in government support for non-governmental organizations: is the ‘golden age’ of the NGO behind us? Programme Paper No. 23. Geneva: UNRISD.

Ahmad, M.M. 2006. The ‘partnership’ between international NGOs (non-governmental organisations) and local NGOs in Bangladesh. *Journal of International Development* 18, 629–38.

Anheier, H. and Salamon, L.M. 1998. *The Nonprofit Sector in The Developing World: A Comparative Analysis*. Manchester and New York: Manchester University Press.

Arellano-Lopez, S. and Petras, J.F. 1994. Non-governmental organizations and poverty alleviation in Bolivia. *Development and Change* 25, 555–68.

Atack, I. 1999. Four criteria of development NGO legitimacy. *World Development* 27, 855–64.

Azam, J. and Laffont, J. 2003. Contracting for aid. *Journal of Development Economics* 70, 25–58.

Barr, A., Fafchamps, M. and Owens, T. 2005. The governance of non-governmental organizations in Uganda. *World Development* 33, 657–79.

Bebbington, A. 1997. New states, new NGOs? Crises and transitions among rural development NGOs in the Andean region. *World Development* 25, 1755–65.

Bebbington, A. 2004. NGOs and Uneven development: geographies of development intervention. *Progress in Human Geography* 28, 725–45.

Bebbington, A. 2005. Donor–NGO relations and representations of livelihood in nongovernmental aid chains. *World Development* 33, 937–50.

Bebbington, A. and Farrington, J. 1993. Governments, NGOs and agricultural development: perspectives on changing inter-organisational relationships. *Journal of Development Studies* 29, 199–219.

Besley, T. and Ghatak, M. 2001. Government versus private ownership of public goods. *Quarterly Journal of Economics* 116, 1343–72.

Bratton, M. 1990. NGOs in Africa: can they influence public policy? *Development and Change* 21, 87–188.

Clark, J. 1991. *Democratizing Development: The Role of Voluntary Organizations*. West Hartford, CT: Kumarian Press.

Clark, J. 1995. The state, popular participation, and the voluntary sector. *World Development* 23, 593–601.

Collier, P. and Dollar, D. 2004. Development effectiveness: what have we learnt? *Economic Journal* 114, F244–F271.

Edwards, M. and Hulme, D. 1996. Too close for comfort? The impact of official aid on nongovernmental organizations. *World Development* 24, 961–73.

Farrington, J. and Lewis, D., eds. (with S. Satish and A. Miclat-Teves). 1993. *NGOs and the State in Asia: Rethinking Roles in Sustainable Agricultural Development*. London: Routledge.

Fowler, A. 2000. NGO futures: beyond aid: NGDO values and the fourth position. *Third World Quarterly* 21, 589–603.

Fyvie, C. and Ager, A. 1999. NGOs and innovation: organizational characteristics and constraints in development assistance work in The Gambia. *World Development* 27, 1383–95.

Gauri, V. and Fruttero, A. 2003. Location decisions and nongovernmental organization motivation: evidence from rural Bangladesh. Policy Research Working Paper No. 3176. Washington, DC: World Bank.

Gauri, V. and Galef, J. 2005. NGOs in Bangladesh: activities, resources, and governance. *World Development* 33, 2045–65.

Hasan, A. 1993. *Scaling-up the OPP's Low-Cost Sanitation Programme*. Karacji: OPP-RTI.

Howes, M. and Sattar, M. 1992. Bigger and better? Scaling-up strategies pursued by BRAC 1972–1991. In *Making a Difference: NGOs and Development in a Changing World*, ed. M. Edwards and D. Hulme. London: Earthscan.

Hulme, D. and Mosely, P. 1996. *Finance Against Poverty*. London: Routledge.

Kilby, P. 2006. Accountability for empowerment: dilemmas facing non-governmental organizations. *World Development* 34, 951–63.

Leonard, K.L. 1998. Institutional structure of health care in rural Cameroun: structural estimation of production in teams with unobservable effort. Discussion Paper No. 9798–16, Department of Economics, Columbia University.

Leonard, K.L. 2002. When both states and markets fail: asymmetric information and the role of NGOs in African health care. *International Review of Law and Economics* 22, 61–80.

Leonard, K.L. and Leonard, D.K. 2004. The political economy of improving health care for the poor in rural Africa: institutional solutions to the principal–agent problem. *Journal of Development Studies* 40, 50–77.

Lewis, D. and Opoku-Mensah, P. 2006. Moving forward research agendas on international NGOs: theory, agency and context. *Journal of International Development* 18, 665–75.

Meyer, C.A. 1995. Opportunism and NGOs: entrepreneurship and green North–South transfers. *World Development* 23, 1277–89.

Morduch, J. 1999. The microfinance promise. *Journal of Economic Literature* 37, 1569–615.

Nelson, P. 2006. Policy arena, the varied and conditional integration of NGOs in the aid system: NGOs and the World Bank. *Journal of International Development* 18, 701–13.

OECD (Organisation for Economic Co-operation and Development). 2005. *Geographical Distribution of Financial Flows to Aid Recipients 2000–2004*. Paris: OECD/DAC.

Phinney, R. 2002. A model NGO? *Global Policy Forum*. Online. Available at http://www.globalpolicy.org/ngos/fund/2002/1205model.htm, accessed 31 March 2007.

Pitt, M. and Khandker, S. 1998. The impact of group-based credit programs on poor households in Bangladesh: does the gender of participants matter? *Journal of Political Economy* 106, 958–96.

Price, M.D. 1999. Non-governmental organizations on the geopolitical frontline. In *Reordering the World. Geopolitical Perspectives on the 21st Century*, ed. G. Demko and W.B. Wood. Boulder, CO: Westview Press.

Riddell, R. and Robinson, M. 1992. The impact of NGO poverty alleviation projects: results of the case study evaluations. Working Paper No. 68. London: Overseas Development Institute.

Scott, R. and Hopkins, C.D. 1999. The economics of non-governmental organizations. Development Economics Discussion Paper No. 15, London School of Economics.

Stiles, K. 2002. International support for NGOs in Bangladesh: some unintended consequences. *World Development* 30, 835–46.

Tendler, J. 1982. Turning private voluntary agencies into development agencies: questions for evaluation. Evaluation Discussion Paper No. 10. Washington DC: USAID.

Tendler, J. 1989. Whatever happened to poverty alleviation? *World Development* 17, 1033–44.

UNRISD (United Nations Research Institute for Social Development). 2000. *Visible Hands: Taking Responsibility for Social Development*. Geneva: UNRISD.

Vakil, A. 1997. Confronting the classification problem: toward a taxonomy of NGOs. *World Development* 25, 2057–70.

Weisbrod, B.A. 1972. Towards a theory of the voluntary nonprofit sector in a three-sector economy. In *Altruism, Morality, and Economic Theory*, ed. E.S. Phelps. New York: Russell Sage Foundation.

Willetts, P. 2002. What is a non-governmental organization? Article 1.44.3.7, UNESCO Encyclopedia of Life Support Systems. Online. Available at http://www.staff.city.ac.uk/p.willetts/CS-NTWKS/NGO-ART.HTM, accessed 7 March 2007.

Wood, G.D. 1997. States without citizens: the problem of the franchise state. In *NGOs, States and Donors: Too Close for Comfort?* ed. M. Edwards and D. Hulme. London: Macmillan.

Woods, A. 2003. *Facts about European NGOs Active in International Development*. Paris: Development Centre, OECD.

World Bank. 1997. *World Development Report 1997: State in a Changing World*. New York: Oxford University Press.

World Bank. 2001. *Attacking poverty. World Development Report, 2000–2001*. Washington, DC: World Bank.

Yontcheva, B. 2003. Hierarchy and authority in a dynamic perspective: a model applied to donor financing of NGO proposals. Working Paper No. 03/157, International Monetary Fund.

How to cite this article

Grown, Caren. "non-governmental organizations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000144> doi:10.1057/9780230226203.1193

The New Palgrave Dictionary of Economics Online

nonlinear panel data models

Ekaterini Kyriazidou
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

conditional maximum likelihood; conditioning variables; fixed effects; heteroskedasticity; incidental parameters; individual effects; logit models; longitudinal data; nonlinear panel data models; quadrature procedures; random effects; random variables; Tobit models

Article

Panel or longitudinal data are becoming increasingly popular in applied work as they offer a number of advantages over pure cross-sectional or pure time-series data. They allow researchers to model *unobserved heterogeneity* at the level of the observational unit, where the latter may be an individual, a household, a firm or a country. This article describes several estimation methods that are available for nonlinear panel data models, that is, models which are nonlinear in the parameters of interest and which include models that arise frequently in applied work, such as discrete choice models and limited dependent variable models, among others.

1 Introduction

Panel or longitudinal data are becoming increasingly popular in applied work as they offer a number of advantages over pure cross-sectional or pure time-series data. A particularly useful feature is that they allow researchers to model *unobserved heterogeneity* at the level of the observational unit, where the latter may be an individual, a household, a firm or a country. Standard practice in the econometric literature is to model this heterogeneity as an individual-specific effect which enters additively in the model, typically assumed to be linear, that captures the statistical relationship between the dependent and the independent variables. The presence of these *individual effects* may cause problems in estimation. In particular in short panels, that is, in panels where the time-series dimension is of smaller order than the cross-sectional dimension, their estimation in conjunction with the other parameters of interest usually yields inconsistent estimators for both. (Notable exceptions are the static linear and the Poisson count panel data models, where estimation of the individual effects along with the finite dimensional coefficient vector yields consistent estimators of the latter.) This is the well-known *incidental parameters* problem (Neyman and Scott, 1948). In linear regression models, this problem may be dealt with by taking transformations of the model, such as first differences or differences from time averages ('within transformation'), which remove the individual effect from the equation under consideration. However they do not apply to nonlinear econometric models, that is, models which are nonlinear in the parameters of interest and which include models that arise frequently in applied work, such as discrete choice models, limited dependent variable models, and duration models, among others.

This article describes several estimation methods that are available for nonlinear panel data models. An approach that is available for estimating certain linear and nonlinear parametric models with individual effects is the *conditional maximum likelihood* approach. This is described in Section 2. Section 3 describes estimation techniques that have been recently developed for several semiparametric nonlinear panel data models. A common feature in the methods discussed in that section is that we do not make any assumptions about the nature of these individual effects, that is, whether they are fixed constants or random variables. Thus, we do not make any assumptions about whether they are related to the conditioning variables and, if so, in what manner. This approach is typically referred to as the *fixed effects* approach. Section 4 describes the so-called *random effects* approach in estimating nonlinear panel data models. In contrast to the fixed effects approach, the random effects approach does make assumptions about the individual effects.

The discussion distinguishes between two types of models, *static* and *dynamic*. In static models, the conditioning set includes past, present and future values of the variables. In this case the conditioning variables are said to be *strictly exogenous*. In dynamic models, the conditioning set may also include lags of the dependent variable and other endogenous variables, that is, variables that are only *weakly exogenous* or *predetermined*.

Our discussion is limited in several aspects. First, we focus only on the case when the time series dimension of the panel (T) is short so that it makes sense to consider the asymptotic properties of the estimators when the cross-sectional dimension (N) is large while T remains fixed. Second, we do not consider estimation of *random coefficient models*, that is, models where all the parameters are varying at the individual level. Finally, we do not discuss the Bayesian approach to estimating panel data models.

2 The conditional maximum likelihood (CML) approach

Suppose that a random variable y_{it} has density $f(\cdot, \theta, \alpha_i)$ where θ is the parameter of interest which is common across all units i, whereas α_i is a nuisance parameter which is allowed to differ across i. A *sufficient statistic* S_i for a_i is a function of the data such that the conditional distribution of the data given S_i does not depend on α_i. However, the conditional distribution may depend on θ. In this case, one can estimate θ by maximizing the *conditional likelihood function*, which conditions on the sufficient statistic(s). Such sufficient statistics are readily available for the exponential family that includes the normal, Poisson, gamma, logistic, and binomial distributions. The CML approach, when it exists, yields consistent and asymptotically normal estimators for parametric panel data models with individual effects (Andersen, 1970). We will next demonstrate how the CML approach works in the case of a static and a dynamic logit model with individual effects.

2.1 The static panel data logit model

Consider the binary choice logit model with individual effects

$$y_{it} = 1\{x_{it}\beta_0 + \alpha_i + \varepsilon_{it} \geq 0\} i = 1, \ldots, N; \quad t = 1, \ldots, T$$

where $1\{A\} = 1$ if A occurs and is 0 otherwise. Let $x_i \equiv (x_{i1}, \ldots, x_{iT})$. Here the error term ε_{it} is distributed i.i.d. over t with a logistic distribution conditional on (x_i, α_i). Note that this assumption implies that ε_{it} is in fact independent of α_i and x_{it} for all t. We can easily calculate that

$$\Pr(y_{it} = 1 | x_i, \alpha_i) = \frac{\exp(x_{it}\beta_0 + \alpha_i)}{1 + \exp(x_{it}\beta_0 + \alpha_i)}.$$

In this model it turns out that $\Sigma_t y_{it}$ is a sufficient statistic for α_i. Indeed, let $T = 2$. Note that

$$\Pr(y_{it} = 1 | y_{i1} + y_{i2} = 0, x_i, \alpha_i) = 0 \Pr(y_{it} = 1 | y_{i1} + y_{i2} = 2, x_i, \alpha_i) = 1$$

that is, individuals who do not switch states (i.e. who are 0 or 1 in both periods) do not offer any information about β_0. But it can be easily shown that

$$\Pr(y_{i1} = 1 | y_{i1} + y_{i2} = 1, x_i, \alpha_i) = \frac{1}{1 + \exp((x_{i2} - x_{i1})\beta_0)}$$

and

$$\Pr(y_{i1} = 0 | y_{i1} + y_{i2} = 1, x_i, \alpha_i) = \frac{\exp((x_{i2} - x_{i1})\beta_0)}{1 + \exp((x_{i2} - x_{i1})\beta_0)}.$$

In other words, conditional on the individual switching states (from 0 to 1 or from 1 to 0), the probability that y_{it} is 1 or 0 depends on β_0 (that is, contains information about β_0) but is independent of α_i.
The conditional log-likelihood is

$$\mathcal{L}_C(\beta) = \sum_{i=1}^{N} 1\{y_{i1} + y_{i2} = 1\} \ln\left(\frac{\exp((x_{i2} - x_{i1})\beta)^{(1-y_{i1})}}{1 + \exp((x_{i2} - x_{i1})\beta)}\right)$$

and may be maximized over β to produce a consistent and root-*N* asymptotically normal estimator of β_0. Note that the approach uses a subset of the data, since only individuals who switch states enter the likelihood. For the expression of the conditional log-likelihood in the general *T* case, see Chamberlain (1984).

2.2 The dynamic panel data logit model

Chamberlain (1985) noticed that the conditional maximum likelihood approach also applies to the 'AR(1)' logit model with individual effects:

$$y_{it} = 1\{\gamma_0 y_{it-1} + \alpha_i + \varepsilon_{it} \geq 0\} i = 1, \ldots, N; \ t = 1, \ldots, T$$

where the error term ε_{it} is distributed i.i.d. with a logistic distribution conditional on α_i and the initial observation of the sample y_{i0}. Note that we are not making any assumption about the distribution of the initial y_{i0}. As we will see, the approach requires at least four observations for each individual (including the initial observation). In fact, let that be the case and consider the events:

$$A = \{y_{i0} = d_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3\} B = \{y_{i0} = d_0, y_{i1} = 1, y_{i2} = 0, y_{i3} = d_3\}$$

where d_0 and d_3 are either 0 or 1. It is rather easy to derive the following probabilities which condition on the individual switching states in the two middle periods

$$\Pr(A|A \cup B, \alpha_i) = \frac{1}{1 + \exp(\gamma_0(d_0 - d_3))} \Pr(B|A \cup B, \alpha_i) = \frac{\exp(\gamma_0(d_0 - d_3))}{1 + \exp(\gamma_0(d_0 - d_3))}.$$

Note that these depend on γ_0 but are independent of α_i. The conditional log-likelihood of the model for four periods is:

$$\mathcal{L}_C(\beta) = \sum_i 1\{y_{i1} + y_{i2} = 1\} \ln\left(\frac{\exp(\gamma(y_{i0} - y_{i3}))^{y_{i1}}}{1 + \exp(\gamma(y_{i0} - y_{i3}))}\right)$$

and maximizing it with respect to γ produces a consistent and root-*N* asymptotically normal estimator. The approach generalizes to logit models with more than one lags of y_{it} (see Magnac, 2000).
It is important to note that the CML approach described above does *not* work in the logit model

$$y_{it} = 1\{\gamma_0 y_{it-1} + x_{it}\beta_0 + \alpha_i + \varepsilon_{it} \geq 0\} i = 1, \ldots, N; \ t = 1, \ldots, T$$

that is, when the conditioning set also includes exogenous variables. Honoré and Kyriazidou (2000a) show that β_0 and γ_0 in the model above are in fact identified both for the case when the errors ε_{it} are logistic and when they are only assumed to have the same distribution over time conditional on (x_i, y_{i0}) (see below). In the logistic case identification is based on the fact that the following probabilities

$$\Pr(A|A \cup B, x_{i2} = x_{i3}, x_i, \alpha_i) = \frac{1}{1 + \exp((x_{i1} - x_{i2})\beta_0 + \gamma_0(d_0 - d_3))} \Pr(B|A \cup B, x_{i2} = x_{i3}, x_i, \alpha_i) = \frac{\exp((x_{i1} - x_{i2})\beta_0 + \gamma_0(d_0 - d_3)}{1 + \exp((x_{i1} - x_{i2})\beta_0 + \gamma_0(d_0 - d_3))}$$

are independent of α_i. Note that the probabilities above condition not only on the individual switching states in the middle two periods so that $y_{i1} + y_{i2} = 1$ but also on the event that $x_{i2} = x_{i3}$. Honoré and Kyriazidou (2000a) propose estimating β_0 and γ_0 by maximizing

$$\sum_i 1\{x_{i2} - x_{i3} = 0\} 1\{y_{i1} + y_{i2} = 1\} \times \ln\left(\frac{\exp((x_{i1} - x_{i2})\beta + \gamma(y_{i0} - y_{i3}))^{y_{i1}}}{1 + \exp((x_{i1} - x_{i2})\beta + \gamma(y_{i0} - y_{i3}))}\right)$$

when $\Pr(x_{i2} = x_{i3}) > 0$. When $x_{i2} - x_{i3}$ is continuously distributed with support around 0, β_0 and γ_0 can be obtained by maximizing

$$\sum_i K\left(\frac{x_{i2} - x_{i3}}{h_N}\right) 1\left\{y_{i1} + y_{i2} = 1\right\} \times \ln\left(\frac{\exp((x_{i1} - x_{i2})\beta + \gamma(y_{i0} - y_{i3}))^{y_{i1}}}{1 + \exp((x_{i1} - x_{i2})\beta + \gamma(y_{i0} - y_{i3}))}\right)$$

where *K* () is a *kernel density function* and h_N is a *bandwidth sequence*, chosen so as to satisfy certain assumptions that guarantee consistency and asymptotic normality of the proposed estimators.

3 The fixed effects approach

The conditional maximum likelihood approach is not always available. For example, there are no sufficient statistics for the binary choice model with individual effects when the errors are normally distributed. Furthermore, like all ML approaches, the approach suffers from the fact that the distribution of the unobserved idiosyncratic errors needs to be parametrically specified. There do exist, however, methods for some *semiparametric* nonlinear panel data models with individual effects where the distribution of the underlying idiosyncratic errors is left unspecified. These include the binary choice model, the censored and truncated regression models, and the sample selection model.

3.1 The semiparametric panel data binary choice model

Manski (1987) considers the model

$$y_{it} = 1\{x_{it}\beta_0 + \alpha_i - \varepsilon_{it} \geq 0\} i = 1, \ldots, N; \quad t = 1, \ldots, T$$

where ε_{it} is identically distributed over time conditional on (x_i, α_i), with distribution function F that is a continuous and strictly increasing function on $\Re$. Note that, in contrast to the models considered above, F here is not assumed to have a specific functional form, hence the characterization of the model as *semiparametric*.
He observes that for T=2 the time invariance of F implies that

$$\Pr(y_{i1} = 1|x_i) \square \Pr(y_{i2} = 1|x_i) \text{if and only if } x_{i1}\beta_0 \square x_{i2}\beta_0$$

or equivalently that

$$sgn(\Pr(y_{i2} = 1|x_i, \alpha_i) - \Pr(y_{i1} = 1|x_i, \alpha_i)) = sgn((x_{i2} - x_{i1})\beta_0)).$$

In fact it can be shown that, under appropriate regularity conditions on the joint distribution of $\Delta x_i \equiv (x_{i2} - x_{i1})$, β_0 uniquely (up to scale) maximizes the so-called population ‘score function’

$$E[\Delta y_i \cdot sgn(\Delta x_i\beta_0)]$$

where $sgn(x)$ equals 1 if $x > 0$, equals −1 if x<0 and is equal to 0 if $x = 0$. This suggests estimating β_0 by the so-called *conditional maximum score estimator* which maximizes the sample analog of the population score function

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \sum_i \Delta y_i \cdot sgn(\Delta x_i\beta).$$

Note that only observations for which $y_{i1} \neq y_{i2}$ are used here, similarly to conditional logit. The estimator is consistent under some additional assumptions but is not asymptotically normal and its rate of convergence is not root-N.
Honoré and Kyriazidou (2000a) show that it is possible to extend the conditional maximum score approach to the dynamic binary choice model:

$$\Pr(y_{i0} = 1|x_i, \alpha_i) = p_0(x_i, \alpha_i)\Pr(y_{it} = 1|x_i, \alpha_i, y_{i0}, \ldots, y_{it-1}) = F(x_{it}\beta_0 + \gamma_0 y_{it-1} + \alpha_i) t = 1, \ldots, T$$

where y_{i0} is assumed to be observed and F is strictly increasing.
We will next demonstrate their identification scheme. Assume $T = 3$ and define the events A and B as above. Then

$$\Pr(A|x_i, \alpha_i, x_{i2} = x_{i3}) = p_0(x_i, \alpha_i)^{d_0}(1 - p_0(x_i, \alpha_i))^{1-d_0} \times (1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i)) \times F(x_{i2}\beta_0 + \alpha_i) \times (1 - F(x_{i2}\beta_0 + \gamma_0 + \alpha_i))^{(1-d_3)} \times F(x_{i2}\beta_0 + \gamma_0 + \alpha_i)^{d_3} \Pr(B|x_i, \alpha_i, x_{i2} = x_{i3}) = p_0(x_i, \alpha_i)^{d_0}(1 - p_0(x_i, \alpha_i))^{1-d_0} \times F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i) \times (1 - F(x_{i2}\beta_0 + \gamma_0 + \alpha_i)) \times (1 - F(x_{i2}\beta_0 + \alpha_i))^{(1-d_3)} \times F(x_{i2}\beta_0 + \alpha_i)^{d_3}.$$

If $d_3 = 0$, then,

$$\frac{\Pr(A|x_i, \alpha_i, x_{i2} = x_{i3})}{\Pr(B|x_i, \alpha_i, x_{i2} = x_{i3})} = \frac{(1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i))}{(1 - F(x_{i2}\beta_0 + \alpha_i))} \times \frac{F(x_{i2}\beta_0 + \alpha_i)}{F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i)} = \frac{(1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i))}{(1 - F(x_{i2}\beta_0 + \gamma_0 d_3 + \alpha_i))} \times \frac{F(x_{i2}\beta_0 + \gamma_0 + \alpha_i)}{F(x_{i2}\beta_0 + \alpha_i)}$$

while if $d_3 = 1$, then,

$$\frac{\Pr(A|x_i, \alpha_i, x_{i2} = x_{i3})}{\Pr(B|x_i, \alpha_i, x_{i2} = x_{i3})} = \frac{(1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i))}{(1 - F(x_{i2}\beta_0 + \gamma_0 + \alpha_i))} \times \frac{F(x_{i2}\beta_0 + \gamma_0 + \alpha_i)}{F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i)} = \frac{(1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i))}{(1 - F(x_{i2}\beta_0 + \gamma_0 d_3 + \alpha_i))} \times \frac{F(x_{i2}\beta_0 + \gamma_0 d_3 + \alpha_i)}{F(x_{i2}\beta_0 + \gamma_0 d_0 + \alpha_i)}.$$

Monotonicity of F implies that

$$sgn(\Pr(A|x_i, \alpha_i, x_{i2} = x_{i3}) - \Pr(B|x_i, \alpha_i, x_{i2} = x_{i3}) = sgn((x_{i2} - x_{i1})\beta_0 + \gamma_0(d_3 - d_0)).$$

This last equation suggests that β_0 and γ_0 can be estimated by conditional maximum score using only the observations satisfying $y_{i1} + y_{i2} = 1$ and $x_{i2} = x_{i3}$, that is, by maximizing

$$\sum_i 1\{x_{i2} - x_{i3} = 0\}\ (y_{i2} - y_{i1})sgn((x_{i2} - x_{i1})\beta + \gamma(y_{i3} - y_{i0})).$$

Similar to the logit case, when $x_{i2} - x_{i3}$ is continuously distributed with support around 0, estimation of β_0 and γ_0 can be obtained by maximizing

$$\sum_i K\left(\frac{x_{i2} - x_{i3}}{h_N}\right)(y_{i2} - y_{i1})sgn((x_{i2} - x_{i1})\beta + \gamma(y_{i3} - y_{i0})).$$

3.2 The semiparametric panel data censored regression model

The standard censored panel data (or Type 1 Tobit) model with individual effects is given by

$$y_{it} = \max\{x_{it}\beta_0 + \alpha_i + \varepsilon_{it}, 0\} i = 1, \ldots, N;\ t = 1, \ldots, T.$$

Estimation of this model was first considered by Honoré (1992) and later by Honoré and Kyriazidou (2000b), who extend the results of the former paper. We will present here Honoré (1992), who assumes that (ε_{it}, ε_{is}) are *pairwise exchangeable* conditional on (x_i, a_i). This implies that ε_{it} and ε_{is} are identically distributed conditional on (x_i, a_i) although it does not require (conditional) independence over time. (Fristedt and Gray, 1997, give the following definition of exchangeability: Let Φ be a countable set. A sequence $(X_i: i \in \Phi)$, finite or infinite, of random variables on a probability space $(\Omega, \mathcal{F}, P)$ is *exchangeable* if, for every permutation ρ of Φ, the distribution of $(X_{\rho(i)}: i \in \Phi)$ and $(X_i: i \in \Phi)$ are identical. Note that a finite or infinite i.i.d. sequence is exchangeable and that exchangeability allows for certain types of serial correlation. Furthermore, exchangeability implies strict stationarity although the converse is not true.)
Consider the 'pseudo-error':

$$e_{ist}(\beta) = \max\{y_{is}, (x_{is} - x_{it})\beta\} - x_{is}\beta.$$

With this definition, at the true β_0

$$e_{ist}(\beta_0) = \max\{y_{is}, (x_{is} - x_{it})\beta_0\} - x_{is}\beta_0 = \max\{\max\{x_{is}\beta_0 + \alpha_i + \varepsilon_{is}, 0\}, (x_{is} - x_{it})\beta_0\} - x_{is}\beta_0 = \max\{\max\{\alpha_i + \varepsilon_{is}, -x_{is}\beta_0\}, -x_{it}\beta_0\} = \max\{\alpha_i + \varepsilon_{is}, -x_{is}\beta_0, -x_{it}\beta_0\}$$

The conditional exchangeability assumption implies that $(e_{ist}(\beta_0), e_{its}(\beta_0))$ is distributed like $(e_{its}(\beta_0), e_{ist}(\beta_0))$ conditional on (x_{it}, x_{is}, a_i) and hence the difference $e_{its}(\beta_0) - e_{ist}(\beta_0)$ is distributed symmetrically around 0 conditional on (x_{it}, x_{is}, a_i). Since this is true for any α_i this symmetry holds conditional only on (x_{it}, x_{is}).
Therefore for any odd function ξ (that is, a function ξ that satisfies $\xi(-d) = -\xi(d)$) we have

$$E[\xi(e_{ist}(\beta_0) - e_{its}(\beta_0))|x_{it}, x_{is}] = 0 \quad (3.1)$$

which also implies the following moment restriction:

$$E[\xi(e_{ist}(\beta_0) - e_{its}(\beta_0))(x_{is} - x_{it})'|x_{it}, x_{is}] = 0.$$

The left-hand side of the moment condition above may be thought of as the first order condition for the following population minimization problem

$$\min_\beta E[q(y_{is}, y_{it}, (x_{is} - x_{it})\beta)|x_{it}, x_{is}]$$

where

$$q(y_i, y_j, \delta) = \begin{cases} \Xi(y_i) - (y_j + \delta)\xi(y_i) & \text{if} \quad \delta \le -y_j \\ \Xi(y_i - y_j - \delta) & \text{if} \quad -y_j < \delta < y_i \\ \Xi(-y_j) - (\delta - y_i)\xi(-y_i) & \text{if} \quad y_i \le \delta \end{cases}$$

and $\Xi(d): \Re \to \Re^+$ is an even function (that is, $\Xi(-d) = \Xi(d)$) which is convex, strictly increasing for $d > 0$ and has $\Xi(0) = 0$, and $\Xi'(d) = \xi(d)$ where $\xi(0) = 0$. Note that for Ξ to be convex, ξ has to be monotone. Obvious choices for Ξ are $\Xi(d) = d^2$ (which corresponds to $\xi(d) = 2d$) and $\Xi(d) = |d|$ (which corresponds to $\xi(d) = sgn(d)$).
The fact that the true β_0 solves the population minimization problem above suggests the following estimator for β_0:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i \sum_{s<t} q(y_{is}, y_{it}(x_{is} - x_{it})\beta).$$

Honoré (1992) shows that the estimators corresponding to $\Xi(d) = d^2$ and $\Xi(d) = |d|$ are root-*N* consistent and asymptotically normal.

Honoré (1993) considers a dynamic version of the model where the lag of the *observed* (censored) dependent variable appears in the model instead of the latent one. Hu (2002) considers the case where one lag of the *latent* (unobserved) dependent variable is included along with the set of exogenous variables x_{it}.

3.3 The semiparametric panel data sample selection model

The standard panel data sample selection (or Type 2 Tobit) model is defined as:

$$y_{it}^* = x_{it}^*\beta_0 + \alpha_i^* + \varepsilon_{it}^* y_{it} = d_{it} \cdot y_{it}^* d_{it} = 1\left\{z_{it}\gamma_0 + \eta_i - u_{it} \geq 0\right\}$$

where $i = 1, 2, \ldots, N$; $t = 1, \ldots, T$. Kyriazidou (1997) considers estimation without any parametric assumptions on the form of the joint distribution of $(\varepsilon_{it}^*, u_{it})$ or on the individual effects (α_i, η_i).
Consider the case where $T = 2$ and only those individuals for whom $d_{i1} = d_{i2} = 1$. Let $\xi_i = (z_{i1}, z_{i2}, x_{i1}^*, x_{i2}^*, \alpha_i^*, \eta_i)$ denote all the information about individual *i*. Note that

$$E(y_{i1} - y_{i2} | d_{i1} = d_{i2} = 1, \xi_i) = (x_{i1}^* - x_{i2}^*)\beta_0 + E(\varepsilon_{i1}^* - \varepsilon_{i2}^* | d_{i1} = d_{i2} = 1, \xi_i)$$

and hence OLS estimation of the first differenced model will not yield consistent estimation of β_0 since in general the so-called 'sample selection bias term'

$$\lambda_{it} \equiv E(\varepsilon_{it}^* | d_{i1} = d_{i2} = 1, \xi_i) = E(\varepsilon_{it}^* | u_{i1} \leq z_{i1}\gamma_0 + \eta_i, u_{i2} \leq z_{i2}\gamma_0 + \eta_i, \xi_i)$$

is not zero. Nor do we have in general that $\lambda_{i1} = \lambda_{i2}$, so that first differencing removes the sample selection bias along with the individual effects. Kyriazidou (1997) makes a *conditional exchangeability assumption* that $(\varepsilon_{i1}^*, \varepsilon_{i2}^*, u_{i1}, u_{i2})$ and $(\varepsilon_{i2}^*, \varepsilon_{i1}^*, u_{i2}, u_{i1})$ are identically distributed conditional on ξ_i. Under this assumption, it is easy to see that if $z_{i1}\gamma_0 = z_{i2}\gamma_0$ then

$$\lambda_{i1} = E(\varepsilon_{i1}^* | u_{i1} \leq z_{i1}\gamma_0 + \eta_i, u_{i2} \leq z_{i2}\gamma_0 + \eta_i, \xi_i) = E(\varepsilon_{i2}^* | u_{i1} \leq z_{i1}\gamma_0 + \eta_i, u_{i2} \leq z_{i2}\gamma_0 + \eta_i, \xi_i) = \lambda_{i2}$$

so that first differencing will eliminate both α_i and λ_{it} simultaneously. So β_0 can be estimated by first difference OLS for the subsample of individuals that are observed in both periods (that is, that have $d_{i1} = d_{i2} = 1$) and also have the selection index, $z_{it}\gamma_0$, constant (that is, $z_{i1}\gamma_0 = z_{i2}\gamma_0$). Of course, this estimation scheme cannot be directly implemented since γ_0 is unknown. And it is quite possible that no observation has $z_{i1}\gamma_0 = z_{i2}\gamma_0$ if $z_i\gamma_0$ is continuously distributed. If, however, λ_{it} is a sufficiently smooth function and $\hat{\gamma}$ is a consistent estimator of γ_0, $z_{i1}\gamma_0 \approx z_{i2}\gamma_0$ implies $\lambda_{i1} \approx \lambda_{i2}$, and the preceding augment holds approximately. Kyriazidou proposes a two-step estimation procedure, in the spirit of Powell (2001), and Ahn and Powell (1993) who consider estimation of cross-section versions of the sample selection model. In the first step, γ_0 is consistently estimated based on the selection equation. In the second step, the estimate $\hat{\gamma}$ is used to estimate β_0 based on those pairs of observations for which $z_{i1}\hat{\gamma}$ and $z_{i2}\hat{\gamma}$ are 'close'. To this end define

$$\hat{\psi}_i = \frac{1}{h_N} K\left(\frac{\Delta z_i \hat{\gamma}}{h_N}\right)$$

where $K()$ is a kernel density function and h_N is a bandwidth sequence. The proposed estimator takes the form:

$$\hat{\beta} = \left[\sum_{i=1}^{N} \hat{\psi}_i \Delta x_i' \Delta x_i d_{i1} d_{i2}\right]^{-1} \sum_{i=1}^{N} \hat{\psi}_i \Delta x_i' \Delta y_i d_{i1} d_{i2}.$$

Under some assumptions and by appropriately choosing h_N, the estimator can be shown to be asymptotically normal although the rate of convergence is slower that the parametric $\sqrt{N}$ rate. Apart from the conditional exchangeability assumption, another important assumption that underlies the approach is that there is at least one variable in z_{it} not contained in x_{it}, which is an exclusion restriction common in semiparametric sample selection models.
A dynamic version of the panel data sample selection model, with the own lagged dependent variable appearing in each equation, is considered by Kyriazidou (2001).

4 The random effects approach

Fixed effects methods and conditional maximum likelihood methods (when they exist) estimate the coefficients of time-varying regressors consistently without making any assumptions on how the individual effects are related to the observed covariates or to the time-varying errors or to the initial observations of the sample. However, these methods do not deliver estimates of coefficients of time-invariant regressors and of the individual effects, and hence cannot be used for prediction, or for computation of marginal effects and elasticities which are often the quantities of interest. Furthermore, none of these approaches allows for non-stationary errors and hence for time-series heteroskedasticity.
These problems do not arise in the random effects approach. The approach essentially consists of treating $(\alpha_i + \varepsilon_{it})$ as a two-component error term and making assumptions about its relationship with the observed covariates and, in the case of dynamic models, with the initial conditions as well. A downside of the approach is that misspecification of any part of the model typically yields inconsistent estimates.

4.1 Static case

In the static panel data linear regression model, the traditional random effects approach (sometimes also called the *uncorrelated random effects approach*) assumes that the individual effects α_i along with the time-varying errors ε_{it} are uncorrelated with the observed covariates x_{it}. Then the coefficients of both time-varying and time-invariant regressors may be estimated consistently (albeit not efficiently) by pooled OLS. In static nonlinear models, the traditional random effects approach apart from parameterizing the conditional distribution of ε_{it} given x_{it}, also assumes that α_i is independent of x_{it} and ε_{it} for all *t*, and has a distribution, say *H*, that depends on a finite set of unknown parameters, say δ_0. For example, in the binary choice model,

$$y_{it} = 1\{x_{it}\beta_0 + \alpha_i + \varepsilon_{it} \geq 0\} i = 1, \ldots, N; \quad t = 1, \ldots, T \tag{4.1}$$

assuming that ε_{it} are i.i.d. over time and independent of x_i and α_i with known distribution F (say, standard normal or logistic), we may estimate the unknown parameters (β_0, δ_0) via ML. The log-likelihood is

$$\ln L(\beta, \delta) = \sum_i \ln \int \prod_{t=1}^{T} F(x_{it}\beta + \alpha)^{y_{it}} (1 - F(x_{it}\beta + \alpha))^{1 - y_{it}} dH(\alpha, \delta)$$

and involves a one-dimensional integral which may be calculated numerically, for example, by *quadrature procedures* (see Butler and Moffitt, 1982).
However, things become quite complicated if we want to allow for arbitrary serial correlation in the ε_{it}'s. Consider the binary choice model

$$y_{it} = 1\{x_{it}\beta_0 - u_{it} \geq 0\}$$

where $u_{it} = \alpha_i + \varepsilon_{it}$ is the composite error term. For $T = 3$ there are 2^3 possible sequences of 0's and 1's. The likelihood for an individual for whom the sequence of observed y_{it}'s is (0,1,0) takes the form

$$\int_{x_{i1}\beta} \int^{x_{i2}\beta} \int_{x_{i3}\beta} f(u_1, u_2, u_3)\, du_1 du_2 du_3$$

where f is the trivariate density of (u_1,u_2,u_3) conditional on x_i. The log-likelihood is

$$\ln L(\beta, \delta) = \sum_{i:(0,0,0)} \ln \int_{x_{i1}\beta} \int_{x_{i2}\beta} \int_{x_{i3}\beta} f(u_1, u_2, u_3)\, du_1 du_2 du_3 + \sum_{i:(0,0,1)} \ln \int_{x_{i1}\beta} \int_{x_{i2}\beta} \int^{x_{i3}\beta} f(u_1, u_2, u_3)\, du_1 du_2 du_3 + \ldots$$

which requires the computation of multiple trivariate integrals. Multivariate integration is basically infeasible for large T. This is where simulation methods come in very handy.
The assumption that α_i is independent of x_i is often found unsatisfactory. A possible solution is to assume a specific functional form for the relationship of α_i with x_i. This approach (recently also called the *correlated random effects approach*) was first proposed by Chamberlain (1984). Suppose that

$$\alpha_i = \sum_{t=1}^{T} x_{it}\gamma_{0,t} + v_i$$

where v_i is independent of x_i, similarly to the time varying error component ε_{it}, and that the composite new error term $v_i + \varepsilon_{it}$ follows a specific distribution, say normal. In the case of the binary choice model, for example, assuming that $\varepsilon_{it} + v_i | x_i$, α_i is $N(0, \sigma_{0,t}^2)$ implies that

$$\Pr(y_{it} = 1 | x_i) = \Phi\left(\frac{x_{it}\beta_0 + \Sigma_{t=1}^{T} x_{it}\gamma_{0,t}}{\sigma_{0,t}}\right) = \Phi(x_{it}\theta_{0,t}).$$

For computational simplicity, Chamberlain proposes to estimate the unknown parameters $\theta_{0,t}$ via period-by-period probit. The 'structural parameters' β_0, $\{\sigma_{0,t}^2\}_{t=1}^{T}$, and $\{\gamma_{0,t}\}_{t=1}^{T}$ can then be recovered by *minimum distance* estimation. Note that the approach allows for time series heteroskedasticity and requires only one normalization e.g. that $\sigma_{0,t}^2 = 1$.
Newey (1994) generalizes Chamberlain's approach by postulating that

$$\alpha_i = \rho(x_{i1}, \ldots, x_{it}) + v_i$$

where ρ () is an unknown function of x_i. Assuming again that v_i and ε_{it} are independent of x_i and that the composite new error term $v_i + \varepsilon_{it}$ follows a specific distribution, say F_t, we obtain

$$\pi_t = \Pr(y_{it} = 1 | x_i) = F_t(\rho(x_i) + x_{it}\beta_0)$$

which for a strictly monotonic F_t implies that

$$F_t^{-1}(\pi_t) = \rho(x_i) + x_{it}\beta_0.$$

For example in the normal case

$$\Phi^{-1}(\pi_t) = \frac{\rho(x_i) + x_{it}\beta_0}{\sigma_{0,t}}.$$

Thus for two periods *t* and *s* we obtain

$$\Phi^{-1}(\pi_t) = \frac{\sigma_{0,s}}{\sigma_{0,t}}\Phi^{-1}(\pi_s) + \frac{\sigma_{0,s}}{\sigma_{0,t}}(x_{it} - x_{is})\beta_0.$$

Normalizing $\sigma_{0,t}=1$ and estimating π_t and π_s nonparametrically, we can recover $\sigma_{0,s}$ and β_0 from the regression of $\Phi^{-1}(\hat{\pi}_t)$ on $\Phi^{-1}(\hat{\pi}_s)$ and $(x_{it} - x_{is})$.

A criticism of all these correlated random effects approaches is that, although in the linear model writing $\alpha_i = \Sigma_{t=1}^{T} x_{it}\gamma_{0,t} + u_i$ where $E(u_i x_{it}) = 0$ for all *t* does not impose $x_{it} - x_{is}$ any restrictions on the joint distribution of α_i and x_i (apart from the requirement that it has second moments) since this is just the best linear projection of α_i on x_i, in the nonlinear model assuming $\alpha_i = \rho(x_{i1}, \ldots, x_{iT}) + u_i$, even without specifying the functional form of ρ, imposes implausible restrictions in the sense that, if this relationship holds for the *T* observations, a similar one will *not* in general hold for $T + 1$.

4.2 Dynamic case

In the case where there are genuine dynamics in the model in the form of lags of the dependent variable or other endogenous regressors, random effects methods become even more complicated and require additional assumptions about the relationship of the individual effects with the initial observations. We next describe a general approach for estimating dynamic random effects models suggested by Wooldridge (2000). For simplicity we will drop the subscripts *i*.

We are interested in the conditional distribution of y_t given a vector of strictly exogenous variables $z^T \equiv (z_1, \ldots, z_T)$, own lags and lags of other endogenous variables $x^{t-1} \equiv (y_{t-1}, w_{t-1}, y_{t-2}, w_{t-2}, \ldots, y_0, w_0)$, and an unobserved scalar or vector random effect α. Here z_t is strictly exogenous in the sense that

$$F(w_t|z^T, x^{t-1}, \alpha) = F(w_t|z_t, x^{t-1}, \alpha).$$

The conditional density of $x_t \equiv (y_t, w_t)$ is

$$f_t(x_t|z^T, x^{t-1}, \alpha) = f_t(x_t|z_t, x^{t-1}, \alpha) = f_t(y_t|w_t, z_t, x^{t-1}, \alpha) \cdot f_t(w_t|z_t, x^{t-1}, \alpha)$$

and the joint density for all *T* periods is

$$f(x_1, x_2, \ldots, x_T|z^T, x_0, \alpha) = \prod_{t=1}^{T} f_t(x_t|z_t, x^{t-1}, \alpha).$$

But *a* is unobserved. We need to integrate it out. One solution is to parameterize the distribution of α conditional on z^T and x_0, say $h(\alpha|z^T, x_0)$. Then

$$f(x_1, x_2, \ldots, x_T|z^T, x_0) = \int \prod_{t=1}^{T} f_t(x_t|z_t, x^{t-1}, \alpha)\, h(\alpha|z^T, x_0)\, d\alpha.$$

Notice that in the traditional random effects approach (in the line of Anderson and Hsiao, 1981) we would have to make assumptions about the conditional distribution of x_0 conditional on *a* and z^T.

See Also

- fixed effects and random effects
- maximum likelihood

Bibliography

Ahn, H. and Powell, J.L. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58, 3–29.

Andersen, E. 1970. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society*, Series B 32, 283–301.

Anderson, T. and Hsiao, C. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76(375), 598–606.

Butler, J.S. and Moffitt, R. 1982. A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica* 50, 761–4.

Chamberlain, G. 1984. Panel data. In *Handbook of Econometrics*, vol 2, ed. Z. Griliches and M. Intrilligator. Amsterdam: North-Holland.

Chamberlain, G. 1985. Heterogeneity, omitted variable bias, and duration dependence. In *Longitudinal Analysis of Labor Market Data*, ed. J.J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Fristedt, B. and Gray, L. 1997. *A Modern Approach to Probability Theory*. Boston: Birkhauser.

Honoré, B.E. 1992. Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60, 533–65.

Honoré, B.E. 1993. Orthogonality conditions for Tobit models with fixed effects and lagged dependent variables. *Journal of Econometrics* 59, 35–61.

Honoré, B.E. and Kyriazidou, E. 2000a. Panel data discrete choice models with lagged dependent variables. *Econometrica* 68, 839–74.

Honoré, B.E. and Kyriazidou, E. 2000b. Estimation of Tobit-type models with individual specific effects. *Econometric Reviews* 19, 341–66.

Hu, L. 2002. Estimation of a censored dynamic panel data model. *Econometrica* 70, 2499–517.

Kyriazidou, E. 1997. Estimation of a panel data sample selection model. *Econometrica* 65, 1335–64.

Kyriazidou, E. 2001. Estimation of dynamic panel data sample selection models. *Review of Economic Studies* 68, 543–72.

Magnac, T. 2000. Subsidised training and youth employment: distinguishing unobserved heterogeneity from state dependence in labour market histories. *Economic Journal* 110, 805–37.

Manski, C. 1987. Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* 55, 357–62.

Newey, W. 1994. The asymptotic variance of semiparametric estimators. *Econometrica* 62, 1349–82.

Neyman, J. and Scott, E.L. 1948. Consistent estimation from partially consistent observations. *Econometrica* 16, 1–32.

Powell, J.L. 2001. Semiparametric estimation of bivariate latent variable models. In *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune and J.L. Powell. Cambridge: Cambridge University Press.

Wooldridge, J.M. 2000. A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. *Economics Letters* 68, 245–50.

How to cite this article

Kyriazidou, Ekaterini. "nonlinear panel data models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000141> doi:10.1057/9780230226203.1194

The New Palgrave Dictionary of Economics Online

nonlinear programming

Michael D. Intriligator

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

activity analysis; concave programming; demand functions; dual problem; firm, theory of the; hyperplanes; Kuhn–Tucker conditions; Lagrange multipliers; linear programming; mathematical programming; nonlinear programming; quadratic programming; saddlepoints; Slater constraint qualification

Article

The problem of *nonlinear programming* is that of maximizing (or minimizing) a given function subject to a set of inequality constraints. Such problems arise in many areas of economics, such as the microeconomic theory of the household and the firm. It has also had wide applicability in game theory and operations research. Historically, the subject developed from the work of mathematicians, primarily John in studying extremum problems with inequalities as side constraints and Kuhn and Tucker who made the fundamental contribution of characterizing the nature of the solution to such problems (John, 1948; Kuhn and Tucker, 1951).

The nonlinear programming problem is a special case of the general *mathematical programming problem* of maximizing a function subject to constraints. The *linear programming problem* can be considered a special case of the nonlinear programming problem, namely one of maximizing a given linear form subject to a set of linear inequality constraints.

1 Mathematical programming: resource allocation in economics

The more general *problem of mathematical programming* is that of maximizing a function subject to constraints. Using standard notation the problem can be written

$$\max_{\mathbf{x}} F(\mathbf{x}) \text{ subject to } \mathbf{x} \in X.$$

(1)

Here **x** is a (column) vector of *n* choice variables $(x_1, x_2, \ldots, x_n)'$ (the prime denoting the transpose of the row vector); *F*(**x**) is a given real-valued function of these variables $F(x_1, x_2, \ldots, x_n)$; and *X* is a given non-empty subset of Euclidean *n*-space (the space of all *n*-tuples of real numbers) (Hadley, 1964; Intriligator, 1971; 1981; Aoki, 1971; Luenberger, 1973; Hestenes, 1975).

In economics the vector **x** is frequently called the vector of *instruments*, the function *F*(**x**) is frequently called the *objective function* (or *criterion function*), and the set *X* of feasible instrument vectors $(\mathbf{x} \text{ satisfying } \mathbf{x} \in X)$ is frequently called the *opportunity set.* The basic economic problem of allocating scarce resources among competing ends can thus be interpreted as one of mathematical programming, where a particular resource allocation is represented by the choice of a particular vector of instruments, the scarcity of the resources is represented by the opportunity set, reflecting constraints on the instruments; and the competing ends are represented by the objective function, which gives the value attached to each of the alternative allocations. Problem (1) can therefore be interpreted in the language of economics as that of choosing instruments within the opportunity set so as to maximize the objective function (Lancaster, 1968; Intriligator, 1971; 1981; Bazaraa and Shetty, 1976; Takayama, 1985).

2 Nonlinear programming

The *problem of nonlinear programming*, a special case of (1), is that of choosing non-negative values of *n* variables so as to maximize a function of these variables subject to inequality constraints. Using the same type of notation the problem is

$$\max_{\mathbf{x}} F(\mathbf{x}) \text{ subject to } \mathbf{g}(\mathbf{x}) \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}.$$

(2)

Here the vector of instruments **x** and the objective function *F*(**x**) are as in (1), where *F*(**x**) is assumed to be a real-valued continuously differentiable function of *n* variables. The vector-valued function **g**(**x**) is a representation of *m* constraint functions, $[g_1(x_1, x_2, \ldots, x_n), g_2(x_1, x_2, \ldots, x_n), \ldots, g_m(x_1, x_2, \ldots, x_n)]'$ and *b* is a column vector of *m* constraint constants $(b_1, b_2, \ldots, b_m)'$, so the *m* inequality constraints in (2) can be written

$$g_i(x_1, x_2, \ldots, x_n) \leq b_i, \; i = 1, 2, \ldots, m.$$

(3)

The *n* non-negativity constraints in (2) state that all *n* instruments are non-negative. Thus the problem of nonlinear programming can be written

$$\max F(x_1, x_2, \ldots, x_n) \quad \text{by choice of} \quad x_1, x_2, \ldots, x_n \tag{4}$$

$$\text{subject to} \begin{cases} g_i(x_1, x_2, \ldots, x_n) \leq b_i, & i = 1, 2, \ldots, m \\ x_j \geq 0, & j = 1, 2, \ldots, n \end{cases}$$

This problem is a special case of (1) in which the opportunity set can be written

$$X = \{\mathbf{x} \in E^n | \mathbf{g}(\mathbf{x}) \leq \mathbf{b}, \ \mathbf{x} \geq \mathbf{0}\} \tag{5}$$

where E^n is Euclidean *n*-space. Thus the problem is one of maximizing a given function subject to *m*+*n* constraints – m inequality constraints and *n* non-negativity constraints (Kuhn and Tucker, 1951; Hadley, 1964; Mangasarian, 1969; Zangwill, 1969; Intriligator, 1971; 1981; Luenberger, 1973; Hestenes, 1975; Martos, 1975; Avriel, 1976; Bazaraa and Shetty, 1979; McCormick, 1983).

3 Linear programming

In spite of the contradictory terminology, the *problem of linear programming* is in fact an important special case of nonlinear programming. Here the problem is that of choosing non-negative values of *n* variables so as to maximize a linear from in these variables subject to *m* linear inequality constraints

$$\max_{\mathbf{x}} \mathbf{cx} \text{subject to} \mathbf{Ax} \leq \mathbf{b}, \ \mathbf{x} \geq 0, \tag{6}$$

where **A** is a given *m* x *n* matrix, **b** is a given *m* × 1 column vector, and **c** is a given 1 × *n* row vector. Thus the linear programming problem represents the special case of nonlinear programming (2) which is doubly linear, being linear in the objective function $\mathbf{cx} = \sum_{j=1}^{n} c_j x_j$ and in each of the constraint functions $g_i(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} a_{ij} x_j$. Thus the problem of linear programming can be written

$$\max \sum_{j=1}^{n} c_j x_j \text{ by choice of } x_1, x_2, \ldots, x_n$$

$$\text{subject to} \begin{cases} \sum_{j=1}^{n} a_{ij} x_j \le b_i, & i = 1, 2, \ldots, m \\ x_j \ge 0, & j = 1, 2, \ldots, n. \end{cases} \tag{7}$$

(Dorfman, Samuelson and Solow, 1958; Gale, 1960; Hadley, 1963; Dantzig, 1963; Intriligator, 1971; Luenberger, 1973; Gass, 1975.)

4 Kuhn – Tucker conditions

The Kuhn–Tucker conditions provide a characterization of the solution to the problem of nonlinear programming (2). These conditions are defined in terms of the *Lagrangian function*

$$L(\mathbf{x}, \mathbf{y}) = F(\mathbf{x}) + \mathbf{y}[\mathbf{b} - \mathbf{g}(\mathbf{x})] = F(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{m} y_i [b_i - g_i(x_1, x_2, \ldots, x_n)]. \tag{8}$$

Here **y** is a (row) vector of *m Lagrange multipliers* (sometimes written as λ's), one for each of the inequality constraints defined by the $g_i(x_1, x_2, \ldots, x_n)$ constraint functions and b_i constraint constants in (3). The *Kuhn–Tucker conditions* are then defined at the point $\mathbf{x}^*$, $\mathbf{y}^*$ as the 2*n*+2•*m* inequalities and 2

equalities

$$\frac{\partial L}{\partial \mathbf{x}}\left(\mathbf{x}^*, \mathbf{y}^*\right) \leq \mathbf{0}, \qquad \frac{\partial L}{\partial \mathbf{y}}\left(\mathbf{x}^*, \mathbf{y}^*\right) \geq \mathbf{0} \quad (n + m \text{ conditions})$$
$$\frac{\partial L}{\partial \mathbf{x}}\left(\mathbf{x}^*, \mathbf{y}^*\right)\mathbf{x}^* = 0, \quad \mathbf{y}^* \frac{\partial L}{\partial \mathbf{y}}\left(\mathbf{x}^*, \mathbf{y}^*\right) = 0 \qquad (2 \text{ conditions})$$
$$\mathbf{x}^* \geq \mathbf{0}, \qquad \mathbf{y}^* \geq \mathbf{0} \quad (n + m \text{ conditions}) .$$
(9)

Half of the inequalities represent the constraints of the original problem

$$\frac{\partial L}{\partial \mathbf{y}}\left(\mathbf{x}^*, \mathbf{y}^*\right) = \mathbf{b} - \mathbf{g}(\mathbf{x}^*) \geq \mathbf{0} (m \text{ conditions})$$
(10)

$$\mathbf{x}^* \geq \mathbf{0} \quad (n \text{ conditions}),$$
(11)

while the added *n*+*m* inequalities require that

$$\frac{\partial L}{\partial \mathbf{x}}\left(\mathbf{x}^*, \mathbf{y}^*\right) = \frac{\partial F}{\partial \mathbf{x}}(\mathbf{x}^*) - \mathbf{y}^* \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}^*) \leq \mathbf{0} (n \text{ conditions})$$
(12)

$$\mathbf{y}^* \geq \mathbf{0} (m \text{ conditions}) .$$
(13)

The first *n* conditions, in (12), state that

$$\frac{\partial F}{\partial x_j} - \sum_{i=1}^{m} y_i^* \frac{\partial g_i}{\partial x_j} \leq 0, \quad j = 1, 2, \ldots, n, \tag{14}$$

and they are written as inequalities because of the non-negativity restrictions on **x**, which allow for the possibility of boundary solutions. The last *m* conditions, in (13), state that the Lagrange multipliers are non-negative, and they stem from the fact that the constraints are written as inequalities rather than as equalities. (If any constraint is imposed as an equality, then the corresponding Lagrange multiplier y_i^* is unrestricted.)
The two equality Kuhn–Tucker conditions

$$\frac{\partial L}{\partial \mathbf{x}}\left(\mathbf{x}^*, \mathbf{y}^*\right)\mathbf{x}^* = \sum_{j=1}^{n}\left[\frac{\partial F}{\partial x_j}\left(\mathbf{x}^*\right) - \mathbf{y}^* \frac{\partial \mathbf{g}}{\partial x_j}\left(\mathbf{x}^*\right)\right] x_j^* = 0 \tag{15}$$

$$\mathbf{y}^* \frac{\partial L}{\partial \mathbf{y}}\left(\mathbf{x}^*, \mathbf{y}^*\right) = \sum_{i=1}^{m} y_i^*\left[b_i - g_i\left(\mathbf{x}^*\right)\right] = 0 \tag{16}$$

together with the other conditions in (9) require that every term in both of these sums vanishes. These *complementary slackness conditions of nonlinear programming* require that when one of the inequality constraints is satisfied at the solution as a strict inequality then the corresponding (dual) variable equals zero at the solution

$$\frac{\partial F}{\partial x_j}\left(\mathbf{x}^*\right) - \mathbf{y}^* \frac{\partial \mathbf{g}}{\partial x_j}\left(\mathbf{x}^*\right) < 0 \text{implies } x_j^* = 0, \quad j = 1, 2, \ldots, n$$

(17)

$$b_i - g_i(\mathbf{x}^*) > 0, \text{ i.e. } g_i(\mathbf{x}^*) < b_i, \text{implies } y_i^* = 0, \quad i = 1, 2, \ldots, m.$$

(18)

At the solution the value of the Lagrangian is the maximized value of the objective function

$$L(\mathbf{x}^*, \mathbf{y}^*) = F(\mathbf{x}^*) = F^*,$$

(19)

and the solutions for the Lagrange multipliers are to be interpreted as the sensitivities of the maximized value of the objective function to changes in the constraint constants

$$\mathbf{y}^* = \frac{\partial F^*}{\partial \mathbf{b}}, \text{ i.e. } y_i^* = \frac{\partial F^*}{\partial b_i}, \quad i = 1, 2, \ldots, m.$$

(20)

In particular, from the complementary slackness conditions (18), if a constraint is met as a strict inequality at the solution then the corresponding Lagrange multiplier is zero, so increasing the constraint constant by a ‘small’ amount will not change the maximized value of the objective function.
If a suitably strong *constraint qualification condition* is satisfied the Kuhn–Tucker conditions are necessary conditions for the nonlinear programming problem in that if **x*** solves (2) then there exists a vector of Lagrange multipliers $\mathbf{y}^*$ satisfying (9). There are, in fact, many alternative forms of the constraint qualification condition. One is the *Slater constraint qualification* requiring that there exist a point $\mathbf{x}^0 \geq \mathbf{0}$ such that $\mathbf{g}(\mathbf{x}^0) < \mathbf{b}$, that is, there exists a non-negative point at which all inequality constraints are satisfied as strict inequalities, thus excluding outward pointing cusps (Arrow, Hurwicz and Uzawa, 1958; 1961; Mangasarian, 1969; Zangwill, 1969; Bazaraa, Goode, and Shetty, 1972; Bazaraa and Shetty, 1976; 1979). For problems not satisfying the constraint qualification condition it is necessary to add another Lagrange multiplier y_0, for the objective function (John, 1948).

As to sufficiency, a sufficient condition for $\mathbf{x}^*$ to solve the nonlinear programming problem (2) is that there exists a $\mathbf{y}^*$ such that $\mathbf{x}^*$, $\mathbf{y}^*$ solves the *saddle point problem*

$$\max_{\mathbf{x}} \min_{\mathbf{y}} L(\mathbf{x}, \mathbf{y}) \text{subject to} \mathbf{x} \geq \mathbf{0}, \mathbf{y} \geq \mathbf{0}, \tag{21}$$

where $\mathbf{x}^*$, $\mathbf{y}^*$ solves this problem if and only if, for all $\mathbf{x} \geq \mathbf{0}$, $\mathbf{y} \geq \mathbf{0}$

$$L\left(\mathbf{x}, \mathbf{y}^*\right) \leq L\left(\mathbf{x}^*, \mathbf{y}^*\right) \leq L\left(\mathbf{x}^*, \mathbf{y}\right). \tag{22}$$

Thus, if a pair of vectors $\mathbf{x}^*$, $\mathbf{y}^*$ satisfies (22) then $\mathbf{x}^*$ solves the nonlinear programming problem. Conversely, assuming both that a suitable constraint qualification condition is met and that the problem is one of *concave programming* in which $F(\mathbf{x})$ is a concave function and each constraint function $g_i(\mathbf{x})$ is a convex function, then if $\mathbf{x}^*$ solves the nonlinear programming problem (2) there exists a non-zero vector $\mathbf{y}^*$ such that $\mathbf{x}^*$, $\mathbf{y}^*$ solves the saddle point problem (21) and the two problems are equivalent. In fact, if the problem is one of concave programming and a suitable constraint qualification condition is met, then the nonlinear programming problem (2), the problem of finding a solution to the Kuhn–Tucker conditions (9), and the saddle point problem (21) are all equivalent in that if $\mathbf{x}^*$ solves (2) then and only then there exists a $\mathbf{y}^*$ such that $\mathbf{x}^*$, $\mathbf{y}^*$ solves both (9) and (21).

Various computational approaches have been developed to solve nonlinear programming problems, and such approaches, in the form of computer codes, are widely available and routinely used to solve particular problems (Mangasarian, 1969; Zangwill, 1969; Polak, 1971; Avriel, 1976; Bazaraa and Shetty, 1979; Schittkowski, 1980; Dennis, 1984).

The Kuhn–Tucker conditions imply that for an interior solution $\mathbf{x}^* > \mathbf{0}$ (or for a problem in which the non-negativity of the *x*'s is not part of the problem)

$$\frac{\partial F}{\partial \mathbf{x}}\left(\mathbf{x}^*\right) = \mathbf{y}^* \frac{\partial \mathbf{g}}{\partial \mathbf{x}}\left(\mathbf{x}^*\right). \tag{23}$$

Thus at the solution the gradient vector of the objective function (the vector of first-order derivatives of

the objective function, $\partial F / \partial \mathbf{x}(\mathbf{x}^*)$) must be a non-negative weighted combination of the gradient vectors of the constraint functions, the weights being the Lagrange multipliers. Geometrically this condition means that the gradient vector of the objective function must, at the solution, lie within the cone spanned by the outward pointing normals to the opportunity set, where the gradient vectors for the constraint functions define the outward pointing normals to the opportunity set.
For the special case of linear programming (6) the saddlepoint problem is

$$\max_{\mathbf{x}} \max_{\mathbf{y}} L(\mathbf{x}, \mathbf{y}) = \mathbf{c}\mathbf{x} + \mathbf{y}(\mathbf{b} - \mathbf{A}\mathbf{x}) \tag{24}$$

and the Kuhn–Tucker conditions (9) are

$$\begin{gathered} \frac{\partial L}{\partial \mathbf{x}} = \mathbf{c} - \mathbf{y}^*\mathbf{A} \le \mathbf{0}, \qquad \frac{\partial L}{\partial \mathbf{y}} = \mathbf{b} - \mathbf{A}\mathbf{x}^* \ge \mathbf{0} \\ \frac{\partial L}{\partial \mathbf{x}}\mathbf{x}^* = \left(\mathbf{c} - \mathbf{y}^*\mathbf{A}\right)\mathbf{x}^* = \mathbf{0}, \quad \mathbf{y}^*\frac{\partial L}{\partial \mathbf{y}} = \mathbf{y}^*\left(\mathbf{b} - \mathbf{A}\mathbf{x}^*\right) = \mathbf{0} \\ \mathbf{x}^* \ge \mathbf{0}, \qquad \mathbf{y}^* \ge \mathbf{0}, \end{gathered} \tag{25}$$

and they characterize the solution. The same conditions form the Kuhn–Tucker conditions for the *dual problem*

$$\min_{\mathbf{y}} \mathbf{y}\mathbf{b} \text{subject to} \mathbf{y}\mathbf{A} \ge \mathbf{c}, \ \mathbf{y} \ge \mathbf{0}, \tag{26}$$

where the variables of the dual problem, **y**, are the Lagrange multipliers of the original (primal) problem. The dual problem uses the same matrix **A** and vectors **b** and **c** as the primal problem, but it is one of minimization, rather than maximization, and the constraint constants of the primal problem become the coefficients of the objective function of the dual problem while the coefficients of the objective function of the primal problem become the constraint constants of the dual problem. Vectors $\mathbf{x}^*$, $\mathbf{y}^*$ satisfying (24) or (25) solve both the primal problem (6) and dual problem (26). Geometrically, in the case of linear programming the opportunity set is a polyhedral closed convex set since it is the intersection of $m+n$ half spaces defined by the m inequality and n non-negativity constraints. The contours of the linear

objective function are hyperplanes, and the problem is solved on the highest hyperplane within the polyhedral set. A solution must occur at a vertex, in which case it is unique, or along a bounding face, in which case it is non-unique. As in the more general case of nonlinear programming, however, the solution always occurs at a point where the gradient vector of the objective function (here **c**) lies in the cone spanned by the outward pointing normals to the opportunity set (here the relevant columns of **A** or the relevant outward pointing unit vectors corresponding to the non-negativity constraints).

Another special case of nonlinear programming, one which subsumes linear programming, is that of quadratic programming. The *problem of quadratic programming* is that of

$$\max_{\mathbf{x}} F(\mathbf{x}) = \mathbf{cx} + \frac{1}{2}\mathbf{x}'\mathbf{Qx} \text{ subject to } \mathbf{Ax} \leq \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \tag{27}$$

where **A**, **b**, **c** are as in the linear programming problem and **Q** is a given *n* x *n* negative semidefinite symmetric matrix. The problem is one of concave programming because **Q** is negative semidefinite and the linear transformation **Ax** is convex. Furthermore the constraint qualification is met. The Kuhn–Tucker conditions are therefore both necessary and sufficient, that is, the vector $\mathbf{x}^*$ solves (27) if and only if there is a vector of Lagrange multipliers $\mathbf{y}^*$ such that the pair $\mathbf{x}^*$, $\mathbf{y}^*$ satisfies the Kuhn–Tucker conditions

$$\begin{gathered}
\frac{\partial L}{\partial \mathbf{x}} = \left(\mathbf{c} + \mathbf{x}^{*\prime}\mathbf{Q} - \mathbf{y}^*\mathbf{A}\right) \leq \mathbf{0}, \qquad \frac{\partial L}{\partial \mathbf{y}} = \mathbf{b} - \mathbf{Ax}^* \geq \mathbf{0} \\
\frac{\partial L}{\partial \mathbf{x}}\mathbf{x}^* = \left(\mathbf{c} + \mathbf{x}^{*\prime}\mathbf{Q} - \mathbf{y}^*\mathbf{A}\right)\mathbf{x}^* = 0, \quad \mathbf{y}^*\frac{\partial L}{\partial \mathbf{y}} = \mathbf{y}^*\left(\mathbf{b} - \mathbf{Ax}^*\right) = 0 \\
\mathbf{x}^* \geq \mathbf{0}, \qquad \mathbf{y}^* \geq \mathbf{0},
\end{gathered} \tag{28}$$

where **L(x**, **y)** is the Lagrangian function

$$L(\mathbf{x}, \mathbf{y}) = \mathbf{cx} + \frac{1}{2}\mathbf{x}'\mathbf{Qx} + \mathbf{y}(\mathbf{b} - \mathbf{Ax}). \tag{29}$$

This case reduces to that of linear programming if **Q**, the matrix defining the quadratic form $\frac{1}{2}\mathbf{x}'\mathbf{Q}\mathbf{x}$ in (27) vanishes.

5 Neoclassical theory of the household and the firm

The problem of nonlinear programming can be applied in economics to the neoclassical theory of both the household and the firm, the two most important units of microeconomics. For the household, assume there are *n* goods (and services) available where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ is the column vector of goods purchased and consumed by the household. Assume further that the household seeks to maximize a utility function, a real-valued function defined on these goods $U(\mathbf{x}) = U(x_1, x_2, \ldots, x_n)$. Assume finally that the household purchases non-negative quantities of each good so as to maximize the utility function subject to a budget constraint that states that expenditure on all *n* goods cannot exceed available income. The *neoclassical problem of the household* is then

$$\max_{\mathbf{x}} \; U(\mathbf{x}) \text{subject to } \mathbf{p}\mathbf{x} \leq I, \mathbf{x} \geq \mathbf{0}. \tag{30}$$

Here **p** is a given row vector of (positive) prices of each of the *n* goods, and *I* is the given (positive) income available to the household. Thus the household chooses non-negative amounts of goods **x** so as to maximize the utility function *U*(**x**) subject to the budget constraint $\mathbf{p}\mathbf{x} \leq I$ which states that expenditure on all *n* goods cannot exceed income.
This problem is one of nonlinear programming, so, introducing the (single) Lagrange multiplier *y* the Lagrangian is

$$L(\mathbf{x}, y) = U(\mathbf{x}) + y(I - \mathbf{p}\mathbf{x}). \tag{31}$$

The Kuhn–Tucker conditions characterize an optimum point. Under the further regularity conditions that $\mathbf{x}^* > \mathbf{0}$ and *U*(**x**) is a twice continuously differentiable function in a neighbourhood of $\mathbf{x}^*$ with a non-singular Hessian matrix of second-order derivatives there then exist solutions for the purchases of goods $\mathbf{x}^*$ as functions of the *n*+1 parameters **p**, *I*, which are the *demand functions* characterizing the optimum point (Hicks, 1946; Samuelson, 1947; Wold and Jureen, 1953; Intriligator, 1971; Barten and Böhm, 1982; Phlips, 1983).

For the firm, assume that the firm uses *n* inputs to produce a single output, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ is a column vector of inputs, *q* is output, and *f*(**x**) is the production function of the firm. Assume further

that the firm seeks to maximize a profit function, given as the difference between revenue and cost. Assume finally that the firm purchases non-negative inputs and produces non-negative output subject to the technology of the given production function so as to maximize profit. The *neoclassical theory of the firm* is then

$$\max_{\mathbf{x}} \pi(\mathbf{x}) = pf(\mathbf{x}) - \mathbf{wx} \text{ subject to } \mathbf{x} \geq \mathbf{0}. \tag{32}$$

Here $pf(\mathbf{x})$ is total revenue, p being the given price of output, and $\mathbf{wx}$ is the cost of production, the total expenditure on all inputs, $\mathbf{w}$ being the vector of given prices (wages) of inputs.

This problem is also one of nonlinear programming, and the Kuhn–Tucker conditions characterize an optimum point. Under the further regularity conditions that $\mathbf{x}^* > \mathbf{0}$ and $f\bullet(\mathbf{x})$ is twice continuously differentiable in a neighbourhood of $\mathbf{x}^*$ with a non-singular Hessian matrix of second-order derivatives there exist solutions for the purchase of inputs $\mathbf{x}^*$ and production of output q as functions of the n+1 parameters $\mathbf{w}$, p which are the *input demand functions* and *output supply function* characterizing the optimum point (Hicks, 1946; Samuelson, 1947; Intriligator, 1971; Nadiri, 1982).

The problem of linear programming can be applied to a firm that produces output using an *activity analysis* technology. In such a case the firm produces n outputs $x_1, x_2, \ldots, x_n$ using m inputs $b_1, b_2, \ldots, b_m$. To produce one unit of output x_j requires a_{ij} units of input i. In the short run all inputs are fixed so the only choice for the firm is that of deciding what mix of outputs to produce given these inputs. The problem is then

$$\max_{\mathbf{x}} \mathbf{cx} \text{ subject to } \mathbf{Ax} \leq \mathbf{b},\ \mathbf{x} \geq \mathbf{0}, \tag{33}$$

as in (6). Here the objective function to be maximized is total revenue, where c_j is the given price of output j, so the problem is one of choosing non-negative outputs so as to maximize profit, given the technology (the a_{ij}) and the inputs (the b_i). The dual problem is

$$\min_{\mathbf{y}} \mathbf{yb} \text{ subject to } \mathbf{yA} \geq \mathbf{c},\ \mathbf{y} \geq \mathbf{0}, \tag{34}$$

as in (26), which can be interpreted as choosing non-negative values (shadow prices) for the inputs $y_1, y_2, \ldots, y_m$ so as to minimize the cost of the inputs $\mathbf{yb} = \Sigma y_i b_i$ where y_i is the chosen value and b_i is the given level of input *i*. The *n* constraints state that the unit cost of good *j*, obtained by summing the cost of producing one unit over all inputs, is no less than the price of this good. The dual to a problem of allocation, the primal problem (33), is one of valuation, the dual problem (34). According to the complementary slackness conditions in (25) if for any output *j* unit costs exceeds price (that is, the output is produced at a loss) then this output is not produced $(x_j^* = 0)$ and if for any input *i* not all of the input is used then it is valued at zero $(y_i^* = 0)$.

In conclusion, the problem of nonlinear programming is one with important applications to the microeconomic theory of the household and the firm, leading to conditions characterizing an equilibrium at an optimum point. The problem also has many other applications throughout economics.

See Also

- calculus of variations
- convex programming
- demand theory
- duality
- functional analysis
- Lagrange multipliers
- linear programming

Bibliography

Aoki, M. 1971. *Introduction to Optimization Techniques*. New York: Macmillan.

Arrow, K.J., Hurwicz, L. and Uzawa, H. 1958. Constraint qualifications in maximization problems. In *Studies in Linear and Nonlinear Programming*, ed. K.J. Arrow, L. Hurwicz and H. Uzawa. Stanford: Stanford University Press.

Arrow, K.J., Hurwicz, L. and Uzawa, H. 1961. Constraint qualifications in maximization problems. *Naval Research Logistics Quarterly* 8, 175–91.

Avriel, M. 1976. *Nonlinear Programming: Analysis and Methods*. Englewood Cliffs, NJ: Prentice-Hall.

Barten, A.P. and Böhm, V. 1982. Consumer theory. Chapter 9 in *Handbook of Mathematical Economics*, vol. 2, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Bazaraa, M.S., Goode, J.J. and Shetty, C.M. 1972. Constraint qualifications revisited. *Management*

Science 18, 567–73.

Bazaraa, M.S. and Shetty, C.M. 1976. *Foundations of Optimization*. Berlin: Springer-Verlag.

Bazaraa, M.S. and Shetty, C.M. 1979. *Nonlinear Programming: Theory and Algorithms*. New York: Wiley.

Dantzig, G. 1963. *Linear Programming and Extensions*. Princeton: Princeton University Press.

Dennis, J.E. 1984. A user's guide to nonlinear optimization algorithms. *Proceedings IEEE* 72, 1765–76.

Dorfman, R., Samuelson, P.A. and Solow, R.M. 1958. *Linear Programming and Economic Analysis*. New York: McGraw-Hill.

Gale, D. 1960. *The Theory of Linear Economic Models*. New York: McGraw-Hill.

Gass, S.I. 1975. *Linear Programming: Methods and Applications*. 4th edn. New York: McGraw-Hill.

Hadley, G. 1963. *Linear Programming*. Reading, MA: Addison-Wesley.

Hadley, G. 1964. *Nonlinear and Dynamic Programming*. Reading, MA: Addison-Wesley.

Hestenes, M.R. 1975. *Optimization Theory: The Finite Dimensional Case*. New York: Wiley.

Hicks, J.R. 1946. *Value and Capital*. 2nd edn. New York: Oxford University Press.

Intriligator, M.D. 1971. *Mathematical Optimization and Economic Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Intriligator, M.D. 1981. Mathematical programming with applications to economics. Chapter 2 in *Handbook of Mathematical Economics*, vol. 1, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

John, F. 1948. Extremum problems with inequalities as side conditions. In *Studies and Essays: Courant Anniversary Volume*, ed. K.O. Friedrichs, O.W. Neugebauer, and J.J. Stoker. New York: Interscience Publishers.

Kuhn, H.W. and Tucker, A.W. 1951. Nonlinear Programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Lancaster, K. 1968. *Mathematical Economics*. New York: Macmillan.

Luenberger, D.G. 1973. *Introduction to Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley.

Mangasarian, O.L. 1969. *Nonlinear Programming*. New York: McGraw-Hill.

Martos, B. 1975. *Nonlinear Programming: Theory and Methods*. Amsterdam: North-Holland.

McCormick, G.P. 1983. *Nonlinear Programming: Theory, Algorithms, and Applications*. New York: Wiley.

Nadiri, M.I. 1982. Producers theory. Chapter 10 in *Handbook of Mathematical Economics*, vol. 2, ed. K. J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Phlips, L. 1983. *Applied Consumption Analysis*. Revised edn. Amsterdam: North-Holland.

Polak, E. 1971. *Computational Methods in Optimization: a Unified Approach*. New York: Academic Press.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Schittkowski, K. 1980. *Nonlinear Programming Codes*. Berlin: Springer-Verlag.

Takayama, A. 1985. *Mathematical Economics*. 2nd edn. New York: Cambridge University Press.

Wold, H. and Jureen, L. 1953. *Demand Analysis*. New York: Wiley.

Zangwill, W.I. 1969. *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall.

How to cite this article

Intriligator, Michael D. "nonlinear programming." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000083> doi:10.1057/9780230226203.1195

The New Palgrave Dictionary of Economics Online

nonlinear time series analysis

Bruce Mizrach
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since the early 1980s, there has been a growing interest in stochastic nonlinear dynamical systems of the form $x_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-p}) + \sigma(x_t)\epsilon_t$, where $\{x_t\}_{t=0}^{\infty}$ is a zero mean, covariance stationary process, $f: R^{p+1} \to R$, σ is the conditional volatility, and $\{\epsilon_t\}_{t=0}^{\infty}$ is an independent and identically distributed noise process. The major recent developments in nonlinear time series are described using this canonical model: (*a*) representation theory; (*b*) nonparametric modelling; (*c*) ergodic properties; (*d*) piecewise linear models; (*e*) volatility modelling; (*f*) hypothesis testing for linearity and normality; (*g*) forecasting.

Keywords

chaos; cointegration; ergodicity; forecasting; GARCH models; inference; kernel estimators; linear models; Lyapunov exponents; Markov switching models; nonlinear time series analysis; nonparametric estimation; piecewise linear models; regime dependence; regime switching; representation theory; stochastic volatility models; testing; threshold autoregressions; Volterra expansion; wavelets

Article

Since the early 1980s, there has been a growing interest in stochastic nonlinear dynamical systems of the form

$$x_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-p}) + \sigma(x_t)\epsilon_t, \tag{1}$$

where $\{x_t\}_{t=0}^{\infty}$ is a zero mean, covariance stationary process, $f: R^{p+1} \to R$, σ is the conditional volatility,

and $\{\varepsilon_t\}_{t=0}^{\infty}$ is an independent and identically distributed noise process. The major recent developments in nonlinear time series are described here using this canonical model. The first section develops representation theory for a third order approximation. Nonparametric approaches follow; these rely on series expansions of the general model. Ergodic properties including path dependence and dimension are considered next. I then consider two widely utilized parametric models, piecewise linear models of *f* and autoregressive models for volatility. I conclude with a discussion of hypothesis testing and forecasting.

Volterra expansion

There is no general causal representation for nonlinear time series as in the linear case. Series approximations rely on the *Volterra expansion*,

$$x_{t+1} \simeq f(0) + \sum_{i=1}^{p} f_{i_1} x_{t-i_1} + \sum_{i_1=1}^{p} \sum_{i_2=i_1}^{p} f_{i_1 i_2} x_{t-i_1} x_{t-i_2} + \sum_{i_1=1}^{p} \sum_{i_2=i_1}^{p} \sum_{i_3=i_2}^{p} f_{i_1 i_2 i_3} x_{t-i_1} x_{t-i_2} x_{t-i_3} + \cdots \quad (2)$$

Brockett (1976) shows any continuous map over $[0, T]$ can be approximated by a finite Volterra series. Mittnik and Mizrach (1992) examine forecasts using generalized polynomial expansions like (2). Potter (2000) shows that in the cubic case, a one-sided Wold-type representation in terms of white noise v_t can be obtained,

$$x_{t+1} \simeq \sum_{i=1}^{\infty} g_{i_1} v_{t-i_1} + \sum_{i_1=1}^{\infty} \sum_{i_2=i_1}^{\infty} g_{i_1 i_2} v_{t-i_1} v_{t-i_2} + \sum_{i_1=1}^{\infty} \sum_{i_2=i_1}^{\infty} \sum_{i_3=i_2}^{\infty} g_{i_1 i_2 i_3} x_{t-i_1} x_{t-i_2} x_{t-i_3}. \quad (3)$$

Koop, Pesaran and Potter (1996) note that the impulse response functions, $E[x_{t+n}|x_t, v_t] - E[x_{t+n}|x_t]$ will depend upon the size and sign of v_t as well as the current state x_t.
I now turn to nonparametric approaches which build on approximations like 2.

Nonparametric estimation

Consider the local polynomial approximation to $f(.)$ around x_0,

$$\hat{f}(x) = \sum_{j=0}^{m} \beta_j (x - x_0)^j. \quad (4)$$

In the case $j = 0$, this corresponds to the *kernel regression* estimator of Nadaraya and Watson,

$$\hat{f}(x) = \frac{\sum_{t=1}^{T} x_{t+1} K_h(x_t - x_0)}{\sum_{t=1}^{T} K_h(x_t - x_0)}. \tag{5}$$

The K_h are *kernels*, usually functions with a support on a compact set, assigning greater weight to observations closer to x_0. *h* is the *bandwidth* parameter, determining the size of the histogram bin. *Nearest neighbours* estimation is the case where *h* is adjusted to find a fixed number of nearby observations *k*. More generally, the *local linear approximation* solves,

$$\alpha_0, \beta_0 \min \left(x_{t+1} - \alpha_0 - \beta_0 (x_t - x_0)\right)^2 K_h(x_t - x_0). \tag{6}$$

The estimator (5) corresponds to the case where the only regressor in (6) is the constant term.
The application of these methods in the time series case is a fairly recent development. Conditions for consistency and asymptotic normality rely on *mixing conditions* where the dependence between x_{t+j} and x_t becomes negligible as *j* grows large.
A closely related approach involves the use of a *recurrent neural network*,

$$\Psi_i(x_t, h_{t-1}) = \Psi(\gamma_{i0} + \gamma_{i1} x_t + \sum_{k=1}^{r} \delta_{ik} h_{k,t-k}),$$

$$x_{t+1} = \Phi\left(\beta_0 + \sum_{i=1}^{p} \beta_i \Psi_i(x_t, h_{t-1})\right). \tag{7}$$

Kuan, Hornik and White (1994) provide convergence results for bounded Ψ (most commonly the logistic) as *p* grows large.
A popular approach in the frequency domain is wavelets. The *discrete wavelet transform* is

$$x_{t+1} = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \gamma(j, k) \Psi_{j,k}(t), \tag{8}$$

where the *mother wavelet* $\Psi(t)$,

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}} \Psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j} \right), \tag{9}$$

is parameterized by scale s_0 and translation τ, and the wavelet coefficients are given by

$$\gamma(j, k) = \left\{ \Psi_{j,k}(t), x(t) \right\}. \tag{10}$$

Daubechies (1992) orthonormal basis functions,

$$E[\Psi_{j,k}(t) \Psi_{m,n}(t)] = 0, \quad \forall j \neq m, k \neq n, \tag{11}$$

have received the widest application.
Even when very little is known about f or σ, nonlinear time series analysis can shed light on the long run average or *ergodic* properties of the dynamical system.

Ergodic properties

Mathematicians have known since Poincaré that even simple maps like (1) can produce very complex dynamics. The nonlinear time series literature has developed tools for estimation of ergodic properties of these systems. Denote by $Df(\bar{x})$ the Jacobian matrix of partial derivatives of (1),

$$\begin{bmatrix} \partial f_1 / \partial x_1 & \cdots & \partial f_1 / \partial x_p \\ \vdots & \ddots & \vdots \\ \partial f_p / \partial x_1 & \cdots & \partial f_p / \partial x_p \end{bmatrix}$$
(12)

evaluated at $\bar{x}$. Replacing 12 with a sample analog,

$$J_t = \begin{bmatrix} \Delta f_1 / \Delta x_{1,t} & \cdots & \Delta f_1 / \Delta x_{p,t} \\ \vdots & \ddots & \vdots \\ \Delta f_p / \Delta x_{1,t} & \cdots & \Delta f_p / \Delta x_{p,t} \end{bmatrix}$$
(13)

we compute eigenvalues V_i,

$$V_i(Q_T' Q_T)$$
(14)

rank ordered from $1, \ldots, p$, where

$$Q_T = J_{T-p} \cdot J_{T-p-1} \cdots J_1$$
(15)

The *Lyapunov exponents* are defined for the positive eigenvalues V_i^+ as

$$T \to \infty \lim \lambda_i = \frac{1}{2(T-p)} \ln V_i^+,$$
(16)

and a single exponent greater than 1 characterizes a system with sensitive dependence. Popularly known as

‘chaos’, this property implies that dynamic trajectories become unpredictable even when the state of the system is known with certainty. Gençay and Dechert (1992) and Shintani and Linton (2004) provide methods for estimating these. Shintani and Linton (2003; 2004) reject the presence of positive Lyapunov exponents in both real output and stock returns.

The sum of the Lyapunov exponents also provides a measure of the Kolmogorov–Sinai *entropy* of the system. This tells the researcher how quickly trajectories separate. Mayfield and Mizrach (1991) estimate this time at about 15•minutes for the S&P 500 index.

A final quantity of interest is the *dimension* p of the dynamical system. Nonlinear econometricians try to estimate the dimension from a scalar m-history. A powerful result due to Takens (1981) says this can be done as long as $m \geq 2p + 1$. Diks (2004) has shown that the scaling of correlation exponents seems to be consistent with the stochastic volatility model.

A great deal of progress has been made with parametric models of (1) as well. I begin with the widely utilized piecewise linear models.

Piecewise linear models

The most widely applied parametric nonlinear time series specification has been the Markov switching model introduced by James Hamilton (1989). The function f is a piecewise linear function,

$$f(x_t) = \left\{ \begin{array}{l} \mu^{(1)} + \sum_{j=0}^{p} \varphi_j^{(1)} (x_{t-j} - \mu^{(1)}, S_t = s_t^{(1)}) \\ \vdots \\ \mu^{(m)} + \sum_{j=0}^{p} \varphi_j^{(m)} (x_t - \mu^{(m)}, S_t = s_t^{(m)}) \end{array} \right\}, \tag{17}$$

where the changes among states are governed by an unobservable regime switching process, $S_t = s_t^{(i)}$, $i = 1, \ldots m$, an $m \times m$ transition matrix Π, and $E[x_t | S_t = s_t^{(i)}] = \mu^{(i)}$. When S_t is unobserved, $Pr(S_t | x_{t-1})$ is nonlinear in x_{t-1}. Hamilton has shown that a two-dimensional switching model describes well the business cycle dynamics in the United States. This model has been extended to include regime dependence in volatility (Kim, 1994) and time varying transition probabilities (Filardo, 1994).

The latent state vector requires forming prior and posterior estimates of which regime you are in. The EM algorithm (Hamilton, 1990) and Bayesian Gibbs sampling methods (Albert and Chib, 1993) have proven fruitful in handling this problem. Hypothesis testing is also non-standard because under the alternative of $m - 1$ regimes, the conditional mean parameters are nuisance parameters. Hansen (1996) has explored carefully these issues.

A closely related framework is the *threshold autoregressive* (TAR) model,

$$f(x_t) = \begin{cases} \left[\mu^{(1)} + \sum_{j=0}^{p} \varphi_j^{(1)} \left(x_{t-j} - \mu^{(1)}\right)\right] I\left(q\left(x_{t-d}, Z_t\right) \le \gamma_1\right) \\ \left[\mu^{(2)} + \sum_{j=0}^{p} \varphi_j^{(2)} \left(x_{t-j} - \mu^{(2)}\right)\right] I\left(\gamma_1 < q\left(x_{t-d}, Z_t\right) \le \gamma_2\right) \\ \vdots \\ \left[\mu^{(m)} + \sum_{j=0}^{p} \varphi_j^{(m)} \left(x_t - \mu^{(m)}\right)\right] I\left(q\left(x_{t-d}, Z_t\right) > \gamma_{m-1}\right) \end{cases}.$$

(18)

$I(\,.\,)$ is the indicator function, and $q(x_{t-d}, Z_t)$, the regime switching variable, is assumed to be an observable function of exogenous variables Z_t and lagged x's. The integer d is known as the *delay parameter*. When q depends only upon x, the model is called *self-exciting*.

Teräsvirta (1994) has developed a two-regime version of the TAR model in which regime changes are governed by a smooth transition function $G(x_{t-d}, Z_t): R^k \to [0, 1]$,

$$f(x_t) = G(x_{t-d}, Z_t) \sum_{j=0}^{p} \varphi_j^{(1)} (x_{t-j} - \mu^{(1)}) + (1 - G(x_{t-d}, Z_t)) \sum_{j=0}^{p} \varphi_j^{(2)} (x_{t-j} - \mu^{(2)}).$$

(19)

Luukkonen, Saikkonen and Teräsvirta (1988) have shown that inference and hypothesis testing in this model is often much simpler than in the piecewise linear models. Van Dijk and Franses (1999) have extended this model to multiple regimes. Applications of this framework have been widespread from macroeconomics (Teräsvirta and Anderson, 1992) to empirical finance (Franses and van Dijk, 2000).

Krolzig (1997) considers the multivariate case where $x_t = (x_{1,t}, x_{2,t}, \ldots, x_{k,t})'$ is $k \times 1$. Balke and Fomby (1997) introduced threshold cointegration by incorporating error correction terms into the thresholds. Koop, Pesaran and Potter (1996) develop a bivariate model of US GDP and unemployment where the threshold depends upon the depth of the recession.

I now turn to models that introduce nonlinearity through the error term.

Models of volatility

Engle and Bollerslev have introduced the generalized autoregressive conditional heteroskedasticity (GARCH) model,

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \sigma^2(x_{t-i}) \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i},$$

(20)

where $h_t = E[(x_t - E[x_t|\Omega_{t-1}])^2|\Omega_{t-1}]$ is the *conditional variance*. This is just a Box–Jenkins model in the squared residuals of 1 of order $(max[p, q], p)$. The model is nonlinear because the disturbances are uncorrelated, but their squares are not.

The GARCH model describes the volatility clustering and heavy-tailed returns in financial market data, and has found wide application in asset pricing and risk management applications.

Volatility modelling has been motivated by the literature on options pricing. Popular alternatives to the GARCH model include the stochastic volatility (SV) model (Ghysels, Harvey and Renault, 1996), and the realized volatility approach of Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002). The discrete-time SV model takes the form,

$$x_t = \sigma_\varepsilon \exp(h_t/2)\varepsilon_t, \tag{21}$$

$$h_t = \beta h_{t-1} + \sigma_h \eta_t,$$

where x_t is the demeaned log asset return, and ε_t and η_t are noise terms. Realized volatility sums high-frequency squared returns as an approximation of lower frequency volatility. Both GARCH and SV have been successful in explaining the departures from the Black–Scholes observed empirically.

The final two sections address the marginal contribution of nonlinear modelling to goodness of fit and forecasting.

Testing for linearity and Gaussianity

There is a large literature on testing the importance of the nonlinear components of a model. The most widely used test is due to Brock, Dechert, Scheinkman and LeBaron (BDSL, 1996). Their nonparametric procedure is built upon U-statistics. Serfling (1980) is a good introduction.

The first step is to form *m*-histories of the data,

$$x_t^m = (x_t, x_{t+1}, \ldots, x_{t+m-1}), \tag{22}$$

with joint distribution $F(x_t^m)$. Introduce the kernel $h: R^m \times R^m \to R$,

$$h(x_t^m, x_s^m) = I(x_t^m, x_s^m, \varepsilon) \equiv I\left[\left\|x_t^m - x_s^m\right\| < \varepsilon\right], \tag{23}$$

where $I(\cdot)$ is the indicator function. The *correlation integral* of Grassberger and Procaccia (1983),

$$C(m, \varepsilon) \equiv \int_X \int_X I(x_t^m, x_s^m, \varepsilon)\, dF(x_t^m)\, dF(x_s^m), \tag{24}$$

is the expected number of *m*-vectors in an ε neighbourhood. A *U*-statistic,

$$C(m, N, \varepsilon) \equiv \frac{2}{N(N-1)} \sum_{t=1}^{N-1} \sum_{s=t+1}^{N} I(X_t^m, X_s^m, \varepsilon), \tag{25}$$

is a consistent estimator of 24. BDSL demonstrate the asymptotic normality of the statistic

$$\sqrt{N} \frac{S(m, N, \varepsilon)}{\sqrt{Var[S(m, N, \varepsilon)]}} \sqrt{N} d \times \to N(0, 1), \tag{26}$$

where

$$S(m, N, \varepsilon) = C(m, N, \varepsilon) - C(1, N, \varepsilon)^m. \tag{27}$$

There is a multi-dimensional extension due to Baek and Brock (1992). De Lima (1997) explores the use of

the BDSL under moment condition failure.

There is a direct relationship between nonlinear and non-Gaussian time series. In the model (1), even if the disturbance term ε_t is normal, nonlinear transformations of Gaussian noise will make x_t non-Gaussian. Testing for Gaussianity is then an instrumental part of the nonlinear time series toolkit.

Hinich (1982) has developed testing in the time domain using the *bicorrelation*,

$$\gamma(r, s) = \sum_{t=1}^{s} x_t x_{t+r} x_{t+s} / (N - s), \quad 0 \le r \le s, \tag{28}$$

and in the frequency domain using the *bispectrum*,

$$B(\omega_1, \omega_2) = \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \gamma(r, s) \exp[-i(\omega_1 r + \omega_2 s)]. \tag{29}$$

For a Gaussian time series, the bicorrelation should be close to zero, and the bispectrum should be flat across all frequencies. Both tests have good power against skewed alternatives.

Ramsey and Rothman (1996) have proposed a related time domain procedure that looks for *time reversibility*,

$$F(X_t, X_{t+1}, \ldots, X_{t+r}) = F(X_{s-t}, X_{s-t-1}, \ldots, X_{s-t-r}) \tag{30}$$

for any r,•s and t, where $F(\cdot)$ is the joint distribution. This condition is stronger than stationarity because of the triple index. The authors find evidence of business cycle asymmetry using this diagnostic.

Forecasting

For many, the bottom line on nonlinear modelling is the ability to generate superior forecasts. In this respect, the results from the nonlinear literature are decidedly mixed. Harding and Pagan (2002) are prominent sceptics. Teräsvirta, van Dijk and Medeiros (2005) provide a very wide set of evidence in favour of nonlinear models.

Aside from the comparison of point forecasts from model i, $u_{i,t+1} = x_{t+1} - f_i(x_t)$, with a particular loss function $g(\cdot)$,

$$H_0: E[g(u_{i,t+1}) - g(u_{j,t+1})] = 0, \tag{31}$$

there has been growing interest in comparing forecast densities $p_i(x_{t+1}|f_i(x_t))$,

$$H_0: \int [p_i(x_{t+1}|f_i(x_t)) - p_j(x_{t+1}|f_j(x_t))]\,dx = 0. \tag{32}$$

Corradi and Swanson (2005) provide a comprehensive overview of available tools.

See Also

- forecasting
- linear models
- stochastic volatility models

I would like to thank Cees Diks, James Hamilton, Sebastiano Manzan, Simon Potter, Phil Rothman, Dick van Dijk and Steven Durlauf for helpful comments.

Bibliography

Albert, J. and Chib, S. 1993. Bayesian analysis via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics* 11, 1–15.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2003. Modeling and forecasting realized volatility. *Econometrica* 71, 579–625.

Barndorff-Nielsen, O.E. and Shephard, N. 2002. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society* , Series B 64, 253–80.

Baek, B. and Brock, W.A. 1992. A nonparametric test for independence of a multivariate time series. *Statistica Sinica* 2, 137–56.

Balke, N. and Fomby, T. 1997. Threshold cointegration. *International Economic Review* 38, 627–45.

Brock, W.A., Dechert, W.D., Scheinkman, J.A. and LeBaron, B. 1996. A test for independence based on the correlation dimension. *Econometric Reviews* 15, 197–235.

Brockett, R.W. 1976. Volterra series and geometric control theory. *Automatica* 12, 167–76.

Corradi, V. and Swanson, N.R. 2005. Predictive density evaluation. In *Handbook of Economic Forecasting*, ed. C.W.J. Granger, A. Timmermann and G. Elliott. Amsterdam: North-Holland.

Daubechies, I. 1992. *Ten Lectures on Wavelets*. 2nd ed. Philadelphia: SIAM.

De Lima, P. 1997. On the robustness of nonlinearity tests due to moment condition failure. *Journal of Econometrics* 76, 251–80.

Diks, C. 2004. The correlation dimension of returns with stochastic volatility. *Quantitative Finance* 4, 45–54.

Filardo, A.J. 1994. Business cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics* 12, 299–308.

Franses, P.H. and van Dijk, D. 2000. *Nonlinear Time Series Models in Empirical Finance*. New York: Cambridge University Press.

Gallant, A.R. and Tauchen, G. 1987. Seminonparametric maximum likelihood estimation. *Econometrica* 55, 363–90.

Gallant, A.R. and Tauchen, G. 1996. Which moments to match? *Econometric Theory* 12, 657–81.

Gençay, R. and Dechert, W.D. 1992. An algorithm for the n Lyapunov exponents of an n-dimensional unknown dynamical system. *Physica D* 59, 142–57.

Ghysels, E., Harvey, A. and Renault, E. 1996. Stochastic volatility. In *Handbook of Statistics* 14: *Statistical Methods in Finance*, ed. G.S. Maddala and C.R. Rao. Amsterdam: North-Holland.

Granger, C.W.J. and Hallman, J. 1991. Nonlinear transformations of integrated time series. *Journal of Time Series Analysis* 12, 207–24.

Grassberger, P. and Procaccia, J. 1983. Measuring the strangeness of strange attractors. *Physica D* 9, 189–208.

Hamilton, J.D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–84.

Hamilton, J.D. 1990. Analysis of time series subject to changes in regime. *Journal of Econometrics* 45, 39–70.

Hansen, B.E. 1996. Inference when a nuisance parameter is not identified under the null hypothesis.

Econometrica 64, 413–30.

Harding, D. and Pagan, A. 2002. Dissecting the cycle: a methodological investigation. *Journal of Monetary Economics* 49, 365–81.

Hinich, M.J. 1982. Testing for Gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis* 3, 169–76.

Kim, C.J. 1994. Dynamic linear models with Markov-switching. *Journal of Econometrics* 60, 1–22.

Koop, G., Pesaran, H. and Potter, S. 1996. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics* 74, 119–48.

Krolzig, H.-M. 1997. *Markov Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis*. Berlin: Springer.

Kuan, C.-M., Hornik, K. and White, H. 1994. A convergence result for learning in recurrent neural networks. *Neural Computation* 6, 620–40.

Luukkonen, R., Saikkonen, P. and Teräsvirta, T. 1988. Testing linearity against smooth transition autoregressive models. *Biometrika* 75, 491–9.

Mayfield, S. and Mizrach, B. 1991. Nonparametric estimation of the correlation exponent. *Physical Review* A 88, 5298–301.

Mittnik, S. and Mizrach, B. 1992. Parametric and seminonparametric analysis of nonlinear time series. In *Advances in GLIM and Statistical Modeling*, ed. L. Fahrmeir, B. Francis, R. Gilchrist and G. Tutz. New York: Springer.

Potter, S. 2000. Nonlinear impulse response functions. *Journal of Economic Dynamics and Control* 24, 1425–46.

Ramsey, J.B. and Rothman, P. 1996. Time irreversibility and business cycle asymmetry. *Journal of Money, Credit, and Banking* 28, 1–21.

Serfling, R.J. 1980. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley.

Shintani, M. and Linton, O. 2003. Is there chaos in the world economy? A nonparametric test using consistent standard errors. *International Economic Review* 44, 331–58.

Shintani, M. and Linton, O. 2004. Nonparametric neural network estimation of Lyapunov exponents and a direct test for chaos. *Journal of Econometrics* 120, 1–33.

Takens, F. 1981. Detecting strange attractors in turbulence. In *Springer Lecture Notes in Mathematics*, vol. 898, ed. D. Rand and L.-S. Young. Berlin: Springer.

Teräsvirta, T. 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* 89, 208–18.

Teräsvirta, T. and Anderson, H.M. 1992. Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics* 7, S119–S136.

Teräsvirta, T., van Dijk, D. and Medeiros, M.C. 2005. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: a re-examination. *International Journal of Forecasting* 21, 755–74.

Van Dijk, D. and Franses, P.H. 1999. Modeling multiple regimes in the business cycle. *Macroeconomic Dynamics* 3, 311–40.

How to cite this article

Mizrach, Bruce. "nonlinear time series analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000152> doi:10.1057/9780230226203.1196

The New Palgrave Dictionary of Economics Online

non-nested hypotheses

M. Hashem Pesaran and M. Rodrigo Dupleich Ulloa
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides an overview of the literature on hypotheses testing when the hypotheses or models under consideration are non-nested. Two models are said to be non-nested if neither can be obtained from the other by some limiting process, including the imposition of equality and/or inequality constrains on one of the model's parameters. Relevant concepts such as closeness measures and pseudo-true values are discussed and alternative approaches to testing non-nested hypotheses, including the Cox procedure, artificial nesting and the encompassing approach, are reviewed. The Vuong approach to model selection is also covered.

Keywords

artificial nesting; Cox's test; encompassing test; hypothesis testing; Kullback–Leibler information criteria; linear regression models; maximum likelihood estimation; model selection; non-nested hypotheses; pseudo-true values

Article

In economics, as in many other disciplines, there are competing explanations of the same phenomena, often characterized by alternative statistical models. Different models may represent, for example, different theoretical paradigms, or could be the result of alternative formulations from the same paradigm. Within the classical framework, the problem of model adequacy is approached through ‘general specification tests’, the ‘diagnostic tests’, and the ‘non-nested tests’. All three approaches can be used to test the same explanation or hypothesis of interest (the null or the maintained hypothesis), but they differ in their consideration of the alternative(s). General specification tests intentionally consider a broad class of alternatives, while the alternatives considered under diagnostic and non-nested testing procedures are much more specific. In the case of non-nested tests the null hypothesis is contrasted with a specific alternative. Non-nested tests are appropriate when rival hypotheses are advanced for the explanation of the same economic phenomenon, and the aim is to devise a powerful test against a specific

alternative.

When the null hypothesis is nested within the alternative, standard classical procedures such as those based on the likelihood ratio, Wald and Lagrange multiplier (or score) principles can be utilized. But if the null and the alternative hypotheses belong to 'separate' families of distributions, classical testing procedures cannot be applied directly and need to be suitably modified.

This article provides an overview of the concepts and some of the most widely used non-nested hypotheses tests and applies these procedures to the classical regression models. Our discussion of non-nested hypothesis testing will necessarily omit many topics. Survey articles on this subject include McAleer and Pesaran (1986), Gourieroux and Monfort (1994), and Pesaran and Weeks (2001).

Non-nested models

Suppose the object of interest is the process generating the random variable Y, observed over a sample of size n, $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$. Assume that the true process generating $\mathbf{y}$ is characterized by a joint probability density function, $f_0(\mathbf{y})$, which is unknown, and two models (hypotheses) are advanced as possible explanations of Y, represented by the joint probability density functions:

$$H_g = \{g(\mathbf{y};\ \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}, H_h = \{h(\mathbf{y};\ \boldsymbol{\gamma}),\ \boldsymbol{\gamma} \in \Gamma\}. \tag{1}$$

These functions are known but depend on a finite number of unknown parameters denoted by $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\gamma} \in \Gamma$, respectively. The sets Θ and Γ represent the 'admissible' parameter space for which the respective densities $g(\mathbf{y};\ \theta)$ and $h(\mathbf{y};\ \gamma)$ are well defined. The aim is to ascertain which of the two alternatives, H_g and H_h, if any, can be viewed as belonging to $f_0(\mathbf{y})$. In this set-up there is no natural null hypothesis; either of the two hypotheses under consideration can be taken as the null. In practice, the analysis of non-nested hypotheses is carried out with both alternatives taken in turn as the null hypothesis. Four outcomes are possible: (i) H_g rejected against H_h and not vice versa, (ii) H_h rejected against H_g and not vice versa, (iii) neither hypothesis is rejected against the other, and finally (iv) both hypotheses are rejected against one another. The first two outcomes are familiar from the classical test results and are straightforward to interpret. The third outcome can arise when the two models are very close to $f_0(\mathbf{y})$, and hence equivalent observationally. The fourth outcome suggests the existence of a third possible model which shares important features from both models under consideration.

Pseudo-true values and closeness measures

Given the observations $\mathbf{y}$, the maximum likelihood (ML) estimators of θ and γ are given by

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} L_g(\boldsymbol{\theta}),\ \hat{\boldsymbol{\gamma}}_n = \arg\max_{\boldsymbol{\gamma}\in\Gamma} L_h(\boldsymbol{\gamma}),$$

where the corresponding log-likelihood functions are defined by $L_g(\boldsymbol{\theta}) = \log(g(\mathbf{y};\ \boldsymbol{\theta}))$ and $L_h(\boldsymbol{\gamma}) = \log(h(\mathbf{y};\ \boldsymbol{\gamma}))$. Throughout we shall assume that probability densities satisfy the usual regularity conditions as established, for example in White (1982), such that $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ have asymptotically normal limiting distributions under the 'true' model, $f_0(\mathbf{y})$. In the general case where neither of the models under consideration coincide with $f_0(\mathbf{y})$, $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ are known as quasi-ML estimators and their probability limits under $f_0(\mathbf{y})$ are referred to as (asymptotic) pseudo-true values, such that

$$\boldsymbol{\theta}_{*f} = \arg\max_{\theta\in\Theta} \mathbb{E}_f\left\{L_g(\boldsymbol{\theta})\right\} \boldsymbol{\gamma}_{*f} = \arg\max_{\gamma\in\Gamma} \mathbb{E}_f\left\{L_h(\boldsymbol{\gamma})\right\}$$

(2)

where $\mathbb{E}_f(\cdot)$ denotes expectations under the true density $f_0(\mathbf{y})$. In what follows, we assume that the above asymptotic pseudo-true values exist; and θ_{*f} and γ_{*f} are the *unique* maxima to the respective optimization problem given in (2), such that *global identifiability* is ensured. For the case in which $f_0(\mathbf{y})$ belongs to H_g, we have that $\boldsymbol{\theta}_{*g} = \boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_{*g} = \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$, where the 'true' value of θ under H_g, is denoted by θ_0. Given the symmetry of our setting, under H_h we have $\boldsymbol{\theta}_{*h} = \boldsymbol{\theta}_*(\boldsymbol{\gamma}_0)$ and $\boldsymbol{\gamma}_{*h} = \boldsymbol{\gamma}_0$, where γ_0 is the 'true' value of γ under H_h. The relationship between the parameters of the two models under consideration is given by the functions $\boldsymbol{\gamma}_{*g} = \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}_{*h} = \boldsymbol{\theta}_*(\boldsymbol{\gamma}_0)$, known as the *binding functions*.

Using closeness measures and pseudo-true values, Pesaran (1987) provides a formalization of the concepts of nested and non-nested hypotheses. The closeness of H_g with respect to H_h is given by

$$C_{gh}(\boldsymbol{\theta}_0) = \boldsymbol{I}_{gh}(\boldsymbol{\theta}_0,\ \boldsymbol{\gamma}_*(\boldsymbol{\theta}_0)) = \min_{\boldsymbol{\gamma}} \boldsymbol{I}_{gh}(\boldsymbol{\theta}_0,\ \boldsymbol{\gamma})$$

(3)

$$= \mathbb{E}_g\{L_g(\boldsymbol{\theta}_0) - L_h(\boldsymbol{\gamma}_*(\boldsymbol{\theta}_0))\}$$

(4)

where $I_{gh}(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*)$ is known as the *Kullback–Leibler information criterion* (KLIC) measure, introduced by Kullback (1959). Similarly, the closeness of H_h to H_g, is defined by $C_{hg}(\boldsymbol{\gamma}_0) = I_{hg}(\boldsymbol{\gamma}_0, \boldsymbol{\theta}_*(\boldsymbol{\gamma}_0))$.

- H_g is *nested* within H_h if and only if $C_{gh}(\boldsymbol{\theta}_0) = 0$, for all values of $\theta_0 \in \Theta$, and $C_{hg}(\boldsymbol{\gamma}_0) \neq 0$ for some $\gamma_0 \in \Gamma$.
- H_g and H_h are *globally non-nested* if and only if $C_{gh}(\theta_0)$ and $C_{hg}(\gamma_0)$ are both non-zero for all values of $\theta_0 \in \Theta$ and $\gamma_0 \in \Gamma$.
- H_g and H_h are *partially non-nested* if $C_{gh}(\theta_0)$ and $C_{hg}(\gamma_0)$ are both non-zero for some values of $\theta_0 \in \Theta$ and $\gamma_0 \in \Gamma$.
- H_g and H_h are *observationally equivalent* if and only if $C_{gh}(\theta_0)=0$ and $C_{hg}(\gamma_0)=0$ for all values of $\theta_0 \in \Theta$ and $\gamma_0 \in \Gamma$.

Tests of non-nested hypotheses

There are three general approaches to testing non-nested hypotheses. The first, due to the pioneering contributions of Cox (1961; 1962), involves centring the log-likelihood ratio statistic under the null hypothesis and then deriving its asymptotic null distribution. This is known as the *Cox test*. A second approach, also suggested by Cox (1962) and explored extensively by Atkinson (1970), is based on an artificially constructed general model. The basic idea is to introduce a third hypothesis in which both H_g and H_h are nested as special cases. A third approach, originally considered by Deaton (1982) and Dastoor (1983), and further developed by Mizon and Richard (1986) known as the *encompassing procedure*, focuses on the ability of one model in explaining particular features of an alternative model. In a related contribution, Gourieroux, Monfort and Trognon (1983) extend the Wald and score-type tests to non-nested models. Their statistics are based on the difference between two estimators of the pseudo-true values.

The Cox test statistic is derived by modifying the log-likelihood ratio statistic, $L_g(\hat{\boldsymbol{\theta}}_n) - L_h(\hat{\boldsymbol{\gamma}}_n)$, so that it is appropriately centred. Specifically, for testing H_g against H_h, the numerator of the Cox statistic is given by

$$S_n^{gh} = \left\{L_g(\hat{\boldsymbol{\theta}}_n) - L_h(\hat{\boldsymbol{\gamma}}_n)\right\} - \hat{\mathbb{E}}_g\left\{L_g(\hat{\boldsymbol{\theta}}_n) - L_h(\hat{\boldsymbol{\gamma}}_n)\right\},$$

(5)

where $\hat{\mathbb{E}}_g\left\{L_g(\hat{\boldsymbol{\theta}}_n) - L_h(\hat{\boldsymbol{\gamma}}_n)\right\}$ is a consistent estimator of $C_{gh}(\theta_0)$. In the case where H_g is nested within H_h we have $C_{gh}(\theta_0)=0$ for all θ_0, and S_n^{gh} reduces to the standard log-likelihood ratio statistic. An application to linear regression models has been proposed by Pesaran (1974) and subsequently extended to simultaneous nonlinear equations systems by Pesaran and Deaton (1978). As pointed out previously, since there is no natural null hypothesis in this testing framework, one also needs to consider the modified log-likelihood statistic for testing H_h against H_g which is denoted by S_n^{hg}. Under a suitable normalization (that is $\sqrt{n}$), both statistics are asymptotically normally distributed under their respective nulls with a zero mean and a finite asymptotic variance. When the null hypothesis of H_g is considered against H_h, we have

$$N_n^{gh} = \frac{\sqrt{n}S_n^{gh}}{\sqrt{V_{gh}}} \overset{a}{\sim} N(0, 1)$$

where V_{gh} is the asymptotic variance of $\sqrt{n}S_n^{gh}$ and $\overset{a}{\sim}$ denotes asymptotical equivalence in distribution (for details see Pesaran and Deaton, 1978). Based on the results of the two statistics, N_n^{gh} and N_n^{hg}, four outcomes are possible:

- reject H_g but not H_h if $\left|N_n^{gh}\right| < c_\alpha$ and $\left|N_n^{hg}\right| \geq c_\alpha$,
- reject H_h but not H_g if $\left|N_n^{gh}\right| \geq c_\alpha$ and $\left|N_n^{hg}\right| < c_\alpha$,
- reject both H_g and H_h if $\left|N_n^{gh}\right| \geq c_\alpha$ and $\left|N_n^{hg}\right| \geq c_\alpha$,
- reject neither H_g and H_h if $\left|N_n^{gh}\right| < c_\alpha$ and $\left|N_n^{hg}\right| < c_\alpha$,

where the $(1-\alpha)$ per cent critical value of the standard normal distribution is denoted by c_α. In the case of non-nested hypotheses, there is no way of ranking the models by the level of their generality. As a consequence, the test results may provide a consistent outcome such as the rejection of H_g (or H_h) by both tests. But it is also not unusual, given the data, for both non-nested models to be simultaneously rejected or fail to be rejected. For the case of a simultaneous rejection of H_g and H_h, we need to find some other model that fits the data better. If neither model is rejected, this may indicate lack of power. The second approach, named *Atkinson's comprehensive* method (Atkinson, 1970), is based on an artificial

nesting of the two models with a general model such as,

$$f_c(\mathbf{y};\ \boldsymbol{\theta},\ \boldsymbol{\gamma},\ \lambda) = \left\{ \frac{(g(\mathbf{y};\ \boldsymbol{\theta}))^{\lambda}(h(\mathbf{y};\ \boldsymbol{\gamma}))^{1-\lambda}}{\int (g(\mathbf{y};\ \boldsymbol{\theta}))^{\lambda}(h(\mathbf{y};\ \boldsymbol{\gamma}))^{1-\lambda}\mathrm{d}\mathbf{y}},\ \ \lambda \in [0,\ 1],\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\gamma} \in \Gamma \right\}$$

(6)

Atkinson's comprehensive approach considers families that are obtained by mixing the probability distributions of H_g and H_h. It requires the existence of the integral appearing in the denominator in eq. (6). This component ensures that the combined function $f_c(\mathbf{y};\ \theta,\ \gamma,\ \lambda)$, is in fact a proper density function. In equation (6), the comprehensive model is based on an exponential combination (that is, a geometric mean); alternatively the compound model can also be derived from an arithmetic mean (see for instance Quandt, 1974). In this set-up, the hypothesis H_g is obtained by imposing $\lambda=1$, while the hypothesis H_h is obtained by imposing $\lambda=0$. Thus, in principle, by testing $\lambda=1$ or $\lambda=0$, we can test H_g or H_h, respectively. The ‘mixing’ parameter, λ, varies over the range [0,1] and measures the relative weights attached to H_g and H_h. As a consequence, tests for the restriction of $\lambda=1$ ($\lambda=0$) against the alternative that $\lambda \neq 1$ ($\lambda \neq 0$) can be performed based on standard techniques from the literature of nested hypothesis testing (see Atkinson, 1970; Pesaran, 1982b).

Atkinson's approach is, however, subject to a number of drawbacks. The first one arises from the fact that under $\lambda=1$ (or $\lambda=0$), the unknown parameter vector γ (or θ) disappears from the combined model written in (6). This is known as Davies' problem (Davies, 1977) which can be circumvented in various ways, as discussed, for example, by Pesaran (1982c). The second limitation is due to the fact that testing $\lambda=1$ against $\lambda \neq 1$ is not equivalent to performing the test of H_g against H_h, which is the primary object of the non-nested testing exercise. Finally, there is some degree of arbitrariness in the choice of the comprehensive model (see Pesaran, 1981).

The *encompassing approach* generalizes Cox's original idea and examines the extent to which H_g explains one or more features of the rival model, H_h. When all the features of the model H_h can be explained by model H_g, then H_g is said to *encompass* H_h. This condition is denoted by

$$H_g \varepsilon H_h\colon \boldsymbol{\gamma}_{*f} = \boldsymbol{\gamma}_*(\boldsymbol{\theta}_{*f}).$$

(7)

Likewise, $H_h\ \varepsilon\ H_g$ implies that all features of model H_g can be explained by the model H_h, that is H_h *encompasses* H_g, such that

$$H_h \varepsilon H_g: \boldsymbol{\theta}_{*f} = \boldsymbol{\theta}_*(\boldsymbol{\gamma}_{*f}). \tag{8}$$

Recall that γ_{*f} and θ_{*f} are the pseudo-true values of γ and θ, with respect to the true model H_f. Moreover, $\gamma_*(\cdot)$ and $\theta_*(\cdot)$ are the binding functions linking the parameters of the models under H_g and H_h. The encompassing hypothesis $H_g \varepsilon H_h$ (resp. $H_h \varepsilon H_g$) can be tested using the statistic $\sqrt{n}(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_*(\hat{\boldsymbol{\theta}}_n))$, respectively $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_*(\hat{\boldsymbol{\gamma}}_n))$. Gourieroux and Monfort (1995) show that under the encompassing hypothesis, $H_g \varepsilon H_h$ and a set of regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_*(\hat{\boldsymbol{\theta}}_n))$ is asymptotically normal with zero mean and a finite covariance matrix. Based on this result, two testing procedures are proposed by Gourieroux and Monfort (1995): the Wald encompassing test (WET) and the score encompassing test (SET). In practice, the implementation of these tests tends to be difficult. First, the binding functions $\gamma^*(\cdot)$ and $\theta^*(\cdot)$ are not easy to derive and, second, the variance–covariance matrices appearing in the test statistics tend to be difficult to compute in practice. Chen and Kuan (2002) suggest the use of ‘pseudo-true score’ as a way of avoiding the need to estimate pseudo-true values.

Vuong's model selection test

Vuong's (1989) criterion is motivated by testing that H_g and H_h are observationally equivalent, using the *Kullback–Leibler information criterion* (KLIC) as a closeness measure. The focus of this approach is to test the hypothesis that the models under consideration are ‘equally’ close to the true unknown model, $H_f: f_0(\mathbf{y})$. It provides a natural link between model selection and hypothesis testing approaches. Under model selection, a model is selected even if the ‘best’ model happens to be very close to the second best model. Vuong's approach allows the statistical significance of the differences between models to be tested using classical testing procedures. It is based on the closeness measures of H_g and H_h with respect to the true model, H_f, namely (for closeness of H_g to H_f)

$$C_{fg}(\boldsymbol{\theta}_{*f}) = \mathbb{E}_f\{L_f(\cdot) - L_g(\boldsymbol{\theta}_{*f})\}$$

and (for closeness of H_h to H_f)

$$C_{fh}(\boldsymbol{\gamma}_{*f}) = \mathbb{E}_f\{L_f(\cdot) - L_h(\boldsymbol{\gamma}_{*f})\}.$$

The null hypothesis of interest, 'H_g and H_h are equivalent', is then formally defined by

$$H_V: C_{fg}(\boldsymbol{\theta}_{*f}) = C_{fh}(\boldsymbol{\gamma}_{*f}) \tag{9}$$

which is equivalent to the unknown quantity $H_V: \mathbb{E}_f\{L_g(\boldsymbol{\theta}_{*f}) - L_h(\boldsymbol{\gamma}_{*f})\} = 0$, that depends on $f_0(\mathbf{y})$, the unknown true distribution. However, the latter difference can be consistently estimated by $T^{-1}\left\{L_g(\hat{\boldsymbol{\theta}}_n) - L_h(\hat{\boldsymbol{\gamma}}_n)\right\}$, an average of the log-likelihood ratio statistic. Vuong derives an asymptotic standard normal distribution for the related test statistic under H_V.
Rivers and Vuong (2002) provide a number of generalizations and show that the test can be applied to nonlinear dynamic models and other closeness measures.

Application to linear regression models

An important application of the non-nested tests in econometrics has been to linear regression models. Consider the following classical normal regression models:

$$H_g: \mathbf{y} = \mathbf{X}\beta + \mathbf{u}_g,\ \ \mathbf{u}_g \sim N(0, \sigma^2\mathbf{I}_n),\ \ 0 < \sigma^2 < \infty, H_h: \mathbf{y} = \mathbf{Z}\alpha + \mathbf{u}_h,\ \ \mathbf{u}_h \sim N(0, \omega^2\mathbf{I}_n),\ \ 0 < \omega^2 < \infty, \tag{10}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are $n \times k_g$ and $n \times k_h$ matrices of observations on the explanatory variables of models H_g and H_h, respectively. These variables are assumed to be distributed independently of the $n \times 1$ disturbance vectors $\mathbf{u}_g$ and $\mathbf{u}_h$. The parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ and $\boldsymbol{\gamma} = (\boldsymbol{\alpha}', \omega^2)'$ are the $(k_g + 1) \times 1$ and $(k_h + 1) \times 1$ vectors of unknown regression coefficients, and $\mathbf{I}_n$ is the identity matrix of order n. It is also assumed that the probability limits of $\widehat{\Sigma}_{xx} = n^{-1}(\mathbf{X}'\mathbf{X})$, $\widehat{\Sigma}_{zz} = n^{-1}(\mathbf{Z}'\mathbf{Z})$ and $\widehat{\Sigma}_{xz} = n^{-1}(\mathbf{X}'\mathbf{Z})$ exist with population values denoted by the non-singular matrices Σ_{xx}, Σ_{zz} and Σ_{xz}. At the same time, define $\widehat{\Sigma}_g = \widehat{\Sigma}_{xx} - \widehat{\Sigma}_{xz}\widehat{\Sigma}_{zz}^{-1}\widehat{\Sigma}_{zx} > 0$ and $\widehat{\Sigma}_h = \widehat{\Sigma}_{zz} - \widehat{\Sigma}_{zx}\widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xz} > 0$. The link between these strict inequality restrictions and the nesting properties of the models in (10) will be made clear below.
Suppose that neither H_g nor H_h belong to the true DGP, and the data is generated by

$$H_f: \mathbf{y} = \mathbf{W}\delta + \mathbf{u}_f,\ \ \mathbf{u}_f \sim N(0, v^2\mathbf{I}_n),\ \ 0 < v^2 < \infty.$$

(11)

As before, assume that $\widehat{\Sigma}_{ww} = n^{-1}(\mathbf{W}'\mathbf{W})$, $\widehat{\Sigma}_{wx} = n^{-1}(\mathbf{W}'\mathbf{X})$ and $\widehat{\Sigma}_{wz} = n^{-1}(\mathbf{W}'\mathbf{Z})$ can be replaced by their population values given by Σ_{ww}, Σ_{wx} and Σ_{wz}. The pseudo-true values given in (2) can be obtained for this case by maximizing $\mathbb{E}_f\left\{n^{-1}L_g(\boldsymbol{\theta})\right\}$ with respect to θ which yields

$$\boldsymbol{\theta}_{*f} = \begin{pmatrix} \boldsymbol{\beta}_{*f} \\ \sigma^2_{*f} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx}^{-1}\Sigma_{xw}\boldsymbol{\delta} \\ \sigma^2 + \boldsymbol{\delta}'\left(\Sigma_{ww} - \Sigma_{wx}\Sigma_{xx}^{-1}\Sigma_{xw}\right)\boldsymbol{\delta} \end{pmatrix}.$$

(12)

Similarly, for model H_h we have

$$\boldsymbol{\gamma}_{*f} = \begin{pmatrix} \boldsymbol{\alpha}_{*f} \\ \omega^2_{*f} \end{pmatrix} = \begin{pmatrix} \Sigma_{zz}^{-1}\Sigma_{zw}\boldsymbol{\delta} \\ \sigma^2 + \boldsymbol{\delta}'\left(\Sigma_{ww} - \Sigma_{wz}\Sigma_{zz}^{-1}\Sigma_{zw}\right)\boldsymbol{\delta} \end{pmatrix}.$$

(13)

Note that, for the case in which H_f belongs to the family of models given by H_g, the latter result can be rewritten as

$$\boldsymbol{\gamma}_{*g} = \begin{pmatrix} \boldsymbol{\alpha}_{*g} \\ \omega^2_{*g} \end{pmatrix} = \begin{pmatrix} \Sigma_{zz}^{-1}\Sigma_{zx}\boldsymbol{\beta} \\ \sigma^2 + \boldsymbol{\beta}'\Sigma_g\boldsymbol{\beta} \end{pmatrix}.$$

(14)

In terms of our previous discussion, these regression models are non-nested if it is not possible to write **X** as an exact linear function of **Z** and *vice versa*, or more formally if $\mathbf{X} \not\subseteq \mathbf{Z}$ and $\mathbf{Z} \not\subseteq \mathbf{X}$. The model H_g is said to be nested in H_h if $\mathbf{X} \subset \mathbf{Z}$ and $\mathbf{Z} \not\subseteq \mathbf{X}$. The two models are observationally equivalent if $\mathbf{X} \subset \mathbf{Z}$ and $\mathbf{Z} \subset \mathbf{X}$. These conditions can be written in terms of the KLIC measure given by (3) as in McAleer and Pesaran

(1986), who derive the closeness measure of H_h with respect to H_g as

$$C_{gh}(\boldsymbol{\theta}) = \frac{1}{2}\log\left[1 + \frac{\boldsymbol{\beta}'\Sigma_g\boldsymbol{\beta}}{\sigma^2}\right].$$

Similarly, the KLIC measure of closeness of H_g with respect to H_h is

$$C_{hg}(\boldsymbol{\gamma}) = \frac{1}{2}\log\left[1 + \frac{\boldsymbol{\alpha}'\Sigma_h\boldsymbol{\alpha}}{\omega^2}\right].$$

It can be easily seen from this example that a necessary and sufficient condition for H_g to be nested within H_h is $\boldsymbol{\beta}'\Sigma_g\boldsymbol{\beta} / \sigma^2 = 0$ for all admissible values of β with $\boldsymbol{\alpha}'\Sigma_h\boldsymbol{\alpha} / \omega^2 \neq 0$ for some α. Note that $\boldsymbol{\beta}'\Sigma_g\boldsymbol{\beta} / \sigma^2 = 0$ is implied by either $\Sigma_g\boldsymbol{\beta} = 0$ or $\Sigma_g = 0$.
Given the linear set-up and using results in Pesaran (1974), the adjusted log-likelihood ratio statistic for testing H_g against H_h is given by

$$S_n^{gh} = \frac{n}{2}\log\left(\frac{\widehat{\omega}^2_{*ng}}{\widehat{\omega}^2_n}\right) \tag{15}$$

where $\widehat{\omega}^2_n$ is the estimate of ω^2 under the alternative H_h, and $\widehat{\omega}^2_{*ng}$ is the estimated pseudo-true value of the residual variance of H_h under H_g, such that

$$\widehat{\omega}^2_n = n^{-1}(\mathbf{y} - \mathbf{X}\widehat{\alpha}_n)'(\mathbf{y} - \mathbf{X}\widehat{\alpha}_n)\widehat{\omega}^2_{*ng} = \hat{\sigma}^2_n + \hat{\boldsymbol{\beta}}'_n\widehat{\Sigma}_g\hat{\boldsymbol{\beta}}'_n$$

in which the estimates under the true model H_g are given by $\hat{\sigma}_n^2 = n^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n')'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n')$, $\hat{\boldsymbol{\beta}}_n' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\widehat{\boldsymbol{\alpha}}_n = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$. As pointed out earlier, since we do not have a natural null hypothesis in this framework, one also needs to evaluate H_g against H_h, for which the modified log-likelihood statistic is given by

$$S_n^{hg} = \frac{n}{2}\log\left(\frac{\hat{\sigma}_{*nh}^2}{\hat{\sigma}_n^2}\right). \tag{16}$$

For the statistic given by (15), the asymptotic variance of $\sqrt{n}S_n^{gh}$, denoted by V_{gh}, can be computed as follows:

$$V_{gh} = \frac{\hat{\sigma}_n^2\left(\hat{\boldsymbol{\beta}}_n'\mathbf{X}'\mathbf{M}_Z\mathbf{M}_X\mathbf{M}_Z\mathbf{X}\hat{\boldsymbol{\beta}}_n'\right)}{n\left(\hat{\sigma}_n^2 + \hat{\boldsymbol{\beta}}_n'\widehat{\Sigma}_g\hat{\boldsymbol{\beta}}_n'\right)} \tag{17}$$

where $\mathbf{M}_X = \mathbf{I}_n - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{M}_Z = \mathbf{I}_n - \mathbf{Z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ are orthogonal projection matrices.

Combining (15) and (17), the associated standardized Cox statistic, $N_n^{gh} = \sqrt{n}S_n^{gh} / \sqrt{V_{gh}}$, can now be calculated as described in Pesaran (1974). Similar derivations lead to the analogue statistic for the test of H_h against H_g, N_n^{hg}.

The application of the comprehensive approach to the above linear regression models yields the following exponential combination as presented in (6):

$$H_\lambda: \mathbf{y} = (1 - \xi)\mathbf{X}\boldsymbol{\beta} + \xi\mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}, \quad \mathbf{u} \sim N(0, \sigma_\lambda^2\mathbf{I}_n) \tag{18}$$

where $\xi = \lambda\sigma_\lambda^2 / \sigma^2$ and $\sigma_\lambda^{-2} = (1-\xi)\sigma^{-2} + \xi\omega^{-2}$. Given that the error variances σ^2 and ω^2 are strictly positive, performing a test of ξ=0 is equivalent to testing λ=0 when we consider the null hypothesis of H_g against H_h. As pointed out earlier, the Davies problem arises when under the null hypothesis of H_g (λ=0), the unknown parameter vector α disappear from the mixture model. The presence of this nuisance parameter results in a Student-type of test statistic associated with λ that depends on the value of α, such that

$$t_\lambda(a) = \frac{a'\mathbf{Z}'\mathbf{M}_X\mathbf{y}}{\hat{\sigma}_\lambda^2 (a'\mathbf{Z}'\mathbf{M}_X\mathbf{Z}a)^{1/2}} \tag{19}$$

where $\hat{\sigma}_\lambda^2$ denotes the usual estimator of the variance of the errors. One possible way to solve this identification problem would be to construct a test statistic based on $F_\lambda = \max_\alpha t_\lambda(\alpha)$.

A different approach to deal with the identification problem was proposed by Davidson and MacKinnon (1981), who propose a J-test by replacing the nuisance parameter α by its estimator, $\widehat{\boldsymbol{\alpha}}_n$, under H_h. An exact version of this test, proposed by Fisher and McAleer (1981) and known as the JA-test (indicating the Atkinson variation of this test), substitutes α by the estimate of its pseudo-true value under H_h given in (14), that is $\widehat{\boldsymbol{\alpha}}_*(\hat{\boldsymbol{\beta}}_n) = \widehat{\Sigma}_{ZZ}^{-1}\widehat{\Sigma}_{ZX}\hat{\boldsymbol{\beta}}_n$. By symmetry of our testing problem, the J and JA versions of the *t*-test can also be calculated for H_h against H_g. Davidson and MacKinnon (1981) show that the *t*-ratio statistic, $t_\lambda(\widehat{\boldsymbol{\alpha}}_n)$, has asymptotically a standard normal distribution under the null.

Based on the application of Roy's union-intersection principle, McAleer and Pesaran (1986) show that the test for ξ=0 in (18) is equivalent to the standard F-statistic for the test of δ_2=0 in the combined model $\mathbf{y} = \mathbf{X}\boldsymbol{\delta}_1 + \mathbf{Z}\boldsymbol{\delta}_2 + \mathbf{u}$.

In order to frame the linear regression models into the encompassing type tests, we can focus on the discrepancy between the OLS estimator of the regression coefficients, denoted by $\widehat{\boldsymbol{\alpha}}_n$, and the estimator of the pseudo-true value in finite samples, such that $\sqrt{n}(\widehat{\boldsymbol{\alpha}}_n - \widehat{\boldsymbol{\alpha}}_*(\hat{\boldsymbol{\beta}}_n)) = \sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{y}$. Using this, we can build an encompassing statistic for testing H_g ε H_h, as follows:

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_n - \widehat{\boldsymbol{\alpha}}_*(\hat{\boldsymbol{\beta}}_n)) = \sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{W}\boldsymbol{\delta} + \sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{u}_f,$$

if H_f is taken as the true model given in (11). As a consequence, under some regularity conditions,

$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_n - \widehat{\boldsymbol{\alpha}}_*(\hat{\boldsymbol{\beta}}_n))$ is asymptotically normally distributed with mean zero and the covariance matrix $v^2\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_{xx}^{-1}$. Using these results the WET statistic for testing H_g ε H_h, is given by

$$\varepsilon_{gh} = \frac{\mathbf{y}'\mathbf{M}_x\mathbf{Z}_1(\mathbf{Z}_1'\mathbf{M}_x\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{M}_x\mathbf{y}}{\hat{v}^2} \tag{20}$$

where $\mathbf{Z}_1$ are the components in **Z** that are orthogonal to **X**. Similarly, a *variance* encompassing test of H_g ε H_h can be constructed for the discrepancy between a consistent estimate of ω^2 and its pseudo-true value ω_{*f}^2, which takes the form of $\widehat{\omega}_n^2 - \widehat{\omega}_*^2(\hat{\boldsymbol{\theta}}_n)$. For the case in which H_g contains the true model, H_h, the variance encompassing test is asymptotically equivalent to the Cox and the J-tests.
Vuong's test criterion for the comparison of H_g and H_h is computed as

$$G_{gh} = \frac{\sum_{i=1}^n d_i}{\left[\sum_{i=1}^n (d_i - \bar{d})^2\right]^{1/2}} \tag{21}$$

where $\bar{d} = n^{-1}\sum_{i=1}^n d_i$; $d_i = -1/2(\hat{\sigma}_n^2/\widehat{\omega}_n^2) - 1/2((\hat{u}_{ig}^2/\hat{\sigma}_n^2) - (\hat{u}_{ih}^2/\widehat{\omega}_n^2))$; and $\hat{u}_{ig}$ and $\hat{u}_{ih}$ are the estimated residuals of the underlying linear models given by (10). Under the null hypothesis H_V, H_g and H_h are equivalent and G_{gh} is approximately distributed as a standard normal variate.

Extensions and empirical applications

Non-nested tests have also been derived for a number of other models, including tests of non-nested linear regression models with serially correlated errors (McAleer, Pesaran and Bera, 1990), regression models estimated by instrumental variables (Ericsson, 1983), models estimated by generalized method of moments (Smith, 1992), generalized empirical likelihood (Ramalho and Smith, 2002), conditional empirical likelihood (Otsu and Whang, 2005), non-nested Euler equations (Ghysels and Hall, 1990), autoregressive versus moving average models (Walker, 1967), autoregressive conditional heteroskedastic models (Bera and Higgins, 1997; McAleer and Ling, 1998), logit and probit models (Pesaran and Pesaran, 1993), non-nested threshold autoregressive models (Altissimo and Violante, 2001; Pesaran and

Potter, 1997), and stochastic volatility models (Kim, Shephard and Chib, 1998).

Further theoretical contributions include a robust version of Cox-type statistics that controls for the effect of contamination in the data (Victoria-Feser, 1997), conditional tests on sufficient statistics (Pace and Salvan, 1990), asymptotic improvements to Davidson and MacKinnon's approach (Royston and Thompson, 1995), score-type statistics which are constructed from linear combinations of the likelihood functions (Santos Silva, 2001) and the enhancement of finite-sample performance of non-nested tests by bootstrap methods (Godfrey, 1998; Davidson and MacKinnon, 2002).

Various economic applications of non-nested hypothesis testing have appeared in the literature. Among them, savings and consumption functions (Deaton, 1982), Keynesian and new classical models of unemployment (Pesaran, 1982a), wage-employment bargaining models (Vannetelbosch, 1996), effects of dividend taxes on corporate investment decisions (Poterba and Summers, 1983), money demand functions (McAleer, Fisher and Volker, 1982; Elyasiani and Nasseh, 1994), autoregressive and moving-average schemes for unanticipated inflation series (Pagan, Hall and Trivedi, 1983), exchange rates models (Backus, 1984), alternative crop response models (Ackello-Ogutu, Paris and Williams, 1985; Frank, Beattie and Embleton, 1990), agricultural marketing margins (Lyon and Thompson, 1993), economic growth models (Ram, 1986; Dowrick and Gemmell, 1991; Bleaney and Nishiyama, 2002), and hedonic house prices (Dubin and Sung, 1990; Goodman and Dubin, 1991). In the literature of empirical industrial organization, non-nested hypothesis testing is applied to compare a Nash and collusive pricing in an industry with vertical product differentiation (Bresnahan, 1987). Non-nested tests are also applied in game-theoretic contexts by Gasmi, Laffont and Vuong (1992) and Sandler and Murdoch (1990), in sociological research by Halaby and Weakliem (1993), and in political science by Clarke (2001).

Non-nested tests for rival linear regression models can be computed using various econometric packages. See, for example, Pesaran and Pesaran (1997).

See Also

- econometrics
- encompassing
- linear models
- maximum likelihood
- model selection
- specification problems in econometrics
- statistical inference

Bibliography

Ackello-Ogutu, C., Paris, Q. and Williams, W.A. 1985. Testing a von Liebig crop response function against polynomial specifications. *American Journal of Agricultural Economics* 67, 873–80.

Altissimo, F. and Violante, G.L. 2001. The nonlinear dynamics of output and unemployment in the US. *Journal of Applied Econometrics* 16, 461–86.

Atkinson, A. 1970. A method for discriminating between models. *Journal of the Royal Statistical Society B* 32, 323–53.

Backus, D. 1984. Empirical models of the exchange rate: separating the wheat from the chaff. *Canadian Journal of Economics* 17, 824–46.

Bera, A.K. and Higgins, M.L. 1997. ARCH and bilinearity as competing models for nonlinear dependence. *Journal of Business and Economic Statistics* 15, 43–50.

Bleaney, M. and Nishiyama, A. 2002. Explaining growth: a contest between models. *Journal of Economic Growth* 7, 43–56.

Bresnahan, T.F. 1987. Competition and collusion in the American automobile market: the 1955 price war. *Journal of Industrial Economics* 35, 457–82.

Chen, Y.T. and Kuan, C.M. 2002. The pseudo-true score encompassing test for non-nested hypotheses. *Journal of Econometrics* 106, 271–95.

Clarke, K.A. 2001. Testing nonnested models of international relations: reevaluating realism. *American Journal of Political Science* 45, 724–44.

Cox, D.R. 1961. Tests of separate families of hypothesis. *Proceedings of the Forth Berkeley Symposium on Mathematical Statistics and Probability* 1, 105–23.

Cox, D.R. 1962. Further results on tests of separate families of hypothesis. *Journal of the Royal Statistical Society B* 24, 406–24.

Dastoor, N.K. 1983. Some aspects of testing nonnested hypothesis. *Journal of Econometrics* 21, 213–28.

Davidson, R. and MacKinnon, J.G. 1981. Several tests for model specification in the presence of alternative hypothesis. *Econometrica* 49, 781–93.

Davidson, R. and MacKinnon, J.G. 2002. Bootstrap J tests of nonnested linear regression models. *Journal of Econometrics* 109, 167–93.

Davies, R.B. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–54.

Deaton, A.S. 1982. Model selection procedures, or, does the consumption function exist? In *Evaluating*

the Reliability of Macro-economic Models, ed. G.C. Chow and P. Corsi. New York: Wiley.

Dowrick, S. and Gemmell, N. 1991. Industrialisation, catching up and economic growth: a comparative study across the world's capitalist economies. *Economic Journal* 101, 263–75.

Dubin, R.A. and Sung, C.-H. 1990. Specification of hedonic regressions: non-nested tests on measures of neighborhood quality. *Journal of Urban Economics* 27, 97–110.

Elyasiani, E. and Nasseh, A. 1994. The appropriate scale variable in the U.S. money demand: an application of nonnested tests of consumption versus income measures. *Journal of Business and Economic Statistics* 12, 47–55.

Ericsson, N. 1983. Asymptotic properties of instrumental variables statistics for testing nonnested hypothesis. *Review of Economic Studies* 50, 287–304.

Fisher, G.R. and McAleer, M. 1981. Alternative procedures and associated tests of significance for nonnested hypothesis. *Journal of Econometrics* 16, 103–19.

Frank, M.D., Beattie, B.R. and Embleton, M.E. 1990. A comparison of alternative crop response models. *American Journal of Agricultural Economics* 72, 597–603.

Gasmi, F., Laffont, J.J. and Vuong, Q. 1992. Econometric analysis of collusive behavior in a soft-drink market. *Journal of Economics and Management Strategy* 1, 277–311.

Ghysels, E. and Hall, A. 1990. Testing nonnested Euler conditions with quadrature based method approximation. *Journal of Econometrics* 46, 273–308.

Godfrey, L.G. 1998. Tests of non-nested regression models: some results on small sample behaviour and the bootstrap. *Journal of Econometrics* 84, 59–74.

Goodman, A.C. and Dubin, R.A. 1991. Sample stratification with non-nested alternatives: theory and a hedonic example. *Review of Economics and Statistics* 72, 168–73.

Gourieroux, C. and Monfort, A. 1994. Testing nonnested hypothesis. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Gourieroux, C. and Monfort, A. 1995. Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* 11, 195–228.

Gourieroux, C., Monfort, A. and Trognon, A. 1983. Testing nested and nonnested hypothesis. *Journal of Econometrics* 21, 83–115.

Halaby, C.N. and Weakliem, D.L. 1993. Ownership and authority in the earnings function: nonnested tests of alternative specifications. *American Sociological Review* 58, 16–30.

Kim, S., Shephard, N. and Chib, S. 1998. Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies* 65, 361–93.

Kullback, S. 1959. *Statistics and Information Theory*. New York: Wiley.

Lyon, C.C. and Thompson, G.D. 1993. Temporal and spatial aggregation: alternative marketing margin models. *American Journal of Agricultural Economics* 75, 523–36.

McAleer, M., Fisher, G. and Volker, P. 1982. Separate misspecified regressions and the U.S. long run demand for money function. *Review of Economics and Statistics* 64, 572–83.

McAleer, M. and Ling, S. 1998. A nonnested tests for GARCH and EGARCH models. Working paper, Department of Economics, University of Western Australia.

McAleer, M. and Pesaran, M.H. 1986. Statistical inference in nonnested econometric models. *Applied Mathematics and Computation* 20, 271–311.

McAleer, M., Pesaran, M.H. and Bera, A.K. 1990. Alternative approaches to testing nonnested models with autocorrelated disturbances: an application to models of U.S. unemployment. *Communications in Statistics A* 19, 3619–44.

Mizon, G. and Richard, J.F. 1986. The encompassing principle and its applications to testing nonnested hypothesis. *Econometrica* 3, 657–78.

Otsu, T. and Whang, Y.J. 2005. Testing for non-nested conditional moment restrictions via conditional empirical likelihood. Discussion Paper No. 1533, Cowles Foundation, Yale University.

Pace, L. and Salvan, A. 1990. Best conditional tests for separate families of hypotheses. *Journal of the Royal Statistical Society B* 52, 125–34.

Pagan, A.R., Hall, A.D. and Trivedi, P.K. 1983. Assessing the variability of inflation. *Review of Economic Studies* 50, 585–96.

Pesaran, M.H. 1974. On the general problem of model selection. *Review of Economic Studies* 41, 153–71.

Pesaran, M.H. 1981. Pitfalls of testing nonnested hypotheses by the Lagrange multiplier method. *Journal of Econometrics* 17, 323–31.

Pesaran, M.H. 1982a. A critique of the proposed tests of the natural rate-rational expectations hypothesis.

Economic Journal 92, 529–54.

Pesaran, M.H. 1982b. Comparison of local power of alternative tests of nonnested regression models. *Econometrica* 50, 1287–305.

Pesaran, M.H. 1982c. On the comprehensive method of testing nonnested regression models. *Journal of Econometrics* 18, 263–74.

Pesaran, M.H. 1987. Global and partial nonnested hypothesis and asymptotic local power. *Econometric Theory* 3, 69–97.

Pesaran, M.H. and Deaton, A.S. 1978. Testing nonnested nonlinear regression models. *Econometrica* 46, 677–94.

Pesaran, M.H. and Pesaran, B. 1993. A simulation approach to the problem of computing Cox's statistic for testing nonnested models. *Journal of Econometrics* 57, 377–92.

Pesaran, M.H. and Pesaran, B. 1997. *Working with Microfit 4.0*. Oxford: Oxford University Press.

Pesaran, M.H. and Potter, S. 1997. A floor and ceiling model of US output. *Journal of Economic Dynamics and Control* 21, 661–96.

Pesaran, M.H. and Weeks, M. 2001. Nonnested hypothesis testing: an overview. In *Companion to Theoretical Econometrics*, ed. B.H. Baltagi. Oxford: Basil Blackwell.

Poterba, J.M. and Summers, L.H. 1983. Dividend taxes, corporate investments, and ‘Q’. *Journal of Public Economics* 22, 135–67.

Quandt, R.E. 1974. A comparison of methods for testing nonnested hypothesis. *Review of Economics and Statistics* 56, 92–9.

Ram, R. 1986. Government size and economic growth: a new framework and some evidence from cross-section and time-series data. *American Economic Review* 76, 191–203.

Ramalho, J.J.S. and Smith, R.J. 2002. Generalized empirical likelihood nonnested tests. *Journal of Econometrics* 107, 99–125.

Rivers, D. and Vuong, Q. 2002. Model selection tests for nonlinear dynamic models. *Econometrics Journal* 5, 1–39.

Royston, P. and Thompson, S.G. 1995. Comparing non-nested regression models. *Biometrics* 51, 114–27.

Sandler, T. and Murdoch, J.C. 1990. Nash–Cournot or Lindahl behavior?: an empirical test for the NATO allies. *Quarterly Journal of Economics* 105, 875–94.

Santos Silva, J.M.C. 2001. A score test for non-nested hypothesis with applications to discrete data models. *Journal of Applied Econometrics* 16, 577–97.

Smith, R.J. 1992. Nonnested for competing models estimated by generalized method of moments. *Econometrica* 4, 973–80.

Vannetelbosch, V.J. 1996. Testing between alternative wage-employment bargaining models using Belgian aggregate data. *Labour Economics* 3, 43–64.

Victoria-Feser, M.-P. 1997. A robust tests for non-nested hypothesis. *Journal of the Royal Statistical Society B* 59, 715–27.

Vuong, Q.H. 1989. Likelihood ratio tests for model selection and nonnested hypothesis. *Econometrica* 57, 307–33.

Walker, A.M. 1967. Some tests of separate families of hypothesis in time series analysis. *Biometrika* 54, 39–68.

White, H. 1982. Regularity conditions for Cox's test of nonnested hypothesis. *Journal of Econometrics* 19, 301–18.

How to cite this article

Pesaran, M. Hashem and M. Rodrigo Dupleich Ulloa. "non-nested hypotheses." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000084> doi:10.1057/9780230226203.1197

The New Palgrave Dictionary of Economics Online

nonparametric structural models

Rosa L. Matzkin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Nonparametric structural models facilitate the analysis of counterfactuals without making use of parametric assumptions. Such methods make use of the behavioural and equilibrium assumptions specified in economic models to define a mapping between the distribution of the observable variables and the primitive functions and distributions that are used in the model. Using these methods, one can infer elements of the model, such as utility and production functions, that are not directly observed. We review some of the latest works that have dealt with the identification and estimation of nonparametric structural models.

Keywords

additivity; average derivative methods; control functions; convergence; curse of dimensionality; endogeneity; estimation; identification; instrumental variables; maximum likelihood; nonadditivity; nonparametric structural models; nonseparable models; observable and unobservable explanatory variables; partial integration methods; quantile structural functions; simultaneous equations

Article

The interplay between economic theory and econometrics comes to its full force when analysing structural models. These models are used in industrial organization, marketing, public finance, labour economics and many other fields in economics. Structural econometric methods make use of the behavioural and equilibrium assumptions specified in economic models to define a mapping between the distribution of the observable variables and the primitive functions and distributions that are used in the model. Using these methods, one can infer elements of the model, such as utility and production functions, that are not directly observed. This allows one to predict behaviour and equilibria outcomes under new environments and to evaluate the welfare of individuals and profits of firms under alternative policies, among other benefits.

To provide an example, suppose that one would like to predict the demand for a new product. Since the product has not previously been available, no direct data exists. However, one could use data on the demand for existent products together with a structural model, as shown and developed by McFadden (1974). Characterize the new product and the existent competing products by vectors of common attributes. Assume that consumers derive utility from the observable and unobservable attributes of the products, and that each chooses the product that maximizes his or her utility of those attributes among the existent products. Then, from the choice of consumers among existent products, one can infer their preferences for the attributes, and then predict what the choice of each of them would be in a situation when a new vector of attributes, corresponding to the new product, is available. Moreover, one could get a measure of the differences in the welfare of the consumers when the new product is available.
Economic theory seldom has implications regarding the parametric structures that functions and distributions may possess. The behavioural and equilibrium specifications made in economic models typically imply shape restrictions, such as monotonicity, concavity, homogeneity, weak separability, and additive separability, and exclusion restrictions, but typically not parametric specifications, such as linearity of conditional expectations, or normal distributions for unobserved variables. Nonparametric methods, which do not require specification of parametric structures for the functions and distributions in a model, are ideally fitted, therefore, to analyse structural models, using as few a priori restrictions as possible. Nonparametric techniques have been applied to many models, such as discrete choice models, tobit models, selection models, and duration models. We will concentrate here, however, on the basic models and on those, indicate some of the latest works that have dealt with identification and estimation.

Nonparametric structural econometric models

As with parametric models, a nonparametric econometric model is characterized by a vector X of variables that are determined outside the model and are observable, a vector ε of variables that are determined outside the model and are unobservable, a vector Υ of outcome variables, which are determined within the model and are unobservable, and a vector Y of outcome variables that are determined within the model and are observable. These variables are related by functional relationships, which determine the causal structure by which Υ and Y are determined from X and ε. The functional relationships are characterized by some functions that are known and some that are unknown. Similarly, some distributions may be known, some are unknown, and the others should be derived from the functional relationships and the known and unknown functions and distributions. Let $\underline{h}^*$ denote the vector of all the unknown functions in the model, $\underline{F}^*$ denote the vector of all unknown distributions, and $\zeta^* = (\underline{h}^*, \underline{F}^*)$. In contrast to parametric models, in nonparametric models, none of the coordinates of ζ^* is assumed known up to a finite dimensional parameter. Only restrictions such as continuity or values of the conditional expectations are assumed. The specification of the model should be such that from any vector $\zeta = (\underline{h}, \underline{F})$, satisfying those same restrictions that ζ^* is assumed to satisfy, one is able to derive a distribution for the observable variables, $F_{YX}(\cdot, \zeta)$.

Nonparametric identification

When specifying an econometric model, we may be interested in testing it, or we may be interested in estimating $\zeta^* = (\underline{h}^*, \underline{F}^*)$ or some feature of ζ^*, such as only one of the elements of $\underline{h}^*$, or even the value of that element at one point. Suppose that interest lies on estimating a particular feature, $\psi(\zeta^*)$ of ζ^*. The first question one must answer is whether that feature is identified. Let Ω denote the set of all possible values that $\psi(\zeta)$ may attain, when ζ is restricted to satisfy the properties that ζ^* is assumed to satisfy. Given $\psi \in \Omega$, define $\Gamma_{Y,X}(\psi)$ to be the set of all probability distributions of (Y, X) that are consistent with ψ. This is the set of all distributions that can be generated by a ζ, satisfying the properties that ζ^* is assumed to satisfy, and with $\psi(\zeta)=\psi$. We say that two values $\psi, \psi' \in \Omega$ are *observationally equivalent* if

$$\left[\Gamma_{Y,X}(\psi) \cap \Gamma_{Y,X}\left(\psi'\right)\right] \neq \varnothing,$$

that is, they are observationally equivalent if there exist a distribution of the observable variables that could have been generated by two vectors ζ and ζ' with $\psi(\zeta)=\psi$ and $\psi(\zeta')=\psi'$. The feature $\psi^*=\psi(\zeta^*)$ is said to be identified if there is no $\psi \in \Omega$ such that $\psi \neq \psi^*$ and ψ is observationally equivalent to ψ^*. That is, $\psi^*=\psi(\zeta^*)$ is identified if a change from ψ^* to $\psi \neq \psi^*$ cannot be compensated by a change in other unknown elements of ζ, so that a same distribution of observable variables could be generated by both, vectors ζ^* and ζ with $\psi^*=\psi(\zeta^*)$ and $\psi=\psi(\zeta)$.
When ψ^* can be expressed as a continuous functional of the distribution of observable variables (Y, X) one can typically estimate ψ^* nonparametrically by substituting the distribution by a nonparametric estimator for it.

Additive and nonadditive models with exogenous explanatory variables

The current literature on nonparametric econometrics methods considers additive and nonadditive models. In *additive models*, the unobservable exogenous variables ε are specified as affecting the value of Y though an additive function. Hence, for some functions m and v and some unobservable η

$$Y = m(X) + v(X, \varepsilon) = m(X) + \eta.$$

In these models, the object of interest is typically the function m. Depending on the restrictions that one may impose on η, m may denote a conditional expectation, a conditional quantile, or some other function. Many methods exist to estimate conditional means and conditional quantiles nonparametrically. Prakasa Rao (1983), Härdle and Linton (1994), Matzkin (1994; 2007b), Pagan and

Ullah (1999), Koenker (2005), and Chen (2007), among others, survey parts of this literature. Some nonparametric tests for these models include Wooldridge (1992), Yatchew (1992), Hong and White (1995), and Fan and Li (1996).

In *nonadditive models*, one is interested in analysing the interaction between the unobservable and observable explanatory variables. These models are specified, for some function *m* as

$$Y = m(X, \varepsilon).$$

Nonparametric identification and estimation in models of this type was studied in Roehrig (1988), Olley and Pakes (1996), Brown and Matzkin (1998), Matzkin (1999; 2003; 2004; 2005; 2006), Chesher (2003), Imbens and Newey (2003), and Athey and Imbens (2006), among others.

Dependence between observable and unobservable explanatory variables

In econometric models, it is often the case that in an equation of interest, some of the explanatory variables are endogenous; they are not distributed independently of the unobservable explanatory variables in that same equation. This typically occurs when restrictions such as agent's optimization and equilibrium conditions generate interrelationships among observable variables and unobservable variables, ε, that affect a common observable outcome variable, *Y*. In such cases, the distribution of the observable outcome and observable explanatory variables does not provide enough information to recover the causal effect of those explanatory variables on the outcome variable, since changes in those explanatory variables do not leave the value of ε fixed. A typical example of this is when *Y* denotes quantity demanded for a product, *X* denotes the price of the product, and ε is an unobservable demand shifter. If the price that will make firms produce a certain quantity increases with quantity, this change in ε will generate an increment in the price *X*. Hence, the observable effect of a change in price in demanded quantity would not correspond to the effect of changing the value of price when the value ε stays constant. Another typical example arises when analysing the effect of years of education on wages. An unobservable variable, such as ability, affects wages and also affects the decision about years of education.

Estimation techniques for additive and nonadditive functions of endogenous variables

The estimation techniques that have been developed to estimate nonparametric models with endogenous explanatory variables typically make use of additional information, which provides some exogenous variation on either the value of the endogenous variable or on the value of the unobservable variable. The common procedures are based on conditional independence methods and on instrumental variable methods. In the first set of procedures, independence between the unobservable and observable explanatory variables in a model is typically achieved by either *conditioning on observable* variables, or *conditioning on unobservable* variables. A *control function* approach (Heckman and Robb, 1985)

models the unobservable as a function, so that conditioning on that function purges the dependence between the explanatory observable and unobservable variables in the model. Instrumental variable methods derive identification from an independence condition between the unobservable and an external variable (an instrument) or function, which is correlated with the endogenous variable and which might be estimable.

Conditioning on unobservable variables often requires the estimation of those unobservable variables. Two-step procedures, where they are first estimated, and then used as additional regressors in the model of interest have been developed for additive models by Ng and Pinkse (1995), Pinkse (2000), and Newey, Powell and Vella (1999), among others. Two-step procedures for nonadditive models have been developed by Altonji and Matzkin (2001), Blundell and Powell (2003), Chesher (2003), and Imbens and Newey (2003), among others. Conditional moment estimation methods or quasi-maximum likelihood estimation methods can also be used (see, for example, Ai and Chen, 2003). Altonji and Ichimura (2000), Altonji and Matzkin (2001; 2005), and Matzkin (2004), among others, considered conditioning on observables for estimation of nonadditive models with endogenous explanatory variables. Matzkin (2004) provides insight into the sources of exogeneity that are generated when conditioning on either observables or unobservables, and which allow identification and estimation in nonadditive models. In particular, if $Y=m(X, \varepsilon)$, with m strictly increasing in ε, and ε is independent of X conditional on W, she shows that there exists functions $s(W, \eta)$ and $r(W, \delta)$ such that δ is independent η conditional on W, $X=s(W, \eta)$ and $\varepsilon=r(W, \delta)$. Hence,

$$Y = m(X, \varepsilon) = m(s(W, \eta), r(W, \delta)).$$

Instrumental variable methods for additive models were considered by Newey and Powell (1989; 2003), Ai and Chen (2003), Darolles, Florens and Renault (2003), and Hall and Horowitz (2003), among others. To develop estimators for m in the model

$$Y_1 = m(Y_2) + \varepsilon E[\varepsilon|Z] = 0.$$

they use the moment condition

$$E[Y_1|Z = z] = \int m(y_2)\, f_{Y_2|Z=z}(y_2)\, dy_2,$$

which depends on the conditional expectation $E[Y_1|Z=z]$ and the conditional density $f_{Y_2|Z=z}(y_2)$, which can be estimated nonparametrically. For nonadditive models, of the form

$$Y_1 = m(Y_2, \varepsilon)\ \varepsilon \text{ independent of } Z$$

where m is strictly increasing in ε, Chernozhukov and Hansen (2005) and Chernozhukov, Imbens and Newey (2007) developed estimation methods using the moment condition that for al τ

$$\tau = E[1(\varepsilon < \tau)] = E[1(\varepsilon < \tau)|Z]$$

from which m can be estimated using the conditional moment restriction

$$E[1(Y_1 < m(Y_2, \varepsilon)) - \varepsilon|Z] = 0.$$

Matzkin (2006) considered the model

$$Y_1 = m_1(Y_2, \varepsilon)\ Y_2 = m_2(Y_1, Z, \eta)$$

where Z is distributed independently of (ε, η). She established restrictions on the functions m_1 and m_2 and on the distribution of (ε, η, Z) under which

$$\left[\frac{\partial r_1(y_1, y_2)}{\partial y_2}\right]^{-1}\left[\frac{\partial r_1(y_1, y_2)}{\partial y_1}\right]$$

can be expressed as a function of the conditional density $f_{Y_1, Y_2|Z=z^*}(y_1, y_2)$, where r_1 is the inverse of m_1 with respect to ε, and the value z^* of the instrument Z is easily identified (see also Matzkin, 2005; 2007a; 2007b).

Estimation of averages and average derivatives

Nonparametric estimators are notorious by their slow rate of convergence, which worsens as the dimension of the number of arguments of the nonparametric function increases. A remedy for this is to consider averages of the nonparametric function. The average derivative method in Powell, Stock and Stoker (1989) and the partial integration methods of Newey (1994) and Linton and Nielsen (1995), for example, show how rates of convergence can increase by averaging a nonparametric function or its derivatives. This approach has been extended to cases where the explanatory variables are endogenous, using additional variables to deal with the endogeneity, and averaging over them. Examples are Blundell and Powell's (2003) *average structural function*, Imbens and Newey's (2003) *average quantile function*, and Altonji and Matzkin's (2001; 2005) *local average response* function.
Suppose, for example, that the model of interest is

$$Y_1 = m(Y_2, \varepsilon)$$

and W is such that Y_2 and ε are independent conditional on W. Then, the Blundell and Powell (2003) average structural function is

$$G(y_2) = \int m(y_2, \varepsilon) f_\varepsilon(\varepsilon) d\varepsilon$$

which can be derived from a nonparametric estimator for the distribution of (Y_1, Y_2, W) as

$$G(y_2) = \int E(Y_1 | Y_2 = y_2, W = w) f_W(w) dw.$$

Imbens and Newey's (2003) quantile structural function is defined for the τ-th quantile of ε, $q_\varepsilon(\tau)$, as

$$r(y_2, y_1) = \Pr(m(Y_2, q_\varepsilon(\tau)) \leq y_1 | Y_2 = y_2)$$

which can be estimated by

$$r(y_2, y_1) = \int \Pr(Y_1 \leq y_1 | Y_2 = y_2, W = w) f_W(w) dv.$$

Altonji and Matzkin's (2001; 2005) local average response function is

$$\beta(y_2) = \int \frac{\partial m(y_2, \varepsilon)}{\partial y_2} f_{\varepsilon | Y_2 = y_2}(\varepsilon) d\varepsilon$$

which can be derived from a nonparametric estimator for the distribution of (Y_1, Y_2, W) as

$$\beta(y_2) = \int \frac{\partial E(Y_1 | Y_2 = y_2, W = w)}{\partial y_2} f_{W | Y_2 = y_2}(w) dw.$$

Conclusions

The literature on nonparametric structural models has been rapidly developing in recent years. The new methods allow one to analyse counterfactuals without making use of parametric assumptions. Estimation of some features of the model rather than the functions themselves may reduce the curse of dimensionality, therefore providing improved properties and reducing the need for large data-sets.

See Also

- endogeneity and exogeneity
- identification
- quantile regression
- simultaneous equations models

Bibliography

Ai, C. and Chen, X. 2003. Efficient estimation of models with conditional moments restrictions containing unknown functions. *Econometrica* 71, 1795–843.

Altonji, J.G. and Ichimura, H. 2000. Estimating derivatives in nonseparable models with limited dependent variables. Mimeo. Northwestern University.

Altonji, J.G. and Matzkin, R.L. 2001. Panel data estimators for nonseparable models with endogenous regressors. NBER Working Paper T0267.

Altonji, J.G. and Matzkin, R.L. 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73, 1053–102.

Athey, S. and Imbens, G. 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74, 431–97.

Blundell, R. and Powell, J.L. 2003. Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics, Theory and Applications, Eighth World Congress*, vol. 2, ed. M. Dewatripont, L.P. Hansen and S.J. Turnovsky. Cambridge: Cambridge University Press.

Brown, D.J. and Matzkin, R.L. 1998. Estimation of nonparametric functions in simultaneous equations models, with an application to consumer demand. Discussion Paper No. 1175, Cowles Foundation, Yale University.

Chen, X. 2007. Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, vol. 6, ed. E. Leamer and J.J. Heckman. Amsterdam: North-Holland.

Chernozhukov, V. and Hansen, C. 2005. An IV model of quantile treatment effects. *Econometrica* 73, 245–61.

Chernozhukov, V., Imbens, G. and Newey, W. 2007. Instrumental variable estimation of nonseparable models. *Journal of Econometrics* 139, 4–14.

Chesher, A. 2003. Identification in nonseparable models. *Econometrica* 71, 1404–441.

Darolles, S., Florens, J.P. and Renault, E. 2003. Nonparametric instrumental regression. IDEI Working Paper No. 228.

Fan, Y. and Li, Q. 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica* 64, 4.

Florens, J.P. 2003. Inverse problems and structural econometrics: the example of instrumental variables. In *Advances in Economics and Econometrics, Theory and Applications*, vol. 2, ed. M. Dewatripont, L.P. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

Hall, P. and Horowitz, J.L. 2003. Nonparametric methods for inference in the presence of instrumental variables. Working Paper No. 102/03, CMMD.

Härdle, W. and Linton, O. 1994. Applied nonparametric methods. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engel and D.F. McFadden. Amsterdam: North-Holland.

Heckman, J.J. and Robb, R. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. J.J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Hong, Y. and White, H. 1995. Consistent specification testing via nonparametric series regression. *Econometrica* 63, 1133–59.

Imbens, G.W. and Newey, W.K. 2003. Identification and estimation of triangular simultaneous equations models without additivity. Mimeo, Massachusetts Institute of Technology.

Koenker, R.W. 2005. *Quantile Regression*. Cambridge: Cambridge University Press.

Linton, O.B. and Nielsen, J.B. 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.

Matzkin, R.L. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engel and D.L. McFadden. Amsterdam: North-Holland.

Matzkin, R.L. 1999. Nonparametric estimation of nonadditive random functions. Mimeo. Northwestern University.

Matzkin, R.L. 2003. Nonparametric estimation of nonadditive random functions. *Econometrica* 71, 1339–75.

Matzkin, R.L. 2004. Unobservable instruments. Mimeo, Northwestern University.

Matzkin, R.L. 2005. Identification in nonparametric simultaneous equations. Mimeo, Northwestern University.

Matzkin, R.L. 2006. Estimation of nonparametric simultaneous equations. Mimeo, Northwestern University.

Matzkin, R.L. 2007a. Heterogenous choice. Invited lecture, Ninth World Congress of the Econometric Society. In *Advanced in Economics and Econometrics, Theory and Applications, Ninth World Congress*,

vol. 3, ed. R. Blundell, W., Newey, and T. Persson. Cambridge: Cambridge University Press.

Matzkin, R.L. 2007b. Nonparametric identification. In *Handbook of Econometrics*, vol. 6, ed. E. Leamer and J.J. Heckman. Amsterdam: North-Holland.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

Newey, W.K. 1994. Kernel estimation of partial means and a general variance estimator. *Econometric Theory* 10, 233–53.

Newey, W. and Powell, J. 1989. Instrumental variables estimation of nonparametric models. Mimeo. Princeton University.

Newey, W. and Powell, J. 2003. Instrumental variables estimation of nonparametric models. *Econometrica* 71, 1565–78.

Newey, W.K., Powell, J.L. and Vella, F. 1999. Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67, 565–603.

Ng, S. and Pinkse, J. 1995. Nonparametric two-step estimation of unknown regression functions when the regressors and the regression error are not independent. Mimeo, CIREQ.

Olley, G.S. and Pakes, A. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64, 1263–97.

Pagan, A. and Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Pinkse, J. 2000. Nonparametric two-step regression estimation when regressors and errors are dependent. *Canadian Journal of Statistics* 28, 289–300.

Powell, J.L., Stock, J.H. and Stoker, T.M. 1989. Semiparametric estimation of index coefficients. *Econometrica* 51, 1403–30.

Prakasa Rao, B.L.S. 1983. *Nonparametric Functional Estimation*. New York: Academic Press.

Roehrig, C.S. 1988. Conditions for identification in nonparametric and parametric models. *Econometrica* 56, 433–47.

Wooldridge, J. 1992. Nonparametric regression tests based on an infinite dimensional least squares procedure. *Econometric Theory* 8, 435–51.

Yatchew, A.J. 1992. Nonparametric regression tests based on an infinite dimensional least squares procedure. *Econometric Theory* 8, 435–51.

How to cite this article

Matzkin, Rosa L. "nonparametric structural models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000143> doi:10.1057/9780230226203.1198

The New Palgrave Dictionary of Economics Online

non-profit organizations

Richard Steinberg and Burton A. Weisbrod

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Non-profit organizations are hybrids – private but with restricted ownership rights. This defining ‘nondistribution constraint’ reduces incentives to exploit underinformed customers and allows non-profits to depart from profit-maximizing behaviour, although costly enforcement of this constraint limits effectiveness. Non-profits' GDP share in the United States is about 30 per cent of the governmental non-defence share. Worldwide they employ about four per cent of the labour force. Non-profits receive public subsidies potentially justifiable by their provision of public goods. Sales of goods and services constitute the main source of non-profit revenues, but government grants and private donations are also important. Extensive research on the economic behaviour of non-profit, for-profit, and governmental organizations in mixed industries has disclosed systematic differences.

Keywords

agency problems; crowding out; free-rider problem; Lindahl prices; nondistribution constraint; non-profit organizations; price discrimination

Article

Variously termed voluntary, philanthropic and charitable, non-governmental, as well as non-profit, these organizations constitute a sizeable and growing share of economic activity. Non-profit organizations contribute some four per cent of GDP in the United States, up from three per cent in 1970 and two per cent during the Second World War, but their GDP share is about 40 per cent of that of government. In the social service sector where they predominate, non-profits account for some 20–25 per cent of outputs. There are at least 1.6 million non-profit organizations in the United States, a number that grew by 27 per cent during the decade 1994–004 alone. The United States is not alone in the prominence or growth of the non-profit sector. Figures gathered by the Johns Hopkins International Comparison Project reveal a ‘global associational revolution’, with paid and volunteer labour in non-profits involving an average of 4.4 per cent of the economically active population in the 35 countries studied, ranging from a

high of 14 per cent in the Netherlands to a low of 0.4 per cent in Mexico.

Non-profits are a form of institutional hybrid, combining attributes of profit-maximizing firms with those of government. Their organization and control are exercised through private initiatives rather than through the political process, and they cannot levy taxes. But, like government, they are constrained from distributing any profit or surplus to managers – the ‘non-distribution constraint’ (NDC). Many non-profits are granted a variety of tax subsidies such as eligibility for tax-deductible donations, special postal rates and exemption from taxation of income, property and sales. The NDC implies that non-profits cannot sell ownership shares, and so they can pursue social objectives other than profit maximization without fearing a hostile takeover. The NDC also implies that non-profits must rely for capital on sources other than equity shares. Thus, non-profits have both advantages and disadvantages in relation to private firms, with which they compete in industries such as health care (hospitals, nursing homes, hospices), education, and the arts.

Non-profits have provided public-type services, similar to those of government, for centuries. Jews created communal soup kitchens for travellers and collective charity funds for the needy in the second century bce. In 16th-century England, private ‘philanthropies’ were engaged in such wide-ranging social services as schools, hospitals, non-toll roads, fire-fighting apparatus, public parks, bridges, dykes and causeways, drainage canals, docks, harbour cleaning, libraries, care of prisoners in jails, and charity to the poor. In short, non-profits supplied the gamut of non-military goods and services that we identify today as governmental responsibilities.

Recent economic theorizing about the role of non-profits has examined both the nature of demand and the source of supply. Research on demand has two strands, one emphasizing failures of private markets, the other emphasizing governmental failures. In markets where valued attributes of the product are hard for consumers to observe and not verifiable by third parties, profit-maximizing firms can exploit their informational advantage. Alone, this outcome is inefficient, but the inefficiency is reduced if consumers deal with non-profit organizations. Non-profits have less incentive to short-change consumers because there are no shareholders or managers who can lawfully profit from this act. Nursing homes, day care for children, blood banks, medical research, environmental protection, and organizations claiming to aid the needy illustrate industries in which consumer information problems are not left to the private market alone. It is difficult for a nursing home patient or family member to determine whether the supplier is providing ‘tender loving care’; it is difficult to determine whether a day care centre is providing the attention that parents expect; and it is sometimes difficult for a patient or even a physician to determine the quality of blood available for a transfusion. Non-profits are a major force in all these industries characterized by informational asymmetries.

The quality assurance provided by the non-profit label, however, is limited. Enforcement of the NDC is spotty, and it is difficult to prevent distribution of profits in non-financial forms. Even when the NDC is well enforced, non-profits may short-change some under-informed customers in order to cross-subsidize missions that are not popular enough to generate direct donations. On the other hand, the sorting of entrepreneurs and consumers across ownership sectors, and the religious affiliation of many non-profits, enhances their credibility. The occasional failure of all these mechanisms is revealed, for example, by the collapse in 1995 of the Foundation for New Era Philanthropy following the revelation that the founder was benefiting from an illegal ‘Ponzi’ scheme in which colleges and universities, religious charities and individual donors were assured that their donations to the Foundation would be matched by a secret donor, resulting in grants that would double their ‘investment’ in three months.

Non-profits are a response to government failures as well as private market failures. The quantity and quality of outputs financed by government represent political decisions. Rather than setting individualized tax shares (Lindahl prices) to equate marginal benefits, governments use generalized systems of taxation, and so few voters get the quantity of governmentally provided goods that they would like, given the price each person confronts in the form of tax rates. Those who prefer less output and lower taxes have little recourse, but high demanders, who prefer more services at the tax prices they pay, and those seeking an alternative type or quality of service, may turn to non-profits to supplement governmental provision. Demanders of high-quality education, for example, often send their children to private non-profit schools. Non-profit schools also accommodate diverse minority demands relating to religion or educational philosophy. Thus, it is understandable that the United States, with a population unusually diverse in religion, culture, and ethnicity, has an unusually large non-profit sector, and not only in education.

Less is known about the *supply* of non-profit services. Non-profits are often created by churches and fraternal organizations, although over their life cycle they may disassociate from their founders. Religious organizations sometimes create health, education and welfare organizations as a way of attracting new adherents, binding the faithful, and meeting the moral obligations of their faith. In addition, religious and fraternal associations form communities of repeated interaction between like-minded individuals that help overcome free-rider problems and the transactions costs involved in creating a new organization. Nonetheless, the non-profit form solves an agency problem between the organization's founders and later donors. Thus, the organization's founder rationally chooses the non-profit form if the value of donor-supported public goods exceeds the value of the option to receive a share of future profits. But the entrepreneur may also be a profit-motivated organizer who sees the patina of a non-profit organization as little more than access to the subsidies and donations non-profits receive, and weak governmental enforcement of the NDC as providing opportunities for reaping greater personal financial gains than would be possible by founding a for-profit firm.

At any point in time, the number of non-profits is also affected by inter-sectoral conversions (particularly in the hospital industry), mergers, and tax and regulatory policies. Very little work has been done on the life cycles of non-profit organizations, but there is some evidence that non-profits are slower than for-profits to enter, expand, exit and contract. A limited supply of socially oriented entrepreneurs, an organizational preference for selectivity over expansion, non-profit inattention or incompetence, a preference to hold ‘excess capacity’ in case of medical emergencies, a reluctance to lay off employees, or differential capital constraints could explain these findings.

The non-profit form is far from a panacea; it, too, can fail. Revenue challenges abound. Non-profits cannot solve the free-rider problem because they cannot compel payments. The NDC encourages donations, but it also eliminates equity sales as a source of finance. Moreover, while the NDC reduces non-profits' incentives to take advantage of their patrons, it also reduces incentives for productive efficiency and responsiveness to changing market demand.

Non-profits rely on a mixture of revenue sources, varying greatly across industries, but sales of goods and services – especially tuition at colleges, patient fees at hospitals, and admission fees at museums, zoos and theatres – together with government grants and private donations are the three predominant sources. Research on donations has been extensive, examining the returns to fundraising expenditures and the efficacy of various fundraising mechanisms, as well as the effect of the charitable income tax deduction for private donors, the extent of crowding-out (or in) of private donations by direct

government expenditures or by governmental service contracts with non-profit organizations, and the determinants of time donations (volunteering). Laboratory and field experiments on fundraising techniques are proliferating, revealing, for example, the positive effects of raffle mechanisms and information disclosures on net funds raised. Studies of donations' crowd-out – the effect on donations of an exogenous change in other revenues – run the gamut from ‘near 100 per cent’ to ‘small but significant’ to ‘crowding in’ (that is, negative crowd-out).

Determining the full effects of any fundraising mechanism is complex. Government grants, for example, are not only a source of revenue but may also certify quality to other donors; private donations may be affected differently for persons who derive utility from their own giving to a non-profit – the ‘warm-glow’, private-good effect – and for those who do not, being indifferent between their own contribution and the same amount given by someone else; and giving of money and volunteering of time are still not clearly identified as complements or substitutes.

Sales of goods and services are the dominant overall source of non-profits' revenue, and they come in diverse forms. First, many non-profits sell services that constitute their charitable mission rather than simply generating revenue, as in tuition charged by non-profit schools or payments for health services by non-profit hospitals, whether paid by the consumer or some third party. This is often accompanied by price discrimination designed to generate revenue when doing so does not compromise the tax-exempt mission, while providing the service at low cost – even free – to the poor or other ‘deserving’ consumers. A second form of programme service revenue has become increasingly prominent around the world – governmental purchase-of-service contracts with non-profits for the delivery of social services. Finally, some non-profits derive income from sales of goods and service that are unrelated to their charitable purpose, denoted ‘unrelated business’ (UB) income. Thus, non-profit universities have become major sellers of computer software; non-profit hospitals have opened pharmacies, hotels, and fitness centres; and non-profit museums’ gift shops have become major purveyors of art objects.

Controversy surrounds the social-welfare impact of having tax-privileged non-profits competing with taxable for-profits in commercial markets – the ‘unfair competition’ issue. In addition, analysts disagree about the impact of UB commercial activity on the social missions of non-profits, as they do about interpretation of the fact that half of all non-profits engaged in UB activities report no profit at all.

The impact of each revenue source on non-profits' social mission remains an area of controversy. Conceptually, the links reflect the non-profit's need to satisfy the wants of whoever is providing revenue – consumers, corporations, governments, alumni, and so on – and the consequences of doing so for the non-profit mission. With that mission typically being quite general, there is concern about ‘mission creep’ – the mutating definition of mission so as to justify taking advantage of a new revenue source. All revenue sources pose this potential problem, but it is particularly acute for collaborative ventures between non-profit and for-profit organizations. The issue often arises between research universities and firms in the pharmaceutical and information sciences fields, but similar issues arise, for example, in non-profit hospital relationships with for-profit medical groups, and in food pantries' relationships with food manufacturers.

Financing non-profits involves more than monetary flows. A major resource for the non-profit sector is volunteer labour – another form of donation. Of trivial importance in the for-profit sector, volunteer labour is, in the United States, similar in value to the total amount of money donations, although controversy persists as to how such labour, with an explicit transaction price of zero, should be valued for various purposes – replacement cost to the organization, opportunity cost to the volunteer, and the

average market wage being the three prominent alternatives. Not counted in official labour force statistics, and generally overlooked as a contributor to output, volunteer labour in the United States equals about five per cent of the hours worked by the entire national labour force. Research on the supply of volunteer time indicates that it is affected by the same type of price, income and income tax rate variables that affect the supply of money donations. The identification of the separate effects of volunteer supply and organization demand, however, remains largely unstudied. The mission of non-profits may be more conducive to the use of volunteers than the profit-oriented goal of private firms. There is some evidence that even paid labour in non-profit organizations is partially volunteered, that is, workers accept a lower salary, in effect donating some of the opportunity cost of their time. However, differences between wages in the non-profit sector and in other sectors appear to be specific to particular industries and job titles.

Hundreds of studies compare the performance of non-profit organizations with similar organizations in other sectors, but severe methodological challenges remain. Reviewing the vast evidence on health care organizations with respect to economic performance, quality of care, and accessibility to unprofitable patients, Schlesinger and Gray (2006) note that some authors conclude there are no clear differences. However, they dispute this interpretation, arguing instead that the literature is persuasive that there are clear differences, but the extent and direction of such differences depend on the nature of the service provided, market conditions, and external constraints on behaviour.

Non-profit organizations sometimes convert to for-profit and vice versa, especially in three industries – hospitals, health maintenance organizations (HMOs) and higher education. When non-profits convert, they first sell their assets to the new for-profit entity, using the proceeds to create or support non-profit organizations with closely related charitable purposes. Controversy surrounds conversions because of the difficulty of establishing a fair market value for these assets, particularly in leveraged conversions by insiders. If the assets are sold too cheaply, the new owners receive windfall profits and the NDC is violated.

Both theory and quantitative evidence suggest that all forms of institutions – non-profits included – fail to be efficient or equitable under particular circumstances. The key public policy questions are: do non-profits behave in systematically different ways from proprietary organizations or governments? If so, under what conditions and in which realms of economic activity should each form be encouraged, mandated, discouraged or prohibited?

See Also

- agency problems
- altruism in experiments
- charitable giving
- health economics
- non-governmental organizations
- public goods

Bibliography

Ben-Ner, A. and Van Hoomissen, T. 1991. Non-profit organizations in the mixed economy: a demand and supply analysis. *Annals of Public and Cooperative Economics* 62, 519–50.

Boris, E.T. and Steuerle, C.E. 2006. Scope and dimensions of the non-profit sector. In Powell and Steinberg (2006).

Bilodeau, M. and Slivinski, A. 1998. Rational non-profit entrepreneurship. *Journal of Economics and Management Strategy* 7, 551–71.

Bilodeau, M. and Steinberg, R. 2006. Donative non-profit organizations. In *Handbook on the Economics of Giving, Altruism, and Reciprocity*, vol. 2, ed. S.C. Kolm and J.M. Ythier. Amsterdam: Elsevier.

Chakravarty, S., Gaynor, M., Klepper, S. and Vogt, W. 2006. Does the profit motive make Jack nimble? Ownership form and the evolution of the US hospital industry. *Health Economics* 15, 345–61.

Clotfelter, C. 1985. *Federal Tax Policy and Charitable Giving*. Chicago: University of Chicago Press.

Duncan, B. 1999. Modeling charitable contributions of time and money. *Journal of Public Economics* 72, 213–42.

Goddeeris, J. and Weisbrod, B. 2006. Why not for-profit? Conversions and public policy. In *Government and Non-profit Organizations: The Challenges of Civil Society*. Revised edn, ed. E.T. Boris and C.E. Steuerle. Washington, DC: Urban Institute.

Hansmann, H. 1980. The role of non-profit enterprise. *Yale Law Journal* 89, 835–98.

Hirth, R.M. 1999. Consumer information and competition between non-profit and for-profit nursing homes. *Journal of Health Economics* 18, 219–40.

James, E. 1983. How non-profits grow: a model. *Journal of Policy Analysis and Management* 2, 350–65.

James, E. 1986. The non-profit sector in comparative perspective. In *The Non-profit Sector: A Research Handbook*, ed. W. Powell. New Haven: Yale University Press.

Krashinsky, M. 1986. Transactions cost and a look at the non-profit organization. In *The Economics of Non-profit Institutions*, ed. S. Rose-Ackerman. New York: Oxford University Press.

Landry, C.E., Lange, A., List, J.A., Price, M.K. and Rupp, N.G. 2006. Toward an understanding of the economics of charity: evidence from a field experiment. *Quarterly Journal of Economics* 121, 747–82.

Leete, L. 2006. Work in the non-profit sector. In Powell and Steinberg (2006).

National Research Council. 2005. *Beyond the Market: Designing Nonmarket Accounts for the United States*. Panel to Study the Design of Nonmarket Accounts, K.G. Abraham and C. Mackie, eds., Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

Newhouse, J. 1973. Toward a theory of non-profit institutions: an economic model of a hospital. *American Economic Review* 60, 64–73.

Pauly, M. and Redisch, M. 1973. The not-for-profit hospital as a physician's cooperative. *American Economic Review* 63, 87–99.

Payne, A. 2001. Measuring the effect of federal research funding on private donations at research universities: is federal research funding more than a substitute for private donations? *International Tax and Public Finance* 8, 731–51.

Powell, W.W. and Steinberg, R. 2006. *The Non-profit Sector: A Research Handbook*, 2nd edn. New Haven: Yale University Press.

Ribar, D.C. and Wilhelm, M.O. 2002. Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy* 110, 425–57.

Robbins, K.C. 2006. The non-profit sector in historical perspective: traditions of philanthropy in the west. In Powell and Steinberg (2006).

Roomkin, M. and Weisbrod, B.A. 1999. Managerial compensation and incentives in for-profit and non-profit hospitals. *Journal of Law, Economics and Organizations* 15, 750–81.

Rose-Ackerman, S. 1982. Charitable giving and ‘excessive’ fundraising. *Quarterly Journal of Economics* 97, 195–212.

Rose-Ackerman, S. 1986. *The Economics of Non-profit Institutions*. New York: Oxford University Press.

Salamon, L.M., Sokolowski, S.W. and List, R. 2003. *Global Civil Society: An Overview*. Baltimore, MD: Johns Hopkins Center for Civil Society Studies.

Schlesinger, M. and Gray, B.H. 2006. Non-profit organizations and health care: some paradoxes of persistent scrutiny. In Powell and Steinberg (2006).

Steinberg, R. 1986. Should donors care about fundraising? In *The Non-Profit Sector: Economic Theory and Public Policy*, ed. S. Rose-Ackerman. New York: Oxford University Press.

Steinberg, R. 2006. Economic theories of non-profit organizations. In Powell and Steinberg (2006).

Tuckman, H.P. and Chang, C.F. 2006. Commercial activity, technological change, and non-profit mission. In Powell and Steinberg (2006).

Vesterlund, L. 2006. Why do people give? In Powell and Steinberg (2006).

Weisbrod, B.A. 1977. *The Voluntary Non-profit Sector*. Lexington, MA: D.C. Heath.

Weisbrod, B.A. 1988. *The Non-profit Economy*. Cambridge, MA: Harvard University Press.

Weisbrod, B.A. 1998. *To Profit or Not to Profit: The Commercial Transformation of the Non-profit Sector*. Cambridge: Cambridge University Press.

Weisbrod, B.A. 1998. Institutional form and organizational behavior. In *Private Action and the Public Good*, ed. W.W. Powell and E.S. Clemens. New Haven: Yale University Press.

Weisbrod, B.A. and Dominguez, N. 1986. Demand for collective goods in private non-profit markets: can fundraising expenditures help overcome free-rider behavior? *Journal of Public Economics* 30, 83–96.

White, M. 1981. *Non-profit Firms in a Three Sector Economy*. Washington, DC: Urban Institute.

Young, D. 1983. *If Not for Profit, for What*? Lexington, MA: D.C. Heath.

How to cite this article

Steinberg, Richard and Burton A. Weisbrod. "non-profit organizations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000164> doi:10.1057/9780230226203.1199

The New Palgrave Dictionary of Economics Online

non-standard analysis

Peter A. Loeb and Salim Rashid

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Non-standard analysis is an area of mathematics that provides a natural framework for the discussion of infinite economies. It is more suitable in many ways than Lebesgue measure theory as a source of models for large but finite economies since the sets of traders in such models are infinite sets which can be manipulated as though they were finite sets. The number system used to describe non-standard economies is an extension of the real numbers R; it is denoted by $^*R$. The set $^*R$ contains 'infinite natural numbers' and their multiplicative inverses, which are positive infinitesimals. It was with the development in 1960 of such a number system that Abraham Robinson (1974) solved an age-old problem by making rigorous the use of infinitesimals in mathematical analysis. Robinson gave a model-theoretic approach to his theory that is relevant to any infinite mathematical structure; that approach starts by listing the basic properties of the new number system. Before taking up this approach, however, it will be helpful to consider a simple nonstandard extension of the real numbers system that is constructed from sequences of real numbers.

The real numbers can be embedded in the set of sequences by associating a constant sequence $\{c_i\}$ with each real number c so that $c_i=c$ for all i. The relation on the set of sequences defined by setting $\{r_i\}>\{s_i\}$ if $r_i>s_i$ for an infinite number of indices i has the property that if $r_i=i$ and $s_i=1/i$ for all i, then $\{r_i\}>c$ and $c>\{s_i\}$ for any positive real number c. Here $\{r_i\}$ represents a positive infinite number and $\{s_i\}$ represents a positive infinitesimal. The relation>is not yet an ordering on a number system since, for example, if $t_i=0$ when i is even and $t_i=3$ when i is odd, then $\{t_i\}>2$ and $1>\{t_i\}$. To fix the situation, one forms an equivalence relation in the set of sequences. The above sequence $\{t_i\}$ for example should either be equivalent to the constant sequence 0 or the constant sequence 3.

An equivalence relation appropriate to the formation of a simple non-standard model of the real numbers from the set of real sequences is obtained by fixing a free ultrafilter U in the natural numbers N (i.e. a collection of subsets of N such that finite intersections of sets in U are in U, but the empty set and singleton sets are not in U, and if a subset of N is not in U, its complement is in U). Two sequences of real numbers $\{r_i\}$ and $\{s_i\}$ are equivalent if $r_i=s_i$ for all i in an element A of U. In this case, we say that $r_i=s_i$ almost everywhere. The equivalence classes form the non-standard real numbers $^*R$. As before, the

constant sequence $r_i=c$ represents the standard real number c, while the sequence $r_i=i$ represents an infinite element of $^*R$ and the sequence $r_i=1/i$ represents a nonzero infinitesimal. In general, a property holds for elements of $^*R$ if for representing sequences it holds on some set in U; one says that the property holds almost everywhere, or 'a.e.'. An element r of $^*R$ is finite if for some standard c in R, $|r_i|$ is smaller than c a.e., and r is infinitesimal if for every positive c in R, $|r_i|$ is smaller than c a.e.; the elements of $^*R$ that are not finite are called infinite.

Properties true for the real numbers are again true for $^*R$, but quantification over sets must be interpreted as quantification over 'internal' subsets of $^*R$. For our simple model, these subsets correspond to equivalence classes of sequences of subsets of R; the element of $^*R$ represented by $\{r_i\}$ is in the set represented by $\{A_i\}$ if and only if r_i is in A_i a.e. Not all subsets of $^*R$ are internal. Those that are not are called external. Some internal sets, called hyperfinite sets, have all of the formal properties of finite sets. Such a set A is represented by a sequence of subsets A_i from R with A_i finite a.e. The 'internal' cardinality of A is represented by the sequence $\{\mathrm{Card}(A_i)\}$ with 0 replacing infinite cardinals in the sequence. Thus, for example, if $A_i=\{1, 2,\ldots, i\}$, then A is the set of all non-standard natural numbers less than or equal to the infinite natural number γ where γ is represented by the sequence $\langle 1, 2, 3, \ldots\rangle$. The internal cardinality of A, which we denote by $|A|$ is in this case equal to γ.

In working with non-standard analysis, it is usually best to ignore any particular construction of non-standard models and think only of the properties they satisfy. For general applications, one starts with a set theoretic structure $V(S)$ where S is a set containing R and $V(S)$ consists of all the sets one can obtain from S in a finite number of steps using the usual operations of set theory. For example, the number 5 and the set of all Lebesgue measurable sets is in $V(S)$; so is the set of all Borel measures on R. Let L be a formal language for $V(S)$; L contains a name for each object in $V(S)$, variables, connectives (such as the symbols $\vee$ for 'and' and $\wedge$ for 'or'), quantifiers, brackets, and sentences formed with these symbols. The main result of Robinson's theory establishes the existence of a (not unique) structure $V(^*S)$ built from a set of individuals $^*S$ with the following properties; (1) Every name of an object in $V(S)$ names something of the same type (i.e. constructed with exactly the same operations) in $V(^*S)$ We write $^*A$ for the object in $V(^*S)$ with the same name as A in $V(S)$; A is called standard and $^*A$, the (nonstandard) extension of A. (2) (Transfer Principle) Every sentence in L that is true for $V(S)$ is true when interpreted in $V(^*S)$; quantification, however, is over 'internal' objects in $V(^*S)$. (3) If $A\in V(S)$ is a set, then there is a 'hyperfinite' set B which is a member of the extension $^*P_{\mathrm{F}}(A)$ of the set of all finite subsets of A such that for each $a\in A$, $^*a\in B$. Thus B contains the extension of each standard element of A.

The extension $^*s$ of an individual s is usually denoted by s instead of $^*s$; one thinks of a subset A of S as being imbedded in $^*A$. Internal objects in $V(^*S)$ are those objects which are members of the extensions of standard objects; the non-internal objects in $V(^*S)$ are called external. Any object that can be described in the formal language L using the names of internal objects is itself internal. The illuminating fact that the set N of finite natural numbers forms an external set in the non-standard natural numbers $^*N$ can be established as follows: If N were an internal set, then by applying the transfer principle to the theorem that every non-empty subset of N has a first element, it would follow that there is a first infinite element of $^*N$, that is, a first element of $^*N–N$, and thus a last element of N.

Hyperfinite sets are internal sets in internal one-to-one correspondence with an initial segment of $^*N$.

Such sets are useful in economics because they are treated like finite sets. To illustrate this fact and Property 3 above, we note that if γ is an infinite element of $^*N$, then the initial segment $T = \{n \in {}^*N : 1 \leq n \leq \gamma\}$ of $^*N$ is a hyperfinite set containing every element of *N*. Every formal property true for an initial segment of the natural numbers is true for the set *T*, whence *T* can be used as the set of traders in a 'hyperfinite economy'.

The set of non-standard real numbers $^*R$ contains infinite elements which are positive, infinite elements which are negative, and finite elements. Any finite element α of $^*R$ is infinitely close to a unique standard real number $a \in R$ that is, $\alpha - a$ is infinitesimal. We write $\alpha \approx a$ and also $\alpha - a \approx 0$ in this case; *a* is called the standard part of α. The standard part of α is denoted by st(α) or ${}^0\alpha$. The set of all points infinitely close to *a* is called the monad of *a*. A subset 0 of *R* is open if and only if for each point *x* in 0, the monad of *x* is contained in ^{*0}.

As an application of these ideas, we note that a real-valued sequence s_n (i.e. a mapping *s* from *N* into *R*) has a limit *l* if and only if for each infinite n, ${}^*s_n \approx L$ where $^*s_n$ is the image of *n* with respect to the non-standard extension $^*s$ of the function s. A real-valued function *f* defined on a subset *A* of *R* is continuous at a point $x \in A$ if and only if for all $y \in {}^*A$ with $y \approx x$, ${}^*f(y) \approx f(x)$; is uniformly continuous on *A* if and only if for all $y \in {}^*A$ and $z \in {}^*A$ with $y \approx z$, ${}^*f(y) \approx {}^*f(z)$. A subset *A* of *R* is compact if and only if for each $y \in {}^*A$ there is a standard *x* in *A* with $y \approx x$, whence *A* is compact if and only if it is closed and bounded. It is immediate that if *f* is continuous on a compact set *A* then *f* is uniformly continuous on *A*.

Brown and Robinson (1975) introduced the use of non-standard analysis in economics as a source of models for infinite exchange economies. The set of traders in non-standard economies is a hyperfinite set $T = \{1, 2, \ldots, \gamma\}$ where γ is an infinite element of $^*N$. The preferences and endowments are internal mappings defined on *T* analogous to the corresponding mappings in finite economies. Each trader's commodity endowment is an infinitesimal part of the market, and so that trader's influence on the formation of prices is infinitesimal but not zero. One can show in such economies, even without the usual convexity assumptions, that approximate competitive equilibria and approximate cores exist and that these cores can be approximately decentralized by the price system.

Given a hyperfinite set *T*, such as the set of traders in a non-standard economy, one can apply to *T* all of the combinatorial methods that are available for finite sets. For example Loeb (1973) obtained a form of the Lyapunov convexity theorem that is appropriate for the hyperfinite economies described above by applying a 'packing theorem' concerning a finite set of vectors in Euclidean space. Using another construction of P.A. Loeb (1975) one can form on *T* a standard measure space which is rich with structures inherited from the underlying point set. This construction starts by noting that the set *M* of all internal subsets of *T* forms an algebra in the usual sense. One obtains a finitely additive probability measure *P* on (*T*, *M*) by setting *P*(*A*) equal to the standard part of the ratio $|A| / |T|$ for each *A* in *M*. One may assume that any ordinary sequence $\{A_i : i \in N\}$ from *M* is the initial segment of an internal sequence $\{A_i : i \in {}^*N\}$ from *M*. This will be the case, for example, if the superstructure is constructed via an ultrafilter as indicated above. Now, if an ordinary sequence $\{A_i : i \in N\}$ from *M* is pairwise disjoint and $\cup A_i$ equals some element *A* in *M*, then all but a finite number of the A_i's are empty. (Extend $\{A_i : i \in N\}$

to $\{A_i : i \in {}^*N\}$ for every infinite $n \in {}^*N$ and therefore for some finite $n \in N$ A is contained in the union of the A_i's, $1 \leq i \leq n$.) The condition one checks to apply the Carathéodory extension theorem to (T, M, P) is thus vacuously satisfied. Therefore P has a σ-additive extension μ to the smallest σ-algebra $\sigma(M)$ generated by M. The space (T, $\sigma(M)$, μ) is a standard probability space which is very close in structure to the internal hyperfinite space (T, M, P).

Rashid (1979) first established the connection between the standard measure spaces that exist on hyperfinite economies and the models of infinite economies using Lebesgue measure. Measure spaces on hyperfinite economies have the great advantage of an underlying structure that closely parallels finite economies. This parallelism has been exploited by H.J. Keisler in his forthcoming work detailing the price adjustment processes in nonstandard exchange economies. Emmons (1984) has obtained results for economies using measure spaces on hyperfinite sets of traders that are not available for general measure space economies. Nonstandard economies also have the advantage of making readily apparent regularities in the asymptotic behaviour of large but finite economies. Anderson's (1978) core equivalence theorem, for example, was obtained by translating a result originally proved with nonstandard analysis. Similarly, a translation of Khan and Rashid (1982) produced the existence theorem of Anderson, Khan and Rashid (1982).

A further advantage inherent in the use of the number system ${}^*R$ in economics is the ability to distinguish behaviour on the finite part of ${}^*R$ from that on the infinite part and to distinguish different orders of infinities and infinitesimals. The first type of distinction was used by K.D. Stroyan (1983) to provide an elegant non-standard characterization of myopia in the evaluation of infinite consumption streams. The second distinction was central in Brown and Loeb's (1976) short, nonstandard proof of Aumann's theorem showing that the Shapley value of infinite economies under appropriate differentiability conditions coincides with the competitive equilibria.

See Also

- existence of general equilibrium
- large economies
- Lyapunov functions
- measure theory
- perfect competition

Bibliography

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46(6), November, 1483–87.

Anderson, R.M., Khan. M. Ali and Rashid, S. 1982. Approximate equilibria with bounds independent of preferences. *Review of Economic Studies* 49, 473–75.

Brown, D.J. and Loeb, P.A. 1976. The values of nonstandard exchange economies. *Israel Journal of Mathematics* 25(12), 71–86.

Brown, D.J. and Robinson, A. 1975. Nonstandard exchange economies. *Econometrica* 43(1), January, 41–55.

Emmons, D.W. 1984. Existence of Lindahl equilibria in measure theoretic economies without ordered preferences. *Journal of Economic Theory* 34(2), December, 342–59.

Khan. M. Ali and Rashid, S. 1982. Approximate equilibria in markets with indivisible commodities. *Journal of Economic Theory* 28(1), October, 82–101.

Loeb, P.A. 1973. A combinatorial analog of Lyapunov's Theorem for infinitesimally generated atomic vector measures. *Proceedings of the American Mathematical Society* 39(3), August, 585–6.

Loeb, P.A. 1975. Conversion from nonstandard to standard measure spaces and applications in probability theory. *Transactions of the American Mathematical Society* 211, 113–22.

Rashid, S. 1979. The relationship between measure-theoretic and non-standard exchange economies. *Journal of Mathematical Economics* 6(2), July, 195–202.

Robinson, A. 1974. *Non-standard Analysis*. Revised edn, Amsterdam: North-Holland.

Stroyan, K.D. 1983. Myopic utility functions on sequential economies. *Journal of Mathematical Economics* 11(3), July, 267–76.

How to cite this article

Loeb, Peter A. and Salim Rashid. "non-standard analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000092> doi:10.1057/9780230226203.1200

The New Palgrave Dictionary of Economics Online

non-substitution theorems

Neri Salvadori

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

joint production; non-substitution theorem; substitution

Article

A non-substitution theorem asserts that under certain specified conditions an economy will have one particular price structure for each admissible value of the profit rate, regardless of the pattern of final demand. The theorem has two forms. As first stated, it applies to an economy with single production and therefore no fixed capital (Arrow, 1951; Koopmans, 1951; Samuelson, 1951; Levhari, 1965). In its later formulation, some special joint products are considered to take account of fixed capital (Samuelson, 1961; Mirrlees, 1969; Stiglitz, 1970).

Consider first the single production form. The non-substitution theorem asserts that if (i) there exists one primary input (call it labour); (ii) all processes of production are perfectly divisible, with constant returns to scale, and have the same production period (this period is taken as the time unit for the analysis); (iii) each process produces one perfectly divisible commodity, making use of definite amounts of produced commodities and, perhaps, perfectly divisible labour; (iv) for each commodity there exists at least one process producing it; (v) labour is indispensable for the reproduction of commodities; (vi) the exchange of commodities takes place at the end of each period in fully competitive markets (that is the profit rate, the wage rate, and the price of each commodity are uniform); (vii) producers operate a process if and only if it is cost-reducing at current prices; then for each admissible value of the profit rate only one vector of relative prices (including the wage rate) is possible for the economy, so that relative prices are independent of demand.

In order to understand why the theorem works, let us denote with vector p and scalar w the equilibrium commodity price vector and the equilibrium wage rate, respectively, when the rate of profit equals r and the net output vector equals vector d. Hence, (a) no process is able to pay extra profit at prices p, wage rate w, profit rate r; (b) for each commodity there exists at least one operable process producing it (an operable processes is a process whose costs, including normal profits, are not larger than the price of the product); (c) operable processes can be operated in such a way to produce net output d.

The non-substitution theorem asserts that

1. (α) if the net output is $\hat{d} \neq d$ then *p* and *w* are still an equilibrium price-vector and an equilibrium wage rate;
2. (β) if more than one solution exists, they are characterized by the fact that price vectors and wage rates are respectively equal each other (if the same *numéraire* is utilized).

To prove statement (α) we just need to prove that operable processes can be operated in such a way to produce net output $\hat{d}$, since statements (a) and (b) hold. This is shown in the following way. Take one operable process for each commodity (they exist because of (b)) to arrange the material input matrix A and the labour input vector 1. It is easily shown that matrix (I–A) is invertible and $(I - A)^{-1} \geq 0$ where I is the identity matrix of appropriate size. Hence, statement (α) is a consequence of the fact that if the operation intensities of these processes are $\hat{d}^{T}(I - A)^{-1} \geq 0$ all the others being zero, then the net output vector equals $\hat{d}$. This procedure can also be utilized if a uniform growth rate not larger than *r* is assumed.
To prove statement (β) let (p_1, w_1) and (p_2, w_2) be the price vector and the wage vector relative to two equilibrium solutions, respectively. Similarly as in the proof of (α) we can arrange material input matrices A_1 and A_2 and labour input vectors 1_1 and 1_2 from the first and the second solution respectively. Hence,

$$p_1 = (1 + r)A_1 p_1 + w_1 l_1 \quad p_2 = (1 + r)A_2 p_2 + w_2 l_2.$$

Moreover, axiom (vii) requires that

$$p_1 \leqq (1 + r)A_2 p_2 + w_1 l_2 \quad p_2 \leqq (1 + r)A_1 p_2 + w_2 l_1$$

and since $[I - (1+r)A_i]$ is invertible and $[I - (1 + r)A_i]^{-1} \geq 0$ (i=1,2),

$$p_1 \leqq w_1 [I - (1 + r)A_2]^{-1/2} = (w_1 / w_2)\, p_2$$
$$p_2 \leqq w_2 [I - (1 + r)A_1]^{-1/1} = (w_2 / w_1)\, p_1 .$$

Thus,

$$p_1 = (w_1 / w_2) p_2.$$

Then, by introducing the common numeraire we obtain that

$$w_1 = w_2$$

and, therefore,

$$p_1 = p_2.$$

More recent formulations of the non-substitution theorem weaken the previously stated assumption (iii) to allow the introduction of some particular joint production cases. Assumptions are introduced to divide commodities into ‘final goods’ and ‘used machines’. Each process is assumed to produce one final good, but some joint products are allowed since used machines are produced jointly with final goods. Used machines are not transferable, that is an oven utilized once to produce bread cannot be utilized later to produce biscuits.

If machines are not used jointly, then a non-substitution theorem is stated as in the single production form. If machines can be used jointly, then the growth rate plays a role in determining prices and the wage rate, as does the profit rate. This fact has been recognized by Stiglitz (1970), who, however, failed to recognize that when this is so the uniqueness of the relative prices and wage rate does not need to hold even if prices are still independent of demand.

The label ‘non-substitution’ is appropriate to these theorems in so far as it assumed that there is only one scarce factor (primary input). Relaxation of any or all of the other assumptions, for example that of constant returns to scale, will mean that prices vary in response to changes in the structure of demand, but *will not* mean that there is ‘substitution’ in any meaningful sense.

In a neoclassical model prices are determined by the relation between demands (direct and derived) for endowment and the magnitude of the components of the endowment (typically conceived as stocks of factor services). The prices of produced commodities are equal to their costs of production, that is to the sum of rentals paid for the factor services used in their production. The possibility of *substitution* between factors, due either to substitution between commodities consumed or substitution in production, or to a combination of both, is the source of variation in derived demand, and hence in relative rentals. If, by comparison with a given situation, preferences were different, relative demands for factor services would typically be different, and hence their rentals and the prices of the commodities in the production

of which they are used would be different.

But, if there is only one factor of production, no substitution is possible whatever the composition of demand or the range of technical possibilities. Hence the relative prices of produced commodities will be determined by the least amounts of the single factor by means of which (directly and indirectly) they are produced. If, as is the case in the examples discussed above, the minimum cost technique is invariant to changes in demand, then prices too will not change as demand changes. If, however, the minimum cost technique does change as demand changes, say because of increasing returns to scale, then prices will change, *but this will not be due to any substitution.* There cannot be any substitution because there is only one factor. Similarly, in those cases of joint production in which a change in the structure of demand does lead to a change in relative prices, the change derives not from substitution between factors but from a change in the minimum cost combination of production processes.

See Also

- input–output analysis

Bibliography

Arrow, K.J. 1951. Alternative proof of the substitution theorem for Leontief models in the general case. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley.

Koopmans, T.C. 1951. Alternative proof of the substitution theorem for Leontief models in the case of three industries. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley.

Levhari, D. 1965. A non-substitution theorem and switching of techniques. *Quarterly Journal of Economics* 79, 98–105.

Mirrlees, J.A. 1969. The dynamic non-substitution theorem. *Review of Economic Studies* 36, 67–76.

Samuelson, P.A. 1951. Abstract of a theorem concerning substitutability in open Leontief models. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley.

Samuelson, P.A. 1961. A new theorem on non-substitution. In *Money Growth and Methodology*. Lund: CWK Gleerup .

Stiglitz, J.E. 1970. Non-substitution theorems with durable capital goods. *Review of Economic Studies* 37, 543–52.

How to cite this article

Salvadori, Neri. "non-substitution theorems." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000093> doi:10.1057/9780230226203.1201

The New Palgrave Dictionary of Economics Online

non-tariff barriers

John Beghin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Non-tariff barriers (NTBs) refer to the wide range of policy interventions other than border tariffs that affect trade of goods, services and factors of production. Most taxonomies of NTBs include market-specific trade and domestic policies affecting trade in that market. Extended taxonomies include macroeconomic policies affecting trade. NTBs have gained importance as tariff levels have been reduced worldwide. Common measures of NTBs include tariff equivalents of the NTB policy(ies), and count and frequency measures of NTBs. These NTB measures are subsequently used in various trade models, including gravity equations, to assess trade and/or welfare effects of the measured NTBs.

Keywords

antidumping; border effects; countertrade; domestic content requirements; gravity equations; nontariff barriers; price control; protection; quantity control; tariff versus quota; tariff-rate quotas; tariffs; technical barriers to trade; trade costs

Article

Nontariff barriers (NTBs) refer to the wide and heterogeneous range of policy interventions other than border tariffs that affect and distort trade of goods, services and factors of production. Common taxonomies of NTBs include market-specific trade and domestic policies such as import quotas, voluntary export restraints, restrictive state-trading interventions, export subsidies, countervailing duties, technical barriers to trade (TBTs), sanitary and phytosanitary (SPS) policies, rules of origin and domestic content requirements schemes. Extended taxonomies also include macro-policies affecting trade. No taxonomy can be complete since NTBs are defined as what they are not (Deardorff and Stern, 1998). This article is complemented by related articles on antidumping, border effects, countertrade, gravity equation, tariff versus quota, and trade costs. Deardorff and Stern (1998) suggest the following taxonomy with five categories.

A first broad category covers quantitative NTBs and similar restrictions. It includes import quotas and their administration methods (licensing, auctions, and other); export limitations and bans; voluntary export restraints, a limit on imports but managed by exporters; foreign exchange controls often based on licensing; prohibitions such as embargoes; domestic content and mixing requirements forcing the use of local components in a final product; discriminatory preferential trading agreements and rules of origin; and countertrade such as barter and payments in kind.

A second category covers fees other than tariffs, and associated policies affecting imports. This category includes variable levies triggered once prices reach a threshold or target level; advanced deposit requirements on imports, antidumping and countervailing duties imposed on landing goods allegedly exported ‘below cost’ or with the help of export subsidies provided by foreign governments; and border tax adjustment such as value-added taxes potentially imposed asymmetrically on imported and domestic competing goods.

A third category is extensive. It collects various forms of government policies including a wide set of macroeconomic policies. This category covers direct governmental participation and restrictive practices in trade, such as state-trading and state-sponsored monopoly and monopsony; government procurement polices with domestic preferences; and industrial policy favouring domestic firms with associated subsidies and aids. In addition, the category extends to macroeconomic and foreign exchange policies, competition policies, foreign direct investment policies, national taxation and social security policies, and immigration policies. Where to draw on the NTB definition is context-dependent.

Two better-targeted categories deal with customs procedure and administrative practices, and technical barriers to trade, which are central to NTBs. The former covers custom valuation methods that may depart from the actual import valuation; customs classification procedures other than the international harmonized system of classification to levy further fees; and customs clearance procedures such as inspections and documentation creating trading cost. Technical barriers to trade relate to health, sanitary, animal welfare and environmental regulations; quality standards; safety and industrial standards; packaging and labelling regulations and other media/advertising regulations. With the exception of export subsidies and quotas, NTBs have become more prominent than tariffs. Tariffs on manufacturing goods have been reduced to low levels through eight successive rounds of the World Trade Organization (WTO) and its predecessor, the General Agreement on Tariffs and Trade (GATT). As of 2005, the unweighted average tariff is roughly three per cent in high-income countries, and 11 per cent in developing countries according to the World Bank, from respective levels at least three times as high in 1980. Exports subsidies have almost disappeared except in a few agri-food markets. Quotas have become less important since they have been converted into two-tier tariff schemes, the so-called tariff-rate quotas. As tariffs have been lowered, demands for protectionism have induced new NTBs, such as TBT interventions. The United Nations Conference on Trade and Development (UNCTAD) estimates that the use of NTBs based on quantity and price controls and finance measures has decreased dramatically from slightly less than 45 per cent of tariff lines faced by NTBs in 1994 to 15 per cent in 2004, reflecting commitments made during the last round of WTO negotiations, the Uruguay Round. However, the use of NTBs other than quantity and price controls and finance measures increased from 55 per cent of all NTB measures in 1994 to 85 per cent in 2004. The use of TBTs almost doubled, from 32 to 59 per cent of affected tariff lines during the same period. The use of quantity control measures associated with TBTs showed a small increase, from 21 to 24 per cent of affected tariff lines, suggesting that trade impediments within TBTs are rising. Kee, Nicita and Olarreaga compute a 9 per cent tariff

equivalent of NTBs including price and quantity controls, finance measures, and TBTs, on average for all goods. The average tariff equivalent is about 40 per cent for the goods affected by these NTBs. Increased consumer demand for safety and environment-friendly attributes have also translated into an increase in the number of TBTs. Many NTBs are regulated by the WTO agreements that came out of the Uruguay Round (the TBT Agreement, SPS Measures Agreement, the Agreement on Textiles and Clothing), and articles of the original GATT among others. NTBs in service industries have recently become more important as trade in services has been expanding (Dee and Ferrantino, 2005).

Most NTBs are intrinsically protectionist whenever they do not address market failures such as externalities and information asymmetries between consumers and producers of goods being traded. Safety standards and labelling requirements are examples of the latter case. Some NTBs may restrict trade but improve welfare in the presence of negative externalities or informational asymmetries. Other NTBs can expand trade as they enhance demand and trade of a good through better information about the good or by enhancing the good's characteristics. Whether an NTB is protectionist is sometimes difficult to identify in the presence of market failure. If an NTB is equal to the measure that a social planner would implement for domestic purposes (that is, all firms are domestic firms or all agents belong to a single economy), that NTB is presumably non-protectionist (Fisher and Serra, 2000).

Measuring NTBs and their effects is a challenge, because of the heterogeneity of policy instruments and lack of systematic data. A unified approach to the measuring of NTBs does not exist. Most measurement methods start from a simple partial equilibrium approach looking at a single commodity, and attempt to develop a producer, consumer or trade tax equivalent to the NTBs, that explains by how much supply, and/or demand, or trade are affected by the policy intervention. Most NTB analyses implicitly rely on a framework that accounts for three economic effects: the regulatory protection effect providing rents to the domestic sector; the ‘supply shift’ effect, that reflects the increased costs of enforcing compliance of the NTBs on foreign and sometime domestic suppliers; and the ‘demand-shift’ effect, that takes into account the fact that a regulation may enhance demand with new information or by reducing an externality.

The measurement of an NTB is hard to disentangle from the measurements of its effects on market equilibrium and trade. Most NTB measures and analyses focus on the increase in the price of imports resulting from the NTB, the resulting import reduction, the change in the price responsiveness of the demand for imports, the variability of the effects of the NTB, and the welfare cost of the NTB (Deardorff and Stern, 1998; Dee and Ferrantino, 2005).

Several NTBs based on a price intervention (for example, export subsidies, countervailing duties), are a tax instrument. More complex NTBs can sometimes be represented by a set of taxes, such as in the case of a domestic content requirement (Vousden, 1990). These NTBs can be analysed as such taxes. To develop a tax equivalent, a basis of equivalence has to be chosen (Vousden, 1990). The tax equivalent has to lead to either an equivalent protection level (same profit under the tax equivalent or the NTB), or to a price increase equivalence (a price wedge), or to consumption, production or trade equivalent. This choice of basis depends on the intended policy analysis.

However, many NTBs do not easily translate into a tax-equivalent instrument. They require more sophisticated and indirect approaches to be measured and to quantify their effects on import volume, price, and welfare. Roundabout approaches are also used because of lack of data on the direct implications of an NTB on cost of production and consumer decisions (Beghin and Bureau, 2001).

Common measurement approaches of NTBs

The price-wedge method measures the impact of an NTB on the domestic price of a good in comparison to a reference price, often the border price of a comparable good. The aim of this method is to derive a tariff/tax equivalent to the NTB as discussed above, and use the tariff/tax equivalent in further analysis that measures implications of the NTB on resource allocation in the given markets affected by the NTB. Deardorff and Stern (1998) provide price-wedge equivalent formulas for an extended coverage of NTBs.

Conceptually, the measure compares the domestic price that would prevail without the NTB to the domestic price prevailing in the presence of the NTB, on the assumption that the price paid to suppliers remains unchanged. However, these prices are practically unobservable. Implementations of the price-wedge measure of an NTB compare the domestic and foreign prices of comparable goods in the presence of the NTB accounting for tariffs, transportation costs, and other known and observed trading costs. Adjustments can be made to recover a price estimate that would prevail in the absence of the NTB, using observed levels of quantities and prices, and own-price elasticities of demand and supply and imported goods.

The price-wedge method has several drawbacks. First, if several NTBs are jointly in place, the price-wedge measures the price effect of these policies without being informative about their respective contributions or even their nature. Second, quality differences are hard to account for precisely although they are a pivotal element of the price-wedge computation. The price-wedge estimate of an NTB is usually sensitive to the assumptions made on the substitution between the imported and domestic goods. This method has also some limitations in large empirical studies for which data are aggregated, resulting in loss of information on quality differences between import and domestic comparable goods. Finally, trading costs may be present but not accounted for and the price-wedge method may falsely attribute these trading costs to a NTB.

Inventory-based frequency measures count the number or frequency of regulations and barriers present in a given market. They are used in both quantitative and qualitative assessments of the incidence of the NTBs. Common measures include the number of regulations and policies, which can be further elaborated to indicators such as the number of pages of national regulations. Frequency of trade detentions at borders is also used, and so are survey-based frequency and number of complaints reported by exporters for perceived discriminatory regulatory practices.

When implemented, quantitative estimates often rely on catalogues of technical barriers (identification and description) using datasets such as UNCTAD's TRAINS data-set. Measures include simple frequency of occurrence of NTBs, frequency ratios for product categories subject to an NTB, and coverage ratio based on the value of imports of products within a category subject to the NTB, expressed as a share of import value of the corresponding category. Relative measures can also be developed comparing the latter frequency measures in a given country with respect to accepted international norms or best practice, for example, for the SPS or food safety regulations. Alternatively, frequency measures can be compared across commodities or across countries to identify large deviations from average frequencies, flagging potential protectionist issues.

NTBs vary in importance across sectors and products. Even for a given NTB type, its effects may vary across products. A major drawback of the frequency measures is that a correlation between the number of NTBs and their effect on trade and welfare may be low in absolute value. International data-sets on

NTBs inventories may also suffer from uneven reporting by countries and heterogeneous coverage of measures across countries and commodities. Survey-based measures focus on effective barriers rather than just an NTBs count. However, they may suffer from various reporting biases as surveys and respondents are often motivated by mercantilism to facilitate exports by the responding exporters. Frequency measures do not identify the trade restrictiveness of NTBs but can be used in gravity equations to identify the effects of NTBs on trade flows. When trying to quantify NTBs, an obvious technique is to consider the forgone trade that cannot be explained by tariffs and known trading costs. The NTB frequency measures, or in certain cases the level of standards themselves, can help identify the trade effects of these NTBs. Provided there is enough variability across countries or over time in the measure (for example, the level of toxic residues) they can explain the variation in trade flow not explained by other explanatory variables included in the gravity equation (respective incomes of trading countries, distance, tariff, and other variables measuring border effects).

Gravity-equation techniques attempt to measure the trade impact of NTBs, not their welfare impact, and may therefore ignore some of the beneficial effect of the regulations that correct negative externalities but restrict trade. NTBs are appropriate if trade is the vector of negative externalities such as unsafe food imports or pest-infested imports. In addition, the direction of the effect of the ‘NTB’ variable on trade flows in the regression is not constrained. It is possible to capture a trade or demand-enhancing effect of regulations and standards. This enhancement occurs when the NTB facilitates trade and induces consumers to consume more of a product although the product's price is higher because of the NTB. Such expansion through standards has been observed in OECD food trade (Disdier, Fontagné and Minouni, 2006).

Risk assessment approaches and scientific knowledge can contribute to gauging a subset of NTBs, especially safety and SPS standards and regulations. The latter approach can contribute to assessing the welfare effects and the potential protectionism of these types of NTBs. Scientific knowledge can determine if a regulation is science-based or not, or if a risk simply does not exist or is negligible. This criterion is used by the WTO in its assessment of TBT and SPS regulations. Cost–benefit calculations combined with risk assessment provide expected cost and benefits of such types of NTBs. Risk-assessment measures provide an economic criterion to gauge the desirability of an NTB and its likely protectionist nature if externalities are small and if its costs greatly exceed its benefits in expected terms. The combined use of scientific knowledge and cost–benefit assessment of an NTB is a demanding process suitable for a detailed analysis of a specific case study, rather than for large-scale multi-market analyses. Another limitation of this approach is the partial knowledge of health, environmental and other risks associated with trade and their economic significance.

NTBs measures are an essential step to computing the welfare effects of the NTBs. Beyond welfare effects, these measures are also useful for policy purposes. WTO disputes frequently arise, alleging that some NTBs impede trade more than necessary to achieve some legitimate objective, or that they are just protectionist. These NTB measures are used in the formal dispute process to estimate export market losses and price-lowering effects of the incriminated policy.

See Also

- antidumping

- border effects
- countertrade
- gravity equation
- tariff versus quota
- trade costs

Bibliography

Beghin, J. and Bureau, J.-C. 2001. Quantitative policy analysis of sanitary, phytosanitary and technical barriers to trade. *Economie Internationale* 87, 107–30.

Deardorff, A.V. and Stern, R.M. 1998. *Measurement of Nontariff Barriers: Studies in International Economics*. Ann Arbor, MI: University of Michigan Press.

Dee, P. and Ferrantino, M. 2005. *Quantitative Methods for Assessing the effects of Non-Tariff Measures and Trade Facilitation*. Singapore: APEC Secretariat and World Scientific.

Disdier, A.-C., Fontagné, L. and Minouni, M. 2006. The impact of regulations on agricultural trade: evidence from SPS and TBT agreements. Working paper. Paris: CEPII.

Fisher, R. and Serra, P. 2000. Standards and protection. *Journal of International Economics* 52, 377–400.

Henson, S. and Wilson, J.S. 2005. *The WTO and Technical Barriers to Trade*. Critical Perspectives on the Global Trading System and the WTO Series. Northampton, MA: Edward Elgar.

Kee, H.L., Nicita, A. and Olarreaga, M. 2006. Estimating trade restrictiveness indices. Policy Research Working Paper No. 3840. Washington, DC: World Bank.

UNCTAD (United Nations Conference on Trade and Development). 2005. Methodologies, classifications, quantification and development impacts of non-tariff barriers. Note by the UNCTAD secretariat, document TD/B/COM.1/EM.27/2, 23 June. Geneva.

Vousden, N. 1990. *The Economics of Trade Protection*. Cambridge: Cambridge University Press.

How to cite this article

Beghin, John. "non-tariff barriers." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000169> doi:10.1057/9780230226203.1202

The New Palgrave Dictionary of Economics Online

North American Free Trade Agreement (NAFTA)

Gordon H. Hanson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The North American Free Trade Agreement (NAFTA) eliminated trade barriers on most products between Canada, Mexico, and the United States. NAFTA included provisions to remove restrictions on cross-border investment, expand service trade, and address environmental and labour standards. Post-NAFTA increases in trade between member countries were matched by comparable decreases in their trade with the rest of the world. Freer trade has brought a shift in economic activity within Mexico and the United States towards their shared border and an increase in direct investment from the United States to Mexico. In Mexico these developments have contributed to greater wage inequality.

Keywords

Canada–United States Free Trade Agreement; computable general equilibrium models; foreign direct investment; international migration; North American Free Trade Agreement (NAFTA); regional and preferential trade agreements; rules of origin; tariffs; trade diversion; trade policy, political economy of; World Trade Organization

Article

The North American Free Trade Agreement (NAFTA), which entered into force in 1994, eliminated trade barriers on most products between Canada, Mexico, and the United States. The agreement culminated a decade of liberalization in North America, which included the Canada–United States free trade agreement in 1989 and Mexico's joining the General Agreement on Trade and Tariffs (GATT) in 1986. (On the impact of the Canada–United States Free Trade Agreement, see Trefler, 2005.) For Mexico, NAFTA was the final step in reversing four decades of protectionist trade policies. For Canada and the United States, NAFTA completed three decades of promoting closer economic ties.
Upon its implementation, NAFTA eliminated tariffs on goods accounting for approximately one-half of trade between the three countries. Tariffs on other goods (primarily those with relatively high pre-

NAFTA tariffs) were phased out over 5 to 15-year periods. Among the industries with the slowest tariff phase-outs were Canadian textiles, Mexican corn and US sugar. A few industries (mainly in agriculture, energy and services) were excluded from NAFTA altogether. The agreement incorporates stringent rules of origin, which apply a country's external tariff to NAFTA imports whose North American content (in terms of the share of value added) fails to meet mandated thresholds (Estevadeordal and Suominen, 2005). Rules of origin prevent three-way trade in which, say, Canada imports a good from Japan at an external tariff that is below that for Mexico and then re-exports the good to Mexico at a zero NAFTA tariff. If allowed, such trade would effectively impose a common external tariff across North America equal to the minimum tariff for each good among the three countries. Content requirements vary across sectors, with those for the auto industry being among the highest.

NAFTA was broad in its scope and included provisions for removing restrictions on cross-border investment between member countries, expanding service trade and protecting intellectual property. A novel feature of the agreement was the adoption of side accords for environmental and labour standards, which created a mechanism under which citizens of member countries can adjudicate disputes over the violation of standards (which in their essence state that NAFTA members are obliged to uphold environmental and labour laws that each has on its books). While the standards were controversial at the time of NAFTA's passage, few cases of significance involving environmental or labour infractions have been resolved under the agreement.

The economic rationale for creating a regional free trade area is that it eliminates price distortions caused by tariffs, quotas and other policy barriers, which induce countries to allocate too many resources to import-competing industries and too few resources to exporting industries. In the early 1990s, results from computable general equilibrium models suggested that NAFTA would raise welfare by an amount equal to between two and four per cent of GDP in Mexico and one per cent or less of GDP in Canada and the United States (Brown, Deardorff and Stern, 1992). Low estimated gains from trade associated with NAFTA are not surprising, given that prior to the agreement Canada and the United States already had a free trade agreement in place, Canadian and US external tariffs on most products were already quite low, and Mexico had begun to unilaterally liberalize its trade following its joining the GATT.

While a free trade area creates trade between member countries, it also diverts trade between the trade bloc and the rest of the world. Between 1993 and 2004, trade between Canada, Mexico and the United States increased by 2.6 times in real terms; over the same period, trade between NAFTA countries and the rest of the world increased by only 1.9 times (Hufbauer and Schott, 2005). In sectors that had the highest protection prior to NAFTA, nearly all of the increase in trade within the NAFTA region was matched by comparable decreases in trade between NAFTA members and the rest of the world (Romalis, 2005), consistent with the agreement causing trade diversion.

Even where the net change in income associated with freer trade is small, gross changes in income for particular groups may be large. Because trade agreements redistribute income, they tend to provoke political conflict. The politics surrounding NAFTA were perhaps most contentious in the United States. President Clinton's support for NAFTA became an issue in the 1996 US presidential campaign, with opposition candidate Ross Perot memorably claiming that increased trade with Mexico would create a ‘giant sucking sound’ as US jobs moved across its southern border.

In the United States, one would expect groups allied with labour to oppose freer trade with a low-wage country and groups allied with capital-intensive industries to support it. NAFTA was narrowly approved by the US Congress, with its outcome uncertain until the final hour. Consistent with standard models of

political economy, US politicians receiving donations from labour groups tended to vote against NAFTA, while those receiving donations from business groups tended to vote for NAFTA (Baldwin and Magee, 2000). There was also a regional dimension to NAFTA's politics, with US politicians representing districts near the US border with Mexico being much more likely to support the agreement. This in part reflects the fact that, as US trade with Mexico has expanded, US border states have seen their manufacturing and trade-related industries grow relative to the rest of the country (Hanson, 2001).

In Mexico, also, NAFTA has had a varied regional impact. Following Mexico's opening to trade, Mexican states near the US border have had high growth in manufacturing employment, exports, and foreign direct investment (FDI) relative to the rest of the country. The shift in economic activity towards Mexico's border region has increased regional income differences in the country, which had been declining until Mexico began to liberalize trade (Chiquiar, 2005).

Economic theory suggests that trade may either complement or substitute for factor flows, depending on the magnitude of transport costs, fixed production costs and cross-country differences in technology and factor supplies. Following NAFTA, there has been a substantial increase in FDI by the United States in Mexico. Much of the FDI has involved US multinational firms setting up export assembly plants, known as *maquiladoras*, in Mexico. FDI in assembly plants has resulted from US firms outsourcing production to Mexico and has created substantial intra-industry trade flows in which the US exports parts and components to Mexico, and Mexico exports finished goods back to the United States. Similar trade patterns have existed between the United States and Canada since the 1960s, when the two countries liberalized trade in the auto industry. By moving labour-intensive production activities out of the United States, NAFTA has decreased the relative demand for less skilled labour in the country. And by moving capital, technology and new production operations into Mexico, NAFTA has increased the relative demand for skilled labour in Mexico. Thus, United States–Mexico economic integration appears to have contributed to a widening of the skilled–unskilled wage gap in both countries (Feenstra and Hanson, 1997).

At the time of NAFTA's signing, the agreement was touted as a means of reducing United States–Mexico wage differences and the incentive for workers in Mexico to migrate to the United States. By the 1980s, Mexico–United States migration had become an important policy issue on both sides of the border. NAFTA was justified in part as a way to reduce international migration flows. However, since NAFTA's implementation there has been an increase rather than a decrease in the flow of labour from Mexico to the United States (Hanson, 2006). At least some of the increased migration appears associated with the collapse of the peso in 1994 and the ensuing economic contraction in Mexico (Hanson and Spilimbergo, 1999). Partly as a result of the peso collapse, the difference in per capita income between the United States and Mexico was larger in 2002 than in 1990 (Tornell, Westermann and Martinez, 2003). Other evidence suggests that, whatever its long-run effects, NAFTA may have contributed to a transitory increase in Mexico-to-United States migration. By contributing to gross job destruction in agriculture and manufacturing, NAFTA may have displaced workers who then migrated to the United States.

For Canada and Mexico, NAFTA helped increase the importance of the US economy for their economic development. For the United States, NAFTA was a milestone in the country's approach to trade policy. Since 1994, the United States has concluded bilateral trade agreements with a dozen other countries, but has not succeeded in helping complete a multilateral trade agreement under the auspice of the World

Trade Organization. One interpretation of this pattern is that NAFTA signalled a shift in US trade policy away from multilateralism and toward bilateralism, perhaps weakening multilateral trade institutions in the process.

See Also

- banking crises
- computation of general equilibria
- factor prices in general equilibrium
- foreign direct investment
- new economic geography
- regional and preferential trade agreements
- supply chains

Bibliography

Baldwin, R. and Magee, C. 2000. Is trade policy for sale? Congressional voting on recent trade bills. *Public Choice* 105, 79–101.

Brown, D., Deardorff, A. and Stern, R. 1992. North American integration. *Economic Journal* 102, 1507–18.

Chiquiar, D. 2005. Why Mexico's regional income convergence broke down. *Journal of Development Economics* 77, 257–75.

Estevadeordal, A. and Suominen, K. 2005. Rules of origin in preferential trading arrangements: is all well with the spaghetti bowl in the Americas? *Economia* 5, 63–9.

Feenstra, R. and Hanson, G. 1997. Foreign direct investment and relative wages: evidence from Mexico's maquiladoras. *Journal of International Economics* 42, 371–94.

Hanson, G. 2001. U.S.–Mexico integration and regional economies: evidence from border-city pairs. *Journal of Urban Economics* 50, 259–87.

Hanson, G. 2006. Illegal migration from Mexico to the United States. *Journal of Economic Literature* 44, 869–924.

Hanson, G. and Spilimbergo, A. 1999. Illegal immigration, border enforcement and relative wages: evidence from apprehensions at the U.S.–Mexico border. *American Economic Review* 89, 1337–57.

Hufbauer, G. and Schott, J. 2005. *NAFTA Revisited: Achievements and Challenges*. Washington, DC: Institute for International Economics.

Romalis, J. 2005. NAFTA's and CUSFTA's impact on international trade. Working Paper No. 11059. Cambridge, MA: NBER.

Tornell, A., Westermann, F. and Martinez, L. 2003. Liberalization, growth, and financial crises: lessons from Mexico and the developing world. *Brookings Papers on Economic Activity* 2003(2), 1–88.

Trefler, D. 2005. The long and the short of the Canada–U.S. Free Trade Agreement. *American Economic Review* 94, 870–95.

How to cite this article

Hanson, Gordon H. "North American Free Trade Agreement (NAFTA)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000145> doi:10.1057/9780230226203.1203

The New Palgrave Dictionary of Economics Online

North, Douglass Cecil (born 1920)

Avner Greif

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A pioneer of the New Institutional Economics, Douglass North has built upon the property rights and transaction cost approaches of Coase and others to explain economic growth in terms not of changes in technology and productive factors but of institutional and organizational change. His most recent work stresses the need to integrate insights from cognitive science into the examination of the interplay among belief systems, institutions, and economic performance. Institutions reduce the uncertainties that would otherwise overwhelm cognitive capacity in complex social situations, but the resulting bias in our beliefs can lead to the persistence of inefficient institutions.

Keywords

Alchian, A.; Cheung, S.; Cliometric Society; cliometrics; Coase, R.; Commons, J.; Demsetz, H.; feudalism; Fogel, R.; growth and institutions; Hayek, F.; industrial revolution; Institutionalism, old; Kuznets, S.; Mitchell, W.; new economic history; new institutional economics; North, D.; patents; peasants; property rights; social norms; Society for the New Institutional Economics; stagnation; state, theory of; tax transaction costs; technical change; West, the

Article

A native of Cambridge, Massachusetts, who was born in 1920, North received his undergraduate and doctoral degrees from the University of California, Berkeley. His 1952 Ph.D. dissertation focused on the history of the American insurance industry. Most of his professional career has been spent at two institutions: the University of Washington in Seattle and, from 1983, Washington University in Saint Louis. North was among the founders of cliometrics (also known as the New Economic History). Later he was a pioneering researcher of the New Institutional Economics. In 1993 North and Robert Fogel shared the Nobel Prize in Economics. In 1997–8 he served as the first president of the Society for the New Institutional Economics.

North's publications are numerous and space limitation precludes presenting all of them or even doing justice to particular ones. Accordingly, the following discussion of his books highlights some of North's main contributions.

Neoclassical analyses

North initially studied American economic growth. Under the influence of Simon Kuznets, he compiled the first quantitative historical series of the US balance of payments. This work, in conjunction with his studies of regional development, led to his first published book in 1961, *The Economic Growth of the United States, 1790–1860*. In this book North developed an export-based growth model to argue that the expansion of one sector (cotton plantations) in the United States stimulated development in other sectors, and led to specialization and interregional trade.

By relying on economic theory and quantitative analysis, this line of work contributed to the rise of cliometrics (or the New Economic History). In contrast to traditional economic historians who relied on narratives and non-qualitative analysis, cliometrics combines economic theory, quantitative methods, hypothesis testing, counterfactual analysis, and traditional techniques of historical analysis to explain economic outcomes, evaluate and develop economic theories, and deepen our historical knowledge. North further fostered this development by helping to found the Cliometric Society and serving as co-editor of the *Journal of Economic History* for five years.

Towards institutional analysis

In the late 1960s North began expanding his analysis of economic growth beyond the confines of neoclassical economics by considering the importance of organizational changes to increasing efficiency. In his 1968 article on productivity in overseas shipping, North argued that organizational changes had more important effects than technological changes in reducing transportation costs between 1600 and 1850. Market integration and growth followed due to organizational rather than technological changes.

More generally, North began to emphasize that, in order to understand growth, one had to go beyond the neoclassical framework, which at that time attributed growth to changes in technology and factors of production. In sharp contrast, North argued that changes in technology and factors of production are not the sources of growth but, in fact, constitute growth. This implies that, to understand growth, we must examine the forces that cause beneficial technological changes and increase the utilization of factors of production. North argued that institutions constitute such forces and his subsequent research focused on them.

In developing the analysis of the relationship between institutions and economic growth, North built on and expanded the property rights and transaction costs approaches advanced by Ronald H. Coase, Armen A. Alchian, Steven N.S. Cheung, Harold Demsetz, and others. His subsequent books were ambitious attempts to place institutions at the centre of economic growth analysis. Good institutions promote growth by bringing private return from economic activities closer to their social return. Economic growth transpires in response to low-cost enforcement of contracts when property rights are secured and when governments pursue growth-oriented policies rather than prey on the wealth of their subjects. Institutions that achieve these goals encourage technological innovations, foster capital

accumulation, and increase labour input. Growth follows as technology improves, capital accumulates, and specialization occurs.
Institutions in the Northian framework consist of rules and regulations which, together with their enforcement mechanisms, determine the incentives faced by economic agents. Similar focus on the relationships between institutions and economic outcomes has also been the hallmark of old institutionalism (associated with such scholars as John R. Commons, Friedrich A. von. Hayek, and Wesley C. Mitchell). Old institutionalism, however, considered institutions either as exogenous and immutable or as reflecting spontaneous, uncontrolled processes. In contrast, North attempted to consider institutions as endogenous and to understand the forces that shaped their development. To do this, he particularly concentrated on the state as setting the rules of the economic game.

Institutions and American growth

North's first book on this issue, *Institutional Change and American Economic Growth*, was co-authored with Lance Davis and published in 1971. It outlines a theoretical perspective on the role and dynamic of institutions. The main theoretical assertion is that new institutions – specifically, new property-rights assignments – arise when groups in the society perceive that there are opportunities for profit that cannot be consummated given the existing institutions, but that would be feasible if these institutions were changed. Perceptions of benefits of institutional change and the details of the political system are what determine whether socially beneficial institutional change will transpire.
The book demonstrates the merits of this assertion by considering growth in the United States during the 19th century. It advances a new interpretation of American economic growth as one that reflects the pursuit of profit opportunities by economic agents through changing politically determined rules. Commodity markets expanded, for example, because canals reduced transportation – and hence transaction – costs. Investments in canals, however, didn't occur automatically. Public investment, state-mandated changes in property rights, and changes in perceptions of the profitability of these investments were prerequisites. Similarly, political decisions and changes in property rights were crucial to other factors that directly contributed to growth: the evolution of capital markets, the rise of large corporations and of the manufacturing sector, investment in human capital and the expansion of service industries. Institutional evolution was central to American economic growth. More generally, North's work illustrates that in order to understand economic growth, the evolution of laws and regulations governing property rights must first be analyzed. Changes in property rights are often required before individuals and societies can gain from increasing the scale and efficiency of production and exchange.

Institutions and the rise of the West

A subsequent book (co-authored with Robert Paul Thomas) published in 1973, *The Rise of the Western World: A New Economic History*, further applied these ideas to explain the performance of various western European economies. By examining economic outcomes from the feudal period to the Industrial Revolution, the book sought answers to two questions. First, do differences in institutions account for patterns of economic growth and stagnation in European economies, and does the rise of the West reflect the efficiency of its property-rights regime? Second, what determines whether more or less efficient institutions will prevail?

The book argues that patterns of growth and stagnation in Europe reflect whether property rights were assigned efficiently and secured. The feudal system ended in economic stagnation and crises because of the misallocation of property rights to land. Peasants had few incentives to increase land productivity because they did not own it. Later, the Dutch Republic and England outpaced Spain and France because their property right's assignments were better designed to close the gap between private and social rates of return from economic activities. England's rising technological superiority, for example, reflected its effective system of patenting. In the long run, other European economies adopted similarly efficient systems of property rights.

Two forces determine institutions' relative efficiency. Institutions' degrees of efficiency respond to changes in relative prices, which, in turn, are due to changes in population and technology. As the relative price of a factor of production increases, property rights will be altered to better align incentives. The collapse of the state in medieval Europe rendered protection a valuable commodity. The feudal system, in which specialists in protection held property rights to land, reflected the relatively high value of protection. The large decline in the European population during the 14th century, however, increased the relative price of labour. This undermined the feudal system, and property rights in land were transferred to the peasants who toiled on it.

The tendency towards efficiency in institutional change is countered by the transaction costs of tax collection. Specifically, a ruler assigns property rights in a manner that maximizes his net revenue rather than efficiency. The transaction costs of tax collection place a wedge between efficient property right regimes and those that are optimal to a ruler. France's geographical scale and diversity, for example, implied that the taxation regime that was optimal to its rulers entailed a high efficiency cost. France's economic growth therefore lagged behind England's.

Given the importance of the state in this analysis, North advanced a theory of the state in his 1981 book, *Structure and Change in Economic History*. He departed from the common view of the state as an efficiency-enhancing social contract aimed at increasing security or providing other public goods, and characterized it as a ruler-predator, utilizing a bargaining framework, to consider a ruler's relationship with his subjects. In a state, citizens contract with a specialist in enforcement to provide them with protection. The terms of the deal – the extent of absolutism and predation – reflect the relative bargaining power of these parties, which, in turn, depends on such factors as military technology and the threat of entry by competing rulers. This analysis contributes to the argument that interstate competition within Europe was growth-enhancing by emphasizing that this competition may have constrained predation by rulers.

The 1981 book is also a departure from North's previous lines of analysis in that it focuses on ideology. North's previous writings noted the importance of informal institutions, such as ideology, social norms, and values, but they had not been explicitly integrated into the analysis. This book, however, claims that ideology develops as a justification for existing institutions and hence it is both endogenous to institutions and a strengthening factor. Although North's earlier analyses were rooted in history, they developed an ahistorical theory of institutions. These analyses sought a deterministic theory of institutions: a mapping from exogenous, contemporaneous conditions (such as population, technology, and geography) to institutions. Subsequently, North developed a more elaborate view of institutional change that attempted to capture how past institutions influence ensuing ones.

Recent theoretical developments

North's 1990 book, *Institutions, Institutional Change and Economic Performance*, develops an historical theory of institutional change. The argument revolves around the interplay between organizations and institutions. Institutions provide the incentives for establishing some organizations – for example, firms or political interest groups – but not others, and influence their activities. Through such activities, the organizations that institutions promote acquire new knowledge and information. This new knowledge enables them to recognize how they can improve their ability to advance their interests through institutional change. Therefore, these organizations act as players in the politics of setting the rules that govern economic interactions. Hence, institutional change is a path-dependent process. Institutions induce the emergence of particular organizations which later engage in institutional change. Such changes are incremental because organizations don't set out to destroy the institutions that gave rise to them. History matters.

Complementary forces that render institutional dynamics a historical process are the focus of North's 2005 book, *Understanding the Process of Economic Change*. More generally, the book emphasizes that economic stagnation emerges when and where institutions fail to adjust efficiently. The focus of the analysis is on the cognitive capabilities and limitations of individuals and how they influence institutional change. Institutions constructed by individuals reflect their understanding of reality and determine the growth of their understanding. Dissimilar initial cognitive views of reality can therefore lead societies to develop distinct institutions in the same objective situation. The different processes of individual and social learning that these initial institutions imply keep each society on a distinct institutional trajectory.

Hence, for example, the establishment of institutions in the Soviet Union was based on a particular concept of reality. Once established, however, these institutions led to particular learning processes as well as the emergence of organizations with vested interests in the institutions. The result was initial economic success followed by decades of decline because the initial concept of reality was wrong but the organizations it led to had an interest in maintaining the system.

While this book provides new answers to an important question, it more generally calls attention to the need to integrate insights from cognitive science into the examination of the interplay among belief systems, institutions, and economic performance. It particularly emphasizes the relevance of theories of connected or embedded cognition, which argue that human cognition is a social phenomenon shaped by man-made constructs. Institutions shape individual cognition by reducing the uncertainties that would otherwise overwhelm cognitive capacity in complex social situations. At the same time, the resulting bias in our beliefs about this environment can lead to a lock-in of these institutions.

See Also

- Austrian economics
- Coase theorem
- economic history
- growth and institutions
- institutionalism, old
- political institutions, economic approaches to

Selected works

1961. *The Economic Growth of the United States, 1790–1860*. Englewood Cliffs, NJ: Prentice-Hall.

1968. Sources of productivity change in ocean shipping, 1600–1850. *Journal of Political Economy* 76, 953–70.

1971. (With L. Davis.) *Institutional Change and American Economic Growth*. Cambridge: Cambridge University Press.

1973. (With R.P. Thomas.) *The Rise of the Western World: A New Economic History*. Cambridge: Cambridge University Press.

1981. *Structure and Change in Economic History*. New York: Norton.

1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

2005. *Understanding the Process of Institutional Change*. Princeton, NJ: Princeton University Press.

How to cite this article

Greif, Avner. "North, Douglass Cecil (born 1920)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000236> doi:10.1057/9780230226203.1204

The New Palgrave Dictionary of Economics Online

North, Dudley (1641 – 1691)

Douglas Vickers
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

free trade; loanable funds; McCulloch, J. R.; monetary circulation; money supply; North, D.

Article

Sir Dudley North, knighted for his service as a sheriff of London in 1682, was born at Westminster in 1641, the third of five sons of the fourth Baron Guilford. He died at Covent Garden on the last day of December 1691. After a highly successful merchant career in the Levant, he returned to England in 1680 and was appointed a Commissioner of the Customs in 1683. He was promoted to Commissioner of the Treasury in 1685, and when that Commission was dissolved a few months later he returned to the Customs where he remained until the Revolution of 1688.

North's place in the history of economic theory is due to his essay *Discourses upon Trade*, published in 1691 (or early 1692). His clear-sighted advocacy of free trade principles, his opposition with John Locke, to the proposals advocated by Sir Thomas Culpeper and Sir Josiah Child for a legal maximum rate of interest, and his advanced views of the beneficial effects of monetary circulation make the *Discourses* a high-water mark in the pre-classical literature.

The *Discourses*, first published anonymously, were summarized in the biography of Sir Dudley published by his brother Roger in 1744. The Preface to the *Discourses*, the concluding paragraph of the second Discourse, and the final paragraph of the Postscript appear to be the work of Roger. The work was rediscovered and evaluated very highly by the classical economists and J.R. McCulloch published a reprint of the *Discourses* in 1822.

Applying a general supply and demand theory of prices to the determination of interest rates, North argued that a law to restrict the interest rate to a specified maximum level would be ineffective. The market rate of interest depended heavily on the availability of loanable funds which depended on the savings made out of income, a ‘surplus’ that provides an accumulation of investable ‘stock’. A fourfold proposition follows. First, ‘as more buyers than sellers raiseth the price of a commodity, so more borrowers than lenders will raise interest’. Second, ‘as the landed man letts his land, so these still let their stock; … thus to be a landlord or a stock-lord is the same thing’. Third, ‘it is not low interest that

makes trade, but trade increasing, the stock of the nation makes interest low.' Fourth, as the largest part of the demand for loanable funds was for consumption purposes (leading to a *prodigality* and thrift theory of interest, rather than one of *productivity* and thrift) 'an ease of interest will rather be a support to luxury than to trade'.

North argued that it was not so much that trade depended on money as that the money supply depended on trade. For 'nations which are very poor, have scarce any money, and in the beginnings of trade have often made use of something else, … as wealth increased, gold and silver hath been introduced and drove out the other'. A money supply adequate to the needs of trade would be assured, moreover, by the 'ebbing and flowing of money', the coining, melting, and recoining of bullion. 'The buckets work alternately.' Emphasizing the significance of monetary expenditure and circulation, and not simply the money supply as such, complaints against a shortage of money were met by the argument that the remedy for a depressed economy was not 'the increase of specific money' but a disposition to spend rather than hoard. 'The nation … never thrives better than when riches are tost from hand to hand.'

See Also

- mercantilism

Selected works

1691. *Discourses upon Trade*. London. Ed. J.H. Hollander, Baltimore: Johns Hopkins University Reprint, 1907.

Bibliography

Letwin, W. 1963. *The Origins of Scientific Economics: English Economic Thought 1660–1776*. London: Methuen.

McCulloch, J.R. 1824. *A Discourse on the Rise, Progress, Peculiar Objects, and Importance of Political Economy*. Edinburgh: Constable.

McCulloch, J.R. 1856. *Early English Tracts on Commerce*. London: Political Economy Club.

North, R. 1744. *The Life of the Honourable Sir Dudley North … and of Dr. John North, Master of Trinity College*. Ed. M. North, London.

Vickers, D. 1959. *Studies in the Theory of Money 1690–1776*. Philadelphia: Chilton.

How to cite this article

Vickers, Douglas. "North, Dudley (1641–1691)." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000099> doi:10.1057/9780230226203.1205

The New Palgrave Dictionary of Economics Online

Nove, Alexander (Alec) N. (1915 – 1994)

P.J.D. Wiles

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Nove, A. N.; planning; socialism; Stalinism, political economy of

Article

Born in Leningrad, of a Polish–Jewish family Novakovski, Alec Nove would say he felt like a Russian. His father was a Menshevik and his uncle a Bolshevik. He used as a small boy to listen to them arguing, and respected his uncle more. His native language was Russian. The family emigrated to London shortly after the Revolution.

Nove was educated at the London School of Economics (B.Sc. Econ. 1936). His first civilian job was at the Board of Trade and he entered academic life in 1958 as Reader at the London School of Economics. He was Professor Emeritus at Glasgow University at the time of his death.

Impatience with orthodox theory and its whole implementarium did not conceal a sharp economic mind. This was applied mainly to Sovietology and to socialism generally. Nove comes after the great pioneers of Sovietological economics: Sergei Prokopovich, Naum Jasny, Solomon Schwarz and (slightly younger) Abram Bergson. Less Soviet or at least less Russian than the first three, he was also less Western than the last, and this from personal choice, since his whole education was Western. But he always cultivated an understanding of the system in its own terms, and this fit in with his anti-neoclassical bent. A flaw here however was his extreme reluctance to master Marxist ideology in its many varieties: it was a strictly practical view of the Soviet system that was taken.

His methodology can only be called breadth of mind, energy and intuition: foraging through the wasteland of the current Russian literature and making new and important insights. Nove was the first to write seriously about the *variety* of the success indicators imposed upon planned enterprises; the first to note that in about 1980 Soviet economists were producing and almost publishing their own price indices (these rose far quicker than the official ones); one of the first to spot the brave scholars who were revising the harvest figures during the collectivization and the famine.

Much of his work was political economy: Trotsky and socialism; Stalinism and planning; the decision as to when to collectivize agriculture; glasnost. There is a also a cornucopia of minor contributions on

Soviet literature; pre-revolutionary Russian opera; Poland; Hungary; and the misuse of economic criteria by the British public sector (this sideline however was flawed by Scottish nationalism, if that is a correct name for the disgruntlement of a Glaswegian globetrotter who finds he must go everywhere via London). Here as everywhere inspired common sense and strong empirical knowledge produced work that was occasionally wrong-headed, usually brilliant, very seldom dull, never unclear.

Selected works

1961. *The Soviet Economy*. London: George Allen & Unwin; 3rd edn, 1969.

1965. *Was Stalin Really Necessary?* London: George Allen & Unwin.

1969. *An Economic History of the USSR*. London: Allen Lane, The Penguin Press, 3rd edn, 1993.

1973. *Efficiency Criteria for Nationalised Industries*. London: Allen & Unwin.

1977. *The Soviet Economic System*. London: Allen & Unwin.

1979. *Political Economy and Soviet Socialism*. London: Allen & Unwin.

1983. *The Economics of Feasible Socialism*. London: Allen & Unwin; revision *The Economics of Socialism Revisited*, London: Routledge, 1992.

1986. *Socialism, Economics and Development*. London: Allen & Unwin.

1989. *Glasnost in Action: Cultural Renaissance in Russia*. London: Routledge.

1991. *Studies in Economics and Russia*, London: Palgrave MacMillan.

1998. *Alec Nove on Economic Theory. Previously Unpublished Writings*, 1, ed. I. D. Thatcher. Cheltenham: Edward Elgar.

1998. *Alec Nove on Communist and Postcommunist Countries. Previously Unpublished Writings*, 2, ed. I. Thatcher. Cheltenham: Edward Elgar.

How to cite this article

Wiles, P. J. D. "Nove, Alexander (Alec) N. (1915–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000107> doi:10.1057/9780230226203.1206

The New Palgrave Dictionary of Economics Online

Novozhilov, Viktor Valentinovich (1892 – 1970)

Holland Hunter and Robert W. Campbell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capital intensity; depreciation; inversely related expenditures; Kantorovich, L. V.; labour theory of value; Nemchinov, V. S.; Novozhilov, V. V.; opportunity cost; resource allocation; value theory

Article

Novozhilov was born in Khar'kov, and died in Leningrad. He was instrumental, along with the mathematician Leonid Vital'evich Kantorovich, in reviving a mathematical approach to economic theory in the USSR after Stalin's death, and in laying a basis for a modern theory of value and allocation. Educated at Kiev University before the revolution, Novozhilov taught at several institutions in the Ukraine, but from 1922 lived in Leningrad, teaching and working in research institutes. From 1935 he taught at the Leningrad Polytechnical Institute, and from 1944 until 1952 was also professor and head of the Department of Statistics at the Leningrad Engineering–Economics Institute. His work with project-making institutes involved Novozhilov in the issue of capital intensity choices, which became the basis for his doctoral dissertation. In illuminating the question of effective allocation of capital among competing projects, he developed a more general theory for allocation of all resources, the centrepiece of which was the concept of ‘inversely related expenditures’ (*zatraty obratnoi sviazi*) equivalent to opportunity cost. His analytic framework was dynamic, incorporating capital allocation over time, as well as the impact of depreciation and obsolescence.

His original and elegant theoretical ideas were presented in papers published in 1939, 1941, 1946 and 1947 that were largely ignored. The most comprehensive exposition of Novozhilov's ideas is a book he was finally able to publish in 1967, which illustrates his ideas on investment choices and the time factor in economics, places his innovative approach in its doctrinal context, and defends it against domestic and foreign critics. His economic theory is expounded within the limits of political orthodoxy. Novozhilov took the structure of demand as given (by the Party), which enabled him to spell out resource-allocating criteria for the Soviet economy very similar to those familiar in the West, except that with the demand blade of the scissors held fixed, only the supply side cut the paper. By casting the resource allocation problem in terms of minimizing labour input (direct and indirect) he sought to

preserve Marx's labour theory of value. Both his contribution and the absence of an explanation of demand were soon recognized abroad (see Grossman, 1953; Campbell, 1961).

In the mid-1950s, when V.S. Nemchinov organized a revival of serious economic analysis in the USSR, Novozhilov, along with Kantorovich, was a central figure in training a new generation of economists. The three men were awarded Lenin Prizes in 1965. As a result of the pioneering work of Novozhilov and Kantorovich, the basis for a correct and comprehensive theory of value has already been to hand for several decades. Additional biographic and bibliographical details, and interpretations of Novozhilov's work may be found in Campbell (1961), Ellman (1973), Grossman (1953), Holubnychy (1982), and Petrakov (1972).

See Also

- economic calculation in socialist countries
- Kantorovich, Leonid Vitalievich

Selected works

1939. Metody soizmereniia narodnokhoziaistvennoi effektivnosti planovykh i proektnykh variantov [Methods of commeasuring the economic effectiveness of variants in planning and project-making]. *Trudy Leningradskogo industrial'nogo institute* [Papers of the Leningrad Industrial Institute] No. 4. Leningrad.

1941a. Metody izmereniia narodnokhoziaistvennoi effektivnosti proektnykh variantov [Methods of measuring the national economic effectiveness of project variants]. Dissertation for the doctoral degree, awarded in 1941.

1941b. Praktikuemye metody soizmereniia sebestoimosti i vlozhenii [Methods used in practice for co-measuring current outlays and investments]. *Trudy Leningradskogo politekhnicheskogo instituta* [Papers of the Leningrad Polytechnical Institute] No. 1. Leningrad.

1946. Metody nakhozhdeniia minimuma zatrat v sotsialisticheskom khoziaistve [Methods of finding the minimum expenditure in a socialist economy]. *Leningradskii politekhnicheskii institut imeni M.I. Kalinina: Trudy* [The Leningrad Kalinin Polytechnical Institute Papers] No. 1. Leningrad.

1947. Sposoby nakhozhdeniia maksimuma effekta kapitalovlozhenii v sotsialisticheskom khoziaistve [Methods for finding the maximum effect of capital investments in the socialist economy]. *Trudy Leningradskogo finansovo-ekonomicheskogo instituta* [Papers of the Leningrad Financial Economic Institute], Vypusk III [Issue 3]. Leningrad.

1967. *Problemy izmereniia zatrat i rezul'tatov pri optimal'nom planirovanii* [Problems of Cost–Benefit Analysis in Optimal Planning]. Moscow. 2nd edn, 1972. Trans. (with title as shown), White Plains, NY:

International Arts and Sciences Press, 1970.

1972. *Voprosy razvitiia sotsialisticheskoi ekonomiki* [Questions of the development of socialist economics]. Moscow.

Bibliography

Campbell, R.W. 1961. Marx, Kantorovich, and Novozhilov: Stoimost’ versus reality. *Slavic Review* 20, 402–18.

Ellman, M. 1973. *Planning Problems in the USSR; The Contribution of Mathematical Methods to Their Solution, 1961–1971*. Cambridge: Cambridge University Press.

Grossman, G. 1953. Scarce capital and Soviet doctrine. *Quarterly Journal of Economics* 67, 311–43.

Holubnychy, V. 1982. V.V. Novozhilov's theory of value. In *Soviet Regional Economics: Selected Works of Vsevolod Holubnychy*, ed. I.S. Koropeckyj. Edmonton: Canadian Institute of Ukrainian Studies, University of Alberta.

Petrakov, N.I. 1972. Nauchnaia i pedagogicheskaia deiatel'nost’ V.V. Novozhilova [The scientific and pedagogical work of V.V. Novozhilov]. In Novozhilov (1972).

How to cite this article

Hunter, Holland and Robert W. Campbell. "Novozhilov, Viktor Valentinovich (1892–1970)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000108> doi:10.1057/9780230226203.1207

The New Palgrave Dictionary of Economics Online

numeraire

Michael Allingham

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

money; numeraire

Article

In general equilibrium theory the price of one good in terms of another is interpreted as the amount of the second which can be exchanged for a given amount of the first. There is thus no essential role for a standard of value, or *numéraire*, though it is frequently helpful to introduce this. Such a *numéraire* is a commodity in terms of which, by convention, other commodities are valued.

The concept seems to have been introduced by Steuart (1767), albeit with some confusion between the properties of ‘money’ and ‘units of account’. Walras (1874–7) clarified the concept, and showed how prices expressed in terms of one *numéraire* could be translated into prices in terms of another, without any introduction of ‘money’. In the present discussion we commence with a justification of the use of a *numéraire*. We then discuss the choice of a *numéraire* and some problems which may arise through the use of this.

We may represent an economy with n commodities by the excess demand function $f: S \to R^n$ where $S = R^n_+ - 0$. The interpretation is that $f(p)$ is the vector of aggregate excess demands (positive) or excess supplies (negative) expressed at the price system p. A basic property of f is that it is homogeneous of degree zero, that is $f(tp)=f(p)$ for all positive t.

It is this property which justifies the use of a *numéraire*. We can, for example, take commodity n to be *numéraire*, that is, ensure that $p_n=1$, by setting the scalar t appropriately. Thus the price system q can be replaced by the *numéraire* price system p with $p_n=1$ by multiplying q by $t=1/q_n$; nothing real changes, since $f(p)=f(q)$. However, this is only possible if we can ensure that q_n is positive; since q is restricted only to S this may prove difficult.

The problem of the price of a chosen *numéraire* possibly being zero may be avoided by using a composite *numéraire*, that is a basket of goods. The scalar t may then be set as $u \cdot q$ where u is the unit vector in R^n: this has the effect of restricting p to the unit simplex in R^n. Alternatively, a nonlinear

normalization may be used, for example setting $t = q \cdot q$: this has the effect of restricting *p* to the surface of a sphere in R^n_+.

However, in reality prices are usually quoted in terms of some single unit of account, or *numéraire*, and it may be useful for the model of the economy to recognize this. Provided that all commodities are desirable, in the sense that $f_i(p)$ is infinite if $p_i=0$, there is no possibility of any price being zero in equilibrium, that is some *p* where $f(p)=0$. But there may be a problem of p_i being zero on some adjustment path of prices. Whether this is indeed a problem will depend on both the nature of *f* and on the adjustment process governing this path. For example, if the adjustment process is given by $\dot{p} = h(f(p))$ where *h* is a continuous sign-preserving function (and a dot indicates differentiation with respect to time) and if *f* has the above desirability property, then there is no problem. Alternatively, if the adjustment process is $\dot{p}_i = 0$ if $p_i \leq 0$ and $f_i(p)<0$, while $\dot{p}_i = h_i(f_i(p))$ otherwise, then again there is no problem, provided of course that initial prices are positive (Arrow and Hahn, 1971). However, if these properties do not apply, and particularly if the adjustment process is discrete, there may be a problem.

Provided we can use a simple *numéraire* it is clear that if equilibrium is unique in terms of one *numéraire* then it will be unique in terms of another. However, the choice of *numéraire* may be relevant to considerations of stability: that is, for some given adjustment process involving a *numéraire* the economy may be stable for some *numéraire* but not for some other. Some sufficient conditions for stability, such as the condition that *f* have the revealed preference property, are clearly independent of any choice of a *numéraire*, while others are not (Hahn, 1982). For example, the diagonal dominance condition that all commodities are normal and that there are some units in which commodities can be measured such that each of their excess demands is more sensitive to a change in its own price than it is to a change in all other non-*numéraire* prices combined, is clearly dependent on the choice of *numéraire*; indeed, because of homogeneity it makes no sense to attempt to extend it to include the *numéraire.* An economy may have this property, which is sufficient for stability, for one *numéraire* but not for some other. Since this condition is not necessary for stability it does not follow that the economy will be unstable with the second *numéraire*, but neither can stability be guaranteed.

The reason why uniqueness, for example, does not depend on the choice of *numéraire* while stability may, is that stability depends on the adjustment process. Strictly speaking, a change of *numéraire* is simply a change of adjustment process: it is quite natural that the economy may be stable under one adjustment process but not under another.

The question of a *numéraire* has a practical as well as a theoretical importance. In many cases ‘money’ is the natural *numéraire* – though the introduction of money in an essential sense, as opposed to simply as a unit of account, introduces its own problems (Clower, 1967).

See Also

- Walras, Léon

Bibliography

Arrow, K.J. and Hahn, F. 1971. *General Competitive Analysis*. San Francisco: Holden Day.

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6(December), 1–8.

Hahn, F. 1982. Stability. In *Handbook of Mathematical Economics*, vol. 2, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Steuart, Sir J. 1767. *Principles* (Book 1). London.

Walras, L. 1874–7. *Eléments d'économie politique pure*. Definitive edn, Lausanne: Corbaz, 1926. Trans. W. Jaffé as *Elements of Pure Economics*, London: George Allen & Unwin, 1954.

How to cite this article

Allingham, Michael. "numeraire." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000110> doi:10.1057/9780230226203.1208

The New Palgrave Dictionary of Economics Online

numerical optimization methods in economics

Karl Schmedders
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Optimization problems are ubiquitous in economics. Many of these problems are sufficiently complex that they cannot be solved analytically. Instead economists need to resort to numerical methods. This article presents the most commonly used methods for both unconstrained and constrained optimization problems in economics; it emphasizes the solid theoretical foundation of these methods, illustrating them with examples. The presentation includes a summary of the most popular software packages for numerical optimization used in economics, and closes with a description of the rapidly developing area of mathematical programs with equilibrium constraints, an area that shows great promise for numerous economic applications.

Keywords

central path; computable general equilibrium models; computational algebraic geometry; constrained optimization; global minimum; homotopy continuation methods; interior-point methods; Karush-Kuhn-Tucker conditions; Kuhn–Tucker conditions; Lagrange multipliers; line search methods; linear independence constraint qualification (LICQ); linear programming; local minimum; local solution; logarithmic barrier method; mathematical programs with equilibrium constraints; metaheuristics; multidimensional optimization problems; Newton's method; Newton–Raphson methods; nonlinear programming; numerical optimization methods; optimality conditions; optimization algorithms; penalty methods; polynomial functions; quadratic programs; sequential quadratic programming; simplex method for solving linear programs; simulated annealing; Smale's global Newton method; steepest descent method; Taylor's th; trust region methods; unconstrained optimization

Article

Optimizing agents are at the centre of most economic models. In our models we typically assume that consumers maximize utility or wealth, that players in a game maximize payoffs, that firms minimize costs or maximize profits, or that social planners maximize welfare. But it is not only the agents in our models that optimize. Econometricians maximize likelihood functions or minimize sums of squares. Clearly optimization is one of the key techniques of modern economic analysis.

The optimization problems that appear in economic analysis vary greatly in nature. We encounter finite-dimensional problems such as static utility maximization problems with a few goods. An optimal solution to such a problem is a finite-dimensional vector. We analyse infinite-dimensional problems such as infinite-horizon social planner models or continuous-time optimal control problems. Here the solution is an infinite-dimensional object, a vector with countably infinitely many elements or even a function over an interval. Our agents may face constraints such as budget equations, short-sale restrictions or incentive-compatibility constraints. There are also unconstrained problems such as nonlinear least-square problems. Decision variables may even be restricted to be discrete. Agents' objective functions may be linear or nonlinear, convex or nonconvex, many times differentiable or discontinuous. Finally, an economic optimization problem may be deterministic or stochastic.

Unless we consider stylized models in theoretical work or make very stringent and often quite unrealistic assumptions in applied models, the optimization problems that we encounter cannot be solved analytically. Instead we need to resort to numerical methods. The numerical methods that we employ to solve economic optimization models vary just as much as the optimization problems we encounter. It is therefore impossible to cover the wide variety of numerical optimization methods that are useful in economics in a short article. For the purpose of the exposition here we focus on deterministic finite-dimensional nonlinear optimization problems including linear programs. This is a natural choice because such

problems are ubiquitous in economic analysis. Moreover, the techniques for these problems play also an important part in many other numerical methods such as those for solving economic equilibrium and infinite-dimensional problems. The interested reader should consult computation of general equilibria (new developments), computational methods in econometrics and dynamic programming.

We first indicate some of the fundamental technical difficulties that we need to be aware of when we apply numerical methods to our economic optimization problems. We then highlight the basic theoretical foundations for numerical optimization methods. The popular numerical optimization methods have strong theoretical foundations. Unfortunately, current textbooks in computational economics, with the partial exception of Judd (1998), neglect to emphasize these foundations. As a result some economists are rather sceptical about numerical methods and view them as rather ad hoc approaches. Instead, a good understanding of the theoretical foundations of the numerical solution methods gives us an appreciation of the capabilities and limitations of these methods and can guide our choice of suitable methods for a specific economic problem. We outline the most fundamental numerical strategies that form the basis for most algorithms. All presented numerical strategies are implemented in at least one of the those computer software packages for solving optimization problems that are most popular in economics. We close our discussion with a look at mathematical programs with equilibrium constraints (MPECs), a promising research area in numerical optimization that has useful applications in economics.

1 Newton's method in one dimension

We start with the one-dimensional unconstrained optimization problem

$$\min_{x \in \mathbb{R}} f(x). \tag{1}$$

Perhaps the first (if any) numerical method that most of us learnt in our calculus classes is Newton's method. Newton's method attempts to minimize successive quadratic approximations to the objective function f in the hope of eventually finding a minimum of f. To start the computations we need to provide an initial guess $x^{(0)}$. The quadratic approximation $q(x)$ of $f(x)$ at the point x^0 is

$$q(x) = f(x^0) + f'(x^{(0)})(x - x^{(0)}) + \frac{1}{2} f''(x^{(0)})(x - x^{(0)})^2$$

where f' and f'' denote the first and second derivative of the function f, respectively. Solving the first-order condition

$$q'(x) = f'(x^{(0)}) + f''(x^{(0)})(x - x^{(0)}) = 0$$

on the assumption that $f'(x^{(0)}) \neq 0$ yields the solution

$$x^{(1)} = x^{(0)} - \frac{f'(x^{(0)})}{f''(x^{(0)})}.$$

Now we repeat this process using a quadratic approximation to f at the point $x^{(1)}$. The result is a sequence of points, $\{x^{(k)}\} = x^{(0)}, x^{(1)}, x^{(2)}, \ldots, x^{(k)}, \ldots$, that we hope will converge to the solution of our minimization problem. This approach is based on the following theoretical result.

Theorem: Suppose x^* is the solution to the minimization problem (1). Suppose further that f is three times continuously differentiable in a neighborhood of x^* and that $f'(x^*) \neq 0$. Then there exists some $\delta > 0$ such that if $|x^* - x(0)| < \delta$, then the sequence $\{x^{(k)}\}$ converges quadratically to x^*, that is,

$$\lim_{k \to \infty} \frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^2} = \kappa$$

for some finite constant κ.□

We illustrate this theorem with a simple example.

Example 1: A consumer has a utility function $u(x, y) = \ln(x) + 2\ln(y)$ over two goods. She can spend $1 on buying quantities of these two goods, both of which have a price of $1. After substituting the budget equation, $x+y=1$, into the utility function the consumer wants to maximize $f(x) = \ln(x) + 2\ln(1-x)$. Setting the first order condition equal to 0 yields the solution $x^* = \frac{1}{3}$. (This quantity is globally optimal because the function f is strictly concave.)

Suppose we start Newton's method with the initial guess $x^{(0)}=0.5$. Then the first Newton step yields

$$x^{(1)} = 0.5 - \frac{f'(0.5)}{f''(0.5)} = 0.5 - \frac{-2}{-12} = \frac{1}{3}.$$

Newton's method found the exact optimal solution in one step. This (almost) never happens in practice. Much more usual is the behaviour we observe when we start with $x^{(0)}=0.8$. Then Newton's method delivers as its first five steps

$$0.63030303,\ 0.407373702,\ 0.328873379,\ 0.333302701,\ 0.333333332.$$

We observe that the sequence rapidly converges to the optimal solution. The corresponding errors $|x^{(k)}-x^*|$,

$$0.2969697,\ 0.07404037,\ 0.00445995,\ 3.0632 \cdot 10^{-5},\ 1.4078 \cdot 10^{-9}$$

converge to but never exactly reach zero. The rate of convergence is called quadratic since $|x^{(k+1)} - x^*| < L|x^{(k)} - x^*|^2$ for some constant L once k is sufficiently large.□

But, of course, contrary to this simple example, we typically do not know x^* and so cannot compute the errors $|x^{(k)} - x^*|$. Instead, we need a stopping rule that indicates when the

procedure terminates. The requirement that $f'(x^{(k)}) < \delta$ may appear to be an intuitive stopping rule. But that rule may be insufficient for functions that are very 'flat' near the optimum and have large ranges of non-optimal points satisfying this rule. Therefore, a safer stopping rule requires both $f'(x^{(k+1)}) < \delta$ and $|x^{(k+1)} - x^{(k)}| < \varepsilon(1 + |x^{(k)}|)$ for some pre-specified small error tolerance $\varepsilon, \delta > 0$. So the Newton method terminates once two subsequent iterates are close to each other and the first derivative almost vanishes.
Observe that Newton's method found a maximum, and not a minimum, of the utility function. The reason for this fact is that this method does not search directly for an optimizer. Note that the key step in the algorithm is finding a stationary point of the quadratic approximation $q(x)$, that is, a point satisfying $q'(x)=0$. Before we can claim to have found a maximum or minimum of f we need to do more work. In this example the strict concavity of the utility function ensures that a stationary point of f yields a maximum. So an assumption of our economic model assures us that the numerical method indeed finds the desired maximum.

Example 2: Consider the simple polynomial function $f(x) = x(x-2)^2$. Starting with $x^{(0)}=1$ leads to the sequence

$$0.5,\ 0.65,\ 0.666463415,\ 0.666666636,\ \ldots$$

converging to $\frac{2}{3}$. Starting with $x^{(0)}=1.5$ leads to the sequence

$$2.75,\ 2.198529412,\ 2.022777454,\ 2.000376254$$

converging to 2. Neither of these two points yields a global optimum, the function f is actually unbounded above and below. The point $\frac{2}{3}$ is a local maximizer ($f''(2/3) = -4 < 0$) while 2 is a local minimizer ($f''(2) = 4 > 0$). The stationary point that we find greatly depends on our initial guess.□

Our simple observations about the behaviour of Newton's method for one-dimensional optimization problems apply in practice to higher-dimensional nonlinear optimization problems and to almost all optimization methods. We will almost always face these fundamental issues in our economic applications. First, most practical optimization methods for unconstrained problems search only for a stationary point (with possibly additional favourable properties). They do not directly attempt to compute an optimizer. Second, as a result, most practical methods may terminate with a non-optimal point. To ensure global optimality we need to perform additional checks. Third, it is rather unusual in practice to explicitly solve for an exact solution. Usually we can only hope for a sequence of points $\{x^{(k)}\}$ generated by an iterative process that converges to a limit having some desired property. Therefore, we need a stopping rule that indicates when the iterative process stops. Fourth, the algorithm may not terminate and diverge even if a globally optimal solution exists.
Newton's method is a special instance of a family of methods for solving multidimensional optimization problems. Before we examine more general methods we provide some basic intuition for the theoretical underpinnings of these solution methods.

2 Theoretical foundation: Taylor's th

The gradient of the function f at a point $x = (x_1, x_2, \ldots, x_n)$ is the column vector

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right)^T$$

of partial derivatives of f with respect to the variables $x_1, x_2, \ldots, x_n$. The Hessian of f at x is the $(n \times n)$-matrix

$$H(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1}^{n}$$

of the second derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ of *f*. The inner product of two (column) vectors $x, y \in \mathbb{R}^n$ is denoted by $x^T y$.
Many numerical methods rely on linear or quadratic approximations of the function *f*. Taylor's theorem provides a justification for this approach. Here we give a simple version of this theorem for functions with Lipschitz continuous derivatives. Consider a function $F: X \to Y$ for open sets $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$. Then *F* is Lipschitz continuous at $x \in X$ if there exists a constant $\gamma(x)$ such that

$$\| F(y) - F(x) \| \leq \gamma(x) \| y - x \|$$

for all $y \in X$, where $\|\cdot\|$ denotes the standard Euclidean norm.
Theorem: Suppose the function $f: X \to \mathbb{R}$ is continuously differentiable on the open set $X \subset \mathbb{R}^n$ and that the gradient function ∇f is Lipschitz continuous at *x* with Lipschitz constant $\gamma^l(x)$. Also suppose that for $s \in \mathbb{R}^n$ the line segment $x + \theta s \in X$ for all $\theta \in [0,1]$. Then, the linear function *l* with $l(s) = f(x) + \nabla f(x)^T s$ satisfies

$$| f(x+s) - l(s) | \leq \frac{1}{2} \gamma^l(x) \| s \|^2.$$

Moreover, if *f* is twice continuously differentiable on *X* and the Hessian *H* is Lipschitz continuous at *x* with Lipschitz constant $\gamma^q(x)$, then the quadratic function *q* with $q(s) = f(x) + \nabla f(x)^T s + \frac{1}{2} s^T H(x) s$ satisfies

$$| f(x+s) - q(s) | \leq \frac{1}{6} \gamma^q(x) \| s \|^3.$$

□

3 Unconstrained optimization

The multidimensional generalization of the unconstrained optimization problem (1) is given by

$$\min_{x \in \mathbb{R}^n} f(x). \tag{2}$$

Solving this optimization problem entails finding a global minimizer x^* satisfying $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$. With the exception of a few algorithms for problems that are either very small or have very special structure, there are no algorithms that are guaranteed to find a global minimum. Thus, we need to think in terms of local minima. A local minimizer is a point x^* that satisfies $f(x^*) \leq f(x)$ for all $x \in \mathcal{N}(x^*)$ where $\mathcal{N}(x^*)$ denotes a neighborhood of x^*. The point x^* is called an isolated local minimizer if it is the only local minimizer in $\mathcal{N}(x^*)$.

All these definitions by themselves are not all that helpful for finding a minimum. Instead, just as Newton's method in one dimension does, all practical numerical methods for unconstrained optimization problems rely on optimality conditions to find candidates for local minima. For functions with sufficient differentiability properties these are the following well-known conditions.

Theorem: [Optimality conditions for unconstrained minimization]

1. 1. If f is continuously differentiable and x^* is a local minimizer of f, then $\nabla f(x^*)=0$.
2. 2. If f is twice continuously differentiable and x^* is a local minimizer of f, then $\nabla f(x^*)=0$ and $s^T H(x^*) s \geq 0$ for all $s \in \mathbb{R}^n$.
3. 3. If f is twice continuously differentiable and if x^* satisfies $\nabla f(x^*)=0$ and $s^T H(x^*) s > 0$ for all $s \in \mathbb{R}^n$, $s \neq 0$, then x^* is an isolated local minimizer of f.□

But when can we be assured that a local minimizer of f is actually a solution to the unconstrained optimization problem (2)? The perhaps easiest sufficient condition is that the function f is convex, that is, $s^T H(x) s \geq 0$ for all $x \in \mathbb{R}^n$ if f is twice differentiable. Then any local minimizer x^* is a solution to problem (2), in fact, any stationary point x^* is a solution to (2).

The optimality conditions provide the foundation for all practical unconstrained optimization methods. The focus of all these algorithms is to find (actually, to approximate) a stationary point of f, that is, a solution to $\nabla f(x)=0$. They do so by generating a sequence of iterates $\{x^{(k)}\}$ that ideally terminates once a stopping rule is satisfied indicating that an approximate solution has been found. The key step for these methods is to generate a new iterate $x^{(k+1)}$ from a current iterate $x^{(k)}$. A vast majority of optimization routines uses one of two basic strategies for moving from $x^{(k)}$ to $x^{(k+1)}$, a line search approach or a trust region method.

3.1 Line search methods

The general set-up of a line search method is as follows. From a point $x^{(k)}$ (with $\nabla f(x^{(k)}) \neq 0$) we look for a search direction $s^{(k)}$ that leads us to lower function values for f. Using the linear approximation l with $l(s) = f(x^{(k)}) + \nabla f(x^{(k)})^T s$ we determine a descent direction $s^{(k)}$ satisfying

$$\nabla f(x)^T s^{(k)} < 0,$$

which in turn implies $l(s^{(k)})<f(x^{(k)})$. Because of Taylor's theorem we hope that along a step in the direction $s^{(k)}$ the function value $f(x)$ will be reduced. We calculate a suitable step length $\alpha_k>0$ to ensure that $f(x^{(k+1)})<f(x^{(k)})$ where

$$x^{(k+1)} = x^{(k)} + \alpha_k s^{(k)}.$$

Observe that at a given point $x^{(k)}$ and for a descent direction $s^{(k)}$ finding the optimal value of α_k requires us to solve a one-dimensional optimization problem. In principle we could apply Newton's method to this problem. In practice, however, this one-dimensional problem does not need to be solved exactly because repeatedly finding the optimal step length is

both unnecessary for convergence of line search methods and computationally rather inefficient. Instead modern line search methods prefer to use inexact line searches that just pick a step length that leads to a sufficient decrease in the objective function value. One such approach is the backtracking Armijo line search, which requires that

$$f(x^{(k)} + \alpha_k s^{(k)}) \leq f(x^{(k)}) + \alpha_k \beta \nabla f(x^{(k)})^T s^{(k)}$$

for some $\beta \in (0,1)$. The idea of this requirement is to link the step size α_k to the decrease in f. The longer the step the larger the decrease must be. Starting with an initial guess for α_k, say 1, we can now stepwise reduce the value of α_k until the above condition is satisfied. At that point we set $x^{(k+1)}=x^{(k)}+\alpha_k s^{(k)}$.
While the basic line search method seems very intuitive, it can fail if the search direction and the gradient tend to a point where they are orthogonal to each other, that is, the product $\nabla f(x^{(k)})^T s^{(k)}$ tends to zero without the gradient itself approaching zero. This kind of failure can be avoided by a proper choice of search direction.

3.1.1 Method of steepest descent

The perhaps most intuitive choice for a descent direction is

$$s^{(k)} = -\nabla f(x^{(k)}),$$

because this search direction gives the greatest possible decrease in the linear approximation l (for a fixed step length). It is thus called the steepest descent direction. And indeed, a line search with the steepest descent direction has very nice theoretical properties.
Theorem: Suppose that f is continuously differentiable and that ∇f is Lipschitz continuous on $\mathbb{R}^n$. Then for the sequence $\{x^{(k)}\}$ of iterates generated by a line search method using the steepest descent direction and the backtracking Armijo line search one of the following three conditions must hold.

- (C1) $\nabla f(x^{(k)})=0$ for some $k\geq 0$.
- (C2) $\lim_{k\to\infty}\nabla f(x^{(k)})=0$.
- (C3) $\lim_{k\to\infty} f(x^{(k)})=-\infty$.□

The method of steepest descent has the global convergence property, that is, independent of the starting point the sequence of gradients will converge to a stationary point (but that does not mean that the sequence $x^{(k)}$ converges, think of $-\ln(x^{(k)})$!) or the function values diverge and indicate that no minimum exists.
Example 3: A consumer has a utility function $u(x_1, x_2, x_3) = \sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}$ over three goods. She can spend \$1 on buying quantities of these three goods, all of which have a price of \$1. After substituting the budget equation, $x_1+x_2+x_3=1$, into the utility function the consumer wants to maximize $\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{1 - x_1 - x_2}$. (We can trivially solve this problem with pencil and paper and find the optimal solution $(\frac{1}{14}, \frac{4}{14}, \frac{9}{14})$.) We solve the consumer's optimization problem by minimizing the function $f(x_1, x_2) = -(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{1 - x_1 - x_2})$ with a steepest descent method (using the optimal step length in each step). Figure 1 indicates some of the early steps and Table 1 lists details of some of the steps. (To show convergence of variable values and the optimal function value we report six digits for these terms. The search direction and norm of the gradient are converging to zero and so for simplicity we report fewer and not always the same number of digits. We abbreviate numbers like $6.7\cdot 10^{-8}$ by 6.7(–8).)
The steepest descent method makes good progress in the first few iterations but then slows down considerably. Note the comparatively little change in the values of $x^{(k)}$ during the last 10 to 15 iterations. The figure shows a lot of ‘zigzagging’ from iterate to iterate.□

Steps of a steepest descent method

k	$x_1^{(k)}$	$x_2^{(k)}$	$s^{(k)}$		$\lVert \nabla f(x^{(k)}) \rVert$	$f(x^{(k)})$
0	0.1	0.5	–0.7906	–0.9575	1.2417	–3.62781
1	0.0358229	0.422272	0.6041	–0.4988	0.7834	–3.69734
2	0.0867861	0.380194	–0.3573	–0.4328	0.5612	–3.71804
3	0.0528943	0.339146	0.2503	–0.2066	0.3245	–3.73387
4	0.0772951	0.318999	–0.1321	–0.1600	0.2075	–3.73858
5	0.0643195	0.303284	0.0853	–0.0704	0.1106	–3.74074
6	0.0732734	0.295891	–0.0414	–0.0502	0.0651	–3.74136
7	0.0691862	0.290940	0.0257	–0.0212	0.0334	–3.74157
⋮						⋮
10	0.0715805	0.286543	–0.0034	–0.0041	0.0054	–3.74166
⋮						⋮
15	0.0714140	0.285747	1.64 (–4)	–1.35 (–4)	2.12 (–4)	–3.74166
⋮						⋮
20	0.0714288	0.285716	–5.94 (–6)	–7.19 (–6)	9.33 (–6)	–3.74166

Figure 1

First steps of a steepest descent method

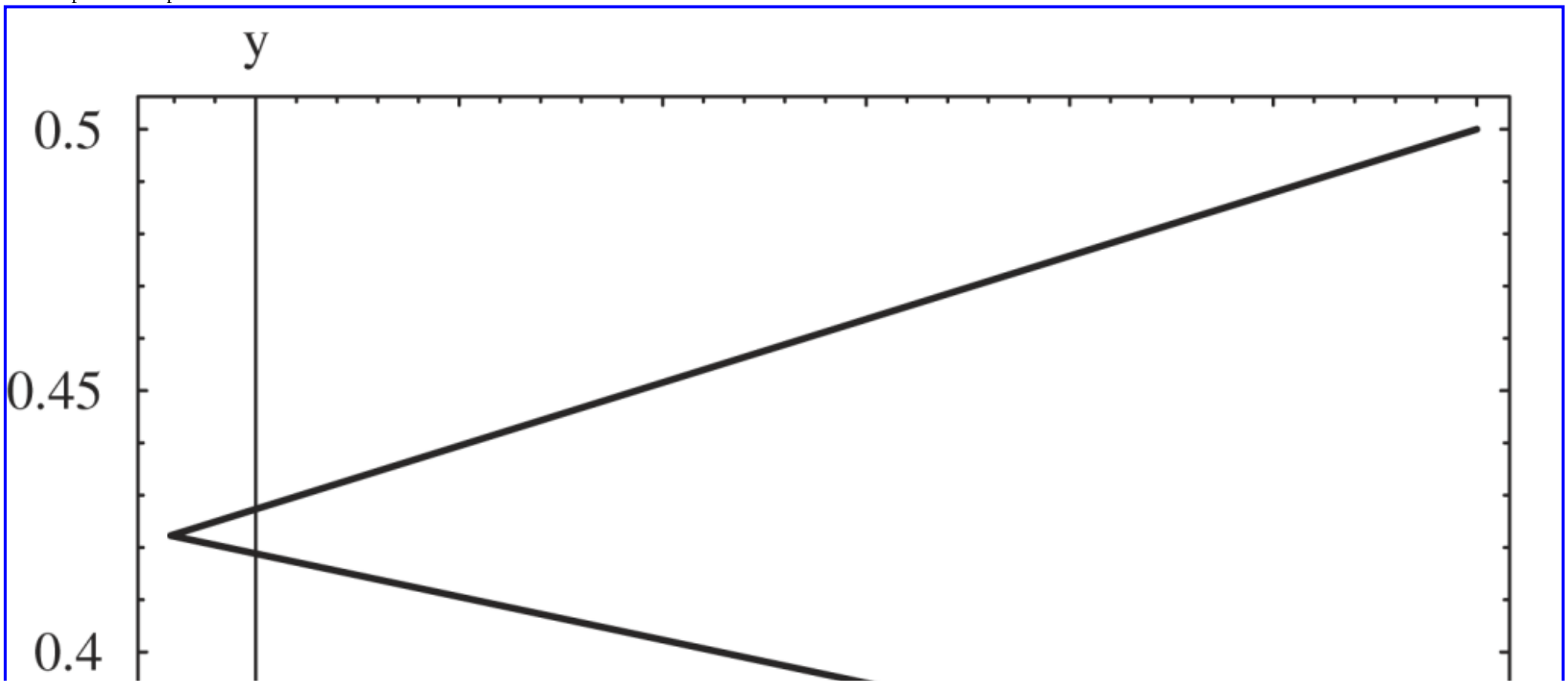

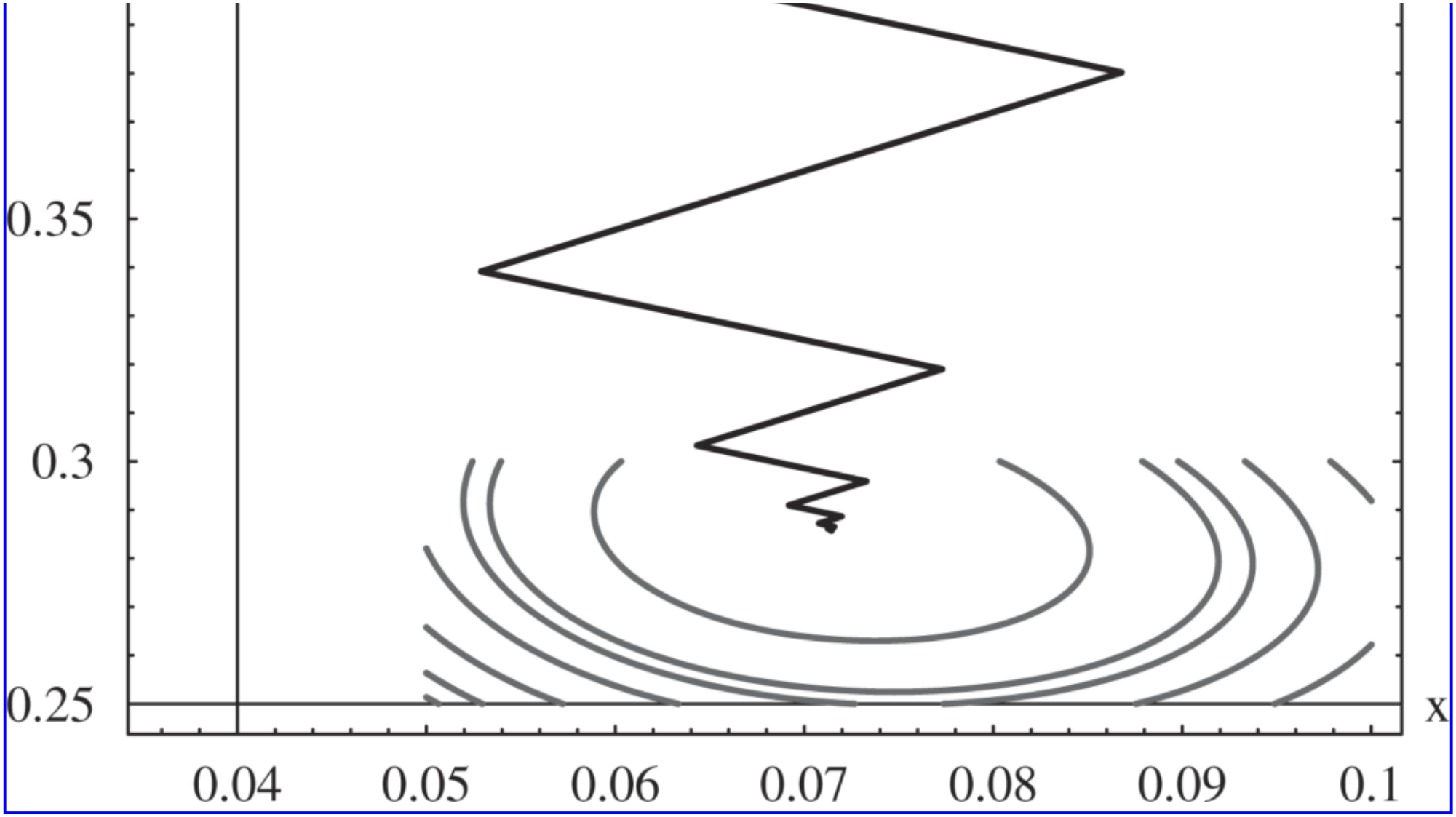

The behaviour of the steepest descent method in the example is quite typical. As a result the convergence of the method is rather slow. And so, despite having the global convergence property, it is useless in practice. The slow convergence (see Nocedal and Wright, 2006, ch. 3) of this method renders it impractical. The convergence problems are essentially due to the reliance on a first-order approximation, which ignores the curvature properties of *f*. Newton's method takes advantage of a second-order approximation.

3.1.2 Newton methods

The quadratic approximation *q* of the objective function *f* at an iterate $x^{(k)}$ is given by

$$q(s) = f(x^{(k)}) + \nabla f(x^{(k)})^{\mathsf{T}} s + \frac{1}{2} s^{\mathsf{T}} H(x^{(k)}) s.$$

The first-order condition $q'(s)=0$ yields the search direction

$$s^{(k)} = -H(x^{(k)})^{-1}\nabla f(x^{(k)}).$$

Only under very strong conditions is Newton's method globally convergent.

Theorem: Suppose that f is continuously differentiable and that ∇f is Lipschitz continuous on $\mathbb{R}^n$. If for the sequence $\{x^{(k)}\}$ of iterates generated by a line search method using the Newton direction and the backtracking Armijo line search the Hessian matrices $H(x^{(k)})$ are positive definite with eigenvalues that are uniformly bounded away from zero, then one of the conditions (C1), (C2), (C3) must hold.□

Example 4: We revisit the consumer's optimization problem from Example 3 and minimize the function $f(x_1, x_2) = -(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{1 - x_1 - x_2})$ with a Newton method (using the optimal step length in each step). Table 2 lists all the steps of this method and Figure 2 displays some of the early steps.

Newton's method converges very rapidly. Unlike the steepest descent method it does not slow down near the solution, instead we see a quadratic rate of convergence just like in the one-dimensional problem in Example 1.□

Steps of a Newton method

k	$x_1^{(k)}$	$x_2^{(k)}$	$s^{(k)}$		$\lVert\nabla f(x^{(k)})\rVert$	$f(x^{(k)})$
0	0.1	0.5	–0.0161	–0.2078	1.2417	–3.62781
1	0.0829896	0.280	–0.0128	0.0062	0.1440	–3.74077
2	0.0714128	0.285450	1.58 (–5)	2.64 (–4)	0.0014	–3.74166
3	0.0714286	0.285714	–2.27 (–8)	1.10 (–8)	3.15 (–7)	–3.74166
4	0.0714286	0.285714			5.46 (–15)	–3.74166

Figure 2

First steps of a Newton method

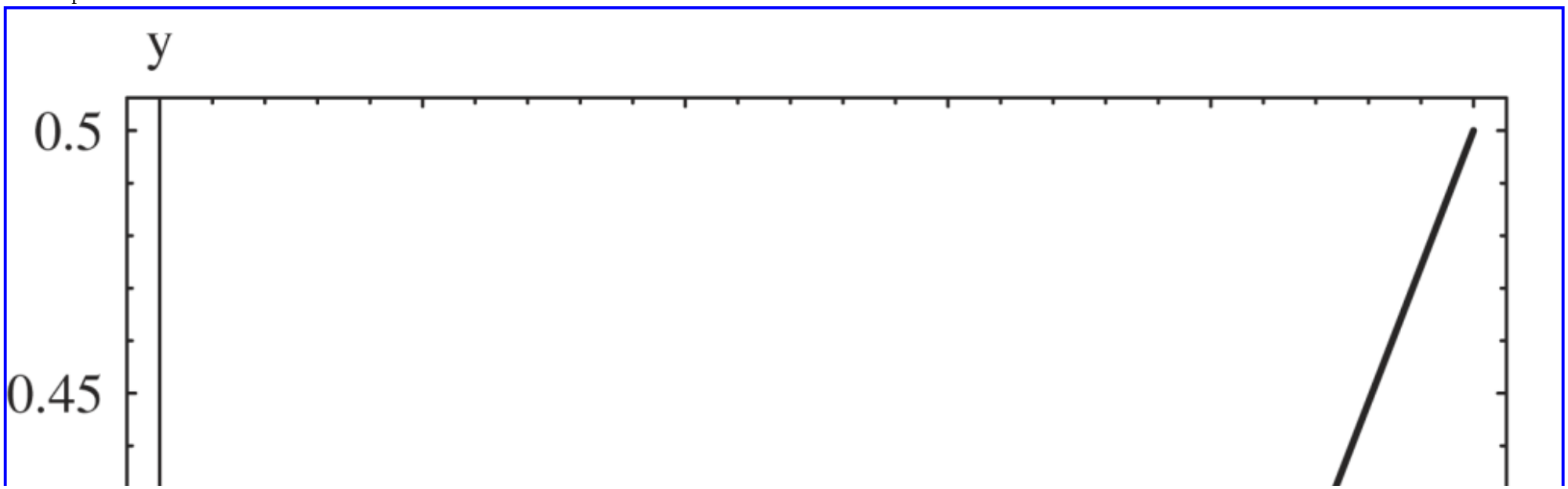

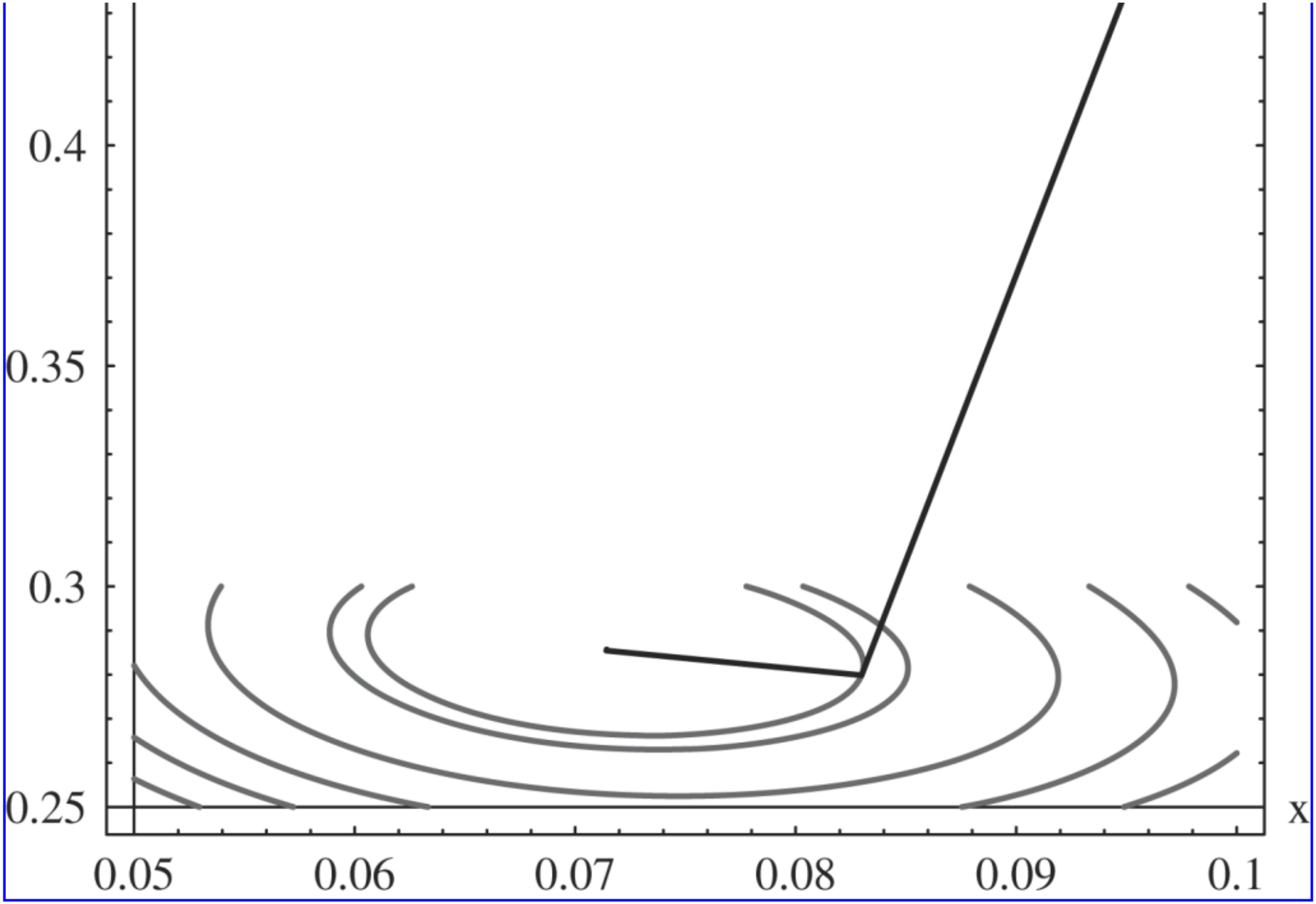

The condition that the Hessian matrix $H(x^{(k)})$ is positive definite for the entire sequence $\{x^{(k)}\}$ is rarely satisfied for general problems. But if the Hessian is not positive definite then the search direction $s^{(k)}$ may be an ascent instead of a descent direction. The modified Newton methods address this problem by modifying the Hessian matrix $H(x^{(k)})$. These methods choose a search direction

$$s^{(k)} = -(H(x^{(k)}) + M(x^{(k)}))^{-1} \nabla f(x^{(k)}),$$

where the matrix $M(x^{(k)})$ is chosen so that $H(x^{(k)})+M(x^{(k)})$ is 'sufficiently' positive definite. If $H(x^{(k)})$ is sufficiently positive definite itself then, of course, $M(x^{(k)})=0$. A proper choice of $M(x^{(k)})$ is crucial for the effectiveness of this approach; see Gould and Leyffer (2002) and Nocedal and Wright (2006) for many more details.

The most tedious task in Newton's method is the computation of the Hessian matrix $H(x^{(k)})$. Therefore, for decades it was fashionable to develop methods, the so-called quasi-Newton methods, that rely on approximations of the exact Hessian matrix. Interest in these methods has somewhat diminished due to the development of automatic differentiation techniques. These techniques allow a very fast and reliable computation of derivatives and so make the task of calculating the Hessian feasible even for large problems. Nocedal and Wright (2006, ch. 6) discuss quasi-Newton methods in detail.

Before we continue our discussion of optimization algorithms we pause for a quick comment on some potential name confusion. In addition to Newton methods for unconstrained optimization there is also a Newton method for solving nonlinear systems of equations. To avoid confusion and for historical reasons the root-finding methods for nonlinear systems of equations are sometimes called Newton–Raphson methods; see Judd (1998) and references therein. In particular, Newton methods for solving unconstrained optimization problems should not be confused with so-called global Newton methods. In economic theory the term 'Smale's global Newton method' appears to be well known. This term refers to a solution method for solving nonlinear systems of equations (see Smale, 1976) which is closely related to homotopy continuation methods. Clearly, we could use methods for nonlinear equations to solve the first-order conditions $\nabla f(x)=0$. This approach, however, does not use other information from the underlying optimization problem and thus is often inefficient. Here we do not discuss methods for solving nonlinear equations, and refer to Allgower and Georg (1979), Judd (1998) and Miranda and Fackler (2002).

3.2 Trust region methods

Line search methods use an approximation of the objective function f to generate a search direction. Subsequently they determine a suitable step length along this direction. Trust region methods also rely on an approximation of f, but they first define a region around the current iterate in which they trust the approximation to be adequate. Then they simultaneously choose the direction and step length.

For the purpose of our discussion here we consider a quadratic approximation of f around $x^{(k)}$,

$$q_k(s) = f(x^{(k)}) + \nabla f(x^{(k)})^T s + \frac{1}{2} s^T B(x^{(k)}) s,$$

where $B(x^{(k)})$ is a symmetric approximation of the Hessian matrix $H(x^{(k)})$. Trust region methods do not require the Hessian matrix of the function q_k to be positive definite. Therefore, we could use $B(x^{(k)})=H(x^{(k)})$. In that case, the algorithm is called a trust region Newton method. Given a trust region radius $\Delta_k>0$ in each iteration, the algorithm seeks an (approximate) solution to the trust region sub-problem

$$\min_{s \in \mathbb{R}^n} q_k(s) \text{subject to} \|s\| \le \Delta_k.$$

Before we discuss how we may solve this sub-problem we need to decide on a proper choice for the trust region radius. Note that $q_k(0)-q_k(s^{(k)})$ is the predicted reduction for a step $s^{(k)}$. Similarly, $f(x^{(k)})-f(x^{(k)}+s^{(k)})$ is the actual decrease in the objective. The ratio

$$\rho_k = \frac{f(x^{(k)}) - f(x^{(k)} + s^{(k)})}{q_k(0) - q_k(s^{(k)})}$$

gives an indication on how well the quadratic approximation predicts the reduction in the function value. Ideally we would like the step $s^{(k)}$ to yield a value of ρ_k of close to or larger than 1. In that case we accept the step and may possibly increase the radius for the next iteration. If, however, ρ_k is close to zero or even negative, then we would decrease the trust region radius, set $x^{(k+1)}=x^{(k)}$, and attempt to solve the sub-problem again.

Recall that line search methods do not require the step length to be chosen optimally in order to be globally convergent. Similarly, it is unnecessary and in fact computationally inefficient to solve the trust region sub-problem exactly. Instead, it suffices to search for a step giving a sufficient reduction in q_k. Such a sufficient reduction is achieved by requiring a decrease that is at least as large at that obtained by a step in the direction of steepest descent. The solution to

$$\min_{\alpha \in \mathbb{R}} q_k(-\alpha \nabla f(x^{(k)})) \text{subject to} \| -\alpha \nabla f(x^{(k)}) \| \le \Delta_k$$

yields the Cauchy point

$$s_k^C = -\tau_k \Delta_k \frac{\nabla f(x^{(k)})}{\| \nabla f(x^{(k)}) \|}$$

where the constant $\tau_k \in (0,1]$ depends on the curvature of q_k and the radius Δ_k; see Nocedal and Wright (2006) for a closed-form solution. The approximate solution $s^{(k)}$ of the trust region subproblem must now satisfy $q_k(s^{(k)}) \le q_k(s_k^C)$.

Theorem: Let q_k be the second-order approximation of the objective function f at $x^{(k)}$ and let s_k^C be its Cauchy point in the trust region defined by $\|s\| \le \Delta_k$. Then

$$q_k(0) - q_k(s_k^C) = f(x^{(k)}) - q_k(s_k^C) \ge \frac{1}{2} \| \nabla f(x^{(k)}) \| \min\left\{ \frac{\| \nabla f(x^{(k)}) \|}{1 + \| B(x^{(k)}) \|}, \Delta_k \right\}.$$

□

The theorem has the typical flavour of results on trust region methods. It relates the reduction in the quadratic approximation, $q_k(0) - q_k(s_k^C)$, to $\|\nabla f(x^{(k)})\|$, which is a measure for the distance to optimality. Once again a global convergence result holds.

Theorem: Consider the sequence $\{x^{(k)}\}$ of iterates generated by the described trust region method. Suppose that f is twice continuously differentiable and both the Hessian of f and the quadratic approximation q_k are bounded for all k. Then one of the conditions (C1), (C2), (C3) must hold.□

The trust region method based on the Cauchy point is effectively a steepest descent (line search) method where the choice of the step length is bounded by the trust region radius. Therefore, this method also suffers from very poor convergence in practice. Better algorithms start from the Cauchy point and try to improve upon it. There is a variety of such

methods that take advantage of additional properties of *f*; see Gould and Leyffer (2002) and Nocedal and Wright (2006). For a comprehensive treatment of trust region methods, see Conn, Gould and Toint (2000).

4 Constrained optimization

Now we consider the constrained optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \geq 0 \; i \in I \qquad \text{(NLP)} \\ & h_j(x) = 0 \; j \in E. \end{aligned}$$

We define the feasible region $\mathcal{F}$ of this optimization problem to be the set of all points that satisfy the constraints, so

$$\mathcal{F} = \{x \in \mathbb{R}^n \mid g_i(x) \geq 0, \;\; i \in I; \; h_j(x) = 0, \;\; j \in E\}.$$

Just as for the unconstrained optimization problem, we can define global and local solution. Of course, a desired optimal solution x^* to this optimization problem satisfies $f(x^*) \leq f(x)$ for all $x \in \mathcal{F}$. A point x^* is a local minimizer if it satisfies $f(x^*) \leq f(x)$ for all $x \in \mathcal{N}(x^*) \cap \mathcal{F}$ for some neighbourhood $\mathcal{N}(x^*)$ of x^*. The vector x^* is an isolated local minimizer if there exists a neighbourhood $\mathcal{N}(x^*)$ in which it is the only local minimizer.

The conditions of these definitions, just like their counterparts for unconstrained optimization problems, are pretty much useless for the computation of optimal solutions – with one major exception. The simplex method for solving linear programming problems relies on the comparison of objective function values at some special points in the feasible region. Most other practical numerical methods, however, rely again on optimality conditions. Penalty methods transform the problem (NLP) into (a sequence of) unconstrained optimization problems and then rely on their respective first-order conditions. Many methods rely directly on optimality conditions for constrained optimization. These optimality conditions require that certain degenerate behaviour does not occur at potential minimizers. Conditions that rule out such degenerate points are called ‘constraint qualifications’. These conditions are important but do not always get the proper attention in economics, but see Simon and Blume (1994) for a rigorous treatment. Numerous such constraint qualifications exist; here we just mention one such condition.

The set of constraints that hold with equality at a feasible point $x \in \mathcal{F}$ is called the active set $\mathcal{A}(x)$. Formally,

$$\mathcal{A}(x) = \{i \in I \mid g_i(x) = 0\} \cup E.$$

The linear independence constraint qualification (LICQ) holds at a point $x \in \mathcal{F}$ if the gradients of all active constraints are linearly independent. Now we can state the well-known first-order necessary conditions, which most of us learnt as Kuhn–Tucker or Karush-Kuhn-Tucker (KKT) conditions.

Theorem: Suppose x^* is a local solution of the problem (NLP) that satisfies the (LICQ). Then there exist unique Lagrange multipliers ν_i^*, $i \in I$, and λ_j^*, $j \in E$, such that the following conditions are satisfied.

$$\nabla f(x^*) - \sum_{i \in I} v_i^* \nabla g_i(x^*) - \sum_{j \in E} \lambda_j^* \nabla h_j(x^*) = 0, \tag{3}$$

$$g_i(x^*) \geq 0, \text{ for all } i \in I, \tag{4}$$

$$h_j(x^*) = 0, \text{ for all } j \in E, \tag{5}$$

$$v_i^* g_i(x^*) = 0, \text{ for all } i \in I, \tag{6}$$

$$v_i^* \geq 0, \text{ for all } i \in I. \tag{7}$$

Again we may ask when we can be assured that a solution to the KKT conditions is actually a solution to the nonlinear optimization problem (NLP). If the feasible region $\mathcal{F}$ is a convex set (see Simon and Blume, 1994), and the objective function f is convex on $\mathcal{F}$, then the problem (NLP) is called a convex programming problem, and any local solution is also a (global) solution of (NLP). For example, if the functions h_j, $j \in E$, are all linear and the functions $-g_i$, $i \in I$, are all convex, then $\mathcal{F}$ is a convex set. In this case, if f is convex, too, indeed any solution to the KKT conditions is a solution to (NLP).

Many of the most popular numerical methods for solving nonlinear constrained optimization problems take advantage of the KKT conditions in one form or another. First, however, we describe the basic version of the simplex method for linear programming which does not rely on first-order conditions.

4.1 The simplex method

When the objective function f and the constraint functions g_i, $i \in I$, and h_j, $j \in E$, are all linear functions in the variables $x \in \mathbb{R}^n$, then the constrained optimization problem is a linear programming problem, or 'linear program' for short. Linear programs have a standard form,

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^T x \\ \text{s.t.} \quad & Ax = b \quad \text{(LP)} \\ & x \geq 0 \end{aligned}$$

where $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and A is an $m \times n$ matrix. We can easily transform any linear programming problem with arbitrary linear inequalities and unbounded variables into this standard form.

The development of the simplex method in the late 1940s (Dantzig, 1949) for solving linear programs is generally regarded as the beginning of the modern era of optimization (Nocedal and Wright, 2006). The simplex method is, however, not only of historical importance but to this day the perhaps most widely used tool in optimization outside economics. Here we describe the fundamental idea of the basic version of the simplex method.

The system of equality constraints, $Ax=b$, has m equations in the n decision variables. For the LP to be an interesting optimization problem it must be the case that $m<n$. If $m>n$ then either the linear system is overdetermined and so the feasible region is empty and the LP has no solution, or the system can be simplified so that the number of equations does not exceed the number of variables. The same conclusions apply for the case $m=n$ if the matrix A is singular. If $m=n$ and A has full rank, then the feasible region consists of at most one point and the LP is trivial. We can therefore assume that the system of equality constraints is underdetermined, that is, it has fewer equations than variables. Modern computer implementations of the simplex method start with a pre-processing phase, which transforms a given linear programming problem by removing redundancies and possibly even also eliminating some variables.

We can easily calculate some of the solutions to the system $Ax=b$. If we choose m of the n variables and set the remaining $n-m$ variables to zero, then the system reduces to a square system of m linear equations, which can be solved via Gaussian elimination. The chosen variables for which we solve the system are called `basic variables', while those variables that we set to zero are called 'non-basic variables'. Solving the m linear equations in the m basic variables can lead to three possible outcomes. First, we may detect that the system has no solution. Second, a solution, called basic solution, may exist and it also satisfies the remaining constraints of the LP, namely the sign restrictions $x \geq 0$. In this case the solution is called a 'basic feasible solution'. Third, the solution to the linear system may entail a negative value for at least one variable and thus violate the sign restriction. Such a solution is called 'basic infeasible'. Two basic solutions are called adjacent if their respective sets of basic variables have all but one element in common. The next theorem explains why the basic feasible solutions are of central importance to the linear program.

Theorem: If the problem (LP) has a non-empty feasible region, then there is at least one basic feasible solution. If the problem (LP) has an optimal solution then it has the following properties.

1. 1. At least one optimal solution is a basic feasible solution.
2. 2. If (LP) has a unique solution, then this optimal solution is basic feasible.
3. 3. If a basic feasible solution x^* has an objective function value that is not larger than the objective function values at all its adjacent basic feasible solutions, then x^* is a solution of (LP).
4. 4. If the feasible region is bounded and a basic feasible solution x^* has an objective function value that is strictly less than the objective function value at all its adjacent basic feasible solutions, then x^* is the unique solution of (LP).□

This theorem provides the foundation for the basic approach of the simplex method. According to the first statement of the theorem, if an optimal solution exists then there must be a basic feasible solution that is optimal. Thus, for solving the problem (LP) it suffices to only examine basic feasible solutions. In principle we could now find a solution to the problem (LP) by simply calculating all its basic solutions and then choosing a basic feasible solution with the smallest objective function value. We would not want to do this in practice, however, since the number of possible basic solutions is $\binom{n}{m}$ and thus is huge for many applications. The simplex method prescribes a smart way of searching through the basic feasible solutions. Starting from some basic feasible solution, the simplex searches for another basic feasible solution with a lower objective function value. From a computational

standpoint it is much quicker to examine only adjacent basic feasible solutions. The information we have from having solved a linear system in, for example, the variables x_1, x_2, x_3, greatly simplifies finding a solution in the variables x_2, x_3, x_5. Therefore, the simplex considers only adjacent feasible solutions and chooses one of them by exchanging one basic variable against one non-basic variable and solving the resulting system of linear equations. On most (but not all) steps of the method the objective function value decreases. This process repeats itself until the method reaches a basic feasible solution without any adjacent basic feasible solutions having a lower objective function value. The third statement of the theorem (which is a special version of the convex programming property for linear programs) then ensures that the simplex method has found an optimal solution.

We illustrate the basic ideas underlying the simplex method in the following example.

Example 5: Consider the following linear programming problem.

$$\begin{aligned} \max_{x_1, x_2} \quad & 3x_1 + 4x_2 \\ \text{s.t.} \quad & x_1 + 2x_2 \le 10 \\ & x_1 + x_2 \le 8 \\ & x_2 \le 4 \\ & x_1 \ge 0 \\ & x_2 \ge 0 \end{aligned}$$

Linear programming problems with two variables allow a beautiful graphical representation, which greatly helps us to gain some intuition for the simplex method. Figure 3 shows the feasible region of this linear programming problem.

This problem has three inequality constraints and two sign restrictions. The introduction of three so-called slack variables transforms the inequalities into equations. This introduction of new variables is just one of several simple transformations that allow us to rewrite any linear programming problem into a linear program in standard form; see Dantzig (1963) or many other linear programming books. Here we obtain the following linear program.

$$\begin{array}{lrcrcrcrcrcl} \min_{x_1, x_2, x_3, x_4, x_5} & -3x_1 & - & 4x_2 & & & & & & & & \\ \text{s.t.} & x_1 & + & 2x_2 & + & x_3 & & & & & = & 10 \\ & x_1 & + & x_2 & + & & x_4 & & & & = & 8 \\ & & & x_2 & + & & & x_5 & & & = & 4 \\ & & x_1, & x_2, & & x_3, & & x_4, & & x_5 & \ge & 0 \end{array}$$

This linear program has $n = 5$ variables and $m = 3$ constraints.

Table 3 lists all $\binom{5}{3} = 10$ possible combinations of basic variables, the corresponding basic solution (if it exists), whether this solution is feasible, and the objective function values $z = -3x_1 - 4x_2$ for the basic feasible solutions. For example, the basic solution (4, 4, –2, 0, 0) is obtained by setting x_4=x_5=0 and then solving the remaining three equations

$$x_1 + 2x_2 + x_3 = 10,\ x_1 + x_2 = 8,\ x_2 = 4,$$

in the three basic variables x_1, x_2, x_3. This basic solution is infeasible since x_3=–2 violates the non-negativity constraint on this variable. The basic variables x_1, x_3, x_4 lead to the three equations

$$x_1 + x_3 = 10,\ x_1 + x_4 = 8,\ 0 = 4,$$

which obviously have no solution. We can relate the nine basic solutions to points in the graph of the feasible region in Figure 3. The five feasible solutions are represented by disks while the four infeasible solutions are given by circles. We can easily identify the coordinates of the nine indicated points with the values of the original variables x_1 and x_2 in the nine basic solutions. But where are the later introduced slack variables? The values of these variables at a basic solution show us where the corresponding point in the figure is in relation to the three constraints. The basic solution (4,4,–2,0,0) is represented by the point (4,4) in the graph. This point lies on the lines representing the second and third constraints, since x_4=x_5=0, and outside the first constraint, since x_3<0.

The simplex method quickly solves this problem. Starting from the basic feasible solution that corresponds to the origin in Figure 3 it moves through the basic feasible solutions ('BFS') listed in Table 4 to find the optimal basic feasible solution (6,2,0,0,2). Figure 4 illustrates the steps of the simplex method. Starting from the point (0,0) it moves upwards to the point (0,4) with an objective function value of z=– 16, then to (2,4) with z=– 22 and finally to (6,2) with z= – 26. The basic feasible solution corresponding to this last point has a strictly lower objective function value than both its adjacent basic feasible solutions at (2,4) and (8,0) and hence it must be the unique optimal solution.In Figure 4 only the visited points are indicated by disks and the iso-objective function lines for the values – z of the original objective function (from the maximization problem) at these points.□

All basic solutions

Basic variables	Basic solution	Property	z
x_1, x_2, x_3	(4, 4, –2, 0, 0)	Infeasible	–
x_1, x_2, x_4	(2, 4, 0, 2, 0)	Feasible	–22
x_1, x_2, x_5	(6, 2, 0, 0, 2)	Feasible	–26
x_1, x_3, x_4	–	No Solution	–
x_1, x_3, x_5	(8, 0, 2, 0, 4)	Feasible	–24
x_1, x_4, x_5	(10, 0, 0, –2, 4)	Infeasible	–
x_2, x_3, x_4	(0, 4, 2, 4, 0)	Feasible	–16
x_2, x_3, x_5	(0, 8, –6, 0, –4)	Infeasible	–
x_2, x_4, x_5	(0, 5, 0, 3, –1)	Infeasible	–
x_3, x_4, x_5	(0, 0, 10, 8, 4)	Feasible	0

Iterates of the simplex method

Basic Variables	BFS	z
x_3, x_4, x_5	(0, 0, 10, 8, 4)	0
x_2, x_3, x_4	(0, 4, 2, 4, 0)	–16
x_1, x_2, x_4	(2, 4, 0, 2, 0)	–22
x_1, x_2, x_5	(6, 2, 0, 0, 2)	–26

Figure 3

Feasible region of the (LP)

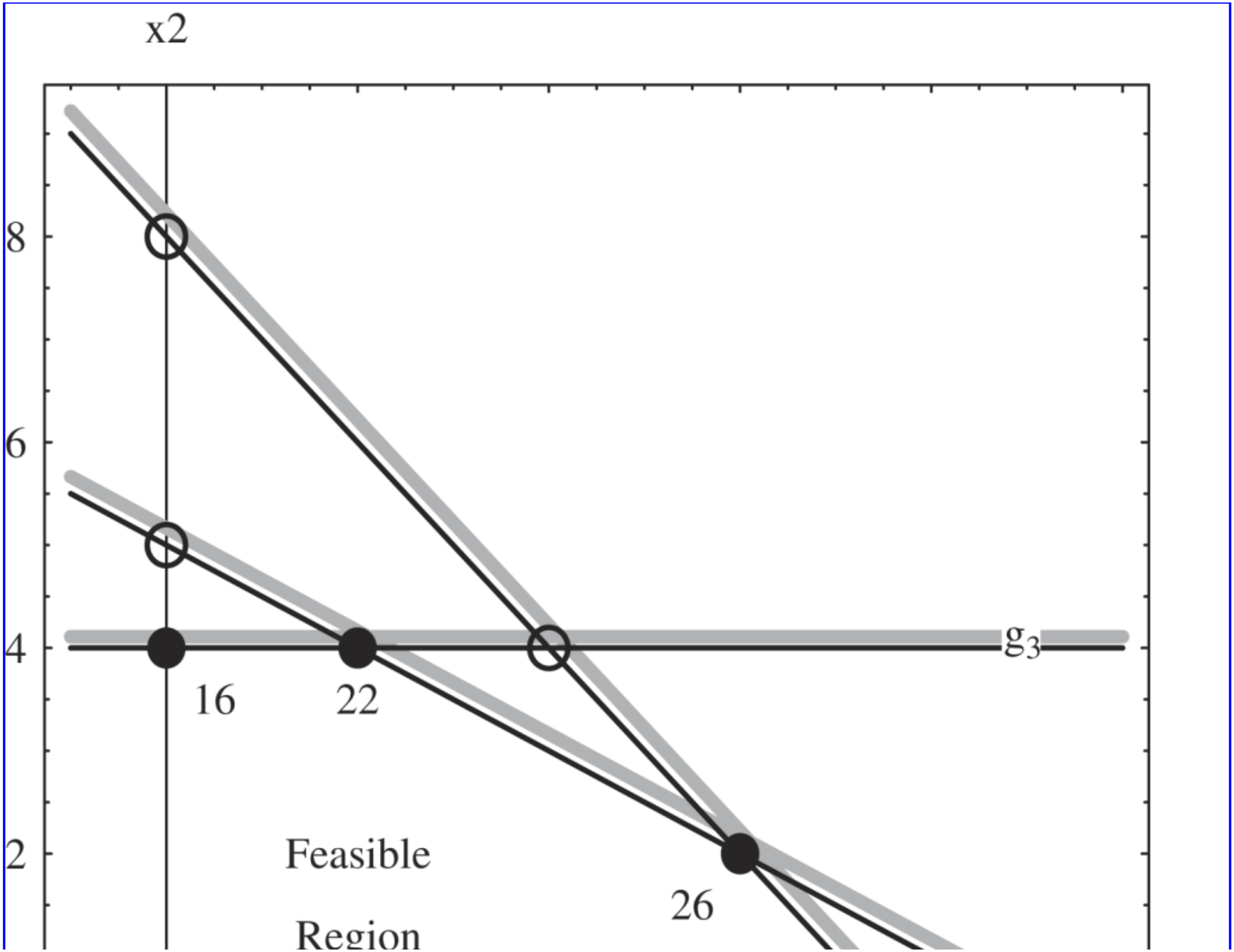
x2
8
6
4
2
16
22
26
g3
Feasible
Region

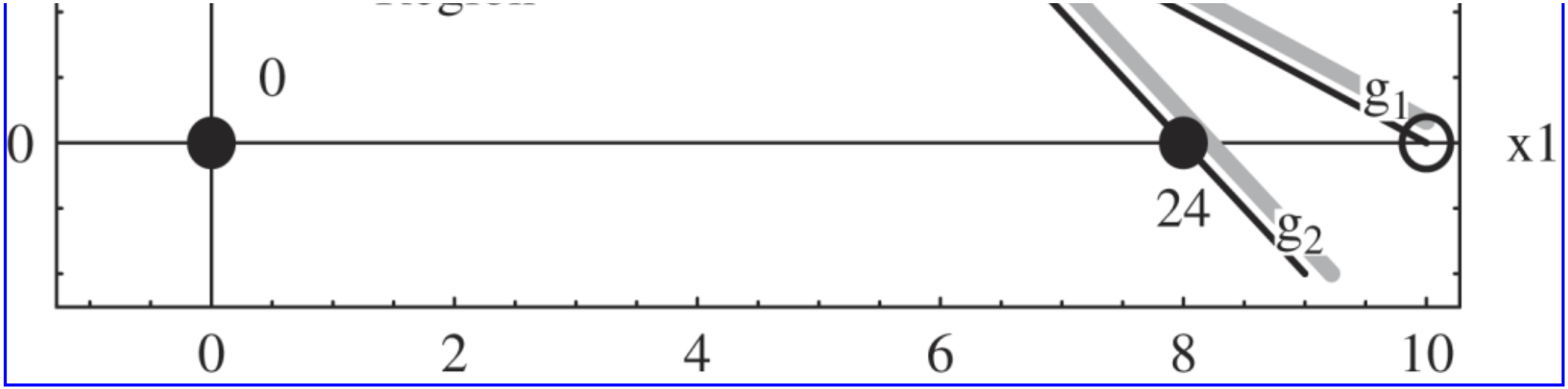

Figure 4
Solving the (LP)

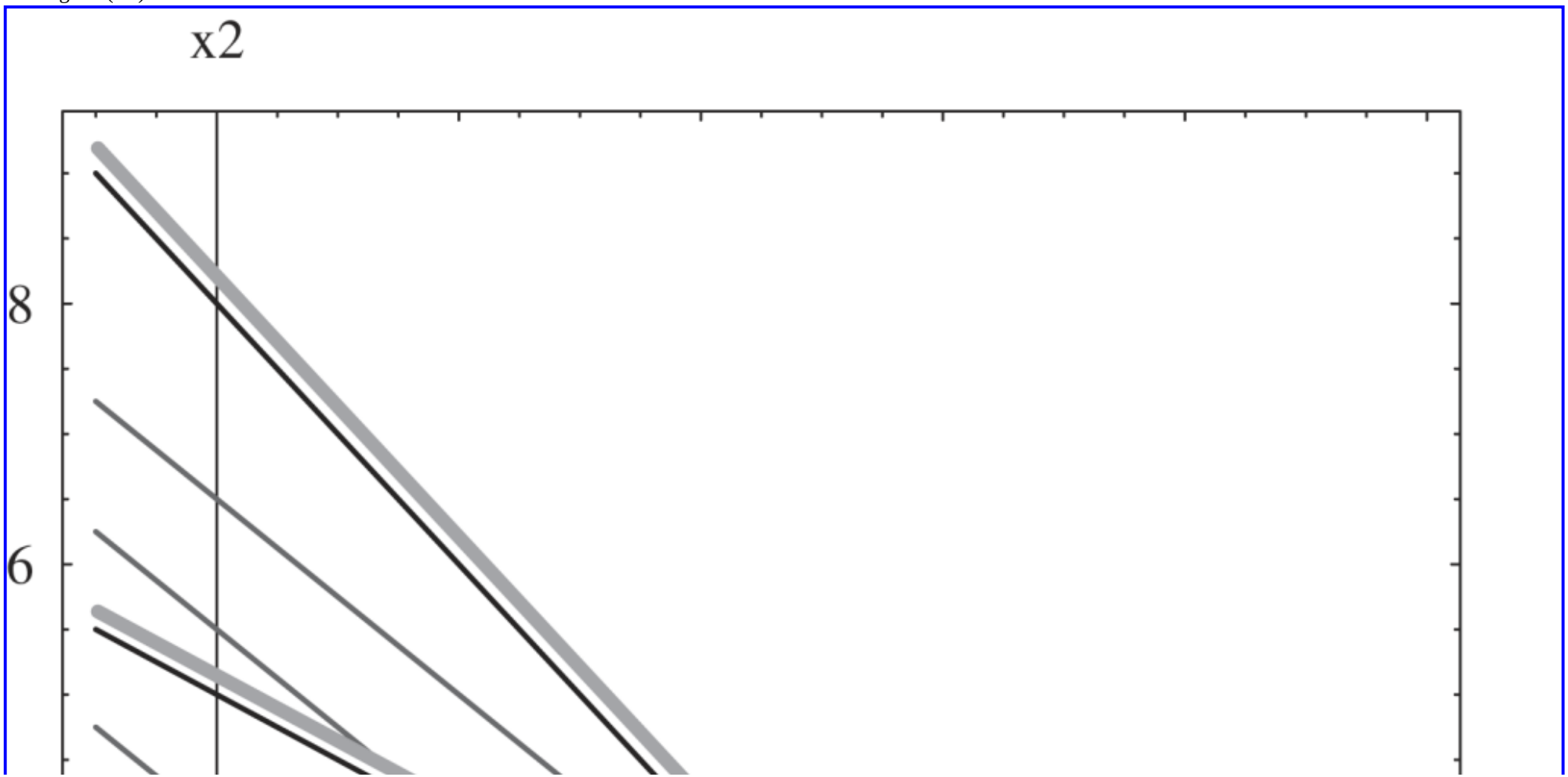

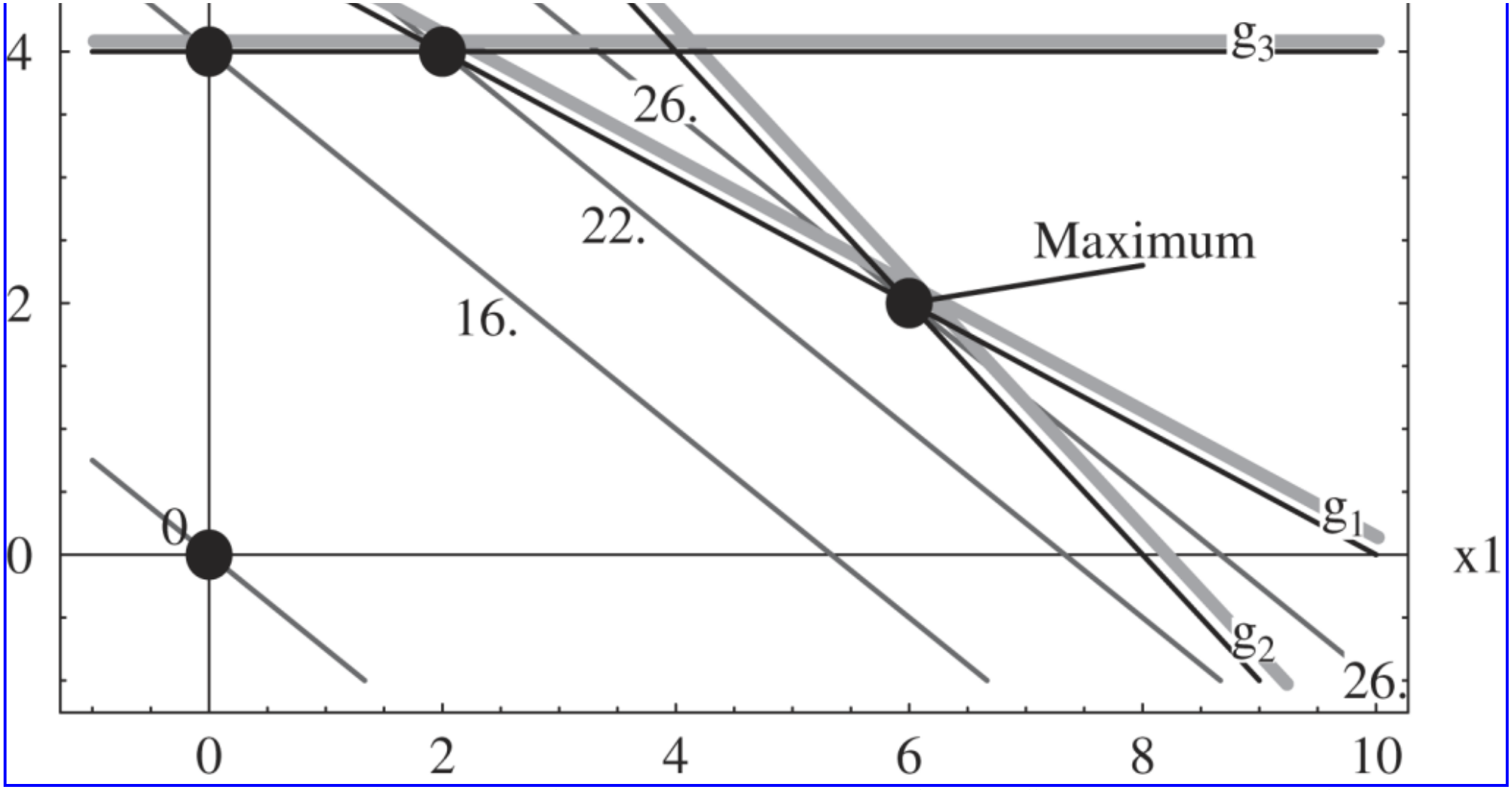

We have conveyed only the basic principle of the simplex method for solving linear programming problems. Of course, an efficient and robust implementation of the simplex algorithm must address many technical details; see Fletcher (1987) or once again Nocedal and Wright (2006). The classical reference for the theory of the simplex method is the book by Dantzig (1963).

The simplex method is highly efficient on virtually all practical problems, but there do exist pathological problems on which it shows very poor performance. In these worst-case problems the running time of the simplex method grows exponentially in the dimension of the problems. (In a nutshell, the method visits far too many basic feasible solutions until it finds the optimal one.) Therefore, the simplex method is of exponential complexity. Although these examples are irrelevant for practical applications, they generated interest in the development of different algorithms that would show better worst-case running times, in particular, that would have running times that grow only polynomially in the size of the problems. The first linear programming algorithm with polynomial complexity was the ellipsoid method of Khachiyan (1979). Although this method has polynomial complexity it is useless for actual computations, and apparently there has never been a serious practical implementation. The projective algorithm of Karmarkar (1984) started what is nowadays called the ‘interior-point revolution’. This algorithm both has polynomial complexity and is of practical use, although the initial claims about its supposedly stellar practical performance were shown to be outrageous. The projective algorithm has long been superseded by more efficient methods, and the field of interior-point methods remains an active area of research to this day.

4.2 The idea of interior-point methods

Primal-dual methods are an important subclass of interior-point methods for solving constrained optimization problems. Here we give a basic outline of such a method for solving linear programs. The Karush–Kuhn–Tucker conditions for a linear programming problem in standard form are as follows.

$$A^T\lambda + s = c \tag{8}$$

$$Ax = b \tag{9}$$

$$x_i s_i = 0,\ i = 1, 2, \ldots, n \tag{10}$$

$$x \geq 0 \tag{11}$$

$$s \geq 0 \tag{12}$$

These first-order conditions characterize both the optimal solution of the given linear program and of its dual. (See Dantzig, 1963, or any book on linear programming for the definition of the dual of a linear program.) That fact motivates the name ‘primal-dual’ method.
Interior-point methods (approximately) solve a sequence of perturbed problems. Consider the following perturbation of the first-order conditions.

$$A^T\lambda + s = c \tag{13}$$

$$Ax = b \tag{14}$$

$$x_i s_i = \mu,\ i = 1, 2, \ldots, n \tag{15}$$

$$x > 0 \tag{16}$$

$$s > 0 \tag{17}$$

Observe that the complementarity condition (10) has been replaced by the equations (15) for some positive scalar $\mu > 0$. Assuming that a solution $(x^{(0)}, \lambda^{(0)}, s^{(0)})$ to this system is given for some initial value of $\mu^{(0)} > 0$, interior-point methods decrease the parameter μ and thereby generate a sequence of points $(x^{(k)}, \lambda^{(k)}, s^{(k)})$ that satisfy the non-negativity constraints on the variables strictly, $x^{(k)} > 0$ and $s^{(k)} > 0$. This property led to the name 'interior-point' method. In the limit, as μ is decreased to zero, a point satisfying the original first-order conditions is reached. The set of solutions to the perturbed system,

$$C = \{x(\mu), \lambda(\mu), s(\mu) | \mu > 0\}$$

is called the central path.

The method is rather intuitive at this point. Given an iterate $(x^{(k)}, \lambda^{(k)}, s^{(k)})$ for some parameter value $\mu^{(k)}$ decrease the parameter to $\mu^{(k+1)} < \mu^{(k)}$ and determine the next iterate $(x^{(k+1)}, \lambda^{(k+1)}, s^{(k+1)})$. Implementing this method requires handling of many details. For example, it is often difficult to find a feasible starting point $(x^{(0)}, \lambda^{(0)}, s^{(0)})$ of the perturbed system. The most important step in the method is to solve the system (13)–(15) in each iteration (while maintaining the inequalities (16, 17)). Observe that this system consists of $2n$

+*m* linear and bilinear equations in as many variables. We can apply a nonlinear equations solver to this model. A popular approach is to use Newton's method for solving nonlinear systems of equations; see Judd (1998) or Miranda and Fackler (2002). The difficulty is to maintain the strict non-negativity constraints on the variables $x^{(k+1)}$ and $s^{(k+1)}$. An alternative approach for solving the parameterized system of equations is the application of path-following methods; see Nocedal and Wright (2006). Intuitively we can think of interior-point methods to be closely related to homotopy continuation methods for solving nonlinear systems of equations; see Allgower and Georg (1979).

Example 6: We revisit the linear program from Example 5. The perturbed first-order conditions (13)–(17) for this (LP) are as follows.

$$
\begin{aligned}
y_1 + y_2 + s_1 + 3 &= 0 \\
2y_1 + y_2 + y_3 + s_2 + 4 &= 0 \\
y_1 + s_3 &= 0 \\
y_2 + s_4 &= 0 \\
y_3 + s_5 &= 0 \\
x_1 + 2x_2 + x_3 - 10 &= 0 \\
x_1 + x_2 + x_4 - 8 &= 0 \\
x_2 + x_5 - 4 &= 0 \\
x_i \cdot s_i &= \mu, \quad i = 1, 2, \ldots, n \\
x_1, x_2, \ldots, x_5 &> 0 \\
s_1, s_2, \ldots, s_5 &> 0
\end{aligned}
$$

Table 5 displays the values for the variables $x_1, x_2, \ldots, x_5$ at some points on the central path for small values of μ. We observe how the central path moves through the interior of the feasible region, see Figure 3, and converges to the optimal solution as $\mu \to 0$.□

Solutions $x^*(\mu)$ for small μ

μ	$x_1(\mu)$	$x_2(\mu)$	$x_3(\mu)$	$x_4(\mu)$	$x_5(\mu)$
1	5.9775	1.5451	0.9323	0.4774	2.4549
0.5	6.0305	1.7311	0.5073	0.2384	2.2689
0.1	6.0029	1.9478	0.1014	0.0492	2.0522
0.01	6.0000	1.9950	0.0100	0.0050	2.0050
0	6	2	0	0	2

By now the conceptual differences between the simplex method and interior-point methods are transparent. In geometric terms, the simplex method moves on specific points around the boundary of the feasible region until it finds a corner point corresponding to an optimal basic feasible solution. Interior-point methods move through the interior (or some methods even through the exterior) of the feasible region but they do not move within the boundary. Instead they approach the boundary only in the limit. In computational terms, the typical iteration of an interior-point method is relatively expensive to compute but can make significant progress towards the solution. Conversely, an iteration of the simplex method is relatively inexpensive but the method often requires a larger number of iterations.

Obviously the question arises of which of these two basic approaches is better for solving linear programs in practice. The answer depends very much on the nature of the problem. Currently the best available computer programs are efficient implementations of the dual simplex method (a special variant of the described standard simplex method) and primal–dual interior-point methods. Simplex method computer programs are usually faster on problems of small or medium size (say, of fewer than a million variables and constraints) while interior-point methods tend to do better on many but certainly not all large problems. If the user has significant prior information about the optimal solution, such as a good initial guess for an optimal basic feasible solution, then the simplex method is often much faster. The reason for this is that the simplex method is much easier to ‘warm-start’ than interior-point methods. In summary, interior-point methods and the simplex method are both important and useful algorithms for solving linear programs in practice.

Before we turn to interior-point methods for nonlinear optimization problems, we outline the basic concepts of another class of optimization algorithms. Penalty methods are quite intuitive and some of their ideas are relevant for interior-point methods but they are also of interest on their own.

4.3 Penalty methods

The basic idea of penalty methods is to replace the constrained optimization problem (NLP) by an unconstrained optimization problem and to solve the new problem instead. The objective function for the new unconstrained problem is the original objective plus a new term for each constraint. The new term is zero when the original constraint is satisfied but is positive if the original constraint is violated. The simplest and perhaps most intuitive penalty function is the quadratic penalty function.
To start we consider a nonlinear optimization problem with only equality but no inequality constraints,

$$\min_{x\in\mathbb{R}^n} f(x)$$
$$\text{s.t. } h_j(x) = 0 \; j\in E.$$

For such a problem we can define a penalty function

$$Q(x;\mu) = f(x) + \mu\left(\sum_{j\in E} h_j^2(x)\right)$$

with a penalty parameter $\mu>0$. The idea of the penalty function method is to minimize the function Q for increasing values of μ. Observe that the function Q inherits its differentiable properties from the functions f and h_j, $j\in E$, of the original problem, and so we can use unconstrained optimization methods for minimizing $Q(x;\mu)$. In addition, as we generate a sequence $\mu^{(k)}$, k=0,1,2,…, we can use the previously calculated minimizer $x^{(k)}(\mu^{(k)})$ as initial guesses for the problem with $\mu^{(k+1)}$. This intuitive approach has a strong theoretical foundation, as the following theorem reveals; see Nocedal and Wright (2006).

Theorem: Consider a sequence $\{\mu^{(k)}\}$ of penalty parameters with $\mu^{(k)}\to\infty$. Suppose that $x^{(k)}$ is the exact global minimizer of $Q(x;\mu_k) = f(x) + \mu(\Sigma_{j\in E} h_j^2(x))$. Then every limit point x^* of the sequence $\{x^{(k)}\}$ is a global solution of the (NLP).□
Although this result is nice from a theoretical viewpoint, it does not directly apply to practical applications. Of course, we typically cannot determine the exact minimizer of the penalty function and have to account for errors in the numerical approximation. The discussion in Nocedal and Wright (2006) shows that things get more complicated in practice once we allow for approximation errors. In addition, the penalty function may have many other stationary points that are not global or even local minimizers. The penalty function may even be unbounded if the penalty parameter μ is too small. At the other extreme, for very large values of μ the unconstrained minimization problem becomes more difficult, and the Hessian of Q gets ill-conditioned. All kinds of numerical problems arise that need to be carefully addressed in robust computer implementations of the quadratic penalty method; see Nocedal and Wright (2006).
For the general problem (NLP) with inequality and equality constraints we can define the penalty function as

$$Q(x;\mu) = f(x) + \mu\left(\sum_{i\in I}(\max(-g_i(x),0))^2 + \sum_{j\in E} h_j^2(x)\right).$$

Now, however, things get more complicated since Q will typically not be twice differentiable. As a result the new unconstrained problem becomes more difficult to solve.
In addition to the quadratic penalty method several other such approaches exist and are used in practice. Nocedal and Wright (2006) describe non-differentiable penalty functions and the augmented Lagrangian method. Here we finish our discussion with an illustration of the quadratic approach.
Example 7: Consider a simple example of the classical portfolio optimization problem (Markowitz, 1952). An investor wants to allocate her entire wealth across three securities with respective expected returns of 4 per cent, 8 per cent and 12 per cent. If she invests the respective portions x_1, x_2, x_3 in the three assets, then the variance of such a portfolio is $x_1^2 + 5x_2^2 + 3x_2x_3 + 10x_3^2$. The investors wants to minimize this variance under the condition that the expected return of her portfolio is at least 9 per cent. To simplify this illustration of the quadratic penalty method we exploit the fact that at the optimal solution the lower bound on the expected return is binding and thus write the investor's portfolio allocation problem as a nonlinear optimization problem with only equality constraints.

$$\begin{aligned} \min_{x_1, x_2, x_3} \quad & x_1^2 + 5x_2^2 + 3x_2x_3 + 10x_3^2 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 - 1 = 0 \\ & 4x_1 + 8x_2 + 12x_3 - 9 = 0 \end{aligned}$$

The quadratic penalty function for the investor's portfolio optimization problem is

$$Q(x; \mu) = x_1^2 + 5x_2^2 + 3x_2x_3 + 10x_3^2 + \mu((x_1 + x_2 + x_3 - 1)^2 + (4x_1 + 8x_2 + 12x_3 - 9)^2).$$

We can easily solve the unconstrained problem with the basic Newton method as described in Section 3.1.2. Table 6 shows the solution to the unconstrained minimization of the penalty function for increasing values of μ.□

Solutions $x^*(\mu)$ for large μ

μ	$x_1^*(\mu)$	$x_2^*(\mu)$	$x_3^*(\mu)$
1	0.78547	0.30659	0.26008
10	0.42484	0.35361	0.36870
100	0.21880	0.37632	0.42571
1000	0.18852	0.37962	0.43403
10000	0.18535	0.37996	0.43490
∞	0.185	0.38	0.435

Observe that the nonlinear optimization problem in this example has a quadratic objective function and linear constraints. Such optimization problems constitute a special and important subclass of problems called quadratic programs. Their special properties give rise to efficient solution methods, and we would not want to solve large quadratic programs with a penalty method. Nocedal and Wright (2006) present several algorithms for quadratic programming. Since solving quadratic programs is comparatively easy, an integral part of some algorithms for more general nonlinear optimization problems, such as the sequential quadratic programming methods, is to repeatedly solve quadratic programs that are derived as approximations for the more general problem.

4.4 The logarithmic barrier method

Logarithmic barrier methods are a particular type of interior-point methods for the solution of nonlinear optimization problems. We illustrate the basic idea of these methods for an inequality-constrained minimization problem.

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} \quad f(x) \\ &\text{s.t.} \quad g_i(x) \geq 0 \quad i\in I. \end{aligned}$$

We can combine the objective function and the constraints and define a penalty function for this optimization problem by

$$P(x;\mu) = f(x) - \mu\sum_{i\in I}\ln g_i(x),$$

where $\mu>0$ is called the barrier parameter and the expression $\sum_{i\in I}\ln g_i(x)$ is called a logarithmic barrier function. Each logarithmic term $-\ln g_i(x)$ tends to infinity as x approaches the boundary given by $g_i(x)\geq 0$ from the interior of the feasible region. This effect of the logarithmic terms will decrease as the barrier parameter μ becomes smaller. The idea of the logarithmic barrier method is now to let the parameter μ converge to zero. Under some conditions the optimal solution $x^*(\mu)$ of the unconstrained optimization problem $\min_{x\in\mathbb{R}^n} P(x;\mu)$ converges to the optimal solution of the original constrained optimization problem as μ tends to zero. Note that the logarithm ensures that $g_i(x^*(\mu))>0$ for all $\mu>0$, that is, the solution to the unconstrained minimization problem is in the strict interior of the original constraints. This property represents a crucial distinction between this variant of an interior-point method and an active-set method such as the simplex method, which always tracks the set of binding constraints at a given iterate.
Observe that the first-order conditions for the penalty function problem are given by

$$\nabla_x P(x;\mu) = \nabla f(x) - \sum_{i\in I}\frac{\mu}{g_i(x)}\nabla g_i(x) = 0.$$

Now define for all $i\in I$

$$v_i(\mu) := \frac{\mu}{g_i(x)}.$$

Note that since $\mu>0$ by definition we have that $\nu_i(\mu)>0$. Thus, at a stationary point of the penalty function the following conditions hold.

$$\nabla f(x) - \sum_{i\in I} v_i\nabla g_i(x) = 0 \quad g_i(x) - s_i = 0 \text{ for all } i\in I \quad v_i s_i = \mu \text{ for all } i\in I \quad v_i > 0 \text{ for all } i\in I \quad s_i > 0 \text{ for all } i\in I.$$

This set of conditions is just the primal-dual interior-point conditions for our original problem. We see that conditions (13)–(17) are just the specialization of these conditions for the linear programming model. And just like in the illustration of the section 4.2 we are interested in taking the parameter μ to zero. Unfortunately we do not have the space here to properly state a formal theorem. To make a long story short, under a few additional technical conditions, most notably second-order conditions of optimality, the following statements hold for a local solution x^* at which the KKT conditions are satisfied for some Lagrange multipliers ν^*.

1. 1. The local minimizer $x^*(\mu)$ of $P(x; \mu)$ in some neighbourhood of x^* with $\lim_{\mu \downarrow 0} x^*(\mu) = x^*$ uniquely defines a continuously differentiable vector function $x^*(\mu)$ for all sufficiently small μ.
2. 2. The function $x^*(\mu)$ yields Lagrange multipliers $\nu(\mu)$ satisfying $\lim_{\mu \downarrow 0} \nu(\mu) = \nu^*$ where $\nu^* g_i(x^*)=0$.

An algorithm for solving the constrained problem is apparent now. For a given value of μ solve the unconstrained optimization problem with the objective function P. Then reduce μ stepwise to zero and follow the path of solutions $x^*(\mu)$. In the limit we can find the local solution x^* of the original problem. While this approach works in principle, it entails various difficulties. For example, the Hessian matrix of P becomes ill-conditioned for small values of μ. For this and many other technical issues see Gould and Leyffer (2002). Here we just illustrate the fundamental idea with an example.

Example 8: We revisit the consumer's utility maximization problem from Example 3 once again. The consumer has a utility function $u(x_1, x_2, x_3) = \sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}$ over three goods and faces the budget constraint $x_1 + x_2 + x_3 \leq 1$. We formulate the consumer's problem as the constrained minimization problem

$$\begin{aligned} \min_{x_1, x_2, x_3} \quad & -\left(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}\right) \\ \text{s.t.} \quad & 1 - x_1 - x_2 - x_3 \geq 0. \end{aligned}$$

The unconstrained function including a logarithmic barrier function for this minimization problem is

$$P(x_1, x_2, x_3; \mu) = -\left(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}\right) - \mu \ln (1 - x_1 - x_2 - x_3).$$

Table 7 displays solutions to this unconstrained problem for a few values of μ. Note that as $\mu \to 0$ the optimal solution approaches the optimal solution of the original utility maximization problem.

In all our examples so far we ignored the sign restriction of the variables. We could do that since the utility functions exhibit an Inada property, that is, $\lim_{x_i \to 0} \frac{\partial u}{\partial x_i} = +\infty$, and so we hope that a solver starting at a strictly positive solution will only iterate through such solutions (although we have to be careful in practice). But, of course, we can easily take the non-negativity constraints explicitly into account and consider the following problem.

$$\begin{aligned} \min_{x_1, x_2, x_3} \quad & -\left(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}\right) \\ \text{s.t.} \quad & 1 - x_1 - x_2 - x_3 \geq 0 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \\ & x_3 \geq 0 \end{aligned}$$

Note that the condition (LICQ) is always satisfied since not all four constraints can be satisfied simultaneously. As long as three constraints are binding (LICQ) holds. The unconstrained function including a logarithmic barrier function for this minimization problem is

$$P(x_1, x_2, x_3; \mu) = -\left(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}\right) - \mu(\ln(1 - x_1 - x_2 - x_3) + \ln(x_1) + \ln(x_2) + \ln(x_3)).$$

Table 8 displays solutions to this unconstrained problem for a few values of μ. Again we observe that $x^*(\mu) \to x^*$ as $\mu \to 0$.□

Solutions $x^*(\mu)$ for small μ

μ	$x_1^*(\mu)$	$x_2^*(\mu)$	$x_3^*(\mu)$
1	0.0421124	0.168450	0.379012
0.5	0.0547198	0.218879	0.492478
0.1	0.0677112	0.270845	0.609401
0.01	0.0710478	0.284191	0.639430
0.005	0.0712379	0.284952	0.641141

Solutions $x^*(\mu)$ for small μ

μ	$x_1^*(\mu)$	$x_2^*(\mu)$	$x_3^*(\mu)$
1	0.219696	0.270563	0.331757
0.5	0.195664	0.278956	0.389721
0.1	0.124696	0.286370	0.543846
0.01	0.0789861	0.285758	0.630009
0.005	0.0753165	0.285727	0.636309

Strangely enough, some of the modern and best interior-point algorithms are based on work predating Karmarkar (1984). For example, Frisch (1955) had already proposed an interior-point method based on logarithmic barrier functions for solving linear programs. A full early history with many results on barrier functions is Fiacco and McCormick (1968).

4.5 Sequential quadratic programming

Sequential quadratic programming (SQP) methods are among the most effective constrained optimization techniques, particularly when nonlinear constraints are present. These algorithms belong to the class of active-set methods that keep track of the binding constraints at each step. For a description of the basic ideas we consider a minimization problem with only equality constraints. (But these methods are much more widely applicable.)

$$\min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } h_j(x) = 0,\ j \in E. \tag{18}$$

The KKT conditions for this problem are as follows.

$$\nabla f(x) - \sum_{j \in E} \lambda_j \nabla h_j(x) = 0, \tag{19}$$

$$h_j(x) = 0,\ j \in E. \tag{20}$$

These conditions are a system of $n+m$ nonlinear equations in the n variables x and the m Lagrange multipliers λ. Newton's method for solving nonlinear equations is now a natural approach for solving this system. The Jacobian of the left-hand side of the system (19)–(20) is given by

$$\begin{bmatrix} H(x) - \sum_{j \in E} \lambda_j H_j(x) & -A(x)^{\mathrm{T}} \\ A(x) & 0 \end{bmatrix},$$

where the matrix $A(x)^{\mathrm{T}} = [\nabla h_1(x), \ldots, \nabla h_l(x)]$ is the collection of the gradient vectors of all constraints $h(x) = (h_j(x))_{j \in E = \{1,2,\ldots,J\}}$. The matrix $H_j(x)$ denotes the Hessian matrix of the constraint function h_j at the point x. For a given point $(x^{(k)}, \lambda^{(k)})$ the Newton step is then determined by the linear system

$$\begin{bmatrix} H(x^{(k)}) - \sum_{j \in E} \lambda_j^{(k)} H_j(x^{(k)}) & -A(x^{(k)})^{\mathrm{T}} \\ A(x^{(k)}) & 0 \end{bmatrix} \begin{bmatrix} s_x^{(k)} \\ s_\lambda^{(k)} \end{bmatrix} = - \begin{bmatrix} \nabla f(x^{(k)}) - \sum_{j \in E} \lambda_j^{(k)} \nabla h_j(x^{(k)}) \\ h(x^{(k)}) \end{bmatrix}$$

resulting in the new iterate $(x^{(k+1)}, \lambda^{(k+1)}) = (x^{(k)} + s_x^{(k)}, \lambda^{(k)} + s_\lambda^{(k)})$. Note that this last system is equivalent to the following linear system.

$$\begin{bmatrix} H(x^{(k)}) - \sum_{j\in E} \lambda_j^{(k)} H_j(x^{(k)}) & A(x^{(k)})^{\mathrm{T}} \\ A(x^{(k)}) & 0 \end{bmatrix} \begin{bmatrix} s_x^{(k)} \\ -\lambda^{(k+1)} \end{bmatrix} = - \begin{bmatrix} \nabla f(x^{(k)}) \\ h(x^{(k)}) \end{bmatrix} \tag{21}$$

Now consider the following quadratic optimization problem (QP).

$$\min_{s\in\mathbb{R}^n} \frac{1}{2} s^{\mathrm{T}} \left(H(x) - \sum_{j\in E} \lambda_j H_j(x) \right) s + \nabla f^{\mathrm{T}}(x) s$$
$$\text{s.t.} \quad A(x) s + h(x) = 0$$

The left-hand side of the constraints are a first-order (Taylor) approximation of the constraint function *h* of the original optimization problem. The objective function of (QP) is a second-order approximation of the difference $f(x+s)-f(x)$. The KKT conditions for the problem (QP) are as follows.

$$\left(H(x) - \sum_{j\in E} \lambda_j H_j(x) \right) s + \nabla f(x) - A^{\mathrm{T}}(x) v = 0 \tag{22}$$

$$A(x) s + h(x) = 0 \tag{23}$$

Observe that these KKT conditions at a point $(x^{(k)}, \lambda^{(k)})$ are equivalent to (21). Solving the first-order conditions of the original optimization problem with Newton's method is, under some technical conditions, equivalent to solving the quadratic optimization problem (QP). A Newton step at a given point $(x^{(k)}, \lambda^{(k)})$ is the same as solving the (QP) at this point. The idea of SQP methods is now to repeatedly solve this quadratic problem to generate a sequence of iterates that converges to a local solution of the original problem. Various good methods for solving quadratic optimization problems exist and can be applied to the problem (QP). Moreover, when combined with line search or trust region methods the approach has useful global convergence properties. Gould and Leyffer (2002) and Nocedal and Wright (2006) discuss details of line search and trust region SQP methods.

5 Global optimization

We emphasized repeatedly that most practical algorithms for solving nonlinear optimization problems search for a solution only to the (necessary) first-order conditions, that is, they search for a local solution. Unless we are solving a convex programming problem or an unconstrained minimization problem of a convex function, we often cannot be sure that a computed local solution is indeed an approximate solution to the problem at hand; recall Example 2. Only occasionally other additional knowledge, perhaps some particular property of an underlying economic model, may assure us that we found an optimal solution. Obviously it would be helpful to have methods for general non-convex problems that may not, or are at least less likely to, get stuck in only locally optimal solutions. Here we lay out two approaches for global optimization. We describe the basic ideas of some popular metaheuristics and, subsequently, the very promising area of research in polynomial optimization, which is likely going to produce powerful tools for economic problems.

5.1 Metaheuristics

Metaheuristics provide a general framework and basic guidelines for developing specific heuristics for solving optimization problems. While the underlying principles are very general, typically a method must be carefully tailored in order to obtain an effective algorithm for the special problem at hand. Most metaheuristics were originally developed for solving discrete optimization problems, such as integer or combinatorial problems. Their principal ideas can also be applied to come up with heuristics for continuous nonlinear optimization problems.

The central problem of most nonlinear optimization methods is the possibility of getting stuck at a locally optimal solution. Many methods allow only for iterative steps that lead to an improvement in the objective function value, but, for an exception, see the discussion on nonmonotone techniques in Conn, Gould and Toint (2000) and Nocedal and Wright (2006). Such methods cannot get away from a locally optimal solution. In order to escape from such a local solution we must allow our search procedure, at least sometimes, to move into a non-improving search directions; that is, temporarily the objective function value of the sequence of iterates may increase (in a minimization problem). Three metaheuristics that are supposed to escape local solution are tabu search, simulated annealing, and genetic algorithms. The latter two methods are examples of stochastic approaches for optimization. Here we give a description of the basic ideas underlying these three methods and refer to Brandimarte (2006), Judd (1998) and the citations in those books for details.

5.1.1 Tabu search

The choice of non-improving search directions must be carefully managed to avoid repeatedly returning to a previously found optimal solution. Such cycling may occur if, after a non-improving step away from a local solution, the algorithm takes an improving step and immediately returns to the previously found local solution. A tabu search procedure imposes at every iteration a list of search directions that the algorithm is not allowed to pursue. For example, if the method just took a step in the direction $s^{(k)}$ then it may not be allowed to examine a neighbourhood of search directions around $-s^{(k)}$ for the next few iterations. In order to avoid memory problems in practical implementations, the tabu list usually consists only of the most recent steps taken. Of course, many technical issues need to be addressed to obtain a robust and efficient algorithm. The treatment of these issues usually depends greatly on the specific problem.

5.1.2 Simulated annealing

Simulated annealing is another metaheuristic that helps an algorithm to escape from locally optimal solutions. Instead of choosing only iterates that decrease the objective function in a minimization problem, simulated annealing methods also accept with some probability new iterates that increase the objective. The probability of accepting an iterate $x^{(k+1)}$ if $f(x^{(k+1)}) > f(x^{(k)})$ is

$$e^{-\frac{f(x^{(k+1)}) - f(x^{(k)})}{T}}$$

with a parameter $T>0$. Simulated annealing methods typically start out with a fairly large value for T and then decrease it to 0. Observe that for large values of T the heuristic is likely to accept non-decreasing iterates, and so it allows the method to explore the feasible region. As T decreases the probability of acceptance of non-decreasing iterates of a fixed size also decreases. In the limit $T \to 0$ the method allows only iterates that decrease the objective function value. The perhaps simplest rule for reducing T is to start from a high value T_0 and

then to set

$$T_{l+1} = \alpha T_l \text{ for some } 0 < \alpha < 1.$$

The basic ideas of simulated annealing are derived from an analogy of minimization with the physical annealing process of slowly cooling metals in order to reach a strong low-energy solid state. This analogy motivates the particular probability function for accepting increasing iterates and explains why the parameter T is called the temperature of the process. The rule of decreasing T is analogously called the cooling schedule. The earliest applications of simulated annealing were combinatorial problems; see Kirkpatrick, Gelatt and Vecchi (1983) as well as Cerny (1985).

5.1.3 Genetic algorithms

Genetic algorithms are derived from the analogy of finding better and better solutions with the theory of biological evolution of selecting fitter and fitter members of a species. As a result the literature on genetic algorithms uses terminology from evolutionary biology. Iterates in tabu search and simulated annealing algorithms are a single point. Contrary to that, genetic algorithms work with a set (‘generation’) of several current points. A genetic algorithm constructs a sequence of such sets. In a given iteration the objective function is evaluated at the points in the set (‘fitness of a member’). The method then chooses elements of the set in a probabilistic fashion in order to build new elements for the next set. Usually the probability of an element being chosen is the higher the better its objective function value. Several ways to construct new elements exist. A standard operation is the so-called crossover. Given two elements (‘parents’) $x^{(k)}$ and $y^{(k)}$ in the set the crossover operation leads to

$$x^{(k+1)} = \left(x_1^{(k)}, \ldots, x_l^{(k)}, y_{l+1}^{(k)}, \ldots, y_n^{(k)}\right), \tag{24}$$

$$y^{(k+1)} = \left(y_1^{(k)}, \ldots, y_l^{(k)}, x_{l+1}^{(k)}, \ldots, x_n^{(k)}\right), \tag{25}$$

where the method chooses some arbitrary break point l in the n-dimensional vectors. The idea behind crossover is to preserve some parts of the original elements and at the same time generate quite arbitrarily new elements (‘children’) that are far away from the original ones, and thereby to escape local solution. Another type of operation aimed at achieving this goal is to randomly exchange an element in a member $x^{(k)}$ by another value (‘mutation’). While these approaches have proven useful in combinatorial optimization, it is quite apparent that they may run into severe difficulties for constrained problems. Many technical details must, therefore, be resolved before these ideas yield a useful heuristic approach for solving an optimization problem.

The monograph by Holland (1975) popularized genetic algorithms. The basic ideas of computer simulations of evolution are much older.

Any heuristic method derived from a metaheuristic will always be an ad hoc approach to the problem at hand. Just like the standard methods of nonlinear optimization presented in this article, they are not guaranteed to find the solution of a problem. And, while such heuristics have proven useful in discrete optimization, they are generally regarded as inferior to the modern standard optimization techniques for continuous optimization. An economist's first choice of a solution method for a continuous optimization problem, particularly when nonlinear constraints are present, should always be one of the standard methods.

5.2 Polynomial functions

A substantial number of prominent economic models involves only polynomial functions, equations or inequalities. Even problems that at first appear to be non-polynomial can sometimes be transformed into having only polynomial expression. For example, the first-order conditions for the standard log-utility maximization problem

$$\max_{x \in \mathbb{R}^n} \sum_{i=1}^{n} \ln(x_i) \text{ s.t. } \sum_{i=1}^{n} p_i(x_i - \omega_i) = 0$$

for prices p_i and endowments ω_i, i=1,…, n, can be written in polynomial form,

$$1 - \lambda p_i x_i = 0, \quad i \in \{1, \ldots, n\}, \tag{26}$$

$$\sum_{i=1}^{n} p_i(x_i - \omega_i) = 0, \tag{27}$$

where λ denotes the Lagrange multiplier. Polynomial functions and equations can be analysed using tools from computational algebraic geometry (Cox, Little and O'shea, 1997). Global optimization with polynomials is an active field of research in mathematics; see, for example Lasserre (2001), Parrilo and Sturmfels (2003), and the book by Sturmfels (2002) and the citations therein. It is possible (at least in theory) to compute all local minima of polynomial functions. Similarly, it is possible to compute all solutions to a polynomial system of equations. With further expected advances in the theory of polynomial optimization and ever increasing speed of modern computers, these tools will soon have an impact in economics. For first results see computation of general equilibria (new developments).

6 Popular optimization software in economics

This section covers software packages and modelling languages that are frequently used in economics to solve optimization problems. This list is by no means exhaustive, and many other software products for solving optimization problems exist.

Perhaps the most popular software for numerical work in economics is MATLAB (MATLAB is a registered trademark of The MathWorks, Inc.). Computational economics and finance textbooks such as Brandimarte (2006), Kendrick, Mercado and Amman (2006) and Miranda and Fackler (2002) use MATLAB to solve economic problems. Other popular packages include GAUSS (GAUSS is a registered trademark of Aptech Systems, Inc.) and Mathematica (Mathematica is a registered trademark of Wolfram Research, Inc.) All three languages offer solvers for nonlinear optimization problems, which are continuously enhanced to solve larger and more difficult problems. Here we just list a few features of these software packages.

MATLAB has an optimization toolbox containing routines for solving both unconstrained and constrained nonlinear optimization problems. Methods for unconstrained problems include quasi-Newton and trust region techniques. The solvers for constrained optimization include an SQP method. MATLAB also has specialized methods for nonlinear least

square problems; however, most of these solvers are considered to be of only mediocre quality. Much better solvers in MATLAB are available through the NAG toolbox (NAG is a registered trademark of The Numerical Algorithms Group, Inc.) The NAG Foundation Toolbox provides access to the large set of numerical routines contained in the Fortran-based NAG Foundation Library, which contains routines for constrained and unconstrained optimization.

The high-level matrix programming language GAUSS includes an applications module for constrained optimization that uses an SQP method in conjunction with several line search methods or a trust region method. GAUSS has some specialized modules for constrained maximum likelihood problems. For Mathematica there exists a global optimization package, which contains various functions for optimization. These functions are designed to search for global optima for problems with hundreds of variables. The monograph by Bhatti (2000) comes with an optimization toolbox for Mathematica that includes all the methods presented in this article.

These high-level languages are popular in economics because they are easy to learn and quickly facilitate solving problems of moderate size. For larger problems with thousands or even hundreds of thousands of variables, however, they are not reliable and certainly too slow. Economists interested in solving large problems need to use alternative software. An excellent alternative is the use of algebraic modelling languages.

The General Algebraic Modeling System (GAMS) is a high-level modelling language designed for mathematical programming and optimization; see Rosenthal (2006) for a user's guide. GAMS consists of a language compiler and a family of integrated high-performance solvers. GAMS is tailored for complex, large-scale modelling applications, and allows the user to build large models. It has a long history of successful applications in economics, particularly in solving large-scale computable general equilibrium (CGE) models. AMPL (Fourer, Gay and Kernighan, 2003) is an algebraic modelling language for mathematical programming, which allows users to set up and solve a great variety of optimization problems. The user has access to many popular and sophisticated solvers.

An exciting environment for solving optimization problems is the Network-Enabled Optimization System (NEOS); see Czyzyk, Mesnierand and Moré (1998) and Ferris, Mesnier and Moré (2000). NEOS is an optimization site that allows users to submit optimization problems over the Internet. The user does not need to download any solver but can just send optimization problems to NEOS and choose from a list of solvers. NEOS has access to many of the most current and powerful optimization routines. NEOS returns a solution and some runtime statistics to the user. Unfortunately, NEOS has been largely ignored by many economists.

7 Mathematical programs with equilibrium constraints

Mathematical programs with equilibrium constraints (MPECs) are currently at the frontier of numerical analysis. Economic models that can be classified as ‘leader-follower’ games are examples of MPECs. Suppose that the economic variables can be partitioned into x, those chosen by the ‘leader’ (government, employer, market maker, mechanism designer, and so on), and y, those chosen by the ‘followers’ (taxpayers, employees, traders, and so on) or determined in equilibrium (such as price). Suppose that the leader's payoff is $f(x, y)$ and that the equilibrium value y given x is represented by a combination of inequality conditions, $c(x,y) \geq 0$, and complementarity constraints, $0 \leq y \perp F(x, y) \geq 0$, where $0 \leq y \perp F(x, y) \geq 0$ if and only if $0 \leq y$, $F(x,y) \geq 0$, and $y^T F(x,y)=0$. Equality constraints can be added without difficulty. The constraints correspond to, for example, budget and incentive constraints, and the complementarity constraints model the optimality conditions of the followers including any Lagrange multipliers. Then the leader's problem and the corresponding equilibrium are given by the solution to the MPEC

$$\begin{aligned} \max_{x,y} \quad & f(x, y) \\ \text{s.t.} \quad & c(x, y) \geq 0 \\ & 0 \leq y \perp F(x, y) \geq 0 . \end{aligned}$$

MPECs present many mathematical challenges; the constraints are non-convex and reformulations as standard nonlinear optimization problems violate fundamental stability assumptions. Despite these facts, nonlinear optimization methods applied to such reformulations have been successful at solving some MPECs. For example, Chen et al. (2006) solve MPECs derived from some large-scale electricity market models. But they also show the limitations of the nonlinear optimization approach, and advocate the development of robust algorithms for solving MPECs that directly exploit the structure of the complementarity constraints. The development of such methods is under way. The ability to solve large and complicated MPECs will greatly enhance economic modelling in many areas and will likely make MPECs a key tool of computational economic analysis in the future.

See Also

- computation of general equilibria
- computation of general equilibria (new developments)
- computational methods in econometrics
- dynamic programming
- linear programming
- nonlinear programming
- operations research
- simplex method for solving linear programs

I am grateful for helpful discussions with Sven Leyffer and am indebted to Ken Judd, Annette Krauss, and in particular Che-Lin Su for detailed comments on earlier drafts. I also thank the editors Larry Blume and Steven Durlauf for a careful review of my initial submission.

Bibliography

Allgower, E.L. and Georg, K. 1979. *Introduction to Numerical Continuation Methods*. New York: John Wiley & Sons. Reprinted by SIAM Publications, 2003.

Bhatti, M.A. 2000. *Practical Optimization Methods: With Mathematica Applications*. New York: Springer-Verlag.

Brandimarte, P. 2006. *Numerical Methods in Finance and Economics: A MATLAB-Based Introduction*. New York: John Wiley & Sons.

Cerny, V. 1985. A thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications* 45, 41–51.

Chen, Y., Hobbs, B.F., Leyffer, S. and Munson, T.S. 2006. Leader–follower equilibria for electric power and NO_x allowances markets. *Computational Management Science* 4, 307–30.

Conn, A.R., Gould, N.I.M. and Toint, P.L. 2000. *Trust-Region Methods*. Philadelphia: SIAM.

Cox, D., Little, J. and O'shea, D. 1997. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. New York: Springer-Verlag.

Czyzyk, J., Mesnier, M.P. and Morè, J. 1998. The NEOS server. *IEEE Journal on Computational Science and Engineering* 5, 68–75.

Dantzig, G.B. 1949. Programming of inter-dependent activities II, mathematical model. *Econometrica* 17, 200–211. Also in Koopmans, T.C., ed. *Activity Analysis of Production and Allocation*. New York: John Wiley & Sons, 1951.

Dantzig, G.B. 1963. *Linear Programming and Extensions*. Princeton: Princeton University Press.

Ferris, M.C., Mesnier, M.P. and Moré, J. 2000. NEOS and Condor: solving nonlinear optimization problems over the Internet. *ACM Transactions on Mathematical Software* 26, 1–18.

Fiacco, A.V. and McCormick, G.P. 1968. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: John Wiley & Sons, Inc. Reprinted by SIAM Publications, 1990.

Fletcher, R. 1987. *Practical Methods of Optimization*. Chichester: John Wiley & Sons.

Fourer, R., Gay, D.M. and Kernighan, B.W. 2003. *AMPL: A Modeling Language For Mathematical Programming*. Pacific Grove, CA: Brooks/Cole–Thomson Learning.

Frisch, R.A.K. 1955. The logarithmic potential method of convex programming. Technical Report, University Institute of Economics, University of Oslo, Norway.

Gould, N.I.M. and Leyffer, S. 2002. An introduction to algorithms for nonlinear optimization. In *Frontiers in Numerical Analysis*, ed. J.F. Blowey, A.W. Craig and T. Shardlow. Berlin, Heidelberg: Springer-Verlag.

Holland, J.H. 1975. *Adaptation in Natural and Artifical Systems*. Ann Arbor: University of Michigan Press.

Judd, K.L. 1998. *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorics* 4, 373–95.

Kendrick, D.A., Mercado, P.R. and Amman, H.M. 2006. *Computational Economics*. Princeton: Princeton University Press.

Khachiyan, L.G. 1979. A polynomial algorithm in linear programming. *Soviet Mathematics Doklady* 20, 191–4.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science* 220, 671–80.

Lasserre, J.B. 2001. Global optimization with polynomials and the problem of moments. *SIAM Journal of Optimization* 11, 796–817.

Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7, 77–91.

Miranda, M.J. and Fackler, P. 2002. *Applied Computational Economics and Finance*. Cambridge, MA: MIT Press.

Nocedal, J. and Wright, S.J. 2006. *Numerical Optimization*. New York: Springer.

Parrilo, P.A. and Sturmfels, B. 2003. Minimizing polynomial functions. *Algorithmic and Quantitative Real Algebraic Geometry, DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 60, 83–99.

Rosenthal, R.E. 2006. *GAMS – A User's Guide*. Washington, DC: GAMS Development Corporation. Online. Available at http://www.gams.com/docs/gams/GAMSUsersGuide.pdf, accessed 7 February 2007.

Simon, C.P. and Blume, L. 1994. *Mathematics for Economists*. New York: W. W. Norton.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3, 107–20.

Sturmfels, B. 2002. *Solving Systems of Polynomial Equations*, vol. 97, CBMS Regional Conference Series in Mathematics, Providence, RI: American Mathematical Society.

How to cite this article

Schmedders, Karl. "numerical optimization methods in economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?

id=pde2008_N000148> doi:10.1057/9780230226203.1209

The New Palgrave Dictionary of Economics Online

Nurkse, Ragnar (1907 – 1959)

Kaushik Basu

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balanced growth; development economics; full employment; international policy coordination; international trade; Nurkse, R.; poverty traps

Article

Nurkse was born on 5 October 1907 on an estate where his father was an overseer, near the village of Viru in Estonia. His father was Estonian and his mother of Swedish origin. Ragnar Nurkse was educated in Tallinn, Tartu, Edinburgh and Vienna. From 1934 to 1945 he worked as an economist with the League of Nations and from 1945 until his death he was a professor at Columbia University. He wrote on international currency questions, trade, vicious circles of poverty and on balanced growth. In 1959 he delivered the Wicksell Lectures in Stockholm. Exhausted by the lectures, he went to Geneva and while taking a stroll on Mont Pèlerin he collapsed and died of a heart attack or stroke on 6 May 1959. The Wicksell Lectures were published posthumously (Nurkse, 1961).

One of Nurkse's two most important books was *International Currency Experience: Lessons of the Inter-War Period* (1944). It was published by the League of Nations, and though it did not carry the name of any author, this was (excepting chapter 6) the work of Nurkse. From this and several other of his writings, what comes out most clearly is Nurkse's pragmatism. Though he was one of the originators of the doctrine of balanced growth, he never minimized the role of international trade. However, he believed that the scope for trade-based expansion for Third World countries was much less in the 20th century than it was in the 19th century. Balanced growth could supplement this and even enlarge the scope for trade. Balanced growth and international trade, Nurkse argued, ‘are really friends, not enemies’ (Haberler and Stern, 1961, p. 257).

Nurkse had a deep concern for full employment. He viewed exchange rate adjustments and trade restrictions as legitimate measures for preventing balance of payments difficulties from translating into unemployment and domestic instability. He stressed that trade restrictions ought to be used as temporary measures. With the emergence of Keynesian macroeconomics, Nurkse came to have faith in effective-demand management as a tool for maintaining employment in the face of trade adversities. This also led

him to argue for some international coordination of domestic policies.

Nurkse's other important (and, in my opinion, more important) book was *Problems of Capital Formation in Underdeveloped Countries* (1953). Here he developed the important idea that though the producer of each commodity may find an expansion unprofitable because of limitations of the market, a coordinated expansion of all productive activities could be profitable for all producers. Hence, atomistic behaviour on the part of producers could trap an economy *within* its production possibility frontier. This idea had been discussed earlier – most notably by Rosenstein-Rodan (1943) and more distantly by Young (1928) – but Nurkse took it further. While this work has been the basis of several debates in development economics (for critiques and formalizations, see Flemming, 1955; Findlay, 1959), it has the scope for further research, especially in the light of recent advances in non-Walrasian equilibrium analysis (see Basu, 1984).

The lack of formalization in Nurkse's work led to much misunderstanding – handsomely contributed to by Nurkse himself – about the policy implications of the poverty-trap doctrine. Nurkse tried to clarify these in his Ankara lectures in 1957 and his posthumously published note in *Oxford Economic Papers* (1959), both reprinted in Haberler and Stern's (1961) collection. The potential of this branch of development economics remains large.

See Also

- balanced growth

Selected works

1944. *International Currency Experience: Lessons of the Interwar Period*. Princeton: League of Nations.

1947. International monetary policy and the search for economic stability. *American Economic Review, Papers and Proceedings* 35, 569–80.

1953. *Problems of Capital Formation in Underdeveloped Countries*. Oxford: Basil Blackwell.

1954. International investment today in the light of nineteenth-century experience. *Economic Journal* 64, 744–58.

1959. Notes on ‘unbalanced growth’. *Oxford Economic Papers* 11(3), 295–7.

1961. *Patterns of Trade and Development*. The Wicksell Lectures, Oxford: Basil Blackwell.

Bibliography

Basu, K. 1984. *The Less Developed Economy: A Critique of Contemporary Theory*. Oxford: Basil Blackwell.

Findlay, R.V. 1959. International specialisation and the concept of balanced growth: comment. *Quarterly Journal of Economics* 73, 339–46.

Flemming, J.M. 1955. External economies and the doctrine of balanced growth. *Economic Journal* 65, 241–56.

Haberler, G. and Stern, R.M., eds. 1961. *Equilibrium and Growth in the World Economy: Economy Essays by Ragnar Nurkse*. Cambridge, MA: Harvard University Press.

Rosenstein-Rodan, P.N. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53, 202–11.

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

How to cite this article

Basu, Kaushik. "Nurkse, Ragnar (1907–1959)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000112> doi:10.1057/9780230226203.1210

The New Palgrave Dictionary of Economics Online

nutrition and development

Paul Glewwe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The nutritional status of children and adults is primarily determined by consumption of foodstuffs that contain macronutrients and micronutrients and by the incidence of gastro-intestinal diseases. Insufficient nutrition among young children has particularly severe negative consequences. Factors that lead to better nourished children include better-educated mothers, higher household income, potable water and sanitary toilet facilities. The most effective nutrition programmes target children during their first two years of life; such programmes increase life cycle income by raising children's levels of education. Economists should focus their research efforts on empirical studies that use panel data and data from randomized trials.

Keywords

bargaining power; education; labour productivity; Malthus, T.; medical care; nutrition; panel data; random experiments; simultaneity bias

Article

Economists have studied human nutrition since Thomas Malthus published his *Essay on the Principle of Population* in 1798. His pessimistic predictions about economic growth and human welfare proved to be incorrect; in today's developed countries the primary nutrition problem for the majority of the population is obesity, not lack of food. Yet in low-income countries the nutritional status of both children and adults can have a substantial effect on the incomes of individuals and on the rate of economic growth. In addition, the nutritional status of poor children remains a policy concern in almost all developed countries. This article reviews recent research on the factors that affect the nutritional status of children and adults, and the causal impact of child and adult nutrition on income and on other economic outcomes. It focuses on developing countries, where nutritional problems are the most severe and thus their consequences are the largest. For recent studies of nutrition in developed countries, see Kenkel and

Manning (1999) and Currie (2000).

Factors that determine child and adult nutritional status

Malnutrition can be defined as the lack of sufficient nutrients for human growth and/or for carrying out daily work and non-work activities. Nutrients can be classified into two broad groups: macronutrients, which are primarily calories and protein; and micronutrients, the vitamins and minerals that are essential for good health. Lack of macronutrients is often caused by insufficient consumption of staple foodstuffs, while lack of micronutrients often reflects an unbalanced diet. The most serious micronutrient deficiencies in developing countries are lack of iron, iodine, vitamin A and zinc. (See Behrman, Alderman and Hoddinott, 2004, and the references therein for further details.)

The nutritional status of children and adults in developing countries (and in developed countries) is primarily determined by consumption of foodstuffs that contain macronutrients and micronutrients and by the incidence of gastro-intestinal diseases that interfere with the body's ability to extract macronutrients and micronutrients from those foodstuffs. By far the most serious manifestation of such diseases is the incidence of diarrhoea among very young children. The consumption of foodstuffs is in turn determined by household income and food and non-food prices, as proposed by standard demand theory. Some countries have programmes that provide households with food rations or food coupons (stamps) that can be used to purchase food items, which effectively loosens households' budget constraints. The incidence of gastro-intestinal diseases is mainly due to three factors: exposure to infectious diseases, the health knowledge of both children and adults, and the availability (and prices) of medicines and medical care services. A final consideration is the allocation of foodstuffs and medical care within the household, which is likely to depend on the relative bargaining power of key household members; in particular, several studies have shown that children are better nourished in households in which their mothers have a relatively high level of bargaining power (see Strauss and Thomas, 1998, for references).

Many economists, nutritionists and other researchers have attempted to identify the most important causes of malnutrition among children in developing countries. This research is motivated in part by estimates indicating that about 30 per cent of the children in those countries are seriously underweight (de Onis et al., 2004) and about 1.7 million children die every year from malnutrition and diarrhoea (WHO, 2003). Careful empirical studies of children's nutritional status in developing countries have provided credible evidence that the following factors have strong causal effects: mother's education; mother's health knowledge; infant breastfeeding; household income; potable water; and modern toilet facilities. Several specific policy interventions have also been shown to have a strong positive impact on child nutrition: oral rehydration therapy (ORT) for children with diarrhoea; monitoring of child growth; programmes that provide health and nutrition information to mothers; and fortification of commonly purchased food items (such as salt and sugar) with selected micronutrients (see World Bank, 2004, and Filmer, 2003, for detailed references).

The factors that determine the nutritional status of adults in developing countries have received less attention from economists and other researchers, primarily because in almost all cases the impact of poor nutrition on adults is thought to be less harmful than the impact of poor nutrition on children. Higher incomes, higher education levels and availability of health care services all have positive impacts on

adult health and nutrition. Yet these impacts are not necessarily very strong; for example, the income elasticity of calorie consumption is quite low (Strauss and Thomas, 1998), although it is somewhat higher for the poorest households. Similarly, Haddad et al. (2003) found that the income elasticity of the rate of child malnutrition is less than one in 11 of the 12 countries they analysed.

While the results summarized in the previous two paragraphs are intuitively plausible, they can be challenged because formidable econometric problems confound attempts to estimate the determinants of the nutritional status of both children and adults. Several recent studies have carefully attempted to overcome problems of simultaneity bias (for example, food intakes, income and medical treatments are all jointly endogenous), but attenuation bias due to measurement error in the explanatory variables has received less attention. The most convincing studies are those based on either panel data or randomized experiments.

Impact of poor nutrition on socioeconomic outcomes

The impacts of poor nutritional status on important socioeconomic outcomes vary according to the age of the individual when he or she is malnourished. It is useful to consider separately the following three age ranges: from birth to about five years (before children are enrolled in primary school), from six years to the early teenage years (when most children are enrolled in school and not working), and from the late teenage years through retirement age (the working-age years).

Empirical evidence suggests that poor nutrition in the first few years of life can have substantial negative consequences for educational outcomes and, eventually, for adult income (see World Bank, 2004, and Glewwe, 2005, for recent reviews). For example, Glewwe, Jacoby and King (2001) show that children in the Philippines who were malnourished during the first two years of their lives start school at a relatively late age and learn less per year while in school. The precise mechanisms are not completely clear, but it is likely that inadequate nutrition during the first years of life affects the physical development of the brain in ways that cannot be easily reversed. For example, iodine deficiency impairs the development of the central nervous system. The reduction in skills obtained from schooling due to poor nutrition in the preschool years almost certainly has large negative impacts on children's incomes when they become adults, and back-of-an-envelope calculations suggest that the benefits (in terms of life cycle income) of programmes to reduce malnutrition among very young children are much higher than the costs (see Glewwe, Jacoby and King, 2001, for an example of such calculations).

Many nutrition programmes in both developed and developing countries are designed to provide nutritious breakfasts and lunches to children on the days they are in school. In developing countries, there is little research on the impact of these programmes on educational outcomes. Almost all of the existing literature suffers from serious estimation problems and/or small sample sizes (see Glewwe, 2005, for a detailed discussion). While it may seem obvious that providing breakfasts and/or lunches to students would increase their learning, parents may reduce the amount of food provided at home in response to provision of meals at school; surprisingly, Jacoby (2002) found no evidence that parents respond in this way. Perhaps the best evidence on the impact of the nutritional status of the learning of school age children is a recent randomized study of Kenyan pre-schools. Vermeersch and Kremer (2005) provide evidence of a positive impact of school feeding programme on learning, although only in schools with more experienced teachers. Further research is needed on the impact on education

outcomes of child nutrition during their years in school. It may be that improvements in schooling outcomes from school feeding programmes are primarily due to increases in daily attendance brought about by those programmes, as opposed to higher nutritional status among children who participate in those programmes.

Finally, many economists have examined the role of nutrition during adulthood on concurrent labour productivity and labour income. Robert Fogel has studied this phenomenon in the United States and Europe in the 18th and 19th centuries (see Fogel, 1999), but the historical data are too incomplete to resolve a host of econometric issues in a convincing way. Some economists have developed ‘efficiency wage’ models in which low nutrition among adults can lead to involuntary unemployment. Strauss and Thomas (1998) provide a recent review of the empirical evidence on the relationship between adult nutrition, labour productivity and income. There is strong evidence that, *Ceteris paribus*, better-nourished adults (as measured by body mass index) are more productive workers. (There is also evidence that taller workers are relatively more productive, but height is primarily determined by nutritional status during childhood.) Yet Swamy (1997) and others have presented strong evidence that the estimated magnitudes of the effect of current nutritional status on worker productivity are far too small to be a cause of unemployment in developing countries.

Much has been learned in recent years about the relationship between nutrition and economic and social outcomes in developing countries, but even more remains to be learned. The evidence to date suggests that the most effective, and most cost-effective, nutrition programmes are those that are targeted to children during their first two years of life, for whom the main benefits are a higher rate of survival into adulthood and an increase in life cycle income brought about by higher levels of education. The most convincing studies are based on either panel data or randomized trials, but such data are available for only a handful of countries. Indeed, the Cebu Longitudinal Health and Nutrition Survey, which covers only one region of the Philippines, is the data source for many of the most convincing studies based on panel data. While randomized trials, such as Gertler's (2004) assessment of the impact of Mexico's *Progresa* programme on children's nutritional status, can be a very effective method for assessing programme and policy impacts, one must wait many years before long-term impacts can be measured. Very little is known about the impact of poor nutrition among school-age children on academic performance and, ultimately, adult income, and the same is true of the impact of policies and programmes designed to improve the nutritional status of adults. A very recent policy option that deserves careful study is the development and provision of genetically modified foodstuffs that contain higher levels of essential nutrients. An example of this is ‘golden rice’ that has been fortified with vitamin A. To provide useful information for policymakers, economists' research efforts in the area of nutrition should not be devoted to developing theoretical models but instead should focus on empirical studies that make careful use of panel data and data from randomized trials. This will require new data collection efforts, but the cost of such data collection is very small compared with the potential benefits.

See Also

- anthropometric history
- child health and mortality
- efficiency wages

- fertility in developing countries
- human capital, fertility and growth
- longitudinal data analysis
- returns to schooling
- treatment effect

Bibliography

Behrman, J., Alderman, H. and Hoddinott, J. 2004. Hunger and malnutrition. In *Global Crises, Global Solutions*, ed. B. Lomborg. Cambridge: Cambridge University Press.

Currie, J. 2000. Child health in developed countries. In *Handbook of Health Economics*, ed. A. Culyer and J. Newhouse. Amsterdam: North Holland.

De Onis, M., Blössner, M., Borghi, E., Frongillo, E. and Morris, R. 2004. Estimates of global prevalence of childhood underweight in 1990 and 2015. *Journal of the American Medical Association* 291, 2600–6.

Filmer, D. 2003. *Determinants of Health and Education Outcomes*. Washington, DC: World Bank.

Fogel, R. 1999. Catching up with the economy. *American Economic Review* 89, 1–21.

Gertler, P. 2004. Do conditional cash transfers improve child health? Evidence from *PROGRESA*'s control randomized experiment. *American Economic Review* 94, 336–41.

Glewwe, P. 2005. The impact of child health and nutrition on education in developing countries: theory, econometric issues and recent empirical evidence. *Food and Nutrition Bulletin* 26, S235–S250.

Glewwe, P., Jacoby, H. and King, E. 2001. Early childhood nutrition and academic achievement: a longitudinal analysis. *Journal of Public Economics* 81, 345–68.

Haddad, L., Alderman, H., Appleton, S., Song, L. and Yohannes, Y. 2003. Reducing child malnutrition: how far does income growth take us?. *World Bank Economic Review* 17, 107–31.

Jacoby, H. 2002. Is there an intrahousehold flypaper effect? Evidence from a school feeding program. *Economic Journal* 112, 196–221.

Kenkel, D. and Manning, W. 1999. Economic evaluation of nutrition policy or, there's no such thing as a free lunch. *Food Policy* 24, 145–62.

Strauss, J. and Thomas, D. 1995. Human resources: empirical modeling of household and family decisions. In *Handbook of Development Economics, Volume 3B*, ed. J. Behrman and T. Srinivasan.

Amsterdam, New York and Oxford: Elsevier Science, North Holland.

Strauss, J. and Thomas, D. 1998. Health, nutrition and economic development. *Journal of Economic Literature* 36, 766–817.

Swamy, A. 1997. A simple test of the nutrition-based efficiency wage model. *Journal of Development Economics* 53, 85–98.

Vermeersch, C. and Kremer, M. 2005. School meals, educational achievement and school competition: evidence from a randomized evaluation. Policy Research Working Paper No. 3523. Washington, DC: World Bank.

WHO (World Health Organization). 2003. *The World Health Report 2003: Shaping the Future*. Geneva: WHO.

World Bank. 2004. *World Development Report 2004: Making Services Work for Poor People*. Washington, DC: World Bank.

How to cite this article

Glewwe, Paul. "nutrition and development." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000113> doi:10.1057/9780230226203.1211

The New Palgrave Dictionary of Economics Online

nutrition and public policy in advanced economies

Janet Currie
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses the measurement of nutritional status of populations and examines two classes of tools that policymakers in advanced economies can use to improve nutrition: targeted food and nutrition programmes, and regulation of the food industry. It presents an overview of the economic rationale for providing nutrition programmes (rather than cash assistance), as well as an analysis of some of the difficulties of providing aid in kind – one of the chief difficulties is low take-up of programme benefits by eligible citizens. The overview of regulations suggests that measures aimed at improving nutrition information may be especially attractive.

Keywords

advertising; asymmetric information; food insecurity; Food Stamp Program (USA); health care regulation; hunger; in-kind transfers; Medicaid (USA); nutrition; nutrition and public policy; technical change; well-being

Article

Measures of nutritional status such as height, body mass index and the prevalence of nutrient-deficiency diseases are now accepted indicators of well-being. Economic development changes nutritional threats to well-being as populations move from scarcity to abundance. Fogel (1994) links the decline of malnutrition to economic growth, and highlights improvements in nutrition as an engine of growth. Cutler, Glaeser and Shapiro (2003) highlight technological change as a factor in reducing the cost of the production and distribution of food: the average household in the United States spent one-third of its income on food in 1960, but spends less than half that amount on food today.
As a result, public policymakers now struggle against a rising tide of obesity and related diseases such as type 2 diabetes using policy tools that were formulated largely to combat the effects of scarcity. The incidence of type 2 diabetes has doubled since 1995 in the United States, where 30 per cent of adults

over the age of 20 are obese. Even in countries like France, which historically had little obesity, rates are increasing rapidly (World Health Organization, 2005). Surprisingly, many people in the United States are both overweight and consuming diets that are deficient in fibre, calcium, potassium, magnesium and vitamin E. This juxtaposition suggests that an excess of calories and a deficit of nutrients may in fact be closely related and reflect poor food choices rather than food scarcity.

This article considers the difficulties involved in tracking the nutritional status of populations and examines two classes of tools that policymakers in advanced economies can use to improve nutrition: targeted food and nutrition programmes and regulation of the food industry.

Measuring nutrition

Tracking the nutritional status of a population over a long period of time is difficult. Much of Fogel's work relies on the records of army veterans, largely because the veterans represent a large group for whom anthropometric measures are available. Birth weight is also available in many populations over long periods of time (cf. Currie and Moretti, 2007).

Going beyond anthropometric measures is generally expensive. Data on food consumption is often collected using food diaries, in which subjects are asked to record everything that they ate (and the amount that they ate) over some specified period such as a day or a week. These entries must then be converted into data about the number of calories from various sources. Clearly, there is likely to be a great deal of measurement error in this type of data, so large sample sizes are needed to uncover any systematic relationships between food intakes and outcomes.

A few data sets such as the National Health and Nutrition Examination Survey (NHANES) in the United States collect information about the levels of specific nutrients using blood and urine tests as well as food diaries. This information is collected as part of a complete physical examination conducted in a mobile clinic. Each wave takes several years to collect, as the mobile examination units travel to interview sites around the country. The expense of collecting the data means that the survey is mounted approximately once a decade. The long intervals between surveys raise additional problems because best practices in terms of ways to measure nutritional status often change between the surveys. Hence, while one can use the NHANES to track changes in body mass index over time, it is difficult to use these data to examine changes in the prevalence of specific nutritional deficiencies.

A fourth source of information about nutrition comes from health surveillance data. Doctors are often required to report the prevalence of specific conditions in their practices to central health agencies. These central agencies in turn can determine how many cases of something like iron deficiency anemia occur in a given population. One suspects that such surveillance systems will tend to underestimate the extent of nutritional deficiencies to the extent that people go untreated, or doctors fail to meet reporting requirements.

Finally, developed countries often produce statistics about the number of people suffering from ‘hunger’. It is important to realize that in advanced economies hunger is a social construct that is not directly related to measures of actual nutritional deficiency. In 1968, a group of physicians issued ‘Hunger in America’, a landmark report documenting appalling levels of malnutrition among poor children. Outright malnutrition is now extremely rare in the developed world. In the United States, people are now classified as hungry if they respond affirmatively to a series of questions in the current population survey. These questions ask whether households are worried about having the money to pay

for food, whether there are times that households go without food because they lack money to pay for it, and whether specific household members go without food. These ‘food insecurity’ questions are inexpensive to ask and can be asked more frequently and consistently than the direct measures of nutritional status can be collected in more episodic surveys.

However, once poverty is controlled for, food insecurity is predictive of poorer nutritional outcomes among older household members, but not among children (Bhattacharya, Currie and Haider, 2004). To say that food insecurity is not a direct measure of nutritional deficiency does not mean that it is unimportant. Food insecurity has been linked to higher levels of hyperactivity, absenteeism, aggression and tardiness as well as impaired academic functioning among children, although these linkages may not be causal.

Targeted food and nutrition programmes

Most advanced economies prefer income support to targeted food and nutrition programmes as a way of improving the nutrition (and overall well-being) of their poorest citizens. In contrast, the United States has an array of food and nutrition programmes targeted to specific low-income groups. School meal (or milk) programmes are an exception, in that they are widespread in advanced economies. Apparently the paternalism involved in creating a feeding programme is acceptable when dealing with children, but not (in many countries) when dealing with adults.

Apart from paternalism, economists have developed an array of rationales for providing benefits (including food) in kind, rather than in cash. One common rationale for government intervention in kind is that malnourished citizens create negative externalities for other citizens, through the psychological distress of those who interact with them, burdens on social programmes and health care systems, or their own inability to work.

A second set of arguments has to do with informational asymmetries. Since the government cannot perfectly identify those who need help, it must create schemes that will encourage self-selection. Such schemes often involve penalizing recipients through stigma or through the imposition of non-trivial transactions costs (see Blackorby and Donaldson, 1988; Besley and Coate, 1991; 1995).

A final rationale is more dynamic: the government fears that cash aid will not be spent as intended, and that recipients will return again and again. The problem is that the government cannot credibly commit to cut off starving people, even if the needy person has squandered past aid (Bruce and Waldman, 1991).

These models shed some light on the question of why in-kind programmes are set up as they are, with often substantial barriers to entry and consequent lack of take-up by the neediest people (see Currie, 2006b, for a discussion of the take-up of these programmes, and of factors that affect it).

A complete survey of the literature assessing US in-kind food and nutrition programmes is beyond the scope of this article, but see Currie (2006a, ch. 3) for more details about the programmes discussed here and evidence regarding their effectiveness. These programmes take various forms and target various groups. The largest and most studied include the Food Stamp Program (FSP), the Supplemental Nutrition Program for Women, Infants, and Children (WIC) and the National School Lunch Program (NSLP). These three programmes have adopted very different approaches to improving nutrition in disadvantaged families.

The NSLP (and the smaller School Breakfast programme) provide free or reduced-price meals

conforming to certain nutritional guidelines directly to their target population. The programme is available in most US government-sponsored schools and serves approximately 27 million lunches every day at a cost of about six billion dollars annually. The FSP provides electronic debit cards that can be redeemed for food with few restrictions on the types of foods which can be purchased. The programme serves about 20 million households at a cost of roughly 19 billion dollars. WIC offers coupons that may be redeemed only for specific types of food (often specific brands), to women, infants, and children under five who are certified to be at nutritional risk. WIC also involves a significant nutrition education component, which is largely absent from the other two programmes. This programme serves about eight million people each month, at a cost of approximately four billion dollars.

WIC packages are tailored to the nutritional needs of each of the target groups. The programme has been credited with virtually eliminating iron deficiency anemia among infants and young children and with improving birth weight and birth outcomes among the most disadvantaged mothers in an extremely cost-effective manner. There is less research available about WIC's effects on young children. In the past, WIC promoted bottle over breast-feeding by giving mothers free infant formula. Ongoing strenuous efforts are being made to promote breast-feeding and give nursing mothers food packages of equal value to those received by mothers getting formula.

The near-unanimous consensus regarding the positive effects of WIC on infant outcomes has been disturbed in recent years by those who argue that there may be unobserved factors that are correlated both with positive infant health outcomes and with WIC participation. While this is true in theory, careful analyses of selection in the WIC programme suggest that it is the most disadvantaged eligible women who participate, and it is unlikely that they have other positive unobserved characteristics that are driving the findings – that is, selection is probably leading to underestimates of the effects of WIC.

At some points WIC has also generated controversy by enrolling more infants than the government estimated to be eligible. A National Academy of Sciences report on the subject found, however, that the number of those eligible was underestimated, and that the programme fell a long way short of full take-up (National Research Council, 2003). The fact that many eligible people do not participate in food and nutrition programmes remains a far more significant problem than participation by ineligible people.

FSP benefits are available to all households with incomes less than 130 per cent of the poverty threshold. FSP benefits can be used to purchase virtually any foods at almost all grocery stores. Since the benefits are generally less than the household's food budget, economic theory suggests that the benefit should be treated in the same way as a cash transfer. But several food stamp ‘cash-out’ experiments in which treatment households were given cash instead of food stamps while control households continued to receive food stamps suggested that the cash-out reduced spending on food. However, Whitmore (2002) re-analysed data from one such experiment and found that only households whose benefits exceeded their food budgets initially reduced spending in response to the cash-out. Thus it appears that the FSP may in fact be no different from a cash transfer. It is thus worth asking whether the FSP plays any role other than serving as an indirect cash safety net that is available to the many US households that do not qualify for any other form of assistance. Given that virtually any type of food can be purchased, the FSP should not be expected to have much impact on the quality of the diet, other than via relaxation of the budget constraint. Evidence that people buy and sell stamps (often doing both within a month) further suggests that FSP benefits are treated like cash.

Studies of the FSP shed a good deal of light on the question of take-up and again suggest that lack of participation by eligible people is a greater problem than participation by those who are ineligible.

Enrolments in the FSP grew rapidly in the early 1990s following the expansion of the federal Medicaid programme. Households could sign up for Medicaid and the FSP at the same office, so households that were attracted by Medicaid also signed up for FSP. Conversely, the 1996 welfare reform in the United States was accompanied by a decline in FSP participation even among those who remained eligible. Those who lost eligibility for cash benefits were no longer automatically eligible for the FSP and the fact that people were now required to go through enrolment procedures for the FSP and to repeat those procedures every three to six months drove many eligible people away. These examples suggest that transactions costs are an important deterrent to enrolment in means-tested transfer programmes.

The NLSP operates in a way that is similar to school meal programmes in many other countries. In the United States the poorest children are eligible for free meals, while slightly better off children are eligible for reduced-price meals, and other children can purchase school meals at ‘full price’. The meals are subject to US government dietary guidelines, which were revised in 1994 to limit the amount of fat and sodium.

Evaluations suggest that the NLSP has successfully raised the consumption of important nutrients. At the same time, meals have been roundly criticized for being high in calories, fat and sodium. Still, the evidence suggests that many American children have extremely unhealthy diets which are improved somewhat through participation in the NLSP.

Like other food and nutrition programmes, the NLSP has been criticized for serving too many ineligible children. The US government conducted several studies of this issue, experimenting with different ways to tighten controls on eligibility. In every case, ‘reforms’ were more likely to discourage eligible children from applying than they were to reduce programme use by ineligible children (see Neuberger and Greenstein, 2003). As a result of these policy experiments, the US government adopted several measures designed to make it easier for poor families to document and maintain eligibility when these programmes were re-authorized in 2004.

Regulation

Traditionally, regulation of the food industry has aimed to ensure the safety of the food supply. However, regulation has been increasingly used as a tool to improve the quality of the diet. Governments in advanced economies have mandated the inclusion of important nutrients such as iodine in salt (which has eliminated goitre), vitamin D in milk (which has helped to eliminate rickets), and folic acid in flour (which has greatly reduced the incidence of neural tube birth defects). Increasingly, regulation is being targeted at the information available to consumers, through labelling and advertising. There is a good deal of evidence that consumers respond to food labels. Ippolito and Mathios (1990) examine the effect of a US government decision to allow cereal makers to advertise the link between fibre and cancer reduction. The change led to increased advertisement of fibre content, as well as other content information, and to increases in the consumption of high-fibre cereals. Ippolito and Mathios (1995) found that consumption of fat had been declining secularly, but that it declined more rapidly after manufacturers were allowed to advertise health claims associated with low-fat products.

It is however, unclear whether food labels have allowed consumers to make food choices that are healthier overall. Marketing studies suggest that few consumers consult labels assiduously and that many are unaware of the nutritional contents of items in their food baskets. Moreover, food-away-from-

home constitutes a large and growing fraction of total consumption and is largely exempt from labelling regulations.

Low socio-economic status households have higher propensities to suffer from nutrition related disorders and are also least likely to use labels. However, some labelling requirements have encouraged manufacturers to reformulate their products in ways that will benefit all consumers, whether or not they read labels. For example, a recent US requirement that manufacturers label ‘transfats’ has led many producers of products such as crackers to substitute transfats with less harmful fats.

Governments have also acted directly in the limitation of the consumption of unhealthy foods. Many US school districts have removed ‘junk food’ and soft drinks from vending machines, and federal legislation that would require this of all school districts has been introduced. France and the United Kingdom have taken similar measures nationally and the UK is going further by banning burgers and processed sausages in schools and requiring two servings of fruit and/or vegetables per day.

The UK also banned the use of celebrities to advertise junk food during children's television programming and the use of film tie-in advertisements in 2006 (*Guardian*, 2006). Several studies indicate that the majority of food advertising directed at children is for relatively unhealthy foods, and a recent report from the National Academy of Sciences (Institute of Medicine, 2006) concluded that children's preferences are significantly swayed by such advertising, and called for either voluntary or regulatory controls on the advertising of food to children.

Given our increasing knowledge about the links between poor food choices and future health, and the rising social costs of providing health care for nutrition-related conditions such as diabetes, additional future regulation is likely. Government intervention can be viewed as a way of reducing the externalities created by poor individual choices, which in turn may be encouraged by food producers who do not bear the social costs created by their products. Economists can contribute to this important public health debate by analysing the costs and benefits of regulation.

See Also

- health outcomes (economic determinants)
- nutrition and development
- poverty alleviation programmes

Bibliography

Bhattacharya, J., Currie, J. and Haider, S. 2004. Poverty, food insecurity, and nutritional outcomes in children and adults. *Journal of Health Economics* 23, 839–62.

Besley, T. and Coate, S. 1991. Public provision of private goods and the redistribution of income. *American Economic Review* 81, 979–84.

Besley, T. and Coate, S. 1995. The design of income maintenance programs. *Review of Economic Studies* 62, 187–221.

Blackorby, C. and Donaldson, D. 1988. Cash versus kind, self-selection and efficient transfers. *American Economic Review* 78, 691–700.

Bruce, N. and Waldman, M. 1991. Transfers in kind: why they can be efficient and nonpaternalistic. *American Economic Review* 81, 1345–51.

Currie, J. 2006a. *The Invisible Safety Net: Protecting Poor Children and Families*. Princeton: Princeton University Press.

Currie, J. 2006b. The take-up of social benefits. In *Poverty, the Distribution of Income, and Public Policy*, ed. A. Auerbach, D. Card and J. Quigley. New York: Russell Sage.

Currie, J. and Moretti, E. 2007. Biology as destiny? Short and long-run determinants of intergenerational transmission of birth weight. *Journal of Labor Economics* 25, 231–64.

Cutler, D., Glaeser, E. and Shapiro, J. 2003. Why have Americans become more obese? *Journal of Economic Perspectives* 17(3), 93–118.

Fogel, R. 1994. Economics growth, population theory, and physiology: the bearing of long-term processes on the making of economic policy. *American Economics Review* 84, 369–95.

Guardian. 2006. Supermarkets feel junk food ad ban bite. 27 November.

Institute of Medicine. 2006. *Food Marketing to Children and Youth: Threat or Opportunity?* ed. M. McGinnis, G. Jennifer and K. Vivica. Washington, DC: National Academies Press.

Ippolito, P. and Mathios, A. 1990. Information, advertising and health choices. *RAND Journal of Economics* 21, 459–80.

Ippolito, P. and Mathios, A. 1995. Information and advertising: the case of fat consumption in the United States. *American Economic Review* 85, 91–5.

National Research Council. 2003. *Estimating Eligibility and Participation in the WIC Program*. Washington, DC: National Research Council.

Neuberger, Z. and Greenstein, R. 2003. *What Have We Learned from FNS' New Research Findings About Overcertification in the School Meals Programs?* Washington, DC: Center on Budget and Policy Priorities.

Whitmore, D. 2002. What are food stamps worth? Working Paper No. 468. Industrial Relations Section, Princeton University.

World Health Organization. 2005. *The challenge of obesity in the WHO European Region. Fact sheet EURO/13/05*. Geneva: World Health Organization.

How to cite this article

Currie, Janet. "nutrition and public policy in advanced economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000165> doi:10.1057/9780230226203.1212

The New Palgrave Dictionary of Economics Online

offer curve or reciprocal demand curve

Harvey Gram
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adjustment process; community indifference curves; continuity; convexity; duality; excess demand; excess supply; external economies; homogeneous programming; income effects; increasing returns; Marshall, A.; Mill. J. S.; monotonicity; offer curve; reciprocal demand curve; recontracting; stability of equilibrium; substitution effects; tatonnement; trading curve; uniqueness of equilibrium; Walras's Law

Article

The offer curve made its first appearance in Alfred Marshall's *Pure Theory of Foreign Trade* (1879), a privately printed paper consisting of the second and third chapters (chosen by Henry Sidgwick) of a four chapter manuscript. Almost 50 years passed before Marshall's analysis became generally available under his own name as Appendix J to *Money, Credit and Commerce* (1923). Thus, it was mainly through the writings of Edgeworth (1894) and others who had read Marshall's original contribution (see Whitaker, 1975, p. 114n), that the offer curve came to be known.

Newman (1965, p. 104) notes the objections raised by Edgeworth (1924) and Wicksell (1925) to the name *offer curve* which was coined by W.E. Johnson (1913) and used by Bowley (1924). They were concerned that *offer curve* might suggest an asymmetry between supply and demand where, in fact, there was none. The alternative name, *reciprocal demand curve*, or *trading curve* (Newman, 1965, pp. 89 ff.), avoids any such suggestion.

Marshall commented that his ‘International Trade curves … were set to a definite tune, that called by [John Stuart] Mill’ (Pigou, 1925, p. 451). It was Mill who had written that

> supply and demand are but another expression for reciprocal demand; and to say that value will adjust itself so as to equalize demand with supply, is in fact to say that it will adjust itself so as to equalize demand on one side with the demand of the other. (Mill, 1852, p. 604)

Mill's purpose was to close Ricardo's trade model by finding prices such that ‘demand will be exactly sufficient to carry off the supply’ (Mill 1844, p. 238). Edgeworth, though giving high praise to Mill's mature statement of his equation of international demand, thought little of Mill's exact solution; and Marshall commented only ‘that the special example which [Mill] has chosen does not illustrate the general problem in question’ (Whitaker, 1975, p. 148). Chipman has argued, however, that Mill, in effect, solved a problem in what would now be called *homogeneous programming*. In claiming Mill's result to be a ‘genuine and correct proof of the existence of equilibrium … [pre-dating the next such proof] by eighty years’, Chipman remarks of Mill's law of international value, or *reciprocal demand*, that in ‘its astonishing simplicity, it must stand as one of the great achievements of the human intellect’ (Chipman, 1965, Part 1, pp. 491 and 486, respectively).

Modern uses of the offer curve in trade theory and other areas (see Cass, Okuno and Zilcha, 1980; Cass, 1980; Grandmont, 1985) have a greater affinity with Mill's analysis of a general equilibrium of supply and demand than with Marshall's original argument. That argument had three parts. The first is directly relevant to modern theory and concerns what would now be called the *income and substitution effects* of relative price changes. The second part deals with *increasing returns in production*, a phenomenon whose formulation and implications for traditional theory remain controversial. And finally, there is the problem of the *adjustment mechanism*, a part of Marshall's theory which, though highly regarded (Whitaker, 1975, p. 115; Kemp, 1964, p. 60), has been almost completely eclipsed by a Walrasian inspired ‘stability’ analysis. What follows is accordingly divided into three sections,

following a brief discussion of the formal basis for the offer curve as it exists in modern theory.

General equilibrium

The traditional offer curve arises in the context of a two-country general equilibrium model. Each country has an endowment of resources and a technology for transforming the associated factor service flows into flows of output of two tradable commodities. Resources are owned by the country's consumers, each of whom has a preference ordering which is continuous, convex, and monotonic. Under constant or decreasing returns to scale, resources and technology generate a *convex* production possibilities set (although some degree of increasing returns to scale is not inconsistent with convexity). The assumptions on preferences guarantee that the set of points ranked ‘at least as good as’ any given point is also *convex*, its boundary defining an indifference curve. If consumers have identical preferences and factor endowments, community indifference curves are simply radial expansions of individual indifference curves (cf. Chipman, 1965, part 2, pp. 690–8).

Geometrical derivations of the offer curve utilize techniques introduced by Leontief (1933), Lerner (1932; 1934), and Meade (1952). Implicit in the derivation is the solution to a pair of problems in constrained optimization. At given commodity prices, P_j, j=1, 2, outputs, Y_j^k, in each country k, k=A, B, are such that the value of production, $\Sigma_j P_j Y_j^k$ is a maximum, subject to resource constraints which define the production possibilities sets. Simultaneously, for given factor supplies, F_i^k, i=1, 2 (assuming two factors for simplicity), rental rates or factor prices, W_i^k, are such that cost of production, $\Sigma_i W_i^k K_i^k$, is a *minimum*, subject to price constraints which state that equilibrium profits are nowhere positive. Duality theory establishes an equality between maximum value and minimum cost. Because consumers own all resources, total cost is equal to total income, and so consumption choices, X_j^k satisfy Walras's Law: $\Sigma_j P_j X_j^k = \Sigma_j P_j Y_j^k$.

Given P_1 and P_2, there may or may not exist a solution or set of solutions, $(Y_1^k - X_1^k, Y_2^k - X_2^k)$, k=A, B. If $Y_1^k - X_1^k$ is positive (negative), Walras's Law ensures that $Y_2^k - X_2^k$ is negative (positive): country k offers an excess supply of good 1 (good 2) in order to satisfy its excess demand for good 2 (good 1). If both prices are positive, $Y_1^k - X_1^k$=0 implies $Y_2^k - X_2^k$=0; while if one price is zero and satiation is ruled out, the corresponding excess demand will be unbounded. In Figure 1, the offer curves therefore occupy quadrants II and IV, passing through the origin and approaching the axes asymptotically.

Figure 1

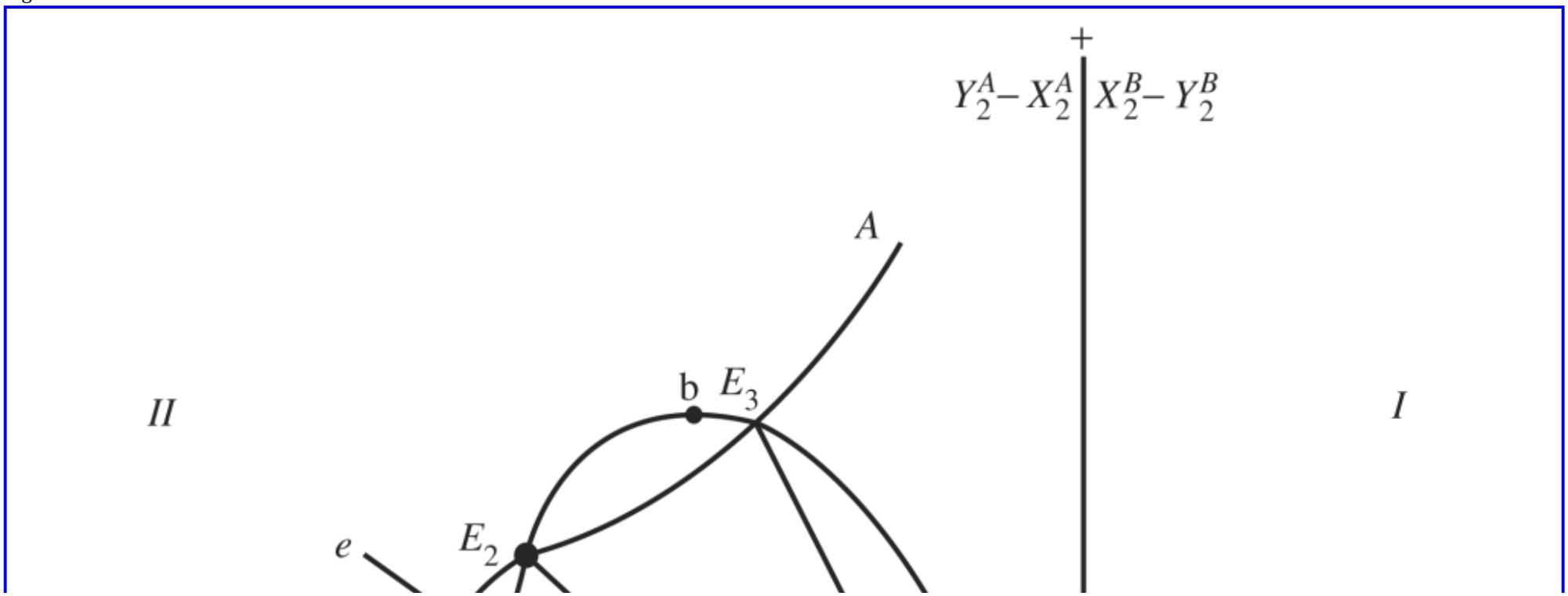

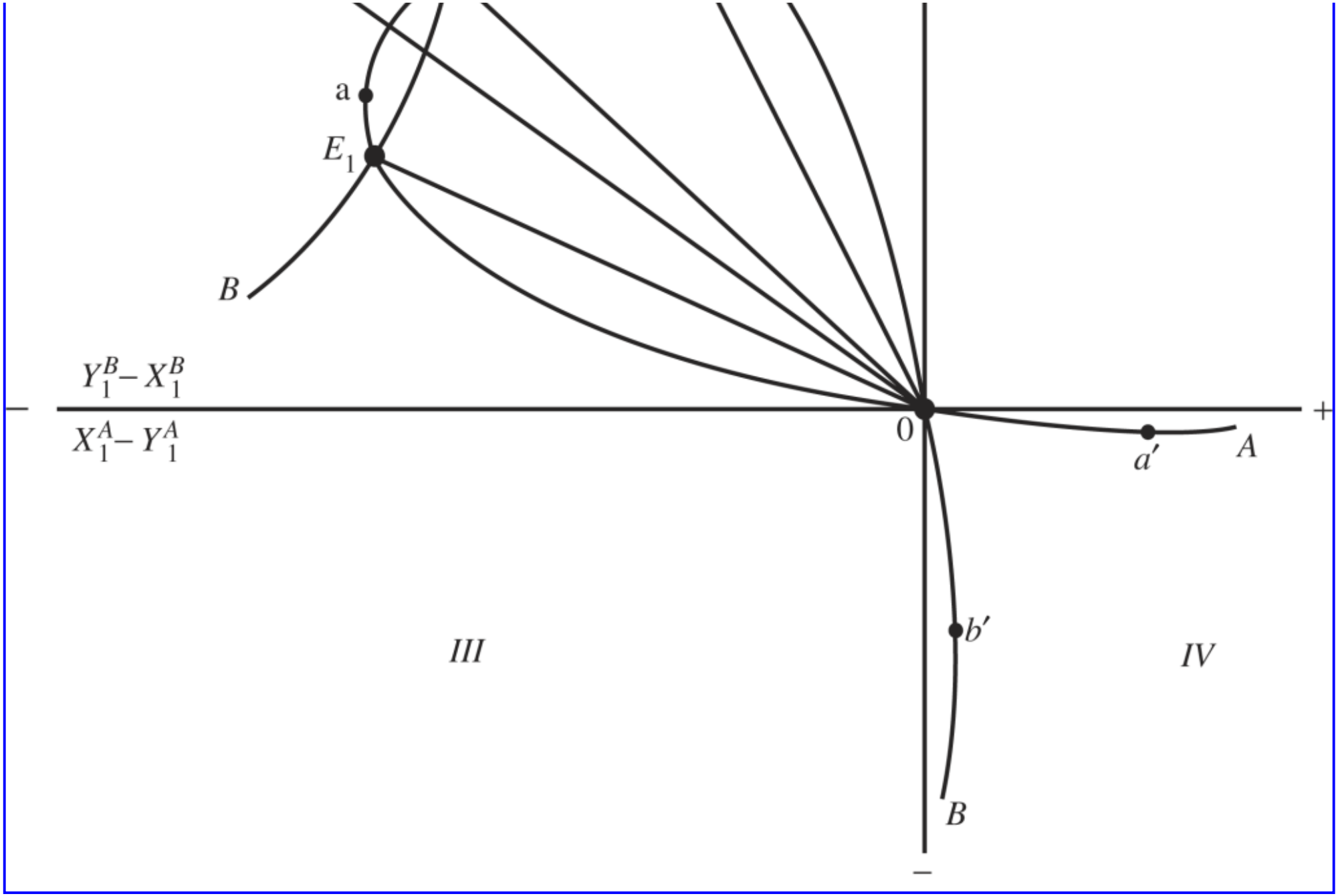

At a given price ratio, measured by the (absolute) slope of a straight line through the origin, the solution for excess supply and excess demand in each country is unique in Figure 1 and therefore (trivially) convex-valued. The solutions are also upper semicontinuous (Chipman, 1965, part 2, p. 717). These two conditions are the basis for the idea of ‘connectedness’ or ‘continuity’ of the offer curve. They follow from the postulates on preferences: continuity, convexity, and monotonicity (where the last can be replaced by the assumption that outputs are strictly positive). The importance of the postulates turns on the fact that when the set of offers by each country, at a given price vector, is closed, convex, and upper semicontinuous, it can be shown that an equilibrium price vector exists. Mill found a *unique* equilibrium for Ricardo's trade model by assuming (implicitly) unitary price and income elasticities of demand for the two commodities in each country (Chipman, 1965, part 1, pp. 483–91). The offer curves in Figure 1 intersect three times, indicating *three*

isolated equilibrium price vectors, OE_1, OE_2, and OE_3. Perpendiculars from E_1, E_2, and E_3 to the axes mark off the matching reciprocal demands of each country.

Income and substitution effects

The shape of each curve in Figure 1 reflects the income and substitution effects of relative price changes. Consider country *A*. When P_1/P_2 is zero output of good 1 is zero and demand is unbounded so that the offer curve shoots off to the right in quadrant IV. As P_1/P_2 increases, convexity of the production possibilities set ensures that Y_1^A increases and Y_2^A decreases (unless both remain constant at a vertex). Assuming hypothetically that country *A* is confined to a given, convex-to-the-origin community indifference curve, X_1^A decreases and X_2^A increases (unless both remain constant at a 'corner' of the curve). These two substitution effects, one in production and one in consumption, reduce excess demand for good 1 (and excess supply of good 2) in quadrant IV, while raising excess supply of good 1 (and excess demand for good 2) in quadrant II. Note, however, that excess supply of good 1 reaches a maximum at *a* in quadrant II, while the same is true for good 2 at a' in quadrant IV. The reason for this is the income effect of the relative price change. In quadrant IV, a higher P_1/P_2 reduces the real purchasing power of country *A* which is an importer of good 1. If both goods are normal, the reduction in X_1^A associated with substitution is reinforced while the increase in X_2^A is offset. In quadrant II, the reverse is true. Country *A*, as an exporter of good 1, gains from an increase in the relative price of good 1. Now the income effect is pushing against the substitution effect in determining X_1^A and reinforcing it in determining X_2^A. Along the offer curve between *a* and a' substitution effects in production and consumption dominate the income effect. Beyond those critical points, the income effect is dominant in the sense that the excess supply of a commodity is lower when its relative price is higher.

Marshall's explanation of the critical point *a* is somewhat different. The independent variable in his analysis is the quantity of imports rather than the relative price ratio. In Marshall's normal class, an increase in imports in the neighbourhood of the origin results in an increase in receipts and, for this reason, the volume of exports which can be produced at normal profits increases. Receipts from imports pay the cost of exports. The slope of the offer curve increases (in absolute value) from the origin to point *a* because demand for imports is elastic. Beyond point *a* import demand turns inelastic, receipts fall off, and so the volume of exports which can be produced at normal profits declines. Marshall referred to this situation as Class I.

A final aspect of the income effect of relative price changes concerns changes in the distribution of purchasing power among consumers within each country, a problem which can only be addressed if consumers have different resource endowments. Assume therefore that country *A* and country *B* in Figure 1 are, in fact, two groups of consumers within a single country. Aggregate excess demand for good 1 and excess supply of good 2 are *positive* for price vectors flatter than OE_1 and for vectors intermediate between OE_2 and OE_3, and *negative* for vectors intermediate between OE_1 and OE_2 and steeper than OE_3. An offer curve defined for the two groups of consumers would therefore pass through the origin three times, tying itself in a bow. Its slope at the origin has three isolated values given by the slopes of OE_1, OE_2, and OE_3 (cf. Johnson, 1959; 1960). A pair of such curves, constructed for two countries, each composed of two differentiated groups of consumers, would intersect at various points in quadrants II and IV. This indicates the possibility of trade pattern reversals as the relative price ratio takes on different equilibrium values.

The one proposition that Marshall insisted upon as 'the only law to which the curves must conform under all circumstances' is violated by offer curves which form loops through the origin. This was his Proposition VI to the effect that country *A*'s offer curve 'cannot in any case be cut more than once by a horizontal line. Similarly [country *B*'s curve] cannot in any case be cut more than once by a vertical line' (Whitaker, 1975, p. 140). Marshall's argument, however, had nothing to do with the income redistribution effects which, upon aggregation of consumer groups (as above), countries (see Chipman, 1965, part 2, p. 217), or generations (see Cass, Okuno and Zilcha, 1980) pp. 25–6), can result in offer curves exhibiting the floral patterns first noted by Johnson (1959; 1960). Rather he was concerned with the problem of increasing returns.

Increasing returns

Marshall put increasing returns under the heading of 'problems of Exceptional Class II' (Whitaker, 1975, p. 144). Where an increase in the production of exports leads to the introduction of extensive economies, a reduction in the volume of exports to a level previously experienced would not require as large a volume of imports to cover their costs of production as had previously been the case (assuming implicitly an elastic demand for imports). Thus, a movement along the offer curve would simultaneously shift the curve *towards* the export axis. Moreover, 'if time was allowed for the development of economies of production on a large scale, time ought to be allowed for the general increase of demand' (Pigou, 1925, p. 49). In that event, a given volume of imports would yield higher receipts thereby shifting the offer curve *away* from the import axis.

Marshall's long period offer curve does not show any maximum level of exports, as does every static curve constructed on the basis of given resources and technology. Moreover, the

slope of the long period curve can *decrease* (in absolute value) as a consequence of technological change. It was in this context that Marshall denied that any given volume of imports would cover the costs of more than one volume of exports. His Proposition VI claimed that economies of scale would never be sufficient to lower the total cost of a larger volume of exports below that of a smaller volume previously produced. Marshall had made the same assumption in the first edition of his *Principles*, but dropped it subsequently (Whitaker, 1975, p. 116).

Modern discussions of the offer curve in the presence of increasing returns are concerned with technological externalities rather than with irreversible economies of large-scale production. A firm's output may depend on the output of the industry to which it belongs. Output in one industry, or the level of employment of particular factors in that industry, may have external effects on the output of another industry. Theoretical questions then arise concerning the convexity of the production possibilities set, the relationship between opportunity cost and relative price, and whether or not production occurs at a limit point of the feasible set. What the models have is common with Marshall's discussion is that the associated offer curves are no longer convex-valued functions of the relative price ratio. Marshall indicated this by drawing offer curves with several inflexion points. In modern treatments of external economies in production, offer curves typically have the shape indicated in Figure 2. Curvature at the origin is opposite to that indicated in Figure 1, changing abruptly at points of complete specialization. (The latter may or may not correspond to the critical points in Figure 1 where excess supply reaches a maximum.)

Figure 2

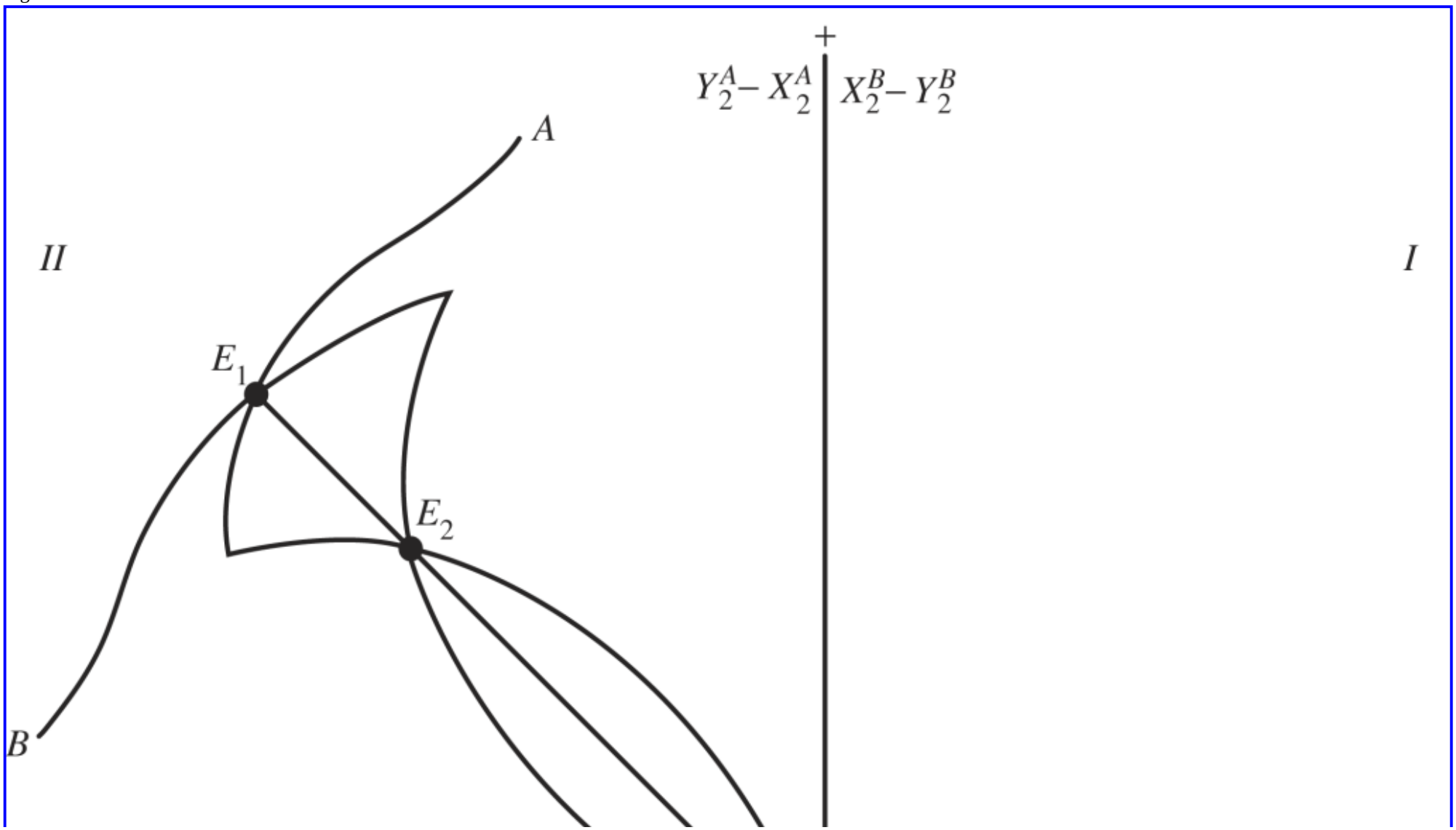

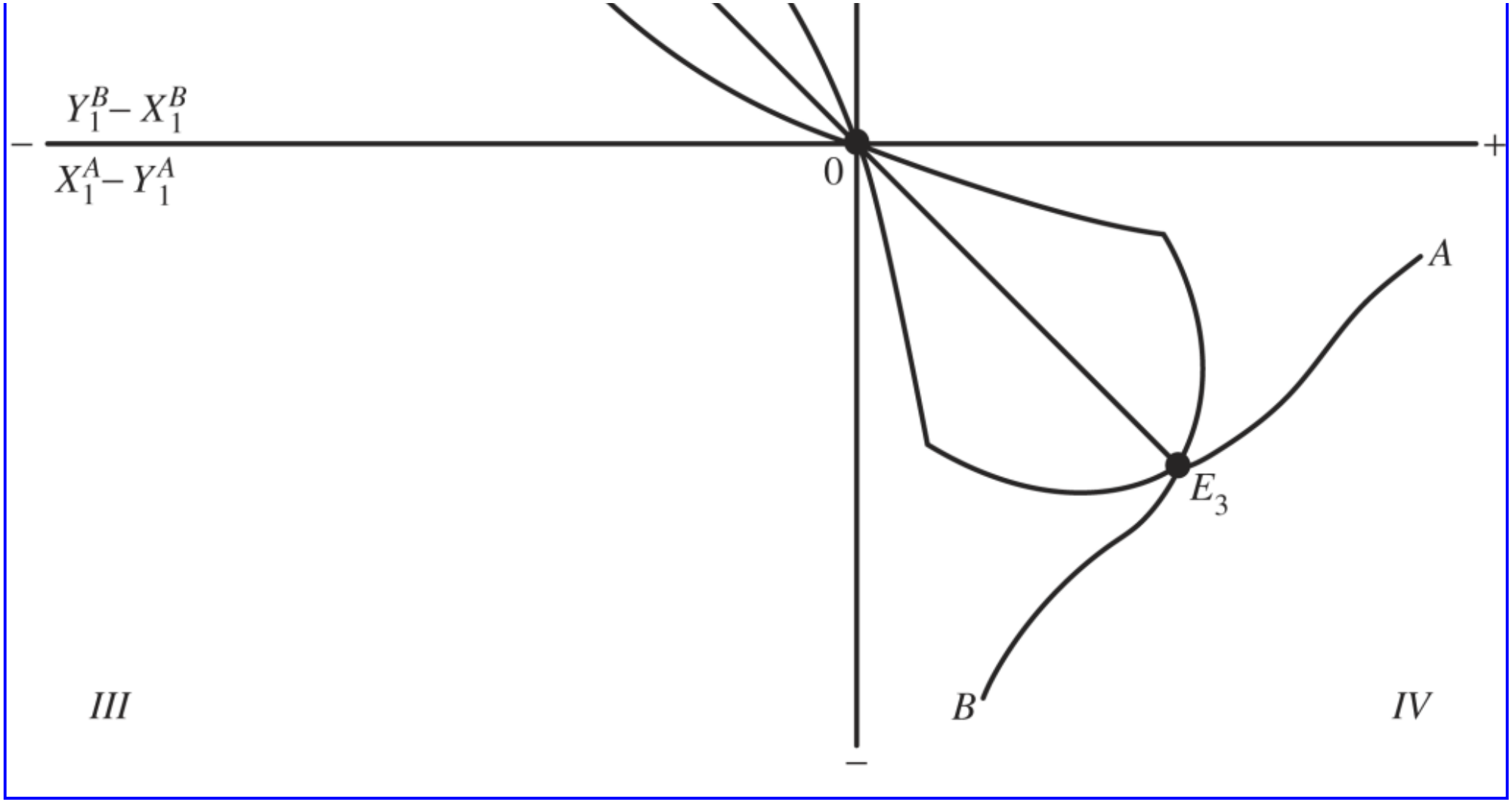

The curves in Figure 2 have three intersections indicating three equilibrium trades. (These may be reduced to two by drawing curves which are mutually tangent at the origin indicating equal pre-trade price ratios which, in Figure 1, would be sufficient to rule out an equilibrium with positive trade.) There is nothing in principle to prevent all three points from falling along a single ray. The resulting indeterminacy in the volume and direction of trade is the main distinguishing feature of trade models with external economies in production. Chacholiades (1978, pp. 197–9) has considered the problem in some detail, arguing that, in general, a country benefits from specialization in the production of the commodity subject to external economies. If this is the same commodity in each country, then, depending on the pattern of demand, one country may lose from trade. This suggests an even sharper conflict of interest than is evident in Figure 1 where a country is clearly better off in that equilibrium in which its exports are smallest and its imports are largest.

Stability

Stability of equilibrium is defined in relationship to a process of adjustment which determines the movement of prices and/or quantities when the system is out of equilibrium. A distinction has been drawn between processes which focus on price changes and those which focus on quantity changes. A frequently considered case is that in which prices respond to differences between *hypothetical* supply and demand (those quantities which would prevail on each side of the market if the current price were an equilibrium price). Transactions, however, only take place in equilibrium. This is a *recontracting process* and its convergence to an equilibrium of supply and demand is often referred to as the Walrasian *tâtonnement* or ‘groping’ process. Walras, in fact, referred to *tâtonnement* in connection with the problem of bringing a set of interrelated markets into equilibrium *sequentially*, and, as such, it

was problematical since ‘few prices will lie quiet at equilibrium while others are brought to heel, and the whole thing may turn out to be like the labour of Sisyphus’ (Newman, 1965, p. 103).

The offer curves in Figure 1 can be used to illustrate the stability of a recontracting process. Consider a price ratio, P_1/P_2, slightly greater than the (absolute) slope of OE_1. Hypothetical exports of good 1 by country *A* exceed hypothetical imports by country 2, while the opposite is true for good 2. If P_1 falls and P_2 rises in this situation, P_1/P_2 falls back towards OE_1. The opposite is true for price ratios slightly lower than the (absolute) slope of OE_1. Thus, E_1 is a stable point. Note that, as prices move, the four substitution effects (in production and consumption in both countries) contribute towards reducing the initial divergence between supply and demand for each good. The same is true of part of the income effect. As the price ratio falls towards OE_1, for example, country *A* is made worse off and country *B* is made better off. As *importers*, country *A* demands less of good 2 and country *B* demands more of good 1 (assuming that the goods are normal), and this reinforces the substitution effects. But, as *exporters*, country *A* demands less of good 1 while country *B* demands more of good 2, thereby exacerbating the initial excess supply (of good 1) and excess demand (for good 2). At stable points, such as E_1 and E_3, exporters’ income effects are not strong enough to swamp importer's income effects plus all substitution effects. At E_2, however, a slight increase in the relative price of good 1 would be associated with a hypothetical excess demand for good 1 and excess supply of good 2. The recontracting hypothesis would therefore result in a further increase in P_1/P_2, reflecting the fact that the initial increase has generated exporters’ income effects which swamp all other income and substitution effects (cf. Caves and Jones, 1985, pp. 492–4).

A variation on the above analysis allows trade to take place out of equilibrium but assumes that demand for imports is always satisfied. A disequilibrium exchange ray, such as *Oe* in Figure 1, cuts the two offer curves in distinct points. Perpendiculars to the axes from these points indicate excess supply of good 1, which causes inventories to rise, and excess demand for good 2, which causes inventories to fall. If P_1 then falls and P_2 rises, P_1/P_2 once again falls back toward OE_1. During the process, however, country *A* must be selling assets to country *B* in order for the trade flow to be financed. If the consequence of this is to alter the position and shape of the offer curves, a more complete and undoubtedly more complex analysis of the convergence to equilibrium would be required (cf. Jones, 1961, p. 203). Marshall's discussion of the adjustment mechanism is concerned neither with a recontracting process nor with inconsistent trades ‘financed’ by changes in inventories. In this theory, profits in export industries are abnormally high at points between a country's offer curve and the axis measuring its imports. On the other side of the curve, profits in exports are abnormally low. Marshall's adjustment mechanism is summed up as follows:

> when the terms on which a country's foreign trade is conducted are such as to afford a rate of profits higher than the rate current in other industries, the competition of traders to obtain these higher profits will lead to an increase in the exportation of her wares: and *vice versa* when the rate of profits in the foreign trade [is] exceptionally low. (Whitaker, 1975, p. 151)

This adjustment in the production of exports (imports and the domestic consumption of exports held constant) appears to have been meant by Marshall to reflect a concomitant change in the production of non-traded goods (Marshall, 1923, pp. 354–5n). Thus, at points off the offer curves

> production is changing in both countries, [and so] the dimensions of the Edgeworth box must be changing, as are also the shapes of the offer curves. The extreme subtlety of the Marshallian conception becomes more apparent the further one probes into it. (Chipman, 1965, Part 2, p. 723)

One can only conclude that efforts to formalize Marshall's ‘dynamics’ (Samuelson, 1947, pp. 266–8; Kemp, 1964, pp. 66–9; Amano, 1968, pp. 326–39) are but valiant attempts to come to terms with an approach to equilibrium which itself moves in an unspecified manner as a consequence of not being attained initially.

Conclusion

Not surprisingly, that part of Marshall's *Pure Theory of Foreign Trade* which is most evident in modern discussions of the offer curve concerns the income and substitution effects which are central to the theory of supply and demand equilibrium. His treatment of increasing returns and his discussion of the adjustment process raise dynamic considerations associated with changes in technology and with changes in the structure of productive capacity. It is just such changes which present the equilibrium theory of supply and demand with some of its greatest difficulties.

See Also

- Marshall, Alfred

- terms of trade

Bibliography

Amano, A. 1968. Stability conditions in the pure theory of international trade – rehabilitation of the Marshallian approach. *Quarterly Journal of Economics* 82, 326–39.

Bowley, A. 1924. *The Mathematical Groundwork of Economics*. New York: Oxford University Press.

Cass, D. 1980. Money in consumption loan type models: an addendum. In *Models of Monetary Economics*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Cass, D., Okuno, M. and Zilcha, I. 1980. The role of money in supporting the Pareto optimality of competitive equilibrium in consumption loan type models. In *Models of Monetary Economies*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis, 13–48.

Caves, R.E. and Jones, R.W. 1985. *World Trade and Payments, An Introduction*. 4th edn. Boston: Little, Brown.

Chacholiades, M. 1978. *International Trade Theory and Policy*. New York: McGraw-Hill.

Chipman, J.S. 1965. A survey of the theory of international trade. *Econometrica* 33, Pt. I, 477–519; Pt. II, 685–760.

Edgeworth, F.Y. 1894. The theory of international values. *Economic Journal* 4, 35–50, 424–43, 606–38. Reprinted in F.Y. Edgeworth, *Papers Relating to Political Economy* [1925], vol. 2, New York: Burt Franklin, 1970.

Edgeworth, F.Y. 1924. Review. *Economic Journal* 34, 430.

Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53, 995–1045.

Johnson, H.G. 1959. International trade, income distribution, and the offer curve. *Manchester School of Economic and Social Studies* 27, 241–60.

Johnson, H.G. 1960. Income distribution, the offer curve, and the effects of tariffs. *Manchester School of Economic and Social Studies* 8, 215–42.

Johnson, W.E. 1913. The pure theory of utility curves. *Economic Journal* 23, 483–513.

Jones, R.W. 1961. Stability conditions in international trade: a general equilibrium analysis. *International Economic Review* 2, 199–209.

Kemp, M.C. 1964. *The Pure Theory of International Trade*. Englewood Cliffs, NJ: Prentice-Hall.

Leontief, W.W. 1933. The use of indifference curves in the analysis of foreign trade. *Quarterly Journal of Economics* 47, 493–503.

Lerner, A.P. 1932. The diagrammatical representation of cost conditions in international trade. *Economica* 12, 346–56.

Lerner, A.P. 1934. The diagrammatical representation of demand conditions in international trade. *Economica*, NS 1, 319–34.

Marshall, A. 1879. *The Pure Theory of Foreign Trade*. Reprinted in Whitaker (1975).

Marshall, A. 1923. *Money, Credit and Commerce*. New York: Augustus Kelley, 1965.

Meade, J.E. 1952. *A Geometry of International Trade*. Reprinted, New York: Augustus Kelley, 1971.

Mill, J.S. 1844. On the laws of interchange between nations; and the distribution of the gains of commerce among countries of the commercial world. In *Essays on Some Unsettled Questions of Political Economy*. Reprinted in *Collected Works of John Stuart Mill*, vol. 4, Toronto: University of Toronto Press, 1967.

Mill, J.S. 1852. *Principles of Political Economy*. 3rd edn. Reprinted in *Collected Works of John Stuart Mill*, vol. 3, Toronto: University of Toronto Press, 1965.

Newman, P. 1965. *The Theory of Exchange*. Englewood Cliffs, NJ: Prentice-Hall.

Pigou, A.C., ed. 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Whitaker, J.K., ed. 1975. *The Early Writings of Alfred Marshall, 1867–1890*. vol. 2, New York: Free Press.

Wicksell, K. 1925. Matematisk nationalekonomi. *Ekonomisk Tidskrift* 27, 103–25. Trans. in *Knut Wicksell, Selected Papers on Economic Theory*, ed. E. Lindahl. Reprinted, New York: Augustus Kelley, 1969.

How to cite this article

Gram, Harvey. "offer curve or reciprocal demand curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000005> doi:10.1057/9780230226203.1213

The New Palgrave Dictionary of Economics Online

Ohlin, Bertil Gotthard (1899 – 1979)

Hans Brems
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

accelerator; adaptive expectations; bonds; budget deficits; Cassel, G.; factor endowments; factor price equalization; factor substitution; Heckscher–Ohlin trade theorem; Hecksher, F. H.; input–output coefficients; international capital flows; international trade theory; interregional trade; Keynesianism; Lindahl, E. R.; liquidity preference; money supply; multiplier; natural rate and market rate of interest; Ohlin, B. G.; preferences; propensity to consume; public works; Stockholm School; Stolper–Samuelson theorem; technology; Wicksell, J. G. K.

Article

Ohlin was born on 23 April 1899 in Klippan, Sweden. He took a degree in mathematics, statistics and economics at the University of Lund in 1917, a degree in economics under Heckscher at the Stockholm School of Business Administration in 1919, an AM degree under Taussig and Williams at Harvard in 1923, and a Ph.D. degree under Cassel at the University of Stockholm in 1924. Ohlin taught at the University of Copenhagen (1925–30) and, as Heckscher's successor, at the Stockholm School of Business Administration (1930–65). He was a visiting professor at the University of California at Berkeley in 1937 and at Columbia and Oxford in 1947.
For the League of Nations Ohlin prepared a report on the world depression in 1931 and for the Swedish government a report on unemployment in 1934. He was a member of the Swedish parliament (1938–70), a member of the Cabinet (1944–45), the leader of the Liberal Party (1944–67); he died on 3 August 1979 in Stockholm.

Trade theory

Ohlin is best known for, and received the 1977 Nobel Prize for, his modernization of the theory of international trade. The modernization was long overdue: discredited in general economic theory after 1870, the labour theory of value was still surviving in the province of international-trade theory half a

century later.

Ohlin's teacher at Stockholm was Gustav Cassel, and his point of departure was Cassel's (1918) version of a Walrasian general equilibrium of a closed economy with perfect mobility of goods and factors. Unlike Walras, Cassel assumed the factor endowments of all households to be fixed. Household income would then be the sum of the products of factor price and all factor endowments of that household. Like Walras, Cassel assumed the input–output coefficients of all goods to be fixed. The competitive price of a good would then be the sum of the products of factor price and all input–output coefficients of that good. Facing such household income and such competitive goods prices, every household would reveal its preference. Goods–market equilibrium would require industry supply and such household demand to be equal for every good. Industry demand for a factor would be the sum of the products of such industry goods supplies and all input–output coefficients of that factor. Factor-market equilibrium would require household supply and such industry demand to be equal for every factor.

The ultimate determinants of all quantities and relative prices in such a general equilibrium were, first, factor endowments; second, technology in the form of the input–output coefficients; and, third, preferences. Inspired by his other teacher at Stockholm, Eli Filip Heckscher (1919), Ohlin (1924; 1933) set out to modify the Cassel model to fit interregional and international trade.

As his first modification Ohlin visualized an economy composed of regions within which factor mobility was perfect but between which it was imperfect or, as a first approximation, non-existent. In the absence of goods trade, isolation would be complete, and such regions would simply constitute a system of miniature Casselian closed economies. Between them relative prices could differ because factor endowments, technology, or preferences differed. As another first approximation, Ohlin assumed regions to differ solely in their factor endowments, not in their technology or preferences. Finally, Ohlin unfroze Cassel's fixed input–output coefficients, thus making room for factor substitution. With such assumptions he had the ingredients to what later became known as the ‘strong’ Heckscher–Ohlin theorem. In the simple case of two factors, two goods and two regions the theorem becomes very tractable. In isolation each region would have a relatively low-priced and a relatively high-priced good. Since nothing else than factor endowments differed between regions, the low-priced good would be low-priced because it required relatively much of that region's relatively abundant, hence low-priced, factor. That good will be a candidate for export once we remove isolation. The high-priced good would be high-priced because it required relatively much of that region's relatively scarce, hence high-priced, factor. That good will be a candidate for import once we remove isolation; but we are not removing it yet. As we know, under profit maximization, pure competition, and factor substitution the physical marginal productivity of either factor in terms of either good will equal the real price of that factor in terms of that good.

Now remove isolation and let goods be traded. Export would expand a region's demand for its abundant factor and import reduce the demand for its scarce factor. Thus trade would raise the price of the abundant factor, reduce the price of the scarce one, and encourage substitution between them: either good would use less abundant factor per unit of scarce factor than in isolation. The abundant factor would then have a higher physical marginal productivity and a higher real price in terms of either good than in isolation. Vice versa for the scarce factor. Does all this mean that trade would eventually equalize real factor prices in terms of either good between regions – although no factor ever crossed the border? Yes, in the absence of transportation costs and in the absence of specialization. One reason for specialization would be increasing returns to scale. Specialization would leave a region with an

unproduced good. Where nothing is produced, no factor can have a marginal productivity. In terms of the unproduced good, then, physical marginal productivity could no longer equal real factor price, and the theorem would fail. So it would in case of transportation costs or in case regions differed, not in factor endowments but in technology or preferences. And so it might if there were more than two factors, goods, or regions.

Few theorems have been as fruitful, that is, few inspired as much later work, theoretical and empirical, as the Heckscher–Ohlin theorem. Neither Heckscher nor Ohlin applied present-day rigour. To Heckscher factor-price equalization would be complete; to Ohlin – more aware of the many qualifications – incomplete. The theorem was first taken up, baptized, and rigourized by Stolper and Samuelson (1941), who examined a scarce factor's case for protectionism but found ‘the definiteness of the Heckscher–Ohlin theorem [beginning] to fade’ with more than two factors. More groundwork was done by Samuelson (1948; 1949). Using his domestic US input–output table with many goods but only two factors, Leontief (1953; 1956) found the capital–labour ratio to be lower in US exports than in US import-competing goods. If the Heckscher–Ohlin theorem were true, then, capital would have to be the scarce and labour the abundant US factor. This Leontief paradox did not make the theorem go away but stimulated new contributions. A good guide to them is the third part of Chipman's (1966) survey of the theory of international trade.

Ohlin's second modification of Cassel saw international trade as a special case of interregional trade. What was special about nations?

First, national differences in factor endowments, technology, and preferences might be rooted in differences in climate, language, cultural, and legal institutions. Of international movements of factors, labour as well as capital, and such obstacles to them Ohlin gave a full account. His account of international capital movements found an early and specific expression (1929) in his discussion with Keynes of the mechanism of the reparation payments imposed upon Germany by the Versailles Treaty. Still influenced by Marshallian tradition, Keynes saw a drastic worsening of Germany's terms of trade as a necessary condition for such payments. To Ohlin reparations were nothing but huge international transfers of ‘buying power’. Against an uncomprehending 1929 Keynes, Ohlin advocated the view of a 1936 Keynes, that is, the income mechanism would do; no price mechanism was needed.

Second, nations were special in having their own currency and monetary authorities. In a two-country world such separate currencies would add a new unknown, that is, the price of one currency in terms of the other – the exchange rate. Fortunately there would also be a new equation, that is, the equilibrium condition that in a pure-trade model the balance of trade would be zero or that in a trade *cum* lending and borrowing model the balance of payments would be zero.

Macroeconomic theory

Less well known to the English-speaking world is Ohlin's macroeconomic theory: its most important work (1934) was never fully translated. Here, Ohlin was inspired by Wicksell and Lindahl.

Wicksell (1893) had restated Böhm-Bawerk mathematically and (1898) wondered how a Böhm-Bawerk ‘natural’ rate of interest was related to the rate of interest observed in markets where the supply of money met the demand for it. If such a ‘money’ rate of interest were lower than the natural rate of interest, entrepreneurs would be induced – and the money supply correspondingly expanded – to pay a

higher money wage rate. Physically speaking, nothing would come of this, for when labour spent the higher money wage rate, prices would rise correspondingly and unexpectedly leave the real wage rate unchanged. There would be a cumulative process of inflation expected by nobody.

Wicksell's answer was made possible by a method fundamentally new in three respects. Wicksell's method was a macroeconomic, dynamic disequilibrium method based upon adaptive expectations whose disappointment constituted the motive force of the system. But Wicksell had applied his method to a model with price as the only variable. Using Wicksell's method and inspired by Lindahl's (1930) refinement of it, Ohlin (1933; 1934) added physical output as an additional variable. Two years ahead of Keynes, Ohlin used three Keynesian tools, that is, the propensity to consume, liquidity preference and the multiplier, and one non-Keynesian tool, that is, the accelerator. The four tools would interact as follows in Ohlin's feedback mechanism. Let consumption demand be stimulated. As a result physical output would rise, generating new income. The propensity to consume would link physical consumption to the *level* of physical output and thus establish a consumption feedback. The accelerator would link physical investment to the *growth* of physical output and thus establish an investment feedback. As did the Wicksellian one, Ohlin's two feedbacks unfolded in a cumulative process along a time axis as a succession of disequilibria: expectations and plans were for ever being revised in the light of new experience. By contrast, Keynes used only the consumption feedback and telescoped it into an instant static equilibrium along an output axis.

Ohlin's relation to Keynesian economics was discussed by Steiger (1976), Patinkin (1978), and Brems (1978). Forty-one years apart Ohlin expressed his own view on the matter in (1937) and (1978).

Ohlin's (1934) analysis appeared in a report on unemployment requested by the Swedish government, and his policy conclusions were quite specific. In times of excess capacity the government should undertake investment projects – say highway construction or the electrification of state railroads – which would not compete with private investment and which should be allowed to generate fiscal deficits. Tax financing would reduce consumption and thus defeat the purpose of public works. Ohlin wrote the government budget constraint: deficits might be financed by expanding either the bond or the money supply. Sale of government bonds would depress bond prices and thus discourage private investment, again defeating the purpose of public works. That left central-bank discounting of treasury bills as the only way which would not deprive private investment of finance. Thus financed, public works would generate income. Such income generation would be magnified by the multiplier and the accelerator.

Except for a nine-page algebraic two-country Cassel general equilibrium, banished to an appendix, Ohlin used neither algebra nor diagrams. But in all his work his style was accurate, cautious and lucid, often enlivened by relevant statistical and historical illustrations.

See Also

- Heckscher–Ohlin trade theory
- Stockholm School

Selected works

1924. *Handelns teori*. Stockholm: Centraltryckeriet.

1929. Transfer difficulties, real and imagined. *Economic Journal* 37(March), 172–8. Reprinted in *Readings in the Theory of International Trade*, ed. H.S. Ellis and L. Metzler. Philadelphia and Toronto: Blakiston, 1949.

1933. *Interregional and International Trade*. Cambridge, MA: Harvard University Press.

1933. Till frågan om penningteoriens uppläggning. *Ekonomisk Tidskrift*. Trans. as ‘On the Formulation of Monetary Theory’, *History of Political Economy* 10 (1978), 353–88.

1934. *Penningpolitik, offentliga arbeten, subventioner och tullar som model mot arbetslöshetbidrag till expansionens teori*. Unemployment Report II, 4. Stockholm: P.A. Norstedt & Söner. Summarized by Brems (1978).

1937. Some notes on the Stockholm theory of saving and investment. I–II. *Economic Journal* 47 (March), 53–69; June, 53–69, 221–40. Reprinted in *Readings in Business Cycle Theory*, ed. G. Haberler. Philadelphia and Toronto: Blakiston, 1951.

1978. Keynesian economics and the Stockholm School: a comment on Don Patinkin's paper. *Scandinavian Journal of Economics* 80(2), 144–7.

Bibliography

Brems, H. 1978. What was new in Ohlin's 1933–34 macroeconomics? *History of Political Economy* 10, 398–412.

Cassel, G. 1918. *Theoretische Sozialökonomie*. Leipzig: Winter. Trans. from the 5th edn as *The Theory of Social Economy*, New York: Harcourt, Brace, 1932.

Chipman, J.S. 1966. A survey of the theory of international trade: Part 3, the modern theory. *Econometrica* 34, 18–76.

Heckscher, E.F. 1919. Utrikeshandelns verkan på inkomstfordelningen. *Ekonomisk Tidskrift*, 497–512. Trans. as ‘The effect of foreign trade on the distribution of income’ in *Readings in the Theory of International Trade*, ed. H.S. Ellis and L. Metzler Philadelphia and Toronto: Blakiston, 1949.

Leontief, W.W. 1953. Domestic production and foreign trade; the American capital position re-examined. *Proceedings of the American Philosophical Society* 97, 332–49.

Leontief, W.W. 1956. Factor proportions and the structure of American trade: further theoretical and empirical analysis. *Review of Economics and Statistics* 38, 386–407.

Lindahl, E. 1930. *Penningpolitikens medel*. Lund: Gleerup. Partially trans. in E. Lindahl, *Studies in the Theory of Money and Capital*. London: Allen & Unwin, 1939.

Patinkin, D. 1978. On the relation between Keynesian economics and the ‘Stockholm School’. *Scandinavian Journal of Economics* 80(2), 135–43.

Samuelson, P.A. 1948. International trade and the equalization of factor prices. *Economic Journal* 58, 163–84.

Samuelson, P.A. 1949. International factor-price equalization once again. *Economic Journal* 59, 181–97.

Steiger, O. 1976. Bertil Ohlin and the origins of the Keynesian Revolution. *History of Political Economy* 8, 341–66.

Stolper, W.F. and Samuelson, P.A. 1941. Protection and real wages. *Review of Economic Studies* 9 (November), 58–73. Reprinted in *Readings in the Theory of International Trade*, ed. H.S. Ellis and L. Metzler. Philadelphia and Toronto: Blakiston, 1949.

Wicksell, K. 1893. *Über Wert, Kapital und Rente*. Jena: G. Fischer. Trans. as *Value, Capital and Rent*. London: Allen & Unwin, 1954.

Wicksell, K. 1898. *Geldzins und Güterpreise*. Jena: G. Fischer. Trans. R.F. Kahn with an introduction by Bertil Ohlin as *Interest and Prices*, London: Macmillan, 1936.

How to cite this article

Brems, Hans. "Ohlin, Bertil Gotthard (1899–1979)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000011> doi:10.1057/9780230226203.1214

The New Palgrave Dictionary of Economics Online

oil and the macroeconomy

James D. Hamilton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Oil price shocks are often followed by economic downturns. Both theory and evidence are consistent with the interpretation that this effect results from the disruption of the pattern of spending by consumers and firms on items whose utilization is sensitive to energy prices and supplies.

Keywords

oil and the macroeconomy; real business cycles; inflation

Article

Nine out of ten of the US recessions since the Second World War were preceded by an upward spike in oil prices. One way to inquire whether this might be just a coincidence is with a statistical regression of real GDP growth rates (quoted at a quarterly rate) on lagged changes in GDP growth rates and lagged logarithmic changes in nominal oil prices. The results from an ordinary least squares (OLS) estimation of this relation for t=1949:II to 1980:IV are as follows (standard errors in parentheses):

$$y_t = \underset{(0.18)}{1.14} + \underset{(0.09)}{0.20}\, y_{t-1} + \underset{(0.09)}{0.05}\, y_{t-2} - \underset{(0.09)}{0.10}\, y_{t-3} - \underset{(0.09)}{0.19}\, y_{t-4} - \underset{(0.026)}{0.004}\, o_{t-1} - \underset{(0.026)}{0.027}\, o_{t-2} - \underset{(0.026)}{0.034}\, o_{t-3} - \underset{(0.027)}{0.065}\, o_{t-4}.$$

The coefficient on the fourth lag of oil prices (o_{t-4}) is negative and highly statistically significant (t-statistic=–2.4), and an F-test leads to a rejection of the null hypothesis that the coefficients on lagged oil prices are all zero with a p-value of 0.005. Quite a few studies have tested and rejected the hypothesis that the relation between oil prices and output could just be a statistical coincidence, including Rasche and Tatom (1977; 1981), Hamilton (1983), Burbidge and Harrison (1984), Santini (1985; 1992), Gisser and Goodwin (1986), Rotemberg and Woodford (1996), Daniel (1997), Raymond and Rich (1997), Carruth, Hooker, and Oswald (1998), and Hamilton (2003).

Another possibility is that the correlation between oil prices and output results from common dependence on some third factor or factors that are the true cause of both the increase in oil prices and the subsequent recession. For example, something about the last stages of an economic expansion may often produce a surge in oil prices just before output is about to turn down, so that both the oil price increase and the subsequent recession result from the same business cycle dynamics. This is difficult to reconcile with the fact that, at least for the early post-war period, oil price changes could not be predicted from earlier movements in other macro variables (Hamilton, 1983), and that most of the oil spikes can be attributed to exogenous events such as military conflicts (Hamilton, 1985). However, Barsky and Kilian (2002; 2004) have recently developed challenges to

the latter claim.

Predicted size of effects

Economic theory suggests that it is the real oil price rather than the nominal price that should matter for economic decisions. It does not make much difference in summarizing the size of any given shock whether one uses the nominal price o_t or the real price of oil, since in most of the shocks discussed here the move in nominal prices is an order of magnitude larger than the change in overall prices during that quarter. However, particularly in the early part of the sample, the nominal oil price would stay frozen for years and then adjust suddenly. To the extent that there is a difference between using nominal and real prices as the explanatory variable in such regressions, the real price results from the confluence of two forces: events such as the Suez crisis, which accounts for almost all of the movement in the nominal price between 1955 and 1965, and the quarter-to-quarter change in inflation, which is completely endogenous with respect to the economy and whose consequences for future output are likely to be quite different from those of an oil shock. In so far as the statistical exogeneity of the right-hand variables is important for interpreting the regression, many researchers have for this reason used the nominal oil price change rather than the real oil price change as the explanatory variable.

One simple framework for thinking about what the effects of energy supply disruptions should be comes from examining a production function relating the output Y produced by a particular firm to its inputs of labor N, capital K, and energy E:

$$Y = F(N, K, E).$$

Suppose that output is sold for a nominal price P dollars per unit, labour is paid nominal wage W, energy's nominal price is Q, and capital is rented at nominal rate r. The profits of the firm are given by

$$PY - WN - rK - QE.$$

A price-taking profit-maximizing firm would purchase energy up to the point where the marginal product of energy is equal to its relative price,

$$F_E(N, K, E) = Q / P,$$

where $F_E(N, K, E)$ denotes the partial derivative of $F(\cdot)$ with respect to E. If we multiply both sides of the above equation by E and divide by Y, we find

$$\frac{\partial \ln F}{\partial \ln E} = \frac{QE}{PY}.$$

In other words, the elasticity of output with respect to a given change in energy use can be inferred from the dollar share of energy expenditures in total output.

This dollar share for the economy as a whole is fairly small. For example, in 2000 the United States consumed about 7.2 billion barrels of oil. At a price of $30 a barrel, that represents only 2.2 per cent of a $9.8 trillion nominal GDP. With the rapid

price increases of 2003–5, that share has risen to 3.8 per cent of GDP. Table 1 reports Hamilton's (2003) values for the size of the supply disruptions associated with the five most important oil shocks, calculated from the magnitude of the drop in production in the affected countries. Kilian (2005) has more modest estimates based on his inference that production might have fallen even in the absence of the indicated events, and neither Hamilton's nor Kilian's figures take into account the fact that typically production increased in other parts of the world to make up part of the gap. Even using the ten per cent figure and a four per cent crude oil share, however, such shocks would by the above calculation be predicted to reduce GDP by only 0.4 per cent. Table 1 also reports the amount by which US real GDP declined between the date of the oil shock and the trough of the subsequent recession, which trough usually was reached a little over a year after the oil shock. Since the US economy would grow 3.4 per cent during a typical year, these numbers imply declines of real GDP relative to trend in excess of four per cent, an order of magnitude greater than predicted by the factor share argument. Furthermore, Bohi (1991) failed to find statistically significant evidence that industries with greater energy factor shares suffered more than others in response to the oil shocks of the 1970s.

Exogenous disruptions in world petroleum supply, 1956–90

Date	Event	Drop as % of world production	Change in US real GDP (%)
Nov. 1956	Suez crisis	10.1	–2.5
Nov. 1973	Arab–Israel war	7.8	–3.2
Nov. 1978	Iranian revolution	8.9	–0.6
Oct. 1980	Iran–Iraq war	7.2	–0.5
Aug. 1990	Persian Gulf war	8.8	–0.1

Source: Hamilton (2003).

One would arrive at a similar prediction if one thought of the oil shock as an exogenous change in the price of oil rather than a decrease in the quantity supplied. Faced with an increase in fuel costs, one option a given consumer would always have would be to keep on buying as much gas as before and just pay the higher price, decreasing other expenditures as needed. The value of what is lost by such behaviour is given by $E \cdot \Delta Q$, or, to express this relative to total income *PY*,

$$\frac{E \cdot \Delta Q}{PY} = \frac{QE}{PY} \cdot \frac{\Delta Q}{Q},$$

in other words, the percentage change in oil prices $\Delta Q/Q$ is again multiplied by energy's value share QE/PY. This actually places an upper bound on the value of what the consumer loses, because, in so far as the consumer opts to reduce *E* rather than hold *E* fixed, it must be because the latter strategy is in fact an inferior option.

If these oil shocks did contribute to economic downturns, this would have to be attributed to the movements they induced in other factors of production rather than to the value of the lost energy input per se. Some modest adjustments of other factors would be anticipated in a frictionless neoclassical model, but these appear to be small. Kim and Loungani's (1992) real business cycle analysis suggested that oil price shocks could explain only a modest component of the variance of US output growth.

One modification that can make a difference is to replace the assumption of perfect competition with mark-up pricing. Rotemberg and Woodford (1996) showed that this can induce a response of labour utilization to an oil price shock that greatly amplifies the effects, with simulations in which a ten per cent increase in energy prices could lead to a 2.5 per cent drop in output six quarters later.

Another important margin is the capital utilization rate, as emphasized by Finn (2000), who was able to arrive at similar quantitative effects as Rotemberg and Woodford even under the assumption of perfect competition.

Other mechanisms

Another explanation offered for the correlation between energy prices and output has to do with the role of monetary policy. Barsky and Kilian (2002; 2004) argued that a monetary expansion was the cause of much of the 1973–4 oil price increase, and that this monetary expansion also set the stage for a subsequent decline in output. Bernanke, Gertler and Watson (1997) took the view that the oil shocks were exogenous, but the Federal Reserve responded to them by raising interest rates in order to control inflation, with this monetary contraction itself the principal cause of the downturns. Hamilton and Herrera (2004) argued that the Bernanke, Gertler and Watson conclusion was due primarily to the fact that these authors omitted the biggest effects of oil shocks corresponding to the coefficients on o_{t-3} and o_{t-4} in the regression above. Leduc and Sill (2004) added sticky prices to a theoretical model generalizing the approach considered by Finn (2000), and concluded that monetary policy makes only a modest contribution. More empirically oriented studies also concluding that the oil shocks were more important than any monetary contraction include Dotsey and Reid (1992), Hoover and Perez (1994), Ferderer (1996), Brown and Yücel (1999), and Davis and Haltiwanger (2001).

A different class of explanations emphasizes the frictions in reallocating labour or capital across different sectors that may be differentially affected by an oil shock. For example, one common consequence of an oil price shock is a sudden drop in demand for certain kinds of cars, which leads to lower capacity utilization at affected plants (Bresnahan and Ramey, 1993). Because labour and capital cannot move costlessly to alternative productive activities, the result is idle resources that can significantly multiply the effects described above. Manufacturing of transportation equipment is one of the industries most affected by oil shocks in the United States but has one of the lowest energy intensities, and thus is part of the reason that Bohi (1991) found no connection between energy intensity and output decline. Lee and Ni (2002) found that oil price shocks tend to reduce supply in oil-intensive industries but reduce demand in other industries such as autos. Davis and Haltiwanger (2001) found oil shocks reduce employment the most in industries that are more capital intensive, more energy intensive, and have greater product durability. Keane and Prasad (1996) documented significant differences across industries in the effects of oil shocks on workers' wages.

Hamilton (1988) and Atkeson and Kehoe (1999) provided theoretical analyses of the way in which technological costs of adjusting capital or labour can result in magnification of the disruptive effects of oil shocks. One of the key predictions of such models is that, unlike the factor share stories, the response of output to oil prices would not be log-linear. When oil prices go up, consumers may postpone their car purchases, but when oil prices go down, they do not go out and buy a second car. In fact, it is a theoretical possibility that, as a result of the output that is lost from trying to reallocate capital and labour, the short-run effect of an oil price decrease would actually be a decline rather than an increase in output.

Linearity

If one estimates a log-linear relation between GDP growth and lagged oil prices, the statistical significance of the relation falls as one adds more data (Hooker, 1996), suggesting at a minimum that a linear relation is either mis-specified or unstable. For example, when the regression described above is re-estimated with data through 2005:II, the result is

$$y_t = \underset{(0.11)}{0.69} + \underset{(0.07)}{0.28}\, y_{t-1} + \underset{(0.07)}{0.13}\, y_{t-2} - \underset{(0.07)}{0.07}\, y_{t-3} - \underset{(0.07)}{0.12}\, y_{t-4} - \underset{(0.006)}{0.003}\, o_{t-1} - \underset{(0.006)}{0.006}\, o_{t-2} - \underset{(0.006)}{0.002}\, o_{t-3} - \underset{(0.006)}{0.015}\, o_{t-4}.$$

Although the t-statistic on o_{t-4} remains statistically significant with a p-value of 0.02, an F-test of the null hypothesis that all four coefficients on lagged oil prices are zero would be accepted with a p-value of 0.11. The size of the effect is substantially smaller as well – whereas the 1949–80 regression would predict that GDP growth would be 2.9 per cent slower (at an annual rate) four quarters after a ten per cent oil price hike, the 1949–2005 regression would predict only 0.7 per cent slower growth. A number of authors have concluded that this instability is due to the nonlinearity of the relationship, with a linear relationship breaking down empirically when the huge oil price drops of 1985 failed to produce an economic boom. Loungani (1986) and Davis (1987a; 1987b) were the first to report evidence of nonlinearity of these relations, which they interpreted as implying that the effects of oil shocks resulted from sectoral shifts with costly reallocation of resources. Mork (1989) estimated separate coefficients on oil price increases and decreases, and found that the latter were statistically insignificantly

different from zero.

To the extent that the oil shocks are operating through an effect on demand for items such as less fuel-efficient cars, the influence would depend not just on the size of the oil price increase but also the context in which it occurred. Lee, Ni and Ratti (1995) found that much better forecasts of GDP growth were obtained if one divided the oil price increase by the standard deviation of recent price volatility. Hamilton (2003) used a flexible parametric model to investigate the nature of this nonlinearity, and found support for the Lee, Ni and Ratti formulation as well as an alternative that looks at how much the oil price might exceed its previous three-year peak; if it does not exceed the previous three-year peak, no oil shock is said to have occurred. An OLS regression of quarterly GDP growth (quoted at a quarterly rate) on lags of this net oil price measure for 1949:II to 2005:II results in the following estimates:

$$y_t = \underset{(0.12)}{0.87} + \underset{(0.07)}{0.24}\, y_{t-1} + \underset{(0.07)}{0.11}\, y_{t-2} - \underset{(0.07)}{0.08}\, y_{t-3} - \underset{(0.07)}{0.13}\, y_{t-4} - \underset{(0.012)}{0.009}\, o^{\#}_{t-1} - \underset{(0.012)}{0.014}\, o^{\#}_{t-2} - \underset{(0.012)}{0.009}\, o^{\#}_{t-3} - \underset{(0.012)}{0.031}\, o^{\#}_{t-4}.$$

Here an *F*-test of the null hypothesis that all coefficients are zero is rejected with a *p*-value of 0.006, and a ten per cent increase in oil prices above their previous three-year high is predicted to reduce quarterly GDP growth (quoted at an annual rate) by 1.4 per cent.

Similar evidence of nonlinearity, with oil price increases reducing real output growth, has also been reported for a number of other countries by Mork, Olsen and Mysen (1994), Cuñado and Pérez de Gracia (2003), and Jimenez-Rodriguez and Sanchez (2005).

Other factors and consequences

As noted by Kilian (2005), civil unrest in Venezuela in December 2002 led to a drop in production of 2.3 million barrels a day, representing 3.4 per cent of world production at the time. The net oil price series $o^{\#}_t$ reflected a surge in crude oil prices 20 per cent above their previous three-year high. Nevertheless, there was no discernible drop in GDP. Another surge in $o^{\#}_t$ of 18 per cent occurred in 2004:III, accompanied by a 1.3 per cent increase in world production, and a third surge of 21 per cent in 2005:I, accompanied by a 0.2 per cent increase in production, with no recession as of the time of this writing (August 2005). It is clear from the last two examples in particular that demand increases rather than supply reductions have been the primary factor driving oil prices over recent years. In so far as these demand increases resulted from global income growth, one wouldn't expect to see the sharp drop in consumer spending on other key items that accompanied the episodes in Table 1. At a minimum, the failure of a recession to result as of the time of this writing from the oil price increases of 2003–5 suggests that there is not simply a mechanical relation, even a nonlinear one, between oil prices and output. The experience is consistent with the claim that the key mechanism whereby oil shocks affect the economy is through a disruption in spending by consumers and firms on other goods and that, if this disruption fails to occur, the effects on the economy are indeed governed by the factor share argument.

Another potential macroeconomic effect of oil price shocks is on the inflation rate. The long-run inflation rate is governed by monetary policy, so ultimately this is a question about how the central bank responds to the oil shock. Hooker (2002) found evidence that oil shocks made a substantial contribution to US core inflation before 1981 but have made little contribution since, consistent with the conclusion of Clarida, Galí and Gertler (2000) that US monetary policy has become significantly more devoted to curtailing inflation.

See Also

- cost-push inflation
- inflation
- real business cycles

Bibliography

Atkeson, A. and Kehoe, P. 1999. Models of energy use: putty-putty versus putty-clay. *American Economic Review* 89, 1028–43.

Barsky, R. and Kilian, L. 2002. Do we really know that oil caused the great stagflation? A monetary alternative. In *NBER Macroeconomics Annual 2001*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Barsky, R. and Kilian, L. 2004. Oil and the macroeconomy since the 1970s. *Journal of Economic Perspectives* 18, 115–34.

Bernanke, B., Gertler, M. and Watson, M. 1997. Systematic monetary policy and the effects of oil price shocks. *Brookings Papers on Economic Activity* 1997:1, 91–124.

Bohi, D. 1991. On the macroeconomic effects of energy price shocks. *Resources and Energy* 13, 145–62.

Bresnahan, T. and Ramey, V. 1993. Segment shifts and capacity utilization in the U.S. automobile industry. *American Economic Review Papers and Proceedings* 83, 213–18.

Brown, S. and Yücel, M. 1999. Oil prices and US aggregate economic activity: a question of neutrality. *Federal Reserve Bank of Dallas Economic and Financial Review* (Second Quarter), 16–23.

Burbidge, J. and Harrison, A. 1984. Testing for the effects of oil-price rises using vector autoregressions. *International Economic Review* 25, 459–84.

Carruth, A., Hooker, M. and Oswald, A. 1998. Unemployment equilibria and input prices: theory and evidence from the United States. *Review of Economics and Statistics* 80, 621–28.

Clarida, R. Galí, J. and Gertler, M. 2000. Monetary policy rules and macroeconomic stability: evidence and some theory. *Quarterly Journal of Economics* 115, 147–80.

Cuñado, J. and Pérez de Gracia, F. 2003. Do oil price shocks matter? Evidence from some European countries. *Energy Economics* 25, 137–54.

Daniel, B. 1997. International interdependence of national growth rates: a structural trends analysis. *Journal of Monetary Economics* 40, 73–96.

Davis, S. 1987a. Fluctuations in the pace of labor reallocation. In *Empirical Studies of Velocity, Real Exchange Rates, Unemployment and Productivity*, ed. K. Brunner and A. Meltzer. Carnegie-Rochester Conference Series on Public Policy, 24. Amsterdam: North-Holland.

Davis, S. 1987b. Allocative disturbances and specific capital in real business cycle theories. *American Economic Review Papers and Proceedings* 77, 326–32.

Davis, S. and Haltiwanger, J. 2001. Sectoral job creation and destruction responses to oil price changes. *Journal of Monetary Economics* 48, 465–512.

Dotsey, M. and Reid, M. 1992. Oil shocks, monetary policy, and economic activity. *Federal Reserve Bank of Richmond Economic Review* 78(4), 14–27.

Ferderer, J. 1996. Oil price volatility and the macroeconomy: a solution to the asymmetry puzzle. *Journal of Macroeconomics* 18, 1–16.

Finn, M. 2000. Perfect competition and the effects of energy price increases on economic activity. *Journal of Money, Credit, and Banking* 32, 400–16.

Gisser, M. and Goodwin, T. 1986. Crude oil and the macroeconomy: tests of some popular notions. *Journal of Money, Credit, and Banking* 18, 95–103.

Hamilton, J. 1983. Oil and the macroeconomy since World War II. *Journal of Political Economy* 91, 228–48.

Hamilton, J. 1985. Historical causes of postwar oil shocks and recessions. *Energy Journal* 6(1), 97–116.

Hamilton, J. 1988. A neoclassical model of unemployment and the business cycle. *Journal of Political Economy* 96, 593–617.

Hamilton, J. 2003. What is an oil shock? *Journal of Econometrics* 113, 363–98.

Hamilton, J. and Herrera, A. 2004. Oil shocks and aggregate macroeconomic behavior: the role of monetary policy. *Journal of Money, Credit, and Banking* 36, 265–86.

Hooker, M. 1996. What happened to the oil price-macroeconomy relationship? *Journal of Monetary Economics* 38, 195–213.

Hooker, M. 2002. Are oil shocks inflationary? Asymmetric and nonlinear specifications versus changes in regime. *Journal of Money, Credit and Banking* 34, 540–61.

Hoover, K. and Perez, S. 1994. Post hoc ergo propter hoc once more: an evaluation of ‘does monetary policy matter?’ in the spirit of James Tobin. *Journal of Monetary Economics* 34, 89–99.

Jimenez-Rodriguez, R. and Sanchez, M. 2005. Oil price shocks and real GDP growth: empirical evidence for some OECD countries. *Applied Economics* 37, 201–28.

Keane, M. and Prasad, E. 1996. The employment and wage effects of oil price changes: a sectoral analysis. *Review of Economics and Statistics* 78, 389–400.

Kilian, L. 2005. Exogenous oil supply shocks: how big are they and how much do they matter for the US economy? Discussion Paper No. 5131. London: CEPR.

Kim, I. and Loungani, P. 1992. The role of energy in real business cycle models. *Journal of Monetary Economics* 29, 173–89.

Leduc, S. and Sill, K. 2004. A quantitative analysis of oil-price shocks, systematic monetary policy, and economic downturns. *Journal of Monetary Economics* 51, 781–808.

Lee, K. and Ni, S. 2002. On the dynamic effects of oil price shocks: a study using industry level data. *Journal of Monetary Economics* 49, 823–52.

Lee, K., Ni, S. and Ratti, R. 1995. Oil shocks and the macroeconomy: the role of price variability. *Energy Journal* 16(4), 39–56.

Loungani, P. 1986. Oil price shocks and the dispersion hypothesis. *Review of Economics and Statistics* 58, 536–39.

Mork, K. 1989. Oil and the macroeconomy when prices go up and down: an extension of Hamilton's results. *Journal of Political Economy* 91, 740–44.

Mork, K., Olsen, Ø. and Mysen, H. 1994. Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy Journal* 15(4), 19–35.

Rasche, R. and Tatom, J. 1977. Energy resources and potential GNP. *Federal Reserve Bank of St. Louis Review* 59(June), 10–24.

Rasche, R. and Tatom, J. 1981. Energy price shocks, aggregate supply, and monetary policy: the theory and international evidence. In *Supply Shocks, Incentives, and National Wealth*, ed. K. Brunner and A. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 14. Amsterdam: North-Holland.

Raymond, J. and Rich, R. 1997. Oil and the macroeconomy: a Markov state-switching approach. *Journal of Money, Credit and Banking* 29, 193–213. Erratum, 29 (November, Part 1), p. 555.

Rotemberg, J. and Woodford, M. 1996. Imperfect competition and the effects of energy price increases. *Journal of Money, Credit, and Banking* 28, 549–77.

Santini, D. 1985. The energy-squeeze model: energy price dynamics in U.S. business cycles. *International Journal of Energy Systems* 5, 18–25.

Santini, D. 1992. Energy and the macroeconomy: capital spending after an energy cost shock. In *Advances in the Economics of Energy and Resources*, vol. 7, ed. J. Moroney. Greenwich, CN: JAI Press.

How to cite this article

Hamilton, James D. "oil and the macroeconomy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000233> doi:10.1057/9780230226203.1215

The New Palgrave Dictionary of Economics Online

Okun, Arthur M. (1928 – 1980)

James Tobin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

budget deficits; equality of opportunity; implicit contracts; inequality; leaky bucket; new classical macroeconomics; Okun, A. M.; Okun's Law; output gap; potential GNP; redistribution of income and wealth; unemployment–inflation trade-off.

Article

Okun was born in Jersey City, New Jersey, on 28 November 1928. He died suddenly in Washington, DC, on 23 March 1980.
Okun received his BA, ranked first in his college class, in 1949 and his Ph.D. in economics in 1956, both from Columbia University. He started teaching at Yale as Instructor in 1952, and advanced up the ladder to the rank of Professor in 1963. From September 1961 to January 1969 Okun was, except for two academic years 1962–4, on leave from Yale at the President's Council of Economic Advisers (CEA) in Washington, first as a staff member, then as a Council Member 1964–8, and finally as Chairman 1968–9. When Administrations changed in 1969, Okun joined the Brookings Institution as a Senior Fellow, an appointment he held the rest of his life.
Prior to his public service in the 1960s Okun was not well known outside Yale. Those who knew him personally appreciated his extraordinary talents and virtues. He was a great and generous teacher, both in the classroom and out. His open-door office was the place for students and colleagues to get things straight, confusions dispelled, errors corrected, models repaired. A thinker of natural integrity and inexhaustible curiosity, he pursued matters in depth, unsatisfied until logic was tight and facts fell into place. His teachings of policy-oriented macroeconomics created an oral tradition that many beneficiaries remember with deep gratitude. But little of it was published, because Art Okun was unduly modest and perfectionist about putting his wisdom into print.
At the CEA Okun found another metier, macroeconomic analysis directly related to the policy issues of the day. It began when President Kennedy's Council, of which one member was from Yale, enlisted Okun as a consultant. The Council wanted to convince the President, his White House staff, the Congress and the public that reduction of unemployment from seven per cent to four per cent would

yield economy-wide benefits much greater than moving from 93 to 96 per cent employment superficially suggested. Okun was asked to estimate the gains of real Gross National Product associated with unemployment reduction. The answer became famous as Okun's Law, one of the most reliable empirical regularities of macroeconomics. Okun found that a reduction of one percentage point of unemployment was associated with a gain of three per cent in real GNP. His research, later published (1962), also provided a methodology for estimating potential GNP, the real output the economy can produce at a full-employment or ‘natural’ rate of unemployment, and the ‘Gap’ between actual and potential output.

These concepts are central to estimates of the ‘high-employment’ or ‘structural’ federal budget deficits implied by tax and spending policies, as distinguished from actual deficits, which depend also on the performance of the economy as indicated by the Gap. The entire apparatus was displayed in the 1962 *Economic Report of the President*, and was a mainstay of subsequent *Reports* for 20 years. Okun himself was a major contributor to all the *Reports* 1962–70.

Okun was the Council's principal forecaster and estimator of the consequences of alternative policies. As he won the confidence of Council chairmen, White House staff, presidents, and even Treasury secretaries, he became the obvious choice for President Johnson to appoint Council Member and then Chairman. The period 1966–9 was difficult for the CEA and for Okun personally. The four per cent unemployment target had been achieved in 1965, with negligible cost in higher inflation. Then came the acceleration of Vietnam spending, overheating the economy and lifting the inflation rate three percentage points by 1969. At the beginning of 1966 Gardner Ackley, CEA Chairman, and Okun urged President Johnson to ask Congress to raise taxes. He would not do so until too late, and even then the temporary income surtax of 1968 had disappointingly small effects. When Okun left the government in 1969, the unemployment–inflation nexus became the foremost problem on his research agenda for the rest of his career.

From 1969 much of his energy and leadership went into his brainchild, the Brookings Panel on Economic Activity, which enlisted able economists from Brookings and elsewhere for research on the major macroeconomic developments and policies of the times. The papers are published in *Brookings Papers on Economic Activity*, which under the painstaking editorship of Okun and George Perry quickly became one of the most admired professional journals in economics. The editors put the contents of every issue in perspective with their analytical summaries of the papers and discussions.

Okun had nearly completed a major treatise on macroeconomics (1981) when he died; it was edited and finished by his colleagues at Brookings. The book is a culmination of his thinking and writing over many years, his search for a coherent model of an advanced capitalist economy in a democratic society, based on his understanding of how businesses, workers and consumers behave and relate to one another. Okun did not believe that the economists’ favourite paradigm, purely competitive markets cleared by flexible prices – Adam Smith's ‘invisible hand’ – provided realistic foundations for macroeconomics. He was impressed by the informal reciprocal expectations and obligations that characterize repeated dealings between sellers and customers or employers and workers. A creative phrase-maker, Okun called this web of implicit contracts the ‘invisible handshake’. His ‘customer markets’ are in many ways efficient substitutes for price-cleared auction markets, but they are also the source of endemic macroeconomic difficulties.

Okun saw no easy resolution of the cruel dilemma policymakers face in the trade-off between unemployment and inflation. All too often, and especially in the 1970s, fiscal and monetary demand

management could achieve acceptable outcomes in one of these two dimensions only at the cost of unacceptable results in the other. Okun had no use for the monetarist view that inflation could be easily prevented or conquered if only the central bank mustered sufficient will and wisdom. Nor did he share the simplistic view of some theorists of various schools that inflations are neutral and innocuous, devoid of real consequences. He advocated structural anti-inflation policies, including wage and price guideposts strengthened by tax-based incentives for compliance, to diminish the unemployment costs of anti-inflationary monetary and fiscal measures (1978).

The intellectual climate of professional macroeconomics was inhospitable to *Prices and Quantities* when it was published. ‘New classical’ models relying on ‘invisible hand’ micro-foundations were the dominant fashion. They are theoretically appealing but have trouble explaining the commonly observed facts of business fluctuations. No one knew those facts better than Okun, whose last published paper (1980) is a masterful litany of the many ways new classical business cycle theories fail to fit them. Fashions change and controversies fade. Okun's macroeconomics will be an important component of whatever new synthesis emerges from contemporary debate.

Arthur Okun was not only an effective adviser and participant in the making of economic policy; he was also a scholar and scientist of *political economy* – the ancient name for our discipline suggests a broader scope of inquiry and concern that most economists essay today. Okun's reflections on the role of the academic policy adviser in government and on the politics and economics of macroeconomic management, published shortly after he returned to private life, are the most thoughtful of the genre (1970).

For his Godkin Lectures at Harvard (1975) Okun chose the broadest and most basic question of political economy: how democratic societies do, can, and should balance the ethical desirability of mitigating inequalities of well-being against the practical utility of the inequalities arising in free markets as incentives for efficient economic performance. Okun coined the metaphor ‘leaky bucket’ for losses in aggregate wealth incident to government interventions to transfer wealth from rich to poor. Citizens will disagree on the tolerable degree of leakage, he says, but both liberals and conservatives should face the trade-offs realistically. They should be able to agree on measures to plug leaks, exploiting opportunities to diminish inequality without impairing incentives (even if such reforms are not Pareto optimal). Okun suggests an agenda of such opportunities, focusing on measures to assure greater equality of opportunity. The book has already become a classic. In its erudition, logic, lucidity, and wisdom, and above all in its humanity, it truly reflects the qualities of its author.

Selected works

1962. Potential GNP: its measurement and significance. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 98–103. Reprinted in (1983).

1970. *The Political Economy of Prosperity*. Washington, DC: Brookings Institution.

1975. *Equality and Efficiency: The Big Tradeoff*. Washington, DC: Brookings Institution.

1978. A reward TIP. In Senate, Banking, Housing and Urban Affairs, *Anti-Inflation Proposals*,

Hearings, 95 Congress 2 Session, Washington: Government Printing Office. Reprinted in (1983).

1980. Rational-expectations-with-misperceptions as a theory of the business cycle. *Journal of Money, Credit and Banking* Part 2, 12, 817–25. Reprinted in (1983).

1981. *Prices and Quantities: A Macroeconomic Analysis*. Washington, DC: Brookings Institution.

1983. *Economics for Policymaking: Selected Essays of Arthur M. Okun*, ed. J.A. Pechman, with editor's preface. Cambridge, MA: MIT Press.

How to cite this article

Tobin, James. "Okun, Arthur M. (1928–1980)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000013> doi:10.1057/9780230226203.1216

The New Palgrave Dictionary of Economics Online

Okun's Law

Jesús Crespo Cuaresma
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Okun's law describes the empirical relationship between changes in unemployment and output at the macroeconomic level and has been regarded since its discovery by Arthur Okun (1962) as a building block of traditional macroeconomic models. This article discusses the interpretation of this relationship and summarizes recent developments in the econometric specification of Okun's law.

Keywords

aggregate demand; demand and supply shocks; Okun's coefficient; Okun's law; output gap; Phillips curve; production functions; unemployment

Article

The term ‘Okun's law’ refers to the empirical relationship between changes in unemployment and output. It is a basic building block of traditional macroeconomic models, where the aggregate supply function is derived from combining Okun's law with the Phillips curve.
In Okun's original contribution (Okun, 1962), the empirical relationship between unemployment and output is introduced in the context of the quantification of potential output and the measurement of the social costs of unemployment in terms of forgone production.
Okun (1962) presents estimates for the United States which are based on three alternative econometric specifications aimed at quantifying empirically the relationship between unemployment and output growth: (*a*) regressing (quarterly) changes in the unemployment rate on (quarterly) percentage changes in production (as proxied by Gross National Product, or GNP), (*b*) regressing the unemployment rate on percentage deviations from potential output, defined as the exponential trend in GNP and (*c*) regressing the (logarithmized) employment rate on a linear time trend and (logarithmized) GNP. The effect of output changes on unemployment is quantified by the estimated parameter associated with the output variable in each of these regressions, whose inverse is usually known as ‘Okun's coefficient’. The results

in Okun's contribution indicate that there exists roughly a three-to-one link between unemployment and output changes, in the sense that an increase/decrease of three percentage points in output (or the output gap, depending on the specification) is associated with a decrease/increase of one percentage point in unemployment. This rule of thumb is proposed as a ‘subjectively weighted average’ (Okun, 1962, p. 100) of the estimates obtained from the three specifications. Estimations based on data including the post-oil crisis period, which can be found in most modern macroeconomic textbooks, tend to reveal an Okun's coefficient that is closer to two than to three.

Okun's arguments (see, for example, Okun, 1962, p. 99) suggest that the link found between unemployment and output is not to be understood as a *ceteris paribus* relationship, but rather as capturing also the effects of simultaneous changes in labour force, hours worked and productivity (see also Friedman and Wachter, 1974). Okun argues that a reduction in the unemployment rate would induce an increase in the labour force by persuading discouraged workers to seek work actively, and also presents estimates of the increase in hours worked per employed person caused by rising output. The analysis carried out in Okun's contribution, based on data for the United States in 1960, assigns approximately 56 per cent of the change in output to the effect of changes in total labour input measured in hours worked, while the rest is attributed to productivity increases. Prachowny (1993) approaches the quantification of the link between unemployment and output by proposing a specification based on a fairly general production function, where the independent effects of changes in unemployment, hours worked, capacity utilization and labour force on output can be estimated separately. In this setting, Okun's empirical specifications would be appropriate only if certain parameter restrictions on the production function are satisfied. Prachowny (1993) therefore proposes labelling Okun's law ‘Okun's theory’ and testing these restrictions directly on the data. The estimates of the direct effect of unemployment on output obtained using this specification are correspondingly smaller than in the original contribution by Okun, although the econometric modelling strategy used (based on estimating the production function in gap form and in first differences) is not without criticism. Attfield and Silverstone (1997) reconsider this approach using cointegration techniques and find estimates that are comparable with the original values in Okun (1962).

Obviously, the relationship observed between changes in output and changes in the unemployment rate is determined by the nature of the shocks hitting the economy. The usual interpretation of the relationship summarized by Okun's law refers to arguments based on shocks to aggregate demand. Blanchard and Quah (1989) emphasize the importance of identifying demand and supply shocks in order to estimate and interpret Okun's coefficient. Using a dynamic system formed by the unemployment rate and output growth, Blanchard and Quah (1989) assess the issue by isolating supply and demand shocks and interpreting the responses of these two variables to each type of shock. The results suggest that the implied Okun's coefficient for demand shocks is slightly above two, while there is no such systematic short-run relationship between unemployment and output changes following a supply shock.

While many macroeconomic textbooks tend to emphasize that the relationship between short-run changes in output and unemployment is a robust and reliable empirical regularity, much of the literature dealing with Okun's law is aimed at evaluating the robustness of this link across countries (Kaufman, 1988; Moosa, 1997; Lee, 2000), in time (Sheehan and Zahn, 1980; Gordon, 1984; Evans, 1989), across econometric specifications (Weber, 1995; Lee, 2000) and across states of the business cycle – recessions

versus expansions – (Lee, 2000; Crespo Cuaresma, 2003). The results of this branch of literature point towards the existence of asymmetric, country-specific Okun's coefficients, with a higher elasticity of unemployment to output changes in recessions than in expansions, and a lower elasticity in continental European countries compared with Canada, the United States and the United Kingdom. The estimates of Okun's coefficient appear to be sensitive to the specification and de-trending method used for retrieving the cyclical component of output and the unemployment rate. Furthermore, this empirical literature usually reports evidence of structural instability, with a break in Okun's coefficient taking place in the 1970s.

See Also

- Okun, Arthur M.
- Phillips curve
- trend/cycle decomposition
- unemployment

Bibliography

Attfield, C.L.F. and Silverstone, B. 1997. Okun's coefficient: a comment. *Review of Economics and Statistics* 79, 326–9.

Blanchard, O.J. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.

Crespo Cuaresma, J. 2003. Okun's law revisited. *Oxford Bulletin of Economics and Statistics* 65, 439–51.

Evans, G.W. 1989. Output and unemployment dynamics in the United States: 1950–1985. *Journal of Applied Econometrics* 4, 213–7.

Friedman, B.M. and Wachter, M.L. 1974. Unemployment: Okun's law, labor force, and productivity. *Review of Economics and Statistics* 56, 167–76.

Gordon, R.J. 1984. Unemployment and potential output in the 1980s. *Brookings Papers on Economic Activity* 15, 537–64.

Kaufman, R.T. 1988. An international comparison of Okun's law. *Journal of Comparative Economics* 12, 182–203.

Lee, J. 2000. The robustness of Okun's law: evidence from OECD countries. *Journal of Macroeconomics* 22, 331–56.

Moosa, I.A. 1997. A cross-country comparison of Okun's coefficient. *Journal of Comparative Economics* 24, 335–56.

Okun, A.M. 1962. Potential GNP: its measurement and significance, in American statistical association. *Proceedings of the Business and Economics Statistics*, 98–104.

Prachowny, M.F.J. 1993. Okun's law: theoretical foundations and revised estimates. *Review of Economics and Statistics* 75, 331–6.

Sheehan, R.G. and Zahn, F. 1980. The variability of the Okun coefficient. *Southern Economic Journal* 47, 488–97.

Weber, C.E. 1995. Cyclical output, cyclical unemployment and Okun's coefficient: a new approach. *Journal of Applied Econometrics* 10, 433–45.

How to cite this article

Cuaresma, Jesús Crespo. "Okun's Law." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000098> doi:10.1057/9780230226203.1217

The New Palgrave Dictionary of Economics Online

oligarchs

Sergei Guriev
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In several countries economic transition was accompanied by the emergence of ‘oligarchs’ –businessmen who amassed fortunes and used them to influence economic policies. At their height in 2003, a few oligarchs controlled much of Russia's economy, as did a similar elite in Ukraine. Oligarchs seem to run their empires more efficiently than other domestic owners. While the relative weight of their firms in the economy is huge, it is not excessive by the standards of the global economy where most of them are operating. Policymakers should therefore focus on ‘political antitrust’ to prevent state capture and subversion of institutions.

Keywords

arbitrage; Berezovsky, B.; bribery; democracy; hold-up problem; Khodorkovsky, M.; loans-for-shares auctions (Russia); market power; oligarchs; ownership and control, separation of; ownership concentration; private property; privatization; Putin, V.; rent seeking; tariffs; total factor productivity; transition and institutions; vertical integration; World Trade Organization

Article

An oligarchy, as discussed in Plato's *Republic* and *Statesman* and Aristotle's *Politics*, is a form of government by a small group. Interestingly, while in Plato's works, *oligarchy* is used as a neutral term, and may include both aristocracy and plutocracy, Aristotle already provides the term with a negative connotation, defining oligarchy (similar to Plato's plutocracy) as a deviant form of the rule by a few (while aristocracy remains the correct one).

In its current meaning in transition economies, the term ‘oligarch’ denotes a businessman who controls sufficient resources to influence national politics. (The lists of oligarchs include only men; the richest Russian businesswoman, Moscow mayor's wife Elena Baturina, ranked outside the top 25 wealthiest Russians in 2004; she entered the *Forbes* billionaires list (Figure 1) only in 2005 but remained the only woman in the list, ranked 27 out of 34 in 2006 (*Forbes*, 2004–6). Such businessmen have played a substantial role in almost all transition countries, although most of the discussion of the role of oligarchs in transition has concerned Russia and Ukraine. The reason for this is also similar to the ideas of Plato and Aristotle, who classified oligarchy as an intermediate form of government between dictatorship/monarchy and democracy. On the one hand, EU accession countries in Central and Eastern Europe have succeeded in building accountable and democratic governments, thus limiting the role of oligarchs. On the other hand, members of the Commonwealth of Independent States (except Russia and Ukraine) have seen the concentration of power in the hands of a single politician rather than a group of rich businessmen. Also, Russian oligarchs have been more prominent than those in Ukraine, in terms of both their wealth (due to Russia's resource richness) and their substantial impact on politics. Actually, in the *Forbes* 2005 and 2006 lists the total wealth of all non-Russian billionaires from transition countries (including China but excluding Hong Kong) was less than that of the single richest Russian. Not surprisingly, Russian oligarchs have been studied in far more detail. This is why this article concentrates on the case study of Russia even though most issues are relevant to Ukraine and other transition countries. (See Aslund, 2006, for a study of Ukrainian oligarchs; Gorodnichenko and Grigorenko, 2005, provide a quantitative analysis.)

Figure 1
Numbers of Russians in the *Forbes* billionaires list, 2002–2006. *Source*: *Forbes* (2002–6).

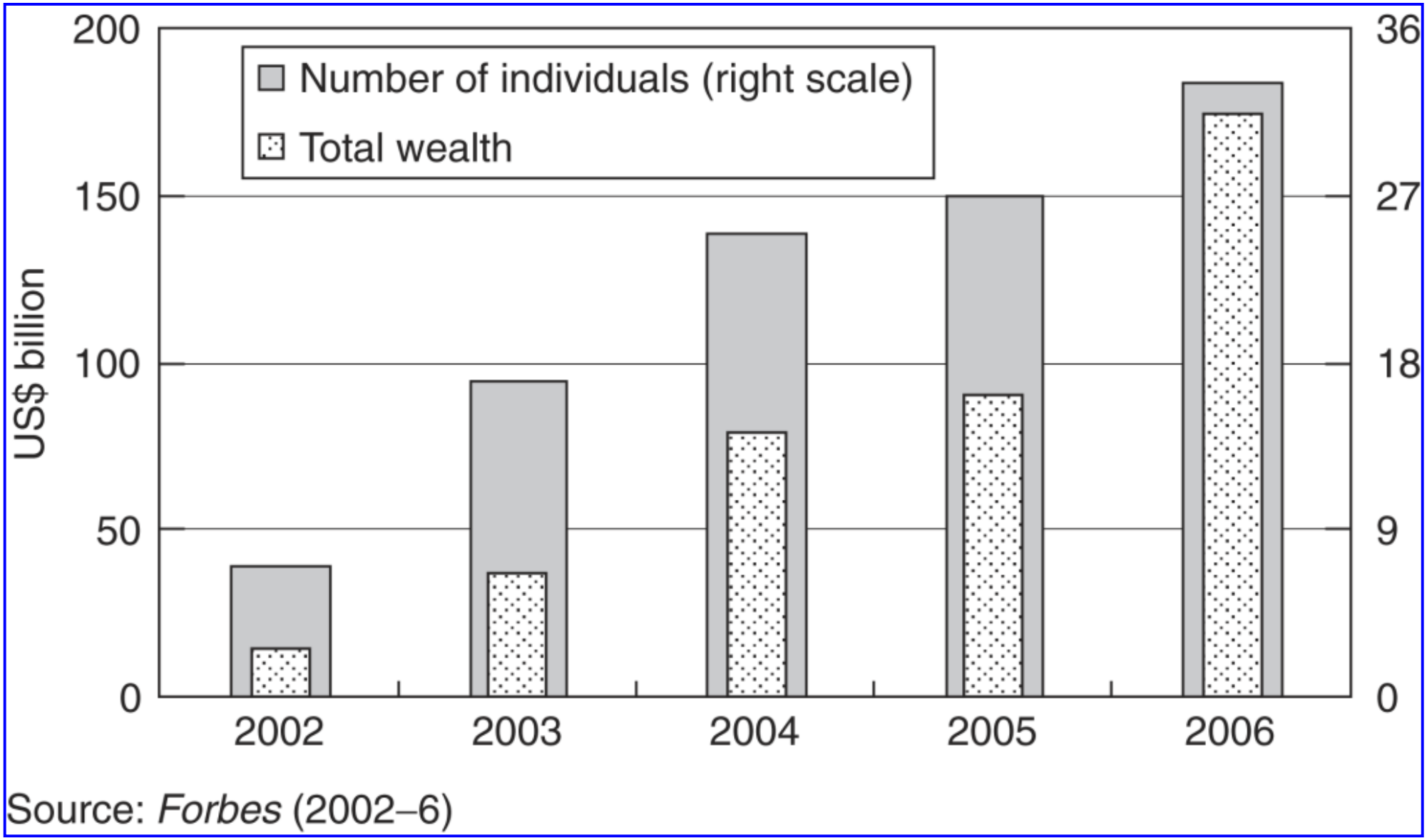

It is not clear who first used the term ‘oligarch’ to describe the newly emerged class of Russian tycoons. *Kommersant* (2003) refers to a pro-market politician Boris Nemtsov (then a governor of Nizhny Novgorod region, later to become a deputy prime minister) and a journalist, Alexander Privalov (then *Izvestiya* daily and *Expert* weekly), both introducing the term in 1994–5. It is also clear that the Russian elite's thinking of oligarchs has been affected by Jack London's *The Iron Heel* (1908), an anti-utopia on the rise of an oligarchy of robber barons, which was widely publicized in Soviet times.

Who are the oligarchs?

There is no complete list of Russian oligarchs. Given the multi-layered and non-transparent ownership structure of Russian companies, compiling such a list would be extremely difficult. On the other hand, any such list has to be constantly updated: there is substantial vertical mobility among Russia's richest. For example, out of seven or eight business groups that dominated President Yeltsin's Russia in the 1990s, two were destroyed by the 1998 crisis (SBS and Inkombank), one took a hit but survived to be later sold to fellow billionaires

(Roskredit-cum-Metalloinvest), two have their leaders (Berezovsky and Gusinsky) in exile, and one (Khodorkovsky) in prison. Other problems are related to the vagueness of the definition of oligarchs. First, there are different views on how to measure tycoons' power rather than wealth (this is especially important for a comparison between oligarchs and US robber barons). Second, it is not clear whether to count public officials and CEOs of large public companies as oligarchs. In what follows, we stick to the definition of oligarchs as private owners, although certain CEOs of state-owned firms and family members of some government officials do resemble oligarchs in many respects.

The first list of oligarchs probably belongs to Boris Berezovsky (by all accounts, an oligarch himself) who, in his 1996 interview in the *Financial Times*, named seven bankers who controlled about 50 per cent of the productive assets of the Russian economy. Since then there have been numerous lists, some even endorsed by the oligarchs themselves. Still, all the oligarch rankings identify similar sets of individuals. Table 1 presents a list that was constructed based on a study of ownership concentration in a substantial subset of Russian economy by the World Bank's 2004 *Country Economic Memorandum* (CEM) for Russia (see Guriev and Rachinsky, 2005, for a detailed description of the data-set and the project). The study refers to summer 2003 – the oligarchs' heyday. While this study has its limitations, it makes it possible to reach some conclusions on who the Russian oligarchs are, and why they matter.

Russian oligarchs as of summer 2003

Senior partner(s)	Holding company/firm, major sector(s)	Employment, '000s (% sample)	Sales, in billions of roubles (% sample)	Wealth, in billions of US dollars	Other rankings[a]	RSPP bureau, head of committee/taskforce (as of June 2004)
Oleg Deripaska	Base Element/RusAl, aluminum, auto	169 (3.9)	65 (1.3)	4.5	P, BR, DS, K, F	B, Railroad reform
Roman Abramovich	Millhouse/Sibneft, oil	169 (3.9)	203 (3.9)	12.5	S, BR, DS, K, H, [b] F	
Vladimir Kadannikov	AutoVAZ, automotive	167 (3.9)	112 (2.2)	0.8	BR, K	
Sergei Popov, Andrei Melnichenko, Dmitry Pumpiansky	MDM, coal, pipes, chemical	143 (3.3)	70 (1.4)	2.9	F	B, Financial markets (Mamut[b])
Vagit Alekperov	Lukoil, oil	137 (3.2)	475 (9.2)	5.6	S, P, BR, DS, K, F	
Alexei Mordashov	Severstal, steel, auto	122 (2.8)	78 (1.5)	4.5	BR, DS, F	B, Customs and WTO accession
Vladimir Potanin, Mikhail Prokhorov	Interros/Norilsk Nickel, non-ferrous metals	112 (2.6)	137 (2.6)	10.8	B, S, P, BR, DS, K, F	B, Social and labour relations (Eremeev[b])
Alexandr Abramov	Evrazholding, steel	101 (2.3)	52 (1.0)	2.4	F	B
Len Blavatnik, Victor Vekselberg	Access-Renova/TNK-BP, oil, aluminum	94 (2.2)	121 (2.3)	9.4	DS, F	B
Mikhail Khodorkovsky[c]	Menatep/Yukos, oil	93 (2.2)	149 (2.9)	24.4	B, S, P, BR, DS, K, H, F	B, International affairs
Iskander Makhmudov	UGMK, non-ferrous metals	75 (1.7)	33 (0.6)	2.1	K	
Vladimir Bogdanov	Surgutneftegaz, oil	65 (1.5)	163 (3.1)	2.2	P, BR, DS, K, F	
Victor Rashnikov	Magnitogorsk Steel, steel	57 (1.3)	57 (1.1)	1.3		
Igor Zyuzin	Mechel, steel, coal	54 (1.3)	31 (0.6)	1.1		
Vladimir Lisin	Novolipetsk Steel, steel	47 (1.1)	39 (0.8)	4.8	F	B
Zakhar Smushkin, Boris Zingarevich, Mikhail Zingarevich	IlimPulpEnterprises, pulp	42 (1.0)	20 (0.4)	1		
Shafagat Tahaudinov	Tatneft, oil	41 (1.0)	41 (0.8)	2.9		

Mikhail Fridman	Alfa/TNK-BP, oil	38 (0.9)	107 (2.1)	5.2	B, S, P, BR, DS, K, F	B, Judiciary reform
Boris Ivanishvili	Metalloinvest, ore	36 (0.8)	15 (0.3)	8.8	P	B, Land reform (Kiselev[b])
Kakha Bendukidze	United Machinery, engineering	35 (0.8)	10 (0.2)	0.3	BR, K	B, Budget and taxes
Vladimir Yevtushenkov	Sistema/MTS, telecoms	20 (0.5)	27 (0.5)	2.1	S, P, BR, DS, K, F	B, Industrial policy, Pension reform (Yurgens[b])
David Yakobashvili, Mikhail Dubinin, Sergei Plastinin	WimmBillDann, dairy/juice	13 (0.3)	20 (0.4)	0.2		
Total		**1,831 (42.4)**	2,026 (39.1)			

Sources: Employment and sales are from World Bank (2004) and Guriev and Rachinsky (2005). The percentages in parentheses are the shares of employment/sales of the World Bank's sample, which in turn covers a substantial share of the economy (yet, as some industries are not represented, the list misses a couple of important candidates, such as Alexander Lebedev of National Reserve Corporation). Wealth is the market value of the oligarchs' stakes in spring 2004, calculated by authors using Forbes (2004) and stock market data. Wealth includes stakes of all the partners identified by the survey. Each entry lists the leading shareholder(s) in a respective business group, the name of the holding company or the flagship asset, and one or two major sectors. Several individuals per group are reported only when there is equal or near equal partnership. Ranking is based on employment in the sample and may therefore be different from the actual, as the sample disproportionally covers assets of different oligarchs. Employment and sales are based on official firm-level data for 2001. The exchange rate was 29 roubles to the US dollar.

RSPP=Russian Union of Industrialists and Entrepreneur, the leading lobbying organization for Russian business. Among other things, RSPP represented the private sector in multiple meetings with President Putin, including the first one where the 2000 pact was allegedly concluded. B=RSPP Bureau membership (in total, the RSPP Bureau includes the President and 24 members); we also list the RSPP committees/taskforces the particular oligarchs are in charge of (in total there are 17 committees/taskforces in the RSPP).

[a]Other oligarch rankings. B: Berezovsky's Group of Seven (*Financial Times*, 1996). BR: Boone and Rodionov (2002). DS: Dynkin and Sokolov (2002). F: *Forbes* (2004). H: Hoffman (2003). K: *Kommersant* (2003). P: Pappe (2000). S: Classified as oligarchs in Freeland (2000, pp. xv–xvii).

[b]Some RSPP committee chairs have retired from active business. Eremeev was an Interros executive prior to the appointment at RSPP. Hoffman discusses Berezovsky rather than Abramovich. In 2000–3, Abramovich took over most of Berezovsky's assets in Russia as Berezovsky went into exile. Kiselev was Metalloinvest Board Chairman at the time of appointment at RSPP. Mamut was MDM Board Chairman at the time of appointment at RSPP. Yurgens was a Sistema executive prior to the appointment at RSPP.

[c]Khodorkovsky remained a Bureau member and a Committee Chair for a while even after he was imprisoned, indicted and even convicted.

How important are the oligarchs?

First, the oligarchs do control a substantial part of Russian economy. In the CEM sample, they account for about 40 per cent of sales and employment – more than all other private owners combined, or more than federal and regional governments combined. (As of June 2006, quite a few of these oligarchs have seen their assets nationalized, so a more relevant figure would be 30 per cent.)

Cross-country comparisons of wealth concentration are usually based on the share of stock market capitalization controlled by a given number (often ten) of families. Certainly, it is not a perfect metric – after all, it doesn't include firms not listed on stock markets, and emerging markets are likely to provide at best an imperfect measure of value. But we are not aware of comparable data-sets on non-listed firms, so we have to rely on the data on the share of the stock market owned by the top ten families. By that measure, ownership concentration in modern Russia is higher than in any other country for which the data are available. The top ten families or ownership groups (a subset of Table 1) owned 60.2 per cent of Russia's stock market in June 2003. This percentage is much higher than in any country in Continental Europe, where the share of the ten largest families is less than 35 per cent in small countries and less than 30 per cent in all large countries. In the United States and the United Kingdom, this share is in single-digit percentages. (A less rigorous approach is to look at the *Forbes* billionaires lists. Even though Russian companies are significantly undervalued relative to their OECD counterparts, *Forbes*, 2004, lists 26 billionaires in Russia; only the United States and Germany have more. The 26 Russian billionaires are worth $81 billion, or 19 per cent of Russia's annual GDP. The 26 richest US citizens are worth four per cent of US GDP; the total wealth of all US billionaires is less than seven per cent of US GDP.) In the East Asian countries before the 1997 crisis, the highest shares of the ten largest families were in Indonesia (58 per cent), Philippines (52 per cent), Thailand (43 per cent) and Korea (37 per cent). The numbers for Indonesia and Philippines include the

holdings of the Suharto and Marcos families, each controlling 17 per cent of total market capitalization in the respective countries. In Russia, the personal wealth of ex-President Yeltsin and President Putin is considered to be very modest.

What do oligarchs control?

Each group in Table 1 controls assets in multiple provinces of Russia and even other countries, and in several industries. Mostly, the oligarchs' conglomerates are horizontally and vertically integrated. (Only Abramovitch, Deripaska, MDM group, and Potanin control major assets in unrelated industries, but even in their empires a single industry accounts for most of the conglomerate's value.) Oligarchs do dominate the largest industrial sectors, in particular natural resources (especially oil and metals) and automotive. The only large sectors not controlled by oligarchs are natural gas, energy, and manufacture of machinery. The gas and energy sectors are dominated by federally owned monopolies Gazprom and RAO UES; machinery production is a diverse sector which is populated by defence equipment suppliers (controlled by the federal government), oligarch firms and smaller firms controlled by non-oligarch private domestic owners.

Do oligarchs exercise excessive market power in the sectors that they control? The sectors controlled by oligarchs are indeed those with the highest concentration ratios in Russia (Guriev and Rachinsky, 2005). However, these are also tradable goods sectors that are subject to global competition. For example, consider the ten sectors where oligarchs control more than 20 per cent of total sales. Except for ore and automotives, all these sectors sell to the global market: they export 30 to 90 per cent of their output; indeed, these sectors account for half of total Russian exports. The first exception, ore production, is mostly owned by oligarchs' vertically integrated conglomerates, where ore is an input. The second exception, the automotive sector, is a classic example of interest group politics. Russian cars are not internationally competitive, and the industry has always relied on protection. Such protection was usually granted, especially in the period in the 1990s when the largest carmaker's CEO, Vladimir Kadannikov, served as the first deputy prime minister in charge of economic policy. Yet, even with high import duties and support for domestic producers through generous tax write-offs and subsidies, import penetration was 25 per cent and rising. As of 2000, Oleg Deripaska consolidated his control over the second largest car producer and almost all of the bus and truck production, and the lobbying for stronger protection reached new heights. Indeed, one of the main reasons Russia is not yet a member of the World Trade Organization is that the WTO requires lowering import duties for cars, and Russia's automotive lobby launched an aggressive (and a very successful) anti-WTO campaign. The lobbyists managed to install increasingly high tariffs on both used and new imported cars.

The large industries where oligarchs play a large role are also those with substantial economies of scale. Indeed, these are exactly the sectors where large business empires originated in many countries in the late 19th century and the early 20th century, including the United States, Japan and Sweden. But, except for the automotive sector, there seems little reason for concern that Russia's oligarchs have excessive market power. Although their conglomerates are large by Russian standards, they are certainly not excessive by global standards. Some oligarchs are important global players in their industries (especially in oil and metals), but none is a dominant market leader. Thus, there is no basis, on efficiency grounds, for antitrust policies aimed at breaking up the oligarchs' companies. Instead, it is more important that Russian competition policy assure a level playing field for all owners without regard to their size and political influence.

How did the oligarchs gain control?

A common belief is that the oligarchs owe their fortunes to the 'loans-for-shares' auctions held in mid-1990s, which are widely regarded as the most scandalous episode of Russian privatization. In the classical loans-for-shares scenario, the government appointed a commercial banker to run an auction that would allocate a controlling stake of a large natural resource enterprise in exchange for a loan to the federal government that the latter never intended to repay. Not surprisingly, the auctioneer always awarded the stake to himself for a nominal bid (usually, slightly above a very low reserve price) by excluding all outside bidders. The scheme was designed to consolidate the bankers' support for Yeltsin's re-election campaign in 1996.

The conventional loans-for-shares story fits Abramovich (in 1995–7, a junior partner of Berezovsky), Khodorkovsky, and especially Potanin. The other two winners were the oil sector insiders Alekperov and Bogdanov, who obtained stakes in firms they already controlled. However, most of those listed in Table 1 did not become oligarchs through the loans-for-shares programme. Some of the 22 largest owners tried to participate in the loans-for-shares programme and even offered more competitive bids, but were excluded by those in charge of respective auctions; some even raised their concerns in public.

Most of the individuals listed in Table 1 are relatively young: nine of them are in their thirties, and 13 are in their forties. (Both mean and median individuals in Table 1 are 44 years old. Russian oligarchs are much younger than their American counterparts. In the *Forbes* 2004, list, the average age of the 25 richest Americans is 64 years; the average age of all 262 US billionaires is the same.) The older oligarchs have typically come from Soviet-era nomenklatura. Prior to transition, they were either managing their respective enterprises or working in government agencies supervising those enterprises. When Soviet-era enterprises were privatized, they successfully converted their de facto control into ownership rights. The younger entrepreneurs started from scratch in the late 1980s, building their initial wealth during President Gorbachev's partial reforms when the coexistence of regulated and quasi-market prices created huge opportunities for arbitrage. In 1992, as price liberalization and privatization began, most of them owned trading companies and/or banks. Thus, when

privatization of industrial enterprises occurred, they had the financial capital available to purchase ownership in privatization auctions. Some of these entrepreneurs were neither industry nor government insiders; yet, they converted Soviet manufacturing enterprises into successful modern capitalist firms. Of course, a cynic might note that such companies are near the bottom of the list in Table 1 in terms of size, while the loans-for-shares winners dominate the top of the list.

Oligarchs ’ dilemmas

Whatever the source of individual oligarchs’ wealth, the Russian public still deems it illegitimate, believing that the oligarchs obtained their initial wealth through connections and furthered it by securing preferential treatment through exerting political influence. (In a July 2003 poll by ROMIR, an independent Russian research and polling agency, 88 per cent responded that all large fortunes were amassed in an illegal way, 77 per cent said that privatization results should be partially or fully reconsidered, and 57 per cent agreed that the government should launch criminal investigations against the wealthy; *Vedomosti*, 2003.) This has created a fundamental problem for Russia's transition: promoting democratic values (that is, respecting the median voter's opinion) may undermine liberal values (private property rights in a substantial part of the economy). This conflict has created a window of opportunity for such a pragmatic politician as President Vladimir Putin, who has managed to play oligarchs and voters off against each other to consolidate his own political power.
Curbing the oligarchs’ political influence was an essential part of Vladimir Putin's presidential campaign in 2000. In his open letter to voters, he promised to treat the oligarchs in the same way as other entrepreneurs; a few days later he announced that all interest groups would be kept at an ‘equal distance’ from his government. In the first meeting with the leading oligarchs on 28 July 2000, President Putin offered them the following pact. As long as the oligarchs paid taxes and did not use their political power (at least not against Putin), Putin would respect their property rights and refrain from revisiting privatization. This pact defined the ground rules of oligarchs’ interaction with central and regional government during Putin's first term (2000–4). Although the pact could have never been written, even the general public was well aware of its existence. A poll by FOM (2000), an independent non-profit Russian polling organization, a week after the meeting showed that 57 per cent Russians knew about it.
Putin's threat to prosecute any oligarch who deviated from the pact was based on the median voter's support for expropriating the oligarchs. Putin carried out his threat in 2003, when the prominent oligarch Mikhail Khodorkovsky, the majority owner of the Yukos oil company, deviated from the pact by openly criticizing corruption in Putin's administration and supporting opposition parties and independent media. He and his partners were soon arrested or forced into exile, and their stakes in Yukos expropriated. It is not clear why Khodorkovsky did not stick to the pact. Perhaps he thought that supporting opposition parties rather than challenging Putin himself was not a violation. Almost certainly, he did not expect Putin to respond so decisively.
The expropriation of the Yukos shareholders certainly involved serious costs for Russian economy – the investment climate worsened and capital flight increased substantially. However, Putin clearly demonstrated that his priority was to establish his credibility even if this damaged his economic agenda. The Yukos affair has clarified the rules of the game between oligarchs and the Kremlin. Oligarchs have learned the risks associated with violating the pact, and so in the future they will be less likely to interfere in national politics. The Yukos affair effectively shifted the bargaining power from oligarchs to bureaucrats. Although outright expropriation of oligarchs will probably remain just a threat, their cash flows will be milked more intensively by bureaucrats in the form of kickbacks, donations to pet projects, and direct bribes (for a discussion of this ‘contract’ between bureaucrats and the entrepreneurs as a ‘viability insurance contract’, see Ickes, 2005). This will in turn undermine oligarchs’ property rights and incentives to invest. To sustain economic growth, Putin has to constrain rent-seeking by his own bureaucrats. This task is certainly not an easy one, given that democratic checks and balances are very weak. Moreover, neither government nor the oligarchs are interested in the development of democracy and civil society. (Actually, oligarchs may also benefit from imperfect property rights protection as there are economies of scale in private rent-seeking; see Glaeser, Scheinkman and Shleifer, 2003; Rajan and Zingales, 2003; Sonin, 2003.) Bureaucrats do not like to cede their control, while oligarchs are afraid of the median voter's redistributive agenda.
The potential exit strategy for any individual Russian oligarch is to sell a large stake to a reputable foreign investor. Indeed, expropriating foreigners is harder for the state because they are more popular than oligarchs, and because of pressures from foreign governments. However, timing the exit properly is a complex problem. Selling too early would bring too little as the assets are initially undervalued. Delaying the sale in order to restructure the company and improve its transparency would raise the price, but would also increase the risk of expropriation by the Russian government. This expropriation may also occur through a seemingly market-based transaction. For example, the government can use public funds to pay the oligarch the market value of his assets in exchange for (hidden) substantial side payments to selected government officials or their pet projects. Given the threat of complete expropriation, this is an offer the oligarch cannot refuse.

Economic performance of the oligarchs

Do oligarchs create value or strip assets? Do they improve the performance of the firms they control or injure their performance?
Most oligarchic groups are horizontally or vertically integrated and are run by active majority owners, so the usual ‘conglomerate discount’ diseconomies of scale are unlikely to apply. A more important problem is, of course, the political risk of expropriation that shortens time horizons and reduces the incentives to invest.
On the other side, several arguments suggest that Russia's oligarchs might improve firm performance. First, the oligarchs’ performance might be superior because they have

successfully overcome the separation of ownership and control. An oligarch who owns a very large majority share should have strong incentives to restructure companies and to seek to improve the value of this asset, rather than for diverting cash flows and stripping the assets. Even if a firm was originally privatized to dispersed shareholders, its ownership structure was quickly consolidated through dilution and, in some cases, outright expropriation of outside investors, including government and foreigners. The current champions of transparency, Mikhail Khodorkovsky and Vladimir Potanin (now chairing Russia's National Council for Corporate Governance), kept expropriating outside investors until as recently as 1999. In our sample, oligarchs do control large stakes in their firms. In an average firm where the largest owner is the oligarch, he controls 79 per cent; in the case of non-oligarch private domestic owners, the corresponding figure is only 74 per cent. The difference is statistically significant but not necessarily economically important. The average degree of control exercised by smaller owners over their companies is also very high. Poor protection of minority shareholders rights has resulted in consolidation of control within most Russian companies. As a result, smaller owners are not investors that hold small stakes in large companies; rather, they hold large stakes in small companies.

Second, vertical integration can mitigate the risk of hold-up problems, where in a situation of relatively few buyers and sellers each party must be concerned that the other will attempt to renegotiate and seize a greater share of the joint surplus. Many oligarch empires have been built to overcome such hold-up problems: for example, all Russian major oil companies are vertically integrated; most steel producers own sources of coal and ore; some companies own ports, fleets of railroad cars and even railroad track. Third, in a situation with underdeveloped financial markets, external finance is costly; larger oligarch-run firms can benefit from their access to internal finance. They can create an internal financial market to finance expansion (see Khanna and Yafeh, 2005, for the discussion of these two benefits for business groups in developing countries). Fourth, Russia lacks a clear rule of law, and the larger conglomerates are certainly more effective than small firms in influencing judicial and political decisions and protecting their property from the predatory ‘grabbing hand’ of federal and local governments.

There is still no convincing test of whether and how oligarchs affect the performance of their firms. Constructing such a test is a significant challenge. Preliminary results (Guriev and Rachinsky, 2005) show that in terms of total factor productivity growth (with industry, region and size controlled for) oligarchs’ firms do perform almost as well as foreign firms and better than other Russian-owned firms. Yet more empirical work is needed to control for endogeneity of oligarch ownership, and to study the long-term effects. In addition, more work is needed to produce a quantitative evaluation of the oligarchs’ effect on social welfare.

Oligarchs and Russia's future

While ownership concentration in Russia is higher than in other countries today, it does not seem unprecedented in historical perspective. Owners of Korean chaebols, Japanese zaibatsu, Sweden's and Italy's largest family controlled firms, and US ‘robber barons’ exercised a similar share of economic and political power. Also, in many of these countries the oligarchs’ wealth was accumulated with substantial support from the state (in direct subsidies, tax breaks, land grants, subsidized credit, and so forth) and was deemed illegitimate by a substantial share of the public at some points in history. Yet these countries have managed to build functioning market economies, although it took much longer for some of them to create functioning democracies. Therefore, it is not clear whether and how soon Russia will succeed in establishing legitimacy of private property rights and whether this will be accompanied by a transition to a sustainable democracy.

See Also

- inequality (international evidence)
- policy reform, political economy of
- privatization
- rent seeking
- state capture and corruption in transition economies
- transition and institutions

This article draws substantially on Guriev and Rachinsky (2005) *and mostly refers to the situation in Russia prior to the renationalization campaign that started in 2004*.

Bibliography

Aslund, A. 2006. *Revolution in Orange: The Origins of Ukraine's Democratic Breakthrough*. Washington, DC: Carnegie Endowment for International Peace.

Boone, P. and Rodionov, D. 2002. *Rent Seeking in Russia and the CIS*. Moscow: Brunswick UBS Warburg.

Dynkin, A. and Sokolov, A. 2002. Integrated business groups in the Russian economy [in Russian]. *Voprosy Ekonomiki* 4, 78–95.

Financial Times. 1996. Moscow's Group of Seven. 1 November, p. 17.

FOM (Fond Obschestvennogo Mnenia [Public Opinion Foundation]). 2000. *Government and Large Business: A Poll of 1500 Russian Citizens Held on 5 August 2000* [in Russian]. Online. Available at http://bd.fom.ru/report/cat/societas/market_economy/economic_reform/private_enterprise/power_and_business/d001621, accessed 16 June 2006.

Forbes. 2002–6. The world's billionaires. Online. Available at http://www.forbes.com/billionaires, consulted 16 June 2006.

Freeland, C. 2000. *Sale of the Century: Russia's Wild Ride from Communism to Capitalism*. New York: Crown Business.

Glaeser, E., Scheinkman, J. and Shleifer, A. 2003. *Journal of Monetary Economics* 50, 199–222.

Gorodnichenko, Y. and Grygorenko, Y. 2005. Are oligarchs productive? Theory and evidence. Mimeo, University of Michigan.

Guriev, S. and Rachinsky, A. 2005. The role of oligarchs in Russian capitalism. *Journal of Economic Perspectives* 19(1), 131–50.

Hoffman, D. 2003. *The Oligarchs: Wealth and Power in the New Russia*. New York: Public Affairs.

Ickes, B. 2005. Economic pathology and comparative economics: why economies fail to succeed. Presidential address. *Comparative Economic Studies* 47, 503–19.

Khanna, T. and Yafeh, Y. 2005. Business groups in emerging markets: paragons or parasites? Discussion Paper No. 5208. London: CEPR.

Kommersant. 2003. *Who Owns Russia?* [in Russian]. Moscow: Vagrius.

London, J. 1908. *The Iron Heel*. London and New York: Macmillan.

Pappe, Y. 2000. *The Oligarchs*. Moscow: Higher School of Economics.

Rajan, R. and Zingales, L. 2003. *Saving Capitalism from the Capitalists: Unleashing the Power of Financial Markets to Create Wealth and Spread Opportunity*. New York: Crown Business.

Sonin, K. 2003. Why the rich may favor poor protection of property rights. *Journal of Comparative Economics* 31, 715–31.

Vedomosti. 2003. Take away and divide: people's aspirations have not changed in 86 years [in Russian]. 18 July.

World Bank. 2004. *From Transition to Development: A Country Economic Memorandum for the Russian Federation*. Moscow: World Bank.

How to cite this article

Guriev, Sergei. "oligarchs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000095>

doi:10.1057/9780230226203.1218

The New Palgrave Dictionary of Economics Online

oligopoly

P. Sylos-Labini
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advertising; barriers to entry; competition; concentration; differentiated oligopoly; duopoly; full cost principle; Great Depression; imperfect competition; increasing returns; innovation; market power; mixed oligopoly; monopoly; obstacles to entry; Okun's Law; oligopoly; price determination; price leaders; price rigidity; price variation; price wars; productivity growth; returns to scale; specialization economies; Sraffian economics; technical change; Verdoorn's Law

Article

No article entitled ‘oligopoly’ appeared in any edition of Palgrave's *Dictionary of Political Economy*. It is true that the simplest case of oligopoly, that is, duopoly, was considered more than a century and a half ago, by Cournot; but such an analysis was motivated by purely theoretical interests. The fact is that only in the 20th century and especially after the Second World War did this market form become important in economic reality, as a result of two processes of economic change: the process of concentration and the process of differentiation. In those branches where the former process has asserted itself – for example, steel, basic chemical products, cement, electricity – concentrated oligopoly with relatively homogeneous products has emerged; where the latter process has prevailed, we find differentiated oligopoly; in those branches where both processes have taken place simultaneously, then mixed oligopoly has emerged. In both processes innovations have played a major role, with the proviso that in the process of concentration innovations have given rise to economies of scale, whereas in the process of differentiation the most important role has been that of technological innovations implying economies of specialization; in this case, technological innovations are combined with commercial innovations. In fact, differentiated oligopoly can be found mainly in those activities in which quality competition, commercial services and advertising have had a particularly relevant role – non-durable consumer goods, such as textiles, tyres, canned foods, soft drinks and cigarettes are often produced in conditions of differentiated oligopoly.
In the past, when the standard of living of the masses of consumers was not much above the subsistence level, there was not much scope for the factors just mentioned. With the gradual increase of per capita

income, consumers' preferences have acquired an increasing space. At the same time the possibility of advertising has been greatly enhanced by particular innovations – modern means of transportation and the so-called mass media, among which radio and television play a special role. Mixed oligopoly (concentration cum differentiation) is typical of several industries producing consumer durables such as automobiles, typewriters, refrigerators, radio and television sets, computers; mixed oligopoly can be found in several important service sectors such as banking and insurance. In addition a large number of non-durable consumers' goods and services – including commercial services – constitute the area where differentiated oligopoly prevails; it is well to point out that as a rule there is no difference between imperfect competition and differentiated oligopoly. Analytically, the former can be seen, as a rule, as a first and the latter as a second approximation; this standpoint becomes natural if we recognize that the imperfect markets are composed by a 'chain of oligopolistic groups'.

After careful reflection, we are bound to admit that in modern industry and in services, oligopoly, in its three varieties, is the rule and competition the exception – to be found in certain industries producing sufficiently homogeneous non-durable goods and in subsidiary activities. Competition, on the other hand, is the rule in most agricultural and mineral raw materials traded in international markets.

1. According to the traditional (neoclassical) conception, markets in competitive conditions are formed by a great number of firms, each of which is so small as to be unable to influence prices. Each firm, then, is bound to accept the market price and pushes output up to the point at which marginal cost – which, after a point, cannot but be increasing – equals price. In fact, the increasing marginal cost, that is, diminishing returns both in the short and in the long run are a necessary feature of traditional theory. In monopoly equilibrium is reached when the decreasing marginal income equals marginal cost. Indeed, according to that theory, only two market forms are worth consideration – competition and monopoly – the former being the rule, the second the exception. (The analytical tools to be used for imperfect competition are those worked out for monopoly.)

The whole analysis is statical and thus presupposes given technology. To work out theoretical models consistent with dynamic analysis, we have to go back to the classical concept of competition, where freedom of entry and not the number and size of firms is crucial. If we adopt this concept, it is easy to shift from competition to non-competitive market forms, by considering obstacles of various relevance to entry. Clearly, when in a given market the obstacles to entry are serious, firms operating in that market are likely to be few; this, however, is to be seen, not as a preliminary datum, but as the likely (not necessary) result of the existence of those obstacles.

When the obstacles to entry are of little importance, then a super-normal profit will attract new firms: supply will increase and the price will fall, so that super-normal profit will tend to disappear: such a profit can persist when obstacles to entry are important.

Having chosen this approach, in a first approximation we have to distinguish, in price analysis, between agriculture and mining, on the one hand, where obstacles to entry as a rule are modest, and industry and services, on the other, where those obstacles are often considerable. Again, in the first approximation, we can state, with Ricardo, that in primary activities in the short run prices depend on demand and supply, whereas in the long run they depend on costs. If we refer to the short run and intend to work out an analysis susceptible of empirical verification, we realize that 'demand' can be variously interpreted; in the case of raw materials traded in the international markets, demand can best be represented by an index of world industrial production. In industry and services, instead, in the short run prices depend principally on changes in direct costs and, in the long run, on changes in total costs per unit.

The reason for this sharp difference as regards short-run variations of prices is as follows. In primary activities firms, owing to the relative freedom of entry, have no outstanding market power and cannot influence prices, which vary according to the variations of aggregate demand and aggregate supply. In the other activities, however, prices are to a non-negligible extent controlled by firms and, in particular, by those that act as price leaders. Starting from a price that is accepted by all firms – that is, from an 'equilibrium price' – the firms acting as leaders will modify it when the conditions of equilibrium change. There are, then, two analytical problems, conceptually different but strictly interrelated: the problem of price determination and that of price variations. In traditional terms, the former problem belongs to the area of statical analysis, the latter to that of dynamics. We can accept such a distinction provided that it implies no cleavage, that is, provided that we can pass without discontinuities from the analysis of price determination to that of price variations.

The problem of price determination implies the analysis of the equilibrium, which includes: the size of the market (that is, the position in a Cartesian diagram of the demand curve, a concept that becomes relevant when firms are no more conceived as atoms); the shape of the demand curve (that is, the elasticity of demand); technology, salaries and other administrative expenses; taxes; and the prices of durable and those of variable means of production. This is not the place to present a formal solution of the problem of price determination. Suffice it to say that the concept of entry-preventing price and elimination price are important analytical tools to be used in the construction of a theoretical model of price determination. Once the price reaches the level acceptable to all firms – the equilibrium level – each firm is in a position to calculate the markup, that is, the ratio between price and cost or, more precisely, direct cost. When the equilibrium conditions change, the price is to be changed. Normally this occurs without a price war, since such wars are costly and major firms are willing to undertake them if only the expected gains (net of risks) are higher than expected costs, an occurrence that does not appear to be frequent.

The analytical steps, then, are two: the first is to understand how the equilibrium price is arrived at; the second is to understand how it varies when the equilibrium conditions change. If in the first step the concepts of entry-preventing and elimination prices are essential, in the second step it is the 'full cost principle' that plays the key role. Empirical enquiries have consistently shown that this principle is generally followed by managers operating in non-agricultural activities. Yet for a long period it has been considered only as a rough rule of thumb, without theoretical relevance. Probably the reasons are twofold. The first is that it contradicts the received doctrine, which is founded on marginal analysis and which, as a condition of equilibrium, assumes a rising marginal cost – the full cost principle, instead, which is based on the markup on direct costs, assumes the marginal cost to be constant and therefore equal to direct cost. The second reason is that that principle has been described as if it were a criterion to determine the price, not to modify it, but it can have a meaning only in the second case. Thus, Hall and Hitch (1939), in their pioneer empirical enquiry, report that 'prime (or "direct") cost per unit is taken as the base, a percentage addition is made to cover overheads•…•and a further conventional addition•…•is made for profit'. However, it is evident from this statement that the crucial theoretical problem is to explain the height of the two percentage additions – that can be unified into one percentage. Thus, given the cost elements, we have to explain the conditions that limit the discretionary powers of managers in choosing a given percentage and not another, that is, we have to explain the equilibrium conditions. Only after having explained the equilibrium price can the markup acquire a meaning. In other words, the full cost principle is theoretically meaningless as regards the problem of price determination and

becomes meaningful as regards the problem of price variations: in fact, barring price wars, the markup appears to be the quickest and most rational way for firms, and particularly for price leaders, to arrive at a new equilibrium price when the equilibrium conditions vary.

The further question is to understand why direct cost and not total unit cost is taken as the term of reference to modify the price in the short run – say, year by year or even in shorter periods. The reason is that the changes in the prices of variable factors affect without much delay all firms, though not necessarily in the same proportions, whereas the changes in the other equilibrium conditions – size of the market, elasticity of demand, technology, salaries and other overhead costs – affect the firms at different degrees and in different times. These changes either affect prices in relatively long periods or do not affect them at all – substantial increases in overhead costs can be offset, not by price increases, but through productivity increases. To be sure, when these are insufficient, the increases in overhead costs can push some of the firms out of the market; this can also be the outcome of unfavourable changes in market conditions.

Changes in direct costs, then, tend to be shifted to prices in the short run. But even for this category of changes a sort of hierarchy is necessary: changes in the prices of raw materials (including the sources of energy) tend to be fully shifted on prices of finished products in both directions, since those changes tend very quickly to affect all firms. This is not so for changes in wage cost per unit, since this cost is given by the ratio between wages and productivity. Now, wage changes – if we except the areas of the so-called submerged economy – affect in a relatively short run all firms, whereas productivity increases due to organizational innovations and to technological changes determined by previous investment tend to take place at different rates in the different firms (declines in productivity are exceptional): only those changes in wage cost per unit of output have to be shifted onto prices that are common in both the upward and the downward direction. However, under contemporary conditions the shift in the downward direction will be more limited than that in the upward direction, since it is unlikely that the prices of finished industrial products in international markets will generally decrease; and it is international competition that, in industry, will limit the market power of the firms of a given country. Briefly, in the short run, the shift of changes in total direct costs will tend to be not only partial but also asymmetrical.

In the case of industrial products, then, short-run variations in prices depend on the variations of direct costs: demand does affect prices, but, as a rule, only in the long run and not in the same direction, as is the case in the short-run variations of prices under competitive conditions, but in the opposite direction, since the long-run expansion of demand makes the entry of new firms easier and opens the possibility of exploiting economies of scale. Thus, an expansion of demand tends, *ceteris paribus*, to reduce and not to raise the price. In the short run demand increases have no significant direct effect on prices of industrial goods; they can have an indirect effect, that is, via the prices of raw materials, when demand pressure is so strong as to affect not only finished products but also raw materials. As for finished products, demand pressure tend to affect not prices but (consistently with the Keynesian conception) the level of activity.

2. If we pass from partial to general analysis and adopt the framework of a Sraffian model, we are bound to distinguish between basic and non-basic (‘luxury’) products. If we decide to consider not only competitive but also non-competitive markets, we have to drop either the assumption of a unique rate of profit or the assumption of a unique wage rate (for a given type of labour). In any case, prices enter into the conditions of simple reproduction. The conditions of expanded reproduction, that is, of accumulation – to use the Marxian expression – imply, in addition, that at least a share of the surplus be employed

productively, that is, invested – the velocity of accumulation being determined by that basic product that has got the lowest surplus. It is important to point out that technological progress is essential not only in the case of accumulation but also in the case of simple reproduction, since mineral products tend gradually to exhaust themselves; it is essential also in the case of a growth proportional to the increase of population, not only due to the reason just mentioned, but also due to the necessity of offsetting the tendency of diminishing returns in agriculture.

3. If we adopt a Sraffian model of general analysis, the study of the effects of technological changes meets with several problems, certainly serious, but, in principle, not insurmountable; in fact, some important steps in this direction have already been made by Sraffa himself. That study, instead, seems to be precluded if we adopt a Walrasian model of general equilibrium that implies a strictly static framework, in which all firms operate in conditions of diminishing returns, that is, of increasing marginal costs. Now, barring special cases, increasing returns are to be related to changes in the methods of production, even in the short run: increases in the productivity of labour can take place as a consequence of quick readjustments of the labour force and of innovating investment carried out in previous periods. A long series of empirical observations – among which may be mentioned Dunlop's 1938 article on the movement of real wages, the ‘Verdoorn Law’ and ‘Okun's Law’ – show that increasing, not diminishing, returns dominate modern economies and, in particular, non-agricultural activities. Thus, to admit that it is not perfect but imperfect competition and oligopoly that is the rule seems to be the only way to reconcile theoretical models and empirical enquires in both partial and general analysis.

4. In the short run technical progress takes mainly the form of increases in productivity of means of production and, in particular of labour; in the long run one has also to consider the production of new goods, that in the short run represent a tiny fraction of the total. The diversification of output, which in fact conditions the growth of all firms, can assume either a prevailingly commercial character, in the case where the goods are already in the market, or also a technological character, if the goods or the process through which they are produced are new. In its turn, the expansion of demand represents the condition for the introduction of two important types of technological innovations – that is, new goods and new processes implying the exploitation of economies of scale – which, after all, is nothing but another way to re-propose the Smithian proposition according to which ‘the division of labour is limited by the extent of the market’.

For the sake of simplicity, we limit ourselves to considering, as the index of technical progress both for the short and the long run, the increase in productivity of labour. The basic consequence of this increase is, at the aggregate level, a systematic divergence between the average variations of nominal incomes and the average variations of prices, with the fall of the relative prices of those goods produced in the most dynamic industries. Referring to average variations, the said divergence can take four different forms:

	Nominal incomes	*Prices*
(*a*)	falling	falling more rapidly
(*b*)	constant	falling
(*c*)	rising	constant
(*d*)	rising	rising more slowly

Cases of falling prices – (*a*) and (*b*) – were frequent in the 19th century, when the process of

concentration and that differentiation in industry and services had not proceeded far enough and competition was still the rule in those sectors. In the 20th century case (*a*) occurred during the first four years of the Great Depression; but, in sharp contrast to what was normally occurring in the 19th century, the level of activity in industry and services fell much more than prices, whereas in agriculture the prices fell violently, but the level of activity remained approximately constant. The comparison with the great depression of the 19th century – which occurred in the years 1873–9 – is illuminating.

Putting aside services, which offer a picture similar to that of industry, the percentage changes in prices and production of agriculture and industry during the two great depressions (I and II) were as follows:

		United Kingdom		United States	
		Prices	Production	Prices	Production
Agriculture	I	–18	+3	+31	+4
	II	–44	0	–54	+2
Industry	I	–29	–5	–33	–5
	II	–21	–16	–23	–48

If we except the period of the great depression of the 20th century, which was in all senses an exceptional event, with productivity varying in a very irregular fashion, in the 20th century cases (*c*) and (*d*) – rising nominal incomes with constant prices or prices rising more closely – were the rule. Now, it is not indifferent that the fruits of technical progress have one type of consequence or the other on prices and incomes.

When prices of all goods fall, the means of production (Sraffa's basic products) become cheaper and this stimulates the expansion of all firms, including those that do not introduce innovations. On the other hand, when prices fall demand increases automatically in real terms.

Let us now consider what happens when productivity rises but prices do not fall. If, in such circumstances, nominal incomes do not rise, the whole increase in productivity tends to translate itself into a decreasing level of employment; to have at least a stable level of employment, nominal incomes should rise in proportion to the increase in productivity; and this is not an automatic process. It is unlikely that wages and salaries rise if there is not a systematic action of trade unions, unless the process of differentiation and the consequent increasing fragmentation in the labour market have become so widespread as to favour wage increases even without a generalized pressure of trade unions. On their side, non-labour incomes will increase only if investment or government expenditure increases, or both. Investment can increase only if new investment opportunities arise, due to technical innovations, whereas government expenditure can increase as a political decision. On the other hand, with stable prices, the firms that do not introduce innovations cannot receive the stimulus arising from the means of production becoming cheaper. As a result, the process of growth tends to become more and more unbalanced, unless a general expansion of demand – originated by innovations and/or by government – takes the place of the stimulus afforded by an overall fall in prices.

In short, owing to the obstacles to entry, in most non-agriculture activities the ‘competitive mechanism’ for the distribution of the fruits of technical progress (falling prices, stable nominal incomes) has been more and more substituted by the ‘oligopolistic mechanism’ (stable prices and increasing nominal incomes). In the new conditions, the process of growth requires increasing intervention of public powers, but not necessarily in the form of increasing public expenditure. That intervention can consist of

taxation (to afford incentives or to put brakes), or can support the prices and the incomes of those activities, like agriculture, least affected by those two processes, or can promote the source of technological innovation, that is, scientific research, or – to give another important example – can create conditions favourable to development of small firms, not only with fiscal and credit incentives, but also by supplying real services – especially commercial and technical assistance. All these measures of public powers can push up the growth of the volume of investment to the velocity required to avoid an increase of unemployment or gradually to reduce it to the frictional level.

If the countervailing influences of public interventions process are not strong enough, in the new conditions the process of growth tends to become more unbalanced not only from the standpoint of the different industries (since those that do not carry out innovations directly have no more the stimulus determined by the declining prices of the means of production they use), but also from the point of view of income distribution. In fact, the downward price rigidity tends to create special margins in certain industries or in certain firms. These rising margins do not necessarily become above-normal profits; they can become, too, above-normal wages or salaries, depending on the relative strength of the opposing parties. Instead of the above-normal incomes, the advantages for workers can also take the form of a greater stability of employment; similarly, the advantages of capitalists can take the form, rather than of above-normal profits, of more stable profits.

5. It seems that in recent times the process of differentiation has become more important than the process of concentration and the economies of specialization seem to have become more important than the economies of scale. This new development in industry has been promoted by at least three changes: (1) the growth of electronics and allied industries; (2) the reaction of increasing masses of workers in advanced countries against the monotony of assembly lines and other methods of mass production; (3) the growing differentiation in consumer preferences originated by the increasing per capita income. In services, differentiation has always been important and in recent times has become even more important; at the same time, services become the most important section of the economy in the so-called post-industrial societies. Considering the declining relative weight of agriculture and mining in advanced societies, we have to conclude that the area of flexible prices tends to shrink and that of rigid prices to expand – I mean flexibility or rigidity in the downward direction. In particular, the area of rigid prices tending to expand refers more and more to services and less and less to industry; this phenomenon, that has important consequences also on the overall behaviour of prices, up to now has received very little attention.

It remains true, however, that the increasing rigidity of prices of goods and services determines the need for an increase in demand large enough to avoid a decline in employment, if population grows. Now, with the diffusion of high education, with the space for a rapidly increasing number of goods opened up by the increasing per capita income, in recent times the potentialities of development have increased. But such potentialities can remain unexploited if they are left to spontaneous market forces; given the rate of interest, all depends on investment stimulated by technological innovations that promise to be profitable and that can be devised and carried out by private firms without the support or the stimulus afforded by public powers. If those investments are not enough to promote an increase of demand capable of generating an increase in income at least equal to that in productivity, unemployment gradually grows. It is well to emphasize that the main obstacles to a policy of economic growth arise not by diminishing returns, but either from the side of the public deficit, if the increasing supply of bonds pushes up the rate of interest; or from the side of the foreign deficit, which pushes up the value of

foreign currencies, giving rise to a special kind of inflationary pressure. Such problems are aggravated by the fact that the two deficits, to some extent, reinforce each other: for instance, large firms tend to borrow abroad, owing to the high internal rate of interest. But these are matters that go beyond the limits of our theme.

See Also

- game theory

Bibliography

Andrews, P.W.S. 1949. *Manufacturing Business*. London: Macmillan.

Bain, J.S. 1956. *Barriers to New Competition*. Cambridge, MA: Harvard University Press.

Baumol, W.J. 1967. *Business Behavior, Value and Growth*. Rev. edn. New York: Harcourt Brace and World.

Bhagwati, J.N. 1970. Oligopoly theory, entry-prevention, and growth. *Oxford Economic Papers* 22, 297–301.

Chamberlin, E.H. 1933. *The Theory of Monopolistic Competition: A Reorientation of the Theory of Value*. Cambridge, MA: Harvard University Press.

Dunlop, J. 1938. The movement of real and money wage rates. *Economic Journal* 48, 413–34.

Eichner, A.S. 1976. *The Megacorp and Oligopoly: Microfoundations of Macrodynamics*. Cambridge: Cambridge University Press.

Fellner, W. 1949. *Competition Among the Few: Oligopoly and Similar Market Structures*. New York: Knopf.

Galbraith, J. 1957. Market structure and stabilization policy. *Review of Economic and Statistics* 39 (May), 124–33.

Hall, R.L. and Hitch, C.J. 1939. Price theory and business behaviour. *Oxford Economic Papers* 2, 12–45.

Kaldor, N. 1935. Market imperfection and excess capacity. *Economica*, NS 2, 33–50.

Kalecki, M. 1943. Costs and prices. Repr. in M. Kalecki, *Selected Essays on the Dynamics of the Capitalist Economy 1933–1970*, Cambridge: Cambridge University Press, 1971.

Modigliani, F. 1958. New developments on the oligopoly front. *Journal of Political Economy* 66, 215–32.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan. New edn, 1969.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Sylos-Labini, P. 1956. *Oligopoly and Technical Progress*. Trans., Cambridge, MA: Harvard University Press, 1969.

Sylos-Labini, P. 1984. *The Forces of Economic Growth and Decline*. Cambridge, MA: MIT Press.

How to cite this article

Sylos-Labini, P. "oligopoly." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000014> doi:10.1057/9780230226203.1219

The New Palgrave Dictionary of Economics Online

Olson, Mancur (1932 – 1998)

Joe A. Oppenheimer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Mancur Olson was one of the small group of economists in the 20th century who laid the foundation of rational choice theorizing about non-market behaviour. He demonstrated self-interested individuals have a great incentive to free ride rather than to contribute to the supply of a public good. He also showed how self-interested group behaviour explained why nations tend to stagnate after periods of growth. Utilizing the notion of profit seeking political entrepreneurs, he argued the benefits of democratic systems were to contain the extractive costs imposed by government and to extend the time horizons for property rights.

Keywords

Arrow, K.; bandits, roving; Baumol. W.; Buchanan, J.; collective action; constitutionalism; democracy, theory of; dictatorship; Downs, A.; economic growth; free rider problem; interest groups; kleptocracy; non-market behaviour; non-market economics; Olson, M.; political entrepreneur; public goods; Samuelson, P.; social dilemmas; time horizon; Tullock, G.: on constitutionalism; Von Neumann, J.; tragedy of the commons

Article

Along with a handful of other economists of the 20th century (Kenneth Arrow, James Buchanan, Anthony Downs, and John von Neumann), Mancur Olson laid the foundation for the adoption of rational choice theorizing about non-market behaviour in the social sciences. His work (1965) on the relationship between the rational choice of individuals and the performance of groups had a revolutionary impact on the fields of sociology and political science. In 1967 he left his first academic job at Princeton University to become Deputy Assistant Secretary of the US Department of Health, Education and Welfare. From there he went to the University of Maryland, where he held the position of Distinguished Professor of Economics, co-founded the University of Maryland's Center for Collective Choice, and

founded the Institute for Research on the Informal Sector (IRIS).

Mancur Olson was born in January 1932 to a Norwegian-American farming family in North Dakota's Red River Valley. The valley contained some of the richest farmland in the state; the family grew mainly flax and did quite well. Neither his parents nor other members of that generation in the Olson family were educated beyond high school, but his father and his uncle were intellectually curious and questioning of society's arrangements. Mancur grew up on the farm and, as the eldest of three sons, he was permitted to be party to the adults' conversations about farming and social problems.

Throughout his life he recalled those early discussions regarding the shared interests of farmers in getting a fair price for their crops, the difficulties in their meeting other common concerns and the many references to the ability of the Scandinavian countries to overcome narrow interests to achieve both social justice and economic growth. He noted these as the part of his inheritance that motivated his life-long research interests in the problems of collective action, social justice and economic prosperity.

Mancur went to college at North Dakota State University on an Air Force Reserve Officer Training Corps (ROTC) scholarship. There he studied agricultural economics and had the good fortune to be mentored by Rainer Schickle (the father of the American composer, Peter Schickle). He won a Rhodes scholarship and went to Oxford, only to discover that Oxford dons could not imagine that a graduate (1954) from North Dakota's Agricultural College could qualify for entry into their graduate programme of Philosophy, Politics and Economics. So, unlike most of the other Americans coming from more prestigious institutions, he was required to get a second BA from Oxford before going on for an M.Phil. At Oxford he met his lifelong companion and wife, Allison, who was also getting her M.Phil. (in history). The Olsons left Oxford together for the environs of Boston, where Allison had a job at Smith College and Mancur was to get a Ph.D. at Harvard. Two barriers were created. First, Mancur's chosen advisors, first Kenneth Galbraith and then also Otto Eckstein, left for Washington to work in the Kennedy administration. Further, Air Force officials discovered that Mancur had yet to do his service for his North Dakota Air Force ROTC contract, and they required him to leave Harvard to do military service. That service was performed between Rand, Brookings and the Air Force Academy for two years, after which he was able to finish his work again at Harvard under the tutelage of Thomas Schelling. During this time their family grew and eventually Allison and Mancur had four children: Elicka, born in 1963, a veterinarian; Severn, born in 1967, a civil servant; Sander, born in 1969, a journalist; and Garth, who died in infancy.

Olson's major contribution to economics and to the social sciences more broadly was in the analysis of 'non-market' economics. He focused both on how individual non-market behaviour and political institutions (broadly understood) affected socio-political and economic outcomes. Many of his most important findings are encapsulated in his three major books *The Logic of Collective Action* (LCA) (1965), *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities* (RD) (1982), and *Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships* (PP) (posthumous, 2001).

His first book, LCA, grew out of his dissertation, and focused on the non-Paretian outcomes one can expect from unorganized groups of individuals in their efforts to secure costly public goods (that is, goods where consumers cannot be excluded and consumption by one does not diminish consumption by another) such as air quality and peace. LCA built on the findings of William Baumol (1967) and Paul Samuelson (1954) who had shown that suboptimality was to be expected from rational self-interested behaviour regarding public goods. Olson expanded their arguments and generalized them by noting that

the satisfying of virtually all shared interests is a form of public good, thereby selling the argument to the non-economist. The crux of the observation is simple: self-interested individuals have a great incentive to free ride rather than to contribute to the supply of a public good. Individuals will, after all, receive the good if others supply it. In LCA, Olson also tried to develop an argument that the size of the group was central to the analysis, but this was later shown to be erroneous (Frohlich and Oppenheimer, 1970; Hardin, 1982). The work spawned a paradigm shift in the study of group behaviour in both political science and sociology.

In RD, Olson built on LCA (and also the 1962 work of Buchanan and Tullock, who argued that one could evaluate constitutional rules by the externalities imposed upon losing subsets of the population by the extraction of resources for redistribution to the winners). In RD Olson argued that narrow-interested lobbying groups, designed to extract rewards from the general population via governmental action, clung to stable political systems much like barnacles to a ship's hull. Such extractive interests were shown to be more harmful the narrower the interests they represented. Newer political systems, built on cataclysmic changes in a society, were likely to be relatively free of such encumbrances and hence would lead to less wasteful extraction. Therefore, their economies would be more likely to exhibit substantial and sustained growth than would those associated with more established, stable political systems. He expanded the analysis (1990) to consider the comparative efficiency of the Scandinavian political systems' foundation on a coalition of a very few, very broad political interests. These welfare states were contrasted with welfare states in other industrialized countries built on a patchwork quilt of narrow, coalesced social-interest groups.

Coupled with Downs's 1957 work *An Economic Theory of Democracy*, LCA also sparked a reconsideration of political leaders as entrepreneurs (Salisbury, 1969; Frohlich, Oppenheimer and Young, 1971) as a way of solving the collective action problem. In the 1990s Olson himself began to mine the profit motive as a tool to understand the motivational characteristics of political leaders, and to reconsider the social gains from democracy. By assuming politics was necessarily based on coercive taxes, he considered the evolution of political systems as a hypothetical history from roving bandits to stationary bandits and then to kleptocratic political leaders constrained by the rules of succession and, more generally, competition. Roving bandits would take what they could. Stationary bandits, who controlled an area (for example, 'war lords' and mafiosi) would find it worth their while to ensure the prosperity of the population they exploited. Rules of succession, such as those that underlie monarchies, were shown to change the time horizon for maximizing the extractive behaviour of the kleptocrat, thereby giving incentives to investments that had longer time horizons. Using the finding that narrower interests impose greater costs on society than wider ones, Olson (1993) and McGuire and Olson (1996) showed the general gain from democratic (majoritarian) systems to be the decrease in imposed external costs by the winning kleptocrats, as well as the extension of the time horizons for property rights. His last book, PP, built upon his kleptocratic entrepreneurial arguments and their relation to the time horizon of politicians. Long-term property and other rights were seen to be a key to the development of more complex financial markets that underlie modern economic development.

Olson's heritage is extraordinarily wide: the general interest in 'social dilemmas' grew directly out of his work via the translation of LCA into the language of *n* person game theory. His sure-handed encouragement of young scholars interested in non-market economics helped foster the multidisciplinary adoption of rational choice theoretic tools in the social sciences in general.

See Also

- Buchanan, James M.
- collective action
- public goods
- rational choice and political science
- tragedy of the commons

Selected works

1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. New Haven: Yale University Press.

1990. *How Bright Are the Northern Lights? Some Questions about Sweden*. Lund: Institute of Economic Research, Lund University.

1993. Dictatorship, democracy, and development. *American Political Science Review* 87, 567–76.

1996. (With M. McGuire.) The economics of autocracy and majority rule: the invisible hand and the use of force. *Journal of Economic Literature* 34, 72–96.

2001. *Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships*. New York: Basic Books.

Bibliography

Baumol, W. 1967. *Welfare Economics and the Theory of the State*, 2nd edn. Cambridge, MA: Harvard University Press.

Buchanan, J. and Tullock, G. 1962. *The Calculus of Consent*. Ann Arbor: University of Michigan Press.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.

Frohlich, N. and Oppenheimer, J. 1970. I get by with a little help from my friends. *World Politics* 23, 104–21.

Frohlich, N., Oppenheimer, J. and Young, O. 1971. *Political Leadership and the Supply of Collective Goods*. Princeton, NJ: Princeton University Press.

Hardin, R. 1982. *Collective Action*. Baltimore, MD: Johns Hopkins University Press.

Salisbury, R. 1969. An exchange theory of interest groups. *Midwest Journal of Political Science* 13, 1–32.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36, 387–89.

How to cite this article

Oppenheimer, Joe A. "Olson, Mancur (1932–1998)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000094> doi:10.1057/9780230226203.1220

The New Palgrave Dictionary of Economics Online

Oncken, August (1844 – 1911)

Jürg Niehans

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

history of economic thought; Oncken, August; Physiocrats; Quesnay, François; *Tableau Economique*

Article

Oncken was born in Heidelberg on 10 April 1844 and died in Schwerin (Mecklenburg) on 10 July 1911. After studies in Munich, Heidelberg and Berlin, Oncken first became a landowner in Oldenburg. Behind his scholarly interest in physiocracy was a life-long interest in agriculture. He began his academic career as university lecturer in economics and statistics at the Vienna School of Agriculture. In 1878, after a brief interlude at the Aachen Institute of Technology he accepted an appointment as professor of economics at the University of Bern, where he taught a wide range of courses until his retirement (because of failing eyesight) at the end of 1909.

As a general economist, Oncken has little claim to our attention. He never had a correct understanding of things like, say, diminishing returns, and he remained an unsophisticated advocate of protection, particularly for agriculture (1901a), applauding Henry Carey as the greatest living economist (1874). As an historian of economic thought, however, he was one of the leading lights between 1870 and 1920.

In his earliest historical paper (1874) Oncken criticized Adam Smith, in the spirit of German economics of that time, for his ‘materialism’ and his radical ‘laissez faire’ doctrines. In *Adam Smith and Immanuel Kant* (1877) he confessed that these criticisms did not survive a careful reading of the original sources. Instead he now stressed the similarities between those two giants of moral philosophy.

In Bern, Oncken's interests shifted to the Physiocrats. The result was a series of masterpieces of archival detective work and historical interpretation. It begins with a paper on the relationship between the Physiocrats and their disciples in Bern (1886a). In the following monograph (1886b), Oncken traces the maxim ‘laissez faire’ to d'Argenson (and not to Boisguillebert, as Stephan Bauer states in the *Encyclopaedia of the Social Sciences*) and further back to the time of Colbert, while ‘laissez passer’ was later added by de Gournay in a conversation with Mirabeau. In this context, Oncken puts forth the startling conjecture (not reiterated in (1902)) that the *Tableau Economique* was originally printed in support of a bid by Quesnay for the premier ministership.

Oncken's edition of Quesnay's writings (1888) became fundamental for all further work in this field. The sought-for completeness, however, eluded Oncken, because his very publication set off a renewed search of archives, culminating in Bauer's discovery of an early (but still not the first) version of the *Tableau Economique* (published by the British Economic Association) and of the article ‘Hommes’ in 1890. A further article, ‘Impôts’, was later published by Schelle. On the other hand, Oncken's collection includes non-economic writings not available in the 1958 edition, as well as the basic biographical sources. A first, hand-written draft of the *Tableau* was later reproduced in Oncken's *History of Political Economy* (1902).

Oncken himself made use of much of the newly discovered material in a succession of essays on Quesnay's life (1894–6) and the history of physiocracy (1893a, 1893b and 1897a). He was well aware that the time was not ripe for a definitive biography, but for brilliance of historical scholarship Oncken's essays are unsurpassed. It is regrettable that, being available only in German and in inaccessible journals, they are usually not given the credit they deserve.

With respect to the circumstances under which the *Tableau Economique* was first printed, we do not seem to have progressed much beyond Oncken. The story that the most famous single page in the history of economics was typeset and printed by a bored Louis XV with his own hands, Oncken regarded as a fable, mainly because of its incompatibility with the known facts about the King's character. Schelle, however, chose to treat the story, despite its implausibility, as historical fact and his view was still accepted by Jacqueline Hecht in 1958.

Of the *History of Political Economy* (1902), only the first volume appeared, dealing with the time before Adam Smith. The first half, reaching from antiquity to mercantilism, is today of little interest. The second half, treating the Physiocrats and their predecessors, is still a valuable source of historical information about men, books and ideas, making an effective case for Quesnay as the ‘founder’ of economic science.

Oncken later returned to Adam Smith by defending him, not without some polemics, against his detractors of the Schmoller School (1897b, 1898). In another paper (1909), he also pointed out that Smith did not borrow from Ferguson, but had valid reasons for feeling that Ferguson had borrowed from his lecture notes.

Selected works

1870. *Untersuchung über den Begriff der Statistik*. Leipzig.

1874. *Adam Smith in der Kulturgeschichte. Ein Vortrag*. Vienna.

1877. *Adam Smith und Immanuel Kant. Der Einklang und das Wechselverhältniss ihrer Lehren über Sitte, Staat und Wirthschaft. Vol. 1: Ethik und Politik*. Leipzig.

1886a. *Der ältere Mirabeau und die ökonomische Gesellschaft in Bern*. Berner Beiträge zur Geschichte der Nationalökonomie No. 1, Bern.

1886b. *Die Maxime Laissez faire et Laissez passer, ihr Ursprung, ihr Werden. Ein Beitrag zur*

Geschichte der Freihandelslehre. Berner Beiträge zur Geschichte der Nationalökonomie No. 2, Bern.

1888. ed. *Oeuvres économiques et philosophiques de François Quesnay, fondateur du système physiocratique*. Berner Beiträge zur Geschichte der Nationalökonomie No. 3, Frankfurt and Paris.

1893a. Zur Geschichte der Physiokratie. *Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft im Deutschen Reich* 17, Leipzig.

1893b. Ludwig XVI und das physiokratische System. *Zeitschrift für Litteratur und Geschichte der Staatswissenschaften* 1, Leipzig.

1894–6. Zur Biographie des Stifters der Physiokratie, François Quesnay. *Zeitschrift für Litteratur und Geschichte der Staatswissenschaften* (from 1896: *Vierteljahrsschrift für Staats- und Volkswirtschaft, für Litteratur und Geschichte der Staatswissenschaften aller Länder*), Vols 2, 3, 4, Leipzig.

1895. Political economy in Switzerland. *Economic Journal* 5(March), 133–7.

1896–8. Letter from Switzerland. *Economic Journal* 6(June), 1896, 308–14; 7, September 1897, 228–93; 8, June 1898, 269–73.

1897a. Entstehen und Werden der physiokratischen Theorie. *Vierteljahrsschrift für Staats- und Volkswirtschaft, für Litteratur und Geschichte der Staatswissenschaften aller Länder*, Vol. 5. Leipzig.

1897b. The consistency of Adam Smith. *Economic Journal* 7(September), 443–50.

1898. Das Adam Smith-Problem. *Zeitschrift für Socialwissenschaft* 1, Berlin.

1901a. *Was sagt die Nationalökonomie als Wissenschaft über die Bedeutung hoher und niedriger Getreidepreise?* Offprint from *Monatliche Nachrichten zur Regulierung der Getreidepreise*. Berlin.

1901b. Quesnay, François. In *Handwörterbuch der Staatswissenschaften*, 2nd edn, Vol. VI, Jena.

1902. *Geschichte der Nationalökonomie. Erster Teil: Die Zeit vor Adam Smith*. Leipzig.

1909. Adam Smith und Adam Ferguson. *Zeitschrift für Socialwissenschaft* 12, Leipzig.

How to cite this article

Niehans, Jürg. "Oncken, August (1844–1911)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_O000016> doi:10.1057/9780230226203.1863

The New Palgrave Dictionary of Economics Online

operations research

Ilan Vertinsky
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Operations research (OR) is both a profession and an academic discipline. It involves the application of advanced analytical methods to improve executive and management decisions. This survey highlights the types of OR models and techniques in common use. It explores the roots of OR and its theoretical and professional evolution, and presents the current trends which shape its future.

Keywords

allocation problems; chance-constrained programming; computational methods; critical path method; Dantzig, G.; data mining; dual method; dynamic pricing; dynamic programming; e-commerce; financial engineering; game theory; globalization; graph theory; information technology; integer programming; inventory theory; Kantorovich, L. V.; lattice programming; linear programming; marketing engineering; Markov processes; multi-criteria programming; network flow optimization; operations research; polynomial algorithms; polynomial submodular set functions; probability theory; production control theory; program evaluation and review technique; quadratic programming; Quesnay, F.; queuing theory; revenue management methods; simplex method for solving linear programs; simulation; stochastic programming; supermodularity; von Neumann, J.; Walras, L.

Article

Operations research is commonly referred to as OR. In the United Kingdom, where the first formally recognized group of practitioners was formed, it is called ‘operational research’. Other names, such as ‘management science’, ‘operational analysis’ and ‘systems analysis’, are frequently used as synonyms. Definitions of OR abound. The differences among these definitions reflect important dimensions of conflict in philosophy and perception of the field among the members of the various communities identifying themselves as operations researchers. It is instructive, therefore, to examine some of these definitions to identify areas of agreement about the distinctive characteristics of the field as well as those

dimensions which are a cause of tension.

The Operational Research Society of the UK, the oldest OR professional society, developed the following official definition (Dando and Sharp, 1978, p. 940):

> Operational Research is the application of the methods of science to the complex problems arising in the direction and management of large systems of men, machines, materials and money in industry, business, government and defense. The distinctive approach is to develop a scientific model of the system, incorporating measurements of factors such as chance and risk, with which to predict and compare outcomes of alternative decisions, strategies or controls. The purpose is to help management determine its policy and actions scientifically.

Its current website suggests that OR is the discipline of ‘applying advanced analytical methods to make better decisions’ (the OR Society). While these definitions see OR as an eclectic, problem-centred approach where scientific methods are employed to help management, definitions proposed in the United States view OR as a science or as a distinctive methodology providing scientific bases for decision-making. The constitution of the Operations Research Society of America (ORSA) referred to OR as ‘the science of operations research’ (House, 1952, p. 28). This view was incorporated in 1982 in the Decision and Management Program of the U.S. National Science Foundation that referred to the emergence of a combined theoretical and empirical science of operational and managerial processes (Little, 1986). The current website of the Institute for Operations Research and Management Science (INFORMS) which has succeeded ORSA, refers to OR as the ‘science of better’.

The scientific view of OR sees its goals as (*a*) the development of models of operations that represent the causal relationship between controlled variables, uncontrolled variables and system performance, and (*b*) the development of the computational means for identifying levels of controlled variables in ways that help managers of a system achieve systems outputs as close as possible to the ones they desire.

A broader and more proactive variant of the scientific view of OR was proposed in the first major textbook to be published on OR (Churchman, Ackoff and Arnoff, 1957). It suggested that the goal of OR is an overall understanding of optimal solutions to executive-type problems in organizations. This comprehensive goal of OR implies a normative prescriptive role with boundaries emancipated from mere reactive problem-solving.

Examination of the various definitions of OR establishes the following features upon which there is almost general agreement: (*a*) OR focuses upon executive and management-type decisions in organized systems; (*b*) a distinct feature of the methodologies used in OR is the development of quantitative models which relate controllable and uncontrollable variables to system performance measures; and (*c*) the outputs of OR models are solutions, that is, suggested levels of control variables that meet some prescribed restrictions. In addition, OR attempts to identify those solutions that are ‘better’ than others or are ‘best’ given an objective function and the validity of solutions ought to be tested empirically. OR, however, is concerned not only with the derivation of solutions but their relevance to management practice and their implementation.

While OR definitions reflect the ideals and aspirations of many leaders of OR communities, a commonly held view is that OR is a collection of techniques (National Academy of Sciences, 1976). In fact, the

success in practice of some OR techniques, such as linear programming and the proliferation of accessible optimization packages, is responsible for this perception. In part, it is also a reflection of imbalances in the work of OR academics. Examination of the content of OR journals and textbooks, for example, would support such a proposition. Indeed, much of the academic effort since the mid-1980s has focused on articulating the supporting mathematical theories of OR models, the development of alternative models and computational methods with a glaring absence of empirical testing (Denizel, Usdiken and Tuncalp, 2003). This, however, was more a result of a natural progression of the life cycle of the field than a paradigm shift.

Disagreements in the OR communities exist with regard to the following questions. First, what is the level of generality that OR models can attain (that is, what are the prospects of OR becoming a science of operations as opposed to an approach to problem-solving in specific organizational contexts)? Second, what is the degree of comprehensiveness of OR missions, in particular the degree to which a systems approach should characterize OR activities (that is, focus of OR methodologies upon overall effects of a proposed solution on an organization rather than a narrower problem-solving focus)? Third, what is the role of interdisciplinary teamwork in OR?

OR models and techniques

Models and computational techniques are key elements in the OR methodology. Models in OR, as opposed to models developed by mathematicians, derive their legitimacy from the real world (as in other sciences) and from their potential uses. Thus one can classify OR models and techniques according to the type of management problems or decision areas they deal with.

Some of the characteristic problem areas that have stimulated OR modeling include:

- Allocation problems
- Inventory problems
- Queuing problems
- Scheduling problems
- Competitive problems
- Renewal and replacement problems
- Search problems
- Revenue management
- Supply chain management
- Financial and marketing engineering
- Data mining

Each of these problem areas is characterized by some typical structures which have stimulated the development of certain classes of mathematical models as well as their supporting mathematical theories. Often, however, a type of mathematical model developed for a specific problem area can be used to model processes with similar structures in other problem areas.

Let us consider, for example, allocation problems. These are the typical economic problems of allocating scarce resources between competing demands so as to maximize net benefits. The allocation, however,

must satisfy some prescribed constraints. The first primitive mathematical programs were formulated by economists late in the 18th century. The typical structure of mathematical programs is the maximization or minimization of an objective function subject to a set of constraints. The properties of the objective function and the special structures of the constraints determine the methods and difficulty of finding optimal values. For example, *linear programming* postulates a system with a linear objective function, linear constraints and non-negative control variables. Thus sub-areas of mathematical programming designate the mathematical structure of the optimization problem at hand: *integer programming* requires integer solutions; *quadratic programming* postulates a quadratic objective function; *stochastic programming* assumes that stochastic parameters describe the objective function; *chance-constrained programming* assumes that the restrictions on a problem are given as probabilities of satisfying each constraint, and so on.

An interesting allocation problem arises in situations with multiple decision units with separate conflicting objectives, when rules for trade-offs or reconciliation of conflict are not given. This problem led to the emergence of *multi-criteria programming*, a technique that postulates several objective functions subject to a set of joint constraints.

As we indicated, while mathematical programming emerged as a means of dealing with allocation problems, its applications cut across most areas of OR endeavour.

Queuing theory evolved primarily to help design service policies to deal with congestion and waiting lines. The theory has its roots in probability theory. The application of the theory demonstrates well a problem which characterizes many OR models – limited empirical validity. Indeed, in many practical situations, the probability distributions which characterize arrival and service time depart from those postulated by the basic theory. In such cases problems become analytically intractable and simulation techniques are used.

Inventory and production control theory can be divided into the tractable but unrealistic deterministic cases, and the more problematic stochastic cases. The theory has contributed important insights as to the shape of optimal policies, but specific solutions to problems arising in the real world are typically obtained by simulation.

Simulation is indeed the most prolific OR technique. It is used in practice especially to model stochastic processes and provide solutions to analytically difficult or intractable problems. A computer model representing the system provides the vehicle for low-cost, fast experimentation with alternative patterns of control variables.

Competitive problems have led to the emergence of *game theory*. While the theory has had some important applications (for example, designing optimal stable policies of inspections associated with international nuclear-testing restrictions), its restrictive assumptions with respect to the rationality of players have limited its usefulness for modelling many competitive business situations. *Gaming* and simulation techniques are often used to improve strategic decisions in competitive situations.

Scheduling problems are typically modelled as *network flow optimization* problems. Two techniques have received great attention and have been employed widely in project planning: the program evaluation and review technique (PERT) and the critical path method (CPM). Network flow optimization is used extensively to deal with many transportation and communication problems.

An important area of OR modelling is the area of Markov and related processes. In a Markov process, knowledge of the present makes the future independent of the past. Markov chains have been used extensively in manpower planning.

Dynamic programming is a method of analysing multi-stage decision processes in which each decision in a sequence depends upon those preceding it as well as exogenous factors. The technique reduces significantly the computational effort by eliminating the need to enumerate and consider the consequences of all possible decision sequences. The method is used in a wide variety of problem areas. In the 1980s *revenue management methods* combining accurate demand forecasts with intelligent dynamic pricing were developed for and adopted by airlines, and their use spread to other sectors. In the 1990s increasing globalization and the emergence of complex business networks of suppliers and producers created the need for better *supply chain management*. Advances in computing power, communications and operation research methods created new modelling opportunities responding to the challenge of finding best overall combinations of suppliers, transportation, production, warehousing and inventory. Recent modelling efforts in this domain incorporate ‘game like’ situations in cooperative networks where incentives of different participants may be misaligned. The late 1990s saw the emergence of e-commerce and powerful information technology applications in business generating high volumes of customer data. Large, high-quality data stimulated the development of *data mining* techniques to use the data to improve business strategies and operations. The proliferation of personal powerful computers created opportunities for the development of OR applications for a variety of business functions. *Financial and marketing engineering* are examples of OR applications to traditional functional fields of business.

The lack of definite boundaries as to what constitutes OR makes it difficult to determine whether some techniques originating in other fields, but frequently used by OR practitioners, should be designated as OR techniques. Statistical analysis, forecasting methodologies and evaluation techniques are good examples.

The roots of OR

The beginnings of OR can be traced to the emergence of the executive function and the complex organization brought about by the Industrial Revolution of the 19th century. The mathematical roots of OR can be traced earlier to the work of Quesnay (1759), who formulated primitive mathematical programming models. This fundamental work was followed by the work of Walras (1883), and by the work of von Neumann (1937) and Kantorovich (1939).

The roots of empirical OR can be traced to the scientific management movement. The work of Taylor, Gantt, Emerson and other pioneers of scientific management began around 1885. They proposed that scientific methods of analysis and measurement could and should be used in production management and business decisions. In 1909, Erlang, a Danish mathematician, published his study of traffic congestion in a telephone network, pioneering the modelling of queues. In 1916 Lanchester published his ‘N-square law’, assessing the fighting power of opposing forces. The theory was tested retrospectively against Admiral Nelson's plan of the battle of Trafalgar.

The appearance of OR as an organized activity is associated with preparation in the UK for the Second World War. In 1936 the British government decided to set up radar stations. The need to study the operational use of radar chains in order to increase their ability to detect aircraft led to the establishment of a study group of scientists called ‘the operational research group’. Their success led to the adoption of OR by other branches of the military. In 1942 an OR section was established by the US Air Force. OR was soon adopted by other branches of the US military. Under the aegis of the US Air Force, a team of

economists and mathematicians began in 1947 to model the military structure and the economy. During this period Dantzig (1963) developed the simplex method of solving linear programs.

The evolution of OR

The diffusion of OR to the industrial world was slow. Only in the early 1950s did the tools and methods of OR begin to be used outside the military. The first important industrial application of OR was the use of linear programming to schedule a petroleum refinery (Charnes and Cooper, 1961).

The Operational Research Club of Britain was formed in 1948, and the Operations Research Society of America (ORSA) was established in 1951. Other national societies for OR soon followed, and in 1957 the International Federation of Operational Research Societies was formed. Books, journals and university programmes specializing in OR proliferated in the 1960s. A gradual process of change in the membership of most OR communities started, bringing a shift towards a higher proportion of university-based members. While the ORSA constitution saw as one of its major missions the establishment and maintenance of professional standards of competence in OR, the evolution of the field caused more emphasis to be placed upon the academic mission of the development of methods and techniques of OR. The tension between practice and theory of OR indeed originated in the 1950s and 1960s. It is interesting to note that this period is viewed by some as the best of times for OR (Miser, 1978) and by others as the worst of times (Churchman, 1979).

The period saw some of the most exciting mathematical developments since the simplex algorithm. Examples are the important paper by Kuhn and Tucker (1951) laying the foundations of nonlinear programming; the paper by Gomory (1958) presenting a systematic computational technique for integer programming; the works of Bellman (1957) developing dynamic programming; the seminal book by Ford Jr and Fulkerson (1962) articulating network flow optimization; and the volume edited by Arrow, Karlin and Scarf (1958) on the mathematical theory of inventory and production processes. Other important developments during the period were the articulation of decision analysis (see, for example, Raiffa, 1968), the development of stochastic programming and chance-constrained programming (see, for example, Charnes and Cooper, 1959) and the development of the dual method (Lemke, 1954) and the linear complementarity algorithm (Lemke, 1965).

Yet, despite these developments, Churchman (1979, p. 13) called the period ‘dreary’, lamenting the separation of theoretical developments from application, describing OR modelling as a ‘study of the delights of algorithms; nuances of game theory; fascinating but irrelevant things that can happen in queues’.

The 1970s presented OR with an important mathematical theory – a theory focusing on its bounds rather than promises: the theory of NP-completeness. The theory presents a framework for the identification of bounds on computational efficiencies (Cook, 1971; Karp, 1972). Important breakthroughs in the early 1980s were associated with possible improvements on the simplex algorithm in solving linear programs – the development of polynomial algorithms by Khachian (1979; 1980) and Karmarkar (1984).

The 1980s also saw a breakthrough development in the inventory management field. Roundy (1985; 1986) found a simple heuristic and proved that it yields schedules within two per cent of the optimal solution; this work anticipated also the coordination problems characterizing supply chain management.

The 1990s saw articulation of the general theory of supermodularity and lattice programming pioneered by Veinott, Edmonds and Topkis (see Topkis, 1998). The theory provides fundamental insights to certain classes of optimization problems and issues related to monotone comparative statics, fundamental in economic analysis. The development of polynomial submodular set functions – an unresolved problem remaining – was solved simultaneously by Iwata, Fleischer and Fujishige and Schrijver (see Fleischer, 2000).

The new millennium also saw a breakthrough in graph theory – the characterization of the strong perfect graphs by Chudnovsky, Robertson, Seymour and Thomas (see Cornuéjols, 2003).

Perhaps more important to the future of operations research has been the great progress achieved since the mid-1990s in computational methods. Advances in computing machinery, software improvements and development combined to increase the practical significance of the various OR methods. The increased speed of computation and the huge increases in computer memory capacity have made it possible to solve much larger problems and use entirely different solution strategies (Bixby, 2002).

Improved software also allowed also better interface with users, increasing the accessibility of OR methods to a wider population of users.

The scope of OR was enlarged while the cohesiveness of its communities reduced. Fragmentation was identified by many as an explanation of the declining memberships of many OR and management science professional societies. OR appeared to some observers to be ‘in danger of losing its identity as a recognized activity and being assimilated into other fields of endeavor’ (Bonder, 1979, p. 218). Thus, while the power of OR methods and their use increased, the period since the mid-1980s has witnessed some trends which are threatening the identity of OR as a distinct profession.

The future of OR

The apparent divorce of OR theory from practice and empirical testing led some leaders in the OR community to wonder whether ‘the future of OR is past’. The microcomputer revolution has increased the benefit–costs ratios of OR methods and increased the direct access of general business users to OR. OR groups and practitioners, however, have lost some of their unique advantages as gatekeepers to the application of OR methods. Much of the diffusion of OR methods to the industry is now accomplished through the sales of packaged programs, and is marketed through demonstration CDs. Many users of OR methods in business do not consider themselves OR practitioners. Thus, the dispersion of OR practice in business has resulted in a loss of professional identity (Geoffrion, 1992).

Loss of professional identity reduces the flow of new recruits to the profession and limits the career opportunities of OR professionals. The success of OR methods may, therefore, entail the decline of the profession. The sustainability and health of the profession depends on its ability to adopt new business models that fit the new environment, turning threats to opportunities for growth.

See Also

- computer science and game theory
- convex programming

- graph theory
- linear programming

Bibliography

Arrow, K., Karlin, S. and Scarf, H. 1958. *Studies in the Mathematical Theory of Inventory and Production*. Stanford: Stanford University Press.

Bellman, R. 1957. *Dynamic Programming*. Princeton: Princeton University Press.

Bixby, R. 2002. Solving real-world linear programs: a decade and more of progress. *Operations Research* 50, 3–15.

Bonder, S. 1979. Changing the future of operations research. *Operations Research* 27, 209–24.

Charnes, A. and Cooper, W. 1959. Chance-constrained programming. *Management Science* 6, 73–9.

Charnes, A. and Cooper, W. 1961. *Management Models and Industrial Applications of Linear Programming*, vols. 1 and 2, New York: Wiley.

Churchman, C. 1979. Paradise regained: a hope for the future of systems design education. In *Education in Systems Science*, ed. B. Bayraktar. London: Taylor and Francis.

Churchman, C., Ackoff, R. and Arnoff, E. 1957. *Introduction to Operations Research*. New York: Wiley.

Cook, S. 1971. The complexity of theorem-proving procedures. *Proceedings of the Association for Computing Machinery Annual Symposium on the Theory of Computing* 3, 151–8.

Cornuéjols, G. 2003. The strong perfect graph theorem. *Optima: The Mathematical Programming Society Newsletter* 70, 2–6 June.

Dando, M.R. and Sharp, R. 1978. Operational research in the UK in 1977: the cases and consequences of a myth? *Journal of the Operational Research Society* 29, 939–49.

Dantzig, G. 1963. *Linear Programming and Extensions*. Princeton: Princeton University Press.

Denizel, M., Usdiken, B. and Tuncalp, D. 2003. Drift or shift? Continuity, change, and international variation in knowledge. *Operations Research* 51, 711–20.

Fleischer, L. 2000. Recent progress in submodular function minimization. *Optima: The Mathematical Programming Society Newsletter* 64, 1–11 September.

Ford, L., Jr. and Fulkerson, D. 1962. *Flows in Networks*. Princeton: Princeton University Press.

Geoffrion, A. 1992. Forces, trends and opportunities in MS/OR. *Operations Research* 40, 423–45.

Gomory, R. 1958. Essentials of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society* 64, 275–8.

House, A. 1952. The founding meeting of the society. *Journal of the Operations Research Society of America* 1, 18–32.

INFORMS (Institute for Operations Research and Management Science). About operations research. Online. Available at http://www.informs.org/index.php?c=49&kat=+About+Operations +Research&p=17|, accessed 22 August 2006.

Kantorovich, L. 1939. *Mathematical Methods of Organising and Planning Production*. Leningrad University [in Russian]. Trans. R. Campbell and W. Marlow, *Management Science* 6(1960), 366–422.

Karmarkar, N. 1984. A new polynomial time algorithm for linear programming. *Combinatorica* 4, 373–95.

Karp, R. 1972. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, ed. R. Miller and J. Thatcher. New York: Plenum Press.

Khachian, L. 1979. A polynomial algorithm in linear programming. *Doklady Akademii Nauk SSSR* 224, 1093–6.

Khachian, L. 1980. Polynomial algorithms in linear programming. *Zhurnal vychisditel'noi matematiki i matematicheskoi* 20, 51–68.

Kuhn, H. and Tucker, A. 1951. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Lemke, C. 1954. The dual method of solving the linear programming problem. *Naval Research Logistic Quarterly* 1, 36–47.

Lemke, C. 1965. Bimatrix equilibrium points and mathematical programming. *Management Science* 11, 681–9.

Little, J. 1986. Research opportunities in the decision and management sciences. *Management Science* 32

(1), 1–13.

Miser, H. 1978. The history, nature and use of operations research. In *Handbook of Operations Research*, vol. 1, ed. J. Moder and S. Elmaghraby. New York: Van Nostrand Reinhold.

National Academy of Sciences. 1976. *Systems Analysis and Operations Research: A Tool for Policy and Program Planning for Developing Countries*. Report of an ad hoc panel. Washington, DC: National Academy of Sciences.

Pocock, J. 1956. Operations research: a challenge to management. In *Operations Research*. Special Report No. 13. New York: American Management Association.

Quesnay, F. 1759. *Tableau économique*, 3rd edn. Paris. Ed. M. Kuczynski and R. Meek. London: Macmillan, 1972.

Raiffa, H. 1968. *Decision Analysis: Introductory Lectures on Choices and Uncertainty*. New York: Addison-Wesley.

Roundy, R. 1985. Effective integer ratio lot-sizing for one warehouse multi-retailer systems. *Management Science* 31, 1416–30.

Roundy, R. 1986. Effective lot-sizing rule for a multi-product multi stage production/inventory system. *Mathematics of Operations Research* 11, 699–727.

The OR Society. What operational research is. Online. Available at http://www.orsoc.org.uk/orshop/(aamlgbmxg1m44xb520nsqw45)/orcontent.aspx?inc=about.htm, accessed 22 August 2006.

Topkis, D.M. 1998. *Supermodularity and Complementarity*. Princeton: Princeton University Press.

von Neumann, J. 1937. A model of general economic equilibrium. In *Ergebnisse eines mathematischen Kolloquiums* 8, ed. K. Menger [in German]. Trans. as ‘A model of general equilibrium’, *Review of Economic Studies* 13 (1945–6), 1–9.

Walras, L. 1883. *Théorie mathématique de la richesse sociale*. Lausanne: Corbaz.

Wolfe, P. 1959. The simplex method for quadratic programming. *Econometrica* 27, 382–98.

How to cite this article

Vertinsky, Ilan. "operations research." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000026> doi:10.1057/9780230226203.1221

The New Palgrave Dictionary of Economics Online

opportunity cost

James M. Buchanan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

choice; opportunity cost; scarcity

Article

The concept of *opportunity cost* (or alternative cost) expresses the basic relationship between scarcity and choice. If no object or activity that is valued by anyone is scarce, all demands for all persons and in all periods can be satisfied. There is no need to choose among separately valued options; there is no need for social coordination processes that will effectively determine which demands have priority. In this fantasized setting without scarcity, there are no opportunities or alternatives that are missed, forgone, or sacrificed.

Once scarcity is introduced, all demands cannot be met. Unless there are ‘natural’ constraints that predetermine the allocation of end-objects possessing value (for example, sunshine in Scotland in February), scarcity introduces the necessity of choice, either directly among alternative end-objects or indirectly among institutions or procedural arrangements for social interaction that will, in turn, generate a selection of ultimate end-objects.

Choice implies rejected as well as selected alternatives. *Opportunity cost is the evaluation placed on the most highly valued of the rejected alternatives or opportunities*. It is that value that is given up or sacrificed in order to secure the higher value that selection of the chosen object embodies.

Opportunity cost and choice

Opportunity cost is the anticipated value of ‘that which might be’ if choice were made differently. Note that it is not the value of ‘that which might have been’ without the qualifying reference to choice. In the absence of choice, it may be sometimes meaningful to discuss values of events that might have occurred but did not. It is not meaningful to define these values as opportunity costs, since the alternative scenario does not represent a lost or sacrificed opportunity. Once this basic relationship between choice and opportunity cost is acknowledged, several implications follow.

First, if choice is made among separately valued options, someone must do the choosing. That is to say, a chooser is required, a person who decides. From this the second implication emerges. The value placed on the option that is not chosen, the opportunity cost, must be that value that exists in the mind of the individual who chooses. It can find no other location. Hence, cost must be borne exclusively by the chooser; it can be shifted to no one else. A third necessary consequence is that opportunity cost must be subjective. It is within the mind of the chooser, and it cannot be objectified or measured by anyone external to the chooser. It cannot be readily translated into a resource, commodity, or money dimension. Fourth, opportunity cost exists only at the moment of decision when choice is made. It vanishes immediately thereafter. From this it follows that cost can never be realized; that which is rejected can never be enjoyed.

The most important consequence of the relationship between choice and opportunity cost is the *ex ante* or forward-looking property that cost must carry in this setting. Opportunity cost, the value placed on the rejected option by the chooser, is the obstacle to choice; it is that which must be considered, evaluated, and ultimately rejected before the preferred option is chosen. Opportunity cost in any particular choice is, of course, influenced by prior choices that have been made, but, with respect to this choice itself, opportunity cost is *choice-influencing* rather than *choice-influenced*.

Other notions of cost

The distinction between opportunity cost and other conceptions or notions of cost is best explained in this choice-influencing and choice-influenced classification. Once a choice is made, consequences follow, and these consequences may, indeed, involve utility losses, either to the person who has made initial choice or to others. In a certain sense it may seem useful to refer to these losses, whether anticipated or realized, as costs, but it must be recognized that these choice-determined costs, as such, cannot, by definition, influence choice itself.

A single example may clarify this point. A person chooses to purchase an automobile through an instalment loan payment plan, extending over a three-year period. The opportunity cost that informs and influences the choice is the value that the purchaser places on the rejected alternative, in that case the anticipated value of the objects which might be purchased with the payments required under the loan. Having considered the potential value of this alternative, and chosen to proceed with the purchase, the consequences of meeting the loan schedule follow. Monthly payments must be made, and it is common language usage to refer to these payments as ‘costs’ of the automobile. The individual will clearly suffer a sense of utility loss as the payments come due and must be paid. As choice-influencing elements, however, these ‘costs’ are irrelevant. The fact that, in a utility dimension, post-choice consequences can never be capitalized is a source of major confusion.

Economists recognize the distinction being made here in one sense. With the familiar statement that ‘sunk costs are irrelevant’, economists acknowledge that the consequences of choices cannot influence choice itself. On the other hand, by their formalized constructions of cost schedules and cost functions, which necessarily imply measurability and objectifiability of costs, economists divorce cost from the choice process.

Essentially the same results hold for accountants, who normally measure estimated costs strictly in the *ex post* or choice-influenced sense. Those ‘costs’ estimated by accountants can never accurately reflect the value of lost or sacrificed opportunities. Numerical estimates could be introduced in working plans

for alternative courses of action prior to decision, but such estimates of opportunity costs would be the accountant's measure of the values for projects not undertaken rather than the value of commitments made under the project chosen.

As suggested, choice-influencing opportunity costs exist only for the person who makes choice. By definition, opportunity costs cannot ‘spill over’ to others. There may, of course, be consequences of a person's choice that impose utility losses on other persons, and it is sometime useful to refer to these losses as ‘external costs’. The point to be emphasized is that these external costs are obstacles to choice, and hence a measure of forgone opportunities, only if the individual who chooses takes them into account and places his own anticipated utility evaluation on them.

Opportunity cost and welfare norms

The source of greatest confusion in the analysis and application of opportunity cost theory lies in the attempted extension of the results of idealized market interaction processes to the definition of rules or norms for decision makers in non-market settings. In full market equilibrium, the separate choices made by many buyers and sellers generate results that may be formally described in terms of relationships between prices and costs. Under certain specified conditions, prices are brought into equality with marginal costs through the working of the competitive process. Further, the general equilibrium states described by these equalities are shown to meet certain efficiency norms.

Prices may be observed; they are objectively measurable. A condition for market equilibrium is equalization of prices over all relevant exchanges for all units of a commodity of service. From this equalization it may seem to follow that marginal costs, which must be brought into equality with price as a condition for the equilibrium of each trader, are also objectively measurable. From this the inference is drawn that, if marginal costs are then measured, ‘efficiency’ in resource use can be established independently of the competitive process itself through the device of forcing decision makers to bring prices into equality with marginal costs.

The whole logic is a tissue of confusion based on a misunderstanding of opportunity cost. The equalization of marginal opportunity cost with price for each trader is brought about by the adjustments made by each trader along the relevant quantity dimension. The fact that the marginal opportunity costs for all traders are all brought into equalization with the relevant uniform price implies only that traders retain the ability to adjust quantities of goods until this condition is met. There is no implication to the effect that marginal opportunity costs are equalized in some objectively meaningful sense independently of the quantity adjustment to price.

Consider an idealized market for a good that is observed to be trading at a uniform price of $1 per unit. The numeraire value of the anticipated lost opportunity is $1 for each trader. But it is only as quantity is adjusted that the trader can bring the numeraire value of his subjectively experienced and anticipated utility sacrifice into equality with the objectively set price that he confronts. The anticipated value of that which is given up in taking a course of action is no more objectifiable and measurable than the anticipated value of the course of action itself. The two sides of choice are equivalent in all respects. Independently of market choice, there is no means through which marginal opportunity costs can be brought into equality with prices. Hence, any ‘rule’ that directs ‘managers’ in non-market settings to use cost as the basis for setting price is and must remain without content. There is, however, a second equally important criticism of the welfare rule that opportunity cost reasoning identifies, quite apart from

the measurability question. Even if the first criticism is ignored, and it is assumed that marginal opportunity cost can, in some fashion, be measured, instructions to ‘managers’ to use cost to set price must rely on ‘managers’ to behave, personally, as robots rather than rational utility-maximizing individuals. Why should a ‘manager’ be expected to follow the rule? Would he not be expected to behave so that marginal cost, that which he faces personally, be brought into equality with the anticipated value of the benefit side of choice? The fact that the ‘manager’ remains in a non-market setting insures that he cannot be the responsible bearer of the utility gains and losses that his choices generate. His own, privately sensed, gains and losses, evaluated either prior to or after choice, must be categorically different from those anticipated for principals before choice and enjoyed and/or suffered by principals after choice.

Opportunity cost and the choice among institutions

As noted earlier, in the absence of ‘natural’ constraints that predetermine allocation, the introduction of scarcity introduces the necessity of choice, either directly among ultimate ‘goods’ or indirectly among rules, institutions, and procedures that will operate so as to make final allocative determinations. Opportunity cost in the second of these choice-settings remains to be examined. In a sense, the use of institutionalized procedures to generate allocations of scarce resources may eliminate ‘choice’ in the familiar meaning used above and is akin in this respect to the ‘natural’ constraints noted. Results may emerge from the operation of some institutional process without any person or group of persons ‘choosing’ among end-state alternatives, and, hence, without any subjectively-experienced opportunity cost. Despite the absence of this important bridge between cost and choice in the ordinary sense, however, values may be placed on the ‘might have beens’ that would have emerged under differing allocations. The patterns of these estimated value losses, over a sequence of institution-determined allocations, may enter, importantly, in a rational choice calculus involving the higher-level choice among alternative institutional procedures for allocation. In this higher-level choice, opportunity cost again appears as the negative side of choice even if ‘choice’ in the standard usage of the term is not involved in the making of allocations, taken singly.

Consider the following extreme example. There are two mutually exclusive thermostat settings for a building, *High* and *Low*. An institution is in being that uses an unbiased coin to ‘choose’ between these two settings each day. It is meaningful for an individual to discuss the potential value to be anticipated if the setting is *High* rather than *Low*, even if the individual does not make the selection, individually or as a member of a collective. The setting that is ‘chosen’ by the coin flip has consequences for individual utility and these consequences may be anticipated in advance of the actual ‘choice’. So long as the institutional procedure remains in effect, however, with respect to a single day's selection, the anticipated value lost by one setting of the thermostat rather than the other cannot represent opportunity cost.

Suppose, now, that instead of the unbiased and equally weighted device, the institution in being is one that allows all persons in the building to vote, each morning, on the thermostat setting with the majority option ‘chosen’ for the day. Assume, further, that the group of voters is large, so that the influence of a single person on the expected majoritarian outcome is quite small. It is important to emphasize that, in this procedure, as with the coin toss, no person really ‘chooses’ among the alternative end-states. Each voter confronts the quite different, intra-institutional choice between ‘voting for High’ and ‘voting for

Low', with the knowledge that any individual has relatively little influence on the outcome. In the choice that he confronts, the voter cannot rationally take into account the anticipated losses from the ultimate alternatives, either for himself or for others, in any full-value sense of the term. The loss anticipated from, say, a Low thermostat setting may be estimated to be valued at $1,000 for the individual. Yet if he considers himself to have an influence on the outcome of the voting choice only in one case out of a thousand, the expected utility value of the anticipated loss will be only $1 in terms of the numeraire. This $1 will then represent the numeraire value of the *opportunity cost* involved in voting for High.

Since these same results hold, with possibly differing values, for all voters, no one 'chooses' in accordance with fully evaluated gains and losses. 'Choices' emerge from the institutional procedure without full benefit – cost considerations being made by anyone, taken singly or in aggregation. In the relevant opportunity-cost sense, effective choice is shifted to that among alternative institutions. The results of the 'choices' made within an institution over a whole sequence of periods (over many days in our thermostat example) may, of course, become data for the choice comparison among institutions themselves. And, to the extent that the individual, when confronted with a choice among institutions, knows that he is individually responsible for the selection, the whole opportunity cost logic then becomes relevant at the level of institutional or constitutional choice. This result is accomplished, however, only if each person in the relevant community does, in fact, become the chooser among institutional rules. Only if, at some ultimate level of institutional-constitutional choice the Wicksellian unanimity rule becomes operative, hence giving any person potential choice authority, can the opportunity cost of alternatives for choice be expected to enter and to inform individual decisions.

Summary

Opportunity cost is a basic concept in economic theory. In its rudimentary definition as the value of opportunities forgone as a result of choice in the presence of scarcity, the concept is simple, straightforward, and widely understood. In the analysis of choices made by buyers and sellers in the marketplace, the complexities that emerge only in rigorous definition of the concept remain relatively unimportant. But when attempts are made to extend opportunity cost logic to non-market settings, either in the derivation of norms to guide decisions or in application to choice within and among institutions, the observed ambiguity and confusion suggest that even so basic a concept requires analytical clarification.

Bibliography

Alchian, A. 1968. Cost. *Encyclopedia of the Social Sciences*, vol. 3. New York: Macmillan.

Buchanan, J.M. 1969. *Cost and Choice*. Chicago: Markham, Republished as Midway Reprint, Chicago: University of Chicago Press, 1977.

Buchanan, J.M. and Thirlby, G.F., eds. 1973. *LSE Essays on Cost*. London: Weidenfeld and Nicholson. Reissued by New York University Press, 1981.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3 (October), 1–44.

How to cite this article

Buchanan, James M. "opportunity cost." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000029> doi:10.1057/9780230226203.1222

The New Palgrave Dictionary of Economics Online

optimal fiscal and monetary policy (with commitment)

Mikhail Golosov and Aleh Tsyvinski

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Optimal fiscal and monetary policy with commitment’ is a policy of choosing taxes and transfers or monetary instruments to maximize social welfare. ‘Commitment’ refers to ability of a policymaker to make binding policy choices.

Keywords

bonds; commitment; disability insurance; Friedman rule; homotheticity; incentive compatibility; inflation targeting; inflation tax; intermediate good; linear taxation; lump sum taxes; Mirrlees, J.; money; nominal interest rates; optimal fiscal and monetary policy with commitment; optimal interest rate; optimal monetary policy; optimal taxation; output gap; partial equilibrium; Ramsey taxation; separability; sticky prices; tax smoothing; taxation of capital income; taxation of consumption; taxation of income; Taylor rule; time preference; uniform commodity taxation

Article

The Ramsey approach to the optimal taxation

‘Ramsey approach to optimal taxation’ is the solution to the problem of choosing optimal taxes and transfers given that only distortionary tax instruments are available.
A starting point of a Ramsey problem is postulating tax instruments. Usually, it is assumed that only linear taxes are allowed. Importantly, lump sum taxation is prohibited. Another assumption crucial to this approach is that all activities of agents are observable.
Given the set taxes, a social planner (government) maximizes its objective function given that agents (firms and consumers) are in a competitive equilibrium. Usually, it is assumed that government's objective is to finance an exogenously given level of expenditures. It is important to note that if the lump sum taxes were allowed than the first welfare theorem would hold, and the unconstrained optimum

would be achieved.

There are two common approaches to solving Ramsey problems. The first is the *primal* approach, which characterizes a set of allocations that can be implemented as a competitive equilibrium with taxes. By ‘implementation we mean’ the following: for a set of taxes find a set of (consumption and labour) allocations and equilibrium prices such that these allocations are a competitive equilibrium given taxes. Conversely, a set of (consumption and labour) allocations is implementable if it is possible to find taxes and equilibrium prices such that these allocations are a competitive equilibrium given these prices and taxes. Implementation often makes it possible to simplify a Ramsey problem by reformulating a problem of finding optimal taxes as the problem of finding implementable allocations. This reformulation is referred to as the *primal approach* to Ramsey taxation.

Main lessons of Ramsey taxation: uniform commodity taxation, zero capital tax in the long run, and tax smoothing

One of the central results of the literature on Ramsey taxation is *uniform commodity taxation* (Atkinson and Stiglitz, 1972). Consider a model with a finite set of consumption goods that can be allocated between government and private consumption. All of these goods are produced with labour. Assume that each consumption good can be taxed at a linear rate. Then, under certain separability and homotheticity assumptions, commodity taxation is uniform, that is, the optimal taxes are equated across consumption goods.

Ramsey taxation provides a compelling argument against taxing capital income in the long run in a model of infinitely lived households. The *Chamley–Judd result* (Chamley 1986; Judd 1985) states that in a steady state there should be no wedge between the intertemporal rate of substitution and the marginal rate of transformation, or, alternatively, that the optimal tax on capital is zero. The intuition for the result is that even a small intertemporal distortion implies increasing taxation of goods in future periods in contrast to the prescription of the uniform commodity taxation. Therefore, distorting the intertemporal margin is very costly for the planner. Jones, Manuelli and Rossi (1997) extend the applicability of the Chamley–Judd result by showing that the return to human capital should not be taxed in the long run. Chari, Christiano and Kehoe (1994) provide the state-of-the art numerical treatment for optimal Ramsey taxation over the business cycle and conclude that the *ex ante* capital tax rate is approximately zero.

There has been a long debate on the optimal composition of taxation and borrowing to finance government expenditures. Barro (1979) considers a partial equilibrium economy and argues that it is optimal to smooth distortions from taxation over time, a policy referred as *tax smoothing*. The implication of this analysis is that optimal taxes should follow a random walk. Lucas and Stokey (1983) consider an optimal policy in a general equilibrium economy without capital, and show that, if government has access to state-contingent bonds, optimal taxes inherit the stochastic process of the shocks to government purchases. Chari, Christiano and Kehoe (1994) extend this analysis to an economy with capital and show the Lucas and Stokey results remain valid in that set-up with or without state contingent debt, as long as the government can use taxes on capital to effectively vary the *ex post* after-tax rate of return on bonds. Finally, Aiyagari et al. (2002) show that, if *ex post* taxation of returns is impossible, the optimal taxes follow a process similar to a random walk. They also show the conditions under which the tax smoothing hypothesis is valid.

The Mirrlees approach to optimal taxation

The Mirrlees approach to optimal taxation is built on a different foundation from Ramsey taxation. Rather than stating an ad hoc restricted set of tax instruments as in Ramsey taxation, Mirrlees (1971) assumed that an informational friction endogenously restricted the set of taxes that implement the optimal allocation. This set-up allows arbitrary nonlinear taxes, including lump-sum taxes.
The informational friction posed in those models is unobservability of agents' skills: only labour income of agents can be observed. Therefore, from a given level of labour income it cannot be determined whether a high-skill agent provides a low amount of labour or effort, or whether a low-skill agent works a prescribed amount. The objective of the social planner (government) is to maximize *ex ante*, before the realization of the shocks, utility of an agent. This objective can be interpreted as either insurance against adverse shocks or as *ex post* redistribution across agents of various skills. An informational friction imposes *incentive compatibility* constraints on the planner's problem: allocations of consumption and effective labour must be selected such that an agent chooses not to misrepresent its type.
In summary, the objective of the Mirrlees approach is to find the optimal incentive–insurance trade-off: how to provide the best insurance against adverse events (low realizations of skills) while providing incentives for the agents to reveal their types (provide high amount of labour).

Main lessons of the Mirrlees approach in a static framework

Theoretical results providing general characterization of the optimal taxes in the static Mirrlees environment are limited. The central result is that the consumption–leisure margin of an agent with the highest skill is undistorted, implying that the marginal income tax at the top of the distribution should be optimally set equal to zero. Saez (2001) is a state-of-the art treatment of the static Mirrlees model in which he derives a link between the optimal tax formulas and elasticities of income. Mirrlees (1971) was also able to establish broad conditions that would ensure that the optimal marginal tax rate on labour income was between zero and 100 per cent.

Main lessons of dynamic Mirrlees literature: distorted intertemporal margin

Recent literature starting with Golosov, Kocherlakota, and Tsyvinski (2003) and Werning (2001) extends the static Mirrlees (1971) framework to dynamic settings. Golosov, Kocherlakota, and Tsyvinski (2003) consider an environment with general dynamic stochastically evolving skills. An example of a large unobservable skill shock is disability that is often difficult to observe (classical example is back pain or mental illness). Golosov, Kocherlakota, and Tsyvinski (2003) show for arbitrary evolution of skills that, as long as the probability of agent's skill changing is positive, any optimal allocation includes a positive intertemporal wedge: a marginal rate of substitution across periods is lower than marginal rate of transformation. The reason for this is that this wedge improves the intertemporal provision of incentives by implicitly discouraging savings. This result holds even away from the steady state and sharply contrasts with the Chamley–Judd result that stems from the exogenous restriction on tax

instruments. Golosov, Kocherlakota, and Tsyvinski (2003) and Werning (2001) show that in a case of constant types a version of uniform commodity taxation holds and the intertemporal margin is not distorted.

Implementation of dynamic Mirrlees models is more complicated than implementation of either static Mirrlees models, which are implemented with an income tax, or Ramsey models of linear taxation. By ‘implementation’ we mean finding tax instruments such that the optimal allocation is a competitive equilibrium with taxes. One possible implementation is a direct mechanism that mandates consumption and labour menus for each date. However, such a mechanism can include taxes and transfers never used in practice. Three types of implementations have been proposed. In Albanesi and Sleet (2006), wealth summarizes agents’ past histories of shocks that are assumed to be i.i.d. and allows us to define a recursive tax system that depends only on current wealth and effective labour. Golosov and Tsyvinski (2006) implement an optimal disability insurance system with asset-tested transfers that are paid to agents with wealth below a certain limit. Kocherlakota (2005) allows for a general process for skill shocks and derives an implementation with linear taxes on wealth and arbitrarily nonlinear taxes on the history of effective labour.

Optimal monetary policy

The theory of the optimal monetary policy is closely related to the theory of optimal taxation. Phelps (1973) argues that the inflation tax is similar to any other tax, and therefore should be used to finance government expenditures. Although intuitively appealing, this argument is misleading. Chari, Christiano and Kehoe (1996) extend the Ramsey approach to analyse optimal fiscal and monetary policy jointly in several monetary models, and find that typically it is optimal to set the nominal interest rate to be equal to zero. Such a policy is called a ‘Friedman rule’, after Milton Friedman, who was one of the first proponents of zero nominal interest rates (Friedman, 1969). To understand intuition for the optimality of Friedman rule, it is useful to think about the distinctive the features that distinguish money from other goods and assets. In most models money plays a special role of providing liquidity services to households that cannot be obtained by using other assets such as bonds. Inefficiency arises if the rates of return on bonds and money are different, since by holding money balances households lose the interest rate. When a nominal interest rate is equal to zero, which in a deterministic economy implies that inflation is negative, with nominal prices declining with the rate of households’ time preferences, the real rates of return on money and bonds are equalized and this inefficiency is eliminated.

The optimality of the Friedman rule stands in a direct contrast with Phelps’ arguments for use of the inflationary tax together with other distortionary taxes such as taxes on consumption or labour income. The reason for this is that money, unlike consumption or leisure, is not valued by households directly but only indirectly, as long as it facilitates transactions and provides liquidity. Therefore, it is more appropriate to think of money as an intermediate good in acquiring final goods consumed by households. Diamond and Mirrlees (1971) established very general results about the undesirability of distortion of the intermediate goods sector, which in monetary models implies that the inflationary tax should not be used despite the distortions caused by taxes on the final goods and services.

The intuition developed above is valid under the assumption that nominal prices are fully flexible, and firms adjust to them immediately in response to changes in market conditions. However, even casual

observation suggests that many prices remain unchanged over long periods of time, and Bils and Klenow (2004) document inflexibility of prices for a wide variety of goods. Inflexible or *sticky prices* lead to additional inefficiencies in the economy that could be mitigated by monetary policy. For example, an economy-wide shock, such as an aggregate productivity shock or change in government spending, may call for readjustment of real prices. If adjustment of nominal prices is sluggish, the central bank can increase welfare by adjusting nominal interest rates and affecting real prices.

It is important to recognize that the government is also able to affect real (after-tax) prices using fiscal instruments instead. In fact, Correia, Nicolini and Teles (2002) show that, if fiscal policy is sufficiently flexible and can respond to aggregate shocks quickly, then the Friedman rule continues to be optimal even with sticky prices, with fiscal instruments being preferred to monetary ones. In current practice, however, it appears that it takes a long time to enact changes in tax rates, while monetary policy can be adjusted quickly. Schmitt-Grohe and Uribe (2004) show that, as long as tax levels are fixed or the government is not able to levy some of the taxes on goods or firms' profits, then the optimal interest rate is positive and variable.

Most of the applied literature on the monetary policy is based on the joint assumption of sticky prices and inflexible fiscal policy. Woodford (2003) provides a comprehensive study of the optimal policy in such settings. This analysis examines how central bank response should depend on the type of the shock affecting the economy, the degree of additional imperfections in the economy, and the choice of policies that would rule out indeterminacy of equilibria. Two common policy recommendations for central banks share many of the features of the optimal policy responses in this analysis. One of such recommendations – a *Taylor rule* (see Taylor, 1993) – calls for the interest rates to be increased in response to an increase in the output gap (the difference between actual and a target level of GDP) or inflation. Another recommendation, *inflation forecast targeting*, requires that the central bank commits to adjust interest rate to ensure that the projected future path of inflation or other target variables does not deviate from the pre-specified targets.

In addition to the analysis set out above, several new, conceptually different approaches to the analysis of monetary policy have emerged in the recent years. For example, da Costa and Werning (2005) re-examine optimal monetary policy with flexible prices in Mirrleesian settings and confirm the optimality of the Friedman rule there. Seminal work by Kiyotaki and Wright (1989) has given rise to a large search-theoretic literature seeking to understand the fundamental reasons that money differs from other goods and assets in the economy. Lagos and Wright (2005) provide a framework for the analysis of optimal monetary policy in such settings.

See Also

- income taxation and optimal policies
- monetary and fiscal policy overview
- new Keynesian macroeconomics
- optimal fiscal and monetary policy (without commitment)
- optimal taxation
- Taylor rules

Bibliography

Aiyagari, S., Marcet, A., Sargent, T. and Seppala, J. 2002. Optimal taxation without state-contingent debt. *Journal of Political Economy* 110, 1220–54.

Albanesi, S. and Sleet, C. 2006. Dynamic optimal taxation with private information. *Review of Economic Studies* 73, 1–30.

Atkinson, A. and Stiglitz, J. 1972. The structure of indirect taxation and economic efficiency. *Journal of Public Economics* 1, 97–119.

Barro, R. 1979. On the determination of the public debt. *Journal of Political Economy* 87, 940–71.

Bils, M. and Klenow, P. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112, 947–85.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54, 607–22.

Chari, V., Christiano, L. and Kehoe, P. 1994. Optimal fiscal policy in a business cycle model. *Journal of Political Economy* 102, 617–52.

Chari, V., Christiano, L. and Kehoe, P. 1996. Optimality of the Friedman rule in economies with distorting taxes. *Journal of Monetary Economics* 37, 203–23.

Correia, I., Nicolini, J.-P. and Teles, P. 2002. Optimal fiscal and monetary policy: equivalence results. Working Paper No. WP-02-16, Federal Reserve Bank of Chicago.

da Costa, C. and Werning, I. 2005. On the optimality of the Friedman rule with heterogeneous agents and non-linear income taxation. Working paper, MIT.

Diamond, P. and Mirrlees, J. 1971. Optimal taxation and public production I: production efficiency. *American Economic Review* 61, 8–27.

Friedman, M. 1969. The optimum quantity of money. In *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

Golosov, M., Kocherlakota, N. and Tsyvinski, A. 2003. Optimal indirect and capital taxation. *Review of Economic Studies* 70, 569–87.

Golosov, M. and Tsyvinski, A. 2006. Designing optimal disability insurance: a case for asset testing. *Journal of Political Economy* 114, 257–79.

Jones, L., Manuelli, R. and Rossi, P. 1997. On the optimal taxation of capital income. *Journal of Economic Theory* 73, 93–117.

Judd, Kenneth L. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 28, 59–83.

Kiyotaki, N. and Wright, R. 1989. On money as a medium of exchange. *Journal of Political Economy* 97, 927–54.

Kocherlakota, N. 2005. Zero expected wealth taxes: a Mirrlees approach to dynamic optimal taxation. *Econometrica* 73, 1587–622.

Lagos, R. and Wright, R. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 113, 463–84.

Lucas, R. and Stokey, N. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12, 55–93.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Phelps, E. 1973. Inflation in the theory of public finance. *Swedish Journal of Economics* 75, 67–82.

Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68, 205–29.

Schmitt-Grohe, S. and Uribe, M. 2004. Optimal fiscal and monetary policy under sticky prices. *Journal of Economic Theory* 114, 183–209.

Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series in Public Policy* 39, 195–214.

Werning, I. 2001. Optimal unemployment insurance with hidden savings. Mimeo, University of Chicago.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Golosov, Mikhail and Aleh Tsyvinski. "optimal fiscal and monetary policy (with commitment)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000100> doi:10.1057/9780230226203.1223

The New Palgrave Dictionary of Economics Online

optimal fiscal and monetary policy (without commitment)

Mikhail Golosov and Aleh Tsyvinski

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Optimal fiscal and monetary policy’ is a policy of choosing taxes and transfers or monetary instruments to maximize social welfare. ‘Absence of commitment’ refers to inability of a policymaker to make binding policy choices.

Keywords

bonds; central bank; commitment; dimensionality; Friedman rule; infinite-horizon models; inflation; inflationary bias; inflationary cap; Markov-perfect equilibria; nominal interest rates; optimal fiscal policy without commitment; optimal monetary policy without commitment; output gap; Ramsey taxation; rational expectations; sustainable equilibrium; taxation of capital; Taylor rule; time consistency of monetary and fiscal policy

Article

Most of the results of optimal taxation literature in the Ramsey framework are derived under the assumption of commitment. *Commitment* is usually defined as ability of a government to bind future policy choices. This assumption is restrictive. A government, even a benevolent one, may choose to change its policies from those promised at an earlier date. The first formalization of the notion of time inconsistency is due to Kydland and Prescott (1977), who showed how timing of government policy may change economic outcomes. Furthermore, equilibrium without commitment can lead to lower welfare for society than when a government can bind its future choices.

An example that clarifies the notion of time inconsistency in fiscal policy is taxation of capital. A classical result due to Chamley (1986) and Judd (1985) states that capital should be taxed at zero in the long run. One of the main assumptions underlying this result is that a government can commit to a sequence of capital taxes. However, a benevolent government will choose to deviate from the prescribed sequence of taxes. The reason is that, once capital is accumulated, it is sunk, and taxing capital is no

longer distortionary. A benevolent government would choose high capital taxes once capital is accumulated.

The reasoning above leads to the necessity of the analysis of time inconsistent policy as a game between a policymaker (government) and a continuum of economic agents (consumers). A formalization of such a game and an equilibrium concept is due to Chari and Kehoe (1990). They formulate a general equilibrium infinite-horizon model in which private agents are competitive, and the government maximizes the welfare of the agents. They define an equilibrium concept – sustainable equilibrium – which is a sequence of history-contingent policies that satisfy certain optimality criteria for the government and private agents.

Recent developments in solving for the set of sustainable government policies use the techniques of the analysis of repeated games due to Abreu (1986) and Abreu, Pearce and Stachetti (1990). Phelan and Stachetti (2001) extend these methods to analyse the equilibria of the Ramsey model of capital taxation. Their contribution is to provide a method in which the behaviour of consumers is summarized as a solution to the competitive equilibrium, thus significantly reducing the dimensionality of the problem. They provide a characterization of the whole set of sustainable equilibria of the game. Their methods are especially relevant for the environments in which the punishment to the deviator is difficult to characterize analytically.

Benhabib and Rusticchini (1997) and Marcet and Marimon (1994) provide an alternative method to solve policy games without commitment. They use the techniques of optimal control in which they explicitly impose additional constraints on the standard optimal tax problem such that a government does not deviate from the prescribed sequence of taxes. Their methods, while easier to use than those of Abreu (1986), Abreu, Pearce and Stachetti (1990) and Phelan and Stachetti (2001), are efficient only if the worst punishment to the deviating government can be easily determined.

Klein, Krusell and Rios-Rull (2004) numerically solve for equilibria where reputational mechanisms are not operative and characterize Markov-perfect equilibria of the dynamic game between successive governments in the context of optimal Ramsey taxation. For a calibrated economy, they find that the government still refrains from taxing at confiscatory rates.

Optimal monetary policy without commitment

The problem of time consistency also arises in monetary economics. Kydland and Prescott (1977) and Barro and Gordon (1983) analyse a reduced form economy with a trade-off between inflation and unemployment. Consider an economy where the growth rate of nominal wages is being set one period in advance. The government can decrease unemployment by having setting the inflation rate higher than the wage rate, thus reducing the real wage; but inflation is socially costly. Suppose that a monetary authority chooses the inflation rate *after* nominal wages were set in the economy to maximize social welfare. Such a rate would equalize the marginal benefits of reducing unemployment and the marginal costs of increasing inflation. But now consider wage determination in a rational-expectations equilibrium. In anticipation of the government's policy, agents will choose a positive growth rate of wages to avoid losses from inflation. Therefore, in equilibrium the monetary authority is not able to affect unemployment, but there is a positive rate of inflation. This outcome is inefficient since by committing not to inflate *ex ante* the monetary authority could achieve the same level of unemployment

but with zero inflation. Therefore, the lack of commitment by the monetary authority will lead to *inflationary bias*, or an inefficiently high level of inflation.

Similar effects are present in many other monetary models. For example, Calvo (1978) shows time inconsistency of the optimal policy in a general equilibrium model. Chang (1998) considers a version of Calvo's model to find the optimal monetary policy without commitment. Similar to Phelan and Stacchetti (2001), he uses tools of repeated game theory to describe the best equilibrium in the game between the central bank and a large group of agents.

A substantial amount of work has been done in finding the ways to overcome time consistency problems. One of the first practical proposals is Rogoff's (1985) suggestion to appoint a ‘conservative’ central banker, whose private valuation of the costs of inflation is higher than the social valuation. Such a banker has less temptation to inflate, and the inflationary bias will be reduced.

Pre-specifying the rules of conduct for monetary policy reduces the discretionary actions a central bank can undertake and improves time consistency. For example, the commonly advocated Taylor rule prescribes that the central bank sets nominal interest rates as a linear function of inflation and the output gap with fixed coefficients (see, for example, Woodford, 2003). On the other hand, it may be desirable to leave some discretion to the central bank, particularly if it has access to information about economic conditions which is impossible or impractical to incorporate into predetermined rules. Athey, Atkeson and Kehoe (2005) consider an example of such an economy where the central bank has private information about the state of the economy, which is unavailable to others. They show that the optimal policy in such settings is an *inflationary cap* that allows discretion to the central bank as long as the inflation rate is below a certain bound.

Following Lucas and Stokey's (1983) analysis, substantial work has been done in determining conditions under which the government can eliminate the time consistency problem by optimally choosing debt of various maturities. Lucas and Stokey themselves point out the fundamental difficulty with this approach in monetary economies since, as long as the government holds a positive amount of nominal debt, it is tempted to inflate in order to reduce its real value. Two recent papers describe some of the conditions under which this problem can be overcome. Alvarez, Kehoe and Neumeyer (2004) consider several monetary models and show that if it is optimal to set nominal interest rates at zero (that is, the optimal monetary policy with commitment is to follow the *Friedman rule*), then the time consistency problem can be solved. By issuing a mixture of nominal and real (indexed) bonds in such a way that the present value of the nominal claims is zero, the temptation for inflation can be removed. Persson, Persson and Svensson (2006) consider a model where the Friedman rule is not optimal, but they still are able to characterize the optimal maturity structure of nominal and indexed bonds that achieve the social optimum with commitment even with time-inconsistent government.

See Also

- monetary and fiscal policy overview
- optimal taxation
- optimal fiscal and monetary policy (with commitment)
- repeated games

Bibliography

Abreu, D. 1986. Extremal equilibria of oligopolistic supergames. *Journal of Economic Theory* 39, 191–225.

Abreu, D., Pearce, D. and Stacchetti, E. 1990. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica* 58, 1041–63.

Alvarez, F., Kehoe, P. and Neumeyer, P. 2004. The time consistence of optimal monetary and fiscal policies. *Econometrica* 72, 541–67.

Athey, S., Atkeson, A. and Kehoe, P. 2005. The optimal degree of monetary policy discretion. *Econometrica* 73, 1431–76.

Barro, R. and Gordon, D. 1983. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91, 589–610.

Benhabib, J. and Rustichini, A. 1997. Optimal taxes without commitment. *Journal of Economic Theory* 77, 231–59.

Calvo, G. 1978. On the time consistence of optimal policy in a monetary economy. *Econometrica* 46, 1411–28.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54, 607–22.

Chang, R. 1998. Credible monetary policy in an infinite horizon model: recursive approach. *Journal of Economic Theory* 81, 431–61.

Chari, V. and Kehoe, P. 1990. Sustainable plans. *Journal of Political Economy* 98, 783–802.

Judd, K. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 28, 59–83.

Klein, P., Krusell, P. and Ríos-Rull, J.-V. 2004. Time consistent public expenditures. Discussion Paper No. 4582. London: CEPR.

Kydland, F. and Prescott, E. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–92.

Lucas, R. and Stokey, N. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12, 55–93.

Marcet, A. and Marimon, R. 1994. Recursive contracts. Working Paper No. 337, Department of Economics and Business, Universitat Pompeu Fabra.

Persson, M., Persson, T. and Svensson, L. 2006. Time consistency of fiscal and monetary policy: a solution. *Econometrica* 74, 193–212.

Phelan, C. and Stacchetti, E. 2001. Sequential equilibria in a Ramsey tax model. *Econometrica* 69, 1491–518.

Rogoff, K. 1985. The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100, 1169–90.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Golosov, Mikhail and Aleh Tsyvinski. "optimal fiscal and monetary policy (without commitment)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000099> doi:10.1057/9780230226203.1224

The New Palgrave Dictionary of Economics Online

optimal tariffs

Nuno Limão
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Optimal tariffs allow a country to exploit its market power in international trade. A country can improve its terms of trade by unilaterally restricting its exports if it faces a downward-sloping demand for them or restricting its imports if it faces an upward-sloping foreign export supply. This argument against unilateral free trade is over 150 years old but it remains central to modern theories that explain trade agreements and their rules. This, along with recent evidence that prior to such agreements countries exploit their market power in trade, shows that optimal tariffs may be an important positive theory of protection.

Keywords

Corn Laws; cross-elasticities; free trade; imperfect competition; marginal rates of transformation; market power; marketing boards; monopolistic competition; monopsony pricing; optimal tariffs; optimal taxation; Smoot Hawley Tariff Act of 1930 (USA); tariffs; terms of trade; Torrens, R.; trade agreements; trade policy, political economy of

Article

A country that faces a downward-sloping demand for its exports has market power and therefore, as a monopolist, can benefit from restricting its export supply. When a country's exporters are perfectly competitive, the government can coordinate this restriction via an export tax, which increases the world price for its exports and so improves its terms of trade. Analogously, a country facing an upward-sloping export supply has market power in imports and can benefit from restricting them via a tariff. Generally, the optimal tariff is defined as the rate that unilaterally maximizes a country's welfare and is given by the inverse elasticity of foreign export supply, as determined by optimal monopsony pricing.
The terms-of-trade argument against unilateral free trade is over 150 years old, yet it remains one of the hardest to refute theoretically. The reason is simple. A country's atomistic consumers impose an

externality on each other since, by increasing import demand, they raise the equilibrium price for all. The optimal instrument to correct an externality must target it at the source (Bhagwati and Ramaswami, 1963), and the optimal tariff does this by reducing import demand. This quantity reduction entails a cost but, for a sufficiently small tariff, it is more than offset by the improved terms of trade. This is one of the only cases when, in the absence of retaliation, the tariff is a first-best instrument. However, if a tariff improves a country's terms of trade, it worsens those of its trading partner, who is therefore likely to retaliate. The typical trade war outcome is to leave both worse off relative to free trade, which explains many economists' opposition to optimal tariffs as a normative theory. The trade war outcome points to the benefits from reciprocal tariff reductions and as such the terms-of-trade argument remains central to modern theories of trade agreements and their rules. This, along with recent evidence that prior to such agreements countries exploit their market power in trade, shows that optimal tariffs may be an important positive theory of protection rather than an irrelevant normative one.

Informal derivation and applications

The standard derivation of the optimal tariff focuses on a standard neoclassical economy with no domestic externalities and available lump-sum transfers to address any resulting redistribution issues (see Graaf, 1949–50). A Pareto optimum for the closed economy requires the domestic marginal rate of substitution between any two goods *i* and *j* to equal their marginal rates of transformation, which in a competitive economy is done via the domestic relative prices, that is, $MRS_{ij} = p_i / p_j = MRT_{ij}$. An open economy can exchange goods at the prevailing world prices, which can be thought of as having access to a new technology or foreign rate of transformation. Now efficient production requires the domestic MRT_{ij} to equal the marginal foreign rate of transformation (MFRT). Optimal ad valorem tariffs, t_i, imposed on the world prices, π_i, ensure that this additional condition for efficiency is met, by introducing a wedge such that the relative price faced by domestic producers is equal to $MFRT_{ij}$, that is, $p_i / p_j = \pi_i(1 + t_i) / \pi_j(1 + t_j)$. The final step is to determine the $MFRT_{ij}$, which simply reflects the relative marginal cost of these goods in the world market. The marginal cost for the importer is $\pi_i + a_i$, the price paid for the unit plus the marginal change in price(s) it causes, $a_i = \Sigma_k m_k \partial p_k / \partial m_i$. Therefore the optimal *ad valorem* tariff rates are determined by

$$\frac{\pi_i(1 + t_i)}{\pi_j(1 + t_j)} = \frac{\pi_i(1 + a_i / \pi_i)}{\pi_j(1 + a_j / \pi_j)} \text{ for all } i \text{ and } j, \tag{1}$$

which is satisfied by any tax structure such that

$$t_i = \frac{a_i}{\pi_i} \text{ for all } i. \tag{2}$$

When all cross-price elasticities are zero, a_i/π_i is simply the inverse of the foreign export supply for i, $1/\varepsilon_i$, and we obtain the standard formula, $t_i = 1/\varepsilon_i$. Otherwise t_i also includes the cross-elasticities, as a_i captures the weighted sum of marginal world price changes in all goods due to the increase in demand for i.

Since the cross-effects in a_i can be negative the optimal tariff may be zero or even a subsidy on any or all goods. However, that can't be the case with only one import and one export good, $i = m$ and e, as is easily shown if their cross-elasticity is zero. To see this, note that with two goods we can attain the same outcome with either a tariff or an export tax (Lerner, 1936), which simultaneously accounts for market power in imports and exports. Solving (1) with $t_e = 0$ we have the import tariff rate

$$t_m = \frac{1/\varepsilon_m + 1/\varepsilon_e}{1 - 1/\varepsilon_e}, \tag{3}$$

which is positive given the positively defined elasticities of foreign export supply, ε_m, and foreign import demand, ε_e, and $1/\varepsilon_e < 1$.

The result extends in several ways under perfect competition settings. If a domestic distortion exists, and is addressed by a first-best instrument, then the rate in (2) is generally still optimal. Graaf (1949–50) shows this for external (dis)economies in production, for example. The equivalence of tariffs and quantity restrictions, that is, quotas, under certainty implies that quotas can be used to the same effect provided that their rents accrue to the country that imposes them (but the welfare and trade volume outcomes of tariff and quota wars differ, as shown by Rodriguez, 1974). Kemp (1966) and Jones (1967) derive the optimal tax structure when capital is mobile and a country has market power in goods and factors trade. Similarly to (2) the optimal tariffs on goods take into account their effect on the price of capital and vice versa.

Tariffs can also affect a country's terms of trade under imperfect competition. However, there are fewer general results in these settings, even in the simpler cases of zero cross-price elasticities. Nonetheless, a few points are worth noting. First, if a country has a monopoly importer or exporter (for example, an agricultural marketing board or a cartel of oil exporters such as OPEC) then there is no first best role for a trade tax – a monopolist would already optimally restrict quantities. A tariff may still be necessary to internalize effects from any cross-elasticities, as Gros (1987) shows for a monopolistic competition model. Second, under imperfect competition a tariff can affect a country's terms of trade even if it has an infinitesimally small share of the world's expenditure. For example, if imposing a small *ad valorem* tariff reduces the import demand elasticity, then it also improves the welfare of a small importer facing a

monopoly exporter (Katrak, 1977; Brander and Spencer, 1984). However, when the country is small the tariff is generally not the first best instrument.

Early contributions and current relevance as a positive theory

There have been four important waves in the development and application of the optimal tariffs idea. They were each about 40 to 50 years apart, and at least three appear to be linked to important policy events.

Several early advances in trade theory arose during the debate over the repeal of British import duties imposed by the Corn Laws of 1815. Robert Torrens, a famous classical economist, initially supported the repeal but eventually turned against unilateral free trade as he understood that countries may gain from tariffs through an improvement in their terms of trade. This basic idea and the intuition for it are found in Torrens (1833; 1844) and Mill (1844). However, a country will actually gain only if the terms-of-trade benefit offsets the cost from lower import volume; in a second phase of development, Edgeworth (1894) shows that this is the case unless the foreign country's offer curve is perfectly elastic, while Bickerdike (1906; 1907) develops the first optimal tariff formula, similar to (3).

Renewed interest in the topic came after the Smoot Hawley Tariff Act of 1930, which raised US average tariffs to about 50 per cent and triggered a cycle of tariff retaliation. The key contributions by Kaldor (1940), Scitovsky (1942) and Johnson (1953–4) focus on the outcomes when countries retaliate. Johnson (1953–4) shows the outcome of a tariff war using tariff reaction curves that summarize a country's best response. He confirms that two *symmetric* countries prefer free trade to a trade war; but otherwise one of them may be better off under a trade war.

The latest developments in the topic also came in the wake of important economic events. Mayer (1981) examines the possible tariff outcomes under the tariff cutting formulas used in the 1973–9 multilateral trade negotiations under the General Agreement on Tariffs and Trade (GATT). Since then numerous authors have relied on the tariff war equilibrium as the threat point for the theoretical analysis of multilateral and bilateral trade agreements. Notably, Bagwell and Staiger (1999; 2002) argue that the purpose of the GATT and its successor, the World Trade Organization, is to allow countries to reciprocally lower protection in a way that eliminates the terms-of-trade component of tariffs, and show that such an economic theory of GATT can explain several of its key rules.

Despite the success of the terms-of-trade motive for tariffs in explaining important features of trade agreements, its power as a positive theory of trade protection is often questioned for two reasons. First, governments do not set tariffs to maximize social welfare. Although governments often set tariffs to redistribute income across interest groups, this does not imply that tariffs will not reflect market power. For example, Johnson (1950) derives the revenue-maximizing tariff rate, which does not maximize welfare but is nonetheless increasing in a country's market power since a given tariff rate yields higher revenue, under a less elastic export supply. Moreover, recent micro-founded political economy models predict that a large country's unilateral tariff reflects its market power, even if the government places no weight on social welfare (Grossman and Helpman, 1995). Thus, even if the primary objective of the government is a political economy one, its tariffs can reflect market power since this allows it to achieve that objective at a lower cost as it captures some income from its trading partners via improved terms of trade.

The second critique is the argument that most countries cannot affect their terms of trade and so it is not an important determinant of protection. Critics concede that certain commodity exporters do have some market power and have at times exerted it (for example, OPEC's oil restrictions or export taxes by marketing boards). But evidence of market power in exports appears to go beyond these obvious cases since aggregate estimates of ε_e in (3) are often found to be low, sometimes close to unity. Nonetheless, there are considerable difficulties in estimating and interpreting such aggregate elasticities, which are often estimated only for countries already setting their tariffs cooperatively. Therefore cross-country comparisons of average tariffs and these aggregate elasticities cannot provide much insight into the empirical importance of the terms-of-trade motive.

There is also growing evidence of market power in imports since when countries change their exchange rates or tariffs part of the effect is absorbed by the foreign exporters (cf. Kreinin, 1961). Broda, Limão and Weinstein (2006) provide compelling evidence that countries have and exploit their market power. They estimate inverse foreign export supply elasticities by good and country, and find that even small countries have some market power, which is increasing in country size and degree of good differentiation. They then examine tariffs for countries that are not setting them cooperatively and find that they are set higher in goods with higher inverse elasticities. They conclude that market power is an economically and statistically important determinant of tariffs.

In sum, optimal tariffs are evolving from a curious normative theory to a positive one. The broad applicability of the terms-of-trade motive for tariffs; its theoretical success in explaining important rules of trade agreements; and the recent evidence that countries exploit their market power, all indicate this will remain a key concept in economics.

See Also

- Bickerdike, Charles Frederick
- Corn Laws, free trade and protectionism
- Johnson, Harry Gordon
- tariffs
- Torrens, Robert

Bibliography

Bagwell, K. and Staiger, R.W. 1999. An economic theory of GATT. *American Economic Review* 89, 215–48.

Bagwell, K. and Staiger, R.W. 2002. *The Economics of the World Trading System*. Cambridge, MA: MIT Press.

Bhagwati, J. and Ramaswami, V.K. 1963. Domestic distortions, tariffs and the theory of optimum subsidy. *Journal of Political Economy* 71, 44–50.

Bickerdike, C.F. 1906. The theory of incipient taxes. *Economic Journal* 16, 529–35.

Bickerdike, C.F. 1907. Review of A.C. Pigou's protective and preferential import duties. *Economic Journal* 17, 98–108.

Brander, J.A. and Spencer, B.J. 1984. Tariff protection and imperfect competition. In *Monopolistic Competition and International Trade*, ed. H. Kicrzkowski. Oxford: Oxford University Press.

Broda, C., Limão, N. and Weinstein, D. 2006. Optimal tariffs: the evidence. Working Paper No. 12033. Cambridge, MA: NBER.

Edgeworth, F.Y. 1894. The theory of international values. *Economic Journal* 4, 35–50.

Graaf, J.V. 1949–50. On optimum tariff structures. *Review of Economic Studies* 17, 47–59.

Gros, D. 1987. A note on the optimal tariff, retaliation and the welfare loss from tariff wars in a framework with intra-industry trade. *Journal of International Economics* 23, 357–67.

Grossman, G. and Helpman, E. 1995. Trade wars and trade talks. *Journal of Political Economy* 103, 675–708.

Johnson, H.G. 1950. Optimum welfare and maximum revenue tariffs. *Review of Economic Studies* 19, 28–35.

Johnson, H.G. 1953–4. Optimum tariffs and retaliation. *Review of Economic Studies* 21, 142–53.

Jones, R.W. 1967. International capital movements and the theory of tariffs and trade. *Quarterly Journal of Economies* 81, 1–38.

Kaldor, N. 1940. A note on tariffs and the terms of trade. *Economica* 7, 377–80.

Katrak, H. 1977. Multi-national monopolies and commercial policy. *Oxford Economic Papers* 29, 283–91.

Kemp, M.C. 1966. The gains from international trade and investment: a neo-Heckscher–Ohlin approach. *American Economic Review* 65, 788–809.

Kreinin, M.E. 1961. Effect of tariff changes on the prices and volume of imports. *American Economic Review* 51, 310–24.

Lerner, A.P. 1936. The symmetry between import and export taxes. *Economica* 3, 306–13.

Mayer, W. 1981. Theoretical considerations on negotiated tariff adjustments. *Oxford Economic Papers* 33, 135–53.

Mill, J.S. 1844. *Essays on some Unsettled Questions of Political Economy*. London: Parker.

Rodriguez, C. 1974. The non-equivalence of tariffs and quotas under retaliation. *Journal of International Economics* 4, 295–98.

Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9, 89–110.

Torrens, R. 1833. *Letters on Commercial Policy*. London: Longman.

Torrens, R. 1844. *The Budget: On Commercial and Colonial Policy*. London: Smith-Elder.

How to cite this article

Limão, Nuno. "optimal tariffs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000033> doi:10.1057/9780230226203.1225

The New Palgrave Dictionary of Economics Online

optimal taxation

Louis Kaplow
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Optimal taxation concerns how various forms of taxation should be designed to maximize social welfare. The task requires an integrated consideration of the revenue-raising and distributive objectives of taxation. The central instrument in developed economies is the labour income tax, the analysis of which was pioneered by Mirrlees (1971). Subsequently, Atkinson and Stiglitz (1976) showed how commodity taxes should be set in the presence of an optimal income tax, the results differing qualitatively from, and in important respects displacing, the teachings derived from Ramsey's (1927) seminal analysis of the pure commodity tax problem.

Keywords

ability; commodity taxation; externalities; income taxation; labour supply; leisure; linear income tax; lump-sum taxes; marginal cost pricing; marginal tax rates; marginal utility of consumption; Mirrlees, J.; nonlinear income tax; optimal government policy; optimal taxation; optimal tax systems; Pigouvian taxes; public goods; Ramsey taxation; redistribution; revelation principle; separable preferences; social preferences; social welfare function; taxation of capital; taxation of income; transfer programmes; uniform taxation; value-added tax

Article

Optimal taxation concerns the question of how various forms of taxation should be designed in order to maximize a standard social welfare function subject to a revenue constraint. The task requires an integrated consideration of the revenue-raising and distributive objectives of taxation. The central instrument in developed economies is the labour income tax. Mirrlees (1971) pioneered the analysis of this challenging problem. Subsequently, Atkinson and Stiglitz (1976) showed how commodity taxes should be set in the presence of an optimal income tax. The results are qualitatively different from —

and in important respects displace – prior teachings that originate in Ramsey's (1927) analysis of the pure commodity tax problem. In addition to setting particular taxes optimally, it is also necessary to choose optimally among tax systems.

Income taxation

Model

The standard optimal income tax model involves a one-period setting in which individuals' only choice variable is their degree of labour effort *l*. There is a single composite consumption good *c*. An individual's utility is given by $u(c, l)$, where $u_c>0$ and $u_l<0$. An individual's consumption is given by

$$c = wl - T(wl), \tag{1}$$

where *w* is the individual's wage rate and *T* is the tax-transfer function.
The motivation for redistributive taxation is that individuals differ, in particular in their wages, that is, their earning abilities. The distribution of abilities will be denoted *F*(*w*), with density $f(w)$. Individuals' wage rates are taken to be exogenous. Their pre-tax earnings *wl* are the product of their wage rate and level of labour effort. More broadly, one can interpret labour effort as including not only hours of work but also intensity and not only productive effort but also investments in human capital.
Taxes and transfers, $T(wl)$, at any income level may be positive or negative. The (uniform) level of the transfer received by an individual earning no income, that is, –*T*(0), is sometimes referred to as the grant *g*. Taxes may be interpreted broadly, to include sales taxes or value-added tax (VAT) payments in addition to income taxes. Transfers include those through the tax system in addition to welfare programmes. The inclusion of transfers is important both practically, since they are in fact significant, and conceptually, since otherwise redistribution would be limited to transfers between the rich and the middle class, once the poor were exempted from the tax system.
Taxes and transfers are taken to be a function of individuals' incomes, assumed to be observable, and it is this dependence of taxes on income that is the source of distortion. If taxes could instead depend directly on individuals' abilities, *w*, individualized lump-sum taxes would be feasible and redistribution could be accomplished without distorting labour supply. Ability, however, is assumed to be unobservable.
The government's problem is to choose $T(wl)$ to maximize social welfare, which can be stated as

$$\int W(u(c(w), l(w)))\, f(w)\, dw, \tag{2}$$

where *c* and *l* are each expressed as functions of *w* to refer to the levels of consumption achieved and labour effort chosen by an individual of ability *w*. If *W* is linear, the welfare function is utilitarian, whereas if *W* is strictly concave, additional weight is given to inequality in utility levels (not just levels of marginal utilities).

This maximization is subject to a revenue constraint and to constraints regarding individuals' behaviour. The former is

$$\int T(wl(w))\, f(w)\, dw = R, \tag{3}$$

where *R* is an exogenously given revenue requirement. Here, revenue is to be interpreted as expenditures on public goods that should be understood as implicit in individuals' utility functions; because these expenditures are taken to be fixed, they need not be modelled explicitly. Regarding the latter constraints, individuals are assumed to respond to the given tax schedule optimally, which determines the functions *c* (*w*) and *l*(*w*).

Mirrlees's (1971) original exposition has been followed by subsequent elaborations, much of which is synthesized and extended in Atkinson and Stiglitz (1980), Stiglitz (1987), Tuomala (1990), and Salanié (2003). Because the problem is formidable, the present discussion will be confined to stating basic results, such as are embodied in first-order conditions and produced by simulations.

Linear income tax

Substantial illumination with greatly reduced complexity is provided by first examining a linear income tax,

$$T(wl) = twl - g, \tag{4}$$

where *t* is the (constant, income-independent) marginal tax rate and *g*, as previously noted, is the uniform per-capita grant. Because of the presence of the grant *g*, a linear income tax can be highly redistributive (consider setting *t* at 100 per cent and *g* equal to mean income net of any per capita revenue requirement – in the absence of incentive constraints) or not at all redistributive (*t* may be 0 per cent and *g* equal to the negative of the per capita revenue requirement). Foreshadowing discussion of the nonlinear income tax, the degree of redistribution is more directly related to the levels of *t* and *g* than to

the shape (deviation from linearity) of the tax schedule.
To derive the optimal linear income tax, the government's maximization problem can be written in Lagrangian form as choosing *t* and *g* to maximize

$$\int [W(u((1-t)wl(w)+g, l(w))) + \lambda(twl(w) - g - R)]\, f(w)\, dw, \tag{5}$$

where λ is the shadow price of revenue, referring to the constraint (3), and expression (4) is substituted into expression (1) so that consumption is expressed in terms of the specific linear tax system under consideration. Following Atkinson and Stiglitz (1980) and Stiglitz (1987), the first-order condition for the optimal tax rate can usefully be expressed as

$$\frac{t}{1-t} = -\frac{\operatorname{cov}(\alpha(w), y(w))}{\int y(w)\,\varepsilon(w)\, f(w)\, dw}, \tag{6}$$

where $y(w)=wl(w)$, income earned by individuals of ability w; $\varepsilon(w)$ is the compensated elasticity of labour effort of individuals of ability w; and $\alpha(w)$ is the net social marginal valuation of income, evaluated in dollars, of individuals of ability w:

$$\alpha(w) = \frac{W' u_c(w)}{\lambda} + tw\left(\frac{\partial l(w)}{\partial g}\right). \tag{7}$$

The numerator of the first term on the right side of expression (7) indicates how much additional (lump-sum) income to an individual of ability w contributes to social welfare (u_c indicates how much utility rises per dollar and W' indicates the extent to which social welfare increases per unit of utility) and this product is converted to a dollar value by dividing by the shadow price of government revenue. The second term takes into account the income effect, namely, that giving additional lump-sum income to an individual of ability w will reduce labour effort ($\partial l(w)/\partial g<0$), which in turn reduces government tax collections by tw per unit reduction in $l(w)$.
Expression (6) indicates how various factors affect the optimal level of a linear income tax. Beginning

with the numerator on the right side, a higher (in magnitude) covariance between α and y favours a higher tax rate. In the present setting, $\alpha(w)$ will (under assumptions ordinarily postulated) be falling with income. Note that a larger covariance does not involve a closer (negative) correlation but rather a higher dispersion (standard deviation) of α and of y. The dispersion of α will tend to be greater the more concave (egalitarian) is the welfare function W and the more concave is utility as a function of consumption (that is, the greater the rate at which marginal utility falls with income). Income, y, will have a higher dispersion (again, under standard assumptions) when the distribution of underlying abilities is more unequal. In sum, more egalitarian social preferences, more rapidly declining marginal utility of consumption, and higher underlying inequality each contribute to a higher optimal tax rate. The denominator indicates that a higher compensated labour supply elasticity favours a lower tax rate. The other terms in the integrand indicate that, ceteris paribus, the labour supply elasticity matters more with regard to high-income individuals and at ability levels where there are more individuals (typically the middle of the income distribution) because of the greater sacrifice in revenue.

The foregoing exposition is incomplete in not emphasizing the various respects in which income effects are relevant (they influence α and also λ) and in ignoring that the values on the right side of expression (6) are endogenous. Especially for the latter reason, the literature has relied heavily on simulations.

The most-reported optimal linear income taxation simulations are those of Stern (1976). For his preferred case — an elasticity of substitution between consumption and labour of 0.4, a government revenue requirement of 20 per cent of national income, and a social marginal valuation of income that decreases roughly with the square of income — he finds that the optimal tax rate is 54 per cent and that individuals' lump-sum grant equals 34 per cent of average income. To illustrate the benefits of redistribution, he finds that a scheme that uses a lower tax rate, just high enough to finance government programmes (that is, with a grant of zero), produces a level of social welfare that is lower by an amount equivalent to approximately 5 per cent of national income. If there is very little weight on equality, the optimal tax rate is only 25 per cent, whereas if there is extreme weight on equality, the optimal tax rate is 87 per cent. Returning to his central case, an extremely low labour supply elasticity implies an optimal tax rate of 79 per cent, and an elasticity as high as had been used in some earlier literature implies an optimal tax rate of 35 per cent. In the absence of the need to finance government expenditures, the optimal tax rate is 48 per cent, and if government expenditures are twice as high, the optimal tax rate is 60 per cent.

Nonlinear income tax

Mirrlees (1971) and subsequent investigators employ control-theoretic techniques to address the more general formulation of the optimal nonlinear income taxation problem, which requires choosing an entire tax schedule $T(w)$ rather than a single tax rate. In this maximization, the constraints regarding individuals' maximizing behaviour entail that no individual of any type w will prefer the choice specified for any other type $w°$. This approach is related to the use of the revelation principle in work on mechanism design, and in similar spirit many researchers following Stiglitz (1982) and others analyse a simpler, discrete variant of the problem, often involving two types, in which the binding incentive constraint is usually that the high-ability type not have an incentive to mimic the low-ability type in

order to pay less tax.

The analysis of the continuous case can be summarized in a first-order condition for the optimal marginal income tax rate at any income level y^*, where w^* and l^* correspond to the ability level and degree of labour effort supplied by the type of individual who would earn y^*. Making the simplifying assumptions that utility is separable between consumption and labour effort and that marginal utility u_c is constant, the condition can be expressed as

$$\frac{T'(w^* l^*)}{1 - T'(w^* l^*)} = \frac{1 - F(w^*)}{\xi^* w^* f(w^*)} \frac{\int_{w^*}^{\infty} \left(1 - \frac{W'(u(w))u_c}{\lambda}\right) f(w)\,dw}{1 - F(w^*)}, \tag{8}$$

where $\xi^* = 1/(1 + l^* u_{ll}/u_l)$ — which, when marginal utility is constant as assumed here, equals $\varepsilon/(1+\varepsilon)$, where ε again is the elasticity of labour supply. For derivations of related expressions, see, for example, Auerbach and Hines (2002), Atkinson and Stiglitz (1980), Dahan and Strawczynski (2000), Diamond (1998), Saez (2001), and Stiglitz (1987). Note that this formulation (like those in recent literature) includes $1-F(w^*)$ in both the numerator and the denominator on the right side. The motivation is that, in the first term, $(1 - F(w^*)) / f(w^*)$ is purely a property of the distribution of w, and, in the second term, because the numerator is an integral from w^* to ∞, the term as a whole gives an average value for the expression in parentheses in the integrand. Both aspects aid intuition, as will be seen in the discussion to follow.

Expression (8), being a first-order condition, should be interpreted by reference to an adjustment that slightly raises the marginal tax rate at income level y^* (say, in a small interval from y^* to $y^* + \delta$), leaving all other marginal tax rates unaltered. There are two effects of such a change. First, individuals at that income level face a higher marginal rate, which will distort their labour effort, a cost. Second, all individuals above income level y^* will pay more tax, but these individuals face no new marginal distortion. That is, the higher marginal rate at y^* is inframarginal for them. Since those thus giving up income are an above-average-income slice of the population (it is the part of the population with income above y^*), there tends to be a redistributive gain.

The right side of expression (8) can readily be interpreted in terms of this perturbation (although it should be kept in mind that this interpretation omits, inter alia, income effects and the endogeneity of variables). Begin with the first term. Revenue is collected from all individuals with incomes above y^*, which is to say all ability types above w^*; hence the $1-F(w^*)$ in the numerator. This factor favours marginal tax rates that fall with income. As there are fewer individuals who face the inframarginal tax, the core benefit of a higher marginal rate declines. In the extreme, if there is a highest known type in the income distribution, the optimal marginal rate at the top would be zero because $1-F$ would be zero: a higher rate collects no revenue but distorts the behaviour of the top individual. However, when there is

no highest type, known with certainty in advance, this result is inapplicable. Furthermore, with a known highest type, simulations suggest that zero is not a good approximation of the optimal marginal tax rate even quite close to the top of the income distribution, so the zero-rate-at-the-top result is of little practical importance.

To continue with the first term, raising the marginal rate at a particular point distorts only the behaviour of the marginal type, which explains the $f(w^*)$ in the denominator. For standard distributions, this factor is rising initially and then falling, which favours falling marginal rates at the bottom of the income distribution and rising rates at the top. The denominator also contains weights of ξ^*, indicating the extent of the distortion, and w^*, indicating how much productivity is lost per unit of reduction in labour effort. The elasticity is often taken to be constant, although some empirical evidence on the elasticity of taxable income supports a rising elasticity due to the greater ability of higher-income individuals to avoid taxes. This consideration may favour marginal rates that fall with income. Finally, w^* is rising, which also favours falling marginal rates: The greater is the wage (ability level), the greater is the revenue loss from a given decline in labour effort.

The second term applies a social weighting to the revenue that is collected. The expression in parentheses in the integrand in the numerator is the difference between the marginal dollar that is raised and the dollar equivalent of the loss in welfare that occurs on account of individuals above w^* paying more tax. As in the interpretation of expression (7), u_c is the marginal utility of consumption to such individuals, W' indicates the impact of this change in utility on social welfare, and division by λ, the shadow price on the revenue constraint, converts this welfare measure into dollars. This integral is divided by $1-F(w^*)$, which as noted makes the second term an average for the affected population. This term tends to favour marginal rates that rise with income. The greater is w, the lower is W' (unless the welfare function is utilitarian, in which case this is constant) and the lower would be the marginal utility of income u_c (had we not abstracted from this effect in the simplifying assumptions). Hence, at a higher w^* the average value of the term subtracted in the integrand is smaller, making the entire term larger. Note further that, if social welfare or utility is reasonably concave, $W'u_c$ will approach zero at high levels of income, at which point this term will be nearly constant in w^*. That is, the term favours rising marginal tax rates when income is low or moderate, but has little effect on the pattern of marginal tax rates near the top of the income distribution.

Because of difficulties in determining the shape of the optimal income tax schedule by mere inspection of the first-order condition (8), analysts beginning with Mirrlees (1971) have used simulations to help join the theoretical analysis with empirical estimates of labour supply elasticities and of the distribution of skills or income in order to provide further illumination. Tuomala (1990) offers a useful survey and set of calculations. In all the cases he reports, marginal tax rates fall as income increases, except at very low levels of income. Mirrlees's (1971) original calculations had displayed a similar tendency, but subsequent researchers had questioned the extent to which this result may have depended on the social preferences he stipulated or the arguably high labour supply response he assumed. Later work, however, suggests that a greater social preference for equality or a lower labour supply response tends to increase the level of optimal marginal tax rates but does not generally result in a substantially different shape. This phenomenon is also illustrated by Slemrod et al. (1994), who examine the optimal two-bracket income tax. In all of their simulations, the optimal upper-bracket marginal rate is lower than the lower-

bracket rate; indeed, this gap widens as the social preference for equality increases because of the additional value of raising the lower-bracket rate in generating funds to increase the grant, which is of greatest relative benefit to the lowest-income individuals.

Subsequent work further explores the circumstances in which optimal marginal tax rates might rise with income. Kanbur and Tuomala (1994) find that, when inequality in individuals' abilities (wages) is significantly greater than previously assumed (but at levels they suggest to be empirically plausible), optimal marginal tax rates do increase with income over a substantial range, although for upper-income individuals optimal marginal rates still fall with income. Diamond (1998) examines a Pareto distribution of skills (instead of the commonly used lognormal distribution), under which the $(1-F)/f$ component of expression (8) rises more rapidly at the top of the distribution, and finds that optimal marginal tax rates are rising at the top. However, Dahan and Strawczynski's (2000) simulations indicate that Diamond's result was driven in large part by his additional assumption that preferences were quasi-linear, thus removing income effects. (Nevertheless, their diagrams do suggest that, consistent with Diamond's claim, moving from a lognormal to a Pareto distribution favours higher rates — still falling, but notably less rapidly — at the top of the income distribution.) Saez (2001), using income distribution data in the United States from 1992 and 1993, finds that the shape of the distribution of $(1-F)/wf$ is such that optimal rates should fall substantially well into the middle of the income distribution, to an income of approximately $75,000, rise until approximately $200,000, and then be essentially flat thereafter.

An additional result from the simulations is that, at the optimum, a nontrivial fraction of the population does not work, and this fraction is larger when social preferences favour greater redistribution and when the labour supply elasticity is higher. This outcome should hardly be surprising because, as the analysis of expression (8) and the simulations suggest, high marginal rates tend to be optimal at the bottom of the income distribution, along with a sizable grant. Relatedly, little productivity and thus little tax revenue is sacrificed when those with very low abilities are induced not to work (whereas substantial revenue is raised from the rest of the population, for whom marginal tax rates on the first dollars of income are inframarginal).

Extensions

Given the central importance of income taxation to the revenue and distributive objectives of government, further exploration of various aspects of the problem should be a high research priority. A number of features have received some, although generally quite limited, attention. For broader discussions and further references, see Atkinson and Stiglitz (1980), Stiglitz (1987), Tuomala (1990), Salanié (2003), and Kaplow (2008).

A critical assumption in optimal income tax analysis is that earning ability is unobservable so that income, a signal of ability, is taxed instead, which is the source of distortion. Hence, it is worth considering the possibilities for basing taxation more directly on ability. To some degree, hours may be observable, and ability (wages) can thus be inferred. But in many occupations (notably, self-employment) hours are difficult to observe, and both hours and wages are manipulable, such as by extending reported hours and lowering the reported wage. Another approach would be to measure proxies of earning ability, such as through testing. Unfortunately, skills measurable by testing explain

only some of the variance in earning ability, and, if taxes were to be based on test results or other ability measures, individuals would adjust their performance and thereby distort the measurement. A third technique — one sometimes employed — is to adjust taxes and transfers for observable personal attributes, such as physical disability, age or family composition.

In general, tax and transfer schedules could be made a function of various imperfect signals of ability (or of other pertinent differences, such as in utility functions). For each value of the signal, there would in essence be a different tax schedule, governed by the first-order condition (8); each of these tax schedules would, however, be linked in a common optimization by the shadow price λ. One might view models like those of Akerlof (1978), in which he assumes that a subset of the lowest-ability group can be identified perfectly ('tagged'), and Stern (1982), in which he examines the usefulness of a noisy signal of ability in a two-type model, as special cases of this more general formulation.

There exist myriad additional complications. One is that income may be a noisy signal of ability, whether because of variations in occupations (for a given ability, one job may pay more to compensate for specific disamenities) or in preferences (an individual may earn more not because of greater ability but rather due to a higher marginal utility of consumption or a lower marginal disutility of labour effort). Another possibility is that individuals may have preferences concerning redistribution itself, perhaps due to altruism or envy. Other topics that have been explored include liquidity constraints, general equilibrium effects of taxation on the distribution of pre-tax wages, uncertainty, interactions with non-tax distortions, and human capital.

Commodity taxation

Commodity taxation with income taxation

To examine optimal commodity taxation with labour income taxation, the foregoing model can be modified as follows. In place of consumption *c*, individuals choose commodity vectors $\boldsymbol{x}$ and, as before, labour effort *l* to maximize the utility function $u(x, l)$. On the left side of individuals' budget constraints (1), *c* is replaced by $\rho \boldsymbol{x}$, where ρ is the consumer price vector equal to $\boldsymbol{p}+\tau$: the sum of a producer price vector (taken to be constant and equal to production costs) and a vector of commodity taxes (which, if negative, are subsidies).

Atkinson and Stiglitz (1976) demonstrate that, when the income tax is set optimally, commodity taxes should be undifferentiated, that is, $\tau=0$, when utility is weakly separable in labour (on which more in a moment). Alternatively, other levels of τ are similarly optimal as long as the ratio of any two consumer prices equals the ratio of producer prices, with the difference in consumer price level being offset by an adjustment to the income tax schedule. (For example, if all commodity taxes are ten per cent rather than zero, the income tax schedule may be reduced so that, at all levels of pre-tax income *wl*, disposable income is ten per cent higher.) Subsequent work extends this uniformity result to examine cases in which the income tax need not be optimal and to assess various partial reforms, one result being that any proportionate reduction in non-uniform commodity taxes can generate a Pareto improvement (see Kaplow, 2006; also Konishi, 1995; Laroque, 2005).

The intuition behind the uniformity result is that, despite the second-best setting (due to the inherently distortionary character of a redistributive labour income tax), there is nothing to be gained — except

distortion of consumption — by differentiating commodity taxes when the utility function is weakly separable in labour. When that assumption is relaxed, one has the qualification — due originally to Corlett and Hague (1953) in a Ramsey setting — that complements to leisure (labour) should be taxed (subsidized). For example, taxing beach attendance or the purchase of novels may make leisure less attractive, encouraging labour effort and thereby reducing the distortion due to the income tax. Other qualifications, including with regard to preferences that depend on ability, other preference heterogeneity, and administrative and enforcement concerns, are catalogued in Kaplow (2008).

The Ramsey problem: commodity taxation alone

The foregoing analysis is usefully contrasted with that of Ramsey (1927), who considered how to set commodity taxes on a population of identical individuals to meet a revenue requirement. The familiar result is that commodity taxes should be inversely proportional to the elasticity of demand, with refinements for demand interdependencies. Introducing nonidentical individuals leads to modifications reflecting distributive concerns that entail higher taxes than otherwise on luxuries and lower taxes on necessities. See generally Atkinson and Stiglitz (1976, 1980), Auerbach and Hines (2002), Salanié (2003), and Stiglitz (1987).

As initially emphasized in Atkinson and Stiglitz (1976) and elaborated in Stiglitz (1987) and Kaplow (2008), however, neither prescription is apt if there is also an income tax. In the original Ramsey model in which all individuals are identical and thus there are no distributive concerns, the optimal tax obviously would be a uniform lump-sum extraction (a limiting case of an income tax), which, it should be noted, neither requires information about individuals' types nor is distributively objectionable in this setting. When differences in earning ability are admitted, the optimal tax is a nonlinear income tax, and in typical cases the lump-sum component involves a uniform lump-sum subsidy. Nevertheless, optimal commodity taxation still is not guided either by the familiar inverse-elasticity rule or by the general preference for harsher treatment of luxuries than of necessities. As noted, in the basic case optimal differentiation is nil regardless of the demand elasticity or how demand changes with income, and qualifications such as that favouring taxation (subsidization) of leisure complements (substitutes) are largely unrelated to the level of the own-elasticity of demand for a commodity or its income elasticity.

Applications

Optimal commodity taxation is, in an important sense, a building block for the analysis of many other important problems. For example, Atkinson and Stiglitz (1976) explain how the analysis of optimal capital taxation can be assimilated into the framework, for it involves nonuniform taxation of consumption in different time periods, which may be interpreted in terms of the model simply as differently indexed commodities. Hence, in the basic case, the optimal tax on capital is zero.

Furthermore, as discussed by Kaplow (2004, 2008), other types of government policy may be analysed in a similar fashion. Allowing for externalities, the no-differential-tax prescription may be interpreted as requiring that consumer price ratios equal not producer price ratios but instead ratios of full social costs; hence, first-best Pigouvian taxes and subsidies (that is, set equal to marginal external effects) are optimal despite second-best concerns about distortionary income taxation and distributive effects. For public

goods, the analogy to differential taxation is a departure from the pure Samuelson rule, so in the basic case, that cost–benefit test also does not require modification on account of income tax distortions and distributive concerns. Likewise, deviations from marginal cost pricing of public production is counter-indicated.

By contrast, much prior and ongoing work examines these problems and others in a Ramsey-like setting. As Stiglitz (1987) observes, this course may be appropriate for developing economies in which income taxation is largely infeasible, but not for developed economies with an income tax.

Optimal tax systems

Most optimal taxation analysis simply assumes that certain tax instruments are available and others are not. Mirrlees (1971), Atkinson and Stiglitz (1980), and Slemrod (1990), however, emphasize the importance of motivating the presumed set of available instruments by administrative and enforcement concerns that indicate what actually is feasible. Ideally, these concerns would not be stipulated but rather would be made endogenous. Often, feasibility is a matter of degree, and one must choose among various imperfect systems, the quality of each being determined by policy choices regarding administration and enforcement and also by how the instrument is used.

To illustrate these trade-offs, note that a nonlinear income tax may be a more fine-tuned redistributive instrument than a linear income tax but is subject to additional types of manipulations that are costly to regulate. Likewise, if nonuniform commodity taxation is employed, there exist incentives to reclassify commodities. More comprehensive tax bases may avoid unnecessary distortions but be more costly to administer. The extent of evasion under any system may depend on the level of tax rates and on what other taxes are in place.

Greater attention to the choice among tax systems seems warranted. Whether or not to have a 20 per cent VAT, relying far less on income taxes, is probably a more important decision than how to set commodity tax differentials in light of subtle qualifications to the uniformity result. System choices are likely to be particularly important for developing countries, where fewer options are feasible and the available instruments are changing over time and in ways that are influenced by other government policies.

See Also

- income taxation and optimal policies
- redistribution of income and wealth
- social welfare function
- taxation of income

Bibliography

Akerlof, G.A. 1978. The economics of ‘tagging’ as applied to the optimal income tax, welfare programs, and manpower planning. *American Economic Review* 68, 8–19.

Atkinson, A.B. and Stiglitz, J.E. 1976. The design of tax structure: direct versus indirect taxation. *Journal of Public Economics* 6, 55–75.

Atkinson, A.B. and Stiglitz, J.E. 1980. *Lectures on Public Economics*. New York: McGraw-Hill.

Auerbach, A.J. and Hines, J.R. 2002. Taxation and economic efficiency. In *Handbook of Public Economics*, vol. 3, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Corlett, W.J. and Hague, D.C. 1953. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21, 21–30.

Dahan, M. and Strawczynski, M. 2000. Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates: comment. *American Economic Review* 90, 681–86.

Diamond, P.A. 1998. Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates. *American Economic Review* 88, 83–95.

Kanbur, R. and Tuomala, M. 1994. Inherent inequality and the optimal graduation of marginal tax rates. *Scandinavian Journal of Economics* 96, 275–82.

Kaplow, L. 2004. On the (ir)relevance of distribution and labor supply distortion to government policy. *Journal of Economic Perspectives* 18(4), 59–75.

Kaplow, L. 2006. On the undesirability of commodity taxation even when income taxation is not optimal. *Journal of Public Economics* 90, 1235–50.

Kaplow, L. 2008. *The Theory of Taxation and Public Economics*. Princeton: Princeton University Press.

Konishi, H. 1995. A Pareto-improving commodity tax reform under a smooth nonlinear income tax. *Journal of Public Economics* 56, 413–46.

Laroque, G. 2005. Indirect taxation is harmful under separability and taste homogeneity: a simple proof. *Economics Letters* 87, 141–4.

Mirrlees, J.A. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 68, 175–208.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 41–61.

Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68, 205–29.

Salanié, B. 2003. *The Economics of Taxation*. Cambridge, MA: MIT Press.

Slemrod, J. 1990. Optimal taxation and optimal tax systems. *Journal of Economic Perspectives* 4(1), 157–78.

Slemrod, J., Yitzhaki, S., Mayshar, J. and Lundholm, M. 1994. The optimal two-bracket linear income tax. *Journal of Public Economics* 53, 269–90.

Stern, N.H. 1976. On the specification of models of optimum income taxation. *Journal of Public Economics* 6, 123–62.

Stern, N.H. 1982. Optimum taxation with errors in administration. *Journal of Public Economics* 17, 181–211.

Stiglitz, J.E. 1982. Self-selection and Pareto efficient taxation. *Journal of Public Economics* 17, 213–40.

Stiglitz, J.E. 1987. Pareto efficient and optimal taxation and the new new welfare economics. In *Handbook of Public Economics*, vol. 2, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Tuomala, M. 1990. *Optimal Income Tax and Redistribution*. Oxford: Clarendon Press.

How to cite this article

Kaplow, Louis. "optimal taxation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000034> doi:10.1057/9780230226203.1226

The New Palgrave Dictionary of Economics Online

optimality and efficiency

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Allais, M.; competitive equilibrium; completness; convexity; core; compensated core; compensated equilibrium; efficiency; optimality and efficiency; Pareto efficiency; quasi-equilibrium; transitivity

Article

An exchange economy consists of a group of people, each of whom has preferences concerning what commodities he or she likes, and initial holdings of the various commodities available. Operating under whatever institutional rules permit freedom of contract, the society redistributes the initial holdings among itself so as to achieve a distribution that is in some sense a *solution* to the exchange problem. But in what sense? Over the years three common meanings of solution have emerged, each with ever greater clarity. In order of increasing structural content rather than historical origin they are: (*a*) optimality in the sense of Edgeworth (1881) and Pareto (1909), or for brevity *EP-optimality*; (*b*) core solutions, which originated wholly with Edgeworth (1881) but had to wait until the advent of game theory before they were properly understood; and (c) competitive equilibria, which owe most to Walras (1874). Diverse as they are these three concepts are linked by a common thread, that each agent's objective is to seek the greatest satisfaction possible within the constraints that bind him.

If the roles of objectives and constraints are interchanged in (*a*), (*b*) and (*c*) we obtain three new concepts of solution, which are in effect mirror images of the earlier ideas. Thus corresponding to EP-optimality there is (*a*)′ efficiency in the sense of Allais (1943, pp. 610–16, 637–44) and Scitovsky (1942), or in brief *AS-efficiency*. Corresponding to the core is the idea (*b*)′ of a *compensated core* (for which see below), and corresponding to competitive equilibria is the concept (*c*)′ of *compensated equilibria* due to Arrow and Hahn (1971, p. 108), although the closely related *quasi-equilibria* were defined earlier by Debreu (1962).

(Curiously enough, the passage in Scitovsky (1941–2) that gives his definition of (*a*)′ was omitted from the reprinted version in his collected essays (1964). Very clear accounts of his approach may however be found in Samuelson (1956) and Graaff (1957), while Allais has published many further

elaborations of his ideas, for example in Allais (1978). Those ideas were clearly at work in the pioneering paper by Debreu (1951), where he used them to overcome the problem ‘that no meaningful metrics exists in the satisfaction space, [that is, that utility is not cardinally measurable]’ (1951, p. 273). Later, Debreu's proof of the Second Fundamental Theorem of Welfare Economics (the ‘pricing-out’ of EP-optima) in Chapter 6 of (1959) also depended quite explicitly on the use of AS-efficiency.)
The interrelations between competitive and compensated equilibria are well recognized (see for example cost minimization and utility maximization) and concern such matters as the existence of locally cheaper points, which in turn necessarily imply market valuations of commodity bundles. The analogous interrelations between EP-optimality and AS-efficiency, and between cores and compensated cores, do not involve market phenomena and perhaps as a consequence are not so well known.

1 Preliminaries

We need appropriate language and notation, and some general assumptions. The exchange economy consists of *m* agents, indexed by *h*, and *n* goods, indexed by *i*. Each agent has a preference relation $\precsim_k$ that is defined over some subset of the non-negative orthant of R^n and whose meaning is ‘at least as good as’. It is assumed to be composed of two disjoint sub-relations, strict preferences $\succ_k$ and indifference $\sim_k$. Completeness and convexity of preferences are never assumed, and only partial transitivity is required, in the sense that the two cases $(z_k^1 \succ_k z_k^2 \text{ and } z_k^2 \succ_k z_k^3)$ and $(z_k^1 \succ_k z_k^2 \text{ and } z_k^2 \sim_k z_k^3)$ are each assumed to lead to the conclusion $z_k^1 \succ_k z_k^3$. In particular, $\sim_k$ need not be transitive.
A *distribution* (or *allocation*) *Z* is any $m \times n$ matrix of the individual holdings z_{hi}, so that Z^0 is the distribution of initial holdings z_{hi}^0 or the *endowment*. If two distributions Z^1 and Z^2 are such that $z_k^1 \succ_k z_k^2$ for every agent *h*, then we write $Z^1 \succ\succ Z^2$ and say that Z^i is *better* than Z^2. Similarly, $Z^1 \succsim Z^2$ means that $z_k^1 \succsim_k z_k^2$ for every *h*; we say that Z^1 *meets* Z^2. If Z^1 meets Z^2 and the number of agents *k* for whom $z_k^1 \sim_k z_k^2$ is at least 1 and not *m*, then we write $Z^1 \succ Z^2$.
An agent's holdings are written $z_h = (z_{h1}, z_{h2}, \ldots, z_{hn})$. The symbolic expression ΣZ means the commodity vector $z = \Sigma z_h$ (*all* summations here are over the index *h* for agents). In particular, $\Sigma Z^0 = z^0$, the vector of total endowments of each good; **O** is the vector with zero amounts of every good. $Z^1 \prec\prec Z^2$ means $(z^2 - z^1) \prec\prec \mathbf{O}$, that is, ΣZ^1 is less in every component (good) than ΣZ^2; Z^1 is then *less* than Z^2. Similarly, $Z^1 \leq Z^2$ means that ΣZ^1 is not greater than ΣZ^2 in any component. If $\Sigma Z^1 \leq \Sigma Z^2$ and the number of goods *j* for which $\Sigma z_{hj}^1 = \Sigma z_{hj}^2$ is at least 1 and not *n*, we write $Z^1 < Z^2$.
The notation $Z^1 \succ\succ Z^2$ means $(z^1 - z^2) \prec\prec \mathbf{O}$. In an exchange economy it is natural to assume $z^0 \succ\succ \mathbf{O}$. A distribution Z is *feasible* if $\Sigma Z = z^0$, the use of=here rather than$\leq$implying that free disposal is not assumed; quantities have to be conserved during exchange.
The assumptions made in this section will be maintained throughout.

2 EP-optimality and AS-efficiency

Purely for simplicity the definition of EP-optimality given by Arrow–Hahn (1971, p. 91) is used here, generalized to allow for incompleteness of preferences but specialized to an exchange economy. For compactness, that exchange economy will always be denoted $E(\succsim_k, Z^0)$.

Definition 1 (D1): A distribution Z^1 is *EP-optimal* for $E(\succsim_k, Z^0)$ if: (*a*) it is feasible; and (*b*) there is no other feasible distribution Z that is better than Z^1.

Notice that D1 depends only on the totals z^0 and not on their distribution Z^0. Applying a weaker meaning of being better, namely: that $Z \succ Z^1$, produces a smaller set of allocations that can withstand such tests, the *strongly EP-optimal allocations*. Proofs of the interrelations between this more usual type of EP-optimality and the strong AS-efficiency defined below are similar to those given here, but with more complication, much as the theory of non-negative matrices is basically similar to but more complicated than the theory of positive matrices.

The following definition of AS-efficiency is implicit in the original works of Allais and Scitovsky.

Definition 2 (D2): A distribution Z^2 is *AS-efficient* for $E(\succsim_k, Z^0)$ if: (*c*) it is feasible; and (d) there is no other distribution Z which meets Z^2 and is less than Z^0.

Again, D2 depends only on z^0 and not on Z^0, since 'less than Z^0' actually involves only z^0. As before, a weaker meaning of being less, namely: that $Z \prec Z^2$, produces a smaller set of allocations that can withstand such tests, the *strongly AS-efficient allocations*.

The first special assumption asserts a kind of monotonicity of preference for the society considered as a whole.

Assumption 1 (A1): For any Z and any commodity vector $s \succ\succ \mathbf{0}$ there exists Z^s such that $\Sigma Z^s = \Sigma Z + s$ and $Z^s \succ\succ Z$.

Theorem 1: Assume A1. If Z^1 is EP-optimal then it is AS-efficient.

Proof: This and all other proofs are by contraposition. If Z^1 is not AS-efficient, there exists Z such that $Z \succsim Z^1$ and $\Sigma Z \prec\prec Z^0$. So there is a vector of surpluses in every commodity, i.e. $s = (z^0 - \Sigma Z) \succ\succ \mathbf{0}$. Hence from A1 there is Z^s such that $\Sigma Z^s = \Sigma Z + s = z^0$ and $Z^s \succ\succ Z$. But then $Z^s \succ\succ Z \succsim Z^1$ implies $Z^s \succ\succ Z^1$, and Z^s is feasible. So Z^1 is not EP-optimal.

The second special assumption does not involve the topology of R^n but nevertheless plays the role of a continuity condition on preferences.

Assumption 2 (A2): For any agent h, $z_h^1 \succ_h z_h^2$ implies the existence of $\mu_h \in (0, 1)$ such that $\lambda z_k^1 \succsim_k z_k^2$ for all $\lambda \in [\mu_h, 1)$.

Theorem 2: Assume A2. If Z^2 is AS-efficient then it is EP-optimal

Proof: If not, there exists Z such that $\Sigma Z = z^0$ and $Z \succ\succ Z^2$. From A2 there exists for each z_h in Z some $\mu_h \in (0, 1)$ such that $\lambda z_k \succsim_k z_k^2$ for all $\lambda \in [\mu_h, 1)$. Put μ equal to the maximum of these μ_h, so that $\mu < 1$ and write μZ for the $m \times n$ matrix of the μz_h. Then $\mu Z \succsim Z^2$. But by construction and the fact that $z^0 \succ\succ \mathbf{0}$, $\Sigma \mu Z = \mu \Sigma Z \prec\prec \Sigma Z = z^0$. So Z^2 is not AS-efficient.

3 Cores and compensated cores

The language and notation of Section 1 need modification to cope with cores. A *coalition C* is any non-empty subset of the *m* agents in the economy, and $|C|$ denotes its cardinality, so that $1 \leq |C| \leq m$. A *distribution over C* is the $|C| \times n$ matrix Z_c whose rows are the *n*-vectors z_k for $k \in C$. The notation ΣZ_c means the sum over the $|C|$ rows of Z_c, and for any Z^0 and any *C* we write $z_c^0 = \Sigma Z_c^0$, the total endowments available to the coalition *C*. Given any distribution *Z* for the whole economy and any coalition *C*, the *C*-section of *Z* is the distribution Z_c over *C*.

The notion and language of Section 1 for preferential and quantitative relations between distributions will be applied freely to *C*-sections. But rather than writing $Z_c^1 \succ\succ_c Z_c^3$ and $Z_0^2 \leq_c Z_c^4$ etc., the simpler notation $Z_c^1 \succ\succ Z_c^3$ and $Z_0^2 \leq Z_c^4$ will be used.

Just as in Section 2 with EP-optimality and AS-efficiency, stronger concepts of core and compensated core could be defined and corresponding results proved for them; but that is not done here.

Definition 3 (D3): A distribution Z^1 is in the *core* of $E(\succsim_k, Z^0)$ if: (i) it is feasible; and (ii) there is no coalition *C* and no distribution Z_c over *C* such that (*a*) $\Sigma Z_z = z_c^0$ and (*b*) Z_c is better than the *C*-section Z_c^1 of Z'.

D1 is the special case of D3 in which the only coalition allowed is the whole society, and similarly for D2 and D4, which is given next.

Definition 4 (D4): A distribution Z^2 is in the *compensated core* of Z_c if: (iii) it is feasible; and (iv) there is no coalition *C* and no distribution Z_c over *C* such that (c) Z_c meets Z_c^2 and (d) $\Sigma Z_c \prec\prec z_c^0$.

The rationale for D4 is clearly similar to that for D2. Equally clearly, it is the appropriate ‘mirrored’ version of the core, in which objectives and constraints are interchanged. For example, it is easy to show that any compensated equilibrium of $E(\succsim_k, Z^0)$ is in its compensated core. A much deeper result, for an exchange economy with a continuum of agents, is a ‘compensated’ version of the core equivalence theorem of Aumann (1964), namely: the set of the compensated equilibria is precisely the compensated core (Newman, 1982). Moreover, the assumptions and proof needed for this result are significantly simpler than in the classic paper of Aumann; in particular, only a non-topological separating hyperplane theorem is needed.

Since a coalition can be of any size, from one agent to every agent, the monotonicity of preference for the society as a whole asserted by A1 is quite inadequate to prove interrelations between cores and compensated cores. Instead, we use a more standard monotonicity assumption:

Assumption 3 (A3): For any agent *h*, $z_h^1 \gg z_h^2$ implies $z_h^1 \succ_h z_h^2$.

Theorem 3: Assume A3. If Z^1 is in the core of $E(\succsim_k, Z^0)$ then it is in its compensated core.

Proof: If not there is a coalition *C* and a distribution Z_c over *C* such that $Z_c \succsim Z_c^1$ and $\Sigma Z_c \prec\prec z_c^0$. So for *C* there is a vector s_c of surpluses in every commodity, that is, $s_c = \left(z_c^0 - \Sigma Z_c^3\right) \succ\succ \mathbf{0}$.

Now form a new distribution Z_C^S over C by adding the vector $(|C|^{-1})_S \succ\succ 0$ to each z_k for $k \in C$, and denote the result by z_k^S. Since $z_k^S \succ\succ z_k$, from A3 $z_k^S \succ_k z_k$. Then $Z_C^S \succ\succ Z_C \succsim Z_C^1$ so that $Z_C^S \succ\succ Z_C^1$. Moreover, by construction, $\Sigma Z_C^S = z_C^0$. Hence Z^1 is not in the core.

Theorem 4: Assume A2. If Z^2 is in the compensated core of $E(\succsim_k, Z^0)$ then it is in its core.

Proof: If not, there exists a coalition C and a distribution Z_c over C such that $\Sigma Z_c = z_C^0$ and $Z_c \succ\succ Z_C^2$. The proof then proceeds as in Theorem 2.

4 Conclusion

There is remarkable symmetry between the solution concepts (*a*) and (*b*) on the one hand, and (a) $'$ and (b) $'$ on the other. However, there is a major asymmetry. The concepts (*a*) and (*b*) implicitly give each member of the society a positive weight, that is, each person ‘counts’ for something. Hence, as Edgeworth first observed (1881, p. 23), it is easy to show that a distribution is (strongly) EP-optimal if and only if it maximizes the satisfaction of any agent picked at random, *given* both the total endowments and the levels of satisfaction of the remaining (m−1) agents.

The corresponding statement for strong AS-efficiency is not so obvious. Suppose (and this is Scitovsky's original argument) that we fix the levels of satisfaction of everyone in the society and the total amounts of all but one commodity chosen at random, say z_i. Then it is tempting to say that a distribution is AS-efficient if and only if it minimizes the usage of z_i. The trouble with this is that, in the situation prevailing z_i just might be a commodity that nobody wants. So it is not scarce, its shadow price is zero, and there is no point in trying to economize on its use. Unlike the case with persons, we cannot be sure that a commodity chosen at random will carry positive weight.

The obvious way of dealing with this point is to put sufficient structure on the problem to make z_i always desired. But then it ceases to be an arbitrary commodity, unless all commodities are always so desired; and that is a strong assumption indeed.

Exactly the same difficulty arises of course with efficient production programmes, if they are defined as allocations that maximize the output of an arbitrary product y_j given the supplies of all the factors and the quantities of all products other than y_j. This is really not surprising, since such ‘Pareto-efficiency’ is the analogue in a production economy of AS-efficiency in an exchange economy.

Bibliography

Allais, M. 1943. *A la recherche d'une discipline économique. Première partie, l’économie pure*. Paris: Ateliers Industria.

Allais, M. 1978. Theories of general economic equilibrium and maximum economic efficiency. In *Equilibrium and Disequilibrium in Economic Theory*, ed. G. Schwodiauer. Dordrecht: D. Reidel.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 22, 39–50.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19, 273–92.

Debreu, G. 1959. *Theory of Value*. New York: John Wiley.

Debreu, G. 1962. New concepts and techniques for equilibrium analysis. *International Economic Review* 3, 257–73.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Graaff, J. de V. 1957. *Theoretical Welfare Economics*. Cambridge: Cambridge University Press.

Newman, P. 1982. Compensated cores and the equivalence theorem. Working Papers in Economics No. 112, Department of Political Economy, Johns Hopkins University.

Pareto, V. 1909. *Manuel d'économie politique*. Paris: Giard.

Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70, 1–22.

Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economics* Studies 9, 89–110.

Scitovsky, T. 1964. *Papers on Welfare and Growth*. Stanford, CA: Stanford University Press.

Walras, L. 1874. *Éléments d'économie politique pure*. Lausanne: Corbaz.

How to cite this article

Newman, Peter. "optimality and efficiency." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000035> doi:10.1057/9780230226203.1227

The New Palgrave Dictionary of Economics Online

optimum quantity of money

Timothy S. Fuerst
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The optimum quantity of money is a normative monetary policy conclusion drawn from the long-run properties of a theoretical model. Most famously associated with Milton Friedman, the optimum calls for a zero nominal rate of interest and thus a steady state of price deflation at the long-run real rate of interest. Although this policy prescription has played a minor role in monetary policy implementation, it has had an enormous influence in monetary theory.

Keywords

bargaining; deflation; dynamic new Keynesian models; fiat money; Friedman rule; Friedman, M.; hold-up problem; inflation; monetary policy; optimal taxation; optimum quantity of money; search-theoretic monetary models; seigniorage; transactions role of money

Article

The optimum quantity of money is most famously associated with Milton Friedman (1969). The optimum is a normative policy conclusion drawn from the long-run properties of a theoretical model. Friedman posited an environment that abstracts from all exogenous shocks and nominal price and wage sluggishness. The basic logic is then straightforward. One criterion for Pareto efficiency is that the private cost of a good or service should be equated to the social cost of this good or service. The service in question is the transactions role of money. The social cost of producing fiat money is essentially zero. Since fiat money pays no interest, the private cost of using money is the nominal interest rate. Hence, one criterion for Pareto efficiency is that the nominal interest rate should equal zero. Since long-run real rates are positive, this implies that monetary policy should bring about a steady deflation in the general price level. This famous policy prescription is now commonly called the Friedman rule.
Although most closely associated with Friedman's (1969) bold statement of the policy conclusion, the basic idea of the optimum quantity can be found in Tolley (1957), who argues, on similar efficiency

grounds, for paying interest on currency. Friedman (1960) credits Tolley with this suggestion, and further notes that an alternative policy would be a steady deflation. It is curious that Friedman (1960) dismisses the 'Friedman rule' deflation as not feasible for practical purposes. Finally, the optimum-quantity result is implicit, but never noted, in Bailey (1956) who examines the welfare cost of inflation but does not consider the welfare gain of deflations.

In practice, the optimum-quantity result has had remarkably little influence on monetary policy implementation. Although many central banks pursue low inflation rates with an eventual goal of price stability, no central bank has advocated a policy that would bring about a steady price deflation. There are likely several reasons, both judgemental and theoretical, that have led to this lack of influence. I will briefly review both types of objections.

One of the first theoretical objections to the optimum-quantity results was made by Phelps (1973), who argued that Friedman's first-best argument ignored the second-best fact that money growth produces seigniorage revenues for a government, and that all forms of taxation produce distortions of some kind. If 'money' or 'liquidity' is a good like any other, then familiar optimal taxation arguments would suggest that it should be taxed via a steady inflation. This argument seems all the more persuasive given empirical estimates of a fairly low money demand elasticity.

This public finance approach spawned a very large literature. Important contributions include Kimbrough (1986), Guidotti and Vegh (1993), Correia and Teles (1996; 1999), Chari, Christiano and Kehoe (1996), and Mulligan and Sala-i-Martin (1997). These analyses were much more explicit than Friedman (1969) and considered a fully dynamic theoretical environment with no nominal rigidities. A key relationship in all these models is the transactions or shopping function. The time spent by households shopping (s_t) is a function of the form: $s_t = \Phi(c_t, m_t)$, where c_t denotes real consumption and m_t denotes real cash balances. The function φ is assumed to be homogenous of degree k, increasing in consumption, and decreasing in real cash balances, the latter effect motivated by the transactions function of money. Money can be thought of as an intermediate good that facilitates consumption purchases. Now suppose a central government needs to finance an exogenous level of spending and can do so only with distortionary taxes on, say, labour income, or the inflation tax on money balances. In this case, is the Friedman rule still optimal?

Most of these papers were supportive of the Friedman rule, concluding that in such a second-best environment the optimal monetary policy is a zero nominal rate. Mulligan and Sala-i-Martin (1997) argued that the result was fragile as it depended on the degree of homogeneity in φ and the alternative tax instruments available to the government, for example, income taxes against consumption taxes. These conflicting results have been usefully explained in DeFiore and Teles (2003), who demonstrated that the reason for the divergent conclusions is an inappropriate specification of how consumption taxes are entered in the transactions cost function. They consider a more general environment in which the government has access to both consumption and income taxes. They also consider the case where money is costly to produce at a constant marginal cost of α. Further, they demonstrate that if φ is linearly homogenous (k=1) then the optimal interest rate is equal to α. This is a modified Friedman rule in that the private cost and social cost of money are set equal to each other, and is analogous to the Diamond and Mirrlees (1971) optimal taxation result: intermediate goods should not be taxed when consumption taxes are available and the technology is constant returns to scale (k=1). If φ is not linearly

homogeneous, then the optimal policy involves a tax (or subsidy) on money proportional to α. Since money is essentially costless to produce ($\alpha=0$) the optimal nominal interest rate is zero. DeFiore and Teles (2003) thus conclude that the Friedman rule is the optimal second-best policy for all homogeneous transactions technologies. Hence, the Phelps (1973) objection appears to be settled in Friedman's favour.

A second theoretical objection to the optimum-quantity result is that, in a world with nominal rigidities, a steady general price deflation would produce unwanted relative price movements since not all nominal prices would be adjusted simultaneously. Strictly speaking this is not a theoretical objection to Friedman (1969), as he assumed a world with perfectly flexible nominal prices and wages. But if one believes that nominal rigidities are important, and that they matter even in the long run, then this is a relevant objection to the Friedman rule. For example, in the dynamic new Keynesian (DNK) class of models (for example, Woodford, 2003) the assumed nominal rigidities have permanent effects so that any departure from price stability causes permanent movements in relative prices. Hence, these models typically suggest that optimal policy is a stable price level, and that a Friedman-rule deflation would be suboptimal. These DNK models typically abstract from the nominal interest rate distortions that are at the heart of the optimum-quantity result. A model that combined the DNK nominal rigidities with the nominal rate distortion would presumably result in a long-run optimal nominal interest rate somewhere between zero and the steady-state real rate.

The principle judgemental objection to the Friedman rule is historical. The instances in US history in which deflations occurred are associated with severe recessions, most famously in the 1929–33 period. A related judgemental concern deals with the zero bound. If the central bank's principal tool to stimulate the economy is a reduction in the nominal rate of interest, then the zero nominal rate prescribed by the Friedman rule apparently leaves no additional ammunition in the monetary policy arsenal (as nominal rates cannot be negative). This nervousness about the Friedman rule was enhanced by the experience of Japan during the 1990s. The Japanese economy performed poorly at a time in which general prices were falling and the short-term nominal rate was zero.

Since central banks have not followed Friedman's (1969) proposal to set the nominal rate to zero, a natural issue is to quantify the welfare costs of being away from Friedman's optimum quantity of money. Following in the footsteps of Bailey (1956), Lucas (2000) uses a theoretical environment similar to that of Correia and Teles (1996; 1999) to address this question. The welfare cost is approximately the area underneath the money demand curve between the optimal zero nominal rate and the interest rate under question. Lucas reports that the welfare cost of a four per cent nominal rate is between 0.2 per cent and one per cent of annual income, the difference depending upon the assumed behaviour of money demand as the nominal rate approaches zero. Since a zero nominal rate has not been observed in the United States in the post-Second World War period, the data cannot determine which estimate is more accurate. But either estimate suggests a fairly modest welfare cost.

Studies analysing the optimality of the Friedman rule have been reignited by the new class of search-theoretic monetary models. These models are micro-based, replacing the function φ in DeFiore and Teles (2003) with a search-based trading environment in which money improves the chances of successfully finding a suitable partner with whom to trade. In an innovative paper, Lagos and Wright (2005) use a search-theoretic environment to address the optimality of the Friedman rule and the welfare consequences of deviating from it. In search models of money the buyer and seller engage in a

bargaining game to determine the transactions price at a given meeting. The buyer is carrying money and has thus postponed previous consumption. If sellers have some bargaining power, then there is a hold-up problem because part of the gain associated with the holding of money is received by the seller. This bargaining distortion leads the buyers to economize on money holdings so that they are below the socially efficient level. Lagos and Wright (2005) demonstrate that the optimal policy in this search environment is the Friedman rule (a similar conclusion is reached by Shi, 1997). But more interestingly, the welfare cost of being away from the Friedman rule, at say a four per cent nominal rate, is significantly higher than calculated by Lucas (2000). This arises because the positive nominal rate exacerbates an already suboptimal level of real balances arising from the hold-up problem.

The search models of money have rekindled interest in the optimality of the Friedman rule at just the time when DeFiore and Teles (2003) appear to have settled the issue in the aggregative monetary models. The coming years will probably see further work on the Friedman rule from this search-theoretic perspective. A key issue is the nature of the bargaining process that arises at trading opportunities. These recent developments testify to the continued prominence of the optimum quantity of money in monetary theory, if not practice. The lasting contribution of the theory is to introduce explicit, utility-based welfare analysis into monetary economics.

See Also

- Friedman, Milton
- monetary policy, history of
- money and general equilibrium
- real bills doctrine

The author would like to thank Charles Carlstrom and John Hoag for their helpful comments.

Bibliography

Bailey, M.J. 1956. The welfare costs of inflationary finance. *Journal of Political Economy* 64, 93–110.

Chari, V.V., Christiano, L.J. and Kehoe, P. 1996. Optimality of the Friedman rule in economies with distorting taxes. *Journal of Monetary Economics* 37, 202–23.

Correia, I. and Teles, P. 1996. Is the Friedman rule optimal when money is an intermediate good. *Journal of Monetary Economics* 38, 223–44.

Correia, I. and Teles, P. 1999. The optimal inflation tax. *Review of Economic Dynamics* 2, 325–46.

DeFiore, F. and Teles, P. 2003. The optimal mix of taxes on money, consumption and income. *Journal of Monetary Economics* 50, 871–88.

Diamond, P.A. and Mirrlees, J.A. 1971. Optimal taxation and public production. *American Economic Review* 63, 8–27.

Friedman, M. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, M. 1969. *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

Guidotti, P.E. and Vegh, C.A. 1993. The optimal inflation tax when money reduces transactions costs. *Journal of Monetary Economics* 31, 189–205.

Kimbrough, K.P. 1986. The optimum quantity of money rule in the theory of public finance. *Journal of Monetary Economics* 18, 277–84.

Lagos, R. and Wright, R. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 46, 463–84.

Lucas, R.E. Jr. 2000. Inflation and welfare. *Econometrica* 68, 247–74.

Mulligan, C.B. and Sala-i-Martin, X. 1997. The optimum quantity of money: theory and evidence. *Journal of Money, Credit and Banking* 29, 687–715.

Phelps, E.S. 1973. Inflation in the theory of public finance. *Swedish Journal of Economics* 75, 37–54.

Shi, S. 1997. A divisible search model of fiat money. *Econometrica* 65, 75–102.

Tolley, G. 1957. Providing for growth of the money supply. *Journal of Political Economy* 65, 465–85.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Fuerst, Timothy S. "optimum quantity of money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000040> doi:10.1057/9780230226203.1228

The New Palgrave Dictionary of Economics Online

options (new perspectives)

Thaleia Zariphopoulou
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides an overview of risk-neutral valuation methodology and presents historical milestones in the development of quantitative finance. It also discusses current challenges and new perspectives in model choice, pricing and hedging.

Keywords

arbitrage; Bachelier, L.; Brace-Gatarek-Musiela model; barrier options; call options; continuous-time models; copulas; credit default obligations; credit default swaps; credit risk; derivatives; exotics; Heath–Jarrow–Morton model; hedging; incomplete markets; Libor; martingales; model calibration; model specification; option valuation; options; real options; reduced-form models of default; risk measures; risk-neutral pricing; risk-neutral valuation; stochastic integration theory; structural models of default; swap market models; term structure models; value at risk; vanilla options; volatility smile; yield curve

Article

In 1973, Black, Scholes and Merton developed a method for the valuation of a European option based on the idea of perfect replication of its payoff. Their approach demonstrates how to act in an uncertain environment so that relevant risks are controlled. Around the same time, trading of options on common stocks started in the Chicago Board Options Exchange. Theory met practice and an exciting and fruitful journey started on the crossroads of economics, finance and mathematics. Its impact was phenomenal in both academia and industry. New areas of research were created, and numerous educational and training activities were established. The derivatives market grew at an unprecedented rate and influenced the development of other markets. Complex mathematical modelling and technical sophistication, predominant elements in theory and applications in engineering and natural sciences, now entered the theory and practice of finance. This was not the first time that stochastic modelling touched finance. At the beginning of the 20th century, in his pioneering doctoral work, Bachelier (1900) proposed a

stochastic model, based on normality assumptions on their returns, for stock prices. In many aspects, however, his work was ahead of its time and had no impact for years to come.

What was the Black, Scholes and Merton option valuation approach? A European call option is a contract that gives its owner the right to buy the underlying stock at a given price, *K* and a given maturity, *T*. Their model, powerful and simple, assumed a liquid market environment consisting of a non-defaultable bond and a stock. The bond yields constant interest rate *r*, while the stock price, S_t, is modelled as a log-normal diffusion process having constant mean rate of return, μ, and volatility parameter, σ. Applying Ito's formula – a fundamental result of modern stochastic calculus – they were able to build a dynamic self-financing portfolio, (α_t, β_t), $0 \leq t \leq T$, that replicates the option payoff, that is, for which $\alpha_T + \beta_T = (S_T - K)^+$. For all *t*, the option price, ν_t, is, then, given by the current portfolio value, $v_t = \alpha_t + \beta_t$. Stochastic and differential arguments yield the price process representation $v_t = C(S_t, t)$, with the function *C* satisfying the partial differential equation

$$C_t + \frac{1}{2}\sigma^2 S^2 C_{SS} + rSC_S = rC \tag{1}$$

and the terminal condition $C(S, T) = (S - K)^+$. The components of the replicating portfolio turn out to be $\alpha_t = S_t C_S(S_t, t)$ and $\beta_t = C(S_t, t) - \alpha_t$, representing the amounts invested, respectively, in the stock and bond.

The construction of the price and hedging policies, as well as the specification of various sensitivity indices (greeks), thus amount to solving linear partial differential equations, a relatively easy task given the existing technical body in mathematical analysis.

The industry rapidly adopted the Black and Scholes model as a standard for the valuation of simple (vanilla) options. Soon after, more complex products were created and traded, like options on fixed-income securities, currencies, indices and commodities. Gradually, the options market experienced great growth and its liquidity reached very high levels (for a concise exposition see, for example, Musiela and Rutkowski, 2005).

In parallel, substantial advances in research took place. In 1979, Harrison and Kreps laid the foundations for the development of the risk-neutral pricing theory. They created a direct link between derivative valuation and martingale theory. For a finite number of traded securities and under general assumptions on their price processes and related payoffs, they established that the price of a replicable contingent claim corresponds to the expected value, calculated under the risk-neutral probability of the (discounted) claim's payoff. These results were further developed and presented by Harrison and Pliska (1981). In the years that followed, the theory was extended and a model-independent approach for pricing and risk management emerged. In a generic derivatives model, the (discounted) prices of primary assets are represented by a vector-valued semi-martingale $\mathbf{S}_S = (S_S^1, \ldots, S_S^m)$, defined in a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ where $\mathbb{P}$ is the historical measure. The (discounted) payoff, C_T, is taken to be an $\mathcal{F}_T$-

measurable random variable.

The derivative price, discounted under the same numeraire as $\boldsymbol{S}$ and C_T, is given by the conditional expectation

$$v_t(C_T) = E_{\mathbb{Q}}(C_T|\mathcal{F}_t). \tag{2}$$

The pricing measure $\mathbb{Q}$ is equivalent to $\mathbb{P}$ and, under it, the (discounted) price processes become martingales, that is, $E_{\mathbb{Q}}(S_s|\mathcal{F}_t) = S_t$, $t \leq s \leq T$. The derivative prices, themselves martingales under $\mathbb{Q}$, are linear with respect to their payoffs, time and numeraire consistent and independent of their holder's risk preferences.

Fundamental questions in risk-neutral valuation are related to existence and uniqueness of the derivative price. Uniqueness turns out to be equivalent to the replicability of all claims in the market. Such a market is classified as complete. Stochastic integration theory was used to establish that market completeness is equivalent to uniqueness of the risk-neutral martingale measure $\mathbb{Q}$. In this case, the price is given by (2) and, thus, exists and is unique. If, however, the market is not complete there is multiplicity of equivalent martingale measures. In this case, perfect replication is abandoned and absence of arbitrage becomes the key requirement for price specification and model choice. In an arbitrage-free model, a judicious choice of the pricing measure is made and the price is still represented as in (2). In many aspects, market completeness and absence of arbitrage are complementary concepts. Their relationship has been extensively studied with the use of martingale theory and functional analysis. Important results in this direction are formulated in the First and Second Fundamental Theorems of Asset Pricing (see, among others, Bjork, 2004; Delbaen and Schachermayer, 2006).

The risk-neutral valuation theory, built on a surprising fit between stochastic calculus and quantitative needs, revolutionized the derivatives industry. But its impact did not stop there. Because the theory provides a universal approach to price and manage risks, the option pricing methodology has been applied in an array of applications. Indeed, corporate and non-corporate agreements have been analysed from an options perspective. Option techniques have also been applied to the valuation of pension funds, government loan guarantees and insurance plans. In a different direction, applications of the theory resulted in a substantial growth of the fields of real options and decision analysis. Complex issues related, for example, to operational efficiency, financial flexibility, contracting, and initiation and execution of research and development projects were revisited and analysed using derivative valuation arguments (see the review article of Merton, 1998).

Since the 1970s, theoretical developments, technological advances, modelling innovations and creation of new derivatives products have been proceeding at a remarkable rate. During this period, theory and practice have been shaping each other in a unique challenging and intense interaction. The rest of the article is, mainly, dedicated to this dimension.

Theory and practice in derivatives markets

The Black and Scholes model included various assumptions that are not valid in practice. Interest rates and volatilities are not constant, trading is not continuous, defaults occur and information is not complete. How did academic research and industry reality react to and handle these issues? Albeit there are very distinct priorities, needs and goals, shortcomings of the theory not only did not limit its applicability but prompted a remarkable progress between the theoretical and the applied worlds. Models were developed and innovative computational techniques were invented, and used in practice, for new complex products (exotics). Progress did not occur simultaneously. While theory developed mostly in bursts, practice continued the use of basic models which often involved self-contradictory assumptions. However, despite internal modelling inconsistencies, industry applications offered valuable intuition and feedback to the abstract theoretical developments.

The first revisited assumption was that the (short) interest rate is constant. Models of stochastic interest rates started appearing, and a major breakthrough occurred in 1992 with the work of Heath, Jarrow and Morton. Moving away from modelling directly the short rate, their novel approach was focused on the dynamics of the entire (instantaneous and continuously compounded) forward curve $f(t, T)$, defined by

$$f(t, T) = -\frac{\partial}{\partial T}\ln B(t, T)$$

where $B(t, T)$ represents the price, at time *t*, of a zero-coupon discount bond with maturity *T*. To facilitate the analysis of the forward curve, Musiela (1993) introduced an alternative parametrization, namely, $r(t, x) = f(t, t + x)$, which exhibited the importance of infinite dimensional diffusions and stochastic partial differential equations in finance. This helped to find answers to a number of practical questions related to the yield curve dynamics. Indeed, the issue of consistency between the yield curve construction and its evolution was resolved. Additionally, the support of the yield curve distribution has been studied and the mean reversion, or, more mathematically, stationarity of the entire yield curve dynamics has been addressed.

Clearly, the infinite dimensional analysis was useful in a study of the dynamics of the forward rates for all maturities. There was, however, still a problem that needed to be looked at, namely, that the forward rates $f(t, T)$ are not traded in the market, and the Libor and swap rates are together with options on them. Moreover, information contained in these option prices should be taken into account in the specification of the yield curve dynamics. Because the market trades caps and swaptions in terms of their Black and Scholes volatilities, it would be advantageous to develop a term structure model that is consistent with such practice, a task seen by many academics at that time, as impossible for its apparent internal inconsistency.

In a series of papers by Miltersen, Sandmann, Sondermann, Brace, Gatarek, Musiela, Rutkowski and Jamshidian (see Part II of Musiela and Rutkowski, 2005, for a detailed exposition of these works), a new modelling framework for term structure dynamics was put in place. The so-called Libor, also known as BGM (Brace–Gatarek–Musiela), and swap market models resolved the outstanding issue of the link between the traded instruments and the mathematical description of their dynamics. In essence, they provided a model-independent framework for the analysis of the interest rates dynamics when coupled

with the advances – taking place in parallel – in the modelling of volatility smile dynamics. The latter issue is discussed next.

The Black and Scholes model assumes constant volatility and hence, within this model, a call option with arbitrary strike is priced with the same volatility. However, call options of different strikes are priced differently by the market which ‘allocates’ into the Black and Scholes formula a strike-dependent volatility generating the so-called volatility smile. This is clearly inconsistent with the assumption of the model. It turns out, however, that a complete collection of call prices, for all strikes and maturities, uniquely determines the one-dimensional distributions of the underlying forward price process, under a probability measure which should be interpreted as a forward measure to the option maturity. In a series of papers, Dupire (1993) shows how to construct martingale diffusions with a given set of one-dimensional distributions, demonstrating, once more, that the market practice is theoretically sound and internally consistent when analysed from the perspective of the appropriate model. The Black and Scholes model is used only to convert the quoted volatility into a price and it is no longer used for the pricing of vanilla options. Moreover, there are many ways of constructing martingales with a given set of one-dimensional marginals, and the question is not so much how to construct one but, rather, which one to choose and under which criteria. The important message here is that, again, one can now look at the problem in a completely model-independent way, provided all objects – namely, the underlying assets, the associated probability measures and the relevant market information – are correctly interpreted.

Obviously, the theory and practice, at least in the equity, foreign exchange and interest rates derivatives markets, have moved to a different level and reached a certain degree of maturity. Of course, important challenges remain but experience since the 1970s defines clearly a path to follow.

Current challenges and perspectives

Credit risk

A fundamental assumption of the Black and Scholes model is that the underlying securities do not default. However, default is a realistic element of financial contracts and very relevant to any firm's performance. Credit-linked instruments have, by now, become a central feature in derivatives markets. These are financial products that pay their holders amounts contingent on the occurrence of a default event ranging from bankruptcy of a firm to failure to honour a financial agreement. Examples include, among others, credit default swaps (CDS), credit default obligations (CDO) and tranches of indices. Their market has grown more than eightfold in recent years and, undoubtedly, credit risk is, today, one of the most active and challenging areas in quantitative finance.

There are various issues that make the problems in credit risk difficult, from both the modelling and the implementation point of view. The first challenge is how to model the time of default. In academic research, there are two well-established approaches, the structural and the reduced. In the structural models, it is postulated that uncertainty related to default is exclusively generated by the firm's value. Modelling default, then, amounts to building a good model for the company's assets and determining when the latter will fall below existing liabilities. However, such default times are, typically, predictable which is not only unrealistic, but, also, difficult to implement due to limited public information about the firm's prospects. In the other extreme, the reduced-form models, the default time is associated with a

point process with an exogenously given stochastic jump intensity. The intensity essentially measures the instantaneous likelihood of default. Reduced models are more tractable for pricing and calibration but the default times are completely sudden (totally inaccessible), a non-realistic feature. Recently, efforts have been made to bridge the two approaches by incorporating the limited information the investors might have about the firm's value. This information-based approach is gradually emerging but a number of modelling and technical serious issues remain to be tackled. See, among others, Bielecki and Rutkowski (2002) and Schönbucher (2003).

Even though the above models are theoretically sound, their practical implementation is so difficult that it makes them, effectively, inapplicable. The main problem stems from the high dimensionality and inability to develop computational methods that track ‘name by name’ the valuation outputs. For this reason, the focus in the industry has shifted to an alternative direction centred on modelling the joint distribution of default times. An important development in this direction is the use of a copula function, a concept introduced in statistics by Sklar (1959). The aim is to define the joint distribution of a family of random variables when their individual marginal distributions are known. Such marginal distributions may be, frequently, recovered from the market, as is the case with CDS that yield implicit information on the underlying name's default time. Today, the most widely used copula is the one-factor Gaussian one, proposed by Li (2000). Its popularity lies in the ability to obtain the sensitivity, and thus information on hedging, of the derivative price in a name by name correspondence.

Model specification

As has been mentioned earlier, the theory has long departed from perfect replication, and practice never relied on it. Absence of arbitrage is the underlying pricing criterion in the derivatives market. However, a plethora of pricing issues and model specifications arise every day. Derivatives markets have been growing very rapidly, and high liquidity in vanilla options on a large number of underlyings including, among others, single stocks and equity indices, interest rates, foreign exchange and commodities, has been achieved. The users benefit from competitive prices, quoted at very tight spreads, for the protection they need. This, in itself, brings another challenge to the providers of such services and products, namely, the models that are currently under development need to reflect this liquidity before they can be used for the pricing of less liquid products. This process is known in the industry as model calibration. To a large extent, one can assume that the market gives the prices for simple derivatives like calls and puts and, hence, pricing considerations dissolve. However, more exotic options need to be priced and this must be done in a way consistent with the basic products (vanilla).

To provide some intuition, consider the case of the so-called first generation exotic, namely, a down and out call option. This is a barrier option that reduces to a simple call option when the likelihood of crossing the barrier is very small. Consequently, a model to price such an option must return the market price of a call in such a scenario. Call prices will be liquid for all strikes up until a certain maturity, say, 18 months or two years for currency options. However, there may be a need to price products with embedded currency options of very long maturity, like up to 50 years in dollar–yen exchange rate. In this case, a suitable model needs to be developed that accommodates short- and long-term issues. On one hand, the model must fit the short-dated foreign exchange (FX) calls and puts. On the other, it has to be consistent with the interest rates volatilities and must capture correctly the dependence structure

between the dollar and yen interest rates curves, their volatilities and the spot FX.

A standard approach for solving such problems consists of writing a continuous-time model and trying to fit it to the liquid prices. This task is often very difficult to complete. Indeed, as more market information must be put into a model, the more complicated the model gets, the more difficult and time consuming the calibration procedure becomes, and the more time it takes to produce accurate prices and stable sensitivity reports. To a large extent, model calibration is identical to the specification of one-dimensional distributions of the underlying process. Model specification, on the other hand, can be identified with the specification of an infinite dimensional copula function defining the joint distribution of the entire path, given the marginal distributions that can be deduced from the call prices. At this point, it is important to recall that, often, option payoffs depend solely on a finite dimensional distribution of the underlying process. Consequently, the need to specify the continuous-time dynamics remains valid only if one wants to link the concept of price with perfect replication of the payoff, a requirement that is, in any case, not met in practice.

Seen from this perspective, a new modelling path emerges, namely, one can take the marginals as given by the call prices and choose a copula function in such a way that the joint distribution is consistent with an arbitrage-free model. For example, if one wants to price a forward start option, the distributions of the underlying asset at two different dates are given. Then, only the joint distribution needs to be specified but in such a way that the martingale property is preserved. Clearly, there is an infinite number of ways to build such a martingale, and the choice should be based on additional information – for example, not on the smile as seen today but on the assumptions one might want to make about the smile dynamics.

Risk measures

As was previously discussed, absence of arbitrage is the fundamental ingredient in derivative pricing. Absence of perfect replication remains, however, a major issue and dictates the creation of financial reserves. To this effect, regulatory policies have been in place for few years now.

These requirements prompted the axiomatic analysis of the so-called risk measures, which are nonlinear indices yielding the capital requirement of financial positions. The theory of coherent risk measures was proposed by Artzner et al. (1999). A popular risk measure is the ‘value at risk’, which, despite its widespread use, neither promotes diversification nor measures large losses accurately. Since the mid-1990s a substantial research effort has been invested in further developing the theory. Relaxing a scaling assumption in the coherent case has led to the development of convex risk measures. The next step has been the axiomatic construction of dynamic risk measures that are time consistent, an indispensable property of any pricing system.

See Also

- hedging
- options

Bibliography

Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. 1999. Coherent measures of risk. *Mathematical Finance* 9, 203–28.

Bachelier, L. 1900. Théorie de la spéculation. Ph.D. dissertation L' Ecole Normale Supérieure. English translation in *The Random Character of Stock Market Prices*, ed. P.H. Cootner. Cambridge, MA: MIT Press.

Bjork, T. 2004. *Arbitrage Theory in Continuous Time*, 2nd edn. Oxford: Oxford University Press.

Bielecki, T.R. and Rutkowski, M. 2002. *Credit Risk: Modeling, Valuation and Hedging*. Berlin: Springer.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Delbaen, F. and Schachermayer, W. 2006. *The Mathematics of Arbitrage*. Berlin: Springer.

Dupire, B. 1993. Pricing and hedging with a smile. *Journées Internationales de Finance*, IGR–AFFI, La Baule.

Harrison, M. and Kreps, D.M. 1979. Martingales and arbitrage in multi-period security markets. *Journal of Economic Theory* 20, 381–408.

Harrison, M. and Pliska, S. 1981. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Applications* 11, 215–60.

Heath, D., Jarrow, R.A. and Morton, A. 1992. Bond pricing and the term structure of interest rates: a new methodology for contingent claim valuation. *Econometrica* 60, 77–105.

Li, D.X. 2000. On default correlation: a copula function approach. *Journal of Fixed Income* 9, 43–54.

Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–83.

Merton, R. 1998. Applications of option-pricing theory: twenty-five years later. *American Economic Review* 88, 323–49.

Musiela, M. 1993. Stochastic PDEs and term structure models. *Journées Internationales de Finance*. IGR-AFFI: La Baule.

Musiela, M. and Rutkowski, M. 2005. *Martingale Methods in Financial Modelling*, 2nd edn. Berlin:

Springer.

Schönbucher, P.J. 2003. *Credit Derivatives Pricing Models. Model, Pricing and Implementation*, Chichester: Wiley.

Sklar, A. 1959. Fonctions de répartition à n dimensions et leures marges. *Publications de l'Institut de Statistique de l' Université de Paris* 8, 229–31.

How to cite this article

Zariphopoulou, Thaleia. "options (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000104> doi:10.1057/9780230226203.1230

The New Palgrave Dictionary of Economics Online

options

Robert C. Merton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An option is a security whose owner has a right to buy (sell) it at a specified price on a specified date (or, with an American-type option, on or before the specified date). Trading of options on common stock began in 1973 and has since spread to other commodities. Option pricing theory provides a unified theory for the pricing of corporate liabilities. Of its more recent extensions, perhaps the most significant is its application in the evaluation of operating or ‘real’ options in the capital budgeting decision problem.

Keywords

arbitrage; Bachelier, L.; Black, F.; capital budgeting; deposit insurance; forward contracts; futures contracts; government loan guarantees; Ito's Lemma; Merton, R. C.; option pricing theory; options; pension fund insurance; probability; Scholes, M.

Article

A ‘European-type call (put) option’ is a security that gives its owner the right to buy (sell) a specified quantity of a financial or real asset at a specified price, the ‘exercise price’, on a specified date, the ‘expiration date’. An American-type option provides that its owner can exercise the option on or before the expiration date. If an option is not exercised on or before the expiration date, it expires and becomes worthless.

Options and forward or futures contracts are fundamentally different securities. Both provide for the purchase (or sale) of the underlying asset at a future date. A long position in a forward contract obliges its holder to make an unconditional purchase of the asset at the forward price. In contrast, the holder of a call option can choose whether or not to purchase the asset at the exercise price. Thus, a forward contract can have a negative value whereas an option contract never can.

The first organized market for trading options was the Chicago Board Options Exchange (CBOE) which

began trading options on common stocks in 1973. The initial success of the CBOE was followed by an expansion in markets to include options on fixed-income securities, currencies, stock and bond indices, and a variety of commodities. Although these markets represent an increasingly larger component of total financial market trading, options are still relatively specialized financial securities. Option pricing theory has, nevertheless, become one of the cornerstones of financial economic theory.

This central role for options analysis derives from the fact that option-like structures pervade virtually every part of the field. Black and Scholes (1973) provide an early example: shares of stock in a firm financed in part by debt have a payoff structure which is equivalent to a call option on the firm's assets where the exercise price is the face value of the debt and the expiration date is the maturity date of the debt. Option pricing theory can thus be used to price levered equity and, therefore, corporate debt with default risk.

Identification of similar isomorphic relations between options and other financial instruments has led to pricing models for seniority, call provisions and sinking fund arrangements on debt; bonds convertible into stock, commodities, or different currencies; floor and ceiling arrangements on interest rates; stock and debt warrants; rights and stand-by agreements. In short, option pricing theory provides a unified theory for the pricing of corporate liabilities.

The option-pricing methodology has been applied to the evaluation of noncorporate financial arrangements including government loan guarantees, pension fund insurance and deposit insurance. It has also been used to evaluate a variety of employee compensation packages including stock options, guaranteed wage floors, and even tenure for university faculty.

Perhaps the most significant among the more recent extensions of option analysis is its application in the evaluation of operating or ‘real’ options in the capital budgeting decision problem. For example, a production facility which can use various inputs and produce various outputs provides the firm with operating options that it would not have with a specialized facility which uses a fixed set of inputs and produces a single type of output. Option-pricing theory provides the means of valuing these production options for comparison with the larger initial cost or lower operating efficiency of the more flexible facility. Similarly, the choice among technologies with various mixes of fixed and variable costs can be treated as evaluating the various options to change production levels, including abandonment of the project. Research and development projects can be evaluated by viewing them as options to enter new markets, expand market share or reduce production costs.

As these examples suggest, option analysis is especially well suited to the task of evaluating the ‘flexibility’ components of projects. These, corporate strategists often claim, are precisely the components whose values are not properly measured by traditional capital-budgeting techniques. Hence, option-pricing theory holds for the promise of providing quantitative assessments for capital budgeting projects that heretofore were largely evaluated qualitatively. Survey articles by Smith (1976) and Mason and Merton (1985) provide detailed discussion of these developments in option analysis along the extensive bibliographies.

The lineage of modern option pricing theory began in 1900 with the Sorbonne thesis, ‘Theory of Speculation’, by the French mathematician Louis Bachelier. The work is rather remarkable because, in analysis the problem of option pricing, Bachelier derives much of the mathematics of probability diffusions; this, five years before Einstein's famous discovery of the theory of Brownian motion. Although, from today's perspective, the economics and mathematics of Bachelier's work are flawed, the connection of his research with the subsequent path of attempts to describe an equilibrium theory of

option pricing is unmistakable. It was not, however, until nearly 75 years later with the publication of the seminal Black and Scholes article (1973), that the field reached a sense of closure on the subject and the explosion in research on option pricing applications began.

As with Bachelier and later researchers, Black and Scholes assume that the dynamics for the price of the asset underlying the option can be described by a diffusion process with a continuous sample path. The breakthrough nature of the Black–Scholes analysis derives from their fundamental insight that the dynamics trading strategy in the underlying asset and a default-free bond can be used to hedge against the risk of either a long or short position in the option. Having derived such a strategy, Black and Scholes determine the equilibrium option price from the equilibrium condition that portfolios with no risk must have the same returns as a default-free bond. Using the mathematics of Ito stochastic integrals, Merton (1973, 1977) formally proves that with continuous trading, the Black–Scholes dynamic portfolio will hedge all the risk of an option position held until price exercise or expiration, and therefore, that the Black–Scholes option price is necessary to rule out arbitrage.

Along the lines of the derivation for general contingent claims pricing in Merton (1977), a sketch of the arbitrage proof for the Black–Scholes price of a European call option on a nondividend-paying stock in a constant interest rate environment is as follows.

Assume that the dynamics of the stock price, $V(t)$, can be described by a diffusion process with a stochastic differential equation representation given by:

$$dV = \alpha V dt + \alpha V dz \tag{1}$$

where α is the instantaneous expected return on the stock; σ^2 is the instantaneous variance per unit time of the return, which is a function of V and t; dz is a standard Wiener process. Let $F[V, t]$ satisfy the linear partial

$$0 = \frac{1}{2}\sigma^2 V^2 F_{11} + rVF_1 - rF + F_2 \tag{2}$$

where subscripts denote the partial derivatives and r is the interest rate. Let F be such that it satisfies the boundary conditions:

$$F / V \leq 1; \quad F(0, t) = 0; \quad F[V, T] = \max[0, V - E] . \tag{3}$$

Note from (3) that the value of *F* on these boundaries are identical to the payoff structure on a European call option with exercise price *E* and expiration date *T*. From standard mathematics, the solution to (2) and (3) exists and is unique.

Consider the continuous-time portfolio strategy which allocates the fraction $w(t) \equiv F_1[V, t]V(t)/P(t)$ to the stock and 1–*w*(*t*) to the bond, where *P*(*t*) is the value of the portfolio at time *t*. Other than the initial investment in the portfolio at there are no contributions or from the portfolio until it is liquidated at $t = T$.

The prescription for the portfolio strategy for each time *t* depends only on the first derivative of the solution to (2)–(3) and the current values of the stock and the portfolio. It follows from the prescribed allocation *w*(*t*) that the dynamics for the value of the portfolio can be written as:

$$dP = w(t)P\,dV/V + [1 - w(t)]rP\,dt = F_1 dV + r[P - F_1 V]dt. \tag{4}$$

As a solution to (2), *F* is twice-continuously differentiable. Hence, we can use Ito's Lemma to express the stochastic process for *F* as:

$$dF = \left[\frac{1}{2}\sigma^2 V^2 F_{11} + \alpha V F_1 + F_2\right]dt + F_1 \sigma V dz \tag{5}$$

where *F* is evaluated at $V = V(t)$ at each point in time *t*. But, *F* satisfies (2). Hence, we can rewrite (5) as:

$$dF = F_1 dV + r[F - F_1 V]dt. \tag{6}$$

Define *Q*(*t*) to be the difference between the value of the portfolio and the value of the function F[*V*, *t*] evaluated at $V = V(t)$. From (4) and (6), we have that $dQ = rQdt$ which is a nonstochastic differential equation with solution $Q(t) = Q(0)\exp[rt]$ and $Q(0) = P(0) - F[V(0), 0]$. Hence, if the initial investment in the portfolio is chosen so that $P(0) = F[V(0), 0]$ then $Q(t) \equiv 0$ and $P(t) = F[V(t), t]$ for all *t*.

Thus, we have constructed a dynamic portfolio strategy in the stock and a default-free bond that exactly replicates the payoff structure of a call option on the stock. The solution of (2) and (3) for *F* and its first

derivative F_1 provides the 'blueprint' for that construction. The standard no-arbitrage condition for equilibrium prices holds that two securities with identical payoff structures must have the same price. It follows, therefore, that the equilibrium price of the call option at time t must equal the Black–Scholes price, $F[V(t), t]$.

The extraordinary impact of the Black–Scholes analysis on financial economic research and practice can in large part be explained by three critical elements: (1) the relatively weak assumptions for its valid application; (2) the variables and parameters required as inputs are either directly observable or relatively easy to estimate, and there is computational ease in solving for the price; (3) the generality of the methodology in adapting it to the pricing of other options and option-like securities.

Although framed in an arbitrage type of analysis, the derivation does not depend on the existence of an option on the stock. Hence, the Black–Scholes trading strategy and price function provide the means and the cost for an investor to create synthetically an option when such an option is not available as a traded security. The findings that the equilibrium option price is a twice continuously differentiable function of the stock price and that its dynamics follow an Ito process are derived results, not assumptions.

The striking feature of (2) and (3) is not the variables and parameters that are needed for determining the option price but rather, those not needed. Specifically, determination of the option price and the replicating portfolio strategy does not require estimates of either the expected return on the stock, α or investor risk preferences and endowments. In contrast to most equilibrium models, the pricing of the option does not depend on price and joint distributional information for all available securities. The only such information required is about the underlying stock and default-free bond. Indeed, the only variable or parameter required in the Black–Scholes pricing function that is not directly observable is the variance rate function, σ^2. This observation has stimulated a considerable research effort on variance-rate estimation in both the academic and practising financial communities.

With some notable exceptions, equations (2) and (3) cannot be solved analytically for a closed-form solution. However, powerful computational methods have been developed to provide high-speed numerical solutions of these equations for both the option price and its first derivative.

As in the original Black and Scholes article, the derivation here focuses on the pricing of a European call option. Their methodology is, however, easily applied to the pricing of other securities with payoff structures contingent on the price of the underlying stock. Consider, for example, the determination of the equilibrium price for a European put option with exercise price E and expiration date T. Suppose that in the original derivation we change the boundary conditions specified for F in (3) so as to match the payoff structure of the put option on these boundaries. That is, we now require that F satisfy $F \leq E$; $F[0, t] = E\exp[-r(T-t)]$; $F[V, T] = \max[0, E-V]$. Once F and its derivative are specified, the development of the replicating portfolio proceeds in identical fashion to show that $P(t) = F[V(t), t]$. With the revised boundary conditions, the portfolio payoff structure will match that of the put option at exercise or expiration. Thus, $F[V(t), t]$ is the equilibrium put option price.

As shown in Merton (1977), the same procedure can be used to determine the equilibrium price for a security with a general contingent payoff structure, $G[V(T)]$, by changing the boundary conditions in (3) so that $F[V, T] = G[V]$. A particularly important application of this procedure is in the determination of pure state-contingent prices.

Let $\pi[V, t; E, T]$ denote the solution of (2) subject to the boundary conditions:

$$\pi / V \leq 1; \quad \pi[0, t; E, T] = 0; \quad \pi[V, T; E, T] = \delta(E - V)$$

where $\delta(x)$ is the Dirac delta function with the properties that

$$\delta(x) = 0 \text{ for } x \neq 0$$

and $\delta(0)$ is infinite in such a way that

$$\int_a^b \delta(x)\mathrm{d}x = 1 \text{ for } a < 0 < b.$$

By inspection of this payoff structure, it is evident that this security is the natural generalization of Arrow–Debreu pure state securities to an environment where there is a continuum of states defined by the price of the stock and time. That is loosely, $\pi[V, t; E, T]\mathrm{d}E$ is the price of a security which pays $1 if $V(T) = E$ at time *T* and $0, otherwise.

As is well known from the Green's functions method of solving differential equations, the solution to equation (2) subject to the boundary condition $F[V, T] = G[V]$ can be written as:

$$F[V, t] = \int_0^\infty G[E]\,\pi[V, t; E, T]\mathrm{d}E. \tag{7}$$

Thus, just as with the standard Arrow–Debreu model, once the set of all pure state-contingent prices, $\{\pi\}$ are derived, the equilibrium price of any contingent payoff structure can be determined by mere summation or quadrature.

To underscore the central importance of call option pricing in the general theory of contingent claims pricing, consider a portfolio containing long and short positions in call options with the same expiration date *T* where each ‘unit’ contains a long position in an option with exercise price $E - \varepsilon$; a long position in an option with exercise price $E + \varepsilon$; and a short position in two options with exercise price *E*. If one takes a position in $1/\varepsilon^2$ units of this portfolio, the payoff structure at time *T* with $V(T) = V$ is given by:

$$\{\max[0, V+\varepsilon-E] - 2\max[0, V-E] + \max[0, V-\varepsilon-E]\} / \varepsilon^2. \tag{8}$$

The limit of (8) as $\varepsilon \to 0$ is $\delta(E-V)$ which is the payoff structure to a pure contingent-state security. If $F[V, t; E, T]$ is the solution to (2) and (3), then it follows from (8) that:

$$\pi[V, t; E, T] = \lim_{\varepsilon \to 0} \{F[V, t; E-\varepsilon, T] - 2F[V, t; E, T] + F[V, t; E+\varepsilon, T]\} / \varepsilon^2 = \frac{\partial^2 F[V, t; E, T]}{\partial E^2}. \tag{9}$$

Hence, once the call-option pricing function has been determined, the pure state-contingent prices can be derived from (9).

For further discussion of options, see especially the January/March 1976 issue of the *Journal of Financial Economics*; the October 1978 issue of the *Journal of Business*; and the excellent book by Cox and Rubinstein (1985).

See Also

- Bachelier, Louis
- finance
- options (new perspectives)

Bibliography

Bachelier, L. 1900. *Théorie de la speculation*. Paris: Gauthier-Villars. English translation in *The Random Character of Stock Market Prices*, ed. P. Cootner, revised edn, Cambridge, MA: MIT Press, 1967, 17–78.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–59.

Cox, J. and Rubinstein, M. 1985. *Options Markets*. Englewood Cliffs, NJ: Prentice-Hall.

Mason, S. and Merton, R.C. 1985. The role of contingent claims analysis in corporate finance. In *Recent Advances in Corporate Finance*. ed. E.I. Altman and M.G. Subrahmanyan. Homewood, Illinois: Richard D. Irwin, 7–54.

Merton, R.C. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, Spring, 141–83.

Merton, R.C. 1977. On the pricing of contingent claims and the Modigliani–Miller theorem. *Journal of Financial Economics* 5, 241–50.

Smith, C.W. 1976. Option pricing: a review. *Journal of Financial Economics* 3(1/2), 3–51.

How to cite this article

Merton, Robert C. "options." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000044> doi:10.1057/9780230226203.1229

The New Palgrave Dictionary of Economics Online

orderings

C. Blackorby

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

choice; compensation criteria; happiness; interpersonal utility comparisons; ordering; preference orderings; preferences; transitivity; utility measurement; welfare economics; well-being

Article

An ordering (also called a complete preordering or a weak ordering) is a binary relation which is reflexive, transitive and complete, that is, it is a preordering that is complete.
A binary relation R defined on a set S is a set of ordered pairs of elements of S, that is, a subset of the Cartesian product of S with itself, $S\times S$. One writes xRy (or $(x, y)\in R$) to mean that $x\in S$ stands in relation R to $y\in S$. An ordering is a binary relation, R, which satisfies three properties: (i) reflexivity: for all $x\in S$, xRx; (ii) transitivity: for $x, y, z\in S$, if xRy and yRz, then xRz; and (iii) completeness for all $x, y\in S$, xRy or yRx, where ‘or’ is used in its non-exclusive sense.
A simple example results from letting S be the real line and R the greater than or equal to relation so that xRy if and only if $x\geq y$. The most common use of orderings in economics is in preference theory where S is a commodity space and R stands for ‘at least as desirable as’. Every ordering can be separated into its symmetric and asymmetric factors, respectively, as follows:

$$xIy \text{ if and only if } xRy \text{ and } yRx$$

and

$$xPy \text{ if and only if } xRy \text{ and not } yRx.$$

In the case of preference theory, these correspond to indifference and strict preference relations.
In consumer theory orderings first appeared in the work of Wold (1943–4). In an attempt to put utility theory on a more solid foundation, Wold posited the existence of an ordering with certain properties and demonstrated that this could be represented by a continuous real-valued function, thus making absolutely clear that this was an ordinal concept. Perhaps the most innovative and useful aspect of Wold's argument was an insightful definition of a continuous ordering. (An ordering is continuous if the sets $x|xRy,y\in S$ and $x|yRy,y\in S$ are closed.)
The first modern treatment of the subject appears in Arrow (1951). Agents as well as society as a whole are characterized by their orderings over spaces of alternative. That the choices of society be consistent with an ordering, and understanding the implications of that requirement, has been particularly important in welfare economics. For example, various compensation criteria have been shown to fail transitivity (see Gorman, 1955) and hence be unsuitable for public decision-making. In addition, by representing agents and society by their orderings, Arrow made the first step toward unravelling a long-standing confusion between the measurability of utility on the one hand and interpersonal comparability on the other. This step was critical if social decision-making was to rest on solid ground; for an accessible discussion of these issues see Blackorby, Donaldson and Weymark (1984).
It is common in economics to represent agents by their preference orderings. This leads to a set of complicated and somewhat unresolved issues: what are the relationships among the notions of preference, choice and happiness or well-being. Either a preference ordering or the choices of an individual may be viewed as a primitive and they may or may not be mutually consistent; the issues at stake can, however, be characterized quite precisely. The relationship between either of these and some notion of happiness or well-being is much less clear; for a good introduction to these problems see Sen and Williams (1982).

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley. 2nd edn, 1963.

Blackorby, C., Donaldson, D. and Weymark, J. 1984. Social choice with interpersonal utility comparisons: a diagrammatic introduction. *International Economic Review* 25, 327–56.

Gorman, W. 1955. The intransitivity of certain criteria used in welfare economics. *Oxford Economic Papers* 7, 25–35.

Sen, A. and Williams, B., eds. 1982. *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Wold, H. 1943–4. A synthesis of pure demand analysis, I–III. *Skandinavisk Aktuarietidskrift* 26, 85–118; 220–63; 27, 69–120.

How to cite this article

Blackorby, C. "orderings." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000048> doi:10.1057/9780230226203.1231

The New Palgrave Dictionary of Economics Online

Oresme, Nicholas (1325 – 1382)

Barry Gordon

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Aquinas, St. Thomas; Biel, G.; Buridan de Bethune, J.; debasement; Langenstein, H. von; money as standard of value; Oresme, N.; William of Ockham

Article

A noted mathematician and physicist and Bishop of Lisieux (1377–82), Oresme was a close friend and adviser of Charles V of France. Because of the frequent currency debasements of his era, he contributed a most influential treatise on money, *Tractatus de origine natura jure, et mutationibus monetarum* (*c*1360). Other works by Oresme of interest to economists are his commentaries on Aristotle's politics, economics, and ethics.

Oresme's tract on money owes much to the ideas of Jean Buridan de Bethune, who was Rector of the University of Paris (*c*1327) and may have taught Oresme. Another to take up Buridan's line of monetary thought was Heinrich von Langenstein (1325–97), a German theologian who taught at both Paris and Vienna. In the next century, the doctrines of Buridan and Oresme were developed by Gabriel Biel (1430–95), a founder of the University of Tübingen and its first professor of theology (from 1484). Biel wrote the outstanding *Tractatus de Potentate et Utilate Monetarum*.

Each of the foregoing Schoolmen wrote in the nominalist tradition deriving from the Oxford philosopher William of Ockham (1285–1347), a tradition that stands in opposition to St Thomas Aquinas on money as on much else. The nominalists question the Thomistic understanding of money as a standard of value established by the Prince (that is, by the ruler of the state). In their view, the Prince's right with respect to setting the standard is a limited right.

According to Oresme and the other nominalists, who are reacting against the princely practice of debasement, a particular currency is likely to be an effective medium of exchange only if the nominated values of the units of that currency are acceptable to the citizens who are the users of the medium. They add that the users of the money are the real owners of it, and so have the right to be consulted by the Prince concerning appropriate arrangements.

Such ideas were revolutionary in terms of much earlier Western thought. One notable aspect of the

revolution is the shifting of the grounds for thinking about money. Earlier scholastic discussion had concentrated on the morality of individual transactions but here the operation of a monetary system as a whole begins to come into view.

Selected works

1956. *The De Moneta of Nicholas Oresme and English Mint Documents*. Introduction and translation by C. Johnson. London and New York: Nelson.

Bibliography

Biel, G. 1930. *Treatise on the Power and Utility of Moneys*. Trans. R.B. Burke, Philadelphia: University of Pennsylvania Press.

Estrup, H. 1966. Oresme and monetary theory. *Scandinavian Economic History Review* 14(2), 97–116.

Gordon, B. 1975. *Economic Analysis before Adam Smith*. London: Macmillan.

Langholm, O. 1983. *Wealth and Money in the Aristotelian Tradition*. Bergen: Universitetsforlaget.

Menut, A.D. 1957. Introduction and English translation to ‘Maistre Nicole Oresme: Le Livre de Yconomique d'Aristotle’. *Transactions of the American Philosophical Society*, N.S. 47, 783–853.

Monroe, A.E. 1923. *Monetary Theory Before Adam Smith*. Cambridge, MA: Harvard University Press.

Noonan, J.T. 1957. *The Scholastic Analysis of Usury*. Cambridge, MA: Harvard University Press.

How to cite this article

Gordon, Barry. "Oresme, Nicholas (1325–1382)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000050> doi:10.1057/9780230226203.1232

The New Palgrave Dictionary of Economics Online

Organization of the Petroleum Exporting Countries (OPEC)

James L. Smith
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since the 1960s, the Organization of the Petroleum Exporting Countries (OPEC) has dominated the world oil market by exercising physical control over a large portion of the world's oil reserves. Coordinated production restraint among OPEC members has artificially limited the supply of oil and succeeded in pushing oil prices far above the competitive level. Despite its past success, OPEC faces three basic problems that, in the long run, tend to undermine all cartels: coordination failures, opportunistic cheating, and the entry of competing producers who manage to find and bring alternative supplies to the market.

Keywords

barriers to entry; cartels; cheating; coordination; Cournot oligopoly; entry; free-rider problem; Organization of Petroleum Exporting Countries (OPEC); Prisoner's Dilemma; Stackelberg dominant-firm models.

Article

The Organization of the Petroleum Exporting Countries (OPEC), an international cartel of oil-producing states, affects the price of nearly all crude oil traded in the world economy and has done so since the early 1970s.

Founded in 1960, OPEC initially consisted of five member states (Iran, Iraq, Kuwait, Saudi Arabia and Venezuela) which together accounted for 38 per cent of total world production of crude oil. The founders sought to coordinate national petroleum policies and forge a more united front in dealings with the multinational oil companies that operated within their borders. Although membership has grown to 12, OPEC's share of global crude oil production still amounts to only about 44 per cent. Coordinated restraints on output (especially since 1973) have deliberately held OPEC's market share in check.

During its first decade (1960–70), OPEC's principal objective was to secure for its members a larger share of the profits derived from the production and sale of their oil – the stated goal being to raise government take from 50 per cent to 80 per cent of total profit. Beginning with the so-called Teheran–Tripoli Agreements of 1970–71, OPEC turned to what has become its main purpose: manipulating the level of world oil prices by restricting productive capacity and output. Initially, this was attempted without assigning individual production quotas to the respective members. Only after the downturn in world oil prices that began in 1982 did OPEC introduce a formal system of production allocations – which remained in force as of 2007. The members meet at regular intervals (and sometimes on an emergency basis) to review market conditions and adjust individual production ceilings as needed to maintain a target price. Adelman (1995) and Parra (2004) describe the intriguing economic and political challenges faced by the members of OPEC in dealing with the market and with each other.

There is no question that OPEC members have restricted production in ways that are unrelated to the physical scarcity of oil. Even though OPEC's proved oil reserves in 2007 were double those of 1973, the cartel initiated sharp output cuts that by 1985 had removed nearly half of their previous production from the market, as shown in Figure 1. Not until 2005 did OPEC production regain (barely) the level of 1973. Over that same period, worldwide consumption of crude oil grew by 50 per cent and production from non-OPEC producers (who faced much higher marginal costs) managed to increase by 70 per cent.

Figure 1

OPEC Production and reserves, 1973–2005. *Sources*: Production, U.S. Energy Information Administration. Reserves, *Oil & Gas Journal*.

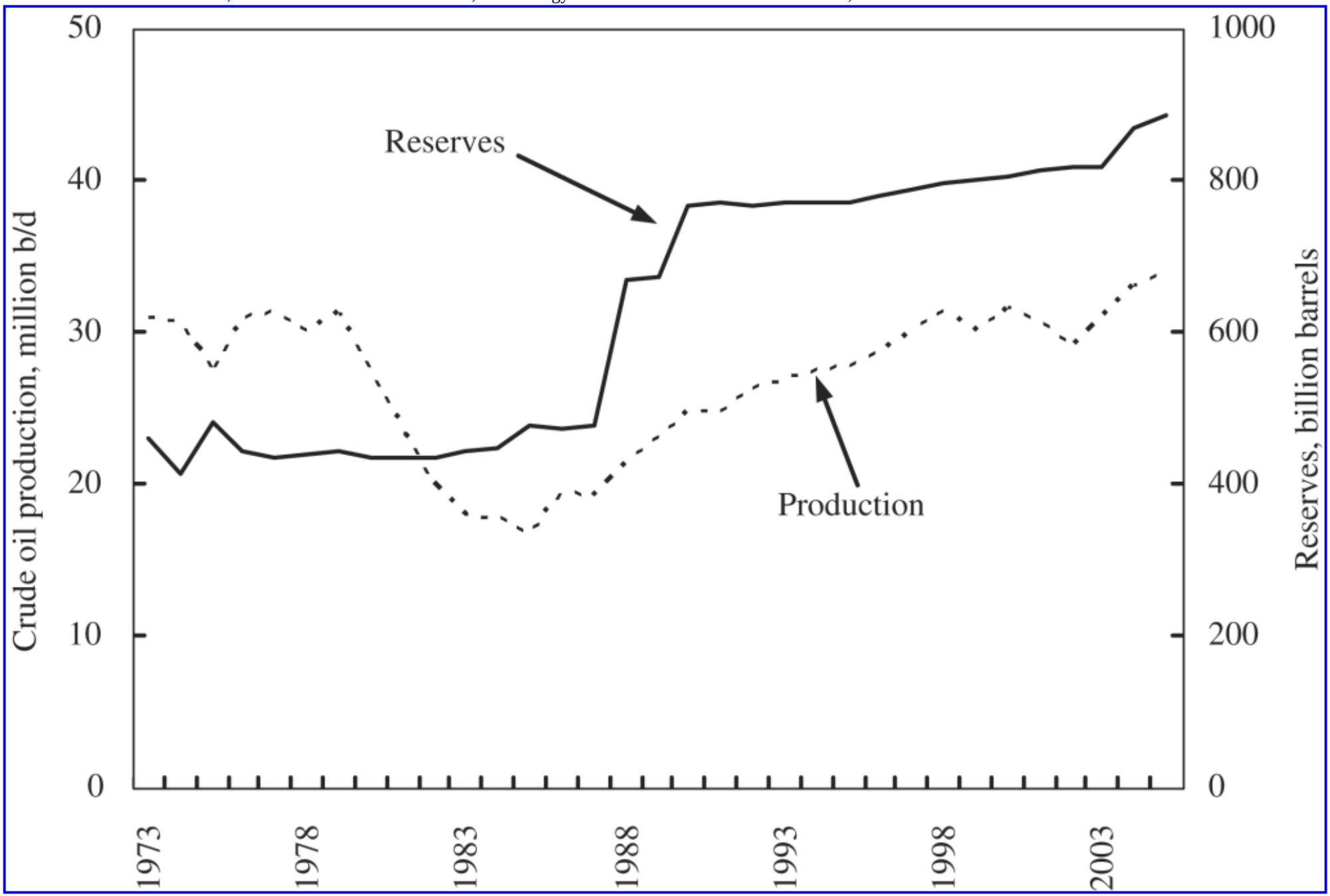

Economic models of OPEC behaviour

Early economic analyses of OPEC behaviour questioned whether the output reductions might reflect competitive or other forms of non-cooperative conduct (for example, oligopoly),

as opposed to outright collusion. Mead (1979) and Johany (1980) proposed a ‘property rights’ explanation that linked the production cuts to the wave of nationalizations that swept through the global oil industry in the early 1970s. Property rights in oil reserves were transferred, via nationalizations, from the multinational corporations (with higher presumed discount rates) to OPEC states (with lower presumed discount rates and therefore greater patience in extracting the oil). However, this explanation is belied by the fact that, throughout the 1960s, these same host governments had repeatedly exhorted the multinational companies to increase, not decrease, their rates of production (Adelman, 1982).

Teece (1982) and Crémer and Salehi-Isfahani (1980) advanced the idea that the limited domestic revenue needs (‘absorptive capacity’) of some OPEC members imposed an indirect restriction on production. The higher the price, the lower the volume of oil exports required to achieve a requisite amount of revenue. The result would be a backward-bending supply curve that links lower oil output to higher prices in a manner that implies no coordination among OPEC members. One problem with this argument, as Adelman (1982) pointed out, is that the absorptive capacities of OPEC members seemed to increase faster than export revenues. Griffin's (1985) subsequent empirical tests found little statistical support for the target revenue hypothesis.

Distinguishing between the various models of OPEC behaviour has been complicated by the fact that cooperative and non-cooperative models share many similar predictions. Thus, the same body of evidence has been interpreted in ways that are consistent with a variety of competing models. By focusing on one aspect of producer behaviour (short-run reactions to cost shocks) that more clearly distinguishes between models, Smith (2005) found a degree of parallelism among OPEC producers that can be accounted for only as the result of cooperative behaviour, not competition or mere interdependence among producers, as in the Cournot oligopoly or Stackelberg dominant-firm models.

Future challenges facing OPEC

Levenstein and Suslow (2006) identify three critical problems that any cartel must solve if it is to endure: coordination, cheating and entry. In the case of OPEC, the last of these has been the easiest. OPEC is protected by barriers to entry that stem from ownership and control of low-cost oil reserves. Roughly 75 per cent of the world's proved reserves of crude oil are located in OPEC nations. Additional reserves are discovered and developed each year, but this process has become increasingly difficult and expensive – even more so outside OPEC than within. Thus, production of crude oil from non-OPEC sources does expand when the cartel cuts production and pushes prices up, but the scope for this is limited and will remain so.

The problem of cheating has been more difficult for OPEC. Any system of output restraints is vulnerable to the free-rider problem. Although OPEC as a whole may benefit by restricting total output, individual members are tempted to produce beyond their assigned quotas. Cartel membership is most beneficial to those members who do *not* cut production. Without a system to detect and punish cheating, the cartel is hampered by a Prisoner's Dilemma in which the dominant strategy for most, if not all, members is to ignore their assigned quotas.

It is common, as in Gately (2004), to distinguish between ‘core’ (low cost, high compliance) and ‘non-core’ (high cost, low compliance) members of OPEC. In fact, compliance with the quota by members of both groups has been sporadic, as shown in Figure 2. Since the inception of the formal quota system in 1983, total OPEC production of crude oil through 2005 has exceeded the ceiling by four per cent on average, but on numerous occasions the excess has run to 15 per cent or more. In general, full compliance has been achieved only during episodes (like 2005–6) when the production ceiling itself tested the limits of each member's available production capacity, such that cheating was not feasible.

Figure 2

OPEC compliance with the production ceiling, 1983–2005. *Sources*: Ceilings, *OPEC Annual Statistical Bulletin*. Actual production, U.S. Energy Information Administration.

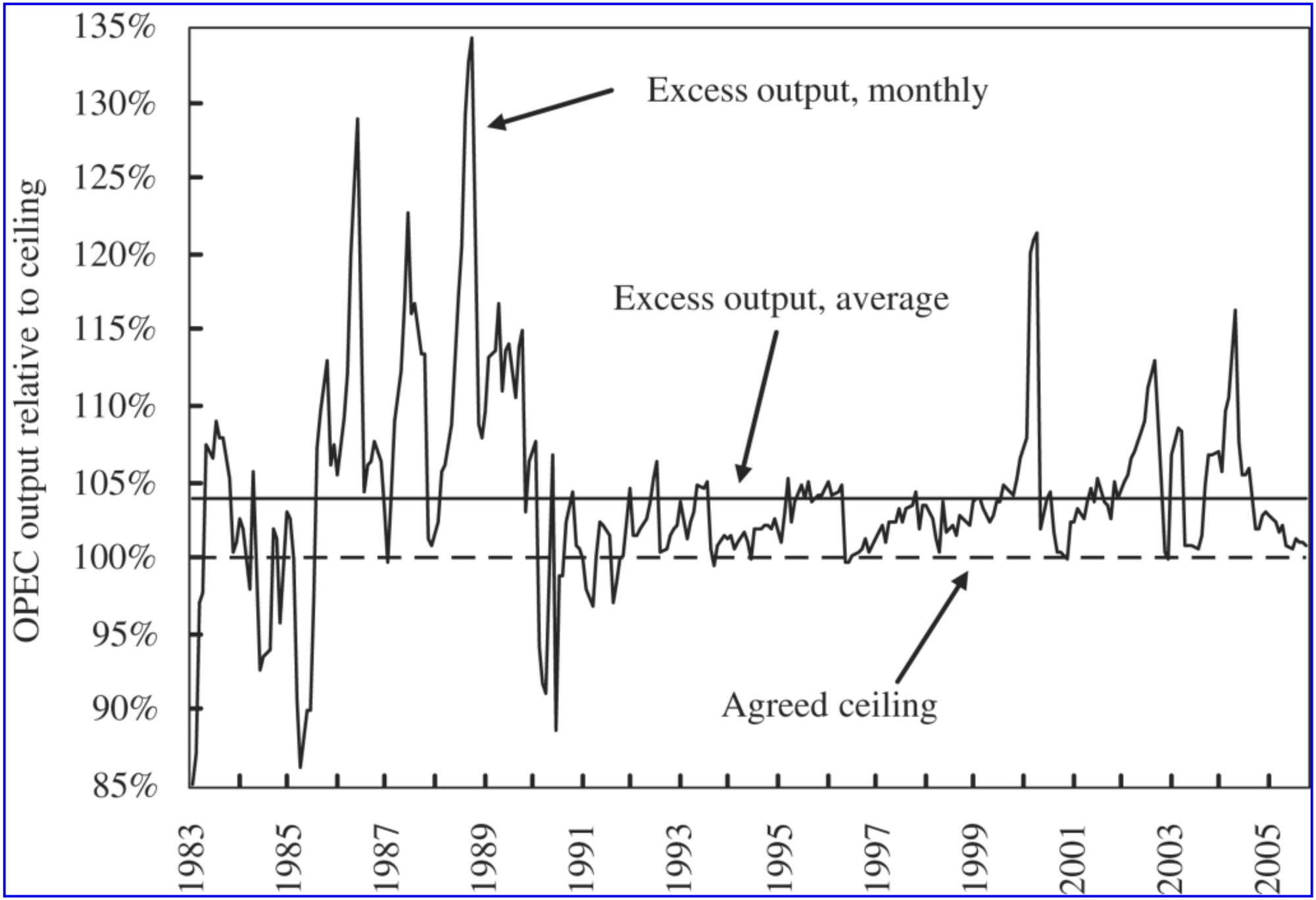

The third problem – coordination among members – presents further difficulties. Due to economic and demographic heterogeneity, the interests of individual OPEC members do not naturally align behind a single ‘correct’ price or production target. In part this is due to the fact that OPEC has limited means by which to redistribute earnings among members. Therefore, any given set of quotas determines not only the overall profit of OPEC but also the individual revenues that accrue to each member. Moreover, coordination requires agreement not only about how aggregate output is parcelled out to individual members, but also about the amount of oil to be produced by OPEC in total. Members with low-cost, long-lived reserves may be more reluctant to have OPEC pursue severe output cuts since too-high prices would induce technological development and new forms of energy (or energy

conservation) that will eventually compete with OPEC. Members that possess smaller reserves and shorter horizons are less affected by this and may prefer deeper production cuts. Internal divisions between ‘price hawks’ and ‘price doves’ have been observed previously and will likely surface within OPEC again.

In terms of longevity, OPEC is already far beyond the mean lifetime (five years) of contemporary international cartels (Levenstein and Suslow, 2006). In terms of economic impact, it is sufficient to note that crude oil is among the most valuable commodities exchanged in international trade, with total daily receipts in 2007 in excess of $1 billion. Thus, by exerting even a small impact on the market price, the cartel effects an enormous transfer of wealth between consumers and producers of crude oil, and creates a substantial allocative inefficiency of the type that arises whenever the price of a product deviates from its marginal cost. As of 2007, no one has attempted to reckon the full magnitude of welfare losses that may be associated with OPEC's manipulation of the world oil market.

See Also

- cartels
- concentration measures

Bibliography

Adelman, M.A. 1982. OPEC as a Cartel. In *OPEC Behavior and World Oil Prices*, ed. J.M. Griffin and D.J. Teece. London: George Allen and Unwin.

Adelman, M.A. 1995. *The Genie Out of the Bottle: World Oil Since 1970*. Cambridge, MA: MIT Press.

Crémer, J. and Salehi-Isfahani, D. 1980. A theory of competitive pricing in the oil market: what does OPEC really do? Working Paper No. 80-4, CARESS, University of Pennsylvania.

Gately, D. 2004. OPEC's incentives for faster output growth. *The Energy Journal* 25(2), 75–96.

Griffin, J.M. 1985. OPEC behavior: a test of alternative hypotheses. *American Economic Review* 75, 954–63.

Johany, A.D. 1980. *The Myth of the OPEC Cartel*. New York: John Wiley & Sons.

Levenstein, M.C. and Suslow, V.Y. 2006. What determines cartel success? *Journal of Economic Literature* 44, 43–95.

Mead, W.J. 1979. The performance of government in energy regulation. *American Economic Review* 69, 352–6.

Parra, F. 2004. *Oil Politics: A History of Modern Petroleum*. London: I.B. Taurus.

Smith, J.L. 2005. Inscrutable OPEC? Behavioral tests of the cartel hypothesis. *The Energy Journal* 26(1), 51–82.

Teece, D.J. 1982. OPEC behavior: an alternative view. In *OPEC Behavior and World Oil Prices*, ed. J.M. Griffen and D.J. Teece. London: George Allen & Unwin.

How to cite this article

Smith, James L. "Organization of the Petroleum Exporting Countries (OPEC)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000103> doi:10.1057/9780230226203.1233

The New Palgrave Dictionary of Economics Online

Ortes, Giammaria (1713 – 1790)

Ugo Rabbeno
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

division of labour; free exchange; Malthus, T. R.; mathematical method; money; Ortes, G.; population growth; wealth

Article

A Venetian monk, Ortes left his cloister on the entreaties of his mother after his father's death, but remained in holy orders and was ever a strenuous defender of the clergy. It is with this purpose that he wrote his *Errori popolari intorno al-l'Economia nazionale*, his *Lettere sulla religione* and his treatise *Dei Fide-commessi a famiglie e a chiese*, with the scope of upholding the existence of clerical property in Mortmain.
In his *Economia nazionale* (vols xxi, xxii, and xxiii, of Custodi's *Scrittori classici italiani di economia politica*, Milan, 1802–1816) Ortes endeavours to demonstrate that as

> the wealth of a nation is determined by the (previous) wants of its members, the riches of one of them cannot increase unless at the expense of another one; the bulk of existing riches is in each nation measured by its wants, and cannot by any means whatever exceed this measure. (*Discorso preliminare*)

From this rather startling proposition, Ortes, who certainly was an original thinker, deduces the condemnation of the principles on which mercantilism was based.

> Money is only a sign of wealth, and must never be considered as being wealth itself. The error of those who mistake money for wealth, proceeds from a confusion between the equivalent of a thing and the thing itself, or between two equivalents which they consider as identical things, although they are not. (ch. ix)

In his *Riflessioni sulla popolazione* (Venice, 1790, and vol. xxiv of Custodi) Ortes controverts the prevailing opinion that an increase of population must necessarily increase the wealth of a nation, and maintains that ‘in any nation whatever the population is compelled to keep within fixed limits, which are invariably determined by the necessity of providing for its subsistence’ (*Prefazione*). In his very first chapter he asserts that, if natural instincts were allowed full play, population would increase in a *geometrical* progression (doubling every 30 years), and calculates that a group of 7 persons composed of three old people, two young men and two young women of 20, would be the ancestors at the end of 150 years of 224 living persons.

150 years of	224 living persons
300 years of	7, 1688 living persons
450 years of	229, 376 living persons
900 years of	7, 516, 192, 768 living persons

Sheer violence keeps down the numbers of animals within the necessary limits, but among men, ‘generation is limited by reason’ (ch. iii), especially by voluntary celibacy, which affords Ortes an occasion of extolling the provident discipline of the Roman Catholic Church. Ortes is a harbinger of Malthus; first by his law of the geometrical increase of population, and secondly by the influence which he ascribes to human reason as a prudential check against over-population.

Ortes was a fervent mathematical student, and expresses himself in algebraical formulae in his *Calcolo sopra il Valore delle Opinioni umane* (vol. xxiv, Custodi). In the same work he illustrates his meaning by curves, which, if not actually traced, are at least minutely described.

Edward Cannan

Ortes is undoubtedly the most eminent of the Venetian economists of the 18th century; his genius, original and sometimes paradoxical, is often opposed to the general tendency of the ideas of his time, and though his researches are occasionally faulty in their method, he has left a deep impress on the history of economic theory. He regards economic laws as immutable, like those of nature; he maintains this in opposition to the opinion usually accepted in his time, which regarded economics only in relation to special interests. Perhaps it is this idea which leads him to distrust the action of the state, considering it is not adapted to promote the wealth of a country.

While Ortes applied a mathematical method to economics, his arguments are based throughout on abstract theory, disregarding the study both of facts and of history as not appertaining to economic science. This detracts from the value of his labours. Still his works are of weight in the history of economic theory. He did not adopt the doctrines of the Physiocrats, and he also recognizes the importance of division of labour, and the important place taken by production in economic theory. Contrary to the prevailing ideas of his day, Ortes upholds universal free exchange.

Selected works

n.d. *Calcolo sopra il valore delle opinioni umane*.

1771. *Errori popolari intorno all'economia nazionale*.

1790. *Reflessioni sulla popolazione delle nazioni per rapporto all'economia nazionale*. Venice.

1802–16. Many works in *Scrittori classici italiani di economia politica*, ed. P. Custodi, Milan.

How to cite this article

Rabbeno, Ugo. "Ortes, Giammaria (1713–1790)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000056> doi:10.1057/9780230226203.1234

The New Palgrave Dictionary of Economics Online

output fall – transformational recession

Barry W. Ickes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A significant decline in GDP has been a common feature in transition economies. This sharp drop in output has been seen as a surprise and puzzle to many observers. Understanding the nature of the output fall is crucial to understanding transition. Analysis is complicated by measurement issues associated with moving from plan to the market. Theoretical models of the output fall are examined, including those that see the output fall as a natural consequence of the legacies of the Soviet-type economic system.

Keywords

arrears; asset specificity; command economy; coordination problems; double marginalization model; hold-up problem; incomplete contracts; incomplete information; monopoly; national income measurement; output fall in transition economies; over-industrialization; planning; price controls; price liberalization; privatization; search frictions; second economy; technical complementarities; trade dependency; uncertainty problems; value destruction; wage rigidity

Article

A significant output decline has been a common feature in transition economies. To some extent this is a surprise: transition represents the removal of (highly significant) distortions. See, for example, Blanchard (1997, p. v): ‘The fact that the transition came with an often large initial decrease in output should be seen as a puzzle. After all, the previous economic system was characterized by a myriad of distortions. One might have expected that removing most of them would lead to a large increase, not a decrease, in output.’ Or as Svejnar (2000, p. 8) notes, ‘The depth and length of the early transition depression was unexpected.’ Similarly, Robert Mundell has written:

> The first and most obvious conclusion is that output contracted by a cumulative percentage never before experienced in the history of capitalist economies (at least in peacetime). Early denials that the contractions were occurring have proved to be incorrect. We observe that cumulative contractions over the 1990–4 period ranged widely, from a low of 18% to a high of more than 80%. (Mundell, 1997, pp. 97–8)

Hence, a simple neoclassical argument would predict that output would rise rather than fall as the transition starts. Yet output fell in each transition economy, and quite significantly. The *officially reported* cumulative output decline for 26 transition economies from 1989 to 1995 was 41 per cent. Of this, the average decline in central Europe was 28 per cent and in the former Soviet Union it was 54 per cent (Fischer and Sahay, 2000, Table 1). By comparison, output in the United States during the Great Depression declined by 34 per cent. The ubiquitous nature of the output fall thus represents an important puzzle for transition economics, and understanding the causes and nature of the output fall is crucial.
Analysis of the output fall is complicated by important measurement issues. In the change of economic systems from plan to market, the valuation of goods and services changes dramatically. This makes it important to distinguish official measures of the output fall from welfare-based measures.

Stylized facts

It is useful to begin with some stylized facts about the output fall. Official GDP measures of output are given in Figures 1 and 2. It is evident from these figures that in all transition economies output follows a U-shaped pattern. This represents another interesting puzzle. A theory based on the chaotic nature of the collapse of planning might predict that output would collapse at the start of transition, but would rise from that point. The pattern displayed by the transition economies, on the other hand, suggests that the peak output fall occurs with a lag of several years. So an additional part of the puzzle is to explain why the output fall intensifies in the early transition.

Figure 1

Official GDP growth in central and eastern Europe. *Source*: International Monetary Fund Dataset.

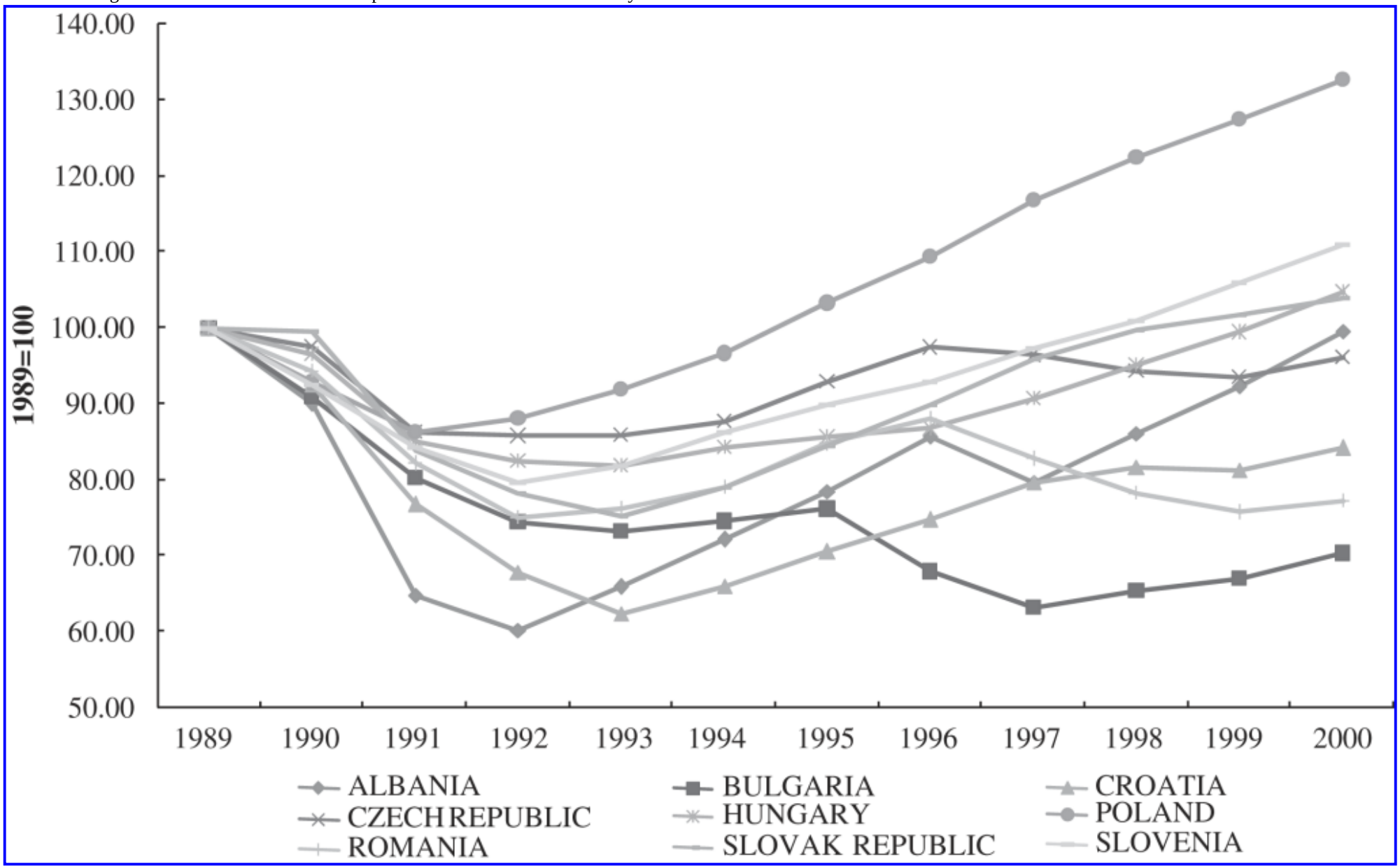

Figure 2

GDP in the former Soviet Union, 1989–2000. *Source*: International Monetary Fund Dataset.

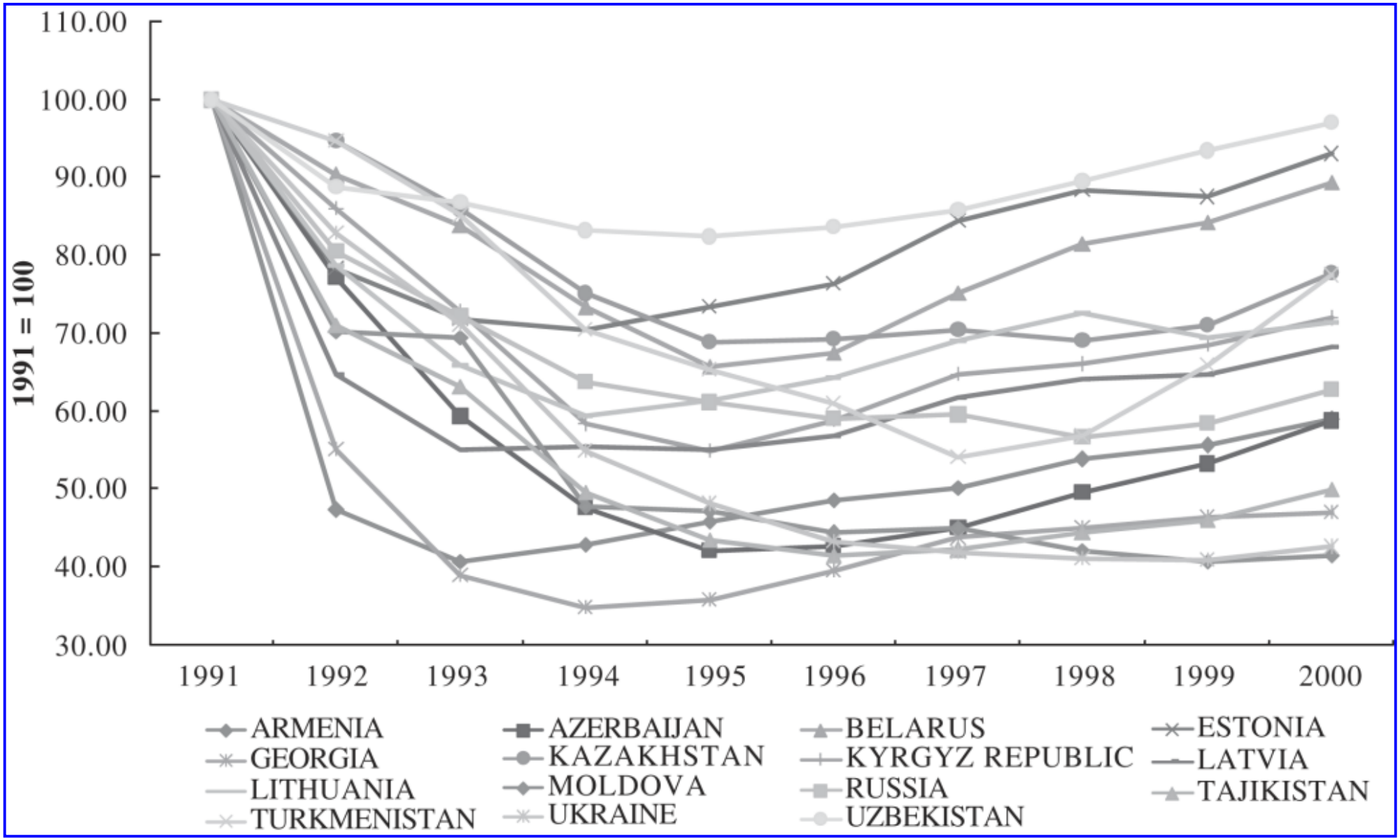

Measured output fell in all transition economies. Generally, the declines are larger in the former Soviet Union (FSU) than in central and eastern European economies (CEEs). For example, using 1989 as the starting point, the falls in Poland (15 per cent), Hungary (18 per cent), and the Czech Republic (21 per cent) were relatively moderate compared with Russia, where from 1991 GDP fell by 40 per cent. Later reformers appear to have larger falls: Ukraine has had a very significant fall in output. (There is, however, a puzzle concerning the output path of Uzbekistan. The output fall was smaller there than in any former Soviet republic, yet it reformed the least. For an analysis, see Zettelmeyer, 1998.)

If we look at industrial output, rather than GDP, the observed declines would be even larger: about 40–50 per cent in central Europe and 50–60 per cent in the FSU. The reason, of course, is that most of the negative value added under planning was in industry, so we would expect a larger contraction there.

The decline in investment, especially in inventories and housing, was even greater than the decrease in GDP. This is especially true for defence. Hence, consumption has fallen less than GDP. In a sense, this is not a surprise as investment is more volatile than output in market economies. Yet transition as an economic process involves restructuring, and this does require investment. The fact that investment absorbed so much of the shock means that the resumption of growth was delayed even further. But it also means that living standards have not fallen as much as GDP. This is important for considering the welfare effects of the output decline.

Measurement issues

Perhaps the output fall is overestimated. (Aslund, 2002, p. 121, considers the output fall to be a myth.) There are many problems with interpreting official data in the context of transition, especially with respect to living standards: too many, indeed, to discuss here. Tracing output dynamics in the transition is complicated by the measurement issues that arise as the economic environment changes from central planning to market forces. Hence, an important issue in understanding the output fall is to gauge the extent to which it is a statistical rather than a real phenomenon.

Some observers (Aslund, 2002; Campos and Coricelli, 2002) argue that the size of the output fall is overstated because of the growth in the size of the shadow economy in early transition. It is argued that the hidden economy grew substantially during the transition period. Hence, actual production fell by less than measured output. The factual basis of this claim is controversial, however. It is also suspect theoretically. The biggest incentive to growth in the second economy is price controls. Hence, price liberalization should result in an immediate drop in the size of the shadow economy. The countervailing pressure could come from tax incentives, but it is hard to believe that this force is stronger than the impact of price controls.

The typical evidence cited in support of the proposition that the hidden economy grew in transition is that measured output fell by more than electricity production. Estimates based on comparing electricity consumption and GDP assume that the elasticity is close to unity. But this elasticity is well below unity in market economies *during recessions*, so employing the unit elasticity assumption amounts to assuming away the phenomenon to be measured. In Finland, for example, real GDP fell by about 11 per cent from 1990 to 1993 while electricity consumption rose by 5.5 per cent (Statistics Finland). By the logic of the advocates of the power consumption thesis we are led to conclude that the hidden economy exploded in size over these three years. For example, if the hidden economy initially was five per cent of total output, then for electricity consumption to rise with no change in intensity of use the hidden economy would have had to grow by 319 per cent! This seems hard to believe. A more likely explanation is the decline in capacity utilization that occurs in recessions causes kilowatt hours of electricity per unit of GDP to increase.

Moreover, as shown by Alexeev and Pyle (2003) the frequently cited estimates of Johnson, Kaufmann and Shleifer (1997) assumed no growth in the size of the shadow economy of the Soviet Union from the late 1970s to the collapse of the system. (The same error is made by Aslund, 2002, p. 122.) This assumption is rejected by all observers of the Soviet economy. Hence, these empirical estimates of the growth in the shadow economy are based on too small an estimate of its initial size.

A second measurement problem in assessing the output fall arises because of the inadequacy of the inherited statistical system to cope with a market economy. Command economies, by their nature, focused on population statistics with regard to output. This is natural in a planned economy where the output produced was the result of a central plan. Indeed, the very nature of command required the planners to coordinate output, hence the statistical system needed to record what each enterprise produced. (Of course, in practice, this was difficult, as discussed in command economy.) The demise of the planning system weakened the authority of central statistical systems. More importantly, new entry became increasingly important in market economies, and the inherited statistical systems are not organized effectively to capture this.

It is also argued that under command systems enterprises had an incentive to overstate output in order to achieve bonuses, while firms in market economies want to hide output in order to avoid taxes (for example, Shleifer and Treisman, 2004). It is thus argued that much output is simply missed by the change in the incentive to report. While it is certainly the case that firms have an incentive to hide output – especially when the financial system is undeveloped so they cannot seek external finance – the incentive to over-report under planning is less clear. Enterprises in planned economies were subject to the notorious ratchet effect. Higher production today meant higher output targets in the future – essentially a highly progressive dynamic tax system. The typical response to the ratchet effect was to produce only as much as needed to satisfy the plan. Hence, it is not at all clear that enterprises over-reported output in the command system.

A more important reason to question the magnitude of the output fall is the contraction in value-destroying activities. Because prices were distorted in planned economies, a portion of economic activity actually destroyed value at market prices. The contraction in these activities represents an increase in welfare, and correctly measured represents an increase in national income as well. The problem is that at the prices that prevailed in command economies this output appeared to be valuable; hence the contraction is measured as a fall in output.

There are two aspects to this decline. First, the separation of domestic from world prices means that activities that produce value added at domestic prices could destroy value at world prices. Given the underpricing of raw materials and overpricing of industrial goods characteristic of planned economies, this was more than a theoretical possibility. External liberalization then leads to a contraction of these activities (McKinnon, 1991). The second aspect is that domestic prices were similarly distorted so that domestic price liberalization has a similar effect. This is discussed below.

To the extent that a reduction of value-destroying activity occurs at the same time as output falls, it is clear that movements in measured output are not consistent with movements in welfare. Indeed, if a greater measured output fall is associated with a faster removal of value-destroying activities, then it is likely that welfare is enhanced by the output fall. In this case the output fall is associated with more reform and quicker removal of welfare destroying activities. (This also means that output recovery could mean a resurgence of value-destroying activities, in which case the upward-sloping part of the U shape is welfare decreasing. Unlikely, but it might be relevant for Belarus under President Lukashenko). Of course, for this to be the case there must a serious distortion in national income measurements. To the extent that output measurements use base-weighted prices this is possible.

It is difficult to measure the extent to which the output fall is overstated by the contraction of value destroying activities. For example Aslund (2002, p. 126) estimates that about 20 per cent of GDP was value destroying in the last years of Communism. He uses, as an indicator, the decline in the share of industry in GDP. Soviet-type economies were over-industrialized, and liberalization led to sectoral shifts as services, which were previously undersupplied, expanded. Moreover, shifts in relative prices, discussed below, also lead to a reduction in the share of value added produced by industry. Thus, one cannot infer value destruction from the change in industrial shares. The general problem is that output may be falling for various reasons so one cannot consider all of the contraction to be previously value destroying. One valuable indicator of the importance of value destruction is given by the comparison of the contraction in industrial output with the rise in consumption that occurred in transition economies. In Russia, for example, industrial output contracted by roughly 35 per cent from January 1992 to January 1994. Real disposable income, on the other hand, increased by almost 70 per cent in the same period (albeit from depressed levels). The fact that real disposable income was growing at the same time as industrial output was contracting suggests that the cessation of value-destroying activity was an important process, and that some of the output fall may be overstated.

A related problem is the shift in preferences. Gaddy and Ickes (2003) argue that a specific index number problem leads to an overstatement of the output fall – the *camellia effect*. The argument is easily understood in terms of an analogy. Consider a flower shop that specializes in the sale of extremely rare camellias. Cultivating these plants is inordinately expensive, but this activity is profitable because the shop has a customer willing to pay very high prices for camellias. Now suppose this customer passes away. The shop can no longer sell rare camellias at a price that covers the cost of production. So camellia cultivation ceases. Resources that were previously devoted to camellia production will now be used for something else, say, roses. Profits at the flower shop fall because camellias were very profitable as long as their special customer lived. But given that there is no longer a market for rare camellias (while there is a market for roses, everyone is better off with rose cultivation than if they continued to cultivate camellias as if nothing had changed.) In the Soviet regime defence output was demanded despite the enormous cost. It had value as long as the Communist Party had command over resources. The special customer of Soviet times made it ‘valuable’ to produce defence output. When the Soviet system collapsed, so did the special customer. Output thus fell – valued at Soviet prices – because at those prices defence output was valued far above cost. After the fall this output is not valued sufficiently and production declines. This is an output fall, but welfare is certainly higher with lower defence production given that the Communist Party is no longer the measure of value.

To see this, suppose that we have two final goods, (x_1, x_2), and that the pre-transition production bundle is (x_1^A, x_2^A), where good 2 is defence output, and *A* represents planners’ preferences. The post-transition allocation is (x_1^B, x_2^B), and reflects social preferences. We might consider, for example, that at point A there is large military production and little civilian production, reflecting planners’ preferences (U_P). The new production bundle is at point B, based on society's preferences. Note that using pre-transition prices to value output, GDP is $Y^A = \Sigma_i p_i^A x_i^A$.

Now suppose that liberalization causes the production bundle to move to point F in Figure 3. This is the most pessimistic outcome – demand for x_2 declines with almost no increase in x_1. Measured in real terms, at the old prices, output falls approximately by the distance AF in units of x_2, or $\Sigma_i p_i^A x_i^F - \Sigma_i p_i^A x_i^A$. But this greatly overestimates the welfare change, because it places a high value on the output that has fallen in valuation.

Figure 3

The camellia effect

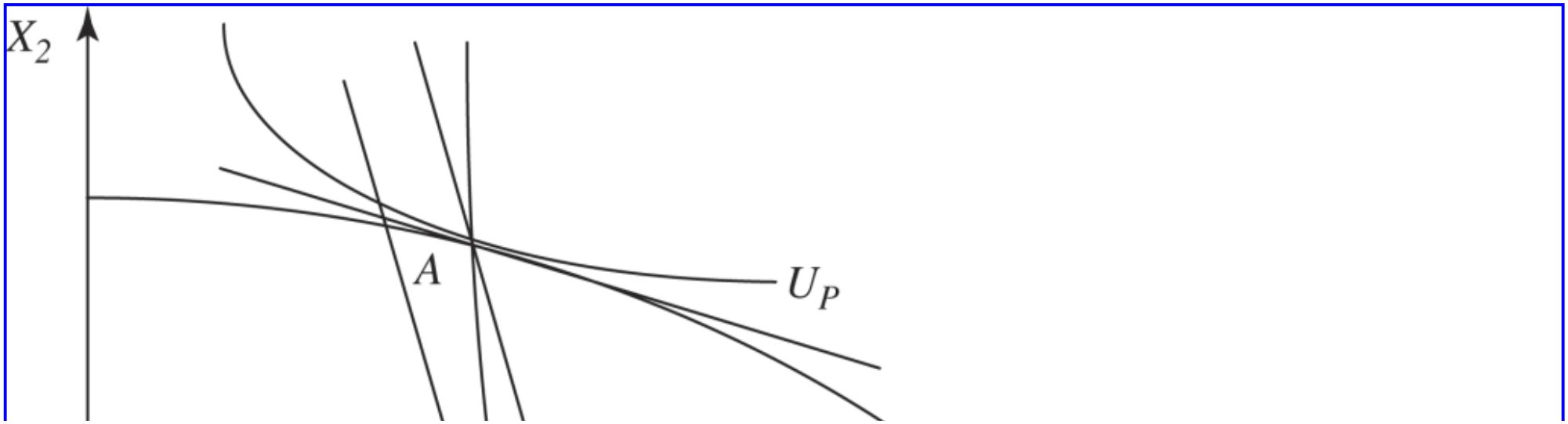

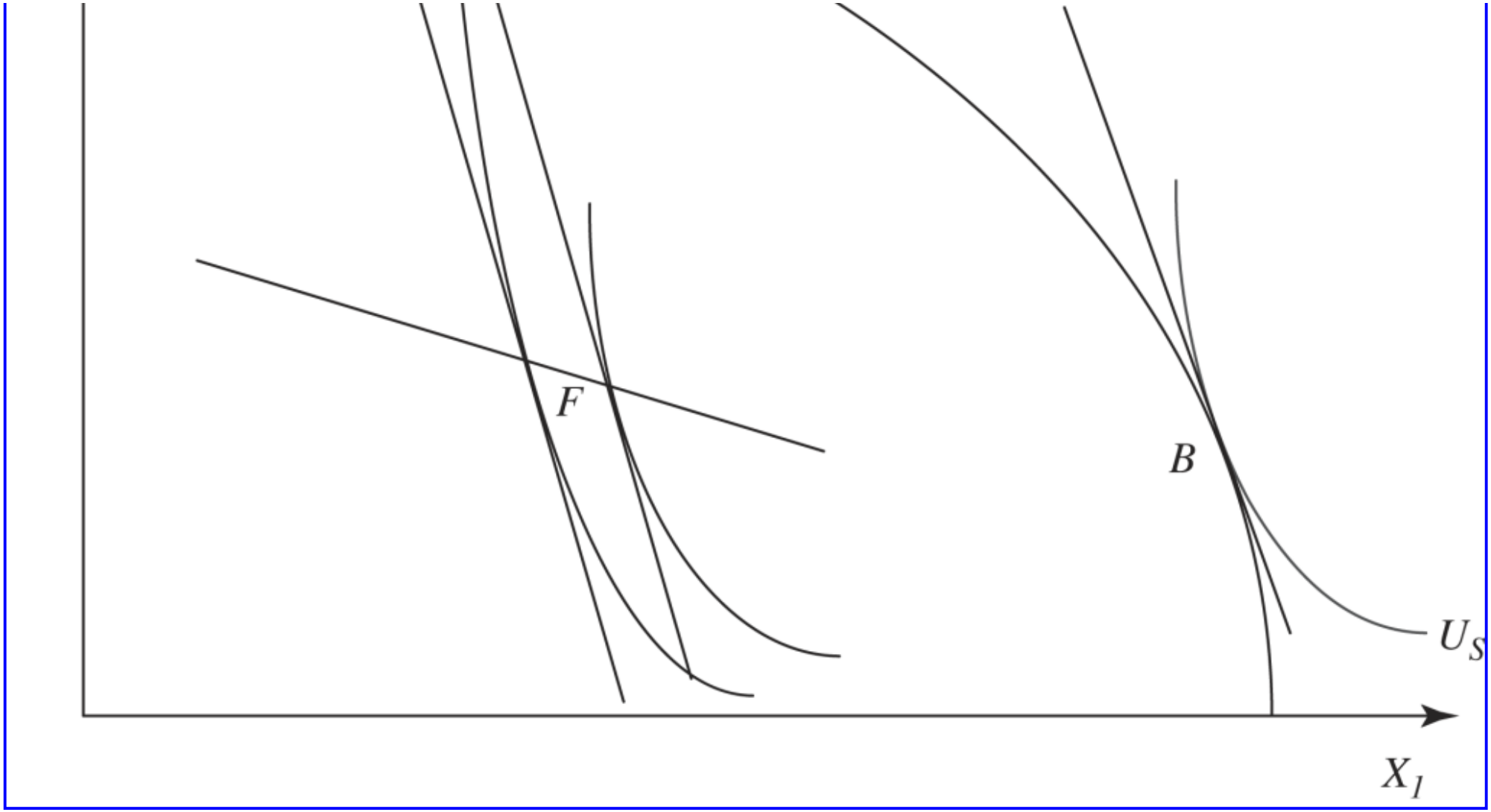

Although output has fallen precipitously at planners' prices, measured at the new prices welfare has clearly increased. The minimum expenditure to achieve the old welfare level $e(p^B, U_A^P)$ is less than the cost of purchasing bundle *F* at the new prices. It is evident that welfare is higher at point *F* than at point *A*. Output has risen at the new prices but has fallen at the old prices.

From Figure 3 we can also distinguish the fall in output due to coordination-type failure and that due to measurement. If resources are fully utilized we would be at point *B*. Hence $\Sigma_i p_i^B x_i^B - \Sigma_i p_i^B x_i^F \equiv \Omega$ measures the fall in output due to coordination-type failure. The measured fall in output, could be larger or smaller than this. The key point, however, is that the measured fall does not measure Ω at all.

Notice that, if the resources devoted to defence production are highly specialized, then there may be great inertia in response to the demand shift. It may be very hard to find alternative uses for these inputs. Output may remain depressed for quite a while. There may also be interesting behavioural issues to think about. A Russian defence enterprise director may expect that the government will soon restore orders and that cuts were temporary. This would lead to inertia in shifting to new activities. Both of these inertial forces could prolong the decline in output.

The importance of the camellia effect for thinking about the output decline is especially important in comparative terms. The camellia effect explains why transitional recessions are observed. But the size of this drop will be proportional to the share of 'camellias' in GDP, and this clearly differs across the post-Communist world. (Even for the former Soviet

Union the differences are dramatic, as Russia had a much larger than average share of Soviet defence industry; see Gaddy, 1996.)

In a country like Russia the size of the defence sector was especially large. This exacerbates the size of the output drop that is due to transitional factors. To measure the pure transition effect we should compare what would have been produced under central planning had planners' preferences not determined production decisions with what happened during transition. Ignoring the camellia effect mixes the two sources of output fall.

Theories of the output fall

Theories of the output fall in transition generally fall into one of two classes. The first class of theories treats this phenomenon as a sign of inefficiency. The output fall is thus welfare decreasing. The second class treats the output fall as a natural feature of liberalization but does not consider the fall to be welfare reducing. (One could also consider the specific negative shocks that have caused output disruptions. For central Europe there is the breakup of Council for Mutual Economic Assistance (CMEA) trade plus the end of subsidized energy from the Soviet Union. For the former Soviet Union there is the disruption in trade caused by the breakup of a common economic space into 15 independent countries. For Russia, there is the decline in oil prices. The importance of movements in the oil price for Soviet and Russian output has been emphasized by Gaddy and Ickes, 2005. The power of this explanation has been fortified by the close timing of the recovery of Russian output with the increase in oil prices starting in the later 1990s.)

A basic framework for thinking about the output fall is the reallocation problem. Consider an economy with two sectors, state (S) and private (P). Initially all labour is employed in the state sector. It is assumed that labour productivity in the private sector (β) exceeds that in the state sector (α), $\alpha < \beta$. The reallocation process occurs as labour moves from the state to the private sector. Per-capita output, y_t, is thus given by

$$y_t = \alpha \frac{L_t^S}{L_t} + \beta \frac{L_t - L_t^S}{L_t};$$

it is immediately apparent that rather than decline, output will increase monotonically in the transition. Hence, to obtain an output fall some unemployment of resources is necessary. If the private sector cannot absorb all the labour released from the state sector then labour will be unemployed, L^U. In that case per-capita output is given by

$$y_t = \alpha \frac{L_t^S}{L_t} + \beta \frac{L_t - L_t^S - L_t^U}{L_t}.$$

This simple framework suggests that to produce an output fall some rigidity or friction is required that prevents smooth reallocation of the labour released from the state sector. The essence of transition suggests that this will be likely. In addition to the normal culprits such as wage rigidity, institutional features play a critical role. For example, prior to the privatization of state sector assets, capital is immobile between sectors. This naturally limits the absorption rate of the private sector. Hence, the exit rate from unemployment will depend on the rate of growth of the private sector. What is important to understand are the determinants of the exit rates from these states. Notice that the growth of the private sector may depend on what is happening in the other sectors. This dependence can occur for several reasons. First, following Aghion and Blanchard (1994), unemployment can cause fiscal deficits which must be financed at the expense of the private sector, limiting its growth. Second, the growth of the private sector may depend on the rate at which complementary resources are released from the state sector. This is especially true for the most basic of resources for production, space. Until privatization of fixed capital takes place it is difficult for new private enterprises to obtain space for production, let alone to lease equipment.

At the most basic level, unemployment can be due to rigidity in real wages. But it is hard to understand how this can explain the output falls that were actually observed, as real wages fell in most transition economies once prices were liberalized. Hence the need for more fully developed theories.

Double marginalization

Li (1999) develops a theory of the output fall in transition based on double marginalization. The basic idea is that the dismantling of central planning or centralized organization of production permits monopolistic and vertically interdependent enterprises to pursue their own monopoly profits by restricting output and intermediate trade to the detriment of the economy as a whole. The basic idea is that the collapse of planning institutions removes constraints on intermediate producers' activities. Intermediate producers now have monopoly power, so they raise prices. This happens all along the supply chain, and results in an increase in the cost of producing final output. So there is less final output available and government output falls. The essential reason is that the enterprises do not consider the consequences of their price increases for the profits of the other enterprises. Since there is less left over for consumers, it is equivalent to a decrease in real wages, and hence labour supply falls.

The essential idea of the double marginalization theory is that output falls because liberalization precedes the development of competition. Entry is a process that takes time. Hence, the theory would predict that output falls would be greater in economies that are less able to 'import' competition through opening the economy. This roughly fits the picture of larger output falls in the FSU than in the CEEs. But the theory also predicts that the output fall should be largest when liberalization first takes place, since that is when market power is most potent. The effect of double marginalization should wane over time. This is harder to reconcile with the paths of output in Figures 1 and 2.

The double marginalization model also predicts that each enterprise will face a contraction in demand and an increase in input prices relative to wage rate. The contraction in demand is attributable to the following factors in this model: the decline in real wage rate, the decline in the government's real income and the decline in input demand. The increase in input prices relative to wage rate is attributable in this model to monopoly pricing by a 'web of monopolies'. The more complex is the web of inter-industry production, the greater the propagation of the price shock. Hence, complexity magnifies any intermediate price markup throughout the economy, resulting in higher input prices relative to wage rate. The sharp increase in input costs is indicative of a sharp supply contraction. This prediction is also consistent with empirical observations.

Disorganization

Blanchard and Kremer (1997) (see also Blanchard, 1997) have developed a model of disorganization that has had great impact. Their argument is that the output fall is a result of the chaos that surrounds the elimination of central planning. They focus on three mechanisms (hold-up problems, coordination, and uncertainty problems) that are greatly magnified as the result of missing institutions likely to be important at the start of transition. The basic idea is that the collapse of planning causes performance to decline during the period when alternative market mechanisms have not yet developed.

The basic idea can be understood in terms of a simple example presented by Blanchard and Kremer. Consider a vertical chain of production. Assume that each step is carried out by a different enterprise. A unit of a primary good is needed at the first step. At the end of the *n* steps one unit of the final good results, and we normalize the price of this good to unity. The value of the intermediate output, at each step, is zero. The supplier of the primary input has an alternative use, which is *c*. This could be much lower than one. It is a private opportunity that could be exporting the good, or selling it for a less fabricated use. Under planning the relations in the chain were directed from above. With liberalization alternative activities may be considered.

The end of planning thus leads to *n* bargaining problems. Each unit must bargain with a supplier and a customer. They assume that there is Nash bargaining at each step, so that the surplus is split given the symmetry of the situation. To see what happens start with the last step. The value of the surplus in the last stage (bargaining between the final producer and the last intermediate producer) is 1. This follows because the value of the good at stage *n* is still zero. So the last intermediate producer gets one half of the surplus. Similar bargaining takes place at all the upstream stages. At the *n*–1 stage there is one half to split … Continue in this fashion and it follows that the first intermediate producer gets $\left(\frac{1}{2}\right)^n$. The surplus available to split at the first stage is $\left(\frac{1}{2}\right)^n - c$, since the first producer must purchase the primary input to produce. It is thus clear that unless $c < \left(\frac{1}{2}\right)^n$ the raw material will be diverted and production will cease. Moreover, *c* does not need to be all that large to trigger defection that results in a fall in output that could be as large as $1 - \left(\frac{1}{2}\right)^n$. Thus rather meagre private opportunities can cause a rather large fall in output.

Blanchard and Kremer interpret *n* as the level of complexity of production. As *n* increases, the likelihood of defection increases exponentially. This is a hold-up problem. Each producer in the chain must produce before bargaining with the next in line. This suggests that the problem would go away if each of the producers could sign an enforceable contract before production takes place. As long as $c < 1$, defection could be avoided and production could take place, if the intermediate producers could sign a contract to split the 1–*c* before production. The problem is thus one of asset specificity and incomplete contracts. Eliminating the ministry before institutions that support contracts are developed is the source of the problem. Vertical integration could help, but this requires ownership to be specified, another problem early in transition. The notion that producers in transition could suffer from this problem is not far-fetched. (It is interesting to compare this outcome with the *double marginalization* case. Notice that in that case the raw materials producer has market power and thus a higher share of the surplus than is the case in the bargaining problem. This makes production in the state sector more likely. Of course, what is not explained is why the producer is able to extract monopoly rents in a situation of bilateral monopoly.)

Blanchard and Kremer consider other examples based on incomplete information. A state-owned enterprise must negotiate with many suppliers that may have outside options. Each of the suppliers produces a key input without which production is impossible. With uncertainty over the magnitude of outside options a state-owned enterprise must guess how much

to pay for the inputs. When outside opportunities are low the possibility that the state-owned enterprise offers too low a price is negligible. But as these outside opportunities rise this probability increases. Even if it is still efficient to sell to the state-owned enterprise because of uncertainty over the size of these options, the price offered may be too low and production falls. The interesting feature of this model is that it produces a U-shaped output path. The key assumptions are technological complementarities and inefficient bargaining.

A coordination example can also be constructed. Suppose that the firm needs *n* workers (it could be supplying firms, but this is easier), and the technology is Leontief. If all workers stay, the firm produces one unit of output per worker. If a worker leaves, a replacement is hired with output per worker equal to $\gamma < 1$. Here again *n* measures the degree of complexity, while γ is an inverse measure of the specificity of the production process or job-specific human capital.

Each worker has an alternative opportunity given by *c*, distributed on $[0, \bar{c}]$, where $\bar{c}$ represents the maximum outside opportunity, which is of course a function of the state of the transition. Draws from this distribution are independent across workers. The distribution is known, but the specific realization is private information. This could be thought of as alternative employment, perhaps in a Western multinational. The firm pays a common wage, *w*, to all workers, equal to output per worker. This simplifies the analysis, but is probably not crucial.

The key assumption of the model is that workers must decide whether to take up the alternative before they know the decision of the other workers. This creates the coordination problem. Workers are risk neutral, so that all we need to look at is expected output. There are thus two potential outcomes: (*a*) all workers stay, output per worker and thus the wage are equal to unity, or; (*b*) one or more workers leave, output per worker and the wage are equal to γ.

The decision problem for the agents boils down to determining some threshold level of outside opportunities, c^*, such that if $c<c^*$, workers stay and vice versa. If a worker leaves he receives *c*. If he stays his expected earnings will depend on what the other *n*–1 workers do. Assume symmetry so that the other workers also have the same c^*. Then the probability that they all stay is $(F(c^*))^{n-1}$, where $F(\cdot)$ is the distribution function so that $F(0) = 0$ and $F(\bar{c}) = 1$. Expected output per worker is thus equal to $(F(c^*))^{n-1} + \gamma[1 - (F(c^*))^{n-1}]$.

The key point is that there may be multiple equilibria, depending on the level of outside opportunities. If alternative opportunities are very low, workers always stay in the firm, and output equals 1. As outside opportunities increase there are two equilibria; in one of these output falls close to γ. With very high outside opportunities production in the state sector ceases. Note the problem here is coordination, not uncertainty. If the outside opportunity were common knowledge, with $\gamma < c < \bar{c}$ there would still be two equilibria.

The essential feature of the disorganization model is that central planning is replaced before the infrastructure of markets is created. The lack of central organization leads to disorganization, and the development of outside opportunities makes this problem more severe. Over time, market infrastructure develops and disorganization problems are lessened.

Roland and Verdier (1999) develop a related model of disorganization, focusing on search frictions rather than bargaining problems. In their model liberalization means that enterprises can search for new suppliers and customers. There are good matches and bad matches. If too many bad clients are searching the productivity of potential matches may fall. What is critical in their model is that relationship-specific investments take place only after long-term matches are formed. If search continues this will not happen, investment demand will fall, and output can fall.

Investment specificity is crucial in this model. Without it output would not fall even with bad matches, since the partners could produce this period and keep on searching. It is the asset specificity that introduces the cost of bad matches.

The Roland–Verdier model is interesting from a theoretical point of view, but one may wonder how relevant it really is for explaining the output fall. The problem is that the initial output fall was associated with very little search for new suppliers. The predominant behaviour was a relationship-conservatism. Agents tried to maintain their relationships as much as possible. Networks of suppliers already had relationship-specific investments. The problem is that they had no customers who would purchase the goods at a price that covered their new costs.

Micro-distortions

A more subtle, but equally important explanation of the output fall focuses on the micro distortions due to Soviet pricing rules. Ericson (1999) has analysed this problem. His focus is on structural problems with Soviet pricing – the arbitrariness and non-uniformity of producers' prices across users of the product within standard commodity aggregates. Ericson shows that Soviet pricing rules hid inefficiency and waste, creating an illusion of capacity and output that wasn't there. The advantage of this theory is that it can explain why prices exploded when output fell. His argument is that post-Soviet 'stagflation' is, to some extent, a consequence of the irrational structure of production hidden in apparently consistent (adjusted) input–output (I–O) matrices and economic statistics.

Soviet pricing rules contained three systematic distortions: (*a*) basic factors were seriously undervalued (land was free, and capital-in-place virtually so); (*b*) raw materials and natural resources were undervalued; and (*c*) highly processed goods – in particular investment products and services – were seriously overvalued. These distortions in the principles of economic valuation used in centrally planned economies systematically hide tremendous waste, exaggerating both net outputs and net income (economic value) produced, while understating the productivity of that most seriously mismeasured factor of production, capital. This implies that the size of the apparent initial collapse in industrial production is evidently exaggerated, even if one ignores new economic activity generated in the wake of the reforms. However, the wasteful production structure can also spur a continuing and deepening collapse, as it is not economically viable in a market environment.

Ericson shows that embedding these distortions in the input–output tables that are used to create national income statistics results in lower prices for inputs than for final uses, and

generates an understatement of the share of gross output used in the production process. Thus, it leads to an overstatement of the share of net output. Furthermore, these distortions cannot be revealed by any consistent input–output framework derived from the 'value' of transactions between sectors; the methodology itself imposes a consistency that hides those distortions. This means that the true nature of the system cannot be revealed until price liberalization takes place. Until then, intersectoral relationships are hidden. This is what creates the 'circus mirror' effect discussed by Gaddy and Ickes (2002). (A circus mirror distorts size and shape. Soviet pricing rules had the same effect, making value added look larger and intermediate input use look smaller). Just as an individual may look taller and thinner in a circus mirror, the Soviet-type economy appeared more productive under Soviet pricing rules. Liberalization revealed the true nature of the economy.

Ericson shows that for the case of Russia the 1991 input–output coefficients were substantially understated, hiding significant materials input use and waste, and hence obscuring much of the inherited inefficiency in the industrial structure. This inefficiency became of consequence for producers when liberalization released them from ministerial tutelage and constraints, and made them primarily responsible for covering their own costs. Because enterprises are initially constrained by existing technological structures, the first impact of liberalization is typically seen in the move to raise prices to cover their full material costs and to compensate for any increases. This led to increases in industrial prices that far exceeded the general rate of inflation, raising the real price of industrial output and consequently real materials costs. As in the double marginalization theory, price increases in the intermediate sector propagate through the economy and result in less final output. But the impulse is different. Ericson's theory does not require any market power on the part of intermediate producers. Price increases are solely due to price liberalization itself in the context of Soviet pricing. Of course, at those increased real prices, demand for many products, now not supported by plan requirements, falls dramatically; producers find they are unable to sell at higher prices and hence unable to recover the full costs of production. Yet they continued to operate and ship output to traditional users of their product.

Ericson's theory is thus consistent with several important aspects of the output fall that are hard to explain in other models. First, his theory explains why the output fall is associated with a rise in the price level. Second, it is consistent with higher wholesale price inflation than consumer price inflation. Third, it is consistent with the explosion of inter-enterprise arrears. Supply and disorganization type theories make no prediction with regard to overall inflation and they are inconsistent with the latter two observations.

Empirical analysis

Most empirical analyses of the output fall has been focused on assessing the role of policies (primarily, stabilization and liberalization) and initial conditions in determining the size of the fall in output. This literature is too large to summarize here (a good summary is Campos and Coricelli, 2002), but a few points can be made. First, results are very dependent on how policies, especially the speed and extent of liberalization, are measured, and how initial conditions are proxied. Measures of liberalization that rely on expert evaluation are subject to performance bias: that is, the liberalization score that is assessed is often inferred from economic performance. The set of initial conditions that are important include the degree of over-industrialization, repressed inflation, dependence on CMEA trade, distance from Frankfurt, years spent under Communism, initial income, and the rate of urbanization. Depending on the set used results can differ dramatically.

One of the most comprehensive studies of the impact of policies versus initial conditions is by Berg et al. (1999). They use a sample of 26 transition economies and use a general to specific modelling approach that allows for differential effects of policies and initial conditions and for time-dependent effects of initial conditions. They find that structural reforms are more important than either policies or initial conditions in explaining the cross-country variation in performance. Initial conditions play the predominant role in explaining the output fall, while structural reforms explain the recovery. The most important initial conditions appear to be the degree of over-industrialization and trade dependency.

Conclusion

Although the size of the output fall indicated by official measures is clearly overstated, the fact that output and incomes did fall in the aftermath of liberalization is not disputed. Moreover, the fact that output followed a U-shaped pattern has had important consequences for transition. Not least of these is the negative effect it had on the political support for many economic reformers. The output decline made it politically difficult to stick with reforms. Hence, the output declines may have altered the course of policy reform in transition. Ironically, it seems that reform reversals were often associated with longer output declines.

See Also

- command economy
- institutional trap
- second economy (unofficial economy)
- soft budget constraint

- virtual economy

Bibliography

Aghion, P. and Blanchard, O. 1994. On the speed of transition in central Europe. In *NBER Macroeconomics Annual*. Cambridge, MA: MIT Press.

Alexeev, M. and Pyle, W. 2003. A note on measuring the unofficial economy in the former Soviet republics. *Economics of Transition* 11, 153–75.

Aslund, A. 2002. *Building Capitalism: The Transformation of the Former Soviet Bloc*. Cambridge: Cambridge University Press.

Berg, A., Borenzstein, E., Sahay, R. and Zettelmeyer, J. 1999. The evolution of output in transition economies: explaining the differences. Working Paper No. 99/73, International Monetary Fund.

Blanchard, O. 1997. *The Economics of Transition in Eastern Europe*. Oxford: Oxford University Press.

Blanchard, O. and Kremer, M. 1997. Disorganization. *Quarterly Journal of Economics* 112, 1091–126.

Campos, N. and Coricelli, F. 2002. Growth in transition: what we know, what we don't know, and what we should. *Journal of Economic Literature* 40, 793–836.

Ericson, R.E. 1999. The structural barrier to transition hidden in input–output tables of centrally planned economies. *Economic Systems* 23(3), 199–224.

Fischer, S. and Sahay, R. 2000. The transition economies after ten years. Working Paper No. 7664. Cambridge, MA: NBER.

Gaddy, C.G. 1996. *The Price of the Past*. Washington: Brookings Institution Press.

Gaddy, C.G. and Ickes, B.W. 2002. *Russia's Virtual Economy*. Washington: Brookings Institution Press.

Gaddy, C.G. and Ickes, B.W. 2003. How to think about the post-Soviet output fall. Working paper, Brookings Institution and Pennsylvania State University. Online. Available at http://econ.la.psu.edu/~bickes/OutputFall.pdf, accessed 20 May 2007.

Gaddy, C.G. and Ickes, B.W. 2005. Resource rents and the Russian economy. *Eurasian Geography and Economics* 46, 559–83.

Johnson, S., Kaufmann, D. and Shleifer, A. 1997. The unofficial economy in transition. *Brookings Papers on Economic Activity* 1997(2), 159–239.

Li, W. 1999. A tale of two reforms. *RAND Journal of Economics* 30, 120–36.

McKinnon, R. 1991. *The Order of Economic Liberalization: Financial Control in the Transition to a Market Economy*. Baltimore: Johns Hopkins Press.

Mundell, R.A. 1997. The great contractions in transitions economies. In *Macroeconomic Stabilization in Transition Economies*, ed. M.I. Blejer and M. Skreb. New York: Cambridge University Press.

Popov, V. 2005. Shock therapy versus gradualism reconsidered: lessons from transition economies after 15 years of reforms. TIGER Working Paper No. 82.

Roland, G. and Verdier, T. 1999. Transition and the output fall. *Economics of Transition* 7, 1–28.

Shleifer, A. and Treisman, D. 2004. A normal country. *Foreign Affairs* 83(2), 20–38.

Statistics Finland. Online. Available at http://www.tilastokeskus.fi/index_en.html, accessed 28 June 2007.

Svejnar, J. 2002. Transition economies: performance and challenges. *Journal of Economic Perspectives* 16(1), 3–28.

Zettelmeyer, J. 1998. The Uzbek growth puzzle. Working Paper No. 98/133, International Monetary Fund.

How to cite this article

Ickes, Barry W. "output fall – transformational recession." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000106> doi:10.1057/9780230226203.1235

The New Palgrave Dictionary of Economics Online

overlapping generations model of general equilibrium

John Geanakoplos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The OLG model of Allais and Samuelson retains the methodological assumptions of agent optimization and market clearing from the Arrow–Debreu model, yet its equilibrium set has different properties: Pareto inefficiency, multiplicity, positive valuation of money, and a golden rule equilibrium in which the rate of interest is equal to population growth (independent of impatience). These properties are shown to derive not from market incompleteness, but from lack of market clearing ‘at infinity’: they can be eliminated with land or uniform impatience. The OLG model is used to analyse bubbles, social security, demographic effects on stock returns, the foundations of monetary theory, Keynesian vs. real business cycle macromodels, and classical vs. neoclassical disputes.

Keywords

agent optimization; Allais, M.; animal spirits; Arrow–Debreu model of general equilibrium; backward induction; bubbles; Cobb–Douglas functions; comparative statics; consumption loan model; continuum of equilibria; cores; demography; double coincidence of wants; equilibrium; existence of equilibrium; expectations sensitivity hypothesis; impatience; incomplete markets; indeterminacy of equilibrium; infinite horizons; involuntary unemployment; Keynesianism; marginal utility of money; market clearing; money; multiple equilibria; new classical macroeconomics; numeraire; overlapping generations models of general equilibrium; Pareto efficiency; Pareto inefficiency; perfect foresight; price normalization; Samuelson, P. A.; sequential equilibrium; social security; Sraffa, P.; sunspots; uncertainty; uniform impatience; uniqueness of equilibrium

Article

The consumption loan model that Paul Samuelson introduced in 1958 to analyse the rate of interest, with or without the social contrivance of money, has developed into what is without doubt the most important and influential paradigm in neoclassical general equilibrium theory outside of the Arrow–Debreu economy. Earlier Maurice Allais (1947) had presented similar ideas which unfortunately did not then receive the attention they deserved. A vast literature in public finance and macroeconomics is based on the model, including studies of the national debt, social security, the incidence of taxation and bequests on the accumulation of capital, the Phillips curve, the business cycle, and the foundations of monetary theory. In this article I give a hint of these myriad applications only in so far as they illuminate the general theory. My main concern is with the relationship between the Samuelson model and the Arrow–Debreu model.

Allais’s and Samuelson's innovation was in postulating a demographic structure in which generations overlap, indefinitely into the future; up until then it had been customary to regard all agents as contemporaneous. In the simplest possible example, in which each generation lives for two periods, endowed with a perishable commodity when young and nothing when old, Samuelson noticed a great surprise. Although each agent could be made better off if he gave half his youthful birthright to his predecessor, receiving in turn half from his successor, in the marketplace there would be no trade at all. A father can benefit from his son's resources, but has nothing to offer in return.

This failure of the market stirred a long and confused controversy. Samuelson himself attributed the suboptimality to a lack of double coincidence of wants. He suggested the social contrivance of money as a solution. Abba Lerner suggested changing the definition of optimality. Others, following Samuelson's hints about the financial intermediation role of money, sought to explain the

consumption loan model by the incompleteness of markets. It has only gradually become clear that the ‘Samuelson suboptimality paradox’ has nothing to do with the absence of markets or financial intermediation. Exactly the same equilibrium allocation would be reached if all the agents, dead and unborn, met (in spirit) before the beginning of time and traded all consumption goods, dated from all time periods, simultaneously under the usual conditions of perfect intermediation. Indeed, in the early 20th century Irving Fisher (1907; 1930) implicitly argued that any sequential economy without uncertainty, but with a functioning loan market, could be equivalently described as if all markets met once with trade conducted at present value prices.

Over the years Samuelson's consumption loan example, infused with Arrow–Debreu methods, has been developed into a full-blown general equilibrium model with many agents, and multiple kinds of commodities and production. It is equally faithful to the neoclassical methodological assumptions of agent optimization, market clearing, price taking, and rational expectations as the Arrow–Debreu model. This more comprehensive version of Samuelson's original idea is known as the overlapping generations (OLG) model of general equilibrium.

Despite the methodological similarities between the OLG model and the Arrow–Debreu model, there is a profound difference in their equilibria. The OLG equilibria may be Pareto suboptimal. Money may have positive value. There are robust OLG economies with a continuum of equilibria. Indeed, the more commodities per period, the higher the dimension of multiplicity may be. Finally, the core of an OLG economy may be empty. None of this could happen in any Arrow–Debreu economy.

The puzzle is: why? One looks in vain for an externality, or one of the other conventional pathologies of an Arrow–Debreu economy. It is evident that the simple fact that generations overlap cannot be an explanation, since by judicious choice of utility functions one can build that into the Arrow–Debreu model. It cannot be simply that the time horizon is infinite, as we shall see, since there are classes of infinite horizon economies whose equilibria behave very much like Arrow–Debreu equilibria. It is the combination, that generations overlap indefinitely, which is somehow crucial. In Section 4 I explain how.

Note that in the Arrow–Debreu economy the number of commodities, and hence of time periods, is finite. One is tempted to think that, if the end of the world is put far enough off into the future, it could hardly matter to behaviour today. But recalling the extreme rationality hypotheses of the Arrow–Debreu model, it should not be surprising that such a cataclysmic event, no matter how long delayed, could exercise a strong influence on behaviour. Indeed, the OLG model proves that it does. One can think of other examples. Social security, based on the pay-as-you-go principle in the United States in which the young make payments directly to the old, depends crucially on people thinking that there might always be a future generation. Otherwise the last generation of young will not contribute; foreseeing that, neither will the second-to-last generation of young contribute, nor, working backward, will any generation contribute. Another similar example comes from game theory, in which cooperation depends on an infinite horizon. On the whole, it seems at least as realistic to suppose that everyone believes the world is immortal as to suppose that everyone believes in a definite date by which it will end. (In fact, it is enough that people believe, for every T, that there is positive probability the world lasts past T.)

In Section 1, I analyse a simple one-commodity OLG model from the present value general equilibrium perspective. This illustrates the paradoxical nature of OLG equilibria in the most orthodox setting. These paradoxical properties can hold equally for economies with many commodities, as pointed out in Section 4. Section 2 discusses the possibility of equilibrium cycles in a one-commodity, stationary, OLG economy. In Section 3, I describe OLG equilibria from a sequential markets point of view, and show that money can have positive value.

In the simple OLG economy of Section 1 there are two steady-state equilibria, and a continuum of non-stationary equilibria. Out of all of these, only one is Pareto efficient, and it has the property that the real rate of interest is always zero, just equal to the rate of population growth, independent of the impatience of the consumers or the distribution of endowments between youth and old age. This ‘golden rule’ equilibrium seems to violate Fisher's impatience theory of interest.

In Section 5 I add land to the one-commodity model of Section 1. It turns out that now there is a unique steady-state equilibrium that is Pareto efficient and that has a positive rate of interest, greater than the population growth rate, that increases if consumers become more impatient. Land restores Fisher's view of interest. In this setting it is also possible to analyse the effects of social security.

In Section 6 I briefly introduce variations in demography. It is well known that birth rates in the United States oscillated every 20 years over the 20th century. Stock prices have curiously moved in parallel, rising rapidly from 1945 to 1965, falling from 1965 to 1985, and rising ever since. One might therefore expect stock prices to fall as the post-war baby boom generation retires. But some authors have claimed that these parallel fluctuations of stock prices must be coincidental. Otherwise, since demographic changes are known long in advance, rational investors would have anticipated the price fluctuations and changed them. In Section 6 I allow the size of the generations to alternate and confirm that in OLG equilibrium land prices rise and fall with demography, even though the changes are perfectly anticipated.

In Section 7 I show that not just land but also uniform impatience restores the properties of infinite horizon economies to those found in finite Arrow–Debreu economies.

Section 8 takes up the question of comparative statics. If there is a multiplicity of OLG equilibria, what sense can be made of comparative statics? Section 8 summarizes the work showing that, for perfectly anticipated changes, there is only one equilibrium in the multiplicity that is 'near' an original 'regular' equilibrium. For unanticipated changes, there may be a multidimensional multiplicity. But it is parameterizable. Hence, by always fixing the same variables, a unique prediction can be made for changes in the equilibrium in response to perturbations. In Section 9 we see how this could be used to understand some of the New Classical–Keynesian disputes about macroeconomic policy. Different theories hold different variables fixed in making predictions.

Section 10 considers a neoclassical–classical controversy. Recall the classical economists' conception of the economic process as a never-ending cycle of reproduction in which the state of physical commodities is always renewed, and in which the rate of interest is determined outside the system of supply and demand. Samuelson attempted to give a completely neoclassical explanation of the rate of interest in just such a setting. It now appears that the market forces of supply and demand are not sufficient to determine the rate of interest in the standard OLG model. In other infinite-horizon models they do.

Section 11 summarizes some work on sunspots in the OLG model. Uncertainty in dynamic models seems likely to be very important in the future.

An explanation of the puzzles of OLG equilibria without land is given in Section 4: lack of market clearing 'at infinity'. By appealing to non-standard analysis, the mathematics of infinite and infinitesimal numbers, it can be shown that there is a 'finite-like' Arrow–Debreu economy whose 'classical equilibria', those price sequences which need not clear the markets in the last period, are isomorphic to the OLG equilibria. Lack of market clearing is also used to explain the suboptimality and the positive valuation of money.

1 Indeterminacy and suboptimality in a simple OLG model

In this section we analyse the equilibrium set of a one-commodity per period, overlapping generations (OLG) economy, assuming that all agents meet simultaneously in all markets before time begins, just as in the Arrow–Debreu model. Prices are all quoted in present value terms; that is, p_t is the price an agent would pay when the markets meet (at time $-\infty$) in order to receive one unit of the good at time t. Although this definition of equilibrium is firmly in the Walrasian tradition of agent optimization and market clearing, we discover three surprises. There are robust examples of OLG economies that possess an uncountable multiplicity of equilibria, that are not in the core, or even Pareto optimal. This lack of optimality (in a slightly different model, as we shall see) was pointed out by Samuelson in his seminal (1958) paper. The indeterminacy of equilibrium in the one-commodity case is usually associated first with Gale (1973). In later sections we shall show that these puzzles are robust to an extension of the model to multiple commodities and agents per period, and to a non-stationary environment. We shall add still another puzzle in Section 3, the positive valuation of money, which is also due to Samuelson.

A large part of this section is devoted to developing the notation and price normalization that we shall use throughout. In any Walrasian model the problem of price normalization (the 'numeraire problem') arises. Here the most convenient solution in the long run is not at first glance the most transparent.

Consider an overlapping generation (OLG) economy $E = E_{-\infty,\infty}$ in which discrete time periods t extend indefinitely into the past and into the future, $t \in \mathbf{Z}$. Corresponding to each time period there is a single, perishable consumption good x_t. Suppose furthermore that at each date t one agent is 'born' and lives for two periods, with utility

$$u^t(\ldots, x_t, x_{t+1}, \ldots) = a^t \log x_t + (1 - a^t) \log x_{t+1}$$

defined over all vectors

$$x = (\ldots, x_{-1}, x_0, x_1, \ldots) \in L = R_+^{\mathbf{Z}}.$$

Thus we identify the set of agents A with the time periods $\mathbf{Z}$. Let each agent $t \in A$ have endowment

$$e^t = (\ldots, e_t^t, e_{t+1}^t, \ldots) \in L$$

which is positive only during the two periods of his life. Note that

$$\sum_{t \in A} e_s^t = e_s^{s-1} + e_s^s \text{ for all } s \in \mathbf{Z}.$$

An equilibrium is defined as a (present value) price vector

$$p = (\ldots, p_{-1}, p_0, p_1, \ldots) \in L$$

and allocation

$$\bar{x} = [x^t = (\ldots, x_t^t, x_{t+1}^t, \ldots); t \in A]$$

satisfying $\bar{x}$ is feasible, that is,

$$\sum_{t \in A} x_s^t = \sum_{t \in A} e_s^t, \text{ for all } s \in \mathbf{Z} \quad (1)$$

and

$$\sum_{s \in \mathbf{Z}} p_s e_s^t < \infty \text{ for all } t \in A \quad (2)$$

and

$$x^t \in \arg\max_{x \in L} \left\{ u^t(x) \middle| \sum_{s \in \mathbf{Z}} p_s x_s \leq \sum_{s \in \mathbf{Z}} p_s e_s^t \right\}. \quad (3)$$

The above definition of equilibrium is precisely in the Walrasian tradition, except that it allows for both an infinite number of traders and commodities. All prices are finite, and consumers treat them as parametric in calculating their budgets. The fact that the definition leads to robust examples with a continuum of Pareto-suboptimal equilibria calls for an explanation. We shall give two of them, one at the end of this section, and one in Section 4. Note that condition (2) becomes necessary only when we consider models in which agents have positive endowments in an infinite number of time periods.

As usual, the set of (present value) equilibrium price sequences displays a trivial dimension of multiplicity (indeterminacy), since, if p is an equilibrium, so is kp for all scalars $k>0$. We can remove this ambiguity by choosing a price normalization $q_t = p_{t+1} / p_t$, for all $t \in \mathbf{Z}$. The sequence $q = (\ldots, q_{-1}, q_0, \ldots)$ and allocations $(x^t; t \in A)$ form an equilibrium if (1) above holds together with

$$x^t \in \arg\max_{x \in L} \left\{ u^t(x) \mid x_t + q_t x_{t+1} \leq e_t^t + q_t e_{t+1}^t \right\}. \tag{4}$$

Notice that we have taken advantage of the finite lifetimes of the agents to combine (2) and (3) into a single condition (4). We could have normalized prices by choosing a numeraire commodity, and setting its price equal to one, say $p_0=1$. The normalization we have chosen instead has three advantages as compared with this more obvious system. First, the q system is time invariant. It does not single out a special period in which a price must be 1; if we relabelled calendar time, then the corresponding relabelling of the q_t would preserve the equilibrium. In the numeraire normalization, after the calendar shift, prices would have to be renormalized to maintain $p_0=1$. Second, on account of the monotonicity of preferences, we know that, if the preferences and endowments are uniformly bounded

$$0 < \underline{a} \leq a^t \leq \bar{a} < 1, \quad 0 < \underline{e} \leq e_t^t, \quad e_{t+1}^t \leq \bar{e} \leq 1 \text{ for all } t \in A,$$

then we can specify uniform a priori bounds $\underline{k}$ and $\bar{k}$ such that any equilibrium price vector q must satisfy $\underline{k} \leq q_t \leq \bar{k}$ for all $t \in \mathbf{Z}$. Third, it is sometimes convenient to note that each generation's excess demand depends on its own price. We define

$$[Z_t^t(q_t), Z_{t+1}^t(q_t)] = (x_t^t - e_t^t, x_{t+1}^t - e_{t+1}^t)$$

for x^t satisfying (4), as the excess demand of generation t, when young and when old. We can accordingly rewrite equilibrium condition (1) as

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0 \text{ for all } t \in \mathbf{Z}. \tag{5}$$

Let us now investigate the equilibria of the above economy when preferences and endowments are perfectly stationary. To be concrete, let

$$a^t = a \text{ for all } t \in A,$$

and let

$$e_t^t = e, \text{ and } e_{t+1}^t = 1 - e, \text{ for all } t \in A,$$

where $e > a \geq 1/2$. Agents are born with a larger endowment when young than when old, but the aggregate endowment of the economy is constant at 1 in every time period. Furthermore, each agent regards consumption when young as at least as important as consumption when old $(a \geq 1/2)$, but on account of the skewed endowment the marginal utility of consumption at the endowment allocation when young is lower than when old:

$$\frac{a}{e} < \frac{1-a}{1-e}.$$

If we choose

$$q_t = \bar{q} = \frac{(1-a)e}{(1-e)a} > 1$$

for all $t \in \mathbf{Z}$, then we see clearly that at these prices each agent will just consume his endowment; $q = (\ldots, \bar{q}, \bar{q}, \ldots)$ is an equilibrium price vector, with $x^t = e^t$ for all $t \in A$. Note that if we had used the price normalization $p_0 = 1$, the equilibrium prices would be described by

$$(\ldots, p_0, p_1, p_2, \ldots) = (\ldots, 1, \bar{q}, \bar{q}^2, \ldots)$$

where $p_t \to \infty$ as $t \to \infty$. With $a = 1/2$ and $e = 3/4$, we get $\bar{q} = 3$ and $p_t = 3^t$.
But there are other equilibria as well. Take $q = (\ldots, 1, 1, 1, \ldots)$, and

$$(x_t^t, x_{t+1}^t) = (a, 1-a) \text{ for all } t \in A$$

This ‘golden rule’ Pareto equilibrium dominates the autarkic equilibrium previously calculated. With $a = 1/2$ and $e = 3/4$, we see that (1/2,1/2) is much better for everyone than (3/4,1/4). This raises the most important puzzle of overlapping generations economies: why is it that equilibria can fail to be Pareto optimal? We shall discuss this question at length in Section 4.
For now, let us observe one more curious fact. We can define the *core* of our economy in a manner exactly analogous to the finite commodity and consumer case. We say that a feasible allocation $x = (x^t; t \in A)$ is in the core of the economy E if there is no subset of traders $A' \subset A$, and an allocation $y = (y^t; t \in A')$ for A' such that

$$\sum_{t \in A'} y^t = \sum_{t \in A'} e^t,$$

and

$$u^t(y^t) > u^t(x^t) \text{ for all } t \in A'.$$

A simple argument can be given to show that the core of this economy is empty. For example, the golden rule equilibrium allocation is Pareto optimal, but not in the core. Since $a<e$, every agent is consuming less when young than his initial endowment. Thus for any $t_0 \in A$, the coalition $A' = \{t \in A | t \geq t_0\}$ consisting of all agents born at time t_0 or later can block the golden rule allocation.

Let us continue to investigate the set of equilibria of our simple, stationary economy. Gale (1973) showed that for any $\bar{q}_0$, with $1 < \bar{q}_0 < \bar{q}$, there is an equilibrium price sequence

$$q = (\ldots, q_{-1}, q_0, q_1, \ldots)$$

with $q_0 = \bar{q}_0$. In other words, there is a whole continuum of equilibria, containing a nontrivial interval of values. Incidentally, it can also be shown that for all such equilibria q, $q_t \to \bar{q}$ as $t \to \infty$, and $q_t \to 1$ as $t \to -\infty$. Moreover, these equilibria, together with the two steady state equilibria, constitute the entire equilibrium set.

This raises the second great puzzle of overlapping generations economies. There can be a non-degenerate continuum of equilibria, while in finite commodity and finite agent economies there is typically only a finite number. Thus if we considered the finite truncated economy $E_{-T,T}$ consisting of those agents born between $-T$ and T, and no others, then it can easily be seen that there is only a unique equilibrium $(q_{-T}, \ldots, q_T) = (\bar{q}, \ldots, \bar{q})$, no matter how large T is taken. On the other hand, in the overlapping generations economy, there is a continuum of equilibria. Moreover, the differences in these equilibria are not to be seen only at the tails. In the OLG economy, as $\bar{q}_0$ varies from 1 to $\bar{q}$, the consumption of the young agent at time zero varies from a to e, and his utility from $a \log e + (1-a) \log (1-e)$ (which for e near 1 is close to $-\infty$), all the way to $a \log a + (1-a) \log (1-a)$. By pushing the 'end of the world' further into the future, one does not approximate the world which does not end. We shall take up this theme again in Section 4.

It is very important to understand that the multiplicity of equilibria is not due to the stationarity of the economy. If we imagined a non-stationary economy with each a^t near a and each (e_t^t, e_{t+1}^t) near $(e, 1-e)$, we would find the same multiplicity. One might hold the opinion that in a steady-state economy one should only pay attention to steady-state equilibria, that is, only to the autarkic and golden rule equilibria. In non-steady-state economies, there is no steady-state equilibrium to stand out among the continuum. One must face up to the multiplicity.

Let us reconsider how one might demonstrate the multiplicity of equilibria, even in a non-stationary economy. This will lead to a first economic explanation of indeterminacy similar to the one originally proposed by Gale. Suppose that in our non-stationary example we find one equilibrium $\hat{q} = (\ldots, \hat{q}_{-1}, \hat{q}_0, \hat{q}_1, \ldots)$ satisfying:

$$Z_t^{t-1}(\hat{q}_{t-1}) + Z_t^t(\hat{q}_t) = 0 \text{ for all } t \in \mathbf{Z}. \tag{6}$$

Let us look for ‘nearby’ equilibria.

We shall say that generation *t* is expectations sensitive at $\hat{q}_t$ if both $[\partial Z_t^t(\hat{q}_t) / \partial q_t] \neq 0$ and $[\partial Z_{t+1}^t(\hat{q}_t) / \partial q_t] \neq 0$. If the first inequality holds, then the young's behaviour at time *t* can be influenced by what they expect to happen at time *t*+1. Similarly, if the second inequality holds, then the behaviour of the old agent at time *t*+1 depends on the price he faced when he was young, at time *t*. Recalling the logarithmic preferences of our example, it is easy to calculate that the derivatives of excess demands, for any $q_t>0$, satisfy

$$\frac{\partial Z_t^t(q_t)}{\partial q_t} = a^t e_{t+1}^t \neq 0$$

and

$$\frac{\partial Z_{t+1}^t(q_t)}{\partial q_t} = \frac{-(1-a^t)e_t^t}{q_t^2} \neq 0.$$

Hence, by applying the implicit function theorem to (1) we know that there is a non-trivial interval I_{t-1}^F containing $\hat{q}_{t-1}$ and a function F_t with domain I_{t-1}^F such that $F_t(\hat{q}_{t-1}) = \hat{q}_t$, and more generally,

$$Z_t^{t-1}(q_{t-1}) + Z_t^t[F_t(q_{t-1})] = 0 \text{ for all } q_{t-1} \in I_{t-1}^F.$$

Similarly there is a non-trivial interval I_t^B containing $\hat{q}_t$, and a function B_t with domain I_t^B such that $B_t(\hat{q}_t) = \hat{q}_{t-1}$, and more generally, $Z_t^{t-1}[B_t(q_t)] + Z_t^t(q_t) = 0$, for all $q_t \in I_t^B$. Of course, if $F_t(q_{t-1}) = q_t \in I_t^B$, then $B_t(q_t)=q_{t-1}$.

These forward and backward functions F_t and B_t, respectively, hold the key to one understanding of indeterminacy. Choose any relative price $q_0 \in I_0^F \cap I_0^B$ between periods 0 and 1. The behaviour of the generation born at 0 is determined, including its behaviour when old at period 1. If $q_0 \neq \hat{q}_0$, and generation 1 continues to expect relative prices $\hat{q}_1$ between 1 and 2, then the period 1 market will not clear. However, it will clear if relative prices q_1 adjust so that $q_1=F_1(q_0)$. Of course, changing relative prices between period 1 and 2 from $\hat{q}_1$ to q_1 will upset market clearing at time 2, if generation 2 continues to expect $\hat{q}_2$. But if expectations change to $q_2=F_2(q_1)$, then again the market at time 2 will clear. In general, once we have chosen $q_t \in I_t^F$, we can take $q_{t+1}=F_{t+1}(q_t)$ to clear the (*t*+1) market. Similarly, we can work backwards. The change in q_0 will cause the period 0 market not to clear, unless the previous relative prices between period –1 and 0 were changed from $\hat{q}_{-1}$ to $q_{-1}=B_0(q_0)$. More generally, if we have already chosen $q_t \in I_t^B$, we can set $q_{t-1}=B_t(q_t)$ and still clear the period *t* market.

Thus we see that it is possible that an arbitrary choice of $q_0 \in I_0^F \cap I_0^B$ could lead to an equilibrium price sequence *q*. What happens at time 0 is undetermined because it depends on expectations concerning period 1, and also the past. But what can rationally be expected to happen at time 1 depends on what in turn is expected to happen at time 2, and so on.

There is one essential element missing in the above story. Even if $q_t \in I_t^F$, there is no guarantee that $q_{t+1}=F_{t+1}(q_t)$ is an element of I_{t+1}^F. Similarly, $q_t \in I_t^B$ does not necessarily imply that $q_{t-1} = B_t(q_t) \in I_{t-1}^B$. In our steady state example, this can easily be

remedied. Since all generations are alike,

$$F_t = F_1,\ \ B_t = B_0,\ \ I_t^F = I_0^F \text{ and } I_t^B = I_0^B \text{ for all } t \in \mathbf{Z}.$$

One can show that the interval $(1, \bar{q}) \subset I_0^F \cap I_0^B$, and that if $q_0 \in (1, \bar{q})$, then $F_1(q_0) \in (1, \bar{q})$, and $B_0(q_0) \in (1, \bar{q})$. This establishes the indeterminacy we claimed.

In the general case, when there are several commodities and agents per period, and when the economy is non-stationary, a more elaborate argument is needed. Indeed, one wonders, given one equilibrium $\hat{q}$ for such an economy, whether after a small perturbation to the agents there is any equilibrium at all of the perturbed economy near $\hat{q}$. We shall take this up in Section 8.

It is worth noting that we can define two more complete markets OLG economies with present value prices. In the economy $E_{0,\infty}$ only agents born at time $t \geq 1$ participate. The definition of OLG equilibrium is the same as before, except that now the set of agents is restricted to the participants, and market clearing is only required for $t \geq 1$. In the q-normalized form, equilibrium is defined by $q=(q_1, q_2,\ldots)$ such that

$$Z_1^1(q_1) = 0 Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0\ \forall t \geq 2.$$

It is immediately apparent (with one agent born per period and one good) that $E_{0,\infty}$ has a unique equilibrium, at which no agent trades and which is Pareto inefficient.

We could also define an economy $E_{0,\infty}^M$ in which only agents $t \geq 1$ participate, but where we require (in the normalized price version) that

$$Z_1^1(q_1) = -M Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0\ \forall t \geq 2.$$

Equilibrium in $E_{0,\infty}^M$ is as if we gave an outside agent who had no endowment the purchasing power of M at time 1, and still managed to clear all markets $t \geq 1$. As long as $0 \leq M \leq Z_1^0(\bar{q})$, $E_{0,\infty}^M$ has an equilibrium. Take q_0 solving $M = Z_1^0(q_0)$, and $q_1=F_1(q_0)$ and $q_t=F_t(q_{t-1})$ for $t \geq 2$. We examine these two models more closely in Section 3.

2 Endogenous cycles

Let us consider another remarkable and suggestive property that one-commodity, stationary OLG economies can exhibit. We shall call the equilibrium $q = (\ldots, q_{-1}, q_0, q_1, \ldots)$ periodic of period n if $q_0, q_1, \ldots, q_{n-1}$ are all distinct, and if for all integers i and j, $q_i=q_{i+jn}$. The possibility that a perfectly stationary economy can exhibit cyclical ups and downs, even without any exogenous shocks or uncertainty, is reminiscent of 1930s–1950s business cycle theories. In fact, it is possible to construct a robust one-commodity per period economy which has equilibrium cycles of every order n. Let us see how.

As before, let each generation t consist of one agent, with endowment $e^t = (\ldots, 0, e, 1-e, 0, \ldots)$ positive only in period t and t+1, and utility $u^t(x)=u_1(x_t)+u_2(x_{t+1})$. Again, suppose that $\bar{q} = u_2'(1-e) / u_1'(e) > 1$. It is an immediate consequence of the separability of u^t, that for $q_t \leq \bar{q}$,

$$Z_t^t(q_t) \leq 0,\ Z_{t+1}^t(q_t) \geq 0,\ \frac{\partial Z_{t+1}^t(q_t)}{\partial q_t} < 0.$$

From monotonicity, we know that $Z_{t+1}^t(q_t) \to \infty$ as $q_t \to 0$. Hence it follows that for any $0 < q_0 < \bar{q}$, there is a unique $q_{-1}=B_0(q_0)$ with

$$Z_0^{-1}[B_0(q_0)] + Z_0^0(q_0) = 0.$$

From the fact that $Z_0^0(q_0) \geq -e$ for all q_0, it also follows that there is some $\underline{q} \leq 1$ such that if $q_0 \in [\underline{q}, \bar{q}]$, then $B_0(q_0) \in [\underline{q}, \bar{q}]$. Now consider the following theorem due to the Russian mathematician Sarkovsky and to the mathematicians Li and Yorke (1975).

Sarkovsky–Li–Yorke Theorem: . Let $B: [\underline{q}, \bar{q}] \to [\underline{q}, \bar{q}]$ *be a continuous function from a nontrivial closed interval into itself. Suppose that there exist a three-cycle for B, that is, distinct points* q_0, q_1, q_2, *in* $[\underline{q}, \bar{q}]$ *with* $q_1=B(q_0)$, $q_2=B(q_1)$, $q_0=B(q_2)$. *Then there are cycles for B of every order n.*

Grandmont (1985), following related work of Benhabib and Day (1982) and Benhabib and Nishimura (1985), gave a robust example of a one-commodity, stationary economy (u_1, u_2, e) giving rise to a three-cycle for the function B_0. Of course a cycle for B_0 is also a cyclical equilibrium for the economy, hence there are robust examples of economies with cycles of all orders.

Theorem: (Benhabib–Day, 1982; Benhabib–Nishimura,1985; Grandmont, 1985). *There exist robust examples of stationary, one-commodity OLG economies with cyclical equilibria of every order n.*

This result is extremely suggestive of macroeconomic fluctuations arising for endogenous reasons, even in the absence of any fundamental fluctuations. Note first, however, that all of the cyclical equilibria, except the autarkic one-cycle $(\ldots, \bar{q}, \bar{q}, \bar{q}, \ldots)$, can be shown to be Pareto optimal (see Section 4), while the theory of macroeconomic business cycles is concerned with the welfare losses from cyclical fluctuations. (On the other hand, the fact that cyclical behaviour is not incompatible with optimality is perhaps an important observation for macroeconomics.) More significantly, it must be pointed out that Sarkovsky's theorem is a bit of a mathematical curiosity, depending crucially on one dimension. And of course non-stationary economies, even with one commodity, will typically not have any periodic cycles. By contrast, the multiplicity and suboptimality of non-periodic equilibria that we saw in Section 1 are robust properties that are maintained in OLG economies with multiple commodities and heterogeneity across time.

The main contribution of the endogenous business cycle literature is that it establishes the extremely important, suggestive principle that very simple dynamic models can have very complicated ('chaotic') dynamic equilibrium behaviour.

In the next section we turn to another phenomenon that can generally occur in overlapping generations economies, but never in finite horizon models.

3 Money and the sequential economy

Money very often has value in an overlapping generations model, but it never does in a finite horizon Arrow–Debreu model. The reason for its absence in the latter model is familiar: money would enable some agents to spend more on goods than they received from sales of their goods. But that would mean in the aggregate that spending on goods would exceed revenue from the sale of goods, contradicting market clearing in goods.

This argument can be given another form. Without uncertainty, Arrow–Debreu equilibrium can be reinterpreted as a sequential equilibrium with contemporaneous prices. But if the number of periods is finite, then in the last period the marginal utility of money to every consumer is zero, hence so is its price. In the second-to-last period nobody will pay to end up holding any money, because in the last period it will be worthless. By induction it will have no value even in the first period.

Evidently both these arguments fail in an infinite horizon setting. There is no last period, so the backward induction argument has no place to begin. And with an infinite number of consumers, aggregate spending and revenue might both be infinite, preventing us from comparing their sizes. On the other hand, there are infinite horizon models where money cannot have value. The difference between the OLG model and these other infinite horizon models will be discussed in Section 7.

Strictly speaking, the overlapping generations model we have discussed so far has been modelled along the lines of Arrow–Debreu: each agent faced only one budget constraint and equilibrium was defined as if all markets met simultaneously at the beginning of time (–∞). In such a model money has no function. However, we can define another model, similar to the first considered by Samuelson, in which agents face a sequence of budget constraints and markets meet sequentially, and where money does have a store-of-value role. Surprisingly, this model turns out to have formally the same properties as the OLG model we have so far considered. To distinguish the two models we shall refer to this latter monetary model as the Samuelson model.

Suppose that we imagine a one-good per period economy in which the markets meet sequentially, according to their dates, and not simultaneously at the beginning of time. Suppose also that there are no assets or promises to trade. In such a setting it is easy to see that there could be no trade at all, since, as Samuelson put it, there is no double coincidence of wants. The old and the young at any date *t* both have the same kind of commodity, so they have no mutually advantageous deal to strike. But as Samuelson pointed out, introducing a durable good called money, which affects no agent's utility, might allow for much beneficial trade. The old at date *t* could sell their money to the young for commodities, who in turn could sell their money when old to the next period's young. In this manner new and more efficient equilibria might be created. The ‘social contrivance of money’ is thus connected to both the indeterminacy of equilibrium and the Pareto suboptimality of equilibrium, at least near autarkic equilibria. The puzzle, we have said, is how to explain the positive price of money when it has no marginal utility.

A closer examination of the equilibrium conditions of Samuelson's sequential monetary equilibrium reveals that, although it appears much more complicated, it reduces to the timeless OLG model we have defined above, but with one difference, namely, that the budget constraint of the generation endowed with money is increased by the value of money. The introduction of the asset money thus ‘completes the markets’ in the sense of Arrow (1953), by which we mean that the equilibrium of the sequential economy can be understood as if it were an economy in which money did not appear and all the markets cleared at the beginning of time (except, as we said, that the incomes of several agents are increased beyond the value of their endowments). The puzzle of how money can have positive value in the Samuelson model can thus be reinterpreted in the OLG model as follows. How is it possible that we can increase the purchasing power of one agent beyond the value of his endowment, without decreasing the purchasing power of any other agent below his, and yet continue to clear all the markets? Before giving a more formal treatment of the foregoing, let me re-emphasize an important point. It has often been said that the paradoxical properties of equilibrium in the sequential Samuelson consumption loan model can be explained on the basis of incomplete markets. Adding money to the model, however, completes the markets, in the precise sense of Arrow–Debreu, but the result is the OLG model in which the puzzles remain.

Let us now formally define the sequential one-commodity Samuelson model with money, $E_{0,\infty}^{M,S}$. Consider a truncated economy in which there is a new agent ‘born’ at each date $t \geq 0$, whose utility depends only on the two goods dated during his lifetime, and whose endowment is positive only in those same commodities. At each date $t \geq 1$ there will be two agents alive, a young one and an old one. Let us suppose that trade does not begin until period 1, so that the date 0 generation must consume its endowment when it is young. To this truncation of our earlier model we now add one extra commodity, which we call money. Money is a perfectly durable commodity that affects no agent's utility. Agents are endowed with money (M_t^t, M_{t+1}^t), in addition to their commodity endowments.

A (contemporaneous) price system is defined as a sequence

$$(\pi, p) = (\pi_1, \pi_2, \ldots; p_1, p_2, \ldots)$$

of contemporaneous money prices π_t and contemporaneous commodity prices p_t for each $t \geq 1$. The budget set for any agent $t \geq 1$ is defined by

$$\left\{(m_t, m_{t+1}, x_t, x_{t+1}) \geq 0 \mid \pi_t m_t + p_t x_t \leq \pi_t M_t^t + p_t e_t^t \text{ and } \pi_{t+1} m_{t+1} + p_{t+1} x_{t+1} \leq \pi_{t+1} M_{t+1}^t + p_{t+1} e_{t+1}^t + \pi_{t+1} m_t\right\}.$$

For agent 0 the budget constraint is

$$\left\{(m_0, m_1, x_0, x_1) \geq 0 | m_0 = M_0^0, x_0 = e_0^0, \text{ and } \pi_1 m_1 + p_1 x_1 \leq \pi_1 M_1^0 + p_1 e_1^0 + \pi_1 m_0\right\}.$$

The budget constraints express the principle that in the Samuelson model agents cannot borrow at all, and cannot save, that is, purchase more when old than the value of their old endowment, except by holding over money m_t from when they were young. Let $m_t^t(\pi, p)$ and $m_{t+1}^t(\pi, p)$ be the utility maximizing choices of money holdings by generation *t* when young and when old. As before, the excess commodity demand is defined by $Z_t^t(\pi, p)$ and $Z_{t+1}^t(\pi, p)$.

To keep things simple, we suppose that agent 0 is endowed with $M_1^0 = M$ units of money when he is old, but all other endowments M_s^t are zero. Since money is perfectly durable, total money supply in every period is equal to *M*. Equilibrium is defined by a price sequence (π, *p*) such that for all $t \geq 1$,

$$m_t^{t-1}(\pi, p) + m_t^t(\pi, p) = M \text{ and } Z_t^{t-1}(\pi, p) + Z_t^t(\pi, p) = 0.$$

At first glance this seems a much more complicated system than before.

But elementary arguments show that in equilibrium either π_t=0 for all *t*, and there is no intergenerational trade of commodities, or π_t>0 for all *t*, or π_t<0 for all *t*. In the case where π_t>0, no generation will choose to be left with unspent cash when it dies, hence $m_{t+1}^t(\pi, p) = 0$ for all *t*, hence money market clearing is reduced to

$$m_t^t(\pi, p) = M \text{ for all } t \geq 1.$$

By homogeneity of the budget sets, if π_t>0, we might as well assume π_t=1 for all *t*. But then the prices p_t become the same as the present value prices from Section 1. From period by period Walras' Law, we deduce that, if the goods market clears at date *t*, so must the money market. So we never have to mention money market clearing.

Moreover, by taking $q_t = (\pi_t p_{t+1}) / (\pi_{t+1} p_t)$ we can write the commodity excess demands for agent $t \geq 1$ just as in Section 1, by

$$[Z_t^t(q_t), Z_{t+1}^t(q_t)]$$

and they are the same as

$$[Z_t^t(\pi, p), Z_{t+1}^t(\pi, p)].$$

The only agent who behaves differently is agent 0, whose budget set must now be written

$$B^0(\mu, M) = \left\{(x_0, x_1) | x_0 = e_0^0, x_1 \leq e_1^0 + \mu M\right\},$$

where

$$\mu = \frac{\pi_1}{p_1}.$$

We can then write agent 0's excess demand for goods at time 1 as

$$Z_1^0(\mu, q, M) = Z_1^0(\mu M) = \mu M.$$

Thus any sequential Samuelson monetary equilibrium can be described by (μ , q), $\mu \geq 0$, satisfying

$$Z_1^0(\mu M) + Z_1^1(q_1) = 0,$$

and

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0 \text{ for all } t \geq 2.$$

But of course that is precisely the same as the definition of an OLG equilibrium for $E_{0,\infty}^{\mu M}$ given in Section 1.

4 Understanding OLG economies as lack of market clearing at infinity

In this section we point out that the suboptimality of competitive equilibria, the indeterminacy of non-stationary equilibria, the non-existence of the core, and the positive valuation of money can all occur robustly in possibly non-stationary OLG economies with multiple consumers and L>1 commodities per period. We also note the important principle that the potential dimension of indeterminacy is related to L. In the two-way infinity model, it is 2L–1. In the one-way infinite model without money it is L–1; in the one-way infinity model with money the potential dimension of indeterminacy is L.

None of these properties can occur (robustly) in a finite consumer, finite horizon, Arrow–Debreu model. In what follows we shall suggest that a proper understanding of these phenomena lies in the fact that the OLG model is isomorphic, in a precise sense, to a '*-finite' model in which not all the markets are required to clear.

One of the first explanations offered to account for the differences between the Arrow–Debreu model and the sequential Samuelson model with money centred on the finite lifetimes of the agents and the multiple budget constraints each faced. These impediments to intergenerational trade (for example, the fact that an agent who is 'old' at time t logically cannot trade with an agent who will not be 'born' until time t+s) were held responsible. But as we saw in the last section, without uncertainty the presence of a single asset like money is enough to connect all the markets. Formally, as we saw, the model is identical to what we called the OLG model in which we could imagine all trade taking place simultaneously at the beginning of time, with each agent facing a single budget constraint involving all the commodities. What prevents trade between the old and the unborn is not any defect in the market, but a lack of compatible desires and resources.

Another common explanation for the surprising properties of the OLG model centres on the ‘paradoxes’ of infinity, as suggested by Shell (1971). In finite models, one proves the generic local uniqueness of equilibrium by counting the number of unknown prices, less 1 for homogeneity, and the number of market clearing conditions, less 1 for Walras’ Law, and notes that they are equal. In the OLG model there is an infinity of prices and markets, and who is to say that one infinity is greater than another? We already saw that the backward induction argument against money fails in an infinite horizon setting, where there is no last period. Surely it is right that infinity is at the heart of the problem. But this explanation does not go far enough. In the model considered by Bewley (1972) there is also an infinite number of time periods (but a finite number of consumers). In that model all equilibria are Pareto optimal, and money never has value, even though there is no last time period. The problem of infinity shows that there may be a difference between the Arrow–Debreu model and the OLG model. In itself, however, it does not predict the qualitative features (like the potential dimension of indeterminacy) that characterize OLG equilibria.

Consider now a general OLG model with many consumers and commodities per period. We index utilities $u^{t,h}$ by the time of birth t, and the household $h \in H$, a finite set. Household (t, h) owns initial resources $e_t^{t,h}$ when young, an L-dimensional vector, and resources $e_{t+1}^{t,h}$ when old, also an L-dimensional vector, and nothing else. As before utility $u^{t,h}$ depends only on commodities dated either at time t or t+1. Given prices

$$q_t = (q_{ta}, q_{tb}) \in \Delta_{++}^{2L-1} = \left\{ q \in R_{++}^{2L} \middle| \sum_{\ell=1}^{L} (q_\ell + q_{L+\ell}) = 2 \right\}$$

consisting of all the $2L$ prices at date t and t+1, each household in generation t has enough information to calculate the relevant part of its budget set

$$B^{t,h}(q_t) = \left\{ (x_t, x_{t+1}) \in R_+^{2L} | q_{ta} \cdot x_t + q_{tb} \cdot x_{t+1} \leq q_{ta} \cdot e_t^{t,h} + q_{tb} \cdot e_{t+1}^{t,h} \right\}.$$

Hence we can write household excess demand $[Z_t^{t,h}(q_t), Z_{t+1}^{t,h}(q_t)]$ and the aggregate excess demand of generation t as $[Z_t^t(q_t), Z_{t+1}^t(q_t)]$, where

$$Z_{t+s}^t(q_t) = \sum_{h \in H} Z_{t+s}^{t,h}(q_t), \; s = 0, 1.$$

Of course we need to put restrictions on the q_t to ensure their compatibility, since q_{tb} and $q_{t+1,a}$ refer to the same period t+1 prices. But this is easily done by supposing that

$$q_{tb} = \lambda_t q_{t+1a} \text{ for some } \lambda_t > 0, \; \forall t \in \mathbf{Z}.$$

Present value OLG prices p can always be recovered from the normalized prices q via the recursion

$$p_1 = q_{1a} p_t = q_{ta}(\lambda_1 \lambda_2 \ldots \lambda_{t-1}) \text{ for } t \geq 2 \; p_t = q_{ta}(\lambda_0^{-1} \lambda_{-1}^{-1} \ldots \lambda_t^{-1}) \text{ for } t \leq 0.$$

We shall now define three variations of the OLG model and equilibrium, depending on when time starts, and whether or not there is money.

Suppose first that time goes from $-\infty$ to ∞. We can write the market clearing condition for equilibrium exactly as we did in the one-commodity, one-consumer case, as

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0, \; t \in \mathbb{Z}. \tag{A}$$

Similarly we can define the one-way infinity economy $E_{0,\infty}$, in which time begins in period 0, but trade begins in time 1. We simply retain the same market clearing conditions for $t \geq 2$,

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0, \; t \geq 2 \tag{A$'$}$$

$$\sum_{h \in H} \tilde{Z}_1^{0,h}(q_{1a}) + Z_1^1(q_1) = 0, \tag{7}$$

it being understood that $Z_1^{0,h}$ has been modified to $\tilde{Z}_1^{0,h}(q_{1a})$ because every agent $(0,h)$ is forced to consume his own endowment at time 0, so that he maximizes over his budget set

$$B^{0,h}(q_{1a}) = \left\{ (x_0, x_1) \in R_+^{2L} | x_0 = e_0^{0,h}, \; q_{1a} \cdot x_1 \leq q_{1a} \cdot e_1^{0,h} \right\}.$$

Finally, let us define equilibrium in a one-way infinity model with money, $E_{0,\infty}^M$, when agents $(0,h)$ are endowed with money M^h, in addition to their commodities, by (μ, q), $\mu \geq 0$, satisfying

$$\sum_{h \in H} \tilde{Z}_1^{0,h}(q_{1a}, \mu M^h) + Z_1^1(q_1) = 0, \tag{A$''$}$$

and

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0, \text{ for } t \geq 2.$$

Again it is understood that the agents (0, *h*) born in time 0 cannot trade in time 0, and they maximize over the budget set

$$B^{0,h}(q_{1a}, \mu M^h) = \left\{(x_0, x_1) \in R_+^{2L} | x_0 = e_0^{0,h}, q_{1a} \cdot x_1 \leq q_{1a} \cdot e_1^{0,h} + \mu M^h\right\}.$$

These are the natural generalizations of the one-good economies defined in Section 1. (There is one small difference. With many agents born per period we can no longer conclude that if one agent holds a positive amount of money when young, then so must every other agent – no matter when he is born. We shall ignore this complication and allow some agents to hold negative money.)
We must now try to understand very generally why there may be many dimensions of OLG equilibria, why they might not be Pareto efficient, and how it is possible that some agents can spend beyond their budgets without upsetting market clearing.
Our explanation amounts to ‘lack of market clearing at infinity’. We illustrate this for the case $E_{0,\infty}$.
Consider the truncated economy $E_{0,T}$ consisting of all the agents born between periods 0 and *T*. Market clearing in $E_{0,T}$ is defined to be identical to that in $E_{0,\infty}$ for *t*=1 to *t*=*T*. But at *t*=*T*+1, we require $Z_{T+1}^T(q_t) = 0$ in $E_{0,T}$. This is a perfectly conventional Arrow–Debreu economy, and so necessarily has some competitive equilibria, all of which are Pareto efficient; generically its equilibrium set is a 0-dimensional manifold.
We have already seen in Section 1 what a great deal of difference there is between the economies $E_{0,T}$ (no matter how large *T* is) and $E_{0,\infty}$. The interesting point is that, by appealing to non-standard analysis, which makes rigorous the mathematics of infinite and infinitesimal numbers, one can easily show that the economy $E_{0,T}$, for *T* an infinite number, inherits any property that holds for all finite $E_{0,T}$. Thus the paradoxical properties of the economy $E_{0,\infty}$ do not stem from infinity alone, since the infinite economy $E_{0,T}$ does not have them. We shall need to modify $E_{0,T}$ before it corresponds to $E_{0,\infty}$. Nevertheless, the economies $E_{0,T}$ do provide some information about $E_{0,\infty}$.
Theorem: (Balasko–Cass–Shell, 1980; Wilson, 1981). *Under mild conditions, at least one equilibrium for $E_{0,\infty}$ always exists.*
To see why this is so, note that $E_{0,T}$ is well-defined for any finite *T*. From non-standard analysis we know that the sequence $E_{0,T}$ for $T \in \mathbf{N}$ has a unique extension to the infinite integers. Now fix *T* at an infinite integer. We know that $E_{0,T}$ has at least one equilibrium, since $E_{0,s}$ does for all finite *s*. But if *T* is infinite, $E_{0,T}$ includes all the finite markets *t*=1,2,…, so all those must clear at an equilibrium q^* of $E_{0,T}$. Taking the standard parts of the prices q_t^* for the finite *t* (and ignoring the infinite *t*) gives an equilibrium *q* for $E_{0,\infty}$.
To properly appreciate the force of this proof, we shall consider it again, when it might fail, in Section 7, where we deal with infinite lived consumers.
In terms of the existence of equilibrium, $E_{0,\infty}$ (and similarly $E_{0,\infty}^M$ and $E_{-\infty,\infty}$) behaves no differently from an Arrow–Debreu economy. But the indeterminacy is a different story.
Definition: A classical equilibrium for the economy $E_{0,T}$ is a price sequence $q^*=(q_1,\ldots, q_T)$ that clears the markets for $1 \leq t \leq T$, but at *t*=*T*+1, market clearing $Z_{T+1}^T(q_T) = 0$ is replaced by

$$Z_{T+1}^T(q_T) \leq \sum_{h \in H} e_{T+1}^{T+1,h}.$$

Thus in a classical equilibrium there is lack of market clearing at the last period. The aggregate excess demand in that period, however, must be less than the endowment the young of period T+1 would have had, were they part of the economy. Economies in which market clearing is not required in every market are well understood in economic theory. Note that in a classical equilibrium the agents born at time T are not rationed at T+1; their full Walrasian (notional) demands are met, out of the dispossessed endowment of the young. But we do not worry about how this gift from the T+1 young is obtained. The significance of our classical equilibrium for the OLG models can be summarized in the following theorem from Geanakoplos and Brown (1982):

Theorem: (Geanakoplos–Brown, 1982). *Fix T at an infinite integer. The equilibria q for $E_{0,\infty}$ correspond exactly to the standard parts of classical equilibria q^* of $E_{0,T}$.*

The Walrasian equilibria of the economy $E_{0,\infty}$, which apparently is built on the usual foundations of agent optimization and market clearing, correspond to the 'classical equilibria' of another finite-like economy $E_{0,T}$ in which the markets at T+1 ('at infinity') need not clear. The existence of a classical equilibrium in $E_{0,T}$, and thus an equilibrium in $E_{0,\infty}$, is not a problem, because market clearing is a special case of possible non-market clearing, and $E_{0,T}$, being finite-like, always has market clearing equilibria.

Thus even though the number of prices and the number of markets in $E_{0,\infty}$ are both infinite, by looking at $E_{0,T}$ it is possible to say which is bigger, and by how much. There are exactly L more prices than there are markets to clear. From Walras' Law we know that if all the markets but one clear, that must clear as well. Hence having L markets that need not clear provides for L–1 potential dimensions of indeterminacy.

Corollary: (Geanakoplos–Brown, 1982). *For a generic economy $E_{0,\infty}$, there are at most L–1 dimensions of indeterminacy in the equilibrium set.*

Though the classical equilibria of $E_{0,T}$ generically have L–1 dimensions of indeterminacy, it is by no means true that there must be L–1 dimensions of *visible* indeterminacy. If we consider any classical equilibrium q^* for a generic economy $E_{0,T}$, then we will be able to arbitrarily perturb some set of L–1 prices near their q^* values, and then choose the rest of the prices to clear all the markets up through time T. But which L–1 prices these are depends on which square submatrix N (of derivatives of excess demands with respect to prices) is invertible. For example, call the economy $E_{0,\infty}$ intertemporally separable if each generation t consists of a single agent whose utility for consumption at date t is separable from his utility for consumption at date t+1. Then the L–1 free parameters must all be chosen at date T+1 (as part of $q_{T,b}$), that is, way off at infinity.

Corollary: (Geanakoplos–Polemarchakis, 1984). *Intertemporally separable economies $E_{0,\infty}$ generically have locally unique equilibria (in the product topology).*

For example, a natural generalization of the example in Section 1 would be to generations consisting of a single Cobb–Douglas consumer of L>1 goods when young and when old. The corollary shows that this economy has no indeterminacy of equilibrium. Since Cobb–Douglas economies seem so central, one might guess that multi-good OLG economies $E_{0,\infty}$ do not generate indeterminacy. But that is incorrect. Separability with one agent drastically reduces the effect expectations about future prices can have on the present, because changes in future consumption do not change marginal utilities today. In the separable case, changing all L prices tomorrow only affects today through the one dimension of income.

Even when the L–1 degrees of freedom may be chosen at time t=1, there still may be no visible indeterminacy, if the matrix N has an inverse (in the non-standard sense) with infinite norm. But when the free L–1 parameters may be chosen at t=1 and also the matrix N has an inverse with finite norm, then all nearby economies must also display L–1 dimensions of indeterminacy.

Theorem: (Kehoe–Levine, 1984; Geanakoplos–Brown, 1982). *In the $E_{0,\infty}$ OLG model there are robust examples of economies with L–1 dimensions of indeterminacy. In the monetary economy, $E^M_{0,\infty}$, there are robust examples of economies with L dimensions of indeterminacy.*

Let us now turn our attention to the question of Pareto optimality.

Definition: An allocation $\bar{x} = (x^{t,h}; 0 \le t \le T)$ is classically feasible for the economy $E_{0,T}$ if $\sum_{(t,h)\in A} x_s^{t,h} \le \sum_{(t,h)\in A} e_s^{t,h}$, for $0 \le s \le T+1$. The classically feasible allocation $\bar{x}$ for $E_{0,T}$ is a classic Pareto optimum if there is no other classically feasible allocation $\bar{y}$ for $E_{0,T}$ with $u^t(y^{t,h}) > u^t(x^{t,h})$ for all $(t,h)\in A$ with $0 \le t \le T$, with at least one inequality $(0,h)$ representing a non-infinitesimal difference.

Theorem: (Geanakoplos–Brown, 1982). *The Pareto-optimal allocations $\bar{x}$ for the OLG economy $E_{0,\infty}$ are precisely the standard parts of classical Pareto-optimal allocations $\bar{x}^*$ for $E_{0,T}$, if T is fixed at an infinite integer.*

The upshot of this theorem is that the effective social endowment includes the commodities e_{T+1}^{T+1} of the generation born at time $s=T+1$, even though they are not part of the economy $E_{0,T}$. Since the socially available resources exceed the aggregate of private endowments, it is no longer a surprise that a Walrasian equilibrium, in which the value of aggregate spending every period must equal the value of aggregate private endowments, is not Pareto optimal.

On the other hand, this does not mean that all equilibria are Pareto suboptimal. If the (present value) equilibrium prices $p_t \to 0$, as $t \to \infty$ (or, more generally, if p_{T+1} is infinitesimal), then the value of the extra social endowment is infinitesimal, and there are no possible non-infinitesimal improvements. To see this, let $(p, \bar{x})$ be an equilibrium in present value prices for the OLG economy $E_{0,\infty}$. Consider the concave–convex programming problem of maximizing the utility of agent $(0, \bar{h})$, holding all other utilities of agents (t, h) with $0 \le t \le T$ at the levels $u^{t,h}(x^t)$ they get with $\bar{x}$, over all possible allocations in $E_{0,T}$ that do not use more resources, even at time $T+1$, than $\bar{x}$. Clearly $\bar{x}$ itself is a solution to this problem. But now let us imagine raising the constraints at time $T+1$ from

$$\sum_{h \in H} x_{T+1}^{T,h} \text{ to } \sum_{h \in H} (e_{T+1}^{T,h} + e_{T+1}^{T+1,h}).$$

What is the rate of change of the utility $u^{0,\bar{h}}$? From standard concave programming theorems, for the first infinitesimal additions to period $T+1$ resources, the rate of change of $u^{0,\bar{h}}$ is on the order of p_{T+1}, assuming p_1 is normalized to equal the marginal utility of consumption for agent $(0, \bar{h})$ at date 1. Additional resources bring decreasing benefits. This shows that if p_{T+1} is infinitesimal, then there are no possible non-infinitesimal improvements with a finite amount of extra resources.

An important example of $p_t \to 0$ occurs when the prices are summable, as they are when they decline geometrically to zero. Thus in a stationary equilibrium with a positive real interest rate, equilibrium must be Pareto efficient. Another proof of efficiency in the case of geometric present value prices is to observe that then the present value of the aggregate endowment must be finite, so the standard proof of Pareto efficiency in a finite horizon model goes through.

If p_t increases geometrically to infinity, then it is evident that equilibrium cannot be Pareto efficient. Thus, in a stationary equilibrium with a negative real interest rate, equilibrium must be Pareto inefficient.

When $p_t \nrightarrow 0$ but also does not increase exponentially to infinity, the calculation becomes much more delicate. An infinitesimal increase ε in resources at time $T+1$ can be used to increase utility of $(0, \bar{h})$ on the order of $p_{T+1}\varepsilon$, which is still infinitesimal if p_{T+1} is non-infinitesimal but finite. As the increases ε get larger, this rate of change could drop quickly, as higher derivatives come into play (assuming that agents have strictly concave utilities), leaving infinitesimal (and thus invisible) increases in utility even with a finite increase in resources. Second derivatives, and their uniformity, come into play. But this subtle case has been brilliantly dealt with:

Theorem: (Cass, 1972; Benveniste–Gale, 1975; Balasko–Shell, 1980; Okuno–Zilcha, 1980). *If agents have uniformly strictly concave utilities, and if the aggregate endowment is uniformly bounded away from* 0 *and* ∞, *then the equilibrium* $(p, \bar{x})$ *with present value prices p for an OLG economy* $E_{0,\infty}$ *is Pareto optimal if and only if* $\sum_{t=0}^{\infty} 1/\|p_t\| = \infty$.

Note that in this theorem it is the present value prices that play the crucial role. It follows immediately from this theorem that the golden rule equilibrium $q = (\ldots, 1, 1, 1, \ldots)$ for the simple one good, stationary economy of Section 1 is Pareto optimal, since the corresponding present value price sequence is also $(\ldots, 1, 1, 1, \ldots)$. In fact, a moment's reflection shows that any periodic, non-autarkic equilibrium must also be periodic in the present value prices p. Hence, as we have said before, but without a proof, the cyclical equilibria of Section 2 are all Pareto optimal.

Having explained the indeterminacy and Pareto suboptimality of equilibria for $E_{0,\infty}$ in terms of lack of market clearing at infinity, let us re-examine the monetary equilibria of OLG economies $E_{0,\infty}^{M}$, where $M = (M^h; h \in H)$ is the stock of money holdings by the agents $(0, h)$ at time 0.

The next theorem shows that any monetary equilibrium allocation of $E_{0,\infty}^{M}$ corresponds to the standard part of a non-monetary economy $E_{0,T}(z)$ obtained from $E_{0,T}$ by augmenting the endowments of the first generation $(0,h)_{h \in H}$ by a vector of goods z at time T

+1.

Definition: Let $z \in R^L$ be a vector of commodities for time T+1. Suppose that $-\sum_{h \in H} e_{T+1}^{T,h} \leq \sum_{h \in H} M^h z \leq \sum_{h \in H} e_{T+1}^{T+1,h}$. Let the augmented non-monetary economy $E_{0,T}(z, M)$ be identical to the non-monetary economy $E_{0,T}$, except that the endowment of each agent (0, h) is augmented by $M^h \cdot z$ units of commodities at time T+1.

Theorem: (Geanakoplos–Brown, 1982). *Fix an infinite integer T. The equilibria q of the monetary economy* $E_{0,\infty}^M$ *are precisely obtained by taking standard parts of full market clearing equilibria* q^* *of all the augmented non-monetary economies* $E_{0,T}(z,M)$.

The above theorem explains how it is possible to give agents (0, h) extra purchasing power without disturbing market clearing in the economy $E_{0,\infty}^M$. The answer is that the purchasing power comes from owning extra commodities at date T+1, and equilibrium in $E_{0,\infty}^M$ does not require market clearing in date T+1 commodities.

The above theorem gives another view of why there are potentially L dimensions of monetary equilibria: the augmenting endowment vector z can be chosen from a set of dimension L. It also explains how money can have positive value: it corresponds to the holding of extra physical commodities. The theorem also explains how the ‘social contrivance of money’ can lead to Pareto-improving equilibria, even in OLG economies where there is already perfect financial intermediation. The holding of money can effectively bring more commodities into the aggregate private endowment. The manifestation of the ‘real money balances’ is the physical commodity bundle z at date T+1. Money plays more than just an intermediation role.

Before concluding this section let us consider a simple generalization. Suppose that agents live for three periods. What plays the analogous role to $E_{0,T}$? The answer is that prices need to be specified through time T+2, but markets are only required to clear through time T. There are therefore 2L–1 potential dimensions of indeterminacy, even in the one-sided economy. In general, we must specify the price vector up until some time s, and then require market clearing only in those commodities whose excess demands are fully determined by those prices.

This reasoning has an important generalization to production. Suppose that capital invested at time t can combine with labour at time t+1 to produce output at time t+1, and suppose that all agents live two periods. Is there any difference between the case where labour is inelastically supplied, and the case where leisure enters the utility? In both cases the number of commodities is the same, but in the latter case the potential dimension of indeterminacy is one higher, since the supply of labour at any time might depend on further prices.

5 Land, the real rate of interest, and Pareto efficiency

Allais and Samuelson argued that the infinity of both time periods and agents radically changed the nature of equilibrium. Samuelson suggested that equilibrium might not be Pareto efficient, and that the real rate of interest might be negative, even if the economy did not shrink over time. In our one-good example from Section 1, the autarkic equilibrium has a negative real interest rate since each q_t<1, and the real interest rate is $1/q_t$–1.

They also thought that a second, new kind of equilibrium would emerge in which the real rate of interest is divorced from any of the considerations like impatience that Irving Fisher had stressed. They thought that in this new kind of equilibrium the real rate of interest would turn out to be equal to the rate of population growth, irrespective of the impatience of the consumers or the distribution of their endowments. Indeed, in the example from Section 1, the ‘golden rule’ equilibrium had real interest rate $1/q_t$ –1=0 in every period, irrespective of the utilities or the endowments, but equal to the population growth rate.

Furthermore, as we saw in Section 3, Samuelson argued that it might not be necessary for an asset to be valued according to the present value of its dividends, contradicting yet another one of Fisher's central concepts. Samuelson suggested that a piece of green paper might be worth a lot, even though it pays no dividends, because the holder might think he could sell it to somebody later, who would buy it on the expectation that he could sell it to somebody else later, ad infinitum. Later authors called this a rational bubble. It turns out that these views are incorrect if one includes in the model infinitely lived assets like land, that do pay dividends in every period.

Imagine an OLG economy as before with

$$u^t(x_t, x_{t+1}) = \frac{1}{2}\log x_t + \frac{1}{2}\log x_{t+1} (e_t^t, e_{t+1}^t) = (3, 1).$$

But let us also suppose there is one acre of *land* in the economy that produces a dividend D_t=1 apple every period for ever. Suppose the economy begins in period 1, with an old agent who owns the land and has an endowment of one apple, and a newly born agent as above. We suppose that buying the land at time *t* gives ownership of all dividends from time *t*+1 up to and including the dividends in the period in which the asset is sold. The apple dividend from the land at time 1 is owned by the old agent at time 1 (who presumably acquired the land at time 0 and hence has the claim on the apple).

At every period *t* we need to find the contemporaneous price q_t of the commodity and the price Π_t of the land.

Every agent in the economy must decide how much to consume when young, and what assets to hold when young, and how much to consume when old. The decision in old age is trivial, since the agent cannot do better than selling every asset he has and using the proceeds to buy consumption goods.

Thus for every $t \geq 1$ we can describe the decision problem of generation *t* by

$$\max_{y,z,\theta} u^t(y, z) = \frac{1}{2}\log y + \frac{1}{2}\log z \text{ such that } q_t y + \Pi_t \theta = q_t e_t^t = q_t 3\, q_{t+1} z = q_{t+1} e_{t+1}^t + \theta D_{t+1} + \Pi_{t+1}\theta = q_{t+1} 1 + \theta 1 + \Pi_{t+1}\theta.$$

For the original old generation, he optimizes simply by setting

$$x_1^0 = e_1^0 + D_1 + \Pi_1 = 1 + 1 = 2 + \Pi_1.$$

Denote the optimal choice of agents $t \geq 1$ by $(x_t^t, x_{t+1}^t, \theta^t)$. Market clearing requires for each $t \geq 2$ that consumption of the old plus consumption of the young is equal to total output of goods, and also that demand equals the supply of land

$$x_t^{t-1} + x_t^t = e_t^{t-1} + e_t^t + D_t = 1 + 3 + 1 = 5\, \theta^t = 1.$$

In period *t*=1 we must have

$$x_1^0 + x_1^1 = e_1^0 + e_1^1 + D_1 = 1 + 3 + 1 = 5,\ \theta^1 = 1.$$

Sequential equilibrium is thus a vector $(x_1^0, (q_t, \Pi_t, (x_t^t, x_{t+1}^t, \theta^t))_{t=1}^{\infty})$ satisfying the above conditions on agent maximization and market clearing.

Fisher's recipe for computing equilibrium with assets is to put the asset dividends into the endowments of their owners, and then find the usual general equilibrium with present value prices ignoring the assets. In this example that means giving agent 0 an endowment e^0=(2,1,…) of two apples in period 1 and one apple every period thereafter, and ignoring the land. Equilibrium with present value prices is then described exactly as in Section 1.

To solve for the present value prices $(p_1, p_2, \ldots)$ we can guess that since the economy is stationary, there will be stationary equilibrium $(p_1, p_2, \ldots) = (1, p, p^2, \ldots)$. For each $t \geq 2$, we must solve

$$\frac{1}{2}\frac{[3 + p1]}{p} + \frac{1}{2}[3 + p1] = 1 + 3 + 1,$$

which gives a quadratic equation

$$p^2 - 6p + 3 = 0$$

which is solved by

$$p = \frac{6 \pm (36 - 12)^{.5}}{2} = .55, r = 1/p - 1 = 81.7\%.$$

The other root is greater than one, and could not be right, because it would give a real interest rate less than zero, which would make the present value of land infinite. Hence consumption when young and old is

$$(y, z) = (1.775, 3.225).$$

Clearly these values clear the consumption market for all $t \geq 2$. We know by Walras's Law that, if all markets but one clears, then the last will as well, so we don't really have to check the period 1 market. But we will check it anyway. The present value of agent 0's endowment is

$$2 + p1 + p^2 1 + \ldots = 2 + p/(1 - p) = 3.225$$

and so indeed the period t=1 market clears.
We can now translate this general equilibrium back into a sequential equilibrium. Taking q_t=1 for every period and the real interest rate solving p=1/(1+r), the present value of land is

$$PVLand = p1 + p^2 1 + \ldots = p/(1 - p) = \frac{1}{1+r}1 + \frac{1}{(1+r)^2}1 + \ldots = 1.225.$$

In every period the old will consume their endowment of 1 plus the dividend of 1 plus the value of the land they will sell, which gives exactly 3.225. The sequential equilibrium is
$(x_1^0, (q_t, \Pi_t, (x_t^t, x_{t+1}^t, \theta^t))_{t=1}^{\infty}) = (3.225, (1, 1.225, (1.775, 3.225, 1))_{t=1}^{\infty})$.
Despite what Allais and Samuelson said, the rate of interest at the unique steady state is positive, higher than the growth rate of population. Moreover, as noted in Geanakoplos (2005), the real interest rate does respond to shocks in exactly the way Fisher argued. Consider the same model as before, but make all the consumers more impatient

$$U(y, z) = \frac{2}{3}\log y + \frac{1}{3}\log z.$$

Then our master equation would become

$$\frac{1}{3}\frac{[3+p1]}{p}+\frac{2}{3}[3+p1]=1+3+1$$

giving

$$p=.419,\ \ r=139\%,\ \ PVLand=.721.$$

As Fisher would have predicted, the real rate of interest does indeed increase, and the price of land decreases.

5.1 Pareto efficiency and bubbles

Observe that in our example the dividends of land represent 20 per cent of all endowments every period. Since the price of land must be finite, that means in any equilibrium the present value of all endowments must be finite. We know that implies equilibrium must be Pareto efficient.
Furthermore, if the value of aggregate endowments is finite, then money cannot have value and there can be no bubbles, because the old argument is correct that markets cannot clear if some agents are spending more than the value of their commodity endowments and nobody is spending less. Land makes the OLG economy look much more like an Arrow–Debreu economy.

5.2 Social security

The overlapping generations model is the workhorse model for examining social security. There is not space here to describe these studies. Observe simply that a pay-as-you-go system amounts to a simple transfer of endowments from each young person to each old person. We can immediately calculate the effects of such a transfer on our steady state interest rate and land value by recomputing the equilibrium for the OLG economy in which endowments are adjusted to (2,2) for every generation $t \geq 1$, and assuming the old generation 0 has an endowment of 2 apples at time 1 plus the land, which pays 1 apple every period. We get

$$p=.38,\ r=161\%,\ PVLand=.62.$$

This also confirms Fisher's contention that decreasing early endowments and increasing later endowments should raise the rate of interest and lower land values.
Notice that the pay-go system gives each agent the same number of apples when he is old that he gave up when he is young, which is a below market return on his original contribution. Social security lowers the utility of every agent except the first generation. Samuelson had argued that social security could make every agent better off. But his conclusion is false in the model with land.
It is often said that if only every generation had more children, social security would give better returns, since the young would be able to share the burden of helping the old. The trouble with that reasoning is that it ignores the fact that higher population and output growth would mean higher real interest rates, which tend to make the social security rate of return as bad as before relative to market interest rates. There is no space to discuss this here.

6 Demography in OLG

In America since the early 20th century, the generations have alternated in size between big and small. Everybody knows about the

baby boom and echo baby boom, but the same pattern happened before. Recently many authors have suggested that the retiring of the baby boom generation will force stock prices to fall. This has been criticized on the grounds that demography is easy to predict. If agents knew that stock prices would fall when the baby boomers retired, they would fall now. These two opposing views can be analysed in the OLG model by allowing generation sizes to fluctuate.

Suppose the small generation is exactly as before, but now we alternate that small generation with a large generation that is identical in every respect, except that it is twice as big

$$u^b(y, z) = \frac{1}{2}\log y + \frac{1}{2}\log z(e_y^b, e_z^b) = (6, 2).$$

As before, suppose that land produces 1 unit of output each period. Begin at time 1 with a small generation of young, and suppose the old owns the land.

We investigate whether the price of land and the real interest rate alternate between periods.

Let r_b be the interest rate that prevails when the big generation b is young, and r_a prevail when the small generation a is young.

Equilibrium can be reduced to two equations. The first describes market clearing for goods in odd periods when the small generation is young and the big generation is old, and the second equation describes market clearing in even periods, when the big generation is young and the small generation is old. As before, we let p_a=1/(1+r_a) and p_b=1/(1+r_b). Then

$$\frac{1}{2}\frac{[6 + p_b 2]}{p_b} + \frac{1}{2}[3 + p_a 1] = 2 + 3 + 1\frac{1}{2}\frac{[3 + p_a 1]}{p_a} + \frac{1}{2}[6 + p_b 2] = 1 + 6 + 1.$$

These can be simultaneously solved to get

$$p_a = .418, r_a = 139\%, PV_{Land\,a} = 1.29$$

$$p_b = .912, r_b = 9.6\%, PV_{Land\,b} = 2.09.$$

It is evident that the price of land is higher in the periods when b is young, since the interest rate is lower. Even though it is perfectly anticipated that when the big generation gets old, the price of land will fall, the price does not fall earlier because the interest rate is so low. (This point has been made by Geanakoplos, Magill and Quinzi, 2004.)

7 Impatience and uniform impatience

We have already suggested that it is useful in understanding the OLG model to consider variations, for example in which consumers live for ever. By doing so we shall also gain an important perspective on what view of consumers is needed to restore the usual properties of neoclassical equilibrium to an infinite horizon setting, a subject to which we return in Section 8.

Let us now allow for consumers $t \in A$ who have endowments e^t that may be positive in all time periods, and also for arbitrary utilities u^t defined on uniformly bounded vectors $x \in L = \mathbf{R}_+^{\mathbf{N}}$. For ease of notation we assume one good per period. A minimal assumption we need about utilities u^t is continuity on finite segments, that is, fixing x_s for all $s>n$, $u^t(x)$ should be continuous in (x_1,

…, x_n). We also need continuity on L, in some topology, but we will not go into these details. We also assume $\Sigma_{t\in A} e^t$ is uniformly bounded. In short, we suppose consumers may live for ever.

We shall find that in order to have Walrasian equilibria the consumers must be impatient. Suppose we try to form the truncated economy $E_{0,T}$ as before, say for T finite. Since utility potentially depends on every commodity, we could not define excess demands in $E_{0,T}$ unless we knew all the prices. To make it into a finite economy, let us call $E'_{0,T}$ the version of $E_{0,T}$ in which every agent is obliged to consume his initial endowment during periods $t \leq 0$ and $t>T$. Clearly $E'_{0,T}$ has an equilibrium. For this to give information about the original economy $E_{0,\infty}$, we need that consumers do not care very much about what happens to them after T, as T gets very far away. This requires a notion of impatience.

For any vector x, let ${}_n\hat{x}$ be the vector which is zero for $t>n$, and equal to x up until n. Thus ${}_n\hat{x}$ is the initial n-segment of x. To say that agent $t\in A$ is impatient means that for any two uniformly bounded consumption streams x and y, if $u^t(x) > u^t(y)$, then for all big enough n, $u^t({}_n\hat{x}) > u^t(y)$. Let us suppose that all consumers are impatient. If these segments can be taken uniformly across agents, then we say the economy is uniformly impatient. Any finite economy with impatient consumers is uniformly impatient. Note that the OLG agents are all impatient, since none of them cares about consumption after he dies, but the economy is not uniformly impatient.

Even with an economy consisting of all impatient consumers, the truncation argument, applied at an infinite $E'_{0,T}$, does not guarantee the existence of an equilibrium. For, once we take standard parts, ignoring the infinitely dated commodities, it may turn out that the income from the sale of an agent's endowed commodities at infinite t, which he used to finance his purchase of commodities at finite t, is lost to the agent. It must also be guaranteed that the equilibria of $E'_{0,T}$ give infinitesimal total value to the infinitely dated commodities. Wilson (1981) has given an example of an economy, composed entirely of impatient agents, that does not have an equilibrium precisely for this reason.

On the other hand, if there are only finitely many agents, even if they are infinitely lived, then we have:

Theorem: (Bewley, 1972). *Let the economy E be composed of finitely many, impatient consumers. Then there exists an equilibrium, and all equilibria are Pareto optimal.*

The Pareto efficiency of equilibria in these Bewley economies can be derived from the standard proof of efficiency: since there is a finite number of agents, the value of the aggregate endowments is a finite sum of finite numbers, and therefore finite itself.

In the special case with separable, commonly discounted utilities of the form $u^h(x) = \Sigma_{t=0}^{\infty} \delta^t v^h(x_t)$, with $\delta<1$, we have:

Theorem: (Kehoe–Levine, 1985). *In finite agent, separable commonly discounted utility economies, there is generically a finite number of equilibria.*

This theorem has been extended by Shannon (1999) and Shannon and Zame (2002).

Returning to the case of an infinite number of consumers, Pareto efficiency of equilibria, if they exist, can be guaranteed as long as a finite number of the agents collectively hold a non-negligible fraction of total endowment. But that also would guarantee the existence of equilibrium, since in the economy $E'_{0,T}$ we would then get the summability of the prices, meaning the endowments at infinity would have zero value, as Wilson (1981) pointed out.

It is extremely interesting to investigate the change in behaviour of an economy that evolves from individually impatient to uniformly impatient. Wilson (1981) considered an example with one infinitely lived agent, and infinitely many, overlapping, finite-lived agents, and showed that equilibria must exist, and all must be Pareto efficient. By the foregoing remarks, no matter what the proportion of sizes of the two kinds of consumers, equilibria must exist and be Pareto efficient. Muller and Woodford (1988) showed in a particular case that, when the single agent's proportion of the aggregate endowment is low enough, there is a continuum of equilibria, but if it is high enough there is no local indeterminacy.

8 Comparative statics for OLG economies

A celebrated theorem of Debreu assets that almost any Arrow–Debreu economy is regular, in the sense that it has a finite number of equilibria, each of which is locally unique. Small changes to the underlying structure of the economy (tastes, endowments, and so on) produce small, unique changes in each of the equilibria.

We have already seen that there are robust OLG economies with a continuum of equilibria. If attention is focused on one of them, how can one predict to which of the continuum of new equilibria the economy will move if there is a small change in the underlying structure of the economy, perhaps caused by deliberate government intervention? In what sense is any one of the new equilibria

near the original one? In short, is comparative statics possible?

It is helpful at this point to recall that the OLG model is, in spirit, meant to represent a dynamic economy. Trade may occur as if all the markets cleared simultaneously at the beginning of time, but the economy is equally well described as if trade took place sequentially, under perfect foresight or rational expectations. Indeed, this is surely what Samuelson envisaged when he introduced money as an asset into his model. Accordingly, when a change occurs in the underlying structure of the economy, we can interpret it as if it came announced at the beginning of time, or as if it appeared at the date on which it actually affects the economy.

We distinguish two kinds of changes to the underlying structure of an economy $\bar{E}_{-\infty,\infty}$ starting from an equilibrium $\bar{q}$. Perfectly anticipated changes, after which we would look for a new equilibrium that cleared all the markets from the beginning of time, represent one polar case, directly analogous to the comparative statics experiments of the Arrow–Debreu economy. At the other extreme we consider perfectly *un*anticipated changes, say at date t=1. Beginning at the original economy and equilibrium $\bar{q} = (\ldots, \bar{q}_{-1}, \bar{q}_0, \bar{q}_1, \ldots)$, we would look, after the change from $\bar{E}_{-\infty,\infty}$ to $E_{-\infty,\infty}$, at time t=1 (say to the endowment or preferences of the generation born at time 1), for a price sequence $q = (\ldots, q_{-1}, q_0, q_1, \ldots)$ in which $q_t = \bar{q}_t$ for $t \leq 0$, and $Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0$ for $t \geq 2$. But at date t=1 we would require q_1 to satisfy $Z_1^0(q_{1a}|\bar{q}_0) + Z_1^1(q_1) = 0$, where $Z_1^0(q_{1a}|\bar{q}_0)$ represents the excess demand of the old at time 1, given that when they were young they purchased commodities on the strength of the conviction that they could surely anticipate prices $\bar{q}_{0b}$ when they got old, only to discover prices q_{1a} instead.

To study these two kinds of comparative statics, we must describe what we mean by saying that two price sequences are nearby. Our definition is based on the view that a change at time t=1 ought to have a progressively smaller impact the further away in time from t=1 we move. We say that q is near $\bar{q}$ if the difference $|q_t - \bar{q}_t|$ declines geometrically to zero, both as $t \to \infty$ and as $t \to -\infty$.

We have already noted in Section 1 that the multiplicity of OLG equilibria is due to the fact that at any time t the aggregate behaviour of the young generation is influenced by their expectations of future prices, which (under the rational expectations hypothesis) depends on the next generation's expectations, and so on. Accordingly we restrict our attention to generations whose aggregate behaviour Z^t satisfies the expectations sensitivity hypothesis:

$$\operatorname{rank}\frac{\partial Z_t^t(p_t, p_{t+1})}{\partial p_{t+1}} = \operatorname{rank}\frac{\partial Z_{t+1}^t(p_t, p_{t+1})}{\partial p_t} = L.$$

For economies composed of such generations we can apply the implicit function theorem, exactly as in Section 1, around any equilibrium q to deduce the existence of the forward and backward functions F_t and B_t. We write their derivatives at $\bar{q}$ as D_t and D_t^{-1}, respectively.

For finite Arrow–Debreu economies, Debreu gave a definition of regular equilibrium based on the derivative of excess demand at the equilibrium. He showed that comparative statics is sensible at a regular equilibrium, and then he showed that a ‘generic’ economy has regular equilibria. We follow the same program.

We say that the equilibrium $\bar{q}$ for the expectations sensitive OLG economy $\bar{E}$ is Lyapunov regular if the long-run geometric mean of the products $D_t^* D_t D_{t-1}^* D_{t-1} \cdots D_1^* D_1$ and $D_{-t}^{-1*} D_{-t}^{-1} \cdots D_{-1}^{-1*} D_{-1}^{-1}$ converge and if to these products we can associate $2L-1$ eigenvalues, called Lyapunov exponents. The equilibrium is also non-degenerate if in addition none of these Lyapunov exponents is equal to 1.

Theorem: (Geanakoplos–Brown, 1985). *Let* $\bar{E} = \bar{E}_{-\infty,\infty}$ *be an expectations-sensitive economy with a regular non-degenerate equilibrium* $\bar{q}$. *Then for all sufficiently small perfectly anticipated perturbations E of* $\bar{E}$ (*including* $\bar{E}$ *itself*) *E has a unique equilibrium q near* $\bar{q}$.

Thus the comparative statics of perfectly anticipated changes in the structure of $\bar{E}$, around a regular, non-degenerate equilibrium, is directly analogous to the Arrow–Debreu model. The explanation for the theorem is that a perfectly anticipated change at time 0 gives rise to price changes that have a forward stable manifold (on which prices converge exponentially back to where they started) and a backward stable manifold, and that there is only one price at time 0 that is on both the forward and the backward stable manifolds. Note, incidentally, that one implication of the above theorem is that neutral policy changes, like jawboning or changing animal spirits, that is, those for which $\bar{q}$ itself remains an equilibrium, cannot have any effect if they are perfectly anticipated and move the economy to nearby equilibria.

Theorem: (Geanakoplos–Brown, 1985). *Let* $\bar{E}$ *be an expectations-sensitive economy with a regular equilibrium* $\bar{q}$. *Then, for all sufficiently small perfectly unanticipated perturbations E of* $\bar{E}$ (*including* $\bar{E}$ *itself*), *the set of unanticipated equilibria q of E near* $\bar{q}$ *is either empty, or a manifold of dimension r,* $0 \leq r \leq L$ (*L*–1 *if there is no money in the economy*), *where r is independent of the perturbation*.

The above theorem allows for the possibility that an unanticipated change may force the economy onto a path that diverges from the original equilibrium; the disturbance could be propagated and magnified through time. And if there are nearby equilibria, then there may be many of them. (Indeed, that is basically what was shown in Section 4.) In particular, an unanticipated neutral policy change could be compatible with a continuum of different equilibrium continuations. The content of the theorem is that, if there is a multiplicity of equilibrium continuations, it is parameterizable. In other words, the same *r* variables can be held fixed, and for any sufficiently small perturbation, there is exactly one nearby equilibrium which also leaves these *r* variables fixed. We shall discuss the significance of this in the next section.

This last theorem was proved first, in the special case of steady-state economies, by Kehoe–Levine, in the same excellent paper to which we have referred already several times. The theorem quoted here, together with the previous theorem on the comparative statics of perfectly anticipated policy changes, refers to economies in which the generations may be heterogeneous across time.

Let us suppose that *A* is a compact collection of generational characteristics, all of which obey the expectations-sensitive hypothesis. Let us suppose that each generation's characteristics are drawn at random from *A*, according to some Borel probability measure. If the choices are made independently across time, then the product measure describes the selection of economies. Almost any such collection will have a complex demographic structure, changing over time. The equilibrium set is then endogenously determined, and will be correspondingly complicated. It can be shown, however, that

Theorem: (Geanakoplos–Brown, 1985). *If the economy E is randomly selected, as described above, then with probability* 1, *E has at least one Lyapunov regular equilibrium*.

Note that the regularity theory for infinite economies stops short of Arrow–Debreu regularity. In the finite economies, with probability one all the equilibria are regular.

9 Keynesian macroeconomics

Keynesian macroeconomics is based in part on the fundamental idea that changes in expectations, or animal spirits, can affect equilibrium economic activity, including the level of output and employment. It asserts, moreover, that publicly announced government policy also has predictable and significant consequences for economic activity, and that therefore the government should intervene actively in the marketplace if investor optimism is not sufficient to maintain full employment.

The Keynesian view of the indeterminacy of equilibrium and the efficacy of public policy has met a long and steady resistance, culminating, in the sharpest attack of all, from the so-called new classicals, who have argued that the time-honoured microeconomic methodological premises of agent optimization and market clearing, considered together with rational expectations, are logically inconsistent with animal spirits and the non-neutrality of public monetary and bond-financed fiscal policy.

The foundation of the new classical paradigm is the Walrasian equilibrium model of Arrow–Debreu, in which it is typically possible to prove that all equilibria are Pareto optimal and that the equilibrium set is finite; at least locally, the hypothesis of market clearing fixes the expectations of rational investors. In that model, however, economic activity has a definite beginning and end. Our point of view is that for some purposes economic activity is better described as a process without end. In a world without a definite end, there is the possibility that what happens today is underdetermined, because it depends on what people expect to happen tomorrow, which in turn depends on what people tomorrow expect to happen the day after tomorrow, and so on.

Consider the simple one-good per period overlapping generations economy with money $E_0^{M,S}$, which we discussed in Section 3. Generation 0 is endowed with money when old, and equilibrium can be described with the contemporaneous commodity prices $\bar{p} = (\bar{p}_1, \bar{p}_2, \ldots)$ where we take the price of money to be fixed at 1. (In this case, as we saw in Section 3, contemporaneous prices are also present value prices.) It is helpful to reinterpret the model as a simple production economy. Imagine that the endowment e_t^t in the first period of life is actually labour, which can be transformed into output, y_t, according to the production function, $y_t=e_t$. We would then think of any purchases of goods by the old generation as demand for real output to be produced by the young. The young in turn now derive utility from leisure in their youth and consumption in their old age. Equilibrium in which consumption of the old is higher can be interpreted as an equilibrium with less leisure and higher output.

The indeterminacy of rational expectations equilibrium has the direct interpretation that optimistic expectations *by themselves* can cause the economy's *output* to expand or contract. In short the economy has an inherent volatility. The Keynesian story of animal spirits causing economic growth or decline can be told without invoking irrationality or non-market clearing.

In fact, the indeterminacy of equilibrium expectations is especially striking when seen as a response to public (but unanticipated)

policy changes. Suppose the economy is in a long-term rational expectations equilibrium $\bar{p}$, when at time 1 the government undertakes some expenditures, financed, say, by printing money. How should rational agents respond? The environment has been changed, and there is no reason for them to anticipate that $(\bar{p}_2, \bar{p}_3, \ldots)$ will still occur in the future. Indeed, in models with more than one commodity (such as we will shortly consider) there may be no equilibrium (p_1, p_2, p_3,…) in the new environment with $p_2 = \bar{p}_2$, $p_3 = \bar{p}_3$, and so on. There is an ambiguity in what can be rationally anticipated.

We argue that it is possible to explain the differences between Keynesian and monetarist policy predictions by the assumptions each makes about expectational responses to policy, and not by the one's supposed adherence to optimization, market clearing, and rational expectations, and the other's supposed denial of all three.

Consider now the government policy of printing a small amount of money, ΔM, to be spent on its own consumption of real output – or equivalently to be given to generation t=0 (when old) to spend on its consumption. Imagine first that agents are convinced that this policy is not inflationary, that is, that $\bar{p}_1$ will remain the equilibrium price level during the initial period of the new equilibrium. This will give generation t=0 consumption level $(M + \Delta M) / \bar{p}_1$. As long as ΔM is sufficiently small and the initial equilibrium was one of the Pareto-suboptimal equilibria described in Section 1, there is indeed a new equilibrium price path p beginning with $p_1 = \bar{p}_1$. Output at time 1 rises by $\Delta M / \bar{p}_1$, and in fact this policy is Pareto improving. On the other hand, imagine instead that agents are convinced that the path of real interest rates $p_t/p_{t+1}-1$ will remain unchanged. In this economy, price expectations are a function of p_1. Recalling the initial period market-clearing equation, it is clear that p_1 and all future prices rise proportionally to the growth in the money stock. The result is that output is unchanged and the old at t=1 must pay for the government's consumption. If the government's consumption gives no agent utility, the policy is Pareto worsening.

This model is only a crude approximation of the differences between Keynesian and monetarist assumptions about expectations and policy. It is quite possible to argue, for example, that holding $p_2 / p_1 = \bar{p}_2 / \bar{p}_1$ (the future inflation rate) fixed is the natural *Keynesian* assumption to make. This ambiguity is unavoidable when there is only one asset into which the young can place their savings. We are thereby prevented from distinguishing between the inflation rate and the interest rate. Our model must be enriched before we can perform satisfactory policy analysis. Nevertheless, the model conveys the general principle that expected price paths are not locally unique. There is consequently no natural assumption to make about how expectations are affected by policy. A sensible analysis is therefore impossible without externally given hypotheses about expectations. These can be Keynesian, monetarist, or perhaps some combination of the two.

Geanakoplos and Polemarchakis (1985) build just such a richer model of macroeconomic equilibrium by adding commodities, including a capital good, and a neoclassical production function. With elastically supplied labour, there are two dimensions of indeterminacy. It is therefore possible to fix both the nominal wage, and the firm's expectations (‘animal spirits’), and still solve for equilibrium as a function of policy perturbations to the economy. These institutional rigidities are more convincingly Keynesian, and they lead to Keynesian policy predictions. Moreover, taking advantage of the simplicity of the two-period lived agents, the analysis can be conducted entirely through the standard Keynesian (Hicksian) IS–LM diagram.

Keynesians themselves often postulate that the labour market does not clear. For Keynesians, lack of labour market clearing has at least a threefold significance, which it is perhaps important to sort out. First, since labour is usually taken to be inelastically supplied, it makes it possible to conceive of (Keynesian) equilibria with different levels of output and employment. Second, it makes the system of demand and supply underdetermined, so that endogenous variables like animal spirits (that is, expectations) which are normally fixed by the equilibrium conditions can be volatile. Third, it creates unemployment that is involuntary. By replacing lack of labour market clearing at time 1 with elastic labour supply and lack of market clearing ‘at infinity’ one can drop what seems to many an ad hoc postulate, yet retain at least the first two desiderata of Keynesian analysis.

10 Neoclassical equilibrium vs. classical equilibrium

The Arrow–Debreu model of general equilibrium, based on agent optimization, rational expectations, and market clearing, is universally regarded as the central paradigm of the neoclassical approach to economic theory. In the Arrow–Debreu model, consumers and producers, acting on the basis of individual self-interest, combine, through the aggregate market forces of demand and supply, to determine (at least locally) the equilibrium distribution of income, relative prices, and the rate of growth of capital stocks (when there are durable goods). The resulting allocations are always Pareto optimal.

Classical economists at one time or another have rejected all of the methodological principles of the Arrow–Debreu model. They replace individual interest with class interest, ignore (marginal) utility, especially for waiting, doubt the existence of marginal product, and question whether the labour market clears. But by far the most important difference between the two schools of thought is the classical emphasis on the long-run reproduction of the means of production, in a never-ending cycle.

Thus the celebrated classical economist Sraffa writes in Appendix D to his book:

> It is of course in Quesnay's *Tableau Economique* that is found the original picture of the system of production and consumption as a circular process, and it stands in striking contrast to the view presented by modern theory, of a one-way avenue that leads from ‘Factors of Production’ to ‘Consumption Goods.’

The title of his book, *Production of Commodities by Means of Commodities*, itself suggests a world that has no definite beginning, and what is circular can have no end.

In the Arrow–Debreu model time has a definite end. As we have seen, that has strong implications. With universal agreement about when the world will end, there can be no reproduction of the capital stock. In equilibrium it will be run down to zero. Money, for example, can never have positive value. Rational expectations will fix, at each moment, and for each kind of investment, the expected rate of profit.

In the classical system, by contrast, the market does not determine the distribution of income. Sraffa (1960, p. 33) writes:

> The rate of profits, as a ratio, has a significance which is independent of any prices, and can well be ‘given’ before the prices are fixed. It is accordingly susceptible of being determined from outside the system of production, in particular by the money rates of interest. In the following sections the rate of profits will therefore be treated as the independent variable.

Other classical writers concentrate instead on the real wage as determined outside the market forces of supply and demand, for example by the level of subsistence or the struggle between capital and labour. Indeterminacy of equilibrium seems at least as central to classical economists as it is to Keynesians.

Like Keynesians, classicals often achieve indeterminacy in their formal models by allowing certain markets not to clear in the Walrasian sense. (Again like Keynesians, the labour market is usually among them.) Thus we have called the equilibrium in Section 4 in which some of the markets were allowed not to clear a ‘classical equilibrium’.

What the OLG model shows is that, by incorporating the classical view of the world without definite beginning or end, it is possible to maintain all the neoclassical methodological premises and yet still leave room for the indeterminacy which is the hallmark of both classical and Keynesian economics. In particular this can be achieved while maintaining labour market clearing. The explanation for this surprising conclusion is that the OLG model is isomorphic to a finite-like model in which indeed not all the markets need to clear. But far from being the labour markets, under pressure to move towards equilibrium from the unemployed clamouring for jobs, these markets are off ‘at infinity’, under no pressure towards equilibrating.

We have speculated that, once one has agreed to the postulate that the resources of the economy are potentially as great at any future date as they are today, then uniform impatience of consumers is the decisive factor, according to Walrasian principles, which may influence whether the market forces of supply and demand determine a locally unique, Pareto-optimal equilibrium, or leave room for extra-market forces to choose among the continuum of inefficient equilibria. In these terms, the Arrow–Debreu model supposes a short-run impatient economy, and OLG a long-run patient economy.

11 Sunspots

So far we have not allowed uncertainty into the OLG model. As a result we found no difference in interpreting trade sequentially, with each agent facing two budget constraints, or ‘as if’ the markets all cleared simultaneously at the beginning of time, with each agent facing one budget constraint. Once uncertainty is introduced these inpts become radically different. In either case, however, there is a vast increase in the number of commodities, and hence in the potential for indeterminacy.

If we do not permit agents to make trades conditional on moves of nature that occur before they are born, then agents will have different access to asset markets. Even in finite horizon economies, differing access to asset markets has been shown by Cass and Shell (1983) to lead to ‘sunspot effects’.

A ‘sunspot’ is a visible move of nature which has no real effect on consumers, on account of preferences, or endowments, or through production. In the Arrow–Debreu model it also could have no effect on equilibrium trade; this is no longer true when access to asset markets differs.

The sunspot effect is intensified when combined with the indeterminacy that can already arise in an OLG economy. Consider the simple one good, steady state OLG economy of Section 2. Suppose that there is an equilibrium two cycle in present value prices $p = (\ldots, p_{-1}, p_0, p_1, \ldots)$ with $p_{2t}=p^S$ and $p_{2t+1}=p^R$, for all $t \in \mathcal{T}$. Now suppose that the sun is known to shine on even periods, and hide behind rain on odd periods. The above equilibrium is perfectly correlated with the sun, even though no agent's preferences or

endowments are. As usual, the same prices for $t \geq 0$ support an equilibrium, given the right amount of money, in $E_{0,\infty}^{M,S}$.
More generally, suppose that the probability of rain or shine, given the previous period's weather, is given by the Markov matrix $\pi = (\pi_{SS}, \pi_{SR}, \pi_{RS}, \pi_{RR})$. A steady state equilibrium for $E_{0,\infty}^{M,S}$, given π, is an assignment of a money price for the commodity, depending only on that period's weather, such that, if all agents maximize their expected utility with respect to π, then in each period the commodity market and money market clears. Azariadis (1981) essentially showed that, if there is a two-cycle of the certainty economy, then there is a continuum of steady state sunspot equilibria.
The sunspot equilibria, unlike the cyclical equilibria of Section 2, are Pareto suboptimal whenever the matrix π is non-degenerate.
The combination of the dynamic effects of the infinite horizon OLG model with the burgeoning theory of incomplete markets under real uncertainty, is already on the agenda for the next generation's research.

See Also

- Arrow–Debreu model of general equilibrium

Bibliography

Allais, M. 1947. *Economie et intérêt*. Paris: Imprimerie Nationale.

Arrow, K.J. 1953. The role of securities in the optimal allocation of risk-bearing. Repr. in K.J. Arrow, *Essays in the theory of Risk Bearing*. Chicago: Markham, 1971.

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25, 380–96.

Balasko, Y., Cass, D. and Shell, K. 1980. Existence of competitive equilibrium in a general overlapping-generations model. *Journal of Economic Theory* 23, 307–22.

Balasko, Y. and Shell, K. 1980. The overlapping-generations model, I: the case of pure exchange without money. *Journal of Economic Theory* 23, 281–306.

Benhabib, J. and Day, R.H. 1982. A characterization of erratic dynamics in the overlapping generations model. *Journal of Economic Dynamics and Control* 4, 37–44.

Benhabib, J. and Nishimura, N. 1985. Competitive equilibrium cycles. *Journal of Economic Theory* 35, 284–306.

Benveniste, L. and Gale, D. 1975. An extension of Cass' characterization of infinite efficient production programs. *Journal of Economic Theory* 10, 229–38.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4, 514–40.

Cass, D. 1972. On capital overaccumulation in the aggregative, nonclassical model of economic growth. *Journal of Economic Theory* 4, 200–23.

Cass, D. and Shell, K. 1983. Do sunspots matter? *Journal of Political Economy* 91, 183–227.

Cass, D. and Yaari, M.E. 1966. A re-examination of the pure consumption loan model. *Journal of Political Economy* 74, 200–23.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Diamond, P.A. 1965. National debt in a neo-classical growth model. *American Economic Review* 55, 1126–50.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Gale, D. 1973. Pure exchange equilibrium of dynamic economic models. *Journal of Economic Theory* 6, 12–36.

Geanakoplos, J. 1978. Sraffa's *Production of Commodities by Means of Commodities*: indeterminacy and suboptimality in neoclassical economics. Working Paper, RIAS. Also Chapter 3 in ‘Four Essays on the Model of Arrow–Debreu’, Ph.D. thesis, Harvard University, 1980.

Geanakoplos, J. 2005. The ideal inflation indexed bond and Irving Fisher's Theory of Interest with Overlapping Generations. In *Celebrating Irving Fisher: The Legacy of a Great Economist*, ed. R. Dimand and J. Geanakoplos. *Journal of Economics and Sociology* 64, 257–305. Oxford: Blackwell.

Geanakoplos, J. and Brown, D. 1982. Understanding overlapping generations economies as lack of market clearing at infinity. Mimeo, Yale University, revised 1985, 1986.

Geanakoplos, J. and Brown, D. 1985. Comparative statics and local indeterminacy in OLG economies: an application of the multiplicative ergodic theorem. Discussion Paper No. 773, Cowles Foundation.

Geanakoplos, J., Magill, M. and Quinzi, M. 2004. Demography and the long-run predictability of the stock market. *Brookings Papers on Economic Activity* 2004(1), 241–325.

Geanakoplos, J. and Polemarchakis, H.M. 1984. Intertemporally separable overlapping generations economies. *Journal of Economic Theory* 34, 207–15.

Geanakoplos, J. and Polemarchakis, H.M. 1985. Walrasian indeterminacy and Keynesian macroeconomics. *Review of Economic Studies* 53, 755–99.

Grandmont, J.M. 1985. Endogenous, competitive business cycles. *Econometrica* 53, 995–1046.

Kehoe, T.J. and Levine, D.K. 1984. Regularity in overlapping generations exchange economies. *Journal of Mathematical Economics* 13, 69–93.

Kehoe, T.J. and Levine, D.K. 1985. Comparative statics and perfect foresight in infinite horizon models. *Econometrica* 53, 433–53.

Kehoe, T.J., Levine, D.K. and Romer, P.M. 1990. Determinacy of equilibrium in dynamic models with finitely many consumers. *Journal of Economic Theory* 50, 1–21.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Li, T.Y. and Yorke, J.A. 1975. Period three implies chaos. *American Mathematical Monthly* 8, 985–92.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 102–21.

Muller, W.J. and Woodford, M. 1988. Determinacy of equilibrium in stationary economies with both finite and infinite lived consumers. *Journal of Economic Theory* 46, 255–90.

Okuno, M. and Zilcha, I. 1980. On the efficiency of a competitive equilibrium in infinite horizon monetary economies. *Review of Economic Studies* 47, 797–807.

Samuelson, P.A. 1958. An exact consumption loan model of interest, with or without the social contrivance of money. *Journal of*

Political Economy 66, 467–82.

Shannon, C. 1999. Determinacy of competitive equilibria in economies with many commodities. *Economic Theory* 14, 29–87.

Shannon, C. and Zame, B. 2002. Quadratic concavity and determinacy of equilibrium. *Econometrica* 70, 631–62.

Shell, K. 1971. Notes on the economics of infinity. *Journal of Political Economy* 79, 1002–11.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Wilson, C. 1981. Equilibrium in dynamic models with an infinity of agents. *Journal of Economic Theory* 24, 95–111.

How to cite this article

Geanakoplos, John. "overlapping generations model of general equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000067> doi:10.1057/9780230226203.1236

The New Palgrave Dictionary of Economics Online

Overstone, Lord [Samuel Jones Loyd] (1796 – 1883)

D.P. O'Brien

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balance of payments; Bank Charter Act (1844); Banking principle; Currency principle; endogenous trade cycle; liquidity preference; monetary base; Overstone, Lord; Palmer rule; specie-flow mechanism

Article

Samuel Loyd (the single ‘l’ seems to have been a device adopted by his father to shake off Welsh relatives), Lord Overstone, was born on 25 September 1796, the son of Lewis Loyd, a Unitarian minister turned banker, and Sarah Loyd (née Jones), the daughter of a Manchester banker. Lewis Loyd's drive and ability transformed an obscure provincial bank into a major concern. An MP from 1819 to 1826, Overstone only began to devote himself seriously to banking after the death of his mother in 1821. Though perhaps lacking his father's flair, he was a shrewd and successful banker, influential with his contemporaries. He retired from business only in 1850, on his elevation to the peerage by Lord John Russell.

In 1837 he entered, with considerable effectiveness, the arena of monetary controversy with his *Reflections suggested by a perusal of Mr. J. Horsley Palmer's pamphlet on the Causes and Consequences of the pressure on the Money Market*. This was not his first statement on the matter; he had been a witness before the 1832 committee on the renewal of the Bank Charter. But it was the start of his pre-eminence as a monetary writer, a pre-eminence which was to prove decisive in the debates leading up to the renewal of the Bank Charter in 1844 and which shaped the institutional framework of British monetary policy from that time until the First World War. Overstone's monetary thought starts from a position that the economy contains an endogenous trade cycle – he was indeed one of the first people to identify the stages of the cycle. Monetary policy could then be procyclic, responding to the needs of customers (the Banking principle) or it could act counter-cyclically so as to stabilize the level of prices and activity (the Currency principle). The theoretical position underlying the latter was as follows. In the upswing of the cycle money income rose, exports were less competitive, and a balance of payments deficit developed. Counter-cyclical contraction of the currency, in line with the loss of specie through the balance of payments deficit, would then moderate the upswing and prevent it getting out of

hand. Conversely, in the lower half of the cycle, with a balance of payments surplus, the money supply would be increased. (O'Brien, 1971; O'Brien, 1975; Wood, 1939)

The origins of this position were threefold: Hume's theory of the balance of payments, positing a direct link between the money supply, the price level, exports, and imports: the Ricardian theory of the equilibrium distribution of the precious metals (that when countries were in relative money income equilibria, there would be no net flows of precious metal) deriving from Hume; and the Ricardian definition of ‘excess’. The last is particularly crucial. If specie was flowing out then, by definition, there was excess currency. This idea leads in turn to the principle, formulated in 1826 by several writers, of ‘metallic fluctuation’: a paper currency should fluctuate in amount exactly as an identically circumstanced metallic one would do.

On this basis Overstone emerged as a critic of the ‘Palmer Rule’ under which the Bank allowed drains of specie to fall on deposits equally with notes: unless deposits were as important as notes in correcting the price level in relation to the balance of payments, the drain might exhaust the specie without correcting the balance of payments. Overstone's emphasis was on control of currency as the high-powered money base, with deposits as part of an inverted credit pyramid lacking any independent effect of their own, and dependent upon the currency base if banks behaved properly with respect to reserve ratios.

Thus, fundamental to monetary control was separation of departments in the Bank: the Banking department followed Banking principles, but the Issue department must follow Currency principles and thus stabilize economic activity, following automatic rather than discretionary procedures.

The role of the rate of interest in all this was twofold: short-run balance of payments correction, although this could only be a palliative if relative money incomes were out of line, and the production of an effect on confidence which in turn affected liquidity preference through increasing precautionary reserve holdings when the rate was raised, thus reducing the effectiveness of a given money supply. This variation in liquidity preference with confidence was an important part of the analysis, and was built into the 1844 Act with weekly publication of the Bank reserves, which were supposed to cause prudent adjustment of other reserves. This in turn would avoid the Bank of England's having to act as lender of last resort, a role which Overstone opposed as incompatible both with inducing the rest of the system to respond counter-cyclically and with the necessary limitation of the high-powered base.

Overstone was a many-sided man. But it is as a monetary theorist that he is chiefly remembered.

See Also

- Banking School, Currency School, Free Banking School

Selected works

1857. *Tracts and other Publications on Metallic and Paper Currency*. London: privately printed. London: Longmans, 1858.

1858. *The Evidence given by Lord Overstone, before the Select Committee of the House of Commons of 1857, on Bank Acts, with additions*. London: Longmans.

Bibliography

O'Brien, D.P., ed. 1971. *The Correspondence of Lord Overstone*, 3 vols. Cambridge: Cambridge University Press.

O'Brien, D.P. 1975. *The Classical Economists*, ch. 6. Oxford: Clarendon Press.

Wood, E. 1939. *English Theories of Central Banking Control, 1819–1858*. Cambridge, MA: Harvard University Press.

How to cite this article

O'Brien, D. P. "Overstone, Lord [Samuel Jones Loyd] (1796–1883)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000074> doi:10.1057/9780230226203.1237

The New Palgrave Dictionary of Economics Online

Owen, Robert (1771 – 1858)

N.W. Thompson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

autarky; class conflict; labour time; Owen, R.; pecuniary and non-pecuniary penalties; poverty alleviation; socialism

Article

Born in Newtown, Montgomeryshire (Powys), in 1771, Robert Owen was in many ways both the child and the victim of his age, making his fortune as a cotton manufacture involved in the industrial transformation of Britain and dissipating it in his efforts to eliminate its evils. With the purchase of the New Lanark cotton mills in 1797 Owen did, for a time, successfully combine the roles of factory owner and social reformer, showing how a humanized working environment might effect a reformation in human character. For the modern social scientist, one interesting innovation Owen implemented was the silent monitor, a four-sided block that was hung next to each worker's machine; a supervisor would turn the block to a colour that reflected the worker's effort during the day; colours were recorded in a ‘book of character’. (See Podmore, 1906, based on Owen's autobiography.) The silent monitor was meant to substitute for corporal punishment as a discipline device; it resonates with recent thinking on social sanctions; see pecuniary versus non-pecuniary penalties.

Owen's success in the New Lanark venture encouraged him to devote his life to the regeneration of mankind and it also provided him with the funds necessary to attempt this. However, further practical experiments proved disastrous. The cooperative communities he established, such as those at New Harmony Indiana in 1824 and Queenwood in Hampshire in 1839, soon collapsed, while his efforts in 1832 to socialize money through a National Equitable Labour Exchange proved equally disastrous. However, such failures never inspired self-doubt and Owen remained to the end of his long life a living embodiment of hope's capacity to triumph over experience.

As a cotton manufacturer Owen grasped the potential for material abundance which industrialization was creating in early 19th-century Britain; yet as an acute observer of economic life he was equally aware of the existence of widespread material impoverishment. His chief concern in his economic writings was, therefore, to investigate this paradox of poverty in the midst of abundance and show how it

might be resolved.

For Owen the realization of economic prosperity for all was obstructed by the tendency, in a competitive market economy, for rapid mechanization to create ‘a most unfavourable disproportion between the demand for and supply of labour’. This resulted in its progressive devaluation which in turn caused a diminution in consumption and a general economic crisis as manufacturers responded to a deficiency of effective demand by reducing output and laying off labour. As Owen phrased it, ‘It is want of a profitable market that alone checks the successful and otherwise beneficial industry of the labouring-classes’.

To remove this constraint upon production and to realize the potentialities of industrial development, Owen believed that ‘Human labour [should] acquire its natural and intrinsic value, which would increase as science advanced’, and to secure this Owen argued in such works as his *Report to the County of Lanark* (1821) that goods should be valued according to the labour time that they embodied and exchanged against labour notes rather than conventional money. Such a socialization of exchange, Owen believed, would give labour its whole product and further ensure that aggregate supply and aggregate demand expanded *pari passu*.

It was these ideas which bore practical fruit in the National Equitable Labour Exchange, where attempts were made to value goods and reward labour in terms of time. As might be expected this institution suffered a speedy demise. However, it was never seen by Owen as more than a stepping stone to his ideal of a ‘new moral world’ of neo-autarkic cooperative communities, where each would contribute to the common stock according to ability and consume according to need. Insulated thus against the exploitation and vagaries of a competitive market economy, material well-being could be assured and the character of man created anew.

Owen's economic writing was only one facet of a more general attempt to construct a science of society – a science which would have both an explanatory and prescriptive power and which could be used to determine the means necessary to transform man from an egotistical, competitive atom into a truly social being. It was this broader intellectual enterprise which enthused and interested British socialist thinkers in the first half of the 19th century, as can be seen, for example, in their redefinition of ‘political’ as ‘social’ or ‘moral’ economy.

Engels in *The Condition of the Working Class in England* (1844) remarked that, ‘English socialism arose with Owen, a manufacturer, and proceeds therefore with great consideration towards the bourgeoisie’, and, undoubtedly, Owen's tendency to stress the socially harmonious future and the ultimate reconcilability of class antagonism, rather than the social hostilities of the present, left its quietistic mark upon Owenite socialism. Yet for socialist writers such as Thompson, Gray and Bray, Owen's real legacy was methodological rather than ideological. What they imbibed from Owen was a particular, social scientific way of approaching the condition of labour rather than any unwillingness to unearth the roots of social antagonism.

Selected works

1813. *A New View of Society, Essays on the Formation of Human Character*. London.

1815. *Observations on the Effect of the Manufacturing System*. 2nd edn. London.

1817. Report to the Committee for the Relief of the Manufacturing Poor. In *The Life of Robert Owen written by Himself*, 2 vols, London, 1857–8.

1818. Two memorials on behalf of the working classes. In *The Life of Robert Owen Written by Himself*. 2 vols. London, 1857–8.

1819. *An Address to the Master Manufacturers of Great Britain*. Bolton.

1821. *Report to the County of Lanark of a Plan for Relieving Public Distress*. Glasgow: Glasgow University Press.

1823. *An Explanation of the Cause of Distress which Pervades the Civilized Parts of the World*. London.

1832. *An Address to All Classes in the State*. London.

1849. *The Revolution in the Mind and Practice of the Human Race*. London.

Bibliography

Beales, H.L. 1933. *The Early English Socialists*. London: Hamish Hamilton.

Beer, M. 1953. *A History of British Socialism*. 2 vols. London: Allen & Unwin.

Butt, J. 1971. Robert Owen in his own time 1771–1858. In *Robert Owen and his Relevance to our Times*. Cooperative College Papers No. 14. Loughborough: Co-operative Union.

Cole, G.D.H. 1930. *Robert Owen*. London: Macmillan.

Cole, G.D.H. 1977. *A History of Socialist Thought*. 5 vols. Vol. 1, *Socialist Thought: The Forerunners, 1789–1850*. London: Macmillan.

Cole, M. 1971. Owen's mind and methods. In *Robert Owen, Prophet of the Poor*, ed. S. Pollard and J. Salt. London: Macmillan.

Engels, F.W. 1844. *The Condition of the Working Class in England*. London: Panther, 1974.

Garnett, R.G. 1971. *Co-operative and Owenite Socialist Communities in Britain and America, 1825–45*. Manchester: Manchester University Press.

Gray, A. 1967. *The Socialist Tradition, Moses to Lenin*. London: Longman.

Harrison, J.F.C. 1969. *Robert Owen and the Owenites in Britain and America: The Quest for the New Moral World*. London: Routledge & Kegan Paul.

Oliver, W.H. 1958. The labour exchange phase of the co-operative movement. *Oxford Economic Papers* 10, 355–67.

Podmore, F. 1906. *Robert Owen: A Biography*. London: Hutchinson. Honolulu, HI: University Press of the Pacific, 2004.

Pollard, S. 1971. Robert Owen as an economist. In *Robert Owen and his Relevance to our Times*. Co-operative College Papers No. 14. Loughborough: Co-operative Union.

Royle, E. 1998. *Robert Owen and the Commencement of the Millennium: The Harmony Community at Queenwood Farm, Hampshire, 1839–1845*. Manchester: Manchester University Press.

Thompson, N.W. 1984. *The People's Science: The Popular Political Economy of Exploitation and Crisis, 1816–34*. Cambridge: Cambridge University Press.

Woodward, L. *The Age of Reform: 1915–1870*. Oxford: Clarendon Press (for broad historical context.)

How to cite this article

Thompson, N. W. "Owen, Robert (1771–1858)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_O000076> doi:10.1057/9780230226203.1238

The New Palgrave Dictionary of Economics Online

Palander, Tord Folkeson (1902 – 1972)

T. Puu

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bonds; choice under uncertainty; duopoly; factor substitution; land use; location theory; monopolistic competition; Palander, T. F.; rent; spatial monopoly; Stockholm School

Article

Palander became an engineer at the Royal College of Technology in Stockholm, before studying economics at Stockholm University. He published his dissertation *Beiträge zur Standortstheorie* (Contributions to Location Theory) in 1935. He was appointed Professor of Economics at the Business School of Gothenburg in 1941 and at Uppsala University in 1947. The dissertation is a standard reference in location theory; it was never published in English, though a Japanese edition was published in 1984.

Chapters III–X of the *Beiträge* contain a very detailed discussion of spatial economics from the classics to recent developments up to the date of Palander's own contribution. The Ricardian and von Thünen land rent theories, the Launhardt–Weber theories of location and market area formation are penetrated, and the spatial facets of Hotelling's and Chamberlin–Robinson's then fresh monopolistic competition theories are for the first time given due regard. Palander aims at an integration of classical location and land use theories with modern developments in general economics. Bringing in profit maximization under price-dependent demand and the possibility of spatial monopoly is one step in this direction. Another is the stress on the importance of factor substitution. Palander is extremely critical of Andreas Predöhl's attempts in this direction. His own planned contribution was withdrawn from the manuscript just before its publication, and the problem first got a satisfactory solution with the contribution by Leon Moses in 1958.

Palander's most original contributions are contained in Chapters XII–XIV. The central theme is to extend the classical models, where transportation is assumed to be along straight lines, to more realistic situations where transport rates are distance dependent, or traffic crosses different media, like land and sea. In the last context, the neat refraction law of traffic was discovered. These contributions set the stage for Martin Beckmann's (1952) continuous model of transportation.

One of the most attractive features of Palander's work is the artwork, which must be considered unsurpassed until the advent of computer graphics. Palander brings graphical analysis combined with simple algebra and analysis to perfection.

Palander remained in the USA as a Rockefeller Fellow during 1936, studying Chamberlin's monopolistic competition theory. His own contributions are a brief abstract in English of a presentation at the Cowles Commission conference at Colorado College 1936 (‘Instability in Competition between Two Sellers’) and an article in Swedish, ‘Konkurrens och marknadsjämvikt vid duopol och oligopol’ (‘Competition and Market Equilibrium in Duopoly and Oligopoly’), published in *Ekonomisk Tidskrift* (later *Scandinavian Journal of Economics*) in 1939. Palander's particular interest was the stability of adjustment processes in classical Cournot and similar types of duopoly.

Palander belonged to the informal group of economists called the ‘Stockholm School’ and wrote an extensive critical review of their methods, ‘Stockholmsskolans begrepp och metoder’ (1941). Palander's remarks concern in particular the lack of rigorous dynamic analysis.

Palander took a great interest in Keynesian macroeconomics. He edited a translation of the *General Theory* into Swedish in 1945 (*sysselsättningsproblemet*), with commentary, and wrote an extensive mathematical and graphical analysis of the work in 1942. This article, in comparison to the similar works by Hicks and Klein, contains a thorough analysis of all the different variants with the relations expressed in monetary, real and wage units.

Palander's later work was mainly pedagogical, and his last research interest concerned monetary theory in connection with choice under uncertainty. He wrote a monograph (in Swedish) on the effects of index bonds upon an inflationary economy in 1957. Among his consultancies the most important was to the Swedish Railway Board on fare tariff policy.

See Also

- location theory

Selected works

1935. *Beiträge zur Standortstheorie*. Uppsala: Almqvist & Wiksell.

1936. Instability in competition between two sellers. In Cowles Commission, Research Conference on Economics and Statistics, Colorado Springs: Colorado College. Chicago: University of Chicago Press.

1939. Koncurrens och Marknadsjämvikt vid duopol och oligopol. *Ekonomisk Tidskrift* 41, Pt. I, 41(2), 123–45; Pt. II, 41(3), 222–50.

1941. Stockholmsskolans begrepp och metoder. *Ekonomisk Tidskrift* 43(1), 88–143. Trans. as: On the concepts and methods of the ‘Stockholm School’, *International Economics Papers* No. 3. London and New York: Macmillan, 1953.

1942. Keynes allmänna teori och dess tillämpning inom ränte-, multiplikator- och pristeorien.

Ekonomisk Tidskrift 44(4), 233–72.

Bibliography

Beckmann, M. 1952. A continuous model of transportation. *Econometrica* 20, 643–60.

Moses, L. 1958. Location and the theory of production. *Quarterly Journal of Economics* 72, 259–72.

How to cite this article

Puu, T. "Palander, Tord Folkeson (1902–1972)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000002> doi:10.1057/9780230226203.1239

The New Palgrave Dictionary of Economics Online

Palgrave, Robert Harry Inglis (1827 – 1919)

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Palgrave, R. H. I.; central banking; Palgrave's Dictionary of Political Economy

Article

Palgrave was born in London, the third of four male children of Francis Palgrave and Elizabeth Turner. He was named after Robert Harry Inglis – an Old Tory, Member of Parliament, and a friend of Palgrave's father. Quite incidentally, this R.H. Inglis edited some works by the economist Henry Thornton. Palgrave was denied the formal education provided for his two elder brothers, instead entering the banking business of Gurney & Co. (in which his maternal grandfather had been a partner) in Great Yarmouth at the age of 16. Palgrave himself subsequently became a partner in the bank, and married in 1859 a daughter of Mr George Brightwen, who was related to the Gurney family.

Family

Palgrave's father was born Francis Cohen in 1788, the son of Meyer Cohen, member of the London Stock Exchange during most of the years that the Ricardo's were members. Francis Cohen altered his name to Palgrave in 1823 upon marriage to Elizabeth Turner – Palgrave being Elizabeth Turner's mother's maiden name. She, in turn, was the daughter of Dawson Turner, a partner in the English country bank Gurney & Co. in Great Yarmouth. Sometime during the second decade of the 19th century, Francis Cohen renounced the Jewish faith and embraced the Christian religion in the form of the teachings of the Church of England. He was a medievalist of some repute, publishing *The Rise and Progress of the English Commonwealth* in 1832, and *The History of the Anglo-Saxons* in 1837. He was Deputy Keeper of H.M. Public Records, was knighted in 1832, and his literary friends included Macaulay and Henry Hallan (who Palgrave was to quote to moving effect in his editorial preface to the final volume of his *Dictionary*).
His first son, Francis Turner Palgrave (1824–1897), is still widely known today for his famous *Golden Treasury of English Lyrics and Verse*, the first edition of which appeared in 1861. He was educated at

Charterhouse and Balliol College, Oxford, going up in 1843, and in 1846 acted as assistant private secretary to Gladstone. Between 1850 and 1855 he directed a government teacher training college near Twickenham. Thereafter he was engaged in the Department of Education in London until his retirement in 1884. In 1885 he was elected into the Professorship of Poetry at Oxford (with the support of Alfred Tennyson, whom he had met in his days near Twickenham).

The second son has William Gifford Palgrave (1826–1888), the least ‘typical’ of the family. After Charterhouse and Trinity College, Oxford, he moved to India, becoming a lieutenant in the 8th Bombay Regiment. He soon converted to Roman Catholicism, entered a Jesuit mission in Madras, and was ordained a priest of the Order. He remained as a missionary in India until 1853 when he was recalled to Rome. Later in that same year he went as a missionary to Syria. The *Dictionary of National Biography* reports that ‘he could and did pass without difficulty for a native of the East’, adding that ‘the often repeated story that he had officiated as Imaum in mosques is without foundation’. When hostilities between the Druse and the Maronite Christians broke out, the Maronites invited him to become their leader – an invitation which it seems he declined. A massacre of Christians in Damascus in June 1881 precipitated his return to Europe. There he reported to Napoleon III on the Syrian situation. This contact led him into an expedition in 1862–3 across the Middle East. This was financed by the French government, to whom he was to ‘report on the state of the Arab attitude’ towards France. Subsequent to this venture he returned to England.

At ‘home’ again, he published his *Narrative of a Year's Journey Through Central and Eastern Asia* in 1865. This was the most widely read narrative on that region until the accounts of T.E. Lawrence appeared on the scene. He broke with the Jesuits, and became a diplomat in the service of the British government. He was dispatched to Abyssinia, and then went as consul to St. Thomas in the West Indies in 1873. There followed postings in Manila (1876), Bangkok (1879), and Uruguay (1884). He died in Montevideo on 30 September 1888.

The youngest son, Reginald Francis Douce Palgrave (1829–1904), was Clerk of the House of Commons. He was the only sibling to survive to see the publication of the *Dictionary* and Palgrave consistently requested Macmillan to forward to him complimentary copies of the work as it appeared.

Works

It is said that ‘as quite a young boy’ Palgrave received from his father a copy of the *Wealth of Nations*, which he treasured throughout his life. That book seems to exert a power so mysterious that few who take it up and study it seriously have been able to avoid the fate of a career in economics. However, his activities at Gurney & Co. in Great Yarmouth, while immersing him in the daily business of economics, delayed the entry of his name into its literature until 1870, when he received the Statistical Society's Taylor Prize for an essay on local taxation in Britain and Ireland.

The work by which Palgrave's name will be perpetuated is, of course, his *Dictionary of Political Economy*, one of the finest achievements of Victorian scholarship. Shortly after the publication of its last appendix he was knighted (1909).

Here, we shall consider only his other writings, all of which dealt with some aspect of banking practice or theory. His publications of 1873, 1874 (a and b) and 1877 typify a kind of statistical analysis of central banking, and their results are largely collated and summed up in *Bank Rate and the Money Market* of 1903. Of this book, Schumpeter commented:

> [It] is a masterpiece of the art of making figures speak … it is very difficult to formulate particular results but he who peruses this book page by page suddenly discovers that he understands its subject. (1954, p. 1080)

On matters of policy, he opposed bimetallism, opposed the monetary policy of the government in India, pushed for stability in Bank Rate, and was a supporter of the kind of regulations embodied in Peel's Act of 1844. However, in a review of Bagehot's *Lombard Street* (1874b) he formulated clearly the idea that the central bank was effectively an arm of government and thus a vehicle through which governments could effectuate monetary policy. The idea of the Bank of England as an autonomous agency governed only by the legislative provisions of its act of establishment was thereby altered. The new conception which was to take root was of a central bank more familiar nowadays than it was at that time.

In 1877, Palgrave became financial editor at the *Economist*, and on Bagehot's death took over its editorship. He remained there until 1883. He also edited the *Banking Almanac* until his death, and was briefly editor of the *Banker's Magazine* to which he contributed regularly after 1880.

Palgrave was also closely involved in the public affairs of the nation. In 1875 he gave evidence before the House of Commons Select Committee on Banks of Issue (George Goschen being the Committee's economic expert) on behalf of the Country Bankers' Association, and in 1885 he was a member of the Royal Commission on Depression of Trade and Industry. In the memorandum of evidence submitted to that Commission, Alfred Marshall remarked that he would not cover matters already dealt with in Mr Palgrave's memorandum since he was in broad agreement with that document.

It is said that as a boy Palgrave dreamed of becoming a Fellow of the Royal Society – a dream which became reality in 1882, thanks in part to the support he received from Jevons. The latter's correspondence with Palgrave from that period (held in the archives of King's College, Cambridge) speaks both to the modesty of Palgrave and the genuine friendship of which Jevons was capable. There is a postcard from Jevons, dated on the day the names of elected Fellows were published in the *Times*, containing no other communication than the name of its adressee: 'R.H. Inglis Palgrave, F.R.S.'

See Also

- Palgrave's Dictionary of Political Economy

Selected works

1870. Local taxation in Great Britain and Ireland. Taylor Prize Essay, Statistical Society of London.

1873. *Notes on Banking in Great Britain and Ireland, Sweden, Denmark and Hambourg*. London.

1874a. *Analysis of the Transactions of the Bank of England for the Years 1844–1872*. London.

1874b. Banking. *Fortnightly Review*, 1 January, 92–108.

1877. The influence of a note circulation in the conduct of the banking business. *Journal of the Manchester Statistical Society*, March.

1894, 1896, 1899., ed. *Dictionary of Political Economy*, 3 vols. London: Macmillan.

1903. *Bank Rate and the Money Market in England, France, Germany, Belgium and Holland: 1844–1900*. London: J. Murray.

Bibliography

Kiddy, A.W. 1919. Obituary – Sir Inglis Palgrave. *Economic Journal* 29, 112–17.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

How to cite this article

Milgate, Murray. "Palgrave, Robert Harry Inglis (1827–1919)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000005> doi:10.1057/9780230226203.1240

The New Palgrave Dictionary of Economics Online

Palgrave's Dictionary of Political Economy

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Edgeworth, F. Y.; Higgs, H.; Marshall, A.; Palgrave, R. H. I.; *Palgrave's Dictionary of Political Economy*

Article

Inglis Palgrave's *Dictionary of Political Economy* appeared in volume form sequentially in 1894, 1896 and 1899. However, 1894 was not the year in which the *Dictionary* began publication. Under an earlier publishing plan, subsequently abandoned, a first part of the *Dictionary* (covering the entries Abatement to Bede) appeared in 1891, followed by two more in the next year (extending the project well into the letter C). Furthermore, 1899 does not accurately represent the completion date of the work. It was not until 1908, when the appendix to the third volume was published, that its publication could be said to have been complete. It took 17 years to effect the publication of the *Dictionary* – better than 20 years of work if one takes into account the fact that the contractual agreement between Palgrave and Macmillan is dated 1888.

Though the original contract called for a work in two volumes, it seems that this plan was subsequently revised to entail publication in parts, each of 120–130 pages in length, and to appear at quarterly intervals. It was envisaged that the entire work would run to between 12 and 14 parts.

The rationale behind the adoption of this plan seems to have derived from a number of considerations. In the first place, the French *Dictionnaire d'économie politique* and the German *Handswortsbuch der Staatswissenschaften* had already been appearing in parts, and Palgrave specifically cited these instances in support to the plan. Closer to home, the *Dictionary of National Biography* was also appearing in parts at the time, and successfully at that. Commercial considerations exerted due influence. Palgrave argued that 'each *part* of the Dictionary, as it comes out, may be expected to be noticed …, each *volume* would only receive a similar notice', so that '*parts* will be more frequently [brought] before public notice'. He was also concerned that any delay in commencing publication might allow competitors to beat him to the market.

The fact of sequential publication, whether in parts or in volumes, went hand in hand with sequential

writing and sequential planning. In 1892, after this process had actually begun, Palgrave argued that it brought two substantial externalities – he might receive ‘a good many valuable suggestions and hints’, and he could easily refer contributors to the later parts back to earlier published parts in order to avoid overlap. But there were diseconomies as well. With the prospect of publication extending over four or five years, even if it was kept to a strict schedule, there was a real danger of contributor exhaustion. Perhaps more importantly, there was no way of making adjustments for recent advances, or for correcting oversights, or for taking advantage of the valuable advice an editor receives, if the relevant material had already gone to press.

The original *Dictionary* does seem to have suffered some of these disadvantages. Soon after 1899 it was in need of revisions substantial enough to require the printing of separate appendices (published separately at first, and bound in with subsequent reprints). Some of the reasons for this will be touched upon later. What is more, just how exhausted its contributors became during the process can be witnessed in the record of one of the most loyal among their number – F.Y. Edgeworth. In the first volume there were 77 entries from his pen. In the second, the tally had fallen to 38. In the third volume it was down to 17 (in the Higgs edition there are 10 more, mostly addenda to existing entries). Not only did Edgeworth's entries shrink in total number, they shrank in average length as well.

While little concrete detail is readily to hand concerning the editorial practices that Palgrave adopted, there is sufficient evidence available from which to make some fairly confident conjectures. To begin with, it is clear that the list of entries was planned well ahead, despite the more immediate horizons imposed by the choice of sequential publication. In November 1889, for example, before anything had been published, Palgrave reported to Macmillan that he had planned the list of entries down to the letter K. In a letter of 16 March 1892, he reported that he had ‘forwarded a considerable number of articles … in the S’’ to the printer. Yet it is equally clear that there was no attempt to generate a list of entries that was in any significant sense ‘complete’ prior to the commissioning of contributors.

Just how Palgrave arrived at the actual entries to be included is not so easily established. It seems to have been a combination of his own ideas, and those of specialist contributors in particular fields. The list of entries classified by contributor to the original *Dictionary* (compiled by K. Newman and appearing as an appendix of the present work) reveals a pattern whereby certain contributors wrote nearly every entry in a given field. It seems likely that Palgrave simply gave them a free hand to generate the key entries in that field. This probably explains some of the singularities of the pattern of contributions in the original work. How else, for example, might one explain the fact that Mr F.E. Allum of the Royal Mint at Perth, Western Australia, contributed over 100 entries on various media of exchange – from the English Angel to the Japanese Yen. Or that A. Courtois *fils* contributed a similar number of biographical entries on (mainly) French writers – from the marquis d’Audiffret to Louis François Michel Raymond Wolowski. How free were the contributors’ hands in determining the length of entries is indicated, perhaps, by the fact that M. Courtois *fils* produced two-and-a-half columns on the obscure Wolowski, and just two on Quesnay.

If the practice of deferring to ‘specialists’ seems to have introduced certain idiosyncrasies, in other cases it bore fruit, that of Edgeworth being exemplary. It is hardly necessary to add that even with his specialists, Palgrave experienced the usual problems of tardiness in the delivery of sacredly promised entries, and of restraining contributors to limits, even if flexible, as to length. On the whole, however, he seems to have handled these with admirable tolerance and forbearance – though at one point he did suggest that Macmillan might consider sending a man round to Robert Giffen's residence to await on the

delivery of his promised essay on Bagehot (which, as it transpired, he did not obtain) – and with not a little creativity in the re-titling of entries so that they would appear further down the alphabet.

The reaction to the *Dictionary* can be considered from two rather different perspectives, that of the economics profession itself and that of the market. As to the latter, there were two phases, the first covering the three parts which appeared before the first ‘recognized’ volume in 1894, the second covering the subsequent period.

The first phase was a clear commercial failure. The first part had gone to the printers in June of 1891 and was published that year. Within a few months Palgrave was already alert to its lack of success in the market. In a letter to Macmillan dated 14 March 1892 he expresses himself ‘extremely disappointed to find that the sale of the Dictionary has been so small’. At about the same time, the publishers began to suggest abandoning the existing publishing plan, in favour of a format more like that which actually appeared. Initially, Palgrave held out against these suggestions. But the similarly disappointing sales of the second and third parts, which appeared in 1892, seems to have reconciled Palgrave to a change of plan. In November of that year little sign remains of the vigorous defence he had made of the earlier plan just a year before. Instead he writes: ‘I do not wish *myself* to suggest a change from parts of volumes, but … as you appear to have this in your mind, I have now planned out the work as far as the next volume would extend, should *you* desire it be dealt with *as a whole*’.

The professional reaction to the *Dictionary* was generally favourable, as might have been expected given the fact that almost all economists of any repute had already endorsed the enterprise by agreeing to contribute. Of course, any encyclopedia is vulnerable to criticism. Why one particular title for an entry, rather than some other? Why include some unimportant subject or author, and neglect other more worthy ones? There was also some critical comment on the work's sequential appearance. Ever a supporter, Edgeworth effectively put paid to this avenue of attack: ‘not even Homer brings forward all his Greeks at once, but makes one the hero of the third, another of the fifth book’ (1892, p. 525).

Probably more to the point, some reviewers were wary of the presence in the *Dictionary* of so much material on legal matters, current commercial practices and international treaty arrangements (a hangover from McCulloch's *Commercial Dictionary* perhaps?) and statistical information. This sentiment was shared even by Henry Higgs, who in the editorial preface to his edition of the *Dictionary* remarked that most of this is ‘only remotely connected to economics’. The presence of these subjects probably reflected in substantial measure the tastes of Palgrave himself, a commercial banker.

Two specific and less favourable reactions to the *Dictionary* must be singled out, the first contained in an essay by E.R.A. Seligman (who would later edit the *Encyclopedia of the Social Sciences*) that appeared in two parts in the *Economic Journal* for 1903 under the title ‘On Some Neglected British Economists’. The second reaction was that of Alfred Marshall.

Seligman's article (in Seligman, 1925, pp. 65 ff.) seems to have been an impetus to some of the revisions which Palgrave began shortly after the third volume had appeared. Seligman used the *Dictionary* to exemplify his claim that insufficient attention had been given to a number of British economists: Torrens, W.F. Lloyd, Bailey, Longfield, Read, Craig, Butt and George Ramsay. The cases of Torrens, Lloyd and Longfield were compelling, and Palgrave sought remedy in the appendices. The cases of Craig, Butt and Ramsay were much less so, and Bailey had in fact been the recipient of a generous notice by Edgeworth in the original (where Read had been noticed by James Bonar).

Marshall's reaction, though contained only in asides in correspondence, was unambiguously negative – at one point he makes a play on Palgrave's initials (RIP) in regard to his expectations for the fate of the

enterprise. This might help explain why Marshall's name is the most glaring absence from the list of contributors to the *Dictionary*. It was certainly not because he was not asked – it is clear from Palgrave's correspondence with Macmillan that Marshall was approached. In the end the *Dictionary* had to wait for a contribution which bore the signature of Marshall until the Higgs edition of Volume I in 1925. Even then, it was merely a note on the teaching of economics at Cambridge, not written originally for the *Dictionary* in any case.

The judgement of the profession on his *Dictionary* is probably best summed up in a letter published in the *Economic Journal* for September 1917, congratulating Palgrave on reaching his 90th birthday. For there even Marshall's name appears among the distinguished list of signatories.

Little need be said here of the Higgs edition of the *Dictionary*, which for the first time formally added Palgrave's name to the title (though Edgeworth had done so informally in his review of its second and third parts in 1892), but which made few changes to its structure or contents. Like its predecessor the Higgs edition also appeared sequentially, though not in alphabetical order. Volume II appeared first, in 1923, Volume I followed in 1925, and Volume III in 1926; just why, is not known.

Palgrave was already in his sixties when he began the *Dictionary*, and was in his eighties when it was done – an act of dedication to the discipline unlikely to be replicated. As Edgeworth presciently remarked when reviewing its second and third parts in the *Economic Journal* for 1892, it ‘will remain a monument of what may be accomplished by individual initiative and energy’.

See Also

- Palgrave, Robert Harry Inglis

Bibliography

All quotations from Palgrave's correspondence over the *Dictionary* are taken from material held in the Macmillan archive in the library of the University of Reading.

Edgeworth, F.Y. 1892. Review of *Dictionary of Political Economy*. Edited by R.H. Inglis Palgrave, F.R.S. *Economic Journal* 2, 524–5.

Kiddy, A.W. 1919. Obituary: Sir Inglis Palgrave. *Economic Journal* 29, 112–17.

Price, L.L. 1891. Review of *Dictionary of Political Economy*. Edited by R.H. Inglis Palgrave F.R.S. *Economic Journal* 1, 605–8.

Seligman, E.R.A. 1925. *Essays in Economics*. New York: Macmillan.

How to cite this article

Milgate, Murray. "Palgrave's Dictionary of Political Economy." The New Palgrave Dictionary of

Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000006> doi:10.1057/9780230226203.1241

The New Palgrave Dictionary of Economics Online

Palmer, John Horsley (1779 – 1858)

Anna J. Schwartz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank Charter Act (1844); Bank of England; bank rate; central banking; Palmer rule; Usury Laws

Article

Governor of the Bank of England from 1830 to 1833, Palmer was a significant participant in 19th-century controversies concerning the Bank's proper management. In 1802 he entered into a partnership with two others as East India merchants and shipowners, and remained active in business until weeks before his death in 1858. ‘[A] vigorous, outspoken man’ (Clapham, 1944, vol. 2, p. 114), he was first elected a director of the Bank in 1811 and was regularly re-elected thereafter except for the usual hiatus every third year before 1828 and again in 1845–6. By 1857, when his service terminated, he was the senior director of the Court.

Palmer's view over the period from 1832 to 1857 may be gleaned from his answers to questions addressed to him by Parliamentary committees, three pamphlets he published in 1836 and 1837, and correspondence. Among the central issues he discussed were the nature of the Bank's responsibilities, its relation to the London money market, the joint stock and country banks, its role in stabilizing domestic economic conditions, and how it operated on the exchanges.

Palmer's initial statement of the principles that guided the Bank's policy became known as the Palmer rule or the rule of 1832: the Bank's duty in ordinary times, when the reserve was at a maximum and exchange rates were at par, was to maintain its bullion reserve at one-third of its liabilities, the sum of deposits and note issues. At such times the Bank should hold interest-earning assets of government stock and other long-dated securities equal to two-thirds of its liabilities. Thereafter the portfolio should be maintained unchanged so that as gold was withdrawn from or brought to the Bank, the public would reduce or increase notes and deposits. Changes in the Bank's liabilities would thus arise at the public’s, not the Bank's initiative. A loss of bullion would be matched by a reduction in notes and deposits with no change required in the portfolio to reduce the supply of funds in the money market. Palmer held that the Bank should set its rate above the market rate so that in normal times it would not be competing for discounts with London bankers. ‘At times of discredit’, however, when the market rate rose to the level

of Bank rate, the Bank should discount bills of exchange, selling off government securities as discounts increased. (Palmer did not recognize that selling securities would offset the provision of funds to the discount market.)

The validity of the Palmer rule was challenged by both contemporary and later critics (Loyd, 1837; Viner, 1937), although some modern students (Horsefield, 1949; Matthews, 1954) regard it as essentially sound. A different line of criticism, occasioned by financial market stringency in 1836–7 and again in 1839, was that in practice the Bank did not observe the rule. Palmer defended the Bank, arguing that in face of deposit declines at other banks associated not with gold drains but with transfers to the Bank – he was referring to East India and other special deposits, 1833–7, and to seasonal Exchequer deposits – it was proper for the Bank to increase its portfolio to offset the extra funds it held. Similarly, the increase in the Bank's portfolio in 1836–7 was the correct response to an internal drain of gold, as it was also in 1839 to an external drain, none of these ordinary years in which the rule applied.

Palmer's arguments failed to convince his interlocutors. He himself retreated from some initial positions. In 1848 he qualified the rule governing reserves in relation to total liabilities. External drains affecting exchange rates, he noted (British Sessional Papers 1847–8, VIII, Pt. 1, 167–8), might be related to political factors abroad rather than to domestic circulation and deposits. As reserves of London bankers gradually came to be held at the Bank rather than as Bank notes, he shifted from a view of Bank rate as the means for influencing note issues to the view that it was the means for influencing the money market. Initially insistent that the Bank must have a monopoly of the note issue – he claimed in 1837 that the issues of many newly formed joint stock banks in 1835–6 had nullified the Bank's contraction of its issues (clearly not true) – in 1848 he did not object to unrestricted country bank note issues provided they were adequately secured.

On other matters, Palmer's views held firm. He believed that changes in Bank rate in relation to rates abroad could control international trade and capital movements. He was a critic of the Bank in the 1840s for too often changing Bank rate and failing to maintain it above market rates. Despite acknowledging the Bank's influence on economic affairs, he denied that it was answerable to anyone but its proprietors, or that publication of a statistical account of its actions was desirable. He opposed separation of the Bank into Issue and Banking Departments before and after the Act of 1844 became law.

Horsley Palmer's name survives as a spokesman for the proper conduct of monetary policy in a period of imperfect understanding of the Bank of England's responsibilities. By asserting that the Bank ought not to compete with other banks in discounting commercial bills in ordinary times, he centred attention on the position of the Bank as distinct from that of other institutions in the money market. He recognized the primacy of its central banking function as lender of last resort during financial stringency. His advocacy of setting Bank rate above market rates hastened the demise of the Usury Laws. For modern observers of the instability that discretionary central bank policies at times have produced, his rule of a constant portfolio has resonance.

See Also

- Banking School, Currency School, Free Banking School

Selected works

1836. *Reasons Against the Proposed Indian Joint Stock Bank, in a Letter to G.G. de H. Larpent Esq*. London: Pelham Richardson.

1837a. *The Causes and Consequences of the Pressure upon the Money-Market*. London: Pelham Richardson.

1837b. *Reply to the Reflections of Mr. Samuel Jones Loyd Suggested by a Perusal of Mr. J. Horsley Palmer's Pamphlet entitled 'Causes and Consequences of the Pressure Upon the Money-Market'*. London: Pelham Richardson.

Bibliography

British Sessional Papers, session year, Parliamentary Paper number, volume number, pages of Palmer's testimony: 1831–2 (722), VI, 7–70. 1840 (602), IV, 103–41. 1847–8 (395), VIII, Pt. 1, 147–68; Pt. 2, 521–3; 1847–8 (565), VIII, Pt. 3, 79–124. 1857, 2d sess. (220), X, Pts. 1, 2, App., 6 (Palmer letter, dated 15 October 1856); Index, 452.

Clapham, J.H. 1944. *The Bank of England: A History, 1694–1914*. Cambridge: Cambridge University Press.

Horsefield, J.K. 1949. The opinions of Horsley Palmer. *Economica* 16(62), 143–58.

Loyd, S.J. 1837. *Reflections Suggested by a Perusal of Mr. J. Horsley Palmer's Pamphlet*. London: Pelham Richardson.

Matthews, R.C.O. 1954. *A Study in Trade-Cycle History: Economic Fluctuations in Great Britain, 1833–1842*. Cambridge: Cambridge University Press.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper.

How to cite this article

Schwartz, Anna J. "Palmer, John Horsley (1779–1858)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000007> doi:10.1057/9780230226203.1242

The New Palgrave Dictionary of Economics Online

panic of 1907

Jon R. Moen and Ellis Tallman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank Panic of 1907; central banking; Federal Reserve Act; National Banking Era

Article

The Bank Panic of 1907 was the final banking crisis of the National Banking Era (1863–1913); it was significant in that it led to the Federal Reserve Act. The panic began when the spectacular attempt by F. Augustus Heinze to corner the stock of United Copper Company collapsed on 16 October 1907. The collapse revealed the extensive links of Heinze to another notorious financier in the New York City banking community, Charles F. Morse, a man O. M. W. Sprague (1910, p. 248) describes as having ‘an extreme character, even by American speculative standards’. Solvency concerns led to a series of bank runs at several national banks controlled by the two men. Yet the turmoil surrounding the Heinze collapse did not produce a systemic panic in New York, because the New York Clearinghouse took prompt corrective actions on the member institutions.

But on Monday 21 October the National Bank of Commerce announced late in the afternoon that it would no longer clear checks through the New York Clearinghouse for the Knickerbocker Trust Company. The following day, Knickerbocker Trust faced a run on deposits that lasted three hours, and it suspended business just before noon after having paid out $8 million in cash. The next day, the *New York Times* reported that the Trust Company of America was the current ‘sore point’ in the panic, a report that magnified the run on the Trust Company of America. Over the next two weeks the Trust Company of America paid out $47 million to depositors.

J. P. Morgan, James Stillman of National City Bank, and George Baker of First National Bank arranged through the New York Clearinghouse to aid the Trust Company of America, other stricken trusts, and the stock market after it had been determined that Knickerbocker could not be saved. On 26 October the New York Clearinghouse authorized the issuance of clearinghouse loan certificates to make available otherwise illiquid reserves and assets as a substitute for currency in transactions between banks, currency that could then be paid out to depositors. In addition, the Clearinghouse authorized restrictions on the payments of cash from deposit accounts, thereby limiting the outflow of the means for payment

finality from the intermediation system.

While private sector efforts succeeded in quelling the panic, it still altered how key New York bankers perceived their ability to manage the New York money market. New risks set 1907 apart from the earlier national banking era panics of 1893, 1890, 1884, and 1873. These new risks arose from the increased participation of trust companies and other intermediaries in the call money loan market, the overnight loan market for the New York Stock Exchange. By 1907 the New York trust companies had nearly 90 per cent of the loan volume of the New York national banks (Moen and Tallman, 1992). As intermediaries outside the Clearinghouse such as foreign banks, national banks from the interior of the United States, and the New York City trust companies became more prominent in the call loan market, the New York bankers realized that their ability to manage that market during panics had diminished considerably.

Under normal financial conditions the call loan market had been reasonably stable, serving as an outlet for excess reserves accumulating in New York through the correspondent banking system. During bank panics, however, the call loan market magnified the scramble for cash as numerous banks tried to call in their loans simultaneously. Typically, interest rates would spike upward and the value of the stock collateral would start to fall as borrowers as a group sold off their collateral to pay call loans. Until 1907, however, such a catastrophe had been averted because the New York national banks, acting through the Clearinghouse, jointly responded to mitigate disruptive contractions of call loans. The collective action of New York national bankers through the issuance of clearinghouse loan certificates and the partial suspension of convertibility of deposits into cash succeeded when clearinghouse banks provided the largest source of funds for the call loan market. The New York banks had been the main suppliers of call loan money since the inception of the stock market, which motivated them to preserve both the stock and call loan markets.

Outside intermediaries like the trusts, having no collective concern for the call loan market, began calling in large numbers of loans on 24 October 1907 (Cleveland and Huertas, 1985). Stock equity values began falling precipitously. Depressed stock values threatened the financial condition of both borrowers and lenders of call loans, including the national banks. The positive correlation between changes in collateral values and the changes in the creditworthiness of borrowers and lenders (sometimes referred to as covariance risk) transmitted the financial shock faced by trusts throughout the financial system to the national banks. Early in the panic there were reports of New York Clearinghouse member banks taking over a large volume of call loans made by trust companies to prevent the collapse of the call loan market and, hence, the stock market. By the end of the panic, loans (and deposits) at the New York national banks increased by nearly ten per cent, in contrast to the 37 per cent contraction of loans at trusts. No similar pattern was seen in the earlier panics (Moen and Tallman, 1992).

Contemporary observers applauded the leadership of New York Clearinghouse banks. Sprague (1910) noted how the national banks forbore on calling in loans that were technically insolvent, expecting them to recover as conditions returned to normal. Woodlock (1908) noted the increasing and destabilizing role of outside lenders in the call loan market, lenders like the trust companies that were outside the influence of the Clearinghouse. New York Clearinghouse bankers lost confidence that they alone could reliably alleviate stress in the call loan market during a panic. The movement to establish a centralized system of reserves or a central bank gained momentum after the Panic of 1907 because the New York Clearinghouse banks were no longer willing to bear all the risks alone.

See Also

- banking crises
- economic history
- Federal Reserve System

Bibliography

Cleveland, H. van B. and Huertas, T. 1985. *Citi bank 1812–1970*. Cambridge, MA: Harvard University Press.

Moen, J. and Tallman, E. 1992. The Bank Panic of 1907: the role of the trust companies. *Journal of Economic History* 52, 611–30.

Sprague, O. 1910. *History of Crises Under the National Banking Era*. National Monetary Commission. Washington, DC: Government Printing Office.

Woodlock, T. 1908. The stock exchange and the money market. Reprinted from *The Currency Problem and the Present Financial Situation*. New York: Columbia University Press.

How to cite this article

Moen, Jon R. and Ellis Tallman. "panic of 1907." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000348> doi:10.1057/9780230226203.1243

The New Palgrave Dictionary of Economics Online

Pantaleoni, Maffeo (1857 – 1924)

G. Becattini

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

marginalism; Pantaleoni, M.; Pareto, V.; public finance; regional economics; value theory

Article

Italian economist and politician, Pantaleoni was born in Frascati (Papal States) on 2 July 1857 and died in Milan on 2 October 1924. His career as a university teacher in economics was rather stormy on account of his impatient rejection of any attempt to interfere with his teaching and the free expression of his thought. Elected in 1900 as a radical to the Italian Parliament, he resigned shortly afterwards.

In 1920 he was appointed to manage the finances of d'Annunzio's Free State of Fiume, and in 1923 was nominated a member of the Italian Senate by the Fascist government, of which he was a supporter. His contribution to scholarship may be divided into three parts. First, a famous textbook, *Principii di economia pura* (1889), which contributed to the introduction of marginalist ideas into Italian economic thought and which, in its English translation (1898), made a considerable impression outside Italy as well. Second, a monograph on applied economics on the fall of the 'Credito Mobigliare', which Piero Sraffa aptly compared to Bagehot's *Lombard Street*. And third, a long series of papers, some of which may be regarded as seminal, on a wide variety of topics in both economics and at the interface with other social sciences. His lectures at the University of Rome, transcribed and published by his students, are also worthy of mention.

Much of Pantaleoni's writing has been brought together in anthologies and Pantaleoni's thought has been the subject of much comment; however, no persuasive, thorough, study has yet appeared. The man and the scholar emerge most vividly from his correspondence with Vilfredo Pareto.

The most distinctive feature of Pantaleoni's theoretical work is his tendency to generalize across disciplinary boundaries. Economics, sociology, anthropology and psychology form a kind of unified field within which Pantaleoni, while still employing the style of reasoning of the economist, moves freely and creatively, without shrinking from paradox and logical extremes. His great friend Vilfredo Pareto reproached him in 1898: 'the advancement of scholarship lies in creating new distinctions and not, as you seek to do, in reducing their number.' In apparent contradiction to this extreme tendency

towards generalization is Pantaleoni's capacity for minute and piercing analyses and broad and brilliant syntheses of given concrete situations.

How far Pantaleoni may be classified as a genuine marginalist is still an open question. Musgrave and Peacock (1967) wrote that ‘one of the first attempts at dealing with the determination of the tax-expenditure plan as a problem of economic value appears in Pantaleoni's essay of 1883’. They refer to the early Pantaleoni paper, ‘Contributo alla teoria del riparto delle spese pubbliche’, later republished in the *Scritti vari* anthology.

It is true that his extreme subjectivism brings him close to the ‘classical’ marginalists (though he was very critical of Menger at times), but his eclecticism – in a half-Marshallian vein – about the theory of value, and his acceptance of many of the concepts typical of evolutionist sociology (for example, his distinction between predatory, parasitical and mutualistic settlements), incline one to define him as unclassifiable except in the historical context. His relationship to the thought of Edgeworth and Marshall comes out clearly from the following letter to Edgeworth: ‘you are the closest approximation of a match for Marshall living in England. You know that to my mind, Marshall is simply a new Ricardo who has appeared in the field.’

If we look at Pantaleoni's mature work, we can conclude, with G. di Nardi, that ‘the Pantaleonian essays following *i Principii*, place him outside orthodox marginalism and made of him a very acute forerunner of contemporary critical schools’.

Pantaleoni's many disciples have helped to consolidate the profound imprint left by him (much deeper than that of Pareto, though the latter is better known nowadays outside Italy) upon Italian economic thought, especially upon general economic theory and the theory of public finance. Pantaleoni can also be considered among the founders of the modern Italian statistical school and a true forerunner of regional economics. A good example of the most typically Pantaleonian style of reasoning is given by his analysis of the concepts of ‘strong’ and ‘weak’ in economics, so alive and thought-provoking a century later.

See Also

- Marshall, Alfred
- Pareto, Vilfredo

Selected works

1889. *Principii di economia pura*. Florence: G. Barbèra. 2nd edn, 1894. Trans. T.B. Bruce as *Pure Economics*, London: Macmillan, 1898.

1890. Letter to Edgeworth on 15 November 1890. In the possession of David E. Butler, Nuffield College, Oxford.

1898. An attempt to analyse the concepts of ‘strong’ and ‘weak’ in their economic connection. *Economic Journal* 8(June), 183–205.

1904–10. *Scritti vari di economia*. 3 vols. Vols. 1 and 2, Milan: R. Sandron; vol. 3, Rome: Libereria Castellani Editrice.

1925. *Erotemi di economia*. 2 vols. Bari: G. Laterza.

Bibliography

Di Nardi, G. 1977. Pantaleoni rivisitato. *Annali della Facoltà' di Giurisprudenza*. Facoltà' di Macerata. Repr. in G. Di Nardi, *L'Economia a una svolta difficile*. Milan: Giuffré, 1984.

Loria, A. and Sraffa, P. 1924. Maffeo Pantaleoni: obituary. *Economic Journal* 34, 653–4.

Musgrave, R.A. and Peacock, A.T., eds. 1967. *Classics in the Theory of Public Finance*. London: Macmillan; New York: St Martin's Press.

Pareto, V. 1960. *Lettere a Maffeo Pantaleoni*, 3 vols, ed. G. De Rosa. Rome: Storia ed Economia.

How to cite this article

Becattini, G. "Pantaleoni, Maffeo (1857–1924)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000009> doi:10.1057/9780230226203.1244

The New Palgrave Dictionary of Economics Online

paradigms

Peter Urbach

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

falsificationism; Kuhn, T. S.; marginalist economics; paradigm; Popper, K.; subjective theory of value

Article

The idea of a paradigm, in the sense of a dominating principle governing a whole area of scientific research, was invented by Thomas Kuhn (1962), a historian and philosopher of science. One of his starting points was Karl Popper's view of science, according to which the best scientific theories are falsifiable, viz., such theories could, at least in principle, be refuted by empirical evidence. Popper has argued that theories such as those of Freud and Marx, and the central assumptions of astrology, were unfalsifiable, and that in consequence they were inferior to Newtonian and Einsteinian physics, for example. Kuhn, however, pointed out the already well-known fact that the leading theories of the physical sciences are not straightforwardly falsifiable. On the contrary, a theory such as Newton's typically requires auxiliary assumptions before it can make any empirical predictions. If such predictions turn out to be false, then logic alone does not determine whether the main theory or one or more of the auxiliaries is at fault, and a person is then at liberty to retain the central theory and to reject one or more auxiliary.

Indeed, according to Kuhn, this is roughly what scientists generally do. Kuhn was particularly impressed by the way in which the Copernican and Ptolemaic theories, which provided rival accounts of the planetary system, were each sufficiently flexible to account for any astronomical observations. New observations that did not match expectation could always be reconciled with either the Copernican or the Ptolemaic theories by adjusting the system of epicycles, equants, and so on, upon which the planets were supposed to revolve. Kuhn concluded from this that no objective method could determine which theory is the right one or the better one, and that the transfer of allegiance from the earth-centred Ptolemaic to the sun-centred Copernican hypothesis was not based on the rational method of comparing the relative empirical support enjoyed by the two theories: ‘the real appeal of sun-centred astronomy was aesthetic rather than pragmatic’ (Kuhn, 1957, p. 172).

Kuhn described the Copernican and Ptolemaic programmes as rival paradigms. A paradigm, for Kuhn, is

a ‘[way] of seeing the world and of practising science in it’ (Kuhn, 1970, p. 4). Thus a paradigm incorporates a set of assumptions, especially assumptions about the fundamental nature of some aspect of the world, such as whether there are atoms or not. These assumptions are accepted unquestioningly by adherents to a paradigm. Scientific research is normally conducted in the context of such a paradigm, but it is not directed towards testing the paradigm; on the contrary, it consists in ‘a strenuous and devoted attempt to force nature into the [paradigm’s] conceptual boxes’ (1970, p. 5).

Paradigm-directed research was dubbed ‘normal science’ by Kuhn to indicate that most work in science is of this kind. Normal science involves fact-gathering activities such as the determination of specific physical quantities (for example, stellar positions), and experiments designed to check theoretical predictions; it also ‘consists of empirical work undertaken to articulate the paradigm theory, resolving some of its residual ambiguities and permitting the solution of problems to which it had previously only drawn attention’ (1970, p. 27). The crucial feature of normal science is that it is directed to ‘puzzle-solving’, and ‘not [to] major substantive novelties’ (1970, p. 35), and predictive failures are normally blamed on the scientist rather than on the paradigm itself.

In addition to a set of assumptions, a paradigm comprises prescriptions for research; however, the precise character of a paradigm is impossible to state. According to Kuhn, the nature of a paradigm resembles the meanings of words, as Wittgenstein characterized them. According to Wittgenstein, a word like ‘game’ cannot be defined fully and explicitly, its meaning can only be intuited or grasped. One learns the meaning of such words by exposure to examples or paradigm uses. The same applies to a scientific paradigm, which is taught through exemplary or paradigmatic applications of a theory to a concrete range of phenomena. A similar view was held by Polanyi (1958).

How is one paradigm supplanted by another? Kuhn claimed that this is not done by the careful weighing of evidence for each and the selection of the one with the greater empirical support. He argued that this would not be possible, first of all because of the inarticulable and elusive character of a paradigm and secondly because paradigms are ‘incommensurable’, in that they are really dealing with quite separate phenomena. Kuhn claimed that the observations made by scientists are never ‘pure’ or ‘theory-free’, but are always interpreted in terms of the prevailing paradigm. In Kuhn's view, this means that there can be no data that would facilitate a comparison between two paradigms (Kuhn, 1970). Hence, Kuhn argued, a rational comparison of the two paradigms is impossible and ‘communication across the revolutionary divide is inevitably partial’ (1970, p. 169).

Kuhn likened a change of paradigm, or scientific revolution, to a gestalt-switch, calling it ‘a transition between incommensurables’ (1970, p. 50). After a prolonged period of repeated failures to resolve anomalies in accordance with a paradigm's internal criteria, a time of ‘crisis’ sets in. Scientists are then prone to change paradigm all of a sudden, after which they see the world in terms of a new paradigm. Sometimes the ‘shape of the new paradigm is foreshadowed’ during the period of crisis. More often, though, ‘the new paradigm, or a sufficient hint to permit later articulation, emerges all at once, sometimes in the middle of the night, in the mind of a man deeply immersed in crisis’ (1970, pp. 89–90).

Some aspects of Kuhn's view of science have been widely accepted, in particular, his claim that theories may be tenaciously upheld in the face of apparently unfavourable evidence and that such theories often generate a long programme of research, in which they are defended and extended to new areas.

Other Kuhnian claims have, however, been contested, especially the idea that the nature of a paradigm is inarticulable and that apparently competing paradigms are incommensurable. Lakatos argued that paradigms, or research programmes, as he called them, could be accurately described. In his view, they

consisted of a set of unfalsifiable theories (or a ‘hard core’), together with a heuristic, or instructions and hints on how to apply the hard core theories to specific explanatory tasks.

The idea that different paradigms are incommensurable has been criticized in a number of ways. The most telling objection is this: although two theories, such as the Newtonian and Einsteinian theories, describe the world in terms peculiar to each and so, in a sense, are disconnected, when combined with appropriate auxiliary assumptions, they may each make predictions that are comparable. Such predictions are not expressed in a pure observation language, but they may be expressed in terms common to both paradigms, and this, contrary to Kuhn's claim, is all that is needed for a comparison of the explanatory powers of rival paradigms. For example, Newton's theory and Einstein's made different predictions over the rate of precession of the perihelion of the planet Mercury and this may be described in a way that is valid and comprehensible, whichever of these theories one favours.

It has been argued that such comparisons permit the relative merits of paradigms to be *rationally* assessed, and allow one to conclude that one of them is ‘objectively’ the better. Thus Lakatos spoke of a research programme that consistently leads to successful novel predictions as ‘progressive’, and those that produce a string of failures as ‘degenerating’. Lakatos believed that he could advance beyond Kuhn by exploiting this dichotomy to supply rational criteria for paradigm choice. However, he was unable to justify this claim.

Treating scientific research in terms of paradigms or programmes is certainly a useful historiographical tool of analysis. However, there are often difficulties in determining the boundaries of a paradigm. This difficulty means that one often cannot be sure whether a change of view amongst scientists constitutes a revolution, that is, a replacement of one paradigm by another, or whether it is merely a change within a single, more general paradigm (Toulmin, 1970).

Perhaps the most significant development in economics that resembles a paradigm is that of marginalism, in particular, the subjective theory of value and the associated methods of marginal analysis. This became the dominant approach of economists from the 1870s on (see Schumpeter, 1954, ch. 7). However, a number of authors have picked out parts of this general approach as constituting separate paradigms, or research programmes. For instance, Latsis has given a detailed account of the hard core and heuristic methods of the neoclassical theory of the firm. And Blaug has outlined the paradigm of ‘economic equilibrium via the market mechanism’, which he traced to Adam Smith. De Marchi has described the Ohlin–Samuelson approach, in which the relative commodity outputs of different nations are connected to factor endowments, as a paradigm. Within macroeconomics there are competing paradigms, namely that which assumes continuous market clearing based on rational expectations (new classical macroeconomics), and that which is based on the old and neo-Keynesian approaches (see Klamer, 1984).

A disadvantage of Kuhn's and Lakatos's views is that they suggest that the rules of science allow any group of theories to be set up as the basis of a paradigm. This has led to some areas of inquiry, especially those dealing with social phenomena, mistakenly being regarded as scientific, simply because they are dominated by some tenet that is dogmatically upheld against all difficulties. That such conditions are insufficient for a theory to be acceptable as science has been argued by Dorling (1979). Proceeding from the assumption that theories are judged by their probabilities, in the light of the available evidence, Dorling demonstrated the conditions under which a theory may remain very probable, even when the combination of that theory with some auxiliary hypothesis has been refuted. Redhead (1980) showed how this fact could lead to some theories forming the basis of a paradigm, and how, after a number of

unsuccessful predictions, confidence in the assumptions of the programme would gradually be eroded. Thus the crucial characteristics of paradigms may be explained and rationalized.

Bibliography

Blaug, M. 1976. Kuhn versus Lakatos or paradigms versus research programmes in the history of economics. In *Method and Appraisal in Economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

De Marchi, N. 1976. Anomaly and the development of economics. In *Method and Appraisal in Economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

Dorling, J. 1979. Bayesian personalism, the methodology of scientific research programmes, and Duhem's problem. *Studies in History and Philosophy of Science* 10, 177–87.

Klamer, A. 1984. *The New Classical Macro-economics: Conversations with the New Classical Economists and the Opponents*. Brighton: Wheatsheaf Books.

Kuhn, T.S. 1957. *The Copernican Revolution*. Cambridge, MA: Harvard University Press.

Kuhn, T.S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press. 2nd edn, 1970.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In Lakatos and Musgrave (1970).

Lakatos, I. and Musgrave, A., eds. 1970. *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Latsis, S.J. 1976. A research programme in economics. In *Method and Appraisal in Economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

Polanyi, M. 1958. *Personal Knowledge*. London: Routledge & Kegan Paul.

Redhead, M. 1980. A Bayesian reconstruction of the methodology of scientific research programmes. *Studies in the History and Philosophy of Science* 11, 341–7.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: George Allen & Unwin.

Toulmin, S. 1970. Does the distinction between normal and revolutionary science hold water? In Lakatos and Musgrave (1970).

How to cite this article

Urbach, Peter. "paradigms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000016> doi:10.1057/9780230226203.1245

The New Palgrave Dictionary of Economics Online

paradoxes and anomalies

N. De Marchi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Allais paradox; anomalies; Arrow's theorem; capital reversal; falsificationism; Gibson paradox; Giffen goods; Giffen paradox; Kuhn, T. S.; Leontief paradox; Mill, J. S.; paradox of thrift; paradoxes and anomalies; parametric paradox; Popper, K.; preference reversal paradox; private vices, public virtues; social choice; St Petersburg paradox; transitivity; voting paradoxes; water–diamonds paradox

Article

Paradox originally meant contrary to accepted opinion. In logic something more precise is usually intended. A paradox is involved, for example, if we are led to a contradiction by sound reasoning. Economists on the whole seem to have stayed closer to the original sense. We can use this fact to claim that there is ground for treating economists' paradoxes simply as puzzling outcomes. There have been rhetorical appeals to 'paradox', as we shall see; but there are also numerous examples of substantive puzzles. They are to be expected as the limits to existing ways of explaining are explored. This usage has the advantage therefore of allowing us to treat paradoxes as a normal aspect of ongoing inquiry, and it shifts the focus of interest in them away from a status as intellectual curiosities to a status as stimulant to further research.
Anomaly, in the physical and biological sciences, is used to refer to an observational irregularity, or an exception. Economists are reluctant to claim a similar implied degree of regularity, or reliability in estimation. There are, however, at all times facts which resist ready incorporation within existing ways of explaining. We may reserve the term anomaly for these, to indicate the empirical origin of the puzzles which they pose. For the rest, however, there seems little reason to distinguish sharply between paradox and anomaly.

An interpretative framework

To make much sense of how paradoxes and anomalies relate to change and progress in economics, we

need a framework. It seems probable that economists, like other scientists, are more disposed to admit corrections that they can deal with without altering too radically their existing theories. Challenges which threaten precipitate depreciation of their human capital are likely to be resisted. Furthermore, challenges to 'hard core' propositions, to use Lakatosian terminology, will be resisted absolutely. (This has certain implications for the way 'core' change, if it comes, will be experienced, but these need not occupy us here.) Challenges within the 'protective belt' (in the 'demi-cores', as Remenyi (1979) refers to them, or at the sub-disciplinary nodes) will be dealt with more flexibly, but we may expect a different response according to whether the challenge is empirical or analytical in *origin*. Theoretical challenges, so long as they are less than 'core'-threatening, actually provide occasions for the display of ingenuity, and are a major vehicle for change. Empirical puzzles, on the other hand, for reasons that are pretty well understood (but will be reviewed below) seem generally to be regarded as less compelling. The accumulation of empirical anomalies which eventually become so numerous as to crush a theory by sheer weight, and to which Thomas Kuhn (1970) points as a major precipitating factor in revolutions in physical science, is not at all familiar in economics. Gross anomalies, of course, such as unemployment in the 1930s or stagflation in the 1970s, may provoke basic rethinking; but these are not our concern here.

If these suggestions broadly capture the situation and behaviour of economists, we would expect to observe little impact of empirical anomalies on preferred ways of thinking, and relative autonomy in theoretical developments. As Lakatos puts it (1970, p. 136): 'if the positive heuristic is clearly spelt out, the difficulties of the programme are mathematical rather than empirical.'

Such expectations merely reflect the rationalistic conception of their discipline that many economists seem to hold. Historically, that mind-set derives from John Stuart Mill. Mill, though in principle a radical empiricist (V.R. Smith, 1985, pp. 269–77), managed nonetheless to formulate an economic methodology in which there is no uncertainty, merely incompleteness. This accords well with the hypothetico-deductive model of explanation which economists have known since Mill and have found attractive as re-formulated in recent decades by Sir Karl Popper. On this view, science progresses by the making of bold conjectures, boldness meaning that there is much in the world of 'facts out there' that could refute them; and by subjecting these conjectures to factual tests to identify and help eliminate falsehood. Economists' Millian inheritance leads them to put a particular gloss on this: given our inability to conduct controlled experiments, we tend to look for certainty in premises that we 'know' to be true, by reason of introspection or casual observation. Hence we reason downward *from* truth. In this model, factual evidence can only be at odds with theory if our variables are incorrectly measured, or if we have failed to incorporate all those which are relevant to an explanation, or if the empirical model supposedly corresponding to our theory is incorrectly specified. These attitudes infused the early work of econometricians (Morgan, 1984, chs 5, 7); and even the sophisticated methodology of the Cowles Commission in the 1940s used economic theory in a peculiarly Millian manner, to provide a priori grounds for rendering the problems of structure and causation operationally tractable.

With these general considerations in mind we turn to paradoxes and anomalies. What follows is not a survey, nor is there space to examine any single instance in detail (though several receive fuller attention elsewhere in this Dictionary). The instances mentioned serve us as illustrative material and are drawn together in this way in the hope of stimulating further exploration. We shall look at three categories: rhetorical paradoxes; 'fact of life' paradoxes, such as the failure of aggregation rules; and the main group, theoretical paradoxes and empirical anomalies. This last we shall split, as far as seems sensible,

into challenges to the hard core and positive heuristics of the dominant neo-Walrasian style of analysis or research programme (for which see Weintraub, 1985, ch. 7) and challenges within the protective belt.

Rhetorical paradoxes

Here use is made of terminological fuzziness, or a premise is left unstated, so as to excite puzzlement and interest in the reader. Adam Smith's *diamonds and water paradox* is of the first sort. Neoclassical economists have, on the whole, viewed the puzzle as emanating simply from a confusion of total with marginal utility. An example of the second sort is again provided by Smith. When he avers that ‘it is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest’ (1776, vol. 1, pp. 26–7), the air of paradox is deliberate. It is dispelled when he goes on to relate the proposition to the principle of occupational specialization. Paradox was a favourite literary device in an earlier age. Donald McCloskey (1985) has recently alerted us to many others in the writings of modern economists.

Paradoxes arising from the absence or failure of an aggregation condition

Examples here are the *paradox of thrift*, *Mandeville's paradoxes* about private vices (for example, profligacy) leading to public virtues (jobs) and *Arrow's impossibility theorem*. The first two, like Smith's paradox of self-interested behaviour leading to socially beneficial results or the Austrian view of social outcomes as complexes of individual choices which interact unpredictably, are instances of unintended consequences. Economists have failed to provide convincing reductionist accounts of aggregative behaviour and tend to take unintended consequences as a fact of economic life. Consistently with the dominant commitment to the neo-Walrasian approach, but paradoxically from any other point of view, this does not stop them from employing micro-motives to account for aggregate relations whose entities they cannot explain. (Excellent discussion of these things is to be found in Elster, 1978, ch. 5, and in Nelson, 1984.) Arrow's theorem, in so far as it is regarded as a generalization of the paradox of voting (as he himself is inclined to view it) creates difficulties at different levels in different problems (Sen, 1985); but far from issuing in defeatism or the rejection of economic rationality it has given rise to a whole new sub-discipline, social choice theory.

Challenges to the neo-Walrasian hard core and positive heuristics

Here we shall consider examples of both theoretical and empirical origin and note responses within the profession.
Take first the possibility of *capital reversal* or *reswitching*. In considering an array of techniques in a two-factor, two-product model, capital reversal occurs if, as the wage rate rises (interest rate falls), a less capital-intensive technique is chosen. A far-reaching implication is contained in this simple possibility. If there is no strictly monotonic relation between interest and the capital–labour or capital–output ratio, then it is conceivable that a more, then a less, then once again a more capital-intensive technique is the more profitable as the interest rate declines. This undermines the traditional demand curve for ‘capital’,

negative in the interest rate, since relative goods prices may differ as between two interest rates at which the same technique may be equally profitable. There is then no unambiguous way to value ‘capital’ (Blaug, 1985, pp. 523–8). The very possibility seems to render unserviceable the traditional aggregate production function. Neo-Walrasian economists in effect concede all of this yet go on using devices like the factor-price frontier, suggesting something like absolute resistance to challenges to the basics of the dominant research programme.

A curious instance with somewhat similar implications is the *Giffen paradox*. The positively sloped demand curve was thought of by Marshall as an empirical anomaly, and ‘discoveries’ of such phenomena by early econometricians led to identification and other ‘correspondence’ problems being defined (Morgan, 1984, ch. 6). It has always been doubtful whether there are any actual observations of Giffen goods, and the strong presumption of theorists and theoretically influenced econometricians has been that, in Stigler's words, ‘experience and common sense are opposed to the idea of a positively sloped demand curve’ (1965, p. 384). Thus even the standard price-theoretic rationalization in terms of a negative income effect dominating weak substitution effects (possibly due to strong rivalry between goods) is quite unaffected by the fact that tests normally turn up positive income effects. This merely confirms the theorist's suspicion. While the case of Giffen goods is not all that significant, the typical theorist's attitude in this instance is interesting because it is wholly in line with what is observed elsewhere: within the programme, theoretical developments are relatively autonomous. Tests of demand theory were reported in 1975, for example, which in the words of the authors ‘make possible an unambiguous rejection of the theory of demand’ (Christensen, Jorgenson and Lau, 1975, p. 381). The authors, however, did not refer to their own results as puzzling or anomalous, and their frontal assault also went unremarked.

A third example, involving experimental evidence, is the *preference reversal paradox*. Experimental trials conducted in the 1970s and 1980s have caused consternation by consistently implying intransitivity between individuals’ direct preference rankings over risky prospects and the respective certainty equivalents they assign to them. Individuals will choose a high probability of low gain over a low probability of high gain while assigning a higher monetary value or certainty price to the second. This evidence appears to undermine all theories of choice which require transitivity (Machina, 1983, pp. 76 ff.). As Imre Lakatos points out, however, ‘there is no falsification before the emergence of a better theory’ (1970, p. 119); and economists who use standard choice theory have remained impassive in the face of this evidence. The literature openly addressing the matter is apt to challenge the experimental design or to argue that intransitivity of certainty equivalents does not imply intransitivity of preferences (Safra and Karni, 1984).

Challenges in the protective belt

Here we are much more likely to see positive responses to puzzling outcomes since there is more room for theoretical manoeuvre. Well-known examples among challenges of this sort include the *St Petersburg paradox*, the *Allais paradox*, the *Gibson paradox* and the *Leontief paradox*.

The Bernoulli solution to the St Petersburg paradox has been amended by placing bounds on the utility function. The Allais challenge to the independence axiom of expected utility theory has produced some modification in the specification of the subjects’ choices – questioning the experimental design is

something of a standard defence – but has issued mainly in the development of non-expected utility models (Schoemaker, 1982). It is perhaps worth noticing that the Allais results might have been taken as a challenge to choice theory as much and therefore as threatening to the hard core; but turning the issue into one of model choice has deflected the threat into the protective belt. The Leontief paradox was explained initially as very largely due to omitted or improperly specified variables (De Marchi, 1976), but more recent work has focused on a full articulation of the theory that would suffice to generate trade-revealed factor abundance (Leamer, 1984, pp. 50 ff.). The Gibson paradox – an observed close correlation between long series of prices and bond yields – was given a complex explanation by Keynes in terms of market interest rate stickiness relative to the natural rate (Keynes, 1930, vol. 6, pp. 182–3). Keynes thought of his account as rejecting the simple quantity theory account of the matter; but more recently Gibson's observations have simply been absorbed into the infinite time-horizon, intertemporal choice models of macroeconomics embodying a modified quantity theory approach.

What we see at work in all four of these cases is the emergence of a paradox being taken as an occasion for further theoretical refinement. Only one challenge was analytic in origin, but the three instances of empirical anomaly were treated as invitations to amend theory rather than to abandon it, and in this sense they were subsumed under the power of the positive heuristic. The difference between the way empirical anomalies are handled when the hard core is threatened and when it is not does not turn on any weakening of the rationalistic presumption that theory is the true arbiter. It is simply a methodological choice that is made: challenges to the hard core are inadmissible; challenges to theories in the protective belt allow of a more positive reaction.

Methodological reasons why empirical results are less compelling

Finally, it is worth asking again why empirical tests result which look ‘wrong’ generally seem to be judged so, rather than to be taken as falsifications of theory. Theory may be modified by empirical challenges in the protective belt, but is unlikely to be rejected.

One problem is the economist's special version of the Duhem–Quine thesis to the effect that one never tests an hypothesis in isolation, but always in combination with a host of background conditions. Because it is often difficult to know the exact translation of theoretical terms in economic theories into empirical counterparts, and because the data often are not quite what the theory requires, tests are mostly joint tests of theory *and* the adequacy of proxies. A major reason, in turn, why the translation is not clear cut is that economic theories tend to be generic (choice theory, for example). It is possible to model such theories in many, many specific ways; but a negative test result for any one such specification has no implications for the generic theory. A third reason is that we do not observe in economics the constants of physical science. Especially in policy contexts it tends to be the case that altering a parameter of interest causes changes in other parameters. This is a particular form of what Klant has called the *parametric paradox* (1984, ch. 4.9). Here, and in much comparative statics analysis, our ‘constants’ are algebraic rather than numerical: they actually function as variables. This robs our models of predictive content unless very special restrictions can be devised (such as cross-equation restrictions in macro policy models) and rendered operational.

In sum, paradoxes, regarded as puzzling outcomes, are normal occurrences in the ordinary line of economic inquiry. Where they impinge on those basic commitments that determine what is ‘acceptable’

in a line of research there is every incentive to dismiss or ignore them. This is true for analytical puzzles (for example, reswitching) as well as for those in the form of anomalous empirical findings (for example, experimental results indicating preference reversal; Giffen goods; and contrary tests of basic demand theory). It is much easier to modify specific theories in the protective belt, though responses to challenges occurring there will still be driven by the positive heuristics of the research programme (as in the case of the Leontief and Allais paradoxes). The presumption even at this level is that theory arbitrates. There are good reasons for this in the generic nature of economic theories, in the specific difficulties of translating theory into an empirical model and in getting appropriate observations, and in the fact that genuine constants have not been found in economics.

Bibliography

Allais, M. and Hagen, O., eds. 1979. *Expected Utility Hypotheses and the Allais Paradox*. Dordrecht: Reidel.

Amihud, Y. 1979. Critical examination of the new foundations of utility. In Allais and Hagen (1979).

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Blaug, M. 1985. *Economic Theory in Retrospect*. 4th edn. Cambridge: Cambridge University Press.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1975. Transcendental logarithmic utility functions. *American Economic Review* 65, 367–83.

De Marchi, N. 1976. Anomaly and the development of economics: the case of the Leontief Paradox. In *Methods and Appraisal in Economics*, ed. S. Latsis. Cambridge: Cambridge University Press.

Elster, J. 1978. *Logic and Society*. London: Wiley.

Harcourt, G.C. 1973. *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.

Hayek, F.A. 1931. The ‘paradox’ of saving. *Economica* 11(May), 125–69.

Keynes, J.M. 1930. *A Treatise on Money*. In *The Collected Writings of John Maynard Keynes*, vols 5 and 6, London: Macmillan for the Royal Economic Society, 1971.

Klant, J.J. 1984. *The Rules of the Game. The Logical Structure of Economic Theories*. Cambridge: Cambridge University Press.

Kuhn, T.S. 1970. *The Structure of Scientific Revolutions*. Revised edn. Chicago: University of Chicago

Press.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the Growth of Knowledge*, ed. I. Lakatos and A. Musgrave. Cambridge: Cambridge University Press.

Leamer, E.E. 1984. *Sources of International Comparative Advantages: Theory and Evidence*. Cambridge, MA: MIT Press.

Leontief, W.W. 1953. Domestic production and foreign trade: the American capital position re-examined. *Proceedings of the American Philosophical Society* 97, 332–49.

Machina, M.J. 1983. The economic theory of individual behavior toward risk: theory, evidence and new directions. Technical Report No. 433, Center for Research on Organizational Efficiency, Stanford University.

McCloskey, D.N. 1985. *The Rhetoric of Economics*. Madison: University of Wisconsin Press.

Morgan, M.S. 1984. The history of econometric thought. Analysis of the main problems of relating economic theory to data in the first half of the twentieth century. Ph.D. thesis, University of London.

Nelson, A. 1984. Some issues surrounding the reduction of macroeconomics to microeconomics. *Philosophy of Science* 51, 573–94.

Remenyi, J.V. 1979. Core demi-core interaction; toward a general theory of disciplinary and subdisciplinary growth. *History of Political Economy* 11, 30–63.

Safra, Z. and Karni, E. 1984. ‘Preference reversal’ and the theory of choice under risk. Working Papers in Economics No. 154, Department of Political Economy, Johns Hopkins University.

Samuelson, P.A. 1977. St Petersburg Paradoxes: defanged, dissected, and historically described. *Journal of Economic Literature* 15, 24–55.

Schoemaker, P.J.H. 1982. The expected utility model: its variants, purposes, evidence and limitations. *Journal of Economic Literature* 20, 529–63.

Sen, A.K. 1985. Social choice and justice: a review article. *Journal of Economic Literature* 23, 1764–6.

Smith, A. 1776. In *An Inquiry into the Nature and Causes of the Wealth of Nations*. 2 vols, ed. R.H. Campbell and A.S. Skinner, Oxford: Clarendon Press, 1976.

Smith, V.R. 1985. John Stuart Mill's famous distinction between production and distribution. *Economics*

and Philosophy 1(2), 267–84.

Stigler, G.J. 1965. *Essays in the History of Economics*. Chicago: University of Chicago Press.

Weintraub, E.R. 1985. *General Equilibrium Analysis: Studies in Appraisal*. Cambridge: Cambridge University Press.

How to cite this article

De Marchi, N. "paradoxes and anomalies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000018> doi:10.1057/9780230226203.1246

The New Palgrave Dictionary of Economics Online

Pareto distribution

Josef Steindl

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Using certain data on personal income V. Pareto (1897) plotted income on the abscissa and the number of people who received more than that on the ordinate of logarithmic paper and found a roughly linear relation. This Pareto distribution or Pareto law may be written as

$$x = ay^{-\alpha} \text{ or } \log x = a' - \alpha \log y \tag{1}$$

where α (the negative slope of the straight line) is called the Pareto coefficient. The density of the distribution is

$$dx = a\alpha y^{-\alpha+1} dy$$

The Pareto coefficient is occasionally used as a measure of inequality: The larger α the less unequal is the distribution. According to Champernowne (1952), α is useful as a measure of inequality for the high income range whereas for medium and low incomes other measures are preferable.
α takes only positive values. If $\alpha<2$ the distribution has no variance; if $\alpha<1$ it has no mean either. In practice the Pareto law applies only to the tail of the empirical distributions i.e. to incomes above a certain size. Thus the law (1) is valid asymptotically as $y\rightarrow\infty$. The range in which the empirical distributions conform to the law is different in different cases. It seems to be larger for wealth than for income (perhaps because we have data only for large wealth) and even larger for towns. In the case of firm sizes only very large firms are covered by the law.

In the case of the distribution of towns by size of population the rank-size relation has been used (Zipf, 1949) which is the same as the Pareto distribution except that it uses rank as a measure of the tail (instead of the number of towns above a certain size) so that the higher the rank (beginning with rank one for the largest town) the smaller the size of the town. Zipf believed (incorrectly) that the coefficient α is always about one so that the product of rank and size is constant. But Pareto, of course, was even more ‘out’ with his belief that the Pareto coefficient for income α always equals unity. In highly industrialized countries today it is above 2 and sometimes above 3.

The main interest of the Pareto distribution lies not in its rather limited use as a measure of inequality but in the explanations it has provoked, naturally so since regular patterns are felt to be a challenge to the mind. There are two types of approach to the problem. That of Champernowne, Yule and Simon explains the characteristic pattern as the steady state of a stochastic process which has been evolving in time, so that the pattern reflects something which has been going on in the past. In contrast, Mandelbrot has been looking for a ‘synchronic’ explanation which does not depend on a process in time. He is mainly concerned with the reproductive quality of the Pareto distribution: If a large number of independent random variables is identically distributed according to Pareto's law then the sum of these random variables will also be distributed according to this law. Thus it could be expected that the income of the various counties in England would be Pareto distributed because it results in each case from the addition of individual incomes which are Pareto distributed.

Champernowne’s pioneering work (1953) in essence goes back to his fellowship dissertation of 1936, published in 1973. He builds on a tradition which explains the normal distribution as the result of the addition of random unit steps (left or right) on the line over a long time (random walk; for the terms and concepts relating to random processes, see Feller, Vol. I). If the random walk takes place on the logarithmic scale the distribution of the sum of steps will tend to log normality. This does not give, however, a stable distribution, because the dispersion will go on increasing all the time. Champernowne chooses the technique of the Markov chain: Each year's income depends only on the previous year's income plus a random increment proportionate to last year's income; the probability of various increments remains constant from one year to the other. This feature is called the law of proportionate effect. Thus the required data will be embodied in a matrix which contains the probabilities of transition from one income in one year to another income in the following year. The number of income receivers remains stable in Champernowne's model because each exit is assumed to be automatically compensated by a new entry. To guarantee that the system reaches a steady state it is assumed that on the average the change of income is downwards; this is necessary to compensate the tendency of the system to diffusion which is characteristic of the unrestrained random walk. The assumption reflects the low income of new entrants. In fact the role of new entry is crucial not only in this model but in other applications as well (size of firms, towns, wealth).

H. Simon (1955) studied the number of times a particular word (vocable) occurs in a text. The number of vocables which occur with a given frequency decreases with that frequency in a Pareto-like fashion. Simon's treatment is based on the work of Yule (1924), who dealt with a biological problem: the frequency of genera with different number of species which is distributed according to Pareto. He explained this pattern by means of a pure birth process deriving from this the Yule distribution with density

$$f(n) = \alpha\Gamma(1+\alpha)n^{-1-\alpha} \text{ as } n \to \infty .$$

The model of evolution assumes that mutations occur randomly with a frequency *g* per time unit, creating new genera, and with a frequency *s* per time unit creating new species, where $g<s$. Since each species has the same chance of creating a new species we have here a proportionate growth, in analogy to the law of proportionate effect. The steady state is produced by the emergence of new genera. The Pareto coefficient equals the ratio of the frequencies with which the two kinds of mutations appear, that is g/s. Simon, whose merit it is to have drawn attention to this brilliant work, has suggested application to incomes (not very convincingly) and has himself applied it to firm sizes (1967). A very direct application relates to the size of towns (Steindl, 1965). If the number of towns grows at the rate of μ and the number of inhabitants of the town grows at the rate of ρ then after a sufficiently long time there will be a steady state distribution with Pareto coefficient μ/ρ.

Mandelbrot (1960, 1961) deals with the problem from the point of view of a mathematician and therefore on a very general level. He starts from the concept of stable laws (compare Feller, Vol. II, ch. VI). If a sum of independent identically distributed random variables is distributed in the same way as its components, except for a scale factor and possibly of a location factor, then this distribution is stable. The best-known example is the normal distribution. It has been shown by *P*. Lévy that there is a class of distributions with infinite variance which are stable and which converge to the law of Pareto when the variable in question (say, income) tends to infinity. The Pareto law in this context is confined to the range $1<\alpha<2$. Mandelbrot surmises, owing to the reproductive quality, in the above sense, of the Pareto law, that its importance empirically must be very great. He also considers that this must have implications for some statistical methods which depend on the assumption of normalcy.

As to income, Mandelbrot suggests that it can be regarded as composed of a number of independent elements which are identically distributed. We can easily imagine a decomposition into a few parts such as earned income, property income and transfer income. Mandelbrot requires, however, in order to assure convergence, a large number of components, and these, as he admits, have hardly any counterparts in reality (1961, p. 525). The explanation is analogous to the well known explanation of the stature of adult men as a random variable composed of a great number of independent small random variables; this explains the normal distribution of height. The precise identity of these small random variables is, here again, not specified and rather speculative. This may perhaps explain why this ‘synchronous’ approach has not, so far, found much resonance among economists.

The interest of the alternative approach (Champernowne or Yule) of explaining the law as a steady state of a stochastic process is that it establishes a relation between the stratification found in a cross section and the past history which has produced it, and which is mapped in the cross section. This is analogous to the stratifications in geology or the rings in the trunk of a tree. Irregularities or shifts in the empirical distributions can according to this view be explained by major disturbances of the process in certain points of time in the past.

Concretely, the Pareto distribution has been shown, in the case of a birth and death process model, to depend on growth; in an economy which has always been stationary it would not exist (Steindl, 1965). The Pareto coefficient in such models is usually a ratio of growth rates; thus in the case of firm size it is

a ratio of the growth rate of the number of firms to the growth rate of the firms themselves (Steindl, 1965). The importance of new entry as a factor making for less inequality has also been shown, *inter alia* in the case of wealth (Steindl, 1972).

The stochastic models have often been criticized for their lack of economic content. Perhaps it has been overlooked that they only represent the first steps in a new and exceedingly difficult terrain. It may be thought that the work of Champernowne, Yule, Simon and Wold and Whittle contains the seed of future studies which will reveal their full potentiality only when they are extended to distributions in several dimensions.

See Also

- Gini ratio
- lognormal distribution
- Lorenz curve

Bibliography

Champernowne, D.G. 1952. The graduation of income distributions. *Econometrica* 20, October, 591–615.

Champernowne, D.G. 1953. A model of income distribution. *Economic Journal* 63, June,318–51 Reprinted in D.G. Champernowne, *The Distribution of Income between Persons*, Cambridge: Cambridge University Press, 1973.

Feller, W. 1950, 1966. *An Introduction to Probability Theory and its Applications* 2 vols. Reprinted, New York: John Wiley & Sons, 1968, 1971.

Ijiri, Y. and Simon, H.A. 1964. Business firm growth and size. *American Economic Review* 54, 77–89.

Mandelbrot, B. 1960. The Pareto–Lévy law and the distribution of income. *International Economic Review* 1(2), May, 79–106.

Mandelbrot, B. 1961. Stable Paretian random functions and the multiplicative variation of income. *Econometrica* 29(4), October, 517–43.

Pareto, V. 1897. *Cours d'économie politique*. Lausanne: Rouge.

Simon, H.A. 1955. On a class of skew distribution functions. *Biometrika* 42, December, 425–40. Reprinted in H.A. Simon, *Models of Man: Social and Rational*, New York: John Wiley, 1957.

Steindl, J. 1965. *Random Processes and the Growth of Firms. A Study of the Pareto Law*. London: Griffin.

Steindl, J. 1972. The distribution of wealth after a model of Wold and Whittle. *Review of Economic Studies* 39(3), July,263–79.

Wold, H.O.A. and Whittle, P. 1957. A model explaining the Pareto distribution of wealth. *Econometrica* 25, October, 591–95.

Yule, G.U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis. *Philosophical Transactions of the Royal Society of London*, Series B 213, May, 21–87.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Reading, Mass.: Addison-Wesley.

How to cite this article

Steindl, Josef. "Pareto distribution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000023> doi:10.1057/9780230226203.1248

The New Palgrave Dictionary of Economics Online

Pareto efficiency

B. Lockwood
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bounded rationality; constrained efficiency; cost–benefit analysis; Efficiency; externalities; incomplete contracts; incomplete markets; lump sum taxes; market failure; market socialism; multiple equilibria; neoclassical economics; optimal taxation; Pareto efficiency; Pareto optimality; Pareto, V.; planning; signalling; socialist calculation debate; welfare economics; X-efficiency

Article

Few concepts are more widely used within economics than that of ‘efficiency’. It usually means not wasteful, or doing the ‘best’ one can with available resources. However, there are specialized usages; for example, the concept of efficient markets in the finance literature, or Leibenstein's concept of X-inefficiency. Not all these meanings, even in academic work, have a common provenance. However, the concept as used in neoclassical economics has a precise but rather narrower meaning, given to it by Pareto, the Italian economist and sociologist, in his works *Course in Political Economy* and *Manual of Political Economy* around the turn of the 20th century. He suggested the following definition: an allocation of resources in the economy was optimal if there existed no other productively feasible allocation which made all individuals in the economy at least as well-off, and at least one strictly better off, than they were initially. Although Pareto actually used the word ‘optimal’, this is really a definition of efficiency, as a Pareto-‘optimal’ allocation of resources is ‘good’ only in the limited sense that not everybody can be made better off. It may in fact be very undesirable in some other way, for example, very unequal. It is not surprising, therefore, that the word ‘Pareto-optimal’ has gradually been replaced by ‘Pareto-efficient’.

There are several points to note about this definition. First, it is only well defined within a neoclassical framework, that is, where the preferences of individuals and the technical possibilities of production are taken as the ultimate data of economic analysis. Secondly, even within this framework, it is an ordinal concept of efficiency, as it does not rely on any intensity of preference, interpersonal comparability of utilities, or commensurability of different inputs or outputs for its definition. This is no accident; Pareto was a convinced ordinalist, who believed that the utilitarian concept of introspective utility was

unscientific (see for example Pareto, 1927, p. 113). Thirdly, while it provides a ranking of allocations of economic goods between individuals, it does not permit a ranking of all such allocations, that is, there are many different allocations that are Pareto-optimal and which differ with respect to the distribution of real income (that is, utility) among the individuals in society.

Simple and limited idea though this is, it has had an enormous influence on the development of neoclassical economics. First and foremost this is because Pareto did not simply present this notion of optimality as an abstract criterion, but showed that competitive equilibrium would yield an optimal allocation of resources in this sense, thus making precise the notion of the 'invisible hand'. It is no exaggeration to say that the entire modern microeconomic theory of government policy intervention in the economy (including cost–benefit analysis) is predicated on this idea. It also stimulated other debates, such as the one over 'market socialism' in the 1930s, which led to modern theories of economic planning. However, it has by its very success inhibited investigation of other criteria for the performance of economic systems. More radical commentators argue that Pareto, and what followed after, also serves an ideological purpose, namely, to show that capitalism is inherently self-regulating, with phenomena such as unemployment being explained as deviations from an 'ideal' equilibrium rather than inherent structural problems. In this article, we briefly review the historical evolution of the idea, and then attempt a critical assessment.

As already remarked above, the context in which Pareto first presented his concept of optimality was in demonstrating that competitive equilibrium was optimal, of efficient. This crystallized the notion, present at least since Adam Smith, that free trade has (possibly unintended) beneficial consequences. His arguments were much refined and extended over the years by figures such as Barone, Lerner, Hicks and Samuelson, although it took some 20 or 30 years after the *Manual of Political Economy* was published for the 'new' welfare economics to become common currency. The current version of the proposition is essentially based on the work of Arrow and Debreu (for example, Debreu, 1959) who generalized and clarified the mathematics of the result. They showed that it is in fact a twofold proposition; under certain conditions, competitive equilibrium is Pareto-efficient, and second, the additional assumption of non-increasing returns to scale, any Pareto efficient allocation of resources may be decentralized as a competitive equilibrium. These statements are known collectively as the two theorems of welfare economics.

Before going on to discuss them, one should note that there is, to begin with, a problem with the notion of 'competitive'. By this, we simply mean here that all firms (or more generally, all agents) take prices as given, not necessarily that they are 'small' relative to the economy. The problem is that the former is not generally plausible behaviour unless the latter is true. Therefore, the result should really be thought of as approximate – that is, with price or quantity-setting firms equilibrium is approximately competitive, and hence approximately optimal when firms are 'small' – although it is not usually presented in this way.

Now, given price-taking behaviour, the sufficient conditions for the first theorem are (i) that there are no externalities and (ii) that there are complete contingent markets for all commodities (apart from externalities), that is, markets at all present and future dates and states in all contingencies. Implicit in (ii) is the assumption that all agents are equally and perfectly informed about all aspects of their environment. The reason why the first condition is sufficient (and, generally, necessary) is simply that externalities such as air pollution, are in this framework goods (or, more properly, 'bads') for which no markets exist, so there is no mechanism for the marginal benefits of the externality-producing activity to

be equated to the marginal damages they impose on others.

The role of complete contingent markets, however, is not so immediately apparent. The reason is as follows. Consider, for example, a two-period economy with spot markets, but no means of transferring income from period to period (that is, no securities or money). Then, clearly, individual marginal utilities of income will not be equalized in the two periods. Equalization is, however, a necessary condition for full Pareto efficiency, as otherwise a reallocation of income between periods can make at least one agent better off than in the original position. The same argument applies *a fortiori* if there are no, or limited, means of transferring income between different states of the world. If there are complete contingent markets, however, this problem cannot arise.

Some, however, have suggested that, with incomplete markets, full Pareto efficiency is too demanding a performance criterion, and have suggested that one should reformulate the concept to take into account the inherent restrictions on allocation of resources when markets and information are incomplete. This is sometimes called constrained Pareto efficiency. The problem with such an approach is that the exact definition of constrained efficiency is often arbitrary. To take an example, Hart (1975) proposed the definition that a competitive equilibrium with incomplete markets was constrained Pareto-optimal if there was no other competitive equilibrium relative to the same allocation of endowments which Pareto-dominated it. This seems a weak criterion, (for a start, it only has force when there are multiple equilibria), but nevertheless he showed surprisingly that not all equilibria were Pareto-efficient even in this sense, that is, that multiple equilibria could be Pareto-ranked. On the other hand Gale (1982) has proposed an even weaker notion of Pareto efficiency relative to which the first theorem is true, and there is no way of deciding which ‘the’ correct measure of performance is. In addition, the issue becomes even more complex in the more interesting case where information is asymmetric, with the concomitant phenomena of signalling, adverse selection, moral hazard, and so on. Here, competitive equilibrium may ‘fail’ in a number of new ways; for example, resources may be wasted on signalling. In summary, Pareto's argument is not general; the invisible hand becomes very shaky when unrealistic assumptions are dropped.

Therefore, very few people take the theorems of welfare economics seriously as descriptions of the real world. The main significance of the two theorems has been in generating a framework for evaluation of government intervention in the economy: this framework has dominated neoclassical thinking about public policy. One can distinguish two types of policy analysis. The first, which we can call ‘market failure’ analysis, abstracts from distributional considerations by supposing that the government can lump-sum tax individuals in the economy. The procedure is to compare the ‘real’ economy to the complete contingent markets economy, which is known to be efficient, and on the basis of this prescribe policies that either mimic or replace markets to some extent, or more generally alleviate the inefficiency. The classic example of this approach is in the externalities literature, which proposes either the creation of artificial markets in the externalities via the assignment of property rights (à la Coase), or the imposition of corrective taxes (à la Pigou). This kind of prescriptive analysis may seem excessively utopian; however, some have used the market failure paradigm descriptively to explain the existence of various institutions as replacements for markets, a classic example being Arrow's (1963) discussion of health care.

The second type of policy analysis is concerned with the ‘problem of redistribution’ or how to redistribute pre-tax incomes to satisfy distributional objectives without the benefit of lump-sum taxation, that is, using income or commodity taxes, and so on. In this case, the first theorem says that the initial no-

tax equilibrium is efficient, but, given that redistribution involves distortionary taxation, with redistributive taxes competitive equilibrium will be Pareto-inefficient. Hence, there is a trade-off between Pareto-efficiency and distributional objectives. The literature has, in the main, been concerned with characterizing those tax structures that are ‘second best’ Pareto-efficient, that is, such that there is no change in the available taxes such that all agents can be made better off. There is now a large literature on such optimal redistributive taxation (see for example Mirrlees, 1986).

There are, of course, major problems with the actual implementation of the policy prescriptions that arise from this analysis. First, usually the policy recommendations depend on the taste/technology specification of the models (for example, optimal taxation formulae depend crucially on the functional forms chosen for labour supply and commodity demands) and the latter are only testable to a very limited extent. Second, as Lipsey and Lancaster (1956) showed, the ‘optimal’ policies are also generally sensitive to assumptions made about the existing ‘distortions’ in the economy (for example, taxes, monopolies, and so on) if these are not also controllable by the government or planner. For, as they showed, it may not be desirable to substitute lump-sum taxation for a tax on one commodity if there is a pre-existing tax on another. Finally, the characterization of the trade-off between Pareto efficiency and distribution needs to be complemented by a distributional judgement to provide concrete policy recommendations, for example, a rate of income tax.

The alternative approach to policy, of course, is to suppose that the planner is interested in pursuing a number of ‘intermediate’ objectives such as maximizing the growth of national income or employment, or reduction of inequality of income, or inflation, or some weighted combination of these. However, while these objectives are undoubtedly more operational, and perhaps more philosophically appealing as they do not commit one to a utility-based view of welfare, the above problems will still arise with intermediate objectives.

Therefore, the Paretian approach to policy may have a role to play, especially (i) when Pareto-efficient policies are relatively robust to the structure of tastes/technology and distributional goals as are for example some shadow-price rules for cost–benefit analysis (see for example Drèze and Stern, 1987, pp. 49–62) and (ii) to critically analyse the basis for the choice of intermediate objectives; for example, why is inflation ‘costly’? When Pareto presented his original proof of the efficiency of competitive equilibrium, he seemed also to assert that these conditions could only be attained in a decentralized economy; ‘if one could know all these equations (which describe the optimum) the only means to solve them which is available to human powers is to observe the practical solution given by the market’ (Pareto, 1927). Nevertheless, Barone pointed out explicitly shortly afterwards that the same efficient allocation of resources could be achieved by an omniscient central planner in a ‘socialist’ economy, that is, one where the means of production were collectively owned. This, and other subsequent contributions provoked a debate between, among others, von Mises, Lange and Hayek (see for example Lange, 1936, or Hayek, 1940) about how – if at all – in practice the Central Planning Board (as Lange called it) could achieve this. Lange, for example, proposed a price-based iterative procedure where the CPB effectively replaced the Walrasian auctioneer. Other solutions were also proposed, and since the 1950s these have been extended and formalized by Arrow, Hurwicz, Malinvaud, Heal and others (see for example Heal, 1986). All these schemes, however, essentially use the price system as a means of transmitting information to the central planner. In the end, though, it is questionable whether the market socialism debate has had any real impact on the adoption of market socialism in centrally

planned economies.

The concept of Pareto optimality, while simple, almost trivial in itself, has had an enormous impact on economics. However, by providing an apparently precise measure of the efficiency of an economic system, independent of distributional questions, it has, in my view, inhibited discussion of distributional questions and alternative criteria of efficiency. One reason for this, however, may be that within the ordinal neoclassical paradigm Pareto's definition is the only tenable concept of efficiency; that is, any other concept of efficiency, once reformulated in a neoclassical model, will eventually reduce to it.

An example of this is Leibenstein's (1966) notion of X-inefficiency. When Leibenstein introduced the idea, he sharply distinguished it from the notion of Pareto efficiency. The former was inefficiency in the process of production due to the fact that contracts for labour are incompletely specified, the ‘production function’ is not known, and so on, and so derived from bounded rationality. However, Hart (1983) attempted to capture the notion of X-inefficiency in a fully rational, maximizing model; he identified it with the loss of output due to the fact that managers’ efforts cannot be perfectly observed by shareholders. In this framework, X-inefficiency reduces to Pareto inefficiency relative to the full-information equilibrium. A similar fate befalls Schumpeter's definition of efficiency (see Schumpeter, 1942, p. 188), which emphasizes the long-run performance of the economy – capital accumulation, technical progress, and so forth. He proposed that perfect competition, which is Pareto-efficient in a static sense, would not be efficient in this long-run sense compared with monopolized industries, as the latter survive better in the ‘gale of creative destruction’, as he describes capitalism. However, this concept of long-run efficiency reduces to Pareto efficiency if one writes down a dynamic model of competition. This is not to say that Schumpeter's ideas can all be adequately modelled in a neoclassical framework – they cannot – but simply that no other concept of efficiency can sensibly be formulated within this framework.

Therefore, Pareto efficiency and the neoclassical paradigm go hand in hand. If one rejects some aspects of the paradigm, then Pareto efficiency may not have much meaning. For example, some (for example, Galbraith) have argued that, in practice, there is little consumer sovereignty; if desires are manipulated and fears exploited by advertising and so on, the Pareto criterion is of little use in gauging how well real wants are satisfied. Some Marxists (see for example Rowthorn, 1980) go further than this, and argue that Pareto's proof of optimality serves an ideological purpose, by presenting a picture of capitalism as a harmonious enterprise and distracting attention from its exploitative nature.

See Also

- optimality and efficiency
- rational behaviour
- welfare economics

Bibliography

Arrow, K.J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53, 941–73.

Debreu, G. 1959. *The Theory of Value*. New York: Wiley.

Drèze, J. and Stern, N. 1987. The theory of cost–benefit analysis. In *Handbook of Public Economics*, vol. 2, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Gale, D. 1982. *Money: in equilibrium*. Cambridge: Cambridge University Press.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11, 418–43.

Hart, O. 1983. The market mechanism as an incentive scheme. *Bell Journal of Economics* 14, 366–82.

Hayek, F.A. von. 1940. Socialist calculation; the competitive solution. *Economica* 7, 125–49.

Heal, G. 1986. Planning. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M. Intriligator. Amsterdam: North Holland.

Lange, O. 1936. On the economic theory of socialism, Part 1. *Review of Economic Studies* 4(1), 53–71.

Leibenstein, H. 1966. Allocative efficiency v. X-efficiency. *American Economic Review* 56, 392–415.

Lipsey, R.G. and Lancaster, K. 1956. The general theory of second best. *Review of Economic Studies* 24, 11–32.

Mirrlees, J. 1986. The theory of optimal taxation. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M. Intriligator. Amsterdam: North-Holland.

Pareto, V. 1927. *Manual of Political Economy*. London: Macmillan, 1971.

Rowthorn, B. 1980. Neo-classicism, neo-Ricardianism, and Marxism. In *Capitalism, Conflict and Inflation*, ed. B. Rowthorn. London: Lawrence & Wishart.

Schumpeter, J.A. 1942. *Capitalism, Socialism, and Democracy*. London: Allen & Unwin.

How to cite this article

Lockwood, B. "Pareto efficiency." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000024> doi:10.1057/9780230226203.1249

The New Palgrave Dictionary of Economics Online

Pareto principle and competing principles

Louis Kaplow
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Pareto principle, the seemingly incontrovertible dictum that if all individuals prefer some regime to another then so should society, may conflict with competing principles. Arrow's impossibility theorem and Sen's liberal paradox are two notable examples. Subsequent work indicates more broadly that the Pareto principle conflicts with all non-welfarist principles. This essay surveys these results, including various extensions thereof, and offers perspectives on the conflict, drawing on classical and contemporary work in political economy and economic psychology.

Keywords

Arrow, K.; Arrow's impossibility theorem; behavioural economics; Bentham, J.; Darwin, C.; Harrod, R.; Harsanyi, J.; Hume, D.; interpersonal utility comparisons; liberal paradox; Mill, J. S.; Pareto principle; philosophy and economics; psychology and economics; Rawls, J.; Robbins, L.; second best; Sen, A.; Sidgwick, H.; Smith, A.; social choice; social welfare function; welfare economics

Article

The Pareto (1906) principle holds that, if all individuals strictly prefer one state, regime, or policy to another, then that selection is deemed socially preferable as well. Because of the power of unanimous endorsement, the Pareto principle has understandably been important in normative economic analysis. Even though strict Pareto dominance is unlikely to prevail when society is deciding among plausible competing alternatives (for this would require that literally each of millions preferred the same outcome), the Pareto principle nevertheless offers important guidance. In particular, the principle may help in choosing among or ruling out various other evaluative notions; principles that turn out to conflict with the Pareto principle may accordingly be rejected. Alternatively, if some competing principles seem compelling, they may raise doubts about the ostensibly incontrovertible Pareto principle.
The first sections to follow review two well-established conflicts between the Pareto principle and

certain competing principles: Arrow's (1951) impossibility theorem and Sen's (1970) liberal paradox. The succeeding section presents more recent work that establishes a general conflict between the Pareto principle and all non-welfarist notions, whether they concern rights, justice, or other conceptions of fairness (apart from those pertaining only to the distribution of welfare itself•). A final section examines classically grounded strands of literature on political economy and economic psychology that help reconcile the tension between the seemingly unimpeachable Pareto principle and conflicting non-welfarist principles, many of which have appeal to the public, policymakers, and economists as well. (The Pareto principle is also important in normative economic analysis, notably with regard to the two fundamental theorems of welfare economics, a subject not considered in this article.)

Arrow's impossibility theorem

Perhaps the most famous instance of conflict between the Pareto principle and competing principles is Arrow's (1951) impossibility theorem. Arrow considered social choice procedures designed to generate a consistent social ordering (a complete and transitive ranking) from purely ordinal information about individuals' preferences. In one formulation of Arrow's theorem, the assumptions of universal domain (no restriction on individuals' preferences), independence of irrelevant alternatives (the social ordering of any two alternatives depends only on individuals' orderings of those two alternatives), non-dictatorship (no one individual's preferences completely determine social preferences), and the Pareto principle imply that such a social ordering is impossible.

A large subsequent literature explores whether relaxing some of Arrow's assumptions modestly would make possible procedures that yield robust social orderings. Of particular relevance here are attempts to weaken the Pareto principle. As surveyed in Campbell and Kelly (2002), these efforts have been largely unsuccessful: either there are frequent violations of the Pareto principle or a single individual will have substantial, even if not completely dictatorial, influence.

Nevertheless, Arrow's theorem does not rule out the class of standard, individualistic social welfare functions (SWFs), mappings from individuals' utilities to a measure of social welfare, that are fully consistent with the Pareto principle. Consider the discrete case, in which there are n individuals, U_i•(x) is the utility of the ith individual, and x is a complete description of the pertinent state. Then we can define $W(U_1(x), \ldots, U_n(x))$ as an individualistic SWF (so called because it depends only on individuals' utilities). Assuming, as is standard, that W is increasing in each individual's utility, it follows that, for any set of individuals' utility functions $\{U_i\bullet(x)\}$, W provides a complete and transitive social ordering of all possible social states that is independent of irrelevant alternatives, non-dictatorial, and satisfies the Pareto principle. The classical utilitarian criterion, $W = \Sigma U_i(x)$, is an example of such an SWF.

The possibility of an SWF is restored by altering Arrow's framework to allow the domain of social choice procedures to consist of individuals' utilities rather than just their orderings. This approach entails interpersonal utility comparisons, which during the mid-20th century (and to an extent thereafter) were eschewed in welfare economics, following the argument of Robbins. As Robbins (1935, vii–x; 1938) himself clarified in the second edition of *An Essay on the Nature and Significance of Economic Science* and a subsequent essay, however, his argument was not that interpersonal comparisons should not be made – indeed, they were inevitable – but rather that they involve value judgements rather than scientifically verifiable statements. Much modern welfare economics has pursued analysis of SWFs that

depend on individuals' utilities and not just orderings, presumably because of a belief that preference intensities matter and that interpersonal comparisons are required if distributive judgements are to be made.

Sen's liberal paradox

In 'The impossibility of a Paretian liberal', Sen considered whether the Pareto principle conflicts with a specific notion of liberalism, subsequently described by many (including, on occasion, Sen himself) as a species of libertarianism. His condition stipulates that there exist certain choices about which the social ranking should reflect that of a particular individual, regardless of other considerations, including effects on the utility of others. This conception and Sen's analysis thereof is well illustrated by considering his much-discussed example. One individual, whom we shall call Prude, abhors erotic literature, and a second, Lewd, adores it. Both individuals' preferences, moreover, are assumed to be meddlesome in the following manner. Prude would be more upset by Lewd's reading a certain lascivious novel than reading it himself, and Lewd would get more pleasure from Prude's reading the novel than reading it herself. Therefore, as between just Prude reading the novel and just Lewd reading it, both prefer the former. However, Sen's liberal principle insists that the latter be the social choice: Prude's preference against his own reading of the book, ceteris paribus, dictates socially that Prude should not read the book, and likewise Lewd's desire that she read the book, ceteris paribus, dictates socially that Lewd should read it. Hence, the choice that Sen's liberal principle deems socially best is one that would be rejected under the Pareto principle.

Analytically, Sen's result can be understood by reference to the familiar concept of externalities. Lewd's reading the book involves a negative externality on Prude, whereas Prude's reading the book involves a positive externality on Lewd. (Compare the case in which Lewd moderately enjoys loud parties that greatly annoy his neighbour Prude, and Prude would rather not bother to replace his weed-ridden garden with flowers that would greatly delight his neighbour Lewd.) Failing to regulate externalities obviously may violate the Pareto criterion. Furthermore, in Sen's example, the two individuals – if left to themselves – would wish to enter a Coasian bargain under which Prude, rather than Lewd, reads the book (just as, in the variation, Lewd should agree to refrain from loud parties if Prude agrees to replace his weeds with flowers). Sen's principle implicitly prohibits both government regulation and private exchange in which individuals mutually relinquish their posited liberal rights. Preventing mutual waiver both by vote and by contract may hardly seem liberal, as argued by Gibbard (1974) and many others in a highly elaborated literature, surveyed by Suzumura (forthcoming). Indeed, any notion that conflicts with the Pareto principle must embody an underlying opposition to freedom since a violation of the Pareto principle entails contravention of unanimous choice. Some of Sen's subsequent writing (for example, 1992, pp. 144–6) defends his original liberal principle on grounds of practicality and concern for governmental abuse of power. As explored below in the final section, however, such Millian (Mill, 1859) justifications for rights may be powerful but are not, at root, inconsistent with the Pareto principle.

Conflict between Pareto principle and all non-welfarist principles

Sen showed that one particular formulation of a libertarian principle, which carries the implication that

externalities of a sort may not be regulated, violates the Pareto principle. Subsequently, it has been asked more broadly which notions of right, justice, and fairness conflict with the Pareto principle. The answer, it turns out, is that essentially all such notions do, as long as they do not depend exclusively on individuals' utilities – that is, unless they are a reformulation of welfarism.

To state the matter more precisely, we can contrast the individualistic SWF introduced previously, $W(U_1(x), \ldots, U_n(x))$, which by construction depends only on individuals' utilities, with the more generalized SWF, $Z(x)$, which also may be written as $Z(U_1(x), \ldots, U_n(x), x)$. Under the latter, social welfare may depend on anything and, in particular, need not depend exclusively on how the pertinent state x affects individuals' utilities. For example, notions of merit or desert concern whether certain actions or attributes are rewarded, principles of corrective or retributive justice demand that specific norm violations be followed by compensation or punishment, and so forth. Under each of these non-welfarist criteria, knowing each individual's utility in state x is insufficient information to form a social judgement.

Kaplow and Shavell (2001) prove that, if an SWF is not individualistic, then it violates the Pareto principle, if one makes a certain continuity assumption. The assumption is not that the SWF is continuous in all respects. (It is allowed, for example, that infinitesimal violation of some right might cause a discrete reduction in social welfare.) Rather, it is assumed that there exists some good that, if all individuals are given more of it, ceteris paribus (for example, holding rights violations constant), all will have a higher utility and, moreover, the value of the SWF changes continuously as the amount of that good is changed.

The proof is roughly as follows. First, if the SWF does not depend only on individuals' utilities, there must exist two states that are evaluated differently despite everyone's utilities being the same. That is, the non-welfarist SWF is supposed, in at least one instance, to rank states differently on account of a non-welfare difference. Now, taking whichever of the two states ranks lower, we can increase slightly everyone's allotment of the aforementioned good. By continuity, if that increase is sufficiently small, the lower-ranking state must still be ranked lower. However, since all individuals had equal levels of utility in the two initial states, every individual in the modified state now has greater utility, making it Pareto preferred despite the fact that the posited non-welfarist SWF ranks it lower. Hence, the Pareto principle is violated.

One way to understand the conflict between the Pareto principle and all non-welfarist principles is to reflect on the fact that a non-welfarist SWF by definition gives some weight in some instances to a factor independent of its effect on individuals' utilities. We can compare a state that is preferred on account of this non-utility factor to a state that is otherwise identical except that all individuals are slightly better off with respect to some commodity. In other words, a non-welfarist SWF, by its nature, sometimes sacrifices welfare, and nothing in logic rules out the possibility that the welfare sacrifice is borne pro rata.

Subsequent work has generalized and extended this theorem. Campbell and Kelly's (2002) survey notes that the proof in Kaplow and Shavell (2001) does not require the SWF to be a function rather than a binary relation; that this relation need not be fully transitive, only acyclic; and that only lower continuity is required. In a different vein, Suzumura (forthcoming) derives a sort of converse, namely, given Pareto indifference (if everyone is indifferent then society is indifferent – a principle implied by welfarism), social choice must respect the weak Pareto principle (the version defined at the outset of this entry) as well as the strong Pareto principle (if everyone weakly prefers one alternative and at least one individual

strictly prefers it, then it is socially preferred). This theorem requires two additional assumptions: positive responsiveness of the social decision to individual preferences, and that, ceteris paribus, any utility level for an individual can be reached by adjusting the amount of a particular divisible good received by that individual.

Kaplow and Shavell (2002) also offer a complementary demonstration of the conflict between all non-welfarist principles and the Pareto principle. If one restricts attention to symmetric settings – those in which all individuals are identically situated – then any non-welfarist principle conflicts with the Pareto principle in every instance in which its ranking differs from a purely welfarist one. Because everyone is affected identically, it must be that, whenever any amount of aggregate welfare is sacrificed, each and every individual's welfare is sacrificed. The significance of this result is that many traditions favour assessing principles for guiding society in hypothetical situations that, because they are designed to create an impartial perspective, have a symmetric character. Consider, for example, the original position of Rawls (1971) – with important prior formulations thereof by Harsanyi (1953) and others – in which individuals are taken to have no knowledge of their own characteristics. Likewise, the injunctions of the Golden Rule and, relatedly, of Kant's (1785) categorical imperative demand, in essence, that one examine rules as if both positive and negative consequences were borne symmetrically by all. Since, as noted, all choices in symmetric settings involve strict Pareto rankings (except in cases in which all are indifferent), admitting a non-welfarist principle entails the view that the socially preferred state is systematically one in which everyone is worse off.

Perspectives on the conflict

The Pareto criterion is a bedrock principle. Yet it conflicts with all non-welfarist principles – whether they pertain to rights, justice, or fairness – and some of these principles have apparent appeal. How may this tension be reconciled? That the Pareto principle should be seen as paramount is suggested by the rhetorical question: To whom is one doing right, providing justice, or being fair if every possible beneficiary is thereby made worse off•? Additionally, as Sidgwick (1907) and others have queried, if something like utility does not underlie rights and related concepts, by what criterion is the proper list of rights determined in the first instance, and how in principle should the inevitable conflicts between different rights be resolved? A possible reconciliation is suggested by lines of thinking that trace their roots to prominent political economists of a prior era (among others), as more recently elaborated in Kaplow and Shavell (2002).

The relationship between the Pareto principle and other seemingly appealing principles can be understood by reference to what are known as two-level moral theories. (Act utilitarianism versus rule utilitarianism comes to mind, although that somewhat problematic distinction is subtly different from the one under consideration.) As suggested by Hume (1751), Mill (1861), and Sidgwick (1907), one can envision a first-level principle (such as utility) that provides our ideal assessment of states (corresponding to an SWF) and also numerous second-level principles (for example, that one should keep promises, tell the truth, not kill others) that are used as guides by individuals in their everyday conduct. Subsequent prominent statements of this view include Harrod (1936), Rawls (1955), and, most extensively, Hare (1981).

Put in a more explicit optimizing framework, the first-level principle serves as the objective function,

and possible second-level principles constitute the universe of feasible policies. This feasible set is assumed to be constrained by limits of human nature and human institutions. Accordingly, the optimal scheme – taken here to consist of the optimal subset of second-level principles – will be only second best. The aforementioned limits render any attempt at direct implementation of the first-best criterion – commanding that everyone in their individual or institutional capacity act always so as to maximize social welfare – inferior to employment of second-best principles that, inevitably, deviate from the first-best criterion (welfare) in some instances. Two sets of rationales for this conception of the social maximization problem have been offered.

The first sort of justification is based on decision-making costs, complexity, limited information, limited self-control (for example, myopia), and so forth. Such considerations imply that all manner of behaviour, including some types that have no interpersonal effects, should be guided by rules. Moreover, given the nature of the problems that such rules are designed to address, it is inevitable that the rules will not require performance of a complete social welfare calculus and hence will sometimes command behaviour that differs from the first-best outcome. This conflict hardly makes the first-best principle any less of an ideal, just one that is not perfectly achievable in practice.

Second, the nature of human motivation, particularly the problem of cabining self-interest, provides another reason that sensible individual and institutional commands sometimes deviate from a pure concern for individuals' utilities, and thus offers another account of the conflict between the Pareto principle (viewed here as an aspect of the first-level social objective) and alluring non-welfarist principles (understood as second-level rules). Emphasized by Hume, Mill, and Sidgwick, and also by Smith (1790) and Darwin (1874), this strand of thinking is rooted in what may be called moral psychology. As a consequence of biological and social evolution, human emotions may help to channel behaviour in a positive fashion. Opportunism – whether through cheating, theft, or aggression – may be constrained by the prospect of guilt feelings or social disapprobation. Cooperation may be encouraged by anticipated positive internal sentiments or praise by others. Two familiar examples are the retributive urge, the prospect of which may deter aggression, and the desire for social approval, which may inhibit opportunism and encourage constructive collaboration. Given the limitations of biological evolution (limits on altruism as well as the tendency of evolved mechanisms to be specialized), constraints on social inculcation (including the fact that much is directed at young children), and the factors mentioned with regard to the first rationale for second-level rules, it is unsurprising that the resulting precepts sometimes deviate from the first best. Once again, this gap does not call into question the supremacy of the first-best ideal as a matter of principle. (Interestingly, however, this second explanation suggests that emotional force will be associated with moral criteria – various notions of what is right, just, or fair – that conflict with the Pareto principle, which helps explain why our intuitions may be in tension with pure welfarism in some settings.)

Both of these enduring strands of thought that help to reconcile the conflict between the Pareto principle and non-welfarist notions are related to the more recent upsurge of interest at the intersection of economics and psychology, often under the rubric of behavioural economics. Just as Tversky and Kahneman (1974) have stimulated research on heuristics and biases in a range of economic settings, Baron (1993) and others have documented similar phenomena – such as overgeneralization – in individuals' moral thinking. Likewise, many researchers, including Frank (1988) – following intervening provocative statements by Darwin (1874) and Wilson (1975) – have reinvigorated Smith's

interest in human emotions as forces that guide human behaviour, although not always in an ideal manner.

The foregoing discussion suggests that, in regulating individuals' behaviour, various normative criteria that conflict with the Pareto principle may nevertheless usefully advance welfare and thus, at root, be consistent with the underlying force for that principle. These non-welfarist notions may also be relevant to the promotion of welfare for other, related reasons. As argued at length by Bentham (1822–23) in his constitutional writings and Mill (1859) in *On Liberty*, second-best rules obviously may play an important role in constraining government officials. In addition, since many of the non-welfarist criteria exist because of their relationship with the promotion of welfare, they may be useful proxy standards in some settings. Finally, due to the affective aspect of many non-welfarist principles, a complete welfarist account would incorporate them because they are in part constitutive of individuals' utilities. Note that, in each instance, because the relevance of non-welfarist criteria lies in the advancement of welfare, there is no conceptual inconsistency with the ultimate motivation for the Pareto principle even though the non-welfarist second-level rules on their face deviate from the posited first-level ideal.

In sum, a complete understanding of the relationship between the Pareto principle and other, possibly competing normative principles involves many dimensions. Formal analysis of these principles reveals the existence of an underlying, logical conflict. Examination of literatures in other fields of economics and in other disciplines, however, suggests a fundamental harmony.

See Also

- Arrow, Kenneth Joseph
- behavioural economics and game theory
- interpersonal utility comparisons
- philosophy and economics
- Sen, Amartya
- social norms

I am grateful to Steven Shavell for comments and the John M. Olin Center for Law, Economics, and Business at Harvard University for financial support.

Bibliography

Arrow, K. 1951. *Social Choice and Individual Values*. New York: Wiley.

Baron, J. 1993. *Morality and Rational Choice*. Boston: Kluwer Academic Publishers.

Bentham, J. 1822–23. *Securities Against Misrule and Other Constitutional Writings for Tripoli and Greece*, ed. P. Schofield. Oxford: Oxford University Press, 1990.

Campbell, D. and Kelly, J. 2002. Impossibility theorems in the Arrovian framework. In *Handbook of*

Social Choice and Welfare, vol. 1, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: Elsevier Science.

Darwin, C. 1874. *The Descent of Man; and Selection in Relation to Sex*, 2nd edn. Amherst, NY: Prometheus Books, 1998.

Frank, R. 1988. *Passions within Reason*. New York: W.W. Norton & Co.

Gibbard, A. 1974. A Pareto consistent libertarian claim. *Journal of Economic Theory* 7, 388–410.

Hare, R.M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.

Harrod, R. 1936. Utilitarianism revised. *Mind* 45, 137–56.

Harsanyi, J. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61, 434–35.

Hume, D. 1751. *An Enquiry Concerning the Principles of Morals*, ed. T. Beauchamp. Oxford: Oxford University Press, 1998.

Kant, I. 1785. *Groundwork of the Metaphysics of Morals*, trans. and ed. M. Gregor. Cambridge: Cambridge University Press, 1997.

Kaplow, L. and Shavell, S. 2001. Any non-welfarist method of policy assessment violates the Pareto principle. *Journal of Political Economy* 109, 281–6.

Kaplow, L. and Shavell, S. 2002. *Fairness versus Welfare*. Cambridge, MA: Harvard University Press.

Mill, J. S. 1859. *On Liberty*. London: J.W. Parker.

Mill, J. S. 1861. *Utilitarianism*, ed. R. Crisp. Oxford: Oxford University Press, 1998.

Pareto, V. 1906. *Manuale D'Economia Politica*. Milan: Società Editrice Libraria.

Rawls, J. 1955. Two concepts of rules. *Philosophical Review* 64, 3–32.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Robbins, L. 1935. *An Essay on the Nature and Significance of Economic Science*, 2nd edn. London: Macmillan.

Robbins, L. 1938. Interpersonal comparisons of utility: a comment. *Economic Journal* 48, 635–41.

Sen, A. 1970. The impossibility of a Paretian liberal. *Journal of Political Economy* 78, 152–7.

Sen, A. 1992. Minimal liberty. *Economica* 59, 139–59.

Sidgwick, H. 1907. *The Methods of Ethics*, 7th edn. Indianapolis: Hackett Publishing Company, 1981.

Smith, A. 1790. *The Theory of the Moral Sentiments*, 6th edn. Oxford: Oxford University Press, 1976.

Suzumura, K.. Welfarism, individual rights, and procedural fairness. In *Handbook of Social Choice and Welfare*, vol. 2, ed. K. Arrow, A. Sen and K. Suzumura. Amsterdam: Elsevier Science (forthcoming).

Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 195, 1124–31.

Wilson, E. O. 1975. *Sociobiology: The New Synthesis*. Cambridge, MA: Harvard University Press.

How to cite this article

Kaplow, Louis. "Pareto principle and competing principles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000315> doi:10.1057/9780230226203.1250

The New Palgrave Dictionary of Economics Online

Pareto, Vilfredo (1848 – 1923)

A.P. Kirman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Pareto made major contributions to a wide range of subjects covering mathematical economics, statistics, sociology and many others. In economics his name is mainly associated with general equilibrium, welfare economics and ordinal utility. Yet he insisted on the need to confront economic theories with empirical data as his work on income distribution shows. Furthermore, he was far from convinced of the rationality of individual economic behavior. Yet these aspects of his work have been put one side and he is now regarded essentially as the forerunner of the axiomatic school which reached its zenith in the Arrow–Debreu model. This is paradoxical for the latter has been shown to provide no empirically falsifiable propositions.

Keywords

adaptive expectations; Arrow–Debreu model of general equilibrium; Barone, E.; cardinal and ordinal utility; central limit theorems; compensation principle; competitive equilibrium; convexity; economic cycles; Edgeworth box; Edgeworth, F.Y.; equivalent surplus; experimental methods in economics; falsificationism; fixed point theorems; free trade; general equilibrium; homo economicus; hyperplanes; imperfect competition; income distribution; inequality; infant-industry protection; integrability; marginal utility of money; Marshall, A.; mathematics and economics; monopolistic competition; multiple equilibria; New Welfare Economics; non-tâtonnement process; ophelimity; ordinal utility; Pantaleoni, M.; Pareto efficiency; Pareto optimality; Pareto, V.; Pareto's law; partial equilibrium; perfect competition; positive economics; positivism; preferences; rationing; social optimum; social welfare function; socialism; surplus; tâtonnement; transitivity; utility; utility measurement; Walras, L.; Walras's Law

Article

Vilfredo Pareto's name is one of the most familiar in economics, with the universal use of ‘Pareto optimality’ and the Pareto distribution. Yet in 1968 Allais said, in his biography of Pareto, ‘His influence on the development of economics as a science was felt only after considerable delay and has largely been confined to Italy and France’. There are several explanations for this neglect. First, as Chipman (1976) in his admirable survey of Pareto's contributions suggested, it has become less and less fashionable to cite early scholars. Second, Pareto spread his net wide and his persistent desire for empirical verification led him to explore the sociology of human behaviour. This part of his work had some success among sociologists, but interest in it has waned. Third, among theoretical economists Pareto's lasting contribution has come to be regarded as that which helped the evolution of economic theory on the path from the contributions of his predecessor Walras to the full axiomatic formulation of the Arrow–Debreu model. Yet this was but a small part of his overall contribution and of interest to an elite of theorists. Finally, Pareto has been wrongly described as the originator of some ideas and yet has not been credited with certain ideas which he did originate.

All of this explains why his reputation in economics was limited for a long time and has fluctuated considerably since his death in 1923.

Vilfredo Pareto was born in Paris in 1848 where his father, a follower of Mazzini, had exiled himself for political reasons and returned to Italy when Pareto was four years old. His family moved to Florence in 1862. In 1864 Vilfredo Pareto finished school at the early age of 16 and entered the University of Turin where he studied mathematics. He then went on to do a doctorate in engineering and his thesis was entitled, ‘Principi fondamentali della teoria della elasticita dei corpi solidi e ricirche fondamentali sulla integrazione delle equazione diffenziali che ne differiscono l’equilibrio’. The fact that it was on the equilibrium of a physical system is, of course, significant for his later economic work.

Pareto then took up a post as an engineer for a railway company in Florence and continued in this work for three years. In Florence he became a member of the Accademia deo Georgofili and during this period he wrote a number of pieces in economics on such varied topics as the comparison of the advantages of publicly and privately owned railway systems, the merits of proportional representation and the state of the Italian industrial system. He was an ardent campaigner against any form of state interference with the market system and was one of the founders of the Adam Smith Society in Ferrara. As a consequence of their activity he developed a network of contacts in the fields of economics and politics. In 1880 and 1882 he was a candidate for member of parliament as an exponent of free trade, but was unsuccessful. In 1875 he was appointed as technical director of an ironworks in Florence. When raising the capital necessary to modernize the plant he travelled widely in Europe meeting bankers and financiers.

Through his intellectual activities, he established a relationship with Pantaleoni, who later held a chair at the University of Geneva and who introduced him to Walras in Switzerland in 1891. Pantaleoni then, in 1892, recommended him as a worthy successor to Walras at Lausanne when it became clear that the latter's health would not allow him to continue teaching. (See Walras, 1965, vol. 2, p. 455, letter no. 1015.) He took up the chair in Lausanne in 1893 and his vision of the state of the field at that time is illustrated by his inaugural lecture in which he said.

> Nous ne connaissons la théorie d’aucun phénomène naturel dans tous ses details; nous connaissons seulement des théories des phénomènes ideaux, qui se rapprochent plus ou moins du phénomène concret.

After eight years at the university during which he produced his first major contributions to economics and was involved both in the activities of the university and of the canton de Vaud, he bought a villa near Geneva at Celigny. Here he progressively isolated himself from the university although, despite various health problems, he continued to teach in Lausanne till 1911. (Despite the fact that Schumpeter, 1949, describes him as having had a ‘vigorous and fertile old age’, it is clear from his correspondence with the Dean of the Faculty at the university that Pareto was constantly preoccupied by health problems, in particular a heart ailment. See the letters from the archives of the law faculty cited by Biaudet, 1975.)

The university organized a ‘jubilee’ in his honour in 1917 which was attended by a large number of distinguished social scientists and included an official delegation from the Italian government. The latter did something to offset Pareto's feeling that his own country had treated him badly.

While it is customarily asserted that Pareto did not start his work in economics until he reached the age of 45, he would obviously not have been offered a chair with no justification as to his ability in that subject. In fact, for well over ten years prior to his nomination in Lausanne Pareto had been interested in and had contributed to economics.

From the outset, Pareto was preoccupied by the idea of the economy as a complete system and by the interaction between the various sectors of the economy. In this, he was completely in line with the approach developed by Walras and far from the predominantly partial equilibrium analysis of his English contemporary, Marshall. What he was interested in was providing rigorous but parsimonious models of individual economic behaviour and then constructing from these a model of the economy as a whole. He was interested in the ‘points of rest’ of his system and their welfare characteristics.

It was his background as engineer and mathematician that led him to adopt a formal approach to the subject and his frustration with his inability to explain empirical facts that later led him to extend his analysis to sociology. This last phase of his career should not be construed as a disillusionment with the mathematical approach but rather as an attempt to include the other phenomena which he thought might account for the failure of economics to explain empirical facts. In particular he sought to include in his analysis the idea that people could make, from an economic point of view, ‘irrational choices’. In so doing he anticipated the modern ‘cognitive’ approach to economics by a century. His overall aim was therefore to broaden his analysis and eventually to construct a system of laws capable of describing the behaviour of society as a whole, an enterprise that Schumpeter (1949) dismissed as a ‘complete delusion’.

As has been observed his work received rather little attention for a long while. Allais (1968) lamented the failure of the economics profession to recognize the pioneering work of Pareto on the idea of social surplus. (This reflected, in part, his frustration with the lack of recognition of his own work in this area, which was later to be corrected by the award of the Nobel prize.) Georgescu-Rogen (1975) said, ‘There is no denying that Pareto's own ideas met with an incredible lack of attention from most economists during his life as well as for many years after.’ Hicks (1932) wondered why economists had been so hesitant to study Pareto's work and suggested that it was perhaps ‘the sheer impressiveness of his achievements’ that discouraged them. However, later Hicks (1975) himself, observed that the origins of Pareto's contributions on the social optimum form could be traced to Edgeworth rather than to Pareto himself. Yet Malinvaud (1993) asserted that Pareto is now unanimously regarded as one of the founders

of the Arrow–Debreu approach to theory. To see how this re-evaluation has taken place, we need to examine Pareto's general approach to economics, certain of his specific contributions and his relationship to the work of his contemporaries and his predecessors.

It is worth observing at the outset that, while he condemned literary economists out of hand and professed to be interested only by a strictly scientific approach to economics, he, nevertheless, frequently made normative judgements and indulged in casual empiricism. This was, in part, a reflection of a world in which academics would feel much freer to express themselves on a wide variety of subjects without the many constraints that govern academic publication today.

Yet, despite a period in the desert, perhaps due to his having formed almost no students, some of Pareto's main contributions have come to be recognized as having had a profound and lasting impact on economics. The three contributions to economics which have best stood the test of time are the *Cours d'économie politique* (1897), the *Manuale d'economia politica* (1906) and his article 'Economie mathématique' in *L'Encyclopédie des Sciences Mathématiques* (1911). In addition to these one has to mention his articles, in particular those collected and published later as *Marxisme et économie pure* in the *Oeuvres complètes* (1964–84) together with the *Trattato di sociologia generale* (1916) which, with *Les systèmes socialistes* (1901c) includes a substantial body of economic analysis.

The Cours

Of these contributions, the first, the *Cours*, originally published in two volumes, contains an exposition of economic theory illustrated with numerous empirical facts. The theory is presented in a more precise and refined way than that of his intellectual predecessor Walras and the emphasis, in the theoretical analysis is consistently and unequivocally on the interdependence of economic phenomena and the idea of general equilibrium. However, it should be noted that only 75 of the 800 pages are devoted to pure theory and that there is very little that is completely original. At least in Bousquet's (1928a) eyes the theory was better presented than in Walras's *Eléments* (1900) where he suggests the exposition is so tedious as to deter anyone from reading it. Nevertheless, the organization of the *Cours* is curious and, as Cirillo (1979) remarks, it is odd that production is treated after banking and social evolution. Of course, given its title and Pareto's new responsibilities it is not surprising that it gives the strong impression of having been assembled from course notes, and that the order and content of these left something to be desired. (Pareto's teaching cannot have been quite as discouraging as some have suggested since he wound up with 56 students in 1893 as opposed to the six who attended Walras's last courses.) It is also somewhat odd that, given Pareto's strong feelings about the importance of a positivist approach to economics he periodically indulges in direct pleading for the 'liberal' cause. It should, however, be remembered that, while so much of the material that Pareto discusses is now standard and has been refined by successive generations of economists, much of it was new, recent, or even original for him and it should be judged in context. What is remarkable is that Pareto, although one of the founders (if not *the* founder) of the school which culminated in the Arrow–Debreu model, did not hesitate to move beyond the strictly theoretical framework. He included empirical observations and examples of economic phenomena for which he was able to develop little satisfactory theory. Much of the statistical material in the *Cours* had, according to Pantaleoni (1924), been gathered while Pareto was a businessman, and this fact may have some bearing on its presentation. However, the *Cours* did also

contain the first material on income distribution, a subject which will be examined below in more detail. This material provoked a great deal of comment and criticism from such authorities as Edgeworth (1925), and the latter was some 20 years later to comment ironically both on the universality of the law and on the character of its proponent.

The Manuel

The *Manuel* marks, as Ingrao and Israel (1990) suggest, a watershed between Pareto's involvement in economics and his move into sociology. (The reader seeking a detailed and rigorous account of Pareto's main theoretical contributions in the *Manuel* need look no further than Malinvaud, 1993.) It illustrates the coexistence of philosophical reflection, empirical observations and rigorous analysis in Pareto's work. It explicitly acknowledges errors in the *Cours*, in particular that of taking too dogmatic a position in favour of free trade. While still accepting the theoretical arguments in favour of the free trade position, Pareto now had doubts as to the practical value of these arguments. He was no longer convinced that ‘homo oeconomicus’ was useful as anything other that a theoretical construction. He goes even further and even suggests that theoretical economics have not had, from a practical point of view, ‘any great utility so far’ (pp. vi–viii). This did not lead him, however, to abandon theory and he continued to develop his ‘successive approximations’ approach which he thought would lead from a highly abstract theory to a closer approximation of reality.

Of this book, it is the last section, the ‘Mathematical Appendix’ (in the French edition, 1909, pp. 538–671) which has come to be thought of as Pareto's basic contribution to the theory of general equilibrium and to what we now call ‘Pareto optimality’ but which he referred to as ‘The maximum of society's ophelimity’. (This appendix was considerably modified and rewritten for the French edition, in large part as a result of Volterra’s, 1906, comments on the Italian edition.) Although the appendix with its formal analysis is the most widely cited part of the *Manuel*, it makes up less than a quarter of the volume. The rest of the work gives an insight into the more general view that Pareto had of economics. The first two chapters give his views on the scientific status of the social sciences. He argued strongly that in economics and in the social sciences in general, there were underlying laws and structures which had to be determined, specified and tested by scientific methods. However, he, himself, was later to become more and more frustrated with the failure of economic theory to explain empirical facts. The third chapter provides an introduction to the idea of equilibrium which he now saw as a sort of balancing between ‘tastes’ and ‘obstacles.’ In this he had moved on from a more static concept as envisaged in the *Cours* and thought of a situation in which the ‘obstacles’ reacted to the ‘tastes’ and sought a resting point for these competing forces. (He does seem to have thought of his own approach, even in the *Cours*, vol. 1, p. 18, as being more dynamic than that of Walras, but it is difficult to see it as other than the solution of a static set of equations.) In Chapters 4 and 5, he then separates consumption and production and reduces the individual's problem, in each, to one of constrained optimization. Consumers follow paths of increasing utility until they are brought to a halt by the resistance of the ‘obstacles’. Producers seek profit but face technological constraints. (These do not, in current terms, define a convex set. The introduction of fixed costs which cause this complication did, however, allow Pareto to reconcile zero profits and profit maximization.) In the sixth chapter, he then brings the markets together to talk of the general equilibrium of the system and to discuss its efficiency properties. Indeed the ‘first theorem of

welfare economics' makes its first clear appearance here. Two things are worth noting in passing. Pareto was well aware that the important thing for the individual was to maximize subject to the constraints that he perceives, which might or might not be the ones he faces in reality. Second, he systematically considered the possibility of monopolistic competition in parallel with that of perfect competition, although it is this treatment of the latter that is remembered today. In the case of imperfect competition the 'obstacles' change as the result of the individual's actions.

The chapters following those on economic equilibrium deal with a variety economic problems. Pareto deals with demographic problems and their consequences for the labour force but does not pursue the analysis of resultant effects on the labour market. He uses many factual illustrations, in dealing with these problems as well of those of natural resources, in particular land. He also treats capital and savings and the theory of the interest rate. The latter reveals an interesting discrepancy between Pareto's treatment of consumption and that of capital. When dealing with capital he clearly expressed the idea that goods are dated and that the rate of interest can therefore be deduced from the differences in successive prices. Yet when dealing with consumption the notion that the dates at which goods are available is part of their definition is much less clear. He then goes on to deal with monetary problems, but is clear that money can only be introduced once the general equilibrium problem has been fully analysed. All the latter problems are dealt with summarily and consist often of observations based on specific facts. The last chapter is devoted to 'concrete economic phenomena', although these already figure largely in the two preceding chapters.

Other contributions by Pareto

There is little point in simply cataloguing Pareto's numerous other contributions but there are one or two specific items which cast light on the evolution of his thought. One aspect of his work that has been lost from sight is that on international trade, and yet his two articles on that subject had a clear influence on the major figures in the field. Ohlin (1924) went as far as to say that had he read Pareto earlier he would have saved himself a great deal of time and effort. Haberler (1965) said, 'But the only important theoretical advance has been the application, notably by Pareto, of general equilibrium analysis to the problems of international trade.' In particular, in marked contrast to the standard view that he was an unalloyed free trader, he used a sort of 'infant industry' argument in favour of protectionism for this, he thought, would lead to the emergence of a vigorous and productive class which would by its activities lead to a long-run gain which would more than offset the short-term loss from the absence of free trade. (This argument may be found in Chapter 9 of the *Manuel* and in the *Trattato*.) This, of course, was strongly related to his sociological theories, particularly that concerning the 'circulation des elites'.

A second and interesting aspect of Pareto's work was his concern with the origin and nature of economic cycles and crises. This does not fit well in a framework which is essentially static but he had the clear idea as early as the *Cours* that there was a certain overshooting in individuals' adaptation of their expectations. Thus he thought that people move too far from optimism to pessimism and that this leads to the sort of cyclical behaviour we observe in economies. In this he was arguing for a vision different from that of the 'rational expectations' school of today and rather more in line with the idea of 'adaptive expectations' but with coefficients which lead to over-adaptation.

Pareto examined socialism as a system for allocating resources in the *Systèmes socialistes* (1901c). Pareto saw socialism on the one hand as a threat to private property with its desire to extend the role of

the state to the detriment of individual liberty but on the other as a force for change in society. He was, for example, not convinced by the redistributive goal of socialism. This was coloured by his own work on income distribution and the constancy that he thought he had found under many different institutional arrangements. The relation of this work to that of Barone is interesting. Pareto saw a role for a ‘Ministry of Production’ as a way of overcoming the difficulty of fixed costs in production. He envisaged a system of taxes on consumers to cover these costs, and goods would then be sold at cost price. This, he maintained, would restore efficiency despite the non-convexities present in the system. The shift in Pareto's ideas away from the purely liberal view towards a more subtle view of the role of the state is intimately linked to the evolution of his thoughts on sociology. He had what has been described as a ‘clientelist’ view of the organization of society. In his view, the mechanics of government intervention are governed by its need to satisfy its clients and not, as in the collectivist state view, to allocate resources efficiently and equitably.

This analysis together with Pareto's later emphasis on the non-rational, from an economic point of view, elements of choice led to his considering the achievement of any sort of economic efficiency, through a market system alone, as a utopian dream. He was, therefore, highly critical of those who insisted on applying economic theory without taking the whole political and social system into account.

From the economics point of view these contributions are not regarded as major and it is those that are regarded as of lasting importance that will be reviewed in the remainder of this contribution.

Ordinal utility, measurable utility and the integrability problem

One of Pareto's major contributions has long been considered as that of establishing that an ordinal notion of utility is sufficient for the construction of equilibrium theory. The importance of this step for the development of modern theory was not immediately recognized by Pareto. It was only with his article (1900) that he started to develop a fully ordinal theory. Whether this led Pareto actually to reject the idea of ‘measurable utility’ is an interesting question. One suggestion is that he still adhered to the idea of some ‘true measure’ of utility but that he thought that is was simply impossible to identify the appropriate function. In fact, there is no logical contradiction between the observation that ordinality suffices to establish equilibrium and the idea that utility has some cardinal sense. Indeed, although he was very clear that any one of the ‘indices’ of utility would suffice for his analysis, he stated that ‘In certain cases they would permit knowledge of the value of ophelimity’ (*Manuel*, Mathematical Appendix).

Before returning to the measurability problem let us first examine how Pareto developed his ordinal approach. In the *Manuel*, he explicitly contrasts his analysis of ‘indifference curves’ which are constructed without reference to any utility function to that of Edgeworth who started with ‘ophelimity’ or ‘utility’ and obtained expressions for the indifference curves (*Manuel*, p. 540, n. 1).

Pareto proceeds as follows. In the case of two goods, consider x and y the quantities consumed of those goods and consider the quantities dx and dy which, when added to x and y, leave the consumers’ satisfaction unchanged. This gives an equation:

$$f_1[x, y]\,dx + f_2[x, y]\,dy = 0.$$

This equation for the ‘indifference line’ gives us the expression: $dF[x, y]=0$ which is satisfied by an infinite number of F, and Volterra (1906) remarked that ‘Ophelimity is one of these functions F'.’ (Whether this remark should be interpreted as meaning that any of these functions would serve as a utility function or whether the idea was that amongst these functions was the ‘true utility function’ is an interesting question.) In any event, as Volterra pointed out, Pareto's treatment required further work if it was to be extended to the case of three or more goods. The problem here is a simple one. The equations that Pareto wrote down define the tangent hyperplanes to the true indifference surfaces at each point in the consumption space, and what Volterra observed was that there may be no utility function compatible with these equations. This ‘integrability problem’ has preoccupied theorists until recently, although Allais (1973) dismisses it as a red herring.

Pareto's well-known (1906b) paper on ‘Ophelimity in non-closed cycles’ has generally been regarded as an attempt to solve this problem. There are many interpretations of this but one is that it revealed that convexity can replace transitivity. (However, Chipman, 1971, suggests that Pareto's real aim was to give a full treatment of the measurable utility problem and Malinvaud, 1993, regards it as a straightforward attempt to deal with the problem of the transitivity of preferences which in turn is directly linked to the problem of the existence of a utility function. Without entering into the details it is worth observing that Sonnenschein, 1971, showed that the assumption of the convexity of preferences and hence of the quasi-concavity of the utility function could be substituted for transitivity. Thus transitivity of preferences is not a necessary condition for equilibrium analysis. Pareto himself, as Malinvaud, 1993, indicates, glimpsed the importance of convexity and discusses it in Section 4 of the *Manuel* and in Sections 44–50 of the mathematical appendix.) Indeed convexity of preferences is, in a certain sense, a natural condition. This is simply because a great deal of theoretical work in economics boils down to looking at the solution of the maximization of a concave or quasi-concave function over a convex set, and for this most economists have settled for an examination of the first-order conditions for such a maximum. Here the function in question is the utility function and the convex set is provided by the budget constraint or, in Pareto's terminology, the ‘obstacles’. Given this, by now standard view, it is easy to understand why Pareto's 1906 contribution seems so convoluted and makes such difficult reading.

One of the problems is that there is, as in several of Pareto's works, a preoccupation with the order in which goods are consumed which, to modern eyes, confuses the discussion. In modern general equilibrium theory, all goods are dated and hence changing the order of consumption changes the specifications of the bundle of goods in question. Indeed, Pareto, unlike the other economists of his time, did, at some points in his analysis, explicitly adopt the idea of dating goods in order to reduce a temporal problem to a static one. But he also devotes considerable time to discussing ‘paths of consumption’. (Detailed analysis of Pareto's analysis of the ‘order of consumption’ problem can be found in Chipman, 1971, and Malinvaud, 1993.) Thus, individuals move along a path improving their welfare until they arrive at the best bundle given their constraints. Was this how individuals actually behave as he originally indicated or did he, as in his 1906 paper, regard consumption paths as only involving time in an abstract and ‘virtual’ sense.

Pareto also considered explicitly in this context the measurability of utility. As it happens, as Chipman (1971) points out, in the specific case discussed by Pareto, utility is independent of the path of

consumption. In this case, consider the marginal utility (‘elementary ophelimity’) of a good as dependent only on the quantity consumed of that good. (The type of result Pareto was aiming at is closely related to another case already treated by Fisher, 1896.) The utility function obtained in this case is measurable, that is, it is invariant up to a linear transformation. Pareto's work here may thus best be regarded as an early attack on the problem of separable utility functions. Given this, it seems that, while Pareto recognized that equilibrium analysis could be carried out using only ordinal utility, he still had not reached the point of abandoning measurable utility as a concept, and still attached importance to the idea that the difference between the utility obtained in two situations had some significance. Thus, perhaps paradoxically to modern eyes, having liberated the theory of general equilibrium from the notion of cardinal utility Pareto continued to concern himself with the idea of measurable and comparable utilities.

The integrability problem mentioned above, that of recovering an underlying utility function from demand behaviour, was already solved in large part by Antonelli (1886) in a paper which Pareto had in his possession, albeit briefly, since he commented on it in a letter to Pantaleoni in 1891 (letter no. 39 in Pareto, 1960). However, Pareto does not seem to have attached much importance to Antonelli's work, although Walras had already praised it. Thus, for some reason, Pareto did not profit from what had already been done and did not make any significant contribution in this particular direction despite assertions in the literature to the contrary. This may have been because of the brief acquaintance that he had with Antonelli's contribution, or may rather be, as Chipman maintains, because he was less concerned with this problem than with that of the measurability of utility.

To conclude the discussion of this part of his contributions, it is worth emphasizing that Pareto arrived at conditions for economic equilibrium using preferences alone and thus clearly marked out the trail for modern economic theory, but that he did not abandon his interest in the nature of utility and its measurement, and indeed it was as a result of this dual preoccupation that he adopted the curious term ‘ophelimity’.

Lastly, it is worth mentioning that Hutchison (1953) asserted that Pareto had anticipated Slutsky's income and substitution effects. This would indeed have been a major achievement, but, as C. Weber (2002) has shown, a careful reading of the appropriate passage shows that this is not the case. This is yet another example of the inappropriate attributions involving Pareto's work.

General equilibrium

There can be few who have studied general equilibrium theory without using the famous Edgeworth box. This graphical trick which was for a period, unjustifiably called the Edgeworth–Bowley box, actually first appears in its modern form in Pareto (*Manuel*, 1906, p. 355). This was used by Pareto to motivate his ‘proofs’ of the welfare theorems in the general case. The box is the simplest case of what was Pareto's constant preoccupation, namely, the economy as a complete system. He can be thought of as regarding the economy as one market rather than as many individual markets which could be studied, as in the Marshallian tradition, separately. He wrote down what are now considered to be the standard equilibrium conditions for the consumer side of economy, showing the equality of the marginal rate of substitution to the price ratio, normalizing the price of money, which he assumed to give direct utility, to one (*Manuel*, Mathematical Appendix). (Paradoxically, as Hildenbrand, 1994, points out, these

conditions are absent from the final culmination of Pareto's theoretical work, the Arrow–Debreu model, and Debreu dismisses the use of calculus and confines himself until much later to convex sets and separating hyperplanes.)

Taking these equations:

$$U_x = \frac{U_y}{P_y} = \frac{U_z}{U_z} \ldots$$

together with Walras's Law:

$$(x - x_0) + P_y(y - y_0) + P(z - z_0) \ldots = 0$$

and differentiating, he found expressions for

$$\frac{\partial y}{\partial P_y} \cdot \frac{\partial z}{\partial P_z} \text{ and so on.}$$

These he had already set out in his 1892 article in the *Giornale degli Economisti*. He then shows that if goods are independent, that is

$$U_{xy} = 0x \ \pi y,$$

then the conditions $U_{xx}<0$ $U_{yy}<0$ and so on. imply that demand for each good is a decreasing function of its own price (*Manuel*, Mathematical Appendix, section 53). This is a forerunner of more general but very recent results.

Pareto's introduction of money into the utility function is in the spirit of his time and in so doing he was able to clarify some of Marshall's analysis. Firstly, he showed in the *Manuel* (Mathematical Appendix, section 56) that, in general, the ‘marginal utility of money’ changes with prices. Thus it cannot arbitrarily be assumed to be constant.

Secondly, in *Cours*, section 83, he showed that the idea of estimating consumer surplus as the area under the consumer's demand curve above the exchange price was wrong unless the marginal utility of money

happened to be constant which, as we have just seen, it is not, in general.

Finally, he showed in his Encyclopaedia article (1911, section 23) that, if the elasticity of demand for all goods is constant, then it is unity, and remarked that, since Marshall had not realized that he had to impose this restrictive condition, his analysis was defective.

An important aspect of Pareto's work that seems to have escaped attention is the effort that Pareto made to discuss what has come to be called ‘monopolistic competition’ and its introduction into equilibrium analysis (*Manuel*, chs. 3, 5 and 6). Recognizing that individuals can influence prices and that this should be taken into account led him to try to take explicit account of the demand with which they are faced. This merits two comments. Only recently has the introduction of ‘monopolistic competition’ into general equilibrium models resurfaced, and the article by Negishi (1961), which uses rather arbitrary assumptions, is generally cited as the first example. Thus Pareto was already dealing with a problem which has still not been really satisfactorily treated. The way in which the individuals have an effect on prices is revealed in his treatment of the ‘obstacles’ or constraints. As an individual modifies his demand he might be thought of as moving to the appropriate point on his budget hyperplane, thus as making a linear movement. But what if his displacement itself influences prices? In this case his movement will be nonlinear. Once again one can think of all these movements as being virtual and only the final result counting. Alternatively, one can think of a non-tâtonnement process in which individuals trade until they hit a constraint. In the Negishi style, non-tâtonnement process prices are still called centrally and then traders exchange and terminate before their desired bundle if they are not at the equilibrium prices. In Pareto's analysis trade is pairwise and there is rationing, leading Malinvaud (1993) to suggest that this is an early foretaste of the rationing literature associated with the names of Benassy, Drèze and Malinvaud. Nevertheless, it is clear that Pareto in this respect made a step in the direction of greater realism of the adjustment process.

A difficulty with all of Pareto's analysis is the confusion as to the nature of time. Indeed, he recognizes this himself, when talking of the passage from an initial position to an equilibrium, when he says (*Manuel*, ch. 3, section 171), that the issues he discusses fall into the domain of dynamics rather than statics. The modern convention that the adjustment process to equilibrium is instantaneous and that long-term dynamics consist of the passage from one equilibrium to another is far from that adopted by Pareto in this part of his analysis.

Pareto did not make a clear distinction between the question of existence and the question of stability. He regarded equilibrium as the terminating point of a process and this is brought out in the *Cours* and particularly in the *Manuel* (ch. 3, sections 110–15, for example). As has been suggested, the time taken for this process was not specified but is certainly not regarded, even conventionally, as negligible. Having described the passage or path from the initial position to the final position under assumptions of perfect competition with tâtonnement and pair-wise non-tâtonnement processes, he then considered the monopolistic competition case, but rejected it as too difficult to handle. He did, however, enter into a discussion as to how individuals could push the economy towards a preferred equilibrium, from their point of view, in the case of multiple equilibria, and showed how they would try to manipulate the terms of exchange along this path (*Manuel*, p. 197) and discussed how individuals would benefit from doing this. Furthermore, in the light of this manipulation he suggested that certain equilibria would be stable. Thus Pareto recognized explicitly that stability is a property of a particular process.

Although he wished to consider equilibrium as the resting point of a process, Pareto did not try to show the existence of equilibrium as such except by counting equations and unknowns (*Manuel*, Appendix).

In effect, he said that, since one could find the conditions for equilibrium, an interesting possibility would be to solve explicitly all the equations necessary to determine the equilibrium. However, as he pointed out (*Manuel*, ch. 3, section 217), ‘If we take into account the fabulous number of equations that a population of forty million individuals and several thousand goods would give, it would not be mathematics that would come to the aid of economics, but economics that would come to the aid of mathematics’ since such a system would be beyond human capacity to solve. Thus Pareto assumes from a simple argument the existence of a solution and simply dismisses the practicality of finding it for a large economy even if all the relevant equations were known. (This is in no way surprising since we now know that proving the existence of equilibrium is equivalent to proving the existence of a fixed point and the first fixed point theorem was proved by Brouwer in 1910.) (Pareto would have much appreciated Scarf's computational approach to the finding of equilibrium; Scarf and Hansen, 1973.)

Thus the preoccupation with the formal establishment of equilibrium which was later to dominate mathematical economics was not shared by Pareto. Although relatively rigorous, he failed to specify various assumptions used for his approach, such as differentiability and only considered interior solutions to the maximum problem.

Before leaving Pareto's treatment of equilibrium, it is interesting to note that he was clearly aware of the possibility of multiple equilibria and in his diagrams in the *Manuel* (p. 192), he seems to have realized that ‘in general, the number of equilibria would be odd’, a result proved only very recently. An antecedent for this can be found in Edgeworth (1881), who explicitly talks about several equilibria and the fact that they will ‘alternate’ in terms of stability.

Pareto also suggests, as mentioned, that a collectivist state would be better able to lead its economy to an equilibrium than an economy based on private property. The reasoning given for this is based on Pareto's particular view of production and his introduction of non-convexities. This assertion, given Pareto's natural aversion to state intervention, is also heavily qualified (*Manuel*, ch. 6, sections 58–61). Nevertheless, it is interesting to note the contrast with the view of Pareto as an unqualified liberal.

Efficiency or ‘Pareto optimality’

Of all Pareto's contributions to economics, it is this notion of ‘optimality’ or efficiency that has made the greatest impact.

Yet it was not he who first gave a definition of a situation corresponding to the modern definition. Edgeworth (1881) clearly defined a situation in which the utility of each individual is maximized given the utilities of the others. Although this definition is given in the context of an exchange economy, its extension to more general cases was not difficult.

It was not so much the introduction of the idea but the use that Pareto made of it which makes his contribution important. Thus, although he had read Edgeworth, his definition, which also includes production, is an integral part of his own work.

Pareto defined a notion of surplus or gain, which is what is now referred to as ‘equivalent surplus’, this is, the amount of a given numeraire good which would leave the individual indifferent between his original bundle together with this quantity of numeraire and his original bundle together with some proposed change in all the commodities. At an optimum or efficient point there is no surplus. Put alternatively, one could think of the economy as moving along paths as individuals seek to profit from

the surplus that exists until no further such change is possible. Thus an efficient situation is one in which no feasible change exists which would correspond to a positive surplus. The originality and correctness of Pareto's contribution has been questioned and Samuelson (1947) suggested that it was Barone (1908) who first dealt with this point correctly. In fact, as Allais (1973) points out, Barone acknowledged Pareto's priority and furthermore developed a less adequate, price-dependent, version of surplus.

The real insight that Pareto had was that this notion of efficiency or optimality was independent of all institutional arrangements and of all distributional considerations (*Cours*, vol. 2). Pareto then went on in the *Manuel* (ch. 6 and Mathematical Appendix, sections 145–52), to establish the ‘first theorem of welfare economics’, that a competitive equilibrium is a Pareto optimum and a tentative version of the ‘second theorem’, that any Pareto optimum can be obtained as a competitive equilibrium from an appropriate distribution of initial resources. The latter result is only suggested and is never clearly stated. Furthermore, both results are incomplete and even incorrect as a result of the confusion in the treatment of production. There are also a number of simple errors that creep into the exposition which confuse the argument. To take a simple example at two points in the *Manuel* (Appendix, sections 45 and 89) he says that some people will necessarily be better off and others worse off. Did he here envisage only movements along the efficient surface and therefore rule out changes which would make everyone worse off? This is not clear for at another point (*Manuel* (ch. 6, section 37) he clearly envisages everyone's welfare as declining. The key here is that he makes the correct statement in the case of a finite move but rules out the possibility of everybody being worse off in an infinitesimal move. It was Wicksell (1897) who pointed this out and Chipman (1976) mounted a rigorous defence of Pareto.

Pareto's ideas on the nature of efficiency evolved over time and in the *Trattato* (sections 2128–39), he showed that the maximization of any social welfare function W which was an increasing function of individual utility functions U_i

$$W = F(U_1, U_2, \ldots)$$

whether the U_i were defined over the consumption of all individuals or just restricted to individual consumption gave an optimum. Now as Pareto states (*Trattato*, pp. 1342–3), it is clear that in defining W a government would have to give weights to the different individuals. The idea of including the consumption of other individuals in the utility functions extends the scope of normal economic analysis to what were considered at the time and are still often thought of as ‘sociological’ considerations. Pareto did not observe that by appropriately modifying F all optima could be generated. As Allais (1968) suggests, it is not clear that Pareto was fully aware of the impact of this contribution.

In addition to these contributions to welfare economics Pareto has been credited with the founding of the ‘New Welfare Economics’. In particular, it is argued that in his 1894b article ‘Il massimo di utilita dato dalla libera concorrenza’ he introduced the Hicks–Kaldor compensation principle. However, as Kemp and Pezanis-Christou (1999) point out, Pareto argued that compensation should only be a consideration if it was actually carried out, and not just potentially possible. He spelled out the way in which it could be achieved by transfers between individuals, but did not go as far as saying that situation X is better

than situation Y if transfers could be made from Y that would make everybody better off than in X.

Income distribution, ‘ Pareto's law ’

One of Pareto's major contributions was to propose a ‘law’ governing the distribution of income. Here by distribution of income is meant the distribution of personal income amongst individual economic units and not the distribution of income between factors of production. The latter line was developed by Ricardo, in particular and was, of course, at the centre of the Marxian, neoclassical, and Keynesian debates. Pareto's interest and motivation were very different. On the one hand, it has often been suggested that his work in this area reflected the search for some sort of universal principles underlying economic behaviour. This would not explain why he chose this particular domain. The reason for his initial interest was rather his disagreement with the socialist proposals to undertake institutional reforms to make the distribution of personal income more equal. His initial work was published in an article in the *Gironale degli Economisti* in 1895 then in a memoire (1896b) on the ‘income distribution curve’. Detailed discussion is given in the *Cours* (sections 957–65) and in the *Manuel* (ch. 7, sections 2–31).
In the course of analysing different data, Pareto was led to believe not only that he had established a functional form for income distributions which was essentially independent of institutional considerations but, even more remarkably, that the parameter of that function might well be the same across all countries and thus also independent of institutional arrangements. This would be enough to make any attempts to achieve a significant redistribution of income impossible. It is not surprising, given its social implications, that his contribution has been the source of controversy. Second, this work can be thought of as a pioneering piece of applied econometrics and not therefore as in keeping with the rather abstract image that is often painted of Pareto's work.
Three formulae were proposed by Pareto and the first and most widely cited of these is given by:

$$N(x) = \frac{A}{x^{a}}$$

where $N(x)$ is the number of people having an income greater than or equal to x. As has been frequently pointed out, it has obvious problems where either x tends to zero, or one increases x so that $N(x)$ goes to zero. Pareto's proposed a second form which mitigated these problems and which was to replace x by $x+q$ where q is a constant. Using his original form, Pareto then estimated values for a in particular for data for the UK collected by Giffen (reproduced in Giffen, 1904). He obtained for 1843: a=1.5 and for 1879/80: a=1.35. Further computations for Prussia, Saxony, Paris and several Italian cities gave values around 1.5 with a maximum of 1.73. Pareto denied that his ‘law’ had the status of a physical law, and stated in an article in the *Journal of Political Economy* (1897b) that ‘I should not be greatly surprised if some day, a well authenticated exception were discovered.’ Nevertheless, he believed that the values of a that he found, a itself being a statistic, were sufficiently close for his law to be ‘provisionally accepted as universal’. This statement is not wholly unambiguous since closeness of the estimated parameter values is not an indication that the functional form itself corresponds well to the data. Nevertheless,

Pareto asserted that the values he obtained which he considered to be remarkably close, despite the different origins of the data, could not be attributed to chance.

Pareto was well aware that other functional forms might also fit the data well; for example, he estimated a distribution of the form:

$$N(x) = \frac{A}{(x+a)^{\alpha}} 10^{-\beta x}$$

where *a* and *b* like *a* are constants. He found a value of *b* so low that he concluded that a distribution of the second form that he had proposed

$$N(x) = \frac{A}{(x+a)^{\alpha}}$$

would suffice.

It is worth noting, in passing, that the three forms proposed by Pareto have a number of particular properties. The third form has finite moments for all r whereas this is only true for the first and second forms for r<*a*. When *a*. is less than or equal to 2 the first form has infinite variance and ‘Pareto's law’ is characterized by a fat right tail. In this case both the first and second forms belong to the Pareto-Lévy class of stable distributions. Indeed, Barbut (2000) has pointed out that the reason that distributions of the Pareto type occur so frequently was shown by Paul Lévy (1937). He showed that stable distributions other than the normal exhibit asymptotic behaviour of Pareto's second form with 0<*a*<2. Hence, a central limit theorem of a certain type exists for heavy tailed distributions to which the standard central limit theorem does not apply. Pareto was thus credited with having removed the stranglehold of the Normal distribution!

It has been widely recognized since Pareto's time that other distributions provide more satisfactory fits for particular income data. Nevertheless, ‘Pareto's law’ gives empirically a satisfactory fit for the upper tail of the income distribution (the top 20 per cent according to Lydall, 1968) but is clearly inconsistent with the lower end. This has resulted in a search for distributional forms which are close approximations of the Pareto form for the upper tail.

However, it is not the adequacy of Pareto's income distribution as a description of empirical data that has been controversial, it is rather the relation between ‘Pareto's law’ and the problem of income inequality that has been the subject of dispute. Pareto says that, if the number of individuals with an income over a certain level *x* in relation to the number of those below that level increases, then inequality diminishes (*Manuel*, ch. 7, section 24). Unfortunately, there was a printer's error in the *Cours*, and there the opposite is stated, although from the footnote (*Cours*, Livre III, section 965), it is clear what Pareto intended. There has since been considerable confusion about what Pareto actually said.

Let $N(h)$ be the number of individuals with income above h (the 'minimum income') and $N(x)$ the number above x with $x>h$. Then, as Pareto says, if we define

$$U_x = \frac{N(x)}{N(h)}$$

then, 'income inequality will decrease as U_x increases' (*Cours*, Livre III, section 965; *Manuel*, p. 390, n. 2).

Allais (1968) interprets Pareto as saying the opposite, perhaps following the error in the *Cours*. Yet if we now proceed and assume that 'Pareto's law' holds, then we have:

$$U_x = \left(\frac{h}{x}\right)^{\alpha}.$$

Since $x>h$ by hypothesis, U_x decreases when a increases and income inequality increases. Allais makes an error in his argument and states that:

$$\frac{N(h)}{N(x)} = \left(\frac{h}{x}\right)^{\alpha},$$

an error identical to that made by Roy (1966). Since both Roy and Allais had started from the original mistake in the *Cours*, this further error should have led them to the same final conclusion as Pareto that income inequality varies in the same direction as a. Roy indeed arrives at this conclusion and contrasts it with the work of Gini and others. Allais made a further error and stated that Pareto believed that income inequality varied inversely with a. All this gives some indication of the sort of confusion that has surrounded Pareto's contribution. An explanation of these different interpretations can, however, be found and is that the authors mentioned were working with different basic hypotheses. If two distributions with the same mean income are compared, then the view that inequality increases with a is correct. If, on the other hand, one compares two distributions with the same minimum income, then Pareto's view is the appropriate one.

If a were a constant, then there would be little hope for policies aimed at reducing income inequality, as Pareto pointed out to those in favour of the socialist position. Lastly, Pareto's law has the peculiar feature that the ratio of the average income above x say $m(x)$ to x itself is a constant given by:

$$\frac{m(x)}{x} = \frac{\alpha}{\alpha - 1}.$$

Allais suggested that this might be taken as Pareto's index of inequality. If this were so, then it would decrease with *a*, the opposite of what Pareto intended.

Economics and physics: Pareto's view

What is clear from both Pareto's analysis and that of many of his contemporaries such as Edgeworth, Jevons and Fisher is that they all shared a conviction that there was an analogy between economic systems and those of classical mechanics. Edgeworth (1881) was quite explicit in suggesting that a ‘mécanique sociale’ would take its place alongside the ‘mécanique celeste’. Jevons (1905) said that economics resembles physics in that ‘the equations employed do not differ in general character from those which are really treated in many branches of physical science’. Another contemporary, Cairnes (1875; the citations from Cairnes and Edgeworth are taken from Cohen, 1994) was even more explicit. He asserted that ‘Political Economy is as well entitled to be considered a “positive science” as any of those physical sciences to which this name is commonly applied.’ He went on to argue that the principles of economics have identical features to those ‘of the physical principles which are deduced from the laws of gravitation and motion.’

The validity and consequences of such assertions have been examined at length by Mirowski (1989), Ingrao and Israel (1990) and Cohen (1994). The extent to which the analogy between physics and economics has ensnared economics in a position which it could have avoided had it found its source of inspiration elsewhere – for example, in biology, as Marshall suggested – is well documented by these authors.

Pareto himself made the remark that when examining the equations which have to be solved to determine an economic equilibrium someone well versed in mathematics or physics would say, ‘These equations do not seem new to me, they are old friends. They are the equations of rational mechanics.’ He went so far, in the *Cours*, as to draw up a table of analogies between the two disciplines. What is most interesting about this table is not the analogies themselves, which are, in some cases, inaccurate and misleading, but rather the caveats that are provided by Pareto. He seemed to be well aware, even at his early stage in his writings, of the dangers of taking the analogy too literally, and in this he distinguished himself from a number of his contemporaries. He understood that, when extended to the full social system, the physical analogy was highly tentative. Yet he had no other formal frame of reference within which to model the socio-economic system. This led to his increasingly cautious attitude in using equations from physics, but did not deter him in his goal of modelling the whole social system rigorously. Given the reservations expressed by Pareto it seems unfair to lay the blame for the domination of classical mechanics as a mathematical framework for economics at his door. This only partially absolves him, for his attitude was essentially that physics provided an analogy but those parts of it that were inappropriate could be put to one side. As Mirowski (1989) points out, it is a common error to believe that all parts of the physics metaphor are equally dispensable. This misunderstanding has led,

in part, to the persistence of the metaphor for it seems that we are free to weaken it as much as we wish till it is suitable for our purposes. Had this been recognized as erroneous economists might have strived harder for an alternative metaphor.

Economics and its relationship with the other social sciences

Pareto's vision of the nature of the social sciences is reflected in his works on sociology (in particular the *Trattato*) and a certain number of his positions mark him out from his contemporaries and his successors. He developed and reinforced his idea that such sciences should be positive and went as far as criticizing his earlier work, taking the ‘author’ of the *Cours* to task for mixing ethical and positive considerations (*Manuel*, Preface). His defence of positivism was clearly associated with Comte's position (1830) and he was interested in developing a ‘positive theory of economic policy’. He argued that ‘laws’ or relations deduced from specific assumptions should be tested empirically against ‘observed statistical laws’. He went further, however, and unlike J.S. Mill (1844), who asserted that to verify hypotheses was not part of the business of science, a position supported by Friedman (1953) and Machlup (1955) and others, later argued that assumptions should be examined to see how reasonable they are (*Trattato*, section 59). The importance of Pareto's statistical work which reflected his standpoint has tended to be overlooked and has been dominated by analysis of his purely theoretical contributions. It cannot be repeated often enough that Pareto insisted on what he called the ‘experimental method’ as the *only* appropriate method appropriate for the social sciences and would not countenance wholly theoretical work which could not be empirically tested.

His approach to economics reflected a double position. Firstly, he shared Marshall's opinion that economic theory should be aimed at examining ‘man as he is’ and should not become an abstract intellectual exercise. Secondly, however, while he wished economics to be a relevant science, he condemned attempts to apply too readily economic theory to real problems. He believed that much harm had been done to the cause of ‘scientific economics’ by such hasty applications. This was, he thought, particularly dangerous since economic considerations could not be isolated from more general sociological concerns and to do so would lead to misleading and erroneous conclusions. His preoccupation with the analysis of non-rational behaviour adds force to this view.

Finally, it should be remembered that, while Pareto was with Weber among the first to expound the principles of ‘positive social science’, his view of the status of economics was ambiguous. He believed fundamentally, and in this he shared Comte's view, that there should be a universal scientific approach to social science. Yet he recognized the need for and desirability of specialized disciplines, although he regarded these as building blocks for a general approach. Thus while he was persuaded that certain aspects of economic phenomena are more quantifiable than many social phenomena, he was not prepared to isolate man's economic activity from his other functions. It should clearly be recognized that Pareto himself (Pareto, 1917) considered his *Trattato* as his most important work and that he came to believe that the non-economic component of social phenomena dominates the economic part; and as he said, ‘The most important error of the so-called liberal economist is not to recognise this.’ Thus Pareto was progressively more convinced of the importance of the non-economic in explaining the evolution of society.

Conclusion

Pareto's economic contribution has acquired an increasing reputation over time, unlike his sociological work. Yet, as was suggested at the outset, it is disappointing that this reputation should be constructed on the basis of such a small part of his work. His strictly theoretical contributions are an essential part of modern general equilibrium theory. Yet here his education and training pushed him towards an equilibrium notion close to those of classical mechanics, and in a certain sense he helped to lock economics into an unhappily rigid framework.

Pareto's work covered a wide range of subjects. Within the field of economic theory, not only did he examine the nature and existence of a general equilibrium but he also considered what we would now refer to as the problem of ‘imperfect competition’, that is, the analysis of directly and consciously conflicting interests. Furthermore, his concern with statistical verification and his constant references to the idea that economic theories should be confronted with economic facts as in his examination of the problem of the form of income distributions are all central to an understanding of his contribution. He was preoccupied with the idea that economic theory fails to explain many phenomena, not because the theory itself is inadequate, but rather because that theory is just one part of a larger theoretical structure which should incorporate all social phenomena.

All of this illustrates the richness and diversity of Pareto's work. It is therefore paradoxical that, as Pareto's stature as one of the major figures in the development of economics has grown in recent years, most of this increased recognition has been based on a limited part of his most formal contributions. As has been observed, Pareto came to emphasize more and more the role of the non-economic in explaining social phenomena, yet he has come to be remembered essentially as the forerunner of the axiomatic school of economics where rationality is rigidly imposed. Perhaps the most ironic aspect of the evolution of Pareto's reputation is the current state of general equilibrium theory. While the most refined version of Pareto's theory, the Arrow–Debreu model, has been shown, thanks to the Sonnenschein–Mantel–Debreu results, to provide no empirically falsifiable propositions, Pareto himself was impatient with the idea of purely theoretical models which were not subject to falsification. How would he have reacted to the idea that his reputation was to be essentially based on his contribution to the construction of such a model? How far he was in spirit from the theory that he is claimed to have founded can be understood from a remark he made to a specialist in mathematical logic.

> I cannot admit that there is any rational method which is superior to the experimental method: I do not accept that one can study what should be; I, on the contrary, try to find out what exists in reality. (Pareto, 1964–84, vol. 19, 1027)

Selected works

1892–93. Considerazioni sui principii fondamentali dell’ecconomia politica pura. *Giornale Degli Economisti* 4, 389–420, 485–512; 5, 119–57; 6, 1–37; 7, 279–321.

1893. La mortalità infantile e il costo dell’uomo adulto. *Giornale Degli Economisti* 7, 451–6.

1894a. Teoria matematica dei cambi forestieri. *Giornale Degli Economisti* 8, 142–73.

1894b. Il massimo di utilita dato dalla libera concorrenza. *Giornale degli Economisti* 9, 48–66.

1895. Teoria del commercio internazional. *Giornale Degli Economisti* 9.

1896a. (A proposito di un libro del dottore Fornasari). Il modo di figurare i fenomeno economici. *Giornale Degli Economisti* 12, 75–87.

1896b. La curve della entrate e le osservazioni del professor Edgeworth. *Giornale Degli Economisti* 13, 439–48.

1896–7. *Cours d'économie Politique*, 2 vols. Lausanne: Librairie de l'Université. (Referred to as *Cours* in the text).

1897a. Aggiunta allo studio curva della entrate. *Giornale Degli Economisti* 14, 15–26.

1897b. The new theories of economics. *Journal of Political Economy* 5, 485–502.

1899. Quelques exemples d'application de la méthode de moindres carrés. *Journal de Statistique Suisse* 121–50.

1899. Tables pour faciliter l'application de la méthode de moindres carrés. *Journal de Statistique Suisse* 121–50.

1900. Sunto di alcuni capitoli di un nuovo trattato di economia pura del prof. Pareto. *Giornale Degli Economisti* 20, 216–35, 511–49. In Chipman et al. (1971).

1901a. Sul fenomeno economico. Lettera a Benedetto Croce. *Giornale Degli Economisti* 22, 131–8.

1901b. Le nuovo teorie economiche. Appunti. *Giornale Degli Economisti* 23, 235–52.

1901c. *Les Systèmes socialistes*, reprinted in Pareto (1964–84).

1902. Di un nuovo errore nello interpretare le teorie dell'economia matematica. *Giornale Degli Economisti* 25, 401–33.

1906. *Manuale d'economia Politica*. Milan: Societa Editrice Libraria. (Referred to as *Manuel* in the text). Revised and translated as *Manuel d'économie politique*, Paris: Giard et Brière, 1909. A new very complete and annotated edition of the *Manuale*, edited by A. Montesano A. Zanni and L. Bruni, Milan: Universita of Bocconi Press, was published in 2006.

1906b. L'ofelimità nei cicli non chiusi. *Giornale Degli Economisti* 33, 15–30. English translation in Chipman et al. (1971).

1907–8. L'interpolazione per la ricerca delle leggi economiche. *Giornale Degli Economisti* 34, 266–85; 36, 423–53.

1910. Walras. *Economic Journal* 20, 138–9.

1911. Economie mathématique. In *Encyclopédie des sciences mathématiques*, I (iv. 4), Paris: Teubner, Gauthier, Villars. Trans as 'Mathematical Economics', *International Economic Papers* 5 (1955), 58–102. (Referred to as 'Economie mathématique' in the text.)

1913. Il massimo di utilità per una collettività in sociologia. *Giornale Degli Economisti* 46, 337–8.

1916. *Trattato di Sociologia Generale*, 4 vols. Florence: Barbera. (Referred to as *Trattato* in the text.) Trans. as *The Mind and Society*, ed. A. Livingston, New York: Harcourt Brace & Co., 1935.

1917. Pratica e Teoria. *L'Economista* 8(July), 540–2.

1918. Economia sperimentale. *Giornale Degli Economisti* 52, 1–18.

1960. *Lettere a Maffeo Pantaleoni 1890–1923*, 3 vols. ed. G. De Rosa. Roma: Banca Nazionale Del Lavoro.

All the above works together with a number of other articles, reviews and comments have been collected and published from 1964 on as Pareto, Vilfredo, *Oeuvres complètes*, 28 vols, ed. G. Busino, Geneva: Librairie Droz, 1964–84.

Bibliography

Accademia Nazionale dei Lincei. 1975. *Convegno Internazionale Vilfredo Pareto (Roma, 25–27 ottobre 1973)*. Roma: Accademia Nazionale dei Lincei.

Allais, M. 1968. Pareto, Vilfredo: contributions to economics. In *Encyclopaedia of the Social Sciences*. New York: Macmillan.

Allais, M. 1973. Inequality and civilizations. *Social Science Quarterly*, Special issue for the 50th anniversary of Pareto's death, September.

Amoroso, L. 1938. Vilfredo Pareto. *Econometrica* 6, 1–21.

Antonelli, G.B. 1886. *Sulla Teoria Matematica della Economia Politica*. Pisa: Tipografia del Folchetto. Trans. J.S. Chipman and A.P. Kirman as ‘On the Mathematical Theory of Political Economy’. In Chipman et al. (1971).

Atkinson, A. 1975. *The Economics of Inequality*. Oxford: Oxford University Press.

Barone, E. 1908. Il Ministro della produzione nello stato collettivista. *Giornale Degli Economisti* 37, 267–93, 391–414.

Barbut, M. 2000. Pareto et la statistique. L’homme extrême de Pareto: sa postérité, son universalité. In *Pareto Aujourd’hui*, ed. B. Alain. Paris: PUF.

Benini, R. 1906. *Principi di statistica metodologica*. Torino: Unione Tipografica Editrice.

Biaudet, J-C. 1975. Vilfredo et Lausanne. In Accademia Nazionale dei Lincei (1975).

Borgatta, G. 1924. I rapporti fra la scienza economica e la sociologia nell’opera Paretiana. *Giornale degli Economisti* 39(64), 81–9.

Borkenau, F. 1936. *Pareto*. London: Chapman & Hall.

Bousquet, G.H. 1927. *Introduction aux Systèmes Socialistes de Pareto*. Paris: Giard.

Bousquet, G.H. 1928a. *The Work of Vilfredo Pareto*. Minneapolis: Sociological Press.

Bousquet, G.H. 1928b. *Vilfredo Pareto, sa vie et son oeuvre*. Paris: Payot. In *The Invisible Hand*, ed. B. Ingrao and G. Israel. Cambridge, MA: MIT Press, 1990.

Bousquet, G.H. 1963. Pareto et ses *Systèmes Socialistes. Cahiers de l’ Institut de Science Appliquée* (Supplement no. 134), 25–32.

Busino, G. 1963. Pareto e le autorità di Losanna. *Giornale degli Economisti NS* 22, 260–303.

Cairnes, J.E. 1875. *The Character and Logical Method of Political Economy*. New York: Harper & Brothers.

Cappa, A. 1924. *Vilfredo Pareto*. Turin: Gobetti.

Chipman, J.S. 1971. Introduction to Part II of Chipman et al. (1971).

Chipman, J.S. 1976. The Paretian heritage. *Revue européene des sciences sociales* 14(37), 65–173.

Chipman, J.S., Hurwicz, L., Richter, M.K. and Sonnenschein, H.F., eds. 1971. *Preferences, Utility, and Demand*. New York: Harcourt Brace, Jovanovich.

Cirillo, R. 1979. *The Economics of Vilfredo Pareto*. London: Frank Cass and Co.

Cohen, I.B. 1994. *Interactions*. Cambridge, MA: MIT Press.

Comte, A. 1830. *Cours de Philosophie Positive*, 6 vols. Paris: Schleicher.

Davis, H.T. 1941. *The Analysis of Economic Time Series*. Bloomington, IN: Principia Press.

Demaria, G. 1949. L'oeuvre économique de Vilfredo Pareto. *Revue d'économie politique* 59, 517–44.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Edgeworth, F.Y. 1896. Supplementary notes on statistics. *Journal of the Royal Statistical Society* 2, 533–4.

Edgeworth, F.Y. 1925. *Papers Relating to Political Economy*, 3 vols. London: Macmillan for the Royal Economic Society.

Eisermann, G. 1961. *Vilfredo Pareto als Nationalökonom und Soziologe*. Tübingen: Mohr.

Fisher, I. 1896. (A review of) 'La courbe de la répartition de la richesse', by Vilfredo Pareto. *Yale Review* 5, 325–8.

Fossati, E. 1949. Pareto dans son et notre temps. *Revue d'économie politique* 59, 585–99.

Frechet, M. 1939. Sur les formules de répartition des revenus. *Revue de l'Institut International de Statistique* 7, 32–8.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Georgescu-Roegen, N. 1975. Vilfredo Pareto and his theory of ophelimity. In Accademia Nazionale dei Lincei (1975).

Giacolone-Monaco, T. 1957. *Pareto–Walras da un Carteggio Inedito (1891–1901)*. Padua: Cedam.

Gibrat, R. 1931. *Les Inégalités économiques*. Paris: Sirey.

Gide, C. 1917. Le jubilé Vilfredo Pareto. *Revue d'économie politique* 31, 426–33.

Giffen, R. 1904. *Economic Inquiries and Studies*. London: G. Bell and Sons.

Haberler, G. 1965. *The Theory of International Trade*. London: William Hodge.

Hicks, J.R. 1932. [Marginal productivity and the Lausanne School.] A reply. *Economica* 12, 297–300.

Hicks, J.R. 1975. Pareto and the economic optimum. In Accademia Nazionale dei Lincei (1975).

Hildenbrand, W. 1994. *Market Demand: Theory and Empirical Evidence*. Princeton: Princeton University Press.

Hutchison, T.W. 1953. *A Review of Economic Doctrines, 1870–1929*. Oxford: Clarendon Press.

Ingrao, B. and Israel, G., eds. 1990. *The Invisible Hand*. Cambridge, MA: MIT Press.

Jevons, W.S. 1905. *The Principles of Science: A Treatise on Logic and Scientific Method*, 2nd edn. London: Macmillan.

Johnson, N.O. 1937. The Pareto law. *Review of Economic Statistics* 19, 20–6.

Kemp, M.C. and Pezanis-Christou, P. 1999. Pareto's compensation principle. *Social Choice and Welfare* 16, 441–4.

Kirman, A. 1998. Vilfredo Pareto. In *Italian Economists of the 20th Century*, ed. F. Meacci. Cheltenham, UK–Northampton, MA: Edward Elgar.

Lévy, P. 1937. *Théorie de l'addition des Variables Aléatoires*. Paris: Gauthier-Villars.

Lydall, H.F. 1968. *The Structure of Earnings*. Oxford: Oxford University Press.

Machlup, F. 1955. The problem of verification in economics. *Southern Economic Journal* 22, 1–21.

Machlup, F. 1960. Operational concepts and mental constructs in model and theory formation. *Giornale degli economisti e annali di economia* 19, 553–82.

Machlup, F. 1964. Professor Samuelson on theory and realism. *American Economic Review* 54, 733–6.

Malinvaud, E. 1993. Le Manuel de Pareto et la Théorie Moderne des Prix. *Revue d' Economie Politique*

103, 157–89.

Mandelbrot, B. 1963. New methods in statistical economics. *Journal of Political Economy* 71, 421–40.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan (first published in 1890).

Mill, J.S. 1844. *Essays on Some Unsettled Questions of Political Economy*. London: J.W. Parker.

Mirowski, P. 1989. *More Heat than Light*. Cambridge: Cambridge University Press.

Mortara, G. 1924. Pareto statistico. *Giornale degli Economisti* 64, 120–5.

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28, 196–201.

Ohlin, B.G. 1924. *Handelns Teori*. Stockholm: Centraltryckeriet.

Pantaleoni, M. 1907–1908. *Pure Economics*. New York: Kelley and Macmillan, 1957. Originally published in Italian as ‘Elementi di economia pura’ and translated into English in 1898. *Giornale degli economisti* 34, 266–85; 36, 423–53.

Pantaleoni, M. 1923. Vilfredo Pareto. *Economic Journal* 33, 582–90.

Pantaleoni, M. 1924. In occasione della morte di Pareto reflessioni. *Giornale Degli Economisti* 44, 1–19.

Pirou, G. 1938. *Les Théories de l’équilibre économique: Walras et Pareto*. Rome: Giornale degli Economisti e Rivista de Statistica.

Roy, R. 1966. Preface to V. Pareto, *Statistique et économie mathématique*. Geneva: Librairie Droz.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge: Harvard University Press.

Samuelson, P.A. 1963. Problems of methodology: discussion. *American Economic Review, Papers and Proceedings* 53, 231–6.

Samuelson, P.A. 1964. Theory and realism: a reply. *American Economic Review* 54, 736–9.

Scarf, H.E. and Hansen, T. 1973. *The Computation of Economic Equilibria*. New Haven, CT: Yale University Press.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. New York: Harper & Brothers.

Schumpeter, J. 1949. Vilfredo Pareto (1848–1923). *Quarterly Journal of Economics* 63, 147–73.

Schumpeter, J. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Sonnenschein, H. 1971. Demand theory without transitive preferences: with applications to the theory of competitive equilibrium. In Chipman et al. (1971).

Spengler, J.J. 1944. Pareto on population 1. *Quarterly Journal of Economics* 58, 593–8.

Stark, W. 1963. In search of the true Pareto. *British Journal of Sociology* 14, 103–12.

Volterra, V. 1906. L'Economia matematica e il Nuovo manuale del Prof. Pareto. *Giornale degli economisti* 32, 296–301. Translated by A.P. De Kirman, revised and edited by J.S. Chipman as 'Mathematical Economics and Professor Pareto's New Manual', in Chipman et al. (1971).

Walras, L. 1900. *Eléments d'économie politique pure ou théorie de la richesse sociale*, 4th edn. Paris: Rouge, Lausanne, and Pichon. Last revised edition, Paris, 1926; English translation, *Elements of Pure Economics*, ed. W. Jaffé. London: Allen & Unwin, 1954. In *The Invisible Hand*, ed. B. Ingrao and G. Israel (1990), Cambridge, MA: MIT Press.

Walras, L. 1965. *Correspondence of Léon Walras and Related Papers*, 3 vols. ed. W. Jaffé. Amsterdam: North-Holland.

Weber, C. 2002. Did Pareto discover income and substitution effects? On an interpretation suggested by Hutchison. *Economics Bulletin* 2(2), 1–6.

Weber, M. 1913. *Wirtschaft und Gesellschaft*. Originally printed as the third part of *Grundriss der Sozialökonomik*, 2 vols. Tübingen: Mohr. Reprinted in *Gesammelte Aufsätze zur Wissenschaftslehre*, Tübingen: Mohr, 1922.

Weber, M. 1949. *The Methodology of the Social Sciences*, trans. and E. Shils and H. Finch. Glencoe, III: Free Press.

Wicksell, K. 1897. Vilfredo Pareto, Cours d'économie politique. *Zeitschrift für Volkswirtschaft, Sozialpolitik und Verwaltung*, 159–66. Reprinted in English in *Selected Papers of Knut Wicksell*, ed. E. Lindahl. Cambridge, MA: Harvard University Press, 1958.

How to cite this article

Kirman, A. P. "Pareto, Vilfredo (1848–1923)." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000026> doi:10.1057/9780230226203.1247

The New Palgrave Dictionary of Economics Online

partial identification in econometrics

Charles F. Manski

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Econometricians long thought of identification as a binary event: a parameter is either identified or not. Empirical researchers combined available data with assumptions that yield point identification, and reported point estimates of parameters. Yet there is enormous scope for fruitful inference using weaker and more credible assumptions that partially identify parameters. Until recently, study of partial identification was rare and fragmented. However, a coherent body of research took shape in the 1990s and has grown rapidly. This research has yielded new approaches to inference with missing outcome data, analysis of treatment response, and other important problems of empirical research.

Keywords

counterfactuals; discrete response analysis; errors in variables; Gini coefficient; identification region; Law of Total Probability; nonparametric methods; parametric prediction; partial identification in econometrics; reverse regression; sampling; statistical inference; treatment response

Article

Suppose that one wants to use sample data to draw conclusions about a population of interest. Econometricians have long found it useful to separately study identification problems and problems of statistical inference. Studies of identification characterize the conclusions that could be drawn if one were able to observe an unlimited number of realizations of the sampling process. Studies of statistical inference characterize the generally weaker conclusions that can be drawn given a sample of positive but finite size. Koopmans (1949, p. 132) put it this way in the article that introduced the term ‘identification’:

> In our discussion we have used the phrase ‘a parameter that can be determined from a sufficient number of observations.’ We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from ‘a

> sufficiently large number of observations' we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect.

For most of the 20th century, econometricians commonly thought of identification as a binary event – a parameter is either identified or it is not. Empirical researchers applying econometric methods combined available data with assumptions that yield point identification and they reported point estimates of parameters. Many economists recognized with discomfort that point identification often requires strong assumptions that are difficult to motivate. However, they saw no other way to perform inference.

Yet there is enormous scope for fruitful inference using weaker and more credible assumptions that partially identify population parameters. A parameter is partially identified if the sampling process and maintained assumptions reveal that the parameter lies in a set, its 'identification region', that is smaller than the logical range of the parameter but larger than a single point. Estimates of partially identified parameters generically are set-valued; a natural estimate of an identification region is its sample analog.

Until recently, study of partial identification was rare and fragmented. Frisch (1934) and Reiersol (1941) developed sharp bounds on the slope parameter of a linear regression with errors-in-variables, with refinement by Klepper and Leamer (1984) and others. Duncan and Davis (1953) used a numerical example to show that the ecological inference problem of political science is a matter of partial identification. Cochran, Mosteller and Tukey (1954) suggested conservative analysis of surveys with missing data due to non-response by sample members, although Cochran (1977) subsequently downplayed the idea. Peterson (1976) initiated study of partial identification of the competing risks model of survival analysis.

For whatever reason, these scattered contributions remained at the fringes of econometric consciousness and did not spawn systematic study of partial identification. However, a coherent body of research took shape in the 1990s and has grown rapidly. The new literature on partial identification emerged out of concern with traditional approaches to inference with missing outcome data. Empirical researchers have commonly assumed that missingness is random, in the sense that the observability of an outcome is statistically independent of its value. Yet this and other point-identifying assumptions have regularly been criticized as implausible. So it was natural to ask what random sampling with partial observability of outcomes reveals about outcome distributions if nothing is known about the missingness process or if assumptions weak enough to be widely credible are imposed. This question was posed and partially answered in Manski (1989), with subsequent development in Manski (1994; 2003, chs.1 and 2), Scharfstein, Manski and Anthony (2004), Blundell et al. (2004) and Stoye (2005).

Study of inference with missing outcome data led naturally to consideration of conditional prediction and analysis of treatment response. A common objective of empirical research is to predict an outcome conditional on given covariates, using data from a random sample of the population. Often, sample realizations of outcomes and/or covariates are missing. Horowitz and Manski (1998; 2000) and Zaffalon

(2002) study nonparametric prediction when nothing is known about the missingness process; Horowitz et al. (2003) and Horowitz and Manski (2006) consider the computationally challenging problem of parametric prediction. Missing data on outcomes and covariates is the extreme case of interval measurement of these variables. Manski and Tamer (2002) study conditional prediction with interval data on outcomes or covariates, while Haile and Tamer (2003) analyse an interesting problem of interval data that arises in econometric analysis of auctions.

Analysis of treatment response must contend with the fundamental problem that counterfactual outcomes are not observable; hence, findings on partial identification with missing outcome data are directly applicable. Yet analysis of treatment response poses much more than a generic missing-data problem. One reason is that observations of realized outcomes, when combined with suitable assumptions, can provide information about counterfactual ones. Another is that practical problems of treatment choice as well as other concerns motivate research on treatment response and thereby determine what population parameters are of interest. For these reasons, it has been productive to study partial identification of treatment response as a subject in its own right. This stream of research was initiated independently in Robins (1989) and Manski (1990). Subsequent contributions include Manski (1995; 1997a; 1997b), Balke and Pearl (1997), Heckman, Smith and Clements (1997), Hotz, Mullin and Sanders (1997), Manski and Nagin (1998), Manski and Pepper (2000), Moinari (2002), and Pepper (2003). The normative problem of treatment choice when treatment response is partially identified is studied in Manski (2000; 2002; 2005a; 2005b; 2006) and Brock (2005).

Another broad subject of study has been inference on the components of finite probability mixtures. The mathematical problem of decomposition of mixtures arises in many substantively distinct settings, including contaminated sampling, ecological inference, and conditional prediction with missing or misclassified covariate data. Findings on partial identification of mixtures have application to all of these subjects and more. Research on this subject includes Horowitz and Manski (1995), Bollinger (1996), Cross and Manski (2002), Dominitz and Sherman (2004), Kreider and Pepper (2004), and Molinari (2004).

There has been other research as well. In discrete response analysis, response-based sampling poses a ‘reverse regression’ problem in which one seeks to learn the distribution of outcomes given covariates but the sampling process reveals the distribution of covariates given outcomes. This problem has been studied in Manski (1995, ch. 4; 2001; 2003, ch. 6) and King and Zeng (2002). In econometric analysis of multi-player games, a long-standing problem has been to infer behaviour from outcome data when the game being studied may have multiple equilibria. Ciliberto and Tamer (2004) address this problem.

Whatever the specific subject under study, a common theme runs through the new literature on partial identification. One first asks what the sampling process alone reveals about the population of interest and then studies the identifying power of assumptions that aim to be credible in practice. This conservative approach to inference makes clear the conclusions one can draw in empirical research without imposing untenable assumptions. It establishes a domain of consensus among researchers who may hold disparate beliefs about what assumptions are appropriate. It also makes plain the limitations of the available data. When credible identification regions turn out to be large, researchers should face up to the fact that the available data do not support inferences as tight as they might like to achieve.

The remainder of this article uses the problem of inference with missing outcome data and the analysis

of treatment response to develop the common theme of recent research on partial identification and to give illustrative findings. Readers who aim to learn more may want to begin with two monographs that provide self-contained expositions with different audiences in mind. Manski (1995) presents basic ideas in a way intended to be broadly accessible to students and researchers in the social sciences. Manski (2003) develops the subject in a rigorous manner meant to provide the foundation for further study by econometricians.

Readers who prefer to learn about econometric methods through the study of empirical applications will find diverse case studies using observational data to analyse treatment response. Manski et al. (1992) investigate the effect of family structure on children's outcomes, and Hotz, Mullin and Sanders (1997) analyse the effect of teenage childbearing. Manski and Nagin (1998) study the effects of judicial sentencing on criminal recidivism. Pepper (2000) examines the intergenerational effects of welfare receipt. Manski and Pepper (2000) and Ginther (2002) analyse the returns to schooling.

There have also been empirical studies of problems of partial identification that arise in analysis of randomized experiments. Horowitz and Manski (2000) study a medical clinical trial with missing data on outcomes and covariates. Pepper (2003) asks what welfare-to-work experiments reveal about the operation of welfare policy when case workers have discretion in treatment assignment. Scharfstein, Manski and Anthony (2004) analyse an educational experiment with randomized assignment to treatment but non-random attrition of subjects.

Inference with missing outcome data

To formalize the missing data problem, let each member *j* of a population *J* have an outcome *y*, in a space *Y*. The population is a probability space and y: $J \rightarrow Y$ is a random variable with distribution $P(y)$. Let a sampling process draw persons at random from *J*. However, not all realizations of *y* are observable. Let the realization of a binary random variable *z* indicate observability; *y* is observable if $z = 1$ and not observable if $z = 0$.

By the Law of Total Probability

$$P(y) = P(y|z = 1)P(z = 1) + P(y|z = 0)P(z = 0). \tag{1}$$

The sampling process reveals $P(y|z = 1)$ and $P(z)$, but is uninformative regarding $P(y|z = 0)$. Hence, the sampling process partially identifies $P(y)$. In particular, it reveals that $P(y)$ lies in the identification region

$$H[P(y)] \equiv [P(y|z = 1)P(z = 1) + \gamma P(z = 0),\ \gamma \in \Gamma_\gamma], \tag{2}$$

where Γ_Y is the space of all probability distributions on *Y*.
The size of the identification region *H*[*P*(*y*)] grows with $P(z = 0)$, which measures the prevalence of missing data. The region is a proper subset of Γ_Y whenever the probability of missing data is less than 1, and it is a singleton when there are no missing data. Thus, *P*(*y*) is partially identified when $0 < P(z = 0) < 1$ and is point-identified when $P(z = 0) = 0$.

Means of bounded functions of y

A common objective of empirical research is to infer parameters of a probability distribution. The identification region for a parameter of *P*(*y*) follows immediately from *H*[*P*(*y*)]. Let $\tau(\cdot): \Gamma_Y \to T$ map probability distributions on *Y* into a parameter space *T* and consider inference on the parameter τ [*P*(*y*)]. The identification region consists of all possible values of the parameter. Thus,

$$H\{\tau[P(y)]\} = \{\tau(\eta), \eta \in H[P(y)]\}. \tag{3}$$

Result (3) is simple but is too abstract to be useful as stated. Research on partial identification has sought to characterize $H\{\tau[P(y)]\}$ for different parameters. Manski (1989) does this for means of bounded functions of *y*, Manski (1994) for quantiles, and Manski (2003, ch. 1) for all parameters that respect first-order stochastic dominance. Blundell et al. (2004) and Stoye (2005) characterize the identification regions for spread parameters such as the variance, inter-quartile range, and the Gini coefficient; these authors apply their findings in empirical research assessing nationwide income inequality using surveys with missing income data.
The results for means of bounded functions are easy to derive and instructive, so I focus on these parameters here. Let *R* be the real line. Let *g*(·) be a function that maps *Y* into *R* and that attains finite lower and upper bounds $g_0 \equiv \min_{y \in Y} g(y)$ and $g_1 \equiv \max_{y \in Y} g(y)$. The problem of interest is to infer *E*[*g*(*y*)].
The Law of Iterated Expectations gives

$$E[g(y)] = E[g(y)|z = 1]P(z = 1) + E[g(y)|z = 0]P(z = 0). \tag{4}$$

The sampling process reveals $E[g(y)|z = 1$ and *P*(*z*), but is uninformative regarding $E[g(y)|z = 0$, which can take any value in the interval [g_0, g_1]. Hence, the identification region for *E*[*g*(*y*)] is the closed interval

$$H\{E[g(y)]\} = [E[g(y)|z=1]P(z=1) + g_0P(z=0), E[g(y)|z=1]P(z=1) + g_1P(z=0)]. \quad (5)$$

$H\{E[g(y)]\}$ is a proper subset of $[g_0, g_1]$ whenever $P(z=0)$ is less than one. The width of the region is $(g_1 - g_0)P(z=0)$. Thus, the severity of the identification problem varies directly with the prevalence of missing data.

Result (5) has many applications. Perhaps the most far-reaching is the identification region it implies for the probability that y lies in any non-empty, proper set $B \subset Y$. Let $g_B(\cdot)$ be the indicator function $g_B(y) \equiv 1[y \in B]$; that is, $g_B(y)=1$ if $y \in B$ and $g_B(y)=0$ otherwise. Then $g_B(\cdot)$ attains its lower and upper bounds on Y, these being 0 and 1. Moreover, $E[g_B(y)]=P(y \in B)$ and $E[g_B(y)|z=1]=P(y \in B|z=1)$. Hence,

$$H[P(y \in B)] = [P(y \in B|z=1)P(z=1), P(y \in B|z=1)P(z=1) + P(z=0)]. \quad (6)$$

Observe that the width $P(z=0)$ of this interval depends only on the prevalence of missing data, not on the form of set B.

When y is real-valued, result (6) immediately yields the identification region for the distribution function of y. Given any $r \in R$, it follows from (6) that

$$H[P(y \le r)] = [P(y \le r|z=1)P(z=1), P(y \le r|z=1)P(z=1) + P(z=0)]. \quad (7)$$

The feasible distribution functions are all increasing functions $F(\cdot)$ such that $F(r) \in H[P(y \le r)]$ for all $r \in R$.

To go further still, result (7) may be used to obtain sharp bounds on quantiles of y, by inverting the bounds on the distribution function. Manski (1994) and Manski (2003, ch. 1) give alternative derivations of the results for quantiles.

Distributional assumptions

Distributional assumptions may enable one to shrink identification regions obtained using the empirical evidence alone. One type of assumption asserts that the distribution $P(y|z=0)$ of missing outcomes lies

in some set $\Gamma_{0Y} \subset \Gamma_Y$. Then the identification region shrinks from $H[P(y)]$ to

$$H_1[P(y)] \equiv [P(y|z=1)P(z=1) + \gamma(P(z=0)),\ \gamma \in \Gamma_{0Y}]. \tag{8}$$

Assumptions of this type are not refutable; after all, the empirical evidence reveals nothing about $P(y|z=0)$. A leading example is the assumption that data are missing at random. Formally, this is the assumption that $P(y|z=0) = P(y|z=1)$, which implies that $H_1[P(y)]$ contains the single distribution $P(y|z=1)$.

A different type of assumption asserts that the distribution of interest, $P(y)$, lies in a set $\Gamma_{0Y} \subset \Gamma_Y$. Then the identification region shrinks from $H[P(y)]$ to

$$H_1[P(y)] \equiv \Gamma_{0Y} \cap H[P(y)]. \tag{9}$$

Assumptions of the latter type may be refutable: if the intersection of Γ_{0Y} and $H[P(y)]$ should be empty, then $P(y)$ cannot lie in Γ_{0Y}. For example, let y be real-valued and consider the assumption that $P(y)$ is a symmetric distribution. Then $H_1[P(y)]$ is composed of all members of $H[P(y)]$ that are symmetric. If $H[P(y)]$ contains no symmetric distributions, the empirical evidence reveals that $P(y)$ is not symmetric.

Statistical inference

The fundamental problem posed by missing data is identification, so it has been convenient in the above discussion to suppose that one knows the distributions that are asymptotically revealed by the sampling process, namely, $P(y|z=1)$ and $P(z)$. An empirical researcher observing a sample of finite size N must contend with issues of statistical inference as well as identification. I shall not dwell on these here, but merely point out that the empirical distributions $P_N(y|z=1)$ and $P_N(z)$ almost surely converge to $P(y|z=1)$ and $P(z)$ respectively. Hence, a consistent estimate of the identification region $H[P(y)]$ is its sample analog

$$H_N[P(y)] \equiv [P_N(y|z=1)P_N(z=1) + \gamma P_N(z=0),\ \gamma \in \Gamma_Y]. \tag{10}$$

Moreover, a natural estimate of the identification region for a parameter τ is $\{\tau(\eta), \eta \in H_N[P(y)]\}$. Sample analogs may also be used in the presence of distributional assumptions.
Confidence intervals (CIs) may be constructed to measure the sampling variation in estimates of identification regions. Considering cases in which the identification region is an interval on the real line, Horowitz and Manski (2000) propose CIs that asymptotically cover the entire region with fixed probability. Chernozhukov, Hong and Tamer (2004) develop methods for construction of such CIs when the identification region is a general finite-dimensional set. Imbens and Manski (2004) develop a conceptually different confidence interval; rather than cover the entire identification region with fixed probability, their interval asymptotically covers the true value of the parameter with this probability.

Analysis of treatment response

Analysis of treatment response poses a pervasive and distinctive problem of missing outcomes. Studies of treatment response aim to predict the outcomes that would occur if different treatment rules were applied to a population. Treatments are mutually exclusive, so one cannot observe the outcomes that a person would experience under all treatments. At most, one can observe the outcome that a person experiences under the treatment he actually receives. The counterfactual outcomes that a person would have experienced under other treatments are logically unobservable.
For example, suppose that patients ill with a specified disease can be treated by drugs or by surgery. The relevant outcome might be lifespan. One may want to predict the lifespans that would occur if all patients were to be treated by drugs. The available data may be observations of the actual lifespans of patients in a study population, some of whom were treated by drugs and the rest by surgery.
To formalize the inferential problem, let each member j of a study population J have a response function $y_j(\cdot)$: $T \rightarrow Y$ mapping the mutually exclusive and exhaustive treatments $t \in T$ into outcomes $y_j(t) \in Y$. Let $z_j \in T$ denote the treatment that person j receives and $y_j \equiv y_j(z_j)$ be the outcome that he experiences. Then $y_j(t)$, $t \neq z_j$ are counterfactual outcomes.
Let $y(\cdot): J \rightarrow Y^{|T|}$ be the random variable mapping the population into their response functions. Let $z: J \rightarrow T$ be the 'status quo treatment rule' mapping the members of J into the treatments that they actually receive. Response functions are not observable, but realized treatments and outcomes may be observable. If so, random sampling from J reveals the status quo (outcome, treatment) distribution $P(y, z)$.

The selection problem

Analysis of treatment response seeks to predict the outcomes that would occur under alternatives to the status quo treatment rule. A leading objective is to predict the outcomes that would occur if all persons were to receive the same treatment. By definition, $P[y(t)]$ is the distribution of outcomes that would occur if all persons were to receive a specified treatment t. Hence prediction of outcomes under a rule mandating uniform treatment requires inference on $P[y(t)]$. The problem of identification of this distribution from knowledge of $P(y, z)$ is commonly called the 'selection problem'.
The selection problem has the same structure as the missing-outcomes problem discussed above. To see

this, write

$$P[y(t)] = P[y(t)|z = t]P(z = t) + P[y(t)|z \neq t]P(z \neq t) = P(y|z = t)P(z = t) + P[y(t)|z \neq t]P(z \neq t). \quad (11)$$

The first equality is the Law of Total Probability. The second holds because $y(t)$ is the outcome experienced by persons who receive treatment t. The sampling process reveals $P(y|z = t)$, $P(z = t)$, and $P(z \neq t)$, but it is uninformative about $P[y(t)|z \neq t]$. Hence, the identification region for $P[y(t)]$ if we use the empirical evidence alone is

$$H\{P[y(t)]\} = \{P(y|z = t)P(z = t) + \gamma P(z \neq t), \gamma \in \Gamma_\gamma\}. \quad (12)$$

This identification region has the same form as the region (2) for inference on outcomes with missing data, with $P(z \neq t)$ being the probability of missing data. Hence, all of the analysis of missing outcomes discussed above applies here as well.

Distributional assumptions

A familiar ‘solution’ to the selection problem is to assume that the status quo treatment rule makes realized treatments statistically independent of response functions; that is,

$$P[y(\cdot)] = P[y(\cdot)|z]. \quad (13)$$

This assumption implies that $P[y(t)] = P(y|z = t)$. The sampling process reveals $P(y|z = t)$. Hence, assumption (13) point-identifies $P[y(t)]$.

Assumption (13) is credible when the status quo treatment rule calls for random assignment of treatments and all persons comply with their assignments. Indeed, the fact that (13) holds is the reason why randomized experiments are held in high esteem. However, the credibility of the assumption in settings without random assignment or full compliance almost invariably is a matter of controversy. This motivates interest in other assumptions that may be better motivated in practice.

There has been much study of assumptions that use an ‘instrumental variable’; that is, an observable covariate whose value varies across the study population. Suppose that outcomes are real-valued.

Manski (1990) poses the mean-independence assumption $E[y(t)] = E[y(t)|v]$. If outcomes are bounded with values normalized to lie in the unit interval, the resulting identification region for $E[y(t)]$ is

$$H\{E[y(t)]\} = [\max_{v \in V} E\{y \cdot 1[z = t] | v = v\}, \min_{v \in V} E\{y \cdot 1[z = t] + 1[z \neq t] | v = v\}]. \quad (14)$$

Manski and Pepper (2000) study identification of $E[y(t)]$ when v is real-valued and the assumption of mean independence is weakened to state that $E[y(t)|v]$ weakly increases in v. Heckman and Vytlacil (2001) combine the mean-independence assumption with some of the structure of an econometric selection model and show that the identification region for $E[y(t)]$ remains (14).

Statistical independence assumptions are stronger than mean independence. Manski (2003, ch. 7) poses the assumption $P[y(t)] = P[y(t)|v]$ and shows that it yields this identification region for $P[y(t)]$:

$$H\{P[y(t)]\} = \bigcap_{v \in V} \{P(y|v = v, z = t)P(z = t|v = v) + \gamma_v \cdot P(z \neq t|v = v), \gamma_v \in \Gamma_y\}. \quad (15)$$

Balke and Pearl (1997) poses the yet stronger assumption $P[y(\cdot)] = P[y(\cdot)|v]$ and characterize its identifying power when outcomes are binary variables.

A different idea, developed in Manski (1995, ch. 6; 1997a) is to place assumptions on the shape of the response functions $y(\cdot)$. One may sometimes believe that treatment response is monotone, in the sense that outcomes increase with the intensity of the treatment. When the set T of treatments is ordered in terms of degree of intensity, the assumption of ‘monotone treatment response’ asserts that, for all persons j and all treatment pairs (s, t), $t \geq s \Rightarrow y_j(t) \geq y_j(s)$. If outcomes are bounded with values normalized to lie in the unit interval, the resulting identification region for $E[y(t)]$ is the interval

$$H\{E[y(t)]\} = [E(y|t \geq z) \cdot P(t \geq z), P(t > z) + E(y|t \leq z) \cdot P(t \leq z)]. \quad (16)$$

A narrower interval results if treatment response is assumed to be concave as well as monotone. Shape restrictions on the response function and assumptions using instrumental variables illustrate the vast middle ground between inference from the empirical evidence alone and analysis predicated on assumptions that are strong enough to achieve point identification. As the study of partial identification continues to broaden and deepen, empirical researchers will be able to choose from a growing menu of

inferential options. One should, however, not expect one uniformly best option to emerge. The appeal of any approach to inference necessarily depends on the objectives of the research, the available data, and the assumptions that are credible to maintain.

See Also

- nonparametric structural models
- statistics and economics
- treatment effect

Bibliography

Balke, A. and Pearl, J. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92, 1171–7.

Blundell, R., Gosling, A., Ichimura, H. and Meghir, C. 2004. Changes in the distribution of male and female wages accounting for employment composition using bounds. Working Paper W04/25. London: Institute of Fiscal Studies.

Bollinger, C. 1996. Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics* 73, 387–99.

Brock, W. 2005. Profiling problems with partially identified structure. Mimeo. Madison, WI: Department of Economics, University of Wisconsin-Madison.

Chernozhukov, V., Hong, H. and Tamer, E. 2004. Parameter set inference in a class of econometric models. Mimeo. Evanston, IL: Department of Economics, Northwestern University.

Ciliberto, F. and Tamer, E. 2004. Evanston, IL: Market structure and multiple equilibria in airline markets. Mimeo. Evanston, IL: Department of Economics, Northwestern University.

Cochran, W. 1977. *Sampling Techniques*, 3rd edn. New York: Wiley.

Cochran, W., Mosteller, F. and Tukey, J. 1954. *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male*. Washington, DC: American Statistical Association.

Cross, P. and Manski, C. 2002. Regressions, short and long. *Econometrica* 70, 357–68.

Dominitz, J. and Sherman, R. 2004. Sharp bounds under contaminated or corrupted sampling with verification, with an application to environmental pollutant data. *Journal of Agricultural, Biological, and Environmental Statistics* 9, 319–38.

Duncan, O. and Davis, B. 1953. An alternative to ecological correlation. *American Sociological Review* 18, 665–6.

Frisch, R. 1934. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo, Norway: University Institute for Economics.

Ginther, D. 2002. Alternative estimates of the effect of schooling on earnings. *Review of Economics and Statistics* 82, 103–16.

Haile, P. and Tamer, E. 2003. Inference with an incomplete model of English auctions. *Journal of Political Economy* 111, 1–51.

Heckman, J., Smith, J. and Clements, N. 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Review of Economic Studies* 64, 487–535.

Heckman, J. and Vytlacil, E. 2001. Localinstrumental variables. In *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune and J. Powell. Cambridge: Cambridge University Press.

Horowitz, J. and Manski, C. 1995. Identification and robustness with contaminated and corrupted data. *Econometrica* 63, 281–302.

Horowitz, J. and Manski, C. 1998. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *Journal of Econometrics* 84, 37–58.

Horowitz, J. and Manski, C. 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95, 77–84.

Horowitz, J. and Manski, C. 2006. Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics* 132(2), 445–59.

Horowitz, J., Manski, C., Ponomareva, M. and Stoye, J. 2003. Computation of bounds on population parameters when the data are incomplete. *Reliable Computing* 9, 419–40.

Hotz, J., Mullin, C. and Sanders, S. 1997. Bounding causal effects using data from a contaminated natural experiment: analyzing the effects of teenage childbearing. *Review of Economic Studies* 64, 575–603.

Imbens, G. and Manski, C. 2004. Confidence intervals for partially identified parameters. *Econometrica*

72, 1845–57.

King, G. and Zeng, L. 2002. Estimating risk and rate levels, ratios and differences in case-control studies. *Statistics in Medicine* 21, 1409–27.

Klepper, S. and Leamer, E. 1984. Consistent sets of estimates for regressions with errors in all variables. *Econometrica* 52, 163–83.

Koopmans, T. 1949. Identification problems in economic model construction. *Econometrica* 17, 125–44.

Kreider, B. and Pepper, J. 2004. Disability and employment: reevaluating the evidence in light of reporting errors. Mimeo. Ames, IA: Department of Economics, Iowa State University.

Manski, C. 1989. Anatomy of the selection problem. *Journal of Human Resources* 24, 343–60.

Manski, C. 1990. Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings* 80, 319–23.

Manski, C. 1994. The selection problem. In *Advances in Econometrics, Sixth World Congress*, ed. C. Sims. Cambridge: Cambridge University Press.

Manski, C. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C. 1997a. Monotone treatment response. *Econometrica* 65, 1311–34.

Manski, C. 1997b. The mixing problem in programme evaluation. *Review of Economic Studies* 64, 537–53.

Manski, C. 2000. Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics* 95, 415–42.

Manski, C. 2001. Nonparametric identification under response-based sampling. In *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune and J. Powell. New York: Cambridge University Press.

Manski, C. 2002. Treatment choice under ambiguity induced by inferential problems. *Journal of Statistical Planning and Inference* 105, 67–82.

Manski, C. 2003. *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C. 2005a. *Social Choice with Partial Knowledge of Treatment Response*. Princeton: Princeton University Press.

Manski, C. 2005b. Search profiling with partial knowledge of treatment response. Mimeo. Evanston, IL: Department of Economics, Northwestern University.

Manski, C. 2006. Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics* f139(1), 105–15.

Manski, C. and Nagin, D. 1998. Bounding disagreements about treatment effects: a case study of sentencing and recidivism. *Sociological Methodology* 28, 99–137.

Manski, C. and Pepper, J. (2000). Monotone instrumental variables: with an application to the returns to schooling. *Econometrica* 68, 997–1010.

Manski, C., Sandefur, G., McLanahan, S. and Powers, D. 1992. Alternative estimates of the effect of family structure during adolescence on high school graduation. *Journal of the American Statistical Association* 87, 25–37.

Manski, C. and Tamer, E. 2002. Inference on regressions with interval data on a regressor or outcome. *Econometrica* 70, 519–46.

Moinari, F. 2002. Missing treatments. Mimeo. Ithaca, NY: Department of Economics, Cornell University.

Molinari, F. 2004. Partial identification of probability distributions with misclassified data. Mimeo. Ithaca, NY: Department of Economics, Cornell University.

Pepper, J. 2000. The intergenerational transmission of welfare receipt: a nonparametric bounds analysis. *Review of Economics and Statistics* 82, 472–88.

Pepper, J. 2003. Using experiments to evaluate performance standards: what do welfare-to-work demonstrations reveal to welfare reformers? *Journal of Human Resources* 38, 860–80.

Peterson, A. 1976. Bounds for a joint distribution function with fixed subdistribution functions: application to competing risks. *Proceedings of the National Academy of Sciences U.S.A.* 73, 11–13.

Reiersol, O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9, 1–24.

Robins, J. 1989. The analysis of randomized and non-randomized AIDS treatment trials using a new

approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, ed. L. Sechrest, H. Freeman and A. Mulley. Washington, DC: NCHSR, US Public Health Service.

Scharfstein, J., Manski, C. and Anthony, J. 2004. On the construction of bounds in prospective studies with missing ordinal outcomes: application to the good behavior game trial. *Biometrics* 60, 154–64.

Stoye, J. 2005. Partial identification of spread parameters when some data are missing. Mimeo. Evanston, IL: Department of Economics, Northwestern University.

Zaffalon, M. 2002. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference* 105, 105–22.

How to cite this article

Manski, Charles F. "partial identification in econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000312> doi:10.1057/9780230226203.1251

The New Palgrave Dictionary of Economics Online

partial linear model

Elie Tamer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

It is popular to summarize the relationship between an outcome variable *y* and a vector (*x*, *z*) through a linear mean regression where the mean of *y* is modelled as a linear function of both *x* and *z*. A more robust specification is called for in some situations where the imposed linear relationship between (the mean of) *y* and *z* is suspect. A partially linear specification allows for a regression function that maintains linearity in *x* but allows the effect of *z* to be nonlinear. This partially linear model has been widely studied in the statistics and the semiparametric econometrics literature.

Keywords

age elasticity; bootstrap; censored selection models; conditional expectations; convergence; cross validation; heteroskedasticity; homoskedasticity; income elasticity; kernel estimators; linear models; linear regression; nonparametric estimation; nonparametric selection models; partially linear models; quantile regression; random variables; semiparametric estimation; semiparametric sieve least squares; sieves; spline regression

Article

A partially linear model requires the regression function to be a linear function of a subset of the variables and a nonparametric non-specified function of the rest of the variables. Suppose, for example, that one is interested in estimating the relationship between an outcome variable of interest *y* and a vector of variables (*x*, *z*). The economist is comfortable modelling the regression function as linear in *x*, but s hesitant in extending the linearity to *z*. One example, considered by Engle et al. (1986), is the effect of temperature on fuel consumption using a time series of cities. To do that, one can consider a regression of average fuel consumption in time *t* on average household characteristic and average temperature in time *t*. The analyst might be more comfortable with imposing linearity on the part of the regression function involving household characteristics but unwilling to require that fuel consumption

varies linearly with temperature. This is natural since fuel consumption tends to be higher at extremes of the temperature scale, but lower at moderate temperatures. The regression function Engle et al. consider is:

$$y = x'\beta + g(z) + u \tag{1}$$

where *x* denotes a vector of household/city characteristics and *z* is temperature and *u* is a mean zero random variable such that is independent of (*x*, *z*). The function g(.) is unspecified except for smoothness assumptions. They term this *the semiparametric regression model*.

Another example is the demand for gasoline model used by Schmalensee and Stoker (1999). The primary interest in this paper is the age income structure of household demand for fuel. In particular, the authors want to estimate demand elasticities of age and income (do richer household consume more gasoline, and, if so, by how much?). Hence, their dependent variable is the logarithm of gasoline consumed and *g*(.) is a function of both age and the logarithm of income. Schmalensee and Stoker also control for a set of other household characteristics. This partially linear model allows one to have a more robust model of the relationship between mean gasoline consumption and age and income.

The partially linear model can arise also as a special case of a censored sample selection model (see Heckman, 1974). There, we are interested in estimating β in the equation $y^* = d^*(x\beta + \varepsilon)$ where *d* is a binary observed random variable that indicates censoring: *d*=1 the outcome *y* is uncensored, and *d*=0 otherwise. The model above can be written as

$$E[y|w, x, d = 1] = x\beta + g(w).$$

If there is no overlap between *x* and *w*, this is an example of a partial linear model with a nonparametric selection mechanism. A more general version of this model is studied in Ahn and Powell (1993).

Partially linear models are more attractive than linear models especially in cases where the linearity assumption on a subset of the regressors is suspect. This more robust model allows for a more flexible parametrization for that part of the regression where the analyst is not convinced of the linearity. On the other hand, the main motivation for modelling the regression function as partially nonparametric, or semiparametric, as opposed to fully nonparametric, is the concern for the precision of the estimates. In particular, with more continuous regressors in the regression the ‘curse of dimensionality’ slows the rate of convergence, effectively reducing the usefulness of the regression in data-sets with moderate sizes. Hence, partially linear models provide another practical tool for analysts to use in regressions where linearity of part of the regression function is questionable and provides a middle ground between a completely linear regression that is less robust and one that is totally nonparametric but less practical.

There are many approaches to estimating β and *g*. For example, one can use a penalized spline

regression similar to the one used in Engle et al. (1986), or use semiparametric sieve least squares by replacing the function *f* with an appropriate sieve that approximates the function space (where *g* lies) as sample size increases. The method we describe here uses kernel smoothing similar to the one used by Robinson (1988) and Speckman (1988). Notice that eq. (1) above implies that

$$E[y|z] = E[x'|z]\beta + g(z). \tag{2}$$

Subtracting (2) from (1) we obtain

$$y - E[y|z] = (x - E[x|z])'\beta + u. \tag{3}$$

Hence, one can consistently estimate β by regressing $(y - E[y|z])$ on $(x - E[x|z])$ if the matrix $E[(x - E[x|z])(x - E[x|z])']$ is full rank. This procedure has some similarities to a linear regression where one is interested in a subset of the slope parameters. One can obtain this by regressing the dependent variable on residuals from a regression of the regressors of interest on the nuisance regressors. It is a regression of the outcome on what remains of the regressors after purging them of their linear component that is common with other regressors.

One problem in our set-up is that the regression in (3) is unfeasible since $E[y|z]$ and $E[x|z]$ are not known. These can be consistently estimated using a variety of methods like kernels or sieves. Robinson (1988), for example, replaces the conditional expectations by appropriate Naradaya–Watson kernel estimators where for a random sample of size *N*,

$$\hat{E}[y|z] = \frac{1}{N}\sum_{i=1}^{N} w_{in}(z)\, y_i \tag{4}$$

where the weight function w_{in} is such

$$w_{in}(z) = \frac{K\left(\frac{z_i - z}{h_n}\right)}{\frac{1}{N}\sum K\left(\frac{z_i - z}{h_n}\right)}. \tag{5}$$

K(.) is a kernel function satisfying certain conditions (see Hardle, 1991, for more on smoothing conditional expectations), and h_n is a bandwidth parameter that is positive and converges to zero as sample size increases. Conditions on the rate of convergence of this bandwidth are obtained to ensure desirable theoretical properties of the estimators (for example, on the conditional expectation case, we have $h_n = \lambda n^{-1/5}$ where $0 < \lambda < \infty$). Robinson then shows that the estimator $\hat{\beta}$ of β is normally distributed asymptotically as sample size increases. (The estimator Robinson considers requires trimming those values of *z* that cause instability in the estimates in the ‘random denominator’ of the conditional expectation.) In particular,

$$\sqrt{n}(\hat{\beta} - \beta) \to_d N(0, \sigma^2 E[(x - E[x|z])(x - E[x|z])']^{-1}). \tag{6}$$

This is derived on the assumption of homoskedasticity $(V(u|x) = \sigma^2)$, and other conditions guaranteeing well behaviour of the kernel estimators as sample size increases. As for estimating the nonparametric function g(.), one can use a feasible version of eq. (2) to get

$$\hat{g}(z) = \hat{E}[y|z] - \hat{E}[x|z]'\hat{\beta}. \tag{7}$$

Under a set of assumptions, it can be shown that $\hat{g}(z)$ is a consistent estimator for *g*(*z*). For example, in the case of scalar *z* that has support on [0,1], it can be shown that under appropriate assumptions,

$$\sup_{t\in[0,1]}|\hat{g}(t) - g(t)| = O(n^{-2/5}\log^{2/5} n).$$

In addition, Hardle, Liang and Gao (2001) provide more consistency results for the nonparametric function g(.).

In practice, to implement a partially linear regression, three additional tasks remain. First, one needs to choose a kernel function. Second, although rates of convergence for the smoothing parameter h_n were given, those provide no guidance for choosing a particular value for this smoothing parameter with a given data-set. Third, and to account for sample variability, one needs to obtain estimates for the variance covariance matrix. As for the choice of the kernel function, one can use $K(u) = (2\pi)^{-1/2} exp(-\frac{1}{2}u^2)$, $K(u) = \frac{1}{2}1[|u| \leq 1]$ or a quartic kernel $15/16((1-u^2)^2 I(|u| \leq 1))$ (see Hardle, 1991, for more on kernel selection. Kernel selection does not seem to make a difference in practice). As for the choice of the smoothing parameter h_n, one method that can be used is *cross validation*. In particular, our estimators of β and the function g(.) obtained from (3) and (7) can be written as $\hat{\beta}(h_n)$ and $\hat{g}(z) = \hat{g}(z; h_n)$ which are functions of the smoothing parameter h_n. So, to choose h_n in practice, one can minimize the cross-validation function cv(h_n) defined as

$$cv(h_n) = \frac{1}{N}\sum_{i=1} (_i - x_i\hat{\beta}(h_n) - \hat{g}(z_i; h_n))^2. \tag{8}$$

Finally, to estimate the variance covariance matrix in the homoskedastic case, one can replace $\sigma^2 E[(x - E[x|z])(x - E[x|z])']^{-1}$ by its sample analog. In particular, an estimator $\hat{\sigma}^2$ of σ^2 can be:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i\hat{\beta} - \hat{g}(z_i)\right)^2.$$

However, since the conditional mean is semiparametric, a better estimator for the variance matrix is one that is heteroskedasticity robust. This estimator is similar to the heteroskedasticity robust estimator in linear regression and can be written as

$$\hat{V} = \frac{1}{N}\sum (x_i - \hat{E}(x|z_i))(x_i - \hat{E}(x|z_i))'(y_i - \hat{\beta}x_i - \hat{g}(z_i))^2.$$

One can also approximate the finite sample distribution of the estimator by a bootstrapped distribution.

After estimating β and $g(.)$, one obtains a set of centred residuals $e_i^*, i = 1, \ldots, N$ with distribution F_n^* from which one can draw a bootstrap sample, and then generate a sample of y's from which one can obtain one's bootstrap estimates. Hardle, Liang and Gao (2001) contains consistency results for the bootstrap procedure in the partially linear model above.

Partially linear models are semiparametric linear regressions where the regression function contains a nonparametric function. These regressions are robust to the linear specification for part of the regressors. In addition, partially linear models provide a good alternative to fully nonparametric regression in settings where the data-set that is available is of moderate sample sizes and/or when one has to smooth over a set of continuous random variables of high dimension. Finally, one can also extend the independence (or mean independence) usually used in estimating partially linear models to conditional quantile restrictions and obtain a partially linear semiparametric quantile regression.

See Also

- semiparametric estimation

Bibliography

Ahn, H. and Powell, J. 1993. Semiparametric estimation of censored selection models. *Journal of Econometrics* 58, 3–29.

Engle, R., Granger, C., Rice, C. and Weiss, J. 1986. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81, 310–20.

Hardle, W. 1991. *Applied Nonparametric Regression*. New York: Cambridge University Press.

Hardle, W., Liang, H. and Gao, J. 2001. *Partially Linear Models (Contributions to Statistics)*. Heidelberg: Physica-Verlag.

Heckman, J. 1974. Shadow wages, market wages and labor supply. *Econometrica* 42, 679–93.

Robinson, P. 1988. Root-n-consistent semiparametric regression. *Econometrica* 56, 931–54.

Schmalensee, R. and Stoker, T. 1999. Household gasoline demand in the United States. *Econometrica* 67, 645–62.

Speckman, P. 1988. Kernel Smoothing in partial linear models. *Journal of the Royal Statistical Society, B* 50, 413–36.

How to cite this article

Tamer, Elie. "partial linear model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000339> doi:10.1057/9780230226203.1252

The New Palgrave Dictionary of Economics Online

Pascal, Blaise (1623 – 1662)

A.W.F. Edwards

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

expectations; Fermat, P. de; Pascal, B.; Probability

Article

Pascal was born on 19 June 1623 in Clermont, France, and died on 19 August 1662. In 1631 his father Etienne Pascal moved to Paris in order to secure his son a better education. In 1635 Etienne was one of the founders of Marin Mersenne's ‘Academy’, to which he introduced his son at the age of 14, and Blaise immediately put this new source of knowledge to good use, producing (at the age of 16) his famous *Essai pour les coniques*.

In the succeeding years the young Pascal designed and had built the first mechanical adding machine (there is now a computer language called ‘Pascal’) and conducted experiments into the nature of a vacuum (the ‘Pascal’ is the S.I. unit of pressure), but his chief mathematical contribution was to lay the foundations of the theory of probability.

Before his time probability calculations amounted to no more than the enumeration of equally probable outcomes in games of chance, but Pascal introduced the important idea of expectation and used recursively the fact that if expectations of gain X and Y are equally probable, the expectation is $\frac{1}{2}(X+Y)$. He also introduced the binomial distribution for equal chances and with its help, and that of mathematical induction applied to expectations, solved the Problem of Points for two players.

This problem was the topic of correspondence between Pascal and Pierre de Fermat in 1654 which, together with Pascal's contemporary *Traité du triangle arithmétique*, includes three methods of solution. Two players stake equal money on being the first to win n points in a game in which the winner of each point is determined by the toss of a coin. If such a game is interrupted when one player still lacks a points and the other b, how should the stakes be divided between them?

Fermat and Pascal independently concluded that the problem could be solved by noting that at most $(a+b-1)$ more tosses will settle the game, and that if this number of tosses is imagined to have been made, the resulting 2^{a+b-1} possible games (each equally probable) may be classified according to the winner in

each case, the stakes then being divided accordingly. Thus the real game, of indeterminate length, is embedded in an imaginary game of fixed length. Apart from this novel idea, however, such a solution by enumeration was straightforward, but Pascal offered both an independent method based on expectations which is valid for any number of players, and, in the *Traité du triangle arithmétique*, the solution for two players in terms of the binomial distribution, proved by induction. He did not give the binomial distribution algebraically, but by reference to the ‘arithmetical triangle’ of binomial coefficients, whose properties he elaborated in his *Traité* (whence the name ‘Pascal's triangle’).

In the *Pensées* Pascal introduced his celebrated wager *‘infini–rien’*, in which he argued that we should wager for the existence of God since the stakes are finite (our lives) but there is an infinite prize (eternal life). The argument is that of modern decision theory, which it may be said to foreshadow. The ‘states of nature’ are the existence and non-existence of God whilst the ‘decisions’ are to act as if God exists and as if he does not. If God does exist and we act as if he does, the ‘utility’ is infinite, and thus so is the ‘expected utility’ of this course of action whether he exists or not, provided there is a non-zero chance that he does.

How to cite this article

Edwards, A. W. F. "Pascal, Blaise (1623–1662)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000034> doi:10.1057/9780230226203.1253

The New Palgrave Dictionary of Economics Online

patent pools

Daniel Quint

From The New Palgrave Dictionary of Economics Online, ,
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A patent pool is an agreement by multiple patentholders to share intellectual property among themselves or to license a portfolio of patents as a package to outsiders. Patent pools were common in the United States from the 1890s to the 1940s; since the mid-1990s, there has been a resurgence of patent pools tied to technological standards. I discuss the history and antitrust treatment of patent pools in the United States, and review the related academic literature (both theoretical and empirical).

Keywords

antitrust legislation; complements; intellectual property; patent pools; patents

Article

A patent pool is an agreement by multiple patentholders to share intellectual property among themselves or to license a portfolio of patents as a package to outsiders.
Patent pools were common in the United States in the first half of the 20th century, and reemerged as an important institution in the mid-1990s; an estimated $100 billion worth of goods sold in 2001 were based at least partly on pooled patents.

History

The first patent pool emerged from infringement lawsuits won by Elias Howe, credited with inventing the sewing machine, who returned from marketing his invention in England in the 1840s to find that others had copied it. Following the lawsuits, Howe, Isaac Singer and two other manufacturers established a pool of sewing machine-related patents in 1856, with Howe receiving the bulk of the royalties.
Patent pools were commonplace in the United States from the 1890s to the 1940s. Lerner, Strojwas and

Tirole (2007) identify 125 pools, most of them from this time; Lerner and Tirole (2007) claim that in the early 20th century, ‘many (if not most) important manufacturing industries had a patent pooling arrangement’. (A partial list from Merges (2001) includes pools covering shoe machinery, automobiles, bathtubs, door parts, seeded raisins, coaster brakes, davenport beds, movie projectors, hydraulic pumps, and swimming pool cleaners; a longer list from Lerner, Strojwas and Tirole includes railroad couplers, television equipment, and plastic artificial eyes.) In 1917, with airplanes needed for the First World War, then Assistant Secretary of the Navy Franklin D. Roosevelt pushed eight aircraft manufacturers into a patent pool because patent litigation had shut down US aircraft production. A 1915 pool containing automobile patents had 146 initial members, but most of the pools examined in Lerner, Strojwas and Tirole started with six members or fewer.

Following Congressional hearings on patent pools in the 1930s and 1940s and several negative antitrust rulings, patent pools essentially vanished from the mid-1950s until the mid-1990s. In 1997, after extensive discussion with regulators, a pool formed containing patents essential to the MPEG-2 digital video standard. This was followed by pools tied to the DVD, Bluetooth, 1394 (Firewire), DVB-T, MPEG-4 (AVC) and 3G-Mobile standards. The MPEG-2 pool alone currently has 26 members, nearly a thousand patents, and over 1,300 licensees and affiliates. Pools have also recently been discussed for the biotech and pharmaceutical industries.

Antitrust treatment

For two decades following the passage of the Sherman Antitrust Act in 1890, patent pools appeared to offer a way to circumvent its prohibitions. In 1902, the Supreme Court upheld the legality of the National Harrow pool, which dominated the market for float spring tooth harrows. Among other things, the licensing terms required licensees to only sell particular products, and fixed the prices for these products. The Court wrote:

> The general rule is absolute freedom in the use or sale of rights under the patent laws of the United States. The very object of these laws is monopoly, and the rule is, with few exceptions, that any conditions which are not in their very nature illegal with regard to this kind of property… will be upheld by the courts.
> *E. Bement & Sons* v. *National Harrow* (186 US 70)

In 1912, however, the Court reversed itself, upholding a lower court's break-up of a pool with similarly restrictive licensing terms (*Standard Sanitary Manufacturing* v. *United States*). In the decades following, the court continued to focus on licensing terms, breaking up pools that fixed downstream prices or production, and allowing pools whose licensing agreements ‘contained no restrictions as the quantity of goods to be produced, or the price to be charged, or the territory in which they might be sold by the licensee’ (*Baker-Cammack Hosiery Mills* v. *Davis*, 181 F.2d 550 1950). In 1945, the Supreme Court ruled against the Hartford-Empire pool, which used licensing terms to set production quotas in the glassware manufacturing industry, claiming, ‘The history of this country has perhaps never witnessed a more completely successful economic tyranny over any field of industry’ (*Hartford Empire Co.* v. *United States*, 323 US 386). Although the Baker-Cammack ruling followed that, several other pools

were broken up in subsequent years (*United States* v. *Line Material*, *United States* v. *U.S. Gypsum*, *United States* v. *New Wrinkle*), and Hartford-Empire was generally seen as signalling the end of favourable treatment toward pools; by the mid-1950s, pool formation had essentially ceased.
This changed following release of the Antitrust Guidelines for the Licensing of Intellectual Property by the Department of Justice and Federal Trade Commission in April 1995. Under the heading 'cross-licensing and pooling arrangements,' the Guidelines stated:

> These arrangements may provide procompetitive benefits by integrating complementary technologies, reducing transaction costs, clearing blocking positions, and avoiding costly infringement litigation. By promoting the dissemination of technology, cross-licensing and pooling arrangements are often procompetitive.

Department of Justice analysis, enunciated in business review letters of several proposed pools, focused on three questions: whether a pool would integrate complementary patent rights (as opposed to patents which would otherwise be in competition); whether it would foreclose competition in related markets; and whether it would discourage further innovation. In the cases of the MPEG-2, DVD, and 3G pools, the DOJ stated after review that it was 'not presently inclined to initiate antitrust enforcement action against the conduct you have described'. In 1998, the FTC did challenge a pool formed by Summit Technology and VISX, the only firms with FDA-approved technology for laser eye surgery, which was viewed to be functioning primarily as a price-fixing arrangement; the pool was dissolved as part of a settlement resolving the case. A 2007 DOJ/FTC report, which followed public hearings held in 2002, summarizes the current regulatory view.

Characteristics of recent pools

To address the first regulatory concern – the integration of only complementary patent rights – recent pools have been limited to patents deemed essential for standard compliance. The business review letter on the proposed MPEG-2 pool reads:

> The Portfolio combines patents that an independent expert has determined to be essential to compliance with the MPEG-2 standard; there is no technical alternative to any of the Portfolio patents within the standard. Moreover, each Portfolio patent is useful for MPEG-2 products only in conjunction with the others. The limitation of the Portfolio to technically essential patents, as opposed to merely advantageous ones, helps ensure that the Portfolio patents are not competitive with each other …. The continuing role of an independent expert to assess essentiality is an especially effective guarantor that the Portfolio patents are complements, not substitutes.
> Joel Klein (Acting Assistant Attorney General), letter to Garrard Beeney, 26 June 1997

Several of the recent pools include grantback provisions – pool participants and licensees agree to add to the pool, or to license to each other at reasonable terms, any future patents they receive that are judged to be essential. The pools also allow for separate licensing of individual patents – that is, licensing through

the pool is not done exclusively. The majority of the recent pools allocate revenue in proportion to the number of essential patents that each firm has contributed to the portfolio, although some of the pools do attempt to account for patents that are more or less valuable.

One unusual case is the 3-G mobile standard. 3-G was designed to use five different radio interfaces, in order to be backward-compatible with five second-generation wireless networks. Antitrust concerns led to the establishment of five separate License Administrators to oversee licensing of patents essential for each interface, rather than a single platform or pool containing all of the relevant patents. (The 3-G platforms are different from traditional pools in that all licensing is done ‘a la carte’, at standardized terms set by each Administrator.)

Theoretical literature

Shapiro (2001) employs a Nash-Bertrand model to show that pools result in lower prices and greater welfare when patents are perfect complements, by correcting the ‘complements problem’ of excessive prices; and higher prices and lower welfare when patents are perfect substitutes, by eliminating competition. Kim (2004) finds that when patents are perfect complements, the case for pools is even stronger in the presence of vertically integrated firms (patentholders who are also downstream producers). Choi (2003), on the other hand, shows that patent pools change the incentive for another patentholder or a potential infringer to challenge questionable patents in court, making pools of complementary but weak patents possibly welfare-destroying.

Lerner and Tirole (2004) introduce a more flexible model than perfect complements and perfect substitutes, and show that when patents are more substitutable, pools are more prone to be welfare-negative. They show that forcing pool participants to also make their patents available individually has a destabilizing effect on welfare-negative pools, but no effect on welfare-positive pools, and therefore propose compulsory individual licensing as a screen for efficient pools. Brenner (forthcoming) examines the equilibrium effects of different pool formation rules in the Lerner and Tirole framework, showing that endogenously occurring pools will be inefficiently small if patentholders can opt out individually without disrupting pool formation. My own work (Quint, 2008) examines pools in a setting with both essential and nonessential patents; I find that pools of essential patents are always welfare-increasing, while pools containing nonessential patents have ambiguous welfare effects, even when they are limited to patents that are perfect complements. I also find that when a pool is welfare-increasing, agreements that ‘bind the pool's hands’ with respect to pricing will reduce, and may even reverse, the welfare gains.

Empirical literature

Merges (2001) discusses the workings of many historical pools. Gilbert (2004) discusses a number of important court rulings and how they hold up under economic analysis. Lerner, Strojwas and Tirole (2007) analyse the licensing rules of 63 patent pools, most from before 1950 but a handful from the 1990s; they find that, consistent with theory, pools containing complementary patents were more likely to allow independent licensing and require grantbacks. Layne-Farrer and Lerner (2008) examine arrangements for dividing pool revenue and its effect on participation; they also find that vertically

integrated firms are more likely to join pools. Lerner and Tirole (2007) review current public policy and suggest certain changes.

See Also

- antitrust enforcement
- intellectual property
- patents

Bibliography

Brenner, S. Forthcoming. Optimal formation rules for patent pools. *Economic Theory*.

Choi, J.P. 2003. Patent pools and cross-licensing in the shadow of patent litigation. CESifo Working Paper #1070. Available at http://ssrn.com/abstract=466062.

Gilbert, R. 2004. Antitrust for patent pools: a century of policy evolution. *Stanford Technology Law Review* 3.

Kim, S. 2004. Vertical structure and patent pools. *Review of Industrial Organization* 25(3), 231–50.

Layne-Farrar, A. and Lerner, J. 2008. To join or not to join: examining patent pool participation and rent sharing rules. SSRN working paper. Available at http://ssrn.com/abstract=945189.

Lerner, J., Strojwas, M. and Tirole, J. 2007. The design of patent pools: the determinants of licensing rules. *RAND Journal of Economics* 38(3), 610–25.

Lerner, J. and Tirole, J. 2004. Efficient patent pools. *American Economic Review* 94(3), 691–711.

Lerner, J. and Tirole, J. 2007. Public policy toward patent pools. In *Innovation Policy and the Economy*, Vol. 8, pp. 157–86, ed. A. Jaffe, J. Lerner and S. Stern. Chicago: University of Chicago Press.

Merges, R. 2001. Institutions for intellectual property transactions: the case of patent pools. In *Expanding the Boundaries of Intellectual Property: Innovation Policy for the Knowledge Society*, pp. 123–66, ed. R. Dreyfuss, D. Zimmerman and H. First. Oxford: Oxford University Press.

Quint, D. 2008. Economics of patent pools when some (but not all) patents are essential. Working paper. Available at http://www.ssc.wisc.edu/~dquint/papers/patent-pools-quint.pdf.

Shapiro, C. 2001. Navigating the patent thicket: cross licenses, patent pools, and standard-setting. In

Innovation Policy and the Economy, Vol. 1, pp. 119–50, ed. A. Jaffe, J. Lerner and S. Stern. Cambridge, Mass.: MIT Press.

Regulatory guidelines and business review letters

US Department of Justice and Federal Trade Commission. 1995. *Antitrust Guidelines for the Licensing of Intellectual Property*. Available at http://www.usdoj.gov/atr/public/guidelines/0558.htm.

US Department of Justice and Federal Trade Commission. 2007. *Antitrust Enforcement and Intellectual Property Rights: Promoting Innovation and Competition*. Chapter 3: Antitrust analysis of portfolio cross-licensing agreements and patent pools. Available at http://www.usdoj.gov/atr/public/hearings/ip/chapter_3.htm.

Letter from Joel Klein, Acting Assistant Attorney General, Department of Justice, to Garrard Beeney, 26 June 1997. Available at http://www.usdoj.gov/atr/public/busreview/215742.htm.

Letter from Joel Klein, Assistant Attorney General, Department of Justice, to Garrard Beeney, 16 December 1998. Available at http://www.usdoj.gov/atr/public/busreview/2121.htm.

Letter from Joel Klein, Assistant Attorney General, Department of Justice, to Carey Ramos, 10 June 1999. Available at http://www.usdoj.gov/atr/public/busreview/2485.htm.

Letter from Charles James, Assistant Attorney General, Department of Justice, to Ky Ewing, 12 November 2002. Available at http://www.usdoj.gov/atr/public/busreview/200455.pdf.

How to cite this article

Quint, Daniel. "patent pools." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000371> doi:10.1057/9780230226203.1871

close print preview

The New Palgrave Dictionary of Economics Online

patent races

Richard A. Jensen

From The New Palgrave Dictionary of Economics Online, , 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A patent race is a competition between two or more inventors, typically firms, to discover an invention first, thereby securing a patent which protects the invention from imitation. The date at which a firm discovers the invention is stochastic, but can be reduced in expectation by increased investment in research and development. Competition to win the patent leads firms to over-invest, compared with the outcome where they invest cooperatively and share the patent equally. However, the expected discovery date is later than socially optimal, so innovation is delayed, on average, compared with the social optimum.

Keywords

patents; research and development; exponential distribution; Schumpeterian hypothesis; hazard function; Nash equilibrium; feedback strategies; subgame perfect equilibrium

Article

A patent race is a situation in which two or more inventors, typically firms, compete to discover an invention first, thereby securing a patent which protects the invention from imitation or infringement. The literature on patent races is predominantly theoretical. These analyses have two fundamental properties. First, for each firm, the discovery date of the invention is stochastic, and depends on the effort expended (or investment) by both itself and its rivals. It is common to assume the discovery date is a random variable that is exponentially distributed with a parameter that depends on the knowledge levels of the firms, which in turn depend on the cumulative research and development (R&D) investments of the firms. If firm i invests an amount $I_i(t)$ at date t, then the growth of its knowledge stock $K_i(t)$ is $K_i'(t) = I_i(t)$. The distribution of firm i's random discovery date t_i is $Pr\{t_i \leq t\}=F(K_i(t))=1-exp\{-\lambda K_i(t)\}$, where $\lambda>0$ is a parameter, noteworthy because the cumulative knowledge needed for discovery is exponentially distributed with mean $1/\lambda>0$. The probability that i discovers at t,

conditional on i not discovering before t, is $Pr\{t_i \in (t, t+dt] | t_i > t\} = \lambda I_i(t)dt$. Given n firms, if their research processes are stochastically independent, then the probability density that i wins the race at t is $\lambda I_i(t) exp\left\{-\lambda \Sigma_{j=1}^{n} K_i(t)\right\}$.

Second, the race is modelled as a game in which each firm chooses its R&D investment (effort) at each t to maximize the present discounted value of its expected profit, subject to knowledge growth $K_i'(t) = I_i(t)$ from an initial stock $K_i(0)=K_0 \geq 0$. If $V>0$ is the value of the invention at its discovery date, $r>0$ is the interest rate, and $c_i(x_i)$ is i's R&D cost function, then this expected present value is

$$\int_0^{\infty} exp\left\{-rt\right\}[\lambda I_i(t) exp\left\{-\lambda \sum_{j=1}^{n} K_i(t)\right\} V - exp\left\{-\lambda \sum_{j=1}^{n} K_i(t)\right\} c_i(x_i(t))]\, dt.$$

Research on patent races initially focused on the Schumpeterian hypothesis regarding the relationship between competition and the pace of innovative activity. In the seminal article on patent races, Loury (1979) assumes each firm chooses a lump-sum R&D expenditure at the start, so $c_i(x_i(t))=x_i(t)$ where $x_i(0)=X_i$ and $x_i(t)=0$ for $t>0$. With no knowledge accumulation over time, the probability density of discovery becomes $h(X_i) exp\left\{-\Sigma_{j=1}^{n} h(X_i) t\right\}$ where the hazard function $h(X_i)$, the probability that i discovers at t, given that it has not discovered before t, depends only on the lump-sum R&D expenditure. Thinking of invention as a stochastic production process, it is natural to assume that this hazard function is increasing in expenditure, possibly with initially increasing returns to scale, but necessarily with decreasing returns eventually. In this model, increased competition (an increase in the number of firms) reduces the Nash equilibrium expenditure of each firm. Given a fixed number of firms, however, each firm spends more than in the outcome where they invest cooperatively and share the patent value equally. Thus, with unrestricted entry, there are too many firms and too much aggregate R&D investment in the Nash equilibrium, compared with the cooperative outcome.

Lee and Wilde (1980) note that Loury's approach does not allow firms to invest in R&D over time. They assume instead that each firm initially chooses a level of R&D expenditure for each date until it or a rival discovers the invention, $c_i(x_i(t))=x_i$ for all $t \geq 0$ before discovery and $c_i(x_i(t))=0$ thereafter. Again, with no knowledge accumulation, the probability density of discovery is $h(x_i) exp\left\{-\Sigma_{j=1}^{n} h(x_i) t\right\}$ where the hazard function $h(x_i)$ now depends only on current R&D expenditure. In this case, increased competition increases the Nash equilibrium expenditure of each firm. For a fixed number of firms, each firm spends more than in the outcome where they invest cooperatively and share the patent value equally. And with unrestricted entry, there are too many firms, each of which spends too much on R&D, in the Nash equilibrium, compared with the cooperative outcome.

One notable difference between these studies is that competition increases R&D investment per firm in Lee and Wilde's approach. This arises from the different R&D strategies. In Loury's model, firms choose the scale of R&D effort with one initial investment, whereas in Lee and Wilde's model, firms choose the

intensity of R&D effort per period. In the latter approach, firms can cut their R&D spending, and so their losses, after a rival discovers.

Both of these patent races are essentially static in that the firms choose the strategies at the beginning of the game, and there is no knowledge accumulation. Reinganum (1982) generalizes this by allowing firms to choose feedback strategies: each firm chooses its R&D investment at each date as a function of the observed knowledge stocks of all firms in the race. When patent protection is perfect, increased competition increases the R&D investment expenditure of each firm. However, the effect of increased competition on a firm's R&D investment is ambiguous when patent protection is not perfect. Finally, when the social value of the patent exceeds the private value to a firm in the race, the noncooperative growth rate of knowledge is less than socially optimal, and so innovation is delayed on average compared with the social optimum.

Subsequent research has sought to understand the effects of two types of asymmetries on patent race outcomes. The first type involves an incumbent who owns the current patent and a group of potential entrants vying for the patent to the next generation technology. There are two conflicting effects. First, there is a dissipation of monopoly rent if the incumbent loses. If the innovation is a non-drastic new process, and M=monoply profit and D=duopoly profit, then the incumbent earns $M–D$ if it wins, and the entrant earns D, where typically $M–D>D$. However, if the monopolist wins, it replaces itself as the monopolist (Arrow, 1962). If the innovation is drastic and pre-innovation profit is π, then the monopolist's gain from winning is $M-\pi$, while the entrant's is $M>M-\pi$. Incumbents have a greater incentive to innovate when the rent dissipation effect dominates (Gilbert and Newbery, 1982), but not when the replacement effect dominates (Reinganum, 1983).

The second asymmetry involves a race in which one firm has a lead in the race. When the lead takes the form of greater accumulated knowledge, $K_i(0)>K_j(0)$ for all $j\neq i$, the laggards simply exit and concede the race to the leader firm i in the unique subgame perfect equilibrium. However, the laggards can and do remain in the race if there is some way that they can leapfrog into the lead, such as when the R&D process requires completion of several successive stages (Fudenberg et al., 1983).

See Also

- intellectual property
- patents

Bibliography

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity*, ed. R.R. Nelson. Princeton, N. J.: Princeton University Press.

Fudenberg, D., Gilbert, R., Stiglitz, J. and Tirole, J. 1983. Preemption, leapfrogging and competition in patent races. *European Economic Review* 22, 3–31.

Gilbert, R. and Newbery, D. 1982. Preemptive patenting and the persistence of monopoly. *American*

Economic Review 72, 514–26.

Lee, T. and Wilde, L.L. 1980. Market structure and innovation: a reformulation. *Quarterly Journal of Economics* 94, 429–36.

Loury, G.C. 1979. Market structure and innovation. *Quarterly Journal of Economics* 93, 395–410.

Reinganum, J.F. 1982. Dynamic game of R and D: patent protection and competitive behavior. *Econometrica* 50, 671–88.

Reinganum, J.F. 1983. Uncertain innovation and the persistence of monopoly. *American Economic Review* 73, 741–8.

How to cite this article

Jensen, Richard A. "patent races." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_P000374> doi:10.1057/9780230226203.1897

The New Palgrave Dictionary of Economics Online

patent valuation

Jonathan Putnam

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Agents routinely appraise and trade individual patents. But small-sample methods (generally derived from basic accounting and finance) are often crude, and their results may bear little relationship to economic fundamentals, especially in litigation. Meanwhile, large-sample methods usually lack much invention-specific data on which to condition value estimates. Regardless of sample size, proper valuation methods require both conceptual delineation and empirical ingenuity.

Keywords

disclosure; inventions; patent citations; patent counts; patent valuation; patents; research and development

Article

The valuation of patent rights sounds like a simple enough concept. It is true that agents routinely appraise and trade individual patents. But small-sample methods (generally derived from basic accounting and finance) are often crude, and their results may bear little relationship to economic fundamentals, especially in litigation. On the other hand, large-sample methods usually lack much invention-specific data on which to condition value estimates. Regardless of sample size, proper valuation methods require both conceptual delineation and empirical ingenuity.

Concepts

Legally, a patent is the right to exclude others from making, using or selling an invention. In economic terms, that right is an asset, yielding a non-negative returns stream while it is enforceable. Because the right is a private means (increased exclusivity) to a public end (increased productivity), a patent's private value only partially conveys its market significance.

Unlike most property rights, patents do not comprise the affirmative right to use the invention. Absent the right to use, patents may generate private value only when combined with complementary assets, such as a licence under other patents. Contracting problems (for example, asymmetric information) may strongly influence value.

A patent may generate private returns apart from the right to exclude rivals. The patentee may use it: to monitor employee performance; to signal otherwise unobservable quality to prospective financiers; to enhance reputation; to signal a willingness to litigate; or to reduce the costs of settlement in the event that litigation occurs (‘defensive patenting’). In large samples, it is usually impossible either to observe the magnitude and timing of these sources of value, or to decompose them.

Patents also impose unobservable private costs on the patentee. Chiefly, the inventor must disclose the means for reproducing the invention. Disclosure reduces the cost to rivals of reproducing the invention (static spillover) and conducting R&D (dynamic spillover). Apart from reducing the incentive to invent, these private costs imply social benefits not captured by the patentee.

Cross-sectionally, patents are usually modelled as having a one-dimensional ‘quality’ (which is either synonymous with, or a monotone function of, the patent's value). More precisely, a patent's private value depends significantly on the exclusivity conferred by its claims, but its uncaptured social value depends significantly on the scope of its disclosure (which must be at least as broad as the claims). For various reasons, including rival use of the patentee's disclosure to develop competing innovations (‘creative destruction’), the social and private values of a patent may diverge. Thus, it is theoretically preferable, but empirically much less tractable, to model patents as having two-dimensional ‘quality’.

Over time, because of ongoing research by the patentee and his rivals, the private returns to patent protection may fluctuate sharply up or down, in response to complementary or competitive discoveries. The variance is likely to be larger in a patent's early years.

Stylized facts

The following stylized facts bear on the calculation of aggregate private patent values:

1. 1. Whether aggregated by firm, industry or country, patent counts do not vary much from one period to the next.
2. 2. The distribution of patent values is skewed.
3. 3. Social and private patent values are imperfectly correlated.
4. 4. *Ex ante* and *ex post* values are imperfectly correlated.
5. 5. Most patents are not traded.
6. 6. Samples are selected (not all innovations are patented; not all applications are filed in any single country; not all applications are granted).

Related research

Proceeding in the direction of generally increasing complexity and structure, the following categories describe large-sample models that economists have developed to value patent rights. Lanjouw, Pakes and Putnam (1998) surveys recent papers.

Patent counts

A variety of models employ simple patent counts to indicate the value of patent rights. Strictly speaking, patent counts indicate quantities rather than values. Under certain assumptions, relative quantities may be proportional to relative values. For example, if two patent samples are drawn from the same value distribution, then the ratio of quantities is an efficient estimator of the ratio of values.

Griliches (1990) reviews a large number of studies that, implicitly or explicitly, rely on this assumption. Griliches' view of 'patent [counts] as economic indicators' is not encouraging ('The food here is terrible.' 'Yes, and the portions are so small.'). Stylized facts 1 and 2 combine to thwart inference. A firm facing a fixed budget constraint may patent its best *N* inventions, which implies little intertemporal variation in patent counts even if their realized quality varies markedly. Thus, patent counts are a *biased* measure of value. Because R&D outcomes are highly variable and skewed, patent counts are an *imprecise* measure of value. For these reasons, the assumption that patent samples are drawn from the same distribution is difficult to test, and often false.

On the other hand, fixed budget constraints for R&D and patenting imply that patent counts may proxy for the value of R&D *inputs*. Hausman, Hall and Griliches (1986) model the lag relationship between patent counts and R&D, and find an approximately contemporaneous relationship.

One may compute implied patent values by associating patent counts with other observable aggregates. On the macro level, McCalman (2005) employs the structural imitation model of Eaton and Kortum (1996) to determine international 'trade' in patents. He estimates that the worldwide value of patent applications filed by US inventors in 1988 was about $12.4 billion ($163,700 per application). The estimates for four other large patenting countries vary: France, $147,200; Germany, $82,200; UK, $53,100; Japan, $47,700.

At the firm level, Pakes (1986) constructs a time series model of patent applications, R&D and the stock market rate of return. Controlling for R&D expenditures, an unanticipated patent application implies an $800,000 increase in market capitalization. This relatively high value also reflects investors' revised expectations of research success, and the selection of publicly traded patentees (which are larger and more successful than average).

Patent citations (weighted patent counts)

Patent examiners cite prior patents when they decide whether to grant a patent application. Analysts count these citations to indicate the value of the cited patent. Patent counts are then weighted by the number of citations. A recent book-length treatment is Jaffe and Trajtenberg (2002).

This branch of the literature divides in two: estimates of the relationship between citations and patent value; and studies that assume that relationship. In the former category, Trajtenberg's (1990) pioneering study showed that citation-weighted patent counts perform better than unweighted counts in explaining aggregate patent value (see Harhoff et al., 1999). However, this and subsequent studies found that citations tend to indicate the social value of the patent rather than the purely private value (stylized fact 3). Private value is better captured by 'self-citations' from the patentee's own later inventions. Hall, Jaffe and Trajtenberg (2005) show that weighted patent counts are associated with – and predict – higher

stock market returns.
Assuming that citations proxy for value, Henderson, Jaffe and Trajtenberg (1998) examine the contribution of university patenting to commercial technology; Trajtenberg, Henderson and Jaffe (1997) find that the 'basicness' of university patents relative to corporate patents has narrowed over time. Jaffe, Trajtenberg and Henderson (1993) model the spatial distribution of dynamic spillovers.

Other indicator-based methods

Lanjouw and Schankerman (2004) construct a composite index of patent quality using several indicators (forward and backward citations, number of claims, and number of filing countries). This combination of *ex ante* and *ex post* measures (stylized fact 4) efficiently aggregates informationally distinct components of patent value. The composite also explains related *ex post* decisions (for example, patent renewal and litigation); forward citations (an *ex post* measure) demonstrate the greatest explanatory power.

Structural models: patent renewals and patent applications

Although most patents are not traded (stylized fact No. 5), patent office rules effectively require patentees to make optimal investments to create and maintain patent rights. These investments reveal information about the expected value of the asset. The information is censored, however, because (conditional on choosing to invest) patentees make the same investment regardless of the expected value. Structural econometric models identify the underlying value distribution.
Most countries require that a patentee pay an increasing fee to keep a patent right in force. Beginning with Pakes and Schankerman (1984), so-called patent renewal models exploit the optimal stopping problem implicit in the annual investment decision. The *ex post* value distribution is identified from the shares of an annual cohort that are renewed each subsequent year when patentees confront known renewal fee schedules, observed over multiple cohorts. In relatively simple *deterministic* models (Schankerman and Pakes, 1986; Sullivan, 1994; Schankerman, 1998), returns are assumed to depreciate at a known rate following an initial draw from the value distribution. In more complex *options* models (Pakes, 1986; Lanjouw, 1998), returns evolve stochastically. In both models, the average patent value is relatively low (for example, less than $20,000 in Europe during the post-war period). Lorenz plots reveal that the top 10 per cent of patents account for about 47 per cent of the total value distribution.
The value distribution may also be identified from cross-sectional information (Putnam,1996). Under international rules, patent applicants typically determine simultaneously whether to file in each jurisdiction outside their home jurisdiction. Applicants file if the capitalized value of net returns exceeds the application cost. Application models capture filing anywhere in the world, conditional on a common information set, which mitigates both intertemporal (stylized fact #4) and sample selection (stylized fact #6) problems. The *ex ante* value distribution is identified from the combination of filing countries, assuming that national returns are the product of a common invention-level 'random effect' and an idiosyncratic national market draw. Putnam (1996) values the mean German patent at about $69,000 in 1974, with the top 10 per cent of patents accounting for about 70 per cent of the value distribution.

Small-sample methods

Small-sample patent valuation typically occurs in a legal or quasi-legal context, such as licensing or litigation. In infringement litigation, the law typically allows one of three measures of damages: the patentee's lost profits; the infringer's incremental profits; or a ‘reasonable royalty’ (conceived as the outcome of a hypothetical licensing negotiation (Weil, Wagner and Frank, 2001)). Typically, parties employ discounted cash flow methods and ‘comparable’ licence transactions to support valuation claims. Both *ex ante* and *ex post* methods are used, not always consistently. The law also allows limited consideration of an infringer's *ex ante* alternatives to infringement, such as inventing a substitute. Generally, the most difficult legal and empirical question is: What fraction of (actual or expected) profits should be imputed to the patent? While much damages jurisprudence remains economically ad hoc, courts are increasingly inclined to require the same market analyses that characterize antitrust law (*Crystal Semiconductor v. TriTech Microelectronics*, 246 F. 3d 1336, (Fed. Cir. 2001)).

See Also

- patents

Bibliography

Eaton, J. and Kortum, S. 1996. Trade in ideas: patenting and productivity in the OECD. *Journal of International Economics* 40, 251–78.

Griliches, Z. 1990. Patent statistics as economic indicators: a survey. *Journal of Economic Literature* 28, 1661–707.

Hall, B., Jaffe, A. and Trajtenberg, M. 2005. Market value and patent citations. *RAND Journal of Economics* 36, 16–38.

Harhoff, D., Narin, F., Scherer, F.M. and Vopel, K. 1999. Citation frequency and the value of patented inventions. *Review of Economics and Statistics* 81, 511–15.

Hausman, J., Hall, B. and Griliches, Z. 1986. Patents and R&D: is there a lag? *International Economic Review* 27, 265–83.

Henderson, R., Jaffe, A. and Trajtenberg, M. 1998. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and Statistics* 80, 119–27.

Jaffe, A., Trajtenberg, M. and Henderson, R. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 43, 578–98.

Jaffe, A. and Trajtenberg, M. 2002. *Patents, Citations and Innovations: A Window on the Knowledge Economy*. Cambridge, MA: MIT Press.

Lanjouw, J.O. 1998. Patent protection in the shadow of infringement: simulation estimations of patent value. *Review of Economic Studies* 65, 671–710.

Lanjouw, J.O., Pakes, A. and Putnam, J. 1998. How to count patents and value intellectual property: uses of patent renewal and application data. *Journal of Industrial Economics* 46, 405–33.

Lanjouw, J.O. and Schankerman, M. 2004. Patent quality and research productivity: measuring innovation with multiple indicators. *Economic Journal* 114, 441–65.

McCalman, P. 2005. Who enjoys ‘TRIPs’ abroad? An empirical analysis of intellectual property rights in the Uruguay Round. *Canadian Journal of Economics* 38, 574–603.

Pakes, A. 1986. Patents as options: some estimates of the value of holding European patent stocks. *Econometrica* 54, 755–84.

Pakes, A. and Schankerman, M. 1984. The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. In *R&D, Patents and Productivity*, ed. Z. Griliches. Chicago: University of Chicago Press.

Pakes, A. and Simpson, M. 1989. Patent renewal data. *Brookings Papers on Economic Acivity: Microeconomics* 1989, 331–410.

Putnam, J. 1996. The value of international patent rights. Ph.D. thesis, Yale University.

Schankerman, M. 1998. How valuable is patent protection? Estimates by technology field. *RAND Journal of Economics* 29, 77–107.

Schankerman, M. and Pakes, A. 1986. Estimates of the value of patent rights in European countries during the post-1950 period. *Economic Journal* 96, 1052–76.

Sullivan, R. 1994. Estimates of the value of patent rights in Great Britain and Ireland, 1852–1876. *Economica* 61, 37–58.

Trajtenberg, M. 1990. A penny for your quotes: patent citations and the value of innovation. *RAND Journal of Economics* 21, 172–87.

Trajtenberg, M., Henderson, R. and Jaffe, A. 1997. University versus corporate patents: a window on the basicness of invention. *Economics of Innovation and New Technology* 5, 19–50.

Weil, R.L., Wagner, M.J. and Frank, P.B. 2001. *Litigation Services Handbook: The Role of the Financial Expert*. New York: Wiley.

How to cite this article

Putnam, Jonathan. "patent valuation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000326> doi:10.1057/9780230226203.1255

The New Palgrave Dictionary of Economics Online

patents

Bronwyn H. Hall

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A patent is the legal right of an inventor to exclude others from making or using a particular invention. This right is sometimes termed an ‘intellectual property right’ and is viewed as an encouragement for innovation. This article gives a brief history of patenting, and discusses the legal and administrative process for obtaining a patent in the major world jurisdictions. Evidence on patent effectiveness in encouraging innovation is surveyed, and the article concludes with a discussion of the use of patent data in economic analysis.

Keywords

biotechnology industry; innovation; intellectual property rights; inventions; patent citations; Patent Cooperation Treaty; patent litigation; patent races; patent valuation; patents; pharmaceutical industry; research and development; Trade-Related Aspects of Intellectual Property Rights (TRIPS); World Intellectual Property Organization

Article

A patent is the legal right of an inventor to exclude others from making or using a particular invention. This right is customarily limited in time, to 20 years from the date of the application submission in most countries. The principle behind the modern patent is that an inventor is allowed a limited amount of time to exclude others from supplying or using an invention in order to encourage inventive activity by preventing immediate imitation. In return, the inventor is required to make the description and implementation of the invention public rather than keeping it secret, allowing others to build more easily on the knowledge contained in his invention.

The economics of patents has two distinct components, one normative and one positive. The first is directed towards questions of optimal patent policy, the existence and strength of patents, and the design of the patent system. The second uses patent data as an indicator of inventive activity, relying on the fact

that patent offices attempt to apply fairly uniform standards of novelty and inventive step when granting patents, so that counts based on them should reflect the innovative activity in a society or in a particular industrial or technology sector. The advantage of patent data is that they are available in great detail over a wide range of time periods, geographic areas, and technological sectors (Griliches, 1990).

Nevertheless, all patents are not equal, and it is important to understand the operation of patent systems throughout their history in order to make effective use of these data.

This article begins with a brief history of patents, followed by a discussion of the legal and administrative processes for obtaining a patent in the three major patent offices, the United States, European, and Japanese. Then the evidence on patent effectiveness in encouraging innovation is surveyed. The final section discusses the use of patent data in economic analysis.

Brief history

Patents have a long history, although some of the earliest patents are simply the grant of a legal monopoly in a particular good rather than protection of an invention from imitation. Early examples of technology-related patents are Brunelleschi's patent on a boat designed to carry marble up the Arno, issued in Florence in 1421, the Venetian patent law of 1474, and various patent monopolies granted by the English crown between the 15th and 17th centuries. The modern patent, which requires a working model or written description of an invention, dates from the 18th century, first in Britain (1718) and then in the United States (1790), followed closely by France (in both the latter two cases one of the consequences of a revolution). Many other Continental European countries introduced patents during the 19th century, as did Japan. During the 20th century, the use of patent systems became almost universal.

The French patent law of 1791 emphasizes the property right aspect of the patent rather than its use in promoting the useful arts: ‘All new discoveries are the property of the author; to assure the inventor the property and temporary enjoyment of his discovery, there shall be delivered to him a patent for five, ten or fifteen years’ (Ladas and Parry, 2003). In contrast, the Japanese law of 1959 states that its goal is to encourage ‘inventions by promoting their protection and utilization and thereby to contribute to the development of industry’ (JPO, 2006). Patents are enshrined in the US constitution with the sentence ‘Congress shall have power … to promote the progress of science and useful arts by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries’ (Article 1, Section 8, clause 8), which implicitly recognizes both goals of a patent system, namely, reward to the inventor and the promotion of inventive progress.

In 1883 the Paris Convention for the Protection of Industrial Property ensured national treatment of patent applicants from any country that was a party to it. Its most important provision gave applicants who were nationals or residents of one member state the right to file an application in their own country and then, as long as an application was filed in another country that was a member of the treaty within a specified time (now 12 months) to have the date of filing in the home country count as the effective filing date in that other country (the ‘priority date’). This is an important feature of the patent system, and enables worldwide priority to be obtained for an invention originating in any one country, in addition to ensuring that in principle all inventors are treated equally by the system, regardless of the country from which they come.

Legal and administrative

Although the process for granting a patent varies slightly according to the jurisdiction for which protection is desired, the adoption of the agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) in 1995 ensures that it is approximately the same everywhere in the world. This agreement requires its member countries to make patent protection available for any product or process invention in any field of technology with only a few specified exceptions. It also requires them to make the term of protection available for not less than a period of 20 years from the date of filing the patent application.

The World Intellectual Property Organization (WIPO) has almost 200 member states and lists an equivalent number of national patent offices and industrial property offices on its website. In general, the patent right extends only within the border of the jurisdiction that has granted it (usually but not always a country). An important exception is the European system, where it is possible to file a patent application at the European Patent Office (EPO) that will become a set of national patent rights in several European countries at the time of issue (EPO, 2006). A similar situation exists with respect to the African Regional Intellectual Property Organization (ARIPO). The exact number and choice of countries is under control of the applicant. Patents granted by the EPO have the same legal status as patents granted by the various national offices that are party to the European Patent Convention (EPC).

The Patent Cooperation Treaty (PCT) came into existence in 1978, and now has 133 countries as contracting signatories. Any resident or national of a contracting state of the PCT may file an international application under the PCT that specifies the office which should conduct the search. The PCT application serves as an application filed in each designated contracting state. However, in order to obtain patent protection in a particular state, a patent needs to be granted by that state to the claimed invention contained in the international application. The advantage of a PCT application is that fewer searches need to be conducted and the process is therefore less expensive. In fact, 87 per cent of the PCT applications go to one of three patent office for search: those in the United States, Europe, and Japan. Most of the other systems rely on them for the search process and follow them in a number of other areas. Therefore the brief account that follows focuses on these three major systems.

EPO patent grants are issued for inventions that are novel, mark an inventive step, are commercially applicable, and are not excluded from patentability for other reasons (Article 52, EPC). The statutory requirements for patentability in the United States are similar: ‘any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof’ may be patented (35 US Code 101-103 and 112). By itself, this definition does not create a subject matter restriction, although it has long been held that laws of nature, physical phenomena, and abstract ideas are not patentable subject matter.

The origins of the Japanese patent system date back to the Meiji Era (1868–1912). Early patent laws in 1885 and 1899 were modeled on French, US, and then German patent law. In 1899, Japan acceded to the Paris Convention for the protection of industrial property. The patent law was completely revised in 1909, 1921, and 1959. Today, in Japan, patent rights are still protected by the Patent Act of 1959, frequently amended since then (JPO, 2006; Kotabe, 1992). Two important recent changes were the introduction of a product patent in 1976 and the switch to allowing multiple claims in a patent in 1987, both of which have the effect of bringing the system closer to those in Europe and the United States (Nagaoka, 2006).

US patent applications must be filed within one year of the invention's public use or publication – this year is called the ‘grace period’, intended to allow researchers some ability to publish their results as soon as possible. In Europe and other jurisdictions, there is no grace period. Alone among the world's patent offices, the US Patent and Trademark Office operates a ‘first-to-invent’ rather than a ‘first inventor-to-file’ system. In either case, the applicant must be the inventor (except in certain special cases such as death or mental incapacity), but in the US system priority is assigned to the inventor who can show that he reduced the invention to practice first. Also unique to the United States is the fact that patent applications are not made public automatically. Ordinarily patent applications are published 18 months after their priority date, but in the United States an applicant may request exemption from this rule if he files an application on the equivalent invention only at the United States Patent and Trademark Office (USPTO) and in no other jurisdiction.

Many patent offices have a provision for challenging patents following their issue. In the United States, any third party may request re-examination of a patent during its lifetime, although for various reasons related to potential subsequent litigation this opportunity is rarely taken up. In Europe and Japan, robust patent opposition systems with limited time frames operate, and these systems are often used by rival firms as an alternative to more expensive litigation (Hall et al., 2003). In Europe this avenue of challenge is particularly attractive because it is the last opportunity to attack a patent at the European-wide level rather than in individual national courts.

Patents are valuable only if they can be enforced and this fact has a number of implications for their use. First, the ability of the courts to reach the ‘correct’ verdict with respect to infringement and validity will matter; in situations or jurisdictions where there is a great deal of uncertainty about the outcome, and even if both parties agree as to the merits of the case, it may be worth pursuing the issue further or in some cases, reaching a private financial settlement to avoid a random outcome in the courts. Second, the costs of litigation will matter: parties with deep pockets can threaten those with less access to resources, or where the opportunity cost of paying attention to a patent suit is high. On the other hand smaller parties with less to lose can also hold up firms with large sunk investments that they might lose. Finally, the threat of litigation may discourage firms from even entering certain areas, thus providing a disincentive rather than an incentive for R&D. Lerner (1995) documented this phenomenon for biotechnology. The degree to which these kinds of threats matter depends to a great extent on the costs and extent of litigation, both of which tend to be higher in the United States than in many other countries.

Research on patent litigation is difficult because of the data collection problem (it frequently requires accessing the records of courts in several different jurisdictions) but in recent years there have been series of studies of US patent litigation (Moore, 2000; Lanjouw and Schankerman, 2001; Bessen and Meurer, 2005) and at least one of the German system (Cremers, 2004). All of these studies document the fact that litigated patents tend to be the more valuable. The US studies also show that only about five per cent of such suits go to trial, with the remainder being settled before going to trial. They also show that whether patent litigation has increased depends on whether it is measured in aggregate or per patent. That is, the increase in patent litigation has roughly paralleled the increase in patenting, at least in the United States.

Economics of patents

The economic view of patents is that they offer a bargain between society and the inventor: in return for a limited period of exclusivity, the inventor agrees to make his invention public rather than keeping it secret. Therefore, one of the central questions that arises when patents are used as a policy tool to encourage innovation is whether this tool is effective. The theoretical literature in this area produces somewhat ambiguous results. In the simplest case, where a patent corresponds to a single product and knowledge is not cumulative, clearly patents do encourage innovation. In fact, the early theoretical industrial organization literature on patent races seemed to suggest that patents produced too much innovation (Wright, 1983; Reinganum, 1989). However, models that incorporate the cumulative nature of innovation or the fact that production of something new frequently relies on patents held by a large number of entities produce more ambiguous results (Judd, 1985; Bessen and Maskin, 2006).

This question has also proved exceedingly difficult to answer empirically, largely because of the absence of real experiments. Some researchers have looked at historical eras when there were changes to the system and examined the consequences for subsequent innovative activity, measured either by patenting in a jurisdiction not affected by the changes to the system or by invention counts obtained independently (Lerner, 2002; Moser, 2005). A second widely used approach is to survey firms and ask about their patent use (Levin et al. 1987; Cohen et al. 2002; Arundel, 2003). Using these kinds of survey data matched to R&D spending and innovation outcomes, more structural approaches have been pursued by Baldwin, Hanl and Sabourin, 2000; Arora, Ceccagnoli and Cohen, 2003; and Bloom, Van Reenen and Schankerman, 2005, among others.

A few conclusions emerge from this body of work. First, introducing or strengthening a patent system (lengthening the patent term, broadening subject matter coverage or available scope, improving enforcement) unambiguously results in an increase in patenting and also in use of patents as a tool of firm strategy (Lerner, 2002; Hall and Ziedonis, 2001). It is much less clear that these changes result in an increase in innovative activity, although they may redirect such activity toward things that are patentable and are not subject to being kept secret within the firm (Moser, 2005). Sakakibara and Branstetter (2001) studied the effects of expanding patent scope in Japan in 1988 and found that this change to the patent system had a very small effect on R&D activity in Japanese firms.

The survey evidence from a number of countries shows rather conclusively that patents are not among the important means to appropriate returns to innovation, except perhaps in pharmaceuticals (Levin et al., 1987; Cohen et al., 2002; Arundel, 2003). More important means of appropriation are usually superior sales and service, lead time, and secrecy. Patents are usually rated as important only for blocking and defensive purposes. Thus, if there is an increase in innovation due to patents, it is likely to be centred in the pharmaceutical and biotechnology areas, and possibly specialty chemicals. Arora, Ceccagnoli and Cohen (2003) found that increasing the patent premium, which they describe as the difference in payoffs to patented and unpatented inventions, does not increase R&D much except in pharmaceuticals and biotechnology. Using aggregate data across 60 countries for the 1960–90 period, Ginarte and Park (1997) found that the strength of the patent system is positively associated with R&D investment in countries with high median incomes (that is, G-7 and others), but not in lower-income countries.

Recently it has been suggested that the existence and strength of the patent system affects the organization of industry by allowing trade in knowledge, which facilitates the vertical disintegration of

knowledge-based industries and the entry of new firms that possess only intangible assets. The argument is that, by creating a strong property right for the intangible asset, the patent system enables activities that formerly had to be kept within the firm because of secrecy and contracting problems to move out into separate entities. Although limited, research in this area supports this conclusion in the chemical and semiconductor industries (Arora, Fosfuri and Gambardella, 2001; Hall and Ziedonis, 2001).

Economic analysis has also been used to address the optimal design of the patent system. The seminal work in this area was Nordhaus (1969), which considered two policy instruments: the length of the patent term and the breadth of the patent, that is, the range or scope of the inventions covered. The broader the scope of a patent, the larger the number of competing products and processes that will infringe the patent, and the larger the market power of the patentholder. Later work by Gilbert and Shapiro (1990) and Klemperer (1990) built on and extend his method of analysis. Unfortunately, even though all three sets of authors simplified the problem by assuming that a patent corresponds to a product and that there is no uncertainty, the welfare conclusions still turn on assumptions about the nature of the product market and the existence of close substitutes for the patented product. The main conclusion from this line of work is that optimal patent design is likely to depend on the nature of the product market and the technology, which is inconsistent with long-standing practice and policy in most patent systems. Historically, the only important exception to the homogeneous treatment of technologies is the extreme one of excluding some of them (such as pharmaceutical products, medical practices, or disembodied software) completely from the system.

Recent theoretical and empirical work on the patent system has focused on a set of questions that have increased in importance because of the complexity of modern technology and the growth in patent use in sectors that traditionally had paid relatively little attention to them. Briefly described, the new setting is one where a single product involves hundreds of patents, and where one innovation builds directly on many others. Neither feature is really new, but both have assumed increasing importance in a number of technology areas such as information technology and biotechnology. At a theoretical level, Scotchmer (1991; 2005) was the first to identify the problem that cumulative innovation creates for the patent system, in the sense that it is difficult if not impossible to set incentives at the correct level for both the first and subsequent innovators.

When development of an innovative product requires multiple patent inputs, Heller and Eisenberg (1998) have argued forcefully that the licensing solution may fail because of transactions costs if a large number of patentholders are involved. One consequence of this fragmentation threat may be increased defensive patenting by the product developer. Empirical evidence for this proposition has been provided by Ziedonis (2004) in the context of the semiconductor industry.

Using patent data

Researchers into the economics of innovation and technical change frequently find themselves in need of measures of innovative output or success, preferably classified by sector or technology. Many would also like measures of knowledge flow between individuals and firms, given the potential importance of spillovers in the production of knowledge. In recent years, the growth in importance of the knowledge economy worldwide has lead to an increased interest in such measures. As was noted long ago by such pioneers in the field as Schmookler (1966), patent data can be very helpful in constructing them. The

primary advantage of patent data is that they are available over a wide range of countries and years, for detailed technology classes, and they contain information on inventor, geographic area, and owner (if there is one other than the inventor). Together, these data provide information on the locus and type of newly created knowledge. The second advantage is that they provide information on links between different quanta of knowledge via the citations to other patents and non-patent documents that they contain (see Jaffe, Trajtenberg and Fogarty, 2000, for further justification of the use of patent citations to model knowledge flow and for the limitations of the measure). With the possible exception of data on scientific paper publication, no other data source comes even close to providing this level and quantity of information about the creation and dissemination of new knowledge.

The use of patent data as a proxy for innovation output in the economic analysis of technological change dates back to the path-breaking analyses of Schmookler (1966) and Scherer (1965). An overview is given in OECD (1994). The availability of information from the US patent office in machine-readable form in the late 1970s enabled research using these data with much larger samples of firms; the resulting early work is reported in Griliches (1984) and then surveyed by Griliches, Pakes and Hall (1987) and Griliches (1990). At the same time, Schankerman and Pakes (1986) pioneered the use of renewal data from the patent offices of several European countries to estimate the value distribution of patents; at the time, such data were not available for the United States owing to the absence of renewal fees in that country.

The results of this early work were, first, to demonstrate a strong correlation between the size of a firm's R&D effort and its patenting output, with little evidence that smaller programmes and firms yielded more output per unit of input, once selection was controlled for. Second, the renewal data, along with pieces of evidence from some specific sectors such as pharmaceuticals (Grabowski and Vernon, 1994) and medical devices (Trajtenberg, 1990), suggested that the value distribution of patents was very skewed, with a few patents worth a lot and most patents worth nothing. Third, there was little evidence that patent outcomes added much predictive power to sales, profits, or market value equations in the presence of R&D expenditure (Griliches, Hall and Pakes, 1991).

With the advent of the personal computer and the increased access to computing power on the part of economic researchers, it became feasible to construct data-sets containing patent citations in the late 1980s, leading to a second wave of research. Similarly to a research paper, the patent document contains a set of references to earlier patents and scientific literature on which it builds; a typical patent referenced approximate five earlier patents in the 1980s, and an increasing number as time passes. These citations can be used to give an indication of the impact of a patented invention on the inventions in subsequent patents and to investigate an additional set of questions related to the flow of knowledge across time, space and organizational boundaries. However, it is important to note that differences exist in citation practice between the US and other patent systems (see Webb et al., 2005, and Harhoff, Hoisl and Webb, 2006, for further discussion of this issue), and most of the validation of this methodology has been done using US data.

Researchers have used these data to explore questions involving spatial spillovers (for example, Jaffe, Trajtenberg and Henderson, 1993), knowledge flows among firms in a research consortium (for example, Ziedonis, Ziedonis and Silverman, 1998), and spillovers from public research (for example, Jaffe and Trajtenberg, 1996; Jaffe and Lerner, 2001). In using citations as evidence of spillovers, or at

least knowledge flows, from cited inventors to citing inventors, it is clearly a problem that many of the citations are added by the inventor's patent attorney or the patent examiner, and may represent inventions that were wholly unknown to the citing inventor. On the other hand, in using citations received by a patent as an indication of that patent's importance, impact or even economic value, the citations that are identified by parties other than the citing inventor may well convey valuable information about the size of the technological ‘footprint’ of the cited patent.

Beginning with Trajtenberg's (1990) study of the welfare impact of CAT scanners, there are by now a number of studies that ‘validate’ the use of citations data to measure economic impact, by showing that citations are correlated with non-patent-based measures of value. Hall, Jaffe and Trajtenberg (2005) investigated the use of citations as an indicator of private invention value in a large sample of publicly traded US manufacturing firms and confirmed that, although patent yield conveys little information beyond that conveyed by R&D spending, citation-weighted patents are strongly related to market value in a nonlinear way, with very highly cited patents worth a great deal more than those with less than average citation.

Recent work by Lanjouw and Schankerman (2004) also uses citations, together with other attributes of the patent (number of claims and number of different countries in which an invention is patented) as a proxy for patent quality. They find that a patent ‘quality’ measure based on these multiple indicators has significant power in predicting which patents will be renewed and which will be litigated. They infer from this that these quality measures are significantly associated with the private value of patents.

Similarly, Harhoff et al. (1999) surveyed 962 holders of German patents that had a priority date of 1977, asking them to estimate at what price they would have been willing to sell the patent right in 1980, about three years after the date at which the German patent was filed. They find both that more valuable patents are more likely to be renewed to full term and that the estimated value is correlated with subsequent citations to that patent. As in Hall, Jaffe and Trajtenberg (2005, p. 23), the most highly cited patents are very valuable, ‘with a single U.S. citation implying on average more than $1 million of economic value’.

See Also

- intellectual property

Bibliography

Arora, A., Ceccagnoli, M. and Cohen, W. 2003. R&D and the patent premium. Working Paper No. 9431. Cambridge, MA: NBER.

Arora, A., Fosfuri, A. and Gambardella, A. 2001. *Markets for Technology: The Economics of Innovation and Corporate Strategy*. Cambridge, MA: MIT Press.

Arundel, A. 2003. *Patents in the Knowledge-Based Economy: Report of the Know Survey*. Maastricht: MERIT, University of Maastricht.

Baldwin, J.R., Hanl, P. and Sabourin, D. 2000. Determinants of innovative activity in Canadian manufacturing firms: the role of intellectual property rights. Working Paper No. 122. Ottawa: Statistics Canada.

Bessen, J. and Maskin, E. 2006. Sequential innovation, patents, and imitation. Working paper, Boston University School of Law and Princeton University.

Bessen, J. and Meurer, M. 2005. The patent litigation explosion. Law and Economics Working Paper No. 05-18, Boston University School of Law.

Bloom, N., Van Reenen, J. and Schankerman, M. 2005. Identifying technology spillovers and product market rivalry. Discussion Paper No. 3916. London: CEPR.

Cohen, W.M., Goto, A., Nagata, A., Nelson, R.R. and Walsh, J.P. 2002. R&D spillovers, patents and the incentives to innovate in Japan and the United States. *Research Policy* 31, 1349–67.

Cremers, K. 2004. Determinants of patent litigation in Germany. Discussion Paper No. 04-72. Mannheim: Centre for European Economic Research (ZEW).

EPO (European Patent Office). 2006. Online. Available at http://www.european-patent-office.org/index.en.php, accessed 4 January 2007.

Federal Trade Commission (FTC). 2003. *To Promote Innovation – The Proper Balance of Competition and Patent Law and Policy*. Washington, DC: FTC.

Gilbert, R. and Shapiro, C. 1990. Optimal patent length and breadth. *RAND Journal of Economics* 21, 106–12.

Ginarte, J.C. and Park, W. 1997. Determinants of patent rights: a cross-national study. *Research Policy* 26, 283–301.

Grabowski, H. and Vernon, J.M. 1994. Returns to R&D on new drug introductions in the 1980s. *Journal of Health Economics* 13, 383–406.

Griliches, Z., ed. 1984. *R&D, Patents and Productivity*. Chicago: University of Chicago Press.

Griliches, Z. 1990. Patent statistics as economic indicators: a survey. *Journal of Economic Literature* 28, 1661–707.

Griliches, Z., Hall, B.H. and Pakes, A. 1991. R&D, patents, and market value revisited: is there a second (technological opportunity) factor? *Economics of Innovation and New Technology* 1(3), 183–202.

Griliches, Z., Pakes, A. and Hall, B.H. 1987. The value of patents as indicators of inventive activity. In *Economic Policy and Technological Performance*, ed. P. Dasgupta and P. Stoneman. Cambridge: Cambridge University Press.

Hall, B.H., Graham, S., Harhoff, D. and Mowery, D.C. 2003. Prospects for improving U.S. patent quality via postgrant opposition. *Innovation Policy and the Economy* 4, 115–43.

Hall, B.H., Jaffe, A.B. and Trajtenberg, M. 2005. Market value and patent citations. *RAND Journal of Economics* 36, 16–38.

Hall, B.H. and Ziedonis, R. 2001. The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979–1995. *RAND Journal of Economics* 32, 101–28.

Harhoff, D., Hoisl, K. and Webb, C. 2006. European patent citations – how to count and how to interpret them? Working paper, University of Munich, CEPR and OECD.

Harhoff, D., Narin, F., Scherer, F.M. and Vopel, K. 1999. Citation frequency and the value of patented inventions. *Review of Economics and Statistics* 81, 511–15.

Heller, M.A. and Eisenberg, R.S. 1998. Can patents deter innovation? The anticommons in biomedical research. *Science* 280, 698–701.

Jaffe, A.B. and Lerner, J. 2001. Reinventing public R&D: patent policy and the commercialization of national laboratory technologies. *RAND Journal of Economics* 32, 167–98.

Jaffe, A.B. and Trajtenberg, M. 1996. Flows of knowledge from universities and federal labs: modeling the flow of patent citations over time and across institutional and geographic boundaries. *Proceedings of the National Academy of Sciences* 93, 12671–7.

Jaffe, A.B., Trajtenberg, M. and Fogarty, M. 2000. The meaning of patent citations: report of the NBER/Case Western Reserve survey of patentees. Working Paper No. 763. Cambridge, MA: NBER.

Jaffe, A.B., Trajtenberg, M. and Henderson, R. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108, 577–98.

JPO (Japanese Patent Office). 2006. A history of system of industrial property rights. Online. Available at http://www.deux.jpo.go.jp/cgi/search.cgi?query=history&lang=en&root=short, accessed 19 December 2006.

Judd, K.L. 1985. On the performance of patents. *Econometrica* 53, 567–95.

Klemperer, P. 1990. How broad should the scope of patent protection be? *RAND Journal of Economics* 21, 113–30.

Kotabe, M. 1992. A comparative study of U.S. and Japanese patent systems. *Journal of International Business Studies* 23, 147–68.

Ladas and Parry LLP. 2003. A brief history of the patent law of the United States. Online. Available at http://www.ladas.com/Patents/USPatentHistory.html, accessed 19 December 2006.

Lanjouw, J.O. and Schankerman, M. 2001. Characteristics of Patent litigation: a window on competition. *RAND Journal of Economics* 32, 129–51.

Lanjouw, J.O. and Schankerman, M. 2004. Patent quality and research productivity: measuring innovation with multiple indicators. *Economic Journal* 114, 441–65.

Lerner, J. 1995. Patenting in the shadow of competitors. *Journal of Law and Economics* 38, 463–95.

Lerner, J. 2002. Patent policy shifts and innovation over 150 years. *American Economic Review* 92, 221–5.

Levin, R.C., Klevorick, A.K., Nelson, R.R. and Winter, S.G. 1987. Appropriating the returns from industrial research and development. *Brookings Papers on Economic Activity* 1987(3), 783–831.

Moore, K.A. 2000. Judges, Juries, and patent cases – an empirical peek inside the black box. *Michigan Law Review* 99(281), 365–409.

Moser, P. 2005. How do patent laws influence innovation? Evidence from nineteenth-century world fairs. *American Economic Review* 95, 1214–36.

Nagaoka, S. 2006. Reform of patent system in Japan and challenges. Paper presented at the conference on 21st Century Innovation Systems for Japan and the United States: Lessons from a Decade of Change. Tokyo: Institute of Innovation Research, Hitotsubashi University.

Nordhaus, W. 1969. *Invention, Growth and Welfare: A Theoretical Treatment of Technological Change*. Cambridge, MA: MIT Press.

OECD. 1994. *The Measurement of Scientific and Technological Activities: Using Patent Data as Science and Technology Indicators*. Paris: OECD.

Reinganum, J.F. 1989. The Timing of innovation: research, development, and diffusion. In *Handbook of Industrial Organization*, vol. 1, ed. R. Schmalensee and R.D. Willig. Amsterdam: North-Holland.

Sakakibara, M. and Branstetter, L. 2001. Do stronger patents induce more innovation? Evidence from the 1988 Japanese patent law reforms. *RAND Journal of Economics* 32, 77–100.

Schankerman, M. and Pakes, A. 1986. Estimates of the value of patent rights in European countries during the post-1950 period. *Economic Journal* 96, 1052–76.

Scherer, F.M. 1965. Firm size, market structure, opportunity, and the output of patented innovations. *American Economic Review* 55, 1097–123.

Schmookler, J. 1966. *Invention and Economic Growth*. Cambridge, MA: Harvard University Press.

Scotchmer, S. 1991. Standing on the shoulders of giants. *Journal of Economic Perspectives* 5(1), 29–41.

Scotchmer, S. 2005. *Innovation and Incentives*. Cambridge, MA: MIT Press.

Trajtenberg, M. 1990. A penny for your quotes: patent citation and the value of innovations. *RAND Journal of Economics* 21, 172–87.

USPTO (United States Patent and Trademark Office). Online. Available at http://www.uspto.gov, accessed 4 January 2007.

Webb, C., Dernis, H., Harhoff, D. and Hoisl, K. 2005. Analysing European and international patent citations: a set of EPO patent database building blocks. STI Working Paper No. 2005/9. Paris: OECD.

WIPO (World Intellectual Property Organization). Online. Available at http://www.wipo.int/patentscope/en/, accessed 4 January 2007.

Wright, B.D. 1983. The economics of invention incentives: patents, prizes, and research contracts. *American Economic Review* 73, 691–707.

Ziedonis, R.H. 2004. Don't fence me in: fragmented markets for technology and the patent acquisition strategies of firms. *Management Science* 50, 804–20.

Ziedonis, A.A., Ziedonis, R.H. and Silverman, B.S. 1998. Research consortia and the diffusion of technological knowledge: insights from SEMATECH. Working paper, University of Michigan and University of Toronto.

How to cite this article

Hall, Bronwyn H. "patents." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven

N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000039> doi:10.1057/9780230226203.1254

The New Palgrave Dictionary of Economics Online

path dependence and occupations

Maristella Botticini and Zvi Eckstein

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Path dependence in occupations refers to the observed occupational distribution in a population or in a sub-population at a point in time that depends on changes that occurred years or centuries earlier. Path dependence in occupations can be the outcome of the cumulative concentration of certain productive activities in specific regions over time, it can emerge through the effect of parental income or wealth on offspring's occupations and incomes, or it can be the outcome of group effects. Some historical cases are selected to illustrate the various mechanisms through which path dependence in occupations can emerge or disappear.

Keywords

affirmative action; black–white labour market inequality (USA); Civil Rights Act 1964 (USA); commercial and trade diasporas; economic geography; ghettoes; Great Migration (USA); high school movement (USA); inequality (explanations); innovation; intergenerational income mobility; intergenerational occupational mobility; intergenerational skill transfers; international migration; Jewish economic history; Kuznets, S.; labour market discrimination; manufacturing belt (USA); migration and labour markets; occupational choice; occupational structure; path dependence; path dependence in occupations; poverty traps; QWERTY keyboard; religion and economic development; residential mobility; rural-urban migration; urbanization; women's work

Article

Path dependence in occupations can be interpreted to mean that the observed occupational distribution in a population or in a sub-population at a point in time depends on changes that occurred years or centuries earlier (for example, a war that siphons off certain types of workers, the enactment of anti-discriminatory labour practices, a technological invention which is not gender- or race-neutral). This definition is consistent with the notion of path dependence suggested in the economic history literature

by David (1985) with the example of the ‘standard QWERTY’ keyboard. Under this definition one can include both the cases in which particular innovations in the economy have permanent consequences and those instances in which particular shocks are not self-correcting, so that they remain permanent in the absence of some countervailing change.

To show that there is path dependence in occupations, one has to describe the exact sequencing of events related to the initial change and show that they had a permanent effect on the occupational choice and distribution observed later. In other words, one has to show that, at a given point in time, multiple occupational distributions were available for selection, and theory is unable to predict or explain the occupational structure that will be chosen. Then, a change occurs and an occupational distribution is favoured over competing ones. Finally, the selected occupational structure capitalizes on initial advantage and is stably reproduced over time.

The economics literature identifies a number of possible sources of path dependence (see for discussions Arthur, 1989; David, 1994; Liebowitz and Margolis, 1995; Blume and Durlauf, 2005). For example, the economic geography literature explains path dependence in occupations as the outcome of the cumulative concentration of certain productive activities in specific regions over time (for example, Krugman, 1991b; Fujita, Krugman and Venables, 1999). This literature highlights the potentially big impact of increasing returns and cumulative processes, which in turn can make the role of historical accidents decisive. Small changes in the parameters of the economy may have large effects. For example, if transportation costs, economies of scale, and the share of non-agricultural goods in expenditure cross a critical threshold, population may start to concentrate and regions to diverge; once started this process will feed on itself.

However, increasing returns are not necessary for path dependent processes (Bowles and Gintis, 2002). For example, in models of intergenerational mobility where individual-level characteristics matter (for example, Becker and Tomes, 1979; Loury, 1981; Banerjee and Newman, 1991; 1993; Galor and Zeira, 1993; Eckstein and Zilcha, 1994; Mookherjee and Ray, 2002; 2003), the existence or absence of path dependence in relative economic status across generations emerges through the effect of parental income or wealth on offspring's occupations and incomes.

In contrast, starting from the seminal work of Shelling (1971), in membership models an individual's economic choices are influenced not only by his or her traits but also by characteristics of the group of individuals with whom the person typically interacts (see Durlauf, 2006, for a discussion of these models and related empirical literature). Groups may differ in average level of schooling, cognitive functioning, occupational structure and wealth level. Some groups are exogenously determined, for example by ethnicity or gender. Other groups are endogenously determined. For example, individuals may be strongly influenced by groups such as residential neighbourhood, the schools attended, and the co-workers at various jobs. Group effects on economic success are well documented and may arise for a number of reasons, including discrimination, conformist effects on behaviour, differential access to information, and complementarities in production.

An exhaustive survey of historical and contemporary examples of path dependence in occupations is beyond the scope of this article. Instead, we selected some examples which illustrate the various mechanisms discussed above through which path dependence in occupations can emerge or disappear.

Jewish economic history in the past two millennia

At the beginning of the first millennium, an exogenous change in the religious and social norm that defined Judaism occurred as a result of the shift in the leadership within the Jewish community. Before the destruction of the Temple in Jerusalem in 70 ce, the Jewish population in Eretz Israel, which consisted mainly of farmers, was segmented in many religious groups. After the destruction of the Temple, many Jewish sects disappeared, whereas the Pharisees became the dominant group. They replaced sacrifices with the study of the Torah in the synagogue. The transformation of the religion created the need for the devoted Jews to be literate and to educate their children. In about 200 ce, the transformation of Judaism reached its full-fledged stage with the compilation of the Mishna. Also, a new social norm came to prevail according to which an illiterate Jewish individual was considered an outcast in the community.

Despite education being very costly and ‘useless’ in production for farmers, religious instruction and primary education became more and more widespread among the Jewish communities in Eretz Israel and Babylonia from the second to the seventh century. The spread of literacy among the Jewish rural population is even more impressive when compared with the literacy rates of the non-Jewish rural population in the same period. In the Roman, Byzantine, Christian and Persian worlds there was no mandatory primary education, and the non-Jewish rural population was almost entirely illiterate.

Before 400 ce almost all Jews in the three main centres of Jewish life in the classical period – Palestine, Babylonia, and Egypt – were farmers, exactly like the rest of the population. The transition away from agriculture into crafts, trade, and moneylending started in the Talmudic period (200–500 ce), especially in Babylon. In the fifth and sixth centuries, some literate Jews abandoning agriculture moved to the towns and became small shopkeepers, craftsmen and artisans. However, given the stagnant economies in the late Roman, early Byzantine and Persian empires in the fourth to the seventh centuries, the growing number of literate Jewish farmers could not find skilled occupations in the existing cities at that time and many of them converted out of Judaism. World Jewry was reduced from about 4.5 million in 70 ce to about 1.5 million in 600 ce, with 80 per cent living in Babylonia.

But in the eighth and ninth centuries, another exogenous event occurred: massive urbanization in the newly established Muslim empire under the Abbasid caliphate vastly increased the demand for urban, skilled occupations. The *literate* Jewish rural population in Iraq and later in the Abbasid empire as a whole moved to urban centres, abandoned agriculture, and became engaged in a wide range of crafts, local and long-distance trade, moneylending, tax-farming and the medical profession. This occupational transition took about 150 years, and by 900 ce almost all Jews in Iraq, Persia, Syria and Egypt, were engaged in urban occupations. In contrast, most non-Jews remained farmers, even though they could engage in any occupation in the regions under Muslim rule. These two facts identify the educational reform in Judaism around 200 ce as the key factor for the occupational transition of the Jewish people (Botticini and Eckstein, 2005).

Judaism, with its costly religious norm regarding education, can thrive in the long run only if the Jews can find occupations in which their earnings significantly gain from literacy (Botticini and Eckstein, 2006). The voluntary diaspora of the Jews to western Europe during the tenth to the 13th centuries, to eastern Europe in the 16th and 17th centuries, and then worldwide supports this argument. Other minorities within the Muslim empire under the Abbasid caliphate did not migrate to western Europe even though no prohibitions prevented them from doing so. The distinctive engine of the Jewish migrations to the West was the incentive to maximize the returns to their investment in education.

Hence, these two facts identify the link between the ‘historical accident’ and the voluntary diaspora of the Jews in search of urban, high-skill and high-income occupations.

The large Jewish population of Iraq and Iran, which amounted to about 800,000 in 1250, almost disappeared when the Mongol invasions brought the Near East back to a subsistence farming economy. In contrast, the small Jewish population in Europe survived, kept its literacy and educational distinctiveness, and through urban and skilled occupations reached high standards of living.

These urban, skilled occupations remained the distinctive mark of the Jewish people throughout their history, as clearly highlighted by the data provided by Kuznets (1960): in the countries which hosted the largest Jewish communities in the early 20th century (countries in eastern Europe, Russia, the United States and Canada), between 96 and 99 per cent of the Jews were engaged in non-agricultural occupations even though no restrictions prevented them from being farmers. Chiswick (2005) documents the same occupational selection of the Jewish population in the United States as late as the year 2000. For example, about 53 per cent of adult Jewish men are engaged in professions such as law, medicine, and academia, whereas the percentage for white non-Jewish men is about 20 per cent. In contrast, only six per cent of adult Jewish men are employed in the construction, transportation, and production sectors in comparison with about 39 per cent of adult non-Jewish men.

Jewish economic history fits very well the multiple features of path dependence outlined in the introduction. On the one hand, two exogenous changes (the transformation of the religious norm in the first and second centuries ce and the urbanization in the Muslim empire in the eighth and ninth centuries) created a permanent effect on the occupational distribution among the Jews. On the other hand, the mechanisms through which these changes worked to affect the occupational structure of the Jews in the long run were twofold: the intergenerational transmission of skills and literacy from parents to children, and the peer pressure (social penalty) that the Jewish communities imposed on those who did not invest in their children's education.

Commercial and trade diasporas

As membership models would predict, ethnic groups can influence the occupational distribution of immigrants in a country and create occupational clustering by ethnicity. One of the most visible examples of this occupational clustering is offered by the so-called commercial and trade diasporas.

A diaspora is any ethnic group without a territorial base within a given polity, and whose social, economic and political networks cross the borders of nation states. In particular, trade and commercial diasporas are those diasporas whose members specialize in trade and commercial activities or, more generally, in urban, skilled jobs. Historical examples include the Jews in the last two millennia, the Parsi (Zoroastrian) diaspora from Iran, the Huguenots in early modern and modern western Europe, the Armenians, the Greeks of the Ottoman Empire, the Germans throughout eastern Europe in modern times, the Chinese in many areas of south-east Asia from the 15th to the 20th century, the Indian middleman minorities of east Africa and Malaya, the Pakistanis in Great Britain, and the Lebanese Christians in 18th-century Egypt and contemporary west Africa (Botticini, 2003).

Commercial and trade diasporas – indeed, diasporas in general – have been characterized by strong linguistic skills, often including the ability to speak and write in both their own and alien languages. This enabled members of a diaspora to maintain communication networks within the group and to use alien languages for practical purposes. Maintaining the common original language is one of the means to

enhance the organization of a diaspora. Others mechanisms include the establishment of communal institutions, such as the commercial coalitions among the Jews in the Mediterranean in the high Middle Ages (Greif, 1989) or the Chinese societies known as *Houei*; the development of a common set of commercial laws or norms whose enforcement is delegated to courts within the communities; and strong endogamous marriage strategies.

In some cases, exogenous changes have created or reinforced occupational selection among ethnic or religious groups. For example, it has been often argued that legal prohibitions and the exclusion of Jews from guild membership in medieval and early modern Europe would account for their occupational selection into moneylending and the medical profession. Similarly, it has been pointed out that, after the revocation of the Edict of Nantes by King Louis XIV in 1685 that made Protestantism illegal, many Huguenots (French Protestants) emigrated to Ireland, England, Prussia, and America, where they contributed to the development of industries and trades. The Agricultural Law of 1870 in Indonesia against land ownership by ethnic Chinese has been cited to explain the exclusion of the Chinese diaspora from farming and agricultural activities.

In other instances, the occupational distribution was altered by rulers who substituted one diaspora for another if they perceived the change to be advantageous for them. Thus, in the Ottoman Empire, Catholic Levantines, who held the leadership in crafts and trade in the 15th century, were replaced by the Jews in the 16th and 17th centuries, followed by the Greeks until the beginning of the 19th century and Armenians during the 19th century.

Geography also played a role in the occupational specialization of some ethnic groups. With the European geographical expansions and the establishment of colonial rule in south-east Asia and west and east Africa during the 19th and 20th centuries, Lebanese Christians, Chinese, and Indians have contributed to the establishment of commercial economies in the European colonial empires.

The manufacturing belt in the United States

The establishment and remarkable persistence of the manufacturing belt in the United States is one of the most prominent examples of geographic concentration which in turn affected the occupational distribution of the US population.

Early in the history of the United States, when most of the population was engaged in agriculture, when transportation costs were high, and when manufacturing was characterized by few economies of scale, no concentration could occur. When the United States started to industrialize, manufacturing first developed in regions where most of the agricultural population outside the South was located. The manufacturing belt developed in the second half of the 19th century when economies of scale in manufacturing increased, transportation costs fell, and the share of the population in non-agricultural occupations rose. The initial advantage of the manufacturing belt was locked in, leading the bulk of US manufacturing to be concentrated in a relatively small part of the north-east and the eastern part of the Midwest. It persisted even as the centre of gravity of agricultural and mineral production shifted to the West. As late as 1957, the manufacturing belt still contained 64 per cent of US manufacturing employment (Krugman, 1991a).

Intergenerational occupational mobility in Britain and the United States since 1850

Unlike today, the United States in the 19th century was ‘exceptional’ in the occupational mobility experienced by its population (as well as in its geographic mobility) compared with Europe. As documented by Long and Ferrie (2005), this contrast is even more striking when 19th-century United States is compared with 19th-century Britain – the country with which it shared legal traditions and property rights systems and sources of labour, capital, and technology.

Differences have been attributed to a number of factors. First, the absence of feudalism and of strong craft guilds has been put forth as one reason for the higher occupational mobility in the United States. Second, at least some of the high mobility in 19th-century United States may result from it being at an earlier stage of development than 19th-century Britain, so its farm sector was relatively larger.

Third, the United States provided considerably more public education than Britain in the middle of the 19th century: the primary school enrolment rate was one and a half times greater in the United States than in Britain. The US educational system in the second half of the 19th century, though less extensive at the secondary and post-secondary levels than European systems, was considerably more egalitarian (Goldin and Katz, 2003). To the extent that intergenerational mobility is greater where fewer parents are wealth-constrained, superior mobility in the United States may well have been a consequence of its educational system, which provided a public alternative to a private education that was outside the reach of many families.

Fourth, residential mobility to places that were growing more rapidly than others may have provided an alternative to direct investment in human capital. Cities (such as Chicago) sprang up initially to provide services demanded as the frontier expanded. Though US labour markets in the North were well-integrated at the regional level by the middle of the 19th century, differences across smaller units of geography may have continued to present opportunities for ‘locational arbitrage’ that provided a route to occupational change through the start of the 20th century (Long and Ferrie, 2005).

The feminization of teaching and clerical work in the United States

Teaching profession

Today in the United States the vast majority of elementary and secondary teachers are women. In 2000, the female proportion among teachers was 76 per cent. Much earlier in American history, however, this was not the case. The feminization of teaching occurred over the course of the 19th century and continued throughout the 20th century. Two exogenous factors changed the social norm and attitude towards female teachers in the United States and, therefore, significantly contributed to the feminization of teaching: (*a*) the ethnic, national, and cultural identity of the European settlers who established their communities in the Northern, Midwestern, and Southern states, and (*b*) the wars (especially the American Civil War and the First World War).

Relatively early in the 19th century, women came to dominate teaching in New England through the establishment of two educational institutions: the so-called ‘dame schools’ and a two-tier system divided into winter and summer sessions. The ‘dame schools’ were an educational institution imported by British settlers, in which women taught very young children as they were considered the natural carers for these children. The division into winter and summer sessions reinforced this gender-specific assignment of teachers to pupils according to age. As winter sessions were geared towards older boys,

male teachers were considered to have greater human capital and skills to enforce discipline among them. Female teachers were considered better equipped to teach summer sessions attended by younger children. As population spread westward in the North, the female percentage in teaching increased in these states (Carter and Margo, 2007).

In contrast, because of the different ethnic and national background of the European settlers who established themselves in the US South, neither ‘dame schools’ nor the two-tier system were developed and the percentage of female teachers remained much lower there until the Civil War. But even within the North itself, the role of culture and institutions in affecting the gender distribution in the teaching profession is illustrated by regional variation. In Illinois counties where the settlers were mainly Yankees, female teachers were quite common, whereas in those counties where the settlers were mainly Southerners, male teachers predominated (Carter and Margo, 2007).

As Perlmann and Margo (2001) have shown, the American Civil War significantly contributed to the feminization of teaching. In 13 Northern and Midwestern states, the average share of female teachers rose from about 57 per cent in 1860 to 67 per cent in 1865 and 79 per cent in 1915. During the war women took jobs in teaching, substituting for men who were at war. When the war ended, there was some mean reversion, but not back to the original equilibrium.

The entry of many women into teaching during the Civil War changed the social norm and attitudes toward female teachers by making the bias against them gradually fade. In the earlier decades, the argument against hiring female teachers had been that, especially in winter classes when adult boys attended school, women lacked the skills to discipline these students. However, the entry of women in the teaching profession during the war to substitute for the male teachers gave them the opportunity to show that they could be as effective as their male colleagues in teaching and maintaining the discipline among students. This changed the social norm and attitude toward hiring female teachers, which increased the feminization of teaching in both the Northern states and in the South, where the share of female teachers reached unprecedented levels, rising from about 35 per cent in 1875 to 73 per cent in 1915 (Perlmann and Margo, 2001, p. 169).

The First War World had a similar effect on the selection of women into the teaching profession, although on a smaller scale. After the Second World War, women entered many other occupations and professions. Yet the predominance of female teachers in primary and secondary schools holds to the present day.

Clerical work

In 1870, fewer than three per cent of all clerical employees were women. In 1930, women made up over half (52.5 per cent) of the total clerical workforce, and today the clerical sector is one of the major employers of women. The most rapid increases occurred in two decades, 1880–90 and 1910–20, as the outcome of two exogenous shocks on the demand side of the labour market coupled with a profound transformation on the supply side of the same market.

On the demand side, Rotella (1981) has argued that the adoption and diffusion of the typewriter in the 1880s, the growth of large firms and the expansion of the government sector in the 1910s created a huge demand for clerical work. Specifically, the diffusion of the typewriter made the skills required of clerical labour no longer firm-specific, as it had been when employers preferred to hire male workers who were

expected to have a long working life within the firm. With the development of the modern, mechanized office, employers could afford to hire young, educated women who had high expected turnover and who desired clean, high-status employment. Later, in the 1910s, the growth of large firms and the expansion of the government sector through regulation and tax laws greatly increased the demand for information and information processing within firms and government offices. Again, this shift in demand was not gender neutral: it favoured women, and women came to dominate office work, basically after about 1910.

On the supply side, Goldin (1986; 1990) has shown that the huge increase in high school attendance around the turn of the 20th century – the so-called ‘high school movement’ – dramatically increased the supply of young, educated women in the labour market. These women offered a relatively cheaper and easier to monitor labour force.

The occupational transition of African-Americans in the 20th century

After the American Civil War, a steady stream of African-Americans moved out of the South to the North. It has been estimated that from 1870 to 1910 about 535,000 blacks emigrated from the South on the net as the outcome of the large wage differentials between the North and the South and of the increased human capital acquired by the first generations of blacks after the Emancipation (Margo, 1990, ch. 7). This migration, though, did not have a huge impact on the overall occupational and residential distribution of African-Americans. In fact, in 1900 approximately 90 per cent of the blacks still lived in the South, and the majority of them worked in agriculture and were very poor.

In contrast, from 1910 to 1950 the Great Migration brought about 3.5 million African-American people out of the South mainly to the urban North. Even when migration occurred to rural areas in the North, it invariably involved a shift out of agriculture. The Great Migration represented a watershed in African-American economic history and implied a profound and permanent transformation of the occupational distribution of the blacks in the United States.

The relevant exogenous shocks that fuelled both the Great Migration and the permanent change in the occupational distribution of the blacks were the two world wars and a combination of government policies.

First, quotas set by the US government on foreign immigration after the First World War greatly accelerated the outmigration of blacks from the South in the 1920s, as blacks were substitutes for the foreign-born immigrants (Collins, 1997).

Second, the Second World War was an even bigger exogenous shock. When the United States entered the war, demand for white workers in the war-industry sector increased at the same time as the military was siphoning off potential workers. US employers were faced with a tough choice: either to follow the prevailing taste for discrimination among employers and white workers and the social norm against hiring black workers, or to expand production and to gain profits by hiring black workers.

The enforcement by President Roosevelt of the anti-discriminatory policy amongst defence contractors through the Fair Employment Practice Committee (FEPC) established in 1941 was the exogenous change in government policy that helped employers choose the second option and hire black workers despite the prevailing taste for discrimination (Collins, 2001). The impact of this government intervention was twofold. It made defence contractors hire black male workers who otherwise would not

have been hired because of the hostility of white male workers towards hiring fellow black workers. At the same time, it started changing the social norm and attitude against hiring black workers in other firms and industries in those instances when the enforcement of the anti-discriminatory policy among the defence contractors sector had spillover effects on other firms' hiring practices.

The combination of the two exogenous shocks – the Second World War and the establishment of the Fair Employment Practice Committee – had a large impact on the occupational and residential transition of African-Americans. Between 1940 and 1950 the proportion of black male workers classified as operatives (semi-skilled) rose from 12.6 to 21.4 per cent, and the proportion in manufacturing industries rose from 16.2 to 23.9 per cent (Collins, 2000). This transition into manufacturing and war-related industries greatly contributed to the economic progress of blacks, as the data on the substantial wage premium these workers earned indicate.

A similar effect occurred ten years later as the outcome of another major change in government policy. The *Brown vs Board of Education* Supreme Court's decision in 1954, which invalidated school segregation in the US South, the enactment of Title VII of the 1964 Civil Rights Act, which forbade discrimination in employment, the establishment of the Office of Federal Contract Compliance (OFCC), which monitored the anti-discrimination and affirmative action responsibilities of government contractors, and the passage of the Voting Rights Act of 1965 were the most famous among several government policies designed to eliminate discrimination against blacks. Donohue and Heckman (1991) show that a significant portion of the sustained improvement in the labour market status of black males from 1965 to 1975 (especially in the US South) was the outcome of these changes in government policies.

Poverty traps

Intertemporal social interactions (that is, social interactions in which choices made at one time affect others made later) can create path dependence in occupations through a variety of mechanisms. Role models and peer group effect models are two examples of these mechanisms. Suppose, as role models do, that the decision to attend college by a young adult depends on the percentage of college graduates among adults in his community. Then two communities, one where the adults are all college graduates and the other where none are, can converge to different levels of college attendance in a steady state, leading to path dependence in occupational and economic segregation across long time periods and generations.

The persistence of ghettoes and poverty traps are the two most visible examples of the intertemporal effect of group membership on individual outcomes (Bowles, Durlauf and Hoff, 2006). Poverty traps are situations where the evolution of individual wealth is governed by a path-dependent process such that, depending on initial conditions, otherwise identical individuals or groups (ethnic, linguistic, religious) may remain for long periods of time 'locked into' poverty. The key characteristic of a poverty trap is that the 'good' and 'bad' outcomes are self-enforcing, so that small interventions or chance events will not alter the long-term outcome. Recent evidence of the persistence of income differences between races, even after some of the structural determinants of inequality (such as colonialism, inequalities of educational opportunity, and de jure segregation) have been removed, point to the importance of historical contingency and 'lock-in effects' in the process that generates inequality (Loury, 2002;

Bowles, 2006).

See Also

- ghettoes
- income mobility
- inequality between nations
- inequality (global)
- inequality (international evidence)
- inequality (measurement)
- intergenerational income mobility
- path dependence
- social networks in labour markets
- social norms

Bibliography

Arthur, W. 1989. Increasing returns, competing technologies and lock-in by historical small events: the dynamics of allocation under increasing returns to scale. *Economic Journal* 99, 116–31.

Banerjee, A. and Newman, A. 1991. Risk-bearing and the theory of income distribution. *Review of Economic Studies* 58, 211–35.

Banerjee, A. and Newman, A. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101, 274–98.

Becker, G. and Tomes, N. 1979. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87, 1153–89.

Blume, L. and Durlauf, S., eds. 2005. *The Economy as an Evolving Complex System III: Current Perspectives and Future Directions*. Oxford and New York: Oxford University Press.

Botticini, M. 2003. Commercial and trade diasporas. In *Oxford Encyclopedia of Economic History*. New York: Oxford University Press.

Botticini, M. and Eckstein, Z. 2005. Jewish occupational selection: education, restrictions, or minorities? *Journal of Economic History* 65, 922–48.

Botticini, M. and Eckstein, Z. 2006. From farmers to merchants, voluntary conversions and diaspora: a human capital interpretation of Jewish history. Discussion Paper No. 5571. London: CEPR.

Bowles, S. 2006. Institutional poverty traps. In *Poverty Traps*, ed. S. Bowles, S. Durlauf and K. Hoff. Princeton, NJ: Princeton University Press.

Bowles, S., Durlauf, S. and Hoff, K., eds. 2006. *Poverty Traps*. Princeton, NJ: Princeton University Press.

Bowles, S. and Gintis, H. 2002. The inheritance of inequality. *Journal of Economic Perspectives* 16(3), 3–30.

Carter, L. and Margo, R. 2007. Feminization of teaching in the United States. In *Gender and Education in the United States*, ed. B. Bank. Westport, CT: Greenwood Press.

Chiswick, R. 2005. The occupational attainment of American Jewry: 1990–2000. IZA Discussion Paper No. 1736. Bonn, Germany: Institute for the Study of Labour.

Collins, W. 1997. When the tide turned: immigration and the delay of the Great Black migration. *Journal of Economic History* 57, 607–32.

Collins, W. 2000. African-American economic mobility in the 1940s: a portrait from the Palmer Survey. *Journal of Economic History* 60, 756–81.

Collins, W. 2001. Race, Roosevelt, and wartime production: fair employment in World War II labor markets. *American Economic Review* 91, 272–86.

David, P. 1985. Clio and the economics of QWERTY. *American Economic Review Papers and Proceedings* 75(2), 332–7.

David, P. 1994. Why are institutions the ‘carriers of history’? Path dependence and the evolution of conventions, organizations, and institutions. *Structural Change and Economic Dynamics* 5, 205–20.

Donohue, J. III and Heckman, J. 1991. Continuous versus episodic change: the impact of the civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29, 1603–43.

Durlauf, S. 2006. Groups, social influences and inequality: a membership theory perspective on poverty traps. In *Poverty Traps*, ed. S. Bowles, S. Durlauf and K. Hoff. Princeton: Princeton University Press.

Eckstein, Z. and Zilcha, I. 1994. The effects of compulsory schooling on growth, income distribution and welfare. *Journal of Public Economics* 54, 339–59.

Fujita, M., Krugman, P. and Venables, A. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Galor, O. and Zeira, J. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60, 35–52.

Goldin, C. 1986. Monitoring costs and occupational segregation by sex: a historical analysis. *Journal of Labor Economics* 4, 1–27.

Goldin, C. 1990. *Understanding the Gender Wage Gap: An Economic History of American Women*. Chicago: University of Chicago Press.

Goldin, C. and Katz, L. 2003. The ‘virtues’ of the past: education in the first hundred years of the new republic. Working Paper No. 9958. Cambridge, MA: NBER.

Greif, A. 1989. Reputation and coalitions in medieval trade: evidence on the Maghribi traders. *Journal of Economic History* 49, 857–82.

Krugman, P. 1991a. History and industry location: the case of the manufacturing belt. *American Economic Review Papers and Proceedings* 81(2), 80–3.

Krugman, P. 1991b. Increasing returns and economic geography. *Journal of Political Economy* 99, 483–99.

Kuznets, S. 1960. Economic structure and life of the Jews. In *The Jews: Their History, Culture, and Religion*, ed. L. Finkelstein. Philadelphia: Jewish Publication Society of America.

Liebowitz, S. and Margolis, S. 1995. Path dependence, lock-in, and history. *Journal of Law, Economics, & Organization* 11, 205–226.

Long, J. and Ferrie, J. 2005. A tale of two labor markets: intergenerational occupational mobility in Britain and the United States since 1850. Working Paper No. 11253. Cambridge, MA: NBER.

Loury, G. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49, 843–67.

Loury, G. 2002. *The Anatomy of Racial Inequality*. Cambridge, MA: Harvard University Press.

Margo, R. 1990. *Race and Schooling in the South, 1880–1950: An Economic History*. Chicago: University of Chicago Press.

Mookherjee, D. and Ray, D. 2003. Persistent inequality. *Review of Economic Studies* 70, 369–93.

Mookherjee, D. and Ray, D. 2002. Contractual structure and wealth accumulation. *American Economic*

Review 92, 818–49.

Perlmann, J. and Margo, R. 2001. *Women's Work? American Schoolteachers, 1650–1920*. Chicago: University of Chicago Press.

Rotella, E. 1981. *From Home to Office: US Women at Work, 1870–1930*. Ann Arbor, MI: UMI Research Press.

Shelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1, 143–86.

How to cite this article

Botticini, Maristella and Zvi Eckstein. "path dependence and occupations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000318> doi:10.1057/9780230226203.1257

The New Palgrave Dictionary of Economics Online

path dependence in technical standards

Douglas Puffert

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The value of standardization leads to the persistence of established technical practices despite, in some cases, imperfect adaptation to current exogenous conditions. Economic explanation of these practices therefore requires reference to history. The conditions leading to path dependence in technical standards, as well as path independence, are examined with reference both to economic theory and to case studies of the QWERTY keyboard, videocassette recorder systems, railway track gauge and other railway standards. The controversy over path dependence is discussed.

Keywords

increasing returns to adoption; information technology; learning by doing; learning by using; learning effects; multiple equilibria; network effects; path dependence; path dependence in technical standards; QWERTY keyboard; railway track gauge; standards; technical interrelatedness; technical standards; transaction costs

Article

Path dependence is the dependence of outcomes on the course of previous outcomes, and thus on past conditions, rather than simply on current exogenous conditions. In a path-dependent process of economic change, choices motivated by transitory conditions can have results that persist long after those conditions change. The early conditions that have persisting effects could be systematic in nature, but the literature has focused more on the role of non-systematic ‘small’ events in selecting one potential path of later outcomes rather than another. A path-dependent process is non-ergodic, that is, its limiting distribution of possible outcomes changes as a function of its specific, evolving history. History matters for later outcomes, and economic explanation is incomplete without accounting for that history.

Early choices can have particularly strong effects in the context of technologies that exhibit ‘increasing returns to adoption’ (Arthur, 1989), in that specific practices become more valuable to each user as the

total number of users rises. These increasing returns often give rise to standards – rules or practices that enable adopters to pursue some sort of value-producing interaction with other adopters. For example, railway companies that adopt a common track gauge (width between the rails) can exchange cars without reloading, while typists who learn a standard keyboard layout can apply their skills in any office that uses standard machines. Increasing returns to adoption can arise either on the demand side or the supply side of a market. On the demand side, as in the cases of railway gauges and typewriter keyboards, adopters gain from participation in physical or virtual networks of adopters. On the supply side, learning effects – learning by doing or learning by using – reduce the cost or improve the characteristics of a product as cumulative adoption increases.

Quite often, a technology embodying increasing returns to adoption offers a range of specific practices that could form the basis for a standard. Railways have used track gauges ranging from about 600•mm (two feet) to 2,140•mm (seven feet), and typewriter or computer keyboards could use 26-factorial (about 10^{26}) different orderings of the letters. These practices may represent different, diverging, potential paths of outcomes. Once early adopters choose a particular practice, later adopters have an incentive to match those choices in order to gain the benefits of compatibility. Thus, increasing returns can give rise to positive feedbacks among agents' choices. The selection process (or allocation process) that results does not converge to a unique equilibrium outcome. Rather, such a process has multiple potential equilibria or, rather, equilibrium paths. These equilibria could vary substantially both in their general efficiency (total payoffs) and in their distribution of payoffs among different agents. Which equilibrium is selected depends in large part on early choices.

Requirements for path dependence

Two things are necessary for path dependence to make a difference for outcomes (David, 1999; 2001). First, the conditions or criteria that determine early choices – and thus one branching path rather than another – must not be closely correlated with the conditions or interests that matter later. Second, the selected path of outcomes must constitute a locally stable equilibrium, so that the selection process does not simply revert to an outcome that is determined by later conditions or interests.

Empirically, the most important reasons that early choices might not reflect later conditions are limited information and limited technical capability. These are common occurrences in the early stages of a new technology. Innovators must often engage in exploratory behaviour to learn the possibilities for both technological development and market application (Nelson and Winter, 1977), and the effective choice of a standard practice may precede much of this learning. For example, the standard track gauge used in most of the world today was chosen during the 1820s, when railway cars were little more than road wagons and when locomotives were little more than small steam engines set on wagons and linked by crank to a wheel. The gauge was not optimized for what railways were soon to become, let alone for what they are 180 years later. Many engineers since the 1830s have believed that a broader gauge would be more efficient for most purposes (Puffert, 2002; 2008). As another example, the 'QWERTY' standard typewriter keyboard was developed by rearranging letter sequences so as to minimize the jamming propensities of one short-lived typewriter design in 1873. Modern eight-finger typing methods emerged a decade later (David, 1986), and keyboards designed for such methods offer about a ten per cent improvement over QWERTY in typing efficiency (Norman, 1990).

A further reason that early choices might not reflect later conditions and interests is that later adopters may have different interests with regard to the content of a standard than early adopters, but high transaction costs (or simple lack of foresight) may prevent the later adopters from influencing the early choices that determine their later options. In addition, the discounting of future payoffs would lead even perfectly foresighted agents to place little value on distant outcomes.

The second requirement for path dependence to have consequential effects, again, is that a selected path of outcomes be locally stable. Part of the cause of such stability is increasing returns to adoption. Increasing returns lead new adopters to adopt a practice simply because it is the established standard with a large installed base, even if they would prefer some alternative practice if that were to have a comparable installed base. In such an instance the established standard is said to be ‘locked in’, both in the economics literature (Arthur, 1989) and in the business world.

Another part of the reason for the stability of a path of outcomes is switching costs – the hardware conversion costs, retraining costs, and transaction costs (that is, information and coordination costs) entailed in converting from an established standard practice to a superior alternative. Users are often restrained from converting not only by irreversible investments in the established practice but also by the technical interrelatedness of system components (David, 1986), which makes piecemeal conversion impractical. In a railway, for example, individual equipment and fixtures of one gauge cannot simply be replaced with items of another gauge when they wear out. Rather, new equipment must continue to match the installed base of old equipment, and a conversion requires that all equipment be converted together. Furthermore, the value of compatibility may mean that any conversion must be coordinated among many agents.

These two requirements are by no means always present as technical standards are formed. Foresight into the technological and market opportunities of a new technology is often sufficient to enable early adopters or product sponsors to choose, in effect, a superior path of later outcomes. As an example, Sony and Philips introduced the standard compact disc (CD) format in the early 1980s after digital audio sampling theory, other relevant technologies, and market requirements were already well known. The standard has served quite well. In such instances path dependence, as such, plays no role in the selection process.

Furthermore, if switching costs are sufficiently low, then a less preferred path does not constitute a stable equilibrium. Thus, the potential inefficiency of a path-dependent process is generally limited to the cost (including transaction costs) of carrying out a remedy for this inefficiency. A less preferred path of outcomes may become unstable through innovations or market developments that either reduce the costs or increase the benefits of transition to a preferred practice (Puffert, 2004). For example, invention of the low-cost rotary electrical converter in the 1880s helped bring an end to regional lock-in to DC electrical power by facilitating the coupling of AC transmission networks with applications that required DC (David, 1991). Similarly, in contemporary information technology, adapters or ‘gateways’ arise frequently to link otherwise incompatible networks (David, 1987). Sometimes these techniques offer a migration path from an inferior or obsolete practice to a superior one, making the selection process ‘path independent’.

The controversy over path dependence

The concept of path dependence first gained widespread attention in economics through Paul David's (1985; 1986) interpretation of the case of QWERTY and through a series of theoretical discussions by W. Brian Arthur (1989; 1994). David's thinking grew out of an earlier literature on how technical interrelatedness can inhibit adaptation to changing conditions (Veblen, 1915; Frankel, 1955; Kindleberger, 1964; David, 1975). Arthur combined mathematical models of non-ergodic processes in the natural sciences with economic theory concerning how increasing returns can give rise to multiple equilibria.

Arthur developed models in which stochastic fluctuations in the market shares of alternative products or practices are magnified by positive feedbacks until one practice gains the whole market, becoming locked in as a de facto standard. In view of the subsequent controversy over path dependence, it is worth noting that Arthur's (1989) primary model used two key assumptions that obviated the need to consider expectations and forward-looking behaviour. First, he assumed that alternative competing practices are unsponsored, rather than promoted by suppliers. Second, he assumed that increasing returns to adoption are based simply on learning effects embodied in a practice at the time of adoption, so that each adopter's payoffs depend only on the number of previous adoptions, not the number of future adoptions. Arthur discussed only briefly how outcomes of his model would differ under alternative assumptions. He acknowledged that, if increasing returns were based on network effects rather than learning effects, then each adopter's payoffs would continue to rise after adoption as the number of adopters increased. He reasoned that expectations would then lead to earlier lock-in, but he did not carry his analysis further.

Stan Liebowitz and Stephen Margolis (1990; 1995) raised a substantive critique of David's and Arthur's writings, based partly on exploring the implications of assumptions other than those of Arthur's model. The central thrust of their argument was that purposeful, profit-seeking, forward-looking behaviour can override the mechanisms of path dependence whenever, in their view, outcomes truly matter. According to Liebowitz and Margolis (1995), if agents can foresee that some potential future outcomes offer higher payoffs than others, then they have a variety of means to steer the selection process toward the preferred outcomes. Suppliers of products that embody superior practices can profit by promoting those practices to become standards. Adopters can also conduct transactions among themselves, by direct communication or market mechanisms, to assure that they realize the highest available payoffs. According to Liebowitz and Margolis, if means such as these are unable to realize a putatively superior outcome, then that is only because the costs (including transaction costs) of pursuing that outcome are greater than the benefits. In other words, they argued, the putatively superior outcome is not really superior. Agents may come to regret that earlier choices, made in the absence of good foresight, had made some conceivable outcome unattainable. However, Liebowitz and Margolis argued, such regret is naive, a crying over spilt milk.

Liebowitz and Margolis concluded that path dependence is likely to affect only features of the economy that no economic agent has a real reason to care about – and that are not worth much attention from economists or economic historians. They set forth a taxonomy of ‘degrees’ of path dependence: first degree, in which alternative outcomes have no consequences for efficiency; second degree, in which different outcomes offer differing payoffs but imperfect foresight and transaction costs prevent purposeful behaviour to attain the highest payoffs; and third degree, in which there is sufficient foresight and scope for forward-looking behaviour to attain the superior outcome, but this outcome is somehow still not attained. Liebowitz and Margolis argued that only the third type of path dependence would offer a real challenge to what they called ‘the neoclassical model of relentlessly rational behavior leading to

efficient, and therefore predictable, outcomes’. They claimed, however, that this type is unlikely to arise empirically.

David (1997; 1999; 2001) responded that Liebowitz and Margolis had mischaracterized several of the issues at stake. Puffert (2000; 2002; 2004; 2008) responded to the critics by incorporating the issues of foresight and forward-looking behaviour explicitly into models and case studies, and he argued that such behaviour is fully compatible with path dependence. He maintained that Liebowitz's and Margolis's taxonomy of ‘degrees’ is incomplete, leaving out a great range of cases where agents are neither fully passive nor fully able to control outcomes. Such cases demonstrate a rich, complex interplay between forward-looking behaviour and the legacy of past events.

Although David and Arthur had not sufficiently examined the issues of foresight and forward-looking behaviour, they also had not fully neglected these matters. David (1986), indeed, attributed path dependence in typewriter keyboards to the lack of perfect futures markets. Kenneth Arrow stated in his foreword to Arthur's collected articles that much of Arthur's analysis applies specifically where ‘expectations are myopic, based on limited information’ (Arthur, 1994). This is not how Arthur himself explicitly interpreted his models, but it is how he applied them. For example, Arthur (1989), as well as David (1987), argued that path dependence may be particularly relevant for policy when early information is imperfect. Government, they argued, might improve information and later outcomes by exploring the potential payoffs of alternative practices before one practice became locked in. Such a policy later proved its value when the US government sponsored a competition among alternative high-definition television systems, resulting in the accelerated development of a superior digital technology while preventing lock-in to a soon-to-be outmoded analog system.

The relevance of all these considerations must be judged empirically. We begin with the disputed case of QWERTY.

The QWERTY keyboard

David (1985; 1986) argued that QWERTY gained a lead over rival keyboard systems due to the happenstance that instruction in eight-finger ‘touch’ typing was developed first for QWERTY during the mid-1880s. The best-trained typists used QWERTY, so office managers hired them and bought QWERTY machines to match. This, in turn, gave budding typists, typing schools, the writers of typing manuals, and typewriter manufacturers a further incentive to focus on QWERTY, to the exclusion of alternative systems. Positive feedbacks reinforced QWERTY's early lead until it gained virtually the whole market. The superior ‘Ideal’ keyboard layout, introduced in 1893, appeared too late to disrupt a lock-in to QWERTY.

David (1986) concluded, ‘competition in the absence of perfect futures markets drove the industry prematurely into standardization on the wrong system’. Critical to both the emergence and persistence of QWERTY was that the ‘larger system of production’, comprising typists, employers, manufacturers, and typing instructors, ‘was nobody's design’; it was characterized by decentralized decision making.

Liebowitz and Margolis (1990) responded, in effect, that ‘design’ rather than positive feedbacks controlled the process that produced the QWERTY standard. Early typewriter manufacturers, they noted, competed vigorously on features of their machines, and they inferred from this that QWERTY succeeded due to a market test of its relative fitness. Positive feedbacks played no role, they argued,

because typewriter suppliers had an opportunity to provide training to offices where they sold their machines. Suppliers could thus internalize, and profit from, the advantages of a superior keyboard. David (1999; 2001) responded in turn that keyboards were never tested by the market in isolation from numerous other features of machines that varied among manufacturers. Furthermore, Liebowitz and Margolis offered no evidence that typewriter manufacturers found it practical to offer training in touch typing before the 1920s, long after QWERTY had become the established standard. Thus their argument has no empirical basis. Still, David's empirical evidence appears less than conclusive in light of the points raised by his critics.

Liebowitz and Margolis devoted most of their article to matters less relevant to David's argument. They refuted a popular account that QWERTY won ‘once and for all’ due to the publicity it received when a touch-typing QWERTY typist won a single typing contest in 1888. As they showed, non-QWERTY typists soon won other typing contests, so a single contest could not have been decisive. Their refutation did not, however, address David's argument that the contest in question had publicized the value of touch typing, which was being taught at the time only for QWERTY. Liebowitz and Margolis also refuted the mistaken story that QWERTY had been designed to slow typists down, but, again, this story was never part of the claims made by theorists of path dependence.

Liebowitz and Margolis did convincingly refute one claim about QWERTY that David had tacitly accepted – that the Dvorak Simplified Keyboard, invented in 1932, was so superior to QWERTY that the cost of retraining could be recovered in a period of weeks. As Liebowitz and Margolis showed, this claim was based on dubious experiments, and it does not stand the test of reasoned inference from users’ behaviour. But David had mentioned the claim, in a single sentence, only to establish the extent to which the legacy of early events had mattered. His argument is little affected if the relative inefficiency of QWERTY is only on the order of ten per cent, as estimated by a leading researcher in industrial design and ergonomics (Norman, 1990). David's claim about history mattering would, however, be affected if QWERTY's relative inefficiency is next to nothing, as Liebowitz and Margolis suggest.

Videocassette recorders and similar cases

Another influential case study in path dependence was the competition between alternative videocassette recording systems from the mid-1970s to the mid-1980s. The VHS system of JVC (Japan Victor Corporation) became the standard, beating out Sony's Betamax. Arthur (1990) explained this as the result of positive feedbacks in the video rental market, as video stores stocked more film titles for the system that accidentally gained a larger user base, while consumers bought the system for which they could rent more videos. Liebowitz and Margolis (1995) pointed out, however, that Sony had actually been first to market. If positive feedbacks had mattered, they argued, then Sony should have won. They attributed the VHS victory to active product promotion and to the advantage of VHS in offering a longer playing time. In their view, purposeful, forward-looking behaviour had overridden positive feedbacks, ensuring the superior outcome. They offered substantial evidence against Arthur's suggestion that the winning system may have been technically inferior.

The extensively documented account of Cusumano, Mylonadis and Rosenbloom (1992) showed, however, that purposeful behaviour did not trump path dependence. There was indeed a positive-feedback dynamic in the video rental market, but this market emerged late, after VHS had already

gained a strong lead. The onset of positive feedbacks turned Betamax's small but stable market share into a fast-declining one, forcing it to exit the market.

More intriguingly, Cusumano, Mylonadis and Rosenbloom attributed the earlier lead of VHS to path dependence in supplier choices. Manufacturers and distributors increasingly supported VHS over Betamax as they saw others doing so, increasing their expectations that VHS, not Betamax, would later become the standard. Ultimately, the authors argued, VHS won as the result of non-systematic differences in the promoters' early strategy choices. First, Sony initially pursued a go-it-alone strategy, while JVC built a coalition of suppliers in order to benefit from positive feedbacks. Second, JVC's partner Matsushita installed a large manufacturing capacity to solidify expectations among other suppliers. Third, Sony opted for a smaller cassette size, while JVC chose a larger cassette with longer playing time. In the event, a longer playing time proved more important to consumers in the early years, when only a VHS tape could record an entire American football game or a long movie. Distributors responded to this temporary advantage by joining the VHS coalition permanently.

This account shows that path dependence is fully compatible with forward-looking behaviour, provided that foresight is imperfect when early choices are made about strategy and product characteristics. Indeed, market participants recognized positive feedbacks, and they sought to influence the early events that would have a disproportionate effect on later outcomes.

Such behaviour is common in advanced-technology industries, and innovators whose forward-looking behaviour takes positive feedbacks into account are more likely to win their markets (Morris and Ferguson, 1993; Shapiro and Varian, 1998). Indeed, according to many observers, either IBM or Apple Computer rather than Microsoft could have become the dominant firm in microcomputers, controlling the key system standard (Rohlfs, 2001; Carlton, 1997). However, only Bill Gates of Microsoft had, and acted on, the foresight that control of a standard would matter. He became the world's richest individual as a result.

Such processes are path dependent when outcomes depend, in part, on non-systematic choices and events. If general foresight is good, however, then systematic considerations may dominate. Market participants may agree on a superior outcome from the start, in the manner of a fulfilled-expectations process (Katz and Shapiro, 1985). An example, again, is the CD standard.

What is at stake in a path-dependent process is not necessarily general efficiency or total payoffs. It may, rather, be the distribution of payoffs to different innovators and suppliers. Furthermore, in a path-dependent process, particular individuals can have lasting effects on later outcomes, for better and for worse.

Railway track gauge

One individual who made a lasting difference was railway pioneer George Stephenson. Stephenson transferred the gauge of four feet eight and a half inches (1,435•mm) from the primitive mining tramways where he gained his early experience to the Liverpool and Manchester Railway. That line became the model of best practice for the earliest railways of Britain, North America, and Continental Europe (Puffert 2000; 2002; 2008). The Stephenson gauge became the standard over wider areas as new railways, interested in compatibility, adopted the gauge of prior neighbouring lines.

Engineers soon came to prefer broader gauges, and they introduced such gauges to new regions. A lack

of foresight into the later importance of large-scale network integration led to the emergence of two regional standard gauges in Britain, six in North America, six in Continental Europe, and multiple gauges in Australia, India, and other intercommunicating regions. The cost of coping with or resolving this diversity was the main path-dependent inefficiency in track gauge, outweighing the minor inefficiency of the prevalent Stephenson gauge. Still, diversity was resolved most easily where it proved most costly, and the mechanism for resolving diversity was frequently the sort of coordinating behaviour discussed by Liebowitz and Margolis (1995). Much of Britain's and North America's diversity, for example, was resolved by emerging interregional rail systems that internalized the benefits of standardization.

Even so, these improvements in outcomes were a matter of ‘path-constrained amelioration’ (David, 2001) rather than a complete break from the historical legacy. Britain made the Stephenson gauge its general standard at a time when the consensus of engineers favoured a gauge of five feet to five feet six inches (1,524•mm to 1,676•mm), and North America did so when the consensus favoured five feet. Japan has long regretted its choice of a narrow standard gauge, three feet six inches (1,067•mm). Australia and India have only recently resolved much of their diversity, while the variant gauges of the Iberian peninsula and the former Russian and Soviet empires are becoming more costly as those regions are integrated economically into the core of Europe. However, the cost of this diversity is being reduced by innovative mechanisms that enable trains to change their gauge en route.

The potential role of government mandates in improving on path-dependent outcomes was proved in Britain, where the 1846 Gauge Act led to some rationalization of gauges.

Other railway standards

Path-dependent diversity in regional standards has also proven costly in such matters as railway electrification systems, clearance dimensions (‘loading gauges’), and train control and signalling systems. This diversity has hindered the formation of international high-speed train links in Europe (Puffert, 1993).

Several railway standards that were well adapted to early conditions proved poorly adapted to later ones, but they continued in use due to the cost of converting the installed base. Examples reportedly include mechanical couplings and air brakes, as electrical systems would now be safer and less labour-intensive (Hilton, 1990, p. 294).

A more famous example is what Veblen (1915) called the ‘silly little bobtailed carriages’ used in British goods traffic. A long literature has addressed how the historical legacy of interrelated freight handling facilities prevented the modernization of coal cars in particular (Kindleberger, 1964, pp. 141–4). Recently, Van Vleck (1997) argued that small coal cars were well adapted to the larger system of distribution, chiefly by reducing the costs of small deliveries. Scott (2001) showed, however, that few coal users benefited from small car-size deliveries. Rather, the cars’ small size, widely dispersed ownership (by collieries), antiquated braking and lubrication systems, and generally poor physical condition made them quite inefficient indeed. Replacing these cars and associated infrastructure with modern, larger wagons owned and controlled by the railways would have offered savings in railway operating costs of about 56 per cent, yielding a social rate of return of 24 per cent on the physical costs of conversion. Nevertheless they were not replaced until both the railways and the collieries were

nationalized after 1945. Until then, regulations forced the railways to accept colliery cars at set rates or else offer high levels of compensation. Due to technical interrelatedness, the railways could not have saved much in operating costs until virtually all the antiquated cars were replaced, so high transaction costs prevented transition to a more efficient practice.

Further cases

Cowan (1990) argued that transitory circumstances led to the establishment of the prevalent ‘light-water’ design for civilian nuclear power reactors. This design, adapted from nuclear submarines, was rushed into use due to the political value of demonstrating peaceful uses for nuclear technology. Thereafter, learning effects arising from engineering experience continued to make the light-water design the rational choice for new reactors. Cowan argued, however, that an equivalent degree of development would likely have made an alternative design superior.

Cowan and Gunby (1996) addressed farmers’ choices between systems of chemical pest control and integrated pest management (IPM), which uses predatory insects to devour harmful ones. As the drift of chemical pesticides from neighbouring fields often makes the use of IPM impossible, IPM must be used on the whole set of farms that are in proximity to one another. Where this set is large, the transaction costs of persuading all farmers to forgo chemical methods often prevent adoption. In addition to these localized positive feedbacks, local learning effects also make the choice between systems path dependent. Local lock-in to each practice is sometimes upset by such developments as invasions by new pests and the emergence of resistance to pesticides.

See Also

- irreversible investment
- learning-by-doing
- network goods (empirical studies)
- network goods (theory)
- path dependence
- technical change
- Veblen, Thorstein Bunde

Bibliography

Arthur, W.B. 1989. Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal* 99, 116–31.

Arthur, W.B. 1990. Positive feedbacks in the economy. *Scientific American* 262 (February), 92–9.

Arthur, W.B. 1994. *Increasing Returns and Path Dependence in the Economy*. Ann Arbor: University of Michigan Press.

Carlton, J. 1997. *Apple: The Inside Story of Intrigue, Egomania, and Business Blunders*. New York: Times Business.

Cowan, R. 1990. Nuclear power reactors: a study in technological lock-in. *Journal of Economic History* 50, 541–67.

Cowan, R. and Gunby, P. 1996. Sprayed to death, path dependence, lock-in and pest control strategies. *Economic Journal* 106, 521–42.

Cusumano, M.A., Mylonadis, Y. and Rosenbloom, R.S. 1992. Strategic maneuvering and mass-market dynamics: the triumph of VHS over Beta. *Business History Review* 66, 51–94.

David, P.A. 1975. *Technical Choice, Innovation and Economic Growth: Essays on American and British Experience in the Nineteenth Century*. Cambridge: Cambridge University Press.

David, P.A. 1985. Clio and the economics of QWERTY. *American Economic Review, Papers and Proceedings* 75, 332–7.

David, P.A. 1986. Understanding the economics of QWERTY: the necessity of history. In *Economic History and the Modern Economist*, ed. W.N. Parker. Oxford: Oxford University Press.

David, P.A. 1987. Some new standards for the economics of standardization in the information age. In *Economic Policy and Technological Performance*, ed. P. Dasgupta and P. Stoneman. Cambridge: Cambridge University Press.

David, P.A. 1991. The hero and the herd: reflections on Thomas Edison and the ‘Battle of the Systems’. In *Favorites of Fortune, Technology, Growth, and Economic Development Since the Industrial Revolution*, ed. P. Higonnet, D.S. Landes and H. Rosovsky. Cambridge, MA: Harvard University Press.

David, P.A. 1997. Path dependence and the quest for historical economics: one more chorus of the ballad of QWERTY. Discussion Papers in Economic and Social History No. 20, University of Oxford.

David, P.A. 1999. At last, a remedy for chronic QWERTY-skepticism! Working paper, All Souls College, Oxford University.

David, P.A. 2001. Path dependence, its critics and the quest for ‘historical economics’. In *Evolution and Path Dependence in Economic Ideas, Past and Present*, ed. P. Garrouste and S. Ioannides. Cheltenham, UK and Northampton, MA: Edward Elgar.

Frankel, M. 1955. Obsolescence and technological change in a maturing economy. *American Economic*

Review 45, 296–319.

Hilton, G.W. 1990. *American Narrow-Gauge Railroads*. Stanford: Stanford University Press.

Katz, M.L. and Shapiro, C. 1985. Network externalities, competition, and compatibility. *American Economic Review* 75, 424–40.

Kindleberger, C.P. 1964. *Economic Growth in France and Britain, 1851–1950*. Cambridge, MA: Harvard University Press.

Liebowitz, S.J. and Margolis, S.E. 1990. The fable of the keys. *Journal of Law and Economics* 33, 1–25.

Liebowitz, S.J. and Margolis, S.E. 1995. Path dependence, lock-in, and history. *Journal of Law, Economics, and Organization* 11, 204–26.

Morris, C.R. and Ferguson, C.H. 1993. How architecture wins technology wars. *Harvard Business Review* 71 (March–April), 86–96.

Nelson, R.R. and Winter, S.G. 1977. In search of a useful theory of innovation. *Research Policy* 6, 36–76.

Norman, D.A. 1990. *The Design of Everyday Things*. New York: Doubleday.

Puffert, D.J. 1993. Technical diversity and the integration of the European high-speed train network. In *High-Speed Trains, Fast Tracks to the Future*, ed. J. Whitelegg, S. Hultén and T. Flink. Hawes, North Yorkshire: Leading Edge.

Puffert, D.J. 2000. The standardization of track gauge on North American railways, 1830–1890. *Journal of Economic History* 60, 933–60.

Puffert, D.J. 2002. Path dependence in spatial networks: the standardization of railway track gauge. *Explorations in Economic History* 39, 282–314.

Puffert, D.J. 2004. Path dependence, network form, and technological change. In *History Matters: Essays on Economic Growth, Technology, and Demographic Change*, ed. T. Guinnane, W. Sundstrom and W. Whatley. Stanford: Stanford University Press.

Puffert, D.J. 2008. *Tracks Across Continents, Paths Through History: The Economic Dynamics of Standardization in Railway Gauge*. Chicago: University of Chicago Press.

Rohlfs, J.H. 2001. *Bandwagon Effects in High-Technology Industries*. Cambridge, MA and London:

MIT Press.

Scott, P. 2001. Path dependence and Britain's ‘coal wagon problem’. *Explorations in Economic History* 38, 366–85.

Shapiro, C. and Varian, H.R. 1998. *Information Rules*. Cambridge, MA: Harvard Business School Press.

Van Vleck, V.N.L. 1997. Delivering coal by road and rail in Britain: the efficiency of the ‘silly little bobtailed’ coal wagons. *Journal of Economic History* 57, 139–60.

Veblen, T. 1915. *Imperial Germany and the Industrial Revolution*. London: Macmillan.

How to cite this article

Puffert, Douglas. "path dependence in technical standards." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000323> doi:10.1057/9780230226203.1258

The New Palgrave Dictionary of Economics Online

path dependence

Steven N. Durlauf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Path dependence refers to the idea that ‘history matters’, that is, that various types of contingent events may have long-term consequences. This article provides some formalization of the concept and assesses its usefulness in elucidating economic phenomena.

Keywords

ergodicity and non-ergodicity in economics; multiple equilibria; multiple steady states; network externalities; path dependence; QWERTY keyboard configuration; reflection problem

Article

Originating with work of Paul David (1985; 1986) and W. Brian Arthur (1989), there have been a number of efforts by economists to argue that path dependence exists in various socio-economic outcomes. Heuristically, path dependence is understood to mean that ‘history matters’ in the sense that certain long-term economic outcomes are contingent on particular events that themselves need not have occurred.

The canonical example of path dependence is that of the QWERTY keyboard configuration for typewriters, studied by David (1985; 1986). David argues that the emergence of the QWERTY configuration as the standard for typewriters was an historical accident in the sense it was the consequence of a set of decentralized, uncoordinated choices by different economic actors whose decisions were driven by network externalities. As such, the standard became locked in even though it was not socially optimal, that is, there existed an alternative configuration, the Dvorak keyboard, which was preferable in terms of typing efficiency. From this perspective, the QWERTY keyboard was simply one of several potential long-run standards that could have emerged, its actual emergence being a function of a particular set of contingent events, that is, shocks. Other cases where path dependence has been argued to occur include nuclear reactor technology (Cowan, 1990) and railway gauge length

(Puffert, 2002). Arthur (1989) provides a formal model of path dependence which captures the logic of the QWERTY example; see also Farrell and Saloner (1985) for an early and important example in the industrial organization literature on network externalities which exhibits path dependence-type phenomena.

From the perspective of current economics, most of the path dependence literature is somewhat anomalous. While path dependence is often informally invoked to describe one phenomenon or another, there has been little systematic research on path dependence outside of economic history; in particular, there is a general dearth of formal theoretical and econometric analyses. As a result, discussions of path dependence are often very imprecise. Such imprecision may first be seen in definitions of path dependence. The term means different things in different writings, so that disagreements on its presence to some extent are simply disagreements about its definition.

Most discussions seem to equate path dependence with non-ergodicity. Consider a set of independent shocks ε_t and an outcome of interest x_t; let μ denote a probability measure. The process x_t is non-ergodic if for a fixed k, $\lim_{j \Rightarrow \infty} \mu(x_{t+j} | \varepsilon_0, \ldots, \varepsilon_k)$ depends on the realization of $\varepsilon_0, \ldots, \varepsilon_k$. Such a definition captures some of the main intuition underlying qualitative discussions of path dependence in that for such processes history matters, that is, particular sets of shocks have long-run consequences. Theoretical models of path dependence such as Arthur (1989) have this property. It is worth noting that, from the perspective of those few formal theories that claim to model path dependence, the phenomenon is typically a form of multiple equilibria or multiple steady states, both of which occur in many contexts. See Blume and Durlauf (2001) for a conceptual discussion on how various deviations from the Arrow–Debreu baseline, when combined with complementarities (those of network externalities are simply one example), can lead to multiple equilibia or multiple metastable states, that is, states from which a system will emerge only after long epochs. What distinguishes theoretical models of path dependence appears to be the explicit attention to the consequences of individuals making decisions sequentially, so that dynamic forms of coordination failure can occur.

However, it is far from clear that such a notion of path dependence as non-ergodicity is sensible for the examples for which path dependence has been claimed to occur. While it is incontrovertible that technological standards are subject to strong network externalities, it is equally true that technological standards evolve over time. One early example of path dependence was the success of the VHS tape standard over Betamax. In light of the rise of the DVD, it is unclear in what sense the success of the VHS tape over a particular time horizon is evidence of anything deeper than network externalities per se. It is possible that a better definition of path dependence relates to whether shocks to a system are self-reversing. Suppose we consider a system where $\varepsilon_l = 0, \ l > k$. If $\lim_{j \Rightarrow \infty} \mu(x_{t+j} | \varepsilon_0, \ldots, \varepsilon_k, \varepsilon_l = 0, \ l > k)$ depends on the realization of $\varepsilon_0, \ldots, \varepsilon_k$, then one has a system in which shocks to a system can persist unless overcome by future shocks. This notion of path dependence may be more sensible for contexts such as technological standards as it respects the role of new technologies in undoing current configurations; in fact, it seems the more appropriate definition from the perspective of various examples in economic history. This definition of path dependence also has the advantage that it is meaningfully different from other mathematical concepts, that is, non-ergodicity, that have separately appeared in the economics literature. This suggests that theories of path dependence should focus on how systems can exhibit long passage times out of local basins of attraction rather than multiple equilibria or multiple steady states per se. This in turn would suggest that analyses

of path dependence should focus on understanding aggregate nonlinearities rather than the persistence of shocks as has occurred historically. The reason for this is that the second definition does not imply that actual shocks have persistent effects, only that they could have them.

The definitional ambiguities associated with path dependence are mirrored in the substantive discussions that have been developed concerning the economic environments in which it is supposed to occur. Liebowitz and Margolis (1990; 1995) have challenged David's claims about path dependence and the QWERTY keyboard and indeed have questioned the general empirical relevance of the concept. The Liebowitz and Margolis arguments, in the context of QWERTY, largely amount to claiming that there is no good evidence that the Dvorak keyboard is superior and that, further, the historical record indicates that the emergence of the QWERTY standard was driven by competitive forces to a much greater extent than acknowledged by David (rebuttals to their claims include David, 2001). This debate has not been resolved and has generally been unproductive. As discussed in Durlauf (2005), the main problem is the lack of careful attention to microeconomic behaviours when analysing the historical evidence. For example, to the extent that evidence of path dependence is equated with the possible stability of a technologically inferior standard, then what matters in evaluating the claim is the level of information that was available to the individual economic actors when they made their standard choices, not what was *ex post* true. This requires much more explicit attention to the decision problems of the individual actors whose choices are collectively said to produce path dependence as well as the way in which an equilibrium configuration of choices occurs at each point in time.

Put differently, the path dependence literature has generally reasoned from aggregate observations towards microeconomic conclusions, whereas a rigorous formulation and empirical evaluation of path dependence as the property of an economic system requires that one start with individual decisions and reason towards aggregate implications. What this means is that resolution of whether network externalities, for example, have produced multiple steady states in a particular market should be understood as claims about the nature of individual decisions and how aggregate equilibria emerge from them. It is well known, as a theoretical matter, that broad claims such as the assertion that markets select for rationality, efficiency and the like (cf. Blume and Easley, 1992) depend on details of the economic environment under study. By implication, formal microeconometric analysis will be necessary for empirical adjudication of claims concerning path dependence.

The existing microeconometric literature makes it very clear that there exist deep difficulties in determining whether path dependence is present in a given environment. For example, without strong assumptions on the decision rules of individual agents, one cannot identify whether the equilibrium of a given model is or is not unique. Indeed, even with individual level data, so that the decision rules of the agents can in principle be estimated, identification of whether an environment can produce multiple equilibria is difficult (cf. Brock and Durlauf, 2008; Tamer, 2003). From the econometric perspective, one basic problem is that any argument that an equilibrium has emerged that is not unique implicitly requires identification of the strength of the interdependences in individual choices, for example network effects. These interdependences cannot be identified unless one is willing to make assumptions about the correlated (across individuals) unobserved components to the costs and payoffs of the choices that are made; some possibilities on how to do this appear in Brock and Durlauf (2008). Further, even when there are no such correlated components, there are cases where the degree of interdependence cannot be identified when there are correlated observables components; this was established in Manski (1993) who

calls this the reflection problem. At the current writing, *none* of these issues has been explicitly considered in the study of path dependence.

Thus, the current path dependence literature has had mixed success. From the perspective of the identification of interesting facts and the description of candidate environments for multiple steady states, the path dependence literature has been quite stimulating. From the perspective of developing a new theoretical view of economic outcomes, something that the more grandiose writings on path dependence sometimes allege they do, the literature is still highly imprecise and speculative. Thus the contributions of path dependence research really amount to the delineation of interesting historical episodes, episodes whose interpretation has yet to be resolved.

See Also

- economy as a complex system
- path dependence and occupations
- path dependence in technical standards
- social interactions (empirics)
- social interactions (theory)

Bibliography

Arthur, W. 1989. Increasing returns, competing technologies and lock-in by historical small events: the dynamics of allocation under increasing returns to scale. *Economic Journal* 99, 116–31.

Blume, L. and Durlauf, S. 2001. The interactions-based approach to socioeconomic behavior. In *Social Dynamics*, ed. S. Durlauf and H. Young. Cambridge, MA: MIT Press.

Blume, L. and Easley, D. 1992. Evolution and market behavior. *Journal of Economic Theory* 58, 9–40.

Brock, W. and Durlauf, S. 2008. Identification of binary choice models with social interactions. *Journal of Econometrics*.

Cowan, R. 1990. Nuclear power reactors: a study in technological lock-in. *Journal of Economic History* 50, 541–67.

David, P. 1985. Clio and the economics of QWERTY. *American Economic Review* 75, 332–7.

David, P. 1986. Understanding the economics of QWERTY: the necessity of history. In *Economic History and the Modern Economist*, ed. W. Parker. Oxford: Blackwell.

David, P. 2001. Path dependence, its critics and the quest for ‘historical economics’. In *Evolution and Path Dependence in Economic Ideas: Past and Present*, ed. P. Garrouste and S. Ioannides. Cheltenham,

UK and Northampton, MA: Edward Elgar.

Durlauf, S. 2005. Complexity and empirical economics. *Economic Journal* 112, 225–43.

Farrell, J. and Saloner, G. 1985. Standardization, compatibility, and innovation. *RAND Journal of Economics* 16, 70–83.

Liebowitz, S. and Margolis, S. 1990. The fable of the keys. *Journal of Law and Economics* 22, 1–26.

Liebowitz, S. and Margolis, S. 1995. Path dependence, lock-in, and history. *Journal of Law, Economics and Organization* 11, 205–26.

Manski, C. 1993. Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60, 531–42.

Puffert, D. 2002. Path dependence in spatial networks: the standardization of the railway track gauge. *Journal of Economic History* 39, 282–314.

Tamer, E. 2003. Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies* 70, 147–65.

How to cite this article

Durlauf, Steven N. "path dependence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000330> doi:10.1057/9780230226203.1256

The New Palgrave Dictionary of Economics Online

Patinkin, Don (1922 – 1955)

Nissan Liviatan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Don Patinkin's main contributions are in monetary theory, including the topics of involuntary unemployment and the interpretation of the writings of J.M. Keynes. He criticized the classical and neoclassical monetary model for its ‘invalid dichotomy’ between the real and the monetary sectors. His main underlying concern was whether capitalism possessed an automatic mechanism for attaining full employment. He claimed that the real interest rate might be insufficiently flexible and the real balance effect insufficiently powerful to allow any rapid convergence to equilibrium, rendering it politically unrealistic to rely on automatic forces to establish full-employment equilibrium in reasonable time.

Keywords

capitalism; Chicago School; Clower constraint; Cowles Commission; Friedman, F.; involuntary unemployment; IS–LM model; Keynes, J.M.; Kuznets, S.; monetary economics, history of; neoclassical monetary theory; Patinkin, D.; quantity theory of money; real vs monetary sector dichotomy; real-balance effect; Samuelson, P.; stability analysis; Walras's Law

Article

Don Patinkin is regarded as the ‘father of the economics profession’ in Israel. Upon his arrival in Israel with his wife Dvora in 1949, he joined the Hebrew University and raised a generation of students trained in modern economics (known as the ‘Patinkin boys’), who were to form the backbone of the economics departments in the various universities, the staff of Treasury and the Bank of Israel, the commercial banks and the other institutions that had a demand for economists. In spite of his young age, he was an economist with outstanding academic achievements, which marked him as a rising star in the economics profession. His choice to live in Israel was a source of much pride to the new state. He passed away in 1995, but the impact of his teaching and of his personality will linger on for many years.

The Chicago days

Don Patinkin was born in 1922 in Chicago to an Orthodox Jewish family that lived in a predominantly Jewish neighbourhood. His early education was a combination of secular and rabbinical studies. In 1943 he enrolled in the economics department of the University of Chicago, obtaining his Ph.D. in 1947. This period left a deep impression on Patinkin and had a great impact on his future work. He was influenced not only by the Chicago Tradition of free-market liberalism, but also by the personalities of his prominent teachers Frank H. Knight, Jacob Viner, Henry C. Simons, Oscar Lange and Lloyd W. Mints. (See his account of his teachers in Patinkin, 1981b.)

In addition, he was awarded a fellowship with the Cowles Commission, then situated in Chicago, which hosted a remarkable number of prominent economists. Patinkin looked back to his Chicago days with pleasure and nostalgia, and considered himself lucky to have benefited from the contact with such ‘giants’. His Ph.D. dissertation, ‘On the consistency of economic models: a theory of involuntary unemployment’, under the supervision of Jacob Marshak (the chairman of the Cowles Commission), is on a topic to which he returned many times over the years without being able to find a satisfactory solution (nor has any other economist).

The years at the Hebrew University

In 1949 he accepted the proposal of the Hebrew University of Jerusalem to serve as a senior lecturer in the economics department. On his very first day Patinkin plunged into this task, directing the transition of the department from the Continental descriptive and institutional framework to the Anglo-Saxon tradition of analytical economics (Barkai, 1993). Professor Alfred Bonne, who chaired the traditional economics department, supported this move. In the first years he taught practically all the courses in microeconomics and macroeconomics, at all levels, and performed this task outstandingly.

It is remarkable that these years of great pressure were also the most fruitful of his career: he completed his monumental book *Money, Interest and Prices* (MIP, 1956) and wrote a number of influential papers in leading journals, usually rebutting criticisms on various topics related to the book (such as the invalid dichotomy, discussed later). In evaluating the numerous reviews of the book, Stanley Fischer (1993) states that practically all the reviewers recognized that they were dealing with a major work.

To build the foundations of the economics profession in Israel, Patinkin sent a group of graduates, whom he considered candidates for an academic career, for Ph.D. studies in the top universities in Britain and the United States. The building of foundations included also the construction of a statistical base of the Israeli economy. To perform this task he was appointed director of the Falk Institute of Economic Research Israel in 1956, where he continued the work of Daniel Creamer and Harold Lubell, the previous directors. It seems that Simon Kuznets, who was involved in formulating the programme of the Falk Institute, had a profound influence on Patinkin's interest in empirical research. In addition to directing numerous research projects at the institute, Patinkin himself wrote on the early years of the Israeli economy (*The Israeli Economy in the First Decade*, 1959a), where the elimination of the monetary overhang associated with repressed inflation fitted well with his model of the real balance effect.

Although he believed in, and represented, the Chicago pro-market creed, he never pushed this approach

forcefully. The very first lesson in his celebrated course ‘Introductory Economics’, modelled on the famous textbook of Samuelson with application to Israel, was about the allocation of scarce resources among competing uses, which could in principle be performed by the market or by a central planning committee.

Patinkin completed his term of office as chairman of the department of economics in 1960, moving on to serve as the Dean of Social Sciences, and from there in 1980 to serve as Rector and finally as President of the Hebrew University. In all these years he maintained his touch with monetary economics, especially from the doctrinal aspect.

Patinkin participated actively in the debates concerning Israel's economic policy problems. In particular, he was critical of the way monetary policy was run. He served on a number of policy committees (for a thorough discussion of this aspect of Patinkin's activity, see Barkai, 1993) and contributed to the daily press of the early 1970s when the inflationary process began. In later years he preferred that the economists that he had raised should handle these matters.

On the occasion of Patinkin's retirement his colleagues organized a conference in his honour. The scientific works of the participants, who included many of the economists that he regarded highly, were published in *Monetary Theory and Thought* (Barkai, Fischer and Liviatan, 1993), which covered topics related to Patinkin's work.

Patinkin's contribution to monetary economics

Patinkin contributed to three main areas in economic theory: his criticism of neoclassical monetary theory, his treatment of involuntary unemployment and his work on the history of economic thought, in particular the writings of Keynes. Patinkin introduced some order into the vague (some may prefer the term ‘chaotic’) state of the monetary model that existed in his time. MIP stands out as a bridge between pre-Keynesian economics, Keynesian economics and the modern economic literature. Its economic rigour, building the macroeconomic model on micro foundations, was unprecedented in the literature on monetary economics.

Patinkin was very critical of the monetary model formulated by the classical and neoclassical theorists. In particular, he claimed that their theory was ‘guilty’ of the ‘invalid dichotomy’ between the determination of relative prices and the absolute price level. More specifically, the dichotomy relates to the separation between the real sector, where relative prices are determined, and the monetary sector, where the absolute price level is determined by some version of the quantity theory of money (the Cambridge equation). He claimed that this dichotomization is invalid because, by Walras's Law, the excess demand for money is just the sum of excess supplies in all other markets and hence must share the same parameters, in particular the money supply. The fact that in the neoclassical formulation the money supply appears only in the money market is self-contradictory. (In the second edition of MIP, 1965, Appendix to ch. 8, Patinkin pointed out that, when the real balance effect is confined to the bond market, it is possible to express the excess demand functions for commodities in terms of relative prices and the interest rate without referring explicitly to the real balance effect.).

To prove that the neoclassical monetary economists adhered in fact to the invalid dichotomy, Patinkin created a ‘database’ of the relevant writings of these economists (summarized in the first and second editions of MIP), and scrutinized carefully the suspect sentences to show that they were unclear and even reckless. There is no doubt that Patinkin was a master of the literature on monetary theory, and he

used it effectively to support his arguments.

Since many people wrote without a formal analytical apparatus in those days, they often said contradictary things concerning the dichotomy, and Patinkin identified and stressed the inconsistencies. The mathematically inclined economists used the formulation of excess demand functions in terms of all the n prices ($p_1,\ldots,p_n$), which can be multiplied by a Lagrange multiplier λ, which could be any positive number. However, in order to reflect the fundamental property of zero degree homogeneity of real excess demand functions with respect to money prices and the nominal money supply, λ has to be set equal to 1/M, where M is the nominal money supply; then it would represent the real balance effect. But Patinkin insisted (and documented) that as a rule they thought of λ as $1/p_n$, that is, they thought of excess demand for commodities as dependent only on relative prices, without taking account of the real balance effect.

The preoccupation with the question of ‘what people really thought’ left a gray area of possible interpretations, which depended on subjective evaluations. Paul Samuelson (1968), who thought that in principle Patinkin's criticism was well taken, nevertheless believed that Patinkin's reading of the earlier theorists was not sympathetic. However, the examples of the articles of Hickman (1950) and Archibald and Lipsey (1958), who tried to defend the invalid dichotomy, made it clear that Patinkin's tough criticism was justified from the point of view of improving professional rigour in economic science (see also Fischer, 1993).

Patinkin's critical evaluation reflects the stringent criteria he applied to the work of his predecessors. He required of monetarist theorists who put money in the utility function to state explicitly the rationale for holding money; he insisted on an explicit reference to the real-balance effect, and he required an understanding of the difference between the individual and market experiments. In addition, he insisted on the incorporation of stability analysis of the money market in the same way as his predecessors analysed the stability of markets for ordinary commodities. He considered the fulfillment of all these criteria necessary for a full integration of money and value theory.

The ‘victims’ of this harsh criticism included such famous names as Walras, Fisher, Pigou and Cassel, in whose writings the presence of the invalid dichotomy was ‘highly probable’, as well as others who were more explicit about it, such as Lange, Modigliani, and Hickman (Patinkin, 1965, p. 175, n.33).

Patinkin enjoyed the role of critical interpreter of texts, which he attributed to his training at the Yeshiva College in Chicago (1994). This perhaps explains his infatuation with Keynes's writings in later years, and his preoccupation with the writings in the Chicago tradition. The former involved mainly the evolution of Keynes's thoughts on effective demand and involuntary unemployment, and the latter focused on the interpretation of the quantity theory of money and the economic philosophy of his famous teachers at the University of Chicago.

The non-technical writings of Keynes (1936) were a fertile ground for interpretations and formulations of formal models attributed to his ideas, and it provided Patinkin with ample room for clarification of Keynes's arguments. For example, he presented a diagrammatic exposition of the Keynesian theory in Patinkin (1982), especially Figures 5 and 6, clarifying the concepts of effective demand and aggregate supply in the Keynesian model. (In Figure 6 it is shown that effective demand is determined at the intersection of aggregate demand and supply – in terms of wage units – as functions of employment. In this diagram the real wage is endogenous to the level of employment on the assumption that firms are on the demand curve for labour. Thus the real wage is indirectly determined by aggregate demand. In this

sense it is not a fixed-price model.) In his analysis of Milton Friedman's statement of the quantity theory of money (Patinkin, 1981) he contrasts it critically with what Patinkin considered the true Chicago tradition.

Involuntary unemployment

While the task of putting the house of neoclassical monetary theory in order involved an in-depth analysis of the early literature, his other major preoccupation was in an area which required his own creativity – involuntary unemployment. This problem, which reflected the realities of the Great Depression of the 1930s, occupied Patinkin's academic interests from his Ph.D. dissertation and throughout his later work. Yet the problem of why the workers could not avoid unemployment by real wage cuts remains basically unresolved to this very day.

Patinkin first approached this problem in his famous early article in the *American Economic Review* (1948), where he claimed that the real interest rate and the real-balance effect might not be sufficiently flexible to allow an equilibrium solution, and even if they did it may take a long time (due to bankruptcies and pessimistic expectations). This may render it politically unrealistic to rely on automatic forces to establish full-employment equilibrium. Patinkin therefore considered unemployment essentially in the context of economic dynamics.

In Chapter 13 of MIP, Patinkin took an additional step in dealing with this issue, arguing that if firms cannot sell their optimal (competitive) output they will not employ their optimal labour input. This gave rise to a new area of research in macroeconomics, namely, disequilibrium models. Barro and Grossman (1971) combined this analysis with the Clower constraint, which postulates (as explained by Barro and Grossman) that, if workers cannot supply their optimal labour services they will not purchase their optimal (competitive) quantity of goods. Barro and Grossman go on to show how equilibrium can be established in the fixed-price model of this type. Over the years, the criticism of these models increased because they required arbitrary rationing rules (Drazen, 1980), and because they were too complicated technically. The disappearance of widespread involuntary unemployment in the post-Second World War era probably had something to do with the growing unpopularity of these models.

It is noteworthy that Patinkin refrained (in the second edition of MIP) from seeking a solution to the problem of involuntary unemployment in the domain of imperfect competition, in spite of Arrow's (1959) remark that in disequilibrium situations the competitive model is problematic. It seems that this is an indication of Patinkin's conservative approach to economic analysis.

Although most of MIP is devoted to the working of Patinkin's model in full employment, the more interesting implications of monetary policy were in connection with unemployment. The latter case gave rise to the fundamental question of whether the capitalist system possesses an *automatic* mechanism for attaining full employment, which is the basic problem that underlies much of Patinkin's work. Perhaps this explains why he was willing to take the risk of dealing with disequilibrium models, although he realized their limitations (1965, ch. 13, n. 9).

Some of the issues which were presented in MIP gave rise to criticism by prominent economists. But in all these confrontations Patinkin had the upper hand. One can cite as an example Hicks's (1957) criticism of Patinkin's interpretation of Keynesian unemployment theory; Patinkin's reply (1959b) in terms of the Hicksian IS–LM model suggested that Hicks did not fully understand Pigou's (1943) mechanism of the

real-balance effect.

Patinkin's early work dealt solely with the static economy, while the profession was concerned in the 1960s with models of economic growth, including monetary growth. This led Patinkin to write a paper, with David Levhari (1968), on monetary growth in the fashion of Tobin's original contribution to these models.

Patinkin's own view of his early work and his critical reflections about the recent developments in economics are interesting. We have a glimpse of these in the introduction to his final, abridged edition of MIP in 1989, 23 years after the publication of the first edition. In this introduction he welcomes the progress that has been made in disequilibrium theory, although he realizes its limitations, since it contradicts some of the tenets of rational expectations. He also welcomes the renewed theoretical work by the neo-Keynesian economists on the rational basis of price and wage rigidities, and discusses the effect of the new developments related to rational expectations. His discussion is certainly very scholarly but short of the original insights that characterized his earlier writings. It seems that rational expectations represent a whole new philosophy that was absent in the writing of the 1950s and 1960s, which one might call the age of innocence.

See Also

- Keynes, John Maynard
- monetary economics, history of.

I am grateful to Akiva Offenbacher and Stanley Fischer for their comments. I also benefited greatly from the comments and suggestions of my veteran colleagues of the economics department of the Hebrew University: Haim Barkai, Nachum Gross, Michael Michaely, Ephraim Kleiman and Gur Ofer.

Selected works

1948. Price flexibility and full employment. *American Economic Review* 37, 543–64.

1956. *Money, Interest and Prices*. Evanston, IL: Row Peterson.

1959a. *The Israeli Economy – The First Decade*. Jerusalem, The Maurice Falk Institute For Economic Research in Israel.

1959b. Keynesian economics rehabilitated: a rejoinder to Professor Hicks. *Economic Journal* 69, 582–7.

1965. *Money, Interest and Prices*, 2nd edn. New York: Harper and Row.

1968. (With D. Levhari.)The role of money in a simple growth model. *American Economic Review* 58, 713–53.

1981a. The Chicago tradition, the quantity theory, and Friedman. In *Essays In and On the Chicago Tradition*. Durham, NC. Duke University Press.

1981b. Introduction: Reminiscences of Chicago 1941–47. In *Essays In and On the Chicago Tradition*. Durham, NC. Duke University Press.

1982. A critique of Keynes' theory of effective demand. In *Anticipations of the General Theory, and Other Essays on Keynes*. Chicago: University of Chicago Press.

1989. *Money, Interest and Prices*, 2nd edn, abridged, with a new introduction. Cambridge, MA: MIT Press.

1994. From Chicago to Jerusalem. *Economic Quarterly*, Hebrew, New Series, 1994(2), 165–90.

Bibliography

Archibald, G. and Lipsey, R. 1958. Monetary and value theory: a critique of Lange and Patinkin. *Review of Economic Studies* 26, 1–22.

Arrow, K. 1959. Towards a theory of price adjustment. In *The Allocation of Economic Resources*, ed. M. Abramovitz et al. Stanford, CA: Stanford University Press.

Barkai, H. 1993. Don Patinkin's contribution to economics in Israel. In *Monetary Theory and Thought*, ed. H. Barkai, S. Fischer and N. Liviatan. London: Macmillan.

Barro, R. and Grossman, H. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

Drazen, A. 1980. Recent developments in macroeconomic disequilibrium theory. *Econometrica* 48, 283–306.

Fischer, S. 1993. Money, interest and prices. In *Monetary Theory and Thought*, ed. H. Barkai, S. Fischer and N. Liviatan. London: Macmillan.

Hickman, W. 1950. The Determinacy of absolute prices in classical economic theory. *Econometrica* 18, 9–20.

Hicks, J. 1957. A rehabilitation of 'classical economics'? *Economic Journal* 47, 278–89.

Keynes, J. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Pigou, A. 1943. The classical stationary state. *Economic Journal* 53, 343–51.

Samuelson, P. 1968. What classical and neoclassical monetary theory really was. *Canadian Journal of Economics* 1, 1–15.

How to cite this article

Liviatan, Nissan. "Patinkin, Don (1922–1955)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_D000243> doi:10.1057/9780230226203.1259

The New Palgrave Dictionary of Economics Online

Patten, Simon Nelson (1852 – 1922)

A.W. Coats

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Patten, S. N.; protection; James, E. J.

Article

One of the most original and idiosyncratic American economists of his generation, Patten was born at Sandwich, Illinois on 1 May 1852 and studied at Jennings Seminary, Aurora, Illinois. There he met Joseph French Johnson, later a colleague at the University of Pennsylvania, whom he followed to Halle in 1876 after spending only 18 months as a freshman at Northwestern University. At Halle Patten obtained the Ph.D. degree remarkably quickly, in 1878, and he encountered two major personal influences, his teacher Johannes Conrad and a fellow American student, Edward Janes James, who was eventually instrumental in securing Patten's appointment at the University of Pennsylvania in 1888, where he remained throughout his academic career. In the intervening period, however, like Thorstein Veblen, Patten had been unable to get a university post despite the publication of his highly original *Premises of Political Economy* (1885), and was obliged to work on a farm and teach in various public schools, partly because of his poor eyesight.

Once at Philadelphia, Patten proved to be a profoundly stimulating pedagogue and author of a series of unusual, even eccentric books that challenged, provoked and sometimes baffled his professional peers. In harmony with the Wharton School tradition, he was an ardent protectionist, believing that trade barriers would stave off the dangers envisaged by Ricardo and Malthus. Adopting an optimistic, teleological view of the prospects for American abundance, provided that crop variations could be developed to counteract soil exhaustion, Patten insisted that economic laws were not natural, but social. His conception of economics was broad, as in the German tradition, yet his own work was abstract and deductive rather than heavily empirical or statistical. Together with James, he tried in 1884 to form a Society for the Study of National Economy, modelled on Conrad's suggestions, but when this failed to gain sufficient support they joined Ely and others in launching the American Economic Association, of which Patten was elected president in (1908–9). Patten's concepts of the laws of pleasure and pain, his theory of consumption, and his idea of the social surplus were intriguing but puzzlingly novel and

unsystematic, yet his awareness of the costs of growth and his concern for the environment anticipated late 20th-century anxieties.

Selected works

1885. *The Premises of Political Economy: being a re-examination of certain fundamental principles of economic science*. Philadelphia: J.B. Lippincott.

1889. *The Consumption of Wealth*. Philadelphia: T. & J.W. Johnson. 2nd edn, Philadelphia: published for the University; Boston: Ginn & Co., 1901.

1890a. *The Economic Basis of Protection*. Philadelphia: J.B. Lippincott.

1890b. *The Educational Value of Political Economy*. Baltimore: American Economic Association.

1896. *The Theory of Social Forces*. Philadelphia: American Academy of Political and Social Science.

1899. *The Development of English Thought: a study in the economic interpretation of history*. New York and London: Macmillan.

1902. *The Theory of Prosperity*. New York and London: Macmillan.

1924. *Essays in Economic Theory*, ed. R.G. Tugwell. New York: A.A. Knopf.

Bibliography

Fox, D.M. 1967. *The Discovery of Abundance: Simon N. Patten and the Transformation of Social Theory*. Ithaca: Cornell University Press.

How to cite this article

Coats, A. W. "Patten, Simon Nelson (1852–1922)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000041> doi:10.1057/9780230226203.1260

The New Palgrave Dictionary of Economics Online

payment systems

William Roberds
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Payment systems are arrangements that allow for the discharging of debts by the transfer of specialized claims. This article illustrates how payment systems can facilitate exchange in economic environments where enforcement of obligations is limited, and collateral is scarce.

Keywords

banking industry; banknotes; bills of exchange; central banking; fiat money; inside money; intermediate goods; net settlement; network effects; outside money; payment systems; settlement; Wicksell triangle

Article

A *payment* occurs when one party, the *payer*, transfers an asset to another party, the *payee*, for the purpose of discharging a debt incurred by the payer. Or, a payment may consist of the payer's instruction to a third party to make such a transfer, as is the case with a cheque payment. While in principle a payment may be made with any asset, in practice virtually all modern payments involve transfers of debt claims on either central banks (including ‘outside money’ in the form of both currency and deposits) or private banks (‘inside money’, today almost always in the form of deposits). Available evidence suggests that most payments are still made in cash, but these transactions tend to be for relatively small amounts. By value, the wide majority of payments involve transfer of bank deposits by various means.

A payment may or may not constitute *settlement*, a legal discharge of a debt. In most countries, for example, a payment by means of a transfer of claims on a central bank unconditionally settles a debt, whereas other types of payment settle a debt only after certain conditions have been fulfilled (for example, after a cheque has been honoured by the bank on which it is drawn).

A payment *system* is a collection of technologies, laws, and contracts that allow payments to occur and determine when a payment effects a settlement. Payment systems include currency, cheques, credit and debit cards, electronic funds transfers, and so on. Developed economies depend critically on the near-flawless operation of such systems. By offering debtors low-cost and trustworthy means of settling their debts, payment systems provide an important stimulus to the use of credit, and to economic activity more generally.

Some simple statistics illustrate these assertions: in the year 2003, 81 billion payments of $824 trillion were recorded in the United States, not counting payments made in currency (Committee on Payment and Settlement Systems, 2005). Another way of framing these numbers is to note that they imply, on average, $75 in non-cash payments for each dollar of final output produced in the United States in 2003. During the same year each US resident made 278

non-cash payments on average. All developed economies display similar levels of payments activity.

Theory of payments

Despite their ubiquity and their obviously central role in modern economies, payments have only recently begun to make their way into mainstream economic theory. Payment systems do not exist in Arrow–Debreu economies, where transfers may always be made in kind, and promises to transfer are enforced by a social planner. In these economies there is no need for specialized assets to allow for payments, technologies for transferring these assets, or rules concerning when such transfers settle a debt.

Even if the planner's ability to enforce promises is limited, payments may still be inessential. Agents will have incentives to honour their obligations so long as they have access to sufficient amounts of collateral that can be attached by creditors after a default. Payment systems become relevant when enforcement is limited and collateral is scarce. In such environments, payment systems serve as devices that allow for enforcement of debts while making efficient use of available collateral.

One commonly available type of collateral is, of course, outside fiat money, but a discussion of the comparative payment roles of inside and outside money is beyond the scope of this essay. Two influential papers in this area have been Freeman (1996; see especially its discussion in Green, 1999) and Cavalcanti and Wallace (1999). For the reminder of this article I will concentrate on payments in private debt.

An illustration

To demonstrate the function of payment systems, I consider some models of payment based on the celebrated ‘Wicksell triangle’ depicted in Figure 1. Each of the three agents is endowed with a unit of a generic numeraire good. Agent *A* has the possibility of converting this good into a ‘customized’ good that is (highly) desired by agent *B*, who can convert his numeraire into a good desired by agent *C*, who can produce a good that is desired by *A*. Barring difficulties in enforcement, efficiency would require each agent to produce the appropriate customized good and deliver it to the next agent. I call this allocation the *full-enforcement efficient* allocation.

Figure 1

Wicksell triangle

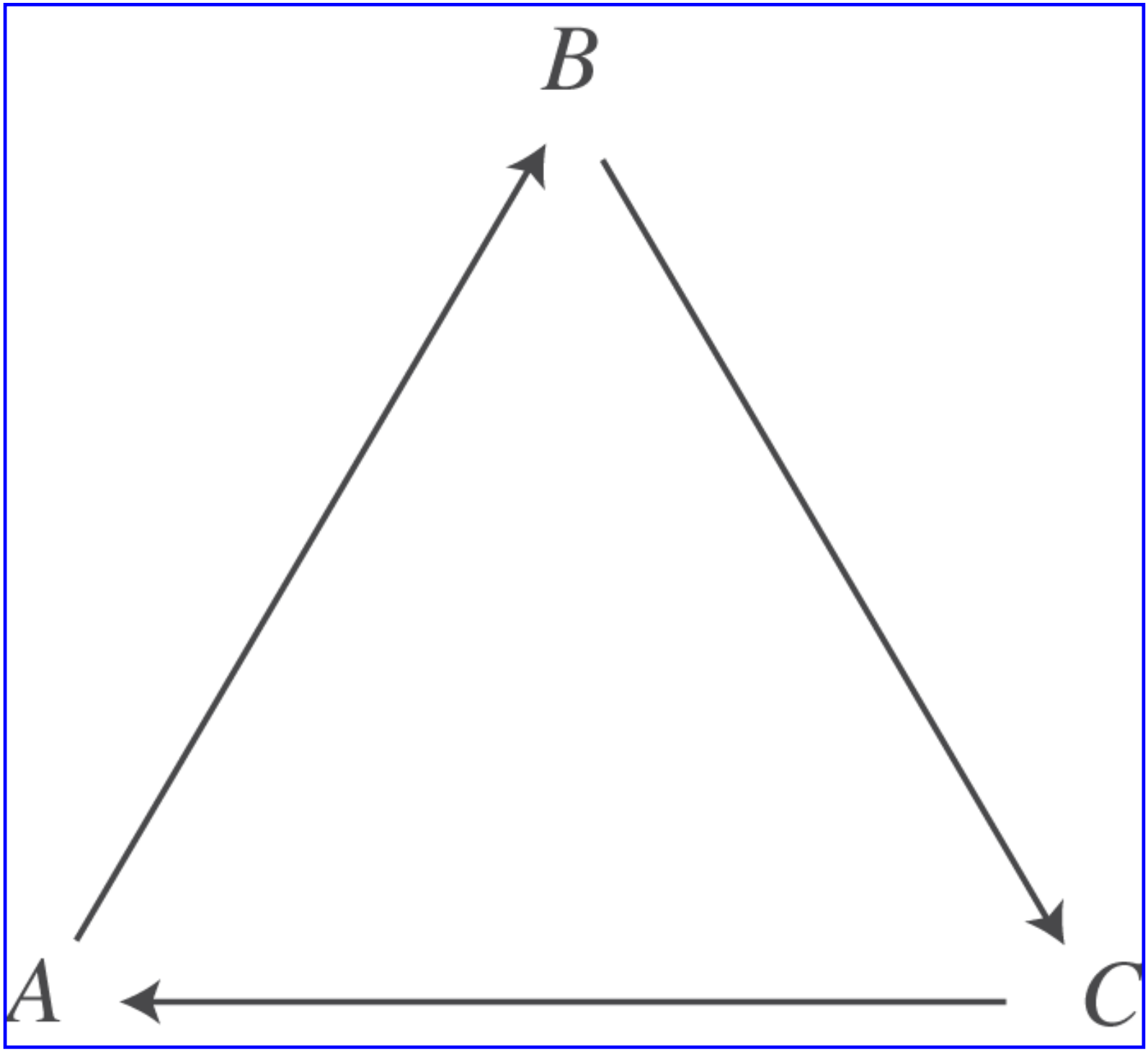

The general goal of payment systems is to deliver an allocation that approximates this allocation, to the extent this is feasible under limited enforcement. I now consider to what extent various payment systems are able to do this. In each of these environments, any enforcement actions will occur through a fourth agent known as the *centre* or ‘central counterparty’, who has a restricted ability to punish agents who default on their obligations. Punishments may include limited fines, attachment of collateral, and public announcements of a default.

Payment model 1: ‘ netting ’

Kahn, McAndrews and Roberds (2003) analyse the following version of the Wicksell-triangle environment. *A*, *B*, and *C* each consists of a buyer-seller pair who live at a separate ‘location’, meaning that trade occurs as bilateral encounters between buyer and seller. Agents are not particularly inclined to keep their promises, but may post some numeraire as collateral before trading begins. There is a single period during which sellers can visit buyers and

transfers of customized goods can occur, and a subsequent period during which numeraire may be transferred. Prices of goods are given in numeraire and are determined through bilateral negotiations.

In this environment, it is easy to show that the amount of collateral required for trade can be minimized by the use of a payment system based on *net settlement*. However, as net settlement typically requires the diversion of resources in order to acquire and post costly collateral, its use will entail a welfare loss, relative to the full-enforcement efficient allocation.

Under net settlement, after trades have occurred, the central counterparty sums for each agent the amount the agent owes *to* the seller he bought from, minus the amount owed *from* the agent from the buyer sold to. If this sum is positive, the agent transfers numeraire to the central counterparty, and if the amount is negative, he receives numeraire from the central counterparty.

Payment in this environment simply consists of an agent's declaration of his intent to settle, and this occurs simultaneous with trade. *Settlement* is the two-stage process of (*a*) replacing gross obligations with net obligations and (*b*) discharging net obligations through transfer of numeraire.

A characteristic feature of net settlement is ‘set-off,’ under which a debt owed *by* party *X* is enforced by cancelling it (‘setting it off’) against its debt owed *to* party *X*. In this fashion, agent *X*'s creditor may exercise a de facto prior claim against *X*, even when other means of exercising priority are costly (such as posting additional collateral). Payment systems incorporating net settlement allow set-off to occur in a regular and predictable fashion.

Netting of obligations is an ancient method of payment, dating at least to the 13th century fairs of Champagne (Kohn, 2001). It continues to be used extensively for settling high-value, recurring obligations such as those that arise between commercial banks (for example, the CHIPS system which operates in the United States). But there are certain limitations that prevent its more widespread use. The first is that there may be an inadequate legal basis for netting (Bliss, 2003). Second, netting works well only if all parties involved are of roughly equal creditworthiness (Kahn and Roberds, 2003). Finally, netting may require too much coordination in the sense that all parties must agree in advance to participate in the netting arrangement. These limitations have given rise to other forms of payment systems which, in effect, allow netting to occur in a more decentralized fashion.

Payment model 2: ‘ banknote ’

Kiyotaki and Moore (2000) discuss a slightly different model from model 1 above. Suppose that preferences and endowments are the same as above, but that bilateral encounters between agents are separated in time: agent *C* first has an opportunity to buy his desired good from agent *B*, who then has an opportunity to buy from *A*, who can then buy from agent *C*. Agent *C* is known to be creditworthy but *A* and *B* are not.

In this model, the full-enforcement efficient allocation can be implemented if *C*'s debt can ‘circulate’. More specifically, *B* receives debt from *C* in return for a customized good. Agent *B* then trades *C*'s debt to *A*, in return for *A*'s customized good. *A* then presents *C*'s debt to *C* for redemption. Finally, agent *C* completes the cycle of trade by transferring a customized good to *A*. *Payment* in this environment corresponds to either the issue (by *C*) or transfer (by *B*) of *C*'s debt. If *C* is sufficiently creditworthy, *B*'s transfer of *C*'s debt will also constitute a *settlement*. Otherwise settlement may not occur until *C* redeems his debt.

Under this arrangement it is not necessary for all parties to be creditworthy for trade to occur. Agent *C* may enjoy some natural advantage in this regard. This advantage could take the form of ownership of attachable assets or, in a dynamic setting, it could be that people have better information on the actions of *C* than on the actions of other agents (Cavalcanti and Wallace, 1999). In this arrangement, *C*'s debt becomes a form of specialized asset for use in payment, a ‘banknote’.

This is the basic model for many transactions using not only privately issued banknotes (which are rarely observed nowadays) but also other means of transferring debt claims. A retail store may not be willing to accept a customer's IOU in exchange for merchandise but is perfectly willing to accept a debt (that is, deposit) claim on a bank, transferred by means of a credit or debit card.

This form of payment also has a long history. One of the most famous early examples is from 15th-century Genoa.

There, payments were commonly made using claims on an institution responsible for managing the debt of the state (the *Casa di San Giorgio*; see Kohn, 1999). Under this arrangement agent *C* became, in effect, an agent of the state, whose creditworthiness derived from the taxation powers delegated to it. People owing taxes could use claims on the *Casa di San Giorgio* to discharge their own tax obligations, which generated a demand for these claims as payment instruments.

Note that model 2, like model 1, involves a form of netting. When a consumer purchases merchandise with, say, a debit card, the consumer is in effect netting out the debt he owes *to* the merchant against debt (deposits) owed him *by* his bank. In contrast to model 1, however, there need be no prior agreement between merchant and consumer, given sufficient trust in the banking system.

Payment model 3: ‘bank loan’

Model 2 illustrates how payment systems allow netting to occur in a decentralized fashion. This model is inadequate for some situations, however, because it does not explain the simultaneous existence of both liquid and illiquid debt. In particular, this model is inappropriate for production economies where a producer may require prompt delivery of an intermediate good now in order to produce a final good that can be sold only later. In such situations, working capital is typically provided by the issue of debt.

To remedy this shortcoming, some studies have attempted to modify model 2 in order to incorporate both transferable (‘liquid’) and non-transferable (‘illiquid’) debt. Kiyotaki and Moore (2000) consider a model which maps into the following variation. Suppose that the timing of the first two transactions in the Wicksell triangle is reversed, so that that agent *B* first has an opportunity to buy from *A*, then *C* from *B*, and finally *A* from *C*. This timing is natural if *B* uses *A*'s good as an intermediate good.

As in model 2, agent *C* is trustworthy but agents *A* and *B* may not be. In addition, Agent *C* enjoys a special privilege as a creditor, that is, an enhanced ability to enforce debts, and serves as ‘banker’ to agent *B*.

In this modified example, it is possible to show that the full-enforcement efficient allocation can sometimes be implemented through use of a combination of transferable and non-transferable debt. Specifically, suppose that *B* has an opportunity to meet with *C* before production of specialized goods can occur, and before trading begins. Agent *B* issues debt to *C*, and *C* in turn issues debt to *B*. When *B* then encounters *A*, he pays for *A*'s specialized good by transferring *C*'s debt to *A*. Agent *B* then has the opportunity to discharge his debt to agent *C* by transferring his specialized good to *C*. Finally, agent *A* presents *C* with his debt, and receives *C*'s specialized good. *Payment* and *settlement* are defined as in model 2.

In short, in this model agent *C* is engaged in ‘liquidity transformation’, which consists of holding *B*'s debt, which would be unenforceable by *A*, while issuing to *B* his own enforceable and therefore transferable debt. In practice, this liquidity transformation is usually provided by banks. This function of banks was already well established by the 14th century (Kohn, 2001).

Payment model 4: ‘bill of exchange’

Model 3 allows for the coexistence of liquid and illiquid debt, but may not be appropriate for all circumstances. In some environments, there may be no agents with special enforcement abilities, such as agent *C* above. This is particularly true for economies with less developed legal and financial systems. Yet through the process of payment it may still be possible to economize on resources devoted to enforcement, by allowing for the discharge of one debt by the transfer of another.

Kahn and Roberds (2001) consider the following variation on model 2. The order of meetings is *A* with *B*, *B* with *C*, and *C* with *A*. The customized good produced by agent *C* is now valued by both *A* and *B*.

The full-enforcement efficient allocation can then be supported as follows. Suppose that agent *B* issues debt to *A* in the first transaction, and that agent *C* issues debt to *B* in the second transaction, which is subsequently passed to *A*. In the final transaction, *A* presents *C*'s debt to *C*, and *C* redeems his debt by providing the appropriate good to *A*.

Payment in this environment again corresponds to the passing of *C*'s debt by *B* to *A*, and *settlement* occurs either simultaneously with payment, or when *C* redeems his debt.

The intuition behind the efficiency of this arrangement is as follows. Suppose that, instead of making use of transferable debt, trade is organized as a ‘credit chain’ (Kiyotaki and Moore, 1997), in which B issues debt to A, C issues debt to B, and B promises to discharge his debt with A once he has collected from C. If enforcement is less than perfect and B also values C's customized good, then B may collect C's debt then ‘take the money and run’, that is, abscond with C's good. But if *A* requires an ‘early’ payment from *B* in the form of a transfer of *C*'s debt, *B*'s default can be averted, provided that *A* can respond to a failure to pay at this stage by preventing *B* from collecting with *C*.

As in the models above, enforcement of *B*'s obligation to *A* occurs through a form of netting. By requiring early payment from *B* in the form of *C*'s transferable debt, *A* is in effect forcing *B* to cancel one debt with another. The key distinction between model 4 and earlier models is that this cancellation is no longer instantaneous. In other words, even potentially bad credits such as *B* are allowed to issue debts as long as they agree to punctually pay them off using the debt of another, possibly stronger credit.

The work of economic historians (see Ashtor, 1972) suggests that model 4 is also an ancient one. Its use in the West (in the form of bills of exchange and similar instruments) dates from the late 12th century, and likely arose from even earlier Middle Eastern precedents. Even in today's advanced economies, this model persists in the form of trade credit that is granted with the understanding it will be repaid in another form of debt, nowadays typically bank funds.

Payments and networks

Payment systems based on the models discussed above have been in use for some time. Successful application of these models, however, requires some information which may not always be present in practice. At a minimum, participants in these arrangements must be able to distinguish the identity of their counterparties, and have some notion of their counterparties’ ability to honour their debts. Historically, these requirements have often worked to limit the use of many forms of non-cash payments to established businesses, wealthy individuals, or parties already well known to each other.

These constraints have become less onerous with improvements in information technology. In particular, the years since 1960 have seen rapid development of electronic payment systems based on the use of cards (Evans and Schmalensee, 1999). A noteworthy distinction between electronic systems and their paper-based counterparts is that the new systems require the use of specialized communications networks.

As is the case with other industries, the presence of ‘network effects’ in payment systems leads to complications (see Weinberg, 1997). Baxter (1983) was the first to point out the essentially ‘two-sided’ nature of the service provided by these networks: that is, that efficiency in these networks may depend critically on the allocation of their costs between buyers and sellers. This insight has been subsequently expanded on by many authors (an authoritative survey is given in Rochet and Tirole, 2004). Nonetheless, as of this writing, no consensus has emerged concerning efficient allocation of services provided by these systems (Evans and Schmalensee, 2005).

Conclusion

Payment systems are an important component of decentralized exchange. This article has illustrated how the fundamental role of these systems is the reduction of chains of obligations to a smaller and more readily enforceable set of obligations. Ongoing improvements in information technology have the potential to increase the scope and efficiency of payment systems, and this will require economists to provide more precise models of their function and essential nature.

See Also

- banking industry
- bankruptcy, economics of
- inside and outside money
- money and general equilibrium
- network goods (theory)

Bibliography

Ashtor, E. 1972. Banking instruments between the Muslim East and Christian West. *Journal of European Economic History* 1, 553–73.

Baxter, W. 1983. Bank exchange of transactional paper: legal and economic perspectives. *Journal of Law and Economics* 26, 541–88.

Bliss, R. 2003. Bankruptcy law and large complex financial organizations: a primer. *Federal Reserve Bank of Chicago Economic Perspectives* 27(1), 48–58.

Cavalcanti, R. and Wallace, N. 1999. A model of private banknote issue. *Review of Economic Dynamics* 2, 104–36.

Committee on Payment and Settlement Systems. 2005. *Statistics on Payment and Settlement Systems: Figures for 2003*. Basel: Bank for International Settlements.

Evans, D. and Schmalensee, R. 1999. *Paying with Plastic: The Digital Revolution in Buying and Borrowing*. Cambridge, MA: MIT Press.

Evans, D. and Schmalensee, R. 2005. The economics of interchange fees and their regulation: an overview. Working Paper No. 18181. Cambridge, MA: Sloan School of Management, MIT.

Freeman, S. 1996. The payments system, liquidity, and rediscounting. *American Economic Review* 86, 1126–38.

Green, E. 1999. We need to think straight about electronic payments. *Journal of Money, Credit, and Banking* 31, 668–70.

Kahn, C. and Roberds, W. 2001. Transferability, finality, and debt settlement. Working paper. Federal Reserve Bank of Atlanta.

Kahn, C. and Roberds, W. 2003. Payments settlement under limited enforcement: public versus private systems. Working paper. Federal Reserve Bank of Atlanta.

Kahn, C., McAndrews, J. and Roberds, W. 2003. Settlement risk under net and gross settlement. *Journal of Money, Credit, and Banking* 47, 299–319.

Kiyotaki, N. and Moore, J. 1997. Credit chains. Mimeo. London School of Economics.

Kiyotaki, N. and Moore, J. 2000. Inside money and liquidity. Mimeo. London School of Economics.

Kohn, M. 1999. The capital market before 1600. Working Paper No. 99-06. Department of Economics, Dartmouth College.

Kohn, M. 2001. Payments and the development of finance in pre-industrial Europe. Working Paper No. 01-05. Department of Economics, Dartmouth College.

Rochet, J.C. and Tirole, J. 2004. Two-sided markets: an overview. Working Paper No. 275. IDEI, University of Toulouse.

Weinberg, J. 1997. The organization of private payment networks. *Federal Reserve Bank of Richmond Economic Quarterly* 83(2), 25–43.

How to cite this article

Roberds, William. "payment systems." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000336> doi:10.1057/9780230226203.1261

The New Palgrave Dictionary of Economics Online

peasant economy

N.H. Stern

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

credit; Green Revolution; land reform; peasant economy; rational behaviour; rural–urban migration; sharecropping; urban unemployment

Article

A peasant is someone who lives in the country and works on the land (the word derives from the French *paysan*). Taking this definition, the topic ‘peasant economy’ concerns the analysis of the economic decisions and interactions of peasants, their relations with other agents and the rest of the economy, the determinants of the general level and distribution of their economic welfare, and how their position might move over time or be affected by policy. As such it is very broad in scope, involving the study of the economic life of around half the world's population. The term ‘peasant’ is sometimes used in a somewhat narrower sense in economics to mean the small farmer (tenant or smallholder) as opposed to the agricultural labourer or very large landowner. The peasant economy would then be one where farming was conducted mainly by tenants and smallholders. Even under this narrower definition it is clear that vast numbers of individuals are included.

There are fundamental differences amongst economists in their views of the way in which the peasant economy functions and these underlie many of the strong disagreements over policy. The main sources of the differences concern views on the ‘rationality’ of economic behaviour by individuals, the competitiveness and efficiency of markets, the importance and implications of the distribution of power and wealth, and the role of institutions, cultures and beliefs. Whilst we cannot provide a detailed description of these general views we shall try to give a flavour of their diversity and focus on the basis of their differences. We shall then examine some specific issues and problems including the objectives of peasants and others in terms of profit, utility and attitudes to risk; labour markets; credit; sharecropping; and relationships between size of holding and productivity. Concentration will be on the literature since the Second World War, although many of the issues concerned and divided some of the outstanding economists of the 19th and early 20th century.

One of the most clearly stated and definite views places the peasant economy firmly within the standard

competitive analysis; see for example Schultz (1964). Within the constraints of their knowledge, it is argued, participants in the peasant economy make the best use of the assets available to them. Each agent makes production, working and spending decisions to maximize utility or profit. This is essentially the notion of rationality in this context: individuals have preferences and act according to them. Markets for labour, land, credit, inputs and outputs, consumer purchases and so on function competitively and efficiently. The outcome for the usual reasons is therefore a Pareto efficient allocation. The role of policy is then to improve knowledge, increase assets and, if desired, to improve the distribution of income.

At another extreme we find the views of those such as Myrdal (1968), who believes that markets and prices play a minimal role. He argues that few people calculate in terms of costs and returns and that, even if they do, such calculations are not the primary determinants of their behaviour. Further, he argues that many transactions are not of the market type at all, and where markets do exist they are very far from perfect. He pleads for an institutional analysis of behaviour and the workings of the economy. Further, he suggests direct controls to implement policy; he calls these non-discretionary controls as opposed to the manipulation of prices, where individuals are left to take their own decisions.

Away from these extremes we have varying emphases on the role of rational behaviour, incentives and market structure. For example, Lewis (1955) regards institutions, legal structures and political and religious attitudes and practices as major determinants of the form of incentives. Thus he suggests that land reform may be a prerequisite to successful agricultural extension if, without it, farmers believe that others will reap the fruits of their improvements. Since the 1970s there has been substantial concentration on the forms of peasant arrangements for cultivation, the incentives which they give and the reasons for their selection. A central example (discussed briefly below) has been the study of sharecropping following the questions and analysis of Marshall in Chapter X, Book VI, of his *Principles of Economics*. In this context individuals are seen as rational but face problems of information and supervision in designing and implementing agreements for the use of land, labour and other inputs. Marxist writers have emphasized property and power. For example, Bhaduri (1973) suggests that landlords manipulate indebtedness over their labourers and tenants to maintain a very tight hold over their freedom. He argues from his model that landlords have an incentive to block technical change and that progress requires expropriation.

These views are generalizations about the world and no single study could provide a conclusive test between them. An empirical judgement should be based on the accumulated experience of detailed studies. Here economists have not been as active as perhaps they should in conducting economic studies of peasant societies to examine how the theories they are discussing fare in the field (compare the many studies by anthropologists; see, for example, Srinivas, 1960 and 1976, and Wiser and Wiser, 1971). Nevertheless, many studies are available (see for example, Bailey, 1957; Epstein, 1962; Haswell, 1975; Bell, 1977; Bliss and Stern, 1982, and for further references Binswanger and Rosenzweig, 1984; see also the bibliographies of village studies prepared at the Institute of Development Studies, Sussex – Lambert, 1976 and 1978). One should not perhaps expect a clear, single picture to emerge; people and societies vary considerably. However, it seems that neither of the simple descriptions of Myrdal and Schultz are remotely adequate as generalizations. The institutional structure and conventions concerning the disposition of land and labour (for example, the form of ownership and duties of owners, structure of tenancy agreements, restrictions on the obligations of labourers and so on) will be of considerable

importance in determining cultivation decisions. Individuals vary greatly in their ability to make the most of their circumstances. Nevertheless, most of the studies point to strong economic responses and these are often rapid and subtle: the Myrdal picture is clearly unacceptable.

We comment briefly on some of the particular positive issues that have been prominent in theory and applied work. The objectives of peasants have been modelled in terms of profit and utility and in varying ways relative to uncertainty (for an early discussion see Chayanov, 1925). Thus, for example, Hopper (1965) suggests that simple maximization of expected profit provided a good description of farming decisions in the village he studied in North India. This seems implausible in a poor society and for a risky activity, and a number of models of behaviour under uncertainty have been considered. These include the standard model of expected utility maximization and ‘survival algorithms’ where individuals attempt to minimize the probability of falling below ‘disaster level’. The implications can be very different from simple profit maximization. Under expected utility maximization with risk aversion the expected value of the marginal product of an input would, in equilibrium, be above the price of the input (possibly well above) whereas with profit maximization we must have equality (see, for example, Bliss and Stern, 1982).

Two central issues in discussion of the labour market have been, first, the relationship between wages and the marginal product of labour and, second, migration. On the former some appear to have argued that the marginal product is zero. This receives little empirical or theoretical support in that an extra hour of work in agriculture usually has some contribution to production. The question of whether the withdrawal of an extra person from agriculture reduces output and by how much depends on the response to the departure by others. Whilst the marginal product of an hour or day is unlikely to be zero, it is quite possible that it may be less than the wage in the case of family labour where there are perceived costs in working for others or of hiring labour (see for example Sen, 1975).

Migration decisions have been examined extensively, both in theory and practice, in terms of expected differences in net incomes or utility from making a move. Of particular influence was the paper by Todaro (1969) in which he proposed a model where the probability of employment in the town was equal to the number of jobs divided by the number of seekers. If rural and urban wages and urban employment are fixed, the number of seekers adjusts to make, in equilibrium, the expected urban wage equal to the rural wage. If we associate the job seekers with the employed plus the unemployed then this is a theory of urban unemployment with the striking implication that an increase in the number of urban jobs increases unemployment. The model has been extended, elaborated and tested by many authors (see, in particular, Fields, 1975; Sabot, 1982; Todaro, 1976).

The role of credit, for example, the much easier access and cheaper rates available to the richer farmers (Griffin, 1974) and its use in manipulation and control (Bhaduri, 1973) have been major issues. It is an area where data are particularly difficult to collect and good empirical studies are rare (a notable exception in the context of fishing is Platteau, Murickan and Delbar, 1985).

Share-cropping was discussed carefully by Marshall in his *Principles*. Following the book by Cheung (1969), it has become a popular issue in recent research. Cheung contrasted his view of sharecropping as an efficient arrangement (with the tenancy contract clearly defined to stipulate inputs) with that of Marshall, who had pointed to the possibility that the tenant who receives half the output may not push the level of an input as far as someone who receives the full amount of the marginal product. Many of Cheung's arguments were, however, anticipated by Marshall in his account which contains a description

of how the landlord might try to enforce higher input levels. More recently attention has been focused on sharecropping as a means of sharing risk between landlord and tenant and as providing incentives for the tenant which would not be present under simple wage labour (see Binswanger and Rosenzweig, 1984, for references).

The proposition that larger holdings may have lower output per acre has been the subject of much theoretical and empirical discussion. In Indian studies it receives more support for comparisons across districts than within villages. Possible reasons for the phenomenon, where it occurs, include more labour input per acre on smaller family plots (where labour may be applied beyond the point where the marginal product is equal to the wage) and faster population growth (and thus greater subdivision of holdings) on fertile land. For further discussion, see Sen (1975).

On the policy side some of the major issues have been land reform, the dissemination of technical change, the pricing of output and the supply and pricing of crucial inputs such as water, fertilizer and draught power. We shall be very brief since our main emphasis has been on the functioning of the peasant economy. Land reform in the sense of redistribution has been very difficult to achieve, in part because many of those who have it will make great efforts to resist losing it. It has sometimes been argued that the (supposed) inverse relationship between size of holding and land productivity will imply that a more egalitarian distribution of land will yield higher total output. Agricultural extension has long been seen as part of government policy, but it has become particularly prominent with the arrival of the newer varieties of seeds (the so-called ‘Green Revolution’) which are particularly responsive to water and fertilizers. Of special concern has been the differential impact of the advances on different groups in the population and how the changes might be influenced to provide greater benefits to the poor.

The relative price of food and the implicit or explicit taxation of peasants have been seen as critical aspects of the availability of food (and its price) to the rest of the economy as well as influencing growth within and outside peasant agriculture. Much turns on the assumed elasticity of response. A further important feature of government policy concerns the pricing and supply of inputs. The effects on agricultural production and on the welfare of peasants and labourers can be substantial, the most obvious example being irrigation.

The study of the peasant economy is a subject for which careful economic theorizing is critical since transactions can have special structures, uncertainty will be central, and economic relations will be strongly influenced by institutional arrangements. And those theories should be tested against, and arise from, detailed empirical observation since the successful application of the theories turns on which of the structures are relevant for the particular peasant economy under examination.

See Also

- agriculture and economic development
- peasants
- sharecropping

Bibliography

Bailey, F.G. 1957. *Caste and the Economic Frontier*. Manchester: Manchester University Press.

Bell, C.L.G. 1977. Alternative theories of share-cropping: some tests using evidence from North-East India. *Journal of Development Studies* 13, 317–46.

Bhaduri, A. 1973. Agricultural backwardness under semi-feudalism. *Economic Journal* 83, 120–37.

Binswanger, H.P. and Rosenzweig, M.R. 1984. *Contractual Arrangements, Employment, and Wages in Rural Labour Markets in Asia*. New Haven: Yale University Press.

Bliss, C.J. and Stern, N.H. 1982. *Palanpur: The Economy of an Indian Village*. Oxford: Oxford University Press.

Chayanov, A.V. 1925. *Organizatsiya krest'yanskogo khozyaistva* [Peasant farm organization]. Moscow. Trans. as *The Theory of Peasant Economy*, ed. D. Thorner, B. Kerblay and R.E.F. Smith, Homewood, IL: Irwin (AEA Translation Series), 1966.

Cheung, S.N.S. 1969. *The Theory of Share Tenancy*. Chicago: University of Chicago Press.

Epstein, T.S. 1962. *Economic Development and Social Change in South India*. Manchester: Manchester University Press.

Fields, G.S. 1975. Rural urban migration, urban unemployment and underemployment, and job search activities in LDCs. *Journal of Development Economics* 2, 165–87.

Griffin, K.B. 1974. *The Political Economy of Agrarian Change*. London: Macmillan.

Haswell, M. 1975. *The Nature of Poverty*. London: Macmillan.

Hopper, W.D. 1965. Allocation efficiency in 'traditional Indian agriculture'. *Journal of Farm Economics* 47, 611–24.

Lambert, C.M., ed. 1976. *Village Studies I*. Institute of Development Studies, University of Sussex.

Lambert, C.M., ed. 1978. *Village Studies II*. Institute of Development Studies, University of Sussex.

Lewis, W.A. 1955. *The Theory of Economic Growth*. London: George Allen & Unwin.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan, 1959.

Myrdal, G. 1968. *Asian Drama: An Enquiry into the Poverty of Nations*. Harmondsworth: Allen Lane.

Platteau, J.-P., Murickan, J. and Delbar, E. 1985. *Technology, Credit and Indebtedness in Marine Fishing*. Delhi: Hindustan Publishing Corporation.

Sabot, R.H., ed. 1982. *Migration and the Labour Market in LDCs*. Boulder, CO: Westview.

Schultz, T.W. 1964. *Transforming Traditional Agriculture*. New Haven: Yale University Press.

Sen, A.K. 1975. *Employment, Technology and Development*. Oxford: Oxford University Press.

Srinivas, M.N., ed. 1960. *India's Villages*. Bombay: Asia Publishing House.

Srinivas, M.N. 1976. *The Remembered Village*. Oxford: Oxford University Press.

Todaro, M.P. 1969. A model of labour migration and urban unemployment in less developed countries. *American Economic Review* 59, 138–48.

Todaro, M.P. 1976. *Internal Migration in Developing Countries: A Review of Theory, Evidence, Methodology and Research Priorities*. Geneva: International Labour Office.

Wiser, W.H. and Wiser, C.V. 1971. *Behind Mud Walls, 1930–1960*. Berkeley: University of California Press.

How to cite this article

Stern, N. H. "peasant economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000047> doi:10.1057/9780230226203.1263

The New Palgrave Dictionary of Economics Online

peasants

Keijiro Otsuka
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

While traditionally peasants are regarded as subsistence-oriented, full-time, and small-scale farmers, many small-farmers are part-time farmers engaged in both cash- and food-crop farming and non-farm jobs. Therefore, peasants may be defined as small-scale, family based farmers, including both owner cultivators and tenants. A major question is whether the peasant mode of production is socially efficient. Because of the absence of scale economies, the advantage of risk sharing under share tenancy contracts, and the inefficiency of agricultural labour contracts due to the difficulty of supervision, small-scale family based farming system, including share tenancy, is a socially efficient system in low-wage economies.

Keywords

access to land; cash crops; contract choice; contract enforcement; credit markets; efficient allocation; family labour; fixed-rent tenancy; food security; human capital; insurance markets; land use rights; market failure; Marshallian inefficiency of share tenancy; monitoring costs; peasants; reputation; risk sharing; scale economies; sharecropping; tenancy markets

Article

Defining who ‘peasants’ are is not an easy task for social scientists interested in rural economies and their transformation over time. The traditional image of peasants may be small-scale, full-time farmers, who have some access to land, depend largely on family labour, and produce food primarily for home consumption. The main characteristics of peasants, however, have changed over time in the process of economic development that has accompanied the penetration of markets into rural areas. Many of them are part-time farmers engaged in both farming and non-farm jobs, and produce cash crops in addition to food crops. Yet family farms continue to dominate throughout the world, contrary to the traditional view that they are remnants of feudal society and are bound to disappear as modernization proceeds (Hayami,

1996). Therefore, in order for the concept of peasants to be relevant to the present world, it seems sensible to define peasants simply as small-scale, family based farmers.

Thus, we exclude agricultural labourers dependent on wage employment, and ‘capitalist’ farmers, large landlords, and plantation owners who operate large farms using hired labour. The major categories of peasants are owner-cultivators and tenants, and tenants can be further classified into leaseholders (or fixed-rent tenants) and share tenants (or sharecroppers). Commonly these peasants are managers of farms, engaged in multiple farm tasks such as land preparation, fertilizer application, and the supervision of wage workers hired for simple tasks such as weeding and harvesting. Tenants are subject to terms of contracts, which are often unwritten and implicit such as the careful maintenance of irrigation facilities and diligent work on assigned tasks. Being small-scale, efficient farm production for food security is a major concern in a peasant society.

We also regard small cultivators in customary land tenure areas as peasants, even though they are neither owner-cultivators nor tenants. These cultivators have the use right on land as long as they continue to cultivate it, but typically they do not possess ownership rights. Thus, for example, once land is put into fallow, the cultivator tends to lose the use right. The future use of land, as well as the inheritance of land use rights, is determined by the leader of the extended family or the village chief. Such insecurity of tenure arising from the uncertain access to land in future may reduce incentives to invest in the long-term improvement of land, because those who invest may not be able to reap the benefits in the future. Although it may appear that the same argument applies to tenancy contracts, so long as the landowner has the right to terminate the contract, it is landowners, but not tenants, who make long-term investment decisions in the case of tenancy. Thus, whether the tenure insecurity results in underinvestment in land improvement is a major empirical question particularly relevant to customary land tenure areas (Besley, 1995). A critical question in the study of customary tenure institutions is whether efforts to invest in land – for example, tree planting and terracing – confer strong individualized land rights *ex post*, so as to provide proper incentives to invest *ex ante* (Otsuka and Place, 2001).

Peasant farms are small because scale economies are absent under the prevailing labour-intensive farming systems, which are characterized neither by indivisibility caused by large-scale mechanization nor by the specialization and division of labour among farm workers. Thus, the optimum farm size is likely to be small. In Asia, the average size of rice-growing farm households seldom exceeds two hectares, and it can be as low as 0.5 hectares in Java, Bangladesh, and China (David and Otsuka, 1994). Extremely large farms, including haciendas, plantations, and estates, were created by force by colonial governments, not by market forces. Once they are created, however, they tend to persist, even though their sizes exceed the optimum, primarily because the land sales market does not function due to imperfect credit markets (Binswanger and Rosenzweig, 1986). The inverse correlation between farm size and productivity, often measured by yield per hectare, is widely observed in South Asia, which indicates the existence of scale diseconomies (Otsuka, 2007). Such scale diseconomies are likely to arise from the difficulty faced by large farmers in supervising hired workers in spatially wide and ecologically diverse farm production environments.

It is widely believed that peasants are poor but efficient in resource allocation – the ‘efficient but poor’ hypothesis of Schultz (1964), which argues that investments in human capital and the dissemination of new technologies are the keys to improving their livelihood. A major challenge to the Schultz thesis is the so-called Marshallian inefficiency of share tenancy. According to this theory, a share tenant does not

work as hard as an owner-cultivator or a leasehold tenant, because he receives only a fraction of the value of the marginal product of labour. Inexplicably, output sharing rate under share tenancy is fifty–fifty not only historically in France, Italy and the antebellum South in 19th century United States, but also in many contemporary developing countries (Hayami and Otsuka, 1993). If sharing rate is not fifty–fifty it is two-thirds for the tenant and one-third for the landlord almost without exception. It is argued by the advocates of the Marshallian thesis that output-sharing is like the imposition of proportional income tax on a tenant, which discourages him from working hard. For this reason, share tenancy is prohibited by land reform laws in a number of countries in Asia.

It must be pointed out that Marshall himself (1890) did not necessarily support the Marshallian thesis: he pointed out the major shortcomings of this argument in a footnote, which was later elaborated upon by Johnson (1951) and Cheung (1969). The main point is that both the landlord and the tenant can be made better off by adopting a fixed-rent contract, which does not distort work incentives as the tenant receives the entire marginal product of labour, and then by sharing the larger ‘pie’ between the two parties. Marshall argued that, if the work effort of the share tenant can be monitored costlessly by the landlord, the share tenant will be forced to work as hard as a fixed-rent tenant. The implication is that share tenants tend to shirk unless they are effectively monitored or provided extra incentives to work harder. It is also generally agreed that, despite such problems, share tenancy is prevalent because of the risk sharing advantage; the production risk is shared between share tenants and landlords, unlike with fixed-rent contracts in which all the risk is shouldered by the tenants. This argument is plausible considering the absence of insurance markets and the existence of substantial production risk in poor agrarian communities.

Because of the existence of monitoring costs, share tenancy is inefficient in the literal sense of the word. One may argue, however, that since we cannot avoid monitoring costs in the real world, it is misleading to argue that share tenancy is inefficient; it is ‘second-best’ efficient, even if a tenant shirks. Unlike other areas of contract studies, there have been a huge number of empirical studies comparing yields per hectare between share tenancy and owner-cultivation or fixed-rent tenancy. According to a summary of earlier empirical studies by Hayami and Otsuka (1993) and a number of subsequent empirical studies, the difference in yield is found to be generally insignificant, suggesting that resource allocation under share tenancy is not significantly different from the ‘first-best’ efficiency. It is true that the differences in land and labour qualities are not properly controlled for in some studies, so that their yield comparisons are not as rigorous as they ought to be. However, since there is no reason to believe that share tenants are endowed with greater human capital and cultivate higher-quality land, the empirical evidence can be taken to imply that share tenancy is efficient, or at least not as inefficient as the Marshallian thesis assumes.

Hayami and Otsuka (1993) argue that significant shirking by share tenants is prevented by multifaceted, enduring personal relationships between tenants and landlords and the community mechanism of contract enforcement. More often than not, the landlord selects the share tenant, who is deeply related by kinship or community ties. Therefore, if the dishonest behaviours of a tenant are detected, he will be penalized not only by the termination of the share contract but also by the discontinuation of multifaceted personal relationships. Furthermore, he will not be able to find other landlords in the same community who are willing to offer new share contracts because of the loss of reputation as an honest and hard-working tenant. In this way shirking is prevented, which supports the Schultz thesis.

The above argument implies that if the share contract is deemed to be short term – for example, one

season or one year – the share tenant is likely to shirk because, regardless of whether he shirks or not, the contract will be terminated at the end of the cropping season. Indeed, according to Hayami and Otsuka (1993), significantly lower crop yields under share tenancy are typically found in India, where large landlords rotate share tenants season after season in order to avoid the implementation of the 'land-to-the-tiller program', which attempts to transfer land to the tenant. Since the presumption of the land-to-the-tiller program is that there is a single tenant on each piece of land, its implementation becomes difficult if there are many tenants.

If share tenancy is not significantly inefficient, we should not observe the inverse correlation between farm size and productivity, because larger and less productive farmers can gain by renting out a part of their lands to smaller and more productive share tenants. Indeed, in general, the inverse correlation is seldom found in South-east Asia, where tenancy markets are generally active, whereas it is often observed in South Asia where tenancy markets tend to be suppressed or discouraged by land reform laws (Otsuka, 2007).

In theory, it is considered that a fixed-rent contract will be chosen only if the tenant is risk neutral, because it provides proper work incentives to tenants who are willing to assume production risks. It is, however, highly unlikely that tenants, who are often landless and poor, do not care about the production and income risks. Although rigorous analysis is required, casual observation as well as a brief literature survey suggest that fixed-rent tenancy is more common than share tenancy in sub-Saharan Africa, unlike Asia where share tenancy is dominant. Since African farmers are poorer, it seems unreasonable to assume the risk neutrality of fixed-rent tenant farmers in sub-Saharan Africa. Indeed, they grow multiple crops presumably to diversify the production risks. One missing factor that possibly affects the contract choice is the cost of metering output under share tenancy. Since a share tenant has an incentive to under-report the amount of output to increase his share of income, the landlord must be able to meter the output effectively in order to prevent the tenant's cheating. In the case of rice farming in Asia, either the landlord himself watches the harvesting or he sends some dependable person, like his son, to the field on the designated harvesting days. The cost of watching the harvest will be high for absentee landlords and widows who have no farming experience. This is one of the reasons why such landowners usually offer fixed-rent contracts.

The importance of the cost of metering output in contract choice is illustrated by the case of the share contract of cocoa farming in Ghana, whose harvesting season lasts for more than a few months; instead of sharing output, the tenant and the landlord share the ownership of the land after the tenant finishes planting the cocoa trees (Otsuka and Place, 2001). One plausible hypothesis is that a precondition for share tenancy to be adopted is a short harvesting season, so that the cost of measuring output for the landlord is reasonably low. This hypothesis may explain why share tenancy is common in Asia, where rice and wheat are the major crops, whereas fixed-rent contracts are common in sub-Saharan Africa where maize, cassava, and other food crops which are harvested for prolonged periods are the major crops. Interestingly enough, in Ethiopia, where wheat, barley, and teff (a uniquely grain grown in this country alone) are the major crops, share tenancy is common (Benin et al., 2005). It is, however, fair to say that how crop choice and contract choice are related is an important empirical question to be investigated further.

As the population pressure on limited land resources increases in developing countries, land becomes scarce and tenancy becomes important in achieving the efficient allocation of land among farm

households by transferring cultivation rights from land-rich to land-poor households. Also becoming increasingly important are non-farm jobs, as small-scale farming subject to seasonality cannot ensure a decent living standard. Thus, the development of the rural non-farm economy is the norm rather than the exception throughout developing countries (Haggblade, Hazell and Reardon, 2006), in which members of small-scale family farms are engaged not only in food production but also the production of cash crops and, more importantly, in non-farm jobs (Bliss and Stern, 1982; David and Otsuka, 1994; Hayami and Kikuchi, 2000; Quisumbing, Estudillo and Otsuka, 2004). Thus, the conventional characterization of peasants as self-sufficient food production units as envisaged by Chayanov (1966) is no longer valid.

This does not imply, however, that markets work competitively in rural economies, so that rural households allocate labour time among food production, cash crop production, and non-farm activities, purchase all factors of production freely so as to maximize profits, and purchase goods and services so as to maximize utility. This separability of production and consumption decisions does not hold if markets are imperfect (Singh, Squire and Strauss, 1986). Therefore, in a Chayanovian world, production and consumption decisions must be made simultaneously.

Although many commodities and factors of production can be purchased and sold at competitive markets, there are also serious market failures. First of all, insurance markets fail to develop, as bad harvests negatively affect the income of all farmers in the locality (Binswanger and Rosenzweig, 1986). Second, credit markets tend to be imperfect primarily because of the lack of collateral, except for owner-cultivators who can use land as collateral. Third, labour markets, in general, do not function efficiently because of the difficulty in labour supervision. Thus, the labour market is typically thin or hired labour is employed only for such simple tasks as weeding and harvesting, activities which can be monitored easily (Hayami and Otsuka, 1993). If hired labour is employed for tasks which require care and judgment, such as water management, land preparation, and fertilizer application, the farm operation becomes inefficient. This is likely to be the main reason for the inverse correlation between farm size and productivity, in view of the fact that the suppression of land tenancy transactions forces large farmers to employ seasonal labour, often called ‘permanent’ labour, for tasks requiring care and judgement in South Asia.

If the labour market fails, the response of peasants to new marketing opportunities and new technologies can become sluggish or even perverse (de Janvry, Fafchamps and Sadoulet, 1991). For example, when the price of cash crops increases, their supply may not increase much, because farmers must depend solely on family labour without employing additional hired labour, which ought to be available at constant wage rates in the presence of a competitive labour market. Similarly, technological change in the food sector may not lead to large increases in the market supply of food if labour markets fail and food markets do not function effectively.

In view of the increasing involvement of peasants in market transactions, it is critically important to strengthen the efficiency of marketing sectors through investing in roads, communication facilities, and marketing information, such as the establishment of quality standards for farm products, in order to improve their well-being. According to Hayami (1996), peasant entrepreneurs significantly contributed to the development of rural commerce and industries in the process of economic development in East Asia.

It must be emphasized that, given the difficulty in labour supervision, we can hardly expect the farm labour markets to function efficiently. In all likelihood, it is more realistic to promote efficient tenancy

transactions, be it share or fixed-rent tenancy, if we hope to develop peasant sectors in the rapidly globalizing world where markets penetrate increasingly into rural areas. Efficient tenancy markets will increase the responsiveness of peasant sectors to new market and technological opportunities by facilitating the reallocation of land from households endowed with meagre family labour relative to land to those with an abundant supply of family labour.

The importance of tenancy transactions will continue to increase as an economy develops further. An efficient farm size expands with an increase in the wage rate, which makes it profitable to introduce large-scale mechanization to save labour. The traditional peasant mode of labour-intensive production on small farms, therefore, will no longer be sustainable in high-wage economies. Because of the scale economies associated with large-scale mechanization, viable farmers accumulate large cultivation areas through land tenancy. The practical question is at what farm size we can legitimately claim that farmers are no longer peasants. Although a clear and unanimously acceptable answer can hardly be given, I would like to propose that the issue of peasants ceases to be relevant when the issue of food insecurity associated with small farm size is resolved through farm size expansion, as well as the development of efficient marketing systems and technological changes in food production.

See Also

- access to land and development
- land markets
- peasant economy
- sharecropping

Bibliography

Benin, S., Ahmed, M., Pender, J. and Ehui, S. 2005. Development of land rental markets and agricultural productivity growth. *Journal of African Economies* 14, 21–54.

Besley, T. 1995. Property rights and investment incentives. *Journal of Political Economy* 103, 913–37.

Binswanger, H.P. and Rosenzweig, M.R. 1986. Behavioral and material determinants of production relations in agriculture. *Journal of Development Studies* 22, 503–39.

Bliss, C.J. and Stern, N.H. 1982. *Palanpur: The Economy of an Indian Village*. Oxford: Oxford University Press.

Chayanov, A.V. 1966. *The Theory of Peasant Economy*. Homewood, IL: Irwin.

Cheung, S.N.S. 1969. *The Theory of Share Tenancy*. Chicago: University of Chicago Press.

David, C.C. and Otsuka, K. 1994. *Modern Rice Technology and Income Distribution in Asia*. Boulder,

CO: Lynne Rienner.

de Janvry, A., Fafchamps, M. and Sadoulet, E. 1991. Peasant household behaviour with missing markets: some paradoxes explained. *Economic Journal* 101, 1400–17.

Haggblade, S., Hazell, P.B.R. and Reardon, T., eds. 2006. *Transforming the Rural Nonfarm Economy*. Wallingford: CAB International.

Hayami, Y. 1996. The peasant in economic modernization. *American Journal of Agricultural Economics* 78, 1157–67.

Hayami, Y. and Kikuchi, M. 2000. *A Rice Village Saga: Three Decades of Green Revolution in the Philippines*. London: Macmillan.

Hayami, Y. and Otsuka, K. 1993. *The Economics of Contract Choice: An Agrarian Perspective*. Oxford: Clarendon Press.

Johnson, D.G. 1951. Resource allocation under share contracts. *Journal of Political Economy* 58, 111–23.

Marshall, A. 1890. *Principles of Economics*. 8th edn. London: Macmillan Press.

Otsuka, K. 2007. Efficiency and equity effects of land markets. In *Handbook of Agricultural Economics*, ed. R.E. Evenson, P.L. Pingali and T.P. Schultz. Amsterdam: North-Holland.

Otsuka, K. and Place, F., eds. 2001. *Land Tenure and Natural Resource Management: A Comparative Study of Agrarian Communities in Asia and Africa*. Baltimore: Johns Hopkins University Press.

Quisumbing, A.R., Estudillo, J.P. and Otsuka, K. 2004. *Land and Schooling: Transferring Wealth across Generation*. Baltimore: Johns Hopkins University Press.

Schultz, T.W. 1964. *Transforming Traditional Agriculture*. New Heaven: Yale University Press.

Singh, I., Squire, L. and Strauss, J., eds. 1986. *Agricultural Household Models*. Baltimore: Johns Hopkins University Press.

How to cite this article

Otsuka, Keijiro. "peasants." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/

article?id=pde2008_P000048> doi:10.1057/9780230226203.1262

The New Palgrave Dictionary of Economics Online

pecuniary versus non-pecuniary penalties

Dan Kahan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Pecuniary’ penalties (fines) seem underutilized relative to ‘non-pecuniary’ penalties such as imprisonment, since they are ceteris paribus cheaper for society to impose. But the public preference for imprisonment over fines might reflect the value that the public attaches to the condemnatory meaning that imprisonment, unlike fines, conveys. An economic theory of punishment should include this sensibility in the social welfare calculus used to appraise the efficiency of various forms of punishment. The expressive utility of imprisonment might more than offset the higher cost of imprisoning offenders who could just as effectively be deterred by fines.

Keywords

Becker, G.; Bentham, J.; fines; imprisonment; pecuniary vs. non-pecuniary penalties; Posner, R.; punishment; social norms

Article

The topic of ‘pecuniary’ and ‘non-pecuniary’ penalties involves a distinction easily grasped but also a puzzle not easily solved. The distinction is between fines and all other types of criminal punishments, most conspicuously imprisonment. The puzzle arises from the seeming underutilization of pecuniary penalties, especially relative to imprisonment, in the American criminal justice system.

An economic theory of law furnishes a straightforward case for the use of pecuniary penalties. From an economic point of view, it is assumed that an individual will refrain from criminality when the expected cost of lawbreaking exceeds the expected gains (Bentham, 1843). The law can raise the expected cost by divesting offenders of either their liberty or their monetary assets. Depriving them of the latter, however, is much cheaper for society: whereas imprisonment demands an immense expenditure of resources, fining involves a transfer of wealth from offenders to the state. Accordingly, whenever a fine of a particular size and a prison term of a particular length would impose equivalent disutility on offenders,

the state, in the interest of efficiency, should select the fine (Becker, 1968).

But as cogent as it might be, this economic defence of pecuniary penalties has had strikingly little influence on the law. Non-violent offenders, who could be fined rather than imprisoned consistent with public safety, make up over half the American prison population. Many of these non-violent offenders are likely to be poor and thus effectively immune to the threat of massive fines (Shavell, 1985).

However, the possibility of ‘day fines’ – a procedure, common in Europe, whereby fines are meted out over time based on ability to pay – would make stiff penalties feasible even for offenders of relatively modest means. Based on these considerations, there is widespread expert consensus that American jurisdictions rely far too heavily on imprisonment relative to pecuniary penalties, particularly for white-collar offenders, who are obviously the least violent and the most credibly threatened with large fines (Morris and Tonry, 1990; Posner, 1980).

Confronted with this tension between theory and practice, one might be tempted to shrug one's shoulders at the seeming economic irrationality of the law and move on. But before doing so, it is worth considering whether the relative underutilization of pecuniary penalties might itself be explained in economic terms – ones that the conventional defence of pecuniary penalties overlooks.

Perhaps surprisingly, the key to a more complete economic analysis is rooted in a distinction that sociologists and philosophers draw based on the *social meanings* that legal impositions convey. ‘Prices’, on this account, refer to pecuniary exactions that connote an intention to levy a tax on an activity that society views as morally permissible; ‘sanctions’, in contrast, connote punishments that the state imposes on activities that are morally forbidden (Cooter, 1984). Criminal fines, particularly for offences that seem to involve a serious flouting of societal norms, often strike members of the public, dissonantly, as mere ‘prices’. Imprisonment, in contrast, unambiguously registers as a ‘sanction’; by virtue of the veneration of individual liberty in American society, taking a person's liberty away conveys a highly condemnatory intent on the part of the law (Kahan, 1996).

Is there any reason, economically speaking, to prefer sanctions to prices? Perhaps. Again, from an economic point of view, a person will refrain from criminality when the expected cost exceeds the expected gain. If that is correct, then the law can discourage criminality not just by increasing an offender's estimation of the costs but also by diminishing his or her valuation of the gains associated with lawbreaking. It is often argued, in fact, that the law plays a vital role in inculcating preferences that conduce to law-abiding behaviour (Andenaes, 1966; Dau-Schmidt, 1990).

On this account, one economic defence of imprisonment in preference to fines would be that sanctions are more effective than mere prices in instilling law-abiding preferences. Imposing a sanction, such as imprisonment, on an act would impart information – that the act is morally frowned upon – whereas a mere price, such as a fine, would not. On the assumption that individuals adapt their values to those expressed in law, the threat of imprisonment would in these circumstances more effectively suppress a potential offender's estimation of the gain associated with a particular criminal act than would the threat of a fine. If this characteristic of imprisonment is sufficiently pronounced, it might result in behavioural effects that more than compensate for the additional cost of imprisonment (Kahan, 1997).

But such an argument is speculative. There is some empirical evidence that the perceived justice of legal outcomes and procedures influences persons’ disposition to obey (Tyler, 1990; Nadler, 2005), but none to show that the form of punishment (abstracted from its severity) does.

In addition, the claim that the law prefers imprisonment to fines because of its superior preference-

shaping effect has a ‘just so’ quality. Aside from the implicit and by now largely rejected assumption that the law tends naturally toward efficiency, the preference-shaping defence offers no explanation of how this supposed feature of imprisonment figures in the political economy of punishment selection. Accordingly, this argument does not offer a particularly satisfying solution to the puzzle of why American jurisdictions so decidedly favour pecuniary over non-pecuniary penalties.

The real contribution the ‘price-sanction’ distinction makes to solving this puzzle consists in its power to illuminate an otherwise obscure element of the public demand for punishment. Criminal punishments, that distinction reminds us, do more than protect society from harm; they also evince a societal attitude towards criminal wrongdoers. The preference for imprisonment over fines, then, might reflect the immense value that the public attaches to the condemnatory meaning that sanctions, relative to mere prices, express.

This hypothesis finds ample empirical support. Some of it is experimental: even when fines are perceived as imposing levels of disutility comparable to particular terms of incarceration, members of the public reject fines as lacking the power to express moral condemnation (Marinos, 1997). Analysis of the reasoning of legislators, judges, and ordinary citizens confirms that it is this sensibility that causes legal decision-makers to resist substituting fines for imprisonment as a punishment for white-collar offences and for other serious but non-violent common crimes (Kahan, 1996).

The public demand for expressively satisfying punishments arguably helps to acquit imprisonment of the charge that it is less efficient than fines. Members of the public clearly value criminal punishment not only as a device for protecting them from harm but also as a ceremonial gesture for proclaiming the deviant status of those who violate societal norms (Garfinkel, 1956). There is no reason, economically speaking, to exclude this sensibility from the social welfare calculus used to appraise the efficiency of various forms of punishment. If the value that members of society obtain from the expressive utility of imprisonment is sufficiently high, that species of well-being might more than offset the higher cost of imprisoning offenders who could just as effectively be deterred by fines (Kahan, 1998).

Even more importantly, the contribution that the ‘prices-sanctions’ distinction makes to solving the puzzle of non-pecuniary penalties suggests insights into how, from an economic perspective, the law might be profitably reformed. Once the full dimensions of the social welfare function of punishment is discerned, it becomes clear that making law more efficient requires identifying relatively cheap punishments that, unlike fines, are comparable to imprisonment in *both* their expressive *and* their deterrent value. Because the expressive inadequacy of fines is also what constrains their political acceptability, expressively adequate alternatives to imprisonment also stand a much better chance than do fines of being adopted in the political process. These arguments have been used to defend the advent of *shaming* punishments – another non-pecuniary penalty – for white-collar criminals and other common offenders (Kahan and Posner, 1999).

The desirability of shaming or any other non-pecuniary penalty is obviously open to debate, economically and otherwise. What should not be, however, is the proposition that the perfection of economic analyses of law depends on their cognizance of the full range of societal benefits, including expressive ones, that the law secures.

See Also

- Becker, Gary S.
- deterrence (theory), economics of
- law, economic analysis of

Bibliography

Andenaes, J. 1966. General preventive effects of punishment. *University of Pennsylvania Law Review* 114, 949–83.

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Bentham, J. 1843. Principles of penal law. In *The Works of Jeremy Bentham*, vol. 1, ed. J. Bowring. Edinburgh, London: W. Tait; Simpkin, Marshall & Co.

Cooter, R. 1984. Prices and sanctions. *Columbia Law Review* 84, 1523–60.

Dau-Schmidt, K. 1990. An economic analysis of the criminal law as a preference shaping policy. *Duke Law Journal* 1990, 1–38.

Garfinkel, H. 1956. Conditions of successful degradation ceremonies. *American Journal of Sociology* 61, 420–4.

Gibbs, J. 1978. Preventive effects of capital punishment other than deterrence. *Criminal Law Bulletin* 14, 34–50.

Kahan, D. 1996. What do alternative sanctions mean? *University of Chicago Law Review* 63, 591–653.

Kahan, D. 1997. Social influence, social meaning, and deterrence. *Virginia Law Review* 83, 349–95.

Kahan, D. 1998. Social meaning and the economic analysis of crime. *Journal of Legal Studies* 27, 609–22.

Kahan, D. and Posner, E. 1999. Shaming white-collar criminals: a proposal for reform of the federal sentencing guidelines. *Journal of Law and Economics* 42, 365–91.

Marinos, V. 1997. Equivalency and interchangeability: the unexamined complexities of reforming the fine. *Canadian Journal of Criminology–Revue Canadienne De Criminologie* 39, 27–50.

Morris, N. and Tonry, M. 1990. *Between Prison and Probation: Intermediate Punishments in a Rational Sentencing System*. New York: Oxford University Press.

Nadler, J. 2005. Flouting the law. *Texas Law Review* 83, 1399–442.

Posner, R. 1980. Optimal sentences for white-collar criminals. *American Criminal Law Review* 17, 409–18.

Shavell, S. 1985. Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* 85, 1232–62.

Tyler, T. 1990. *Why People Obey the Law*. New Haven: Yale University Press.

How to cite this article

Kahan, Dan. "pecuniary versus non-pecuniary penalties." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000319> doi:10.1057/9780230226203.1264

The New Palgrave Dictionary of Economics Online

Pennington, James (1777 – 1862)

Murray Milgate
From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Bank of England; currency principle; Fullarton, John; Pennington, James; Tooke, Thomas

Article

Pennington may be credited with having been among the first to produce a concise statement of the so-called currency principle which formed the basis of the thinking behind the Bank Charter Act of 1844. Pennington's proposal appeared in the form of a privately printed *Memorandum* issued in 1827. This tract actually contained two memoranda, separated by a reply to the first (of 1826) from Huskisson. Much of the material from the memoranda was subsequently reissued by Pennington himself in 1840 as part of his larger *Letter to Kirkman Finlay, Esq., on the Importation of Foreign Corn*. It seems likely that the first memorandum was written at the suggestion of Thomas Tooke.

Pennington's argument was that by bringing under the direct control of the Bank of England the entire note issue, and by restricting that issue to the amount of specie reserves of the central bank so that, as Pennington put it, ‘in all cases paper would contract and expand according to the increase or diminution of its bullion’, monetary stability would be ensured. The similarity between this proposition and the practices which were emerging in the Bank of England itself at roughly the same time is worth noting. The so-called ‘Palmer rule’ differed from Pennington's only in as much as that under its operation the monetary magnitude that was to be tied to the Bank's specie reserves included not only notes and coin but also deposits. However, unlike the Palmer rule, Pennington's proposal entailed control by the central bank over the independent note-issuing activities of the country banks.

Though Pennington's proposal did not gain much public notoriety at the time, by the early 1830s Pennington had become an occasional adviser to ministers of state and to government departments. By 1844, it would seem that Pennington was sufficiently close to the government to have been asked to assist in drafting the technical details of the Bank Charter Act. The evidence currently available, however, suggests that this assistance was requested after Peel had decided upon the main provisions of that Act.

It must be remembered that although an advocate of the currency principle, Pennington actually opposed

the division of the Bank of England into separate Issue and Banking departments (as was done under the provisions of the Act of 1844). A note to this effect written by Pennington was appended by Tooke to the first volume of his *History of Prices*. The curious fact that the most famous opponent of the currency principle should have appended to his celebrated study a note by the originator of that principle, was used to humorous effect by some of Tooke's adversaries (in particular, Torrens) in subsequent controversy over the consistency of Tooke's arguments. There is a comment on this aspect of the debate by Fullarton in his *Regulation of Currencies* (1844; 2nd edn 1845, p. 191).

Pennington was born at Kendal on 23 February 1777, and died at Clapham Common on 23 March 1862. There is an admirable and thorough survey of Pennington's life and work written by R.S. Sayers to accompany his edition of Pennington's economic writings, in which can be found all of the tracts referred to above. It may be of anecdotal interest to record that Hayek has conjectured that Pennington's brother may have been the apothecary who attended Henry Thornton during his final illness in 1815 (1939, p. 33n).

Selected works

1829. Deposits with bankers. In T. Tooke, *A Letter to Lord Grenville … on the Value of the Currency*, London: J. Murray, 1829.

1838–57. Appendix C. In T. Tooke, *A History of Prices*, Vol. 2, London: Longman, Orme, Brown, Green & Longmans.

1840. *A Letter to Kirkman Finlay, esq., on the Importation of Foreign Corn … .* London: Simpkin, Marshall & Co.

1848. *The Currency of the British Colonies*. London: W. Clowes & Sons.

1963. *The Economic Writings of James Pennington*, ed. R.S. Sayers, London: London School of Economics and Political Science.

Bibliography

Fullarton, J. 1844. *On the Regulation of Currencies*. London: John Murray.

Hayek, F.A. 1939. Introduction to Henry Thornton, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802). London: George Allen & Unwin. Reprinted New York: Kelley, 1962.

Tooke, T. 1840. *A History of Prices and of the State of the Circulation in 1838 and 1839, with remarks on the corn laws and some of the alterations in our banking system*. London: Longman, Orme, Brown, Green & Longmans.

How to cite this article

Milgate, Murray. "Pennington, James (1777–1862) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000027> doi:10.1057/9780230226203.1886

The New Palgrave Dictionary of Economics Online

Penrose, Edith Tilton (1914 – 1996)

R.L. Marris

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

firm, growth of the; firm, theory of the; managerial capitalism; Penrose effect; Penrose, E. T

Article

Edith Penrose had a distinguished career in economics teaching, research and administration in the USA and the UK. From Johns Hopkins University she went to London University, where she became a professor; later she was Associate Dean of Research and Development at INSEAD in France. In administration she also rendered valuable service to the UK as chairman of the economics committee of the British Social Sciences Research Council. She retired in the mid-1980s.

In research, in the final part of her career, she concentrated on the oil industry and on multinational companies generally. Her place in the history of economic thought, however, lies in a single book *The Theory of the Growth of the Firm*, published in 1959. The review in *The Economic Journal* (1961) predicted that the book would prove one of the most influential books of the decade: this proved an understatement.

In Edith Penrose's conception, a firm is an administrative organization representing a collection of human and material resources for the purpose of producing goods and services for sale on the markets. It is essentially directed and controlled by its managers who will for various reasons be strongly motivated towards growth. The firm is not confined to any one product or market, but may diversify as its managers think fit. Profits, as seen by Penrose in her original book, were essentially a means to that end, a necessary condition for expansion.

There were, however, important administrative restraints on the rate of growth. Human resources required for *the management of change* (*growth*) were firm-specific and therefore, at any one moment, internally scarce. Expansion, however, included recruitment of additional high level human resources, that is, recruitment of additional growth-creating capacity. Therefore, subject to the dynamic constraint, there need be no ultimate limit on size. More generally, the relationship can be stated as a proposition that the *level* of current efficiency will, beyond a point, diminish with the rate of change of size: fast growth has a price.

The book was also rich in many associated and diverse ideas which cannot be set out in detail. The administrative ‘Penrose effect’ has been generally accepted and incorporated into a variety of micro-and macroeconomics, especially in the field known as ‘the Corporate Economy’. The idea was most especially used in Robin Marris in *The Economic Theory of Managerial Capitalism* (1964) and by Hirofumi Uzawa in a significant contribution to macroeconomics a few years later (Uzawa, 1969).

The total effect of Edith Penrose's work was that of destruction of the neoclassical model of the firm, followed by reconstruction. In the following years, however, despite the wide recognition the work received, classroom microeconomic theory, and also classroom industrial organization, often seemed to continue as if nothing had happened.

Selected works

1959. *The Theory of the Growth of the Firm*. Oxford: Basil Blackwell.

Bibliography

Marris, R.L. 1964. *The Economic Theory of Managerial Capitalism*. London: Macmillan; New York: Free Press.

Uzawa, H. 1969. Time preference and the Penrose effect in a two-class model of economic growth. *Journal of Political Economy* 77, 628–52.

How to cite this article

Marris, R. L. "Penrose, Edith Tilton (1914–1996)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000052> doi:10.1057/9780230226203.1265

The New Palgrave Dictionary of Economics Online

pensions

Zvi Bodie
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Pensions are benefit contracts that replace a person's earnings after she reaches old age and retires from the labour force. Pension systems vary widely across countries, but everywhere the government's role is to provide a minimum through a mix of cash and medical benefits. Governments often provide tax incentives for employers and unions to sponsor occupational pension plans that complement the government-run system. The nature of the pension benefits promised and the assets that back them have profound effects on social welfare, on the development of a country's domestic asset markets, and on the global financial system.

Keywords

adverse selection; annuities; defined-benefit pension plans; defined-contribution pension plans; equipment trusts; free-rider problem; funding of pensions; informational inefficiencies; life-cycle saving model; longevity insurance; moral hazard; occupational pensions; pay-as-you-go pension plans; pensions; retirement; risk sharing; Social Security (USA)

Article

Pensions are retirement income contracts, and their manifest function is to replace a person's earnings after she reaches old age and retires from the labour force. Prior to the Industrial Revolution, the extended family was the primary institution that performed this function. Elderly family members lived and worked with offspring on a family-owned farm, and all drew a common livelihood from it. In many of today's less developed countries, this family-based pattern for old-age support still holds true.

Over time, urbanization and other fundamental economic and social changes gave rise to new institutional structures for the care and support of the elderly in much of the industrialized world. An often-used metaphor for describing developed countries' pension systems is that of the 'three-legged stool'. The first leg consists of government-provided old-age assistance and insurance programmes; the second leg is comprised of employer or labour union-provided pensions; and the third is individual and family support. There is substantial variation in the mix of the three sources of retirement income, both across households in a given country and across different countries (Bodie and Davis, 2000).

Pensions should be analysed in the context of a life-cycle model of saving. In this framework, people save during their working years so that they can consume in their non-working retirement period. Some simplifying assumptions can quickly convey the essence of the life-cycle approach. Assume for the sake of illustration that an individual enters the labour force at age 20, works until retiring at age 65, and dies at age 80. His initial wealth is zero. During the working years, he earns constant real labour earnings, a portion of which is saved for retirement. The saving includes personal saving and the accrual of benefits under social security and employer-sponsored pension plans. We assume that the individual chooses to save an amount during the working years sufficient to make his level of real consumption after retirement equal to what it was before retirement. These savings earn a zero real rate of interest. At retirement, a constant real retirement benefit is paid, and at death there is nothing left over as a bequest.

These assumptions imply that the ratio of consumption to earnings must equal the ratio of years of work to total years of work and retirement:

$$\text{Years of work} \times (\text{earnings} - \text{consumption}) = \text{years of retirement} \times \text{consumption}$$

$$\text{Years of work} \times \text{earnings} = (\text{years of work} + \text{years of retirement}) \times \text{consumption}$$

$$\frac{\text{Consumption}}{\text{Earnings}} = \frac{\text{years of work}}{\text{years of work} + \text{years of retirement}}$$

In this example, there are 45 years of work and 15 years of retirement, so the ratio of consumption to earnings is equal to 45/60 or 75 per cent, and the individual's ‘gross saving’ rate during his working years is 25 per cent. The benefits received during retirement come from three sources corresponding to the components of gross saving: social security, employer-provided pensions, and personal saving.

The government's role in providing retirement income

The government's role in providing retirement income varies considerably across countries, but despite these variations there is a common theme: in virtually every country the government provides a ‘floor’ of income protection for the elderly, with the aged population's needs met by some mix of national insurance and national welfare systems, in the form of cash and medical insurance. This floor (or ‘safety net’) is usually mandatory and cannot be transferred.

Several economic arguments justify the government's provision of a layer of retirement benefits for everyone (Merton, 1983). The first deals with *informational inefficiencies*. It is costly to acquire the knowledge necessary to prepare and carry out long-run plans for income provision. Although peoples’ lifetime financial plans depend on their individual preferences and opportunities, their goals may be similar enough that a standard retirement savings plan can prove suitable to many. By providing a basic plan that supplies at least a minimum level of old-age support, the government is likely to help people save more efficiently than they could on their own.

The second argument revolves around *adverse selection* problems, There is considerable ‘longevity risk’ that people will outlive their retirement savings because their date of death is not known with certainty, in contrast to the simplified version of the life-cycle model we described earlier. One way to insure against the risk of exhausting one's savings during retirement is to purchase a life annuity contract. But the private market for life annuities suffers from adverse selection because people with a higher-than-average life expectancy have a high demand for this kind of insurance. As a consequence, an average individual will find the equilibrium price for privately purchased life annuities too high, and will tend to self-insure against longevity risk by having an extra reserve of retirement savings. Universal and mandatory social security is one way of overcoming this adverse selection problem. Making participation in the national plan mandatory and not giving anyone a choice about the form of benefit payouts creates more complete pooling of longevity risk.

A third reason for a government-mandated universal retirement income system is to address the *free-rider* problem, which arises when the citizenry collectively feels an obligation to offer a universal ‘safety net’. If this collective commitment were well understood by all, some people would avoid saving for their own retirement, intending instead to rely on benefits provided by others when they are old. Similarly, some might take on more risk in investing their retirement savings than they would in the absence of a safety net. In such an environment, mandating universal participation simply forces people to pre-pay in the form of social security taxes for benefits they ultimately will receive from the system. Therefore, the purpose of a mandatory system is to protect society against free riders.

While these three arguments explain why governments might believe it important to mandate a minimum level of universal participation in a national retirement programme, they are silent about what the particular level of government benefits should be. These arguments are also silent on whether the government might merely mandate a plan, leaving it to the private sector to manage it. For example, in several countries the other two legs of the retirement-income stool are encouraged by government regulation as an alternative to government provision. Governments often use tax policy to provide incentives for employers and unions to sponsor pension plans that, like the government-run plan, are mandatory and non-assignable. In some of those countries, tax incentives are also given to self-employed individuals and households

(who are not otherwise covered) to create a retirement fund for themselves. Use of such funds for other purposes is discouraged by imposing penalties on early withdrawal of money from the fund.

The role of occupational pensions

Pension plans sponsored by employers or unions – also known as occupational pensions – are often integrated with the government-run plan, either explicitly or implicitly. When combined with the government-provided retirement benefit, these plans are usually designed to replace 70–100 per cent of pre-retirement earnings of lower- and middle-income employees in developed nations. Benefits are usually lower for higher-income workers, who then must rely on direct personal savings for a larger part of their retirement income.

Why are employers and/or trade unions logical sponsors of retirement plans for their employees? There are at least four good reasons (Bodie, 1990). First, they make for *efficient labour contracting*. Pension plans are an incentive device in labour contracts because they affect employee hiring and turnover patterns, work effort, and the timing of retirement.

Second, they promote *informational efficiencies*. Employment-based plan sponsors often have better access than the plan's beneficiaries to information needed for preparing long-run financial plans tailored to the needs of the employees. In particular, sponsors may have better knowledge of the probable path of future labour income for their employees. By providing a basic plan that saves enough to provide for replacement of anticipated future labour earnings, the corporate sponsor can potentially save more efficiently than each employee acting individually. In order for the sponsor to provide efficiently for future wage and salary replacement of employees, it is enough to have accurate forecasts of the earnings of the group as a whole and not the individual earnings of each member of the group. It is probably easier (although by no means simple) to forecast group earnings than it is to forecast an individual's future earnings.

Third, employment-based plans can avoid *principal–agent problems*. While plan sponsors and beneficiaries may have conflicting economic interests, in many respects their interests coincide. Employers who acquire a reputation for taking care of their employees' retirement needs may find it easier to recruit and retain higher-quality employees. If employees' trust and goodwill towards their employers develop, then motivation and labour productivity may also be enhanced. Employers therefore have some economic incentive to act in the best interests of their employees.

Other possible providers of retirement planning services may be less suitable as beneficial agents of employees. Insurance agents, stockbrokers, and others who are often engaged in providing these services to individual households may be less trustworthy than employers because they could be interested in selling individuals some product or service that those individuals might not choose were they well-informed. These other agents may be motivated to persuade individuals to save too much for retirement or to invest in inappropriate ways. Anyone who has ever tried to find competent and impartial personal financial planning or investment advice is aware of the difficulties.

Fourth, plan sponsors often have *access to capital markets* that is unavailable to their employees acting as individual savers. Employees may not be able to buy certain kinds of insurance individually, but might be able to do so as members of an employee group. In addition, sponsoring firms can take advantage of scale economies while individual employees cannot. Financial intermediaries such as insurance companies can provide a suitable vehicle for the insurance needs of employees. But often a financial intermediary will not be willing to provide enough of the insurance desired by the individual at an efficient price because of problems of adverse selection and moral hazard.

Longevity insurance is an important example of this. In principle longevity risk is diversifiable and can be largely eliminated through risk pooling and sharing. But, as explained earlier, the problem of adverse selection can make the private insurance market for life annuities inefficient. Group insurance through pension plans is often seen as a solution to this problem.

Defined benefit and defined contribution pension plans

Pension plans are usually classified in terms of what is promised to the beneficiaries. There are two basic categories: *defined contribution* and *defined benefit* plans. In a defined contribution plan, a formula specifies the amount of money that must be contributed to the plan, but does not specify benefit payouts. Contribution rules are usually a predetermined fraction of salary (for example, the employer contributes ten per cent of the employee's annual wages to the plan), although that fraction need not be constant over an employee's career. The pension fund consists of a set of individual investment accounts, one for each covered employee. Pension benefits are not specified, other than that at retirement the employee gains access to the total accumulated value of the contributions and the earnings on those contributions. These funds can be used to purchase an annuity or can be taken in the form of a lump sum.

In a defined contribution plan, the participating employee frequently has some choice over both the level of contributions and the way the account is invested. In principle,

contributions could be invested in any security, although in practice most plans limit investment choices to bond, stock, and money-market funds. The employee retirement account is, by definition, fully funded by the contributions, and the employer has no legal obligation beyond making its periodic contributions. Therefore, in a defined contribution plan much of the task of setting and achieving retirement income replacement goals falls on the employee. In some defined contribution plans, employees have the option of transferring some of the risks to an insurance company.

In a defined benefit plan, by contrast, the pension plan specifies formulae for the cash benefits to be paid after retirement. The benefit formula typically takes into account years of service for the employer and level of wages or salary (for example, the employer pays a retired worker an annuity from retirement to death, the amount of which might be equal to one per cent of his final annual earnings multiplied by years of service). Contribution amounts are not specified, and the employer (called the ‘plan sponsor’) or an insurance company hired by the sponsor guarantees the benefits and thus absorbs the investment risk. The obligation of the plan sponsor to pay the promised benefits is similar to a long-term debt liability of the employer.

In the United States, the United Kingdom, and many other countries the trend since the mid-1990s has been away from defined benefit towards the defined contribution form. The two plan types are not, however, mutually exclusive. Many sponsors have defined benefit plans as a ‘primary’ plan, in which participation is mandatory, and supplement them with voluntary defined contribution plans. Moreover, some plan designs are ‘hybrids’ combining features of both plan types. For example, in a ‘cash-balance’ plan each employee has an individual account that accumulates interest. Each year, employees are told how much they have accumulated in their account and, if they leave the firm, they can take that amount with them. If they stay until retirement age, however, they receive an annuity determined by the plan's benefit formula. A variation on this design is a ‘floor’ plan, which is a defined contribution plan with a guaranteed minimum retirement annuity determined by a defined benefit formula. These plan designs usually take into account the benefits provided by the government-run system.

Why does funding matter?

The *pension plan* is the contractual arrangement setting out the rights and obligations of all parties; the *pension fund* is a pool of assets set aside to provide collateral for the promised benefits. In defined contribution plans, the value of the benefits equals that of the assets and so the plan is always exactly fully funded. In contrast, defined benefit plans have a continuum of possibilities. There may be no assets dedicated to the pension plan in a separate fund, in which case the plan is said to be *unfunded*. When there is a separate fund but assets are worth less than the present value of the promised benefits, the plan is *underfunded*. If the plan's assets have a market value that exceeds the present value of the plan's liabilities, it is said to be *overfunded*.

Why and how does funding matter? The assets in a pension fund provide collateral for the benefits promised to the pension-plan beneficiaries. A useful analogy is that of an equipment trust. In an equipment trust, such as one set up by an airline to finance the purchase of airplanes, the planes serve as specific collateral for the associated debt obligation. The borrowing firm's legal liability, however, is not limited to the value of the collateral. By the same token, if the value of the assets serving as collateral exceeds the amount required to settle the debt obligation, any excess reverts to the borrowing firm's shareholders. So, for instance, if the market value of the equipment were to double, this would greatly increase the security of the promised payments, but it would not increase their size. The residual increase in value would accrue to the shareholders of the borrowing firm.

The relation among the shareholders of the firm sponsoring a pension plan, the pension fund, and the plan beneficiaries is similar to the relation among the shareholders of the borrowing firm in an equipment trust, the equipment serving as collateral, and the equipment-trust lenders. In both cases, the assets serving as collateral are ‘encumbered’ (that is, the firm is not free to use them for any other purpose as long as that liability remains outstanding), and the liability of the firm is not limited to the specific collateral. Any residual or ‘excess’ of assets over promised payments belongs to the shareholders of the sponsoring firm. Thus the greater the funding, the more secure the promised benefits. However, whether the plan is underfunded, fully funded, or overfunded, the size of the *promised* benefits does not change.

Why do employers fund their defined benefit plans? Reasons appear to vary across countries. First, funding offers benefit security if there is no government insurance of pension benefits, or only partial insurance. Employees may demand that the future pension promises made to them by their employer be collateralized through a pension fund. In the United Kingdom, for example, there is no government pension insurance beyond the minimum guaranteed pension of the State Earnings Related Pension Scheme (SERPS). Pension funding in this case provides an important cushion of safety for retirement income.

Second, some countries impose minimum funding standards by law. These standards seek to insure that promised pension benefits are paid even in the event of default by the corporate sponsor and also aim to protect the government (and the taxpayer) from abuse of government-supplied pension insurance. In the United States, for example, the Pension Benefit Guaranty Corporation (PBGC) must continue pension payments offered by defined benefit pension plans if their sponsoring corporations become bankrupt with an underfunded pension plan. Recent changes in United States pension law mandate that the PBGC insurance premium must depend on the plan's extent of underfunding, and have also eliminated the possibility of voluntary termination of an underfunded pension plan.

Third, there may be tax incentives for plan sponsors to fund their defined benefit plans. Black (1980) and Tepper (1981) have shown that the tax advantage to pension funding stems from the ability of the sponsor to earn the pre-tax rate of return on pension investments. It is no accident that in Germany, where employers face a tax disadvantage if they fund their pension plans, pensions are predominantly unfunded.

Finally, funding a pension plan may provide the sponsoring firm with financial ‘slack’ that can be used in case of possible financial difficulties the firm may face in the future. In the United States, pension law allows plan sponsors facing financial distress to draw upon excess pension assets by reduced funding or, in the extreme case, voluntary plan termination. The pension fund therefore effectively serves as a tax-sheltered contingency fund for the firm.

Funding of pensions in the public sector

In a strictly unfunded *pay-as-you-go* government-operated pension system, retirees’ benefits depend entirely on the stream of revenue generated by taxes levied on currently active workers. If this were exactly true, benefits would fluctuate with changes in economic fortunes, rising when tax collections rose, and falling in recessions. In practice this does not happen because most government pensions are of the defined-benefit variety and promise to deliver retirement benefits according to a specified benefit formula. Nevertheless, without funding, benefit payouts are susceptible to cuts when the public sector experiences a rising ratio of retired to active workers and/or large government deficits. In this event benefits accrued under that formula may be altered as a way of reducing this form of government debt.

As a case in point, consider the 1983 reform of the United States Social Security system. A changing demographic structure for workers led many to become concerned that the future benefits in a pure pay-as-you-go system could be dramatically reduced. Hence, a key provision of that reform was to require substantial pre-funding of future benefits. To do this, the Social Security payroll tax rate was raised and the excess of current revenues over current benefit payments was invested in government bonds held in a trust fund.

While this reform apparently funds the plan, some are less sure about the result. In a private plan, funding is used to insure against default by the plan sponsor. Under Social Security, the promise to pay benefits seemingly has the same level of full faith and credit of the government as the bonds used to fund the plan. Yet there seems to be a belief that pre-funding will ensure that when workers reach retirement they will indeed receive benefits approximating those promised under the current benefit formula (that is, the one in effect when they were active in the labour force).

A problem with this view is that there remains a potential risk associated with benefits promised under a government-run retirement income system. Even if the current government is committed to maintaining the current schedule of promised benefits, it cannot credibly fully bind future governments to do so. Indeed, it has become evident in many countries that the benefit formula and the method of financing those benefits can be and often are changed. In the United States, for example, the Congress has changed both in the past and it can surely do so again in the future. Perhaps more strikingly, public pensions in Chile were radically restructured in the early 1980s, replacing the defined benefit public social security system with a mainly private defined-contribution plan. In the 1990s Australia followed Chile's lead, and several eastern European countries have done so too.

These examples bring out an important difference between government and private-sector obligations. A private-sector plan sponsor cannot unilaterally repudiate its legal liability to make promised payments. It can default because of inability to pay, but it cannot repudiate its legal obligations without penalty. On the other hand, a government – because it has the power to legislate changes in the law – can sometimes find ways to repudiate such obligations without immediate and obvious penalty. Indeed, an integrated system in which private plan sponsors supplement government-provided pension benefits to achieve a promised ‘replacement ratio’ of pre-retirement earnings can be seen as a type of private-sector insurance against the political risks of the government-run system.

In sum, a mixed public–private system of retirement income provision is a way of reducing the risks of each separate component through diversification across providers. Public-sector pension plans can change the law to reduce promised benefit levels. Private-sector pension plan sponsors are committed by law (and perhaps reputation) to pay promised benefits, but they may default. And sometimes, as an additional linkage reinforcing the first two legs of the retirement income stool, the government may insure private pension benefits against the risk of default (Bodie and Merton, 1993).

See Also

- population ageing
- retirement

Bibliography

Black, F. 1980. The tax consequences of long-run pension policy. *Financial Analysts Journal* 36, 25–31.

Bodie, Z. 1990. Pensions as retirement income insurance. *Journal of Economic Literature* 28, 28–49.

Bodie, Z. and Davis, E., eds. 2000. *The Foundations of Pension Finance*. Cheltenham: Edward Elgar.

Bodie, Z. and Merton, R. 1993. Pension benefit guarantees in the United States: a functional analysis. In *The Future of Pensions in the United States*, ed. R. Schmitt. Philadelphia: University of Pennsylvania Press.

Merton, R. 1983. On consumption indexed public pension plans. In *Financial Aspects of U.S. Pension System*, ed. Z. Bodie and J. Shoven. Chicago: University of Chicago Press.

Tepper, I. 1981. Taxation and corporate pension policy. *Journal of Finance* 36, 1–13.

How to cite this article

Bodie, Zvi. "pensions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000054> doi:10.1057/9780230226203.1266

The New Palgrave Dictionary of Economics Online

perfect competition

M. Ali Khan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article attempts a critical appraisal of the literature on perfect competition as it has evolved since the work of Debreu–Scarf and Aumann in the 1960s, following papers of Debreu–Scarf and Aumann. It focuses on mathematical techniques that have been garnered to cope with the presuppositions of the classical theory relating to *finitude*, *convexity* and *agent-independence*.

Keywords

additivity; aggregation; Arrow–Debreu model of general equilibrium; central limit theorems; competitive equilibrium; consumptions sets; convexity; core equivalence; cores; correspondences; equilibrium; existence of general equilibrium; externalities; fixed-point theorems; game theory; Hayek, F. A.; hyperplanes; idealized limit economy; incentive compatibility; independence; infinite commodities; large games; law of large numbers; Lindahl equilibrium; Lyapunov theorem; marginal cost pricing; measure spaces; negligibility; non-convexity; perfect competition; probability; public goods; returns to scale; Shapley–Folkman theorem; Steinitz's theorem; Wald, A.

Article

An allocation of resources generated under perfect competition is an allocation of resources generated by the pursuit of individual self-interest and one which is insensitive to the actions of any single agent. Self-interest is formalized as the maximization of profits over production sets by producers and the maximization of preferences over budget sets by consumers, both sets of actions being taken at a price system which cannot be manipulated by any single agent, producer or consumer. An essential ingredient then in the concept of perfect competition, that which gives the adjective *perfect* its thrust, is the idea of *economic negligibility* and, in a set of traders with *many* equally powerful economic agents, the related notion of *numerical negligibility*. Perfect competition is thus an idealized construct akin (say) to the mechanical idealization of a frictionless system or to the geometric idealization of a straight line.

Following the lead of Wald, a mathematical formalization of perfect competition in a setting with an exogenously given finite set of commodities and of agents was developed in the early 1950s in the pioneering papers of Arrow, Debreu and McKenzie. It was shown that convexity and independence assumptions on tastes and technologies guarantee that a competitive equilibrium exists, and that a Pareto-optimal allocation can be sustained as a competitive equilibrium under appropriate redistribution of resources. It was also shown, drawing on the tacit assumption that markets are universal but by avoiding any convexity assumptions, that, with local non-satiation, every competitive allocation is Pareto-optimal. Relegating precise definitions to the sequel, we refer the reader to Koopmans (1961) for a succinct statement of the theory; Debreu (1959) and McKenzie (2002) remain its standard references, Fenchel (1951) and Rockafellar (1970) its mathematical subtexts, and Weintraub (1985), Ingrao and Israel (1987) and Mirowski (2002) its sources of historical appraisal.

However, in its exclusive focus on drawing out the implications of convexity and agent-independence for a formalization of perfect competition, the theory remained silent about environments with increasing marginal rates, in production and in consumption, as well as those where private and social costs and benefits do not coincide, to phrase this silence in Pigou's (1932) vocabulary of a preceding period. In particular, the notion of perfect competition that was fashioned by the initial theoretical development had no room for economic phenomena emphasized, for example, in the papers of Hotelling (1938), Hicks (1939) and Samuelson (1954). It took around two decades to show that, at least as far as collective consumption and public goods were concerned, the theory had within it all the resources for an elegant incorporation, but of course within the confines and limitations of its purview (see Foley, 1970 and his followers). Non-convexities in production and consumption were a different matter entirely; they required mathematical tools that went beyond convexity, and further development had to await the invention of non-smooth calculus of Clarke and his followers; see Rockafellar and Wets (1998) and Mordukhovich (2006) for a comprehensive treatment.

A robust formalization of the idea of perfect competition for non-convex technological environments in the specific form of marginal cost pricing equilibria, with the regulation of the increasing returns to scale producer(s) given an explicit emphasis, can be outlined under each of the three headings of the theory identified by Koopmans: existence and the two welfare theorems. Marginal cost pricing equilibria exist under suitable survival and loss assumptions, but are not globally Pareto-optimal even under the assumption of universality of markets. Finally, Pareto-optimal allocations can be sustained as marginal cost-pricing equilibria under appropriate redistribution of resources. Moreover, under the terminology of Lindahl–Hotelling equilibria, Khan and Vohra (1987) provide the existence of an equilibrium concept that incorporates both public goods and increasing returns to scale in one sweep. This work on perfect competition in the presence of individualized prices stemming from collective consumption and a regulated production sector (or sectors) merits an entry in its own right, and rather than a detailed listing of the references, we refer the reader to Vohra (1992) and Mordukhovich (2006, ch. 8) for details and references.

Three observations in connection with this recent, but already substantial, literature are worth making. First, in the attempts to generalize the second fundamental theorem of welfare, one can discern a linguistic turn whereby both the Arrow–Debreu emphasis on decentralization and the Hicks–Lange–Bergson–Samuelson–Allais equality of marginal rates are seen as special cases within a synthetic

treatment emphasizing the intersection of the cones formalizing marginal rates; Khan's (1988) introduction is an emphatic articulation of this point of view. Second, a canonical formulation of the notion of marginal rates, despite fits and starts, now seems within reach, though a notion that works well for the necessary conditions may not be the one equally suited for the question of existence; see Hamano (1989) Khan (1999). Finally, conceptual clarity requires an understanding of circumstances when this type of non-convex theory bears a strong imprint of its finite-dimensional, convex counterpart, as detailed in Khan (1993), as opposed to when its higher reaches require a functional-analytic direction totally different from that charted out in the pioneering papers of the 1950s; see Bonnisseau and Cornet (2006) for reference to recent work.

With price-taking assumed rather than endogenously deduced, there is no overriding reason why a formalization of perfect competition must limit itself to a setting with a finite, as opposed to an unbounded (infinite), number of (perfectly divisible) commodities. Indeed, another set of pioneering papers of Debreu, Hurwicz and Malinvaud, written in the 1950s with an eye to a theory of intertemporal allocation but over a time horizon that is not itself arbitrarily given, fixed and finite so to speak, did consider the decentralization of efficient production plans as profit-maximizing ones. But again, it was only two decades later that the work of Bewley, Peleg-Yaari, Gabszewicz and Mertens inaugurated sustained attempts to provide a general formalization of perfect competition over infinite-dimensional commodity spaces (see Khan and Yannelis, 1991). The work can again be categorized under Koopmans's three headings of the theory, but relative to its finite-dimensional counterpart, it noted that the separation of disjoint convex sets, and the use of aggregate resources to furnish a bound on the consumption sets to ensure compactness, proved to be matters of somewhat greater subtlety. In short, even a norm-compact set of an infinite-dimensional commodity space is ‘rather large’ and its cone of non-negative elements ‘rather small’. Indeed, as Negishi's method of proof attained dominance, the imbrication of the convexity assumption in a clear demarcation of fixed-point theorems for issues of existence and separating hyperplane theorems for those of decentralization, no longer obtained. The subject is surveyed in Mas-Colell and Zame (1991), but another survey is perhaps overdue as exploration of individual mathematical structures, ordered structures in particular, reveals hitherto unforeseen essentials, and increasing returns to scale and other non-classical phenomena are inevitably accommodated; see the references of Aliprantis, Cornet and Tourky (2002) and Aliprantis, Florenzano and Tourky (2006), on the one hand, and those of Shannon (1999) and Bonnisseau (2002) on the other.

However, the question persists as to what meaning can be given to the study of perfect competition in a setting with an exogenously given infinite-dimensional commodity space where markets open only once and there is no room for the correction of mistakes and unfulfilled plans. If the extension of the theory requires additional technical assumptions, how do they translate into *desiderata* that are of relevance for the formalization of the coherence of decentralized, self-interested decision-making of independent agents acting independently of each other? Even if, for example, the *uniform properness* assumption of Mas-Colell (1986) and his followers could be pinned down as a formalization of bounded marginal rates of substitution (see the notion of a Fatou cone in Araujo, Martins-da-Rocha and Monteiro, 2004, and one failed attempt in Khan and Peck, 1989), what does it say about the set-up of the model itself that lifts this up to be a limitation as fundamental as that of convexity or independence? If the underlying motivation for the extension to infinite-dimensional commodity spaces is time, risk, quality, information or location, how do these considerations manifest themselves in the infinite dimensionality of the

commodity space, in a situation that necessities (or precludes) one commodity in an economy being numerically negligible relative to the entire set? More sharply, why ought not the resulting problems be more squarely faced in simpler partial equilibrium models, rather than studied under the limitation of a construction whose primary emphasis is the viability and desirability of static interaction? We defer these issues to turn to our principal theme, namely, the formalization of the perfectness of perfect competition.

The point is that the assumption of a finite number of agents embodied in all of this work is an explicit admission of the fact that the economic non-negligibility of each agent, at least in principle, and therefore her non-manipulation of, and corresponding submission to, the price system furnishes a somewhat muted maximization of her self-interest. In terms of the emphasis on *negligibility* as a prerequisite for a rigorous formalization of perfect competition, as is being emphasized in this article, the postulated behaviour of individual agents in the so-called Arrow–Debreu–McKenzie model of perfect competition, with or without infinite commodities, externalities and increasing returns to scale, leads to the rather natural puzzlement as to what it is precisely that guarantees an agent's passive acceptance of the price system, let alone individualized pricing rules, and that too in a construction whose primary motivation is consistency and generality. In the vernacular due to Hurwicz (1972), one that has gained increasing currency since the 1980s, what is it that makes this model of the economic system *incentive-compatible*? How is its gloss of the intuitive notions of *negligibility*, *large* and *many* to be made precise?

Six conceptually separate attempts to answer this question are distinguished here; these alternative but interrelated formalizations of perfect competition draw their meaning from two early conjectures: (i) Edgeworth's (1881) conjecture on the shrinking of the *core* to its set of competitive allocations (again, precise definitions to follow), and (ii) Farrell's (1959) conjecture on the existence of competitive equilibrium in a environment that is not necessarily convex. Interpreted literally, both conjectures are clearly false for a given finite economy, but the first can be distinguished from the second in not being simply a case of dispensing with an assumption in a result whose basic contours are well-established, but rather in going beyond Koopmans's categorization of perfect competition to include a solution concept other than that of Pareto optimality. It is in the reliance of the core notion as a test for the perfectness of competition, in working with a third fundamental theorem of welfare economics, so to speak, and in giving precision to the ambiguity inherent in the term *shrinking*, that allows an entry into the formalization of the negligibility of individual agents. However, at this point, the discussion demands the rigour of notation and definitions; and since the essence of the ideas can be adequately communicated in the context of an economy without producers, that is, in an *exchange economy*, we confine ourselves to this case.

An *exchange economy* consists of a commodity space L, a set of traders T, a space of trader characteristics P defined on the commodity space, and a mapping $\mathcal{E}$ from T into P with the value of $\mathcal{E}$ at a particular t in T being given by the triple $\mathcal{E}(t) \equiv ((X(t), \succsim_t, e(t))$ specifying the characteristics of agent t in T. The space of characteristics is thus a product space constituted by consumption sets $X(t) \subseteq L$, by binary relations $\succsim_t$ over $X(t)\times X(t)$, preferences over the consumption set read ‘preferred or indifferent to’, and by initial endowments $e(t) \in X(t)$.An *allocation* x: $T\rightarrow L$ is an assignment of commodity bundles such that $x(t) \in X(t)$ for all t in T and such that the summation, suitably formalized, of $(x(t) - e(t))$ over T is zero, or, in the case of *free-disposal*, less than or equal to zero. In either case, the

fundamental economic problem facing a particular exchange economy, as discussed above and being given symbolic formulation here, is the choice of an allocation.

An allocation x: $T \to L$ is said to be in the *core* if there does not exist any other allocation y and a coalition $S \subseteq T$, suitably formalized, such that $y(t) \succsim_t x(t)$ and not $x(t) \succsim_t y(t)$ for all $t \in S$, and that the summation of $(y(t) - e(t))$ over S is zero, or again with free disposal, less than or equal to zero. A perfectly competitive allocation of resources is a price-based allocation where a *price system* is a non-zero, continuous linear function on the commodity space L. A *competitive equilibrium* is a pair (p, x) where p is a price system and x an allocation such that for all t in T, $x(t)$ is a maximal element for $\succsim_t$ in the budget set $\{y \in X(t) : (y, p) \leq (e(t), p)\}$. Here (y, p) denotes the valuation of the commodity bundle y by the function p and, in case L is the Euclidean space $\mathbb{R}^\ell$, the Riesz representation theorem allows it to be given a simple accounting interpretation of an inner product $(y, p) = \sum_{i=1}^{\ell} p_i y_i$; see Rudin (1974) for this theorem and for other unspecified terminology. For any competitive equilibrium (p, x), x is referred to as a *competitive allocation*. In terms of the earlier discussion of infinite-dimensional commodity spaces, the commodity space L has presumed on it enough mathematical structure so as to give meaning to the ordering ‘less than or equal to’, to the summation operator in the notion of an allocation and of a blocking coalition, and to linearity and continuity in the notion of a price system. Conceptually, what is of consequence here is that competitive allocations can be viewed as making precise the idea of some sort of individual rationality, and core allocations as making precise the idea of some sort of group rationality.

In Aumann's (1964) formulation of perfect competition, the set of traders is the Lebesgue unit interval, the commodity space is the Euclidean non-negative orthant $\mathbb{R}^\ell_+$, the set of admissible coalitions the Borel σ-algebra on the unit interval, and summation, Lebesgue integration. Under the assumption of Lebesgue measurability of preferences $\succsim_t$, and of Lebesgue integrability of the initial endowments $e(\cdot)$, he proved that the set of competitive allocations of such an economy coincides with its set of core allocations and, in Aumann (1966), that neither set is empty. These precise and elegant affirmations of the conjectures of Edgeworth and Farrell did not require any convexity hypotheses on preferences, and, what is perhaps of equal significance, they furnished a precise formulation of an idealized limit economy in which price-taking is rendered theoretically reputable: every agent is numerically and economically negligible in that the effect of his or her action, not only on the price system but also on the equilibrium allocation, is precisely zero. An agent has a negligible weight very much akin (say) to the probability of a particular point on a dartboard being hit by a dart.

The seminal nature of Aumann's conception was quickly realized and incorporated in to the mainstream. The metaphor of a continuum of agents is now routinely (but not incorrectly) invoked to validate the removal of idiosyncratic uncertainty by aggregation even in models of a representative agent in theoretical work in macroeconomics and other, so-called applied, fields. This work that is nothing if not an investigation of competition, perfect or otherwise. Two observations are worth making. First, whereas Aumann's assumption of a Lebesgue unit interval was only a simplifying one, and that the results hold for any arbitrary *atomless* and finite measure space, Lebesgue (rather than Riemann–Stieltjes) integration is essential to a theory based on T as the set of agent-names, and therefore free of any topological considerations; see Khan and Sun (2002, Introduction) for a detailed exposition of this point. Indeed, Shapley has even questioned the postulate of measurability, leave alone continuity, for a notion

of an allocation whose very *raison d'être* is a formalization of independent individual self-interest. Second, since the theory is based on a neglect of sets of measure zero, it is a conception of an allocation as an equivalence class of functions, rather than of functions themselves, that is identified by the theory. Put more sharply, Pareto-optimal allocations in an economy with a continuum of agents do not exist if their definition is taken verbatim from that of a finite economy, and not recast in terms of coalition of positive measure. In any case, the theory of an economy *E* conceived as a measurable map, at least in its finite-dimensional embodiment, is a testimony to the power of the Lyapunov theorem on the range of an atomless vector measure and to a powerful mathematical theory of the integration of correspondences that emerges as its corollary; Hildenbrand (1974) is the relevant reference.

A contemporaneous formulation of Vind (1964) short-circuits some of these issues concerning sets of zero measure by ignoring agents altogether, and focusing instead on coalitions, each with its own preferences and endowments, as the primitive data of the economy. Allocations then are measures on a non-atomic measure-space, and the notions of core and competitive allocations, correspondingly defined, can be shown to be identical solution concepts. This is a formulation of perfect competition that is also measure-theoretic, but one, alternative to that of Aumann, that explicitly does away with mathematical integration as its necessary microfoundation. However, by assuming countable additivity, Vind enabled Debreu (1967) to draw on Radon–Nikodym differentiation to effect a reconciliation. It took subsequent work of Armstrong and Richter to give fuller autonomy to this alternative point of view by first eliminating countable additivity, and then in setting the discussion in the framework of non-atomic Boolean algebras; see Armstrong and Richter (1986) and their references. Whereas the technical underpinning of this approach is now clearly seen to be the Armstrong and Prikry (1981) extension of the Lyapunov theorem, it is perhaps fair to say that the conceptual ramifications of this alternative (perhaps syndicalist) vision have yet to be fully explored and understood; see Avallone and Basile (1998) and Basile and Graziano (2001) for references to current research.

The formulation of perfect competition due to Brown and Robinson (1975), the third to be discussed here, returns to the methodological individualism of Aumann, and requires the set of agent-names *T* to be an internal star-finite set, the commodity space to be ${}^*\mathrm{IR}^{\ell}_{+}$, the nonstandard extension of IR^{ℓ}_{+} based on manipulable infinitely large and infinitesimally small numbers, the summation in the definitions of allocations and core to be summation over internal sets, the set of admissible coalitions to be the set of all internal subsets of *T* and $\mathcal{E}$ to be an internal map from *T* to ${}^*\mathcal{P}$, the set of agent characteristics modelled on ${}^*\mathrm{IR}^{\ell}_{+}$. Such a formulation utilizes methods of nonstandard analysis, a specialization in mathematical logic due to A. Robinson; see Loeb and Wolff (2000) for details and references. On replacing equality by equality modulo infinitesimals in the definitions of allocation and the core, Brown and Robinson (1975) and, without their ad hoc standardly bounded assumption on allocations, Brown and Khan (1980) showed the equivalence (and Brown, 1976, and Khan, 1975, the existence) of core and competitive allocations of a non-standard economy without any convexity assumptions on preferences. Loeb's (1973) combinatorial analogue of Lyapunov's theorem provided the mathematical underpinning of the theory. This alternative affirmation of the conjectures of Edgeworth and Farrell is another way of making precise the concepts of *many* agents and of their individual *negligibility*: meaning can be given to an individual trader's actions having a positive, but infinitesimal, effect on the price system and on an

allocation. Even though an initial motivation of this work was to explore a formulation of perfect competition and of a *large* economy in a vernacular alternative to that of measure theory, it was heavily influenced by measure-theoretic formulations, but with an added emphasis on asymptotic implementation (discussed below), something clear even in the earliest papers of Brown-Robinson and Khan; see Rashid (1987), Anderson (1991) for details and references.

Relative to the classical theory brought to a culmination by Arrow, Debreu, McKenzie, Uzawa, Gale, Nikaido and Negishi, and succinctly surveyed in Koopmans (1961), the literature discussed above can be read as an exploration of the structural analytics of the set of agents of a stylized economy. Where Aumann takes the replicated sequence of Debreu–Scarf to a countably additive atomless measure space of agents, Brown–Robinson take it to a star-finite internal set each of whose points (agents) is given the same weight, and Armstrong–Richter, following Vind's cue, to a finitely-additive atomless measure space of coalitions. A fourth direction, intriguing and not yet fully synthesized in and with the other three, is represented in the work of Kaneko–Wooders (1986; 1989) and Hammond–Kaneko–Wooders (1989); also see Hammond (1995), Kaneko–Wooders (1994; 1996), Winter and Wooders (1994) and their references. The heart of this approach is to grapple with absolute and proportional magnitudes within the same framework, to focus on finite coalitions chosen from a continuum, through the notion of a *measure-consistent* partition. It concerns an atomless countably additive measure space of agents in which a single agent (and therefore a finite set of agents) is closed and thereby measurable, and a set of measure-preserving isomorphisms. A notion of an *f*-core is formulated and shown to be equivalent to the set of competitive equilibria *even* with externalities, and to the so-called Aumann core, without externalities; Wooders (1997) focuses on public goods. This approach yields its own particular way of looking at the continuum as a idealized limit of a finite economy, one that revolves around finer and finer measure-consistent partitions of an atomless continuum. It is thereby different in spirit from the more conventional way that asymptotic implementation has been formalized. We refer the reader to the references for details, and turn to what is seen here as the fifth formulation of perfect competition.

Strange as it may seem in retrospect, the idealizations of Aumann and Brown–Robinson were criticized on grounds of realism, on the observation that there do not exist economies with uncountably many agents; see Koopmans (1974) and the Georgescu-Roegen–Rashid exchange discussed in Khan (1998). The work categorized here as an asymptotic implementation of the idealized limiting versions of perfect competition was motivated, in part, by this criticism (ironically also used by Armstrong–Richter as their stated motivation for finitely additive measures), and, in part, by a methodological curiosity as to whether the results established for non-standard and measure-theoretic economies are artifacts of the way *negligibility* and *large* economies were being modelled. Taking its point of departure from the replicated sequences of Debreu and Scarf (1963), the response is to consider a sequence $\mathcal{G} = \{\mathcal{E}_k\}_{k=1}^{\infty}$ of finite economies based on the commodity space IR_+^{ℓ} where $\mathcal{E}_k$ is an economy with a set of agents T_k of cardinality k. For each finite economy $\mathcal{E}_k$, competitive and core allocations can be defined in the conventional way without encountering any technical difficulties in the formalization of summation or of a coalition. It is clear that agents in $\mathcal{E}_k$ get increasingly numerically negligible with an increase in k, and given a uniformly bounded assumption on initial endowments, also get increasingly economically negligible. For this *perfectly competitive sequence* of economies, one can ask: for any $\varepsilon > 0$, however

small, does there exist an integer k_o such that core allocations of all $\mathcal{E}_k \in \mathcal{G}$, $k \geq k_0$, can be sustained as approximate competitive equilibria, and whether such equilibria exist, with ε indicating in either instance, the degree of approximation? In short, are the formulations of perfect competition in idealized limit economies capable of an asymptotic implementation, with an arbitrarily fine degree of approximation, in economies of arbitrarily large but finite cardinality?

Asymptotic equivalence and existence theorems under varying degrees of generality followed quickly once the problem was posed. We shall not touch upon the various elaborations and refinements except to note that they have been obtained under two disparate techniques, both drawing on the results for an idealized limit economy. The first, associated especially with Hildenbrand, is to conceive of an economy as a measure on the space of characteristics and to utilize Skorokhod's theorem and the theory of weak convergence of measures on a topological space (typically metrizable) of characteristics $\mathcal{P}$. Under Debreu's rather vivid terminology of ‘neighboring economic agents’, such topologies were formulated by Debreu, Kannai, Hildenbrand–Mertens, Grodal and others, and surely have independent interest; see Hildenbrand (1974). The second approach is based on the observation that ‘any sentence which is true in the standard universe is true for internal entities in the nonstandard universe’, and as such, results pertaining to a nonstandard exchange economy can be ‘flipped over’, as it were, to a corresponding result for a large but finite economy. The differences between the two approaches are interesting from a methodological point of view: the fact that one approach is, in principle, not inherently dependent on any topology on the space of preference relations or on their continuity (as in Khan and Rashid, 1976; 1982) and applies as readily to core as to competitive allocations (as in Khan, 1974), suggests a further look as to how the other may be extended; see Anderson (1992) for a comprehensive treatment. In any case, we have two mutually supporting ways of extracting information for large but finite economies from idealized limit economies, even of the mixed type with atoms that generated the scepticism about idealized limit economies in the first place; see Gabszewicz and Shitovitz (1992) and their references. This claim is further underscored by a development due to Loeb (1975), but before turning to it, we discuss what may be seen as fifth formulation of *negligibility* and thereby of perfect competition.

The asymptotic interpretation of the perfectness of perfect competition concerns sequences of economies, and a question arises as to whether, given an arbitrary economy rather than an arbitrary degree of approximation, one can find the error, independent of the number of agents, with which the equivalence and existence theorems hold. Thus, rather than ask how large is large enough, one asks how small is small enough for the assumption of price-taking behaviour to be unjustified. For the question posed in this way, initially by Starr (1969), it was the definitive result of Anderson (1978) that capped initial explorations of Arrow–Hahn, Henry, Shaked and others. With the shedding of compactness and continuity assumptions under the nonstandard approach, Anderson observed that the argument in Khan and Rashid (1976) could be based on the Shapley–Folkman theorem instead of that of Loeb (1973) (itself based on Steinitz's theorem), and carried out entirely in standard terms to obtain an elementary equivalence theorem. This yields the asymptotic results as corollaries, and also furnishes them with a rate of convergence, a consideration emphasized by Shapley (1975). The same observation applied to Khan and Rashid (1982) led to an elementary existence theorem; see Geller's (1986) extension of Anderson, Khan and Rashid (1982).

In the prominence that it gives to a fixed finite economy, this sixth and final fifth formulation of perfect

competition connects directly to the results whose introduction began this entry; it emphasizes that the equalities in the results surveyed by Koopmans, and the counter-examples implicitly underlying them, perhaps ought to be given a probabilistic cast rather than taken completely literally. In his alternative proof of the Shapley–Folkman theorem, Cassels (1975) had already emphasized this connection. Mas-Colell deepened it further by appealing to results of especial sophistication concerning the law of large numbers and the central limit theorem, and by noting that his refinement of the equivalence theorem has ‘no analogue in Aumann's continuum of traders model’, and that the precise probabilistic estimates that this approach offers have no counterpart in the continuum framework (see Anderson, 1992, Sections 8 and 9 for details and precise references). However, it is undeniable that it is the exact results for the idealized limit economies that generally indicate the directions of pursuit of the approximations for a finite economy: approximations and numerical algorithms come into play once the exact has been exactly identified. Thus, from a substantive point of view, modulo fine technicalities, how a particular issue pertaining to perfect competition is set, measure-theoretic or nonstandard or asymptotic, is largely a contextual matter of analytical convenience and preference.

This conclusion is further sharpened by the methodological unification offered in Loeb (1975) (see Khan and Sun, 1997b, for exposition). It is the central claim of this article that Loeb probability spaces go a long way towards settling the question of how the perfectness of perfect competition is to be given a precise mathematical formulation. It is already clear in Aumann's pioneering papers that perfect competition draws from the atomlessness rather than any other particularities of the measure space of agents: the metric on the unit interval, or the topology of any topological measure space, is not, indeed cannot be, of any direct relevance. What is presumably of the essence is that the space of agents’ names be hospitable to measurability as well as to independence (the latter term now being used in its precise probabilistic sense rather than as a reference to an absence of externalities), that it generates results capable of straightforward asymptotic implementation, and that, for concepts that revolve only on distributions of the allocations as in Hart–Kohlberg, it yields solutions that are insensitive to a permutation of agent names. In the context of *large* games (discussed below), Khan and Sun (1996; 1999b) make the case for Loeb spaces on the basis of these *desiderata* and emphasize their dual identity in the ‘pushing down’ and ‘lifting up’ theorems: being standard, measure spaces, any result on an abstract measure-space (Aumann) economy applies to them, and thereby to an internal non-standard (Brown–Robinson) economy and hence can be asymptotically interpreted; or alternatively, any approximate result can be translated, as indicated above, to a non-standard economy, and thereby pushed down to its standard Loeb measure-theoretic counterpart. As such, Loeb spaces go a considerable way in obliterating the sixfold categorization of perfect competition that marks this entry.

Going beyond method to mathematical substance, atomless Loeb spaces are ideally suited for operations ensuring that aggregation removes the irregularities that arise from non-convexities as well as from idiosyncratic uncertainty. In a systematic and far-reaching development, Sun established that the integrals and distributions of correspondences defined on Loeb spaces and taking values in a separable infinite-dimensional Banach space, in the first instance, and into Polish spaces (separable and completely metrizable) in the second, have all the properties that the theory of perfect competition requires of them. Moreover, a perfectly satisfactory law of large numbers for a continuum of random variables is obtained, and for a such a continuum, the notions of *independence* and of *exchangeability* are dual in a very elegant sense, and yields, as in Duffie and Sun (2007), the existence of an independent

random matching. Supplementing the notion of an economy as a random variable, the measurability of the map noted above, a stochastic economy can now be formalized as a stochastic process on a product space, the space of agent names T and an atomless Loeb space of states of nature, Ω, to reveal circumstances under which the distributions of core and competitive allocations of a sampled economy coincide, or approximately coincide in the case of a large economy, with those of the deterministic (population) economy; see Sun (1999). Further application of this substantial theory is noted below; here the reader is referred to Sun's chapters in Loeb and Wolff (2000, chs. 7 and 8) for exposition and full mathematical references. (For references to work on random economies that does not rely on Loeb spaces, see Radner, 1982, Section 7.6, and Majumdar and Rotar, 2000.)

In taking stock at this stage, we underscore the fact that even though six robust and logically related methods of studying perfect competition have been illustrated through the conjectures of Edgeworth and Farrell, the discussion could, in principle, equally well have been conducted through alternative tests based on alternative solution concepts: the value (Hart, 2002 and his references), or the bargaining set (Anderson, 1998 and his references), or Cournot's conjecture (Mas-Colell, 1986; Novshek and Sonnenschein, 1983 and their references), all now conceived in a setting where individual agents are negligible. Alternatively, we could discuss applications, particularly in mathematical finance where Arrow markets and ideas of negligibility find concrete expression in derivative financial instruments and in *well-diversified* portfolios (see Anderson and Raimondo, 2006; Khan and Sun, 1997a, respectively, for references). However, rather than turn to them and make this article unmanageable, we draw on the rich and diverse formulation of perfect competition at our disposal to consider the substantive issues broached earlier: public goods, externalities, increasing returns to scale and infinite commodities, all under the rubric of static interaction. Ironically, non-convexities in idealized limit economies have concerned consumptions sets and survival assumptions rather than increasing returns to scale technologies (see Trockel, 1984; Hammond, 1993 and their references); research efforts have been most active in the study of public goods and externalities, and here the theory dovetails, from a technical point of view, into work on infinite-dimensional commodity spaces.

The formalization and defence of perfect competition has, from the very beginning, proceeded on the independence assumption: the fact that individual agents are not related other than through the price system, with a 1952 paper of McKenzie's being the sole exception. Thus Hayek (1948, pp. 96–7) quotes Stigler in emphasizing the ‘explicit and complete exclusion from the theory … of all personal relationships existing between the parties’. Such relationships are *external* to the perfected concept, and, to the extent that positive and normative content can be cleanly distinguished, externalities, and the Pigovian private–social divergences that they entail, have strong and negative implications for its normative content. If the non-convexities identified by Starrett (1972) are ignored (but also see Otani and Sicilian, 1977), Arrow's universality requirement for the first fundamental theorem of welfare economics can always be met by the creation of markets, fictitious or otherwise, but it clearly leans on a particularly acute form of myopia. Arrow securities and Lindahl prices for public goods, and more generally, prices for contingent commodities, and personalized prices for more pervasive externalities, bring out an obvious tension between incentive compatibility and efficiency. As emphasised in Starrett (1971), if there is a commodity that reflects a particular agent's dependence on my consumption, why should she or I, let alone the others, take the price of that commodity to be given and non-manipulable? or take myself to be economically negligible?

Of course, one response to these difficulties is to face the future as a future without the fiction of a complete set of commodity markets or, equally imaginatively, existing markets for securities that span *all* contingencies. Under this alternative, one can regard the price system itself as a means of fostering a relationship between the parties, and to conceive of a rationality which explicitly incorporates the informational resources of the *others* in the economy. This is to look on a price system as an instrument of solidarity as well as an instrument of allocation, a keeping up with the Joneses, not so much in their actions as in the individualized information that undergirds their actions, a move reminiscent of Veblen in the space of information rather than that of conspicuous consumption. This is a move inaugurated by Radner (1967), and it leads to a notion of equilibrium, a *rational expectations equilibrium* so to speak, in which both aspects of the price system are taken into account while not necessarily departing from the purview of the static Arrow–Debreu–McKenzie theory. One can only wonder what mathematical form such a theory will take when it is set in the framework in which individual agents are *negligible;* we point the reader to Radner (1982) and Jordan and Radner (1982) for details and references, and revert to the idealized limit economy.

There is also a technical problem in the consideration of pervasive externalities in an idealized limit economy. Since the individualistic, as opposed to the coalitionally based, approach to perfect competition works with an equivalence class of functions from the space of agent-names to agent-actions rather than the function itself, it is difficult to give meaning to one agent's dependence on the actions of another. In a context of a Lindahl equilibrium of an idealized limit economy, even one with a finite number of commodities and a single public good, one has to reckon with the fact that public goods enjoin equality instead of aggregation, and thereby force the analysis out of a finite-dimensional Euclidean space, as in the Aumann–Brown–Robinson limit theory, to a search for a suitably tractable space of equivalence classes of functions of individualized prices. It is these attendant functional-analytic difficulties, perhaps as much as the fact that the incentive-compatibility problems are most acute in this setting, that have discouraged the initial exploratory attempts of Roberts, Emmons and Khan and Vohra from being followed up; see Khan and Vohra (1985) for references. And it is precisely difficulties of this kind that also prevent a successful theory for idealized limit economies with non-ordered preferences; see Balder's (2000) interpretive use of the argument in Khan and Papageorgiou (1987), originally due to Grodal, to turn a positive proof into a negative claim of inconsistency, a claim that apparently derails the initial exploration of Khan and Vohra (1984) and their followers. Externalities, rather than being widespread, need to be controlled and confined in an idealized limit economy. This previous sentence, as well as the tone of this entire paragraph so far, runs counter to the fourth approach to perfect competition associated with Hammond, Kaneko and Wooders, but, as emphasized above, the integration of this fourth approach with the other five has not yet been fully achieved. The theory is under active development, and it is too early to say that a formulation sufficiently robust as to be deemed canonical has been achieved (see Balder, 2007b; Cornet and Topuzo, 2005; Hammond, 1995; Kaneko and Wooders, 1994; Noguchi, 2005; Noguchi and Zame, 2006 and their references).

In its dissociation of the study of perfect competition from its roots in welfare theory, the inclusion of externalities makes explicit its connection to game theory. Competitive equilibria with externalities take their place next to marginal-cost pricing and Cournot–Nash equilibria in violating Pareto optimality, but do allow one to ask whether decentralized self-interested decision-making is consistent in the aggregate if it is taken with respect to certain measurable indices of societal responses rather than solely with

respect to a price system. Such a formulation of perfect competition goes back to the early 1950s in the papers of McKenzie and Debreu, and to the 1970s in Chipman's formulations of Marshallian parametric externalities. Indeed, the original proof of Arrow–Debreu of the existence of competitive equilibrium revolved around viewing the economy as a game in which the only ‘personal relationship’ between the parties relates to that with a fictitious auctioneer, a point of view that finds fuller expression in the Shafer–Sonnenschein notion of an *abstract economy*. In more recent investigations of a *large game*, the literature takes another turn towards probability theory, and conceives of an agent's actions as resulting from maximization that takes as given the distribution, or individual moments, of the random variable summarizing societal responses. The question then reduces to the existence of such equilibrium distributions, but with social interaction, however limited, recourse has to be made to assumptions on ideal types, and on the *conditional* or *mutual independence* of these types (see the prescient remarks of Hayek, 1948, p. 47). This is a theory of competition in which Loeb spaces, and the Dvoretsky–Wald–Wolfowitz extension of the Lyapunov theorem play a dominant role; see Khan and Sun (1999b; 2002); Khan, Rath and Sun (2006) Loeb and Sun (2006) and their references to the work of Schmeidler, Radner–Rosenthal, Milgrom–Weber and Mas-Colell. (Balder (2007a) offers a perspective based on Young measures.)

The technical machinery forged through the study of large games enables a broadened notion of economic negligibility, one that includes *informational negligibility* in an environment with asymmetric information. In a 1936 article on ‘Economics and Knowledge’, Hayek (1948, pp. 43–44) had already supplemented Adam Smith's emphasis on the *division of labour* by the principle of the *division of knowledge* and asked

> whether, in order that we can speak of equilibrium, every single individual must be right, or whether it would not be sufficient if, in consequence of a compensation of errors in different directions, quantities of the different commodities coming on the market were the same as if every individual had been right. A fuller discussion of this problem would have to consider the whole question of the significance which some economists (including Pareto) attach to the law of great numbers in this connection.

The issue is: ‘right’ about what? The problem devolves on anticipations and expectations, beliefs about beliefs regarding each other and the price system, and it does not require more than a mild degree of scepticism to abandon fictional markets responding to predetermined and universally agreed upon states of nature. There is a need for viable notions of independence and aggregation to eliminate idiosyncratic risk and nullify ‘combination of fragments of knowledge existing in different minds’. Sun (2006) and Sun and Yannelis (2007a; 2007b) give pride of place to the Fubini property in idealized limit economies, and consolidate earlier applications of Loeb spaces for a successful resolution of Malinvaud's work on insurance markets, and that of Gul, McLean and Postlewaite on the compatibility of efficiency and incentive compatibility; also see Jackson and Manelli (1997). Khan and Sun (1999a) and Sun (2006) also present compelling arguments why finitely additive measures and the conventional product measure cannot respond to the technical difficulties.

The problems arising from asymmetric information are, at their root, problems of agent interdependence that cannot be internalized through markets, and as such represent particularly recalcitrant externalities;

the assumptions that Sun–Yannelis impose on their signal process can be seen as one successful attempt to subdue them. And in an idealized limit economy with *many* commodities, each commodity seen on its own rather than through the externalities' lens, one has to cope with the fact that Lyapunov's theorem is false for an infinite dimensional vector measure, in addition to all of the problems discussed earlier. It is the *thinness* of its target space, as proposed by Kingman–Robertson in the late 1960s, that allows an atomless probability space of agents to work its magic in the form of the existence and equivalence theorems; see Kluvanek and Knowles (1976) and Diestel and Uhl (1977) for necessary and sufficient conditions for the validity of the Lyapunov theorem. There is a hidden assumption, to adopt the postmodern flourish of Tourky and Yannelis (2001), in the Aumann–Brown–Robinson formulations of perfect competition, and the equivalence theorem can fail when the qualitative relationship between the cardinalities of agents and commodities fails; in addition to Muench's example, see Forges, Heifetz and Minelli (2001) and Serrano, Vohra and Volij (2001). More generally, if the intricacies of reaching binding agreements in coalition formation cannot be bracketed away, how can a concept embodying group rationality coincide with one hinging on individual rationality? An option, but one that goes against the very grain of this article, is to dissociate competition from price-taking entirely and derive it as a consequence, as in the no-surplus characterizations of Makowski and Ostroy (2001, Section 9) and Serrano and Volij (2000). The field is under active development; in addition to the papers of Sun, Tourky and Yannelis, see Forges, Minelli and Vohra (2002), Herves-Beloso, Moreno-Garcia and Yannelis (2005), Martins-da-Rocha (2003; 2004), and Podczeck (1997; 2001; 2004) and their references. In his classic 1936 tour de force, Hayek deconstructed the Arrow–Debreu–McKenzie construction before it was constructed, so to speak, by distinguishing between an a priori 'pure logic of choice' and an empirical science. In so far as this article, in its focus on existence and core equivalence, has concentrated on the adjective *perfect*, and avoided questions of cardinality, computability, learning and stability of a perfectly competitive allocation of resources, it has neglected the noun *competition* as being outside its scope. For this, the reader could perhaps begin with Morgan (1993), and move from there to Arrow (1986), Buchanan (1987) and Radner (1991), and from there, if she is still so inclined, to the entire gamut of economic theory.

See Also

- competition
- large economies
- measure theory
- non-standard analysis
- Shapley–Folkman theorem

Bibliography

Aliprantis, C.D., Cornet, B. and Tourky, R. 2002. Economic equilibrium: optimality and price decentralization. *Positivity* 6, 205–41.

Aliprantis, C.D., Florenzano, M. and Tourky, R. 2006. Production equilibria. *Journal of Mathematical Economics* 42, 406–21.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46, 1483–7.

Anderson, R.M. 1991. Non-standard analysis with applications to economics. In *Handbook of Mathematical Economics*, vol. 4, ed. W. Hildenbrand and H. Sonnenschein. New York: North-Holland.

Anderson, R.M. 1992. The core in perfectly competitive economies. In *Handbook of Game Theory with Economic Applications*, vol. 1, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Anderson, R.M. 1998. Convergence of the Aumann–Davis–Maschler and Geanakopolos bargaining sets. *Economic Theory* 11, 1–37.

Anderson, R.M., Khan, M.A. and Rashid, S. 1982. Approximate equilibria with bounds independent of preferences. *Review of Economic Studies* 49, 473–5.

Anderson, R.M. and Raimondo, R. 2006. Equilibrium in continuous-time financial markets: endogenously dynamically complete markets. Working paper, University of California at Berkeley.

Araujo, A., Martins-da-Rocha, V.F. and Monteiro, P. 2004. Equilibria in reflexive Banach lattices with a continuum of agents. *Economic Theory* 24, 469–92.

Armstrong, T.E. and Prikry, K. 1981. Liapunoff's theorem for nonatomic, finitely additive, bounded, finite-dimensional vector-valued measures. *Transactions of the American Mathematical Society* 266, 499–514. Erratum in the same journal (1982), 272, 809.

Armstrong, T. and Richter, M.K. 1986. Existence of nonatomic core-Walras allocation. *Journal of Economic Theory* 38, 137–59.

Arrow, K.J. 1986. Economic theory and the hypothesis of rationality. *Journal of Business* 59, S385–99.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

Aumann, R.J. 1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34, 1–17.

Avallone, A. and Basile, A. 1998. Lyapunov–Richter theorem in B-convex spaces. *Journal of Mathematical Economics* 30, 109–18.

Balder, E.J. 2000. Incompatibility of usual conditions for equilibrium existence in continuum economies

without ordered preferences. *Journal of Economic Theory* 93, 110–7.

Balder, E.J. 2007a. Comments on purification in continuum games. *International Journal of Game Theory*, forthcoming.

Balder, E.J. 2007b. More on equilibria in competitive markets with externalities and a continuum of agents. *Journal of Mathematical Economics*, forthcoming.

Basile, A. and Graziano, M.G. 2001. Restricted coalition formation mechanisms in finitely-additive economies. *Journal of Mathematical Economics* 36, 219–40.

Bonnisseau, J.M. 2002. The marginal pricing rule in economies with infinitely many commodities. *Positivity* 6, 275–96.

Bonnisseau, J.M. and Cornet, B. 2006. Existence of equilibria with a tight marginal pricing rule. *Cahier de la MSE*, Université Paris 1.

Brown, D.J. 1976. Existence of competitive equilibrium in a non standard exchange economy. *Econometrica* 44, 537–47.

Brown, D.J. and Khan, M.A. 1980. An extension of the Brown–Robinson equivalence theorem. *Applied Mathematics and Computation* 6, 167–75.

Brown, D.J. and Robinson, A. 1975. Non standard exchange economies. *Econometrica* 43, 41–55.

Buchanan, J.M. 1987. *Economics: Between Predictive Science and Moral Philosophy*. College Station: Texas A&M University Press.

Cassels, J.W.S. 1975. Measures of the non-convexity of sets and the Shapley–Folkman–Starr theorem. *Mathematical Proceedings of the Cambridge Philosophical Society* 78, 433–6.

Cornet, B. and Topuzu, M. 2005. Existence of equilibria for economies with externalities and a measure space of consumers. *Economic Theory* 26, 397–421.

Debreu, G. 1959. *The Theory of Value*. New York: Wiley.

Debreu, G. 1967. Preference functions on measure spaces of economic agents. *Econometrica* 35, 111–22.

Debreu, G. and Scarf, H. 1963. A limit theorem on the core of an economy. *International Economic Review* 4, 235–46.

Diestel, J. and Uhl, J.J. 1977. *Vector Measures*. Providence, RI: American Mathematical Society.

Duffie, D. and Sun, Y.N. 2007. Existence of independent random matching. *Annals of Applied Probability* 17, 386–419.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan-Paul.

Farrell, M.J. 1959. The convexity assumption in the theory of competitive markets. *Journal of Political Economy* 67, 377–91.

Fenchel, H. 1951. *Convex Cones, Sets and Functions*. Lecture Notes. Princeton, NJ: Princeton University Press.

Foley, D. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38, 66–72.

Forges, F., Heifetz, A. and Minelli, E. 2001. Incentive compatible core and competitive equilibria in differential information economies. *Economic Theory* 18, 349–65.

Forges, F., Minelli, E. and Vohra, R. 2002. Incentives and the core of an exchange economy: a survey. *Journal of Mathematical Economics* 38, 1–41.

Gabszewicz, J. and Shitovitz, B. 1992. The core in imperfectly competitive economies. In *Handbook of Game Theory with Economic Applications*, vol. 1, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Geller, W. 1986. A improved bound for approximate equilibria. *Review of Economic Studies* 53, 307–8.

Hamano, T. 1989. On the non-existence of the marginal cost pricing equilibrium and the Ioffe normal cone. *Zeitschrift für National Ökonomie* 50, 47–53.

Hammond, P.J., Kaneko, M. and Wooders, M.H. 1989. Continuum economies with finite coalitions: core equilibria and widespread externalities. *Journal of Economic Theory* 49, 113–134.

Hammond, P.J. 1993. Irreducibility, resource-relatedness, and survival in equilibrium with individual non-convexities. In *General Equilibrium, Growth and Trade II*, ed. R. Becker, M. Boldrin, R. Jones and W. Thomson. New York: Academic Press.

Hammond, P.J. 1995. Four characterizations of constrained Pareto efficiency in continuum economies with widespread externalities. *Japanese Economic Review* 46, 103–24.

Hart, S. 2002. Values of perfectly competitive economies. In *Handbook of Game Theory with Economic Applications*, vol. 3, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Hayek, F.A. 1948. *Individualism and Economic Order*. Chicago: Chicago University Press.

Herves-Beloso, C., Moreno-Garcia, E. and Yannelis, N.C. 2005. An equivalence theorem for a differential information economy. *Journal of Mathematical Economics* 41, 844–56.

Hicks, J.R. 1939. The foundations of welfare economics. *Economic Journal* 49, 696–712. Also, Prefatory note to the 1984 reprint in *Wealth and Welfare: Collected Essays in Economic Theory*, vol. 1. Oxford: Basil Blackwell.

Hildenbrand, W. 1974. *Core and Equilibria of a Large Economy*. Princeton, NJ: Princeton University Press.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6, 242–69.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and Organization*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Ingrao, B. and Israel, G. 1987. *La Mano Invisibile*. Roma-Bari: Gius. Laterza & Figli Spa. English translation, *The Invisible Hand: Economic Equilibrium in the History of Science*. Cambridge: MIT Press, 1990.

Jackson, M. and Manelli, A. 1997. Approximate competitive equilibria in large economies. *Journal of Economic Theory* 77, 354–76.

Jordan, J. and Radner, R. 1982. Rational expectations in microeconomic models: an overview. *Journal of Economic Theory* 26, 201–23.

Kaneko, M. and Wooders, M.H. 1986. The core of a game with a continuum of players and finite coalitions: the model and some results. *Mathematical Social Sciences* 12, 105–37.

Kaneko, M. and Wooders, M.H. 1989. The core of a continuum economy with widespread externalities and finite coalitions: from finite to continuum economies. *Journal of Economic Theory* 49, 135–68.

Kaneko, M. and Wooders, M.H. 1994. Widespread externalities and perfectly competitive markets: examples. In *Imperfections and Behavior in Economic Organizations*, ed. R.P. Gilles and P.H.M. Ruys. Boston: Kluwer.

Kaneko, M. and Wooders, M.H. 1996. The nonemptiness of the *f*-core of a game without side-payments. *International Journal of Game Theory* 25, 245–58.

Khan, M.A. 1974. Some remarks on the core of a ‘large’ economy. *Econometrica* 42, 633–42.

Khan, M.A. 1975. Some approximate equilibria. *Journal of Mathematical Economics* 2, 63–86.

Khan, M.A. 1988. Ioffe's normal cone and the foundations of welfare economics: an example. *Economics Letters* 28, 15–9.

Khan, M.A. 1993. Lionel McKenzie and the existence of competitive equilibrium. In *General Equilibrium, Growth and Trade II*, ed. R. Becker, M. Boldrin, R. Jones and W. Thomson. New York: Academic Press.

Khan, M.A. 1998. Representation, language and theory: Georgescu-Roegen on methods of economic science. Paper presented at the Colloque International L’Oeuvre Scientifique de Nicholas Georgescu-Roegen, Strasbourg, 7–9 November.

Khan, M.A. 1999. The Mordukhovich normal cone and the foundations of welfare economics. *Journal of Public Economic Theory* 1, 309–38.

Khan, M.A. and Papageorgiou, N. 1987. On Cournot–Nash equilibrium in generalized qualitative games with an atomless measure space of players. *Proceedings of the American Mathematical Society* 100, 505–10.

Khan, M.A. and Peck, N.T. 1989. On the interiors of production sets in infinite dimensional spaces. *Journal of Mathematical Economics* 18, 29–39.

Khan, M.A. and Rashid, S. 1976. Limit theorems on cores with costs of coalition formation. Johns Hopkins Working Paper No 24. Abridged version published as ‘A limit theorem for an approximate core of a large but finite economy’, *Economics Letters* 1 (1978), 297–302.

Khan, M.A. and Rashid, S. 1982. Approximate equilibria in markets with indivisible commodities. *Journal of Economic Theory* 28, 82–101.

Khan, M.A., Rath, K.P. and Sun, Y.N. 2006. The Dvoretzky–Wald–Wolfowitz theorem and purification in atomless finite-action games. *International Journal of Game Theory* 34, 91–104.

Khan, M.A. and Sun, Y.N. 1996. Non-atomic games on Loeb spaces. *Proceedings of the National Academy of Sciences of the United States of America* 93, 15518–21.

Khan, M.A. and Sun, Y.N. 1997a. The capital-asset-pricing model and arbitrage pricing theory: a unification. *Proceedings of the National Academy of Sciences of the United States of America* 94, 4229–32.

Khan, M.A. and Sun, Y.N. 1997b. On Loeb measure spaces and their significance for non-cooperative game theory. In *Current and Future Directions in Applied Mathematics*, ed. M. Alber, B. Hu and J. Rosenthal. Berlin: Birkhäuser.

Khan, M.A. and Sun, Y.N. 1999a. Weak measurability and characterizations of risk. *Economic Theory* 13, 441–60.

Khan, M.A. and Sun, Y.N. 1999b. Non-cooperative games on hyperfinite Loeb spaces. *Journal of Mathematical Economics* 31, 455–92.

Khan, M.A. and Sun, Y.N. 2002. Non-cooperative games with many players. In *Handbook of Game Theory with Economic Applications*, vol. 3, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Khan, M.A. and Vohra, R. 1984. Equilibrium in abstract economies with out ordered preferences and with a measure space of agents. *Journal of Mathematical Economics* 13, 133–42.

Khan, M.A. and Vohra, R. 1985. On the existence of Lindahl equilibria in economies with a measure space of non-transitive consumers. *Journal of Economic Theory* 36, 319–32.

Khan, M.A. and Vohra, R. 1987. Lindahl–Hotelling equilibria. *Journal of Public Economics* 34, 143–58.

Khan, M.A. and Yannelis, N.C. 1991. *Equilibrium Theory in Infinite Dimensional Spaces*. New York: Springer.

Kluvanek, I. and Knowles, G. 1976. *Vector Measures and Control Systems*. Amsterdam: North-Holland.

Koopmans, T.C. 1961. Convexity assumptions, allocative efficiency, and competitive equilibrium. *Journal of Political Economy* 69, 478–9.

Koopmans, T.C. 1974. Is the theory of competitive equilibrium with it? *American Economic Review* 69, 325–9.

Loeb, P.A. 1973. A combinatorial analog of Lyapunov's theorem for infinitesimally generated atomic vector measures. *Proceedings of the American Mathematical Society* 39, 585–6.

Loeb, P.A. 1975. Conversion from nonstandard to standard measure spaces and applications in probability theory. *Transactions of the American Mathematical Society* 211, 113–22.

Loeb, P.A. and Sun, Y.N. 2006. Purification of measure-valued maps. *Illinois Journal of Mathematics* 50, 747–62.

Loeb, P.A. and Wolff, M. 2000. *Nonstandard Analysis for the Working Mathematician*. Dordrecht: Kluwer.

Majumdar, M. and Rotar, V. 2000. Equilibrium prices in a random exchange economy with dependent agents. *Economic Theory* 15, 531–50.

Makowski, L. and Ostroy, J.M. 2001. Perfect competition and the creativity of the market. *Journal of Economic Literature* 39, 479–535.

Martins-da-Rocha, V.F. 2003. Equilibria in large economies with a separable Banach commodity space and non-ordered preferences. *Journal of Mathematical Economics* 39, 863–89.

Martins-da-Rocha, V.F. 2004. Equilibria in large economies with differentiated commodities and non-ordered preferences. *Economic Theory* 23, 529–52.

Mas-Colell, A. 1986. The price equilibrium existence problem in topological vector lattices. *Econometrica* 54, 1039–54.

Mas-Colell, A. and Zame, W.R. 1991. Equilibrium theory in infinite dimensional spaces. In *Handbook of Mathematical Economics*, vol. 4, ed. W. Hildenbrand and H. Sonnenschein. Amsterdam: North-Holland.

McKenzie, L.W. 2002. *Classical General Equilibrium Theory*. Cambridge: MIT Press.

Mordukhovich, B.S. 2006. *Variational Analysis and Generalized Differentiation, I & II*. Berlin: Springer.

Mirowski, P. 2002. *Machine Dreams*. Cambridge: Cambridge University Press.

Morgan, M. 1993. Competing notions of ‘competition’ in late nineteenth-century American economics. *History of Political Economy* 25, 563–604.

Noguchi, M. 2005. Interdependent preferences with a continuum of agents. *Journal of Mathematical Economics* 41, 665–86.

Noguchi, M. and Zame, W.R. 2006. Competitive markets with externalities. *Theoretical Economics* 1, 143–66.

Novshek, W. and Sonnenschein, H. 1983. Walrasian equilibria as limits of noncooperative equilibria.

Journal of Economic Theory 30, 171–87.

Otani, Y. and Sicilian, J. 1977. Externalities and problems of nonconvexity and overhead costs in welfare economics. *Journal of Economic Theory* 14, 239–52.

Pigou, A.C. 1932. *The Economics of Welfare*, 4th edn. London: Macmillan. 1st edn, 1920.

Podczeck, K. 1997. Markets with infinite commodities and a continuum of agents with non-convex preferences. *Economic Theory* 9, 385–426.

Podczeck, K. 2001. Core and Walrasian equilibria when agents' characteristics are extremely dispersed. *Economic Theory* 22, 699–725.

Podczeck, K. 2004. On Core–Walras equivalence in Banach spaces when feasibility is defined by the Pettis integral. *Journal of Mathematical Economics* 40, 429–63.

Radner, R. 1967. Equilibre des marchés à terme et au comptant en cas d'incertitude. *Cahiers d'econométrie* 9, 30–47.

Radner, R. 1982. Equilibrium under uncertainty. In *Handbook of Mathematical Economics*, vol. 2, ed. K. J. Arrow and M.D. Intriligator. New York: North-Holland.

Radner, R. 1991. Intertemporal general equilibrium. In *Value and Capital: Fifty Years Later*, ed. L.W. McKenzie and S. Zemagni. London: Macmillan.

Rashid, S. 1987. *Economies with Many Agents*. Baltimore, MD: Johns Hopkins Press.

Rockafellar, R.T. 1970. *Convex Analysis*. Princeton: Princeton University Press.

Rockafellar, R.T. and Wets, J.-B. 1998. *Variational Analysis*. New York: Springer.

Rudin, W. 1974. *Real and Complex Analysis*. New York: McGraw-Hill Book Co.

Samuelson, P.A. 1954. The pure theory of public expenditures. *Review of Economics and Statistics* 36, 387–9.

Serrano, R., Vohra, R. and Volij, O. 2001. On the failure of core convergence with asymmetric information. *Econometrica* 69, 1685–96.

Serrano, R. and Volij, O. 2000. Walrasian allocations without price-taking behavior. *Journal of Economic Theory* 95, 79–106.

Shannon, C. 1999. Increasing returns in infinite horizon economies. *Review of Economics Studies* 64, 73–96.

Shapley, L.S. 1975. An example of a slow-converging core. *International Economic Review* 16, 345–51.

Starr, R.M. 1969. Quasi-equilibria in markets with non-convex preferences. *Econometrica* 37, 25–38.

Starrett, D.A. 1971. A note on externalities and the core. *Econometrica* 41, 179–83.

Starrett, D.A. 1972. Fundamental nonconvexities in the theory of externalities. *Journal of Economic Theory* 4, 180–99.

Sun, Y.N. 1999. The complete removal of individual uncertainty: multiple optimal choices and random economies. *Economic Theory* 14, 507–44.

Sun, Y.N. 2006. The exact law of large numbers via Fubini extension and the characterization of insurable risks. *Journal of Economic Theory* 126, 31–69.

Sun, Y.N. and Yannelis, N.C. 2007a. Perfect competition in asymmetric information economies: compatibility of efficiency and incentives. *Journal of Economic Theory* 134, 175–94.

Sun, Y.N. and Yannelis, N.C. 2007b. Core, equilibria and incentives in large asymmetric information economies. *Games and Economic Behavior*, forthcoming.

Trockel, W. 1984. *Market Demand: An Analysis of Large Economies with Non-convex Preferences*. Berlin: Springer.

Tourky, R. and Yannelis, N.C. 2001. Markets with many more agents than commodities: Aumann's hidden assumption. *Journal of Economic Theory* 101, 189–221.

Vind, K. 1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5, 165–77.

Vohra, R. 1992. Marginal cost pricing under bounded marginal returns. *Econometrica* 60, 859–76.

Weintraub, E.R. 1985. *General Equilibrium Analysis: Studies in Appraisal*. Cambridge: Cambridge University Press.

Winter, E. and Wooders, M.H. 1994. An axiomatization of the core for finite and continuum games. *Social Choice and Welfare* 11, 165–75.

Wooders, M.H. 1997. Equivalence of Lindahl equilibrium with participation prices and the core: an axiomatization of the core for finite and continuum games. *Economic Theory* 9, 115–27.

How to cite this article

Khan, M. Ali. "perfect competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000056> doi:10.1057/9780230226203.1267

The New Palgrave Dictionary of Economics Online

perfect foresight

Margaret Bray
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cobweb model; expectations; imperfect information; intertemporal general equilibrium; overlapping generations models; perfect foresight; probability distribution; rational expectations equilibrium; stationary state; uncertainty

Article

Perfect foresight is an occasionally convenient theoretical assumption whose total lack of realism is undisputed, and perhaps unrivalled. There are two elements to perfect foresight; firstly that people have definite point expectations, allowing no uncertainty, of future variables, and secondly that these expectations are correct. In practice, as these fortunate perfectly foresightful individuals generally inhabit models with instantaneously clearing perfectly competitive markets, they only need to forecast prices. The pioneering work by Hicks (1939) on intertemporal general equilibrium theory provides a framework in which the issues associated with perfect foresight can be explored. Writing prior to the development of the expected utility theory of choice under uncertainty (von Neumann and Morgenstern, 1944), Hicks had no alternative to a deterministic model in his discussion. He acknowledges the existence and importance of uncertainty in expectation formation, but argues in a somewhat unsatisfactory fashion that point predictions can be interpreted as risk-adjusted summaries of underlying probability distributions. Hicks divides time into weeks. Trade takes place weekly. Supply and demand in each week depend upon decisions made in the past, expectations of spot prices in future weeks, and current spot prices. In temporary equilibrium these spot prices adjust to clear markets, but expectations may be wrong. In the situation which Hicks terms ‘Equilibrium over Time’, markets clear at each date, and, crucially, everyone has perfect foresight; price expectations are fulfilled.
Hicks's insight that perfect foresight is an equilibrium concept is important. If people have non-equilibrium expectations, the temporary equilibrium prices in the current spot markets differ from the prices in full equilibrium over time, and the effects of current investment and production decisions based on mistaken expectations reverberate through the future. This can be illustrated in the simplest model of

supply and demand in which expectations play a part: the cobweb model, used by Kaldor (1934) in discussing disequilibrium adjustments, and by Muth (1961) in the paper which gave us the phrase ‘rational expectations’. In the cobweb model, demand at *t* D_t, depends upon the price at *t* p_t, $D_t = a - bp_t$. Supply depends upon point expectations p_t^e formed before *t* about p_t; $S_t = cp_t^e$. In temporary equilibrium supply equals demand, $a - bp_t = cp_t^e$, so $p_t = (a - cp_t^e) / b$. In the perfect foresight equilibrium expectations are correct $p_t^e = p_t = a / (b + c)$. If the price is and has been at the perfect foresight equilibrium level for a long time people will, quite reasonably, expect this price to persist. In an economy in a long-run stationary state with unchanging prices perfect foresight is plausible. Difficulties arise when a shift in an exogeneous variable changes the perfect foresight equilibrium price. Suppose that in the cobweb model an increase in costs causes the supply curve to shift to $S_t = c' p^e$. If people are aware of the change, and understand fully the working of their economy, they may at once calculate and expect the new equilibrium price; alternatively they may all believe the forecast generated by the brilliant economist who knows it all. Less well-informed people may be forced to use past prices in forming their expectations. If these expectations are not at the new equilibrium value $a / (b + c')$ actual prices also differ from equilibrium prices; the economy will take some time to adjust to its new equilibrium and may, as Kaldor shows, fail to get there at all. The dynamic adjustment process, as people try to learn from their mistakes, depends very much upon how they learn, and is not understood in any generality (Bray, 1983).

As Hicks argued, equilibrium over time with perfect foresight is most plausible when people expect prices to remain steady, and they do remain steady at the expected level. In the long-run stationary state with no uncertainty there is no need to distinguish between current prices and price expectations. Supply and demand can be thought of as relating to either. In this context the atemporal textbook theory of production and consumption can be reinterpreted to describe a world where production takes time, and is determined by price expectations as well as prices.

In the long-run stationary state tastes and technology must be unchanging and the size of the population and supplies of natural resources static, or possibly in a semi-stationary state growing steadily. These conditions are demanding and implausible. Further, they are not always sufficient for steady prices. As Grandmont (1985) shows, a very simple overlapping generations model has a constant price equilibrium, but may have other perfect foresight equilibria in which the price follows a very complicated, possibly chaotic, path. Unless people know precisely the underlying nonlinear difference equation generating prices they may have great difficulty in inferring prices from past prices.

Postulating perfect foresight allows another reinterpretation of an atemporal general equilibrium model to allow for time (Debreu, 1959). In the atemporal model there is a list of different commodities, and a market and price for each commodity. These markets all operate simultaneously; in general equilibrium they all clear. The same mathematical formalism can be used to describe an intertemporal model, by distinguishing commodities by their date of delivery as well as by their characteristics. Commodities may be produced or consumed at a number of different dates, but all trade takes place at the initial date, in a complete set of spot and contingent futures markets. This of course strains credibility; only a very limited number of futures markets exist. However, as Bliss (1975) shows, the same trades, production and consumption can take place if there is a futures market for one good at each date and spot markets for all other goods, provided everyone foresees the full equilibrium over time prices perfectly. There is

little to be gained in realism by exchanging the myth of complete markets for the fantasy of perfect foresight. The value of this approach lies in the handle which the well-understood atemporal general equilibrium theory gives in seeking to understand those aspects of intertemporal economics where mistakes in expectation formation appear unimportant.

The most obvious limitation of perfect foresight models is the absence of uncertainty; but the concept has been extended to allow for uncertainty, in the form of the ‘rational expectations hypothesis’. This allows expectations to take the form of a probability distribution rather than a point, and requires the distribution to be correct. This begs the question of what is meant by a correct probability distribution. In a theoretical model this is conceptually straightforward. Writing down a theoretical model quite naturally generates a probability distribution describing people's beliefs about certain variables, and another describing the actual probability distribution of these variables. In a rational expectations equilibrium these are the same. In simple cases it may be easy to show that a rational expectations equilibrium exists, by solving the equations equating the distributions.

Consider, for an example, a slight generalization of the cobweb model discussed earlier, in which demand $D_t = a - bp_t + \varepsilon_t$ where ε_t is a normal random variable with mean 0 and variance 1. The price p_t is now a random variable; suppliers believe that it is normal with mean $\hat{p}_t^e$ and variance $\hat{\sigma}_t^2$ and want to supply $S = c\hat{p}_t^e - \hat{\sigma}_t^2$. In temporary equilibrium supply equals demand, $a - bp_t + \varepsilon_t = c\hat{p}_t^e - \hat{\sigma}_t^2$ so $p_t = (a - c\hat{p}_t^e - \hat{\sigma}_t^2 + \varepsilon_t) / b$. Given the $N(0, 1)$ distribution of ε_t the price p_t is indeed normally distributed with mean $Ep_t = (a - c\hat{p}_t^e + \hat{\sigma}_t^2) / b$ and variance $1/b^2$. The suppliers have rational expectations if $Ep_t = \hat{p}_t^e$ and $\hat{\sigma}_t^2 = 1 / b^2$ in which case $Ep_t = \hat{p}_t^e = (a + 1 / b^2) / (b + c)$.

In more complex theoretical models the mathematics is more difficult, but the concept is clear enough. But is it plausible? The very name ‘rational expectations equilibrium’ is based on the presumption that this is how rational, optimizing economic agents form expectations. This requires, minimally, that they should, at some point, be able to tell whether their beliefs are correct or not. Apart from examples of the card-choosing or coin-tossing type which have little economic relevance, empirical knowledge of the probability distribution of, for example, a price, depends upon repeated observations of that price. If the probability distribution is stationary, given enough observations of the price, the statistical frequency distribution of past prices reveals the underlying probability distribution. But stationarity is a very strong condition to require. Even if the exogenous random variable (ε_t in the example) is stationary, the distribution of p_t will change as beliefs change. As noted above, we know very little about dynamic adjustment processes outside perfect foresight or rational expectations equilibria.

Knight (1921) uses the term ‘risk’ to describe situations where probabilities can be inferred from data giving the results of repeated observations of similar events, or symmetry arguments (for example coin-tossing). He reserves ‘uncertainty’ for situations concerning unique events where there is no such basis for numerical probability assessments. It is a matter of some philosophical debate whether it is in fact possible to interpret probability numerically in situations which Knight calls uncertainty; subjectivists claim that it is possible, but make no claim that different people will make the same probability assessments. Whatever the outcome of this debate the rational expectations hypothesis is in trouble in situations of Knightian ‘uncertainty’ because there is no single ‘correct’ probability distribution.

Knight argues that economies with risk, but no uncertainty, are essentially identical to economies with perfect foresight, whereas uncertainty (which he claims is all pervasive in business decisions) has a very great effect on the workings of the economy, accounting for imperfect competition and the existence of profit. Risk is unimportant because its effects are nullified by the ability to hedge, to diversify through stock markets, and most importantly because all risks can be perfectly insured. In the light of more recent theory, Knight is clearly wrong, but his argument anticipates recent developments in a fascinating way.

The formalism of the Arrow–Debreu model can be extended to allow for risk and uncertainty, as well as time, by assuming a complete set of contingent futures markets. Commodities are distinguished by the contingencies in which they are available as well as the date. This provides complete insurance. This model has all the properties of the Arrow–Debreu model without risk (existence of equilibrium and Pareto efficiency); thus far Knight's intuition is correct. Knight is also correct in his observation that in practice complete insurance is not available for many contingencies; we do not live in a world of complete markets. His grand theme is that the presence of uncertainty as opposed to risk renders complete insurance impossible; but in his detailed discussion ‘moral hazard’ plays a key role. Moral hazard is due to the incentive insurance gives to take less care to avoid accidents, and explains why complete insurance is rarely available. As Knight points out, it is a very widespread phenomenon; any implicit or explicit contract which allows one of the parties discretion whose exercise cannot be observed by the other is subject to moral hazard. It is, as Knight argues, all-pervasive in business. But it does not require uncertainty in Knight's sense; if there is risk and imperfect information there is moral hazard. The economics of information has been an enormously active area of theoretical research in recent years; considerable progress has been made by formal modelling of situations with imperfect information, giving us a much clearer view of its considerable importance and implication. We know that these make for an economics which is qualitatively quite different from that of the Arrow–Debreu model. We would not have learnt this if theorists had not been willing to make assumptions which cannot be taken literally or completely defended, in order to pursue questions. Quantitative probability, perfect foresight and rational expectations have been crucial tools in developing our understanding of economics.

See Also

- uncertainty and general equilibrium

Bibliography

Bliss, C.J. 1975. *Capital Theory and the Distribution of Income*. Amsterdam: North-Holland.

Bray, M.M. 1983. Convergence to rational expectations equilibrium. In *Individual Forecasting and Aggregate Outcomes*, ed. R. Frydman and E.S. Phelps. Cambridge: Cambridge University Press.

Debreu, G. 1959. *Theory of Value: an axiomatic analysis of economic equilibrium*. Coules Foundation Monograph No. 17. New York: Wiley.

Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53, 995–1045.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Kaldor, N. 1934. A classificatory note on the determinateness of equilibrium. *Review of Economic Studies* 1, 122–36.

Knight, F.H. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

How to cite this article

Bray, Margaret. "perfect foresight." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000057> doi:10.1057/9780230226203.1268

The New Palgrave Dictionary of Economics Online

perfect information

Leonard J. Mirman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Arrow–Debreu model of general equilibrium; common knowledge; competitive equilibrium; complete information; cooperation; cooperative games; cores; Cournot–Nash equilibrium; general equilibrium; partial equilibrium; perfect information; price system; stationary state; tâtonnement; uncertainty

Article

Perfect information is usually thought of as complete knowledge of a person's economic environment. It is clear that nobody in a real economy has perfect knowledge about every aspect of the economy. However it has been argued that perfect knowledge is unnecessary since the price system summarizes all necessary information. Under this line of reasoning the only information that economic agents need are their own tastes and prices. This seems like a very naive argument. However, the real world is more complicated than this argument suggests. Even the prices system itself is not so simple: there are nonlinear prices, for example quantity discounts, as well as different prices for exactly the same commodity. Moreover the economy would function quite differently if the information structure was different, for example if all agents had more knowledge about economic variables. Hence the question arises: how are prices and information used in ideal models of the economy where many very complicated real world relationships have been simplified? In the following discussion the effect of information and the value of prices in conveying and summarizing this information in economic models is described. It appears that in economic models of the economy the ‘information content’ of prices is not as valuable as it appears on the surface. A well-functioning economy needs much more information than is contained in the price system.

In the quest for the effect of information on the economic environment two basic models come to mind. These are the general equilibrium and partial equilibrium models. The remarks in this article will be aimed basically at the Walrasian general equilibrium model without production. However, many of the points dealing with equilibrating prices – and information – can be made about partial equilibrium models as well as general equilibrium models with production.

The Walrasian paradigm envisioned an economy consisting of a large number of agents trading many

goods. Each person, at each point in time, knowing their own tastes and stock of resources (or endowments) decides how much of each good to buy or sell at each possible price, that is, excess demands can be calculated on the basis of each person's environment (tastes and endowments) and the market price. Walras envisioned a steady state or stationary economy. Prices were thought to be generally in equilibrium, known to the consumers or economic agents but with perhaps slight, insignificant fluctuations. In this stable environment the price system regulates the supply to the market and dictates market clearing. In fact each individual agent reacts to the price system which summarizes all necessary information for this agent. Hence if any agent found himself confronted with an equilibrium price vector, then knowledge of only his own tastes and endowments would be sufficient to find the demands which equilibrate the market. However, to actually find the equilibrium price vector requires considerably more information. This difficulty also occurs in a partial equilibrium environment in which the price regulates the market clearing quantity and even the long-run number of producers.
This perception of the Walrasian economy translates into the well-known modern general equilibrium model. In this model there are generally assumed to be a finite number of commodities and a finite number of agents. Each agent is assumed to be a price taker. Equilibrium is found from the market clearing condition on the basis of the aggregate excess demand functions. The price taking behaviour is somewhat unnatural with a finite number of agents since each agent has some market power and therefore would naturally be expected to use strategic behaviour rather than passively taking prices as given. However, there is a very important extension of this model in which there is a continuum of agents. In this extension price taking behaviour is natural since no agent has any market power.
In this economy agents, in deriving their excess demands, need only information about their own tastes and endowments as well as prices. This model is analogous to the Walrasian paradigm described above. However, the analogy does not hold exactly since in the mathematical model there is no historical equilibrium price vector. Hence it is necessary to use an agent outside of the model, for example an auctioneer, to set equilibrium prices. In order to do this, aggregate excess demands must be known to the auctioneer. As a result, although each agent needs to know prices only, any equilibrating mechanism can work only if it has information about all the agents. If no auctioneer is used then it is necessary to design some sort of tâtonnement or groping mechanism to find equilibrating prices. However for such a mechanism to work and to converge to equilibrium prices, it must take account of all agents, In particular, information about excess demands must be available to make the mechanism work.
In order to make clear exactly how information is used in an economic environment, consider a simple economy consisting of two goods and two individuals. Each agent is assumed to take prices as given. For each of these agents only the price is required to describe excess demands while knowledge of both consumers is needed for an equilibrium. Suppose now that agent 2 can have two possible endowments. The first possibility corresponds to a good year while the second corresponds to a lean year. The good year results in high endowments and the lean year results in low endowments. The process of determining the excess demand function for agent 1 remains the same as before; no knowledge of agent 2 is necessary. Agent 2 on the other hand has two possible excess demand functions – one corresponding to the good year, the other to the lean year. These two excess demand functions will in general be quite different, leading to two quite different equilibria. Since he is a price taker no knowledge of agent 1 is needed by agent 2. To find equilibrium prices and allocations, however, the exact characteristics of each agent must be known, no matter what means is chosen to find an equilibrium.
The ideas discussed above can be illuminated by studying various equilibrium concepts and their

informational requirements from the theory of games. In particular the information requirements in the general equilibrium model can be highlighted using the core concept of a cooperative game.

First consider the various notions of information in the game context. A distinction is made between games with perfect information and games with complete information. Perfect information in the game theoretic sense pertains to knowledge of the previous history of the game; that is, for perfect information all previous actions of the agents and equilibrium outcomes of the game are known. The notion of complete information in a game theoretic setting pertains to knowledge about the environment. In the general equilibrium context, complete information means that each agent knows his own taste and endowments as well as the tastes and endowments of all other agents. An even sharper notion of information is used in game theoretic models. This is the notion of common knowledge. Common knowledge implies not only that each agent knows his own environment – complete information – but each agent knows that the other agents know that the first agent has complete information and so on ad infinitum.

To see the importance of the common knowledge requirement in a noncooperative game consider a duopolistic market structure using the Cournot–Nash equilibrium concept. In this model each firm maximizes profits given the behaviour of the other firm. An equilibrium is a pair of outputs which is optimal for each firm given that the other firm is playing its equilibrium strategy (or output). In this model, each firm must know its own and its opponents' payoff function but each firm must also know that the opponent knows this information. This is clearly the case since the opponent's strategy will depend upon whom he thinks he is playing against. Moreover the opponent should know that the first firm knows that the opponent has this information. This chain must be continued indefinitely in order to achieve a Cournot–Nash equilibrium. Clearly for a Cournot–Nash equilibrium to obtain, that is, for the common knowledge requirement to be valid, a great deal of information is required.

Another game theoretic equilibrium concept is the core of an economy. The general equilibrium model is a very natural setting for the cooperative notion of the core. The relationship between the purely game-theoretic idea of the core and the general equilibrium concept using prices again illustrates the importance and role of information in a Walrasian general equilibrium model. The core of a general equilibrium economy is defined as the set of outcomes or allocations which cannot be improved upon by any coalition or group of agents. This means that, for any allocation in the core, no subset of agents can band together, trade among themselves using their own endowments and make each agent as well off and at least one agent better off than with the allocation in the core. The core is a cooperative game with complete information. Since the idea of a core involves coalitional or cooperative behaviour the core and competitive equilibrium are quite different. In particular the price taking assumption is incompatible with cooperative behaviour. Hence it is not surprising that more information seems to be needed to find the set of core allocations. The surprising result is that for economies with a continuum of players the set of core allocations coincide with the set of competitive allocations. The use of a continuum of agents is a natural way to model price taking behaviour since no individual agent has power to affect prices. The notion of a core for large economies involves the use, by each agent, of considerably more information than the competitive economy, and yet for large economies the informational content of both notions is exactly the same. Moreover even for finite economies a similar, although not identical, statement can be made. This result is surprising since the core does not contain any explicit reference to prices. However, the relationship between competitive equilibrium and the core does show that prices are implicitly contained in the idea of a core. The relationship also underlines the fact that more information than

contained in prices is needed to find a general competitive equilibrium.

The discussion thus far has centred on perfect information in a general equilibrium model without uncertainty. Putting uncertainty into the model involves changing the specification of the market structure and the informational flow of the model. It is now necessary to know when the uncertainty is resolved to specify how the market reacts. Moreover, it is also necessary to specify the agent's subjective beliefs about the likelihood of the various states of nature. Although the advent of uncertainty raises many interesting questions about imperfect or incomplete information – for example, moral hazard problems when actions are unobservable or adverse selection problems when information is unobservable – questions remain about perfect information in models with uncertainty. In particular, consider an Arrow–Debreu world under uncertainty. In this model the information requirements are analogous to the requirements in a general equilibrium model under certainty with perfect information. In this economy trading takes place for contingent claims or Arrow–Debreu commodities. More precisely, since each state of the world can be distinguished, trading for commodities occurs for each commodity for each state of the world. This increases considerably the number of markets and the number of trades. However, except for information about which state of the world has occurred there are no extra informational requirements in this model. Each agent, knowing his own tastes and endowments in each state of the world, must know only prices. To actually find equilibrium prices, however, excess demands must be known in each possible state of the world.

Perhaps a more reasonable economy under uncertainty is to allow trading to take place on the basis of expectations or beliefs about the likelihood of the states of the world and not to assume that the state of the world is known after trading occurs, that is, not to allow contingent trades. The informational requirement in this model is quite different than in the Arrow–Debreu model. In this model there is only one market clearing price for each commodity, rather, as in the Arrow–Debreu world, than a price for each commodity in each state of the world. The agents (or auctioneer) need not know which state of the world actually occurred. However, they must know which states are possible. Finally, the equilibrium in this model depends crucially on the subjective beliefs of the agents, whereas in the Arrow–Debreu model subjective beliefs do not affect the equilibrium outcomes.

This difference in market structure and information requirement in these two models leads to a loss in efficiency. In the Arrow–Debreu model equilibrium is always Pareto optimal but in the non-contingent claims model it will, in general, not be Pareto optimal. Non-contingent claims equilibrium will in general be *ex ante* but not *ex post* Pareto optimal. In fact, if the market were to reopen after the realization of the state of the world and trading were allowed to take place, a Pareto optimal Arrow–Debreu equilibrium would result.

See Also

- uncertainty

How to cite this article

Mirman, Leonard J. "perfect information." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000058> doi:10.1057/9780230226203.1269

The New Palgrave Dictionary of Economics Online

period of production

G.O. Orosel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate capital; Böhm-Bawerk, E. von; capital intensity; capital measurement; capital theory; non-substitution theorem; period of production; rate of profit; reswitching; stationary state; time preference

Article

The period of production is, or purports to be, a measure of aggregate capital per head. More specifically, it is a theoretical concept which tries to measure an economy's (heterogeneous) capital stock per head in (homogeneous) units of time.

Necessarily the concept of the period of production is based on an Austrian, or temporal, view of production. In this view production is conceptualized as a sequence of primary inputs on the one hand and a corresponding sequence of consumption outputs on the other. Produced means of production (capital goods) are reduced to dated primary inputs and consumption outputs. This implies that the approach is suited best to the analysis of steady states, where specific properties of capital goods are irrelevant, whereas it will be misleading, in general, if applied to problems of transition or disequilibrium. In particular, this approach is inadequate for business cycle analysis.

Although the temporal view can be traced back to Thünen, Senior, Rae and Jevons, it was Böhm-Bawerk (1889) who made it a cornerstone of his theory. This theory was directed at a fundamental problem of political economy: why is the (net) rate of profit positive? A related problem concerns the measurement of heterogeneous capital goods in homogeneous units which are independent of distribution.

A sketch of Böhm-Bawerk's theory is as follows. According to Böhm-Bawerk the fundamental feature of an economy using capital is that there is a temporal distance, called *period of production* (or *period of investment*), between primary inputs and corresponding consumption outputs. Capital is, in its essence, a fund of means of subsistence which allows for consumption during this period. In a steady state this subsistence fund consists of different ‘layers’ of goods which are distinguished by their respective degree of maturity, such that each period's consumption can be provided by the layer which has just become ready for consumption. A longer period of production is equivalent, in this view, to more capital

per head. Hence the per capita stock of heterogeneous capital goods can be measured in homogeneous units of time. Adding to this (*a*) the technological hypothesis that consumption output per head increases with the period of production, and (*b*) the psychological hypothesis of a positive time preference gives, in a nutshell, Böhm-Bawerk's explanation of the positivity of the rate of profit.

From the beginning, Böhm-Bawerk's theory, and in particular the concept of the period of production, has caused heated debates (involving, among others, J.B. Clark, Irving Fisher, Schumpeter, Wicksell, Hayek, Kaldor and Knight). The contributions to these debates, not all of them to the point, are not reviewed here (see, however, Kaldor, 1937; Weston, 1951). Instead, we will analyse the period of production from a fundamentalist and from a pragmatic point of view. In a *fundamentalist* view the period of production is seen as an important component of the theory sketched above and, therefore, must have properties which make it consistent with this theory. In particular, it must be a technological parameter. In a *pragmatic* view the period of production is just a conventionally measured distance between primary inputs and consumption outputs and need not have any definite properties.

In order to give a more rigorous presentation of the period of production and the problems associated with it, we make the following assumptions. Unless stated otherwise time is measured continuously and it is assumed that primary inputs and consumption outputs can each be measured in homogeneous units. A technique is assumed to be representable by a pair (a, b) of non-negative, continuous functions a: $R \to R_+$ and b: $R \to R_+$ where $a(t)$ is the amount of primary inputs expended at t and $b(t)$ is the amount of consumption outputs delivered at t (note that such a representation where (a, b) is independent of the rate of growth may not be possible for technologies with joint production; cf. the non-substitution theorem). The primary input will be called ‘labour’; ‘per head’ (or ‘per capita’) will mean per unit of labour. It is assumed that

$$\lim_{t \to -\infty} a(t) = \lim_{t \to \infty} a(t) = \lim_{t \to -\infty} b(t) = \lim_{t \to \infty} b(t) = 0,$$

that a and b are not identically zero, that there are constant returns to scale (that is, for any feasible technique (a, b) and any $\lambda > 0$ the technique (λa, λb) is also feasible) and that there exist some real numbers $H<0$, $G>0$ such that the improper Riemann-integrals

$$\int_{-\infty}^{\infty} e^{-\gamma t} a(t) \mathrm{d}t$$

and

$$\int_{-\infty}^{\infty} e^{-\gamma t} b(t)\mathrm{d}t$$

converge for $\gamma \in (H, G)$. The analysis is restricted to steady states with technique (*a*, *b*), a rate of growth $g \in (H, G)$ and a rate of interest $r \in (H, G)$ and to conditions of zero excess profits, implying

$$w\int_{-\infty}^{\infty} e^{-rt} a(t)\mathrm{d}t = \int_{-\infty}^{\infty} e^{-rt} b(t)\mathrm{d}t \quad (1)$$

where *w* is the steady state price of the primary input, henceforth called (real) wage, and the price of the consumption good is set equal to 1. Given a technique (*a*, *b*) and any point of time *s*, let $\lambda(s, t)$ denote the activity level, at *s*, of the techniques which are in 'stage' *t*. For a steady state $\lambda(s, t) = e^{-gt}\lambda(s, 0)$ and total labour inputs at *s* are

$$A(s) = \int_{-\infty}^{\infty} \lambda(s, t)\, a(t)\mathrm{d}t = \lambda(s, 0)\int_{-\infty}^{\infty} e^{-gt} a(t)\mathrm{d}t.$$

Similarly, total consumption outputs at *s* are

$$B(s) = \lambda(s, 0)\int_{-\infty}^{\infty} e^{-gt} b(t)\mathrm{d}t.$$

This implies for per capita consumption $c := B(s)/A(s)$

$$c\int_{-\infty}^{\infty} e^{-gt} a(t)\mathrm{d}t = \int_{-\infty}^{\infty} e^{-gt} b(t)\mathrm{d}t \quad (2)$$

which is dual to (1). For the value of capital k per head the steady state identity $c+gk=w+rk$ implies

$$k = \begin{cases} \dfrac{c-w}{r-g} & \text{for } r \neq g \\ -\dfrac{dc}{dg} = -\dfrac{dw}{dr} & \text{for } r = g \end{cases}. \tag{3}$$

The fundamentalist view

In a fundamentalist view the period of production T must have two properties: first, it must be a technological parameter, that is, for each technique (a, b) and rate of growth g it must be uniquely determined by the associated flows of labour inputs and consumption outputs (which are proportional to $e^{-gt}a(t)$ and $e^{-gt}b(t)$ respectively); second, as a subsistence fund for T periods with per period consumption c the steady state value of capital per head k must be given by $k=cT$. But this leads to an inconsistency: since it implies $T=k/c$ and since in general k/c varies with the rate of interest, T cannot be a technological parameter. Hence the period of production in the fundamentalist sense does not exist. An analogous inconsistency occurs, if one follows Böhm-Bawerk in (wrongly) identifying consumption with wages and therefore postulates $k=wT$ rather than $k=cT$.

Part of the fundamentalist perspective can be rescued if one gives up the idea that the period of production is one-dimensional. This has been shown by Orosel (1979) within the context of a flow input–point output model where time is measured discretely.

The basic idea can be sketched for a stationary state. With time measured discretely a (flow input–point output) technique can be described by one consumption output $\beta(0) > 0$ and corresponding labour inputs $\alpha(t) \geq 0$, t=0, –1, –2, …, where, for some G>0,

$$0 < \sum_{t=-\infty}^{0} (1+\gamma)^{-t}\alpha(t) < \infty$$

for $\gamma \in (-1, G)$. To the sequence of labour inputs $\{\alpha(t)\}_{-\infty}^{0}$ is associated a sequence of wage payments $\{z_1(t) = w\alpha(t)\}_{-\infty}^{0}$; a sequence of simple interest payments on these, that is

$$\left\{z_2(t) = r\sum_{\tau=-\infty}^{t-1} z_1(\tau)\right\}_{-\infty}^{0};$$

a sequence of simple interest payments on $z_2(t)$, that is

$$\left\{z_3(t) = r\sum_{\tau=-\infty}^{t-1} z_2(\tau)\right\}_{-\infty}^{0},$$

and so on, that is

$$z_{i+1}(t) = r\sum_{\tau=-\infty}^{t-1} z_i(t),$$

i=1, 2, … To each sequence $\{z_i(t)\}_{-\infty}^{0}$ we can define a ‘period of production’ T_i as the ‘average distance’ of $\{z_i(t)\}_{-\infty}^{0}$ from output $\beta(0)$, that is, from t=0, by

$$T_i := \frac{\sum_{t=-\infty}^{0}(-t)z_i(t)}{\sum_{t=-\infty}^{0} z_i(t)}.$$

Further, with each period of production T_i we can associate a (per capita) subsistence fund s_i which makes it possible to consume the incomes $\{z_i(t)\}_{-\infty}^{0}$ generated by the technique before the technique generates a consumption output. These funds are given by $s_1=wT_1$ for wages, by $s_2=(rs_1)T_2$ for simple interest on wages, and so on, that is, $s_{i+1}=(rs_i)T_{i+1}$, i=1, 2, …, for simple interest on s_i during the period of production T_{i+1}, associated with these interest incomes. The total per capita subsistence fund is given by

$$s = \sum_{i=1}^{\infty} s_i = wT_1 + \sum_{i=1}^{\infty} (rs_i)T_{i+1},$$

which is a sum of consumption terms (w and rs_i respectively) each of which is multiplied by the associated period of production. It can be shown that for $r \in [0, G)$ all series converge and (i) all T_i are technological parameters; (ii) $k=s$, that is, the value of capital (per head) equals the subsistence fund (per head); (iii)

$$c = w + \sum_{i=1}^{\infty} rs_i,$$

that is, the consumption terms in s add up to per capita consumption. These results can be generalized to steady states with a positive rate of growth (Orosel, 1979).
The periods of production T_i are fundamentalist in the sense that they are technological parameters and that the subsistence fund corresponding to them equals the value of the capital stock. They lead to a consistent reformulation of some of Böhm-Bawerk's main ideas, but they do not give a measure of aggregate capital. In fact, in the 1960s the debates in the theory of capital have made clear that such a measure does not exist.

The pragmatic view

There are three prominent proposals as to how to measure the time interval between primary inputs and consumption outputs. They are associated with the names of (i) Böhm-Bawerk (1889), (ii) Hicks (1939) and von Weizsäcker (1971), and (iii) Dorfman (1959). Although only von Weizsäcker's analysis is directly applicable to steady states with a given (flow input–flow output) technique (a, b), all three proposals can be generalized accordingly. These three (generalized) concepts of the period of production, denoted by T^B, T^H and T^D respectively, are defined as follows (all integrals being improper Riemann-integrals):

$$T^B(g) := \frac{\int_{-\infty}^{\infty} te^{-gt}b(t)\mathrm{d}t}{\int_{-\infty}^{\infty} e^{-gt}b(t)\mathrm{d}t} - \frac{\int_{-\infty}^{\infty} te^{-gt}a(t)\mathrm{d}t}{\int_{-\infty}^{\infty} e^{-gt}a(t)\mathrm{d}t},$$

$$g \in (H, G)$$
(4)

$$T^{\mathrm{H}}(r) := \frac{\int_{-\infty}^{\infty} t e^{-rt} b(t) \mathrm{d}t}{\int_{-\infty}^{\infty} e^{-rt} b(t) \mathrm{d}t} - \frac{\int_{-\infty}^{\infty} t e^{-rt} w a(t) \mathrm{d}t}{\int_{-\infty}^{\infty} e^{-rt} w a(t) \mathrm{d}t},$$

$$r \in (H, G)$$
(5)

$$T^{\mathrm{D}}(g, r) := \frac{k(g, r)}{c(g)}, \; g \in (H, G), \; r \in (H, G)$$
(6)

where $k(g, r)/c(g)$ is, in value terms, the capital–consumption ratio (if, as in Dorfman's analysis, a stationary state is considered, it is also the capital–output ratio). Given our assumptions all integrals are convergent. Definitions (4) and (5) measure the difference between two points of gravity, or mean values of time, associated with outputs and inputs respectively (the densities being

$$e^{-gt} b(t) \; / \int_{-\infty}^{\infty} e^{-gt} b(t)$$

and so on). In (4) the densities applied are given by the respective steady state *quantities*, in (5) they are

given by the steady state *values*. The justification of (6) is less obvious. Dorfman's argument is that, given g and r, k is a constant stock (of value) with a constant outflow c; therefore, the average time a unit of c remains in k is k/c ('bathtub theorem'). Alternatively, (6) can be derived from the postulate that the (per capita) subsistence fund associated with T^{D}, that is, cT^{D}, equals k.

What are the properties of T^{B}, T^{H} and T^{D}, and how are the three concepts related to each other? First, it is interesting, though not shown in the literature, that T^{D} can also be represented as a difference between points of gravity. Without loss of generality, let the level of activity associated with t be e^{-gt}. Then to a point of time t there corresponds a technique $(e^{-gt}a, e^{-gt}b)$ and therefore wages $we^{-gt}a(t)$, profits $r\kappa(t)$ and investments $g\kappa(t)$ where

$$\kappa(t) := \int_{-\infty}^{t} e^{r(t-\tau)}[we^{-g\tau}a(\tau) - e^{-g\tau}b(\tau)]d\tau$$

is the accumulated value of capital, at t, associated with process $(e^{-gt}a, e^{-gt}b)$. Therefore, in a steady state with technique (a, b), growth rate g and interest rate r there is associated to each t an amount $q(t)$ of consumption claims (wages plus profits minus investments)

$$q(t) := e^{-gt}wa(t) + (r-g)\kappa(t). \tag{7}$$

It is possible to prove that these claims sum up to total consumption, that is,

$$\int_{-\infty}^{\infty} q(t)dt = \int_{-\infty}^{\infty} e^{-gt}b(t)dt \tag{8}$$

and that T^{D} is the 'temporal distance' between consumption outputs and consumption claims, that is

$$T^{D} = \frac{\int_{-\infty}^{\infty} te^{-gt}b(t)dt}{\int_{-\infty}^{\infty} e^{-gt}b(t)dt} - \frac{\int_{-\infty}^{\infty} tq(t)dt}{\int_{-\infty}^{\infty} q(t)dt} \tag{9}$$

In (9) T^D has a structure analogous to T^B and T^H. Because of (4), (5), (7) and (9)

$$T^B = T^H = T^D, \text{ for } r = g. \tag{10}$$

Differentiation of (2) gives

$$T^B = -\frac{1}{c}\frac{dc}{dg},$$

of (1)

$$T^H = -\frac{1}{w}\frac{dw}{dr}.$$

Therefore, using (3), $k=cT^B=wT^H=cT^D$ for $r=g$. For $r \neq g$ we have

$$\frac{1}{r-g}\int_g^r c(\gamma)T^B(\gamma)d\gamma = -\frac{1}{r-g}\int_g^r \frac{dc(\gamma)}{d\gamma}d\gamma = \frac{c(g)-w(r)}{r-g} = k(g, r)$$

since $c(r)=w(r)$. Similarly

$$\frac{1}{r-g}\int_g^r w(\rho)T^H(\rho)d\rho = k(g, r).$$

Hence k can be interpreted as an average of subsistence funds of the form cT^B and wT^H respectively. Finally, if for two techniques (a^1, b^1) and (a^2, b^2) *one* of the three periods of production, $T^B(g)$, $T^H(r)$ or

$T^D(g, r)$, is *for all* feasible g and r greater for (a^1, b^1) than for (a^2, b^2), then for these techniques no reswitching or other paradoxa can occur and (a^1, b^1) can be regarded as unambiguously more capital intensive than (a^2, b^2). However, in general the ranking of techniques according to their period(s) of production will depend on the chosen g and r. Therefore, none of the pragmatic concepts of the period of production gives an unambiguous and generally applicable measure of capital intensity. In the light of the so-called reswitching debate this result is to be expected.

Conclusions

The period of production purports to be a measure of capital intensity. Although it is a useful concept for clarifying the relation between capital and time, it is not, and cannot be, a rigorous measure of aggregate capital per head because even in a restricted model with only one primary input and one consumption output such a measure does not exist. As a *fundamentalist* concept the period of production fails because it cannot simultaneously be a technological concept and explain capital as a subsistence fund; as a *pragmatic* concept it fails because it is not possible to rank techniques according to their period of production independently of the rate of growth and the rate of interest. Hence the period of production cannot avoid the inconsistencies (pointed out in the capital controversies of the 1960s) which are associated with the concept of aggregate capital.

See Also

- Austrian economics
- Böhm-Bawerk, Eugen von

Bibliography

Böhm-Bawerk, E. von. 1889. *Kapital und Kapitalzins*. Zweite Abteilung: *Positive Theorie des Kapitals*. Innsbruck: Wagnersche Universitäts-buchhandlung. 4th edn, Jena: Gustav Fischer, 1921. Trans. as *Capital and Interest*, vol. 2: *Positive Theory of Capital*, South Holland, IL: Libertarian Press, 1959.

Dorfman, R. 1959. Waiting and the period of production. *Quarterly Journal of Economics* 73, 351–72.

Hicks, J.R. 1939. *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory*, 2nd edn. Oxford: Clarendon Press.

Kaldor, N. 1937. Annual survey of economic theory: the recent controversy on the theory of capital. *Econometrica* 5, 201–33.

Orosel, G.O. 1979. A reformulation of the Austrian theory of capital and its application to the debate on reswitching and related paradoxes. *Zeitschrift für Nationalökonomie* 29(1–2), 1–31.

Weizsäcker, C.C. von. 1971. *Steady State Capital Theory*. Lecture Notes in Operations Research and Mathematical Systems 54. Berlin: Springer.

Weston, J.F. 1951. Some perspectives on capital theory. *American Economic Review, Papers and Proceedings* 41, 129–44.

How to cite this article

Orosel, G. O. "period of production." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000062> doi:10.1057/9780230226203.1270

The New Palgrave Dictionary of Economics Online

Perlman, Selig (1888 – 1959)

M. Donnelly

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Commons, J. R.; Perlman, S.; socialism; trade unionism

Article

Selig Perlman was born in 1888 at Bialystok, Poland, then a part of Tsarist Russia. His father was a yarn spinner. Perlman grew up in an atmosphere shaped at once by the labour movement, socialism and Zionism. He emigrated to the United States in 1908, and took up studies with John R. Commons at the University of Wisconsin. He received his Ph.D. there in 1915, joined the teaching faculty and became professor of economics in 1927. He collaborated on the four-volume *History of Labor in the United States* compiled by Commons (1918–35), and published *A History of Trade Unionism in the United States* (1922). His most important work, the influential *A Theory of the Labor Movement*, appeared in 1928. Perlman died in 1959.

Perlman's early sympathies were Marxist, but his views were shaken considerably upon his going to America and coming under the influence of Commons. He came to regard the ideas of socialism as essentially the creation of intellectuals, fundamentally at odds with manual workers' own aspirations and experience. Where the labour movement is weak, Perlman argued, it is more susceptible to control by intellectuals; where, on the other hand, political conditions allow it to become strong, the labour movement is better able to outgrow its early ideological trappings and advance to maturity. Late 19th- and early 20th-century America seemed to Perlman the clearest case of a labour movement 'emancipated from the hegemony of intellectual revolutionists' and expressing its own 'philosophy of organic labor'. The key to successful trade unionism, in Perlman's view, was a limited, practical 'job-consciousness', struggling toward collective control of employment opportunities but not otherwise challenging the prerogatives of capitalists.

Selected works

1918–35. (With J.R. Commons et al.) *History of Labor in the United States*. New York: Macmillan.

1922. *A History of Trade Unionism in the United States*. New York: Macmillan.

1928. *A Theory of the Labor Movement*. New York: Macmillan.

How to cite this article

Donnelly, M. "Perlman, Selig (1888–1959)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000064> doi:10.1057/9780230226203.1271

The New Palgrave Dictionary of Economics Online

permanent-income hypothesis

Mark Aguiar and Erik Hurst
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The permanent income hypothesis (PIH) is a theory that links an individual's consumption at any point in time to that individual's total income earned over his or her lifetime. The hypothesis is based on two simple premises: (1) that individuals wish to equate their expected marginal utility of consumption across time and (2) that individuals are able to respond to income changes by saving and dis-saving. In this article we present the intuition and empirical implications of the PIH in several standard contexts.

Keywords

buffer stocks; consumption insurance; Euler equations; impatience; liquidity constraints; marginal utility of consumption; martingales; permanent income hypothesis; precautionary wealth; preferences; retirement; retirement consumption puzzle; uncertainty

Article

The permanent income hypothesis (PIH) is a theory that links an individual's consumption at any point in time to that individual's total income earned over their lifetime.
The PIH is based on two simple premises: (1) that individuals wish to equate their expected marginal utility of consumption across time and (2) that individuals are able to respond to income changes by saving and dis-saving. Because consumers are making their consumption decisions based on lifetime resources, the PIH implies that today's consumption will respond differently to changes in today's income depending on whether the income changes are expected as opposed to unexpected, or temporary as opposed to permanent. The PIH provides a sharp contrast to Keynesian consumption rules, which assume consumers make their consumption decisions based only upon current income.
The major insights of the PIH originated in Friedman (1957). They are closely related to the ideas expressed in Modigliani and Brumberg's (1954) life-cycle hypothesis (see Carroll, 2001, for a summary of Friedman's original work). Since the 1950s there have been many additional theoretical and empirical

contributions. This article presents the intuition and empirical implications of the PIH that have evolved since the 1950s in several standard contexts.

The canonical model

Consider the canonical model in which an individual lives $T+1$ periods and earns y_t in period t=0,…,T. For now, we assume that the income stream is known at time zero. The canonical model assumes that the individual can borrow and lend freely at an interest rate r. The standard model also assumes that the future is discounted at the rate $\beta < 1$ and utility is additively separable across time and additively separable across consumption and leisure. For simplicity, we treat leisure as fixed and treat income as exogenous to the consumer. We revisit these assumptions below. Let $u(c)$ represent the period utility enjoyed from consumption, where $u' > 0$, $u'' < 0$. The consumer's problem is therefore:

$$\max_{\{c_t\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t u(c_t) \tag{1}$$

subject to $\Sigma_{t=0}^{T}(1+r)^{-t}c_t \leq \Sigma_{t=0}^{T}(1+r)^{-t}y_t + A_0$, where A_0 represents initial assets.

A necessary condition for an interior optimal consumption plan is $u'(c_t) = \beta(1+r)u'(c_{t+1})$, for all $0 \leq t \leq T-1$. Therefore, the relationship between consumption in two periods is independent of the relationship between income in those two periods. For example, suppose that individual's discount the future at the rate of interest such that $\beta(1+r) = 1$. With such a restriction on preferences, the individual will consume the same amount each period. Also for simplicity, let $T \to \infty$ and A_0= 0 (and impose the 'no-Ponzi-game' condition $\lim_{t\to\infty} A_t/(1+r)^t \geq 0$). The budget constraint then implies that consumption in each period equals the annuity value of the present discounted value of income, or 'permanent income,' such that:

$$c = r\sum_{t=0}^{\infty}(1+r)^{-t}y_t. \tag{2}$$

Note that consumption is a function only of permanent income, and not how that income is allocated across periods. The ability to borrow and lend is key to the permanent income hypothesis. This allows the individual to transfer income across periods at the rate (1+r). Access to such an asset makes the

present discounted value of income the only relevant constraint on consumption.

The result has a natural implication in a life-cycle model. Suppose individuals work for $S < T$ periods and then retire. Aside from a potential trend due to time discounting, the PIH implies that consumption should not respond to the drop in income at a known period of retirement. Rather, assets built up over the working years are used to finance retirement consumption. Similar examples are plentiful. For example, a teacher on a 9-month salary consumes steadily over 12 months, or a year-end bonus is used for purchases throughout the year. The fact that income is expected to change tomorrow should already be incorporated into today's consumption plan.

In the above model, there was no uncertainty about future income. This is reasonable for predictable changes to income such as retirement or seasonal work, but less useful in understanding consumption's response to unexpected ‘shocks’ such as an unemployment spell or changes in business cycle conditions. We extend the model to the case of uncertainty by assuming that income follows a stochastic process. In particular, let Y_t denote the random variable of income at time t=0, …,T. We continue our assumption that individual's have access to a risk-free bond. Let E_t denote expectations conditional on information as of time t. At any point in time, t, the consumer's problem can be expressed as the following:

$$\max_{\{c_\tau\}_{\tau=t}^{T}} E_t \sum_{\tau=t}^{T} \beta^{\tau-t} u(c_\tau) \tag{3}$$

subject to the period-by-period budget constraint: $A_{t+1} = (1+r)(A_t + Y_t - c_t)$. Notice that (3) differs from (1) in that individuals in (3) are maximizing expected utility. The first-order conditions imply the following ‘Euler equation’:

$$u'(c_t) = \beta(1+r)E_t u'(c_{t+1}). \tag{4}$$

The marginal utility of consumption varies in a predictable way due only to the interest rate and the subjective discount rate. All other movements are unpredictable (with respect to information available prior to time t). Jensen's inequality implies that consumption will be a martingale when $\beta = 1 + r$ only if marginal utility is linear in consumption (that is, quadratic utility). In many standard utility functions, marginal utility is convex, implying that consumption trends upward in expectation when marginal utility is a martingale. Moreover, all else equal, consumption will respond more to unanticipated permanent innovations to income than to transitory innovations.

Empirical tests of the canonical model

Equation (4) states that, aside from r and δ, information known at time t should not affect the change in the marginal utility of consumption between t and $t+1$. Estimating (4) has been the focus of numerous empirical studies, beginning with seminal paper of Hall (1978). Using aggregate data, Hall finds that lagged consumption and lagged income have minimal predictive power for changes in current consumption growth between t and $t+1$. This, by itself, may be interpreted as a victory for the PIH. However, Hall also finds that a lagged index of stock prices does have predictive power for future consumption changes, an apparent violation of (4). Hall's study was followed by a large empirical literature exploiting aggregate consumption data to test whether innovations to consumption are predictable using information available in prior periods. However, a consensus has emerged that aggregation issues undermine the validity of tests using aggregate data.

A large literature has emerged testing (4) using micro data. For example, Attanasio and Weber (1995) and Attanasio and Browning (1995) find support for the PIH using data from the US Consumer Expenditure Survey and the UK Family Expenditure Survey, respectively. Additionally, Shea (1995), Parker (1999), Souleles (1999), Browning and Collado (2001), and Hsieh (2003), among others, have used micro data to examine how consumption responds to anticipated changes in income. These results, however, have been mixed. The conclusion of this literature is that, at least in some instances, consumption responds to predictable changes in income. This excess sensitivity of consumption to predictable income changes has been seen as a violation of the canonical model of the PIH outlined above.

Moving beyond the canonical model

Depending on the context, the ability to freely borrow and lend may be considered too restrictive or not restrictive enough. On the one hand, it rules out state-contingent insurance contracts between consumers. On the other hand, the ability to borrow against future income is often limited in practice due to lack of enforcement. We now briefly describe how the canonical PIH differs from optimal consumption patterns in models with complete insurance markets or models with borrowing constraints.

Perfect insurance in an economy inhabited by agents that enjoy utility as given by (3) implies that individual consumption depends only on aggregate income rather than how that income is distributed across individuals. That is, consumption depends only on aggregate shocks and not on idiosyncratic shocks. This contrasts with the PIH's statement that consumption responds to idiosyncratic permanent income shocks. The difference reflects the limits of the insurance provided by a risk-free bond. However, there is a parallel as noted by Cochrane (1991). The implication that consumption should not respond to idiosyncratic income shocks was formalized and tested by Townsend (1994) using data from Indian villages and Cochrane (1991) using US data. While Townsend rejects perfect risk sharing, he presents evidence that there is significant insurance of idiosyncratic shocks within villages in India. Cochrane rejects perfect insurance in the case of long illness and involuntary job loss, but fails to reject in the case of several other idiosyncratic shocks.

Another alternative to the standard PIH asset market structure is limiting the amount one can borrow against future income. The inability to borrow implies that eq. (4) may not hold. When constrained, a consumer may be forced to adjust consumption in response to a transitory or predictable shock to

income. For example, if an individual receives a temporary income decline, the inability to borrow against future income may necessitate that consumption moves with contemporaneous income. Zeldes (1989) argues that liquidity constraints do bind for a significant fraction of consumers. Moreover, the inability to borrow presents consumers with the risk that a series of negative income shocks may force consumption down to extremely low levels. To mitigate this risk, potentially constrained consumers build up a ‘buffer stock’ of savings. See precautionary saving and precautionary wealth for a discussion of the accumulation of wealth for precautionary reasons.

Life-cycle consumption

While liquidity constraints can explain the empirical fact that consumption is excessively sensitive to changes in predictable income, empirical critiques remain about the ability of individuals to rationally make consumption decisions today based on their expectations of future income realizations. Two of the strongest critiques are that consumption expenditures are hump-shaped over the life cycle (peaking when households are in their mid-forties) and that there is a significant decline in consumption expenditures at the time of retirement. The latter fact has been referred to as the ‘retirement consumption puzzle’ and has been documented and discussed by, among others, Bernheim, Skinner and Weinberg (2001).

The two empirical critiques are related. According to the standard permanent income hypothesis outlined above, individuals should be smoothing their marginal utility of consumption over their lifetimes. Researchers have been trying to modify the PIH so that it matches these two additional empirical facts. For example, Attanasio et al. (1999) find that, if preferences are a function of demographics, the life-cycle profile can be matched. Alternatively, Gourinchas and Parker (2002) find that a model with a properly calibrated income process can match the hump-shaped consumption profile if households are liquidity constrained and sufficiently impatient.

Aguiar and Hurst (2005; 2007) adopt a different approach from those above by appealing to the intuition of Becker (1965). They argue that the PIH theory concerns consumption while the data reports expenditure. The distinction is important because consumption requires time as well as market goods. In particular, households may substitute time for expenditure and maintain a constant level of consumption as expenditures fall. This margin of substitution is suppressed in the canonical form of the model, but Aguiar and Hurst (2005; 2007) document that it is empirically important and reconciles the PIH with both the life-cycle profile of expenditure and the changes in expenditure associated with retirement.

In summary, the current state of literature has expanded on the insights of Friedman's original discussion of the PIH by building in additional features to the canonical model to match a wide variety of empirical regularities. However, this discussion highlights the broader point that any empirical test of the PIH is always a joint test of the hypothesis itself as well as the specific restrictions the researcher places on preferences (for example, whether utility is non-separable between consumption and leisure, the curvature of marginal utility, or the extent to which individuals are impatient), information (for example, assumptions about the income process), or technologies (for example, the existence of liquidity constraints, a home production sector, or complete markets) used to construct the hypothesis’ empirical counterpart.

See Also

- Friedman, Milton
- Modigliani, Franco
- precautionary saving and precautionary wealth

Bibliography

Aguiar, M. and Hurst, E. 2005. Consumption vs. expenditure. *Journal of Political Economy* 113, 919–48.

Aguiar, M. and Hurst, E. 2007. Lifecycle prices and production. *American Economic Review* (forthcoming).

Attanasio, O. and Browning, M. 1995. Consumption over the life cycle and over the business cycle. *American Economic Review* 85, 1118–37.

Attanasio, O.P. and Weber, G. 1995. Is consumption growth consistent with intertemporal optimization? Evidence from the consumer expenditure survey. *Journal of Political Economy* 103, 1121–57.

Attanasio, O., Banks, J., Meghir, C. and Weber, G. 1999. Humps and bumps in lifetime consumption. *Journal of Business and Economic Statistics* 17, 22–35.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–508.

Bernheim, B.D., Skinner, J. and Weinberg, S. 2001. What accounts for the variation in retirement wealth among U.S. households? *American Economic Review* 91, 832–57.

Browning, M. and Collado, D. 2001. The response of expenditures to anticipated income changes: panel data estimates. *American Economic Review* 91, 681–92.

Carroll, C. 2001. A theory of the consumption function, with and without liquidity constraints. *Journal of Economic Perspectives* 15, 23–45.

Cochrane, J. 1991. A simple test of consumption insurance. *Journal of Political Economy* 99, 957–76.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Gourinchas, P.-O. and Parker, J. 2002. Consumption over the life-cycle. *Econometrica* 70, 47–89.

Hall, R. 1978. Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86, 971–87.

Hsieh, C.-T. 2003. Do consumers react to anticipated income changes? Evidence from the Alaska permanent fund. *American Economic Review* 93, 397–405.

Modigliani, F. and Brumberg, R. 1954. Utility analysis and the consumption function: an interpretation of the cross section data. In *Post-Keynesian Economics*, ed. K. Kurihara. New Brunswick, NJ: Rutgers University Press.

Parker, J. 1999. The reaction of household consumption to predictable changes in payroll tax rates. *American Economic Review* 89, 959–73.

Shea, J. 1995. Union contracts and the life-cycle/permanent income hypothesis. *American Economic Review* 85, 186–200.

Souleles, N. 1999. The response of household consumption to income tax refunds. *American Economic Review* 89, 947–58.

Townsend, R. 1994. Risk and insurance in village India. *Econometrica* 62, 539–91.

Zeldes, S. 1989. Consumption and liquidity constraints: an empirical investigation. *Journal of Political Economy* 97, 305–46.

How to cite this article

Aguiar, Mark and Erik Hurst. "permanent-income hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000066> doi:10.1057/9780230226203.1272

The New Palgrave Dictionary of Economics Online

Perron – Frobenius theorem

Hukukane Nikaido
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Brouwer's fixed point theorem; Hawkins-Simon conditions; Leontief system; Perron-Frobenius theorem

Article

A linear transformation mapping (x_1,x_2, …, x_n) to (y_1,y_2, …, y_n) by

$$y_i = \sum_{j=1}^{n} a_{ij} x_j (i = 1, 2, \ldots, n)$$

all of whose coefficients a_{ij} are non-negative has special properties not shared by the general linear transformation. In a matrix form the transformation takes the form

$$x \to y = Ax,$$

where A is the n-dimensional square matrix with elements a_{ij} in the ith row and the jth column, x is the vector having x_j in the jth component and y is the vector y_i in the ith component. The non-negativity of elements of the coefficients matrix A obviously implies that a vector x with all its components non-negative is mapped to a vector y with all its components non-negative in the transformation. This peculiar nature gives rise to special properties of the eigenvalues and associated eigenvectors of the matrix A. Among them, those found and proved by Frobenius (1908; 1909; 1912), also already noticed for a special case by Perron (1907), are the most relevant to linear economic models in which variables are non-negative. The Perron–Frobenius theorem states them in several propositions.

1. (1) A has real non-negative eigenvalues. With the largest $\lambda = \lambda(A)$ of the non-negative eigenvalues is associated an eigenvector x having non-negative components fulfilling

$$\lambda x = Ax.$$

2. (2) The absolute value $|\omega|$ of any eigenvalue ω of A, either real or complex, is bounded by $\lambda(A)$ so that $|\omega| \leq \lambda(A)$.
3. (3) The matrix $\rho I - A$ where I is the identity matrix and ρ is a real number, has an inverse matrix with all its elements non-negative if and only if ρ is larger than $\lambda(A)$.
   Alternative methods of proving the propositions (1), (2) and (3) are available. Some of them are given below.
   Proof of (1). The proof is straightforward for A with all its elements positive. Among all the pairs (θ, y) of a real number θ and a

nonzero vector y having all its components non-negative that fulfil the n inequalities, the ith component of $Ay \geq$ the ith component of $\theta y (i = 1, 2, \ldots, n)$ there is one (λ, x) with λ being the largest of all such θ. Then $Ax = \lambda x$. For otherwise, some components are larger than the corresponding components of λx while the other components of Ax are not less than the corresponding ones of λx. Whence all the components of $A(Ax)$ are larger than the corresponding ones of λ (Ax) so that λ can be further increased to get another pair (θ, Ax), $\theta > \lambda$ fulfilling the inequalities, contrary to the maximum property of λ. Generally A can be approximated from above by $A_\varepsilon = A + \varepsilon T$ where ε is a small positive number and T is the matrix all of whose elements are one. A_ε with all its elements positive has a special pair $(\lambda_\varepsilon, x_\varepsilon)$ satisfying $A_\varepsilon x_\varepsilon = \lambda_\varepsilon x_\varepsilon$ and maximizing θ at λ_ε over all pairs (θ , y) fulfilling the inequalities, the ith component of $A_\varepsilon y \geq$ the ith component of $\theta y (i = 1, 2, \ldots, n)$. Then, for $\varepsilon' \geq \varepsilon$ the ith component of $A_{\varepsilon'} x_\varepsilon \geq$ the ith component of

$$A_\varepsilon x_\varepsilon = \lambda_\varepsilon x_\varepsilon (i = 1, 2, \ldots, n),$$

implying $\lambda_{\varepsilon'} \geq \lambda_\varepsilon$ Whence, λ_ε converges monotonically to a non-negative λ as ε decreases toward zero. The corresponding eigenvector x_ε if so normalized that its components sum up to 1, converges to a nonzero vector having all its components non-negative for a subsequence $\varepsilon(s)$ of positive numbers tending monotonically to zero when $s \to \infty$. Hence $A_{\varepsilon(s)} x_{\varepsilon(s)} = \lambda_{\varepsilon(s)} x_{\varepsilon(s)}$ becomes $Ax = \lambda x$ in the limit when $s \to \infty$. λ is the largest of θ of all pairs (θ, y) fulfilling the inequalities, the ith component of $Ay \geq$ the ith component of θy (i=1,2, …, n). For $\lambda_\varepsilon \geq \theta$ by construction, which becomes $\lambda \geq \theta$ in the limit.
Alternatively, this proposition can be proved by virtue of Brouwer's fixed point theorem as a fixed point of the mapping that transforms each vector x with non-negative components x_i (i=1,2, …, n) adding up to unity to a vector y with components

$$y_i = \left(x_i + \sum_{j=1}^{n} a_{ij} x_j \right) / \left(1 + \sum_{k,j=1}^{n} a_{kj} x_j \right), \ (i = 1, 2, \ldots, n).$$

At a fixed point x^* that is transformed to itself these equations can be rearranged to

$$\lambda x^* = Ax^*, \lambda = \sum_{k,j=1}^{n} a_{kj} x_j^*.$$

Then $\lambda(A)$ is obtained as the largest of λ's of all such fixed points.
Proof of (2). For an eigenvalue ω of A and an associated eigenvector z with components z_j, the equations

$$\omega z_i = \sum_{j=1}^{n} a_{ij} z_j, \ (i = 1, 2, \ldots, n).$$

hold by definition. Then the absolute values of ω, z_i satisfy

$$\sum_{j=1}^{n} a_{ij} |z_j| \geqq |\omega| |z_i|, \ (i = 1, 2, \ldots, n)$$

Whence $\lambda(A) \geq |\omega|$ by the maximum property of $\lambda(A)$. In particular, $\lambda(A)$ is the largest of all non-negative eigenvalues of *A*.

Proof of (3). Necessity. If $\rho I - A$ has an inverse matrix having all its elements non-negative, $p'(\rho I - A)$ has all its elements positive for some vector *p* having all its components positive, where the prime stands for transposition. Then, $Ax = \lambda x, \lambda = \lambda(A)$ with *x* an associated eigenvector having all its components non-negative, becomes, when premultiplied by p', $\rho p' x > \lambda p' x,\ p' x > 0$, which implies $\rho > \lambda$.

Sufficiency. First note that $\lambda(A) \geq \lambda(C)$ for any principal minor matrix *C* of *A*. For, if $\lambda(C) y = Cy$ for an eigenvector *y* with all its components non-negative associated with $\lambda(C)$ the inequalities, the *i*th component of $Az \geq$ the *i*th component of $\lambda(C) z\ (i = 1, 2, \ldots, n)$ hold for the vector *z* augmented from *y* by putting zero in the missing components, so that $\lambda(A) \geq \lambda(C)$ by the maximum property of $\lambda(A)$. *If* $\rho > \lambda = \lambda(A)$, the determinant of $\rho I - A$ must not be zero, for otherwise ρ would be a positive eigenvalue of *A* larger than $\lambda(A)$. Hence $\rho I - A$ is nonsingular and invertible. For any vector *c* with non-negative components $x = (\rho I - A)^{-1} c$ must have all its components non-negative. Otherwise *x* would have some components negative, and an identical, simultaneous renumbering of equations and variables would bring the relation between *x* and *c* to the form

$$\begin{aligned} \rho x_i - \sum_{j=k+1}^{n} a_{ij} x_j &= \sum_{j=1}^{k} a_{ij} x_j + c_j, && (i = k+1, \ldots, n) \\ x_j &\geqq 0, && (j = 1, 2, \ldots, k) \\ x_j &< 0, && (j = k+1, \ldots, n) \end{aligned}$$

which are non-negative on the right side. Whence

$$\begin{aligned} \sum_{j=k+1}^{n} a_{ij} y_j &\geqq \rho y_i, && (i = k+1, \ldots, n) \\ y_j &= -x_j > 0, && (j = k+1, \ldots, n) \end{aligned}$$

so that $\rho > \lambda(A) \geq \lambda(C)$ contrary to the maximum property of $\lambda(C)$ for the principal minor matrix *C* of *A* obtained by deleting the first *k* rows and columns of *A*. This shows that the components of *x* are non-negative, which ensures the non-negativity of all the elements of $(\rho I - A)^{-1}$.

The condition in (3) that $\rho I - A$ has an inverse matrix with all its elements non-negative can be paraphrased as the positivity of all the principal minor determinants of $\rho I - A$ the so-called *Hawkins–Simon conditions*.

The Perron–Frobenius theorem pertains to the possibility of special solutions of linear economic models and to the ‘good behaviour’ of those solutions. The most typical instance of such models is the Leontief system. In a Leontief system consisting of *n* sectors, each of which produces a single good, without joint products, under constant returns to scale, using *n* goods as current input and as capital, let a_{ij} and b_{ij} be the amounts of the *i*th good consumed as input and used as capital, respectively, which are necessary to produce one unit of the *j*th good in the *j*th sector (*i*,*j*=1,2, …, *n*).

Let the levels of sectoral output $x_j(t)$ at time *t* (*j*=1,2, …, *n*) be so determined that net outputs are invested to increase capital. Then

$$x(t) = Ax(t) + B(x(t+1) - x(t)),$$

where *A* and *B* are the input coefficients matrix and the capital coefficients matrix having elements a_{ij} and b_{ij} in the *i*th row and the *j*th column, respectively. A special time path of output $x_i(t)=(1+g)^t x_i$ (*i*=1,2, …, *n*), called a balanced growth path, on which the levels of sectoral output grow at the equal positive rate *g* is generated by an eigenvector *x* with non-negative components x_i associated with the Perron–Frobenius eigenvalue $\lambda(=1/g)$ of the matrix $(I - A)^{-1}B$ having all elements non-negative, provided the system is *productive* enough for *A* to have its Perron–Frobenius eigenvalue less than 1. On the dual side a row eigenvector *p*’ with

non-negative components p_j (j=1,2,…, n) associated with the Perron–Frobenius eigenvalue $\lambda(=1/r)$ of $B(I-A)^{-1}$, equal to that of $(I-A)^{-1}B$, gives a special set of prices determined by

$$p' = p'A + rp'B$$

at which the sectoral rates of profit are equalized to the common rate $r = 1/\lambda$.
In the system of Sraffa (1960) in which input–output correspondences are

$$(A_a, B_a, \ldots, K_a) \rightarrow A(A_b, B_b, \ldots, K_b) \rightarrow B(A_k, B_k, \ldots, K_k) \rightarrow K$$

the standard commodity is constructed by non-negative multipliers q_a, q_b, …, q_k that fulfil

$$(A_a q_a + A_b q_b + \ldots + A_k q_k)(1+R) = Aq_a(B_a q_a + B_b q_b + \ldots + B_k q_k)(1+R) = Bq_b(K_a q_a + K_b q_b + \ldots + K_k q_k)(1+R) = Kq_k.$$

These multipliers are obtained as components of an eigenvector associated with the Perron–Frobenius eigenvalue $\lambda[=1/(1+R)]$ of the matrix

$$\begin{pmatrix} A_a/A & A_b/A \ldots A_k/A \\ B_a/B & B_b/B \ldots B_k/B \\ K_a/K & K_b/K \ldots K_k/K \end{pmatrix}.$$

More specific information is available about the Perron–Frobenius eigenvalue of those matrices having all their elements non-negative in such a way as to be *indecomposable*, in the sense that no identical, simultaneous renumbering of its rows and columns can be put into the form

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

where A_{11} and A_{22} are square submatrices while A_{12} is a rectangular submatrix and 0 is a rectangular submatrix having zero in all its elements.

4. (4) If A is indecomposable, the Perron–Frobenius eigenvalue $\lambda(A)$ is positive, and with it is associated an eigenvector x having all its components positive. Any eigenvector associated with $\lambda(A)$ is a scalar multiple of x.
5. (5) $\lambda(A)$ is a simple root of the characteristic equation. If A has s eigenvalues of moduli equal to $\lambda(A)$, they give all the roots of the equation

$$\omega^s = \lambda(A).$$

6. (6) By an identical, simultaneous renumbering of the rows and columns A can be put in a form

$$\begin{bmatrix} 0 & . & . & . & 0 & A_{15} \\ A_{21} & 0 & & & & 0 \\ 0 & A_{32} & 0 & & & \\ \vdots & & & & & \\ 0 & . & . & 0 & A_{35-1} & 0 \end{bmatrix},$$

where s is the number of eigenvalues of moduli equal to $\lambda(A)$ and $A_{1s}, A_{21}, \ldots, A_{ss-1}$ are rectangular submatrices, while all the other elements are zero.

A standard reference compiling the main results centring around the Perron–Frobenius theorem is Debreu and Herstein (1953).

See Also

- linear models

Bibliography

Debreu, G. and Herstein, I.N. 1953. Nonnegative square matrices. *Econometrica* 21, 597–607.

Frobenius, G. 1908. Über Matrizen aus positiven Elementen I. *Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften*.

Frobenius, G. 1909. Über Matrizen aus positiven Elementen II. *Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften*.

Frobenius, G. 1912. Über Matrizen aus nichtnegativen Elementen. *Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften*.

Perron, O. 1907. Zur Theorie der Matrizen. *Mathematische Annalen* 64, 248–63.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

How to cite this article

Nikaido, Hukukane. "Perron–Frobenius theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000068> doi:10.1057/9780230226203.1273

The New Palgrave Dictionary of Economics Online

Perroux, François (1903 – 1987)

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

domination effect; economic space; inequality; market power; Perroux, F.; poles of development; regional development; structural change; unbalanced economic growth

Article

French economist, best known for his construction of a theoretical system of economic power. He was born in Lyon and his academic career led him to the Sorbonne and from 1955 to 1975 to the Collège de France. A critic of neoclassical economics, Perroux shared some of the concerns of the American institutionalists, but went beyond them by constructing a system of economic analysis – the only one at his time – that rivals conventional equilibrium economics. This system, comprehensive and consistent, is grounded in an all-pervasive ‘domination effect’ that reflects the inequality of economic agents with respect to their economic power. In equilibrium economics the actions of the economic agents are considered coordinated by an adequate amount of equality, leading to mutual concessions that in turn bring about adjustments and the removal of disturbances. Such an approach, according to Perroux, is contradicted by the facts of economic life and fails to reveal the role of economic power in the market. Where conventional economics stresses coordination among equals and their functional interdependence, Perroux sees a relationship of subordination among economic agents, with the latter either dominating or dominated. Just as Schumpeter, who influenced Perroux and about whom he wrote a book, had revealed the dynamics of innovation, so Perroux disclosed the dynamics of inequality. He acknowledged that there were other theories of monopolistic market situations that shared features of his own ideas, and himself introduced Chamberlin's work to French readers, but pointed out that these theories covered only special cases that would be more adequately handled by a general theory such as that developed by him.

Perroux described his domination effect as asymmetrical and irreversible and as not presuming any intention on the part of the dominating agent. Unlike conventional economics, the domination effect does not produce equilibrium but protracted and cumulative changes. It operates at the level of the firm, of the industry and of the national economy. A dominant firm, for example, can integrate its operations

and earn a surplus from increasing sales to and declining purchases from the outside, and from a market position that yields it favourable prices. The surplus adds to the power of the dominant firm by providing it with means for internal financing, for mergers and acquisitions, and for financing or manipulating the demand for its products. At the international level Perroux's domination effect yields new insight into the position of the dominant economy. His theory differs from the theories of imperialism by not requiring an intention on the part of the dominating power.

Perroux's general theory of economic power was developed during the 1940s and 1950s, not long after the new theories of Keynes, input–output analysis, mathematical programming and game theory had been absorbed into mainstream economics. The economics profession was not ready for still another profound change. Thus, Perroux's general theory of economic domination did not upset conventional analysis. However, from his general theory Perroux derived theories of ‘economic space’ and ‘poles of development’, which in turn yielded theories of structural change, unbalanced economic growth and regional development that continue to be widely discussed and applied in regional planning.

Selected works

1950a. Economic space: theory and applications. *Quarterly Journal of Economics* 64(1), 89–104.

1950b. The domination effect and modern economic theory. *Social Research* 17(2), 188–206.

1969. *L’économie du XXe siècle*. 3rd edn. Paris: PUF.

Bibliography

Bocage, D. 1985. *The General Economic Theory of François Perroux*. Lanham, MD: University Press of America.

Perroux, F. 1980. Peregrinations of an economist and the choice of his route. *Banca Nazionale del Lavoro Quarterly Review* 33(June), 147–62.

Uri, P. 1984. Perroux. In *Contemporary Economists in Perspective*, vol. 2, ed. H.W. Spiegel and W.J. Samuels. Greenwich, CT: JAI Press.

How to cite this article

Spiegel, Henry W. "Perroux, François (1903–1987)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000069> doi:10.1057/9780230226203.1274

The New Palgrave Dictionary of Economics Online

personnel economics

Edward Lazear

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Personnel economics is the application of economic and mathematical approaches and econometric and statistical methods to traditional questions in human resources management. Many of the issues studied by personnel economists can be found in traditional textbooks written by organizational behaviour scholars and other human resources specialists. Economists have something new to say about these issues, however, primarily because economics provides a rigorous, and in many cases more straightforward, way to think about these human resources questions than do the more sociological and psychological approaches.

Keywords

agency problems; human resources management; incentive theory; labour economics; personnel economics; prospect theory; risk aversion; stock options; tournament theory

Article

Personnel economics is the application of economic and mathematical approaches and econometric and statistical methods to traditional questions in human resources management. Many of the issues studied by personnel economists can be found in traditional textbooks written by organizational behaviour scholars and other human resources specialists. Economists have something new to say about these issues, however, primarily because economics provides a rigorous, and in many cases more straightforward, way to think about these human resources questions than do the more sociological and psychological approaches. Certain questions, especially those dealing with compensation, turnover and incentives, are inherently economic. Others, like those associated with non-monetary aspects of the job, norms, teamwork, worker empowerment and peer relationships, while seemingly non-economic, are capable of being informed by economic reasoning. Economists have the advantage of knowing how to strip away extraneous detail and focus on the essentials. This allows them to provide precise and reasoned answers that are testable and refutable and thereby follow the scientific method used by the

physical sciences. One drawback of the economic approach, when applied to human resources (and other) issues, is that sometimes its simplifications miss some of the descriptive detail that gives depth and understanding to a situation.

What are the main goals of personnel economics? The primary goal is to provide positive analysis of human resources practices and methods. When do firms choose to use one form of compensation over another? When are teams important? When is job rotation effective? When are certain benefits or stock grants given to workers? The list extends. But in addition to being able to describe what is, personnel economics is more normative than most fields of economics. Perhaps because the subject was taken up by business school economists whose job is to teach managers what to do, personnel economics has not shied away from being somewhat prescriptive. In part, personnel economics is an attempt to look inside the black box. It is an imperialistic attempt by economists to do what Alfred Marshall (1890) said that ‘economists do not do’: Marshall's famous statement that it is not the economist's business to tell the brewer how to brew beer has not been adhered to when it comes to personnel economics. Personnel economists often attempt to do precisely that; namely, to use the tools of economics to understand and sometimes even to guide practitioners and consultants in their trade.

From a practical point of view, personnel economics is important. Labour accounts for approximately 70 per cent of costs and this number has been reasonably stable over time. Changes that affect labour productivity, turnover, or aspects of compensation can have quite dramatic effects on company profits. In one recent example (see Lazear, 2000b), a company altered its method of pay and consequently experienced a 44 per cent increase in productivity in a period of about six months. Such large shifts in productivity are extremely rare and come about mostly with major innovations in technology. Although changes of this magnitude are likely to be unusual even in the realm of personnel economics, the point remains that action on the cost front is likely to involve labour issues because labour is the primary component of cost for most firms.

Personnel, which has become more fashionably known as human resources management, has been around as an academic and practical subject for at least the last 50 years. But personnel economics takes a different view of many of the same questions and issues that are part of standard human resources management. How does personnel economics differ from ‘old-style’ personnel analyses? Primarily, the difference lies in the rigour associated with the economic approach, which is absent from traditional analyses. Personnel economics is, above all, economics. As such, it follows the approach used by economists. This approach is described in Lazear (2000a) and again in Lazear (2000c). Much of the material in the next few paragraphs is taken directly from Lazear (2000c).

First, personnel economics assumes that the worker and firms are rational maximizing agents. Constrained maximization is the basic building block of all theories in personnel economics. Empirical analyses focus on tests of rational, maximizing models. When evidence contradicts a model, the approach of personnel economists is to think more carefully about the nature of the model set-up, rather than to drop the assumption of rationality. The assumption of maximizing rational behaviour in personnel economics is in large part done in order to allow the analyst to express complicated concepts in relatively simple, albeit abstract, terms.

In many respects, this is the main virtue of personnel economics. The typical human resources text eschews generalization, arguing that each situation is different. The economist's approach is the opposite, following the scientific method that places a premium on discovering the underlying general principle.

A second distinguishing feature is that personnel economists focus on equilibrium. Like the physical sciences, almost all theories in personnel economics are consistent with some notion of equilibrium. This differs dramatically from the approaches used in other social sciences, primarily psychology and sociology. Psychologists are interested in individual behaviour and so equilibrium at the market level is not central. But when discussing issues at the level of the firm, especially those that are imbedded in a market context, equilibrium is essential. Personnel economics differs from other approaches to studying personnel in that, as in all branches of economics, there is no free lunch. Firms hire workers in a competitive labour market and cannot simply take advantage of them. Workers cannot be induced to do things that they do not want to do without appropriate compensation, either in the form of money or some other non-monetary reward.

Consider, for example, the provision of incentives. A psychologist might argue that a particular compensation structure offers stronger incentives than another – the best known is Kahneman and Tversky's (1979) prospect theory, which argues that losses impose more disutility than an equivalent gain produces utility. This implies that penalties are more powerful incentive providers than are bonuses – and might suggest that, as a result, firms should adopt the more powerful form of compensation. This ignores the fact that effort is costly and in equilibrium firms that induce more effort must pay higher wages. It is possible that too much effort results because the additional output from the effort may be smaller than the additional amount necessary to compensate the worker for the increased effort.

Third, efficiency is a central concept of personnel economics. Adam Smith's early notion of the invisible hand makes its way into personnel economics. Individuals who maximize their own utility and interact with firms that maximize profits generate behaviour that usually makes both parties better off. When efficiency suffers, say as a result of moral hazard problems that arise in the agency literature, the economist pushes the analysis to another level, asking what actions might firms and/or workers take to alleviate such inefficiency. Taking this further step assists in making better positive predictions and also normative prescriptions for the business student.

In an analogous vein, personnel economists think in terms of substitution, where other human resources specialists do not. For example, most firms have a benefits department that is distinct from the compensation department and compensation is defined specifically to include monetary remuneration only. There is no explicit recognition of trade-offs, and non-economists frequently think in terms of providing some market level of each job attribute rather than thinking in terms of a total package that guarantees some reservation utility.

Some basic theory

Much of the early work in personnel economics was on the theory of compensation. This was a natural outgrowth of the agency literature that dates back to 1950 (see Johnson, 1950, and later Cheung, 1969). The early modern treatments of the agency problem are found in Ross (1973), who lays out the fundamental agency analysis and later Stiglitz (1975) and Bergson (1978). The basic idea in this early work is that the owner and the worker are not the same individual, and so their interests may not be aligned. In particular, the worker wants money, but does not like to put forth effort. The owner wants output, but would rather not pay for it.

The standard agency problem is solved by a piece-rate compensation scheme. In the simple, risk-neutral

worker case, the optimal scheme pays the worker the full value of his output on the margin, setting the piece rate equal to the (net) value of output. Generally coupled with this is a rent-sharing parameter so that

$$\text{Wage} = a + b\,q, \tag{1}$$

where q is output, b=1 and a is set so that the worker is just indifferent between taking the job and taking his next best alternative.

The analysis becomes more complicated, but not fundamentally different, when noise in production and risk aversion are introduced. The most complete early analysis of this is contained in Hölmstrom (1979). The primary result is that there is now a trade-off. Because workers do not like risk, the firm must dampen the relation of wages to q. There is a trade-off between insurance and incentives. In the context of (1), full insurance can be provided by setting b=0, but as a result, the worker has no incentive to put forth. Were b=1, incentives are provided but the worker bears the full risk. The solution, which generally uses a nonlinear compensation scheme, forces the worker to bear some risk and sacrifices effort relative to the risk-neutral case. Another variant on this scheme is presented by Gibbons (1987). Gibbons considers the case where only the worker knows the difficulty of the job and only the worker knows his true action. Under these circumstances, Gibbons shows that workers will restrict output.

Although piece-rate incentive pay characterizes part of the labour market, especially those jobs where output is easily measured, other jobs, perhaps most, do not lend themselves to piece-rate pay. In such cases firms pay salaries, defined as pay based on an input measure, like hours of work, rather than an output measure, like sales. Lazear (1986) describes the factors that lead firms to choose between paying on output or paying on input. But salaries are not fixed and motivation is provided to workers by altering salaries over time based on performance. When absolute output is difficult to observe, workers are ranked, one relative to another, and promotions that are awarded to the better workers serve as motivation. This logic forms the basis of tournament theory, which shows that a well-designed promotion scheme based on rank alone is a perfect substitute for a piece-rate scheme. See Lazear and Rosen (1981).

There are three basic principles of tournament theory. First, prizes are fixed in advance and depend on relative rather than absolute performance. Second, larger spreads in wages at different levels of the hierarchy motivate those at lower levels to put forth more effort. Third, there is an optimal spread. Although a greater spread increases effort, at some point the additional wages necessary to compensate workers for the increased effort is larger than the additional output generated. The formal analysis sets up a problem in which workers maximize

$$\max_{\mu_j} W_1 P + W_2(1 - P) - C(\mu_j) \tag{2}$$

where W_1 is the wage to the winner who gets the promotion, W_2 is the wage to the loser who is not promoted, P is the probability that a worker gets promoted to the high paying job by out-performing his rival, and μ_j is effort, having cost $C(\mu_j)$. The first-order condition to the worker's problem is

$$(W_1 - W_2)\frac{\partial P}{\partial \mu_j} - C'(\mu_j) = 0 \tag{3}$$

A firm takes (3) into account and sets wages, W_1 and W_2, so as to maximize profits subject to paying enough on average to attract workers to the job.

Some implications follow from (3). First, an increase in $W_1 - W_2$ implies a higher equilibrium level of effort, since $C'(\mu_j)$ is increasing in μ. Larger rises associated with promotion increase the equilibrium level of effort. If promotion is valuable, workers work hard to obtain a promotion.

Second, a decrease in $\partial P/\partial \mu_j$ lowers effort. It is straightforward to show that an increase in noise or luck lowers $\partial P/\partial \mu_j$. Volatile industrial environments generally have highly skewed earnings, which serve to offset the tendency for workers to give up when there is too much randomness associated with the promotion decision.

Additional implications follow. Because nepotism reduces the effect of effort on changing the probability of winning, nepotism kills off effort in an organization. Additionally, if too many workers are competing for a given promotion, incentives are weak because effort does not alter the probability of winning very much. This provides a rationale for limiting the competition in a promotion race.

Some workers will never be promoted again and know it, but it may nevertheless be important to keep them motivated. Upward-sloping experience-earnings profiles that result in backloaded compensation provide incentives. Workers are paid less than they are worth in the early years of their job, but more than they are worth when senior. The higher-than-alternative wage that they receive in the latter years keeps them performing on the job because they do not want to lose the (*ex post*) rents associated with satisfactory performance on the current job. To clear the market, they accept lower wages when young so that over their working life, wages add up to their productivity. Unlike the efficiency wage literature (see Shapiro and Stiglitz, 1984; and Akerlof, 1984 for the classic reference) that focuses on how unemployment can emerge, the thrust in personnel economics has been to ask whether less constrained compensation schemes can remove the excess supply of labour.

The theoretical literature in personnel economics is now quite rich. Topics of hiring and firing, the trade-off between money and benefits, evaluation and worker empowerment and delegation of authority are only a few of the topics analysed.

Empirical literature

Theory is most valuable when it provides predictions that can be verified or refuted by real-world experiences. Personnel economics has an array of implications and some have been analysed in the context of data from businesses that try different aspects of personnel and compensation policy.
There are many examples, but only a few are listed here. Most obvious are tests of incentive theory. Compensation variations provide fertile ground on which to examine worker responses to incentives. A number of recent papers have examined piece-rate pay and its implications for worker behaviour, both in terms of incentives and sorting (see Lazear, 2000; Paarsch and Shearer, 1999; Fernie and Metcalf, 1999; and Erikkson and Villeval, 2004). The finding is that a move from hourly pay to piece-rate pay generally increases output and attracts a more productive workforce. The incentive models of the theoretical personnel economics literature are excellent predictors of real behaviour. They do not imply that piece-rate pay is superior to hourly wages. Higher output comes at a cost (higher wages, sometimes lower quality) and the choice of compensation scheme depends on the factors described in the theory, such as measurement costs and quality–quantity trade-offs. Other examples that tie pay to output involve evidence on the nature of executive compensation and formulae that link pay to measures of output. In most cases, earnings of top executives are tied to a measure of team performance instead of, or in addition to, individual performance. The metric is stock or bonuses that are based on earnings (the best known is Jensen and Murphy, 1990).
Stock and stock options have become an important part of compensation for high-level managers and for knowledge workers in general. Stock may provide incentives, but for most workers the incentive effects of stock ownership must be quite small because they own only a small part of the firm and capture a trivial part of the returns to their effort. Some argued that stock ownership, because of its gradual vesting structure, provides incentives to stay on the job (Oyer, 2004; and Oyer and Shaeffer, 2005). Recently, evidence has become available that demonstrates the significant effect of non-vested stock options and certain types of bonus payments in employee retention (Russell, 2005). It is important to point out that the fact that non-vested compensation provides incentive effects does not imply that they should be used. Again, this is part of thinking about equilibrium. Inefficient retention provides benefits to the firm that fall short of worker costs and in equilibrium vanish as firms find that they must pay workers too much when they create excessive retention incentives.
There is also empirical support for the tournament view of labour markets. Larger prize spreads induce more effort; wage structures seem consistent with tournament structures, and workers behave selfishly and fail to cooperate when relative performance pay is too strong (see for example Ehrenberg and Bognanno, 1990; Drago and Garvey, 1998; Erikkson, 1999; Falk and Fehr, 2006; and Knoeber, 1989).
Related to tournaments, other empirical evidence provides support that upward-sloping experience-earnings profiles are used to motivate workers. These studies use the implications of the theory with respect to variations in use of the method across demographic groups and job complexity. More complex jobs with harder-to-measure output must seek forms of incentive pay other than pure piece rates. The evidence suggests that steeper profiles are used in jobs where measurement is less straightforward (Hutchens, 1986; 1987; 1989). Others have pointed out that long-term employment incentives can only be used for workers who are permanently attached to the labour market. Those who have shorter expected employment duration should be more likely to be paid piece rates; those with permanent attachments should be relatively more likely to see upward-sloping experience-earnings profiles. The

evidence supports this claim (see Goldin, 1986).

Another kind of evidence relates to how human resource practices affect productivity. (The best known are a series of papers by Ichniowski and Shaw, for example, Ichniowski, Shaw and Pernoushi, 1997; Ichniowski, Shaw and Boning, 2001; Ichniowski, Shaw and Bartel, 2007). Compensation is only one way that worker productivity can be altered. The actual organization of work can matter. Working in teams, using job rotation, providing training, sharing information and a number of other practices have been shown to have significant effects on worker productivity. Interestingly, the more modern human resource practices are always coupled with some kind of (team) incentive pay (Jensen and Kevin, 1990). Apparently, the practices themselves, without the incentives to use and implement them, do not produce the desired effects on output.

Conclusion

Personnel economics has been among the most active fields in labour economics since the 1980s. There are three reasons. First, the questions it raises are fundamentally important. Labour is the key factor of production and understanding the ways by which labour productivity can be altered is central to the economics of business. Second, there has been an abundance of theoretical insights that are satisfying not only at the intellectual level, but that seem inherently sensible and able to explain the real world. Third, the theories provide specific implications that can be tested: when the analyses are brought into contact with real data, the theories are confirmed.

See Also

- efficiency wages
- experimental labour economics
- labour economics
- labour economics (new perspectives)
- labour market institutions.

*Much of this article is excerpted from a larger overview on personnel economics in R. Gibbons and J. Roberts (eds), **Handbook of Organizational Economics**, Princeton: Princeton University Press, 2008.*

Bibliography

Akerlof, G. 1984. Gift exchange and efficiency wage theory: four views. *American Economic Review* 74, 79–83.

Bergson, A. 1978. Managerial risks and rewards in public enterprises. *Journal of Comparative Economics* 2, 211–25.

Cheung, S.N.S. 1969. *The Theory of Share Tenancy: With Special Application to Asian Agriculture and*

the First Phase of Taiwan Land Reform. Chicago: University of Chicago Press.

Drago, R. and Garvey, G.T. 1998. Incentives for helping on the job: theory and evidence. *Journal of Labor Economics* 16, 1–25.

Ehrenberg, R.G. and Bognanno, M.L. 1990. Do tournaments have incentive effects? *Journal of Political Economy* 98, 1307–24.

Eriksson, T. 1999. Executive compensation and tournament theory: empirical tests on Danish data. *Journal of Labor Economics* 17, 262–80.

Eriksson, T. and Villeval, M.-C. 2004. Other-regarding preferences and performance pay. An experiment on incentives and sorting. Discussion Paper No.1191, Institute for the Study of Labor (IZA), Bonn.

Falk, A. and Fehr, E. 2006. The power and limits of tournament incentives. Mimeo, University of Bonn.

Fernie, S. and Metcalf, D. 1999. It's not what you pay it's the way that you pay it and that's what gets results: jockeys’ pay and performance. *Labour* 13, 385–411.

Goldin, C. 1986. Monitoring costs and occupational segregation by sex: a historical analysis. *Journal of Labor Economics* 4, 1–27.

Gibbons, R. 1987. Piece-rate incentive schemes. *Journal of Labor Economics* 5, 413–29.

Hölmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Hutchens, R.M. 1986. Delayed payment contracts and a firm's propensity to hire older workers. *Journal of Labor Economics* 4, 439–57.

Hutchens, R.M. 1987. A test of Lazear's theory of delayed payment contracts. *Journal of Labor Economics* 5, S153–70.

Hutchens, R.M. 1989. Seniority, wages and productivity: a turbulent decade. *Journal of Economic Perspectives* 3(4), 49–64.

Ichniowski, C., Shaw, K. and Prennushi, G. 1997. The effects of human resource management practices on productivity. *American Economic Review* 86, 291–313.

Ichniowski, C., Shaw, K. and Boning, W. 2001. Opportunity counts: teams and the effectiveness of production incentives. Working Paper No. 8306. Cambridge, MA: NBER.

Ichniowski, C., Shaw, K. and Bartel, A. 2007. How does information technology really affect productivity? Plant-level comparisons of product innovation, process improvement and worker skills. *Quarterly Journal of Economics* (forthcoming).

Jensen, M.C. and Kevin, J.M. 1990. Performance pay and top management incentives. *Journal of Political Economy* 98, 225–64.

Johnson, D.G. 1950. Resource allocation under share contracts. *Journal of Political Economy* 58, 111–23.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Knoeber, C.R. 1989. A real game of chicken: contracts, tournaments, and the production of broilers. *Journal of Law, Economics & Organization* 5, 271–92.

Lazear, E.P. 1986. Salaries and piece rates. *Journal of Business* 59, 405–31.

Lazear, E.P. 2000a. Performance pay and productivity. *American Economic Review* 90, 1346–61.

Lazear, E.P. 2000b. Economic imperialism. *Quarterly Journal of Economics* 115, 99–146.

Lazear, E.P. 2000c. The future of personnel economics. *Economic Journal* 110, F611–F639.

Lazear, E.P. and Rosen, S. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89, 841–64.

Marshall, A.P. 1890. *Principles of Economics*. 9th edn, with annotations by C.W. Guillebaud. London, New York: Macmillan, for the Royal Economic Society, 1961.

Oyer, P. 2004. Why do firms use incentives that have no incentive effects? *Journal of Finance* 59, 1619–49.

Oyer, P. and Schaefer, S. 2005. Why do some firms give stock options to all employees: an empirical examination of alternative theories? *Journal of Financial Economics* 76, 99–133.

Paarsch, H.J. and Shearer, B. 1999. The response of worker effort to piece rates. *Journal of Human Resources* 34, 643–67.

Ross, S.A. 1973. The economic theory of agency: the principal's problem. *American Economic Review* 63, 134–9.

Russell, K. 2005. Deferred compensation and employee retention. Ph.D. thesis, Department of Economics, Stanford University.

Shapiro, C. and Stiglitz, J.E. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Stiglitz, J.E. 1975. Incentives, risk, and information: notes toward a theory of hierarchy. *Bell Journal of Economics and Management Science* 6, 552–79.

How to cite this article

Lazear, Edward. "personnel economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000327> doi:10.1057/9780230226203.1275

The New Palgrave Dictionary of Economics Online

Persons, Warren Milton (1878 – 1937)

George S. Tavlas

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business-cycle measurement; correlation coefficient; equation of exchange; Persons, W. M.; quantity theory of money; time series analysis

Article

Persons was born on 12 March 1878 in West De Pere, Wisconsin; he died on 11 October 1937 in Cambridge, Massachusetts. After studying mathematics and economics at the University of Wisconsin, he taught at several universities, including Harvard where he became, in 1919, the first editor of the *Review of Economics and Statistics*.

Persons' primary contribution was in the application of statistical methods to the analysis and measurement of economic fluctuations. Early on in his career, he was involved in the debate regarding the empirical validity of the quantity theory of money. He introduced the use of the correlation coefficient into the quantity theory literature as a means of testing the relationships among the variables in the equation of exchange (Persons, 1908) and was the first to employ first-differencing in the quantity-theory debate to remove trend from his data (Persons, 1910).

At Harvard, Persons set out to put differing numerical series into a form which comparisons could be made both among the various series and between different points of time in a given series. In this regard, he devised the 'Harvard Barometer' technique of eliminating seasonal and trend influences from time series. Comparisons of the timing of the adjusted series showed systematic differences among them, and led Persons to emphasize the short-run, periodic, nature of business fluctuations. Consequently, in *Forecasting Business Cycles* (1931), he predicted an end to the business downturn then under way by March 1931. His prediction of an early end to the Great Depression, combined with his advocacy of fiscal retrenchment to combat the depression, may have served to deflect the profession from his substantial contribution to the literature on business-cycle measurement.

Selected works

1908. The quantity theory as tested by Kemmerer. *Quarterly Journal of Economics* 22(February), 274–89.

1910. The correlation of economic statistics. *Quarterly Publications of the American Statistical Association* 12(92), 287–322.

1919. *Indices of General Business Conditions*. Boston: Harvard University Committee on Economics.

1928. *The Construction of Index Numbers*. Boston: Houghton Mifflin.

1931. *Forecasting Business Cycles*. New York: John Wiley & Sons.

How to cite this article

Tavlas, George S. "Persons, Warren Milton (1878–1937)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000072> doi:10.1057/9780230226203.1276

The New Palgrave Dictionary of Economics Online

peso problem

Karen K. Lewis

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

If market participants expect a future discrete change in asset fundamentals, then rational forecast errors may be correlated with current information and have a mean different from zero in finite samples. This statement may seem inconsistent with the standard assumption that forecast errors are orthogonal to current information and have a mean of zero. By contrast, this article describes how this phenomenon may be rational using the example of the Mexican peso market in which it was first noted. It then illustrates how the peso problem applies more generally to a wide range of asset prices.

Keywords

efficient markets hypothesis; foreign exchange risk premium; Friedman, M.; German hyperinflation; learning; martingales; peso problem; rational expectations; regime-switching models; risk neutrality; stock price volatility; term premium; white noise

Article

Asset prices are determined by expectations about the paths of future economic variables. Therefore, anticipated discrete changes in the distribution of these variables directly affect asset price behaviour. The ‘peso problem’ focuses upon how asset prices behave when market traders have expectations about infrequent discrete shifts in economic determinants. With these expectations, the discrete switches can induce behaviour in asset prices that apparently contradicts conventional rational expectations assumptions. The fundamental shifts are rare events and typically occur infrequently, even in relatively large samples. As such, the term ‘peso problem’ is interchangeably with the small-sample inference problems arising from these expected events.

The specific currency reference used in the term ‘peso problem’ may seem at odds with its general potential effects on asset prices. The origins of the term therefore deserve further explanation. The phenomenon is called the ‘peso problem’ because it was first noted in the Mexican peso market. The

original source of the term is unknown, though some economists have attributed it to Milton Friedman. The empirical phenomenon was originally mentioned in writing in the dissertation by Rogoff (1977; 1980) and in publication form by Krasker (1980). Based upon evidence from the Mexican peso futures market from June 1974 to June 1976, Rogoff used the relationship between futures contracts and spot contracts to test market efficiency under rational expectations and risk neutrality. He found that the implications of market efficiency were rejected, but that the behaviour of futures contracts could be explained by the market's persistent belief that the Mexican peso might be devalued. Consistent with this explanation, the peso was devalued in August 1976.

The peso problem in the Mexican currency crisis

To illustrate the effects upon asset prices during this period, consider the relationship between the spot and forward rate of a contract for future delivery. If we define S_{t+1} as the logarithm of the future spot rate (dollars per peso) at date t+1 and F_t as the logarithm of the forward rate contracted at date t for delivery at date t+1, the relationship between the two variables may be written

$$S_{t+1} - F_t = r_t + u_{t+1} \tag{1}$$

where r_t is the risk premium, the forecast error on the spot rate is $u_{t+1} \equiv S_{t+1} - E_t S_{t+1}$, and E_t is the expectations operator conditional on information available at time t. Through covered interest parity, the difference between the spot and forward rate also equals the return on holding peso deposits over the same period and converting the proceeds back into dollars at date t + 1. In order to focus on the effect of expectations, the analysis below will ignore the risk premium effect. This assumption is not necessary, however, and much of the literature described below includes models of the risk premium term, r_t.

From April 1954 to August 1976, the spot peso exchange rate was fixed at 0.08 dollars per peso. During this period, which covered over 20 years, the exchange rate was constant. If we use the notation above, therefore, S_{t+1} was equal to a constant, call it S^0. Nevertheless, futures and forward contracts sold at a discount for much of the early 1970s. For example, the year ahead contract on June 1975 and June 1976 futures contracts sold at a discount of 2.6 and 2.7 per cent respectively. Similarly, Mexican peso deposit rates traded higher than dollar deposit rates over this period, implying a forward rate in (1) that was less than the *ex post* spot rate. Therefore, the *ex post* rate of return on holding Mexican peso accounts, $S^0 - F_t$, was systematically positive. Under risk neutrality, this behaviour contradicts the assumption of rational expectations since it implies that the market's forecast errors, $S_{t+1} - E_t S_{t+1}$, were biased and serially correlated.

At the end of this period, on 31 August 1976, the authorities allowed the Mexican peso to float. Subsequently, the peso fell to 0.05 dollars per peso, implying a devaluation of about 46 per cent. If we define the logarithm of the spot rate associated with this level as S^1, the implied forecast error over this

event was $S^1 - F_t \cong -46$ per cent. If one takes account of this large negative observation together with the many small positive observations over the early 1970s the implication is an average forecast error close to zero, which explains the apparent Mexican peso paradox.

Examining how traders with rational expectations would have formed their forecasts helps to define the peso market phenomenon further. Lizondo (1983) postulated that the expected future peso exchange rate could be written as:

$$E_t S_{t+1} = (1 - p_t)S^0 + p_t S^1 \tag{2}$$

where p_t is the market's assessed probability that the authorities will devalue the peso to S^1 during the next period. Therefore, as long as the peso remains fixed at S^0, the forecast error is

$$u_{t+1} = S^0 - E_t S_{t+1} = p_t(S^0 - S^1) \tag{3}$$

Since the Mexican spot rate over the early period was greater than the devalued August 1976 rate, the initial spot rate S^0 was greater than the anticipated rate if devaluation were to occur, S^1. As such, *ex post* forecast errors were systematically positive. The *ex post* bias observed in forecast errors depended upon both the probability of the devaluation, p_t, and the expected size of the fall in the exchange rate, $S^0 - S^1$. On the other hand, for the period when the devaluation occurred, the forecast error was a large negative number, $(1 - p_t)(S^1 - S^0)$.

In a sample with many observations of similar devaluations, forecast errors would be persistently positive with infrequent large negative observations. The frequent small positive forecast errors and the infrequent large negative forecast errors will tend to cancel each other out. Over a sufficiently large sample with enough of the rare events, the forecast errors would roughly sum to zero, as implied by rational expectations. However, the market would appear to make systematic forecast errors between the episodes of discrete changes, even though the forecasts will be unbiased in sufficiently large samples. Even in large samples, therefore, rational forecast errors with a ‘peso problem’ may be serially correlated.

The peso problem in general asset prices

Although first noted in the period of the fixed Mexican peso rate, this phenomenon can be found in any forward-looking asset price when market traders anticipate a discrete change in the distribution of its

economic determinants. A simple example serves to illustrate the peso problem in general. Suppose that agents rationally anticipate a switch in the process of an economic variable from its current process, R^0, to an alternative, R^1. In this case, rational forecasts of asset prices that depend upon this variable include forecasts of the price conditional upon each regime process. Denote the general asset price as S_t to preserve the same notation as above. Then the expected future value of the asset price is:

$$E_t S_{t+1} = (1 - p_t) E_t(S_{t+1} | R^0) + p_t E_t(S_{t+1} | R^1) \quad (2')$$

where p_t is the market's assessed probability conditional upon time t information that the process will switch to process 1; and where $E_t(S_{t+1} | R^i)$ for $i = 0, 1$ is the expected value conditioned upon time t information and upon process i generating the asset's determining variables.

A few examples of peso problem studies serve to illustrate the breadth of its application in diverse settings. Salant and Henderson (1978) considered the effects upon the price of gold from the market's assessed probability that governments might sell their gold holdings in large discrete amounts. In this case, the spot rate S_t represents the price of gold, $E_t(S_{t+1} | R^i)$ are the expected future gold prices conditional upon $i = 0, 1$, no government sales or government sales, respectively, and p_t is the market's assessed probability that the government will sell gold. Flood and Garber (1980) examined the price level effects resulting from anticipated monetary reforms in hyperinflation-era Germany. In this case, the spot rate represents the price level, $E_t(S_{t+1} | R^i)$ are the expected future price levels conditional upon no reform and reform, alternatively, and p_t is the market's assessed probability that the reform will take place. Lewis (1991) evaluated the term structure of US interest rates following the 1979 change in Federal Reserve operating procedures to determine whether the market believed a shift in policy to lower interest rates was possible. In this case, S_t represents the interest rate, $E_t(S_{t+1} | R^1)$ is the expected future interest rates conditional upon on shift to lower rates, and p_t is the marker's assessed probability that this shift will take place. Bates (1991) used option prices to estimate the market's beliefs that the US stock market might crash before October 1987. In this case, S_t represents the stock price, $E_t(S_{t+1} | R^i)$ is the expected future stock prices conditional upon no crash or crash, respectively, and p_t is the market's assessed probability that the crash will occur. Bekaert, Hodrick and Marshall (2001) analysed international term structure returns using expectations of discrete shifts in short-term interest rate regimes. In this case, S_t is the excess return of long bonds over short-term bonds, and R^i refer to different short-term interest rate regimes. Ang, Gu and Hochberg (2007) examine the effects upon long-horizon initial public offering (IPO) returns based upon uncertainty about which performance regime determines a given initial listing. In this case, S_t refers to the abnormal returns and R^i dictate whether they follow

under- or over-performance.

In general, when traders believe a future shift may occur in determinants of asset prices, expectations will have the form given in (2′), as the above examples demonstrate. Now suppose that no change in regime occurs in the sample. Define $(S_{t+1}|R^0)$ as observations drawn from the current regime process. Then, the forecast errors become:

$$u_{t+1} = (S_{t+1}|R^0) - E_t S_{t+1} = [(S_{t+1}|R^0) - E_t(S_{t+1}|R^0)] + p_t[E_t(S_{t+1}|R^0) - E_t(S_{t+1}|R^1)] \quad (3')$$

As long as the process does not change, the first term represents the forecast error conditioned on the current regime and therefore has mean zero. By contrast, the second term captures the effect of an expected switch to process R^1 that does not materialize in the sample. If the expected price conditioned on process R^0 is on average greater, say, than the price conditioned on regime R^1, the mean of the forecast errors within the sample will tend to be positive. Note that, for the Mexican peso example, the conditional expectations are simply constants where $E_t(S_{t+1}|R^i) = S^i$, for $i = 0, 1$, so that eqs (2) and (3) are equivalent to (2′) and (3′) in this case. In general, however, the expectation conditional upon each regime varies over time as new information arrives to the market.

The example in (3′) illustrates the peso problem effects upon realized returns when no switches occur in the sample. Of course, the forecast error will include this event when the switch occurs. If the switches do not occur with sufficient frequency in the sample, however, forecast errors may continue to appear to be biased. Moreover, even with sufficient occurrences of these shifts, the forecast errors may be serially correlated since they weight the difference between the two expected processes, given by the second term on the right-hand side of eq. (3′). When the probabilities or the differences in expectations under the two regimes are serially correlated, these components of the forecast errors are serially correlated as well. In this case, the difference between the spot rate and the forward rate as in (1) will be serially correlated even in the absence of risk premia. This explanation is consistent with the observation in Rogoff (1977) that Mexican peso futures prices before the devaluation did not follow a martingale as they should have by the efficient markets hypothesis.

The peso problem and Bayesian learning

The simple intuition of the Mexican foreign exchange devaluation example casts the peso problem as a problem arising from anticipated *future* shifts in fundamentals. More generally, the peso problem phenomenon has also come to encompass the asset price implications due to uncertainty about *past* discrete changes. To see why the asset price behaviour is similar, consider a simple example. Suppose that market participants believe that the regime may have shifted in some past time period, $\tau < t$. Given priors about the probability of a change, they will then update their assessed probabilities of living in a

new regime as new information arrives. If they learn through Bayesian inference, the forecast errors will depend upon expectations conditioned on each regime process and upon the updated probabilities of being in each regime.

The form of these forecast errors is isomorphic to equation (3′). To illustrate, suppose that in fact the process changed at time τ. In this case, the current regime R^0 is the new regime, and the alternative regime R^1 is the old regime. The probability p_t represents the market's assessed probability that no change took place. Over time, as the market learns the truth, the probability of no change goes to zero and the second component in the forecast error (3′) vanishes. Clearly, these forecast errors converge to mean-zero, white-noise levels even though they may appear biased during the learning process. Similar results hold when the market does not know the parameters of the new distribution but learns them over time. For example, Lewis (1989) relates the US dollar foreign exchange rate behaviour in the early 1980s to the market's uncertainty about whether a past shift to tighter monetary policy took place. Similarly, Timmermann (1993) shows how the learning can help explain the excess volatility in stock markets.

Despite the similarity of expectations based on learning about past discrete changes and on anticipating future discrete changes, their implications for forecast error behaviour in sufficiently large samples can be somewhat different. A once-and-for-all shift in the asset process with subsequent learning will induce forecast errors that are biased and serially correlated over the learning period. However, as the market learns, the probability of the old regime continuing will go to zero and the effect from the second term on the right-hand side of (3) will vanish. Thus, with sufficient observations, forecast errors following learning will behave according to the standard rational expectations assumptions; that is, they will be mean zero and serially uncorrelated. By contrast, with sufficient observations of the discrete shifts in processes, forecast errors arising from anticipated future discrete events will remain serially correlated in general but will be unbiased.

Empirical approaches to the peso problem in asset prices

As this description makes clear, the peso problem is inherently a problem of identifying a low probability event in a given sample. Many researchers simply acknowledge that this small sample problem may be an issue in their results. Other researchers examine the potential for peso problems to explain anomalous asset price behaviour by using different approaches to identify the peso problem in sample.

These approaches can be divided into three main groups. The first group uses a calibrated asset pricing model to consider whether a peso problem explanation can explain a given empirical regularity. For example, Rietz (1988) uses this approach to consider whether the equity premium can be explained by rare adverse events. More recently, Barro (2006) examines the plausibility of this explanation using data over the 20th century.

The second group identifies the peso problem by using dates of known discrete changes in fundamentals to empirically back out expectations from asset prices. This group of studies focuses upon easily observable shifts in fundamentals. Examples include exchange rate realignments (Bertola and Svensson, 1993; Campa and Chiang, 1996; Campa, Chiang, and Refalo, 2002; Mundaca, 2004) and announced

shifts in monetary policy targeting (Lewis, 1991; Hallwood, MacDonald and Marsh, 2000).

The third group analyses the peso problem by directly estimating regime-switching models of fundamentals to explain anomalous behaviour in their asset prices. This approach has the advantage that the fundamentals process can be estimated from the available data and does not require the researcher to take a stand on the timing of the events. As a result, the analysis can be conducted in a wide range of applications where the dating of events is not known a priori. Many different asset prices have been studied using this approach, including floating spot exchange rates (Engel and Hamilton, 1990; Kaminsky, 1993), the equity premium (Cecchetti, Lam and Mark, 1993), the real interest rate (Evans and Lewis, 1995a), the foreign exchange risk premium (Evans and Lewis, 1995b), the term premium (Bekaert, Hodrick and Marshall, 2001), and IPO abnormal returns (Ang, Gu and Hochberg, 2006).

Summary

In summary, as long as agents anticipate occasional discrete changes in the process of economic variables that affect asset prices, and these changes occur infrequently, asset prices contain the potential for the peso problem. If so, then forecast errors will be serially correlated. Furthermore, unless the sample contains many observations of the discrete shifts, forecast errors will appear biased when observed *ex post* even though traders may have rational expectations. Despite this problem, empirical financial studies frequently measure the risk premium as the predictable component of the realized spot rate less the forward rate, described in (1). Therefore, if the ‘peso problem’ is present in the sample, researchers may incorrectly attribute asset price behaviour to anomalies rather than to the market's rational forecasts of discrete events.

See Also

- Bayesian statistics
- currency crises
- financial market anomalies
- finite sample econometrics
- international finance
- rational expectations
- regime switching models

Bibliography

Ang, A., Gu, L. and Hochberg, Y. 2007. Is IPO underperformance a peso problem? *Journal of Financial and Quantitative Analysis* (forthcoming).

Barro, R. 2006. Rare disasters and asset markets in the 20th century. *Quarterly Journal of Economics* 121, 823–66.

Bates, D.S. 1991. The Crash of 87: was it expected? The evidence from options markets. *Journal of Finance* 46, 1009–44.

Bekaert, G., Hodrick, R.J. and Marshall, D.A. 2001. Peso problem explanations for term structure anomalies. *Journal of Monetary Economics* 48, 241–70.

Bertola, G. and Svensson, L.E.O. 1993. Stochastic devaluation risk and the empirical fit of target-zone models. *Review of Economic Studies* 60, 689–712.

Campa, J.M. and Chang, P.H.K. 1996. Arbitrage-based tests of target-zone credibility: Evidence from ERM cross-rate options. *American Economic Review* 86, 726–40.

Campa, J.M., Chang, P.H.K. and Refalo, J.F. 2002. An options-based analysis of emerging market exchange rate expectations: Brazil's real plan, 1994–1999. *Journal of Development Economics* 69, 227–53.

Cecchetti, S.G., Lam, P. and Mark, N.C. 1993. The equity premium and the risk-free rate: Matching the moments. *Journal of Monetary Economics* 31, 21–45.

Engel, C. and Hamilton, J.D. 1990. Long swings in the dollar: are they in the data and do markets know it? *American Economic Review* 80, 689–713.

Evans, M.D.D. and Lewis, K.K. 1995a. Do expected shifts in inflation affect estimates of the long-run Fisher relation? *Journal of Finance* 50, 225–53.

Evans, M.D.D. and Lewis, K.K. 1995b. Do long-term swings in the dollar affect estimates of the risk premia? *Review of Financial Studies* 8, 709–42.

Flood, R. and Garber, P. 1980. An economic theory of monetary reform. *Journal of Political Economy* 88, 24–58.

Hallwood, C.P. MacDonald, R. and Marsh, I.W. 2000. Realignment expectations and the US dollar, 1890–1897: was there a ‘Peso Problem’? *Journal of Monetary Economics* 46, 605–20.

Kaminsky, G. 1993. Is there a peso problem? Evidence from the dollar/pound exchange rate: 1976–1987. *American Economic Review* 83, 450–72.

Krasker, W.S. 1980. The peso problem in testing the efficiency of forward exchange markets. *Journal of Monetary Economics* 6, 269–76.

Lewis, K.K. 1989. Changing beliefs and systematic forecast errors. *American Economic Reviews* 79,

621–36.

Lewis, K.K. 1991. Was there a peso problem in the U.S. interest rate term structure: 1979–1982. *International Economic Review* 32, 159–73.

Lizondo, J.S. 1983. The efficiency of the Mexican peso market. *Journal of International Economics* 14, 69–84.

Mundaca, G. 2004. Modeling probabilities of devaluations. *Economica* 71, 13–37.

Rietz, T.A. 1988. The equity risk premium: A solution. *Journal of Monetary Economics* 22, 117–31.

Rogoff, K. 1977. Rational expectations in the foreign exchange market revisited. Unpublished manuscript, Massachusetts Institute of Technology.

Rogoff, K. 1980. Tests of the martingale model for foreign exchange futures markets. In Essays on expectations and exchange rate volatility. Ph.D. dissertation, Massachusetts Institute of Technology.

Salant, S. and Henderson, D. 1978. Market anticipations of government policies and the price of gold. *Journal of Political Economy* 86, 627–48.

Timmermann, A.G. 1993. How learning in financial markets generates excess volatility and predictability in stock prices. *Quarterly Journal of Economics* 108, 1135–45.

How to cite this article

Lewis, Karen K. "peso problem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000074> doi:10.1057/9780230226203.1277

The New Palgrave Dictionary of Economics Online

Petty, William (1623 – 1687)

Alessandro Roncaglia
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cantillon, R.; classical value theory; demography; economic statistics; expenditure tax; fiscal pressure; inflation tax; just price; labour theory of value; market price; Marx, K. H.; maximum wages; money; natural price; optimal taxation; Petty, W.; political anatomy; political arithmetic; Quesnay, F.; rent; subsistence wage; surplus; velocity of circulation; wealth

Article

Sir William Petty was born on 26 May 1623 in the village of Romsey, Hampshire, and died on 26 December 1687 in London. His life was hectic: son of a clothier, he was a cabin boy on a merchant ship at 13, admitted to the Jesuit college in Caen (France) at 14; after serving in the Royal Navy he sought refuge in the Netherlands (1643) and Paris (1645), where he studied medicine and (with Hobbes) anatomy. He returned to Romsey in 1646 to revive his father's business; became a doctor of medicine in Oxford University in 1648, and, after an impressive academic career, Professor of Anatomy in 1650, but moved immediately – in 1651 – to the Chair of Music at Gresham College, London. He was also appointed chief medical officer to the English army in Ireland in 1651, and was responsible in 1655–8 for the topographical survey of Irish lands destined for English soldiers, from which he himself emerged with a large landed estate. From then until his death, he was engaged in the management of his estate and in endless litigation over titles of property and taxes, constantly travelling between England and Ireland. Petty also managed to participate, in 1660–2, in the founding of the Royal Society (in full: the Royal Society for the Improving of Natural Knowledge) and in furthering its activities. He married Elizabeth Waller in 1667, and had five children by her and at least one illegitimate child.
Only a small part of Petty's written work was published under his own name during his lifetime. The main essays concerned with economic issues were published after his death, soon after the Glorious Revolution of 1688 made the political climate more favourable to the reception of Petty's ideas. The *Verbum Sapienti* and the *Political Arithmetick* were published in 1690, *The Political Anatomy of Ireland* in 1691, and the *Quantulumcumque concerning Money* in 1695, though they were written respectively in 1664, 1676, 1672 and 1682. Among the writings published during Petty's lifetime, the *Natural and*

Political Observations upon the Bills of Mortality appeared in 1662 under the name of John Graunt, a good friend, although it seems certain that Petty authored at least part of it.

This work is generally considered as marking the birth of the science of demography. A collection of Petty's economic writings, containing some unpublished material, appeared in 1899 as *The Economic Writings of Sir William Petty*, edited by Charles Hull. In 1927 and 1928 other unpublished material appeared (*The Petty Papers*, in two volumes; and *The Petty–Southwell Correspondence*), edited by the sixth Marquis of Lansdowne, a descendant of Petty. Unpublished material (known as ‘the Bowood Papers’) is still extant at the Bodleian Library in Oxford. An important item – *A Dialogue on Political Arithmetic* – edited by S. Matsukawa was published in 1977 (Matsukawa, 1977). (For Petty's complete bibliography, see Keynes, 1971; for a bibliography on Petty, see Roncaglia, 1985.)

Petty's contribution to the origins of classical political economy is threefold, involving method, conceptual framework, and analysis. These aspects are interconnected and often implicit in Petty's writings, which specifically refer to policy issues of his time. We will consider the three aspects separately for the sake of clarity, summarizing Petty's ideas on each of them from his several writings.

Petty refers to his method as ‘political arithmetick’, which comprises the following principles:

> To express my self in Terms of Number, Weight or Measure; to use only Arguments of Sense, and to consider only such Causes, as have visible Foundations in Nature; leaving those that depend upon the mutable Minds, Opinions, Appetites and Passions of particular Men, to the Consideration of others. (Petty, 1899, p. 244)

This method recalls Hobbes's *logica sive computatio*, and Bacon's inductive method (to which Petty explicitly refers). It points to a rejection of the then prevailing qualitative approach to science, based on the description of the quality of the sensations associated with physical objects and human events, in favour of the newly rising quantitative–objectivistic approach. The physical sciences were experiencing this shift during the 17th century; the foundation of the Royal Society marked a decisive step in the transition from the old to the new methodology.

In this respect, the commonly held idea that Petty's ‘political arithmetick’ simply marks the origin of modern economic statistics should be rejected. Petty aims at something more than *recording* and *describing* reality ‘in terms of number, weight or measure’. He aims at *expressing* reality in such terms, since this allows him to identify ‘such causes, as have visible foundations in nature’, that is, to identify the laws intrinsic in reality. Petty thus adopts a point of view which was embedded in the new quantitative approach to science. As Galileo expresses it: ‘This great book which is open in front of our eyes – I mean the Universe – … is written in mathematical characters’ (Galilei, 1623, p. 232). According to this point of view reality *contains* natural laws, so that the task of the scientist is to *discover* these natural laws lying beneath the surface of the apparently erratic phenomena experienced by our senses. Petty himself recognizes that as a description of reality political arithmetick is necessarily imperfect; however, his aim is to locate reality's *inner structure*, not the descriptive details.

Petty's methodological contribution to the development of political economy consists precisely in this: he brings the new quantitative method into the political science dealing with the nature and causes of social wealth. As already stressed, this method means more than an intention to *measure* social phenomena: it means a systematic search for the main characteristics of human societies – a fact well

expressed by Petty's other favourite term for the object of his enquiries, ‘political anatomy’.

In the Preface to *The Political Anatomy of Ireland* (1691) Petty recalls Francis Bacon's parallel between the ‘body natural’ and the ‘body politick’. (This parallel has a long tradition indeed, going back to Menenio Agrippa's apology in Ancient Rome; but in Petty's writings it loses its old moral connotation, no longer suggesting the need for diverse social groups to cooperate.) Petty then goes on: ‘as Anatomy is the best foundation of the one [the body natural], so also of the other [the body politick]’, and points to the need of ‘knowing the Symmetry, Fabrick, and Proportion’ of the ‘Body Politick’ (Petty, 1899, p. 129).

There is a clear parallel between the triad ‘Symmetry, Fabrick and Proportion’ of Political Anatomy, and the triad) ‘Number, Weight or Measure’ of Political Arithmetick. The new science, of which Petty claims to be the founder, is thus characterized both by its quantitative nature and by its objectivistic approach. Also, Petty's reference to the Political Body points to the ‘systemic nature’ of the new approach (or, in other terms, to the ‘holistic nature’ which from Petty onwards characterizes classical political economy).

The vision of society as a political body, comparable to the human body, was probably influenced by Petty's medical career, which earned him the Oxford chair in Anatomy (another illustrious example of a doctor–economist is provided by the founder of the Physiocratic school, François Quesnay). Thus the human body–political body comparison constitutes the background to the specification of the conceptual framework of the new science, to which Petty makes an important contribution.

Petty's notion of money, for instance, is specified through a human-body metaphor:

> Money is but the Fat of the Body-politick, whereof too much doth as often hinder its Agility, as too little makes it sick … As Fat lubricates the motion of the Muscles, feeds in want of Victuals, fills up uneven Cavities, and beautifies the Body, so doth Money in the State quicken its Action, feeds from abroad in the time of Dearth at Home; evens accounts by reason of its divisibility. (Petty, 1899, p. 112)

This metaphor shows that Petty perceived the three functions of money: unit of measure, means of exchange, and store of value. It also shows that Petty did not consider money as constituting the wealth of nations. In fact, Petty's notion of wealth is well expressed through another body-politick metaphor (in all likelihood influenced by William Harvey's then recent discovery of the circulation of the blood: ‘the blood and nutritive juices of the Body Politick’ are ‘the product of Husbandry and Manufacture’ (Petty, 1899, p. 28).

In relation to money, we can also recall that Petty clearly perceived the notion of the velocity of circulation (which is measured through reference to the customary payment intervals for taxes, rents, wages; and which is used for estimating the optimal quantity of money required to finance a given volume of income and trade). Furthermore, Petty stresses that banks, in creating paper money, allow society to save on the cost of acquiring the precious metals necessary for ensuring the required monetary circulation.

The idea of the ‘body-politick’ is also relevant for Petty's analysis of the fiscal system, which mainly concerns its impact on the economic development of society. Petty confronts his ideas on the optimal conditions for a system of taxation, considered as a coherent whole, with the chaotic situation then prevailing, and spells out the preconditions for modern fiscal institutions. He also introduces the notion

now known as fiscal pressure, frequently referring in his works to the ratio between the amount of taxation and the level of national income (or of national expenditure: in fact, Petty favoured an expenditure tax over an income tax system). Interestingly, but not uncommonly, the devaluation of the currency (that is, inflation) is considered as a particular kind of tax.

The idea of the ‘body-politick’ implies a connection between the concept of the economic system and the concept of the nation-state. Machiavelli's work most probably influenced Petty in this respect, as well as in the shift from the moral judgement of human actions to the objective analysis of social events. Machiavelli and Petty also share common limits in their notions of the nation-state and the economic system, since they seem not to perceive the productive interrelationships connecting city and countryside, industry and agriculture: productive interrelationships on which Richard Cantillon was to focus attention, and which would constitute the main analytical contribution of Quesnay's *Tableau économique* in the 18th century.

The notion of the surplus is generally regarded as one of Petty's most relevant contributions. Petty expresses the surplus in physical terms, as the amount of product (corn) exceeding the required means of production, and identifies it with rent. In this way Petty avoids the problem of the determination of the profit rate, which in turn involves the problem of relative prices, since relative prices are required for evaluating both capital advances and the net product. (Such problems were to be taken up later by classical economists like Ricardo and Marx, and then, more recently, by Piero Sraffa.) Interestingly, Petty also expresses the surplus in terms of the number of unemployed persons who can be maintained by a group of labourers who are producing the strict necessaries for both groups, workers and non-workers alike: shades of Marx's surplus labour notion? Like the production of services and luxury goods, unemployment thus appears as a particular way of utilizing the surplus. Wages are not included in the surplus, since they correspond to the necessary subsistence of the workers (and Petty, who considers the workers as nothing else but a produced means of production, considered the subsistence wage not as the result of some automatic mechanism, but as an objective to be reached through laws regulating maximum wages).

Petty's strictly analytical contributions to the origins of classical political economy are more limited than his methodological and conceptual contributions, but are nonetheless relevant.

Petty was credited, by Marx and others, with a labour-embodied theory of value. However, the passages usually quoted to support this interpretation are in fact simplifications of a more complex (and less useful) labour-cum-land theory, based on the idea that the price of each commodity depends on the quantities of the various means of production required to obtain it. In particular, the absence of any consideration of profits and the profit rate suggests that Petty's theory of prices must be considered as very primitive. However, it provided a starting point for subsequent developments. Richard Cantillon's posthumously published *Essay* (1755), for example, dwells on the problem of the ‘par’ between labour and land, and this is derived from Petty's attempt to find a way of expressing one of the two ‘originary’ means of production in terms of other, in order to obtain a single magnitude expressing the difficulty of production of any commodity.

But what is especially relevant for all subsequent analyses of price is Petty's distinction between ‘actual’ and ‘natural’ prices or, in other terms, between exchange relationships actually taking place, and theoretical prices which express the most relevant factors influencing current prices. Petty clearly identifies (in the *Dialogue of Diamonds*, first published in 1899) the factual preconditions – the existence of a regular market, namely, of repeated acts of exchange following regular patterns –

necessary for the notion of ‘natural price’ to be meaningful. This is an *objective* notion of natural price, distinct from the notion of the ‘just price’, the determination of which, as a moral rule of behaviour for sellers and buyers, was one of the main purposes of the writers dealing with economic issues for centuries before Petty's time (for example, Pufendorf). Thus, once again, Petty's contribution to the development of classical political economy relates more to concepts and method than to specific analytic propositions. Nevertheless, it is difficult to overestimate his contribution to classical value theory, for which the notion of natural price (as well as the surplus) represents a necessary prerequisite.

Petty's importance for 17th-century culture is undeniable. His search for an ‘objective’ science contributes to the paradigm shift that was taking place at the time. In this regard his part in the creation of the Royal Society went hand in hand with his development of the new science of ‘political arithmetick’. The ‘human body–political body’ comparison provides a much-needed ‘systemic’ background to the emerging objective analysis of economic events. On both levels (‘political arithmetick’ and ‘political anatomy’), his influence on subsequent developments was decisive: his immediate followers (for example, Gregory King and Charles Davenant) definitively established the sciences of demography and economic statistics, while Petty's conceptual framework, adopted by Cantillon, exerted a decisive influence on the development of Quesnay's economic thinking. In this way (as well as through other less direct channels) Petty influenced both Smith and Ricardo, even if they do not refer directly to his writings. Petty's relevance for the development of classical political economy is emphasized by Karl Marx, who considers Petty to be the ‘founder of Classical political economy’. Later economists limit reference to some specific aspect of Petty's ideas: for instance Keynes (1936, pp. 359, 362) quotes with approbation his ideas on the use of public works as a tool of employment policy; and Luigi Einaudi (1941) refers with enthusiasm to Petty's preference for expenditure taxes. However, these aspects, while testifying to Petty's brilliant intelligence, should not obscure what are in fact his main contributions to economic science: the emphasis on the ‘objective’ method, and the establishment of certain key concepts which later became so basic to economic science as to be unconsciously but consistently accepted as part of our scientific background.

See Also

- Davenant, Charles
- political arithmetic

Selected works

1690a. *Verbum Sapienti*. Reprinted in Petty (1899).

1690b. *Political Arithmetick*. Reprinted in Petty (1899).

1691. *The Political Anatomy of Ireland*. Reprinted in Petty (1899).

1695. *Quantulumcumque concerning Money*. Reprinted in Petty (1899).

1899. *Economic Writings*. 2 vols, ed. C. Hull. Reprinted, New York: A.M. Kelley, 1964.

1927. *Papers*. 2 vols, ed. H. Lansdowne. London: Constable.

1928. *The Petty–Southwell Correspondence, 1676–1687*, ed. H. Lansdowne. London: Constable.

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Ed. H. Higgs, London: Macmillan, 1931.

Einaudi, L. 1941. *Saggi sul risparmio e l'imposta*. Turin: Einaudi.

Fitzmaurice, E. 1895. *The Life of Sir William Petty*. London: Murray.

Galilei, G. 1623. *Il Saggiatore*. Reprinted in *Le opere*, vol. 6, ed. A. Favaro. Florence, 1890–1909.

Graunt, J. 1662. *Natural and Political Observations upon the Bills of Mortality*. Reprinted in Petty (1899).

Keynes, G. 1971. *A Bibliography of Sir William Petty, F.R.S. and of Observations on the Bills of Mortality by John Graunt, F.R.S*. Oxford: Clarendon Press.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Matsukawa, S. 1977. Sir William Petty: an unpublished manuscript. *Hitotsubashi Journal of Economics* 17(2), 33–50.

Roncaglia, A. 1985. *Petty: The Origins of Political Economy*. New York: Sharpe.

How to cite this article

Roncaglia, Alessandro. "Petty, William (1623–1687)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000077> doi:10.1057/9780230226203.1278

The New Palgrave Dictionary of Economics Online

pharmaceutical industry

Sara Fisher Ellison

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

agency problems; asymmetric information; entry; industrial organization; innovation; patents; pharmaceutical industry; regulation; research

Article

The pharmaceutical industry comprises firms which manufacture medicines, including vaccines. Many of these firms also perform some or all of the following functions: conducting basic scientific research to identify (patentable) chemical compounds with medicinal properties, developing those chemical compounds into safe, effective, and commercially viable medicines, gaining government approval to sell those medicines, and marketing those medicines to potential consumers and prescribers. This industry has been widely studied by those interested in the analysis of health care systems. However, I focus here on industrial organization research which seeks to explain general economic phenomena by using the pharmaceutical industry as a setting.

Certain salient features of the pharmaceutical industry have made it a popular focus of research in industrial organization. First, asymmetric information and agency problems are present. Second, innovation plays a central role. Third, entry of new products is common. Fourth, data are unusually available. Finally, the industry is regulated along many dimensions.

Economists have used the pharmaceutical industry to study how asymmetric information and agency problems can affect demand for products in a differentiated product setting. Properties of medicines are not always easily verified or understood by consumers. Furthermore, some medicines are available only through a physician's prescription, and consumers may not be the ones making purchase decisions or paying for the medicine once the decision is made. Hellerstein (1998) and Stern and Trajtenberg (1998) study the role of the prescribing physician in the type of medicine dispensed. Ellison et al. (1997) measure price sensitivity of various agents involved in prescribing and dispensing medicines. Berndt, Pinkyck, and Azoulay (2003) study the impact of incomplete product information on the diffusion of medicines after initial release.

Researchers have used the industry to study determinants of innovation, including incentives provided to firms by various patent systems, incentives faced by researchers within a firm, the size of the firm's research effort, the diversity of its research portfolio, and the geographic proximity of other research centres. Work in this area includes Henderson and Cockburn (1996) and Azoulay (2004).

Most pharmaceutical products are initially patent-protected because they are based on the discovery or synthesis of some new chemical compound. When patents expire, then, the potential exists for entry by chemically identical products, or ‘generics’. The large number of similar markets with observable dates of potential entry has proven a boon to researchers studying entry. Caves, Whinston, and Hurwitz (1991) and Scott Morton (1999) identify factors important in generic manufacturers’ decisions to enter a market, and Ellison and Ellison (2000) look for empirical evidence of strategic entry deterrence by incumbent producers. Pervasive entry has also made the industry a natural setting for critiquing how government price indices handle product introductions. Griliches and Cockburn (1995) and Berndt, Griliches, and Rosett (1993) influenced the Boskin Commission report (Boskin et al., 1996), which suggested alternative ways of computing those indices.

Study of vertical relationships is often hampered by the proprietary nature of the transactions between firms. But pharmaceutical wholesale transactions data are often available, enabling studies such as Ellison and Snyder (2001), which tests various theories of buyer size effects.

Past regulation has shaped the industry, and significant effort is expended by the industry to shape future regulation in turn. Ellison and Mullin (2001) demonstrate the effect on the industry of proposed regulatory reform in the early 1990s, while Ellison and Wolfram (2000) provide evidence of actions the industry took to forestall reform. Also, Scott Morton (1997) studies the distortionary effect of government procurement regulations on firms’ pricing decisions. Much of the research on the pharmaceutical industry has focused on the United States, but interesting questions involving international comparisons of regulatory regimes have been addressed by Danzon and Chao (2000), focusing mainly on price differences, and Kyle (2005), focusing on firms’ entry, or ‘launch’ decisions.

Bibliography

Azoulay, P. 2004. Capturing knowledge across and within firm boundaries: evidence from clinical development. *American Economic Review* 94, 1591–612.

Berndt, E., Griliches, Z. and Rosett, J. 1993. Auditing the producer price index: micro evidence from prescription pharmaceutical preparations. *Journal of Business and Economic Statistics* 2, 251–64.

Berndt, E., Pindyck, R. and Azoulay, P. 2003. Consumption externalities and diffusion in pharmaceutical markets: antiulcer drugs. *Journal of Industrial Economics* 51, 243–70.

Boskin, M., Dulberger, E., Gordon, R., Griliches, Z. and Jorgenson, D. 1996. *Toward a More Accurate Measure of the Cost of Living*. Final report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index. Washington, DC: Senate Finance Committee.

Caves, R., Whinston, M. and Hurwitz, M. 1991. Patent expiration, entry and competition in the US pharmaceutical industry: an exploratory analysis. *Brookings Papers on Economic Activity* 1991, 1–48.

Danzon, P. and Chao, L. 2000. Cross-national price differences for pharmaceuticals: how large and why? *Journal of Health Economics* 19, 159–95.

Ellison, G. and Ellison, S. 2000. Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration. Mimeo. Cambridge, MA: MIT.

Ellison, S. and Mullin, W. 2001. Gradual incorporation of information: pharmaceutical stocks and the evolution of President Clinton's health care reform. *Journal of Law and Economics* 44, 89–129.

Ellison, S. and Snyder, C. 2001. Countervailing power in wholesale pharmaceuticals. *Working Paper 01-27*. Cambridge, MA: MIT.

Ellison, S. and Wolfram, C. 2000. Pharmaceutical prices and political activity. *Working Paper 8482*. Cambridge, MA: NBER.

Ellison, S., Cockburn, I., Griliches, Z. and Hausman, J. 1997. Characteristics of demand for pharmaceutical products: an examination of four cephalosporins. *RAND Journal of Economics* 28, 426–46.

Griliches, Z. and Cockburn, I. 1995. Generics and new goods in pharmaceutical price indexes. *American Economic Review* 84, 1213–32.

Hellerstein, J. 1998. Importance of the physician in the generic versus trade-name prescription decision. *RAND Journal of Economics* 29, 108–36.

Henderson, R. and Cockburn, I. 1996. Scale, scope and spillovers: determinants of research productivity in the pharmaceutical industry. *RAND Journal of Economics* 27, 32–59.

Kyle, M. 2005. Pharmaceutical price controls and entry strategies. Mimeo, Duke University.

Scott Morton, F. 1997. The strategic response by pharmaceutical firms to the Medicaid most-favored-customer rules. *RAND Journal of Economics* 28, 269–90.

Scott Morton, F. 1999. Entry decisions in the generic drug industry. *RAND Journal of Economics* 30, 421–40.

Stern, S. and Trajtenberg, M. 1998. *Empirical implications of physician authority in pharmaceutical decisionmaking*. *Working Paper 6851*. Cambridge, MA: NBER.

How to cite this article

Ellison, Sara Fisher. "pharmaceutical industry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000329> doi:10.1057/9780230226203.1279

The New Palgrave Dictionary of Economics Online

Phelps Brown, (Ernest) Henry (1906 – 1994)

Guy Routh

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

inequality; Phelps Brown, E. H.; wage rates

Article

Born in Calne, Wiltshire, on 10 February 1906, Phelps Brown was educated at Taunton School and then at Wadham College, Oxford, where he was a Scholar and gained First Class Honours in Modern History (1927) and in Philosophy, Politics and Economics (1929). He was a Fellow of New College from 1930 to 1947. In 1936 he published *The Framework of the Pricing System*, an orthodox exposition of marginal theory notable for its clarity.

After distinguished war service with the Royal Artillery (which provided material for *The Balloon*, a novel published in 1953), he became the first Professor of the Economics of Labour at the University of London, teaching at the London School of Economics from 1947 until 1968, when he retired as Emeritus Professor. His lecture courses, ‘Applied Economics’ and ‘The Economics of Labour’, were well attended and valued for their incisiveness and lucidity. *A Course in Applied Economics* was published in 1951, frequently reprinted and issued in a second edition, with J. Wiseman, in 1964. The reader was invited to apply economic analysis to practical problems ‘seen in the many-sidedness that calls for more insights than those of the economist alone’. *The Economics of Labor* appeared in 1962 as the first of the Yale University Studies in Comparative Economics. At the LSE, Phelps Brown carried out a series of studies in the tradition of the great British sociologist-statisticians. These were republished in Henry Phelps Brown and Sheila V. Hopkin, *A Perspective of Wages and Prices* (1981). They are characterized by a scrupulous assembly of data from various countries and, in the case of building wages and the price of consumables, extended over seven centuries. A remarkable stability was found in building wage rates, with no sustained change in 500 out of 690 years, and in differentials between craftsmen and labourers, with a failure by supply and demand to overcome ‘the inertia of convention’ (p. 8). An ability to combine history, sociology and statistics to illuminate economics is demonstrated in *The Growth of British Industrial* Relations*: A Study from the Standpoint of 1906–14* (1959), in *A Century of Pay* (1968) and in *The Inequality of Pay* (1977). In this last, a mass of data from

many countries is marshalled and analysed to assess the relative significance of market and sociological factors in determining inequalities of pay. The conclusion is that these differences are ‘better explained by the play of market forces than by that of custom, convention, status, or power’ (p. 325). In *The Origins of Trade Union Power* (1983), however, the emphasis is on socio-psychological forces: ‘because attitudes govern responses, they are among the basic determinants of the course of history•…•. At the last we are left with the paradox of historical understanding, that we can trace past happenings to their causes without thereby gaining the power to predict’ (pp. 300 and 302). Phelps Brown pursued his work on inequality in the wide ranging *Egalitarianism and the Generation of Inequality* (1988).

Phelps Brown served on a number of public bodies: as one of the ‘Three Wise Men’ (the Council on Prices, Productivity and Income) in 1959; on the National Economic Development Council, 1962; on the OECD Working Party on Wages and Labour Mobility, 1963–4; and on the Royal Commission on the Distribution of Income and Wealth, 1974–8. He was awarded a knighthood in 1976, and became a Fellow of the British Academy in 1960. He was President of the Royal Economic Society, 1970–2, and in his presidential address (published in the *Economic Journal*, 1972) presented his credo on the nature and methods of economics, joining other critics who had independently arrived at similar conclusions: training in advanced economics might be actively unhelpful to those concerned with the application of policy, for ‘it is impaired from the first by being built upon assumptions about human behaviour that are plucked from the air’ (p. 3). His remedies were the removal of the traditional boundary between economics and the other social sciences; a clinical commitment to diagnose and prescribe for particular economic ailments, beginning with practice and working back to theory; the study of history as an essential part of economic training; more observation of actual behaviour, ingenuity in devising methods, accumulating facts, seeking connections and significant detail (p. 9). This analysis was further developed in ‘The Radical Reflections of an Applied Economist’ (1980), reinforcing and extending the arguments of 1972.

A detailed obituary is Hancock and Isaac (1998); see as well his own ‘Autobiographical Notes’ (1996).

Selected works

1936. *The Framework of the Pricing System*. London: Chapman & Hall.

1951. *A Course in Applied Economics*. London: Sir Isaac Pitman & Sons.

1959. *The Growth of British Industrial Relations: a study from the standpoint of 1906–14*. London: Macmillan.

1962. *The Economics of Labor*. New Haven and London: Yale University Press.

1964. (With J. Wiseman.) *A Course in Applied Economics*. 2nd edn. London: Sir Isaac Pitman & Sons.

1968. (With M.H. Browne.) *A Century of Pay: The Course of Pay and Production in France, Germany, Sweden, the United Kingdom, and the United States of America, 1860–1960*. London: Macmillan.

1972. The underdevelopment of economics. *Economic Journal* 82, 1–10.

1977. *The Inequality of Pay*. Oxford: Oxford University Press.

1980. The radical reflections of an applied economist. *Banca Nazionale del Lavoro Quarterly Review* 132(March), 3–14.

1981. (With S.V. Hopkins.) *A Perspective of Wages and Prices*. London: Methuen.

1983. *The Origins of Trade Union Power*. Oxford: Clarendon Press.

1988. *Egalitarianism and the Generation of Inequality*. Oxford: Clarendon Press.

1996. Autobiographical notes. *Review of Political Economy* 8, 129–39.

Bibliography

Hancock, K. and Isaac, J.E. 1998. Henry Phelps Brown 1906–1994. *Economic Journal* 108, 757–78.

How to cite this article

Routh, Guy. "Phelps Brown, (Ernest) Henry (1906–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000078> doi:10.1057/9780230226203.1280

The New Palgrave Dictionary of Economics Online

Phelps, Edmund (born 1933)

Gylfi Zoega

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Edmund Phelps is a Nobel Prize winner in economics who has contributed to our understanding of the supply side of the macroeconomy. He showed that there is no stable trade-off between inflation and unemployment. He derived the socially optimal level of saving and the socially optimal level of research into new technologies and showed how technological progress depended on the size of the population and its level of education. In recent years, Phelps has developed models of the equilibrium unemployment rate, what he calls structural unemployment, that can explain the long swings of unemployment as well as differences across countries.

Keywords

asymmetric information; demand shocks; dynamic inefficiency; economic growth in the very long run; education; efficiency wages; endogenous growth theory; expectations; golden rule; hyperbolic preferences; hysteresis; imperfect information; inflation expectations; intertemporal investment; Learning; Lucas, R.; matching; microfoundations; monetary shocks; natural rate of unemployment; natural rate of unemployment; new classical economics; New Keynesian macroeconomics; path dependence; Phelps, E.S.; Phillips curve; productivity growth; rational expectations; stabilization policy; statistical discrimination; structural unemployment; technical change; unemployment–inflation trade-off

Article

Edmund S. Phelps was born in Evanston, Illinois, on 26 July 1933 and grew up in Hasting-on-Hudson, New York. He attended Amherst College as an undergraduate, where he took a second-year economics course, at his father's suggestion, which sparked an interest in economics. After receiving his BA degree in 1955, Phelps started graduate studies at Yale, where he was influenced by, among others, James Tobin, William Fellner, Henry Wallich and Thomas Schelling. He received his Ph.D. from Yale in 1959. After a short spell at the Rand Corporation, Phelps accepted a research post at the Cowles Foundation in

1960. In 1966 he left Yale for the University of Pennsylvania where he stayed until 1969. There followed a year visiting Stanford University and then in 1971 a move to Columbia University, where he was later made McVickar Professor of Political Economy. At Columbia he met his wife, Viviana Montdor Phelps.

Phelps was elected to the National Academy of Sciences (USA) in 1981 and was made a Distinguished Fellow of the American Economic Association in 2000. He is also a former vice-president of the Association, a fellow of the Econometric Society, the American Academy of Arts and Sciences and the New York Academy of Sciences. A Festschrift conference in his honour was held at Columbia University in October 2001 and the volume published by Princeton University Press in 2003 (Aghion et al., 2003). Phelps received the Nobel Prize in economics in 2006 for his work on intertemporal trade-offs in macroeconomics.

The economics of Phelps – a brief outline

During a telephone conversation with journalists in Stockholm, after being told that he had been awarded the Nobel Prize in economics, Phelps described his contribution as that of introducing people to macroeconomics. This was indeed the tenet of the ‘Phelps volume’ of 1970, which contained a selection of path-breaking papers, all providing microeconomic foundations for macroeconomics (Phelps et al, 1970c). By convincing others to follow suit in explaining macroeconomic relationships with models that describe the behaviour of firms, workers and consumers, Phelps began to transform macroeconomics. Moreover, he also contributed to bringing economic theory closer in line with 20th-century economic life by emphasizing imperfect information and imperfect knowledge with its accompanied market failures into macroeconomics. This heralded another transformation of the field.

Since the publication of his well-known paper on the golden rule of accumulation (Phelps, 1961), Phelps has introduced new ways of thinking about such diverse issues as the effect of monetary policy on output and employment; equilibrium unemployment and efficiency wages; the sources of economic growth in the long run; imperfect competition; discrimination in the workplace; and optimal inflation targeting. His work can be divided chronologically into four distinctive phases. In the early to mid-1960s he wrote extensively on growth theory and produced the golden rules of growth and models of technological progress that were genuine precursors to what we now call endogenous growth theory. In the late 1960s and early 1970s his attention turned to the unemployment–inflation trade-off. Phelps showed how an increase in the supply of money would make firms raise output in the short run while in the long run only wages and prices were affected. The rejection of the notion of a stable Phillips curve – providing policymakers with a menu of unemployment/inflation pairs – was one of the most significant achievements in the history of macroeconomic thought, which changed the practice of monetary policy profoundly. It also opened the avenue to research on the optimal design of monetary policy, which essentially became an intertemporal optimization problem (see Phelps, 1967; 1972b), as well as towards studying the determinants of the steady-state equilibrium unemployment rate, which Friedman dubbed the natural rate of unemployment (Friedman, 1968). A third phase in Phelps's research consisted of a reaction to the rational expectations revolution and its challenge to the effectiveness of monetary policy. In the 1970s Phelps and colleagues – mainly at Columbia University – constructed models with rational expectations but also having wage and price contracts, wages and prices set for longer periods than it

takes to change the course of monetary policy (see Phelps and Taylor, 1977; Taylor, 1980; Calvo, 1983). These papers showed that systematic monetary policy was possible in spite of agents having rational expectations. This was followed by a direct attack on rational expectations, when Phelps and Roman Frydman challenged the idea by demonstrating the implications of each agent having a distinct model of the world in mind (Frydman and Phelps, 1983). During the fourth phase, Phelps responded to another challenge, this one that to his natural rate theory presented by the persistent elevation of unemployment in much of the Organisation for Economic Co-operation and Development (OECD) countries in the 1970s, 1980s, and 1990s. He proposed a set of generating general-equilibrium models of equilibrium unemployment (Phelps, 1994) that can explain the long swings in the unemployment rate for a given country as well as differences in average unemployment across countries.

Capital accumulation and endogenous growth

One of Phelps's first influential papers was his ‘The Golden Rule of Accumulation: A Fable for Growthmen’, which was published in the *American Economic Review* in 1961. In the paper he follows in the footsteps of Ramsey (1928) in deriving the golden rule of capital accumulation that maximizes the long-run level of consumption per capita. According to his golden rule, the savings rate should be set equal to the share of capital in national income. Shortly after writing this paper Phelps went on to introduce the notion of dynamic inefficiency, which is characterized by a state where lowering the rate of saving would raise the utility of all generations, both current and future (Phelps, 1965). In contrast, Phelps and Pollack (1968) showed how inefficiently low levels of saving might arise if each generation discounted its own future utility at a lower discount rate than the utility of future generations. This idea later became know as ‘hyperbolic preferences’ and has been applied to the study of diverse phenomena. Phelps also introduced many of the ideas that later became a part of what is now known as endogenous growth theory. He went beyond the neoclassical framework and gave people an explicit role in the generation and adoption of ideas. In his 2006 Nobel Prize Lecture he describes the neoclassical growth model in the following words:

> Neoclassical growth theory was conspicuous in having no people in it. It explained the accumulation and investment of physical capital yet the driving force in that story – increases in knowledge, called ‘technology’ – rains down exogenously, like manna from heaven – and the selection among new technologies is instantaneous, costless and error-free. Nowhere were people required except in the production functions. It would have been better to suppose that machines do all the producing and that people are deployed over the vast range of activities involving management, judgment, insight, intuition and creativity. (Phelps, 2006)

Phelps, knowing that technological progress requires people doing research, explicitly modelled technological progress as a function of the number of workers doing research. For constant exponential growth his model calls for an exponential growth of labour inputs into research, which makes the long-run rate of growth of technology ultimately determined by the growth rate of the population. Two implications followed. First, there is a golden rule level of research effort that maximizes the level of

consumption per capita, similar to the golden rule of investment. In other words, a society can have excessive research in that consumption per capita is lower than it would be if more people were producing and fewer engaged in research – the gains in technology cannot compensate for lost consumption. The other implication is that a larger population provides a larger number of people doing research and hence makes it possible to climb to a higher technology path. The following quotation is revealing:

> One can hardly imagine, I think, how poor we would be today were it not for the rapid population growth of the past to which we owe the enormous number of technological advances enjoyed today… Another instance of external economies is parallel. Our artistic heritage is much like our technology; it is a part of our 'public capital'. If I could re-do the history of the world, halving population size each year from the beginning of time on some random basis, I would not do it for fear of losing Mozart in the process. No improvement of our dirty air and our traffic congestion could compensate me for that! (Phelps, 1968b, pp. 511–2)

The adoption of new technology also requires people. Nelson and Phelps (1966) study the implications of managers needing to have an idea about the expected value (net of costs) of a technological innovation and the probability of a successful adoption. They propose the idea that education helps managers in this regard; education enhances the ability to learn, understand and adopt what others have discovered. Accordingly, economic growth in the long run depends on the level of education, not its change, as confirmed by recent empirical work. The Nelson and Phelps paper also introduces the concept of a technology gap between each country and a technology leader, an idea that has become important in recent work on endogenous growth. The steady-state technology gap is shown by Phelps to be a decreasing function of the level of education and a positive function of the rate of change of leading technology.

Inflation and unemployment

The rejection of a stable inflation–unemployment trade-off is perhaps Phelps's greatest achievement. He did this essentially by bringing expectations into macroeconomic models of inflation and unemployment. By turning expectations into a state variable, reflecting past unemployment/inflation choices, Phelps showed how monetary policy had an intertemporal dimension. By increasing the supply of money and lowering unemployment today inflation is increased, which eventually raises expectations about future inflation and makes the inflation–unemployment trade-off worse – there is no long-run trade-off between unemployment and inflation, contrary to what the economics profession had believed. In Phelps's 1968 paper in the *JPE* he sets himself the task of explaining why an increase in the supply of money has a positive effect on output in the short run, instead of just raising prices and wages (Phelps, 1968a). The paper provides microeconomic foundations for wage setting, introduces the notion of efficiency wages and equilibrium unemployment, and provides a model of the labour market with job search. Each of these contributions opened up paths for others to research.

In that 1968 paper Phelps models the labour market using a search framework where heterogeneous firms and workers are searching for a suitable match and they meet randomly at a rate determined by the

number of unemployed workers searching and the number of vacancies that need to be filled. The frequency of matches is described by a matching function, which makes the paper a forerunner of the matching theory of Diamond, Mortensen and Pissarides. However, as Phelps pointed out later, the existence of unemployment in equilibrium was essentially not dependent on the heterogeneity of workers and labour market search; all that was needed was rising marginal training costs and job heterogeneity that made workers quit their jobs occasionally (see Phelps, 1995). In the model, firms and employees have to make their decisions before learning about the decisions made by others. An expectational disequilibrium is created when a positive monetary shock drives unemployment below its equilibrium level and firms experiencing higher quit rates respond by raising their money wages, thinking that this will raise their relative wages. Here Phelps spearheaded the work on efficiency wages, an idea later developed by Steven Salop, Guillermo Calvo, Carl Shapiro and Joseph Stiglitz. But, to continue the present story, observed wage inflation rises when every firm raises its wages and this is soon reflected in expectations of higher wage inflation which makes each firm raise wages even more, hence further increasing actual wage inflation. The only non-inflationary point is at the equilibrium rate of unemployment where expected wage inflation equal actual wage inflation. The paper has the seeds of a model of an endogenous natural rate because the rate of equilibrium unemployment is shown to be a function of the rate of growth of the labour force – an increase in the rate of growth of the labour force raises the level of the equilibrium unemployment rate due to rising marginal costs of hiring.

In a separate paper, Phelps proposed a parable of an economy in which output is produced on separate islands, each having its own labour market. When wages and prices are set on one island, this is done without the knowledge of what is happening on the other islands. When demand goes up, due to a loose monetary policy, individual producers do not realize that this is happening; instead they think this is at least partly caused by the changed preferences of consumers and hence do not raise wages and prices fully to neutralize any output effects. Only gradually do their expectations about prices and wages adjust, making them raise wages further, thus eliminating the output effects (Phelps, 1970b).

Phelps treats expectations of wages and prices as a state variable affecting output and employment. What matters for output and unemployment is the deviation of actual wages and prices from their expected values. The implication is that a monetary stimulus has only a short-run effect on employment and output; in the long run both are determined by the structure of the economy (that is, non-monetary factors). This is the *natural-rate hypothesis*. The policy implication that follows is that central banks must be concerned about the effects of their actions on inflationary expectations. If they reduce interest rates today, they may stimulate output and employment but at the cost of higher expected inflation – an upward shift of the short-run Phillips curve – which requires higher interest rates in the future, hence lower employment and output. Monetary policy becomes an intertemporal planning problem (see Phelps, 1967; 1972b). This intertemporal dimension of monetary policy is taken quite seriously by independent central banks that target inflation. The intertemporal dimension of policymaking was emphasized by the Nobel committee when explaining its decision to choose Phelps for the prize.

Phelps's treatment of the labour market as plagued by various imperfections and market failures was mirrored in his description of goods markets. Phelps proposed an early model of imperfectly competitive goods markets in a joint paper with Sidney Winter in the Phelps volume of 1970. The basic idea is that consumers have imperfect information about prices and therefore become customers where they believe prices to be lower. However, information about prices gradually spreads between consumers and when a consumer learns about lower prices elsewhere he leaves his present supplier. In this set-up firms treat

their market share as an asset, comparable to their stock of capital and trained employees. The markup decision becomes an intertemporal investment decision; a price increase, while raising current profits, gradually causes customers to drift elsewhere, hence reducing future profits. The implication is that the markups decision is affected by macroeconomic variables such as the rate of interest and the expected rate of growth of sales to each customer. A fall in the rate of interest, as well as a rise in expected sales per customer, would make firms cut markups in order to invest in an expanded market share. Similarly, when firms expect an imminent recession they have an incentive to raise prices so that price inflation precedes recessions. The customer market model later played an important role in the general equilibrium models of the natural rate developed in the 1994 book, *Structural Slumps*.

New Keynesian economics

In the early 1970s Robert Lucas combined Phelps's island parable and the assumption of rational expectations to generate what became known as new classical economics (Lucas, 1972). In these models only unexpected demand shocks affect output and employment and, more controversially, the deviations of these variables from their equilibrium values only persist as long as expectations remain incorrect; hence anticipated stabilization policy is ineffective. Phelps responded – often in collaboration with his colleagues at Columbia, John Taylor and Guillermo Calvo – by showing that a firm's expectational errors could have real effects even in models having rational expectations, where there is no lack of understanding or perception of what other firms were up to, because of staggered wage and price contracts. The objective was to establish microfoundations for the Keynesian prediction that a permanent demand shock causes a persistent slump and that monetary stabilization policy can be effective. The proposed models are based on the simple observation that wages and prices are never adjusted continuously, but by convention they are set periodically and the timing of wage and price changes is staggered across firms. However, money is neutral in the long run in the staggering models and output tends towards an equilibrium level, unemployment towards its equilibrium level. Phelps was the first to express the view that a model combining rational expectations and wages and prices being set at regular intervals could give Keynesian results (Phelps, 1974). This work later became known as New Keynesian economics (see Phelps and Taylor, 1977; Fischer, 1977; Taylor, 1980; Calvo, 1983).

Phelps continued his attacks on new classical economics in the 1980s when, in collaboration with Roman Frydman, he challenged the very notion of rational expectations by expressing scepticism about their relevance when agents’ actions depended not only on their beliefs about aggregate variables but also about other agents’ beliefs. Individual agents, when acting on their understanding of an economic model, may not converge to a rational-expectations equilibrium because they need to continuously re-estimate while other agents are doing exactly the same. Frydman and Phelps (1983) claim that individual rationality does not guarantee the coordination of beliefs that is assumed in a rational expectations equilibrium, and emphasize the need for a model of learning as an integral part of a model of macroeconomic dynamics.

The changing natural rate

In the late 1980s and early 1990s, Phelps's attention turned to explaining the persistent rise of

unemployment in most OECD countries in the previous two decades. While unemployment had been lower in Europe than in the United States in the 1950s and 1960s, European unemployment started its ascent in the 1970s and moved to a higher plateau, where the big continental economies still find themselves. This experience turned out to be a challenge to Phelps's important work on equilibrium unemployment. How come, asked the critics, that unemployment does not revert back to its pre-shock levels? What happened to the natural rate? In spite of Phelps's 1968 *JPE* paper having the seeds of a model of an endogenous natural rate, this model was inadequate when it came to explaining the persistent rise of unemployment from the early 1970s on.

It is a testimony of Phelps's pervasive influence on the theory of unemployment and inflation that initial attempts by others to explain the persistently high unemployment in the 1980s were to a large extent based on ideas taken from his 1972 book, *Inflation Policy and Unemployment Theory*. Here he introduced the concept of ‘hysteresis’ to economics: there is hysteresis when an equilibrium point depends on the path taken by prices and quantities towards the equilibrium. In the labour market context, the level of equilibrium unemployment may depend on the path taken toward it, that is, a temporary recession may have a permanent effect. In the same book, Phelps went on to consider some possible hysteresis channels. He suggested that unemployment might adversely affect the human capital and work habits of those affected and also that hysteresis could arise due to the dynamics of union membership: a recession reduces the number of union members and this makes those remaining push for higher wages, thus preventing employment from recovering. The hysteresis effects of long-term unemployment working through human capital depreciation were emphasized and developed further by, amongst others, Layard, Nickell and Jackman (1991), while Lindbeck and Snower (1989) extended and developed the idea of hysteresis arising from insider–outsider dynamics.

Phelps disagreed with those who believed that hysteresis could explain the failure of unemployment to fall in the 1980s following the steep recessions at the beginning of the decade. He responded with a series of papers that gradually built a general equilibrium model of the determination of equilibrium unemployment – in steady state the natural rate of unemployment. This work culminated in the publication of *Structural Slumps* in 1994. This book has three prototype models that are non-monetary and emphasize the role of various market imperfections in goods and labour markets. The key imperfection in the labour market is asymmetric information, which leads firms to use wages to reduce quitting and shirking. There arises an upward-sloping wage curve in the wage-employment plane that reflects efficiency-wage considerations. Three models described labour demand. The first uses the customer market set-up of Phelps and Winter (1970) where firms set markups so as to maximize the present discounted value of future profits and customers have imperfect information about prices charged by different suppliers. In this model, a rise in the real rate of interest – or a fall in the expected rate of growth of sales per customer – makes firms disinvest in market share by raising the markup of price over marginal cost, which effectively lowers the real demand wage and raises the natural rate of unemployment. In the second model, firms are concerned about quitting because of the cost of training replacements. Managers use wages to deter quits and the hiring decision also becomes an intertemporal investment decision. Higher interest rates and lower expected productivity growth both make firms reduce hiring as well as lowering wages, which raises the quit rate. The third and final set of models has two sectors: a labour-intensive capital goods sector and a capital-intensive consumer goods sector. A rise in real interest rates makes the relative price of the labour intensive capital good fall, which then lowers the price of labour, raising the natural rate of unemployment. Fitoussi and Phelps (1988) – a precursor to

Structural Slumps – provide a monetary exposition of some of these effects.

In *Structural Slumps*, as well as in the papers that followed, the elevation of unemployment in the OECD countries in recent decades – particularly on the Continent of Europe – is explained by the simultaneous fall in the rate of productivity growth and the rise of world real interest rates in the early 1980s (see Hoon and Phelps, 1997; Phelps and Zoega, 1998; Fitoussi, Jestaz, Zoega and Phelps, 2000). Europe did well in the first two decades following the Second World War because of a combination of low world interest rates and higher productivity growth. Productivity grew at a brisk pace because Europe could imitate the United States – adopt technology that had been developed there in the pre-war years – and meanwhile enjoy high productivity growth in spite of the economic models of the large Continental economies, which otherwise stifled entrepreneurship, initiative and innovation. The closing of the gap and a simultaneous rise in world real interest rates caused a structural slump that monetary policy could not remedy. Analogously, the non-inflationary boom in the United States at the end of the century can be explained by the effect of an anticipated productivity increase in the labour demand wage.

With the passing of time the view that long swings of unemployment require a theory of a changing natural rate of unemployment has gained acceptance. The current debate is focused on the importance of labour market institutions per se and macroeconomic shocks in determining the natural rate of unemployment. Recent work by Phelps has described the adverse effects of Europe's economic model on entrepreneurship, innovation and growth, stemming from its culture as well as the institutions of financial and labour markets, which foster rent seeking and protect vested interests instead of promoting initiative, risk taking and innovation.

Phelps has been interested not only in the macroeconomic causes of unemployment; his interests also extend to the fate of the disadvantaged in modern societies. In the 1990s he wrote a book titled *Rewarding Work* (Phelps, 1997) on the problems facing low-skilled workers. This book emphasizes the importance of having a stable job for self-realization, mental stimulation, lending a rhythm to daily life and participation in society as well as income to support one's family and to share in the consumption and leisure activities of others. He describes the worsening of job prospects for the lowest-skilled American workers and proposes a scheme of general subsidies for the lowest paid. The book demonstrates a genuine commitment to help improve society, as do his frequent articles in the *Financial Times* and the *Wall Street Journal*.

Other contributions

The literature on statistical discrimination originates with Phelps (1972a) and Arrow (1972; 1973). Again we start with asymmetric information, in this case about an individual worker's productivity. Given a statistical correlation between a worker's group attributes and average productivity in the group, an employer may wish to discriminate on the basis of which group the worker belongs to. Unequal treatment of identically productive workers may give a result that does not depend in any way on the employer's preferences or prejudice.

In the field of public finance, Phelps (1973a) found that inflation, being a source of tax revenue, should be chosen optimally along with other forms of taxation. A positive rate of inflation is required to minimize the distortions from different forms of taxation. Finally, there is the ‘Phelps–Sadka result’, namely, that the marginal tax rate should approach zero at the top of the income distribution because policymakers can observe only wage incomes, not wage rates per hour (Phelps, 1973b).

Last but not least, one should mention his *Seven Schools of Macroeconomic Thought* that offers a very personal description of the genesis and distinctive characteristics of the macroeconomics of Keynes, monetarism, the New Classical School, the New Keynesian School, supply-side economics, real business cycle theory, and what he called the Structuralist School, which includes the work done on endogenizing the natural rate of unemployment using nonmonetary models in the 1980s and 1990s.

Concluding thoughts

This overview of Phelps's work is a testimony to his impact on the history of macroeconomic thought. From the microfoundations of macroeconomics, the attack on the Phillips curve trade-off and equilibrium unemployment, to efficiency wages, optimal monetary policy, staggered contracts and a theory of moving equilibrium rate, Phelps has helped shape our view of the macroeconomy. Moreover, as a person he has clearly inspired and motivated a host of other well-known economists in their work. A surprising number of important contributions trace their origins to his influence. One can say that this particular contribution of his is more subtle yet no less real. For almost half a century Edmund Phelps has contributed to economics, driven by the excitement of discovery and the joys of creativity.

See Also

- efficiency wages
- endogenous growth theory
- golden rule
- microfoundations
- natural rate of unemployment
- new Keynesian macroeconomics
- structural unemployment

Selected works

1961. The golden rule of accumulation: a fable for growthmen. *American Economic Review* 51, 638–43.

1965. Second essay on the golden rule of accumulation. *American Economic Review* 55, 793–814.

1966. Models of technical progress and the golden rule of research. *Review of Economic Studies* 33, 133–45.

1966. (With R.R. Nelson.) Investment in humans, technological diffusion and economic growth. *American Economic Review* 56, 69–75.

1967. Phillips curves, expectations of inflation and optimal unemployment over time. *Economica* 34, 254–81.

1968. (With R.A. Pollack.) On second-best national saving and game-equilibrium growth. *Review of Economic Studies* 35, 185–99.

1968a. Money wage dynamics and labor market equilibrium. *Journal of Political Economy* 76, 687–711.

1968b. Population increase. *Canadian Journal of Economics* 1, 511–2.

1970a. (With S.G. Winter, Jr.) Optimal price policy under atomistic competition. In Phelps et al. (1970c).

1970b. The new microeconomics in inflation and employment theory. *American Economic Review* 59, 147–60; revised version in Phelps et al. (1970c).

1970c. (With A.A. Alchian, C.C. Holt, D.T. Mortensen, G.C. Archibald, R.E. Lucas, L.A. Rapping, S.G. Winter, J.P. Gould, D.F. Gordon, A. Hynes, D.A. Nichols, P.J. Taubman, and M. Wilkinson.) *Microeconomic Foundations of Employment and Inflation Theory*. New York: W.W. Norton.

1972a. The statistical theory of racism and sexism. *American Economic Review* 62, 659–61.

1972b. *Inflation Policy and Unemployment Theory*. New York: W.W. Norton.

1973a. Inflation in the theory of public finance. *Swedish Journal of Economics* 75, 67–82.

1973b. Taxation of wage income for economic justice. *Quarterly Journal of Economics* 87, 331–54.

1974. Remarks on monetary policy-making under rational expectations. Unpublished paper, Conference on rational expectations and Monetary Policy. Minneapolis: Federal Reserve Bank of Minneapolis.

1977. (With J.B. Taylor.) Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85, 163–90.

1983. (With R. Frydman.) *Individual Forecasting and Aggregate Outcomes: 'Rational Expectations' Examined*. Cambridge: Cambridge University Press.

1988. (With J.P. Fitoussi.) *The Slump in Europe: Open Economy Theory Reconstructed*. Oxford: Basil Blackwell.

1990. *Seven Schools of Macroeconomic Thought*. Oxford: Clarendon Press.

1994. *Structural Slumps: The Modern-Equilibrium Theory of Unemployment, Interest and Assets*. Cambridge, MA: Harvard University Press.

1995. The origins and further development of the natural rate of unemployment. In *The Natural Rate of Unemployment: Reflections on 25 years of the Hypothesis*, ed. R. Cross. Cambridge: Cambridge University Press.

1997. *Rewarding Work: How to Restore Participation and Self-Support to Free Enterprise*. Cambridge, MA: Harvard University Press.

1997. (With H.T. Hoon.) Growth, wealth and the natural rate: is Europe's jobs crisis a growth crisis? *European Economic Review* 41, 548–57.

1998. (With G. Zoega.) Natural-rate theory and OECD unemployment. *Economic Journal* 108, 782–801.

2000. (With J.P. Fitoussi, D. Jestaz, and G. Zoega.) Roots of the recent recoveries: labor reforms or private sector forces. *Brookings Papers on Economic Activity* 2000(1), 237–91.

2006. My kind of macroeconomics: modern economies and their policy choices. Prize Lecture, Nobel Prize in Economics, 8 December 2006. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/2006/phelps-lecture.html, accessed 5 June 2007.

Bibliography

Aghion, P., Frydman, R., Stiglitz, J.E. and Woodford, M. 2003. *Knowledge, Information and Expectations in Modern Macroeconomics*. Princeton: Princeton University Press.

Arrow, K.J. 1972. Models of job discrimination. In *Racial Discrimination in Economic Life*, ed. A.H. Pascal. Lexington, MA: D.C. Heath.

Arrow, K.J. 1973. The theory of discrimination. In *Discrimination in Labor Markets*, ed. O. Ashenfelter and A. Rees. Princeton, NJ: Princeton University Press.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85, 191–205.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Layard, R., Nickell, S. and Jackman, R. 1991. *Unemployment: Macroeconomic Performance and the Labour Market*. Oxford: Oxford University Press.

Lindbeck, A. and Snower, D. 1988. *The Insider–Outsider Theory of Employment and Unemployment*. Cambridge, MA: MIT Press.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

How to cite this article

Zoega, Gylfi. "Phelps, Edmund (born 1933)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000369> doi:10.1057/9780230226203.1281

The New Palgrave Dictionary of Economics Online

Phillips curve (new views)

Jonas D.M. Fisher
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A *Phillips curve* is an equation which relates the unemployment rate, or some other measure of aggregate economic activity, to a measure of the inflation rate. Since there is a significant correlation between inflation and unemployment over some horizons, understanding this correlation should yield insight into the impulses the economy faces and the mechanisms that propagate their effects. Since the 1990s, research has focused on making progress in three main areas: forecasting, microeconomic foundations and empirical tests of the microfoundations.

Keywords

business cycles; Cobb–Douglas functions; expectations-augmented’ Phillips curve; inflation; inflation forecasting; labour's share of GDP; menu costs; microfoundations; Phillips curve; price indexation; state-dependent models; sticky prices; technology shocks; unemployment

Article

A *Phillips curve* is an equation which relates the unemployment rate, or some other measure of aggregate economic activity, to a measure of the inflation rate. This equation continues to prompt a lot of research in macroeconomics, as it has for most of the years since the influential Phillips (1958) and Samuelson and Solow (1960) articles. The early work documents a negative relationship between the unemployment rate and either nominal wage growth or inflation. Equations relating the unemployment rate to inflation were the first to be called Phillips curves. Samuelson and Solow (1960) were bold enough to posit a stable and exploitable structural relationship between unemployment and inflation. The viability of a policy of using inflation to combat unemployment was debunked theoretically in Friedman's (1968) classic presidential address and empirically in the subsequent decade.

The rise of inflation over the 1970s came along with a breakdown in the inflation unemployment relationship and gave birth to the ‘expectations-augmented’ Phillips curve. This formulation allows the relationship between unemployment to shift due to changes in inflation expectations. Figure 1 shows how such a formulation can be used to fit the data. This shows scatter plots of unemployment and NIPA personal consumption deflator inflation for different sub-periods over the years 1948–2004 along with regression lines. Table 1 reports the regression coefficients, R^2 for the regressions and the means of inflation and unemployment. For the whole sample there is a significant *positive* relationship. However there is always a sequence of consecutive dates where the regression line is negative. The slope coefficient is also highly significant in all cases but one. The movements in the regression line occur as changes in the mean inflation and unemployment rates. Another way to verify that there is a strong association between inflation and unemployment is to focus on business cycle frequencies, in the bottom right hand corner of Figure 1 and the second row of Table 1. Clearly inflation and unemployment are highly correlated at business cycle frequencies.

The US Phillips curve, 1948:1–2004:4

Sample	Intercept	Slope	R^2	Mean inflation	Mean unemployment
Full sample	2.06^{***}	0.29^{**}	0.02	3.70	5.64

Business cycle frequencies	0.32****	****	0.39	0.00	0.00
1948:1–1969:4	6.55****	****	0.16	2.20	4.67
1970:1–1973:4	20.0****	****	0.52	5.07	5.35
1974:1–1984:4	20.7****	****	0.49	7.54	7.49
1985:1–1994:4	10.2****	****	0.27	3.52	6.41
1995:1–1996:4	5.51	–0.48	0.01	2.89	5.50
1997:1–2001:4	11.1****	****	0.56	2.24	4.47
2002:1–2004:4	15.7****	*	0.17	2.46	5.57

Note: The number of asterisks from one to four denotes significance at the 10, 5, 1, and 0.1 per cent levels of the constant and slope terms of a regression of inflation on unemployment. Business cycle frequencies means the data have been subjected to Christiano and Fitzgerald's (2003) band pass focusing on a 2–8 year horizon. *Sources*: authors' calculations; unemployment rate: US Bureau of Labor Statistics; personal consumption expenditure deflator: US Department of Commerce.

Figure 1

The US Phillips curve, 1948:1–2004:4. *Sources*: authors' calculations; unemployment rate: US Bureau of Labor Statistics; personal consumption expenditure deflator: US Department of Commerce.

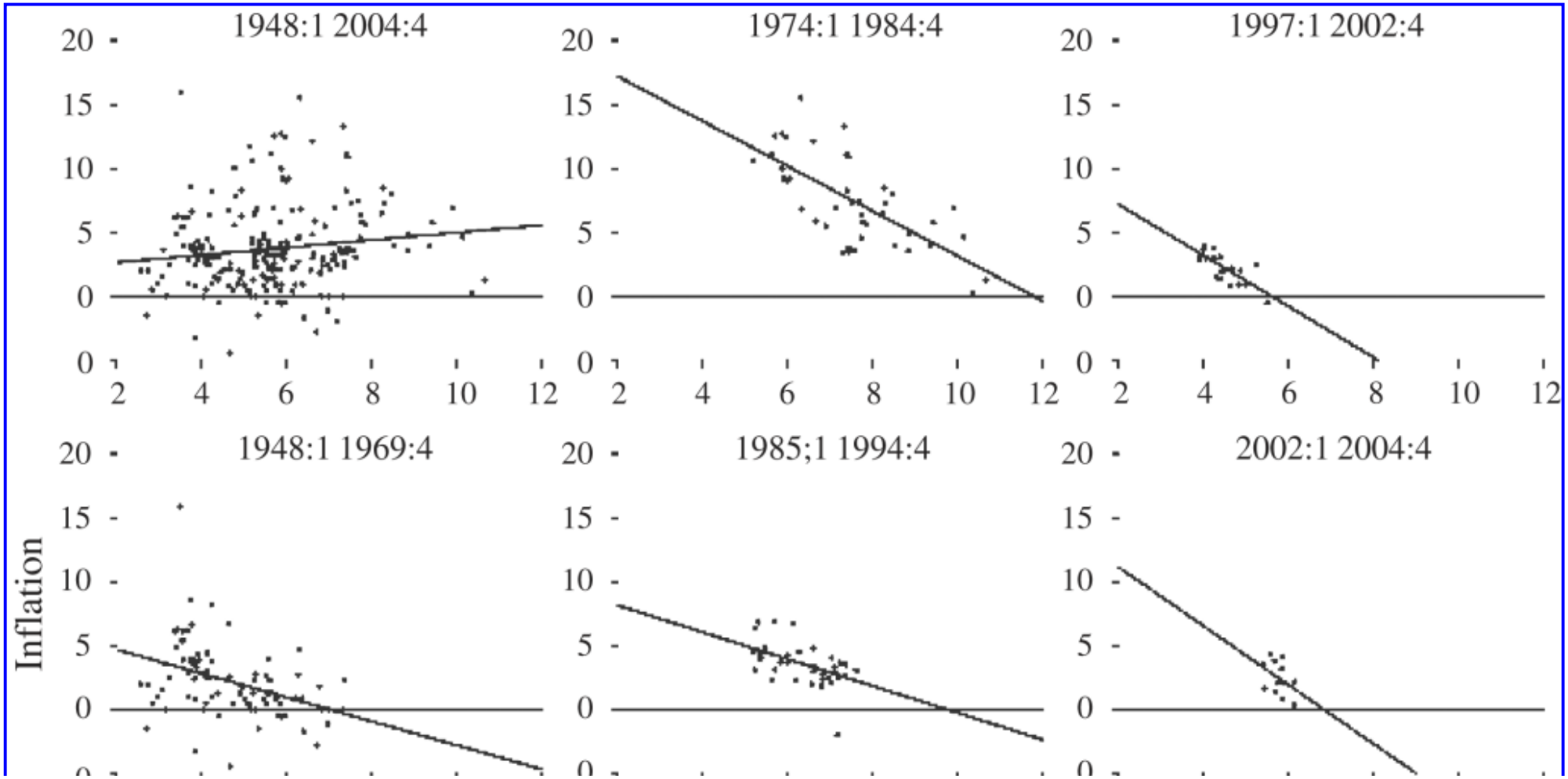

Sources: authors' calculations; unemployment rate: US Bureau of Labor Statistics; personal consumption expenditure deflator: US Department of Commerce.

Since there is a significant correlation between inflation and unemployment over some horizons, understanding this correlation should yield insight into the impulses the economy faces and the mechanisms that propagate their effects. Since the 1990s, research has focused on making progress in three main areas: forecasting, microeconomic foundations and empirical tests of the microfoundations. This article reviews the recent research in each of these areas.

Forecasting with the Phillips curve

Inflation forecasting models rely heavily on the Phillips curve. For many years, even as the traditional Phillips curve relationship evaporated, variables such as the unemployment rate have continued to be very useful predictors of future inflation. Stock and Watson (1999) argued they could do better. They proposed using principal components of large numbers of data series to aid in forecasting macroeconomic variables. The idea was that this approach uses the information in a large number of variables, which is impossible with traditional regression-based forecasting. One of their most interesting findings involves the first principal component of roughly 80 macroeconomic variables, including measures of production and income, employment, unemployment and hours, personal consumption and housing, and sales, orders and inventories. They argued that this 'activity index' variable is more useful than even unemployment for predicting inflation. Such a finding strongly suggests a connection between current activity and future inflation, essentially the Phillips curve relationship.

Atkeson and Ohanian (2001) argued that the success of the Phillips curve in forecasting is just as illusory as a stable Phillips curve. They argued that, for forecasting one year ahead, a simple random walk suffices – the best predictor of one-year-ahead inflation is current inflation. Atkeson and Ohanian's (2001) finding has proven to be remarkably robust (see Brave and Fisher, 2004; Fisher, Liu and Zhou, 2002). However, the random walk result does depend on the sample period considered by Atkeson and Ohanian (2001), which is 1984–99. Beginning the sample in 1984 is justified by evidence of a major structural change around that time (see Fisher, 2006). However, as the sample is extended, the random walk loses some of its lustre. The poor performance of Phillips curve-based forecasting models is mainly confined to the period 1984–93. Since the mid-1990s the traditional variables such as unemployment have been useful forecasters. These findings are easily explained by noting that inflation was generally falling from 1984 to the mid-1990s as the economy adjusted to the Federal Reserve's stronger willingness to fight inflation. It is natural for old models to fail after a major structural change. Moreover, in an environment where output is growing strongly while inflation is falling, it is not surprising the random walk model does well between 1984 and 1993.

Microfoundations of the Phillips curve

Since Lucas (1972) economists have known how to formulate models in which inflation and activity are correlated but there is not a policy-exploitable Phillips curve. The focus of much of the recent literature has been on the Calvo–Yun Phillips curve, which arises from one particular model. The Phillips curve in this model is named after Calvo (1983) and Yun (1996). Most of the literature uses the hopelessly ambiguous term 'New Keynesian' to describe this model of the Phillips curve.

Phillips curves arise naturally in models where firms set prices and at least some of those prices do not respond to every shock to the economy. Calvo's contribution is a very simple model of sticky prices. He assumed monopolistically competitive firms could re-optimize their price with a fixed probability, θ, each period so that firms re-optimize prices on average every $1/(1-\theta)$ periods (usually quarters of a year). This formulation can be taken literally, in which case prices are fixed until the next opportunity to re-optimize. Alternatively, firms might follow simple pricing rules at high frequencies and occasionally adjust these rules overtime. Under this interpretation firms can index their prices to inflation.

Yun derived a Phillips curve by introducing the Calvo model of price adjustment into an otherwise standard monetary model with monopolistic competition and constant markups. The result is the Calvo–Yun Phillips curve:

$$\hat{\pi}_t = \beta E_t \hat{\pi}_{t+1} + \frac{(1-\beta\theta)(1-\theta)}{\theta} AD\hat{s}_t. \tag{1}$$

The variable $\hat{\pi}_t$ is the deviation of the log of the gross inflation rate from its steady state value, $\hat{s}_t$ is the log deviation of real marginal cost for the representative firm, and β is the time discount factor of the representative household. The *A* and *D* terms are equal to unity in the Yun paper. This equation is derived from the log-linearized necessary conditions of the equilibrium. To linearize around a steady state with positive inflation, firms must index their prices to inflation.

Eichenbaum and Fisher (2007) describe how Kimball's (1995) extension to variable markups implies that $0<A\bullet 1$, where *A* depends on the shape of the firm's demand curve and equals unity in the constant markup case. Eichenbaum and Fisher also study Woodford's (2003; 2005) model of capital adjustment and describe how this yields.

$0<D\bullet 1$ where *D* depends on the firm's supply curve. Generally, marginal cost is increasing in output. Since β and θ also lie between zero and unity, the coefficient in front of marginal cost is positive and (1) is an equilibrium relationship where output and inflation are positively related. In most of the literature assumptions are such that $A=D=1$. This literature generally predicts reasonably large effects of monetary shocks if firms adjust their prices once a year.

The Calvo model is called a *time-dependent* model because the opportunity to change prices depends only on the passage of time. Taylor's (1980) model where firms rotate changing their prices is also a time-dependent model. The main alternative is *state-dependent* models, where changing the price is a choice of the firm which depends on both firm-level variables such as productivity and aggregate variables like the interest rate. The dominant state-dependent model involves menu costs. Studying state-dependent models is more difficult than time-dependent models because the price distribution is endogenous.

Five papers make major progress toward understanding menu cost models. Dotsey, King and Wolman (1999) study a model with random menu costs and Taylor-style staggering. A key advantage of their model is that it can be linearized like a simple real business cycle model. Klenow and Krsystov (2005) calibrate this model to US consumer price index (CPI) micro data for the years 1988–2003. They find that matching the micro data yields a model which behaves very much like the Calvo–Yun model. Golosov and Lucas (2003) study a menu cost model with a constant menu cost but where firms face exogenous technology and/or preference shocks. Under the assumption that the shocks are Gaussian, Golosov and Lucas find that firms choose to adjust their prices a lot when there is a monetary shock, and this makes prices flexible enough that monetary shocks have small affects. Midrigan (2005) uses scanner data to determine the distribution of technology or preference shocks in the Golosov–Lucas model. He estimates this distribution to be non-Gaussian with fat tails. With the estimated distribution monetary shocks have affects similar to models with a Calvo–Yun Phillips curve. Gertler and Leahy (2005) develop an analytically tractable state-dependent model which also behaves like the Calvo–Yun model.

Another key area of research involves building fully specified dynamic general equilibrium models with Phillips curves which fit the data well. This work has focused on the Calvo–Yun Phillips curve instead of more deeply motivated models because of its simplicity. The key contribution is Christiano, Eichenbaum and Evans (2005). Their model also includes portfolio rigidities, adjustment costs in capital, and a Calvo-style version of nominal wage setting. They find their model does a good job matching the evidence on how the economy responds to a monetary shock, with a small amount of price stickiness, but wages must be more rigid. There is a growing amount of research which reaches the same basic conclusion that the wage–activity relationship is more important for understanding macroeconomic dynamics than the traditional Phillips curve (cf. Galí, Gertler and López-Salido, 2007).

Empirical evaluation of the microfoundations

Equation (1) is to the empirical macro literature as the Lucas (1978) asset pricing relationship is to empirical finance. In recent years it has come under considerable empirical scrutiny. Galí and Gertler (1999) were the first to use Hansen's (1982) generalized method of moments (GMM) to estimate θ and test (1). They measured marginal cost using labour's share of GDP, which is true if firms use a Cobb–Douglas production technology. Gagnon and Kahn (2005) consider other production structures where marginal cost is not measured with labour's share and conclude that the Galí and Gertler (1999) findings hold up. Galí and Gertler estimate of θ implies more than a year between price changes, but they cannot reject the equation. Micro price data might be useful to identify θ, but, under the frequency of *re-optimization* interpretation of the Calvo model, estimates of θ over a year for the United States seem too high (cf. Blinder et al., 1998). Galí and Gertler consider an alternative model with ‘rule-of-thumb’ firms who use lagged inflation to update their prices when they have the opportunity to re-optimize. This model is motivated by the fact that lagged inflation enters significantly in the empirical version of (1). The model with rule-of-thumb firms is not rejected and the estimates of θ imply prices are re-optimized every two or three quarters. The latter estimates are within the range of plausibility. Galí and Gertler also estimate the number of rule-of-thumb firms to be small and emphasize that (1) holds approximately. Bakhshi, Kahn and Rudolf (2005) argue that the Galí–Gertler ‘hybrid’ model is a good approximation to Dotsey, King and Wolman's (1999) menu cost model.

It is clear that θ is not identified separately from A and D in (1). However, A and D can be identified with auxiliary information. Sbordone (2002) identifies D by assuming the stock of capital is fixed exogenously at each firm for all time. Under the usual assumption of a Cobb–Douglas production function, auxiliary information on the share of labour income in GDP can be used to identify D. Sbordone considers the forward looking solution to (1) as well as the solution to a similar equation for the labour market. The expected present-value calculations needed to implement this estimation are implemented with a vector autoregression. This empirical strategy is analogous to Abel and Blanchard's (1986) approach to estimating investment adjustment costs. Sbordone estimates prices are re-optimized every one to two quarters. Galí, Gertler and López-Salido (2001) apply Sbordone's fixed capital assumption to their rule-of-thumb model and estimate the frequency of re-optimization to be a little higher than in Galí and Gertler (1999), and significant small positive numbers of rule-of-thumb firms continue to be estimated.

Eichenbaum and Fisher (2007) explore Woodford's (2003; 2005) dynamic version of Sbordone's (2002) model in an environment which also includes Kimball's (1995) variable markup. As in Galí and Gertler (1999) and Galí, Gertler and López-Salido. (2001), they adopt a GMM estimation and testing strategy. To improve the power of their tests, Eichenbaum and Fisher (2004) impose the restrictions eq. (1) place on the moving average structure of the Euler equation errors and reduce the number of instruments compared to the previous papers. They easily reject (1) assuming the Euler error is an MA(0). This motivates them to include the auxiliary assumption that firms make decisions based on lagged information. With one such *implementation* lag, this yields an MA(1) structure which is not rejected and which the re-optimization frequency is about two years, if $A=D=1$. With empirically motivated values for the curvature of the demand curve and the size of capital adjustment costs, re-optimization every two quarters cannot be ruled out at conventional significance levels. Eichenbaum and Fisher include dynamic indexation (prices indexed to the most recent inflation rate) of the prices of firms that do not re-optimize in a given period. This is an alternative to rule-of-thumb firms as away of introducing a lagged inflation term into (1). Eichenbaum and Fisher (2004) find they cannot reject the possibility that there are no rule-of-thumb firms under dynamic indexation.

Recently much work has been done to document prices at the microlevel. Blinder et al. (1998) survey actual price setters and find that, among firms reporting regular price reviews, annual reviews are by far the most common. Other key contributions include Bils and Klenow (2004), Klenow and Krystov (2005) and work done with European data for example by Stahl (2005). Much of this literature emphasizes the frequency of price changes. For example, Blinder et al. (1998) report that the median time between price changes among the firms that they survey is roughly three quarters. Comparing the Calvo–Yun Phillips curve with these findings is delicate. With price indexation the model implies that prices change *too* frequently relative to the micro data because all prices are changing all the time. Also, just because firms are changing prices does not mean that they have re-optimized those prices: a subset of the price changes being recorded could reflect various forms of time-dependent pricing rules.

Integrating over all the micro evidence, with a low inflation economy like the United States, versions of the Calvo–Yun Phillips curve with an implementation lag, dynamic indexation, capital adjustment costs and time-varying markups can be reconciled with the macro data without requiring implausible degrees of rigidities in price-setting behaviour at the micro level. Of course this model is not literally ‘true’. For instance, the model also has the implausible implication that any CPI observation for which $P_{i,t}/P_{i,t-1}$ is not equal to π_{t-1} involves re-optimization. Developing tractable models that are fully consistent with the salient macro facts and the emerging literature on the behaviour of individual good prices is a key challenge going forward.

See Also

- adaptive expectations

- adjustment costs
- generalized method of moments estimation
- inflation
- inflation dynamics
- macroeconomic forecasting
- Phillips curve

Bibliography

Abel, A. and Blanchard, O. 1986. The present value of profits and cyclical movements in investment. *Econometrica* 54, 249–74.

Atkeson, A. and Ohanian, L.E. 2001. Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review* 25(1), 2–11.

Bakhshi, H., Kahn, H. and Rudolf, B. 2005. The Phillips curve under state-dependent pricing. Manuscript, Carleton University.

Bils, M. and Klenow, P. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112, 947–85.

Blinder, A., Canetti, E., Lebow, D. and Rudd, J. 1998. *Asking About Prices: A New Approach to Understanding Price Stickiness*. New York: Russell Sage Foundation.

Brave, S. and Fisher, J.D.M. 2004. In search of a robust inflation forecast. *Economic Perspectives* 28(4), 12–31.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–45.

Christiano, L. and Fitzgerald, T. 2003. The band pass filter. *International Journal of Economics* 44, 435–65.

Dotsey, M., King, R.G. and Wolman, A.L. 1999. State-dependent pricing and the general equilibrium dynamics of money and output. *Quarterly Journal of Economics* 114, 655–90.

Eichenbaum, M. and Fisher, J. 2004. Evaluating the Calvo model of sticky prices. Working Paper No. 10617. Cambridge, MA: NBER.

Eichenbaum, M. and Fisher, J. 2007. Estimating the frequency of price re-optimization in Calvo-style models. *Journal of Monetary Economics* (forthcoming).

Erceg, C., Henderson, J., Dale, W. and Levin, A.T. 2000. Optimal monetary policy with staggered wage and price contracts. *Journal of Monetary Economics* 46, 281–313.

Fisher, J.D.M. 2006. The dynamic effects of neutral and investment-specific technology shocks. *Journal of Political Economy* 114, 413–51.

Fisher, J.D.M., Liu, C. and Zhou, R. 2002. When can we forecast inflation. *Economic Perspectives* 26(1), 30–42.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Gagnon, E. and Kahn, H. 2005. New Phillips curve under alternative production technologies for Canada, the United States, and the Euro area. *European Economic Review* 49, 1571–602.

Galí, J. and Gertler, M. 1999. Inflation dynamics: a structural econometric analysis. *Journal of Monetary Economics* 44, 195–222.

Galí, J., Gertler, M. and López-Salido, D. 2001. European inflation dynamics. *European Economic Review* 45, 1237–70.

Galí, J., Gertler, M. and López-Salido, D. 2007. Mark-ups, gaps and the welfare costs of business cycles. *Review of Economics and Statistics* 89, 44–59.

Gertler, M. and Leahy, J. 2005. A Phillips curve with an Ss foundation. New York, University manuscript.

Golosov, M. and Lucas, R.E., Jr. 2003. Menu costs and Phillips curves. Working Paper No. 101187. Cambridge, MA: NBER.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Kimball, M. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27, 1241–77.

Klenow, P. and Krystov, O. 2005. State-dependent or time-dependent pricing: does it matter for recent US inflation? Working paper, Bank of Canada.

Lucas, R.E., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Lucas, R.E., Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

Midrigan, V. 2005. Menu costs, multiproduct firms and aggregate fluctuations. Manuscript, Ohio State University.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25, 283–99.

Samuelson, P.A. and Solow, R.M. 1960. Analytical aspects of anti-inflation policy. *American Economic Review* 50, 177–94.

Sbordone, A. 2002. Prices and unit labor costs: a new test of price stickiness. *Journal of Monetary Economics* 49, 265–92.

Stahl, H. 2005. Price setting in German manufacturing: new evidence from new survey data. Working Paper No. 561, European Central Bank.

Stock, J. and Watson, M. 1999. Forecasting inflation. *Journal of Monetary Economics* 44, 293–335.

Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, NJ: Princeton University Press.

Woodford, M. 2005. Firm-specific capital and the New Keynesian Phillips curve. *International Journal of Central Banking* 1(2), 1–46.

Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37, 345–70.

How to cite this article

Fisher, Jonas D. M. "Phillips curve (new views)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000353>

doi:10.1057/9780230226203.1284

The New Palgrave Dictionary of Economics Online

Phillips curve

Edmund S. Phelps
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In 1957 A.W. Phillips argued that, other things equal, the rate at which the nominal wage level was changing was a decreasing function of the level of the unemployment rate. Further, the rate of unemployment required to keep the rate of wage inflation down to the normal level was certainly positive in the United Kingdom, the domain of Phillips's data, had remained stable for nearly a century. Milton Friedman and Edmund Phelps criticized the concept of a stable Phillips curve for having treated wage-setters' behaviour, which presumably involved their expectations of the general wage movement, as a mechanical toy.

Keywords

aggregate demand; cost inflation; equilibrium price level; excess demand and supply; Fellner, W. J.; Fisher, I.; Friedman, M.; hyperinflation; inflation; Keynes, J. M.; labour supply; Lipsey, R. G.; Lucas, R.; monetary theory; Muth, J.; natural rate of unemployment; nominal wages; Phelps, E. S.; Phillips curve; Phillips, A. W.; political business cycles; rational expectations; Samuelson, P. A.; Sargent, T.; unemployment; wage inflation

Article

By the 1950s there was achieved a working synthesis, despite some unsolved problems, of the contributions of Keynes to monetary theory with the older truths of his several predecessors Marshall, Pigou, Wicksell and Fisher. Given the supply of money, there is a nominal price level that is in some suitable sense the equilibrium price level; more generally, there is an equilibrium path of the price level. The equilibrium price level in the current period, given next period's price level, is just high enough to reduce the real value of this period's cash balances down to the quantity demanded – figured at the corresponding nominal rate of interest (which is a decreasing function of this period's price level) and output level (which was taken to be independent of the price level if nominal wages were also taken as finding their equilibrium level). If people expect that the general level of prices and nominal wages is higher, and we assume that the actual price level at first equals this expected level, the result will be disappointment – an unexpected weakening of sales. Presumably, the price and wage levels will then tend to adjust, and perhaps employment will detour from its equilibrium level in the process.
The disequilibrium dynamics of the adjustment process, however, remained *terra incognita*. Suppose that is a sudden and unexpected disturbance that displaces upwards or downwards the path of the equilibrium price level. Keynes had declared in his 1936 book that the money wages set by producers would not generally take the downward jumps occasionally necessary for continued maintenance of equilibrium, hence the need for a more general theory of interest and employment in which the nominal wage level was not on the equilibrium track. (He further opined that lessened wage inflexibility would be destabilizing.) By the 1950s it was agreed that wages would *gradually* move from the former equilibrium path, if we assume they were originally in equilibrium, toward the new and lower equilibrium path, whether or not there would be later overshooting, and further that, if there is such gradualness, the result will be a bulge of unemployment during the process of wage adjustment. Similarly, an upward displacement of the equilibrium path would likewise engender only a gradual adjustment of money wages, accompanied in this case by a dip of the unemployment rate below its equilibrium, or normal, level. Increasingly, economists spoke of buying a spell of abnormally low unemployment by generating a round of inflation. (Yet, some economists of Austro-Hungarian or German schooling, notably William Fellner, argued that successive doses of (equal) inflation would lose their effectiveness, so that the same effect on unemployment would require ever increasing doses, as anticipations of higher demand came to be built into wage contract increases.) The term cost inflation arose to refer to the sort of inflation the avoidance of which needed the

discipline, and social waste, of unemployment above what could be achieved through high demand.

It was against this background that A.W. Phillips's extraordinary article, scholarly yet accessible, appeared in the academic journal *Economica* in 1957. Phillips changed the terms of discourse of the subject from the qualitative and discontinuous to ordinary quantitative terms. Other things equal, such as the rate of change of unemployment, the rate at which the nominal wage level is changing – the (algebraic) rate of wage inflation – is a decreasing function of the level of the unemployment rate. Further, the rate of unemployment required to hold down the rate of wage inflation to the level of normal experience – the average, and accustomed, rate – is certainly positive, perhaps 2 to 3 per cent in the United Kingdom, the domain of Phillips's data, and has not shifted notably over nearly a century of observation. Almost overnight the Phillips curve (so named in a discussion by Samuelson and Solow) invaded the language of macroeconomics.

Phillips uncovered another fact about past wage inflation. Among years with the same (annual) level of the unemployment rate there tended to be a higher rate of wage inflation when the annual unemployment rate was falling, as in a cyclical recovery or developing boom, than when the annual unemployment rate was rising. Phillips drew a counterclockwise loop around the downward-sloping Phillips curve, to indicate the typical motion of the wage inflation rate in relation to the unemployment rate over the typical historical cycle. (See the lower Phillips curve and the loop around it in Figure 1.) It remained for R.G. Lipsey, also of the London School of Economics at that time, to express this historical phenomenon in quantitative terms too. Lipsey in 1960 published estimates obtained by regression analysis of the coefficients of a linear rate-of-wage-change equation in which the explanatory righthand-side variables were the level of the unemployment rate and its rate of change. The negative sign of Lipsey's estimate of the latter coefficient reflected the above loop. The statistical estimation of such Phillips–Lipsey equations rapidly developed from a cottage activity using electric calculators to a booming computerized industry.

Figure 1

Wage and unemployment dynamics

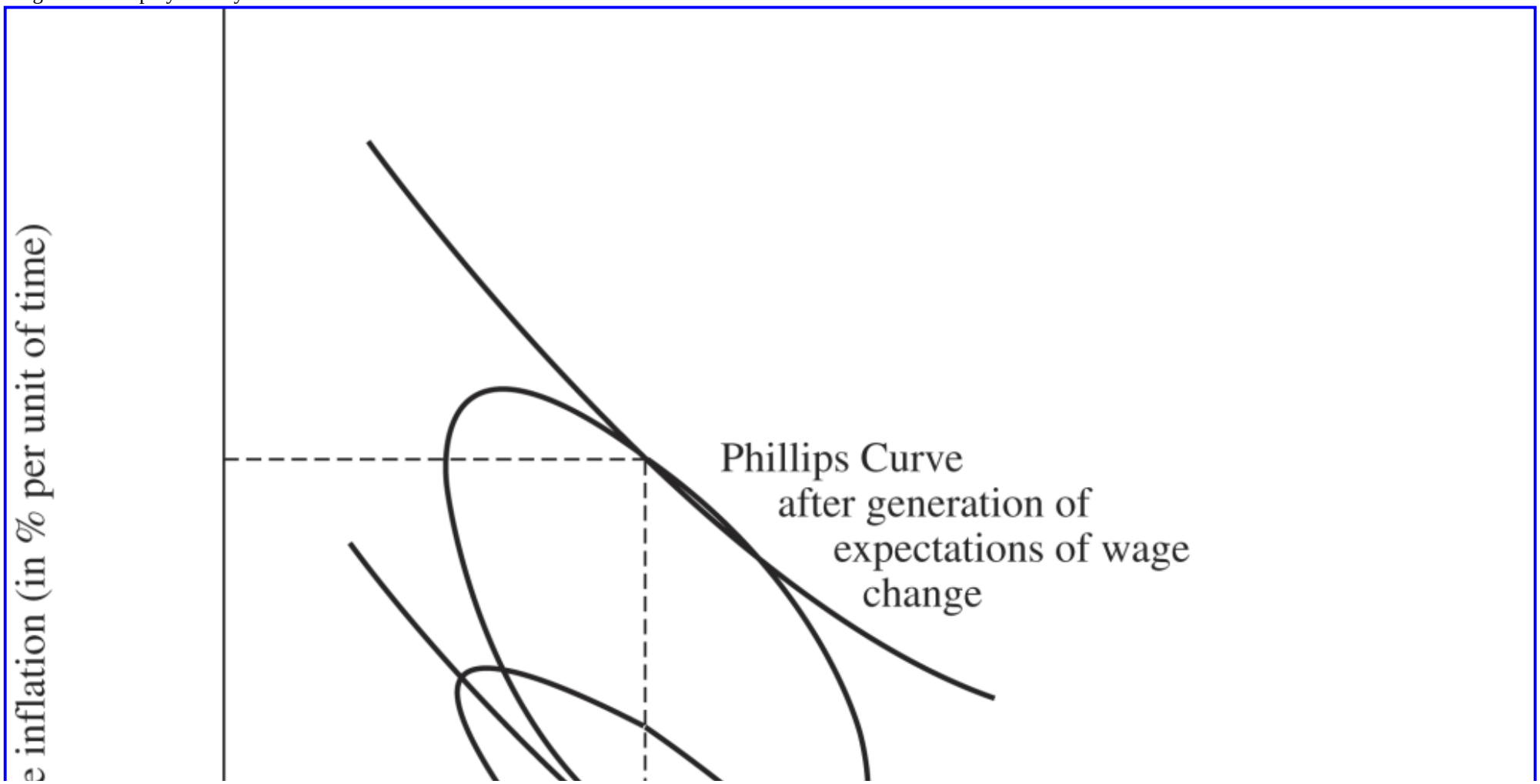

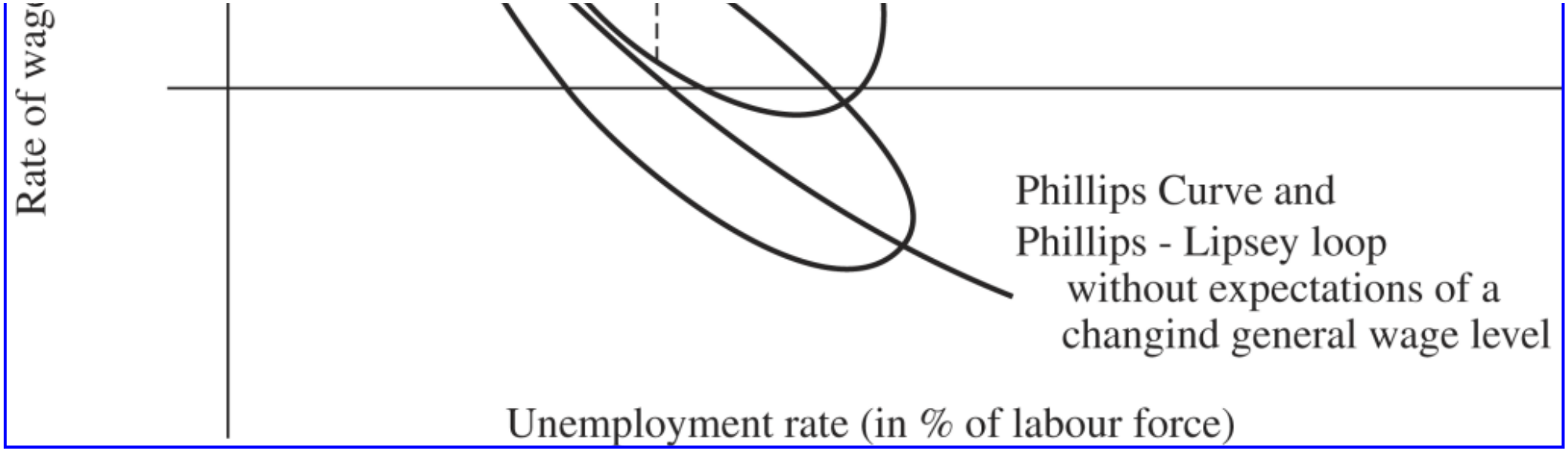

In a way, the new and developing fact book seemed to contain information that was entirely reasonable and surely in keeping with existing theoretical (or pretheoretical) notions. (Indeed, a remarkably early anticipation of the Phillips curve was later unearthed in an obscurely placed paper in 1926 by Irving Fisher.) It seemed to say, essentially, that if there was an aggregate excess supply then nominal wages would be found falling and employment would be depressed – as long as wages remained too high to eliminate the excess supply – and both effects of the excess supply would be larger the greater was the size of the excess supply. More exactly, the sudden appearance of an excess supply that is maintained at a given level for a while would first generate a positive rate of change of unemployment alongside falling wages and only later, in a sort of disequilibrium steady state, a higher level of the unemployment rate without a positive rate of change. This part of the Phillips curve story seemed unsurprising and unpuzzling.

Yet some theoretical problems that had long lain submerged and unnoticed when the subject of disequilibrium adjustment was still muddy and relatively quiet came to surface once the Phillips–Lipsey formulation had stirred things up. Among these was the problem of explaining why nominal wages did not jump down to their new equilibrium level (with prices jumping after them) and, beyond that, the problem of determining the pace with which wages fell. The same theoretical void had been created more than a decade before Phillips's article when Samuelson in his *Foundations*, addressing Walrasian stability, simply postulated that the rate at which the price of a commodity falls is an increasing function of the excess supply of it. This was a macroeconomic hypothesis, perhaps a kind of theory by the behavioural standards of the day, but not a microeconomic theory running in terms of the motives and perceptions of the individual actors operating in the economy.

If the first problem was explaining that the Phillips curve was sloping, the second problem was explaining its remarkably rightward position: Money wage rates tended to be rising over a range of positive unemployment rates, including rates exceeding the lower bound obtainable by high-pressure aggregate demand levels. If nominal wages tend to be rising as long as the unemployment rate stays above non-depression levels, then the Samuelsonian hypothesis explains that markets have normally operated in a state of considerable excess demand. But is that likely? Is a state of zero excess demand (and excess supply) really marked by a zero rate of wage change, or is something missing here? Somehow, it was evident, the factors of productivity growth and inflation needed to be brought into the analysis, but not just as incantations to make the problem go away.

For many economists there was the further problem of reconciling the empirical regularity depicted by the Phillips curve, which seemed to possess an extraordinary stability, with the older Continental, or Austro-Hungarian, doctrine, propounded by Fellner and others, that below-normal unemployment constantly fuelled by a permissive monetary–fiscal policy will soon cause wages (and hence prices) to rise in ever-accelerating fashion until the hyperinflation finally brings collapse or structural change. This was the further problem of understanding in microeconomic terms the shiftability of the Phillips curve.

The solution of the first problem, that of explaining the gradualness of the wage adjustment and the attendant slump of employment, led theorists in the 1960s in the same direction in which Keynes had been led in his search for an explanation of slumps. A key element of the solution was the fact that there is no coordination, to use Keynes's term, among the managers deciding upon wages and employment (inter alia) at the various production sites. If there is a weakening of aggregate demand – here, a curve in the output–price level plane – in a previously normal and equilibrium situation, the resulting fall in the demand curve facing the individual manager, or producer, even if seen by him as permanent, would not induce the workers employed there (or unemployed there) to accept the job-preserving money wage cut unless they were expecting workers elsewhere at the same moment to be facing and accepting the very same percentage wage cut; and they would have not reason to have that expectation unless there was news bearing on the scale of the decline in demand and such news was observed to have produced job-preserving wage cuts. Pending such news, then, there would be only an insufficient wage cut, so the supply price of output would fall by less than the demand price, and hence output and employment would decline. These impact effects would show a negative correlation between wage change and

unemployment level (though here the true correlation is with the change of employment).

In the 1960s, however, a number of theorists pointed out the theoretical existence of a deeper Phillips curve relation. The higher unemployment level comes about because ‘expected wages’ in the economy as a whole exceed ‘actual wages’, and as information comes in that actual wages elsewhere are lower than expected the ensuing downward revision of expectations will induce workers to accept still lower actual wages. This latter wage fall grows out of the disequilibrium situation, like the higher unemployment. If one were to go so far as to posit static expectations, so that each observed wage decline is thought to be the last, there would exist a disequilibrium steady-state relationship between the size of the (swelling of the) unemployment rate and the magnitude of the rate of wage change. A 1969 Pennsylvania conference developed these points in a variety of models, and the conference volume published a year later served to popularize these expectational microeconomic foundations of unemployment and wage-price behaviour (Phelps et al., 1970).

In the 1970s theorists moved toward rational expectations in the sense of Muth. In this case, the news of the initial fall of wages (together with any news on the unemployment front) is enough for workers to expect that the general wage level will now fall to exactly the job-preserving level, so that the unemployment rate will return to the equilibrium level; otherwise workers are implied to be repeatedly misforecasting the wage level, contrary to rational expectations. Here, too, the high unemployment precedes a wage fall (though large enough to eliminate the high unemployment), so that there is again a negative correlation between unemployment level and wage change. A microtheoretic model along these lines, involving known stationary stochastic processes, was developed by R.E. Lucas (1972, 1973) and an intertemporal model with which to show, as a corollary, the ineffectiveness of preannounced monetary policy in stabilizing output or employment was analysed by T.J. Sargent (1973).

The rational expectations postulate seemed at first to point to the conclusion that, following an unexpected drop of aggregate demand, nominal wages would indeed jump – though too late to prevent a recession – once the news of the economic indicators signalling a slump was out, and that with that jump the unemployment rate would jump back to its steady-state equilibrium (and normal) level. But that would have been jumping to conclusions, and fortunately so for the rational expectations hypothesis since there is convincing econometric evidence that the unemployment rate displays statistical persistence. It was soon remembered, however, that the antecedent literature on the costs of recruitment or training provided the basis for an equilibrium path of recovery from a downturn along which both the unemployment rate and the nominal wage level decline continuously, or gradually in discrete-time terms, rather than with a jump. There was also a development of the point made in the earlier literature that firms’ wage commitments are apt to be durable and non-synchronous, so that the respective firms in the economy take turns over the wage-setting cycle, or ‘year’, in resetting their ‘annual’ wage scales. In such a nonsynchronous wage-setting context, the average level of nominal wages cannot jump and hence employment will not recover from a recession with a jump. Further, a model of wage staggering, though quite different from the preceding types, likewise produces an explanation of the negative correlation between wage change and the unemployment rate, as shown by Taylor (1980).

The second Phillipsian problem, that of explaining the coexistence of rising nominal wages with above-minimum unemployment, had two answers, independent and additive. One answer lay in divorcing ourselves from thinking of the unemployment rate – or even the excess of the unemployment rate over the minimum rate achievable by stimulating aggregate demand – as a satisfactory measure of downward pressure on nominal wages. If the unemployment rate (or, more accurately, the aforementioned excess rate) were driven to zero, quitting would presumably be rampant and so the representative firm would endeavour to pay a wage premium – a positive differential over the wages paid elsewhere. If this average wage level is expected to be unchanged, the firm will therefore raise its wage to a level in excess of that average, with the consequence that the average wage will actually rise – resulting in an excess of ‘actual’ over ‘expected’, thus a disequilibrium. It is only when the unemployment rate (or the excess rate) is positive and high enough that the quit rate will be damped sufficiently to encourage the representative firm to content itself with paying the representative wage, that the average wage will remain flat as expected. (The argument is implicit in Phelps, 1968, and the explicit focus of Stiglitz, 1974, and Salop, 1979.) In this equilibrium there is involuntary unemployment in a natural sense of the term, since wages exceed the market-clearing level, and this unemployment may very well exceed job vacancies (if any), so there may be considerable excess supply. (See also Calvo, 1979, for another model.)

The other answer to the problem lay in realizing that wages do not rise only when firms (or at least the representative firms) want to be more competitive than the others. Wages may also rise because the firms believe they must raise their wages just to avoid losing any of their present competitiveness. The same point can be made in terms of the excess-demand framework of Samuelson: the error in Samuelson's formulation was in excluding the possibility that wages will be increased in an anticipatory move that serves to prevent the emergence of an excess demand, not just in response to excess demands that are not previously expected and forestalled by intervening wage increases. Hence, nominal wages may be rising not because the labour market is in disequilibrium, marked by mutually inconsistent desires among the firms for superior competitiveness in the labour market, but rather because the prospect of productivity growth or of inflation or of both generates expectations that the general level of wages is going to increase (Phelps, 1968).

With the latter insight our third problem, that of explaining the possible shift of the Phillips curve, is also solved. When governments seek to exploit the Phillips curve by trading off price stability in hopes of obtaining reduced unemployment in retun, they ultimately engender expectations of regularly increasing wages. Such an increase in the expected rate of wage inflation (at each level of the unemployment rate) shifts up the Phillips curve; a new one arises corresponding to the new expected rate of wage inflation. In Figure 1 see the upper Phillips curve, which has been driven higher by expectations of a rising general wage level. It is now evident that a political business cycle, by alternately lifting and depressing the Phillips curve, would generate the clockwise loop shown in the figure.

If we posit, as a plausible approximation, that the expected wage inflation variable takes its place among the explanatory right-hand variables (alongside the Phillips–Lipsey terms) with a unitary coefficient, the implication is that the steady-state equilibrium unemployment rate – at which expectations are borne out – is the same number independently of the inflation rate. Then, maintaining a steady unemployment rate below that constant equilibrium rate would entail rising inflation without bound (Phelps, 1968; see also Friedman, 1968,

discussed below). With this coefficient value of one (or any larger value) the model gives algebraic expression to the abiding accelerationist fears of the Austro-Hungarian school. The notion that the equilibrium unemployment rate was a constant, as above, also emerged from a quite different formulation by Friedman (1968), where the constant was dubbed the natural rate of unemployment. There the rate of wage change is postulated to be a function of the unemployment rate plus the expected rate of *price* inflation. The implicit rationale was that the amount of labour supplied as an increasing function of the expected real value of the nominal wage. A way to synthesize the above wage–wage model (in which expected real-wage changes are captured in the Phillips–Lipsey terms) with the wage–price model is to add to a quasi-Phillips *em*ployment term a weighted average of the expected rates of wage inflation and price inflation where the latter weight is positive, zero, or negative as the labour supply curve is forward rising, vertical, or backward sloped (Phelps, 1979).

See Also

- neoclassical synthesis

Bibliography

Calvo, G.A. 1979. Quasi-Walrasian theories of unemployment. *American Economic Review* 69, 102–7.

Fellner, W.J. 1959. Demand inflation, cost inflation and collective bargaining. In *The Public Stake in Union Power*, ed. P.D. Bradley, Charlottesville: University of Virginia Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Lipsey, R.G. 1960. The relation between unemployment and the rate of change of money wage rates: a further analysis. *Economica* 27, 1–31.

Lucas, R.E., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4(2), 103–24.

Lucas, R.E., Jr. 1973. Some international evidence on output–inflation tradeoffs. *American Economic Review* 63, 326–34.

Phelps, E.S. 1968. Money-wage dynamics and labor-market equilibrium. *Journal of Political Economy* 76(4), Pt II, July–August, 678–711.

Phelps, E.S. 1979. *Studies in Macroeconomic Theory. Vol. 1: Employment and Inflation*. New York: Academic Press.

Phelps, E.S. et al., eds. 1970. *Microeconomic Foundations of Employment and Inflation Theory*. New York: Norton.

Phillips, A.W. 1958. The relation between unemployment and the rate of change in money wage rates in the United Kingdom 1861–1957. *Economica* 25, 283–99.

Salop, S.C. 1979. A model of the natural rate of unemployment. *American Economic Review* 69, 117–25.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. and Solow, R.M. 1960. Analytical aspects of anti-inflation policy. *American Economic Review, Papers and Proceedings* 50, 177–94.

Sargent, T.J. 1973. Rational expectations, the real rate of interest, and the natural rate of unemployment. *Brookings Papers on Economic Activity* 1973(2), 429–72.

Stiglitz, J.E. 1974. Alternative theories of wage determination and unemployment in LDC's: the labor turnover model. *Quarterly Journal of Economics* 88(2), 194–227.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–23.

How to cite this article

Phelps, Edmund S. "Phillips curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000081> doi:10.1057/9780230226203.1283

The New Palgrave Dictionary of Economics Online

Phillips, Alban William Housego (1914 – 1975)

C.A. Blyth

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

identification; inflation; Phillips curve; Phillips, A. W. H.; stabilization policy; unemployment

Article

Bill Phillips was born on 18 November 1914 into a farming family in Te Rehunga, near Dannevirke, in southern Hawkes Bay in the North Island of New Zealand, and died at Auckland, New Zealand on 4 March 1975. He came to economics after a career as an electrical engineer and following military service and imprisonment by the Japanese in the Second World War; in 1946 he became a Member of the Order of the British Empire for his military services. His rise in the profession was rapid. He was appointed an Assistant Lecturer at the London School of Economics in 1950 and to a Readership in 1954. In 1958 he became Tooke Professor, resigning in 1967 to take a Chair at the Institute of Advanced Studies, Australian National University in Canberra. A crippling stroke in 1969 forced his retirement and he lived in Auckland until his death. In his short career in economics he made major contributions to problems of dynamic stabilization, estimation and, most notoriously, empirical economics, where he gave his name to the ‘Phillips curve’.

Phillips's Ph.D. at LSE was on the problems of stabilizing or controlling an economy. Before this he had built a hydraulic model in perspex of a dynamic Keynesian-type economy and sold commercial versions of the model to academic and other institutions in Britain and the United States. (Some machines had a bottle of water named the Bank of England, used for ‘topping up’. Richard Goodwin is credited with introducing an accelerator in Cambridge's machine, but leakages were always a problem.) In a seminal paper in 1954 Phillips dealt with response lags and the problems they presented for stabilization policy. He distinguished between proportional, integral and derivative policies, depending on whether policy changes responded to current errors, cumulated deviations or rates of change of objectives. Optimal policy depends on the lag properties of the economy, and would consist of a mixture of proportional, integral and derivative components. Subsequent analysis of stabilization policy has used this scheme (for example, Meade, 1971).

This early work convinced Phillips that proper econometric modelling was a precondition for dynamic

stabilization. He turned to the problem of empirical description of the lag structures and their statistical estimation. All his later papers are concerned with the problems and difficulties of describing and estimating the dynamic relationships embodied in time forms of economic responses. His retreat to Canberra and Chinese economic studies has been seen by Lancaster (1979) as an acknowledgement that he found the problems, as he posed them, of estimating the relationships required for dynamic control beyond his capacity to solve. In Canberra, however, he continued to work on problems of identification (for example, 1968). In this later work he foreshadowed subsequent thinking (for example, that of Lucas) by explaining how the application of stabilization policy through a model results in the model becoming underidentified or, more generally, how policy changes relationships. The ‘Phillips dilemma’ remains: in the absence of adequate econometric modelling, stabilization policy is an empty box.

But before he left macroeconomics Phillips made in 1958 his epochal contribution of the Phillips curve, which mesmerized economists for the next decade and continues to attract attention. Responding to Dennis Robertson's criticism of the Keynesian mathematical model of his Ph.D. thesis, Phillips later used a relation between the rate of price change and capacity utilization, but without being able to give it any satisfactory empirical foundation. The lengthy time series of British wages produced by Henry Phelps Brown and Sheila Hopkins gave him the opportunity to experiment with the long series of British unemployment statistics from 1861 to 1957. What began as an attempt to derive a simple relationship between rate of change of wage rates and unemployment emerged as a nonlinear long term relationship with a complex short period lagged response. The famous paper is striking for the informal estimation method and the ad hoc theorizing and Phillips admitted an excessive haste to publish while also acknowledging that A.J. Brown had almost got the results earlier but without the lags (Blyth, 1975).

Apart from an enquiry into Australian statistics, Phillips did not enter into the subsequent international controversy over the theoretical and empirical foundations of the ‘Phillips curve’. It is necessary to read Phillips's original paper to understand its intentional exploratory character, and the deliberate absence of theoretical generalization. For an enquiry which in the eyes of, for example, Samuelson and Solow ‘closed’ the Keynesian system, it is remarkably but typically modest and tentative. Furthermore, on the controversial issue of the high cost of the trade-off between inflation and unemployment, Phillips refers briefly to the 5 per cent unemployment level necessary to maintain stable wage rates without expanding on it. If there were close intellectual connections with the Joan Robinson–Kalecki–Beveridge approach to the full employment–trade union problem, they are not disclosed.

Scientific problems with stabilization theory, family reasons and reaction to student revolt at LSE all may have contributed to Phillips's move to Australia, where he energetically began to develop Chinese economic studies. His interest in China had begun in the 1930s and he began to learn Chinese while a prisoner-of-war in Java. In the short period before he retired he saw the firm establishment of a Centre for Contemporary Chinese Studies in Canberra, while his final academic activity in Auckland was appropriately enough to start a lecture course in Chinese economic history. In 1974, on his 60th birthday, his colleagues and friends presented him with a subsequently published Festschrift (Bergstrom et al., 1978).

In a profession which accepts many strays from other disciplines, Phillips was outstanding. His formal education ended in New Zealand at the age of 15, and after apprenticeship as an electrician he qualified as an electrical engineer in London in 1938 after working in Australia and travelling to Britain via Japan and Siberia. An authentic hero of the Second World War, his introduction to economics was as he said through a poor degree in sociology. James Meade sponsored his hydraulic model, and Phillips's

‘launching’ is by tradition associated with a famous Robbins seminar in which he successfully explained his machine. A New Zealand influence and connection is not intellectually evident, but may have contributed to the willingness to ‘do it himself’. Who else but a New Zealander would have learnt his differential equations at a Queensland goldmine?

See Also

- Phillips curve
- Phillips curve (new views)

Selected works

1950. Mechanical models in economic dynamics. *Economica* 17, 283–305.

1954. Stabilization policy in a closed economy. *Economic Journal* 64, 290–323.

1956. Some notes on the estimation of time-forms of reactions in interdependent dynamic systems. *Economica* 23, 99–113.

1957. Stabilization policy and the time-forms of lagged responses. *Economic Journal* 67, 265–77.

1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25, 283–99.

1959. The estimation of parameters in systems of stochastic differential equations. *Biometrika* 46(June), 67–76.

1968. Models for the control of economic fluctuations. In Scientific Control Systems, *Mathematical Model Building in Economics and Industry*. London: Griffin.

Bibliography

Bergstom, A.R., Catt, A.J.L., Peston, M.H. and Silverstone, B.D.J., eds. 1978. *Stability and Inflation*. Chichester: Wiley.

Blyth, C.A. 1975. A.W.H. Phillips. *Economic Record* 51, 303–7.

Lancaster, K. 1979. A. William Phillips. In *International Encyclopaedia of the Social Sciences*, vol. 18. New York: Free Press.

Meade, J.A. 1971. *The Controlled Economy*. London: Allen & Unwin.

How to cite this article

Blyth, C. A. "Phillips, Alban William Housego (1914–1975)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000082> doi:10.1057/9780230226203.1282

The New Palgrave Dictionary of Economics Online

philosophy and economics

D. Wade Hands

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The literature on philosophy and economics has traditionally been divided into two areas: economic methodology, which connects economics and epistemology/philosophy of science, and the literature on economics and moral philosophy/ethics. Recent developments in both of these areas are discussed in detail.

Keywords

altruism; bargaining; capabilities approach to social welfare; consumer choice theory; critical realist research programme; economics of scientific knowledge; empirical macroeconomics; epistemology and economics; ethics and economics; evolutionary biology; experienced utility; experimental economics; falsificationism; free-rider problem; happiness; hedonism; history of economic thought; Human Development Index; Hume, D.; Hutchison, T.; interpersonal utility comparisons; justice; Kahneman, D.; material welfare school; methodology of economics; models; naturalism; neuroeconomics; ontology and economics; Pareto efficiency; philosophy and economics; philosophy of science; Popper, K.; positive–normative dichotomy; positivism; postmodernism; preferences; rational choice theory; Rawls, J.; Robbins, L.; Sen, A.; social contract; ultimatum game; utilitarianism; value judgements; welfare economics

Article

The essential interdependency of philosophical and economic ideas was a prominent feature of classical economics. Adam Smith was the author of *The Theory of Moral Sentiments* as well as *Wealth of Nations*. John Stuart Mill was an extremely wide-ranging scholar, as well known as the author of *A System of Logic* as of *The Principles of Political Economy*. And of course Karl Marx's *Capital* also drew on intellectual resources from economics, philosophy and a number of other fields. Classical political economy was deeply influenced by philosophy – different philosophies for different economists, but

influenced nonetheless – and ideas also flowed freely in the opposite direction, from political economy to various areas of philosophical inquiry.

This changed significantly in the first third of the 20th century. The abandonment of ‘political economy’ and the self-conscious development of ‘scientific economics’ coincided with a major change in the relationship between the two disciplines. Although philosophy never completely disappeared from economic theorizing, it systematically came to play a less and less obvious role. There are undoubtedly many reasons for this. Two of the more important include the overall professionalization of disciplinary economics and the general acceptance of a more narrow, positivist-inspired notion of legitimate ‘scientific’ inquiry. John Stuart Mill directed his arguments at the general educated public and wrote confidently about the ‘moral sciences’; by the first half of the 20th century fewer economists were doing the former and almost no professional economist would be comfortable doing the latter.

Although there were different versions of positivism, one common theme was that ‘meaningful’ discourse comes in only two forms: the synthetic knowledge of empirical science and the analytic knowledge of logic and mathematics. During the period of positivist dominance (roughly from the early 1930s through the 1950s), many, perhaps most, of the lines of inquiry that had previously travelled under the label of ‘philosophy’ – including, ethics, ontology, metaphysics, and aesthetics – were dismissed from the realm of meaningful discourse. Science ceased to be a generic category that included any rational, non-faith-based inquiry, and instead came to designate only the natural sciences (or modes of inquiry that follow the same scientific method). Economics clearly had scientific aspirations, and in such a regime fulfilling those aspirations required jettisoning the profession's old philosophical ways. Many of the significant developments in economic theory during the first half of the 20th century can be understood in precisely these terms: as an attempt to systematically discard the old metaphysical and utilitarian baggage, and replace it with more appropriate scientific concepts. Moral philosophy, for example, might still make an appearance in discussions about economic theory, but it almost always played a disparaging role: either to indict another theory for retaining some ethical residuum, or to emphasize that one's own theory was entirely free of such normative influences. Such an environment was certainly not conducive to forging new links between philosophy and economics, and for much of the 20th century very few were.

A particularly good example of the rejection of philosophy is the development of welfare economics during the second quarter of the 20th century. From the hedonism of many early neoclassicals to the so-called ‘material welfare school’ (Cooter and Rappoport, 1984) of Alfred Marshall and Arthur Pigou, welfare economics (and applied microeconomics in general) had traditionally been associated with utilitarianism: policy A was better than policy B if A increased total utility by more than B. During the 1930s, as a result of the work of Lionel Robbins (1952) and others, most economists came to view this type of ‘interpersonal’ utility comparison as unscientific and thus inappropriate for economic analysis. Moral values were simply raw, subjective or ‘emotive’ preferences that were not amenable to scientific analysis, and must therefore be kept out of economic science.

As economists moved away from the earlier utilitarian notions of ‘good’ economic policy, they increasingly turned to the Pareto criterion as an alternative evaluative standard. It was argued (and still is) that Pareto efficiency – an allocation of resources such that no one person can be made better off without making someone else worse off – does not require making interpersonal utility comparisons and is therefore an entirely appropriate standard for scientific economics. The most important theoretical results of modern welfare economics – the first and second fundamental theorems – are based on a direct

application of the Pareto criterion to questions about the welfare implications of competitive equilibrium. Although the norm-free credentials of Pareto efficiency have repeatedly been challenged (Blaug, 1980; Hausman and McPherson, 2006; Robertson, 1952), the standard interpretation among practising economists remains that such judgements, and thus any policy recommendations based on them, are fundamentally value free. But it is not necessary to take sides in the debate over whether Pareto efficiency is or is not an ethical criterion in order to recognize that the entire discussion is couched in terms of whether moral concepts are properly kept out of economic science, and to note that such a discussion does not provide a very fertile environment for the cultivation of new relationships between economics and moral philosophy.

Economic methodology has traditionally been the one exception to economists' general rejection of philosophy. Although ethics and metaphysics were shunned by economists, epistemology and philosophy of science were often consulted for guidance regarding the proper scientific method. This said, even within methodology the use of philosophical resources varied greatly from economist to economist. Some of the classical works in economic methodology (Milton Friedman, 1953, for example) hardly mentioned philosophy at all; others (Robbins, 1952, and Hutchison, 1938, for example) drew on selected aspects of the philosophy of science, while still others (Blaug, 1980; Samuelson, 1963) tried to apply the arguments of particular philosophers of natural science directly to economics. Thus, even in methodology economists focused on only a relatively small portion of the philosophical literature and employed even those resources in a less than systematic way.

Although the traditional methodological literature is both extensive and ongoing, it is not the focus of the following discussion. There are at least two reasons for this. First, this literature has been effectively surveyed in a number of contemporary works (Blaug, 1980; Caldwell, 1994; Hands, 2001; Hausman, 1992) and second, things have again changed. Since the mid-1980s there has been a renaissance in the interaction between economics and philosophy. The traditional approach to economic methodology continues to produce viable research, but economics and philosophy are also interacting in many other, new and important ways. Philosophy of natural science is no longer the only relevant set of philosophical ideas – ethics and ontology have both returned to the scene – and the intellectual dynamic is now one of bilateral exchange rather than economists simply borrowing ideas from one corner of the philosophical shelf.

In addition to the revival of the interplay between economics and philosophy there has been an increase in the traffic between economics and a number of other fields that compete for some of the same intellectual space that philosophy has traditionally occupied. For example, resources from the sociology of science and science studies (Mirowski, 2002; Sent, 1998; Weintraub, 2002; Yonay, 1998), the rhetoric of science (McCloskey, 1998), postmodernism (Ruccio and Amariglio, 2003), feminism (Ferber and Nelson, 2003; Nelson, 1996), and variety of other fields have provided new tools for the examination of (and often confrontation with) modern economic theory. Although these works frequently overlap with the literature on philosophy and economics, they also involve ideas sufficiently removed from disciplinary philosophy that they fall outside of the work considered here.

The discussion is divided into two parts; the first examines recent developments in the relationship between economics and scientific philosophy. Some of this work has much in common with traditional economic methodology, while other contributions approach the relationship in entirely new ways. In the interest of brevity, only five of the many possible areas of significant research are examined. The second

section examines the recent literature that combines economics and moral philosophy. Ethical questions are again back on the table, and an extensive literature has grown up relating various issues in moral philosophy to developments within economic theory. Some of this research challenges the received view of the relationship between economics and ethics established during the first half of the 20th century, while other parts of the literature develop totally new connections. Again, as with the methodological literature, only a few examples are discussed. The final section briefly considers some points of convergence between contemporary work on economics and epistemology and that on economics and ethics. Throughout the discussion, the emphasis is on microeconomics and rational choice theory (rather than, say, macroeconomics or econometrics).

Economics, epistemology, and philosophy of science

The first area of research to be examined goes back to Terence Hutchison (1938); it is the literature relating the philosophical ideas of *Karl Popper* (1965; 1968) to economics. Popper is best known as an advocate of falsificationism, a philosophy that has two main theses: one demarcating science from non-science and the other characterizing the growth of scientific knowledge. For a theory to be scientific it must be at least potentially falsifiable by empirical evidence (in Popperian language, be falsifiable by at least one empirical basic statement). Scientific knowledge grows as the scientific community rejects falsified theories and retains those that have survived attempted falsifications (that is, by ‘bold conjecture and severe test’). The body of accepted science at any point in time consists of all scientific theories that have survived such severe empirical tests. Elements of such a methodology were present in Hutchison (1938), and elaborated in more detail in his later work. The position has been most articulately defended in the methodological writings of Mark Blaug (1980). Although many economists continue to endorse a falsificationist approach to methodological questions, there is also an extensive critical literature on the subject (Caldwell, 1991; 1994; Hands, 1993; Hausman, 1988; 1992).

If the only research connecting the Popperian tradition to economics was the literature on falsificationism, then the subject would probably not be included in this discussion of recent developments. But that is not the case. During the last few decades the Popperian tradition has engaged economics on a number of different fronts, and currently consists of much more than just the literature defending (or criticizing) falsificationism (Caldwell, 1991). At least three other developments should be noted. The first involves Popper's own brief discussion of economic methodology (Popper, 1994). This work is controversial because Popper's statements about economics – and social science more generally – differ from what he said about the (falsificationist) methodology of natural science. The second concerns the so-called ‘critical rationalist’ interpretation of Popper's overall philosophical programme: an interpretation that goes back in the economics literature to Kurt Klappholz and Joseph Agassi (1959), but has its best contemporary representation in the work of Lawrence Boland (1997). Supporters of critical rationalism argue that Popper's main philosophical contribution was not (empirical) falsificationism but rather a more general view of the growth of knowledge through open debate and rational criticism – of which falsification by empirical evidence is simply one, albeit a very important, special case. Although the discussion of critical rationalism has remained primarily an in-house debate among Popperians, it has much broader implications because it opens the door to characterizing the growth of knowledge as a product of particular social institutions rather than as the result of following

fixed methodological rules, a view that has become increasingly important in general philosophy of science. Finally, there has been an extensive discussion of the work of Popper's student Imre Lakatos (1970) and his ‘methodology of scientific research programs’ (Backhouse, 1997; Blaug and De Marchi, 1991; Latsis, 1976). Economists have focused on two different aspects of Lakatos's work: his historical framework for understanding the evolution of economic research programmes (his concepts of hard core, protective belt, and so on) and his specific methodological framework for appraising scientific research programmes as progressive or degenerating. Even though there exists a critical literature on both of these issues, the Lakatosian framework has produced important case studies and also encouraged a re-examination of the general relationship between economic methodology and the history of economic thought.

The second area to consider involves the revival of interest in *ontology* and *metaphysics* in the philosophy of economics. There now exists a burgeoning literature on ‘economics and ontology’ (Mäki, 2001), something that would have been next-to-impossible only a few decades ago. During the heyday of positivism any mention of such (occult) notions as essential natures, underlying causal powers, or ontological necessity all but disappeared from academic discussions about economics. Ontological discussion continued to some extent within certain heterodox, particularly Marxist, research programmes, but among mainstream economists, even philosophically informed ones, such concepts had no place in professional discourse. Although many things have contributed to this revival, three issues seem to be particularly important.

One factor contributing to this ontological renewal has clearly been the development of the ‘critical realist’ research programme, an anti-empiricist approach to the philosophy of social science that focuses on uncovering the hidden underlying causal mechanisms at work in social life. The most prolific defender of critical realism within economics has been Tony Lawson (2003), and his writings have generated an extensive secondary literature. A second factor involves changes that have taken place within the philosophy of natural science. Although there were many reasons for the decline of positivist-inspired philosophy of science, one of the most important was the perception that serious problems had developed within the Humean-inspired ‘empiricist’ component of the programme. Although debate continues about whether the founders of positivism were actually as empiricist as the standard view suggests (Michael Friedman, 1999), it is certainly clear that the programme was perceived that way by both critics and supporters, and that it was this aspect of the programme that was most effectively targeted by the criticism that descended upon it in the last quarter of the 20th century. Some of the efforts to reconfigure our reigning philosophical conceptions in light of these developments – particularly about scientific laws (Cartwright, 1989) and causality (Hoover, 2001) – draw directly on insights from economics. Finally, the literature on economics and ontology has benefited from recent changes that have taken place within the discipline of economics itself. A discipline that is more willing to entertain theoretical pluralism is more likely to be willing to entertain philosophical, even ontological, pluralism as well. The bottom line is that ontology and metaphysics are back and they are opening up a number of new (and renewed) lines of inquiry relevant to the philosophy of economics.

The third set of changes to consider involves border crossings between economics and certain other scientific fields – *cognitive science*, *neuroscience*, and related disciplines – that have influenced the recent literature on the philosophy of mind. This literature is relatively new and rapidly growing, so much so that no appellative convention has emerged. Until such a consensus has been reached it is

perhaps best to be inclusive and simply call it the literature on ‘the mind, the brain, rationality, agency and economics’. Examples would include such disparate works as Davis (2003), Glimcher (2003), Mirowski (2002), and Ross (2005). Although the arguments of the various contributors are quite different, there is some agreement about the main issues, as well as about the requirements for any adequate approach to these issues. These requirements concern consistency with recent developments in fields such as cognitive science, neurophysiology and artificial intelligence. The common concern is the core rational choice framework of modern economics: explaining economic behaviour as the outcome of rational constrained optimization of well-ordered preferences. Consumer choice theory is the paradigm case of such an explanatory strategy, but it is standard throughout economics (traditionally microeconomics, but increasingly macroeconomics as well).

Such rational choice explanations have recently been subject to a variety of criticisms. Some of these relate to the abundance of contrary empirical evidence that has appeared in the experimental literature – in both economics and psychology (Kahneman and Tversky, 2000) – and some of it has to do with the well-known philosophical problems associated with ‘intentional’ or ‘folk psychological’ (belief-desire-action) explanations (Rosenberg, 1992). Although much of the impetus comes from critiques of rational choice theory, this does not mean that all of the resulting literature advocates doing away with it. Some authors clearly do, but others interpret these recent theoretical developments as a way of defending standard practice. In either case, whether its authors defend or attack rational choice theory, the literature embodies a fundamental change in the rules of engagement. It is too early to know how it will develop, or the various turns it might take along the way, but it is clear that both in its use of resources from other disciplines and in its overall mode of argumentation it has moved economics and philosophy in a substantially different direction.

The fourth area to consider overlaps substantially with previous section on minds, brains, cognitive science, and such. It concerns the tendency towards ‘*naturalism*’ in epistemology and philosophy of science. The standard interpretation of both positivist and falsificationist philosophy of science puts ‘philosophy before science’ in the sense that philosophers first decide what scientists must do to produce theories that are cognitively significant – constitute legitimate scientific ‘knowledge’ – and then evaluate specific scientific practices on the basis of this philosophical analysis. Naturalism – and there are many specific versions, but here we consider its most generic form – reverses this relationship. Instead of starting with a priori philosophical analysis about what scientific knowledge must be, naturalism starts with science, that is, the best current scientific practice, and uses this best practice to inform our epistemological inquiries about knowledge in general. Much of the philosophical literature discussed in the previous section – the literature that employs contemporary cognitive science and neuroscience in the investigation of knowledge in general – is naturalist in this sense. Such naturalism raises a host of questions, particularly questions about how it is possible to have a ‘normative’ philosophy of science, one that explains what ought to be done in science, when the ‘philosophy’ in question is based on descriptions of scientific practice. Such questions are the subject of much current debate and do not have easy or simple answers. Fortunately, such answers are not required for a discussion of how naturalism has affected research in the philosophy of economics.

Much of the recent research in the history and philosophy of economics is broadly naturalist in spirit. Naturalism informs some of the work on traditional methodological questions (Hausman, 1992) as well as research in general philosophy of science that draws heavily on economics (Cartwright, 1989). It also

provides the backdrop for a number of recent studies on specific research programmes within economics, including the role of models (Morgan, 1999; 2001), the practice of empirical macroeconomics (Hoover, 2001), and the development of experimental economics (Guala, 2005). Although the boundary that separates such naturalist-inspired research from similar work informed by science studies is somewhat blurred, it is often possible to categorize a particular piece of work as primarily one or the other. If the main question is the philosophical justification of the particular economic tool or theory – even if the standards for such justification are naturalistically or historically grounded – then the research is in the spirit of naturalistic philosophy; but if the explanation of the acceptance or rejection of particular economic tools or theories is based primarily on the influence of social, political, or individual (non-epistemic) interests, then it falls more into science studies.

The final category of literature to be considered, the *economics of scientific knowledge*, reverses the standard relationship between a particular social science like economics and the philosophy of natural science. As discussed above, the traditional relationship between philosophy of science and economics has been that philosophy comes first (laying the foundations for knowledge), economic methodology then translates those philosophical ideas into the context of economic science, and finally particular economic theories are appraised on the basis of the methodological rules so acquired. In the economics of scientific knowledge this process is reversed. Certain areas of economic theory – for example, industrial organization (IO) economics – examine how the institutional organization of a particular industry contributes to economic efficiency. Shifting this type of reasoning from the production of goods and services to the production of scientific knowledge is the basis for one way of thinking about the economics of scientific knowledge. The scientific community has a particular institutional structure; if the goal of this scientific ‘industry’ is the production of (reliable, justified, …) scientific knowledge, then an obvious question is the degree to which the industrial organization contributes to the growth of knowledge (that is, epistemic efficiency). Since the goal is the growth of knowledge within the community, it might be the case that all of the individual scientists following the same methodological rule may not be the optimal way to arrange the available epistemic resources; perhaps the greatest production of scientific knowledge comes about as the result of a ‘cognitive division of labor’ (Kitcher, 1993) rather than methodological homogeneity. It is easy to see how such an approach opens up new ways of thinking about the growth of scientific knowledge, and does so by employing economic theory as a resource (in the spirit of naturalism) to address general questions about the growth of knowledge and the optimal design of scientific institutions.

It can be argued that such research on the economics of scientific knowledge goes back to Charles Sanders Peirce in 1879 (Wible, 1998), but regardless of its origins it has expanded rapidly during the last few years, with contributions coming from both economists and philosophers (Dasgupta and David, 1994; Goldman and Shaked, 1991; Kitcher, 1993; Wible, 1998). As one might expect, the literature has also generated a variety of critical responses (Hands, 1997; Mirowski, 2004). In addition, many other contributions to the economics of scientific knowledge are quite different from the version of epistemic IO discussed above (Mirowski and Sent, 2002). But in all of its various forms this work clearly represents a significant change in the interaction between economics and philosophy of science.

Economics and moral philosophy

One of the many changes that have taken place in the relationship between economics and moral philosophy has been a re-examination of economists' traditional stance on the 'positive–normative dichotomy'. This change is sufficiently complex that it is examined in two parts. First, there has been a substantive reconsideration of the general place of 'the normative' within the science of economics (where 'normative' does not necessarily concern ethics), and second, ethical norms are increasingly being considered in the causal explanation of economic phenomena.

Enforcing the prohibition against value judgements in economics requires maintaining a strict dichotomy between positive statements about what 'is' and normative statements about what 'ought to be'. These two issues – dichotomization and prohibition – are certainly related, but they can also be separated. The first asserts that a dichotomy should be maintained – 'ought' should be kept separate (and cannot be derived) from 'is' – while the second asserts that separate is not equal – things on the normative/'ought' side of the dichotomy have no place within scientific economics. Although the first (dichotomy) is necessary for the second (prohibition), it is clearly not sufficient; one could argue, as, say, Mill and Marshall did, that there is a difference between positive and normative economics, and yet also leave room for a version of normative economic science.

Debate over the strict dichotomy and the prohibition against deriving 'ought' from 'is' has a long history. It was popularized by David Hume in the 18th century, labelled the 'naturalistic fallacy' by G.E. Moore early in the 20th century, and is the subject of a long and contentious debate within philosophy (Putnam, 2002). Although many economists have been concerned with these issues, the one who probably played the most important role in the profession's ultimate establishment of the principle of strict separation was Lionel Robbins. Robbins (1952, p. 149) endorsed a strict dichotomy – 'Propositions involving the verb "ought" are different in kind from propositions involving the verb "is"' – but he went beyond mere separation to prohibition, advocating complete exclusion of normative analysis from scientific economics. In particular, he criticized the normative welfare economics of the Marshallian school because it relied on 'interpersonal' utility comparisons. For Robbins, the normative economics resulting from such analysis was 'illegitimate' and 'lacking in scientific foundation' (1952, p. 141).

By and large Robbins's position on these matters has become the conventional wisdom among practising economists as well as among most contributors to the methodological literature. Where methodological commentators often differ is not over whether normative concerns should be kept out of scientific economics but rather on the factual question of whether most practising economists have actually done so. For example, two well-known contributors to economic methodology, Mark Blaug (1980) and Milton Friedman (1953), both endorse the dichotomy and prohibition, but differ on the question of whether the economics profession has in fact been successful at keeping normative propositions out of its scientific practice.

The core of standard microeconomics continues to be rational choice theory; economic agents are assumed to have well-ordered preferences and make optimal choices given those preferences and the various constraints they face. Such rational choice explanations involve two parts: preferences (goals/ends) are assumed to be rational (that is, well-ordered, satisfying conditions such as transitivity and completeness) and the agent is presumed to act in the most efficient way to achieve those given ends (that is, to act in an instrumentally rational way). Philosophers have traditionally called such rationality 'practical rationality' to distinguish it from 'theoretical' or 'epistemic' rationality. In general practical rationality involves what it is rational to do, or at least intend to do, while theoretical or epistemic rationality involves what it is rational to believe.

The literature on practical rationality leads to a very different characterization of the positive–normative dichotomy than the one standard in economics. Although most practising economists continue to view rational choice theory as a positive theory about the behaviour of economic agents (at least under ideal conditions), most philosophers writing on the subject consider it a normative theory in the sense that it involves norms and obligations. Practical rationality, and thus rational choice theory as a particular instantiation of it, is a normative theory because it tells agents what they 'ought' to do in order to act rationally. In the contemporary philosophical literature this view is often associated with the work of Donald Davidson (2001), but it has a long history and continues to be debated (Searle, 2001).

Philosophers have certainly not closed the book on the question of how a theory of practical rationality could be a descriptive theory, or how, if it is normative, it might relate to associated descriptive theories. The point is simply that it is increasingly the case, in both philosophy and economics, that the discussion of rational choice theory starts from the presumption that it is a particular instantiation of the theory of normative rationality, and as a result, the description of actual economic agents – whether in the laboratory or in 'the wild' – is coming to be seen as something to be compared with, or reconciled with, this theory of normative rationality. It is still possible to discuss the ways in which rational choice theory is or is not an adequate scientific theory, but the starting point of the discussion has changed substantially (Hausman and McPherson, 2006; Mongin, 2006; Ross, 2005).

The second change to be examined requires us to step back from the previous discussion of normative rationality. Suppose we use 'normative' to mean 'ethically normative', and view rational choice theory as a strictly positive, not a normative, theory, then *there are still a number of arguments for increasing the normative content of positive economic science*. Although these arguments are less of a challenge to the conventional wisdom, they still constitute a potentially significant change in the relationship between economics and moral philosophy.

Many of the arguments for increasing the (ethically) normative content of economic science come from the experimental literature, either experimental economics or experimental psychology. Researchers in these fields often reaches similar conclusions about the behaviour of the agents they study, although they differ regarding experimental protocols (particularly the role of cash payments) and how such results are to be interpreted (as a critique of rational choice theory or as a critique of the standard assumptions of rational choice theory). One of the systematic results of the literature has been that moral beliefs matter to decision making in experimental environments, and are sufficiently important that such morality often provides better empirical predictions than self-interested rational choice. For example, one of the earliest counter-intuitive experimental results was the tendency for individuals to over-contribute to (that is, not free ride on) public goods (Isaac, Walker and Thomas, 1984). One explanation for this over-contribution is an ethical 'taste for fairness'. Another example involves the 'ultimatum game', a game where a self-interested rational agent should offer the smallest possible amount to the other player. The experimental evidence indicates that individuals do not generally behave as rational choice theory suggests, but rather give the other player a more 'fair' distribution. Since rational choice theory allows for the possibility of 'moral' (or otherwise non-self-interested) preferences, these results do not constitute a direct falsification of the core theory of rational choice (Guala, 2005), but they certainly do challenge profession's traditional view of the positive and the normative. Instead of ethical norms interfering with the scientific investigation, these are cases where including ethical beliefs in the analysis improves the theory's descriptive accuracy.

The next two developments shift attention away from the positive–normative dichotomy but still

challenge key features of the view passed down from Robbins and the ordinal revolution. According to the standard history of demand/choice theory, three (good) things happened as the theory of consumer choice progressed from the hedonistic cardinalism of the late 19th century, through the ordinal revolution of the 1930s, and on to the revealed preference/consistency interpretation in contemporary textbooks. First, all vestiges of hedonistic psychology were finally abandoned; second, all interpersonal comparisons of utility were eliminated; and finally, these changes brought about a steady improvement in the scientific foundations of the theory.

In recent years there has been serious reconsideration of at least two of these aspects of choice theory: hedonism and the impossibility of interpersonal utility comparisons. There have, of course, always been critics of the move away from hedonism and interpersonal utility comparisons (Harsanyi, 1955; Robertson, 1952), but the goal of such criticism has traditionally been to defend utilitarian ethics as the normative basis for economic policy. Appeals on such grounds certainly continue, but in recent years support for a return to hedonism and interpersonal utility comparisons has come from a number of new directions. Although these two topics are closely related, it is useful to discuss them separately.

Hedonism in rational choice theory is the idea that an agent's preference for a particular bundle of goods is based on the psychological feeling of satisfaction the agent receives when the bundle is purchased or consumed. This is clearly the notion of utility present in 19th century utilitarianism, and, even though it has been replaced by a non-hedonistic notion of preference in modern economics, it is still heard in casual conversation and in the classroom. One criticism of the move away from such psychological hedonism – a criticism from an earlier generation as well (Little, 1957; Robertson, 1952) – is that the move enervated the theory's ability to provide any real explanation of observed behaviour. Although this criticism has been a theme in a number of important recent studies (Davis, 2003; Giocoli, 2003; Mandler, 1999), these authors do not generally recommend returning to a version of the earlier hedonist doctrine. On the other hand, some recent research does reach such neo-hedonist conclusions.

One research programme that endorses a return to hedonism is the work of the 2002 Nobel Prize winner in economics, the experimental psychologist Daniel Kahneman (Kahneman and Tversky, 2000). Although the research of Kahneman and his associates is wide-ranging, and perhaps not every participant would support this particular aspect of the programme, the argument for a return to hedonism – what is called ‘experienced utility’ – has been a key aspect of Kahneman's approach (Kahneman, 1994; 1999; Kahneman, Wakker and Sarin, 1997). There are two main parts to the argument for experienced utility, one philosophical and the other based on recent changes in our scientific tools. The philosophical argument is simply that weakening the positivist grip on experimental practice has opened the door to a number of new and fruitful possibilities; the more practical argument is that new tools for measuring experienced utility are becoming, and will continue to become, more available over time.

> The methodological strictures against a hedonistic notion of utility are a relic of an earlier period in which a behavioristic philosophy of science held sway. Subjective states are now legitimate topic of study, and hedonic experiences such as pleasure, pain, satisfaction or discomfort are considered open to useful forms of measurement. (Kahneman, 1994, p. 20)

Paralleling such neo-hedonist arguments from experimental psychology are similar arguments from economics, particularly the literature endorsing ‘happiness research’ as a source of useful, and

measurable, data for applied economic theory (Frey and Stutzer, 2002). Economists appear to be more willing than psychologists to accept measures of happiness based on survey data, but the hedonistic themes are very much the same. Finally, there is a literature on the relationship between economic rationality and evolutionary biology that also suggests a hedonistic characterization of utility is scientifically appropriate (Robson, 2001). It does not seem, as yet, that these newer interdisciplinary arguments defending hedonism have been integrated into the more traditional defence of utilitarian-based ethics as the basis for economic policy, but it is an obvious next step and is therefore extremely important for the relationship between economics and moral philosophy.

To turn from hedonism to a fourth change in the recent economics and ethics literature, there are similar (and often overlapping) arguments endorsing the revival of *interpersonal utility comparisons* in economics. Although the two issues – hedonism and interpersonal comparisons – are closely related, it is important to keep them separate. Hedonism is about feelings of pleasure and pain, and interpersonal comparisons are about having a common unit of comparison between the preferences of different agents (Mandler, 1999). One can compare the current running through two different electrical appliances, but it is reasonable to conclude that such appliances do not ‘feel’ anything; similarly, two individuals could possess subjective, even cardinal, feelings about various goods and yet there would exist no way for a third party to measure or compare those feelings.

As in the case of hedonism, there have been consistent defenders of the legitimacy of interpersonal comparisons within economics, even when it was out of favour with most of the profession; many of these defenders came from the Marshallian tradition (Pigou, 1920), but that is not exclusively the case (Harsanyi, 1955; 1982). Often the argument was simply that economists should start with the observable facts of everyday life, and the fact is that humans make interpersonal comparisons all the time (Little, 1957). Such defences continue, but in addition – again, as in the hedonism case – a number of new arguments are being made that draw on a range of interdisciplinary resources.

One source of evidence for interpersonal utility comparisons comes from recent research on neuroeconomics, part of the literature on ‘the mind, the brain, rationality, agency and economics’ discussed above. Neuroeconomics is a research programme that combines contemporary neuroscience and economics in the investigation of the microfoundations of decision making (Glimcher, 2003). Imaging studies from neuroeconomic research suggest that humans have the capacity to both represent the mental states of others and to empathize, that is, share the feelings of others. These abilities, it is argued, were selected for in human evolution because they ‘enable people to predict others’ behavior and, therefore, help them meet their individual goals’ (Singer and Fehr, 2005, p. 343). Neuroeconomics is not the only source of such arguments for the reliability, and survivability, of interpersonal utility comparisons. Similar arguments have also been made in the literature on the philosophy of mind. For example Alvin Goldman (1995) combines a reliabilist approach to the philosophy of science with various arguments from cognitive psychology to make the case for individuals having the ability to mirror, or simulate, the mental states of others in a reliable way, including interpersonal utility comparisons. In addition to the obvious support such research provides for moral theorizing within the utilitarian tradition, it also seems to provide a naturalistic explanation for the sympathy that played such an important role in Adam Smith's moral theory. At the very least, moral, economic and cognitive theorizing are simply different parts of a single intellectual exercise – as they were for Smith and Mill – rather than being hermetically isolated, as they were for most of the 20th century.

The fifth and final research to examine carries us outside the boundaries of the previous topics. Whether one is considering rational choice theory as normative theory, using moral preferences to explain observed behaviour in experimental economics, or defending hedonistic psychology and interpersonal utility comparisons, the discussion continues to be broadly within the research programme that identifies welfare with the satisfaction (or feelings received from the satisfaction) of individual preferences. In all of these cases, regardless of how much the recent literature conflicts with the mainstream view on such matters, the bottom line is still that individuals have preferences (hedonistic or not) and the individual 'good' is to have those preferences satisfied. But not all moral and political philosophy, even all that involves economics, follows this tradition.

John Rawls's *A Theory of Justice* (1971) is arguably one of the most important books on moral philosophy of the 20th century; it, and the philosophical discussion surrounding it, set the stage for many of the changes discussed above. Although Rawls's theory of justice falls squarely within the contractarian tradition – defining 'justice' as a property of the social contract that would emerge from the interaction of rational self-interested agents – he imposed strong restrictions on the context in which such contractual bargaining takes place; the decisions must be made in 'the original position' behind a 'veil of ignorance'. The principles of justice are those that would emerge from the bargaining of rational agents if those agents did not have any information about the position they would ultimately occupy (professional, class, gender, level of health, …) within the society governed by the contract, or even about what their preferences would be. Rawls goes on to argue for specific rules of justice that would emerge from such a context – including the much-debated 'difference principle' – but it is possible to separate his general approach to the question of justice from his specific distributional answers.

Although it is impossible to discuss the extensive literature surrounding Rawls's work in the space available here, it is important to consider the related contribution of one economist. The economist is Amartya Sen, the 1998 winner of the Nobel Prize in Economics. Sen has long been a critic of standard rational choice theory (Sen, 1977), but his critical writings have come to be overshadowed by his own *capabilities* approach to social welfare and related issues (Sen, 1985; 2002). The core idea of the capabilities approach to social welfare is to focus on the capabilities that people have, that is, on the things that people are effectively able to do or be – the functionings they are free to achieve – rather than on the satisfaction of individual preferences. Such capabilities are obviously multifaceted; they depend on the person's mental and physical characteristics as well as his or her social context and opportunities. One may have the capability to ride a bike, to find meaningful work, to express oneself artistically, or to participate in the governance of one's society; alternatively, one may have none, or only a few, of these capabilities. For Sen, such capabilities should be the proper focus for both the analysis of social welfare and the theory of economic development. The point of both welfare and development is to increase the capabilities of the population – to give them the freedom and opportunity to be better able to live the kind of life they find valuable. This, of course, does not rule out increasing the quantity of goods and services they have available, but it is at best only part of the story. In this sense Sen's approach actually moves us farther away from the traditional preference-based notion of social welfare than Rawls. Rawls's concept of justice is still based on the notion of a distribution of preference-satisfying goods (albeit primary social goods), while Sen shifts the focus away from individual preferences towards freedom and functioning.

Needless to say, Sen's approach has many critics, but his work has also generated an extensive supporting, extending and implementing literature. An important example of support and extension is

Martha Nussbaum's (2000) research on women and development, which provides a specific list of the most important ‘central human capabilities’; an example of implementation is the United Nations Development Program's Human Development Index, which builds on Sen's capabilities approach. Undoubtedly the capabilities literature will continue to evolve, but, regardless of the eventual shape it takes, it is an important contribution that has substantially changed the discourse on economics and moral philosophy.

Convergences

In closing, it is important to note the change that has taken place in the general way that various questions in philosophy and economics are approached in the recent literature compared with the way they were approached, at least by economists, for most of the 20th century. The traditional view considered ‘the philosophical’, whether it be epistemology or ethics, as something ‘out there’ with respect to economics. In the case of epistemology it was appropriate to seek methodological advice from philosophers about the character and practice of science, but the border crossing remained sporadic and one-way. In the case of ethics, the traditional view was simply to be aware of such ideas in order to prevent them from influencing the discipline's scientific practice.

Things have indeed changed. This is not to say that there is any consensus about specifics in the contemporary literature on either economics–epistemology or economics–ethics – in fact there has been an explosion of diversity and debate, and as such there is far less consensus on such matters than among economists in the past – but rather that the style of discussion has changed in both fields, and in a sense converged. Although a much longer list could be constructed, there seem to be three features of the debates in philosophy and economics discussed above that were effectively absent from the previous discussions: the interdisciplinarity, the naturalism, and the two-way relationship involved. The literatures discussed above all draw on a wide range of resources: economics and disciplinary philosophy certainly, but also cognitive psychology, neuroscience, the history and sociology of science, ideas from evolutionary biology, and a host of others. They are also broadly naturalist in focus in the sense that the relevant philosophical questions – whether epistemological or ethical – are on equal footing with the science, social or natural, that is employed in, and constrains, the philosophical discussion. Finally, and perhaps most obviously, work in philosophy and economics is much more of a two-way street. It is not simply that a shelf of scientific philosophy is ‘applied’ to economic methodology, or that a shelf of moral philosophy is used to cull normative concepts from economic science, but rather that economic notions of agency, choice, efficiency and equilibrium now condition the discussions in philosophy in the same way that alternative philosophical ideas, and ‘normativity’ more broadly, are increasingly involved in discussions within economic theory. On the one hand, these are substantive changes; on the other hand, such interconnections were present in the work of Smith, Mill and others. Perhaps these changes in the relationship between philosophy and economics are not so new after all; perhaps what needs explanation is not recent developments but the aberration of the 20th century.

See Also

- conventionalism
- epistemic game theory: complete information
- ethics and economics
- experimental economics
- explanation
- falsificationism
- happiness, economics of
- instrumentalism and operationalism
- interpersonal utility comparisons (new developments)
- Methodenstreit
- methodology of economics
- positivism
- scientific realism and ontology
- theory appraisal
- value judgements.

Bibliography

Backhouse, R. 1997. *Explorations in Economic Methodology: From Lakatos to Empirical Philosophy of Science*. London: Routledge.

Blaug, M. 1980. *The Methodology of Economics: Or How Economists Explain*, 2nd edn. Cambridge: Cambridge University Press, 1992.

Blaug, M. and De Marchi, N., eds. 1991. *Appraising Economic Theories: Studies in the Methodology of Scientific Research Programs*. Aldershot: Edward Elgar.

Boland, L. 1997. *Critical Economic Methodology: A Personal Odyssey*. London: Routledge.

Caldwell, B. 1991. Clarifying Popper. *Journal of Economic Literature* 29, 1–33.

Caldwell, B. 1994. *Beyond Positivism: Economic Methodology in the Twentieth Century*, 2nd edn. London: Routledge.

Cartwright, N. 1989. *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Cooter, R. and Rappoport, P. 1984. Were the cardinalists wrong about welfare economics? *Journal of Economic Literature* 22, 507–30.

Dasgupta, P. and David, P. 1994. Toward a new economics of science. *Research Policy* 23, 487–521.

Davidson, D. 2001. *Essays on Actions and Events*, 2nd edn. Oxford: Oxford University Press.

Davis, J. 2003. *The Theory of the Individual in Economics*. London: Routledge.

Ferber, M. and Nelson, J. 2003. *Feminist Economics Today: Beyond Economic Man*. Chicago: University of Chicago Press.

Frey, B. and Stutzer, A. 2002. What can economists learn from happiness research? *Journal of Economic Literature* 40, 402–35.

Friedman, Michael. 1999. *Reconsidering Logical Positivism*. Cambridge: Cambridge University Press.

Friedman, Milton. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Giocoli, N. 2003. *Modeling Rational Agents: From Interwar Economics to Early Modern Game Theory*. Cheltenham: Edward Elgar.

Glimcher, P. 2003. *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. Cambridge, MA: MIT Press.

Goldman, A. 1995. Simulation and interpersonal utility. *Ethics* 105, 709–26.

Goldman, A. and Shaked, M. 1991. An economic model of scientific activity and truth acquisition. *Philosophical Studies* 63, 31–55.

Guala, F. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Hands, D. 1993. *Testing, Rationality, and Progress: Essays on the Popperian Tradition in Economic Methodology*. Lanham, MD: Rowman and Littlefield.

Hands, D. 1997. Caveat emptor: economics and contemporary philosophy of science. *Philosophy of Science* 64(Proceedings), S107–S116.

Hands, D. 2001. *Reflection Without Rules: Economic Methodology and Contemporary Science Theory*. Cambridge: Cambridge University Press.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics and Interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Harsanyi, J. 1982. Morality and the theory of rational behaviour. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams. Cambridge: Cambridge University Press.

Hausman, D. 1988. An appraisal of Popperian economic methodology. In *The Popperian Legacy in Economics*, ed. N. De Marchi. Cambridge: Cambridge University Press.

Hausman, D. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Hausman, D. and McPherson, M. 2006. *Economic Analysis, Moral Philosophy, and Public Policy*. Cambridge: Cambridge University Press.

Hoover, K. 2001. *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Hutchison, T. 1938. *The Significance and Basic Postulates of Economic Theory*. London: Macmillan.

Isaac, J., Walker, J. and Thomas, S. 1984. Divergent evidence on free-riding: an experimental examination of possible explanations. *Public Choice* 43, 113–49.

Kahneman, D. 1994. New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics* 150, 18–36.

Kahneman, D. 1999. Objective happiness. In *Well-Being: The Foundations of Hedonic Psychology*, ed. D. Kahneman, E. Diener and N. Schwarz. New York: Russell Sage.

Kahneman, D. and Tversky, A., eds. 2000. *Choices, Values, and Frames*. Cambridge: Cambridge University Press.

Kahneman, D., Wakker, P. and Sarin, R. 1997. Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics* 112, 374–405.

Kitcher, P. 1993. *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford: Oxford University Press.

Klappholz, K. and Agassi, J. 1959. Methodological prescriptions in economics. *Economica* 26, 60–74.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the Growth of Knowledge*, ed. I. Lakatos and A. Musgrave. Cambridge: Cambridge University Press.

Latsis, S. (ed.) 1976. *Method and Appraisal in Economics*. Cambridge: Cambridge University Press.

Lawson, T. 2003. *Reorienting Economics*. London: Routledge.

Little, I. 1957. *A Critique of Welfare Economics*, 2nd edn. Oxford: Oxford University Press.

Mäki, U. (ed.) 2001. *The Economic World View: Studies in the Ontology of Economics*. Cambridge: Cambridge University Press.

Mandler, M. 1999. *Dilemmas in Economic Theory*. Oxford: Oxford University Press.

McCloskey, D. 1998. *The Rhetoric of Economics*, 2nd edn. Madison: University of Wisconsin Press.

Mirowski, P. 2002. *Machine Dreams: Economics Becomes a Cyborg Science*. Cambridge: Cambridge University Press.

Mirowski, P. 2004. The scientific dimensions of social knowledge and their distant echoes in 20th-century American philosophy of science. *Studies in History and Philosophy of Science* 35, 283–326.

Mirowski, P. and Sent, E.-M., eds. 2002. *Science Bought and Sold*. Chicago: University of Chicago Press.

Mongin, P. 2006. Value judgments and value neutrality in economics. *Economica* 73, 257–86.

Morgan, M. 1999. Learning from models. In *Models as Mediators*, ed. M. Morgan and M. Morrison. Cambridge: Cambridge University Press.

Morgan, M. 2001. Models, stories and the economic world. *Journal of Economic Methodology* 8, 361–84.

Nelson, J. 1996. *Feminism, Objectivity and Economics*. London: Routledge.

Nussbaum, M. 2000. *Women in Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.

Pigou, A. 1920. *The Economics of Welfare*, 1st edn. London: Macmillan.

Popper, K. 1965. *Conjectures and Refutations*, 2nd edn. New York: Harper and Row.

Popper, K. 1968. *The Logic of Scientific Discovery*, 2nd edn. New York: Basic Books.

Popper, K. 1994. Models, instruments, and truth: the status of the rationality principle in the social sciences. In *The Myth of the Framework: In Defense of Science and Rationality*. London: Routledge.

Putnam, H. 2002. *The Collapse of the Fact/Value Dichotomy and Other Essays*. Cambridge, MA: Harvard University Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Robertson, D. 1952. *Utility and All That and Other Essays*. London: Allen and Unwin.

Robbins, L. 1952. *An Essay on the Nature & Significance of Economic Science*, 2nd edn. London: Macmillan (2nd edn first published in 1935).

Robson, A. 2001. Why would nature give individuals utility functions? *Journal of Political Economy* 109, 900–14.

Rosenberg, A. 1992. *Economics – Mathematical Politics or Science of Diminishing Returns?* Chicago: University of Chicago Press.

Ross, D. 2005. *Economic Theory and Cognitive Science*. Cambridge, MA: MIT Press.

Ruccio, D. and Amariglio, J. 2003. *Postmodern Moments in Modern Economics*. Princeton, NJ: Princeton University Press.

Samuelson, P. 1963. Problems of methodology – discussion. *American Economic Review* 53, 231–6.

Searle, J. 2001. *Rationality in Action*. Cambridge, MA: MIT Press.

Sen, A. 1977. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6, 317–44.

Sen, A. 1985. *Commodities and Capabilities: The Professor Dr. P. Hennipman Lectures in Economics*, vol. 7. Amsterdam: Elsevier Publishing.

Sen, A. 2002. *Rationality and Freedom*. Cambridge, MA: Harvard University Press.

Sent, E.-M. 1998. *The Evolving Rationality of Rational Expectations*. Cambridge: Cambridge University Press.

Singer, T. and Fehr, E. 2005. The neuroeconomics of mind reading and empathy. *American Economic Review* 95, 340–5.

Weintraub, E. 2002. *How Economics Became a Mathematical Science*. Durham, NC: Duke University

Press.

Wible, J. 1998. *The Economics of Science: Methodology and Epistemology as if Economics Really Mattered*. London: Routledge.

Yonay, Y. 1998. *The Struggle for the Soul of Economics*. Princeton, NJ: Princeton University Press.

How to cite this article

Hands, D. Wade. "philosophy and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000084> doi:10.1057/9780230226203.1285

The New Palgrave Dictionary of Economics Online

physiocracy

G. Vaggi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advances; agrarian capitalism; agriculture; *avances annuelles* ; *avances foncières* ; *avances primitives* ; Baudeau, N.; Cantillon, R.; Chalmers, T.; circular flow; circulating capital and fixed capital; circulation of commodities; civil society; class; Colbert, J.-B.; Diderot, D.; Du Pont de Nemours, P. S.; Ephémérides du citoyen ou chronique de l'esprit national; fixed capital; Forbonnais, F. V. D. de; free trade; Galiani, F.; Gournay, Marquis de; Gray, J.; hoarding; input–output analysis; laissez-faire; Le Trosne, G. F.; luxury; Mercier de la Rivière, P.-P.; Mirabeau, V. R., Marquis de; natural laws; net product; Pattullo, H.; peasants; Petty, W.; Physiocracy; population growth; productive and unproductive labour; property rights; Quesnay, F.; rent; reproduction schemes; single tax; Smith, A.; Spence, W.; Sraffa, P.; surplus; *Tableau économique* ; Turgot, A. R. J.; underconsumptionism; Vauban, S. le P.; Voltaire, F. M. A. de; wealth

Article

The Physiocrats lived and worked in France in the middle of the 18th century. The name derives from the title of a collection of some of the most important writings of their master François Quesnay, *Physiocratie, ou constitution naturelle du gouvernement le plus avantageux au genre humain* published in 1767 by P.S. Du Pont de Nemours. The term Physiocracy indicates the importance ascribed by these authors to natural forces, and derives from the Greek: *phýsis*, nature, and *kràtos*, power. The Physiocrats can be regarded as the first school of economists. They acted as an organized group of thinkers who intended to influence the French government's economic policy. They were accused of being sectarian because of their strict allegiance to the economic theories and opinions of their master, Quesnay. He provided the most important and original ideas, Victor Riqueti, Marquis de Mirabeau, was his first disciple, and included among the best known Physiocrats were Du Pont de Nemours, l'Abbé Nicolas Baudeau, Le Mercier de La Riviére and François Guillaume Le Trosne. One should also mention Henry Pattullo, an Irishman, who was deeply influenced by Quesnay's early articles (see Hecht, 1958, vol. 1, p. 257). These French authors can be regarded as the 'inner circle' of the Physiocrats.

Another group of writers, sometimes confused with the Physiocrats, was Vincent de Gournay and his followers, the most famous of whom is Turgot. Gournay was appointed *Intendant* of commerce in 1751, and, like Quesnay, favoured laissez-faire. However, Gournay and his school never followed the Physiocratic programme and, in particular, disagreed on such important points as the idea that agriculture was the only productive sector of the economy.

Physiocracy covers a period of 20 years, from 1756 when Quesnay published his first economic articles in the *Encyclopédie* of Diderot and D'Alembert, until 1777 when Le Trosne's book appeared.

After a period of relative prosperity at the end of the 17th century and the beginning of the 18th century France experienced many bad years, mainly due to the backwardness of her agriculture. Often the Physiocrats recalled the age of Sully, Prime Minister to Henry IV, as the golden period of French agriculture and of the whole country. But now farmers were poor and could not implement the best methods of cultivation; the fiscal system was inefficient and unjust; and there were many different taxes and duties, both on the peasants themselves and on their products (Loménie, 1879, vol. 2, p. 218). For instance, one had to pay an excise in order to take products from one province to another. Trade in agricultural products was greatly hindered by these impediments to the free circulation of commodities. There were also taxes that were levied on the number of people in the family – the various forms of *capitation*.

On top of these duties there were taxes which had to be paid to the Church and to the King, the *dîme* for the Church and the *taille* for the government. These taxes were levied on the revenue of lands, but their collection was extremely inefficient. The government used to sell the right to collect the *taille* in one province to some wealthy people who became tax collectors. This was the system of the *ferme général*, and was opposed by the Physiocrats because the peasants were oppressed by the *fermiers généraux*, who, having paid the government in advance, tried to make as much money as possible. They were allowed to keep all the taxes for themselves, and thus the King received much less money than that paid by the peasants.

There was a huge public deficit, and at the same time the peasants and the farmers were deprived of the fruits of agriculture.

The fiscal systems and the various barriers to the domestic and foreign trade of agricultural products discouraged the farmers from improving farming and agricultural productivity. During the first half of the 18th century there were many years of misery and famine (see Meek, 1962, p. 46). According to Quesnay, during that period the population of France decreased from 24 million to 16 million (INED, 1958, vol. 2, p. 506). He was too pessimistic (Eltis, 1984, p. 39), but certainly French agriculture was unable to sustain a growing population. The Physiocrats compared the farming conditions in France with those in England, where farmers were rich and productivity was very high (INED, 1958, vol. 2, pp. 440–41). The backward economic situation in France was made worse by the almost continuous wars, which absorbed human and financial resources. The Physiocratic movement must be examined in the light of this situation of recurring economic crises. The purpose of Physiocracy was to bring changes to certain characteristics in the French economy and in the political system of the *ancien régime*. They were a group of reformers, who tried to convince the rulers and the sovereign that some changes were needed to make the country more wealthy and politically stronger.

The history of the school can be divided into three periods: the years in which the main ideas appeared from 1756 to 1760, mostly in the works of Mirabeau and Quesnay; from 1760 there was a period of

almost three years silence; the third period, from 1764 to 1777, saw a flourishing of writings and enterprises, thanks to the younger Physiocrats. Quesnay published his first economic articles in the *Encyclopédie*, to which he had been asked to collaborate on matters of agrarian economics. In 1756 *Evidence* and *Fermiers* appeared and in 1757 he published *Grains*. These works present most of the new ideas of the school and in particular they stress the view that agriculture is the most important sector in the economy. This is the corner stone of the Physiocratic theory of the nature and causes of national wealth. Quesnay identified the entire social product with the annual output of agriculture, and maintained that neither industry nor trade could increase the country's wealth, a doctrine which won him many enemies. In 1757 Quesnay met his first disciple: the Marquis de Mirabeau. This member of the French aristocracy had become famous because of his book *L'ami des hommes, ou traité de la population*, in which he stated that the wealth of a country depended upon the size of her population; like the title of his book he was called 'the friend of mankind', because of his liberal and reformist views. On July 1757 Quesnay and Mirabeau met at Versailles, where Quesnay was one of the King's physicians. Quesnay convinced Mirabeau that the products of land were more important than people because they secured the survival of the peasants and their families, who had to be regarded as the most important element in the economy (Weulersse, 1910, vol. 1, pp. 55–6). Mirabeau was won over to the cause of Physiocracy, and in a couple of years he wrote many important works. In 1758 Quesnay wrote his famous *Tableau économique*, which was printed in three different editions between the end of 1758 and the first months of 1759. The analytical structure of Physiocratic economics was an enormous step forward. French society was divided into three main classes: the landlords – including the King and the Church – the farmers, and finally the artisans; the last two groups were respectively in charge of agricultural and industrial production. The *Tableau* outlines the main features of the process of circulation of commodities, at the end of the productive process, and gives a precise definition of the means of production and the net product. To illustrate his main economic ideas Quesnay used some rather obscure diagrams, which nevertheless greatly impressed the Versailles aristocrats. To make the *Tableau* more understandable to the public Mirabeau wrote some explanation in three further books of his *L'ami des hommes*, which were published between 1758 and 1760, and in which Quesnay's influence is very strong. Always in strict collaboration with the master, Mirabeau wrote a treatise on one of the major economic problems of the time: the reform of the fiscal system. The *Théorie de l'impôt* appeared in 1760 and presented one of the Physiocrats' most famous proposals: the single tax on rent. Fiscal reform must abolish all taxes and duties which are levied either on the peasants or on their products. This tax burden is one of the main reasons why cultivation cannot become profitable. The financial needs of the Kingdom must be met by a single general tax, which has to be paid in proportion to the net product of agriculture. This recommendation was the logical consequence of Quesnay's division of the social product into two parts: capital and surplus. The capital consists in the *avances* for farming, and must be preserved to maintain the same level of agricultural output. Any form of taxation falling on farmers' advances, *les avances*, would reduce the amount of capital employed in agriculture, and this would have disastrous effects on the whole country. Thus, only the surplus is really disposable for taxation, because it does not affect reproduction of output.

The largest part of agricultural net product accrued to the landlords in the form of rent. Quesnay and Mirabeau's single tax on the net revenue meant the abolition of all the fiscal privileges of the ruling classes, the Church, and the aristocracy. Mirabeau and Quesnay tried to convince the nobles that in the following years their rents, net of taxes, would be much higher than before. In fact, the farmers, freed

from the previous fiscal burden, would invest more money in the cultivation of land. The productivity of agriculture would rise, as would the surplus. But these arguments did not impress the nobility. Moreover, Mirabeau also violently attacked the tax collectors. The state must collect its taxes without the intermediation of these merchants and businessmen. But for many members of the aristocracy and the merchant bourgeoisie the role of tax collector meant power and wealth. Their reaction to Mirabeau's book was so strong that he was imprisoned for a few days, and then exiled to his countryside estate for some months (Loménie, 1879, vol. 2, p. 226).

Here ended the formative period of Physiocratic school. Quesnay and Mirabeau did not publish anything for two and a half years. Du Pont wrote that Mirabeau's misfortune delayed the development of enlightenment (Du Pont, 1769, *Ephémérides*, vol. 2). Quesnay and Mirabeau spent this period of silence working towards a new book which was to be a fundamental text for Physiocratic doctrine. It appeared at the end of 1763 in three volumes, with the title *Philosophie rurale ou économie générale et politique de l'agriculture*. 1763 saw renewed interest in Physiocratic ideas; the government accepted the principle of free trade for corn inside France, which was one of the main reforms advocated by the Physiocrats. New followers joined Quesnay and Mirabeau; Du Pont de Nemours became an enthusiastic propagator of Physiocracy and in 1764 published a pamphlet in favour of free foreign trade for French corn. The mid-1760s were the period when Physiocracy had most influence on French economic policy.

In 1764 Du Pont became chief editor of a famous periodical, the *Journal de l'agriculture, du commerce et des finances*, which became an important vehicle for Physiocratic propaganda for some years. In the same year two new followers joined the school: Le Trosne and the less famous Saint Péravy. In 1765 Mercier de La Rivière was converted to Physiocracy. The school was now powerful enough to try to gain more influence on political and economic matters. After six years during which he mostly collaborated with Mirabeau's work Quesnay wrote again on his own, and from 1765 to 1768 he published many important articles intended to explain the principles of Physiocracy further, and to defend them from growing attack. At least three articles must be mentioned: 'Le droit naturel', which was written in 1765, the 'Analyse de la formule arithmétique du Tableau économique', which was written in 1766, both of which appeared in the *Journal de l'agriculture*. The latter is particularly important because it provides an easy explanation of the *Tableau économique* and in fact became its best-known version. The third work is the 'Dialogue sur les travaux des artisans', published in 1767 in the *Ephémérides*. Here Quesnay defended his view that only agriculture was capable of yielding a net product, while industrial activity was sterile because it only replaced the value of the raw materials and necessaries which had been used up in production.

During this period other Physiocrats contributed to the development of the school. In 1767 Mercier de La Rivière published his book *L'ordre naturel et essentiel des sociétés politiques*, in which he elaborated the political doctrines of Physiocracy. In the same year Du Pont published a collection of some of Quesnay's work, entitled *Physiocratie*, where this term appears for the first time. The Physiocrats met every Tuesday in Mirabeau's palace, and became a political group (Weulersse, 1910, vol. 1, p. 132).

The abbé Baudeau too became a Physiocrat. In 1767 Baudeau founded an influential periodical, the *Ephémérides du citoyen* in which several Physiocrats collaborated. In the same year Du Pont started losing power in the *Journal de l'agriculture*. Since it was important for the Physiocrats to publish in a friendly periodical, they tried to win over Baudeau. After a few months of discussion the *Ephémérides* became the official periodical of Physiocracy. Many powerful people looked favourably on this group of intellectuals. Among them were Traudaine de Montigny and, above all, Turgot.

Physiocracy was also exerting some influence abroad. Catherine II invited Mercier de La Rivière to St. Petersburg to spread the new ideas. The Margrave of Baden also became a Physiocrat and exchanged letters with Du Pont. At home the Physiocrats had good relationships with the *encyclopédistes*; Diderot personally admired Mercier, but never shared the Physiocratic opinion that the wealth of a country derives from agriculture. The school also received support from the *Sociétés d'agriculture*, coalitions of wealthy farmers who tried to defend their interests and gain power over landlords. For these bourgeois farmers the doctrines of the Physiocrats were a powerful instrument of propaganda and political influence on the government.

The growing prestige and power of the Physiocrats also gained them new enemies including many of the aristocrats, and all those merchants who had exclusive trading privileges granted by the government.

In 1767 and 1768 many authors wrote against the Physiocrats. Grimm, Forbonnais and Mably, a disciple of Rousseau, attacked different aspects of Physiocracy. In his pamphlet *L'homme à 40 écus*, Voltaire ridiculed the Physiocrats' fixation with numerical examples. The *encyclopédistes* became less friendly towards the Physiocrats. Some critics accused Quesnay and his disciplines of trying to mitigate the most unjust aspects of the *ancien régime* and to improve its inefficiencies only to prevent any major change in the French political system.

One particular point of Physiocratic theory came under attack at the end of the 1760s: the doctrine of the exclusive productivity of agriculture and the sterility of industry. (Notice that the Physiocrats regarded as productive not only the cultivation of soil but all the activities directly connected with agriculture, such as 'grasslands, pastures, forests, mines, fishing' (Kuczynski and Meek, 1972, p. i).) Nobody questioned the importance of agriculture, but the attacks focused on the view that trade and above all industry were regarded as sterile occupations. This was the crucial point of the Physiocrats' definition of wealth, and all their policy measures depended upon this doctrine. The liberalization of the corn trade, the reform of the fiscal system, and the attack on expenditures on luxury goods all depended upon the Physiocrats' identification of national wealth with agricultural production. Many contemporary authors rejected the idea that national wealth could only be increased through land.

Veron de Forbonnais, a former pupil of Gournay, defended the productive role of commerce and industry (see Weulersse, 1910, vol. 1, pp. 121–2). The strongest attack on the doctrine of the exclusive productivity of agriculture came from a Neapolitan priest, the abbé Ferdinando Galiani. With the help of Diderot, at the end of 1769, he published the *Dialogues sur le commerce des bléds*, in which he used brilliant prose to ridicule the supposed superiority of agriculture over industry. Galiani gave very simple and straightforward examples to show that increases in productivity were much more likely to take place in industrial production than in agriculture. Good and bad weather does not influence the output of manufacturing, and the advantages of the division of labour which derive from increases in capital stock are not limited by the existence of a fixed amount of soil (see Galiani, 1770, p. 142). The decisive element which undermined the influence of Physiocratic views on government policy was growing opposition to the deregulation of the corn trade. From 1763 the commercial policy for the products of land, and in particular corn, had become one of the main economic issues in French society. We have already seen that the Physiocrats had some success with the 1763 declaration of the free circulation of corn inside France. In July 1764 an edict authorized the exportation of corn under certain circumstances. According to Quesnay laissez-faire in domestic and foreign trade was designed to favour the circulation of corn and increase its demand. A free trade policy implied the abolition of all the rights and the rules which hampered the corn market (Mirabeau, 1764, vol. 2, p. 343). The merchants and all the people who

had been granted some ‘exclusive privileges’ in corn trade were damaged, because they lost the position as middlemen between consumers and producers (INED, 1958, vol. 2, p. 532). The Physiocrats wanted to favour direct contact between consumers and cultivators. The final outcome of a laissez-faire policy would have been an increase in the price received by farmers without damaging consumers, thanks to the squeeze, or the abolition, of the earnings of all intermediate agents. Free corn exportation would have further contributed to sustaining its demand and its price on the French market. The establishment of a *bon prix* for primary commodities was meant to raise farming's profitability (INED, 1958, vol. 2, p. 529). Farmers would have been able to make new investments, the productivity of French agriculture would have risen and the gross and net output of the primary sector would have been larger than before. This was the Physiocratic road to welfare and prosperity for the French Kingdom (Mirabeau, 1760, vol. 2, p. 143).

In the second half of the 1760s the price of corn rose, but unfortunately this happened both in the wholesale and in the retail markets. It is difficult to ascribe this rise to free corn exports: most likely the price increases were due to a series of bad harvests. But the Physiocrats were accused of having contributed to the worsening of the living conditions of the French people, for whom corn was the most important consumption good. In 1768 there were popular uprisings against the high price of corn both in Paris and in the countryside.

Part of public opinion began to consider Physiocratic theory as a dangerous attack on poor people, and some parliaments, in particular those of Paris and Rouen, called for the reintroduction of the restrictions on the corn trade. Between 1768 and 1770 there were many discussions for and against laissez-faire for primary commodities; the public and the government itself came gradually to oppose the Physiocratic views. At the end of 1769 the abbé Terray, one of the fiercest opponents of Physiocracy, was appointed *contrôleur général*, a sort of minister in charge of all economic matters. There were more uprisings and more declarations by provincial parliaments against the free exportation of corn.

During this period relationships between the Physiocrats and the *encyclopédistes* deteriorated notably. Grimm made fun of the *secte* of the *philosophes économistes*, who were accused of presenting a reactionary doctrine, designed to favour the landlords and the rural classes against the people in the cities (Weulersse, 1910, vol. 1, pp. 230–1). In this hostile climate Turgot and Morellet rejected the invitation to join the Physiocrats. After a period of irregular publication the *Ephémérides* were put under censorship. At the end of 1770 corn trade legislation was completely changed and strict regulations were introduced both in foreign and domestic trade. The period of political influence of Physiocracy was almost over, and Physiocratic doctrines rapidly disappeared from public debate and the political arena. A final glimpse of the Physiocrats’ impact on French economic policy was due to Turgot. On becoming *contrôleur général* in 1774 he restored internal free trade in corn, with the exception of Paris (Groenewegen, 1977, p. xxxii). But this policy had many powerful opponents, and caused Turgot's fall after two years.

At the end of the 1760s Physiocracy was on the wane. After 1768 Quesnay wrote nothing on economic matters, lost interest in economic problems, and spent the last years of his life studying geometry; he died in 1774. In the 1770s there were only two works by Physiocrats. In 1772 Du Pont published an *Abrégé des principes de l’économie politique*. In 1777 Le Trosne published his book *De l’intérêt social, par rapport à la valeur, à la circulation, à l’industrie et au commerce intérieur et extérieur*. These attempts to renew interest in Physiocracy failed to influence either the policies of the government or

discussions in French society.

Quesnay and his followers must be regarded as part of that cultural phenomenon which was the French Enlightenment. Many authors had already pointed out France's disastrous economic circumstances in the first half of the 18th century. Moreover, most of these writers did not limit their denunciations to the unjust and inefficient aspects of French society, but extended their investigations to the problem of the origin and nature of civil societies and the analysis of the best rules and laws which should regulate the relationships between individuals.

Quesnay and his disciples contributed to the French enlightenment. They concentrated their efforts on the economic and social reforms which were needed to make France more efficient. But they were not very much interested in the analysis of the fundamental principles of the civil societies and the role of subjects and the state; such issues have no prominence in their works. The only exception is the book by Mercier de La Rivière, which is mainly dedicated to the analysis of the political system. In general the Physiocrats never questioned the existence of the absolute monarchy and the political organization of the *ancien régime*. This is one of the main reasons why they were accused of being too hesitant in the defence of individual rights against state power. Montesquieu's *L'Esprit des lois* was one of the philosophical works which most influenced the Physiocrats; this book appeared in 1748. A year later Rousseau published his *Discours*, with which the Physiocrats were much less in agreement.

In the first half of the century, several authors had already analysed the economic and social conditions in France, paying particular attention to the agricultural sector. In different ways these writers can be considered as the forerunners of Physiocracy. We have already seen that at the beginning of the 17th century, French agriculture was prosperous thanks to Sully, Henry IV's Prime Minister. The years of Louis XIV were marked by Colbert's attempt to favour industrial activities by keeping the prices of subsistence goods low. At the end of the 17th century there was a reaction to Colbert's policy and the role of agricultural production was again emphasized. Among the authors who influenced the Physiocrats Vauban and Boisguillebert should be recalled. In 1707 Vauban published *Dîme royale*; he said that a single tax on agricultural output was the best solution to France's fiscal problems.

In 1695 Boisguillebert published the *Détail de la France*, a collection of statistical information on the French economy, and in 1707 his *Dissertations sur la nature des richesses, de l'argent et des tributs* appeared. He stressed the importance of agriculture among various economic activities, and above all described the production and exchange of commodities in terms of a circular flow, a sort of self-regenerating circuit. Two other works which deserve mention are Melon's *Essai politique sur le commerce* (1734) and Herbert's *Essai sur la police des grains* (1754). But of the writers who exerted a major influence on Physiocracy, a special place is occupied by Cantillon. Other British authors were well known in France in those days, for instance Child, Tucker and Hume, but Cantillon's impact on Physiocratic theory was much deeper.

At the beginning of the 1740s Mirabeau had a copy of Cantillon's *Essai sur la nature du commerce en général*, which he regarded as the fundamental text on economic matters, an opinion shared by Quesnay. Mirabeau published Cantillon's *Essai* in 1755. In many ways the Physiocrats followed Cantillon's general approach to economic analysis. Cantillon gave a framework in which to build up a theoretical model of the working of the whole economic system. Economics would no longer be a subject for pamphleteers, merchants and practical men, but would become a topic of separate theoretical speculation. Practical matters would be examined in terms of the theories' fundamental principles. Cantillon's analysis left clear marks on Physiocratic economics, such as, for example, his classification

of the people of a kingdom into three main classes: landlords, entrepreneurs and workers. His analysis of the distribution of income was related to these classes; he spoke of the farmers' 'three rents', which make up the value of the products and each rent is the income of one class (Cantillon, 1755, p. 43). Like the Physiocrats Cantillon emphasized the productive role of the farmers as entrepreneurs. Finally, Cantillon considered expenditures of revenue as the most important element in the determination of the prices of commodities and the level of activity of the sectors other than agriculture. It is thanks to the landlord's expenditures of their revenues that these activities can exist. Through Cantillon the Physiocrats were also influenced by Sir William Petty.

With regard to Physiocratic theory it must be remarked that almost all the main contributions are due to Quesnay. As to their philosophical views they believed that civil societies were only a mirror of natural order. The Physiocrats believed that societies are characterized by the existence of laws which govern the relationships between individuals. These natural laws can be studied, and their knowledge provides the foundation for the proper administration of the country. It must be noticed that the Physiocrats' attitude towards natural laws and natural order is somewhat different from most of their French contemporaries and from Adam Smith. Natural laws operate quite independently of men's will (INED, 1958, vol. 2, p. 526), but at the same time they are not so powerful as to be ignored. These laws have been inscribed in nature by God himself (Mirabeau, 1764, vol. 2, pp. 9–11; INED, 1958, vol. 2, p. 934), but their working can be hampered and their effects can be modified by unwise ruling of society and by powerful social groups. Therefore natural laws do not necessarily overwhelm men's actions, and civil societies cannot be analysed as if they were a mechanical system which always gives the same results. The Physiocratic concept of natural order is a peculiar mixture of objective laws and of socio-historical modifications. Natural laws exist, and can be studied and precisely singled out, but there is also room for active human intervention. This view of the natural order has far reaching implications. The Physiocrats believed that societies evolve through definite specific stages (Meek, 1976, pp. 72, 99). But this evolutionary process can be stopped for long periods. They regarded England as the country where natural law displayed its positive effects, and which reached the highest stage of economic development. However, in France civil laws and historical traditions prevented the full unfolding of natural laws and the country was still in a backward condition. Thus, the Physiocrats did not take a deterministic approach to the study of societies, even if they believed in the existence of objective natural laws. Natural order is a sort of normative situation which describes the features of an ideal society.

How can these natural laws be discovered? Quesnay wrote an article entitled *Evidence* in which he maintained that the laws of natural order reveal themselves in day-to-day events. The Physiocrats were also influenced by Descartes, a fact which helps to explain their belief in knowledge through evidence. In some way natural laws seem to be inborn in men, and this is why the system of natural order should be clear to everyone.

Which are the fundamental principles of natural order? Here too the Physiocrats' answer shares some features of contemporary French culture but also presents some peculiarities. Quesnay and Mercier de La Rivière contributed to the development of the philosophical and political views of the school. In 1765 Quesnay wrote *Le droit naturel*, where he mentions the natural rights of men, which, however, are discussed mainly in relation to the economic features of society. Thus, freedom implies the abolition of privileges and regulations in all markets. Free competition must rule in the labour market as well as in domestic and foreign trade; people must be entirely free to decide how to spend their revenues. For the Physiocrats freedom meant universal competition, and was regarded as the basis for the increase in

private and public wealth.

They regarded private property as a fundamental right of men, but by this they meant that land ownership was part of the natural order of societies, and the King was considered to be co-owner of all French soil. But the Physiocrats also emphasized the importance of guaranteeing the farmers and their families the ownership of the capital employed in agriculture and the fruits of farming.

Private ownership excludes the possibility of equality between men; indeed development of the economy will cause more inequalities. Differences among people are necessary in order to have an efficient economic system, capable of yielding a high net product. Therefore the political structure of a country reflects its economic and social circumstances. For the Physiocrats the major forces which explain historical changes in societies must be sought in their economic structure. This economic interpretation of history (Meek, 1962, p. 376) underlines the fact that economic systems are based on the existence of different social groups which have separate economic functions. The Physiocrats distinguish French subjects into three classes: the landlords, including the King and the Church who represent the First and Second Estate; the people working in agriculture; and the industrial workers. This tripartite distinction is a hybrid since it is based partly on intersectoral differences and partly on property relationships. But in Physiocracy there is also a more detailed class analysis. In agriculture there are both farmer entrepreneurs and salaried peasants; with some ambiguity the same distinction between employers and employees exists in the industrial sector too. Then there are the merchants and all the people related to trade, and here too a whole class is identified with a sector of the economy. However unsatisfactory this approach may be, it was to be extremely important in the development of economic theory. First, following Petty, Boisguillebert and Cantillon the Physiocrats consider the economy as a system which is made up of different social groups, and which tends to reproduce both its economic and social relationships. Second, these classes are defined according to their role in the process of production and circulation of commodities. These two features are typical of the whole of classical political economy. Of course, the main limitation of the Physiocratic analysis of classes is the fact that they tend to identify social groups with the sectors of the economy, even if there are also hints of a distinction based on political and economic power relationships.

The Physiocrats' concept of natural order deeply affects their political views. In general they argue that the principles of political order must accomplish those of the natural order. The particular way in which this connection between the two orders comes about is through the form of government which they call *despotisme légal*. The supreme authority is that of the absolute hereditary monarchy which does not need to be legitimatized by the subjects. Of course this view was criticized by many authors of the time. According to the Physiocrats the only authority was that of the sovereign and that came directly from God. The King was also the natural owner of all the territory; *despotisme* was also patrimonial; the King was also the highest tutor of all forms of property. He was a legal despot, because he guaranteed security and freedom in property. Property is the key notion in the foundations of a political order. But according to the Physiocrats the authority of the King was moderated by the fact that he had to exert his power as an enlightened sovereign. By this the Physiocrats meant that the King had to be aware of natural laws and had to favour their implementation in civil society. The King's knowledge of natural laws was the decisive element which had to secure the existence of appropriate civil laws and just administration. No confrontation could exist between the sovereign and his subjects because the King, properly instructed about natural order, knew that his interests coincided with those of citizens.

The Physiocrats envisaged only two limits to the King's power. On the one hand there was public

opinion, which was also instructed in the principles of natural order, so that the people could react to a situation where the King ignored natural laws. On the other hand the fact that the King was the owner of the whole country did not entail exploitation of his subjects. Customs and habits about the fiscal system could not be modified by the sovereign's own decision alone.

A final peculiarity of the Physiocrats' political thinking is their view that only agriculture produces a surplus. In an agricultural country like France the merchants and all the owners of monetary and financial wealth are not part of the nation because their interests are in opposition to those of the state. The only true citizens are the landowners, the wealthy cultivators, and the other people directly linked to agricultural production; the artisans of the industrial sector were somehow tolerated. It is clear that Physiocrats aimed at a political system based on the alliance of all social groups linked to agriculture and the King.

The distinctive feature of Physiocratic economics was the doctrine of the exclusive productivity of agriculture; only activities directly linked to nature could yield a net product over costs. To justify their views Quesnay and the Physiocrats used many different arguments. Agriculture was superior to other economic sectors because it produced the raw materials and the necessities for all other occupations. The subsistence of all people could only come from farming (INED, 1958, vol. 2, p. 775). Industrial and commercial activities could exist only because the peasants were producing more foodstuffs than was required for their own subsistence. Moreover, France had been endowed by nature with a large and fertile territory and was surrounded by countries whose soil was much less suitable for farming and who were potential buyers of French products (Le Trosne, 1777, p. 988; INED, 1958, vol. 2, pp. 600–1).

The fiercest attacks by Physiocracy concerned the view that industrial and commercial activities were sterile. The Physiocrats believed that in all trading activities there was only an exchange of commodities of equal value, but these values had already been produced elsewhere.

All merchants and middlemen who operated in 'resale trade' were a burden to society, since they had to be maintained without adding anything to national wealth (INED, 1958, vol. 2, p. 947; Mercier, 1767, p. 278). The Physiocrats saw that some traders were making large monetary fortunes, but this was not a proof of their productiveness; on the contrary, this was the result of a violation of natural laws. Merchants could become rich thanks to unequal exchanges due to exclusive trading privileges. These regulations contradicted the natural principle of free and unobstructed competition in all markets. Industrial activities simply transformed the products of agriculture into different types of commodities, whose exchange values had already been determined (INED, 1958, vol. 2, pp. 496, 865). The sterility of industry was then explained by the fact that, according to the Physiocrats, the value of its product was equal to the value of its expenses and there was no net product left.

The profitability of farming is the most important requirement for the accumulation of capital in agriculture. Hence commercial policy must be designed to sustain the exchange value of the products of land. Free trade was the main way to raise the prices of primary commodities and induce the farmers to reinvest their profits in farming (INED, 1958, vol. 2, p. 602).

It is important to notice the Physiocrats considered laissez-faire instrumental in the establishment of favourable trading conditions for French farmers. Quesnay and his disciples were not in favour of a generalized free trade, and they were not particularly interested in the commercial conditions of manufactures; their only aim was the achievement of high exports of primary products. They looked to a positive balance of trade for French agriculture, since France should have become the granary of Europe. Moreover, Physiocrats regarded foreign trade as necessary only because the French domestic market

was too small and too poor to guarantee the profitable sale of French corn (INED, 1958, pp. 848–9). With a larger domestic market there would be no need to export corn.

Quesnay was quite aware of the important role of markets; a large consumption was necessary to sustain the prices of agricultural commodities. The Physiocrats believed that there was no lack of potential demand for corn, since it was a fundamental item. The main problem of French agriculture was not the lack of potential consumers but the lack of effective consumption (INED, 1958, pp. 528, 963). The exchange value of corn was affected by the number of effective consumers and by their wealth (p. 824): these were the true causes which determine the price of corn. The demand of those people who were not rich enough to pay for corn at its proper price was of no interest for the economy, according to the Physiocrats. They argued that landlords should spend most of their revenues in the purchase of agricultural products to increase the effective demand for French foodstuffs. Landlords were the social class which received most of the surplus, as rent, and all activities depended on the expenditure of this revenue.

The Physiocrats noted that the way in which revenue is spent influences society's economic structure. For instance, if the landlords buy many primary commodities and few manufactures, agriculture grows at a faster pace than industry (Kuczynski and Meek, 1972, p. 12). Of course the Physiocrats were in favour of high consumption of agricultural products which they called *luxe de subsistence*, and were against the purchase of industrial goods, *luxe de décoration* (Baudeau, 1767, pp. 190, 217). The Physiocrats attacked luxury because they wanted to encourage the profitable sale of agricultural products. They also opposed savings and the hoarding of money which would end up in monetary stocks to be lent at interest (Mirabeau, 1764, vol. 2, p. 343). Monetary and financial fortunes were not a true form of wealth, but represented a deduction from the process of circulation of agricultural commodities.

The concept of the net product is the Physiocrats' main contribution to economic theory. This notion is related to that of advances, a term they used to indicate the means of production. The social product must include all the goods which make up the advances, and for each of them the quantity produced must be at least equal to the quantity which has been used as input.

Physiocratic analysis of the different types of advances is the first classification of the means of production, or capital, in the history of economic theory. The *avances foncières*, or land advances, included all the operations necessary to prepare a piece of land for farming. *Avances annuelles* are another important type of advances, this time annual ones. They are made by farmers and consist of products which must be invested in cultivation at each productive cycle because they are completely consumed during the process of production. These commodities include raw materials and necessaries which allow the peasants and their families to work during the year, but some interpreters of Physiocracy maintain that they also include some manufactured goods (Meek, 1962, pp. 274–5; Eltis, 1984, pp. 29–31). Annual advances are a typical kind of circulating capital.

The original advances, *avances primitives*, are made up of instruments and equipment which last for more than one year; they also include livestock (Eltis, 1975, p. 189). All these commodities must be regarded as fixed capital lasting for many years (INED, 1958, vol. 2, p. 798). In fact Quesnay indicated that the average life cycle of the *avances primitives* lasted ten years. Productivity increases are closely related to capital accumulation (ibid., pp. 427 ff.). A prosperous economy is characterized by large-scale farming, where agriculture employs a large stock of *avances primitives*. This view of the ideal economic system has been called 'agrarian capitalism', since agriculture is the most advanced capitalist sector

(Hoselitz, 1968).

According to Quesnay, in ideal agricultural production the value of the fixed capital must be five times that of the annual capital and is assumed to be ten ‘milliards’. Given an annual rate of decay of ten per cent, the farmer must repay a fixed amount of capital equal to a half of the circulating capital. Therefore, the overall *réprises*, or returns, which make up the value of all the means of production annually consumed is given by the sum of the whole annual advances plus ‘1 *millard livres*’ depreciation of *avances primitives* (Meek, 1962, p. 154).

At the end of the 1770s the political reactions to the economic policy suggested by the Physiocrats caused a decline in their intellectual influence. Echoes of Physiocratic economics survived in some European countries such as Russia, Poland, Germany, and Tuscany, and in the United States.

But Quesnay and his followers left important marks in the history of economic theory. At the end of the 18th century and at the beginning of the 19th century several British economists looked quite favourably on many ideas of the Physiocrats. In different ways, John Gray, William Spence and Thomas Chalmers defended the superiority of agriculture over industrial activities (Meek, 1962, pp. 345 ff.). Because of the way in which they stressed the importance of demand and consumption in sustaining economic activities, the Physiocrats were also regarded as forerunners of underconsumption theories. Lack of consumption and the excess of expenditure on luxury goods could cause economic crises. Thus, the Physiocrats recognized the possibility of economic breakdown. From this point of view Physiocracy can be related to Sismondi and Malthus (Meek, 1962, pp. 313 ff.). The *Tableau économique* does not only describe the necessary economic relationships between some economic magnitudes, but it can also indicate why and how the ideal conditions of production could break down. Quesnay himself provided several examples of *Tableau* ‘in disequilibrium’ (Eltis, 1975).

The major merit of the Physiocrats is that of having given a fundamental contribution to the rise of that stream of thought which was classical political economy. They precisely defined the concepts of surplus and capital; they introduced the distinction between productive and sterile activities. The Physiocrats clearly distinguished the social classes according to their role in production. Therefore the Physiocrats can properly be regarded as the first inspiration of that economic theory which goes by the name of the surplus approach. In the *Theories of Surplus Value* (Marx, 1864–5, vol. 2, ch. 2), Marx indicated the Physiocrats as the first authors who adopted this approach for the analysis of economic systems. One aspect of Marxian economics which is derived from Physiocracy is the description of the economy by means of reproduction schemes. It must be noticed that the first two sectors in Marx's reproduction schemes coincide with those of Quesnay, that is, agriculture and industry (Marx, 1867–74, vol. 2, part 2).

The Physiocratic distinction between productive and unproductive labour can be found in all major classical economists, from Smith to Malthus and Ricardo, even though they gave different solutions to this problem.

The surplus approach, which was characteristic of classical economists and of Marx, was again brought to the fore in the 1960s thanks to Piero Sraffa's book *Production of Commodities by Means of Commodities*. Sraffa refers to the Physiocrats as one of his sources (see Sraffa, 1960, appendix D).

Physiocratic economics also influenced other aspects of modern economic theory. Leontief's input–output analysis finds an important forerunner in the *Tableau économique*, while the distinction between productive and unproductive labour has been the focus of renewed interest and has been used to investigate the failures of some modern economic systems (Bacon and Eltis, 1976, preface).

The influence of the Physiocrats on Adam Smith deserves special attention. Smith was in France for three years between 1763 and 1766, and was in touch with some Physiocrats and with Turgot. Certainly Smith was well aware of the debates which were taking place during those years about Physiocratic economics, and it is generally admitted that he borrowed some specific concepts from Quesnay. These are the concepts of net product, its difference with the capital advanced, and the distinction between production and unproductive labour. These concepts did not appear in Smith's economic writings before his visit to France, but played an important role in the *Wealth of Nations*. The Physiocrats' influence on Smith is further proof of their important place in the building of classical political economy. In the *Wealth of Nations* Smith dedicates many pages to explain Physiocratic economics (Smith, 1776, book 2, chapter 9). He criticized many aspects of Physiocracy; for instance, while accepting the idea that agriculture was the most important economic sector of the country, he did not agree that industry was sterile. For Smith many features of Physiocratic economics were not appropriate to explain the workings of modern commercial societies like England. Physiocracy was too influenced by the economic conditions of 18th-century France, and was thus particularly useful to study agricultural societies. But for Smith, Physiocracy was greatly superior to mercantilism and it was the necessary basis on which to found the new economic science, or as Smith wrote 'the nearest approximation to truth' (Smith, 1776, vol. 2, p. 199).

See Also

- Du Pont de Nemours, Pierre Samuel
- Ephémérides du citoyen ou chronique de l'esprit national
- Quesnay, François

Bibliography

Bacon, R. and Eltis, W. 1976. *Britain's Economic Problem: Too Few Producers*. London: Macmillan.

Baudeau, N. 1767. *Explication du Tableau économique à Madame de ***, par l'auteur des Ephémérides*. In *Ephémérides du citoyen*, vols 11 and 12.

Beer, M. 1939. *An Inquiry into Physiocracy*. London: Allen & Unwin.

Boisguillebert, P.P.S. 1707. Dissertation sur la nature des richesses, de l'argent et des tributs. In *Economistes et financiers du dix-huitième siècle*, ed. E. Daire. Paris: Guillaumin, 1843.

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Ed. H. Higgs, London: Frank Cass, 1959.

Daire, E., ed. 1846. *Physiocrates*. Paris: Guillaumin.

Du Pont de Nemours, P.S. 1767. *Physiocratie, ou Constitution naturelle du gouvernement le plus avantageux au genre humain*. Leyden and Paris.

Du Pont de Nemours, P.S. 1772. *Abrégé des principes de l'économie politique*. In Daire (1846).

Eltis, W.A. 1975. François Quesnay: a reinterpretation. 1. The Tableau économique. *Oxford Economic Papers* 27(2), 167–200.

Eltis, W.A. 1984. *The Classical Theory of Economic Growth*. London: Macmillan.

Galiani, F. 1770. *Dialogues sur le commerce des bléds*. Milan and Naples: Riccardo Ricciardi, 1959.

Groenewegen, P.D. 1977. *The Economics of A.R.J. Turgot*. The Hague: Nijhoff.

Hecht, J. 1958. La vie de François Quesnay. In INED, vol. 1.

Hoselitz, B.F. 1968. Agrarian capitalism, the natural order of things: François Quesnay. *Kyklos* 21, 637–62.

INED (Institut Nationale d'Etudes Démographiques). 1958. *François Quesnay et la Physiocratie*, 2 vols. Paris: INED.

Kuczynski, M. and Meek, R.L. 1972. *Quesnay's Tableau Economique*. London: Macmillan.

Le Trosne, G.F. 1777. *De l'intérêt social, par rapport à la valeur, à la circulation, à l'industrie et au commerce intérieur et extérieur*. In Daire (1846).

Loménie, L. 1879. *Les Mirabeau – Nouvelles études sur la société française au XVIII siècle*, vol. 2. Paris: Dentu.

Marx, K. 1864–5. *Theories of Surplus Value*. London: Lawrence & Wishart, 1963.

Marx, K. 1867–94. *Capital*. London: Lawrence & Wishart, 1970.

Meek, R.L. 1962. *The Economics of Physiocracy*. London: Allen & Unwin.

Meek, R.L. 1968. Ideas, events and environment – the case of the French Physiocrats. In *Events, Ideology and Economic Theory*, ed. R.V. Eagly. Detroit, MI: Wayne State University Press.

Meek, R.L. 1976. *Social Science and the Ignoble Savage*. Cambridge: Cambridge University Press.

Mercier de la Rivière, P. 1767. L'ordre naturel et essentiel des sociétés politiques. In *Collection des économistes et des réformateurs sociaux de la France*, Paris: Geuthner, 1910.

Mirabeau, V.R. 1758–60. *L'ami des hommes, ou Traité de la population*. Aalen: Scientia Verlag, 1970.

Mirabeau, V.R. 1760. *Theorie de l'impôt*. Aalen: Scientia Verlag, 1972.

Mirabeau, V.R. 1764. *Philosophie rurale, ou Economie générale et politique de l'agriculture*. Aalen: Scientia Verlag, 1972.

Pattullo, H. 1758. *Essai sur l'amélioration des terres*. Paris: Durand.

Petty, W. 1662. A Treatise of Taxes and Contributions. In *The Economic Writings of Sir William Petty*, vol. 1, ed. C.H. Hull. Cambridge: Cambridge University Press, 1899.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. E. Cannan, London: Methuen, 1961.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Voltaire, F.-M. 1767. *L'homme aux quarante écus*. In *Oeuvres complètes de Voltaire*, vol. 51, Paris: Baudouin, 1825.

Weulersse, G. 1910. *Le mouvement Physiocratique en France (de 1756 à 1770)*. Paris: Félix Alcan.

How to cite this article

Vaggi, G. "physiocracy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000085> doi:10.1057/9780230226203.1286

The New Palgrave Dictionary of Economics Online

Pierson, Nichlaas Gerard (1839 – 1909)

Arnold Heertje
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Cohen Stuart, A. J.; Edgeworth, F. Y.; education; Pierson, N. G.; taxation, pure theory of

Article

Born in Amsterdam, 7 February 1839; died in Heemstede, 24 December 1909. A Dutch economist of international reputation, Pierson dominated economics in the Netherlands during the second half of the 19th century. He started his career in the commercial and banking world of Amsterdam. He became President of the Dutch Central Bank, Minister of Finance and Prime Minister. As an economist he was a self-educated man, just like David Ricardo, but he was nevertheless invited to become Professor of Economics at the University of Amsterdam. He taught in the Faculty of Law from 1877 onwards until 1885. Broadly speaking, he advocated the main ideas of the Austrian school of thought in economic theory, although he maintained a material concept of welfare and production. On money, banking and taxation he was a well-known authority, who stimulated Cohen Stuart to write his famous dissertation on the application of utility theory to taxation. His knowledge of the history of ideas was outstanding and he was one of the first to recognize the significance of the Italian authors of the 17th and 18th centuries. As a political economist, Pierson was basically in favour of a market economy. He was a critic of Marxism, but still not against a modest degree of state intervention. He advocated, in particular, the importance of high-level education, organized by the government, in order to improve the condition of the working class.

In 1863 he wrote a booklet on the future of the Dutch Central Bank, in which he strongly defended the monopolistic position of the Bank with regard to the creation of banknotes (Pierson, 1863). An English translation of his very popular textbook also appeared (1902a).

Pierson's analysis of value in a socialist society (1902b) is of lasting significance.

See Also

- Cohen Stuart, Arnold Jacob

Selected works

1863. *De toekomst der Nederlandsche Bank*. Haarlem: Kruseman.

1902a. *Principles of Economics*. London: Macmillan.

1902b. The problem of value in the socialist society. In *Collectivist Economic Planning*, ed. F.A. von Hayek. London: Routledge, 1935.

Bibliography

Edgeworth, F.Y. 1925. *Papers Relating to Political Economy*. London: Macmillan, vol. 1, 351–6; vol. 3, 77–85.

Heertje, A. 1989. Nicolaas Gerard Pierson. In *Perspectives on the History of Economic Thought*, vol. 1, ed. D.A. Walker. Aldershot: Edward Elgar.

Hennipman, P. 1964. *Handwörterbuch der Sozialwissenschaften*. Tübingen: Mohr.

Van Maarseveen, J.G.S.J. 1981. *Nicolaas Gerard Pierson*. Rotterdam: Erasmus University.

How to cite this article

Heertje, Arnold. "Pierson, Nichlaas Gerard (1839–1909)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000086> doi:10.1057/9780230226203.1287

The New Palgrave Dictionary of Economics Online

Pigou, Arthur Cecil (1877 – 1959)

Nahid Aslanbeigui

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Arthur Cecil Pigou founded welfare economics by synthesizing Marshall's theoretical framework and Sidgwick's categories of market failure and imperfections. His view of welfare economics was expansive, including resource allocation, income redistribution, business cycles, and unemployment. Pigou made important contributions to other areas of economics as well: the theory of value, public finance, index numbers, and evaluation of real national income. The most neglected aspect of Pigou's work is his investigation of a remarkable range of labour-market phenomena explored by subsequent economists – implicit contracts, internal labour markets, wage rigidity, labour market segmentation, human capital theory, and collective bargaining.

Keywords

bandwagon effect; business cycles; Cambridge School; Coase, R. H.; collective bargaining; cost–benefit analysis; double taxation; externalities; factor prices; fixed factors; happiness; Harris–Todaro model; ideal output; increasing returns; inheritance tax; interdependent utility; interpersonal utility comparisons; involuntary unemployment; Kahn, R.; Keynesian Revolution; Marshall, A.; mathematical economics; money; money illusion; monopoly; natural monopoly; natural rate of unemployment; Pigou, A. C.; Pigou effect; positive economics; price discrimination; principal and agent; public goods; public works; real balances; real wages; redistribution of income; relative income; Robbins, L. C.; Robertson, D.; Robinson, A.; Robinson. J. V.; Sidgwick, H.; snob effect; social cost; Sraffa, P.; stationary state; sticky wages; technical change; trade unions; utilitarianism; welfare economics; Young, A. A.

Article

Arthur Cecil Pigou was founder of welfare economics, long-time occupant of the Chair of Political Economy at Cambridge University (1908–43), and author of hundreds of articles, pamphlets and books. As Alfred Marshall's successor, he embraced, refined and extended the analytical framework that his

master had painstakingly constructed. He also lived long enough to witness its disintegration at the hands of a generation of economists who had lost their tolerance for its limitations.

Life and career

A.C. Pigou was born on 18 November 1877 at Ryde, Isle of Wight, England, and died in Cambridge on 7 March 1959. He attended Harrow (1891–96), emerging as a brilliant scholar and athlete who harboured a shyness of women that bordered on panic. Contrary to common belief, Pigou was no misogynist. He advocated paid maternity leaves for factory workers, voted for women's degrees at Cambridge University, and played a decisive role in creating a lectureship for the young Joan Robinson. Pigou entered King's College, Cambridge on a Minor Scholarship in History and Modern Languages (1896). However, his interests spanned poetry, moral philosophy, politics and economics. His achievements were stunning: a First in Part I of the History Tripos (1899) and another in Part II of the Moral Sciences Tripos with special distinction in political economy (1899), the Chancellor's Medal for English Verse (1899), the Burney Prize (1901), the Cobden Prize (1901) for an essay that secured him a fellowship at King's (1902), the Adam Smith Prize (1903) for work that formed the basis of his Jevons Memorial Lectures at University College, London (1903–4), and a Girdlers' lectureship (1904) that he held until his election to the Chair of Political Economy (30 May 1908).

Although significantly influenced by Henry Sidgwick, Pigou was the foremost disciple of Alfred Marshall, who was impressed by his protégé on several grounds. Pigou's 'exceptional genius', evident in his masterful thesis, foretold a future as 'one of the leading economists of the world'. He knew the proper role of economic theory: an instrument for social betterment, not intellectual gymnastics. Pigou fought for Marshall's brainchild, the independent Economics Tripos (established in 1903), and personally funded lectureships, prizes and book acquisitions. He shared Marshall's commitment to free trade, using his publications (Pigou, 1904; 1906) and superb oratory skills – honed at the Cambridge Union Society of which he was President (1900) – to promote it. Together with Marshall, he signed the notorious Economists' Manifesto that rejected the Tariff Reform Proposal (1903) of Joseph Chamberlain. It is not surprising that Marshall's face beamed with delight when Pigou was chosen as his successor. He had manipulated the election in favour of the 30 year-old Pigou, embittering his old friend H.S. Foxwell, a serious contender.

Pigou's *Wealth and Welfare* (1912) – a synthesis of Marshall's engine of analysis and Sidgwick's categories of market failure and imperfections – laid the foundation for *Economics of Welfare* (1920), *Industrial Fluctuations* (1927a), and *A Study in Public Finance* (1928b). Taken together, these books covered most of the territory of general economics. *Industrial Fluctuations* was later complemented by *The Theory of Unemployment* (1933), which received a harsh and sophistical critique at the hands of J. M. Keynes in *The General Theory of Employment, Interest, and Money* (1936). Although faithful to the classical doctrine, *Employment and Equilibrium* (1941) – arguably the first textbook in macroeconomics – employed an IS–LM version of *The General Theory* and offered a careful analysis of the differences between Keynesian and classical economics. Pigou's other works included *Unemployment* (1913), *The Political Economy of War* (1921), *The Economics of Stationary States* (1935), many collections of essays, as well as books and pamphlets that he characterized as 'low-brow', among them the highly successful *Socialism versus Capitalism* (1937b), *Lapses from Full Employment* (1945) and *Income: An Introduction to Economics* (1946). The rise of a Cambridge School of economics was in large measure

due to Pigou's articulation of Marshall's organon (see, for example, Pigou's classic exposition of Cambridge monetary theory, 1917). Generations of economists – among them Dennis Robertson, Joan and Austin Robinson, and Richard Kahn – learned Marshall in Pigou's lectures, which were legendary for their clarity and logical rigour.

Pigou was not a public man. His aversion to discussions of economics outside ‘the home’ extended to a distaste for conferences. Acting on his sense of public obligation, he served on several government committees – among them the Chamberlain Committee on the Currency and Bank of England Note Issues (1924–5), which recommended a return of sterling to its pre-war level, imposing immense costs on British labour. Disillusioned by British economic policies in the 1930s, he withdrew from public life, making only occasional ritually obligatory appearances before commissions.

Pigou's personal life also became increasingly hermetic. By the 1940s, the high-spirited, companionable young man of the Edwardian era was regarded as a recluse. As a conscientious objector, he never recovered from the experience of the carnage of the Great War, which he observed first-hand as a driver in the Friends’ Ambulance Unit, commanded by his student and friend Philip Noel-Baker. Beginning in the mid-1920s, severe cardiac fibrillation (irregular heartbeat) curtailed his mountaineering – he was a deft climber introduced to the sport by the economic historian J.H. Clapham. This condition left him permanently anxious over his health. Finally, Pigou watched with dismay as the Keynesian Revolution destroyed the Edwardian intellectual culture of high civility in Cambridge economics. In time, he rose above his own angry response to Keynes's gratuitous depiction of classical economists as ‘a gang of incompetent bunglers’ (Pigou, 1936, p. 115). But as relations between Keynes's disciples and Dennis Robertson became increasingly hostile, he grew more remote and diffident. In his judgement, Joan Robinson's dogmatic instruction of Keynesian economics turned undergraduates into ‘identical sausages’, and under Keynes's stewardship in the 1930s the *Economic Journal* violated its mission of representing different schools of thought ‘with equal impartiality’.

Theoretical contributions

Pigouvian economics is grounded in utilitarian moral philosophy: creating the greatest good – Pigou's cognate of welfare – for the greatest number of people. Its analysis is limited to economic welfare: satisfactions that, directly or indirectly, can be related to the measuring rod of money. Up to a point, money, which measures the intensity of desires, performs well as a proxy for satisfaction. However, the human ‘telescopic faculty’ irrationally discounts future satisfactions, resulting in inadequate savings, insufficient investment in tunnels or forests, depletion of natural resources, and extinction of animal species. Pigou assumes that, as a rule, economic and total welfare are positively related. Anticipating contemporary research on happiness, he also recognizes the importance of factors that contribute to non-economic welfare such as relative status, social capital, political freedom, and moral quality of life. Economic welfare may improve if its objective counterpart, the national product, is increased in size, distributed more evenly, and made more stable.

Optimal resource allocation

Integrating Marshall's marginal analysis and Sidgwick's distinction between private and public interests, Pigou produces some of the most important concepts (1912) and diagrams (1910) of welfare economics:

marginal private and social net products (benefits and costs in contemporary parlance). In the absence of ‘costs of movement’ – associated with geographic and occupational reallocation of resources – the allocation of resources by competitive markets achieves universally equal marginal private net products. However, the production of ideal output requires equality of marginal social net products. Where private and social net products diverge, there is a prima facie case for reallocation of resources (1932, p. 136).

In Pigou's competitive economy, social and private benefits diverge in three different respects. First, a principal–agent problem arises when owners of land contract out its use to tenants. Since some benefits of the agent's investment accrue to the principal on termination of the contract, investment levels are not socially optimal. Pigou's remedies are limited to modifying contractual specifications between the two parties, presumably because low transactions costs render government action unnecessary.

Second, economic transactions between two agents may render incidental services or disservices to third parties, who cannot be forced to pay for the benefits or compensated for the costs. Unlike contemporary economists, Pigou does not distinguish public goods and externalities. Positive spillovers are a combination of public goods and beneficial externalities: lighthouses that benefit free-riding ships; private parks and forests that improve air quality; roads and tramways that improve the value of neighbouring land; privately owned lamps that shed light on streets; items of smoke-prevention equipment that benefit buildings, vegetables, clothes, and air quality; and ‘most important of all’ scientific research that leads to inventions, innovations and ‘discoveries of high practical utility’ (1932, p. 185). Negative spillovers are harmful externalities: a landlord raises rabbits that overrun a neighbour's property; a firm builds a factory in a densely populated area, destroying its amenities and injuring family health and productivity; automobile operators drive cars that wear out the surface of roads; and producers sell alcoholic beverages that increase crime. The ‘crowning illustration’ of negative externalities is women's factory work, especially immediately before and after childbirth, which damages the health of the fetus and increases infant mortality (1932, pp. 185–7).

Since it is difficult to internalize positive or negative externalities through contractual modifications, the state may offer ‘extraordinary encouragements’ or ‘extraordinary restraints’ as remedies, most obviously taxes and ‘bounties’. In Pigou's era, a variety of taxes had already been imposed on alcoholic beverages, roads, gasoline and car licences. Bounties ranged from complete government provision (police protection and cleaning slums) to grants for scientific research. Pigouvian solutions went beyond taxes and subsidies to include patent enforcement, provision of information and training, and paid maternity leaves. In cases such as urban planning, where ‘the inter-relations of the various private persons affected [are] highly complex’, the state may have to exercise ‘authoritative control’ because the invisible hand fails to ‘tackle the collective problems of beauty, of air and of light’ (1932, pp. 193–6; also see 1947, pp. 94–100).

Careful readers of Pigou will note that much of Ronald Coase's critique of his analysis (Coase, 1960) is misplaced. Pigou stressed that on issues of policy he always spoke with an ‘uncertain voice’ (Pigou, 1932, p. 10), carefully considering the costs and benefits of proposed solutions. Government action entails allocative, administrative and political costs. Redeployment of labour, land and capital is also costly. It follows that the goal of achieving ideal output should be subjected to a cost–benefit analysis that shows ‘at which point the advantage of getting closer is outweighed by the complications, inconvenience and expense involved in doing so’ (Pigou, 1932, p. 315).

Third, in his early work (1912), Pigou argued that private and social benefits diverge if industries exhibit increasing or decreasing costs. Under decreasing returns, a small increase in the output of one firm

creates external diseconomies for the industry by increasing the price of fixed factors. Under increasing returns, a small rise in the output of one firm creates external economies for the industry. A prima facie case could therefore be made for taxing increasing-cost and subsidizing decreasing-cost industries. Pigou's critics – Allyn Young and Dennis Robertson – pointed out that the two types of returns are essentially different phenomena: external economies – technological change and managerial breakthroughs – are irreversible social gains. External diseconomies – increased factor prices – are not social costs since they merely transfer purchasing power from producers to factor owners. The second edition of *The Economics of Welfare* (1924) conceded this point, with the proviso that foreign owners do not capture the increased rents.

In 1926, Piero Sraffa argued that increasing and decreasing returns are incompatible with Marshall's competitive, partial-equilibrium assumptions. Under increasing costs, for instance, a marginal increase in the output of a firm in a given industry increases the price of fixed factors for all industries that use them. Relative prices may change as a result, rendering Marshallian assumptions logically incoherent since industry supply and demand become interdependent. Although economies and diseconomies that are external to the firm but internal to the industry do not generate the same logical problem, they are rare empirically. Pigou (1927b) concluded that, although increasing costs were incompatible with his framework, he could not logically rule out external economies. In 1928, he published the standard textbook analysis of stable equilibrium in a competitive firm (1928b). The costs of the equilibrium firm (a theoretical entity based on Marshall's representative firm) are a function of its own output and that of the industry. Although the industry may experience increasing or constant returns, the equilibrium firm is always at equilibrium when industry price is equal to its marginal and (the minimum of) average costs. U-shaped average and marginal cost curves for the equilibrium firm complemented the mathematical treatment, perhaps the first time that such diagrams were published in English. External economies shift the equilibrium firm's cost curves.

As a rule, monopolistic conditions create discrepancies between private and social benefits. Pigou argues that their implications for welfare must be evaluated on a case-by-case basis. The incidence of discrepancies depends on whether a monopoly practices price discrimination of the first, second or third degree. State control and state operation of natural monopolies have different ramifications for welfare. Oligopolistic market structures, however, create unequivocal social costs irrespective of output: wasteful advertising, exploitation of workers – defined as payment below the value of marginal product – customer deception, reduction of upward mobility by forcing small entrepreneurs out of the market, constraints on inventions and innovations, and Tayloristic practices that dull worker initiative. Pigouvian remedies range from taxes and prohibitions to encouragement of small business.

Income redistribution

Redistribution schemes that favour the poor but leave the national product intact are likely to improve economic welfare. However, both the expectation and the fact of such transfers may produce disincentives that reduce the national product. The implication is not inaction. Rather, the state should design redistributive measures based on a comprehensive knowledge of legal, psychological and institutional factors. If capital is subject to double taxation, its flight is less probable. If economic actors target a specific level of savings, inheritance taxes may not affect investment activity. If redistributed income is used to train workers with uncommon abilities, its rate of return may surpass the return on

investment in physical capital. Finally, taxation may not discourage the rich if it leaves their relative income intact. Pigou's theoretical analysis of interdependent utility (welfare) – based on reference groups, relative income, snob and bandwagon effects – anticipates Duesenberry's and Leibenstein's by some 45 years (Pigou, 1903).

Transferring one dollar from the rich to the poor increases economic welfare because ‘it enables more intense wants to be satisfied at the expense of less intense wants’ (Pigou, 1932, p. 89). This proposition assumes that representative members of different income groups have equal capacities for satisfaction. In 1932, Lionel Robbins claimed that such interpersonal comparisons are normative judgments and have no place in science. The ensuing attempts to establish a positivist welfare economics engaged such luminaries as Hicks, Kaldor, Scitovsky, Little, Bergson and Arrow. The results produced a sophisticated theoretical apparatus but confirmed Pigou's belated response to Robbins that without such comparisons every ‘apparatus of practical thought’ will collapse (Pigou, 1951, p. 292). In recent decades, the recognition that all sciences make normative claims has become received wisdom in the philosophy of science. With the demise of doctrinaire positivism, economists seem more willing to venture into the territory of interpersonal comparisons, as contemporary happiness research suggests. This research provides new grounds for reconsidering the unexploited resources of Pigouvian welfare economics.

Industrial fluctuations and unemployment

Long spells of unemployment have serious deleterious effects – malnutrition, permanent damage to the capabilities of youth, loss of skills and work ethic, alcoholism, a ‘haunting’ sense of insecurity and uncertainty, and the destruction of self-respect and self-confidence – that cannot be reversed in good times. Thus a prima facie case for macroeconomic stability is evident.

Pigou's theory of unemployment can be elucidated by using the language of supply and demand. Aggregate labour supply is vertical, even though individual labour supply curves may be upward sloping or backward bending. Aggregate labour demand – difficult to construct due to sectoral interdependence – is downward sloping and dependent on marginal product. Since unemployment is always positive, it can be explained only by movements in wages and the demand for labour.

Pigou distinguishes two types of unemployment. Short-run involuntary unemployment – a term he may have coined in 1913 – occurs because of frequent changes in labour demand and real wages. Although prices vary, real wages fluctuate because nominal wages remain sticky. (*a*) A perpetually flexible nominal wage is impracticable due to high administrative costs, which become more significant if ‘elaborate and formal arbitration proceedings’ are instituted to resolve capital–labour conflicts (Pigou, 1913, pp. 92–3). (*b*) Some wage rigidity is preferred: while workers want stable living standards, employers are obliged to deliver products at prices previously negotiated. (*c*) The duration of recessions and recoveries is unpredictable; it is not worthwhile to alter wages if the state of the economy is ephemeral. (*d*) Due to mutual mistrust, workers and firms alike resist wage changes, fearing that they may be irreversible. (*e*) Employees and employers suffer from money illusion, the latter resisting wage increases and the former refusing wage cuts.

Contrary to Keynes's straw-man depiction, Pigouvian labour demand fluctuates due to general and wave-like swings in expectations of profits. Three sets of factors affect expectations: real causes such as crop size or technological breakthroughs; monetary variables, which are restricted to exogenous shifts in credit under the gold standard; and psychological factors, which occur spontaneously or as a

consequence of the other two variables. Undue pessimism or optimism may be magnified because psychology, output, and debt–credit linkages create sectoral interdependence.

The amplitude of business cycles depends on the institutional structure of the economy: monetary policy, the pricing strategy of firms, income maintenance programmes, wage policy and unions. Although limited by the quality and quantity of data at his disposal, Pigou tries to quantify factors that cause business cycles or affect their amplitude. Removing monetary or psychological factors would each reduce the amplitude by one-half, crop variation by one-quarter, wage rigidity by one-eighth and price rigidity by one-sixteenth. It is clear that Pigou does not regard high real wages as the single or even the most important cause of short-run unemployment. In many cases, high wages and unemployment are both effects of factors such as ‘bursting of a gigantic bubble of unwarranted optimism, with a heavy fall in price’ (Pigou, 1929, pp. 200–1). Short-run unemployment can be reduced proactively through distribution of information, price stability, or interest-rate manipulation. Reactive policies that dampen the impact of unemployment range from (public work projects to guarantees of interest or subsidies for employers. Although Pigou favours wage flexibility at the theoretical level, he does not consider it a viable political option.

Pigou analyses long-run unemployment on stationary-state assumptions, ruling out changes in expectations, tastes, net investment, productivity, and technology. The only conceivable unemployment under these conditions is an ‘intractable minimum’ that resembles the natural rate of unemployment. It is caused by frictions, immobility, public opinion, the practical impossibility of setting wages according to marginal productivity, and unions. Collective bargaining introduces indeterminacy, which he analyses in a quasi-game-theoretic framework (Pigou, 1905). Employers and employees negotiate money wages within a ‘range of indeterminateness’. The upper limit depends on unions’ reluctance to demand a wage so high that it would result in layoffs. The lower limit is determined by employers’ recognition that a wage that reduces the available supply of labour is too low. Peaceful wage bargains are conducted within a narrower range determined by the ‘sticking point’ of each party: a certain minimum below which workers would rather strike and a maximum above which employers would prefer shutdowns. Firms often have bargaining power to exploit workers but may choose not to, recognizing that low wages affect the productivity of workers they want to retain for the long period. This results in unemployment in a casual labour market. The magnitude of joblessness is determined by a Harris–Todaro comparison of the expected wage – ‘the wage-rate multiplied by the chance of employment’ (Pigou, 1913, p. 55) – with wages elsewhere. Unemployment is not an inevitable outcome if outsiders (low-wage workers) know that insiders (high-wage workers) are irreplaceable.

To reduce long-run unemployment, the state may attempt to educate the unskilled and try to improve wage flexibility. The effective demand ramification of wage flexibility was a major point of contention between Keynes and Pigou (see Pigou, 1937a; Keynes, 1937). Although Pigou was finally persuaded by Kaldor (1937) to take such effects into account (Pigou, 1938; 1941), he discounted them based on the well-known Pigou effect: lower money incomes and prices would increase the value of real balances, reducing and ultimately eliminating the individual's desires to save out of any assigned real income (Pigou, 1943, p. 349). In Pigou's opinion, Keynes's true contribution was not substantive but analytical: no one before him had constructed a model of the aggregate economy that incorporated both real and monetary factors. But Pigou (1950) also maintained that Keynes's analytical framework was too limited to be suitable for direct practical application.

Legacy

Economists have generally judged Pigou's work on Robbinsian, Keynesian or Coasean premises, ignoring his important contributions to the theories of value, distribution, business cycles, public finance, index numbers, and evaluation of real national income. Pigou's neglected contributions to labour economics, which anticipate Hicks's work by a quarter of a century, are especially noteworthy. *Wealth and Welfare*, hailed by Schumpeter as ‘the greatest venture in labor economics ever undertaken by a man who was primarily a theorist’ (1954, p. 948), and his numerous other works on labour and unemployment demonstrate an acute understanding of the importance of a remarkable range of phenomena explored by subsequent economists – implicit contracts, internal labour markets, labour market segmentation, wage rigidity, human capital theory, and collective bargaining. Alfred North Whitehead famously held that ‘a science which hesitates to forget its founder is lost’. Economists have not found it difficult to forget the founder of welfare economics, with regrettable consequences that Whitehead did not envision.

See Also

- financial intermediation
- interpersonal utility comparisons
- Keynesian revolution
- labour economics
- unemployment
- welfare economics

Selected works

1903. Some remarks on utility. *Economic Journal* 13, 58–68.

1904. *The Riddle of the Tariff*. London: R. Brimley Johnson.

1905. *Principles and Methods of Industrial Peace*. London: Macmillan.

1906. *Protective and Preferential Import Duties*. London: Macmillan.

1910. Producers’ and consumers’ surplus. *Economic Journal* 20, 358–70.

1912. *Wealth and Welfare*. London: Macmillan.

1913. *Unemployment*. London: William and Norgate.

1917. The value of money. *Quarterly Journal of Economics* 32, 38–65.

1920. *The Economics of Welfare*. London: Macmillan.

1921. *The Political Economy of War*. London: Macmillan.

1927a. *Industrial Fluctuations*. London: Macmillan.

1927b. The laws of diminishing and increasing costs. *Economic Journal* 37, 188–97.

1928a. An analysis of supply. *Economic Journal* 38, 238–57.

1928b. *A Study in Public Finance*. London: Macmillan.

1929. *Industrial Fluctuations*, 2nd edn. London: Macmillan.

1932. *The Economics of Welfare*, 4th edn. London: Macmillan.

1933. *The Theory of Unemployment*. London: Macmillan.

1935. *The Economics of Stationary States*. London: Macmillan.

1936. Mr. J.M. Keynes' *General Theory of Employment, Interest and Money Economica* 3, 115–32.

1937a. Real wages and money wage rates in relation to unemployment. *Economic Journal* 47, 405–22.

1937b. *Socialism versus Capitalism*. London: Macmillan.

1938. Money wages in relation to unemployment. *Economic Journal* 48, 134–8.

1941. *Employment and Equilibrium*. London: Macmillan.

1943. The classical stationary state. *Economic Journal* 53, 343–51.

1945. *Lapses from Full Employment*. London: Macmillan.

1946. *Income: An Introduction to Economics*. London: Macmillan.

1947. *A Study in Public Finance*, 3rd edn. London: Macmillan.

1950. *Keynes's General Theory: A Retrospective View*. London: Macmillan.

1951. Some aspects of welfare economics. *American Economic Review* 41, 287–302.

Bibliography

Aslanbeigui, N. 1990. On the demise of Pigouvian economics. *Southern Economic Journal* 56, 616–27.

Aslanbeigui, N. 1992. Pigou's inconsistencies or Keynes's misconceptions? *History of Political Economy* 24, 413–33.

Aslanbeigui, N. 1996. The cost controversy: Pigouvian economics in disequilibrium. *European Journal of the History of Economic Thought* 3, 275–95.

Aslanbeigui, N. 1998. Unemployment through the eyes of a classic. In *Keynes and the Classics Reconsidered*, ed. J.C.W. Ahiakpor. Boston: Kluwer.

Aslanbeigui, N. and Medema, S.G. 1998. Beyond the dark clouds: Pigou and Coase on social cost. *History of Political Economy* 30, 601–25.

Aslanbeigui, N. and Oakes, G. 2007. The editor as scientific revolutionary: Keynes, *The Economic Journal*, and the Pigou affair, 1936–1938. *Journal of the History of Economic Thought* 29, 1–34.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Collard, D. 1981. A.C. Pigou, 1877–1959. In *Pioneers of Modern Economics in Britain*, ed. D.P. O’Brien and J.R. Presely. Totowa, NJ: Barnes and Noble Books.

Collard, D. 1996. Pigou and modern business cycle theory. *Economic Journal* 106, 912–24.

Cooter, R. and Rappoport, P. 1984. Were the ordinalists wrong about welfare economics? *Journal of Economic Literature* 22, 507–30.

Kaldor, N. 1937. Prof. Pigou on money wages in relation to unemployment. *Economic Journal* 47, 745–53.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1937. Prof. Pigou on money wages in relation to unemployment. *Economic Journal* 47, 743–5.

O'Donnell, M.G. 1979. Pigou: an extension of Sidgwickian thought. *History of Political Economy* 11, 588–605.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Robertson, D.H. 1924. Those empty economic boxes: a rejoinder. *Economic Journal* 34, 16–30.

Saltmarsh, J. and Wilkinson, P. 1960. *Arthur Cecil Pigou, 1877–1959*. Cambridge: Cambridge University Press.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Shiller, R.J. Ultimate sources of aggregate variability. *American Economic Review* 77, 87–92.

Solow, R.M. 1980. On theories of unemployment. *American Economic Review* 70, 1–11.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36, 535–50.

Young, A. 1913. Pigou's Wealth and Welfare. *Quarterly Journal of Economics* 27, 672–86.

How to cite this article

Aslanbeigui, Nahid. "Pigou, Arthur Cecil (1877–1959)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000088> doi:10.1057/9780230226203.1288

The New Palgrave Dictionary of Economics Online

Pigouvian taxes

Agnar Sandmo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Pigouvian taxes are taxes designed to correct for negative external effects. The idea is originally due to Pigou (1920), and has received increased attention in recent years because of the concern with environmental issues. This article sets out the basic theoretical argument and considers the modifications of the theory that have to be made when these taxes are seen in the context of an otherwise distortionary tax system. It also briefly considers the issue of the ‘double dividend’ from a green tax reform.

Keywords

distortionary tax; double dividend; externalities; lump sum taxes; marginal cost of public funds; optimal taxation; partial equilibrium; payroll tax; Pigou, A. C.; Pigouvian taxes; Ramsey tax; substitutes and complements; tax base; tax wedge

Article

‘Pigouvian taxes’ is the generic term for taxes designed to correct inefficiencies of the price system that are due to negative external effects. In partial equilibrium terms, the basic idea can be presented as follows: under competitive conditions, utility-maximizing consumers will equate their marginal benefit to the market price Q; we may write this as $MB=Q$. Similarly, profit-maximizing producers will set their marginal private cost equal to the price, so that $MPC=Q$. In the absence of externalities, marginal private and social costs coincide: $MPC=MSC$. Consequently, market equilibrium implies that $MB=MSC$, which is the condition for efficient resource allocation. If there are negative external effects related to the production or consumption of the good in question, the marginal social cost is higher than the marginal private cost: $MSC>MPC$. If the market prices facing producers and consumers are identical, this implies that $MB<MSC$. To restore efficiency, we may levy a tax on the commodity, so that the consumer price is Q while the producer price is $Q–t$. In the new equilibrium we have that $MB=Q$ and $MPC=Q–t$; it follows that $MB=MPC+t$. Since we wish the equilibrium to satisfy the condition that $MB=MSC$, we

must have $t=MSC–MPC$, which we may define as the marginal social damage. Accordingly, the optimal Pigouvian tax internalizes the externality; producers act as if they took account of the marginal social damage associated with the production of the commodity.

This idea was first expressed by Pigou, especially in his *Economics of Welfare* (1920). He mentions a number of examples of what he calls divergence between ‘social and private net product’, for example, production activities generating smoke from factory chimneys that create adverse consequences for consumers in the form of damage to buildings, increased expenses for washing clothes, house-cleaning and indoor lighting. These inefficiencies can be corrected, he says, by ‘imposing appropriate rates of tax on resources that tend to be pushed too far’; he also points out that cases of positive externalities where $MSC<MPC$ can be corrected by means of subsidies or ‘bounties’ (Pigou, 1920; 1932, p. 184). In his later book, *A Study in Public Finance*, he claims that

> [there] will necessarily exist a certain determinate scheme of taxes and bounties, which, in given conditions, distributional considerations being ignored, would lead to the *optimum* result. (Pigou, 1928; 1947, p. 99)

An interesting and important question concerns the choice of the tax base. On what should the Pigouvian tax be levied? From a theoretical point of view, the correct tax base is the one that affects the crucial margin of decision. In the factory smoke example, the best tax base is actually the amount of smoke emission. A tax on coal is an imperfect instrument to the extent that it also affects margins that are irrelevant for smoke emission, and this is even more true for a tax on the output produced by the factory. Some would therefore reserve the term ‘Pigouvian tax’ for the tax on smoke emission, but in the literature it has become common to use the concept to refer to all cases where the policy motivation is to correct for negative externalities.

For a long time, Pigouvian taxes led an obscure life in the public economics literature; thus, in the famous treatise by Musgrave (1959), the subject is barely mentioned. However, with the increased concern for the environment that rapidly gained ground from the late 1960s, economists became much more interested in this form of tax policy both as a tool for environmental policy and as an efficient source of revenue for the public sector.

Distortionary taxes

The partial equilibrium approach is based on some simplifying assumptions. First, it focuses solely on the market for the ‘commodity’ (final good, factor of production or emission) that gives rise to the externality, while neglecting the interconnections with other markets. Second, it assumes, rather implicitly, that there are no other violations of the efficiency conditions in the economy, so that the design of Pigouvian taxes does not need to take into account the presence of other distortions. Third, as also emphasized by Pigou, it ignores distributional concerns.

All these simplifications must be overcome if one wishes to analyse Pigouvian tax policy within the context of the overall tax system. There is actually one tax system in which the partial equilibrium analysis is valid, and that is the assumption that the rest of the requirement for public sector revenue can be satisfied by means of individualized lump sum taxes. This leads to a ‘first-best’ allocation: tax

revenue is raised without distortions of the price mechanism, and the desired income distribution can be achieved without loss of efficiency. The only commodity taxes that are used are the Pigouvian taxes on commodities that generate negative external effects. But lump sum taxes are not policy instruments that can be used realistically. Instead, governments have to rely on direct and indirect taxes, and these will create tax wedges and distortions of private incentives. What is the role of Pigouvian taxes within the context of an otherwise distortionary tax system?

One might perhaps come to think that in such a setting Pigouvian considerations should affect the taxes on all goods: for example, there might be a case for subsidizing substitutes and taxing complements to the harmful commodities. However, it was shown in Sandmo (1975) that in an optimal system of commodity taxes the integration of Pigouvian taxes with the Ramsey (1927) objective – minimizing efficiency loss for a given tax revenue – takes a strikingly simple form. If there is one commodity that creates a negative externality, the tax on this commodity can be expressed as a weighted average of Ramsey and Pigou terms, while other taxes contain only a Ramsey term. Formally, suppose that there are a number of taxed goods (i=1,…,n) and that the externality is generated by the nth good. Suppose for simplicity that all cross-elasticities between the taxed goods are zero, so that Ramsey taxes can be characterized by the inverse elasticity formula. Then the optimal tax system can be written as follows:

$$t_i = \alpha(-1/\varepsilon_i)\ (i = 1, \ldots, n-1).$$

$$t_n = \alpha(-1/\varepsilon_n) + (1-\alpha)\delta_n.$$

Here ε_i is the own price elasticity of commodity i, and δ_n is the marginal social damage of commodity n. α is a parameter that characterizes the tightness of the government budget constraint. If the budget is extremely tight, all weight is on the need for revenue. Then α =1, the tax rates are chosen so as to maximize revenue, and Pigouvian taxes play no role in the tax structure. However, in the happy situation where the revenue from Pigouvian taxes is exactly sufficient to meet the government's revenue requirement, α =0, and no other taxes are desirable. It can be shown that the ‘additivity’ property of the optimal tax system continues to hold when distributional considerations are incorporated into the model, but in that case the weights on the inverse elasticity and the marginal social damage will have to reflect distributional concerns in addition to those of efficiency.

The double dividend and the marginal cost of funds

In recent years there has grown up a strong interest in ‘green tax reforms’. Such reforms would reduce conventional distortionary tax rates and compensate for the loss of tax revenue by introducing more

Pigouvian taxes. A popular view of the gain from this kind of reform is that society would reap a ‘double dividend’. First, higher Pigouvian taxes would create an improved environment; second, lower distortionary taxes would imply a more efficient tax system. This argument has a strong appeal to economic intuition; however, as often happens, when one comes to study it more closely, it turns out to contain some complicating elements. The crucial point to note is that the effects of Pigouvian taxes interact with those of the distortionary taxes. If for example the existing tax system has a high marginal tax rate on labour income, an increase of Pigouvian taxes together with a lowering of other indirect tax rates might exacerbate the labour market distortion if the externality-creating goods are complementary with labour supply. This argument does not imply that the argument in favour of the double dividend is groundless. It simply means that one has to be careful in taking account of the interaction between markets for taxed goods before predicting a double dividend.

Another version of the double dividend argument focuses on unemployment. If the basic cause of unemployment is that employers’ labour cost is above the market-clearing wage, a promising tax reform might be to reduce the payroll tax while increasing Pigouvian taxes. The double dividend in this case would be a better environment and lower unemployment. Again, the consensus of professional opinion seems to be that this is indeed a possible outcome, but it is by no means assured. For example, in a unionized labour market much will depend on the combined incidence of a reduced payroll tax and higher indirect taxes on union wage demands. For further discussion of both versions of the double dividend, see Bovenberg (1999) and Sandmo (2000).

Related to the question of the double dividend is the relationship between Pigouvian taxes and the marginal cost of public funds (MCF), a concept whose origin can also be traced to Pigou: see Atkinson and Stern (1974). With distortionary tax finance, the direct resource cost of public goods should be multiplied with an MCF adjustment factor which exceeds one. Since Pigouvian taxes actually increase the efficiency of the market mechanism, one might expect that for this type of tax finance one would have MCF<1. Theoretical analysis has shown that this is indeed likely to be true in a number of cases, but that here too one needs to pay attention to the interaction of distortionary and Pigouvian taxes.

See Also

- environmental economics
- optimal taxation

Bibliography

Atkinson, A.B. and Stern, N. 1974. Pigou, taxation and public goods. *Review of Economic Studies* 41, 119–28.

Bovenberg, A.L. 1999. Green tax reforms and the double dividend: An updated reader's guide. *International Tax and Public Finance* 6, 421–43.

Musgrave, R.A. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

Pigou, A.C. 1920. *The Economics of Welfare*, 4th edn. London: Macmillan, 1932.

Pigou, A.C. 1928. *A Study in Public Finance*, 3rd edn. London: Macmillan, 1947.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Sandmo, A. 1975. Optimal taxation in the presence of externalities. *Swedish Journal of Economics* 77, 86–98.

Sandmo, A. 2000. *The Public Economics of the Environment*. Oxford: Oxford University Press.

How to cite this article

Sandmo, Agnar. "Pigouvian taxes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000351> doi:10.1057/9780230226203.1289

The New Palgrave Dictionary of Economics Online

planning

Rajiv Vohra
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

convexity; decentralization; excess demand; increasing returns; informational efficiency; Lange, O. R.; marginal rate of substitution; monotonicity; planning; shadow pricing; socialist calculation debate; tâtonnement.

Article

Formally, planning in an economic context can be identified with a constrained maximization problem. The objective, whether it is simply social welfare or multiple individual utilities, is maximized subject to the resource and technological constraints. It needs to be emphasized that the planning problem is not simply one of characterizing the solution to the maximization problem but also of defining a computational procedure to obtain the solution. A planning process can be defined as an iterative procedure which, through successive approximations, finds a solution to the maximization problem.

The literature on planning processes goes back at least to the debate of the 1920s and 1930s on the possibility of economic calculation in a socialist state. While the formal versions of the welfare theorems, as presented by Arrow (1951) and Debreu (1954), were not available then, it was fairly well recognized that the competitive mechanism would, in equilibrium, satisfy the marginal conditions in terms of the equality of prices and the relevant rates of substitution and that this would constitute an efficient method of allocating resources. In what seems, at least in retrospect, to be an argument one may well be tempted to make if one was aware of the second welfare theorem, Mises (1922) argued that since the markets for capital goods, and hence their prices, would not exist in a socialist economy, it would be impossible for such an economy to allocate its resources rationally. However, Pareto (1897), in comparing the market to a computing machine, had already pointed out that a procedure similar to the competitive process of the market could be used to determine a plan. His argument had been further elaborated by Barone (1908). The focus of Mises's criticism was somewhat changed by Hayek (1935) who did not rule out the theoretical possibility of a planned economy being able to allocate resources rationally. The scepticism was centred around the ability of the planning authority, say the Central planning Board (CPB), to solve the ‘hundreds of thousands’ of equations necessary to achieve the objective. Partly in response to this criticism, iterative processes were presented, in what are now famous papers by Taylor (1929) and Lange (1936–7), to show that a planned economy could allocate resources in much the same way as the competitive system. They formalized a planning process which would follow the competitive rules to allocate resources; the trial and error method for finding the optimal allocation was similar to Walrasian tâtonnement. The arguments presented by the sceptics were turned on their head; the planned economy could play the competitive game just as well as the market, perhaps better.

While, in the classical environment, a process which imitates the competitive market has the clear advantage of leading, in equilibrium, to a Pareto optimal allocation, the dynamic properties of such a process were analysed much later. Samuelson (1949) showed that in a linear economy, such a mechanism led to indefinite oscillations. Arrow and Hurwicz (1960) rigorously formalized Lange's process, for an economy with a single utility function, and showed that strict concavity of the utility function and the technological constraints were crucial in establishing the convergence of the dynamic process.

In the subsequent development of this literature considerable attention has been paid to developing processes which converge to an optimal plan. Other criteria for comparing different processes have also been formalized (see for example Hurwicz, 1960, and Malinvaud, 1967) and we shall discuss these in more detail in Section 1. At this stage it is, however, worthwhile to point out that a planning process which mimics the competitive process has considerable appeal. In the classical environment it leads to an allocation which is Pareto optimal. It also retains the attractive informational processing properties of the competitive mechanism; the CPB is not required to collect all the information on the economic environment nor does it need to solve the entire programming problem by itself since various stages of the optimization process are conducted at the individual level. Subsequent

literature on planning processes has, quite justifiably, concentrated on processes which are in some sense decentralized. Processes applicable to non-classical environments have also been formulated.

There is also a considerable literature on general allocation processes in which the CPB is not assigned a distinguished role (see for example Arrow and Hurwicz, 1977). In this article we shall concentrate only on planning processes. In particular, we consider an economy with many firms and a CPB. Except for the section on public goods where we consider many consumers, the CPB is assumed to have the objective of maximizing a single utility function. Section 1 will set out the model and the criteria which may be used to compare different processes. Processes designed for the classical environment are considered in Section 2. Sections 3 and 4 deal respectively with economies with increasing returns and with public goods. Due to limits on space, we shall not deal with other non-classical environments that have also been studied in the literature on allocation process (see for example Section III of Hurwicz, 1973). Another important aspect of planning which is not covered here is that of incentive compatibility. Moreover, the discussion is not intended to cover all the details of the processes under consideration and the reader may find it useful to consult the cited papers. Notable among the surveys in this area are Heal (1973), Hurwicz (1973) and Tulkens (1978).

1 The formal model and definitions

We shall consider an economy with k commodities, indexed by l, and n+1 agents, n firms and the CPB. Agents will be indexed by

$$j, j = 1, \ldots, n, n+1.$$

We shall also find it convenient to index the firms by j, j=1, …, n. Firm j's technology is represented by production set $Y^j \subseteq R^k$. The environment of firm j is simply $e^j = Y^j$. The CPB has a continuous utility function $U: X \to R$ where X is the consumption set. The aggregate endowment of the economy is denoted $\omega \in R^k$. The economic environment of the CPB is $e^{n+1} = (X, U, \omega)$. The economy can be described in terms of its environment $e = (X, (Y^j), U, \omega)$.

D.1. A *program* (x, y) consists of a consumption plan $x \in X$ and a collection of production plans $y = (y^j) \in Y = \Pi_j Y^j$.

D.2. A program (x, y) is said to be *feasible* if $x = \Sigma_j y^j + \omega$.

D.3. A program (x, y) is said to be *Pareto optimal* if it is feasible and there does not exist another feasible program $(\bar{x}, \bar{y})$ such that $U(\bar{x}) > U(x)$.

A planning process is an iterative process in which messages are exchanged between the firms and the CPB. Agent i chooses a message m^i from a set M, taking into account the environment and the messages received in the previous period. Let m_t^i refer to agent i's message in time period t and

$$m_t = \left(m_t^1, \ldots, m_t^j, \ldots, m_t^{n+1}\right).$$

The response of agent i may then be defined in terms of a response function $f^i: M^{n+1} \to M$ where M^{n+1} refers to the n+1 fold Cartesian product of M and

$$m_{t+1}^i = f^i(m_t; e).$$

An equilibrium message is simply defined as a stationary message. The equilibrium of the process is determined by an outcome function h which translates the equilibrium message

into the equilibrium program or plan. We can now formally define these concepts:

D.4. Given an environment *e*, a *planning process* is defined as $\pi = (M, f, h)$ where $f: M^{n+1} \to M$ and $h: M^{n+1} \to X \times Y$.

D.5. An *equilibrium message* for a process π is an $m \in M^{n+1}$ such that $m=f(m; e)$.

D.6. An *equilibrium program* (or an *equilibrium plan*) for a process π is a program (x, y) such that $(x, y)=h(m; e)$ and m is an equilibrium message.

We shall now discuss some of the desirable properties that a planning process may have. These properties may be broadly classified in terms of the performance of the process and its informational efficiency. We begin by presenting the performance criteria introduced in Malinvaud (1967).

Clearly convergence to a Pareto optimal allocation is a requirement that any planning process ought to satisfy.

D.7. A process π is said to be *convergent* if an equilibrium program exists and is Pareto optimal, that is, as $t \to \infty$ $\mathrm{Lim}_t h(m_t; e)$ is Pareto optimal.

Malinvaud also stresses the importance of the following properties which may, in practice, be even more important if the process needs to be terminated before equilibrium is reached.

D.8. A planning process $\pi = (M, f, h)$ is said to be *feasible* if $f(m, e)$ and $h(m, e)$ are non-empty and $h(m, e)$ is feasible for all $m \in M^{n+1}$.

D.9. A planning process π is said to be *monotonic* if $U(x_{t+1}) \geq U(x_t)$ for all t, where x_t is the consumption plan corresponding to $h(m_t; e)$. It is *strictly monotonic* if it is monotonic and $U(x_{t+1})=U(x_t)$ implies that $h(m_t; e)$ is Pareto optimal.

Hurwicz (1960) and (1969) formalized the notion of informational efficiency associated with a process. His definitions are applicable to general allocation processes in which a CPB is not assigned a distinguished role and we shall suitably modify his concepts to apply specifically to planning processes. The definitions which follow are aimed at formally defining a decentralized process, a definition which is intended to include but not be synonymous with the competitive process. An important characteristic of the competitive system is that initial information is dispersed among the agents: firm *j* knows only its own environment Y^j while a consumer knows only his or her utility function and endowment. A process in which the *i*th agent's response functions depends only on e^i is said to be *external*. A process is *anonymous* if the agents do not know the source of their messages. Since there is only one planning authority, this requirement may not be relevant for the firms; if certain kinds of messages are transmitted only by the CPB the firms would know the source of these messages. As far as the CPB is concerned, it would be desirable if messages did not have to be identified with particular firms. In particular, if the aggregate response of the firms is all that the CPB needs to determine its message, this must be considered a significant advantage. Clearly, this would be a stronger requirement than anonymity and a process satisfying this requirement will be called *aggregative*. Another informational requirement that Hurwicz (1969) imposes on a decentralized process is that the message space M be R^k. Calsamiglia (1977) considers a somewhat less restrictive condition on the amount of information that needs to be transmitted. He defines a process to be *point valued* if M is some finite dimensional Euclidean space.

D.10. A process $\pi = (M, f, h)$ is *informationally decentralized* if it is external, aggregative and point valued, that is if

$$m_{t+1}^j = f^j\left(\sum_{)j(} m_t^j, m_t^j, m_t^{n+1}; e^j\right), m_{t+1}^{n+1} = f^{n+1}\left(\sum_j m_t^j, m_t^{n+1}; e^{n+1}\right)$$

and

$$M = R^s,$$

where $\Sigma_{)j(} m^j$ refers to the summation across the messages of all except the *j*-th firm and *s* is a positive integer.

While most of the planning processes in the literature are external and aggregative, many of them are not point-valued in the above sense. In particular, Malinvaud's (1967) process is one in which the agents transmit point-valued messages in every time period but the response of the CPB depends also on the messages received in the past. Such a process would not be informationally decentralized according to the above definition; however, there is something to be said for making a distinction between messages and memory, and between a process in which messages at each point in time are infinite dimensional and one in which finite dimensional messages are transmitted but the CPB has a memory of past messages.

Processes of the later variety have also been termed decentralized and while this may not be unreasonable, it has led to some confusion (see Cremer, 1978).

2 The classical environment

In this section we discuss planning processes designed for the classical environment in which there are no externalities, production sets are convex and the utility function is quasi-concave. In this setting, the competitive allocation has the attractive welfare properties that it is Pareto optimal and any Pareto optimal allocation can be sustained as a competitive allocation with a redistribution of initial resources. In an economy with a single utility function, a competitive allocation is unambiguously optimal.

We begin by considering Lange's process as formalized by Arrow and Hurwicz (1960). As mentioned earlier, the Lange–Arrow–Hurwicz (LAH) process is closely related to the Walrasian tâtonnement. Arrow and Hurwicz consider the following process: the CPB announces prices p and the firms choose profit maximizing production plans. The CPB computes a consumption plan to maximize $U(x)–px$. The prices are then varied in proportion to excess demand.

Arrow and Hurwicz formulate the planning problem as a programming problem and apply the gradient method (or method of steepest ascent) to find its solution. They formulate their process in continuous time in the activity analysis framework. We now formally describe the LAH process in its discrete version, as transposed by Malinvaud (1967) to the model presented in Section 1 above. Given prices p, firms choose their profit maximizing plans and the CPB chooses the consumption plan which maximizes $U(x)–px$. The price of a commodity is then increased by an amount proportional to its aggregate excess demand, the coefficient of proportionality being a positive constant ρ provided this change does not make the price negative. The responses of the agents can now be defined formally:

$$\text{(i)} \qquad y_t^j = \left\{ y_t^j \in Y^j \mid p_{t-1} y_t^j \geq p_{t-1} z \text{ for all } z \in Y^j \right\} j = 1, \ldots, n,$$

$$\text{(ii)} \qquad x_t = \{ x_t \in X \mid U(x_t) - p_{t-1} x_t \geq U(z) - p_t z \text{ for all } z \in X \},$$

$$\text{(iii)} \quad p_{t,l} = \max \left\{ 0,\ p_{t-1,l} + \rho \left(x_{t-1,l} - \sum_j y_{t-1,l}^j - \omega_l \right) \right\} l = 1, \ldots, k.$$

It is clear that in order for the process to be convergent, a Pareto optimal allocation must exist. This in turn will be guaranteed if, for example, all the production sets and the consumption sets are compact. The assumption that the above mappings are all single-valued, that is, there is a unique production plan that maximizes profits for each firm, given p and a unique consumption plan that maximizes $U(x)–px$, also turns out to be important for the convergence properties of the process. While the process in continuous time is convergent (see Theorem 12 in Arrow and Hurwicz, 1960), Uzawa (1958) showed that the discrete version of the LAH process converges only approximately. The following result is the version presented in Malinvaud (1967).

Theorem 1: If there is a unique Pareto optimal allocation $(\bar{x}, \bar{y})$ and the functions defined by (i) and (ii) are single-valued, the process defined by (i), (ii) and (iii) is approximately convergent in the following sense: for any $\varepsilon > 0$ there exist ρ_0 and t_0 both depending on ε such that if $\rho \leq \rho_0$ then for $t \leq t_0$ the distance between $h(m_{t:e})$ and $(\bar{x}, \bar{y})$ is no greater than the distance between $h(m_{t-1}, e)$ and $(\bar{x}, \bar{y})$ and for $t \geq t_0$ the distance between $h(m_t; e)$ and $(\bar{x}, \bar{y})$ is no greater than ε.

It is easy to see that this process is not feasible since, out of equilibrium, aggregate excess demand for some commodities may be positive. There is also the problem that, since the function $\rho(\varepsilon)$ is not known, it is not possible, given some ε, to choose the value of ρ to be most efficient.

Malinvaud (1967) proposes two other processes which are feasible, monotonic, and convergent but are not decentralized according to D.10 since the CPB is required to remember the messages conveyed by the firms in the past. Malinvaud's first process is designed only for a linear economy and is based on Taylor's (1929) proposal. Each firm is assumed to have a set of fixed coefficient techniques that can be operated under constant returns to scale. The CPB announces prices corresponding to which firms respond with a cost minimizing technique. The CPB then solves the open Leontief model to obtain prices which would make firms' proposed techniques earn zero profits and a consumption plan which maximizes utility at these prices. This process is then shown to satisfy Malinvaud's criteria under certain conditions. Malinvaud's second process covers a more general environment and we shall now discuss this in somewhat more detail.

This process is an application of the Dantzig–Wolfe (1961) decomposition algorithm to the planning problem. The CPB builds up an approximation of the firms' production sets based on messages received from them in the past. At each stage firms reveal their profit maximizing production plans, given the prices conveyed by the CPB. Assuming that all the

production sets are convex, the CPB can construct a subset of a firm's production set by taking the convex combination of all the production plans revealed by that firm in the past. The CPB then solves the programming problem of maximizing utility subject to the resource constraints and the technological constraints as given by its construction of the firms' production sets. In the next stage the shadow prices obtained from the programming problem are announced as prices and the process continues. For the process to start it is assumed that the CPB has initial information about at least one feasible production plan for each firm.

We can now define the process formally. Let Δ denote the k–1 dimensional simplex and $Y_t^j = \mathrm{Con}(y_t^j, y_{t-1}^j)$ where Con denotes convex hull. Let $(\bar{x}_t, \bar{y}_t)$ denote the allocation which solves the programming problem at t, that is,

$$(\bar{x}_t, \bar{y}_t) = \left\{(x, y) \in X x Y | x_t \leq \sum_j y_t^j + \omega \text{ and } U(x_t) \geq U(z) \text{for all } z \in \sum_j \mathrm{con}\left(Y_t^j\right) + \omega\right\}.$$

We shall say that $p \in \Delta$ *supports* the allocation $(\bar{x}_t, \bar{y}_t)$ if

1. (a) $U(x) \geq U(\bar{x}_t)$ implies $px \geq p\bar{x}_t$
2. (b) $y \in Y^j$ implies that $py \leq p\bar{y}_t^j$ for all j.

The process is defined by the following equations:

1. (i) $y_t^j = \left\{y \in Y^j | py \geq p\bar{y} \text{ for all } \bar{y} \in Y^j\right\}, \; j = 1, \ldots, n$.
2. (ii) $p_t = \{p \in \Delta | p \text{ supports } (\bar{x}_t, \bar{y}_t)\}$.

The plan at stage t is simply defined as $(\bar{x}_t, \bar{y}_t^j)$.

The following assumptions are sufficient for this process to satisfy Malinvaud's criteria. (A1) X is closed, convex and bounded from below. $U(x)$ is continuous, quasi-concave and locally non-satiated (A2) Y^j is convex and compact for all j. (A3) the CPB knows a feasible program (x_1, y_1). We can now state:

Theorem 2: (Malinvaud, 1967) If (A1), (A2) and (A3) are satisfied the process defined by (i) and (ii) is feasible, monotonic and convergent.

To see that this process is feasible notice that, given (A2), (i) always has a solution and, given (A1) and (A2), the programming problem, for any t, has a solution and this solution constitutes the plan for that time period. We can appeal to the second welfare theorem (see, for example Theorem 6.4 in Debreu, 1959) to assert that (ii) also has a solution.

Monotonicity is an obvious property of this process since the constraint sets in the programming problem of time t are contained in the constraint sets of time t+1. Convergence is established by considering a limit argument, using the fact that all the plans lie in a compact set. We refer to Malinvaud (1967) for the proof.

As the above theorem shows, this process has better performance properties than the LAH process. However, its information requirements are much stronger. The CPB is required to have memory and to know of a feasible allocation. It also solves a rather complicated programming problem at each stage. Moreover, to implement the plan the firms are instructed to follow the production plans computed by the CPB. While these plans are consistent with profit maximization, a specific instruction has to be issued to each firm. But this problem can be avoided in the simpler case where all production sets are strictly convex. In this case, the equilibrium plan can be implemented simply by announcing shadow prices and letting the firms find their unique profit maximizing production plans.

Weitzman (1970) proposed a process which is in a sense a dual of Malinvaud's process. The CPB has a belief about a firm's production set, which is not necessarily correct, and, given these imaginary production sets, it solves the programming problem and provides each firm with a production plan as a target. If the firm finds that this target is not feasible it responds with an efficient plan and a corresponding marginal rate of substitution. The CPB then constructs a new production set which is the intersection of its previous one with the half space determined by the firm's announced efficient plan and marginal rate of substitution. The CPB again solves the programming problem and announces new targets (see Figure 1). Not only is this process convergent, if the production sets are polyhedral, convergence is achieved in a finite number of steps. However, it is not feasible since the CPB's targets may not be feasible for the firms.

Figure 1

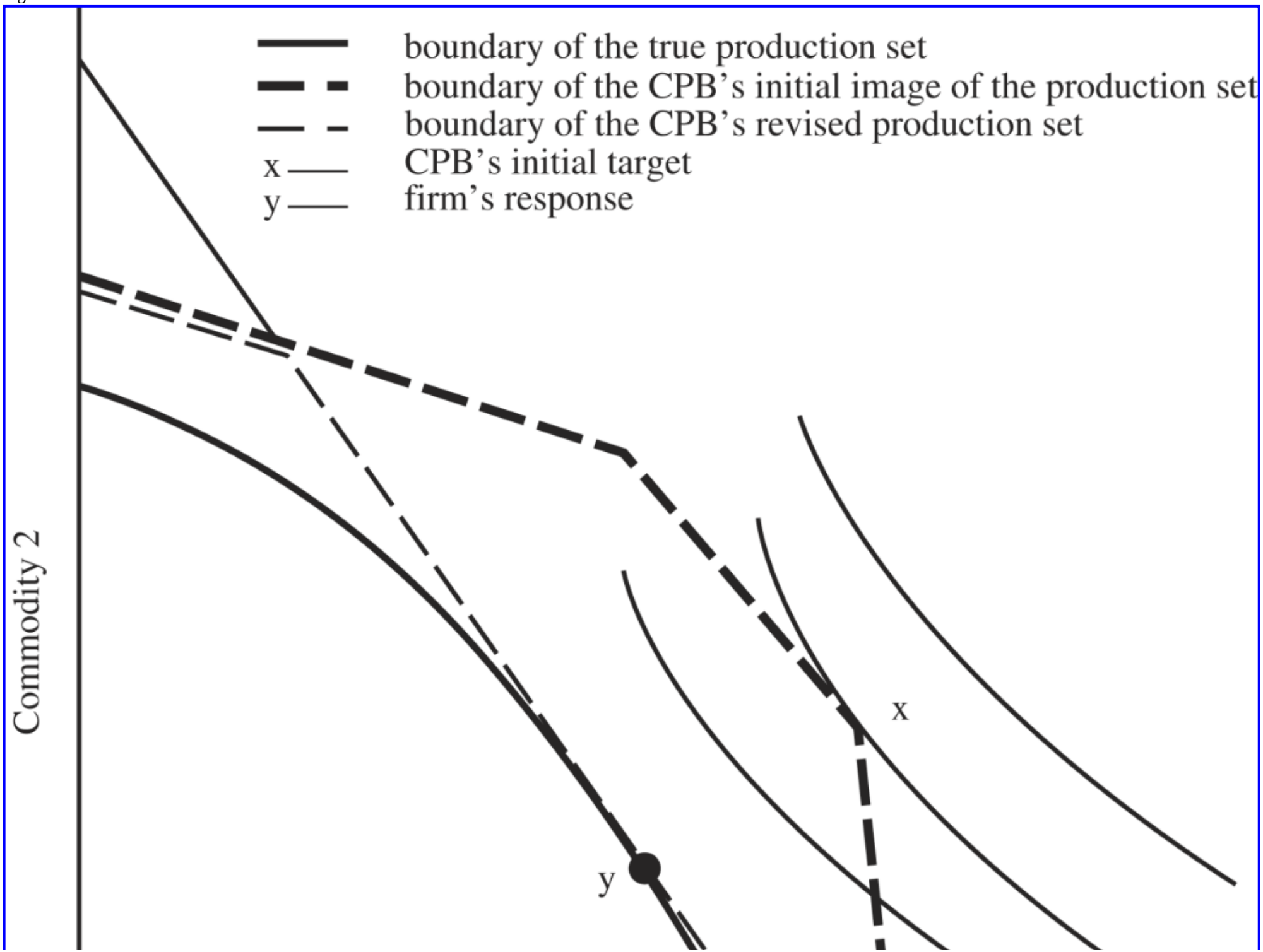
boundary of the true production set
boundary of the CPB's initial image of the production set
boundary of the CPB's revised production set
x
CPB's initial target
y
firm's response
Commodity 2
x
y

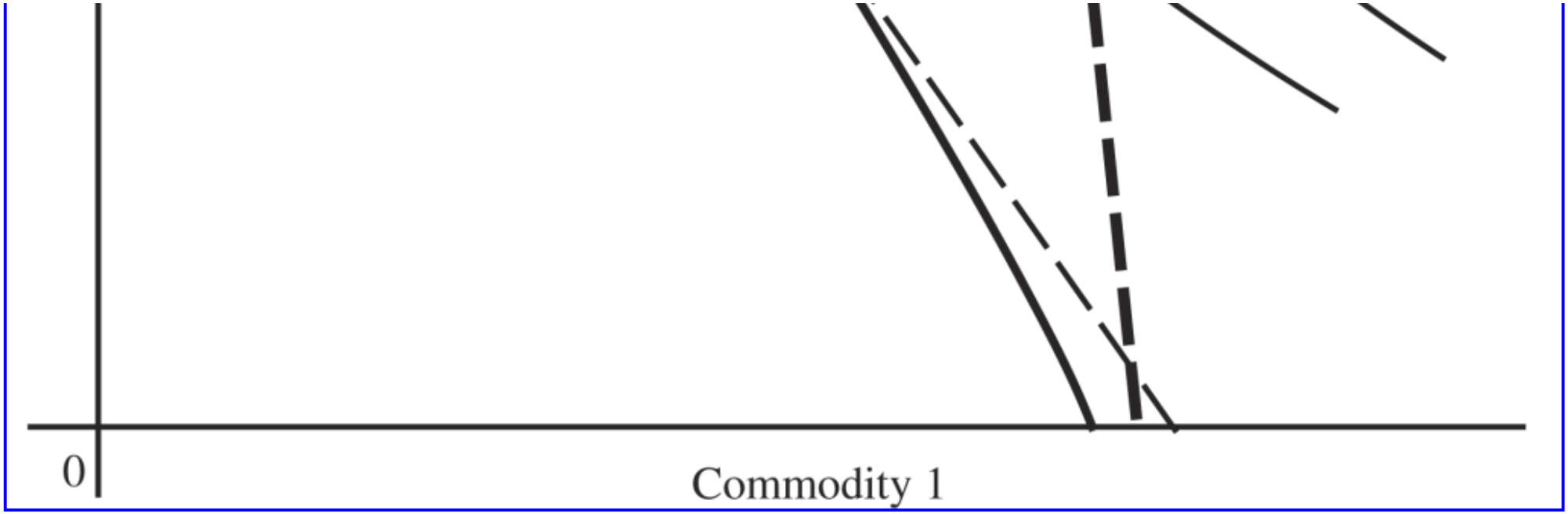

Another process which uses production quotas rather than prices as signals is one due to Kornai and Liptak (1965). Their process is formulated for a linear economy in which the CPB's utility function is separable among the firms' outputs. The CPB allocates resources to the firms which respond with rates of substitution and the CPB reallocates resources in response to the value of the allocated resources at the shadow prices. They model the interaction between the CPB and the firms as a game and show that the process is convergent.

3 Increasing returns

Heal (1969) proposed a non-price gradient process which locates local maxima even for economies with increasing returns. The CPB allocates the inputs among firms which then respond with efficient output levels and marginal productivities. The CPB then reallocates the inputs towards the firms with higher marginal productivities. The process can be most easily understood in the simple setting in which all firms produce an identical output using m primary resources. Departing from our notation of Section 1, we shall denote by y_i the amount of output produced by firm i and by f_i firm i's production function. The amount of input j used by firm i is denoted x_{ij} and the technological constraints may be stated as follows,

$$y_i = f_i(x_{i1}, \ldots, x_{im}) i = 1, \ldots, n, x_{ij} \geq 0, \text{ for all } i \text{ and } j.$$

Let R_j denote the aggregate endowment of the jth resource. The resource constraints can be stated as

$$\sum_i x_{ij} \leq R_j \text{for all } j.$$

The objective of the planning process is to find $((x_{ij}))$ to maximize $\Sigma_i y_i$ subject to the technological and resource constraints. Let f_{ij} denote firm i's marginal productivity of the j-th

input and let a dot over a variable denote its rate of change. The process starts with the CPB allocating x_{ij} to the firms subject to the resource constraints. The CPB then raises the allocation of input j to a firm if its marginal productivity is greater than a certain average productivity and lowers it if it is lower than the average, subject to the non-negativity constraints. Formally, the rate of adjustment is determined by the following equations

$$\dot{x}_{ij} = f_{ij} - \mathrm{Av}(K_j) f_{ij}, \text{ if } i \in K_j 0 \text{ otherwise,}$$

where $\mathrm{Av}(K_j)f_{ij}$ denotes the average of f_{ij}'s contained in the set K_j. The set K_j is constructed (see Heal, 1969) so that the non-negativity constraints are not violated in applying the adjustment equations and it satisfies the following property

$$K_j = \{i | x_{ij} > 0 \text{ or } x_{ij} = 0 \text{ and } f_{ij} > \mathrm{Av}(K_j) f_{ij}\}.$$

K_j includes firms with positive allocations of input j or firms with a zero allocation but a marginal productivity higher than the average.

We can now state the following theorem, which applies to the simple model we are considering but also extends to the more general case where firms produce different commodities.

Theorem 3: (Heal, 1969) If all f_i have continuous, finite first derivatives and the initial allocation is feasible, the process defined above is feasible and monotonic. Moreover, every limit point of the process satisfies the necessary conditions for Pareto optimality. If the initial allocation is not a local minimum, then the limit points are to local minima.

To see that the process is feasible, notice that

$$\sum_i \dot{x}_{ij} = \sum_{i \in K_j} \left[f_{ij} - (1 / K_j) \sum_{i \in K_j} f_{ij} \right] = 0, \text{ for all } j.$$

Thus, if the initial allocation is feasible so are all other allocations. To establish monotonicity, we consider

$$\dot{V} = \sum_i \sum_j f_{ij} \dot{x}_{ij}$$

which can be written as

$$\dot{V} = \sum_j \sum_{i \in K_j} f_{ij} \left[f_{ij} - (1 / |K_j|) \sum_{i \in K_j} f_{ij} \right] = \sum_j \sum_{i \in K_j} \left[f_{ij} - (1 / |K_j|) \sum_{i \in K_j} f_{ij} \right]^2 \geq 0.$$

Thus, $\dot{y} \geq 0$ and $\dot{y} = 0$ if and only if $f_{ij}=f_{kj}$ for all i and $k \in K_j$ for all j. It is easy to see that the equality of f_{ij} for all i in K_j is the necessary condition for optimality. This implies that y increases monotonically except when the necessary conditions for optimality are satisfied. In particular, if the initial allocation is not a local minimum, the equilibrium allocation, arrived at through monotonic increases, cannot be a local minimum. Hori (1975) showed that the convergence to a point of inflection is unlikely in a well defined sense. A discrete version of this gradient process would also be approximately convergent in the sense described in Theorem 2 above.

Since this process requires the CPB to respond with allocations to firms, the informational requirements are much stronger than those of a price guided process in which a common price vector is given out to the entire production sector. In the general case where firms produce many commodities the CPB uses marginal valuations not only to allocate inputs but also output combinations to each firm (see Heal, 1973, ch. 8). However, unlike the Malinvaud or the Weitzman process, the CPB is not required to have a memory. It is also possible to modify this process to take advantage of the informational efficiency which is characteristic of the price guided processes. Such a mixed planning process was formulated by Heal (1971) and is similar to one proposed by Marglin (1969). In Heal's (1971) process the CPB allocates resources to the firms and also provides them with prices of the final goods. The firms inform the CPB of their profit maximizing output bundles and also of the marginal productivity of the inputs. The CPB reallocates inputs as in the previous process and announces new output prices which reflect the marginal rates of substitution in consumption. The performance of this process is similar to that of the previous one with the important difference that the CPB does not determine the complete allocation at each step. The substitution of one output for another is carried out by the firms depending on the common price vector for outputs announced by the CPB. Aoki (1971a) proposed a mixed planning process which combines the LAH process with Heal's (1969) process. He considers an economy with increasing returns in which there is one input such that if this is fixed, each firm faces decreasing returns with respect to all the other inputs. The CPB allocates this input to the firms in accordance with its marginal profitability and the LAH process is then used to allocate all the other resources. This process is clearly more complex since the LAH process is used at each step in which the essential input is reallocated, but it does converge to a local maximum.

Another approach to planning in economies with increasing returns is the modified LAH process. Arrow and Hurwicz (1960) showed that their process could deal with linearities and non-convexities if the Lagrangian is suitably modified so that it becomes strictly concave and the gradient method is then applied to locate a saddlepoint of this concavified Lagrangian. There is, however, a significant difference. The modified Lagrangian expression is no longer a sum of functions each involving a different variable and it is no longer simply possible to determine demands and supplies given the prices. The CPB and firms need the entire price schedule and this makes this modified process less informationally decentralized than the original LAH process.

All the processes that we have so far considered in this section, depending as they do on first order properties of the relevant functions, cannot guarantee convergence to a global optimum. They also seem to be less informationally decentralized than processes for the classical environment. The natural question to be raised at this stage is whether it is possible to formulate a decentralized process which converges to a global optimum in an environment with increasing returns. Calsamiglia (1977) showed the answer to this question is no. He begins by making a rather important point about the interpretation of a local maximum. He provides an example of an allocation which is a local maximum but does not satisfy aggregate production efficiency. While at a local maximum it is not possible to make marginal changes to increase utility, it may be possible to increase utility simply by reorganizing production among the firms to produce more of each commodity. But, as he then proves, even in simple economies with increasing returns there does not exist a decentralized process which converges to a global optimum.

It is however, possible to construct a process which has nice convergence properties at the cost of giving up decentralization as defined in D.10. This was shown by Cremer (1977). He considers a quantity–quantity algorithm in which the CPB, as in Malinvaud (1967) and Weitzman (1970), possesses a memory and builds up successive approximations of the firms' production sets and solves the programming problem. This process is in many respects similar to Weitzman's process. Convexity of the production sets was used crucially in Weitzman's process to ensure that when the CPB constructs a new production set, by considering the announced marginal rate of substitution, it knows that no point above the corresponding hyperplane need be considered again. In the presence of increasing returns this is no longer true and in Cremer's process firms do not respond with marginal rates of substitution. The CPB only knows that if a firm responds with a feasible production plan then all production plans which are greater than it can be ruled out of further consideration.

Figure 2 shows how the CPB revises its information about the firm's technology. It is assumed that the CPB knows that the optimal production plan $y^* \leq w$. It announces w as a target. If this is not feasible the firm responds with some y^1 which is feasible and strictly less than w. The CPB then knows that it must now consider only points less than or equal to either v^1 or v^2. If the utility at v^2 is higher than at v^1 the CPB considers its new approximate production set to be the set of all points in Y^1 but equal to or less than v^2. Under certain boundedness conditions it can be shown that this process converges to a global optimum. Since the targets are not necessarily feasible nor is the process.

Figure 2

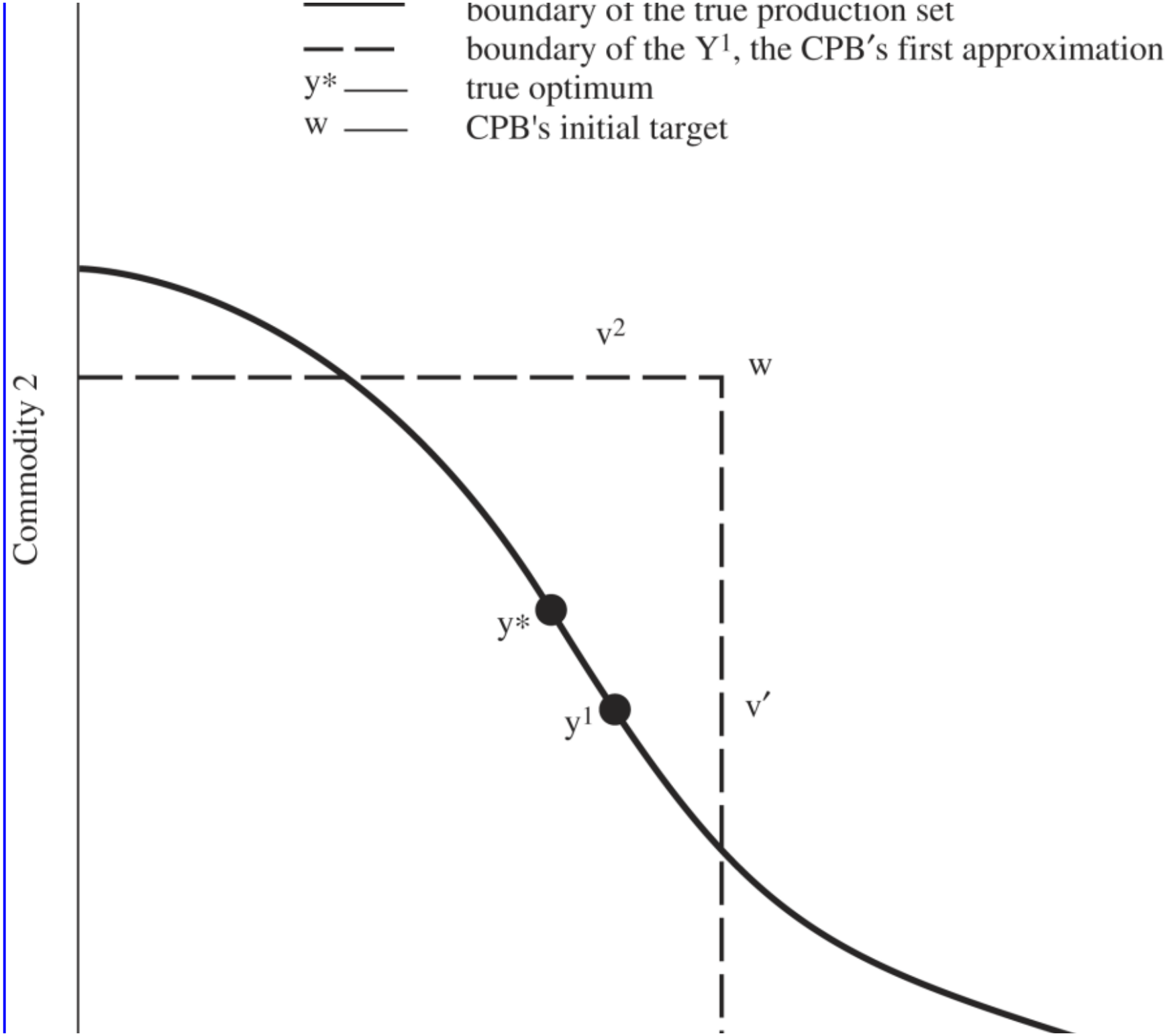
boundary of the true production set
boundary of the Y1, the CPB′s first approximation
y* true optimum
w CPB's initial target
Commodity 2
v2
w
y*
y1
v′

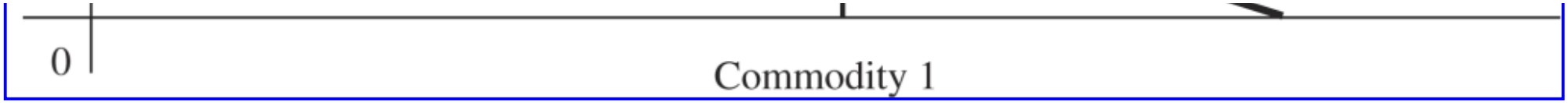

4 Public goods

This section draws heavily on Tulkens (1978). We begin by considering the simple setting of an economy with a single private good *y* and a single public good *z*. There are *n* consumers with continuously differentiable and strictly quasi-concave utility functions $U^i(x^i)$, where $x^i=(y^i, z^i)$. The public good is produced according to the technology of the form $w=g(z)$, where *w* represents the private good input and *g* is assumed to be convex. A feasible allocation $((x^i))$ satisfies the conditions that

$$\sum_i y^i + w = \sum_i \omega_i,\ z^i = z \text{ for all } i \text{ and } w = g(z).$$

Lindahl (1919) in his positive solution to the public goods problem proposed a process, the convergence properties of which were analyzed by Malinvaud (1971). The Lindahl process concerns a two consumer economy in which the public good is produced under constant marginal cost γ. Each consumer is assigned a share, θ^i in the price of the public good so that $\theta^1 + \theta^2 = \gamma$. Consumers take as given their personalized prices or unit taxes θ^i to determine their demands for the public good. The supply of the public good is made equal to the lower of these two demands and the CPB adjusts the unit taxes by raising the tax on the consumer with the higher demand and lowering it for the other. The process continues as long as the utilities of both the consumers rise. Malinvaud (1971) showed that utilities would not rise monotonically until the two demands become identical and, therefore, this process does not converge to a Lindahl equilibrium. He suggested a modification which ensures convergence to a Pareto optimal allocation (though not necessarily to a Lindahl allocation). In this modified process the CPB announces not only unit taxes but also lump-sum taxes, T^i such that $\Sigma_i T^i = 0$. Let $d^i(\theta^i, T^i)$ refer to consumer *i*'s demand for the public good and $\bar{d}$ the corresponding average demand. The CPB adjusts *i*'s unit tax in proportion to the difference between d^i and $\bar{d}$. Supply is made equal to the average demand and T^i is adjusted to compensate *i* for the change in θ^i. Formally the adjustment equations are, (i) $\dot{\theta}^i = a[d^i - \bar{d}]$ for all *i*, (ii) $\dot{T}^i = -(1/n)\Sigma_i d^i \theta^i$ for all *i*, where *a* is a positive constant. While this process converges to a Pareto optimal allocation, it is neither feasible nor monotonic.

An alternative would be to consider a process in which the CPB responds with quantities rather than prices. The Malinvaud–Drèze–de la Vallée Poussin (MDP) process, formulated by Malinvaud (1970–71) and Drèze and de la Vallée Poussin (1971), is a quantity guided process in which the CPB announces an allocation and the agents respond with rates of substitution. Starting with a feasible allocation, the firm reports its marginal cost γ and each consumer reports his or her marginal rate of substitution of the public good for the private good π^i. The adjustment takes place according to the following differential equations, (i) $\dot{z}_t = \dot{z}_t^i = a(\Sigma_i \pi_t^i - \gamma_t)$ for all *i*, (ii) $\dot{w}_t = \gamma_t \dot{z}_t$ (iii) $\dot{y}_t^i = -\pi_t^i \dot{z}_t + \delta^i a(\Sigma_i \pi_t^i - \gamma_t)^2$ for all *i*, where *a* is a positive constant and $\delta^i \geq 0$ for all *i* and $\Sigma_i \delta^i = 1$.

Since the process starts at a feasible allocation, (ii) ensures that the process is feasible. It has also been shown that it converges to an allocation at which the first order conditions for optimality are satisfied, that is, the sum of the marginal rates of substitution equals the marginal cost. The MDP process is also monotonic. To see this consider

$$\dot{U}^i = U^i y^i \left(\dot{y}^i + \pi^i \dot{z}\right).$$

Using (i) and (ii) this can be rewritten as

$$\dot{U} = U^i \gamma^j \delta^i a \left(\sum_i \pi^i - \gamma \right)^2 \geq 0.$$

While the MDP process converges to some Pareto optimal allocation depending on the choice of the distribution profile ((δ^i)), Champsaur (1976) has shown that the process is neutral in the sense that given any initial allocation and any Pareto optimal allocation which is Pareto superior to this allocation, there exists a distribution profile with which the MDP process converges to the given optimum. A discrete time version of the MDP, with the same performance properties, was provided by Champsaur, Drèze and Henry (1977).
Malinvaud (1970–71) and Drèze and de la Vallée Poussin (1971) also extend the MDP process to an economy with many private and public goods by considering the MDP process as described above for public goods and a quantity guided process for the private goods. Another alternative, considered in Aoki (1971b), Malinvaud (1972) and Champsaur, Drèze and Henry (1977), is to construct a process which combines the MDP process with a price guided process for private goods. These processes, however, have to deal with a well known problem, namely one of ensuring convergence of a price guided process in an economy with many consumers without making the gross substitutability assumption.
Aoki (1971b) considers an economy with many private and public goods and many firms and consumers. He avoids the income distribution problem by specifying a social welfare function. The CPB announced prices of the private goods and quantities of the public goods. Firms maximize profits and report input demands and marginal costs for public goods. The CPB increases private goods prices according to the difference between marginal utilities and prices and the public goods levels are adjusted according to the difference between marginal utilities and marginal costs. This process is feasible, monotonic and convergent.
Malinvaud (1972) formulates a price guided process for allocating not only private goods but also public goods. The gross substitutability assumption is avoided by specifying individual incomes as proportions of aggregate income and revising them during the process (notice that Malinvaud's, 1971, price guided process also made use of lump-sum transfers). This process converges locally but is neither feasible nor monotonic.
Champsaur, Drèze and Henry (1977) present a process which combines in a sequential way an MDP process for public goods allocation with a price guided process for private goods allocation. Given public goods' levels a price guided process is used to allocate private goods. Then keeping fixed the levels of all except one numeraire private good, the MDP process is applied to allocate public goods. This process is shown to be feasible, monotonic and convergent to some Pareto optimal allocation.
Given the difficulty in using a price guided process when there are many consumers, it is perhaps not surprising that a satisfactory process which converges to a Lindahl equilibrium has not been established, although some results are available in this direction (see Milleron, 1974).

See Also

- decentralization
- Hotelling, Harold

Bibliography

Aoki, A. 1971a. An investment planning process for an economy with increasing returns. *Review of Economic Studies* 38, 273–80.

Aoki, A. 1971b. Two planning processes for an economy with production externalities. *International Economic Review* 12, 403–413.

Arrow, K. 1951. An extension of the basic theorems of welfare economics. In *Proceedings of the Second Berkeley Symposium*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J. and Hurwicz, L. 1960. Decentralization and computation in resource allocation. In *Essays in Economics and Econometrics in Honor of Harold Hotelling*, ed. R.W. Pfouts. Chapel Hill: University of North Carolina Press; reprinted in K.J. Arrow and L. Hurwicz (1977).

Arrow, K.J. and Hurwicz, L., eds. 1977. *Studies in Resource Allocation Processes*. Cambridge: Cambridge University Press.

Barone, E. 1908. The Ministry of Production in the collectivist state. In *Collectivist Economic Planning*, ed. F.A. von Hayek. London: Routledge, 1935.

Calsamiglia, X. 1977. Decentralized resource allocation and increasing returns. *Journal of Economic Theory* 14, 263–83.

Champsaur, P. 1976. Neutrality of planning procedures in an economy with public goods. *Review of Economic Studies* 43, 293–300.

Champsaur, P., Drèze, J.H. and Henry, C. 1977. Stability theorems with economic applications. *Econometrica* 45, 273–94.

Cremer, J. 1977. A quantity–quantity algorithm for planning under increasing returns to scale. *Econometrica* 45, 1339–48.

Cremer, J. 1978. A comment on 'Decentralized planning and increasing returns'. *Journal of Economic Theory* 19, 217–21.

Dantzig, G.B. and Wolfe, P. 1961. The decomposition algorithm for linear programs. *Econometrica* 29, 767–78.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40, 588–92.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. Cowles Foundation Monograph No. 17. New York: Wiley.

Drèze, J.H. and de la Vallée Poussin, D. 1971. A tâtonnement process for public goods. *Review of Economic Studies* 37, 133–50.

Hayek, F.A. von. 1935. The present state of the debate. In *Collectivist Economic Planning*, ed. F.A. von Hayek. London: Routledge.

Heal, G.M. 1969. Planning without prices. *Review of Economic Studies* 36, 346–62.

Heal, G.M. 1971. Planning, prices and increasing returns. *Review of Economic Studies* 38, 281–94.

Heal, G.M. 1973. *The Theory of Economic Planning*. Amsterdam: North-Holland.

Hori, H. 1975. The structure of the equilibrium points of Heal's process. *Review of Economic Studies* 42, 457–67.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical Methods in Social Sciences*, ed. K.J. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press. Reprinted in Arrow and Hurwicz (1977).

Hurwicz, L. 1969. On the concept and possibility of informational decentralization. *American Economic Review* 59, 513–34.

Hurwicz, L. 1973. The design of resource allocation mechanisms. *American Economic Review* 58, 1–30; reprinted in Arrow and Hurwicz (1977).

Kornai, J. and Liptak, T. 1965. Two-level planning. *Econometrica* 33, 141–69.

Lange, O. 1936–37. On the economic theory of socialism. *Review of Economic Studies*. Reprinted in *On the Economic Theory of Socialism*, ed. B. Lippincott, Minneapolis: University of Minnesota Press, 1938.

Lindahl, E. 1919. Just taxation – a positive solution. In *Classics in the Theory of Public Finance*, ed. R.A. Musgrave and A. Peacock, London: Macmillan, 1958.

Malinvaud, E. 1967. Decentralized procedures for planning. In *Activity Analysis in the Theory of Growth and Planning*, ed. M.O.L. Bacharach and E. Malinvaud. London: Macmillan.

Malinvaud, E. 1970–71. Procedures pour la determination d'un programme de consommation collective. *European Economic Review* 2, 187–217.

Malinvaud, E. 1971. A planning approach to the public goods problem. *Swedish Journal of Economics* 73, 96–112.

Malinvaud, E. 1972. Prices for individual consumption, quantity indicators for collective consumption. *Review of Economic Studies* 34, 385–406.

Marglin, S.A. 1969. Information in price and command systems of planning. In *Public Economics*, ed. J. Margolis and H.K. Guitton. London: Macmillan.

Milleron, J.C. 1974. Procedures to a Lindahl equilibrium corresponding to a given distribution of income. Paper presented at the European Meeting of the Econometric Society in Grenoble.

Mises, L. von. 1922. *Socialism*. London: Jonathan Cape, 1936.

Pareto, V. 1897. *Cours d'économie politique*, vol. 1, Lausanne: Librarie de l'Université.

Samuelson, P.A. 1949. Market mechanisms and maximization. In *Collected Scientific Papers of Paul A. Samuelson*, ed. J.E. Stiglitz. Cambridge, MA: MIT Press, 1966.

Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19, 1–8. Reprinted in *On the Economic Theory of Socialism*, ed. B. Lippincott, Minneapolis: University of Minnesota Press, 1938.

Tulkens, H. 1978. Dynamic processes for public goods: an institution-oriented survey. *Journal of Public Economics* 9, 163–201.

Uzawa, H. 1958. Iterative methods for concave programming. In *Studies in Linear and Non-Linear Programming*, ed. K.J. Arrow, L. Hurwicz and H. Uzawa. Stanford: Stanford University Press.

Weitzman, M. 1970. Iterative multi-level planning with production targets. *Econometrica* 38, 50–65.

How to cite this article

Vohra, Rajiv. "planning." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000094> doi:10.1057/9780230226203.1290

The New Palgrave Dictionary of Economics Online

Plekhanov, Georgii Valentinovich (1856 – 1918)

M. Falkus

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; Lenin, V. I.; Marxism; Plekhanov, G. V.; socialism

Article

Plekhanov was a major figure in the development of Marxist economic and political philosophy during the late 19th century. His importance springs from four principal sources. He was the first Russian intellectual to apply Marxist theory to Russian conditions. In so doing, he undermined the intellectual foundations of the Populists (*Narodniki*) and showed the relevance of Marxist economic determinism to Russia. Secondly, he exerted a profound influence upon the Russian revolutionary intelligentsia, persuading many of them to abandon Populism in favour of Marxism. Plekhanov was one of the founders of the Marxist Russian Social Democratic Party. Thirdly, the originality and perception shown in Plekhanov's own voluminous and wide-ranging writings show him to be an outstanding Marxist theoretician. Finally, the approval given Plekhanov's writings by Marx and, especially, Lenin (despite their later disagreements) has assured Plekhanov of an honoured place in Soviet histories of the development of socialist philosophy. Indeed, Plekhanov was one of the two figures whose writings were specifically acknowledged by Lenin as leading to his own conversion to Marxism; the other was Marx. Plekhanov was born on 29 November 1856 in the village of Gudalovka in what was then the province of Tambov (Lipetsk Oblast). He was the son of a wealthy nobleman and attended military college in Voronezh and the Konstantin Cadets' College in St Petersburg in 1873–4 before entering the St Petersburg Institute of Mines. Here he became influenced by the revolutionary movements of the time and was eventually expelled in 1876 for his part in such activities. In 1875 he had joined the *Narodniki* and in the following year he joined the newly formed *Zemlya i Volya* (Land and Liberty) *Narodnik* organization – Russia's first political party. This group believed that Russia's future lay with the peasant masses, and that the peasants should be given land. Plekhanov soon became one of the leading *Narodnik* writers and activists, and took part in the 'going to the people' movement. He also gave a speech at a major demonstration organized in 1876 by *Zemlya i Volya* in front of St Petersburg's Kazan Cathedral. In 1879 the *Narodnik* movement split, the majority faction advocating the use of terrorist tactics.

Plekhanov favoured a more moderate approach, and together with a small group of other leading *narodniks* (including Pavel Axelrod and Leo Deutsch) formed the non-violent *Cherny Peredel* movement (Black Repartition – that is, the movement wanted repartition of the fertile Black Soil lands to the peasantry).

In January 1880 Plekhanov emigrated to Europe to escape persecution from the tsarist authorities. He remained in exile until 1917, living in Switzerland, France, Italy and elsewhere, travelling widely throughout the continent. In western Europe he made contact with numerous other Russian revolutionary exiles and also became deeply interested by Marxist thought. From about 1882 he became a fervent advocate of Marxism, and in his writings he now sought to establish the relevance of Marxism to Russian conditions and to undermine the intellectual foundations of Russian Populism. In 1883 Plekhanov founded in Geneva Russia's first Marxist Social Democratic organization, the Liberation of Labour. The group translated into Russian and published many works by Marx and Engels, Plekhanov himself translating the *Communist Manifesto*.

During the 1880s and 1890s Plekhanov wrote his most influential works, denouncing not only the Populists but the Legal Marxists and the Economists (Marxist factions which developed after 1895), and he put forward his own interpretation of the path towards socialism which Russia was to follow. The root of his philosophy was in what he termed ‘scientific’ historical materialism, exposing the *narodniks* as ‘unscientific’. In Plekhanov's view, revolution could not succeed unless it has the support of the class-conscious masses. Revolution could not come from the agrarian peasantry, and must come from the urban proletariat. As he argued in *Socialism and the Political Struggle* (1883) and *Our Differences* (1885a), the utopian socialists (Blanquists) were mistaken in their reliance on intellectual conspiracy alone: revolution could succeed only as a result of a class struggle emanating from the working classes. It therefore became important for Plekhanov to demonstrate that Russia's path towards socialism could not come, as the *narodniki* argued, from the village-based commune (*mir*) and the peasantry. Capitalism in Russia was a necessary phase of historical development and was not ‘accidental’ or ‘non-Russian’. Indeed, in Russia of the 1880s capitalism was already a reality.

To be sure, Plekhanov's theories contained many obscurities and contradictions. Fundamental were the dichotomies between economic determinism and the role of the revolutionary, and also between the reliance on the class-conscious urban masses and the evident industrial backwardness of Russia. Plekhanov ‘solved’ the problems, albeit unsatisfactorily, by arguing that the Russian revolution could be accelerated by the role of the revolutionary intelligentsia, whose activities were to compensate for the lack of a middle class. He wrote in *Our Differences:* ‘Our capitalism will fade without ever having flowered.’

Particularly influential was Plekhanov's *The Development of the Monistic View of History*, which was brought to Russia by the Marxist publisher Potresov in 1894. Here Plekhanov elevated the ‘objectivism’ of Marx in contrast to the subjective values of the *Narodniki*. He wrote, ‘the criterion of truth lies not in me, but in the relations which exist outside of me’. Thus, objectivity was possible in social theory. Plekhanov drew from Marx, and from the traditions of the English economists and German historicists, the fundamental principle that economic forces determine social development.

Plekhanov was active in the Second International (1889) and attended its Congresses in Zurich (1893), Amsterdam (1904) and Copenhagen (1910). Together with Lenin, Martov and Potresov, Plekhanov founded *Iskra* (The Spark) in 1900 – the first Russian Marxist newspaper. In 1903 he worked jointly with Lenin to draw up the programme adopted at the famous Second Congress of the Social Democratic

Party, but it was shortly after this that Plekhanov broke with Lenin and the Bolsheviks and sided with the Mensheviks. During the Revolution of 1905 Plekhanov advocated an ‘opportunist’ alliance with the liberals, while in 1914 he supported the war against Germany for the defence of Russia (in opposition to Lenin and the Bolshevik position). In that year he formed the *Yedinstvo* (Unity) group, which was designed to bring together the Mensheviks and the anti-Lenin Bolsheviks, but its influence was negligible.

After the Revolution of February 1917 Plekhanov returned to Russia, supporting the Provisional government and the continuation of the war. He denounced the Bolshevik coup of October 1917, and shortly afterwards fell ill with tuberculosis. Ostracized by Lenin and terrorized by the Cheka, Plekhanov's wife took him to Finland, where he died on 30 May 1918.

Despite his differences with Lenin and the Bolsheviks after 1903, Plekhanov's writings continued to be highly regarded and widely studied in the Soviet Union. During the 1920s his library and archives were gathered from a number of European centres and taken to Leningrad, where the Plekhanov Library was established, and his complete writings were published.

Selected works

1883. *Socialism and Political Struggle.* In *Selected Philosophical Works*, vol. 1. Moscow: Progress Publishers, 1974.

1885a. *Our Differences*. In *Selected Philosophical Works*, vol. 1. Moscow: Progress Publishers, 1974.

1885b. *The Development of the Monistic View of History*. In *Selected Philosophical Works*, vol. 1. Moscow: Progress Publishers, 1974.

1895. *Anarchism and Socialism*. Trans. E. Marx Aveling. Chicago: C.H. Kerr.

1897. *Contributions to the History of Materialism*. Trans. R. Fox as *Essays in the History of Materialism*, London: Lane, 1934.

1923–7. *Sochineniya* [Works]. 24 vols. Moscow.

1987. *Selected Works*. 5 vols. London: Lawrence and Wishart.

Bibliography

Baron, S. 1963. *Plekhanov: The Father of Russian Marxism*. London: Routledge & Kegan Paul.

Baron, S. 1996. *Plekhanov in Russian History and Soviet Historiography*. Pittsburgh: University of Pittsburgh Press.

Chagin, B.A. 1963. *G.V. Plekhanov i ego rol' v razvitii Marksistskoi Filosofii*. Moscow and Leningrad: Akademii Nauk.

Harding, N. 1977. *Lenin's Political Thought, Volume 1: Theory and Practice of the Democratic Revolution*. New York: St. Martins.

Harding, N. 1981. *Lenin's Political Thought, Volume 2: Theory and Practice of the Socialist Revolution*. New York: St. Martins.

Keep, J.L.H. 1963. *The Rise of Social Democracy in Russia*. Oxford: Clarendon Press.

Kolakowski, L. 1978. *Main Currents of Marxism. Volume II: The Golden Age*. Oxford: Oxford University Press.

How to cite this article

Falkus, M. "Plekhanov, Georgii Valentinovich (1856–1918)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000101> doi:10.1057/9780230226203.1291

The New Palgrave Dictionary of Economics Online

pluralism in economics

Sheila C. Dow

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article explores the meaning of pluralism in economics and the arguments put forward in support of it. In particular, the distinction is drawn between methodological pluralism (support for variety in methodological approach) and a pluralist methodology (one which employs a variety of methods). Methodological pluralism usually takes the form of arguing that it is in the nature of knowledge about social systems that there will be variety of methodological approaches. But prescriptive arguments for a particular pluralist methodology may accompany the argument that this is the single best methodology (methodological monism).

Keywords

critical realism; methodological pluralism; monism; open systems; paradigms; pluralism in economics; pluralist methodology; positivism; postmodernism; rhetoric

Article

Pluralism is the advocacy of plurality (Mäki, 1997), or variety, and has been featured increasingly in discussions of economic methodology. Indeed, the International Confederation of Associations for Pluralism in Economics (ICAPE) is an umbrella organization for around 40 international economics organizations.

The term ‘pluralism’ was first used in modern methodological discourse by Bruce Caldwell (1982) in *Beyond Positivism*. Here he charted the growing dissent from a positivist methodology that had been put forward within a monist approach, that is, the advocacy of a single, best methodology. Positivism had prescribed progress in knowledge by means of empirical testing of propositions. Yet it had proved impossible to express all propositions in testable form, and difficult to derive definitive empirical tests for those propositions which were quantifiable (partly because of the so-called ‘Duhem–Quine problem’). Caldwell concluded that, rather than searching for some other, elusive, monist

methodological approach, economists should accept that a range of approaches could legitimately be sustained (though he later expressed the hope that there would some day be agreement on one best methodological approach: Caldwell, 1989). We will use Caldwell's term ‘methodological pluralism’ for pluralism at this level of methodological approach.

The implication of methodological pluralism is that methodologists should study (critically) a range of methodologies rather than seek to identify one (universally) best methodology. Practising economists must, however, choose one methodological approach or another, but may nevertheless support methodological pluralism by accepting that, while they may have good reason for their choice of approach, they accept that its superiority cannot be demonstrably proven. This chosen methodology itself may (or may not) be pluralist, that is, employ a variety of methods. Similarly, there is scope for plurality at the theoretical level, whether or not there is methodological pluralism, or a pluralist methodology. Thus, for example, Colander (2000) draws the distinction between the theoretical pluralism of mainstream economics and its monism in terms of formalist method (see also Goodwin, 2000). There has been considerable confusion in the literature between the meanings of pluralism at these different levels. No doubt this stems in part from the fact that many who support pluralism at one level also support it at another; but one does not necessarily entail either of the others.

Since Caldwell's initial proposal, a range of further arguments has been developed for methodological pluralism. Samuels (1997) has been a consistent exponent in practice of methodological pluralism, even before it had been explicitly identified (Caldwell, 1997). His support arises from a critique of prescriptive epistemology, on the constructivist grounds that our knowledge of the economy is situated, and the economy itself is a social construction, so that there is no scope for a common methodological approach to knowledge. Samuels's argument is consistent with the postmodern critique that emphasizes the absence of independent facts by which to test theories, which follows from the subjective and fragmented perceptions of experience (a plurality of understandings). Samuels is explicitly prescriptive at the meta-methodological level – he positively advocates methodological pluralism (but that is the limit of his prescription). Postmodernists agree with Samuels that there is no basis for any form of limitation on plurality, but go further in arguing that there is no role for prescriptive methodology at all. It follows that methodological pluralism has no meaning for them, since it is a prescriptive position.

Weintraub (1989) and McCloskey (1983) draw the distinction between prescriptive, ‘large M’, Methodology and descriptive, ‘small m’, methodology. Thus the science studies and rhetoric approaches, respectively, see a role for the second in providing descriptive accounts of different methodological approaches. This takes the form of recognizing methodological plurality as a feature of the subject matter, while not advocating it. Although not prescriptively methodological pluralist themselves, Weintraub and McCloskey nevertheless support a weak form of pluralism, which takes the form of an ethical argument. If there is a plurality of methodological approaches to economics (that is, economics is not defined by a particular methodological approach), then our discourse should be structured in such a way as to take that on board. Thus there are injunctions to practising economists to respect the legitimacy of expressions of methodological difference, be polite and so forth (McCloskey, 1996; Screpanti, 1997).

While much of the support for pluralism has focused on knowledge limitations, others (including Caldwell, 1997) have extended the argument to the nature of reality. Dating back to Keynes's (1921) *Treatise on Probability*, it has been argued that the organic nature of the social world means the absence

of law-like behaviour. It is not just that the capacity for knowledge is limited, but that the basis for laws is not there to be found. Since the social world is believed to be open, it requires an open system of knowledge, one which allows both for change and variety in methodological approach (Chick and Dow, 2005).

Keynes had argued further that reliable knowledge is derived from accessing a range of sources, given that there is so little scope for establishing demonstrably true knowledge. In the absence of certainty as to premises, classical logic is of limited use, so reason employs ‘human logic’. While rationalism is inadequate as a basis for action, human logic rather involves drawing on evidence and theoretical knowledge, supplemented by conventional opinion and intuition. The weight of argument is greater, the greater the number of sources of knowledge which support the hypothesis. This is to be distinguished from the probability that the hypothesis is correct, which probability may rise or fall with new evidence. This is an argument for a pluralist methodology, that is, a methodology that employs a range of methods. Inevitably this means methods beyond mathematical formalism, since one of the main attractions of that method is that all arguments can be expressed commensurately and can therefore be collapsed into one formal argument, or model. But the reasoning can also be applied to the meta-methodological level, to support methodological pluralism. If a policymaker perceives support for a particular policy from a range of methodological approaches, weight is added to the view that the policy is a good one.

If no single methodological approach can be demonstrated to be universally the best one, then it is to be expected that there will be a range of methodological approaches, even without the added, Keynesian, argument that this is desirable. This range is bound to have some limits, given that science operates within loose communities, requiring shared understandings of reality (ontologies) and of meanings of terms, and shared views as to how to proceed to build up (fallible) knowledge. These communities can be understood in Kuhnian terms as paradigms (Kuhn, 1962). This advocacy of variety of methodological approach limited according to the requirement for viable scientific communities is termed variously ‘critical pluralism’ (Caldwell, 1997), ‘principled relativism’ (Davis, 1999) or ‘structured pluralism’ (Dow, 2004). Each of the range of methodological approaches involves a different set of views as to how best to build up knowledge. Some will be pluralist, advocating the use of a range of methods, but this does not necessarily follow from the application of methodological pluralism. There are trade-offs involved in whatever methodology is chosen, and the benefits of a single commensurate method may be judged to outweigh the epistemological costs. But those approaches which adopt a pluralist methodology will be distinguished by the selection of methods employed. This in turn follows from ontology, the understanding of the nature of the subject matter. An understanding of economic relations in terms of class implies one set of methods, in terms of competitive markets another, and of individual entrepreneurial creativity another, for example.

The focus on the ontological level for the case for pluralism owes much to critical realism (Lawson, 1997; 2003). However, critical realists themselves take a particular position on pluralism. The case is made that the real world is open, requiring an open-system epistemology. The open-system nature of the real world means that knowledge is situated and contestable, and therefore there is likely to be a range of methodologies; to that extent critical realism is methodologically pluralist. But, since there is one external real world, there is only one ontology, and one open-system meta-methodology. Different methodologies simply reflect different ‘commitments’ with respect to that common ontology; beyond that no judgement is to be expressed on the content of these methodologies. This approach differs from

the structured pluralist position outlined above, whereby different methodological approaches follow ultimately from the plurality of ontological understandings (with respect to a common external real world). However, critical realists also support the idea that methodologies should be pluralist, adopting a range of methods suited to the chosen focus of analysis.

Finally, the argument that plurality of methodology is not only inevitable but desirable is supported by means of a biological metaphor (Hodgson, 1997). The view that the real social world is open involves the view that it undergoes structural change. Even if it were generally agreed that a particular methodological approach is best suited to the current economic structure, it is not unlikely that this approach would not be capable of addressing change in that structure. In the biological world, a dominant strain of a species may not be able to survive an environmental shock. Unless there is a range of alternative strains, including one better suited to the new environment, the species will die out. Since the nature of environmental shocks in general cannot be predicted, it is necessary for the survival of the species for there to be a range of alternative strains available at any time. The same argument can be made for different methodological approaches to economics.

See Also

- Keynes, John Maynard
- methodology of economics
- paradigms
- scientific realism and ontology
- theory appraisal

Bibliography

Caldwell, B.J. 1982. *Beyond Positivism*. London: Allen & Unwin.

Caldwell, B.J. 1989. Post-Keynesian methodology: an assessment. *Review of Political Economy* 1, 43–64.

Caldwell, B.J. 1997. Comment. In Salanti and Screpanti (1997).

Chick, V. and Dow, S.C. 2005. The meaning of open systems. *Journal of Economic Methodology* 12, 363–81.

Colander, D. 2000. New Millennium economics: how did it get this way, and what way is it? *Journal of Economic Perspectives* 14(1), 121–32.

Davis, J.B. 1999. Postmodernism and identity conditions for discourses. In *What Do Economists Know? New Economics of Knowledge*, ed. R.F. Garnett Jr., London: Routledge.

Dow, S.C. 2004. Structured pluralism. *Journal of Economic Methodology* 11, 275–90.

Goodwin, C. 2000. Comment: It's the homogeneity, stupid! *Journal of the History of Economic Thought* 22, 179–84.

Hodgson, G.M. 1997. Metaphor and pluralism in economics: mechanics and biology. In Salanto and Screpanti (1997).

Keynes, J.M. 1921. *A Treatise on Probability. Collected Writings*, vol. 3. London: Macmillan for the Royal Economic Society, 1973.

Kuhn, T.S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lawson, T. 1997. *Economics and Reality*. London: Routledge.

Lawson, T. 2003. *Reorienting Economics*. London: Routledge.

Mäki, U. 1997. The one world and the many theories. In Salanti and Screpanti (1997).

McCloskey, D.N. 1983. The rhetoric of economics. *Journal of Economic Literature* 21, 434–61.

McCloskey, D.N. 1996. *The Vices of Economists: The Virtues of the Bourgeoisie*. Amsterdam: Amsterdam University Press.

Salanti, A. and Screpanti, E., eds. 1997. *Pluralism in Economics*. Cheltenham: Edward Elgar.

Samuels, W.J. 1997. The case for methodological pluralism. In Salanti and Screpanti (1997).

Screpanti, E. 1997. Afterword: can methodological pluralism be a methodological canon? In Salanti and Screpanti (1997).

Weintraub, E.R. 1989. Methodology doesn't matter, but the history of thought might. *Scandinavian Journal of Economics* 91, 477–93.

How to cite this article

Dow, Sheila C. "pluralism in economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000389> doi:10.1057/9780230226203.1292

The New Palgrave Dictionary of Economics Online

Polanyi, Karl (1886 – 1964)

George Dalton

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

machine technology; managed capitalism; Polanyi, K.; pre-industrial economies

Article

Polanyi was born in Vienna in 1886 and grew up in Budapest, where he studied law and philosophy. He served as an officer in the First World War, after which he turned to economic journalism as foreign editor of Vienna's *Osterreichisiche Volkswirt* throughout the 1920s. He emigrated to England in 1933, where he worked in adult education, as a lecturer on world affairs for the Workers' Educational Association and for the Extramural Delegacies of the Universities of Oxford and London. He became intensely interested in the origins of the British Industrial Revolution and the enormity of its economic and social consequences, the subject of his book, *The Great Transformation* (1944), written while he was a resident scholar at Bennington College in Vermont between 1940 and 1943.

John Maurice Clark was sufficiently impressed by the book to invite Polanyi to Columbia University as a visiting professor of economic history in 1947 (when Polanyi was already 61). Polanyi remained at Columbia until his retirement in 1953. He continued doing research until his death in 1964 at his home in a suburb of Toronto.

The Great Transformation remains in print 45 years after its publication. It argues a triple thesis: (i) that in Great Britain and Western Europe, the coming of machine technology to mercantilistic national economies that contained governmentally regulated markets induced enormous growth in all input and output markets and the removal of governmental controls from some of them – what Polanyi calls an attempt to create a 'self-regulating' market system; (ii) that nationally integrated market systems in which labour, land, and money as well as produced goods were transacted as market commodities were historically unique (that such full-blooded capitalism dominated by market transactions for factor inputs as well as produced outputs was a new kind of economic system markedly different from any that preceded it); (iii) although machine technology producing within a market system was enormously productive – an 'unbound Prometheus' in the vivid phrase used by David Landes – its destructive consequences (that is, sporadic unemployment, the business cycle, large inequalities in income and

wealth) culminating in the Great Depression of the 1930s, forced governments from the early 19th century onwards to initiate market controls, monetary and fiscal policy to mitigate its destructive consequences, what we now call ‘managed’ and ‘welfare state capitalism’.

Polanyi's second big book, *Trade and Market in the Early Empires* (1957), which also remained in print after 30 years and which has also been translated widely, created a theory of pre-industrial, non-market economies of interest to economic archaeologists, economic anthropologists, and those economic historians who study early, pre-industrial economies throughout the world. Polanyi invented a conceptual vocabulary to specify the core attributes of such early and primitive economies much of which is employed today in standard fashion: ‘reciprocity’, ‘redistribution’, ‘special-purpose money’, ‘port of trade’, ‘politically administered trade’, ‘economy embedded in society’. This part of Polanyi's work is widely thought to illuminate the nature of early money, early foreign trade, and the economic organization of early kingdom-states.

Polanyi's continuing significance is reflected in the Karl Polanyi Institute of Political Economy, founded at Concordia University, Canada, in 1987, which in addition to scholarly activity that is motivated by Polanyi's thought, maintains an archive of his works.

Selected works

1944. *The Great Transformation*. New York: Rinehart.

1957. (With C.M. Arensberg and H.W. Pearson, eds.) *Trade and Market in the Early Empires*. Glencoe, IL: Free Press.

1966. *Dahomey and the Slave Trade*. Seattle: University of Washington Press.

1971. *Primitive, Archaic, and Modern Economies: Essays of Karl Polanyi*, ed. G. Dalton. Boston: Beacon Press.

1977. *The Livelihood of Man*. New York: Academic Press.

1996. *Uncollected Works of Karl Polanyi*. New York: St. Martins.

Bibliography

Adelman, I. and Morris, C.T. 1978. Patterns of market expansion in the nineteenth century: a quantitative study. In *Research in Economic Anthropology*, vol. 1, ed. G. Dalton. Greenwich, CT: JAI Press.

Dalton, G. and Köcke, J. 1983. The work of the Polanyi group: past, present, and future. In *Economic Anthropology*, ed. S. Ortiz. New York: University Press of America.

Polanyi-Levitt, K. (ed.) 1990. *The Life and Work of Karl Polanyi*. Montreal: Black Rose Books.

Stanfield, J. 1986. *The Economic Thought of Karl Polanyi: Lives and Livelihood*. London: Palgrave Macmillan.

Valensi, L. et al. 1981. Economic anthropology and history: the work of Karl Polanyi. In *Research in Economic Anthropology*, vol. 4, ed. G. Dalton. Greenwich, CT: JAI Press.

How to cite this article

Dalton, George. "Polanyi, Karl (1886–1964)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000106> doi:10.1057/9780230226203.1293

The New Palgrave Dictionary of Economics Online

polarization

Gordon Anderson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Polarization means the tendency of economic agents to form different groups and acquire identities that enhance differences from other groups. It is both cause and consequence of much economic behaviour. It has been employed, for example, in describing the diminution of the middle class in wage, income and wealth distributions, in studying growth and convergence issues, and in examining the plight of the poor. Although polarization is closely associated with trends in inequality, increased polarization can correspond to an increase, a reduction or no change in inequality.

Keywords

Akerlof, G.; alienation; clubs; endogeneous growth; Gini coefficient; identity; inequality; polarization

Article

Polarization, or the tendency of economic agents to collect into different groups and to feel increasingly different from members of other groups, is both cause and consequence of much economic behaviour. The essence of informal and formal club formation, polarization is born of an increasing sense of identity within the group members and an increasing sense of distance from members of other groups. The terminology is gaining increasing currency in economics. Akerlof (1997), Anderson (2004a; 2004b), Beach and Slotsve (1996), Beach, Chaykowski and Slotsve (1998), Bossert, D'Ambrosio and Peragrine (2004), Corak (2004), D'Ambrosio and Wolff (2001), Dinardo and Lemieux (1997), Foster and Wolfson (1992), Jenkins (1996), Jones (1997), Keefer and Knack (2002), Levy and Murnane (1992), Quah (1997) and Wolfson (1994; 1997) constitute an extensive but not exhaustive list of its use. In the list will be seen applications in describing the diminution of the middle class in wage, income and wealth distributions, in studying growth and convergence issues, and in examining the plight of the poor; it has also been used in the study of inter-generational income relationships and in discriminating between competing matching models of marriage partners. These literatures broadly interpret polarization as the disappearance of mass at the centre of an empirical distribution of a characteristic, or the increasing distance between, and intensity of, multiple points of modality of the distribution as it evolves through time. It is inherently a dynamic process involving the comparison of the anatomy of states at different points in time, essentially examining how the shape of the distribution of a characteristic (or a collection of characteristics) has evolved during the process. Thus, the objective is to detect trends in shapes of distributions over time that reflect the polarization or de-polarization (sometimes referred to as 'convergence') of that group of agents. Although polarization is closely associated with trends in inequality, it is distinguishable and quite different from changes in inequality in that increased polarization can induce an increase, a reduction or no change in inequality.

The concept need not be confined to the study of changes within a population's distribution of a particular characteristic, but can be used in assessing the relative movements of two or more distributions as they evolve (for example, polarization between ethnic groups, genders, and nations). In this context polarization takes the form of distributions becoming 'less alike' in a particular fashion; as such it involves comparison of complete distributions, not just their location or scale characteristic. While the identification of polarization within and between populations presents quite distinct empirical challenges, polarizations have many common features which can be exploited in understanding the nature of the phenomenon. Indeed, it is convenient to contemplate within-distribution polarization phenomena as the consequence of that population distribution being a mixture of sub-population distributions which are themselves polarizing (this notion is at the heart of the initial formalization of a Polarization index in Esteban and Ray, 1994, and Duclos, Esteban and Ray, 2004). Such a

construction will highlight why within distribution polarization is sometimes hard to detect in the absence of sub-population information.

Polarization: an axiomatic foundation

Indices of polarization were formulated in Esteban and Ray (1994) and Duclos, Esteban and Ray (2004) (see also Wang and Tsui, 2000) by positing a collection of axioms whose consequences should be reflected in a Polarization measure. The axioms are founded upon a so-called Identification–Alienation nexus wherein notions of polarization are fostered jointly by an agent's sense of increasing within-group identity and between-group distance or alienation. The four axioms may be loosely summarized as follows:
Axiom 1: : A mean preserving reduction in the spread of a distribution cannot increase polarization.
Axiom 2: : Mean preserving reductions in the spread of sub-distributions at the extremes of a density cannot reduce polarization.
Axiom 3: : Separation of two sub-densities towards the extremes of the distributions range must increase polarization.
Axiom 4: : Polarization measures should be population-size invariant.
The polarization index developed for discrete distributions as a consequence of these axioms (Esteban and Ray, 1994) may be written as:

$$P_\alpha = K\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|\pi_i^{1+\alpha}\pi_j \quad (1)$$

Here K is a normalizing constant, π_i is the sample weight of the i'th observation and where $\alpha \geq 0$ is a polarization sensitivity factor chosen by the investigator. It may readily be seen that $\alpha = 0$ yields a sample weighted Gini coefficient.
The continuous distribution analogue (Duclos, Esteban and Ray, 2004) may be written as:

$$P_\alpha(F) = \int_y f(y)^\alpha \int_x |y - x| dF(x) dF(y) \quad (2)$$

Again, α is the polarization sensitivity factor which in this case is confined to [0.25,1].

The anatomy of polarized states

For expositional simplicity in exploring the anatomy of polarized states, a population is represented by the equi-probable mixture of two sub-distributions (in reality more than two subgroups are possible), representing two subgroups or clubs which make up the population. The initial sub-distributions, which are identical except for having different means, are subjected to various transformations which are characterized in Figures 1a to 4b.
Figure 1
(a) Divergence in means between population polarization. (b) Divergence in means within population polarization

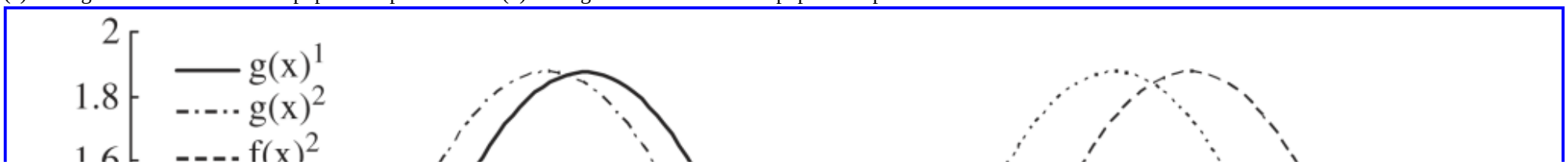

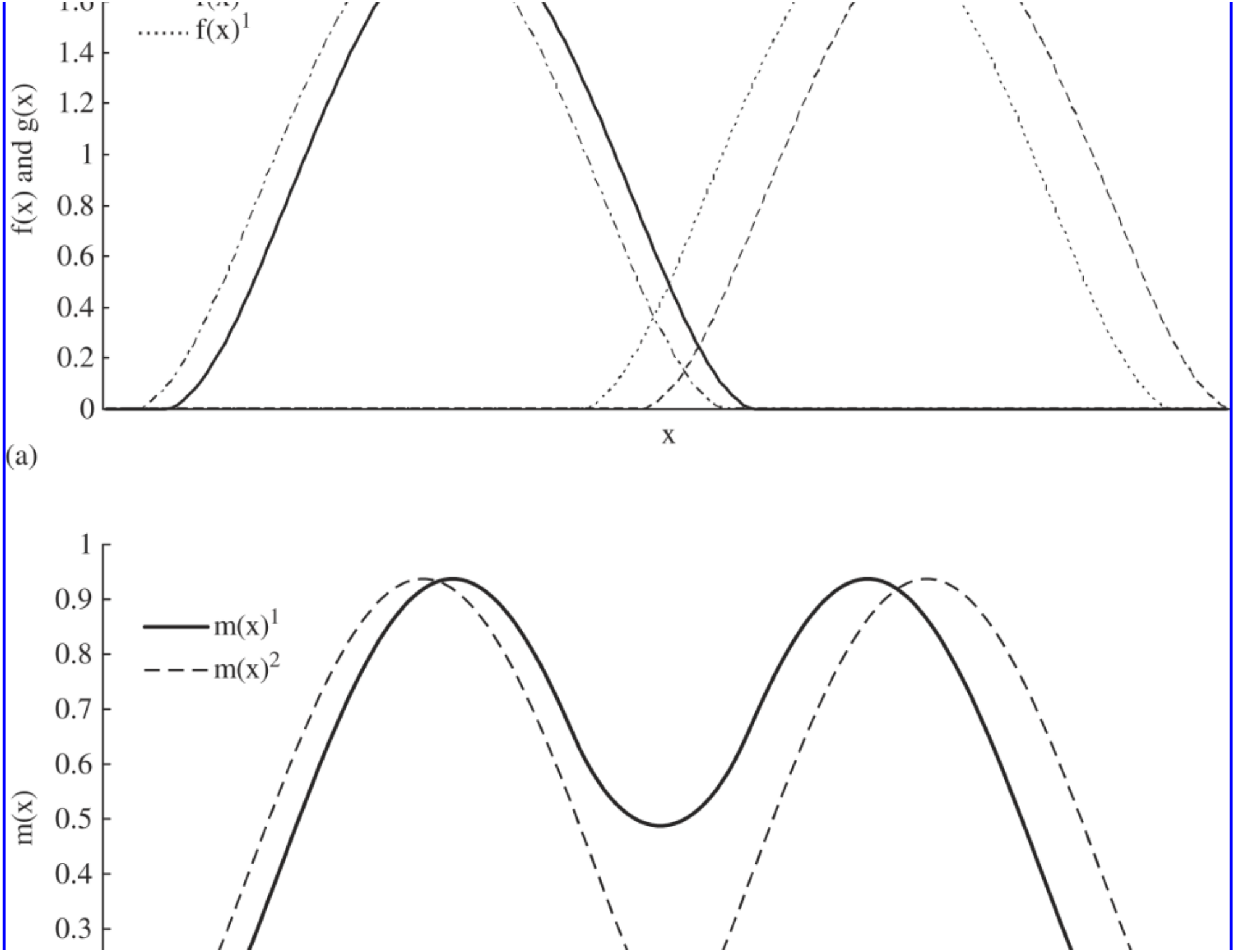
f(x)1
f(x) and g(x)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
x
(a)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
m(x)1
m(x)2
m(x)

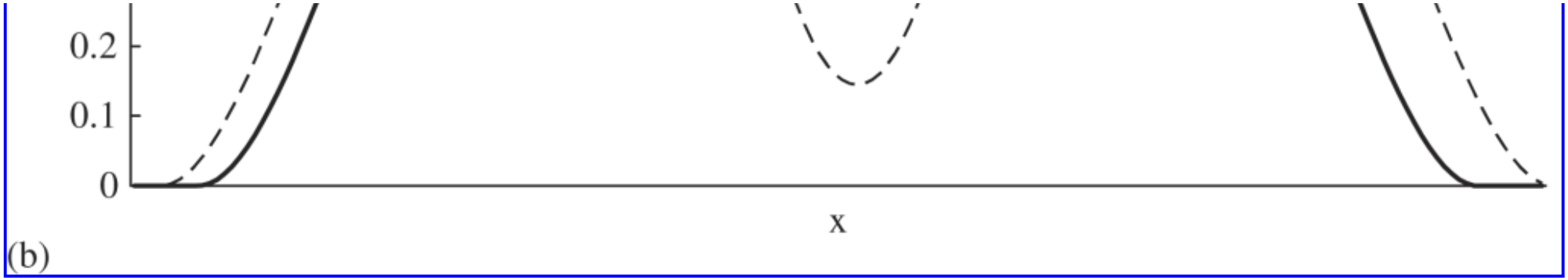

Following the spirit of Axiom 3, the simplest form of polarization occurs when the sub-populations exhibit divergence in their means. Here there is no increase in within-group identity but there is an increase in the distance between members of different groups (alienation). Figure 1a exemplifies this situation in terms of the sub-populations and its consequence for the mixture is illustrated in Figure 1b. As may be seen, the overlap of the two sub-distributions diminishes, and the centre of the mixture becomes hollowed out. Note that when the means are relatively close together there is no hollowing out but simply a flattening of the unimodal peak of the mixture distribution, implying that polarized or polarizing states need not be characterized by the existence or emergence of bimodality. (For example, for mixtures of equal-variance normal distributions, bimodality will not emerge under any mixing scheme until the difference in means exceeds $4.5^{0.5}$ standard deviations). Thus, bump-hunting techniques available in the statistics literature (Good and Gaskins, 1980; Hartigan and Hartigan, 1985) which seek out inflections in the probability density function will not necessarily be useful in the analysis of polarization.

Figure 2a and 2b illustrate another form of polarization when sub-population means remain constant but their variances diminish. This is much in the spirit of Axiom 2 and characterizes a situation of increased identification within the groups without an increased sense of alienation between them. Again, the overlap of the sub-populations diminishes and the centre of the mixture becomes hollowed out, but in this case the anatomical change is not unequivocal. Finally, when both locations and spreads remain constant but the lower distribution skews left and the upper distribution skews right, the overlap again diminishes and the centre of the mixture distribution again hollows out, as Figure 3a and 3b illustrate.

Figure 2

(a) Increased concentration between population polarization. (b) Increased concentration within population polarization

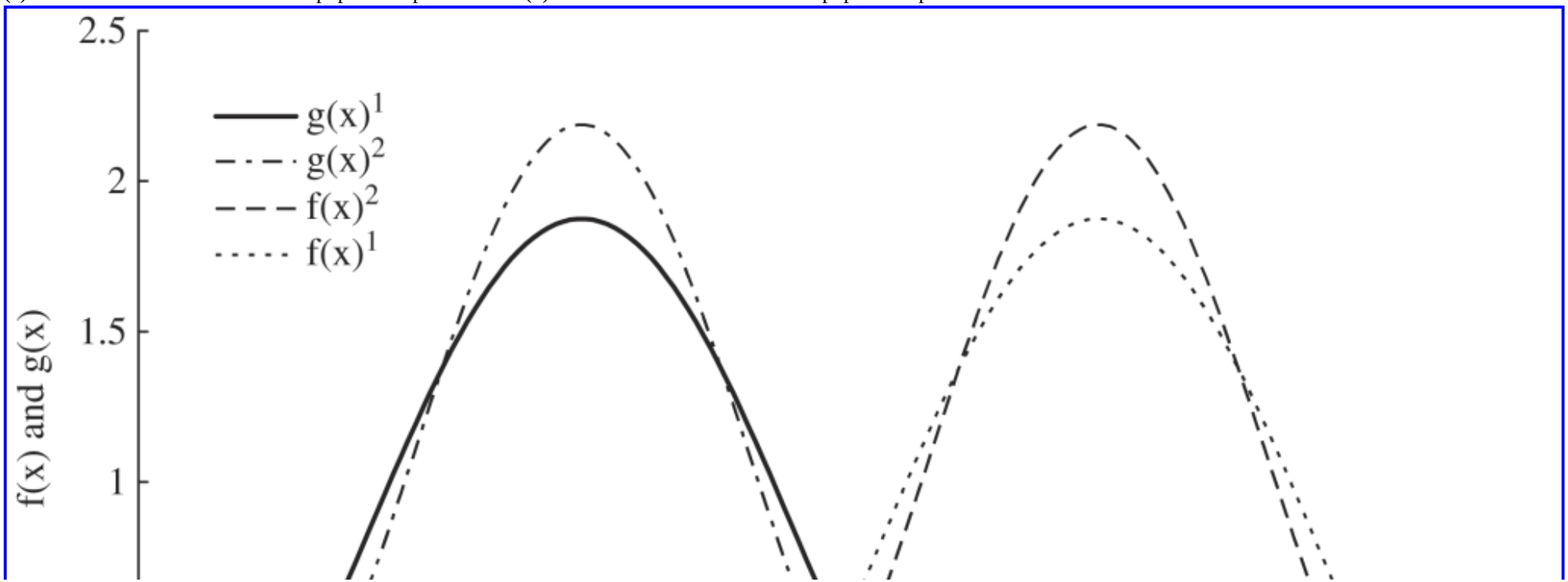

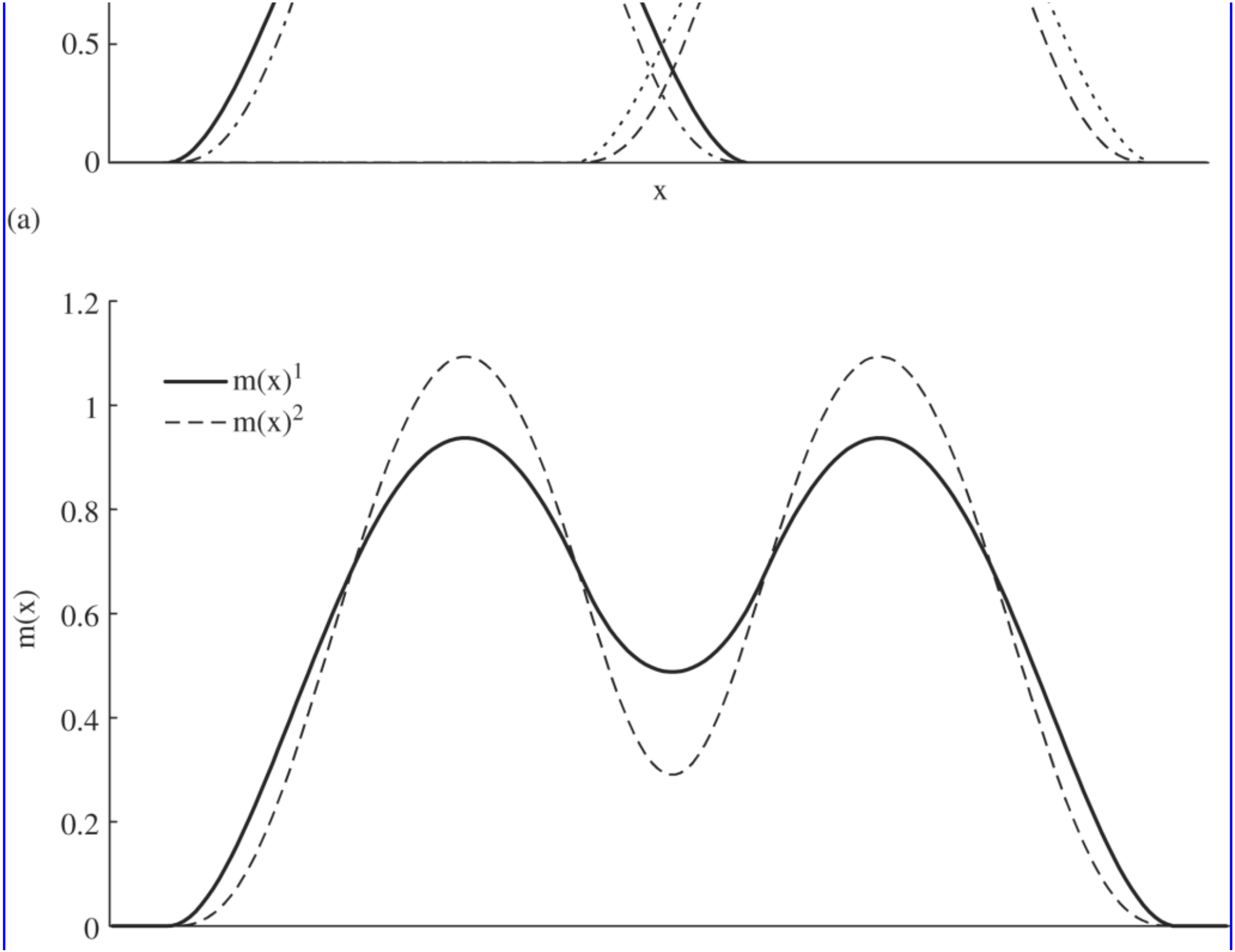
0.5
0
x
(a)
1.2
1
0.8
0.6
0.4
0.2
0
m(x)
m(x)1
m(x)2

Figure 3

(a) Opposite skewness between population polarization. (b) Opposite skewness within population polarization

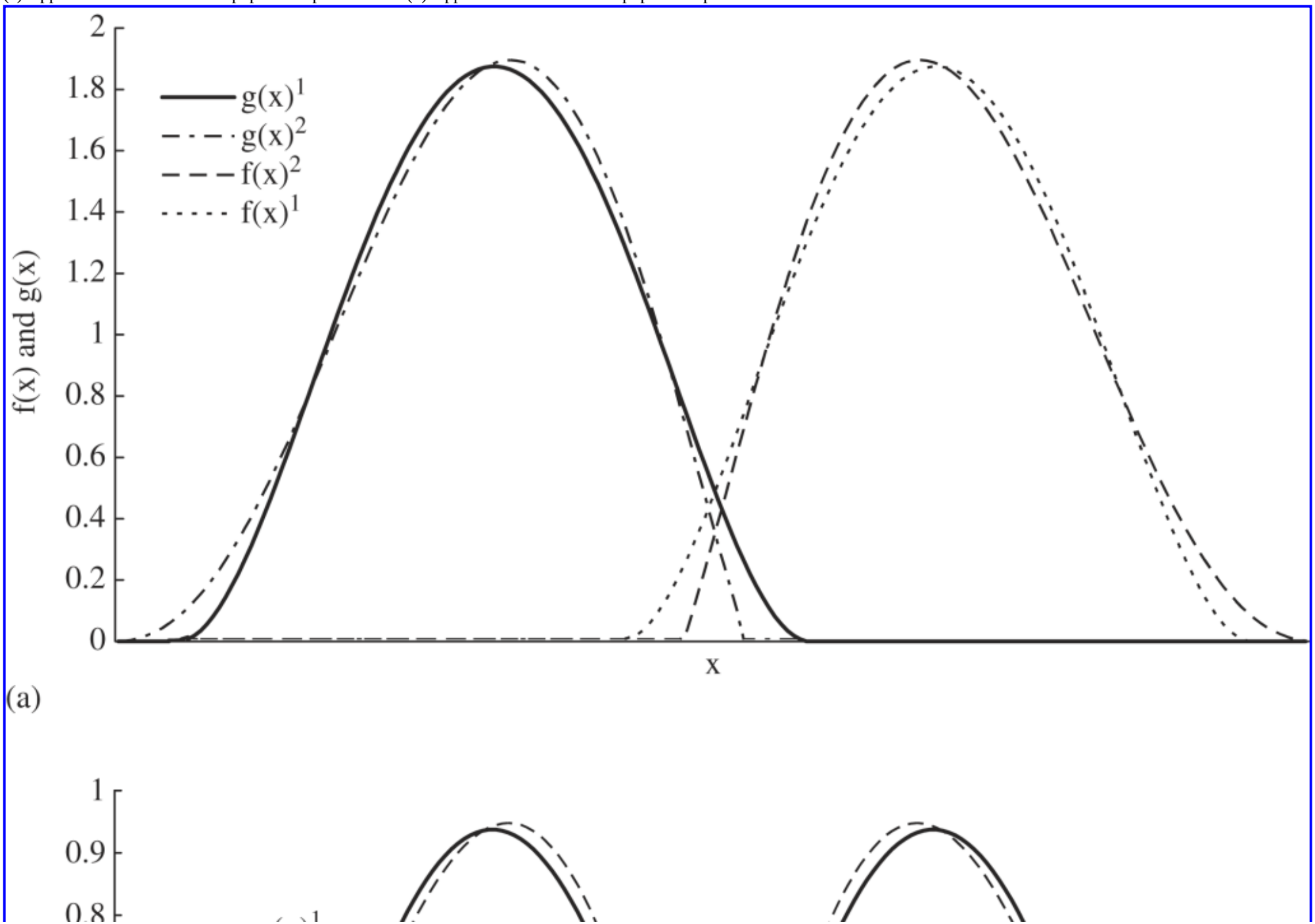

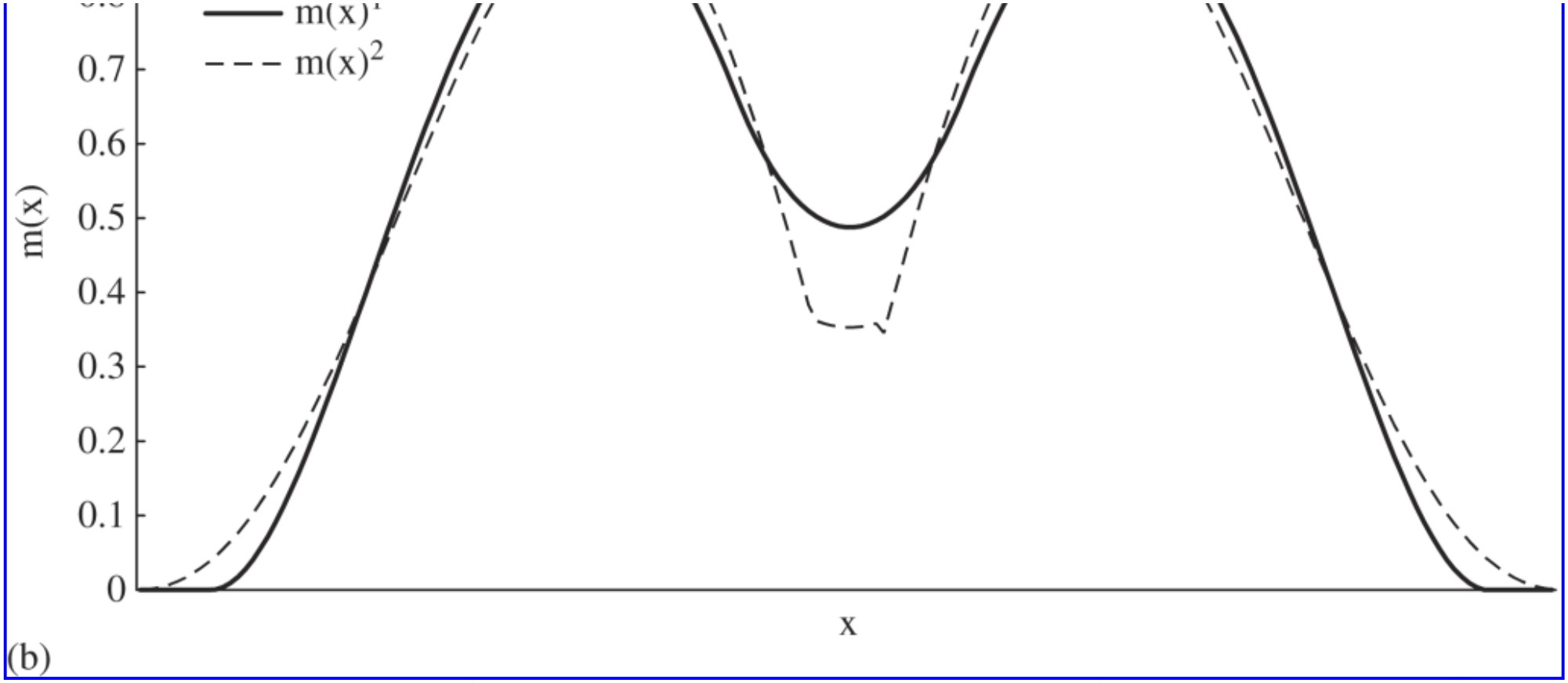

One thing these examples demonstrate is the potential for changes in the overlap measure to provide a general test or indicator of polarization. But this is not without qualification. Figure 4a and 4b return us to the increased identification case and demonstrate that, if the subgroup means happen to be close together, the extent of overlap can increase and the mass at the centre of the mixture distribution is enlarged with increased polarization. This highlights what is in effect a potential statistical identification problem associated with polarization when only the mixture of sub-distributions is observed. The tenuous link between polarization and inequality is also illustrated in this example. If we consider the mixture *f*(*x*) to be an equally weighted mixture of normal distributions $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, then the variance of *x* (for our purposes a measure of inequality) which will be $0.5(\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2)$ can be seen to either increase, decrease or remain unchanged with an increase in polarization interpreted as any combination in reductions of sub-population variances and increases in the difference of sub-population means.

Figure 4

(a) Increased concentration between population polarization close means. (b) Increased concentration within population polarization close means

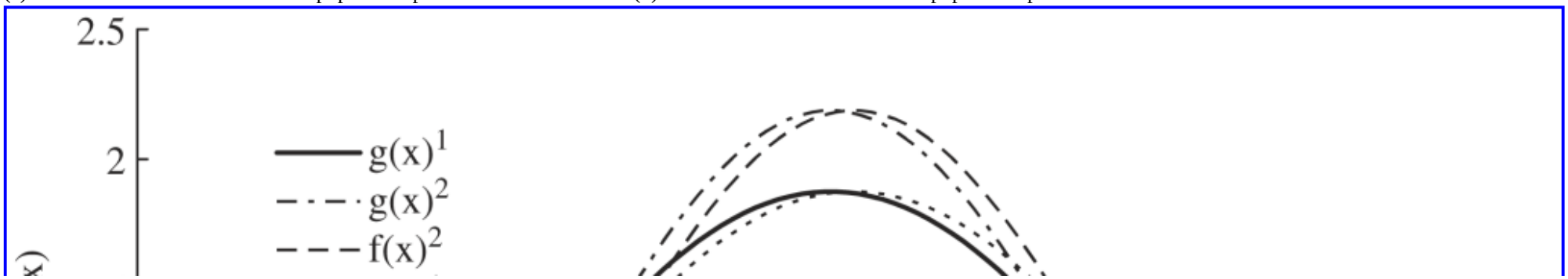

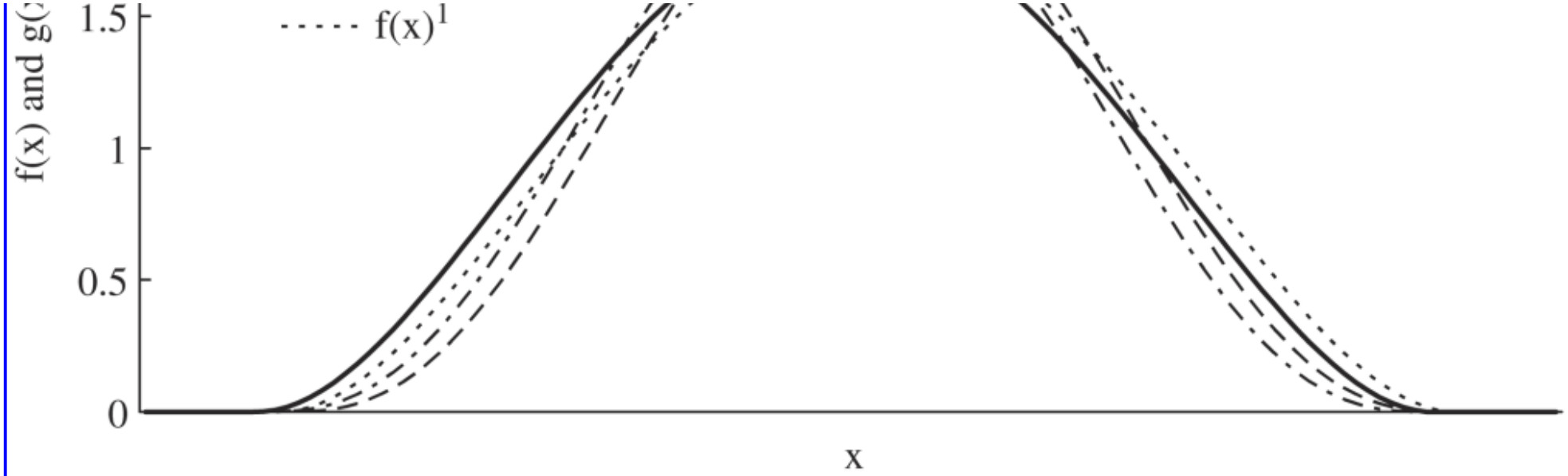

(a)

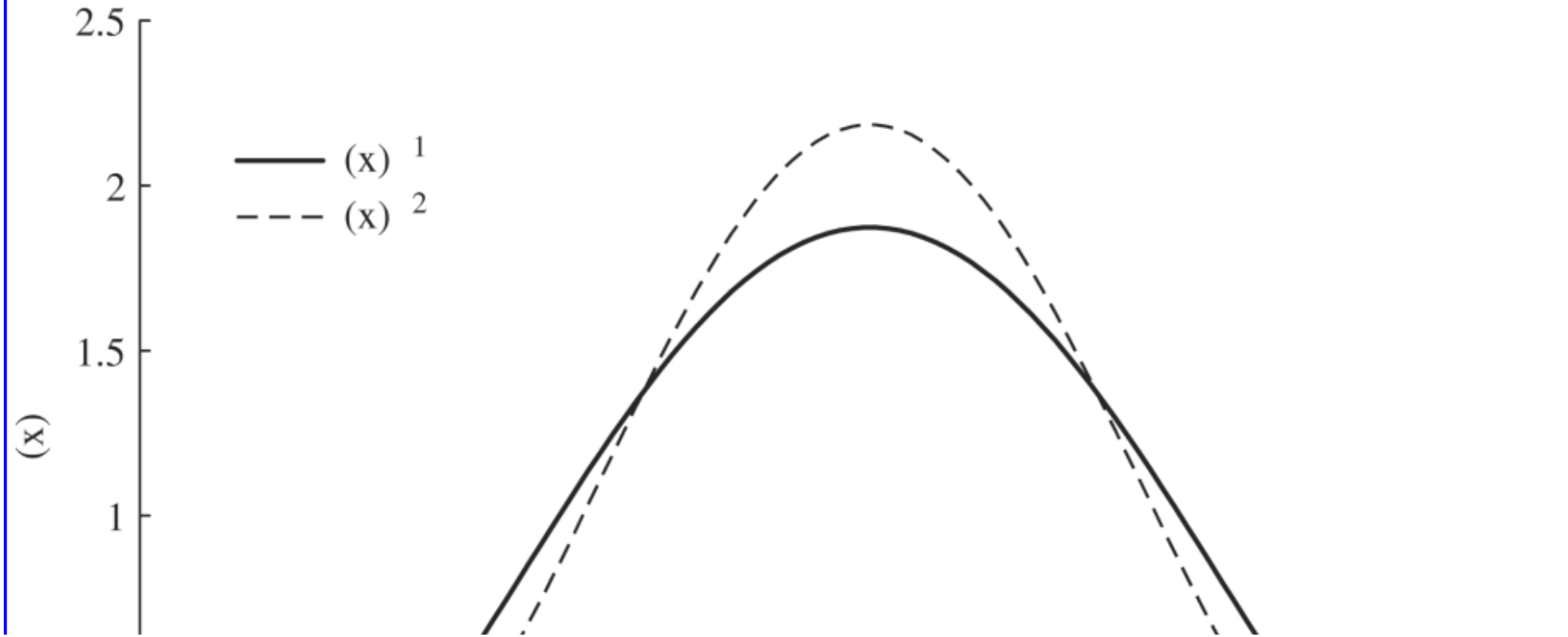

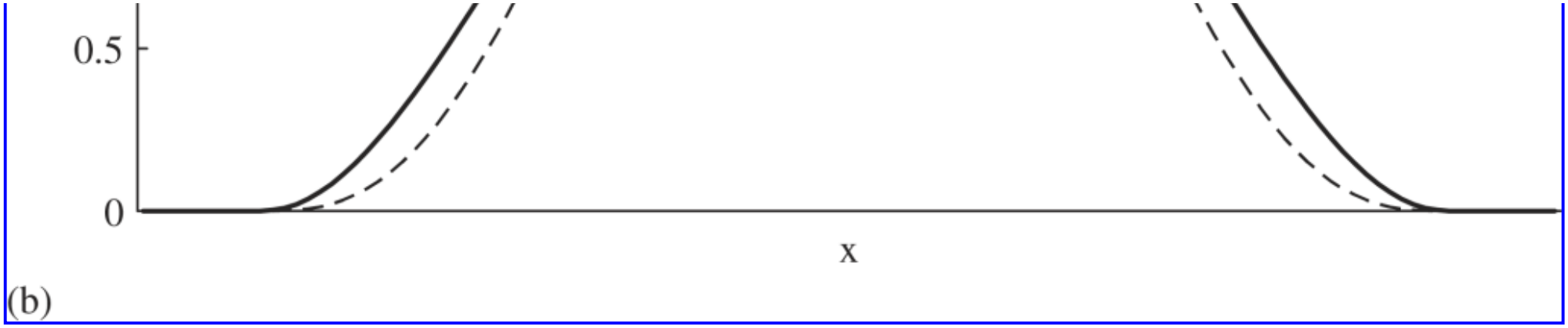

Alternative between-group polarization measures

Distributional overlap

The anatomy analysis suggests that one technique for assessing polarization between two groups is to evaluate how much they have in common. Such a measure corresponds to non-alienation, and its negative (or some negative function of it) corresponds to a degree of alienation. Anderson (2004a; 2004b) proposes an overlap measure as an index of convergence and a function of its negative as a measure of alienation. The extent to which two distributions $f(x)$ and $g(x)$ overlap is given by:

$$OV = \int_{-\infty}^{\infty} \min(f(x), g(x))\,dx \tag{3}$$

Clearly it is a number between 0 and 1 with 0 corresponding to no overlap and 1 to the perfect matching of the two distributions. It follows that $1-OV$ is a measure of the extent to which the distributions do not match or are alienated. When $f(x)$ and $f(y)$ are specified to the extent that all of their parameters can be estimated and the intersection points of $f(x)$ and $g(x)$ calculated, OV can be estimated parametrically (see Anderson and Ge, 2004). When $f(\cdot)$ and $g(\cdot)$ are unknown, given independent samples from $f(\cdot)$ (represented by x) and $g(\cdot)$ (represented by y) of sizes n_x and n_y respectively, its empirical counterpart may be implemented by choosing $K+1$ mutually exclusive and exhaustive partitions of the range of x whose upper bound is defined by x^k, $k = 1, \ldots, K+1$ and calculating

$$OV^e = \sum_{i=1}^{k+1} \min\left(\frac{\sum_{j=1}^{n_x} I(x_j, x^i)}{n_x}, \frac{\sum_{j=1}^{n_y} I(y_j, x^i)}{n_y}\right) \tag{4}$$

where $I(z,w^i)$ is an indicator function equal to 1 when z is in the interval (w^{i-1}, w^i) and 0 otherwise. The statistical properties of such an estimator are discussed in Anderson (2005). The nonparametric measure is prone to two sources of bias: the first, due to the intersection points of the underlying distributions not coinciding with partition points, is actually not that large provided k is not small and the partition points are chosen judiciously; the second, due to the estimator being implicitly a conditional estimator, can be large when the

measure is either close to 0 or 1. However, these biases do not appear to impede its use in calibrating changes in overlap. The main problem with this particular instrument arises when distributions do not actually overlap; for that purpose the following measures may prove useful.

Gini-based between-group polarization measures

Starting with the classic Gini inequality coefficient which, with x_i being the income of the i'th agent for agents i=1,…,n and where for convenience and without loss of generality incomes are arranged in ascending rank order, may be written as:

$$Gini = \frac{1}{2n^2\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j| \tag{5}$$

where μ is the mean of the x's. Suppose the rich and poor groups are defined by a poverty cut-off somewhere between x_p and x_{p+1} where $1<p<n$ (what Yitzhaki 1994 refers to as perfect stratification of groups, that is, no overlapping): then Gini may be thought of as the sum of the average mean normalized differences between agents in the poor group, between agents in the rich group, and between poor- and rich-group agents. In measuring alienation it is only the last group of comparisons that are relevant, that is, the average normalized difference between the rich-group and poor-group agents. In this case the new 'PGini' index could be written as:

$$PGini = \frac{1}{p(n-p)\mu}\sum_{j=p+1}^{n}\sum_{i=1}^{p}(x_j - x_i) \tag{6}$$

Clearly this is still a number greater than 0 (but it is no longer guaranteed to be less than 1), which reflects the mean normalized average distance between the poor group and the rich group, and as such it is easy to show that it is the overall mean normalized difference between the subgroup means. Indeed, the formulae can be generalized to general group differences where there is not perfect segmentation, that is, where the subgroups overlap.

Observe that the same index would be arrived at if one were to work with x_i–z and x_j–z, the corresponding distances from the poverty line z which facilitates a link to the well-known FGT family of poverty and welfare indices introduced by Foster, Greer and Thorbecke (1984), as follows. The formal representation of this family is given by:

$$POV_\theta(x, z) = \int_0^z \left(\frac{z-x}{z}\right)^\theta dF(x) \tag{7}$$

where $F(x)$ is the cumulative density function (with p.d.f. $f(x)$) describing the population of incomes, z is the maximum of the poor, and θ (≥ 0) is a parameter defining the nature of the poverty index and corresponds to a measure of poverty aversion. As a consequence POV_0 corresponds to the proportion of people in the poverty group, POV_1 is a normalized measure of the intensity of relative deprivation and so on. POV_i/POV_0 may be construed as the expected value of a weighted function of the normalized income deficiency where the weights are the (i–1)'th power of the normalized income deficiency itself. Thus increasing i increases the weights attached to those furthest from the poverty line. Interestingly, as i becomes very large the index becomes a Rawlsian measure in focusing almost entirely on the poorest agent. All of these measures obey the focus axiom, which holds that poverty

measures should depend only upon the incomes of the poor. As such they are not in any sense related to the status of the rich.
Along similar lines *RIC*(*x*,*z*), an index of weighted relative distances of incomes above the poverty line, may be contemplated whose theoretical representation is of the form:

$$RIC_{\theta}(x, z) = \int_{z}^{x\max} \left(\frac{x - z}{z}\right)^{\theta} dF(x) \quad (8)$$

In this case *x* max corresponds to the maximum possible income. Here RIC_0 corresponds to the proportion of the population above the poverty line, RIC_1 is a normalized measure of relative well-being of the non-poor, RIC_2 is a measure of the intensity of the relative well-being of those above the poverty line, and so on. In this case as α becomes very large the index becomes almost entirely focused on the richest person, RIC_1/RIC_0 corresponds to the expected normalized income surplus over the maximum poverty income, and so on. For all $\theta > 0$ all of these indices are essentially measuring relative weighted distances from the poverty line, and it is in this sense that they are considered relative measures. However, both *RIC* and *POV* are completely uninformed with respect to the distribution of incomes in the other group, which accords with the focus axiom mentioned above. For the purposes of reflecting the notion of alienation between the poor and non-poor groups, this axiom needs to be violated. Indeed the population analogue of PGini can be shown to be a specific member ($\theta = 1$) of a general class of polarization measures defined by

$$POL(z, \theta) = \left(\frac{z}{\mu}\right)^{\theta} \left(\frac{RIC_{\theta}}{RIC_0} - \frac{POV_{\theta}}{POV_0}\right) \quad (9)$$

where $\theta \geq 1$.

Tests for polarization

Given the distribution of the above indices, tests for increases or decreases in polarization in terms of movements in the indices can be readily established. But, although indices provide complete orderings, much like the Gini coefficient with which they are associated, they can be ambiguous. Direct tests of the anatomy of polarization based upon degrees of separation or stochastic dominance between density functions can provide an unambiguous (though not complete) orderings of the states of polarization. These tests can be developed by employing combinations of stochastic-dominance conditions, tests for which have been proposed by Anderson (1996; 2004a), Davidson and Duclos (2000), Barrett and Donald (2003), Linton, Maasoumi and Whang (2002), and McFadden (1989). The conditions can be used in combination to compare the right separation of the upper distribution with the left separation of the lower distribution and thus establish a statistical criterion for polarization both within and between distributions. Anderson (2004b) provides a taxonomy of such tests.

An alternative approach: the growth and convergence literature

The endogenous growth literature has for a long while been concerned about issues of polarization specifically in the form of convergence or de-polarization. Early attempts at identifying the phenomenon via panel data regression techniques (see, for example, Barro, 1998) ran into difficulties in interpretation (Bernard and Durlauf, 1996). The phenomenon has been studied via the use of probability transition matrices implicit in the Markov chain methods employed by Quah (1997) (see also Durlauf and Quah, 1999). (These techniques have also been applied to the problem of intergenerational Income relationships; see Corak, 2004, and city sizes, Dobkins and Ioannides, 2000, Anderson and Ge, 2004). If we let *f*(*y*) be the distribution of income *y* in some future period, and *f*(*x*) the distribution of income *x* in the present period, the issue to be addressed is the relationship between the two distributions, that is, the extent to which, and manner in which, *f*(*y*) and *f*(*x*) are related. If for the moment we think of *x* and *y* as having the joint distribution *f*(*y*,*x*) so that *f*(*y*) and *f*(*x*)

are the respective marginal distributions, at one extreme there is a sense of no relationship – that is to say, when x and y are independent $f(y,x)=f(y)f(x)$ – at the other there is the completely deterministic environment whereby $y=a+bx$ and the joint distribution is degenerate. If y and x are partitioned into k mutually exclusive and exhaustive regions where $p(y)$ and $p(x)$ are respectively the vectors of marginal probabilities of falling into those regions, interest centres on the elements of the square matrix T defined by $p(y)=T(y,x)p(x)$, the matrix in the square brackets in the following equation. T is of course the matrix of conditional probabilities formed by the product of the two square matrices in the equation, so that:

$$\begin{Bmatrix} p_1(y) \\ p_2(y) \\ . \\ p_k(y) \end{Bmatrix} = \left[\begin{Bmatrix} p_{11}(y,x) & p_{12}(y,x) & . & p_{1k}(y,x) \\ p_{21}(y,x) & p_{22}(y,x) & . & p_{2k}(y,x) \\ . & . & . & . \\ p_{k1}(y,x) & p_{k2}(y,x) & . & p_{kk}(y,x) \end{Bmatrix} \begin{Bmatrix} p_1(x) & 0 & & 0 \\ 0 & p_2(x) & & 0 \\ . & . & . & . \\ 0 & 0 & . & p_k(x) \end{Bmatrix}^{-1} \right] \begin{Bmatrix} p_1(x) \\ p_2(x) \\ . \\ p_k(x) \end{Bmatrix}$$

(10)

which is a matrix of conditional probabilities – that is, $T = \| p_{ij}(y,x) \,/\, p_j(x) \| i,\ j = 1, \ldots, k -$ familiar in the convergence literature and made popular by Quah (1997). As such its properties are well known, as are the techniques for its estimation. The i'th column of T is a conditional probability density function describing the distribution or reallocation over states of the i'th element of $p(x)$ the initial income distribution, to the elements of $p(y)$, the resultant income distribution, after one period. If this process is thought to be time invariant, then letting p^s be the vector of $p_i(x)$'s s periods ahead, $p^s = T^s p$ corresponds to the s period ahead distribution, and the solution to $p^\infty = T p^\infty$ (if it exists) is what is known as the long-run ergodic mass function. By interpreting these ergodic distributions as 'characterizations of tendencies', one can infer a tendency towards polarization if they display multiple peaks. Polarization can be examined in two ways in this context. When the diagonal elements of T are large relative to the off-diagonal elements, the system is said to exhibit persistence; if the diagonal is particularly large in the high and low ends it indicates a tendency towards polarization. Alternatively, one could compare p^∞, the long-run distribution, with $p(x)$, the initial distribution. If the former exhibits multiple peaks whereas the latter does not, a polarizing tendency may be inferred. One difficulty here is that no theory of inference has as yet been outlined for examining the 'multiple peakedness' of these ergodic functions.

Multivariate polarization

When agents are characterized in terms of more than one characteristic, their polarization or otherwise will be reflected in more than one dimension. The empirical problem is then altogether much more challenging; the extension of the analysis to a multivariate measures can be somewhat problematic. Multivariate Gini coefficients have been developed (see Anderson, 2004c, and Koshevoy and Mosler, 1997) but adapting them to the current context is complex; it requires defining a poverty cut-off for each characteristic (or some poverty boundary in multidimensional space), but even then extending the analogy to multivariate measures of FGT indices is not possible. One simple approach is to take a weighted geometric mean of the various Gini coefficients in each dimension; but then the weights have to be determined in an inevitably arbitrary fashion.
On the other hand, extension of the overlap measure OV to a multivariate overlap measure MOV is very straightforward, since MOV is of the form:

$$MOV = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \min(f(x, y \ldots, z), g(x, y \ldots, z))\, dx dy \ldots dz$$

(11)

In the corresponding empirical measure MOV^e, given suitable partitions in each dimension, the indicator function would simply be modified to a multivariate version accordingly (Anderson, 2005, provides an example) and $1-MOV$ would provide an appropriate polarization measure.

See Also

- alienation
- convergence
- Gini ratio
- inequality between nations
- wage inequality, changes in

Bibliography

Akerlof, G. 1997. Social distance and social decision. *Econometrica* 65, 1005–27.

Anderson, G. 1996. Nonparametric tests for stochastic dominance in income distributions. *Econometrica* 64, 1183–93.

Anderson, G. 2004a. Making inferences about the polarization, welfare and poverty of nations: a study of 101 countries 1970–1995. *Journal of Applied Econometrics* 19, 537–50.

Anderson, G. 2004b. Toward an empirical analysis of polarization. *Journal of Econometrics* 122, 1–26.

Anderson, G. 2004c. The empirical assessment of multidimensional inequality: sample weighted multivariate generalizations of the Gini Coefficient and Kolmogorov-Smirnov two sample tests for stochastic dominance. Mimeo. Department of Economics, University of Toronto.

Anderson, G. 2005. Indices and tests for alienation based upon Gini type and distributional overlap measures. Mimeo. Department of Economics, University of Toronto.

Anderson, G. and Ge, Y. 2004. A new approach to convergence, city types and the ‘true’ convergence of post-reform Chinese urban income distributions. Mimeo. Department of Economics, University of Toronto.

Barrett, G. and Donald, S. 2003. Consistent tests for stochastic dominance. *Econometrica* 71, 71–104.

Barro, R. 1998. *Determinants of Economic Growth: A Cross Country Empirical Study*. Cambridge, MA: MIT Press.

Beach, C., Chaykowski, R. and Slotsve, G. 1998. Inequality and polarization of male earnings in the US 1968–1992. *North American Journal of Economics and Finance* 8, 135–51.

Beach, C. and Slotsve, G. 1996. *Are We Becoming Two Societies? Income, Polarization and the Middle Class in Canada*. Toronto: C.D. Howe Institute.

Bernard, A. and Durlauf, S. 1996. Interpreting tests of the convergence hypothesis. *Journal of Econometrics* 71, 161–74.

Bossert, W., D’Ambrosio, C. and Peragrine, V. 2004. Deprivation and social exclusion. Paper presented at the International Association for Research in Income and Wealth, Cork, Ireland, August.

Corak, M. 2004. *Generational Income Mobility in North America and Europe*. Cambridge: Cambridge University Press.

D’Ambrosio, C. and Wolff, E. 2001. *Is Wealth Becoming More Polarized in the United States?* Working Paper 330. New York: Levy Economics Institute, Bard College.

Davidson, R. and Duclos, J.-Y. 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* 68, 1435–64.

Dinardo, J. and Lemieux, T. 1997. Diverging male wage inequality in the United States and Canada 1981–1988: Do institutions explain the difference? *Industrial and Labour Relations Review* 50, 629–51.

Dobkins, L. and Ioannides, Y. 2000. Dynamic evolution of the U.S. city size distribution. In *Economics of Cities*, ed. J.M. Huriot and J.F. Thisse. Cambridge: Cambridge University Press.

Duclos, J.-Y., Esteban, J. and Ray, D. 2004. Polarization: concepts, measurement, estimation. *Econometrica* 72, 1737–73.

Durlauf, S. and Quah, D.T. 1999. The new empirics of economic growth. In *Handbook of Macroeconomics*, vol. 1A, ed. J.B. Taylor and M. Woodford. Amsterdam: North Holland.

Esteban, J.-M. and Ray, D. 1994. On the measurement of polarization. *Econometrica* 62, 819–51.

Foster, J., Greer, J. and Thorbecke, E. 1984. A class of decomposable poverty measures. *Econometrica* 52, 761–66.

Foster, J. and Wolfson, M. 1992. Polarization and the decline of the middle class: Canada and the US Mimeo. Nashville: Vanderbilt University.

Good, I. and Gaskins, R. 1980. Density estimation and bump-hunting by the penalised likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association* 75, 42–56.

Hartigan, J. and Hartigan, P. 1985. The dip test of unimodality. *Annals of Statistics* 13, 70–84.

Jenkins, S. 1996. Recent trends in the U.K. income distribution: what happened and why? *Oxford Review of Economic Policy* 12, 29–46.

Jones, C. 1997. On the evolution of the world income distribution. *Journal of Economic Perspectives* 11(3), 19–36.

Keefer, P. and Knack, S. 2002. Polarization, politics and property rights: links between inequality and growth. *Public Choice* 111, 127–54.

Koshevoy, G. and Mosler, K. 1997. Multivariate Gini indices. *Journal of Multivariate Analysis* 60, 252–76.

Levy, F. and Murnane, R. 1992. US earnings inequality: a review of recent trends and proposed explanations. *Journal of Economic Literature* 30, 1333–81.

Linton, O., Maasoumi, E. and Whang, Y.-J. 2002. Consistent testing for stochastic dominance: a subsampling approach. Discussion Paper No. EM/02/433. STICERD, London School of Economics.

McFadden, D. 1989. *Testing for Stochastic Dominance in Studies in the Economics of Uncertainty* (in honor of Josef Hadar), ed. T. Fomby and T. Seo. New York: Springer-Verlag.

Quah, D. 1997. Empirics for growth and distribution: stratification, polarization, and convergence clubs. *Journal of Economic Growth* 2, 27–59.

Wang, Y.Q. and Tsui, K.Y. 2000. Polarization orderings and new classes of polarization indices. *Journal of Public Economic Theory* 2, 349–63.

Wolfson, M. 1994. When inequalities diverge. *American Economic Review Papers and Proceedings* 84, 353–8.

Wolfson, M. 1997. Divergent inequalities: theory and empirical results. *Review of Income and Wealth* 43, 401–21.

Yitzhaki, S. 1994. Economic distance and overlapping of distributions. *Journal of Econometrics* 61, 147–59.

How to cite this article

Anderson, Gordon. "polarization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000317> doi:10.1057/9780230226203.1294

The New Palgrave Dictionary of Economics Online

policy reform, political economy of

Sharun W. Mukand

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Policymakers face political constraints that make enacting reform difficult. Since the late 1980s economists have developed a framework to analyse the deeper political underpinnings of policy inefficiency. This article develops a framework for delineating the key findings of this literature. It then briefly sketches out the role of institutions in facilitating policy reform.

Keywords

adjustment costs; commitment; compensation; democracy; distributional conflict; government failure; institutional design; new political economy; policy persistence; policy reform; public choice; rational choice political economy; rent seeking; reputation; stabilization policy; time inconsistency

Article

Good policymaking is difficult. Inefficiencies abound in all areas of policymaking, due to constraints faced by the policymaker, be they informational, administrative or political. The subject of the political economy of policy reform is concerned with the *political* factors that make it difficult to reform policies and institutions. This field has focused on examining the impact of different political institutions on exacerbating or alleviating the ability to carry out reform, the consequences of these political constraints on policy outcomes as well as normative issues of institutional design that impinge on a policymakers' choice of policy.

The systematic exploration of the political economy underpinnings of policy reform began with two developments. First, there was an attempt to understand why governments in many developing countries failed to reform policies and institutions, despite low growth and stagnation and overall inefficiency (Rodrik, 1996). Second, there were new developments in rational choice political economy. In particular, there was growing recognition of the power of the public choice critique of traditional policy analysis due to Buchanan and Tullock (1962). This literature emphasized that policymakers' preferences

may be quite distinct from those of social planners and result in inefficient policy choices – that is, government failure. Therefore, policymakers' choices may be driven by a desire to retain office or may give very different weight to the preferences of some individuals (or groups) than might those of social planners. However, the insights from the public choice tradition were not explicitly grounded in a rational choice framework. This is where the literature on time-inconsistency spawned by the classic contribution of Kydland and Prescott (1977) played a crucial role. This literature demonstrated the importance of clearly specifying the policymaker's objectives and the constraints within a framework of optimization. However, the study of the political underpinnings of policy reform took off when insights from this 'new' political economy were used to deepen our understanding of economic crises, poverty traps and institutional inefficiencies in developing and transition economies.

Policy reform is difficult to achieve. Indeed, understanding the persistence of inefficient policy choices has been one of the central themes of much of the literature on policy reform. We can delineate the mechanisms proffered by the literature by focusing on two kinds of conflicts that make all policy reform more or less difficult. The first is the distributional conflict between different groups of citizens and individuals, be it due to differences in income, occupation, ethnicity or even religion. Given that much of policymaking is an attempt to balance these competing interests, the ability of a society to resolve this conflict is likely to affect its ability to reform. Second, the ability of a society to reform a failed policy may be due to conflict of interests between the politician–policymaker and those of the public. Institutions lie at the heart both these conflicts. Therefore, much recent work on the political economy of reform has focused on the interaction between political institutions, policy choices and inefficiency.

Distributional conflicts and policy inefficiency

During the 1980s Latin America witnessed a number of macroeconomic crises due to delays in the enactment of any stabilization policy (Rodrik, 1996). The puzzle was, why were these stabilizations delayed? In a near classic in this field, Alesina and Drazen (1991) addressed this issue. They showed that policy reform can be delayed due to a 'war of attrition' between two groups. Given uncertainty about the other group's willingness to bear a disproportionate burden of the adjustment costs, each group delays adjustment in the hope that the other caves in first. The economic crises worsens before one of the sides caves in and reform takes place.

At its broadest level inefficiency in policymaking in democracies arises from a commitment problem. Governments, which are vulnerable to losing power, are unable to commit to future policy outcomes. This failure in commitment can result in inefficient policy choices due to a variety of reasons (see Besley and Coate, 1998). In particular, most policy reforms have distributional consequences, resulting in winners and losers. However, there is a time-inconsistency problem with promises of future compensation (see Acemoglu, 2003; Robinson, 1998, for an elaboration). Therefore, what is key is the inability of a government to credibly commit to compensate losers from economic reform. Not surprisingly, if losers are in a majority (or politically influential) they will be not only opposed but will be able to prevent the implementation of such a policy reform – even if it is efficient. Now, if a government through some form of taxes and transfers could credibly commit to compensate losers for their losses, then policy reform would be much easier to achieve. In part, the difficulty in making credible promises to compensate losers is that the gains and losses from policy choices are spread out

over time, while the winners may not have enough resources to compensate the losers up front for their subsequent losses (see Dixit and Londegran, 1996).

However, inability to compensate losers is not in itself sufficient for there to be a failure to adopt policy reform. If individuals are risk neutral, then they may well be willing to adopt a policy which has winners and losers. For example, consider an economy where 100 risk-neutral voters face the prospect of voting for or against a policy reform. If enacted this policy reform will result in 51 winners, each of whom gains five dollars, and 49 losers, each of whom stands to lose one dollar. We may suspect that since (in expected terms) all individuals stand to gain from the adoption of this policy reform, it will always be adopted, whether or not winners compensate the losers. In an important contribution, Fernandez and Rodrik (1991) suggest that this is not the case. They argued that, even in a world with risk-neutral agents, individual specific uncertainty about the identity of winners and losers from a reform may prove to be crucial. In particular, to continue with our example, consider the case where the identity of 49 of the 51 winners is common knowledge. In this case there is individual-specific uncertainty amongst the remaining majority about their identity as a winner or a loser. Observe that this uncertain majority has a negative expected payoff from the reform and will vote it down. Therefore under individual-specific uncertainty a majority may vote against a policy, despite the fact that a majority stands to benefit from it. In an extension, Jain and Mukand (2003) show that policy reform may fail to get enacted, despite the existence of tax-transfer compensation instruments.

Therefore, social conflict across groups coupled with a failure of credible and efficient means of conflict resolution result in a persistence of inefficient policies. In contrast to the above, the other class of models in this literature has focused on the agency conflict between the politician and the citizen.

Political losers, agency and policy inefficiency

Once in power, politicians earn both economic and non-economic rents. As such there may be a failure to enact any policy reform if it adversely affects the rents earned by the current incumbent politician. A number of mechanisms have been studied.

The prospect of earning rents from a status quo policy can make the adoption of policy reform by the politician much more difficult. Coate and Morris (1999) show that the mere introduction of a policy encourages the affected parties to make investments that increase their willingness to pay for retaining these policies in the future. If, in the future, the efficient policy is no longer the status quo policy then there may be a problem. Any government attempting to reform the status quo policy is likely to be vulnerable to lobbying by the now entrenched firms.

Indeed, in the presence of uncertainty about the policy choice, this inefficiency is exacerbated. For instance, many commentators wondered why US President Johnson persisted with military escalation in Vietnam though it was apparent to him (and most others) that such a policy was unlikely to work. Similarly, many analysts have been puzzled by the persistence of policymakers in many Latin American countries with extreme neoliberal policies, despite the fact that they seemed to not work. Majumdar and Mukand (2004) suggest that the reason is perhaps reputational. In particular, suppose that the initial choice of policy is a function of the policymaker's ability. In this case even if the policy seems to be failing, the policymaker may persist with it even if it is not efficient to do so. This is because a policy reversal by the incumbent will call into question the incumbent's competence in choosing this course of

action in the first place. Fear of the adverse reputational (and electoral) consequence that such a policy reversal entails results in inefficient policy persistence.

We have delineated above a number of political mechanisms that make policy reform difficult. In response to the difficulty of reforming policies, recent work has focused on the implications of institutional innovations.

Institutions and policy reform

A country's institutional structure affects the political context of policymaking. Any inefficiencies in policy reform are likely to arise from an inability of existing institutions to mediate and resolve conflicts between groups or politicians. This central point about the role of institutional structures in stifling policy reform and growth has been made by Dixit (2004), Rodrik, Subramaniam and Trebbi (2004) and Acemoglu, Johnson and Robinson (2004) among others. Accordingly, recent work has focused on issues of institutional design and the appropriate intervention most likely to change political (and economic) outcomes.

Broadly described, there have been two kinds of institutional interventions that can potentially alter the political equilibrium. The first is directed at resolving policy inefficiency arising out of distributional conflicts. The classic institution that facilitates conflict resolution is of course democratization and regular elections. While democratic elections bring their own inefficiencies, they have a positive first-order effect in that they give the electorate an opportunity to replace a policymaker who fails to undertake policy reform. Alternatively, a constitutional change in the nature of local government can help resolve policy inefficiencies. For instance, explicit political reservations for women and disadvantaged groups can directly alter an existing political equilibrium to one where policy reforms that benefit these groups will be more likely to take place (see Duflo and Chattopadhya, 2004). The second kind of institutional intervention is one that helps limit the possibility of rent-seeking activity by the government. This may involve insulation of some of the policymaking apparatus from political pressures (as in an independent judiciary or central bank) or it may involve term limits or greater decentralization.

Final remarks

Since the late 1980s the study of the political economy of policy reform has become an active area of research in political economy. Indeed, many of the seminal contributions to the area of the ‘new political economy’ were first made in an attempt to understand policy inefficiencies in developing countries and their failure to reform. Many empirical puzzles such as the inefficient delay in enacting reforms and the reversal of optimal policies have been addressed. In addition, the study of policy reform has given the initial impetus to the literature on institutions and underdevelopment.

However, much remains to be done. At an empirical level, we need to understand what institutions are likely to facilitate reform and prevent inefficiency. The need for micro-based studies is particularly important given the vast differences in history and socio-cultural norms that may result in both economic and political markets functioning in unexpected ways. Furthermore, more work needs to be done to understand the political economy of institutional change and its impact on policy reform. In particular, the role of leadership in catalysing institutional change and policy reform is poorly understood.

See Also

- political competition
- political economy
- political institutions, economic approaches to
- public choice

Bibliography

Acemoglu, D. 2003. Why not a political coase theorem? Social conflict, commitment and politics. *Journal of Comparative Economics* 31, 620–52.

Acemoglu, D., Johnson, S. and Robinson, J. 2004. Institutions as the fundamental cause of long run growth. In *Hand book of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Alesina, A. and Drazen, A. 1991. Why are stabilizations delayed? *American Economic Review* 81, 1170–88.

Besley, T. 2006. *Principled Agents: The Political Economy of Good Government*. Oxford: Oxford University Press.

Besley, T. and Coate, S. 1998. Sources of inefficiency in a representative democracy: a dynamic analysis. *American Economic Review* 88, 139–56.

Buchanan, J. and Tullock, G. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.

Coate, S. and Morris, S. 1999. Policy persistence. *American Economic Review* 89, 1327–36.

Dixit, A. 2004. *Lawlessness and Economics: Alternative Modes of Governance*. Princeton: Princeton University Press.

Dixit, A. and Londregan, J. 1996. The determinants of success of special interests in redistributive politics. *Journal of Politics* 58, 1132–55.

Duflo, E. and Chattopadhyay, R. 2004. Women as policy makers: evidence from a randomized policy experiment in India. *Econometrica* 72, 1409–43.

Fernandez, R. and Rodrik, D. 1991. Resistance to reform: status-quo bias in the presence of individual-specific uncertainty. *American Economic Review* 81, 1146–55.

Grossman, G. and Helpman, E. 1994. Protection for sale. *American Economic Review* 84, 833–50.

Jain, S. and Mukand, S.W. 2003. Redistributive promises and the adoption of economic reform. *American Economic Review* 93, 256–64.

Kydland, F. and Prescott, E. 1977. Rules rather than discretion: the time inconsistency of optimal plans. *Journal of Political Economy* 85, 473–90.

Majumdar, S. and Mukand, S.W. 2004. Policy gambles. *American Economic Review* 94, 1207–22.

Robinson, J.A. 1998. Theories of bad policy. *Policy Reform* 1, 1–46.

Rodrik, D. 1996. Understanding economic policy reform. *Journal of Economic Literature* 34, 9–41.

Rodrik, D., Subramanian, A. and Trebbi, F. 2004. Institutions rule: the primacy of institutions over geography and integration in economic development. *Journal of Economic Growth* 9, 131–65.

How to cite this article

Mukand, Sharun W. "policy reform, political economy of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000363> doi:10.1057/9780230226203.1295

The New Palgrave Dictionary of Economics Online

political arithmetic

Phyllis Deane
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chalmers, G.; Colquhoun, P.; Davenant, C.; demography; Graunt, J.; King, G.; Massie, J.; national income; Petty, W.; political arithmetic; political economy; Young, A.

Article

The term ‘political arithmetic’ predates the term ‘political economy’. It was coined by Sir William Petty, a founder member of the Royal Society, who – being a scientist by education and a government economic adviser by career choice – deliberately set out to apply the new scientific methodology of the 17th century to the practical economic problems of the modern nation state. For the leading spirits of the scientific revolution which reached a climax in the second half of the 17th century, the common article of faith was a belief in the unity of theory and practice, combined with a conviction that the first step in the advancement of human understanding in any sphere of knowledge – whether in astronomy or in chemistry or in industrial or social technology – was to lay a foundation of direct, empirical observations. To quote Bacon's *Novum Organum:*

> The roads to human power and to human knowledge lie close together, and are nearly the same; nevertheless, on account of the pernicious and inveterate habit of dwelling on abstractions, it is safer to begin and raise the sciences from those foundations which have relation to practice and let the active part be as the seal which prints and determines the contemplative counterpart.

That was the inspiration which underlay the foundation of the Royal Society and allowed men like Graunt and Petty, Newton and Boyle, Flamsteed and Hooke, to feel themselves part of the same intellectual community. That was also the inspiration for the first exercises in political arithmetic. Sir William Petty was a medical practitioner (as was his contemporary Locke, and Quesnay a century later), and his interest in economic problems had been stimulated in Ireland, to which he went in the early 1650s as physician to the Cromwellian army of occupation. There he persuaded the civil

authorities, faced with the problem of consolidating the conquest by making an orderly distribution of forfeited lands, to give him the task of organizing a comprehensive land survey. It was on the basis of this massive research project that he wrote *The Political Anatomy of Ireland* in the 1670s. But by then he had already published his *Treatise on Taxes and Contributions* (1662), which contained a miscellany of sharp observations, incisive economic analysis and forthright policy advice, mainly focused on English problems of public finance. He had also written (during the 1665–7 conflict with Holland) an essay in a similar analytical mould, concerned with the practical problems of financing the war, and it was in this connection that he produced his first estimates of national income and wealth for England and Wales (published posthumously as *Verbum Sapienti*).

Most of Petty's pamphlets on economic questions were circulated privately and published posthumously, for the second half of the 17th century was an age in which giving politico-economic advice to governments was a perilous occupation. However, the distinctive message running through these writings was the importance of basing public economic policies on systematically compiled empirical evidence and reasoned quantitative estimates of the nation's human and material resources. In the 1670s he spelt out and developed this theme in his most path-breaking work, *Political Arithmetick*, subtitled: ‘A discourse concerning the extent and value of lands, people, buildings, husbandry, manufacture, commerce, fishery, artisans, seamen, soldiers, public revenues, interest, taxes, superlucration, registries, banks, valuation of men, increasing of seamen, of militias, harbours, situation, Power at sea, etc. As the same relates to every country in general, but more particularly to the Territories of His Majesty of Great Britain and His Neighbours of Holland Zealand and France.’ Written in order to rebut those commentators who were lamenting the nation's economic decline, Petty's *Political Arithmetick* was an explicit attempt to apply a Baconian methodology to economic analysis. According to his preface:

> The Method I take to do this is not yet very usual; for instead of using only comparative and superlative Words and intellectual Arguments, I have taken the course (as a Specimen of the Political Arithmetick I have long aimed at) to express myself in terms of *Number*, *Weight* or *Measure*; to use only Arguments of Sense and to consider only such Causes as have visible Foundations in Nature, leaving those that depend on the Mutable Minds, Opinions, Appetites and Passions of particular men, to the Consideration of others.

Petty concluded his preface in the self-consciously undogmatic spirit of the ‘new science’ by inviting other seekers after truth to confront his results with rational criticism and new data:

> I hope all ingenious and candid Persons will rectifie the Errors, Defects and Imperfections which probably may be found in any of the Positions upon which these Ratiocinations were grounded. Nor would it misbecome Authority itself to clear the Truth of those matters which private Endeavours cannot reach to.

Petty is sometimes credited with having founded the first ‘school’ of economic thought. But it would be more accurate to say that he had launched the first scientific research programme in political economy. Indeed, in carrying over a Baconian scientific ideology to an analysis of public financial issues, he was simply epitomizing the spirit of his age, for political arithmetic was not narrowly economic in its scope. It was his friend John Graunt, for example, a London draper, who in 1662 published a pamphlet entitled

Natural and Political Observations Made upon the Bills of Mortality and who, by applying a logical technique of coordination and deduction to the limited vital statistics that had been collected for London over the preceding century, took the first steps in the formulation of the modern science of demography. Using the death returns and other data regularly published in the London Bills of Mortality, plus some personally assembled data for a few country parish records, Graunt made the first reasoned estimates of total population, not only for the metropolis, but also for the country as a whole, and even set up the first life table. Significantly, Graunt's election to the Royal Society was made within a month of the publication of the first edition of his pamphlet, and was strongly supported by Charles II who (according to Sprat, the first secretary and historian of the Royal Society) ‘gave his particular charge to His Society, that if they found any more such tradesmen, they should be sure to admit them all, without any more ado’.

It was indeed Graunt rather than Petty who inspired Gregory King to produce his demographic estimates in the 1690s, for Petty's way with figures was somewhat impressionistic. In particular, his results were less likely to be systematically cross-checked against the results suggested by alternative data sources, or by different sets of assumptions, than was the case for either Graunt or King. By the same token, Gregory King's estimates of national income were more meticulously justified, more detailed, more internally consistent, and hence more credible in their delineation of the dimensions of the economy and in international or intertemporal comparisons than were Petty’s. Graunt and King, that is to say, were both more sophisticated economic statisticians than Petty. On the other hand, it was Petty's imaginative and ambitious use of his estimates as a basis for economic analyses and policy prescriptions that earned him his reputation as the leading political arithmetician. It is doubtful whether King's estimates, for example, would have had more than an ephemeral currency had they not been so brilliantly applied in the course of the polemical analyses of Charles Davenant – an MP and a public official of some weight, who had held inter alia the posts of commissioner of the excise 1683–9, inspector general of exports and imports 1705–15, and secretary to the commission set up to negotiate the Union with Scotland.

The half-century following the Restoration was the golden age of political arithmetic, but as a method of economic analysis it failed to develop appreciably during the next two centuries. Petty's (or occasionally King’s) estimates of national income were often cited by mercantilist pamphleteers without any attempt at updating. True, there were from time to time new aggregative valuations of the nation's total income or product designed to put into perspective a polemical argument relating to a particular sector. For example, in 1760 Joseph Massie updated King's ‘Scheme of the Income and Expence of the several Families of England Calculated for the Year 1688’, in order to establish a framework for his own estimates of excise tax incidence in a polemic robustly entitled *A Computation of the Money that hath been exorbitantly Raised upon the People of Great Britain by the Sugar Planters in one Year from January 1759 to January 1760; shewing how much Money a Family of each Rank Degree or Class hath lost by that rapacious Monopoly* … Similarly, Arthur Young, who was mainly concerned to prescribe for and defend the economic interests of the agricultural sector, made some reasonably careful and well-informed estimates of the nation's agricultural output in the course of his reports on his *Tours* of the northern and eastern counties of England, and associated these estimates with more casual calculations of value added in manufacturing, commerce and various other industries. Young, who had a deservedly high reputation as an agricultural economist, but was undistinguished as a general economist, was probably the last of the political arithmeticians in the original sense of the term. Certainly he was the last economist to write under that banner, which by the 1930s had been annexed by the demographers. His

treatise, entitled *Political Arithmetic, Containing Observations on the Present State of Great Britain and the Principles of her Policy in the Encouragement of Agriculture, to Which is Added a Memoir on the Corn Trade*, was a commentary on current agricultural issues which incidentally summarized and reconsidered some of his earlier national product estimates originally published in the *Tours*.

No doubt because the new discipline of political economy that took shape in the late 18th and early 19th centuries took off in more theoretical, less Baconian directions, political arithmetic lost its capacity to attract innovative exponents. Perhaps Adam Smith gave the *coup de grâce* to the whole approach when he announced in *The Wealth of Nations* that he had ‘no great faith in political arithmetick’ – though he was not above borrowing some of the political arithmeticians’ estimates when they served the purpose of his argument. The third (1787) edition of the *Encyclopedia Britannica* (which did not include an entry for political economy) contained a lengthy piece on political arithmetic, defined as ‘the art of reasoning by figures upon matters relating to government, such as the revenues, number of people, extent and value of land, taxes, trade, etc. in any nation’. The explanation went on: ‘These calculations are generally made with a view to ascertain the comparative strength, prosperity etc. of any two or more nations.’ Most of the entry was devoted to citations from Petty, although it referred also to Davenant, King, Graunt, Halley (the astronomer who had constructed a life table) and to various contributors to the demographic debates which flared up in the second half of the 18th century – such as Brakenridge and Price. After the 18th century, however, political arithmetic ceased to rate an entry in the *Encyclopedia Britannica* in its own right, and even the entries on political economy failed to notice it as a substantial episode in the history of economic ideas.

True, the idea of quantifying the total national income or wealth did not die, and when decennial population estimates were introduced, from 1801 onwards, the bases for such calculations became less speculative. The significance of Petty's role was that he had broken new ground in setting his analyses and associated policy prescriptions within a framework of national aggregates on whose structural relationships and absolute magnitudes the nation's productive strength and taxable capacity evidently hinged. A long, if sporadic, stream of national income estimates was accordingly produced by diligent researchers following in Petty's or King's footsteps – usually in relation to questions of war finance or taxable capacity or comparative economic strength. At the turn of the century, for example, Pitt's plans to raise an income tax to finance the French war stimulated a flurry of national income estimates. At about the same time, George Chalmers put Gregory King's *Natural and Political Observations* in an appendix to the fourth edition of his bestselling *Estimate of the Comparative Strength of Great Britain* (1802); and the appearance in print (for the first time ever) of King's famous table of incomes by families inspired Patrick Colquhoun, then researching the poverty problem, to update King's 1688 results for his *Treatise on Indigence* (1806). Less than a decade later, Colquhoun carried his statistical enterprise even further (with the aid of the first two population censuses) by publishing elaborate and detailed estimates of national income and wealth for the United Kingdom as a whole. This set the stage for the subsequent stream 19th-century estimates of national income which began essentially with Pebrer in 1833 and ended with Mulhall's *Dictionary of Statistics* (1890). Most of these, however, were exercises in descriptive statistics rather than in economic analysis.

In effect, then, Petty's aggregative approach to quantifying and analysing the nation's resources fell into disuse among leading economists as the abstract notion of a self-equilibrating economic system gradually took precedence over the essentially political concept of the royal domain as the central object of economic analysis. What 17th- or early 18th-century economic advisers typically addressed

themselves to were the practical problems of the nation state, and these were seen as analogous to the problems of managing a household. Petty, for example, had dedicated his *Political Arithmetick* to the king, because it was the royal domain whose resources he was endeavouring to assess, and its management problems that the quantification was designed to inform. No doubt it was inevitable that when economists founded their theories on the assumption that ‘things will have their course’ in the politico-economic as in the natural universe, and while the role of the state within the wider economic system was assumed to be constrained by the sheer futility of legislating against the ‘laws of nature’, there was little incentive to extend national income analysis beyond the rudimentary levels it reached in the golden age of political arithmetic. Accordingly, the analytical approach to the study of national income was largely ignored by economists until the middle decades of the 20th century when J.M. Keynes's macroeconomic theorizing revolutionized their discipline.

See Also

- King, Gregory
- Petty, William

Bibliography

Colquhoun, P. 1806. *Treatise on Indigence*. London: J. Hatchard.

Colquhoun, P. 1814. *Treatise on the Wealth, Power and Resources, of the British Empire*. London: Joseph Mawman.

Davenant, C. 1771. In *The Political and Economic Works of that celebrated writer Charles D’Avenant*. 2 vols, ed. Sir Charles Whitworth. London.

Deane, P. 1955. Implications of early national income estimates for the measurement of long-term economic growth in the United Kingdom. *Economic Development and Cultural Change* 4(November), 3–38.

Deane, P. 1956–7. Contemporary estimates of national income in the nineteenth century. *Economic History Review*, Pt I, 8 (1956), 339–54; Pt II (1957), 451–61.

Graunt, J. 1662. *Natural and Political Observations upon the Bills of Mortality*. Reprinted in *The Economic Writings of Sir Wm. Petty*, vol. 2, ed. C.H. Hull, Cambridge: Cambridge University Press, 1899.

Hogben, L. 1938. *Political Arithmetic: A Symposium of Population Studies*. London: Allen & Unwin.

King, G. 1936. In *Two Tracts by Gregory King*, ed. G.E. Barnett. Baltimore: Johns Hopkins Press.

Massie, J. 1760. *A Computation of the Money that has been Exorbitantly Raised upon the People of Great Britain by the Sugar Planters*. London.

Mathias, P. 1957. The social structure in the eighteenth century: a calculation by Joseph Massie. *Economic History Review* 10, 30–45.

Mulhall, M. 1890. *Dictionary of Statistics*. London: G. Routledge & Sons.

Pebrer, P. 1833. *Taxation, Power, Statistics and Debt of the Whole British Empire*. London.

Petty, W. 1899. In *The Economic Writings of Sir William Petty*. 2 vols, ed. C.H. Hull. Cambridge: Cambridge University Press.

Young, A. 1770. *A Six Months Tour Through the North of England*. 4 vols. London: W. Strahan.

Young, A. 1771. *The Farmer's Tour Through the East of England*. 4 vols. London: W. Strahan.

Young, A. 1774. *Political Arithmetic*. Pt I. London: W. Nicoll.

Young, A. 1779. *Political Arithmetic*. Pt II. London: T. Cadell.

How to cite this article

Deane, Phyllis. "political arithmetic." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000291> doi:10.1057/9780230226203.1296

The New Palgrave Dictionary of Economics Online

political budget cycles

Allan Drazen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Theoretical and empirical research on political budget cycles is surveyed and discussed. Significant political budget cycles are seen to be primarily a phenomenon of the first elections after the transition to a democratic electoral system.

Keywords

asymmetric information; competence; democracy; electoral cycles; electoral systems; fiscal manipulation; fiscal policy; fiscal transparency; imperfect information; moral hazard; political budget cycles; political expenditure cycles; political revenue cycles; rational voting; separating equilibria; signalling; targeted public spending

Article

Political budget cycles are cycles in some component of the government budget induced by the electoral cycle. More specifically, the term most often refers to increases in government spending or the deficit or decreases in taxes (including changes relative to long-term trends) in an election year which are perceived as motivated by the incumbent's desire for re-election for himself or his party. Though political budget cycles may be seen as just one type of political cycle in macroeconomic variables, most research on cycles in economic variables induced by elections now focuses on budget cycles, and it is useful to study such cycles independent of political cycles in economic activity (the political business cycle). The shift in focus is due in part to the lack of strong empirical evidence for the existence of a political business cycle in many countries.
In contrast to the literature on the political business cycle – where development of formal models preceded the bulk of empirical testing – much empirical research on political budget cycles is based not on explicit models but on more conceptual arguments, with sophisticated formal models being developed later to show how the existence of cycles could be consistent with rational voters. In this

article, we first review the basic conceptual arguments and then the formal models before considering the empirical research. There are two key empirical questions. The first is whether political budget cycles in fact exist in a large number of countries. Recent evidence, discussed below, suggests that they do not on the aggregate budget level, except for new democracies. The second key question, which underlies the first, is whether manipulation of the budget is an effective tool in gaining votes. Though it is widely believed that deficit spending in an election year in general gains votes for the incumbent, empirical research does not support this view.

Basic conceptual arguments

There are two main (and contradictory) views of pre-electoral fiscal manipulation. One is that politicians may be expected to engage in such manipulation and that empirically it is widespread. A simple argument supporting this view is that voters like low taxes and high government expenditures, and vote for incumbents who provide them. Opportunistic incumbents will therefore use expansionary fiscal policy before elections to increase the probability of re-election.

However, this simple argument is inconsistent with rational, forward-looking voters who are aware of government budget constraints both at a point in time and intertemporally. Since the non-smooth paths of taxes and government expenditures implied by election-year deficits are presumably costly, voters should dislike deficits in general and especially those seen as electorally motivated. They would therefore not reward incumbents who engage in election-year manipulation. Hence, the alternative view is that voters (especially in developed countries) are ‘fiscal conservatives’ who punish rather than reward fiscal manipulation. Evidence, discussed in greater detail below, suggests that this is the case in developed countries with established democracies.

A second argument is that if voters respond to good economic conditions by being more likely to vote for the incumbent, he will use expansionary fiscal policy to try to manipulate macroeconomic outcomes and provide higher growth. Hence, expansionary fiscal policy will help an incumbent's re-election prospects. However, even if good economic conditions help an incumbent's chances of re-election, it is not clear that fiscal manipulation will be effective – politicians may have very limited ability to manipulate the economy successfully, both because of a lack of technical ability to time the expansion accurately enough to happen just before the elections and because, as discussed above, rational, well-informed voters should not support such policies.

A more sophisticated argument on why rational voters may respond to pre-electoral fiscal expansions is that they have imperfect information about candidates’ abilities or about the environment, and that a fiscal expansion signals incumbent ability or some other characteristic which voters value, so that it is effective in gaining votes. This was first formalized in the work of Rogoff, which is summarized below.

An alternative is that, if voters do punish election-year deficits or spending increases (as the data indicate for developed countries), electoral manipulation takes the form of changes in the composition of the budget rather than in its overall level (or the overall deficit). This may take the form of increases in spending that voters as a whole favour at the expense of those types of spending that voters may be believed to like less (or are less visible), or the form of expenditures targeted at some voters at the expense of other voting groups who are seen as electorally less valuable.

Signalling models

The basic competence model

Formal modelling of the signalling role of a pre-election fiscal expansion under asymmetric information was introduced by Rogoff and Sibert (1988) and Rogoff (1990). The models are based on unobserved ‘competence’, that is, the ability to deliver more public goods for the same level of taxes. Hence, more competent policymakers can generate higher welfare and so are preferred by voters. Competence is correlated over time, so that a candidate who is believed by voters before an election to be more competent than average (the presumed competence of his randomly drawn challenger, who is unable to signal) is expected to be more competent than average after the election as well. Voters therefore rationally prefer a candidate who delivers higher expenditures before an election, since this is a signal of higher competence.

The basic ideas can be represented by a simple version of the model in Rogoff (1990). There is an election at the end of the first period, with the leader who is elected remaining in office thereafter. Voters will choose the leader on the basis of any information they gather in the first period. The utility of the representative voter as of period *t* may be represented by

$$\Gamma_t = \sum_{s=t}^{T} \beta^{s-t}(g_s + v(k_s)) + \eta_t \tag{1}$$

where g_s is public consumption and k_s is public investment. The function $v(.)$ is assumed to be increasing, concave and satisfying the Inada conditions on its first derivatives as k goes to zero or infinity. The term η_t is a random shock in the election period $t = 1$ such that the outcome is not known *ex ante* to the incumbent setting policy. The voter maximizes the expected value of utility by choosing a candidate in an election at the end of the first period.

The production of public goods is represented as follows. If a leader has an ‘administrative ability’ or ‘competence’ ε, he can produce public goods at time *t* according to:

$$\varepsilon = g_t + k_{t+1} \tag{2}$$

where it is assumed that ε is not directly observable. Investment *k* must be chosen one period in advance, so that it is not currently observable. Hence, if a voter observes a high value of g_t, he does not know whether this reflects high ability of the policymaker (high ε) or high current public consumption ‘bought’ at the expense of a cut in some other component of public spending (here, public investment) at

some point in the future. This is meant to represent the basic inference problem a voter faces when he observes high government spending before an election – does high observable government expenditure represent fiscal manipulation, in the sense of implying that taxes will be raised or other programmes cut in the future, or does it represent the ability of the leader to provide more goods or services without cutting future goods services?

Potential leaders are assumed to differ in their unobserved ability. Suppose there are two possible levels of ε: ε^H and $\varepsilon^L < \varepsilon^H$, where ability ε_j is expected to persist after the election. Let the prior probability that $\varepsilon = \varepsilon^H$ be $0 < \rho < 1$. The voter's inference problem is to use an observation of g to try to infer the probability that the leader is high-ability, that is, to form a posterior $\hat{\rho}(g)$.

The utility of the incumbent leader is given by:

$$E_t\Gamma_t + \left(\chi + q\sum_{s=t+1}^{T}\beta^{s-t}\chi\right)$$

where χ is the value of holding office and q is the probability of being re-elected at the end of the first period. A key point is that a policymaker's utility depends both on social welfare (the first term) and on his own private payoffs (the second term). If it depended only on social welfare, incumbents would choose the socially optimal fiscal policy and there would be no signalling. If it depended only on private payoffs, low-ability incumbents would mimic whatever high-ability incumbents do and there would only be a pooling equilibrium with no signalling.

At the beginning of period 1, the incumbent observes his ε^j, sets g_1 and k_2 (where k_1 is predetermined). Voters then observe g_1 and f_1 and then vote at the end of the period for either the incumbent or a randomly drawn challenger (who cannot signal his competence, which is average expected competence $\bar{\varepsilon}$ given the prior ρ.) In subsequent periods, the elected policymaker chooses g_t and k_{t+1} to maximize social welfare, given his competence ε. This first-best solution is given by maximizing (1) subject to (2), yielding $k^* = v'^{(-1)}(1/\beta)$ and $g^*(\varepsilon^j) = \varepsilon^j - k^*$. (This would also be the solution in period 1 if voters knew the incumbent's ε.) Since higher-ability incumbents provide more public goods, and thus higher utility, voters prefer a high-ability incumbent to the challenger of expected ability $\bar{\varepsilon}$, but prefer the challenger to a low-ability incumbent.

Under asymmetric information (that is, when the representative voter does not observe the incumbent's ε before voting, or cannot infer it because of imperfect information about the components of the budget), a voter's beliefs about an incumbent's ability are conditioned on his observation of g_1. These beliefs can be summarized as the posterior probability $\hat{\rho}(g_1)$ the voter assigns to the incumbent being of ability ε^H conditional on the value of g_1 observed. Given the voters' rational voting rule, an incumbent has an incentive to appear to be of high ability.

The equilibrium is a separating equilibrium in which the level of spending reveals the incumbent's

competence type. A high-ability incumbent will spend just enough so that the low-ability incumbent will not find it optimal to mimic him. (Since a high-ability incumbent can invest $\varepsilon^H - \varepsilon^L$ in k_2 for the same level of g_1, and since politicians care about social welfare, concavity of $v(k)$ implies that the high-ability type can cut back on k at a lower marginal cost to himself than the low-ability type can, the signal of raising g_1 is less costly for him to send.) The low-ability incumbent will choose the first-best solution for his type, namely, $g_1 = g^*(\varepsilon^L) = \varepsilon^L - k^*$. Since this reveals his type he loses the election almost certainly.

If the values of ε^H and ε^L are far enough apart, then the high-ability incumbent can signal his type by choosing his first-best $g^*(\varepsilon^H)$, which the low-ability type won't mimic. However, if ε^H and ε^L are sufficiently close, then a high-ability incumbent can signal his type only by choosing $g_1 > g^*(\varepsilon^H)$. With a continuum of ability types, then each type separates from the type immediately ‘below’ him by choosing a $g_1 > g^*(\varepsilon^j)$, except for the lowest-ability type who plays his first best. Hence, there is the general result that there will be a fiscal expansion in an election year relative to non-election years, not because voters are naive but because they are sophisticated.

Timing of signals

A question often raised about election-year expansions as a signal of competence (or some other desirable characteristic of a politician) is why the signal should be sent just before an election, rather than earlier in the politician's term. The argument in this sort of model is that information about such characteristics evolves over time, so that there is new information to be signalled in the time period before an election. At the same time the desirable characteristic must have some persistence, so that its pre-electoral value provides information about its post-electoral value. (Formally, Rogoff modelled this by assuming there was an election at the end of every other period, with ability ε assumed to be the sum of the current period and previous period's i.i.d. shock, that is, an AR(1) structure. Therefore, information signaled by g_t in period t before an election was relevant for the post-electoral period $t+1$, but not for the subsequent election at $t+2$. This makes the incumbent's choice problem for choice of g_t fairly simple.)

Observability of fiscal policy

A key ingredient of this type of signalling model focusing on competence is voters’ inability to observe the overall level of spending or of the deficit, for otherwise they could perfectly infer his competence. The reliance of this result on voters’ lack of information is consistent with Brender and Drazen's (2005a) empirical finding of no statistically significant aggregate deficit or expenditure cycle in established democracies, where voters may be well-informed about fiscal outcomes. Gonzalez (2002) and Shi and Svensson (2002) extend the Rogoff model to study the effect of transparency on the magnitude of fiscal cycles. The basic result is that the higher the degree of transparency, the lower is the amount of distortion away from the first best in the political budget cycle. Shi and Svensson include a similar measure of transparency. Shi and Svensson further argue that, while the proportion of uninformed voters

– who may be influenced by fiscal manipulation – is initially large, it is likely to decrease over time, thus decreasing the magnitude of budget cycles. They create a measure of the availability of information and show that as voters become more informed the magnitude of the cycle decreases. A key innovation of Shi and Svensson (2002) is that the policymaker chooses fiscal policy before he knows his competence level, so that all ‘types’ choose the same level of expansion. That is, the model focuses on moral hazard rather than signalling, as the other models do. An implication is a cycle in the aggregate deficit.

Unobserved politician preferences

The argument that, with high transparency, political cycles in aggregate expenditures or deficits are likely to be weak or non-existent (combined with empirical evidence on the absence of political cycles in budget aggregates in countries where transparency is seen as high) has led to alternative signalling models. If voters are fiscal conservatives, election-year fiscal manipulation may take the form of changes in the composition of the budget with overall spending and deficits held constant. These compositional changes may be either in categories of expenditures or in expenditures or transfers targeted to some voters at the expense of others.

Drazen and Eslava (2005; 2006) argue that, if it is the composition of spending or transfers, rather than their overall level, that is manipulated for electoral purposes, rational voters may be trying to infer something other than (or in addition to) competence from election-year fiscal policy. Voters who are targeted before an election want to know whether they will be similarly favoured after the election. They therefore suggest that a key unobserved characteristic of an incumbent politician is his preferences over groups of voters or types of expenditure. As in the Rogoff competence models, these preferences have some persistence over time, so that a voter who believes that the incumbent favours him before the election rationally expects some similarity in the composition of expenditures after the election as well. A voter thus faces an inference problem – whether receiving high targeted expenditures before the election signals a greater weight of his group in the incumbent's objective function than other voters or non-targeted expenditures, or whether it signals simply how ‘swing’ his demographic group is, meaning how many votes the incumbent can raise by targeting his group with expenditures. In both papers, Drazen and Eslava show the existence of an equilibrium in which voters rationally respond to election-year expenditures and politicians allocate expenditure on the basis of this behaviour. Politicians increase spending targeted to electorally attractive groups before elections, while they reduce other types of expenditure to satisfy the no-deficit constraint. As mentioned, a key result is that electoral manipulation arises even with fully rational voters. Drazen and Eslava (2006) further show that even when voters know how ‘swing’ their group is a political cycle may still arise.

There are several key differences between competence as the crucial unobserved characteristic and the approach of Drazen and Eslava, where a politician's preferences are unobserved and spending is targeted to some groups of voters or types of expenditure at the expense of others. First, in the latter approach, manipulation may occur even without affecting the aggregate deficit, consistent with empirical findings discussed below. Second, electoral fiscal manipulation arises even if voters can perfectly monitor the fiscal choices of an incumbent. Finally, political budget cycles in the Drazen and Eslava models arise even if all politicians are equally able to provide public goods.

Empirical studies of political budget cycles

Empirical studies of political budget began with the work of Tufte (1978) for the United States, followed by numerous other empirical studies for both developed and developing countries, as summarized in Drazen (2001). Political budget cycles were widely believed to be strongest for developing countries. More recently, a number of papers have argued that, while these cycles are stronger in developing countries, they characterize democracies at all levels of economic development, and even non-democracies. Shi and Svensson (2002) find that, in a large panel of both democracies and non-democracies over the period 1975–95, the government deficit rises significantly in an election year in both developing and developed countries. (They show that the effect is far stronger in developing countries, consistent with earlier studies.) The economic effect is significant for the sample as a whole, the fiscal surplus falling on average in their full sample by one half to one per cent in an election year, depending on the estimation method they use. Persson and Tabellini (2003) restrict their sample to a group of 60 democracies from 1960 to 1998. They find a political revenue cycle (government revenues as a percentage of GDP decrease before elections), but no political cycle in expenditures, transfers, or the overall budget balance across countries or political systems. They argue that the electoral system (proportional versus majoritarian) and the governmental system (presidential versus parliamentary) is a key determinant of the nature of the cycle across countries.

However, Brender and Drazen (2005a) argue that the political deficit cycle in democracies is a phenomenon of recently democratized countries, that is, are found to be statistically significant only in the first few elections after a country has made a transition from being a non-democracy to a democracy (which holds true whether or not the formerly socialist economies are included). It is the strong political budget cycle in these countries that accounts for the political budget cycle in larger samples including these countries. Once these countries are removed from the larger sample, the political fiscal cycle disappears. This is true in both developed and developing countries. Hence, the stronger results previously found for developing countries reflect the fact that new democracies comprise a larger fraction of developing than developed country democracies. The ‘new democracy’ effect also helps explain previous findings of a stronger political cycle in weaker democracies (new democracies are a larger fraction of ‘weak’ than ‘strong’ democracies, with no significant cycle found in weak, old democracies). They also find that helps account for differences in the political cycle across government or electoral systems.

There is also a significant political expenditure cycle in the new democracies, with the very similar positive coefficients on the fiscal deficit and on expenditures in the analogous equations, while there does not appear to be a statistically significant revenue cycle. The deficit cycle in the new democracies thus appears to be driven by higher election-year expenditures.

Brender and Drazen suggest several explanations for their ‘new democracy’ finding. One is that fiscal manipulation may be used in new democracies because voters are inexperienced with electoral politics or may simply lack the information needed to evaluate fiscal manipulation that is produced in more established democracies. This suggests one way to reconcile the two contradictory views of pre-electoral manipulation. The argument that politicians may be expected to engage in such manipulation may apply to new democracies, where it is possible to carry out such manipulation. The alternative that voters punish fiscal manipulation is applicable to established democracies, where voters have the ability to

identify fiscal manipulation and punish such behaviour, so that politicians avoid it.

This is consistent with work by Gonzalez, Shi, and Svensson, discussed above, that focuses on information asymmetries in explaining budget cycles when voters are not naive. It is also consistent with findings by Akhmedov and Zhuravskaya (2004), who find similar evidence in regional elections in Russia after its transition to democracy. Using monthly data between 1996 and 2003, they found sizable but short-lived political budget cycles in local fiscal spending, which became significantly smaller over time and disappeared for most (but not all) fiscal instruments after two rounds of elections. Akhmedov and Zhuravskaya (2004) find similarly that measures of the freedom of the regional media and the transparency of the regional governments were important predictors of the magnitude of the cycle. Alt and Lassen (2006a) find that in OECD countries higher fiscal transparency also lowers the magnitude of the electoral cycle.

The absence of political cycles in budget aggregates in established democracies as a group does not, however, mean there are no electoral effects on fiscal policy. Established democracies appear to be characterized by cycles in the composition of spending rather than cycles in its overall level. Several papers find evidence of electoral composition changes in government spending at the sub-national level, including the United States (Peltzman, 1992), Canada (Kneebone and McKenzie, 2001), Colombia (Drazen and Eslava, 2005), India (Khemani, 2004), and Israel (Brender, 2003). Drazen and Eslava (2005) present a signalling model of composition cycles with rational voters where the unobserved characteristic of politicians is their preferences for different types of expenditure, specifically those types of expenditure that voters as a whole prefer.

A second possible explanation for the new democracy effect follows from the Brender and Drazen (2005b) finding that fiscal balance has no significant effect on the probability of re-election, a surprising finding given the existence of a political budget cycle in new democracies. The authors suggest that these two findings may be reconciled by the possibility that fiscal expansions in election years in new democracies do not represent an attempt to gain voter support for the leader but reflect expenditures incurred in an attempt to consolidate democracy. Democracy is often not ‘consolidated’ in new democracies, that is, it is not accepted unconditionally by all citizens. An election year may be an especially dangerous time for the existence of the democracy itself, and thus may be a time when leaders have to spend money to retain popular support for the democratic regime to prevent its overthrow or subversion and the return to an autocratic system. One might then observe higher expenditures and deficits in an election year, but without fiscal expansion necessarily gaining votes for the incumbent over the challenger.

The effect of deficits on re-election

In contrast to the fairly extensive direct tests of overall macroeconomic performance on election outcomes in the literature on political business cycles, there are few tests of fiscal performance on election outcomes, primarily at the sub-national level. These include Peltzman (1992), Brender (2003), and Drazen and Eslava (2005), who examine the direct effect of fiscal performance on re-election at the state and local levels in a single country (the United States, Israel, and Colombia respectively), and find that voters punish – rather than reward – loose fiscal policies in general, as well as in election years.

The only large cross-country study is by Brender and Drazen (2005b), who look at the effects of fiscal performance on re-election in a sample of 74 democracies (comprising 350 election campaigns) over the period 1960 to 2003. They estimate probit regressions giving the probability of an incumbent's re-election as a function of macroeconomic and fiscal variables. They find no evidence that expansionary fiscal policy helps a leader to get re-elected; in fact, it is likely to reduce the chances of re-election. In developed countries, especially established democracies, deficits lower the probability of re-election, with an effect that is both statistically and economically significant. In developing countries, the effect of deficits on re-election is close to zero and is not statistically significant. While voters in developing countries may be more tolerant of an expanding budget deficit in election years, even in these countries voters do not reward election-year deficits at the polls. Brender and Drazen find no statistically significant difference between the effect of deficits that are created by higher expenditures and of those that are created by lower revenue, although in the developed countries the effect of revenue reductions (as a share of GDP) is somewhat larger.

They also find that in established democracies in developed countries voters punish election-year deficits and deficits over the incumbent's term of office. The effects are quite substantial quantitatively. An increase of one percentage point in the ratio of the central government surplus to GDP over the term can increase the probability of re-election by 3–4.5 percentage points in the developed, established democracies, and an increase of one percentage point in the surplus during an election year increases the probability of reelection by between seven and nine percentage points.

The Brender–Drazen results indicate that controlling for the type of political system (parliamentary versus presidential) or the type of electoral system (majoritarian versus proportional) does not change the effect of the election year deficit and growth, nor does whether elections were held at their scheduled date or early. Similarly, they find no significant effect of the level of democracy on the finding that deficits do not help re-election chances of an incumbent.

See Also

- political business cycles

Bibliography

Akhmedov, A. and Zhuravskaya, E. 2004. Opportunistic political cycles: test in a young democracy setting. *Quarterly Journal of Economics* 119, 1301–38.

Alt, J. and Lassen, D. 2006a. Transparency, political polarization, and political budget cycles in OECD countries. *American Journal of Political Science* 50, 530–50.

Alt, J. and Lassen, D. 2006b. Fiscal transparency, political parties, and debt in OECD countries. *European Economic Review* 50, 1403–39.

Brender, A. 2003. The effect of fiscal performance on local government election results in Israel: 1989–1998. *Journal of Public Economics* 87, 2187–205.

Brender, A. and Drazen, A. 2005a. Political budget cycles in new versus established democracies. *Journal of Monetary Economics* 52, 1271–95.

Brender, A. and Drazen, A. 2005b. How do budget deficits and economic growth affect reelection prospects? Evidence from a large cross-section of countries. Working Paper No. 11862. Cambridge, MA: NBER.

Drazen, A. 2001. The political business cycle after 25 years. In *NBER Macroeconomics Annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Drazen, A. and Eslava, M. 2005. Electoral manipulation via voter-friendly spending: theory and evidence. NBER working paper 11085. Cambridge, MA: NBER.

Drazen, A. and Eslava, M. 2006. Pork barrel cycles. NBER working paper 12190. Cambridge, MA: NBER.

Gonzàlez, M. 2002. Do changes in democracy affect the political budget cycle? Evidence from Mexico. *Review of Development Economics* 6, 204–24.

Khemani, S. 2004. Political cycles in a developing economy: effect of elections in the Indian states. *Journal of Development Economics* 73, 125–54.

Kneebone, R. and McKenzie, K. 2001. Electoral and partisan cycles in fiscal policy: an examination of Canadian provinces. *International Tax and Public Finance* 8, 753–74.

Peltzman, S. 1992. Voters as fiscal conservatives. *Quarterly Journal of Economics* 107, 327–61.

Persson, T. and Tabellini, G. 2003. *The Economic Effects of Constitutions*. Cambridge, MA: MIT Press.

Rogoff, K. 1990. Equilibrium political budget cycles. *American Economic Review* 80, 21–36.

Rogoff, K. and Sibert, A. 1988. Elections and macroeconomic policy cycles. *Review of Economic Studies* 55, 1–16.

Shi, M. and Svensson, J. 2002. Conditional political budget cycles. Discussion Paper No. 3352. London: CEPR.

Tufte, E. 1978. *Political Control of the Economy*. Princeton, NJ: Princeton University Press.

How to cite this article

Drazen, Allan. "political budget cycles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000346> doi:10.1057/9780230226203.1297

The New Palgrave Dictionary of Economics Online

political business cycles

Allan Drazen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Theoretical and empirical research on political business cycles, both opportunistic and partisan, is surveyed and discussed. The evidence for the existence of empirically significant opportunistic political business cycles is argued to be mixed.

Keywords

central bank independence; competence; imperfect information; infinite horizons; inflationary expectations; natural rate of unemployment; opportunistic vs. partisan business cycles; Phillips curve; political budget cycles; political business cycles; rational expectations; rational voting; signaling

Article

Political business cycles are cycles in macroeconomic variables – output, unemployment, inflation – induced by the electoral cycle. (Political cycles in fiscal policy variables, termed ‘political budget cycles’, are treated in a separate article.) Key questions this literature addresses include the following. Are such cycles observed in the data? What are the political and economic mechanisms that lead to such cycles? What do they imply about voter behaviour?
There are two basic types of models. ‘Opportunistic’ political business cycles are expansions in economic activity induced by an opportunistic incumbent before an election meant to increase his chances of re-election. ‘Partisan’ political business cycles are fluctuations in macroeconomic variables over or between electoral cycles resulting from leaders having different policy objectives.

Opportunistic models

Formal models of the opportunistic business cycle began to appear in the mid-1970s, the most influential of which was that of Nordhaus (1975). The structure of the economy is summarized by a downward-

sloping Phillips curve, yielding a trade-off between unemployment and unexpected inflation. Inflation expectations are formed adaptively on the basis of past observed inflation. Identical voters base their voting decisions on aggregate inflation and unemployment outcomes relative to their most preferred outcomes. They have a preference for both low unemployment and low inflation, but, in evaluating incumbents on the basis of macroeconomic performance, they have short memories and no foresight. An opportunistic incumbent policymaker has no preferences over inflation and unemployment *per se* and cares only about re-election. The slow adjustment of inflation expectations to economic stimulation, combined with myopic voters, allows an opportunistic incumbent to manipulate macroeconomic time paths to his electoral benefit. He stimulates the economy before the election to reduce unemployment, with the inflationary cost of such a policy coming only after.

More formally, the basic opportunistic model may be simply represented as follows. The objective of the policymaker is to maximize his probability of re-election, where voting behaviour is retrospective in that it depends on economic performance under the incumbent in the past. Economic performance in a period is measured by the behaviour of current inflation π_t and unemployment U_t, so that voter dissatisfaction in any period can be represented by a loss function which is increasing in these two variables. Consider, for simplicity:

$$L(U_t, \pi_t) = U_t + \theta\frac{(\pi_t)^2}{2} \tag{1}$$

where θ is the relative weight the electorate puts on inflation deviations relative to unemployment and where (for simplicity of exposition) it is assumed that the representative voter's most preferred rate of inflation is zero.

One may then posit a retrospective voting function for an election at the end of period *t*, of the form:

$$V_t = \Psi\left(\sum_{s=0}^{T-1}\gamma(s)L(U_{t-s}, \pi_{t-s})\right) \tag{2}$$

yielding the number of votes V_t for the incumbent as a decreasing function of loss from economic outcomes $(\Psi' < 0)$. The exogenous length of time between elections is *T* periods, and $\gamma(s)$ is the weight voters put on a loss *s* periods in the past. $\gamma(s)$ is assumed to be decreasing in *s*, that is, past economic outcomes have a smaller effect on votes at *t* the further in the past they are. If $\gamma(s)$ is rapidly decreasing in *s*, very recent events are weighted most heavily. In the extreme, if $\gamma(s) = 0$ for $s > 0$, then only economic outcomes in the year of the election affect voting. The electoral mechanism is not made more

specific. One could add a stochastic element to allow for the possibility of an incumbent losing the election.

In the Nordhaus model, the structure of the economy is summarized by an expectations-augmented Phillips curve relating the difference between the actual and the natural rates of unemployment U_t^N to the difference between actual and expected inflation π_t^e:

$$U_t = U_t^N - (\pi_t - \pi_t^e) \tag{3}$$

To close the model one must specify the formation of expectations. Crucial to the main results of the above models is some form of backward-looking expectations, so that inflationary policy in an election period is not fully anticipated and can therefore lower the unemployment rate. A standard formulation of adaptive determination of the expected rate of inflation:

$$\pi_t^e = \pi_{t-1} + \alpha(\pi_{t-1}^e - \pi_{t-1}) \tag{4}$$

where α is a coefficient between 0 and 1 representing the speed with which expected inflation adapts to past expectational errors. This may be solved to yield π_t^e as a weighted declining sum of past inflation rates.

This four-equation system may then be solved for unemployment and inflation over the electoral cycle. When voters have ‘short memories’ (γ (s) small for $s > 0$) a political business cycle will emerge if the incumbent wants to maximize his probability of re-election. In the period immediately after the election the government engineers a recession via contractionary monetary policy to bring down inflationary expectations. The incumbent keeps economic activity low to keep expected inflation low until the period immediately before the next election, so that a given rate of economic expansion (induced by a monetary surprise) can be obtained at a relatively low rate of inflation. The government then stimulates the economy via expansionary monetary policy, unemployment falling due to high unanticipated money growth. The levels of monetary expansion and unemployment are those which maximize voter satisfaction in the election period. In the next election cycle the same behaviour is repeated, with contractionary monetary policy to bring down inflation expectations. Hence, the possibility of influencing the probability of re-election, combined with the structure of the economy, yields a cycle in economic activity which would not be present with a planner with an infinite horizon. The political cycle thus induces a cycle in economic activity and inflation.

Though these models capture the incentive for opportunistic policymakers to manipulate policy and the macroeconomic cycle that may result, a number of conceptual and empirical objections may be raised.

First, incumbents running for re-election do not control monetary policy in countries with independent central banks. However, there is evidence that nominally independent central banks often accommodate the executive branch's pressures for monetary policy during election years in order to prevent sharp movements in interest rates (see, for example, Woolley, 1984, for evidence for the United States). Hence, politically motivated monetary policy in an election year may be a good approximation to reality. Second, one may question whether voters are really as unsophisticated as the basic models assume, both in the way they form expectations of inflation and in the way they assess government performance. Voters realize that ‘election-year economics’ may be used to win their votes and hence may be sceptical of an economic upturn in the months before an election. More formally, their expectations of inflation should take the possibility of an election-year monetary expansion into account (which would then nullify its effects since it is no longer a surprise). An intermediate view is that voters have less-than-perfect information about the causes of economic fluctuations and take good economic performance as indicating incumbent competence. Hence voting for the incumbent when times are good is consistent with rationality when voters have imperfect information. This has been argued by Nordhaus (1989) and has been formalized using signalling models, as discussed below.

Partisan models

In partisan models, cycles are induced by differences among parties in their ideology and their economic goals. The basic partisan model is due to Hibbs (1977), based on different preferences over inflation and unemployment across parties. One replaces the voters’ loss function (1) with one representing the preferences of a party *j*, for example,

$$L^j(U_t, \pi_t) = \frac{(U_t - \tilde{U}^j)^2}{2} + \theta^j \frac{(\pi_t - \tilde{\pi}^j)^2}{2} \tag{5}$$

where $\tilde{\pi}^j$ is party *j*'s target rate of inflation, $\tilde{U}^j$ is party *j*'s target unemployment rate, and θ^j is the weight party *j* puts on deviations of inflation from target inflation relative to deviations of unemployment from target. The two parties, say a right-wing party *R* and a left-wing party *L*, are characterized, for example, by $\tilde{U}^L \leq \tilde{U}^R$, $\theta^L \leq \theta^R$, and $\tilde{\pi}^L \geq \tilde{\pi}^L$. Thus, the left-wing party will pursue a more expansionary monetary policy throughout its term. Using the same specification of the relation between unemployment and inflation as in (3) and a similar specification of backward-looking expectations (4), one may derive a cycle in which the level of economic activity and inflation varies with the ideology of the incumbent.

Rational voters

Early models in both strands of the literature were often criticized in their modelling of expectations, since the backward-looking nature of expectations was crucial for some of the results. Hence, in both strands the focus has shifted to models in which voters form their expectations rationally, with the question being whether a political budget cycle will still exist with rational, forward-looking voters.

In the context of an opportunistic political budget cycle, the key argument is that some characteristic of policymakers is unobserved, and the voters' inference problem over an incumbent's 'type' will imply it is optimal to vote more heavily for the incumbent when economic outcomes are favourable. A leading unobserved characteristic is the incumbent's 'competence'. More competent policymakers produce better outcomes, and competence has some persistence over time. Therefore, good outcomes in the time period before the election may signal high competence of the incumbent (relative to a challenger who cannot signal), which is expected to persist after the election. Hence, when competence cannot be observed directly, it may be optimal for voters to vote more heavily for the incumbent if times are good.

This argument may be formalized in an imperfect information framework. The first formal models concerned political budget cycles in work by Rogoff (for example, Rogoff, 1990). Persson and Tabellini (1990) and Lohmann (1998) present similar models of unobserved policymaker ability as applied to cycles in economic activity. High economic activity before an election signals a high-ability incumbent, that is, higher than the average expected ability of the challenger. Since ability has a persistent component, voters expect better economic performance from the incumbent than from the challenger after the election as well, and hence vote for him.

Alesina (1987) introduces rational expectations into the original partisan model of Hibbs, so that fluctuations in inflation and unemployment are driven by partisan differences combined with uncertainty about election outcomes. Close elections imply the sort of fluctuations Hibbs found, but because expansionary monetary policy by a left-wing policymaker (for example) is not fully anticipated before an election and therefore will lead to a fall in unemployment after the election. A key difference from the Hibbs model is that any effect on unemployment will no longer be present after inflation expectations are adjusted. Hence, the effects on unemployment will be concentrated early in a leader's term of office and disappear in the latter part of the term once the leader's preferences are known.

Empirical testing

The existence of opportunistic political business cycles has been subject to extensive empirical testing. There are two key questions: are election years characterized by economic expansions? Do voters respond to 'good times'?

The standard test for the existence of a political cycle is to run an autoregression of an economic performance measure on itself, a small set of economic variables, and political dummies, that is, a regression:

$$Y_t = \sum_{i=1}^{s} a_i Y_{t-i} + b_0 + \sum_{j=1}^{k} b_j X_{jt} + d\,PDUM_t + \varepsilon_t \tag{6}$$

where Y is an outcome variable such as output growth, the X_j are control variables, and $PDUM$ is a political dummy variable (or set of variables) meant to represent a given political model. The autoregressive specification for Y_t is adopted as a parsimonious representation of the time series behaviour of Y, instead of using a structural model. The hypothesis that output growth, for example, is higher in election years would be represented by setting $PDUM_t$ equal to 1 in election years and zero otherwise, and testing whether the coefficient d is statistically significant.

The evidence for a political cycle in outcomes is quite mixed, with most studies finding little evidence of opportunistic political cycles in developed countries. Much of this evidence is summarized in Alesina, Roubini and Cohen (1997) and Drazen (2000).

The evidence on voter response to economic conditions is also mixed. Generally, the effect of growth on re-election probabilities was found to be insignificant in most cross-section studies in developed countries (see Brender and Drazen, 2005, for a summary). The United States seems to be an exception to these findings. The most influential paper on voter response in the United States is probably that of Fair (1978), who found that an increase in real economic activity in the year of the election, as measured either by the change in real per capita GNP or the change in unemployment in the election year, has a strong positive effect on the incumbent's vote total in US presidential elections. Alesina and Rosenthal (1995) find similar results.

Brender and Drazen (2005) confirm the insignificant effect of growth on re-election probabilities in developed countries in a large cross-section study of a sample of 74 democracies over the period 1960 to 2003. In contrast, they find that in less developed countries higher growth in real GDP has a positive and statistically significant effect on the probability of re-election. They then remove from the overall growth rate the part that voters might attribute to global developments and find that in the less developed countries it is the component of growth associated with domestic influences that accounts for the highly significant effect of growth on re-election, while the part attributable to global economic growth has no statistically significant effect on the probability of re-election. In the developed countries they find that neither the effect of global growth nor the effect of domestically induced growth is statistically significant.

There has been less empirical testing of the partisan political business cycle. The striking empirical regularity in the United States since the Second World War is that economic activity is substantially higher under Democrats than under Republicans in the first part of their four-year terms, but more similar in the second part of their terms, consistent with the Alesina model. However, Faust and Irons (1999) argue that the data do not give strong support to any partisan model. For the OECD, Alesina, Roubini, and Cohen (1997) find supporting evidence for the rational partisan model in a number of countries.

Overall, the focus of both theoretical and empirical research has shifted to political budget cycles, in large part due to the weak empirical evidence for the existence of an opportunistic political business cycle in many countries, combined with the widespread view that, nonetheless, election year manipulation of some sort is a common phenomenon.

See Also

- political budget cycles

Bibliography

Alesina, A. 1987. Macroeconomic policy in a two-party system as a repeated game. *Quarterly Journal of Economics* 102, 651–78.

Alesina, A. and Rosenthal, H. 1995. *Partisan Politics, Divided Government, and the Economy*. Cambridge: Cambridge University Press.

Alesina, A., Roubini, N. and Cohen, G. 1997. *Political Cycles and the Macroeconomy*. Cambridge, MA: MIT Press.

Brender, A. and Drazen, A. 2005. How do budget deficits and economic growth affect reelection prospects? Evidence from a large cross-section of countries. Working Paper No.11862. Cambridge, MA: NBER.

Drazen, A. 2000. *Political Economy in Macroeconomics*. Princeton, NJ: Princeton University Press.

Fair, R. 1978. The effect of economic events on votes for president. *Review of Economics and Statistics* 60, 159–72.

Faust, J. and Irons, J. 1999. Money, politics, and the post-war business cycle. *Journal of Monetary Economics* 43, 61–89.

Hibbs, D. 1977. Political parties and macroeconomic policy. *American Political Science Review* 71, 1467–87.

Lohmann, S. 1998. Rationalizing the political business cycle: a workhorse model. *Economics and Politics* 10, 1–17.

Nordhaus, W. 1975. The political business cycle. *Review of Economic Studies* 42, 169–90.

Nordhaus, W. 1989. Alternative approaches to the political business cycle. *Brookings Papers in Economic Activity* 1989(2), 1–49.

Persson, T. and Tabellini, G. 1990. *Macroeconomic Policy, Credibility, and Politics*. London: Harwood.

Rogoff, K. 1990. Equilibrium political budget cycles. *American Economic Review* 80, 21–36.

Tufte, E. 1978. *Political Control of the Economy*. Princeton, NJ: Princeton University Press.

Woolley, J. 1984. *Monetary Politics*. Cambridge: Cambridge University Press.

How to cite this article

Drazen, Allan. "political business cycles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000112> doi:10.1057/9780230226203.1298

The New Palgrave Dictionary of Economics Online

political competition

David Austen-Smith

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article is limited to interaction between candidates and voters and examines the cases of two-candidate competition and multiple candidate competition. It employs the spatial model of elections introduced to study single-issue politics and generalized to study multiple-issue politics in order to explain the alternatives strategically offered to voters by candidates or parties competing for electoral office.

Keywords

Black, D.; Downs, A.; Duverger's Law; electoral competition; First Welfare Theorem; Hotelling, H.; median voter theorem; mixed strategy equilibrium; myopia; Nash equilibrium; political competition; rational choice models; voting

Article

In its most general form, political competition concerns the struggle of ideas for organizing societies. This article, however, focuses explicitly on one concrete manifestation of this struggle, namely, electoral competition. Any convincing and general explanation of electoral competition must account for the role of money in campaigns, for the behaviour of interest groups and the implications of party organization. This article addresses only the interaction between candidates and voters. Although not the only framework for studying the topic, the spatial model of elections introduced by Hotelling (1921) and Downs (1957) for single-issue politics and generalized by Davis and Hinich (1966; 1967) to multiple-issue politics, is surely the most widely used. The principal goal of the theory is to explain the alternatives strategically offered to voters by candidates or parties competing for electoral office.

Two-candidate competition

A benchmark model

There are two candidates *A*,*B*, and a large finite set of voters $N = \{1, \ldots, n\}$. The policy space is a convex and compact set $X \subset R^k$, where $x \in X$ is a typical feasible policy. Each voter $i \in N$ has policy preferences on *X* representable by a continuous and strictly quasi-concave utility function $u_i: X \to R$; let $\mathbf{u} = (u_1, \ldots, u_n)$ denote the preference profile over *X*. Candidates too have preferences although they need not be defined directly on the policy space, *X*; candidates, for example, might plausibly be more interested in winning office than in policy per se, or in some combination of winning and the policy eventually implemented, irrespective of who wins. On the assumption of complete information on the part of voters regarding candidates' motivations and policy platforms, and on the part of candidates regarding voters' preferences, however, introducing policy motivations for candidates leads to no essential change in the predictions of the model with purely office-motivated candidates (Calvert, 1985; Duggan and Fey, 2005). Some implications of assuming that candidates are policy motivated when there is some uncertainty about payoffs are considered later; for now, suppose that each candidate is motivated solely by the desire to win office.

The election is for a single office and is determined by a plurality rule. On the assumption of no abstention, a voting strategy for any citizen $i \in N$ is a mapping $v_i: X^2 \to [0, 1]$, where $v_i(a, b)$ is the probability that *i* votes for candidate *A* when *A* chooses electoral platform $a \in X$ and *B* chooses a platform $b \in X$. Given a profile of vote strategies $\mathbf{v} = (v_i)_{i \in N}$, $V_j(a, b|\mathbf{v}) \in [0, n]$ is the expected number of votes cast for candidate *j*. Let $\Pi(a, b|\mathbf{v}) = V_A(a, b|\mathbf{v}) - V_B(a, b|\mathbf{v})$ denote *A*'s expected plurality, so $-\Pi(a, b|\mathbf{v})$ is *B*'s expected plurality. Then candidate *j*'s payoff under plurality rule is 1 if her realized plurality is strictly positive, –1 if it is strictly negative, and zero otherwise.

The strategy space for each candidate is *X*. Maximizing *j*'s payoffs in this setting is equivalent to maximizing *j*'s expected plurality. Thus, *A* chooses $a \in X$ to maximize $\Pi(a, b|\mathbf{v})$ and *B* chooses $b \in X$ to minimize, $\Pi(a, b|\mathbf{v})$. Under the assumptions of no abstention and two candidates, maximizing $\Pi(a, b|\mathbf{v})$ is equivalent to maximizing $V_A(a, b|\mathbf{v})$. Later, I consider some implications of admitting abstention and multiple candidates where the equivalence fails. An equilibrium to the game is a vector of undominated strategies $(a^*, b^*, \mathbf{v}^*)$ such that each voter $i \in N$ is maximizing u_i conditional on (a^*, b^*) and the voting strategies of all $j \in N \setminus \{i\}$ and, given $\mathbf{v}^*$,

$$\Pi(a^*, y|\mathbf{v}^*) \geq \Pi(a^*, b^*|\mathbf{v}^*) \geq \Pi(y, b^*|\mathbf{v}^*)$$

for all $y \in X$.

Existence of equilibrium in the model when candidates use pure strategies is a problem. The majority core, that is, the set of alternatives $x \in X$ such that no alternative is strictly preferred by a majority to *x*, is guaranteed to be non-empty only when the dimensionality of the policy space, *k*, is 1 (Plott, 1967; McKelvey, 1979; Schofield, 1983) and it is not hard to see that the set of equilibria in pure candidate strategies coincides with the majority core. The most familiar example of this coincidence is the median

voter theorem for $X \subseteq R$ (Downs, 1957; Black, 1958), predicting candidate convergence on the median most preferred policy. On the other hand, if a policy space of any finite dimension is approximated with a finite grid, irrespective of how fine the grid might be, the classical Nash equilibrium existence theorem implies the existence of an equilibrium in mixed candidate strategies. Furthermore, in the finite case with no majority indifference, the mixed strategy equilibrium is unique and symmetric (Laffond, Laslier and Le Breton, 1993): both candidates adopt the same mixed strategy and thus, *ex ante*, the candidates are equivalent from a policy perspective, just as they are under the median voter theorem.

The difficulty in proving a general mixed strategy equilibrium existence result for the spatial voting model with a continuum of alternatives lies in the absence of sufficient continuity in the mapping that connects pairs of policy positions to vote shares: a small unilateral change in one candidate's position can result in the candidate's vote share changing from less to more than one-half of the electorate, thus inducing a discrete jump in her payoff. But these discontinuities often arise as a result of the presumption that indifferent individuals in the spatial model necessarily vote for each candidate with probability one-half. If this assumption is relaxed and individuals are restricted only to symmetric voting strategies, thus allowing the probability of indifferent individuals voting for one or other candidate to be sensitive to the platforms offered, then existence of a mixed strategy equilibrium is guaranteed (Duggan and Jackson, 2004). Characterizing mixed candidate strategies, however, is not easy. McKelvey (1986) and Banks, Duggan and Le Breton (2002) provide some insight by showing that the support of any mixed strategy equilibrium (essentially) lies within the closure of the uncovered set: say that an alternative *x* is covered by an alternative *y* if *x* is strictly majority preferred to *y* and, further, that any alternative *z* that defeats *x* also defeats *y*; the uncovered set is then the set of alternatives that are not covered (Miller, 1980). The uncovered set generally exists in the spatial model and, moreover, if a sequence of (continuous and strictly quasi-concave) preference profiles converges uniformly to a profile $\mathbf{u}^*$ at which the majority core is non-empty, then (loosely speaking) the associated sequence of uncovered sets converges to the core at $\mathbf{u}^*$. Thus, for any profile **u** ‘close’ to a profile $\mathbf{u}^*$ supporting pure strategy equilibria to the election, the realized policy platforms offered to the electorate at **u** are ‘close’ to the (pure strategy) equilibrium policies offered at $\mathbf{u}^*$.

Results on the uncovered set notwithstanding, a convincing interpretation of mixed candidate strategies in electoral competition is elusive. A satisfactory theory of elections therefore seems to require more structure than that presumed in the benchmark model. Important approaches in this regard are to introduce various informational limitations on the part of candidates and voters, to allow voter abstention and to admit the possibility of policy-motivated candidates.

Candidate uncertainty and abstention

Candidates for electoral office clearly do not know the details of every individual's preferences or voting criteria. Adding idiosyncratic non-policy characteristics to voters’ decisions (for example, their attitude towards the social background of the candidates) and assuming candidates know at best the distribution from which these characteristics are drawn can induce sufficient continuity in candidates’ assessments of how policy positions map into vote shares to admit a general equilibrium existence result in pure strategies. Specifically, let $p_i(u_i(a), u_i(b))$ be the probability that voter *i* votes for candidate *A* given platforms $a, b \in X$. Then A's expected plurality, given (a, b) and no abstention, is

$$\Pi(a, b) = 2\sum_{i \in N} p_i(u_i(a), u_i(b)) - n.$$

If we assume that p_i is strictly concave increasing (respectively, convex decreasing) in $u_i(a)$ (respectively, $u_i(b)$) and that u_i is strictly concave, there exists a unique equilibrium in pure candidate strategies and the equilibrium platforms coincide (Enelow and Hinich, 1982; Coughlin, 1992). More importantly, it turns out that the policy on which both candidates converge in equilibrium maximizes weighted aggregate utility (Couglin and Nitzan, 1981; Banks and Duggan, 2005).

On the other hand, if candidates $j \in \{A,B\}$ are policy-motivated and seek to maximize their expected utility defined by

$$E[u_j|a, b] = \Pr[A\text{wins}|a, b]\, u_j(a) + \Pr[B\text{wins}|a, b]\, u_j(b),$$

then, under suitable regularity conditions on the probabilities of winning, candidate convergence is not assured (Wittman, 1977; Calvert, 1985). Unfortunately, such regularity conditions are unlikely: for instance, if there is no abstention, then

$$\Pr[A\text{wins}|a, b] = \sum_{\substack{M \subseteq N \\ |M| > n/2}} \left[\prod_{i \in M} p_i(u_i(a), u_i(b)) \times \prod_{j \in N \setminus M} [1 - p_j(u_j(a), u_j(b))] \right]$$

which is not at all nicely behaved.

A conceptual difficulty with the probabilistic voting approach is that it seems ad hoc. Although idiosyncratic components to voters' decision calculus are plausible, the assumptions required on the distributions of such idiosyncrasies to insure existence – that they are uncorrelated with individuals' policy preferences and induce the appropriate concavity properties – are stringent. In particular, if the candidates' uncertainty regards voter policy preferences rather than some non-policy idiosyncrasies, then candidate objective functions again become discontinuous, leading to a breakdown in equilibrium (Ball, 1999). An alternative approach in the same spirit is to say nothing about voter idiosyncrasies at the individual level but rather to assume that the winner depends on policy-oriented voting over platforms and the true state of the world, known only up to its distribution at the time of the election (Roemer, 2001). The interpretation here is that the realized preference profile of voter preferences is conditional

on the state. For example, if those who live closest to voting stations are the most likely to vote in bad weather, then the effective distribution of voter preferences is conditional on the weather. To insure electoral equilibria then requires imposing particular conditions directly on the distribution of states.
Voting is not costless and this fact gives rise to a problem for rational choice theoretic models of participation in large elections: given that the probability of being pivotal is negligible in large elections, the net benefit of voting is negative (Feddersen, 2004, provides a recent survey of the literature on turnout). However, assuming that voting costs and (possibly) policy preferences are private information to individuals, with only their joint distribution being common knowledge, induces uncertainty on the part of both voters and candidates sufficient to yield existence of equilibrium in a model with all agents being fully instrumentally rational (Ledyard, 1984; Myerson, 2000). The idea is to note that, for any fixed and distinct pair of platforms $(0,b)\in X^2$, voters are confronted with a strategic decision whether to abstain or to vote for their favoured candidate. In a voter equilibrium relative to (a, b), the probability of being pivotal induces a level of expected turnout which in turn justifies the probability of being pivotal. Candidates then choose their platforms to maximize their respective expected pluralities, recognizing the implications for turnout from any pair of platforms. In this setting, under various assumptions on the joint distribution of preferences and voting costs, there exists a unique equilibrium when voter preferences are concave and, in equilibrium, both candidates adopt the policy that maximizes aggregate utility and turnout is zero. In effect, this model produces an efficiency result for electoral competition analogous to the First Welfare Theorem for competitive markets.

Voter uncertainty and commitment

Although voters are uncertain about the behaviour of other voters in the costly voting model discussed above, they are fully informed about the candidates' platforms and the policy that the winner will implement. Developing an understanding of how voter uncertainties about candidate policies and intentions affect electoral competition in multidimensional policy spaces has proved very difficult. Instead, most of the insights to date derive from one-dimensional models where, under complete information, the median voter theorem applies.
Suppose candidates for office have policy preferences over a one-dimensional policy space just like voters, but that these preferences are unknown to the electorate. Specifically, a candidate's type $t\in R$ parameterizes the candidate's preferences over policies and is private information; voters know only the distribution from which t is drawn. In a two-candidate election, the candidates, knowing their own types, simultaneously choose policy platforms on which to campaign; voters observe the platforms, update their beliefs about the likely types of the candidates, and vote accordingly. The winner is then free to implement any policy she chooses as government policy; in particular, she has an incentive to implement her ideal point. Assume that there is a cost to implementing any policy other than that on which the winner is elected and, further, that this cost increases with the distance between the electoral platform and the implemented policy. Then Banks (1990) shows that, in any (appropriately refined) sequential equilibrium, relatively extreme candidate types, that is, those with preferences far from those of the median voter, offer revealing platforms but there is pooling on the median voter's most preferred outcome by an interval of relatively moderate types.
The model therefore predicts candidate divergence in equilibrium precisely in competitions in which at

least one extremist is running for office; indeed, the observed equilibrium platforms in such instances can be far from the median. More importantly, as the cost of implementing any policy distinct from that on which the election was won goes to zero, the interval of types pooling on the median voter's preferred policy expands to include the entire type space. Conversely, if the cost goes to infinity, the interval of types pooling on the median voter's most preferred policy contracts to the median's ideal point. *Inter alia*, this result highlights the central role of policy commitment in models of electoral competition. If we leave aside concerns with legislative coalition formation and so forth, elected candidates are free to implement any policy they choose once in office. In the benchmark model with full information and purely office-motivated candidates, there is no reason for an elected official to implement anything other than her electoral platform. This is not so if candidates' motivations or preferences are unclear to voters. Unless candidates can make credible commitments to implement their electoral platforms conditional on being elected, electoral platforms are at best signals of candidate intentions; and there is no obvious way for candidates to make such commitments.

If candidates are assumed to have adopted distinct platforms and members of the electorate have private and (possibly) asymmetric information about which candidate is most likely to be their *ex post* preferred candidate, then Feddersen and Pesendorfer (1996; 1997) prove that, as the electorate becomes arbitrarily large, the winning candidate is almost surely the winning candidate under complete information. This is a remarkable result and suggests that questions of commitment and voter uncertainty need not be problematic in large elections. On the other hand, allowing for some strategic platform choice by candidates attenuates the result (Razin, 2003; Gul and Pesendorfer, 2004).

Dynamics

Parties, if not candidates, are often long-lived, and winning platforms in one election may be empirically hard to change for a following election. Assume the benchmark model of elections above is iterated over an uncountably infinite number sequence of elections t=1,2,… with all voters and the two candidates being myopic. Assume in addition that the electoral platform on which the period-t incumbent won office is necessarily the platform on which the incumbent contests election $t+1$. The opposition candidate's choice of platform in period t+1, however, is unconstrained. Then the non-existence results for equilibria discussed above imply that the two candidates alternate in office over time. More interesting is the fact that the winning platform converges to a neighbourhood of the minmax set, a centrally located set of alternatives that coincides with the majority core when the latter is non-empty: for any alternative x, let $\gamma(x)$ denote the maximal number of votes that any alternative policy y could attract in a pairwise vote against x (equilibrium non-existence implies $\gamma(x) > n/2$ for all $x \in X$); then the minmax set is the set of policies $x \in X$ for which $\gamma(x)$ is minimal (Kramer, 1977).

The myopia assumption and the constraint on an incumbent's policy choice underlying the preceding result are not very satisfactory. A strategically richer framework is proposed by Banks and Duggan (2002). In their set-up, all individuals are far-sighted and instrumentally rational. Individual preferences are private information up to the common knowledge that utilities are continuous and strictly concave over the multidimensional policy space. The $t = 1$ incumbent is chosen randomly and implements a policy $x_1 \in X$; the incumbent's name and the policy choice are observed by all voters. In the period $t = 2$ election, the incumbent faces a randomly chosen challenger whose name is observed by voters. Voters

then vote for one or other of the competitors. The plurality winner becomes the incumbent and (freely) implements a policy $x_2 \in X$; there is no restriction on the $t = 1$ incumbent's choice should he or she win a second term in office. This process then repeats for $t > 2$. The authors prove the existence of (stationary subgame perfect) equilibria in which voters employ a simple cut-off rule, that is, vote for the incumbent if her previous policy choice at least achieves an endogenously determined reservation utility value. The main result here is that there is eventual policy persistence in that the distribution of policy outcomes converges to a fixed platform as *t* goes to infinity. However, while this platform is necessarily centrally located because it must be acceptable to a majority of the population, there is no assurance in the multidimensional policy space that it is uniquely defined.

Multiple candidate competition

Fixed number of candidates

Political competitions with exactly two candidates are relatively unusual. In general there are at least two candidates in any election, and the possibility of multiple electoral competitors raises a variety of issues that are largely irrelevant when considering elections with two given candidates seeking a single office. For example, questions of candidate participation, or the number of candidates, in an election are finessed by assuming a given two-candidate contest, and proportional representation schemes for determining electoral success are irrelevant when only one office is at stake. As a result, important questions about the relative merits of various electoral rules cannot be addressed. Similarly, if the election is for a legislature and legislative policy decisions require, as is typical, majority support of the elected legislators, then rational voters and candidates make their respective electoral decisions taking account of the subsequent legislative bargaining and committee decision-making (Austen-Smith and Banks, 1988; Baron and Diermeier, 2001). Addressing these issues, among others, requires admitting a more general class of electoral rule for multi-candidate competition and providing a more complex analysis of voter behaviour.

There is fixed set of candidates $M = \{1, \ldots, m\}$ who compete for $l \geq 1$ elected offices, $m < l$, in an election decided by a normalized rank scoring rule. A normalized rank scoring rule for a fixed number of candidates *m* is defined by a vector $s = (s_1, \ldots, s_m)$ such that $1 = s_1 \geq s_2 \geq \ldots \geq s_{m-1} \geq s_m = 0$ and a mapping that assigns a set of $l \geq 1$ winners for any profile of permissible ballots, where a permissible ballot is any permutation of the vector $(1, s_2, \ldots, s_{m-1}, 0)$. The normalization here refers to the joint restrictions $s_1 = 1$ and $s_2 = 0$ and is purely a convenience; the defining characteristics of rank scoring rules are that $s_t \geq s_{t+1}$ for all $t = 1, \ldots, m-1$ and $s_1 > s_m$. Not all rules of interest are rank scoring rules. For instance, approval voting is a scoring rule but not a rank scoring rule; under approval voting, the restriction that $s_1 > s_m$ is not required so voters may vote for, or approve of, any and all candidates should they so choose. Examples of rank scoring rules include single non-transferable voting ($s_t = 0$, all $t > 1$), a generalization of the simple plurality rule (where $l = 1$); single negative voting ($s_t = 1$, all $t < m$); and the Borda rule ($s_t = [m - t] / [m - 1]$, all $t = 2, \ldots, m - 1$). In each of these examples, the *l* top-scoring candidates are the winners, with ties being broken randomly.

An individual votes sincerely under a scoring rule if she always assigns higher scores to more preferred candidates. If we assume a one-dimensional policy space, say *X*=[0,1], sincere voting and that

candidates maximize expected plurality, now defined to be the difference between their vote share (aggregate score) and the maximum vote share among their competitors, Cox (1987; 1990) establishes a connection between the existence of equilibria in which all candidates converge and the average score $C(s; m) = \Sigma_{j=1}^{j=m} s_j / m$, the Cox threshold (Myerson, 1999). Specifically, suppose there is a continuum of voters with interior most-preferred policies (or ideal points) distributed over a policy space according to a strictly increasing and differentiable cdf, *F* and let α_t be the quantile implicitly defined by $\int_0^{\alpha_t} dF(x) = t$, $t \in [0, 1]$; then there exists an equilibrium in which all candidates adopt the platform α_t if and only if $[1 - C(s; m)] \leq t \leq C(s; m)$.

The Cox threshold for single non-transferable voting is 1/*m*, it is 1/2 for the Borda rule and it is (*m*–1)/*m* for single negative voting. The median voter theorem is therefore a direct implication of the theorem. Moreover, writing the Cox threshold equivalently as $C(s; m) \equiv \bar{s} = 1 - (s_1 - \bar{s}) / (s_1 - s_m)$ makes clear that the threshold is decreasing in the extent to which the scoring rule provides an incentive to be ranked first rather than average relative to being ranked first rather than last. Hence, Cox's result implies that, the greater this incentive is, the greater is the incentive for candidates to differentiate their platforms.

While sincere voting is the unique undominated strategy for individuals when there are only two candidates, it is not obviously rational when there are the more than two candidates. Indeed, the fact that people typically try to avoid ‘wasting’ their vote by voting for an almost sure loser suggests that strategically rational voting is substantively significant. An important observation in this respect generalizes Duverger's Law, namely, that single non-transferable voting for a single office promotes only two-candidate competition. Assuming there are $m > 2$ candidates with fixed and distinct platforms competing for $l < m$ offices, Cox (1994) proves that voting equilibria in undominated strategies have the following form: the top *l* candidates receive identical (strictly positive) vote shares, with all other candidates receiving either no votes or the same vote share as the candidate with the $(l+1)^{th}$ highest vote share, which in turn is less than or equal to that of the candidates with the *l* highest vote shares. Moreover, although not proved formally, equilibria in which the vote-share of the $(l+t)^{th}$ most successful candidate, $t > 1$, is positive are almost surely not robust, as any shock will move votes away from this candidate to one of the $l + 1$ top-ranked competitors.

Candidate entry

The last result on Duverger's Law raises a question as to why candidates who are almost surely going to lose enter the electoral competition at all. In view of the canonical model of electoral competition, one natural starting point for understanding candidate entry is to fix a set of potential candidates, assume each candidate is concerned only with winning office, and ask what platforms such candidates would adopt if entry is costly and voters are strategically rational rather presumed than to vote sincerely under all circumstances. Then purely office-motivated candidates have no incentive to enter an electoral contest if they are sure to lose. On the assumption that voters have strictly concave preferences over a one-dimensional issue space, it can be shown that the number of entrants is constrained only by the ratio of benefits from holding office to the costs of entry and that all entrants adopt the median voter's most

preferred policy as their platform (Feddersen, Sened and Wright, 1990). This result is in stark contrast to the implications discussed earlier of assuming a fixed number of competing candidates and sincere voting. As such, the value of this benchmark entry model derives from its appeal less as a reflection of any empirical reality (which is limited at best) and more as an analytical robustness check on models of multi-candidate elections that presume fixed numbers of candidates and sincere voting. While the latter certainly illuminate some incentives facing strategic candidates with several competitors, equilibrium predictions supported by the models appear fragile.

An alternative, and more plausible, approach to presuming a fixed pool of potential candidates is to recognize that candidates are voters too and to suppose the pool is exactly the electorate itself: every individual voter is eligible to run for office should he or she so choose (Osborne and Slivinski, 1996; Besley and Coate, 1997). In such a model candidate preferences derive directly from individual policy preferences, implying that the entry decision and the decision on which platform to run for office conditional on entering an election are equivalent. At least in the static setting without commitment, citizens who enter the race implement their respective ideal policies should they win office; it immediately follows that policy convergence is impossible if multiple citizens declare a candidacy, unless all such individuals share the identical ideal point. An equilibrium in this framework is a mutually consistent list of best-response decisions for individuals regarding whether or not to enter the election and, conditional on the realized slate of entrants, on how to vote. Given complete information on individuals' preferences and the inability of entrants to commit to implement any policy other than their ideal points, establishing a fairly general existence theorem for multidimensional policy spaces is straightforward. Furthermore, since candidates are also citizens with policy preferences, there are circumstances under which it is more rational for an individual to contest a costly election than when individuals are concerned only with winning office. Specifically, despite being sure that he or she will not win, an individual might enter an electoral race to affect the expected policy outcome by implicitly blocking the entry or exit of other potential or declared candidates.

The citizen-candidate perspective on political competition is quite intuitive and simultaneously gives rise to a theory of candidate entry and a theory of candidate preferences. On the other hand, the implication that, other things being equal, declared candidates are locked into implementing their ideal points should they win is restrictive. As a matter of fact, candidates do credibly adjust their policy positions and, even should they not do so, a theory of platform selection predicated exclusively on an individual's exogenously given policy preferences effectively reduces any account of collective policy outcomes to an account of which particular individuals choose to run for office. Although understanding who chooses to seek election is clearly important to a full theory of electoral politics, an account of collective choice based entirely on such a foundation is less compelling.

A different approach within the spirit of the citizen-candidate theory is to eschew candidates and parties altogether. Instead, the set of 'potential candidates' is assumed to be the full set of feasible policy outcomes, with each individual permitted to vote for any one policy or not at all. Assuming away candidates as agents or, equivalently, assuming there is a candidate for every possible alternative in the policy space, seems, at least prima facie, to be unreasonable for any sort of theory of electoral competition with a large number of voters. In so far as the focus of a model of elections is on understanding the behaviour of particular agents with given objectives (for example, maximizing the probability of winning), or on understanding the behavior of historically established political parties in a constrained (for example, two-party) environment, such a presumption is justified. However, if the focus

is on understanding the deeper implications of using a given voting scheme to determine collective policy choices, then dispensing at the outset with the intermediary steps involved in candidates choosing platforms is reasonable. And in this respect the implications of plurality rule under costly voting and abstention are subtle and striking: in every voting equilibrium with risk-averse (and strategically rational) voters and a multidimensional policy space, exactly two divergent policy positions receive any votes at all (Feddersen, 1992). Despite the absence of candidates or parties, strategic and costly voting under plurality rule leads to equilibria with two distinct platforms and positive turnout. The precise location of any given pair of equilibrium platforms, however, is in general indeterminate.

Concluding comments

Despite the large literature on the spatial theory of elections, our understanding of political competition is still relatively primitive in many respects. There is, for example, much to be learned from dynamic models, and a compelling theory of candidate entry has yet to be developed. Similarly, a tractable theory of strategic voting in large populations, an essential component of a satisfactory theory of political competition, remains elusive.

See Also

- collective rationality
- political institutions, economic approaches to

Bibliography

Austen-Smith, D. and Banks, J. 1988. Elections, coalitions and legislative outcomes. *American Political Science Review* 82, 405–22.

Ball, R. 1999. Discontinuity and nonexistence of equilibrium in the probabilistic spatial voting model. *Social Choice and Welfare* 16, 533–56.

Banks, J. 1990. A model of electoral competition with incomplete information. *Journal of Economic Theory* 50, 309–25.

Banks, J. and Duggan, J. 2002. A multidimensional model of repeated elections. Working paper. Department of Political Science, University of Rochester.

Banks, J. and Duggan, J. 2005. Probabilistic voting in the spatial model of elections: the theory of office-motivated candidates. In *Social Choice and Strategic Decisions: Essays in Honor of Jeffrey S. Banks*, ed. D. Austen-Smith and J. Duggan. Berlin: Springer.

Banks, J., Duggan, J. and Le Breton, M. 2002. Bounds for mixed strategy equilibria and the spatial

model of elections. *Journal of Economic Theory* 103, 88–105.

Baron, D. and Diermeier, D. 2001. Elections, governments and parliaments in proportional representation systems. *Quarterly Journal of Economics* 116, 933–67.

Besley, T. and Coate, S. 1997. An economic model of representative democracy. *Quarterly Journal of Economics* 112, 85–114.

Black, D. 1958. *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.

Calvert, R. 1985. Robustness of the multidimensional voting model: candidate motivations, uncertainty and convergence. *American Journal of Political Science* 29, 69–95.

Coughlin, P. 1992. *Probabilistic Voting Theory*. Cambridge: Cambridge University Press.

Couglin, P. and Nitzan, S. 1981. Electoral outcomes with probabilistic voting and Nash social welfare maxima. *Journal of Public Economics* 15, 113–21.

Cox, G. 1987. Electoral equilibrium under alternative voting institutions. *American Journal of Political Science* 31, 82–108.

Cox, G. 1990. Centripetal and centrifugal incentives in electoral systems. *American Journal of Political Science* 34, 903–35.

Cox, G. 1994. Strategic voting equilibria under the single nontransferable vote. *American Political Science Review* 88, 608–21.

Davis, O. and Hinich, M. 1966. A mathematical model of policy formation in a democratic society. In *Mathematical Applications in Political Science II*, ed. J. Bernd. Dallas: Southern Methodist University Press.

Davis, O. and Hinich, M. 1967. Some results related to a mathematical model of policy formation in a democratic society. In *Mathematical Applications in Political Science III*, ed. J. Bernd. Dallas: Southern Methodist University Press.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper.

Duggan, J. and Fey, M. 2005. Electoral competition with policy-motivated candidates. *Games and Economic Behavior* 51, 490–522.

Duggan, J. and Jackson, M. 2004. Mixed strategy equilibrium and deep covering in multidimensional

electoral competition. Working Paper. Department of Political Science, University of Rochester.

Enelow, J. and Hinich, M. 1982b. Nonspatial candidate characteristics and electoral competition. *Journal of Politics* 44, 115–30.

Feddersen, T. 1992. A voting model implying Duverger's Law and positive turnout. *American Journal of Political Science* 36, 938–62.

Feddersen, T. 2004. Rational choice theory and the paradox of not voting: a review. *Journal of Economic Perspectives* 18, 99–112.

Feddersen, T. and Pesendorfer, W. 1996. The swing voter's curse. *American Economic Review* 86, 408–24.

Feddersen, T. and Pesendorfer, W. 1997. Voting behavior and information aggregation in elections with private information. *Econometrica* 65, 1029–58.

Feddersen, T., Sened, I. and Wright, S. 1990. Rational voting and candidate entry under plurality rule. *American Journal of Political Science* 34, 1005–16.

Gul, F. and Pesendorfer, W. 2004. Partisan politics and aggregation failure with ignorant voters. Working Paper. Princeton, NJ: Department of Economics, Princeton University.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39, 41–57.

Kramer, G. 1977. A dynamical model of political equilibrium. *Journal of Economic Theory* 16, 310–34.

Laffond, G., Laslier, J.-F. and Le Breton, M. 1993. The bipartisan set of a tournament game. *Games and Economic Behavior* 5, 182–201.

Ledyard, J. 1984. The pure theory of large two-candidate elections. *Public Choice* 44, 7–43.

McKelvey, R. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47, 1086–112.

McKelvey, R. 1986. Covering, dominance and institution-free properties of social choice. *American Journal of Political Science* 30, 283–314.

Miller, N. 1980. A new solution set for tournaments and majority voting. *American Journal of Political Science* 24, 68–96.

Myerson, R. 1999. Theoretical comparisons of electoral systems. *European Economic Review* 43, 671–97.

Myerson, R. 2000. Large Poisson games. *Journal of Economic Theory* 94, 7–45.

Myerson, R. 2002. Comparison of scoring rules in Poisson voting games. *Journal of Economic Theory* 103, 217–51.

Osborne, M. and Slivinski, A. 1996. A model of political competition with citizen-candidates. *Quarterly Journal of Economics* 111, 65–96.

Plott, C. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57, 787–806.

Razin, R. 2003. Signaling and election motivations in voting model with common values and responsive candidates. *Econometrica* 71, 1083–120.

Roemer, J. 2001. *Political Competition: Theory and Applications*. Cambridge, MA: Harvard University Press.

Schofield, N. 1983. Generic instability of majority rule. *Review of Economic Studies* 50, 695–705.

Wittman, D. 1977. Candidates with policy preferences: a dynamic model. *Journal of Economic Theory* 14, 180–9.

Wittman, D. 1983. Candidate motivation: a synthesis of alternatives. *American Political Science Review* 77, 142–57.

How to cite this article

Austen-Smith, David. "political competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000322> doi:10.1057/9780230226203.1299

The New Palgrave Dictionary of Economics Online

‘ political economy ’

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Baran, P.A.; Bentham, J.; Cairnes, J. E.; classical economics; Dobb, M.H.; economics; Engels, F.; Hearn, W. E.; Jevons, W.S.; MacLeod, H.D.; Marshall, A.; Marshall, M.P.; Marx, K. H.; McCulloch, J. R.; Mill, J.; Mill, J. S.; Mirabeau, V. R., Marquis de; normative economics; Physiocracy; political arithmetic; political economy; Quesnay, F.; Robbins, L.C.; Robinson, J.V.; Senior, N. W.; Steuart, J.; surplus approach; Torrens, R.

Article

This article provides a survey of the origin of the term ‘political economy’ and its changes in meaning, emphasizing in particular its first modern usage in the 18th century, its demise from the end of the 19th century, when it was gradually replaced by the word ‘economics’, and its revival in a variety of forms, largely during the 1960s, which have altered its meaning from more traditional usage. What follows is therefore largely definitional and etymological, designed to indicate the lack of precise meaning associated with both the term, ‘political economy’ and its more modern synonym, ‘economics’.
The origin of words starting with ‘econom’ is Greek, from *eco* meaning ‘house’ and *nom* meaning ‘law’ in the sense appropriate to astronomy when it deals with ‘the law and order of the stars’ (Cannan, 1929, p. 37). The traditional meaning of *oikonomike* or economics, was therefore ‘household management’. Aristotle (1962, p. 30) used it in this sense when analysing households as ‘three pairs: master and slave, husband and wife, father and children’. This meaning persisted in moral philosophy until the middle of the 18th century, for example, in Hutcheson (1755) and Smith (1763, p. 141). The Latin *oeconomia* likewise meant management of household affairs, extended to management in general including orderly arrangement of speech and composition. The French *oeconomie* or *économie* took over this wider meaning of management from the Latin and when combined with *politique* it signified public administration or management of the affairs of state. Arthur Young (1770) applied this wider meaning in the title of a treatise on agricultural management. Using ‘economy’ as a synonym for ‘thrift’, ‘frugality’ and careful management of the finances of households and other organizations also derives from the

Latin adaptation. 17th-century concern with nation building gave the term) ‘public administration’ a wider scope, and given developments in France under Henry IV and Richelieu it is not surprising that the term ‘political economy’ made its first appearance there. This first use is generally attributed to Montch•etien (1615), but King (1948) indicates prior use in Mayerne-Turquet (1611). Because the relationship between state and economy it signified was so appropriate to the times, King suggests that other, perhaps earlier, uses may be found. Petty (1691, p. 181; cf. 1683, p. 483) used the term in England. As Cannan (1929, p. 39) surmised, he could as well have used ‘political economy’ as ‘political anatomy’ to describe his analysis of the Irish economy, considering he used ‘political arithmetick’ for the art of making more precise statements on the political economy of nations, interpreted as their comparative strengths (cf. Verri, 1763, pp. 9–10, who speaks of the science of political economy in this manner). Cantillon (1755, p. 46) referred to an ‘oeconomy’ in the sense of an economic organism in which classes exist as interdependent units, but his book remained an ‘Essay on Commerce’.

More precise formulations of political economy as a science of economic organization, though with continuing connotations of management, regulation and even orderly natural laws, are found in Physiocracy. Quesnay's early usage generally implies the traditional meanings, but in addition he applied the term to include discussions of the nature of wealth, its reproduction and distribution. This double meaning is particularly evident in his *Tableau économique*. It is therefore no accident that Mirabeau (1760) spoke of *économie politique* ‘as if it consisted of a dissertation on agriculture and public administration as well as on the nature of wealth and the means of procuring it’ (Cannan, 1929, p. 40). During the subsequent decades the second meaning became more dominant, the word ‘science’ was added to it (an innovation attributed to Verri, 1763, p. 9) and by the 1770s it almost exclusively referred to the production and distribution of wealth in the context of management of the nation's resources.

Sir James Steuart (1767) is the first English economist to put ‘political economy’ into the title of a book. Its introductory chapter explained that just as ‘Oeconomy in general, is the art of providing for all the wants of the family’, so the science of political economy seeks ‘to secure a certain fund of subsistence for all the inhabitants, to obviate every circumstance which may render it precarious; to provide every thing necessary for supplying the wants of the society, and to employ the inhabitants … in such a manner as naturally to create reciprocal relations and dependencies between them, so as to make their several interests lead them to supply one another with reciprocal wants’ (1767, pp. 15, 17). Steuart's full title gave the subject matter to be covered: ‘population, agriculture, trade, industry, money, coin, interest, circulation, banks, exchange, public credit and taxes’. In 1771 Verri published *Reflections on Political Economy*, the preface of which referred to a new department of knowledge called political economy. Although Smith did not use ‘political economy’ in his title the introduction and plan of his book refers to ‘different theories of political economy’ and at the start of Book IV he defined the term as ‘a branch of the science of a statesman or legislator’ with the twofold objectives of providing ‘a plentiful revenue or subsistence for the people … [and] to supply the state or commonwealth with a revenue sufficient for the publick services’ (Smith, 1776, pp. 11, 428). Elsewhere (1776, pp. 678–9) Smith indicated that he saw political economy as an inquiry into the nature and causes of the wealth of nations or, as the Physiocrats had initially suggested, the science of the nature, reproduction, distribution and disposal of wealth.

The association of the science, political economy, with material welfare proved to be particularly hardy, as was its association with the art of legislation. Bentham (1793–5, p. 223) put the matter concisely

when he argued, ‘Political Economy may be considered as a science or as an Art. But in this instance as in others, it is only as a guide to the art that the science is of use’. Torrens (1819, p. 453) also called it ‘one of the most important and useful branches of science’ while James Mill (1821, p. 211) and McCulloch (1825, p. 9) defined it as a systematic inquiry into the laws regulating the production, distribution, consumption and exchange of commodities or the products of labour. ‘Confounding’ the art with the science was criticized by Senior (1836, p. 3) as being detrimental to its development, a position likewise taken by John Stuart Mill (1831–3) and which also reaffirmed its moral and social nature. In this influential essay, Mill (1831–3, p. 140) defined political economy as ‘the science which traces the laws of such of the phenomena of society as arise from the combined operations of mankind for the production of wealth, in so far as those phenomena are not modified by the pursuit of any other object’. This position was more or less adhered to in his later *Principles* (1848, p. 21), when he defined its subject matter as ‘the laws of Production and Distribution, and some of the practical consequences deducible from them …’. Cairnes (1875, p. 35) condensed this to the statement that ‘Political Economy … expounds the laws of the phenomena of wealth.’

The middle of the 19th century saw two criticisms of this meaning of political economy. Marx (1859, p. 20) identified the study of political economy with a search for ‘the anatomy of civil society’ or, as Engels (1859, p. 218) put in his review of this book, ‘the theoretical analysis of modern bourgeois society’. This preserved the name but criticized the scope and method of political economy. Others suggested the name be changed because it had become misleading. Hearn (1863) put forward *Plutology* or the theory of efforts to satisfy human wants; MacLeod (1875) proposed ‘economics’, defining it as the ‘science which treats of the laws which govern the relations of exchangeable quantities’, a nomenclature of whose virtues he successfully persuaded Jevons (Black, 1977, p. 115). When in 1879 the Marshalls published an elementary political economy text, they called it *The Economics of Industry*. The new name of MacLeod and the Marshalls was favourably referred to in the second edition of Jevons's *Theory* (1879, p. xiv) because of convenience and scientific nicety (it matched mathematics, ethics and aesthetics) and Jevons's last published book (Jevons, 1905) bore the title *Principles of Economics*. Although Cannan (1929, p. 44) claimed Marshall (1890) induced acceptance of the new name, this only came with the later editions, and the change was not completed until the early 1920s (Groenewegen, 1985). Even then, Marshall (1890, p. 1) appeared to treat the two names as synonyms: ‘Political Economy or Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of well-being.’

Just as J.S. Mill (1831–3, pp. 120–1) had attempted retrospective codification of scope and method in the 1820s, so Robbins (1932, p. 16) redefined economics in its marginalist form as ‘the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses’. This did more than supply a meaning for the new term, ‘economics’. It destroyed the view classical economists had of their science, as Myint (1948) clearly pointed out. Others (for example, Knight, 1951, p. 6) complained that Robbins's definition neglected the link between economics and the ‘individualistic or “liberal” outlook on life, of which “capitalism”, or the competitive system, or free business enterprise, is the expression upon the economic side, as democracy on the political’. However, the major drawback of the Robbins definition was its irreconcilability with Keynes's work with its proof of the possibility of

unemployment equilibrium and hence contradicting Robbins's requirement for the existence of an economic problem that resources have to be scarce. Modern mainstream definitions of economics (Rees, 1968; Samuelson, 1955, p. 5) have simply combined the Robbinsian resource allocation problem with the new economics of employment, inflation and growth developed from Keynes's work.

Robbins's definition also aimed to make economics a ‘system of theoretical and positive knowledge’ (Fraser, 1937, p. 30), preferring to reserve the older name, ‘political economy’, for applied topics such as monopoly, protection, planning and government fiscal policy, subjects included in his essays on political economy (Robbins, 1939). Although Schumpeter (1954) held a similar opinion he was careful to warn that ‘political economy meant different things to different writers, and in some cases it meant what is now known as economic theory or “pure” economics’ (p. 22). These views of political economy conflict with the pragmatic Cambridge outlook on economics, derived from Marshall's description of economics as ‘an engine for the discovery of concrete truth’, encapsulated by Keynes (1921, p. v) in his famous introduction to the Cambridge Economics Handbooks: ‘Economics is a method rather than a doctrine, an apparatus of the mind, a technique of thinking which helps its possessor to draw correct conclusions.’ This sentiment is concisely summarized by Joan Robinson's view of economics (1933, p. 1) as ‘a box of tools’.

Marxists had never abandoned the older terminology of political economy. Dobb (1937, p. vii) defended ‘political economy’ against the new term ‘economics’ because its controversies ‘have meaning as answers to certain questions of an essentially practical kind’, associated with the ‘nature and behaviour’ of the capitalist system. Likewise, Baran (1957, p. 131) argued for a ‘political economy of growth’ because an ‘understanding of the factors responsible for the size and the mode of utilization of the social surplus … [is] a problem, not even approached in the realm of pure economics’. For the classical economists, use of the surplus had been a major research question. Political economy is therefore a very appropriate title for the endeavours of some contemporary economists to resurrect both practical and theoretical aspects of the classical tradition in what they describe as the surplus approach.

By the 1960s the radical libertarian right from Chicago and the Center for the Study of Public Choice at Virginia Polytechnic appeared to have appropriated the title ‘political economy’ for their wide application of Robbins's (1932) injunction that analysis in terms of ‘*alternatives*’ is the key distinguishing feature of economics. This effectively replaced Robbins's question ‘what is or is not economic in nature’ with the far wider one of ‘what can economics contribute to our understanding of this or that problem?’ This opens up the way for an economics of ‘family life, child rearing, dying, sex, crime, politics and many other topics’ which some of its practitioners identify with Adam Smith's research agenda (McKenzie and Tullock, 1975, p. 3). Others continue to associate the term ‘with the specific advice given by one or more economists … to governments or to the public at large either on broad policy issues or on particular proposals’ or, alternatively, as another term for ‘normative economics’ (Mishan, 1982, p. 13).

At the approach of the 21st century, both terms – ‘political economy’ and ‘economics’ – survive. During their existence, both have experienced changes of meaning. Nevertheless, they can still essentially be regarded as synonyms, a feature of this nomenclature reflecting an interesting characteristic of the science it describes. In its sometimes discontinuous development, economics or political economy has invariably experienced difficulties in discarding earlier views, and traces of old doctrine are intermingled with the latest developments in the science.

Bibliography

Aristotle. 1962. *The Politics*. Trans. J.E. Sinclair. Harmondsworth: Penguin Classics.

Baran, P.A. 1957. *The Political Economy of Growth*. Harmondsworth: Penguin Books, 1973.

Bentham, J. 1793–5. Manual of political economy. In *Jeremy Bentham's Economic Writings*, ed. W. Stark. London: George Allen & Unwin, 1952.

Black, R.D.C., ed. 1977. *Papers and Correspondence of William Stanley Jevons: Correspondence 1873–78*. London: Macmillan for the Royal Economic Society.

Cairnes, J.E. 1875. *The Character and Logical Method of Political Economy*. London. Reprinted, New York: Kelly, 1965.

Cannan, E. 1929. *A Review of Economic Theory*. London: P.S. King & Son.

Cantillon, R. 1755. In *Essay on the Nature of Commerce in General*. Ed. H. Higgs, London: Macmillan, 1931.

Dobb, M.H. 1937. *Political Economy and Capitalism*. London: Routledge.

Engels, F. 1859. Karl Marx's ‘A Contribution to the Critique of Political Economy’. *Das Volk*, No. 14, Berlin, August. In Marx (1859).

Fraser, L.M. 1937. *Economic Thought and Language*. London: A. & C. Black.

Groenewegen, P.D. 1985. Professor Arndt on political economy: a comment. *Economic Record* 61, 744–51.

Hearn, W.E. 1863. *Plutology*. Melbourne: Robertson.

Hutcheson, F. 1755. *A System of Moral Philosophy*. Glasgow: Robert and Andrew Foulis.

Jevons, W.S. 1879. *The Theory of Political Economy*, 2nd edn. London: Macmillan; Preface in 4th edn, London, 1910.

Jevons, W.S. 1905. *Principles of Economics*. London: Macmillan.

Keynes, J.M. 1921. Introduction to Cambridge economic handbooks. In *Money*, ed. D.H. Robertson.

London and Cambridge: Cambridge Economic Handbooks.

King, J.E. 1948. The origin of the term ‘political economy’. *Journal of Modern History* 20, 230–31.

Knight, F.H. 1951. Economics. In *On the History and Method of Economics*, ed. F.H. Knight. Chicago: University of Chicago Press, 1963.

McCulloch, J.R. 1825. *Principles of Political Economy with Sketch of the Rise and Progress of the Science*. London: Murray, 1870.

MacLeod, H.D. 1875. What is political economy? *Contemporary Review* 25, 871–93.

McKenzie, R.B. and Tullock, G. 1975. *The New World of Economics: Explorations into the Human Experience*. Homewood, IL: Irwin.

Marshall, A. 1890. *Principles of Economics*. 9th variorum edn, ed. C.W. Guillebaud, London: Macmillan, 1961.

Marshall, A. and Marshall, M.P. 1879. *The Economics of Industry*. London: Macmillan.

Marx, K. 1859. *A Contribution to the Critique of Political Economy*. Introduction by M. Dobb, London: Lawrence & Wishart, 1971.

Mayerne-Turquet, L. de. 1611. *La Monarchie Aristodémocratique; ou le Gouvernement composé et meslé des trois formes de légitimes républiques*. Paris.

Mill, J. 1820. *Elements of Political Economy*, 3rd edn. London, 1926. Reprinted in James Mill, *Selected Writings*, ed. D. Wisen. Edinburgh: Oliver & Boyd for the Scottish Economic Society, 1966.

Mill, J.S. 1831–3. On the definition of political economy; and on the method of investigation proper to it. Essay V in J.S. Mill, *Essays on Some Unsettled Questions of Political Economy*, London, LSE Reprint, 1948.

Mill, J.S. 1848. Principles of political economy with some of their applications to social philosophy. In *Collected Works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Mirabeau, V.R., Marquis de. 1758–60. *L’ami des hommes ou traité de la population*. Avignon and Paris.

Mishan, E.J. 1982. *Introduction to Political Economy*. London: Hutchinson.

Montch•etien, A. de. 1615. *Traité de l’économie politique*. Ed. Th. Funck-Brentano, Paris: Plon, 1889.

Myint, H.L.A. 1948. *Theories of Welfare Economics*. London: Longmans, Green & Co.

Petty, Sir W. 1683. Observations upon the Dublin bills of mortality and the state of that city. In *Economic Writings of Sir William Petty*, ed. C.H. Hull. Reprinted, New York: Kelley, 1963.

Petty, Sir W. 1691. The political anatomy of Ireland. In *The Economic Writings of Sir William Petty*, ed. C.H. Hull. New York: Kelley, 1963.

Rees, A. 1968. Economics. In *International Encyclopaedia of the Social Sciences*, vol. 4, ed. D.L. Sills. New York: Macmillan.

Robbins, L.C. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan. 2nd edn, 1935.

Robbins, L.C. 1939. *The Economic Basis of Class Conflict and Other Essays in Political Economy*. London: Macmillan.

Robinson, J.V. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

Samuelson, P.A. 1955. *Economics*, 3rd edn. New York: McGraw-Hill. 7th edn, 1967.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: George Allen & Unwin.

Senior, N.W. 1836. *An Outline of the Science of Political Economy*. London: Unwin Library of Economics, 1938.

Smith, A. 1763. *Lectures on Jurisprudence*. Ed. R.L. Meek, D.D. Raphael and P.G. Stein, Oxford: Oxford University Press, 1978.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. R.H. Campbell and A.S. Skinner, Oxford: Oxford University Press, 1976.

Steuart, J. 1767. *An Inquiry into the Principles of Political Economy*. Ed. A.S. Skinner, Edinburgh and London: Oliver & Boyd for the Scottish Economic Society, 1966.

[Torrens, R.] 1819. Mr Owen's plans for relieving the national distress. *Edinburgh Review* 32, October, Article XI.

Verri, P.In 1763. Memorie storiche sulla economia pubblica dello stato di Milano. In *Scrittori Classici Italiani di Economia Politica*, Parte Moderna, vol. 17, Milan, 1804.

Verri, P. 1771. In *Reflections on Political Economy*. Trans. B. McGilvray, ed. P. Groenewegen, Reprints of Economic Classics, Series 2, No. 4, Sydney: University of Sydney, 1986.

Young, A. 1770. *Rural Oeconomy, or Essays on the Practical Parts of Husbandry*. London.

How to cite this article

Groenewegen, Peter. "‘political economy’." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000114> doi:10.1057/9780230226203.1300

The New Palgrave Dictionary of Economics Online

political institutions, economic approaches to

Timothy Besley and Torsten Persson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Political institutions affect the rules of the game in which politics is played. Economists now have theoretical approaches to explain the impact of institutions on policy, and empirical evidence to support the relevance of the theory. This article sketches a framework to inform discussions about how political institutions shape policy outcomes. It does so using four examples: majoritarian versus proportional elections; parliamentary versus presidential government; whether to impose term-limits on office holders; and the choice between direct and representative democracy. Each example illustrates how theory and data can be brought together to investigate a specific issue.

Keywords

accountability; adverse selection; agency; citizen initiative and referendum; coalition government; corruption; direct democracy; majoritarian elections; moral hazard; parliamentary government; plurality rule; political competition; political institutions, economic approaches to; presidential government; proportional representation; rent seeking; representative democracy; representation; separation of powers; spatial voting models; targeted public spending; term limits; treatment effects

Article

1 Introduction

Political institutions play a key role in shaping economic policies. Economists now have theoretical approaches to explain this claim and empirical evidence to support it. Political institutions affect the rules of the game in which politics is played. For the most part the term ‘institutions’ is taken to mean formal rules as embodied in constitutions, and other forms of legislation. However, it may also refer to norms and informal rules.
Two basic categories of political institutions are electoral rules and forms of government. The former

term refers to features such as district magnitudes and electoral formulas that translate votes into seats. It also refers to the rules for selecting candidates and for governing their tenure in office. The latter category refers to such questions as whether the systems is presidential or parliamentary, how decision making powers are divided between central and local governments or between executive and legislature, and whether citizens have a direct say in policymaking via referenda.

Our aim in this article is to sketch an intellectual framework that informs discussions about how political institutions may shape policy outcomes. We do this by way of specific examples, referring to recent research on the topic – we do not attempt to provide a comprehensive overview of theoretical modelling or empirical knowledge. In each case, the example illustrates the potential for theoretical frameworks to shape thinking on the topic backed up with empirical analysis.

When political scientists debate democratic institutions, they frequently use two metrics for their performance – accountability and representation. The former refers to the way in which political institutions make politicians (and to some degree bureaucrats) answerable for their actions. The second refers to whether the policies and/or policymakers fairly reflect the population as a whole.

Translated into the language of economics, these two performance dimensions correspond well to two main conflicts of interest that arise in representative democracies – those between politicians and citizens and those between groups of citizens with competing economic interests. Accountability deals predominantly with the former and representation with the latter. As normative criteria, the welfare underpinnings of these metrics are somewhat vague, but they do provide a useful way of thinking about the positive effects of political institutions.

Economic models for studying accountability are mostly based on some form of *agency* approach. Such models assume that there exist problems of hidden actions (moral hazard) and hidden types (adverse selection) in politics. Politicians typically have career concerns which lead them to seek re-election. Voters decide whether or not to re-elect based on the record of politicians. To make the problem interesting, there has to be some conflict of interest between politicians and voters. The simplest (and most widely used model) supposes that this is due to opportunities for rent seeking (or effort avoidance) among politicians. The question is then how much of this conflict of interest rubs off on to policy choice in equilibrium, that is, when voters and politicians are behaving rationally and optimally. There is now a large body of literature using such models. Political institutions can affect policy in such models in three main ways: affecting the information that voters have to assess politician performance, directly affecting incentives of politicians to extract rents, and affecting the kinds of people of who are selected for public office. (See Besley, 2006, for a broad survey of agency models and their uses.)

Economic models for studying representation rely on some kind or another of a spatial framework. These models envisage citizens being located at different points in the space according to their underlying economics interests (such as their age or ability) and their social interests (such as ethnicity). The classic Downsian model of political competition (Downs, 1957) falls in this class and many subsequent developments have built on its insights. More recent work has tried to make the framework more tractable by supposing that voting is probabilistic – there is a random element in the ballots cast by voters, and politicians can therefore not be exactly sure how policies translate into voting outcomes. In standard models, competition is directly over policies without regard to who is being asked to carry these policies out. More recent approaches have looked at the problem of picking policymakers to deliver these policies. This is particularly important when modelling the credibility of policies being offered. (See Persson and Tabellini, 2000, for a broad survey of spatial models of policymaking and

their uses.)

In this article, we illustrate the main themes of the recent literature by focusing on four examples of how political institutions shape policy. Two of these examples deal with electoral rules, two with forms of government, broadly defined. Two of the examples are motivated mainly from cross-country empirical applications, while the other two are motivated more from studies of within-country variation. (Persson and Tabellini, 2003, discuss empirical work on cross-country studies of political institutions, while Besley and Case, 2003, survey within-country (cross-state) studies for the United States.) Thus, Sections 2–5 discuss, in turn, the policy consequences of adopting proportional or plurality elections, the effects of parliamentary or presidential forms of government, the consequences of term limits for elected politicians, and the impact of direct or representative democracy. Section 6 concludes.

2 Proportional or majoritarian elections

Political scientists often describe a key trade-off in electoral systems: electoral formulas based on plurality rule promote accountability at the expense of representation, while formulas based on proportional representation (PR) errs on the other side of the trade-off. Recent theoretical work by economists has analysed the consequences for governments spending of having legislative seats awarded by plurality rules rather than PR – an issue closely related to representation. The key idea is relatively straightforward (see Persson and Tabellini, 1999; Lizzeri and Persico, 2001; Milesi-Feretti, Perotti and Rostagno, 2002). If candidates with the highest vote shares win every seat at stake in a district, rather than seats in proportion to their vote shares, it becomes more attractive to target spending to small and geographically concentrated groups of voters. (The same will hold true if each district has small magnitude, that is, represents a small share of the electorate.) This tilts equilibrium policy towards spending programmes with benefits targeted to particular geographical groups, not the electorate at large, and (perhaps) towards higher overall spending.

Empirical work has sought to evaluate these predictions using cross-national data. Long-term inertia in the broad features of electoral systems makes it necessary to rely on the cross-sectional variation in the data, which, together with the non-random selection of electoral systems, raises a number of statistical issues. These issues are tackled by a variety of methods in Persson and Tabellini (2003; 2004), who classify actual electoral systems according to their electoral formula (classifying by district magnitude gives similar results) and approximate geographically non-targeted spending by welfare-state programmes, such as pensions and unemployment insurance. Their results indicate that a reform from an all-PR to an all-plurality-rule system would cut welfare spending by about two per cent of GDP in the long run. Such an electoral reform would cut overall government spending by a substantial five per cent of GDP.

The underlying theory works off the incentives of politicians and takes party structure as given. Yet it is a well documented fact that PR promotes a more fractionalized party system than plurality rule (see, for example, Lijphart, 1990). Austen-Smith (2000) studies a model where redistributive tax policy is set in post-election bargaining, assuming that the number of parties is, exogenously, higher under PR than plurality rule. He shows that this produces higher taxes and spending under PR. Bawn and Rosenbluth (2006) and Persson, Roland and Tabellini (2005) obtain a similar prediction but endogenize the number of parties. In their models of parliamentary democracy, they show that coalition governments spend

more than single-party governments under each electoral rule. We should still observe higher spending in PR systems, but this is an indirect effect of a larger number of parties increasing the incidence of coalition government. Persson, Roland and Tabellini (2005) derive an empirical way of discriminating between the indirect effect and the direct effect via the incentives of politicians. Using panel data for parliamentary democracies since 1960, they find that the higher overall spending observed under PR is entirely due to its more fractionalized party systems and hence more frequent coalition governments than under plurality rule.

A second body of theory relates to the accountability of politicians under alternative electoral systems. The key idea here is that extraction of rents – or, more generally, corruption – is better deterred the more swiftly the probability of re-election responds to performance (see Myerson, 1993; Persson and Tabellini, 2000). Large district magnitude achieves this by allowing easier entry and a larger number of candidates than small districts. Personal ballots impose individual accountability and stronger incentives than party-list ballots, which impose only collective accountability. In other words, systems where a larger number of lawmakers are elected in each district, and systems where they are elected on personal rather than party-list ballots, are both expected to reduce rent extraction by politicians. Empirically, Persson and Tabellini (2003) find quite sizeable effects in the hypothesized direction on different perception indexes of corruption, or on inefficiency in the delivery of government services.

3 Presidential or parliamentary government

How well voters can hold politicians accountable also depends on the form of government. This insight goes far back in political writing. For example, James Madison insightfully discussed various aspects of the separation of powers in his contributions to *The Federalist Papers*. Economists have recently produced modern versions of the argument as to how separation of powers across political offices may serve to limit conflicts of interest between voters and their elected representatives. Extending the agency model of Ferejohn (1986), Persson, Roland and Tabellini (1997) show that separating the proposal powers over taxes and spending creates a conflict between politicians that enables voters to better discipline their power to extract rents when in office.

This approach is extended to include issues of representation by Persson, Roland and Tabellini (2000), who analyse how different forms of government shape fiscal policy by embedding different forms of legislative bargaining in spatial voting models. They assume that presidential systems have a more extensive separation of powers across legislators than parliamentary systems. On the other hand, as in Huber (1996) and Diermeier and Feddersen (1998), parliamentary systems make the government subject to a confidence requirement of the legislature, whereas a presidential system does not (the president is directly elected). These two institutional features shape the legislative bargaining, such that legislative majorities in presidential systems become less stable than in parliamentary regimes. If majorities re-form, issue by issue, different minorities are pitted against each other for different issues on the legislative agenda. As a result, broad spending programmes suffer at the expense of targeted spending. Moreover, the lack of a stable legislative majority means that there is no well-defined residual claimant on government revenue. This reduces the incentives to boost overall taxation and spending. Overall, we should thus expect presidential regimes to be associated with lower total spending and smaller broad (non-targeted) spending programmes than parliamentary regimes.

Persson and Tabellini (2003; 2004) confront these predictions with data, in which real-world forms of government are classified as parliamentary or presidential, depending on whether the executive is subject to the continual confidence of the legislature. For broad welfare state programmes, they find the hypothesized result only among long established democracies, among which presidential regimes spend less, by about two per cent of GDP. For overall spending the results are very robust across samples and in line with the basic hypothesis. Whether the results are obtained by OLS, instrumental variables or matching methods, the finding is that presidential regimes have smaller governments by at least five per cent of GDP – again, a large number.

4 Term limits or no term limits

Political accountability is achieved in part by re-election chances responding to performance while in office. This resembles the kind of contractual relations that arise in a market context and provide workers with incentives. However, the relationships between politicians and voters are not contractual – they resemble something closer to a fiduciary relationship. While political parties may have a role in disciplining politicians, the ultimate sanction is an electoral one: poorly performing incumbents are removed from office by the voters.

The frequency of re-election and the number of terms that a politician can serve become important institutional choices in shaping electoral accountability. The agency model of politics referred to above provides a tool to approach these issues. The theory suggests two ways of thinking about term limits: incentive effects and selection effects. Incentive effects arise because politicians who face a shorter time horizon are less obliged to please voters. Whether this increases or reduces the quality of policy is moot. On the one hand, politicians facing term limits may have less incentive to please voters and hence may follow their private agendas. But they may also pander to voters, eschewing hard decisions that impose short-run costs in exchange for long-run benefits. This latter effect can lead term-limited politicians to act more in the voters' interests. Either way, if electoral incentives matter, then we should expect term limits to shape political decisions. Terms limits will also induce a selection effect. Politicians have to be elected to lame-duck terms. Rational voters should anticipate this when deciding whether to (re)elect them, which will make politicians elected to lame-duck terms better than average. Such positive selection may counteract any adverse incentive effect.

US states provide a natural experiment for looking at the impact of term limits, because governors are subject to such limits in around half the states. This allows two kinds of comparisons: across time – governors when they are up against a term limit versus their first (non-term limited) period in office – and across states – term-limited versus non-term-limited governors.

Besley and Case (1995) identify the effect of a term limit from the difference between first and second terms in office for incumbents facing term limits. Controlling for state fixed effects and year effects, and using annual data from the 48 continental US states from 1950 to 1986, they find that a variety of policy measures are affected by term limits. Specifically, state taxes and spending are higher in the second term when term limits bind in states that have them. Such limits tend to induce a fiscal cycle, with states having lower taxes and spending in the first gubernatorial term than in the second. More recently, List and Sturm (2006) have applied these ideas to environmental policies at the US state level and also find evidence of a term-limit effect. They observe that the way in which environmental interests are

represented in policy may depend on whether the governor is in his last term in office.
Term limits have also been advocated as solutions to institutional distortions in legislatures. A good example is the committee system in the US Congress, which puts a premium on seniority of politicians and thus, effectively, a lower performance threshold for incumbents with a resulting diminution in accountability (see Dick and Lott, 1993, for development of this argument).
A host of studies look for effects of announced retirements on voting behaviour in Congress. On the whole, it has been difficult to find evidence of a last-period effect. For example, Lott and Bronars (1993) analyse Congressional voting data from 1975 to 1990 and find no significant change in voting patterns in a representative's last term in office. McArthur and Marks (1988) look at Congressional behaviour in a lame-duck session of Congress: in post-election sessions, members who have not been re-elected are at times called upon to vote on legislation before the swearing in of the new Congress. They find that lame-duck representatives were significantly more likely in 1982 to vote against automobile domestic content legislation than were returning members.

5 Direct or representative democracy

Whether polities should use some element of direct democracy as part of their political institutions is widely debated. The two most famous examples are US states and Swiss cantons, which display considerable variation in their reliance on citizen initiatives and referenda. From a theoretical point of view, issues of accountability and representation are important in thinking through these issues.
Some commentators (for example, Denzau, Mackay and Weaver, 1981) emphasize the role of initiatives in reducing rent-seeking by government and hence enhancing accountability in the political process. This underpins a number of studies investigating whether jurisdictions that permit initiatives have smaller governments. For example, Matsusaka (1995) regresses government expenditures and revenues on a number of control variables for a panel of 49 US states (Alaska excluded) sampled over a 30-year period at five-year intervals from 1960 to 1990. He includes year effects, but not state fixed effects, since the presence of initiatives is largely fixed within states over time. His main finding is a strong negative effect on expenditures of access to the initiative. Matsusaka (1995) also finds some evidence that the effect is strongest where the number of citizen signatures required for a referendum is low. Similarly, Pommerehne (1990) shows that Swiss cantons using the initiative indeed have smaller governments.
Others emphasize the fact that initiatives can change the representation of policy preferences. A large body of empirical evidence from political science supports the lack of congruence of policy and voter preferences on a variety of issues (see Besley and Coate, 2000, for references).
Gerber (1999) considers how, given a set of policy preferences in a legislature, the availability of the initiative could change the equilibrium policy bargain. Moreover, the legislature may make such a change pre-emptively, that is, it is sufficient for legislators to anticipate the *possibility* of an initiative at a later date. Hence, the possibility of initiatives forces a greater agreement between voter preferences and policy outcomes, on the assumption that representatives elected to the legislature have views that are out of step with the citizens as large. Similar conclusions follow from the theoretical analysis of Besley and Coate (2000) but for quite different reasons. They develop a model in which initiatives affect electoral outcomes. They argue initiatives have an impact via *issue unbundling*. In general elections,

many issues are decided at once, which may result in non-salient issues being distorted away from the preference of a majority. Initiatives allow such issues to be unbundled from other issues in the election. Besley and Coate show that this can change the probability distribution of a range of policy outcomes and the composition of candidates who are chosen to run. Both of these theoretical approaches, as well as many popular discussions of initiatives, imply that citizen initiatives are a device for bringing policy into line with public opinion.

One strand of empirical literature on initiatives has used data from US states to test whether public opinion and policy outcomes are closer together in initiative states. For example, Lascher, Hagen and Rochlin (1996) and Camobreco (1998) investigate whether the link between aggregate measures of policy outcomes and public opinion is closer when states allow citizens' initiatives. They find no significant effect. With respect to specific policy issues, Gerber (1999) uses cross-sectional state variation from the 1990s and compares stances on an array of policies. She finds significant differences (at the ten per cent level) for personal income taxes (initiative states lower); highway, natural resources and hospital spending (initiative states higher in all cases); and the implementation of three-strike legislation (initiative states lower). Gerber looks in greater detail at the death penalty and parental consent laws for abortion, using public opinion data to estimate median voter preferences. With cross-sectional data for 1990, she runs a logistic regression that interacts whether a state has an initiative with public opinion, and finds that states with initiatives mirror public opinion on abortion and the death penalty more closely, even though these policies are not directly determined via initiatives.

6 Final remarks

The examples discussed above illustrate how knowledge in the field has benefited from research targeted towards understanding specific issues, even though these issues can be nested in broader debates about accountability and representation. Theoretical and empirical research on the boundary between economics and political science has uncovered systematic relationships between political institutions and policy outcomes, and is currently being extended to new domains of economic policymaking.

One challenge for the future is to study what determines changes in institutions over time. It is evident that studying how political institutions work, the focus of the discussion here, is a necessary part of research on institutional change. From a theoretical point of view, it is important to understand whose interests are served by particular institutional arrangements and how policies change as a consequence of them. For practical purposes, this will likely be a piecemeal agenda dealing with specific constitutional arrangements rather than examining constitution design from the ground up. This is why the kind of nuts and bolts issues illustrated in our four examples provides the basis for further progress in the field.

Much of the empirical research, so far, has adopted a relatively simple approach, in which political institutions are taken as given and the hypothesized institutional impact is the same across political, social and economic conditions. As is well known from the microeconometric treatment literature, this can easily lead to biased estimates. Current research has started to address non-random selection of political institutions as well as the likely existence of heterogenous treatment effects, where the effect of a specific institutional reform depends on social and historical preconditions. Measurement and econometric testing of these complex issues would benefit greatly from new theoretical research on the

endogeneity and conditional effects of institutional reform.

See Also

- political competition

We are grateful to Jenny Mansbridge for helpful comments.

Bibliography

Austen-Smith, D. 2000. Redistributing income under proportional representation. *Journal of Political Economy* 108, 1235–69.

Bawn, K. and Rosenbluth, F. 2006. Short versus long coalitions: electoral accountability and the size of the public sector. *American Journal of Political Science* 50, 251–65.

Besley, T. 2006. *Principled Agents: The Political Economy of Good Government*. Oxford: Oxford University Press.

Besley, T. and Case, A. 1995. Does political accountability affect economic policy choices? Evidence from gubernatorial term limits. *Quarterly Journal of Economics* 110, 769–98.

Besley, T. and Case, A. 2003. Political institutions and policy choices: evidence from the United States. *Journal of Economic Literature* 41, 7–73.

Besley, T. and Coate, S. 2000. Issue unbundling via citizens' initiatives. Working Paper No. 8036. Cambridge, MA: NBER.

Camobreco, J.F. 1998. Preferences, fiscal policies and the initiative process. *Journal of Politics* 60, 819–29.

Denzau, A.T., Mackay, R.J. and Weaver, C. 1981. On the initiative-referendum option and the control of monopoly government. In *Tax and Expenditure Limitations*, ed. H.F. Ladd and T.N. Tideman. Washington, DC: The Urban Institute.

Dick, A.R. and Lott, L., Jr. 1993. Reconciling voters' behavior with legislative term limits. *Journal of Public Economics* 50, 1–14.

Diermeier, D. and Feddersen, T. 1998. Cohesion in legislatures and the vote of confidence procedure. *American Political Science Review* 92, 611–21.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper.

Ferejohn, J. 1986. Incumbent performance and electoral control. *Public Choice* 50, 5–25.

Gerber, E. 1999. *The Populist Paradox: Interest Group Influence and the Promise of Direct Legislation*. Princeton, NJ: Princeton University Press.

Huber, J. 1996. The vote of confidence procedure in parliamentary democracies. *American Political Science Review* 90, 269–82.

Lascher, E.L., Jr., Hagen, M.G. and Rochlin, S.A. 1996. Gun behind the door? Ballot initiatives, state policies and public opinion. *Journal of Politics* 58, 760–95.

Lijphart, A. 1990. The political consequences of electoral laws, 1945–1985. *American Political Science Review* 84, 481–96.

List, J. and Sturm, D. 2006. How elections matter: theory and evidence from environmental policy. *Quarterly Journal of Economics* 121, 1249–81.

Lizzeri, A. and Persico, N. 2001. The provision of public goods under alternative electoral incentives. *American Economic Review* 91, 225–45.

Lott, J.R., Jr. and Bronars, S.G. 1993. Time series evidence on shirking in the U.S. House of Representatives. *Public Choice* 76, 125–49.

Matsusaka, J.G. 1995. Fiscal effects of voter initiative: evidence from the last 30 years. *Journal of Political Economy* 103, 587–623.

McArthur, J. and Marks, S.V. 1988. Constituent interest vs. legislator ideology: the role of political opportunity cost. *Economic Inquiry* 26, 461–70.

Milesi-Feretti, G.-M., Perotti, R. and Rostagno, M. 2002. Electoral systems and the composition of government spending. *Quarterly Journal of Economics* 117, 609–57.

Myerson, R. 1993. Effectiveness of electoral systems for reducing government corruption: a game theoretic analysis. *Games and Economic Behavior* 5, 118–32.

Persson, T., Roland, G. and Tabellini, G. 1997. Separation of powers and political accountability. *Quarterly Journal of Economics* 112, 1163–202.

Persson, T., Roland, G. and Tabellini, G. 2000. Comparative politics and public finance. *Journal of*

Political Economy 108, 1121–61.

Persson, T., Roland, G. and Tabellini, G. 2005. Electoral rules and government spending. Mimeo, Stockholm University.

Persson, T. and Tabellini, G. 1999. The size and scope of government: comparative politics with rational politicians. 1998 Alfred Marshall lecture. *European Economic Review* 43, 699–735.

Persson, T. and Tabellini, G. 2000. *Political Economics: Explaining Economic Policy*. Cambridge, MA: MIT Press.

Persson, T. and Tabellini, G. 2003. *The Economic Effect of Constitutions*. Cambridge, MA: MIT Press.

Persson, T. and Tabellini, G. 2004. Constitutional rules and fiscal policy outcomes. *American Economic Review* 94, 25–64.

Pommerehne, W.W. 1990. The empirical relevance of comparative institutional analysis. *European Economic Review* 34, 458–69.

How to cite this article

Besley, Timothy and Torsten Persson. "political institutions, economic approaches to." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000252> doi:10.1057/9780230226203.1301

The New Palgrave Dictionary of Economics Online

pollution haven hypothesis

Arik Levinson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The pollution haven hypothesis, or pollution haven effect, is the idea that polluting industries will relocate to jurisdictions with less stringent environmental regulations. Empirical studies of the phenomenon have been hampered by the difficulty of measuring regulatory stringency and by the fact that stringency and pollution are determined simultaneously. Early studies based on cross sections of data found no significant effect of regulations on industry locations. Newer studies that use panels of data to control for unobserved heterogeneity or instrumental variables to account for simultaneity have found statistically significant, reasonably sized effects.

Keywords

environmental regulations; fixed-effects models; free trade; Heckscher–Ohlin trade theory; inter-jurisdictional competition; not in my backyard (NIMBY); pollution abatement costs and expenditures; pollution haven hypothesis; Porter hypothesis; protectionism; race to the bottom; regulatory stringency; tariffs; unobserved heterogeneity

Article

The pollution haven hypothesis (or pollution haven effect) posits that jurisdictions with weak environmental regulations – ‘pollution havens’ – will attract polluting industries relocating from more stringent locales. The premise is intuitive: environmental regulations raise the cost of key inputs to goods with pollution-intensive production, and reduce jurisdictions’ comparative advantage in those goods. The Heckscher–Ohlin model provides the theoretical foundations by showing that regions will export goods that use locally abundant factors as inputs. Empirically, however, robust evidence that industries shift production to less stringent jurisdictions has proven elusive.

Econometric studies of the pollution haven effect have typically focused on reduced-form regressions of a measure of economic activity on some measure of regulatory stringency and other covariates:

$$Y_i = \alpha R_i + X_i' \beta_i + \varepsilon_i \quad (1)$$

where Y is economic activity, R is regulatory stringency, X is other characteristics that will affect Y, and ε is an error term. The pollution haven hypothesis is that estimates of $\partial Y/\partial R$ will be negative ($\widehat{\alpha} < 0$). The empirical literature contains a wide variety of implementations of (1). Some studies focus on international trade, where Y_i represents, say, net exports from country i, and the right-hand side contains country characteristics. Others focus on employment, foreign direct investment, or new manufacturing plant births. Equation (1) has also been used to examine the pollution haven hypothesis at the level of sub-national jurisdictions, such as US states or counties. Some studies have further disaggregated Y by industry, in the expectation that environmental regulations have a larger effect on polluting industries than on clean ones.

On the right-hand side of (1), finding an appropriate measure of regulatory stringency (R) is not simple. The problem is not merely one of collecting the appropriate data; merely conceiving of data that would represent R is difficult. What we want to know is how much more costly production is in a given jurisdiction relative to others, due to the jurisdiction's environmental regulations. These environmental compliance costs could take many forms: environmental fees or taxes, permitting costs, regulatory delays, emissions limits that require installation of costly technology, the threat of lawsuits, product or process redesign, forgone output, and so forth. Some attempts to measure these costs involve creating indices by weighting various country or state characteristics such as environmental agencies' budgets, public awareness of environmental problems, the number of international environmental agreements the country has joined, states' congressional delegations' voting on environmental issues, or other general indicators. Other studies have used measurements of pollution directly, arguing that, for example, high sulphur emissions are evidence of lax regulations. Studies based on US data have used measures of manufacturers' pollution abatement expenditures by state or industry, using the US Census Bureau's *Pollution Abatement Costs and Expenditures* (PACE) survey, which ran from 1973 to 1994 and resumed in 2005.

None of these measures of R is ideal for testing the pollution haven hypothesis. The compiled indices of stringency are inherently ad hoc, and typically not available in more than one cross section. Using pollution directly as a proxy for stringency is also problematic. High levels of pollution could be symptomatic of lax of regulations, or could mean that the jurisdiction, finding itself with a poor environment, must enact stringent regulations to reduce pollution. This is true in the United States, where counties that are out of compliance with national air-quality standards are required by the federal Clean Air Act to enforce stricter emissions laws. Even direct measures of abatement costs from the PACE are troublesome. States with the highest average abatement costs are those with the most polluting industrial compositions. Estimates of (1) in which average abatement costs proxy for R find that more polluting industries locate in places with higher abatement costs – the opposite of the pollution haven effect.

Even if we had available an ideal measure of regulatory stringency, R, two further econometric issues

complicate estimates of eq. (1): unobserved heterogeneity and simultaneity. The first problem is that some unobserved characteristics of the jurisdictions or industries being studied are likely to be correlated with both economic activity and regulatory stringency. A country with an unobserved comparative advantage in a polluting good (abundant high-sulphur coal or proximity to markets) is likely to both export that good and enact strict environmental regulations. This means that R and ε are correlated in (1), and estimates of $\hat{\alpha}$ will be biased. In fact, cross-section comparisons sometimes find that countries with higher stringency have more polluting activity, which is in turn easily mistaken for evidence of the Porter hypothesis that environmental regulations promote competitiveness (Porter and van der Linde, 1995).

The simplest solution to the problem of unobserved heterogeneity is to estimate a panel-data version of (1) and include fixed effects by jurisdiction or industry, whatever the relevant unit of observation:

$$Y_{it} = v_i + \alpha R_{it} + X'\beta_{it} + \varepsilon_{it} \quad (2)$$

These fixed effects (v_i) capture the unobserved characteristics of jurisdictions or industries that make them likely to have both strict environmental regulations and high levels of activity. However, including fixed effects requires panel data on regulatory stringency, which makes measuring stringency in the first place even more difficult.

The second econometric issue confronting estimates of (1) and (2) is that economic activity and pollution regulations may be determined simultaneously. The pollution haven hypothesis suggests that environmental regulations affect exports, but the reverse may also be true: exports may affect regulations. If trade increases incomes, and environmental quality is a normal good, trade could increase voters' demand for strict environmental regulations. Or, increased pollution caused by trade could increase local demand for strict environmental regulations. In theory the straightforward solution to this problem is to use instrumental variables. In practice this means finding instruments for a variable, R, that is difficult to measure in the first place. In the panel context (2), it means finding something that changes over time, is correlated with R_{it}, and is uncorrelated with ε_{it}.

The empirical studies that employ these techniques span more than 30 years, and are growing in number. While enumerating them here would be impractical, their broad lessons are becoming clear. The first generation of empirical work on the pollution haven hypothesis used cross sections of data and made no attempt to control for unobserved heterogeneity or simultaneity. Most of them found small insignificant effects of environmental regulations, a few found counter-intuitive positive effects, and none found robust significant support for the pollution haven hypothesis. This early literature is summarized in Jaffe et al. (1995, p. 157): 'Overall, there is relatively little evidence to support the hypothesis that environmental regulations have had a large adverse effect on competitiveness.'

In recent years, economists have begun to use panels of data and fixed-effects models to control for unobserved heterogeneity, and instrumental variables to control for simultaneity. In contrast to the earlier cross-section studies, this newer work has tended to find statistically significant, reasonably sized

evidence of pollution havens. It is catalogued in detail by Brunnermeier and Levinson (2004), and summarized in Copeland and Taylor (2004, p. 48), who write that ‘after controlling for other factors affecting trade and investment flows, more stringent environmental policy acts as a deterrent to dirty-good production’.

One example of this recent literature exploits the US Clean Air Act, which mandates that every county in the United States achieve the same minimum level of ambient air quality. Federal law requires counties that fail to attain this standard to implement more stringent regulations. A convenient aspect of this law for pollution haven research is that from the perspective of any single county the law is exogenous. Neither the law's first enactment in 1970 nor any subsequent tightening of the air quality standards has been a function of any one county's characteristics. This suggests that an indicator for whether a particular county is in compliance with the national standards makes a good instrument for the stringency of that county's environmental regulations. Non-compliance changes over time, is correlated (positively) with stricter regulations, and is unlikely to be correlated with ε_{it}. Using this strategy, Becker and Henderson (2000) find that a county's failure to meet the national air quality standards reduces the number of new plants being built by four heavily polluting industries by between 26 and 45• percent. Greenstone (2002) shows that these non-attainment counties had about 590,000 fewer jobs, $37 billion lower capital stock, and $75 billion lower output (in 1987 US dollars) between 1972 and 1987 than counties that met the national standards.

An important caveat should accompany findings of this type: they are positive, or descriptive, rather than normative. These tests of the pollution haven hypothesis merely measure whether industry relocates to less stringent jurisdictions; they have no welfare implications. Nevertheless, advocacy groups with widely varying agendas have seized on the issue. Some environmental groups express concern about pollution increases, resulting either from the trade-induced change in the pollution havens’ industrial compositions or from the increase in overall economic activity due to trade. Manufacturing interests and labour unions in developed countries worry that the pollution haven effect means a loss of domestic profits and jobs. Free trade advocates fear that protectionist interests will use environmental regulations as a justification for trade barriers, or as a direct protectionist mechanism by lobbying for lower environmental standards as a form of subsidy to manufacturers. Anti-globalization protestors claim that trade liberalization will exacerbate all of these outcomes: degrading environmental quality in developing countries, weakening manufacturing in developed countries, and deterring all countries from setting sufficiently strict environmental standards.

In some cases these diverse parties have different or related interpretations of the pollution haven hypothesis. The most straightforward interpretation, represented by $\alpha < 0$ in eqs. (1) and (2), is that environmental regulations cause polluting activity to shift to less stringent jurisdictions. Although virtually all of the empirical literature tests this descriptive hypothesis, much of the policy debate revolves around tangential issues with more normative implications.

One such related issue is whether trade liberalization exacerbates the pollution haven effect. Note the subtle difference. The straightforward pollution haven hypothesis is that environmental regulations affect trade. This extension claims that trade barriers disproportionately affect trade in polluting goods, and hence the environment. It seems that would be true only if the trade barriers had a larger effect on polluting industries than on clean industries. An empirical test of this extension would rewrite eq. (2) to include trade barriers and an interaction between trade barriers and regulations:

$$Y_{it} = v_i + \alpha R_{it} + \gamma T_{it} + \theta R_{it} T_{it} + X' \beta_{it} + \varepsilon_{it} \quad (3)$$

where T_{it} represents trade barriers such as tariffs (Ederington, Levinson and Minier, 2004). The straightforward pollution haven effect is now $\partial Y / \partial R = \alpha + \theta T$. The indirect effect of trade barriers on the pollution haven effect is $\partial \partial Y / [\partial T \partial R] = \theta$. Given the difficulties in measuring both regulatory stringency and trade barriers, and the likely endogeneity of both, few studies have attempted to estimate this indirect effect of trade liberalization on pollution havens. Nevertheless, it is important to be clear that the basic empirical estimates of the pollution haven effect do not address this more complex extension.

A second concern related indirectly to the pollution haven hypothesis is that governments will engage in inefficient competition to attract polluting industries by weakening their environmental standards. A welfare-maximizing government should set standards so that the benefits justify the costs at the margin. This does not mean that environmental standards will be equal everywhere. Jurisdictions have different assimilative capacities, costs of abatement, and values regarding the environment. So heterogeneity in pollution standards is to be expected, and by extension industry migration to less stringent jurisdictions does not necessarily raise efficiency concerns.

There might be cause for concern, however, if jurisdictions compete for investment from polluting industries by setting environmental regulations below Pareto-efficient levels. They might do so, for example, if there were cross-border spillovers, and the benefits of hosting a polluting manufacturer outweighed the *local* costs. Alternatively, if the industry is concentrated and pays rents to outside shareholders, jurisdictions may compete away their ability to capture some of the industry's rents. In these types of case, countries may lower their regulations below the Pareto-optimal levels in a ‘race to the bottom’ in environmental standards. Depending on the costs and benefits of hosting a polluting industry, they may also raise their standards above the Pareto-optimal levels in what has been called the ‘not-in-my-backyard’ (NIMBY) phenomenon. Levinson (2003) summarizes the theoretical and empirical literature on inter-jurisdictional environmental competition.

These questions of trade liberalization and inter-jurisdictional competition, however, extend the central issue of the pollution haven hypothesis. Most empirical studies of the pollution haven hypothesis ask the straightforward, descriptive question: have pollution-intensive industries become concentrated in jurisdictions with less stringent regulations? Early analyses based on cross sections of data typically found that environmental regulations had small or statistically insignificant effects on industry location. However, recent studies using panel data to control for unobserved heterogeneity or instrumental variables to control for the simultaneity of regulations have found statistically significant, reasonably sized pollution haven effects.

See Also

- environmental economics

- international trade theory
- strategic trade policy
- trade policy, political economy of

Bibliography

Becker, R. and Henderson, J. 2000. Effects of air quality regulations on polluting industries. *Journal of Political Economy* 108, 379–421.

Brunnermeier, S. and Levinson, A. 2004. Examining the evidence on environmental regulations and industry location. *Journal of Environment & Development* 13, 6–41.

Copeland, B. and Taylor, M. 2004. Trade, growth, and the environment. *Journal of Economic Literature* 42, 7–71.

Ederington, J., Levinson, A. and Minier, J. 2004. Trade liberalization and pollution havens. *Advances in Economic Analysis & Policy* 4(2) Article 6. Berkeley Electronic Press.

Greenstone, M. 2002. The impacts of environmental regulations on industrial activity: evidence from the 1970 and 1977 Clean Air Act amendments and Census of Manufactures. *Journal of Political Economy* 110, 1175–219.

Jaffe, A., Peterson, S., Portney, P. and Stavins, R. 1995. Environmental regulations and the competitiveness of US manufacturing: what does the evidence tell us? *Journal of Economic Literature* 33, 132–63.

Levinson, A. 2003. Environmental regulatory competition: a status report and some new evidence. *National Tax Journal* 56, 91–106.

Porter, M. and van der Linde, C. 1995. Toward a new conception of the environment–competitiveness relationship. *Journal of Economic Perspectives* 9(4), 97–118.

U.S. Census Bureau. Various years. *Pollution Abatement Costs and Expenditures, MA200*. Washington, DC: U.S. Government Printing Office.

How to cite this article

Levinson, Arik. "pollution haven hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000230> doi:10.1057/9780230226203.1302

The New Palgrave Dictionary of Economics Online

pollution permits

Ted Gayer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Government can reduce pollution by issuing permits to polluters in numbers below existing emission levels. Under a tradable permit programme, a firm with high abatement costs can buy permits from another firm with low abatement costs, leading to a reduction in the total cost of abating relative to a system where reduction levels are strictly assigned. For tradable permits to work effectively, the emissions must come from discrete point sources and be relatively easy to monitor. Aside from issuing the permits, the government's role is to enforce compliance and establish optimal penalties for non-compliance.

Keywords

auction hot spot; carbon emissions; Clean Air Act Amendments of 1990.; Coase Theorem; externalities; market failure; pollution permits; property rights; transaction costs

Article

The government issues pollution permits to designate how many units of a given pollutant the permit owner is legally allowed to emit in a given period. The government can therefore reduce pollution from these sources by setting the total number of permits below their total existing emission levels. The cost savings of this approach result from allowing the pollution permits to be traded (Dales, 1968). Under such a tradable permit programme (also known as a cap-and-trade programme), a firm that has high abatement costs can buy permits from another firm that has low abatement costs, leading to a reduction in the total cost of abating relative to a system where reduction levels are strictly assigned. For tradable permits to work effectively, the emissions must come from discrete point sources and be relatively easy to monitor. Aside from issuing the permits, the government's role is to enforce compliance and establish optimal penalties for non-compliance.

Tradable pollution permits can help address welfare losses caused by pollution. In a free market system

goods are exchanged voluntarily. Buyers and sellers engage in trade only if both parties believe they will benefit from the exchange. These trades are coordinated by market prices, which convey information to all parties on the demand for the good and the cost of supplying the good. This system of mutual improvement results in an efficient allocation. However, inefficiency may result if a voluntary transaction between two parties imposes involuntary costs on a third party. These third-party costs are known as externalities.

The root of this market failure is that there are no clear property rights for the surrounding air. Consider an example of a firm which emits air pollution that imposes costs on its neighbours. If the firm's neighbours owned the rights to clean air, then the firm would need to compensate the neighbours in order to use the air in its production process. Similarly, if the firm owned the rights to pollute the air, the neighbours could pay the firm to reduce its emissions. In either setting, the market would incorporate both the costs of pollution to the neighbours and the benefits of pollution to the firm, resulting in an efficient outcome. Indeed, the key interpretation of the Coase Theorem (Coase, 1960) is that efficiency results no matter who is legally assigned the property right to the air, so long as free exchange is possible and there are no transaction costs.

The necessary condition of no (or even low) transaction costs is likely to be violated when there are many sources of pollution and when many people bear the external costs of the pollution. This presents an economic justification for government involvement, since the absence of a working market for clean air will lead to an externality-induced inefficiency.

If high transaction costs preclude efficiency-enhancing bargaining between third parties and polluters, then the government can assume the role of the property right owner for the air. Because the ‘government’ is not an individual cost-bearing entity in the same manner as the affected third parties, and because government agents may have goals other than efficiency, it is not assured that government regulation will lead to an efficient outcome. Ideally, the role of government would be to assess the external costs associated with the production process and to determine the pollution reduction level that maximizes net benefits. The government could then issue tradable permits that yield this efficient level of pollution reduction.

It is undoubtedly difficult to determine the efficient amount of pollution reduction. However, no matter which target level is chosen, a system of tradable permits can help achieve the goal in the least costly way. In a cap-and-trade system, a firm with a high cost of reducing an additional unit of emissions could purchase a pollution permit from a firm with a lower marginal abatement cost. This trading will continue until the marginal abatement costs are equal across firms, thus minimizing the total cost. The cost-savings occur no matter if the government initially gives out the permits to firms for free (known as grandfathering), or if the government decides to auction the permits. Given a competitive market for permits, the initial allocation of permits has only distributional consequences, not efficiency consequences.

The freedom to trade permits across firms or to bank (or even borrow) permits across time results in cost savings without violating the long-term total pollution reduction goal. Additionally, by creating a property right for pollution, a cap-and-trade system establishes a market price for pollution and therefore provides firms with an incentive to find less expensive ways to reduce emissions (Carlson et al., 2000). In contrast, a regulation that rigidly sets technological standards for a firm does not provide this incentive.

Some environmental problems are difficult to address with tradable permits. For example, if the

marginal damage of emissions varies by location (for example, due to variation in existing ambient concentrations or due to differences in the number of people exposed to the pollutant), then a tradable permit system might shift emissions from a low-damage to a high-damage location and thus increase total damages. A congestion of emissions in one location, known as a ‘hot spot’, could result in greater damage than if pollution were reduced uniformly across polluters. A cap-and-trade system can address this problem by making the required number of permits per unit of emissions a function of the marginal damage, or by establishing separate permit markets by region. However, these options do add a level of complexity. In addition, in order for a tradable permit market to work efficiently, it must be a competitive market composed of informed buyers and sellers. While tradable permits can minimize costs, in practice such programmes are grafted on to existing command-and-control regulations, which can affect the cost savings (Hahn, 1989).

Since around 1985 the United States has adopted a number of tradable permit programmes to address a variety of pollution problems (Stavins, 2000). These include the phase-out of leaded gasoline, chlorofluorocarbon trading, the Regional Clean Air Incentives Market (RECLAIM) to address sulfur dioxide and nitrogen oxides, and the recent Nitrogen Oxides State Implementation Plan (SIP) Call. The most notable example of a cap-and-trade system is the sulphur dioxide programme for electricity generating units, which was enacted under the Clean Air Act Amendments of 1990. This programme has achieved its pollution-reduction goals with estimated cost savings of approximately $1 billion a year compared with costs under a hypothetical command-and-control regulatory alternative.

See Also

- auctions (empirics)
- environmental economics
- environmental Kuznets curve

Bibliography

Carlson, C., Burtraw, D., Cropper, M. and Palmer, K. 2000. Sulfur dioxide control by electric utilities: what are the gains from trade? *Journal of Political Economy* 108, 1292–326.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3(1), 1–44.

Dales, J. 1968. *Pollution, Property and Prices*. Toronto: Toronto University Press.

Hahn, R. 1989. Economic prescriptions for environmental problems: how the patient followed the doctor's orders. *Journal of Economic Perspectives* 3(2), 95–114.

Stavins, R. 2000. Market-based environmental policies. In *Public Policies for Environmental Protection*, 2nd edn., ed. P. Portney and R. Stavins. Washington, DC: Resources for the Future.

How to cite this article

Gayer, Ted. "pollution permits." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000311> doi:10.1057/9780230226203.1303

The New Palgrave Dictionary of Economics Online

Pontryagin's principle of optimality

M.I. Zelikin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

The meaning of the word ‘economics’ is closely related with that of ‘optimality’. It is for this reason that methods used in the theory of optimal control find their natural practical application in economics.
In this entry we deal with the statement of Pontryagin's maximum principle and give an exposition of the results and the perspectives of its applications to macroeconomic optimizational problems. We are concerned with two lines in the development of macroeconomics – that of Ramsey and that of von Neumann. Pontryagin's maximum principle embraced both lines – they now coexist in the principle, being inseparable and yet unmergeable. We begin with the classical formulation of Pontryagin's principle.
Let the state of the given system be described by the vector $x = (x_1, \ldots, x_x);\ x \in X \subset \mathbb{R}^x$ (X is an open domain). Control is described by the vector $u = (u_1, \ldots, u_y);\ u \in U \subset \mathbb{R}^y$. The independent variable t is time. For control one can choose any piecewise continuous function $u(t)$, whose values belong to U. The dynamics of the system are described by the equations $\dot{x}_i = \varphi_i(t, x, u(t)),\ (i - 1, \ldots, n);\ x(t_0) = a$. The pair consisting of the control $u(t)$ and the corresponding path $x(t)$ is called the process. A smooth manifold M in the space (t,x) is given, and the first hitting time of this manifold M is taken as the moment of termination of the process. The hitting time is the moment of first arrival of the point x at the manifold M, i.e. $T = \inf\{t | x(t) \in M\}$. In the case when M is the hyperplane $t=T$=Const one says that this is a fixed time and free end problem. The criterion is the functional

$$x_0 = F(T, x(T)) + \int_{t_0}^{T} f(t, x(t), u(t))\mathrm{d}t \to \sup .$$

In the case $f \equiv 0$ and T=Const the functional x_0 is said to be terminal. It is assumed that the functions f, F and ϕ are smooth. To formulate Pontryagin's maximum principle let us consider a dual (or an adjoint) vector $\psi = (\psi_0, \psi_1, \ldots, \psi_n)$ and the Pontryagin function

$$H(t, \psi, x, u) = \psi_0 f(t, x, u) + \sum_{\alpha=1}^{n} \psi_\alpha \varphi_\alpha(t, x, u) .$$

Pontryagin's maximum Principle

If $u^*(t)$,$x^*(t)$ is the optimal process, then there exists a nontrivial, continuous vector-function $\psi(t) = (\psi_0, \psi_1(t), \ldots, \psi_n(t))$

with the following properties

1. (1) The adjoint equations:

$$\dot{\psi}_0 = 0;\ \dot{\psi}_i = -\frac{\partial H}{\partial x_i}(t, \psi, x^*(t), u^*(t)),\ (i = 1, \ldots, n).$$

2. (2) The transversality conditions:
   $\psi_0 \geq 0$ the (n+1)-dimensional vector

$$\left\{\psi_1(T) - \psi_0\frac{\partial F}{\partial x_1}\left(T, x^*(T)\right), \ldots, \psi_n(T) = \psi_0\frac{\partial F}{\partial x_n}\left(T, x^*(T)\right), -H\left(T, \psi(T), x^*(T), u^*(T)\right) - \psi_0\frac{\partial F}{\partial T}\left(T, x^*(T)\right)\right\}$$

   is orthogonal to the manifold M at the point $(T, x^*(T))$
3. (3) The maximum condition:

$$\max_{u \in U} H(t, \psi(t), x^*(t), u) = H(t, \psi(t), x^*(t), u^*(t)).$$

In the case of fixed time and free end, the transversality conditions reduce to

$$\psi_0 \geq 0;\ \psi_i(T) = \psi_0\frac{\partial F}{\partial x_i}(t, x^*(T)),\ (i = 1, \ldots, n)$$

(if F=0, we have $\psi_i(T) = 0,\ (i = 1, \ldots, n)$). Let us remark that the vector ψ is defined up to multiplication by a positive constant, and in the case $\psi_0 \neq 0$ it can be normalized by dividing by ψ_0. As soon as the optimal value u^* at the point (t, x) depends only on that point, we can seek the optimal control as a feedback control, i.e. in the form $u^*=u^*(t, x)$. This function defines the optimal control at each point of the space (t, x), and thus it is called the optimal synthesis. The variables H and ψ can also regarded as functions of t and x. Let us denote the optimal value of the functional, corresponding to the initial point (t, x), by $x_0(t, x)$. This function is called Bellman's function.

The main idea of the economic interpretation of Pontryagin's maximum principle (which goes back to L.V. Kantorovich) is to consider the variables ψ_i as shadow prices. To explain, let us assume that the problem is regular $(\psi_0 \neq 0)$, and that the optimal synthesis $u^*(t, x)$ and the dual variables $\psi(t, x)$ are smooth. In this case $x_0(t, x)$ is also smooth and Bellman's equation

$$\max_{u \in U}\left\{\frac{\partial x_0}{\partial t}(t, x) + f(t, x, u) + \sum_{\alpha=1}^{n}\frac{\partial x_0}{\partial x_\alpha}(t, x)\varphi_\alpha(t, x)\right\} = 0$$

is fulfilled. The relationship between Bellman's equation and Pontryagin's maximum principle can be expressed by the equations

$$\frac{\partial x_0}{\partial x_i}(t, x) = \frac{\psi_i(t, x)}{\psi_0(t, x)}; \frac{\partial x_0}{\partial t}(t, x) = -\frac{H(t, x)}{\psi_0(t, x)},$$

i.e. the normalized value of the dual variable ψ_i / ψ_0 is the marginal effect of the factor x_i on the optimal value of the functional x_0 and that is exactly the shadow price of x_i. The economic meaning of Pontryagin's maximum principle is as follows. For the optimal process $u^*(t), x^*(t)$ there exist shadow prices ψ the adjoint equations and the transversality conditions being fulfilled, such that the optimal value of the control $u^*(t)$ at each moment t maximizes the flow of the profit, which is calculated in accordance with the shadow price. It is worth remembering, that in the irregular case $(\psi_0 = 0)$, as well as in the case of discontinuous optimal synthesis $u^*(t, x)$ Bellman's function $x_0(t, x)$ is often nonsmooth, in spite of all the functions defining the statement of the problem being smooth. Bellman's equation breaks down, but Pontryagin's maximum principle is fulfilled. In that case the notion of ‘prices’ loses its natural meaning. The search for general enough conditions, guaranteeing the smoothness of Bellman's function, is a difficult and only partially explored mathematical problem.

The creation of Pontryagin's principle stimulated the two aforementioned lines of macroeconomics. Before listing the corresponding results, let us note some significant obstacles in the way of application of these optimizing methods to mathematical economics. To formulate an optimization problem, we have to choose a criterion. It is only natural to take as a criterion some function of the final state $x(T)$ or the profits over some interval of the time $[t_0,T]$. But the choice of the moment T (the horizon of the plan) is arbitrary from an economic point of view. Meanwhile it is highly desirable to define economically reasonable behaviour independently of such arbitrariness. Two approaches to overcome this obstacle are known–that of F.P. Ramsey (and his collaborator J.M. Keynes) and that of J. von Neumann.

Ramsey's approach is to take eternity as the horizon of the plan. He applied the calculus of variations, which can be regarded as a version of Pontryagin's maximum principle, to the problem of resource allocation between consumption and saving, aiming to maximize the benefits of society during the entire infinite period of its possible existence, and proved the Golden Rule of saving. From the mathematical point of view the problem is to minimize the integral over the half-open interval $[t_0, \infty]$ from the difference between absolute welfare (Bliss) and immediate welfare (the utility function), which tends to Bliss and depends on the solution of the differential equation containing the policy of saving as control. The principal part of the right-hand side of this equation is the production function. The naivety of this model lies in the conception of stationary and absolutely stable economic Universe, rather than in the assumption of the possibility of complete aggregation. I hope to be indulged in using such unusual (in economic context) terms as ‘Universe’. But in fact we deal with the closed macroeconomic models, purporting to describe all basic economic phenomena, and in this sense the situation is closely related to that of physical models of the Universe; hence the reason for the proposed usage.

Later on there were attempts to modernize this model and to make it nonstationary. On the one hand, Hicks, Harrod and Solow among others varied the production function, aiming to include in it the effect of technical change. On the other hand, T.C. Koopmans broke off the relationship between the production and the utility functions (which was so essential in the views of Ramsey) and introduced a discounting factor in the integrand which guaranteed the convergence of the functional for any choice of control. The stationarity of Ramsey's economic Universe was slightly shaken.

The path of the Golden Rule in such models appears to be a singular path of Pontryagin's maximum principle. The path of the principle is called singular, if the maximum of the Pontryagin's function (3°) at all points of this path is attained at several distinct points of the set U. In the case of problems which are linear in the control the non-singular (band-bang) control uses only the extreme points of the set U. Such a control corresponds to the economic policy with sharp changes (switches). The characteristic feature of bang-bang optimal control is instantaneous switches from one vertex of the polyhedron U onto another. On the contrary, the singular control, using the internal points of U, as a rule does not need switches and in this sense seems to be more acceptable than the bang-bang control from the economic point of view.

The application of Pontryagin's principle to such models calls for its generalization to the case of infinite time intervals. More

precisely, the transversality conditions at infinity need generalization, and only those. But even that question turns out to be difficult enough. There were incorrect attempts to formulate these conditions in the form: $\psi(t) \to 0$ for $t \to \infty$. For the case of linear differential constraints Aubin and Clarke found out and proved a correct version of transversality conditions, which requires the convergence of the integral of $|\psi(t)|^q e^{\delta t}$ for some *q* and δ. A complete and correct solution of the problem of transversality conditions at infinity for nonlinear differential constraints is still absent.

Another line of evolution of macroeconomics begins with the work of von Neumann. He introduced a discrete model of the expansion of production, which was defined by two given matrices – that of input and that of output. Von Neumann seeks the balanced development of economics when the input vector is proportional to the output one. Among these rays he seeks those yielding maximal growth. He introduces dual variables – prices of the optimal plan – and gave optimality conditions in dual terms.

Later on it turned out (in accordance with the hypothesis of Samuelson) that the ray of maximal growth (now called) ‘Neumann's ray’ or ‘the turnpike’) plays the leading part in exploring the optimal paths of this model. The corresponding theorems (turnpike theorems) assert that this ray defines the asymptotic behaviour of the optimal paths, in the increase of the horizon of the plan (for $T \to \infty$), independently of the choice of terminal functional, and it is precisely this fact which makes it possible to overcome the aforementioned obstacle. Continuous versions of the turnpike theorems, which require Pontryagin's principle or the related methods for their proofs, were obtained by, among others, A.N. Ducalov, A.E. Ilutovich and L.F. Zelikina. Let us note that the turnpike in these models is the singular path of the principle. In the work of Zelikina, for certain optimal resource allocation problems, Neumann's concept of optimal policy, independent of the choice of functional, was brought to its logical conclusion. By constructing the optimal synthesis in *n*-dimensional space, it was shown that the shadow prices for increasing (with $T \to \infty$) initial segments of optimal paths are invariant relative to such a choice. The search for a complete system of invariants of optimal synthesis relative to a choice of functional (in some appropriate class of the latter) for the general economical optimizational problem still remains an open question.

A hint for an infinite-dimensional version of the theory of duality is contained in the formulation of Pontryagin's maximum principle itself. A.M. Ter-Krikorov (1977) introduces the dual problem to the linear optimizational problem, in which the dual variables ψ_i turn into the state variables and vice versa.

The techniques of turnpike theorems and Pontryagin's principle gives rise to a series of optimization models, which become more and more universal and finally develop into the concept of expanding economic Universe. The economic analogue of the physical concept of the oscillating Universe is yet developed only on the phenomenological level, in spite of the empirical evidence for the corresponding economic phenomenon. The reason for this fact is (as it seems) the lack of satisfactory optimization models taking into account the specific effect of money.

It is worth noting that, like the physical models of Universe, all these concepts of economic Universe are devoid of its substance – thought and ethics – which naturally bring the question out of the competence of pure economics. But without this substance, the economical Universe, as soon as it claims to be universal, cannot be explained in principle.

See Also

- Hamiltonians
- stochastic optimal control

Bibliography

Aubin, J.P. and Clarke, F.H. 1979. Shadow prices and duality for a class of optimal control problems. *SIAM Journal of Control and Optimization* 17(5), September, 567–86.

Diukalov, A.N. and Iliutovich, A.E. 1973. Asimptoticheskie svoistva optimal’nikh traektorii ekonomicheskoi dinamiki (Asymptotic properties of optimal trajectories in economic dynamic). *Automatika i Telemakhanika* No. 3.

Harrod, R.F. 1937. Remark. *American Economic Review* 27.

Hicks, J.R. 1932. *The Theory of Wages*. London: Macmillan.

Koopmans, T.C. 1957. *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38, December, 543–59.

Solow, R.M. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, August, 312–20.

Ter-Krikorov, A.M. 1977. *Optimal'noe Upravlenie i Matematicheskaia Ekonomika*. Moscow: Nauka.

von Neumann, J. 1945–6. A model of general economic equilibrium. *Review of Economic Studies* 13(1), 1–9.

Zelikina, L.F. 1975. Universal manifolds and turnpike theorems for a class of optimal control problems. *Soviet Matematics Doklady* 16(5), 1136–40.

Zelikina, L.F. 1977. Mnogomernyi sintez i teoremy o magistrali v zadachkh optimal'nogo upravleniia (Multi-dimensional feed-back control and the turnpike theorems in the optimal control problems). In *Veroiatnostnye Problemy Upravleniia v Ekonomike* (Stochastic problems of control in economics), Moscow: Nauka.

How to cite this article

Zelikin, M. I. "Pontryagin's principle of optimality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000118> doi:10.1057/9780230226203.1304

The New Palgrave Dictionary of Economics Online

Poor Law, new

George R. Boyer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The New Poor Law refers to the system of local public assistance in England and Wales initiated by the passage of the 1834 Poor Law Amendment Act. This act attempted to restrict relief outside of workhouses for the able-bodied, but was evaded for three decades. The Crusade against Outrelief of the 1870s marked a major shift in administration and the increased use of workhouse relief. Numbers on relief fell sharply thereafter, although the elderly continued to rely heavily on the Poor Law. The Liberal welfare reforms of 1906–11 paved the way for the 1948 abolition of the Poor Law.

Keywords

Charity Organization Society; New Poor Law; outdoor relief; poor rate

Article

The New Poor Law refers to the welfare policy in England and Wales initiated by the passage of the Poor Law Amendment Act in 1834. All destitute individuals were eligible for poor relief from their local Poor Law union. Those granted assistance were either given cash or in-kind payments in their homes (outdoor relief) or were relieved in workhouses (indoor relief). Although the Poor Law remained in existence until 1948, the Crusade against Outrelief in the 1870s and the adoption of the Liberal welfare reforms in the decade before the First World War significantly reduced its role as a safety net for the poor.

The Poor Law Amendment Act was an outgrowth of the Report of the Royal Commission to Investigate the Poor Laws (1834), which called for sweeping reforms to the existing system of poor relief, including the grouping of parishes into Poor Law unions, the abolition of relief for the able-bodied and their families outside workhouses, and the appointment of a centralized Poor Law Commission to direct the administration of relief. The Act implemented some of the report's recommendations, but left the regulation of outdoor relief to the Poor Law Commissioners.

By 1839 most rural parishes had been grouped into Poor Law unions, which had built or were building

workhouses. However, the Poor Law Commission met with strong opposition when it attempted to set up unions in the industrial north, and the implementation of the New Poor Law was delayed in several industrial cities. The Commission and its 1847 replacement, the Poor Law Board, issued orders in 1842, 1844 and 1852 to restrict the payment of outdoor relief to able-bodied males, but these were evaded by both rural and urban unions. Thus, while real per capita relief expenditure fell by 43 per cent from 1831 to 1841, and remained at least 20 per cent below its 1831 level for the remainder of the 19th century (see Table 1), many Poor Law unions continued to grant outdoor relief to needy able-bodied males after 1834 (Rose, 1970; Digby, 1978). Data for three London parishes and six provincial towns in the years around 1850 indicate that large numbers of prime-age males continued to apply for relief, and that a majority of those assisted were granted outdoor relief (Lees, 1998). The Poor Law played an important role in assisting the unemployed and their families in urban districts during cyclical downturns (Boot, 1990; Boyer, 2004). Moreover, the New Poor Law, like its predecessor, provided a major source of support for the non-able-bodied poor. From the 1840s to the 1860s, in much of rural England a large share of those aged 70 and over received regular poor relief payments, although these often did not provide full maintenance (Thomson, 1984).

Relief expenditures and numbers on relief, 1831–1936.

Year	Expenditures on Relief (1,000 £s)	Real expenditure per capita 1831=100	Number relieved (official) 1,000s	Share of population relieved (official)	Number relieved (revised) 1,000s	Share of population relieved (revised)	Share of paupers relieved indoors
1831	6,799	100.0					
1836	4,718	75.1					
1841	4,761	57.2					
1846	4,954	64.3					
1851	4,963	62.8	941	5.3	2,108	11.9	12.1
1856	6,004	57.5	917	4.9	2,054	10.9	13.6
1861	5,779	55.6	884	4.4	1,980	9.9	13.2
1866	6,440	60.2	916	4.3	2,052	9.7	13.7
1871	7,887	67.9	1,037	4.6	2,323	10.3	14.2
1876	7,336	58.2	749	3.1	1,678	7.0	18.1
1881	8,102	64.0	791	3.1	1,772	6.9	22.3
1886	8,296	66.7	781	2.9	1,749	6.4	23.2
1891	8,643	66.9	760	2.6	1,702	5.9	24.0
1896	10,216	78.4	816	2.7	1,828	6.0	25.9
1901	11,549	78.5	777	2.4	1,671	5.2	29.2
1906	14,036	89.8	892	2.6	1,918	5.6	31.1
1911	15,023	86.7	886	2.5	1,905	5.3	35.1

1921	31,925	69.7	627	1.7	35.7
1926	40,083	118.9	1,331	3.4	17.7
1931	38,561	124.0	1,090	2.7	21.5
1936	44,379	153.5	1,472	3.6	12.6

Notes: Relief expenditure data are for the year ended on 25 March. In calculating real per capita expenditures, I used cost of living and population data for the previous year.

Sources: Columns 1, 3, 4, and 7 from Williams (1981). Estimates in columns 5 and 6 constructed by the author following Lees (1998). Estimates in column 2 constructed by the author.

Data on the number of persons receiving poor relief are available for two days a year, 1 January and 1 July, beginning in 1849; the official estimates of the annual number relieved in Table 1 are the average of the number relieved on these two dates. Studies conducted by Poor Law administrators in 1892 and 1906–07 found that the day counts significantly underestimated the number assisted during the year. The ‘revised’ estimates in Table 1 are based on these studies, and assume that the ratio of actual to counted paupers was 2.24 for 1851–96 and 2.15 for 1901–11. These estimates indicate that from 1850 to 1870 about ten per cent of the population was assisted by the Poor Law each year. Lees (1998) contends that over a three-year period as much as 25 per cent of the population made use of the Poor Law.

Relief expenditures were financed by a local property tax, known as the poor rate. Up to 1865, each parish within a Poor Law union was responsible for relieving its own poor. As a result, tax rates were often significantly different across parishes within Poor Law unions, and were especially high in working-class districts. Economic crises put enormous financial strain on parishes that were already poor. The ‘basic weaknesses’ of the poor relief system were exposed in the 1860s, when the Poor Law ‘was subjected to an almost continual series of shocks’ (Rose, 1981). The two major shocks of the decade were the Lancashire cotton famine of 1862–4 and the East London crises of 1860–1 and 1867–9. The collapse of raw cotton imports from the United States during the American Civil War forced Lancashire cotton textile factories to shut down or severely curtail production. The resulting unemployment caused a huge increase in demand for relief, which the hardest-hit parishes were unable to meet, and led several Poor Law unions to appeal to private relief committees for charitable assistance. During the severe winters of 1860–1, 1867–8 and 1868–9, Poor Law unions in London's East End were also forced to turn to private charities for assistance in meeting the high demand for relief.

The problems associated with Poor Law finance led parliament to adopt the Union Chargeability Act in 1865, and similar acts relating to London in 1867, 1869 and 1870. These acts placed the cost of poor relief on the Poor Law union rather than on each parish within it, and thus shifted a large share of the cost of relief from working-class parishes (which had low tax bases and many paupers) to middle-class parishes (with higher tax bases and fewer paupers). The tax-shifting eased the financial burdens that had plagued the Poor Law, but also led to the revolt of middle-class taxpayers in many areas.

The Union Chargeability Act was one of the catalysts of the Crusade against Outrelief in the early 1870s. Encouraged by the Local Government Board (LGB), Poor Law unions throughout England and Wales curtailed outdoor relief for all types of paupers. In December 1871 the LGB issued a circular concluding that generous outdoor relief was destroying self-reliance among the poor. In the circular's words: ‘a certainty of obtaining outdoor relief in his own home whenever he may ask for it extinguishes in the mind of the labourer all motive for husbanding his resources, and induces him to rely exclusively

upon the rates instead of upon his own savings for such relief as he may require' (quoted in Englander, 1998, p. 107). The Charity Organization Society (COS), founded in 1869, aided the Board in convincing the public of the need for reform. It argued that most low-skilled workers earned enough to be able to set aside some income in anticipation of future interruptions in earnings caused by unemployment or sickness. The LGB and the COS maintained that the restriction of outdoor relief would improve the moral and economic condition of the poor in the long run. The COS also believed that most applicants for relief would refuse to enter workhouses and would remove themselves from relief roles, so that a shift from outdoor to workhouse relief would significantly reduce Poor Law expenditures. Most Poor Law unions found it difficult to resist a policy that promised to raise the morals of the poor *and* reduce taxes (MacKinnon, 1987).

The effect of the Crusade against Outrelief can be seen in Table 1. Real per capita relief expenditures and the share of the population receiving relief both fell sharply from 1871 to 1876. The decline in numbers on relief was largely a result of the deterrent effect of the workhouse: as the COS predicted, many of those offered indoor relief refused it. From 1871 to 1881, the number of paupers receiving outdoor relief fell by 282,000 (a 33 per cent decline), while the number relieved in workhouses rose by only 21,000.

Real per capita relief expenditures increased after 1876, mainly because the Poor Law provided increasing amounts of medical care for the poor. Otherwise, the role played by the Poor Law declined in the last quarter of the 19th century. The share of the population receiving relief fell from seven per cent in 1876 (revised estimates) to 5.2 per cent in 1901. The decline was due in large part to improvements in living standards, which increased workers' ability to save and to join friendly societies – mutual help associations providing sickness, accident, death, and (sometimes) old age benefits. However, part of the decline in numbers on relief was a result of the Crusade against Outrelief and of a change in the attitude of the poor towards relief. Prior to 1870, a large share of the working class regarded access to public relief as an entitlement, although they rejected the workhouse as a form of relief. Partly as a result of COS propaganda, by the end of the century most within the working class viewed poor relief as stigmatizing, and went to great lengths to avoid applying for relief. Thus, the decline in the share of the population receiving poor relief from 1871 to 1901 overestimates the decline in the share living in poverty.

One section of the working class continued to rely heavily on the Poor Law – the elderly. Table 2 shows that, for the 12-month period from March 1891 to March 1892, 29.3 per cent of those aged 65 and over received poor relief, as compared with 5.1 per cent of children and 3.7 per cent of those aged 16–64. Most elderly paupers received only partial maintenance, which they combined with wage income, savings, friendly society or trade union benefits, and help from relatives or friends to achieve a subsistence income. The ability of the elderly to support themselves declined with age; Booth (1894) estimated that 40 per cent of those aged 75 and over received poor relief, as compared to 20 per cent of those aged 65–70.

Pauperism in the early 1890s: 1 January 1892 and March 1891–March 1892

		1 January			12 months' count		
Ages	Population 1891	Indoor paupers %	Outdoor paupers %	Total paupers %	Indoor paupers %	Outdoor paupers %	Total paupers %

Under 16	10,762,808	0.5	1.6	2.1	1.0	4.1	5.1
16-64	16,867,116	0.5	0.7	1.2	1.4	2.3	3.7
65 and older	1,372,601	4.6	14.9	19.5	8.3	21.0	29.3
Total	29,002,525	0.7	1.7	2.4	1.3	3.8	5.4

Source: Report of the Royal Commission on the Aged Poor, Parliamentary Papers (1895), vol. XIV, pp. xii–xiii.

Despite improvements in living standards, many manual workers still experienced ‘acute financial’ distress at some point in their lives (Johnson, 1985). The inability of low-skilled workers to protect themselves from financial insecurity was the catalyst for the Liberal welfare reforms, several pieces of social welfare legislation adopted between 1906 and 1911. Acts of 1906 and 1907 provided free meals and medical inspections (later treatment) for needy schoolchildren. The 1908 Old Age Pension Act granted weekly pensions to persons aged 70 and over whose annual income was below a certain level, and the National Insurance Act of 1911 established compulsory systems of health insurance (covering all manual workers) and unemployment insurance (covering workers in a limited number of industries). The Liberal welfare reforms provided assistance to the working class that was outside the Poor Law and therefore did not involve ‘the stigma of pauperism’, and they paved the way for the eventual abolition of the Poor Law.

During the inter-war period the Poor Law served as a residual safety net, assisting those who fell through the cracks of existing social insurance policies. A large share of those on relief, especially in the mid-1920s, were unemployed workers who either did not qualify for unemployment benefits or had exhausted their benefits. The Local Government Act of 1929 abolished the Poor Law unions, and transferred the administration of poor relief to the counties and county boroughs. Finally, from 1945 to 1948, Parliament adopted a series of laws that together formed the basis for the welfare state, and made the Poor Law redundant. The National Assistance Act of 1948 officially repealed all existing Poor Law legislation.

See Also

- Poor Law, old
- poverty
- poverty alleviation programmes
- social insurance
- social insurance and public policy
- welfare state

Bibliography

Boot, H.M. 1990. Unemployment and poor law relief in Manchester, 1845–50. *Social History* 15, 217–28.

Booth, C. 1894. *The Aged Poor in England and Wales*. London: Macmillan.

Boyer, G.R. 1990. *An Economic History of the English Poor Law, 1750–1850*. Cambridge: Cambridge University Press.

Boyer, G.R. 2004. The evolution of unemployment relief in Great Britain. *Journal of Interdisciplinary History* 34, 393–433.

Brundage, A. 2002. *The English Poor Laws, 1700–1930*. New York: Palgrave.

Crowther, M.A. 1981. *The Workhouse System, 1834–1929: The History of an English Social Institution*. London: Batsford.

Digby, A. 1978. *Pauper Palaces*. London: Routledge & Kegan Paul.

Englander, D. 1998. *Poverty and Poor Law Reform in Britain: From Chadwick to Booth, 1834–1914*. London: Addison Wesley Longman.

Fraser, D. 1976. *The New Poor Law in the Nineteenth Century*. London: Macmillan.

Humphreys, R. 1995. *Sin, Organized Charity, and the Poor Law in Victorian England*. New York: St Martin's.

Humphreys, R. 2001. *Poor Relief and Charity, 1869–1945: The London Charity Organization Society*. New York: Palgrave.

Johnson, P. 1985. *Saving and Spending: The Working-class Economy in Britain, 1870–1939*. Oxford: Clarendon Press.

Kidd, A. 1999. *State, Society and the Poor in Nineteenth-Century England*. London: Macmillan.

Lees, L.H. 1998. *The Solidarities of Strangers: The English Poor Laws and the People, 1770–1948*. Cambridge: Cambridge University Press.

MacKinnon, M. 1986. Poor law policy, unemployment, and pauperism. *Explorations in Economic History* 23, 299–336.

MacKinnon, M. 1987. English poor law policy and the crusade against outrelief. *Journal of Economic History* 47, 603–25.

Rose, M.E. 1970. The new poor law in an industrial area. In *The Industrial Revolution*, ed. R.M. Hartwell. Oxford: Oxford University Press.

Rose, M.E. 1981. The crisis of poor relief in England, 1860–1890. In *The Emergence of the Welfare State in Britain and Germany, 1850–1950*, ed. W.J. Mommsen. London: Croom Helm.

Rose, M.E. 1985. *The Poor and the City: The English Poor Law in its Urban Context*. Leicester: Leicester University Press.

Thomson, D. 1984. The decline of social security: falling state support for the elderly since early Victorian times. *Ageing and Society* 4, 451–82.

Webb, S. and Webb, B. 1929. *English Poor Law History*. Part II: *The Last Hundred Years*, 2 vols. London: Longmans.

Williams, K. 1981. *From Pauperism to Poverty*. London: Routledge.

How to cite this article

Boyer, George R. "Poor Law, new." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000210> doi:10.1057/9780230226203.1305

The New Palgrave Dictionary of Economics Online

Poor Law, old

George R. Boyer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Old Poor Law was the system of local public assistance that existed in England and Wales from 1597 until 1834. It provided an important safety net for labouring households that were unable to protect themselves against income loss, assisting the elderly, widows, children, the sick, and the unemployed. Relief expenditures increased sharply from 1750 to 1820, as did the share of paupers who were adult able-bodied males. Parliament responded to the increase in spending with the Poor Law Amendment Act (1834), which recommended that poor relief be granted to able-bodied males and their families only in workhouses.

Keywords

Chadwick, E.; enclosures; labour mobility; Malthus, T.; Old Poor Law; Poor Law Amendment Act 1834; poor rate; poverty alleviation programmes; Senior, N; Settlement Law; Smith, A.; Speenhamland system

Article

The Old Poor Law was the system of public assistance in England and Wales from the Tudor era through the passage of the Poor Law Amendment Act in 1834. Parliamentary acts of 1597–98 and 1601 (43 Eliz. I c. 2) established a compulsory system of poor relief administered and financed at the parish (local) level. Overseers of the poor assessed a compulsory property tax, known as the poor rate, to assist those within the parish ‘having no means to maintain’ themselves. The overseers were to put the able-bodied poor to work, give apprenticeships to poor children, and provide ‘competent sums of money’ to relieve the aged or non-able-bodied.
The Elizabethan Poor Law was an attempt by Parliament both to prevent starvation and to ensure public order. It was adopted in response to a sharp deterioration in workers’ living standards in the 16th century, combined with a decline in traditional forms of charitable assistance. The dissolution of the

monasteries, religious guilds, almshouses, and hospitals under Henry VIII had eliminated many of the traditional sources of charity for the poor.

The Settlement Act of 1662 stated that individuals were guaranteed relief only in their parish of settlement (typically their parish of birth). The act gave parishes the right to remove within 40 days of arrival any newcomer deemed ‘likely to be chargeable’ as well as any non-settled applicant for relief. Adam Smith believed that the Settlement Law put a serious brake on labour mobility, but available evidence suggests that parishes used it selectively, to keep out economically undesirable migrants such as single women, older workers and men with large families. The Removal Act of 1795 amended the Settlement Law so that no non-settled person could be removed from a parish unless he or she applied for relief.

The Old Poor Law constituted ‘a welfare state in miniature’, relieving the elderly, widows, children, the sick, the disabled, and the unemployed and underemployed (Blaug, 1964). It provided an important safety net for labouring households who were unable to accumulate enough savings to protect themselves against income loss. While only a small share of the labouring population received relief at any point in time, the life-cycle nature of poverty meant that a much larger share required Poor Law assistance during their lifetimes. Slack (1990) estimates that in the late 18th century one-fifth or more of the inhabitants of a typical parish received poor relief over a five-year period. In years of exceptionally high food prices, the share on relief could exceed 25 per cent.

During the 17th century the bulk of relief recipients were elderly, orphans, or widows with young children. In many parishes a majority of those collecting regular weekly pensions were aged 60 or older. Female pensioners far outnumbered males. On average, the payment of weekly pensions made up about two-thirds of relief spending; the remainder went to casual benefits, often to able-bodied males in need of short-term relief because of sickness or unemployment.

Growth in relief expenditures, 1750 – 1820

The 18th century witnessed an explosion in relief expenditures, as can be seen from Table 1. Real per capita expenditures increased by 80 per cent from 1696 to 1748–50, more than doubled from 1750 to 1803, and then remained at a high level until the Poor Law was amended in 1834. Relief expenditures increased from 0.8 per cent of GDP in 1696 to a peak of 2.7 per cent of GDP in 1818–20. The demographic characteristics of the ‘pauper host’ changed considerably in the late 18th and early 19th centuries, especially in the rural south and east of England. There was a sharp increase in numbers receiving casual benefits, as opposed to regular weekly pensions. The share of paupers aged 20–59 increased significantly, and the share aged 60 and over declined. Finally, the share of relief recipients in the south and east who were male increased from about a third in 1760 to nearly two-thirds in 1820. In the north and west there also were shifts toward prime-age males and casual relief, but the magnitude of these changes was far smaller than elsewhere.

Poor relief expenditures 1696–1841

Year	Expenditures on relief (£1,000s)	Real expenditures per capita, 1803=100	Expenditures as % of GDP
1696	400	24.9	

1748–50	690	45.8	1.0
1776	1,530	64.0	1.6
1783–85	2,004	75.6	1.8
1803	4,268	100.0	2.2
1813	6,656	91.8	2.6
1818	7,871	116.8	
1821	6,959	113.6	2.7
1826	5,929	91.8	
1831	6,799	107.9	2.0
1836	4,718	81.1	
1841	4,761	61.8	1.1

Note: Relief expenditure data are for the year ended on 25 March. In calculations of real per capita expenditures, cost of living and population data for the previous year were used.

Sources: Data in column 1: Slack (1990: 30) and Mitchell (1988: 605). Data in column 2: author's calculations. Data in column 3: Lindert (1998: 114).

What caused the sharp increase in the number of able-bodied males on relief? In the second half of the 18th century, a large share of rural households in southern England suffered significant declines in real income, resulting from the combination of a decline in agricultural labourers' real wage rates, an increase in seasonal unemployment, a decline in employment opportunities for women and children in cottage industry and, in some villages, the loss of access to land for growing food, grazing animals, and gathering fuel (common rights) as a result of enclosures. The situation was different in the north and midlands, where real wages of day labourers in agriculture increased from 1770 to 1820. Moreover, while some areas experienced a decline in cottage industry, in Lancashire and the West Riding of Yorkshire the concentration of textile production led to increased employment opportunities for women and children.

Forms of relief and regional differences in relief spending

Relief for able-bodied males and their families took various forms, the most important of which were: allowances-in-aid-of-wages (the so-called Speenhamland system), child allowances for labourers with large families, and payments to seasonally unemployed agricultural labourers. Under the allowance system, a household head (whether employed or unemployed) was guaranteed a minimum weekly income, the level of which was determined by the price of bread and by the size of his or her family. The most famous allowance scale was that adopted by Berkshire magistrates at Speenhamland in May 1795. Such scales typically were instituted during years of high food prices, such as 1795–6 and 1800–1, and removed when prices declined. Child allowance payments were widespread in the rural south, which suggests that labourers' wages were too low to support large families. The typical parish paid a small weekly sum to labourers with four or more children under age 10 or 12. Seasonal unemployment had been a problem for agricultural labourers long before 1750, but the extent of seasonality increased in the

second half of the 18th century as farmers in southern and eastern England responded to the sharp increase in grain prices by increasing their specialization in grain production. The increase in seasonal unemployment, combined with the decline in other sources of income, forced many agricultural labourers to apply for poor relief during the winter.

Table 2 reports data for 15 counties located throughout England on per capita relief expenditures for the years ending in March 1803, 1812 and 1831, and on relief recipients in 1802–3. Per capita expenditures were higher on average in agricultural counties than in more industrial counties, and were especially high in the grain-producing south-eastern counties. The share of the population receiving poor relief in 1802–3 varied significantly across counties, being 15–23 per cent in the grain-producing south and less than 10 per cent in the north. The demographic characteristics of those relieved also differed across regions. The share of relief recipients who were elderly or disabled was higher in the north and west than in the south, while the share who were able-bodied was higher in the south-east than elsewhere. These regional differences in relief expenditures and numbers on relief largely were caused by differences in economic circumstances; poverty was more of a problem in the agricultural south and east than it was in the pastoral south-west or in the more industrial north (Blaug, 1963; Boyer, 1990). Recently, King (2000, pp. 267–8) has argued that the regional differences in poor relief were determined by ‘very different welfare cultures on the part of both the poor and the poor law administrators’.

County-level poor relief data, 1802–1831

County	Per capita relief spending (shillings per year) 1802–3	Per capita relief spending (shillings per year) 1812	Per capita relief spending (shillings per year) 1831	% of population relieved 1802–3	Share of recipients over 60, or Disabled 1802–3
North					
Durham	6.5	9.9	6.8	9.3	22.8
Northumberland	6.7	7.9	6.3	8.8	32.2
Lancashire	4.4	7.4	4.4	6.7	15.0
West Riding	6.5	9.9	5.6	9.3	18.1
Midlands					
Stafford	6.9	8.5	6.5	9.1	17.2
Nottingham	6.3	10.8	6.5	6.8	17.3
Warwick	11.3	13.3	9.6	13.3	13.7
South-east					
Oxford	16.2	24.8	16.9	19.4	13.2
Berkshire	15.1	27.1	15.8	20.0	12.7
Essex	12.1	24.6	17.2	16.4	12.7
Suffolk	11.4	19.3	18.3	16.6	11.4
Sussex	22.6	33.1	19.3	22.6	8.7

South-west					
Devon	7.3	11.4	9.0	12.3	23.1
Somerset	8.9	12.3	8.8	12.0	20.8
Cornwall	5.8	9.4	6.7	6.6	31.0
England and Wales	8.9	12.8	10.1	11.4	16.0

Sources: Data for columns 1–3: Blaug (1963: 178–9). Data for columns 4–5: *Abstract of Returns relative to the Expense and Maintenance of the Poor*, H.C. 175 (1803–4), xiii.

Political economy of poor relief

From 1795 to 1834 relief expenditures as a share of national product were significantly higher in England than on the European continent. However, differences in spending between England and the continent were relatively small before 1795 and after 1834 (Lindert, 1998). The increase in relief spending in late 18th and early 19th century England overstates the increase in poverty, because it was partly a result of politically dominant farmers taking advantage of the poor relief system to shift some of their labour costs onto other taxpayers (Boyer, 1990). Most rural parish vestries were dominated by labour-hiring farmers as a result of the system of plural voting introduced by Gilbert's Act in 1782 and extended in 1818 by the Parish Vestry Act, which gave large property holders (typically labour-hiring farmers) up to six votes in local elections. Relief expenditures were financed by a tax levied on all parishioners whose property value exceeded some minimum level. A typical rural parish's taxpayers can be divided into two groups: labour-hiring farmers and non-labour-hiring taxpayers (tithe recipients, family farmers, shopkeepers, and artisans).

In grain-producing areas, where there were large seasonal variations in the demand for labour, labour-hiring farmers anxious to secure an adequate peak season labour force were able to reduce costs by laying off unneeded workers during slack seasons and having them collect poor relief. Tithe recipients and other non-labour-hiring taxpayers paid part of the relief benefits that went to seasonally unemployed labourers. Thus, some share of the increase in relief spending in the early 19th century represented a subsidy to labour-hiring farmers rather than a transfer from farmers and other taxpayers to agricultural labourers and their families. In pasture farming areas, where the demand for labour was fairly constant over the year, it was not in farmers' interests to shed labour during the winter, and the number of able-bodied labourers receiving casual relief was smaller.

Reform of the Poor Law

The sharp increase in relief spending after 1780 sparked a major debate on the Poor Laws. Most participants in the debate were critical of the granting of outdoor relief to able-bodied males, on the grounds that such aid created serious work disincentives. Among the sharpest critics was Thomas Malthus, who argued in *An Essay on the Principle of Population* (1798, pp. 40–1) that the Poor Laws, by guaranteeing parish assistance to able-bodied labourers, 'diminish both the power and the will to save among the common people, and thus … weaken one of the strongest incentives to sobriety and industry, and consequently to happiness'.

In 1832 the government appointed the Royal Commission to Investigate the Poor Laws to examine the operation of the Poor Law and suggest methods for improving the administration of relief. The commission's report (1834, pp. 261–3), written by economists Nassau Senior and Edwin Chadwick, called for sweeping reforms, including the abolition of outdoor relief for the able-bodied and their families. The report urged the adoption of a policy of ‘less eligibility’ whereby the condition of paupers would be worse than that of the lowest-paid independent labourers. To achieve this, Senior and Chadwick recommended that relief should be granted to able-bodied labourers and their families only in well-regulated workhouses; they predicted that the use of workhouses would restore the industry and ‘frugal habits’ of the poor, and improve their ‘moral and social condition’.
The era of the Old Poor Law ended with the adoption of the Poor Law Amendment Act of 1834, which implemented many of the report's recommendations.

See Also

- Eden, Frederick Morton
- Malthus, Thomas Robert
- Poor Law, new
- poverty alleviation programmes
- Senior, Nassau William
- welfare state

Bibliography

Blaug, M. 1963. The myth of the Old Poor Law and the making of the New. *Journal of Economic History* 23, 151–84.

Blaug, M. 1964. The Poor Law Report re-examined. *Journal of Economic History* 24, 229–45.

Boyer, G. 1990. *An Economic History of the English Poor Law, 1750–1850*. Cambridge: Cambridge University Press.

King, S. 2000. *Poverty and Welfare in England, 1700–1850: A Regional Perspective*. Manchester: Manchester University Press.

Lees, L. 1998. *The Solidarities of Strangers: The English Poor Laws and the People, 1770–1948*. Cambridge: Cambridge University Press.

Lindert, P. 1998. Poor relief before the welfare state: Britain versus the Continent, 1780–1880. *European Review of Economic History* 2, 101–40.

Malthus, T. 1798. *An Essay on the Principle of Population*. Oxford: Oxford University Press, 1999.

Mitchell, B. 1988. *British Historical Statistics*. Cambridge: Cambridge University Press.

Royal Commission to Investigate the Poor Laws. 1834. *Report on the Administration and Practical Operation of the Poor Laws*. London: B. Fellowes.

Slack, P. 1990. *The English Poor Law, 1531–1782*. London: Macmillan.

Webb, S. and Webb, B. 1927. *English Poor Law History. Part I: The Old Poor Law*. London: Longmans.

Williams, K. 1981. *From Pauperism to Poverty*. London: Routledge.

How to cite this article

Boyer, George R. "Poor Law, old." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000209> doi:10.1057/9780230226203.1306

The New Palgrave Dictionary of Economics Online

population ageing

David N. Weil
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Population ageing is primarily the result of past declines in fertility, which produced a decades-long period in which the ratio of dependents to working-age adults was reduced. Rising old-age dependency in many countries represents the inevitable passing of this ‘demographic dividend’. Societies use three methods to transfer resources to people in dependent age groups: government, family, and personal saving. In developed countries, families are predominant in supporting children, while government is the main source of support for the elderly. The most important means by which ageing will affect aggregate output is the distortion from taxes to fund public pensions.

Keywords

bequests; capital accumulation; capital–labour ratio; demographic dividend; demography; dependency; family transfers; fertility in developed countries; fertility in developing countries; generational crowding; India; international capital flows; Japan; labour supply; mortality; old-age pensions; optimal saving; payroll taxes; population ageing; retirement; saving; Social Security in the United States

Article

Population ageing is the shift in the distribution of a country's population towards older ages. An increase in the population's mean or median age, a decline in the fraction of the population composed of children, or a rise in the fraction of the population that is elderly are all aspects of population ageing.

Population ageing is occurring in most parts of the world, but is most advanced in the richest countries. Among the countries currently classified by the United Nations as more developed (which had a population of 1.2 billion in 2005), the median age of the population rose from 29.0 in 1950 to 37.3 in 2000, and is forecast to rise to 45.5 by 2050. The corresponding figures for the world as a whole are 23.9 for 1950, 26.8 for 2000, and 37.8 for 2050. In Japan, one of the fastest-ageing countries in the world, in 1950 there were 9.3 people younger than 20 for every person older than 65. By the year 2025, the ratio is forecast to be 0.59 people younger than 20 for every person older than 65 (United Nations, 2004).

The sources of population ageing lie in two demographic phenomena: rising life expectancy and declining fertility. An increase in longevity raises the average age of the population by raising the number of years that each person is old relative to number of years in which he is young. A decline in fertility increases the average age of the population by changing the balance of people born recently (the young) to people born further in the past (the old). Of these two forces, it is declining fertility that is the dominant contributor to population ageing in the world today (Weil, 1997). More specifically, it is the large decline in the total fertility rate since the 1950s that is primarily responsible for the population ageing that is taking place in the world's most developed countries. Because many developing countries are going through faster fertility transitions, they will experience even faster population ageing than the currently developed countries in the future.

While the economic underpinnings of the demographic processes that cause population ageing – in particular declining fertility – are interesting topics in and of themselves, this article instead concentrates on how ageing affects the economy.

The economic effects of population ageing

Population ageing has economic effects whenever some economic interaction (the sale of a good or service, the provision of a government benefit, and so on) brings together people whose participation is a function of their age. In such a situation, a change in the relative size of two age groups will require a change in behaviour by members of at least one group. For example, babies demand strollers, which are produced by working-age adults. Thus, a reduction in the ratio of babies to adults will mean more strollers per baby, fewer adults working in stroller production, or both. The changes in behaviour required to restore equilibrium in the face of demographic change are induced through either prices or institutions. If individuals on at least one side of the transaction respond elastically to price changes (as would be the case in getting working-age adults to move from stroller manufacture into the wheelchair business), then the effects of population ageing will be little worth commenting on. But when individuals on both sides of the interaction are not easily induced to change their behaviour, the economic effects of population ageing will be dramatic. Old-age pensions, child rearing, and the combining of old people's capital with young people's labour are all cases where a change in the relative numbers on either side of the equation will have important effects.

The simplest analysis of the economic effects of population ageing starts with the notion of age-based dependency: people of some ages produce less than they consume, and are dependent on the rest of society for their support. Consider a division of the population into three age groups: working age adults, dependent youths, and dependent elderly. We temporarily ignore the question of how resources are transferred from working-age adult to dependent children and elderly. For simplicity, we assume that people of all ages have the same consumption, although the analysis can easily be extended to allow for age-varying consumption needs (see Weil, 1999). Finally, we assume that output is produced solely by the labour of working-age adults, with no additional factors of production such as capital.

The consumption possibilities of our idealized society can be analysed in a diagram like Figure 1. The horizontal axis plots youth dependency ratio (population aged 0–19 divided by population aged 20–64); the vertical axis plots the old-age dependency ratio (population aged 65+ divided by population aged 20–64). A society's demographic structure is represented by a point in this space. For example, a newly planted colony might be represented by a point in the lower left-hand corner, with youth and old age dependency ratios of zero. For a normal society, however, the demographic processes of ageing, mortality and fertility will determine predictable movements of the age structure through the space of Figure 1. A set of points of particular interest are what demographers call stable populations. These are populations in which age-specific mortality and fertility rates have been constant for sufficient time that the relative number of people of each age is constant. Figure 1 shows a typical locus of stable populations, generated using age-specific mortality rates for the United States in 2000 and varying the level of fertility. The labels show the population growth rate consistent with different points on the locus of stable populations.

Figure 1

Stable populations and iso-dependency lines. *Source*: Author's calculations.

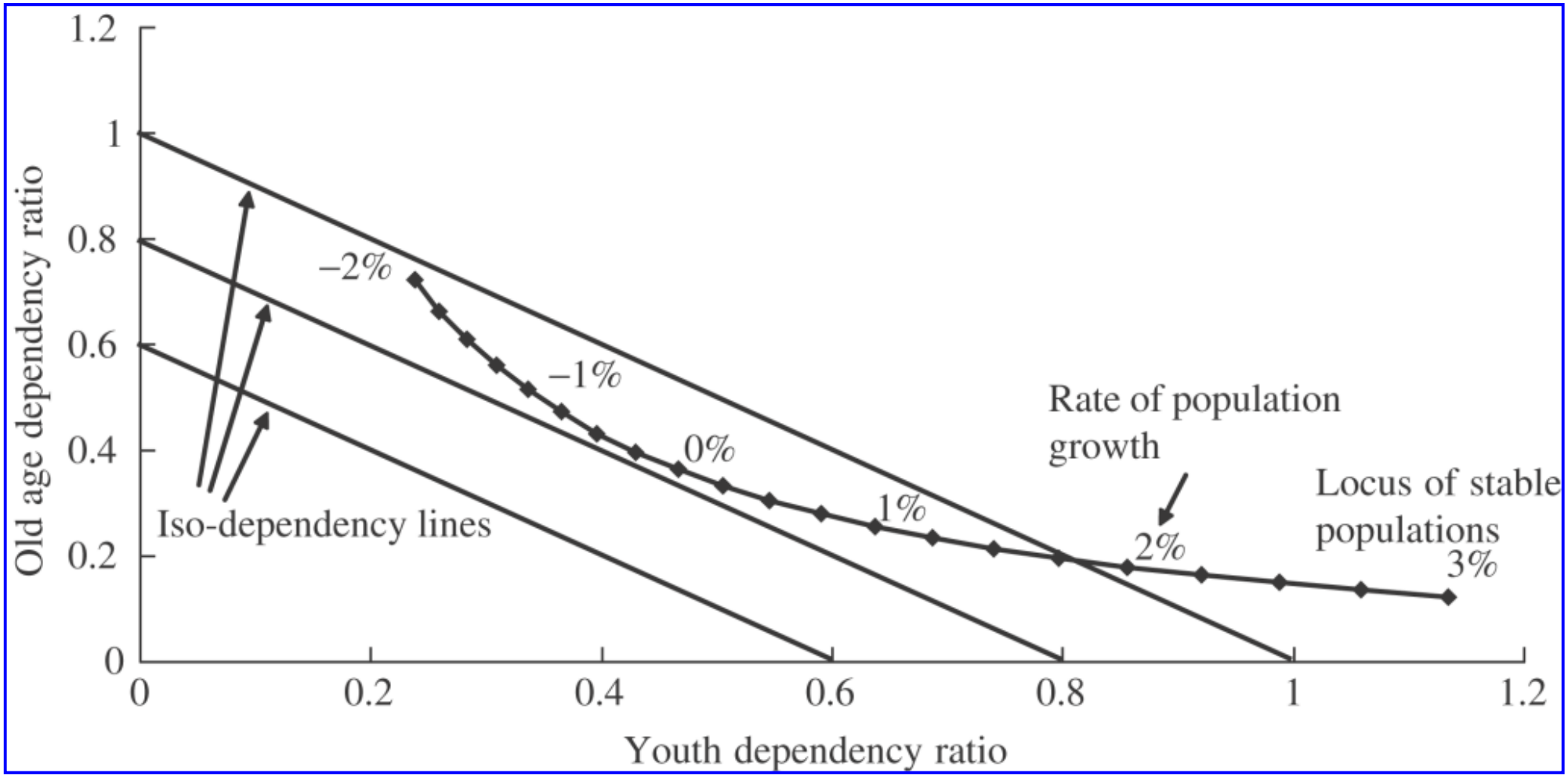

We can also represent in this space the effect of demographic structure on the consumption possibilities of the society through a series of iso-dependency lines. These are lines along which the sum of youth and old-age dependency is constant – in other words, combinations of old-age and youth dependency that yield constant levels of consumption per capita. Iso-dependency lines closer to the origin represent age structures which allow for higher consumption per capita. The tangency between the locus of stable populations and an iso-dependency line shows the stable population with the lowest dependency ratio.

Reductions in fertility will lead to clockwise movements of the point representing a country's demographic structure through the space of Figure 1. Falling fertility reduces the youth dependency ratio immediately, and only raises the old age dependency ratio with a lag of several decades. For this reason, a country experiencing fertility transition will be able to move temporarily below the locus of stable populations.

Figures 2a–2c show data on population age structure for the United States, Japan, and India over the period 1950–2050. In all cases, the clockwise motion and period of temporarily low dependency due to fertility transition are visible, although the countries differ in how far along they are and how severe the process of ageing is forecast to be. In Japan, the total dependency ratio (youth plus old age) will rise from 0.64 to 1.17 over the period 2005–50, implying, *ceteris paribus*, that GDP per capita will grow 0.6 per cent per year more slowly than GDP per worker (see Weil, 2005, ch. 5, for details of this calculation). By contrast, India, like many developing countries, is in the process of receiving a large ‘demographic dividend’ from reduced fertility (Bloom and Williamson, 1998).

Figure 2A

Demographic dynamics in the United States. *Source*: United Nations.

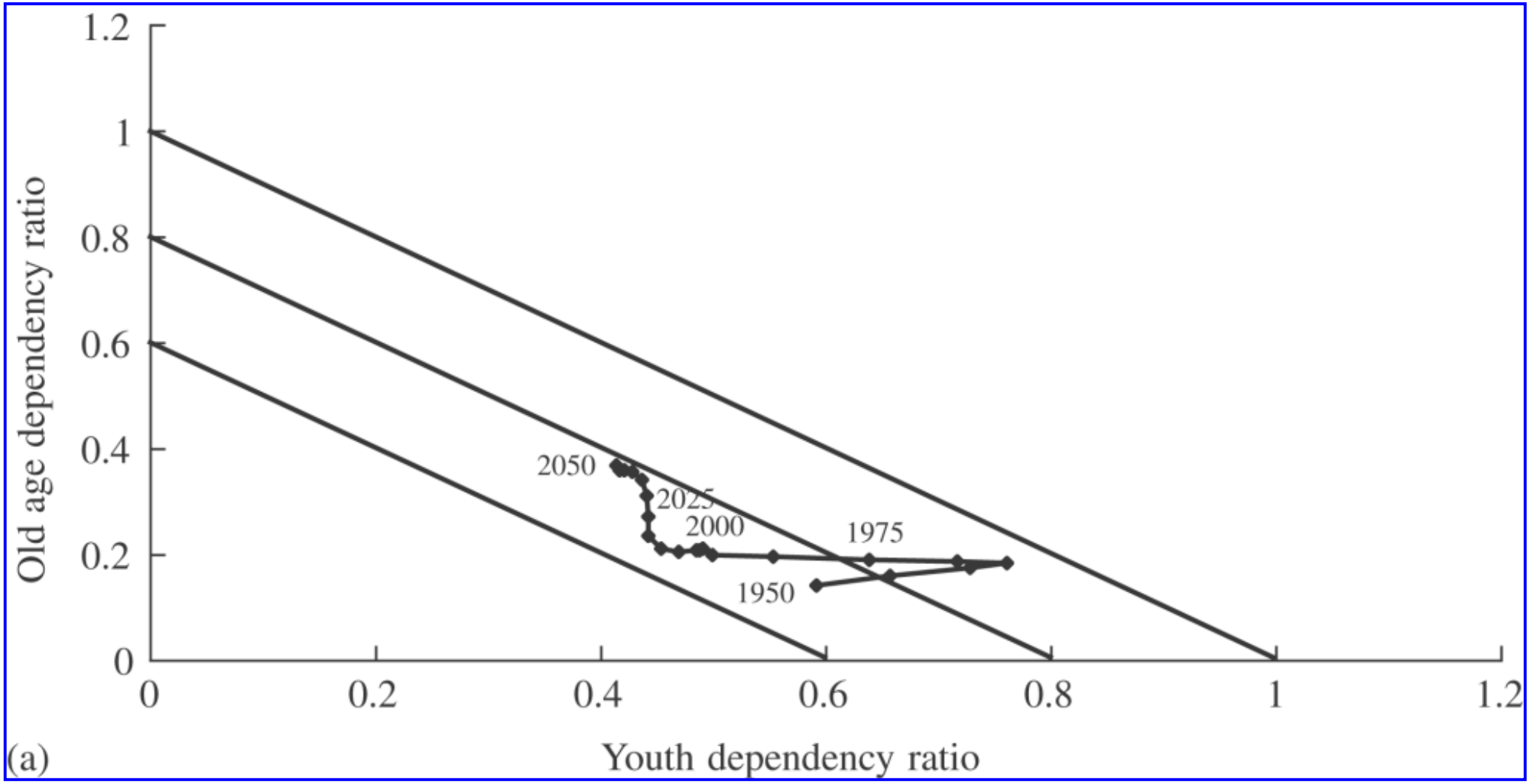

Figure 2B
Demographic dynamics in Japan. *Source*: United Nations.

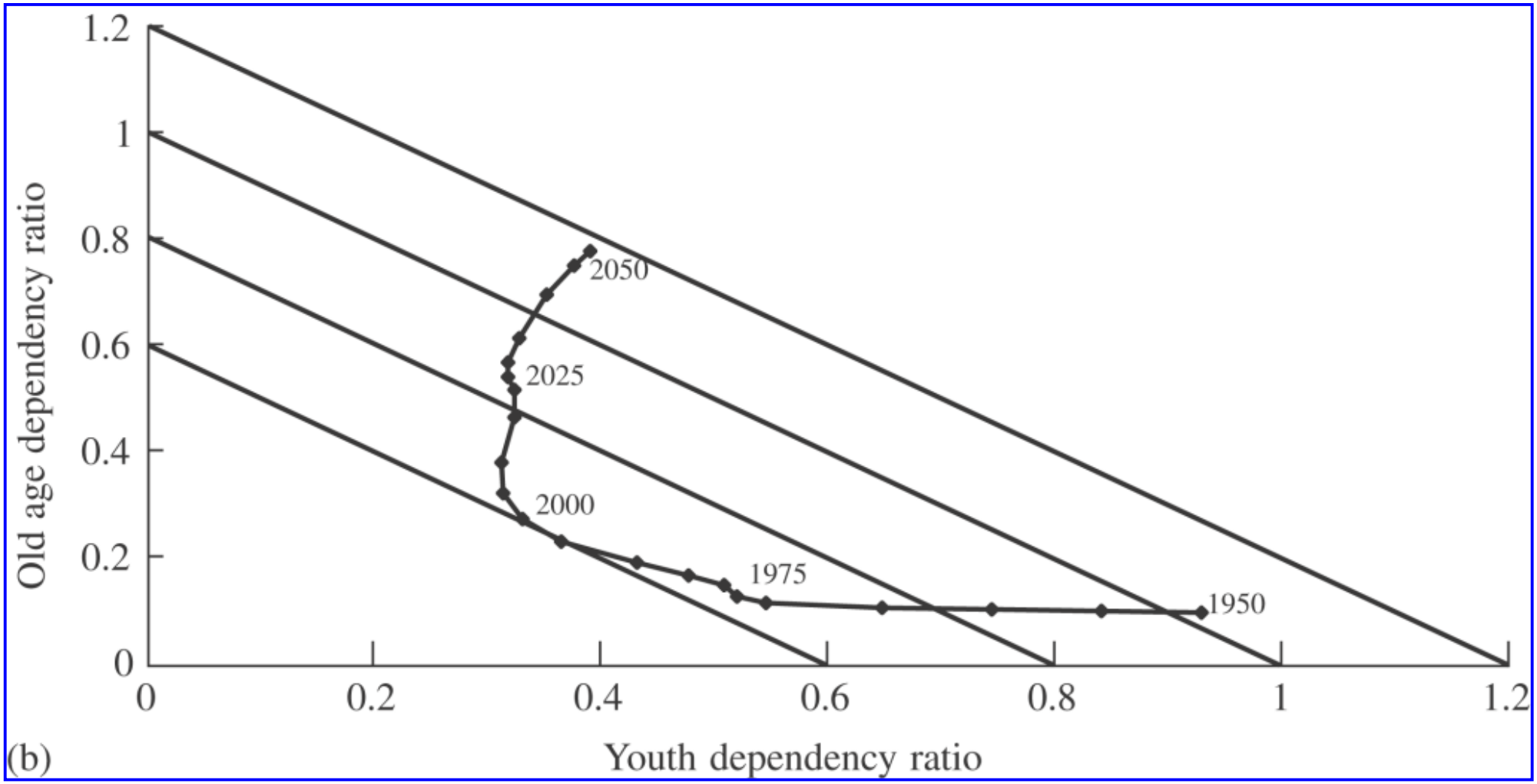

Figure 2C
Demographic dynamics in India. *Source*: United Nations.

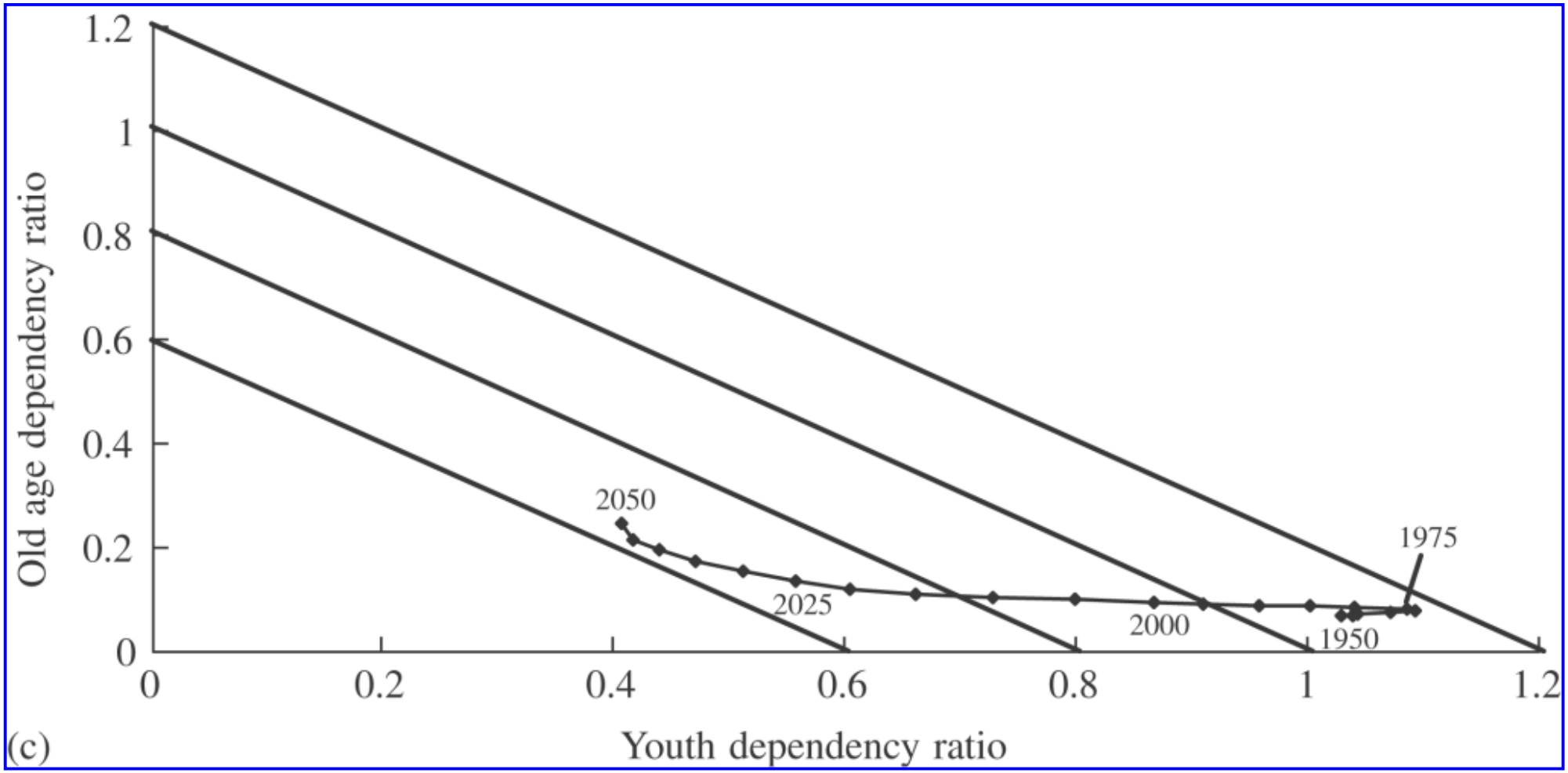

The lesson from this analysis of dependency is that, from the point of view of society as a whole, the period of rapid increase in old-age dependency that is in store for the world's richest countries is to a large extent simply the passing of the transitory benefit derived from a decrease in fertility. A second lesson is that any change in fertility that will in the long run undo the effects of population ageing will, in the short run, lead to an *increase* in total dependency by moving the point representing age structure above and to the right of the locus of stable populations.

The model discussed above ignores the means by which dependent members of society are supported. In practice, there are three mechanisms by which this takes place: through their own past savings; through institutions (primarily the government) that transfer resources between unrelated people of different ages; and through their own families. Lee (2002) refers to these various means by which resources are transferred among age groups as a ‘reallocation system’. We shall see that the nature of the reallocation system affects the overall burden of ageing as well as the distribution of that burden.

Ageing, savings and capital

Capital is important in analyses of population ageing for two reasons. First, accumulation of capital allows either individuals or society as a whole to break the temporal link between production and consumption: an individual, for example, can save some of her wages when she is working, and then use the accumulated capital to fund consumption during retirement. Second, as a factor of production complementary to labour, capital helps determine the quantity of output to be divided among workers and dependents. Analyses of ageing and capital accumulation proceed down both normative and positive channels.

The normative approach asks how society should respond to a looming change in demographics. Although there is in practice no social planner who makes saving decisions for

society as a whole, the solution to the social planner's problem can inform the response of a government that influences national saving through fiscal policy and tax incentives. Common sense would suggest that a country that is undergoing population ageing should ‘save for its old age’, that is, accumulate extra capital during the period of low dependency in order to maintain a smooth path of consumption into the period of high dependency. As stressed by Cutler et al. (1990), however, there is a countervailing effect: population ageing due to lower fertility implies that the working-age population will grow more slowly, reducing the amount of investment required to supply new workers with capital. The flip side of this decrease in required investment is that, if a country did attempt to save sufficient capital to smooth consumption in the face of ageing, the result would be a rise in the capital–labour ratio, lowering the return on capital, which would lead households (or a social planner) to want to raise consumption. Elmendorf and Sheiner (2000) calculate that an optimizing social planner would want to make relatively small changes in saving rates in response to the population ageing currently forecast in the United States.

A positive alternative to the social planner approach is to consider the equilibrium of an economy in which consumers make privately optimal saving decisions taking as given the expected paths of interest rates and wages as well as taxes and government benefits. Forecasting the effects of demographic change on output or capital per worker, interest rates, and so on requires a fully articulated, rational expectations general equilibrium model. Kotlikoff, Smetters and Walliser (2007) use such a model to analyse demographic change in the United States, under the assumption that the Social Security benefit regime does not change, and that payroll taxes adjust accordingly. They find that the capital deepening that would normally accompany a shift of the population into its peak asset-holding years is undone by rising payroll taxes. They forecast ‘capital shallowing’ that will raise the real return on capital by one percentage point by 2030 and a further two percentage points over the rest of the 21st century, as well as a dramatic slowing of real wage growth.

Rather than fitting an optimizing model of saving, another approach is to look empirically at the age pattern of actual behaviour. Poterba (2005) shows that individual net worth follows a hump-shaped path over the course of the lifetime, peaking between ages 65 and 69. Unlike the classical life-cycle model, however, the decline in average net worth is relatively slow, so that average net worth at death is significant. This life-cycle pattern of asset accumulation in turn implies that shifts in demography will shift asset demands and potentially asset prices. In particular, the movement of the baby-boom generation into its high accumulation years was widely cited as a potential explanation for the run-up in stock prices during the last decades of the 20th century (Abel, 2003). Similarly, some analysts have suggested that, as the balance between age groups actively accumulating and running down wealth shifts in the period after 2010, there will be a corresponding meltdown of asset prices. However, Poterba (2005) finds little evidence of demographic effects on asset returns in time series data from the United States, Canada and the United Kingdom. Lim and Weil (2003) show that in a forward-looking asset pricing model it would require an unreasonably large adjustment cost for capital to produce a large asset price meltdown in response to projected population ageing. The shift in population towards the elderly will also lead to a significant increase in the flow of bequests relative to either income or wealth of the younger generation; Weil (1994) argues that this increased flow of bequests will reduce the saving of the receiving generation.

The above discussion implicitly considered the case of an economy closed to international capital flows. In an open economy, the mismatch between the demographically induced demand for asset holding and the capital requirements of the labour force can be channelled into capital flows abroad. For example, a country like India, where the working-age population is forecast to grow an annual rate of 1.8 per cent per year between 2000 and 2025, would be a natural recipient of investment from Japan, where the working-age population will shrink at an annual rate of 0.6 per cent per year over this period. In practice, however, net financial flows among countries tend to be far smaller than a model of perfectly open capital markets would imply, and movements in current accounts seem to bear little resemblance to those predicted by demographic change (see Brooks, 2003).

Ageing and government

In the developed countries that are ageing most rapidly, government transfer programmes are a major source of support for dependent elderly. In Germany, for example, transfers net of taxes and inclusive of public health benefits make up 65 per cent of the income of people aged 65 and older (Burtless, 2006). Correspondingly, one of the most important functions of government is transferring resources to elderly people. In 2005, US federal outlays were 18.9 per cent of GDP. Almost 60 per cent of that amount was spent on direct transfers attributable to specific age groups (Medicare for the elderly, unemployment insurance for working age, and so on). Of such transfers, 58 per cent (6.5 per cent of GDP) was directed toward those aged 65 and older. On a per person basis, the elderly received close to eight dollars in direct transfers for every dollar of transfers received by working-aged persons. In sharp contrast, children received just 35 cents per dollar of transfers awarded to workers. Assuming constant transfers per person by age group, a shift of ten per cent of the population out of the workforce and into retirement would increase federal transfer outlays by 4.7 per cent of GDP (calculations based on data underlying Gokhale and Smetters, 2006). In addition to raising spending, population ageing also reduces government revenue. Putting these tax and spending effects together, Burtless (2006) calculates that the effect of population ageing would raise the tax rate required to pay for government transfers on a PAYGO basis in from 16 per cent in 2000 to 21 per cent in 2030 in the United States. In Germany, where transfers are larger and ageing more extreme, the increase in the tax rate would be from 28 per cent to 40 per cent. In the United States, the effect of ageing on the government budget is greatly exacerbated by the fact that the *price* of health care for the elderly is rising at the same time as the fraction of the population that is elderly (Elmendorf and Sheiner, 2000).

One important way in which transfers to dependents (either children or elderly) that are channelled through the government differ from those mediated by either the family or through one's own saving is in how workers perceive the benefits resulting from their forgone consumption. People give money and other resources to their children or aged parents because

they care about them. And when people save for their own old age, it is because they care about their future selves. But few people are so altruistic that they value the taxes that are taken from their pay in order to fund transfers to the elderly. For this reason, there is an efficiency loss associated with government support of the elderly that is not present in other forms of transfers to dependent age groups. Prescott (2004) argues that differences in marginal labour tax rates explain large cross-sectional differences and changes over time in labour supply among the G-7 countries. For example, in the early 1990s his calculations show the French average marginal tax rate (inclusive of consumption taxes) being 48 per cent larger than that in the United States; correspondingly, French adults aged 15–64 worked only 68 per cent as many hours as their US counterparts. The large elasticity of labour supply that Prescott estimates implies that deadweight losses will increase dramatically as populations age, as long as government old-age pensions continue to be funded through taxes that are largely divorced from the benefits that the individuals paying them will receive. Thus an economy that could function smoothly with a high level of youth dependency funded through family transfers, or a high level of old-age dependency funded through savings, might collapse if a similar level of old-age dependency were funded through taxes.
Because government transfers are so heavily weighted toward the elderly, the adjustment in government finances required to deal with population ageing will be proportionally much larger than the overall change in consumption in the economy as a whole. Roughly put, ageing is a much bigger problem for the government than for the economy as a whole. Most conceivable reforms in government old-age pensions will represent net losses to cohorts who are near or beyond retirement at the time of reform. Bohn (2005) calculates that, based on current participation rates, the fraction of voters aged 65 and over in the United States will rise from 19.8 per cent to 30.5 per cent between 2003 and 2030; over the same period the age of the median voter will rise from 47 to 52. Thus, as the fiscal strain from population ageing becomes acute it will be increasingly difficult for policymakers to solve their problems by reducing transfers to the elderly.

Ageing and families

Transfers within families represent the final channel whereby dependents are supported. For the large majority of old people in developed countries, family transfers are the second or third most important source of support, behind their own past savings and/or transfers from the government. This is a relatively new pattern. Prior to the 20th century, the period of old-age dependency was much shorter, government transfers to the elderly were minimal, and cohabitation of elderly with their children was the norm. In the United States, for example, the fraction of elderly widows who lived with their adult children fell from 67 per cent in 1920 to 20 per cent in 1990 (McGarry and Schoeni, 2000). Only 2.7 per cent of people aged 60 and over in the United States reported support from children as their main source of income in 2001. Even in Japan, where such transfers have traditionally played a much larger role, the fraction of people 60 and over reporting their children as their main source of support fell from 29.8 per cent to 12.0 per cent between 1981 and 2001 (United Nations, 2005, Table I.2). In contrast to the elderly, the burden of supporting young dependents lies foremost on their own families. Mason et al. (2005) calculate that 57 per cent of consumption of people under 20 in the United States in 2000 was financed by transfers from family members. Thus, unlike governments, families headed by working-age adults find their budget constraints relaxed by the low fertility that causes population ageing.
An important distinction between support for elderly dependents and support for child dependents concerns the degree of choice that those doing the support enjoy. Working-age adults cannot choose how many siblings they share the burden of supporting their parents with, much less the size of the working-age cohort relative to the elderly population, which determines the level of taxation required to fund public pensions. But working-age adults *can* choose the number of children they produce and support, and their choices about fertility may respond to economic conditions. Of particular interest in the present context, population ageing itself may feed back to affect fertility. The best-known mechanism whereby population age structure affects fertility was identified by Easterlin (1987), who hypothesized that members of large birth cohorts would suffer from labour market crowding, earn wages that are low relative to the standard of living that they had grown up with, and would adjust fertility downward to partially restore their standard of living. The rise in taxes required to fund transfers to the elderly that will result from population ageing could have effects on after-tax income that are as large or larger than those from Easterlin-style generational crowding; thus, ageing could lead to lower fertility and, down the road, even more ageing (Hock and Weil, 2006).

See Also

- economic demography
- fertility in developed countries
- retirement
- Social Security in the United States

I am grateful to Jagadeesh Gokhale for helpful comments.

Bibliography

Abel, A. 2003. The effects of a baby boom on stock prices and capital accumulation in the presence of Social Security. *Econometrica* 71, 551–78.

Bloom, D. and Williamson, J. 1998. Demographic transitions and economic miracles in emerging Asia. *World Bank Economic Review* 12, 419–56.

Bohn, H. 2005. Will Social Security and Medicare remain viable? In *Social Security Reform: Financial and Political Issues in International Perspective*, ed. R. Brooks and A. Razin. Cambridge, UK: Cambridge University Press.

Brooks, R. 2003. Population ageing and global capital flows in a parallel universe. *IMF Staff Papers* 50:2, 200–221.

Burtless, G. 2006. Cross-national evidence on the burden of age-related public transfers and health benefits. Working Paper No. 2006-6. Center for Retirement Research, Boston College.

Cutler, D., Poterba, J., Sheiner, L. and Summers, L. 1990. An ageing society: opportunity or challenge? *Brookings Papers on Economic Activity* 1990(1), 1–73.

Easterlin, R. 1987. *Birth and Fortune: The Impact of Numbers on Personal Welfare*, 2nd edn. Chicago: University of Chicago Press.

Elmendorf, D. and Sheiner, L. 2000. Should America save for its old age? Fiscal policy, population ageing, and national saving. *Journal of Economic Perspectives* 14(3), 57–74.

Gokhale, J. and Smetters, K. 2006. Fiscal and generational imbalances: an update. In *Tax Policy and the Economy, Volume 20*, ed. J. Poterba. Cambridge, MA: MIT Press.

Hock, H. and Weil, D. 2006. The dynamics of age structure, dependency, and consumption. Mimeo, Brown University.

Kotlikoff, L., Smetters, K. and Walliser, J. 2007. Mitigating America's demographic dilemma by pre-funding social security. *Journal of Monetary Economics* 54, 247–66.

Lee, R. 2002. A cross-cultural perspective on intergenerational transfers and the economic life cycle. In *Sharing the Wealth: Demographic Change and Economic Transfers between Generations*, ed. A. Mason and G. Tapinos. Oxford: Oxford University Press.

Lim, K.-M. and Weil, D. 2003. The baby boom and the stock market boom. Scandinavian *Journal of Economics* 105:3, 359–378.

Mason, A., Lee, R., Tung, A.-C., Lai, M.-S. and Miller, T. 2005. Population ageing and intergenerational transfers: introducing age into national accounts. Mimeo, University of Hawaii.

McGarry, K. and Schoeni, R. 2002. Social Security, economic growth, and the rise of independence of elderly widows in the 20th century. *Demography* 37, 221–36.

Poterba, J. 2005. Demographic structure and asset returns. In *Social Security Reform: Financial and Political Issues in International Perspective*, ed. R. Brooks and A. Razin. Cambridge, UK: Cambridge University Press.

Prescott, E. 2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28(1), 2–14.

United Nations. 2004. *World Population Prospects: The 2004 Revision Population Database*. New York: Population Division, United Nations.

United Nations. 2005. *Living Arrangements of Older Persons Around the World*. New York: Department of Economic and Social Affairs, United Nations.

Weil, D. 1994. The saving of the elderly in micro and macro data. *Quarterly Journal of Economics* 109, 55–81.

Weil, D. 1997. The economics of population ageing. In *Handbook of Population and Family Economics*, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Weil, D. 1999. Population growth, dependency, and consumption. *American Economic Review* 89, 251–5.

Weil, D. 2005. *Economic Growth*. Boston: Addison-Wesley.

How to cite this article

Weil, David N. "population ageing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000224> doi:10.1057/9780230226203.1307

The New Palgrave Dictionary of Economics Online

population and agricultural growth

James Roumasset
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Thinking about population as a driver of agricultural development provides insights into induced technical and institutional change, whether it be Ester Boserup's declining fallow period, modern crop varieties, or the horizontal and vertical specialization that arise in labour-intensive agriculture. The non-convexities of research and development, infrastructure investments, and specialization imply that modest population pressure does not necessarily exert downward pressure on wages. As agricultural growth stimulates industrialization, the non-convexities of specialization become ever more compact. The combination of these and the increased demand for human capital, if not inhibited by policy failures, tends to promote a virtuous circle of human progress.

Keywords

Black Death; capital intensity; child care; development economics; dismal science; division of labour; elasticity of substitution; endogenous growth theory; endogenous population growth; exchange labour; fertility; governance; human capital; immiserizing growth; indivisibilities; induced institutional change; induced technical change; industrialization; innovation possibility frontier; intergenerational neutrality; Kuznets curve; labour intensification; labour market contracts; labour productivity; learning-by-doing; leisure; Malthus's theory of population; mortality; natural capital; new classical economics; new household economics; non-convexity; population and agricultural growth; population growth; poverty alleviation; property rights; rent seeking; research and development; resource depletion; shadow pricing; sharecropping; specialization; spot markets; steady state; subsistence; sustainability criteria; sustainable development; transaction costs

Article

That economics became known as the ‘dismal science’ can largely be attributed to the theory of population and agricultural growth as developed by Malthus and Ricardo, notwithstanding the term's origin in another context. Starting from a point of relatively high wages, for example at the end of the Black Death in Europe, or after some exogenous technological improvement, population increases geometrically. The additional population is assimilated by agricultural growth at the extensive and intensive margins, both of which result in diminishing returns to labour. Extensive growth occurs through the expansion of cultivated land, which Ricardo (1817) presumed to be more distant from or of poorer quality than land already in use. Growth at the intensive margin likewise results in diminishing returns, due to the greater amount of labour and other inputs employed on the fixed quantity of previously cultivated land. As a consequence, Ricardo (1817) and Malthus (1798) theorized that wages would eventually decline towards a subsistence level, where population growth would cease due to ‘positive checks’ such as starvation and disease.

Modern economists still use this dismal theory to explain why growth in levels of living among the working classes was never sustained for long periods until the advent of the Industrial Revolution. Each technological improvement was subsequently ‘eaten up’ by population growth and the subsequent diminishing returns. The belief in this theory is so strong that Lucas (2002, ch. 3) wrote that he could look at a picture of a Korean peasant farm in an unknown century and confidently guess household income. Recent interest in ‘sustainable development’ has augmented resource pessimism. In this view, the conventional Malthusian vicious circle between population growth and poverty is exacerbated by resource depletion and environmental degradation. Expanding numbers of poor people in developing countries put more pressure on limited natural resources and fragile ecosystems,

and the falling resource base makes the Malthusian circle even more vicious than with a fixed resource endowment.

Malthus famously argued that unchecked population growth is exponential while food production at best grows linearly, thus implying the inevitability – in the absence of sufficient *preventative* checks – of *positive* checks such as pestilence, plague, famine and war and of subsistence levels of income in the long run. Ironically, food supply has outstripped population growth ever since the publication of Malthus's *Essay on the Principle of Population*. Technological and institutional change has been more rapid than he envisioned and preventative checks more robust.

Boserup effects

Boserup (1965; 1981) takes a different tack by taking population growth as the exogenous variable and enquiring into the consequences thereof for agricultural technology and institutional change. I follow Boserup's lead in most of what follows, eventually returning to a more integrated view. Boserup focused on the effects of physiological population density on an additional intensive margin – the fallow period. As population (and other demand factors) grow, the predominant agricultural system gradually transitions from long to short fallow to annual cropping to multiple cropping. Table 1 describes these systems and illustrates the rough correspondence between the frequency of cropping and population density in less developed economies. Other authors have extended the correlation between population density and cropping frequency to European countries, both over time and country.

Boserup's frequency of cropping by population density

System	Description of cropping system	Frequency of cropping	Person per km^2	Density
Hunting and gathering	Wild plants, roots, fruits and nuts are gathered	0%	0–2	Very sparse
Forest fallow (w/ astoralism)	One or two crops followed by 15–25 years' fallow	0–10%	1–4	Very sparse
Bush fallow (w/ pastoralism)	Two or more crops followed by 8–10 years' fallow	10–40%	4–64	Sparse to Medium
Short fallow (w/ domestic animals)	One or two crops followed by one or two years' fallow	40–80%	16–64	Medium
Annual cropping (w/ intensive animal husbandry)	One crop each year with only a few months' fallow	80–100%	64–256	Dense
Multi-cropping	Two or more crops in the same fields each year without any fallow	200–300%	≥256	Very dense

Source: Boserup (1981, pp. 9, 19 and 23).

Boserup's insight can be partly understood from the perspective of *induced technical change* (Ahmad, 1966). Absent industrial growth, population pressure makes land increasingly scarce relative to labour, thus inducing land-saving technical change. In the era of modern economic growth, the same tendency would influence whether capital was used to save labour or land. This was exemplified by labour-abundant Japan developing land-saving biological innovations and the United States developing labour-saving mechanical innovations (Hayami-Ruttan, 1985). As represented with standard neoclassical analysis, however, induced innovation simply increases the elasticity of factor substitution (especially between land and labour). In the *very* long run, that is, allowing for induced technical change, the elasticity of substitution, such as between land and labour, is higher than without technical change. Similarly, decreasing the fallow period allows the marginal product to decline more slowly than otherwise. For example, suppose that 100 workers cultivate 100 hectares with a 50 per cent cropping frequency (short fallow) and that the population doubles. Even though the additional labour can be productively employed, for instance by better weeding and more thorough land preparation, the marginal product of labour will suffer a large decline if the cropping frequency remains unchanged (perhaps by a half or more). By switching to annual cropping, however, it may be possible to accommodate the additional labour with only a small decline in its marginal product, even in the steady state. The optimal solution involves some conservation of soil fertility over time, for example through the use of animal manure and crop rotation (Barrett, 1991).

Boserup contends that it is even possible that population pressure increases the productivity of agricultural labour. More intense farming systems require more fixed costs. For example, forest fallow systems require minimal land preparation. The slash and burn method leaves the land both fertile and weed-free. In the tropical African context that she describes, however, once the land has been burned and cropped, it is taken over by grasses and is no longer suitable for slash and burn agriculture until 20 or more years later, when the forest has returned. Consequently, land preparation requires time-intensive ploughing. Because of these fixed costs, the average product of labour rises over some range.

Other investments associated with intensification, such as irrigation and terracing, similarly increase labour productivity. This is illustrated in Figure 1. Once population has reached point C, the average product of the extensive and intensive techniques is equalized and it becomes worthwhile to switch to the intensive method. As labour increases beyond C, the average product rises until D, where diminishing returns just offset the gains from spreading the fixed costs, and average product begins to decline. In this sense, population eventually overcomes the transitory gains from switching techniques and causes productivity to fall.

Figure 1

Average product of labour under different farming techniques. *Source*: adapted from Krautkraemer (1994).

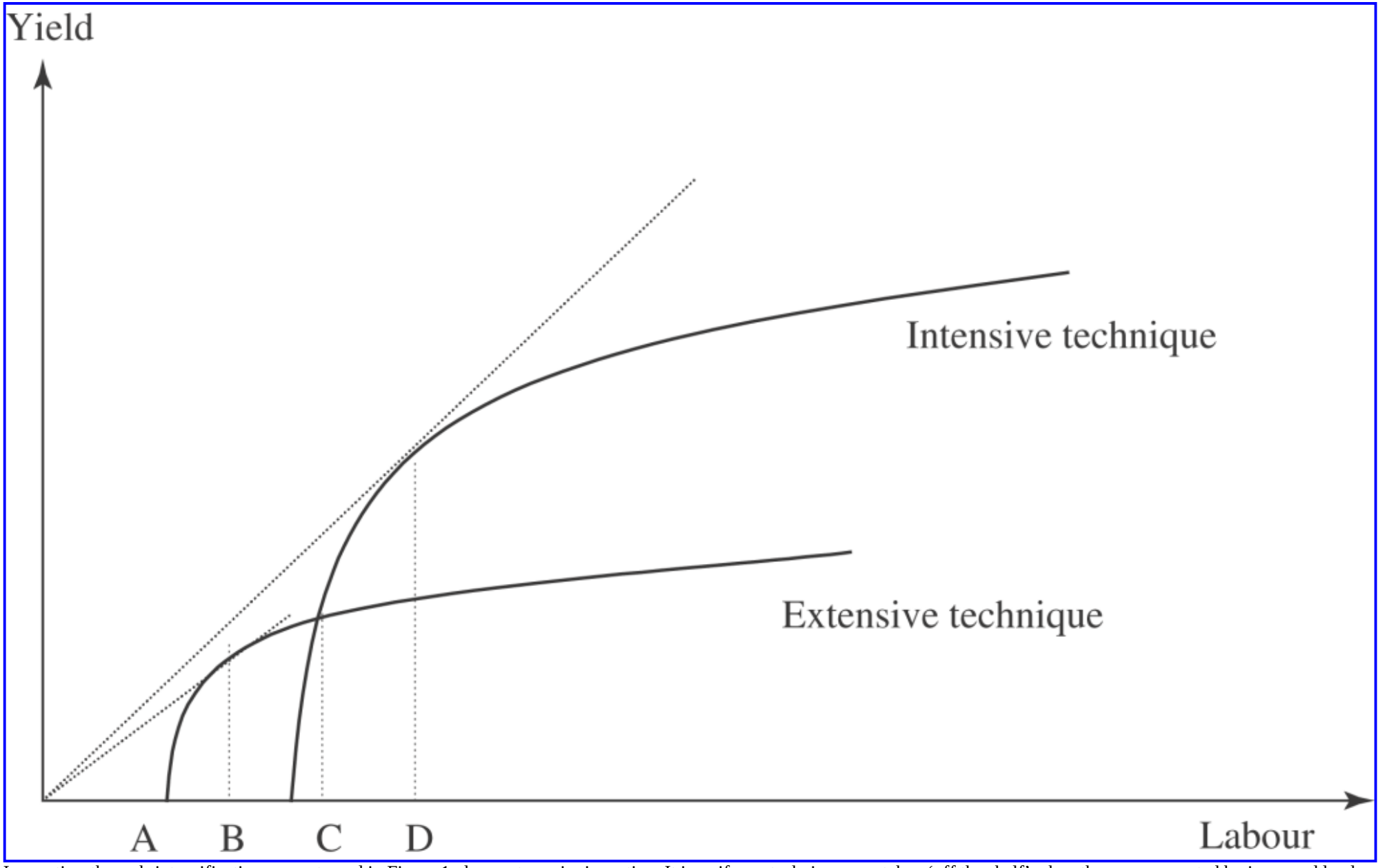

Innovation-through-intensification, as portrayed in Figure 1, does not require invention. It is as if new techniques are taken ‘off the shelf’ when they are warranted by increased land scarcity. Genuinely new technology, developed through invention or imported from other areas, may provide additional positive effects. The same population increase that warrants the fixed cost of intensification also warrants increased expenditures on experimentation and research. This research shifts the *innovation possibility frontier* (IPC) between land and labour inwards. In modern settings, R&D becomes an important source of productivity growth.

For example, the high-yielding, or modern, wheat and rice varieties (MVs) developed in the 1960s were in large part induced by population pressure on increasingly scarce land. In the extensive phase of agricultural development, cultivated hectarage is increasing. Eventually, cultivated area reaches a maximum and declines as towns and industrial areas encroach on agricultural land. At this point, land scarcity is exacerbated by both rising food demand and falling land supply, and intensification accelerates.

One of the effects of intensification is to increase the demand for land-saving technology. According to the ‘political Boserup effect’ (Evenson, 2004), increasing population densities

induce countries to invest more in the genetic improvement of both crops and animals. By first characterizing existing technology by the unit requirements of land, labour, and capital, optimal investment by a country in new technology can be described by the amount of research and its factor-saving bias. In one version of this theory, a given research expenditure allows a country to pick any point on the IPC, the envelope of all unit isoquants in the land–labour plane, that said research expenditure affords. If it is assumed that the IPC shifts in a neutral fashion towards the ultimate IPC, wherein the marginal benefit of research is zero, then the factor-saving bias is in accordance with changes in relative factor prices. For example, if population growth results in a decrease in the wage rate and an increase in the land rental rate, both relative to the price of capital, then technical change will be land-saving and labour-using relative to capital (Binswanger and Ruttan, 1978, chs 2 and 4).

Inasmuch as the IPC shifts in a non-neutral fashion, however, these results will be modified. It is natural to assume, for example, that technical change is inherently capital-using, that the unit isoquant (net of capital costs) can be shifted inward more cheaply by increasing capital per unit of output than by increasing labour or land. Moreover, it may be that inventing technology that uses capital to save labour is cheaper than technology that saves on land. This may explain why the modern rice and wheat varieties have been found to be mildly labour-saving, in addition to being land-saving and capital-using (fertilizer responsive), even though their demand was created by falling wages relative to land rents. But even though labour per unit of output fell, output per hectare increased enough such that MVs had a positive effect on wages (for example, Evenson, 1982). Overall, MVs have had a beneficial effect on poverty reduction by decreasing food prices and increasing wages relative to what they would have otherwise been given population growth and labour demand in other sectors.

Boserup's other ‘secondary effects’ of population growth may also cause productivity to rise, even in the absence of agricultural research. Among these are property rights, work habits, division of labour, education, and the infrastructure for transport and communication. Changing property rights exemplifies how institutions can change in response to population pressure and other changes in factor scarcities. This insight led to the theory of *induced institutional change* as a complement of the theory of induced technical change. For example, as population pressure increased the demand for land-saving investments, private property sometimes emerged as a more efficient substitute for top-down land management by community leaders or feudal lords (see, for example, North and Thomas, 1973). Indeed, the first legal enforcement of the early English enclosures was effected by the Statute of Merton (1235), which noted the need to improve the land in order to generate greater rent. The subsequent waves of English enclosures beginning before the 17th and 19th centuries also appear to have followed increases in the rate of population growth, although the timing is not without dispute.

Population induced specialization in agriculture

While population growth potentially augments the benefits of private property, potential efficiency gains do not automatically induce institutional change. In particular, rent seeking may lead to a ‘race’ such that private property is created before it actually increases efficiency (Lueck, 1998). On the other hand, political costs may retard institutional change beyond the time that its benefits warrant. The advent of private property in Hawaii in 1848 was exceptional in two regards. First, the benefits of private property resulted from the increased profitability of sugar and pineapple production, even in the face of population decline. Second, the timing of private property accorded roughly with its efficiency benefits; the delaying effects of the political costs of change were offset by the expediency of governmental land sales.

A more profound institutional change that may be induced by population pressure and other sources of intensification is that of economic organization. The division of labour has fascinated economists since the time of Adam Smith, but was sidelined during the era of neoclassical economics. The theme of specialization has been resurrected, implicitly in endogenous growth theory and explicitly in the New Classical Economics (as in Yang, 2003). In Yang's model, population growth lowers the relative price of labour, thereby increasing the use and number of intermediate capital goods, which are produced with labour. This in turn increases production and the number of manufactured goods, and further bolsters the value of total output through learning-by-doing. In this model, agricultural growth is only indirectly stimulated, for example through the lower cost of manufactured fertilizer – a land-saving input.

Population growth can also facilitate specialization by lowering unit transaction costs. For example, the fixed costs of transport and communication infrastructure per capita may fall sufficiently to warrant additional infrastructure investment. Falling unit transaction costs, in turn, lower the friction that inhibits both horizontal and vertical specialization. In this case, learning-by-doing can directly bolster agricultural productivity.

A primary vehicle for increased specialization is hired labour. To see how population growth can induce hired labour, consider a hypothetical land-surplus economy wherein food is produced by family farms and where clearing costs are negligible. If we assume for the moment that output per hectare is a function of labour, farm size is efficiently determined where the marginal product of land is zero and the marginal product of labour is equal to the shadow price of household leisure. Once population growth brings lower quality, or sufficiently distant, land into production, intensification begins – lowering labour productivity. As the optimal land-to-labour ratio falls, the size of the average family farm declines. This process is efficiently halted, however, due to indivisibilities such as those associated with ploughs and draft animals. Eventually, farm size shrinks to a point where the economies of scale lost from further shrinkage are just offset by the transaction costs of hired labour. At this fundamental turning point, increases in labour per hectare induced by population growth are accommodated by hired labour instead of falling farm size. In this sense, the change in agricultural organization – known as the emergence of the rural proletariat – is not necessarily an indication of exploitation or inefficiency.

But hired labour is not a perfect substitute for family labour. Transaction costs are different, and, since hired labour is not necessarily tied to a particular farm, it can specialize in

particular skills instead of adjusting to the attributes of that farm. In the common case where family labour has a higher shadow price of leisure, hired labour has a comparative advantage in arduous and well-defined tasks wherein transaction costs are manageable (for instance, because the results of the work are readily observable) and wherein speed and quality are enhanced by training and repetition. Family members have a comparative advantage in management-intensive tasks such as chemical applications that require knowledge of farm attributes and for which shirking is harder to control. The advent of hired labour stimulates horizontal specialization across tasks, as with men in the Philippines who specialize in transplanting rice and move from village to village to do so. The resultant learning-by-doing increases productivity – for example, in producing straighter rows of rice, which raise the productivity of workers through the use of rotary weeders. Vertical specialization also increases. For example, landowners may specialize in land improvements, such as irrigation, and employ tenants who specialize in management-intensive labour and who employ and monitor workers who specialize in arduous and more easily supervised tasks. Further vertical and horizontal specialization is illustrated by the institution of piece-rate by teams. A team is hired to complete a task, such as transplanting, which is easily monitored by *ex post* inspection. In this sense, the task is equivalent to an intermediate good. The team may produce, for example, a stack of cane stalks that are of uniform length and ready for planting. Moreover, the team constitutes a separate firm. Its chief executive officer is the team manager, who contracts with the sugar grower and who bears the adverse reputational effects of any sub par performance. In this sense, the capacity for specialization in industry may be quantitatively greater than that of agriculture but not necessarily qualitatively different. Thus it is neither inevitable that population growth decreases or increases productivity in an agricultural economy.

The following stylized pattern of hired labour, based on Philippine rice farming in the 1960s to the 1980s, may serve to epitomize the evolution of specialization as labour intensification follows population growth. Once population density warrants clustered villages of farm families, the institution of exchange labour emerges for transplanting, harvesting, threshing, and often ploughing. Boserupian intensification increases the value of timeliness, and exchange labour allows these tasks to be completed in a day or less for one farm. The first widespread form of hired labour was for harvesting. Harvesters were paid a share of the harvest, typically one-sixth. This later evolved into the gama system, whereby a family or small group was assigned a portion of the farm to weed and later harvest, albeit for the same one-sixth share. This corresponded to a fall in wages relative to rents. In Java, Indonesia, where population pressure was even more intense, this same institution emerged – for the same one-sixth share – but the work requirement expanded even further, typically including transplanting.

When wage labour first appeared in Philippine rice farming, a given worker would typically perform a myriad of tasks over the cropping season. As intensification proceeded and the man-hours of hired labour increased, this undifferentiated wage-worker system was partially replaced by one involving specialized piece-rate workers who were paid according to their performance of a specific task. This evolved further into the piece-rate-by-team system described above. As per-hectare yields continued to increase, piece-rates were often converted back to wage contracts – due to the increased value of quality shirking – but task-by-task specialization was retained.

A common assertion in development economics is that large farms that rely primarily on hired labour are at a transaction-cost advantage relative to small, family farms. This view implicitly takes the distribution over farm size as exogenous, however. In the efficiency view sketched above, farm size is endogenous and responds to changes in population. Indeed, efficient farm size may actually increase as the increased incidence of hired labour warrants new contracting institutions that lower transaction costs. The transaction costs that remain are the necessary cost of retaining economies of scale and facilitating specialization. Whether productivity gains from specialization are enough to offset diminishing returns to more labour on a fixed amount of aggregate land cannot be determined a priori.

The view that share tenancy is inefficient is similarly incomplete. In the canonical view, share contracts are a pair-wise efficient institution for mitigating both the labour-shirking disadvantages of wage contracts and the risk-bearing disadvantages of rent contracts. Nonetheless, share tenancy is said to be socially inefficient because of the Marshallian labour shirking that remains under the common 50 per cent sharing. This view fails to explain how share tenancy fits into the evolution of agricultural organization in response to population pressure and other forces of intensification. Specialization is warranted by intensification and is facilitated by the evolution of contracts and other institutions. In particular, share tenancy facilitates vertical specialization between the landowner, the tenant, and the hired labour that the tenant supervises. It also facilitates the horizontal division of labour described above. On the other hand, share tenancy is primarily a type of family farm and may become less appropriate as agriculture becomes more capital-intensive. In any case, assessing the consequences of institutions without considering their causes, especially intensification, runs a risk of misplaced exogeneity.

A third example of questionable exogeneity concerns the view that the *modernization triad* – population pressure, technical change and commercialization – has inevitably immiserizing consequences. The case made against the new varieties of rice and wheat that emerged in the mid- to late 1960s is illustrative. Modern rice varieties are said to be most profitable on irrigated, highly productive land and for farmers facing relatively low shadow prices of credit and close connections with the money economy. These characteristics tend to favour wealthy landowners over small farm families. As the rich get richer, small farmers and tenants are allegedly disenfranchised, thus accelerating Ricardian forces of population and polarizing society into a class of landlords and the proletariat. Commercialization further augments proletariatization, breaking down safety-net customs such as gleaning rights for the poor, and setting the stage for violent conflict.

The Boserupian and induced innovation perspectives provide a compelling counterweight to the neo-Marxian view. Technical change induced by population growth is primarily land-saving and offsets downward wage pressure, whereas Marxian technical change is strongly labour-saving and exacerbates the downward effect of population. Like induced technical change, induced institutional change in the form of ‘commercialization’ has a positive effect on wages. The efficient emergence of landless workers helps to avoid the immiserizing effects that would occur from a growing population being accommodated by shrinking farm sizes. This class division in turn creates both a supply and a demand for hired labour. As labour markets emerge, new institutions such as piece-rate contracts and work teams with team leaders emerge to lower contracting costs, thereby lowering the transaction cost wedge between effective wage paid, including costs of recruitment, training and supervision, and effective wage received, net of the costs of search, required tools, and the journey to work.

As the unit-transaction-cost wedge shrinks, workers move up their supply curves and employers down their demand curves for labour, resulting in more hired labour and increased net wages. From this perspective, induced innovation at least partially offsets the downward pressure that population pressure puts on wages.

These efficiency patterns are by no means inevitable, but serve to counter the view that the modernization triad is inevitably impoverishing. The efficiency view also provides a theoretical starting point for explaining agricultural growth or the lack thereof. Rent-seeking and policy distortions may induce arbitrary and inefficient patterns of ownership and farm size, thereby inhibiting the efficiency forces described. A challenge for economic historians and agricultural development theorists is to explain the political-economy forces that have facilitated induced innovation in some cases and inhibited it in others.

The positive Boserupian forces of induced innovation and specialization move in the opposite direction of the classical Malthusian effects. To summarize the above, even a small family farm can have four levels of vertical specialization – landowner, share-tenant farm manager, work team leader, and worker – as well as horizontal specialization across the array of farm tasks. The advent of each new form of specialization can be modelled along the lines of Figure 1. Because of the non-convexity associated with the fixed cost of each advance in organizational complexity, population-induced specialization gives rise to increased labour productivity, but only over a limited range of additional labour. In the absence of other effects and changes, we would expect to see the marginal and average products of labour initially rising after each increase in specialization; then, as labour per hectare increases further, to a decline until the next innovation is made. Adding learning-by-doing to the picture increases the chances of sustained productivity gains. Nonetheless, the theory cannot tell us whether the positive forces will outweigh the negative Malthusian forces in the long run.

A historical perspective

The history of agricultural growth is informative. As documented by Evans (1998), the long-run rate of agricultural growth closely matched that of population until 1825, when world population reached one billion people. The corresponding increase in food production was almost entirely sourced in an increase in cultivated area, that is, it was *extensive* in nature. In contrast, since world population reached five billion late in the 20th century, the increase in food production has been almost entirely driven by increased productivity. During the intervening period, when world population increased by four billion, growth in food production was increasingly intensive in nature (due to increased inputs) with increased productivity becoming more important as the period progressed. That is, as intensification led to diminishing returns, increased productivity became increasingly important.

This broad-brush generalization about the nature of agricultural growth is consistent with the induced innovation perspective. As population growth increases land scarcity, the Ricardian gradient, which depicts the proportion of agricultural growth due to intensification, is monotonically rising. Intensification increases the relative scarcity of land further, relative to labour and capital, thus stimulating induced productivity increases, both from technical and institutional progress. Ironically, food supply has grown ‘geometrically’ since 1938 (averaging 2.2 per cent per year) and population has grown nearly ‘arithmetically’ since 1959 (with one billion being added to world population roughly every 13.3 years). Technological and institutional change has seemingly inverted Malthusian theory.

This does not imply that all technological change is demand-induced. Even the theory of induced innovation admits supply-side innovations. For example, knowledge capital produced in the defence industry may lead to better communications technology. Irrigation systems in ancient Mesopotamia and Egypt were presumably not induced by increasing land scarcity but because someone figured out how to produce more with less. Economic history in the United States suggests that demand was partly induced by labour scarcity, but, once certain types of farm equipment had been invented, they were adapted even in areas where land prices were increasing faster than labour prices. Kremer (1993) even suggests that until the late 18th century the Malthusian argument was so predominant that population could be viewed as a proxy for technological change.

On the other hand, the agricultural and industrial ‘revolutions’ are now viewed less as bursts in productivity spurred by invention and more as induced technical change. For example, the four-field system, whereby wheat, barley, turnips and clover were grown in separate fields and rotated the following year, was once viewed as an essential part of the English agricultural revolution during the 18th century. But the system was developed in land-scarce Flanders two centuries before and popularized in England only once it was warranted by sufficient population-induced land scarcity.

Even the mechanism of induced technical change is not entirely governed by factor prices, however. For example, the replacement of the fallow period in the medieval ‘three field’ rotation by beans or another leguminous crop appears to have been indirectly induced by the population decline in 14th-century western Europe. Higher wages and farm incomes, resulting from the lower population and decreased land scarcity, increased the demand for meat. Complemented with the Flemish demand for wool, this incentivized farmers to increase sheep production, and they responded by both converting some lands to pasture and growing legumes in place of fallow on much of the remaining lands.

The extent to which technical change in English agriculture was induced has been the subject of intense historical debate. Historians reporting that agricultural productivity increased rapidly, say in the late 18th and early 19th centuries, tend to see an agricultural revolution stimulated by exogenous technical change. Economic historians who estimate productivity increases to be quite gradual view changes in rotation and other innovations as induced. As suggested by the discussion of Figure 1, induced changes do not by themselves reverse the price and income trends that induced them in the first place and therefore tend not to be associated with dramatic increases in productivity.

Sustainable development

Resource depletion adds another negative dimension to the never ending debate between the development optimists and pessimists. Even before *sustainable development* became fashionable, neo-Malthusians argued that unbridled population growth in poor countries and economic growth in rich countries must inevitably cause severe pressure on the earth's limited resources, resulting in burgeoning poverty and international conflict. The only solution was said to be the *steady state economy* with constant population, capital stock, and output.

After the Brundtland Commission's 1987 report, resource depletion was broadened to include pollution and other environmental threats. Environmental degradation, including increasing water scarcity, soil erosion, deforestation, desertification, salinization, and global warming, as well as diminishing energy and marine resources, was viewed as exacerbating the Malthusian vicious circle. Accordingly, the Brundtland Commission called for a simultaneous assault on population growth, poverty and environmental degradation, thus giving rise to the modern movement for *sustainable development*. Economists have had limited success in modelling sustainable development, however. One notable review and synthesis (Arrow et al., 2004) was unable to settle on positive principles of sustainability and settled on the negative *sustainability criterion* – an injunction not to deplete the value of natural capital more than the additional value of produced capital.

Even if we abstract from technical change, expanding models of economic growth to include environmental degradation does not produce a necessarily dismal outlook, however. If we represent concern for future generations by *intergenerational neutrality* and assume that population grows exponentially at a constant rate, optimal per capita consumption grows to its *golden rule* level, under plausible assumptions about substitutability, both between renewable and non-renewable resources and between natural and produced capital. Adding technical change provides even rosier possibilities (Weitzman, 1997). Whether these possibilities are realized depends largely on the effectiveness of private and public governance structures in facilitating specialization and exchange while guarding against unproductive rent seeking (Greif, 2006).

The co-evolution of specialization and governance

The economic history of Hawaii provides a relatively recent, pre-industrial example of how specialization and governance in agriculture co-evolve with changes in population. During the ‘colonization’ period ad 300–600, population growth, including further migration of Polynesian peoples, was slow. Agricultural expansion was extensive. The population began to increase more rapidly towards the latter part of the ‘development’ period (600–1100), and agriculture began to intensify with the advent of irrigation. There was little if any division of labour among the commoners. During the expansion period (1100–1650) population accelerated and intensification greatly increased with a decreased fallow period, a major expansion in irrigation and with the development of fishponds. Horizontal specialization among workers became commonplace, with fishing more of a distinct occupation. Evolving from a system of somewhat separate extended families units, social and production relations became increasingly stratified, eventually with a distinct hierarchy from local chief upwards to governor (*ali’i*) of the watershed to district head (see Kirch, 1985).

This stylized history is suggestive of a *governmental Kuznets curve*. During the extensive (pioneer) stage of development, family or extended family units are largely autonomous and decision-making is decentralized accordingly. During the intensive development stage, decision-making and governance are centralized at a higher, albeit intermediate level (for example, communal governance of the commons). As intensification and specialization continue, efficiency favours a further centralization of governance, at least for the minimal functions of defence and the justice system, but a decentralization of decision-making as facilitated by private property. This last stage occurred in Hawaii after Western contact in 1778. New trade opportunities raised the value of irrigation and other investments in plantation agriculture, initially for sugar and later pineapple. Private property provided the assurance that planters needed to commit to these investments and also facilitated specialization between districts that was warranted by international trading opportunities. Graphing this historical progression of increasing governmental centralization on the horizontal axis, and rising and then falling centralization of decision-making on the vertical axis, completes the governmental Kuznets curve. Viewing government intervention in these two dimensions provides a useful antidote to the misleading question of ‘how much government’ that sometimes arises in policy circles.

Smith to Malthus to Solow

A largely unexplored area of enquiry involves combining the theory of endogenous population growth with the theory of sustainable growth outlined above. Perhaps the simplest model of endogenous growth can be found in two-sector growth models of economic development wherein the birth rate is exogenous and the death rate declines to minimum as per capita income increases. The birth rate may also be made endogenous following the Chicago School's new household economics. The increased opportunity cost of child care is one pervasive cause of the decline in fertility with economic development. Moreover, as the capital intensity of the economy increases, the returns to human capital are raised, thus creating incentives for families (individually or collectively) to invest in human capital, a partial substitute for increased fertility.

Malthus's emphasis on the supply of food determining population and Boserup's focus on exogenous population growth increasing the demand for land and inducing supply side changes in agricultural production are clearly complementary. Focusing on one or the other is a device for dealing with the shortcomings of human imagination and the fact that models with both forces are indeterminate without further, possibly arbitrary, restrictions added to the model. Indeed, due to the endogeneity of population, enquiring into the impact of population levels involves something of a category mistake. In light of this, the World Bank statement (1984; see also Kelley, 1988) that population growth in excess of two per

cent per annum tends to have a negative impact on per capita income warrants reinterpretation. A more accurate statement would be that population growth in excess of two per cent tends to be associated with negative growth in per capita income after partially controlling for (imperfectly measured) positive effects. In particular, where high population growth occurs in the face of policy failures that cause an anti-labour bias, population growth tends to exacerbate the Brundtland vicious circle described above.

More generally, the effects of population growth on agricultural and economic development may be different depending on the population density and the stage of economic development, as illustrated in Figure 2. For the early American frontier and for parts of Africa today, physiological population density may be sufficiently sparse for Smithian economies of specialization and Boserupian economies in infrastructure to afford increasing labour productivity, as shown by the rising segment of the average product of labour curve. There is no labour market, at least in the sense of a competitive spot market, in such economies because paying labour its marginal product would more than exhaust total output. When the extensive land frontier nears economic exhaustion, population density becomes high, and the economy is still dominated by agriculture (as on the Indonesian island of Java in the 1960s and early 1970s), real wages fall, along with the average product of labour. Once the ‘structural transformation’ takes place, such that the growth rate of the agricultural labour force (if any) is but a small fraction of that of the industrial labour force, the marginal product of labour begins to rise, causing wage rates to rise and pulling up average labour productivity soon thereafter. Accumulation of produced capital and the relative increase of the industrial sector generate the transition to modern economic growth.

Figure 2

Stages of economic development

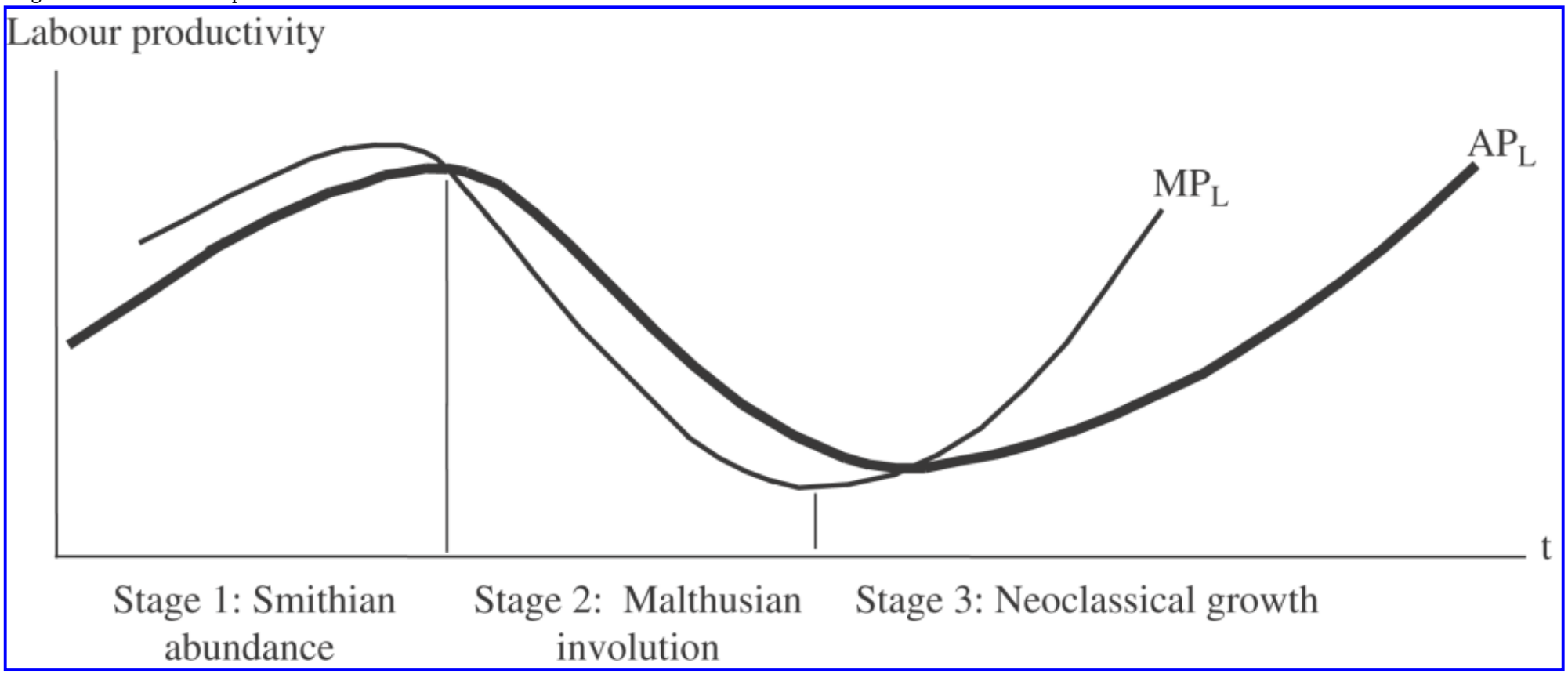

These stages are not inevitable forces of history. Some economies may be able to bypass the Malthusian stage altogether. For example, economic policies in Taiwan during the 1950s and 1960s encouraged labour intensity in agriculture. This and the investments in physical infrastructure, a gradual transition to processing and high value-added agricultural production and an efficient system of marketing cooperatives kept the demand for labour and wages rising. Hong Kong and Singapore were able to skip the Malthusian stage by early industrialization that relied on trade instead of the Johnston–Mellor linkages whereby agricultural development increases incomes (thus stimulating demand for industrial products), mobilizes savings for industrial investments, and provides a market for manufactured farm inputs (Johnston, 1970). Korea was similarly able to bypass an extended Malthusian stage by allowing investment coordination through *chaebols* (business groups)and focusing on manufactured exports. In contrast, the negative force of policy failures can extend Malthusian involution and even prevent the transition to modern economic growth. Finally, because of policy failures and exogenous shocks, history may record more than two

turning points. For example, after going through a Malthusian period during the ‘long 16th century’, wages in England rose between approximately 1640 and 1740, but then fell again before entering a ‘Solovian’ period of increase starting slightly after the advent of the 19th century and accelerating after the American Civil War.

Nonetheless, we may meaningfully enquire into the mechanics of the two turning points shown, after abstracting from policy failures and exogenous shocks. While the first turning point has clear Ricardian underpinnings, the second has generated substantial controversy. How does an economy go from ‘Malthus to Solow?’ Forward linkages from agriculture are important in explaining the relative growth of industry, but they do not, in and of themselves, explain the rapid and sustained growth in labour productivity during modern economic growth.

Note first that there is an implicit Kuznets curve corresponding to Figure 2. During the Malthusian period, wages fall and Ricardian rents increase, worsening income distribution. Even as industrialization begins to pull up wages, income distribution may continue to worsen for some time as the total returns to capital increase faster than the wage bill. Eventually, as the returns to human capital induce the substitution of ‘quality for quantity’ in fertility decisions, widely distributed human capital accumulates and even produced capital becomes less concentrated. These forces cause a more equal income distribution in the model.

Were it only for Ricardian landlords accumulating an agricultural surplus and financing industrialization and the production of goods for a landed aristocracy, industrialization would have not have been as robust as that witnessed in modern economic growth. Indeed, increasing wages stifle the labour-intensive production that characterizes the early stages of industrialization, decrease the agricultural surplus, and detract from the rental incomes of capitalists and landlords that finance capital formation. What saves the day are the non-convexities inherent in industrialization.

While there are numerous possibilities for specialization and other non-convexities in agriculture, these are still few in comparison with those in industry. In industry, there is more horizontal specialization through proliferation in the number of products and more vertical specialization through multiple stages of intermediate production. In agriculture, the number of products is more limited, and vertical specialization without industry tends to be limited to separation of management and labour. With industry, agriculture can take advantage of land-saving intermediates such as fertilizer and tractors. Thus it is plausible that technological and institutional changes in agriculture have not been frequent enough to overcome the inexorable Malthusian force of increased food affording greater population growth.

In contrast, once industry becomes a major part of the economy, non-convexities may be sufficiently compact in the course of development to dominate the negative force of lower death rates. The resultant increase in per capita income in turn invokes a positive feedback mechanism whereby Engel effects increase the demand for manufactures, thus increasing capital formation and the returns to human capital, thereby contributing to the decline in the demand for child numbers described above. Greater product specialization and falling unit transport costs afford a further inducement to international trade, an additional positive feedback mechanism. This theory supports the revisionist interpretation that the agricultural and industrial ‘revolutions’ were misreadings of a gradual process of economic change (see Clark, 2007).

The role of industrial development in sustaining increased wages and per capita incomes does not imply that the appropriate development policy requires pushing industrial development while ‘squeezing’ or neglecting the agricultural sector. Indeed, for countries with a preponderance of the labour force in agriculture, economic development can be sustained only by ‘pushing’ on the agricultural sector with R&D, infrastructure, and non-confiscatory prices (Pingali, 2006). It does mean, however, that stimulating the agricultural sector alone – that is, relying on automatic linkages from the agricultural to the industrial sector – is not sufficient for sustained economic development. External economies of labour-market pooling, human capital, technological spillovers and other network externalities imply that there are aspects of investment coordination that are not internalized by spot markets. This leaves an important role for government in facilitating the requisite economic cooperation.

See Also

- agriculture and economic development
- institutional economics

Bibliography

Ahmad, S. 1966. On the theory of induced invention. *Economic Journal* 76, 344–57.

Arrow, K.J., Dasgupta, P., Goulder, L.H., Daily, G.C., Ehrlich, P., Heal, G., Levin, S., Maler, K.-G., Schneider, S.H., Starrett, D. and Walker, B. 2004. Are we consuming too much? *Journal of Economic Perspectives* 18(3), 147–72.

Barrett, S. 1991. Optimal soil conservation and reform of agricultural pricing policies. *Journal of Development Economics* 36, 167–87.

Barrett, C.B., Reardon, T. and Webb, P. 2001. Non-farm income diversification and household livelihood strategies in rural Africa: concepts, dynamics and policy implications. *Food Policy* 26, 315–31.

Binswanger, H.P. and Ruttan, V.W. 1978. *Induced Innovation: Technology, Institutions, and Development*. Baltimore, MD: Johns Hopkins University Press.

Boserup, E. 1965. *The Conditions of Agricultural Growth*. London: Allen & Unwin.

Boserup, E. 1981. *Population and Technological Change: A Study of Long-term Trends*. Chicago: University of Chicago Press.

Boserup, E. 1987. Agricultural growth and population change. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Brundtland Commission. 1987. *Our Common Future*. Oxford: Oxford University Press.

Clark, G. 2007. *A Farewell to Alms: A Brief Economic History of the World*. Princeton: Princeton University Press.

Evans, L.T. 1998. *Feeding the Ten Billion: Plants and Population Growth*. Cambridge: Cambridge University Press.

Evenson, R.E. 1982. The Green Revolution in North Indian agriculture: ex post assessment. Mimeo, Yale University.

Evenson, R.E. 2004. Food and population: D. Gale Johnson and the Green Revolution. *Economic Development and Cultural Change* 52, 543–70.

Greif, A. 2006. *Institutions and the Path to Economic Modernity: Lessons from Medieval Trade*. Cambridge: Cambridge University Press.

Hayami, Y. and Ruttan, V.W. 1985. *Agricultural Development: An International Perspective*. Baltimore, MD: Johns Hopkins University Press.

Johnston, B.F. 1970. Agriculture and structural transformation in developing countries: a survey of research. *Journal of Economic Literature* 8, 369–404.

Kelley, A.C. 1988. Economic consequences of population change in the Third World. *Journal of Economic Literature* 26, 1685–728.

Kirch, P.V. 1985. *Feathered Gods and Fishhooks: An Introduction to Hawaiian Archaeology and Prehistory*. Honolulu, HI: University of Hawaii Press.

Kremer, M. 1993. Population growth and technological change: one million bc to 1990. *Quarterly Journal of Economics* 108, 681–716.

Krautkraemer, J. 1994. Population growth, soil fertility, and agricultural intensification. *Journal of Development Economics* 44, 403–28.

Lucas, R.E. 2002. *Lectures on Economic Growth*. Cambridge, MA: Harvard University Press.

Lueck, D. 1998. First possession. In *The New Palgrave Dictionary of Economics and the Law*, ed. P. Newman. New York: Macmillan.

Malthus, T.R. 1798. *An Essay on the Principle of Population*. Oxford: Oxford University Press, 1999.

North, D.C. and Thomas, R.P. 1973. *The Rise of the Western World: A New Economic History*. New York: Cambridge University Press.

Pender, J. 2001. Rural population growth, agricultural change and natural resource management in developing countries: a review of hypotheses and some evidence from Honduras. In *Population Matters: Demographic Change, Poverty and Economic Growth in Developing Countries*, ed. N. Birdsall, S. Sinding and A. Kelley. Oxford: Oxford University Press.

Pingali, P. 2006. Agricultural growth and economic development: a view through the globalization lens. Presidential Address, International Association of Agricultural Economists 26th Conference, Queensland, Australia.

Ricardo, D. 1817. *On the Principles of Political Economy and Taxation*. London: John Murray.

Roumasset, J. 2004. Rural institutions, agricultural development, and pro-poor economic growth. *Asian Journal of Agriculture and Development* 1, 61–82.

Ruttan, V.W. and Hayami, Y. 1973. Technology transfer and agricultural development. *Technology and Culture* 14, 119–151.

Weitzman, M.L. 1997. Sustainability and technological progress. *Scandinavian Journal of Economics* 99, 1–13.

World Bank. 1984. *World Bank Development Report: Population Change and Development*. Oxford: Oxford University Press.

Yang, X. 2003. *Economic Development and the Division of Labor*. New York: Blackwell.

How to cite this article

Roumasset, James. "population and agricultural growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000065> doi:10.1057/9780230226203.1308

The New Palgrave Dictionary of Economics Online

population dynamics

Ronald D. Lee

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Population dynamics are the patterns of change over time in populations. Populations fluctuate in response to fluctuating external forces, or because of the internal structure of the process of demographic renewal. Damped cycles one generation long may result from the interaction of random perturbation and the age distribution of reproduction. So-called Easterlin cycles two generations long, either damped or self-exciting, may arise from the lag between birth and labour force entry when fertility responds sensitively to labour market conditions. Longer-term dynamics arise from the interactions of population growth, capital, endogenous technology, and income.

Keywords

Easterlin cycles; Easterlin hypothesis; fertility; Kondratieff cycles; Kuznets swings; Malthus, T.; Malthus's theory of population; net maternity function; net reproduction rate; population and economic growth; population density; population dynamics; Solow, R.; stable population theory; technical progress

Article

Population dynamics are the patterns of change over time in populations, ranging from fluctuations to long-term trends, and the underlying principles that govern these changes.

Population fluctuations

All human populations exhibit fluctuations in their vital rates and consequent irregularities in their age distributions to a greater or lesser degree. Analyses of such fluctuations are of interest for many reasons – for historical understanding, as a basis for forecasting, for a deeper understanding of underlying social processes – but perhaps most intriguing is the possibility that they may afford some insight into more fundamental aspects of population dynamics and may illuminate the very process of demographic

renewal. More specifically, we may be able to learn from the occurrence or absence of longer cycles whether a population is subject to negative feedback of a Malthusian sort and perhaps to place bounds on its sensitivity if it occurs. To Malthus (1798) it seemed obvious that populations would perpetually oscillate about equilibrium. This notion is taken seriously as an interpretation of the long swings in the fertility of many contemporary developed countries, as we discuss in more detail below.

Fluctuations may come about in three ways (or through combinations of these ways). First, they may simply be imposed on a series of births or deaths by fluctuations in some driving force such as prices or the weather. In this case, both the amplitude and the period of the fluctuation depend entirely on the driving series. Second, damped fluctuations may be created by the internal structure of a demographic process, as it responds to random and non-cyclic external shocks; in this case the cycles will die out if the external disturbance stops. The period of such cycles depends entirely on the nature of the renewal process, not on the driving force; however, the amplitude of the cycles depends on the amplitude (variance) of the disturbing force.

The third possibility is that limit cycles occur. Like the aforementioned cycles, these are generated by the internal structure of the reproductive process, but unlike them they are self-sustaining or ‘self-exciting’ and would continue indefinitely even in the absence of outside shocks. In this case, both the amplitude and the period depend only on the reproductive process. When a dynamic equilibrium is unstable, such that trajectories tend to explode away from the equilibrium path, then one of three things may happen: explosive fluctuations may lead to extinction; the non-repeating fluctuations of chaos cycles may occur; or the system may settle down to a limiting pattern of cycles, called limit cycles. There are many examples of animal populations exhibiting such behaviour. In human demography, it is a matter of controversy whether such cycles have ever actually occurred, but if they have it is presumably through the kind of mechanism proposed by Easterlin (1968), a sort of Malthusian cycle about equilibrium.

Imposed cycles

There are well known non-seasonal cycles in fertility and mortality at or below the annual frequency (obstetricians avoid deliveries on Sundays; people have lower mortality just before elections, compensated for by increased mortality thereafter). Seasonality is strong in fertility, mortality, nuptiality and migration, particularly in traditional agricultural societies and in those less insulated by their dwellings from the variations of climate (in the extreme case of Bangladesh in the 1970s, for example, the seasonal peak in fertility was two to three times the seasonal trough). In the case of mortality, nuptiality and migration the causes of seasonal variation are fairly well understood to be rooted in identifiable biological, institutional and economic influences. In the case of fertility, the causes of seasonality are much less well understood.

There are also somewhat longer fluctuations in vital rates, in the range of 2–15 years. These have been quite thoroughly studied and found to be associated with business cycle indicators in the developed world and with the harvest cycle in pre-industrial conditions. Lower agricultural prices and less unemployment are associated with higher fertility and nuptiality and with lower mortality, with lag patterns of response indicating that much of the variation is confined to changes in the timing of events. Fluctuations in temperature also are important, with colder winters and hotter summers raising mortality and reducing fertility, with an appropriate lag. In the case of mortality, exogenous epidemiological

variation historically played a larger role (Wrigley and Schofield, 1981). These relationships have continued to hold at least until a few decades ago in the developed countries and are still evident in the Third World countries where they have been investigated.

Much longer fluctuations in population variables are also visible in the historical record. Kuznets cycles, of 15–25 years, include a pro-cyclical response of migration, both internal and international. Some of the birth-rate series of 19th century Europe show signs of the Kondratieff cycle. But most striking are the waves lasting two or three centuries in the demography of Europe and of China, from at least the 12th century up to the 18th (Wrigley and Schofield, 1981). These are evident in population growth rates and in mortality; their existence in fertility is problematic. The cause of these very long waves is not clear, although a case can be made for the influence of climatic variation and for the effects of intercontinental exchange of diseases through conquest or trade. Whatever their cause, such demographic fluctuations played a critical role in economic history, driving rents, wages and other relative prices, and possibly inflation. It is possible that such fluctuations were generated internally by the economic demographic system as Malthusian fluctuations about equilibrium; in the present state of knowledge, however, it appears more likely that the cycles were imposed.

Cycles arising from the internal age and temporal structure of reproduction

A characteristic pattern of delay between an event and its recurrence can act as a filter which creates quasi-cyclic behaviour in the series of events when the timing is subject to continual random perturbation. In this way, the typical spacing of a mother's births two to three years apart tends to generate cycles of this length, as was first pointed out by Yule (1906). Such cycles are visually discernible in many birth and fertility series and show up in the empirical power spectra.

More importantly, the typical delay between a woman's own birth and the time she herself gives birth to female children leads to cycles of 25–35 years, or the approximate length of a generation, when fertility is randomly perturbed (see Coale, 1972; Lee, 1974). This may be shown as follows. Let $B(t)$ be the number of births in year t, and let $\varphi(a)$ be the expected number of births to each of these births at age a, net of mortality ($\varphi(a)$ is known as the ‘net maternity function’). $\varphi(a)$ typically rises from zero at an age around 15 years to a peak in the twenties and declines again to zero at around age 45; its mean, μ, is the mean age at child-bearing and falls between 25 and 35 years depending on the population. The renewal process is written:

$$B(t) = \sum \varphi(a) B(t-a) \tag{1}$$

where the sum is taken over the reproductive years. Such a process will settle down to a stable exponential growth path if the characteristic roots of ϕ lie within the unit circle. But as the B series converges to this growth path from an irregular past, it will fluctuate, and the fluctuations can be characterized by further examination of the characteristic roots of ϕ. There will generally be one real

root, describing the steady state growth rate, and the others will come in pairs of complex conjugates. The pair with the largest modulus is the only one of substantive interest; it will describe a damped oscillation with length roughly equal to the mean age of child-bearing, μ. Any initially distorted age distribution, if subsequently subjected to fixed vital rates described by ϕ, will generate a birth sequence which moves in waves one generation long as it converges towards exponential growth.

The argument can easily be generalized to cover the case of a population whose net maternity function is subject to constant stochastic disturbance of any autocovariance structure; the age structure of reproduction, described by the mean values of ϕ, will amplify variation in the neighbourhood of frequencies corresponding to cycle length μ, leave them unchanged at higher frequencies, and attenuate them in the neighbourhood of cycle length 2μ. Thus, a population in a random environment will tend to exhibit cycles one generation long or to superimpose these on whatever pattern of variation is forced on it by the environment. The birth series of many pre-industrial populations, particularly at the parish level, indeed do reveal such waves; whether the mechanism described above suffices to account for them has not yet been established empirically.

Some scholars have seen a major economic influence in such population waves, but this view now appears exaggerated; waves generated in this way are generally quite mild; they have low amplitude, and they damp fairly rapidly following an identifiable disturbance.

Cycles arising from economic-demographic interaction

Interest in dynamic economic-demographic models of population renewal, stressing fluctuations arising from age distributions, was prompted by the long ‘cycle’ in US fertility, with a trough in the 1930s, a peak in the late 1950s, and a trough in the 1970s. A number of scholars, most notably Easterlin, suggested around 1960 that the fertility fluctuations might reflect the economic conditions faced by young labour market entrants, conditions which in turn were worse for large cohorts and better for smaller ones. This insight led them to forecast correctly the sharp decline in fertility occurring in the 1960s, as larger cohorts aged into the labour market. Easterlin (1968) developed a detailed theory, buttressed by extensive empirical investigation, leading to a tentative prediction of self-generating demographic cycles two generations long, as small birth cohorts had high fertility and gave birth to large cohorts, who in turn reared small cohorts, and so on. Such cycles are known as ‘Easterlin cycles’. A considerable empirical literature has since appeared on the subject, lending considerable support at the aggregate time series level in the United States and some other countries, but very little at the micro level.

I now briefly review the theoretical literature on economic-demographic cycles. The account of the renewal process given above implicitly assumed that net maternity at time t, $\phi(a,t)$, was independent of the population age distribution at time t, or equivalently of the preceding series of births. But it is entirely possible that this is not so. Suppose, for example, that a Malthusian model is appropriate, such that fluctuations in labour supply lead to inverse fluctuations in wages, and that fertility depends positively on the wage level. This leads to different dynamic possibilities and a modified renewal equation.

Suppose that the net maternity function, $\phi(a)$, depends on some set of economic variables, let us say wages for concreteness. Suppose that these in turn depend on some set of economic variables, Z, which are independent of age distribution, as well as on the current population age distribution, which thus in

conjunction with *Z* determines wages. If mortality is constant and the population closed to migration, as we here assume, then the current age distribution is completely determined by past births. We can then write:

$$B(t) = \sum \varphi^{*}[B(t), Z(t)]B(t-a), \tag{2}$$

where *B*(*t*) denotes the vector of past births; this replaces the purely demographic renewal eq. (1) introduced above (Lee, 1974).

The renewal process will have an exponential equilibrium growth path, $B^{*}(t) = B \cdot \exp(nt)$, which satisfies (2) for all *t*. For simplicity, suppose that *Z* is such that $n = 0$, so that the equilibrium path is stationary. It is helpful to consider the process of proportional deviations about this equilibrium path, denoted *b*(*t*). Let $\varphi(a)$ be the value of $\varphi^{*}[\ ]$ evaluated at equilibrium. In this case, the sum of $\varphi(a)$ over all *a*, known as the net reproduction rate or NRR, is unity when evaluated at the equilibrium age distribution. Let $\Gamma(a)$ be the elasticity of the NRR with respect to the size of age group *a*, or equivalently with respect to births *a* years previously, *B*(*t*–*a*); these elasticities are readily derived from the original function ϕ. Then the renewal process for fluctuations about the equilibrium growth path of births is simply:

$$b(t) = \sum [\varphi(a) + \Gamma(a)]b(t-a). \tag{3}$$

The smaller the effect of the current age distribution on fertility (Γ), the more the population renewal process resembles the purely demographic version of (1). In any event, exactly the same procedures can be used to study the dynamic behaviour of birth fluctuations in this model as were used previously.

The first step is to check the characteristic roots to assess stability. If the oscillations of the process tend to explode away from the equilibrium growth path, then a different kind of analysis, discussed below, is called for. If the roots indicate that oscillations are damped, then the analysis of dynamic behaviour in the neighbourhood of equilibrium will be informative.

We can now consider specifications of the model which have been proposed in the literature. The first is the simplest Malthusian model, in which all age groups in the labour force are assumed to be perfect substitutes in production, and fertility at each age is assumed to be negatively related to the size of the potential labour force, through an hypothesized effect on wages. In this case, $\Gamma(\alpha) = \beta k(a)$, where β is independent of age, and expresses the sensitivity of response (elasticity of the net reproduction rate with respect to labour force size at equilibrium), while the *k*(*a*) depend only on mortality conditions and equilibrium age specific labour supply and are therefore easily calculated from data at hand. Depending

on values of θ, this model will generate cycles ranging from one generation (as in the purely demographic model) to a century and a half or more. For $\theta = 7.5$, which is the empirical estimate from US data, 1917–1973, a cycle corresponding to the observed time path of births may be produced (Lee, 1974; Wachter, 1991).

Another model which is often used makes the fertility of a birth cohort depend only on the size of the cohort and makes it independent of all other age group sizes. The simplest form of this specification leads to:

$$b(t) = (1 - \alpha)\sum \varphi(a)b(t - a), \tag{4}$$

where α is the elasticity of each age's fertility with respect to cohort size. For α less than 1, there is a generation-long cycle; for α greater than 1 but less than 2, there is a damped two-generation cycle, and for α greater than 2 an explosive two-generation cycle occurs.

Specifications reflecting other degrees of substitutability of age groups of labour could of course be tried. Easterlin typically has used a ratio of younger to older workers to drive fertility (this could be derived from a CES model with two age groups of labour as separate factors, for example). The general expression can be used to explore dynamics under a wider variety of specifications. For example, the burden of supporting the elderly retired population might lead to a reduction in fertility; this would be expressed as a suitable negative $\Gamma(\alpha)$ for $a = 65$ and over. If couples were led to desire larger families when they observed other couples' children, then $\Gamma(\alpha)$ would be positive for ages zero to ten.

When the cyclic behaviour near equilibrium is found to be explosive, then we need to consider behaviour further from equilibrium, at which point nonlinearities become important (unless, of course, the behaviour is truly linear, in which case population extinction results). Dynamic behaviour can be 'chaotic', an endless series of non-repeating fluctuations; for many models, however, limit cycles will occur, with amplitude and period determined not by the pattern of disturbances but rather by the functional relations themselves. Such cycles are observed in animal populations in laboratories and occasionally in the wild; in human populations their occurrence is conjectural: Samuelson (1976) considered a particular three-age group model leading to limit cycles.

Long-term population trends and economic growth

Longer-term trends in population have also been viewed in the context of processes related to economic growth. Solow (1956) studied the behaviour of a population whose growth varied first positively and then negatively with respect to per capita income. Combining this study with his neoclassical growth model, he showed there was a stable low-level equilibrium at which per capita income was constant and population grew at the rate of technological progress, but also a second equilibrium at a high per capita income which was unstable. If the capital–labour ratio could be raised slightly above this equilibrium level, then per capita income would rise without limit while the population growth rate fell lower and

lower.

In Solow's approach, as in Malthus's, technological progress was taken as exogenous. Boserup (1981) and others have suggested that larger denser populations would be more likely to experience technological progress in the long run, for reasons related to both the supply of innovations and the demand for them. She suggested that, combined with a Malthusian endogenous response of population growth to economic progress, an upward spiral of population growth and technological progress might occur, with positive feedback. A number of scholars have developed formal models of this process (Lee, 1986; Kremer, 1993), in a literature that overlaps slightly with the endogenous growth literature (Jones, 2003).

See Also

- Easterlin hypothesis
- Kondratieff cycles
- Kuznets swings
- stable population theory

Bibliography

Boserup, E. 1981. *Population and Technological Change*. Chicago: University of Chicago Press.

Coale, A. 1972. *The Growth and Structure of Human Populations: A Mathematical Investigation*. Princeton: Princeton University Press.

Easterlin, R. 1968. *Population, Labor Force, and Long Swings in Economic Growth*. New York: National Bureau for Economic Research.

Jones, C. 2003. Population and ideas: a theory of endogenous growth. In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, ed. P. Aghion et al. Princeton: Princeton University Press.

Kremer, M. 1993. Population growth and technological change: 1,000,000 B.C. to 1990. *Quarterly Journal of Economics* 108, 681–716.

Lee, R. 1974. The formal dynamics of controlled populations and the echo, boom and the bust. *Demography* 11, 563–85.

Lee, R. 1986. Malthus and Boserup: a dynamic synthesis. In *The State of Population Theory*, ed. D. Coleman and R. Schofield. Oxford: Basil Blackwell.

Lee, R. 1997. Population dynamics: equilibrium, disequilibrium, and consequences of fluctuations. *Handbook of Population and Family Economics*, v.1B, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Malthus, T. 1798. *An Essay on the Principle of Population*, ed. A. Flew. Harmondsworth: Penguin, 1970.

Samuelson, P. 1976. An economist's non-linear model of self-generated fertility waves. *Population Studies* 30, 243–7.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Wachter, K. 1991. Elusive cycles: are there dynamically possible Lee–Easterlin Models for US births? *Population Studies* 45, 109–35.

Wrigley, E. and Schofield, R. 1981. *The Population History of England 1541–1871: A Reconstruction*. Cambridge, MA: Harvard University Press.

Yule, G. 1906. Changes in the marriage and birth rates in England and Wales during the past half century. *Journal of the Royal Statistical Society* 69(1), 18–132.

How to cite this article

Lee, Ronald D. "population dynamics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000124> doi:10.1057/9780230226203.1309

The New Palgrave Dictionary of Economics Online

population health, economic implications of

David Canning and David E. Bloom

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Population health is not only a consequence but also a cause of a high level of income. Healthier people are more productive in work. Healthy children have better school attendance and cognitive development, while longer prospective working lifespans encourage investments in education. Longer lifespans can also increase saving and wealth accumulation as an extended retirement becomes more likely. The beneficial effects of population health can be seen both at the individual and macroeconomic levels, while the continuing high burden of disease in sub-Saharan Africa poses a substantial challenge to its economic development.

Keywords

age structure; Black Death; cognitive ability; demographic transition; Dependency; economic growth; education; family planning; fertility; foreign direct investment; health; health policy; HIV/AIDS; human capital; income; instrumental variables; investment; labour productivity; labour supply; learning; life expectancy; morbidity; mortality; population growth; population health; retirement; savings; school attendance; social security; value of life; well-being

Article

Population health and a high level of income go hand in hand. Higher incomes promote better health through improved nutrition, better access to safe water and sanitation, and increased ability to purchase more and better quality health care. There is also, however, an effect of health on income. This can work through several mechanisms (Bloom and Canning, 2000). The first is the role of health in labour productivity. Healthy workers lose less time from work due to ill health and are more productive when working. The second is the effect of health on education. Childhood health can have a direct effect on cognitive development and the ability to learn. In addition, because adult mortality and morbidity (sickness) can lower the prospective returns to investments in schooling, improving adult health can

raise the incentives to invest in education. The third is the effect of health on savings. A longer prospective lifespan can increase the incentive to save for retirement, generating higher levels of saving and wealth, and a healthy workforce can increase the incentives for business investment. We examine the evidence for these mechanisms and find that there are potentially large effects of health on economic outcomes at both individual and macroeconomic levels.

Improved population health has a large impact on population numbers and age structure, and we examine the economic implications of this induced demographic change. The global population explosion of the 19th and 20th centuries was caused not by a rise in fertility but by a fall in mortality. Lower mortality and improved survival rates increased population numbers, but also led to significant increases in the number of young people since the largest improvements in mortality are initially in infant mortality rates. In the long run, reductions in infant mortality lead to a fall in desired fertility, creating a one-time baby-boom cohort. As this large cohort ages, the resultant changes in population age structure can have significant economic implications.

The issue of population health and economic outcomes is particularly acute in sub-Saharan Africa. This region has a high burden of tropical and other infectious disease, such as malaria, tuberculosis, and intestinal worms, and it also suffers from the HIV/AIDS pandemic. We examine the impact of this disease burden on the prospects for economic development in sub-Saharan Africa.

Although we focus on the economic implications of population health, there is clearly two-way causality as health is partly a consequence of income levels. Preston (1975) demonstrated a positive correlation between national income levels and life expectancy. One reason for this link is that higher income levels allow greater access to inputs that improve health, such as food, clean water and sanitation, education, and medical care. Fogel (2004) emphasizes the role of access to food while Deaton (2006) puts more weight on public health measures such as clean water and sanitation (see Cutler and Miller, 2005). Cutler and McClellan (2001) examine the increasing contribution of medical care to health outcomes. Pritchett and Summers (1996) use the relationship between income levels and health to argue for an emphasis on economic growth in poor countries as a method of increasing population health. However, the findings of Easterly (1999) weaken this argument. Easterly finds that, although income levels and population health are closely related, the effect of changes in income on population health over reasonable time spans appears to be quite weak. By contrast, relatively inexpensive public health interventions and policies can have remarkable impacts on population health even in very poor countries. In practice, the major force behind health improvements has been improvements in health technologies and public health measures that prevent the spread of infectious disease, and not higher incomes (Cutler, Deaton and Lleras-Muney, 2006).

We examine the role of health as an instrument to generate economic well-being. However, any reasonable view of the contribution of health to human welfare would also include the direct welfare benefits of a long lifespan and good health. Estimates of the monetary value of life (as measured by the willingness to pay to avoid a small risk of death) are often very large (Viscusi and Aldy, 2003). We can use these estimates of the value of life to compare the welfare improvements that have come about due to improvements in population health and the improvements due to economic growth and higher incomes. Such comparisons suggest that in many countries the value of health gains has been comparable to, or has even surpassed, the value of income gains (Nordhaus, 2003; Becker, Philipson and Soares, 2005).

Health as human capital

The idea of health as a form of human capital has a long history (for example, see Mushkin, 1962). Grossman (1972) develops a model in which illness prevents work so that the cost of ill health is lost labour time. However, there may also be an effect of ill health on worker productivity in employment. A major difficulty in measuring the economic effect of health is the two-way causality between wealth and health (Smith, 1999). Another difficulty is the lack of consensus on what is meant by health. Different studies use different health measures: self-assessments of health, biomarkers, medical records, limitations on physical functioning, and anthropometric measurements have all been used as health indicators. Each of these approaches may fail to provide a complete picture of an individual's health status, giving rise to a problem of measurement error. In addition, it is necessary to separate out the effect of investments in health from the effect of natural or genetic variation in health (Schultz, 2005). One solution to these problems in measuring the effect of health on worker productivity is to establish the causal paths in panel data through the use of timing of health shocks and income or wealth responses (for example, Adams et al., 2003). Case, Fertig and Paxson (2005), controlling for parental influences and education, find that childhood health has a significant impact on adult health and earnings. Yet another approach to establishing causality is to use instrumental variables. For example, Schultz (2002) instruments adult height with childhood health and nutrition to argue that each centimeter gain in height due to improved inputs as a child in Ghana and Brazil leads to a wage increase of between eight and ten per cent (Strauss and Thomas, 1998, provide a survey of studies in this area).

Thomas and Frankenberg (2002) caution against drawing inferences from observational studies and instead advocate an experimental approach. A randomized experiment using iron supplementation to reduce iron deficiency anemia led to sizeable effects on worker productivity in Indonesia (Basta, Soekirman and Scrimshaw, 1979). Quasi-experiments can be used where it is possible to treat changes to health as if such changes were randomly generated. Bleakley (2003) considers the effects of the eradication of hookworm and malaria in the United States in the 1910s and 1920s. These diseases were pandemic in many counties of the American South prior to eradication. Bleakley, controlling for normal wage gains in areas that were not infected, shows that children not exposed to these diseases due to their eradication had improved incomes as adults relative to those born before eradication.

This body of research on health and human capital generally supports the idea that health affects worker productivity. However, it lacks a good appreciation of which types of health intervention are most important and what rate of return can be achieved by investing in health as a form of human capital. In many developing countries, relatively inexpensive activities designed to prevent the spread of infectious disease (for example, vaccination) can increase population health at low cost, suggesting that even modest income gains from health will generate very high rates of return. By comparison, treating chronic non-infectious disease in developed countries is often costly. There is evidence that susceptibility to chronic disease in later life is determined by health and nutrition as a fetus and in infancy (Barker, 1992; Behrman and Rosenzweig, 2004), suggesting that early health investments are crucial for adult productivity.

Health and education

Education is widely agreed to affect economic outcomes, and health affects education through two mechanisms. The first is the effect of better child health on school attendance, cognitive ability, and learning. Bleakley (2003) finds that deworming of children in the American South had an effect on their educational achievements while in school. Miguel and Kremer (2004) find that deworming of children in Kenya increased school attendance.

The second mechanism is the effect of lower mortality and a longer prospective lifespan on increasing incentives to invest in human capital. This effect occurs for the individual for whom the benefits of education are now greater (Kalemli-Ozcan, Ryder and Weil, 2000). In addition, lower infant mortality may encourage parents to invest more resources in fewer children, leading to low fertility but high levels of human capital investment in each child (Kalemli-Ozcan, 2002). Evidence for this effect is limited, though Bils and Klenow (2000) do find an effect of life expectancy on investments in education at the national level.

Health and saving

Poor health affects both the ability to save and the impetus to save. Sickness can have a large effect on out-of-pocket medical expenses, which can reduce current and accumulated household savings. This occurs in developed countries (Smith, 1999) but is of particular concern in developing countries where families may be thrown into poverty if productive assets such as land or animals must be sold to pay for medical expenses.

Because poor health tends to be associated with a short lifespan, increasing population health and expected longevity will have an effect on the planning horizon and will influence life-cycle behaviour. With a fixed retirement age, a longer lifespan elicits greater savings for retirement. Blanchard (1985) considers the theoretical effect of a longer lifespan in a macroeconomic model. Hurd, McFadden and Gan (1998) find that increased expectation of longevity leads to greater wealth-holding at the household level in the United States. Bloom, Canning and Graham (2003) find an effect of life expectancy on national savings, using cross-country data. Lee, Mason and Miller (2000) argue that rising life expectancy can account for the boom in savings in Taiwan since the 1960s. But the effect of a longer lifespan need not be increased saving for retirement; people could instead choose to work longer. The behavioural response to longer lifespans depends on social security arrangements and retirement incentives (Bloom, Canning, Mansfield and Moore, 2007).

In a life-cycle model with a stable age structure and no population growth or economic growth, the dissaving of the old will exactly match the saving of the young at any level of life expectancy. This suggests that the aggregate effect of longer lifespans on savings is temporary and occurs when life expectancy rises. In the long run, the high savings rates of the working age population will be off set by the dissaving of a large cohort of elderly.

An effect on saving may lead to higher investment if capital markets are not perfectly open. In addition, a healthy population and workforce may increase productivity and encourage foreign direct investment (Alsan, Bloom and Canning, 2006).

Health and demography

Improvements in health and decreases in mortality rates can catalyse a transition from high to low rates of fertility and mortality – the ‘demographic transition’ (Lee, 2003). Population growth is the difference between birth and death rates (ignoring migration) and the global population explosion in the 20th century is attributable to improvements in health and falling death rates. In developing countries, health advances tend to lower infant and child mortality rates, leading initially to a surge in the number of children. Reduced infant mortality, increased numbers of surviving children, and rising wages for women can lower desired fertility (see Schultz, 1997) leading to smaller cohorts of children in future generations. Better access to family planning can also help couples achieve match more closely their fertility desires and realizations. This process creates a ‘baby boom’ generation that is larger than both preceding and succeeding cohorts. Subsequent health improvements tend primarily to affect the elderly, reducing old-age mortality and lengthening the lifespan.

In many theoretical models a population explosion reduces income per capita by putting pressure on scarce resources and by diluting the capital–labour ratio. In these models population declines spur economic growth in per capita terms. For example, the very high death rates, and decline in population, due to the Black Death in 14th century Europe appear to have caused a shortage of labour, leading to a rise in wages and the breakdown of the feudal labour system (Herlihy, 1997). However, in modern populations there appears to be little connection between overall population growth and economic growth; indeed the 20th century saw both a population explosion and substantial rises in income levels. Although it is difficult to find significant effects of overall population growth on economic growth, it is possible to consider the components of population growth separately. High birth and low death rates both generate population growth, but seem to have quite different effects on economic growth (Bloom and Freeman, 1988; Kelley and Schmidt, 1995). This may be because, while both forces increase population numbers, they affect the age structure quite differently. The effect of changing age structure due to a baby boom has large effects as the baby boomers enter the workforce and then as they eventually retire. While the baby boomers are of working age, economic growth may be spurred by a ‘demographic dividend’ if the baby boom generation can be productively employed. Bloom, Canning and Sevilla (2004) find that the demographic dividend increases the potential labour supply but its effect on economic growth depends on the policy environment.

There is a worry that health improvements and population aging will lead to high dependency rates and a slowdown in economic growth. In addition to longer lifespans, however, we are seeing a compression of morbidity; the period of sickness towards the end of life is falling as a proportion of overall lifespan (Fries, 1980; 2003). The idea that old-age dependency starts at 65 is essentially a result of social security retirement arrangements (Gruber and Wise, 1998) and healthy aging means that physical dependency now often occurs at much later ages.

Health and economic growth

In growth models, population health is usually taken to be life expectancy, or some other mortality measure, as opposed to the morbidity measured used at the individual level. This disjunction can be bridged by assuming a one-to-one relationship between mortality and morbidity rates in a population;

however it is not clear that such a relationship holds, making comparison of the macroeconomic relationship and microeconomic relationships difficult. In addition, calculating life expectancy requires age-specific mortality rates that are unavailable for many developing countries and published life-expectancy figures from the World Bank and United Nations are often constructed from quite incomplete raw data (Bos et al., 1992). There is a need to improve our measures of population health and to expand them to measures that correspond to morbidity and not just mortality.

The effect of health on individual productivity implies a relationship between population health and aggregate output. Shastry and Weil (2003) calibrate a production function model of aggregate output using microeconomic estimates of the return to health. They find that cross-country gaps in income levels can be explained in part by differential levels of physical capital, education, and health, with these three factors being roughly equal in terms of their contribution to differences in income levels. (A little over half of cross-country income gaps are explained by these factors; the remainder of the gap is ascribed to differences in total factor productivity.)

Another approach estimates the effect of population health on economic growth. Estimating the effect of the current level of population health on current income levels is subject to the problem of reverse causality; income also affects health. One way around this problem is to look at the effect of population health on subsequent economic growth, arguing that the timing can determine the direction of causality. This requires the absence of reverse causality through an expectation effect (so that current health is not caused by expected future economic growth).

Growth regressions show that the initial levels of population health are a significant predictor of future economic growth (Bloom, Canning and Sevilla, 2004, provide a survey of this literature). Sala-i-Martin, Doppelhofer and Miller (2004) find that the predictive power of health (as measured by life expectancy and malaria prevalence) is robust to the specification of the growth regression. Bhargava et al. (2001) argue that the effect of health on economic growth is larger in developing countries than in developed countries.

While population health measures are highly predictive of future economic growth, there is a debate about how to interpret the link. The health effect could be interpreted as the macroeconomic counterpart of the worker productivity effect found in individuals. However, Acemoglu, Johnson and Robinson (2003) argue that health differences are not large enough to account for much of the cross-country difference in incomes, and that the variations in political, economic and social institutions are more central factors. They argue that health does not have a direct effect on growth, but serves in growth regressions as a proxy for the pattern of European settlement, which was more successful in countries with a low burden of infectious disease.

Even if a causal interpretation of the effect of health on individual productivity and economic growth is accepted, the argument for using health as an input depends on there being low-cost health interventions that can increase population health without first having a high income level. There are, however, a large number of such interventions that can be implanted (Commission on Macroeconomics and Health, 2001).

Tropical disease and HIV/AIDS

Sub-Saharan Africa suffers from poor health due to the widespread presence of tropical disease. Malaria and tuberculosis cause high illness and death rates, while parasitic diseases such as schistosomiasis and

intestinal worms can cause anemia and reduced energy levels and productivity. In addition to these tropical diseases, the high prevalence of HIV/AIDS is causing life expectancy to decline dramatically in many countries in the region. Poor health status is one cause of sub-Saharan Africa's economic stagnation (Bloom and Sachs, 1998). Malaria appears to have an effect on economic growth over and above that created through higher mortality, suggesting that its effects on productivity with a given mortality burden are greater than other diseases (Gallup and Sachs, 2001).

Although HIV/AIDS has increased mortality rates dramatically, its impact on income per capita is unclear. HIV/AIDS is associated with high mortality but the period of sickness before death is relatively short. This mutes the worker productivity effects of the disease. Bloom and Mahal (1997) find that HIV/AIDS does not seem to lower the growth rate of income per capita; lower output is matched by lower population numbers due to high death rates. Young (2005) goes further and argues that AIDS mortality reduces fertility significantly, and that this will lower population pressure and increase the income per capita of the survivors of the pandemic in South Africa.

Many authors, however, argue that AIDS mortality has significant indirect effects that will reduce economic growth in the long term. Deaths from HIV/AIDS are concentrated among young adult men and women, leading to a higher dependency ratio. Bell, Devarajan, and Gersbach (2004) argue that the creation of a generation of AIDS orphans may lead to lack of care and education for children and to low productivity in the future. This effect may be compounded by fatalism induced by high AIDS mortality and shortened expected lifespan, which reduce the return to education. The high level of stigma associated with HIV/AIDS can reduce trust in the community, while high mortality and the strains imposed by extreme ill health before death can weaken families, community groups, firms, and government agencies, with long-term consequences for social capital (Haacker, 2004).

It is important to remember that income per capita is not a complete measure of welfare. Resources devoted to preventing and treating HIV/AIDS are part of measured income but reduce consumption of other goods, reducing welfare even as measured GDP per capita may remain steady (Canning, 2006). A more comprehensive welfare measure that included the welfare gain derived from a long lifespan, as well as annual income, would show a large welfare reduction due to HIV/AIDS (Crafts and Haacker, 2004). The main welfare effect of HIV/AIDS is the sickness and death of its victims and the impact of these on the victims’ families; the effect on the average income level of the survivors is decidedly secondary.

See Also

- child health and mortality
- demographic transition
- education in developing countries
- fertility in developing countries
- health outcomes (economic determinants)
- human capital, fertility and growth
- Malthusian economy
- population aging

- retirement
- returns to schooling
- value of life

Bibliography

Acemoglu, D., Johnson, S. and Robinson, J. 2003. Disease and development in historical perspective. *Journal of the European Economic Association, Papers and Proceedings* 1, 397–405.

Adams, P., Hurd, M.D., McFadden, D.L., Merrill, A. and Ribeiro, T. 2003. Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status. *Journal of Econometrics* 112, 3–56.

Alsan, M., Bloom, D.E. and Canning, D. 2006. The effect of population health on foreign direct investment inflows to low- and middle-income countries. *World Development* 34, 613–30.

Barker, D.J.P. 1992. *The Fetal and Infant Origins of Adult Disease*. London: BMJ Books.

Basta, S., Soekirman, K. and Scrimshaw, N. 1979. Iron deficiency anemia and productivity of adult males in Indonesia. *American Journal of Clinical Nutrition* 32, 916–25.

Becker, G.S., Philipson, T.J. and Soares, R.R. 2005. The quantity of life and the evolution of world inequality. *American Economic Review* 95, 277–91.

Behrman, J.R. and Rosenzweig, M.R. 2004. The returns to birthweight. *Review of Economics and Statistics* 86, 586–601.

Bell, C., Devarajan, S. and Gersbach, H. 2004. Thinking about the long-run economic costs of AIDS. In *The Macroeconomics of HIV/AIDS*, ed. M. Haacker. Washington, DC: International Monetary Fund.

Bhargava, A., Jamison, D., Lau, L. and Murray, C. 2001. Modeling the effects of health on economic growth. *Journal of Health Economics* 20, 423–40.

Bils, M. and Klenow, P.J. 2000. Does schooling cause growth? *American Economic Review* 90, 1160–83.

Blanchard, O.J. 1985. Debt, deficits, and finite horizons. *Journal of Political Economy* 93, 223–47.

Bleakley, H. 2003. Disease and development: evidence from the American south. *Journal of the European Economic Association* 1, 376–86.

Bloom, D.E. and Canning, D. 2000. The health and wealth of nations. *Science* 287, 1207–8.

Bloom, D.E., Canning, D. and Graham, B. 2003. Longevity and life-cycle savings. *Scandinavian Journal of Economics* 105, 319–38.

Bloom, D.E., Canning, D. and Sevilla, J. 2004. The effect of health on economic growth: a production function approach. *World Development* 32, 1–13.

Bloom, D.E. and Freeman, R.B. 1988. Economic development and the timing and components of population growth. *Journal of Policy Modeling* 10(1), 57–82.

Bloom, D.E. and Mahal, A.S. 1997. Does the AIDS epidemic threaten economic growth? *Journal of Econometrics* 77, 105–24.

Bloom, D.E. and Sachs, J. 1998. Geography, demography, and economic growth in Africa. *Brookings Papers on Economic Activity* 1998(2), 207–73.

Bloom, D.E., Canning, D., Mansfield, R., Moore, M. 2007. Demographic Change, Social Security Systems, and Savings. *Journal of Monetary Economics* 54, 92–114.

Bos, E., Vu, M.T., Stephens, P.W. and Patience, W. 1992. Sources of World Bank estimates of current mortality rates. Policy Research Working Paper Series 851. Washington, DC: World Bank.

Canning, D. 2006. The Economics of HIV/AIDS in Low-Income Countries: The Case for Prevention. *Journal of Economic Perspectives* 20, 121–142.

Case, A., Fertig, A. and Paxson, C. 2005. The lasting impact of childhood health and circumstance. *Journal of Health Economics* 24, 365–89.

Commission on Macroeconomics and Health. 2001. *Macroeconomics and Health: Investing in Health for Economic Development*. Geneva: World Health Organization.

Crafts, N. and Haacker, M. 2004. Welfare implications of HIV/AIDS. In *The Macroeconomics of HIV/AIDS*, ed. M. Haacker. Washington, DC: International Monetary Fund.

Cutler, D.M., Deaton, A.S. and Lleras-Muney, A. 2006. The determinants of mortality. *Journal of Economic Perspectives* 20(3), 71–96.

Cutler, D.M. and McClellan, M. 2001. Productivity change in health care. *American Economic Review* 91, 281–86.

Cutler, D.M. and Miller, G. 2005. The role of public health improvements in health advances: the

twentieth-century United States. *Demography* 42, 1–22.

Deaton, A. 2006. The great escape: a review essay on Fogel's The Escape from Hunger and Premature Death, 1700–2100. *Journal of Economic Literature* 44, 106–14.

Easterly, W. 1999. Life during growth. *Journal of Economic Growth* 4, 239–76.

Fogel, R.W. 2004. *The Escape from Hunger and Premature Death, 1700–2100: Europe, America, and the Third World*. Cambridge: Cambridge University Press.

Fries, J.F. 1980. Aging, natural death, and the compression of morbidity. *New England Journal of Medicine* 303, 130–5.

Fries, J.F. 2003. Measuring and monitoring success in compressing morbidity. *Annals of Internal Medicine* 139, 455–9.

Gallup, J.L. and Sachs, J.D. 2001. The economic burden of malaria. *American Journal of Tropical Medicine and Hygiene* 64(1, 2 Supplement), 85–96.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80, 223–55.

Gruber, J. and Wise, D. 1998. Social security and retirement: an international comparison. *American Economic Review* 88(2), 158–63.

Haacker, M. 2004. HIV/AIDS: the impact on the social fabric and the economy. In *The Macroeconomics of HIV/AIDS*, ed. M. Haacker. Washington, DC: International Monetary Fund.

Herlihy, D. 1997. *The Black Death and the Transformation of the West*. Cambridge, MA: Harvard University Press.

Hurd, M., McFadden, D. and Gan, L. 1998. Subjective survival curves and life-cycle behavior. In *Inquiries in the Economics of Aging*, ed. D. Wise. Chicago: University of Chicago Press.

Kalemli-Ozcan, S. 2002. Does mortality decline promote economic growth? *Journal of Economic Growth* 7, 411–39.

Kalemli-Ozcan, S., Ryder, H.E. and Weil, D.N. 2000. Mortality decline, human capital investment, and economic growth. *Journal of Development Economics* 62, 1–23.

Kelley, A.C. and Schmidt, R.M. 1995. Aggregate population and economic growth correlations: the role

of the components of demographic change. *Demography* 32, 543–55.

Lee, R. 2003. The demographic transition: three centuries of fundamental change. *Journal of Economic Perspectives* 17(4), 167–90.

Lee, R., Mason, A. and Miller, T. 2000. Life cycle saving and the demographic transition: the case of Taiwan. *Population and Development Review* 26(Supplement), 194–219.

Miguel, E. and Kremer, M. 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72, 159–217.

Mushkin, S.J. 1962. Health as an investment. *Journal of Political Economy* 70(5), 129–57.

Nordhaus, W. 2003. The health of nations: the contribution of improved health to living standards. In *Measuring the Gains from Medical Research: An Economic Approach*, ed. K.H. Murphy and R.H. Topel. Chicago: University of Chicago Press.

Preston, S. 1975. The changing relation between mortality and level of economic development. *Population Studies* 29, 231–48.

Pritchett, L. and Summers, L. 1996. Wealthier is healthier. *Journal of Human Resources* 31, 841–68.

Sala-i-Martin, X., Doppelhofer, G. and Miller, R.I. 2004. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94, 813–35.

Schultz, T.P. 1997. The demand for children in low income countries. In *Handbook of Population and Family Economics*, vol. 1A, ed. M.R. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Schultz, T.P. 2002. Wage gains associated with height as a form of human capital. *American Economic Review, Papers and Proceedings* 92, 349–53.

Schultz, T.P. 2005. Productive benefits of health: evidence from low income countries. In *Health and Economic Growth: Findings and Policy Implications*, ed. G. Lopez-Casasnovas, B. Riveras and L. Currais. Cambridge, MA: MIT Press.

Shastry, G.K. and Weil, D.N. 2003. How much of cross-country income variation is explained by health? *Journal of the European Economic Association* 1, 387–96.

Smith, J.P. 1999. Healthy bodies and thick wallets: the dual relation between health and economic status. *Journal of Economic Perspectives* 13(2), 145–66.

Strauss, J. and Thomas, D. 1998. Health, nutrition and economic development. *Journal of Economic Literature* 36, 766–817.

Thomas, D. and Frankenberg, E. 2002. Health, nutrition and prosperity: a microeconomic perspective. *Bulletin of the World Health Organization* 80(2), 106–13.

Viscusi, W.K. and Aldy, J.E. 2003. The value of a statistical life: a critical review of market estimates from around the world. *Journal of Risk and Uncertainty* 27, 5–76.

Young, A. 2005. The gift of the dying: the tragedy of AIDS and the welfare of future African generations. *Quarterly Journal of Economics* 120, 243–66.

How to cite this article

Canning, David and David E. Bloom. "population health, economic implications of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000259>
doi:10.1057/9780230226203.1310

The New Palgrave Dictionary of Economics Online

Portugal, economics in

António Almodovar and José Luís Cardoso

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The development of economics in Portugal has been marked by intellectual curiosity coupled with pragmatism. Both characteristics are explained by the long-standing feeling that, although Portugal was lagging in terms of social and economic development, the situation could be overcome by means of an appropriate economic policy. This feeling motivated a continuing effort to find answers to economic and financial problems by careful analysis of other countries' experiences – both the principles discussed by economists and the policies eventually implemented by governments. Portuguese experience thus well illustrates the international diffusion of the ideas associated with different schools of economic thought.

Keywords

Austrian School; corporatism; free trade; Harrod–Domar growth model; Keynes, J. M.; Lausanne School; List, F.; Manchester School; marginal revolution; mercantilism; neoclassical synthesis; ophelimity; Physiocracy; Portugal, economics in; protection; Quesnay, F.; Say, J.-B.; Smith, A.; utility

Article

Mercantilism

The first interesting examples of a concern to establish first principles to explain economic reality emerged in Portugal in the 16th century. Extending the spirit of the Discoveries by the early Portuguese explorers to the scantily studied areas of economic knowledge, Portuguese authors showed a certain pioneering spirit. In contrast with former prejudices about the harmfulness of trade, commerce began to be considered as the principal cause of the wealth: commerce dynamically connected the different sectors of economic activity and brought individuals and communities together.
Portuguese economic literature of the second half of the 16th century exhibits innovative analyses with regard to: (*a*) an abstract conceptualization of the market as a space wherein to promote individual and

public interests, and as a mechanism to reveal the value of goods exchanged; (*b*) a comparison of the advantages and disadvantages of a monopolistic organization of trading circuits; (*c*) the link between the real and the monetary spheres of the economy and an early version of a quantity theory of money; and (*d*) the doctrinal legitimization of individual gains arising from mercantile activity, for example in the case of exchange and insurance contracts. Handling such different subjects required new ways of thinking. However, in this literature, produced by merchants, theologians and court counsellors, there was no change in the theological and ethical foundations or the method on which the new elements were based. Despite the adaptations needed to interpret the new realities presented by the Discoveries, Portuguese thought continued to rest on moral and religious ideas.

During the period of the dynastic union that for 60 years (1580–1640) kept Portugal under the direct control of the Spanish crown, economic ideas started to be based on the standards of the so-called bullionist literature. However, the political restoration movement that started in 1640 initiated a search for economic strategies for consolidating independence and political sovereignty.

The tactics recommended were varied. Some favoured aiming for either balanced trade or a surplus, others favoured the introduction of monetary regulation (considering an intuitive approach to the relationship between money flows and prices), while yet others favoured the growth of the population to ensure an increase in output and tax revenues. Portuguese authors managed to receive and disseminate, almost simultaneously, different types of foreign contemporary economic literature. They adapted and used analytical constructs and economic policy proposals provided not just by Spanish but also by Italian mercantilists (particularly their proposals regarding population as a means to increase wealth), the English (the balance of trade doctrine and proposals to set up regulated companies for foreign commerce) and the French (manufacturing policy). The policy of protection for manufacturing proposed by Duarte Ribeiro de Macedo (1675), adopted at the close of the 17th century, clearly illustrates this process of assimilating ideas and economic guidelines into a national economic development strategy. The new commercial framework imposed by the Methuen Treaty with England in 1703 did not silence supporters of this strategy. Protectionism continued to attract support, and contributed to the shaping of an entrenched tradition. During the government of Marquis of Pombal (1750–77), a protectionist economic policy was extensively applied, especially through the establishment of monopolistic companies in both commercial and productive economic activities.

Enlightened political economy

From the late 18th century, the development of political economy reflects the wave of economic, social, cultural and political transformations taking place throughout Europe, known as the Enlightenment. During this period, the discourse of Portuguese economists – particularly that represented by the publications of the Royal Academy of Sciences of Lisbon (Cardoso, 1990–91) – reveals some familiarity with Physiocratic doctrines and principles. The primary aim of these discourses, which helped create a climate receptive to laissez-faire ideology, was the abolition of the internal barriers and excessive regulations of the *ancien régime*, which were considered as obstacles to the smooth working of the domestic market.

The dissemination of Smithian political economy was furthered by the same concerns, particularly after 1803. The ideas of both Smith and the French Physiocrats were valued as possible guidelines for a successful state-led process of social and economic change; for this reason, the reading that was made of

Smith's works in Portugal by J.J. Rodrigues de Brito (1803–1805) and José da Silva Lisboa (1804) focused mainly upon the feasibility of the systems of political economy that were encouraged by François Quesnay and by the *Wealth of Nations*. Aided by Jean-Baptiste Say, Adam Smith was rapidly acknowledged as the true founder of modern political economy.

Notwithstanding the unanimous acknowledgement of the importance of Smith's economic thought in the subsequent spread of classical political economy, which led to its eventual institutionalization as a separate area of study, the English classical school did not have a significant effect on Portuguese economics. As English economic success was undeniable, and as England continued to be Portugal's principal commercial, political and military ally, the lack of a marked preference for English economics may, at first sight, seem strange. However, several factors explain it. Portugal's problematic political and diplomatic circumstances, after the first signs of the Brazilian desire for independence (1814), made it clear that England's support for free trade could be harmful. After this date, a compromise was gradually established regarding the appropriateness of the principles supported by English political economists: Acúrsio das Neves (1814–1817) made a clear distinction between the virtues of domestic liberalization and the need for prudence at the international level. On the other hand, the need to simplify and popularize political economy was seemingly better fulfilled through Continental political economy. Although some French, German and Italian works might be less vigorous analytically, they provided both an explanation and a critical assessment of many of the English school's doctrines.

Given that Continental political economists were as concerned as the Portuguese with the consequences of the English system, it was only natural that they eventually had a greater impact in Portugal than the more specific, abstract ideas of Ricardo and his followers. This is particularly evident in the first discussions regarding the choice of a political economy handbook, for no one suggested the use of an English author. The first attempts to write a Portuguese text were based either on Say's work (Manuel de Almeida) or on Storch's *Treatise* (José Ferreira Borges). In the 1840s, Forjaz de Sampaio (1841) was to write a handbook inspired by the approach developed by Karl Friderich Rau, while Marnoco e Sousa (1910), the most celebrated early 20th-century professor at the University of Coimbra, came under the spell of E.R.A. Seligman.

Establishing a canon

The overthrow of the *ancien régime* in 1820 in a liberal revolution did not change the previous misgivings about some aspects of classical political economy, particularly those regarding international free trade.

Concern with national economic development, coupled with a suspicion that English political economy was biased in favour of English interests, led to the prevalence of an approach that was quite similar to the one that was later to be developed in Friedrich List's national system of political economy. Between 1820 and 1850, a significant number of Portuguese authors insisted that several principles of classical economic theory and policy were abstruse and therefore inappropriate for steering the development of their own country. As a result, not just the doctrine of free trade but also the theories of population, rent, diminishing returns and the stationary state were dismissed either as wrong or as being solely applicable to the more advanced English circumstances.

When, in 1836, political economy was eventually accepted as an academic discipline, and situated

within legal studies, the critical stance regarding the selection of both authors and doctrines to be taught was reinforced. On the one hand, since political economy was mainly taught for the benefit of lawyers, teachers were naturally expected to emphasize the relations between economic laws and legislative action. On the other hand, since they were scholars and not pamphleteers, teachers were meant to adopt an unbiased approach to political economy, not supporting any single school of economic thought. The ensuing eclecticism had the beneficial consequence of allowing for a continuous updating process, teachers such as Adrião Forjaz de Sampaio (between the late 1830s and the early 1870s) or Marnoco e Sousa (between 1900 and 1916) always being ready to mention each and every new school of economic thought that came to their knowledge. This same concern eventually led to significant space being allotted in Portuguese law schools to the study of the history of economic thought.

When coupled with a constant awareness of the national conditions that could make some of the doctrines of political economy unworkable, and a mistrust regarding excessively abstract formulae, this eclecticism also helps us to explain why marginalism and neoclassical economics had less impact than the sociological and historical schools of the second half of 19th century.

The rejection of mainstream economics

At the start of the 20th century, investment in the development of theoretical abstractions was still generally deemed by Portuguese authors not to be an essential part of their role as economists, for they thought that they should concentrate on the task of identifying and solving present economic and social problems. Therefore, even if the conceptual advances made by the marginalist revolution and the theoretical apparatus developed by neoclassical economists both in Europe and in the United States of America did have repercussions, these did not lead to any effort towards furthering economic analysis per se. Jevons was partly translated, and the doctrines of the Manchester, Austrian and Lausanne schools were closely summarized and scrutinized (either approvingly or disapprovingly). But on most occasions, these foundations of the modern canon of economic analysis were laid in Portugal in an eclectic manner, devoid of any noticeable tendency to claim that there were undisputable economic principles (see Almodovar and Cardoso, 2001).

In the rare cases where this attitude did not prevail, the reaction was quite vigorous. António Horta Osório (1911) aimed to establish the importance of the mathematical method in political economy in the context of the development of the general equilibrium theory of the Lausanne School.

The Paretian distinction between utility and ophelimity was used by Osório as the foundation of his view of economic science: economics was defined as a disciplinary field restricted to the study of a small part of human behaviour, while psychology was portrayed as the global science pertaining to the study of human action. Consequently, economics would be no more than one of the branches of the general study of mankind. For him, pure economics was an abstract and experimental science which had to evaluate its scientific character, like all the exact sciences, not through the practical utility of its conclusions but chiefly by establishing the exactness of its formal internal logic.

When Osório was writing, the comprehensibility of this methodological attitude was problematic, even for those learned in economic matters, not to mention the fact that general equilibrium in exchange was far from being acknowledged by mainstream economics – something that happened only in the 1930s with the neoclassical synthesis.

In an environment that was clearly unsympathetic both to pure theory, and to any claims for the

supremacy of one school, Osório was condemned to oblivion and his book was dismissed. Such an outcome symbolized the reiteration by early 20th century Portuguese economists of traditional views regarding the usefulness and role of political economy. At a time when economics was moving decisively towards establishing itself as a science, Portuguese authors stood aloof from that process. At a time when Portugal's political situation was characterized by a fair degree of cosmopolitanism, economic discourse remained focused on its own political implications, musing on topics of an ideological nature.

From corporatism to Keynesianism

The traditional Portuguese unwillingness to abide by any single school of economic thought faded away only when faced with the ambitious state-driven project of building up a new type of political economy, that of the corporatist state (see Almodovar and Cardoso, 2005).

The political and economic experiment of corporatism represents one of the most interesting stages in the study of the historical evolution of economic thought in Portugal. The corporatist movement in itself is part of the broader movement of authoritarian experiments that took place on the European and international political scene, especially after the First World War. The restlessness of many social and political sectors, which were discontented with the performance of liberal and socialist regimes, paved the way for a search for an alternative to both capitalism and socialism. Corporatism was therefore offered as a third way between existing regimes, and its supporters claimed that it provided the sole reliable answer to the ongoing social, political and economic turmoil. Portuguese authors like Pires Cardoso and Marcelo Caetano joined their French and Italian counterparts in the effort to develop an economic point of view that could match the ethical and philosophical base provided by the philosophy of Thomas Aquinas. The final outcome was an economic doctrine openly against utilitarianism and in favour of the gradual establishment of a new type of economic agent – the so-called *homo corporativus* – that would be capable of re-embedding social values and aims into its own scale of preferences. This quest did not stop the reception of Keynes, and particularly those ideas that could be seen as a critique of the idea of a self-regulated market and as an appeal to some state intervention in favour of a more socialized economy. Keynes was fairly well-known in Portugal from the mid-1920s onwards. However, the reception and assimilation of the *General Theory* occurred more than ten years after its publication, when Fernando Pinto Loureiro (1948) and Luis Simões de Abreu (1949) published the first extended reviews and digests of J.M. Keynes's major work. Therefore, it was only after 1950, under the impulse provided by the works of António Pinto Barbosa (1950), Jacinto Nunes (1956) and F. Pereira de Moura (1964), that Keynesian concepts began to play a significant role in newly established courses on macroeconomics, public finance, development economics and econometrics.

Despite this process, the reception given to the propositions of the *General Theory* at the political level was ambiguous and much less enthusiastic. One of the most notable aspects of the economic strategy followed in Portugal throughout the 1930s and 1940s was the enactment of legislation to set up industrial companies and to control industrial activity. The basic aim was not just to prevent Portuguese industry from being disturbed by internal or external competition but to keep in check and organize industrial growth and the overall process of economic development. In such a context, where the pace of development was restricted and a balanced budget and the preservation of the country's gold reserves

were praised, Keynesian economics could hardly figure prominently.

This vision of economic life was shared by the various political assemblies and executive bodies responsible for directing economic and financial policy. As a result, whenever they incorporated Keynesian ideas, they did so in a watered-down and superficial manner. In fact, it can be said that in the post-war period no type of short-run macroeconomic policy was ever developed: the factors which would normally justify it – such as unemployment and external disequilibrium – did not represent real problems for the Portuguese economy. Something similar occurred in relation to long-run economic policy. The first five-year development plan (1953) was totally insensitive to the assessment of its impact in macroeconomic terms. The second plan in 1958 contained, in the explanation given for its design, projections based on a Harrod–Domar growth model, but this was no more than a rhetorical device.

Throughout the period that we have been considering here, Portuguese economic policy remained faithful to corporatist principles, coupled with a traditional model of empirical, and essentially descriptive, economic studies, without any visible influence of Keynesian concepts.

Concluding remarks

The first Portuguese university institution created specifically for the teaching of economics was formally founded in 1933. A profound reform of its curricular and pedagogical structure took place in 1949, involving the replacement of essentially technical courses in the fields of commerce, bookkeeping, accounting, customs and diplomatic services with more general courses in economics and finance. Only after that did the full incorporation of a neoclassical approach begin to take place, in the form of a synthesis with Keynesian thought. However, integration into the international mainstream was held back by the resilience of the traditional Portuguese attitude regarding economic knowledge, which was to try to take over the doctrinal and political ingredients that best fitted the search for a specifically Portuguese route to economic development. At all of the most significant moments in the evolution of economic thought in Portugal, we find this attempt to select and adapt existing economic ideas to Portuguese circumstances. Inquisitiveness regarding alternative routes to economic progress, coupled with a pragmatic view regarding economic policy guidelines, favoured a continuous oscillation between schools of economic thought and the emergence of eclecticism. As a consequence, Portuguese economic thinking retained its links with law, ethics, politics, and sociology; and it took a long time to accept the autonomy and analytical competence of economics (see Almodovar and Cardoso, 1998).

A further example of the tendency to eclecticism prevalent among Portuguese economists was the impact of the structuralist and developmentalist economic ideas in the 1960s and in the 1970s, through the influence of the Latin American economists concerned with the problems of underdevelopment and with the political responses to overcome it. However, this influence was superseded by the process of harmonization resulting from the democratic revolution of 1974 and Portugal's integration into the European economy in 1986. The considerable institutional changes that then occurred have largely contributed to the smooth reception and institutionalization of both macroeconomic and microeconomic principles and applications. As a result of this integration process, eclecticism has gradually given way to approaches that conform much more closely with the international mainstream.

See Also

- France, economics in (before 1870)
- France, economics in (after 1870)
- heterodox economics
- Historical School, German
- Italy, economics in
- Keynesianism
- mercantilism
- physiocracy
- Spain, economics in

Bibliography

Abreu, L.S. 1949. Algumas notas sobre a economia de Keynes [A few notes on the economics of Keynes]. *Revista de Economia* 1(3), 15–25.

Almodovar, A. and Cardoso, J.L. 1998. *A History of Portuguese Economic Thought*. London and New York: Routledge.

Almodovar, A. and Cardoso, J.L. 2001. From learned societies to professional associations: the establishment of the economist profession in Portugal. In *The Spread of Political Economy and the Professionalisation of Economists*, ed. M. Augello and M. Guido. London and New York: Routledge.

Almodovar, A. and Cardoso, J.L. 2005. Corporatism and the economic role of government. In *The Role of Government in the History of Economic Thought*, ed. S. Medema and P. Boettke. Annual Supplement to *History of Political Economy*, vol. 37. Durham, NC, and London: Duke University Press.

Barbosa, A.M.P. 1950. *Economia*. Lisboa: ISCEF.

Brito, J.J.R. 1803–1805. *Memórias Políticas sobre as Verdadeiras Bases da Grandeza das Nações e Principalmente de Portugal* [Political Memoirs on the True Basis of the Greatness of Nations, Especially of Portugal], ed. J.E. Pereira, Lisboa: Banco de Portugal, 1992.

Cardoso, J.L., ed. 1990–91. *Memórias Económicas da Real Academia das Ciências de Lisboa, para o Adiantamento da Agricultura, das Artes, e da Indústria em Portugal, e suas Conquistas, 1789–1815* [Economic Transactions of the Lisbon Academy of Sciences for the Advancement of Agriculture, Industry and Commerce in Portugal and its Domains], 5 vols. Lisboa: Banco de Portugal.

Lisboa, J.S. 1804. Princípios de Economia Política. In *Escritos Económicos Escolhidos (1804–1820)*

[Selected Economic Writings], vol. 1, ed. A. Almodovar, Lisboa: Banco de Portugal, 1993.

Loureiro, F.P. 1948. *Sobre a Introdução ao Estudo da Nova Economia Keynesiana* [Introduction to the Study of the New Keynesian Economics]. Coimbra: Coimbra Editora.

Macedo, D.R. 1675. Discurso sobre a Introdução das Artes no Reino [Discourse on the Introduction of Manufactures in the Kingdom]. In *Antologia dos Economistas Portugueses – Século XVII* [Anthology of Portuguese Economists – 17th Century], ed. A. Sérgio. Lisboa: Biblioteca Nacional, 1924.

Moura, F.P. 1964. *Lições de Economia* [Lectures in Economics]. Lisboa: Clássica Editora.

Neves, J.A. 1814–1817. *Variedades sobre Objectos Relativos às Artes, Comércio e Manufacturas, Consideradas Segundo os Princípios da Economia Política* [Varieties on Subjects Concerning Industry and Commerce, Considered Under the Principles of Political Economy], ed. A. Almodovar and A. Castro, Porto: Afrontamento, 1994.

Nunes, M.J. 1956. *Rendimento Nacional e Equilíbrio Orçamental* [National Income and Budgetary Equilibrium]. Lisboa: Editorial Império.

Osório, A.H. 1911. *A Matemática na Economia Pura: a Troca* [Mathematics in Pure Economics: Exchange], ed. M. Farto. Lisboa: Banco Portugal, 1996.

Sampaio, A.F. 1841. Elementos de Economia Política e Estadística [Elements of Political Economy and Statistics]. In *Estudos e Elementos de Economia Política (1839–1874)* [Studies and Elements of Political Economy], vol. 2, ed. A. Pedrosa. Lisboa: Banco de Portugal, 1995.

Sousa, J.M. 1910. *Ciência Económica* [Economics], ed. M. de Fátima Brandão, Lisboa: Banco de Portugal, 1997.

How to cite this article

Almodovar, António and José Luís Cardoso. "Portugal, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000361> doi:10.1057/9780230226203.1311

The New Palgrave Dictionary of Economics Online

positive economics

Richard G. Lipsey

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Positive economics’ refers to the view that economic theories consistent with all conceivable observations are empirically empty and that empirically useful theories need to be consistent with existing observations (thus passing the ‘sunrise test’) and predict something new. It is neither logical positivist, nor operationalist, nor naïve falsificationist; nor is it based on strict dichotomies between positive and normative statements and between positive analysis and normative advice. It rejects the views that theories can assist understanding the world without making refutable statements about it; that theories can be criticized only on their own terms; and that all distinctions inhibit useful discourse.

Keywords

Agassi, J.; Archibald, G.; assumptions; barriers to entry; Bekar, C.; Blaug, M.; Caldwell, B.; Carlaw, K.; de Marchi, N.; economic history; falsificationism; Friedman, M.; game theory; Hahn, F.; Hodgson, G.; Hutchison, T.; industrial organization; Klappholz, K.; labour economics; Lind, H.; Lipsey, R. G.; logical positivism; methodology; monopolistic competition; operationalism; policy advice; Popper, K.; positive economics; positive–normative distinction; Robbins, L.; Samuelson, P.; scale effects; sunrise test; testing; theory appraisal; value judgements; verificationism; Wong, S.

Article

The term ‘positive economics’ refers to some specific views about what makes economics a science. According to some of the most influential 19th century English economists, positive economics consisted of propositions or ‘laws’ concerning real-world events that were derived from intuitively self-evident assumptions. Facts were to be used, therefore, as illustrations of theories, not as tests. To give policy advice, the propositions of positive economics had to be combined with value judgements. An elegant 20th century statement of this view of ‘scientific economics’ was given by Robbins (1935). Not surprisingly, these economists were, as Blaug (1992) has argued, ‘verificationists’ who shielded their

theories from empirical refutation.
The term was used by such 20th century writers as Friedman (1953) and Lipsey (1963) to refer to what they regarded as scientific economics: non-normative theories whose assumptions were not necessarily self-evident and whose implications were to be judged by empirical observations. Karl Popper provided the methodological underpinning of these works, underpinnings that were either implicit (as in Friedman's case) or explicit (as with most other writers). Terence Hutchison (1938) was the first to introduce Popper's ideas to economists, although he did not describe his work as positive economics. Blaug (1992) and Hutchison (1992) provide excellent formulations of the main tenets of modern positive economics, along with criticisms of both its main detractors and advocates of other views.
Friedman (1953, pp. 7–8) stated the sense in which he understood the term ‘positive’ when he wrote: ‘The ultimate goal of positive science is the development of a “theory” or “hypothesis” that yields valid and meaningful (i.e., not truistic) predictions about phenomena not yet observed. … only factual evidence can show whether it is …tentatively “accepted” as valid or “rejected”•’. Shortly after my textbook appeared, I wrote:

> I tried to break away from the treatment of economic theory as revealed truth and to emphasize from the outset the very tentative nature of much of our economics…to say … to the student ‘you cannot have both certainty and empirical relevance’… . The adjective ‘positive’ in the title of the book was [partly an allusion to the positive–normative distinction and] partly an allusion to the distinction between positive (i.e., empirical) Versus a priori methods of judging between theories. (Lipsey, 1964, pp. 370–1)

To this end, the concluding chapters in several parts of my textbook discussed ‘measurement’, ‘tests’, and ‘criticisms’ of the theories already presented.

Some criticisms and misunderstandings

The history of positive economics at the London School of Economics’ Staff Seminar on Methodology, Measurement, and Testing (M^2T seminar) has been well described by de Marchi (1988) – although I disagree with most of his conclusions on pages 162–3. Ours was a crusade for making economic theories empirically relevant and for rejecting intuition as the test of validity, replacing it with empirical testing. Positive economics, as we conceived it, had two main messages. First, if an economic theory is to be about the real world, it must be possible to imagine observations that would be conflict with it. If conflicting observations cannot even be imagined, the theory is compatible with all states of the world and hence empirically empty. A great advance in making theory more relevant would be achieved if today's editors insisted that each author state what factual observations would conflict with his or her theory, and, if there were none, to state the theory's purpose. Second, a new theory should be compatible with (‘explain’) some existing facts and suggest some new one(s).
We were subject to misunderstandings, as well as to criticisms, from those who disagreed with our main message. Here are some of the most influential.

1. 1. Its philosophical base was thought by many critics to be logical positivism, which we rejected

in favour of Popper's methodology.

2. 2. Samuelson (1948) was partly responsible for another confusion when he stated a similar view on testability but espoused a version of operationalism. We never took an operationalist position, arguing only that *all* those parts of a theory that did say something empirically should be open to empirical testing. Subsequently, Wong (1978) argued – correctly in our view – that, since even such simple ‘entities’ as prices and quantities are theoretical concepts that do not exactly correspond to real-world entities, theories cannot be expressed solely in operational terms.
3. 3. Friedman set off a long debate on the testability of assumptions that helped to discredit positive economics to many. We disagreed with Friedman, arguing that, if empirically correct predictions were deduced from a set of empirically false assumptions, this called for further serious study, not complacency. See Blaug (1992, pp. 91–7) for a full discussion.
4. 4. Contrary to the tenets of positive economics, Friedman used his essay on methodology to dismiss monopolistic competition as adding nothing that could not be learned from a judicious combination of perfect competition and monopoly. That this was not our position was shown when this, and similar arguments of others in the Chicago school, were criticized by members of the M^2T seminar (Klappholz and Agassi, 1959; Archibald, 1961).
5. 5. We were accused of being naive falsificationists. Although we may have been at the outset, we soon refined our position as a result of experience and accepted that ‘we cannot get a categorical disproof of an hypothesis’ (Lipsey, 1975, p. 46), which statement was followed by a long passage on what can be learned from apparent refutations. Another member of the M^2T seminar, Archibald (1967), argued that empirical testing could at best establish the balance of probabilities between two conflicting theories rather than refuting either categorically. We did not, however, as implied by de Marchi (1988, p. 162) give up on positive economics just because we abandoned naive falsification; indeed, many of us went on to do significant empirical work.
6. 6. Some critics argued that positive economics was merely what Mark Blaug calls ‘conformatism’, asking only that a theory be consistent with known facts. From the very outset we accepted Popper's criticism that theories that explained only already known facts were being subjected to a ‘sunrise test’ from which we could only learn that the theorist was ingenious enough to build a theory that jumped through predetermined hoops.
7. 7. Others argued that we naively accepted the earlier economists’ strict dichotomy between positive and normative statements. We quickly discarded this view. Lipsey (1963, p. 4, n. 1) introduces a discussion of this matter thus: ‘Philosopher friends have persuaded me that when pushed to its limits, the distinction between positive and normative becomes blurred or breaks down completely.’ However, the blurring did not stop us from arguing that the ability to distinguish what one thinks is true from what one would like to be true is critical to all science.
8. 8. In a similar but not identical vein, yet others argued that we naively accepted the strict division that the earlier economists had made between positive economic analysis and normative advice. My first exposure to policy advising in 1962 disabused me of that idea. As I later put it: ‘The economic adviser and the policy-maker are involved in a complex human relationship, entangled in various uncertainties and communicating with each other through an inevitable haze of emotional reactions. Economists may strive towards an ideal of communicating their knowledge as objectively as possible, but objectivity remains an ideal that guides their actions, not a reality

that fully describes them' (Lipsey, 1981, p. 35). For detailed discussions of the history of the distinction and its defence on pragmatic grounds, see Blaug (1998, pp. 370–4).

We rejected many other methodological approaches that were either implicit or explicit criticisms of positive economics, of which the following are examples. First was the view of many pure theorists, such as Hahn (1984, pp. 44–5), criticized by Blaug (1992, pp. 164–5) and Hutchison (1992, p. 43), that theories can somehow add to our understanding of the world without making testable statements about it. Second, there was the view subsequently articulated by Caldwell (1982) that falsification is too strong and that we can criticize each school of thought only on its own terms. As Hutchison pointed out, this amounts to rejecting in principle any method of discriminating between alternative theories. Third, in reaction to the view that all distinctions inhibit full discourse, we maintained that distinctions help to structure arguments and, without them, there is anarchy of discourse.

Positive economics today

What is the fate of positive economics today? While many economists pay lip service to the view that economic theories should make testable predictions about the world and that the ultimate arbiter of different theories is empirical evidence, many research programmes do not show this as their revealed preference. Theoretical articles that do no more than state and pass sunrise tests abound.

The modern version of industrial organization has had most of its empirical content eliminated. Students who used to learn institutional material about such 'practical' matters as competition policy, and who studied empirical information about scale effects and entry barriers, often today know little more than game theory. (See Lipsey, 2001, for full discussion.)

The new formalism has given rise to the belief among some theorists (although not typically among those who do game theory) that the more general a theory is, the better it must be. But this assumption ignores the fact that the more general a theory is, the less empirical content it is likely to have since, by ignoring the specific context in which many problems arise, it becomes impossible to analyse them in depth. (See Hodgson, 2001, for a full discussion.) One set of examples, criticized at length in Lipsey, Carlaw and Bekar (2005, pp. 466–7), is found in those modern growth theories that use an aggregate production function devoid of institutions or anything that distinguishes economies with various degrees of development.

Not a few theories are devoted to explaining mere possibilities. Typically someone develops a simple 'Mark I model' on some matter such as the effects of rent controls and draws strong policy conclusions from it; someone else comes along with a Mark II model saying 'if I add some not implausible complexity, the model's predictions and policy conclusions are altered'. Then someone does the same to the Mark II model, and so on. (For a case study see Lind, 2007.) Although it is possible to learn something from all exercises, this sort of research programme tells us little more than that very simple and more complex theories on the same issue do not usually have identical predictions.

Many research programmes are 'internally driven', by which I mean that they are driven by their own internal logic. Investigators seek to understand problems created by the models that they are using rather than deriving their problems from observations. In contrast, an 'externally driven research program' (EDRP) is one that is driven and constrained by observed facts. A perusal of the literature will

show IDRPs to be at least as common as EDRPs. (For examples, see Lipsey, 2001.)

On a personal level, the revolution that I tried to create in textbook writing through *An Introduction to Positive Economics* has slowly dwindled – in spite of its being the dominant text book in the UK for decades, being widely used throughout the Commonwealth, and having significant sales in its US adaptation, initially co-authored with Peter Steiner. As a result of constant criticism from teachers who wanted to present only mainline economics, the criticism and testing chapters were slowly eroded – much faster in the US editions than in the UK ones. Today, all too many modern theory textbooks at all levels, from basic to advanced, present current economic theories as if they were revealed truth, paying little attention to controversies and alternative theories.

Finally, I ask what the real successes of positive economics are. As already mentioned, most economists pay a least lip service to the ideal that economic theories are meant to tell us something about the real world through potentially testable hypotheses. The journals are full of empirical observations, many of which are extremely useful – although many others are used in the non-informative types of theory mentioned earlier. In some empirically oriented fields, such as economic history and labour economics, the ideal of positive economics does come close to realization. For example, much work in labour economics seeks to establish empirical relations, such as those between the characteristics of a person's schooling and his or her lifetime earnings. It is a matter of taste whether one interprets these studies as hypothesis testing or just establishing statistical relations, but either way they are important.

So the ideas of positive economics are still present – more strongly in some fields than in others – though many economists reject them, as shown either explicitly by their methodological pronouncements or implicitly by their research practices.

See Also

- economics, definition of
- Friedman, Milton
- methodology of economics
- positivism
- Robbins, Lionel Charles
- value judgements

Bibliography

Archibald, G. 1961. Chamberlain versus Chicago. *Review of Economic Studies* 29, 316–27.

Archibald, G. 1967. Refutation or comparison? *British Journal for the Philosophy of Science* 17, 279–96.

Blaug, M. 1992. *The Methodology of Economics*. Cambridge: Cambridge University Press.

Blaug, M. 1998. The positive–normative distinction. In *The Handbook of Economic Methodology*, ed. J. Davis, D. Hands and U. Maki. Cheltenham: Edward Elgar.

Caldwell, B. 1982. *Beyond Positivism: Economic Methodology in the Twentieth Century*. London: Allen and Unwin.

de Marchi, N. 1988. Popper and the LSE economists. In *The Popperian Legacy in Economics*, ed. N. de Marchi. Cambridge: Cambridge University Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Hahn, F. 1984. *Equilibrium and Macroeconomics*. Oxford: Basil Blackwell.

Hodgson, G. 2001. *How Economists Forgot History: The Problem of Historical Specificity in Social Science*. London: Routledge.

Hutchison, T. 1938. *The Significance and Basic Postulates of Economic Theory*. London: Macmillan.

Hutchison, T. 1992. *Changing Aims in Economics*. Oxford: Blackwell.

Klappholz, K. and Agassi, J. 1959. Methodological prescriptions in economics. *Economica* 26, 60–74.

Lind, H. 2007. The story and the model done: an evaluation of mathematical models of rent control. *Regional Science and Urban Economics* 37, 183–98.

Lipsey, R.G. 1963. *An Introduction to Positive Economics*, 1st edn. London: Weidenfeld and Nicolson.

Lipsey, R.G. 1964. Positive economics in relation to some current trends. *Journal of the Economics Association* 5, 365–71. Reprinted in Lipsey (1997).

Lipsey, R.G. 1975. *An Introduction to Positive Economics*, 4th edn. London: Weidenfeld and Nicolson. The page referenced is reprinted in Lipsey (1997).

Lipsey, R.G. 1981. Economists, policy makers and economic policy. In *Economic Policy Making in Canada*, ed. D. Smith. Montreal: C.D. Howe Institute. Reprinted in Lipsey (1997).

Lipsey, R.G. 1997. *The Selected Essays of Richard Lipsey. Volume I: Microeconomics, Growth and Political Economy*. Cheltenham: Edward Elgar.

Lipsey, R.G. 2001. Successes and failures in the transformation of economics. *Journal of Economic Methodology* 8, 169–201.

Lipsey, R.G., Carlaw, K. and Bekar, C. 2005. *Economic Transformations: General Purpose Technologies and Long-Term Economic Growth*. Oxford: Oxford University Press.

Robbins, L. 1935. *An Essay on the Nature and Significance of Economic Science*, 2nd edn. London: Macmillan.

Samuelson, P. 1948. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Wong, S. 1978. *The Foundations of Paul Samuelson's Revealed Preference Theory: A Study by the Method of Rational Reconstruction*. London: Routledge and Kegan Paul.

How to cite this article

Lipsey, Richard G. "positive economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000130> doi:10.1057/9780230226203.1312

The New Palgrave Dictionary of Economics Online

positivism

Bruce Caldwell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The article identifies the major tenets of logical positivism and its successor, logical empiricism, two important movements within 20th-century philosophy of science. It then documents some of the arguments that led to the decline of positivism in the latter half of the 20th century. The impact of positivist ideas on the work of economists writing about economic methodology is examined in a final section.

Keywords

Blaug, M.; Braithwaite, R.; Carnap, R.; Comte, A.; confirmationism; falsifiability; Friedman, M.; general equilibrium; Hahn, H.; Hempel, C.; Hume, D.; Hutchison, T.; hypothetico-deductive model; induction; logical empiricism; logical positivism; Mach, E.; Machlup, F.; methodology of economics; Mill, J. S.; Mises, L. von; Nagel, E.; Neurath, O.; Pareto, V.; philosophy of science; Popper, K.; positivism; Robbins, L.; Samuelson, P.; theory appraisal; verifiability; Vienna Circle

Article

Positivism and the philosophy of science

The term ‘positivism’ was coined in the second quarter of the 19th century by one of the founders of sociology, Auguste Comte. Comte believed that human reasoning passes through three distinct historical stages: the theological, the metaphysical, and the scientific. In the theological stage, natural and social phenomena are explained by reference to spiritual forces. In the metaphysical stage, ‘ultimate causes’ are sought to explain such phenomena. In the scientific stage, attempts to explain phenomena are abandoned, and scientists seek instead to discover correlations among phenomena (Comte, 1830–42). Another important figure in the development of *classical positivism* was the physicist Ernst Mach (1886), who propounded a ‘fictionalist’ view of theories. Scientific theories are useful mnemonic

devices, but progress in science occurs only when such useful fictions are replaced by statements which contain only observation terms. Though both Comte and Mach had some influence on the writings of economists (Comte influenced J.S. Mill and Pareto, Mach was mentioned in passing by Samuelson and Machlup), their primary influence was on the ideas of certain 20th century philosophers of science, the logical positivists.

Logical positivism

The major tenets of *logical positivism* were developed in the 1920s by Moritz Schlick, Herbert Feigl, Kurt Gödel, Hans Hahn, Otto Neurath, Friedrich Waismann, Rudolf Carnap and other members of the famous Vienna Circle. Logical positivism was a radically empiricist philosophical position, and its founders believed it marked a new beginning for philosophical inquiry. The goal of all philosophical analysis was henceforth to be the logical analysis of the knowledge claims of the positive, or empirical, sciences: hence the label ‘logical positivism’.

The first task facing the logical positivists was to define what constitutes a knowledge claim. Their solution was to analyse the logical form of statements. Only statements that are either analytic (such as definitions) or synthetic (testable statements of fact) qualify as cognitively significant, or meaningful. All other statements lack cognitive significance: they are meaningless, metaphysical, non-scientific. Analyses that make use of such statements may express emotional stances, or ‘general attitudes towards life’, or moral valuations, but they do *not* express knowledge claims.

To put their programme into operation, the logical positivists needed an objective criterion of cognitive significance which could be used to distinguish synthetic statements from meaningless ones. One early solution was the principle of verifiability: a synthetic statement has meaning only if it is verifiable. Unfortunately, statements of universal form (for example, ‘all ravens are black’), which are frequently encountered in science, are unverifiable. Other criteria included falsifiability, Ayer's weak verifiability, Carnap’s translatability into an empiricist language, and confirmability. None of these was able to resolve the problem conclusively, however. Another dilemma was posed by the presence of theoretical terms in statements made by scientists. Some positivists followed Mach in insisting that they should be eliminated from science, while others argued that such statements should be retained. A final element of the logical positivist programme was an emphasis on the unity of science, variously defined as meaning that all true sciences share a common method, that the results of all sciences should ultimately be expressible in a common physicalist language, or that the results of the various sciences should be integrated, better to assist the scientific planning of society.

Logical empiricism

Hahn died in 1934, and Schlick was murdered in 1936 by an insane student. But it was Hitler's rise to power, and the subsequent flight of intellectuals, that primarily caused the disintegration of the Vienna Circle in the 1930s. Logical positivism was modified and ultimately replaced over the next two decades by a more analytically austere form of positivist thought, *logical empiricism*. Though differences exist in their analyses, philosophers who have contributed to this later tradition include Carnap, Ernest Nagel, Carl Hempel and Richard Braithwaite.

There were six major tenets of the logical empiricist programme. First, the *unity of science thesis* was

narrowed to mean only a unity of scientific methods. The next three had to do with the structure and appraisal of theories. The *hypothetico-deductive model of theory structure* states that all sciences employ theories, which may be represented formally as axiomatic, hypothetico-deductive structures. Such structures have no empirical import until some of their elements (usually the deduced theorems, or predictions of the theories) are given an empirical interpretation via the use of correspondence rules. Not every statement will have an empirical interpretation. Those containing theoretical terms, in particular, will not be interpretable. Are such sentences then meaningless? Not at all; according to the *indirect testability thesis* such sentences gain cognitive significance indirectly when the theories in which they are embedded are confirmed. Finally, concerning the questions of demarcation and theory assessment, logical empiricists settled on *confirmationism* as their primary criterion of theory appraisal. A theory is scientific if it is testable; test instances confirm or disconfirm the theory; the acceptability of the theory depends on its degree of confirmation. Degree of confirmation is measured by such things as the quantity and precision of favourable test outcomes, the precision of procedures of observation and measurement, the variety of supporting evidence, and whether new test situations support the hypothesis. Additional non-empirical criteria of appraisal (for example, simplicity, elegance, fruitfulness, generality, extensibility) may also be invoked if theory choice on empirical grounds yields no preferred theory. The last two tenets of logical empiricism concerned the logic of scientific explanation. All explanations in science must be expressible in the form of a deductive argument in which an explanandum, a sentence describing the event to be explained, is logically deduced from an explanans. The explanans contains a group of sentences, some of which express initial conditions, and at least one of which states either a general or a statistical law. The *deductive-nomological* and *inductive-probabilistic covering law models of scientific explanation* take their names, then, from the types of laws (general or statistical) used in the explanations. Additionally, logical empiricists believed in the *symmetry thesis:* explanation and prediction are structurally symmetrical, the only difference between them being one of temporality. In the case of an explanation the phenomenon described in the explanandum has already taken place, whereas in the case of a prediction it has not yet occurred.

As documented in Suppe (1977), logical empiricist ideas (sometimes dubbed ‘the received view’) came under heavy attack in the mid-20th century. The viability of both the hypothetico-deductive model of theory structure and the indirect testability thesis depended on one’s ability to draw a clear distinction between observational terms (terms that refer to observables, to ‘brute, atomic facts’) and non-observational, theoretical terms. Unfortunately, in many sciences there are degrees of observability, and no hard division can be drawn between theoretical terms that refer to non-observables and non-theoretical terms that refer to observables. Furthermore, because observation itself is not a neutral activity but requires both data selection and interpretation, it was argued (by critics like Karl Popper and Norwood Hanson) that all observation is theory-dependent. Regarding confirmationism, the failure to solve Hume's problem of induction and a number of paradoxes of confirmation undercut attempts to construct an inductive logic of confirmation. In addition, Popper (1959) challenged the desirability of making statements that have a high inductive probability. Finally, many explanations in a variety of sciences could not be reconciled with the two covering law models of scientific explanation.

The naturalistic turn

The influence of positivism within the philosophy of science declined considerably through the 1960s

and 1970s. As noted by Hands (2001), its apparent successor has been dubbed the *naturalistic turn*, an approach that, rather than laying out a priori criteria for identifying appropriate scientific practice, instead employs the tools of the sciences themselves to investigate scientific practice. There are, of course, many different scientific disciplines from which to draw such tools; some that have been used are cognitive psychology, evolutionary biology, sociology, and economics. Depending on which scientific practice is analysed, reflexivity issues may appear (for example, in using economic analysis to explain the development of economics and the behaviour of economists). Other important issues facing the naturalistic turn are choosing among the various tools on offer, and deciding whether the ensuing analysis has prescriptive implications in addition to descriptive merits. Another movement that has had less impact in philosophy of science proper, but great influence in a number of sciences including economics, derives from the work of Karl Popper. A critic of inductivism and confirmationism, the father of falsifiability and of critical rationalism, Popper had sufficient insight, foresight and longevity to influence a number of generations of philosophers of science, among them J. Agassi, W.W. Bartley III, P.K. Feyerabend and Imre Lakatos. Within economics, the work of T.W. Hutchison (for example, 1997), Mark Blaug (1992) and Lawrence Boland (2003) most directly reflect Popper's influence, while that of Wade Hands (1993) and Bruce Caldwell (1991) reflect a critical reappraisal.

In the 1990s an historical dehomogenization of the writings of the logical positivists of the Vienna Circle began. A rehabilitation of Otto Neurath, whose anti-foundationalism, advocacy of pluralism, and emphasis on scientific practice led many to see him as a precursor of the naturalistic turn, was the most notable result (Uebel, 1991). Some historians and philosophers also praised his willingness to advocate the scientific planning of society and of science, to employ the philosophy of science as a tool in the restructuring of society. For these interpreters, the emergence of a more austere logical empiricism in the 1950s represented not a scientific advance but a retreat to more neutral formalism in response to the ideological pressures of McCarthyism and the cold war (for example, Reisch, 2005). This interpretation parallels Philip Mirowski's (2002) historical account of the development of formalism in economics during the same period.

Positivism and economics

There are various ways to describe the influence of positivist thought in economics. If one focuses on the period in which positivist philosophy of science was invoked by economists, the positivist epoch spanned roughly 40 years, from the late 1930s to the late 1970s. This is not to say that during this period economists self-consciously adopted the philosophical positions outlined above. As shown in Caldwell (1994), what in fact occurred was that certain economists writing about methodology borrowed, usually somewhat haphazardly, from the language of positivism, while others invoked various positivist positions to defend or to criticize theories and practices in economics.

Four economists from this period whose writings most reflect the influence of positivism are T.W. Hutchison, Fritz Machlup, Paul Samuelson, and Milton Friedman. In the 1938 book, *The Significance and Basic Postulates of Economic Theory*, Hutchison launched an empiricist attack on the pure logic of choice, a doctrine that had been espoused and defended by Lionel Robbins six years earlier in his *The Nature and Significance of Economic Science* (1932). For more than 50 years, Hutchison was to continue to criticize all forms of economics that were based on untestable foundations, his targets

ranging from the apriorism of Ludwig von Mises to the elaborate mathematical models of general equilibrium theory. Fritz Machlup offered one response to Hutchison with his 1955 paper, ‘On the problem of verification in economics’, where he invoked the indirect testability thesis to defend the use of theoretical constructs in economics against what he dubbed Hutchison's ‘ultra-empiricism.’ In the Introduction of his *Foundations of Economic Analysis*, Paul Samuelson (1947) borrowed from the work of physicist Percy Bridgman when he insisted that economists search for operationally meaningful theorems. The intent of Samuelson's revealed preference approach to demand theory was to place consumer theory on an observational basis. Finally, Milton Friedman's influential 1953 piece ‘The methodology of positive economics’ contained the famous argument that the realism of the assumptions of a theory is irrelevant; what counts in the assessment of a theory is its relative predictive adequacy and its simplicity. Though Friedman's unique brand of instrumentalist methodology owes more to the American pragmatists than to positivism, his approach came to be viewed as synonymous with positivism through the 1950s and 1960s.

Though economists today rarely invoke positivist philosophy of science in defending their preferred practices, there is plentiful evidence of its continued influence, mostly in terms of what is considered to be ‘appropriate’ or ‘legitimate’ practice, with ‘positivist’ often being equated with ‘truly scientific’. Thus, important areas like game theory and transactions cost analysis initially encountered substantial opposition from mainstream economists because such analyses, though rich in terms of explaining diverse economic phenomena, often did not produce the sort of testable hypotheses demanded by positivist doctrine. (Strangely, during its period of dominance, general equilibrium theory was much less affected by such critiques.) Similarly, the positivist belief in the cumulative development of science tends to render less important both heterodox approaches to the discipline and the study of doctrinal history. Finally, the insistence on defining progress in terms of ‘the discovery of law-like relationships’ or ‘better predictive ability’ has fuelled a sustained growth in data collection and in computing power, the development of new econometric techniques, and a staggering increase in empirical studies. That all this has resulted in at best meagre progress (see Backhouse, 1997) in establishing robust economic ‘laws’ and in improving forecasting power has typically engendered not a reassessment of the goals but a redoubling of resources committed to reaching them, with the attendant opportunity costs. It will be interesting to see what the entry on ‘positivism’ in the third edition of *The New Palgrave* reveals about its legacy in economics.

See Also

- methodology of economics
- philosophy and economics
- theory appraisal

Bibliography

Backhouse, R. 1997. *Truth and Progress in Economic Knowledge*. Cheltenham: Edward Elgar.

Blaug, M. 1992. *The Methodology of Economics: Or How Economists Explain*, 2nd edn. Cambridge:

Cambridge University Press.

Boland, L. 2003. *The Foundations of Economic Method: A Popperian Perspective*. London: Routledge.

Caldwell, B. 1991. Clarifying Popper. *Journal of Economic Literature* 29, 1–33.

Caldwell, B. 1994. *Beyond Positivism: Economic Methodology in the Twentieth Century*. London: Routledge.

Comte, A. 1830–42. *Cours de philosophie positive*. Paris: Bachelier.

Friedman, M. 1953. The methodology of positive economics. In M. Friedman, *Essays in Positive Economics*. Chicago: University of Chicago Press.

Hands, D. 1993. *Testing, Rationality and Progress: Essays on the Popperian Tradition in Economic Methodology*. Lanham, MD: Rowman and Littlefield.

Hands, D. 2001. *Reflection without Rules: Economic Methodology and Contemporary Science Theory*. Cambridge: Cambridge University Press.

Hutchison, T. 1938. *The Significance and Basic Postulates of Economic Theory*. Reprinted, with a new preface, New York: Kelley, 1960.

Hutchison, T.W. 1977. *Knowledge and Ignorance in Economics*. Chicago: University of Chicago Press.

Mirowski, P. 2002. *Machine Dreams: Economics Becomes a Cyborg Science*. Cambridge: Cambridge University Press.

Mach, E. 1886. *The Analysis of Sensations*, ed. S. Waterlow, trans. C. Williams. New York: Dover, 1959.

Machlup, F. 1955. The problem of verification in economics. *Southern Economic Journal* 22, 1–21.

Popper, K. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

Reisch, G. 2005. *How the Cold War Transformed Philosophy of Science: To the Icy Slopes of Logic*. Cambridge: Cambridge University Press.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*, 2nd edn. London: Macmillan, 1935.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge: Harvard University Press.

Suppe, F. (ed.) 1977. *The Structure of Scientific Theories*. 2nd edn. Urbana: University of Illinois Press.

Uebel, T. (ed.) 1991. *Rediscovering the Forgotten Vienna Circle: Austrian Studies on Otto Neurath and the Vienna Circle*. Dordrecht: Kluwer.

How to cite this article

Caldwell, Bruce. "positivism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000131> doi:10.1057/9780230226203.1313

The New Palgrave Dictionary of Economics Online

Post Keynesian economics

J.E. King
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Post Keynesian economics is a dissident school in macroeconomics based on a particular interpretation of Keynes. A brief intellectual history of Post Keynesian ideas is provided, along with a discussion of some important methodological questions. Three short-period macro models are outlined: Paul Davidson's aggregate supply–aggregate demand model, Michal Kalecki's two-class model, and Hyman Minsky's financial instability hypothesis. The Post Keynesian approach to economic growth is shown to focus on the expansion of aggregate demand, with a distinctive approach to monetary, fiscal and other dimensions of macroeconomic policy. In conclusion the future prospects of Post Keynesian economics are assessed.

Keywords

administered prices; American Economic Association; animal spirits; asset price inflation; balance of payments constraint; Baran, P. A.; barter; behavioural macroeconomics; bounded rationality; bubbles; budget deficits; business cycle theory; Cambridge capital controversies; capital accumulation; capitalism; capital–labour substitution; central banks; class; cost-push inflation; critical realism; deflation; effective demand; endogenous growth; endogenous money; entrepreneurship; ergodicity and non-ergodicity in economics; European Central Bank; evolutionary economics; explanation; financial regulation; fiscal consolidation; floating exchange rates; functional finance; fundamentalist Keynesians; general equilibrium; government failure; gross substitution; Harcourt, G.; Harrod–Domar growth model; hedging; Hicks, J. R.; hysteresis; inequality; income distribution; income effects; inflation; institutional economics; international financial institutions; investment decisions; involuntary unemployment; IS–LM; justice; Kahn, R. F.; Kaldor, N.; Kalecki, M.; Keynes, J. M.; labour supply; lending controls; lexicographic preferences; liquidity preference; long run and short run; Lucas, R.; marginal productivity theory; market failure; market imperfections; market power; Marx, K. H.; Meade, J. E.; Means, G.; methodology; microfoundations; Minsky, H.; monetarism; monetary policy; money illusion; myopia; national debt; neoclassical growth theory; neoclassical synthesis; neutral money axiom; New Classical Economics; New Classical Macroeconomics; New Keynesian Economics; open-system thinking; paradox of costs; paradox of thrift; Phillips curve; political business cycles; Ponzi finance; Post Keynesian economics; post-Walrasian theory; price rigidity; prices and incomes policies; profit rate; quantity theory of money; rational expectations; representative agents; reserve requirements; Robinson, J. V.; savings propensities; Say's Law; Solow, R.; speculation; Sraffa, P.; Stability and Growth Pact; Steindl, J.; stocks and flows; structural adjustment; structural change; stylized facts; substitution effects; Sweezy, P. M.; Taylor rule; trade unions; transformational growth; uncertainty; wage rigidity; Washington Consensus; Weintraub, S.

Article

Post Keynesian economics is a dissident school of macroeconomic thought based on a particular interpretation of John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936).

Post Keynesian economics developed in the 1950s and 1960s in Cambridge (UK) and in the United States in the course of a critique of the so-called ‘neoclassical synthesis’ (sometimes also described as Old or Bastard Keynesianism). It represents both a recovery and an extension of Keynes's ideas (Palley, 1996): a recovery, because Post Keynesians believe that the neoclassical interpretation of Keynes is profoundly misleading, and an extension, since they deal with important questions that Keynes neglected or

ignored, including income distribution, social conflict, economic growth and inflation. Post Keynesian economics involves a distinctive approach to methodology, theory and policy (Holt and Pressman, 2001; King, 2003).

At the heart of Post Keynesian theory is the principle of effective demand, according to which output and employment are generally demand-constrained rather than supply-constrained. Post Keynesians claim to take the principle of effective demand more seriously do than mainstream macroeconomists, even those who describe themselves as ‘Keynesians’. For Post Keynesians, demand constraints upon output and employment are not restricted to short period and are not the result of market imperfections or wage and price rigidities, but must be explained instead in terms of the characteristics of money and the pervasive influence of fundamental uncertainty. The six central messages of Keynes's vision may be summarized as follows. First, output and employment are determined in the product market, not in the labour market. Second, involuntary unemployment exists. Third, an increase in savings does not automatically generate an equivalent increase in investment. Fourth, a monetary economy is fundamentally different from a barter economy. Fifth, the quantity theory of money holds only under full employment, but cost-push forces may generate inflation well before this point is reached. Sixth, capitalist economies are driven by the ‘animal spirits’ of entrepreneurs, which determine the decision to invest (Thirlwall, 1993).

It follows that Say's Law is false, and capitalism will normally not achieve or sustain full employment without government intervention. Post Keynesians therefore advocate the systematic use of fiscal and monetary policy to regulate aggregate demand, and deny the policy ineffectiveness propositions of mainstream macroeconomics. They advocate prices and incomes policies, rather than restrictive monetary policy, to control inflation.

A brief intellectual history

The origins of Post Keynesian economics may be traced back to the publication of the *General Theory* in 1936, since Keynes's masterpiece was open from the outset to alternative interpretations (King, 2002). One of them, the IS–LM model developed by J. R. Hicks, James Meade and others, subsequently formed the core of the neoclassical synthesis model of output and employment in the short run. However, the Cambridge (UK) Post Keynesians, including Richard Kahn, Nicholas Kaldor, Joan Robinson and Piero Sraffa, directed their early criticisms against the long-run component of the neoclassical synthesis, the Solow growth model, in which full employment was ensured by capital–labour substitution along a well-behaved aggregate production function. The ‘Cambridge capital controversies’ of the late 1950s and early 1960s demonstrated the analytical failure of neoclassical growth theory, and were an important episode in the emergence of the Post Keynesian school (Mata, 2004). Subsequently Robinson, Kaldor and the American Sidney Weintraub attacked the monetarist theory of inflation, emphasizing the causal role of the rate of change of money wages and arguing that monetary growth was the effect of inflation, not its cause. Kaldor, Weintraub and another American, Paul Davidson, were early advocates of the theory of endogenous money (Kaldor, 1970).

Robinson conducted a lengthy correspondence with yet another American dissident, Alfred Eichner (Lee, 2000). When, in December 1971, she gave the keynote Richard T. Ely lecture at the American Economic Association meeting in New Orleans to a large and enthusiastic audience, the defeat of the orthodox paradigm seemed to be only a matter of time (Robinson, 1972). By the mid-1970s the term ‘Post Keynesian’ was widely used to describe the emerging school of thought (Eichner and Kregel, 1975), which had broadened to include a systematic critique of the neoclassical synthesis. The IS–LM model was rejected, since uncertainty and animal spirits rendered the IS curve unstable, and endogenous money undermined the LM function. The Phillips curve model of wage inflation was rejected in favour of a socio-political analysis of distributional conflict and its resolution in a class society where capitalists enjoyed product market power and workers were highly unionized. And the marginal productivity theory of income distribution, discredited in the capital controversies, was replaced by a macroeconomic model that focused on the different savings propensities of capitalists and workers, or companies and households.

The critique of Chicago School monetarism was soon extended to New Classical Economics (‘monetarism Mark II’). Post Keynesians objected to the principle of rational expectations, since it ignored the existence of fundamental uncertainty, and they denied the claim of Lucas and his associates that fiscal and monetary policy could have no effect on output or employment. They saw only slightly more merit in New Keynesian Economics, with its emphasis on market imperfections as the source of all macroeconomic problems. More was involved than disagreements on economic theory; important methodological and policy issues were also at the heart of these criticisms.

The mainstream never accepted the Post Keynesian critique. Orthodox Keynesian macroeconomists like Robert Solow accused the Post Keynesians of incoherence; they were united, he suggested, only by what they were against. Insiders distinguished three Post Keynesian schools, the Kaleckians, the Sraffians and the ‘fundamentalist Keynesians’, with a number of prominent individualists who belonged to none of them (Harcourt, 1987; Hamouda and Harcourt, 1988). Divisions remain on the respective virtues of a ‘big tent’ and a ‘small tent’ definition of Post Keynesianism.

In 2006 Post Keynesians were a small, embattled minority, strongest in France, Italy and a few institutions in the United States (especially the University of Missouri at Kansas City), with outposts in Britain and Australia. They published in a range of heterodox journals, especially in the *Journal of Post Keynesian Economics*, founded by Davidson and Weintraub in 1977, in the *Cambridge Journal of Economics*, *Journal of Economic Issues* and *Review of Political Economy*.

Methodology

As is common with dissenting schools of thought, Post Keynesians have always recognized the importance of methodology. They have been strongly influenced by Keynes's philosophical writings (O'Donnell, 1989), identifying with his insistence on 'open-system thinking' and organic rather than atomistic models of human behaviour, his distrust of formalism (and of econometric modelling in particular), and his belief that economics was nothing if not a policy science – or an art, perhaps.

On many specific methodological questions, Post Keynesians stress their differences with the mainstream (Dow, 1996). They doubt the relevance of equilibrium models, which fail to allow for cumulative causation, the role of history or the importance of hysteresis. Their emphasis on uncertainty leads them to reject the 'rational expectations' principle and to assert the importance of habit, convention and social institutions in the formation of business expectations. Post Keynesians criticize the mainstream insistence that 'microfoundations' must always be provided for macroeconomic theory, because this denies the existence of emergent properties of macro systems that cannot be inferred from their micro components, and thus involves a fallacy of composition. Microeconomic theory needs *macrofoundations*, they maintain. Finally, Post Keynesians take a quite distinctive approach to long-run theory. Since the principle of effective demand applies in the long run, no less than in the short run, Post Keynesian growth theory stresses the role of demand as a determinant of economic growth, and does not impose a condition that resources (including labour) are always fully employed, and output constrained solely by supply, in the long run.

Many (though not all) Post Keynesians are attracted by *critical realism* as a unifying methodological position (Lawson, 2003). There are many points of contact, including the critical realists' endorsement of open-system thinking; their denial of 'event regularities' of the type needed if standard econometric estimation techniques are to be generally reliable; and their stress on the importance of ontology and the identification of causal processes and mechanisms as the key to explanation in social science. Critical realism has become a significant point of contact between Post Keynesians and other schools of heterodox economic thought.

Macroeconomic theory: the short period

The short-period theory of output and employment is the core of Post Keynesian macroeconomics. In the short period, the capital stock is held constant. This is done purely for analytical convenience; there is no presumption that the theory of effective demand is irrelevant to the long period, when the accumulation of capital is brought into the analysis. Thus the Post Keynesian treatment of the short and long periods must be distinguished from the neoclassical analysis of the 'short run' (in which demand matters) and the 'long run' (when it does not).

There is no single canonical short-period Post Keynesian model. The three most influential models are those of Paul Davidson, Michal Kalecki and Hyman Minsky, which differ in some important respects. They are not, however, entirely incompatible. All agree on the central role of the principle of effective demand; the defects of the IS–LM model; the importance of uncertainty, money and finance (this is largely implicit in the Kalecki version, and quite explicit in the other two); the consequent repudiation of rational expectations; and the policy implications, which include the need for government intervention to stabilize the economy and to maintain full employment. They differ on some questions of microeconomics (there is no question of providing microfoundations), on the detailed treatment of money and finance, and most obviously on the social and political context, above all on the class-driven or class-blind nature of the analysis.

The fundamentalist Keynesian model

This is an elaboration of the aggregate supply-aggregate demand model set out by Keynes himself in the early chapters of the *General Theory* (Keynes, 1936, ch. 3). (Note that this is emphatically *not* the textbook model in price level/real output space, which is a teaching version of the neoclassical synthesis and would be repudiated by all Post Keynesians.) Originating in the 1950s with Weintraub, it has been propagated tirelessly over several decades by Davidson in a series of books beginning with *Money and the Real World* in 1973 and culminating in *Financial Markets, Money and the Real World* (Davidson, 2002). In Figure 1 the aggregate supply function Z_w links total employment to total expected sales; it slopes upwards since total costs of production (including gross profits) increase as employment rises. The aggregate demand function D_w does not coincide with Z_w, as it would in an economy where Say's Law prevails. It, too, slopes upwards, as planned spending also rises with employment. The point of effective demand is A, where the two curves intersect, and aggregate employment is given by N_a. The labour market implications are illustrated in Figure 2 (which, like Figure 1, comes from Davidson, 1999, not directly from Keynes). Employment is determined in the product market, in Figure 1. The real wage can be established from the market equilibrium curve of labour-hire (MECL) in Figure 2; it is W_{ra}. Davidson emphasizes that MECL is *not* the labour demand curve; employment depends on aggregate demand and is therefore determined in the product market, not the labour market. Involuntary unemployment is N_aN_b; it is not due to (real or money) wage rigidity, but to deficient effective demand. For full employment to be achieved, the aggregate demand curve would need to shift to D'_w, increasing employment to N_f, with a movement from A to F in Figure 1 and a corresponding move from A' to F' in Figure 2. Without such a shift there will be no increase in employment, no matter how willing workers might be to accept a cut in either money or real wages. In fact Figure 2 reveals that W_{ras} is the reservation wage of the marginal unemployed worker, but in the absence of an adequate level of effective demand this is simply not relevant.

Figure 1

Aggregate supply and demand. *Source*: Davidson (1999, p. 582).

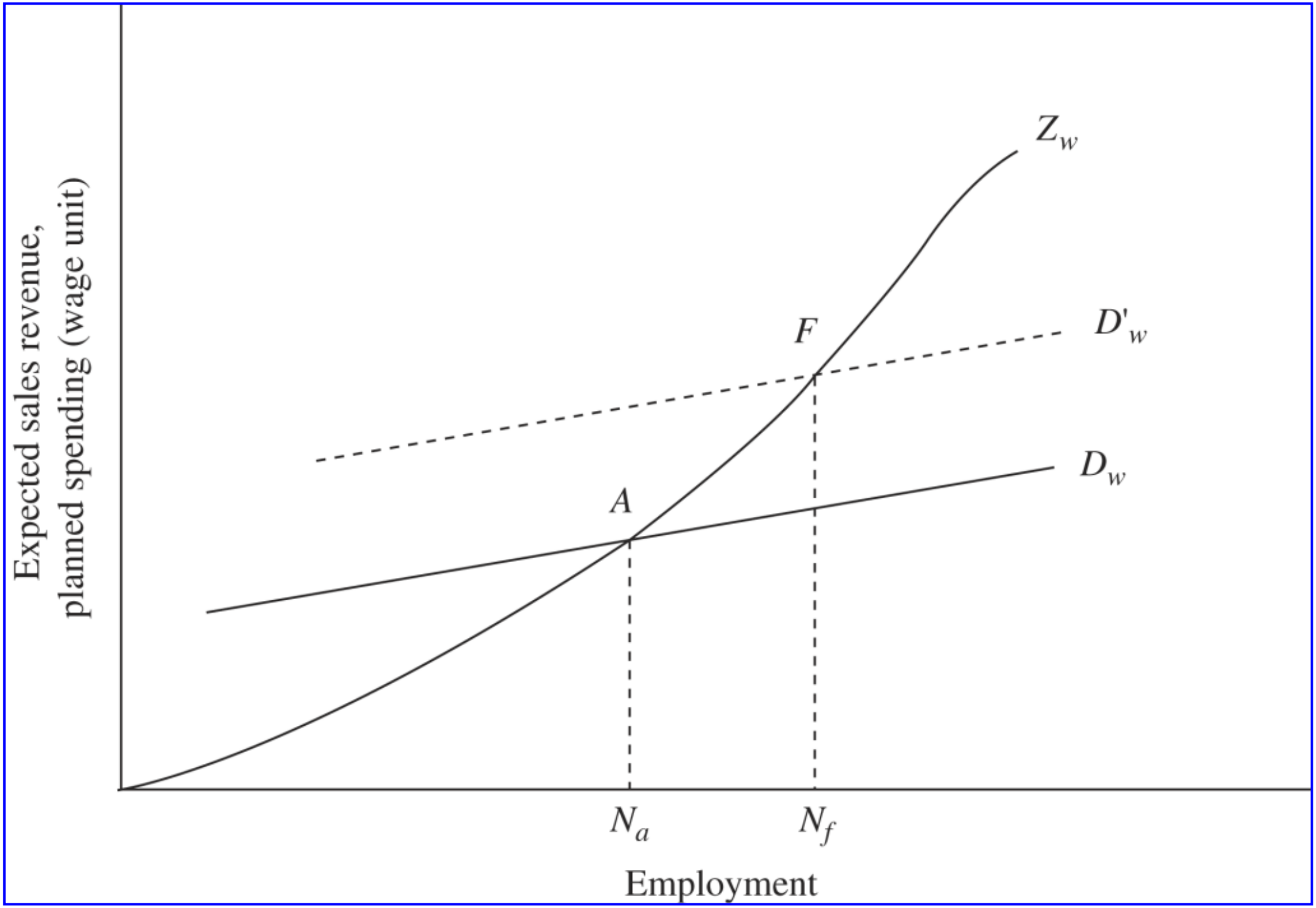

Figure 2
Wages and employment. *Source*: Davidson (1999, p. 582).

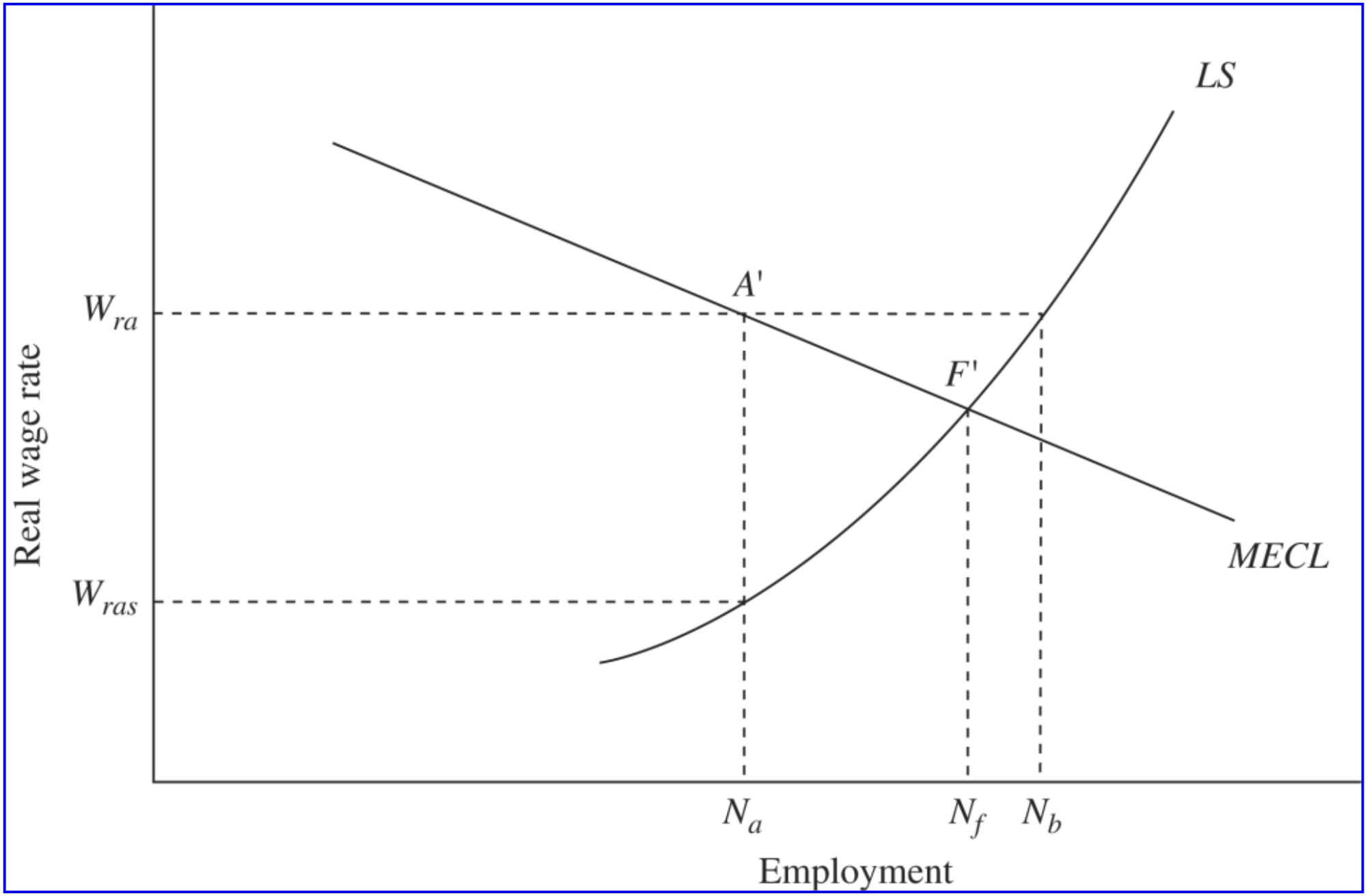

Underlying Keynes's model, Davidson argues, is a rejection of the three fundamental axioms of mainstream macroeconomics. The *axiom of ergodicity* asserts that the future can be reliably inferred from the past. The *axiom of gross substitution* asserts that flexibility in relative prices will ensure that all markets clear. The *neutral money axiom* ensures that changes in the stock of money have no permanent effects on real output or employment. Non-ergodicity creates radical uncertainty, which induces people to hold money; since goods are not perfect substitutes for money, money is not neutral, even in the long run, and Say's Law is false. The non-neutrality of money does not require ‘money illusion’ on the part of any agent. Involuntary unemployment is not the result of wage or price rigidity, and can be eliminated only by increases in effective demand. Wage reductions will prove futile, or even counter-productive (since deflation depresses business and household confidence and increases real interest rates).

The Kaleckian model

Michal Kalecki discovered the principle of effective demand in the 1930s independently of Keynes under the influence of Rosa Luxemburg and, through her, of Karl Marx. The class distinction between capitalists and workers, which is only implicit in the *General Theory*, occupies centre stage in Kalecki's analysis. In place of a single consumption function, there are two. Workers save nothing from their wages, while capitalists save a constant proportion of their profit income. Kalecki's famous aphorism that ‘capitalists get what they spend, while workers spend what they get’, can be derived from the simple income-expenditure model set out in his 1939 *Essays in the Theory of Economic Fluctuations* (1990, pp. 233–318), itself an elaboration of his 1933 model of the business cycle (1990, pp. 65–108). In the simplest case, if we neglect both the government and the foreign sector, total wage (W) and profit (P) income is equal to the sum of consumption expenditure by workers (C_w) and capitalists (C_c) plus investment spending (I):

$$W + P = C_w + C_c + I. \quad (1)$$

Since $W = C_w$, by assumption, it follows that

$$P = C_c + I, \quad (2)$$

so that profits are equal to the sum of capitalists’ consumption and investment expenditure. If investment is a positive function of expected profits, which are themselves closely related to recent past profits, this leads directly to a demand-driven model of the trade cycle. If we incorporate the government and overseas sectors, eq. (1) can be replaced with

$$W + P + T + M = C_w + C_c + I + G + X \quad (3)$$

where taxes (T) and imports (M) are added to the income side of the equation and government expenditure (G) and exports (X) to the expenditure side. It follows that

$$P = C_c + I + (G - T) + (X - M), \quad (4)$$

so that capitalists profit from both government deficits (G–T) and trade surpluses (X–M).

Unlike Keynes, Kalecki had no time for the marginal productivity theory of distribution. In his model the share of profits in total output is determined by the degree of monopoly in the product market. Outside agriculture, oligopoly rather than perfect competition is the rule. Firms set prices by marking up their average variable costs of production, the markup varying inversely with the degree of competition that they face. This, Kalecki argues, establishes a strong tendency for a chronic deficiency in effective demand, since the wage share will normally be too small (and the profit share too high) to generate enough consumption expenditure to maintain full employment. This aspect of Kalecki's analysis was emphasized by ‘left Keynesians’ like Josef Steindl and the neo-Marxists Paul Baran and Paul Sweezy in their work on *monopoly capital*.

Kalecki also highlights the class nature of capitalist society in the context of macroeconomic policy. Capitalists will resist deficit-financed spending by the government, even though it

might be expected to increase total profits. This is only partly due to an unthinking attachment (encouraged by orthodox economists) to the principles of sound finance. It has more rational roots in their concern to avoid competition from state-owned enterprises, and more especially in their well-founded fear that full employment will prove inconsistent with ‘discipline in the factories’. As early as 1943 Kalecki was predicting the emergence of a *political business cycle*, in which fiscal and monetary policy is repeatedly eased before elections and tightened (under pressure from business interests) soon afterwards (Kalecki, 1990, pp. 347–56). In the 1950s, viewing the Cold War from his native Poland, he criticized the ‘military Keynesianism’ of Western governments – capitalists had welcomed demand-boosting armaments spending while resisting more socially useful civilian expenditures.

The financial instability hypothesis

The Kaleckian model is characterized by a relative neglect of money and finance. Hyman Minsky's version of the short-period Post Keynesian model quite explicitly aimed to put finance back into business cycle theory (Minsky, 1986). His ‘Wall Street vision’ of capitalism focuses on the relationship between investment bankers and their customers, by contrast with the ‘village fair’ conception of exchange between individual small producers that underpins mainstream theory. Like Davidson and Kalecki, Minsky sees fluctuations in investment expenditure as the principal cause of economic fluctuations. The investment decisions of capitalists are constrained by their ability to pay for them, and this is conditioned by lenders’ estimates of their ability to repay. Minsky distinguishes three phases of the cycle. In the immediate aftermath of a crisis lenders are very cautious, and accommodate only those borrowers who can demonstrate an ability to service their loans and repay the principal on time; this is the phase of *hedge finance*. As the upswing gathers pace, and memories of previous difficulties begin to fade, it becomes possible to borrow for more questionable projects, where interest payments are covered by expected profits, but not repayments of principal; this is the phase of *speculative finance*. In the final stages of the boom caution is thrown to the winds and lenders now provide *Ponzi finance* (the term is derived from a notorious early 20th-century swindler): new borrowing is now required to enable borrowers to make interest payments on previous loans. When lenders’ confidence collapses, borrowers are unable to obtain refinance and are forced to sell securities and other assets at ‘fire sale’ prices. In the ensuing financial crisis, real investment falls and the economy moves into recession. The early stages of recovery are again characterized by the provision of hedge finance, and so the cycle repeats itself, over and over again. Memories are short, and expectations are far from rational.

Rather late in his career Minsky discovered Kalecki, and added the Kaleckian theory of profits to his own model. This gave him a theory of firms’ financial resources to set against his original analysis of their financial commitments, and reinforced the policy implications that he had drawn from the financial instability hypothesis. It can be seen from eq. (4) that aggregate profits are increased by higher budget deficits, and reduced by fiscal conservatism. Deficits, then, are good for business. There is a stock dimension as well as a flow dimension to this conclusion. The United States financial system was much less fragile after 1945 than it had been in 1929, Minsky argued, in large part because of the cumulative impact of wartime and post-war deficits. The huge growth in the federal government debt had provided the private sector with massive quantities of risk-free government securities, thereby rendering their asset portfolios much more robust than they had previously been. Minsky was therefore a supporter of big government. He also advocated tight and intrusive regulation of financial markets, and argued that central banks should recognize their duty to act as lender of last resort to Wall Street, no matter how much this increased the dangers of moral hazard. But he doubted whether the inherent instability of the capitalist economy could ever be completely overcome.

Some comparisons

The similarities between these three models are much more important than their differences, especially when they are contrasted with the ‘new consensus’ model of mainstream macroeconomics. There are no ‘microfoundations’, and certainly no attempt is made to ground the analysis in any form of multi-period utility-maximizing model of general equilibrium under rational expectations. This is ruled out by the non-ergodicity axiom in the Fundamentalist Keynesian model, and by the cyclical myopia of borrowers and lending institutions that is central to the financial instability hypothesis (Kalecki's analysis of ‘lender's risk’ and ‘borrower's risk’ has affinities both with Minsky and with Keynes's treatment of fundamental uncertainty). There are no ‘representative agents’: capitalists and workers in Kalecki, borrowers and lenders in Minsky (and bulls and bears in Keynes's analysis of liquidity preference) are structurally and behaviourally heterogeneous. Deflation is viewed as part of the problem – a very important part, at least for Minsky – not as the solution to macroeconomic difficulties. Cyclical fluctuations originate in the private sector, due to the volatility of business investment decisions, not in the policy errors of the public sector. And government intervention is essential to ‘stabilize an unstable economy’, to paraphrase the title of Minsky's last (1986) book – though neither he nor Kalecki minimized the obstacles that it would encounter.

These are not the only short-period Post Keynesian models, though they remain the most influential. They are, it must be repeated, all inconsistent with the ‘new consensus’ in macroeconomics, which can be encapsulated in three equations. Post Keynesians dispute all three. They reject the aggregate demand curve, on the grounds that interest rates are less important, and uncertainty-induced shifts in the curve much more important, than the mainstream is willing to admit. They are equally critical of the Phillips curve, since it neglects socio-political institutions and denies any role for class conflict over income distribution. And they criticize the Taylor rule that underpins the monetary policy response function, as it uses a single instrument (the short-term interest rate) instead of many to influence the wrong objective (output price inflation instead of employment, neglecting asset price inflation). More will be said about Post Keynesian thinking on money and inflation in a later section.

Macroeconomic theory: the long period

In the *General Theory* Keynes analysed the effects of investment in a short-period model in which, by definition and purely as a simplifying assumption, the capital stock was held constant. ‘Generalising the *General Theory*’ to the long period, Post Keynesian theories of capital accumulation take as their starting point the Harrod–Domar growth model (which Kalecki extended further to apply to socialist economies). There is no requirement that capital or labour are fully employed or that the growth path will be stable; this can be expected, as Robinson put it in her *Accumulation of Capital* (1956), only in a mythical ‘golden age’. The Cambridge capital controversies demonstrated that the neoclassical adjustment mechanism – capital–labour substitution in response to changes in relative factor prices – is not in general a reliable one. Differences between the actual, equilibrium (or ‘warranted’) and maximum possible (or ‘natural’) rates of growth might be eliminated through changes in the average propensity to save induced by changes in income distribution. But, Post Keynesians maintain, there are no grounds for supposing that effective demand is unimportant in the long period, or for the neoclassical belief that economic growth is entirely supply-determined.

There are a number of Post Keynesian models of demand-driven growth (Setterfield, 2002). All of them invoke ‘Say's Law in reverse’, according to which aggregate supply (and potential output) responds to the growth of aggregate demand (and actual output). *Kaldorian* models treat exports as the only truly exogenous source of demand and highlight the balance of payments constraint on economic growth, which is especially (but not exclusively) relevant to developing economies. *Kaleckian* models focus on the connection between wages, consumption and aggregate demand, adding to the familiar *paradox of thrift* (in which an increased propensity to save reduces income and keeps the volume of saving unchanged) a *paradox of costs*, in which an increase in the real wage increases workers’ consumption, raises the level of capacity utilization and thereby leads to a higher rate of profit. Finally, there are models of *transformational growth* associated with Luigi Pasinetti and Edward Nell, in which capital accumulation is inextricably linked to structural change. Once again attention is concentrated on demand conditions; this time, however, it is investment demand that plays the crucial role.

These Post Keynesian growth models are all radically different from neoclassical theories, including both the canonical Solow model and the more recent ‘New’ or ‘endogenous growth’ models (though they share with the latter a denial of diminishing returns in the manufacturing and advanced service sectors). The Post Keynesians assert the continuing importance of the principle of effective demand and the irrelevance or reversal of Say's Law, since in the long period demand tends to create its own supply. They have no truck with marginal productivity theory or with the use of aggregate production functions of any description.

There are connections between Post Keynesian growth theory and the treatment of capital accumulation in other heterodox traditions, especially the radical-Marxian focus upon the class nature of capitalist society, the critical role of the profit rate and the instability of the capitalist growth path. Equally, the emphasis placed in evolutionary and Schumpeterian theory on the role of entrepreneurs, the importance of finance and the cyclical nature of growth is fully consistent with the Post Keynesian approach.

Post Keynesian microeconomics

Post Keynesian microeconomics is relatively underdeveloped. There are methodological reasons for this, since (as we have seen) Post Keynesians reject the neoclassical requirement that rigorous microfoundations be provided for macroeconomic theory. Although microeconomics is not needed as the basis for serious macroeconomic thinking, Post Keynesians are nevertheless highly critical of many aspects of mainstream microeconomic analysis, including the modelling of equilibrium, the elimination of uncertainty by expressing all relevant magnitudes in certainty-equivalents, and the reliance on identical or ‘representative’ rather than heterogeneous agents.

As in macroeconomics, in their microeconomics Post Keynesians are concerned with the real world, and insist that formal models must bear a close relation to the ‘stylized facts’ of modern capitalism. Thus Post Keynesian pricing theory (Lee, 1998) addresses itself to the large oligopolistic corporation, not to an imaginary world of perfect competitors. Drawing on the work of Kalecki, Philip Andrews, Gardiner Means, Paolo Sylos-Labini and Alfred Eichner, it models the formation of *administered prices*, with firms first adding a markup to their variable costs of production and then selling as much as they can given the prevailing demand conditions. Prices increase only if costs rise, or under quite exceptional demand pressure. This also provides a Post Keynesian theory of income distribution, since the average degree of monopoly, which determines markups, is the most important determinant of the income shares of wages and profits. Changes in the degree of monopoly have other macroeconomic consequences, for both inflation and aggregate demand.

Post Keynesian consumer theory is still in its infancy. It places more emphasis on income effects than on substitution effects and replaces the neoclassical axioms of rational choice with a theory of lexicographic preferences where habit, custom and social convention are important constraints on individual behaviour (Lavoie, 1992, pp. 61–92). The full implications for labour supply decisions have yet to be fully worked out. In their microeconomics Post Keynesians have drawn heavily on insights from other more or less heterodox schools of thought, especially institutional and evolutionary economics. Much remains to be done to extend and deepen Keynes's early exploration of the role of habit, conventions, bounded rationality and rules of thumb in individual and corporate decision-making. The lack of a distinctive Post Keynesian welfare economics is a particularly important weakness, which has hindered the emergence of a coherent approach to environmental issues (Winnett, 2003).

Economic policy

There is, however, a very clear Post Keynesian position on matters of macroeconomic policy. Since Say's Law is rejected, in both the short period and the long period, the principle of effective demand is the foundation for monetary and fiscal policy. This leads Post Keynesians to a broadly social democratic position, not far removed from that of the Old Keynesian advocates of the 1950s–1960s neoclassical synthesis. Thus they favour big government, since it is more likely than small government to be able to stabilize the level of economic activity and achieve full employment. Post Keynesians worry much less about state failure than about market failure. Unlike both Old and New Keynesians, however, they are very clear that market imperfections, and the associated wage and price rigidities, are not at the root of macroeconomic problems. There is no point in using an imaginary world of perfect competition as a reference point. Deflation, even if it were practicable, would be undesirable and counter-productive. Increased inequality is also likely to have adverse macroeconomic consequences, notably if the Kaleckian ‘paradox of costs’ applies, so that stabilization policy need not conflict with the imperatives of social justice. Post Keynesian rejection of neoliberal policies carries over to a comprehensive critique of the ‘Washington Consensus’ on policy for developing countries. At the same time Post Keynesians are not Stalinists; they aim to make markets work better, not to eliminate them. This, they argue, requires wide-ranging government intervention, with a number of macroeconomic targets and a variety of instruments.

Monetary policy

Post Keynesian thinking on monetary policy developed out of opposition to monetarism in the early 1970s and to the policy prescriptions of New Classical macroeconomics in later years. Post Keynesians insisted that, since money was endogenous, the stock of money was not a control variable or a feasible policy instrument. Thus monetary policy must necessarily operate via central bank control over the (short-term) rate of interest, and would inevitably have consequences for output and employment as well as for the inflation rate. In this they were proved to be entirely correct, and they are entitled to view the treatment of monetary policy in the ‘new consensus’ as a vindication of the Post Keynesian critique. However, they also criticize the Taylor rule on the grounds that it is aimed at the wrong target and relies upon a single, very blunt instrument. There is a strong case, they argue, for reviving prices and incomes policy to combat the danger of inflation, using monetary policy to target and output and employment, asset price inflation and financial fragility. Post Keynesians regard stock market and housing bubbles, and rising levels of household and corporate debt, as serious problems that need policy solutions. No single-instrument approach to these problems will succeed. Alternative instruments include the (re-)introduction of direct controls over lending, the tightening of financial regulations, and more market-friendly measures such as asset-based reserve requirements that would operate as a tax on types of lending that the authorities wished to discourage. Post Keynesians have been particularly critical of the high-interest policy adopted by the European Central Bank and the apparent absence of a lender of last resort within the Eurozone.

Fiscal policy

Post Keynesians are no less critical of recent mainstream thinking on fiscal policy. Here the Old Keynesian principle of *functional finance* has given way to the pre-Keynesian principle of *sound finance*, which is invariably interpreted as requiring balanced budgets in the short run and fiscal consolidation (budget surpluses, and a reduction in government debt) in the long run. For Post Keynesians, the principle of effective demand should govern fiscal policy, and governments should run deficits, or surpluses, or (exceptionally) balanced budgets, depending solely on the macroeconomic requirement of achieving full employment with an acceptably low inflation rate. Some would go further in the direction of ‘unsound finance’, since in the Kaleckian short-period model permanent deficits mean permanently higher business profits and hence higher – not lower – levels of investment expenditure; private spending is *crowded in* by government spending, not crowded out. Hyman Minsky added a stock dimension to this argument: the government debt accumulated as the sum of past deficits serves to render private sector balance sheets more robust and thus to reduce the danger of financial instability. The European Union's ‘Stability and Growth pact’ is in fact a recipe, Post Keynesians argue, for stagnation and instability.

Prices and incomes policy

Post Keynesians deplore the way in which one entire area of macroeconomic policy has disappeared completely from the mainstream agenda: prices and incomes policy is no longer taken seriously as an anti-inflationary instrument, and indeed is rarely discussed. In the 1960s and 1970s Post Keynesians led the way in proposing innovative alternatives to deflationary monetary and fiscal policies, including the tax-based incomes policy proposed by Wallich and Weintraub (1971) and more centralized neo-corporatist social agreements that worked well for many years in much of northern Europe (Cornwall and Cornwall, 2003). Weaker unions, reduced levels of industrial conflict and much lower rates of wage and price inflation have reduced the attraction of measures such as these, but Post Keynesians regard them as a valuable policy resource should the inflation dragon once more rear its ugly head.

Other policy issues

On questions of international economic policy Post Keynesians tend to be sceptical of the benefits of unregulated free trade and free capital movements (Blecker, 2003). Many regard floating exchange rates as a major source of macroeconomic instability, urging instead a reconsideration of Keynes's ambitious plans for an International Clearing Union (Davidson, 2002; Milberg, 2003; Vernengo, 2003). For the long period they focus on demand-side policies for economic growth and criticize the deflationary bias of the structural adjustment programmes imposed on developing countries by the international financial institutions. As already noted, there is no specifically Post Keynesian welfare economics, so that there is also no genuinely distinctive position on most microeconomic issues (including environmental questions, industry policy, labour market regulation and antitrust). However, their social democratic sympathies do tend to bring Post Keynesians close to the institutionalist or ‘left neoclassical’ position on many of these questions.

Assessment and prospects

In the 1970s many Post Keynesians believed that mainstream economics was in a state of Kuhnian crisis, with the very real prospect of a paradigm shift in their favour. This confidence proved to be misplaced, and by the first decade of the 21st century Post Keynesian economics had been thoroughly marginalized. This had something to be with the sociology of the profession, which displayed an increasing intolerance of alternative perspectives and methods of research. Some Post Keynesians suspected that they had to share part of the blame, since there was some truth in the accusation that the Post Keynesian church had become too broad, with a message that was lacking in coherence. Part of the problem was that the mainstream itself had changed, with the New Keynesians adopting some Post Keynesian positions (for example on endogenous money and the consequent rejection of the LM schedule) while rejecting many others. New Keynesian economics offered a rather less clear target for Post Keynesian criticism than the neoclassical synthesis, or monetarism, or New Classical economics. Subsequent developments in ‘behavioural macroeconomics’ and ‘post-Walrasian theory’ confused the picture still further, making the oppositional character of Post Keynesian economics less easy to define. The future relationship between Post Keynesianism and heterodox economics more generally was also unclear, with familiar fault lines developing here, too, between sectarians and pluralists. For all that, Thirlwall's (1993) six propositions with which this entry began are clear enough, and important enough, to provide a secure intellectual future for this unusually persistent and perceptive dissenting minority.

See Also

- functional finance
- heterodox economics
- Keynesianism
- macroeconomics, origins and history of
- money
- unemployment

Bibliography

Blecker, R.A. 2003. International economics. In *The Elgar Companion to Post Keynesian Economics*, ed. J. King. Cheltenham: Elgar.

Cornwall, J. and Cornwall, W. 2003. *Capitalist Development in the Twentieth Century: an Evolutionary Keynesian Analysis*. Cambridge: Cambridge University Press.

Davidson, P. 1999. Keynes’ principle of effective demand versus the bedlam of the New Keynesians. *Journal of Post Keynesian Economics* 21, 571–88.

Davidson, P. 2002. *Financial Markets, Money, and the Real World*. Cheltenham: Elgar.

Dow, S. 1996. *The Methodology of Macroeconomic Thought*. Cheltenham: Elgar.

Eichner, A. and Kregel, J. 1975. An essay on Post Keynesian theory: a new paradigm in economics. *Journal of Economic Literature* 13, 1293–314.

Harcourt, G. 1987. Post–Keynesian economics. In *The New Palgrave: A Dictionary of Economics*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Hamouda, O. and Harcourt, G. 1988. Post Keynesianism: from criticism to coherence. *Bulletin of Economic Research* 40, 1–33.

Holt, R. and Pressman, S., eds. 2001. *A New Guide to Post Keynesian Economics*. London: Routledge.

Kaldor, N. 1970. The new monetarism. *Lloyd's Bank Review* 97(July), 1–18.

Kalecki, M. 1990. *Collected Works of Michal Kalecki. Volume I: Capitalism: Business Cycles and Full Employment*. Oxford: Clarendon Press.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

King, J. 2002. *A History of Post Keynesian Economics Since 1936*. Cheltenham: Elgar.

King, J., ed. 2003. *The Elgar Companion to Post Keynesian Economics*. Cheltenham: Elgar.

Lavoie, M. 1992. *Foundations of Post Keynesian Economic Analysis*. Aldershot: Elgar.

Lawson, T. 2003. *Reorienting Economics*. London: Routledge.

Lee, F. 1998. *Post Keynesian Price Theory*. Cambridge: Cambridge University Press.

Lee, F. 2000. On the genesis of Post Keynesian economics: Alfred S. Eichner, Joan Robinson and the founding of Post Keynesian economics. *Research in the History of Economic Thought and Methodology* 18c, 1–258.

Mata, T. 2004. Constructing identity: the Post Keynesians and the capital controversies. *Journal of the History of Economic Thought* 20, 241–60.

Milberg, W. 2003. Globalization. In *The Elgar Companion to Post Keynesian Economics*, ed. J. King. Cheltenham: Elgar.

Minsky, H. 1986. *Stabilizing an Unstable Economy*. New Haven, CT: Yale University Press.

O'Donnell, R. 1989. *Keynes: Philosophy, Politics and Economics*. London: Macmillan.

Palley, T. 1996. *Post Keynesian Economics*. London: Macmillan.

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Robinson, J. 1972. The second crisis of economic theory. *American Economic Review* 62(2), 1–10.

Setterfield, M., ed. 2002. *The Economics of Demand-led Growth*. Cheltenham: Elgar.

Thirlwall, A. 1993. The renaissance of Keynesian economics. *Banca Nazionale del Lavoro Quarterly Review* 186, 327–37.

Vernengo, M. 2003. Bretton Woods. In *The Elgar Companion to Post Keynesian Economics*, ed. J. King. Cheltenham: Elgar.

Wallich, H. and Weintraub, S. 1971. A tax-based incomes policy. *Journal of Economic Issues* 5, 1–19.

Winnett, A. 2003. Environmental economics. In *The Elgar Companion to Post Keynesian Economics*, ed. J. King. Cheltenham: Elgar.

How to cite this article

King, J. E. "Post Keynesian economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000135> doi:10.1057/9780230226203.1314

The New Palgrave Dictionary of Economics Online

Postan, Michael Moïssey (1899 – 1981)

Robert Brenner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic demography; feudalism; peasants; Postan, M. M.; serfdom

Article

M.M. Postan, who was born in Tighina, Bessarabia, in 1899 and who died in Cambridge in 1981, was one of most distinguished economic historians of the 20th century. After briefly studying natural sciences and sociology at the University of St Petersburg he moved on to study law and economics at the University of Odessa, and then at the University of Kiev. He came to England in 1920 and between 1921 and 1926 took his first degree, his MA, and his Ph.D. at the London School of Economics. Between 1927 and 1937 he held lectureships, successively, at University College, London, at the London School of Economics, and at Cambridge University. In 1938 he was appointed to succeed Sir John Clapham in the chair of economic history at Cambridge, a position he retained until his retirement.

A specialist in medieval economic history, Postan originally made his reputation during the late 1920s and early 1930s on the basis of his studies on medieval trade and finance. His joint volume, with Eileen Power, Studies in English Trade in the Fifteenth Century (1933), became a standard work. He also published such seminal articles as ‘Credit in Medieval Trade’ (1928) and ‘Recent Trends in the Accumulation of Capital’ (1935).

From the later 1930s, Postan began to present his own distinctive interpretation of long-term trends in the medieval economy. In ‘The Chronology of Labour Services’ (1937) and ‘The Rise of the Money Economy’ (1944) Postan advanced devastating critiques of the hitherto-dominant unilineal evolutionist interpretation of pre-industrial European economic development, as advanced by distinguished medievalists such as Henri Pirenne. According to that view, it was the more or less steady expansion of commerce which drove the European economy forward, leading first to the decline of serfdom, next to the differentiation of the peasantry and the rise of agrarian capitalism, and ultimately to the growth of manufacturing and modern industry. Postan showed, in contrast, that trade, in itself, could as easily lead to the strengthening of the old, pre-capitalist forms as to their dissolution. He illustrated his point with a detailed discussion of the fluctuations of labour services in medieval England, showing that they rose

and fell in direct proportion to lordly demands for labour and, in particular, that demesne production grew, serfdom was intensified and labour services increased in precisely those areas of the country which were most commercialized and most exposed to the London market. Postan also pointed out that the spectacular rise of serfdom in late medieval and early modern Europe took place in large part in response to the rise of the international grain market.

In his 1950 report to the Ninth Congress of Historical Sciences, Postan put forward the initial version of his own population-centred interpretation of medieval economic history as an alternative to the trade-centred interpretation. In this and later work, Postan demonstrated that the pre-industrial economy of Europe was marked by a succession of long cycles of demographically driven expansions and contractions, following a basically Malthusian dynamic. He then went on to argue, in Ricardian fashion, that during the up phase of these cycles declining returns in agriculture (declining productivity) determined rising rents, falling wages, and terms of trade running in favour of agricultural and against industrial goods, while in the down phase, rising returns in agriculture determined just the opposite trends. Postan's interpretation followed lines which had begun to be sketched by the German demographic historian Wilhelm Abel and it influenced, in turn, the work of the French agrarian historian of the early modern period, Emmanuel Le Roy Ladurie. By the later 1950s, Postan's demographic view already had been so widely accepted as the key to the interpretation of pre-industrial economic change, that H.J. Habakkuk could reasonably conclude, in a synthetic essay on ‘The Economic History of Modern Britain’ for the *Journal of Economic History* in 1958, that

> For those who care for the overmastering pattern, the elements are evidently there for a heroically simplified version of English history before the nineteenth century in which the long-term movements in prices, in income distribution, in real wages, and in migration are dominated by changes in the growth of population.

Postan further developed his interpretation in a long series of specialized studies on all aspects of the medieval economy – agricultural technique, agricultural investment, the legal status of the peasantry, and so on – as well as in a number of major syntheses. In all these works, he remained guided by the conviction that the best results would come by linking, as closely as possible, generalizations derived from economic theory with the results of exhaustive primary research.

See Also

- feudalism

Selected works

1928. Credit in medieval trade. *Economic History Review* 1, 234–61.

1933. (With E. Power.) *Studies in English Trade in the Fifteenth Century*. London: G. Routledge & Sons.

1935. Recent trends in the accumulation of capital. *Economic History Review* 6(1), 1–12.

1937. The chronology of labour services. *Transactions of the Royal Historical Society, 4th series* 20, 169–93.

1939. *The Historical Method in Social Science*. London: London University Press.

1944. The rise of the money economy. *Economic History Review* 14.

1954. *The Famulus: The Estate Labourer in the XIIth and XIIIth Centuries*. London and New York: Cambridge University Press.

1966. Medieval agrarian society in its prime: England. In *The Cambridge Economic History of Europe*, vol. 1, *The Agrarian Life of the Middle Ages*, 2nd edn., ed. M.M. Postan and H.J. Habakkuk. Cambridge: Cambridge University Press.

1971. *Facts and Relevance: Essays on Historical Method*. London: Cambridge University Press.

1972. *The Medieval Economy and Society: An Economic History of Britain in the Middle Ages*. London: Weidenfeld & Nicolson.

1973a. *Medieval Trade and Finance: Collected Essay*. Cambridge: Cambridge University Press.

1973b. *Essays on Medieval Agriculture and General Problems of the Medieval Economy*. London: Cambridge University Press.

Bibliography

Habakkuk, H.J. 1958. The economic history of modern Britain. *Journal of Economic History* 18, 486–501.

How to cite this article

Brenner, Robert. "Postan, Michael Moïssey (1899–1981)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000134> doi:10.1057/9780230226203.1315

The New Palgrave Dictionary of Economics Online

Postlethwayt, Malachy (1707 – 1767)

William Darity, Jr.

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Postlethwayt, Malachy; slavery

Article

Malachy Postlethwayt gave vent to the most comprehensive expression of mercantilist thought on behalf of British imperial interests. Fay (1934, p. 3) justifiably called Postlethwayt, alongside Joshua Gee, a major ‘spokesman’ for 18th-century England. Postlethwayt's mercantilist vision emphasized (1) the slave trade to Africa and slavery in the Caribbean as vital stimuli to development of British manufactures; (2) the Royal African Company as an instrument of management of ‘the African trade’; (3) the necessity of competition with France for control of the slave trade; and (4) the general principle that government must promote trade and industry.

His monumental *Universal Dictionary of Trade and Commerce*, 20 years in the making before its first edition was published in instalments over the interval 1751–55, included an entry entitled ‘Africa’, summarizing his views on the relationship between African slavery and British industry. Despite acknowledging the brutality of the trade and allusion to some future date when a ‘Christian spirit’ might be moved to end the trade, Postlethwayt was wholly pragmatic. After all, he concluded, the gains for Britain from the slave trade were substantial – being a ‘trade (that) is … all profit’ and a trade that ‘occasionally gives so prodigious employment to our people both by sea and land’.

This perspective resonated throughout Postlethwayt's pamphlets (see his *Selected Works*). Sir James Steuart may have been the ‘last’ British mercantilist, but he certainly was not the purest. For that we must turn to Postlethwayt, whose vision was undiluted by vestiges of humanitarism.

Although foreign trade, with the slave trade as a key component, was Britain's engine of growth for Postlethwayt, there was great breadth in the matters he viewed as relevant to British economic development. Scientific and technical advances, maintenance of low or zero interest rates (see Viner, 1937, p. 47), sport and leisure (Dorfman, 1971, p. 7), the public debt (Johnson, 1937, pp. 190–5), agricultural policy (Johnson, 1937, pp. 196–201), maintenance of low wages, and development of securities markets were among the many factors he identified as influences on the rate of economic

expansion. Nevertheless, the overseas ‘plantations’ or ‘colonies’ lay at the heart of Postlethwayt's mercantile system, and, for Postlethwayt, full development of the plantations required slaves. Indeed, Postlethwayt's writings provided compelling evidence for Eric Williams's view in *Capitalism and Slavery* (1944) that British mercantile strategists were aware of slave-trading and slavery's ramifications as a spur to British industrialization.

Postlethwayt's *Universal Dictionary* (4th edn, 1774) purported to be a translation of Jacques Savary's *Dictionnaire universal du commerce*, but as Schumpeter (1954, pp. 156–7) noted, it was really much more. Nevertheless Schumpeter (p. 372, n.15) viewed Postlethwayt as a writer whose name survived despite ‘substandard performance’. Schumpeter added that E.A.J. Johnson's careful bibliographic efforts ‘reduced to its proper proportions the charge of plagiarism that has been frequently leveled against Postlethwayt, though the case remains bad enough’ (Schumpeter, 1954, pp. 156–7). But Johnson himself concluded that his efforts ‘relieve[d] Postlethwayt, at least partially, from an ill-founded charge’ (Johnson, 1937, p. 405). Nonetheless, substantial portions of Richard Cantillon's *Essai* first appeared in English in Postlethwayt's *Dictionary* (Higgs, 1905, pp. ix–xiii) without acknowledgement. Postlethwayt apparently sought, with mixed results, to become a well-heeled sycophant to British royalty through his work (Johnson, 1937, pp. 186–7). Johnson even speculated that Postlethwayt may have been a paid agent of the Royal African Company. He died abruptly in relative poverty in 1767 and was buried in Old Street churchyard in the Clerkenwell section of London. It is probable that he was the brother of James Postlethwayt, author of a major history of British public revenue.

Selected works

1757a. *Britain's Commercial Interest Explained Improved*. 2 vols, New York: Augustus M. Kelley, 1968.

1757b. *Great Britain's True System*. New York: Augustus M. Kelley, 1967.

1774. *The Universal Dictionary of Trade and Commerce*. 4th edn, New York: Augustus M. Kelley, 1971.

1968. *Selected Works*. Vol. 1, *1745–1757;* Vol. 2, *1746–1759*. Farnborough, Hants: Gregg International Publishers.

Bibliography

Dorfman, J. 1971. Postlethwayt's pioneer British Commercial Dictionary. Preface to M. Postlethwayt, *The Universal Dictionary of Trade and Commerce*, New York: Augustus M. Kelley.

Fay, C.R. 1934. *Imperial Economy and its Place in the Formation of Economic Doctrine 1600–1932*. Oxford: Clarendon Press.

Higgs, H. 1905. Preface to W.S. Jevons, *The Principles of Economics: A Fragment of Treatise on the*

Industrial Mechanism of Society and Other Papers. London: Macmillan.

Johnson, E.A.J. 1937. *Predecessors of Adam Smith: The Growth of British Economic Thought*. New York: Prentice-Hall.

‘Malachy Postlethwayt.’ In *Dictionary of National Biography*, ed. L. Stephen and S. Lee, London: Oxford University Press, Vol. 16. Reprinted 1949–50.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Viner, J.A. 1937. *Studies in the Theory of International Trade*. New York: Harper and Brothers.

Williams, E. 1944. *Capitalism and Slavery*. Chapel Hill: University of North Carolina Press.

How to cite this article

Darity, William, Jr. "Postlethwayt, Malachy (1707–1767)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000292> doi:10.1057/9780230226203.1864

The New Palgrave Dictionary of Economics Online

postmodernism

M. Klaes

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Postmodernism resists encyclopaedic definition. On the level of economic phenomena, debates have centred on postmodernity as a separate historiographic period. On the conceptual level, the work of prominent economists has been argued to resonate with postmodernist themes. Certain parts of behavioural and experimental economics have begun to display key postmodernist features. A small self-consciously postmodernist literature draws from economics, literary criticism, and Continental philosophical traditions in its analysis of economic phenomena.

Keywords

animal spirits; choice theory; computational functionalism; critical rationalism; economic modernism; Enlightenment; experimental economics; Keynes, J. M.; postmodernism; rhetoric of economics; Becker, G. S.; rhetoric of economics; methodology of economics; relativism

Article

Postmodernism is a concept that escapes encyclopedic definition, to the extent that mischievous commentators have described postmodernists as a club of individuals who tacitly collude in a refusal to collectively define what postmodernism is about. This should strike a chord with any economists: who have been accused of leaving central notions such as market, firm, competition or equilibrium ill-defined (for example, Coase, 1937; Clower, 1995), despite having good grounds for doing so (compare Popper, 1945: p. 18).

On the level of economic phenomena, debates have centred on whether or not one can consistently speak of postmodernity as a separate historiographic period. Advocates of postmodernity in this epochal sense assume that profound changes in the constitution of contemporary society have brought an end to the modern period, the close of which has variously been located from the last quarter of the 19th century to the last quarter of the 20th century. On the conceptual level, it has been argued that the work of several

prominent economists, including Keynes (1936) and Becker (1976) for example, resonates with postmodernist themes. Broader strands of research in economics have begun to display key postmodernist features, most notably as a result of critical examination of the notion of the rationally unified individual. A small, self-consciously postmodernist literature draws from economics, literary criticism, and Continental philosophical traditions in its analysis of economic phenomena.

Postmodernity

The postmodern found its initial motivation in postmodernity viewed as a historiographic category, commonly attributed to Arnold Toynbee (for example 1954, pp. 234–8). Toynbee suggested that the modern period in Western history, as the period immediately following the Middle Ages, had come to an end by the 1870s. He associated modernity with social stability, Enlightenment rationalism and progress. A ‘post-Modern’ period in turn was characterized by social unrest and the collapse of rationalism. This cultural pessimism in regard to the advent of the postmodern propagated by Toynbee and others contrasts with positive assessments of the move from industrialization to a post-industrial knowledge economy, where new technologies would replace ideology as key drivers of social change (for example, Toffler, 1970). Both the culturally pessimistic and optimistic views share the acceptance of postmodernity as a particular historical phase with a distinct set of postmodern or ‘late capitalist’ socio-economic features, a perspective which has found its apex in neo-Marxist stage theories of capitalist development (Mandel, 1975; Jameson, 1991).

The postmodern as a historiographic category rests on an epochal interpretation of history, which assumes that historic junctures separate adjacent periods. Many historians are not prepared to accept that modernity has been superseded by a qualitatively different period however. Interpreting the postmodern as ‘post modernity’ suffers here from the limitations that plague epochal categorization in general. To the extent that sceptics of postmodernity are not in fact sceptics regarding epochal categories and historiographies, they face a dilemma. They can either argue that modernity is the end of history (Fukuyama, 1992), or propose an alternative successor to it. But how can one conceive of such alternatives as anything else than a particular interpretation of ‘post modernity’?

Postmodern authors seek to avoid being trapped in binary oppositions of this kind. The work of Jean-François Lyotard, for example, has served as a prominent point of reference. His *Postmodern Condition* (1979) defined modernity in terms of a style of thought or epistemological outlook characterized by grand ‘meta-narratives’ centred on the ideas of scientific progress and individual emancipation, or the rationalist Enlightenment project *tout court*. Inverting these characteristics, Lyotard associated the postmodern with fragmented personal identities and a pervasive heterogeneity and indeterminacy of knowledge. But by doing so, he in fact affirmed the ahistorical dimensions of an ultimately bimodal categorization of contemporary society. The ‘postmodern’ turns thus into the less well-recognized face of the modern: ‘[p]ostmodernity is not a new age, but the rewriting of some of the features claimed by modernity … [T]hat rewriting has been at work, for a long time now, in modernity itself’ (Lyotard, 1987, p. 34).

Postmodernist economics

In contrast to other social sciences, the notion of a postmodernist kind of economics has only relatively recently begun to gain currency, both as a label for the work of several prominent economists, including Keynes (1936) and Becker (1976) for example, and in terms of methodological features displayed in several strands of current research. To speak of postmodernist economics along these lines requires a concept of economic modernism to begin with. Largely unrecognized, two different understandings of economic modernism have sprung up, with different implications for the understanding of postmodernism in economics.

Economic modernism has been understood either as the manifestation in economics of modernism more generally understood as a widely recognized 20th-century socio-cultural style, or as the methodological face of modernity epochally conceived (see above). As a socio-cultural style, modernism is commonly thought to have flourished in the early 20th century, although, depending on the particular context, its influence may be traced from the late 19th century to the first decade of the 21st century and quite likely beyond it (compare Weston, 1996). Across fields as diverse as literature, painting, music, architecture and design, proponents of modernism have questioned individual identity, displayed profound scepticism towards realist accounts of the world, and embraced dissonance and uncertainty as defining aspects of social life, developing ever more sophisticated forms of representation and a display of formal technique. What the many guises of modernism share is a profound reaction to the conditions of modernity.

In contrast to this avant-garde notion of modernism as the pursuit and transcendence of the limits of modernity, the concept of modernism first entered economics in a more restricted and conservative way, encapsulating a rejection of the methodological face of modernity. Economists by and large see themselves as adhering to the broad outlines of a critical rationalist methodology. This ‘official’ methodology of economics has been characterized by some methodologists as ‘modernist’, in the sense that it is committed to a belief in scientific progress through the formulation and empirical testing of hypotheses, to the rational actor paradigm, and to mathematical formalism (McCloskey, 1983; Dow, 1991).

Most economists will, of course, find these assumptions innocuous. It is thus no coincidence that economic modernism, in the sense described, is typically employed by authors who dissent from the 20th-century neoclassical tradition in economics. Samuelson's (1939) article on the multiplier–accelerator model, itself a central contribution to Keynesian business cycle theory, has been cited as a *pièce de résistance* in this regard, illustrating the modernist spirit underlying the neoclassical school (Klamer, 1995). The article covers barely four journal pages. Packed with mathematical notation, tables and graphs it keeps discursive elements to a minimum. Compared with Samuelson's treatment of the business cycle, Keynes's (1936) original analysis engages in an exuberance of narrative in his explanation of the business cycle, coming to a head in the well-known passages of Chapter 22 of the *General Theory*. Keynes's portrayal there of the uncertainties of the world of markets as being beyond the reach of rational analysis and containable only within a domain of ‘animal spirits’ (though channelled by social conventions) has prompted some authors to regard these aspects of his work as indicative of important postmodernist currents in 20th-century economics (in particular Ruccio and Amariglio, 2003, ch. 2), which reflect concerns comparable to the appreciation of the heterogeneity and indeterminacy of knowledge as it can be found in the work of Lyotard (1979), for example.

Reading the work of the mature Keynes as an expression of postmodern currents in the economics of the

1930s requires us to regard the neoclassical orthodoxy of that time as the prime manifestation of economic modernism, thereby allowing a rapprochement between dissenting schools of thought and a postmodern kind of economics. Not all who have identified postmodern aspects in economics have found compelling the association of economic postmodernism with dissenting, and of economic modernism with consenting, approaches vis-à-vis a putative mainstream, however. Characterizing Keynes's work as postmodernist can be challenged on historiographic grounds along the lines discussed above in the context of epochal interpretations of the postmodern. Moreover, this characterization rests on interpreting economic modernism as the methodological face of modernity. If modernism is instead understood as a broadly based early 20th-century socio-cultural style, there are good grounds for regarding the *General Theory* and other works of Keynes as a prime expression of economic modernism (Klaes, 2006).

Rather than depicting it as a caricature of orthodox approaches in economics, the appreciation of an economic modernism in its own right may help to account for a range of departures from the neoclassical tradition in early 20th-century economics, including both Keynes's G*eneral Theory* and the work of Samuelson and others who were at the forefront of the formalist turn in economics. Conversely, postmodernist dimensions in economics may be sought not only in dissenting approaches, a point most prominently expressed by Jameson (1991, pp. 263–71), who argues that Becker's (1976) work, in its treatment of children, companionship and health as conventional commodities, displays a deep affinity with the postmodernist notion of consumption as an all pervasive cultural pattern, sharing the ambition of reducing all human interaction to market exchange.

Decentred economic selves

As an illustration of how close contemporary theorizing in economics has come to key postmodernist concerns, let us consider how individuals are portrayed in economics. According to Davis (2003), there no longer exists a coherent account of the individual in contemporary economics following its de-psychologization and reduction to a rational preference ordering. With no concept of the individual beyond this ordering, choice theory has become equally applicable to individual persons and supra-person individuals like firms. Increasingly, however, economists also entertain the possibility of multiple sub-person objectives, with fascinating challenges to the notion of a unified economic self (for example, Schelling, 1984).

Ever since the publication of Berle and Means (1932), economists have been attuned to the split personality of multi-person individuals. Senior management follow their own objectives that do not necessarily coincide with the objective function of the corporation. A similar line of argument can be applied to the notion of a coherent self. At the sub-person level we may well consist of a range of competing selves. The de-psychologization of the individual in economics leads therefore to a postmodern critique of integrated individual identities.

Economists have begun exploring the implications of a decentred economic self (Kavka, 1991; Steedman and Krause, 1986), which rests on the proposition that the market without is matched by a market within. Warding off a postmodernist disintegration of the unified self amounts to solving the internal ‘social choice’ problem through the imposition of a dictator. To the extent that the internal and external worlds of choice are formally equivalent however, this literature has revealed a curious

asymmetry whereby the desirability of dictatorial solutions to the internal choice problem is taken for granted in the same vein as its undesirability regarding the external choice problem is taken for granted in economics.

Upon closer inspection, choice theory exhibits further postmodernist dimensions. Pursuing his basic argument from another angle, Davis (2003) suggests that economics, in its rejection of early neoclassical subjectivism, has subscribed to computational functionalism in its conception of the abstract individual. Computational functionalism, as a theory of mind, holds that brain states are computational states of mental algorithms, and that two individuals share the same type of mental state if they function in a causally equivalent way in respect of their physical environment. The abstract individual is therefore a preference computing algorithm, boundedly rational or not, that can be implemented in different entities without prejudice as to whether these entities are individual human beings, particular ‘modules’ within a human brain, economic institutions such as firms and markets, or non-humans (animals, machines and other ‘aliens’).

Mirowski (2002) has cast this ontological indifference of the economic individual in respect of its range of potential actualizations (human decision-makers, various subsets of brain tissue, animals, computers, and so forth) into the postmodern motif of the cyborg, a cyber organism that is half man and half machine. Recent work in experimental economics has unwittingly come up with an interesting illustration of this proximity between man and machine: on the level of convergence and efficiency, double auction behaviour of experimental subjects and computational agents programmed as random number generators turns out to be rather similar (Gode and Sunder, 1993). The cognitive capacities of market participants matter much less than the market algorithm itself. This allows the provocative suggestion that markets use us simply as pawns to further their algorithmic life, with the possibility of endogenous evolution of cyborg-like market automata in a decidedly post-humanist and thereby postmodernist fashion.

While postmodernist dimensions become apparent in a range of current strands of research in economics once they are read with an eye sensitive to debates in other social sciences and the humanities, self-consciously postmodernist work in economics has remained relegated to its fringes (see the collections by Woodmansee and Osteen, 1999; Cullenberg, Amariglio and Ruccio, 2001; and Zein-Elabdin and Charusheela, 2004). Its most influential impetus has come from the rhetoric of economics tradition. Initially concerned with the rhetoric found in the texts of academic economists, rhetoricians of economics have generalized their approach to include economic conversation more generally conceived (McCloskey, 1994, pp. 367–78). Prices are carriers of information only because they are part of a conversation. Entrepreneurs succeed only if they can persuade others to provide the capital necessary for turning their inventions into marketable products, which only sell if consumers can be persuaded to buy them. Stock markets epitomize this conversational feature of economic life (see Shiller, 1989, pp. 56, 387).

The resulting suggestion, of reading the economy as a text (Brown, 1994), leads to the prospect of a postmodernist economics that approaches the economy on the premise that all economic texts should be treated alike in this ongoing overarching conversation by which resources are allocated, be they authored by Nobel Laureates or the man or woman on the street. Opinion is divided even among rhetoricians regarding the merit of so radical a revision of the traditional hierarchy between economic analyst and agent (Mehta, 1999; McCloskey, 1999). It points to the implied relativism present in most postmodernist

perspectives as a major and recurring point of contention (see Backhouse, 1998), although relativism as such, though not popular among practitioners and methodologists alike, is not the unanimously discredited philosophical position that some make it out to be (Kusch, 2002).

See Also

- culture and economics
- economic history
- methodology of economics
- pluralism in economics
- rhetoric of economics

Bibliography

Backhouse, R.E. 1998. Should economists embrace postmodernism? In *Explorations in Economic Methodology. From Lakatos to Empirical Philosophy of Science*, ed. R.E. Backhouse. London: Routledge.

Becker, G.S. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Berle, A.A. and Means, G.C. 1932. *The Modern Corporation and Private Property*. New York: Legal Classics Library, 1993.

Brown, V. 1994. The economy as a text. In *New Directions in Economic Methodology*, ed. R.E. Backhouse. London: Routledge.

Clower, R.W. 1995. Axiomatics in economics. *Southern Economic Journal* 62, 307–19.

Coase, R.H. 1937. The nature of the firm. *Economica* NS 4, 386–405.

Cullenberg, S., Amariglio, J. and Ruccio, D.F. 2001. *Postmodernism, Economics and Knowledge*. London: Routledge.

Davis, J.B. 2003. *The Theory of the Individual in Economics*. London: Routledge.

Dow, S. 1991. Are there any signs of postmodernism with economics? *Methodus* 3, 81–5.

Fukuyama, F. 1992. *The End of History and the Last Man*. London: Hamilton.

Gode, D. and Sunder, S. 1993. Allocative efficiency of markets with zero-intelligence traders. *Journal of*

Political Economy 101, 119–37.

Jameson, F. 1991. *Postmodernism: Or, the Cultural Logic of Late Capitalism*. Durham, NC: Duke University Press.

Kavka, G.S. 1991. Is individual choice less problematic than collective choice? *Economics and Philosophy* 7, 143–65.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Klaes, M. 2006. Keynes between modernism and post modernism. In *The Cambridge Companion to Keynes*, ed. R.E. Backhouse and B.W. Bateman. Cambridge: Cambridge University Press.

Klamer, A. 1995. *The Conception of Modernism in Economics: Samuelson, Keynes and Harrod*. Aldershot: Elgar.

Kusch, M. 2002. *Knowledge by Agreement*. Oxford: Clarendon.

Lyotard, J.-F. 1979. *La Condition Postmoderne*. Paris: Éditions de Minuit.

Lyotard, J.-F. 1987. Rewriting modernity. In *The Inhuman: Reflections on Time*, ed. J.-F. Lyotard, trans. G. Bennington and R. Bowlby Cambridge: Polity, 1991.

Mandel, E. 1975. *Late Capitalism*. Rev. edn. trans. J. de Bres. London: Verso.

McCloskey, D.N. 1983. The rhetoric of economics. *Journal of Economic Literature* 21, 48–517.

McCloskey, D.N. 1994. *Knowledge and Persuasion in Economics*. Cambridge: Cambridge University Press.

McCloskey, D.N. 1999. Jack, David, and Judith looking at me looking at them. In *What Do Economists Do? New Economics of Knowledge*, ed. R.F.J. Garnett. London: Routledge.

Mehta, J. 1999. Look at me look at you. In *What Do Economists Do? New Economics of Knowledge*, ed. R.F.J. Garnett. London: Routledge.

Mirowski, P. 2002. *Machine Dreams. Economics Becomes a Cyborg Science*. Cambridge: Cambridge University Press.

Popper, K.R. 1945. *The Open Society and Its Enemies*. London: Routledge.

Ruccio, D.F. and Amariglio, J. 2003. *Postmodern Moments in Modern Economics*. Princeton, NJ: Princeton University Press.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics* 21, 75–8.

Schelling, T.C. 1984. Self-command in practice, in policy and in a theory of rational choice. *American Economic Review* 74, 1–11.

Shiller, R.J. 1989. *Market Volatility*. Cambridge, MA: MIT Press.

Steedman, I. and Krause, U. 1986. Goethe's *Faust*, Arrow's possibility theorem and the individual decision-taker. In *The Multiple Self*, ed. J. Elster. Cambridge: Cambridge University Press.

Toffler, A. 1970. *Future Shock*. New York: Random House.

Toynbee, A. 1954. *A Study of History*, vol. 9. New York: Oxford University Press.

Weston, R. 1996. *Modernism*. New York: Phaidon.

Woodmansee, M. and Osteen, M. 1999. *The New Economic Criticism. Studies at the Intersection of Literature and Economics*. London: Routledge.

Zein-Elabdin, E.O. and Charusheela, S. 2004. *Postcolonialism Meets Economics*. London: Routledge.

How to cite this article

Klaes, M. "postmodernism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000345> doi:10.1057/9780230226203.1316

The New Palgrave Dictionary of Economics Online

poverty alleviation programmes

Martin Ravallion

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the issues and evidence concerning a class of policies that aim to reduce poverty by providing direct current relief to those in need and/or by compensating for market and governmental failures that help perpetuate poverty. The article focuses on programmes found in developing countries. Poverty proxies or self-targeting mechanisms are typically used and the specific policies discussed include contingent transfers, community-based programmes, social funds and workfare programmes.

Keywords

affirmative action; agency costs; conditional transfers; deadweight losses; equity–efficiency trade-off; factor mobility; indicator-based targeting; intrahousehold welfare; learning-by-doing; poverty alleviation programmes; poverty proxies; poverty traps; programme evaluation; redistribution of income; workfare

Article

Rapid poverty reduction is widely seen to call for a combination of policies that on the one hand promote economic growth and on the other help poor people share in, and contribute to, the opportunities of a growing economy. There is wide agreement that the latter set of policies should include universal provision of adequate basic health care and schooling. There is less agreement on the scope for ‘poverty alleviation programmes,’ typically entailing transfers in cash or kind targeted to poor people. This article provides an overview of such programmes. First their objectives and the factors constraining their performance are discussed. Then the focus turns to the types of programmes found in developing countries.

Objectives and constraints

The generally agreed objective of this class of policies is to increase the standard of living of those with

low levels of living, that is, to reduce ‘absolute poverty’. While recognizing that high levels of inequality can impede prospects for reducing poverty, the objective of this class of policies is poverty reduction, not redistribution per se. Trade-offs underlie this objective. Inequality-reducing interventions can come at a cost to efficiency, such as through effects on the work effort or savings of beneficiaries. While these costs can be serious for specific programmes in specific contexts, it should not be presumed that there will necessarily be an equity–efficiency trade-off. In a world of market failures and ‘poverty traps’ direct interventions against poverty can also promote aggregate efficiency and (hence) growth. (On how there can be too much inequality and risk from the point of view of aggregate output see, *inter alia*, Bénabou, 1996; Aghion, Caroli and Garcia-Penalosa, 1999; and Bardhan, Bowles and Gintis, 2000. On poverty traps, see, *inter alia*, Dasgupta, 1993; Banerjee and Newman, 1994; and Hoff, 2001. Policy implications are examined in Ravallion, 2005a, and World Bank, 2001; 2006.) For example, credit constraints leave unexploited investment opportunities, notably for the poor (who have little or no collateral). Agency costs are probably also borne more heavily by the poor. (Agency costs arise when an agent, such as a worker or tenant farmer, makes key decisions relevant to a principal – the capitalist or land owner – who faces high supervision costs. Such models can generate efficient redistributions, from principal to agent; see Bowles and Gintis, 1996.)

There can be other trade-offs. The programmes that are best for reducing current poverty need not coincide with those that are best for reducing future poverty; examples will be given later. And the policies that are good for reducing chronic poverty (such as promoting the adoption of new farming technologies) may matter little to, or even exacerbate, transient poverty (by exposing poor farmers to greater downside risk).

There are a number of constraints in formulating effective anti-poverty programmes. Governmental budgets figure prominently. Interventions in the name of poor people that require less public spending on other things that matter to their welfare, or are financed in distortionary or inflationary ways that reduce growth, may well increase poverty. The political economy will also constrain the feasible set of anti-poverty policies. What is feasible in practice will of course depend on the specific context.

The scope for these policies is naturally constrained by the information available and administrative capabilities for acting on that information. Problems of information and incentives are at the heart of programme design. Addressing these problems can increase administrative costs, depleting the net resource transfer to the poor. Informational constraints are particularly relevant in the rural and urban informal sectors of developing countries, where policies such as a progressive income tax are seldom feasible (though such policies are themselves second-best responses to information constraints even in rich countries).

Programmes differ in the emphasis given to enhancing the assets of poor people as opposed to raising their current incomes. In principle, poverty-creating inefficiencies due to credit market failures or agency costs can be ameliorated by asset redistributions. However, governments face political-economy constraints on their ability to redistribute wealth. Certain asset-based interventions tend to be more feasible than others. Reducing inequalities of opportunity by improving the schooling and health of children from poor families is often politically easier than reducing inequalities in the ownership of non-human capital or land. And even when asset redistribution is feasible, state-contingent income transfers may also be needed to help address failures in the provision of private insurance. It is likely that anti-poverty policy will continue to call for a mix of efforts to redress inequalities of opportunity (probably emphasizing human resource development) and specific transfers in cash or kind.

Another issue is how finely targeted anti-poverty programmes should be. Policy discussions often call for better 'targeting' – a higher share of total spending going to the poor. However, the most finely targeted programme need not be the one with the greatest impact on poverty. Fine targeting can increase administrative costs, yield deadweight losses (as will be illustrated later) and undermine political support for the programme. (On the political economy of targeting see Gelbach and Pritchett, 2000. On deadweight losses see Ravallion and Datt, 1995. On administrative costs see Grosh, 1995. More general discussions of these issues can be found in Besley and Kanbur, 1993, and van de Walle, 1998.) Uncertainties about the measures used in practice to identify the poor can also lead one to question the benefits of fine targeting.

Reliable monitoring and *ex post* evaluation is crucial. Our knowledge about the performance of these programmes has traditionally been poor, but this is changing as more resources and better data and methods go into impact evaluations. (An impact evaluation measures impacts on outcomes relative to explicit counterfactuals. Ravallion, 2005b, reviews methods and results on the impact evaluation of this class of policies.) These have revealed both successes and failures, often depending crucially on the context; the same type of programme can achieve very different outcomes in different settings including at different scales of operation. (Theory and evidence indicating that targeting performance tends to improve as a programme expands can be found in Ravallion, 2005c.) Thus greater emphasis is now given to adapting programmes to their context – 'learning-by-doing' – as well as to broader reforms in governance and new, more pro-poor, institutions that can help assure better policy-making and implementation.

The following discussion briefly examines the main types of programmes found in practice, which will illustrate some of the generic points above. While 'targeting' per se is not the objective, existing programmes can usefully be classified according to the way they try to target the poor. The focus is on programmes that rely on transfers in cash or kind. (Lipton and Ravallion, 1995, review the full range of anti-poverty policies found in practice, which, in addition to transfers, include various forms of direct support to smallholders, better instruments for credit and insurance, tenancy reforms and titling programmes to enhance security of access to land, and removing biases against the poor in taxation, spending and regulatory – including migration – policies.)

Indicator targeting

The problems of observing incomes and the incentive effects of means-testing have led to various schemes that make transfers in cash or kind according to 'poverty proxies' such as living in a poor area, age (both children and the elderly) and rural landlessness. Everyone with the same value of the indicator (or some combination) is treated the same way. Tools exist for finding optimal allocations to minimize a poverty index based on poverty proxies and for measuring the impact on poverty (Ravallion, 1993). Naturally, the more information that is available, the better indicator targeting works. Significant advances have been made in our ability to exploit sample survey information for the purposes of informing policy-making. For example, reasonably reliable and quite detailed 'poverty maps' can now be formed by combining sample survey data with census data; see, for example, Elbers, Lanjouw and Lanjouw (2003).

Policy-makers have often been overly optimistic about how well they can reach the poor based on

readily observable indicators. Here there are some sobering lessons from empirical research. Even using comprehensive, high-quality household sample surveys we have rather modest ability to account for differences in the levels of measured consumption or incomes in terms of the sorts of readily observed covariates that are typically used for targeting. There appears to be sizable heterogeneity in living standards within target groups identified by poverty proxies. Further sources of targeting errors arise from the fact that one must base actual policies on data for the whole population (not just a sample survey) and that respondents will naturally face incentives to distort the data when it is known why it is being collected. Thus, one can expect (possibly large) errors in practice when using indicator targeting to fight poverty.

Performance in reaching the transiently poor through indicator-based targeting appears to be generally worse than performance in reaching the chronically poor (see Ravallion, van de Walle and Gautam, 1995; Lokshin and Ravallion, 2000; and van de Walle, 2004). This is not too surprising given that widely used poverty proxies have even less ability to explain changes over time in levels of living (with the use of panel data in which the same households are interviewed over time). And stakeholders naturally resist changes to a programme's allocations. Despite the potential in theory, targeted transfers in practice have not responded rapidly to changing household circumstances, as would be required for effective insurance.

Such observations have prompted efforts to find ‘smart policies’ that rely more on incentives in their design and can adapt more rapidly, and on institutional changes that can help assure that poor people are better represented in decision-making; examples are given below.

There are also reasons for thinking that the benefits of indicator targeting are sometimes underestimated. Policy discussions have typically viewed targeting in a static non-behavioural way; location or ethnicity are simply poverty proxies. Recent research has offered a new perspective, pointing to the potential for efficiency gains from targeting groups being locked out of economic opportunities by market or political mechanisms. For example, residential stratification in the presence of externalities can generate persistent inequality (Bénabou, 1993; Durlauf, 1996). There is evidence of ‘geographic poverty traps’ in underdeveloped rural economies, such that living in a poorly endowed area reduces prospects of escaping poverty at given individual (non-geographic) characteristics; see Jalan and Ravallion (2002), who use data for China. Poor-area development programmes in such a setting can thus secure long-term gains (Jalan and Ravallion, 1998). There is also evidence that the political economy can generate biases against specific groups defined by location or ethnicity and that affirmative actions favouring these groups can enhance the impact of an anti-poverty programme (Besley et al., 2004). Specific demographic groups (both children and the elderly) have also been targeted, given the evidence of strong demographic correlates of poverty found in household survey data, though the robustness of these empirical findings to measurement assumptions is questionable. (Allowing for scale economies in consumption can readily reverse the common finding that larger households tend to be poorer based on consumption or income per person; Lanjouw and Ravallion, 1995.) Here too there can be efficiency gains. South Africa has a pension scheme that gives cash transfers to the elderly; Duflo (2003) reports that these pensions have positive external benefits for child health within recipient families. The upshot of these findings is that targeting certain groups can have a greater long-term impact on poverty than suggested by a purely statistical poverty profile.

Finding that transfers based on indicators of current poverty can bring long-term benefits (given factor

market imperfections) does not, however, mean that they are the best policy option for this purpose. Policies to increase factor mobility can also have a role. Incentives to attract private capital into poorly endowed areas or to encourage labour migration out of them could well be more poverty reducing than transfers targeted to those areas. There has been very little work on these policy choices, and one often hears unsubstantiated claims by advocates.

Securing the efficiency gains from targeted transfers will often require complementary programmes or reforms. This has been emphasized in the context of redistributive land reforms, where impediments in access to credit and technologies can greatly attenuate the efficiency gains (Binswanger, Deininger and Feder, 1995; World Bank, 2003). Recognition of the need to combine transfers (of specific assets or incomes) with other initiatives to help foster the productivity of the poor has prompted recent interest in a class of conditional transfers that we now turn to.

Conditional transfers

Many anti-poverty programmes impose conditions on recipients that attempt to change their behaviour. The (more or less explicit) rationale is that some form of market failure has entailed that current behaviours are not socially optimal. (On the efficiency arguments for conditionality requirements on transfer schemes see Das, Do and Ozler, 2004.) In the 1990s, a number of new conditional transfer programmes emerged that required recipients to satisfy schooling (and sometimes child health-care) requirements. An example is Bangladesh's Food-for-Education (FFE) programme, which relies on community-based targeting of food transfers that aim to create an incentive for reducing the costs to the poor of market failures. Other examples are PROGRESA (renamed Oportunidades) in Mexico and Bolsa Escola in Brazil; in these programmes cash transfers are targeted to certain demographic groups in poor areas, conditional on regular school attendance and visits to health centres.

If one was concerned solely with current income gains to participants then one would not want to make transfers conditional on school attendance, which imposes a cost on poor families by inducing them to withdraw children from the labour force, thus reducing the (net) income gain to the poor from the programme. Rather, this type of programme is aiming to balance a current poverty reduction objective against an objective of reducing future poverty. Given the credit market failure, the incentive effect on labour supply of the programme (often seen as an adverse outcome of transfers) is now judged to be a benefit – to the extent that a well-targeted transfer allows poor families to keep the kids in school rather than sending them to work. Notice too that concerns about distribution *within* the household often underlie the motivation for such programmes; the programme conditionality makes it likely that relatively more of the gains accrue to children. This can also be interpreted as a policy response to the deficiency of traditional poverty proxies in reflecting distribution within the household.

There is evidence of significant gains from Bangladesh's FFE programme in terms of school attendance, with only modest income forgone through displaced child labour (Ravallion and Wodon, 2000). The programme was able to appreciably increase schooling, at modest cost to the current incomes of poor families. Mexico's PROGRESA programme has also been found to increase schooling, though the gains appear to be lower than for FFE (Behrman, Sengupta and Todd, 2002; Schultz, 2004; Skoufias, 2005). This is probably because primary schooling rates are higher in Mexico, implying less value-added over the (counterfactual) schooling levels that would obtain otherwise. Sadoulet and de Janvry (2002) argue

that there would be greater efficiency gains (through higher schooling) from PROGRESA if the programme had concentrated on children less likely to attend school in the absence of the programme, notably by focusing on the transition to secondary school. However, the policy choice will depend critically on the weight one attaches to current income gains for the poor as against future gains through schooling.

Community-based programmes

In recent times, community participation in programme design and implementation has been advocated as an institutional change that can help relieve informational constraints, and possibly tilt the balance of power toward the poor. A common form of this idea in practice is that the central government sets up a ‘social fund’ that provides financial support to a potentially wide range of community-based projects, with strong emphasis on local participation in proposing and implementing the specific projects. Community (governmental or non-governmental) organizations are assumed to be better informed about what is needed. The centre retains control over how much goes to each locality. A useful overview of what is known about this class of programs can be found in Mansuri and Rao (2004).

While ‘empowerment’ of the poor has motivated such community-based efforts, capture by local elites has been a continuing concern. Reliable generalizations are as yet elusive. There are reasons to expect heterogeneity across communities in the impacts of the same programme. Relevant sources of heterogeneity identified in the literature include local asset inequality (Bardhan and Mookerjee, 2000; Galasso and Ravallion, 2005) and the extent of interlinkage in local social networks (Spagnolo, 1999).

In the design of Bangladesh's FFE programme, economically backward areas were supposed to be chosen by the centre, leaving community groups – exploiting idiosyncratic local information – to select participants within those areas. Galasso and Ravallion (2005) use survey data to assess FFE incidence within and between villages. They found that targeting performance – measured by the difference between the realized per capita allocation to the poor and the non-poor – varied greatly between villages. Higher allocations from the centre to a village tended to yield better targeting performance, but there was no sign that poorer villages were any better or worse at targeting their poor.

The results also point to the role played by antecedent inequalities within villages in determining the relative power of the poor in local decision-making. Galasso and Ravallion found that more unequal villages are worse at targeting the poor – consistent with the view that greater land inequality comes with less power for the poor in village decision-making. (This echoes the view that inequalities can persist through their influence on the institutions that develop. For example, Engerman and Sokoloff, 2005, argue that this is why high initial inequality persisted in colonized countries. Also see World Bank, 2006.) This suggests a mechanism whereby inequality is perpetuated through the local political economy; the more unequal the initial distribution of assets, the better placed the non-poor will be to capture the benefits of external efforts to help the poor.

Self-targeting

The informational constraints on anti-poverty programmes have strengthened arguments for using self-targeting mechanisms. There are numerous ways to use incentives in programme design to assure self-

targeting of the poor. For example, the rationing of food or health subsidies by queuing can be self-targeting (Alderman, 1987), as can subsidizing inferior food staples or packaging in ways that are unappealing to the non-poor. However, the classic example of a self-targeted anti-poverty programme is a workfare programme, in which work requirements are imposed on welfare recipients with the aim of creating incentives to encourage participation only by the poor and to reduce dependency on the programme. (Besley and Coate, 1992, provide a formal model of the incentive arguments.)

An example is the famous Employment Guarantee Scheme (EGS) in Maharashtra, India. This aims to assure income support in rural areas by providing unskilled manual labour at low wages to anyone who wants it. The scheme is financed domestically, largely from taxes on the relatively well-off segments of Maharashtra's urban populations. The employment guarantee helps support the insurance function, and is also seen to help empower poor people. In practice, however, most workfare schemes have entailed some administrative rationing of the available work, often in combination with geographic targeting. Workfare schemes generally have a good record in screening the poor from the non-poor, and providing effective insurance against both covariate and idiosyncratic shocks. (For evidence on this point see Ravallion and Datt, 1995; Subbarao, 1997; Jalan and Ravallion, 2003; Coady, Grosh and Hoddinott, 2004.) They have provided protection when there is a threat of famine (Drèze and Sen, 1989) or in the wake of a macroeconomic crisis (see, for example, Pritchett, Sumarto and Suryahadi, 2003, for Indonesia's crisis in 1998, and Galasso and Ravallion, 2004, for Argentina's crisis in 2002). Design features are crucial, notably that the wage rate is not set too high. For example, Ravallion, Datt and Chaudhuri (1993) provide evidence on how the EGS responds to aggregate shocks, and on how its ability to insure the poor was jeopardized by a sharp increase in the wage rate. Low-wage workfare schemes have been advocated as a core element of a ‘permanent safety net’ for risk-prone economies (Ravallion, 2005c).

Self-targeted schemes face a trade-off between targeting performance (meaning their ability to concentrate benefits on the poor) and net income gains to participants, given that these programmes work by deliberately imposing costs on participants. Self-targeting requires that the cost of participation is higher for the non-poor than the poor (so that it is the poor who tend to participate), but it may not be inconsequential for the poor.

A potentially important cost to workfare participants is forgone income. This is unlikely to be zero; the poor can rarely afford to be idle. An estimate for two villages in Maharashtra, India, found that the forgone income from employment on the EGS was quite low – around one quarter of gross wage earnings; most of the time displaced was in domestic labour, leisure and unemployment (Datt and Ravallion, 1994). By contrast, for Argentina's Trabajar Program (a combination of workfare and social fund), it was estimated that about one half of gross wage earnings was taken up by forgone incomes (Jalan and Ravallion, 2003; Ravallion et al., 2005).

Workfare schemes also illustrate the potential trade-off in policy design between short-term income gains to the poor and longer-term gains through asset creation. Workfare programmes have not traditionally emphasized the value to the poor of the assets created, which appear often to mainly benefit the non-poor or to be of remarkably little value to anyone (see, for example, Gaiha, 1996, writing about Maharashtra's EGS.) The Trabajar Program illustrates the scope for a new wave of workfare programmes that emphasize asset creation in poor communities. The programme's design gave explicit incentives (through the *ex ante* project selection process) for targeting the asset creation to poor areas,

again compensating for the market failures that help create poor areas in the first place. There is typically much useful work to do in poor neighbourhoods – work that would probably not get financed otherwise.

The choice between the goal of raising current incomes of the poor and reducing future poverty will never be a straightforward. The choice will naturally depend on circumstances. For example, in macroeconomic or agro-climatic crises it is to be expected that the emphasis will shift to current income gains, away from asset creation – implying, for example, more labour-intensive sub-projects on workfare programmes.

See Also

- income taxation and optimal policies
- poverty lines
- social insurance

Bibliography

Aghion, P., Caroli, E. and Garcia-Penalosa, C. 1999. Inequality and economic growth: the perspectives of the new growth theories. *Journal of Economic Literature* 37, 1615–60.

Alderman, H. 1987. Allocation of goods through non-price mechanisms: evidence on distribution by willingness to wait. *Journal of Development Economics* 25, 105–24.

Banerjee, A. and Newman, A. 1994. Poverty, incentives and development. *American Economic Review, Papers and Proceedings* 84, 211–15.

Bardhan, P., Bowles, S. and Gintis, H. 2000. Wealth inequality, wealth constraints and economic performance. In *Handbook of Income Distribution*, vol. 1, ed. A. Atkinson and F. Bourguignon. Amsterdam: North-Holland.

Bardhan, P. and Mookherjee, D. 2000. Capture and Governance at local and national levels. *American Economic Review, Papers and Proceedings* 90, 135–9.

Behrman, J., Sengupta, P. and Todd, P. 2002. Progressing through PROGRESSA: an impact assessment of a school subsidy program. Mimeo. University of Pennsylvania.

Bénabou, R. 1993. Workings of a city: location, education and production. *Quarterly Journal of Economics* 108, 619–52.

Bénabou, R. 1996. Inequality and growth, In *National Bureau of Economic Research Macroeconomics Annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Besley, T. and Coate, S. 1992. Workfare vs. welfare: incentive arguments for work requirements in poverty alleviation programs. *American Economic Review* 82, 249–61.

Besley, T. and Kanbur, R. 1993. Principles of targeting. In *Including the Poor*, ed. M. Lipton and J. van der Gaag. Washington, DC: World Bank.

Besley, T., Pande, R., Rahman, L. and Rao, V. 2004. The politics of public good provision: evidence from Indian local governments. *Journal of the European Economic Association* 2, 416–26.

Binswanger, H., Deininger, K. and Feder, G. 1995. Power, distortions, revolt and reform in agricultural and land relations. In *Handbook of Development Economics*, vol. 3, ed. J. Behrman and T. Srinivasan. Amsterdam: North-Holland.

Bowles, S. and Gintis, H. 1996. Efficient redistribution: new rules for markets, states and communities. *Politics and Society* 24, 307–42.

Coady, D., Grosh, M. and Hoddinott, J. 2004. Targeting outcomes redux. *World Bank Research Observer* 19, 61–86.

Das, J., Do, Q.-T. and Ozler, B. 2004. A welfare analysis of conditional cash transfer schemes: implications for policy. *World Bank Research Observe* 20, 57–80.

Dasgupta, P. 1993. *An Inquiry into Well-Being and Destitution*. Oxford: Oxford University Press.

Datt, G. and Ravallion, M. 1994. Transfer benefits from public works employment: evidence from rural India. *Economic Journal* 104, 1346–69.

Drèze, J. and Sen, A. 1989. *Hunger and Public Action*. Oxford: Oxford University Press.

Duflo, E. 2003. Grandmothers and granddaughters: old age pension and intrahousehold allocation in South Africa. *World Bank Economic Review* 17, 1–26.

Durlauf, S. 1996. A theory of persistent income inequality. *Journal of Economic Growth* 1, 75–93.

Elbers, C., Lanjouw, P. and Lanjouw, J. 2003. Micro-level estimation of poverty and inequality. *Econometrica* 71, 355–64.

Engerman, S. and Sokoloff, K. 2005. Colonialism, inequality and long-run paths of development. Working Paper No. 11057. Cambridge, MA: NBER.

Gaiha, R. 1996. How dependent are the rural poor on employment guarantee scheme in India? *Journal of Development Studies* 32, 669–94.

Galasso, E. and Ravallion, M. 2004. Social protection in a crisis: Argentina's Plan Jefes y Jefas. *World Bank Economic Review* 18, 367–99.

Galasso, E. and Ravallion, M. 2005. Decentralized targeting of an anti-poverty program. *Journal of Public Economics* 85, 705–27.

Gelbach, J. and Pritchett, L. 2000. Indicator targeting in a political economy: leakier can be better. *Journal of Policy Reform* 4, 113–45.

Grosh, M. 1995. Toward quantifying the trade-off: administrative costs and incidence in targeted programs in Latin America. In *Public Spending and the Poor: Theory and Evidence*, ed. D. van de Walle and K. Nead. Baltimore, MD: Johns Hopkins University Press.

Hoff, K. 2001. Beyond Rosenstein-Rodan: the modern theory of coordination problems in development. *Proceedings of the Annual World Bank Conference on Development Economics 2000*. Washington, DC: World Bank.

Jalan, J. and Ravallion, M. 1998. Are there dynamic gains from a poor-area development program? *Journal of Public Economics* 67, 65–86.

Jalan, J. and Ravallion, M. 2002. Geographic poverty traps? A micro model of consumption growth in rural China? *Journal of Applied Econometrics* 17, 329–46.

Jalan, J. and Ravallion, M. 2003. Estimating the benefit incidence of an anti-poverty program by propensity-score matching. *Journal of Business and Economic Statistics* 21, 19–30.

Lanjouw, P. and Ravallion, M. 1995. Poverty and household size. *Economic Journal* 105, 1415–35.

Lipton, M. and Ravallion, M. 1995. Poverty and policy. In *Handbook of Development Economics*, vol. 3, ed. J. Behrman and T. Srinivasan. Amsterdam: North-Holland.

Lokshin, M. and Ravallion, M. 2000. Welfare impacts of Russia's 1998 financial crisis and the response of the public safety net. *Economics of Transition* 8, 269–95.

Mansuri, G. and Rao, V. 2004. Community-based and -driven development. *World Bank Research Observer* 19, 1–40.

Pritchett, L., Sumarto, S. and Suryahadi, A. 2003. Safety nets or safety ropes? Dynamic benefit

incidence of two crisis programs in Indonesia. *World Development* 31, 1257–77.

Ravallion, M. 1993. Poverty alleviation through regional targeting: a case study for Indonesia. In *The Economics of Rural Organization: Theory, Practice and Policy*, ed. K. Hoff, A. Braverman and J. Stiglitz. Oxford: Oxford University Press.

Ravallion, M. 2005a. Transfers and safety nets in poor countries: revisiting the trade-offs and policy options. In *Understanding Poverty*, ed. A. Banerjee, R. Bénabou and D. Mookerjee. New York: Oxford University Press.

Ravallion, M. 2005b. Evaluating anti-poverty programs. In *Handbook of Agricultural Economics*, vol. 4, ed. R. Evenson and T. Schultz. Amsterdam: North-Holland.

Ravallion, M. 2005c. Who is protected? On the incidence of fiscal adjustment. In *Macroeconomic Policies and Poverty Reduction*, ed. A. Mody and C. Pattillo. London: Routledge.

Ravallion, M. and Datt, G. 1995. Is targeting through a work requirement efficient? In *Public Spending and the Poor: Theory and Evidence*, ed. D. van de Walle and K. Nead. Baltimore, MD: Johns Hopkins Press.

Ravallion, M., Datt, G. and Chaudhuri, S. 1993. Does Maharashtra's Employment Guarantee Scheme guarantee employment? Effects of the 1988 wage increase. *EconomicDevelopment and Cultural Change* 41, 251–75.

Ravallion, M., van de Walle, D. and Gautam, M. 1995. Testing a social safety net. *Journal of Public Economics* 57, 175–99.

Ravallion, M. and Wodon, Q. 2000. Does child labor displace schooling? Evidence on behavioral responses to an enrolment subsidy. *Economic Journal* 110, C158–C176.

Ravallion, M., Galasso, E., Lazo, T. and Philipp, E. 2005. Do workfare participants recover quickly from retrenchment? *Journal of Human Resources* 40, 208–30.

Sadoulet, E. and de Janvry, A. 2002. Alternative targeting and calibration schemes for educational grants programs: lessons from PROGRESA. Mimeo. Berkeley: University of California.

Schultz, T. 2004. School subsidies for the poor: evaluating the Mexican PROGRESA poverty program. *Journal of Development Economics* 74, 199–250.

Skoufias, E. 2005. PROGRESA and its impact on the welfare of rural households in Mexico. Research Report No. 139. Washington, DC: International Food Research Institute.

Spagnolo, G. 1999. Social relations and cooperation in organizations. *Journal of Economic Behavior and Organization* 38, 1–25.

Subbarao, K. 1997. Public works as an anti-poverty program: an overview of cross-country experience. *American Journal of Agricultural Economics* 79, 678–83.

van de Walle, D. 1998. Targeting revisited. *World Bank Research Observer* 13, 231–48.

van de Walle, D. 2004. Testing Vietnam's public safety net. *Journal of Comparative Economics* 32, 661–79.

World Bank. 2001. *World Development Report: Attacking Poverty*. New York: Oxford University Press.

World Bank. 2003. *Land Policies for Growth and Poverty Reduction*. New York: Oxford University Press.

World Bank. 2004. *World Development Report: Delivering Public Services to the Poor*. New York: Oxford University Press.

World Bank. 2006. *World Development Report: Equity and Development*. New York: Oxford University Press.

How to cite this article

Ravallion, Martin. "poverty alleviation programmes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000313> doi:10.1057/9780230226203.1318

The New Palgrave Dictionary of Economics Online

poverty lines

Martin Ravallion
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The article provides welfare-economic definitions of poverty lines and critically assesses the main methods of setting poverty lines found in practice. These can be interpreted as ways of expanding the information set used in applied work to address some long-standing problems in measuring welfare. Objective methods draw on information from outside economics on the commodities needed for normative activity levels. Subjective methods extend the information base by drawing on self-reported perceptions of consumption adequacy, allowing estimation of an endogenous social subjective poverty line.

Keywords

consumers' expenditure; cost-of-basic-needs poverty measure; equivalent income functions; food-energy-intake poverty measure; objective poverty lines; poverty; poverty lines; relative income; relative poverty line; Rowntree, B.; self-assessed welfare; social subjective poverty line; subjective poverty lines; utility; welfare economics; well-being

Article

Knowing how many people live in households with income or consumption expenditure below the 'poverty line' has helped focus attention on the extent of poverty and has informed policymaking for fighting poverty. But how are poverty lines defined and calculated? This article first provides a theoretical definition, and then describes the main methods found in practice.

Poverty lines in theory

People in different circumstances – with different household sizes or demographic compositions or living in different places – naturally have different levels of economic welfare at the same level of income. They have different needs. A poverty line should reflect these differences. But how should that be done? To answer this question we must first define the conceptual ideal against which the methods found in practice are to be judged.

The poverty line for a given individual can be defined as the money the individual needs to achieve the minimum level of 'welfare' to not be deemed 'poor', given its circumstances. Everyone at the poverty line is taken to be equally badly off, and all those below the line are worse off than all above it. The next question is: what concept of 'welfare' should serve as the anchor for the poverty line? For economists the obvious answer is 'utility'. A justification for utility-consistent poverty lines can be found by applying standard welfare-economic principles to poverty measurement (Ravallion, 1994). These principles are that assessments of social welfare should depend solely on utilities, that people with the same initial utility should be treated the same way, and that social welfare should not be decreasing in any utility.

To formalize this definition, consider individual i with characteristics x_i (a vector). The interpersonally comparable utility function is $u(q_i, x_i)$. The quantity vector q_i is utility maximizing, giving demand functions $q(p_i, y_i, x_i)$ at total expenditure y_i and corresponding utility maximum, $v(p_i, y_i, x_i)$. The utility-consistent poverty line is the point on the consumer's expenditure function corresponding to a common reference utility level. (As always, the monetary measurement of welfare requires a reference utility level.) The

consumer's expenditure function is $e(p_i, x_i, u)$, giving the minimum cost of utility *u* when facing the price vector p_i. Let u_z denote the minimum utility deemed necessary to escape poverty. Consistency requires that this is a constant across all *i*. The money metric of u_z defines the *utility-consistent poverty lines*:

$$z_i^u = e(p_i, x_i, u_z) \text{ for all } i = 1, \ldots n \quad (1)$$

This is closely related to a number of other economic concepts. Eq. (1) is ‘money-metric utility’ at a specific reference utility, interpretable as the poverty line in utility space. (Textbook treatments of money-metric utility functions – sometimes called ‘equivalent income functions’ – can be found in Varian, 1978, and Deaton and Muellbauer, 1980.) The value of $z_i^u / e(p_r, x_r, u_z)$ for reference individual *r* gives the ‘true cost-of-living index’ (see, for example, Deaton and Muellbauer, 1980). The value of y_i / z_i^u is the ‘welfare ratio’ (Blackorby and Donaldson, 1987). On exploiting the properties of the expenditure function, eq. (1) can be written in a more instructive form:

$$z_i^u = p_i q(p_i, x_i, u_z) \quad (2)$$

Thus, the poverty line is the cost of a bundle of goods, namely, the vector of utility-compensated (Hicksian) demands, $q(p_i, x_i, u_z)$. This bundle can be interpreted as ‘basic consumption needs’.

Note that an absolute poverty line in terms of utility can have the properties of a *relative poverty line* in the income space, in that it rises with the mean income of a relevant reference group. This is possible if individual utility depends on both ‘own income’ and income relative to others in that reference group. In other words, the indirect utility function has the form, $v(p_i, y_i, y_i / \bar{y}_i^r, x_i)$ where $\bar{y}_i^r$ is the mean income of the reference group(s); the poverty line takes the form $z_i = z(p_i, \bar{y}_i^r, x_i, u_z)$ (which solves $u_z = v(p_i, z_i, z_i / \bar{y}_i^r, x_i)$). Thus, we can view poverty as absolute in the space of welfare, but relative in the space of commodities (paraphrasing Sen, 1983). However, the extreme case in which the poverty line is a constant proportion of the mean (as used by poverty measures derived by Eurostat, for the European Commission) requires the seemingly implausible assumption that utility depends *solely* on relative income, given by the ratio of own income to the mean. By this measure, if all incomes increase by the same multiplicative factor then the proportion of people living below the poverty line will be unchanged. Cross-country comparisons of poverty lines suggest that relative income is valued more highly as mean income rises; in the poorest countries, absolute income levels are the dominant consideration, so then the poverty lines tend to have a low elasticity to the mean (Ravallion, 1994; 1998). There is micro empirical evidence supporting this view, based on self-assessed welfare (Ravallion and Lokshin, 2005).

For economists, utility is the obvious anchor for setting poverty lines, but it is not the only approach. Functioning-based concepts of welfare offer an alternative foundation for poverty lines. The approach can be characterized in the terms of Sen's (1985) argument that ‘well-being’ should be thought of in terms of a person's capabilities, that is, the functionings (‘beings and doings’) that a person is able to achieve. On this view, poverty means not having an income sufficient to support specific normative functionings. Utility – as the attainment of personal satisfaction – can be viewed as one such functioning relevant to well-being (Sen, 1992, ch. 3). But it is possibly only one of the functionings that matter. Independently of utility, one might say that a person is better off if she is able to participate fully in social and economic activity.

From this starting point, a more general theoretical formalization of the definition of a poverty line can be proposed as follows. Let a person's functionings be determined by the goods she consumes and her characteristics, giving the vector of functionings:

$$f_i = f(q_i, x_i) \quad (3)$$

where *f* is a vector-valued function. One can postulate that a person derives utility directly from her functionings. We can then interpret $u(q_i, x_i)$ as a derived utility function, obtained by substituting (3) into the (primal) utility function defined over functionings.

Functioning consistency for a set of poverty lines requires that certain normative functionings are reached at the poverty line. Let f_z denote the vector of critical functionings needed to not be deemed poor. These are normative judgements, just as u_z is a normative judgment. Assume that there is a bundle q_i^c such that no functioning is below its critical value:

$$f_z \leq f(q_i^c, x_i) \tag{4}$$

This yields the poverty lines: $z_i^c = p_i q_i^c$. There can be multiple solutions for q_i^c. Two ways to pick a unique poverty line can be identified. The first is to define z_i^c as the minimum *y* such that (4) holds. Notice that one (or more) specific functionings will be decisive; that is, the functioning that is the last to reach its critical value as income rises. In this sense, the lowest priority functioning for the individual will be decisive. The alternative approach is to treat attainments as a random variable (that is, with a probability distribution) and take a mean conditional on income and other identified covariates, including group membership. Then poverty lines are deemed to be functioning consistent if f_z is reached in expectation.

Implementing these concepts empirically requires that we solve two problems. The first can be called the *referencing problem*: what is the reference level of utility that anchors the poverty line (u_z in eq. 1)? It is tempting to say this choice is arbitrary, and to hope that it is innocuous. But the choice of the reference is far from arbitrary, and (in general) it affects the resulting poverty measure. This speaks to the importance of testing the sensitivity of poverty comparisons (such as between groups or over time) to the choice of reference, as it determines the level of the poverty line. Tests exist for the robustness of ordinal comparisons using stochastic dominance criteria; on this approach see Atkinson (1987) and Ravallion (1994).

The second problem is the *identification problem*. Even if we can readily agree on what the poverty line is in welfare space, there is a further problem in identifying the expenditure function in eq. (1). Standard practice is to calibrate the parameters of the cost function from consumer demand behaviour. The problem is that individuals vary in characteristics, such as their size and demographic composition, which influence welfare in ways that may not be evident in consumer demand behaviour. Then there is a fundamental problem of identification (Pollak and Wales, 1979; Pollak, 1991).

Poverty lines in practice

The methods of setting poverty lines found in practice fall under two headings: objective poverty lines and subjective poverty lines.

Objective poverty lines

The main methods found in practice are the food-energy-intake (FEI) method and the cost-of-basic-needs (CBN) method. It is known that these methods give radically different results; using data for Indonesia, Ravallion and Bidani (1994) found virtually zero correlation between the regional poverty profiles (given the poverty rates across geographic areas) produced by these two methods. Since policy choices (such as in regional targeting) could depend critically on which method is used, it is important to probe carefully into the choice.

The FEI method can be interpreted as a special case of the functioning-based approach described above. The specialization is to focus on just one functioning, namely food-energy intake. The method finds the consumption expenditure or income level at which food-energy intake is just sufficient to meet predetermined food-energy requirements for good health and normal activity levels. (Such caloric requirements are given in WHO, 1985, for example.) To deal with the fact that food-energy intakes naturally vary at a given income level, the FEI method typically calculates an expected value of intake at given income. Figure 1 illustrates the method. The vertical axis is food-energy intake, plotted against income (or expenditure) on the horizontal axis. A line of ‘best fit’ is indicated; this is the expected value of caloric intake at given income (that is, the nonlinear regression function). By simply inverting this line, one finds the income *z* at which a person typically attains the stipulated food-energy requirement. This method, or something similar, has been used often, including by Dandekar and Rath (1971), Osmani (1982), Greer and Thorbecke (1986), and Paul (1989), and by numerous governmental statistics offices. It is often found in practice in developing countries.

Figure 1
The food-energy intake method of setting poverty lines

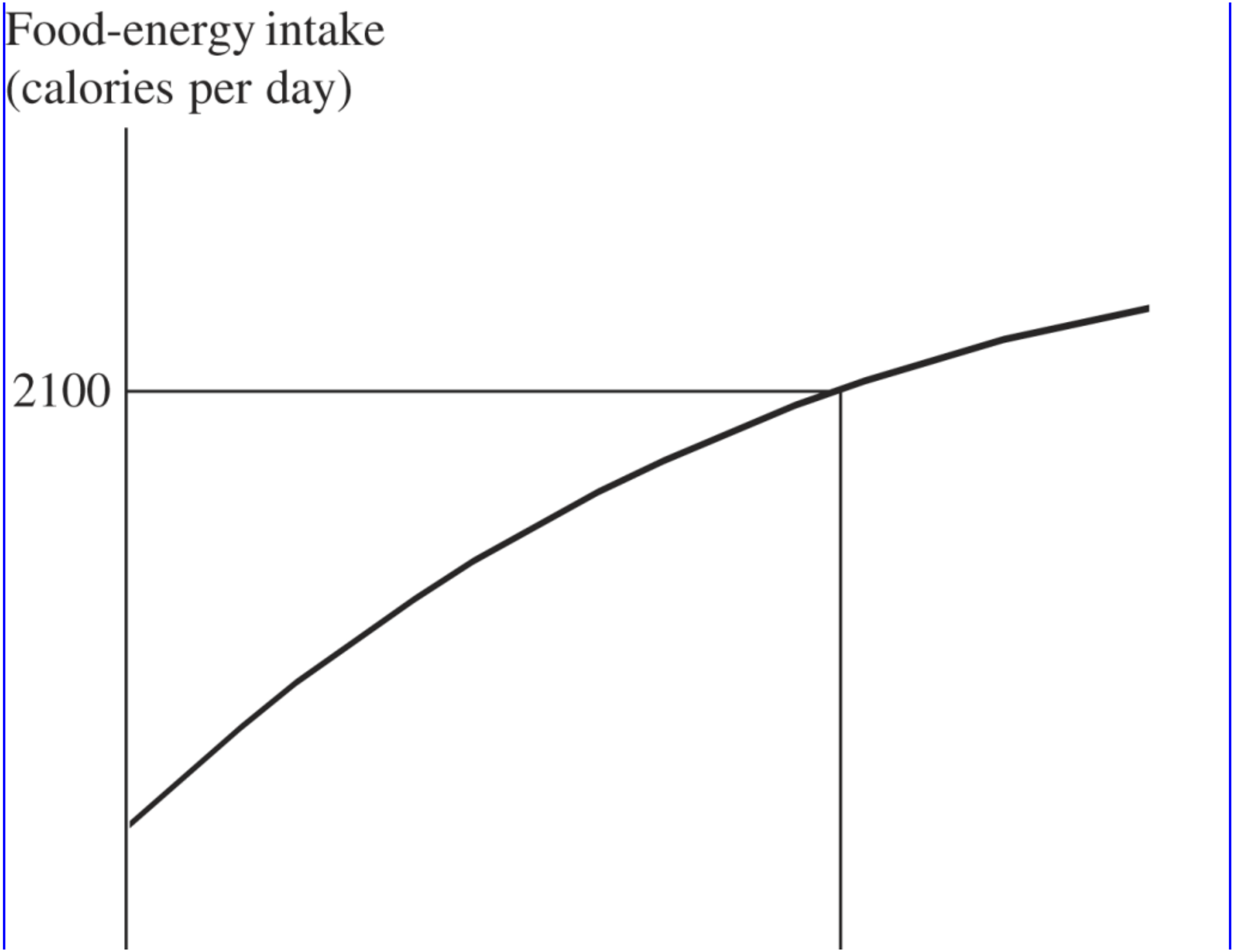
Food-energy intake
(calories per day)
2100

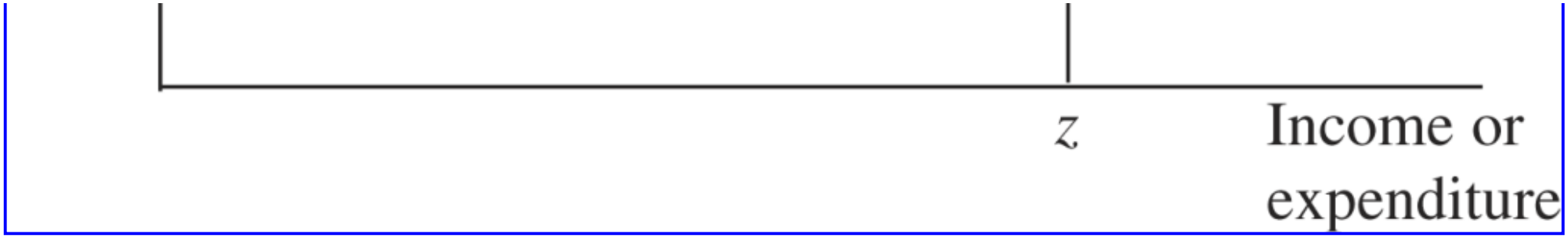

One concern about this method is that the resulting poverty lines need not be consistent in terms of utility or capabilities more generally (Ravallion, 1994; Ravallion and Lokshin, 2006). Consider first how FEI poverty lines respond to differences in relative prices, which can of course differ across the subgroups (such as regions) being compared in the poverty profile and over time. For example, the prices of many non-food goods relative to food are likely to be lower in urban than in rural areas. This will probably mean that the demand for food and (hence) food-energy intake will be lower in urban than in rural areas, at any given real income. But this does not, of course, mean that urban households are poorer. The relationship between food-energy intake and income will shift according to differences in tastes, activity levels and publicly provided goods. There is nothing in the FEI method to guarantee that these differences are ones that would normally be considered relevant to assessing welfare. Indeed, it is quite possible to find that the ‘richer’ sector (by the agreed metric of utility) tends to spend so much more on each calorie that it is deemed to be the ‘poorer’ sector. That has been found to be the case in studies of the properties of FEI poverty profiles for Indonesia (Ravallion and Bidani, 1994) and Bangladesh (Ravallion and Sen, 1996; Wodon, 1997).

Problems also arise in comparisons over time. Suppose that all prices increase, so the cost of a given utility must rise. There is nothing to guarantee that the FEI-based poverty line will increase. That will depend on how relative prices and tastes change; the price changes may well encourage people to consume cheaper calories, and so the FEI poverty line will fall. Wodon (1997) gives an example of this problem in data for Bangladesh. The FEI poverty line fell over time even though prices generally increased. The potential utility inconsistencies in FEI poverty lines are worrying when there is mobility across the subgroups of the poverty profile, such as due to inter-regional migration. For example, it is possible that a process of economic development through urban sector enlargement, in which none of the poor are any worse off and at least some are better off, would result in a measured *increase* in poverty.

The CBN method stipulates a consumption bundle deemed to be adequate for ‘basic consumption needs’, and then estimates its cost for each of the subgroups being compared in the poverty profile. This is the approach of Rowntree (1901) in his seminal study of poverty in York, England, in 1899, and there have been numerous examples since, including the official poverty lines for the United States (Orshansky, 1963; also see Citro and Michael, 1995). Some form of functioning consistency is assured by construction, since various valued functionings are essentially the starting point for defining ‘basic consumption needs’. The poverty bundle is typically anchored to food-energy requirements consistent with common diets in the specific context. However, allowances for non-food goods are also included, to assure that basic non-nutritional functionings are assured.

The CBN method is utility consistent *if* the right bundle is used, corresponding to the relevant points on the utility-compensated demand functions (eq. 2). However, there is nothing to guarantee that the bundles of goods built into CBN poverty lines lie on the compensated demand functions, at the (common) reference level of utility. Thus it is important to have some way of assessing a set of CBN poverty bundles. Ravallion and Lokshin (2006) propose an approach to testing the utility consistency of CBN poverty lines across households with common preferences using Samuelson's (1938) theory of revealed preference. However, this can be applied only within subgroups deemed to have common preferences. In practice utility functions can vary, due to differences in climate, for example.

In some cases a complete vector of normative (food and non-food) goods is set, as in Russia's poverty lines (Ravallion and Loskhin, 2006). However, it is more often the case that only food needs are set, based on nutritional requirements. To include an allowance for non-food needs, a common practice is to divide the food poverty line by some budget share for food. For example, the US poverty line assumes a food share of one third, so the total poverty line is three times the food line (Orshansky, 1963). However, the basis for setting a food share is rarely transparent. Why use the average share, as in the US line? Whose food share should be used?

Arguably, a more appealing approach is to set an allowance for non-food goods that is consistent with demand behaviour at (or in a region of) the food poverty line. Ravallion (1994) proposes two methods. The first divides the food component of the poverty line by the mean food share of households whose actual food spending is in a neighbourhood of the food poverty line. The second method uses mean non-food spending of households whose total sending is in a neighbourhood of the food poverty line. Ravallion argues the first method gives a reasonable upper bound to the allowance for non-food needs while the second gives a lower bound.

Subjective poverty lines

There is an inherent subjectivity and social specificity to any notion of ‘basic needs’, including nutritional requirements. Psychologists, sociologists and others have argued that the

circumstances of the individual relative to others influence perceptions of well-being at any given level of individual command over commodities. (Runciman, 1966, provided an influential exposition, and supportive evidence. Also see the discussions in Easterlin, 1995, and Oswald, 1997.) By this view, 'the dividing line … between necessities and luxuries turns out to be not objective and immutable, but socially determined and ever changing' (Scitovsky, 1978, p. 108).

Subjective poverty lines have been based on answers to the 'minimum income question' (MIQ), such as the following (paraphrased from Kapteyn, Kooreman and Willemse, 1988): 'What income level do you personally consider to be absolutely minimal? That is to say that with less you could not make ends meet.' (This can be thought of as a special case of Van Praag's, 1968, 'income evaluation question', which asks what income is considered 'very bad', 'bad', 'not good', not bad', 'good', 'very good'.) One might define as poor all whose actual income is less than the amount they give as an answer to this question. However, this would almost certainly lead to inconsistencies in the resulting poverty measures, in that people with the same income, or some other agreed measure of economic welfare, will be treated differently. Clearly an allowance must be made for heterogeneity, such that people at the same standard of living may well give different answers to the MIQ, but must be considered equally 'poor' for consistency. Past empirical work has found that the expected value of the answer to the MIQ conditional on actual income tends to be an increasing function of actual income. (Contributions include Groedhart et al., 1977; Danziger et al., 1984; and Kapteyn, Kooreman and Willemse, 1988.) Furthermore, past studies have tended to find a relationship such as that depicted in Figure 2, which gives a stylized representation of the regression function on income for answers to the MIQ. The point z^* in the figure is an obvious candidate for a poverty line; people with income above z^* tend to feel that their income is adequate, while those below z^* tend to feel that it is not. We can call z^* the 'social subjective poverty line' (SSPL).

Figure 2

The social subjective poverty line (z^*)

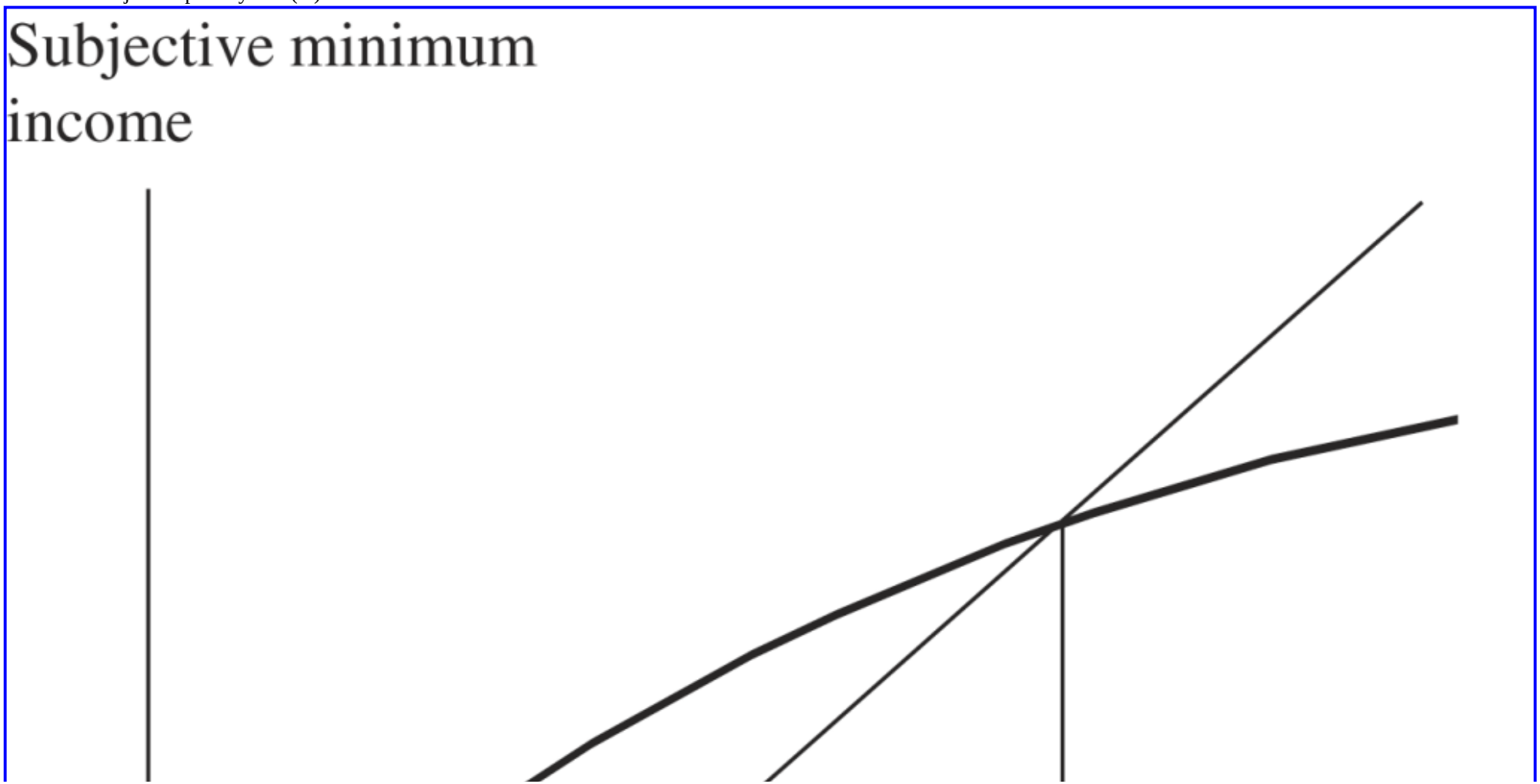

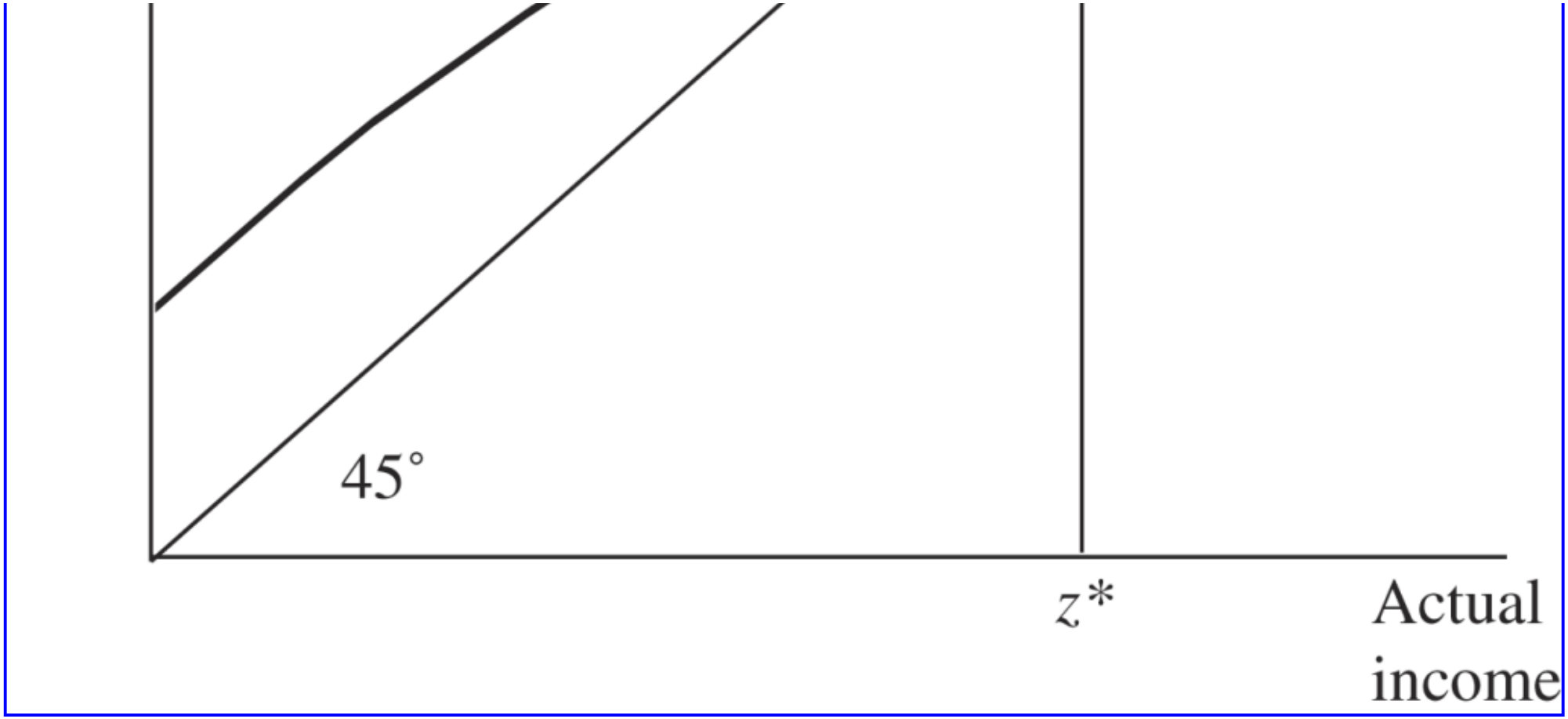

It is recognized in the literature that there are other determinants of economic welfare which should shift the SSPL, such as family size and demographic composition. Indeed, the answers to the MIQ are interpretable as points on the consumer's expenditure function at a point of minimum utility (eq. 1). Under this interpretation, subjective welfare assessments provide a means of overcoming the well-known problem of identifying utility from demand behavior alone when household attributes vary (Kapteyn, 1994).

While the MIQ has been applied in a number of OECD countries, there have been few attempts to apply it in a developing country. There are a number of potential pitfalls. ‘Income’ is not a well-defined concept in most developing countries, particularly (but not only) in rural areas. It is not at all clear whether one could get sensible answers to the MIQ. The qualitative idea of the ‘adequacy’ of consumption is a more promising one in a developing-country setting, and (arguably) many developed counties.

Pradhan and Ravallion (2000) propose a method for estimating the SSPL based on qualitative data on consumption adequacy, as given by responses to appropriate survey questions. Instead of asking respondents what the precise minimum consumption is that they need, one simply asks whether their current consumptions are adequate. This provides a multidimensional extension to the one-dimensional MIQ. The SSPL is the level of total spending above which respondents say (on average) that their expenditures are adequate for their needs. For empirical implementation, the probability that a sampled household will respond that its actual consumption of each type of commodity is adequate can be modelled as a probit regression. Under certain technical conditions, a unique solution for the subjective poverty line can then be obtained from the estimated parameters of the probit regressions for consumption adequacy. Pradhan and Ravallion provide empirical examples for Jamaica and Nepal; the SSPL gave a similar overall poverty rate to pre-existing objective poverty lines for both countries, though the structure of the poverty profile was different in some respects: for example, while the objective poverty lines implied that larger households tended to be poorer, this was not the case with the subjective approach.

Subjective data also offer a test of objective poverty lines, by regressing self-rated welfare on income normalized by the poverty line *plus* the variables that went into the construction of the poverty line, which should be jointly insignificant if those lines accord with subjective welfare. This approach is outlined in Ravallion and Lokshin (2002) and illustrated using Russia's poverty lines.

See Also

- consumer expenditure
- consumer expenditure (new developments and the state of research)
- inflation measurement
- poverty
- poverty alleviation programmes

Bibliography

Atkinson, A. 1987. On the measurement of poverty. *Econometrica* 55, 749–64.

Blackorby, C. and Donaldson, D. 1987. Welfare ratios and distributionally sensitive cost-benefit analysis. *Journal of Public Economics* 34, 265–90.

Browning, M. 1992. Children and household economic behavior. *Journal of Economic Literature* 30, 1434–75.

Citro, C. and Michael, R. 1995. *Measuring Poverty: A New Approach*. Washington, DC: National Academy Press.

Dandekar, V. and Rath, N. 1971. *Poverty in India*. Pune: Indian School of Political Economy.

Danziger, S., van der Gaag, J., Smolensky, E. and Taussig, M. 1984. The direct measurement of welfare levels: how much does it take to make ends meet. *Review of Economics and Statistics* 66, 500–5.

Deaton, A. and Muellbauer, J. 1980. *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.

Easterlin, R. 1995. Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior and Organization* 27, 35–47.

Greer, J. and Thorbecke, E. 1986. A methodology for measuring food poverty applied to Kenya. *Journal of Development Economics* 24, 59–74.

Groedhart, T., Halberstadt, V., Kapteyn, A. and van Praag, B. 1977. The poverty line: concept and measurement. *Journal of Human Resources* 12, 503–20.

Kapteyn, A. 1994. The measurement of household cost functions: revealed preference versus subjective measures. *Journal of Population Economics* 7, 333–50.

Kapteyn, A., Kooreman, P. and Willemse, R. 1988. Some methodological issues in the implementation of subjective poverty definitions. *Journal of Human Resources* 23, 222–42.

Orshansky, M. 1963. Children of the poor. *Social Security Bulletin* 26, 3–29.

Osmani, S. 1982. *Economic Inequality and Group Welfare*. Oxford: Oxford University Press.

Oswald, A. 1997. Happiness and economic performance. *Economic Journal* 107, 1815–31.

Paul, S. 1989. A model of constructing the poverty line. *Journal of Development Economics* 30, 129–44.

Pollak, R. 1991. Welfare comparisons and situation comparisons. *Journal of Econometrics* 50, 31–48.

Pollak, R. and Wales, T. 1979. Welfare comparison and equivalence scale. *American Economic Review* 69, 216–21.

Pradhan, M. and Ravallion, M. 2000. Measuring poverty using qualitative perceptions of consumption adequacy. *Review of Economics and Statistics* 82, 462–71.

Ravallion, M. 1994. *Poverty Comparisons*. Chur, Switzerland: Harwood Academic Press.

Ravallion, M. 1998. Poverty lines in theory and practice. Living Standards Measurement Study Working Paper No. 133. Washington, DC: World Bank.

Ravallion, M. and Bidani, B. 1994. How robust is a poverty profile? *World Bank Economic Review* 8, 75–102.

Ravallion, M. and Lokshin, M. 2002. Self-rated economic welfare in Russia. *European Economic Review* 46, 1453–73.

Ravallion, M. and Lokshin, M. 2005. Who cares about relative deprivation? Policy Research Working Paper 3782. World Bank, Washington, D.C.

Ravallion, M. and Lokshin, M. 2006. Testing poverty lines. *Review of Income and Wealth* 52, 399–421.

Ravallion, M. and Sen, B. 1996. When method matters: monitoring poverty in Bangladesh. *Economic Development and Cultural Change* 44, 761–92.

Runciman, W. 1966. *Relative Deprivation and Social Justice*. London: Routledge and Kegan Paul.

Rowntree, B. 1901. *Poverty: A Study of Town Life*. London: Macmillan.

Samuelson, P. 1938. A note on the pure theory of consumer behaviour. *Economica* 5, 61–71.

Scitovsky, T. 1978. *The Joyless Economy*. Oxford: Oxford University Press.

Sen, A. 1983. Poor, relatively speaking. *Oxford Economic Papers* 35, 153–69.

Sen, A. 1985. *Commodities and Capabilities*. Amsterdam: North-Holland.

Sen, A. 1992. *Inequality Re-examined*. Oxford: Oxford University Press.

Van Praag, B. 1968. *Individual Welfare Functions and Consumer Behavior*. Amsterdam: North-Holland.

Varian, H. 1978. *Microeconomic Analysis*. New York: Norton.

WHO (World Health Organization). 1985. *Energy and Protein Requirements*. Technical Report Series 724. Geneva: WHO.

Wodon, Q. 1997. Food energy intake and cost of basic needs: measuring poverty in Bangladesh. *Journal of Development Studies* 34, 66–101.

How to cite this article

Ravallion, Martin. "poverty lines." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000349> doi:10.1057/9780230226203.1319

The New Palgrave Dictionary of Economics Online

poverty traps

Kiminori Matsuyama
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A poverty trap is a self-perpetuating condition, in which an economy suffers from persistent underdevelopment, vicious circle of poverty, created by circular causation due to the presence of some external economies and/or strategic complementarities. We discuss the concept in a dynamic setting, and review some models of poverty traps in the literature. The policy prescriptions of such models should be treated with caution, since each model identifies one cause; but as many causes are likely to coexist, attempts to pull an economy out of one trap may push it into another.

Keywords

poverty traps; stochastic shocks; human capital; division of labour; market size; distortions

Article

A poverty trap is a self-perpetuating condition whereby an economy, caught in a vicious circle, suffers from persistent underdevelopment. Although it is often modelled as a low-level equilibrium in a static model of coordination failures, we discuss the concept in a dynamic setting. This is because, in a static setting, we would be unable to distinguish poverty traps from (possibly temporary) bad market outcomes, such as recessions and financial crises, that are also often modelled as low-level equilibriums in a static model of coordination failures.

On the mechanics of poverty traps

Imagine that the state of the economy in period t is represented by a single variable, x_t, where a higher x means that the economy is more developed, and that the equilibrium path follows a deterministic one-dimensional difference equation, $x_{t+1} = F(x_t)$. Once the initial condition, x_0, is given, this law of motion can be applied iteratively to obtain the entire trajectory of the economy.
In Figure 1a, $F(x)$, stays above the 45° line everywhere, hence the economy grows forever (as in the endogenous growth models). In Figure 1b, for any x_0, the economy converges to x^* (as in the Solow growth model). In either case there is no poverty trap, since the long-run performance of the economy is independent of the initial condition, no matter how underdeveloped the economy is initially. (Confusion sometime occurs because a few authors use the term ‘trap’ to describe the situation depicted in Figure 1b, in the sense that growth is not sustainable. However, this should more appropriately be called ‘the limit to growth’. This limit is not caused by the initial poverty of the economy.)
Figure 1
(a), (b)

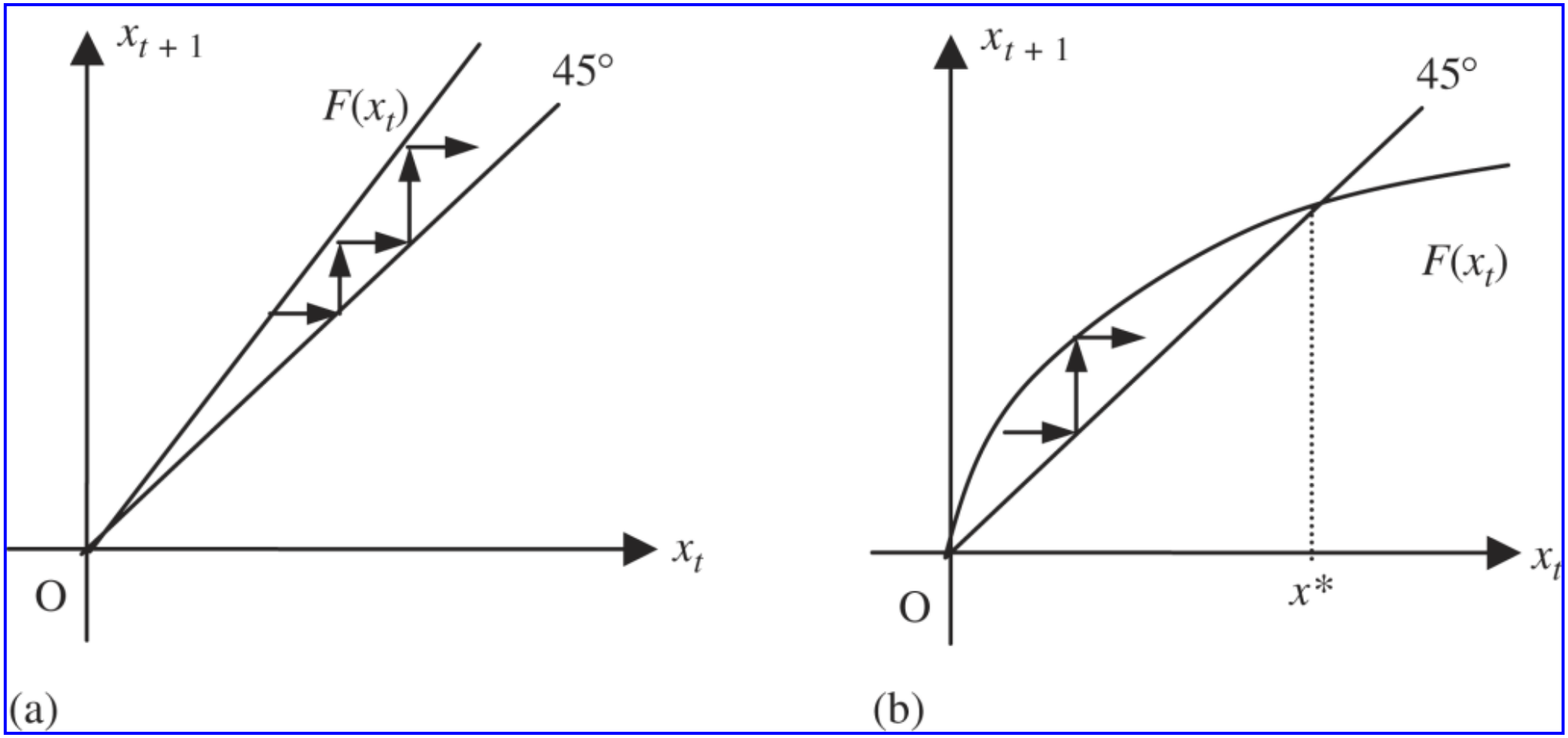

In Figures 2a and 2b, on the other hand, the long-run performance depends on the initial condition. When the economy starts above x_c, it will stay above x_c and may either grow forever or reach a higher stationary state. However, if it starts below x_c, it will be trapped forever below x_c. In this sense, both figures exhibit a poverty trap in its strong form. In Figure 2a, the economy caught in the trap will converge to the low-level stationary state. In Figure 2b, it will fluctuate below x_c. In both cases, the economy will remain poor only because it is poor. Thus, the poverty becomes its own cause. It is this self-perpetuating nature that sets ‘the poverty trap’ apart from ‘the limit to growth’.

Figure 2
(a), (b)

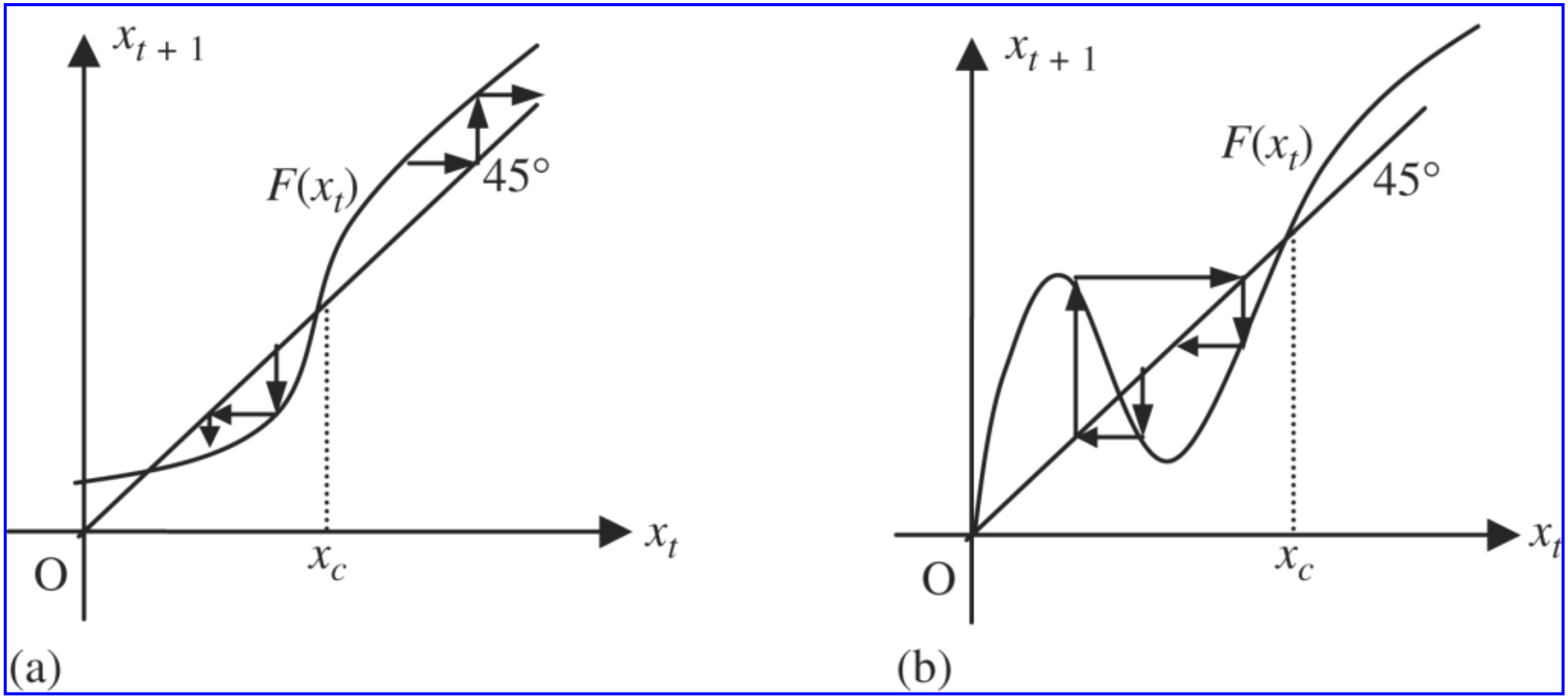

Both Figures 2a and 2b project the very stark view that the economy can never escape from the poverty trap. This should not be taken too literally. The essential message of poverty traps is that poverty tends to persist, and that it is difficult, but not necessarily impossible, for the economy to escape from it. Poverty traps in their weak form are depicted in Figures 3a and 3b. In Figure 3a, the economy has to experience stagnation for long time as it travels through the 'narrow corridor' between $F(\cdot)$ and the 45° line, before eventually succeeding in taking off. In Figure 3b, the economy may or may not manage to escape the trap after experiencing (possibly many) periods of volatility. For all practical purposes, the situations depicted in Figures 2a and 2b and Figures 3a and 3b are difficult to separate, but the message is the same: the self-perpetuating nature of poverty.

Figure 3

(a), (b)

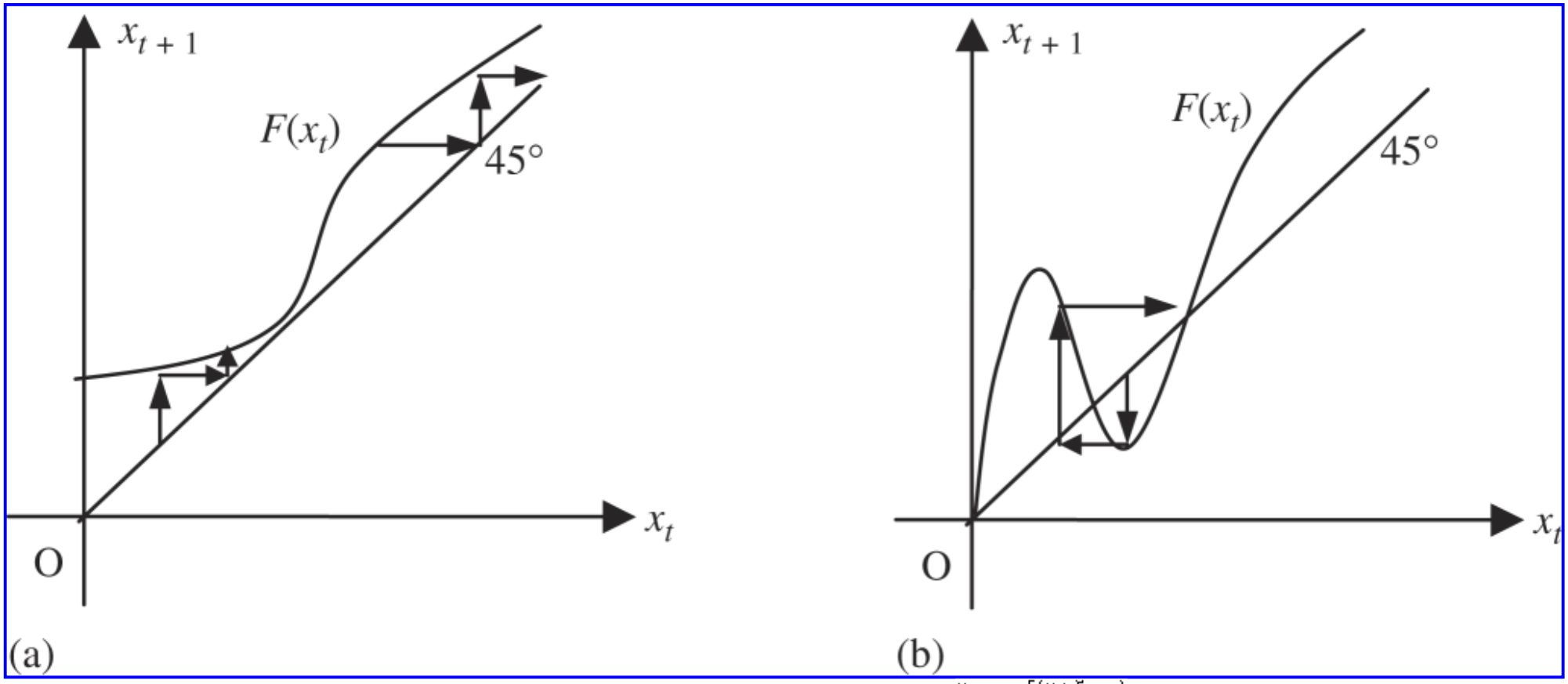

The above analysis can be extended in many directions. First, one could add stochastic shocks to the system, as $x_{t+1} = F(x_t; \xi_{t+1})$. Such shocks perturb the map, which may switch the graph back and forth between Figures 2a (or 2b) and Figures 3a (or 3b). This can be viewed as a jump in the state variable in the case of the additive shocks, $x_{t+1} = F(x_t) + \xi_{t+1}$. (For example, natural disasters, plagues and wars could cause the capital–labour ratio to jump up and down.) In the presence of such stochastic shocks, the economy may occasionally and recurrently escape or fall into the trap. Hence, the analysis has to be described in terms of the stochastic kernel; see Azariadis and Stachurski (2005) for a detailed discussion of stochastic poverty trap models.

Second, the above analysis assumes that x_{t+1} is uniquely determined as a function of x_t. If the underlying economic models permit multiple equilibria, as often is the case with models of external economies and strategic complementarity, then $F(\cdot)$ becomes a correspondence, and the (deterministic) equilibrium path follows the difference inclusion, $x_{t+1} \in F(x_t)$. See Matsuyama (1997) for some examples. Figure 4 depicts one possibility, suggesting that the economy is stuck in a low-level stationary state, in part due to coordination failures. In this case, the economy could escape the poverty trap if it succeeded in coordinating on a higher equilibrium, as indicated by the dotted arrow. (If such coordination takes place through a realization of some coordination devices, ‘sunspots’, it can be viewed as a model of endogenous stochastic shocks.)

Figure 4

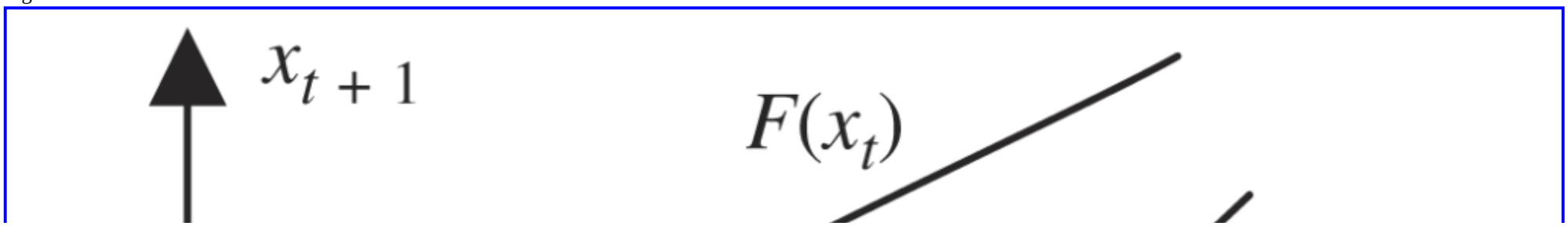

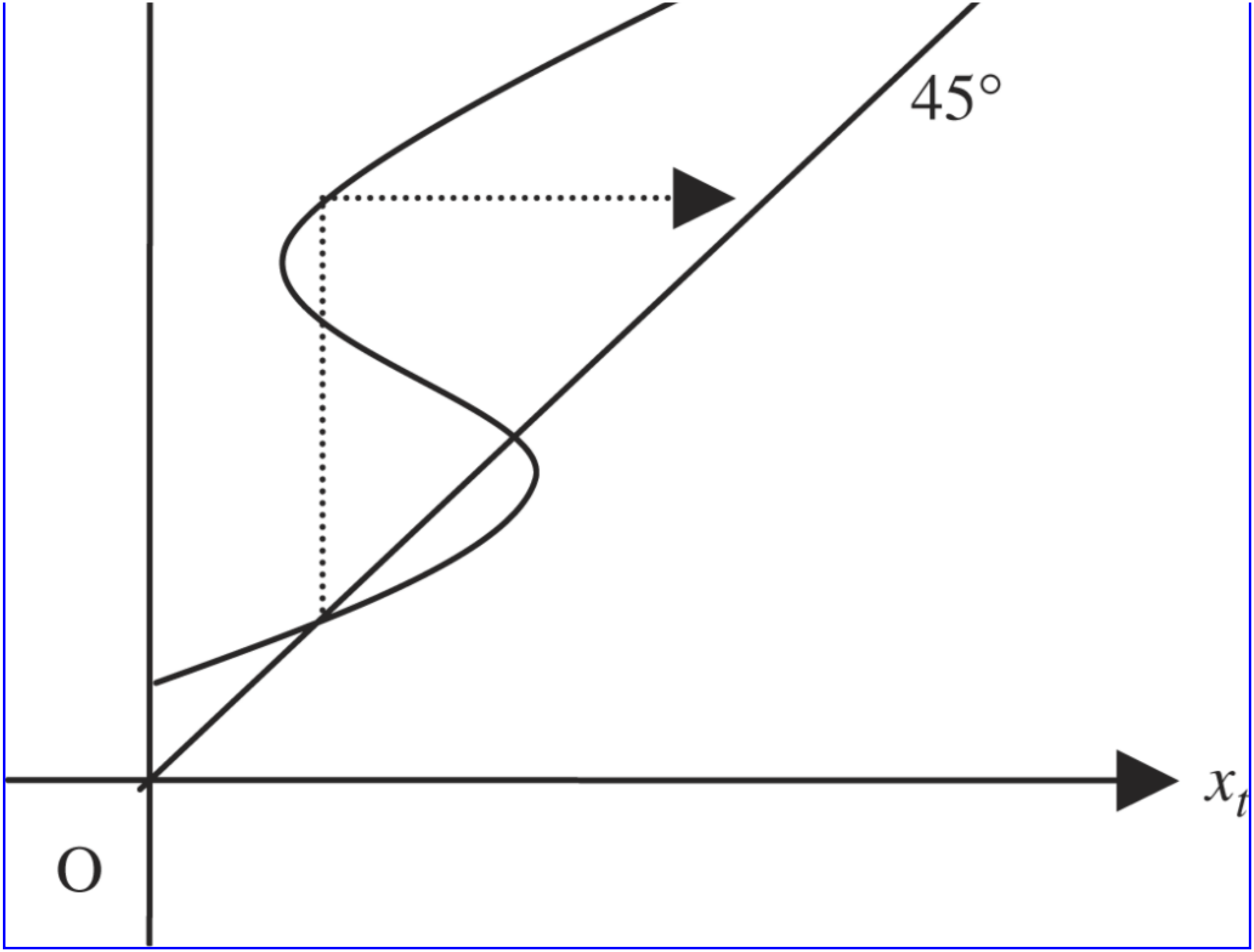
45°
x_t
O

Third, the underlying economic model may imply that the law of motion be described in a multi-dimensional system. For example, the state space may be two-dimensional, (x, q), where x is the state (or backward-looking) variable, such as the capital stock, and q is the co-state (or forward-looking) variable, such as the asset price or consumption, and the law of motion is given by a two-dimensional difference equation, $(x_{t+1}, q_{t+1}) = F(x_t, q_t)$. In this case, for a given initial condition, x_0, the equilibrium condition may not uniquely pin down the initial value, q_0. That is, there may be multiple equilibrium paths, with self-fulfilling expectations, which suggests another way in which the economy may escape from the poverty trap; see Matsuyama (1991). Or the dimensionality of the state space may be equal to the number of industries in a multi-industry model, or to the number of countries in a multi-country world economy model. In such a high-dimensional system, one could encounter a much richer set of dynamics, where the long-run behaviour can depend on the initial condition in a much more complex manner.

Some models of poverty trap

Many (dynamic) models of poverty traps have been proposed in the literature. The common feature of these models is the presence of some external economies or strategic complementarities that give rise to the circular causation. Here is a highly selective list.

Learning-by-doing externalities

The infant industry argument for protection (see Corden, 1977, for a synopsis) is a classic example. When firms are inexperienced and unproductive, they cannot offer wages high enough to attract workers from other sectors, and hence are not able to accumulate experience. Temporary protection has been suggested as a way to break the vicious cirlce. Helping some industries accumulate experience to escape from a poverty trap, however, may end up pushing the economy into another poverty trap, as it could prevent other (new and possibly more promising) industries from growing. If the scope of productivity improvement in any industry is limited, then the only way of avoiding poverty traps and achieving sustainable growth is to keep the delicate balance so that production will shift constantly from one industry to another, as existing industries become mature and new industries are born; see Stokey (1988); Brezis, Krugman and Tsiddon (1993); Matsuyama (2002).

Search externalities

The difficulty of finding business partners can discourage many from entering an industry, which in turn makes it even harder for others to find business partners. See Diamond (1982).

Human capital externalities

Following the Lucas (1988) model of endogenous growth based on human capital accumulation, Azariadis and Drazen (1990) showed how it could lead to the existence of poverty traps, when human capital is subject to threshold externalities.

Market size and division of labour

Adam Smith argued that ‘the division of labour is limited by the extent of the market’. Young (1928) argued that the extent of the market is also limited by the division of labour. That is, economic growth can be achieved by means of greater specialization, which was formalized by Romer (1987) and others. Building on this body of work, Ciccone and Matsuyama (1996) showed how the economy can be caught in a poverty trap. The basic mechanism is that advanced technologies require the use of highly specialized equipment and producer services. In the underdeveloped economy, the limited availability of specialized inputs forces downstream industries to rely on less advanced technologies, which do not require the use of specialized inputs. This in turn leads to a small market size for specialized firms in upstream industries. Hence, the economy is caught in the vicious circle of limited market size and limited division of labour.

Financial developments

In countries with limited opportunities to diversify risk, entrepreneurs are discouraged from making productive but risky investments. This in turn leads to a limited set of traded

financial assets, which reduces the opportunity to diversify risk. See Saint-Paul (1992) and Acemoglu and Zilibotti (1997).

Low wealth/low investment

When external finance is more costly than internal finance, a decline in borrower net worth leads to a higher investment distortion. In Bernanke and Gertler (1989), this leads to a decline in the investment, which in turn leads to a decline in the net worth of the next generation of entrepreneurs, hence generating persistence in the aggregate investment dynamics. In Matsuyama (2004), the same mechanism could make some (but not all) countries in the world caught in the vicious circle of low net worth–low investment. Matsuyama (2007) showed how the trap can sometimes take the form of greater volatility (as shown in Figure 2b). In a set-up that allows for wealth distribution to evolve over time, Banerjee and Newman (1993) suggested that greater initial wealth inequality, to the extent that it increases the number of entrepreneurs rich enough to finance their investments, can lead to a higher aggregate investment, which in turn could help the poor in the long run, thereby breaking the vicious circle.

Demographic trap

Nelson (1956) is among the first to argue that underdeveloped countries are caught in the vicious circle of high population growth and low per capita income. Becker, Murphy and Tamura (1990) showed how the economy may be caught in the vicious circle of high fertility–low human capital. Basu (1999) and Doepke and Zilibotti (2005) discussed child labour traps. In Matsuyama (2000), inter-generational persistence of a high labour force participation rate by the elderly could lead to a poverty trap.

Contagious social norms

Tirole (1996) showed how corruption or other unethical behaviour can be contagious and persistent. He considered the setting where, in the presence of imperfect information, the reputation of a member of the group (say, a firm in the industry) depends not only on his own past behaviour, but also on the past behaviour of other group members. Then, when the group has the reputation of being dishonest, it would be difficult for the member to establish a reputation for honesty. This induces him to behave dishonestly, thereby contributing to the bad reputation of the group.

Modelling inertia

Underdevelopment is often modelled as a Pareto-dominated equilibrium in a static game of strategic complementarities. Murphy, Shleifer and Vishny (1989) is the best-known example. By adding some inertia, which restricts the ability of the players to switch their strategies, one can convert virtually any static game of strategic complementarities into a dynamic model of poverty traps, where both the initial condition and expectations can play a role in determining the long-run performance of the economy. See the techniques developed by Matsuyama (1991) and Matsui and Matsuyama (1995).

Some cautionary remarks on interpretations

The poverty trap is often interpreted as an explanation for *cross-country* income difference. As such, it is frequently viewed as an *alternative* to the models that attribute *cross-country* income difference to the *cross-country* difference in, say, TFP and/or investment distortions. This is a misinterpretation. First, the message of poverty trap models is the self-perpetuating nature of poverty. It suggests that the long-run performance of an economy could be much better if its initial condition were better. It does *not* mean that the cross-country difference in the long-run performance is due mostly to the difference in their initial conditions. Second, the notion of poverty trap does not contradict the observation that low income is often associated with low TFP and/or high investment distortions. Indeed, many poverty trap models attempt to explain the two-way causality between low-income and low TFP and/or high investment distortions. By endogenizing TFP and/or investment distortions, these poverty trap models go one step further than the models that treat these variables as exogenously given.

Many calls for foreign assistance for underdeveloped countries can be understood using the notion of poverty trap; see, for example, Sachs et al. (2004). Indeed, the poverty trap is often viewed as a powerful case for policy activism. However, one should be careful when using any particular model of the poverty trap to make policy proposals. It is important to keep in mind that each model of the poverty trap is designed to highlight one particular feedback mechanism behind the vicious circle. To this end, other sources of the poverty trap are deliberately assumed away. In reality, of course, many sources of the poverty trap are likely to coexist. If there is one important lesson from the literature reviewed above, it

should be that there are hundreds of traps that the economy can fall into, and any policy intervention that attempts to pull the economy out of one trap may end up pushing it into another. As we know, any attempt to solve a problem can often become a source of another, even bigger problem. For more on this issue, see Matsuyama (1996), which discusses economic development as ‘complex’ coordination problems.

See Also

- structural change
- symmetry breaking

Bibliography

Acemoglu, D. and Zilibotti, F. 1997. Was Prometheus unbounded by chance?: Risk, diversification and growth. *Journal of Political Economy* 105, 709–51.

Azariadis, C. and Drazen, A. 1990. Thresholds externalities in economic development. *Quarterly Journal of Economics* 105, 501–26.

Azariadis, C. and Stachurski, J. 2005. Poverty traps. In *Handbook of Economic Growth*, vol. 1, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Basu, K. 1999. Child labor: causes, consequences and cures, with remarks on international labor standards. *Journal of Economic Literature* 37, 1083–119.

Banerjee, A. and Newman, A. 1993. Occupational choices and process of development. *Journal of Political Economy* 101, 274–98.

Becker, G., Murphy, K. and Tamura, R. 1990. Human capital, fertility, and economic growth. *Journal of Political Economy* 98, S12–S37.

Bernanke, B. and Gertler, M. 1989. Agency costs, net worth, and business fluctuations. *American Economic Review* 79, 14–31.

Brezis, E., Krugman, P. and Tsiddon, D. 1993. Leapfrogging in international competition: a theory of cycles in national technology leadership. *American Economic Review* 83, 1211–19.

Ciccone, A. and Matsuyama, K. 1996. Start-up costs and pecuniary externalities in economic development. *Journal of Development Economics* 49, 33–57.

Corden, W. Max. 1977. *Trade Policy and Economic Welfare*. Oxford: Clarendon Press.

Diamond, P. 1982. Aggregate demand management in search equilibrium. *Journal of Political Economy* 90, 881–94.

Doepke, M. and Zilibotti, F. 2005. The macroeconomics of child labor regulation. *American Economic Review* 95, 1492–524.

Lucas, R., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Matsui, A. and Matsuyama, K. 1995. An approach to equilibrium selection. *Journal of Economic Theory* 65, 415–34.

Matsuyama, K. 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. *Quarterly Journal of Economics* 106, 617–50.

Matsuyama, K. 1996. Economic development as coordination problems. In *The Role of Government in East Asian Development: Comparative Institutional Analysis*, ed. M. Aoki, H. Kim and M. Okuno-Fujiwara. New York: Oxford University Press.

Matsuyama, K. 1997. The 1996 Nakahara Lecture: complementarity, instability and multiplicity. *Japanese Economic Review* 48, 240–66.

Matsuyama, K. 2000. Economic development with endogenous retirement. CMS-EMS Discussion Paper No. 1237R. Northwestern University.

Matsuyama, K. 2002. The rise of mass consumption societies. *Journal of Political Economy* 110, 1035–70.

Matsuyama, K. 2004. Financial market globalization, symmetry-breaking, and endogenous inequality of nations. *Econometrica* 72, 853–84.

Matsuyama, K. 2007. Credit traps and credit cycles. *American Economic Review* 97, 503–16.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Industrialization and the Big Push. *Journal of Political Economy* 97, 1003–26.

Nelson, R. 1956. A theory of the low level equilibrium trap in underdeveloped economies. *American Economic Review* 46, 894–908.

Romer, P. 1987. Growth based on increasing returns due to specialization. *American Economic Review* 77, 56–62.

Sachs, J. et al. 2004. Ending Africa's poverty trap. *Brookings Papers on Economic Activity* 1:2004, 117–240.

Saint-Paul, G. 1992. Technology choice, financial markets and economic development. *European Economic Review* 36, 763–81.

Stokey, N. 1988. Learning-By-Doing and the Introduction of New Goods. *Journal of Political Economy* 96, 701–17.

Tirole, J. 1996. A theory of collection reputations (with applications to the persistence of corruption and to firm quality. *Review of Economic Studies* 63, 1–22.

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

How to cite this article

Matsuyama, Kiminori. "poverty traps." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000332>
doi:10.1057/9780230226203.1320

The New Palgrave Dictionary of Economics Online

poverty

A.B. Atkinson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Concern for poverty has been expressed over the centuries, even if its priority on the agenda for political action has not always been high. Its different meanings and manifestations have been the subject of study by historians, sociologists and economists. Its causes have been identified in a wide variety of sources, ranging from deficiencies in the administration of income support to the injustice of the economic and social system. The relief, or abolition, of poverty has been sought in the reform of social security, in intervention in the labour market, and in major changes in the form of economic organization.

Poverty today is most obvious – and has the most pressing claim on our attention – on a world scale. The unequal distribution of income between countries, and the disparities within countries, mean that there are large numbers of people in Africa, Asia and Latin America whose standard of living would be agreed by everyone to be poor. The World Bank has suggested that there is ‘a global total of close to 1 billion people living in absolute poverty’ (World Bank, 1982, p. 78), of whom about 400 million are thought live in South Asia, about 150 million in China, and some 100 million in East/South-East Asia and Sub-Saharan Africa. At such levels of living, the risks of death through hunger or cold, and vulnerability to disease, are of a quite different order from those in advanced countries. This has manifested itself most urgently in the occurrence of famine. Whatever the immediate cause of such disasters, whether inadequate total supply of food or whether unequal distribution, the severity of the situation in areas such as the Sahel and Ethiopia is an indicator of the precariousness of survival in many low income countries.

Such mass poverty in poor countries is quite different from poverty in advanced countries. The target of the American War on Poverty, launched in 1964, was the minority of Americans with incomes below a poverty line of $3000 a year for a family of four (in 1962 prices), which was many times the average income of India. The basis for the US official poverty line is to be found in a food consumption standard (the Department of Agriculture economy food plan), but its level reflects the prevailing living conditions in that society. It might well be argued that concern with poverty in advanced countries, at a time when other countries face disaster, is unjustified and that the term ‘poverty’ cannot legitimately be applied. The parallel may be drawn with rearranging the deckchairs on the *Titanic* as the ship goes down. This

does not, however, seem fully apposite. A closer parallel is with the position of those on ships steaming to the aid of the stricken vessel. The overriding objective should be to get to the rescue as rapidly as possible, but those on the rescuing ships should also be concerned that their steerage passengers do not die of exposure on the way. The relief of famine, and the redistribution of income to alleviate poverty on a world scale, should have priority, but the problem of poverty in advanced countries, defined in their terms, may legitimately come next on the list of concerns.

The fact that the term ‘poverty’ is being used in different senses highlights the need to clarify the underlying concept and the discussion so far has touched on several aspects which need to be elaborated. After a brief historical review of studies of poverty in section 1, we examine some key conceptual issues. What is the indicator of resources which should be employed in measuring poverty? What is the underlying notion of poverty and how is it related to inequality? These issues are discussed in section 2. The determination of the poverty standard is a crucial question. Here we need to consider approaches based on such ‘absolute’ concepts as food requirements and those poverty scales which are explicitly) ‘relative’. We must consider the treatment of families with differing needs. These topics are the subject of section 3. Once we have established the extent of poverty, its causes become a central concern. Here we are led first to ask ‘who are the poor?’ This is examined in section 4. Is poverty concentrated in particular classes or particular sections of society? How far is it associated with particular stages of the life-cycle? The composition of the poor provides in turn a starting point for the investigation of the underlying causes of poverty, and an analysis of policies to combat poverty. These are the subject of section 5.

1 Historical review of studies of poverty

The scientific study of poverty in the Anglo-Saxon world is usually taken to date from the investigations of Booth and Rowntree at the end of the 19th century. In Britain it is true that King and others had given estimates of the number of paupers; and that *The State of the Poor* by Eden (1797) contained a great deal of material collected from over 100 parishes and giving details of family budgets. Engels and Mayhew provided insight into the condition of the poor in urban England. But it was Booth’s *Life and Labour* (1892–7) survey of London, started in the East End in the 1880s, that combined the elements of first-hand observation with a systematic attempt to measure the extent of the problem. Taking the street as his unit of analysis, he drew up his celebrated map of poverty in London.

The study of Rowntree (1901) was intended to compare the situation in York, as a typical provincial town, with that found by Booth in London, but his method represented a significant departure in that it was concerned with individual family incomes and in that he developed a poverty standard based on estimates of nutritional and other requirements. The development of survey methods was taken further by Bowley (1912–13) who pioneered the use of sampling in his 1 in 20 random sample of working-class households in Reading. A great many local studies were subsequently conducted, including Bowley's Five Towns survey in 1915, replicated in the early 1920s, and the new Survey of London Life and Labour published in the early 1930s. Rowntree himself repeated his survey of York in 1936 and 1950. The latter became the standard source of information as to the effectiveness of the post-1948 welfare state, with most commentators concluding that poverty had been effectively abolished in Britain by the combination of full employment and the new social benefits. Doubt began to be cast on this conclusion

by the work of empirical sociologists and came to the fore with the publication of *The Poor and the Poorest* by Townsend and Abel-Smith (1965). This showed, using secondary analysis of a national survey, that in 1960 about two million people fell below the social security safety net level. This finding was confirmed in official estimates which began to be published by the Department of Health and Social Security in the 1970s, and by Townsend's own major survey (1979).

As in many fields, the United States entered later and has taken the subject further. The definition of a poverty line was attempted by Hunter in 1904 and this was developed in a series of studies, such as the ‘minimum comfort’ and other budgets produced for New York City. There was the 1949 report on low income families by the Joint Committee on the Economic Report. It was not however until the 1960s that the problem of poverty received systematic study, with a few notable exceptions such as the work of Lampman (1959). *The Other America* by Harrington (1962) and *The Affluent Society* by Galbraith (1958) did much to arouse the attention of the public, politicians and academics. The 1964 report of the Council of Economic Advisers set out the $3000 poverty level, drawing heavily on the research of Orshansky (1965), and this was subsequently refined to form the official poverty line, which has been applied since that date (with modifications, such as the addition of alternative measures including the value of transfers in kind).

Similar studies have been carried out in many countries, and researchers have become increasingly interested in cross-country comparisons. The OECD made an early attempt at such comparisons and a more extensive exercise is being carried out in the Luxembourg Income Study. Any assessment of world poverty depends on the availability of information about the distribution of living standards within individual countries; and here both the World Bank and the International Labour Organization have made significant contributions. In some low income countries, there has been extensive research on poverty, India being an example, where there has been a great deal of discussion as to whether poverty has increased or decreased over time. The ILO and the World Bank have also been influential in the widespread interest, reflected in the Brandt Report (1980), in the concept of ‘basic needs’, or a minimum set of specific goods and environmental conditions.

2 Poverty: living standards and rights

Concern about poverty may take the form of concern about such basic needs: for example, food, housing and clothing. In this case, we can identify clearly the items of consumption in which we are interested. This approach leads to poverty being measured in a multidimensional way, where a family may be deprived in one but not other respects, although particularly serious will be situations where families suffer deprivation in several dimensions, or what is referred to typically as ‘multiple deprivation’.

This approach is concerned with specific deprivation, but we may also seek to record disadvantage in a single index of living standards, such as total expenditure, a household being said to be in poverty if it has total expenditure below a specified amount. This is not however the approach followed in most studies of poverty in advanced countries, which record poverty on the basis of total *income*. Income may *understate* the level of living. A family may be able to dissave or to borrow, in which case its current level of living is not constrained by current income and expenditure may be the more appropriate index. (Although in the short run there may be a divergence between *consumption* and *expenditure*, as families use up stocks of goods, etc.) The level of living may exceed that permitted by income where the family

is able to share in the consumption of others. An elderly person living with his or her children may benefit from their expenditure. Income may, conversely, *overstate* the level of living. This may happen where money alone is not sufficient to buy the necessary goods: where there is rationing, or unavailability of goods. It is also possible that people choose a low level of consumption. This latter reason has led to its being argued that income *should* be the indicator of poverty, since it is a measure of the opportunities open to a family and is not influenced by the consumption decisions made.

In considering the choice between income and expenditure, it is helpful to distinguish two rather different conceptions of poverty: that concerned with *standards of living* and that concerned with *minimum rights* to resources. On the former approach, the goal is that people attain a specified level of consumption (or consumption of specific goods); on the latter approach, people are seen as entitled, as citizens, to a minimum income, the disposal of which is a matter for them. In practice, the two notions are often confounded, but the distinction is important, and it has obvious implications for the choice of poverty indicator. Income is the focus of the rights approach, but its use on a standard of living approach must be seen as a proxy for consumption.

The reference to ‘rights’ raises the question of the relation between poverty and inequality. Here four different schools of thought may be distinguished. There are those who are concerned only with poverty, attaching no weight to income inequalities above the poverty line. There are those who attach weight to the reduction of inequality as a goal of policy but give priority to the elimination of poverty, so that we have a lexicographic objective function. There are those who are concerned about both goals and who are willing to trade gains in one direction against losses in the other. Finally, there are those who attach no especial significance to poverty, simply regarding it as a component of the wider cost of inequality.

In this context, reference should be made to the choice of *poverty measures*. Where poverty puts survival in doubt, it is natural to take as one's measure the proportion of the population at risk. Concern for minimum rights may also make the ‘head count’ the most relevant measure. But we may also be concerned, particularly on a standard of living approach, with the severity of poverty, in which case measures such as the poverty deficit (the total shortfall from the poverty line) may be more appropriate. One can indeed go further, as proposed by Sen (1976), and take account of the distribution of income within the poor population: for example, with the poverty index depending on the Gini coefficient for this distribution.

3 Setting the poverty line

The most straightforward approach to the determination of the poverty line is to specify a basket of goods, denoted by the vector $\mathbf{x}^*$, purchasable at prices $\mathbf{p}$, and to set the poverty standard as:

$$(1 + h)\mathbf{p} \cdot \mathbf{x}^*$$

where h is a provision for inefficient expenditure or waste, or a provision for items not included in the list $\mathbf{x}^*$. This was in effect the method adopted by Rowntree, whose diet for Tuesdays was porridge for

breakfast, bread and cheese for lunch, and vegetable broth for dinner. It was the method followed by Orshansky, where $\mathbf{x}^*$ represented food requirements and h (=2) made allowance for spending on other goods. This approach is often referred to as an ‘absolute’ poverty standard, and contrasted with a ‘relative’ approach that relates the poverty line to contemporary levels of living: for example the proposal of Fuchs in the United States that the poverty line should be one-half the median family income. It is sometimes suggested that the absolute standard is less problematic than the relative approach and less dependant on value judgements.

The term ‘absolute’ can, however, scarcely be used in the same sense as in the physical sciences and there is scope for a great deal of disagreement about where the line should be drawn. This is most evident in the case of the rights approach, where the determination of the minimum level of income is explicitly a social judgement, but it applies also to the standard of living approach. In the case of food requirements, where a physiological basis may appear to provide a firm starting point, it is in fact difficult to determine $\mathbf{x}^*$ with any precision. There is no one level of food intake required to survive, but rather a broad range where physical efficiency declines with a falling intake of calories and protein. Nutritional needs depend on where people live and on what they are doing. They vary from person to person, so that any statement can only be probabilistic: at a certain level of consumption there is a certain probability that the person is inadequately fed. Even if these problems could be resolved, there is the difficulty of the disparity between expert recommendations and actual consumption behaviour. The factor h is intended to allow for this, but the precise allowance will depend on the judgement of the investigator. Rowntree, for example, included an allowance for tea, which has little or no nutritional value but which formed a staple item of consumption.

In the case of non-food items, there is even greater scope for judgement. This applies whether we seek to include the goods in the vector $\mathbf{x}^*$ or whether we allow for non-food items via the multiplier h. For example, the procedure of Orshansky has been criticized as under-stating the proportion of income spent on food and hence overstating the value of h. More fundamentally, the role of goods in the determination of the poverty line needs reconsideration. The literature on ‘household production’ has pointed to the role of goods as an input into household activities, with the level of activities being our main concern rather than the purchase of goods as such. On this basis, if we denote the target level of activities by $\mathbf{z}^*$, and if there is an input–output matrix A, relating goods inputs to activity levels, then the necessary level of expenditure becomes:

$$Y = (1 + h)\mathbf{p}A\mathbf{z}^*$$

The significance of this view is that poverty may be measured in absolute terms, in the sense that the vector $\mathbf{z}^*$ is fixed, but the required bundle of goods may be changing because the input–output matrix is affected by developments in the particular society. If the activity is ‘attending school’, then the demands in terms of clothing, books and equipment are quite different today from those of a century ago. This does not mean that there is no distinction between absolute and relative concepts. There is a clear difference in principle between taking the vector $\mathbf{z}^*$ as fixed and allowing it to be influenced by the

living patterns of the rest of society, as in the work of Townsend (1979), who is concerned with the extent to which families can participate in the ‘community's style of living’.

The notion of a fixed absolute poverty standard, applicable to all societies and at all times, is therefore a chimera. Nor is it evident that a poverty standard, once set, can be compared across time by simply adjusting by an index of consumer prices. In the case of both absolute and relative approaches, we have to face the problems of judgement. Here several lines of attack may be discerned. There are studies which take the *official* poverty standards as embodying social values, which seems natural on the minimum rights approach and which at least provides a measure of governmental performance. There are studies which base the poverty line on the views expressed in surveys of the population as a whole. In the United States, the Gallup Poll has regularly asked the question: ‘What is the smallest amount of money a family of four needs to get along in this community?’ These, and other approaches, will produce a range of poverty lines, and it seems unlikely that we can reach universal agreement. There are therefore strong reasons for recognizing such differences of view explicitly and using a *range* of poverty lines. This means that we may not be able to reach unambiguous conclusions – it may be that poverty will be shown to have increased according to one line but not according to another – but it will avoid a total impasse. In the same way, when making a comparison over time, we may want to compare 1950 with two alternative lines for 1980, one updated by the price index and the other adjusted to allow for rising real incomes, thus generating a ‘confidence interval’ around the 1980 estimate.

To this point, the poverty line has been discussed as though it were a single number, but families of different types and different sizes will receive different treatment. In Britain, for example, the social security safety net is typically some 60 per cent higher for a couple than for a single person. The relationship between the poverty lines for different family types is usually referred to as an *equivalence scale*. However, a prior question before the equivalence scales are determined is the choice of the *unit of analysis*. Here the distinction between the standard of living and rights approaches is important. In the latter case, the notion of rights must be essentially individualistic. The case for considering a wider unit must rest on there being within-family transfers which cannot be adequately observed. The family is taken when measuring poverty because we do not accept that a large number of those with zero recorded cash income are in fact without resources. At the same time, little is known about the distribution of income within the family. Certainly, it would be quite wrong to treat all married couples as having equal rights to the joint income. On a standard of living approach, the logical unit is that which shares consumption; and we may wish to go beyond the inner family to the household as a whole. This would take account of the fact that items of expenditure may have ‘public good’ characteristics for the family members. Again, however, it may be that there are unequal living standards within the household.

Several approaches have been adopted to the determination of the equivalence scales for different-sized units. Survey information about individual assessments of what is needed ‘to get along’ has been used for this purpose. More commonly, the basis has been sought for observation of actual behaviour. One of the early methods provides an illustration. By taking a commodity consumed only by adults (e.g. men's clothing), one can observe the level of income at which a family with one child, say, can attain the same level of consumption of that commodity as a family with no children. This method, and other more sophisticated implementations of the idea, have been the subject of considerable debate. The underlying difficulty is that one is assuming, in the example given, that preferences for the commodity are independent of family composition: the arrival of the child may mean that the couple go out less and spend less on clothing. With other methods based on observed consumption behaviour, identifying

restrictions are similarly needed. At a more fundamental level, the ethical status of such scales is far from transparent. Not only is it impossible to draw conclusions about welfare levels with different family compositions, but also society may wish to modify the implied judgements: for example, to vary the parental evaluation to take account of the interests of the children.

4 The composition of the poor

One of the main aims of those investigating poverty has been to establish who the poor are. Popular opinion is often coloured by vivid, but not necessarily representative, accounts of life below the poverty line. For this reason, the Council of Economic Advisers stressed at the start of the War on Poverty in the US that poverty should not be seen as a minority phenomenon: ‘Some believe that most of the poor are found in the slums of the central city, while others believe that they are concentrated in areas of rural blight. Some have been impressed by poverty among the elderly, while others are convinced that it is primarily a problem of minority racial and ethnic groups. But objective evidence indicates that poverty is pervasive … the poor are found among all major groups in the population and in all parts of the country’ (1964, pp. 61–2).

Poverty in advanced countries affects a minority in terms of numbers but it is not confined to specific marginal groups. At the same time, certain groups are much more at risk. In 1983, the poverty rate for blacks in the United States was nearly three times that for whites, and that for Hispanics was more than twice. Compared with the average, the rate for families with children is nearly double, and that for families with a female head is much higher. The evidence for other countries equally shows large differences in the incidence of poverty between groups: for instance, in Malaysia, recorded poverty among Malays is much higher than among the Indian or Chinese ethnic groups. The World Bank has argued that poverty in low income countries is very much a rural problem; and the evidence from India shows poverty to be much higher in rural than urban areas.

If we seek to probe further into the composition of the poor, then the dynamics of poverty must be taken into account. Is poverty a largely transitory phenomenon, in that the families poor today will quite probably be above the poverty line next year? Is poverty associated with particular periods of the life cycle? Transitory poverty may occur for a variety of reasons. Income may be temporarily reduced because of ill-health or unemployment or because wages are cut. It may be a bad harvest. Families may split up, leaving one parent with the family responsibilities but inadequate income. The evidence from panel surveys, where the same families are interviewed on a continuing basis (as, for example, in the Michigan Panel Study of Income Dynamics), has shown the extent of mobility in the incomes and circumstances of the poor. A sizeable fraction of those recorded as poor in one year are above the poverty line next year. This does not mean that their poverty is not a matter for concern, since low current incomes may impose severe hardship, but it means that these people do not constitute a permanent ‘under class’.

Such mobility does however require careful interpretation. It may arise on account of the life cycle. In Rowntree's 1899 survey he found that the life of the labourer was marked by ‘five alternating periods of want and comparative plenty’, the periods of want being childhood, when he himself had children, and old age. The impact of such life-cycle factors depends on the extent to which income support is provided by state or private transfers. In this respect the situation in Britain has changed dramatically since 1899, with the introduction of state pensions, a large increase in private pensions, and the payment of child and

other benefits. In other countries too there has been major growth in transfers: between 1960 and 1981 social expenditure as a percentage of GDP rose in the United States from 7 per cent to 15 per cent, in West Germany from 18 per cent to 27 per cent, and in Japan from 4 to 14 per cent (Institute for Research on Poverty, 1985). Transfers, and other programmes, such as health care, must have reduced the extent of life-cycle poverty. The incomes of the elderly in the United States, for example, are considered to have risen relative to those of the population as a whole. But there remains concern about certain stages of the life cycle, particularly among families with children; and while the poverty rate among the elderly in the US has fallen, that among the non-elderly has risen.

To the extent that poverty is a life-cycle phenomenon, this means that more people experience poverty at some point in their lives but that its duration is limited. At the same time, poverty at one stage of the life cycle may lead to poverty at a subsequent stage. Those who are hard-pressed when they are bringing up children may have little savings on which to draw in retirement. Those who grow up in low-income families may themselves be more likely to be below the poverty line, as was found in the follow-up in the 1970s of the children of the families interviewed by Rowntree in 1950 (Atkinson, Maynard and Trinder, 1983). Moreover, we should not however lose sight of the fact that for some people poverty persists. Agricultural labourers, or farmers with small plots, may be in poverty even in ‘good’ years. Among industrial workers, there are those whose earnings are inadequate to support even themselves; there may be a problem of *low pay*. And the low paid may be more vulnerable to the transitory factors such as ill-health and unemployment.

5 Causes and policies

In 1913, R.H. Tawney argued for the restatement of the problem of poverty: ‘the diversion to questions of social organization of much of the attention which, a generation ago, was spent on relief’. The problem of poverty, he said, was ‘primarily an industrial one’. In terms of the composition of the poor described above, this means that the causes of poverty were sought not in the failure of income support but in the reasons why income was inadequate in the first place.

Tawney recognized the importance of personal factors in causing poverty, but laid principal stress on the position of groups and classes and their economic situation, factors which may equally be relevant today. Workers may be locked into low-paying industries where techniques and machinery need to be modernized; they may live in depressed regions to which private capital cannot be attracted. There may be a low level of unionization and employers may be able to hold wages down. These aspects, which have been emphasized in theories of ‘segmentation’ in the labour market, point to the need for government intervention. This may take the form of minimum wage legislation, to guarantee minimum levels of earnings, coupled with measures to offset any adverse effect on employment and to modernize the sectors or regions concerned. At a macro-economic level, the government has an important responsibility. Studies in the United States have identified unemployment as a much more serious problem than inflation for low income groups. There can be little doubt, for example, that the recession of the 1980s has increased the incidence of poverty in advanced countries.

The counterpart of this structural explanation in the context of less developed, primarily agricultural economies is to be found in the role of land tenure and its distribution and in the nature of labour and capital markets. Rural poverty is high among landless labourers and those farmers with small or

unproductive holdings. Their difficulties may be intensified by the terms on which they have to borrow or purchase intermediate goods. Here too policy requires government intervention, whether to redistribute land holdings, or to facilitate the introduction of new methods, or to eliminate extortionate lending practices, or to provide non-farm employment. Measures such as land reform raise major political issues, and in both developing and advanced countries it can be argued that basic changes in the form of economic system are necessary to eradicate poverty. The World Bank has noted, for example, the role played by the Chinese food security policy in the reduction of poverty and the way in which it is tied into China's collective system.

The industrial explanation of poverty may be contrasted with the ‘supply side’ explanation which has seen low pay as attributable to workers lacking productive skills, because they have been unable to complete education or training. This ‘human capital’ interpretation leads in turn to the policy recommendation that training and educational programmes should be expanded, a proposal that is congruent with the goal of reducing inequality of opportunity. Education and training had a central role in the United States War on Poverty, with schemes such as the Job Corps and the Neighborhood Youth Corps. A characteristic of individual workers also identified in the United States is that of race. Discrimination may lead to otherwise equally qualified workers receiving lower pay, as where black workers were prevented from entering certain occupations. The civil rights legislation and the operations of the Equal Employment Opportunity Commission may have reduced the direct effect of discrimination (as well as the indirect effect via unequal opportunities in education, etc.), but although the policy implications are clear in principle, experience suggests that they are not easily made effective.

Policies to improve job and earnings prospects must be central to the elimination of poverty, but they cannot succeed without complementary income maintenance provisions. The growth of transfers has not succeeded in providing a completely effective income guarantee for those without incomes from work or with additional needs. This is because of incomplete coverage, particularly where new needs develop, because of the inadequate levels of benefits (for example, those paid to people with poor employment records) and the incomplete take-up of income-tested benefits. In the last case, there is evidence that complexity or stigma deters families from claiming the transfers to which they are entitled, and hence they fall through the safety net.

To this end, proposals have been made for major reform of the transfer systems in advanced countries. One front-runner for many years in the United States has been the ‘negative income tax’, which would pay an income-related supplement using the income tax machinery. There are those reformers who would like to integrate fully the income tax and social security systems, as with the basic income guarantee scheme, where everyone receives a basic income and is then taxed on all income. Such a reform would mean that income maintenance largely ceased to be categorical: for example, there would not be separate treatment for the unemployed or the sick. An alternative would be to preserve the categorical nature of social insurance but to make the insurance benefits more extensive in their coverage and sufficient to avoid the necessity to depend on public assistance or other forms of means-tested benefits. In considering the feasibility of such reforms, one must have regard both to the arithmetic of the redistribution and to the reasons why they have not been enacted in the past. As the ‘public choice’ school of public finance economists has stressed, the actions of the government are themselves to be explained by economic and other motives. The reasons why governments have failed to enact successful anti-poverty policies is a subject of great importance.

The policies discussed in this section are solely concerned with the poverty *within* countries, and would

do nothing to redistribute between countries. Indeed, some of the policies designed to help the low paid in advanced countries may actually have adverse consequences for low income countries. The income transfers which rich countries have so far made are of minuscule size when viewed against the magnitude of the problem of world poverty, and there can be little doubt that redistribution on a world scale is of the highest priority.

See Also

- equality
- Poor Law, new
- Poor Law, old

Bibliography

Abel-Smith, B. and Townsend, P. 1965. *The Poor and the Poorest*. London: Bell.

Atkinson, A.B., Maynard, A.K. and Trinder, C.G. 1983. *Parents and Children*. London: Heinemann.

Booth, C. 1892–7. *Life and Labour of the People of London*. 9 vols, London: Macmillan.

Bowley, A.L. 1913. Working class households in Reading. *Journal of the Royal Statistical Society* 76, June, 672–701.

Brandt, W. (Chairman). 1980. *North–South*. London: Pan.

Council of Economic Advisers. 1964. *Annual Report*. Washington, DC: Government Printing Office.

Eden, F.M., Sir. 1797. *The State of the Poor*. London: Cass.

Galbraith, J.K. 1958. *The Affluent Society*. Boston: Houghton Mafflin.

Harrington, M. 1962. *The Other America*. New York: Macmillan.

Institute for Research on Poverty. 1985. Antipoverty policy: past and future. *Focus*, Summer.

Lampman, R.J. 1959. *The Low Income Population and Economic Growth*. Study Paper No. 12, US Congress Joint Economic Committee. Washington, DC: Government Printing Office.

Orshansky, M. 1965. Who's who among the poor: a demographic view of poverty. *Social Security Bulletin* 28(7), July, 3–32.

Rowntree, B.S. 1901. *Poverty*. London: Macmillan.

Sen, A.K. 1976. Poverty: an ordinal approach to measurement. *Econometrica* 44(2), March, 219–31.

Tawney, R.H. 1913. *Poverty as an Industrial Problem*. London: London School of Economics.

Townsend, P. 1979. *Poverty in the United Kingdom*. London: Penguin.

World Bank. 1982. *World Development Report, 1982*. New York: Oxford University Press.

How to cite this article

Atkinson, A. B. "poverty." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000138> doi:10.1057/9780230226203.1317

The New Palgrave Dictionary of Economics Online

power laws

Xavier Gabaix
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A power law is the form taken by a remarkable number of regularities in economics, and is a relation of the type $Y = kX^{\alpha}$, where *Y* and *X* are variables of interest, α is called the power law exponent, and *k* is a constant. Many economic laws take the form of power laws, in particular macroeconomic scaling laws, the distribution of income, wealth, size of cities and firms, and the distribution of financial variables such as returns and trading volume. This article surveys the empirical evidence and the theoretical explanations for the occurrence of power laws.

Keywords

cities; GARCH effects; Gibrat's law; matching; networks; Pareto laws; power laws; proportional random growth; quantity theory of money; scaling laws; stock market volatility; stylized facts; superstars, economics of; trading volume; universality; urban economics; Zipf's law

Article

A power law (PL), also known as a scaling law, is the form taken by a remarkable number of regularities or ‘laws’ in economics, and is a relation of the type $Y = kX^{\alpha}$, where *Y* and *X* are variables of interest, α is called the power law exponent, and k is a typically unremarkable constant.

A special type is the distributional PL, also called a Pareto law. For instance, the probability that a firm has more than x employees is proportional to $1 / x^{\zeta}$, for some positive number ζ: $P(S > x) = k / x^{\zeta}$, for some *k*, at least in the upper tail or most of it. The exponent ζ is independent of the units in which the law is expressed. A special case is Zipf's law, which is a Pareto law with $\zeta \simeq 1$.

Understanding what gives rise to the scaling law, and explaining the precise value of the exponent (for example, why it is equal to 1 rather than any other number) is a challenge that has fascinated successive generations. Schumpeter (1949, p. 155) wrote: ‘Few if any economists seem to have realized the possibilities that such invariants hold for the future of our science. In particular, nobody seems to have realized that the hunt for, and the interpretation of, invariants of this type might lay the foundations for an entirely novel type of theory.’ Champernowne (1953) and Simon (1955) made great strides towards realizing Schumpeter's vision, and the quest continues.

Power laws are also of great interest outside of economics. Understanding PLs is a large part of the theory of critical phenomena, in which many materials behave identically around phase transitions – a phenomenon physicists call ‘universality’, and which is still only partially understood. Power laws have proven useful for describing and understanding networks. Biology has also many scaling regularities; for example, the daily energy intake of an animal of mass *M* is proportional to the $M^{3/4}$. This regularity was explained (Brown and West, 2000) via simple physical reasoning, which eschews the need to talk about the feathers and the hair of animals. Simpler and deeper principles underlie the regularities instead. The same holds for economic laws. Power laws give the hope of robust, detail-independent economic laws.

Theory: forces that generate power laws

Proportional random growth

Getting a power law. To explain distributional PLs, a central mechanism is proportional random growth (Sornette, 2001). The process was developed in economics by Champernowne (1953) and Simon (1955). Things are more tractable in continuous time (see Gabaix, 1999).

Take the example of cities in an economy with a constant number of cities and a fixed total population. When the system grows, the same reasoning applies after normalization – *S* is the normalized size of a city, for example as a multiple of the median city population. Suppose that each city *i* has a population S_t^i and, between *t* and $t+1$, increases by a growth rate Y_{t+1}^i:

$$S_{t+1}^i = Y_{t+1}^i S_t^i, \tag{1}$$

and suppose that the Y_{t+1}^i are identically and independently distributed, with density $f(Y)$, at least in the upper tail. Call $G_t(x) = P(S_t^i > x)$ the counter-cumulative distribution function. The equation of motion of *G* is:

$$G_{t+1}(x) = P\left(S_{t+1}^i > x\right) = P\left(S_t^i > x / Y_{it}\right) = E[G_t(x / Y_{it})].$$

Hence:

$$G_{t+1}(S) = \int_0^\infty G_t\left(\frac{S}{Y}\right) f(Y)\, dY.$$

Its steady state distribution *G*, if it exists, satisfies

$$G(S) = \int_0^\infty G\left(\frac{S}{Y}\right) f(Y)\, dY. \tag{2}$$

One can try the functional form $G(S) = a / S^\zeta$, where *a* is a constant. Plugging it in (2) gives: $1 = \int_0^\infty Y^\zeta f(Y)\, dY$, that is

$$E\left[Y^\zeta\right] = 1.$$

(3)

The steady state distribution is (in the upper tail) Pareto, with an exponent ζ that satisfies eq. (3).
To make sure that the steady state distribution exists, one needs some friction, for example a force that prevents small cities from becoming too small.
Getting a Zipf's law. We see that proportional random growth leads to a PL. Why should the exponent $\zeta = 1$ appear in so many economic systems? An answer is the following (see Gabaix, 1999; Luttmer, 2007; Rossi-Hansberg and Wright, 2007). Suppose that the random growth process (1) holds through most of the distribution, and that the system has constant size. Then, $E[S_{t+1}] = E[Y]E[S_t]$. As the system has constant size, then we need $E[S_{t+1}] = E[S_t]$, hence $E[Y] = 1$. That means that $\zeta = 1$ is a solution of eq. (3). In other words, to get Zipf's law we need a random growth process with small frictions.
In sum, proportional random growth with frictions leads to PLs, and proportional random growth with small frictions leads to a special type of PL, namely Zipf's law.

Inheritance via algebraic transformation

Power laws have excellent inheritance and aggregation properties. The property of being distributed according to a PL is conserved under addition, multiplication, power transformation, min, and max. The general rule is that, when we combine two PL variables, the fatter-tailed (that is, the one with the smaller exponent) dominates. Call ζ_X the PL exponent of *X*, with $\zeta_X = +\infty$ if *X* is thinner than any PL, for example is a Gaussian. For *X* and *Y* independent random variables, and $\theta > 0$ a constant, we have: $\zeta_{X+Y} = \zeta_{X \cdot Y} = \zeta_{\max(X,Y)} = \min(\zeta_X, \zeta_Y)$, $\zeta_{\min(X,Y)} = \zeta_X + \zeta_Y$, $\zeta_{\alpha X} = \zeta_X$, $\zeta_{X^\alpha} = \zeta_X / \alpha$ (see Jessen and Mikosch, 2006). Those properties generate new PLs from old ones. For instance, if mutual funds are PL distributed, then many of their actions (for example, trading volumes, or the price movements they create) will be PL distributed (Gabaix et al., 2006).

Equilibrium economic mechanisms

Optimization with PL objective function. The early example is the Allais–Baumol–Tobin model of demand for money (see also Mulligan and Shleifer, 2005; Gabaix et al., 2003). Costs and benefits are power functions of the variables of interest, so that maximization also yields a PL – there, money demand is proportional to the interest rate to the power $-1/2$. PL in, PL out.
Matching talents in the upper tail. Another way to generate PLs is in matching the talent of individuals with large firms or audiences. For instance, Gabaix and Landier (2008) study the market for executives. They derive that, in the upper tail of all well-behaved distributions, if $T(x)$ is the talent of an individual in the x upper quantile, then $T'(x)$ is approximately a power function x^α. As a result, the competitive matching process generates a PL relation between CEO pay and firm size, and a PL of the pay distribution. Huge differences in pay reward minuscule differences in talent. The PL form of T' is likely to be useful in other superstars markets.

Empirics: the main power laws of economics

Old macroeconomic scaling laws

The first quantitative law of economics is probably the quantity theory of money, which, not coincidentally, is a scaling relation. It states that the price level *P* is proportional to the mass of money in circulation *M*, divided by the gross domestic product *Y*, times a pre-factor *V*: $P = VM/Y$. If the money supply doubles while GDP remains constant, prices double – a nice scaling law, relevant to policy.
More modern, we have Kaldor's stylized facts on economic growth: with *K* the capital stock, *Y* GDP, *L* population, *r* the interest rate, K/Y, wL/Y, and *r*, are roughly constant across time and countries. Explaining these facts led Solow to his growth model.

Reasonably old and well-established laws

Income and wealth. The first PL is the Pareto law of income or wealth, which states that the tail distribution of income (or, respectively, wealth), is PL. The tail exponent of income

seems to vary between 1.5 and 3, while the tail exponent of wealth is more stable. While, starting with Champernowne (1953), many models have been proposed to explain it (mainly along the lines of random growth), it is intriguingly unclear why the exponent is rather stable across economies.

Firm sizes. The bulk of the distribution of firm sizes is well described by a Zipf's law (Figure 1). This severely constrains models of firm growth, and means that idiosyncratic shocks of large firms may affect GDP (Gabaix, 2006). Zipf's law holds for different measures of firm sizes and countries (Axtell, 2001; Fujiwara et al., 2004; Gabaix and Landier, 2008).

Figure 1

Note: Log frequency $\ln f(S)$ vs. log size ln *S* of US firm sizes for 1997. OLS fit gives a slope of $1 + \zeta = 2.059$(s.e.=0.054; R^2=0.992). This corresponds to a frequency $f(S) = kS^{-2.059}$, that is, a power law distribution with exponent $\zeta = 1.059$. Indeed, if $P(\text{Size} > S) = kS^{-\zeta}$ the density is $f(S) = k\zeta S^{-(\zeta+1)}$. This is very close to Zipf's law, which says that $\zeta = 1$. *Source*: Reprinted with permission in Fig. 1 from Robert L. Axtel, *Science* 293, 1818–20 (7 September 2001).

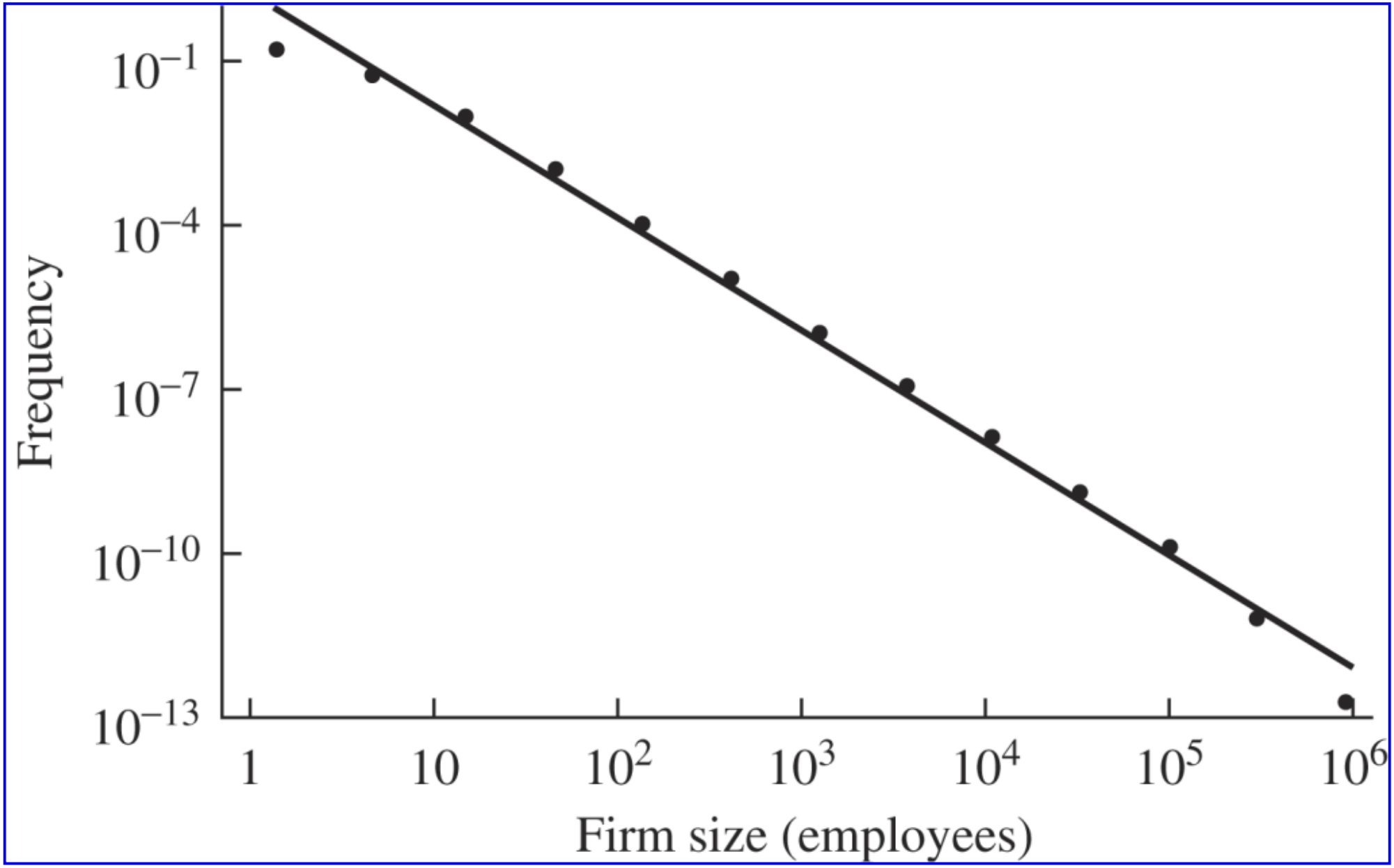

City sizes. In the upper tail, Zipf's law holds generally well across times and countries (Gabaix and Ioannides, 2004).
Gibrat's law for the growth rate of cities is shown in the United States by Ioannides and Overman (2003).
Roberts's law for executive compensation. Across times and countries, an executive heading a firm of size *S* earns an amount proportional to S^{κ}, for a κ around $1/3$. Superstars models explain the presence of this scaling (Gabaix and Landier, 2006), but the reason for the 1/3 value remains a mystery.

More recently proposed laws

Power law of stock market activity: returns, trading volume and trading frequency. Following Mandelbrot, the following regularities have been found. Stock market returns (over one minute to one week) have PL tails, with an exponent around three (Gopikrishnan et al., 1999). Individual trades have a PL exponent around 1.5 (Gopikrishnan et al., 2000). The number of trades executed over a short horizon has an exponent close to three (Plerou et al., 2000). There is no consensus about the origins for those regularities. The fat tails of the returns might come from GARCH effects. One view (Gabaix et al., 2003; 2006) attributes it to the trades of large institutional investors in relatively illiquid markets, which creates spikes in returns and volume, and generates empirically found exponents.
Supply of regulations. Mulligan and Shleifer (2005) establish another candidate law. In US states, the quantity of regulation is a PL of population.

Estimation of power laws

How does one estimate a distributional PL? We take the example of *n* cities in the upper tail, ordering them by size, $S_{(1)} \geq \cdots \geq S_{(n)}$. One method is Hill's estimator:

$$\hat{\zeta}^{Hill} = (n-1) / \sum_{i=1}^{n-1} \left(\ln S_{(i)} - \ln S_{(n)}\right)$$

which has a standard error $\hat{\zeta}^{Hill} n^{-1/2}$. The second method is a ‘log rank log size regression’, where $\hat{\zeta}$ is the slope in the regression of the log rank *i* on the log size:

$$\ln(i-s) = \text{constant} - \hat{\zeta}^{OLS} \ln S_{(i)} + \text{noise}$$

which has a standard error of $\hat{\zeta}^{OLS} \cdot (n/2)^{-1/2}$. *s* is a shift, $s = 0$ is typical, but $s = 1/2$ is optimal (Gabaix and Ibragimov, 2006). Both methods have pitfalls, as true errors are often larger than nominal standard errors (Embrechts, Kluppelberg and Mikosch, 1997; Gabaix and Ioannides, 2004).

See Also

- ARCH models
- econophysics
- inequality (measurement)
- quantity theory of money
- superstars, economics of
- systems of cities

- wealth

Bibliography

Axtell, R. 2001. Zipf distribution of U.S. firm sizes. *Science* 293, 1818–20.

Brown, J.H. and West, G.B. 2000. *Scaling in Biology*. Oxford: Oxford University Press.

Champernowne, D. 1953. A model of income distribution. *Economic Journal* 83, 318–51.

Embrechts, P., Kluppelberg, C. and Mikosch, T. 1997. *Modelling Extremal Events for Insurance and Finance*. New York: Springer.

Fujiwara, Y., Di Guilmi, C., Aoyama, H., Gallegati, M. and Souma, W. 2004. Do Pareto–Zipf and Gibrat laws hold true? An analysis with European firms. *Physica A* 335, 197–216.

Gabaix, X. 1999. Zipf's law for cities: an explanation. *Quarterly Journal of Economics* 114, 739–67.

Gabaix, X. 2006. The granular origins of aggregate fluctuations. Working paper, MIT.

Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H.E. 2003. A theory of power law distributions in financial market fluctuations. *Nature* 423, 267–30.

Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H.E. 2006. Institutional investors and stock market volatility. *Quarterly Journal of Economics* 121, 461–504.

Gabaix, X. and Ibragimov, R. 2006. Log (Rank-1/2): a simple way to improve the OLS estimation of tail exponents. Working paper, Harvard University.

Gabaix, X. and Ioannides, Y. 2004. The evolution of the city size distributions. In *Handbook of Regional and Urban Economics*, vol. 4, ed. V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Gabaix, X. and Landier, A. 2008. Why has CEO pay increased so much? *Quarterly Journal of Economics* 123.

Gopikrishnan, P., Plerou, V., Amaral, L., Meyer, M. and Stanley, H.E. 1999. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E* 60, 5305–316.

Gopikrishnan, P., Plerou, V., Gabaix, X. and Stanley, H.E. 2000. Statistical properties of share volume traded in financial markets. *Physical Review E* 62, R4493–96.

Ioannides, Y.M. and Overman, H.G. 2003. Zipf's law for cities: an empirical examination. *Regional Science and Urban Economics* 33(2), 127–37.

Jessen, A.H. and Mikosch, T. 2006. Regularly varying functions. *Publications de l'Institut Mathématique, Nouvelle Série* 80, 171–92.

Luttmer, E.G.J. 2007. Selection, growth, and the size distribution of firms. *Quarterly Journal of Economics* 122(3).

Mulligan, C. and Shleifer, A. 2005. The extent of the market and the supply of regulation. *Quarterly Journal of Economics* 120, 1445–73.

Plerou, V., Gopikrishnan, P., Amaral, L., Gabaix, X. and Stanley, H.E. 2000. Economic fluctuations and anomalous diffusion. *Physical Review E* 62 (3), R3023–6.

Rossi-Hansberg, E. and Wright, M. 2007. Urban structure and growth. *Review of Economic Studies* 74, 597–624.

Schumpeter, J. 1949. Vilfredo Pareto (1848–1923). *Quarterly Journal of Economics* 63, 147–72.

Simon, H. 1955. On a class of skew distribution functions. *Biometrika* 44, 425–40.

Sornette, D. 2001. *Critical Phenomena in Natural Sciences*. New York: Springer.

How to cite this article

Gabaix, Xavier. "power laws." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000324> doi:10.1057/9780230226203.1322

The New Palgrave Dictionary of Economics Online

power

Samuel Bowles and Herbert Gintis

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We consider the exercise of power in competitive markets for goods, labour and credit. We offer a definition of power and show that if contracts are incomplete it may be exercised either in Pareto-improving ways or to the disadvantage of those without power. Contrasting conceptions of power including bargaining power, market power, and consumer sovereignty are considered. Because the exercise of power may alter prices and other aspects of exchanges, abstracting from power may miss essential aspects of an economy. The political aspect of private exchanges challenges conventional ideas about the appropriate roles of market and political competition in ensuring the efficiency and accountability of economic decisions.

Keywords

bargaining power; Coase, R.H.; consumer sovereignty; firm, theory of; incomplete contracts; labour market contracts; labour market search; market power; monopolistic competition; Nash equilibrium; Pareto efficiency; power; principal and agent; purchasing power; rent; reservation wage; sanctions; short-side power; technical efficiency

Article

Power is exercised in the competitive markets for goods, labour and credit. We consider this aspect of economic power, setting aside the widely recognized exercise of power by members of governments and other coercive bodies and the influence of economic groups on governmental policy.

Background

‘An economic transaction is a solved political problem …’, wrote Abba Lerner (1972, p. 259) ‘… economics has gained the title Queen of the Social Sciences by choosing solved political problems as its

domain'. Prior to the development of modern contract theory, the standard approach to power among economists was aptly summed up by Paul Samuelson (1957, p. 894), 'Remember that in a perfectly competitive market, it really does not matter who hires whom; so have labor hire capital'. As if responding to Samuelson, John Kenneth Galbraith (1967, p. 47), chided economists for not having asked 'why power is associated with some factors [of production] and not with others?' But with some notable exceptions (for example, Zeuthen, 1930; Shapley and Shubik, 1967; Samuels, 1973; Lindblom 1977; Basu, 1986; Takada, 1995; Hirshleifer, 1991; Chichilnisky and Heal, 1984; Lundberg and Pollak, 1994; Rotemberg, 1993; Pagano, 1999; Bardhan, 2005; Aghion and Tirole, 1997) economists have treated power as the concern of other disciplines and extraneous to economic explanation. The term does not appear among the 1,300 or so index entries of the leading graduate microeconomics text (Mas-Colell, Whinston and Green, 1995).

The reason is that Samuelson's claim is true in the Walrasian model: if contracts are complete, 'hiring' simply means 'buying'. 'What does it mean', Oliver Hart (1995) asked, 'to put someone "in charge" of an action or decision if all actions can be specified in a contract?' But as an empirical matter, as Marx (1867), Coase (1937), Simon (1951) and others have stressed, the firm is a political institution in the sense that some members of the firm routinely give commands while others are constrained by the threat of sanctions to obey. To say that the manager has the right to decide what the worker will do means only that he has the legitimate authority to do this, not the power to secure compliance. Given that in a liberal economy management is sharply restricted in the kinds of punishment they can inflict, and given that the employee is free to leave, the fact that orders are typically obeyed is a puzzle. Why, in Coase's initial formulation, is the command of the manager (to move 'from department Y to department X') obeyed (Coase, 1937)?

Noticing the lack of a good answer, Alchian and Demsetz (1972) challenged the Coasean idea that the firm is a mini 'command economy', suggesting that the employment contract is no different in this respect from other contracts.

> The firm … has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people…. Wherein then is the relationship between a grocer and his employee different from that between a grocer and his customer? (1972, p. 777)

Hart (1989, p. 1771) offered the following response to Alchian and Demsetz:

> … the reason that an employee is likely to be more responsive to what his employer wants than a grocer is that the employer…can deprive the employee of the assets he works with and hire another employee to work with these assets, while the customer can only deprive the grocer of his customer and as long as the customer is small, it is presumably not very difficult for the grocer to find another customer.

Hart motivates the difference between the grocer and the employer by the assumption that the employee needs access not just to a job (and hence *some* assets) but to *this particular employer's assets*. This might

be the case due to a complementarity between the two (the employee may have made an investment in training which is of value only when combined with this particular asset, for example). Other less obvious (and probably more important) examples come to mind. Excluding an employee from access to a particular asset may require the employee to relocate, disrupting family and friendships. The loss of a job may also harm the employee's reputation.

While transaction-specific investments of this type undoubtedly explain some authority relationships – in company towns, and for some professional jobs and managers, for example – the explanation seems insufficiently general to provide an adequate explanation of the entire authority structure of the firm, especially in large urban labour markets and for non-professional employees. We thus need a complementary explanation based on the fact that the employee excluded from access to *her current employer's asset* may not find access to *any asset* even in a competitive economy in which transaction-specific assets are absent. This will require clarity about what we mean by power.

Power as a political means to gain economic advantage in private exchange

Because of its close connection to value-laden words such as ‘coercion’ and ‘freedom’ the term itself has proven to be controversial among philosophers and political theorists (Nozick, 1969; Lukes, 1974; Bachrach and Baratz, 1962; Barry, 1976; Taylor, 1982). Nonetheless, common usage suggests several characteristics that must be present when power is said to be exercised. First, power is *interpersonal*, an aspect of a relationship among people, not a characteristic of a solitary individual. Second, the exercise of power involves the *threat and use of sanctions*. Indeed, many political theorists regard sanctions as the defining characteristic of power.

Lasswell and Kaplan (1950, p. 75) make the use of ‘severe sanctions … to sustain a policy against opposition’ a defining characteristic of a power relationship, and Parsons (1967, p. 308) regards ‘the presumption of enforcement by negative sanctions in the case of recalcitrance’ a necessary condition for the exercise of power. Third, the concept of power should be *normatively indeterminate*, allowing for Pareto-improving outcomes (as has been stressed by students of power from Hobbes to Parsons), but also susceptible to abuse in ways that harm others in violation of ethical principles. Finally, power must be *sustainable as a Nash equilibrium* of an appropriately defined game. Power may be exercised in disequilibrium situations, of course, but, as an enduring aspect of social structure, it should be a characteristic of an equilibrium.

The following sufficient condition for the exercise of power captures these four desiderata: *For B to have power over A, it sufficient that, by imposing or threatening to impose sanctions on A, B is capable of affecting A's actions in ways that further B's interests, while A lacks this capacity with respect to B* (Bowles and Gintis,1992).

The fact that sanctions are essential to the exercise of power in our sense makes it distinct from other means of influencing the behaviour of others that may operate even in the complete absence of strategic interaction, as in a Walrasian market setting. Consider, for example, the standard definition due to Robert Dahl (1957, pp. 202–3): ‘A has power over B to the extent that he can get B to do something that B would not otherwise do.’ But one can affect the behaviour of another in ways that do not involve power in the usual sense of that term. If we buy a commodity, there will be a whole series of market effects through the economy which entail others doing things they would not otherwise have done. But

to say that our purchase of bread is an exercise of power over some unknown wheat farmer with whom we do not interact strategically is to expand the concept of power beyond recognition. By making the threat of sanctions a necessary aspect of power we also exclude forms of interpersonal influence such as persuasion and the provision of information.

Short-side power in labour, credit and goods markets

The power that may be exercised by an economic actor depends on the actor's position in the institutions of society. Power may be exercised by economic actors who are on the short side of a non-clearing market, namely, the side of the market on which the number of desired transactions is less, that is, employers in a labour market with unemployment, lenders in a loan market with borrowers facing credit constraints, and so on. Because those holding power in these cases are those on the short side of the market, we term this ‘short-side power’. This clarifies the difference between the employer and the grocer in Hart's response to Alchian and Demsetz: the sanctions imposed on the employee by depriving him of access to the capital good are severe because, in a labour market with perpetual excess supply of labour, finding another job will be difficult, while the costs imposed on the grocer by the departing customer are negligible or zero. The reason why the consumer, in switching to another seller, does not impose a sanction on the grocer is that the grocer (in competitive equilibrium) was maximizing profits by selecting a level of sales that equates marginal cost to the exogenously given price, and, this being the case, a small variation in sales has only a second-order effect on profits.

Let us check to see that this conception of power applies to the employment relationship in which transaction specificity is absent. We know that in a standard labour-discipline model (Gintis, 1976; Shapiro and Stiglitz, 1984; Bowles, 1985), in equilibrium the worker receives a rent: the present value of the job exceeds her next-best alternative (job search) and, because she fears losing his job, she works harder than she would have in the absence of the employer's incentive strategy. These results together imply that the employer has caused the worker to act in the employer's interest by credibly threatening to sanction the worker. The employee lacks this capacity with respect to the employer for, were the employee to threaten the employer with a sanction should he not raise the wage (to damage his machinery or beat him up or simply to work less hard), the threat would not be credible. The employer would simply refuse to respond, knowing that it would not be in the interest of the employee to carry out the threat.

Note that the exercise of power allows a Pareto improvement over a counterfactual condition in which power cannot be exercised, namely, that the worker is hired at her reservation wage and works at the reservation effort level. This follows directly as we know from the fact that the worker receives an equilibrium rent at the wage offered by the employer. Both expected worker lifetime utility and firm profits are higher in equilibrium (with power being exercised) than at the (power-absent) reservation position. This is yet another example of a situation in which the exercise of power helps to address coordination failures, albeit sometimes with objectionable consequences those without power. An example from Bowles (2004) follows.

Suppose the employer determines (in addition to the wage) some aspect of the job affecting workplace amenities, including not only such innocuous things as the quality of the music on the office sound system but also management practices affecting the employee's dignity, such as not being subjected to racial insults, sexual harassment or other on-the-job indignities. If the firm sets these amenities to

maximize profits, it follows that the employer can inflict first-order costs on the worker (by reducing the amenity a small amount) at second-order cost to himself (the costs are second-order because due to profit maximization the derivative of profits with respect to the level of amenities is zero). Thus the competitive equilibrium in an employment relationship gives the employer the capacity not only to exercise power to attenuate coordination problems but also to exercise power arbitrarily, that is, to inflict costs on another at virtually no cost to himself. When this power is exercised in unethical ways it may be termed coercive.

Thus the strategic interaction between the employer and employee allows the exercise of power in a manner conforming to the four desiderata outlined above: sanctions are credibly threatened (and used) in a strategic interaction describing a Nash equilibrium, and the resulting exercise of power is Pareto-improving over a reasonable counterfactual but may also be used coercively.

It is easy to check that power in the sense defined may be exercised in the standard principal–agent model of the credit market as well. The lender offers the borrower terms that are preferred to the borrower's reservation position, promising to make additional loans in the future if the borrower repays the loan. In this contingent renewal model, the borrower pursues a less risky strategy than would have been the case had the lender not offered a rent. Where the borrower's participation constraint holds as an equality, power in the sense defined cannot be exercised for the simple reason that the borrower is indifferent between the current transaction and the next-best alternative, so the only sanction permitted in a liberal economy – termination of the contract – has no force.

Short-side power may be contrasted with the ‘markets and hierarchies’ approach pioneered by Oliver Williamson (1985). Rather than seeing firms simply as ‘islands of conscious power in this ocean of unconscious cooperation’, in Robertson's (1923, p. 85) apt words, the incomplete contracts approach traces the exercise of power to both the structure of markets and the structure of firms. The firm is an important venue in which power is exercised, but, as the credit market model makes clear, power may be exercised in the absence of firms or indeed any organizational structure whatsoever. Short-side power is exercised *in* markets, not simply outside markets or despite markets.

Wealth, power and ‘consumer sovereignty’

Thus an agent's location in the economic structure of a society – on the short side of a non clearing market – may make it possible for him to exercise power over others. How are agents assigned to these positions of short-side power? Given that employing others requires capital and that borrowing substantial amounts typically requires that the borrower have sufficient wealth to invest in the project or to provide collateral, an important determinant of an individual's assignment to a position of short-side power is the individual's wealth. The wealthy may exercise power over those to whom they lend, who in turn may exercise power over those (managers or other employees) whom they hire. As a result, power cascades downward from the loan market to the market for managers to the market for non-managerial employees (Bowles, 2004).

A less obvious case concerns the power of the consumer, sometimes summarized by the term ‘consumer sovereignty’. Consider a principal–agent model involving difficult-to-measure product quality (Klein and Leffler, 1981; Gintis, 1976). In equilibrium, the buyer pays the seller a price exceeding the seller's next-best alternative and promises continued purchases contingent on the seller providing high-quality

goods. The seller's prospect of losing the resulting rent conferred by the buyer induces the seller to provide higher quality than would have been provided in the absence of the threatened sanction. Thus the buyer has exercised power over the seller in the sense just defined.

As the example suggests, buyers may exercise power over sellers whenever the buyer's threat to switch to an alternative seller is credible and inflicts a cost on the seller. Consider two monopolistically competitive sellers (that is, firms facing downward-sloping demand functions) and a consumer who is indifferent between purchasing from one or the other. Both sellers have chosen a level of output to maximize profits, setting marginal cost equal to marginal revenue (which is less than the price because the demand curve is downward sloping). For both sellers, price thus exceeds marginal cost, and as a result the consumer's choice confers a rent on one and deprives the other of the rent. The reader may wonder how the rent can arise if the firm has chosen the output level to maximize profits, each setting the derivative of profits with respect to sales equal to zero. But the buyer's switch from one to the other seller is not a movement *along* a demand function (the basis of the firm's output choice), but rather is a horizontal *shift* in the demand function (inwards for the firm the consumer rejected, outwards for the firm to which he switched). As a result of the switch, for the fortunate firm it is profit maximizing to sell one more unit at the going price.

Ironically, the idealized Walrasian conditions under which consumer sovereignty is said to hold give the consumer no power in the sense defined here, while deviations from the canonical competitive assumption that price equals marginal cost (because firms face downward sloping demand functions) create an environment in which the consumer may exercise power. Of course, the strategic position of the consumer as one of many principals facing a single agent is quite unlike that of the employer facing many potential employees or the lender facing many potential borrowers. As Hart observed about the consumer and the grocer, a single consumer will not generally be in a position to command the supplier to improve the product quality and expect the supplier to obey. The power of consumers is thus limited by the difficulties the many principals face in acting in a coordinated fashion.

Non-clearing markets and inefficient competitive equilibria

Where power is exercised by a principal who confers a rent on an agent and monitors the agent's actions – as in the markets for labour, credit and goods just analysed – the equilibrium allocation will generally be neither Pareto-efficient nor technically efficient. The reason for the first is that the principal is constrained not by the agent's reservation utility but by the agent's best-response function. As a result, small changes in the instruments controlled by the principal – the wage, the rate of interest or the price – incur only second-order costs or benefits for the principal but first-order benefits and costs for the agent. For the actions controlled by the agent the reverse is true. Therefore, there must exist some set of small variations away from the equilibrium allocation that improve the utility of both principal and agent. A labour market example of such a Pareto improvement is a small increase in the wage accompanied by a small increase in worker effort.

The allocation will be technically inefficient because the principal chooses the enforcement strategy with respect to the private costs (the costs of both the rent conferred on the agent and the monitoring) while there is no social cost associated with the rent (because, unlike the monitoring costs, it is a pure transfer and is not resource using). From the equilibrium allocation, therefore, there must exist a technical efficiency-improving increase in the agent's rent and a reduction in monitoring.

Exploiting these potential efficiency gains requires changes in the information and incentive structure of the interaction, for example by making the agent the residual claimant on his or her non-contractible actions, if this is possible.

The three cases for which we have analysed the exercise of power – by the buyer over the seller, the lender over the borrower, and the employer over the employee – are members of a generic class of power relationships which are sustainable in the equilibrium of a system of voluntary competitive exchanges. In all three, those with power are transacting with agents who receive rents and hence are not indifferent between the current transaction and their next-best alternative. This being the case, there must exist other identical agents who are quantity constrained, namely, the unemployed, those excluded from the loan market or restricted in the amount they can borrow, and sellers who fail to make a sale. For this situation to characterize an equilibrium it must be that markets do not clear, which, as we have seen will be the case.

Power as we have defined it can be exercised in other ways, even when markets clear. An interesting (if perhaps not empirically important) example is provided by the case of optimal job fees, in which the fee eliminates the job rent *ex ante* so the market clears, the worker being indifferent between taking the job and paying the fee or not. But an *ex post* rent nonetheless exists, giving the employer the ability to sanction the employee. A job fee of this type is a pure case of an employee's transaction-specific investment, and the basis of the power of the employer in this case is an example of Hart's reasoning, above.

All three of those exercising power in the above examples – buyer, lender, employer – have in common that the party that contributes money to the transaction – the buyer's purchase price, the lender's loan, the employer's wage offer – is the one exercising power. This may seem an analytical foundation for the familiar adage that ‘money talks’, but the conclusion is misleading. Recall that in the centrally planned Communist economies it was generally the case that consumer durables (and many other consumer goods) sold below market-clearing prices. The resulting excess demand was allocated through a process of queuing and by other means (Kornai, 1980). In this case the producers (sellers) were on the short side of the market, and those bringing money to the transaction – the buyers – were the long-siders, some of whom failed to make a trade. The notorious inferiority in the quality of consumer goods in centrally planned economies to those in capitalist economies may be explained in part by the fact that consumers were long-siders in the former and short-siders in the latter. Or, to put it more graphically, one reason why Fords were better cars than their Cold War era Russian equivalents is that in Russia customers waited in line to purchase Volgas while in the United States Ford salesmen lined up to sell customers cars. Another reason is that in the United States workers waited in line to get jobs at Ford.

Other conceptions of power

Other uses of the term ‘power’ are common in economics. (We do not address the concept of ‘coalitional power’ advanced by Shapley and Shubik, 1954, as it has found application primarily in the analysis of committees voting and other arenas addressed by political scientists.) ‘Purchasing power’ is just another word for the position of one's budget constraint (or wealth), and it does not concern the exercise of sanctions or indeed any strategic interaction at all. ‘Market power’ arises in thin markets in which an actor can benefit by varying a price. In the standard monopolistic competition case the seller is said to have market power. The seller is less constrained in the sense that he faces a downward sloping

demand rather than horizontal demand function, while the consumer is more constrained in that there may be less choice among suppliers. But we have just seen that in this case the consumer who switches from one seller to another confers a rent on his favoured firm. (This why Ford salesmen line up to sell you cars.) Thus, if the buyer can credibly threaten to withdraw the rent he may be able to exercise short-side power over the seller. It thus is not clear how to reconcile usual notions of power – the use of sanctions to gain advantage – with the statement that the monopolist has power over the consumer.
Finally, there is ‘bargaining power’, typically meaning the share of the joint surplus which a party gains in a bargain (Binmore, Rubinstein and Wolinsky, 1986). Reflecting this usage, the exponents used in the ‘Nash product’ to solve the generalized Nash bargaining model are said to refer to the bargaining power of the two parties. Used this way, bargaining power refers to outcomes – to how much advantage one may gain – rather than to any particular means of attaining it (for example by threatening a sanction). If the bargaining problem is embedded in an ongoing interaction, then bargaining power and short-side power appear not only unrelated but even opposed. In the competitive equilibrium of the standard principal–agent model of the labour market, for example, the principal receives his reservation return (given by the zero profit condition) while the agent receives a rent. Therefore, the bargaining-power perspective would say that the employee has *all* the bargaining power. But the short-side power perspective would conclude that, far from a sign that the employee is powerful, the rent conferred on the employee as a profit-maximizing choice of the employer is the reason why the employer has power over the employee. The employee receives the rent because his services cannot be costlessly contracted for, and the employer profits in this case by paying to exercise power over the employee.
The fact that the exercise of power is ubiquitous in private exchange shows that it is mistaken to think of society as composed of a political sphere, meaning governments and other bodies with formal powers of coercion, and a private economic sphere in which the exercise of power is absent. The rejection of this public–private division raises important issues concerning the appropriate scope of for democratic political competition (in addition to market competition) as a guarantor of accountability in the economy (Dahl, 1977; Bowles and Gintis, 1993).

See Also

- bargaining
- labour market institutions

We are grateful to Maria Olivella Rizza, Ceren Soylu, Elisabeth Wood and Peyton Young for comments, and to the Behavioral Sciences Program of the Santa Fe Institute, the University of Siena and the MacArthur Foundation for financial support.

Bibliography

Aghion, P. and Tirole, J. 1997. Formal and real authority in organizations. *Journal of Political Economy* 105, 1–29.

Alchian, A.A. and Demsetz, H. 1972. Production, information costs, and economic organization. *American Economic Review* 62, 777–95.

Bachrach, P. and Baratz, M. 1962. The two faces of power. *American Political Science Review* 56, 947–52.

Bardhan, P. 2005. *Scarcity, Conflicts and Cooperation*. Cambridge, MA: MIT Press.

Barry, B., ed. 1976. *Power and Political Theory: Some European Perspectives*. New York: John Wiley.

Basu, K. 1986. One kind of power. *Oxford Economic Papers* 38, 259–82.

Binmore, K., Rubinstein, A. and Wolinsky, A. 1986. The Nash bargaining solution in economic modelling. *RAND Journal of Economics* 17, 176–88.

Bowles, S. 1985. The production process in a competitive economy: Walrasian, neo-Hobbesian, and Marxian models. *American Economic Review* 75, 16–36.

Bowles, S. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton: Princeton University Press.

Bowles, S. and Gintis, H. 1992. Power and wealth in a competitive capitalist economy. *Philosophy and Public Affairs* 21, 324–53.

Bowles, S. and Gintis, H. 1993. A political and economic case for the democratic enterprise. *Economics and Philosophy* 9, 75–100.

Chichilnisky, G. and Heal, G. 1984. Patterns of power: bargaining and incentives in two-person games. *Journal of Public Economics* 23, 333–49.

Coase, R.H. 1937. The nature of the firm. *Economica* 4, 386–405.

Dahl, R.A. 1957. The concept of power. *Behavioral Science* 2, 201–15.

Dahl, R.A. 1977. On removing certain impediments to democracy in the United States. *Political Science Quarterly* 92, 1–20.

Galbraith, J.K. 1967. *The New Industrial State*. Boston: Houghton Mifflin.

Gintis, H. 1976. The nature of the labor exchange and the theory of capitalist production. *Review of Radical Political Economics* 8(2), 36–54.

Gintis, H. 1989. The power to switch: on the political economy of consumer sovereignty. In *Unconventional Wisdom: Essays in Honor of John Kenneth Galbraith*, ed. S. Bowles, R. Edwards and W.G. Shepherd. New York: Houghton-Mifflin.

Hart, O. 1989. An economist's perspective on the theory of the firm. *Columbia Law Review* 89, 1757–74.

Hart, O. 1995. *Firms, Contracts, and Financial Structure*. Oxford: Clarendon Press.

Hirshleifer, J. 1991. The paradox of power. *Economics and Politics* 3, 177–200.

Klein, B. and Leffler, K. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89, 615–41.

Kornai, J. 1980. *Economics of Shortage*. Amsterdam: North-Holland.

Lasswell, H. and Kaplan, A. 1950. *Power and Society: A Framework for Political Enquiry*. New Haven, CT: Yale University Press.

Lerner, A. 1972. The economics and politics of consumer sovereignty. *American Economic Review* 62, 258–66.

Lindblom, C.E. 1977. *Politics and Markets: The World's Political-Economic Systems*. New York: Basic Books.

Lukes, S. 1974. *Power: A Radical View*. London: Macmillan.

Lundberg, S. and Pollak, R. 1994. Noncooperative bargaining models of marriage. *American Economic Review* 84, 132–7.

Marx, K. 1867. *Capital: A Critique of Political Economy. I: The Process of Capitalist Production*. New York: International Publishers, 1967.

Mas-Colell, A., Whinston, M.D. and Green, J.R. 1995. *Microeconomic Theory*. New York: Oxford University Press.

Nozick, R. 1969. Coercion. In *Philosophy, Science and Method*, ed. S. Morgenbesser, P. Suppes and M. White. New York: St. Martins Press.

Pagano, U. 1999. Is power an economic good? Notes on social scarcity and the economics of positional goods. In *The Politics and Economics of Power*, ed. S. Bowles, M. Franzini and U. Pagano. London: Routledge.

Parsons, T. 1967. On the concept of political power. In *Sociological Theory and Modern Society*, ed. T. Parsons. New York: Free Press.

Robertson, D. 1923. *The Control of Industry*. Cambridge: Cambridge University Press.

Rotemberg, J.J. 1993. Power in profit-maximizing organizations. *Journal of Economics & Management Strategy* 2, 165–98.

Samuels, W.J. 1973. The economy as a system of power and its legal bases: the legal economics of Robert Lee Hale. *University of Miami Law Review* 27, 262–371.

Samuelson, P. 1957. Wages and interest: a modern dissection of Marxian economics. *American Economic Review* 47, 884–921.

Shapiro, C. and Stiglitz, J. 1984. Unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Shapley, L.S. and Shubik, M. 1954. A method for evaluating the distribution of power in a committee system. *American Political Science Review* 48, 787–92.

Shapley, L.S. and Shubik, M. 1967. Ownership and the production function. *Quarterly Journal of Economics* 81, 88–111.

Simon, H. 1951. A formal theory of the employment relation. *Econometrica* 19, 293–305.

Takada, Y. 1995. *Power Theory of Economics*. New York: St. Martin's Press.

Taylor, M. 1982. *Community, Anarchy, and Liberty*. Cambridge: Cambridge University Press.

Williamson, O.E. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

Zeuthen, F. 1930. Economic warfare. In *Problems of Monopoly and Economic Warfare*. New York: A. M. Kelly, 1968.

How to cite this article

Bowles, Samuel and Herbert Gintis. "power." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000343> doi:10.1057/9780230226203.1321

The New Palgrave Dictionary of Economics Online

Prebisch, Ra ú l (1901 – 1986)

José Gabriel Palma
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Prebisch was concerned with four stylized facts of commodity-exporting middle-income countries: (*a*) failure to ‘catch-up’; (*b*) recurrent balance of payments disequilibrium; (*c*) unstable and deteriorating terms of trade; and (*d*) persistent unemployment. At the core of Prebisch's analysis lies his differentiation of the economic and export structures of the centre and periphery. Those of the centre were seen as *homogeneous* and *diversified*, those of the periphery as *heterogeneous* and *over-specialized*. Prebisch associates the second and third problems above primarily with export over-specialization (too few homogeneous, unbranded and price-volatile commodities), and the first and fourth with structural heterogeneity (which hindered industrialization).

Keywords

balance of payments; commodity-exporting countries; comparative advantage; cooperative games; excess demand; Kaldor, N.; import substitution; industrialization; long-term economic growth; periphery; Prebisch, R.; Robinson, J. V.; structuralism; terms of trade

Article

Prebisch was born on 17 April 1901 in Tucumán, Argentina, and died at the age of 84 in Santiago de Chile. He graduated in Economics at the University of Buenos Aires in 1923 having already published six papers in academic journals.

He was Professor of Political Economy at the University of Buenos Aires from 1925 to 1948. In 1930 he became Under-Secretary of Finance at the age of 29, and soon afterwards the first Director General of the Argentine Central Bank (1935–43). He then moved to the United Nations, being appointed Executive Secretary of the Economic Commission for Latin America and the Caribbean (ECLAC) in 1950. In 1963 he moved to the United Nations Conference on Trade and Development (UNCTAD) as its first Secretary General.

Although his main intellectual concern was always the understanding of the specific development obstacles facing commodity-exporting, middle-income, peripheral countries, he always acknowledged that early in his career he had viewed them from a mainstream perspective. His approach only changed when he witnessed the Great Depression (including the heterodox response to it of many industrialized countries) and read the *General Theory*. After writing several articles and an influential book on Keynes, in the 1950s he led his ECLAC team (which eventually included Fernando Henrique Cardoso, Enso Faletto, Celso Furtado, Aníbal Pinto and Osvaldo Sunkel) in the formulation of the ‘structuralist approach’ (see dependency; Furtado, Celso; structuralism).

In this approach, Prebisch was basically concerned with four stylized facts of commodity-exporting, middle-income countries: (*a*) their growing income gap with industrialized countries (failure to ‘catch up’); (*b*) their recurrent balance of payments disequilibrium; (*c*) the instability and the tendency to deterioration of their terms of trade; and (*d*) their persistent unemployment (often coexisting with inflationary pressures). At the core of Prebisch's analysis lies his differentiation of the economic and export structures of the centre and the periphery. Those of the centre were seen as *homogeneous* and *diversified*, those of the periphery as *heterogeneous* and *over-specialized*. Heterogeneous because economic activities with remarkably different productivity-growth dynamics existed side by side – namely, a modern export sector coexisting with a backward agriculture and an undersized manufacturing sector. Over-specialized because the range of exports was limited to just a few (homogenous, unbranded and price-volatile) commodities, and their process of production had very limited backward- and forward-linkages with the rest of the economy (see structuralism).

The recurrent cyclical problems of the balance of payments and the instability and the tendency to deterioration of the terms of trade are associated primarily with an excessive degree of export specialization (due to a narrow Ricardian understanding of comparative advantage); the problems of slow growth (and failure to ‘catch up’ with industrialized countries) and persistent unemployment with the constraints created by structural heterogeneity interacting with export over-specialization (which, among other things, hindered industrialization and created inflationary pressures).

His best-known thesis is the tendency to deterioration in the terms of trade of the periphery, the development of which coincided with (and owed much to) Hans Singer's work (1950).

It is not clear whether he saw this as his most important contribution (or even as the most significant problem of a commodity-exporting country), but by its own nature the ‘Prebisch–Singer’ thesis was a seductive empirical challenge to that part of the academic world which is ever anxious for one-dimensional hypotheses referring to clearly established variables. Prebisch was in fact much more concerned with the lack of impetus for industrialization resulting from a narrow and static Ricardian integration into the world economy. His main hypothesis is that there are reinforcing elements that – left to unregulated markets – would tend to work against the periphery's growth and welfare.

The tendency towards deterioration in the terms of trade of the periphery could be synthesized as in Figure 1.

Figure 1

X=exportable of the periphery (primary commodity); M=importable of the periphery (manufacturing good); ABC=transformation curve of the periphery; ODE=the periphery's ‘neutral’ consumption path; ODE′=its ‘biased-for-trade’ consumption path; ADF=the periphery's ‘neutral’ production path; ADF′=its ‘biased-for-trade’ production path. For the periphery, at point D, OA=TT=terms of trade, DH=consumption of M, DG=local production of M, GH=AI=imports of M, IH=production of X, OH=local consumption of X; and IO=exports of X.

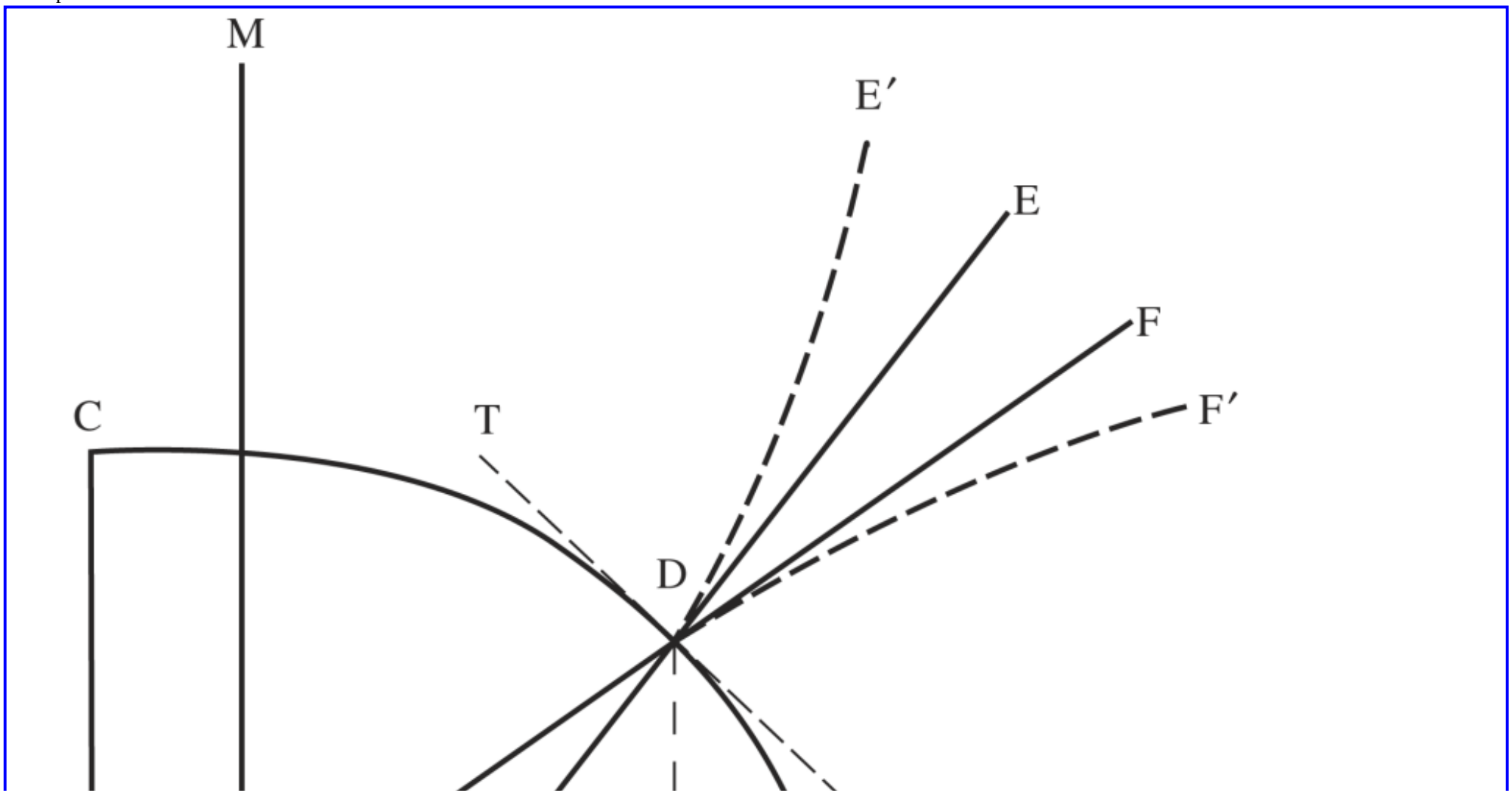

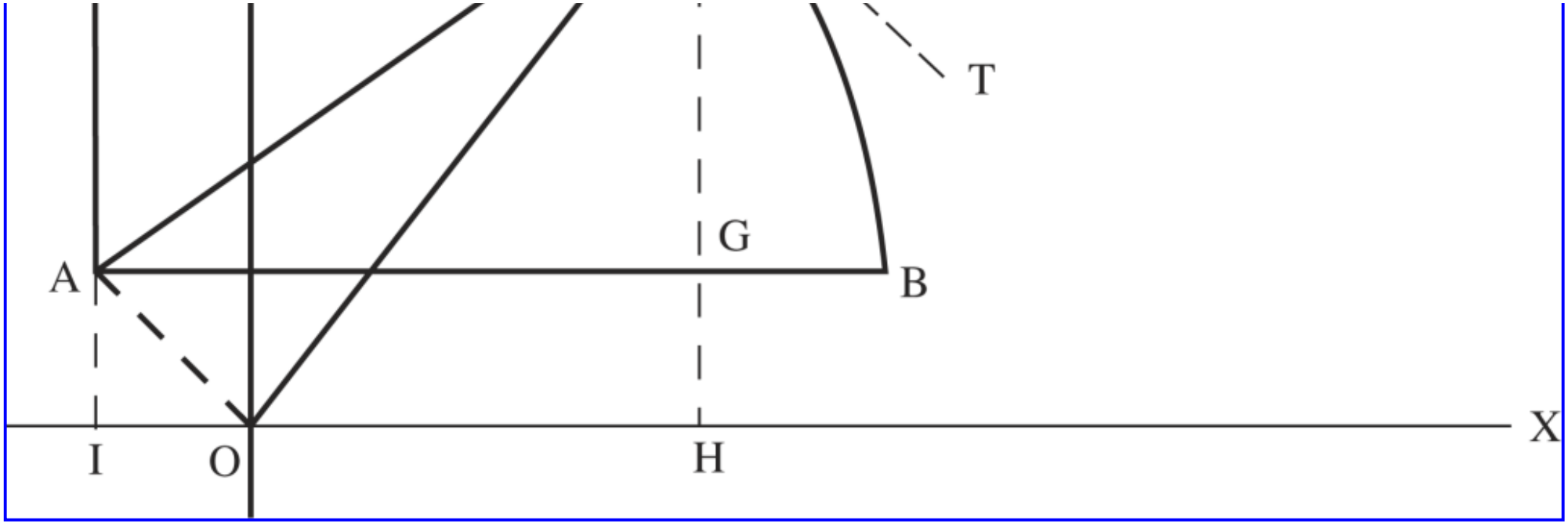

From the point of view of the periphery's consumption path, the Prebisch–Singer hypothesis is that the income elasticity of the periphery's imports from the centre (manufactures) is not only greater than 1 but also much greater than that of the centre for products of the periphery (commodities). Therefore, left to unregulated markets the long-term trend of the periphery's consumption path would be biased towards trade with the centre (say, ODE′ instead of the 'trade-neutral' ODE); that is, as incomes grow the proportion of importables (from the centre) in the periphery's consumption would increase. This would not be the case for the centre in terms of the share of its commodity imports from the periphery in total consumption (their income and price elasticities for commodities are low).

The same long-term trade bias would tend to happen in the production path – the periphery would tend to move along the ADF′ path instead of the 'trade-neutral' ADF one. That is, as output grows the share of (low price- and low income-elasticity) commodities for export in total output would increase, not least because of the additional foreign exchange needed to finance the trade-biased consumption path (something that often turns out to be a self-defeating endeavour). Therefore, vis-à-vis each other's products, the periphery would tend to have a more trade-biased path than the centre in terms of both consumption and production. There would consequently be a tendency for an excess demand from the periphery for imported manufactures from the centre, and an excess supply of commodities, resulting in a tendency towards a deterioration of the terms of trade.

This tendency would be reinforced because of a similarity and a difference in productivity growth between commodities and manufactures. The similarity is that productivity growth can be relatively high in both types of products (commodities and manufactures, although is likely to be faster in the latter). The difference is that productivity increases in the centre's manufacturing do not tend to be transferred into lower prices as much as those of (homogenous and unbranded) primary production in the periphery (due to market imperfections in the centre's product and labour markets, mainly oligopolistic firms operating in product-differentiated markets and strong unions).

The end result would be that, if both poles were to grow at the same rate of per capita income, the periphery's more trade-biased path in both consumption and production (vis-à-vis each other's products) would tend to generate a deficit in its trade balance with the centre. Therefore, a long-term equilibrium in their balance of payments would require the income per capita of the periphery to grow systematically at a lower rate than that of the centre (the opposite of a 'catching up' scenario).

Further, Prebisch adds that this growth asymmetry would be reinforced by the fact that productivity growth in manufactures tends to have higher positive externalities and spillover effects, stronger linkages with the domestic economy, steeper technology ladders, and so on. Within this context, Prebisch argues that (for reasons of supply and demand) the periphery could achieve a higher sustainable growth path only by substituting highly income-elastic manufacturing imports with domestic production, and diversifying its exports towards more income- and price-elastic, productivity-enhancing products – that is, if it were to embark on a deeper and faster process of industrialization than one that would 'spontaneously' emerge from a Ricardian integration into the world economy.

Therefore, Prebisch's arguments for forcing the pace of industrialization are not only based on differences in income elasticities of demand for imports and in price elasticities of demand for exports (arguments at the level of the circulation of commodities), but are also due to the growth-enhancing nature of manufacturing activities (an argument at the level of production; see Kaldor, Nicholas).

Prebisch's theory challenges Ricardo's comparative advantage premises – in fact, for Prebisch the higher the rate of growth of productivity in the periphery's primary commodity

export sector, the greater the need for import-substituting industrialization (1983, p. 1082). It also challenges the classical terms-of-trade approach – for example, Mill (1848) and Keynes (1920) – which argued that in the long term they are bound to move in favour of commodities (mainly due to hypothetical diminishing returns in commodity production). Prebisch's logic would later influence Joan Robinson (1979) when she argued that in Ricardo's example Portugal ends up with a low rate of accumulation, and having destroyed its promising textile industry, while England in contrast had an industrial revolution (see Robinson, Joan Violet). It would also influence Kaldor's arguments in favour of manufacturing-led growth (1967), Pasinetti's multi-sector macro-dynamics framework (1983), Ajit Singh's concept of an optimal degree of industrialization (1977), Rowthorn and Wells' seminal work on de-industrialization (1987), and Thirlwall's balance-of-payments-constrained growth multiplier (2003).

For criticisms of Prebisch's ideas, see structuralism and dependency. Some additional issues to which the literature on Prebisch has not given due consideration are as follows:

1. 1. Although there is no evidence that this was Prebisch's intention, for many years his ideas led in many intellectual and policymaking circles to a strong bias against commodity production per se.
2. 2. The asymmetric trade liberalization that has taken place since globalization (the periphery opening up to manufacturing imports, but the centre not reciprocating for commodities) has deepened the problems identified by Prebisch.
3. 3. Prebisch's Argentinian background is undoubtedly responsible for his focusing mainly on the relative decline of middle-income, commodity-rich countries, and for his scepticism regarding the role of an inelastic supply of agricultural products in explaining the regions' persistent inflationary pressures (Argentina simply did not fit the pattern of the structuralist theories of inflation developed by many of his colleagues at ECLAC; see structuralism).
4. 4. The recent remarkable export drive of basic (homogenous and unbranded) manufacturing products in some developing countries has led to their 'price commoditization', leading to a similar terms of trade problem vis-à-vis their imports of more technologically advanced manufactures (see Palma, 2005).
5. 5. There are significant fallacy-of-composition issues among commodity-exporting countries (for example, actual price elasticity of demand crucially depends on market shares), making cooperative games among producers difficult.
6. 6. Probably because of his own 'institutional constraints' (inevitable when working in international organizations), Prebisch never addressed properly some crucial institutional issues associated with the often poor macroeconomic performance of many mineral-exporting economies – such as those analysed (not always successfully) by the 'resource course' literature (see de-industrialization, 'premature' de-industrialization and the Dutch Disease).
7. 7. Prebisch's preferences for a 'stage' approach (first an import-substituting phase, then a manufacturing export-oriented one towards a regional custom union, then one to the rest of the world) did not account for institutional and political path-dependency inertias that would create almost insurmountable hurdles for the transition even to the second stage; East Asia's 'simultaneous' approach was far more successful (Palma, 2007).
8. 8. Finally (and crucially), the contrasting experiences of Latin America and East Asia show that, while it is one thing to use trade and industrial policies to create incentives (rents) to divert resources towards more 'dynamic' products (that is, income- and price-elastic manufacturing products, with deeper linkages, higher productivity growth potential, stronger externalities and spillover effects, useful for technology ladders, and so on), it is quite another to have the institutional capabilities necessary to ensure that the capitalist elite uses those rents effectively (Khan, 2000).

See Also

- de-industrialization, 'premature' de-industrialization and the Dutch Disease
- dependency
- Díaz-Alejandro, Carlos
- Furtado, Celso
- Kaldor, Nicholas
- Robinson, Joan Violet
- structuralism
- terms of trade

Selected works

1949. The economic development of Latin America and its principal problems. *Economic Bulletin for Latin America* 7, 1962, 1–22 (first published by ECLAC in 1949).

1951a. The spread of technical progress and the terms of trade. In *Economic Survey of Latin America, 1949*. New York: UN-DESA.

1951b. *Problemas teóricos y prácticos del crecimiento económico*. Santiago: UN-ECLAC.

1959. Commercial policy in the underdeveloped countries. *American Economic Review* 49, 251–73.

1962. El falso dilema entre desarrollo económico y estabilidad monetaria. *Boletín Económico de América Latina* 6(1), 1–26.

1963. *Towards a Dynamic Development Policy for Latin America*. New York: UN.

1968. A new strategy for development. *Journal of Economic Studies* 3(1), 3–14.

1971a. *Latin America: A Problem in Development*. Austin: Institute of Latin American Studies, University of Texas.

1971b. *Change and Development – Latin America's Great Task*. New York: Praeger; Mexico: Fondo de Cultura Económico.

1976. A critique of peripheral capitalism. *CEPAL Review* 1, 9–76.

1981. *Capitalismo periférico: crisis y transformación*. México: Fondo de Cultura Económico.

1983. Tres etapas de mi pensamiento económico. *El Trimestre Económico* 50, 1077–96.

Bibliography

Di Marco, L.E. 1972. The evolution of Prebisch's economic thought. In *International Economics and Development: Essays in Honour of Raul Prebisch*, ed. L.E. Di Marco. New York and London: Academic Press.

Kaldor, N. 1967. Problems of industrialization in underdeveloped countries. In *Strategic Factors of Economic Development*. Cornell: Cornell University Press.

Keynes, J.M. 1920. *Economic Consequences of the Peace*. London: Macmillan.

Khan, M. 2000. Rent-seeking as process. In *Rents, Rent-Seeking and Economic Development*, ed. M. Khan and K. Jomo. Cambridge: Cambridge University Press.

Love, J. 1995. Economic ideas and ideologies in Latin America since 1930. In *Cambridge History of Latin America*, vol. 6, part I. Cambridge: Cambridge University Press.

Mill, J.S. 1848. *Principles of Political Economy*. London: Longmans.

Ocampo, J.A. and Parra, M.A. 2003. The terms of trade for commodities in the twentieth century. *ECLAC Review* 29, 7–35.

Palma, J.G. 2005. The six main stylised facts of the Mexican economy since NAFTA. *Journal of Industrial and Corporate Change* 14, 941–91.

Palma, J.G. 2007. Flying geese and waddling ducks: the different capabilities of East Asia and Latin America to ‘demand-adapt’ and ‘supply upgrade’ their export productive capacity. In *Industrial policy in Developing Countries*, ed. M. Cimoli, G. Dosi and J. Stiglitz. Oxford: Oxford University Press.

Pasinetti, L. 1983. The accumulation of capital. *Cambridge Journal of Economics* 7, 405–11.

Robinson, J.V. 1979. *Aspects of Development and Underdevelopment*. Cambridge: Cambridge University Press.

Rodriguez, O. 1980. *La teoría del subdesarrollo de la CEPAL*. Mexico: Siglo XXI Editores.

Rodriguez, O. 2007. *El estructuralismo latinoamericano*. Mexico: Siglo XXI Editores.

Rowthorn, R. and Wells, J. 1987. *De-industrialisation and Foreign Trade*. Cambridge: Cambridge University Press.

Singer, H. 1950. The distribution of gains between investing and borrowing countries. *American Economic Review* 40, 473–85.

Singh, A. 1977. UK industry and the world economy: a case of de-industrialisation? *Cambridge Journal of Economics* 1, 113–16.

Thirlwall, A. 2003. *Trade, the Balance of Payments and Exchange Rate Policy in Developing Countries*. Cheltenham: Edward Elgar.

How to cite this article

Palma, José Gabriel. "Prebisch, Raúl (1901–1986)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000141> doi:10.1057/9780230226203.1323

The New Palgrave Dictionary of Economics Online

precautionary principle

Christian Gollier and Nicolas Treich

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The precautionary principle (PP), as it appears in international treaties or in some countries' legal systems, suggests that the prospect of scientific progress should not justify the delay of preventive measures. Three effects identified in the economics literature – the irreversibility, the precautionary and the ambiguity aversion effects – may be consistent with the normative content of the PP. A difficult question is how then the PP can be implemented. Several social actors may want to take advantage of a current lack of scientific evidence to promote their own interests. The PP can also be misused, for example, for demagogy or protectionism.

Keywords

ambiguity and ambiguity aversion; asymmetric information; availability heuristic; Cartagena Protocol on Biosafety (2000); Convention on Biological Diversity (1992); Convention on Climate Change (1992); environmental economics; irreversibility effect; Maastricht Treaty (EU); precautionary effect; precautionary principle; protection; risk; risk perception; scientific progress; scientific uncertainty

Article

The precautionary principle (PP) is a recent notion. It has its roots, some believe, in the early 1970s as the German principle of *Vorsorge*, or foresight (see, for example, O'Riordan and Cameron, 1994; Morris, 2000; Sunstein, 2005). It is often said that the PP was first introduced in 1984 at the International Conference on Protection of the North Sea. Its popularity increased after the Conference of Rio in 1992; Principle 15 of the Rio Declaration states, 'Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation' (UNGA, 1992). Similar definitions have been proposed in international statements of policy, including the 1992 Convention on Climate Change, the 1992 Convention on Biological Diversity, the Maastricht Treaty in 1992/93, and the 2000 Cartagena Protocol

on Biosafety. The PP has also been enacted in the national law of several countries, especially in Europe. The PP is the most notable anticipatory principle with special relevance for human-induced environmental problems under conditions of scientific uncertainty. Although devoid of practical content, the main message of the PP is conceptually clear: the prospect of anticipation of scientific progress should not justify the delay of measures preventing environmental degradation. In practice, its scope has became wider and there are reasonable grounds for applying it to regulate the protection of human, animal and plant health issues (see, for example, Commission of European Communities, 2000).

The economic analysis of the PP has mostly studied the tension between two effects: (*a*) developing an economic activity that is profitable now but may pose risks to the society in the future, and (*b*) not developing this activity until conclusive scientific information is forthcoming about its harmlessness. The PP is said to be socially efficient if the benefit of postponing the risky activity is greater than its cost. To put it differently, the PP is efficient if the net social benefit of early prevention efforts is positive. The economic conditions for efficiency were first analysed in the 1970s in the literature on the ‘irreversibility effect’.

The irreversibility effect

Let us consider a model of economic decisions represented by the following optimization program

$$\max_{x_1 \in D_1} E_{\tilde{y}} \max_{x_2 \in D(x_1)} E_{\tilde{\theta}/\tilde{y}} V(x_1, x_2, \tilde{\theta}) \tag{1}$$

The timing of the model is the following. At date 1, the decision-maker chooses x_1 in a set D_1. Between date 1 and date 2, he observes the realization of a signal *y* correlated to $\tilde{\theta}$. At date 2, before the realization of $\tilde{\theta}$, he chooses x_2 in a set $D(x_1)$. Finally $\tilde{\theta}$ is realized and the decision-maker gets a utility payoff $V(x_1, x_2, \theta)$. The problem is to determine the effect of a ‘better information structure’ $\tilde{y}$ on the optimal decision at date 1.

We first solve this problem when $V(x_1, x_2, \theta) = x_1 + x_2\theta$ with $D_1 = \{0, 1\}$ and $D(x_1) = \{x_1, 1\}$. This special case can be interpreted as a simple investment problem. The development of a ‘risky’ project – like the exploitation of a forest in which the value of biodiversity is unknown – is considered. If the project is implemented today ($x_1 = 1$), it yields a net benefit of 1 today and of $\tilde{\theta}$ in the future. The project is irreversible in the sense that once it is developed it cannot be stopped ($x_2 = 1$ if $x_1 = 1$). The stakeholders are assumed to be risk neutral.

Consider first the case in which no scientific progress is expected, that is, when $\tilde{y}$ is independent of $\tilde{\theta}$. In this case, program (1) becomes

$$\max_{x_1 \in \{0,1\}, x_2 \in \{x_1, 1\}} E_{\tilde{\theta}}(x_1 + x_2\tilde{\theta}) = \max(1 + E_{\tilde{\theta}}\tilde{\theta}, 0)$$

(2)

Either the project is implemented today if its expected net present value (ENPV) is positive, that is if $1 + E_{\tilde{\theta}}\tilde{\theta} \geq 0$, or it is never implemented. Consider alternatively the case of scientific progress that yields perfect information about $\tilde{\theta}$. This is equivalent to assuming perfect correlation between $\tilde{V}$ and $\tilde{\theta}$. In this case, program (1) becomes

$$\max_{x_1 \in \{0,1\}} E_{\tilde{\theta}} \max_{x_2 \in \{x_1, 1\}} (x_1 + x_2 \tilde{\theta}) = \max(1 + E_{\tilde{\theta}}\tilde{\theta}, E_{\tilde{\theta}} \max(0, \tilde{\theta}))$$

(3)

Viewed today, the ENPV of postponing the decision to develop the project equals $V = E_{\tilde{\theta}} \max(0, \tilde{\theta})$. The project will be initiated today only if it yields a larger ENPV than that obtained if the decision is postponed to the future: $1 + E_{\tilde{\theta}}\tilde{\theta} \geq V$. The quantity *V* has been coined the (quasi-)option value (Arrow and Fisher, 1974).

The comparison between (2) and (3) shows that scientific progress has the effect of increasing the ENPV of the best alternative option from 0 to $V \geq 0$. Consistent with the PP, this example shows that the prospect of scientific progress may lead to the postponement of the development of the risky project. The prospect of receiving information in the future increases the cost of choosing the irreversible decision today. This decision would prevent the decision maker from taking advantage of the information in the future. This is the ‘irreversibility effect’ (Henry, 1974).

The literature has studied the generalization of this effect in several directions, including partial resolution of uncertainty, relative flexibility, continuous decision variables, non-separable preferences and risk aversion. This example relied on two extreme information structures: one structure gives no information and the other gives perfect information. The appropriate general notion of a ‘better information structure’ was introduced by Blackwell (1951). This general notion was used and developed in a systematic way by Epstein (1980) under some differentiability assumptions. Epstein then demonstrated that the irreversibility effect does not hold for most payoff functions $V(x_1, x_2, \theta)$. Jones and Ostroy (1984) have generalized Epstein's result to non-differentiable problems and to a more general characterization of adjustment costs.

The precautionary effect

The subsequent literature has mostly used Epstein's approach to examine the effect of better information for various payoff functions, on the assumption of continuous decisions, differentiability and that the conditions for optimization in (1) were satisfied. Ulph and Ulph (1997) consider a payoff function of the

form $v(x_1, x_2, \theta) = u_1(x_1) + u_2(x_2) - \theta d(\delta x_1 + x_2)$ and interpret x_t as the emissions of CO_2 in period *t* and $\theta d(.)$ as the uncertain climate damage that depends on the sum of emissions up to a decay parameter δ. They show that a better information structure may lead to an increase, not a decrease, in emissions at date 1. Gollier, Jullien and Treich (2000) analyse a similar model with monetary damages $v(x_1, x_2, \theta) = u_1(x_1) + u_2(x_2 - \theta(\delta x_1 + x_2))$. They show that that emissions at date 1 decrease if and only if $u_2(.)$ has a constant relative risk aversion lower than 1, or a derivative 'sufficiently' convex. This latter condition suggests that the coefficient of prudence (Kimball, 1990) is instrumental in signing the effect of a better information structure on x_1. This is not surprising since in this model x_1 affects future utility u_2, no longer by reducing the future set of choices but directly by changing the risk borne in the future $\theta(\delta x_1 + x_2)$. This is the 'precautionary effect'. Overall these results suggest that the qualitative effect of a better information structure strongly depends on functional forms, in particular on the risk attitude of the decision maker.

The ambiguity aversion effect

The Ellsberg paradox tells us that many people do not behave according to the expected utility criterion when facing (scientific) uncertainty, contrary to what we assumed above. Gilboa and Schmeidler (1989) proposed an alternative decision criterion that performs better in this context. Under their model of ambiguity aversion, for each possible choice *ex ante*, the decision maker computes the expected utility conditional to each plausible scientific theory, and takes the minimum to evaluate the welfare generated by that choice. Agents who behave according to this maxmin model exhibit a form of choice-sensitive pessimism, which is called 'ambiguity aversion'. As shown for example by Chen and Epstein (2002) for financial markets, this ambiguity aversion reinforces risk aversion to induce people to adopt a more precautionary behaviour in the case of (scientific) uncertainty, as suggested by the PP.

Positive aspects of the precautionary principle

The economic approach of the PP has been mostly normative so far. Under which conditions is the PP socially efficient? How should scientific uncertainty affect risk management? An equally important approach involves discussion on how the PP has been or should be implemented. We briefly turn to these more positive aspects.

A general argument is that scientific uncertainty may exacerbate, or even trigger, some market or regulatory failures (Gollier and Treich, 2003). With a global pollution problem such as climate change, there are incentives for countries to free ride on other countries' reduction of emissions. Coalitions formations may reduce this inefficiency but coalitions are less likely to form if there is scientific progress (Na and Shin, 1998). At a political level, an argument used by governments is that the problem is 'too uncertain' to abate emissions. Early commitments may help, but there are incentives for some governments, once information reveals low levels of damage in their own country, to refuse to abate emissions at a level announced by previous governments.

A difficult question is that of the most efficient policy to induce firms to internalize the risks they pose for the economy. In a market with imperfect legally enforceable property rights, firms may not take up

the option of waiting for better information when high profits are guaranteed to first-movers. How to set binding legal incentives for firms' past actions made under conditions of scientific uncertainty is a big issue in law. This issue is augmented by the classical limited liability problem.

Another issue is that of international relations and the different approaches to safety and precaution across countries (Hammitt et al., 2005). One possibility is to leave states to decide how to account for scientific uncertainty in their safety policy. The problem is that such a discretionary power may be the source of disguised protectionism.

Scientific uncertainty may also increase the cognitive biases of the public in their perception of risks, like the standard 'availability heuristic'. Citizens often deem an event to be more probable when its occurrence can be easily recalled or visualized. As a result, they may overreact to highly publicized risks. Interest groups may exploit this bias, as well as politicians. A critical interpretation of the PP is to view it as a demagogic response to citizens' perceptions of risks (Sunstein, 2005).

More generally, scientific uncertainty may favour, through the multiple channels of decision-making, opportunistic behaviours. Scientific uncertainty creates space for discretion in the risk regulatory process. Several social actors (entrepreneurs, lobbies, experts, politicians, media, and so on) may take advantage of the lack of scientific evidence to promote their own interests. The PP may be viewed as a soft safeguard against opportunistic behaviours in situations of asymmetric and evolving information. Yet designing stronger mechanisms needs a more detailed analysis of the sources of market failures, of risk management institutions and of citizens' behavioural responses. This may partly explain the existing voluminous literature on the PP in the social sciences, and may occupy economists in the future.

See Also

- ambiguity and ambiguity aversion
- cost–benefit analysis
- irreversible investment
- risk

Bibliography

Arrow, K. and Fischer, A. 1974. Environmental preservation, uncertainty and irreversibility. *Quarterly Journal of Economics* 88, 312–19.

Blackwell, D. 1951. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Chen, Z. and Epstein, L. 2002. Ambiguity, risk, and asset returns in continuous time, *Econometrica* 70, 1403–43.

Commission of the European Communities. 2000. *Communication from the Commission on the Precautionary Principle*. Brussels: Commission of the European Communities. Online. Available at

http://europa.eu.int/eur-lex/en/com/cnc/2000/com2000_0001en01.pdf, accessed 23 March 2006.

Epstein, L. 1980. Decision-making and the temporal resolution of uncertainty. *International Economic Review* 21, 269–84.

Gilboa, I. and Schmeidler, D. 1989. Maximin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Gollier, C., Jullien, B. and Treich, N. 2000. Scientific progress and irreversibility: an economic interpretation of the Precautionary Principle. *Journal of Public Economics* 75, 229–53.

Gollier, C. and Treich, N. 2003. Decision-making under scientific uncertainty: the economics of the Precautionary Principle. *Journal of Risk and Uncertainty* 27, 77–103.

Hammitt, J., Wiener, J., Swedlow, B., Kall, D. and Zhou, Z. 2005. Precautionary regulation in Europe and in the United States: a quantitative comparison. *Risk Analysis* 25, 1215–28.

Henry, C. 1974. Investment decisions under uncertainty: the ‘irreversibility effect’. *American Economic Review* 64, 1006–12.

Jones, J. and Ostroy, R. 1984. Flexibility and uncertainty. *Review of Economic Studies* 6, 13–32.

Kimball, M. 1990. Precautionary savings in the small and in the large. *Econometrica* 61, 53–73.

Morris, J. 2000. Defining the Precautionary Principle. In *Rethinking Risk and the Precautionary Principle*, ed. J. Morris. Oxford: Butterworth-Heinemann.

Na, S. and Shin, H. 1998. International environmental agreements under uncertainty. *Oxford Economic Papers* 50, 173–85.

O’Riordan, T. and Cameron, J., eds. 1994. *Interpreting the Precautionary Principle*. London: Earthscan Publications.

Sunstein, C. 2005. *Laws of Fear: Beyond the Precautionary Principle*. Cambridge: Cambridge University Press.

Ulph, A. and Ulph, D. 1997. Global warming, irreversibility and learning. *Economic Journal* 107, 636–50.

UNGA (United Nations General Assembly). 1992. *Report of the United Nations Conference on Environment and Development. Annex I: Rio Declaration on Environment and Development*. Rio de

Janeiro, 3–14 June. A/CONF.151/26 (Vol. I) of 12 August. Online. Available at http://www.un.org/documents/ga/conf151/aconf15126-1annex1.htm, accessed 30 March 2006.

How to cite this article

Gollier, Christian and Nicolas Treich. "precautionary principle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000347> doi:10.1057/9780230226203.1324

The New Palgrave Dictionary of Economics Online

precautionary saving and precautionary wealth

Christopher D. Carroll and Miles S. Kimball
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Precautionary saving measures the consequences of uncertainty for the rate of change (and therefore the level) of wealth. The qualitative aspects of precautionary saving theory are now well established: an increase in uncertainty will increase the level of saving, but will reduce the marginal propensity to save. Quantitatively, theory combined with empirical estimates of risk aversion suggests that precautionary saving and precautionary wealth should be quite large. More direct empirical evidence on precautionary saving suggests that precautionary effects on saving are substantial, but the magnitude of the effects is disputed, and the different estimates are not all expressed in comparable units.

Keywords

calibration; consumption function; elasticity of intertemporal substitution; Euler equations; impatience; liquidity constraints; perfect foresight; precautionary saving; precautionary savings; precautionary wealth; preferences; risk aversion; time consistency; uncertainty

Article

Precautionary saving is additional saving that results from the knowledge that the future is uncertain.
In principle, additional saving can be achieved either by consuming less or by working more; here, we follow most of the literature in neglecting the ‘working more’ channel by treating non-capital income as exogenous.
Before proceeding, a terminological clarification is in order. ‘Precautionary saving’ and ‘precautionary savings’ are often (understandably) confused. ‘Precautionary saving’ is a response of current spending to future risk, conditional on current circumstances. ‘Precautionary savings’ is the additional wealth owned at a given point in time as the result of past precautionary behaviour. That is, precautionary savings at any date is the stock of extra wealth that has resulted from the past flow of precautionary saving. To avoid confusion, we advocate use of the phrase ‘precautionary wealth’ in place of ‘precautionary savings’.

Strength of the precautionary saving motive

In the standard analysis, originally formulated in a two-period model by Leland (1968), and extended to the multi-period case by Sibley (1975) and Miller (1976), precautionary saving is modelled as the outcome of a consumer's optimizing choice of how to allocate existing resources between the present and the future. Additional interest in precautionary saving was stimulated by numerical solution of a benchmark model by Zeldes (1989) and the connection made in Barsky, Mankiw and Zeldes (1986) between precautionary saving and the effects of government debt. (We assume time-invariant preferences in order to sidestep the important issues of time consistency recently explored by Laibson, 1997, and others. That literature opens up a rich and interesting field of further behavioural possibilities beyond the basic logic outlined here.)
To clarify the theoretical issues, we break down the consumer's problem into two steps: the transition between periods, and the choice within the period. A consumer who ends period t with assets a_t receives capital income in period t+1 of $a_t r$. The consumer's immediate resources (‘cash-on-hand’) in period t+1 consist of such capital income, plus the assets that

generated it, plus labour income y_{t+1}:

$$m_{t+1} = a_t r + a_t + y_{t+1} \quad (1)$$

$$= \underbrace{(1+r)}_{\equiv R} a_t + y_{t+1}. \quad (2)$$

The simplest interpretation of *m* is as the contents of the consumer's bank account immediately after receipt of the paycheck and interest income (‘cash-on-hand’). *R* is the real interest *factor*, as distinct from the real interest *rate*, lower case *r*. a_t reflects the consumer's accumulated assets at the end of period *t*, after the spending decision for period *t* has been made. The transition from the beginning to the end of period *t* reflects the fact that spending is paid for by drawing down *m*:

$$a_t = m_t - c_t. \quad (3)$$

To decide how to behave optimally in period *t*, the consumer must be able to judge the value of arriving in period *t*+1 in any possible circumstance. This information is captured by the value function $\nu_{t+1}(m_{t+1})$. Here, we simply assume the existence of some well-behaved ν_{t+1}; below we show how to construct ν_{t+1}.
Standard practice assumes that consumers in period *t* weight future value by the factor β; if $\beta=1$ the consumer today cares equally about current and future pleasure, while if $\beta < 1$ the consumer prefers present to future pleasure. Given β, and assuming that the consumer's period-*t* beliefs about future distribution of income are captured by the expectations operator E_t, we can define the value of ending period *t* with accumulated assets a_t as

$$\omega_t(a_t) = \beta E_t[v_{t+1}(Ra_t + \tilde{y}_{t+1})], \quad (4)$$

where the ~ over the *y* indicates that period-(*t*+1) income is uncertain from the perspective of period *t*. Think of $\omega_t(a)$ as the end-of-period value function.
The consumer's goal is to optimally allocate beginning-of-period resources between current consumption and end-of-period assets; the value function for period *t* is defined as the function that yields the value associated with the optimal choice:

$$v_t(m_t) = \max_{ct} \left\{ u(c_t) + \omega_t(m_t - c_t) \right\}.$$

(5)

By definition, the optimal choice will be a level of c_t such that the consumer does not wish to change spending. Under standard assumptions, this implies that the marginal utility of consumption must be equal to the marginal value of assets:

$$u'(\overbrace{m_t - a_t}^{c_t}) = \omega_t'(a_t), \tag{6}$$

since if this were not true the consumer would be able to improve his well-being (value) by reallocating some resources between *a* and *c*.

Figure 1 depicts the consumer's problem graphically. For given initial m_t, the consumer's goal is to find the value of *a* such that (6) holds. The left-hand side of (6) is the upward-sloping locus. As for the two downward-sloping loci, the lower one reflects expected marginal value if the consumer is perfectly certain to receive the mean level of income $E_t[\tilde{y}_{t+1}]$, while the higher downward-sloping function corresponds to the case where income is uncertain.

When the risk is added, the optimal choice for end-of-period assets moves from a^* to a^{**}. Since $c_t = m_t - a_t$, the increase in *a* in response to risk corresponds to a reduction in consumption. This reduction in consumption is the precautionary saving induced by the risk.

Figure 1

Marginal utility of assets and of consumption

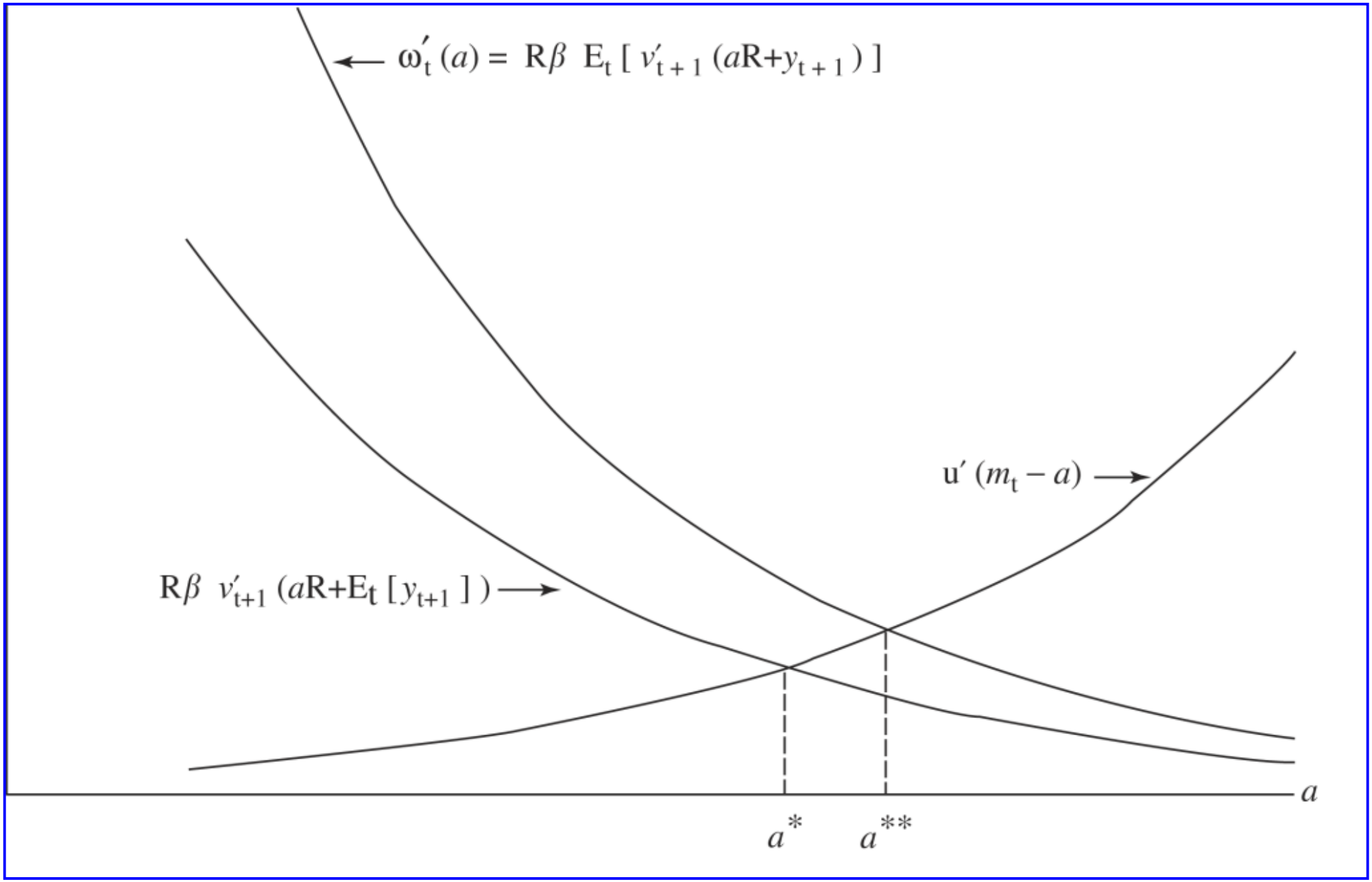

For a given $\nu_{t+1}(m_{t+1})$, the exercise captured in the diagram can be conducted for every possible value of m_t, implicitly defining a consumption function $c_t(m_t)$.

Kimball (1990) shows that the index of absolute prudence $\frac{-v'''_{t+1}(m_{t+1})}{v''_{t+1}(m_{t+1})}$ and the index of relative prudence $\frac{-v'''_{t+1}(m_{t+1})m_{t+1}}{v''_{t+1}(m_{t+1})}$ are good measures of how much a risk of given size will shift the marginal value of assets curve $\omega'_t(a)$ to the right. For a constant relative risk aversion value function, relative prudence is equal to relative risk aversion plus 1. Kimball and Weil (2004) look at the strength of the precautionary saving motive when Kreps and Porteus (1978) preferences are used to break the usual equation $\varsigma=1/\rho$ where ς is the elasticity of intertemporal substitution and ρ is relative risk aversion. In this more general case, the counterpart to relative prudence $\mathcal{P}$ is given by $\mathcal{P} = (1+\varsigma\varepsilon)\rho$, where ε is the

elasticity with which absolute risk aversion declines and absolute risk tolerance increases.

Note that, given the basic properties $c > 0$ and $b > 0$, a positive wealth elasticity of risk tolerance implies that $p > b$. This is a special case of a much more general result first hinted at by Drèze and Modigliani (1972). Even for very exotic objective functions, the precautionary saving motive will always be stronger than risk aversion whenever ownership of more a_t due to a small forced reduction in consumption would lead an optimizing investor to bear more risk (a property that Drèze and Modigliani, 1972 call ‘endogenously decreasing absolute risk aversion’). This general result holds because, if ownership of extra a_t due to a small forced reduction in consumption were to lead an optimizing investor to bear risks she was previously indifferent to, then reduced consumption must be complementary with bearing near-indifferent risks. The symmetry of complementarity then implies that, given a free choice of consumption levels, taking on an additional near-indifferent risk will lead an optimizing consumer to reduce consumption. For example, consider an agent with additive habit formation (as distinct from multiplicative habits, compare Carroll, 2000), for whom reduced consumption not only increases assets but reduces the size of the consumption habit, and so unambiguously leads to more willingness to bear risks. Such an agent will want to reduce consumption if induced to take on an additional risk by a compensation that makes her indifferent to the risk. The size of the compensation is determined by risk aversion. Yet the compensation for the agent's risk aversion is not enough to cancel out the precautionary saving effect of the risk.

Buffer stock wealth

The above discussion suggested that precautionary behaviour can be understood by considering a trade-off between the present (captured by $u(c_t)$) and the future (captured by $\omega_t(m_t - c_t)$).

That analysis was incomplete in a crucial respect: it took the initial level of resources, m_t, as given exogenously. But, arguably, the most important question about precautionary behaviour is how large an effect it has on the prevailing level of m. This cannot be answered using a framework that treats m as exogenous.

The framework can be extended to address this problem, by defining the problem in such a way that the functions ν and ω reflect the discounted value of an infinite number of future periods. This is often accomplished by making assumptions under which optimal behaviour in every future period is identical to optimal behaviour in the current period; it is then possible to solve for a ‘consumption function’ that provides a complete characterization of the relationship between resources and spending.

The critical extra assumption is ‘impatience’, broadly construed as a condition on preferences that prevents wealth (or the wealth to income ratio) from growing to infinity. In the simplest version of the model where income does not grow, the required condition is $R\beta < 1$; for the appropriate condition in models with income growth, see Carroll (2004).

The exact nature of income risk turns out to be less important than the assumption of impatience. Here, we analyse a particularly simple case (which is an adaptation of a model by Toché, 2005). There are two kinds of consumers: workers and retirees. Retirees have no labour income, and must live off their assets. Workers earn a fixed amount of labour income in each period, but face a constant danger of being exogenously forced into retirement. (Exogenous forced retirement is the sole source of risk in the model.)

Under these assumptions, if the utility function is of the standard constant relative risk aversion form $u(c) = c^{1-\rho}/(1-\rho)$, optimal behaviour for retirees is very simple: they spend a constant fraction of m in each period, where the fraction depends on the degree of impatience and intertemporal substitution ($1/\rho$).

The situation for workers is more interesting; it is depicted in Figure 2. The simplest element of the figure is the line labelled ‘Perm inc’. This shows, for any m, the level of spending that would leave expected m unchanged; it is equal to labour income plus the interest on capital income, and is upward sloping because a consumer with more m earns more capital income.

Figure 2

The consumption function

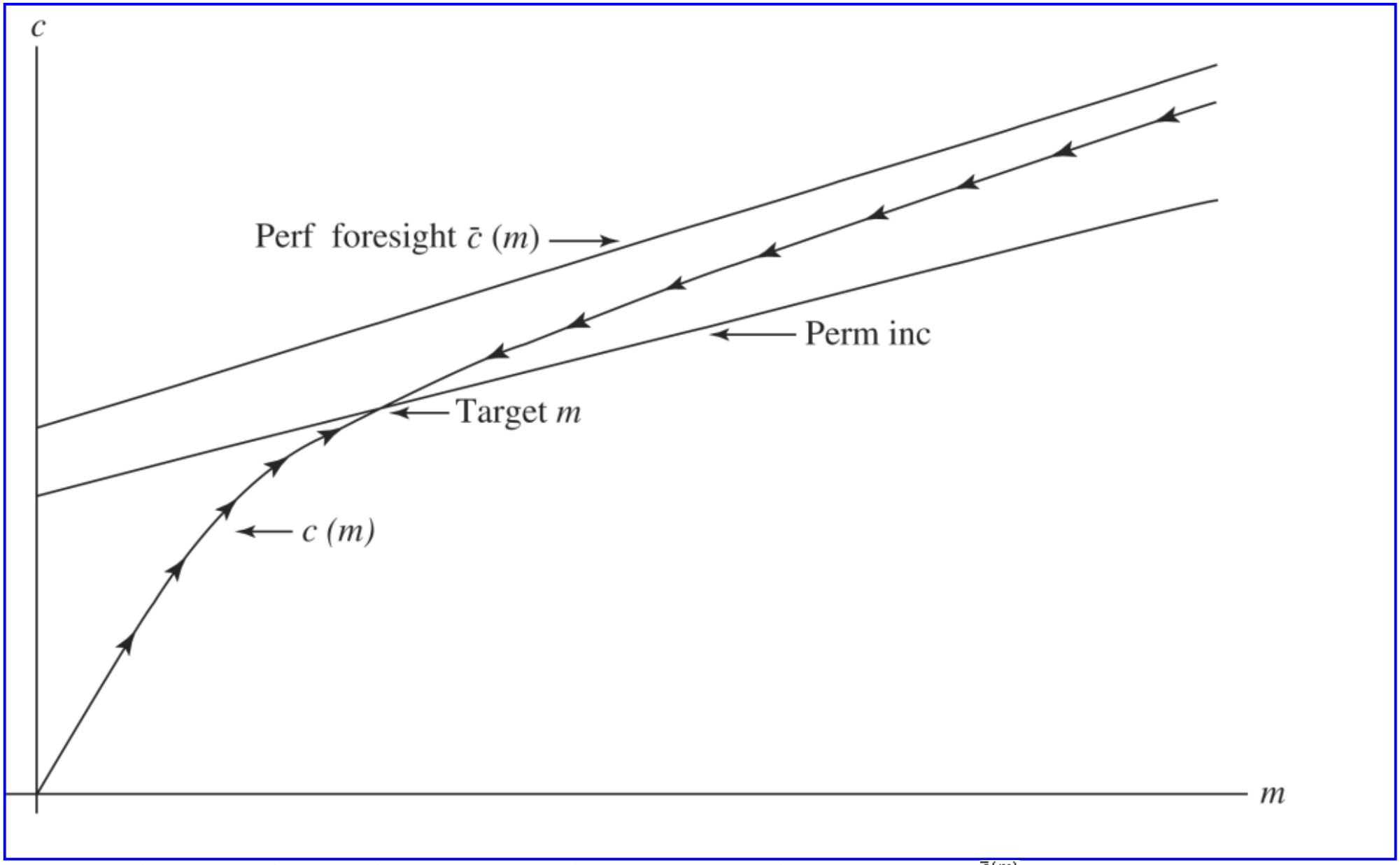

The assumption of impatience is reflected in the fact that the consumption function that would apply if uncertainty did not exist, $\bar{c}(m)$, is everywhere above the level of permanent income (income of the perfect-certainty consumer is adjusted downwards so that the reduction in unemployment risk does not cause an increase in mean income). In other words, an impatient consumer facing no uncertainty would choose to spend at a rate that cannot be sustained indefinitely.

The locus with arrows is the consumption function, which indicates the optimal level of spending (in the presence of uncertainty) for any given level of *m*. Since the difference between $c(m)$ and $\bar{c}(m)$ is purely the consequence of risk, that difference $\bar{c}(m) - c(m)$ constitutes the amount of precautionary saving associated with any specific *m*.

Standard assumptions about preferences and uncertainty imply that there will be an intersection between the permanent income locus and the consumption function. (For a proof that there will be only one intersection, see Carroll, 2004.) The intersection defines a ‘target’ level for the buffer stock of wealth *m*: the level such that an employed consumer with this amount of resources today will end up with the same *m* next period. Dynamics are captured by the arrows, which indicate that, for initial values of *m* below the target, consumption is below permanent income, so *m* is increasing and consumption crawls upwards along the consumption function towards the target. For initial values of *m* above the target,

consumption is above permanent income, so *m* is falling. The consumer holds a 'buffer stock' of wealth in an attempt to reach the 'target' level of wealth as defined above.

The existence of a target level of resources has many interesting implications. Perhaps the most surprising is that in long-run equilibrium the expected growth rate of consumption for employed consumers is unrelated to the interest rate or the degree of impatience.

To understand this point better, and to relate it to the literature, we restate it in a slightly more general form: The equilibrium expected growth rate of consumption for employed consumers is approximately equal to their predictable rate of income growth,

$$E_t[\Delta \log c^e_{t+1}] \approx g. \tag{7}$$

In many respects, the equilibrium equality of consumption growth and permanent income growth seems intuitive. However, it appears to conflict with a standard way of analysing consumption growth, which relies on the first-order condition from the optimization problem (the 'Euler equation'), which is often approximated by an equation of the form

$$E_t[\Delta \log c^e_{t+1}] \approx \rho^{-1}(r - \tau) + \varphi \tag{8}$$

where ρ is the coefficient of relative risk aversion and τ is the geometric rate at which future utility is discounted (related to the time preference factor β); ϕ is a term that reflects the contribution of precautionary motives to consumption growth.

The resolution of the apparent contradiction is that the precautionary component of consumption growth is endogenous; combining (7) and (8) permits us to solve for the equilibrium value of the precautionary contribution to consumption growth:

$$\varphi \approx g - \rho^{-1}(r - \tau). \tag{9}$$

We return to this point below.

We can characterize the effect of uncertainty by noting three facts about Figure 2: $c(m) < \bar{c}(m)$ (consumption is lower in the presence of uncertainty); $\lim_{m \to \infty} \bar{c}(m) - c(m) = 0$ (as wealth approaches infinity the effect of uncertainty in labour income vanishes); and $c(m)$ is strictly concave, so that the marginal propensity to consume out of a windfall increase in income, $c'(m)$, is greater for poor people than for rich people.

The concavity of the consumption function bears further comment. Intuitively, it can be understood in a similar light to the effect of liquidity constraints. A consumer who is subject to a currently binding liquidity constraint is someone for whom a marginal increase in cash will result in an immediate one-for-one increase in spending (a marginal propensity to consume, MPC, of 1). However, if the same consumer happened to have a large windfall transfer of cash (say, he wins the lottery), he would no longer be currently constrained, and his MPC would (presumably) be less than 1. In the case of precautionary saving, the ownership of an extra unit of wealth relaxes the suppression of consumption due to risk; this relaxation is more powerful for low-wealth consumers living on the edge of (precautionary) fear than for high-wealth consumers with plenty of resources. Thus, either liquidity constraints or precautionary motives or both will cause the consumption function to become concave (Carroll and Kimball, 2005). Huggett (2004) shows that consumption concavity in turn implies greater equilibrium wealth.

Empirical evidence indicates that the wealth distribution is highly concentrated. This means that the owners of much of the aggregate capital stock probably inhabit the portion of the consumption function to the far right, where it approaches the linear consumption function that characterizes the perfect foresight solution. Note, however, that this does not necessarily imply that aggregate consumption behaviour will resemble that of a perfect foresight consumer, because a large proportion of aggregate consumption is accounted for by

households with small amounts of market wealth. Spending of such households is probably determined much more by their permanent income than by their meagre wealth, and so it remains possible that a high proportion of consumption is performed by households inhabiting the more nonlinear part of the consumption function.

Empirical evidence

Euler equation methods

The early literature relevant to identifying the strength of precautionary motives tended to rely on Euler equation estimation (see Browning and Lusardi, 1996 for a survey), often by estimating regression equations of the form

$$\Delta \log C_{t+1} = \alpha_0 + \alpha_1 E_t[r_{t+1}] \tag{10}$$

and interpreting the coefficient on the interest rate term as an estimate of the inverse of the coefficient of relative risk aversion (CRAA) (which holds true under time-separable CRRA utility, as in equation (8)). However, this analysis did not take into account the dependence of higher-order terms like ϕ on the independent variables (see (9)). Some papers like Dynan (1993) attempted to account for precautionary contributions to consumption growth; but see Carroll (2001) for a critique of the whole Euler equation literature (including the second-order approach).

Structural estimation using micro data

A new methodology for estimating the importance of precautionary motives was pioneered by Gourinchas and Parker (2002) and Cagetti (2003) (with a related earlier contribution by Palumbo, 1999). Their idea was to calibrate an explicit life-cycle optimization problem using empirical data on the magnitude of household-level income shocks, and to search econometrically for the values of parameters such as the coefficient of relative risk aversion that maximized the model's ability to fit some measured feature of the empirical data. Gourinchas and Parker (2002) matched the profile of mean consumption over the lifetime; Cagetti (2003) matched the profile of median wealth. The intensity of the precautionary motive emerges, in each case, as an estimate of the coefficient of relative risk aversion, which Gourinchas and Parker (2002) put at about 1.4 and Cagetti (2003) finds to be somewhat larger (a value of 1 corresponds to logarithmic utility). One important caveat about these quantitative results is that the method's estimates of relative risk aversion depend on the model's assumption about the degree of risk households face. Recent work by Low, Meghir and Pistaferri (2005) that attempts to correct for measurement problems caused by job mobility suggest that the estimates of the magnitude of permanent shocks in Carroll and Samwick (1997) used for calibration by Gourinchas and Parker (2002) and Cagetti (2003) may be overstated by as much as 50 per cent. Re-estimation of the structural parameters using the Low, Meghir and Pistaferri (2005) calibration would generate larger estimates of relative risk aversion.

Regression evidence

A separate literature attempts direct empirical measurement of the relationship between uncertainty and wealth. To fix notation, index individual households by *i* and assume that uncertainty for household *i* in period *t* can be measured by some variable $\sigma_{t,i}$. Then in its simplest form the idea is to perform a regression of cash-on-hand on its determinants along the lines of

$$\log m_{t,i} = \sigma_{t,i}\gamma + Z_{t,i}\alpha + \varepsilon_{t,i} \tag{11}$$

where Z is some set of variables that capture life cycle, time series, and other non-precautionary effects. In principle, one can then calculate the predicted magnitude of m if everyone's uncertainty were set to zero (or some more sensible alternative like the minimum measured value of σ in the population).

This method permits the data to speak in a much less filtered way than the structural estimation approach. A drawback is that even if the magnitude of precautionary wealth could be estimated reliably and precisely, it would not be clear how to translate those estimates into a measure of relative risk aversion or some other set of behavioural parameters that could be used for analysing policy questions such as the optimal design of unemployment insurance or taxation.

A further disadvantage is that the method does not reliably yield the same answer in different data. Using a measure of subjective earnings uncertainty from a survey of Italian households, Guiso, Jappelli and Terlizzese (1992) estimate the precautionary component of wealth at only a few per cent, while Kazarosian (1997) and Carroll and Samwick (1998) estimate the precautionary component of wealth for typical US households to be in the range of 20–50 per cent. Hurst et al. (2005) argue that estimates of α are inordinately sensitive to whether business owners are included in the dataset; and work by Lusardi (1997; 1998) and Engen and Gruber (2001) implies much smaller precautionary wealth. Such large variation in empirical estimates is not plausibly attributable to actual behavioural differences across the various sample populations.

A problem that plagues all these efforts is identifying exogenous variations in uncertainty across households. The standard method has been to use patterns of variation across age, occupation, education, industry and other characteristics. This runs the danger that people who are more risk tolerant may both choose to work in a risky industry and choose not to save much, biasing downwards the estimate of the effect of an exogenous change in risk.

One recent paper attempts to get around this problem by using a natural experiment: Fuchs-Schündeln and Schündeln (2005) show that, before the collapse of the Berlin Wall, East German civil servants had similar income uncertainty to that faced by other East Germans. However, after the collapse of Communism, income uncertainty went up dramatically for most East Germans – but not for civil servants, who were given essentially the same risk-free jobs in the new merged government that they had had before the collapse. Fuchs-Schündeln and Schündeln (2005) show that, in accord with a model that includes substantial precautionary effects, saving rates of most East Germans increased sharply after unification, but saving rates of civil servants did not. By contrast, the West Germans – who would have been subject to more selection into jobs based on risk preferences – exhibited little difference in saving rates between civil servants and others with riskier jobs, either before or after reunification.

Survey evidence

Given the difficulties of obtaining reliable quantitative measures of precautionary motives using the revealed preference econometric techniques sketched above, some researchers have turned to approaches that involve asking survey participants more direct questions.

Kennickell and Lusardi (2005) find that, when respondents for the 1995 and 1998 US *Survey of Consumer Finances* are asked their target level of precautionary wealth, most have little difficulty in answering the question: desired precautionary wealth represents about eight per cent of total net worth and 20 per cent of total financial wealth. They find that respondents cite a broad array of risks in making their precautionary targets: in addition to labour income risk, they face health risk, business risk, and the risk of unavoidable expenditures (such as home repairs). (Consumers are clearly aware of the theoretical point that a given dollar of wealth can provide self-insurance against multiple different kinds of risks, since the risks are not likely to be perfectly correlated with each other.)

Carefully designed survey questions can in principle also be used to elicit information on the strength of underlying preferences (like risk aversion) that determine precautionary behaviour. The principle that whenever risk-bearing increases with assets, the precautionary saving motive (prudence) must be stronger than risk aversion provides an important theoretical lower bound on the degree of prudence. Using survey responses to hypothetical gambles over lifetime income in the *Health and Retirement Study*, Kimball, Sahm and Shapiro (2005) estimate that relative risk aversion has a median of 6.3 and a mean of 8.2. (Note that because of Jensen's inequality, the mean of relative risk aversion $E\rho$ is larger than the reciprocal of the mean of relative risk tolerance $\frac{1}{E(1/\rho)}$.) These estimates of relative risk aversion imply precautionary saving motives much stronger than those that have been used empirically to match observed wealth holdings. This discrepancy remains unresolved.

Conclusion

The qualitative and quantitative aspects of the theory of precautionary behaviour are now well established. Less agreement exists about the strength of the precautionary saving motive and the magnitude of precautionary wealth. Structural models that match broad features of consumption and saving behaviour tend to produce estimates of the degree of prudence that are less than those obtained from theoretical models in combination with risk aversion estimates from survey evidence. Direct estimates of precautionary wealth seem to be sensitive to the exact empirical procedures used, and are subject to problems of unobserved heterogeneity. Thus, establishing the intensity of the precautionary saving motive and the magnitude of precautionary wealth remain lively areas of debate.

See Also

- ambiguity and ambiguity aversion
- elasticity of intertemporal substitution
- Euler equations
- intertemporal choice
- permanent-income hypothesis
- time preference

Bibliography

Barsky, R.B., Gregory Mankiw, N. and Zeldes, S.P. 1986. Ricardian consumers with Keynesian propensities. *American Economic Review* 76, 676–91.

Browning, M.J. and Lusardi, A. 1996. Household saving: micro theories and micro facts. *Journal of Economic Literature* 34, 1797–855.

Cagetti, M. 2003. Wealth accumulation over the life cycle and precautionary savings. *Journal of Business and Economic Statistics* 21, 339–53.

Carroll, C.D. 2000. Solving consumption models with multiplicative habits. *Economics Letters* 68, 67–77.

Carroll, C.D. 2001. Death to the log-linearized consumption Euler equation! (and very poor health to the second-order approximation). *Advances in Macroeconomics* 1, Article 6.

Carroll, C.D. 2004. Theoretical foundations of buffer stock saving. Working Paper No. 10867. Cambridge, MA: NBER.

Carroll, C.D. and Kimball, M.S. 2005. Liquidity constraints and precautionary saving. Manuscript, Johns Hopkins University. Online. Available at http://econ.jhu.edu/ccarroll/liquidRevised.pdf, accessed 31 May 2007.

Carroll, C.D. and Samwick, A.A. 1997. The nature of precautionary wealth. *Journal of Monetary Economics* 40, 41–71.

Carroll, C.D. and Samwick, A.A. 1998. How important is precautionary saving? *Review of Economics and Statistics* 80, 410–19.

Drèze, J.H. and Modigliani, F. 1972. Consumption decisions under uncertainty. *Journal of Economic Theory* 5, 308–35.

Dynan, K.E. 1993. How prudent are consumers? *Journal of Political Economy* 101, 1104–13.

Engen, E. and Gruber, J. 2001. Unemployment insurance and precautionary saving. *Journal of Monetary Economics* 47, 545–79.

Fuchs-Schündeln, N. and Schündeln, M. 2005. Precautionary savings and self-selection: evidence from the German reunification ‘experiment’. *Quarterly Journal of Economics* 120, 1085–220.

Gourinchas, P.-O. and Parker, J. 2002. Consumption over the life cycle. *Econometrica* 70, 47–89.

Guiso, L., Jappelli, T. and Terlizzese, D. 1992. Earnings uncertainty and precautionary saving. *Journal of Monetary Economics* 302, 307–37.

Huggett, M. 2004. Precautionary wealth accumulation. *Review of Economic Studies* 71, 769–81.

Hurst, E., Lusardi, A., Kennickell, A. and Torralba, F. 2005. Precautionary savings and the importance of business owners. NBER Working Paper No. 11731.

Kazarosian, M. 1997. Precautionary savings – a panel study. *Review of Economics and Statistics* 79, 241–7.

Kennickell, A. and Lusardi, A. 2005. Disentangling the importance of the precautionary saving motive. Working paper, Dartmouth College.

Kimball, M.S. 1990. Precautionary saving in the small and in the large. *Econometrica* 58, 53–73.

Kimball, M.S., Sahm, C.R. and Shapiro, M.D. 2005. Imputing risk tolerance from survey responses. Manuscript, University of Michigan.

Kimball, M.S. and Weil, P. 2004. Precautionary saving and consumption smoothing across time and possibilities. Manuscript, University of Michigan.

Kreps, D.M. and Porteus, E.L. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46, 185–200.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112, 443–77.

Leland, H.E. 1968. Saving and uncertainty: the precautionary demand for saving. *Quarterly Journal of Economics* 82, 465–73.

Low, H., Meghir, C. and Pistaferri, L. 2005. Wage risk and employment over the life cycle. Manuscript, Stanford University.

Lusardi, A. 1997. Precautionary saving and subjective earnings variance. *Economics Letters* 57, 319–26.

Lusardi, A. 1998. On the importance of the precautionary saving motive. *American Economic Review* 88, 449–53.

Miller, B.L. 1976. The effect on optimal consumption of increased uncertainty in labor income in the multiperiod case. *Journal of Economic Theory* 13, 154–67.

Palumbo, M.G. 1999. Uncertain medical expenses and precautionary saving near the end of the life cycle. *Review of Economic Studies* 66, 395–421.

Sibley, D.S. 1975. Permanent and transitory effects of optimal consumption with wage income uncertainty. *Journal of Economic Theory* 11, 68–82.

Toché, P. 2005. A tractable model of precautionary saving in continuous time. *Economics Letters* 87, 267–72.

Zeldes, S.P. 1989. Optimal consumption with stochastic income: deviations from certainty equivalence. *Quarterly Journal of Economics* 104, 275–98.

How to cite this article

Carroll, Christopher D. and Miles S. Kimball. "precautionary saving and precautionary wealth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000359> doi:10.1057/9780230226203.1325

The New Palgrave Dictionary of Economics Online

predatory pricing

Janusz A. Ordover

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Predatory pricing is a response to a rival that sacrifices part of the profit that could be earned under competitive circumstances were the rival to remain viable, in order to lessen competition and gain consequent monopoly profit. The presence of intertemporal cost and/or demand linkages as well as network effects complicates the formulation of pricing rules that would distinguish legitimate from exclusionary pricing behaviour, and suggests that standard (non-strategic) models of markets do not necessarily offer much help in gauging the rationality of predation.

Keywords

above-cost pricing; antitrust policies; barriers to entry; chain-store paradox; entry; exit; incomplete information; increasing returns; intertemporal scope economies; marginal and average cost pricing; natural monopoly; network goods; predatory pricing; returns to scale; standardization; two-sided platforms

Article

Although neither courts nor legal and economic scholars agree on a broad definition of predatory behaviour, the minimal consensus (if such exists) is that predatory pricing entails selling a product ‘below cost’ in order to induce a rival's exit, or deter future entry or competition. More broadly, ‘predatory behaviour is a response to a rival that sacrifices part of the profit that could be earned under competitive circumstances, were the rival to remain viable, in order to [lessen competition] and gain consequent monopoly profit’ (Ordover and Willig, 1981, pp. 9–10).

The broader definition is necessary, at least in part because in many market scenarios, comparisons of prices to marginal cost offer little guidance as to what constitutes competitive, as opposed to predatory, pricing. A multi-product firm might offer one of a pair of complementary products at a price above incremental cost, and yet still be engaged in predation if the price at which it offers the pair as a bundle

is sufficiently higher than the incremental cost of the second component (see, for example, Baumol and Sidak, 1994, ch. 7). Another scenario is markets with intertemporal scope economies, as in Cabral and Riordan (1994), in which two firms race to exploit learning economies or establish their respective products as industry standard. In such markets, it may be profitable for the ‘leading’ firm to price below cost in order to induce the rival's exit. Discouraging such pricing would damage competition for the market for the sake of protecting competition in the market.

Similar issues arise in markets of network goods, where the product is more valuable to a user the more other people use it (for example, fax machines). These markets are characterized by increasing returns, and may be subject to a ‘tipping’ point at which one firm achieves natural monopoly, which is an efficient outcome since it increases consumer welfare by increasing network benefits through standardization. Farrell and Katz (2005) find that although rules to prevent predation, such as the Ordover–Willig rule, can improve welfare, they can also harm it in network markets ‘by preventing firms from internalizing the benefits of increasing returns to scale’.

In these cases, aggressive pricing is not designed to drive the rival into bankruptcy, but to make it realize that the ‘game’ is over from a strategic standpoint. When the market participants jockey for market leadership, pricing below short-run marginal cost (SRMC) could be a rational, non-predatory strategy. Evans and Schmalensee (2002) go so far as to advocate an approach under which, ‘if a defendant can establish that the relevant market is characterized by winner-take-all competition, then they have provided a complete defense against a charge of predatory behaviour’.

Another setting in which simple pricing rules can lead to wrong inferences involves pricing by so-called two-sided platforms, intermediaries that link two distinct groups of customers (for example, Rochet and Tirole, 2003; 2006; Armstrong, 2007). Such intermediaries frequently subsidize customers on one side in order to induce the other side to join the platform and enhance the value to all participants. Thus, below-cost pricing is compensated by above-cost pricing on the other side and by its impact on the overall level of activity on the platform. Such pricing may, of course, harm rivals who only operate on one side of the platform.

The presence of intertemporal cost and/or demand linkages as well as network effects of various kinds complicates the formulation of pricing rules that would sort out legitimate from exclusionary pricing behaviour. It also suggests that standard (non-strategic) models of markets do not necessarily offer much help in gauging the rationality of predation.

1 The apparent irrationality of predatory pricing

The Chicago School critique of traditional views of predatory pricing rested on the hypothesis that losses sustained during the predatory campaign will ordinarily exceed the more speculative gains from attempted supercompetitive pricing following the elimination of the prey. In his examination of *Standard Oil of New Jersey v. US* case, McGee (1958) pointed out that, in order for the predator to succeed in driving out an equally efficient rival, it must be prepared to serve the whole market by itself at an unremunerative price, while the prey can temporarily shut down its operations and restart them during the recoupment phase. Moreover, even if the prey exists permanently, productive assets may remain and could be purchased at scrap value by an opportunistic buyer. Easterbrook (1981) further observed that customers might protect themselves against post-predation exploitation by keeping the

prey in business, even if the product is available from the predator at a lower price, thereby denying the predator an opportunity to drive the rival out. And the prey may also be financed by lenders who (correctly) anticipate that, once the predator gives up, additional profits will be generated with which to repay the loan. McGee also noted that it is generally cheaper to purchase the rival rather than to prey on it. Hence, according to McGee, even if feasible, predation is irrational.

There are several problems with McGee's merger argument: buying a single rival may induce others to enter solely to be bought out at a premium (Rasmusen, 1985); the acquisition price itself may depend on the predator's established reputation for aggressive pricing (Burns, 1986; Saloner, 1987); there may be legal constraints on mergers so that when it is most advantageous, it is also likely to violate anti-merger legislation (Posner, 1976).

McGee's critique significantly influenced antitrust policies regulating pricing conduct, but stopped well short of offering a rigorous model in which predation was irrational. Selten's (1978) *chain store paradox* does so. Intuition suggests that an incumbent operating in a sequence of markets, each with an entrant, may predate in the first few markets to establish a ‘reputation’ for toughness and thereby deter the remaining entrants. The intuition fails, however, by ‘backward induction’: the entrant in the last market will correctly disregard the incumbent's behaviour in the preceding markets and conclude that its entry will be accommodated because, with no reputation to be concerned about, the incumbent has no reason to predate. The penultimate entrant reasons additionally that its predation will not deter entry into the next, and so it too enters, expecting to be accommodated. Inexorable logic leads to the conclusion that the incumbent will not predate and entry will occur in all the markets (see Ordover and Saloner, 1989; Phlips, 1995).

2 Economic models of rational predation

In settings that dispense with some of Selten's assumptions, predation can emerge as a rational strategy.

2.1 The long purse

One typical predatory pricing story involves an incumbent with a ‘deep pocket’, who by pricing aggressively can drive out a financially constrained rival (see Telser, 1966). In order to induce exit, the incumbent drops the price to the *rival’s* (*not necessarily its own*) *variable cost*. The rival, who also incurs fixed costs, soon exhausts its financial resources and leaves the market, enabling the incumbent to raise its price to monopoly level to recoup the costs of the predatory campaign.

Here, the mere threat of rational predation drives the opponent out. Clearly, a rational rival should leave at the first indication of predation, and not squander resources when exit is inevitable. In fact, rational firms with limited resources ought to stay out of a market occupied by an incumbent with a long purse. The ‘long purse’ story is not an entirely plausible basis for rational predation, but rather of entry deterrence, via a *credible threat* of post-entry predation – indeed, a costless one, as Benoit (1984) shows. The long-purse model also ignores the possibility of profit-seeking investors financing the preyed-upon firm, in order to extend its purse. In Bolton and Scharfstein (1990) and Fudenberg and Tirole (1986), the predator imposes losses on its prey in order to signal to investors that the prey is financially troubled. Even when everyone knows it is profitable for the rival to remain in the industry, Bolton and Scharfstein

argue that financial market predation induces exit because agency problems in financial contracting mean that reducing the sensitivity of the re-financing decision to the firm's performance exacerbates managerial incentive problems.

2.2 Predation for reputation

Other models operate by making ‘predation for reputation’ a *rational strategy* (see Ordover and Saloner, 1989; Milgrom and Roberts, 1990; Phlips, 1995, for more detailed analyses). For example, the game may have no ‘end’ from which to reason backwards (see Milgrom and Roberts, 1982b). Or there may be incomplete information, with different incumbent ‘types’, as in the seminal papers by Kreps and Wilson (1982), Milgrom and Roberts (1982b), and Kreps et al. (1982). A ‘weak’ incumbent, who would otherwise prefer to share a market, can falsely establish a ‘tough’ reputation by fighting at the first opportunity, and so convince all possible future entrants of its toughness and deters future entry, since if every incumbent were to predate, the reputational value of fighting would be dissipated and entry would occur, the probability of equilibrium predation must be positive, but less than one. This predatory story can be enriched in several ways: see, for example, Milgrom and Roberts (1982b) and Easley, Masson and Reynolds (1985), whose work is reviewed in Phlips (1995).

2.3 Signalling predation

Under imperfect information, predation can also be used to induce the rival's exit. For example, the rival may not have perfect knowledge of the incumbent's costs or its new product's demand. In these plausible market settings, the better-informed incumbent may price low in order to *signal* to the rival that exiting the market is preferable to staying (see for example, Milgrom and Roberts, 1982a). Even if low pricing does not deter entry, it may convince the rival to curtail its competitive ardour. Or, as Saloner (1987) shows, turning McGee's merger argument on its head, it may improve the terms of a buyout offer, by convincing the quarry that accepting a cheap offer is preferable to sharing a market with a low-cost competitor.

Signalling predation will be especially effective when the rival firm tries to gauge a new product's profitability from its reception in the ‘test market’. Firms with competing products will wish to ‘jam’ the signal (see Salop and Shapiro, 1980; Scharfstein, 1984; Roberts, 1986).

3 Empirical evidence

Recent empirical work has supported the rational predation models. A broad survey found that predatory pricing was present in 27 of 40 litigated cases in which the legal record was sufficiently informative (Zerbe and Mumford, 1996). In addition, several case studies, taken collectively, provide evidence that dominant firms have engaged in predatory behaviour, thereby undermining the Chicago School's claims about its irrationality.

Weiman and Levin (1994) provide evidence of predatory behaviour by Southern Bell Telephone Company from 1894 to 1912 when independent phone companies threatened entry. Genesove and

Mullin (2006) provide evidence of predatory behaviour in the American sugar refining industry before the First World War. They show that the price wars following two episodes of entry were predatory by comparing price to marginal costs and by constructing predicted competitive cost margins that they show to exceed observed margins. Granitz and Klein (1996) re-examine McGee's Standard Oil case and find evidence that Standard had in fact acted as a predator, by threatening to withhold crude shipments from any railroad that did not participate in a railroad cartel, in return receiving discounted shipping rates that left Standard's competitors to sell out at depressed prices.

Von Hohenbalken and West (1984) and West and Von Hohenbalken (1984) provide evidence that a leading Canadian supermarket chain engaged in a predatory location strategy. In a subsequent study (1986), they show that the strategy also gave the chain a reputation for aggressive pricing that deterred future entrants, thereby supporting the reputation model of Kreps and Wilson (1982) and Milgrom and Roberts (1982b), among others.

Burns (1986) similarly finds systematic empirical evidence supporting the theory that firms can acquire a reputation for following through on predatory commitments from past predatory behaviour. He finds that the American Tobacco Company from 1891 to 1906 set up bogus independents that it secretly controlled to sell at low prices in the prey's territories, thereby allowing the predator to maintain its monopoly by acquiring the assets of the prey, as well as other competitors not yet preyed upon, at artificially low prices. This study lends considerable credence to the view that predation can improve the terms of a takeover.

In contrast, Lott (1999) argues on the basis of an empirical survey of firms accused of predation between 1963 and 1982, that such firms did not have the necessary contractual and non-contractual arrangements to provide managers with incentives to engage in costly predatory behaviour, which should be necessary to lend credibility to the strategies. This critique is quite powerful but it goes deeper than just an attack on the credibility of predation. It suggests that, unless the principals (owners) can induce agents (managers) to forgo current profits for the sake of *any* future monopoly profits, managers may simply decide not to implement such strategies. On the other hand, even the most casual empiricism also suggests that such obstacles need not be insurmountable.

4 Legal tests for price predation

Price predation is easily confused with intense competition. Sharp demarcation lines are particularly difficult in strategic environments in which price predation could prove profitable. What should the public policy response be to the inherent difficulties in formulating antitrust rules governing dominant firm pricing? The legal–economic literature offers three distinct responses.

At one extreme, the Chicago School has urged removal of virtually all constraints on single firm pricing behaviour (as well as other forms of unilateral conduct). The rationale is simple: firms should not be discouraged from aggressively competing for and protecting their market positions. Further, since markets are quickly self-correcting (unless protected by governmental grants of monopoly), marketplace advantages unrelated to superior skill and efficiency are quickly driven away by competition. Consequently, anti-competitive conduct – including price predation – is, in general, irrational, and any attempts to forbid such conduct are likely to do more harm than good.

At the other extreme lies an open-ended, rule-of-reason analysis without any specific rules (see, for

example, Scherer, 1976; Comanor and Frech, 1984). There are serious problems with this approach, however. First, it is not clear whether it can be implemented effectively in the adversarial setting of antitrust litigation. Second, because it offers no standards for what constitutes lawful conduct, this approach complicates business planning and may increase incentives for the abuse of antitrust laws.
The third public policy response is consistent with most legal–economic commentary and judicial practice. Although the US courts have rarely found price predation, they have been unwilling to rule it out completely. Instead, the courts have adopted a set of ‘filters’ designed to screen potentially meritorious claims of anticompetitive pricing conduct from those that are probably without merit (see, for example, Joskow and Klevorick, 1979; Easterbrook, 1984; Baker, 1994; Elzinga and Mills, 1994).
The rest of this section discusses these filters, first reviewing proposed direct tests for predatory pricing and then addressing the question of ‘recoupment’ as a precondition for a finding of price predation.

4.1 Pricing tests

4.1.1 The Areeda – Turner test (Areeda and Turner 1975; 1978)

Areeda and Turner proposed that any price above ‘reasonably anticipated’ SRMC should be lawful, and any below, deemed predatory. US courts rapidly embraced this test, which is now a dominant test (see, for example, Areeda and Hovenkamp, 1993; Denger and Herfort, 1994; Green et al., 1996).
Because SRMC is difficult to estimate, Areeda and Turner recommend the average variable cost (AVC) as a workable surrogate. However, AVC is a good surrogate only when it does not diverge significantly from SRMC. Indeed, Areeda and Turner's analysis of the appropriate measures of AVC is inadequate because it does not derive correct cost concepts from the analysis of the predatory conduct itself and, consequently, fails to provide adequate guidance on how to treat such important components of costs as capital and advertising expenses (see Ordover and Willig, 1981; Baumol, 1996). The main problem with the Areeda–Turner test, however, is that it is based on an analysis of a firm's behaviour in a market situation – a temporary price cut by a single-product single-market firm – in which profitable predation is unlikely.

4.1.2 The Areeda-Turner paper

The Areeda and Turner paper generated a flow of alternative tests. For example, the *Joskow–Klevorick test* (1979) offers a two-tier test for price predation. The first step examines whether the structural preconditions for successful and rational predation exist. Because the first step eliminates many baseless claims, Joskow and Klevorick tighten the price comparison in the second step, and propose that any price below average *total* cost be presumptively illegal. The rationale is that in a competitive market, the equilibrium, long-run price will equal AVC and that, furthermore, it is unlikely that a post-entry price in a market predisposed to predation would be so low as to impose losses on the incumbent dominant firm. Some courts have used the Joskow–Klevorick test as an alternative to the Areeda–Turner test, especially when entry barriers are high. Moreover, an analysis of structural and other requirements for rational predation is now central to the analysis of a predatory pricing case.
Williamson (1977) and Baumol (1979) propose tests that isolate the *strategic* aspects of the incumbent's

responses to entry. Both would condemn ‘window shade’-type behaviour by the incumbent, that is, low price (high output) when the rival is in, followed by high price (low output) when the rival is out, and require that the dominant incumbent stick with its aggressive strategy for a prescribed period of time. These tests have not, however, been adopted by the courts.

Both these proposals can be criticized on various grounds (see Ordover and Saloner, 1989; Phlips, 1995). Areeda and Hovenkamp (1993) offer a spirited defence of the original Areeda–Turner rule against its critics and review the alternatives, which they find less desirable than the Areeda–Turner rule.

4.1.3 Above-cost predation and the Edlin test

In recent years a debate has ensued whether ‘above-cost’ pricing can also be predatory. In 1999, for example, the United States sued American Airlines on the theory that it was predatory to respond to entry with business practices that, even if above cost, ‘clearly’ sacrificed profits because it allegedly shifted airplanes from profitable routes to routes on which it was fighting the low-cost carriers. Edlin (2002) supports the move to prohibit above-cost predation because ‘[a]n incumbent monopoly with a significant cost or noncost advantage over entrants … can use these advantages to drive entrants from the market by pricing below their cost, but above its own’ and proposes a rule that would prohibit an incumbent monopoly, when faced with an entrant charging at least 20 per cent below the prevailing price, from cutting its own prices for 12–18 months or until it loses its monopoly position.

According to Edlin, this rule means that matching competitors’ prices after entry is no longer a cheap substitute for actually charging low prices in the first place, so consumers benefit. He explains that existing predation law means that the incumbent will not lower prices until there is an entrant and, since the potential entrant will not, in fact, enter, consumers always pay high prices to the incumbent. The predatory problem, he explains, occurs not after exit, but before entry. His rule, he argues, would address the problem by allowing firms that would otherwise fear being driven out of the market with above-cost predation to enter profitably, and it would benefit consumers because incumbents would charge lower prices to limit entry and because there would be more entry of competitors.

Elhauge (2003) responds that an above-cost pricing rule is ill-advised for three reasons: (*a*) it can often penalize efficient pricing behaviour when incumbents do not even have the market power to restrict output, for example when above-cost price cuts are an efficient response to deviations from the output-maximizing price-discrimination schedule in competitive markets; (*b*) it has mainly undesirable effects, such as raising post-entry prices and harming consumer welfare when the entrant is less efficient than the incumbent; and (*c*) it suffers from unavoidable implementation difficulties, such as ascertaining the moment of entry, dealing with quality changes designed to evade the restriction, and defining a post-entry price floor that will cause inefficiencies. He argues that part of the reason for the debate about whether to expand predation to above-cost pricing is ambiguity over the definition of ‘costs’ in the legal tests. He concludes that costs should be defined functionally as whichever cost measure assures that prices above costs cannot deter or drive out equally efficient rivals, a definition which he argues would resolve apparent anomalies in current predatory pricing law. Of course, from the standpoint of basic economics, it is always the ‘opportunity cost’ that provides the right measure of cost to be used. But this may be too much for the courts as calculations of opportunity costs are far from simple.

4.1.4 The recoupment test

The recoupment test is a potentially useful step in a summary judgement proceeding because it enables the fact-finder to dismiss allegations of predation without engaging in an extensive (and time-consuming) investigation of price-cost margins and other indicia of predatory *conduct*. On the other hand, the evidence that price is below the pertinent cost floor should perhaps obviate the need to enquire whether recoupment is feasible or not: the firm's conduct *reveals* its belief that recoupment is possible. In essence, the recoupment test substitutes the court's assessment of the likelihood of success for the independent business judgement of the alleged predator. Likewise, Hemphil (2001) argues that the recoupment analysis should not consider the firm's conduct at all, but rather should limit itself to an analysis of the structural features of the market, such as asymmetric information and linkages across markets, that might deter entry and allow the predator to profit from its ill-gotten monopoly once the competitor has been eliminated (see also Ordover and Willig, 1981). The recoupment test has also been criticized by Edlin and Farrell (2004) on the grounds that quantifying how the predator might benefit from its behaviour is difficult; courts should thus pay more attention, they argue, to serious consumer harm or harm to economic efficiency – ‘recoupment as harm’ rather than ‘recoupment as reality check’.

5 Critiques of the current legal test of predation

The Supreme Court's two-prong test in *Brooke Group* (1993) (price-cost and recoupment) created a high burden of proof for plaintiffs that solidified the Court's embrace of the Chicago School view that predatory pricing is ‘rarely tried, and even more rarely successful’. In the ensuing six years after *Brooke Group*, plaintiffs had not won a single predatory pricing case in federal court and, more striking, all but three of 39 reported decisions were dismissed or failed to survive summary judgment (see Brodley et al., 2000).

Brodley et al. (2000) criticize the courts for adhering to this ‘static, non-strategic view of predatory pricing’ at the same time that modern economic theory and empirical evidence have demonstrated the prevalence of predatory pricing. Based on this modern strategic theory, they propose a legal rule that, they argue, would augment existing practice in two critical respects: (1) it would explicitly permit proof of predation based on modern economics, and (2) it would expand the standard efficiencies and business justification defences to encompass pro-competitive dynamic gains, such as the learning-by-doing and network markets discussed earlier.

In reply, Elzinga and Mills (2001) fault the proposal for ignoring that predatory pricing is in practice very uncommon. They also argue that such theory lacks factual support and is not yet well developed enough to incorporate into antitrust rules. As a result, to permit predation to be proven by reference to modern strategic theory risks over-enforcement. (Bolton, Brodley and Riordan, 2000, respond by noting that the heavy factual burdens on the defendant and fully developed efficiencies defence available to the defendant mitigate the risk of over-enforcement.)

Marx and Shaffer (1999) argue that, for intermediate goods markets, the Supreme Court's two-prong test in *Brooke Group* may be over-inclusive, as low-cost pricing and recoupment can both occur with the rival supplier, although harmed, remaining in the market, and welfare actually increasing (because the increase in consumer surplus outweighs the reduction in overall joint profit associated with the pricing

distortion).

6 Conclusion

The three decades of research on predatory pricing since Areeda and Turner (1975) lead to the following policy lessons.
First, the strategic approach to modelling pricing debunked the comfortable position that predation is more costly to predator than prey, and hence irrational and unlikely to occur.
Second, given the non-competitive structure and asymmetries of information in the relevant markets, there is no bright line standard for predatory pricing that both proscribes pricing behaviour that reduces economic welfare and does not discourage pro-competitive behaviour.
Third, the focus on price predation to the exclusion of other types of business conduct seems misplaced, given the richness of strategies used by firms in their battles for market share (Ordover and Willig, 1981). Many of these strategies are likely to be more successful than price predation in inducing the exit of efficient rivals, and do not require sustained periods of losses.
Fourth, the courts' shift from vague inquiries into the 'intent' of the alleged predator's actions to more rigorous price and cost comparisons and assessments of the likelihood of recoupment has not benefited plaintiffs in predatory pricing litigation.

See Also

- game theory
- monopoly

Bibliography

Areeda, P.E. and Hovenkamp, H. 1986. *Antitrust Law: 1986 Supplement*. Boston: Little, Brown.

Areeda, P.E. and Hovenkamp, H. 1993. *Antitrust Law* (*Supplement*). Boston: Little, Brown.

Areeda, P.E. and Turner, D.F. 1975. Predatory pricing and related practices under section 2 of the • Sherman Act. *Harvard Law Review* 88, 697–733.

Areeda, P.E. and Turner, D.F. 1978. *Antitrust Law*. Boston: Little Brown.

Armstrong, M. 2007. Two-sided markets: economic theory and policy implications. In *Recent Developments in Antitrust*, ed. J.P. Choi. Cambridge, MA: MIT Press.

Baker, J.B. 1994. Predatory pricing after *Brooke Group*: an economic perspective. *Antitrust Law Journal* 64, 585–604.

Baumol, W.J. 1979. Quasi-permanence of price reductions: a policy for prevention of predatory pricing. *Yale Law Journal* 89, 1–26.

Baumol, W.J. 1996. Predation and the logic of the average variable cost test. *Journal of Law and Economics* 39, 49–72.

Baumol, W.J., Panzar, J.C. and Willig, R.D. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt, Brace Jovanovich.

Baumol, W.J. and Sidak, J.G. 1994. *Toward Competition in Local Telephony*. Cambridge, MA: MIT Press.

Benoit, J.P. 1984. Financially constrained entry in a game with incomplete information. *RAND Journal of Economics* 15, 490–99.

Bolton, P., Brodley, J.F. and Riordan, M.H. 2000. Predatory pricing: strategic theory and legal policy. *Georgetown Law Journal* 88, 2239–330.

Bolton, P. and Scharfstein, D. 1990. A theory of predation based on agency problems in financial contracting. *American Economic Review* 80, 93–106.

Bork, R. 1978. *Antitrust Paradox*. New York: Basic Books.

Brodley, J.F., Bolton, P. and Riordan, M.H. 2000. Predatory pricing: strategic theory and legal policy. *Georgetown Law Journal* 88, 2239.

Brooke Group Led v. Brown & Williamson Tobacco Corp., 113 S Ct. 2578 (1993).

Burns, M.R. 1986. Predatory pricing and the acquisition cost of competitors. *Journal of Political Economy* 94, 266–96.

Cabral, L.M.B. and Riordan, M.H. 1994. The learning curve, market dominance, and predatory pricing. *Econometrica* 62, 115–40.

Comanor, W.S. and Frech III, H.E. 1984. Strategic behavior and antitrust analysis. *American Economic Review* 74, 372–6.

Denger, M.L. and Herfort, J.A. 1994. Predatory pricing claims after *Brooke Group*. *Antitrust Law Journal* 62, 541–58.

Easley, D., Masson, R.T. and Reynolds, R.J. 1985. Preying for time. *Journal of Industrial Organization*

33, 445–60.

Easterbrook, F.H. 1981. Predatory strategies and counterstrategies. *University of Chicago Law Review* 48, 263–337.

Easterbrook, F.H. 1984. The limits of antitrust. *Texas Law Review* 63, 1–40.

Edlin, A. 2002. Stopping above-cost predatory pricing. *Yale Law Journal* 111, 941–91.

Edlin, A. and Farrell, J. 2004. The American airlines case: a chance to clarify predation policy. In *The Antitrust Revolution*, 4th edn, ed. J.E. Kwoka, Jr. and L.J. White. Oxford: Oxford University Press.

Elhauge, E. 2003. Why above-cost price cuts to drive out entrants are not predatory: and the implications for defining costs and market power. *Yale Law Journal* 112, 681–827.

Elzinga, K.G. and Mills, D.E. 1994. Trumping the Areeda–Turner test the recoupment standard in *Brooke Group*. *Antitrust Law Journal* 62, 559–84.

Elzinga, K.G and Mills, D.E. 2001. Predatory pricing and strategic theory. *Georgetown Law Journal* 89, 2475–93.

Evans, D.S. and Schmalensee, R. 2002. Some economic aspects of antitrust analysis in dynamically competitive industries. In *Innovation Policy and the Economy*, vol. 2, ed. A.B. Jae, J. Lerner and S. Stern. Cambridge, MA: NBER and MIT Press.

Farrell, J. and Katz, M. 2005. Competition or predation? Consumer coordination, strategic pricing and price floors in network markets. *Journal of Industrial Economics* 53, 203–31.

Fudenberg, D. and Tirole, J. 1986. A ‘signal-jamming’ theory of predation. *RAND Journal of Economics* 17, 366–76.

Genesove, D. and Mullin, W. 2006. Predation and its rate of return: the sugar industry, 1887–1914. *RAND Journal of Economics* 17, 366–76.

Granitz, E. and Klein, B. 1996. Monopolization by ‘raising rivals’ costs’: the standard oil case. *Journal of Law and Economics* 39, 1–47.

Green, W. and Ordover, J.A. 1996. *Predatory Pricing*. Monograph No. 22, Chicago: American Bar Association.

Green, W. et al. 1996. *Predatory Pricing*. Chicago: American Bar Association.

Hemphil, S. 2001. Note, the role of recoupment in predatory pricing analysis. *Stanford Law Review* 53, 1581–612.

Joskow, P.L. and Klevorick, A.K. 1979. A framework for analyzing predatory pricing policy. *Yale Law Journal* 89, 213–70.

Kreps, D. and Wilson, R. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27, 253–79.

Kreps, D., Milgrom, P., Roberts, J. and Wilson, R. 1982. Rational cooperation in the finitely-repeated Prisoner's Dilemma. *Journal of Economic Theory* 27, 245–52.

Lott, J., Jr. 1999. *Are Predatory Commitments Credible?* Chicago: University of Chicago Press.

Marx, L.M. and Shaffer, G. 1999. Predatory accommodation: below-cost pricing without exclusion in intermediate goods markets. *RAND Journal of Economics* 30, 22–43.

McGee, J. 1958. Predatory price cutting: the *Standard Oil* (NJ) case. *Journal of Law and Economics* 1, 137–69.

McGee, J. 1980. Predatory pricing revisited. *Journal of Law and Economics* 23, 289–330.

Milgrom, P. and Roberts, J. 1982a. Limit pricing and entry under incomplete information: an equilibrium analysis. *Econometrica* 50, 443–59.

Milgrom, P. and Roberts, J. 1982b. Predation, reputation and entry deterrence. *Journal of Economic Theory* 27, 280–312.

Milgrom, P. and Roberts, J. 1990. The new theories of predatory pricing. In *Industrial Structure in the New Industrial Economics*, ed. G. Bonnano and D. Brandolini. Oxford: Clarendon Press.

Ordover, J.A. and Saloner, G. 1989. Predation, monopolization and antitrust. In *Handbook of Industrial Organization*, ed. R. Schmalensee and R.D. Willig. New York: North-Holland.

Ordover, J.A. and Willig, R.D. 1981. An economic definition of predation: pricing and product innovation. *Yale Law Journal* 91, 8–53.

Ordover, J.A. and Willig, R.D. 1995. Economists’ view: the Department of Justice draft guidelines for the licensing and acquisition of intellectual property. *Antitrust Magazine* 9, 29–36.

Phlips, L. 1995. *Competition Policy: A Game-Theoretic Perspective*. Cambridge: Cambridge University

Press.

Posner, R. 1976. Predatory pricing. In *Antitrust Law: An Economic Perspective*, ed. R. Posner. Chicago: University of Chicago Press.

Rasmusen, E. 1985. Entry for buyout. Working paper, Graduate School of Management, UCLA.

Roberts, J. 1986. A signalling model of predatory pricing. *Oxford Economic Papers* (Suppl.), NS 38, 75–93.

Rochet, J-Ch. and Tirole, J. 2003. Platform competition in two-sided markets. *Journal of European Economic Association* 1, 990–1029.

Rochet, J-Ch. and Tirole, J. 2006. Two-sided markets: a progress report. *RAND Journal of Economics* 21, 172–87.

Saloner, G. 1987. Predation, mergers and incomplete information. *RAND Journal of Economics* 18, 165–86.

Salop, S.C. and Shapiro, C. 1980. A guide to test market predation. Working paper, Federal Trade Commission, Bureau of Economics.

Scharfstein, O. 1984. A policy to prevent rational test-market predation. *RAND Journal of Economics* 15, 229–43.

Scherer, F.M. 1970. *Industrial Markets Structure and Economic Performance*. Chicago: Rand McNally; 2nd edn, 1980.

Scherer, F.M. 1976. Predatory pricing and the Sherman Act: a comment. *Harvard Law Review* 89, 869–90.

Selten, R. 1978. The chain-store paradox. *Theory and Decision* 9, 127–59.

Standard Oil Co. of New Jersey v. US, 221 US 1 (1911).

Telser, L.G. 1966. Cutthroat competition and the long purse. *Journal of Law and Economics* 9, 259–77.

Von Hohenbalken, B. and West, D. 1984. Predation among supermarkets: an algorithmic locational analysis. *Journal of Urban Economics* 15, 244–57.

Von Hohenbalken, B. and West, D. 1986. Empirical tests for predatory reputation. *Canadian Journal of*

Economics 19, 160–78.

Weiman, D. and Levin, R. 1994. Preying for monopoly? The case of Southern Bell Telephone Company, 1894–1912. *Journal of Political Economy* 102, 103–26.

West, D. and Van Hohenbalken, B. 1984. Spatial Predation in a Canadian retail oligopoly. *Journal of Regional Science* 24, 415–29.

Williamson, O.E. 1977. Predatory pricing: a strategic and welfare analysis. *Yale Law Journal* 87, 284–340.

Zerbe, R., Jr. and Mumford, M. 1996. Does predatory pricing exist? Economic theory and the courts after *Brooke Group*. *Antitrust Bulletin* 41, 949–85.

How to cite this article

Ordover, Janusz A. "predatory pricing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000147> doi:10.1057/9780230226203.1326

The New Palgrave Dictionary of Economics Online

prediction formulas

Charles H. Whiteman and Kurt F. Lewis
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Prediction formulas for multi-step forecasts and geometric distributed leads of stationary time series are derived using classical, frequency domain methods. Starting with the Wold representation, optimal squared-error loss predictions are derived using the analytic function theory approach of Whittle. This approach is easily adapted to the problem of making predictions that are robust under model misspecification. Forecasts and expected present value calculations are illustrated under both objectives for low-order autoregressive and moving average processes.

Keywords

Blaschke factors; contour integral; cross-equation restrictions; distributed leads; frequency domain problems; least squares; linear least squares projection; minimum norm interpolation problem; min-max problem; misspecification; prediction formulas; rational expectations; Riesz–Fisher th; robustness; squared-error loss optimal prediction; time domain problems; Wiener–Hopf equation; Wold decomposition th; Wold representation

Article

1 Introduction

This article reviews the derivation of formulas for linear least squares and robust prediction of stationary time series and geometrically discounted distributed leads of such series. The derivations employed are the classical, frequency-domain procedures employed by Whittle (1983) and Whiteman (1983), and result in nearly closed-form expressions. The formulas themselves are useful directly in forecasting, and have also found uses in economic modelling, primarily in macroeconomics. Indeed, Hansen and Sargent (1980) refer to the cross-equation restrictions connecting the time series representation of driving variables to the analogous representation for predicting the present value of such variables as the ‘hallmark of rational expectations models’.

2 The Wold representation

Suppose that $\{x_t\}$ is a covariance-stationary stochastic process and assume (without loss of generality) that $Ex_t=0$. Covariance stationarity ensures that first and second unconditional moments of the process do not vary with time. Then, by the Wold decomposition theorem (see Sargent, 1987, for an elementary exposition and proof), x_t can be represented by:

$$x_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{1}$$

with

$$a_0 = 1, \sum_{j=0}^{\infty} a_j^2 < \infty$$

and

$$\varepsilon_t = x_t - P(x_t | x_{t-1}, x_{t-2}, \ldots), E\varepsilon_t^2 = \sigma^2$$

where $P(x_t | x_{t-1}, x_{t-2}, \ldots)$ denotes the linear least squares projection (population regression) of x_t on $x_{t-1}, x_{t-2}, \cdots$ Here, ‘represented by’ need not mean ‘generated by’, but rather ‘has the same variance and covariance structure as’. By construction, the ‘fundamental’ innovation ε_t is uncorrelated with information dated prior to t, including earlier values of the process itself: $E\varepsilon_t\varepsilon_{t-s} = 0 \quad \forall s > 0$. This fact makes the Wold representation very convenient for computing predictions. The convolution in (1) is often written $x_t=A(L)\,\varepsilon_t$ using the polynomial $A(L) = \Sigma_{j=0}^{\infty} a_j L^j$ in the ‘lag operator’ L, where $L\varepsilon_t = \varepsilon_{t-1}$.

3 Squared-error loss optimal prediction

The optimal prediction problem under squared-error loss can be thought of as follows. Given $\{x_t\}$ with the Wold representation (1) we want to find the stochastic process y_t,

$$y_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j} = C(L)\varepsilon_t$$

that will minimize the squared forecast error of the *h*-step ahead prediction

$$\min_{\{y_t\}} E(x_{t+h} - y_t)^2.$$

Equivalently, the problem can be written as

$$\min_{\{y_t\}} E(L^{-h}x_t - y_t)^2$$

or

$$\min_{\{c_j\}} E\left(L^{-h}\sum_{j=0}^{\infty} a_j \varepsilon_{t-j} - \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}\right)^2. \quad (2)$$

The problem in (2) involves finding a *sequence* of coefficients in the Wold representation of the unknown prediction process y_t, and is referred to as the *time domain problem*. By virtue of the Riesz–Fisher theorem (see again Sargent, 1987, for an exposition), the time-domain problem is equivalent to a *frequency domain problem* of finding an analytic function $C(z)$ on the unit disk $|z| \leq 1$ corresponding to the 'z-transform' of the $\{c_j\}$ sequence

$$C(z) = \sum_{j=0}^{\infty} c_j z^j$$

that solves

$$\min_{C(z)\in H^2} \frac{1}{2\pi i}\oint |z^{-h}A(z) - C(z)|^2 \frac{dz}{z} \tag{3}$$

where H^2 denotes the Hardy space of square-integrable analytic functions on the unit disk, and $\oint$ denotes (counterclockwise) integration about the unit circle. The requirement that $C(z) \in H^2$ ensures that the forecast is causal, and contains no future values of the ε 's; this is equivalent to the requirement that $C(z)$ have a well-behaved power series expansion in non-negative powers of z.

Each formulation of the problem is useful, as often one or the other will be simpler to solve. This stems from the fact that convolution in the time domain becomes multiplication in the frequency domain and vice versa. To see this, consider the two sequences $\{g_k\}_{k=-\infty}^{\infty}$ and $\{h_k\}_{k=-\infty}^{\infty}$. The convolution of $\{g_k\}$ and $\{h_k\}$ is the sequence $\{f_k\}$, in which a typical element would be:

$$f_k = \sum_{j=-\infty}^{\infty} g_j h_{k-j}.$$

The z-transform of the convolution is given by

$$\begin{aligned}
\sum_{k=-\infty}^{\infty} f_k z^k &= \sum_{k=-\infty}^{\infty}\left(\sum_{j=-\infty}^{\infty} g_j h_{k-j}\right) z^k \\
&= \sum_{k=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} g_j z^j h_{k-j} z^{k-j} \\
&= \sum_{(k-j)=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} g_j z^j h_{k-j} z^{k-j} \\
&= \sum_{s=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} g_j z^j h_s z^s \text{ (Substituting } s = k - j) \\
&= \sum_{s=-\infty}^{\infty} h_s z^s \sum_{j=-\infty}^{\infty} g_j z^j \\
&= g(z)h(z) .
\end{aligned}$$

Thus the ‘*z*-transform’ of the convolution of the sequences $\{g_k\}$ and $\{h_k\}$ is the product of the *z*-transforms of the two sequences. Similarly, the *z*-transform of the product of two sequences is the convolution of the *z*-transforms:

$$\sum_{k=-\infty}^{\infty} g_k h_k z^k = \frac{1}{2\pi i}\oint g(p)h(z/p)\frac{dp}{p}.$$

To see why this is the case, note that

$$g(p)h(z/p)p^{-1} = \sum_{j=-\infty}^{\infty} g_j p^j \sum_{k=-\infty}^{\infty} h_k z^k p^{-k-1},$$

implying

$$\frac{1}{2\pi i}\oint g(p)h(z/p)p^{-1}dp = \frac{1}{2\pi i}\oint \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} g_j h_k z^k p^{j-k-1}dp.$$

But all of the terms vanish except where *j*=*k* because

$$\frac{1}{2\pi i}\oint z^k \frac{dz}{z} = 0$$

except when *k*=0. To see why, let $z = e^{i\theta}$. As θ increases from 0 to 2π, *z* goes around the unit circle. So, since $dz = ie^{i\theta}d\theta$, we have that

$$\frac{1}{2\pi i}\oint z^k \frac{dz}{z} = \frac{i}{2\pi i}\oint e^{i\theta k}d\theta = \begin{cases} 1 & \text{if } k = 0 \\ \frac{1}{2\pi}\frac{1}{ik}e^{i\theta k}\Big|_0^{2\pi} = 0 & \text{otherwise}. \end{cases}$$

Thus,

$$\frac{1}{2\pi i}\oint g(p)h(z/p)p^{-1}dp = \sum_{j=-\infty}^{\infty} g_j h_j z^j \frac{1}{2\pi i}\oint \frac{dp}{p} = \sum_{j=-\infty}^{\infty} g_j h_j z^j$$

by Cauchy's Integral formula.

The frequency domain formulas can now be used to calculate moments quickly and conveniently. Consider Ex_t^2:

$$Ex_t^2 = E(A(L)\varepsilon_t)^2 = E\left(\sum_{j=0}^{\infty} A_j \varepsilon_{t-j}\right)^2 = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} A_j^2. \tag{4}$$

The result in eq. (4) comes from the fact that $E\varepsilon_t\varepsilon_{t-s} = 0, \quad \forall s \neq 0$. Using the product-convolution relation, we see that

$$\sum_{j=0}^{\infty} A_j^2 = \sum_{j=0}^{\infty} A_j^2 z^j|_{z=1} = \frac{1}{2\pi i}\oint A(p)A(z/p)\frac{dp}{p}|_{z=1} = \frac{1}{2\pi i}\oint A(p)A(p^{-1})\frac{dp}{p} = \frac{1}{2\pi i}\oint |A(z)|^2\frac{dz}{z}. \tag{5}$$

Returning to the prediction problem, the task is to choose $c_0, c_1, c_2, \ldots$ to

$$\min_{\{c_j\}} \frac{1}{2\pi i}\oint \left| z^{-h}A(z) - \sum_{j=0}^{\infty} c_j z^j \right|^2 \frac{dz}{z}. \tag{6}$$

The first order conditions for the optimization in expression (7) are

$$0 = \frac{1}{2\pi i}\oint \left\{ z^j[z^h A(z^{-1}) - C(z^{-1})] + z^{-j}[z^{-h}A(z) - C(z)] \right\}\frac{dz}{z} = \frac{1}{2\pi i}\oint z^{-j}[z^{-h}A(z) - C(z)]\frac{dz}{z} - \frac{1}{2\pi i}\oint p^{-j}[p^{-h}A(p) - C(p)]\frac{dp}{p} \tag{7}$$

for $j = 0, 1, 2, \ldots$, where the second integral is the result of a change of variable $p = z^{-1}$ so that $dp = -z^{-1}dz$, resulting in

$$\frac{dp}{p} = z\left(-z^{-2}dz\right) = -\frac{dz}{z}.$$

The result is that in the second integral, the direction of the contour integration is clockwise. Multiplying by −1 and integrating counterclockwise, the second integral becomes identical to the first, and we can write the set of first-order conditions as

$$0 = \frac{1}{\pi i}\oint z^{-j}[z^{-h}A(z) - C(z)]\frac{dz}{z}\ j = 0, 1, 2, \ldots$$

(8)

Define $F(z)$ such that

$$F(z) = z^{-h}A(z) - C(z) = \sum_{j=-\infty}^{\infty} F_j z^j.$$

From eq. (8), it must be the case that all coefficients on non-negative powers of z equal zero:

$$F_j = 0,\ j = 0, 1, 2, \ldots.$$

Multiplying by z^j and summing over *all* $j = 0, \pm 1, \pm 2, \ldots$, we obtain

$$F(z) = \sum_{-\infty}^{-1}$$

(9)

where the term on the right-hand-side of (9) represents an unknown function in negative powers of z. Thus

$$z^{-h}A(z) - C(z) = \sum_{-\infty}^{-1},$$

which is an example of a ‘Wiener–Hopf’ equation. Now apply the (linear) ‘plussing’ operator, $[\cdot]_+$, which means ‘ignore negative powers of z’. The unknown function in negative powers of z is ‘annihilated’ by this operation, resulting in

$$C(z) = \left[z^{-h}A(z)\right]_+ = \left[z^{-h}a_0 + z^{-h+1}a_1 + z^{-h+2}a_2 + \ldots\right]_+ = \left[z^0 a_h + z^1 a_{h+1} + z^2 a_{h+2} + \ldots\right] = \sum_{j=h}^{\infty} a_j z^{j-h} = z^{-h}A(z) - pr[z^{-h}A(z)]$$

where $pr[z^{-h}A(z)]$ is the principal part of the Laurent expansion of $z^{-h}A(z)$ about z=0. (The principal part of the Laurent expansion about z=0 is the part involving negative powers of z.) This provides a very simple formula for computing forecasts.

3.1 AR(1) example

Suppose that $x_t = ax_{t-1} + \varepsilon_t$. This means that $A(z)=1/(1-az)$. In this case:

$$C(z) = \left[z^{-h}A(z)\right]_+ = \left[z^{-h}(1 + az + a^2z^2 + \ldots)\right]_+ = a^h(1 + az + a^2z^2 + \ldots) = \frac{a^h}{(1 - az)}$$

and the least squares loss predictor of x_{t+h} using information dated t and earlier is

$$P_t^{LS}x_{t+h} = y_t = C(L)\varepsilon_t = C(L)A^{-1}(L)x_t = a^h x_t.$$

The forecast error is

$$x_{t+h} - a^h x_t = \varepsilon_{t+h} + a\varepsilon_{t+h-1} + \ldots + a^{h-1}\varepsilon_{t+1},$$

which is serially correlated (for $h \geq 2$), but not correlated with information dated t and earlier.

3.2 MA(1) example

Supposed that $x_t = \varepsilon_t - \alpha\varepsilon_{t-1}$, meaning $A(z)=1-az$. Thus,

$$C(z) = \left[z^{-h}A(z)\right] = \left[z^{-h}(1-\alpha z)\right] = \begin{cases}\alpha \text{ if } h = 1, \\ 0 \text{ otherwise}.\end{cases}$$

So, the best one-step ahead predictor is

$$\alpha\varepsilon_t = \alpha(1 + \alpha L + \alpha^2 L^2 + \ldots)x_t$$

and the best predictor for forecasts of horizon two or more is exactly zero. For two-step-ahead (and beyond) prediction, the forecast error is x_{t+h} itself, which is serially correlated but not correlated with information dated *t* and earlier.

4 Least squares prediction of geometric distributed leads

A prediction problem that characterizes many models in economics involves the expectation of a discounted value. Perhaps the most common and widely studied example is the present value formula for stock prices. Abstracting from mean and trend, suppose the dividend process has a Wold representation given by

$$d_t = \sum_{j=0}^{\infty} q_j \varepsilon_{t-j} = q(L)\varepsilon_t \quad E(\varepsilon_t) = 0, \; E(\varepsilon_t^2) = 1. \tag{10}$$

Assuming that the constant discount factor is given by γ, we have the present value formula

$$p_t = E_t \sum_{j=0}^{\infty} \gamma^j d_{t+j} = E_t\left(\frac{q(L)}{1-\gamma L^{-1}}\varepsilon_t\right) = E_t(p_t^*). \tag{11}$$

The least-squares minimization problem the predictor faces is to find a stochastic process p_t to minimize the expected squared prediction error $E(p_t - p_t^*)^2$. In terms of the information known at date t, the agent's task is to find a linear combination of current and past dividends, or, equivalently, of current and past dividend innovations ε_t, that is 'close' to p_t^*. Writing $p_t = f(L)\varepsilon_t$, the problem becomes one of finding the coefficients f_j in $f(L) = f_0 + f_1 L + f_2 L^2 + \ldots$ to minimize $E(f(L)\varepsilon_t - p_t^*)^2$. Using the method described in the previous section, the problem has an equivalent, frequency-domain representation

$$\min_{f(z)\in H^2} \frac{1}{2\pi i}\oint \left| \frac{q(z)}{1-\gamma z^{-1}} - f(z) \right|^2 \frac{dz}{z}. \tag{12}$$

The first-order conditions for choosing f_j are, after employing the same simplification used in (7),

$$-\frac{2}{2\pi i}\oint z^{-j}\left[\frac{q(z)}{1-\gamma z^{-1}} - f(z)\right]\frac{dz}{z} = 0,\ j = 0, 1, 2, \ldots. \tag{13}$$

Now define

$$H(z) = \frac{q(z)}{1-\gamma z^{-1}} - f(z)$$

so that (13) becomes

$$-\frac{2}{2\pi i}\oint z^{-j} H(z)\frac{dz}{z} = 0.$$

Then multiplying by z^j and summing over all $j = 0, \pm 1, \pm 2, \ldots$ as above, we obtain

$$H(z) = \frac{q(z)}{1 - \gamma z^{-1}} - f(z) = \sum_{-\infty}^{-1},$$

the Wiener–Hopf equation for this problem. Applying the plussing operator to both sides yields

$$\left[\frac{q(z)}{1 - \gamma z^{-1}}\right]_+ - [f(z)]_+ = 0$$

implying

$$f(z) = \left[\frac{q(z)}{1 - \gamma z^{-1}}\right]_+ = \left[\frac{zq(z)}{z - \gamma}\right]_+$$

because $f(z)$ is, by construction, one-sided in non-negative powers of z. As in the previous section,

$$[A(z)]_+ = A(z) - P(z)$$

where $P(z)$ is the principal part of the Laurent series expansion of $A(z)$. To determine the principal part of $[(z - \gamma)^{-1} zq(z)]$, note that $zq(z)$ has a well-behaved power series expansion about $z = \gamma$, where 'well-behaved' means 'involving no negative powers of $(z - \gamma)$'. Thus $[(z - \gamma)^{-1} zq(z)]$ has a power series expansion about $z = \gamma$ involving a single term in $(z - \gamma)^{-1}$:

$$\left(\frac{zq(z)}{z - \gamma}\right) = \frac{b_{-1}}{z - \gamma} + b_0 + b_1(z - \gamma)^1 + b_2(z - \gamma)^2 + \ldots.$$

The principal part here is the part involving negative powers of $(z - \gamma)$: $b_{-1}(z - \gamma)^{-1}$. To determine it, multiply both sides by $(z - \gamma)$ and evaluate what is left at

$z = \gamma$ to find $b_{-1} = \gamma\, q(\gamma)$. Thus

$$f(z) = \left[\frac{q(z)}{1 - \gamma z^{-1}}\right]_+ = \left[\frac{zq(z)}{z - \gamma}\right]_+ = \frac{zq(z) - \gamma q(\gamma)}{z - \gamma}. \tag{14}$$

The ‘cross-equation restrictions’ of rational expectations refer to the connection between the serial correlation structure of the driving process (here dividends) and the serial correlation structure of the expected discounted value of the driving process (here prices). That is, when dividends are characterized by $q(z)$, prices are characterized by $f(z)$, and $f(z)$ depends upon $q(z)$ as depicted in (14).

To illustrate how the formula works, suppose detrended dividends are described by a first-order autoregression; that is, that $q(L) = (1 - \rho L)^{-1}$. Then

$$p_t = f(L)\varepsilon_t = \frac{Lq(L) - \gamma q(\gamma)}{L - \gamma}\varepsilon_t = \left(\frac{1}{1 - \rho\gamma}\right)d_t. \tag{15}$$

It is instructive to note that, while the pricing formula (15) makes p_t the best least squares predictor of p_t^*, the prediction errors $p_t - p_t^*$ will not be serially uncorrelated. Indeed

$$p_t - p_t^* = \gamma\left\{\frac{Lq(L) - \gamma q(\gamma)}{L - \gamma} - \frac{q(L)}{1 - \gamma L^{-1}}\right\}\varepsilon_t = \frac{-\gamma^2 q(\gamma)}{L - \gamma}\varepsilon_t = -\gamma^2 q(\gamma)\frac{L^{-1}}{1 - \gamma L^{-1}}\varepsilon_t = -\gamma^2 q(\gamma)\left\{\varepsilon_{t+1} + \gamma\varepsilon_{t+2} + \gamma^2\varepsilon_{t+3} + \dots\right\}.$$

Thus the prediction errors will be described by a highly persistent (γ is close to unity) first-order autoregression. But because this autoregression involves *future* ε_t’s, the serial correlation structure of the errors cannot be exploited to improve the quality of the prediction of p_t^*. The reason is that the predictor ‘knows’ the *model* for price setting (the present value formula) and the dividend process; the best predictor $p_t = E_t p_t^*$ of p_t^* ‘tolerates’ the serial correlation because the (correct) model implies that it involves *future* ε_t's and therefore cannot be predicted. If one only had data on the errors (and did not know the model that generated them), they would appear (rightly) to be characterized by a first-order autoregression; fitting an AR(1) (that is, the best *linear* model) and using it to ‘adjust’ p_t by accounting for the serial correlation in the errors $p_t - p_t^*$ would decrease the quality of the estimate of p_t^*. The reason is the usual one that the Wold representation for $p_t - p_t^*$ is not the economic model of $p_t - p_t^*$, and (correct) models always beat Wold representations. This also

serves as a reminder of circumstances under which one should be willing to tolerate serially correlated errors: when one knows the model that generated them, and the model implies that they are as small as they can be made.

5 Robust optimal prediction of time series

The squared-error loss function employed to this point is appropriate for situations in which the model (either the time series model or the economic model) is thought to be correct. But in many settings the forecaster or model builder may wish to guard against the possibility of misspecification. There are many ways to do this; an approach popular in the engineering literature and recently introduced into the economics literature by Hansen and Sargent (2007) involves behaving so as to minimize the maximum loss sustainable by using an approximating model when the truth may be something else. The ‘robust’ approach to this involves replacing the squared-error loss problem

$$\min_{\{C(z)\}} \frac{1}{2\pi i} \oint \left| z^{-h} A(z) - C(z) \right|^2 \frac{dz}{z}$$

with the ‘min-max’ problem

$$\min_{\{C(z)\}} \sup_{|z|=1} \left| z^{-h} A(z) - C(z) \right|^2,$$

so that minimizing the ‘average’ value on the unit circle has been replaced by minimizing the max. This problem can also be written

$$\min_{\{C(z)\}} \sup_{|z|=1} \left| A(z) - z^{h} C(z) \right|^2.$$

This is known as the ‘minimum norm interpolation problem’ and amounts to finding a function $\phi(z)$ to

$$\min \| \varphi(z) \|_{\infty}$$

subject to the restriction that the power series expansion of $\phi(z)$ matches that of $A(z)$ for the first h–1 powers of z. This means that the following must hold:

$$\sum_{j=0}^{h-1} \varphi_j z^j = \sum_{j=0}^{h-1} a_j z^j. \tag{16}$$

Theorem 5.1: The minimizing $\phi(z)$ *function is such that* $|\varphi(z)|^2$ *is constant on* $|z| = 1$. *Moreover,*

$$\varphi(z) = M \prod_{j=1}^{h} \frac{z - \alpha_j}{1 - \overline{\alpha}_j z}$$

where $M, \alpha_1, \alpha_2, \ldots, \alpha_n$ *are chosen to ensure that* (*16*) *holds*.
Proof: : see Nehari (1957).
To see that $\phi(z)$ must be of the indicated form, note that the 'Blaschke factors' in the product have unit modulus:

$$\frac{z - \alpha_j}{1 - \alpha_j z}\left(\frac{z^{-1} - \overline{\alpha}_j}{1 - \alpha_j z^{-1}}\right) = \left(\frac{z - \alpha_j}{1 - \alpha_j z}\right)(z^{-1} z)\left(\frac{z^{-1} - \overline{\alpha}_j}{1 - \alpha_j z^{-1}}\right) = \left(\frac{1 - \alpha_j z^{-1}}{1 - \alpha_j z}\right)\left(\frac{1 - \alpha_j z}{1 - \alpha_j z^{-1}}\right) = 1,$$

so that $|\varphi(z)|^2 = M^2$.
In the general *h*-step-ahead prediction problem, we have that

$$\varphi(z) = M \prod_{j=1}^{h-1} \frac{z - \alpha_j}{1 - \overline{\alpha}_j z} = A(z) - z^h C(z),$$

meaning that

$$C(z) = \frac{1}{z^h}\left(A(z) - M\prod_{j=1}^{h-1}\frac{z-\alpha_j}{1-\bar{\alpha}_j z}\right).$$

This is analogous to the solution in the least-squares case, but, instead of subtracting the principal part of $z^{-h}A(z)$, we subtract a different function from $z^{-h}A(z)$. Note also that because

$$M\prod_{j=1}^{h-1}\frac{z-\alpha_j}{1-\bar{\alpha}_j z}$$

matches the power series expansion of $A(z)$ up to the power z^{h-1}, $C(z)$ is of the form

$$C(z) = c_0 + c_1 z + c_2 z^2 + \ldots$$

Finally, note that the forecast error is serially uncorrelated because $\phi(z)$ is constant on $|z| = 1$.

5.1 Example. AR(1)

Let

$$A(z) = \frac{1}{1-az}.$$

For h=1, we see that $\phi(z)=A(z)-zC(z)$ must be constant on $|z| = 1$, and that $\phi(0)=A(0)=1$. Thus, $\phi(z)=M=1$, so that

$$C(z) = \frac{A(z)-1}{z} = \frac{az}{(1-az)z} = \frac{a}{1-az},$$

which implies that the robust one-step ahead forecast is

$$y_t^R = ax_t,$$

which coincides with the best least-squares forecast. This equivalence between the robust and least-squares one-step ahead forecasts is to be expected because the best one-step-ahead least-squares forecast also has serially uncorrelated errors. For *h*=2, we have that

$$\varphi(z) = \frac{M(z-\alpha)}{1-\bar{\alpha}z}$$

where (again) $\phi(0)=1$, but now we also see that $\phi'(0)=a$. Thus,

$$\varphi(0) = 1 = -\alpha M \Rightarrow M = -\frac{1}{\alpha},$$

and furthermore

$$\varphi'(0) = a = \frac{(1-\bar{\alpha}z)M - M(z-\alpha)(-\bar{\alpha})}{(1-\bar{\alpha}z)^2}\Big|_{z=0} = M - M(\alpha\bar{\alpha}) = M(1-\alpha\bar{\alpha}).$$

Therefore, the solution will have the property that

$$a = -\frac{1}{\alpha}(1-\alpha\bar{\alpha}) - a\alpha = 1 - \alpha\bar{\alpha}0 = 1 + a\alpha - \alpha\bar{\alpha}.$$

That is, the roots are reciprocal pairs. Notice that the discriminant is positive $(a^2\alpha^2 + 4\alpha\bar{\alpha} > 0)$, meaning that we will always have a real solution, and we

choose $|\alpha| < 1$. Then, we have that

$$C(z) = \frac{1}{z^2}\left[\frac{1}{1-az} - \frac{M(z-\alpha)}{1-\alpha z}\right] = \frac{1}{z^2}\frac{1-\alpha z-(1-az)(1-\frac{1}{\alpha}z)}{(1-az)(1-\alpha z)} = \frac{1-\alpha z-1+az+\frac{1}{\alpha}z-\frac{a}{\alpha}z^2}{z^2(1-az)(1-\alpha z)} = \frac{-\frac{a}{\alpha}}{(1-az)(1-\alpha z)}.$$

So, the robust prediction is given by

$$P_t^R x_{t+2} = -\frac{a}{\alpha}\sum_{j=0}^{\infty}\alpha^j x_{t-j},$$

in contrast to the least-squares prediction

$$P_t^{LS} x_{t+2} = a^2 x_t.$$

5.2 Example. MA(1)

Suppose that the process follows an MA(1), $x_t = \varepsilon_t - \beta\varepsilon_{t-1}$, and therefore $A(z)=1-\beta z$. The analysis from the previous example still holds, and all of the following are true:

$$\varphi(z) = \frac{M(z-\alpha)}{1-\bar{\alpha}z}$$

while

$$\varphi(0) = 1 = -\alpha M \Rightarrow M = -\alpha^{-1}$$

and

$$\varphi'(0) = -\beta = -\frac{1}{\alpha}(1-\alpha\overline{\alpha}).$$

Therefore,

$$0 = 1 - \alpha\beta - \alpha\overline{\alpha},$$

meaning that, again, we have real roots which are reciprocal pairs and we can choose $|\alpha| < 1$. Of course, α will depend upon the value of β, and we write $\alpha(\beta)$. Thus

$$C(z) = \frac{1}{z^2}\left[1 - \beta z - \frac{M(z-\alpha(\beta))}{(1-\alpha(\beta)z)}\right] = \frac{1}{z^2}\left[\frac{(1-\beta z)(1-\alpha(\beta)z)) - M(z-\alpha(\beta))}{1-\alpha(\beta)z}\right] = \frac{1}{z^2}\left[\frac{1-\beta z-\alpha(\beta)z+\beta\alpha(\beta)z^2 - Mz + M\alpha(\beta)}{1-\alpha(\beta)z}\right] = \frac{\beta\alpha(\beta)}{1-\alpha(\beta)z}.$$

Therefore, we have the robust prediction

$$P_t^R x_{t+2} = \frac{\beta\alpha(\beta)}{1-\alpha(\beta)L}\varepsilon_t = \frac{\beta\alpha(\beta)}{1-\alpha(\beta)L}[x_t + \beta x_{t-1} + \beta x_{t-2} + \ldots],$$

while the least-squares prediction is the standard

$$P_t^{LS} x_{t+2} = 0.$$

6 Robust prediction of geometric distributed leads

Following the excellent treatment in Kasa (2001), a robust present-value predictor fears that dividends may not be generated by the process in (10), and so, instead of choosing an $f(z)$ to minimize the average loss around the unit circle, chooses $f(z)$ to minimize the maximum loss:

$$\min_{f(z)\in H^{\infty}} \sup_{|z|=1} \left| \frac{q(z)}{1-\gamma z^{-1}} - f(z) \right|^2 \Leftrightarrow \min_{f(z)\in H^{\infty}} \sup_{|z|=1} \left| \frac{zq(z)}{z-\gamma} - f(z) \right|^2.$$

Unlike in the least squares case (14), where $f(z)$ was restricted to the class H^2 of functions finitely square integrable on the unit circle, the restriction now is to the class of functions with finite maximum modulus on the unit circle, and the H^2 norm has been replaced by H^{∞} norm.
To begin the solution process, note that there is considerable freedom in designing the minimizing function $f(z)$: it must be well-behaved (that is, must have a convergent power series in non-negative powers of z on the unit disk), but is otherwise unrestricted. Recalling the Laurent expansion

$$\frac{zq(z)}{z-\gamma} = \frac{b_{-1}}{z-\gamma} + b_0 + b_1(z-\gamma) + b_2(z-\gamma)^2 + \ldots,$$

while in the least squares case $f(z)$ was set to 'cancel' all the terms of this series except the first, here $f(z)$ will be set to do something else. Now define the Blaschke factor $B_{\gamma}(z) = (z-\gamma)/(1-\gamma z)$ and note that, because of the unit modulus condition, the problem can be written

$$\min_{\{f(z)\}} \sup_{|z|=1} \left| \frac{zq(z)}{1-\gamma z} - \frac{z-\gamma}{1-\gamma z} f(z) \right|^2.$$

Defining

$$T(z) = \frac{zq(z)}{1-\gamma z}$$

we have

$$\min_{f \in H^{\infty}} \sup_{|z|=1} |T(z) - B_{\gamma}(z) f(z)| \Leftrightarrow \min_{f \in H^{\infty}} \|T(z) - B_{\gamma}(z) f(z)\|_{\infty}.$$

Define the function inside the ||'s as

$$\varphi(z) = T(z) - B_{\gamma}(z) f(z)$$

and note that $\phi(\gamma)=T(\gamma)$. Thus the problem of finding $f(z)$ reduces to the problem of finding the smallest $\phi(z)$ satisfying $\phi(\gamma)=T(\gamma)$:

$$\min_{\varphi \in H^{\infty}} \|\varphi(z)\|_{\infty} \text{ s.t. } \varphi(\gamma) = T(\gamma)$$

Theorem 6.1: (Kasa, 2001). *The solution to* (*17*) *is the constant function* $\phi(z)=T(\gamma)$.
Proof: . To see this, first note that the norm of a constant function is the modulus of the constant itself. This is written as

$$\|\varphi(z)\|_{\infty} = \|T(\gamma)\|_{\infty} = |T(\gamma)|^2. \quad (17)$$

Next, suppose that there exists another function $\Psi(z) \in H^{\infty}$, with $\Psi(\gamma) = T(\gamma)$ and also

$$\|\Psi(z)\|_{\infty} < \|\varphi(z)\|_{\infty}. \quad (18)$$

Recall the definition of the H^{∞} norm, and using equations (17) and (18):

$$\|\Psi(z)\|_\infty = \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2.$$

The maximum modulus theorem states that a function *f* which is analytic on the disk *U* achieves its maximum on the boundary of the disk. That is

$$\sup_{z\in U} |f(z)|^2 \le \sup_{z\in \partial U} |f(z)|^2.$$

Therefore, we can see that

$$\sup_{|z|<1} |\Psi(z)|^2 \le \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2.$$

However, one of the values on the interior of the unit disk is $z=\gamma$, which can be inserted into the far left-hand-side of eq. (6) to get the result

$$|\Psi(\gamma)|^2 \le \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2 \Rightarrow |\Psi(\gamma)|^2 < |T(\gamma)|^2.$$

This contradicts the requirement that $\Psi(\gamma) = T(\gamma)$. Therefore, we have verified that there does not exist another function $\Psi(z) \in H^\infty$ such that $\Psi(\gamma) = T(\gamma)$ and $\|\Psi(z)\|_\infty < \|\varphi(z)\|_\infty$.□

Given the form for $\phi(z)$, the form for *f*(*z*) follows. After some tedious algebra, we obtain

$$f(z) = \frac{T(z) - \varphi(z)}{B_\gamma(z)} = \frac{zq(z) - \gamma q(\gamma)}{z - \gamma} + \frac{\gamma^2}{1-\gamma^2} q(\gamma)$$

which is the least squares solution plus a constant. Thus the robust cross-equation restrictions likewise differ from the least squares cross-equation restrictions. After the initial period, the impulse response function for the robust predictor is identical to that of the least squares predictor. In the initial

period, the least squares impulse response is $q(\gamma)$, while the robust impulse response is larger: $q(\gamma)/(1-\gamma^2)$.
Because γ is the discount factor, and therefore close to unity, the robust impulse response can be considerably larger than that of the least squares response. Relatedly, the volatility of prices in the robust case will be larger as well. For example, in the first-order autoregressive case studied above,

$$p_t = f(L)\varepsilon_t = \frac{1}{1-\rho\gamma}d_t + \frac{\gamma^2}{(1-\gamma^2)(1-\rho\gamma)}\varepsilon_t \tag{19}$$

from which the variance can be calculated as

$$\sigma^2(p_t) = \left(\frac{1}{1-\rho\gamma}\right)^2\sigma^2(d_t) + \frac{2\gamma^2-\gamma^4}{(1-\rho\gamma)^2(1-\gamma^2)^2}.$$

When the discount factor is large and dividends are highly persistent, the variance of the robust present value prediction can be considerably larger than that of the least squares prediction (the first term on the right alone).

Finally, recall that the least-squares present-value predictor behaved in such a way as to minimize the *variance* of the error $p_t - p_t^*$. Here, robust prediction results in an error with Wold representation

$$p_t - p_t^* = \gamma\left\{\frac{Lq(L)-\gamma q(\gamma)}{L-\gamma} + \frac{\gamma^2}{1-\gamma^2}q(\gamma) - \frac{q(L)}{1-\gamma L^{-1}}\right\}\varepsilon_t = -\frac{\gamma q(\gamma)}{1-\gamma^2}\left\{\frac{1-\gamma L}{L-\gamma}\right\}\varepsilon_t.$$

The term in braces has the form of a Blaschke factor. Applying such factors in the lag operator to a serially uncorrelated process like ε_t leaves a serially uncorrelated result; thus the robust present value predictor has behaved in such a way that the resulting errors are white noise. Of course this comes at a cost: to make the error serially uncorrelated, the robust predictor must tolerate an error variance that is larger than the least squares error variance by a factor of $a^2/(1-\gamma^2)$, which can be substantial when γ is close to unity.

See Also

- forecasting
- robust control

Bibliography

Hansen, L.P. and Sargent, T.J. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2, 7–46.

Hansen, L.P. and Sargent, T.J. 2007. *Robustness*. Princeton: Princeton University Press.

Kasa, K. 2001. A robust Hansen–Sargent prediction formula. *Economic Letters* 71, 43–8.

Nehari, Z. 1957. On bounded bilinear forms. *Annals of Mathematics* 65(1), 153–62.

Sargent, T.J. 1987. *Macroeconomic Theory*. New York: Academic Press.

Whiteman, C.H. 1983. *Linear Rational Expectations: A User's Guide*. Minneapolis: University of Minnesota Press.

Whittle, P. 1983. *Prediction and Regulation by Linear Least-square Methods*, 2nd edn. Minneapolis: University of Minnesota Press.

How to cite this article

Whiteman, Charles H. and Kurt F. Lewis. "prediction formulas." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000344> doi:10.1057/9780230226203.1327

The New Palgrave Dictionary of Economics Online

prediction market design

David C. Croson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys (*a*) the challenges of transitioning the results from prediction-market experiments under laboratory conditions to outside-world conditions devoid of laboratory controls, (*b*) the abilities of current (as of this writing, in 2007) implementations of prediction markets to address these challenges, and (*c*) opportunities for research into future market designs which are robust to these challenges.

Keywords

experimental methods in economics; Goodhart's Law; information aggregation; Iowa Electronic Markets; laboratory markets; moral hazard; planning; prediction market design; prediction markets; subjective probability

Article

prediction markets have made the jump from the laboratory to the field. In these markets, participants bid on Arrow securities, which pay one dollar in one state of the world and zero dollars in others. Since the pioneering work of Plott and Sunder (1982), experiments have generated results that are consistent with the idea that prices in controlled laboratory settings track predicted probabilities formed from the aggregated information of all participants (Hayek, 1945). Emboldened by this general correspondence of theory and the general ability of laboratory experiments to test and validate these theories (see Sunder, 1995; Plott, 2000), prediction markets have escaped the laboratory. Markets offering opportunities to make real-money investments in predicting financial indices, political election results, entertainment awards, world events, and even the minutiae of sporting contests (now viewed with some suspicion as a highbrow form of gambling) have appeared in strength.
The history and general theory of such prediction markets is covered in prediction markets. This article combines (*a*) the challenges of transitioning the results from prediction-market experiments under laboratory conditions to outside-world conditions devoid of laboratory controls, (*b*) the abilities of

current (as of this writing, in 2007) implementations of prediction markets to address these challenges, and (*c*) opportunities for research into future market designs which are robust to these challenges.

While there will necessarily be differences between strictly controlled laboratory studies and the real-world phenomena that they model, four particular cautions should be noted when attempting to extend the predictive abilities of laboratory-generated results to real-world prediction markets. The incidence of any of these four conditions will frustrate our ability to 'read' participants' collective estimation of probabilities from the equilibrium market prices for their associated securities: (*a*) extended duration of capital commitments; (*b*) differing levels of capital commitment across participants; (*c*) strategic objectives other than trading profits; and (*d*) influence by market participants over events on which the contracts are conditioned.

First, laboratory markets generally clear in short periods of time, so that participants face little or no opportunity costs from not investing their capital elsewhere for the duration of the experiment. In an outside prediction market, the lack of this quick-clearing condition means that prices will not generally sum to unity even when the alternatives are a complete partitioning of the possibility space. Participants' capital is tied up in their investment until its resolution, and the opportunity cost of tying up capital while awaiting resolution may be substantial. The common practice is for the exchange provider to capture the float by collecting deposits in time t dollars and paying in time $t+1$ dollars at a 1:1 ratio, rather than a $1{:}1+r_f$ ratio; this distorts prices away from their associated probabilities. The problem is particularly acute when the time to expiry is relatively long over a high proportion of the contract life. An attempted resolution would be for the market organizer to pay the risk-free rate of return on all such committed capital and to allocate a fixed proportion $(1-v)$ of the total amount collected (and earned) pro rata to the winners, with the result that the sum of the prices should converge towards $(1-v)$; probabilities can be renormalized accordingly. In a first step towards addressing this problem, InTrade, a trade exchange market for social, political, and financial events, offered credit interest (at three per cent per annum) on all committed balances beyond a certain threshold account size, thereby reducing the time-value handicap faced by early investors in long-dated contracts.

Second, experimental market participants are allocated fixed amounts of capital, with any variation deliberate on the experimenters' part. When the capital is not uniform across participants, the resulting market prices may diverge from an unbiased estimate of participants' subjective probabilities, a situation hotly debated in the prediction-markets literature (Wolfers and Zitzewitz, 2004; Manski, 2006). The objection to interpreting prices as probabilities arises because the market price is both an input to decisions outside the market and the result of equilibration among these traders. Risk-neutral investors with subjective probabilities above (below) the current market price would wish to buy (sell) the contract, and thus the equilibrium price must balance the total capital of the players on the two sides of the current price. The price will reflect the subjective probability of the investor of the marginal dollar rather than the average subjective probabilities of the potentially large number of participants. With equal capital endowments, the side that prefers the low-probability (inexpensive) side of the contract demands more contract quantity, driving the price down; with unequal capital endowments, this problem can be either masked or exacerbated. The fact that 'heavy hitters' with bigger budgets (or optimistic beliefs about long-shot events) get more 'votes' in this system obscures the information gathered from the other participants. At best, market prices in these situations can be interpreted as the dollar-weighted averages of participants' beliefs, rather than the simple median.

Consider, then, the task of extracting information from such a market. A prospective decision-maker

does not observe the budget weights and will thus be unable to correct for them in the weighted average of collective opinion. A simple solution would make public the holdings of investors with concentrated positions of more than five per cent of outstanding contracts (similar to the 13-D filings required in US securities markets to indicate heavy individual or institutional ownership), or to report measures of concentration in contract ownership or short position. The Iowa Electronic Markets (IEM) has, from its inception, imposed strict capital-inflow restrictions through limiting account funding to 500 dollars (Berg et al., 2006) allowing the IEM to avoid this problem. Substantially larger sums (often in the thousands, if not tens of thousands, of dollars) can be instantly committed at other markets to back a financial investor's probability estimates (for example, those based on the Dow Jones Industrial Average at InTrade).

The third issue is the possibility of strategic manipulation of market prices to achieve an outside goal not shared by all market participants. In a laboratory, the incentives (monetary and otherwise) may be substantial or tiny, but they are by design completely separable from any participant-specific objectives in the outside world. In the absence of such separability, agents with objectives other than capital gain in the prediction market may participate strategically, conveying distorted information to those who rely on unbiased market prices. In a political campaign, for example, it may be worth substantially more to candidates to generate the impression of public support for their preferred campaign than to make a profit on their investments in the information market. Paradoxically, the more credence is given by the general public to the prices-as-probabilities predictions, the higher the incentives for strategic investors to distort price by devoting relatively modest amounts of their private capital to moving the market. (This is an application of Goodhart's Law, wherein indirect measures targeted as policy goals lose their predictive ability.) In extreme cases, this strategic investment may affect voters’ decisions on participation and on candidate selection, thus becoming a self-fulfilling prophecy. This could not only achieve election goals, but also generate capital gains from the information-distorting investment. Therefore, the cost of such a manipulation campaign can be zero or even negative, increasing its attractiveness but sapping the market's predictive power. The design of a prediction market to aggregate opinions and subjective probabilities without encouraging such strategic behaviour remains an open problem.

The fourth issue is the possibility of hidden control of seemingly random events. Under experimental conditions, experimenters control many basic aspects of the study, including randomization of events that are supposed to be random. Since real-world prediction markets generally lack such controls, we must thus be especially cautious in interpreting prices-as-probabilities when certain individuals can profit by exploiting their disproportionate ability to influence the occurrence of the event that is being predicted. As Croson and Kunreuther (2000) note in the analogous situation in the insurance market for natural catastrophic disasters against those caused by terrorism or war, moral hazard can destroy the risk-hedging functions of these markets. The social desirability of such a prediction market subject to moral hazard depends crucially on whether the efficiency value of early warning (caused by the propagation, through price changes, of the inside information) outweighs the equity or efficiency costs resulting from perverse incentives.

Several Fortune 500 companies (for example, HP and Google) have recently implemented prediction markets within the firm. These markets are designed to aggregate information among many employees and thereby produce reasonably accurate estimates that would otherwise be difficult (or impossible) for any single decision-maker to form. Such attempts effectively illustrate Hayek's famous argument (1945)

that central planning cannot replicate the effects of distributed information; one can hardly fault these firms' desire for an unbiased, distributed information-gathering mechanism that communicates a clear message to decision-makers. Corporate motivations for these markets seem primarily aimed at the noble goals of selecting among several alternative investments, informing investment decisions in complementary activities, or attempting to predict future competitive opportunities; they conspicuously and simultaneously risk deviating from controlled experiments, however, in *all four* of the dimensions offered warningly above. Until such markets can be designed to resolve quickly, enforce equal participation among organization members at different ranks or economic stations, disentangle 'in-market' gains from 'out-of-market' gains, and disallow participation by employees who can influence the outcome of the events on which contracts are conditioned, the equilibrium prices shown by these markets will be suspect as a measure of participants' subjective probabilities. Accordingly, unless we develop tools to separate participants' choices from their jobs (effectively creating a 'virtual laboratory' inside the firm) or to correct for the biases induced by these deviations (extracting accurate and useful decision-support information from an unavoidably distorted market), the aggregate value of corporate use of these admittedly promising tools in non-laboratory conditions will be limited, and corporate successes using this powerful technique will be determined as much by chance as by economic science. In strictly controlled laboratory studies, these four divisive effects can be minimized: such studies are completed over short periods of time (making the discounting problem minuscule); participants can be allocated exogenously fixed amounts of capital; the payoffs from successful experimental investments can be separated from outside gains, and the incidence of random events can be kept unpredictable. As prediction markets gain wider acceptance, more active participants, and public credence in the world outside the laboratory, however, the ability to interpret prices as the participants' subjective probabilities of Arrow events becomes increasingly tenuous. To profit from prediction markets outside the laboratory, corporations and investors must combine their knowledge of the established economics of such markets with skills in evaluating investor psychology, probability models more accurate than those of rival participants, and methods of extracting information from noisy and potentially biased signals.

See Also

- experimental economics
- experimental methods in economics
- information aggregation and prices
- moral hazard
- prediction markets

Bibliography

Arrow, K.J. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31, 91–6.

Berg, J., Forsythe, R., Nelson, F. and Rietz, T. 2006. Results from a dozen years of election futures markets research. In *Handbook of Experimental Economic Results*, ed. C. Plott and V. Smith.

Amsterdam: North-Holland.

Croson, D. and Kunreuther, H. 2000. Customizing indemnity contracts and indexed cat bonds for natural hazard risks. *Journal of Risk Finance* 1, 24–41.

Hayek, F. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

InTrade. Online. Available at http://www.intrade.com, accessed 22 June 2007.

Iowa Electronic Markets. Online. Available at http://www.biz.uiowa.edu/iem/, accessed 23 June 2007.

Manski, C.F. 2006. Interpreting the predictions of prediction markets. *Economic Letters* 91, 425–29.

Plott, C. 2000. Markets as information gathering tools. *Southern Economic Journal* 67, 1–15.

Plott, C. and Smith, V. 2006. *Handbook of Experimental Economic Results*. Amsterdam: North-Holland.

Plott, C. and Sunder, S. 1982. Efficiency of experimental security markets with insider information: an application of rational expectations models. *Journal of Political Economy* 90, 663–98.

Roth, A. and Kagel, J. 1995. *Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Sunder, S. 1995. Experimental asset markets: a survey. In *Handbook of Experimental Economics*, ed. A. Roth and J. Kagel. Princeton, NJ: Princeton University Press.

Wolfers, J. and Zitzewitz, E. 2004. Prediction markets. *Journal of Economic Perspectives* 18(2), 107–26.

How to cite this article

Croson, David C. "prediction market design." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000370> doi:10.1057/9780230226203.1329

The New Palgrave Dictionary of Economics Online

prediction markets

Justin Wolfers and Eric Zitzewitz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Prediction markets, sometimes referred to as ‘information markets’, ‘idea futures’ or ‘event futures’, are markets where participants trade contracts whose payoffs are tied to a future event, thereby yielding prices that can be interpreted as market-aggregated forecasts. This article summarizes the recent literature on prediction markets, highlighting both theoretical contributions that emphasize the possibility that these markets efficiently aggregate dispersed information, and the lessons from empirical applications which show that market-generated forecasts typically outperform most moderately sophisticated benchmarks. Along the way, we highlight areas ripe for future research.

Keywords

constant absolute risk aversion (CARA); decision markets; efficient market hypothesis; favourite-longshot bias; forecasting; Gallup Poll; information aggregation; Iowa Electronic Market; prediction; prediction markets; probability; spread betting

Article

Prediction markets, sometimes referred to as ‘information markets,’ ‘idea futures’ or ‘event futures’, are markets where participants trade contracts whose payoffs are tied to a future event, thereby yielding prices that can be interpreted as market-aggregated forecasts. For instance, in the Iowa Electronic Market traders buy and sell contracts that pay one dollar if a given candidate wins the election. If a prediction market is efficient, then the prices of these contracts perfectly aggregate dispersed information about the probability of each candidate being elected. Markets designed specifically around this information aggregation and revelation motive are our focus in this article.

Types of prediction market

The most famous prediction markets are the election forecasting markets run by the University of Iowa (Berg et al., 2006). Election forecasting provides a useful way to introduce a variety of different contract types, and Table 1, adapted from Wolfers and Zitzewitz (2004a), shows how different contracts can be designed to reveal various types of forecasts.

Contract types: estimating uncertain quantities or probabilities

Contract	Details	Example	Reveals market expectation of…	More general application
Binary option	Contract costs \$$p$ Pays \$1 if and only if event x occurs.	Event x: George Bush wins the popular vote.	Probability that event y occurs, $p(x)$.	Defining many events, $x_1, x_2, \ldots, x_n$ reveals probability distribution $F(x)$.
Index futures	Contract pays \$$x$.	Contract pays \$1 for every percentage point of the popular vote won by George Bush.	Mean value of outcome x: E[x].	Contract pays some function of x: \$$g(x)$. Reveals specific moments, E[$g(x)$].

Spread betting	Contract costs \$1 Pays \$2 if $x>x^*$ Pays \$0 otherwise. Bid according to the value of x^*.	Contract pays even money if Bush wins more than x^* % of the popular vote.	Median value of outcome, x.	\$1 contract pays \$($1/q$) if $x>x^*$. Reveals specific quantile, $F_{1-q}(x)$.

The three main types of contract link payoffs to the occurrence of a specific event (the incumbent wins the election), to a continuous variable (the vote share of the incumbent), or to a combination of the two, such as in spread betting (the vote share of the incumbent exceeds x per cent). In each case, the relevant contract will reveal the market's expectation of a specific parameter: a probability, a mean or a median, respectively. More complex contract designs can also be used to elicit alternative parameters. For instance, a family of winner-take-all contracts – each linked to different states of nature – can reveal the full probability distribution.

Prediction markets have been used to forecast elections, movie revenues, corporate sales, project completion, economic indicators and Saddam Hussein's demise. New corporate applications have emerged as firms have looked to markets to predict research and development outcomes, the success of new products, and regulatory outcomes. In the US public sector, the Pentagon attempted to use markets designed to predict geopolitical risks, although negative publicity stopped the project (Hanson, 2006). An intriguing attempt to apply prediction markets to forecasting influenza outbreaks is detailed in Nelson, Neumann and Polgreen (2006). Rhode and Strumpf (2004) have detailed the existence of large-scale election betting as far back as the election of President Grant in 1868.

Prediction market contracts have been traded in a variety of market designs, including continuous double auctions (both with and without market-makers), pari-mutuel pools, and bookmaker-mediated betting markets, or implemented as market-scoring rules.

Prediction markets in theory: information aggregation

The claim that prediction markets can efficiently aggregate information is based on the efficient market hypothesis. In certain cases, existing theoretical results regarding efficient capital markets can be applied directly. Grossman (1976) documents a set of sufficient conditions for the equilibrium price of index futures to summarize private information perfectly: in a market where traders with constant absolute risk aversion (CARA) utility functions each receive independent draws from a normal distribution about the true value of the asset, the market price fully summarizes their information.

Manski (2004) notes that much of the analysis of the price of binary options simply assumes that these revealed a market-based probability estimate, but that appropriate theoretical results are lacking. He illustrates the importance of this issue by way of an example where prediction market prices fail to aggregate information appropriately. In his model all traders are willing to risk exactly \$100. Thus, if a contract paying \$1 if an event occurs is selling for \$0.667, then buyers each purchase 150 contracts, while sellers can afford to sell 300 contracts (at a price of \$0.333). This can be an equilibrium only if there are twice as many buyers as sellers, implying that the market price must fall at the 33rd percentile of the belief distribution, rather than the mean. The same logic suggests that a prediction market price of π implies that $1-\pi$ per cent of the population believes that the event has less than a π per cent chance of occurring. Clearly, the driving force in this example is the assumption that all traders are willing to risk a fixed amount.

Wolfers and Zitzewitz (2005a) provide sufficient conditions under which prediction market prices coincide with average beliefs among traders (and hence aggregate all information in the Grossman set-up). They consider individuals with log utility and initial wealth, y, who must choose how many prediction market securities, x, to purchase at a price, π, given that they believe that the probability of winning their bet is q:

$$\underset{\{x\}}{Max}\, EU_j = q_j Log[y + x_j(1-\pi)] + (1-q_j) Log[y - x_j \pi]$$

$$yielding: \quad x_j^* = y\frac{q_j - \pi}{\pi(1-\pi)}$$

The prediction market is in equilibrium when supply equals demand:

$$\int_{-\infty}^{\pi} y\frac{q-\pi}{\pi(1-\pi)} f(q)\,dq = \int_{\pi}^{\infty} y\frac{\pi-q}{\pi(1-\pi)} f(q)\,dq$$

If beliefs (q) and wealth (y) are independent, then this implies:

$$\pi = \int_{-\infty}^{\infty} qf(q)\,dq = \bar{q}.$$

Thus, under log utility, the prediction market price equals the mean belief among traders. If wealth is correlated with beliefs, then the prediction market price is equal to a wealth-weighted average belief. This finding is general in the sense that no assumptions are required about the distribution of beliefs, but it is also quite specific in that it holds only under log utility. Experimenting with a range of alternative utility functions and distributions of beliefs typically yields prediction market prices that diverge from the mean of beliefs by only a small amount.

Both the Manski and the Wolfers–Zitzewitz models are silent as to the sources of the different beliefs across traders, which allows them to sidestep the theoretical difficulty posed by Milgrom and Stokey (1982), namely, that under common beliefs no trade will occur. The logic of the 'no trade theorem' is simply that traders should always be wary that anyone seeking to trade with them possesses an information advantage, and hence should moderate their beliefs accordingly. Why there should be any trade in prediction markets remains an important open theoretical question. Wolfers and Zitzewitz (2006) provide a simple adaptation of the Kyle (1985) model in which trade is driven by uninformed outsiders with either hedging- or entertainment-driven demand for the prediction security, or by manipulators attempting to influence market prices.

Another important role of prediction markets is that potential trading profits provide an incentive for *information discovery*. Grossman and Stiglitz (1976) consider the case where information is expensive to garner. They point to the impossibility of prices being fully efficient: if prices fully reflect information, then there is no incentive for any trader to gather that information. Instead, they construct a model in which prices never fully reflect all of the information possessed by informed traders; in equilibrium the inefficiency in pricing is just sufficient to induce a proportion of traders to become informed.

Another key advantage of prediction markets over alternative approaches to information aggregation is that they provide incentives for *truthful revelation* of beliefs. If prediction markets are to be used as inputs into future decisions, this may provide a countervailing incentive to trade dishonestly to manipulate prices. While such manipulation would typically lead the manipulator to lose money, Hanson and Oprea (2005) have shown that these losses increase the rewards for informed trading, which may ultimately increase the accuracy of prediction market prices.

Prediction markets in practice

While we are still accumulating evidence on the behaviour of prediction markets in different contexts, already a few generalizations can be drawn from existing, albeit piecemeal, evidence.

First, market prices tend to respond rapidly to new information. Figure 1 draws an interesting example from Snowberg, Wolfers and Zitzewitz (2006): movements in the price of the Tradesports contract on the re-election of US President Bush, around election day, 2004. Early exit polls suggesting victory by John Kerry, the Democrat candidate, were leaked at around 3 p.m., and prices started to move immediately. Indeed, the figure shows that they moved in lockstep with prices on the much larger equity markets. As the count proceeded, it became clear that these early polling numbers were wrong, and the market reversed course sharply. This is only a single anecdote but is representative of the rapid incorporation of new information by prediction markets observed in many domains.

Figure 1

Bush's re-election prospects and the stockmarket. *Source*: Snowberg, Wolfers and Zitzewitz (2006).

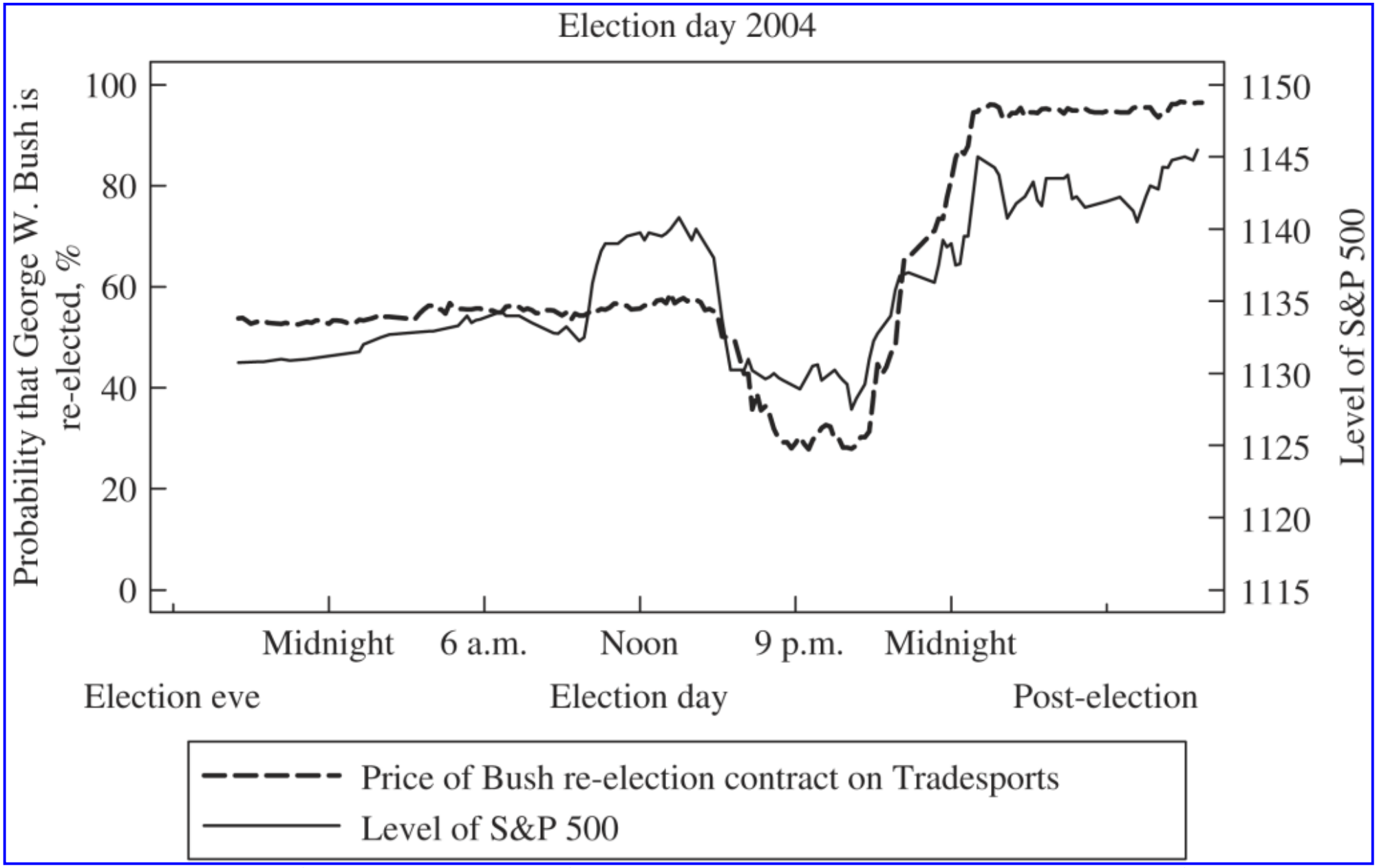

Second, in most cases, the time series of prices in these markets appears to follow a random walk, and simple betting strategies based on publicly available information appear to yield no profit opportunities. That is, these markets appear to meet the standard definition of weak-form efficiency.

Third, the law of one price appears to (roughly) hold, and the few arbitrage opportunities that arise in these markets are fleeting and involve only small potential profits.

Fourth, attempts to manipulate these markets typically fail. Camerer (1998) attempted to manipulate pari-mutuel betting on horse races by canceling $500 or $1,000 bets at the last moment. Rhode and Strumpf (2006) report attempts by specific political campaigns to manipulate the election betting odds on their candidates in the large-scale betting markets operating in the early 20th century. They also analyse an attempt to manipulate the price of a Kerry victory on Tradesports in 2004, as well as their own attempts to manipulate prices on the Iowa Electronic Markets in 2000. Hanson, Oprea and Porter (2006) created experimental prediction markets in which several traders were given an incentive to raise the price. None of these attempts at manipulation had a discernible effect on prices, except during a short transition phase.

Finally, prediction markets usually provide quite accurate forecasts and have typically outperformed alternative prediction tools.

Figure 2 shows evidence collected by Gürkaynak and Wolfers (2005) on the relative performance of a prediction market (the 'Economic Derivatives' market established by Goldman Sachs and Deutsche Bank) and a survey of economists in predicting economic outcomes. They show that the market-based forecast encompasses the information in the survey-based forecasts. Moreover, the behavioural anomalies that have been noted in survey-based forecasts are not evident in the market-based forecasts.

Figure 2

Forecasting economic outcomes. Graphs by economic data series. *Source*: Gürkaynak and Wolfers (2005).

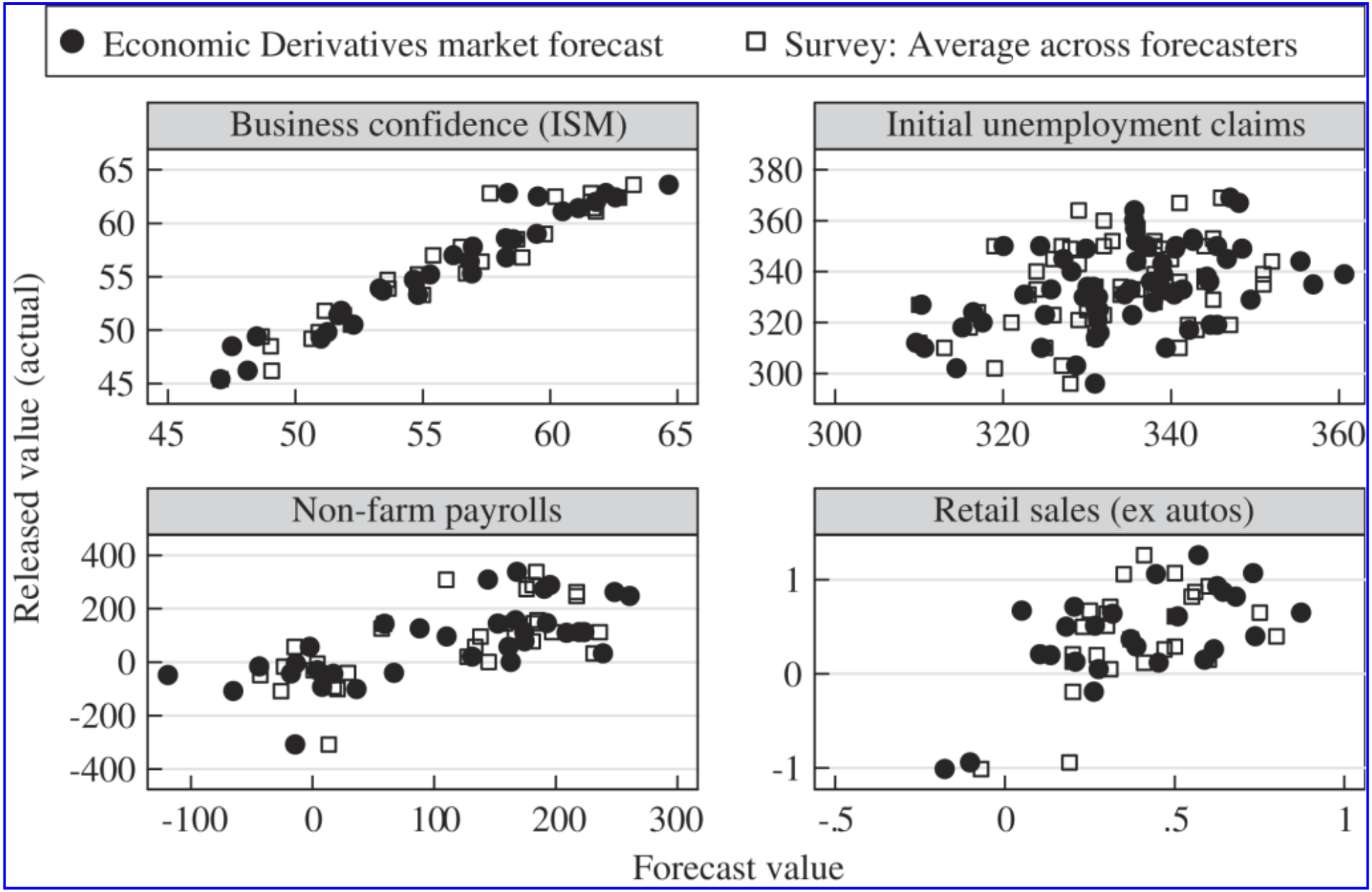

Figure 3 compares the forecasting performance of the Iowa Electronic Markets and the Gallup Poll in predicting the outcomes of presidential elections in the United States. Over the

13 major candidacies from 1988 to 2004, the average absolute error of the market-based forecasts was 1.6 percentage points, while the corresponding number for the Gallup Poll was 1.9 percentage points. As Berg, Nelson and Rietz (2003) discuss, the forecasting advantage of markets over the polls is probably even larger over long horizons, as polling numbers tend to be excessively volatile through the electoral cycle. The initial success of these forecasting methods in the United States has led to similar analysis of election forecasting markets in Austria, Australia, Canada, Germany, the Netherlands and Taiwan.

Figure 3

Forecasting presidential elections. *Note*: Market forecast is closing price on election eve; Gallup forecast is final pre-election projection.

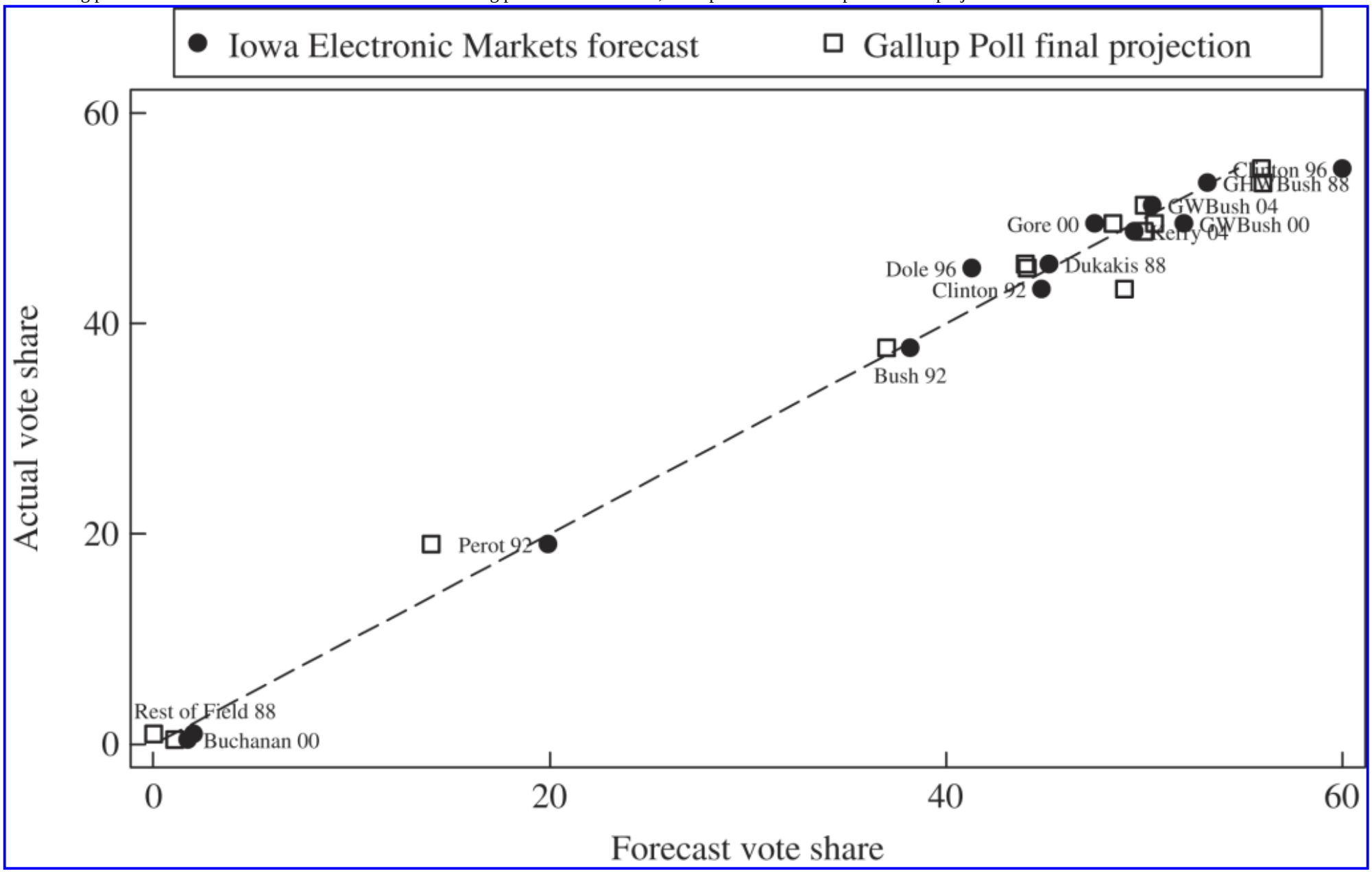

Tests of prediction markets and expert opinions have also been conducted in a range of other domains. The Hollywood Stock Exchange has generated forecasts of box-office success and of Oscar winners that have been more accurate than expert opinions (Pennock et al., 2001). Both real and play-money markets have generated more accurate forecasts of the

likely winners of NFL football games than all but a handful among 2,000 self-professed experts (Servan-Schreiber et al., 2004). In the corporate context, the market established by Chen and Plott (2002) within Hewlett-Packard yielded more accurate sales forecasts than the firm's internal experts. Similarly, Ortner (1998) reports that an internal market correctly predicted that the firm would definitely fail to deliver on a software project on time, even when traditional planning tools suggested that the deadline could be met.

Despite this impressive evidence, there remain a number of documented pathologies in prediction markets. Figure 4 shows evidence from Snowberg and Wolfers (2005) of the 'favourite-longshot bias', which describes a tendency to overprice low-probability events. A similar tendency has been documented in a range of other market contexts, suggesting that some caution is in order in interpreting the prices of low probability events.

Figure 4

Favourite-longshot bias: rate of return at different odds. *Source*: Trackmaster, Inc. Sample is all horse races in the United States, 1992–2002, n=5,067,832 starts in 611,807 races.

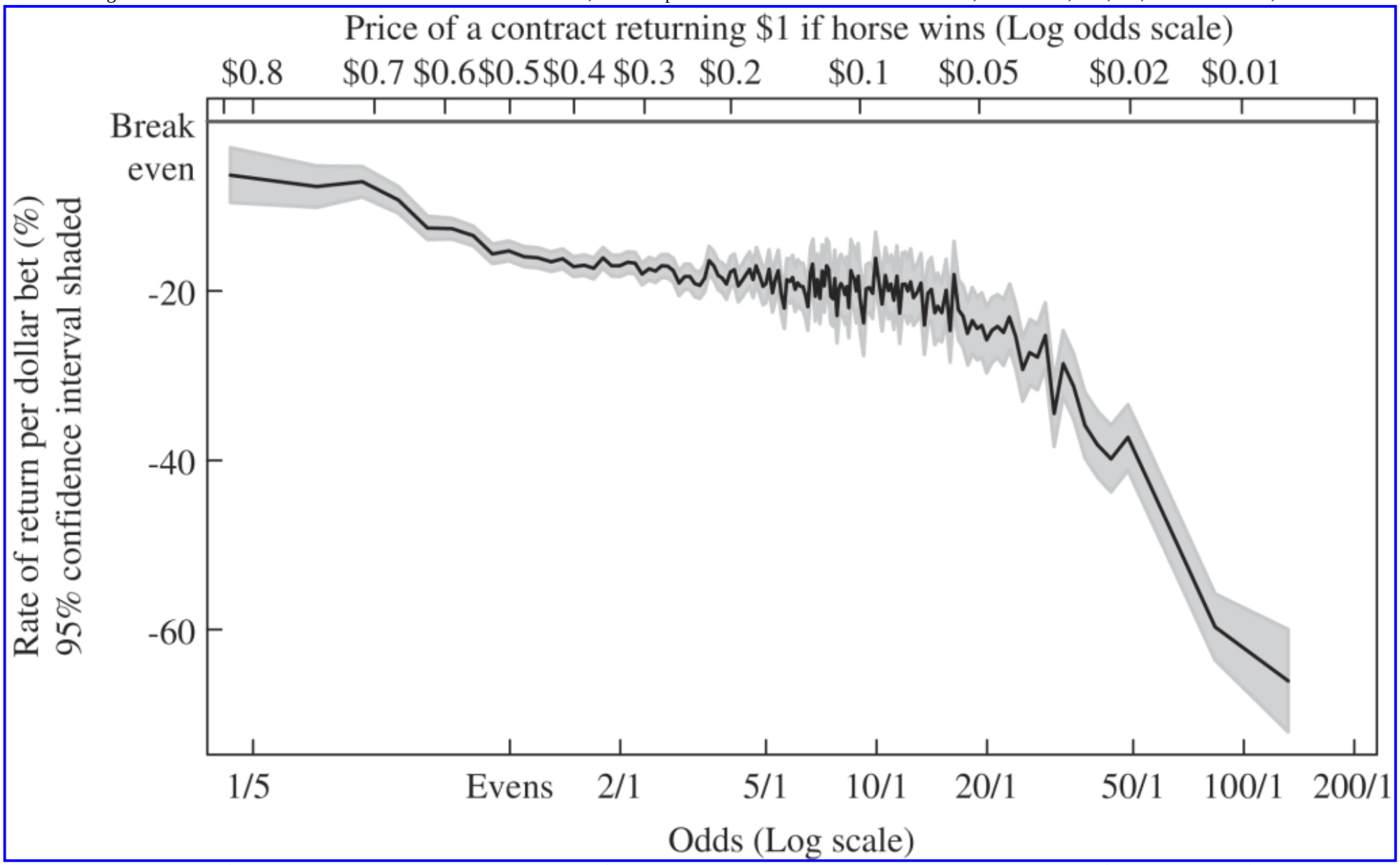

Laboratory experiments also find that, while prediction markets can be successful in some contexts (Plott and Sunder, 1982), in others they may fail to aggregate information (Plott

and Sunder, 1988). Sunder (1995) and Plott (2000) provide excellent reviews of experimental prediction markets, including experiments showing market designs that lead to the appearance of bubbles, false equilibria or excess volatility.

Economic analysis of prediction market prices

Prediction markets are a useful way to elicit predictions, but how might they be used? The most direct form of inference involves simply using these predictions directly. For instance, forecasts of election outcomes may be of intrinsic interest.

Some analyses have tried to link the time series of expectations elicited in prediction markets with time series of other variables, so as to isolate a causal influence. For instance, Roberts (1990) analyses changes in the betting odds posted by Ladbrokes on US President Ronald Reagan's re-election in 1984 and the returns to holding stocks in defence firms, inferring that Reagan led to more robust defence spending. Likewise, Herron et al. (1999) and Knight (2006) analyse the correlation of industry stock indices and individual stocks with movements in the 1992 and 2000 Iowa Electronic Markets US presidential election markets. Snowberg, Wolfers and Zitzewitz (2006) conduct a similar analysis for the aggregate equity and bond markets at an intraday frequency, using the data shown in Figure 1, to infer partisan impacts of the 2004 election. Slemrod and Greimel (1999) examine the effect on municipal bond prices of changes in the probability of a 1996 Republican nomination for Steve Forbes, whose ‘flat tax’ would have eliminated the tax exemption for municipal bond interest.

To move beyond *ex post* studies of elections, Wolfers and Zitzewitz (2005b) report on an *ex ante* analysis of the co-movement of oil and equity prices with a contract tracking the probability of a US attack on Iraq in 2002–3 (Figure 5). The results suggest that a substantial war premium was built into oil prices (and a discount built into equities).

Figure 5

Risk of war in Iraq. Prediction markets, export opinion and oil markets. *Source*: Trade-by-trade Saddam Security data provided by Tradesports.com; Saddameter from Will Saletan's daily column in Slate.com.

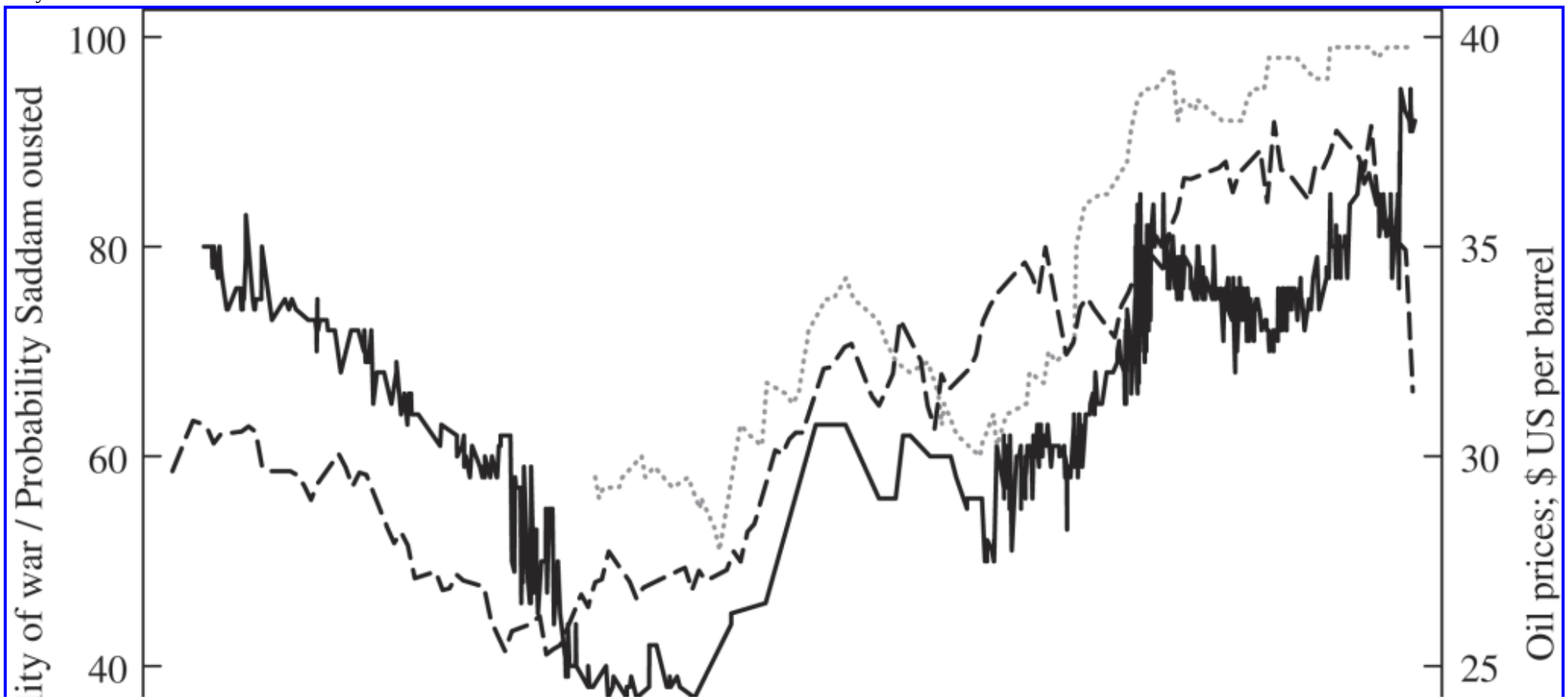

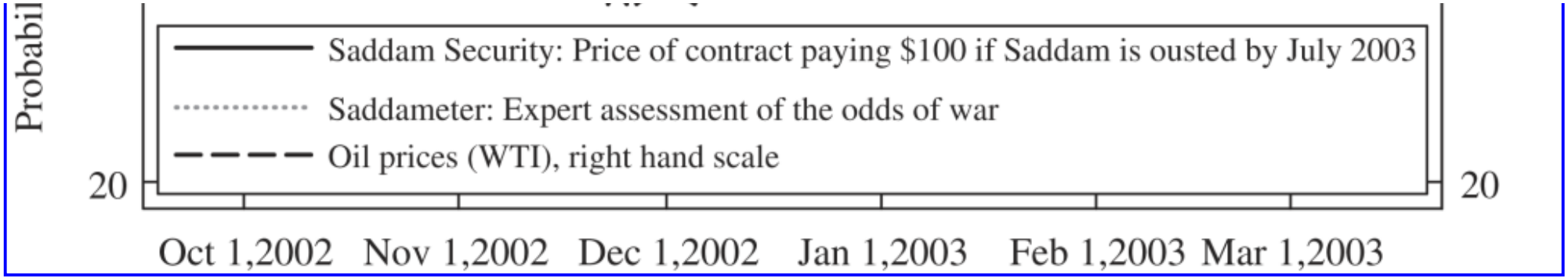

The contracts we have described thus far have depended on only one outcome. The same principles can be applied to contracts tied to the outcomes of more than one event. These contingent contracts potentially provide insight into the correlation between events. For instance, Wolfers and Zitzewitz (2004b) ran experimental markets on the online betting exchange Tradesports.com in the run-up to the 2004 US presidential election. In one example, they ran markets linked to whether George W. Bush would be re-elected, whether Al-Qaeda leader Osama bin Laden would be captured prior to the election, and whether *both* events would occur. These markets suggested a 91 per cent chance of Bush being re-elected *if* Osama had been found, but a 67 per cent unconditional probability. Berg and Reitz (2003) report on contracts whose payoff was linked to 1996 Democratic vote shares conditional on different potential Republican nominees; on the basis of these prices they argue that alternative nominees, such as Colin Powell, would have outperformed Bob Dole, the actual nominee.

The potential to apply these markets to determine the consequences of a range of contingencies has led Hanson (1999) to term these ‘decision markets’. Indeed, Hanson (2003a) has suggested that such markets could be used to remove technocratic policy implementation issues from the bureaucracy, a suggestion endorsed in Hahn and Tetlock (2006). Moreover, while the previous example involves only one contingency, Hanson (2003b) suggests that market scoring rules can allow traders to simultaneously predict many combinations of outcomes. The basic intuition of his proposal is that, rather than betting on each contingency, traders bet that the sum of their errors over all predictions will be lower.

However while contingent markets can be used to estimate the joint probability of choice A and outcome B, care must be taken before inferring that choice A should be made because it will maximize the probability of outcome B. That is, while these markets can highlight the correlation between events, the difficulty of inferring causation remains.

Conclusion

The healthy bibliography below attests to the fact that interest in prediction markets has boomed in recent years. Many questions remain. Theoretical research holds the promise of better understanding the institutional design features that yield optimal information aggregation and efficient pricing. The practical agenda includes developing new ideas about how and when prediction markets can aid decision-making by business and government.

See Also

- capital asset pricing model
- cheap talk
- contingent commodities
- efficient markets hypothesis
- forecasting
- futures markets, hedging and speculation
- hedging
- information aggregation and prices
- noise traders
- terrorism, economics of

Thanks to Colin Camerer, Robin Hanson, Ryan Oprea, Charles Plott, Paul Rhode, and Erik Snowberg for very helpful comments.

Bibliography

Berg, J., Forsythe, R., Nelson, F. and Rietz, T. 2006. Results from a dozen years of election futures markets research. In *Handbook of Experimental Economic Results*, ed. C. Plott and V. Smith. Amsterdam: North-Holland.

Berg, J., Nelson, F. and Rietz, T. 2003. Accuracy and forecast standard error in prediction markets. Mimeo. University of Iowa.

Berg, J. and Rietz, T. 2003. Prediction markets as decision support systems. *Information System Frontiers* 5, 79–93.

Berg, J. and Rietz, T. 2006. The Iowa Electronic Market: lessons learned and answers yearned. In *Information Markets: A New Way of Making Decisions in the Public and Private Sectors*, ed. R. Hahn and P. Tetlock. Washington, DC: AEI–Brookings Joint Center.

Camerer, C. 1998. Can asset markets be manipulated? A field experiment with racetrack betting. *Journal of Political Economy* 106, 457–82.

Chen, K.-Y. and Plott, C. 2002. Information aggregation mechanisms: concept, design and implementation for a sales forecasting problem. Working Paper No. 1131. Division of Humanities and Social Sciences, California Institute of Technology.

Grossman, S. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 31, 573–85.

Grossman, S. and Stiglitz, J. 1976. Information and competitive price systems. *American Economic Review* 66, 246–53.

Gürkaynak, R. and Wolfers, J. 2005. Macroeconomic derivatives: an initial analysis of market-based macro forecasts, uncertainty, and risk. In *NBER International Seminar on Macroeconomics*, ed. C. Pissarides and J. Frankel. Cambridge, MA: NBER.

Hahn, R. and Tetlock, P. 2006. Using information markets to improve decision making. *Harvard Journal of Law and Public Policy* 29, 213–89.

Hanson, R. 1999. Decision markets. *IEEE Intelligent Systems* 14(3), 16–19.

Hanson, R. 2003a. Shall we vote on values, but bet on beliefs? Mimeo, George Mason University.

Hanson, R. 2003b. Combinatorial information market design. *Information Systems Frontiers* 5(1), 105–19.

Hanson, R. 2006. Designing Real Terrorism Futures. *Public Choice* 128, 257–74.

Hanson, R. and Oprea, R. 2005. Manipulators increase information market accuracy. Mimeo, George Mason University.

Hanson, R., Oprea, R. and Porter, D. 2006. Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior and Organization* 60, 449–59.

Herron, M., Cram, D., Lavin, J. and Silver, J. 1999. Measurement of political effects in the United States economy: a study of the 1992 presidential election. *Economics and Politics* 11, 51–81.

Knight, B. 2006. Are policy platforms capitalized into equity prices? Evidence from the Bush/Gore 2000 presidential election. *Journal of Public Economics* 90, 751–73.

Kyle, A. 1985. Continuous auctions and insider trading. *Econometrica* 53, 1315–36.

Manski, C. 2004. *Interpreting the predictions of prediction markets. Working Paper No. 10359*. Cambridge, MA: NBER.

Milgrom, P. and Stokey, N. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26, 17–27.

Nelson, F., Neumann, G. and Polgreen, P. 2006. Operating with doctors: results from the 2004 and 2005 influenza markets. Mimeo, University of Iowa.

Ortner, G. 1998. Forecasting markets – an industrial application. Mimeo, Technical University of Vienna.

Pennock, D., Lawrence, S., Nielsen, F. and Giles, C. 2001. Extracting collective probabilistic forecasts from web games. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press.

Plott, C. 2000. Markets as information gathering tools. *Southern Economic Journal* 67, 1–15.

Plott, C. and Sunder, S. 1982. Efficiency of experimental security markets with insider information: an application of rational expectations models. *Journal of Political Economy* 90, 663–98.

Plott, C. and Sunder, S. 1988. Rational expectations and the aggregation of diverse information in a laboratory security markets. *Econometrica* 56, 1085–118.

Rhode, P. and Strumpf, K. 2004. Historical presidential betting markets. *Journal of Economic Perspectives* 18(2), 127–42.

Rhode, P. and Strumpf, K. 2006. Manipulating political stock markets: a field experiment and a century of observational data. Mimeo, University of North Carolina.

Roberts, B. 1990. Political institutions, policy expectations, and the 1980 election: a financial market perspective. *American Journal of Political Science* 34, 289–310.

Servan-Schreiber, E., Wolfers, J., Pennock, D. and Galebach, B. 2004. Prediction markets: does money matter? *Electronic Markets* 14, 243–51.

Slemrod, J. and Greimel, T. 1999. Did Steve Forbes scare the municipal bond market? *Journal of Public Economics* 74, 81–96.

Snowberg, E. and Wolfers, J. 2005. Explaining the favorite-longshot bias: is it risk-love, or misperceptions? Mimeo, University of Pennsylvania.

Snowberg, E., Wolfers, J. and Zitzewitz, E. 2006. Partisan impacts on the stockmarket: evidence from prediction markets and close elections. Mimeo, University of Pennsylvania.

Sunder, S. 1995. Experimental asset markets: a survey. In *Handbook of Experimental Economics*, ed. A. Roth and J. Kagel. Princeton, NJ: Princeton University Press.

Wolfers, J. and Zitzewitz, E. 2004a. Prediction markets. *Journal of Economic Perspectives* 18(2), 107–26.

Wolfers, J. and Zitzewitz, E. 2004b. Experimental political betting markets and the 2004 election. *The Economists Voice* 1(2). Online. Available at http://www.bepress.com/ev/vol1/iss2, accessed 19 February 2006.

Wolfers, J. and Zitzewitz, E. 2005a. Interpreting prediction market prices as probabilities. Mimeo, University of Pennsylvania.

Wolfers, J. and Zitzewitz, E. 2005b. Using markets to inform policy: the case of the Iraq war. Mimeo, University of Pennsylvania.

Wolfers, J. and Zitzewitz, E. 2006. Five open questions about prediction markets. In *Information Markets: A New Way of Making Decisions in the Public and Private Sectors*, ed. R. Hahn and P. Tetlock. Washington, DC: AEI–Brookings Joint Center.

How to cite this article

Wolfers, Justin and Eric Zitzewitz. "prediction markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000340> doi:10.1057/9780230226203.1328

The New Palgrave Dictionary of Economics Online

preference reversals

Chris Starmer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Preference reversal is a widely observed behavioural tendency for the preference ordering of a pair of alternatives to depend on the process used to elicit it. The phenomenon appears to be both a robust and a systematic departure from conventional preference theory. Competing theoretical explanations variously interpret it as a violation of procedure invariance (the presumption that preferences should be independent of the method of eliciting them); a failure of transitivity; or a consequence of loss-averse (and reference-dependent) preferences. This article discusses these interpretations, the related evidence, and reflects on some of the broader implications of the phenomenon.

Keywords

Allais paradox; decision processes; expected utility hypothesis; expected utility theory; intransitivity; loss aversion; preference reversal; preferences; procedure invariance; regret; Savage's subjective expected hypothesis

Article

Preference reversal (PR) is a widely observed behavioural tendency for the preference ordering of a pair of alternatives to depend, in a predictable way, on the process used to elicit it.
The existence of preference reversal sets an empirical challenge to fundamental assumptions of conventional economic theory: PR is an apparent failure of procedure invariance (that is, the traditional presumption that preferences should be independent of the method of eliciting them). Some see it as a challenge to the very idea that human decisions are governed by preferences.
Much of the empirical PR literature has examined decisions relating to pairs of simple gambles. One of the gambles (typically called the ‘P-bet’) will offer a relatively good chance of winning a modest prize, otherwise nothing (or sometimes a small loss); the other bet (the ‘$-bet’), offers a relatively small chance of winning a larger prize. In classic PR experiments, subjects are required to make straight choices

between such pairs of bets and to provide separate (usually monetary) valuations for each bet. For any individual and gamble pair, conventional economic theory implies that the chosen gamble would also be the more highly valued of the pair. But while many individuals are so consistent, a significant proportion, typically, are not. The existence of some such inconsistency, by itself, is not especially surprising. People might, for instance, make a mistake in one or more task, leading to some level of inconsistency in comparisons of rankings. Interest in PR, however, stems largely from the fact that observed inconsistencies tend to be patterned in a highly predictable way: the typical finding is that considerable numbers of subjects choose the P-bet and value the $-bet more highly (let us call this the standard reversal), while very few commit the opposite reversal ($-bet chosen and P-bet valued more highly). It is this *asymmetric* pattern of inconsistencies between rankings based on choice and valuation that constitutes the intriguing PR phenomenon.

Evidence

PR was first predicted and then observed by psychologists (Lichtenstein and Slovic, 1971; Lindman, 1971). It was later brought to the attention of economists by Grether and Plott (1979) who described its potential significance for economics in the following passage:

> Taken at face value the data are simply inconsistent with preference theory and have broad implications for research priorities within economics. The inconsistency is deeper than mere lack of transitivity or even stochastic transitivity. It suggests that no optimisation principles of any sort lie behind even the simplest of human choices. (Grether and Plott, 1979, p. 623)

Like many economists who have followed in their footsteps, Grether and Plott did not immediately accept this face-value interpretation and, instead, looked for ways of explaining PR while retaining the assumption that individuals do have a unique preference ordering over gambles. A substantial body of research in this spirit has examined whether PR might be an experimental artefact arising from imperfectly designed experiments. Early research of this genre – including Grether and Plott (1979); Reilly (1982) and Pommerehne, Schneider and Zweifel (1982) – investigated issues such as whether PR might be a consequence of subjects failing to understand the tasks confronting them, or of having insufficient motivation to take those tasks seriously. But a large body of evidence now shows that PR is a highly replicable phenomenon, robust to many variations in experimental procedures. Seidl (2002) provides a review.

A more subtle critique of PR experiments and evidence emerged in the late 1980s with the publication of a series of theoretical papers (Holt, 1986; Karni and Safra, 1987; Segal, 1988) arguing that PR might be a spurious artefact of experimental design after all. These papers shared a common strategy, pointing to a potential weakness of two experimental procedures which had been commonly used to incentivize decision tasks in PR experiments: the Becker–DeGroot–Marschak (1964) mechanism and the random lottery incentive system. The thrust of these papers is to show that, if individuals have non-expected utility preferences (violating either the independence axiom of expected utility theory, or the reduction of compound lotteries principle, or both), these standard incentive mechanisms could be biased and

might generate the spurious appearance of PR. On this interpretation, PR would not be evidence against procedure invariance: instead it would be evidence of consistent, but non-expected utility, preferences interacting with specific features of experimental design. This interpretation has, however, been largely discounted in the light of subsequent research (including Tversky, Slovic and Kahneman, 1990 and Cubitt, Munro and Starmer, 2004) which reproduces the PR phenomenon in experiments using incentive mechanisms immune to this critique of earlier studies.

Theory

There remains considerable interest in trying to find a satisfactory explanation of PR. In what follows, we discuss three types of theory that may contribute to that objective: *regret theory*, *reference-dependent theory*, and *constructed preference theory*.

Regret theory (Loomes and Sugden, 1982; 1983) explains PR as a form of intransitivity. In this theory preferences are defined over pairs of acts which map from states of the world to consequences (as in Savage, 1954). Suppose A_i and A_j are two potential acts that result in, respectively, outcomes x_{is} and x_{js}, in state of the world s. If A_i is chosen, the resulting utility in each state is given by a 'modified utility function' $M(x_{is},x_{js})$. Notice that this function allows the consequences of the chosen act to depend upon those that *might have been* experienced under the forgone act A_j. In particular, the utility from having x_{is} may be suppressed by 'regret' when x_{is} is worse than x_{js}. Regret theory assumes that individuals attempt to maximize the expectation of modified utility $\Sigma_s p_s.M(x_{is},x_{js})$ where p_s is the probability of state s. Regret theory reduces to expected utility theory in the special case where $M(x_{is}, x_{js}) = u(x_{is})$ and $u(.)$ is a von Neumann–Morgenstern utility function.

Loomes and Sugden (1982) show that, if preferences in this theory satisfy particular restrictions, then regret theory provides a possible explanation of several well-known violations of expected utility theory including some cases of the famous Allais paradox. The most important of these restrictions is a property (subsequently) called regret aversion and, in a follow-up paper, Loomes and Sugden (1983) show that regret aversion may also explain PR. The argument works roughly as follows. Consider the following three acts labelled \$, P and M with monetary consequences $x > y > m > 0$ defined over three states.

	State 1	State 2	State 3
\$	x	0	0
P	y	y	0
M	m	m	m

The acts labelled \$ and P have the structure of typical \$- and P-bets: they are binary gambles where \$ has the higher prize, and P the higher probability of 'winning'; the third act gives payoff m for sure. Regret theory allows choices over acts with this structure to be non-transitive and, if preferences are regret averse, if a cycle occurs it will be in a specific direction: P chosen over \$; M over P; and \$ over M. Now recall that, in a typical PR experiment, the standard reversal occurred when a subject chose P over \$ but valued \$ more highly than P. So, if we interpret choices from {\$, M} and {P, M} as analogues of valuation tasks asking 'is \$ (or P) worth more or less than M?', then the cycle predicted by

regret theory can be interpreted as a form of PR.

This explanation for PR has been tested via experiments designed to look for the pure choice analogue of PR by confronting subjects with pairwise choices among triples of bets with the structure of $, P and M above. The outcome of this strand of research has produced good and bad news for regret theory. The good news is that the non-transitive choice cycles predicted by it have been observed and replicated (Loomes, Starmer and Sugden, 1991). Since these choice cycles occur in studies that involve no valuation tasks at all, this is evidence for the intransitivity interpretation of PR. The bad news is that subsequent research (Starmer and Sugden, 1998) has cast considerable doubt on regret theory's account of these choice cycles. The current state of play appears to be that regret theory has led to the discovery of a surprising new choice phenomenon, but it turns out not to be the right explanation for it! It remains possible that these intransitive choice cycles are manifestations of regret-type influences at work but that formal models of regret must be refined to properly account for them. Another possibility is that they have nothing to do with ‘regret’ and that their discovery, as a consequence of testing regret theory, was just accidental.

A new account of PR has emerged in the form of reference-dependent subjective expected utility theory (Sugden, 2003). In this model, preferences are again defined over acts. The key structural departure from Savage's (1954) subjective expected utility theory is that consequences in each state are modelled as gains and losses relative to a *reference act* (the status quo). The resulting theory is a formulation of expected utility (that is, a model that is linear in probabilities) that can accommodate loss aversion (that is, losses of a given size being weighted more highly than corresponding magnitude gains). Sugden demonstrates that, when preferences are loss averse, this model predicts standard PR in experiments where values are elicited as selling prices (which they usually are). This prediction depends on the assumption that, in selling tasks, an agent's reference act is the lottery being sold: given this, seemingly reasonable, assumption, $ valuations become particularly ‘inflated’ by consideration of the large $ prize which becomes a (probabilistic) loss if the $-bet is given up for a certain amount of cash. Hence, on this account, PR is the consequence of loss aversion operating through selling tasks. As yet, there have been no direct tests of this explanation, though the evidence of loss aversion operating in other contexts (see Starmer, 2000, for some discussion) perhaps gives it some initial credibility.

Thus far we have discussed various preference-theoretic accounts of PR. The final type of explanation we discuss is the oldest and belongs to a class of theory that has evolved in the psychology literature. From the outset, most psychologists accepted PR as evidence against the very thing that economists have invested their efforts in defending: the presumption that behaviour can be adequately explained in terms of unique underlying preferences. Psychologists have, instead, focused on accounts of PR which attribute it to aspects of human *decision processes*. Viewed from this perspective, there is nothing fundamentally surprising about the fact that rankings delivered via choice and valuation tasks differ; those working within this paradigm will, typically, attempt to read such inconsistencies as clues to the, potentially distinct, mental heuristics invoked in those different tasks.

Numerous theories in this spirit have been proposed as putative accounts of PR, and one of the best known examples is the scale-compatibility hypothesis due to Tversky, Sattath and Slovic (1988). The general hypothesis assumes that the way in which an individual is required to respond to a task (‘the response mode’) can affect the weights that he or she places on particular dimensions of alternatives being evaluated. In application to PR, the hypothesis implies that, because valuation tasks require a money amount as output, individuals place particularly high (low) weight on the money (probability)

dimension, leading to relatively ‘inflated’ values for $ bets. Some recent support for this particular hypothesis is reported in Cubitt, Munro and Starmer (2004). There is, however, a vast theoretical and empirical literature connecting PR with the constructed preference approach and, for those interested in pursuing it, an excellent source is Lichtenstein and Slovic (2006).

Developing themes

One developing theme in empirical PR research examines the persistence of PR in environments where individuals receive feedback on the consequences of their decisions. A famous experiment by Chu and Chu (1990) exposed preference reversers to ‘money pumps’: subjects who committed PR had their stated preferences implemented across a series of trades which ultimately resulted in monetary losses. Individuals quickly learned to avoid PR in this environment. While this is an interesting finding, since Chu and Chu use such an explicit method for disciplining inconsistent preferences, it would be a mistake to view this as persuasive evidence that PR would be eroded in any naturally occurring market. There is some limited evidence to suggest that PR may decay in some specific experimental markets (Cox and Grether, 1996) but the findings here are both tentative and mixed, and further investigation is warranted before any firm conclusions can be drawn.

Another theme of current research explores the implications of preference anomalies (including PR) for the formulation of economic policy. A discussion of this topic is contained in Braga and Starmer (2005).

See Also

- Allais paradox
- expected utility hypothesis
- learning and evolution in games: adaptive heuristics
- paradoxes and anomalies
- prospect theory
- rational behaviour
- rationality
- rationality, bounded
- Savage's subjective expected utility model
- transitivity.

Bibliography

Becker, G.M., DeGroot, M.H. and Marschak, J. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9, 226–32.

Braga, J. and Starmer, C. 2005. Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics* 32, 55–89.

Chu, Y.P. and Chu, R.L. 1990. The subsidence of preference reversals in simplified and marketlike experimental settings: a note. *American Economic Review* 80, 902–11.

Cox, J.C. and Grether, D.M. 1996. The preference reversal phenomenon: response mode, markets and incentives. *Economic Theory* 7, 381–405.

Cubitt, R.P., Munro, A. and Starmer, C. 2004. Testing explanations of preference reversal. *Economic Journal* 114, 709–26.

Grether, D. and Plott, C.R. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69, 623–38.

Holt, C.A. 1986. Preference reversals and the independence axiom. *American Economic Review* 76, 508–15.

Karni, E. and Safra, Z. 1987. ‘Preference reversal’ and the observability of preferences by experimental methods. *Econometrica* 55, 675–85.

Lichtenstein, S. and Slovic, P. 1971. Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89, 46–55.

Lichtenstein, S. and Slovic, P. 2006. *The Construction of Preference*. New York: Cambridge University Press.

Lindman, H.R. 1971. Inconsistent preferences among gambles. *Journal of Experimental Psychology* 89, 390–7.

Loomes, G., Starmer, C. and Sugden, R. 1991. Observing violations of transitivity by experimental methods. *Econometrica* 59, 425–39.

Loomes, G.C. and Sugden, R. 1982. Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal* 92, 805–24.

Loomes, G.C. and Sugden, R. 1983. A rationale for preference reversal. *American Economic Review* 73, 428–32.

Pommerehne, W.W., Schneider, F. and Zweifel, P. 1982. Economic theory of choice and the preference reversal phenomenon: a re-examination. *American Economic Review* 73, 569–74.

Reilly, R.J. 1982. Preference reversal: further evidence and some suggested modifications in

experimental design. *American Economic Review* 73, 576–84.

Segal, U. 1988. Does the preference reversal phenomenon necessarily contradict the independence axiom? *American Economic Review* 78, 233–36.

Seidl, C. 2002. Preference reversal. *Journal of Economic Surveys* 6, 621–55.

Savage, L. 1954. *The Foundations of Statistics*. New York: Wiley.

Starmer, C.V. 2000. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38, 332–82.

Starmer, C. and Sugden, R. 1998. Testing alternative explanations of cyclical choices. *Economica* 65, 259–347.

Sugden, R. 2003. Reference-dependent subjective expected utility. *Journal of Economic Theory* 111, 172–91.

Tversky, A., Sattath, S. and Slovic, P. 1988. Contingent weighting in judgement and choice. *Psychological Review* 95, 371–84.

Tversky, A., Slovic, P. and Kahneman, D. 1990. The causes of preference reversal. *American Economic Review* 80, 204–17.

How to cite this article

Starmer, Chris. "preference reversals." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000152> doi:10.1057/9780230226203.1330

The New Palgrave Dictionary of Economics Online

Preobrazhensky, Evgenii Alexeyevich (1886 – 1937)

Michael Ellman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Preobrazhensky was an Old Bolshevik and an original and perceptive Marxist theorist. His main contribution to Marxist political economy concerned the building of socialism in a predominantly agrarian country at a low level of economic development. He argued that socialist accumulation in such a country would require an initial period of original socialist accumulation. That is, economic growth on the basis of investment generated within industry would have to be preceded, in backward Russia with its limited industry, by a period of economic growth on the basis of investment resources obtained from outside the state sector.

Keywords

accumulation of capital; agriculture and economic development; law of socialist accumulation; Marx's economics; New Economic Policy (NEP) (USSR); over-accumulation; Preobrazhensky, E.; socialism; Stalin, J.; Trotsky, L.

Article

An Old Bolshevik and a distinguished Marxist theoretician, Evgenii Alexeyevich Preobrazhensky joined the Russian Social Democratic Workers’ Party (which split into Bolshevik and Menshevik factions) in 1903 and became a professional revolutionary, being repeatedly arrested and twice subject to internal exile. He led the local party organization in the Urals during the October Revolution. In 1918 he was a member of the Left Communist group within the party which opposed the treaty of Brest-Litovsk (which ended the Russian–German war by an agreement with ‘imperialist’ Germany rather than by a revolution within Germany). He played an active role in the Civil War (1918–20). He was a full member of the Central Committee of the Russian Communist Party (Bolsheviks) and also Central Committee Secretary in 1920–1. In 1921–2 he was critical of the New Economic Policy (NEP – a mixed-economy policy which permitted peasant households to utilize freely the land they cultivated and also permitted small-

scale private enterprise in both villages and towns, while at the same time reserving the railways, large-scale industry, banking and international trade for the state). He was worried about concessions to the peasantry and their implications for rural stratification and Soviet power. A signatory to the Platform of the 46 (October 1923), he was an active oppositionist in 1924–7; he was expelled from the party in December 1927 and exiled to Siberia. Under the influence of Stalin's move to the Left, he broke with the Opposition and in July 1929 accepted Stalin's leadership. He attended the Seventeenth Party Congress (1934) where he praised Stalin and collectivization, denounced both himself and Trotsky (Stalin's chief political opponent), and advocated unity and unconditional acceptance of the party line and Stalin's leadership. Arrested in 1935, he served as a prosecution witness at the trial of Zinoviev (the former Politburo member and former chair of the executive committee of the Communist International) in 1936. Arrested again in 1936, he was not brought to a public trial, probably because of his refusal to confess to non-existent crimes. He was shot in 1937. In 1988 he was rehabilitated.

Preobrazhensky was the author of a large number of books and articles. They covered the exposition of Marxist-Leninist theory, financial and monetary questions, economic policy in France and economic policy in the USSR. Preobrazhensky's most original and important work concerned the problem of building socialism in a backward, overwhelmingly agrarian country.

Marx and Engels did not analyse how a future socialist economy would be organized and strongly opposed utopian socialism with its speculations divorced from current reality. Nevertheless, from their criticism of the anarchy of production under capitalism and their analysis of the views of rivals in the socialist movement, it is possible to draw inferences about how they expected a socialist economy to function. At the end of the 19th century Marxists had worked out some preliminary ideas for the transition to socialism and the organization of a socialist economy, as can be seen, for example, from the 1891 Erfurt Programme of the German Social Democratic Party and Kautsky's *Das Erfurter Programm* (1892), which is a commentary on it. They assumed, however, that the country concerned would be predominantly working-class and have a highly developed industry. In the 1920s, however, the Bolsheviks found themselves in power in a predominantly agrarian country at a low level of economic development. How should they build socialism in these circumstances? It is in answering this question that Preobrazhensky made his main contribution.

In *Novaia ekonomika* (1926a) he argued that, just as capitalist accumulation had required an earlier period of original accumulation as analysed in Marx (1867, vol. 1, part 8), so socialist accumulation would require an initial phase of original *socialist* accumulation. That is, economic growth on the basis of investment generated within industry would have to be preceded, in backward Russia with its limited industrial apparatus, by a period of economic growth on the basis of investment resources obtained from outside the state sector. He generalized his argument into a fundamental law of socialist accumulation which runs as follows:

> The more backward economically, petty-bourgeois, peasant, a particular country is which has gone over to the socialist organization of production, and the smaller the inheritance received by the socialist accumulation fund of the proletariat of this country when the social revolution takes place, by so much the more, in proportion, will socialist accumulation be obliged to rely on alienating part of the surplus product of pre-socialist forms of economy and the smaller will be the relative weight of accumulation on its own production basis, that is the less will it be nourished by the surplus product of the workers

> of socialist industry. Conversely, the more developed economically and industrially a country is, in which the social revolution triumphs, and the greater the material inheritance, in the form of highly developed industry and capitalistically organized agriculture, which the proletariat of this country receives from the bourgeoisie on nationalization, by so much the smaller will be the relative weight of pre-capitalist forms in the particular country; and the greater the need for the proletariat of this country to reduce non-equivalent exchange of its products for the products of the former colonies, by so much the more will the centre of gravity of socialist accumulation shift to the production basis of the socialist forms, that is, the more will it rely on the surplus product of its own industry and its own agriculture. (1926a, 1965 translation, p. 124)

As methods to obtain investment resources from the non-state sector (predominantly peasant agriculture), Preobrazhensky recommended the state monopoly of foreign trade, price policy, railway tariffs, taxation and state control of the banking system. He paid particular attention to the advantages of price policy as opposed to the use of coercion.

Preobrazhensky's analysis was very controversial when it was first published and led to a very heated debate. The reason for this is that the political basis of the Soviet regime in the 1920s was the precarious compromise between the Bolsheviks and the peasantry represented by the NEP. In addition, economic policy was based on the encouragement by the Bolsheviks for the peasants to ‘enrich yourselves’. It was hoped that the development of peasant agriculture, in a mixed economy in which the commanding heights were in the hands of the state, would provide the food, raw materials, exports, internal market and labour force necessary for Soviet economic development. Hence Preobrazhensky's argument, with its presentation of the case for accumulation at the expense of peasant agriculture, was both politically and economically very disturbing. In particular, the analogy with original capitalist accumulation was distinctly ominous. According to Marx, original capitalist accumulation was based mainly on force, in particular on the use of force to expropriate the land from the peasantry. In the minds of the supporters of NEP, Preobrazhensky's analysis raised the spectre of a revival of the methods of War Communism (that is, requisitioning based on direct coercion, rationing, and attempted state control of the whole economy, rather than market economy methods).

Preobrazhensky's ideas evolved over time. In a paper of 1921 (1980, pp. 3–19), the very year the NEP was introduced, he anticipated an armed conflict between the Soviet state and the kulaks. He regarded this as inevitable and argued in good Stalinist style that ‘the outcome of the struggle will depend largely on the degree of organization of the two extreme poles, but especially on the strength of the state apparatus of the proletarian dictatorship’. He concluded his argument, which was published at a time of serious famine and disease, partly caused by the class-war policies of the Bolsheviks, by warning his readers ‘to prepare for everything that will ensure victory in the inevitable class battles that are to come’. In a paper of 1924, the thesis about the inevitable conflict between the state and the peasantry still plays a central role, but economic levers (for example, price policy) rather than coercion play the key role in resolving the conflict in the interests of socialist accumulation.

In a paper of 1927, attention has shifted to the conditions for growth equilibrium. The Harrodian conclusion about the essential precariousness of dynamic equilibrium is reached. The lesson is drawn that ‘The sum of these contradictions shows how closely our development towards socialism is connected with the necessity – for not only political but also for economic reasons – to make a break in

our socialist isolation and to rely in the future on the material resources of other socialist countries.'
In an unpublished paper of 1931 he criticized over-investment and pointed out the danger of an 'overaccumulation crisis'. His argument that 'socialism is production for consumption's sake' was unacceptable during the frenzy of the Soviet Great Leap Forward and was condemned as heretical. His position in 1931 seems to have been similar to that of Rakovsky, another Left Communist intellectual, who in an article of 1930 (published in 1931 and translated into English in 1981) warned against the coming Soviet economic crisis (which shook the whole economy in 1931–3) and stressed the wasteful and inefficient methods of Stalinist industrialization.
The accumulation that Preobrazhensky theorized about was *socialist* accumulation, that is, accumulation leading to the development of socialist relations of production. It is entirely natural, for example, that the imaginary author of Preobrazhensky's book *From NEP to Socialism* (1922), which takes the form of lectures supposedly given in 1970, is simultaneously a university professor and a fitter in a railway workshop. This reflected Preobrazhensky's expectation that the division of labour would be sharply reduced under socialism.
Preobrazhensky's work has had an enormous influence throughout the world. In the USSR in the 1920s he played a major role in the debate about the main directions of economic policy. In the West he was rediscovered in Erlich's famous paper in the *Quarterly Journal of Economics* (1950) and has been much discussed ever since. In the Third World his ideas play an important role in theoretical discussions and policy debates. He is rightly considered one of the outstanding Marxist economists of the 20th century.

See Also

- agriculture and economic development
- development economics
- Marx's analysis of capitalist production

Selected works

1920. (With N. Bukharin.) *Azbuka kommunizma*. Petrograd. Trans. E. and C. Paul as *The ABC of Communism*. Ann Arbor: University of Michigan Press, 1966.

1921. *Bumazhnye den'gyi v epokhu proletarskoi dictatury* [Paper money in the epoch of the proletarian dictatorship]. Tbilisi: Gosizdat.

1922. *Ot nepa k sotzializmu*. Moscow. Trans. B. Pearce as *From NEP to Socialism*. London: New Park Publishers, 1973.

1924. *Ekonomicheskie krizisy pri NEP'e* [Economic crises under NEP]. Moscow: Izdatel'stvo Sotsialisticheskoi Akademii.

1926a. *Novaia ekonomika*. Moscow. Trans. B. Pearce as *The New Economics*. Oxford: Clarendon Press,

1965.

1926b. *Ekonomika i finansy sovremennoi frantsii* [The economics and finances of contemporary France]. Moscow: Izdatel’stvo Kommunisticheskoi Akademii.

1927. Khozyaistvennoe ravnovesie v sisteme SSSR [Economic equilibrium in the system of the USSR]. *Vestnik kommunisticheskoi akademii*, No. 22.

1930. *Teoriia padaiushchei valiuty* [The theory of a depreciating currency]. Moscow: Gosizdat.

1931. *Zakat kapitalizma*. Moscow. Trans. R. Day as *The Decline of Capitalism*. New York: M.E. Sharpe, 1985.

1980. *The Crisis of Soviet Industrialization*, ed. D. Filtzer. London: Macmillan; New York: M.E. Sharpe. (This book of selected articles contains on pp. 237–40 a select bibliography of Preobrazhensky's works).

Bibliography

Day, R. 1975. Preobrazhensky and the theory of the transition period. *Soviet Studies* 27(2), 196–219.

Erlich, A. 1950. Preobrazhenski and the economics of Soviet industrialization. *Quarterly Journal of Economics* 64(1), 57–88.

Erlich, A. 1960. *The Soviet Industrialization Debate*. Cambridge, MA: Harvard University Press.

Filtzer, D. 1978. Preobrazhensky and the problem of the Soviet transition. *Critique* 9, 63–84.

Karshenas, M. 1995. *Industrialization and the Agricultural Surplus: A Comparative Study of Economic Development in Asia*. Oxford: Oxford University Press.

Kautsky, K. 1892. *Das Erfurter Programm*. Stuttgart: Dietz.

Marx, K. 1867. *Das Kapital*, vol. 1. Hamburg: Otto Meissner.

Millar, J. 1978. A note on primitive accumulation in Marx and Preobrazhensky. *Soviet Studies* 30(3), 384–93.

Rakovsky, C. 1931. Na s”ezde i v strane [At the congress and in the country]. *Byulleten’ Oppozitsii*, No. 25–26, pp. 9–32. Trans. D. Filtzer as ‘The Five Year Plan in crisis’. *Critique* (1981), No. 13.

Sah, R. and Stiglitz, J. 1984. The economics of price scissors. *American Economic Review* 74, 125–38.

Sah, R. and Stiglitz, J. 1987. Price scissors and the structure of the economy. *Quarterly Journal of Economics* 102(1), 109–34.

Sah, R. and Stiglitz, J. 1992. *Peasants Versus City-dwellers: Taxation and the Burden of Economic Development*. Oxford: Oxford University Press.

How to cite this article

Ellman, Michael. "Preobrazhensky, Evgenii Alexeyevich (1886–1937)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000156> doi:10.1057/9780230226203.1331

The New Palgrave Dictionary of Economics Online

preordering

C. Blackorby

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

difference principle (J. Rawls); preordering; reflexivity; transitivity; utilitarianism

Article

A preordering (also called a weak ordering or a quasi-ordering) is a reflexive and transitive binary relation which is not necessarily complete.

A binary relation *R* defined on a set *S* is a set of ordered pairs of elements of *S*, that is, a subset of the Cartesian product of *S* with itself, *S*x*S*. One writes *xRy* (or $(x, y) \in R$) to mean that $x \in S$ stands in relation R to $y \in S$. A preordering is a binary relation, *R*, which satisfies two properties: (i) reflexivity: for all $x \in S$ *xRx*, and (ii) transitivity: for $x, y, z \in S$, if *xRy* and *yRz*, then *xRz*.

A simple example is given by the binary relation weak vector dominance which we denote *V*. Suppose *S* is Euclidean *N*-space, then *xVy* if and only if $x_n \geq y_n, n = 1, \ldots, N$. *V* is clearly reflexive and transitive; it is just as clearly not complete, that is, not all elements of *S* are ranked. For example if *N*=2, *x*=(1, 2), and *y*=(2, 1) then it is not the case that *xVy* or that *yVx*.

Quasi-orderings have played their largest role in welfare economics where consistency in decision making is a desirable requirement but where one may be dubious about being able to rank all possible outcomes. Two examples follow for which the notion of a subrelation is useful. Suppose *R* and *S* are binary relations: *S* is a subrelation of *R* if *xSy* implies *xRy*. For example, strong vector dominance, $\bar{V}$ is the binary relation which results when the above weak inequality is replaced with a strict inequality. Clearly, $\bar{V}$ is a subrelation of *V*.

Interpreting the elements of *N*-space as vectors of utilities, it is possible to define a quasi-ordering which is a subrelation of both the utilitarian and the Rawls criteria: Define the binary relation M by *xMy* if and only if $\Sigma_{i=1}^{N} x_i \geq \Sigma_{i=1}^{N} y_i$ and $\min\{x_1, \ldots, x_N\} \geq \min\{y_1, \ldots, y_N\}$. *M* is clearly reflexive, transitive and not complete. The distributional insensitivity of the utilitarian principle is tempered by the Rawls's difference principle.

As an alternative, consider evaluating social states by weighted utility sums where the weights represent

utility comparisons but these comparisons are not precisely fixed. Instead, the weights are drawn from a subset of *N*-dimensional Euclidean space, say *B*. More formally define the quasi-ordering *F* by *xFy* if and only if $\Sigma_{i=1}^{N} b_i x_i \geq \Sigma_{i=1}^{N} b_i y_i$ for all $(b_1, \ldots, b_N) \in B$. Suppose that we try to evaluate the desirability of burning down Rome while Nero fiddles. The quasi-ordering *F* may show a gain for burning Rome only if the set of interpersonal weights is such that Nero is given extreme consideration. (These examples are taken from the articles listed below.)

See Also

- lexicographic orderings
- orderings
- transitivity

Bibliography

Blackorby, C. and Donaldson, D. 1977. Utility vs. equity: some plausible quasi-orderings. *Journal of Public Economics* 7, 365–82.

Sen, A.K. 1970. Interpersonal aggregation and partial comparability. *Econometrica* 38, 393–409.

How to cite this article

Blackorby, C. "preordering." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000157> doi:10.1057/9780230226203.1332

The New Palgrave Dictionary of Economics Online

Prescott, Edward Christian (born 1940)

Stephen L. Parente

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Edward Prescott was awarded the Nobel Prize in Economics in 2004 with Finn Kydland for their contributions to dynamic macroeconomics. Prescott is a member of a small group of economists who, starting in the 1970s, revolutionized macroeconomics by challenging the Keynesian consensus. While he is best known for his research on business cycles and the optimal design of economic policy, he has made important contributions to other applied fields, such as finance and development, as well as to economic theory. He has also made important contributions to methodology, having pioneered many of the standard contemporary techniques and tools in macroeconomics.

Keywords

American Academy of Arts and Sciences; American Economic Association; Bayesian analysis; business cycle; clubs; consumer surplus; Debreu, G.; discretion vs rules; dynamic programming; Econometric Society; economic development; endogenous growth models; equity premium; financial economics; firms; Friedman, M.; Gibrat's law; Haavelmo, T.; Hodrick–Prescott filter; imperfect information; industrial organization; international inequality; investment equation; Klein, L.; lotteries; Lucas, R.; microeconomic foundations; misperceptions theory; model calibration vs. statistical estimation; Modigliani, F.; moral hazard; Muth, J.; neoclassical growth model; New Classical Economics; organizational capital; partial equilibrium; Prescott, E.; productivity shocks; rational expectations theory; real business cycle theory; recursive competitive equilibrium; recursive methods; relative incomes vs relative growth rates; rules vs discretion; statistical estimation vs. model calibration; stochastic process; time inconsistency; total factor productivity

Article

Edward Christian Prescott is one of the leading macroeconomists of our time. He received the Noble

Prize in Economics in 2004, an award he shared with Finn Kydland. Prescott's influence on the evolution of macroeconomics has been profound and far-reaching.

Prescott is a member of a small group of economists who, starting in the 1970s, revolutionized macroeconomics by challenging the Keynesian consensus that had held sway for half a century. This revolution, known as the 'Rational Expectations Revolution', occurred primarily at Carnegie Mellon University, the University of Chicago, the University of Minnesota, the University of Pennsylvania and Rochester University. Initially, this revolution was seen as the start of a new school of economic thought – New Classical Economics – that was based on the assumptions of market clearing and rational expectations. Today this revolution is seen simply as the start of an alternative approach to macroeconomics, one that advocates dynamic general equilibrium models with strong microeconomic foundations. This approach now dominates macroeconomics.

Prescott is best known for his applied research on business cycles, economic development and growth, and financial markets. In addition to his applied work, Prescott has produced a number of theoretical papers. A major line of this research demonstrates how to apply classical competitive analysis to economies with frictions, economies that previously had been regarded as outside the realm of such analysis. Within the profession, however, some of Prescott's most lasting impacts have come in the methodology used for macroeconomic research. Many of the standard contemporary tools and techniques in macroeconomics were pioneered by Prescott. Finn Kydland and Edward Prescott together introduced calibrated models to macroeconomics; in doing so, they fundamentally changed the way applied macroeconomics is carried out. Prescott's work has also been important for the development and diffusion of dynamic general equilibrium techniques and recursive methods. His work has also altered the ways in which economists handle data; for example, the so-called Hodrick–Prescott filter has become a standard tool of those working with time series data displaying trends. Thus, through applied work, theory, and methodology, Prescott has made lasting contributions to economics.

History

Edward C. Prescott was born in Glenn Falls, New York on 26 December 1940. He graduated from Glenn Falls High School in 1958 and Swarthmore College in 1962, with a BA in mathematics. In 1963 he received a Masters degree in Operational Research from Case University, which later became Case Western University. Thereafter, he enrolled in the Ph.D. programme at the Graduate School of Industrial Administration at Carnegie Mellon University (CMU), completing his doctorate in 1967.

Prescott started his academic career in 1967 as an assistant professor in the economics department at the University of Pennsylvania. In 1971 he left and accepted a position at CMU at the rank of assistant professor. He was promoted to the level of associate professor in 1972 and full professor in 1975. In 1974 he visited the Norwegian School of Business and Economics for a year at the invitation of Finn Kydland, who had written his dissertation under Prescott's supervision at CMU. The visit is significant, as it was the occasion for much of the work for which the pair were later awarded the Nobel Prize. Prescott officially left CMU in 1980. Between 1978 and 1982, Prescott was a visiting professor at both the University of Chicago and Northwestern University. In 1981 he accepted a position at the University of Minnesota, where (with the exception of 1998, when he was a professor at the University of Chicago) he remained until 2003. At Minnesota Prescott was appointed a Regent's Professor in 1996 and the McKnight Presidential Chair in Economics in 2003. In 2003 he left Minnesota to become the W.P.

Carey Professor at the W.P. Carey Business School at Arizona State University. In addition to these academic positions, Prescott has served as an advisor to the Federal Reserve Bank of Minneapolis since 1981.

Prescott has received numerous honours in his distinguished career, with the Nobel Prize in 2004 being the most prestigious. He was elected to the Econometric Society in 1980 and the American Academy of Arts and Sciences in 1992. He is the recipient of the 2002 Erwin Plein Nemmers Prize in Economics, awarded by Northwestern University biannually to an outstanding economist. He was chosen to give the First Lionel McKenzie Lecture at the University of Rochester in 1990, the Third Walras–Pareto Lecture at the Univeristé de Lausanne in 1994, and the First Lawrence Klein Lecture at the University of Pennsylvania in 1997. In addition, he was chosen to give the Richard T. Ely Lecture to the American Economic Association in 2002.

Research

Prescott's graduate training was largely in statistics. His thesis advisor was Mike Lovell. In addition, Maurice de Groot, one of the greatest Bayesian statisticians, was involved in the supervision of Prescott's work. Prescott's dissertation, titled ‘Adaptive decision rules for macroeconomic planning’, was an exercise in Bayesian statistical decision theory.

By his own admission Prescott was more of a statistician than an economist in the first years of his career. This changed in 1969, the year, he wrote ‘Investment under uncertainty’ with Robert E. Lucas (1971), whom Prescott had met while he was a graduate student at CMU. The paper studied the optimal investment decision of firms in an industry faced with stochastic demand. The paper was still in the tradition of the Keynesian ‘system of equations’ approach pioneered by Lawrence Klein that dominated macroeconomics at that time; its purpose was to derive a better investment equation to be used in large macro models. As such, it followed the trend established by Milton Friedman, Franco Modigliani, James Tobin and Trygve Haavelmo, who sought to base the individual equations in these systems on microeconomic theory.

The paper marks a watershed in the development of macroeconomics, however. It is one of the first to incorporate John Muth's hypothesis of rational expectations. The assumption of rational expectations forced Lucas and Prescott to develop and apply a new set of tools and concepts that have since become standard in macroeconomic research. For example, the paper introduced the concept of an equilibrium as a stochastic process. The paper is also important in the development of dynamic general equilibrium analysis; although the paper studied a partial equilibrium problem, the rational expectations assumption required that Lucas and Prescott simultaneously study the optimization problems of agents as well as the industry equilibrium. The paper also demonstrated how the competitive equilibrium could be solved from the social planner's problem of maximizing consumer surplus.

The ‘Investment under uncertainty’ paper is also important in that it showed how dynamic programming techniques developed in statistics and operations research could be successfully applied by economists to solve complicated optimization problems. In this respect, Prescott's previous training at Case University proved extremely valuable. The paper was to some extent a precursor to the concept of a ‘recursive competitive equilibrium’, that is, a set of time-invariant decision rules and prices that are functions of limited number of state variables.

In the 1970s Prescott continued to develop the recursive methods that now are commonplace in

macroeconomics. In 1980, he published a paper with Rajish Mehra that extended and generalized recursive equilibrium theory. He also collaborated with Nancy Stokey and Robert Lucas in writing a comprehensive and complete book on the subject, *Recursive Methods in Economic Dynamics*.

Classical competitive analysis

The problems that Lucas and Prescott encountered in the ‘Investment under uncertainty’ paper necessitated that they read a number of papers in mathematical economics. One of these papers was Debreu's classic (1954) paper on competitive equilibrium analysis, ‘Valuation equilibrium and Pareto Optimum’. This paper had a great impact on Prescott's thinking. First, the paper taught Prescott that many apparent market failures disappeared if mutually beneficial trades were permitted. Additionally, the paper showed to Prescott the power and importance of framing economic problems using the correct mathematical and verbal language.
Since the 1980s Prescott has written several papers that show how classical competitive analysis can be applied to a number of economies with frictions once the commodity space (that is, the set of tradable objects) is appropriately reframed. For many economies, the appropriate commodity space is defined over lotteries, namely, contracts with random components over goods and actions. Prescott and Townsend (1984) apply this approach to economies with moral hazard. Prescott and Rios-Rull (1992) do this in economies where people move between locations or occupations, and where at any one location there is imperfect information on the state of the other occupations. Hornstein and Prescott (1993) demonstrate how a number of potentially important production structures can also be mapped into this structure. Finally, Cole and Prescott (1997) show how this can be done in a class of economies where agents voluntarily form associations or clubs that carry out joint activities. These theoretical contributions were all intended to allow macroeconomists to address applied issues of policy relevance.

Industrial organization

Prescott stopped teaching macroeconomics for a period in the 1970s, saying that it made little sense to teach a subject that one did not understand. During this period he primarily taught graduate courses in industrial organization (IO). A number of IO papers grew out of this teaching, such as Prescott (1973) and Prescott and Visscher (1977; 1980). Prescott and Visscher (1980) is an extremely important work. The paper shows that the acquisition of information, or organizational capital, by firms acts as an important cost of adjustment that limits firm growth. The paper makes an important contribution to the IO literature because it explains a number of empirical regularities, including Gibrat's law that firm size and growth are independent. It also makes an important contribution to the macroeconomic literature because the concept of organizational capital, which the paper introduced, has been used by a large number of researchers to understand a variety of phenomena.

Rules and real business cycle theory

During this period Prescott did not abandon research in the field of macroeconomics; in fact, he wrote with Finn Kydland two of the most important papers in macroeconomics: ‘Rules rather than discretion:

the time inconsistency of optimal plans', published in 1977, and 'Time to build and aggregate fluctuations,' published in 1982. The Nobel Prize Committee pointed to these two papers as the basis for awarding Kydland and Prescott the Noble Prize in 2004.

The substance of those papers is well known. Almost every undergraduate macroeconomic textbook written since the mid-1980s provides extensive coverage of both topics. The 'Time inconsistency' paper showed that people were made better off if the policymaker were to use a good rule instead of his discretion. Discretion – the ability of the policymaker to change his mind – leads to a worse outcome because the announced policy is typically not the optimal one to follow at the date of implementation. The 'Time to build' paper showed that productivity shocks account for roughly two-thirds of the volatility of US output over the business cycle in the post-war period. This productivity-driven view of the business cycle has come to be known as 'real business cycle theory'.

The idea that productivity shocks, which correspond to changes in the economy's stock of knowledge as well as changes in regulation or institutional factors, account for most of the US business cycle initially met fierce resistance. This was not surprising as it challenged the Keynesian view that the business cycle was a demand-driven phenomenon that called for government intervention on account of frictions. Many people objected to the theory on the grounds that the model contained no monetary side. These critics not only missed the point but they missed the fact that the precursor of the 'Time to build' paper (Kydland and Prescott, 1980) did contain a monetary side based on Lucas's (1972) misperceptions theory. As that paper found that monetary shocks were quantitatively unimportant for understanding the US business cycle, Kydland and Prescott abstracted from money in their 1982 paper.

Over the years, resistance to Kydland and Prescott's theory of the business cycle has waned. The theory has been examined intensively, in fact probably more so than any other theory in economics to date. Attempts to discredit it have proven unsuccessful. Today, the idea that most of the US business cycle is a supply side phenomenon driven by productivity shocks is almost universally accepted.

The contribution of the 'Time to build' paper to the field of macroeconomics goes well beyond this substantive point. It also makes an important methodological point. Specifically, it lays the foundation for the use of deductive inference or model calibration to the application of macroeconomic issues. In this approach, the model is viewed as a measuring device, a so-called thermometer, to deduce or derive the quantitative implications of theory. This is in contrast to the inductive inference approach, or statistical estimation, that dominated the Keynesian 'system of equations' approach. There, the researcher attempts to discover the model out of a class of models that is the one to have most likely generated the data.

In effect, the 'Time to build' paper was an attempt to derive the quantitative implications of neoclassical growth theory for business cycles. Kydland and Prescott posed the question of whether the widely studied neoclassical growth model could be used not only to analyse long-run growth in the US economy but also to study business cycles. Starting with a standard neoclassical model, Kydland and Prescott restricted the values of the model parameters so that it quantitatively matched the trend growth of the US economy. They then modified the model so that productivity did not grow mechanically from year to year but instead followed a stochastic process that was based on the properties of 'Solow residuals' calculated from the data. Kydland and Prescott then reinserted these stochastic productivity shocks into the model and computed the equilibrium properties of the model. A striking result was that the model economy displayed business cycles that mirrored those found in the macro data, provided the labour supply was reasonably elastic.

Model calibration has become the dominant methodology in macroeconomics. It is widely used to test and develop theory as well as to evaluate policy. It is particularly useful in evaluating policy scenarios that are far ‘out of sample’ compared with historical experience. By using calibrated models, economists can conduct experiments on entire economies in a way that is not generally possible (or desirable!) with real economies.

Since 1982 Prescott and Kydland have continued to develop this line of research. They have subsequently written a number of articles to educate the profession in the use of model calibration (Kydland and Prescott, 1991a; 1996; Prescott, 2001). They have also written a number of articles that modify the model in their 1982 paper in order to explore further the implications of growth theory for understanding business cycles (Kydland and Prescott, 1988; 1991b; Cooley, Hansen and Prescott, 1995).

The equity premium and international income differences

The application of the calibration methodology to a large number of macroeconomic questions has yielded important insights. As its founder, no person has used this methodology more effectively than Prescott. In addition to business cycles, Prescott's work has produced important insights in finance and economic development and growth.

Prescott's paper, ‘The equity premium: a puzzle’, co-authored with Rajnish Mehra and published in 1985, is a seminal work in financial economics. The paper sought to determine how much of the 6.2 percentage point difference between the average historical after-tax real rate of return on equity and the average historical after-tax return on bonds in the United States could be attributed to a premium for bearing non-diversifiable aggregate risk. The paper shows convincingly that households’ aversion to risk cannot account for most of the difference in real rates of returns between bonds and equity. Prescott and Mehra's work has spawned an entire literature in financial economics, the goal of which is to solve the puzzle they uncovered.

Prescott's research has also fundamentally changed the way we think about the wealth and poverty of nations, and has set the direction of subsequent research in the area of economic development and growth. Prescott is among the very first researchers to argue that a theory of relative income levels, rather than relative growth rates, is the goal of development economics. Prescott did not start out with this view. Prescott's first paper on economic growth co-authored with John Boyd (Boyd and Prescott, 1987) is an endogenous growth model whereby differences in policies or preferences across countries generate permanent differences in growth rates. After examining the development and growth facts over the period 1950–85 (Parente and Prescott, 1993), however, Prescott concluded that the data did not support such a theory. The switch to a theory of relative income levels is evident in Parente and Prescott (1994). Today, the vast majority of papers that attempt to explain the huge disparity in international incomes take this relative income approach. Prescott is also one of the first researchers to have argued rigorously that differences in total factor productivity (TFP; that is, the efficiency with which a country uses its resources) account for most of the differences in international incomes. This is the main message of his 1997 Lawrence R. Klein Lecture, ‘Needed: a theory of total factor productivity’, published in 1998. Today, this view is almost universally accepted.

It should be no surprise that Prescott himself went on to provide a theory of TFP. In *Barriers to Riches*, Parente and Prescott (2000) demonstrate how a country's TFP is determined by policies that effectively constrain firms in their choice and use of technologies. Parente and Prescott (1999) completed the theory

by arguing that these constraints typically exist to protect groups who stand to lose through the introduction of better technology.

No decrease in TFP

Prescott has continued to remain highly productive; if anything, his productivity has increased in recent years. This recent research is mostly applied in nature. Like much of Prescott's previous work, it derives the implications of neoclassical growth theory for a variety of macroeconomic phenomena. One branch of this recent research uses the growth model to examine several long-standing hypothesis and puzzles in financial economics, including the equity premium (McGrattan and Prescott, 2000; 2003; 2004). A different branch of this recent work uses the growth model to understand the reasons for the different experiences of OECD countries in the postwar period (Prescott, 2003; 2004; Hayashi and Prescott, 2002).

Teaching and mentoring

It would be a serious omission not to mention Prescott's long-lasting commitment to teaching and advising. Prescott has supervised over 55 dissertations in his career, and many of his students are well-known economists such as Tom Cooley, Costas Aziaridis, Finn Kydland, Charlie Holt, Ed Green, Rajnish Mehra, Barbara Spencer, V.V. Chari, Hugo Hopenhayn, Richard Rogerson, Gary Hansen, Rody Manuelli, Gerhard Glomm, Jim Schmitz, Ayse Imrohoglu, Andreas Hornstein, Victor Rios-Rull, Fernando Alvarez, Dirk Krueger and Betsy Caucutt. Prescott's popularity as mentor reflects his philosophy that students should be treated as colleagues and that a good teacher has as much to learn from his students as they have to learn from him. The success of his students clearly speaks for the rigour that Prescott demands as well as the independence and confidence he instils in them. Prescott's students feel extremely fortunate to have had such a generous, engaging and inspiring advisor.

Conclusion

Edward C. Prescott is one of the most influential economists in the history of macroeconomics. His work has fundamentally changed the way economists conduct macroeconomic research and altered the way economists think about a large number of macroeconomic issues. Perhaps Robert E. Lucas in his introduction to Prescott's 2002 Richard T. Ely lecture best summarized Prescott's effect on economics, when he wrote

> We can remember the way we thought about the issues before Prescott's analysis, and the comparison with the way we think about them now gives a measure of the enormous effect each paper has had on our thinking… [Prescott’s] papers met with resistance, but in the end [he has] caused us to rearrange important parts of our vision of how the economy works, to start over in many respects.

See Also

- calibration
- dynamic programming
- general equilibrium
- growth and cycles
- inequality between nations
- rational expectations models, estimation of
- recursive competitive equilibrium
- total factor productivity

Selected works

1971. (With R. Lucas.) Investment under uncertainty. *Econometrica* 39, 659–81.

1973. Market structure and monopoly profits: a dynamic theory. *Journal of Economic Theory* 6, 546–57.

1977. (With F. Kydland.) Rules rather than discretion: the time inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

1977. (With M. Visscher.) Sequential location among firms with foresight. *Bell Journal of Economics* 8, 378–93.

1980. (With R. Mehra.) Recursive competitive equilibrium. *Econometrica* 48, 1365–79.

1980. (With M. Visscher.) Organizational capital. *Journal of Political Economy* 88, 445–61.

1982. (With F. Kydland.) Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

1984. (With R. Townsend.) Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica* 52, 21–45.

1985. (With R. Mehra.) The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

1987. (With J. Boyd.) Dynamic coalitions, growth and the firm. *American Economic Review* 67, 63–7.

1988. (With F. Kydland.) The work week of capital and its cyclical implications. *Journal of Monetary Economics* 21, 343–60.

1989. (With R. Lucas.) *Recursive Methods Economic Dynamics*. Cambridge, MA: Harvard University Press.

1990. (With F. Kydland.) The workweek of capital and its cyclical implications. *Journal of Monetary Economics* 21, 340–60.

1991a. (With F. Kydland.) The econometrics of the applied general equilibrium approach. *Scandinavian Journal of Economics* 93, 161–78.

1991b. (With F. Kydland.) Hours and employment variations in business cycle theory. *Economic Theory* 1, 63–92.

1992. (With J. Rios-Rull.) Classical competitive analysis with islands. *Journal of Economic Theory* 57, 73–98.

1993. (With A. Hornstein.) The plant and the firm in general equilibrium theory. In *General Equilibrium, Growth, and Trade II: The Legacy of Lionel McKenzie*, ed. R. Becker et al. San Diego, London, Sydney and Toronto: Harcourt Brace.

1993. (With S. Parente.) Changes in the wealth of nations. *Federal Reserve Bank of Minneapolis Quarterly Review* 17, 3–16.

1994. (With S. Parente.) Barriers to technology adoption and development. *Journal of Political Economy* 102, 298–321.

1995. (With T. Cooley and G. Hansen.) An equilibrium analysis of idle resources and varying capacity utilization rates. *Economic Theory* 6, 35–50.

1996. (With F. Kydland.) The computational experiment: an econometric tool. *Journal of Economic Perspectives* 10, 68–86.

1997. (With H. Cole.) Valuation equilibrium with clubs. *Journal of Economic Theory* 74, 19–39.

1998. Needed: a theory of total factor productivity. *International Economic Review* 39, 525–52.

1999. (With S. Parente.) Monopoly rights: a barrier to riches. *American Economic Review* 89, 1216–33.

2000. (With E. McGrattan.) Is the stock market overvalued? *Federal Reserve Bank of Minneapolis Quarterly Review* 24, 20–40.

2000. (With S. Parente.) *Barriers to Riches*. Cambridge, MA: MIT Press.

2001. Business cycle theory: methods and problems. In *Cycles, Growth and Structural Change: Theories and Empirical Evidence*, ed. L.P. Punzo. London: Routledge.

2002. (With F. Hayashi.) The 1990s in Japan: a lost decade. *Review of Economic Dynamics* 5, 206–35.

2003. (With E. McGrattan.) Average debt and equity returns: puzzling? *American Economic Review* 93, 392–97.

2003. Prosperity and depression. *American Economic Review* 92, 1–15.

2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28, 2–13.

2004. (With E. McGrattan.) The stock market crash of 1929: Irving Fisher was right! *International Economic Review* 45, 991–1009.

Bibliography

Debreu, G. 1954. Valuation equilibrium and Pareto Optimum. *Proceedings of the National Academy of Sciences* 40, 588–92.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

How to cite this article

Parente, Stephen L. "Prescott, Edward Christian (born 1940)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000321> doi:10.1057/9780230226203.1333

The New Palgrave Dictionary of Economics Online

present value

Stephen F. LeRoy

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The present value relation says that, under certainty, the value of a capital good or financial asset equals the summed discounted value of the stream of revenues which that asset generates. Otherwise arbitrage would be possible. Under uncertainty, and if risk neutrality is assumed, the future payoffs are replaced by their conditional expectations. Under risk aversion either the natural probability measure under which expectations are taken must be replaced by a ‘risk-neutral measure’, or the discount factor must be modified by a factor that reflects risk. The present value relation leads to bubbles if a convergence condition is not satisfied.

Keywords

arbitrage; bubbles; capital asset pricing model; capital budgeting; capital market efficiency; excess volatility tests; Fisher's separation principle; martingales; present value; risk aversion; risk neutrality; risk premium; risk-neutral probabilities; speculative bubbles; uncertainty; wealth-maximization decision rule

Article

The present value relation says that, under certainty, the value of a capital good or financial asset equals the summed discounted value of the stream of revenues which that asset generates. The discount factor is that determined by the interest rate over the relevant period. The justification for the present value relation lies in the fact that (in perfect capital markets) an asset must earn a rate of return exactly equal to the interest rate. Otherwise arbitrage opportunities emerge, which is inconsistent with equilibrium.

Derivation of the present value relation

If r_t is the one-period interest rate at t, p_t is the (ex-dividend) price of an asset and d_t is its dividend, it

must be true that

$$r_t = (d_{t+1} + p_{t+1}) / p_t - 1, \tag{1}$$

since the right-hand side equals the rate of return on the asset. Solving for p_t,

$$p_t = \frac{d_{t+1} + p_{t+1}}{1 + r_t}. \tag{2}$$

Replacing t by $t + 1$, (2) becomes an equation expressing p_{t+1} as a function of r_{t+1}, d_{t+1} and p_{t+2}. If the resulting expression is substituted to eliminate p_{t+1} in (2), and if this operation is repeated n times, it follows that

$$p_t = \sum_{i=1}^{n} \frac{d_{t+i}}{\Pi_{j=0}^{i-1}(1 + r_{t+j})} + \frac{p_{t+n}}{\Pi_{j=0}^{i-1}(1 + r_{t+j})}. \tag{3}$$

If speculative price bubbles are assumed not to occur (see below), the right-most term in (3) converges to zero as n goes to infinity, so there results the present value equation

$$p_t = \sum_{i=1}^{\infty} \frac{d_{t+i}}{\Pi_{j=0}^{i-1}(1 + r_{t+j})}. \tag{4}$$

If in addition the interest rate is constant at $r_t = \rho$, (3) may be written as

$$p_t = \sum_{i=1}^{n} (1+\rho)^{-i} d_{t+i} + (1+\rho)^{-n} p_{t+n}. \tag{5}$$

or, if the convergence condition is satisfied, as

$$p_t = \sum_{i=1}^{\infty} (1+\rho)^{-i} d_{t+i}. \tag{6}$$

In the special cases in which d_{t+i} is constant at d, or grows from d_t at rate g, (6) simplifies to

$$p_t = \frac{d}{\rho} \tag{7}$$

or

$$p_t = \frac{d_t(1+g)}{\rho - g}, \tag{8}$$

respectively.

Present value in capital budgeting

In introductory finance courses, the present value relation makes an early appearance in the chapter on capital budgeting, where it is taught that corporations should accept any investment project that promises a positive present value (net of costs), and only these. This wealth-maximization decision rule is the correct one independent of agents' preferences because, regardless of preferences, the consumption set that it generates dominates that generated by any other capital budgeting criterion. This is Irving Fisher's separation principle. Other criteria, such as accepting that project with the shortest payback period, or

that with the highest internal rate of return, are either equivalent to present value maximization, ambiguous (sometimes, for example, a single project may have no real internal rate of return, or more than one) or wrong, depending on the characteristics of the project's returns.

Present value under uncertainty

Under uncertainty, but on the assumption of risk neutrality, the present value relation may be written as

$$p_t = \sum_{i=1}^{\infty} (1+\rho)^{-i} E_t(d_{t+i}), \tag{9}$$

which differs from (6) only in that future dividends is replaced by its conditional expectation. This version of the present value relation has received extensive study, especially in the early finance literature. It is easily shown to imply

$$E_t(r_t) = \rho, \tag{10}$$

saying that the conditional expectation of the rate of return on the asset equals a constant independent of the conditioning set (Samuelson, 1965; 1970). Here r_t is defined in (1); use of the same notation for the interest rate above and the rate of return on any asset here reflects the fact that under certainty the return on any asset equals the interest rate. This strong restriction provides the basis for most empirical tests of what has been called ‘capital market efficiency’ (Fama, 1970; LeRoy, 1989): if (10) is true, no information publicly available at *t* should be correlated with the rate of return on the asset from *t* to $t+1$. In this sense prices ‘fully reflect’ all publicly available information.

The present value relation may also be interpreted from the vantage point of its martingale implication: if the asset is priced according to (9), the value x_t of a mutual fund which holds the asset and reinvests all of its dividend income will follow a martingale with drift, defined by

$$E_t(x_{t+1}) = (1+\rho)x_t. \tag{11}$$

To see this, assume that the mutual fund holds h_t shares of the asset so that

$$x_t = h_t p_t \text{ and } x_{t+1} = h_{t+1} p_{t+1}. \tag{12}$$

When dividend income is reinvested, h_{t+1} is given by the value that solves

$$h_{t+1} p_{t+1} = h_t(p_{t+1} + d_{t+1}). \tag{13}$$

Then

$$E_t(x_{t+1}) = E_t(h_{t+1} p_{t+1}) = h_t E_t(d_{t+1} + p_{t+1}) = x_t(1 + \rho). \tag{14}$$

Here we used (1) and (10). To see that the correction for dividends payout is needed, observe that (10) implies that

$$\rho = \frac{E_t(d_{t+1})}{p_t} + \frac{E_t(p_{t+1})}{p_t} - 1, \tag{15}$$

so that changes in the expected dividend yield are always offset one-for-one by changes in the expected rate of capital gain. If p_t by itself were a martingale the expected rate of capital gain would be a constant, implying that p_t is a constant multiple of expected dividends. But this is not an implication of the present value relation (take dividends as given by a first-order autoregressive process, for example). Hence p_t by itself does not follow a martingale.

The present value–martingale model appears in many contexts in finance. If a futures price is assumed equal to the conditional expectation of the relevant spot price, then the futures price will follow a martingale (Samuelson, 1965). If owners of an exhaustible resource like petroleum extract it at optimal

rates, then in some settings the price of reserves will appreciate according to a martingale with drift equal to the interest rate. Finally, the expected present value relation has implications for the volatility of asset prices. Informally, the expected present value relation implies that stock prices are like a moving average of the dividend stream to which they give title. Since a moving average is smoother than its components, it follows that stock prices should show less volatility than dividends. Volatility tests along these lines were originally reported by Shiller (1981) and LeRoy and Porter (1981). A number of subsequent papers extended and criticized the finding of excess volatility.

Equation (10), which requires that the conditional expectation of the rate of return does not depend on the value taken on by the conditioning variables, is very restrictive. Unlike its certainty analogue (1), which reflects only the assumption of zero transactions costs, (10) constitutes a strong restriction on the equilibrium probability distribution of the endogenously determined stock prices – much stronger than anything implied by the idea of capital market efficiency alone. The question becomes: what restrictions on preferences and the production technology are needed to derive (10)? LeRoy (1973) showed that, if agents are risk neutral, then in an exchange economy (10) will be satisfied (see also LeRoy, 1982, for discussion in a more general setting). The result is a consequence of the obvious fact that if agents are risk neutral they will ignore moments in the distribution of rates of return higher than the first. Under non-zero risk aversion, however, the conditional expected rate of return will contain a risk premium which generally depends on the realizations of the conditioning variables. Hence (10) will generally not be true. LeRoy (1973) and Lucas (1978) discussed a class of models in which the expected present value property fails except as a special case.

If the assumption of risk neutrality is relaxed, the valuation equations must be changed. This can be achieved either by modifying the characterization of expected cash flows or by respecifying the discount factor. Modifying the characterization of expected cash flows involves distorting (relative to natural probabilities) the probability measure used to take expectations so as to put greater (lesser) weight on states in which agents have high (low) marginal utility. Such distorted probabilities always exist in the finite case, and exist under weak assumptions generally. When these ‘risk-neutral probabilities’ are used to compute expectations, security prices equal expected payoffs discounted at the interest rate, just as under risk neutrality (hence the name).

Alternatively, one can retain the natural probabilities but adjust the discount factor to allow for risk aversion. Under the capital asset pricing model, the risk premium on any security or portfolio is proportional to a beta coefficient, which equals the regression coefficient of the security's return on that of the market portfolio. The constant of proportionality is the risk premium on the market portfolio. The idea is that risk-averse agents require high expected returns on high-beta securities since such securities increase portfolio risk on the margin.

Rational bubbles

To return to the certainty case, if the rate of return on an asset is constant at ρ but the convergence condition

$$\lim_{n \to \infty} (1 + \rho)^{-n} p_{t+n} = 0$$
(16)

is not satisfied, then asset prices are characterized by a rational speculative bubble. For a non-technical introduction to rational bubbles, see LeRoy (2004). The asset's price is higher than the present value of the stream of dividends the asset is title to, but nonetheless investors are willing to hold the asset because its price is expected to rise in the future. The definition of speculative bubbles under uncertainty is analogous (whether speculative bubbles exist or not has nothing directly to do with uncertainty). If speculative bubbles can occur, the present value eq. (6) must be generalized to

$$p_t = \sum_{i=1}^{\infty} (1 + \rho)^{-i} d_{t+i} + \gamma (1 + \rho)^{t},$$
(17)

where γ is an arbitrary non-negative constant capturing the magnitude of the speculative bubble. Equation (17) is the class of solutions to the difference equation

$$\rho = \frac{(d_{t+1} + p_{t+1})}{p_t} - 1,$$
(18)

where γ is the constant of integration ($\gamma \geq 0$ results from the requirement that asset prices be always non-negative, a consequence of free disposal). In the special case $\gamma = 0$ speculative bubbles are absent and the present value relation results.
Bubbles cannot occur on a security that has a payoff only at one date, such as a zero-coupon bond. By induction, the same is true of securities that have payoffs at a finite number of dates. Existence of a bubble on such assets would imply an arbitrage opportunity: investors could sell the security short and purchase claims for its payoff at a cost equal to the present value of those payoffs. In the case of securities with payoffs at an infinite number of dates, it may or may not be possible to rule out bubbles on theoretical grounds. A few of the many papers dealing with this question are Tirole (1985), Gilles and LeRoy (1992a; 1992b), Santos and Woodford (1997) and Huang and Werner (2000).

See Also

- bubbles
- excess volatility tests
- finance
- martingales
- speculative bubbles

Bibliography

Fama, E. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25, 383–417.

Gilles, C. and LeRoy, S. 1992a. Asset price bubbles. In *The New Palgrave Dictionary of Money and Finance*, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Gilles, C. and LeRoy, S. 1992b. Bubbles and charges. *International Economic Review* 33, 323–39.

Huang, K. and Werner, J. 2000. Asset price bubbles in Arrow–Debreu and sequential equilibrium. *Economic Theory* 15, 253–78.

LeRoy, S. 1973. Risk aversion and the martingale model of stock prices. *International Economic Review* 14, 436–46.

LeRoy, S. 1982. Expectations models: a survey of theory. *Journal of Finance* 37, 185–215.

LeRoy, S. 1989. Efficient capital markets and martingales. *Journal of Economic Literature* 27, 1583–621.

LeRoy, S. 2004. Rational exuberance. *Journal of Economic Literature* 42, 783–804.

LeRoy, S. and Porter, R. 1981. The present value relation: tests based on implied variance bounds. *Econometrica* 49, 555–74.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 41–9.

Samuelson, P. 1970. The fundamental approximation theorem of portfolio analysis in terms of mean, variances and higher moments. *Review of Economic Studies* 37, 537–41.

Santos, M. and Woodford, M. 1997. Rational asset pricing bubbles. *Econometrica* 65, 19–57.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71, 421–36.

Tirole, J. 1985. Asset bubbles and overlapping generations. *Econometrica* 53, 1499–528.

How to cite this article

LeRoy, Stephen F. "present value." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000158> doi:10.1057/9780230226203.1334

close print preview

The New Palgrave Dictionary of Economics Online

pretesting

Jan R. Magnus
From The New Palgrave Dictionary of Economics Online, , 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article briefly discusses the meaning and dangers of pretesting in estimation procedures. It outlines the proof of the equivalence theorem, and compares the pretest estimator with three other estimators: the ‘usual’ estimator, the ‘silly’ estimator and the ‘Laplace’ estimator.

Keywords

estimation; model selection; pretesting

Article

1 Model selection versus estimation

Suppose data are generated by a linear relationship

$$y = X\beta + \gamma z + u, \quad u \sim \mathrm{N}(0, \sigma^2 I_n), \tag{1}$$

where X is an $n\times k$ matrix of explanatory variables and z is an additional $n\times 1$ vector of explanatory variables. In our role as investigator we do not know this relationship. Our interest is in the effect of X on y, that is, we want to estimate β. Since we don't know that y is generated by (1), we formulate a model that will serve as a vehicle to estimate β. Let us assume that we know that the relationship is linear, that X is certainly in the model, and that z is perhaps in the model. For simplicity we assume also that σ^2 is known. Thus our ‘model space’ consists of only two models: the unrestricted model (where $\gamma \neq 0$) and the restricted model (where $\gamma = 0$).
Our interest could be in finding the ‘true’ model, in which case we are concerned with *model selection*. In that case we should select the unrestricted model, however small γ turns out to be. Our interest, however, is in the *estimation* of β and the model is not of interest *per se* – it is only a means towards our goal. Even if we knew that γ is nonzero, this would not necessarily mean that we should include z in our regression. This is because, if γ is close to zero, a small bias in the estimates of β will result if we use the restricted model, but their variances may increase substantially, and hence the mean squared error will also increase substantially. (Recall that the bias depends on the value of γ but the variance does not.) So even if we know the truth, it is typically wise to simplify for the purposes of estimation.

2 What is pretesting?

The ordinary least squares (OLS) estimator for β in the restricted model is of course

$$b_r = (X'X)^{-1}X'y. \tag{2}$$

If we define

$$M = I_n - X(X'X)^{-1}X', \quad q = \frac{\sigma}{\sqrt{z'Mz}}(X'X)^{-1}X'z, \quad \theta = \frac{\gamma}{\sigma/\sqrt{z'Mz}},$$

then we can write the OLS estimators for β and γ in the unrestricted model as

$$b_u = b_r - \hat{\theta}q, \quad \hat{\gamma} = \frac{z'My}{z'Mz}, \tag{3}$$

where

$$\hat{\theta} = \frac{\hat{\gamma}}{\sigma/\sqrt{z'Mz}} = \frac{z'My}{\sigma\sqrt{z'Mz}} \sim \mathrm{N}(\theta, 1) \tag{4}$$

denotes the t-ratio, which is normally distributed in this case because σ^2 is assumed known. We call θ the *theoretical* t-ratio.
Since we don't know which of the two models we should use in order to estimate β, the typical econometric practice is to perform a preliminary test (pretest) on γ, and to include z in our regression if the t-ratio $\hat{\theta}$ is ‘large’ and exclude it if $\hat{\theta}$ is ‘small’. This leads to the so-called *pretest estimator*

$$b = \begin{cases} b_r & \text{if } |\hat{\theta}| \le c, \\ b_u & \text{if } |\hat{\theta}| > c, \end{cases} \tag{5}$$

where c is some positive number such as 1.96. We can also write (5) as

$$b = \lambda b_u + (1-\lambda) b_r, \quad \lambda = \begin{cases} 0 & \text{if} \quad |\hat{\theta}| \le c, \\ 1 & \text{if} \quad |\hat{\theta}| > c, \end{cases} \tag{6}$$

which emphasizes that the pretest estimator is a weighted average of the estimators in the available models. The weights, however, are random variables because they depend on $\hat{\theta}$. The pretest estimator is therefore a complicated nonlinear estimator.
The problem with pretesting is not so much that people do it, but that they ignore the consequences. In typical econometric practice, model selection takes place using t-ratios and other diagnostics, after which a single model is selected (stage 1). Then estimates and standard errors are obtained in the selected model (stage 2), and these are reported. It is then tacitly assumed that the reported estimates are unbiased and that their standard errors are given by the usual OLS formulae. This assumption, however, is incorrect. The estimates are biased and their standard errors are not given by the usual OLS formulae. This is the pretest problem.

3 The equivalence theorem

Things are made simpler by the equivalence theorem, originally proved by Magnus and Durbin (1999), and improved and extended by Danilov and Magnus (2004a).
Theorem 1: **(Equivalence theorem):** Let $b = \lambda b_u + (1-\lambda) b_r$, where $0 \le \lambda \le 1$ and $\lambda = \lambda(My)$.
Then, letting $\tilde{\theta} = \lambda\hat{\theta}$, we have

$$E(b) = \beta - E(\tilde{\theta} - \theta) q, \quad var(b) = \sigma^2 (X'X)^{-1} + var(\tilde{\theta}) qq'$$

and hence

$$MSE(b) = \sigma^2 (X'X)^{-1} + MSE(\tilde{\theta}) qq'.$$

Proof: We know from (3) that $b_u = b_r - \hat{\theta} q$, so that

$$b = \lambda b_u + (1-\lambda) b_r = b_r - \lambda\hat{\theta} q = b_r - \tilde{\theta} q.$$

The crucial ingredient is that b_r and My are independent, so that

$$E(b_r | My) = E(b_r), \quad var(b_r | My) = var(b_r).$$

Also, since both λ (by assumption) and $\hat{\theta}$ as given in (4) depend only on My, we see that $\tilde{\theta} = \lambda\hat{\theta}$ depends only on My. Hence,

$$E(b|My) = E(b_r) - E(\tilde{\theta}|My) q = \beta + \theta q - \tilde{\theta} q = \beta - (\tilde{\theta} - \theta) q$$

and

$$var(b|My) = var(b_r|My) = var(b_r) = \sigma^2 (X'X)^{-1}.$$

Now using the well-known relationships between conditional and unconditional moments, we obtain

$$E(b) = E(E(b|My)) = \beta - E(\tilde{\theta} - \theta) q,$$

and

$$var(b) = E(var(b|My)) + var(E(b|My)) = \sigma^2 (X'X)^{-1} + var(\tilde{\theta}) qq',$$

and hence

$$MSE(b) = var(b) + E(b-\beta)(b-\beta)' = \sigma^2 (X'X)^{-1} + var(\tilde{\theta}) qq' + (E(\tilde{\theta} - \theta))^2 qq' = \sigma^2 (X'X)^{-1} + MSE(\tilde{\theta}) qq'.$$

This completes the proof. ||
The equivalence theorem is important because it tells us that if we have a ‘good’ estimator for θ, say $\ddot{\theta}$, then this defines $\lambda = \tilde{\theta} / \hat{\theta}$ and *the same* λ will provide a good estimator for β, namely $b = \lambda b_u + (1-\lambda) b_r$. The pretest estimator chooses

$$\tilde{\theta} = \begin{cases} 0 & \text{if} \quad |\hat{\theta}| \le c, \\ \hat{\theta} & \text{if} \quad |\hat{\theta}| > c, \end{cases}$$

which is not a good choice as we shall see.

4 Moments of the pretest estimator

In the previous section we have seen that the pretest estimator is, in essence, of the form

$$t(x) = \begin{cases} 0 & \text{if} \quad |x| \le c, \\ x & \text{if} \quad |x| > c, \end{cases} \tag{7}$$

where $x \sim N(\theta, 1)$. When studying this estimator, we confront it with three other estimators: the 'usual' estimator $t(x) = x$, the 'silly' estimator $t(x) = 0$, and the 'Laplace' estimator introduced in Magnus (2002). The four estimators are graphed in Figure 1 for $|x| < 4$.

Figure 1
Four estimators $t(x)$ of θ.

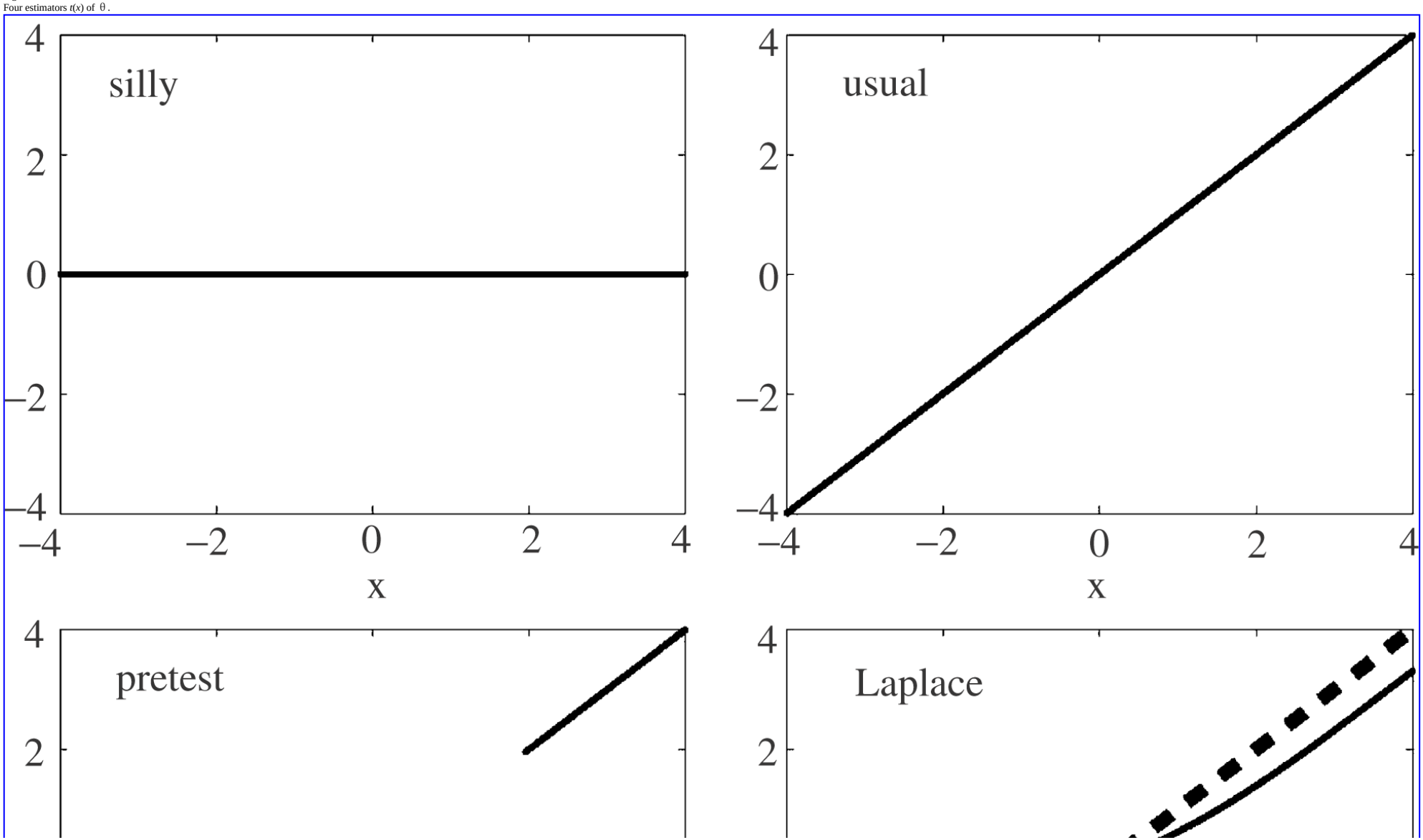

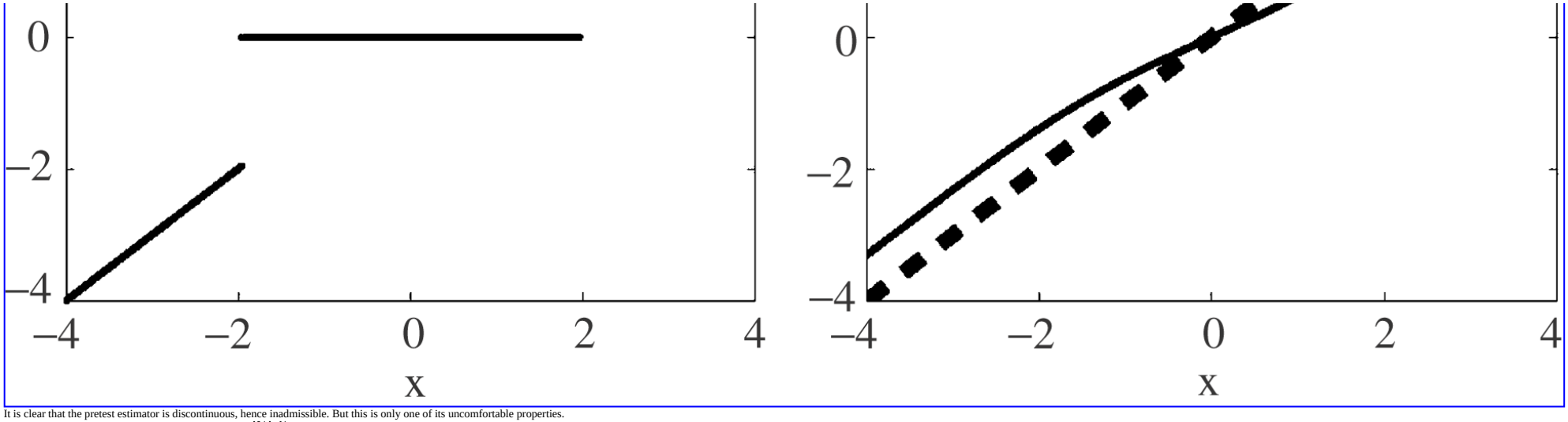

It is clear that the pretest estimator is discontinuous, hence inadmissible. But this is only one of its uncomfortable properties.

Theorem 2: **(Moments of pretest estimator):** Let $X \sim N(\theta, 1)$ and let $t(x)$ be the pretest estimator defined in (7). Then,

$$E(t-\theta) = \phi(c-\theta) - \phi(c+\theta) - \theta P$$

and

$$E(t-\theta)^2 = 1 + (c+\theta)\phi(c+\theta) + (c-\theta)\phi(c-\theta) + (\theta^2 - 1)P,$$

where φ denotes the standard-normal density and $P = \int_{-\theta-c}^{-\theta+c} \phi(u)\,du$.

Proof: Letting $S = \{u: -\theta - c < u < -\theta + c\}$, we have

$$E(t(x)) = \int_{-\infty}^{\infty} t(x)\phi(x-\theta)\,dx = \int_{|x|>c} x\phi(x-\theta)\,dx = \theta - \int_{|x|<c} x\phi(x-\theta)\,dx = \theta - \int_S (u+\theta)\phi(u)\,du = \theta - \int_S u\phi(u)\,du - \theta\int_S \phi(u)\,du = \theta + [\phi(u)]_S - \theta P = \theta + \phi(-\theta+c) - \phi(-\theta-c) - \theta P,$$

using the fact that $\phi'(u) = -u\phi(u)$. Similarly, using the fact that $\phi''(u) = (u^2 - 1)\phi(u)$, we obtain the second result. ||

The bias, standard error and root mean squared error of the pretest estimator are graphed in Figure 2.

Figure 2

Moments of the pretest estimator.

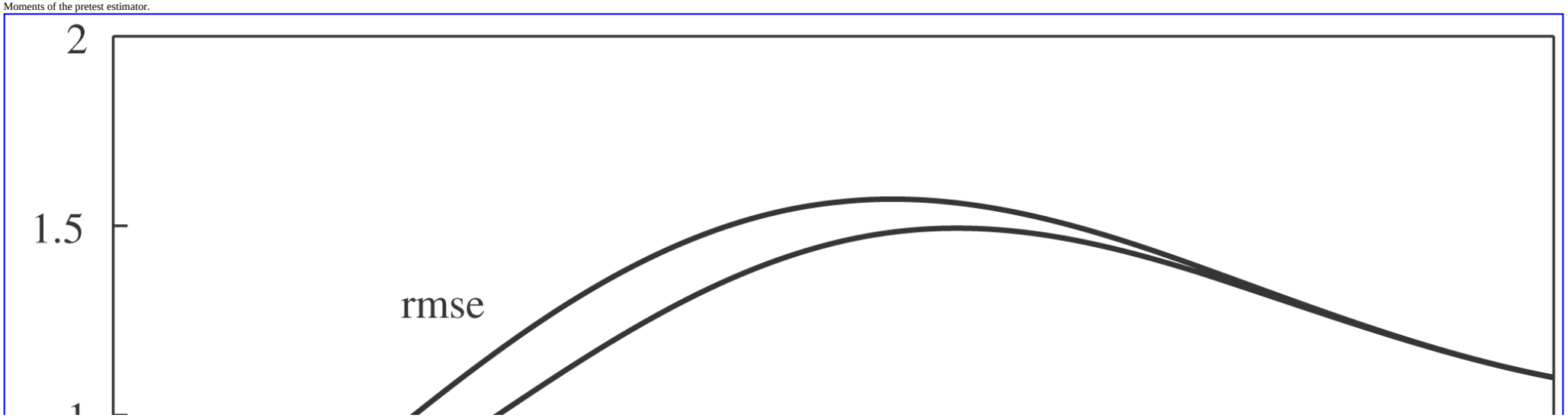

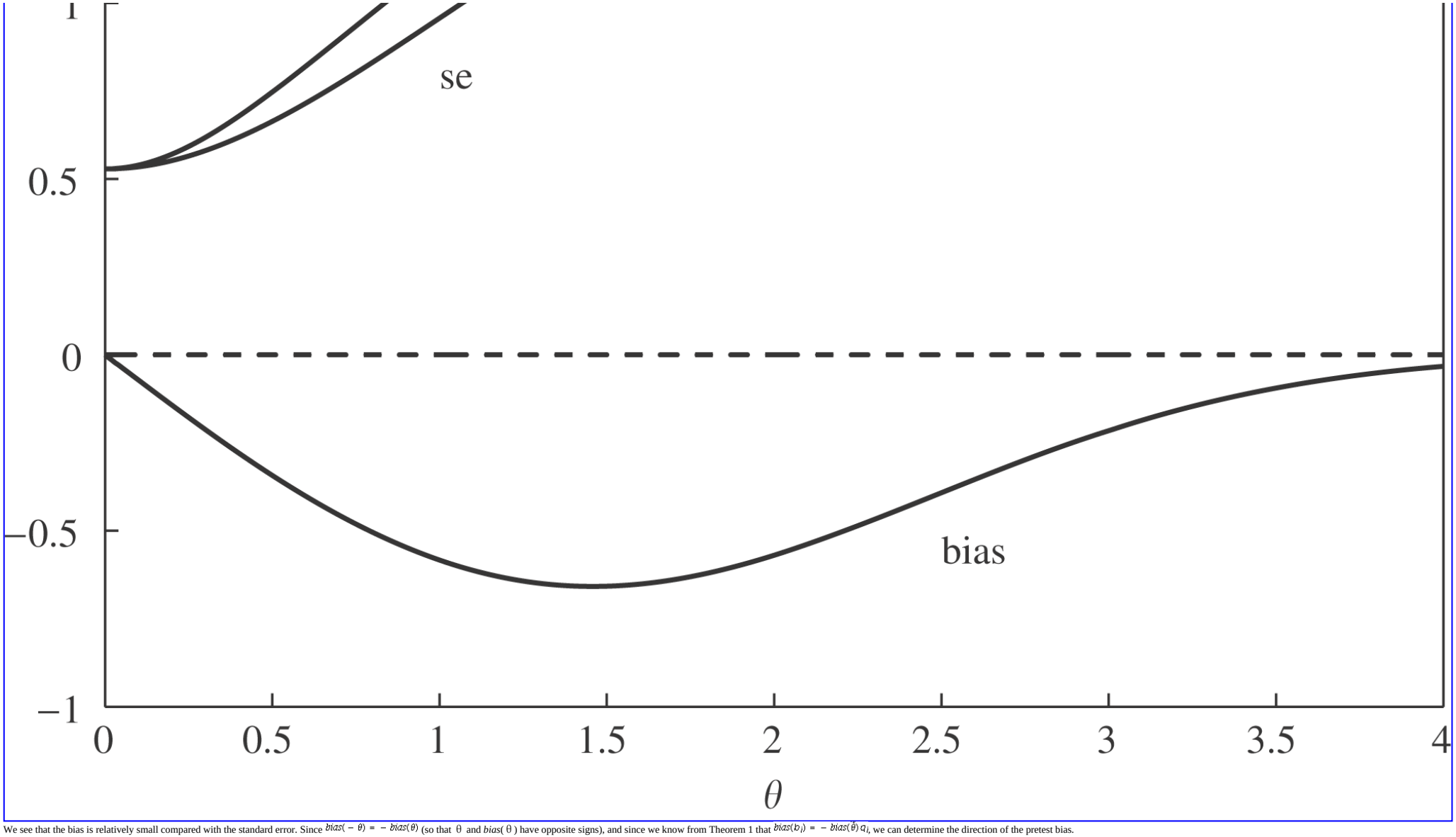

We see that the bias is relatively small compared with the standard error. Since $bias(-\theta) = -bias(\theta)$ (so that θ and $bias(\theta)$ have opposite signs), and since we know from Theorem 1 that $bias(b_i) = -bias(\hat{\theta})q_i$, we can determine the direction of the pretest bias.

Theorem 3: **(Sign of pretest bias):** Let $w_i = (X'X)^{-1}X'z_i$ with components w_{ij} ($j = 1, \ldots, k$). Then the pretest bias of b_i is positive (that is, $E(b_i) > \beta_i$) if and only if $Y_i w_i > 0$. As a consequence we can estimate the sign of the pretest bias of b_i by $sign(w_i \hat{Y}_i)$.

For purposes of exposition we have concentrated on the simplest case, but considerable generalization is possible to more than one additional z-variable, to unknown σ^2, and to general variance matrix.

5 Alternatives

We now compare the pretest estimator with the four estimators in Figure 1. We graph the root mean squared error (RMSE) of each of the four estimators in Figure 3.

Figure 3

Root mean squared error of the four estimators.

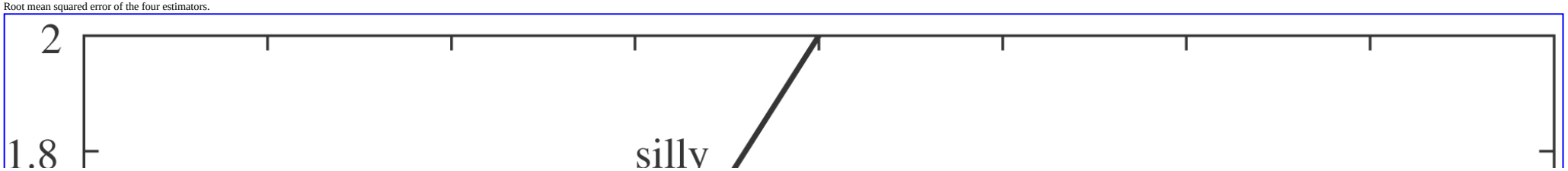

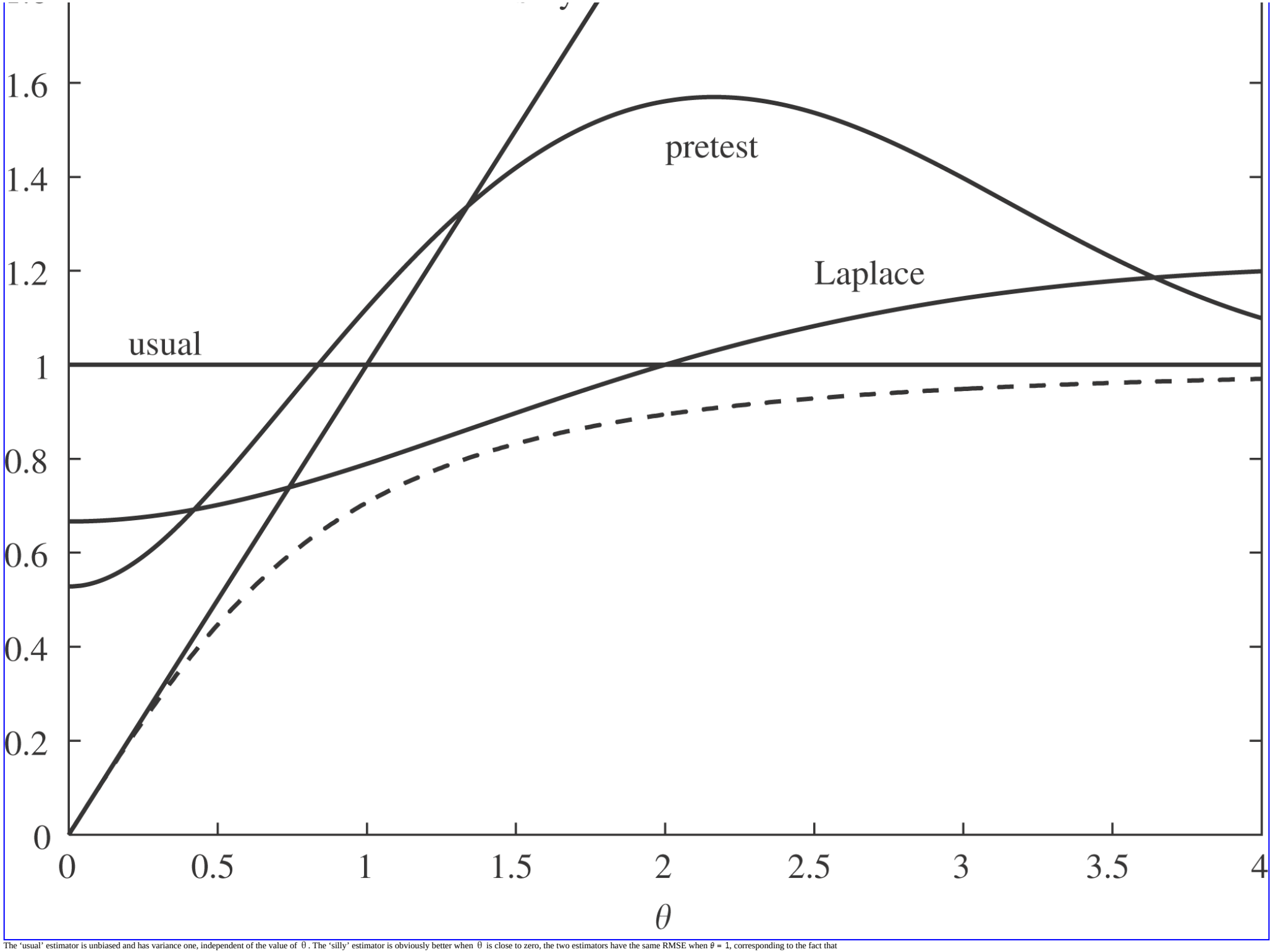

The ‘usual’ estimator is unbiased and has variance one, independent of the value of θ. The ‘silly’ estimator is obviously better when θ is close to zero, the two estimators have the same RMSE when $\theta = 1$, corresponding to the fact that

$$MSE(b_r) - MSE(b_u) = (\theta^2 - 1)qq',$$

but the RMSE of the 'silly' estimator is unbounded. The pretest estimator lies in-between the silly and the usual estimator, except in the important interval around $\theta = 1$ where the pretest estimator is *worse* rather than better than either of the two naive alternatives. This is a most unwelcome property of the pretest estimator, and it has given rise to thought about alternatives. An attractive alternative is the so-called Laplace estimator, which has a Bayesian and a non-Bayesian interpretation, is admissible, is based on a 'neutral' prior, and has good properties around $\theta = 1$. The dotted line $|\theta| / \sqrt{1 + \theta^2}$ denotes the theoretical lower bound of the root mean squared error.

6 History

The implications of model selection on the estimators of the parameters of interest were already being discussed following Tinbergen's (1939) study for the League of Nations. Both Keynes (1939) and Friedman (1940), in their respective critiques on Tinbergen, focused on the method of model selection when the estimation procedure repeatedly uses the same data to discriminate between plausible competing theories. The same point was made in Haavelmo (1944, Section 17). Koopmans (1949) suggested that a completely new theory of inference was required to solve the dilemmas implied by the model selection problem.
Early work on the pretest estimator includes Bancroft (1944, 1964), Huntsberger (1955), Larson and Bancroft (1963), Cohen (1965), Wallace and Ashtar (1972), Sclove, Morris and Radhakrishnan (1972), Bock, Yancey and Judge (1973), and Bock, Judge and Yancey (1973). The harm of ignoring the effects of pretesting was analysed by Danilov and Magnus (2004a, 2004b). Important surveys are provided by Judge and Bock (1978, 1983), Judge and Yancey (1986), Giles and Giles (1993), and Magnus (1999).

Acknowledgements

I am grateful to Chris Müris for preparing the figures.

See Also

- model selection
- robust estimators in econometrics
- semiparametric estimation
- testing

Bibliography

Bancroft, T.A. 1944. On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics* 15, 190–204.

Bancroft, T.A. 1964. Analysis and inference for incompletely specified models involving the use of preliminary tests of significance. *Biometrics* 20, 427–42.

Bock, M.E., Judge, G.G. and Yancey, T.A. 1973. Some comments on estimation in regression after preliminary tests of significance. *Journal of Econometrics* 1, 191–200.

Bock, M.E., Yancey, T.A. and Judge, G.G. 1973. The statistical consequences of preliminary test estimators in regression. *Journal of the American Statistical Association* 68, 109–16.

Cohen, A. 1965. Estimates of linear combinations of the parameters in the mean vector of a multivariate distribution. *Annals of Mathematical Statistics* 36, 78–87.

Danilov, D. and Magnus, J.R. 2004a. On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122, 27–46.

Danilov, D. and Magnus, J.R. 2004b. Forecast accuracy after pretesting with an application to the stock market. *Journal of Forecasting* 23, 251–74.

Friedman, M. 1940. Review of Jan Tinbergen, Statistical Testing of Business Cycle Theories, II: Business Cycles in the United States of America. *American Economic Review* 30, 657–61.

Giles, J.A. and Giles, D.E.A. 1993. Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys* 7, 145–97.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(supplement), 1–115.

Huntsberger, D.V. 1955. A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics* 26, 734–43.

Judge, G.G. and Bock, M.E. 1978. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.

Judge, G.G. and Bock, M.E. 1983. Biased estimation. In *Handbook of Econometrics*, Vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland, Chapter 10.

Judge, G.G. and Yancey, T.A. 1986. *Improved Methods of Inference in Econometrics*. Amsterdam: North-Holland.

Keynes, J.M. 1939. Professor Tinbergen's method. *Economic Journal* 49, 558–68.

Koopmans, T. 1949. Identification problems in economic model construction. *Econometrica* 17, 125–44.

Larson, H.J. and Bancroft, T.A. 1963. Biases in prediction by regression for certain incompletely specified models. *Biometrika* 50, 391–402.

Magnus, J.R. 1999. The traditional pretest estimator. *Theory of Probability and Its Applications* 44, 293–308.

Magnus, J.R. 2002. Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5, 225–36.

Magnus, J.R. and Durbin, J. 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67, 639–43.

Sclove, S.L., Morris, C. and Radhakrishnan, R. 1972. Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* 43, 1481–90.

Tinbergen, J. 1939. *Statistical Testing of Business Cycle Theories*, 2 vols. Geneva: League of Nations.

Wallace, T.D. and Ashtar, V.G. 1972. Sequential methods in model construction. *Review of Economics and Statistics* 54, 172–8.

How to cite this article

Magnus, Jan R. "pretesting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 14 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_P000372>
doi:10.1057/9780230226203.1887

The New Palgrave Dictionary of Economics Online

price control

John Kenneth Galbraith

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In modern times price control has been used to keep down food prices, as part of prices and incomes policies, in wartime economic management, to help governments win elections, and to tackle inflation. Along with macroeconomic restraint and specific commodity restraint by rationing, price controls used by the Allies in the Second World War succeeded in countering inflation. The ineffectiveness of price control in Latin America has helped give it a bad name. There are radically different forms of controls in greatly differing contexts; price control should be seen as a diversely applicable policy, sometimes advisable, sometimes not.

Keywords

black market; corporations; craft guilds; depressions; fiscal policy; Galbraith, J. K.; incomes and prices policy; inflation; Latin America; monetary policy; price control; rationing; rent control; Second World War; trade unions; wage–price inflation

Article

The fixing of prices by public action is of exceedingly ancient origin; popular economic cliché associates it with the Edict of Diocletian, and economic history dwells at length on the controls exercised and imposed by the medieval guilds. Only in modern times, roughly the last 200 years, has it fallen under the general disapproval and interdict of orthodox economic attitudes and has it been seen therein as a temporary or aberrant departure from free-market principles.

In a more adequate view, controls have not one but several forms, some of which are, in their context, a reflection of necessary and appropriate policy, as other designs in other contexts are not.

Specifically, some five employments of price controls can be identified, apart from public-utility and like regulation which reflects the different and largely accepted need to maintain a public surveillance and restraint on natural or legislated monopoly power. There is:

1. (1) The use of controls to address particular wartime pressures of demand on supply, as in the United States and other countries in World War II, and to keep down the price of food for urban dwellers, as now in African countries and elsewhere. These can perhaps be called episodic or casual controls.
2. (2) The use of controls as part of what has come to be called a prices and incomes policy. They here act on the specific problem of wage/price inflation.
3. (3) The use of controls as part of a comprehensive exercise in wartime economic management, backed by rationing of consumers' goods, allocation of materials and labour and a general restraint on aggregate demand.
4. (4) The use of controls as a highly temporary expedient to get by an election.
5. (5) The use of controls in the face of an enduring inflationary movement propelled by a persisting excess of aggregate demand, as recently in Latin America and of late in Israel.

Two of the above employments of controls – to limit the wage/price dynamic and as an adjunct to a comprehensive mobilization of economic resources, as in World War II – have modern policy relevance. In the highly organized modern economy of strong corporations and viable and effective trade unions, price inflation can come from the microeconomic effect of prices and living costs on wages and of wage demands on prices. Much recent experience shows that this wage/price dynamic can be arrested by conventional monetary or fiscal action only by the restraining force of substantial unemployment on wage demands and much idled plant capacity on prices. In other words, conventional monetary and fiscal policy arrest wage/price inflation only as they cause a recession or depression.

Accordingly, attention has focused on direct restraint by the state. Avoiding the unduly blunt reference to wage and price controls, this has come to be called 'an incomes and prices policy'. Austria, Germany, Japan and other industrial countries have, formally or informally, resorted extensively to such restraints. The English-speaking countries and their economists, businessmen and unions have been more reluctant. Market forces must not be impaired. Still, by a growing minority such restraints are viewed as a necessary alternative to economically and politically more painful designs for restraining wage/price inflation. There continue to be repetitive suggestions that such intervention distorts the market allocation of resources. Mention is not made of the way that strong unions and strong corporations in the modern economy have already invaded resource-allocation procedure and accommodated it to their purposes.

A more serious problem lies outside the field of economics. Where fiscal and monetary restraints need only a negligible administrative apparatus, any effective form of price and wage control requires a substantial administrative one. And, of course, the public intervention to limit price or wage increases is highly visible. Thus it invokes the ever-present suspicion or dislike of government intervention and bureaucracy. Fiscal and monetary action, even when more painful in overall effect, encounters far less resistance.

The second acceptable form of controls was the comprehensive design used in all of the industrial countries in World War II. In combination with macroeconomic restraint on demand by fiscal policy and specific commodity restraint by rationing, such controls successfully countered the threat of price inflation in Britain, the United States, Canada and other participants in those years. In the case of non-rationed non-essentials, the controls substituted shortage or non-availability for rationing by price. Some evasion of controls by black market operations was present, but this, though greatly publicized and deplored, was, in general, relatively limited.

Once controls were fully in place in the United States, price increases were nominal, and, overall, there is no memory of inflation from the war years. Increases then or following the removal of controls were insignificant as compared with the double-digit inflation, as it was called, of the 1970s.

The circumstances, especially in the United States during World War II, were, however, exceptionally and perhaps uniquely favourable for successful use of price controls. After ten years, depression had come to be considered by the early 1940s a normal and inevitable peacetime phenomenon. Accordingly, after the war, unemployment and associated hardships would recur. From this came a powerful incentive to save – to save, among other reasons, for the cars and others durables that would only later be available. Labour, in effect, was employed against the promise of future consumption. At the same time there was in the United States the large increase in the supply of non-durables as previously unemployed plant and labour were drawn into production. Overall civilian consumption increased, and this further reduced the pressure of demand on the controls. A similar *general* use of controls following a period of high employment and serious or even incipient inflation with associated expectations would be a far more difficult matter.

Of the other uses of controls there is less to be said. Isolated or piecemeal controls (in contrast with a broad-based incomes and prices policy) can, indeed, have the effect of diverting resources from the area of control and into uncontrolled and thus more remunerative employments. This has been a consequence of one of the more persistent manifestations of isolated or piecemeal controls, that of rents. Its yet more serious manifestation has been in the poor countries, notably of Africa. There the use of price controls to keep down food costs has been an important contributing cause of the food distress and famine.

Controls in the face of a massive excess of demand have been a frequent resort in Latin America. A case can be made for such action to alter expectations, themselves a cause of inflationary pressures of demand, before instituting strong macroeconomic restraints. More frequently, such controls have been a separate and often desperate response to demand-induced inflation. The resulting evasion, ineffectiveness and eventual collapse have contributed notably to the poor reputation of controls in general.

In 1971–3, President Richard Nixon used general controls with great effect to suppress wage–price inflation and allow of companion fiscal and monetary support to employment. Largely, if not principally, in consequence, he carried every state but Massachusetts and the District of Columbia in the election of 1972. Such success for controls must, however, be accounted for and judged in the field of politics, not economics. The removal of the controls after the election restored with some precision the circumstances that had led to their being involved with a strong recurrence of inflation.

A common tendency of orthodox economics has been to deal with all forms of price control as a homogeneous exercise. This, it will be evident, is a grave oversimplification. In fact, there are radically different forms of controls in greatly differing contexts. Reasonable and, indeed, necessary sophistication requires that these differences be recognized, that price control be seen as a diversely applicable policy, sometimes greatly advised, sometimes wholly the reverse.

See Also

- command economy

Bibliography

Galbraith, J.K. 1952. *A Theory of Price Control*. Cambridge, MA: Harvard University Press.

Keynes, J.M. 1940. *How to Pay for the War*. London: Macmillan; New York: Harcourt, Brace and World.

Mitchell, H. 1947. The Edict of Diocletian: a study of price fixing in the Roman empire. *Canadian Journal of Economics and Political Science* 13, 1–12.

Rockoff, H. 1984. *Drastic Measures: A History of Wage and Price Controls in the United States*. Cambridge: Cambridge University Press.

Taussig, F.W. 1919. Price-fixing as seen by a price-fixer. *Quarterly Journal of Economics* 33, February, 205–41.

How to cite this article

Galbraith, John Kenneth. "price control." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000163> doi:10.1057/9780230226203.1335

The New Palgrave Dictionary of Economics Online

price discrimination (empirical studies)

Frank Verboven

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Price discrimination occurs when the prices of similar products sold by the same firm show variation that cannot be attributed to cost variation. Recent empirical work has identified the presence of both direct and indirect price discrimination, after cost-based explanations have been accounted for. Furthermore, there is increasing evidence on the sources of price discrimination. The extent of price discrimination has often been found to increase as competition intensifies, in contrast to conventional wisdom but consistent with new theoretical insights. Finally, various empirical studies have considered the effects of price discrimination on profits, consumer welfare and efficiency.

Keywords

competition and price discrimination; cross-price elasticities; Gini coefficient; monopoly price discrimination; multi-product firms; poaching; price discrimination; price elasticity; single-product firms

Article

Price discrimination occurs when the prices of similar products sold by the same firm show variation that cannot be attributed to variation in marginal costs. Direct (or third-degree) price discrimination serves to exploit observed differences in consumer characteristics; indirect (or second-degree) price discrimination exploits unobservable consumer heterogeneity. While price discrimination has been studied extensively by economic theorists, and illustrated with numerous textbook examples (for example, Scherer and Ross, 1990), it has only recently become an area of rigorous empirical research. Empirical studies have focused on several questions: (*a*) the measurement or identification of price discrimination; (*b*) the sources of price discrimination, notably the role of competition; and (*c*) the effects of price discrimination on profits, consumer welfare and efficiency.

Measurement of price discrimination

The identification of price discrimination can be introduced in a simple framework in which a firm sells two products. The price difference Δp between the two products (assumed positive) can be decomposed in a cost difference Δc and a margin difference Δm, so $\Delta p = \Delta c + \Delta m$. Price discrimination exists to the extent that the observed price difference Δp is due to the margin difference Δm rather than a possible cost difference Δc. (An alternative definition is based on percentage rather than absolute margin differences. To consider this, reinterpret the variables in logs, Clerides, 2004, compares the two approaches in empirical studies.) Identifying margin differences from cost differences is not an obvious task. Lott and Roberts (1991) provide plausible cost-based explanations for commonly viewed price discrimination cases. Several recent empirical studies have attempted to properly account for cost differences before drawing conclusions about the presence of price discrimination.

There have been two methodological approaches. The first approach uses direct cost information. Sometimes the cost difference can be derived from industry information about the production technology. An early example is Benston (1964), who finds that 76 per cent of the higher interest rates charged to small businesses can be attributed to additional costs. In contrast, Clerides (2002) attributes only five per cent of the average price difference between hardback and softback books to higher production costs. In other cases, the production technology is not known, but the cost difference Δc is reasonably assumed to be zero or negative, so that the observed positive price difference Δp provides a lower bound for the extent of price discrimination. Graddy (1995) finds that Asians pay seven per cent less at a fish market, while there are no reasons to believe that these customers have lower servicing costs. Degryse and Ongena (2005) find that bank customers pay lower interest rates as their distance from the bank increases, whereas costs, if anything, are expected to be increasing in distance. Shepard (1991) provides a neat variation on this theme. As in the above framework, she observes the price difference Δp between a high-quality and a low-quality product sold by multi-product firms (full service and self-service at gas stations). In addition, she essentially also observes the analogue price difference Δp^S for single-product firms (selling either full-service or self-service). She defines the extent of price discrimination as the difference between the markup difference for multi-product firms Δm and that of single-product firms Δm^S. Because her qualitative evidence indicates that the cost difference Δc between the high-quality and low-quality product for multi-product firms is no larger than the corresponding cost difference Δc^S for single-product firms, the difference between Δp and Δp^S provides a lower bound for the extent of price discrimination. She finds that the extent of price discrimination for full-service versus self-service gasoline amounts to at least nine cents a gallon.

The second approach to identifying price discrimination does not use cost information, but instead infers the price–cost margin difference Δm from a model of pricing behaviour. This approach essentially replaces cost-side information by demand-side information. For example, Verboven (2002) finds evidence of indirect price discrimination between high-mileage and low-mileage drivers. He uses information on the relative popularity of high-quality and low-quality products (diesel and gasoline cars) and the distribution in consumers' willingness to pay for quality (mileage). He infers that 75–90 per cent of the price premium for the high-quality products can be attributed to a higher margin, a finding that is confirmed by direct cost information.

Sources of price discrimination

Several empirical studies have gone beyond the basic question of identifying price discrimination to uncover its sources, in particular the role of competition. Theoretical work has revealed that competition does not necessarily reduce the incentives to price discriminate. The extent of direct price discrimination depends on both the price elasticity of market demand and the cross-price elasticities with respect to competing products; it is therefore not necessarily smaller under competition. For example, Borenstein (1991) looks at price discrimination in the competitive retail gasoline market. Margins on unleaded gas were initially higher than margins on leaded gas. The decline in the number of competing stations offering leaded gas caused an increase in the margins on leaded gas relative to the margins on unleaded gas, hence a reduction in price discrimination. This illustrates that competition can be a source of price discrimination: stations take into account the buyers' possibilities to substitute to competing stations when setting their prices. Borenstein and Rose (1994) take up a similar question for the US airline industry. Since they observe more than two prices on a given airline/route, they use the Gini coefficient as a summary measure of price dispersion (rather than the price difference Δp for every product pair). They find that the expected price difference for two randomly selected passengers on a given airline/route is 36 per cent of the average ticket price. An increase in the number of competitors raises the extent of price dispersion by a large amount. Goldberg and Verboven (2001) measure margins based on the estimated own- and cross-price elasticities. They find that car manufacturers earn higher margins in their domestic markets than in their foreign markets, because markets are segmented according to country of origin and there is more competition in the foreign segments. Asplund, Eriksson and Strand (2002) find that newspaper subscriptions in Sweden are more often sold at (often introductory) discounts in duopoly regions than in monopoly regions. They interpret this as evidence of poaching, that is, discrimination to attract new customers from rival firms.

The existence of indirect price discrimination is not obvious under competition, as shown in theoretical work. Nevertheless, empirical work has documented that competition may strengthen indirect price discrimination. Verboven (1999) finds a significant percentage price premium for optional engine power in the more competitive car segments, and percentage discounts in the less competitive segments (the latter being consistent with a monopoly discrimination). Busse and Rysman (2005) compare the prices of large and small ads in Yellow Pages directories. Their identification strategy relies on the assumption that cost differences between the two types of ads do not depend on the degree of competition. As such, they do not measure the extent of price discrimination per se, but instead ask how it varies with competition. They find that competition raises the discounts to large buyers: adding one competitor lowers the price of small ads by only six per cent, whereas it lowers the price of large ads by 12 per cent.

Economic effects of price discrimination

Several empirical studies have also assessed the economic implications of price discrimination for profits, consumer welfare, tax revenues and economic efficiency. Leslie (2004) considers monopoly price discrimination. He finds that direct price discrimination for a Broadway theatre play, in the form of a currently observed 50 per cent discount at the discount booth known as the TKTS, raises profits five per cent above the profits under a uniform price strategy. However, he also finds that the current 50 per cent discount is too large to maximize profits, thereby generating too much substitution out of the full-

price tickets. Lowering the discount to 30 per cent would raise profits seven per cent above the profits under a uniform price strategy. Leslie also estimates the aggregate consumer welfare effects from price discrimination, and finds them to be relatively small.

Under competition, the effects of direct price discrimination on profits are ambiguous even if the discriminatory prices are chosen optimally. The possibility to discriminate may lead to a situation of all-out competition, in which all discriminatory prices are lower than the uniform prices, thereby reducing profits. This occurs when the weak (elastic) market of one firm is the strong (inelastic) market of the other firm. Nevo and Wolfram (2002) find suggestive evidence of all-out competition, documenting that price discrimination (in the form of coupons) may lower the prices of all products, and may hence lower profits. Besanko, Dubé and Gupta (2003) consider a situation of uniform pricing (for ketchup), and compute the new equilibrium under the assumption that firms would be able to discriminate between three (latent) customer segments. They find that all firms perceive the same customers as weak or strong. Price discrimination thus does not lead to all-out competition; quite the contrary, it increases profits. Brenkers and Verboven (2006) consider the reverse case in which car manufacturers currently discriminate between consumers from different countries, and would no longer be able to do this in the future (because of improved market integration). In their application, all-out competition appears more likely a priori, since domestic and foreign firms have the reverse strong and weak markets. Nevertheless, they find no evidence of all-out competition: an elimination of price discrimination would lower the prices of domestic firms, but raise the prices of foreign firms. Price discrimination correspondingly has relatively modest effects on industry profits and welfare (unless the high prices in the United Kingdom would be due to collusion).

The effects of indirect price discrimination under competition have also received attention recently. Miravete and Röller (2003) find that a single two-part tariff achieves 94 per cent of the potential profits and 63 per cent of potential welfare under a fully nonlinear tariff. McManus (2004) assesses the extent to which coffee shops distort their qualities (cup sizes) from the efficient levels, as a way to segment customers based on willingness to pay for quality. Consistent with economic theory, he finds that there are quality distortions, tending towards zero for the top qualities. Crawford and Shum (2007), using a somewhat different approach, also find evidence of quality degradation in the cable television industry.

See Also

- price discrimination (theory)

Bibliography

Asplund, M., Erikkson, R. and Strand, N. 2002. Price discrimination in oligopoly: evidence from Swedish newspapers. Discussion Paper No. 3269. London: CEPR.

Benston, G. 1964. Commercial bank price discrimination against small loans: an empirical study. *Journal of Finance* 19, 631–43.

Besanko, D., Dubé, J.-P. and Gupta, S. 2003. Competitive price discrimination strategies in a vertical channel using aggregate retail data. *Management Science* 49, 1121–38.

Borenstein, S. 1991. Selling costs and switching costs: explaining retail gasoline margins. *RAND Journal of Economics* 22, 354–69.

Borenstein, S. and Rose, N. 1994. Competition and price dispersion in the U.S. airline industry. *Journal of Political Economy* 102, 653–83.

Busse, M. and Rysman, M. 2005. Competition and price discrimination in Yellow Pages advertising. *RAND Journal of Economics* 36, 378–90.

Brenkers, R. and Verboven, F. 2006. Liberalizing a distribution system: the European car market. *Journal of the European Economic Association* 4, 216–51.

Clerides, S. 2002. Book value: intertemporal pricing and quality discrimination in the U.S. market for books. *International Journal of Industrial Organization* 20, 1358–408.

Clerides, S. 2004. Price discrimination with differentiated products: definition and identification. *Economic Inquiry* 42, 402–12.

Crawford, G. and Shum, M. 2007. Monopoly quality degradation and regulation in cable television. *Journal of Law and Economics* 50, 181–219.

Degryse, H. and Ongena, S. 2005. Distance, lending relationships and competition. *Journal of Finance* 60, 231–6.

Goldberg, P. and Verboven, F. 2001. The evolution of price dispersion in the European car market. *Review of Economic Studies* 68, 811–48.

Graddy, K. 1995. Testing for imperfect competition at the Fulton fish market. *RAND Journal of Economics* 26, 75–2.

Leslie, P. 2004. Price discrimination in Broadway theatre. *RAND Journal of Economics* 35, 520–41.

Lott, J. and Roberts, R. 1991. A guide to the pitfalls of identifying price discrimination. *Economic Inquiry* 29, 14–23.

McManus, B. 2004. Nonlinear pricing in an oligopoly market: the case of specialty coffee. Mimeo. Olin School of Business, Washington University.

Miravete, E. and Röller, L. 2003. Competitive nonlinear pricing in duopoly equilibrium: the early U.S. cellular telephone industry. Discussion Paper No. 4069. London: CEPR.

Nevo, A. and Wolfram, C. 2002. Why do manufacturers issue coupons? An empirical analysis of breakfast cereals. *RAND Journal of Economics* 33, 319–39.

Scherer, F. and Ross, D. 1990. *Industrial Market Structure and Economic Performance*. Boston: Houghton Mifflin Company.

Shepard, A. 1991. Price discrimination and retail configuration. *Journal of Political Economy* 99, 30–53.

Verboven, F. 1999. Brand rivalry and market segmentation, with an application to the pricing of optional engine power on automobiles. *Journal of Industrial Economics* 47, 399–425.

Verboven, F. 2002. Quality-based price discrimination and tax incidence – the market for gasoline and diesel cars in Europe. *RAND Journal of Economics* 33, 275–97.

How to cite this article

Verboven, Frank. "price discrimination (empirical studies)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000328> doi:10.1057/9780230226203.1336

The New Palgrave Dictionary of Economics Online

price discrimination (theory)

Eugenio J. Miravete

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Price discrimination comprises a wide variety of practices aimed at extracting rents from a base of heterogeneous consumers. When consumer types are private information and only their distribution is known to the monopolist, finding the optimal nonlinear tariff involves solving a constrained variational problem that characterizes the optimal markup for each purchase level so that consumers of different types have no incentive to imitate the behaviour of others. Fully separating equilibrium is ensured when the distribution of types fulfills the increasing hazard rate property and individual demands can be unambiguously ranked. Outside this framework, optimal tariffs are difficult to characterize.

Keywords

arbitrage; exclusive agency; incentive compatibility constraints; individual pricing; inverse elasticity rule; market segmentation; mechanism design; Mirrlees, J.; multidimensional variational problem; multiple products; nonlinear pricing; nonlinear taxation; peak-load pricing; Pigou, A.; price discrimination; Ramsey pricing; reservation price; sequential screening

Article

A monopolist price discriminates when he sells two identical units of a good at different prices, either to two different buyers or to the same customer. Two basic elements serve to classify the numerous methods whereby firms price identical units of the product differently: the amount of information available to the seller regarding how different the valuations of consumers are, and the seller's ability to avoid arbitrage. Avoiding arbitrage when firms sell personal services is easy and inexpensive, and thus price discrimination becomes a common practice in such industries. Conversely, in the absence of restrictions on the transferability of commodities, low-valuation customers could certainly benefit from reselling to higher-valuation customers, thus effectively impeding the seller from actually charging two different prices for the product.

Classification of price discrimination practices

Pigou (1922) distinguished between first-, second-, and third-degree price discrimination depending on the amount of information regarding consumers' preferences that is available to the seller. In the case of first-degree price discrimination, the seller observes the actual valuation of each consumer and, provided that individual pricing is feasible, he could ask each consumer for her individual reservation price. Individual pricing is, however, rarely observed in reality, but such a pricing strategy has the theoretical appeal of leading to the efficient competitive outcome, although obviously with a quite different distribution of rents. This efficiency result vanishes when the seller knows only the distribution of consumers' valuations, as in second-degree price discrimination, or when he knows even less – just a signal about consumers' valuations – as in the third-degree price discrimination case.

Market segmentation, either geographical or personal, may serve as a way to avoid arbitrage. Price differentials across countries are likely to be larger than across neighbourhoods of a city as consumers move more freely in the latter case. Thus, the ability to price-discriminate will be partially determined by the importance of consumers' transaction costs in purchasing from different markets. Similarly, the cost of enforcing market segmentation may lead to different pricing schemes. Charging different individuals a different price for a service depending on their location, age, gender or race is far less expensive in terms of monitoring costs than tying prices to the income of each individual. In some circumstances, when third-degree price discrimination is used, location, age, gender, race or any other observable characteristics can be used in an economically efficient (although sometimes morally rotten) way to infer average individual valuations of products and increase profits by extracting a larger share of the consumer surplus of those individuals with higher valuations. Thus, in the third-degree price discrimination case, solving the price discrimination problem comes down to finding the optimal monopoly price in several independent markets. If there were numerous firms instead of a single seller, the well-known inverse elasticity rule should be modified to account for the existence of substitute goods.

More interesting is the case of second-degree price discrimination, when the seller needs to avoid the possibility of transferability of demand among consumers of different valuations. Since only the distribution of valuations is known, and not the valuation of individual consumers, finding the optimal pricing scheme requires one to solve a complex problem where the monopolist attempts to extract as much rent as possible from each consumer while at the same time ensuring that they do not imitate the behaviour of other consumers with lower valuations. In other words, price discrimination becomes a mechanism-design problem where a nonlinear tariff charging a different unit price for each unit sold maximizes the expected profits of the monopolist, while ensuring incentive compatibility.

Technical issues of single-dimensional price discrimination

To solve this problem, consumers' preferences are assumed to be fully described by $U(q, \theta)$ where q represents the amount of good purchased by a consumer of type θ. This single-dimensional index captures the relevant difference in demand of diverse consumers, and leads to non-price-related shifts of individual demands. Type θ remains private information for each consumer while the monopolist knows only its distribution $F(\theta)$ on $\Theta = [\underline{\theta}, \overline{\theta}]$. The variational problem that the monopolist faces

consists of finding the optimal nonlinear tariff function $T(q)$ that maximizes his expected profits with respect to the distribution $F(\theta)$ provided that in their choices consumers are guided to maximize the net utility $U(q, \theta) - T(q)$. A fully separating equilibrium exists whenever individual demands can be ranked unambiguously, $U_{q\theta}(q, \theta) > 0$, and when the distribution of consumer types $F(\theta)$ fulfills the common increasing hazard rate property (these are sufficient, not necessary, conditions). In such a case, the optimal nonlinear tariff $T(q)$ is a concave function leading to quantity discounts that assigns different quantities and payments to consumers of different types. Maskin and Riley (1984) and Mussa and Rosen (1978) (in a framework of quality discrimination) first fully characterize the solution to this canonical version of the price discrimination problem. Contrary to the first-degree price discrimination case, now only the highest consumer type, $\bar{\theta}$, is efficiently priced – the efficiency at the top result – while all other consumers are charged a positive markup that induces them to self-select the optimal level of consumption according to the intensity of their preferences, θ. The magnitude of this markup depends on how difficult is to enforce the incentive compatibility condition, which is summarized by the hazard rate of the distribution $F(\theta)$. And the difficulty of separating different consumers depends on how these consumer types are distributed. Thus, the more numerous the consumers with a high valuation are, the larger is the average markup that low-valuation consumers should face in order to minimize the incentive of high-valuation types to purchase a small amount of the good. Intuitively, the more numerous high-valuation consumers are, the more likely some of them will be to attempt to behave as low-valuation consumers. To prevent it, a higher markup charged to low-valuation consumers is needed in order to reduce sufficiently the outside option of those more numerous high-valuation consumers. Consequently, if all consumers are alike, the distribution of consumer types, $F(\theta)$, becomes degenerate, and the optimal nonlinear tariff is a two-part tariff with a slope equal to the marginal cost of production and a fixed fee equal to the individual consumer surplus of a buyer.

Engineers (Dupuit, 1849; Hadley, 1885) rather than economists discovered long ago the advantages of charging different prices to different customers in order to cover the fixed costs of operating transportation services. The solution to the second-degree price discrimination problem described above attracted the attention of economists only after the contribution of Mirrlees (1971) in the area of nonlinear taxation. His approach to finding the optimal tax that maximized a social welfare function could easily be adapted to analyse the Ramsey pricing problem of regulated industries contemplated by Ramsey (1927) and Boiteux (1956). With the development of incomplete information games, the nonlinear pricing problem was rapidly reformulated as a direct revelation mechanism (Goldman, Leland and Sibley, 1984; Guesnerie and Laffont, 1984), thus helping to uncover the technical assumptions that ensured well-behaved solutions of the canonical single-product, single-parameters case.

Extensions of monopoly pricing

The solution to this canonical price-discrimination problem serves as a point of departure for many extensions that have attempted to incorporate either a more general theoretical approach or particular features of specific industries where nonlinear pricing is used to cover fixed costs or to fulfill distributional objectives set by regulators.

A first extension included the possibility that income effects were non-negligible and that consumers could be risk averse. Effectively, this means that the net utility of consumers is not additively separable

in payments. Extensions in this direction includes the work of Mirrlees (1976), Roberts (1979), and Wilson (1993, ch. 7).

Another early extension addressed the rationing of stochastic individual demands in the presence of capacity constraints. The nonlinear tariff now attempts to distribute the cost of installed capacity among consumers according to their usage, as consumers with different loads contribute differently towards the cost of providing the service. But this peak-load pricing solution also provides the firm with incentives to reduce the size of consumers' loads in order to minimize the cost of distributing efficiently the existing capacity among all consumers. Oren, Smith and Wilson (1985) and Panzar and Sibley (1978) are the two basic references on capacity pricing.

More recently, the basic canonical model of price discrimination has been modified to contemplate the possibility of sequential screening, a process common in many industries where consumers first subscribe to one of the many optional tariffs available and later decide on their optimal level of consumption. The canonical model is modified to allow consumers to learn about their valuation of the product, thus distinguishing between *ex ante* and *ex post* types – the valuation of customers before and after contracting with the seller – as well as *ex ante* and *ex post* incentive compatibility constraints. Courty and Li (2000) consider the case where the *ex ante* type determines the distribution from where the *ex post* type will be drawn, while Miravete (1996) considers a framework in which the *ex post* type is the sum of the *ex ante* type and an independently distributed shock. Both approaches lead to ambiguous results that can be somewhat qualified depending on the stochastic dominance of the composition or convolution distribution, respectively, of the *ex post* relative to the *ex ante* valuation. Miravete (2005) further evaluates the welfare performance of nonlinear tariff options using data directly linked to *ex ante* and *ex post* types of consumers.

The most challenging extension of the canonical problem is multidimensional types. Wilson (1996) presents a concise description of the difficulties that arise when types are multidimensional or the monopolist sells several products. Type dimensions capture different features of consumer demands (intercept, curvature, or others) that are independent of prices but are relevant to capturing consumer heterogeneity. Multiple products introduce the possibility of accounting for complementarity and substitution effects and thus designing optimal discounts for bundles that include a variable proportion of each good. The difficulty arises because the multidimensional screening problem imposes a continuum of boundary conditions that translate into a large number (the number of type dimensions minus one) of additional partial differential equations that constrain the multidimensional variational problem. Explicit solutions do not exist beyond particular cases such as those studied by Armstrong (1996), Laffont, Maskin and Rochet (1987), or Wilson (1993, chs. 13, 14). A common result reported by Armstrong (1996) and Rochet and Choné (1998) is that low-valuation customers are always excluded, thus leading to bunching at the bottom.

Competitive nonlinear pricings

Besides the technical difficulties in solving multidimensional price-discrimination problems, numerical solutions show that tariffs may become non-monotone and that even the efficiency at the top result may not hold depending on the support of the distribution of types and the interaction among type dimensions given by the specification of the utility function. Perhaps because of these unsurmountable difficulties,

the generalized one-dimensional nonlinear pricing framework of Rochet and Stole (2002) offers the most promising alternative for advancing in this area of research. Their approach consists of adding a second independent type dimension that enters additively into consumers' utility function. This little modification of the canonical problem disassociates the participation and consumption decisions. While in the canonical problem higher-valuation consumer types always participate and purchase more than low-valuation customers if lower types participate, now participation is driven by consumer-specific outside options. Now, relative to the canonical price-discrimination case, the monopolist loses some ability to extract consumer surplus as profit maximization requires him to balance informational rent extraction from high-valuation customers with the participation of low-valuation customers. Characterizing the optimal tariff solution in this model with endogenous participation becomes more involved (although much more feasible) than the general multidimensional case, and it requires solving a two-point boundary problem instead of a simpler recursive first-order differential equation with a boundary condition given by the marginal consumer type that decides to participate in the market. Bunching may also occur at the bottom, but only at the bottom, and the tariff is well behaved, continuously approaching the fully efficient solution on the one hand and the solution to the canonical pricing problem on the other. Furthermore, the efficiency at the top result survives, and the efficiency at the bottom arises in cases where all consumer types are served.

The model of Rochet and Stole (2002) is also appealing because it offers the possibility of addressing competitive environments where firms' tariff are the best response to each other's offering and where the tariff offered by the competitor defines the outside option of consumers. This is a model of exclusive agency where consumers subscribe to only one of the firms competing in the industry. The most important result of this literature, also documented by Armstrong and Vickers (2001), is that, in industries with full market coverage and where all firms face the same marginal cost, the equilibrium tariff solution is a simple cost-plus tariff (Coasian two-part) leading to an efficient allocation of consumption among buyers.

See Also

- mechanism design
- mechanism design (new developments)
- price discrimination (empirical studies)

Bibliography

Armstrong, M. 1996. Multiproduct nonlinear pricing. *Econometrica* 64, 51–75.

Armstrong, M. and Vickers, J. 2001. Competitive price discrimination. *RAND Journal of Economics* 32, 579–605.

Boiteux, M. 1956. Sur la gestion des monopolies publics astreint à l'équilibre budgétaire. *Econometrica* 24, 22–40.

Courty, P. and Li, H. 2000. Sequential screening. *Review of Economic Studies* 67, 697–717.

Dupuit, J. 1849. On tolls and transport charges. *International Economic Papers* 11 (1962), 7–31.

Goldman, M., Leland, H. and Sibley, D. 1984. Optimal nonuniform pricing. *Review of Economic Studies* 51, 305–19.

Guesnerie, R. and Laffont, J.-J. 1984. A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm. *Journal of Public Economics* 25, 329–69.

Hadley, A. 1885. *Railroad Transportation: Its History and Its Laws*. New York: G.P. Putnam's Sons.

Laffont, J.-J., Maskin, E. and Rochet, J.-C. 1987. Optimal nonlinear pricing with two-dimensional characteristics. In *Information, Incentives, and Economic Mechanisms*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Maskin, E. and Riley, J. 1984. Monopoly with incomplete information. *RAND Journal of Economics* 15, 171–96.

Miravete, E. 1996. Screening consumers through alternative pricing mechanisms. *Journal of Regulatory Economics* 9, 111–32.

Miravete, E. 2005. The welfare performance of sequential pricing mechanisms. *International Economic Review* 46, 1321–60.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Mirrlees, J. 1976. Optimal tax theory: a synthesis. *Journal of Public Economics* 6, 327–58.

Mussa, M. and Rosen, S. 1978. Monopoly and product quality. *Journal of Economic Theory* 18, 301–17.

Oren, S., Smith, S. and Wilson, R. 1985. Capacity pricing. *Econometrica* 53, 545–66.

Panzar, J. and Sibley, D. 1978. Public utility pricing under risk: the case of self-rationing. *American Economic Review* 68, 888–95.

Pigou, A. 1922. *The Economics of Welfare*, 4th edn. London: Macmillan.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Roberts, K. 1979. Welfare considerations of nonlinear pricing. *Economic Journal* 89, 66–83.

Rochet, J.-C. and Choné, P. 1998. Ironing, sweeping, and multidimensional screening. *Econometrica* 66, 783–826.

Rochet, J.-C. and Stole, L. 2002. Nonlinear pricing with random participation. *Review of Economic Studies* 69, 277–311.

Wilson, R. 1993. *Nonlinear Pricing*. New York: Oxford University Press.

Wilson, R. 1996. Nonlinear pricing and mechanism design. In *Handbook of Computational Economics*, vol. 1, ed. H. Amman, D. Kendrick, and J. Rust eds. Amsterdam: North-Holland.

How to cite this article

Miravete, Eugenio J. "price discrimination (theory)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000333> doi:10.1057/9780230226203.1337

The New Palgrave Dictionary of Economics Online

price dispersion

Ed Hopkins

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Price dispersion occurs when different sellers offer different prices for the same good. Empirical studies have identified price dispersion as widespread and persistent. The most frequent explanation for this is that consumers do not have perfect information about prices. Only recently have economists succeeded in modelling price dispersion as an equilibrium phenomenon: that is, where consumers' decisions to acquire price information are a best response to the distribution of prices, and sellers' pricing decisions are a best response to consumers' search behaviour.

Keywords

clearinghouse models; price discrimination; price dispersion; sequential search

Article

Price dispersion occurs when different sellers offer different prices for the same good in a given market. Thus, it differs from price discrimination under which a single seller offers different prices to different groups of buyers or in different geographical locations. A simple explanation for price dispersion is that it arises from imperfect information on the part of consumers, who do not all buy from the lowest price seller because some at least do not know who the lowest priced seller is. It is an important topic in the field of the economics of information in that there is considerable empirical evidence that price dispersion is widespread and significant. Yet it has proven surprisingly difficult for economists to derive satisfactory models that support price dispersion as an equilibrium phenomenon.

The rise of electronic commerce at the end of the 20th century gave new impetus to empirical studies of pricing behaviour. Baye, Morgan and Scholten (2004) analyse detailed information on prices of 1,000 items collected from a price comparison site. Price dispersion is found to be significant and persistent. Baye, Morgan and Scholten find an average coefficient of variation of about nine per cent for goods sold online. This is comparable with the results of Lach (2002) for conventional retailers who finds a lower

coefficient for the price of refrigerators, but higher variation for grocery items such as coffee or flour. Such empirical work on price dispersion is often disputed on the basis of two arguments, both of which claim that any apparent price dispersion is largely illusory. First, variance in prices might be explained by hidden heterogeneity in the good being offered for sale. For example, a retailer that charges high prices might survive not because of consumer ignorance of cheaper sellers, but because it offers superior service, something not captured by evidence on prices alone. A second line of scepticism is that dispersion in posted prices may not be inconsistent with uniformity in prices actually paid. Those who post high prices may not in fact sell anything. Certainly, one would expect low-priced sellers to sell more than those charging high prices, so that prices weighted by market share will be less dispersed than if all sellers are given equal weight.

The first criticism is addressed by Baylis and Perloff (2002) who find that, in fact, some online sellers persistently offer both high prices and poor service. The second is answered at least in part by Baye, Morgan and Scholten (2004) who in their empirical study concentrate on the difference between the lowest and second lowest price, rather than the difference from lowest to highest or standard deviation, as their measure of dispersion. Furthermore, their data comes from a price comparison site where listings are costly for sellers. Why pay to list a price at which you think there will be no sales?

In any case, it is certainly possible to construct theoretical models in which prices are dispersed and yet high prices yield positive sales. Such theory is recent, however. In an influential survey, Rothschild (1973) identified serious difficulties with the then existing models of price dispersion. At that time, no one had produced a model where price dispersion was shown to be the result of equilibrium behaviour. The challenge was to show that charging a range of prices could be a rational response by sellers to the search behaviour of consumers, and vice versa.

It took some years for this challenge to be met. The difficulty in doing so is illustrated by the earlier work of Diamond (1971), who found that once one introduces imperfect information for consumers, a natural outcome is not price dispersion, but monopoly pricing by sellers. The essence of Diamond's result is the following. Suppose there are a large number of identical buyers who each want to buy one unit of a good from one of a large number of identical sellers, provided it costs no more than a maximum price p^*. The buyers know the distribution of prices but each only knows the price currently being charged by one seller. Each must then must decide whether to learn the price of one more seller at a fixed cost (imagine searching on foot, or by telephoning a succession of sellers). The optimal search policy in this situation of sequential search is to buy the first time one sees a price that is equal or below a reservation level r, which varies with the unit search cost s and distribution of prices $F(p)$. Now, if all consumers have the same unit search cost, then for a given distribution of prices, they will have the same reservation price r. The optimal price for all sellers must then be r. But if there is no dispersion in prices, it cannot be optimal to learn more than one price. Thus, the only equilibrium is where all sellers charge p^* and all buyers do not search, even when the unit cost of search is arbitrarily small. Ironically, this equilibrium satisfies Rothschild's criteria. Consumer behaviour is optimal since, when prices are identical, paying to learn additional prices is a waste of effort; pricing at the monopoly level is optimal since, when there is no search, there is no incentive for sellers to cut prices to increase sales.

Not surprisingly, therefore, many of the earliest successful equilibrium price dispersion models (Salop and Stiglitz, 1977; Varian, 1980) take a different route from Diamond and do not assume sequential search. Instead, they are what have been called by Baye, Morgan and Scholten (2004) ‘clearinghouse’

models. By buying a newspaper or by visiting a price comparison website, a consumer can obtain information about the prices of a number of sellers all at once. The simplest clearinghouse assumption is that it is possible for consumers to become informed of all current prices. Suppose a proportion q of consumers remain uninformed and hence pick a seller at random. The other $1 - q$ consumers are informed and only purchase from the lowest priced seller. All consumers wish to buy one unit of the good if the price does not exceed a common maximum price p^*. Then, given n sellers and L consumers, if one seller charges a price strictly lower than all others, she sells to both informed and uninformed, a total of $qL/n+(1 - q)L$. The other sellers sell only to the uninformed and expect sales of qL/n. That is, demand is decreasing but discontinuous in price.

For simplicity, let us follow Varian (1980) and assume that sellers have constant marginal cost c. There is then no pure strategy equilibrium for sellers as long as there are both informed and uninformed consumers, that is if $q \in (0,1)$. To see this, note that if all sellers charged the same price, it would generally be profitable for an individual to undercut this price in order to attract the informed buyers. However, because of the presence of uninformed consumers who are not price sensitive, charging the monopoly price p^* gives a guaranteed minimum profit of $(p^*-c)qL/n$, and so when the prices of other sellers are close to c, the most profitable price may be p^*. There is a symmetric mixed equilibrium in which all sellers randomize according to the same continuous distribution. This mixed equilibrium is a dispersed price equilibrium, because since sellers randomize over the prices they charge, realized prices will vary over sellers.

However, to have an equilibrium that fully satisfies Rothschild's challenge, it is necessary to make the consumer's decision to become informed endogenous. Varian (1980) assumed differing information costs, with high-cost consumers remaining uninformed, and low-cost consumers paying for information. However, Burdett and Judd (1983) showed that it is possible to close a clearinghouse model even with identical buyers. For example, given the symmetric mixed equilibrium described above, consumers who pay to become informed will buy from the lowest-priced seller whose expected price is equal to the expected value of the lowest of n independent draws from the equilibrium price distribution. In contrast, those who remain uninformed expect to pay the simple expectation of the distribution. If q is zero or 1, the equilibrium price distribution will collapse on c or p^* respectively. However, for interior values of q, the difference between the price paid by the informed and uninformed will be positive. Thus, it can be shown that for a value of s sufficiently low, there is at least one interior value of q such that the resulting equilibrium distribution of prices is sufficiently dispersed such that consumers are indifferent as to whether they pay or remain uninformed.

That is, there is at least one internally consistent dispersed price equilibrium. The proportion of informed consumers generates exactly the right amount of expected price dispersion such that consumers are indifferent between being informed and uninformed. This is an elegant but delicate construction. In contrast, the Diamond outcome (no consumers pay to be informed, all firms charge p^*) is a simple pure equilibrium of this game that coexists with any dispersed price equilibria. Thus the Varian model and the similar models of Salop and Stiglitz (1977) and of Burdett and Judd (1983) have multiple equilibria (though the Bertrand outcome where all consumers pay to be informed and all firms charge marginal cost is not an equilibrium here, since consumers have no incentive to pay to be informed if all prices are the same).

A reasonable question is whether introducing heterogeneity, either under sequential search or in

clearinghouse models, makes dispersed price equilibria more robust. However, consumer heterogeneity does not remove the Diamond paradox as an alternative equilibrium. Even if consumers have a range of search costs, if there is no price variation at all, then there is no incentive to search (unless one makes the implausible assumption that a mass of consumers have zero search costs). That is, if all sellers share the same monopoly price, then all charging that price can be an equilibrium if consumer search is costly. But if instead there is sufficient seller heterogeneity, an outcome where all sellers charge their monopoly price may not be an equilibrium. Suppose no consumer searches, each seller would then charge his or her monopoly price. However, suppose all consumers have the same continuous increasing demand function (in contrast to the unit demand assumed up to now), then a dispersion of costs amongst sellers would lead to heterogeneity in monopoly prices. This could be sufficiently diverse so that consumers would have an incentive to search. Thus, in the equilibrium of Reinganum (1979), low-cost sellers charge their monopoly price, but high-cost sellers must charge less than their monopoly price to make sales.

Finally, when one has heterogeneity of both buyers and sellers, there are two advantages. First, by the above argument, a Diamond-type outcome cannot be an equilibrium and so uniqueness of the dispersed price equilibrium is possible (Benabou, 1993). Second, the dispersed price equilibrium can be pure and strictly monotonic: higher-cost firms charge higher prices. This follows because sufficient buyer heterogeneity can make demand to be continuous in prices, unlike the discontinuous demand in Varian's clearinghouse model. For example, if there is a continuum of buyers who search sequentially and have a continuous density of unit search costs, then there is the possibility of a continuous density of reservation prices. So, demand will increase smoothly as a seller lowers the price.

What are the major conclusions that can be drawn from these equilibrium models of price dispersion? The first is that both social and consumer welfare are typically decreasing with search costs. A reduction in search costs for some consumers can have a positive externality on other consumers, as increased search brings down prices for all. Other predictions can sometimes be counterintuitive. For example, an increase in the number of sellers actually raises the average price charged in the Varian model. However, this result does not hold for all price dispersion models. Further, Baye, Morgan and Scholten (2004) find empirically that both average prices and the degree of price dispersion fall with an increase in seller numbers. Finally, we have seen that models with homogenous sellers give rise to mixed equilibria, while models with bilateral heterogeneity can generate pure equilibria. Randomization over prices would imply regular change in price order amongst sellers. That is, sometimes a given seller would have the highest price, sometime the lowest, and sometimes in the middle. A monotone pure equilibrium would give rise to a stable price ranking. Baylis and Perloff (2002) find that price ranking on online sales of electronic goods are very stable. In contrast, Lach (2002) finds that price ranking in data on prices charged by different Israeli supermarkets is highly variable.

One possibility is that the difference arises because Lach's data are for groceries that are purchased with greater frequency than the electronic goods in Baylis and Perloff's data set. But this highlights that the current theoretical literature on price dispersion has rarely addressed the related issues of repeat purchases, frequency of purchase and search patterns that depend on past experience, for example returning to sellers that have had low prices before. This would seem the area that is in most the need of further research.

See Also

- oligopoly
- price discrimination (theory)
- search theory

Bibliography

Baye, M.R., Morgan, J. and Scholten, P.A. 2004. Price dispersion in the small and in the large: evidence from an internet price comparison site. *Journal of Industrial Economics* 52, 463–96.

Baylis, K. and Perloff, J.M. 2002. Price dispersion on the internet: good firms and bad firms. *Review of Industrial Organization* 21, 305–24.

Benabou, R. 1993. Search market equilibrium, bilateral heterogeneity and repeat purchases. *Journal of Economic Theory* 60, 140–58.

Burdett, K. and Judd, K. 1983. Equilibrium price dispersion. *Econometrica* 51, 955–69.

Diamond, P. 1971. A model of price adjustment. *Journal of Economic Theory* 3, 156–68.

Lach, S. 2002. Existence and persistence of price dispersion: an empirical analysis. *Review of Economics and Statistics* 84, 433–44.

Reinganum, J.F. 1979. A simple model of equilibrium price dispersion. *Journal of Political Economy* 87, 851–8.

Rothschild, M. 1973. Models of market organization with imperfect information: a survey. *Journal of Political Economy* 81, 1283–308.

Salop, S. and Stiglitz, J. 1977. Bargains and rip-offs: a model of monopolistically competitive price dispersion. *Review of Economic Studies* 44, 493–510.

Varian, H. 1980. A model of sales. *American Economic Review* 70, 651–9.

How to cite this article

Hopkins, Ed. "price dispersion." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/

article?id=pde2008_P000368> doi:10.1057/9780230226203.1338

The New Palgrave Dictionary of Economics Online

Price Revolution

John Munro
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Price Revolution was a unique period of inflation in European economic history, enduring for 130 years, from the early 16th to the mid-17 century. It was fundamentally monetary in origins and character, having commenced with a fivefold increase in silver supplies from the central European mining boom and then sustained and expanded both by a financial revolution in negotiable credit instruments and then by the great influx of silver from the Spanish Americas. The extent of the inflation was, however, influenced by various real factors, especially demographic, which had their greatest impact on the income velocity of money.

Keywords

Bodin, J.; Cambridge cash balances equation; coinage debasement; commodity money; consumer price index; deflation; demography; income velocity of money; industrial revolution; inflation; Malestroit, J.; money; population growth; price revolution; Spanish-American silver; urbanization

Article

The Price Revolution, dating from about 1515 to the 1650s, was a long period of persistent inflation in Europe that was unique for the pre-20th-century economy. The sustained rise in prices, or rather in the Consumer Price Index (CPI) is clearly visible in Figure 1 for English prices from 1266 to 1954 (base 1451–75=100), and in Figure 2 for prices in southern England, the southern Low Countries (Brabant), and Spain, from 1501 to 1650. With a common base of 1501–10=100 (CPI) for all three regions, we find that, over the next century and a half to 1646–50, the index number for Spanish prices rose to 343; for Brabantine prices, to 845; and for English prices, to 698.

Figure 1
Consumer prices and wages for master building craftsmen in southern England, in quinquennial means: 1266–1954 (Phelps Brown and Hopkins indices). *Source*: Phelps Brown and Hopkins (1981, pp. 13–59).

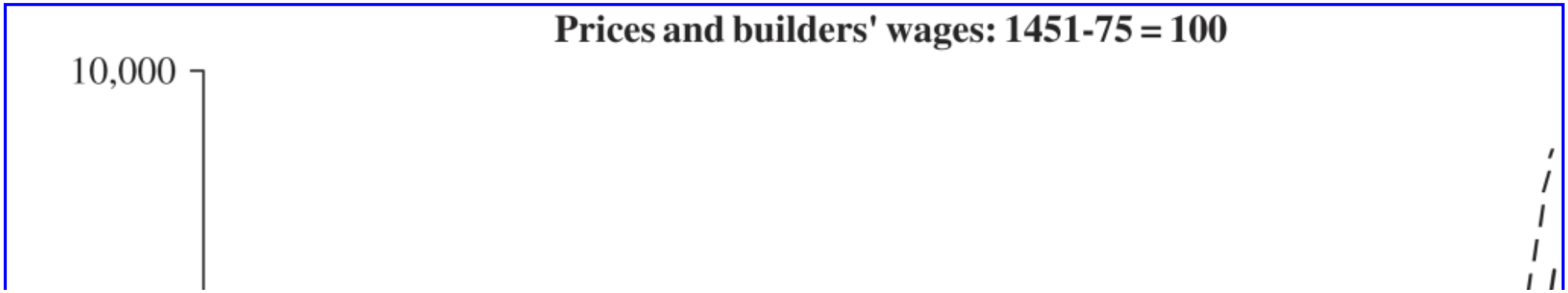

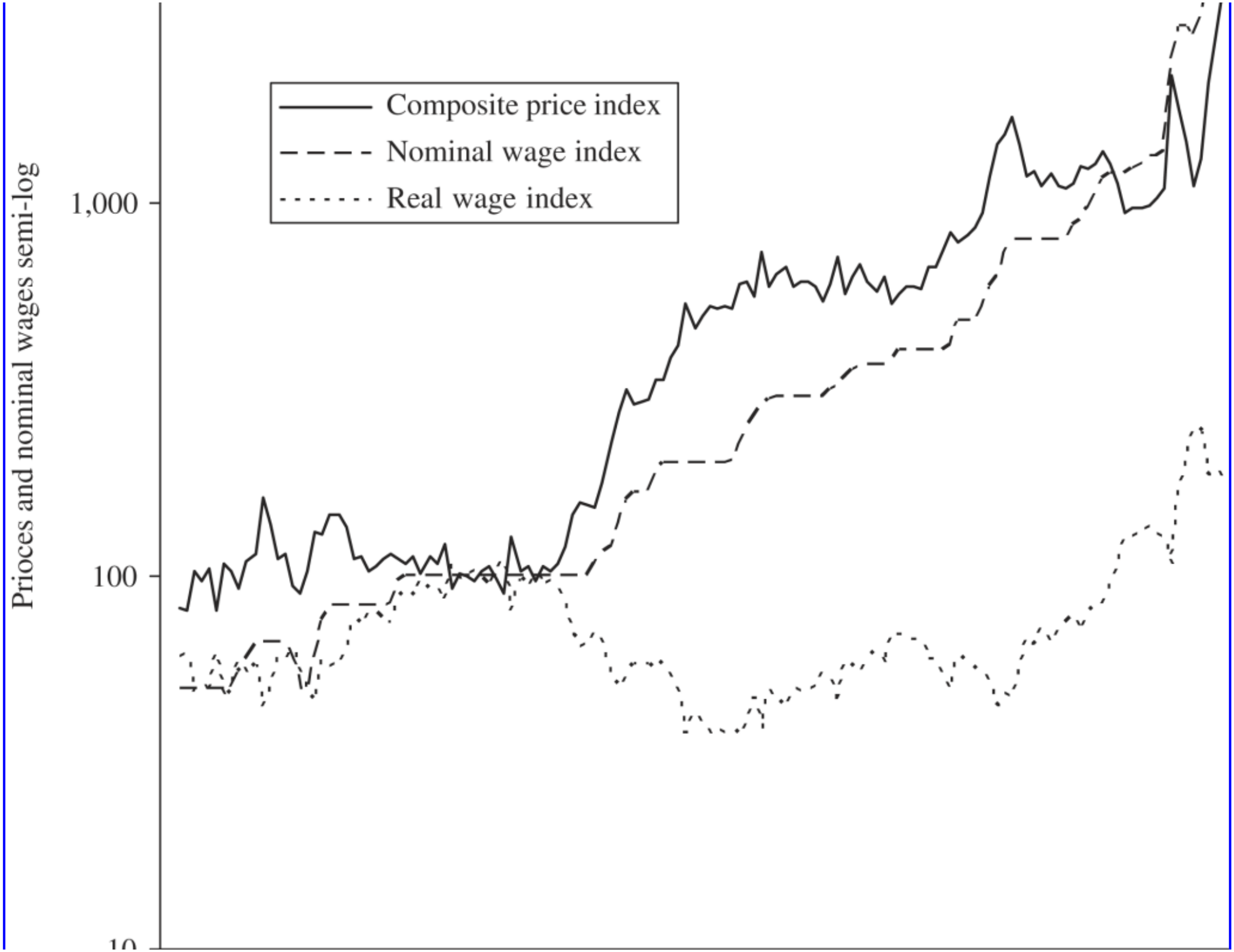
Composite price index
Nominal wage index
Real wage index
Prioces and nominal wages semi-log
1,000
100
10

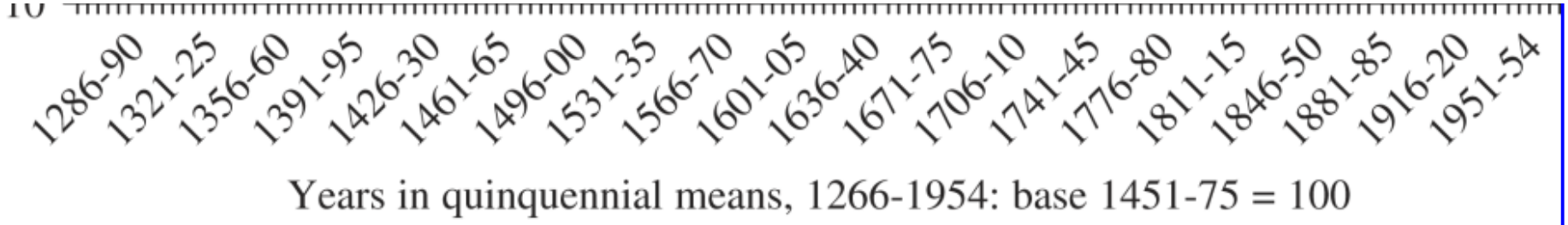

Source: Phelps Brown and Hopkins (1981, pp. 13–59).

Figure 2

Price indexes: England, Brabant, and Spain, 1401–1650. *Source*: England, as Figure 1. Brabant, Van der Wee (1975, pp. 413–35). Spain, Hamilton (1934; pp. 262–403).

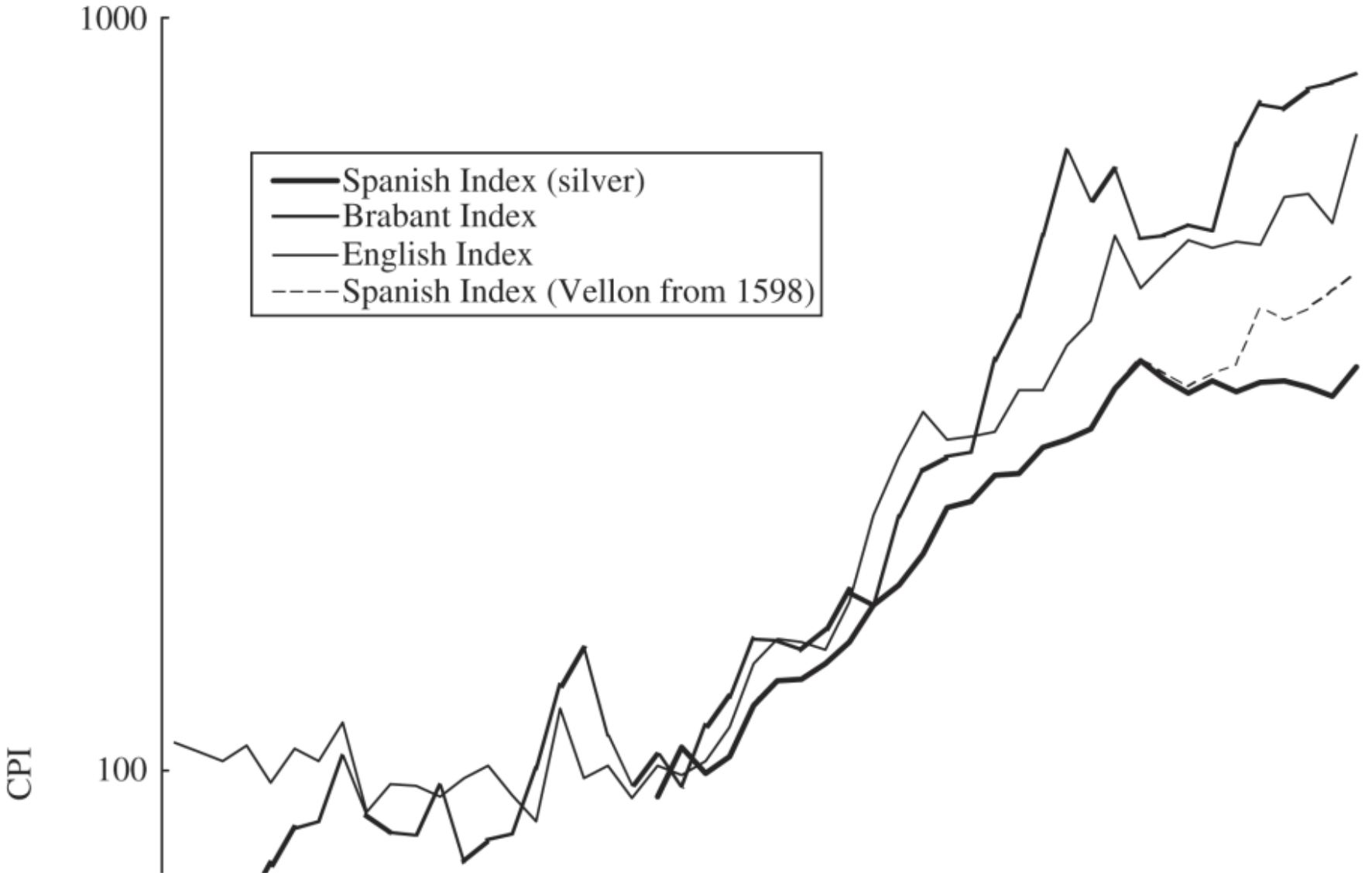

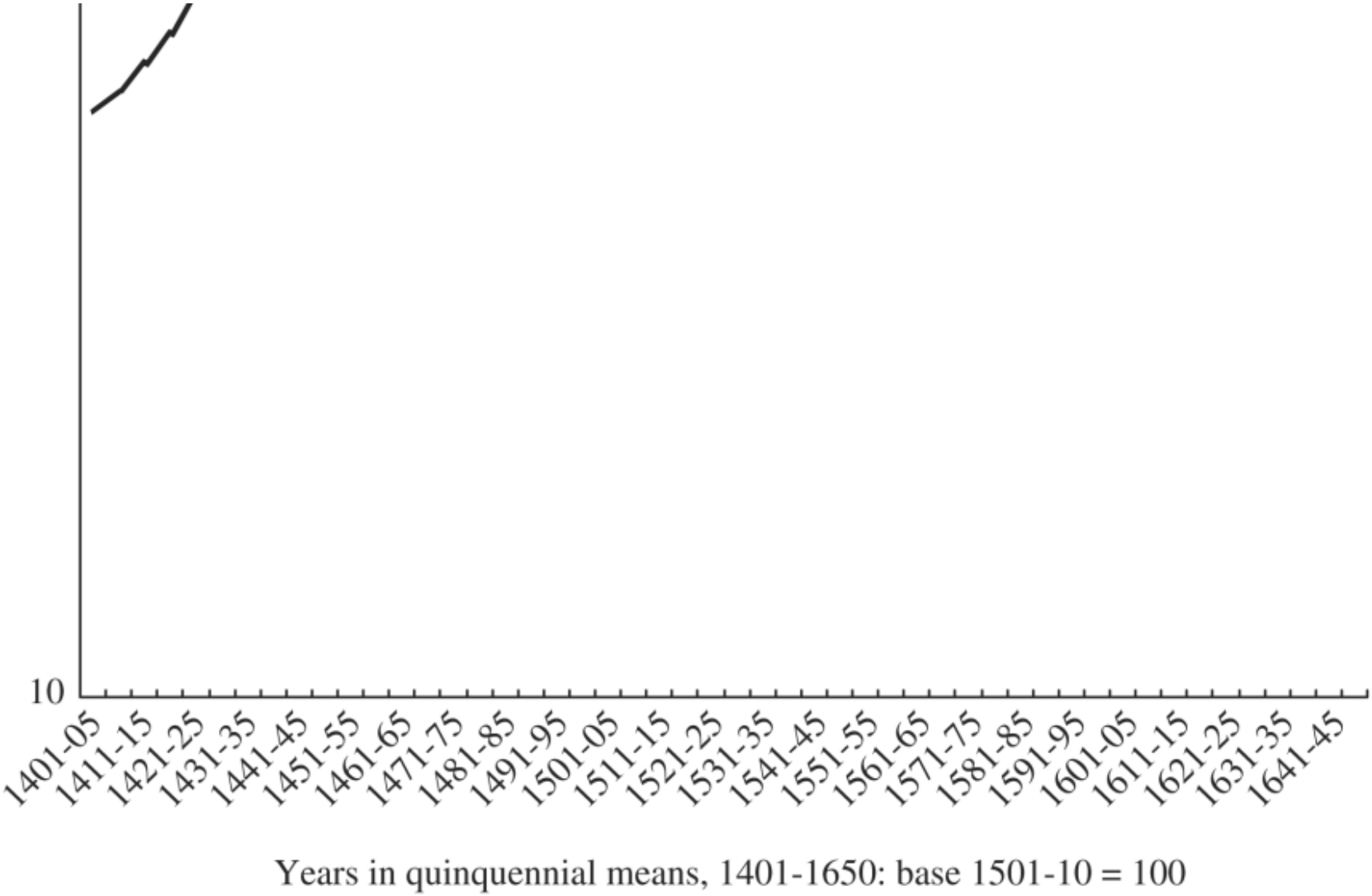

Source: England, as Figure 1. Brabant, Van der Wee (1975, pp. 413–35). Spain, Hamilton (1934, pp. 262–403).

Average annual rates of price increases of less than two per cent in the Price Revolution era may have been mild in comparison with 20th-century inflations: but all pre-20th century inflations were based on commodity moneys, not government issues of fiat money, as in the modern world. Before 1914, western Europe experienced, to be sure, other periods of long-term inflation, particularly, if only periodically, during the ‘long’ 13th century (1180–1315) and in the early Industrial Revolution era (1760–1815). But these produced price-level changes that were far smaller than those of the Price Revolution.

All long-term inflations are fundamentally monetary in nature, even though secondarily influenced by real factors. That may be best understood through the Cambridge cash balances equation, $M = k.P.y$, in which k indicates the quantity of cash balances (high-powered money M) held as a proportion of net national income (y). It is also the inverse of V, the income velocity of money, in the more familiar quantity-theory equation: $M.V = P.y$. Since the opportunity cost of holding cash is forgone interest income, changes in k should therefore depend partly on interest rates. Though an increase in M (money stocks) may prove inflationary, the equation indicates why the extent of such inflation is unpredictable. For such an increase in P can be offset by a rise in k (especially if an increased M reduces interest rates), that is, a fall in V, or by an increase in y, stimulated by increased spending and falling interest rates.

Demographic versus monetary explanations

Regrettably, amongst historians population growth has provided the most popular explanation for the Price Revolution. Contemporary explanations for the Price Revolution, especially in the debate between the French philosophers Jean Bodin and Jean Malestroit (1568), were instead purely monetary: that is, concerning the influx of Spanish-American silver during the 16th and 17th centuries. Modern opponents of this thesis have, however, rightly pointed out that virtually no American silver was imported before the 1530s, and only insignificant volumes were received before the 1560s, while inflation was clearly under way by 1515–20.

Yet to assume that consequently demographic factors provide the only possible alternative explanation is an absurd non-sequitur. There are two major problems with the demographic thesis. First, its most common form confuses microeconomic with macroeconomic changes. Although population growth, with fixed amounts of land and a static technology, should lead to a rise in the *relative* price of grains compared with those for manufactures, it can not explain a rise in the general price level. Second, the populations of both England and the Low Countries in the 1520s were at their late-medieval nadir, about half of what they had been around 1300 (when the English CPI was only 102); and thus any demographic recovery from such a low level could not possibly have provided the initial cause of an inflation that was under way in that very same decade.

The actual origins of the European Price Revolution lie instead in alternative monetary explanations, commencing with the central European silver-copper mining boom in the 1460s. This was an era of severe deflation (in silver-based prices) that had thereby augmented the purchasing power of silver and provided the key profit incentives for two crucial technological innovations: (*a*) in mechanical engineering: water-powered piston drainage pumps that permitted deeper mining, reaching richer ores; and (*b*) in chemical engineering, the *Saigerhütten* process using lead to smelt silver–copper ores, thus for the first time separating the two metals, which were present in vastly larger ore bodies than those of silver alone. The resulting silver–copper mining boom increased aggregate output of European mined silver about fivefold by the 1540s, producing far more silver than was imported from the Americas until the 1580s. By my own conservative estimates, central European silver production itself rose from an average annual of 12,873•kg in 1471–5 to 55,704•kg in 1536–40.

As late as 1556–60, only 27,145•kg of American silver were imported yearly into Seville; but in 1566–70 annual mean imports jumped to 83,274•kg, thanks to another technological innovation: the mercury amalgamation method, employed first at Potosi (Peru) and Zacatecas (Mexico). Thereafter, rising imports, reaching a maximum of 273,821•kg per year in 1591–5, but still amounting to an impressive 223,023•kg per year in 1621–5, continued to fuel the inflation. When the Price Revolution ended in 1656–60, silver imports had diminished to an annual mean of just 27,965•kg. Spanish-American mines were then experiencing severely diminishing returns, while far more metal was being retained for use in the Americas, and more and more silver was being exported across the Pacific, in trade with the Philippines and China.

There was one additional monetary factor to explain the European Price Revolution, namely, a veritable financial revolution in the Habsburg Netherlands, whose towns (from 1507) and then the Estates General (1539–43) established all the legal requirements for negotiability, including legalization of interest and discounting, to protect the rights of third parties in transferable bills, so that bills obligatory and bills of exchange could circulate from hand to hand in commercial and financial transactions as though they were paper money. This financial revolution also established full-fledged negotiability and thus far wider use of government debt instruments, internationally traded on the Antwerp *beurse* from 1531, as perpetual annuities known as *rentes* or *juros*. One measure of their vastly growing importance is the increased issue of Spanish *juros*, from 3.6 million ducats in 1516 to 80.4 million ducats in 1598, most of them held abroad. This financial revolution also increased the income velocity of high-powered money.

Demography and the income velocity of money

Just the same, demographic factors are not irrelevant to our understanding of the dynamics of the Price Revolution, not when population growth became so much more dramatic from the 1540s to the 1640s. First, in various ways that have been elaborated by Harry Miskimin (1975), Jack Goldstone (1984) and Peter Lindert (1985), that population growth, combined with more urbanization, the development of more complex commercial and financial networks, and changes in the age pyramid (with more dependants), may have increased the income velocity of money. Furthermore, as Nicholas Mayhew (1995) has shown, the Keynesian predictions of a fall in income velocity with continued expansions in monetary stocks (and falls in interest rates) seems to hold true from the 13th to the 20th century, with one singular exception: the Price Revolution era.

The role of coinage debasements

Finally, what explains the differences in the inflation rates revealed across the three countries in Figure 2: why did Spanish prices rise less than English, and English prices rise less than Brabantine? Coinage debasement (depreciation) seems to have played a role in these differences. Spain experienced no silver coinage debasement. England experienced one mild coinage debasement, in 1526, and one set of very severe debasements between 1542 and 1552 (though the silver coinage was only partially restored, in 1560–61); but none thereafter. Brabant, on the other hand, suffered a long series of coinage debasements, during the 16th and 17th centuries. Thus, the explanations for the European Price Revolution involve a complex set of monetary and real factors, though monetary factors predominated.

See Also

- Bodin, Jean
- commodity money
- cost-push inflation
- demand-pull inflation
- depreciation
- economic demography
- Fisher, Irving
- hyperinflation
- inflation
- inflation measurement
- Keynes, John Maynard
- monetary economics, history of
- money
- money supply
- population dynamics

Bibliography

Goldstone, J. 1984. Urbanization and inflation: lessons from the English Price Revolution of the sixteenth and seventeenth centuries. *American Journal of Sociology* 89, 1122–60.

Gould, J. 1971. The Price Revolution reconsidered. In *The Price Revolution in Sixteenth-Century England*, ed. P. Ramsey. London: Methuen and Co.

Hamilton, E. 1934. *American Treasure and the Price Revolution in Spain, 1501–1650*. Cambridge, MA: Harvard University Press.

Hamilton, E. 1971. American treasure and Andalusian prices, 1503–1660: a study in the Spanish Price Revolution. In *The Price Revolution in Sixteenth-Century England*, ed. P. Ramsey. London: Methuen and Co.

Lindert, P. 1985. English population, wages, and prices: 1541–1913. *Journal of Interdisciplinary History* 15, 609–34.

Mayhew, N. 1995. Population, money supply, and the velocity of circulation in England, 1300–1700. *Economic History Review*, 2nd series, 48, 238–57.

Miskimin, H. 1975. Population growth and the price revolution in England. *Journal of European Economic History* 4, 179–85.

Munro, J. 2003. The monetary origins of the ‘Price Revolution’: South German silver mining, merchant-banking, and Venetian commerce, 1470–1540. In *Global Connections and Monetary History, 1470–1800*, ed. D. Flynn, A. Giráldez and R. von Glahn. Aldershot and Brookfield, VT: Ashgate.

Outhwaite, R. 1982. *Inflation in Tudor and Early Stuart England*. 2nd edn . London: Macmillan Press.

Phelps Brown, E. and Hopkins, S. 1981. *A Perspective of Wages and Prices*. London and New York: Methuen.

Van der Wee, H. 1975. Prijzen en lonen als ontwikkelingsvariabelen: Een vergelijkend onderzoek tussen Engeland en de Zuidelijke Nederlanden, 1400–1700. In *Album offert à Charles Verlinden à l'occasion de ses trente ans de professoriat*, ed. Jan. Craeybeckx. Wetteren: Universum. Republished as 'Prices and wages as development variables: a comparison between England and the Southern Netherlands, 1400–1700', in H. Van der Wee, *The Low Countries in the Early Modern World*, trans. L. Fackelman, Aldershot and Brookfield, VT: Variorium-Ashgate Publishing, 1993, but without the statistical tables in the original Dutch publication.

How to cite this article

Munro, John. "Price Revolution." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000170> doi:10.1057/9780230226203.1339

The New Palgrave Dictionary of Economics Online

primary and secondary labour markets

Paul Ryan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

dual economies; duality; labour market discrimination; primary and secondary labour markets; productivity growth; segmentationism; uneven development; wage dispersion

Article

The primary/secondary distinction involves an application of the concept of economic dualism to the labour markets of advanced capitalist economies.

In the initial formulation, the primary and secondary segments of a dual labour market were distinguished principally by job characteristics. The rewards of primary jobs, in terms of earnings, working conditions, job security, training opportunities and career prospects, are high; those of secondary jobs, low. Increases in a worker's schooling and work experience lead to higher job rewards in the primary segment but not in the secondary one. Inter-segment mobility is limited, the working poor being confined to secondary jobs. A separate dichotomy in worker traits parallels that in jobs. Secondary workers are those with weak attachment to employment, a consequence of social roles in either the household (youths and married females) or the locality (inner-city minorities; Piore, 1970).

Two important differences soon emerged in dualist interpretation. The central difference between the segments for some authors involves stability of employment; for others, pay levels (Piore, 1970; Bluestone, 1970). Some see the dualist classification as partial; others seek to classify all jobs and workers within an exhaustive schema. Exhaustiveness has become predominant, with the ensuing heterogeneity of an enlarged primary segment leading to further dualisms (upper/lower tier and core/periphery, by occupation and industry respectively) within primary employment (Bluestone, 1970; Edwards, Reich and Gordon, 1975; Piore, 1975).

Dualist interpretations originate from two sources. The first is the failure in the 1960s of a manpower policy oriented to the enhancement of individuals’ job skills to move large numbers of US inner-city residents into stable and well-paid work. The explanation was sought in the characteristics not of workers but of jobs, with the primary/secondary duality building upon the antecedent structured/

unstructured one (Kerr, 1954). The second source is the concern of radical economists to understand the political disunity of the US labour force, a painful anomaly for Marxist analysis. The key to political fragmentation has been sought in the differentiation of work experiences in a dual labour market (Edwards, 1979; Gordon, Edwards and Reich, 1982).

Dualism is a variant of segmentationism, sharing with it three attributes which distinguish both from orthodox labour theory. The first is the widening of the analytical scope beyond comparative statics with given preferences and indeterminate public policy. Thus the instability of inner-city employment is attributed to an interaction between worker attitudes and job attributes, with attitudes thereby made endogenous. The secondary status of female and youth labour is understood in terms not of autonomous preferences but rather of power relations within family and state (Humphries and Rubery, 1984).

Secondly, the labour market is seen as systematically differentiating the job rewards achieved by comparable individuals. The market then becomes a source of inequality in its own right. Thus dualism in employment stability is understood to result not so much from the aversion of secondary workers to steady work as from their discriminatory exclusion from stable jobs. Similarly, the low pay of secondary workers is explained not so much in terms of low labour quality as of denial of access to the primary jobs which convert high potential into high actual productivity (Ryan, 1981).

Finally, labour market outcomes such as pay and turnover are seen as determined principally by such product market attributes as demand variability, employer power and production technology. The part played by labour market influences, including trade unions, is a subsidiary one. An important role is given to competitive forces in determining labour outcomes, but such forces derive more from the product than from the labour market (Wilkinson, 1981).

These three attributes rebut the criticism that the dualist and segmentationist approach is largely descriptive, taxonomic and compatible with competitive theory (Wachter, 1974; Cain, 1976).

Considering dualism as a subset of segmentationism, two interpretations may be placed upon their relationship. The first is descriptive. *Heuristic duality* describes vividly the idea of differential treatment in labour markets without implying discontinuity or universality. Thus to distinguish good and bad jobs for comparable workers need not rule out large numbers of medium jobs and unclassified jobs. Heuristic dualism is also seen in the distinction between sheltered and exposed sectors, familiar in 1920s Britain, when currency overvaluation depressed relative wages according to exposure to foreign competition (Dobb, 1928). To postulate a sheltered/exposed dualism is to dramatize the issue without requiring that exposure itself be dichotomous or that a comprehensive theory of labour outcomes be built on such a limited basis.

The second interpretation of dualism is more demanding. *Strict duality* requires not just a substantial dispersion of job rewards for comparable individuals but also the polarization of their distribution into two clearly separate segments, each with low internal heterogeneity; a substantial distance between average job rewards in the two segments; and few cases falling in the intermediate range. Such bimodality must maintain and reproduce itself over time, while individuals and jobs should show low rates of mobility across a clear intervening boundary. Such conditions may fail to be realized literally in practice but strong tendencies towards them are required for strict dualism to be sustained.

Although the heuristic and the strict formulations of duality are frequently confused, leading dualist writers have explicitly espoused strict duality. The causes of a postulated strict dualism in job rewards have been sought in underlying dichotomies in three dimensions of industrial structure. The first

explanation sees labour dualism in terms of employment stability. Selective worker organization in pursuit of job security leads to primary jobs in firms producing for the stable portion of product demand, with unstable secondary jobs where employers sell to the variable or unpredictable part (Berger and Piore, 1980). The second approach sees dualism in terms of earnings, relating it to an underlying dichotomy amongst firms and industries in market and political power (Averitt, 1968). The third explanation distinguishes firms whose organizational structures motivate and control their employees by providing stable jobs, career prospects and high pay from those which rely upon the traditional methods of low pay, insecurity and discipline. This dichotomy in control techniques overlaps with the preceding one by producer power, it being the large and powerful corporations which adopt sophisticated control strategies (Gordon, Edwards and Reich, 1982).

These three theories of strict dualism all capture important sources of segmentation in labour markets. An empirical role is most evident for producer power, in the shape of significant associations between employee rewards and such power correlates as seller concentration, firm size and ties to the state. However, strict dualism oversimplifies the links between industrial structure and labour outcomes. These theories offer no reason for the distributions of demand variability, producer power or control strategies to become polarized in the first place. In practice, the nexus between product and labour markets proves empirically multidimensional and complex (Wallace and Kalleberg, 1981; Hodson and Kaufman, 1982). Moreover, while bimodality has been found in some attributes of industrial structure, this typically appears in only one of a set of several attributes; is found in data-sets which exclude more than half of national employment; and even then does not lead to any strict dualism in labour outcomes (Oster, 1979; Buchele, 1983).

The empirical status of segmentation (and heuristic duality) remains controversial, reflecting the difficulty of measuring labour quality and market structure. The evidence concerning strict duality, is, however, distinctly unfavourable. No clear boundary emerges between segments. Definitions of the secondary segment vary widely in size and composition from one study to another, with intermediate groups proving numerous and difficult to classify. The difference in average job rewards between segments in most dualist definitions proves only moderate in earnings and erratic in employment stability. Mobility between segments appears too high to support an inference of wholesale confinement to secondary employment.

The empirical failure of strict dualism in the domestic economy may be understood by considering a more promising candidate: the world labour market, treated as a potential whole (Singer, 1970). The gap between the earnings of comparable workers in the two poles of advanced and developing countries is great; intermediate cases (the newly industrializing countries) are certainly numerous, but bimodality is still expected; while the distance in earnings between the two poles has proved not only durable but at times even increasing, with the attainment of higher rates of growth in productivity and earnings in advanced than in developing countries (Brandt Commission, 1980).

Similar forces for dualist divergence function within the labour markets of both the advanced economies and the world economy. The international phenomena of uneven development and cumulative divergence, resting upon the attainment of higher rates of investment and productivity growth in advanced then in developing countries, have as their national counterpart the large and persistent differences in productivity growth across sectors (Salter, 1960). Unequal exchange, or the systematic overvaluation of the output of advantaged countries at the expense of that of weaker ones, also finds its

domestic analogue in the output prices of primary and secondary segment employers.

One reason why strict dualism applies more to the international than to the national labour market involves the greater obstacles to factor mobility across than within national boundaries. Two other influences are potentially more important. First, the dispersion of rates of growth of value productivity within advanced economies is limited relative to its international counterpart. In the domestic economy, sectors with low rates of growth of physical productivity experience either the transfer of production to developing countries (in the case of tradables) or the revaluation of their output by increases in relative price (in the case of non-tradables). The former mechanism has no counterpart in the international context. Second, the world economy lacks the institutions which prevent differences in rates of growth of value productivity from producing increasing dispersion (let alone polarization) within the distributions of job rewards of the advanced economy. Relativity bargaining (for the organized), statutory wage minima and indexed social security provision (for the unorganized) prevent substantial widening of the gap between earnings in high and low productivity growth employment. The only counterpart to these forces in the world economy is development aid, a pale reflection of social security in the domestic economy. The polarization of labour outcomes is therefore possible in the world labour market to an extent inconceivable in the domestic one.

The factors which curb the dispersion of labour outcomes within advanced economies have been weakened lately by the growth of unemployment and anti-regulatory sentiment. They remain nevertheless sufficiently powerful to restrain domestic tendencies toward dualist divergent development. The dual labour market provides a tenable account of labour market segmentation within advanced economies only in its weaker, heuristic formulation.

See Also

- labour market institutions

Bibliography

Averitt, R.T. 1968. *The Dual Economy*. New York: Norton.

Berger, S. and Piore, M. 1980. *Dualism and Discontinuity in Industrial Societies*. Cambridge: Cambridge University Press.

Bluestone, B. 1970. The tripartite economy. *Poverty and Human Resources Abstracts* 5(4), 15–35.

Brandt Commission. (Independent Commission on International Development Issues). 1980. *North–South: A Programme for Survival*. London: Pan Books.

Buchele, R.K. 1983. Economic dualism and employment stability. *Industrial* Relations 22, 410–18.

Cain, G.C. 1976. The challenge of segmented labor market theories to orthodox theory: a survey.

Journal of Economic Literature 14, 1215–57.

Dobb, M. 1928. *Wages*. London: Nisbet, 1956.

Doeringer, P.B. and Piore, M. 1971. *Internal Labor Markets and Manpower Analysis*. Lexington, MA: D.C. Heath.

Edwards, R.C. 1979. *Contested Terrain*. New York: Basic Books.

Edwards, R.C., Reich, M. and Gordon, D.M., eds. 1975. *Labor Market Segmentation*. Lexington, MA: D. C. Heath.

Gordon, D.M., Edwards, R.C. and Reich, M. 1982. *Segmented Work, Divided Workers*. Cambridge: Cambridge University Press.

Hodson, R. and Kaufman, R. 1982. Economic dualism: a critical review. *American Sociological Review* 47, 727–39.

Humphries, J. and Rubery, J. 1984. The reconstitution of the supply side of the labour market. *Cambridge Journal of Economics* 8, 331–46.

Kerr, C. 1954. The balkanization of labor markets. In *Labor Mobility and Economic Opportunity*, ed. E. W. Bakke. Cambridge, MA: MIT Press.

Oster, G. 1979. A factor analytic test of the theory of the dual economy. *Review of Economics and Statistics* 61, 33–9.

Piore, M. 1970. The dual labor market; theory and implications. In *The State and the Poor*, ed. R. Barringer and S.H. Beer. Cambridge, MA: Winthrop.

Piore, M. 1975. Notes for a theory of labour market stratification. In Edwards, Reich and Gordon (1975).

Ryan, P. 1981. Segmentation, duality and the internal labour market. In Wilkinson (1981).

Salter, W.E.G. 1960. *Productivity and Technical Change*. Cambridge: Cambridge University Press.

Singer, H.W. 1970. Dualism revisited: a new approach to the problems of the dual society in developing countries. *Journal of Development Studies* 7(1), 60–75.

Wachter, M. 1974. Primary and secondary labor markets: a critique of the dual approach. *Brookings Papers on Economic Activity* 1974(3), 637–80.

Wallace, M. and Kalleberg, A. 1981. Economic organization of firms and labor market consequences: towards a specification of dual economic theory. In *Sociological Perspectives on Labor Markets*, ed. I. Berg. New York: Academic Press.

Wilkinson, F., ed. 1981. *Dynamics of Labor Market Segmentation*. New York: Academic Press.

How to cite this article

Ryan, Paul. "primary and secondary labour markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000175> doi:10.1057/9780230226203.1340

The New Palgrave Dictionary of Economics Online

principal and agent (i)

J.E.C. Munro

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

asymmetric information; principal and agent (i); contract law; Munro, J. E.

Article

An agent is a person who is employed to do an act on behalf of another called the principal, so that as a rule the principal himself becomes bound. That one person can represent another is a doctrine that has developed but slowly. In Roman law it was a general principle that no one could enter into a contract by stipulation on behalf of another, and in the case of mandate the mandatarius or quasi-agent incurred a personal liability towards their parties. The modern principle is that contracts entered into by an agent are regarded as entered into by the principal, provided the contract is within the scope of the agent's authority.

No special form of words is required to appoint an agent, and agency may be inferred from the conduct of the parties. An agent is required to conduct the business entrusted to him with as much skill as is generally possessed by persons engaged in a similar business, to act with reasonable diligence, to display the utmost fidelity, to keep proper accounts, and to pay over all moneys received less any expenses and his own remuneration.

Directors, managers, clerks, and servants, having power to act for their principals or masters, are agents. Besides these, the chief classes of agents are (*a*) factors; (*b*) brokers; (*c*) auctioneers ; and (*d*) ship masters. Each class is subject to the usages of the trade relating to the class. An agent cannot as a rule delegate his powers, but by the custom of certain trades sub-agents may be employed. The relation of principal and agent is terminated by mutual consent, by revocation, by the agent renouncing, by the expiration of the time agreed upon by the completion of the business, by the death or lunacy of either principal or agent, and by the bankruptcy of the principal.

Reprinted from *Palgrave's Dictionary of Political Economy*.

How to cite this article

Munro, J. E. C. "principal and agent (i)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000182> doi:10.1057/9780230226203.1341

The New Palgrave Dictionary of Economics Online

principal and agent (ii)

Joseph E. Stiglitz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The principal–agent literature is concerned with how the principal (say an employer) can design a compensation system (a contract) which motivates another individual, his agent (say the employee), to act in the principal's interests. A principal–agent problem arises when there is imperfect information concerning what action the agent either has undertaken or should undertake. It arises in insurance and credit relationships because of their intertemporal nature, when it is known as ‘moral hazard’. It also arises where opportunities exist for the principal to extract as much rent as possible from the agent.

Keywords

asymmetric information; compensation; competitive equilibrium; credit; employer–employee relationship; implicit contracts; individual rationality constraints; insurance; intertemporal relationships; Mirrlees, J.; moral hazard; optimal taxation structures; Pareto efficient tax structures; partially discriminating monopoly; pooling equilibrium; principal and agent (ii); rent seeking; reservation utility constraint; risk; risk sharing; Ross, S; separating equilibrium; sharecropping; Stiglitz, J. E.

Article

The principal–agent literature is concerned with how one individual, the principal (say an employer), can design a compensation system (a contract) which motivates another individual, his agent (say the employee), to act in the principal's interests. The term principal–agent problem is due to Ross (1973). Other early contributions to this literature include Mirrlees (1974, 1976) and Stiglitz (1974, 1975).

A principal–agent problem arises when there is imperfect information, either concerning what action the agent has undertaken or should undertake. In many situations, the actions of an individual are not easily observable. It would be very difficult for a landlord to monitor perfectly the weeding activity of his tenant. A bank cannot monitor perfectly the actions of those to whom it lends money. The employer cannot travel on the road with his salesman, to monitor precisely the effort he puts into his salesmanship.

In each of these situations, the agent's (tenant's, borrower's, employee's) action affects the principal (landlord, lender, employer). Clearly, if an individual's actions are unobservable, then compensation cannot be based on those actions. In some cases, even if an individual's actions are not directly observable, it may be possible to infer his actions. Thus, if output were a function just of effort $[Q = F(e)]$ then even if effort were unobservable, if output were observable, and the relationship between output and effort were known, then effort could be inferred with perfect accuracy.

The principal–agent literature focuses on situations where an individual's actions can neither be observed nor be perfectly inferred on the basis of observable variables; thus, for instance, it is usually assumed that output is a function of effort and an unobservable random variable, θ: $Q = F(e, \theta)$.

Moreover, in many circumstances, the principal wishes the agent to take actions based on information which is available to the agent, not the principal. Indeed, this is the very reason that individuals delegate responsibility. Because of the asymmetry of information, the principal does not know whether the agent undertook the action the principal would himself have undertaken, in the given circumstances. Hence, even if the principal can observe the action, he may not know whether that action was appropriate.

Since, in general, the pay-offs to the agent will differ from those to the principal, the agent will not in general take the action which the principal would like him to take, or that they would contract for in the presence of perfect information. For instance, the employee may not adjust his effort as the situation requires, or he may engage in too much or too little risk taking.

The principal–agent problem is, then, the central problem of economic incentives.

In spite of the importance attached to *economic incentives*, until recently economic theory had little to say on the matter. In the standard theory, individuals were paid for performing a particular task. If they performed the task, they received their compensation; if they failed to perform the task, they did not. Individuals thus always had an incentive to perform the contracted-for service. Only if the employer were so foolish as to pay the worker whether he performed the task or not would an incentive problem arise.

The standard theory was based on the assumption that what action the 'principal' wished his agent to perform was perfectly known, and that the action could be perfectly and costlessly monitored. Neither assumption is plausible and, indeed, relatively few workers are paid solely on the basis of their observed inputs.

Origins of principal – agent problems

Principal–agent problems arise whenever one individual's actions have an effect on another individual. The question arises, then, why cannot economic relationships be designed to avoid this kind of dependency? Under what circumstances do these interdependencies arise? For instance, if a landlord were to sell or rent his land to his tenant, then the workers' effort would have no effect on him. If an employer were to sell or rent his capital to his worker, then the workers' effort would again have no effect on him. Traditional neoclassical analysis emphasized the symmetry in economic relationships: one could describe the employer–employee relationship as the employee hiring capital just as well as one could describe it as the employer hiring labour. (This Wicksellian description of economic relationships always seemed peculiar to me; it seemed to suggest the absence within neoclassical analysis of certain important aspects of economic relationships; it is those aspects which are the subject of scrutiny here.)

There are three important reasons for the existence of principal–agent problems. Two have to do with

the essential intertemporal nature of certain relationships: insurance and credit. When two individuals enter into an insurance contract, one individual (*a*) promises to pay the other (*b*) a certain amount if event A occurs, while the other (*b*) promises to pay (*a*) a certain amount if event B occurs. If there are actions which one of the individuals can undertake between the date of the contract and the event which will affect the outcome, then there is a principal–agent relationship between the two. This particular form of the principal–agent problem is referred to within the insurance literature as the *moral hazard problem* (see Arrow, 1965), and, by extension, the term has been applied to the principal–agent problem more generally.

Similarly, in credit relationships, one individual gives another some resource (money), in return for a promise to repay that money at some later date. So long as there is some probability of default, which can be affected by the actions of the borrower, there is a *moral hazard* or *principal–agent* problem (provided that that action cannot be perfectly monitored by the lender).

Many economic relationships have an important element of insurance within them. The landlord–tenant sharecropping relationship can be viewed as if the tenant pays a fixed rent, and then receives an insurance policy from the landlord, in which the landlord agrees to pay the tenant a certain amount if output is low (equal to the difference between his share and the fixed rent); and the tenant agrees to pay a premium equal again to the difference between the share and the fixed rent, when output is high.

Indeed, the credit ‘problem’ can be viewed as a special form of an insurance relationship: the lender provides an insurance policy, such that if the borrower's resources are less than the amount owed, the lender agrees to pay the borrower the difference (which the borrower then immediately repays to the lender). The premium is the difference between the rate of interest on a perfectly safe loan and the rate of interest charged on this risky loan.

Insurance (spreading and transferring risk) provides one of the explanations of sharecropping; were workers to rent the land, they would have to absorb all the risk associated with output variations. With sharecropping, the risk is shared between the landlord and the tenant. Since the wealth of tenants is usually much less than that of landlords, there is some presumption that the landlords are better able to absorb this risk.

But even if the tenants were risk neutral, there might be a principal–agent problem. We suggested above that if the landlord were to rent his land to the tenant, there would be no principal–agent problem. But this is not quite correct. If the tenant did not have sufficient resources to pay the rent before production, then the landlord would have to lend the tenant the money. (If he receives the rent at the end of the period, then it is as if he is lending the individual the money.) And then, if there are actions which the individual can undertake which affect the likelihood of not being able to repay the debt (pay the rent), *then* there is a moral hazard problem.

There is a second reason that renting land might not solve the moral hazard problem. There may be actions which the tenant can take which affect the quality of the land. To the extent that those actions are monitorable, the rental agreement may specify the actions to be undertaken (e.g. concerning what crops are to be grown, or grazing patterns). But these actions are not perfectly monitorable, and thus, even with rental agreements there are important principal–agent problems. (The same issues arise, of course, with the rental of any durable good.)

Again, one should ask, cannot these principal–agent problems be alleviated, e.g. by *selling* the asset. But this entails precisely the two problems we identified before as giving rise to principal–agent relationships: The agent (tenant, employee) may not have sufficient capital (and thus must borrow to

make the purchase, creating a credit principal–agent problem); and if there is any risk associated with the future value of the land, it imposes a risk on the agent. Any attempt to alleviate those risks (through insurance) again gives rise to a moral hazard problem.

The third major source of principal–agent relationships is rather different. It arises from the attempt of the principal to extract as much *rent* (surplus) from the agent as possible. The employer does not know how difficult the task is that he would like the worker to perform. He could pay the worker the full value of his output, but that would leave him no profits. He might pay much less, but that might result in the worker refusing to work, if the task is in fact quite difficult; and thus he would lose profits that he might otherwise obtain. This rent extraction problem has been particularly well studied in the context of public utilities: the government does not know the minimum amount of compensation required to keep the utility producing. The rent extraction problem may be alleviated within competitive environments by holding auctions: the individual for whom the asset (franchise) is most valuable will bid the most. But there may not be enough bidders to extract all the rents through an auction mechanism; and at least in the case of utilities, the government may care not only about the rents received, but also about the actions undertaken by the franchisee. (In some cases, the rent extraction problem and the insurance problem are closely related: the average value of rents received may be increased if rents can be varied with the weather, the state of nature; again, we can think of decomposing the rent payment into a fixed rent and an insurance payment.)

This list of reasons for the origins of principal–agent relations is not meant to be exhaustive; yet many of the other reasons cited may be reduced to one of these explanations. For instance, consider the problem of a production line on which there are many workers; the output of the production line depends on all of their efforts. In the absence of risk aversion and credit problems, the incentive problem could be solved by giving each worker the total value of net output. He would purchase the right to the job by paying a fixed fee. With such a compensation scheme, the worker would have full incentives for maximizing the firm's output. But such a compensation scheme imposes on the worker an intolerable level of risk; and the fixed fee he would be required to pay necessitates his borrowing large amounts of money.

The basic principal – agent problem

In the standard principal–agent problem, one looks for that contract (compensation scheme) which maximizes the expected utility of the principal, given that (a) the agent will undertake the action(s) which maximizes his expected utility, given the compensation scheme; and (b) given that he must be willing to accept the contract.

The second set of constraints (which are nothing more than the standard reservation utility constraints) are sometimes referred to as the individual rationality constraints.

There are two standard mathematical formulations. One is a direct generalization of the insurance–moral hazard problem. There are a set of observable events, such as whether an accident occurs. The probability that an event *i* occurs is a function of the actions undertaken (effort at accident avoidance):

$$p_i = p_i(\mathbf{e}),$$

where **e** may be a vector. The wealth of the individual in state i, in the absence of insurance, is w_i, and with insurance it is y_i. Thus

$$h_i = y_i - w_i$$

is the net payment from (to) the insurance company (the principal) in state i. The expected utility of the insured (the agent) is then just

$$U = \sum_i U_i(y_i, \mathbf{e})\, p_i(\mathbf{e})$$

while that of the principal is

$$V = \sum_i V_i(h_i)\, p_i(\mathbf{e}).$$

$\{h_i\}$ is chosen to maximize V subject to $U \geq \bar{U}$.

Notice that the employer–employee relationship may be cast in this form: the observable events are the levels of output. Assume for simplicity, that we measure outputs in round numbers (say, bushels of wheat). Then state i refers to the number of bushels produced. p_i then is the probability that i bushels will be produced. Assume that the individual's wealth, apart from this contractual arrangement with his employer, is zero. Then y_i is the individual's pay if output is i. If the employer is risk neutral,

$$V_i(h_i) = qi - h_i = qi - y_i,$$

where q is the price of output (of a bushel of wheat), assumed to be independent of i.

Although the employer–employee relationship can be cast in this form, it is more naturally represented by a formulation in which the probabilities of the states (weather) are fixed, where the states are unobservable, but where what is affected by the employee is the output in each state.

We can represent this formally in the following way. Let S be a set of state variables (like weather)

observable to the agent. Let Q be a set of output variables (assumed observable to the principal and agent). And let A be a set of inputs (actions) by the agent assumed observable only by the agent. Then a compensation scheme is a payment from the principal to the agent which is a function of all variables that are observable to both the agent and principal.

$$Y = \varphi(Q)$$

The agent chooses his actions to maximize his expected utility which depends both on his income and his actions, given

$$\max EU(Y, A, S)$$

where outputs (actions), A, are related to the inputs by a production function

$$Q = Q(A, S)$$

We denote the solution to this by

$$A = H(S).$$

Finally, we can calculate the expected utility of the principal; his utility depends on the agent's actions, the payments he makes to the agent, and his state (the actions may affect the principal either directly, or via their effect on outputs, or via their effects on payments).

$$EV = EV(\varphi(Q), Q, A, S).$$

The principal's problem is to choose ϕ to maximize his expected utility

$$\max EV$$

recognizing the dependence of the agent's action on ϕ and recognizing that he must pay the agent enough to induce him to accept the job

$$EU \geq \bar{U}$$
(RU)

Pooling versus separating equilibrium

Much of the literature has focused on situations where the principal wishes to induce the agent to take different actions in different states. That is, in the simplest case where only output is observable by the principal, if $A^*(S)$ is the action desired in state S, then the compensation scheme must be such that

$$EU[\varphi(Q(A^*, S)), A^*, S] > EU[(Q(A, S)), A, S] \text{ for all feasible } A$$

These constraints are referred to as the self-selection or incentive compatibility constraints.
When the individual takes actions in two different states, so that the observable variables are the same, i. e. so that the principal cannot distinguish which of the two states has occurred, we say that there is a *pooling* equilibrium. When the individual takes actions so that the principal can identify which state has occurred, we say that there is a *separating* equilibrium. (This terminology was introduced within the context of the adverse selection literature by Rothschild and Stiglitz (1976).) A basic result of the principal–agent literature establishes conditions under which the optimal contract involves complete or partial separation.

Adverse selection

The variable S can be thought of as a characteristic of an individual, rather than as the state of nature. Then the self-selection constraint says that individuals of type S prefer action $A(S)$ to any other feasible action. If the self-selection constraints are satisfied, we can identify who is of what type. The action may consist of nothing more than making a choice. In the adverse selection interpretations of the model, the constraint (*RU*) needs to be replaced by the set of constraints,

$$U(\varphi(Q(A, S)), A, S) \geq \bar{U}(S), \text{ for all } S$$

that is, there is a reservation utility level for each individual (an individual rationality constraint for each type). (Note that a similar set of constraints is relevant if the contractual arrangement between the principal and agent is not binding, i.e. the individual can quit after he sees what the state of nature is.) Some examples follow.

1. (i) *The partially discriminating monopoly* (see, e.g. Salop, 1977; Stiglitz, 1977). The firm knows that different individuals have different indifference curves between the good he sells and other goods, and different reservation utility levels, but he does not know who is of which type. Q may be the quantity of some commodity chosen by an individual, in which case $\phi(Q)$ can be interpreted as the payment to the monopolist. (If one individual unambiguously has stronger preferences for the good, in the sense that at any quantity and payment, the extra amount he is willing to pay for a marginal unit is greater, then some separation is always desirable; this property is called the single crossing property.)
2. (ii) *Optimal tax structures* (Mirrlees, 1971). The government wishes to impose differential taxation on different individuals; it may want to impose a higher tax on the more able, but cannot tell who is the more able. Neither the individual's productivity nor the number of hours a week he works is observable, but his income is observable. The income tax schedule specifies a level of consumption corresponding to each level of income. The individual chooses (by the amount of work he undertakes) a point on that schedule. A schedule which results in the more able earning (choosing) higher incomes is one which separates. This will be desirable if the indifference curves between consumption and income are flatter for the more able – they require less of an increase in consumption to compensate for an increase in income. This will be true, for instance, if the underlying indifference curves between hours worked and consumption are the same for all individuals.
3. (iii) *Pareto efficient tax structures* (Stiglitz, 1982a). In the previous problem, the government maximized the sum of utilities, subject to the self-selection constraints, the revenue constraints, and the individual rationality constraints (which simply required that the individual desire to work). The revenue constraint was equivalent, in this problem, to the profits (revenues) of the landlord; that is, while in the landlord problem we maximize the revenue, subject to the expected utility of the individual satisfying a certain constraint, here the dual of this problem is analysed. The ‘sum of utilities’ is equivalent to ‘expected utility’ – where the probability of each state S is identical. We can directly generalize this by imposing constraints on the level of utility attained by all individuals other than the first; we then maximize the first individual's utility subject to these constraints (and subject to the self-selection constraints, and the revenue constraints). This is the problem of Pareto efficient taxation. It is equivalent to the problem of maximizing a weighted sum of individuals’ utilities.
4. (iv) *Implicit contracts with asymmetric information*. (For surveys, see Hart, 1983; Stiglitz, 1986; Azariadis and Stiglitz, 1983.) With perfect information, the employer would provide insurance to the employee, to stabilize the employee's income. If, for instance, the workers’ utility function was separable between hours worked, l, and income y

$$U = u(y) - v(l),$$

then with complete information, and risk neutral firms, *y*will be the same in all states, but *l*will be higher in states where labour productivity is higher. Thus, if the employer knew the state, but the worker did not, the employer would have an incentive always to say that it was a good state (since what he paid the worker was the same, but workers are required to work more in good states). The optimal contract will induce the employer to announce that it is bad when it is in fact bad, i.e. it will separate (at least partially).

Qualitative results

It is clear that many economic relationships fall within the scope of the ‘principal–agent’ model. Many of the basic qualitative results emerge from a detailed analysis of the insurance model:

1. (a) There is a risk-incentive trade-off; since the risks undertaken will be a function of the quantity of insurance purchased, if the latter is observable, the premium will depend on it, and in equilibrium, there will be quantity rationing, i.e., the individual would like to purchase more insurance, at the going benefit premium ratio (Pauly, 1968). The amount of insurance will be greater, the more risk averse the individual.
2. (b) Indifference curves (between benefits and premia) are not generally quasi-concave, nor feasibility sets (the set of insurance premia satisfying the non-negative profit constraint) convex ; this has important consequences for the existence of competitive equilibria. The amount of insurance purchased may not be a continuous function of the price of insurance; and the level of effort may not be a continuous function of the amount of insurance purchased.
3. (c) Competitive equilibrium, when it exists, will not in general be Pareto efficient (Arnott and Stiglitz, 1986; Greenwald and Stiglitz, 1986); the profits of one insurance firm are affected both by the terms at which other firms offer insurance contracts (whether for similar accidents or not), and by the prices at which goods (whether complements or substitutes for accident avoidance or accident inducing activities) are sold; there exist a set of Pareto improving subsidies and taxes. In some instances, firms may attempt to internalize some of these ‘externalities.’ This leads to interlinkage of markets, both across time (the same insurance firm insures the individual over time), and at the same time (the same insurance firm insures the individual for many different risks) (Braverman and Stiglitz, 1982). The frequently observed interlinkage between credit and land markets in less developed countries has been interpreted in this light.

Variants of the general model

Further results have been obtained for various variants of the general model. We discuss a few of the

more important versions below:

1. (i) Adverse selection model. The major qualitative results of this model (other than the specification of the conditions under which pooling or separation occurs, discussed above) entail an analysis of the distortions (relative to perfect information) engendered by the self-selection constraints; in the optimal income tax, the reduction in work (income) of the less able (associated with a positive marginal tax rate); in the asymmetric information implicit contract model, in the existence of overemployment in good states (with the separable utility function and risk neutral firms), or underemployment in bad states (with very risk averse firms). To discriminate among individuals, firms may engage in socially wasteful activities, such as random pricing or long queues. Generally, one group in the population (the most risk averse in the insurance model, the highest ability in the optimal income tax model) chooses a contract which does not distort its behaviour.
2. (ii) Incentive model with actions taken before state is known. When the random elements have bounded support, then a first best can be achieved simply by imposing a large enough penalty for performances below a given threshold. The individual will exert enough effort to avoid this. (See Mirrlees, 1974; Stiglitz, 1975.)
3. (iii) Theory of contests. If the output of others performing similar tasks in similar situations is observable, then one will employ compensation schemes based on relative performance; these will do better than individualistic compensation schemes. If there are enough individuals, simple schemes, based only on individuals' rankings, can approximate the first-best outcomes.
4. (iv) Models in which the utility constraint is not binding. In some cases, when the principal maximizes his expected utility, subject to the workers' reservation utility constraint, the latter constraint will not be binding. Such models give rise to unemployment. A particularly important variant of these models is described next.
5. (v) Models in which quality is affected by price. If the probability of default increases with the rate of interest charged (either because individuals undertake more risks when the interest rate is higher, or because those who are less risky stop applying for loans at high interest rates), then banks may not raise interest rates, even in the presence of an excess demand for loans. Similarly, if the productivity of a worker increases with the wage paid (either because individuals exert greater effort at higher wages or because those who are recruited at higher wages are more productive), then firms may not lower wages, even in the presence of an excess supply of labour.
6. (vi) Terminations. In multiperiod models, it has been shown that the optimal contract may entail the termination of a relationship when performance is unsatisfactory; this is shown to be preferable to the imposition of other penalties. (See Stiglitz and Weiss, 1983.)
7. (vii) Infinite period models. Long-term relationships may ameliorate some of the incentive problems (see Radner, 1981). Over an infinite lifetime, the principal (insurer) can make good inferences concerning the actions of the agent (insured); the relative frequency of accidents will converge to the accident probability corresponding to the individual's effort level. Not surprisingly, then, with low enough discount rates, incentive schemes can be designed which approximate the first best outcomes. The interpretation of this result is, however, subject to some controversy. Since with low discount rates, the change in lifetime income which would be associated with the individual bearing the full risk of the outcome for any period is negligible, it

is as if the individual is risk neutral; and with risk neutrality we know that first best optimum can be obtained (if bankruptcy is ignored).

The set of admissible contracts

One of the important and general results to emerge from the principal–agent literature is that the nature of the equilibrium contract depends on the set of admissible contracts. Contracts can depend only on the available information; typically, it is desirable to use all of the available information, though in practice, many variables which ought to be relevant (have information value) are not included within the compensation scheme.

Similarly, if one could costlessly implement a non-linear incentive scheme, such schemes would, in general, be preferable to linear schemes. Though in practice, again, most observed schemes seem relatively simple (linear, piece-wise linear, etc.), much of the literature has been concerned with characterizing in admittedly simple situations the optimal non-linear scheme.

In a variety of situations, if one could make pay a stochastic function, it would be desirable to do so, even with risk averse individuals. (Arnott and Stiglitz (1988) and Holmstrom (1979) show that with separable utility functions, this will not happen.) The intuition behind this, in the case when actions have to be taken prior to the agent obtaining information about the state is that the possibility that he receives a low compensation so induces him to work hard that the employer (landlord) can reduce the dependence (on average) of pay on output, and thus reduce the variability of income.

Though optimal schemes may thus appear to be fairly complex, in practice most schemes employed are relatively simple. There is an ongoing controversy between those who seek to consider increasingly complex schedules, dismissing work which has analysed simple linear schedules as ad hoc; and those who seek to explain the kinds of compensation schedules actually employed; these dismiss the complex solutions as being irrelevant. They would argue that efforts should be devoted to understanding why actually employed schemes take on the form they do.

One possible explanation of the use of simple schedules is that they may be more *robust*. That is, as technology changes or the probability distribution of states changes (the exogenous parameters in the principal–agent problem) the optimal compensation scheme changes. But in practice, revisions to compensation schemes are costly, and one must find a scheme that works under a variety of situations. Simple, linear schemes may possess this property of robustness.

Another important characteristic of the set of admissible schemes relates to commitments. Can, for instance, the worker commit himself not to leave, or can the employer commit himself not to terminate the relationship?

A closely related issue is the set of punishments (rewards) which are admissible. It makes a great deal of difference if there are limits on the negative compensations that can be provided in the presence of bad outcomes.

We have noted the role of observability in the design of contracts. In some cases an important distinction may arise between observability and verifiability. The question is associated with how a contract is to be enforced. If the contract is to be enforced through the courts, it must be the case that any violation can be verified by an outside third party. Both the principal and the agent might know that the contract has been violated i.e. they both may observe the S (and not S') has occurred, and therefore that the payment should be that corresponding to S (and not S'). But unless it can be proved, the principal might attempt

to cheat the agent. Knowing this, the agent would refuse to sign a contract based on unverifiable variables.
On the other hand, if the contract is enforced by a reputation mechanism, good behaviour may be enforced so long as the state is observable by both parties.

Concluding remark

We have focused here on a discussion of general principles. It should be emphasized, however, that the principal–agent model has provided important insights into the nature of a variety of economic relationships, in labour, land, credit, and product markets. These detailed applications of the general theory represent an important area of on-going research.

See Also

- incomplete contracts

Bibliography

Allen, F. 1981. The prevention of default. *Journal of Finance* 36, 271–76.

Arnott, R. and Stiglitz, J.E. 1982. The welfare economics of moral hazard. Discussion Paper No. 465, Queen's University, Kingston, Ontario, March.

Arnott, R. and Stiglitz, J.E. 1985. Labor turnover, wage structures, and moral hazard: the inefficiency of competitive markets. *Journal of Labor Economics* 3, 434–62.

Arnott, R. and Stiglitz, J.E. 1986. Moral hazard and optimal commodity taxation. *Journal of Public Economics* 29(1), 1–24.

Arnott, R. and Stiglitz, J.E. 1988. Randomization with asymmetric information: a simplified exposition. *Rand Journal of Economics* 19, 344–62.

Arrow, K.J. 1965. *Aspects of the Theory of Risk-Bearing*. Helsinki: Yrjö Jahnsson Foundation.

Azariadis, C. and Stiglitz, J.E. 1983. Implicit contracts and fixed price equilibria. *Quarterly Journal of Economics* 98(3), Supplement, 1–22.

Braverman, A. and Stiglitz, J.E. 1982. Sharecropping and the interlinking of agrarian markets. *American Economic Review* 72, 695–715.

Eaton, J. and Gersovitz, M. 1981. Debt with potential repudiation; theoretical and empirical analysis.

Review of Economic Studies 48, 289–309.

Fellingham, J.C., Kwon, Y.K. and Newman, D.P. 1982. Ex ante randomization in agency models. *Rand Journal of Economic Studies* 15, 290–301.

Gjesdal, F. 1982. Information and incentives: the agency information problem. *Review of Economic Studies* 49, 373–90.

Green, J. and Stokey, N. 1983. A comparison of tournaments and contests. *Journal of Political Economy* 91, 349–64.

Greenwald, B. and Stiglitz, J.E. 1986. Externalities in economies with imperfect information and incomplete markets. *Quarterly Journal of Economics* 101(4), 229–64.

Grossman, S. and Hart, O. 1983. An analysis of the principal–agent problem. *Econometrica* 51(1), 7–45.

Hart, O. 1983. Optimal labor contracts under asymmetric information: an introduction. *Review of Economic Studies* 50, 3–35.

Helpman, E. and Laffont, J.J. 1975. On moral hazard in general equilibrium. *Journal of Economic Theory* 10, 8–23.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13, 324–40.

Keeton, W.R. 1979. *Equilibrium Credit Rationing*. New York: Garland Publishing.

Lazear, E. and Rosen, S. 1981. Rank order tournaments as optimum labor contracts. *Journal of Political Economy* 89, 841–64.

Mirrlees, J. 1971. An exploration of the theory of optimum income taxation. *Review of Economic Studies* 38(2), 175–208.

Mirrlees, J. 1974. Notes on welfare economics, information, and uncertainty. In *Contributions to Economic Analysis*, ed. M.S. Balch, D.L. McFadden and S.Y. Wu. Amsterdam: North-Holland.

Mirrlees, J. 1976. The optimal structure of incentives and authority within an organization. *Bell Journal of Economics* 7(1), 105–31.

Nalebuff, B. and Stiglitz, J.E. 1983a. Information, competition and markets. *American Economic Review*

72, 278–84.

Nalebuff, B. and Stiglitz, J.E. 1983b. Prizes and incentives: towards a general theory of compensation and competition. *Bell Journal of Economics* 14, 21–43.

Pauly, M.V. 1968. The economics of moral hazard: comment. *American Economic Review* 58, 531–6.

Radner, R. 1981. Monitoring cooperative agreements in a repeated principal–agent relationship. *Econometrica* 49, 1127–48.

Radner, R. and Stiglitz, J.E. 1983. A nonconcavity in the value of innovation. In *Bayesian Models in Economic Theory*, ed. M. Boyer and R. Kihlstrom. Amsterdam: North-Holland.

Ross, S. 1973. The economic theory of agency: the principal's problem. *American Economic Review* 63, 134–9.

Rothschild, M. and Stiglitz, J.E. 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90, 629–49.

Salop, S. 1977. The noisy monopolist: imperfect information, price dispersion and price discrimination. *Review of Economic Studies* 44, 393–406.

Salop, S. and Salop, J. 1976. Self-selection and turnover in the labor market. *Quarterly Journal of Economics* 90, 619–28.

Sappington, D. and Stiglitz, J.E. 1987. Information and regulation. In *Public Regulation: New Perspectives on Institutions and Policies*, ed. E. Bailey. Cambridge, MA: MIT Press.

Spence, A.M. and Zeckhauser, R. 1971. Insurance, information, and individual action. *American Economic Review* 61, 380–7.

Stiglitz, J.E. 1974. Incentives and risk sharing in sharecropping. *Review of Economic Studies* 41, 219–55.

Stiglitz, J.E. 1975. Incentives, risk, and information: notes toward a theory of hierarchy. *Bell Journal of Economics* 6, 552–79.

Stiglitz, J.E. 1977. Monopoly non-linear pricing and imperfect information: the insurance market. *Review of Economic Studies* 44, 407–30.

Stiglitz, J.E. 1982a. Self-selection and Pareto efficient taxation. *Journal of Public Economics* 17, 213–40.

Stiglitz, J.E. 1982b. Utilitarianism and horizontal equity: the case for random taxation. *Journal of Public Economics* 18, 1–33.

Stiglitz, J.E. 1986. Theories of wage rigidity. In *Keynes' Economic Legacy*, ed. J. Butkiewicz, K. Koford and J. Miller. New York: Praeger Publishers, 153–221.

Stiglitz, J.E. 1987. On the causes and consequences of the dependence of quality on price. *Journal of Economic Literature* 271–48.

Stiglitz, J.E. and Weiss, A. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

Stiglitz, J.E. and Weiss, A. 1983. Incentive effects of terminations: applications to the credit and labor markets. *American Economic Review* 73, 912–27.

Weiss, L. 1976. On the desirability of cheating, incentives and randomness in the optimal income tax. *Journal of Political Economy* 84, 1343–52.

How to cite this article

Stiglitz, Joseph E. "principal and agent (ii)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000183> doi:10.1057/9780230226203.1342

The New Palgrave Dictionary of Economics Online

privatization impacts in transition economies

Saul Estrin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article addresses the large-scale privatization processes in central and eastern Europe. It explains why reformers placed such emphasis on privatization and the practical problems posed by the scale of state ownership under communism, leading to the widespread use of mass privatization. As a result ownership changes were huge and extremely rapid but the improvement in corporate governance was more questionable. The empirical findings about the impact on enterprise performance are patchy, though on balance the effect has been positive, especially in countries with stronger institutions or where the new owners have been foreigners.

Keywords

agency chains; asymmetric information; capital market; company performance; concentration of ownership; corporate governance; corruption; foreign direct investment; free rider problem; growth; insider ownership; intermediate agents; management–employee buyouts; monitor; opportunity costs; outsider ownership; ownership and control; privatization; privatization in transition economies; privatization vouchers; productive efficiency; property rights; rent seeking; restitution; soft budget constraint; takeovers; transition economies

Article

Privatization is the process whereby the ownership of the state's productive assets, often utilities or large industrial enterprises, is transferred into private hands. This has been a major activity for governments in both the developed and the developing worlds since Prime Minister Thatcher's first modern privatization programme in the UK between 1979 and 1984. The cumulative revenues raised from the process globally probably exceeds $1.25 trillion dollars, while the role of state-owned enterprises in the economies of high income countries has declined from around 8.5 per cent GDP on average in 1984 to around six per cent in 1991 and probably below five per cent in 2005 (see Megginson, 2005). The

reduction in state ownership has probably been even more dramatic in less developed countries, from around 16 per cent GDP in 1981 to around five per cent in 2004. Privatization is intended to improve corporate efficiency and generate revenues for the state, and there is now probably sufficient experience in different economic and institutional environments to evaluate its impact relative to expectations.

Privatization has been a particularly important phenomenon in the transition process in central and eastern Europe from planning to a market system. This is because Communist regimes had placed almost all the productive assets of the economy in state hands for ideological reasons, and to facilitate the planning process. As a result, countries like Czechoslovakia and the Soviet Union contained virtually no private sector at all – typically in excess of 90 per cent of assets were state owned – and even in countries with slightly larger private sectors, like Poland or Hungary, private ownership was concentrated in agricultural and handicraft activities; industrial firms were all in state hands. This meant that privatization was a central aspect of building a market economy in all the transition economies. Indeed, to quote Dusan Triska in 1992, ‘privatization is not just one of the many items on the economic program. It is the transformation itself’ (see Estrin, 2002).

This article addresses the privatization process in central and eastern Europe, focusing on the objectives, the methods and, most importantly, the impact of the ownership changes. Privatization always had some ideological content in the transition economies, especially in the early years, when the reformers wished to create a ‘capitalist class’ supporting the radical changes that were required to build a market economy. But the fundamental objective of privatization in transition economies, as in developed and developing ones, has been to enhance company performance. We enquire whether privatization has succeeded in this objective in the remainder of this article.

Why privatize?

We begin by identifying why reformers in the transition economies placed such emphasis on privatization. Transforming state-owned assets into private hands can improve corporate efficiency (see Vickers and Yarrow, 1985), and, particularly with the privatization of infrastructure, the benefits can spill over to the rest of the economy. To understand why, one must compare company objectives and corporate governance under state and private ownership. It is normally argued that the fundamental difference between state-owned and private firms rests in their objectives: the latter focus exclusively on profit, which generates close attention to costs and to the demands of customers. State-owned firms may be interested in profits too, but they will almost certainly be expected by their owners to satisfy other objectives as well, for example, politically determined targets such as creating or maintaining employment in economically depressed regions or holding prices below average costs for redistributive reasons. In this situation, profits become a secondary criterion, or indeed an irrelevance, and business decisions become politicized. Inefficiencies can thrive because they are not a central concern of the owner, and managers can exploit the lack of clarity in company objectives to ensure an easy life for themselves and employees (see Shleifer and Vishny, 1994).

Therefore, an important motive for privatization is to focus attention on profits as the sole objective for the enterprise sector. But the problems of state ownership go beyond just diffuse and non-commercial objectives. In a socialist economy, the system of administered prices also means that privatization and market liberalization are needed to reveal opportunity costs. Moreover, even in a market economy, when a public-sector firm operates in a competitive market and the government tries to enforce an objective of

profit maximization on its management, weaknesses in corporate governance can still cause inferior performance to what might be achieved under private ownership. The problem is centred on the asymmetry of information held by managers and owners; outside owners – private or state – can never have full access to the information about corporate performance that is in the hands of managers. Thus, it is hard for them to establish whether poor results are a consequence of unforeseen circumstances or managers exploiting firm profits for their own purposes. Whenever ownership and control are separated, firm-specific rents can be used to satisfy management's aim – for example, lower effort or managerial power, via the size of the firm – rather than profits. However, a private ownership system places more effective limits than does state ownership on their discretionary behaviour, via external constraints from product and capital markets which largely operate through the market for corporate control, and through the internal constraints imposed via statutes and monitoring by the owners themselves (see Estrin and Perotin, 1991).

In Anglo-Saxon countries, the constraints on managerial discretion in large part derive from stock markets (see Megginson, 2005). The quality of managerial decision-making and the extent of managerial discretion are an input in the choices of traders in equity markets, whose judgement on company performance is summarized in the share price. If the managerial team is thought to be incompetent or inefficient, the share prices will be reduced, putting pressure on managers to improve their performance. A persistently poor showing by a quoted company may also generate external pressure by encouraging a takeover bid. In this case, the stock market can be viewed as a market for corporate control, with alternative teams vying for the right to manage the enterprise. However, the effectiveness of these disciplines relies to some extent on the concentration of ownership. If ownership is highly dispersed, each individual owner has only a slight incentive to monitor effectively, and as there is a free rider problem monitoring may be inadequate.

Governance also comes from the way that the managerial market operates, with managerial performance, pay and job prospects assessed by movements in share prices. Payment mechanisms such as management stock option schemes can also be put in place to align the incentives of owners and managers. In countries such as Japan or Germany, however, the mechanisms can be different, with less reliance on an adversarial market for corporate control and more extensive use of internal governance constraints. Ownership is typically highly concentrated in the hands of banks, funds or families who are granted board representation and undertake close monitoring of managerial performance directly, and use the managerial market and management incentive schemes.

Either way, it is hard for the state to imitate these market-based constraints. State-owned firms are not subject to private capital market disciplines, so neither the competitively driven informational structure nor the market-based governance mechanisms can be substituted for in full. State employees are usually civil servants and do not compete in the wider managerial market, though Western governments have recently tried to reduce the labour market segmentation between the public and private sectors. Moreover, though the government's ownership stake is concentrated, the state is rarely directly represented on the boards of public sector companies and usually does not have the capacity in the supervisory ministries to undertake the necessary scale and quality of monitoring (see Vickers and Yarrow, 1985).

These arguments have particular resonance in the transition economies of central and eastern Europe. The economic problems of the socialist system were largely a result of the impact of state ownership and

planning on investment allocation, incentives and efficiency (see Gregory and Stuart, 2004). Firms did not attempt to maximize profits, and productive efficiency was a low priority. Instead, weak monitoring of managers by the state as owner and the absence of external constraints gave management almost total discretion to follow their own objectives – rent absorption, asset stripping, employment, social targets. The softness of budget constraints (Kornai, 1990) that goes with the political determination of resource allocation was a further source of incentive problems, since managers did not have to bear the consequences of their own actions. Mistakes were condoned and losses were subsidized.

Methods of privatization in transition economies

It is therefore clear why privatization was so important in the transition process. Nonetheless, reforming governments might in principle have left privatization until the track records of particular firms in the market environment had become firmly established and until the stock of domestic savings in private hands was sufficient to ensure the success of a competitive bidding process for the assets. But the state was probably not able to manage its assets effectively in the intervening period, and managers and workers began very rapidly to steal the assets (Canning and Hare, 1994). The collapse of communism had left state-owned firms with limited internal structure to handle the new requirements of the market-place and no mechanisms to monitor or enforce governance on state-owned firms (see Blanchard et al., 1991). The authorities had either quickly to create structures whereby the state as owner could control enterprise decisions or face a gradual dissipation of the net worth of the enterprise sector by consumption, waste or theft. These stark alternatives persuaded many reforming governments and their Western advisors to consider rapid privatization. (Boycko, Shleifer and Vishny, 1995 – the first of these an insider to Russia policymaking at the time – make a similar point concerning Russia. They argue that Russia had to undertake a massive and speedy ownership change in order to break the tradition of rent-seeking behaviour and the long-standing links between the state and the enterprise sector. They argue that, in order to gain political support for the privatization process, substantial stakes had to be given to insiders – the managers and workers in firms – so that they did not block the process.)

The sheer scale of privatization required in the transition economies posed considerable practical problems. As we have seen, the Communist heritage meant that the majority of firms in the economy needed to be privatized. At the aggregate level, the stock of domestic private savings in these countries was too small to purchase the assets being offered. This led the reformers to innovate with privatization methods.

For selected firms, many transition economies used auction or public tender, as have been the norm in the West. Such sales could in principle be to domestic or foreign purchasers but, in practice, only Hungary and Estonia were willing or able to sell an appreciable share of former state-owned assets to foreigners. Foreign capital ended up purchasing about 20 per cent of the privatized assets in Hungary and up to 50 per cent in Estonia, but even in these countries the preponderance of foreign ownership gave rise to public disquiet. Moreover, foreign direct investment flows to the transition economies were modest in the early years, when privatization was taking place, and were highly concentrated towards the Czech Republic, Hungary and Poland (United Nations, 2004; Meyer, 1998). In practice, sales of state-owned enterprises have mainly been to a country's own citizens: either to external capital owners or to insider management–employee buyouts. Managers and employees were the more common initial

buyers, perhaps because they had insider knowledge about their company's business prospects. Some governments, such as in Romania, actively encouraged the emergence of insider-owned firms.
Some countries also experimented with restitution to former owners; the former East Germany, Hungary, the former Czechoslovakia and Bulgaria are prominent examples. Restitution has the advantage that it immediately creates a property-owning middle class and re-establishes ‘real ownership’. However, the process of restitution entails legal complexities. For example, suppose that a factory has been built on a plot of land formerly owned by a noble. Does the noble receive the land back, and therefore rental for the factory? Or should the noble be compensated for the value of the property at the time of its seizure and, if so, how is such an evaluation to be made some 80 years later? Restitution also raises the deep question of how the assets accumulated during the Communist era, when consumption levels were held down for national capital accumulation, should be distributed. Since the burden of lower consumption was imposed on everyone, the argument that the distribution of the resulting assets should be egalitarian has been a powerful one.
To increase the pace of privatization, a number of transition countries began to experiment with ‘mass privatization’. This entails placing into private hands nominal assets of a value sufficient to purchase the state firms to be privatized. To avoid the inflationary consequences of such wide-scale ‘money’ creation, the new assets must be non-transferable and not valid for any transaction other than the purchase of state assets. This was largely achieved using the instrument of privatization vouchers or certificates. It was hoped that any deficiencies in the resulting corporate governance mechanism arising from the fact that the ownership structure was initially diffuse would be addressed by capital market pressures leading to increased ownership concentration (Boycko, Shleifer and Vishny, 1995).
Mass privatization has been carried out in a number of different ways, but the differences can be summarized around two issues. The first was whether the vouchers or certificates were distributed on an egalitarian basis to the population as a whole or whether, as in Russia and many other countries of the former Soviet Union, management and employee groups received many of the shares, perhaps to diffuse potential opposition to privatization. Second, policymakers needed to determine whether vouchers were intended to be exchanged directly for shares in companies, or whether the vouchers should be in funds that own a number of different companies. In the Czech and Slovak republics and in Russia, vouchers were exchanged directly for shares, although financial intermediaries soon developed in the market. In the Polish scheme, vouchers were exchanged for shares in government-created funds that jointly owned former state-owned enterprises.
Every country used a variety of privatization methods and everywhere different sorts of firms were sold in different ways. For example, in most transition economies small firms were usually sold to the highest bidder, and utilities were often floated on stock markets. However, it was possible by the time the bulk of privatization was completed in the late 1990s to discern the predominant method used in each country, and we report the most widely used summary in Table 1 from the EBRD's *Transition Repor*t, 1998. Mass privatization was the most common privatization method across the transition economies; 19 of the 25 countries listed used some form of mass privatization as either a primary or secondary method. Moreover, management–employee buyouts (MEBOs) also proved important, perhaps because transition governments sometimes did not have the authority to take on entrenched insiders in firms. Thus, nine countries used MEBOs as their primary method, and six as their secondary method. Most transition economies therefore eschewed the conventional method of privatization, by direct sale. In fact, only five countries used this as their primary privatization method, though these were among the most developed

transitional economies.

Methods of privatization

	Primary method			Secondary method		
Country	Direct sales	MEBOs*	Vouchers	Direct sales	MEBOs*	Vouchers
Albania		+				+
Armenia			+		+	
Azerbaijan			+	+		
Belarus		+				+
Bulgaria	+					+
Croatia		+				+
Czech Republic			+	+		
Estonia	+					+
FYR Macedonia		+		+		
Georgia			+	+		
Hungary	+				+	
Kazakhstan			+	+		
Kyrgyzstan			+		+	
Latvia			+	+		
Lithuania			+	+		
Moldova			+	+		
Poland	+				+	
Romania		+		+		
Russia			+	+		
Slovak Republic	+					+
Slovenia		+				+
Tajikistan		+				+
Turkmenistan		+		+		
Ukraine			+		+	
Uzbekistan		+		+		

*Management-employee buyouts.

•*Source*: EBRD (1998).

The scale of privatization

There was an extremely speedy ownership change in most transition economies. Few countries had

contained a private sector of any significance in 1990. Exceptions were Hungary and Poland, where there had been long-standing private firms in agriculture and crafts, and the private sector already represented over 30 per cent of GDP (see Estrin, 1994). But in the transition economies as a whole the private sector contribution to GDP was usually less than 20 per cent. The growth in the private sector share during the 1990s, reported in Table 2, is extraordinary. As early as 1995, the private sector share was above 50 per cent in nine countries, though in eight former republics of the Soviet Union it remained below 30 per cent. By 2002, the private sector in 13 additional nations had reached at least 50 per cent of GDP and in only two laggards, Belarus and Turkmenistan, was private sector activity still below 25 per cent of GDP. Thus the privatization process in the transition economies was in many countries effective in transferring the bulk of economic activity from state to private hands in the space of hardly more than a decade.

Private sector percentage shares in GDP and employment, 1991–2002

	In GDP			In employment		
	1991	1995	2002	1991	1995	2001
Albania	24	60	75	—	74	82
Armenia	—	45	70	29	49	—
Azerbaijan	—	25	60	—	43	—
Belarus	7	15	25	2	7	—
Bosnia and Herzegovina	—	—	45	—	—	—
Bulgaria	17	50	75	10	41	81
Croatia	25	40	60	22	48	—
Czech Republic	17	70	80	19	57	70
Estonia	18	65	80	11	—	—
Fyr Macedonia	—	40	60	—	—	—
Georgia	27	30	65	25	—	—
Hungary	33	60	80	—	71	—
Kazakhstan	12	25	65	5	—	75
Kyrgyz Republic	—	40	65	—	69	79
Latvia	—	55	70	12	60	73
Lithuania	15	65	75	16	—	—
Moldova	—	30	50	36	—	—
Poland	45	60	75	51	61	72
Romania	24	45	65	34	51	75
Russia	10	55	70	5	—	—
Serbia and Montenegro	—	—	45	—	—	—
Slovak Republic	—	60	80	13	60	75

Slovenia	16	50	65	18	48	—
Tajikistan	—	25	50	—	53	63
Turkmenistan	—	15	25	—	—	—
Ukraine	8	45	65	—	—	—
Uzbekistan	—	30	45	—	—	—
Means	20	44	62			

Source: EBRD (1999; 2003).

This remarkable performance should not conceal real concerns raised at the time about the quality of privatization, and therefore about its consequences for enterprise restructuring. First, there are questions about how real the privatization has been. In many transition economies, the state continued to own golden shares or significant shareholdings in companies. For example, the Russian state retained more than a 20 per cent share in 37 per cent of privatized firms, and kept more than a 40 per cent share in 14 per cent of the firms that it privatized. Only in half of privatized firms did the Russian government sell its entire holding. Thus, the clean break between the state as owner and the enterprise sector has perhaps been more notional than real. In a survey of privatized firms undertaken by the EBRD in 1999, reported in Table 3, we show that in 20 of the 23 countries the state has retained some shares post-privatization. On average, the state retained some shares in around 20 per cent of privatized firms, with more than a 20 per cent shareholding in around 12 per cent of the firms. It is suggestive that retained state shareholdings are negligible in some of the leading transition economies – for example, the Czech Republic, Hungary and Latvia – but the state has tended to keep a larger share in less advanced transition economies: more than 15 per cent of privatized firms in Albania, Belarus, Georgia, Lithuania, Poland, Romania, Russia and Ukraine, and more than 30 per cent of privatized firms in Bulgaria, Croatia, Slovenia and Uzbekistan. State ownership has also been retained in many developed OECD economies, including via the use of ‘golden shares’. According to Bortolotti and Faccio (2006), governments were actually the largest stakeholder or held special control powers (golden shares) in 62.4 per cent of privatized OECD companies.

Percentage of privatized firms with retained state shareholdings

	Percentage of shares retained by the state		
Country	0%	1–30%	>30%
Albania	83.9	0	16.2
Armenia	97.1	2.9	0
Azerbaijan	94.1	5.9	0
Belarus	80.4	10.7	8.9
Bulgaria	30.8	61.6	7.7
Croatia	59.1	33.4	7.6
Czech Republic	100	0	0
Estonia	92.3	3.9	3.9
Georgia	79.3	7	13.8

Hungary	100	0	0
Kazakhstan	93.6	2.1	4.2
Kyrgyz Republic	91.2	1.8	7.1
Latvia	100	0	0
Lithuania	80.8	19.3	0
Macedonia (FYR)	92.9	0	7.1
Moldova	87.7	5.4	7.1
Poland	71.7	13.3	15.1
Romania	80	13.4	6.7
Russia	82.6	8.7	8.7
Slovak republic	92.3	7.7	0
Slovenia	63	24.1	13
Ukraine	83.6	9.6	6.8
Uzbekistan	69.2	27	3.9
Total	80.9	11.8	6.3

Source: Unpublished EBRD survey, used by Bennett, Estrin and Urga (2007).

But widespread retained state ownership is not the only indication that privatization may not have ensured the establishment of effective corporate governance mechanisms in transition economies. The long ‘agency chains’ implicit in mass privatization may not provide appropriate incentives for corporate governance. Voucher privatization led to ownership structures that were highly dispersed (Coffee, 1996). Typically the entire adult population of the country, or all insiders to each firm, were allocated vouchers with which to purchase the shares of the company. The desire for equitable and politically acceptable outcomes dominated the need to create concentrated external owners who would have a large enough stake to be motivated to maintain oversight of management. However, it was possible that financial intermediaries could aggregate individual voucher holdings and carry out effective monitoring of management, and in Czech Republic, Poland, Slovenia and Slovakia some effort was made to ensure such concentrated intermediate agents did emerge. This was often associated with fraud and the outright theft of assets by managers to avoid their use by the new owners – so-called ‘tunnelling’ (Johnson et al., 2000).

The way that mass privatization was carried out in many countries also sometimes led to majority ownership that was not best suited to accelerate restructuring, for example by insiders. This was probably largely for political reasons, especially in countries where the pro-reform forces were politically weak. According to Earle, Estrin and Leschenko (1996), insiders held a majority shareholding in 75 per cent of firms in Russia immediately post-privatization (1994) and outsiders only nine per cent. Insider ownership was predominantly in the hands of workers. However, this created little problem for management because worker ownership was so highly dispersed. Indeed Blasi, Kroumova and Kruse (1997) argue that control was effectively in the hands of management in Russian employee-owned firms. Outsider ownership is also typically highly dispersed, with much of it in the hands of banks, suppliers,

other firms and an assortment of investment funds. In Russia, it appears from a variety of studies (see Estrin and Wright, 1999, for a survey) that outside shareholding has increased at the expense of the state and insiders during the 1990s, but ownership is also becoming increasingly dispersed and the greater degree of outside ownership may largely represent the fact that former insider voucher owners have left the firm but retained their shares.

This pattern of extensive employee ownership seems broadly consistent with the evidence for other CIS countries. In Ukraine, insiders owned 51 per cent of shares in all privatized firms in 1997 – managers eight per cent and workers 43 per cent – while outsiders held 38 per cent and the state residue share was 11 per cent. In Ukraine, insiders have actually increased their shareholdings, while managers have been buying shares from workers. Thus, rather than evolving towards the structure of firms owned by a concentrated group of outsiders, as was hoped by reformers, enterprises in the CIS appear to have remained primarily owned by dispersed groups of employees or outsiders. However, the situation appears to have been somewhat different in central Europe, where many of the most important firms in the economies are now quoted on the relevant national stock exchanges or owned by large foreign firms – for example, Skoda and Volkswagen. As we have seen, foreign ownership was predominant in Hungary, ownership by new entrepreneurs was common in Poland, while investment-fund ownership predominated in the Czech Republic.

The impact of privatization

In this section, we analyse the impact of privatization on economic and company performance in the transition economies. This can be considered from the macro-economic and the microeconomic sides, and we provide some information on both. We start by considering the effects on government resources, and exploring the relationship between private sector shares, privatization methods and revenues and economic growth. We then summarize the findings of the very large literature about the effects of privatization on company performance.

In Table 4, we present the cumulative revenues from privatization in each of the transition economies, from 1995 to 2002. The sums were relatively modest in most countries in 1995; cumulative revenues were less than two per cent of GDP in 15 countries of the 23 covered, and exceeded 20 per cent in only one country, namely, Hungary. The situation had changed appreciably by 2002. Cumulative revenues from privatization exceeded five per cent of GDP in 14 countries, exceeded ten per cent of GDP in eight and were greater than 30 per cent in Hungary and Slovakia. Thus, even in countries which used mass privatization the selling of state assets proved to be a significant source of government revenue through the financially demanding period of early transition, and may therefore have contributed to macro-stability and growth. However, there is no empirical evidence linking growth to the private sector share. Bennett, Estrin and Urga (2007) explore the impact of privatization on growth, but, while they identify a positive effect from the use of the mass privatization method, they do not find any significant relationship between growth and the private sector share. There is a limited amount of academic work for other economies, which explores the impact of privatization on growth rates. In an early study, Plane (1997) looks at the effects of divestiture on growth in a sample of 35 developing countries. He also controls for the problem of reverse causality by identifying separately the factors that determine a successful privatization programme. He finds that the impact of privatization on economic growth is

indeed positive, and is strengthened when privatization occurs in infrastructure or in industrial sectors., Zinnes, Eilat and Sachs (2001) use a fairly short sample period to undertake an aggregate growth study for the transition economies. They conclude that, while privatization does not actually increase growth, there is a positive impact when the privatization process is accompanied by institutional reforms.

Privatization revenues (cumulative, in percentage of GDP), 1995–2002

	1995	1996	1997	1998	1999	2000	2001	2002
Albania	3.1	3.3	3.6	3.6	3.9	7.0	9.1	9.1
Armenia	3.4	3.4	3.4	5.6	6.7	8.8	9.4	9.7
Azerbaijan	0.0	0.1	0.3	0.9	1.5	1.7	2.0	2.4
Belarus	0.5	0.7	0.9	1.0	1.1	1.1	1.2	2.9
Bosnia and Herzegovina	0.0	0.0	0.0	0.0	0.7	2.0	2.8	2.9
Bulgaria	0.7	1.5	4.6	6.2	8.4	9.7	10.3	11.2
Croatia	0.9	1.4	2.0	3.6	8.2	10.2	13.5	15.8
Czech Republic	4.6	6.3	7.1	7.9	9.3	10.3	13.1	18.7
Estonia	0.0	0.0	0.2	0.3	4.2	5.2	7.2	7.6
Georgia	19.1	19.8	20.5	21.8	22.7	23.0	23.1	—
Hungary	20.8	23.4	27.5	28.6	29.8	30.2	30.6	30.6
Kazakhstan	3.7	5.9	9.2	13.0	14.8	15.6	16.1	16.6
Kyrgyz Republic	0.9	1.3	1.4	1.6	1.9	2.1	2.5	2.7
Latvia	0.7	0.8	2.2	3.3	3.5	4.1	4.7	5.4
Lithuania	1.4	1.4	1.6	6.8	8.0	9.8	10.8	11.3
Moldova	0.8	1.3	3.6	4.4	5.4	11.1	11.1	—
Poland	2.6	3.6	5.1	6.4	7.7	11.4	12.2	12.6
Romania	1.2	2.2	4.6	6.4	7.6	8.2	8.9	9.0
Russia	1.5	1.7	2.7	3.4	3.5	3.8	4.2	4.5
Slovak Republic	8.4	10.2	10.8	11.5	11.8	16.3	20.1	35.1
Slovenia	0.4	0.9	1.4	2.2	2.5	2.5	2.7	4.9
Tajikistan	1.5	1.7	2.3	2.8	3.6	4.6	4.8	5.8
Turkmenistan	0.2	0.2	0.2	0.2	0.3	0.6	0.6	0.6

Source: EBRD (2004).

To turn to the microeconomic evidence, there have been a large number of studies of how privatization affects the performance of firms in transition economies. The most complete of these is by Djankov and Murrell (2002), which surveys the findings of more than 100 empirical studies of transition economies and uses a meta-analysis of the results to draw conclusions. Despite the plethora of material, the overall findings remain ambiguous. This is partly because the studies employ a variety of data-sets, measurements and methods which produce contradictory results. For example, there are many ways of

measuring company performance, including profitability, productivity, sales growth, export growth, and restructuring; and their findings differ. To begin with productivity, there is a wide variance in results across countries and samples, with private ownership found to yield positive, zero or negative effects. There is, however, convincing evidence that sale to foreign owners yields a positive effect, and that privatization is more likely to improve performance in central Europe than in the former Soviet Union. More recent literature strongly confirms the results with respect to foreign direct investment (for example, Sabirianova, Svejnar and Terrell, 2005). However, it is harder to discern positive effects from privatization when profitability is the performance measure, though once again some studies find a positive impact when the firm is sold to a foreign owner, and very few studies isolate a positive significant effect of privatization on revenues. There are fewer studies of the impact of privatization on exports, and these tend to be positive, especially when foreigners take over the former state-owned firm, and restructuring activity seems to have been significant in privatized firms in central Europe, but not in Russia, Ukraine and other countries of the former Soviet Union.

The variation in results is not merely a consequence of the wide variety of measures and countries with which the effects of privatization have been tested. Some serious methodological problems bedevil work of this sort, most importantly that of selection. This is the situation when firms with particular characteristics – for example, superior performance – were systematically chosen for privatization. In such a case, while one observes what appears to be superior performance among firms that have been privatized, the correct interpretation is not that privatization enhances performance but that it was the better firms that were chosen for privatization. The converse applies if the state chooses to keep the best firms for itself and to sell only the less productive ones; in this case, privatization will appear to lead to worse performance. Unfortunately, very few studies of privatization in the transition economies have been able to do much to address this problem of reverse causality. The data-sets upon which the empirical work has been based have been small and usually derived from sample survey questionnaires that did not contain sufficient information to control for the selection problem.

Even so, Djankov and Murrell (2002) conclude on the basis of the weight of the evidence that the impact of privatization on company performance has probably been positive and significant, though not in every circumstance. Two factors are usually cited as being particularly influential in determining whether privatization acts to enhance company performance. The first is the nature and characteristics of the new private owners. We noted that foreign owners lead to an improvement on most measures of performance. There is also some evidence, though it is less convincing, that sale to domestic private owners also improves performance, though it can be important for the ownership shares to be concentrated. However, there is almost no evidence that company performance is improved when firms are sold to insiders, either managers or workers. This is probably because insiders have exploited their control to resist the changes in behaviour required to make firms competitive in the market environment, rather than to promote them. We observed above that insider ownership was a fairly common phenomenon, especially in the former Soviet Union, and this probably goes some way to explain why economic performance in many of those countries was weaker than in, for example, much of central Europe, such as Hungary, Poland and the Czech Republic, where foreign direct investment flows were much greater.

The second factor is the institutional and business environment in which privatization takes place. We noted above that privatization relies on improved corporate governance, but that in turn depends on a competitive market environment and the enforcement of property rights. In countries where the legal

system is not functioning effectively, and businesses face high level of corruption and weak standards of financial discipline, it is hard to imagine how private ownership on its own might be expected to improve company performance. For example, sharper performance is meant to come from tighter financial disciplines to eradicate waste and reduce cost, but these will not bind in situations when budget constraints remain soft, as occurred post-privatization in many countries of the former Soviet Union, with firms financing their deficits not through direct government subsidies but by not paying their bills, especially to their workers, to the government in taxes, and to the state-owned utility companies. These two limiting factors affect some privatizations in all transition economies, but on average have been more likely to pertain in the economies of the former Soviet Union than those of central and eastern Europe. Thus while the macroeconomic work suggests a clear positive impact from privatization on economic growth, the results from the microeconomic literature are more modulated. The positive effects from privatization are found not to be automatic. They depend on to whom the firm was sold – foreigners, outsiders or insiders – and on the broader business environment in which the firm operates. The latter in particular tends to be better in central Europe and especially in the new accession economies to the European Union. Privatization methods may also have played an important role (see Bennett, Estrin and Urga, 2007).

Conclusion

The most impressive feature of privatization in the transition economies has been the speed and scale at which it occurred. The reforming governments of the late 1980s and early 1990s managed successfully to transfer the huge state-owned sector into largely private hands in a time period of hardly more than a decade, and to do so they had to use innovative privatization methods. However, this led them to introduce private ownership into situations where other crucial aspects of the business environment were not yet sufficiently developed to support the private economy. We find that privatization appears to have provided governments with much-needed revenues. However, at the enterprise level the results on performance are more patchy, though on balance the effects of privatization have probably been positive, especially when the new owners were foreigners. The most serious problem for privatization as a policy has been its use in a weak legal and institutional environment. In such cases, it rarely appears to have improved company performance.

See Also

- corporate governance
- privatization
- transition and institutions

Bibliography

Bennett, J., Estrin, S. and Urga, G. 2007. Privatization and economic growth in transition economies. *Economics of Transition* 15, 661–83.

Bevan, A. and Estrin, S. 2004. The determinants of foreign direct investment into European transition economies. *Journal of Comparative Economics* 32, 775–87.

Blanchard, O., Dornbusch, R., Krugman, P., Layard, R. and Summers, L.H. 1991. *Reform in Eastern Europe*. Cambridge, MA: MIT Press.

Blasi, J., Kroumova, M. and Kruse, D. 1997. *Kremlin Capitalism: Privatizing the Russian Economy*. Ithaca, NY: Cornell University Press.

Bortolotti, B. and Faccio, M. 2006. Reluctant privatization. Working Paper No. 40/2004, ECGI.

Boycko, M., Shleifer, A. and Vishny, R. 1995. *Privatizing Russia*. Cambridge, MA: MIT Press.

Canning, A. and Hare, P. 1994. Hungary. In Estrin (1994).

Coffee, J. 1996. Institutional investors in transitional economies: lessons from the Czech experience. In *Corporate Governance in Central Europe and Russia*, vol. 1, ed. R. Frydman, C. Gray and A. Rapaczynski. Budapest: Central European University Press.

Djankov, S. and Murrell, P. 2002. Enterprise restructuring in transition: a qualitative survey. *Journal of Economic Literature* 40, 739–93.

Earle, J., Estrin, S. and Leschenko, L. 1996. Ownership structures, patterns of control and enterprise behaviour in Russia. In *Enterprise Restructuring and Economic Policy in Russia*, ed. S. Commander, Q. Fan and M. Schaffer. Washington, DC: Economic Development Institute and World Bank.

EBRD (European Bank for Reconstruction and Development). 1998. *Various years. EBRD Transition Report*. London: EBRD.

Estrin, S., ed. 1994. *Privatization in Central and Eastern Europe*. London: Longman.

Estrin, S. 2002. Competition and corporate governance in transition. *Journal of Economic Perspectives* 16(1), 101–24.

Estrin, S. and Perotin, V. 1991. Does ownership always matter? *International Journal of Industrial Organization* 9, 55–72.

Estrin, S. and Wright, M. 1999. Corporate governance in the former Soviet Union: an overview. *Journal of Comparative Economics* 27, 398–421.

Gregory, P.R. and Stuart, R.C. 2004. *Comparing Economic Systems in the Twenty-First Century*. New

York: Houghton Mifflin.

Johnson, S., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. 2000. Tunneling. *American Economic Review* 90, 22–7.

Kornai, J. 1990. *The Road to a Free Economy. Shifting from a Socialist System: The Example of Hungary*. New York: Norton, Budapest: HVG Kiadó.

Megginson, W.L. 2005. *The Financial Economics of Privatization*, 1st edn. New York: Oxford University Press.

Meyer, K. 1998. *Direct Investment in Economies in Transition*. Aldershot: Edward Elgar.

Plane, P. 1997. Privatization and economic growth: an empirical investigation from a sample of developing market economies. *Applied Economics* 29, 161–78.

Sabirianova, K., Svejnar, J. and Terrell, K. 2005. Distance to the efficiency frontier and foreign direct investment spillovers. *Journal of the European Economic Association* 3, 576–86.

Shleifer, A. and Vishny, R. 1994. Politicians and firms. *Quarterly Journal of Economics* 46, 995–1025.

United Nations. 2004. *World Investment Report*. New York: United Nations.

Vickers, J. and Yarrow, G. 1985. *Privatisation and the Natural Monopolies*. London: Public Policy Centre.

World Bank. 1996. *World Development Report*. Washington, DC: World Bank.

Zinnes, C., Eilat, Y. and Sachs, J. 2001. The gains from privatization in transition economies: is change of ownership enough? *IMF Staff Papers* 48, 146–70.

How to cite this article

Estrin, Saul. "privatization impacts in transition economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000342> doi:10.1057/9780230226203.1344

The New Palgrave Dictionary of Economics Online

privatization

John Vickers
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Privatization is the transfer from government to private parties of the ownership of firms. Privatization programmes have been carried out worldwide since the mid-1980s, with important consequences for economic efficiency, public finance, and distribution. In competitive industries privatization generally has positive effects on incentives and performance. The economic consequences of privatizing firms with market power depend on the effectiveness of regulation and competition policy. These points are illustrated by experience in Britain, a leading exponent of privatization policies.

Keywords

allocative efficiency; bankruptcy; barriers to entry; bonds; commitment; competition law and policy; competitive tendering; control rights; cross-subsidy; duopoly policy; expropriation; firm, theory of; hard budget constraint; interconnection; liberalization; market power; merger policy; ownership; present value; price cap regulation; price control; principal and agent; privatization; public finance; public–private partnerships; rate-of-return regulation; redistribution of wealth; regulation of monopoly; share options; state-owned enterprises; takeovers; universal service obligations

Article

‘Privatization’ is defined here as the transfer from government to private parties of the ownership of firms. This definition is not so broad as to embrace, for example, the sale of publicly owned housing and natural resources, contracting out the supply of publicly financed services, or the introduction of user charges for services previously provided at public expense. However, some of the economic principles for privatizing firms apply more generally.

This article is in two parts. The first part addresses some economic and financial principles of privatization, beginning with the basic question: how does ownership matter for economic efficiency? It is concluded that, at least for firms with significant market power, this question must be addressed in

conjunction with the framework of regulation and competition that accompanies public or private ownership. The second part examines some aspects of privatization in practice, particularly in Britain, the leading exponent of the policy in the 1980s.

Privatization: principles

Ownership and economic efficiency

If privatization is defined as the transfer of ownership, the first question is: what is ownership? According to the incomplete contracts view of the firm (see Hart, 1995), ownership of an asset is to be identified with residual rights of control – rights to make decisions in the domain not already subject to contractual obligations. No such rights would exist in a world of complete contracts, where ownership, and hence privatization, would therefore be irrelevant.

The ultimate owners of sizable firms typically delegate the exercise of residual control rights to professional managers (whose identity may or may not be affected by privatization). Privatization affects principal–agent relationships between owners and managers by changing (*a*) the principals and hence their objectives, (*b*) the means of monitoring and giving incentives to the agents, and (*c*) the scope and incentives for action by the former public principals.

As to (*a*), a limitation to the economic theory of privatization is that there is no definitive theory of the firm under public ownership. In some sense the ultimate owners are the general public, but, even if their preferences could satisfactorily be aggregated into a welfare measure, it would be pious to suppose that government ministers or bureaucrats would necessarily exercise their authority over public firms to maximize welfare, avoiding distraction by political considerations, influence by well-organized vested interests, and so on. With private firms the usual assumption that owners seek to maximize profit or share value seems a tolerable approximation for present purposes, except perhaps if workers or consumers have large ownership stakes.

Since private, unlike public, ownership claims are generally tradable, privatization can alter the monitoring and incentives of managers by changing information conditions. For example, managers' rewards can be related to share price performance. In so far as share prices reflect the value of the firm, managers can thereby be given incentives to enhance firm value. Stock market investment analysts become a new source of managerial monitoring. However, free-rider considerations imply that monitoring by private owners might be limited, especially if share ownership is diffuse.

The tradability of ownership claims also means that privatized firms, unless they are given special protection, are potentially open to takeover threats, whereas publicly owned firms obviously are not. It is a matter for debate whether such threats from the market for corporate control are effective in disciplining managers of private firms to act in shareholder interests. Private firms also face the possibility of bankruptcy, in which case residual control rights shift to debt-holders.

Privatization changes the relationship between government and the firm. Thus public officials may lose power to intervene in the running of the firm. Moreover, and perhaps most important, the credibility of government commitment not to intervene may be enhanced by privatization so that, for example, managers face harder budget constraints and hence stronger incentives (see Schmidt, 1996).

Nevertheless, privatization might not make government commitment not to intervene completely credible, especially if the firm remains subject to regulation or dependent on public subsidy. In any

event, the government retains powers of taxation, and ultimately there is the possibility that privatization might be reversed, possibly on terms disadvantageous to private owners. To the extent that these factors give rise to risk of more or less subtle expropriation by government, private investment incentives may be adversely affected.

From the considerations above it follows that the consequences of privatization are likely to be influenced by the extent of market power enjoyed by the firm in question. For a firm that operates in competitive conditions, the shift from ‘public’ to profit objectives raises no concerns about the exercise of market power, and, since no special apparatus to regulate market power is required, opportunities for expropriation are limited. In these circumstances one may expect private ownership to be superior to public ownership in terms of economic efficiency, and indeed that is what the empirical evidence shows. For a firm with market power, however, it may be desirable for reasons of allocative efficiency, and inevitable for political reasons, for privatization to be accompanied by monopoly regulation. But regulation risks blunting the very incentives – for example, for cost reduction and efficient investment – that privatization is usually intended to sharpen.

A complementary approach to the problem of privatized (and in principle also nationalized) market power is liberalization – the removal of legal and other barriers to competition, and accompanying measures to contain anti-competitive behaviour by the incumbent firm. Among other things, liberalization may expose and undermine patterns of cross-subsidy practised under public monopoly. Therefore, in contrast to the competitive market case, it would appear that no general claim can be made as to the economic desirability of privatizing firms with market power. The accompanying regimes of regulation and competition policy are crucial determinants of the consequences of their privatization.

Privatization and public finance

In addition to microeconomic efficiency, considerations of public finance have motivated privatization policies in a number of countries, including Britain. By raising government revenue, privatization reduces the immediate need for public sector borrowing. It may also release firms from financial constraints resulting from government macroeconomic policy commitments. But the economic, as distinct from public accounting, significance of these points is unclear.

Selling public firms indeed raises government revenue, but the same is true of selling government bonds: in both cases the public sector receives a lump sum in return for a stream of future profit or interest payments. The deeper question is how privatization differs from government bond issue in terms of its effect on the net worth of the public sector.

If privatization leads to economic efficiency gains (which would not otherwise have been achieved) – or to greater exercise of monopoly power, which is akin to a tax increase in public finance terms – then the firm's profits are greater with privatization than in the public sector. If the firm is sold at a fair price, then the public sector captures the net present value of the profit gain (less the transactions costs of privatization, which are likely to exceed those of bonds). If, however, the firm is underpriced, then any gain to the net worth of the public sector is reduced by the extent of underpricing. Competition among potential buyers and a pre-existing market for the firm's shares are factors likely to assist more accurate pricing of privatization share issues.

Privatization can also affect the net worth of the public sector, compared with selling government bonds, if risk-adjusted discount rates differ. For example, a government with poor inflationary credibility may

have to cede a large interest rate premium when selling bonds. Shares in privatized firms are not so vulnerable to expropriation via inflation (neither are index-linked bonds). However, as discussed above in relation to regulatory credibility, some privatized firms, especially those with monopoly power, may face serious risks of expropriation via regulation or even renationalization. The relative sizes of these risks of default on debt and of 'default on equity' are likely to vary by industry as well as by country. The nature of the private shareholders – for example, their nationality or whether they are small individual investors – might also be an influence upon the probability of expropriation.

Self-imposed public finance constraints by government can provide efficiency rationales for privatization if they prevent publicly owned firms from making desirable investments. In macroeconomic terms, it ought to matter little whether a firm is in public or private ownership when it does a given amount of borrowing; it appears, however, that governments seeking to adhere to public borrowing commitments may view matters differently.

Privatization and distribution

Privatization, and the financial and industrial policies that accompany it, can have large distributional consequences. First, if public firms are sold to private investors for less than their market value – for example as part of a plan to promote 'wider share ownership' – then, relative to the situation with more accurate pricing, wealth is redistributed away from the general taxpayer to the investors who succeed in getting shares. Employees and managers of privatized firms gain from such redistribution if, as has often happened, they are allocated shares on favourable terms. Managers may benefit also from share option schemes and from being released from public sector pay constraints.

Second, if privatization hardens the firm's budget constraint, then it may diminish rents enjoyed by those within the firm to the benefit of the general taxpayer. Third, widespread cross-subsidy – for example, of small customers by large customers, and/or of suppliers of certain inputs – is a common feature of publicly owned monopoly. Privatization entails redistribution in so far as it undoes such cross-subsidies, but, here as elsewhere, the accompanying regime of regulation and competition is likely to be more important. Thus liberalization tends to be a more potent enemy of cross-subsidy than privatization itself, and, in the case of privatized monopoly, regulation can be a major determinant of the extent of redistribution among consumer groups as well as between consumers and shareholders.

Finally, it has been suggested (see Biais and Perotti, 2002) that privatization policies may be designed in part so that their distributional consequences alter political preferences – in particular by giving voters a stake in the avoidance of political parties whose policies would undermine the value of shares in privatized firms.

Privatization in practice

Privatization worldwide

Principally since the mid-1980s, privatization policies have been pursued, to varying degrees, around the world – for example in Argentina, Brazil, Chile, France, Germany, Jamaica, Japan, Malaysia, Mexico, the Philippines, Singapore, Spain, the formerly Communist countries of central and eastern Europe and

the former Soviet Union. Privatization sales proceeds worldwide are estimated to have exceeded a trillion dollars. The extensive survey by Megginson and Netter (2001) concludes that the state-owned enterprise share of global output fell from more than ten per cent in 1979 to below six per cent by 2000. The following account concentrates on Britain, which was a leader of the worldwide privatization movement in terms of both the scale of its programme and its embrace of monopoly industries. Further details may be found in Vickers and Yarrow (1988) and Newbery (2000).

Privatization in Britain

Nationalization by the post-war Labour government and subsequently had led to a situation in 1979 where the public sector in Britain dominated the supply of energy (gas, electricity, coal and some oil), transport (air, rail and bus), communications (post and telecommunications) and water, and also had substantial interests in manufacturing (for example, in aircraft, shipbuilding, steel and cars).

In the 18 years of Conservative government from 1979 to 1997, the proportion of GDP accounted for by state-owned firms fell from 11 per cent to below two per cent. At the peak of the privatization programme, between the mid-1980s and the early 1990s, sales proceeds typically exceeded one per cent of GDP and were sometimes of the order of three per cent of public expenditure.

The watershed in the British privatization programme was the sale of British Telecom (BT) in 1984, an event motivated in good part by a desire to free BT from macro-economic policy restrictions on public sector borrowing. Before that, privatization policies were relatively modest in scale and confined to firms in more or less competitive industries such as oil and manufacturing. By extending the programme to utility monopolies, the sale of BT marked a key shift in the nature, as well as the scale, of the British privatization programme. In particular, it required the development of a system for regulating private monopoly.

Privatization – with accompanying regulation – was subsequently extended to gas (1986), airports (1987), water in England and Wales (1989), electricity (1990–91) and the railways (1996). By 1997, when the Labour Party returned to power (having abandoned its traditional commitment to public ownership), the main activities remaining in the public sector were the Royal Mail, the BBC, London Underground, British Nuclear Fuels, Air Traffic Control, and the water industry in Scotland. In 2001 National Air Traffic Services was partly privatized as a public–private partnership (PPP). London Underground remains in public ownership but since 2003 infrastructure renewal and maintenance has been procured under long-term PPP contracts.

Methods of sale

The main ways of privatizing a firm are (*a*) offer for sale of shares to the general public, (*b*) sale to another firm, and (*c*) management/employee buyout. The third method was used in parts of the transport sector, including road haulage, some bus companies, and rail rolling stock leasing companies. The Rover car group was an example of privatization by sale to another firm (British Aerospace, which later sold Rover to BMW). However, by far the most important method used in Britain was offer for sale to the general public.

With this method, questions include (*a*) whether to sell the firm in two or more stages, or all at once; (*b*) whether the share price is set administratively or by competitive tendering among prospective purchasers

of the shares; and (*c*) whether incentives and bonus schemes are created to encourage small investors to buy (and hold) privatization shares. Before the BT sale, privatizations were mostly in stages (as was BT's), use was often made of competitive tendering, and no great inducements to wider share ownership were given. These methods are conducive to reasonably accurate share pricing. Thus selling a firm's shares in stages enables accurate pricing after the first stage because the market value of the shares is known.

In the latter part of the 1980s, however, some large firms (such as British Gas) were sold in one go, tendering methods were eschewed, and there were strong incentives for small investors to buy shares. This pattern suggests that wider share ownership was a primary objective of privatization policy. In the 1990s tendering methods came back into use, albeit with discounts for small investors, thus combining the objectives of revenue maximization and wider share ownership to some extent. However, Railtrack, the railway infrastructure company, was floated on the stock market in one go in 1996.

Even judged relative to the discounts that are typical with private initial public offerings, the government revenue forgone in pursuit of the objective of wider share ownership appears to have been very large. The number of British individuals directly owning shares rose sharply but the proportion of the stock market owned directly by individuals has continued its long-run decline. If it is thought to be an appropriate policy goal, wider share ownership might be better pursued by reforms to the taxation of saving and investment generally rather than by privatization policies.

Regulation

The regulatory framework for the privatized BT was established by the Telecommunications Act 1984. A similar framework was subsequently adopted for gas, electricity, water and railways. Regulatory powers and duties were divided between the government minister, who granted licences containing regulatory provisions; an industry-specific regulator (for example, the Director General for Telecommunications), who enforced and reviewed licence conditions; and the Monopolies and Mergers Commission, which considered disputes about licence modification.

This regulatory model developed over time. Powers were transferred from individual directors general to boards, and some regulatory bodies were combined. Thus the Utilities Act 2000 created the Gas and Electricity Markets Authority, and under the Communications Act 2003 a new body, Ofcom, took over the roles of several regulators including the Director General for Telecommunications. The regulators gained powers under new UK competition law (see below). And the wider European context grew in importance, with EC directives for the liberalization of network industries such as telecommunications and energy.

For firms with market power, perhaps the most important aspect of regulation concerns price control. When it embarked on the privatization of BT the British government was anxious to avoid perceived deficiencies of rate-of-return regulation. Instead, following the report of Professor Stephen Littlechild (1983), it adopted the form of price cap regulation known as 'RPI minus *X*', which requires an index of the firm's regulated prices to fall by *X* per cent per annum in real terms (that is, relative to the retail price index) for a period of years. This was intended to be 'regulation with a light hand' and to wither away over time. However, price regulation in several industries at first became tighter and more detailed, and rate-of-return considerations were soon seen to be of prime importance at points of regulatory review. Nevertheless, even if RPI minus *X* price cap regulation is akin to rate-of-return regulation with long lags,

this may well have substantial advantages over rate-of-return regulation as traditionally practised. After a time, as competition took hold after liberalization, some price controls were lifted, notably from domestic energy retail prices in 2002 (though transmission and distribution remain regulated) and from BT's retail prices in 2006.

Industry restructuring

Restructuring is an important instrument of competition policy when firms with market power are privatized. (Forced restructuring after privatization may seriously jeopardize regulatory credibility.) Both BT and British Gas were privatized without restructuring as vertically integrated firms with nationwide dominance. However, after a decade of competition problems arising from the vertical integration of British Gas, and, in view of accelerated liberalization of retail supply, the company divided itself into separate pipeline and supply companies in 1997.

By contrast, the government radically restructured the electricity and railway industries before privatization. In 1990 the Central Electricity Generating Board in England and Wales was divided into a transmission company (National Grid) and three generators (National Power and PowerGen; and Nuclear Electric, which was eventually privatized as British Energy in 1996). Vertical separation meant that a new mechanism had to be devised to coordinate transmission and generation, and a wholesale auction market, the Pool, was established (and later reformed by the introduction of New Electricity Trading Arrangements in 2001). In all, 12 Regional Electricity Companies (RECs) were privatized with responsibility for distribution and retail supply, which was progressively liberalized in the late 1990s and finally deregulated in 2002.

Major restructuring and ownership changes have occurred in the energy sector since privatization. The generators, National Power and PowerGen, had to divest substantial generation capacity following concerns about their market power, which was largely due to the concentrated structure for generation chosen by government in an unsuccessful effort to privatize nuclear power at the outset. Initially National Grid was jointly owned by the RECs but it became an independent company in 1995, and in 2002 merged with the gas pipeline company. After the lifting of takeover protections in the mid-1990s, most RECs were acquired, and ten years later six companies, supplying both gas and electricity (often in combined deals), accounted for nearly all energy supply – British Gas and five electricity suppliers, of which one is French- and two are German-owned. Thus, depending on merger policy, industry structure and ownership can alter substantially after privatization.

British Rail was restructured before privatization to separate network infrastructure from train operation. Railtrack, which took over network infrastructure, including track and stations, was privatized in 1996. The company went into administration in 2001 and its assets were acquired by Network Rail, a company limited by guarantee that has no shareholders. Three rolling stock leasing companies were also privatized in 1996 (and soon resold at a profit). Private train operating companies run train services under franchises. Large public subsidy to rail services continues in the privatized regime.

Liberalization of competition

Statutory monopoly typically accompanied public ownership in the utility industries. Among other things this served to facilitate extensive cross-subsidy between groups of customers, and sometimes of

input suppliers – for example, the nationalized electricity industry effectively subsidized British Coal. The removal of statutory barriers to entry in telecommunications, gas and electricity began in the early 1980s, before privatization policies were adopted, but then had little competitive effect. Liberalization has generally gone further since privatization – as illustrated above by the energy sector – and over time more attention has been given to economic, as well as legal, barriers to entry.

In telecommunications, liberalization of apparatus supply and value-added services began in 1981, when BT was split from the Post Office, and in 1982 Mercury was licensed as a competing network operator. However, for the rest of the decade the government adopted a ‘duopoly policy’ of allowing no further entry into fixed-link network operation. A parallel duopoly policy applied to mobile telecommunications. When the duopoly policy was ended in 1991, the interconnection question – on what terms can rivals gain access to BT's local network? – became and has remained a focus of controversy. On the one hand it was argued that rivals could inefficiently ‘cream-skim’ BT's more profitable business while BT remained restricted by controls on its tariff structure and universal service obligations. On the other hand, it was argued that rivals faced entry barriers. These tensions eased somewhat over time as tariff rebalancing diminished cross-subsidies in BT's pricing structure, and as entry barriers (such as the lack of number portability) were tackled directly by the regulator. But the advent of broadband, with BT still an integrated incumbent operator, brought the inter-connection question back into sharp focus. Faced with the prospect of an investigation under competition law, BT agreed in 2005, 20 years after privatization, to operational separation of its local access infrastructure.

A major weakness of UK policy towards privatized firms with market power had been the absence of effective competition law against anti-competitive agreements and abuse of dominance. However, that gap was filled in March 2000 when the Competition Act 1998 – which mirrors Articles 81 and 82 of the EC Treaty – came into force and was followed by the Enterprise Act 2002. The regulators can now apply (non-merger) competition law in their sectors. Over time, then, following the shift from state monopoly to regulated private monopoly, there has been increasing availability of competition policy instruments to address market power in historically monopolized industries such as energy and telecommunications in Britain. Nevertheless, the regulatory regimes have remained the principal means of controlling market power.

The performance of privatization

Privatization policies have undoubtedly had major economic and financial effects. Have they generally been positive? Answering this question properly requires the specification of evaluation criteria, performance measures, statistical methods and the counterfactual: what would have happened without privatization?

Megginson and Netter (2001, section 5) review 38 empirical studies of privatization covering both developed market economies and transition economies. Privatized firms are generally found to become more efficient and profitable, and to invest more. There are mixed results on employment effects, though job cuts appear to be associated with corresponding productivity gains. Direct evidence on effects on consumers is limited. In their survey of studies of transition economies, Djankov and Murrell (2002) conclude that privatization, especially to outside investors as distinct from managers and workers, is robustly associated with enterprise restructuring and growth, and that competition has a significant

positive effect on enterprise performance.

In competitive industries, improvements in the corporate performance of privatized firms imply overall economic gain, and there is ample evidence that privatization has been a success. For firms with market power, however, corporate performance can improve at the expense of the public as well as by enhanced efficiency. Moreover, it is hard to isolate the effects of privatization in hitherto monopolized industries from those of accompanying regulatory and competitive reforms. In Britain, methods of privatization and regulatory reform have at times been seriously flawed. But privatization was probably necessary for liberalization and for the creation of a system of independent economic regulation, augmented in time by effective competition policy. Though far from perfect, these are major improvements upon the nationalized monopoly of old.

See Also

- competition
- policy reform, political economy of
- public utility pricing and finance
- regulation

Bibliography

Biais, B. and Perotti, E. 2002. Machiavellian privatization. *American Economic Review* 92, 40–258.

Djankov, S. and Murrell, P. 2002. Enterprise restructuring in transition: a quantitative survey. *Journal of Economic Literature* 40, 739–92.

Hart, O. 1995. *Firms, Contracts, and Financial Structure*. Oxford: Oxford University Press.

Littlechild, S.C. 1983. *Regulation of British Telecommunications Profitability*. London: HMSO.

Megginson, W.L. and Netter, J.M. 2001. From state to market: a survey of empirical studies on privatization. *Journal of Economic Literature* 39, 321–89.

Newbery, D. 2000. *Privatization, Restructuring and Regulation of Network Utilities*. Cambridge, MA: MIT Press.

Schmidt, K.M. 1996. The costs and benefits of privatization: an incomplete contracts approach. *Journal of Law, Economics and Organization* 12, 1–24.

Vickers, J. and Yarrow, G. 1988. *Privatization: An Economic Analysis*. Cambridge, MA: MIT Press.

How to cite this article

Vickers, John. "privatization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000193> doi:10.1057/9780230226203.1343

The New Palgrave Dictionary of Economics Online

procurement

Yeon-Koo Che

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Firms and government agencies rely increasingly on goods and services procured from outside suppliers. How to assure desired quality at a minimal cost in the procurement is often challenging and warrants carefully devised contracting policies. This article reviews several problems arising in procurement and policies designed to remedy them.

Keywords

adverse selection; auction contests; auctions; collusion; corruption; cost-plus contracts; fixed-price contracts; mechanism design; option contracts; outsourcing; procurement; quality assurance; reputation; risk aversion; risk sharing; scoring auctions; virtual cost

Article

Virtually all businesses, both public and private, rely on procurement of numerous goods and services, ranging from such routine jobs as food and custodial services to the complex job of building high-tech fighter jets and high-speed train systems. Rapid progress of communication technologies – most notably the emergence of the Internet – has made outsourcing both cheaper and more efficient, thus altering the traditional boundaries of ‘make-or-buy’ decisions by many firms and government agencies in favour of more outsourcing. Thus, designing efficient mechanisms for procuring goods and services has become ever more important.

Procurement of standardized parts and services is relatively straightforward as a competitive market or standard bidding would produce an efficient outcome. In many procurement settings, however, the quality of the procured job is an important concern, and it is not easy to assure the desired level of quality, since a high-quality job may entail a cost that is unknown or privately known by the supplier or require his special effort that is not observable to the buyer, or the quality of the job provided is not easy to verify or is simply unobservable to the buyer. The present article reviews some of the answers

economics research has provided for optimal responses to these procurement problems.

Contractible quality

When quality is verifiable, the terms of the contract can be made contingent on the quality of the system. If the cost associated with delivering the job is unobservable, it presents two problems. First, ensuring the supplier's participation may require paying more than the true cost, that is, information rents to elicit the cost, so the quality level must be decided based on the overall cost paid to the supplier. Second, the buyer must identify the supplier who can deliver the good at the minimal cost to her. We sketch the method for finding the optimal mechanism that deals with these issues.

To begin, suppose a buyer derives utility of $v(q) - t$ when she procures a job of quality $q \in \mathbb{R}_+$ and pays $t \in \mathbb{R}_+$, where the gross surplus function, $v: \mathbb{R}_+ \mapsto \mathbb{R}_+$, is strictly increasing, differentiable and strictly concave. Suppose there are $n \geq 1$ potential suppliers. Supplier $i \in N := \{1, \ldots, n\}$ can deliver quality at unit cost of θ_i, which is drawn from $[\underline{\theta}, \overline{\theta}] =: \Theta$ according to the cumulative distribution function $F_i(\cdot)$ which has a positive density over $(\underline{\theta}, \overline{\theta})$. Assume also that $\theta + \frac{F_i(\theta)}{f_i(\theta)}$ is non-decreasing in θ. A supplier *i* receives $t - \theta_i q$ from a contract that pays him *t* for delivering *q*.

If the suppliers' costs are observable, then the procurer's decision will be straightforward. She can pick the most efficient one, $i = \arg\min_{i \in N}\{\theta_i\}$, and have him deliver the job at cost, so she will pick the first-best quality, $q_i^{**}(\theta_i) \in \arg\max_{q \in \mathbb{R}_+} v(q) - \theta_i q$.

When the suppliers' costs are unobservable, it is not possible to procure at the actual cost, since suppliers can pretend to have higher than actual cost. Nor is it easy or necessarily desirable to pick the least-cost supplier, as will be seen. To illustrate the optimal procurement decision, suppose first there is only one potential supplier, $n = 1$. By the revelation principle, there is no loss in restricting attention to a direct revelation contract that determines the quality and the payment, $\{(q_i(\theta), t_i(\theta))\}_{\theta \in \Theta}$, as a function of the cost reported by the supplier.

The optimal contract $(q_i^*(\cdot), t_i^*(\cdot))$ must solve

$$\max_{q_i(\cdot), t_i(\cdot)} \int_{\underline{\theta}}^{\overline{\theta}} [v(q_i(\theta)) - t_i(\theta)]\, dF_i(\theta)$$

[P]

subject to

$$U_i(\theta) := t_i(\theta) - \theta q_i(\theta) \geq 0, \ \forall \theta \in \Theta$$

(IR)

$$U_i(\theta) \geq t_i(\tilde{\theta}) - \theta q_i(\tilde{\theta}), \ \forall \theta, \tilde{\theta} \in \Theta,$$
(IC)

where (IR) and (IC) ensure, respectively, the supplier's participation and his incentive to report truthfully his type.
By the well-known method, (IR) and (IC) constraints can be simplified to a pair of conditions:

$$q(\cdot) \text{ is non-increasing.}$$
(M)

and

$$U_i(\theta) = \int_{\theta}^{\bar{\theta}} q(\tilde{\theta})\, d\tilde{\theta},$$
(Env)

or equivalently

$$t_i(\theta) = \theta q_i(\theta) + \int_{\theta}^{\bar{\theta}} q(\tilde{\theta})\, d\tilde{\theta}.$$
(Env′)

The constraint (M) will be seen not to bind, so it can be ignored. Substituting (Env′) into the objective function of [*P*] and switching the order of expectations yield

$$\int_{\underline{\theta}}^{\overline{\theta}} [v(q_i(\theta)) - J_i(\theta) q_i(\theta)] \, dF_i(\theta),$$

where $J_i(\theta) := \theta + \frac{F_i(\theta)}{f_i(\theta)}$ is the so-called 'virtual' cost. The additional cost $\frac{F_i(\theta)}{f_i(\theta)}$ reflects the additional rents that must be given away to the types more efficient than θ when its quality is raised marginally. To see this more intuitively, suppose the quality for type θ is raised by Δq towards the efficient level. Then, there is an efficiency gain (to be captured by the procurer) of $[v'(q) - \theta] \Delta q f(\theta)$. At the same time, the raising of quality enables each type $\theta' < \theta$ to command extra rents of Δq by mimicking (or choosing the contract intended for) type θ, so the same amount must be given to them to dissuade them from doing so. Since the measure of those types is $F(\theta)$, the marginal cost of quality increase is $F(\theta) \Delta q$. The optimal quality $q_i^*(\theta)$ balances these two marginal effects, so $v'(q) - \theta = \frac{F_i(\theta)}{f_i(\theta)}$ if $q_i^*(\theta) > 0$, or more generally

$$q_i^*(\theta) \in \arg\max_{q \in \mathbb{R}_+} [v(q) - J_i(\theta) q].$$

Clearly, $q_i^*(\theta) < q_i^{**}(\theta)$, for $\theta > \underline{\theta}$, whenever $q_i^{**}(\theta) > 0$. In other words, it is optimal for the buyer to choose less than the first-best quality. In particular, the buyer may not procure at all even though procuring is socially efficient, for instance when $\theta < v'(0) < J_i(\theta)$. In practice, the optimal procurement policy can be implemented by a *menu* of quality-transfer pairs, $(q_i^*(\theta), t_i(\theta))$, or by a *nonlinear pricing scheme* $T(q) := t_i(q_i^{*-1}(q))$.

Now suppose $n \geq 2$ so there are multiple candidate suppliers. The selection of the supplier, which can be studied using the same mechanism design approach (see Myerson, 1981; Laffont and Tirole, 1987; McAfee and McMillan, 1987; Riordan and Sappington, 1987), extends the above insight naturally. What ultimately matter to the buyer are suppliers' virtual costs, not their actual costs. Hence, the supplier with the lowest virtual cost, $i \in \arg\min_{j \in N} \{J_j(\theta_j)\}$, must be selected, and the selected supplier must choose the 'downward distorted' quality level, $q_i^*(\theta)$. If supplier *i* has *ex ante* higher cost than *j*, say in terms of conditional stochastic dominance: $\frac{F_i}{f_i} < \frac{F_j}{f_j}$, then the optimal selection rule favours *i*. Favouring the 'underdog' can be seen as a way of *handicapping* the top dog to make him compete more aggressively.

When the suppliers are *ex ante* symmetric, that is, $F_i = F_j$ for $i \neq j$, then the optimal selection is also

efficient, and the optimal procurement policy can be implemented by the so-called *scoring auction* (see Che, 1993). Specifically, there is a quasi-linear scoring function,

$$S(q, t) := v(q) - \Delta(q) - t,$$

for some $\Delta(\cdot)$ increasing, that implements the optimal outcome if the suppliers are asked to make two-dimensional bids, (q, t), and the supplier who achieves the highest score according to $S(q, t)$ is selected to produce his proposed quality and receive his payment. The term $\Delta(q)$ serves as a penalty against 'quality bid' so as to implement the downward distortion feature of the optimal contract. The scoring auction resembles the procedures used in the procurement of weapons, transportation, construction, and a multitude of other goods and services. *Quasi-linear scoring auctions* are analytically tractable and can implement a broad range of outcomes, even when the quality is multidimensional (so q is a vector of attributes) and the suppliers may have heterogeneous costs with these attributes (Asker and Cantillon, 2004). (A quasi-linear scoring auction may not implement the optimal direct revelation mechanism for the buyer if the suppliers have multidimensional costs, but it does implement the socially efficient quality mix. See Asker and Cantillon, 2005.)

In many procurement settings, the monetary expenditures are observable, but the suppliers' inherent capabilities as well as their effort to reduce the cost may not be observable. In this case, how a supplier's cost should be reimbursed becomes an important issue. A *fixed-price contract* that pays the same price to the supplier regardless of his realized cost provides a strong incentive for cost-reduction but requires the supplier to bear the risk of cost shocks. By contrast, a *cost-plus contract*, which reimburses the supplier's cost fully, provides weak incentives for cost-reduction effort but imposes no risk on the part of the supplier. McAfee and McMillan (1986) show a mixture of the two contract forms — that is, a *partial reimbursement rule* — to optimally balance the trade-offs between the cost reduction incentive, adverse selection and risk sharing. (Laffont and Tirole, 1986, obtained a similar result without risk aversion of the agent.) Bajari and Tadelis (2001) focus on the trade-off between the cost reduction incentive and *ex post* renegotiation inefficiencies, and study how the complexity of the procurement job affects the choice of contract form. They argue that fixed-price contracts are optimal for standard jobs, whereas cost-plus contracts are optimal for complex projects. Empirical findings on the contract choice appear consistent with this latter finding (Crocker and Reynolds, 1993; Corts and Singh, 2004; Bajari, McMillan and Tadelis, 2002).

Uncontractable quality

Often the quality enjoyed by the buyer is unobservable to the supplier and/or unverifiable to the court, so it is difficult to contract on it *ex ante*. Book publishing, advertising, film production, development of new (such as pharmaceutical) technologies, procurement of new weapons systems and hiring new talents all involve some difficulty in specifying the quality of jobs. While *ex post* signals about quality are often available after the procurement (for example, sale of a book or of an advertised product), the fixed cost

associated with procurement may be so high that quality assurance is needed before full-scale production begins. We discuss several methods for assuring quality.

To illustrate, suppose a buyer values the good at *q* if a supplier makes an effort $c(q) = \frac{1}{2}q^2$. It would be ideal for the buyer to obtain the quality $q^* = 1$ at the price of $c(q^*) = \frac{1}{2}$. If quality is unverifiable, it would be difficult to specify contractually the level of quality. The buyer would argue that the quality provided is lower than specified, and the supplier would argue the opposite. In any case, the supplier would have little incentive to provide high quality, since there would be little reward for it.

A simple *option contract* can solve the problem of unverifiable quality. Suppose the buyer signs a contract that requires the supplier to pay a (non-refundable) upfront fee of $q^* - c(q^*) = \frac{1}{2}$ to the buyer and gives the buyer an option either to accept the good at the price of $p = q^* = 1$ or to reject it at no penalty. If the supplier produces quality of *q*, then the buyer would receive $q - \frac{1}{2}$ from accepting the good and $\frac{1}{2}$ from rejecting the good. Hence, the buyer will purchase the good if and only if $q - \frac{1}{2} \geq \frac{1}{2}$, or the quality is at least $q^* = 1$. Knowing this, the supplier will produce $q^* = 1$. The supplier has the incentive to provide adequate quality since the buyer has an option to reject the good if the quality is not to his liking. These option contracts, known by such names as *purchase upon approval* and *delivery-contingent* contracts, are common in situations where quality assurance is important (Taylor, 1993; Che and Hausch, 1999). For instance, advertising agencies must often develop acceptable pilot campaigns before they are paid in full; real estate agencies and other brokers are typically not paid until they find an acceptable match between buyers and sellers; book publishers often reserve the right to recover an advance in the event that they find the book unacceptable. The *up-or-out* contracts well-known for academic tenure and law partnerships are a form of an option contract, presumably motivated to deal with the ‘unverifiable quality’ problem (Kahn and Huberman, 1988).

A problem with the option contract is that it requires the supplier to pay an upfront fee, that is, to buy in. Often, suppliers have limited liability or are liquidity constrained, which can make the option contract infeasible. In the above example, for instance, if the supplier cannot be induced to ‘buy in’, the buyer will receive zero net surplus. (This is attributable to the deterministic nature of quality. If quality is stochastic, the quality accepted will generally exceed the option price. Even with stochastic quality, however, the option contract will be of limited value to the buyer if the suppliers cannot buy in.) This problem can be solved by a *pilot/research contest*, that is, by having multiple suppliers compete for a reward. To be concrete, suppose the buyer invites two suppliers, each with the same technology described above, and suppose the buyer promises a fixed prize $P = \frac{8}{25}$ (which turns out to be the optimal level) for the supplier who offers the higher quality. It is then equilibrium behaviour for each supplier to randomize quality over [0,•4/5] according to the CDF, $F(q) = \frac{25}{16}q^2$, yielding a surplus of $\frac{8}{25}$ to the buyer. (If the other supplier follows the randomization strategy, a given supplier receives a payoff of $F(q)P - c(q) = \frac{1}{2}q^2 - \frac{1}{2}q^2 = 0$, when choosing $q \in [0, 4/5]$, so randomizing according to *F* is a best response.)

Fixed-prize contests have been used for developing new innovations, some historically important, such

as the *longitude technology* and the *steam engine design*. But recent procurement contests have allowed suppliers some freedom in specifying their own rewards. For instance, most of the defence procurement competitions as well as grant competitions allow suppliers to adjust the size of their prizes and compete along that dimension as well. Such *auction contests* can in fact be justified, as a contest that allows suppliers to bid on their reward is optimal (see Che and Gale, 2003). Suppose the buyer in the above example lets suppliers bid prices for their innovations and then procures from the one offering the highest net surplus (the difference between the quality offered and payment demanded). This auction contest induces each supplier to randomize over *q* uniformly from [0,1] and to bid $\frac{1}{2}q$ for his prize, yielding a net expected surplus of $\frac{1}{3}(>\frac{8}{25})$ to the buyer.

The purpose of employing competition here is not to select a supplier efficiently (recall both suppliers are equally efficient *ex ante*) but rather to provide incentives for unverifiable quality. The buyer has an incentive to select the supplier that offers the best value (quality minus payment), and this option to select from suppliers – just like the option to reject in the option contract – creates incentives for quality from the suppliers, and assures the surplus for the buyer even without supplier buy-in. Such incentives come at the expense of duplicative investment, however, since the buyer procures from only one supplier. This suggests that limiting the number of competitors – often to two – is optimal (see Che and Gale, 2003; Taylor, 1995; Fullerton and McAfee, 1999). If the quality of the innovation/good is stochastic, competition will serve the additional purpose of identifying an efficient supplier.

The non-verifiable quality problem can be overcome if the buyer procures the good repeatedly. In such a situation, a *reputation* – more specifically, the promise of granting rents in exchange for an agreed-upon quality – combined with the threat of terminating a relationship for a sub-par quality, can create the supplier's incentives for quality (Klein and Leffler, 1981). For instance, in the above example the buyer and a supplier can make an implicit agreement such that the latter provides the quality of $q^* = 1$ in exchange for a payment $p \in (\frac{1}{2}, 1)$ from the former, as long as both honour the agreement; if one deviates unilaterally, both terminate the relationship. The threat of termination is credible, since it is a Nash equilibrium in each round for the supplier to provide zero quality and for the buyer to pay nothing. If nobody deviates, the buyer and the supplier would obtain $\frac{1-p}{1-\delta}$ and $\frac{p-0.5}{1-\delta}$, respectively, where $\delta \in [0, 1)$ is a common discount factor. The two parties can get at most 1 and *p*, respectively, from a unilateral deviation. If $\delta \geq \left\{p, \frac{1}{2p}\right\}$, then the first-best quality can be implemented in a subgame perfect equilibrium.

So far, we have assumed that quality of procurement is observable to the buyer (albeit non-verifiable to the court). Often, the quality of good supplied may not be observable to the buyer at the time she procures from a supplier. Development of new weapons or transportation systems are subject to this problem, as the quality of new features is learned long after the procurement. If the buyer is unsure about the quality supplied, a standard auction based solely on price performs poorly (and unobservability of quality precludes the use of multi-dimensional competition such as scoring auctions). In such a case, it may be socially optimal for the buyer to bargain with one of many potential suppliers, instead of inviting them to compete for a job. To illustrate, suppose there are two potential suppliers, each with cost *c* drawn independently and uniformly from [0,1]. A supplier with *c* can deliver a good with quality

$v(c) = 3c$ to the buyer, so the quality is not only unknown to the buyer but also positively correlated with the supplier's cost. In this case, competition based only on price will result in the selection of the low quality, with the buyer obtaining only $\frac{1}{3}$ in expectation. By contrast, if the buyer selects a supplier at random and makes a *take-it-or-leave-it offer* of price 1, then the offer will be accepted, and the buyer will enjoy an expected surplus of $\frac{1}{2}$. Notice that bargaining dominates competition also in social surplus in this case (see Manelli and Vincent, 1995).

Procurement irregularities

The difficulty with verifying quality may require a buyer to hire agents with special expertise to evaluate the proposals. This added bureaucracy can introduce agency costs to the procurement. In particular, there is a potential for the agents evaluating proposals to favour a certain supplier in exchange for a bribe or kickback. *Corruption* is a serious problem in both public and private procurement, particularly across national borders. (Between 1994 and 1999, bribery was allegedly a factor in the awarding of nearly 300 contracts worldwide worth $145 billion and caused US firms to lose as many as 77 contracts worth $24 billion.) Burguet and Che (2004) analyse this problem via a scoring-auction model where quality score is measured imperfectly and is manipulable by the procurement agent in exchange for a bribe, and show that bribery competition – unlike standard auction competition – leads to allocational inefficiencies (see also Celentani and Ganuza, 2002; Burguet and Perry, 2002; Compte, Lambert and Verdier, 2005).

Another type of procurement irregularity is *collusion* among bidders in procurement competition. Bidding cartels in procurement auctions account for a significant portion of antitrust cases. McAfee and McMillan (1992) show that standard auctions are vulnerable to collusion but that the outcome will depend crucially on whether the cartel can exchange transfers. If the cartel members can exchange transfers, they can organize a ‘knock-out’ auction to achieve an efficient allocation, whereas if transfers cannot be used (for fear of detection, say) a member will be chosen randomly to win without any competition, meaning that allocation will be inefficient. Subsequent work has shown that repeated interaction allows asymmetrically informed cartel members to sustain collusion via a ‘bid rotation’-type scheme, and that the scheme can be refined to attain a degree of allocational efficiency (see, for example, Aoyagi, 2003; Athey and Bagwell, 2001; Athey, Bagwell and Sanchirico, 2004; Blume and Heidhues, 2002; Skrzypacz and Hopenhayn, 2004). To what extent inefficiencies result from procurement irregularities and how they can be remedied by procurement policies remain open questions. (For some promising lead for the latter question, see Che and Kim, 2006a; 2006b; Dequiedt, 2005; Marshall and Marx, 2003; Pavlov, 2006).

See Also

- auctions (applications)
- auctions (theory)
- cartels

- defence economics
- hold-up problem
- incomplete contracts
- mechanism design
- mechanism design (new developments)

Bibliography

Aoyagi, M. 2003. Bid rotation and collusion in repeated auctions. *Journal of Economic Theory* 112, 79–105.

Asker, J. and Cantillon, E. 2004. Properties of scoring auctions. Discussion Paper No. 4734. London: CEPR.

Asker, J. and Cantillon, E. 2005. Optimal procurement when both price and quality matter. Discussion Paper No. 5276. London: CEPR.

Athey, S. and Bagwell, K. 2001. Optimal collusion with private information. *RAND Journal of Economics* 32, 428–65.

Athey, S., Bagwell, K. and Sanchirico, C. 2004. Collusion and price rigidity. *Review of Economic Studies* 71, 317–49.

Bajari, P. and Tadelis, S. 2001. Incentives versus transaction costs: a theory of procurement contracts. *RAND Journal of Economics* 32, 387–407.

Bajari, P., McMillan, R. and Tadelis, S. 2002. Auctions versus negotiations in procurement: an empirical analysis. Working paper, Department of Economics, Stanford University.

Baron, D. and Myerson, R. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50, 911–30.

Blume, A. and Heidhues, P. 2002. Modeling tacit collusion in auctions. Mimeo, University of Pittsburgh.

Burguet, R. and Che, Y.-K. 2004. Competitive procurement with corruption. *RAND Journal of Economics* 35, 50–68.

Burguet, R. and Perry, M. 2002. Bribery and favoritism by auctioneers in sealed bid auctions. Working paper, Department of Economics, Rutgers University.

Celentani, M. and Ganuza, J.-J. 2002. Corruption and competition in procurement. *European Economic*

Review 46, 1273–303.

Che, Y.-K. 1993. Design competition through multidimensional auctions. *RAND Journal of Economics* 24, 668–80.

Che, Y.-K. and Gale, I. 2003. Optimal design of research contests. *American Economic Review* 93, 646–71.

Che, Y.-K. and Hausch, D. 1999. Cooperative investments and the value of contracting. *American Economic Review* 89, 125–47.

Che, Y.-K. and Kim, J. 2006a. Robustly collusion-proof implementation. *Econometrica* 74, 1063–108.

Che, Y.-K. and Kim, J. 2006b. Optimal collusion-proof auctions. Discussion Paper No 0506-22, Department of Economics, Columbia University.

Compte, O., Lambert, A. and Verdier, T. 2005. Corruption and competition in procurement auctions. *RAND Journal of Economics* 36, 1–15.

Corts, K. and Singh, J. 2004. The effects of relationships on contract choice: evidence from offshore drilling. *Journal of Law, Economics, and Organization* 20, 230–26.

Crocker, K. and Reynolds, K. 1993. The efficiency of incomplete contracts: an empirical analysis of Air Force engine procurement. *RAND Journal of Economics* 24, 126–46.

Dequiedt, V. 2005. Optimal collusion and optimal auctions. Mimeo, University of Toulouse.

Fullerton, R. and McAfee, R. 1999. Auctioning entry into tournaments. *Journal of Political Economy* 107, 573–605.

Kahn, C. and Huberman, G. 1988. Two-sided uncertainty and up-or-out contracts. *Journal of Labor Economics* 6, 423–43.

Klein, B. and Leffler, K. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89, 615–41.

Laffont, J.-J. and Tirole, J. 1986. Using cost observations to regulate firms. *Journal of Political Economy* 94, 614–41.

Laffont, J.-J. and Tirole, J. 1987. Auctioning incentive contracts. *Journal of Political Economy* 95, 921–37.

Manelli, A. and Vincent, D. 1995. Optimal procurement mechanisms. *Econometrica* 24, 668–80.

Marshall, R. and Marx, L. 2003. Bidder collusion. Mimeo, Penn State University.

McAfee, R. and McMillan, J. 1986. Bidding for contracts: a principal agent analysis. *RAND Journal of Economics* 17, 326–38.

McAfee, R. and McMillan, J. 1987. Competition for agency contracts. *RAND Journal of Economics* 18, 296–307.

McAfee, R. and McMillan, J. 1992. Bidding rings. *American Economic Review* 82, 579–99.

Myerson, R. 1981. Optimal auction design. *Mathematics of Operations Research* 6, 58–73.

Pavlov, G. 2006. Colluding on participation decisions. Working paper, Department of Economics, Boston University.

Riordan, M. and Sappington, D. 1987. Awarding monopoly franchises. *American Economic Review* 77, 375–87.

Skrzypacz, A. and Hopenhayn, H. 2004. Tacit collusion in repeated auctions. *Journal of Economic Theory* 114, 153–69.

Taylor, C. 1993. Delivery-contingent contracts for research. *Journal of Law, Economics and Organization* 9, 188–203.

Taylor, C. 1995. Digging for golden carrots. *American Economic Review* 85, 872–90.

How to cite this article

Che, Yeon-Koo. "procurement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000337> doi:10.1057/9780230226203.1345

The New Palgrave Dictionary of Economics Online

product differentiation

Simon P. Anderson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Product differentiation is pervasive in markets. It is at the heart of structural empiricism and it smoothes jagged behaviour that causes paradoxical outcomes in several theoretical models. Firms differentiate their products to avoid ruinous price competition. Representative consumer, discrete choice and location models are not necessarily inconsistent, but performance depends crucially on the degree of localization of competition. With (symmetric) global competition, rents are typically small and market variety near optimal. With local competition, profits may be protected because entrants must find profitable niches. These rents lead firms to competitively dissipative them, and performance may be poor.

Keywords

Bertrand competition; Bertrand paradox; business stealing; chain linking; characteristics; circle model; constant elasticity of substitution (CES) model; Diamond paradox; discrete choice models; endogenous growth; general probit model; horizontal and vertical differentiation; intra-industry trade; local competition; location models; market power; menu costs; monopolistic competition; nested logit model; network externalities; New Keynesian macroeconomics; product differentiation; quality ladders; Representative consumer models; spatial competition; vertical differentiation

Article

1 Overview

Consumer goods are available in a variety of styles and brands. Product differentiation refers to such variations within a product class that (some) consumers view as imperfect substitutes. The store Foods of all Nations in Charlottesville, Virginia, USA (area population 120,000) carries 118 varieties of hot pepper sauce, 41 balsamic vinegars and 121 different olive oils (these figures include variations such as flavourings and different package sizes from the same manufacturer). There are 82 other retail grocers

listed in the area. Charlottesville is served by 23 rated radio stations which differ by format choices (18 are commercially operated).

Product differentiation offers firms market power. This enables them to transcend the Bertrand paradox for pricing homogeneous products. In the Bertrand paradox, two or more firms sell goods that consumers perceive as identical, so goods are perfect substitutes. Assume that marginal costs are common and constant, and market demand has a finite price intercept. Then one good cannot carry a price premium over another while retaining positive sales. Any lowest price above marginal cost would then profitably be undercut. This logic impels us to marginal cost pricing as the only equilibrium under Bertrand competition.

Product differentiation resolves the paradox naturally. When products are imperfect substitutes, a price-cutting firm cannot take all its rivals' customers with an infinitesimally small price cut. This means that firms have some market power (due to the special features that distinguish them from their rivals' products); they can set prices without a completely elastic response by consumers. It also means that the product itself becomes a choice variable and firms differentiate to avoid the Bertrand outcome.

However, many models of product differentiation do not treat this choice explicitly, and instead assume a framework (representative consumer, discrete choice, and symmetric location models) that generates a demand system. It is not so much the framework used but rather the structure of product differentiation that is critical to the predictions and results. Indeed, common models of one type may be recast within another framework and be formally equivalent. Instead, the important feature for performance is whether each product is equally substitutable with all others or if each has only few close substitutes which are chain-linked to other products in the industry. Equal substitutability describes *global competition* where each firm competes with each other firm. Chain-linking corresponds to *local competition*. Local competition models naturally apply in geographical space since nearby stores are closer substitutes for consumers than distant ones. Likewise, in a characteristics setting, a consumer with a sweet tooth will find sugary products closer substitutes for any sweet product than for a saltier one.

The next section describes models of product location (in geographical space or its characteristics analogue) and distinguishes horizontal from vertical differentiation. Section 3 compares the common approaches to product differentiation used to analyse the market provision of variety. In these models, product decisions are suppressed and product selection is determined by entry. Section 4 describes how the market variety diverges from the equilibrium one. Section 5 elaborates on this theme for local competition. Section 6 indicates how product differentiation is used elsewhere in economics.

2 Product choice

Hotelling (1929) wrote the seminal paper treating the product specification as endogenous. Applications beyond industrial organization include marketing, economic geography (spatial competition), political science (the 'Hotelling–Downs' model), and media economics. The basic paradigm is that consumers are differentiated by their locations ('addresses') and dislike distance. Products, too, are locations in this space (geographic, characteristics and so on). When products are priced at marginal cost, consumers differ by which they like best, a situation known as *horizontal differentiation*. The simplest version of the model has two ice-cream sellers locating on a beach (with fixed prices). The Nash equilibrium is back-to-back pairing at the median of the consumer distribution, a result christened the *principle of*

minimum differentiation. It has been used to explain striking similarities in colas, petrol station location, political parties' platforms and the timing of television programmes.

However, the principle dissolves when firms locate in rational expectation of ensuing equilibrium prices (that is, seeking a subgame perfect equilibrium to a two-stage price-then-location game). Indeed, if two products were collocated, Bertrand competition would drive prices to marginal cost. Firms will avoid this ruinous result by differentiating to retain market power attributable to location advantage. The equilibrium trades off two opposing factors. Getting closer to a rival provokes more intense price competition, so firms differentiate in order to relax price competition, but getting close to a rival attracts more customers.

The equilibrium locations are outside the optimum ones (which are at the quartiles for a uniform consumer density) for the central case of quadratic distance disutility costs, but otherwise there is no fundamental reason for excessive market differentiation. More elaborate models can rapidly become quite intractable and are hamstrung by non-existence of (pure-strategy) price equilibria due to fundamental failure of quasi-concavity of the profit functions in prices.

The case above of horizontal differentiation has consumers with fundamental preference differences across different varieties. In *vertical differentiation*, all consumers have the same preference ordering (when goods are priced at marginal cost). More preferred goods are often described as having higher 'quality' (with different individuals having different willingness to pay for quality). In vertical differentiation models, firms are to choose their product qualities. Choosing the same quality is avoided because of ruinous price competition, and the same trade-off operates as under horizontal differentiation. Under vertical differentiation though, the firm producing a higher-quality product earns more profit than a firm with lower quality. This result is an extension of the Bertrand paradox. One firm differentiates itself by a low quality, but this puts it at a disadvantage. Indeed, it may not be able to escape the shadow of the high-quality firm and earn a positive profit in equilibrium. This result implies the finiteness property that only a finite number of firms can survive in equilibrium even as fixed costs become arbitrarily small. By contrast, in a horizontal model, a firm may always find a niche between existing firms that gives it an advantage over some consumers (so that finiteness cannot hold). Finally, if the costs of improving quality are mainly sunk, a firm may invest more heavily in quality in a larger market because the benefit accrues over a larger consumer base (so sunk costs are endogenous).

Quite similar in spirit to the above approaches, Lancaster's (1966) model of *characteristics* was a quite revolutionary approach to consumer theory. It posited that consumers care about the characteristics intrinsic to goods and purchase goods because they deliver the desired characteristic mix, adjusting appropriately for prices. Lancaster's theory answers the question of why goods are desirable by formulating fundamental preferences over characteristics. The approach is intuitively appealing and is at the heart of hedonic models in econometrics, state preference and mean variance models in portfolio choice problems in finance, and structural econometric work in industrial organization. However, the approach is rather cumbersome for generating much theoretical insight into firms' location decisions, that is, the choice of which characteristics to embody in products.

3 Modelling (horizontal) product differentiation

There are three basic families of product differentiation models that are typically used for modelling equilibrium with free entry and comparing optimal to equilibrium diversity.

Representative consumer models start by positing a utility function intended to portray aggregate preferences. This preference ordering generates the demand system for differentiated products and it measures welfare for the optimality analysis. Such functions typically embody global competition insofar as demands for varieties of the differentiated product are symmetric substitutes. Models in this class include the often-used constant elasticity of substitution (CES) preference formulation and the quadratic utility that gives rise to a linear demand system. These are parameterized utility functional forms that embody taste for variety in that more variety raises welfare even when total consumption is fixed.

The *discrete choice approach* is founded in econometric and probabilistic models of consumer behaviour. Each individual has an idiosyncratic taste (or ‘match value’) for each product. Aggregating individual choices yields the demand function and aggregating the surpluses yields the welfare function. Any i.i.d. tastes yield global competition in that products are symmetric substitutes (for example, the logit model).

Discrete choice models are not constrained to symmetric substitutability among variants. Models such as the nested logit embody closer substitutability between products within the same nest and the general probit model embodies quite elaborate substitutability patterns through the variance–covariance matrix of the match terms. These models are commonly used in the new structural empirical industrial organization literature.

Location models explicitly describe product specifications and consumer preferences as addresses and assume that consumers dislike distance ‘travelled’ between ideal type and product. Location models may also be viewed as discrete choice models because individuals make discrete choices and have idiosyncratic match values. There is a difference in interpretation: location models typically assume the population of consumers to be given and deterministic, while discrete choice models suppose that an individual's taste is a realization from a probability distribution.

In models such as the circle model, the emphasis is on the number of products produced in equilibrium and exogenous symmetric locations are effectively imposed: however, the standard symmetric location pattern can be proved to be a location equilibrium under some circumstances.

One major benefit of discrete choice and location models is that the explicit micro foundations indicate how to introduce some economic phenomenon of interest. For example, network externalities may be incorporated into consumer utility and a consistent set of demands is then generated. Representative consumer models are less satisfactory since they do not start with a population of differentiated individuals.

The different approaches are not necessarily inconsistent with or substitutes for each other. Rather, they may frequently be twinned and one approach may be reinterpreted within the setting of the others. The CES model is a variant of the logit model, and a representative consumer exists for the circle model and for probabilistic discrete choice models. Indeed, although global competition is typically generated from models such as the CES representative consumer or models of discrete choice with i.i.d. errors, it can also be derived from a spatial model if there are sufficiently many dimensions (so that each good can be a ‘neighbour’ to each other).

These models are also useful for comparative static analysis of changing patterns in industries in response to structural changes in cost structures, population growth, transport costs and consumer tastes. These descriptions are useful for urban economics, industrial organization, international trade, and economic geography.

4 Monopolistic competition and optimal variety

In Chamberlin's (1933) monopolistic competition model, products have downward sloping individual demands, yet there are so many firms that a free entry condition reasonably applies. With increasing returns to scale in production, there is a social trade-off between the benefits from variety and the costs of producing further varieties. The market equilibrium roughly embodies the same type of trade-off in so far as more firms enter if fixed costs are lower. Chamberlin concluded (although without explicit analysis) that the market equilibrium would reach ‘a sort of ideal’.

Under symmetric global competition, each entrant carves out its market share equally from established firms. Then, the number is the largest whole number at which profits are positive. This number of firms is tied down uniquely and zero profit is a reasonable approximation. Strategic behaviour by firms is scarcely relevant since there are virtually no profits to be had.

Later work showed Chamberlin to be right that the market would settle on the same amount of product diversity as the (zero-profit constrained) optimum in the central case of CES preferences (and for the logit model). Other discrete choice models lead to over-entry: this is exacerbated with asymmetric product qualities. The market may also bias against products with high fixed costs and inelastic demands. Multi-product firms choose inefficiently narrow product ranges in order to relax price competition: this effect exacerbates excessive entry of firms.

Although the symmetric CES/logit results (asymmetries aside) suggest that product differentiation is not much cause for performance concern, the alternative framework of the circle model typically generates substantial over-entry of firms.

The divergence between equilibrium and optimum product variety depends on the balance between two opposing forces. When a firm chooses to enter, it does not consider that its entry will benefit consumers. This *non-appropriation of consumer surplus* is therefore a positive externality that the firm does not internalize insofar that it cannot capture this surplus in its revenue. This force favours insufficient entry into the marketplace. It is the only force governing a multiproduct monopolist's choice of how many products to introduce, so it provides too few products. However, in a competitive setting, a firm's entry can also reduce the profits of existing firms. This *business stealing* is also not accounted for by the firm in its entry calculus because other firms’ profits do not affect its own bottom line. This negative externality encourages too many firms in equilibrium. For the CES model, these two forces exactly cancel out. For the circle model of localized competition, business stealing dominates and so there are too many firms. Loosely, prices fall quite slowly with entry in the circle model, meaning that too many firms are attracted.

5 Localized competition

Vickrey (1964) can be credited with developing several important themes of spatial competition. He formulated the circle model, finding over-entry at the equilibrium, and noting that there may be multiple equilibria under localized competition because a new entrant must fit in a niche between existing firms. An entrant's expected market space is substantially smaller than an incumbent’s. This effect is exacerbated because entrants rationally expect incumbent firms will react to new entry (in a new Bertrand–Nash price equilibrium) by cutting prices. Incumbents may earn substantially higher gross

profits than the cost of entry that would be incurred by an entrant. There are then multiple equilibria. These range from the tightest packing at which incumbents just earn zero profits (and so are not induced to exit), to a loosest packing at which incumbents earn substantial profits (and entrants will not wish to set up).

The normative economics are very sensitive to the particular equilibrium selected. Typically, the equilibrium where the incumbents just make zero profits involves too many firms, while the loosest packing equilibrium involves too few firms. It is therefore crucial to determine which equilibrium is the reasonable description. The possibility of positive profits is also very important for market conduct because firms will strive to capture the rents attributable to advantageous locations. The deadweight losses due to rent seeking should be added to any inefficiency in location choice per se. Firms may commit capital early to a market that is growing over time in order to stake claim to locations that will later be profitable. Such capacity may be sunk before it is economically viable in terms of flow profit. The equilibrium locations are those of minimum packing (maximal spacing). However, a subsidy to encourage more entry might simply raise the amount of rents that are dissipated.

Thus, while performance under global competition may not generate much cause for concern, there may be substantial welfare losses in situations characterized by a strong degree of localized competition.

6 Further applications

Product differentiation explains and resolves some other paradoxical results that obtain when products are assumed to be perfect substitutes. The *Diamond paradox* holds that the monopoly price prevails in the presence of small search costs even with many firms. Suppose consumers expected the monopoly price to be charged everywhere. Any firm pricing lower can raise its price and not lose consumers: a lower price attracts no consumers from other firms because a lower price is not expected. Any (rationally expected) price below the monopoly one is not an equilibrium because a firm can raise its price by an amount up to the search cost without losing any consumer who encounters it first. There is thus a striking discontinuity between the Bertrand and Diamond paradoxes as the search costs go from zero to a small positive value. Product differentiation smoothes the transition by allowing the consumers to shop for attributes other than purely price. A consumer may indeed find the price she expected at the first store sampled but still search further if she expects to find a better match. Her continued searching effectively brings firms into competition with each other. Firms therefore reduce prices to retain consumers who search for better matches.

The existence of a (pure strategy) price equilibrium in the original Hotelling model can be restored (through restoring profit function quasiconcavity) if there is sufficient non-locational product differentiation (through idiosyncratic preferences for products). This mechanism can restore the principle of minimum differentiation in locations, even with endogenous prices.

The standard Bertrand–Edgeworth pricing problem treats capacity constraints and, with homogenous products, has only mixed strategy equilibria. With sufficient product heterogeneity, pure strategy equilibria re-emerge since the benefits from undercutting are reduced. Likewise, standard models of positive network externalities typically exhibit multiple equilibria or no pure strategy equilibria. Unique pure strategy equilibria result with enough differentiation of products.

In international trade, product differentiation explains the empirical paradox of intra-industry trade; much bilateral trade is in the same product class. Furthermore, product differentiation is a source of

gains from trade (in addition to the traditional comparative advantage in production and factor difference reasons) because of access to larger markets supporting more variety. Endo-genous growth theory relies on product differentiation (typically with CES preferences or ‘quality ladders’ based on vertical differentiation) to rationalize continued research and development of new varieties. It is also a predominant feature as an agglomerative force in recent models of new economic geography. In macroeconomics, product differentiation models have been used to introduce imperfect competition. This is useful for providing micro-underpinnings to New Keynesian analysis. For example, in conjunction with ‘menu costs’ of switching prices, it gives rise to real effects to monetary policy.

See Also

- Chamberlin, Edward Hastings
- oligopoly
- spatial economics
- Spence, A. Michael
- Vickrey, William Spencer

Bibliography

Anderson, S.P., Andre, de P. and Thisse, J.-F. 1992. *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.

Archibald, G.C., Eaton, B.C. and Lipsey, R.G. 1986. Address models of value theory. In *New Developments in the Analysis of Market Structure*, ed. J. S. Stiglitz and G. Frank Mathewson. Cambridge, MA: MIT Press.

Berry, S., Levinsohn, J. and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–90.

Chamberlin, E. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

D’Aspremont, C., Gabszewicz, J.J. and Thisse, J.-F. 1979. On Hotelling's ‘Stability in Competition’. *Econometrica* 47, 1145–50.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39, 41–57.

Lancaster, K.J. 1966. A new approach to consumer theory. *Journal of Political Economy* 74, 132–57.

Spence, A.M. 1976. Product selection, fixed costs and monopolistic competition. *Review of Economic Studies* 43, 217–35.

Vickrey, W.S. 1964. *Microstatics*. New York: Harcourt, Brace and World. [The section ‘Spatial competition and monopolistic competition’ repr. in *International Journal of Industrial Organization* 17 (1999), 953–63.]

How to cite this article

Anderson, Simon P. "product differentiation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000201> doi:10.1057/9780230226203.1346

The New Palgrave Dictionary of Economics Online

product life cycle

Steven Klepper
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The product life cycle connotes the idea that, comparable to humans and other organisms, new industries evolve through distinct and predictable stages. When industries are young, they are subject to high product innovation, rapid output growth, a build-up in the number of producers, and flux in firm market shares. As industries age, product innovation gives way to process innovation, output growth declines, the number of producers goes through a shakeout, and firm market shares stabilize. Evidence supporting this characterization is discussed and three alternative theoretical accounts of it are reviewed.

Keywords

innovation; market share; product life cycle

Article

The product life cycle (PLC) connotes the idea that industries evolve through distinct and predictable cycles, similar to the way humans and other organisms pass through distinct stages in their lives. Originally, the idea of a PLC was proposed in the marketing literature (Levitt, 1965). Subsequently it became a rallying point for how a number of disciplines view the evolution of new industries, especially ones with rich opportunities for product and process innovation.

Typically, three stages of evolution are distinguished in the PLC (see Williamson, 1975, pp. 215–16; Clark, 1985; and Drew, 1987, for prototypical depictions). In the first stage, uncertainty about user tastes and the means to satisfy them is high, product design is primitive, unspecialized machinery is used in production, and the volume of production is low. Many firms enter and compete on the basis of product innovation, offering different variants of the industry's product. In the next stage, output growth is high, the design of the product begins to stabilize, product innovation declines, specialized machinery is substituted for labour, and the production process becomes more refined. Entry slows and exit exceeds entry, leading to a shakeout of producers. In the final stage, the industry becomes mature. Output growth

slows, product innovation becomes less significant, diversity in product offerings declines, firm market shares stabilize, entry remains low and the number of firms may continue to decline, and management, marketing and manufacturing techniques become more refined. The firms that are left in the market disproportionately tend to be those that entered early.

Industries

This conceptualization of the PLC has been heavily influenced by the history of the US automobile industry, which is summarized in Klepper and Simons (1997). Commercial production of automobiles in the United States began in 1895. By 1904 annual sales were only 22,800 cars, but from 1909 to 1919 the average annual growth in the number of automobiles sold was 25.8 per cent, and well over a million cars were sold in 1919. Subsequently annual growth slowed to 11.5 per cent through 1929 and then declined to lower levels after the Second World War and the recovery from the Great Depression. Entry initially was slow but then rose steadily through 1907, when it peaked at 82 firms. Two years later the number of firms peaked at 272. Subsequently entry declined precipitously, the number of firms steadily fell, and by 1941 only nine firms were left in the industry. Counts of product and process innovations indicate that product innovation peaked early in 1905 but process innovation increased steadily into the 1930s, with innovations such as the moving assembly line revolutionizing the production process. Firm market shares initially fluctuated greatly, but after 1910 the industry was dominated by Ford and General Motors. Based on data for 1895–1966 (Klepper, 2002), early entry provided a decided advantage. Thirteen of the 219 entrants from 1895 to 1904 survived at least 30 years, as against 3 of the 271 entrants from 1905 to 1909 and none of the subsequent 275 entrants from 1910 through 1966.

The shakeout of producers in automobiles was extreme, but Klepper and Simons (1997) document similarly severe shakeouts in tyres, television receivers and penicillin. All experienced rapid initial output growth that subsided over time. All experienced considerable entry that eventually became negligible, after which the number of firms declined for many years. In all three products firm market shares stabilized over time and the long-term survival rate was decidedly greater for earlier entrants (Klepper, 2002). The record of innovation in the three products is less well documented than in autos (Klepper and Simons, 1997). Trends in product and process innovation in tyres were similar to autos. In televisions there were only two major product innovations, both of which occurred early, but labour productivity grew steadily, suggesting no decline over time in process innovation. In penicillin process preceded product innovation, but this was largely due to a government-orchestrated war effort to reduce the cost of production of penicillin rather than market forces.

These three products were part of a larger sample of 46 new products studied by Gort and Klepper (1982) and later by Klepper and Graddy (1990), Agarwal and Gort (1996), and Agarwal (1998). The products were typically characterized by high initial growth in output that declined over time. Pronounced shakeouts were common in a majority of the 46 products. Agarwal and Gort (1996) found that early entrants had greater survival rates, although this was also true of very late entrants. No systematic evidence was compiled on product and process innovation, but Agarwal's (1998) findings suggest that products subject to greater technological change were more likely to experience shakeouts.

Theory

Numerous theories have been proposed to explain ‘excessive entry’ and shakeouts, but three stand out in their emphasis on technological change and thus their ability to address all aspects of the PLC.
Jovanovic and MacDonald (1994) develop a model of a new industry that is created by a major invention. Initially firms enter to develop the invention until expected profits are driven to zero. Subsequently another major invention occurs that opens up the possibility of an increase in the minimum efficient scale of production. It may induce immediate entry, but entrants are at a disadvantage relative to incumbents in developing the invention because of their lack of experience. Successful innovators expand their output to the new minimum efficient scale, pushing down the price of the product until non-innovators are forced to exit, which triggers a shakeout.
Utterback and Suarez (1993) envision that firms enter a new industry based on innovative designs for the industry's product. Eventually consumers and producers coalesce around a particular design for the industry's product that becomes a de facto product standard known as a dominant design. Product innovation is limited to incremental improvements in the dominant design, which makes entry more difficult. Process innovation increases because firms become less fearful that product innovation will make investments in the production process obsolete. Firms less successful at process innovation, which tend to be later entrants with less experience, exit. Coupled with the decline in entry, this gives rise to a shakeout.
Klepper (1996) develops a model of the evolution of a new industry in which firm growth is costly and firm size conditions the returns from process R&D. This imparts a competitive advantage to earlier entrants. Over time, entry and incumbent expansion causes industry output to rise and price to fall. Eventually this renders entry unprofitable and it ceases. Continued decreases in price compromise the profitability of the latest entrants, forcing them to exit, giving rise to a steady decline in the number of producers. As incumbents expand, they increase their expenditures on process R&D, causing a rise in process relative to product innovation. Firm price–cost margins also get compressed, which diminishes the incentives of firms to grow, causing firm market shares to become more stable and industry output growth to decline.
Testing of alternative accounts of the PLC is in its incipiency. Klepper and Simons (2005) found that in autos, tyres, televisions and penicillin, innovation was dominated by the leading firms and was a key determinant of firm survival, as predicted by all three theories. During the shakeouts in the four products, exit rates of early and late entrants generally did not converge, which is not consistent with the first two theories.

Variations

Not all technologically progressive industries follow the PLC (Klepper, 1997). Notably, some industries have not experienced any sign of a shakeout after 35 years and show no sign of the decline in entry that is characteristic of shakeout industries. Two examples that were in Gort and Klepper's (1982) sample of 46 new products are styrene, which is a petrochemical, and lasers. In both industries firms have ended up specializing either vertically (styrene) or horizontally (lasers). In other industries, technological developments led to turnover in the leading firms, undermining the advantages of early entry. This occurred in the disk drive industry before it went through a shakeout (Christensen, 1993). It also

occurred in autos, tyres and televisions well after their shakeouts had begun, with long-time US leaders displaced by Japanese and European firms that capitalized on small cars, radial tyres and the use of semiconductors in televisions.

These examples make clear that the PLC is a composite that only describes the prototypical evolutionary path followed by new industries. It will be an ongoing challenge to document systematic departures from the PLC and to understand the forces that contribute to them. This process is just beginning.

See Also

- competition and selection
- evolutionary economics

Bibliography

Agarwal, R. 1998. Evolutionary trends of industry variables. *International Journal of Industrial Organization* 16, 511–25.

Agarwal, R. and Gort, M. 1996. The evolution of markets and entry, exit, and survival of firms. *Review of Economics and Statistics* 78, 489–98.

Christensen, C. 1993. The rigid disk drive industry: a history of commercial and technological turbulence. *Business History Review* 67, 531–88.

Clark, K. 1985. The interaction of design hierarchies and market concepts in technological evolution. *Research Policy* 14, 235–51.

Drew, P. 1987. Despite shakeout, imaging industry not doomed to being Greek tragedy. *Diagnostic Imaging*, November, 95–9.

Gort, M. and Klepper, S. 1982. Time paths in the diffusion of product innovations. *Economic Journal* 92, 630–53.

Jovanovic, B. and MacDonald, G. 1994. The life cycle of a competitive industry. *Journal of Political Economy* 102, 322–47.

Klepper, S. 1996. Entry, exit, growth, and innovation over the product life cycle. *American Economic Review* 86, 562–83.

Klepper, S. 1997. Industry life cycles. *Industrial and Corporate Change* 6, 145–81.

Klepper, S. 2002. Firm survival and the evolution of oligopoly. *RAND Journal of Economics* 33, 37–61.

Klepper, S. and Graddy, E. 1990. The evolution of new industries and the determinants of market structure. *RAND Journal of Economics* 21, 27–44.

Klepper, S. and Simons, K. 1997. Technological extinctions of industrial firms: an inquiry into their nature and causes. *Industrial and Corporate Change* 6, 379–460.

Klepper, S. and Simons, K. 2005. Industry shakeouts and technological change. *International Journal of Industrial Organization* 23, 23–43.

Levitt, T. 1965. Exploit the product life cycle. *Harvard Business Review* 88, 81–94.

Utterback, J. and Suarez, F. 1993. Innovation, competition, and industry structure. *Research Policy* 22, 1–21.

Williamson, O. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

How to cite this article

Klepper, Steven. "product life cycle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000331> doi:10.1057/9780230226203.1347

The New Palgrave Dictionary of Economics Online

production functions

Dale W. Jorgenson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Traditionally, the production function was assumed to be additive and homogeneous. The constant elasticity of substitution (CES) production function adds flexibility by treating the elasticity of substitution as an unknown parameter, but retains the assumptions of additivity and homogeneity and imposes very stringent limitations on patterns of substitution. The dual formulation of production theory characterizes the production function by means of a dual representation such as a price or cost function, and generates explicit demand and supply functions as derivatives of the price or cost function.

Keywords

bias of technical change; CES production function; Cobb–Douglas functions; cost flexibility; cost function; demand function; economies of scale; elasticity of substitution; Hicks, J.; implicit function theorem; input–output analysis; Jorgensen, D. W.; price function; production functions; simultaneous equations models; supply function; technical change

Article

1 Introduction

The economic theory of production – as presented in such classic treatises as Hicks's *Value and Capital* (1946) and Samuelson's *Foundations of Economic Analysis* (1983) – is based on the maximization of profit, subject to a production function. The objective of this theory is to characterize demand and supply functions, using only the restrictions on producer behaviour that arise from optimization. The principal analytical tool employed for this purpose is the implicit function theorem.

The traditional approach to economic modelling of producer behaviour begins with the assumption that the production function is additive and homogeneous. Under these restrictions demand and supply functions can be derived explicitly from the production function and the necessary conditions for producer equilibrium. However, this approach has the disadvantage of imposing constraints on patterns of producer behaviour – thereby frustrating the objective of determining these patterns empirically.

The traditional approach was originated by Cobb and Douglas (1928) and was employed in empirical research by Douglas (1948, 1967, 1976) and his associates for almost two decades. The limitations of this approach were made strikingly apparent by Arrow, Chenery, Minhas and Solow (1961, henceforward ACMS), who pointed out that the Cobb–Douglas production function imposes a priori restrictions on patterns of substitution among inputs. In particular,

elasticities of substitution among all inputs must be equal to unity.

The constant elasticity of substitution (CES) production function introduced by ACMS adds flexibility to the traditional approach by treating the elasticity of substitution as an unknown parameter. However, the CES production function retains the assumptions of additivity and homogeneity and imposes very stringent limitations on patterns of substitution. McFadden (1963) and Uzawa (1962) have shown, essentially, that elasticities of substitution among all inputs must be the same.

The dual formulation of production theory has made it possible to overcome the limitations of the traditional approach to econometric modelling. This formulation was introduced by Hotelling (1932) and later revived and extended by Samuelson (1953, 1960) and Shephard (1953, 1970). The key features of the dual formulation are, first, to characterize the production function by means of a dual representation such as a price or cost function, and second, to generate explicit demand and supply functions as derivatives of the price or cost function.

Patterns of producer behaviour can be described most usefully in terms of the behaviour of the derivatives of demand and supply functions. For example, measures of substitution can be specified in terms of the response of demand patterns to changes in input prices. Similarly, measures of technical change can be specified in terms of the response of these patterns to changes in technology. The classic formulation of production theory at this level of specificity can be found in Hicks's *Theory of Wages* (1963).

Hicks (1963) introduced the elasticity of substitution as a measure of substitutability. The elasticity of substitution is the proportional change in the ratio of two inputs with respect to a proportional change in their relative price. Similarly, Hicks introduced the bias of technical change as a measure of the impact of changes in technology on patterns of demand for inputs. The bias of technical change is the response of the share of an input in the value of output to a change in the level of technology.

By treating measures of substitution and technical change as fixed parameters the system of demand and supply functions can be generated by integration. Provided that the resulting functions are themselves integrable, the underlying price or cost function can be obtained by a second integration. As we have already pointed out, Hicks's elasticity of substitution is unsatisfactory for this purpose, since it leads to arbitrary restrictions on patterns of producer behaviour.

The introduction of a new measure of substitution, the share elasticity, by Christensen, Jorgenson and Lau (1971, 1973) and Samuelson (1973) has made it possible to overcome the limitations of parametric forms based on constant elasticities of substitution. Share elasticities, like biases of technical change, can be defined in terms of shares of inputs in the value of output. The share elasticity of a given input is the response of the share of that input to a proportional change in the price of an input.

2 Models of producer behaviour

The purpose of this section is to present the simplest form of the economic theory of production. We base this theory on a production function with constant returns to scale. Producer equilibrium implies the existence of a price function, giving the price of output as a function of the prices of inputs and the level of technology. The price function is dual to the production function and provides an alternative and equivalent description of technology.

An econometric model of producer behaviour takes the form of a system of simultaneous equations, determining the distributive shares of the inputs and the rate of technical change as functions of the input prices and the level of technology. Measures of substitution and technical change give the responses of the distributive shares and the rate of technical change to changes in prices and technology. To generate an econometric model of producer behaviour we treat these measures as unknown parameters to be estimated.

In order to present the theory of production we first require some notation. We denote the quantity of output by *y* and the quantities of *J* inputs by $x_j (j = 1, 2, \ldots, J)$. Similarly, we denote the price of output by *q* and the prices of the *J* inputs by $p_j (j = 1, 2, \ldots, J)$. We find it convenient to employ vector notation for the input quantities and prices:

$$x = (x_1, x_2, \ldots, x_J) - \text{vector of input quantities}. \quad p = (p_1, p_2, \ldots, p_J) - \text{vector of input prices}.$$

We assume that the technology can be represented by a *production* function, say *F*, where:

$$y = F(x, t), \tag{1}$$

and *t* is an index of the level of technology. In the analysis of time series data for a single producing unit the level of technology can be represented by time. In the analysis of cross-section data for different producing units the level of technology can be represented by one-zero dummy variables corresponding to the different units.
We can define the *shares* of inputs in the value of output by:

$$v_j = \frac{p_j x_j}{qy}, \ (j = 1, 2, \ldots, J).$$

Under competitive markets for output and all inputs the necessary conditions for producer equilibrium are given by equalities between the share of each input in the value of output and the elasticity of output with respect to that input:

$$v = \frac{\partial \ln y(x, t)}{\partial \ln x}, \tag{2}$$

where:

$v = (v_1, v_2, \ldots, v_I)$ – vector of vaue shares . $\ln x = (\ln x_1, \ln x_2, \ldots, \ln x_I)$ – vector of logarithms of input quantities.

Under constant returns to scale the elasticities and the value shares for all inputs sum to unity:

$$i' v = i' \frac{\partial \ln y}{\partial \ln x} = 1,$$

where *i* is a vector of ones. The value of output is equal to the sum of the values of the inputs.
Finally, we can define the *rate of technical change*, say v_t, as the rate of growth of the quantity of output holding all inputs constant:

$$v_t = \frac{\partial \ln y(x, t)}{\partial t}. \quad (3)$$

It is important to note that this definition does not impose any restriction on patterns of substitution among inputs.
Given the identity between the value of output and the value of all inputs and given equalities between the value share of each input and the elasticity of output with respect to that input, we can express the price of output as a function, say *Q*, of the prices of all inputs and the level of technology:

$$q = Q(p, t). \quad (4)$$

We refer to this as the *price* function for the producing unit.
The price function *Q* is dual to the production function F and provides an alternative and equivalent description of the technology of the producing unit. We can formalize this description in terms of the following properties of the price function:

1. (1) *Positivity*. The price function is positive for positive input prices.
2. (2) *Homogeneity*. The price function is homogeneous of degree one in the input prices.
3. (3) *Monotonicity*. The price function is increasing in the input prices.
4. (4) *Concavity*. The price function is concave in the input prices.

Given differentiability of the price function, we can express the value shares of all inputs as elasticities of the price function with respect to the input prices:

$$v = \frac{\partial \ln q(p, t)}{\partial \ln p}, \quad (5)$$

where:

$$\ln p = (\ln p_1, \ln p_2, \ldots, \ln p_I) - \text{vector of logarithms of input prices}.$$

Since the price function is increasing in the input prices the value shares must be non-negative.
We can express the negative of the rate of technical change as the rate of growth of the price of output, holding the prices of all inputs constant:

$$-v_t = \frac{\partial \ln q(p, t)}{\partial t}. \tag{6}$$

Since the price function *Q* is homogeneous of degree one in the input prices, the value shares and the rate of technical change are homogeneous of degree zero and the value shares sum to unity.
We have represented the value shares of all inputs and the rate of technical change as functions of the input prices and the level of technology. We can introduce measures of substitution and technical change to characterize these functions in detail. For this purpose we differentiate the logarithm of the price function twice with respect to the logarithms of input prices to obtain measures of substitution:

$$U_{pp} = \frac{\partial^2 \ln q(p, t)}{\partial \ln p^2} = \frac{\partial v(p, t)}{\partial \ln p}. \tag{7}$$

We refer to the measures of substitution (7) as *share elasticities*, since they give the response of the value shares of all inputs to proportional changes in the input prices. If a share elasticity is positive, the corresponding value share increases with the input price. If a share elasticity is negative, the value share decreases with the input price. Finally, if a share elasticity is zero, the value share is independent of the price.
Second, we can differentiate the logarithm of the price function twice with respect to the logarithms of input prices and the level of technology to obtain measures of technical change:

$$u_{pt} = \frac{\partial^2 \ln q(p, t)}{\partial \ln p \, \partial t} = \frac{\partial v}{\partial t} = -\frac{\partial v_t(p, t)}{\partial \ln p}. \tag{8}$$

We refer to these measures as *biases of technical change*. If a bias of technical change is positive, the corresponding value share increases with a change in the level of technology and we say that technical change is *input-using*. If a bias of technical change is negative, the value share decreases with a change in technology and technical change is *input-saving*. Finally, if a bias is zero, the value share is independent of technology; in this case we say that technical change is *neutral* (in the sense of Hicks).
Alternatively, the vector of biases of technical change u_{pt} can be employed to derive the implications of changes in input prices for the rate of technical change. If a bias of technical change is positive, the rate of technical change decreases with the input price. If a bias is negative, the rate of technical change increases with the input price. Finally, if a bias is zero so that technical change is neutral, the rate of technical change is independent of the price.
To complete the description of technical change we can differentiate the logarithm of the price function twice with respect to the level of technology:

$$u_{tt} = \frac{\partial^2 \ln q(p, t)}{\partial t^2} = -\frac{\partial v_t(p, t)}{\partial t}$$

(9)

We refer to this measure as the *deceleration* of technical change, since it is the negative of rate of change of the rate of technical change. If the deceleration is positive, negative, or zero, the rate of technical change is decreasing, increasing, or independent of the level of technology.

The matrix of second-order logarithmic derivatives of the logarithm of the price function Q must be symmetric. This matrix includes the matrix of share elasticities U_{pp}, the vector of biases of technical changes u_{pt}, and the deceleration of technical change u_{tt}. Concavity of the price function in the input prices implies that the matrix of second-order derivatives, say H, is nonpositive definite, so that the matrix $U_{pp} + vv' - V$ is nonpositive definite, where:

$$\frac{1}{q} N \cdot H \cdot N = U_{pp} + vv' - V;$$

the price of output q is positive and the matrices N and V are diagonal:

$$N = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & p_J \end{bmatrix}, \quad V = \begin{bmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & v_J \end{bmatrix}$$

We can define substitution and complementarity of inputs in terms of the matrix of share elasticities U_{pp} and the vector of value shares v. We say that two inputs are *substitutes* if the corresponding element of the matrix $U_{pp} + vv' - V$ is negative. Similarly, we say that two inputs are *complements* if the corresponding element of this matrix is positive. If the element of this matrix corresponding to the two inputs is zero, we say that the inputs are *independent*. The definition of substitution and complementarity is symmetric in the two inputs, reflecting the symmetry of the matrix $U_{pp} + vv' - V$. If there are only two inputs, nonpositive definiteness of this matrix implies that the inputs cannot be complements.

To generate an econometric model of producer behaviour a natural approach is to treat the measures of substitution and technical change as unknown parameters to be estimated. For this purpose we introduce the parameters:

$$B_{pp} = U_{pp}, \; \beta_{pt} = u_{pt}, \; \beta_{tt} = u_{tt},$$

(10)

where B_{pp} is a matrix of constant share elasticities, β_{pt} is a vector of constant biases of technical changes, and β_{tt} is a constant deceleration of technical change.

We can regard the matrix of share elasticities, the vector of biases of technical change, and the deceleration of technical

change as a system of second-order partial differential equations. We can integrate this system to obtain a system of first-order partial differential equations:

$$v = \alpha_p + B_{pp} \ln p + \beta_{pt} \cdot t, \quad -v_t = \alpha_t + \beta'_{pt} \ln p + \beta_{tt} \cdot t, \tag{11}$$

where the parameters– α_p, α_t–are constants of integration.

To provide an interpretation of the parameters – α_p, α_t – we first normalize the input prices. We can set the prices equal to unity where the level of technology *t* is equal to zero. This represents a choice of origin for measuring the level of technology and a choice of scale for measuring the quantities and prices of inputs. The vector of parameters α_p is the vector of value shares and the parameter α_t is the negative of the rate of technical change where the level of technology *t* is zero.

Similarly, we can integrate the system of first-order partial differential equations (11) to obtain the price function:

$$\ln p = \alpha_0 + \alpha'_p \ln p + \alpha_t \cdot t + \frac{1}{2} \ln p' B_{pp} \ln p + \ln p' \beta_{pt} \cdot t + \frac{1}{2} \beta_{tt} \cdot t^2, \tag{12}$$

where the parameter α_0 is a constant of integration. Normalizing the price of output so that it is equal to unity where *t* is zero, we can set this parameter equal to zero. This represents a choice of scale for measuring the quantity and price of output.

For the price function (12) the price of output is a transcendental or, more specifically, an exponential function of the logarithms of the input prices. We refer to this form as the *transcendental logarithmic* price function or, more simply, the translog price function, indicating the role of the variables. We can also characterize this price function as the *constant share elasticity* or CSE price function, indicating the role of the fixed parameters. In this representation the scalars – α_t, β_{tt} – the vectors – α_p, β_{pt} – and the matrix B_{pp} are constant parameters that reflect the underlying technology. Differences in levels of technology among time periods for a given producing unit or among producing units at a given point of time are represented by differences in the level of technology *t*.

3 Economies of scale

In section 2 we have considered producer behaviour under constant returns to scale. In this section we consider producer behaviour under increasing returns to scale. Under increasing returns and competitive markets for output and all inputs, producer equilibrium is not defined by profit maximization, since no maximum of profit exists. However, in regulated industries the price of output is set by regulatory authority. Given demand for output as a function of the regulated price, the level of output is exogenous to the producing unit.

With output fixed from the point of view of the producer, necessary conditions for equilibrium can be derived from cost minimization. Where total cost is defined as the sum of expenditures on all inputs, the minimum value of cost can be expressed as a function of the level of output and the prices of all inputs. We refer to this function as the cost function.

We have described the theory of production under constant returns to scale in terms of properties of the price function (3); similarly, we can describe the theory under increasing returns in terms of properties of the cost function.

Utilizing the notation of section 2, we can define total cost, say *c*, as the sum of expenditures on all inputs:

$$c = \sum_{j=1}^{J} p_j x_j.$$

We next define the shares of inputs in total cost by:

$$v_j = \frac{p_j x_j}{c}, \ (j = 1, 2, \ldots, J).$$

With output fixed from the point of view of the producing unit and competitive markets for all inputs, the necessary conditions for producer equilibrium are given by equalities between the shares of each input in total cost and the ratio of the elasticity of output with respect to that input and the sum of all such elasticities:

$$v = \frac{\dfrac{\partial \ln y}{\partial \ln x}}{i' \dfrac{\partial \ln y}{\partial \ln x}}, \tag{13}$$

where *i* is a vector of ones and:

$$v = (v_1, v_2, \ldots, v_I) - \text{vector of cost shares}.$$

Given the definition of total cost and the necessary conditions for producer equilibrium, we can express total cost, say *c*, as a function of the prices of all inputs and the level of output:

$$c = C(p, y). \tag{14}$$

We refer to this as the *cost* function. The cost function *C* is dual to the production function F and provides an alternative and equivalent description of the technology of the producing unit.
We can formalize the theory of production in terms of the following properties of the cost function:

1. (1) *Positivity*. The cost function is positive for positive input prices and a positive level of output.
2. (2) *Homogeneity*. The cost function is homogeneous of degree one in the input prices.
3. (3) *Monotonicity*. The cost function is increasing in the input prices and in the level of output.

4. (4) *Concavity*. The cost function is concave in the input prices.

Given differentiability of the cost function, we can express the cost shares of all inputs as elasticities of the cost function with respect to the input prices:

$$v = \frac{\partial \ln c(p, y)}{\partial \ln p}. \tag{15}$$

Since the cost function is increasing in the input prices, the cost shares must be non-negative.
We can define an index of returns to scale as the elasticity of the cost function with respect to the level of output:

$$v_y = \frac{\partial \ln c}{\partial \ln y}(p, y). \tag{16}$$

Following Frisch (1965), we can refer to this elasticity as the *cost flexibility*. The cost function is increasing in the level of output, so that the cost flexibility is positive. Since the cost function C is homogeneous of degree one in the input prices, the cost shares and the cost flexibility are homogeneous of degree zero and the cost shares sum to unity.
The cost flexibility v_y is the reciprocal of the *degree of returns to scale*, defined as the elasticity of output with respect to a proportional increase in all inputs:

$$v_y = \frac{1}{i' \frac{\partial \ln y}{\partial \ln x}}. \tag{17}$$

If output increases more than in proportion to the increase in inputs, cost increases less than in proportion to the increase in output.
We have represented the cost shares of all inputs and the cost flexibility as functions of the input prices and the level of output. We can characterize these functions in terms of measures of substitution and economies of scale. We obtain *share elasticities* by differentiating the logarithm of the cost function twice with respect to the logarithms of input prices:

$$U_{pp} = \frac{\partial^2 \ln c(p, y)}{\partial \ln p^2} = \frac{\partial v(p, y)}{\partial \ln p}. \tag{18}$$

These measures of substitution give the response of the cost shares of all inputs to proportional changes in the input

prices.
Second, we can differentiate the logarithm of the cost function twice with respect to the logarithms of the input prices and the level of output to obtain measures of economies of scale:

$$u_{py} = \frac{\partial^2 \ln c(p, y)}{\partial \ln p \, \partial \ln y} = \frac{\partial v(p, y)}{\partial \ln y} = \frac{\partial v_y(p, y)}{\partial \ln p}. \tag{19}$$

We refer to these measures as *biases of scale*. The vector of biases of scale u_{py} can be employed to derive the implications of economies of scale for the relative distribution of total cost among inputs. Alternatively, this vector can be employed to derive the implications of changes in input prices for the cost flexibility. To complete the description of economies of scale we can differentiate the logarithm of the cost function twice with respect to the level of output:

$$u_{yy} = \frac{\partial^2 \ln c(p, y)}{\partial \ln y^2} = \frac{\partial v_y(p, y)}{\partial \ln y}. \tag{20}$$

The matrix of secon-order logarithmic derivatives of the logarithms of the cost function C must be symmetric. This matrix includes the matrix of share elasticities U_{pp}, the vector of biases of scale u_{py}, and the derivative of the cost flexibility with respect to the logarithm of output u_{yy}. Concavity of the cost function in the input prices implies that the matrix of second-order derivatives, say H, is nonpositive definite, so that the matrix $U_{pp} + vv' - V$ is nonpositive definite, where:

$$\frac{1}{c} N \cdot H \cdot N = U_{pp} + vv' - V.$$

Total cost c is positive and the diagonal matrices N and V are defined in terms of the input prices p and the cost shares v, as in Section 2.
We say that the cost function C is *homothetic* if and only if the cost function is separable in the prices of all J inputs $\{p_1, p_2, \ldots, p_J\}$, so that:

$$c = C[P(p_1, p_2, \ldots, p_J), y], \tag{21}$$

where the function P is homogeneous of degree one and independent of the level of output y. The cost function is homothetic if and only if the production function is *homothetic*, where:

$$y = F[G(x_1, x_2, \ldots, x_l)], \tag{22}$$

where the function *G* is homogeneous of degree one.
Since the cost function is homogeneous of degree one in the input prices, it is homogeneous of degree one in the function *P*, which can be interpreted as the price index for a single aggregate input; the function *G* is the corresponding quantity index. Furthermore, the cost function can be represented as the product of the price index of aggregate input *P* and a function, say *H*, of the level of output:

$$c = P(p_1, p_2, \ldots, p_l) \cdot H(y). \tag{23}$$

Under homotheticity, the cost flexibility v_y is independent of the input prices:

$$v_y = \frac{\partial \ln H}{\partial \ln y}(y). \tag{24}$$

If the cost flexibility is also independent of the level of output, the cost function is homogeneous in the level of output and the production function is homogeneous in the quantity index of aggregate input *G*. The degree of homogeneity of the production function is the degree of returns to scale and is equal to the reciprocal of the cost flexibility. Under constant returns to scale the degree of returns to scale and the cost flexibility are equal to unity.
We can generate an econometric model of cost and production by introducing the parameters:

$$B_{pp} = U_{pp}, \beta_{py} = u_{py}, \beta_{yy} = u_{yy}, \tag{25}$$

where B_{pp} is a matrix of constant share elasticities, β_{py} is a vector of constant biases of scale, and β_{yy} is a constant derivative of the cost flexibility with respect to the logarithm of output. We can treat the matrix of constant parameters as a system of second-order partial differential equations, obtaining:

$$v = \alpha_p + B_{pp}\ln p + \beta_{py}\ln y, v_y = \alpha_y + \beta'_{py}\ln p + \beta_{yy}\ln y, \tag{26}$$

where the parameters – α_p, α_y – are constants of integration.

We can integrate the system (26) to obtain the cost function:

$$\ln c = \alpha_0 + \alpha_p \ln p + \alpha_y + \frac{1}{2}\ln p' B_{pp} \ln p + \ln p' \beta_{py} \ln y + \frac{1}{2}\beta_{yy}(\ln y)^2, \quad (27)$$

where the parameter α_0 is a constant of integration. We can refer to this form as the translog cost function, indicating the role of the variables, or the constant share elasticity (CSE) cost function, indicating the role of the parameters. Under homotheticity the cost flexibility is independent of the input prices. A necessary and sufficient condition for homotheticity is given by:

$$\beta_{py} = 0; \quad (28)$$

the vector of biases of scale is equal to zero. Under homogeneity the cost flexibility is also independent of output, so that:

$$\beta_{yy} = 0;$$

the derivative of the flexibility with respect to the logarithm of output is zero. Finally, under constant returns to scale, the cost flexibility is equal to unity; given the restrictions implied by homotheticity, constant returns requires:

$$\alpha_y = 1. \quad (29)$$

4 Summary and conclusion

The econometric modelling of producer behaviour requires parametric forms for demand and supply functions. Patterns of production can be represented in terms of unknown parameters that specify the responses of demands and supplies to changes in prices, technology and scale. New measures of substitution, technical change and economies of scale have provided greater flexibility in the empirical determination of production patterns. These innovations have arisen from the dual formulation of the theory of production.

We can conclude by suggesting possible directions for future research. The primary focus of our discussion has been on the characterization of technology for individual producing units. Application of the results typically involves models for both demand and supply of a given commodity. The ultimate objective of econometric modelling of production is to construct general equilibrium models encompassing demands and supplies for a wide range of products and factors of production, along the lines suggested by Jorgenson (1983).

Our exposition of the theory of production has emphasized models where the econometric methodology has crystallized.

An important area for future research is the implementation of dynamic models of technology. These models are based on substitution possibilities among outputs and inputs at different points of time. The simplest intertemporal model of production is based on capital as a factor of production. This model is treated in a companion paper by Jorgenson (1986). A number of promising avenues for further investigation have been suggested in the literature on the theory of production summarized in the entry on vintages.

See Also

- CES production function
- Cobb–Douglas functions

Bibliography

Arrow, K.J., Chenery, H.B., Minhas, B.S. and Solow, R.M. 1961. Capital–labor substitution and economic efficiency. *Review of Economics and Statistics* 63(3), 225–47.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1971. Conjugate duality and the transcendental logarithmic production function. *Econometrica* 39(3), 255–6.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55(1), 28–45.

Cobb, C.W. and Douglas, P.H. 1928. A theory of production. *American Economic Review* 18, 139–65.

Douglas, P.H. 1948. Are there laws of production? *American Economic Review* 38, 1–41.

Douglas, P.H. 1967. Comments on the Cobb–Douglas production function. In *The Theory and Empirical Analysis of Production*, ed. M. Brown. New York: Columbia University Press, 15–22.

Douglas, P.H. 1976. The Cobb–Douglas production function once again: its history, its testing, and some empirical values. *Journal of Political Economy* 84, 903–16.

Frisch, R. 1965. *Theory of Production*. Chicago: Rand McNally. (English translation from the 9th edn of lectures published in Norwegian; the 1st edn of the lectures dates from 1926.)

Hicks, J.R. 1946. *Value and Capital*. 2nd edn. Oxford: Oxford University Press.

Hicks, J.R. 1963. *The Theory of Wages*. 2nd edn. London: Macmillan.

Hotelling, H.S. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40, 577–616.

Jorgenson, D.W. 1983. Modeling production for general equilibrium analysis. *Scandinavian Journal of Economics* 85 (2), 101–12.

Jorgenson, D.W. 1986. Econometric methods for modeling producer behavior. In *Handbook of Econometrics*, vol. 3, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

McFadden, D. 1963. Further results on CES production functions. *Review of Economic Studies* 30(2), 73–83.

Samuelson, P.A. 1953–4. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21(1), 1–20.

Samuelson, P.A. 1960. Structure of a minimum equilibrium system. In *Essays in Economics and Econometrics*, ed. R. W. Pfouts. Chapel Hill: University of North Carolina Press, 1–33.

Samuelson, P.A. 1973. Relative shares and elasticities simplified: comment. *American Economic Review* 63, 770–1.

Samuelson, P.A. 1983. *Foundations of Economic Analysis*. 2nd edn. Cambridge, MA: Harvard University Press.

Shephard, R.W. 1953. *Cost and Production Functions*. Princeton: Princeton University Press.

Shephard, R.W. 1970. *Theory of Cost and Production Functions*. Princeton: Princeton University Press.

Uzawa, H. 1962. Production functions with constant elasticity of substitution. *Review of Economic Studies* 29(4), 291–9.

How to cite this article

Jorgenson, Dale W. "production functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000204> doi:10.1057/9780230226203.1348

The New Palgrave Dictionary of Economics Online

profit and profit theory

Meghnad Desai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

abstinence theory of profits; aggregate production function; Arrow–Debreu model of general equilibrium; capital accumulation; capital theory; Clark, J.B.; Cobb–Douglas functions; constant and variable capital; contingent commodities; depreciation; division of labour; entrepreneurship; exchange value and use value; exploitation; falling rate of profit; innovation; interest; investment rate; joint production; Kaldor; Kalecki, M.; Keynes, J. M.; Kondratieff cycles; labour power; Lamberton, D. M.; Luxemburg, R.; marginal productivity theory; Marx, K. H.; organic composition of capital; Physiocracy; population growth; produit net; profit and profit theory; quasi-rent; rent; risk; saving and investment; Schumpeter, J. A.; Shackle, G. L. S.; Sraffa, P.; surplus; surplus value; Taylorism; technical progress; technology; uncertainty; use value; variable capital; von Neumann, J.

Article

A theory of profit should address itself to at least three questions – about the size (volume) of profit, its share in total income and about the rate of profit on capital invested. Each of these three issues (size, share, rate, hereafter) can be examined at three separate levels of aggregation, the firm, the industry or the economy.

Theories of profits of size, share or rate can be classified as to whether they treat profits as an equilibrium or as a disequilibrium phenomenon, and of course as to whether the equilibrium is a static or a dynamic one. Theories then deal with the role of profits, that is, the effects of the size, share or rate on other economic variables, and the source of profits, that is, the variables which cause profits to be what they are. An issue related to the cause of profits is the moral justification of the claimant to profits. This issue though prominent in the classical and Marxian economics, disappears in neoclassical economics with the triumph of the marginal productivity theory of distribution. But the issue reappeared during the controversy surrounding capital theory in the 1960s.

Before we go into these issues, we have to ask whether profit is a pure category in economics since a theory presumes the existence, in the abstract at least, of a definable category to which the theory is

addressed. With profits, there has been a frequent problem of conflating it with interest and rent.
One class of theories have taken profits as synonymous with interest. Theories have failed to distinguish consistently between profits and interest as categories of non-labour income. Thus, theories for the existence of positive rates of interest are often thought of as theories of profit rate. The abstinence theory of profits, attributed to Nassau Senior, is as much a theory of interest as of profits. In early classical writing, in Adam Smith for example, profits and interest do not appear as separate income categories. Wages, rent and profits are the three divisions of total income. When the interest rate does make its separate appearance, it is as the limit to which the rate of profit can fall. The ambiguous relationship between the two reappears in the marginal productivity theory where the return to capital (real rental on capital) is not sufficiently distinguished from interest to be a separate category of income.
There are also other conventions which lead to a loose definition of profits. Profits are sometimes used as synonymous with all non-labour income, subsuming rent and interest incomes. This is very much a Marxian tradition which has come down into modern-day Cambridge growth theory via Kalecki's theory of distribution. Alternatively, Marshallian theories subsume profits under a general category of quasi-rent. This leads to the further distinctions made between normal and supernormal profits or pure profits.

Neoclassical theory

In neoclassical theory, the competitive firm (the entrepreneur) maximizes the quantity of profits to decide the level of output and inputs. This gives the price equal marginal cost condition. In equilibrium, the *size* of profits is indeterminate as are the *rate* and *share*. In long-run competitive equilibrium of the industry (if such an equilibrium can be shown to exist) the firm has zero (actual or if there is uncertainty, expected) profits; price equals average cost and the output is produced at the lowest point of the U-shaped average cost curve. Thus, zero profit is an equilibrium condition and also a signal that output is produced under efficient conditions. This zero profit condition is often qualified by adding that this does not rule out ‘normal’ profits. This not only leaves normal profits indeterminate in size but could easily lead to the condition of zero profits being a tautology. In this paradigm, non-zero profits are an indication of (long-run) disequilibrium or of non-competitive conditions (for example, barriers to entry). A third alternative is that in the presence of uncertainty any observed non-zero profits may be the random deviation of actual from expected profits. The zero profit condition is best thought of not as a descriptive prediction but as a rule to check for consistency in any model that assumes competitive behaviour.
In terms of the *rate of profit* the condition for competitive equilibrium is that the rate of profit be equal in all activities (industries, sectors, and so on). Here again the rate of profit is itself indeterminate but it is the interindustrial differential in the rate which needs to be zero in equilibrium. Again in analogy with the size of profits, a persisting non-zero differential indicates either disequilibrium or imperfectly competitive elements.
Normal non-zero level of profits or persisting differential in a particular firm or industry can be reconciled with competitive equilibrium by an appeal to Marshall's doctrine of quasi-rent. Ricardo's theory of rent compels the marginal land to have zero rent but supramarginal lands to earn positive rent. Marshall extends this logic to other factor incomes with the doctrine of quasi-rent relying on restrictions on elasticity of factor supply in the particular case where differential rates of return or non-zero surplus incomes are found. Normal non-zero profits could then be a quasi-rent.

But if so, what is the factor of production whose income is profit (as quasi-rent)? This raises the contentious question of the moral and economic basis of profit as income. Is profit the return to capital or is it the remuneration of the capitalist as a manager/entrepreneur? Attempts to provide a justification for profits as income invoke the abstinence doctrine, or the ‘residual claimant’ argument in the 19th century. But abstinence could provide a theory of reward for savings, that is, a theory of the interest rate but not for a reward for capital unless all savers are also capitalists. The residual claimant theory, that is, profits are what is left over after every other input has been paid its due, is hardly a theory. It needs to be supplemented by a theory of how all other inputs are rewarded, that is, how the residual is determined. It was the achievement of J.B. Clark's marginal productivity theory of distribution to provide such a link by an attempt to integrate production and distribution via the marginal principle. Clark's theory claimed to explain distribution at each of the three levels of aggregation. The equating of the marginal value product of a factor to its unit price is a principle which treats labour and non-labour inputs symmetrically. In this theory, capital appears as equipment and its marginal revenue product is equated to its unit price. If one could further identify the profit per unit of capital with the unit price of capital, then the return to the capitalist (profits) is the reward of the productivity of capital (the rental for the factor). This brings profits in an analogous relation with rent. The capitalist is one who owns the factor of production capital. It is the structure of property rights as well as the productivity of capital which combine to make the owner of capital the recipient of its fruits.

In equilibrium, the price of the capital equipment will be the present value of its future net income stream. This is the contribution of Irving Fisher. Under certain restrictive assumptions – known future income stream, constant relative price of capital and constant and known discount factor – the real rental of capital is equal to the sum of the rate of discount (the rate of interest) and the rate of depreciation of the capital good. The presence of the rate of depreciation in the formula is required only in the case of a physical durable good. The rate of depreciation may of course not be constant but variable; worse, it could be endogenous and dependent on the forces determining the rental on capital (the rate of investment, for example). If, on the other hand, one were to take an infinitely lived capital good, that is, zero depreciation rate, then the real rate of return (the rate of profit) ‘degenerates’ to the rate of interest. In such a case an infinitely lived capital good becomes like a financial asset – a consol.

This illustrates the difficulty of separating profits from rent or interest. The problem here is that capital can be a physical good (a machine/a building), a financial asset (bond/consol) or an abstract attribute (human capital/skills). The marginal productivity notion relates to capital only when it is a material input to production. This is why the specification of the production function used to explain the rate of profit as arising from the marginal productivity of capital as factor becomes a contentious issue. Once capital is a physical good, its durability and heterogeneity entail an assumption of malleability if we wish to add up the disparate units of capital to arrive at an aggregate capital stock. This aggregate may be at the level of a firm, or an industry, but it is at the highest level of the economy that the problem is serious.

The need to have an aggregate measure arises from the practice of using an aggregate production function in terms of labour and capital, both assumed to be homogeneous aggregates. The use of the aggregate production function had begun in the 1920s and 1930s where on the one hand, Paul Douglas and Charles Cobb had used their well-known function to explain the constancy of the share of labour in total income (Cobb and Douglas, 1928; Douglas, 1948). Theorists such as Hicks, Harrod and Joan Robinson had also used the device of an aggregate production function to define various measures of technical progress which would reconcile the stylized facts of the constancy of the share labour, with a

rising real wage and a trendless rate of profit. Thus it was that technical progress measures such as Hicks-neutral, Harrod-neutral, and so on, were proposed as simple constructions to explain these stylized facts. The theorists' work did not require the Cobb–Douglas form as such but the latter proved a convenient way for expressing these forms explicitly in econometric work in the 1950s when economic growth and technical progress engaged the attention of economists. It was Solow's work both in proving the stability of growth equilibrium and in measuring the contribution of technical progress to economic growth which generated the veneration of the Cobb–Douglas production function (Solow, 1956; 1957).

This use of the aggregate production function in neoclassical growth theory in the 1950s accomplished two things. It could link the rate of growth of the economy to the rate of profit; in some cases the two could be equal. At the same time, the Cobb–Douglas form could be used to link the rate of profit to the marginal productivity theory. Thus, a microeconomic firm theoretic proposition – the equality of marginal productivity and factor prices – could be exploited to explain macroeconomic magnitudes of factor shares and growth.

This brilliant device linking the marginal productivity theory of distribution with a macroeconomic growth theory soon ran into difficulties and unravelled itself. The capital theory (or Cambridge–Cambridge) controversy is dealt with elsewhere in this dictionary. For the purpose of this essay, it suffices to note that while the factor prices could be taken as given by the firm, they become endogenous at the macroeconomic level. The return to capital – the real rental or the profit rate – depends on the rate of investment. To explain the latter in terms of demand for capital in terms of factor–price substitution involves circular reasoning. Further, the existence of an aggregate capital stock presumes the prior existence of equilibrium prices for the heterogeneous capital goods comprising such an aggregate. The aggregate then cannot be used to 'explain' equilibrium profit rate. There seemed to be some insuperable logical problems in the notion of an aggregate capital stock. As an explanation of the profit rate, the macroeconomic theory of distribution proved to be a cul-de-sac.

An alternative treatment of the return to capital would be to abstract from such complications and treat capital as a commodity like any other. Using Debreu's very general definition of a commodity, new capital and old capital become different commodities, heterogeneous capital goods retain their differences and do not need to be aggregated. The task of the theory is then to show that a set of non-negative prices will clear markets for all the commodities. Given the facts of time and uncertainty we create dummy markets for contingent commodities. The price of the commodity 'capital' can then be determined for each time period and each state of nature.

But while the price of capital good can be determined by this method, it has no further link with profits. Profits in the Arrow–Debreu equilibrium are zero. Owners of capital goods will receive the price as their reward but this is not profit. Nor need the price of capital good have any specific connection with the rate of interest; as an intertemporal price the interest rate links the price of any commodity to that of its substitute available at a different date.

The Arrow–Debreu theory invokes a very stylized sort of uncertainty. Going back to a distinction made by Frank Knight between risk and uncertainty, it is risk rather than uncertainty which is involved in the Arrow–Debreu notion of contingent commodities. It assumes that states of nature can be described fully and the probabilities of various outcomes under different states of nature calculated in advance of determining the excess demand functions (or correspondences) for such contingent commodities. Such risk being previsible and insurable against, will yield only such return which cannot be arbitraged away. There can be no pure profit in such an equilibrium.

Classical theories of profit (Smith and Ricardo)

In classical economics profits are important as a quantity. Together with rent, they constitute the economic surplus, wages being the *faux frais* of production. The distribution of the surplus as between the owners of stock (capitalists) and the landlords becomes a central issue. This is because the two classes were presumed to have different propensities for productive consumption (accumulation). The division of surplus then had growth consequences.

The Physiocrats regarded land (nature) as the only source of surplus. Profits were only a recycled part of the *produit net*. Adam Smith could be read as regarding a surplus producing agriculture as a necessary but not the sole source of surplus, since productive labour was another source. The productivity of both land and labour depended on constant improvements in the manner of their utilization – agricultural practices and division of labour. Thus behind the 'factors' was the incessant propensity for progress – technical progress in the narrow sense as well as general innovations and improvements in practices.

The size of profits is ambiguous in Smith as there are no clear rules about the division of the surplus between rent and profit. The rate of profit was expected by Smith to fall. The reason for this was not diminishing returns in agriculture, as it was to be for Ricardo. The division of labour and effects of trade could counter the limits imposed by the soil. It was however that the stock of profitable investment opportunities was limited within a country. Smith could see that the fabulous profits made by trading companies were shrinking as merchants began to compete. Profits on industrial activities were on the moderate side. Thus the falling rate of profit hypothesis emerges to recognize an empirical fact though Smith had no reason for believing that technical progress could not stave off the tendency indefinitely.

It is David Ricardo who relates profits and rent antagonistically and has a clear view of the rate of interest as defining the lower limit of the rate of profit. In Ricardo, there is the first rigorous attempt to define the rate of profit, as a pure number which will be free of problems of valuation. By defining a one-good economy, corn, inputs and outputs can be measured in the same commodity. This is done by defining the wage in terms of corn and any equipment as product of labour. Given these two assumptions, the profit rate is defined as the ratio of net surplus – output of corn less inputs defined in terms of corn – to the capital advanced also measured by the inputs defined in terms of corn. But given the conflict in the economy between rent and profits, how was rent to be accounted for?

Ricardo defined the marginal land as zero rent yielding land. Thus the pure rate of profit could be defined on the marginal land free of complications introduced by rent. If we then add that profit rates equalize everywhere, the rent yielding land does so by achieving a superior output–input ratio compared with the marginal land. The difference is rent. Rent is an unearned income accruing to the landlords as a result of the progress of accumulation and the growth of population.

Ricardo integrates the conflict between rent and profit with a theory of growth which gives an explanation of the falling rate of profit. The size of profit as shared out with rent, determines the rate of accumulation since capitalists have a high propensity to accumulate. As accumulation proceeds, there are diminishing returns to land as well as labour. If the real wage were to be taken as constant, then the surplus above wage shrinks due to diminishing productivity of labour. But at the same time, rents rise. Hence profit is squeezed out. The rate of profit falls with accumulation.

But on the zero rent land, profits are antagonistic to wages. Thus any tendency for wages to go up will reduce the rate of profit. The determinants of real wages were accumulation (demand for labour) and

growth of population (supply of labour). The Malthusian mechanism to keep real wages constant worked at least in the long run. But even in the short run, although theoretically real wages could rise with the force of accumulation, the empirical facts of population growth since the 1780s and the potentially high participation rate of men, women and children, meant that for the period in which he was writing, Ricardo could easily assume rapid adjustment of the supply of labour to a rise in the real wage rate.

New classical theories of profit (Sraffa, von Neumann)

There are at least two directions in which Sraffa (1960) generalizes Ricardo. First he drops the single-good assumption but defines a standard commodity in terms of which the rate of profit is invariant to relative price changes. He also drops the assumption of a constant subsistence wage. By taking wages and profits as constituting surplus, that is, wages no longer being merely the costs of production but part of value added, the conflict between the rate of profit and the share of wages in total surplus is made explicit. The wage–profit frontier derived from the structure of the input–output information illustrates this conflict.

If we were to take Ricardo's choice of corn as the single good as a substantive and not merely a methodological device, one must attribute a large role to agriculture as the source of surplus. Ricardo is less explicit about the role of technical progress either in generating surplus or in staving off the effects of diminishing returns. By dropping land from the general model altogether as an essential input, Sraffa implicitly takes the technical conditions of production (the matrix of the input coefficients) as guaranteeing that a surplus exists. He does not however pursue the question of the disposal of profits, that is, accumulation.

In a parallel but independent work, von Neumann put forward a linear model of the economy with joint production, in which the timing of input and output was articulated. In this system, there are several processes available for producing a commodity but only those are chosen which at least yield a certain rate of surplus (say g) of the output above the inputs. Labour is one of the inputs. The converse of this proposition is that given the input and output prices, only those activities are chosen which yield the minimum rate of profit (p). If the linear system of production coefficients is indecomposable, from the duality of prices and quantities, it follows that $g=p$. If we now regard g as the rate of growth between this year's inputs and next year's output, we have in von Neumann's model, an equality between the rates of growth and profit. All profits need to be accumulated and wages have to be at subsistence level. The classical lineaments of von Neumann's model are thus clear (von Neumann, 1945–6).

Von Neumann's system leaves the size and share of profits indeterminate while making its rate determinate. It is his device of defining the production process as jointly producing final output and one year older capital goods, that is, the joint production technology that has proved a fertile innovation. In one way this device avoids the problem of heterogeneity as well as of durability of capital. Each item of capital equipment can be defined as a separate commodity and can be given a one period length of life after which it becomes another commodity, albeit a one year older version of itself. The problems of measurement of capital which plague the neoclassical aggregate production function are thus avoided. Also a strict separation is made between price of capital goods and profits. Finally the source of profits is seen as the technology which permits growth.

Technological progress is not endogenized in neoclassical economics, nor in classical economics except in a loose sense in the grand design of Adam Smith. Von Neumann as well as Sraffa leaves the question

well alone. It is with Karl Marx that an ambitious model is attempted of a fully endogenous model of growth, accumulation and technological progress.

Profits as exploitation (Marx)

Marx makes an explanation of the source of profits a central part of his theory. There are two ways of understanding this central role of profit in Marx. The *measure* of economic surplus being labour in the classical theory, an immediate question arose among certain socialist followers of Ricardo as to whether labour was also the *source* of all surplus and hence should be its *sole recipient*. While this is not the case in Smith or Ricardo, for Marx labour becomes a measure of value, and the source of surplus value. Surplus value takes the money form of profits via the mediation of prices, which are formed on the basis of values.

A second motivation of Marx's theory could be seen as extending Ricardo's critique of rent as an unearned income to the category of profits. The scarcity of fertile land is not a natural but a social phenomenon in Ricardo, caused by the progress of accumulation and population. Was the scarcity of capital and the fact that it commanded a surplus equally social? Marx asked.

Marx's theory of profits is that profits are the money form of surplus value produced by labour in the production process but appropriated by the owners of means of production. The capitalist advances capital to buy labour and means of production. But what he buys is labour power, the capacity for work. This is because the labour contract is a voluntary, *ex ante* agreement on the part of the labourer to work a fixed period of time – the length of the working day in return for a wage. The wage is the *exchange value* of the commodity labour power. The *use value* of the labour power is whatever productivity the capitalist can extract from the worker during the working day. There is a gap between the use value and exchange value of labour power but this gap cannot be seized by the sellers of labour power, but only by the buyers. This is because the buyers of labour power, the capitalists, enjoy a class monopoly of ownership of the means of production. Without finding a buyer for the labour power, the labourer cannot reproduce himself, that is, he cannot survive for any length of time with his working capacity intact. Thus, there is an asymmetry in the positions of the workers and the capitalists, as a result of a historical process that has deprived workers of direct access to means of production.

The gap between use value and exchange value present in the case of labour power is not present in the case of capital goods. These are bought and sold by capitalists and hence there is no scope for unrealized gaps to exist for any length of time without being captured by the seller. This is why Marx called the flow of input services from capital goods *constant* (*c*) capital, i.e. capital which had the same value in the beginning as at the end of the production process. For labour, the flow of input services – the use value realized – was made up of necessary labour measured by the exchange value, i.e. wage and surplus labour which accrued to the capitalist. This is why labour was *variable* (*v*) capital, it changed value between the time it was bought/paid for and the end of the production process.

The rate of profit, for Marx, was then the ratio of surplus value (*s*) to the sum of constant and variable capitals $(c + v)$; $[\rho = s / (c + v)]$. All the three terms of the ratio are measured in labour time and are commensurate. Thus, Marx also attempted to arrive at a measure of the rate of profit which would be invariant to relative price changes. But, unlike Ricardo, he did not assume a single good economy. What is needed is that the wage rate can be converted into labour values. The same has to be done about flow of services emanating from the means of production. This requires that the production technology be of

a form which makes such conversion of goods into their labour values problem free.

The source of profits in Marx is the exploitation of labour by the capitalists, although it is subsumed that the technology is such as to yield a surplus, that is, a gap between the use value and exchange value of labour power. Positive surplus value is seen as a necessary and sufficient condition for profits to be positive (Okishio, 1963; Morishima, 1973). Profits are accumulated, and put into ever improving technology as a result of the competition among capitalists. It is in the constant search for higher profits, partly immanent and partly as a response to factors threatening the rate of profit, that the incessant improvement in technology takes place. Capitalists have to find better ways of increasing profits, by any means which can increase the gap between the exchange value and use value of labour power. They may do this by improving working methods (Taylorism), by extending hours of work without increasing the wage (absolute exploitation) or by investing in improved machinery (relative exploitation).

Marx in common with the classical economists also has a theory of the falling rate of profit. The progress of accumulation raises the wage rate and hence with a given technology lowers the ratio of surplus value to variable capital, that is, lowers the rate of surplus value (s/v). In order to stave off this danger, capitalists are compelled to use labour-replacing technology. This raises the organic composition of capital, that is, the share of constant capital in total capital [$c/(c+v)$]. The rate of profit varies directly with the rate of exploitation and inversely with the organic composition of capital. In the pure theoretical model, Marx reasoned that the balance would be such that the rate of profit would fall. He also discussed a number of countervailing tendencies such as the growth of monopolies in particular and a high degree of concentration in the industrial structure which may arrest this tendency of the rate of profit to fall.

Accumulation in Marx proceeds incessantly but not at a constant or equilibrium rate. The search for higher profits drives accumulation and accumulation in turn increases surplus value by being embodied in better techniques. But accumulation acts on a labour force that ultimately even exhausts its reserve army and thus wages threaten to rise. On the other hand, markets have to be found for the larger quantities of goods being produced at prices which will yield a profit, that is, surplus value has to be *realized* in the market, not just extracted from labour. The result is that the accumulation process facing these two limits results in a cyclical growth pattern. This pattern yields cycles around a declining rate of profit.

Marx's attempt to obtain a pure (relative price invariant) measure of the rate of profit has been criticized mainly due to the problem of evaluating the different types of skilled labour used in production. Marx's theory requires that all types of labour be reducible to homogeneous labour. Such reduction cannot be made without a measure of relative value productivity of different types of labour, independently of their market rates of remuneration. Despite much ingenious work, this problem has proved intractable.

Another problem arises from the durability of capital. In a joint production formulation, it can be shown that positive surplus value is neither necessary nor sufficient for positive profits. The theoretical formulation has to be amended to rule out non-convexities which lead to the curiosum that negative surplus value can lead to positive profits (Steedman, 1976; Morishima, 1974). A third problem arises from the fact that an example of accumulation, with balanced growth and a constant rate of profit, was provided by Marx himself in his Scheme for Expanded Reproduction contradicting the necessity of cyclical accumulation or of a falling rate of profit.

Profit as disequilibrium (Schumpeter, Keynes)

Another ambitious attempt to combine profits and growth was made by Schumpeter. With Schumpeter, profits become a *disequilibrium* phenomenon. He advances a theory of the size of profits, especially the source of profits but none of either the rate or the share of profit. Schumpeter's theory is also the only one where there is a clear link between the monetary system that finances production and the real system that generates profits. Schumpeter's is also the only theory which makes the agency of the profit earner – the entrepreneur – an explicitly central part of the theory.

The source of profits is innovation. Innovations can comprise introduction of a new good, of a new method of production, the opening of a new market, discovery of a new source of supply of raw material, or the carrying out of a new organization of an industry. The economy is supposed to be in a state of stationary or steady-state equilibrium before the innovation occurs. An entrepreneur as a visionary innovates by launching a new product or a new technique, and so on. Such a new product may have a long gestation lag before it earns revenue and, as such, may be risky. Thus in the financing of innovations, credit plays an active role. Such credit will be excess to current goods supply and will cause inflation. Once the innovation appears on the market, the entrepreneur makes monopoly profits. The credit initially created can be liquidated out of profits but the innovation causes further ripples via backward and forward linkages as well as by attracting imitators. Innovations occurring singly or in a cluster set off a long wave, a Kondratieff cycle. In the rising phase, prices, profits and output rise. But eventually the monopoly profits are bid away by competitors and the system returns to equilibrium, with profits tending to zero.

The innovation process is discontinuous and disequilibrating. It is accompanied by a credit boom and a cyclical upturn. Innovations are unanticipated. The economy exists in cycles caused by innovations but tending towards an equilibrium of zero profits, once the innovation has spent itself. The history of capitalism was for Schumpeter made up of successive long waves caused by clusters of innovations. Thus for Schumpeter the source of profits is the superior productivity achieved by innovations but the agent of change is the entrepreneur. Neither the conventional industrialist nor labour generates profits. Profits are by nature abnormal, disequilibrium phenomena. They do not persist but dissipate in equilibrium.

Schumpeter thus reconciles a zero profit stationary equilibrium with observed facts of profits. But while the theory is an appealing one, it has lacked sufficient analytical detail to prove either a source of further developments in profit theory or a tool for empirical investigation.

While Schumpeter put forward a dynamic theory of the disequilibrium role of profits, his model is sparse in details. Keynes in his *Treatise on Money* also treats profits as disequilibrium and insists that national income calculations exclude profits. The emergence of profits as a disequilibrium category comes from the gap between savings and investment. The famous Fundamental Equations of the *Treatise* describe a two sector model with consumption goods and investment goods. For each sector, the price level is made up of unit labour cost and a disequilibrium item. For the consumption goods sector, this item consists of the cost of production of investment goods less savings. Following a similar procedure for the other sector and aggregating over the two sectors, Keynes gets the result that

$$P = w + \pi \tag{1}$$

$$\pi = (I - S) / Y. \quad (2)$$

P is the overall price level, *w* earnings per unit of output (unit labour cost) and π is profit per unit of output, that is, the share of profit. This identity becomes an equation only because via the expenditure equations, profits per unit of output are derived as the gap between investment expenditure (*I*) and savings (*s*) both per unit of output (*Y*) as in equation (2). In equilibrium, Keynes expects π to be zero and *w* to include 'normal' profits or remuneration of non-labour inputs as well as labour. When profits are non-zero this is because of investment exceeding savings and in a Wicksellian process this gap drives profits to drive the gap wider still. This process is not sufficiently articulated due to the fact that Keynes concentrates on conditions for price stability rather than on disequilibrium dynamics.

Post-Keynesian theories of profit and growth

Kalecki's theory marks a bridge between Marxian and Keynesian traditions and is the seminal contribution to what is now called the Post-Keynesian or Cambridge theory of income distribution. Paradoxically its points of contact with the *Treatise on Money* have not been sufficiently brought out. Kalecki's route was via Rosa Luxemburg's critique of Marx's Schemes of Expanded Reproduction (SER). The SER is also a two-sector model but of a growing economy. It has a two goods/two class configuration which is similar to the Fundamental Equations of the *Treatise*. While Marx's formulation of the SER make the model an equilibrium one, Luxemburg was seeking to find roots of dynamic disequilibrium within it. There are several strands which Kalecki weaves into this story.
Kalecki has a macroeconomic theory of pricing which yields a determinate share of profits in total output. He does this by exploiting the marginal revenue equals marginal cost conditions of equilibrium for the neoclassical firm. By then exploiting the simple idea that the ratio of price to marginal revenue departs from one to the extent that the price elasticity of demand is below infinity he connects price to marginal cost via the demand elasticity. Thus

$$p = mc(1 + \eta^{-1}) \quad (3)$$

where *mc* is the marginal cost and η is the elasticity of demand. The coefficient $(1 + \eta^{-1})$ is called the degree of monopoly. To the extent that η^{-1} departs from zero, the firm is a monopolistic one.

This is a partial equilibrium, microtheoretic derivation of the *p/mc* ratio and its generalization to a macro-economic level has proved to contain problems (Mitra, 1954). The main problem is that if (3) is supposed to refer to a specific firm, its elasticity of demand is not a constant but a function of the firm's own and its rivals' strategies. A determinate and tractable aggregation procedure for many jointly dependent *p*/*mc* ratios is not possible. It has however been found possible and empirically fruitful to interpret pricing decision as a mark-up above average cost.

$$p = (1 + k)ac \tag{4}$$

where *ac* is average cost and *k* is the mark-up ratio. The similarity of (4) to Keynes's Fundamental Equation in (1) is striking, that is, $\pi = k/(1 + k)$. But while (1) is an identity, (4) could be thought of as an equation where the profits come from producers' price setting behaviour.

But how are these profits sustained or in Marx's terminology realized? This is where the aggregate demand relations become important. It would be through the spending behaviour of the profit receivers that profits can be sustained. This was already clear in Keynes's invocation of the widow's cruse parable whereby a Wicksellian cumulative dynamic process can sustain growing profits as long as capitalists spend (that is, dis-save) while keeping up their investment expenditure. By starting with the Marxian SER, Kalecki was able to derive this as an *equilibrium* relation.

Kalecki's macroeconomic theory is best seen in terms of Kaldor's generalization. Kaldor takes the two class/two good model and integrates profits into a theory of growth and distribution. Let *R* be total profits $(\equiv \pi Y)$ and *W* be the total wage bill $(= wL)$. Then

$$Y = R + W \tag{5}$$

$$I = S = s_W W + s_C R. \tag{6}$$

Equation (5) is a national income identity, whereas (6) combines the saving–investment equality with a decomposition of total savings into workers' savings ($s_W W$) and capitalists' savings ($s_C R$) with the $s_C s_W$ being saving propensities and $s_C > s_W$. From (5) and (6), we can derive

$$R / Y = \pi = (s_c - s_w)^{-1}(I / Y) - s_w(s_c - s_w)^{-1}$$
(7a)

and

$$R / K = \rho = (s_c - s_w)^{-1}(I / K) - s_w(s_c - s_w)^{-1}(Y / K).$$
(7b)

Equation (7a) gives the share of profits in terms of the investment income ratio and (7b) gives the rate of profits in terms of the rate of growth of capital stock (*I*/*K*) and the output–capital ratio (*Y*/*K*). To specialize the equation, set s_w=0, that is, assume workers do not save. Then

$$\pi = s_c^{-1}(I / Y)$$
(8a)

$$\rho = s_c^{-1}(I / K).$$
(8b)

These two equations show how the profit share is determined by the capitalists' investment behaviour, i. e. capitalists determine their own profits. If we take the output–capital ratio to be a constant, then the rate of growth of income (*g*) is equal to the rate of growth of capital stock (*I*/*K*). Thus from (8b), we have

$$\rho = s_c^{-1} g.$$
(9)

If we now put s_c=1, we have the von Neumann result reproduced in the Kalecki–Kaldor models. The restriction that s_w=0 is of course arbitrary and thus makes the result under (8b) somewhat unrealistic. Pasinetti (1962) generalized the Kaldor argument by allowing workers as well as capitalists to save and own capital. Thus total capital K could be held either by capitalists K_c or by workers K_w but since capitalists make output and investment decisions workers were assumed to have loaned K_w to capitalists. In terms of the distinction we made above, capital as productive equipment is *controlled* by capitalists but capital as a financial asset is *owned* by both workers and capitalists, and capitalists pay workers a rate of interest i on the loaned capital. Thus instead of (7a) and (7b), we get

$$R / Y = \pi = (s_c - s_w)^{-1}\left[(I / Y - s_w) + r(s_w s_c (I / K)^{-1} - s_w (Y / K)^{-1})\right]$$
(9a)

$$R / K = \rho = (s_c - s_w)^{-1}\left[(I / K - s_w (Y / K)) + r(s_w s_c (I / K)^{-1} - s_w)\right].$$
(9b)

If we now put $r = \rho$, (9a) and (9b) degenerate to

$$\pi = s_c^{-1}(I / K)$$
(10a)

$$\rho = s_c^{-1}(I / K).$$
(10b)

The only condition needed for this result is that (I–$s_w Y$)=0. But while (10b) is similar to (8b), now it is independent of whether s_w is zero or not.

The similarity of the Kalecki–Kaldor result to the von Neumann result, as we noted above, is striking. The Pasinetti result seems to reinforce it. It is a one-good model and hence problems of relative price or aggregation or measurement of capital which plague other theories are completely avoided here. It is also not clear as to whether causality proceeds from growth to profits or profits to growth. There is an implicit assumption that the economy must have adequate resources and technology to generate surplus but the source of the surplus is not clear. There is no specification of the production conditions and a neoclassical aggregate production function is deliberately avoided.

The Pasinetti result has been derived by an alternative route by Samuelson and Modigliani (1966) who do use a neoclassical aggregate production function. Their purpose was to point out that the Pasinetti result was a special case of a more general result and that a dual to Pasinetti's theorem – an anti-Pasinetti theorem – could be derived from a slightly alternative formulation. All the assumptions of Pasinetti's theory are retained except that profits and wages are now derived from the marginal productivity conditions and a constant return to scale, two factor production function.

Let the production function be

$$\bar{Y} = f(\bar{K}) \quad f' > 0, \quad f'' > 0. \tag{11}$$

Here $\bar{Y} = Y/L$, $\bar{K} = K/L$, that is, output per worker and capital per worker. By the standard rules of marginal productivity theory we have that wage and rate of profit are determined as

$$\rho = f'(\bar{K}) \tag{12a}$$

$$w = f - \bar{K} f'(\bar{K}). \tag{12b}$$

In the production function, there is no distinction as to who owns the total capital stock – capitalists or workers. The savings augment the amount of capital owned by workers and capitalists,

$$S_c = \dot{K}_c = s_c f'(K) K_c$$
(13a)

$$S_c = \dot{K}_w = s_w [Y - f'(K) K_c].$$
(13b)

The equilibrium condition in the Samuelson–Modigliani model is that the relative rates of growth of capital stock owned by workers and capitalists be the same, i.e. constancy of shares in productive wealth. This is not an obvious condition for equilibrium but it does have the dramatic consequence that in such an equilibrium (of balanced growth of K_c and K_w), the rate of profit is independent of the saving propensity of the worker. If n is the constant growth rate of the capital stock, we get from the above after some manipulation

$$\dot{K}_w = s_w [f(K) - K_c f'(K)] - nK_w$$
(14a)

$$\dot{K}_c = [s_c f'(K) - n] K_c.$$
(14b)

In steady state $\dot{K}_c = \dot{K}_w = 0$, so (14b) gives

$$\rho = f'(K) = n / s_c.$$
(15)

Thus, the Pasinetti result can be derived from a neoclassical logic. This should not be too surprising

though much was made of this paradox at the time (Pasinetti, 1966; Kaldor, 1966). Neither a condition such as $\rho = r$ (what we have called the degeneracy result) nor that K_c/K_w is a constant tell us very much about the mechanisms by which an economy can arrive at such results. Our economic world is a world of many heterogeneous goods – capital as well as labour, of uncertainty, of financial constraints, of the persistent possibility of technical progress, of mergers and takeovers. There is in these models no decision making agency and time is eliminated in any meaningful sense since no *ex ante* versus *ex post* distinctions can be made. The Kalecki–Kaldor–Samuelson–Modigliani propositions are simple parables of pedagogic value no doubt but they do not tell us much about the origins or the role of profits in a modern economy.

Behavioural theories of profit

We can move in the final section of our essay to theories where the behavioural context is much more explicit. The neoclassical firm is a black box, where all the allocative rules could be followed by a computer which can be programmed to equate a derivative to a price. But the modern economy consists of corporations which operate in a world of financial markets and profits and are in this world both a signal of managerial performance and a facilitator for future expansion. It is this cluster of theories which supply the missing dimension in the theories hitherto surveyed.

Entrepreneurial theories of profit

In this cluster of theories, the level of aggregation is the firm and the agency of decision making is identified as the entrepreneur. It is assumed that the firm operates in a noncompetitive but stable environment where the entrepreneur has to form expectations about economic variables within his control as well as rivals' strategies. The other set of variables is the macroeconomic one where entrepreneurs may hold similar short run expectations. The level of profit sustained by the firm will depend in the short run on macroeconomic factors, common to all firms but in the long run on the decisions to invest so as to keep rivals at bay (Keirstead, 1953).

Knight's definition of uncertainty and Keynes's view of the difficulty of rational calculus in forming long-run expectation have been synthesized by Shackle in a series of works (Shackle, 1954; 1969; 1970) which purport to relate investment and *ex ante* profit in a decision theoretic framework. Shackle's decision theory is not however that of von Neumann–Morgenstern. He puts forward the notion of potential surprise and a surprise function relating the 'size' of potential surprise to the profit or loss attached to the project. Along with the surprise functions there is an ascendancy function which relates the entrepreneur's engagement with a project given the size of gain or loss anticipated.

Shackle then describes an optimizing exercise on the part of the entrepreneur, yielding optimal potential surprise. The size of gain or loss attached to the optimal potential surprise is then called primary focus gain or loss. The zero surprise (that is, certainty) equivalent of this primary gain is what Shackle calls profits. Profits are thus an *ex ante* certainty equivalent measure of the likely gain of the optimally chosen investment project. This definition unlike all previous ones makes the subjective and *ex ante* concept of profits clear. The problems with Shackle's theory are that it is determinedly resistant to aggregation even over individuals and while emphasizing the difficulty of using the calculus of subjective probability it

assumes that the surprise function and the ascendancy functions are continuous and differentiable enough to define a unique maximum. There is obviously a contradiction here.

Mention should be made in this respect of Lamberton's work (Lamberton, 1965). Lamberton prefers a satisficing to a maximizing approach and explicitly introduces the entrepreneur's income–leisure (inactivity) trade-off as determining the choice of investment projects and the associated profit outcome. In later work, Lamberton rationalizes suboptimizing behaviour in terms of the costs of information gathering and processing. Entrepreneurial activity to maintain or enhance profits is triggered off subject to threshold effects.

A corporate theory of profit

Authors such as Keirstead and Shackle view the firm as a single-owner entity. Modern firms are not by and large individually owned. In terms of share of output, employment or sales, it is the corporation with team management and hierarchical command structures which is the dominant mode of firm organization. It is not surprising therefore that one class of theories deals with profits in the context of corporate or managerial behaviour.

While in general Galbraith can be said to have made economists aware of the corporate form, corporate theories could be said to have benefited from the contribution of the behavioural theory of firm of Simon, Cyert and March, the hypothesis attributed to Baumol that firms maximized sales rather than profits and the notion due to Marris among others that growth of corporate size was the aim of managers of corporations (Baumol, 1967; Cyert and March, 1963; Galbraith, 1967; Marris, 1964; Simon, 1957). These various developments in industrial economics brought forth a view that managerial behaviour was growth oriented, that operating in monopolistic competitive environments managers could choose profits and growth combination subject to the constraint of external sources of finance.

Wood (1975) presents the most complete theory in this respect. There are no details on the production side in his theory of profits. Profits arise from the need of the corporation to grow. The choice variables are the retention ratio, that is, the proportion of investment financed from external sources and the amount of liquid financial assets desired by the firm to be held as some relation of gross investment expenditure.

The firm is assumed to be facing a convex opportunity frontier between the profit share (profit margin as Wood calls it) and the growth rate of its sales revenue. Its desired (or target) values of the retention ratio (γ) the external finance ratio (λ) and the liquid financial holdings ratio (δ) give a simple relation between profit share (π) and the growth rate g. We have

$$I + \delta l \leq \gamma R + \lambda l. \tag{16}$$

In equation (16), the left-hand side represents the uses of funds and the right-hand side the sources of funds. A slight rearrangement gives us

$$R / Y = \pi \geq \gamma^{-1}(1 + \delta - \lambda)\sigma g \quad (17)$$

where $g = \Delta Y / Y$, the growth rate of sales and $\sigma = I / \Delta Y$ is investment to increase in output (or the incremental capital output) ratio. Now if one could accept that the coefficients δ, λ and γ are fixed parameters and σ is also constant, then π and *g* are linearly related. The firm will start with a desired value of *g* and seek the appropriate $\prod$ and by an iterative process arrive at a $\prod$, *g* combination which satisfies (17) as an equality rather than an inequity. This $\prod$, g combination will also be on the boundary of the opportunity frontier.

Wood's theory treats the corporation in isolation and with some control over its pricing and revenue situation. The presence of competing monopolistic firms is ignored here as in Kalecki's theory. But his theory does bring out the interrelationship between profits and the financing of investment. Note also that in (16) all the variable are in nominal terms. If the parameters δ, λ and γ were identical across firms, we would aggregate (16) across firms. Since *Y* is sales revenue in nominal terms equation (17) would make sense at an aggregate level though it would be harder to swallow that σ would be a constant.

Conclusion

A satisfactory theory of profits is still elusive. For neoclassical economics, profit is a non problem and the only problem is to assign any observed net income above costs to the category of interest, quasi-rent or managerial wages for risk bearing. But problems persist for other theories as well. The classical theory neglects uncertainty and is vague about the microfoundations of profits. If von Neumann and Sraffa detail the technical conditions of production, neither uncertainty nor demand considerations figure prominently in their theories. Most theories, with the exception of the corporate theories and Schumpeter’s, neglect the financial aspects of business operations. The firm is a black box in neoclassical theory whereas for Knight, Keynes and Shackle the individual decision maker's expectations are crucial to business behaviour. To allow for persistent positive profits in a dynamic equilibrium microeconomic model with subjective uncertainty and expectations with the possibility of technical substitution and technical change in production remains a challenge. If such a theory could be cast in terms to allow aggregation over firms to the level of the economy, the puzzle of profits would be solved.

See Also

- worker participation and profit sharing

Bibliography

Baumol, W.J. 1967. *Business Behaviour, Value and Growth*, 2nd edn. New York: Macmillan.

Carter, C.F. and Ford, J.L., eds. 1972. *Expectations and Uncertainty in Economics*. Oxford: Blackwell.

Cobb, C. and Douglas, P. 1928. A theory of production. *American Economic Review* 18, 139–65.

Cyert, R. and March, J.G. 1963. *A Behavioural Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Douglas, P. 1948. Are there laws of production? *American Economic Review* 38, 1–41.

Galbraith, J.K. 1967. *The New Industrial State*. Boston: Houghton Mifflin.

Kaldor, N. 1955. Alternative theories of distribution. *Review of Economic Studies* 23(2), 83–100.

Kaldor, N. 1966. Marginal productivity and the macroeconomic theories of distribution. *Review of Economic Studies* 33, 309–20.

Kalecki, M. 1939. *Essays in the Theory of Economic Fluctuations*. London: Allen & Unwin.

Keirstead, B.S. 1953. *An Essay in the Theory of Profits and Income Distribution*. Oxford: Blackwell.

Keynes, J.M. 1930. *A Treatise on Money*, vol. 1. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Lamberton, D.M. 1965. *The Theory of Profit*. Oxford: Blackwell.

Lamberton, D.M. 1972. Information and profit. In Carter and Ford (1972).

Marris, R. 1964. *The Economic Theory of Managerial Capitalism*. London: Macmillan.

Meek, R.L. 1954. Adam Smith and the classical concept of profit. *Scottish Journal of Political Economy* 1(2), 138–53.

Mitra, A. 1954. *The Share of Wages in National Income*. Rotterdam: Netherlands Central Planning Bureau, limited circulation edition; Calcutta: Oxford University Press, 1980.

Morishima, M. 1973. *Marx's Economics: A Dual Theory of Value and Growth*. Cambridge: Cambridge University Press.

Morishima, M. 1974. Marx in the light of modern economic theory. *Econometrica* 42, 611–32.

Okishio, N. 1963. A mathematical note on Marxian theory. *Weltwirthschaftliches Archiv* 91(2), 287–99.

Pasinetti, L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29, 267–79.

Pasinetti, L. 1966. New results in an old framework. *Review of Economic Studies* 33, 303–6.

Samuelson, P.A. and Modigliani, F. 1966. The Pasinetti paradox in neoclassical and more general models. *Review of Economic Studies* 33, 269–302.

Schumpeter, J.A. 1912. *Theorie der Wirtschaftlichen Entwicklung, Eine Unterschung uber Unternehmergewinn, Kapital, Kredit, Zins und den Konjukturzyklus*. Munich and Leipzig: Duncker & Humblot. Trans. as *Theory of Economic Development*. Cambridge, MA: Harvard University Press, 1934.

Shackle, G.L.S. 1954. Professor Keirstead's theory of profits. *Economic Journal* 64(March), 116–23.

Shackle, G.L.C. 1969. *Decision, Order and Time in Human Affairs*, 2nd edn. Cambridge: Cambridge University Press.

Shackle, G.L.C. 1970. *Expectation, Enterprise and Profit*. London: Allen & Unwin.

Simon, H.A. 1957. *Models of Man: Social and Rational*. New York: Wiley.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Solow, R.M. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–30.

Sraffa, P. 1960. *Production of Commodities by Mean of Commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1976. *Marx After Sraffa*. London: New Left Books.

von Neumann, J. 1945–6. A model of general equilibrium. *Review of Economic Studies* 13, 1–9.

Wood, A.J.B. 1975. *A Theory of Profits*. Cambridge: Cambridge University Press.

How to cite this article

Desai, Meghnad. "profit and profit theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000213> doi:10.1057/9780230226203.1349

The New Palgrave Dictionary of Economics Online

progressive and regressive taxation

Original 1987 article by William Vickrey, revised by Efe A. Ok
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Progressive (resp. regressive) taxation refers to a taxation scheme in which the amount of tax paid as a proportion of the tax base rises (resp. declines) with that base. While progressive taxation has been justified in terms of the ‘principle of equal sacrifice’ and mitigating the inequality of market outcomes, no general political theory of income taxation provides theoretical support for the observed prevalence of progressive taxation schemes. Nor has the theory of optimal income (and consumption) taxation shed any light on the nature of progressive taxation, either normatively or positively.

Keywords

end-point theorem; equal sacrifice principle; equity–efficiency trade-off; income mobility; Jakobsson–Fellman Theorem; Lorenz curve; mixed strategy equilibrium; optimal taxation; progressive and regressive taxation; redistribution of income and wealth; Reynolds–Smolensky progressivity index; tax base; tax burden; tax incidence

Article

Progressive (resp. regressive) taxation refers to a taxation scheme (applied to a monetary base such as income, consumption, wealth and so on) in which the amount of tax paid as a proportion of the tax base rises (resp. declines) with that base. As such, characterizing taxes as progressive or regressive according to the relative degree to which they impose burdens on the wealthy and on the poor seems at first blush a fairly straightforward matter. Unfortunately, when one wishes to understand the nature of *effective* progressivity of a taxation scheme, and its redistributive properties thereof, and especially to compare the relative degree of progressivity or regressivity of alternative taxes, a number of problems arise. Before formalizing the notion of progressivity, we thus have a quick look at these issues.

Tax base

An important and often overlooked question is that of the base to which the tax burden is to be related. Usually the base is taken to be income, which in practice, whenever an attempt is made to produce actual figures, means some version of annual monetary income (which perforce includes salaries, interest and dividends, but which may or may not include capital gains). An alternative basis for evaluating progression would be consumption. This is not often used, and is subject to the same deficiencies as income as a result of omissions of non-monetary items such as imputed incomes from consumer durables (like owned homes).

Tax burden

There are various difficulties in determining the actual tax burden levied on a taxpayer. First, application of tax rates to incomes that fluctuate (due to, say, the stage of one's life cycle) makes this difficult to measure. As advocated by Vickrey (1947), it appears that some form of income averaging is needed (to serve as a proxy for the expected permanent incomes) but this has been found to be too complicated to administer in practice. Second, it is not obvious how to incorporate family size in the computation of tax burdens. Simply taking the aggregate would misclassify the larger and smaller units relative to their level of welfare or ability to pay, while taking a per capita measure overstates the importance of children relative to their needs. (See Pechman, 1987, pp. 78–133, for a careful discussion of these issues and other structural problems with tax assessment in general.) Third, the market mechanism often allows the tax burden to be passed from the taxpayer to other units in the economy, with the eventual consequence that the burden of taxes is not necessarily borne by those upon whom they are levied. In particular, the imposition of taxes on income or sales changes the budget sets of individuals, thereby altering equilibrium prices in the economy. The issue becomes particularly pressing in the case of corporate income tax, as tax incidence in this case varies widely according to fiscal, monetary or activity level changes that are associated with an alteration in the tax. For instance, a tax imposed on firms for hiring labour is more than likely to be ‘shifted’ to workers through lower wages and to consumers through higher prices, upon the adjustment of employment decisions on the part of the firms. In general, then, the ultimate distribution of the tax burden – the so-called *economic incidence* – is different from *statutory incidence*, that is, the initial distribution of tax liabilities. Unfortunately, it is a highly non-trivial matter, both empirically and theoretically, to determine precisely who bears the tax burden, and ultimately to what extent. It is thus not uncommon in practice to encounter discussions on tax progressivity that ignore tax incidence problems and concentrate instead simply on statutory incidence. (See Musgrave and Musgrave, 1989, chs 12 and 13, for an introductory account of tax incidence analysis, and Kotlikoff and Summers, 1989, and Atkinson, 1994, for advanced treatments.)

It is worth noting that the issues concerning the base and burden of taxes are integral to any sort of fiscal analysis, and are not particular to the analysis of tax progressivity. To flesh out the elements of the latter, therefore, we shall abstract from these difficulties in what follows and assume that a notion of monetary outcome, which we shall simply refer to as *income*, is determined as the arbiter of ability to pay. Moreover, for the most part we shall work under the (uncomfortable) supposition that the amount of tax charged on a given level of income corresponds to the actual tax burden of the taxpayer with that amount of income. At the very least, this will allow us to focus properly on certain facets of the *theory* of progressive taxation.

Progressive (regressive) tax functions

Formally speaking, a *tax function* is a right-differentiable and strictly increasing map $T:\mathbb{R}_+ \to \mathbb{R}_+$ such that $T(0)=0$ and $T(x)<x$ for all $x \in \mathbb{R}_{++}$, and $0 < T'(x) < 1$ for all $x \in \mathbb{R}_+$. (Here we denote the right-derivative of T by T'.) This formulation maintains that (i) zero income earners do not pay any taxes; (ii) if a person earns positive income, the amount of taxes imposed on her must be less than her taxable income base; (iii) higher-income earners pay a higher level of tax than lower-income earners; and (iv) taxation is non-confiscatory in that the ranking of taxpayers by pre-tax income and post-tax income is the same. (We rule out here negative taxation to simplify our exposition, and view T as modelling a *statutory* tax scheme.) A tax function T is said to be *progressive* if the map $x \mapsto \frac{T(x)}{x}$ is increasing on $\mathbb{R}_{++}$, that is, if the amount of income tax paid as a proportion of the tax base (say, income) rises with that base. In turn, T is *regressive* if $x \mapsto \frac{T(x)}{x}$ is decreasing on $\mathbb{R}_{++}$. Finally, T is *marginal-rate progressive* (*regressive*) if the tax rate T' is itself an increasing (decreasing) function. In practice, statutory taxes on income and spendings are always progressive – in fact, they are almost always marginal-rate progressive – while payroll and sales taxes possess a flat statutory tax rate (but economic incidence analyses frequently reveal that such taxes are, effectively, regressive).

Normative basis for progressivity

The most well-known equity principle – advanced originally by John Stuart Mill – that provides a normative basis for progressive taxation is the *principle of equal sacrifice*. The modern formulation of this principle demands that there be a social norm, represented by a continuous, concave and strictly increasing (social) utility function $U:\mathbb{R}_+ \to \mathbb{R}$ relative to which the income tax T imposes equal sacrifice upon all taxpayers, that is,

$$U(x) - U(x - T(x)) = \text{constant for all } x > 0$$

(see Young, 1987; 1990; Ok, 1995). We may now ask: does a progressive, or a marginal-rate progressive, tax necessarily satisfy the principle of equal sacrifice? Conversely, does this principle necessitate progressivity?

The answers are, unfortunately, not very clear. First, the good news: it can be shown that a marginal-rate progressive tax function surely satisfies the principle of equal sacrifice. The bad news is that mere progressivity of a tax function is not enough for it to satisfy this principle (Mitra and Ok, 1997). To make matters worse, even for some non-progressive taxes T we can find a (social) utility function U that satisfy the properties above, that is, the principle of equal sacrifice need not imply, or be implied by, tax progressivity. At the very least, one needs to assume more about T and U to be able to relate these

principles more closely. For instance, if we demand that *U* be differentiable (at least near the origin), then a piecewise linear tax function T satisfies the principle of equal sacrifice (as we formulated above) *if, and only if, T* is marginal-rate progressive (Mitra and Ok, 1996). If one is prepared to accept this set-up, therefore, the principle of equal sacrifice can be thought of as characterizing marginal-rate progressivity of a (statutory) tax function, thereby necessitating its progressivity.

An additional caveat here is, of course, that this account ignores the disincentive effects of taxation. It is partial relief that Berliant and Gouveia (1993) have shown that, when the individual utility functions over income and leisure are additively separable, the link between the principle of equal sacrifice and progressivity would prevail even in the presence of such effects. Unfortunately, little is known about this matter in the (more realistic) non-separable case.

Redistributive consequences of progressivity

One of the traditional arguments for progressive taxation is that such schemes redress the highly inegalitarian outcomes of the market system, thereby acting as social insurance against inequality. As colourful as it may be, this argument needs to be formalized properly.

Let us first agree to model an income distribution as a continuous and increasing distribution function $F: \mathbb{R} \to [0, 1]$ with $F(0)=0$ and $F(1)=1$. This sort of a specification is, for instance, frequently adopted in macroeconomic models of income distribution. For any such *F*, we let $\mu_F := \int_0^1 x dF(x)$, which is the total income in the society. (Since incomes are distributed on [0,1], that is, we effectively concentrate on relative incomes, μ_F also corresponds to the per-capita income in this model.) In what follows we naturally assume that $\mu_F>0$.

For any such income distribution *F*, the *pseudo-inverse* of *F* is defined as the function $F^{-1}: (0, 1) \to \mathbb{R}_+$ with

$$F^{-1}(t) := \inf\{x \geq 0 : F(x) \geq t\}, \; 0 < t < 1.$$

Intuitively, we may think of $F^{-1}(t)$ as the income level of the person who belongs to the poorest 100*t* per cent in the income distribution. We next define the map $L_F: [0, 1] \to \mathbb{R}$ by

$$L_F(p) := \frac{1}{\mu_F}\int_0^p F^{-1}(t)\, dt, \; 0 \leq p \leq 1,$$

which corresponds to the cumulative share of income held by the poorest 100 per cent of the population. The graph of the map $p \mapsto L_F(p)$ is called the *Lorenz curve* of the distribution *F*.

We say that the income distribution *F Lorenz dominates* another income distribution *G* whenever $L_F(p) \geq L_G(p)$ holds for all $p \in [0,1]$, with strict inequality for at least one *p*. It is well known that this happens if, and only if, *G* can be obtained from *F* by means of finitely many mean-preserving spreads (Rothschild and Stiglitz, 1970). This is one of the reasons why Lorenz dominance is generally accepted as an *unambiguous* method of making ordinal inequality comparisons. Its welfare basis is identified in the seminal works of Kolm (1969) and Atkinson (1970).

Now take any income distribution *F*. A tax function *T* applied to this distribution induces the *post-tax income distribution* F^T where $F^T(x):=F(x-T(x))$ for any $x \in \mathbb{R}$. A celebrated theorem of public economics – often called the *Jakobsson–Fellman theorem* – maintains that F^T Lorenz dominates *F*, that is, tax is inequality-reducing, if, and only if, *T* is progressive (see Fellman, 1976; Jakobsson, 1976). That is, progressive taxes, *and only progressive taxes*, possess the property of reducing the level of income inequality no matter to which pre-tax income distribution they are applied. (For variations on this theme, see Eichhorn, Funke and Richter, 1984, and Thon, 1987.) This shows, in a nutshell, why the progressivity of a taxation scheme may be justified on the basis of desire for inequality reduction.

The Jakobsson–Fellman theorem also leads to a natural method of quantifying the redistributive effect of a tax function *T* that is applied to an income distribution *F*. To see this, let us consider the function $R_{F,T}: [0, 1] \to \mathbb{R}$ defined by

$$R_{F,T}(p) := L_{F^T}(p) - L_F(p).$$

In words, $R_{F,T}(p)$ measures the income share of the poorest 100*p* per cent in excess of what they would obtain under an equal yield flat tax. Obviously, the Jakobsson–Fellman Theorem says that $R_{F,T} \geq 0$ for *any* income distribution *F* if, and only if, *T* is progressive. (But, of course, we may have $R_{F,T} \geq 0$ for *some F* even if *T* is not progressive.) Now the discussion above suggests to declare a tax function T^1 to be *more redistributive than* T^2 – due to the Jakobsson–Fellman theorem, one often says T^1 *is more progressive than* T^2 – relative to the income distribution *F* if $R_{F,T^1} \geq R_{F,T^2}$. This is, of course, a partial ordering, and when it does not apply one may wish to resort to a compatible index that sizes up the redistributive effects of T^1 and T^2. The most widely used index to this effect is the *Reynolds–Smolensky progressivity index* defined (as a function of a tax function *T*) by

$$I_F^{RS}(T) := 2\int_0^1 R_{F,T}(p)\,\mathrm{d}p,$$

which is the difference between the Gini coefficients of the distributions *F* and F^T. Other indices are proposed in the literature to compare the progressivity of tax functions (which are based, for instance, on

the departure of a tax function from the equal-yield proportional tax). For an extensive discussion of such indices, and further results on the redistributional effects of progressive taxes, we refer the reader to the excellent survey by Lambert (1999).

Political economy of progressive taxation

Now that we have examined a number of normative rationales for progressivity of income taxes, let us turn to the strand of literature that has attempted to explain the prevalence of such tax schemes from the viewpoint of behavioural political economy. This literature maintains that, given that her views about income tax policy is one of the most important traits of a political candidate, it is natural to expect this prevalence to reflect (however indirectly) the majority support in the population. In fact, this way of thinking seems to suggest a straightforward explanation of the empirically observed popularity of marginal-rate progressivity, provided that one subscribes to the ‘one-man one-vote’ rule. Since the income distribution of a country is always globally right-skewed (in the sense that the median income is strictly smaller than the mean income for any right truncation of the income distribution), the number of poorer voters always exceeds the number of richer voters, regardless of how one defines the cut-off that separates the poor from the rich. Since poorer voters are typically the supporters of progressive policies, so the argument goes, there would then be a natural tendency for the marginal-rate progressive tax policies to be favoured by the majority. Even though the actual political processes are far more complex than the scenario in which people vote directly over policies, this argument appears to suggest a convincing reason for why progressive tax policies are so widely adopted.

While there are a few direct democracy models in the literature that provide support for this argument (cf. Romer, 1975; Roberts, 1977; Cukierman and Meltzer, 1991; Gouveia and Oliver, 1996; Marhuenda and Ortuño-Ortín, 1998; Roemer, 1999), these models are either confined to rather specific settings (in which a tax function is characterized by means of at most two parameters) or are not couched in a political equilibrium framework. A natural direct democracy model of voting over income taxes would be a two-party voting game in which each party (whose objective is to win the elections) proposes a tax function from an exogenously given set of admissible tax functions (that raise a given amount of revenue), and voters vote selfishly for the tax function that taxes them less. To make transparent the difficulties that pertain to the political economic approach to progressive taxation, we now describe such a voting model in precise terms.

Let F be a strictly increasing income distribution (as modelled above), and assume that the median income is strictly less than the average income according to F, that is, $m_F := F^{-1}\left(\frac{1}{2}\right) < \mu_F$. (This assumption is but a straightforward formalization of the heuristic statement that ‘the number of poor people in the society is strictly less than that of rich people’.) To focus on the issue of *redistribution*, it is assumed that tax policies are designed to collect an exogenously given amount of revenue $0 < \alpha < \mu_F$, or put differently, an *admissible tax function T* is defined in the model as one with the property $\int_0^1 T\,dF = \alpha$. We denote the class of all such tax functions by $\mathcal{T}_{F,\alpha}$.

Consider a two-party voting model in which each party advocates an income tax policy in $\mathcal{T}_{F,\alpha}$ which is to be put in effect in case this party obtains the support of the majority. Citizens evaluate proposals from a selfish point of view. Put precisely, an individual with income x regards the tax function T^1 as more

desirable than the tax function T^2 if $T^1(x)<T^2(x)$, that is, if this individual's tax liability is lower under T^1. It is assumed that indifferent voters abstain from voting. Thus, if party 1 proposes tax policy t_1 and party 2 proposes tax policy t_2, the share of votes obtained by the first party is determined as

$$\omega(T^1, T^2) := p_F\{x \in [0, 1] : T^1(x) < T^2(x)\},$$

where p_F is the probability measure induced by F on [0,1]. Of course, in this case the share obtained by party 2 is $\omega(T^2, T^1)$. The formal model takes the form of a two-person strategic game in which the two players are the parties both of whose action spaces equal $\mathcal{T}_{F,\alpha}$. There is a multitude of ways of modelling the objectives of the parties here. We follow Carbonell-Nicolau and Ok (2007), and presume that the goal of party i is the maximization of the net plurality defined as the difference between the vote shares obtained by the candidates. (For instance, the payoff function of party 1 is the map $(T^1, T^2) \mapsto \omega(T^1, T^2) - \omega(T^2, T^1)$ on $\mathcal{T}_{F,\alpha} \times \mathcal{T}_{F,\alpha}$).

While this model is one of the simplest of its kind, it readily exhibits the familiar difficulty of (infinite-dimensional) voting games: it does *not* have a Nash equilibrium (for any given F and α). Intuitively speaking, this is because, given any admissible tax T in $\mathcal{T}_{F,\alpha}$, one can always find another tax function which is below T over an interval of p_F measure greater than one half. Not all hope is lost here, however, as it can be shown that there is at least one mixed strategy equilibrium of this game. The question then becomes whether or not the support of any such equilibrium consists only of progressive taxes. Curiously, Carbonell-Nicolau and Ok (2007) show that, if F and α satisfy a certain condition, then, generically, in at least one equilibrium the probability of parties proposing a non-progressive tax function is positive. All in all, after a lot of work, and absent a good economic reason for working with a particular way of restricting the set of admissible tax functions, one is left with the feeling that the prevalence of tax progressivity that we find in all industrial democracies cannot be attributed solely to the right-skewedness of income distributions.

In passing, we should note that, in the context of *representative* democracies (à la Osborne and Slivinski, 1996; Besley and Coate, 1997), one is sometimes able to escape from the equilibrium existence problem discussed above. Indeed some positive results on the majority support for progressive taxation are obtained in this setting (cf. Carbonell-Nicolau and Klor, 2003). Unfortunately, it is not known whether or not these results would survive the inclusion of disincentive effects of taxation into the model. Furthermore, if we add to the picture dynamic considerations of voters, things would get even more complicated. Indeed, we know from Benabou and Ok (2001) that if the income mobility process has a particular property (called *concavity in expectation*) that is often met in reality, then individuals who are currently poor may oppose redistribution because they (rationally) expect to move up in the income ladder in the future. Indeed, it is possible that this *prospect of upward mobility* (*POUM*) *hypothesis* may be strong enough to overturn the majority support of progressivity (even at the steady state of the underlying mobility process).

To wit, it seems at present that we do not have a general political theory of income taxation that provides theoretical support for the observed prevalence of progressive taxation schemes. Nevertheless, this is an area of active research, and it is not unreasonable to expect that lasting contributions on this theme will be made in the near future.

Optimal tax structures

Any discussion of progressive taxation would be incomplete without putting on record the large amount of work conducted towards the end of 20th century on the optimal design of income (and commodity) tax schemes. Indeed, after the seminal contribution of Mirrlees (1971), there has been an immense amount of work in this area, which has, however, slowed down significantly in more recent times. Roughly speaking, the canonical model of optimal income taxation works with a population with a given distribution of income earning ability and with a utility function having disposable income and effort as arguments, pre-tax income being a function of ability and effort. Individuals then act to maximize their individual utility subject to an income tax schedule which is to be determined to raise a given revenue while resulting in a maximum of utility for the population which is obtained as an aggregation of the individual utilities. Unfortunately, despite the promise of the early work on this topic – see, for instance, Sadka (1976), Stern (1976) and Seade (1977) – this model is found not to produce robust qualitative results. (See Stiglitz, 1982, for a careful discussion of this issue.) One exception to this is the infamous *end-point theorem* that states that, when the distribution of skills has a known upper limit, the marginal tax rate should vanish at the level of income of the highest-income earners. (Informally speaking, the argument here is that there is no point to deterring the highest-income earner from earning the last dollar of her income, since if she does not earn it there will be no revenue from it.) This has the unsettling implication that an ‘optimal’ income tax is perforce non-progressive.

There is reason not to take this conclusion too seriously, however. First, simulations show that the end-point theorem is very local (cf. Tuomala, 1990; Saez, 2001). Second, the assumption that the constituent individuals are identical in all aspects but ability is a rather stringent requirement (which is key for the validity of the end-point theorem). Third, if there is uncertainty in the model that results in the expected income distribution having unbounded support, then the result fails. (See Haveman, 1994, for a variety of other critical comments on optimal income taxation theory.) All in all, while it has duly brought the incentive problems into the forefront of public finance, and has provided a new means of evaluating the notion of equity–efficiency trade-off, it appears that the theory of optimal income (and consumption) taxation has not realized its promise in shedding light on the nature of progressive taxation, either normatively or positively.

Perhaps it is best to conclude our discussion, as Haveman (1994) does as well, with the words of Joseph Pechman, taken from his (posthumous) 1990 presidential address to the American Economic Association: ‘Most people support tax progressivity on the ground that taxes should be levied in accordance with ability to pay, which is assumed to rise more than proportionately with income. Economists have … had trouble with the “ability to pay” concept … I believe that the person on the street is right and that we should continue to rely on the income tax to raise revenue in an equitable manner.’

See Also

- optimal taxation
- tax incidence

Bibliography

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 244–63.

Atkinson, A. 1994. The distribution of tax burden. In *Modern Public Finance*, ed. J. Quigley and E. Smolensky. Cambridge: Harvard University Press.

Benabou, R. and Ok, E.A. 2001. Social mobility and the demand for redistribution: the POUM hypothesis. *Quarterly Journal of Economics* 116, 447–87.

Berliant, M. and Gouveia, M. 1993. Equal sacrifice and incentive compatible income taxation. *Journal of Public Economics* 51, 219–40.

Besley, T. and Coate, S. 1997. An economic model of representative democracy. *Quarterly Journal of Economics* 112, 85–114.

Carbonell-Nicolau, O. and Klor, E. 2003. Representative democracy and marginal rate progressive income taxation. *Journal of Public Economics* 87, 2339–66.

Carbonell-Nicolau, O. and Ok, E.A. 2007. Voting over income taxation. *Journal of Economic Theory* 134, 249–86.

Cukierman, A. and Meltzer, A. 1991. A political theory of progressive taxation. In *Political Economy*, ed. A. Meltzer, A. Cukierman and S. Richard. Oxford: Oxford University Press.

Eichhorn, W., Funke, H. and Richter, W. 1984. Tax progression and inequality of income distribution. *Journal of Mathematical Economics* 13, 127–31.

Fellman, J. 1976. The effects of transformations on Lorenz curves. *Econometrica* 44, 823–4.

Gouveia, M. and Oliver, D. 1996. Voting over flat taxes in an endowment economy. *Economics Letters* 50, 251–8.

Haveman, R. 1994. Optimal taxation and public policy. In *Modern Public Finance*, ed. J. Quigley and E. Smolensky. Cambridge: Harvard University Press.

Jakobsson, U. 1976. On the measurement of degree of progression. *Journal of Public Economics* 5, 161–68.

Kolm, S-C. 1969. The optimal production of social justice. In *Public Economics*, ed. J. Margolis and H. Guitton. London: Macmillan.

Kotlikoff, L. and Summers, L. 1989. Tax incidence. In *Handbook of Public Economics*, vol. 2, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Lambert, P. 1999. Redistributional effects of progressive income taxes. In *Handbook of Income Inequality Measurement*, ed. J. Silber. Boston: Kluwer.

Marhuenda, F. and Ortuño-Ortín, I. 1998. Income taxation, uncertainty and stability. *Journal of Public Economics* 67, 285–300.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Mitra, T. and Ok, E.A. 1996. Personal income taxation and the principle of equal sacrifice revisited. *International Economic Review* 37, 925–48.

Mitra, T. and Ok, E.A. 1997. On the equitability of progressive income taxation. *Journal of Economic Theory* 73, 316–34.

Musgrave, R. and Musgrave, P. 1989. *Public Finance in Theory and Practice*, 5th edn. New York: McGraw Hill.

Ok, E.A. 1995. On the principle of equal sacrifice in income taxation. *Journal of Public Economics* 58, 453–68.

Osborne, M. and Slivinski, A. 1996. A model of political competition with citizen candidates. *Quarterly Journal of Economics* 111, 65–96.

Pechman, J. 1987. *Federal Tax Policy*, 5th edn. Washington, DC: Brookings Institution.

Pechman, J. 1990. The future of income tax. *American Economic Review* 80, 1–20.

Roberts, K. 1977. Voting over income tax schedules. *Journal of Public Economics* 8, 329–40.

Roemer, J. 1999. The democratic political economy of progressive income taxation. *Econometrica* 67, 1–19.

Romer, T. 1975. Individual welfare, majority voting, and the properties of a linear income tax. *Journal of Public Economics* 4, 163–85.

Rothschild, M. and Stiglitz, J. 1970. Increasing risk I: a definition. *Journal of Economic Theory* 2, 225–43.

Sadka, E. 1976. On income distribution, incentive effects and optimal income taxation. *Review of Economic Studies* 42, 261–8.

Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68, 205–29.

Seade, J. 1977. On the shape of optimal income tax schedules. *Journal of Public Economics* 7, 203–36.

Stern, N. 1976. On the specification of models of optimum income taxation. *Journal of Public Economics* 6, 123–62.

Stiglitz, J. 1982. Self-selection and Pareto efficient taxation. *Journal of Public Economics* 17, 213–40.

Thon, D. 1987. Redistributive properties of progressive taxation. *Mathematical Social Sciences* 14, 185–91.

Tuomala, M. 1990. *Optimal Income Tax and Redistribution*. Oxford: Clarendon Press.

Vickrey, W. 1947. *Agenda for Progressive Taxation*. New York: Ronald Press.

Young, P. 1987. Progressive taxation and equal sacrifice principle. *Journal of Public Economics* 32, 203–14.

Young, P. 1990. Equal sacrifice and progressive taxation. *American Economic Review* 80, 253–66.

How to cite this article

Vickrey, William and Efe A. Ok. "progressive and regressive taxation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000218> doi:10.1057/9780230226203.1350

The New Palgrave Dictionary of Economics Online

propensity score

Jinyong Hahn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

asymptotic theory; conditional probability; covariate matching; dimensionality; evaluation research; finite sample theory in econometrics; matching; nonparametric models and methods; propensity score

Article

Propensity score is an object often discussed in evaluation studies. It is defined as the conditional probability of treatment given covariates. It has attracted attention for its potential to control for the bias in the presence of high dimensional covariates.

Evaluation research typically begins by comparing the treated group with the control group. For example, estimates of the effect of training programmes on earnings compare the earnings of those who receive training with a candidate control sample of untrained people. Because typically trainees are not chosen randomly, a simple comparison of the two groups may not provide a very accurate picture of what would have happened to the trainees had they not been trained. Under some conditions, such problems can be avoided by comparing the treated and the control groups with identical covariate values. For more formal discussion, denote the covariate vector for person *i* by X_i, treatment status by D_i such that $D_i = 1$ if the *i*th person is treated and $D_i = 0$ otherwise, and define the conditional probability of treatment, or propensity score, as $p(X_i) \equiv \Pr[D_i = 1|X_i]$. Let Y_{1i} denote the potential, or counter-factual, outcome if the *i*th person receives the treatment, and let Y_{oi} denote potential earnings if he or she does not receive the treatment. Note that Y_{1i} is observed only when $D_i = 1$. Likewise, Y_{oi} is observed only when $D_i = 0$. This implies that $Y_{1i} - Y_{0i}$ is not observed by the researcher, and therefore the average treatment effects, which are defined to be $\beta = E[Y_{1i} - Y_{0i}]$, cannot be estimated by the sample analog of $E[Y_{1i} - Y_{0i}]$. Because D_i is not usually assigned randomly, a simple comparison of the two groups, that is, the sample analog of $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$, does not provide a consistent estimate of β, either. On the other hand, if (Y_{0i}, Y_{1i}) is independent of D_i given X_i, that is,

$$Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | X_i \tag{1}$$

then the sample analog of

$$E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0]\}$$

will provide a consistent estimator for β. In other words, β can be consistently estimated by 'matching' the treated and the control groups with identical covariate values.

A problem that often arises in studies of this type is the need to control for continuously distributed and/or high-dimensional covariates. In many evaluation studies, the sample sizes are small, there are many covariates, and some of the covariates are continuous. A number of variations on exact covariate-matching schemes have been developed to deal with such situations. These typically involve approximate matching, or nonparametric smoothing, of some sort.

An alternative strategy to control for covariates begins with Rosenbaum and Rubin's (1983) observation that bias can be eliminated by controlling for a scalar-valued function of the covariates, namely, the propensity score. Rosenbaum and Rubin's propensity-score theorem states that, if (1) is true, then it must also be true that conditioning on $p(X_i)$ eliminates selection bias, that is,

$$Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | p(X_i). \tag{2}$$

This implies that the β can be consistently estimated by the sample analog of

$$E\{E[Y_{1i}|p(X_i), D_i = 1] - E[Y_{0i}|p(X_i), D_i = 0]\}.$$

It is easier to estimate $E[Y_{1i}|p(X_i), D_i = 1]$ than $E[Y_{1i}|X_i, D_i = 1]$, because the former requires the nonparametric regression of Y_{1i} on a scalar object $p(X_i)$ whereas the latter requires the nonparametric regression on a multi-dimensional object X_i. (Such difficulty is often called the curse of dimensionality in the nonparametrics literature.) The value of propensity score matching is in the 'dimension reduction' generated by regions where $p(X_i)$ is constant while $E[Y_{1i}|X_i]$ or $E[Y_{0i}|X_i]$ are not constant. The value of the propensity score is not clear, though, when the applied researcher does not have any information

about the treatment assignment. Without such information, the propensity score needs to be estimated, which requires a nonparametric regression of D_i on X_i. Because this estimation suffers from the curse of dimensionality, the propensity score theorem seems to have little practical value when the propensity score itself needs to be estimated nonparametrically. On the other hand, it can be quite useful when applied researchers may have more information or are willing to make stronger assumptions about treatment assignment than about the relationship between covariates and outcomes. A number of empirical examples using the propensity score suggest that this approach works reasonably well (see, for example, Rosenbaum and Rubin, 1984; 1985; Dehejia and Wahba, 1999; Imbens, Rubin and Sacerdote, 2001; Heckman, Ichimura and Todd, 1998).

This evidence of practical utility notwithstanding, from an asymptotic theory point of view propensity-score-based estimators present a puzzle. Hahn (1998) shows that the propensity score is ancillary for estimates of average treatment effects, in the sense that knowledge of the propensity score does not lower the semiparametric efficiency bound for this parameter. Moreover, covariate matching is asymptotically efficient, that is, it attains the semiparametric efficiency bound, while propensity score matching does not. These results based on conventional asymptotic arguments seem to offer no justification for anything other than full control for covariates in estimation of average treatment effects. The propensity score may still be a useful device. First, the propensity score may enhance finite sample efficiency. Angrist and Hahn (2004) use a non-standard asymptotic argument to point out that the traditional first-order asymptotic theory misses some of the subtleties in finite sample property, and observe that an estimator based on propensity score matching may be superior to the one based on covariate matching. They note, though, that the finite sample efficiency gain becomes smaller as the sample size grows, which is in accordance with the prediction from the traditional asymptotic theory. Second, when the estimated propensity scores are used as a weight in a certain way but not as a basis of matching, the estimator of the average treatment effects based on estimated propensity score is as asymptotically efficient as the covariate matching (Hirano, Imbens and Ridder, 2003).

See Also

- finite sample econometrics
- matching estimators

Bibliography

Angrist, J. and Hahn, J. 2004. When to control for covariates? Panel-asymptotic results for estimates of treatment effects. *Review of Economics and Statistics* 86, 58–72.

Dehejia, R. and Wahba, S. 1999. Propensity score matching methods for non-experimental causal studies. *Journal of the American Statistical Association* 94, 1053–62.

Hahn, J. 1998. On the role of the propensity score in the efficient estimation of average treatment effects. *Econometrica* 66, 315–32.

Heckman, J., Ichimura, H. and Todd, P. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65, 261–94.

Hirano, K., Imbens, G. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–89.

Imbens, G., Rubin, D. and Sacerdote, B. 2001. Estimating the effect of unearned income on labor supply, earnings, savings and consumption: evidence from a sample of lottery players. *American Economic Review* 91, 778–94.

Rosenbaum, P. and Rubin, D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Rosenbaum, P. and Rubin, D. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–24.

Rosenbaum, P. and Rubin, D. 1985. Constructing a control group using multi-variate matching methods that include the propensity score. *American Statistician* 39, 33–8.

How to cite this article

Hahn, Jinyong. "propensity score." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000335> doi:10.1057/9780230226203.1351

The New Palgrave Dictionary of Economics Online

property law, economics and

Dean Lueck

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This entry shows how the economics of property rights can be used to understand fundamental features of property law and related extra-legal institutions. It examines both the rationale for legal doctrine and the effects of legal doctrine regarding the exercise, enforcement, and transfer of rights. It also examines various property rights regimes including open access, private ownership, common property, and state property. Property law is understood as a system of societal rules designed to create incentives for people to maintain and invest in assets, which in turn leads to specialization and trade.

Keywords

Akerlof, G.: and land markets; asymmetric information; and adverse selection; and servitudes; average product rule; Coase Theorem; common interest communities; common property; compensation; and investment decisions; contracting: costs; Epstein, R.: and alienability of rights; Externalities; common law responses; neighbourhood; first possession; land use: servitudes; liability rules: vs. property rules; negligence: vs strict liability; open access property; paradox of compensation; Posner, R.: and eminent domain; private property; property law: externalities; property rights; regimes; origins; alieneability; property rules: vs liability rules; public trust; rule of capture; servitudes; and asymmetric information; social norms: and property rights; state property; strict liability: vs negligence; takings: and US Constitution; regulatory; transaction costs; and Coase Theorem; trespass: vs nuisance; zoning

Article

Property law is the body of court enforced rules that governs the establishment, use and transfer of rights to land and those assets attached to it such as air, minerals, water, and wildlife. In economic terms, property rights are defined as the (expected) ability of an economic agent to freely use an asset (Allen, 1999; Barzel, 1997; Lueck and Miceli, 2007; Shavell, 2004) and represent a social institution that

creates incentives to use, to maintain, and to invest in assets. Property rights may or may not be enforced by courts; and because the actions of courts are costly legal rights are but a subset of economic property rights. In addition to law and regulations, property rights may be enforced by custom and norms (see, for example, Ellickson, 1991) and by markets through repeated transactions.

Property rights, transaction costs, and the Coase Theorem

Consider Coase's (1960) famous example of the rancher and farmer. The rancher's cattle stray onto the farmer's land causing crop damage. The rancher's profit, $\pi(h)$ and the amount of crop damage $d(h)$ are functions of the rancher's herd size h, so the first-best optimal herd size, h^* maximizes $\pi(h) - d(h)$ and h^* solves $\pi'(h) = d'(h)$. This is also the choice made by a single farmer-rancher, Coase's ‘sole owner’ case. If the rancher initially has the economic (and legal) right to impose crop damage without penalty, he would choose the herd size to maximize $\pi(h)$, adding cattle until $\pi'(h) = 0$, which implies $h^r > h^*$. The farmer would be willing to pay up to $d'(h)$, his marginal damage, for each steer that the farmer removes from the herd in order to avoid crop damage, while the rancher would accept any amount greater than his marginal profit, $\pi'(h)$.

If transaction costs are zero, the parties will instantly contract to reduce the herd to the efficient size. The farmer will purchase the rights to the straying cattle, and if the farmer had the initial rights the situation would be reversed: either way the outcome is first-best. This is the Coase Theorem: *When transaction costs are zero the allocation of resources will be efficient regardless of the initial assignment of property rights*. But transaction costs are not zero and thus property rights are not perfectly defined (Allen, 1999; Barzel, 1997; Lueck and Miceli, 2007) so property law becomes important in defining rights and determining the allocation of assets. Indeed, Coase's (1960) discussion of nuisance law suggests an economic logic to the law in its assignment of property rights among various parties to these disputes.

Property rights: taxonomy and models

Property law recognizes several fundamental property rights regimes: private property, open access, common property, and state property (Lueck and Miceli, 2007). Property law also recognizes mixed regimes. Consider a fixed asset (such as a plot of land) used with a variable input (x) to produce a market output $(Y = f(x))$. If the input price is w, then the first-best use $(x^*(w))$ must maximize $R = f(x) - wx$ and satisfy $f'(x)=w$. The first-best value of the land is $V^* = \int_0^\infty R^*(x^*, t)e^{-rt}dt$, where r is the discount rate.

If there is ‘open access’ for n individuals, then output is $Y = f(\sum_{i=1}^{n} x_i)$ where x_i is the effort of the ith individual, $f'(\cdot) > 0$ and $f'(\cdot) < 0$, and the opportunity cost of effort is w_i. Each person can only capture (and thus own) the output in proportion to his share of effort, so each solves:

$$\max_{x_i} R_i = f^i(x_i) - w_i x_i \text{ subject to } f^i = \left[x_i / \sum_{i=1}^{n} x_i \right] f\left(\sum_{i=1}^{n} x_i \right) \tag{1}$$

On the assumption that users are homogeneous ($w_i = w_j$. for all $i \neq j$), the Nash open access equilibrium is $x = x^{oa}(n, w_1, \ldots, w_n)$, which satisfies,

$$(n - 1/n)\left(f\left(\sum_{i=1}^{n} x_i \right) / \sum_{i=1}^{n} x_i \right) - (1/n) f'\left(\sum_{i=1}^{n} x_i \right) = w_i, \quad i = 1, \ldots n. \tag{2}$$

In the limiting case as n→∞, (2) becomes $f(\Sigma_{i=1}^{n} x_i) / \Sigma_{i=1}^{n} x_i = w$ which is the famous ‘average product rule’ (Gordon, 1954; Cheung, 1970; Brooks et al., 1999). The limiting case implies that rents are completely dissipated, or $\Sigma_{i=1}^{n} R_i = \Sigma_{i=1}^{n} [f^i(x^{oa}) - wx^{oa}] = 0$ and the present value of the asset is also zero, $V^{oa} = \int_0^{\infty} R(x^{oa}, t) e^{-rt} dt = 0$. With heterogeneous costs, the infra-marginal users earn rents and have incentives to maintain open access regime (Libecap, 1989).

With private property the owner chooses $x^{*} < x^{oa}$ and generates $V^{*} > V^{oa} = 0$. Private ownership also creates incentives for optimal asset maintenance and investment (Bohn and Deacon, 2000). Let future output be $Y_{t+1} = f(x_t)$, where x_t is current investment, available at a market wage of w. and the interest rate is r. The first-best use of the input (x_t^{*}) must maximize $R = f(x_t) / (1 + r) - w_t x_t$ and satisfy $f'(x_t) / (1 + r) = w_t$. If $\pi \in [0, 1]$ is the probability of expropriation (because of imperfect rights) of the future output, then an owner will maximize $R = f(x_t)[(1 - \pi) / (1 + r)] - w_t x_t$. The solution $(x_t^{\pi} < x_t^{*})$ satisfies $f'(x_t)[(1 - \pi) / (1 + r)] = w_t$ and implies less than first-best investment. Pure open access means that no investor could claim future output $(\pi = 1)$, so $x_t^{oa} = 0$, and the rent from investment also equals zero. This lack of incentive to invest is essentially the problem of the ‘anti-commons’ described by Heller (1998) and formalized by Buchanan and Yoon (2000).

Common property is exclusive ownership by a group and may arise out of explicit private contracting (for example, unitized oil reservoirs) or out of custom (for example, common pastures); it may have legal (for example, riparian water rights) or regulatory (for example, hunting regulations) bases that have implicit contractual origins. Common property is well documented for natural resource stocks in less developed economies (Bailey, 1992; Ostrom, 1990). It is also seen in modern ‘common interest communities’ (such as condominiums, homeowner's associations) where residents use quasi-governments to maintain common areas (such as pools, open space) and provide local public goods

(Dwyer and Menell, 1998). Contracting to form common property creates a group that can realize economies of enforcing exclusive rights. Equal sharing is a typical internal allocation rule; it avoids costs of measuring and enforcing individual use but still leads to overuse compared with first-best. With equal sharing rules a homogeneous membership maximizes the present value of a common property resource (Lueck, 1994; 1995).

Governments own vast amounts of land, buildings, and capital equipment. State property rights are governed by administrative agencies, and the range of property rights regimes incorporates aspects of the three major types: private property, common property, and open access. State property rights commonly – and often severely – limit the transferability of rights, perhaps to limit the moral hazard incentives of agency bureaucrats. The relevant law for state property has its origins in common law (for example, mining on federal land is a first-possession rule) but is primarily governed by statutes and regulations, all shaped by bureaucrats, interest groups and politicians.

Real property regimes tend to mix the four fundamental types: open access, private property, common property and state property (Barzel, 1982; 1997; Eggertsson, 1990; Ellickson, 1993; Kaplow and Shavell, 1996; Merrill and Smith, 2000; Rose, 1998; Stake, 1999), implicitly recognizing that assets are a collection of valuable attributes. A rancher's land is not typically completely private: the streams running through the property may be open access for fishing or recreation; the grass may be a lease from a federal agency with mineral rights held by yet another private party. Similar scenarios are found in residential and commercial real estate, and Bailey (1992) found a mixture of ownership regimes among aboriginal peoples. Smith's (2000) study of the common field system of medieval Europe is a rare study of the underlying economic logic of a mixed property regime.

Origin of property rights

In law and custom, first possession is the dominant method of establishing rights, be it to the *flow* of output from a *stock* or to the stock itself (Lueck, 1995). Let $R(x(t))$ be the flow of benefits from an asset, where $x(t)$ is a variable input supplied at time t, r is the interest rate, and $g < r$ is the rate at which $R(t)$ grows over time. The first-best, full-information outcome is

$$V^{FB} = \int_{t=0}^{\infty} R(x^{*}(t))\, e^{-(r-g)t} dt, \tag{3}$$

where $x^{*}(t)$ is the optimal input level and $t^{*} = 0$ since production begins immediately.

Under first possession the asset's first claimant obtains exclusive rights to the temporal flow of rents, $\int_0^{\infty} R^{*}(t)\, dt$. Since establishing a bona fide claim will be costly and because $g < r$. property rights to the asset will emerge as the value of the asset increases (Demsetz, 1967). Along these lines an entire literature has developed to explain the ‘evolution of property rights’ or, more generally, the determinants

– both temporal and cross section – of property rights regimes (Lueck and Miceli, 2007; Rose, 1998). This literature, mostly empirical, notes that property rights regimes can move in both directions (to and away from private property), that property rights regimes can move among mixed regimes, and that political and other institutions also shape the choice of property regimes.

Returning to first possession, a single claimant will choose the claiming time to maximize

$$V^S = \int_t^{\infty} [R(x^*(t))e^{-(r-g)t}dt] - Ce^{-rt}, \tag{4}$$

where C is the cost of enforcing the claim and t is the time at which ownership of the stock (and the temporal flow of output) is established. The optimal time to establish ownership is when the present value of the asset's flow equals the present value of the opportunity cost of establishing rights at t^S, or $R^*e^{-(r-g)t^S} = rCe^{-rt^S}$. The asset value falls short of first-best, or $V^S < V^{FB}$, because the costs of establishing ownership delay ownership and production to t^S from $t = 0$.

First possession can dissipate value when there is unconstrained competition among homogenous claimants (Barzel, 1968; Mortensen, 1982). A competitive rush to claim rights causes ownership to be established at exactly the time t^R when the present value of the rental flow at t^R equals the present value of the entire costs of establishing ownership at t^R, or when $R^*e^{-(r-g)t^R} / (r-g) = Ce^{-rt^R}$. In this 'race equilibrium' rights are established at t^R, where $t^R < t^S$ since $t^R = (\ln(r-g) + \ln C - \ln R) / g$ and $t^S = (\ln r + \ln C - \ln R) / g$, and the rental stream is fully dissipated; or

$$V^R = \int_{t^R}^{\infty} [R(x^*(t))e^{-(r-g)t^R}dt] - Ce^{-rt^R} = 0. \tag{5}$$

Heterogeneity among claimants can reduce, or eliminate, dissipation (Barzel, 1994; Lueck, 1995). If there are two competitors (i and j) with possession costs $C_i < C_j$, and neither party knows the other's costs, then i gains ownership just before j makes a claim, at $t^i = t^R - \varepsilon$, and earns rent equal to the present discounted value of his cost advantage. The key implication is that, as the differential between the two lowest cost claimants $(C_j - C_i)$ increases, the level of dissipation will decrease. With complete information there is no dissipation because only the low-cost claimant has a positive expected payoff in a race (Fudenberg et al., 1983; Harris and Vickers, 1985).

If the costs of enforcing a claim to the asset are prohibitive, ownership may be established only by

capturing or ‘reducing to possession’ a flow from the asset. The legal term ‘rule of capture’ describes this derivative of the rule of first possession. Wildlife and crude oil are the classic examples: ownership is established only when a hunter bags a pheasant or when a barrel of oil is brought to the surface. The stock itself (that is, the pheasant population or oil reservoir) remains unowned. The new ‘race’ is to claim the present flow $R(t)$ and leads to open access dissipation (Epstein, 1986; Lueck, 1995) since no one owns the asset's entire stream of flows, $\int_0^\infty R(t)\,dt$. The formal analysis is static rather than inter-temporal as in the asset claim race, and is identical to the open access model developed above in equation (1).

Property law implicitly recognizes the two potential paths of dissipation – racing and over-exploitation – and is structured to limit such dissipation (Dharmapala and Pitchford, 2002; Lueck, 1995; 1998). Where first possession rules establish ownership in a resource stock, first possession tends to be defined so that valid claims are made at low cost and before dissipating races begin, thus exploiting claimant heterogeneity. Also, the transfer of rights to the resource is allowed, routinely reflecting security of ownership in the corpus. Where the rule of capture emerges (for example, oil and wildlife) access to the resource tends to be limited through legal, contractual or regulatory methods. As well, the transfer of rights to capturable flows tends to be restricted in order to limit overuse of the asset itself.

Externalities and property law: nuisance, trespass and zoning

Externalities arise because property rights to at least some of the attributes of an asset will be imperfect and thus generate problems of open access or moral hazard. Land externalities are ubiquitous because any parcel (except an island) will have neighbouring owners and because related resources (for example, air, noise, minerals, water) do not tend to coincide with the surface ownership boundaries. Property law addresses externalities through doctrines of trespass, nuisance, servitudes, and through regulatory zoning. Consider, à la Coase (1960), a railroad whose trains emit sparks that occasionally set fire to adjacent farmland. The number of trains is n_T and the number of farms is n_F, resulting in crop damage of $n_T n_F D(x, y)$, where D is the damage (reduced crop value per acre) each train causes, x is the cost of precaution per train, and y is the cost of precaution by each farmer. Assume $D_x < 0,\ D_y < 0\ D_{xx} > 0$, and $D_{yy} > 0$. The marginal benefits are $b_T(n_T)$ and $b_F(n_F)$, where $b_j{}' < 0$. $j=T,F$. The total value of the two activities is

$$W = \int_0^{n_T} b_T(n_T)\,du + \int_0^{n_F} b_F(n_F)\,dz - [n_T n_F D(x, y) + n_T x + n_F y] \tag{6}$$

If the numbers of trains and farms are fixed, as in tort models (Shavell, 1980) that hold ‘activity levels’ fixed, the optimal precaution choices (x^*, y^*) that maximize (6) are $n_F D_x(x, y) + 1 = 0$ and $n_T D_y(x, y) + 1 = 0$. If the number of trains and farms (n_T, n_F) is endogenous, the resulting first-order

conditions are $b_T(n_T) - [n_F D(x, y) + x] = 0$ and $b_F(n_F) - [n_T D(x, y) + y] = 0$.

Remedies for externalities can be viewed as a choice between ‘property rules’ and ‘liability rules’ (Calabresi and Melamed, 1972; Polinsky, 1980). Under property rules, rights holders can refuse any unwanted infringements of their rights, enforceable by injunctions (or criminal sanctions in the case of theft). Property rules thus form the legal basis for voluntary (market) exchange of rights. With liability rules, however, owners can only seek monetary compensation in the form of damages. Liability rules thus form the basis for court-ordered or non-consensual transactions. The choice between the two rules turns on transaction costs, particularly the costs of contracting, the costs of court adjudication, and legal administration. When contracting costs are relatively low, property rules are preferred because they ensure that all transactions are mutually beneficial. When contracting costs are high (for example, in public nuisance cases), property rules may prevent otherwise efficient transactions from occurring. Liability rules have an advantage because courts can force an efficient transfer. This advantage of liability rules must be weighed against the possibility of court error in setting damages, and, because liability rules require courts to establish the initial terms of a transaction by setting damages, the administrative costs of using this rule will likely be higher than under a property rule (Kaplow and Shavell, 1996).

In the railroad–farmer case, if liability is strict the railroad must pay full compensation regardless of its level of precaution. Strict liability induces efficient precaution by the railroad, but farmers are fully compensated and thus have no incentive for precaution. Negligence, which holds the railroad liable for damages only if it takes less than the efficient level of abatement, will induce both parties to take efficient care. Neither rule, however, will achieve first-best railroad and farm activity levels. In general, liability rules cannot create first-best incentives because of the constraint that what one party pays the other must receive. This is an example of the *paradox of compensation* which is also found in tort law and contract law remedies (Cooter, 1985). It can be avoided by ‘decoupling’ liability and compensation, or by using a contract or compensation mechanism that defines and enforces the optimal choices for both parties.

Trespass (for example, squatting, boundary encroachment) and nuisance (for example, air, water, noise pollution) doctrines are the primary common law responses to externalities. The primary remedy under trespass is an injunction, a property rule. The remedy under nuisance law is more complicated. A landowner can obtain relief only if the invasion is substantial, and even then he may have to be satisfied with money damages (a liability rule). If a landowner wishes the harm to be enjoined, he must meet the further legal standard of showing that the harm outweighs the benefit of the nuisance-creating activity. The trespass–nuisance distinction can be understood as a property–liability rule choice (Merrill, 1998). Trespass ordinarily involves a small number of parties where the intruder is easily identifiable, so contracting costs tend to be low and property rules are likely optimal. Nuisance often involves large numbers or sources of harm that are difficult to identify, so liability rules are likely optimal.

Zoning is a common legal response to urban land externalities. The economic rationale for zoning is that ‘similar land uses have no (or only small) external effects on each other whereas dissimilar land uses may have large effects’ (White, 1975), creating what the common law calls a ‘public nuisance’. Ellickson (1973) argues that zoning may have administrative and enforcement costs that often exceed the saved ‘nuisance costs’. A private alternative to zoning is the use of land use servitudes (for example, covenants, easements) that impose limits on what landowners can do with their property. Such

restrictions are frequently observed in condominiums, homeowner associations, and other ‘common interest communities’ (Dwyer and Menell, 1998; Hansmann, 1991). The economic function of these restrictions is twofold: to overcome free rider problems in the provision of certain jointly consumed amenities; and to internalize neighbourhood and rental externalities.

Public trust, public property and public use

The ancient doctrine of ‘public trust’ grants ownership of navigable rivers, shorelines, and the open sea to the public. It is judicially created common property, or sometimes open access. In its traditional application the public trust asset was a public good. When an asset is a public good, unrestricted access will not cause dissipation from overuse of the resource, but it could lead to underinvestment. When the resource has private good characteristics, unrestricted access can trigger the rule of capture and creates a classic open access problem, possibly causing resource degradation through overuse. Some courts have recently extended the doctrine into environmental assets, such as beaches, lakes, stream access and wildlife.

Large-scale projects like dams, railroads and highways often involve the assembly of a large contiguous parcel of land from relatively small and separately owned parcels. Developers face a potential holdout problem because, once assembly becomes public information, parcel owners might hold out for prices in excess of their true valuations, endangering completion of an otherwise efficient project. One solution is to force sales by replacing property rule protection of each owner's land with liability rule protection. This is the economic justification for the eminent domain power of the state (Posner, 2003), which has common law origins. The ‘takings’ clause of the Fifth Amendment of the US Constitution explicitly grants such eminent domain power for ‘public use’ but requires ‘just compensation’, which courts have interpreted to mean ‘fair market value’. Since subjective value is part of the opportunity cost of a taking, failure to compensate for it potentially results in excessive acquisition of land by the government, though one study (Munch, 1979) found that high-valued properties were overcompensated, while owners of low-valued properties were undercompensated.

A large literature has studied the link between compensation and investment decisions of landowners (Blume, Rubinfeld and Shapiro, 1984; Fischel and Shapiro, 1988). Suppose there are many parcels, each worth $V(x)$ if the landowner makes an irreversible investment x, where $V' > 0$ and $V'' < 0$. The land also yields a public benefit of $B(y)$, where y is the number of parcels taken and $B' > 0$, $B'' < 0$. Compensation of $C(x)$ will be paid for each parcel taken, where $C(x) \geq 0$, $C' \geq 0$, and total compensation is $yC(x)$. Landowners choose x given the anticipated behaviour of the government and the compensation rule; then the government chooses y and pays $C(x)$. The first-best choices (x^*, y^*) maximize $B(y) + (1 - y)V(x) - x$, the sum of private and public benefits, and must satisfy $(1 - y)V'(x) - 1 = 0$ and $B'(y) - V(x) = 0$. If the taking is exogenous, y is fixed and the landowner will maximize $(1 - y)V(x) + yC(x) - x$, which must satisfy $(1 - y)V'(x) + yC'(x) - 1 = 0$ and also gives x^l as the solution. This means that compensation must be lump sum $(C' = 0)$ to ensure that $x^l = x^*$; a positive relationship between x and compensation creates over-investment moral hazard (another example of the paradox of compensation). Thus no compensation ($C(x) \equiv 0$ for all x) is actually

efficient, although any lump sum rule is consistent with efficiency. The efficiency of zero compensation, however, depends on assumptions about government behaviour.

Government regulations often restrict land uses without depriving the owner of title (for example, zoning laws, environmental regulations). Historically, courts have granted broad powers to enact such regulations but, when a regulation becomes especially burdensome, the affected landowner may claim that a ‘regulatory taking’ has occurred and seek compensation. As above, the trade-off for regulatory takings concerns the efficiency of the land use decision on the one hand and the regulatory decision on the other. Miceli and Segerson (1994; 1996) propose the following compensation rule, where *y* is a landowner's lost value from the regulation:

$$C = \begin{cases} 0 & \text{if } y \le y^* \\ V(x), & \text{if } y > y^* . \end{cases} \tag{7}$$

Like a negligence rule in tort law, this rule requires full compensation if the government over-regulates $(y > y^*)$ but requires no compensation otherwise $(y \le y^*)$. It also establishes a standard that is economically equivalent to the common law definition of a nuisance (an activity that is efficiently prohibited), and hence is consistent with the threshold for compensation implied by the nuisance exception.

Inalienability of property rights

Posner (2003, p. 75) notes, ‘the law should, in principle, make property rights freely transferable in order to allow resources to move to their most highly valued uses and to foster the optimal configuration of assets.’ Yet there are many legal restrictions that limit the alienability of property: body parts, children, voting, military service, cultural artifacts, endangered animal species, the right to freedom (laws against slavery), certain natural resources and state property.

The dominant economic reason for restrictions on alienability is that externalities can arise from transfers (Barzel, 1997; Epstein, 1985; Rose-Ackerman, 1985; Posner, 2003) if the rights to the assets are not well-defined with respect to the stock (and its stream of flows over time). This generates a rationale for limiting, even prohibiting, certain transfers of the claimed flows in order to protect the asset and its value. For example, the widespread prohibition on trade in wild game is likely to be such a case (Lueck, 1989; 1998), though even here limits on markets can potentially deter the formation of property rights. Restrictions on the sale of children may have a similar rationale: a market for children (or game) would lead to ‘poaching’ of kids (or animals) for which property rights enforcement is extremely costly.

Another reason for restricting transfers is asymmetric information, particularly that leading to adverse selection (Rose-Ackerman, 1998). Adverse selection can potentially dry up markets where product quality cannot be observed prior to purchase. Similar restrictions on the types of property servitudes

allowed (such as limits on ‘negative and in gross’ easements) might be explained by reference to asymmetric information (Dnes and Lueck, 2006). Legal scholars have argued that limitations on servitudes prevent ‘clogging title’ (Gray and Gray, 2000). Consider the market for land of two types: fee simple (that is, unencumbered) and land encumbered with a servitude. Assume that only the seller knows whether the land is encumbered. Buyers do not have this information but know only that one-half of the land is encumbered. The value of an unencumbered plot is V^f, while the value of the encumbered plots is $V^s < V^f$. Given the information asymmetry, buyers will pay only the expected value of a plot, $EV = (V^s + V^f)/2 < V^f$. Following Akerlof (1970) and related literature, this means there will be no market equilibrium for the unencumbered plots; that is, only ‘low-quality’ encumbered plots will be present in the market. Institutions that provide information (such as title recording and registration systems) could eliminate asymmetry and even alter the law of property by allowing an expanded set of servitudes.

Summary

Economic analysis reveals a fundamental logic to the main doctrines and features of property law (Lueck and Miceli, 2007). The observed structure of property rights and property law can be best understood as a system designed to create incentives for people to maintain and invest in assets, which in turn leads to specialization and trade. Among the most important remaining issues for study is a systematic analysis of how the law addresses the use and transfer of complex assets.

See Also

- Akerlof, George Arthur
- contract theory
- law, economic analysis of
- tragedy of the commons

Bibliography

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Allen, D. 1999. Transaction costs. In *Encyclopedia of Law and Economics*, vol. I, ed. B. Bouckaert and G. DeGeest. Cheltenham: Edward Elgar.

Bailey, M. 1992. Approximate optimality of aboriginal property rights. *Journal of Law and Economics* 35, 183–98.

Barzel, Y. 1968. The optimal timing of innovations. *Review of Economics and Statistics* 50, 348–55.

Barzel, Y. 1982. Measurement cost and the organization of markets. *Journal of Law and Economics* 25, 27–48.

Barzel, Y. 1994. The capture of wealth by monopolists and the protection of property rights. *International Review of Law and Economics* 14, 393–409.

Barzel, Y. 1997. *Economic Analysis of Property Rights*, 2nd edn. Cambridge: Cambridge University Press.

Besley, T. 1998. Investment incentives and property rights. In *The New Palgrave Dictionary of Law and Economics*, vol. 2, ed. P. Newman. New York: Stockton Press.

Blume, L., Rubinfeld, D. and Shapiro, P. 1984. The taking of land: when should compensation be paid? *Quarterly Journal of Economics* 99, 71–92.

Bohn, H. and Deacon, R.T. 2000. Ownership risk, investment, and the use of natural resources. *American Economic Review* 90, 526–49.

Brooks, R., Murray, M., Salant, S. and Weise, J. 1999. When is the standard analysis of common property extraction under free access correct? *Journal of Political Economy* 107, 843–58.

Buchanan, J. and Yoon, Y. J. 2000. Symmetric tragedies: commons and anticommons. *Journal of Law and Economics* 43, 1–14.

Calabresi, G. and Melamed, A. 1972. Property rules, liability rules, and inalienability: one view of the cathedral. *Harvard Law Review* 85, 1089–128.

Cheung, S. 1970. The structure of a contract and the theory of a nonexclusive resource. *Journal of Law and Economics* 13, 49–70.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Cooter, R. 1985. Unity in tort, contract, and property: the model of precaution. *California Law Review* 73, 1–51.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57, 347–59.

Dharmapala, D. and Pitchford, R. 2002. An economic analysis of ‘riding to hounds’: Pierson v. Post revisited. *Journal of Law, Economics, and Organization* 18, 39–66.

Dnes, A. and Lueck, D. 2006. An economic analysis of servitudes on land. Unpublished manuscript.

Dwyer, J. and Menell, P. 1998. *Property Law and Policy: A Comparative Institutional Perspective*. Westbury, NY: The Foundation Press.

Eggertsson, T. 1990. *Economic Behavior and Institutions*. Cambridge: Cambridge University Press.

Ellickson, R. 1973. Alternatives to zoning: covenants, nuisance rules, and fines as land use controls. *University of Chicago Law Review* 40, 681–782.

Ellickson, R. 1991. *Order without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

Ellickson, R. 1993. Property in land. *Yale Law Journal* 102, 1315–400.

Epstein, R. 1979. Possession as the root of title. *Georgia Law Review* 85, 1221–43.

Epstein, R. 1985. Why restrain alienation? *Columbia Law Review* 85, 970–90.

Epstein, R. 1986. Past and future: the temporal dimension in the law of property. *Washington University Law Quarterly* 64, 667–722.

Fischel, W. and Shapiro, P. 1988. Takings, insurance, and Michelman: comments on economic interpretations of ‘just compensation’ law. *Journal of Legal Studies* 17, 269–93.

Fudenberg, D., Gilbert, R., Sitglitz, J. and Tirole, J. 1983. Preemption, leapfrogging, and competition in patent races. *European Economic Review* 22, 3–31.

Gordon, H. 1954. The economic theory of a common-property resource: the fishery. *Journal of Political Economy* 62, 24–42.

Gray, K. and Gray, S. 2000. *Elements of Land Law*, 3rd edn. London: Butterworths.

Hansmann, H. 1991. Condominium and cooperative housing: transactional efficiency, tax subsidies, and tenure choice. *Journal of Legal Studies* 20, 25–71.

Harris, C. and Vickers, J. 1985. Perfect equilibrium in a model of a race. *Review of Economic Studies* 52, 193–209.

Heller, M. 1998. The tragedy of the anti-commons: property in transition from Marx to markets. *Harvard Law Review* 111, 621–88.

Kaplow, L. and Shavell, S. 1996. Property rules versus liability rules. *Harvard Law Review* 109, 713–90.

Libecap, G. 1989. *Contracting for Property Rights*. Cambridge: Cambridge University Press.

Lueck, D. 1989. The economic nature of wildlife law. *Journal of Legal Studies* 18, 291–324.

Lueck, D. 1994. Common property as an egalitarian share contract. *Journal of Economic Behavior and Organization* 25, 93–108.

Lueck, D. 1995. The rule of first possession and the design of the law. *Journal of Law and Economics* 38, 393–436.

Lueck, D. 1998. Wildlife law. In *The New Palgrave Dictionary of Economics and the Law*, vol. 3, ed. P. Newman. New York: Stockton Press.

Lueck, D. and Miceli, T. 2007. Property rights and property law. In *Handbook of Law and Economics*, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland.

Merrill, T. 1998. Trespass and nuisance. In *The New Palgrave Dictionary of Economics and the Law*, vol. 3, ed. P. Newman. New York: Stockton Press.

Merrill, T. and Smith, H. 2000. Optimal standardization in the law of property: the *numerus clausus* principle. *Yale Law Journal* 110, 1–70.

Miceli, T. 1997. *Economics of the Law: Torts, Contracts, Property, Litigation*. New York: Oxford University Press.

Miceli, T. and Segerson, K. 1994. Regulatory takings: when should compensation be paid? *Journal of Legal Studies* 23, 749–76.

Miceli, T. and Segerson, K. 1996. *Compensation for Regulatory Takings: An Economic Analysis with Applications*. Greenwich: JAI Press.

Mortensen, D. 1982. Property rights and efficiency in mating, racing, and related games. *American Economic Review* 72, 968–79.

Munch, P. 1976. An economic analysis of eminent domain. *Journal of Political Economy* 84, 473–97.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Polinsky, A. 1980. On the choice between property rules and liability rules. *Economic Inquiry* 18, 233–46.

Posner, R. 2003. *Economic Analysis of Law*, 6th edn. New York: Aspen Law and Business.

Rose, C. 1985. Possession as the origin of property. *University of Chicago Law Review* 53, 73–88.

Rose, C. 1986. The comedy of the commons, custom, commerce, and inherently public property. *University of Chicago Law Review* 53, 711–81.

Rose, C. 1998. Evolution of property rights. In *The New Palgrave Dictionary of Economics and the Law*, vol. 2, ed. P. Newman. New York: Stockton Press.

Rose-Ackerman, S. 1985. Inalienability and the theory of property rights. *Columbia Law Review* 85, 931–69.

Rose-Ackerman, S. 1998. Inalienability. In *The New Palgrave Dictionary of Economics and the Law*, vol. 2, ed. P. Newman. New York: Stockton Press.

Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 9, 1–25.

Shavell, S. 2004. *Foundations of Economic Analysis of Law*. Cambridge, MA: Harvard University Press.

Smith, H. 2000. Semicommon property rights and scattering in the open fields. *Journal of Legal Studies* 29, 131–69.

Stake, J. 1999. Decomposition of property rights. In *Encyclopedia of Law and Economics*, vol. 2, ed. B. Bouckaert and G. DeGeest. Cheltenham: Edward Elgar.

White, M. 1975. Fiscal zoning in fragmented metropolitan areas. In *Fiscal Zoning and Land Use Controls*, ed. E. Mills and W. Oates. Lexington, MA: Lexington Books.

How to cite this article

Lueck, Dean. "property law, economics and." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000219> doi:10.1057/9780230226203.1352

The New Palgrave Dictionary of Economics Online

property rights

Armen A. Alchian
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Keywords

corporations; exchange value; firm, theory of; firm-specific resources; fisheries; government property rights; joint production; limited liability; moral hazard; mutual property rights; ownership and control; private property rights; property rights; quasi-rent; specialization; torts; value.

Article

Private property rights

A property right is a socially enforced right to select uses of an economic good. A Private property right is one assigned to a specific person and is alienable in exchange for similar rights over other goods. Its strength is measured by its probability and costs of enforcement which depend on the government, informal social actions, and prevailing ethical and moral norms. In simpler terms, no one may legally use or affect the physical circumstances of goods to which you have Private property rights without your approval or compensation. Under hypothetically perfect Private property rights none of my actions with my resources may affect the physical attributes of any other person's private property. For example, your Private property rights to your computer restrict my and everyone else's permissible behaviour with respect to your computer, and my Private property rights restrict you and everyone else with respect to whatever I own. It is important to note that it is the physical use and condition of a good that are protected from the action of others, not its exchange value.
Private property rights are assignments of rights to choose among inescapably incompatible uses. They are not contrived or imposed restrictions on the feasible uses, but assignments of exclusive rights to choose among such uses. To restrict me from growing corn on my land would be an imposed, or contrived, restriction denying some rights without transferring them to others. To deny me the right to grow corn on my land would restrict my feasible uses without enlarging anyone else's feasible physical

uses. Contrived or unnecessary restrictions are not the basis of Private property rights. Also, because those restrictions typically are imposed against only some people, those who are not so restrained obtain a ‘legal monopoly’ in the activity from which others are unnecessarily restricted.

Under Private property rights any mutually agreed contractual terms are permissible, though not all are necessarily supported by governmental enforcement. To the extent that some contractual agreements are prohibited, Private property rights are denied. For example, it may be considered illegal to agree to work for over ten hours a day, regardless of how high a salary may be offered. Or it may be illegal to sell at a price above some politically selected limit. These restrictions reduce the strength of private property, market exchange and contracts as means of coordinating production and consumption and resolving conflicts of interest.

Economic theory and Private property rights

A successful analytic formulation of Private property rights has resulted in an explanation of the method of directing and coordinating uses of economic resources in a private property system (that is, a capitalistic or a ‘free enterprise’ system). That analysis relies on convex preferences and two constraints: a production possibility and a private property exchange constraint, expressible biblically as ‘Thou Shalt Not Steal’, or mathematically, as the conservation of the exchange values of one's good.

For the decentralized coordination of productive specialization to work well, according to the well known principles of comparative advantage, in a society with diffused knowledge, people must have secure, alienable Private property rights in productive resources and products tradable at mutually agreeable prices at low costs of negotiating reliable contractual transactions. That system's ability to coordinate diffused information results in increased availability of more highly valued goods as well as of those becoming less costly to produce. The amount of rights to goods one is willing to trade, and in which Private property rights are held, is the measure of value; and that is not equivalent to an equal quantity of goods not held as private property (for example, government property). It probably would not be disputed that stronger Private property rights are more valuable than weaker rights, that is, a seller of a good would insist on larger amounts of a good with weaker Private property rights than if Private property rights to the goods were stronger.

Firms, firm-specific resources and the structure of property rights

Though Private property rights are extremely important in enabling greater realization of the gains from specialization in production, the partitionability, separability and alienability of Private property rights enables the organization of cooperative joint productive activity in the modern corporate firm. This less formally recognized, but nevertheless important, process of cooperative production relies heavily on partitioning and specialization in the components of Private property rights. Yet this method is often misinterpreted as unduly restrictive and debilitating to the effectiveness and social acceptability of Private property rights. To see the error, an understanding of the nature of the firm is necessary, especially in its corporate form, which accounts for an enormous portion of economic production. The ‘firm’, usually treated as an output-generating ‘black box’, is a contractually related collection of resources of various cooperating owners. Its distinctive source of enhanced productivity is ‘team’ productivity, wherein the product is not a sum of separable outputs each of which is attributable to

specific cooperating inputs, but instead is a non-decomposable, non-attributable value produced by the group. Thus, for something produced jointly by several separately owned resources, it is not possible to identify or define how much of the final output value each resource could be said to produce separately. Instead, a marginal product value for each input is definable and measurable.

Whereas specialized production under comparative advantage and trade is directed in a decentralized process by market price and spot exchanges, productivity in the team, called the firm, relies on long-term, constraining contracts among owners who have invested in resources specialized to the group of inputs in that firm. In particular, some of the inputs are specialized to the team in that once they enter the firm their alternative (salvage) values become much lower than in the firm. They are called ‘firm-specific’. In the firm, firm-specific inputs tend to be owned in common or else contracts among separate owners of the various inter-specific resources restrict their future options to those beneficial to that group of owners as a whole rather than to any individual. These contractual restrictions are designed to restrain opportunism and ‘moral hazard’ by individual owners, each seeking a portion of each other's firm-specific, expropriable composite quasi-rent. Taking only extremes for expository brevity, the other ‘general’ resources would lose no value if shifted elsewhere. A firm, then, is a group of firm-specific and some general inputs bound by constraining contracts, producing a non-decomposable end-product value. As a result, the activities and operation of the team will be most intensively controlled and monitored by the firm-specific input owners, who gain or lose the most from the success or failure of the ‘firm’. In fact, they are typically considered the ‘owners’ or ‘employers’ or ‘bosses’ of the firm, though in reality the firm is a cooperating collection of resources owned by different people.

Firm-specific resources can be non-human. Professional firms – in law, architecture, medicine – are comprised of teams of people who would be less valuable elsewhere in other groups. They hire non-human general capital, for complex example building and equipment. The contract, which defines ‘hiring’, depends on the specificity and generality, not on human or non-human attributes nor on who is richer. Incidentally, ‘industrial democracy’ arrangements are rare, because the owners of more general resources have less interest in the firm than those of specific resources.

The corporation and specialization in Private property rights

In a corporation the resources owned by the stockholders are those the values of which are specific to the firm. The complexities in specialization in exercise of the components of property rights and the associated contractual restraints have led some people to believe that the corporation tends to insulate (for example, ‘separate’) decisions of use from the bearing of the consequences (that is, control from ownership) and thereby has undermined the capacity of a private property system to allocate resources to higher market value uses. For example, it has been argued that diffused stock ownership has so separated management and control of resources from ‘ownership’ that managers are able to act without sufficient regard to market values and the interests of the diffused stockholders. Adam Smith was among the first to propound that belief. Whatever the empirical validity, the logical analysis underlying those charges rests on misperceptions of the structure of Private property rights in the corporation and the nature of the competitive markets for control and ownership which tend to restrain such managers. What individual managers seek, and what those who survive are able successfully to do in the presence of competition for control, are very different things.

An advantage of the corporation is its pooling of sufficient wealth in firm-specific resources for large-

scale operations. Pooling is enabled if shares of ownership are alienable private property, thereby permitting individuals to eliminate dependence of their time path of consumption on the temporal pattern of return from firm-specific investments. Alienability is enabled if the shares have limited liability, which frees each stockholder from dependence on the amount of wealth of every other stockholder. The resultant ability to tolerate anonymity, that is, disinterest in exactly who are the other shareholders, enables better market alienability.

When voluntary separability of decision authority over firm-specific resources from their market value consequences is added to alienability, the ability to specialize in managerial decisions and talent (control) without also having to bear the risk of all the value consequences, enables achievement of beneficial specialization in production and coordination of cooperative productivity. Specialization is not necessarily something that is confined to the production of different end products; it applies equally to different productive inputs or talents. Voluntary partitionability and alienability of the component rights enable advantageous specialization (sometimes called ‘separation’) in (*a*) exercise of rights to make decisions about uses of resources and (*b*) bearing the consequent market or exchange values. The former is sometimes called ‘control’ and the latter, ‘ownership’. Separability enables the achievement of the gains from specialization in selecting and monitoring uses, evaluating the results, and bearing the risk of consequent future usefulness and value. Because different uses have different prospective probability distributions of outcomes, and because outcomes are differentially sensitive to monitoring the prior decisions, separability and alienability of the component rights permit gains from specialization in holding and exercising the partitionable rights.

Thus, the modern corporation relies on limited liability to enhance alienability and on partitionability of components of Private property rights in order to achieve gains from large-scale specialization in directing productive team activity and talents. Rather than destroying or undermining the effectiveness of Private property rights, the alleged ‘separation’ enables effective, productive ‘specialization’ in exercising Private property rights as methods of control and coordination.

Government property rights

It might be presumed that Government property rights in a democracy are similar to corporate property with diffused stockholdings and should yield similar results. The analogy would be apt if each voting citizen had a share of votes equivalent to one's share of the wealth in the community, and if a person could shift wealth among governments, as one can among different corporations. If, for example, one could buy and sell land (as assets capturing essentially most of the value of whatever the government does in that particular state) in several different governments and could vote in each in proportion to the value of that ‘land’ then government property would be closer to private property in its effects. But it is difficult to take that possibility seriously. The nature of government, public or communal property rights surely depends on the kind of government. Because these are so vaguely and indefinitely defined, attempts to deduce formally the consequences of resource allocation and behaviour under each have been hampered.

Non-existent property rights

Not all resources are satisfactorily controlled by Private property rights. Air, water, electromagnetic

radiation, noises and views are some examples. Water under my land flows to yours. Sounds and light from my land impinge on yours. Other forms of control are then designed, for example, political or social group decisions and actions, though these other forms are sometimes employed for ideological or political purposes, even where Private property rights already exist.

If these other forms permit open, free entry with every user sharing equally and obtaining the average return, use will be excessive. Extra uses will be made with an increased realized total value that is less than the cost added, that is, the social product value is not maximized. This occurs because the marginal yield is less than the average to each user, to which each user responds. So, use occurs to the point where the average yield is brought down to marginal cost, with the consequence that the marginal yield is less than the marginal cost – often exampled as excessive congestion on a public road or public park, or over-fishing of communal, free access fishing areas. The classic ‘communal property’ implication that apples on the public apple tree are never allowed to ripen is an extreme example of the proposition that property rights, other than private, reduce conformity of resource uses to market revealed values.

Alternatively, if communal property rights mean that incumbent users can block more users, the resource will be underutilized as incumbents maximize their individual yield, which is the average, not the marginal. This results in fewer users. Though more users or uses would lower the average value to the incumbents and hence dissuade a higher rate of use, the addition to the total group value (of the extra use) exceeds the extra costs. Examples are public, low tuition colleges that restrict entry to maximize the ‘quality’ of those who are educated – that is, to maximize the average yield of those admitted. Some labour unions (that is, teamsters) are examples of similar situations.

A mistaken inference commonly suggested by the example of fishermen who overfish unowned lakes is that independent sellers with open access to customers will ‘over-congest’ in product variety and advertising to catch customers, with unheeded costs borne by other sellers. If, for example, Pall Mall cigarettes attract some customers from Camel, the loss to Camel is the reduced value of Camel-specific resources, not its lost sales revenue. General resources will be released from making Camels for use elsewhere with no social loss. But Camel-specific resources fall in value by the extent to which Pall Mall's product is better or cheaper. Camel's loss is more than offset by the sum of Pall Mall's increased net income plus the transfer gain to customers from lower prices or better quality. The loss to Camel is not from new entry itself, but from its incorrect forecasts of its earlier investment value. It is presumed here that mistaken forecasts should not be protected by prohibiting the unexpected future improvements.

This differs from the over-fishing case in that consumers, in contrast to fish, have property rights in what they pay and what they buy. If every fish had a separate owner or owned itself, none would allow it to be caught unless paid enough, and over-fishing would not occur. One owner of all the fish is unnecessary; it suffices that each fish (or potential customer) be owned by someone who can refuse to buy. (Of course, unless the lake were owned, the lake surface might be over-congested with too many fishermen, each fishing to a lesser area, even if the fish were owned.)

Ownership of tradable rights by customers is the feature that is missing in the over-fishing, over-congestion case. Because rights to (or ‘of’) the fish or whales need not be bought, over-fishing does not imply over-customering where customers own rights to what the competing sellers are seeking. Otherwise, customers could be caught like fish, wherein sellers would be competing both to (1) establish property rights over the customers and to (2) possess those rights. Costly redundant competition for initial establishment of rights could be avoided simply by establishing customers’ rights to themselves, as is in fact done. If the preceding seems fanciful, replace ‘fish’ with people and the lake surface with

streets on which taxi-drivers cruise for customers. Excessive costs will be incurred in competition for use of unowned, valuable resources, in this case, the streets.

Mutual property rights

‘Mutual’ forms of organization are used apparently in order to sustain the maximum average per member, or to reserve for the incumbent members any greater group value from more members. Mutual private property, a form that has barely been analysed, does not permit anonymous alienability of interests in what are otherwise Private property rights. A ‘mutual’ member can transfer its interest to other people only upon permission of the other mutually owning members or their agents. Fraternal, social and country clubs are examples. These activities have not typically been viably organized and their services sold, as for example, in restaurants and health and exercise gymnasia. The intragroup-specific resources are themselves the members (erstwhile customers) who interact and create their social utility. More members affect each incumbent's realized utility in two ways: by social compatibility and by congestion. An outside, separate owner interested in the maximum value of the organization, but not the maximum average-per-member, could threaten to sell more memberships which, although enabling a larger total social value with more members, would reduce the average value to the existing members. This is an example of the earlier analysed difference between maximizing the average yield per input rather than the total yield by admitting more members, who while they would be made better off than if not admitted nevertheless reduce the average value to the incumbent members. In addition, the ability of newcomers to compensate incumbents for any loss in the individual (average) value to incumbent members is restrained if the membership fee were to go instead to an outside owner of the club. To the extent that a pecuniary compensation, via an initiation fee, were paid to an outside owner and exceeded the reduction in their average individual and total group utility, newcomers would be admitted, and the outside owner would gain, but incumbent members would lose their composite quasi-rent of their interpersonal sociability. (It is not yet well understood why, aside from tax reasons, the mutual form occurs in savings and loans and insurance firms.)

Torts, conditional and unassigned property rights

Private property rights may exist in principle, but, quite sensibly, not be blindly and uncompromisingly enforced against all possible ‘usurpers’. For example, situations arise in which someone's presumed Private property rights do not exclude an ‘invader’s’ use. Accidental or emergency use of some other person's private property without prior permission constitutes an example, sometimes called a ‘tort’. Another possibility is that the property rights are so ill-defined that whether a right has been usurped or already belonged to the alleged ‘usurper’ is unclear. For example, my newly planted tree may block the view from your land. But did you have a right to look across my land? If the rights to views (or light rays) were clearly defined and assigned, we could negotiate a price for preserving the view or my putting up a tree, depending upon which was more valuable to the both of us and with payment going to whoever proved to have the rights. Or, while sailing on a lake, to escape a sudden storm and save my boat and life, I use your dock without your prior permission. Did I violate any of your rights, or did your rights not include the right to exclude users in my predicament? If such emergency action is deemed appropriate, then rights to use of the dock are not all yours, as you may have thought. Whereas in the

tree and view case, where a prior negotiation might have avoided a ‘tort’ (except that initially we did not agree about who had what rights), in the emergency use of the dock, prior negotiation was unfeasible. If prior negotiation is uneconomic, rights to that emergency use ‘should’ and will exist if that use is the most valuable use of the resource under the postulated circumstances. And compensation may or may not be required to the erstwhile ‘owner’. The principle underlying such a legal principle seems straightforward and consistent with principles of efficient economic behaviour. It suffices for present purposes merely to call attention to this aspect of economic efficiency underlying the law.

See Also

- clubs
- Coase theorem
- common property resources
- fisheries
- law, economic analysis of
- property law, economics and
- taking (eminent domain)

How to cite this article

Alchian, Armen A. "property rights." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000226> doi:10.1057/9780230226203.1353

The New Palgrave Dictionary of Economics Online

property taxation

George R. Zodrow

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Property taxation of both residential and non-residential land and structures is the most common form of wealth taxation worldwide, and is often the revenue instrument of choice for local governments. Despite widespread use of the property tax and a voluminous academic literature examining the tax, its incidence and economic effects are still contentious issues, with the debate centring around whether the capital portion of the tax should be viewed as distorting the allocation of capital or as an efficient benefit tax or user charge for local public services.

Keywords

assessment process; benefit tax; capital tax; capitalization; efficient allocation; equity; excise tax; fiscal differential; fiscal equalization; George, H.; general equilibrium; Harberger, A.C.; Hamilton, B.W.; head tax; housing market; interjurisdictional competition; land tax; land use zoning; local government; local public finance; Mieszkowski, P.; partial equilibrium; profits tax; progressive and regressive taxation; property taxation; property tax limitations; returns to capital; tax incidence; Tiebout, C.; Tiebout model; underprovision of public services; user fees

Article

Property taxation of both residential and non-residential land and structures – or ‘real property’ – is the most common form of wealth taxation worldwide, and is often the revenue instrument of choice for local governments. For example, in the United States property taxes account for over 70 per cent of local own-source tax revenues. Property tax liability is calculated as the product of the statutory rate and the assessed tax base, subject to a variety of adjustments, such as partial exemptions for primary residences and ‘circuit breakers’ designed to reduce tax burdens for certain groups, especially relatively poor elderly homeowners. Although vagaries in the assessment process have long been a source of inequity in the administration of the tax, recent advances in computer-based assessment practices have mitigated

this problem. More recently, rapid growth in residential home values and the concomitant increase in property tax burdens and the share of total local taxes paid by homeowners have led to increasing dissatisfaction with the tax in some regions, culminating in the passage of numerous property tax limitation measures. In addition, concerns about the equity implications of financing primary and secondary public education with the property tax when the tax base, including non-residential property, is unequally distributed across school jurisdictions have led to reduced reliance on the tax as well as equalization mechanisms that redistribute property tax revenues across school districts (Anderson, 1994).

The incidence debate – the three views of the property tax

The academic literature on the property tax – as reviewed by Mieszkowski and Zodrow (1989), Ladd (1998), Ross and Yinger (2000) and Netzer (2001) – has focused on the incidence and effects on the allocation of resources of the residential property tax. There is general agreement that the land component of the property tax is capitalized into land values, is borne by landowners at the time of the imposition of the tax, and – since land is immobile – does not distort the allocation of resources. Indeed, the efficiency advantages of taxing land values, coupled with a belief that most increases in land values reflect the benefits of public services, have led some observers, most prominently Henry George (1879), to advocate replacing property taxes with taxes on land values.

By comparison, the incidence and economic effects of property taxation of the capital or structures component of real property are among the most contentious issues in state and local public finance. Three views dominate the debate. The ‘traditional’ view dates back to Simon (1943) and Netzer (1966), who focused on the partial equilibrium effects of increasing the tax in a local housing market. From this perspective, one can make the standard ‘open economy’ assumption that the national return to capital is fixed. This in turn implies that local capital bears none of the local property tax, as capital in the long run migrates from the jurisdiction until the local after-tax return to capital equals the national value. The burden of the tax is thus borne by local factors and/or consumers, with the traditional view holding that the entire burden is borne by local housing consumers. The traditional view thus implies that the property tax inefficiently reduces the size of the local housing stock and that its burden is borne in proportion to housing consumption – and is thus somewhat regressive with respect to annual income but, more importantly, roughly proportional with respect to lifetime income.

A second popular theory is the ‘benefit tax’ view, developed by Hamilton (1975; 1976); Fischel (2001a; 2001b) provides a recent discussion. This view is an extension of the renowned Tiebout (1956) model, which argues that consumer mobility (‘voting with the feet’) and inter-jurisdictional competition in the provision of local public services can ensure efficiency of resource allocation in the local public sector. Although Tiebout assumed the existence of benefit/head taxes, Hamilton extended the analysis by deriving conditions under which the property tax can be converted into the head tax assumed by Tiebout. Specifically, following Tiebout, Hamilton assumes that individuals sort into local jurisdictions according to their demands for local public services, and that there are enough local tax-expenditure packages to accommodate all tastes. In addition, Hamilton (1975) assumes that local jurisdictions are homogeneous with respect to house values, and that there are enough jurisdictions to accommodate all desired housing and government service/tax packages. Finally, Hamilton assumes the existence of binding zoning constraints that established a minimum house value for each community. Under these circumstances,

individuals are precluded from purchasing homes with a value below the minimum, and would never elect to pay taxes in excess of benefits received by purchasing a home with a value greater than the minimum; all individuals in a given community thus pay exactly the same property tax, which becomes a benefit tax.

Hamilton (1976) extends this model to the more realistic case in which house values within a community are heterogeneous. In this case, Hamilton assumes all communities are fully developed, effectively precluding any tax-induced changes in the housing stock, and that alternative communities which are homogeneous with respect to both demands for public services and housing are available. Under these circumstances, Hamilton shows that ‘perfect capitalization’ converts the property tax into a benefit tax, as a relatively expensive home sells at a discount equal to its ‘fiscal differential’ or the present value of all future taxes in excess of benefits received, while a relatively inexpensive home sells at a premium reflecting its fiscal differential, the present value of all future benefits in excess of future taxes. The implications of the benefit tax view are striking, as it implies that the property tax is effectively a non-distortionary user charge that has no direct effects on income distribution.

Finally, the ‘capital tax’ view (or ‘new’ view) of the property tax, developed by Mieszkowski (1972), subsequently extended by Zodrow and Mieszkowski (1983; 1986b), and reviewed in Zodrow (2001a; 2001b), argues that the property tax is a distortionary tax on the local use of capital that results in a misallocation of the national capital stock across local jurisdictions. Mieszkowski (1972) stresses that earlier partial equilibrium analyses ignored the fact that the property tax is used by virtually all local jurisdictions and applies to a large fraction of the capital stock, including most non-residential capital. Adapting the Harberger (1962) general equilibrium model of tax incidence, he models the economy as having a fixed national capital stock and two types of local jurisdictions, characterized by ‘high’ tax rates and ‘low’ tax rates. In this context, Mieszkowski shows that property tax rates that exceed the national average drive capital out of high-tax jurisdictions into low-tax jurisdictions, with opposing effects occurring in relatively low tax jurisdictions. Property tax differentials thus result in an inefficient allocation of capital across jurisdictions. In addition, concern about the extent to which use of the property tax may drive capital out of a jurisdiction creates a tendency for local governments to under-provide public services (Zodrow and Mieszkowski, 1986b; Wilson, 1986). In terms of incidence, the ‘average burden’ of all of the property taxes imposed across the nation – known as the ‘profits tax’ effect of the tax – is borne by capital owners generally, implying that the tax is relatively progressive (with respect to annual income). In addition, Mieszkowski stresses that property tax differentials about the national average result in ‘excise tax effects’ in the form of housing and commodity price increases and wage and land price declines in relatively high-tax jurisdictions, with offsetting effects in relatively low-tax jurisdictions.

Differentiating among the three views

Much of the subsequent literature has focused on reconciling or differentiating among these three views, and the issue of which view most accurately describes the effects of the property tax is still contentious. Matters are simplified somewhat because the traditional view has been shown to be a special case of the capital tax view. Specifically, the traditional view can be interpreted as a partial equilibrium analysis that focuses exclusively on the excise tax effects of the capital tax view, while neglecting its general

equilibrium profits tax effects. Moreover, the traditional view that these excise tax effects are fully reflected in higher housing prices is accurate only under special circumstances; more generally, excise tax effects are borne in some combination by housing consumers and the owners of labour and land in the taxing jurisdiction (Wildasin, 1986). In addition, the profits tax effect still obtains when one takes a general equilibrium perspective of the use of the property tax by a single small jurisdiction facing a perfectly elastic supply of capital. Specifically, although the capital outflow caused by an increase in the property tax by a small local jurisdiction depresses the overall return to capital only very slightly, this reduction affects a large capital stock; under certain circumstances, the overall reduction in national capital income precisely equals the revenue raised by the taxing jurisdiction, yielding the profits tax result (Zodrow and Mieszkowski, 1983; Brown, 1924; Bradford, 1978). At the same time, the burden of a property tax increase in a single jurisdiction is borne entirely by local residents as higher prices or lower factor returns (with offsetting effects in all other jurisdictions). A critical implication is that even under the capital tax view there is a close link between local public services and the burden of the property tax, as the cost of financing local expenditures largely falls on local factor owners and consumers; thus, this interpretation of the capital tax view clearly has a strong benefit view flavour as local residents ‘pay for what they get’ in public services.

Nevertheless, the debate between proponents of the benefit tax view and the capital tax view is still far from resolved. The original Mieszkowski (1972) derivation, based on a model of national tax incidence, has been criticized for ignoring many of the features of local government service provision stressed by Tiebout and Hamilton. However, Zodrow and Mieszkowski (1986a) present an expanded derivation of the capital tax view that includes most of these aspects, including interjurisdictional competition, varying tastes for local public services, individuals sorting into differing communities according to tastes for local public services, and a simple form of land use zoning. They conclude that these factors thus do not distinguish between the two views; instead, the key factor in determining the incidence of the property tax is whether housing consumption can vary in response to the imposition of the tax. Moreover, although zoning requirements are pervasive, take a wide variety of forms, and can have a significant impact on property values (Fischel, 1992), these facts do not demonstrate that zoning ordinances are sufficiently binding on housing consumption choices to ensure the validity of the benefit view (Rubinfeld, 1987; Ross and Yinger, 2000). In addition, although empirical evidence suggests that intrajurisdictional and intrajurisdictional capitalization of differences in property taxes and local expenditures is widespread (Oates, 1969; Yinger et al., 1988; and Fischel, 2001a; 2001b, who concludes that evidence of ‘capitalization is everywhere’), capitalization is consistent with both the assumption of fixed housing stocks that underlies the benefit tax view and the tax-induced reallocations of capital that underlie the capital tax view (Zodrow, 2006); moreover, some observers have argued that in the long run capitalization is inconsistent with the benefit view (Ross and Yinger, 2000). Finally, although some recent empirical tests are consistent with the capital tax view (Carroll and Yinger, 1994; Wassmer, 1993), these results are quite tentative. The primary empirical issue remaining to be resolved is whether the zoning restrictions or other mechanisms stressed by proponents of the benefit tax view are sufficiently binding to preclude the long-run adjustments in housing capital predicted by the capital tax view. This question promises to be a fertile topic for future research, which hopefully will help clarify the answer to the long-standing and critical questions of the incidence and economic effects of the property tax.

See Also

- tax incidence
- taxation of wealth

Bibliography

Anderson, J.E., ed. 1994. *Fiscal Equalization for State and Local Government Finance*. Westport, CT: Praeger.

Bradford, D.F. 1978. Factor prices may be constant but factor returns are not. *Economics Letters* 1, 199–203.

Brown, H.G. 1924. *The Economics of Taxation*. New York: Holt.

Carroll, R. and Yinger, J. 1994. Is the property tax a benefit tax? The case of rental housing. *National Tax Journal* 47, 295–316.

Fischel, W.A. 1992. Property taxation and the Tiebout model: evidence for the benefit view from voting and zoning. *Journal of Economic Literature* 30, 171–7.

Fischel, W.A. 2001a. Municipal corporations, homeowners, and the benefit view of the property tax. In *Property Taxation and Local Government Finance*, ed. W.E. Oates. Cambridge, MA: Lincoln Institute of Land Policy.

Fischel, W.A. 2001b. Homevoters, municipal corporate governance, and the benefit view of the property tax. *National Tax Journal* 54, 157–73.

George, H. 1879. *Progress and Poverty*. New York: Dutton.

Hamilton, B.W. 1975. Zoning and property taxation in a system of local governments. *Urban Studies* 12, 205–11.

Hamilton, B.W. 1976. Capitalization of intrajurisdictional differences in local tax prices. *American Economic Review* 66, 743–53.

Harberger, A.C. 1962. The incidence of the corporate income tax. *Journal of Political Economy* 70, 215–40.

Ladd, H.F. 1998. *Local Government Tax and Land Use Policies in the United States*. Northampton, MA:

Edward Elgar Publishing.

Mieszkowski, P. 1972. The property tax: an excise tax or a profits tax? *Journal of Public Economics* 1, 73–96.

Mieszkowski, P. and Zodrow, G.R. 1989. Taxation and the Tiebout model: the differential effects of head taxes, taxes on land rents, and property taxes. *Journal of Economic Literature* 27, 1098–146.

Netzer, D. 1966. *Economics of the Property Tax*. Washington, DC: Brookings Institution Press.

Netzer, D. 2001. Local property taxation in theory and practice: some reflections. In *Property Taxation and Local Government Finance*, ed. W.E. Oates. Cambridge, MA: Lincoln Institute of Land Policy.

Oates, W.E. 1969. The effects of property taxes and local public spending on property values: an empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy* 77, 957–61.

Ross, S. and Yinger, J. 2000. Sorting and voting: a review of the literature on urban public finance. In *Handbook of Regional and Urban Economics*, vol. 3, ed. P. Cheshire and E.S. Mills. Amsterdam: North-Holland.

Rubinfeld, D.L. 1987. The economics of the local public sector. In *Handbook of Public Economics*, ed. A.J. Auerbach and M.S. Feldstein. Amsterdam: North-Holland.

Simon, H.A. 1943. The incidence of a tax on urban real property. *Quarterly Journal of Economics* 59, 398–420.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Wassmer, R.W. 1993. Property taxation, property base, and property value: an empirical test of the ‘new view’. *National Tax Journal* 46, 135–60.

Wildasin, D.E. 1986. *Urban Public Finance*. New York: Harwood.

Wilson, J.D. 1986. A theory of inter-regional tax competition. *Journal of Urban Economics* 19, 296–315.

Yinger, J., Bloom, H.S., Boersch-Supan, A. and Ladd, H.F. 1988. *Property Taxes and House Values: The Theory and Estimation of Intrajurisdictional Property Tax Capitalization*. San Diego, CA: Academic Press.

Zodrow, G.R. 2001a. Reflections on the new view and the benefit view of the property tax. In *Property*

Taxation and Local Government Finance, ed. W.E. Oates. Cambridge, MA: Lincoln Institute of Land Policy.

Zodrow, G.R. 2001b. The property tax as a capital tax: a room with three views. *National Tax Journal* 54, 139–56.

Zodrow, G.R. 2006. The property tax incidence debate and the mix of state and local finance of local public expenditures. Working paper, Rice University.

Zodrow, G.R. and Mieszkowski, P. 1983. The incidence of the property tax: the benefit view vs. the new view. In *Local Provision of Public Services: The Tiebout Model after Twenty-five Years*, ed. G.R. Zodrow. New York: Academic Press.

Zodrow, G.R. and Mieszkowski, P. 1986a. The new view of the property tax: a reformulation. *Regional Science and Urban Economics* 16, 309–27.

Zodrow, G.R. and Mieszkowski, P. 1986b. Pigou, Tiebout, property taxation and the under-provision of local public goods. *Journal of Urban Economics* 19, 356–70.

How to cite this article

Zodrow, George R. "property taxation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000228> doi:10.1057/9780230226203.1354

The New Palgrave Dictionary of Economics Online

proportional hazard model

Jerry A. Hausman and Tiemen M. Woutersen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews proportional hazard models and how the thinking about identification and estimation of these models has evolved since the mid-1970s.

Keywords

baseline hazard; convergence; duration dependence; duration measurements; duration models; gamma distribution; heterogeneity; mixed proportional hazard (MPH) model; partial likelihood estimators; proportional hazard models; semiparametric estimation; time-varying regressors; Weibull model

Article

The estimation of duration models has been the subject of significant research in econometrics since the late 1970s. Cox (1972) proposed the use of proportional hazard models in biostatistics and they were soon adopted for use in economics. Since Lancaster (1979), it has been recognized among economists that it is important to account for unobserved heterogeneity in models for duration data. Failure to account for unobserved heterogeneity causes the estimated hazard rate to decrease more with the duration than the hazard rate of a randomly selected member of the population. Moreover, the estimated proportional effect of explanatory variables on the population hazard rate is smaller in absolute value than that on the hazard rate of the average population member and decreases with the duration. To account for unobserved heterogeneity Lancaster proposed a parametric mixed proportional hazard (MPH) model, a partial generalization of Cox's proportional hazard model, that specifies the hazard rate as the product of a regression function that captures the effect of observed explanatory variables, a baseline hazard that captures variation in the hazard over the spell, and a random variable that accounts for the omitted heterogeneity. In particular, Lancaster (1979) introduced the mixed proportional hazard model in which the hazard is a function of a regressor X unobserved heterogeneity v, and a function of time $\lambda(t)$,

$$\theta(t|X, v) = ve^{X\beta_0}\lambda(t). \quad (1)$$

The function $\lambda(t)$ is often referred to as the baseline hazard and $v|X$ has a gamma distribution. The popularity of the mixed proportional hazard model is partly due to the fact that it nests two alternative explanations for the hazard $\theta(t|X)$ to be decreasing with time. In particular, estimating the mixed proportional hazard model gives the relative importance of the heterogeneity, *v*, and genuine duration dependence, $\lambda(t)$ (see Lancaster, 1990, and Van den Berg, 2001, for overviews). Lancaster (1979) uses functional form assumptions on $\lambda(t)$, which were not required by the Cox model, and distributional assumptions on *v* to identify the model. Examples by Lancaster and Nickell (1980) and Heckman and Singer (1984), however, show the sensitivity to these functional form and distributional assumptions. Thus, Lancaster's MPH model is fully parametric and from the outset questions were raised about the role of functional form and parametric assumptions in the distinction between unobserved heterogeneity and duration dependence. (Heckman, 1991, gives an overview of attempts to make this distinction in duration and dynamic panel data models.) This question was resolved by Elbers and Ridder (1982), who showed that the MPH model is semi-parametrically identified if there is minimal variation in the regression function. A single indicator variable in the regression function suffices to recover the regression function, the baseline hazard, and the distribution of the unobserved component, provided that this distribution does not depend on the explanatory variables. Semi-parametric identification means that semi-parametric estimation is feasible, and a number of semi-parametric estimators for the MPH model have been proposed that progressively relaxed the parametric restrictions.

Nielsen et al. (1992) showed that the partial likelihood estimator of Cox (1972) can be generalized to the MPH model with gamma-distributed unobserved heterogeneity. Their estimator is semi-parametric because it uses parametric specifications of the regression function and the distribution of the unobserved heterogeneity. The estimator requires numerical integration of the order of the sample size, as originally discussed by Han and Hausman (1990), which further limits its usefulness and makes it impractical for most situations in econometrics. Heckman and Singer (1984) considered the non-parametric maximum likelihood estimator of the MPH model with a parametric baseline hazard and regression function. Using results of Kiefer and Wolfowitz (1956), they approximate the unobserved heterogeneity with a discrete mixture. The rate of convergence and the asymptotic distribution of this estimator are not known. As a result, these estimators that use discrete mixture with an increasing number of support points cannot be used to test hypotheses. Another estimator that does not require the specification of the unobserved heterogeneity distribution was suggested by Honoré (1990). This estimator assumes a Weibull baseline hazard and uses only very short durations to estimate the Weibull parameter.

Han and Hausman (1990) and Meyer (1990) propose an estimator that assumes that the baseline hazard is piecewise-constant, to permit flexibility, and that the heterogeneity has a gamma distribution. Both papers find that the hazard rate, conditional on heterogeneity, is non-monotonic so that the Weibull

model cannot hold. Hausman and Woutersen (2005) present simulations and a theoretical result that show that using a nonparametric estimator of the baseline hazard with gamma heterogeneity yields inconsistent estimates for all parameters and functions if the true mixing distribution is not a gamma, which limits the usefulness of the Han–Hausman–Meyer approach. Thus, Hausman and Woutersen (2005) find it important to specify a model that does not require a parametric specification of the unobserved heterogeneity.

Horowitz (1999) was the first to propose an estimator that estimates both the baseline hazard and the distribution of the unobserved heterogeneity nonparametrically. His estimator is an adaptation of the semi-parametric estimator for a transformation model that he introduced in Horowitz (1996). In particular, if the regressors are constant over the duration, then the MPH model has a transformation model representation with the logarithm of the integrated baseline hazard as the dependent variable and a random error that is equal to the logarithm of a log standard exponential minus the logarithm of a positive random variable. In the transformation model the regression coefficients are identified only up to scale. As shown by Ridder (1990), the scale parameter is identified in the MPH model if the unobserved heterogeneity has a finite mean. Horowitz (1999) suggests an estimator of the scale parameter that is similar to Honoré's (1990) estimator of the Weibull parameter and is consistent if the finite mean assumption holds so that his approach allows estimation of the regression coefficients (not just up to scale). However, the Horowitz approach permits estimation of the regression coefficients only at a slow rate of convergence and it is not $N^{-1/2}$ consistent, where *N* is the sample size. The reason for the slower than $N^{-1/2}$ convergence is that the information matrix of the MPH model is singular under Horowitz assumptions (see Hahn, 1994; Ishwaran, 1996a). In particular, Horowitz (1999) assumes that the first three moments of the heterogeneity distribution exist, and Ishwaran (1996b) shows that the fastest possible rate of convergence is $N^{-2/5}$ for that case and Horowitz's (1999) estimator converges arbitrarily close to that rate. In other words, the slow rate of convergence is implied by the assumptions and is not a peculiarity of the estimator.

Subsequent research has focused on strengthening the assumptions of the MPH model so that $N^{-1/2}$ convergence is possible. Ridder and Woutersen (2003) derive a $N^{-1/2}$ consistent estimator for the MPH model by assuming that the baseline hazard rate is constant over a small interval, $\lambda(t) = \lambda$ for $0 \leq t \leq \varepsilon$ for any $\varepsilon > 0$ while allowing for a nonparametric baseline hazard function for $t > \varepsilon$. For parametric baseline hazards, Ridder and Woutersen (2003) assume that $\lim_{t \downarrow 0} \lambda(t) = \lambda$ for $0 < \lambda < \infty$ and derive another $N^{-1/2}$ consistent estimator. Hausman and Woutersen (2005) derive an estimator for the mixed proportional hazard model (with heterogeneity) that allows for a nonparametric baseline hazard and uses time-varying regressors. No parametric specification of the heterogeneity distribution or nonparametric estimation of the heterogeneity distribution is necessary. Intuitively, Hausman and Woutersen (2005) condition out the heterogeneity distribution, which makes it unnecessary to estimate it. Thus, they eliminate the problems that arise with the Lancaster (1979) approach to MPH models. In this model the baseline hazard rate is nonparametric, and the estimator of the integrated baseline hazard rate converges at the regular rate, $N^{-1/2}$ where *N* is the sample size. This convergence rate is the same rate as for a duration model without heterogeneity. The regressor parameters also converge at the regular

rate. A nice feature of the estimator is that it allows the durations to be measured on a finite set of points. Such discrete measurement of durations is important in economics; for example, unemployment is often measured in weeks. In the case of discrete duration measurements, the estimator of the integrated baseline hazard converges only at this set of points, as would be expected.

It may be argued that the bias in the estimates of the regression coefficients is small if the estimates of the MPH model indicate that there is no significant unobserved heterogeneity. The problem with this argument is that estimates of the heterogeneity distribution are usually not very accurate. Given the results in Horowitz (1999), this finding should not come as a surprise. The simulation results in Baker and Melino (2000) show that it is empirically difficult to find evidence of unobserved heterogeneity, in particular if one chooses a flexible parametric representation of the baseline hazard. However, Han and Hausman (1990) and applications of their approach have found significant heterogeneity using a flexible approach to the baseline hazard. Bijwaard and Ridder (2002) find that the bias in the regression parameters is largely independent of the specification of the baseline hazard. Hence, failure to find significant unobserved heterogeneity should not lead to the conclusion that the bias due to correlation of the regressors and the unobservables that affect the hazard is small.

Because it is empirically difficult to recover the distribution of the unobserved heterogeneity, estimators that rely on estimation of this distribution may be unreliable. Therefore, it may be advisable to avoid estimating the unobserved heterogeneity distribution and the remainder of the MPH model simultaneously. Nevertheless, after estimating the baseline hazard and regression function, one can usually identify the mixing distribution. In particular, Horowitz (1999) uses the following equation to estimate the mixing distribution,

$$\ln\{\Lambda(T)\} + X\beta - \ln(Z) = -\ln(v)$$

where $\Lambda(T)$ and β can be estimated and the unobserved Z has an exponential distribution with mean one. Thus, Horowitz (1999) solves a deconvolution problem and the speed of convergence depends on the assumptions on the distribution of v.

A hazard model is a natural framework for time-varying regressors if a flow or a transition probability depends on a regressor that changes with time since a hazard model avoids the curse of dimensionality that would arise from interacting the regressors at each point in time with one another. A non-constructive identification proof for the duration model with time-varying regressors can be produced using techniques similar to Honoré (1993b), and Honoré (1993a) gives such a proof. (A non-constructive identification proof is an identification proof that does not suggest an estimator.) In particular, Honoré (1993a) does not assume that the mean of the heterogeneity distribution is finite (nor does Honoré, 1993a, assume a tail condition as in Heckman and Singer, 1984). Ridder and Woutersen (2003) argue that it is precisely the finite mean assumption that makes the identification of Elbers and Ridder (1982) ‘weak’ in the sense that the model of Elbers and Ridder (1982) cannot be estimated at rate $N^{-1/2}$. As in Honoré (1993a), Hausman and Woutersen (2005) do not need the finite mean $N^{-1/2}$

assumption which gives an intuitive explanation of why Hausman and Woutersen (2005) can estimate the model at rate $N^{-1/2}$.

Bibliography

Baker, M. and Melino, A. 2000. Duration dependence and nonparametric heterogeneity: a Monte Carlo study. *Journal of Econometrics* 96, 357–93.

Bijwaard, G. and Ridder, G. 2002. Efficient estimation of the semi-parametric mixed proportional hazard model. Working paper.

Cox, D. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* B, 34, 187–220.

Elbers, C. and Ridder, G. 1982. True and spurious duration dependence: the identifiability of the proportional hazard model. *Review of Economic Studies* 49, 402–09.

Hahn, J. 1994. The efficiency bound of the mixed proportional hazard model. *Review of Economic Studies* 61, 607–29.

Han, A. and Hausman, J. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* 6, 1–28.

Hausman, J. and Woutersen, T. M. 2005. Estimating a semi-parametric duration model without specifying heterogeneity. Working paper, UCL: CeMMAP, Institute for Fiscal Studies.

Heckman, J. 1991. Identifying the hand of the past: distinguishing state dependence from heterogeneity. *American Economic Review* 81, 75–9.

Heckman, J. and Singer, B. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52, 271–320.

Honoré, B. 1990. Simple estimation of a duration model with unobserved heterogeneity. *Econometrica* 58, 453–73.

Honoré, B. 1993a. Identification results for duration models with multiple spells or time-varying regressors. Northwestern working paper.

Honoré, B. 1993b. Identification results for duration models with multiple spells. *Review of Economic Studies* 60, 241–6.

Horowitz, J. 1996. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* 64, 103–7.

Horowitz, J. 1999. Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* 67, 1001–28.

Ishwaran, H. 1996a. Identifiability and rates of estimation for scale parameters in location mixture models. *Annals of Statistics* 24, 1560–71.

Ishwaran, H. 1996b. Uniform rates of estimation in the semiparametric Weibull mixture model. *Annals of Statistics* 24, 1572–85.

Kiefer, J. and Wolfowitz, J. 1956. Consistency of maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* 27, 887–906.

Lancaster, T. 1979. Econometric methods for the duration of unemployment. *Econometrica* 47, 939–56.

Lancaster, T. 1990. *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

Lancaster, T. and Nickell, S. 1980. The analysis of re-employment probabilities for the unemployed. *Journal of the Royal Statistical Society, A*, 143, 141–65.

Meyer, B. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58, 757–82.

Nielsen, G., Gill, R., Andersen, P. and Sørensen, T. 1992. A counting process approach to maximum likelihood estimation in frailty models. *Scandanavian Journal of Statistics* 19, 25–43.

Ridder, G. 1990. The non-parametric identification of generalized accelerated failure time models. *Review of Economic Studies* 57, 167–82.

Ridder, G. and Woutersen, T. 2003. The singularity of the information matrix of the mixed proportional hazard model. *Econometrica* 71, 1579–89.

Van den Berg, G. 2001. Duration models: specification, identification, and multiple duration. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

How to cite this article

Hausman, Jerry A. and Tiemen M. Woutersen. "proportional hazard model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave

Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000334>
doi:10.1057/9780230226203.1355

The New Palgrave Dictionary of Economics Online

prospect theory

Graham Loomes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Prospect theory sought to provide a descriptive model of risky choice which could accommodate a number of seemingly systematic violations of conventional ‘expected utility’ analysis. Although there are phenomena which the model cannot explain (even in its later ‘cumulative’ form), it constitutes a landmark in the development of alternative theories which have modified standard theory and/or have tried to incorporate psychological factors into decision theories.

Keywords

Allais paradox; behavioural economics; choice; common consequence effect; common ratio effect; cumulative prospect theory; decision weights; expected utility hypothesis; isolation effect; Kahneman, D.; probability distribution; prospect theory; rational behaviour; reference point; reflection effect; stochastic dominance; subjective probability; transitivity; Tversky, A.

Article

Prospect theory (PT) was developed by psychologists Daniel Kahneman and Amos Tversky to try to account for a number of patterns of response to risky choices which departed systematically from the conventional wisdom about rational decision making in the form of von Neumann and Morgenstern's (1944) expected utility (EU) hypothesis.
Kahneman and Tversky's (1979) paper ‘Prospect Theory: An Analysis of Decision Under Risk’ has proved to be enormously influential. According to Kim, Morse and Zingales (2006), it is the second most frequently cited paper published in economics journals since 1970, with more than 4,000 citations in the 25 years since its publication. It provided a major stimulus to the development of a number of other ‘non-expected-utility’ theories in the 1980s and 1990s – see Starmer (2000) for a survey and review. It has also inspired much work in behavioural economics and in economic and psychological experiments exploring individual decision making under risk and uncertainty.
The following subsections consider what PT set out to do and how it did it. There then follows a discussion of the importance of the theory as well as its possible limitations.

Background

In the 1950s and 1960s, evidence had begun to accumulate which suggested that EU failed as a general *descriptive* model of risky choice. Two of the most influential ‘paradoxes’ had been identified by Maurice Allais in the early 1950s (see Allais, 1953). These were renamed by Kahneman and Tversky (henceforth K&T) and are now widely known as the ‘common ratio effect’ and the ‘common consequence effect’. Briefly, they are as follows, starting with the common ratio effect.
Consider the choice between two ‘prospects’ A and B where A offers a sum of money x with probability p (and 0 with probability $1-p$) while B offers a smaller sum y with a larger probability q (and 0 with probability $1-q$). An extreme form of this might involve setting $q=1$, so that B offers the certainty of y: in an example used by K&T, B offered the certainty of 3,000 Israeli pounds while A offered a 0.8 chance of 4,000 (and a 0.2 chance of 0). EU does not predict which of A and B an individual will choose – that depends on the individual's personal tastes concerning risk – but what the independence axiom of EU *does* entail is that, if p and q are scaled down by the same factor so that the ratio of ‘winning’

probabilities is maintained, the preference between the scaled-down prospects will be consistent with the preference between A and B.

So – to continue the example used by K&T – suppose that both *p* and *q* are scaled down to a quarter of their original values, generating prospects C and D, where C offers a 0.2 chance of 4,000 and a 0.8 chance of 0, while D offers a 0.25 chance of 3,000 and a 0.75 chance of 0. Then EU entails that anyone who prefers A over B should also prefer C over D, and vice versa. However, the common ratio effect form of the Allais paradox is manifested when a substantial proportion of those who choose the safer option B in the first case switch to the riskier prospect C in the second case, while the combination of choosing A in the first case and D in the second case is relatively rare.

In the above case, the scaling down operated on the *magnitudes* of the winning probabilities, while *maintaining the ratio* between them. Another way of manipulating the prospects could work in terms of replacing some probability of a particular sum common to both prospects by the same probability of a different sum. Consider another example used by K&T. This time, E offers 2,500 with probability 0.33, 2,400 with probability 0.66 and 0 with probability 0.01, while F offers 2,400 with certainty. Now, for both prospects, replace the 0.66 probability of 2,400 by a 0.66 probability of 0: this transforms E into a prospect G which offers a 0.33 chance of 2,500 and a 0.67 chance of 0, and transforms F into a prospect H which offers a 0.34 chance of 2,400 and a 0.66 chance of 0. Once again, EU entails that individuals should either choose E in the first case and G in the second, or else they should choose F and H. However, the common consequence effect form of Allais paradox involves many more individuals switching from safer to riskier (that is, choosing F and G) than switch from riskier to safer (that is, choose E and H).

In addition to the common ratio and common consequence effects, two other ‘effects’ were influential in the formulation of PT. One of these is the ‘isolation effect’. Consider again the ‘scaled-down’ pair of prospects from the common ratio example. In the way they were presented there, C offered a 0.2 chance of 4,000 together with a 0.8 chance of 0, while D offered a 0.25 chance of 3,000 alongside a 0.75 chance of 0. In this case, the implication is that the uncertainty is resolved in a single stage: perhaps a 20-sided die is rolled, and if a number from 1 to 4 comes up C pays 4,000 (and 0 otherwise), whereas D pays 3,000 if the number is anything in the range 1 to 5.

However, there is another way of presenting this choice which EU would regard as amounting to exactly the same thing, but which PT suggests people are likely to treat differently. Suppose that the uncertainty is resolved in two stages, as follows. In the first stage, there is a 0.75 chance of being ‘knocked out’ and getting 0, and there is a 0.25 chance of getting through to the second stage – at which point the choice is between, on the one hand, a 0.8 chance of 4,000 and, on the other hand, the certainty of 3,000. The logic of EU entails that the two stages can be ‘reduced’ to a single stage by multiplying through the probabilities: a 0.25 chance of getting through and facing a 0.8 chance of 4,000 can thus be reduced to a 0.2 chance of 4,000, as offered by prospect C; and a 0.25 chance of getting through to receive the certainty of 3,000 is regarded as just the same as a direct 0.25 chance of 3,000, as offered by prospect D.

Put another way, a 0.25 chance of what was prospect A in the common ratio example is equivalent to C, and a 0.25 chance of prospect B is regarded as the same as D. Yet the evidence of what K&T called the ‘isolation effect’ shows that people do not process the two-stage game in the way presumed by EU. When faced with such a two-stage problem and told that they have to make a commitment ahead of the first stage, most individuals appear to disregard (or isolate) the common first stage, focus on the alternatives that are contingent on getting through to the second stage, and then make much the same choices as they do when presented with the simple one-stage choice between A and B. In other words, when asked to commit ahead of this two-stage resolution of uncertainty, there is a much stronger tendency to pick the safer option than when presented with the one-stage choice between C and D where the calculus of probability ‘reduction’ has already been applied.

The fourth regularity that played a significant role in the formulation of PT was the ‘reflection effect’. Essentially, this refers to the observations that changing payoffs from gains to losses (relative to the status quo) tended to reverse individuals’ choices. Thus if A and B above were transformed into A′ and B′ such that A′ offered a 0.8 probability of *losing* 4,000 (and a 0.2 probability of losing nothing) while B′ entailed the certainty of a 3,000 loss, the modal preference for B over A would often be ‘reflected’ into a modal preference for A′ over B′. Thus, what appears as a predominant pattern of risk aversion in the choice between prospects such as A and B which involve gains seems to transform into a predominant pattern of risk seeking when the non-zero payoffs are losses.

Combining the reflection effect with the isolation effect can produce striking ‘framing’ effects. For example, consider first a scenario where an individual is given a lump sum of 1,000 and then asked to choose between a further 500 for sure or else a risky prospect offering a 50–50 chance of either 0 or an extra 1,000. If the individual isolates the initial 1,000 and displays risk aversion towards the 50–50 gamble involving gains, she will end up with a sure 1,500 rather than a portfolio consisting of a 0.5 chance of a net 1,000 and a 0.5 chance of a net 2,000. But now consider a scenario framed somewhat differently. The individual is given a lump sum of 2,000 and then asked to choose between the certainty of losing 500 or else a 50–50 chance of either losing 1,000 or losing 0. If the individual again isolates the lump-sum but now displays risk seeking towards the 50–50 gamble involving losses, she will end up with exactly the opposite portfolio preference: that is, she will choose the portfolio consisting of a 0.5 chance of a net 1,000 and a 0.5 chance of a net 2,000 rather than 1,500 for sure. K&T presented evidence which showed that this was indeed a strong tendency among those who answered their hypothetical questions framed in these various ways.

The aims and structure of prospect theory

PT can be seen as offering a descriptive (rather than a prescriptive/normative) model of a particular area of decision making. K&T were careful to specify the domain to which their model applied: it was a theory of *choice* over pairs of prospects each involving *no more than two non-zero payoffs* where the *objective probabilities were given* to decision makers. As formulated in the 1979 paper, the theory did *not* apply to valuation tasks (for example, tasks that asked people how much they would pay or accept in exchange for some risky

prospect), nor to prospects involving larger numbers of possible payoffs, nor to cases where there was ambiguity about the likelihood of different events occurring (although in their concluding remarks K&T expressed some optimism that the model could be extended to accommodate the latter two features, while relevant valuations might be inferred via some iterative procedure involving a series of choices between a prospect and different sure sums). Most importantly, because it set out to provide an account of *actual* behaviour rather than a prescription for how decision makers *ought* 'rationally' to behave, PT allowed the possibility of patterns of behaviour that decision makers might wish to modify if they ever became aware of the 'inconsistencies' involved (although, in the absence of opportunities for such realization, the 'anomalies' implied by PT could be expected to occur and persist).

To *some* extent, the elimination of some potentially undesirable possible implications of PT was handled by dividing the modelling of people's decision processes into two phases: first, the editing phase, which involved simplifying prospects and screening out transparent transgressions of reasonable behaviour; and then the evaluation phase, in which the preferred alternative was identified.

The editing phase prepared the ground for the evaluation phase in various intuitively appealing ways. It involved the detection of *transparent* dominance and the discarding/rejection of dominated alternatives in such cases (while allowing the possibility that dominance might be violated if more complicated ways of presenting the prospects obscured the dominance relationship). There was also scope for some *simplification* of prospects (for example, *rounding* of payoffs and/or probabilities). In cases where there were transparently common and/or riskless components, these were liable to be *segregated* and/or *cancelled*. It was also supposed that, when a prospect offered the same payoff contingent on different events with separately expressed probabilities, those probabilities would be added together. For example, suppose a prospect offered a payoff of 100 if a card drawn at random from a standard pack of playing cards turned out to be a spade, and offered the same payoff if the card turned out to be a heart: then the probabilities of these two events – each 0.25 – would be *combined* to give an overall 0.5 chance of receiving 100. Finally, all payoffs were *coded* into gains or losses relative to some reference point – this latter normally being the status quo, although in some circumstances it might be otherwise (as discussed in the penultimate subsection of the 1979 paper).

The evaluation phase involved the interaction of two components: the *value* function, and the *decision weight* function.

A careful reading of the 1979 exposition makes it clear that the subjective value associated with a particular payoff should, strictly speaking, be expected to be a function of *two* factors: the asset position that constitutes the individual's reference point, and the positive or negative change from that point represented by the payoff in question. However, K&T argued that, over quite broad ranges of initial asset positions and for many practical purposes, it is sufficient to focus just on one argument, namely, the size of the gain or loss entailed by any particular payoff.

Drawing on existing evidence, including a substantial body of work from the realm of psychophysics, K&T argued that such a value function is characterized by two key characteristics.

First, the marginal value of both gains and losses is presumed to diminish as the magnitudes increase. Thus the difference between a gain (or loss) of 100 and a gain (or loss) of 150 registers more strongly than the difference between a gain (loss) of 1,100 and a gain (loss) of 1,150. Such diminishing sensitivity means that the gradient of the value function becomes progressively less steep as payoffs are located further from the reference point. Denoting the value of any monetary payoff x by $v(x)$, diminishing sensitivity in the domain of gains can be more formally represented as

$$v(x + a) - v(x) > v(x + a + k) - v(x + k) \quad \text{for all } x,\ a,\ k > 0;$$

in the domain of losses, it entails

$$v(-x) - v(-x - a) > v(-x - k) - v(-x - a - k).$$

Second, the marginal value of losses is modelled as being greater than the marginal value of gains of the same magnitude: that is, for all x, the gradient of the function is steeper at $-x$ than at x. More formally, $v'(-x) > v'(x)$ wherever the derivative of x exists. In conjunction with the first characteristic, this implies a value function as shown in Figure 1: that is, concave in the domain of gains, convex and steeper in the domain of losses, and kinked at the (0) reference point.

Figure 1

Prospect theory's value function

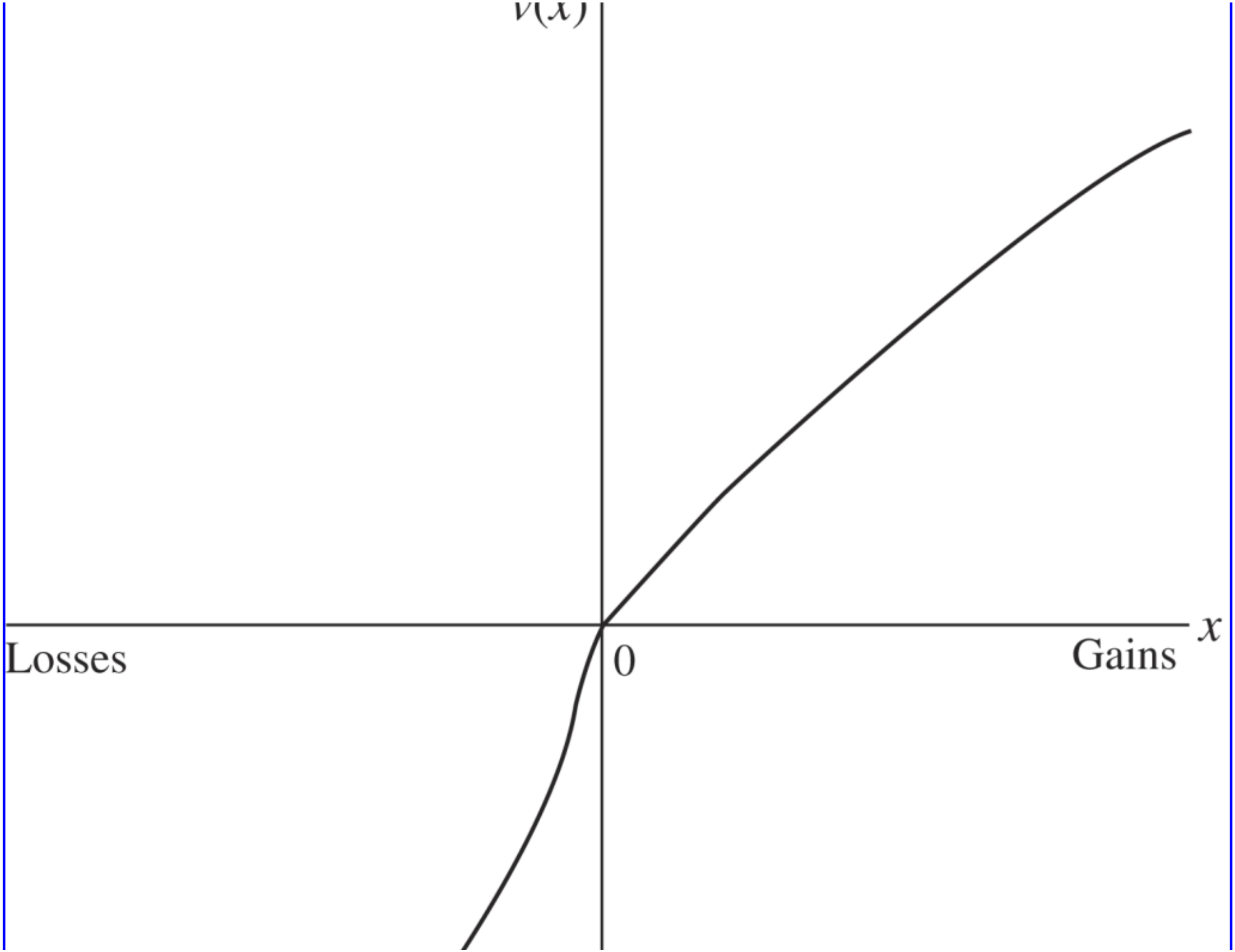
v(x)
x
Losses
0
Gains

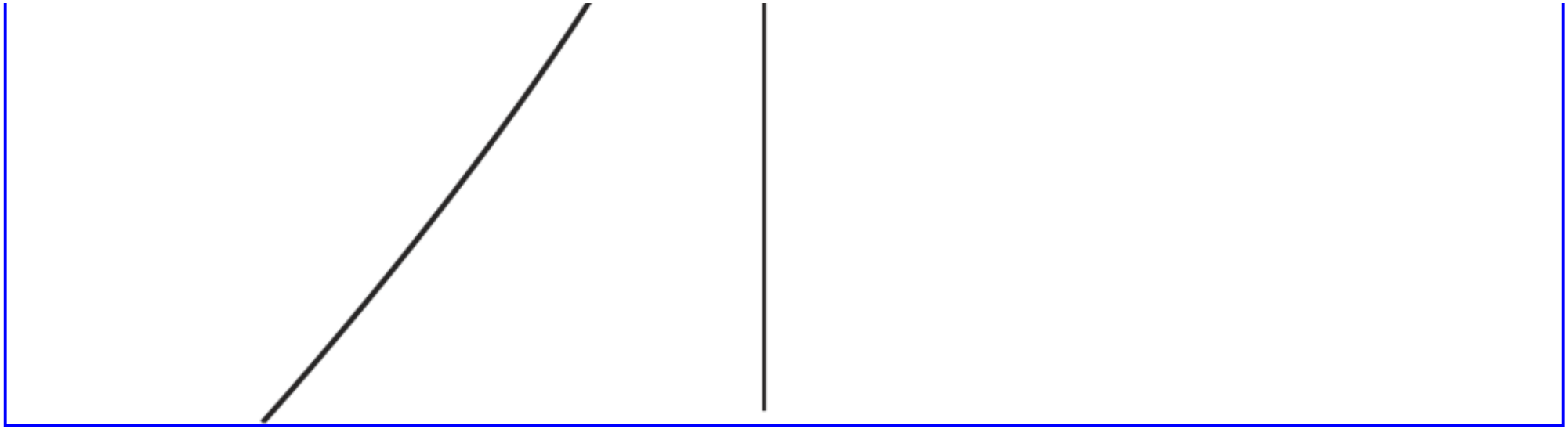

Thus one part of the evaluation of any prospect involves converting money payoffs to their values via the $v(\cdot)$ function as specified above. The other part of the evaluation requires decision weights to be attached to the various values. These decision weights are not the objective probabilities, nor even degrees of belief of the kind that are conventionally supposed to constitute subjective probabilities. In the context of the 1979 exposition, they represent a psychological transformation, or modification, of the objectively given probabilities, with the weighting function being denoted by $\pi(\cdot)$.

The key assumptions about $\pi(\cdot)$ are as follows. First, the weight attached to a zero probability event is 0, and the weight attached to a certainty is 1: that is, $\pi(0)=0$ and $\pi(1)=1$. Second, for low probability events, $\pi(p)>p$; but for higher probability events, $\pi(p)<p$; the 'crossover point', where $\pi(p)=p$, may vary from one individual to another, but is often depicted as being somewhere in the region of $p=0.15$. Third, it is generally supposed that $\pi(p)+\pi(1-p)<1$: this property, labelled *subcertainty*, conveys the idea that complementary intermediate probabilities are jointly disadvantaged relative to certainty.

Taken together, the above assumptions are consistent with a decision weighting function of the kind depicted in Figure 2. Over most of its range, the fact that $\pi(\cdot)$ is flatter than the 45° line suggests that the evaluation of a prospect is less sensitive to changes in the probability of its non-zero payoff(s) than would be the case under EU where the utilities of payoffs are weighted in exact proportion to their respective probabilities of occurring. It also has the implication that for any given ratio of probabilities, the ratio tends to get closer to 1 as the magnitudes of the probabilities fall: more formally, $\pi(pq)/\pi(p) \leq \pi(pqr)/\pi(pr)$ for all p, q, $r<1$.

Figure 2

Prospect theory's decision weighting function

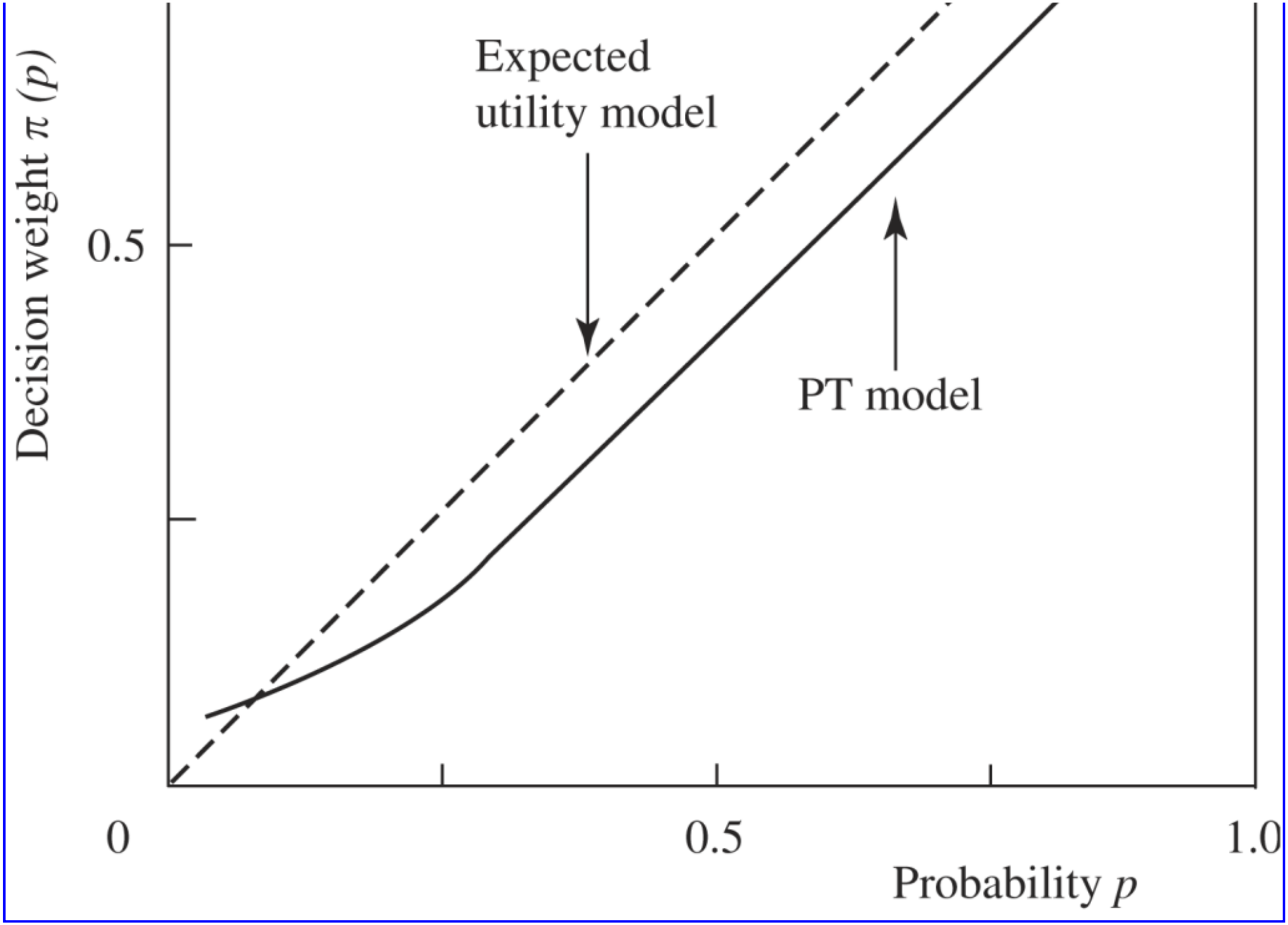

However, such a formulation also has the property that it entails abrupt changes/jumps in the vicinities of p=0 and p=1. It might be said that the function is not 'well-behaved' – or indeed, not defined – in those regions, where there are 'quantal effects'. And this allows at least one pattern of behaviour that many decision theorists would find normatively undesirable/unacceptable, as follows. Consider a case where prospect C offers the certainty of some gain x, while A offers x with probability p and $x+a$ with probability $1 - p$, where a

is some (small) positive amount. Evaluating each prospect separately,

$$v(\mathrm{C}) = v(x) \text{ and } v(\mathrm{A}) = \pi(p)v(x) + \pi(1-p)v(x+a).$$

However, because $\pi(p)+\pi(1-p)$ is liable to sum to less than 1, v(A) may be less than v(C), even though A dominates C. In a direct choice between the two, PT supposes that this dominance (if transparent) will be detected as part of the editing process, so that A will be chosen. But it should be possible to construct some other prospect B which neither dominates nor is dominated by either A or C but whose value lies between the two, so that $v(\mathrm{C})>v(\mathrm{B})>v(\mathrm{A})$. Hence in separate pairwise choices, C will be preferred to B and B will be preferred to A, while A will be chosen over C on the basis of transparent dominance, thereby giving a violation of transitivity.

More recent developments in theory and evidence

Because PT *does* allow such violations of principles that many decision theorists regard as normatively compelling, various modifications have been proposed to ‘fix’ this supposed defect: in particular, a method of deriving decision weights which ensured that they summed to 1 and disallowed any violations of stochastic dominance or transitivity was proposed by Quiggin (1982) and was subsequently incorporated into a revised and extended form of PT known as cumulative prospect theory (CPT) (see Tversky and Kahneman, 1992).
The essence of Quiggin's proposal involved ranking the possible outcomes $x_1 \ldots x_n$ offered by a prospect according to their values and then assigning weights to each of the cumulative probabilities that the prospect pays at least x_i, for all i=1 … n. (Hence this kind of model came to be labelled as ‘rank-dependent’.) The function used to transform cumulative probabilities – call it $w(\cdot)$ to distinguish it from the $\pi(\cdot)$ discussed above – is fully defined in [0, 1] space, with $w(0)=0$ and $w(1)=1$. Like $\pi(\cdot)$, it is usually supposed to have an increasing inverse-S shape (although by contrast with $\pi(\cdot)$, the ‘crossover point’ in CPT is more often regarded as lying in the 0.3–0.4 region – see Figure 3.

Figure 3
CPT's cumulative probability transformation function

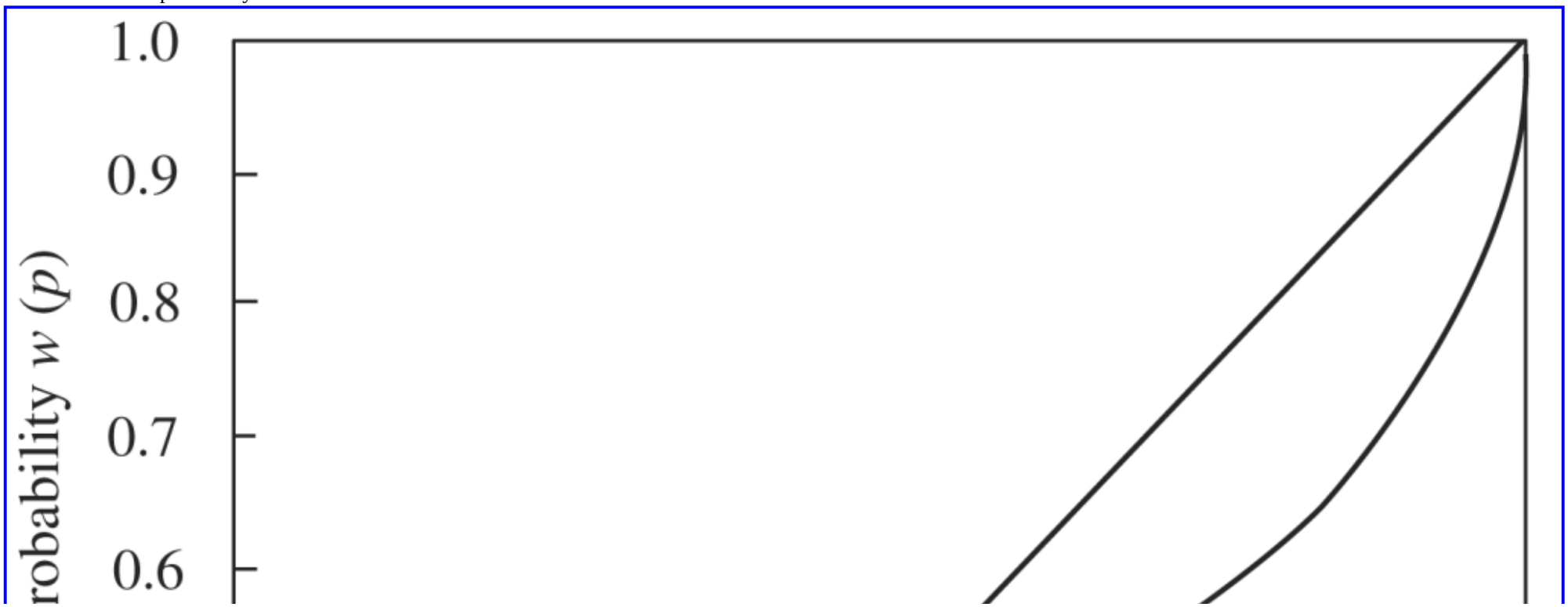

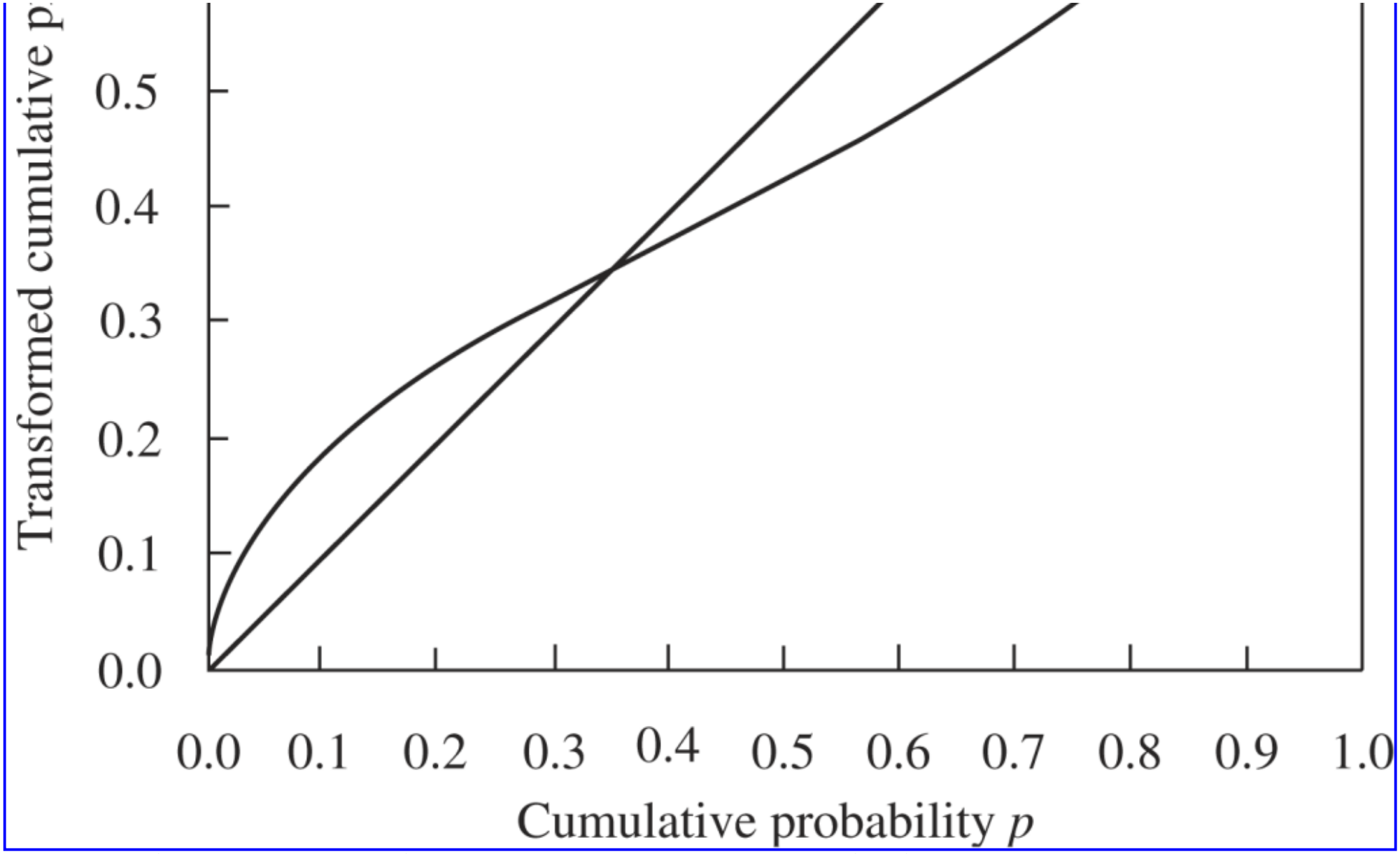

As a consequence of being steeper in the vicinities of 0 and 1, and less steep across the intermediate range, this form of $w(\cdot)$ gives greater weight to extreme than to intermediate outcomes. Although it may be psychologically implausible that most individuals transform probabilities *strictly* according to the rather cumbersome procedure specified by CPT and other rank-dependent models, the approach captures the general intuition that extreme outcomes may attract more attention and receive relatively more weight in decisions. And it appeals to those theorists who are inclined towards models that respect what are perceived to be 'fundamental' requirements of rationality such as transitivity, while also having the advantage of being applicable to prospects involving probability distributions over any number and range of outcomes.

However, the spirit of PT was to provide a *descriptive* model of risky choice, so that violations of transitivity of the kind outlined earlier were an implication of the model; and, to the extent that they occur in practice, PT can claim to be descriptively successful. And indeed there is evidence of such violations (see Starmer, 1999).

On the other hand, as K&T acknowledge, there are limitations to the scope of PT. As discussed above, the domain of the theory was very specific and excluded a number of tasks that are of economic significance (such as the formulation of certainty equivalence values). Moreover, certain assumptions made in the model are open to question. For example, the

phenomenon of ‘event-splitting’ (see Humphrey, 1995) suggests that people may only imperfectly add the probabilities of the same payoff under different ‘states of the world’, contrary to the supposed process of combination in the editing phase. There may also be questions about just how transparent dominance needs to be before it is detected in the editing phase. And some researchers – see, for example, Birnbaum (2006) – have amassed evidence of patterns of choice which appear to run counter to the claims of prospect theories to provide a satisfactory description even of the behaviour which should lie within their domain.

All that having been said, there can be no doubt whatsoever of the success of PT in focusing the attention of decision theorists on patterns of behaviour that do not conform with the conventional (and still predominant) wisdom of EU, and in stimulating a very substantial body of experimental, empirical and theoretical work exploring behaviour outside of the strictures of standard economic ‘rationality’.

See Also

- expected utility hypothesis
- experimental economics
- Kahneman, Daniel
- non-expected utility theory

Bibliography

Allais, M. 1953. Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’école américaine. *Econometrica* 21, 503–46.

Birnbaum, M. 2006. Evidence against prospect theories in gambles with positive, negative and mixed consequences. *Journal of Economic Psychology* 27, 737–61.

Humphrey, S. 1995. Risk aversion or event-splitting effects? More evidence under risk and uncertainty. *Journal of Risk and Uncertainty* 11, 263–74.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Kim, E., Morse, A. and Zingales, L. 2006. What has mattered to economics since 1970. *Journal of Economic Perspectives* 20(4), 189–202.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3, 323–43.

Starmer, C. 1999. Cycling with rules of thumb: an experimental test for a new form of non-transitive behaviour. *Theory and Decision* 46, 139–57.

Starmer, C. 2000. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38, 332–82.

Tversky, A. and Kahneman, D. 1992. Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5, 297–323.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

How to cite this article

Loomes, Graham. "prospect theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000367>
doi:10.1057/9780230226203.1356

The New Palgrave Dictionary of Economics Online

protoindustrialization

Sheilagh Ogilvie

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Proto-industrialization’ is the name given to the massive expansion of export-oriented handicrafts which took place in many parts of Europe between the 16th and the 19th centuries. An influential theory holds that these proto-industries generated the capital, labour, entrepreneurship, agricultural commercialization, and consumer demand needed for factory industrialization. Protoindustrialization, it is argued, also transformed traditional economic mentalities and institutions. However, deeper empirical study has cast doubt on most of these claims. Theories of protoindustrialization have stimulated much excellent research, but do not explain the significant economic growth, demographic change, and institutional transformation that occurred in Europe before the Industrial Revolution.

Keywords

agriculture; fertility; industrial revolution; industrialization; nuptiality; population growth; protoindustrialization; putting-out system; surplus labour; urbanization; workshop system

Article

‘Protoindustrialization’ is the name given to the massive expansion of export-oriented handicrafts which took place in many parts of Europe between the 16th and the 19th centuries.
Often, although not always, such proto-industries arose in the countryside where they were practised alongside agriculture; usually, they expanded without adopting new techniques or centralizing production into factories. This growth of pre-factory industry in early modern Europe has long been a subject of specialized study. But in the 1970s it began to attract much wider interest, when several influential works christened it ‘protoindustrialization’ and argued that it was a major cause of industrialization and capitalism.

Protoindustrialization as the first stage of industrialization

The term ‘protoindustrialization’ was invented by Franklin Mendels, who first used it in his 1969 dissertation on the Flemish linen industry (published in 1981) and popularized it in a now famous article based on that research (Mendels, 1972). Mendels claimed that protoindustrialization was the first phase of industrialization. In the 18th century, seasonally underemployed European country-dwellers moved massively into cottage crafts, exporting their wares beyond the immediate region. This, Mendels argued, broke down traditional urban institutions such as guilds that had previously limited industrial growth. Mendels contended that it also weakened rural institutions such as inheritance systems, communes, and manorial systems that had traditionally calibrated population growth to economic resources. Mendels claimed that this made nuptiality (and thus fertility) ratchet upwards: proto-industrial upswings saw more marriages, but downswings did not see fewer. High protoindustrial fertility fuelled rapid population growth, Mendels argued, in turn causing further industrial expansion. This self-sustaining proto-industrial spiral, according to Mendels, generated the capital, labour, entrepreneurship, agricultural commercialization, and consumer demand needed for factory industrialization.

Protoindustrialization and proletarianization

Mendels's arguments were initially widely adopted, giving rise to several schools of protoindustrial theory. One emanated from David Levine, whose study of two villages in 19th-century Leicestershire appeared to confirm that proto-industry led to population growth (Levine, 1977). For Levine, proto-industry was important mainly because he believed it broke down rural social structure and land ownership, creating a large group of landless people who had to work for wages. This broader process of ‘proletarianization’ was, Levine argued, crucial for capitalism and industrialization.

Protoindustrialization and surplus labour

A third view of protoindustrialization was put forward by Joel Mokyr (1976), who rejected almost all the arguments advanced by Mendels and Levine but argued that proto-industries provided the cheap ‘surplus’ labour to fuel a ‘dualistic’ growth of the European economy as modelled for modern less developed countries (LDCs) by Lewis (1954) and Fei and Ranis (1964). The key empirical problem for the Lewis–Fei–Ranis model was whether ‘surplus’ labour existed and where it came from. Mokyr argued that in pre-industrial Europe surplus labour came from protoindustry, creating a flat labour supply curve and hence very low wages for early factory industry. This ‘dualistic labour surplus’ view of protoindustrialization has hardly been pursued empirically, but is important because of its links with development economics and with Jan De Vries's influential theory of European urbanization (De Vries, 1984).

Protoindustrialization and the transition to capitalism

The protoindustrialization debate was intensified by the publication of a massive book by Peter Kriedte, Hans Medick and Jürgen Schlumbohm (German original 1977, English translation 1981). Combining Mendels's and Levine's findings with the voluminous earlier literature on cottage industries, these

scholars turned the theory of protoindustrialization into a general model of European economic transformation between the medieval and modern periods.

For them, protoindustrialization was the 'second phase' of this transformation process. The first phase, they claimed, was a loosening of feudalism caused by commutation of feudal burdens from labour or grain dues into money rents, polarizing the rural population into two classes: well-off peasants with enough land to live solely from farming, and land-poor or landless strata who had to seek work outside agriculture. The second phase, in their view, was the 16th-century growth in supra-regional and international trade, creating a growing demand for manufactures which the new rural proletariat could satisfy more cheaply than guild-regulated urban craftsmen. So protoindustries arose in the countryside.

These scholars proposed a stage theory according to which rural protoindustries then gradually transformed industrial organization. The first stage, they claimed, was the *Kaufsystem* (artisanal or workshop system), in which rural producers retained autonomy over production and selling. The second stage, the argument continues, was the *Verlagssystem* (putting-out system), in which merchants bought raw materials, 'put them out' to the rural producers who processed them in return for a wage, and then collected the output for transfer either to the finishing stages of production or to the final consumer market. This ultimately led to a third stage, it was claimed: the concentration and mechanization of production in centralized, mechanized factories.

Extensions to the theories of protoindustrialization

By 1977 at the latest, therefore, protoindustrialization had generated a family of different theories, based on differing definitions of protoindustry and differing explanations of economic development. Almost all they had in common was to emphasize the significance of European economic and demographic growth before factory industrialization, and to ascribe such growth to changes in a certain economic sector – export-oriented cottage industry. Over the following decades, these various branches of protoindustrialization theory stimulated a huge outpouring of research into pre-industrial manufacturing, not just in Europe but also in the non-European world, including modern LDCs.

By 1982 protoindustrialization had become such an influential concept that Franklin Mendels and Pierre Deyon were invited to convene one of the three main sessions of the Eighth International Economic History Congress in Budapest, with protoindustrialization as their theme. They pre-circulated a set of hypotheses, 48 researchers contributed papers (Deyon and Mendels, 1982), and Mendels summarized the session with a report, a revised definition, and a set of hypotheses for subsequent debate (Mendels, 1982).

This new 1982 definition of protoindustrialization stressed five key characteristics. First, protoindustrialization occurred not nationally or internationally, but *regionally*: 'within a small radius around a regional capital'. Second, protoindustries must be distinguished from traditional crafts: they produced not for local or regional consumption, but for sale to *export* markets outside the region. Third, protoindustry was mainly *rural* and *part-time* – only in its final or extreme phase did it involve full-time industrial employment. Fourth, protoindustrialization arose symbiotically with *agricultural commercialization*. Finally, protoindustrialization was 'dynamic': it was defined as a *growth* over time in the industrial employment of rural workers.

Deyon and Mendels also proposed four central hypotheses about the *effects* of protoindustrialization.

First, protoindustry led to population growth and land fragmentation because it broke down traditional demographic regulation by communes, landlords and inheritance systems. Second, protoindustrial profits created the capital for factory industrialization. Third, protoindustry trained merchants and workers in the skills needed for factory industrialization. Finally, protoindustrialization caused agriculture to commercialize, thereby feeding urbanization and industrialization. Through these four mechanisms, proto-industry led to factory industry – although the authors admitted that sometimes it led to de-industrialization instead (Mendels, 1982).

Criticisms of the theories of protoindustrialization

Somewhat more slowly than they attracted support, the theories of protoindustrialization also began to draw criticism.

For one thing, the precise size and structure of the unit that qualified as a protoindustrial *region* was unclear. Protoindustries could and often did extend beyond the radius around a single market town, or alternatively were sometimes found in only one or two communities in such a radius. One pragmatic solution was to define the region as simply the area within which a certain proto-industry was practised. But this seemed to leach the concept of the region of much of its analytical content. Second, there was no agreement about how large a proportion of the regional labour force must have been employed in protoindustry, or how fast or sustained the growth of this labour force must have been, in order to qualify as ‘protoindustrialization’ (Ogilvie and Cerman, 1996).

There was also confusion about the precise importance of *export markets* for protoindustrialization. First, why were export markets uniquely important? Second, what proportion of production had to be exported in order for any given industry to qualify as a proto-industry instead of a craft? Third, how distant did final markets have to be to qualify as ‘supra-regional’ rather than ‘local’? The demarcation between local crafts and export-oriented proto-industries was thus very unclear and its analytical importance remained obscure.

The *neglect of other forms of industry* was another weak point. The theories of protoindustrialization concentrated solely on one sort of pre-industrial manufacturing: cottage industry. But what justified this emphasis? Did manufacturing really develop just because of this single sort of industry, which was often technologically very primitive? What about highly skilled and technologically innovative crafts, export-oriented urban industries, or centralized manufactories? Mainstream historians of pre-industrial manufacturing argued that all these branches of the secondary sector should be included in any analysis of industrialization before the Industrial Revolution (Schremmer, 1981; Coleman, 1983; Mager, 1993).

Others argued that large *urban* export industries, and those involving centralized production units, should also be included under the rubric of protoindustrialization (Cerman, 1993).

The *neglect of industrial technology and physical geography* was also criticized. Mendels referred in passing to industrial production functions and transportation costs, but neither he nor other proponents of the theory explored these factors further. Critics argued that any coherent view of protoindustry must consider the technical requirements of different branches of industry and the geographical and physical characteristics of the region (Mager, 1993). Others urged that protoindustry, like any economic activity, be analysed in terms of ‘opportunity costs’, and pointed out that this would imply taking into account a whole array of technological, geographical, and institutional variables (Ogilvie, 1993; 1997).

The theories adopted strong assumptions about the '*traditional societies*' transformed by proto-industry, and these assumptions began to be questioned (Coleman, 1983; Houston and Snell, 1984; Schremmer, 1981; Ogilvie and Cerman, 1996; Ogilvie, 1997). Protoindustrialization theorists had uncritically accepted the theories of Alexander Chayanov, who regarded peasants as unable and unwilling to calculate costs, seek profits, use money, or transact in markets (Chayanov, 1966). But was this really true of the early modern European rural population? The subsistence-orientation assumed for rural domestic workers was not confirmed by empirical studies, and was inconsistent with the fact that proto-industrial producers often became traders, middlemen, putters-out and even manufactory-operators. The demographic decisions and productive choices of protoindustrial workers, rather than being governed by 'traditional mentalities', began to look highly rational (Ogilvie, 1997).

The *demographic* predictions of the theories were widely falsified as empirical studies proliferated. It emerged that pre-industrial demographic behaviour was influenced by such a wide array of variables that proto-industry could have highly divergent effects on nuptiality, fertility, mortality and migration in different European societies. Case studies showed that not all protoindustrial regions had greater population density, faster demographic growth, lower ages of marriage, higher fertility rates, larger households, or a breakdown in the family and gender division of labour – all of which had been postulated in the original theories. Furthermore, many – even all – of these demographic changes could also be observed in some primarily agricultural regions (Schremmer, 1981; Coleman, 1983; Houston and Snell, 1984; Ogilvie and Cerman, 1996; Ogilvie, 1997).

The relationship between *commercial agriculture* and protoindustry was also disputed. Protoindustries arose alongside many different kinds of agriculture, including subsistence cultivation, market farming, and even large feudal domains worked by serf labour. Protoindustries derived food and raw materials not just from commercial agriculture but from local cultivation by proto-industrial workers themselves. Simultaneous employment in proto-industry and agriculture was common but not universal in proto-industrial regions. While traditional agrarian institutions and rural social structure broke down in some protoindustrial regions, in others they survived unaltered for centuries (Houston and Snell, 1984; Ogilvie and Cerman, 1996; Ogilvie, 1997).

The role of *social and political institutions* in theories of protoindustrialization has also been critically revised (Ogilvie, 1993; Ogilvie and Cerman 1996; Ogilvie, 1997; Ogilvie, 2004). The original theorists assumed that protoindustrialization both required and furthered the replacement of 'traditional' social institutions with markets. But deeper research has shown that urban privileges, craft guilds, monopolistic merchant companies, village communities and manorial institutions remained important in many European protoindustries, and crucially influenced economic, demographic and social change in proto-industrial regions.

A final major criticism questioned the role of proto-industry in causing *factory industrialization*. Each of the mechanisms by which protoindustrialization is supposed to have led to industrialization has been subject to sceptical re-evaluation. Research shows that the demographic effects of protoindustrialization were extremely various, as was its impact on the fragmentation of landholdings. Protoindustry appears to have been only one of many sources of capital invested in the early factories, and in many cases proto-industrial profits flowed into agriculture, landholding or socio-political investments. Proto-industry was also only one of many sources of entrepreneurial skills for industrialization, and sometimes did not encourage entrepreneurship at all. There is little evidence that it was proto-industry that led to

commercial agriculture rather than that agricultural surpluses made possible both proto-industrial regions and urbanization. It is now widely acknowledged by both the theorists and their critics that proto-industry often led not to factories but to de-industrialization and a return to agriculture. The critics argue that this finding denudes the theory of most of its empirical content (Coleman, 1983; Houston and Snell 1984; Clarkson, 1985; Ogilvie and Cerman, 1996). Although, therefore, the theory of protoindustrialization has stimulated much excellent research, it does not explain the significant economic growth, demographic change, and socio-institutional transformation that indisputably occurred in Europe well before the industrial revolution.

See Also

- agriculture and economic development
- capitalism
- development economics
- dual economies
- economic demography
- economic history
- growth and institutions
- historical demography
- Industrial Revolution
- labour surplus economies
- Lewis, W. Arthur
- medieval guilds
- peasants

Bibliography

Cerman, M. 1993. Protoindustrialization in an urban environment: Vienna, 1750–1857. *Continuity and Change* 8, 281–320.

Chayanov, A. 1966. The Theory of Peasant Economy, ed. D. Thorner, B. Kerblay and R. E. F. Smith. Homewood, IL: Richard A. Irwin.

Clarkson, L. 1985. *Proto-Industrialization: The First Phase of Industrialization?* Houndmills: Macmillan.

Coleman, D. 1983. Protoindustrialization: a concept too many. *Economic History Review*, 2nd ser. 36, 435–48.

De Vries, J. 1984. *European Urbanization 1500–1800*. Cambridge, MA: Harvard University Press.

Deyon, P. and Mendels, F., eds. 1982. *La protoindustrialisation: théorie et réalité*. Lille: Université de Lille.

Fei, J. and Ranis, G. 1964. *Development of the Labour Surplus Economy: Theory and Policy*. Homewood, IL: Richard A. Irwin.

Houston, R. and Snell, K. 1984. Proto-industrialisation? Cottage industry, social change and Industrial Revolution. *Historical Journal* 27, 473–92.

Kriedte, P., Medick, H. and Schlumbohm, J. 1977. *Industrialisierung vor der Industrialisierung. Gewerbliche Warenproduktion auf dem Land in der Formationsperiode des Kapitalismus*. Göttingen: Vandenhoeck & Ruprecht.

Kriedte, P., Medick, H. and Schlumbohm, J. 1981. *Industrialization before Industrialization: Rural Industry in the Genesis of Capitalism*. Cambridge: Cambridge University Press.

Levine, D. 1977. *Family Formation in an Age of Nascent Capitalism*. New York and London: Academic Press.

Lewis, W. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economics and Social Studies* 22, 139–91.

Mager, W. 1993. Protoindustrialization and proto-industry: the uses and drawbacks of two concepts. *Continuity and Change* 8, 181–216.

Mendels, F. 1972. Protoindustrialization: the first phase of the industrialization process. *Journal of Economic History* 32, 241–61.

Mendels, F. 1981. *Industrialization and Population Pressure in Eighteenth-Century Flanders*. New York: Arno Press.

Mendels, F. 1982. Protoindustrialization: theory and reality. General Report, in *Eighth International Economic History Congress, Budapest 1982: 'A' Themes*, ed. F. Mendels and P. Deyon. Budapest.

Mokyr, J. 1976. Growing-up and the Industrial Revolution in Europe. *Explorations in Economic History* 13, 371–96.

Ogilvie, S. 1993. Protoindustrialization in Europe. *Continuity and Change* 8, 159–79.

Ogilvie, S. 1996. Social institutions and protoindustrialization. In *European Proto-Industrialization*, ed. S. Ogilvie and M. Cerman. Cambridge: Cambridge University Press.

Ogilvie, S. 1997. *State Corporatism and Proto-Industry: The Württemberg Black Forest, 1580–1797*. Cambridge: Cambridge University Press.

Ogilvie, S. 2004. Guilds, efficiency and social capital: evidence from German proto-industry. *Economic History Review* 57, 286–333.

Ogilvie, S. and Cerman, M. 1996. Protoindustrialization, economic development and social change in early modern Europe. In *European Proto-Industrialization*, ed. S. Ogilvie and M. Cerman. Cambridge: Cambridge University Press.

Schremmer, E. 1981. Proto-industrialisation: a step towards industrialisation? *Journal of European Economic History* 10, 653–70.

How to cite this article

Ogilvie, Sheilagh. "protoindustrialization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000341> doi:10.1057/9780230226203.1357

The New Palgrave Dictionary of Economics Online

psychological games

Martin Dufwenberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Traditional game-theoretic models assume that utilities depend only on actions. This is not sufficient for describing the motivations and choices of decision makers who care about reciprocity, emotions, or social rewards. Psychological games allow utilities to depend directly on beliefs (about beliefs) in addition to which actions are chosen, and they can capture a wider range of motivations. This article contains several examples and it is indicated where research on psychological games is headed.

Keywords

Allais paradox; belief-dependent motivation; commitment; decision theory; emotions; extensive game forms; game theory; guilt aversion; psychological forward induction; psychological games; reciprocity; signalling; trust; von Neumann and Morgenstern

Article

Traditional game-theoretic models presume that utilities depend on actions. While this framework is quite general (it can, for example, accommodate profit-maximization, altruism, inequity aversion and Rawlsian maximin preferences) it is not rich enough to adequately describe several psychological or social aspects of motivation which depend directly on beliefs (about beliefs) in addition to which actions are chosen. The following example illustrates.

Karen feels guilty if she lets others down. When paying her landscaper (Jim), this influences her tipping. The more she believes Jim believes he will receive as a tip, the more she gives. More precisely, she gives just as much as she believes Jim believes he will get, in order to avoid the feelings of guilt that will plague her if she gives less.

Beyond depicting something arguably realistic, the example illustrates in the simplest possible way how one may have to transcend traditional game theory to model a belief-dependent motivation. Consider a standard game form where Karen chooses a tip t such that $0 \leq t \leq w$, where w is the number of dollars in

her wallet, and where the landscaper has no choice (his strategy set is modelled as a singleton $\{x\}$). Karen's choice of tip thus pins down a strategy profile (t,x). In traditional game theory, payoffs are defined on strategy profiles (or on endnodes induced by strategy profiles), so Karen's best choice (or choices) would be independent of her belief about Jim's belief about her choice of tip. This runs counter to the example.

Gilboa and Schmeidler (1988) and Geanakoplos, Pearce and Stacchetti (1989) present several examples that illustrate the inadequacy of traditional methods of representing preferences that reflect various forms of belief-dependent motivation. Geanakoplos, Pearce and Stacchetti develop a new analytical framework, in which the centrepiece is the notion of a *psychological game*, which may be seen as a generalization of a traditional game and which can model some of the desired effects. A psychological game differs from a traditional game in that utilities are defined on beliefs (about actions and beliefs), as well as on which actions are chosen. (The term ‘game with belief-dependent motivation’ would be more descriptive than the term ‘psychological game’, but I stick with the latter, which has become established.)

Reciprocity

The best-known example of a psychological games-based application is Rabin's (1993) highly influential model of reciprocity, according to which players wish to act kindly (unkindly) in response to kind (unkind) actions. The key notion of kindness depends on beliefs in such a way that reciprocal motivation can be described only by using psychological games. To see why, suppose that I jump out in front of your car, blocking your way, so that you can't cross a bridge and therefore arrive late to an important meeting. Am I kind? Clearly one cannot say without knowing what my beliefs are. If I believe the bridge is as sturdy as bridges usually are and I am just goofing around, then I am unkind. However, if I believe the bridge is about to collapse, then I am kind. Arguably, I would be kind even if I mistook a sturdy bridge for a dangerous one. So should you be kind or unkind in return? The answer depends on your beliefs about my kindness, and hence on your beliefs about my beliefs. It takes a psychological game to model that. (The example given here is similar in spirit to another example given in Rabin, 1998, p. 23. Rabin's model is normal-form based. See Dufwenberg and Kirchsteiger, 2004, for an extension to extensive game forms. See Fehr and Gächter, 2000, and Sobel, 2005, for general discussions of why reciprocity has important economic consequences.)

Emotions

Reciprocity is but one form of motivation that can be modelled by means of psychological games. Many emotions are good candidates. In his article ‘Emotions and Economic Theory’, Elster (1998) argues that a variety of emotions have important economic consequences, and he laments how little attention economists have paid to this. He argues that a key characteristic of emotions is that ‘they are triggered by beliefs’ (1998, p. 49). He discusses anger, hatred, guilt, shame, pride, admiration, regret, rejoicing, disappointment, elation, fear, hope, joy, grief, envy, malice, indignation, jealousy, surprise, boredom, sexual desire, enjoyment, worry, and frustration. He asks (1998, p. 48): ‘[H]ow can emotions help us explain behavior for which good explanations seem to be lacking?’ Psychological games may be useful for providing answers.

But little work has been done. One exception is Caplin and Leahy's (2004) health care model in which a physician is concerned with a patient's belief-dependent anxiety (compare also Caplin and Leahy, 2001). Another exception is the emotion of guilt for which a string of results, both theoretical and experimental, have been established for the specific context of trust games (see Huang and Wu (1994), Dufwenberg (1995; 2002), Dufwenberg and Gneezy (2000), Bacharach, Guerra and Zizzo (2007), and Charness and Dufwenberg (2006). I shall elaborate in some detail on these latter findings (borrowing eclectically from the cited works), since they may be suggestive of the importance of psychological games more generally in a variety of ways.

Consider the game in Figure 1, where payoffs reflect money income (first for player *A*, then for player *B*) but not the players' preferences which may depend also on guilt as will be indicated.

Figure 1

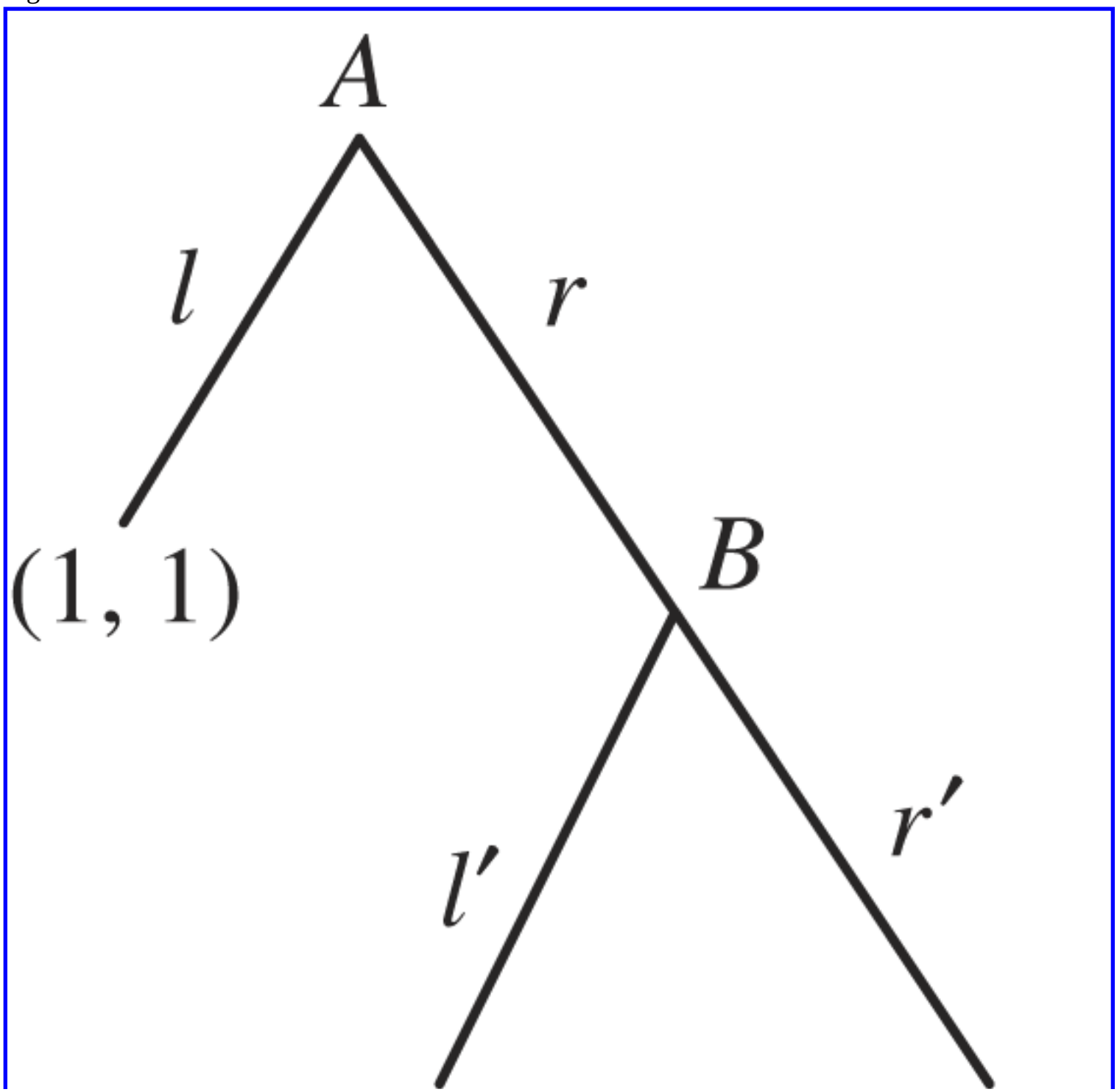

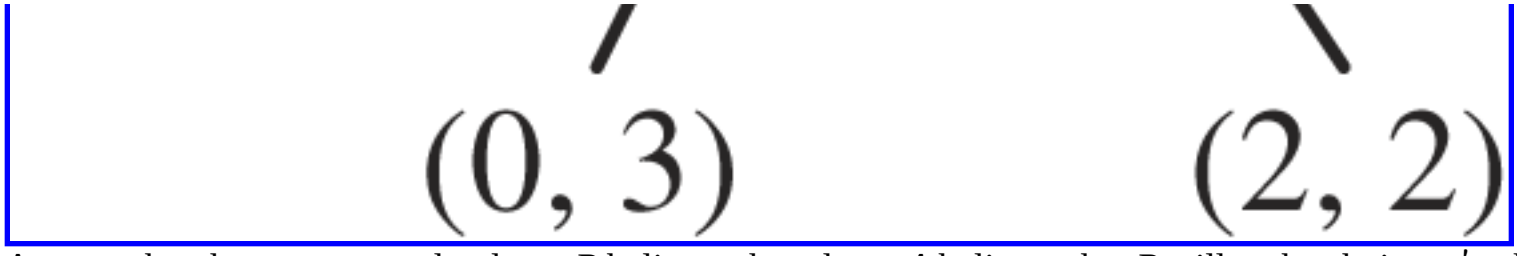

Assume that the more strongly player *B* believes that player *A* believes that *B* will make choice r', the more guilt *B* would feel making choice l' and the more likely *B* is to make choice r'. Specifically, the players' utilities at the various end nodes in the game form of Figure 1 coincide with the monetary payoffs, *except* following the choice sequence (r, l') where *B*'s utility is $3\cdot(1-\beta)$ rather than 3, and where β is a measure of *B*'s belief (with range from 0 to 1) about *A*'s belief that *B* will choose r'. (More specifically, *B* has a probability measure describing her beliefs about which probability *A* assigns to the choice r' conditional on *A* choosing *r*; β is the mean of that measure.) Say that *B* is *guilt averse*. This is all modelled in the psychological game in Figure 2:

Figure 2

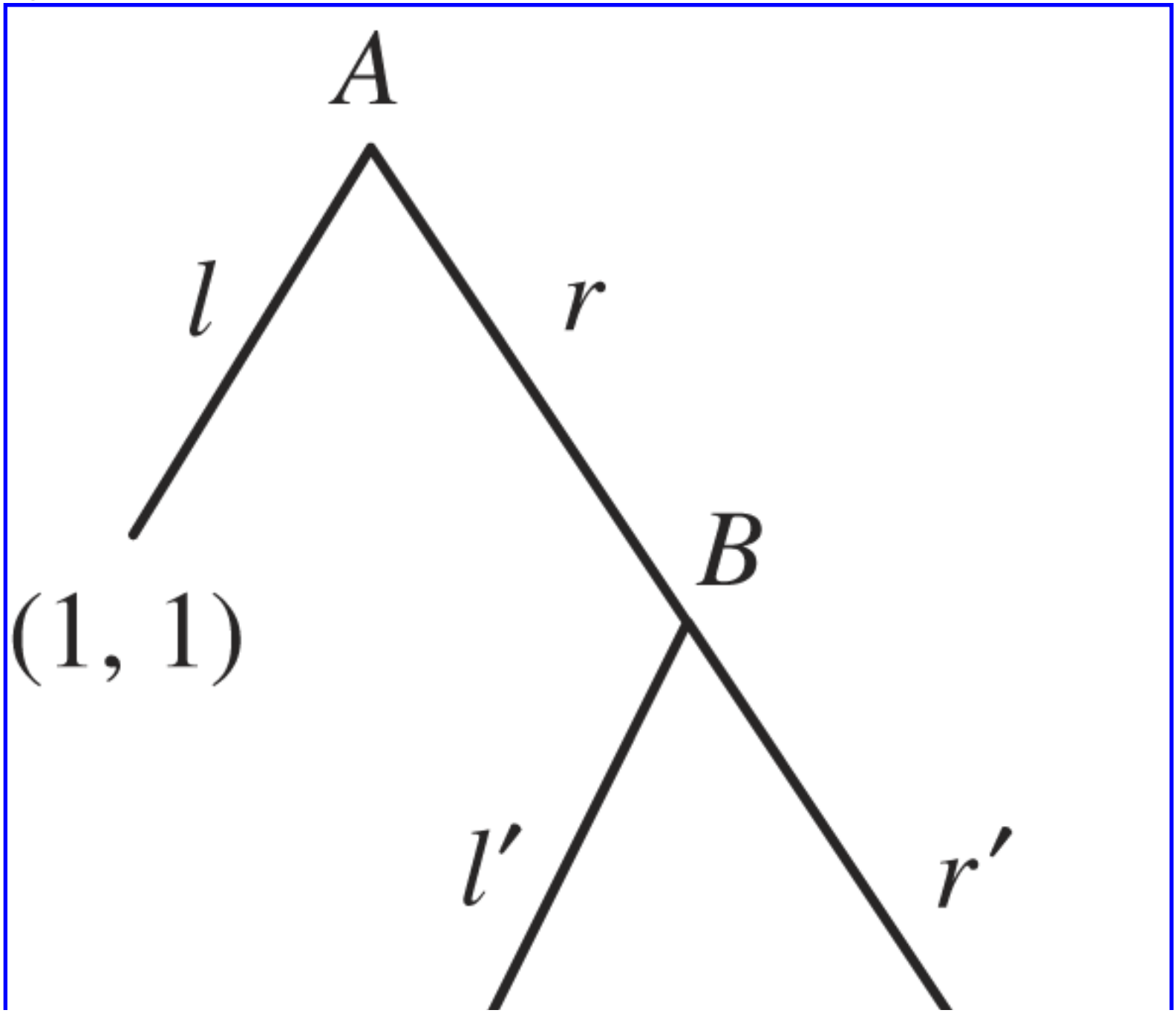

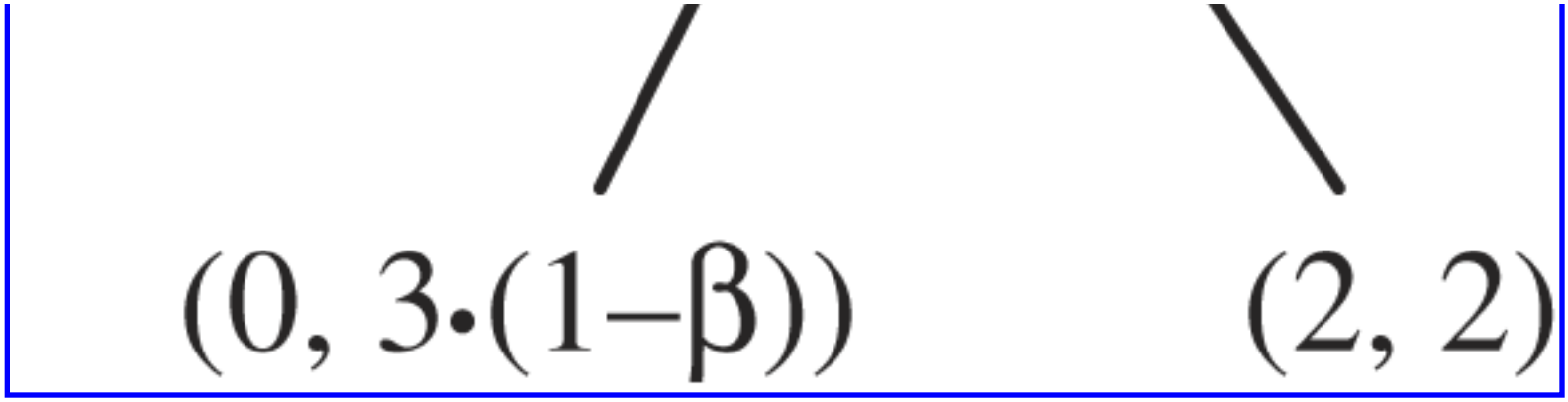

I wish to make several points. First, the guilt aversion modelled in Figure 2 is similar to that involved in the above example featuring Karen and Jim. In fact, the idea that people feel guilty in proportion to the degree to which they do not live up to another's expectations can be extended to any game. (See Battigalli and Dufwenberg, 2007, for a recent attempt at doing this.)

Second, one can test for guilt aversion experimentally, but this requires one to measure *B*'s belief β. This can be done by inviting subjects to make guesses about one another's choices and guesses, rewarding accuracy in the guesswork. Such experimental tests have indicated that the prediction of guilt aversion is empirically supported in trust games. (The involved form of belief elicitation could conceivably be usefully complemented by two other forms of measurement: emotional self-reports and neurological methods such as functional magnetic resonance imaging.)

Third, guilt aversion may provide the seeds of a theory why communication can help foster trust and cooperation. To illustrate with reference to Figure 2, suppose that, before play, *A* and *B* meet and talk. Player *B* looks player *A* in the eye and *promises* to choose r'. If *A* believes this, and if *B* believes that *A* believes this, then guilt aversion would make *B* live up to her promise. A promise by *B* can thus feed a self-fulfilling circle of beliefs about beliefs that r' will be chosen. In combination with guilt aversion, words may be tools that create commitment power, which may in turn foster trust and cooperation.

Fourth, even without communication between *A* and *B*, one may argue that if *B* is guilt averse (as described above) then trust and cooperation will ensue. If *A* is rational and maximizes his expected monetary income (recall that we have assumed that *A* is selfish in this way) then by choosing *r* he *signals* a certain strength of belief in *B* choosing r'; if *A* did not assign a probability of at least 1/2 to *B* choosing r' then he would rather chose *l*. If *B* figures this out, it puts a lower bound of 1/2 on β. So *B* is forced to hold a belief such that she would feel so guilty if she choose l' that she prefers r'; in numbers, with $\beta \geq 1/2$ we get $3\cdot(1-\beta)<2$. If *A* figures this out, he should of course choose *r*. The illustrated phenomenon has been labelled *psychological forward induction*.

To sum up: the idea of guilt aversion is intuitively plausible, experimentally testable, empirically supported, relevant for explaining why communication matters to economic behaviour, and suggestive of intriguing signalling issues that may shape emotions and behaviour. These insights concern a very special emotion and a very special psychological game, but seem profound given that limited scope. One may reasonably suspect that exciting conclusions are in store also for other emotions and other strategic settings, and that these conclusions may in part concern communication or belief signalling.

Social rewards

The discussion so far may have been misleading with its rather heavy emphasis on reciprocity and

emotions. Psychological game theory may be relevant also for describing certain social rewards (norms, respect and status), where decision makers somehow care about the opinions or views of others. Bernheim (1994) and Dufwenberg and Lundholm (2001) present models that bear this out. These authors do not make explicit mention of psychological games, but if one takes a close look at the mathematical details one can discover connections.

Developing the theory

One might hope that Geanakoplos, Pearce and Stacchetti's framework is appropriate for tackling all the interesting problems to which psychological games may be relevant. However, this is not the case. A careful scrutiny reveals that their approach is too restrictive to handle many plausible forms of belief-dependent motivation (as they acknowledge themselves; see 1989, pp. 70, 78). There are several reasons, including the following:

1. (1) Geanakoplos, Pearce and Stacchetti only allow initial beliefs to enter the domain of a player's utility, while many forms of belief-dependent motivation require *updated* beliefs to play a role.
2. (2) Geanakoplos, Pearce and Stacchetti only allow a player's own beliefs to enter the domain of his utility, while there are conceptual and technical reasons to let *others'* beliefs matter.
3. (3) Geanakoplos, Pearce and Stacchetti follow the traditional extensive-games approach of letting strategies influence utilities only in so far as they influence which end node is reached, but many forms of belief-dependent motivation become compelling in conjunction with preferences that depend on strategies in ways not captured by end nodes.
4. (4) Geanakoplos, Pearce and Stacchetti restrict attention to equilibrium analysis, but in many strategic situations there is little compelling reason to expect players to coordinate on an equilibrium, and one may wish to explore alternative assumptions.

(1) is manifest, for example, in the above psychological forward induction argument which hinges crucially on B's motivation depending on an updated belief. (2) is relevant, for example, for modelling social rewards (compare the above comments on Bernheim's and Dufwenberg and Lundholm's models). As regards (3), one can show that the issue comes up if one wants to model, for example, regret, disappointment or guilt. (4) echoes considerations relevant also for traditional games; equilibrium play is not a self-evident proposition in many contexts, for example if one assumes (only) that there is common belief in rationality or in learning scenarios.
The list (1)–(4) is adapted from Battigalli and Dufwenberg (2005), who elaborate in more detail on each issue and take first steps towards developing psychological game theory in the indicated directions. Their approach draws crucially on Battigalli and Siniscalchi's (1999) work on how to represent hierarchies of conditional beliefs.

Decision-theoretic foundations

The decision-theoretic foundations of psychological game theory are not well understood. Classical decision theory (say, von Neumann and Morgenstern) does not apply straightforwardly. Too see this,

take the emotion of disappointment as an example. It is plausible that disappointment is a belief-dependent emotion. To exemplify, I have I just failed to win a million dollars and I am not at all disappointed, which, however, I clearly would be if I were playing poker and knew I would win a million dollars unless my opponent got extremely lucky drawing to an inside straight, and then he hit his card. Another example could be based on the lotteries used in the so-called Allais paradox. In both cases the level of disappointment, which if anticipated might affect choice behaviour, may plausibly depend on the strength of the prior belief that a decision maker will win a lot of money. It follows that, unless consequences are described so as to include a specification of disappointment, the so-called ‘independence axiom’ will not make much sense for decision makers who are prone to disappointment. Decision theorists have given related matters some attention, but not a lot. Machina (1981, pp. 172–3; 1989, p. 1662) presents examples in spirit related to the one million dollar example above. Bell (1985), Loomes and Sugden (1986), Karni (1992), and Karni and Schlee (1995) go on to develop models in which utility may depend directly on beliefs; the latter two references take axiomatic approaches. Robin Pope has written extensively, over many years, about how conventional decision theory excludes various forms of belief-dependent motivation; Pope (2004) expounds her programme and gives further references. Caplin and Leahy (2001) develop a model of ‘psychological expected utility’ that admits belief-dependent motivation. However, these contributions mainly develop perspectives for settings with single decision-makers, and more will be needed to address games more generally.

Conclusion

Research on psychological games is still in its infancy. This is true for all facets of investigation: the development of basic classes of games and solution concepts, the investigation of decision-theoretic underpinning, tests of empirical (most likely experimental) validity, and finally applied theoretical work which uses psychological game theory to analyse various economic models. In each of these domains some work has been done which is indicative of the viability of the line of research, and there is good reason to be thrilled about the prospects for future research.

See Also

- behavioural economics and game theory
- behavioural game theory
- expectations
- expected utility hypothesis
- game theory
- neuroeconomics
- reciprocity and collective action
- social norms

Bibliography

Bacharach, M., Guerra, G. and Zizzo, D. 2007. The self-fulfilling property of trust: An experimental study. *Theory and Decision* 63, 349–88.

Battigalli, P. and Dufwenberg, M. 2005. Dynamic psychological games. Working Paper No. 287, IGIER, Bocconi University.

Battigalli, P. and Dufwenberg, M. 2007. Guilt in games. American Economics Review, 97, 170–6.

Battigalli, P. and Siniscalchi, M. 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory* 88, 188–230.

Bell, D. 1985. Disappointment in decision making under uncertainty. *Operations Research* 33, 1–27.

Bernheim, B.D. 1994. A theory of conformity. *Journal of Political Economy* 102, 841–77.

Caplin, A. and Leahy, J. 2001. Psychological expected utility and anticipatory feelings. *Quarterly Journal of Economics* 116, 55–79.

Caplin, A. and Leahy, J. 2004. The supply of information by a concerned expert. *Economic Journal* 114, 487–505.

Charness, G. and Dufwenberg, M. 2006. Promises and Partnership. *Econometrica* 74, 1579–601.

Dufwenberg, M. 1995. Time consistent wedlock with endogenous trust. In doctoral dissertation, *Economic Studies* 22, Uppsala University.

Dufwenberg, M. 2002. Marital investment, time consistency, and emotions. *Journal of Economic Behavior and Organization* 48, 57–69.

Dufwenberg, M. and Gneezy, U. 2000. Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior* 30, 163–82.

Dufwenberg, M. and Kirchsteiger, G. 2004. A theory of sequential reciprocity. *Games and Economic Behavior* 47, 268–98.

Dufwenberg, M. and Lundholm, M. 2001. Social norms and moral hazard. *Economic Journal* 111, 506–25.

Elster, J. 1998. Emotions and economic theory. *Journal of Economic Literature* 36, 47–74.

Fehr, E. and Gächter, S. 2000. Fairness and retaliation: the economics of reciprocity. *Journal of*

Economic Perspectives 14, 159–81.

Geanakoplos, J., Pearce, D. and Stacchetti, E. 1989. Psychological games and sequential rationality. *Games and Economic Behavior* 1, 60–79.

Gilboa, Y. and Schmeidler, D. 1988. Information dependent games: can common sense be common knowledge? *Economics Letters* 27, 215–21.

Huang, P. and Wu, H.-M. 1994. More order without more law: a theory of social norms and organizational cultures. *Journal of Law, Economics & Organization* 10, 390–406.

Karni, E. 1992. Utility theory with probability dependent outcome valuation. *Journal of Economic Theory* 57, 111–24.

Karni, E. and Schlee, E. 1995. Utility theory with probability dependent outcome valuation: extensions and applications. *Journal of Risk and Uncertainty* 10, 127–42.

Loomes, G. and Sugden, R. 1986. Disappointment and dynamic consistency in choice under uncertainty. *Review of Economic Studies* 53, 271–82.

Machina, M. 1981. ‘Rational’ decision making versus ‘rational’ decision modeling. *Journal of Mathematical Psychology* 24, 163–75.

Machina, M. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27, 1622–68.

Pope, R. 2004. Biases from omitted risk effects in standard gamble utilities. *Journal of Health Economics* 23, 1029–50.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83, 1281–302.

Rabin, M. 1998. Psychology and economics. *Journal of Economic Literature* 36, 11–46.

Sobel, J. 2005. Interdependent preferences and reciprocity. *Journal of Economic Literature* 43, 392–436.

How to cite this article

Dufwenberg, Martin. "psychological games." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

dictionaryofeconomics.com/article?id=pde2008_P000357> doi:10.1057/9780230226203.1358

The New Palgrave Dictionary of Economics Online

psychology of social networks

Philippa Pattison, Garry Robins and Yoshi Kashima
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We review the psychological processes underpinning network formation and network-based processes, focusing first on the nature of relationships and their formation, and then on the consequences of networks for individual outcomes and behavior. We argue that it is important to develop methodological approaches that allow us to regard these processes as hypotheses to be tested rather than as unquestioned assumptions. We suggest that different types of networks and processes are likely to lead to different conclusions about these hypotheses, and that the development of models for networks and network processes should therefore be grounded in careful empirical analysis.

Keywords

social networks, psychology of; network formation; network models; network ties; relational models; relational schemas; social roles; tie interdependencies; exponential random graph models

Article

The significance of social networks for an understanding of the structure and dynamics of our contemporary social world is now widely acknowledged. Different types of networks serve as channels through which, for example, knowledge is diffused, opportunities are recognized, cooperation is garnered and actions are coordinated. Networks have been invoked in many disciplines in order to explain the nature and consequences of these channelling effects, spawning interest in the capacity to model network structure. This capacity is important because of the potentially powerful interplay between network structures and the dynamics of the social transactions that they support.
Two broad strategies for network model building have been identified (Jackson, 2005): a statistical approach, in which networks are seen as the outcome of locally interactive and self-organizing tie-formation processes; and a deterministic, game-theoretic approach, in which networks are seen as the outcome of self-interested behaviour of utility-maximizing actors. In association with the statistical

approach, there has been a dramatic increase in our capacity to build theoretically defensible statistical models for social networks whose parameters can be estimated from empirical observations and that have the capacity to reproduce many important global characteristics of observed social networks (for example, Snijders et al., 2006). The game-theoretic approach, on the other hand, is responsible for an impressive accumulation of theoretical results linking strategic activities of pairs of actors to the emergence of ‘efficient’ network structures.

These two modelling approaches differ in two important respects. The first is in the use of deterministic or stochastic models (but see Snijders, 2001). The second distinction is a deeper one, contrasting a conceptualization of actors as self-interested rational decision makers, on the one hand, with more socialized and enculturated actors, on the other. The latter may sometimes be driven by personal utility, but may also engage in non- or extra-rational behaviours that are enabled and constrained by local social network configurations. Indeed, whereas the game-theoretic approach largely assumes that the actor's decision to form or discontinue a relationship is mediated by the actor's conscious computation of utility, the statistical approach treats this as a hypothesis, and empirically examines *whether* an actor's local network configuration appears to constrain or enable tie formation or maintenance. Thus, this second distinction encapsulates a fundamental empirical question: can network structures and processes be explained in terms of the rational activity of self-interested individuals, or are extra-rational affective, social and/or cultural processes systematically at work in the formation and impact of networks?

With this question in mind, we outline what is known about the social and psychological processes that underpin the formation of networks and the dynamics of network-based processes. We first discuss the nature of relationships and their development, and then examine their consequences for individual outcomes and behaviour. We finish by drawing some implications for model-building.

What is a network tie and why do relationships form?

A network relationship, or tie, is assumed to have some continuity in time, with a relevant past and a somewhat predictable future. This temporal continuity is facilitated by cognitive representation and the use of culturally laden relational descriptors, such as friend and friendship, or partner and partnership, features that also facilitate communication about the nature of ties. Relationships, in other words, entail complex socio-cultural schema that not only frame interpretations of past interactions but also shape expectations concerning future ones by the actors concerned as well as by third-party onlookers.

Although tie formation has sometimes been explained in terms of the utility that ties bring for tie partners (for example, Cook and Whitmeyer, 1992), recent psychological research suggests that significant extra-rational psychological processes may also underlie tie formation. Andersen and Chen (2002) have invoked the concept of *relational self* to explain the pervasive impact of relationships with significant others on the way in which individuals interpret and respond to interpersonal encounters, and therefore form *future* interpersonal ties. This concept is founded on a demonstrable ‘transference’ effect, in which past experiences with significant others can be shown to influence new relationships, often outside of conscious awareness. Holmes (2000) has argued that the concept of relationship is best seen as grounded in the interaction between interdependent actors, and hence as an emergent property of dynamic interaction and influence processes. Generic cognitive representations about relationships called *relational schemas* are postulated to represent actors’ developing knowledge of self, partner and

expected sequences of interaction (for example, Baldwin, 1992).

More generally, Fiske (2004) has proposed that four elementary and universal cognitive schemas frame *all* interpersonal relationships. These four schemas are proposed to structure potential interactions between two actors in terms of: collective belonging or solidarity, as in family membership (the *communal sharing* schema); asymmetrical difference, as in hierarchies based on skill, knowledge or social class (*authority ranking*); an egalitarian relationship, based on turn-taking and exchange, as in many friendship ties (*equality matching*); and a rational analysis of costs and benefits, as in a payment-for-service regime (*market pricing*). Fiske claims that any actual social relationship is constituted by some mixture of these forms, and that socially transmitted interpretive guidelines link these universal forms to specific relationship characteristics in a particular culture.

While relational schemas have advanced our understanding of the nature and variety of dyadic relationships, there is a recognized need to understand the specific ways in which they depend on social situations and give rise to interdependencies among relationships *across* a network. As Haslam (2004, p. 297) has observed, Fiske's relational models theory posits ‘a universal grammar of social relations … out of whose rules and representations the myriad local forms of social life can be generated’. Haslam argues that the categorical nature of the relational models may underpin the complexity of human social organization by facilitating the required coordination of interpersonal obligations, rights and responsibilities. Social roles are important intermediate-level constructs in this account, and are seen as ‘distinctively implemented admixtures’ of relational models, a view that is supported by evidence that relational models mediate the effects of social roles on social cognition. More generally, social roles have been invoked by many theorists to explain interdependencies among multiple relationships across multiple actors (White, Boorman and Breiger, 1976).

It follows from these claims that tie formation processes should depend on the type of tie under consideration. While some ties may be consciously negotiated at a dyadic level, independently of other ties, others may be subject to subtle influences arising from the embedding of the potential tie in a local social setting that comprises ties to and among third-party actors with their own mix of potentially competing and potentially cooperative goals. As a result, Fiske's communal sharing relations, for example, should exhibit much stronger interdependencies across pairs of relation partners than relations of the market pricing kind. Such effects have been empirically demonstrated. For example, Granovetter's influential hypothesis concerning the ‘strength’ of weak ties was based on the distinction between the highly clustered structure of ‘strong’ tie networks and the less structured, more open, spreading character of ‘weak’ tie networks, an hypothesis that has received empirical support (Granovetter, 1982). There is also evidence that different types of network tie can be mutually interdependent and subject to generalized forms of exchange and interlock through ties to third parties (for example, Lomi and Pattison, 2006). Indeed, changes in patterns of cross-network interdependence have been invoked in explanations of social and economic innovation (for example, Padgett and McLean, 2006).

In addition to the interdependencies just described, there is also compelling evidence for the impact of individual actors’ characteristics on the formation of network ties. Tie partners are more likely to share socio-demographic characteristics such as gender, age, ethnicity and religion (for example, McPherson, Smith-Lovin and Cook, 2001). The formation of relationships is also clearly a function of social settings that affect the probability of any two actors having an opportunity to interact (Feld, 1981). The psychological literature on relationship formation emphasizes the importance of more psychological

similarities among potential tie partners, a premise that has been extended by Robins and Boldero (2003) to include a comparison of potential tie partners' aspirations and obligations.

Taken together, these structuring influences on tie formation can be seen as operating at multiple levels, with broad socio-demographic factors at work on a larger scale, and more micro-social and psychological factors at work at more local scales. Whereas the broader factors can often be regarded as exogenous influences on tie formation, the more micro-level factors are usually best seen as endogenous, with the tie formation processes for one pair of actors having consequences for tie formation among their network partners, and their partners, and so on. While some of these interdependencies may operate outside the awareness of individual actors, there are also circumstances in which actors seek out institutional settings and particular relationships precisely for the strategic 'networking' opportunities that they provide. Moreover, expected interdependencies can themselves be countered: Padgett and Ansell (1993) coined the term 'robust action' to describe behaviour that is open to multiple interpretations, and therefore has the capacity to elide third-party influences on tie formation.

How do relationships affect individual outcomes?

An actor's location in a social network is an important aspect of social context and hence potentially plays an important role in determining many types of future behaviour apart from the development of social ties. There are a number of related mechanisms by which such effects might occur (for example, Pattison, 1994). First, network ties serve as a conduit for information, and hence specific network locations can have a dramatic impact on the information or other resources that any one actor possesses (for example, Burt, 2004). This information can itself be subject to subtle filtering effects, such as the suppression of information perceived to be inconsistent with shared understanding at a community level (Lyons and Kashima, 2003). Second, network ties can influence actors' understanding of relationships among others as well as expectations about the future behaviour of others. For example, they are likely to be more certain (and more accurate) in judging the relationships involving their network partners and, to a lesser extent, their network partners' partners (Kumbasar, Romney and Batchelder, 1994). Finally, social influence effects may be brought to bear as actors weigh up – consciously or unconsciously – the views of others in forming or modifying their own beliefs (for example, Friedkin, 1998; Robins, Pattison and Elliott, 2001).

Implications for model-building

Models for network structure and network evolution almost certainly need to accommodate many of the exogenous and endogenous influences on tie formation just described. An appropriate model class is the exponential family of random graph models that was first introduced by Frank and Strauss (1986), building on the general formulation of statistical models for interacting systems of variables by Besag (1974). The use of principled approaches to specifying potential tie interdependencies (Pattison and Robins, 2002) has led to models that yield impressive fits to even large observed network structures (for example, Goodreau, 2006). Interestingly, these models combine a Markov dependence assumption (that potential network ties with an actor in common are dependent, conditional on the state of all other

potential ties in a network) with a ‘longer range’ form of assumed dependence (Snijders et al., 2006) in which, in some circumstances, ties involving discrete pairs of actors are also conditionally dependent. The necessity of these assumptions in many empirical contexts (Robins et al., 2006) suggests that interdependencies among ties within local social contexts are indeed an important influence on observed network forms.

Models for individual states and choices (including beliefs and actions) may likewise need to accommodate subtle network-based interdependencies among the states and choices of their tie partners. Robins, Pattison and Elliott (2001) have developed a general social influence modelling framework for this purpose, akin to the network modelling framework just mentioned.

It is important in the context of the question posed earlier – whether network structures and processes can be explained in terms of the rational activity of self-interested individuals, or whether there are extra-rational processes at work in the formation and impact of networks – to develop observational designs and analytic methods that allow us to regard tie formation and network processes as hypotheses to be tested rather than unquestioned assumptions. We might speculate that different types of networks and different types of social processes are likely to lead to different conclusions about these guiding hypotheses, and that, as a consequence, models should continue to be developed from multiple perspectives and to be grounded in careful empirical analysis. Finally, it is worth noting that the development of methods for estimating models from longitudinal observations may prove particularly helpful in sifting among alternative approaches to model building, not just for models of network evolution and social influence dynamics but also for new approaches to modelling the co-evolution of networks and behaviour (Snijders, Steglich and Schweinberger, 2006).

See Also

- emergence
- learning and information aggregation in networks
- mathematics of networks
- network formation
- social interactions (empirics)
- social interactions (theory)
- social networks in labour markets
- psychology of social networks

Bibliography

Anderson, S. and Chen, S. 2002. The relational self: an interpersonal social-cognitive theory. *Psychological Review* 109, 619–45.

Baldwin, M. 1992. Relational schemas and the processing of social information. *Psychological Bulletin* 112, 461–84.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* 36, 96–127.

Burt, R. 2004. Structural holes and good ideas. *American Journal of Sociology* 110, 349–99.

Cook, K. and Whitmeyer, J. 1992. Two approaches to social structure – exchange theory and network analysis. *Annual Review of Sociology* 18, 109–27.

Feld, S. 1981. The focused organization of social ties. *American Journal of Sociology* 86, 1015–35.

Fiske, A. 2004. Relational models theory 2.0. In *Relational Models Theory: A Contemporary Overview*, ed. N. Haslam. Hillsdale, NJ: Lawrence Erlbaum.

Frank, O. and Strauss, D. 1986. Markov graphs. *Journal of the American Statistical Association* 81, 832–42.

Friedkin, N. 1998. *A Structural Theory of Social Influence*. New York: Cambridge University Press.

Goodreau, S. 2006. Advances in exponential random graph (p^*) models to a large social network. *Social Networks* (forthcoming).

Granovetter, M. 1982. The strength of weak ties: a network theory revisited. In *Social Structure and Network Analysis*, ed. P. Marsden and N. Lin. Beverly Hills, CA: Sage.

Haslam, N. 2004. Four grammars for primate social relations. In *Evolutionary Social Psychology*, ed. J. Simpson and D. Kenrick. Hillsdale, NJ: Lawrence Erlbaum.

Holmes, J. 2000. Social relationships: the nature and function of relational schemas. *European Journal of Social Psychology* 30, 447–96.

Jackson, M. 2005. A survey of models of network formation: stability and efficiency. In *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders. New York: Cambridge University Press.

Kumbasar, E., Romney, A. and Batchelder, W. 1994. Systematic biases in social perception. *American Journal of Sociology* 100, 477–505.

Lomi, A. and Pattison, P. 2006. Manufacturing relations: an empirical study of the organization of production across multiple networks. *Organization Science* 17, 313–32.

Lyons, A. and Kashima, Y. 2003. How are stereotypes maintained through communication? The

influence of stereotype sharedness. *Journal of Personality and Social Psychology* 85, 989–1005.

McPherson, M., Smith-Lovin, L. and Cook, J. 2001. Birds of a feather: homophily in social networks. *Annual Review of Sociology* 27, 415–44.

Padgett, J. and Ansell, C. 1993. Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology* 98, 1259–319.

Padgett, J. and McLean, P. 2006. Organizational invention and elite transformation: the birth of partnership systems in renaissance Florence. *American Journal of Sociology* 111, 1463–568.

Pattison, P. 1994. Social cognition in context: some applications of social network analysis. In *Advances in Social Network Analysis in the Social and Behavioral Sciences*, ed. S. Wasserman and J. Galaskiewicz. Newbury Park, CA: Sage.

Pattison, P. and Robins, G. 2002. Neighbourhood-based models for social networks. *Sociological Methodology* 32, 300–37.

Robins, G. and Boldero, J. 2003. Relational discrepancy theory. *Personality and Social Psychology Review* 7, 56–74.

Robins, G., Pattison, P. and Elliott, P. 2001. Network models for social influence processes. *Psychometrika* 66, 161–90.

Robins, G., Snijders, T., Wang, P., Handcock, M. and Pattison, P. 2006. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks* (forthcoming).

Snijders, T. 2001. The statistical evaluation of social network dynamics. In *Sociological Methodology 2001*, ed. M. Sobel and M. Becker. Boston and London: Basil Blackwell.

Snijders, T., Steglich, C. and Schweinberger, M. 2006. Modeling the co-evolution of networks and behavior. In *Longitudinal Models in the Behavioral And Related Sciences*, ed. K. van Montfort, H. Oud and A. Satorra. Hillsdale, NJ: Lawrence Erlbaum.

Snijders, T., Pattison, P., Robins, G. and Handcock, M. 2006. New specifications for exponential random graph models. *Sociological Methodology* 36, 99–153.

White, H., Boorman, S. and Breiger, R. 1976. Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology* 87, 517–47.

How to cite this article

Pattison, Philippa, Garry Robins and Yoshi Kashima. "psychology of social networks." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000352>
doi:10.1057/9780230226203.1359

The New Palgrave Dictionary of Economics Online

public capital

David A. Aschauer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews recent research emphasizing the potential importance of public capital (or infrastructure) to aggregate economic performance, and provides a survey of empirical estimates of the productivity of public capital and of the impact of public capital investment on economic growth.

Keywords

Cobb–Douglas functions; endogenous growth; long-term economic growth; neoclassical growth theory; output elasticity of public capital; productivity growth; public capital; public finance; returns to public capital; total factor productivity

Article

Public capital (or often ‘infrastructure’) encompasses the publicly provided capital facilities which form the basis for private sector economic activity.
Empirically, public capital typically is defined as a net (of depreciation) stock of non-military structures and equipment and is often decomposed into *core* public capital (consisting of transportation facilities – such as streets and highways, mass transit, rail, and airports, water and sewer systems, and electrical and gas facilities), and *other* public capital (comprising educational structures, public hospitals, courthouses and the like).

The productivity of public capital

Beginning at the end of the 1980s, a significant research effort has focused on estimating the contribution of public capital to macroeconomic performance. The research initiative seems to have been the result of the recognition of certain facts about public capital expenditures in the United States. First, infrastructure capital accumulation, when expressed as a fraction of output, began to decline

toward the end of the 1960s and, as a result, was seen as a potential factor in explaining the productivity growth slowdown of the 1970s and 1980s. Second, during the same period, the United States devoted a smaller share of output to infrastructure than did other industrialized economies (such as those in the Group of Seven), which was taken as possible force in explaining the relatively low rate of productivity growth in the United States vis-à-vis other countries such as Japan and Germany.

The first stage of the research effort centred on estimating the contribution of infrastructure to private sector productivity, where infrastructure is taken as another factor of production, along with private capital and labour, in an aggregate production function of the form

$$Y = A \cdot F(L, K, K^G)$$

where Y denotes the aggregate level of economic output, A an index of total factor productivity, L, the labour force or employment, K private capital (usually restricted to business fixed capital), and K^G the stock of public capital. The basic goal of the research was to ascertain the value of the output elasticity of public capital

$$\varepsilon = \frac{K^G}{Y} \cdot \frac{\partial Y}{\partial K^G}$$

in order to determine the ‘productivity of public capital’.

The early empirical results, typically employing level data in estimating a Cobb–Douglas production function, indicated (strikingly) high elasticity estimates, in the range of 0.25 to 0.50 for the United States and even higher for countries such as Canada and Sweden. These elasticity estimates, in turn, implied very high rates of return to public capital investments which some took as implausible. For example, Gramlich (1994) used Aschauer's (1989) elasticity estimate of 0.39 to generate an estimate of the marginal productivity of public capital in the range of 0.70 to 1.00, which, in his view, was implausible since it implied that investments in government capital generate enough extra output to pay for themselves in a year.

Later, a number of researchers estimated the production function using first-differenced data, arguing that the initial results were ‘spurious’ because (*a*) variables such as output and public capital were first-order integrated series and (*b*) the production function did not serve as a cointegrating relation between output and the various factors of production (including public capital). These studies (for example, Tatom, 1991) often generated much lower, and less reliable, estimates of the output elasticity of public capital.

Recently, Kamps (2004) has developed new estimates of public capital stocks for 22 OECD countries over the period 1960–2001, and has estimated the output elasticity of public capital. The point estimates are positive in 20 of 22 cases and statistically significant in 12 of 22 cases. A panel regression

employing first-differenced data leads to a reasonable elasticity estimate equal to 0.22 which leads the author to conclude that public capital is productive on average in the countries comprising the OECD.

Public capital and economic growth

The finding that public capital is productive is not, in and of itself, evidence that increasing public capital investment will raise long-term economic growth. There are at least three considerations which must be addressed. First, there is the question of whether a permanent increase in public investment will induce a permanent or transitory increase in growth. The traditional neoclassical growth model predicts that an increase in national savings and investment rates will have only a transitory effect on growth; more recent endogenous growth models, on the other hand, would predict permanent effects. Second, given the level of national savings, the effect of public investment on economic growth depends not just on a positive output elasticity of public capital, but on the relative marginal productivities of private and public capital; an increase in public investment at the expense of public investment will raise or lower the economic growth rate depending on whether the marginal product of public capital exceeds, or is exceeded by, the marginal product of private capital. Third, the effect of public capital on economic growth will depend on the method of public finance – whether by current taxes, debt, or (potentially) money creation.

One approach which allows tentative answers to all three questions is that of Aschauer (2000), who extends the Barro (1990) model of productive government spending to explicitly include public investment. This model, which assumes (*a*) that public investment is debt-financed and (*b*) a production function which displays constant returns to scale across private and public capital stocks (per worker) generates endogenous growth in per worker output at the rate

$$\gamma_y = \frac{1}{\sigma} \cdot \left[(1-\tau) \cdot (1-\varepsilon) \cdot \left(\frac{K^G}{K} \right)^{\varepsilon} - \rho \right]$$

where y is the level of output per worker, $(1/\sigma)$ the intertemporal elasticity of substitution, τ the tax rate necessary to service the public debt associated with public capital, and ρ a rate of time preference. Evidently, increases in the tax rate lower economic growth, while increases in the ratio of public capital to private capital raise economic growth. It turns out that increases in public capital will raise or lower economic growth depending on whether the tax rate is lower or higher than the output elasticity of public capital – that is, there is a nonlinear relationship between public capital and growth and an ‘optimal’ level of public capital. Using US state level data, Aschauer finds robust evidence that the relationship between public capital and growth is, indeed, nonlinear and that public capital is underprovided – that is, the ‘optimal’ ratio of public capital to private capital is in the range of 0.60 while the actual average ratio equals 0.44. As a consequence, a ten per cent increase in the public capital ratio is estimated to raise economic growth by approximately one percentage point per year.

See Also

- monetary and fiscal policy overview
- optimal fiscal and monetary policy (without commitment)

Bibliography

Aschauer, D.A. 1989. Is public expenditure productive? *Journal of Monetary Economics* 23, 177–200.

Aschauer, D.A. 2000. Do states optimize? Public capital and economic growth. *Annals of Regional Science* 34, 343–64.

Barro, R.J. 1990. Government spending in a simple model of endogenous growth. *Journal of Political Economy* 98, 103–25.

Gramlich, E.M. 1994. Infrastructure investment: a review essay. *Journal of Economic Literature* 32, 1176–96.

Kamps, C. 2004. New estimates of government net capital stocks for 22 OECD countries 1960–2001. Working Paper 04/67, International Monetary Fund.

Tatom, J.A. 1991. Should government spending on capital goods be raised? *Federal Reserve Bank of St. Louis Review* 73(2), 3–15.

How to cite this article

Aschauer, David A. "public capital." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000366> doi:10.1057/9780230226203.1360

The New Palgrave Dictionary of Economics Online

public choice

Gordon Tullock

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

By assuming that voters, politicians and bureaucrats are mainly self-interested, public choice uses economic tools to deal with the traditional problems of political science. Its findings revolve around the effects of voter ignorance, agenda control and the incentives facing bureaucrats in sacrificing the public interest to special interests. The design of improved governmental methods based on the positive information about how governments actually function has been an important part of public choice. Constitutional reforms advocated variously by public choice thinkers include direct voting, proportional representation, bicameral legislatures, reinforced majorities, competition between government departments, and contracting out government activities.

Keywords

agenda control; Arrow, K.; bicameral legislatures; Black, D.; Buchanan, J.; bureaucracy; constitutions, economic approach to; democracy; direct voting; Downs, A.; Niskanen, W.; proportional representation; public choice; public interest; reinforced majorities; single-peaked preferences; social choice; special interests; Tullock, G.; voting

Article

In the 18th and 19th centuries a number of mathematicians (Condorcet, Borda, Laplace and Lewis Carroll) became interested in the mathematics of the voting process; their work was forgotten until Duncan Black rediscovered it (see, e.g., Black, 1958). Black can be called the father of modern Public Choice, which is in essence the use of economic tools to deal with the traditional problems of political science. Historically, economics (political economy) dealt to a very large extent with the choice of government policies with respect to economic matters. Whether protective tariffs were or were not good things would be a characteristic topic of traditional economics and in examining the question, it was assumed, of course, that the government was attempting essentially to maximize some kind of welfare

function for society.

We do not expect businessmen to devote a great deal of time and attention to maximizing the public interest. We assume that, although they will of course make some sacrifices to help the poor and advance the public welfare, basically they are concerned with benefiting themselves. Traditionally economists did not take the same attitude towards government officials, but public choice theory does. To simplify the matter, the voter is thought of as a customer and the politician as a businessman/entrepreneur. The bureaucracy of General Motors is thought to be attempting to design and sell reasonably good cars because that is how promotions and pay rises are secured. Similarly, we assume that the government bureaucracy will be attempting mainly to produce policies which in the views of their superiors are good because that is how their promotions and pay rises are secured.

In all these cases, of course, the individual probably has at least some willingness to sacrifice for the public good. Businessmen contribute both time and money to worthy causes and politicians on occasion vote for things that they think are right rather than things which will help them get re-elected. In both cases, however, this is a relatively minor activity compared to maximizing one's own well-being.

The only surprising thing about the above propositions is that they have not traditionally been orthodox either in economics or political science. Writers who did hold them, like Machiavelli in parts of *The Prince*, were regarded as morally suspect and tended to be held up as bad examples rather than as profound analysts.

Public Choice changes this, but even more important, by using a model in which voters, politicians and bureaucrats are assumed to be mainly self-interested, it became possible to employ tools of analysis that are derived from economic methodology.

As a result, fairly rigorous models have been developed which can be tested with the same kind of statistical procedures that are used in economics, although their data are drawn from the political sphere. The result is a new theory of politics which is more rigorous, more realistic, and better tested than the older orthodoxy.

While the basic thrust of the Public Choice work has been positive (directed towards understanding politics), from the very beginning it has also had a strong normative component. Students of Public Choice might modify Marx to read that ‘the problem is to understand the world so that we can improve it’. Thus the design of improved governmental methods based on the positive information about how governments actually function has been an important part of Public Choice work, and is usually referred to as the theory of constitutions.

Before discussing this, it is necessary to outline briefly related discoveries in four general areas, viz: voters, politicians, the voting process which relates voters to politicians, and the theory of bureaucracy.

We begin with voters. One of the earliest discoveries of the new Public Choice (see Downs, 1957, pp. 207–78) was that a rational voter would not bother to be very well informed about the votes that he cast. The reason is simply that the effect of his vote on his well-being is trivially small (see Tullock, 1967a, pp. 100–14). Apparently voters have always known this, since empirical studies of voter knowledge show them extremely ignorant, but it was something of a revelation to traditional professors of Political Science. Further, this general ignorance of the voter is not symmetrical. The voter is likely to know a good deal about any special interest which he has. Further, organized special interest groups will put effort into propagandizing the voter in such areas. Thus the voter is not only badly informed, but what information he has tends to be biased very heavily in the direction of his own occupation or avocation. The farmer is much more likely to know the views of the candidates on farm programmes than their

views on nuclear war. It could be said that even on the farm programme he is probably not very well informed, just better informed.

One should not exaggerate of course. The voter, simply by living and following current events in newspapers and on television, does acquire a certain amount of general information about politics. Not much of it seems to stick, however, and in any event it is very heavily affected by temporary fads. It should also be emphasized that some kinds of special interests of the voter are not in any real sense selfish. For example, in the USA many people are influenced in their vote by such institutions as Common Cause and Liberty Lobby and make voluntary cash contributions to them. Clearly, this is an expression by those people of their interest in good government, even though the two groups define this in a radically different way. There is no doubt, however, that a well organized special interest is apt to have more impact on any specific issue than either the general media or so-called public interest groups like Common Cause or Liberty Lobby, even though in the very long run, considering what one might call the ‘general mystique’ of government, the media are very important.

Consider next the politician. A politician is a person who makes a living by being elected by voters of the kind described above. Further, many politicians are themselves voters as, let us say, members of the House of Representatives. While in the latter capacity, although it is not true that politicians’ information is as bad as that of the voter, a similar effect is still at work. An individual member of the House of Representatives or the House of Commons who switches one hour a week from general study of the issues on which he must vote to constituency service will normally reduce only trivially the quality of the legislation as it affects his constituency. On the other hand, by so re-allocating his time, he may materially improve his relations with his electors. Thus we would expect that politicians will be less well-informed on general matters than we would like.

This is simply one example of a large number of cases in which politicians’ behaviour is not necessarily that which maximizes the public welfare: they vote in Congress and seek public positions in terms of what they think the voters *will* reward, not in terms of what they think the voters *should* reward. Since a politician knows that his constituents are badly informed, these two positions can be radically different. Nevertheless, if we are believers in democracy, which literally means popular rule, then the government should do what the people want and not what some wiser person feels that they should want. In any event, ‘in order to be a great Senator, one must first of all be a Senator’.

Obviously the cost to the public of this kind of behaviour is quite considerable. It is particularly so when we think of the investment of resources and influence in the government which are, to a considerable extent, wasted. However, if we contrast functioning democracies with the other types of government which we observe, we are not likely to feel that democracies are markedly less efficient.

We now turn to the voting process, which connects the public to the politicians and the latter to the actual policy outcomes. Uninformed people think that this is basically a trivial problem, you simply count the votes. Unfortunately, this does not follow, even though the author of this essay is one of the few Public Choice theorists who regards the problems to be discussed next as being possibly illusory. Condorcet, Borda, Laplace and Lewis Carroll and, in the 20th century, mathematical economists like Black and Kenneth Arrow discovered a set of mathematical problems sufficiently difficult to be taken as proof that democracy is either an illusion or a fraud. Basically, if we assume that all individuals can order various policy proposals, producing a personal ranking from top to bottom (indifference between alternatives being permitted) and that these orderings differ from person to person (and do not fall into a set of narrowly specified and rather unlikely patterns), then one of the following three phenomena can

occur under any conceivable system of voting:

1. 1. Endless cycling with A beating B and B beating C then C beating A.
2. 2. An outcome which is dependent on the order in which the various proposals are voted on. (It should be pointed out in this connection that if this is so, and the people are well informed, voting on the order of voting reproduces the same problem.)
3. 3. A situation in which the choice between alternative A and alternative B depends on whether alternative C (which in itself has no chance of winning) is or is not entered into the voting process. Most legislatures follow procedures which fall under the second of these possibilities.

If there is a possibility of arranging all of the alternatives in a single dimension with individuals having an optimal point and their preferences falling away monotonically as one moves away from that optimal point in either direction (single peakedness), then the problem is avoided. Unfortunately, most choices involve policies that differ from each other in more than one dimension and so cannot be arrayed in such a one-dimensional continuum. Furthermore, voting on them one aspect at a time reintroduces the second of the problems above. Nevertheless, the assumption of single peaks (whose validity is probably due to voter ignorance) has been successfully used in much empirical work.

While there is no doubt about the mathematical accuracy of the proofs of the above propositions, the real problem is whether they are of great practical significance in voting. Unfortunately, this turns out to be an extremely difficult question whose solution is unlikely to be found in the near future. In essence there are two possibilities when we observe such voting bodies as the House of Representatives and look at the outcome. The first is that the outcome is essentially random, that is, matters are taken up in some order, that order determines the voting outcome and the members of the House do not realize that they could then change that outcome by changing the order in which the propositions are voted on. This possibility would imply that luck plays an immense role in democracy.

The alternative is to say that the outcome is manipulated by somebody who understands the situation and who has control over the agenda. The House majority leader, or the chairman of the Rules committee, is sometimes suggested as that person. This implies that we really have a dictatorship, one that is well concealed.

In my opinion, the indeterminacy thrown into the outcome by these propositions of social choice theory is actually quite small in practical terms. Thus the Chairman of the House Rules Committee may be able to change an appropriation bill by, say, one hundred thousand dollars, but not by an amount which (given the size of these appropriations) is particularly relevant (see Tullock, 1967b). Among Public Choice theorists mine is a minority point of view. The majority, although it is deeply concerned about these problems, tends to ignore the implications of its point of view on the desirability of democracy as a form of government.

Empirical evidence has clearly demonstrated that agenda control can to some extent affect the outcome. This of course is going to surprise nobody. One does not need the complex mathematics of voting in order to realize that those members of any assembly who are in a position to control the order upon which things are voted have power. Similarly the control of what propositions are actually put before the voters can have considerable impact on the outcome. The demonstration of the empirical impact from agenda control, however, does not really support the theorems given above. Of course, we cannot say that the failure to find clearcut proofs that the outcome in a democracy is essentially either random or

fraudulent (as would be implied by the mathematical work on voting) proves that it is not. The problem is difficult and subtle and in the present state of our knowledge must be left for further research. Meanwhile, we all go on with faith that the voting process produces an acceptable outcome even though mathematical investigation raises grave doubts.

Turning now to the theory of bureaucracy, once again Public Choice thought has worked a revolution. The traditional view was either that bureaucrats followed the orders of their political superiors or alternatively that they simply did what was right. Public Choice theorists, following the work of Tullock (1965), Downs (1967) and Niskanen (1971), believe that these are not proper statements about the bureaucrats' motives, although to some extent the bureaucrats do attempt to do what is right – including obedience to the views of their superiors. However, in modern societies where civil service legislation makes it all but impossible for the superiors either to dismiss them or even to reduce their salaries, the degree to which the bureaucrats are so compelled is moderate. Furthermore, in most civil service situations the power of a political appointee to reward his inferiors by promotion is very much restricted. Promotion decisions are to a considerable extent controlled by both legal and public-relations considerations which may compel a superior to promote someone whom he actually thinks has been sabotaging his policy.

While this is a characteristic of most modern civil service structures, there is no law of nature which says that government should be organized in this way. Traditionally, higher officials have been free to promote, demote or dismiss this subordinates. Even here, however, the fact that the higher official cannot possibly know everything that is going on at the lower ranks means that his control gradually diminishes as one moves away from his position down the pyramid of ranks.

For example, in the USA it was recently discovered that it is not possible for the Secretary of Defense to know the specifications which a civil servant, located at a vast distance down the pyramid, produced for a new coffee pot for military aircraft. In this case, the civil servant who specified a coffee pot capable of withstanding a crash that would kill the entire crew of the plane was neither dismissed nor even reprimanded. Indeed the newspapers that reported the story did not even mention his name, but instead concentrated on the Secretary of Defense. In 1870 a military procurement agent who make a mistake like this (and which got into the newspapers) would have found it necessary to hunt for a new job within an hour or so.

Basically the average employee in a bureaucracy is interested in retaining his job and gaining promotion and for this purpose wants to please his superiors. Under the old-fashioned system where he had little job security, and where promotion was determined strictly by his superiors, there was considerable pressure on him. In present circumstances, where to all intents and purposes he cannot be dismissed and where even his promotion is to some extent protected from political intervention by his superiors, this pressure is less important. However, even in a different case, in which he did indeed want to please his superiors, this would not necessarily lead to activity which is in the public interest. That would depend on the political situation of the party or individual who at that time was in control of his branch of the government.

This attenuation of control, in which much of what is done by lower-ranking officials is simply unknown to those of higher rank, is characteristic of all bureaucracies. There are however various ways by which the higher ranks can become, to some extent, aware of what is being done by the lower ranks. Undoubtedly the most efficient of these is simply an accounting system. In the case of a private company, whose motive is making money, the accounts do a reasonably good job (no more) of

signalling what the various lower ranking officials are contributing to that goal. When we turn to government, however, we have the combination of a set of objectives that are either vague or not clearly specified, and a situation where there is no accurate way of measuring the contribution of each person to those objectives. Under such circumstances, control is much more severely attenuated.

When we have a civil service structure which separates the individual from much of the control power of his superiors, the problem is even more severe. Whether an individual bureaucrat works hard or not, prepares himself or herself well or not, is largely a matter of individual choice. As a rough rule of thumb, those people who do work hard and prepare themselves well are those people who have their own idea of what government should do in their particular division and work hard at that. In a way they are hobbyists. It should be said however, that their hobby is normally motivated by a desire on their part to maximize what they think is the public good. In other words, they are usually well-intentioned individuals who can be criticized only in that their idea of the public good may or may not coincide with that of their superiors. If it does not coincide, this does not prove that they are wrong and the superiors right, but it does mean that the government is not apt to follow a coordinated policy. In times past, it used to be normal to refer to the US Department of State as ‘a loose confederation of tribal chieftains’. The phrase is not used any more, but as far as I can see this is only because the confederation itself has broken down.

Bureaucrats normally have several private motives. One is, of course, simply not to work too hard – a motive which does not seriously affect the hobbyist described above. Another is to expand the size of one's own department and in the process of so doing, being willing to go along with the expansion of all the rest. A third is to improve the ‘perks’ that accompany the particular position (see Migue and Balageur, 1974).

Note that this is not intended as criticism of the bureaucrat. We would expect anyone who is given the kind of opportunities that are given to bureaucrats to do more or less what they do. However, the consequence is that large bureaucracies tend to grow larger, tend as they grow larger to follow less in the way of integrated policies and more in the way of policies that develop in the lower reaches of the pyramid, and tend in fact not to work terribly hard (see Bennett and Orzechowski, 1983).

The problem is multiplied when bureaucracies become very large, because the members of the bureaucracy can vote. Furthermore, empirical evidence (see Bennett and Orzechowski, 1983) shows they vote more frequently than non-bureaucrats. Thus their percentage in the voting population is somewhat larger than their percentage in the actual population (see Frey and Pommernhe, 1982). Thus, the political superior must consider the people working for him as in part his employers rather than his employees. He may not be able to fire them, but in the mass they can fire him. Altogether, the system is not well designed and does not work very well.

So far we have been talking about Public Choice and what has been learned, but not of the lessons of a normative nature that have been drawn i.e. the theory of constitutions. It is to this that I now turn.

Not all students of Public Choice favour the same reforms in each area. Further, some have not specifically said what reforms they would prefer because they believe that not enough is yet known about the process to be able to suggest improvements. Nevertheless, there are several rather general propositions which most students would agree upon as ways of improving the functioning of government. In a discussion as brief as this, it is not possible to include all the differences of opinion and all the modifying clauses which would be appended to each suggestion for reform. Thus the reader should not assume that everyone studying Public Choice agrees with all the propositions which follow.

To begin with the voter, no student of the subject has any idea of how to improve the voters' information. With respect to voting itself there have been some proposals for improved voting methods, but no widespread support exists for any particular improvement. In spite of this, I think it can be said fairly that most students would like to see voters vote more than they do now, favouring more direct voting on issues, and legislatures with larger membership (so that the connection of an individual voter and his representative is closer).

The basic desire to give voters more control of the mechanism is not based on any false idea of how well the voters are informed. It is simply that the voters are the only people in the whole process who do not have an element of systematic bias in their decision process. They may be badly informed, but what they want is their own well-being. The well-being of its citizens should be the objective of the state. When we turn to other parts of the government invariably we find at least some conflict between the interests of the officials and the interests of the average man. Thus increasing the average man's control is not particularly likely to improve the efficiency of the government using some abstract definition of efficiency. But it is likely to make the government more in accord with the preferences of the common man; i.e. it brings us a little closer to the objective of popular rule which is supposed to be what democracy is about. Those who do not favour popular rule would not regard this as desirable, but there are few elitists among the students of Public Choice.

The actual decision-making procedures used in the legislatures have been widely discussed and some proposed improvements command wide acceptance. First, many would like to have at least one house of the legislature elected by proportional representation. Secondly, Buchanan and Tullock's arguments in *The Calculus of Consent* (1962) for bicameral legislatures have generally been accepted. The further suggestion there that more than a simple majority is desirable for most legislation is seldom directly criticized, but is not so widely approved. The argument that this higher-than-majority requirement would change the structure of the log-rolling process in a favourable way has seldom been directly criticized, but the asymmetrical effect of such a rule (i.e., the status quo is retained unless a reinforced majority can be obtained to change it) offends some people.

Turning to the bureaucracy, there is much more agreement on reform. First, that a bureaucracy should be brought more firmly under the control of the political leaders is, I think, uniformly accepted. The dangers of this are recognized – but there are various ways in which the higher officials could be given the right to discipline civil servants while still reducing their power to fill the government with their cousins.

Apart from such straightforward proposals for changes in the personnel structure there are other ways of putting pressure on the government. The first is to work some competition into the system. Currently, not only do most government departments have a monopoly over whatever function they perform, but almost every proposal to increase the efficiency of government takes the form of eliminating what little competition has popped up. Competition between government departments should be encouraged rather than discouraged.

Finally, it may be possible to 'contract out' government activities or literally transfer them wholely to the market. The mere threat of this will frequently lower the cost of government activity. Having several private companies bidding for a government service, however, is better.

It can be seen that at the concrete level, those who study Public Choice have been able to provide more in the way of suggestions for reform within the bureaucratic structure than in the higher level parts of democracy where the voters control the legislature, and the legislature and executive then control the

bureaucracy. This is unfortunate but not surprising. Nevertheless, there are suggestions for improving the whole structure of government and with time, it is hoped, there will be both more ways of making improvements and better scientific evidence that the ‘improvements’ are indeed improvements. Public Choice is a new and radical approach to government, but its firm foundations in economic methodology mean that we have more confidence in its accuracy than with most new ideas. Further, it has by now been empirically tested very thoroughly. Government is the solution to some problems and the source of others. Public Choice shows strong promise of being able to reduce significantly the difficulties we now have with democratic government.

See Also

- Black, Duncan
- constitutions, economic approach to
- social choice

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: John Wiley & Sons.

Bennett, J.T. and Orzechowski, W.P. 1983. The voting behavior of bureaucrats: some empirical evidence. *Public Choice* 41(2), 271–84.

Black, D. 1958. *The Theory of Committees and Elections*. Cambridge: Cambridge University Press.

Buchanan, J. and Tullock, G. 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press.

Downs, A. 1957. *An Economic Analysis of Democracy*. New York: Harper & Row.

Downs, A. 1967. *Inside Bureaucracy*. Boston: Little, Brown.

Frey, B.S. and Pommernhe, W.W. 1982. How powerful are public bureaucrats as voters? *Public Choice* 38(3), 253–62.

Migue, J.L. and Balageur, G. 1974. Towards a general theory of managerial discretion. *Public Choice* 17, 27–43.

Niskanen, W. 1971. *Bureaucracy and Representative Government*. Chicago: Aldine-Atherton.

Tullock, G. 1965. *The Politics of Bureaucracy*. Washington, DC: Public Affairs Press.

Tullock, G. 1967a. *Toward a Mathematics of Politics*. Ann Arbor: University of Michigan Press.

Tullock, G. 1967b. The general irrelevance of the general impossibility theorem. *Quarterly Journal of Economics* 81, 256–70.

Tullock, G. 1981. Why so much stability. *Public Choice* 37(2), 189–202.

How to cite this article

Tullock, Gordon. "public choice." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000240> doi:10.1057/9780230226203.1361

The New Palgrave Dictionary of Economics Online

public debt

James M. Buchanan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Classical principles of public debt limited debt financing to non-recurrent, extraordinary or temporary demands. Keynesian macroeconomics, viewing budget deficits as the only means of financing demand-increasing deficits during depressions, overlooked the exchange between government and lenders in debt-financed public expenditure. In the post-Keynesian 1970s and 1980s, governments explicitly used debt to finance ordinary public consumption, including transfers, which was equivalent to a destruction in national capital value and raised the prospect of default. Fiscal responsibility demands that the classical principles of public debt must eventually return to general acceptance.

Keywords

assets and liabilities; Barro, R.; Buchanan, J. M.; budget deficits; capital value; default; fiscal responsibility; Keynesian revolution; neutrality theorem; new classical macroeconomics; public debt; Ricardian equivalence theorem; Ricardo, D.; taxation

Article

Public debt (government debt) is a legal obligation on the part of a government to make interest and/or amortization payments to holders of designated claims in accordance with a defined temporal schedule. Public debt is created through government borrowing from individuals, corporations, institutions, and other governments. Borrowing is part of a bilateral exchange process in which lenders transfer funds to government and government, in turn, transfers to lenders designated instruments that represent claims on government revenues over a series of periods subsequent to that in which the borrowing occurs. In simple balance-sheet terms, public debt is a liability item on the government's account, and an asset on the combined accounts of the holders of the debt instruments.

In its essential respects, public debt is not different from the debt of individuals or non-governmental institutions. The positive analysis is equivalent over the several settings, and the normative principles for

the use of debt are the same. These two propositions are not universally accepted by economists, whose understanding and analysis of established classical principles were eroded in the emergence of the macroeconomic mind-set of the postKeynesian era. Because of the continuing confusion and ambiguity in the basic economic theory of public debt, extended discussion is required on what should otherwise seem quite elementary analytics.

I Taxes, money and public debt

In order to finance spending programmes (including transfers) government must first secure revenues. Three means are available to governments with independent monetary systems: taxation, money issue, and public debt. Only two means are available to subordinate governments without money creation powers or to national governments that tie domestic currencies to external forces in the international economy: taxation and public debt.

It is useful to make two critical distinctions between taxation on the one hand, and public debt on the other. With taxation there is only one possible ‘exchange’ embodied in the combined fiscal process, that effectuated via government between individuals as taxpayers and individuals as beneficiaries of government spending programmes. This two-sided fiscal operation is ‘exchange’ only in the aggregative sense that persons in the community secure benefits from the taxes paid by persons, whether or not the taxpayer and beneficiary groups are overlapping. However, even if the fiscal process satisfies the aggregative and the individualistic efficiency standards such that all persons pay tax-prices, at the relevant margins, equivalent to public-goods benefits, coercion is required to implement the solution. The basic fiscal ‘exchange’ is not, and cannot be, voluntary.

With public debt issue, by comparison, *two* ‘exchanges’ are involved in the combined fiscal operation, one, the political ‘exchange’ analogous to that in taxation, and the other the whole set of privately negotiated and wholly voluntary exchanges between government and those who lend funds to government. Failure to recognize the double set of exchanges that public borrowing combined with spending of the proceeds embodies is the source of major confusion in determining the location of the burden of debt, to be discussed below.

A second critical distinction between taxation and debt issue lies in the temporal difference in the politically determined imputation of the liabilities that are made necessary by the fact of government spending. With taxation, these liabilities are imputed to those persons and institutions that make the funds available directly to government in the period of the spending operation. With public debt, by contrast, there is no current or initial-period imputation of fiscal liability. Revenues are secured from those who lend voluntarily, who do so in exchange for promises of future period interest and amortization payments, and not in ‘political exchange’ for the benefits of spending programmes, even indirectly. With government borrowing (debt issue) the ultimate fiscal liability made necessary by the spending programme in the initial period is postponed. This liability is placed, in the aggregate, on taxpayers in periods subsequent to that in which the debt is issued. The aggregate liability is not, however, imputed or assigned to individuals or to groups of individuals. The postponement of liability implies, therefore, postponement of payments for public programmes.

II The Ricardian theorem on the equivalence between taxation and government borrowing

David Ricardo (1817, 1820) advanced a theorem to the effect that taxation and government borrowing are logically equivalent. This Ricardian theorem was rediscovered by macroeconomists in the 1970s, notably by Robert Barro, and it became an important element in the ‘new classical macroeconomics’ of that period.

On its face, the theorem seems to reject the critical distinctions between taxation and debt discussed above. In what respects are taxation and government borrowing alike rather than different? Both extract revenues from private citizens and transfer these to government for spending on public programmes. The basic Ricardian logic does not reject the difference in status between the taxpayer, who faces governmentally imposed coercive levies, and the bond purchaser, who voluntarily transfers funds to government in private exchange for future-period interest and amortization payments. The theorem does, however, reject the second distinction noted above, that which involves any postponement of the costs of public spending. In the Ricardian model, individuals recognize that any issue of debt by government embodies a commitment to meet payments in future periods. These payments can be converted into individualized shares in the aggregate liability, discounted, and capitalized into a present-value measure, which can then be reckoned as a liability item on individual initial-period balance sheets. If persons think that they will live forever, or if they have intergenerational bequest motives that cause them to act as if their lives are infinite, the liability items, summed over all persons, will just offset the value of the debt in the balance sheets of those who hold debt instruments.

If the equivalence theorem is valid, there are important macroeconomic consequences. If persons treat government debt equivalently with taxation in all respects other than the actual timing of the payments, they will make portfolio adjustments as required by this differential timing. When debt is issued, persons will, knowing that payments must be made in future periods rather than currently, put aside some funds to facilitate such payments. There will be no difference between tax and debt financing in their effect on consumption and investment spending. Individuals, as taxpayers-citizens, will when debt is issued, increase savings to allow a share of the future-period debt obligations to be met. However, this increase in saving will equal the full value of the debt only if the governmental outlays take the form of ideally efficient transfers. This *neutrality theorem*, which we may associate with Barro, is more restrictive than the Ricardo theorem. If the governmental outlays are made for the provision of real goods and services, the neutrality theorem may not hold although debt financing and tax financing of these outlays may still exert equivalent effects (the Ricardo theorem).

The Ricardo equivalence theorem (along with the stronger neutrality theorem) is best evaluated as an extreme model of rational individual behaviour under idealized sets of circumstances. Ricardo himself recognized that persons did not, in fact, treat taxation and debt in the same way. All persons do not act as if they live for ever; persons differ in age as well as interest in future-period tax obligations. Further, taxes are not lump sum, interpersonally or intertemporally. More importantly, there is no assignment of present-value liability among persons such as would allow portfolio adjustments to be made in the manner postulated in the equivalence theorem setting. Even if individuals do recognize that public debt does embody future-period tax liabilities, they cannot reduce this aggregate to individualized shares in any plausible reckoning.

The central flaw in the equivalence theorem stems from the logic of debt itself, which may be illustrated by analogy with private behaviour. Why does a person borrow? He does so in order to rearrange spending temporally. If borrowing and current payment (the private analogue to taxation) are equivalent, there is no point in the exercise. As an institution, borrowing has as its purpose the adjustment of

spending flows through time. Governments, as agents are citizens, borrow for analogous reasons. There is no *raison d'être* for public debt if this instrument is behaviourally equivalent to taxation.
The empirical evidence gained from straightforward observation of modern politics points clearly toward rejection of the equivalence theorem. Taxation and debt are not treated as identical by voting constituents, as is suggested by the fact that politicians responsive to constituents are not indifferent as to the mix between these instruments. Within broad threshold limits, debt financed outlay arouses less political opposition than tax-financed outlay of comparable magnitude. The observed US Federal deficits of the 1980s could not have been eliminated by tax increases without generating significant political struggle.

III Classical principles of public debt

Both the positive analysis of and the normative precepts for public debt were broadly understood by the classical economists, and these principles were carefully articulated in the dominant theory of public debt developed in the 19th century. There is no essential difference between the government account and the account of an individual or private firm in the classical model. Borrowing is a means of raising revenues that allows the borrower to put off or to postpone payments. It is a means of adjusting spending needs to revenue flows over time; in effect, borrowing allows intertemporal trades to be made.
For the government, as for the individual or firm, there would be no basis for borrowing unless the burden of payment could be delayed in time. By the very meaning of debt, therefore, there must be a shifting forward of burden intertemporally. The ultimate payments for the enhanced spending programme during the initial period when debt is issued must be borne exclusively in later periods.
From this straightforward and indeed simple analysis, normative principles for public debt creation emerge. Resort to debt financing is indicated only with nonrecurrent or extraordinary demands, or requirements for public spending that are expected to be temporary. Traditionally, such demands were associated with war emergencies, and the principles of fiscal prudence dictated that debts accumulated during was periods would be retired when the emergency spending demands were past. In addition to these extraordinary spending justifications for resort to borrowing, government is also within bounds under classical norms when debt is issued to finance genuinely productive capital projects, analogously with private firms making capital investments.
When capital spending is debt financed by government, the principles suggested that a scheme for debt retirement be put in place to insure that the pay-off period corresponds to the income-yielding period from the investment asset.

IV Public debt in Keynesian macroeconomics

The classical principles of public debt were not understood by the pre-classical mercantilists and these principles were also questioned by a few fiscal and monetary expansionists in the pre-Keynesian period. Only with the 'Keynesian revolution' in economic thinking in the middle of the 20th century, however, did a rejection of these classical principles become a 'new orthodoxy'. An analysis of public debt, along with relevant normative implications, came to be dominant during the 1940s and 1950s that sought to contradict basic elements of the classical model.
As noted earlier, the classical principles are starkly simple, and are based on the essential similarity

between the government and the individual account. The Keynesian logic rejected this analogy. Specifically, the argument denied that public debt embodies any shift of burden onto taxpayers in periods of time subsequent to debt issue, a temporal shift that was acknowledged to occur in both private debt and external public debt.

The conclusion that public debt could involve no intertemporal shift of burden emerged from an undue concentration on the macroeconomic aggregates and an overlooking of individual adjustments to macroeconomic instruments. Again, the logic is seemingly quite straightforward. Resources are used up only in periods when spending programmes take place; if government borrows to finance spending on guns, the resources that go into producing these guns must be given up by some persons during the period and not later. With internal public debt, therefore, the burden of war spending could not, by definition, be transferred forward.

The basic flaw in the argument is clear from the discussion in section I. The two-part exchange that debt-financed public spending embodies is overlooked. The argument fails to recognize that those who actually give up claims over resources during the period of the spending do so because they voluntarily exchange funds for promises of interest in future periods. These purchasers of bonds do not, in any sense, ‘pay for’ the benefits of the public spending programme. The fact that these persons may also be members of the community of citizens-taxpayers is irrelevant for the temporal location of burden. Taxpayers in later periods are faced with claims against their incomes that must be met, and which exist only because of the initial-period debt issue. If, indeed, the Keynesian orthodoxy of public debt were valid, economists and finance ministers would have discovered the fiscal equivalent of the perpetual motion machine. No non-voluntary transfers of revenues are required to finance spending in the initial period, and, if there is no burden on future-period taxpayers, the spending that is carried out would have required no burden on anyone.

The Keynesian argument was driven by a stance on policy that viewed public debt as the only means of financing demand-increasing deficits during periods of depression. The primary policy instrument of Keynesian economic policy was the budget deficit, and there was an elementary failure on the part of pro-Keynesian economists to recognize that demand-enhancing deficits could be financed with non-interest bearing money creation. If this macroeconomic objective is the only justification for the creation of budgetary deficits, it becomes totally unnecessary to impose the future-period taxes that debt interest reflects. Money creation in such settings carries with it no future-period burden.

V Public debt and deficits in postKeynesian politics

The Keynesian replacement of classical principles of public debt was never total, and in the late 1950s and early 1960s there was a reemergence of economists’ support for the earlier analysis, along with its normative implications. There was not, however, a ‘paradigm shift’ at all comparable to the earlier overthrow of the classical analysis. Somewhat begrudgingly perhaps, economists of the 1960s came to recognize the deficiency in the simplistic Keynesian logic, but there was no general reaffirmation of classical principle. The discussion of public debt that characterized the 1960s, 1970s and 1980s remained confused by an admixture of the two contradictory models of analysis. Economists seemed to concentrate attention on the secondary and tertiary macroeconomic consequences of debt-financed deficits; their considerable formal skills were directed toward attempts to extend the ancient Ricardian theorem.

While confusion and ambiguity described economists' discussion of public debt, the politicians had learned the Keynesian policy lessons with roughly a two-decade lag. By the early 1960s, the 'old-time fiscal religion', based on adherence to the normative precepts of the classical analysis, had lost its constraining influence. The political leaders of the 1960s and beyond had learned that demand-enhancing deficits may be justified in some economic settings. Their natural proclivities to spend without the levy of taxes on constituents caused them to look on economic settings in a biased or one-sided fashion. The idealized Keynesian policy set – deficits in depression, surpluses in booms – proved to be unworkable in democratic politics.

The regime of apparently permanent debt-financed deficit spending was born. During the 1970s and 1980s, for the first time in modern fiscal history, governments explicitly used debt to finance ordinary public consumption outlay, including transfers. This fiscal operation, considered in isolation, is equivalent to a destruction in national capital value. When persons, privately or publicly, abstain from consuming current income, capital value is created. When persons, privately or publicly, consume more than current-period income, capital value (defined as the discounted present value of anticipated future incomes) is destroyed. For both individuals and governments, resort to borrowing allows a 'using up' of future-period income as a means of increasing current consumption, just as resort to saving allows a 'using up' of current income (in an opportunity cost sense) to increase future-period consumption. Only if borrowing is used to finance genuine capital investment, private or public, will the net effects be intertemporally neutral; only in this case will the capital value of anticipated income be unchanged. With debt-financed public consumption, the present value of anticipated future incomes of persons in the polity is reduced relative to that which might have been maintained in the absence of the combined fiscal operation. The fact that some or all of the debt is held by foreigners rather than citizens does not modify this conclusion.

Consider the case where debt instruments are purchased exclusively by domestic citizens, who may also be future period taxpayers. Securities are purchased voluntarily; hence, there is no change on the asset side of purchasers' balance sheets at the time of debt issue. Purchasers could, alternatively, hold or buy privately issued securities with equivalent yields. On the other hand, the fiscal operation does place debt-interest claims against the anticipated private income flows of citizens as taxpayers. There is a net increase in the present value of liabilities on properly calculated individual balance sheets. This increase in the value of liabilities is, of course, equivalent to the value of the debt instruments. But these two items are *not* offsetting since there are two fully offsetting items on the asset side of the accounts, leaving no net change from this side. The combined fiscal operation necessarily reduces net worth, or capital value, in the economy, so long as the government outlay does not generate anticipated income flows from public assets, a result that is ruled out with pure public consumption.

VI Return to classical principles?

Public debt is a topic in political economy in which the level of understanding experienced serious retrogression over the course of the middle decades of the 20th century. Policy-motivated macroeconomic confusion generated political spillovers that remained in the 1980s. Economists seemed unable to contribute to clarification in analysis, in part because they were reluctant to drop either suprarational models of individual behaviour or erratic manipulation of data. The classical analysis, out of which emerged precepts that offered simple guidelines for governmental fiscal authorities, no longer

commanded widespread support, either among political economists or among politicians and, indirectly, their constituencies. Governments in the 1980s were observed to be financing sizeable shares of their public consumption outlays by interest-bearing debt. Interest outlays made up ever-increasing proportions of total government budgets.

The simple logic of compound interest guaranteed that the budgetary regimes observed in the 1980s were not sustainable. Default on government's debt obligations becomes increasingly attractive to politicians as interest charges mount and as borrowing rates for new issues of debt simultaneously increase. Default on public debt has occurred often in history, both through explicit destruction of real value obligations and by means of inflation.

The ultimate prospects for default may be generally recognized, but the political difficulties in restoring some adherence to classical norms cannot be overlooked. Once debt-financed deficit spending for public consumption came to be an element in the quasi-permanent *status quo*, attempts to restore budgetary balance faced enormous political opposition, as indeed the events of the 1980s demonstrated. To reduce the deficits, and hence merely to reduce the rate of increase in public debt issue, governments must resort to tax increases or to spending cuts, both of which arouse political opposition. Those taxpayers who must bear the burden of continuing debt are, at best, only partially and indirectly represented in the decision structure of democratic politics.

Restoration of the classical principles of public debt seemed unlikely from the temporal perspective of the 1980s. The old-time or pre-Keynesian ‘fiscal religion’ did exert an influence on the behaviour of politicians, and through them, on governments. Public debt, as a revenue-raising instrument, has an appropriate and well-defined use as a means of allowing governments to alter the time stream of payments for extraordinary outlays. There was nothing comparable to a ‘fiscal religion’ in postKeynesian politics. Public debt, as it was actually used in the years after the 1960s, became a mere balancing element between proclivities of politicians to tax on the one hand and spend on the other. The absence of immediate fiscal breakdown was explained by some residual carry over of classical norms, some introduction of a Ricardian-like consciousness of future tax liabilities, and some fear of default risk on the part of prospective lenders. But the situation observed over the decades of the 1970s and 1980s could not have represented temporal stability.

The classical principles of public debt, whether they be labelled as such, must eventually return to general acceptance if the fiscal responsibility of governments is to be maintained. Whether or not this acceptance comes before or after a sequence of default-engendered fiscal crises could not be predicted from the temporal perspective of the middle 1980s.

See Also

- government budget constraint
- public finance
- Ricardian equivalence theorem

Bibliography

Adams, H.C. 1893. *Public Debts*. New York: Appleton.

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1095–117.

Bastable, C.F. 1895. *Public Finance*. 2nd edn. London: Macmillan.

Buchanan, J.M. 1958. *Public Principles of Public Debt*. Homewood, IL: Richard D. Irwin.

Buchanan, J.M. and Wagner, R.E. 1977. *Democracy in Deficit*. New York: Academic Press.

Ferguson, J.M., ed. 1964. *Public Debt and Future Generations*. Chapel Hill: University of North Carolina Press.

Harris, S.E. 1947. *The National Debt and the New Economics*. New York: McGraw-Hill.

Lerner, A.P. 1948. The burden of the national debt. In *Income, Employment, and Public Policy: Essays in Honor of Alvin H. Hansen*, New York: Norton, 255–75.

Leroy-Beaulieu, P. 1912. *Traité de la science des finances. Vol. 2: Le budget et le crédit public*. 8th edn. Paris: Alcan.

Ricardo, D. 1817, 1820. *Works and Correspondence*, ed. P. Sraffa, vol. 1, vol. 4. Cambridge: Cambridge University Press for the Royal Economic Society, 1951.

How to cite this article

Buchanan, James M. "public debt." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000241> doi:10.1057/9780230226203.1362

The New Palgrave Dictionary of Economics Online

public finance

R.A. Musgrave

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Public finance may well be the oldest branch of economics. It concerns not only the effects of fiscal operations on the market but also principles of public sector economics, which address a distinct set of issues and are linked closely to the perspectives of political and social science. This article covers (*a*) expenditure on public goods and transfers, including its macroeconomic effects as analysed by the classical and the Keynesian schools, and (*b*) taxation, including issues of tax equity and tax efficiency, definitions of income, consumption and expenditure, and tax shifting and incidence.

Keywords

ability to pay tax; ad valorem taxation; aggregate demand; Bentham, J.; bequests; budgetary policy; capital budgeting; capital gains taxation; capital measurement; consumption base; consumption taxation; cost–benefit analysis; Cournot, A.; deadweight loss; depreciation; Edgeworth, F. Y.; elasticity of substitution; entitlement; equal absolute sacrifice theory of tax equity; equal marginal sacrifice theory of tax equity; equal proportional sacrifice theory of tax equity; equality; excess burden of taxation; exchange; expenditure base; expenditure tax; externalities; factor mobility; factor pricing; fairness; fiscal location; free rider problem; Hobbes, T.; horizontal and vertical equity; human capital; Hume, D.; income base; income effect of taxation; income subsidies; inflation; intergenerational equity; international capital flows; interpersonal utility comparisons; interpersonal utility interdependence; justice; Keynesianism; land tax; leisure; Lindahl, E. R.; local public finance; marginal rates of substitution; marginal utility of income schedules; market failure; merit goods; Mill, J. S.; Musgrave, R. A.; neoclassical growth models; neoclassical synthesis; net income; opportunity cost; optimal commodity taxation; optimal taxation; paternalism; payroll tax; Physiocracy; Pigou, A. C.; preference revelation; primary distribution; progressive and regressive taxation; propensity to save; property tax; public choice; public expenditure; public finance; public goods; public policy failure; quasi-rent; Rawls, J.; rent; retail sales tax; Ricardo, D.; Samuelson, P. A.; Scandinavian model; shadow pricing; Smith, A.;

social discount rate; social welfare function; subsistence hypothesis; substitution effect of taxation; tax avoidance; tax base definition; tax burden; tax efficiency; tax equity; tax expenditures; tax incidence; tax prices; tax shifting; taxation of corporate profits; taxation of income; Tiebout hypothesis; transfers; utilitarianism; value-added tax; voting rules; Wagner's law; Wicksell, K.

Article

Concern with public finances may well be the oldest branch of economics. Statesmen have needed advice, and writing on fiscal affairs dates back to antiquity. It concerned the scholastics of the 16th century and occupied the mercantilists of the 17th. Systematic study of the public household by the Cameralists followed, and the ‘L’impôt unique’ was a central part of Physiocrat doctrine. In England, the writings of Petty, Locke, and Hume preceded Book V of Adam Smith's *Wealth of Nations*, the first ‘modern’ statement of the field. Thereafter, fiscal analysis followed (and in some cases led) the advances of economic science. Ricardo, Mill, the marginalists, Marshall, Pareto, and Pigou all left their stamp on the economics of public finance, not to mention the impact of Keynes and the emergence of stabilization as a goal of budget policy. But fiscal economics also added to the general body of economic analysis. Its concern is not limited to the effects of fiscal operations on the market, and market responses thereto. There remains the more basic question of why a public sector is needed and what rules should be applied to its conduct. Principles of public sector economics are required to provide the answer. These principles, to be sure, are coordinated with those of the market by the broader frame of economic welfare, but they address a distinct set of issues and, by their very nature, are linked more closely to the perspectives of political and social science.

Public expenditures

We begin with the expenditure side of the picture, and public goods in particular. Thereafter, transfers are examined.

Public goods

Here the basic issue is why certain goods and services have to be provided for through the budget, i.e., paid for by taxes and made available free of direct charge. Such goods and services may be produced under public or private management, e.g., government may install traffic lights itself or engage a private firm. Which is done does not matter here. The crucial point is that traffic signals are provided free of direct charge to the individual consumer when passing the intersection. Given a general presumption that consumer preferences should be met by the *quid pro quo* of the market, why should budgetary provision be chosen in the case of public goods?

A modern answer was anticipated by Hume's early insight. Two neighbours, so he argued, might agree to drain a meadow, but no such agreement can be reached by a thousand persons, as each will try to place the burden on others (Hume, 1739, p. 539). Adam Smith also examined why certain services must be provided by the Prince (Smith, 1776, Book V). These included the upkeep of the court, defence, police, and minimal education for the poor. More generally, the Prince was to provide for ‘those

institutions and public works which, though they may be highly advantageous to a great society, are, however, of such a nature that the profit could never repay the expense to any individual' (Smith, 1776, vol. 2, p. 539). The issue was joined but Hume's earlier insight had been lost. The reason why private provision will not work remained unanswered. John Stuart Mill subsequently came a bit closer. He noted that individual preferences, under certain conditions, cannot be met without common concert and legal sanction. The lighthouse illustration appeared (Mill, 1848, p. 976), and the difficulty of collecting tolls for the use of its services was stressed. Market failure due to inapplicability of exclusion was thus noted, but not as yet the more basic proposition that exclusion would be inefficient (absent crowding) even if it could be applied.

The discussion took a new turn in the 1880s, when utility analysis grounded value theory on the demand side. Public no less than private services are provided to meet the preferences of individual consumers, so it was argued, and this provision through the budget may be viewed in analogy to the exchange mechanism of the market. Taxes may be viewed as price payments, offered by the consumer. Thus a voluntary exchange model of the tax-expenditure process emerged (Sax, 1887; Mazzola, 1890). This vision of the fiscal process was rejected at once by Wicksell (1896). While public provision should be in line with individual preferences, it could not be implemented by voluntary exchange. As Hume had recognized a century and a half before, exchange will not work in the large number case. The level of public services available to A will not be affected significantly by his own contribution. Hence A will not reveal his preference for public services. Unlike the case of private goods, where the individual must bid to obtain his share in the auctioning process of the market, the consumer will now act as a free rider. A political process of budget determination by voting, combined with a legal sanction of its enforcement, is needed so that preferences are revealed. Thus the basis was laid for the modern discussion of public choice and the voting rules by which an efficient solution may be approximated. Though the simplistic hypothesis of voluntary exchange was rejected, the exchange formulation as developed by Lindahl (1919) nevertheless had an important role to play. With any given supply of the public service available to both A and B, A's demand curve may be viewed by B as a supply curve. Equilibrium is then reached where the vertically aggregated demand curves of A and B intersect the supply schedule for the product, i.e., where the tax prices paid by the two consumers add up to the social cost of the service. Lindahl's formulation, with its vertical addition of demands, thus anticipated a significant feature of Samuelson's later formulation.

The next phase of the development of public goods theory emerged with Pigou's analysis of externalities (1920). External costs and benefits remain unaccounted for by the market and thus call for correction by public policy. Pigou did not develop this theme in his treatise on public finance (1928), where the proper sphere of public services is defined only in general terms, calling for extension to the point where marginal social costs and benefits will be equal. However, the concept of external benefits might be extended readily to that of public goods, where externalities appear not as a by-product of internal gains but all benefits are external. Introduction of the Scandinavian model into the English-language literature followed (Musgrave, 1939), but the crucial implication of public goods for efficient resource use was not drawn clearly until Samuelson's statement (1954). Whereas efficient use of resources in the provision of private goods calls for an equating of the marginal rates of substitution in production and consumption (with the latter equal for all consumers), that of public goods calls for equality of the marginal rate of substitution in production with the *sum* of the marginal rates of substitution (differing among

consumers) in consumption. Lindahl's earlier exchange solution, arrived at by the vertical addition of demand curves, was compatible with this outcome; but, as shown in Samuelson's model, an efficient solution did not require the Lindahl-type determination of tax shares.

The basic difference in the two formulations should be noted. Lindahl, following Wicksell, began with a distribution of money income and then proceeded to assign efficient tax prices. This is accomplished by charging each consumer in line with his marginal evaluation and setting total supply so as to equate the sum of these charges with cost. Moreover, by postulating a just state of distribution to prevail to begin with, a requirement also advanced by Wicksell, the resulting burden distribution of tax shares would also be just. Samuelson, like Wicksell, rejected Lindahl's ‘pseudo demand curves’ as unrealistic. But where Wicksell proceeded to examine the process of preference revelation, Samuelson provided a more general definition of the efficient solution. Preference revelation is disregarded as the model visualizes an omniscient referee to whom preferences are known. He then establishes a utility frontier, showing various mixes of public and private goods, as well as private good distributions among individual consumers. The optimum optimorum (bliss point) on the utility frontier is then chosen on the basis of a social welfare function. The Lindahl solution becomes a special case only; but given the assumption of known preferences, there is no particular reason why it should be selected.

This more general formulation resolves the allocation and distribution aspects of the problem simultaneously, and deals with distribution in its basic welfare rather than income terms. The Lindahl model by comparison separates allocation and distribution issues. Since welfare is a function of real, not money income, it is open to the objection that the just distribution of money income cannot be determined without also setting tax prices, thus suggesting circular reasoning. However, this critique may be met by adding the determination of the voting rule (designed so as to best secure preference revelation) as a further equation to the model. Moreover, the Lindahl formulation provides a closer linkage to the real world. While tax prices cannot be seen as voluntary offers, there exists no omniscient referee to whom preferences are known. Preferences, as Wicksell noted, must be revealed through a voting process and this presumes a distribution of money income to begin with. The Lindahl price thus remains as a benchmark against which the quality of the voting process can be measured. Separation of the allocation and distribution phase of budget policy was extended subsequently to include the stabilization function as a third branch (Musgrave, 1959).

Subsequent developments in the theory of public goods recognized the fact that particular goods and services may not meet the polar conditions of purely private or public goods, but fall in between. A's consumption of X may provide benefits internal to A and hence be undertaken by him. But it may also generate external benefits or costs for B, C, and D. Thus, partial provision through the budget, that is, a subsidy solution, may be called for. Or, the number of consumers may be sufficiently small to permit a bargaining solution without voting; yet there is no assurance that an efficient outcome will emerge. Moreover, it may be possible to satisfy certain needs via the provision of public or of private goods, thus permitting a choice between the two modes.

A special problem arises also from the fact that the benefits of public services may be subject to spatial limitation. The resulting distinction between ‘local’ and ‘national’ public goods provided the basis for a theory of fiscal location (*see* local public finance) according to which the provision of public services should be arranged so that each jurisdiction will provide and pay for the public services the benefits of which accrue within its borders. Moreover, the spatial limitation of benefits led to the proposal (Tiebout, 1956) that preference revelation may occur through ‘voting by feet’. Individuals with similar public

goods preferences would find it advantageous to congregate. At the same time, individuals with lower incomes will find it advantageous to congregate with higher income individuals, so as to generate an unstable distribution across jurisdictions.

The general theory of public goods has thus been extended and qualified to deal with particular situations. It should be noted, however, that all these variants assume public goods to be provided in line with the preferences of individual consumers. The theory of public goods is thus similar in its psychological underpinnings to that of private goods. The concept of communal needs or goods which the community considers meritorious offers an alternative perspective not included in the mainstream view of expenditure theory (*see* merit goods).

Cost – benefit analysis

Moving from general theory to practical application, the development of cost–benefit analysis has attempted to design an operational framework by which the appropriateness of particular expenditure projects may be evaluated and ranked. In the process, the present value of the expected benefit stream is balanced against that of its opportunity cost. For this purpose, a social discount rate has to be determined, a rate which may or may not be taken to equal that of the market. Moreover, opportunity cost is found to depend upon whether the resource withdrawal is from consumption or from capital formation. Since the income tax enters as a wedge between gross and net return, capital formation in the private sector falls short of the optimal level so that resource withdrawal from private investment carries a higher social cost than does resource withdrawal from consumption. Shadow prices are applied to measure the social cost of labour and other inputs, thus correcting for further distortions in market prices. Finally, distributional weights, based on a social welfare function, may be applied to the resulting costs and benefits. Thus, a framework is provided by which the value of alternative expenditure projects may be assessed and ranked. However, the analysis assumes that the value of the benefit stream can be determined. This involves no difficulty where the output of the public project is sold at the market, but approximations have to be used where the services are in the nature of public goods. Thus, the value of a park may be measured by the opportunity cost reflected in the visitor's travel time or by similar proxies.

Transfers

While economic analysis has focused on the provision of public goods and services, transfers have come to claim an increasing share of total spending. Aimed at correcting the distribution of income, they may be viewed as negative taxes. While resource use for the provision of public goods may be fitted into the Paretian mould of allocation efficiency, transfers pose a more difficult problem. To be sure, a theory of giving, or Pareto optimal redistribution, may be developed in the context of interpersonal utility interdependence. If the donor's satisfaction is derived from the pleasure of individual giving, giving remains a private good. But if it is derived from the welfare of others, giving assumes a social-good quality and calls for budgetary provision. Yet the outcome reflects the initial distribution of income and thus does not resolve the more basic problem of primary distribution. This transcends considerations of Pareto efficiency, and broader grounding in a theory of justice is required, be it a Lockean rule of entitlement, a utilitarian concept of maximum welfare, or a Rawlsian sense of justice as fairness. But notwithstanding this inherent link to a theory of justice, economic analysis retains a decisive role. The

size of the pie is linked to its distribution, and redistribution involves an efficiency cost. The one, therefore, cannot be determined without the other.

Turning to the form in which transfers should be given, economists have traditionally argued in favour of a general income subsidy, rather than selective subsidies designed to support particular uses of income. The general subsidy will be more valuable to the recipient as it does not interfere with his choice among income uses. But various exceptions may be noted. Transferors, consenting to a transfer, may do so on condition that the income is put to specified uses. Giving may take a paternalistic form. Moreover, distributive justice may be viewed in categorical terms, applying different standards of equality to basic items than to other income uses (*see* merit goods). Beyond this, the logic of optimal commodity taxation, calling for differential rates of tax on various goods and services, may also be shown to call for differential subsidy rates to be applied to various commodities.

Macroeconomic aspects

The preceding discussion has dealt with the role of public expenditures in providing for public goods and for adjustments in distribution. In the process, the fiscal system may add to capital formation via public investment and detract therefrom via reduced capital formation in the private sector. Public finances thus have an important bearing on the rate of economic growth, a fact which has been dealt with throughout the literature, and which was central to Ricardo's critique of the public sector. Moreover, the choice between tax and loan finance becomes an instrument of inter-temporal burden distribution. Since loan finance falls more heavily on capital formation, it leaves future generations with a smaller capital stock. Considerations of intergenerational equity thus permit public investment, the benefits from which accrue to future generations, to be loan financed, while calling for current services to be tax financed. This establishes the rationale for a dual budget system, with balance in the current and loan finance in the capital budget. This reasoning, in turn, calls for the inclusion of depreciation in the current budget, with corresponding debt retirement over the useful life of the asset. It should be added that inclusion of outlays in the capital budget does not require public acquisition of real assets (with its fictitious analogy to a balance sheet) but simply the creation of future benefits, including these of investment in human capital through outlays on health and education.

A quite different macro-perspective on public expenditures emerged with the Keynesian model. While the classical framework had left the budget aggregate-demand neutral, deficit and surplus finance now became a source of demand expansion and restriction. With initial emphasis on fiscal expansion (restriction) directed at increase (decrease) in public spending, tax reduction (increase) subsequently entered as an alternative way of achieving similar results. The early Keynesian model, which left money impotent and viewed fiscal policy as all-powerful, was modified in the neoclassical synthesis of the 1950s and 1960s, and attention moved to the correct mix of fiscal and monetary constraint. Moreover, the supremacy of aggregate demand controls become questionable as attention shifted from full employment to inflation. Nevertheless, aggregate demand effects of fiscal operations have remained of major concern, joining the earlier issues of allocation and distribution as a third dimension of fiscal economics.

Fiscal politics

The preceding discussion, in line with the tradition of fiscal economics, has dealt with normative issues of expenditure policy, that is, why such expenditures are needed and how they should be designed to obtain efficient results. More recently, a new perspective has been added. Not resting on the assumption that prescription of correct policy will be followed once laid down by economic analysis, attention has turned to how public policy does in fact behave and how its behaviour is determined. In the process, emphasis has shifted from concern with market failure to focus on failure in public policy (Buchanan and Tullock, 1962). Early efforts to develop a theory of public-sector behaviour had been made in a Marxist framework, viewing the state as an instrument of exploitation by the ruling class. Recent analysis proceeds in analogy to microeconomics, involving the self-interested behaviour of voters, bureaucrats and politicians. An important focus in this analysis has been the growth of the public sector, the extent to which it reflects the inherent needs of modern society as expounded in Wagner's Law (1883), or a malfunctioning of the fiscal system based on an inherent bias towards over-expansion (*see* public choice).

Taxation

We now turn to the tax side of the fiscal picture, beginning with the normative requirements for a good tax system.

Criteria for equity

From Adam Smith on, students of taxation have been concerned with the qualities of a good tax system. One such requirement, traditionally ranked first, is that the tax burden should be distributed in an equitable fashion. This requirement has taken two forms, one calling for taxation in line with benefits received, and the other for taxation in line with ability to pay. Both approaches were reflected in Smith's maxim that ‘the subject of every state ought to contribute towards the support of the government, as nearly as possible in proportion to their respective abilities, that is, in proportion to the revenue which they respectively enjoy under the protection of the state’ (Smith, 1776, vol. II, p. 310). The benefit principle has the advantage that it links the tax and expenditure side of the budget and thus relates to the theory of public goods. The Lindahl price, after all, was the benefit tax par excellence. But benefits are not readily assigned, thus leaving the benefit rule inoperative in most cases. Moreover, as noted before, fee finance, related to the level of individual consumption of public goods, interferes with their efficient provision; nor does the benefit principle admit redistributional uses of the fiscal process.
The ability to pay approach in turn has the disadvantage that it views the distribution of the tax burden (or the resulting change in the distribution of the tax income) as independent of the expenditure side of the budget. Nevertheless, this approach has received primary attention. Beginning with J.S. Mill (1848, p. 804), taxation was viewed as imposing a sacrifice and the problem was how to distribute this sacrifice in an equitable fashion. Justice requires that people in equal positions be taxed equally so as to undergo an equal sacrifice, people in unequal positions, however, are to pay unequal amounts of tax, differentiated so as to involve an equal sacrifice. Underlying this formulation was the assumption of declining, similar, and comparable marginal utility of income schedules. Subsequent refinement by

Edgeworth (1897) and Pigou (1928) differentiated between equal absolute, equal proportional, and equal marginal (least total) sacrifice. Equal marginal sacrifice calls for maximum progression provided only that the utility schedule is declining. Equal absolute sacrifice calls for progressive proportional or regressive taxation, depending upon whether the elasticity of the marginal utility of income schedule falls short of, equals, or exceeds unity. No simple rule, finally, can be given for the case of proportional sacrifice. Authors such as Edgeworth and Pigou, as had Bentham (1802), opted for the equal marginal sacrifice rule, given the utilitarian premise that least total sacrifice (or a maximum level of remaining welfare) should be the goal of rational conduct. Given the further assumption of equal utility schedules, the formulation calls for a move towards equal distribution. But having drawn this basic conclusion, it is then qualified to allow for incentive effects and the resulting shrinkage in the overall level of income which is available for distribution.

The sacrifice theory of tax equity nicely fitted the framework of the ‘old welfare economics’, which was willing to assume inter-personal utility comparison. As this assumption was discarded in the 1930s, equal sacrifice rules became inoperative but subsequently were replaced by the hypothesis of a social welfare function, assigning marginal social utilities to various levels of income. People with equal income should still pay the same tax (the principle of horizontal equity), while differential taxation at different levels of income would be determined in line with society's view of declining social utility as income rises. Notwithstanding Arrow's demonstration that a social welfare function cannot be derived in an unambiguous fashion, such a concept is now widely used in policy evaluation, including cost–benefit analysis and the setting of optimal tax rates.

Definition of income

Dating back to Adam Smith, the analysis of tax equity has focused on income as the index of ability to pay. While expenditures or consumption have entered as alternatives, income has received the major attention. But the definition of income for purposes of taxation is not obvious, especially not in the context of a highly complex financial and industrial economy where income may be received and used in a variety of forms. A large part of tax analysis has thus been concerned with the definition of income and its application in this complex setting.

The analysis has proceeded from the basic concept of net income (Schanz, 1896; Simons, 1938) as accretion to wealth or, which is the same, increase in net worth plus consumption. On this basis, a host of specific issues are dealt with, including the treatment of unrealized gains, depreciation, interest paid, income in kind, imputed income, as well as many other items. While economists have argued for a broad and comprehensive income base which would permit the needed revenue to be obtained at lower rates, they have had only limited success. The tax base has been diluted by an expanding net of tax preferences and it remains to be seen whether a political consensus for base-broadening can be reached. The problem is complicated by the fact that not all omissions from the tax base reflect gross efforts at tax avoidance. Others may be viewed as providing incentives to secure policy objectives which may be valid on their own terms. Economists have opposed such use of ‘tax expenditures’, noting that the underlying policies, if valid, may be pursued more efficiently through the expenditure side of the budget, including subsidies to solicit private sector responses. More recently, the problem of tax base definition has been complicated further by the impact of inflation. With ability to pay relating to real rather than nominal income, inflation adjustments are appropriate, both with regard to the indexing of rate brackets and the

measurement of capital gains and interest income.

A further central issue relates to the tax treatment of the corporation. The purist position has been that equity in taxation refers to the tax treatment of individuals, and that all income ultimately belongs to them, be it directly or via their ownership of legal persons such as corporations. From this it is concluded that corporate source income should be taxed to its owners, and not be subjected to an additional or absolute tax at the corporate level. Dividend and interest disbursement by the corporation should be taxed to the shareholder, as should undistributed earnings. For purposes of administration, shareholders' taxes on corporate source income may be withheld at the corporate level, but they would then be credited under the individual income tax. This approach, however, is rarely followed. Instead, a separate corporate tax is imposed and corporate source income if retained is allowed for but imperfectly at the individual shareholder level.

Finally, and of increasing importance, the structural problems of the income tax are complicated by international capital flows and multiple jurisdictions. Techniques have been devised to protect capital income against multiple taxation, and the question of which jurisdiction (e.g. origin or destination) is entitled to a particular tax base has been debated.

Consumption base

While primary focus has been on the nature of income as the tax base, consumption has been considered as an alternative thereto. Hobbes (1651) argued at an early point that a person should be taxed on what he takes out of the pot (i.e. consumes) and not on what he adds (i.e. saves). Also, economists from Mill to Marshall, Pigou and Fisher, have held that the income tax involves a 'double taxation' of saving. By taxing income when saved and then taxing interest thereon, the tax differentiates unfairly against savers and in favour of consumers, and an excess burden or efficiency cost results from such tax discrimination against saving. The case for a consumption base was impeded, however, by the assumption that it would have to be applied in the form of 'in rem' taxation, i.e., through excises or general growth income taxes such as the retail sales or value added tax (*see* Kay, 1987). Because of their impersonal nature, such taxes would not be acceptable on equity grounds. This objection no longer applies, as the case for the consumption base has been reformulated in the context of a personal expenditure tax (Kaldor, 1955), with personal exemptions and progressive rates applicable as under the income tax.

Much recent attention has been given to the way in which a comprehensive expenditure base would be computed. Determined as cash income plus net borrowing minus net investment, consumption would be arrived at as a residual, rather than by attempting the aggregation of outlays. Determination of the expenditure base would bypass certain central difficulties of income determination (especially the treatment of postponed income, unrealized gains, and depreciation), but new problems would arise as well, and pressures for base preferences would not disappear. Difficulties would have to be met, especially in relation to the transition from income to expenditure taxation.

There is also the question whether consumption is indeed the correct base. The expenditure tax avoids disincentive to saving and gives equal treatment to individuals who consume early and late in their lives. But preference is given to those who make gifts and leave bequests. Thus critics hold that the index of equality should be defined as equal present value of potential (not only actual) consumption. It then follows that bequests and gifts should be included in the testator's expenditure base. Also, it may be

argued that the gain from saving not only consists in increased future consumption but that the holding of wealth itself carries utility, allowance for which may call for a supplementary wealth tax. Viewed as the equivalent of a wage income tax, the expenditure tax stands in uneasy contrast to the traditional view that if anything income from capital should be taxed more heavily than income from wages.

The preceding discussion of the tax base has focused on income and expenditures as the primary options. In a fuller treatment, other forms of taxation, in particular the property tax and payroll tax, would have to be considered as well. Indeed, the growth of tax structures has reflected the changing patterns of economic institutions and availability of ‘tax handles’. What are good taxes for a highly developed financial economy such as the US of today were not feasible when selective property taxes provided the main revenue source under colonial conditions. Nor can the same tax rules be applied to developing countries of today. Moreover, the choice of appropriate taxes differs at the central and local levels of government, all of which renders the problem of tax structure design richer and more complex than can be accounted for here.

Efficiency rules

An equitable distribution of the tax burden is one important attribute of a good tax structure. But it is not the only one. We now turn to the further and related issue of efficiency in taxation. As Adam Smith noted (1776, p. 310), taxes ought to be designed so ‘as to take out of the pockets of the people as little as possible over and above what it brings into the public treasury of the state’. Compliance and collection costs should be minimized, but this is not all. At a more subtle level, as later discussion has shown, a given revenue should be drawn from any one taxpayer so as to impose the least welfare loss. Taxes other than lump sum taxes impose an efficiency cost, i.e., leave the taxpayer with a loss which exceeds the value of revenue which government obtains. In the extreme case, a taxpayer may be burdened while there is no revenue gain: for example, a person may cease to consume a taxed product, leaving the treasury without gain and forcing the taxpayer into a less satisfactory consumption mix. Or, imposition of an income tax may induce the taxpayer to substitute leisure for income, thereby reducing the tax base while burdening him with a less satisfactory work–leisure choice.

The measurement of deadweight loss or loss of consumer surplus as a triangle under the demand curve was anticipated by Dupuit (1844) and Jenkin (1871), and was subsequently developed by Marshall (1890, Book III, Chapter 6). Modern discussion of deadweight loss begins with Pigou's treatment of announcement effects (1928). Assuming leisure to be fixed, the optimal solution (which minimizes deadweight loss) calls for all products to be taxed at a uniform ad valorem rate, but the problem is more difficult if leisure is allowed to vary. Since leisure as such cannot be taxed, the taxation of products complementary to leisure must take its place. As first shown by Ramsey (1927), deadweight loss is minimized by imposing a set of differential ad valorem rates, such as to reduce the production of all commodities in equal proportion. After an interval of nearly fifty years, this rule then laid the basis for the theory of optimal taxation (Diamond and Mirrlees, 1971). Discussed elsewhere (*see* optimal taxation), it will not be expanded on here.

Further problems arise in moving from the optimal treatment of a particular taxpayer to that of the group. If the utility function of all taxpayers is assumed to be the same, no difficulty arises. But if it is allowed to differ, the ideal pattern of optimal taxation would call for the tailoring of differential rates of

tax to the particular preferences of each taxpayer. Since this is impossible, a general tax formula has to be used, based on representative behaviour. This bypasses issues of horizontal equity, issues which arise precisely because behaviour patterns differ.

Shifting and incidence

Economists have for long been aware that there exists a difference between the point at which a tax is imposed (its statutory incidence) and that at which its final incidence comes to rest. The in-between process or shifting has filled the largest chapter in the history of public finance, and by its nature as market economics has developed in close linkage to the general body of economic theory.

In the context of the Physiocratic model, only a tax on land could be productive as only land was a true source of income. The classics continued to focus on the division of output among factor shares, but the addition of capital to land and labour provided a three-factor model. This expanded model not only fitted the analytical scheme but also reflected the social structure of the times. Ricardo in particular devoted a large part of his treatise to this aspect of taxation. He agreed with the Physiocrats that a tax on rent cannot be shifted, but replaced the view of land as the basic source of income by recognition of rent as an intra-marginal return which does not affect price. A tax on wages must in the short run be borne by profits, so he held, since wages are at subsistence and cannot be reduced. But accumulation declines in the longer run, forcing a reduction in population. The same holds for a tax on necessities. Taxes on luxury products are absorbed by the payee as are taxes on profits. But the latter once more reduce accumulation, and hence the demand for labour. In the end only luxury consumption remains as a solid tax base.

This simple solution crumbled with the subsistence hypothesis. Replaced by a generalized theory of factor pricing, based on marginal products, no factor share remained immune to taxation, and tax incidence had to be viewed in the context of a general equilibrium system of competitive factor and product pricing (Walras, 1874). With factor and product pricing interacting in a general equilibrium setting, a product tax might come to affect the position of households from the sources as well as from the uses side of their account, just as a tax on factor income might come to be felt from the uses as well as from the source side. Advancing in many directions, the theory of incidence came to distinguish between partial and general taxes, short and long run results, as well as outcomes in competitive and imperfect markets. Moreover, while the classics had been concerned primarily with the distribution of the burden among factor shares, subsequent concern turned to the more complex issue of burden allocation among income groups.

Analysis of partial product taxes in terms of supply and demand curves and their elasticities were first undertaken by Jenkin (1871), developed by Marshall (1890), and extended in detail by Edgeworth (1897). As concluded later (U. Hicks, 1947), a tax would be divided between buyers and sellers in inverse relation to the elasticity of substitution in supply and demand. As first suggested by Barone (1899) and later developed by J.R. Hicks (1939), a distinction was drawn between the income and substitution effects of a tax. As the two effects contradict each other, the net effect of an income tax on factor supply was no longer evident. Moreover, it no longer followed that a progressive tax schedule must depress work effort more than would a proportional one. The former, to be sure, involves higher marginal rates at high levels of income, and hence imposes a more severe substitution effect on such

taxpayers; but the latter requires higher rates further down, so that the net effect is not evident.

While most incidence analysis has been conducted in competitive markets, attention has also been given to non-competitive conditions. Cournot (1838) early on showed that a tax on monopoly profits cannot be shifted, but later analysis dealing with other forms of market imperfection showed that profits taxes may indeed be passed on. Returning to the competitive case, analysis has focused on the incidence of profits taxes imposed on selected industries only. As Marshall (1890) had pointed out, such profits in the short run are quasi-rents so that the tax stays put; but given sufficient factor mobility, such is not the case in the longer run. A reduction in the return to capital in any particular industry eventually comes to be shared by capital at large. As capital moves from the taxed to tax-free industries, net returns are equalized. In the process, consumers and other factors may come to share part of the burden (Harberger, 1974).

The emergence of neoclassical growth models (Solow, 1956) soon invited a reformulation of long run incidence in the classical tradition (Krzyzaniak, 1967; Feldstein, 1974). Incidence under steady growth is shown to depend on savings propensities as well as the elasticities of factor supplies. Thus substitution of a tax on capital income for an equal tax on labour income will leave part of the burden on labour, even if factor supplies are inelastic, provided that the propensity to save out of capital income is higher; and capital income will bear the entire burden, even if labour supply is elastic, provided that the propensities to save are the same.

As is not infrequently the case, theory advanced more rapidly than its empirical verification. Econometric testing of incidence outcomes has been undertaken but rarely (Musgrave and Krzyzaniak, 1963) and has led to controversial results. Instead, two more hypothetical approaches have been undertaken towards quantitative estimation of tax-burden distribution. One approach has relied on what seem reasonable assumptions regarding the shifting of particular taxes, which assumptions are then implemented on the basis of available income and expenditure data (Colm and Tarasov, 1940; Pechman and Okner, 1974). The other involves simulation of a general equilibrium model, reflecting the observed structure of the economy. This model is then made to respond to the introduction of particular taxes (Shoven and Whalley, 1984), and the resulting changes in household positions are observed. The former approach has the advantage that the implications of various shifting hypotheses can be tested, but it fails to allow for second round effects. The latter has the advantage of accounting for a full sequence of adjustments and includes deadweight losses in the burden estimation, but it has the disadvantage that the result are drawn from the premise of perfectly competitive markets. In comparing the two, much depends on the weight of first round effects.

See Also

- neutral taxation
- optimal taxation
- progressive and regressive taxation
- Ricardian equivalence theorem
- welfare economics

Bibliography

Barone, E. 1899. Di alcuni teoremi fondamentali per la teoria matematica dall' imposta. *Giornale degli Economisti*, Second Series 60. Trans. as 'About fundamental theorems on the mathematical theory of taxation' in Musgrave and Shoup (1959).

Bentham, J. 1802. *Principles of the Civil Code*. In J. Bentham, *The Theory of Legislation*, ed. C. Ogden. London: Kegan, 1931.

Buchanan, J. and Tullock, G. 1962. *The Calculus of Consent*. Ann Arbor: University of Michigan Press.

Colm, C. and Tarasov, H. 1940. *Who Pays the Taxes?* Washington, DC: Temporary National Economic Committee.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Diamond, P. and Mirrlees, J. 1971. Taxation and public production I: Production and efficiency; and II: Tax rules. *American Economic Review* 61, Pt I, March, 8–27; Pt II, June, 261–78.

Dupuit, J. 1844. De la mésure de l'utilité des travaux publics. *Annales des Ponts et chausées*, Second Series, Vol. 8. Reprinted in *International Economic Papers No. 2* (1952), London: Macmillan, 83–110.

Edgeworth, F.Y. 1897. The pure theory of taxation. *Economic Journal* 7, Pt I, March, 46–70; Pt II, June, 226–38; Pt III, December, 550–71.

Feldstein, M. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82, 905–26.

Harberger, A.C. 1974. *Taxation and Welfare*. Boston: Little, Brown.

Hicks, J. 1939. *Value and Capital*. Oxford: Oxford University Press.

Hicks, U. 1947. *Public Finance*. New York: Pitman.

Hobbes, T. 1651. *Leviathan*. Harmondsworth: Penguin Books, 1968.

Hume, D. 1739. *A Treatise of Human Nature*, ed. L. Selby-Bigge. Oxford: Oxford University Press, 1975.

Jenkin, F. 1871. On the principles which regulate the incidence of taxes. *Proceedings of the Royal*

Society of Edinburgh. Reprinted in Musgrave and Shoup (1959).

Kaldor, N. 1955. *An Expenditure Tax*. London: Allen & Unwin.

Kay, J. 1987. Indirect taxes. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Krzyzaniak, M. 1967. The long-run burden of a general tax on profits in a neoclassical world. *Public Finance* 22, 472–91.

Lindahl, E. 1919. *Die Gerechtigkeit der Besteuerung*. Lund: Gleerup.

Marshall, A. 1890. *Principles of Economics*. 8th edn. London: Macmillan, 1920.

Mazzola, U. 1890. *I dati scientifici della finanza pubblica*. Rome: E. Loescher & Co.

Mill, J.S. 1848. *Principles of Political Economy*, ed. W. Ashley, London: Longmans, 1921.

Musgrave, R.A. 1939. The voluntary exchange theory of taxation. *Quarterly Journal of Economics* 53, 213–37.

Musgrave, R.A. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

Musgrave, R.A. and Krzyzaniak, M. 1963. *The Shifting of the Corporation Tax*. Baltimore: Johns Hopkins Press.

Musgrave, R.A. and Shoup, C.S., eds. 1959. *Readings in the Economics of Taxation*. Homewood, Ill.: Irwin.

Pechman, J. and Okner, B. 1974. *Who Bears the Tax Burden?* Washington, DC: Brookings Institution.

Pigou, A. 1920. *The Economics of Welfare*. London: Macmillan.

Pigou, A. 1928. *A Study in Public Finance*. London: Macmillan.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Ricardo, D. 1817. *The Principles of Political Economy and Taxation*, ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36,

387–89.

Sax, E. 1887. *Grundlegung der theoretischen Staatswirtschaft*. Vienna: Holder.

Schanz, G. 1896. Der Einkommensbegriff und die Einkommenssteuergesetze. *Finanzarchiv* 13, 1–87.

Shoven, J. and Whalley, J. 1984. Applied general-equilibrium models of taxation and international trade. *Journal of Economic Literature* 22, 1007–51.

Simons, H. 1938. *Personal Income Taxation*. Chicago: University of Chicago Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. E. Cannan. New York: Putnam, 1904.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Wagner, A. 1883. *Lehr und Handbuch der Politischen Oekonomie, Vierte Hauptabeilung: Finanzwissenschaft*. 3rd edn. Leipzig: Winter.

Walras, L. 1874–7. *Elements of Pure Economics*. Trans. W. Jaffé, Homewood, Ill.: Irwin, 1954.

Wicksell, K. 1896. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens*. Jena: Fischer.

How to cite this article

Musgrave, R. A. "public finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000244> doi:10.1057/9780230226203.1363

The New Palgrave Dictionary of Economics Online

public goods experiments

Rachel T.A. Croson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic theory often cites the existence of goods with externalities as justification for government intervention, either as taxation to fund goods with positive externalities which would otherwise be underprovided, or as regulation on goods with negative externalities which would otherwise be overprovided. A series of experiments tests these predictions of under- or over-provision. This article describes the landscape of public goods experiments, identifying similarities and differences between them and summarizing the broad findings.

Keywords

altruism; common pool resource game; cooperation; coordination; externalities; learning; linear public goods game; other-regarding preferences; Prisoner's Dilemma; production functions; provision point mechanism; public good experiments; public goods; reciprocity; rent seeking; repeated games; self-interest; social cost; tragedy of the commons; voluntary contribution mechanism

Article

Since the vivid description of the Prisoner's Dilemma game and the accompanying tension between self-interest and efficiency, economists, political scientists, psychologists, sociologists and others have wondered about how individuals resolve these conflicting motivations. Many have investigated this question by experimentally examining behaviour when individual interest and group interest conflict. This research goes under different names – for example, social dilemmas in psychology, commons dilemmas in political science and public goods problems in economics.

This article does not provide a comprehensive review of this research; other excellent reviews aimed at economists (Ledyard, 1995), psychologists (Dawes, 1980) or sociologists (Kollock, 1998) exist. Instead, I highlight the different categories of public goods problems and introduce their commonalities.

The rest of this article is organized as follows. First, I offer some definitions of characteristics that public goods problems share. Next I discuss some dimensions on which they differ, and how these dimensions translate into different equilibrium and efficient outcomes. Then I describe three specific public goods problem types that have been extensively studied in the economics literature: the voluntary contribution mechanism, the provision point mechanism and the common pool resource. I conclude with a description of other games that have been studied, but where more work could be done.

Similarities

The common feature in public goods settings is the existence of externalities. In public *goods* problems, individuals can use private resources to provide goods that have positive externalities for others. Since some social benefits are not captured by the individual making the decision, this results in under-provision relative to the socially optimum level. Self-interested economic theory, then, argues that these goods will be under-provided, justifying taxation as a role for government.

Parallel to the public goods situation is the case of public *bads*. Here, individuals receive private resources by producing goods that have a negative externality for others. Since some social costs are not captured by the individual making the decision, this results in overprovision of public bads relative to the socially optimum level. Self-interested economic theory,

then, argues that these goods will be over-provided, again justifying a role for government, here in regulation. (A number of models have extended the existing theory to internalize the externalities. In models of altruism – for example, Becker, 1974; Andreoni, 1989 – the utility function of one party includes the consumption of others. Thus, when an action creates positive externalities, the value from it is increased over the self-interested model. Charness and Rabin, 2002, posit that individuals care not only about their own consumption but also about social welfare directly. These other-regarding preferences can internalize some of the externalities in public goods problems, but typically the over- or under-provision problems are not eliminated.)

The defining characteristic of the public goods problems described here, however, is the existence of positive externalities. My actions affect others, and I do not take this effect (sufficiently) into account in my own maximization problem. Often these problems are symmetric; each individual faces the identical conflict, but this is not necessary for there to be a public goods problem. For a problem to exist, it must only be that the individual's welfare and the group's (social) welfare conflict.

Differences

The largest, most important and least-recognized difference in public goods situations is the production function. How does an individual's action create positive or negative externalities for others? Different production functions have different implications for equilibrium predictions as well as socially optimal outcomes.

A second important dimension on which these situations differ is the decision space. Public goods problems can involve acting to provide goods with positive externalities at a private cost, refraining from acting so as to avoid imposing negative externalities on others at a personal cost, or acting to capture what would otherwise be public benefits for private consumption. Unlike the first dimension, these differences in the decision space have only a superficial impact on the equilibria; that is, one can easily describe ‘not polluting’ as ‘producing a public good’. However, they may affect how individuals think about (and act in) these problems.

To clarify these differences, let's examine the three classic games discussed below in terms of the production function and decision space. In the voluntary contribution mechanism (VCM), the decision space involves *giving*. Individuals are given an endowment, which they can use for their private consumption or to produce the public good. Their allocations toward the public good provide value for others in the experiment (positive externalities). The production function of the VCM game most extensively studied is *linear* (thus it is sometimes also called a linear public goods game). The more that is allocated toward the public good, the greater are the social benefits in a linear fashion. This linearity means that (with appropriate parameters as discussed below) this game has a unique Nash equilibrium in which no participant allocates any resources towards producing the public good. Deviations from the equilibrium are both welfare-enhancing and represent deviations from pure self-interest maximization. They are thus referred to as ‘cooperation’, and concepts like altruism, warm glow and reciprocity are offered as their explanation.

In contrast, consider the provision point mechanism (PPM). Again, this game typically has a *giving* decision space. Individuals are given an endowment which can be allocated to private consumption or towards providing the public good. But, in contrast with the VCM, in the PPM the production function involves a *threshold*. If enough resources are collected, then the public good is provided and all receive its benefits. If too few resources are collected, then the public good is not provided and no positive externalities are enjoyed. The threshold nature of this production function has critical implications for the equilibria of this game. With appropriate parameters (discussed below), the full free-riding equilibrium still exists. But there also exist a set of efficient equilibria, in which the public good is exactly provided with each individual contributing a share of its cost. In each of these equilibria, the share contributed by each individual varies. (For example, there may be one equilibrium in which I contribute 80 per cent and you contribute 20 per cent of the cost of providing the public good, and another where I contribute 20 per cent and you contribute 80 per cent.) The problem then becomes one not of cooperation but primarily one of coordination; how do we select among these efficient equilibria? Formally, this game can be thought of as a large battle-of-the-sexes game (a game of impure coordination), with multiple equilibria each of which is somebody's favourite.

Finally, consider the common pool resource (CPR) game. Here the decision space involves *taking*; for example, individuals can harvest grass from the commons for personal gain. Surprisingly, these experiments are often described as *giving* games, with negative externalities rather than positive, as in the VCM or PPM. So the decision is made on ‘how many hours to spend grazing’ with the resulting negative externalities as cattle eat more grass. The production function used in CPR games is typically *nonlinear*. A small amount of harvesting creates more benefit for the individual than harm to the society (and is thus socially efficient). However, as the individual harvests more, the personal benefits decrease and the social costs increase until societal costs outweigh private benefits (the socially optimal point). With appropriate parameters (described below), however, private benefit is still above private costs, leading individuals to continue harvesting past the socially efficient point. Eventually, private benefits equal private costs, leading individuals to stop harvesting. These equilibria are thus internal (individuals typically harvest more than zero but less than the full amount) but still suboptimal (total harvesting is larger than the socially optimum level).

There are many additional dimensions on which these games can vary. For example, experimenters have varied the number of players from two (the classic Prisoner's Dilemma game) to as many as 100 (Isaac, Walker and Williams, 1994). The particular parameters can vary, subject to constraints that preserve the public goods nature of the problem. The institutional rules can vary, participants can decide simultaneously or sequentially, they can discuss the game in advance or not, and so on. The games can be (finitely) repeated or one-shot. I discuss some of these variations in the sections below, but their impacts on the equilibria of the games are straightforward.

In summary, the set of public goods games is broad. When one looks at a game, however, it is critical to understand the production function that is being used to translate decisions

into outcomes (positive/negative, linear/threshold/nonlinear), and the decision space that participants face (giving/taking/refraining from action). These dimensions have important impacts on the equilibrium predictions, the observed behaviour and the attributions that one can make about the causes of differences between the two.

The voluntary contribution mechanism (VCM)

The work of Marwell and Ames (1979; 1980; 1981) is often cited as the earliest VCM experiments. Unfortunately these early experiments did not involve a linear production function. Instead the return from the public account was discrete (chunky) although some of the experiments involved a linear approximation (for example, 1981, study I). Furthermore, the experiments were relatively uncontrolled; subjects had instructions mailed to them at home, were individually called and had the instructions explained to them, and then called back one week later and made their (one-shot) decision by phone.

The first paper using a linear VCM in a controlled lab setting was Isaac, Walker and Thomas (1984). This paper set a number of precedents for how such experiments are run. In this experiment, participants were brought into the lab and arranged into fixed groups of four. In each period, each group member was given tokens, which he could allocate between a private account and a group account. Tokens allocated to the private account earned 1¢ per token. Tokens allocated to the group account earned 0.3¢ per token for each member of the group, whether or not he had contributed to the group account. As the production function is linear, these parameters remain constant regardless of how much is contributed.

More generally, for there to be a public goods problem in these linear games, a few conditions must be satisfied. First, the return from the public good to the individual must be lower than the return from the private good (0.3<1). This ensures that individuals do not have an individual incentive to contribute, and that the dominant strategy equilibrium in the stage game is thus to contribute zero tokens. Furthermore, the social benefit from contributing toward the public good must be greater than the social cost (0.3*4=1.2>1). This ensures that contributing toward the public good is socially efficient.

The game is finitely repeated for ten periods, to allow for convergence to (and learning of) the equilibrium. In the finitely repeated game, backward induction results in the unique Nash equilibrium of zero contributions. Deviations from that equilibrium are attributed to cooperation, altruism, reciprocity or various other-regarding preferences.

A number of precedents set in this original article have been used in subsequent research. Many papers use a group size of four, although some have gone as low as two and others as high as 100. Most experiments have participants ‘allocate’ tokens between multiple funds rather than ‘contribute’ towards a public good, as this experiment did. Participants typically have multiple tokens to allocate rather than simply one. Most papers use repetition with fixed groups, and many choose ten periods.

The results from this wide variety of experiments are quite robust. First, on average, contributions to the public good begin at about half the endowment of tokens. Second, there is considerable variation in the decisions of individuals. Third, those contributions reduce over time until the contributions in the final round are 10–20 per cent of the endowment. An example of this pattern of contributions is depicted in Figure 1. A number of interesting papers have hypothesized and tested for the source of these regularities. Some explanations include errors (Palfrey and Prisbrey, 1997), confusion (Andreoni, 1995b), strategies and learning (Andreoni and Croson, 2008), and reciprocity or conditional cooperation (Croson, 2007), among others.

Figure 1

Average contributions to public account in VCM. *Source*: Croson (2007).

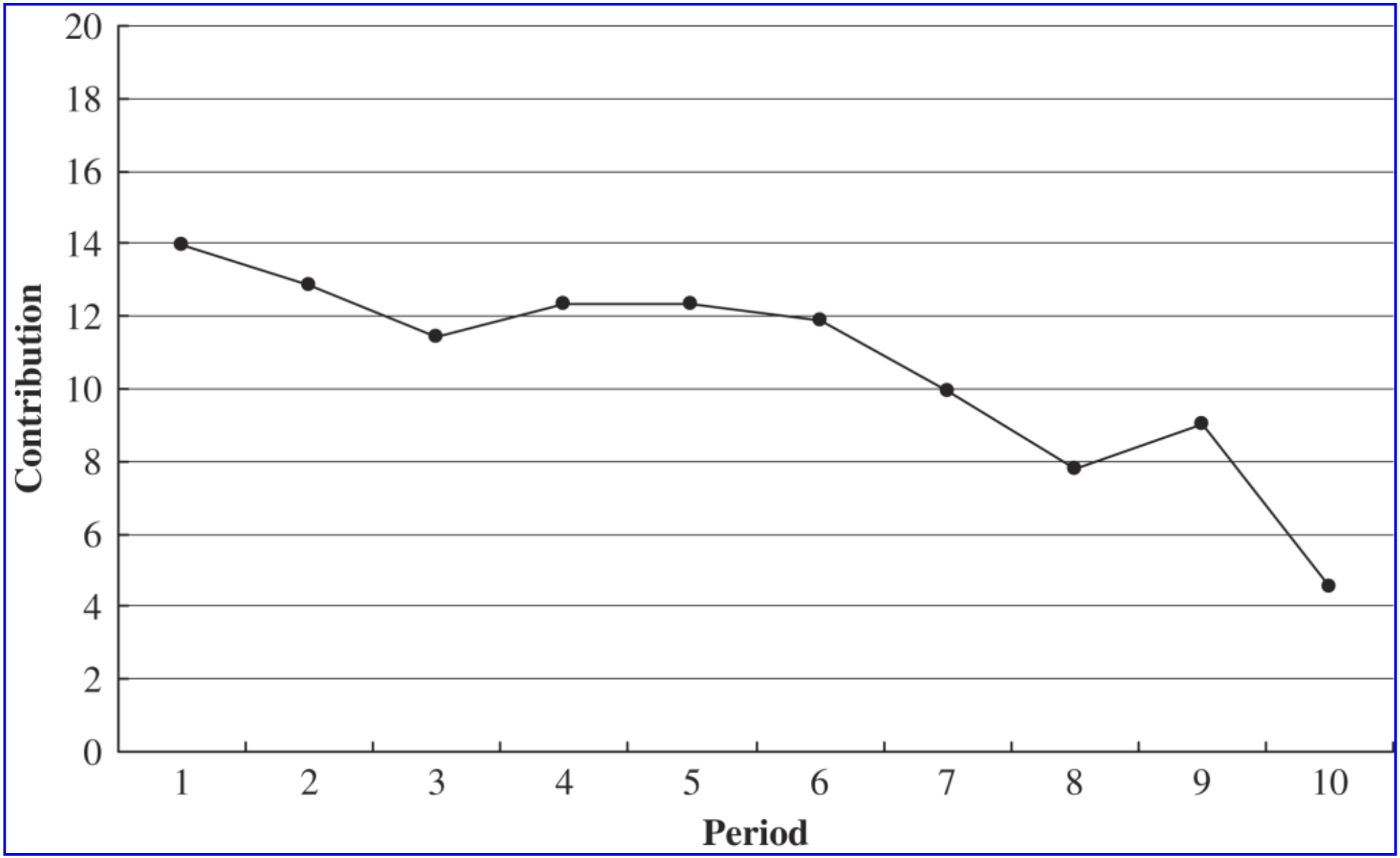

Variations in the parameters have been explored as well; individual papers manipulate group size (Isaac and Walker, 1988b), the ratio of the return from the public good to the return from the private good (Isaac and Walker, 1988b), the existence of communication (Isaac and Walker, 1988a), fixed groups (Andreoni and Croson, 2008), anonymity (Laury, Walker and Williams, 1995) and framing (Andreoni, 1995a). Recent work in this area extends the paradigm to incorporate more realistic assumptions, including heterogeneity of players (Buckley and Croson, 2006), endogenous group formation (Croson, Fatas and Neugebauer, 2005), and punishment/reward (Fehr and Gächter, 2000). Data has been collected from various subject pools, including children (Krause and Harbaugh, 2000) and residents of Asian slums (Carpenter, Daniere and Takahashi, 2004). (For a fascinating look into underappreciated but related psychology literature, see research on *social loafing*, reviewed in Karau and Williams, 1993.)

In summary, the VCM captures the pure tension between individual gains and social efficiency. It is thus used in many settings and by many researchers to investigate the causes (and consequences) of this tension, as well as to describe behaviour in the world.

The provision point mechanism (PPM)

One concern with the VCM is that in equilibrium the public good is not provided at the socially efficient level. Bagnoli and Lipman (1989) discuss a logical response to this problem: add a threshold (or provision point) to the production process. The threshold needed to provide the public good is announced. If at least that much is allocated to the group account, then the public good is produced; if not, no public good is produced.

It is straightforward to see that a VCM can be 'discretized' to the PPM by adding a threshold. With the appropriate parameters (discussed below) this game now has a set of efficient Nash equilibria in which the public good is exactly provided. There are also inefficient equilibria of this game, in which the public good is not provided, but this mechanism nonetheless represents a theoretical improvement over the VCM.

There are some parameter values necessary for the existence of these efficient equilibria. In particular, imagine the threshold is T, the value from private consumption is 1 and individual endowments are E_i. Define v_i as an individual's value from the public good. For an efficient equilibrium there must exist a set of contributions $\{\sigma_i\}$ such that $\Sigma \sigma_i \geq T$. Furthermore, the individual rationality constraints must be satisfied $(\forall i)\sigma_i \leq \min\{E_i, v_i\}$ and providing the public good must be efficient $T \leq \Sigma v_i$.

Additional assumptions are needed before this mechanism is completely described. When the threshold is not reached, the resources contributed to it can be returned or can be lost. This feature has been called the 'money back guarantee' in psychology, and in economics is the *refund* (Isaac, Schmidtz and Walker, 1989; Bagnoli and McKee, 1991). The existence of a refund does not affect the set of efficient equilibria, but does change the set of inefficient equilibria. With no refund, there is one (unique) inefficient equilibrium of zero contribution. With a refund, there are many (weak) inefficient equilibria in which some is contributed towards the public good, but not so much that any player can supplement to reach the threshold. Those contributions are then refunded, making the contributors indifferent between these strategies and contributing zero.

The second dimension is the disposition of resources above the threshold. This is referred to as the *rebate* (Marks and Croson, 1998). Experiments have been run including no rebate (excess contributions are lost), proportional rebates (excess contributions are returned proportionally based on contributions), and utilization rebates (excess contributions are used to provide the public good in a VCM fashion). None of these changes the set of equilibria.

While the PPM has the advantage of the existence of efficient equilibria, it has the disadvantage of too many equilibria. For example, in a typical parameterization used by Croson and Marks (1998), five players each had 55 tokens to allocate. Tokens allocated to the private account earned 1¢ each. If there were at least 125 tokens allocated to the public account, each participant in the group received 50¢. These parameters satisfy the conditions above; the collective benefit from the public good (5 people × 50¢=$2.50) is greater than the social cost of provision ($1.25). There exists a set of allocations such that the public good is provided; one is the unique symmetric equilibrium in which each player allocates 25 tokens, the threshold is exactly met, and each participant receives their value of 50¢, strictly greater than their costs of 25¢.

Unfortunately, this is not the only efficient equilibrium. In particular, the set of allocations {25, 25, 25, 24, 26} is also an equilibrium, as is {25, 25, 25, 26, 24}, although player 4 prefers the former and player 5 the latter. All told, there are 4,052,751 efficient equilibria using these parameters. Thus the main problem in the PPM is not one of *cooperation*; avoiding the inefficient outcome as in the VCM. It's a problem of *coordination*, of choosing which of the many efficient equilibria the group will play.

The coordination problem is difficult enough in the stage game. However, in the lab this game is typically finitely repeated. In the repeated game, the number of potential equilibria grows exponentially, as any sequence of stage-game equilibria are themselves an equilibrium of the repeated game.

In practice, almost no instances of the inefficient equilibria are observed. Group contributions tend to cycle around the efficient equilibrium level, although they are almost equally likely to be above the threshold as below. Examples of group contributions in the PPM can be seen in Figure 2.

Figure 2

Group contributions to public account in PPM, four groups. *Source*: Croson and Marks (2000).

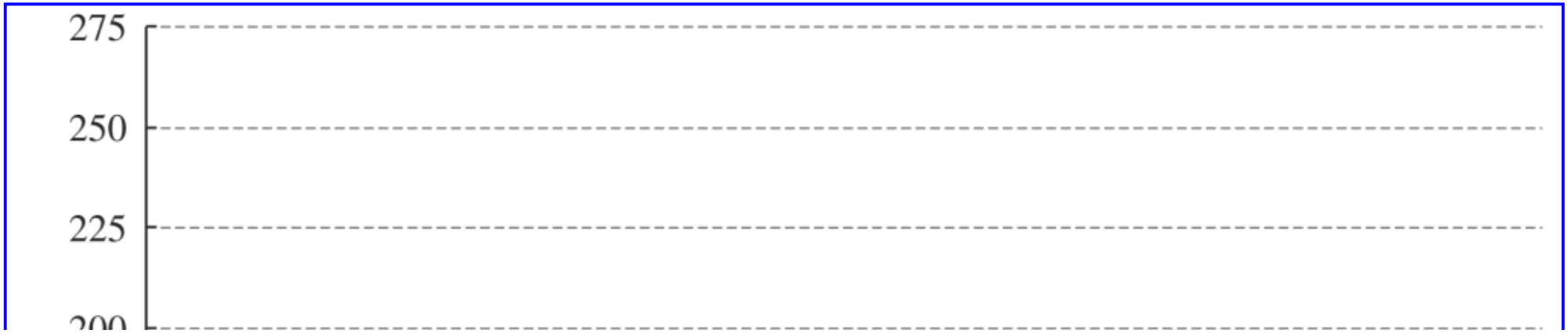

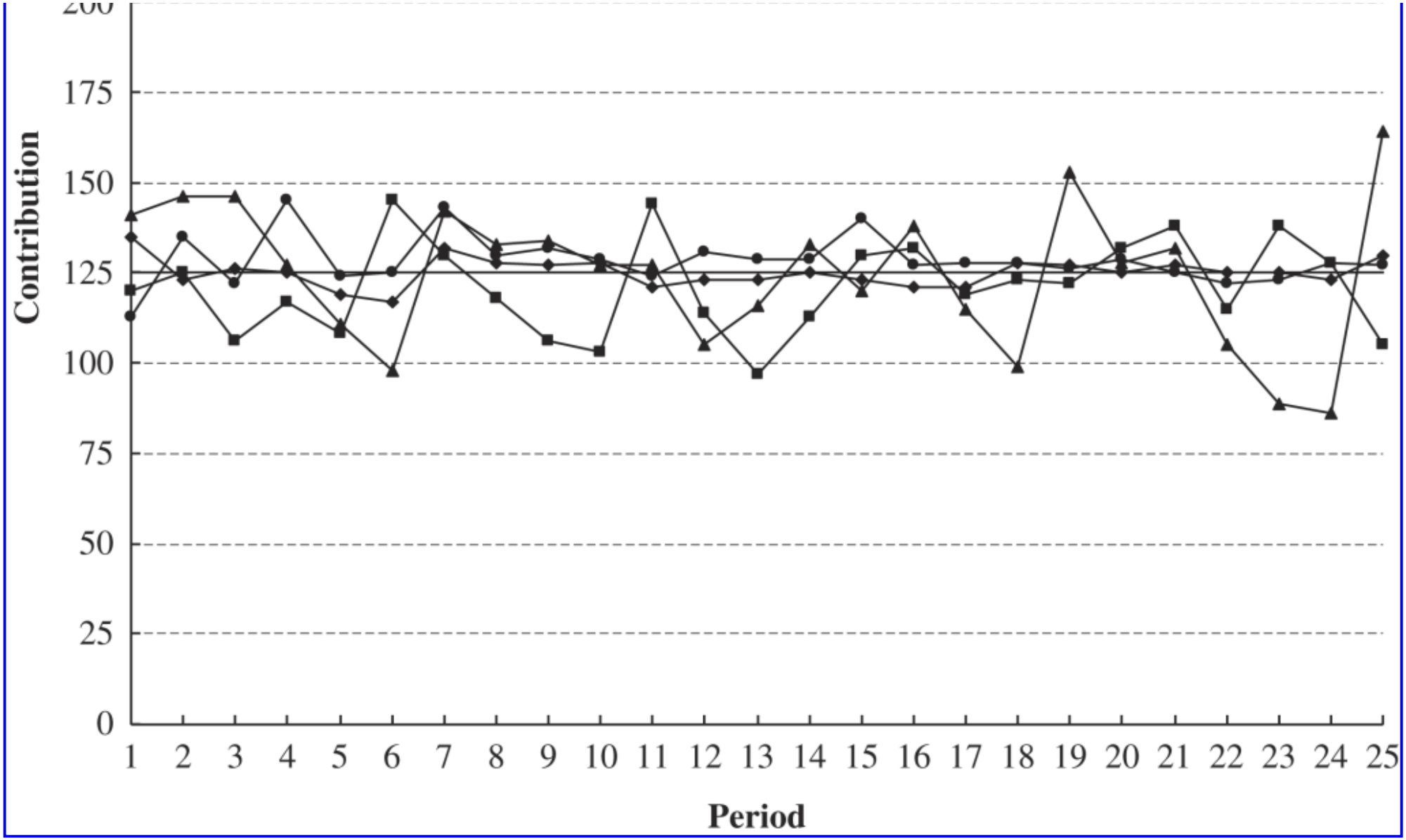

Further research has investigated other dimensions of the PPM. These include the effect of subject pool (Cadsby and Maynes, 1998), binary versus continuous giving (Cadsby and Maynes, 1999), heterogeneous valuations (Croson and Marks, 1999), identifiability of contributions (Croson and Marks, 1998), incomplete information (Marks and Croson, 1999), and framing (Sonnemans, Schram and Offerman, 1998).

The PPM has a number of important and interesting properties. It allows for efficient equilibria, thus to some extent ‘solving’ the public goods problem. However, this solution brings costs: too many equilibria and the need to coordinate among them. This distinction, between the cooperation motive of the VCM and the coordination motive of the PPM is a critical and often-overlooked one.

The common pool resource game

The structure of the CPR game is based on work by Gordon (1954) and Hardin (1968) on the tragedy of the commons. In the typical tragedy, ranchers graze their herds either on their

private land or on the commonly owned land in each village. Since grazing on the commons is free, individuals prefer it to using their own land, which can be used to grow cash crops. However, grazing imposes a negative externality on others; if my cows eat the grass, there is less left for your herd. The CPR game, thus, is a *continuous taking* game; each unit of grass that I take exerts a negative externality on the rest of the village.

Unlike the VCM, the externalities imposed are typically nonlinear, with public costs initially being lower than private benefit, but rising until the two cross. Thus the game has internal equilibria, in which more grazing than is optimal is predicted. (These games are similar to a class of *rent-seeking* games, which have recently been experimentally explored. Rent-seeking games are beyond the scope of this article; but see Önçüler and Croson, 2005, for some recent work.)

In the first CPR economics experiment, Walker, Gardner and Ostrom (1990) arranged subjects into groups of eight. Each participant was given a homogeneous endowment and was told he could allocate this endowment between two markets. Like the VCM, the private market paid a fixed amount, 5¢ per token. The public market (the common pool) had externalities for other group members' consumption. Unlike the VCM this externality was negative rather than positive. Also unlike the VCM, the externality was nonlinear, with increasing social cost. Conceptually, allocating resources to the public market captures the idea of grazing the herd on public land.

When x_i is the amount player *i* allocates to the public market, the earnings from the public market for player *i* are:

$$x_i / \sum x_i \left(23 \sum x_i - 0.25 \left(\sum x_i\right)^2\right).$$

The negative squared term creates the nonlinearity. If no one is allocating resources to the public market, an individual earns more from that market than the private one (the first token allocated there earns 22.5¢ versus 5¢ in the private market). However, this return quickly diminishes, so the value from investing in the public market falls below the value from investing in the private market as the number of tokens increases. This captures the negative externalities. For each token that player *i* invests in the public market, the marginal value of player's *j*'s investment in that market is lowered.

The self-interested, symmetric Nash equilibrium in this game is for each player to invest eight tokens in the public market (for a total investment of 64 tokens). (When each participant invests nine tokens in the public market, the return for that marginal token is exactly 5¢. The authors assume that, when indifferent participants choose not to impose negative externalities on others, thus the equilibrium of eight tokens is used.) This equilibrium prediction is parallel to the prediction of full free-riding in the VCM. In contrast, the symmetric, socially efficient solution is for each participant to invest five tokens in the public market. This is not an equilibrium, however, since each individual privately captures more by investing further in the public market. This capturing is at the expense of the other players, who suffer the negative externality imposed. So five tokens is the socially optimal level, and is parallel to the prediction of full contributing in the VCM.

If behaviour in the CPR were parallel to that in the VCM, we should see allocations to the common market of between eight tokens (the equilibrium) and five tokens (the social optimum). As in the VCM, the stage game described above is repeated finitely many times, either 20 or 30 rounds, depending on the particular parameters. (A parallel literature in CPR games examines *dynamic* versions of the game, in which the resource replenishes itself round to round, with the replenishment rate being dependent on the harvesting rate observed. These are sometimes referred to as *renewable* CPR games. Equilibria in these dynamic games are more complicated, and Herr, Gardner and Walker, 1997, experimentally compare the different games.)

The results from the experiment can be seen in Figure 3. The solid line represents the equilibrium prediction, while the dotted line represents the social welfare maximizing outcome. Unlike the VCM, where contributions lay between these two, here contributions lie on the opposite side of the equilibrium. This indicates excessive allocation to the public market, and excessive negative externalities, over and above the equilibrium prediction.

Figure 3

Average resources allocated to the common pool (CPR). *Source*: Ostrom, Gardner and Walker (1994).

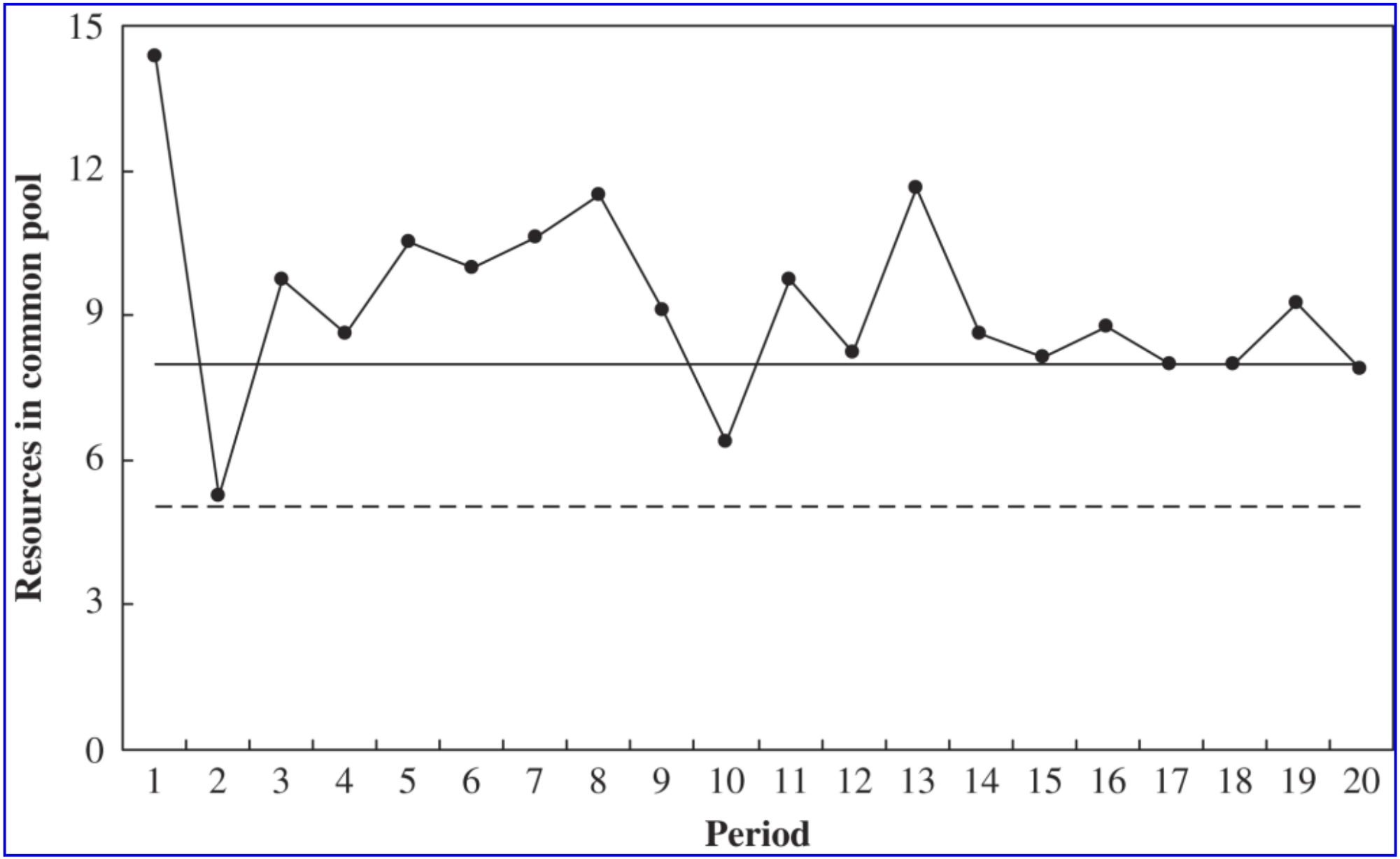

This result of less-than-Nash levels of cooperation is replicated in other experiments, reviewed in Ostrom, Gardner and Walker (1994). Other work also reviewed there examines other questions in CPR games, including probabilistic destruction, communication, monitoring and sanctions, voting and heterogeneity.

One lingering puzzle remains: why are subjects more generous/cooperative than the equilibrium in the VCM and less generous/cooperative than the equilibrium in the CPR game? A number of studies have investigated this question by adding complexity to the VCM to make it resemble the CPR (for example, the stream of research on nonlinear VCM games below). Others investigate framing, suggesting it is the difference between providing a positive externality in the VCM and a negative externality in the CPR game. Unfortunately no study has offered either a definitive experiment or compelling data to explain why the outcomes from these games differ.

Other public goods settings

In addition to the games described above, a small literature explores different types of public goods games. A number of papers examine nonlinear VCMs, with internal equilibria (see Laury and Holt, 2008, for a review). Here the production of public good is nonlinearly related to the amount allocated to the public account. This yields an internal social optimum and Nash equilibrium level of contributions. As before, parameters are set so that the public good is under-provided in equilibrium.

Others have explored markets with externalities rather than public goods per se (for example, Plott, 1983). Still other researchers combine these games in creative ways, for example a PPM with a VCM for excess contributions (as in the utilization rebate of Marks and Croson, 1998), or a PPM with a VCM for under-contributions (as in Vesterlund, Duffy and Ochs 2005).

Finally, a number of papers have experimentally tested other proposed mechanisms for solving the public goods problem. For example, Chen and Plott (1996) provide a test of the Groves–Ledyard mechanism (a mechanism designed to elicit individuals values for public goods). Reviews of experiments using incentive-compatible mechanisms can be found in Chen (2008). These literatures are less developed than the previous three games, a disadvantage when trying to summarize a stream of research but an advantage when seeking a new contribution.

Commonalities and puzzles

The underlying similarity between all public goods experiments is the existence of externalities. These externalities can be positive or negative, and they can be linear, nonlinear or involve thresholds. The decisions participants make can be described as giving or taking. These varying situations affect the equilibrium predictions of the games.

Individuals are ‘cooperative’ in the VCM; they contribute more towards the public good than equilibrium behaviour would predict. There are many explanations for why this may be the case, including altruism, reciprocity (conditional altruism), warm-glow and errors, but no one causal factor has emerged as dominant.

In the PPM, the issue is not one of cooperation but coordination. On average the efficient equilibrium outcomes describe the data. However, there is also ‘gaming’, with groups sometimes failing to provide the public good as one individual attempts to move towards a more attractive equilibrium. Thus, while outcomes from these mechanisms are more efficient than those from the VCM, the coordination problem is severe and unsolved.

Finally, individuals harvest *more* than the Nash equilibrium predictions in CPR games. This result contrasts with the VCM; here individuals are more competitive than the equilibrium prediction. The source of these differences is still unexplored and represents an excellent direction for future research.

Summary

The tension between self-interest and social efficiency is one we experience every day. Experiments like those discussed in this article have been developed to explore how humans resolve this tension. Results from these experiments highlight the impact of different public goods structures, institutional arrangements and repeated interactions on human behaviour. Ultimately they help us to design mechanisms to better provide public goods, and allow for a deeper understanding of human motivations in the wide set of activities involving externalities for others.

See Also

- altruism in experiments
- common property resources
- coordination problems and communication
- experimental economics
- public goods
- reciprocity and collective action

Bibliography

Andreoni, J. 1989. Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy* 97, 1447–58.

Andreoni, J. 1995a. Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110, 1–21.

Andreoni, J. 1995b. Cooperation in public-goods experiments: kindness or confusion? *American Economic Review* 85, 891–904.

Andreoni, J. and Croson, R.T.A. 2008. Partners versus strangers: random rematching in public goods experiments. In Plott and Smith (2008).

Bagnoli, M. and Lipman, B.L. 1989. Provision of public goods: fully implementing the core through private contributions. *Review of Economic Studies* 56, 583–601.

Bagnoli, M. and McKee, M. 1991. Voluntary contribution games: efficient provision of public goods. *Economic Inquiry* 29, 351–66.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82, 1063–93.

Buckley, E. and Croson, R.T.A. 2006. Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics* 90, 935–55.

Cadsby, B. and Maynes, E. 1998. Choosing between a socially efficient and free-riding equilibrium: nurses versus economics and business students. *Journal of Economic Behavior and Organization* 37, 183–92.

Cadsby, B. and Maynes, E. 1999. Voluntary contribution of threshold public goods with continuous provisions: experimental evidence. *Journal of Public Economics* 71, 53–73.

Carpenter, J.P., Daniere, A.G. and Takahashi, L.M. 2004. Cooperation, trust, and social capital in Southeast Asian urban slums. *Journal of Economic Behavior and Organization* 55, 533–51.

Charness, G. and Rabin, M. 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117, 817–69.

Chen, Y. 2008. Incentive-compatible mechanisms for pure public goods: a survey of experimental literature. In Plott and Smith (2008).

Chen, Y. and Plott, C.R. 1996. The Groves–Ledyard mechanism: an experimental study of institutional design. *Journal of Public Economics* 59, 335–64.

Croson, R.T.A. 2000. Feedback in voluntary contribution mechanisms: an experiment in team production. *Research in Experimental Economics* 8, 85–97.

Croson, R.T.A. 2007. Theories of commitment, altruism and reciprocity: evidence from linear public goods games. *Economic Inquiry* (forthcoming).

Croson, R.T.A., Fatas, E. and Neugebauer, T. 2005. Excludability in three public goods games. Working paper, Wharton School, University of Pennsylvania.

Croson, R.T.A. and Marks, M. 1998. Identifiability of individual contributions in a threshold public goods experiment. *Journal of Mathematical Psychology* 42, 167–90.

Croson, R.T.A. and Marks, M. 1999. The effect of heterogeneous valuations for threshold public goods: an experimental study. *Risk, Decision and Policy* 4, 99–115.

Croson, R.T.A. and Marks, M. 2000. Step returns in threshold public goods: a meta- and experimental analysis. *Experimental Economics* 2, 239–59.

Croson, R.T.A. and Marks, M. 2001. The effect of recommended contributions in the voluntary provision of public goods. *Economic Inquiry* 39, 238–49.

Dawes, R. 1980. Social dilemmas. *Annual Review of Psychology* 31, 169–93.

Fehr, E. and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90, 980–94.

Gordon, H.A. 1954. The economic theory of a common property resource: the fishery. *Journal of Political Economy* 62, 124–42.

Hardin, G. 1968. The tragedy of the commons. *Science* 162, 1243–8.

Herr, A., Gardner, R. and Walker, J.M. 1997. An experimental study of time-independent and time-dependent externalities in the commons. *Games and Economic Behavior* 19, 77–96.

Isaac, R.M., Schmidtz, D. and Walker, J.M. 1989. The assurance problem in a laboratory market. *Public Choice* 62, 217–36.

Isaac, R.M and Walker, J.M. 1988a. Communication and free-riding behavior: the voluntary contribution mechanism. *Economic Inquiry* 26, 585–608.

Isaac, M.R. and Walker, J.M. 1988b. Group size effects in public goods provision: the voluntary contributions mechanism. *Quarterly Journal of Economics* 53, 179–200.

Isaac, R.M., Walker, J.M. and Thomas, S.H. 1984. Divergent evidence on free riding: an experimental examination of possible explanations. *Public Choice* 43, 113–49.

Isaac, R.M., Walker, J.M. and Williams, A.W. 1994. Group size and the voluntary provision of public goods: experimental evidence utilizing large groups. *Journal of Public Economics* 54, 1–36.

Karau, S.J. and Williams, K.D. 1993. Social loafing: a meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology* 65, 681–706.

Kollock, P. 1998. Social dilemmas: the anatomy of cooperation. *Annual Review of Sociology* 24, 183–214.

Krause, K. and Harbaugh, W.T. 2000. Children's contributions in public good experiments: the development of altruistic and free-riding behaviors. *Economic Inquiry* 38, 95–109.

Laury, S.K. and Holt, C.A. 2008. Voluntary provision of public goods: experimental results with interior Nash equilibria. In Plott and Smith (2008).

Laury, S.K., Walker, J.M. and Williams, A.W. 1995. Anonymity and the voluntary provision of public goods. *Journal of Economic Behavior & Organization* 27, 365–80.

Ledyard, J. 1995. Public goods: a survey of experimental research. In *Handbook of Experimental Economics*, ed. J. Kagel and A. Roth. Princeton: Princeton University Press.

Marks, M. and Croson, R.T.A. 1998. The effect of alternative rebate rules in the provision point mechanism of voluntary contributions: an experimental investigation. *Journal of Public Economics* 67, 195–220.

Marks, M. and Croson, R.T.A. 1999. The effect of incomplete information and heterogeneity in the provision point mechanism of voluntary contributions: an experimental investigation. *Public Choice* 99, 103–18.

Marwell, G. and Ames, R.E. 1979. Experiments on the provision of public good I: resources, interest, group size, and the free-rider problem. *American Journal of Sociology* 84, 1336–60.

Marwell, G. and Ames, R.E. 1980. Experiments on the provision of public goods II: provision points, stakes, experience, and the free-rider problem. *American Journal of Sociology* 85, 926–37.

Marwell, G. and Ames, R.E. 1981. Economists free ride, does anyone else? Experiments on the provision of public goods, IV. *Journal of Public Economics* 15, 295–310.

Önçüler, A. and Croson, R.T.A. 2005. Rent-seeking for a risky rent: a model and experimental investigation. *Journal of Theoretical Politics* 17, 403–29.

Ostrom, E., Gardner, R. and Walker, J.M. 1994. *Rules, Games, and Common-Pool Resources*. Ann Arbor: University of Michigan Press.

Palfrey, T.R. and Prisbrey, J.E. 1997. Anomalous behavior in public goods experiments: how much and why? *American Economic Review* 87, 829–46.

Plott, C.R. 1983. Externalities and corrective policies in experimental markets. *Economic Journal* 93, 106–27.

Plott, C.R and Smith, V.L., eds. 2008. *Handbook of Experimental Economics Results*. Amsterdam: North-Holland (forthcoming).

Sonnemans, J., Schram, A. and Offerman, T. 1998. Public good provision and public bad prevention: the effect of framing. *Journal of Economic Behavior & Organization* 34, 143–61.

Vesterlund, L., Duffy, J. and Ochs, J. 2005. Giving little by little: dynamic voluntary contribution games. Working Paper, University of Pittsburgh.

Walker, J.M., Gardner, R. and Ostrom, E. 1990. Rent dissipation in a limited-access common-pool resource: experimental evidence. *Journal of Environmental Economics and Management* 19, 203–11.

How to cite this article

Croson, Rachel T. A. "public goods experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000350> doi:10.1057/9780230226203.1365

The New Palgrave Dictionary of Economics Online

public goods

Agnar Sandmo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides a mathematical and diagrammatic exposition of the theory of public goods as originally formulated by Paul Samuelson. It describes the extension of the model to take account of the costs of distortionary taxation, and discusses the concept of the marginal cost of public funds. Different types of public goods (such as mixed goods and local and global public goods) are discussed before a survey of the incentive problems related to preference revelation.

Keywords

benefit theory of taxation; clubs; cost–benefit analysis; direct democracy; distortionary taxation; externalities; free-rider problem; global public goods; international migration; Lindahl equilibrium; local public finance; local public goods; lump-sum taxes; marginal cost of public funds; marginal rate of substitution; marginal rate of transformation; market failure; mixed goods; non-rivalry in consumption; optimal taxation; Pigou, A. C.; public factors of production; public finance; public goods; pure public goods; redistribution of income and wealth; revealed preferences; Samuelson, P. A.; second best; shadow pricing; single-peaked preferences; tax incidence; Tiebout hypothesis; voluntary provision of public goods; voting; willingness to pay

Article

The development by Paul Samuelson (1954; 1955) of the modern theory of public goods must be counted as one of the major breakthroughs in the theory of public finance. In two very short papers Samuelson posed and partly solved the central problems in the normative theory of public expenditure:

1. (a) How can one define analytically goods that are consumed collectively, that is, for which there is no meaningful distinction between individual and collective consumption?
2. (b) How can one characterize an optimal allocation of resources to the production of such goods?
3. (c) What can be said about the design of an efficient and just tax system which will finance the expenditures of the public sector?

None of these questions was entirely new to the literature of public finance. Indeed, more than 250 years ago David Hume (1739) noted that there were tasks which, although unprofitable to perform for any single individual, would yet be profitable for society as a whole, and which could therefore only be performed through collective action. The theme was later taken up by Hume's friend Adam Smith, who maintained that one of the duties of the state consisted in

> erecting and maintaining certain publick works and certain publick institutions, which it can never be for the interest of any individual or small number of individuals, to erect and maintain; because the profit would never repay the expense to any individual or small number of individuals, though it may frequently do much more than repay it to a great society. (Smith, 1776, pp. 687–8)

Apart from this insight, however, the progress made over the next centuries, certainly with regard to problems (a) and (b), was rather modest. From the point of view of the history of ideas, this is hardly surprising. What is required is a satisfactory theory of market failure. But this presupposes a clear understanding of the optimality properties of the market allocation of resources, which was not established until the modern development of Paretian welfare economics in the late 1930s. More was undoubtedly achieved with respect to problem (c), reflecting the fact that problems of tax incidence had been a central area of theoretical analysis ever since the time of the classical economists, and that criteria of just taxation had developed independently of any analysis of the expenditure side of the public budget. Still, Samuelson's formulation was in every respect a great leap forward, presenting an integrated solution to all three problems, and determining the research agenda for the years to come. It is therefore natural to begin by setting out the basic elements of his model.
In a short article it is of course impossible to do justice to the large literature in this field. For more comprehensive surveys the reader is referred to the textbooks by Atkinson and Stiglitz (1980, lectures 16–17) and Myles (1995, ch. 9), and the article by Oakland (1987).

The Samuelson model

The aim of the model is to derive conditions for optimal resource allocation in an economy in which there are two types of goods, private and public. It is worth emphasizing that these terms do not prejudge the respective tasks of the private and public sectors; the analysis at this stage is institution-free and can best be considered as representing the problems of a planner who knows the production possibilities of the economy, the preferences of the consumers and his or her own ethical values. The definition of the two types of goods is technological, not institutional.
The nature of the two types of goods is defined by the equations which give the relationship between individual and aggregate consumption. For private goods the total quantity consumed is equal to the sum of the quantities consumed by the individuals, so that

$$x_j = \sum_{i=1}^{I} x_j^i, \quad (j = 0, \ldots, J) \tag{1}$$

where the superscript refers to individuals and the subscript to commodities. For public goods the corresponding relationship is one of *equality* between individual and total consumption, namely

$$x_k = x_k^i, \quad (i = 1, \ldots, I;\ k = J + 1, \ldots, J + K). \tag{2}$$

Individual preferences, represented by utility functions, are then defined over the quantities consumed of private and public goods, so that we can write the utility of individual *i* as

$$U^i = U^i\left(x_0^i, \ldots, x_J^i, x_{J+1}^i, \ldots, x_{J+K}^i\right) = U^i\left(x_0^i, \ldots, x_J^i, x_{J+1}, \ldots, x_{J+K}\right), \quad (i = 1, \ldots, I). \tag{3}$$

The definition (2) has given rise to some confusion and controversy. Are there actually any goods which can be described by this definition? The usual answer is that there are some cases of ‘pure’ public goods, like national defence, which can indeed be so described; in such cases consumer benefits are directly related to the total availability of the good in question, and the consumption benefits of any one individual do not depend on the benefits enjoyed by others. This property of public goods is usually referred to as non-rivalry in

consumption; given the supply of the good in question, the consumption possibilities of one individual do not depend on the quantities consumed by others as they do in the case of private goods. However, many goods that one naturally thinks of as public turn out on closer inspection to have elements of rivalry. A road may satisfy the definition of a public good as long as volume of traffic is low, but with higher density and consequent congestion this will no longer be the case. Accordingly, several studies have been devoted to the analysis of 'impure' public goods, combining in some way the properties of private and public goods in the original Samuelson definition; we shall return to this below. It should be observed, however, that the Samuelson formulation does not assume that the *benefits* derived from the supply of the public good are the same for all, even though *availabilities* are the same. Neither does it assume that the benefits from public goods are independent of the quantities consumed of private goods. And the elements of rivalry in the road congestion example may be captured by introducing externalities in the consumption of a private good – car use – whose benefits depend on the supply of a public good – the road. Thus, the original Samuelson formulation offers great flexibility of interpretation, and we have been provided with an answer to the first of the main problems noted above.

We now turn to the problem of optimality of resource allocation and begin by characterizing a Pareto optimum for this kind of economy. Since the interesting special features of the model are on the consumption side only, we assume that the conditions for efficient production are satisfied, so that the production possibilities for the economy can be summarized in the transformation or production possibility equation

$$F(x_0, \ldots, x_J, x_{J+1}, \ldots, x_{J+K}) = 0. \tag{4}$$

The problems of Pareto optimality may now be formulated as follows: of all allocations satisfying equation (4), find the allocation which maximizes utility for consumer 1, given arbitrary but feasible utility levels for all other consumers. As shown by Samuelson (1955), the solution can be given an instructive graphical solution in the two-dimensional case. We therefore begin with the case where there are two consumers and one private and one public good. In the upper panel of Figure 1 we have drawn the production possibility curve as well as an indifference curve corresponding to the fixed level of utility for consumer 2; since the two curves intersect, there are obviously a number of allocations which satisfy these two constraints. In the lower panel the curve ab shows the consumption possibilities for consumer 1, the points a and b corresponding to the points of intersection in the upper panel. For any point on U^2 between a and b, it must be the case that the two individuals consume the same amount of the public good, while consumer 1's private good consumption is equal to the vertical difference between the production possibility curve and consumer 2's indifference curve. The best allocation from 1's point of view is then given by the tangency between his indifference curve and the consumption possibility curve in the lower panel. This determines the optimum supply of the public good (x_1^*) and consumer 1's consumption of the private good (x_0^{1*}) as well as the consumption of consumer 2 (x_0^{2*}).

Figure 1

Pareto optimality with one private and one public good

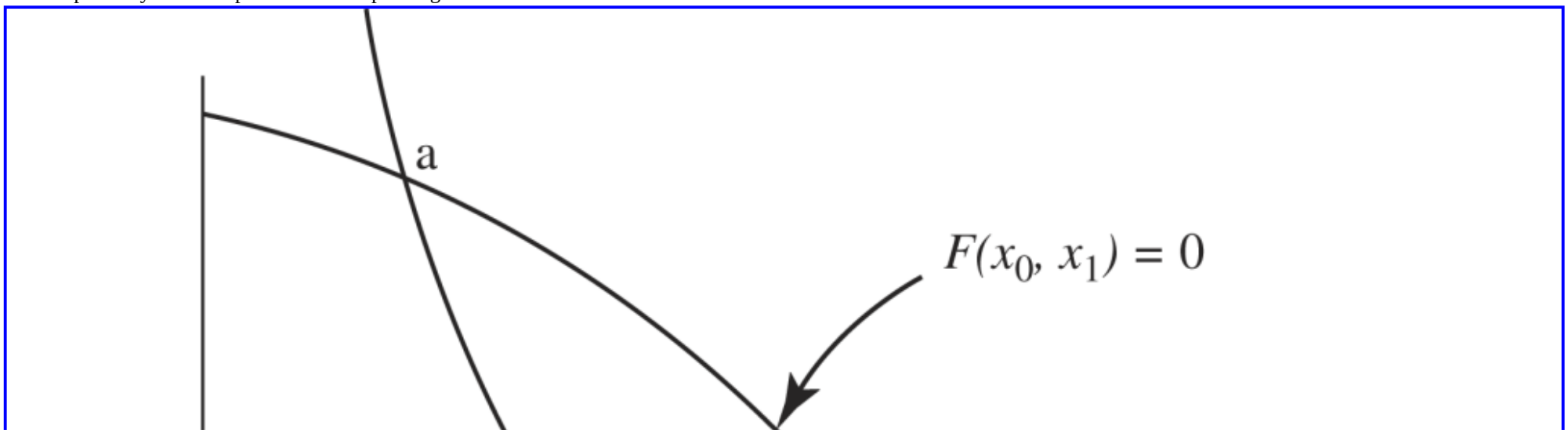

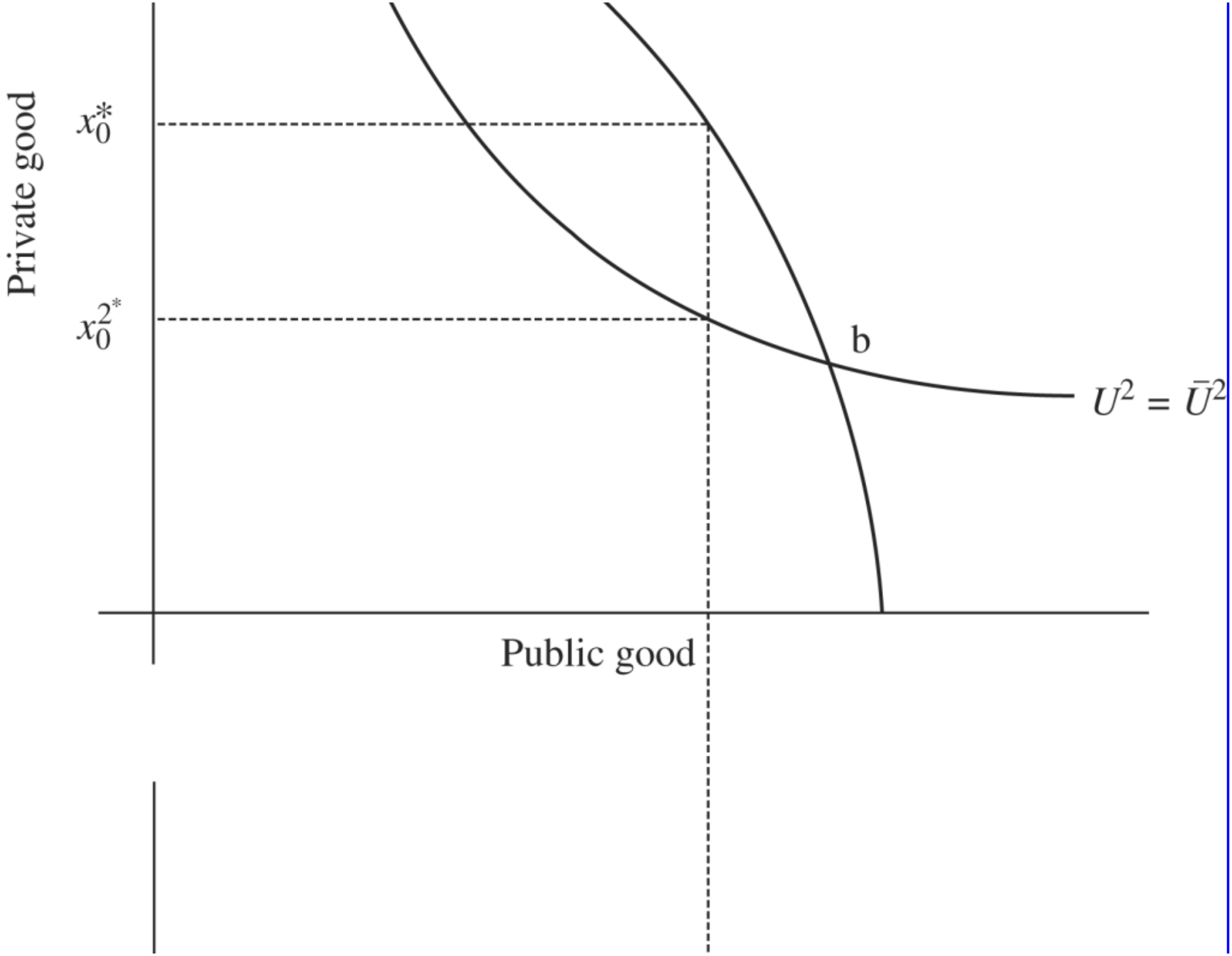
Private good
x_0^*
x_0^{2*}
b
$U^2 = \bar{U}^2$
Public good

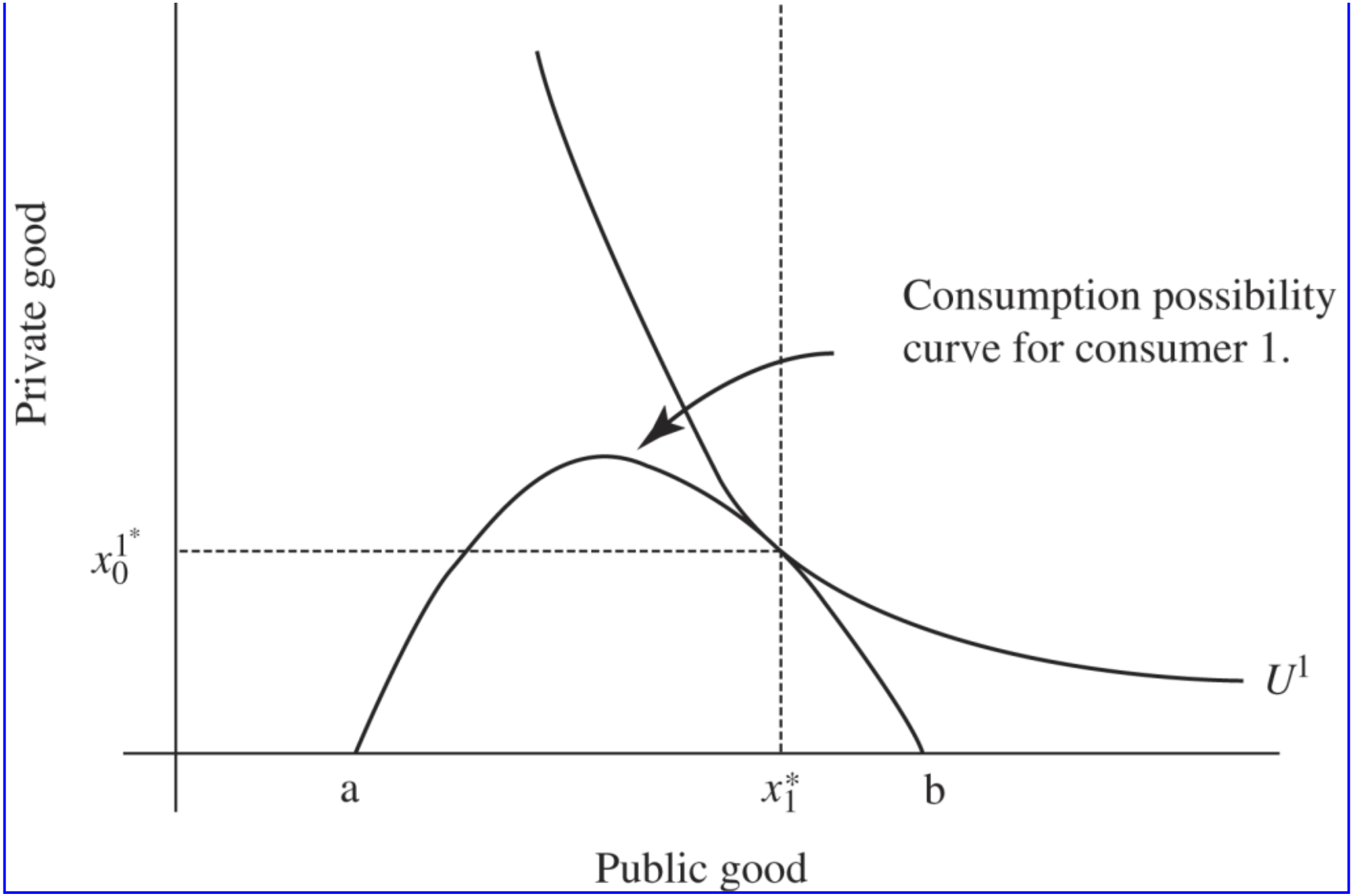

The slope of the consumption possibility curve must of course be equal to the difference of the slopes of the two curves from which it is derived. The tangency point can therefore be characterized in terms of marginal rates of substitution and transformation as

$$MRS^1 = MRT - MRS^2, \text{ or } MRS^1 + MRS^2 = MRT.$$

In more precise mathematical terms this condition can be rewritten (if we let subscripts denote partial derivatives) as

$$\frac{U_1^1}{U_0^1} + \frac{U_1^2}{U_0^2} = \frac{F_1}{F_0}. \tag{5}$$

In words: the sum of the marginal rates of substitution should be equal to the marginal rate of transformation between the public and the private good. Or, since the private good may be taken as a numeraire commodity, the sum of the marginal willingness to pay for the public good should be equal to the marginal cost of production. The intuition should be clear: an extra unit of supply benefits both consumers simultaneously; to find the total marginal benefit we have to take the sum of the marginal benefits accruing to all consumers. Problem (2) has been solved.

The mathematical derivation of the corresponding condition in the general case need not occupy us here. To extend the analysis to more than two consumers, we have only to add more terms on the left-hand side of (5). An increase in the number of public goods simply requires us to introduce similar conditions for every such good. To generalize to an arbitrary number of private goods, we note that for any given allocation of public goods, the allocation of private goods should be a Pareto optimum relative to this, so that the usual marginal conditions must hold. This gives us two sets of first order conditions for Pareto optimality, namely:

$$\frac{U_j^i}{U_0^i} = \frac{F_j}{F_0}, \quad (i = 1, \ldots, I,\ j = 1, \ldots, J) \tag{6}$$

$$\sum_{i=1}^{I} \frac{U_k^i}{U_0^i} = \frac{F_k}{F_0}, \quad (k = J + 1, \ldots, J + K). \tag{7}$$

In the two-dimensional case the first order conditions could be taken to describe a true maximum because the diagrams introduced the required convexity–concavity conditions. In the more general case one has to assume quasi-concavity of the utility functions as well as convexity of the transformation surface for the second order conditions to be satisfied.

There is of course an element of arbitrariness in the concept of Pareto optimality, corresponding to the arbitrary location of consumer 2's indifference curve in Figure 1. The model can be closed by assuming the existence of a social welfare function, and the usual assumption is that this is of the Bergson–Samuelson type, where the arguments of the function are the individual utility levels. Maximizing the welfare function $W(U^1, \ldots, U^I)$ gives as the optimality conditions first (6) and (7) – since a welfare optimum must be a Pareto optimum – and then a set of conditions for optimal distribution of consumption between individuals. These can be written as

$$W_i U_0^i = W_h U_0^h, \quad (i, h = 1, \ldots, I). \tag{8}$$

The marginal social utility of consumption should be the same for all. (Note that although the conditions as stated here refer to the consumption of private good 0, they can be converted, by using conditions (6), to express the equality of the marginal social utility of consumption in terms of any private good.)

Suppose now that private goods are allocated through a system of perfectly competitive markets, and that the allocation of resources to public goods also satisfies the efficiency conditions (7) as the result of some decision procedure which is yet to be specified. Imagine further that at least part of the provision of public goods is undertaken by the public sector, and that taxes are needed to finance this. What is the ideal tax system for this purpose? We wish the tax system to satisfy conditions (8), but these are conditional on the remaining first-order conditions being satisfied. Under competitive conditions the marginal rates of substitution will be equal to consumer prices, if we take commodity 0 to be the numeraire good, while marginal rates of transformation will correspond to producer prices. Thus, conditions (6) will be satisfied in a competitive economy provided that consumer prices are equal to producer prices. But this means that there must be no distortionary taxation; the only taxes that are consistent with a fully optimal solution are lump-sum taxes in amounts independent of all components of demand and supply for consumers and firms. This insight is of course well known from the standard competitive model with private goods only, but it is worth restating in the present context as the answer to problem (c).

This exposition of the basic elements of the Samuelson model can be used to put his contribution into historical perspective. Earlier writers on public finance, for example Mazzola (1890), Sax (1924) and Pigou (1928), did in fact apply marginal utility theory to the problem of the optimal supply of public goods, emphasizing the optimality rule that marginal benefit at the optimum should be equal to marginal cost. They failed, however, to develop a definition of public goods that could be used to characterize the difference between such goods and private goods. For the same reason they were also vague about the nature of the marginal benefit and how to measure it in the absence of market prices. Finally, although there is much interesting discussion by the older writers of the ability to pay and benefit theories of taxation, the efficiency aspect of taxation played a very minor part in their writings, and so they were unable to face the basic problem of how to reconcile the objectives of a just distribution and economic efficiency. With the Samuelson formulation all these issues had been clarified, and the foundation had been laid for further progress.

Distortionary taxation

The above optimality rules hold for the case where taxation is non-distortionary, that is, where taxes are imposed to raise revenue and to redistribute incomes without disturbing the efficiency properties of the price mechanism. For a variety of reasons such taxes are hardly feasible, and it is interesting to consider the modifications that will have to be made if taxes are distortionary. Pigou (1928) argued that the cost of tax distortions should be taken into account in balancing the costs and benefits of public goods supply:

> Where there is indirect damage, it ought to be added to the direct loss of satisfaction involved in the withdrawal of the marginal unit of resources by taxation, before this is balanced against the satisfaction yielded by the marginal expenditure. (Pigou, 1928, p. 34)

As pointed out by Atkinson and Stern (1974), however, this argument is not necessarily correct. Their analysis is an interesting exercise in the theory of the second best.

To abstract from problems of redistribution, consider the case where all individuals are identical. There are two private goods, numbered 0 and 1, and one public good, identified as commodity 2. The representative consumer maximizes his or her utility function $U(x_0, x_1, x_2)$ subject to the budget constraint

$$x_0 + P_1 x_1 = 0. \tag{9}$$

Thus, there is no lump-sum income, and commodity 0 serves as the numeraire. Given the optimum of the consumer, the government maximizes the sum of the utility functions (a special case of the welfare function in the previous section) subject to the constraint that the resource cost of public goods supply equals the tax revenue. Thus, the government maximizes $IU(x_0, x_1, x_2)$ subject to

$$t_1 x_1 = p_2 x_2. \tag{10}$$

Here I is as before the number of consumers, t_1 is the tax per unit of commodity 1 such that $P_1 = p_1 + t_1$. The small p's denote producer prices, which for convenience are taken to be constant, corresponding to constant unit costs of production in terms of the numeraire. The government determines t_1 and x_2 simultaneously.

The analytical details of the model need not concern us here. To understand the result, one should note that from the formulation of the consumer's problem it follows that demand for the taxed good depends on the supply of the public good, so that the demand function can be written as $x_1 = x_1(P_1, x_2)$. Thus, when the supply of the public good is increased, there will be two effects on the demand for private goods. One is the effect via increased availability of the public good, another is the price effect via increased taxation. It can be shown that the condition corresponding to the Samuelson eq. (7) in this case becomes

$$\sum_i \mathrm{MRS}^i = \frac{p_2 - t_1 I(\partial x_1 / \partial x_2)}{1 + (t_1 / x_1)(\partial x_1 / \partial t_1)}. \tag{11}$$

If there is no distortionary taxation, the right-hand side becomes simply p_2, which is the marginal rate of transformation, and we are back to the original Samuelson case. An increase in the tax rate lowers the demand for the taxed good, and the corresponding term in the denominator shows that this ‘blows up’ the cost of the public good; this is the effect alluded to by Pigou. On the other hand, the additional term in the numerator can in principle be of either sign and may therefore reverse Pigou's conclusions. Suppose that $\partial x_1 / \partial x_2$ is positive, meaning that increased supply of the public good increases the demand for the taxed good. Then the relevant social marginal cost of the public good may in fact be lower than the pure resource cost. The point is that in this case the effect of the public good on the demand for the private good serves to counteract the tax effect. The commodity tax is distortionary because it lowers consumption and production of the taxed good. If an increase of the amount of the public good serves to push the quantity of the taxed good back towards its first best optimal level, this could lower the economic cost of production.

This analysis has inspired a considerable literature about the concept of the *marginal cost of public funds* (MCF). To start from the insight provided by the formula (11), it has been suggested that practical calculations of the optimal amount of public expenditure should be based on the formula

$$\sum \mathrm{MRS}^i = \mathrm{MCF} \cdot \mathrm{MC},$$

where MC corresponds to p_2 and the presumption is that $\mathrm{MCF} > 1$. The use of the MCF for practical cost–benefit analysis of public goods provision – one of the more important applications of the pure theory of public goods – would therefore tend to depress the provision of public goods below the level indicated by the Samuelson rule.

This conclusion may be disputed, however. First, it is not clear that eq. (11) supports the hypothesis that the marginal cost of public funds exceeds 1. Even if we assume, which seems reasonable, that the tax elasticity is negative, complementarity between private and public goods $(\partial x_1 / \partial x_2 > 0)$ might lead the right-hand side of (11) to become less than p_2. However, since the sign and magnitude of the complementarity term must be expected to differ between different types of public sector projects, there is a good case for considering this term to be project specific and therefore not to include it in a general measure of the cost of distortionary tax finance. In this view, it is the tax elasticity of demand that is important for the MCF.

Second, there is one feature of the Atkinson–Stern analysis which calls for particular caution in practical application. This is the assumption that the government optimizes with respect to both public goods supply and the tax rate. In principle, therefore, their results are valid only for an optimal tax system, although it can be shown that the formal expression for the MCF is the same also for a non-optimal tax system (see, for example, Sandmo, 1998). More importantly, however, in the more realistic case where there are many tax rates

which have not been chosen optimally, there is no reason to expect that the MCF will be the same for all sources of tax finance. It will therefore be misleading to speak about *the* marginal cost of public funds, as if it were a general characteristic of the whole complex system of direct and indirect tax rates.

Third, in order to focus on the efficiency aspects of the problem, Atkinson and Stern made the assumption that all consumers are identical. But one of the reasons why we have distortionary taxation is that they are not, and that governments try to achieve some measure of redistribution through the design of the tax system. As shown by Sandmo (1998), an explicit modelling that takes account of the redistributive objective leads to a measure of the marginal cost of public funds where the efficiency loss from taxation may, depending on the distributional preferences embedded in the government's policies, be partly or wholly offset by distributive gains.

Types of public good

In line with the original Samuelson formulation we have so far limited the discussion to pure public consumption goods. Various alternative formulations have been discussed in the literature, and we shall briefly discuss some of these.

We have already observed that many consumption goods that may be classified as public turn out also to have important elements of ‘privateness’. This has two aspects. In the first case it may be argued that a public good like a national park cannot really be enjoyed by the individual without expenditure on private goods such as hiking equipment, and that even such an apparently clear case of a public good should be analysed as a mixed case of a private and a public good. To some extent this argument is based on a misunderstanding of the theory. There is no presumption that the benefit that an individual derives from the availability of a public good be independent of his or her consumption of private goods. Still, it may sometimes be useful to model the interaction between private and public goods consumption in a more explicit manner than is done in the standard formulation. One way this can be done is to take as the point of departure the consumption technology approach and assume that there are some final goods such as road trips and nature hikes that are intrinsically private but that are produced by the individual consumer by means of private and public goods inputs. The second aspect of mixed goods is that the benefits enjoyed by any one individual may depend on the consumption of others, as in the cases of a crowded road or a congested national park. This aspect, too, may be handled by the consumption technology approach by letting other people's consumption of complementary private goods enter every individual's production function for the final good in question. This would be a special case of the Samuelson formulation when in addition it is assumed that some private goods create externalities in consumption. Thus, the advantage of the consumption technology approach to the theory of public goods lies not in greater generality, but in a formulation that captures in a more intuitive fashion a natural way of thinking about public goods. An additional advantage is that the theory becomes more closely related to the practice of cost–benefit analysis, where willingness to pay is typically computed not by observing preferences directly, but by calculating the private cost reductions that would follow from an increase in the provision of a public good. The theory is further elaborated in Sandmo (1973); for an alternative formulation of similar ideas see Bradford and Hildebrandt (1977).

Not all public goods are naturally analysed as consumption goods. One of the classical examples, the lighthouse, is more easily interpreted as a producer good or a factor of production. Public factors of production were first introduced into the theoretical literature by Kaizuka (1965), who derives the efficiency conditions analogous to Samuelson's for the production case. Sandmo (1972) shows how the formulation can be used to derive shadow prices for such goods when the private sector is competitive.

The Samuelson formulation implies that the availability of any public good is the same for all individuals and independent of their decisions about private goods consumption – although, as we have noted, the benefit is not. This ignores the fact that many public goods are available only to individuals residing in a particular location, and that an individual may therefore select the amount available of the public good by changing his or her place of residence. This was first pointed out by Tiebout (1956) in a paper which has since given rise to a rich literature on the important topic of local public goods and, more generally, local public finance. We shall return below to the demand-revealing aspects of mobility between communities. But it is worth noting here that, although the original application of the basic idea was to individual choice among residential communities, there are possibilities of application to other interesting areas as well. In the labour market, workers’ choice among firms might be affected by public good aspects of the working environment which are specific to the individual firm. Following Buchanan (1965), ‘clubs’ has become the generic term for voluntary associations of individuals whose purpose is to provide the members with a public good. Internationally, country-specific public goods might influence the pattern of international migration; in this perspective, almost all public goods would be local, and the original formulation becomes a special case characterized by geographical immobility of the population. For surveys of the theory of clubs and local public goods the reader is referred to Rubinfeld (1987) and Scotchmer (2002).

At the other end of the scale from local public goods are global public goods, goods that provide benefits to the whole of the world's population. Examples of such goods are international security, global environmental quality and scientific knowledge. One might perhaps think that in this case the theory is directly applicable, since the complications associated with geographical mobility are ruled out by assumption. On the other hand, additional problems arise because the world is not one jurisdiction but composed of a number of independent nation-states. In the original Samuelson formulation, the economy is at its production possibility frontier; this is evidently a strong assumption even for a national economy, and it becomes even more unrealistic when applied to the world as a whole. Moreover, Samuelson assumed redistribution in the form of individualized lump-sum taxes and transfers; this also is an assumption which is much farther from reality when considered in a global context. Even the assumption of redistribution via progressive taxation, which is a more realistic description of national redistribution policy, is far from the economic realities of the international community of countries.

It can be shown that the problems of global production efficiency and redistribution are in fact interrelated, as one would in fact expect on the basis of the theory of the second best;

see Sandmo (2003). If one takes the viewpoint of global welfare maximization and assumes that there are perfect lump-sum transfers both within and between countries, the Samuelson optimality conditions must hold for the world as a whole. In particular, there will be global production efficiency, and the social marginal utility of income must be the same for all individuals. However, if for some reason the international transfers are not made, then production efficiency is in general not desirable. If one assumes that the global welfare function displays inequality aversion, poor countries should not be required to contribute as much to the production of global public goods as their comparative advantage would otherwise call for. But the model also points to a serious problem of incentives, because each country, in deciding how much to contribute to the production of global public goods, finds itself in a strategic situation similar to that of the single individual in the nation state, who has an incentive to be a free rider on the contributions of others (see below). At least if one assumes that national governments are motivated by a fairly narrow concept of national self-interest, there is likely to be an under-supply of global public goods.

Equilibria with public goods

We have concentrated on the theory of public goods as an extension of welfare economics; the central question has been how to characterize optimal or efficient allocations in economies with public goods. But just as in the case of private goods it is interesting to go on from there to consider the equilibrium allocations that would follow from particular institutional arrangements in the economy and to compare these with the optimality conditions. Thus, the theory of public goods ought to be positive as well as normative, a view emphasized strongly in the influential contributions by Buchanan (for example, 1968).

The first clear formulation of a theory of public expenditure which can be given a positive interpretation was presented by Erik Lindahl (1919), who in turn was inspired by Wicksell (1896); an important modern exposition is that of Johansen (1963). In this formulation, individuals bargain over the level of public goods supply simultaneously with the distribution of the cost between them. The bargaining equilibrium is Pareto optimal, implying that the efficiency conditions (7) are satisfied. In addition, each individual pays a price in terms of private goods which is equal to his or her marginal willingness to pay. Formally, let π^i_{J+k} be the price which individual *i* pays for public good *k*, and let p_{J+k} be the producer price or marginal cost. Then the Lindahl equilibrium will be characterized by the condition

$$\sum_i \pi^i_{J+k} = p_{J+k}, \quad (k = 1, \ldots, K) . \tag{12}$$

Thus, at first glance the concept of a Lindahl equilibrium seems to establish an analogue to competitive markets for private goods with the interesting difference that prices should differ from one individual to another, depending on their marginal willingness to pay. This also ties in with older notions of the benefit theory of taxation, according to which taxes were seen as payments for public goods, to be levied in accordance with the benefits which each individual derived from them.

At the technical level it may be noted that there is an interesting ‘duality’ between the definitions of private and public goods on the one hand and the properties of equilibrium prices on the other. In terms of quantities, for private goods the sum of individual quantities consumed *adds up* to the quantity produced, while for public goods individual consumption *equals* aggregate production. In terms of prices, on the other hand, for private goods each consumer price *equals* the producer price, while for public goods individualized consumer prices *add up* to the producer price.

There is, however, one crucial difference between a Lindahl equilibrium and a competitive equilibrium for private goods. With private goods, individuals facing given prices have clear incentives to reveal their true preferences by equating their marginal rates of substitution to relative prices. Without paying, the individual is excluded from enjoying the benefits of consumption. With public goods this no longer holds. Because individuals have the same quantity of public goods available to them whether they pay or not, they have an incentive to misrepresent their preferences and be free-riders on the supply paid for by others. Moreover, this problem is likely to be particularly severe when the number of individuals is large, since an individual contribution will then make little difference to the total supply. The connection between Lindahl equilibria and the game theoretic concept of the core was discussed by Foley (1970); see also the survey by Milleron (1972).

The equilibrium of the Lindahl model is not compatible with individual incentives to reveal preferences truthfully; for this reason Samuelson (1969) has referred to the individual Lindahl prices as pseudo-prices and to the equilibrium as a pseudo-equilibrium. In this case one would conjecture that, because all individuals have the same incentives to understate their true marginal willingness to pay, the Lindahl mechanism would result in equilibrium levels of public goods supply which would be too low relative to the optimum. But there is really no need to associate the problem of preference revelation with this procedure alone; as another extreme, one might think of the case where individuals are asked to state their preferences on the assumption that the cost to them is completely independent of their stated willingness to pay, but there is a positive association between this and the quantity

supplied. Then there will be incentives to exaggerate the willingness to pay and a consequent tendency towards oversupply. Thus, the general problem which arises is how to design a mechanism that will allow the decision-maker to implement the efficiency condition.

Various solutions to this problem have been discussed in the literature. The most practically oriented solution is that of cost–benefit analysis, which takes as its point of departure that people's preferences for public goods are revealed in the market through their demands for complementary private goods (see above). But in theoretical terms it has been shown that this will be true only on certain rather restrictive assumptions about technology and preferences. Another solution is represented in the literature on local public goods, where it has been suggested that people reveal their preferences for public goods by moving to the community offering them their most preferred combination of taxes and public goods. But whether this process will result in an optimum satisfying the efficiency conditions must clearly depend first on how the supply of public goods is determined within each community and second on whether there are enough communities to satisfy the variations of preferences in the population as a whole. Thus, in general, observation neither of the consumption of private goods nor of individuals' mobility between local communities provides reliable information on preferences.

Presumably as a response to the problem of market failure, decisions on public goods supply are largely made by political processes. In a democracy, the natural decision-making process to study is that of voting, and there is by now a substantial literature on this. Most of this is concerned with the stylized situation where public goods supply is determined by majority voting with the consumers themselves being the voters; thus, 'direct democracy' is assumed. The first paper in this area was that of Bowen (1943), who also considered the question of when a voting equilibrium would be Pareto optimal. Later contributions have emphasized that very restrictive assumptions on preferences are sometimes required for a voting equilibrium to exist, and these – like the so-called single-peakedness assumption – are not always attractive in the public goods context. Nevertheless, voting models have become quite popular in descriptive analyses of public goods decisions, particularly at the local government level.

There has also been a great deal of interest in studying planning procedures whereby individuals find it in their own interest to reveal truthfully their preferences for public goods. The first discussion of such a procedure – although in a somewhat different context – was that of Vickrey (1961), but the more recent developments are based on the work of Clarke (1971) and Groves (1973). It is shown there that truthful preference revelation will result if individuals pay a tax on the marginal unit demanded of the public good which is equal to the difference between the marginal cost and the sum of the marginal benefits received by all other individuals. These procedures are of great theoretical interest, perhaps mainly because they clarify the nature of the free-rider problem. However, at present they seem rather far from the state where they could be implemented in practical situations; they would probably be administratively costly to operate, and they also make heavy demands on individual consumers' ability to understand and participate in the process. For surveys of this area see Tulkens (1978) and Laffont (1987).

Doubts have occasionally been voiced on whether the free-rider problem has been given too much prominence in the theoretical literature. Johansen (1977) has argued that there is no clear evidence that this is seen as a major problem in practical public sector decision-making, and suggests that individuals are much more likely to reveal their true willingness to pay than the literature indicates. This is so, he argues, both because truthfulness is a strong social norm and because it is a simple strategy that does not rely on complicated strategic considerations. There is also some empirical evidence from experimental situations to suggest that the revealed willingness to pay is not very sensitive to the associated method of cost distribution; see Bohm (1972).

The point of view taken in most of the literature considered here is that the incentive revelation problem requires decisions on public goods supply to be taken by some governmental body. However, starting with Olson (1965), there has emerged a literature on the voluntary provision of public goods. This literature is perhaps most naturally interpreted as concerned with relatively small groups, in which the incentive to free ride is limited, and not with public goods provision on a national scale. In the framework of this theory, as formulated for example by Bergstrom, Blume and Varian (1986), the decision to contribute to a public good is formulated in the standard framework of consumer demand theory. Consumers allocate their incomes between private goods and contributions to public goods, which are made under assumption that the contributions of all other consumers are taken as given, and one can then study the properties of the resulting Nash equilibrium. Particular attention has been given to the effect on contributions of a redistribution of income; as first shown by Warr (1983), under some assumptions this will change individual contributions in such a way that the aggregate supply of the public good is unaffected.

Perspectives

The Samuelson theory of public goods has been of decisive influence for the theory of public expenditure, which was developed in a number of directions during the second half of the 20th century. The extensions and reinterpretations of the original theory to the cases of public factors of production, mixed (public–private) goods and local and global public goods have significantly increased the applicability of the theory. Much has also been achieved to enrich our understanding of the incentive problems that arise in actual allocation mechanisms for public goods supply; however, it is probably fair to say that the normative theory of public goods has become much more satisfactory from a theoretical point of view than the positive theory. This state of affairs may in fact be unavoidable. The normative theory has little need to model institutional details and can thus be given a more unified appearance. A positive theory, on the other hand, must to a greater extent model economic and political institutions, and there is no single institution corresponding to the competitive market in the private goods case which can serve as a unifying benchmark for the analysis. Moreover, development of the positive theory of public goods must necessarily be closely tied to the progress of the positive theory of public sector behaviour in general; it will be interesting to see whether this theory can be developed to provide descriptive models of public goods provision that are both realistic and reasonably simple.

See Also

- incentive compatibility
- Lindahl equilibrium
- public choice
- public goods experiments

Bibliography

Atkinson, A.B. and Stern, N. 1974. Pigou, taxation and public goods. *Review of Economic Studies* 41, 119–28.

Atkinson, A.B. and Stiglitz, J.E. 1980. *Lectures on Public Economics*. New York: McGraw-Hill.

Auerbach, A.J. and Feldstein, M., eds. 1987. *Handbook of Public Economics*, vol. 2, Amsterdam: North-Holland.

Bergstrom, T., Blume, L. and Varian, H. 1986. On the private provision of public goods. *Journal of Public Economics* 29, 25–49.

Bohm, P. 1972. Estimating demand for public goods: an experiment. *European Economic Review* 3, 111–30.

Bowen, H.R. 1943. The interpretation of voting in the allocation of economic resources. *Quarterly Journal of Economics* 58, 27–48.

Bradford, D.F. and Hildebrandt, G.G. 1977. Observable preferences for public goods. *Journal of Public Economics* 8, 111–31.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* 33, 1–14.

Buchanan, J.M. 1968. *The Demand and Supply of Public Goods*. Chicago: Rand McNally.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11, 17–33.

Foley, D.K. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38, 66–72.

Groves, T. 1973. Incentives in teams. *Econometrica* 41, 617–31.

Hume, D. 1739. *A Treatise of Human Nature*. London: J. Noon.

Johansen, L. 1963. Some notes on the Lindahl theory of determination of public expenditure. *International Economic Review* 4, 346–58.

Johansen, L. 1977. The theory of public goods: misplaced emphasis? *Journal of Public Economics* 7, 147–52.

Kaizuka, K. 1965. Public goods and decentralization of production. *Review of Economics and Statistics* 47, 118–20.

Laffont, J.-J. 1987. Incentives and the allocation of public goods. In Auerbach and Feldstein (1987).

Lindahl, E. 1919. *Die Gerechtigkeit der Besteuerung*. Lund: Gleerup. Partial translation as ‘Just taxation: a positive solution’ in Musgrave and Peacock (1958).

Mazzola, U. 1890. The formation of prices of public goods. In Musgrave and Peacock (1958).

Milleron, J.-C. 1972. Theory of value with public goods: a survey article. *Journal of Economic Theory* 5, 419–77.

Musgrave, R.A. and Peacock, A.T., eds. 1958. *Classics in the Theory of Public Finance*. London: Macmillan.

Myles, G.D. 1995. *Public Economics*. Cambridge: Cambridge University Press.

Oakland, W.H. 1987. Theory of public goods. In Auerbach and Feldstein (1987).

Olson, M. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Pigou, A.C. 1928. *A Study in Public Finance*, 3rd edn. London: Macmillan, 1947.

Rubinfeld, D.L. 1987. The economics of the local public sector. In Auerbach and Feldstein (1987).

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36, 387–9.

Samuelson, P.A. 1955. Diagrammatic exposition of a theory of public expenditure. *Review of Economics and Statistics* 37, 350–6.

Samuelson, P.A. 1969. Pure theory of public expenditure and taxation. In *Public Economics*, ed. J. Margolis and H. Guitton. London: Macmillan.

Sandmo, A. 1972. Optimality rules for the provision of collective factors of production. *Journal of Public Economics* 1, 149–57.

Sandmo, A. 1973. Public goods and the technology of consumption. *Review of Economic Studies* 40, 517–28.

Sandmo, A. 1998. Redistribution and the marginal cost of public funds. *Journal of Public Economics* 70, 365–82.

Sandmo, A. 2003. International aspects of public goods provision. In *Providing Global Public Goods*, ed. I. Kaul et al. New York: Oxford University Press.

Sax, E. 1924. The valuation theory of taxation. In Musgrave and Peacock (1958).

Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of Public Economics*, vol. 4, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Oxford University Press, 1976.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Tulkens, H. 1978. Dynamic processes for public goods: an institution-oriented survey. *Journal of Public Economics* 9, 163–201.

Vickrey, W.S. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16, 8–37. Repr. in W.S. Vickrey, *Public Economics*, ed. R. Arnott, K. Arrow, A. Atkinson and J. Drèze. Cambridge: Cambridge University Press, 1994.

Warr, P. 1983. The private provision of a public good is independent of the distribution of income. *Economics Letters* 13, 207–11.

Wicksell, K. 1896. *Finanztheoretische Untersuchungen*. Partially translated as ‘A new principle of just taxation’. In Musgrave and Peacock (1958).

How to cite this article

Sandmo, Agnar. "public goods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000245>
doi:10.1057/9780230226203.1364

The New Palgrave Dictionary of Economics Online

public infrastructure

Teresa Garcia-Milà and Therese J. McGuire

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Numerous empirical studies have investigated the contribution of public infrastructure (the stock of publicly provided physical capital) to private economic productivity and growth. Using aggregate time-series data to estimate a production function with private capital, labour and public capital as inputs, the authors found substantial elasticities of private output with respect to public infrastructure. The result did not withstand scrutiny. Studies using disaggregated data (by region and industry), employing econometric diagnostics testing for nonstationarity, fixed effects and endogeneity, and using natural-experiment techniques found public infrastructure's contribution to economic growth to be minor.

Keywords

aggregate production functions; Cobb–Douglas functions; cost functions; output elasticity of capital; productivity growth; public infrastructure

Article

Public infrastructure (the stock of publicly provided physical capital comprising highways, sewage and sanitation systems, water systems, school buildings, hospitals and so forth) comprises an important component of the US economy. In 2004, there were just under seven trillion dollars of public capital in the United States, including 1.7 trillion dollars of highways and streets. By comparison, the stock of private capital stood at 27 trillion dollars.

Public infrastructure has figured in several related economic enquiries. What are the causes of private-sector productivity growth and decline? What causes lesser developed regions or countries to grow? Do country growth rates tend to converge over time? Research into each of these questions has examined the role of public infrastructure.

The contribution of public infrastructure to economic productivity and growth has been the focus of many empirical studies in recent years. The idea that public infrastructure should be considered an input in the aggregate production function, together with labour and private capital, was introduced in early

theoretical models, but it was not taken into account in empirical work until the late 1980s. The increased attention to empirical analysis was linked to concerns about the decrease in productivity observed in the United States after 1970. In Europe much of the infrastructure literature has examined the role of public investment in boosting the growth of less developed regions.

The theoretical framework

The most widely used framework for studying the impact of public infrastructure on productivity and growth has been estimation of aggregate production functions where public capital is considered a production input along with the standard inputs, labour and private capital.
The general form of the aggregate production function can be written as follows:

$$Y_{it} = A_{it} F(L_{it}, Kp_{it}, Kg_{it})$$

where *Y* is a measure of output, *L* represents labour, *Kp* is the stock of private capital, *Kg* is the stock of public capital, *A* is total factor productivity and the subscripts allow for regional and time variation. Although the production function can take many functional forms, most empirical studies have estimated a Cobb–Douglas production function. Under that specification, and taking a logarithmic transformation, the estimating equation can be written as follows:

$$y_{it} = a_{it} + \alpha kp_{it} + \beta kg_{it} + \gamma l_{it} + \varepsilon_{it}$$

where the variables are measured in natural logarithms, and ε is an error term.
The aim of an important part of the public infrastructure literature is to estimate the output elasticity of public capital β, so as to measure its contribution to private productivity. An alternative approach to obtaining similar parameters, based on the duality of production and cost functions, is to estimate a cost function.

The empirical evidence

The first widely known results were those obtained by Aschauer (1989), who estimated a production function using aggregate post-war time series data for the United States. He estimated an output elasticity of public capital of 0.39, larger than the corresponding value for private capital (0.35). These estimates imply large returns for public investment (above 60 per cent, double of those for private capital), and were challenged by many authors, who considered them implausibly high.
Some authors attributed the high output elasticity of public capital found by Aschauer (and by Munnell, 1990a) to a spurious correlation between output and public capital due to a common time trend. Aaron

(1990) and Tatom (1991) corrected for the common time trend by first-differencing the data and obtained small and statistically insignificant coefficients.

The incorporation of state or metropolitan level data adding cross-section information to the time-series data opened up new possibilities for handling the spurious correlation and getting more accurate estimates of the contribution of public capital. Eberts (1986) focused on the manufacturing sector for 38 metropolitan areas and obtained a significant but much smaller estimate of the output elasticity of public capital, with a value of 0.03. Munnell (1990b) and Garcia-Milà and McGuire (1992) pooled state and time variation and obtained public capital elasticity estimates that range between 0.04 to 0.15. These estimates struck researchers as being more reasonable; however, they were subject to criticism because of endogeneity problems related mainly to omission of state-specific characteristics and to reverse causality.

Holtz-Eakin (1994), Evans and Karras (1994), and Garcia-Milà, McGuire and Porter (1996) used panel-data techniques not only to take into account state-specific productivity differences in the estimation, but also to explore non-stationarity of the data and possible endogeneity of the production factors. In all cases the estimates of the output elasticity of public infrastructure dropped dramatically compared with the time-series and pooled-data estimates, with values close to zero (and sometimes even negative) and not statistically significant.

By disaggregating by industry, Fernald (1999) avoided the endogeneity problem: if road infrastructure grew as a result of overall economic growth, and therefore the causality were reversed, one would not expect to find a relationship between increases in road infrastructure and the productivity of some industries but not others. He found that an increase in road infrastructure enhances productivity growth of vehicle-intensive industries much more than other industries. Fernald concluded that the US interstate building of the late 1950s and 1960s was one important factor in explaining the productivity increases up to the early 1970s, but the impact of road-building after the main network was built was small. A productivity burst because of road-building is a one-time effect and cannot be historically repeated. This is also the view of Hulten and Schwab (1993), who argue that, once the basic network is constructed, which has a major impact on the economy of the country, additional road construction has little, if any, effect on private productivity.

The results of Mas et al. (1996) support the idea that the impact of infrastructure investment is greater at earlier stages of development of the infrastructure network. They examine regions in Spain over the period 1964–91 and find that the output elasticity of productive infrastructure is 0.14 in the first ten years of the sample, but falls as more recent years are added to the sample, with a value of 0.08 when the whole period is considered. As the highway network in Spain was not yet completed in 1991, their results for Spain are compatible with those obtained for the United States, where the highway network was virtually complete, which showed that highway construction produced little or no effect.

Another possible response to the endogeneity problem is to estimate aggregate cost functions. The estimation of aggregate cost functions avoids the endogeneity bias if one can assume that prices of inputs are exogenous. Although it is reasonable to assume that individual firms are input price takers, it may not be plausible to assume that input prices are exogenous when considering aggregate (state or national level) cost functions. In spite of this shortcoming, there are several interesting papers that estimate aggregate cost functions and obtain, through the duality between cost and production functions, estimates of the output elasticity of private and public inputs. Berndt and Hansson (1992), Lynde and

Richmond (1992), Nadiri and Mamuneas (1994), and Morrison and Schwartz (1996) are good examples of cost function estimations. They differ in the functional specification, the geographical and industrial scope, and the scope of public infrastructure, but in all cases they find that public capital reduces costs and therefore improves productivity. The size of the effect tends to be quite small, along the lines of the production-function estimates obtained by Eberts (1986), Munnell (1990b) and Garcia-Milà and McGuire (1992).

A very different way to avoid endogeneity bias is to look for a natural experiment related to infrastructure investment. The idea is to compare the economic performance of two otherwise similar areas except that in one area, the control group, there has not been any highway construction whereas the other area, the treated group, has experienced highway construction. Rephann and Isserman (1994) apply matching techniques to analyse the effectiveness of US interstate highways as an economic development tool. Results vary significantly depending on the characteristics of the counties considered. Counties close to a large city or containing small cities (more than 25,000 residents) benefit from new investments in interstate highways, while rural counties without these characteristics do not experience economic growth when interstate highways are built within them. Chandra and Thompson (2000) exploit the fact that in the United States much interstate highway construction was designed to link major metropolitan areas. Thus, the rural counties in between the metropolitan areas through which the interstates run can be considered the treatment group, while the counties adjacent to the treatment group, which in essence just missed having an interstate highway run through them, can be considered as a suitable control group. The authors find that counties that have a new interstate highway running through them experience an increase in overall earnings, while earnings fall in the adjacent counties. The authors conclude that interstate highway construction affects the spatial allocation of economic activity, but has no net effect on the economic development of non-metropolitan areas as a whole.

The question of whether investment in public capital yields a net positive social return has also been addressed. Morrison and Schwartz (1996) calculate a measure of the net social return to public infrastructure investment as the difference between the cost savings to manufacturing firms minus the cost to society of the public capital investment. Their results range from small positive values to negative estimates of the net social return depending on how the price of public capital is adjusted for taxation and for the marginal cost of public funds. Haughwout (2002) examines the impact of public infrastructure on both productivity and consumer utility in a sample of large US cities. He finds that the local benefits of public capital are largely realized by households rather than firms and that the aggregate benefits of large investments in public infrastructure are not likely to be sufficient to offset the costs.

Concluding remarks

Based on the aggregate analysis, we can conclude very little. The most credible aggregate production–function estimates of the impact of public infrastructure on private output hover around zero, as do estimates of the net social benefit of public infrastructure investment. However, when the focus is on sub-aggregates such as particular industries or certain areas or incomplete networks, researchers tend to find that public capital investment boosts private output and productivity for some industries, some areas, and some networks. What is clear from the accumulated evidence is that public infrastructure is not a panacea for all that ails economies, but rather a tool that when properly targeted can be effective at

enhancing growth.

Bibliography

Aschauer, D.A. 1989. Is public infrastructure important? *Journal of Monetary Economics* 23, 177–200.

Aaron, H.J. 1990. Discussion of ‘why is infrastructure important?’ In *Is There a Shortfall in Public Capital Investment?* ed. A.H. Munnell. Boston: Federal Reserve Bank of Boston.

Berndt, E.R. and Hansson, B. 1992. Measuring the contribution of public infrastructure capital in Sweden. *Scandinavian Journal of Economics* 94(Supplement), 151–68.

Chandra, A. and Thompson, E. 2000. Does public infrastructure affect economic activity? Evidence from the rural interstate highway system. *Regional Science and Urban Economics* 30, 457–90.

Eberts, R.W. 1986. Estimating the contribution of urban public infrastructure to regional growth. Working Paper No. 8610, Federal Reserve Bank of Cleveland.

Evans, P. and Karras, G. 1994. Are government activities productive? Evidence from a panel of U.S. states. *Review of Economics and Statistics* 76, 1–11.

Fernald, J.G. 1999. Road to prosperity? Assessing the link between public capital and productivity. *American Economic Review* 89, 619–38.

Garcia-Milà, T. and McGuire, T.J. 1992. The contribution of publicly provided inputs to states’ economies. *Regional Science and Urban Economics* 22, 229–41.

Garcia-Milà, T., McGuire, T.J. and Porter, R.H. 1996. The effect of public capital in state-level production functions reconsidered. *Review of Economics and Statistics* 78, 177–80.

Haughwout, A.F. 2002. Public infrastructure investments, productivity, and welfare in fixed geographic areas. *Journal of Public Economics* 83, 405–25.

Holtz-Eakin, D. 1994. Public-sector capital and the productivity puzzle. *Review of Economics and Statistics* 76, 12–21.

Hulten, C.R. and Schwab, R.M. 1993. Infrastructure spending: where do we go from here? *National Tax Journal* 46, 261–73.

Lynde, C. and Richmond, J. 1992. The role of public capital in production. *Review of Economics and Statistics* 74, 37–44.

Mas, M., Maudos, J., Perex, F. and Uriel, E. 1996. Infrastructures and productivity in the Spanish regions. *Regional Studies* 30, 641–9.

Morrison, C.J. and Schwartz, A.E. 1996. State infrastructure and productive performance. *American Economic Review* 86, 1095–111.

Munnell, A.H. 1990a. Why has productivity growth declined? Productivity and public investment. *New England Economic Review*, January/February, 3–22.

Munnell, A.H. 1990b. How does public infrastructure affect regional economic performance? In *Is There a Shortfall in Public Capital Investment?* ed. A.H. Munnell. Boston: Federal Reserve Bank of Boston.

Nadiri, M.I. and Mamuneas, T.P. 1994. The effects of public infrastructure and R&D capital on the cost structure and performance of U.S. manufacturing industries. *Review of Economics and Statistics* 76, 22–37.

Rephann, T. and Isserman, A. 1994. New highways as economic development tools: an evaluation using quasi-experimental matching methods. *Regional Science and Urban Economics* 24, 723–51.

Tatom, J.A. 1991. Public capital and private sector performance. *St. Louis Federal Reserve Bank Review* 73, 3–15.

How to cite this article

Garcia-Milà, Teresa and Therese J. McGuire. "public infrastructure." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000364> doi:10.1057/9780230226203.1366

The New Palgrave Dictionary of Economics Online

public utility pricing and finance

Frank A. Wolak

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The theory of public utility pricing provides clear recommendations when the regulator and utility have same information about the underlying economic environment – the structure of demand and the production process. In reality, the utility has private information about the underlying economic environment, and the incentives created by the regulatory process can cause it to exploit this information by producing in an inefficient manner. This insight complicates virtually all aspects of the theory of public utility pricing, and has led to theoretical characterizations of the public utility price-setting process as the solution to a mechanism design problem.

Keywords

asymmetrical information; Averch–Johnson effect; cost functions; cost-of-service regulation; increasing returns to scale; inverse elasticity pricing rule; multi-part tariffs; multi-product firms; natural monopoly; optimal pricing; price cap regulation; principal and agent; privatization; public utility pricing and finance; Ramsey pricing; regulatory contracts; two-part tariffs

Article

Public utilities typically provide goods and services using a physical or virtual network infrastructure under a legal monopoly status. Public utilities can be privately owned, government-owned and customer-owned. Products provided by public utilities include electricity, natural gas, water, sewage treatment, waste disposal, public transport, telecommunications, cable television and postal delivery services. In the United States, all the different ownership forms can exist within the same industry. For example, in the electricity supply industry, there are privately owned, investor-owned and municipally owned utilities, and cooperative utilities owned by their customers.

Many explanations have been offered for the public utility industry structure. The standard economic efficiency argument is that the industry is natural monopoly, meaning that a single cost-minimizing firm

is the least-cost way to serve the current level of demand. However, this logic relies on the implicit assumption that the single firm will produce in a cost-minimizing manner, which is unlikely to occur under government ownership or government regulation, for the reasons discussed below. In addition, although the current level of demand may be served at least cost by one cost-minimizing firm, this is unlikely to be that case for all future levels of demand as the number of customers or their purchasing power grows. Recognizing that public safety and health concerns argue for universal access to many of these services and the fact that the demand is very inelastic with respect to its own price leads to political economy explanations for this public utility industry structure. As Waterson (1988) notes, a government-owned or -regulated monopoly may better ensure that all customers have access to these services at reasonable prices.

Over the 100 years or more of state and federal regulation of public utilities in the United States there has been debate over what constitutes a reasonable price for goods and services of public utilities. A price that recovers the firm's operating costs including return on its capital stock is generally considered to meet the legal standard of a reasonable price. This form of price regulation in the United States is often referred to as ‘cost-of-service’ regulation. However, as Joskow (1974) has persuasively demonstrated, the price-setting process for privately owned utilities in the United States does not guarantee the firm a fixed rate of return on its capital stock or full operating cost recovery. In that sense, to call this regulatory price-setting process ‘cost-of-service’ regulation is a misnomer. Joskow (1974, p. 325) states: ‘The rate of return aspect of regulation is merely a method by which a regulatory commission justifies its approval of price increases or major changes in rate structures. Without such triggering mechanisms the rate of return constraint is essentially inoperative.’ When the cost-of-service regulatory process operates it sets a price that allows the public utility an opportunity to recover its operating costs and the regulated rate of return on its capital stock through prudent operation.

If the firm earns a higher rate of return at this price because of superior management, then it is allowed to keep the revenue. If the firm earns a lower rate of return because of poor management, then shareholders must accept a lower rate of return. Only when the regulatory commission has overwhelming evidence that the higher or lower rate of return is due to extraordinary events beyond the control of the firm and not anticipated at the time the regulatory commission set the price will it make *ex post* adjustments to alter the public utility's regulated rate of return. A well-known example of an extraordinary event is a price change for fossil fuels used to produce electricity. The extreme volatility in oil, natural gas and coal prices since 1977 has led many regulatory commissions in the United States to implement fuel price adjustment clauses that automatically pass through in any input fuel price changes in the price of electricity. Baron and De Bondt (1979) discuss the impact of these fuel adjustment mechanisms on the investment and operating decisions of regulated electricity and natural gas utilities.

The terms and conditions surrounding this promise of full cost recovery through prudent operation is often referred to as the ‘regulatory contract’ between the regulatory commission and the public utility. This implicit contract requires the utility to serve all demand at the price set by the commission in exchange for a price that allows the utility the opportunity to recover its operating costs and a reasonable return to its capital stock. A major challenge to this regulatory contract is determining when imprudent operation is the cause of a failure to achieve full cost recovery.

For a number of reasons, unexpected events outside the control of the regulatory commission or utility and *ex post* opportunism by the regulator are often very difficult to distinguish from valid reasons for the regulatory commission to disallow price increases. Utilities typically require substantial investments in a

network infrastructure that has limited alternative uses. The future demand for the public utility's services is uncertain, so there is a substantial risk that investments in network infrastructure will not be needed to serve the demand that exists when the investment is completed.

Several aspects of the regulatory process in the United States are designed to address the problem of the regulator setting a price that is insufficient to provide a reasonable return on past investments. The concept of a ratebase and the requirement that the regulatory commission sets a price that recovers operating costs and a reasonable return on the entire ratebase limits opportunistic behaviour on the part of the regulatory commission. To a first approximation, the ratebase is the sum of all past investments judged as prudent and therefore worthy of cost recovery by the regulatory commission. Phillips (1993, ch. 8) provides a detailed discussion of this concept. The requirement that the entire ratebase earns the regulated rate of return ensures that the current regulatory commission compensates the utility for investments that previous commissions have deemed prudent.

Gilbert and Newbery (1994) construct a dynamic model of the regulatory price-setting process where the commitment to allow the firm to earn a reasonable rate of return on a ratebase composed of past prudent investments results in a socially efficient level of investment by the regulated firm. Lyon and Mayo (2005) investigate the empirical relevance of regulators' opportunistic behaviour by examining the investment behaviour of regulated electric utilities and the propensity of the relevant state regulatory commissions to disallow investments by these utilities from entering the ratebase. Lyon and Mayo (2005) find little evidence that these cost allowances by the state regulatory commissions were due to opportunistic behaviour, and instead argue they were motivated by a desire to punish poorly managed firms.

Optimal pricing of public utility services with full information

Prices that adhere to the implicit regulatory contract of allowing full cost recovery only impose one restriction on the set of possible prices. For the case of a single-product utility that must set the same price for all customers, this restriction implies that the regulated price is equal to average total cost. However, virtually no public utilities sell a single product or are required to set a single price for all customers, so that regulatory commissions are free to pursue additional goals, besides the promise of cost recovery, in setting regulated prices.

This section discusses methods for setting economically efficient prices – those that maximize some social welfare function – under the simplifying assumption that the utility and the regulatory commission have the same amount of information about the utility's production process and demand. The remainder of this section assumes symmetric information between the regulated utility and the regulatory commission, so the commission can credibly set a price that only recovers the firm's minimum cost of serving its demand. Although the assumption of symmetric information about the production process and nature of demand between the utility and regulatory commission is unrealistic, the literature on optimal pricing for public utilities described in this section relies on this assumption.

Two-part tariffs relax the assumption that a single uniform price is charged to all customers for each unit of output. If the production of the good or service is subject to increasing returns to scale, setting price equal to the marginal cost of the last unit sold violates the legal requirement that the firm has an opportunity to recover total production costs. Coase (1946) addresses this problem by considering a

regulated public utility producing a homogenous product with a monthly fixed cost of production, F, and a constant marginal cost, c. Coase (1946) argues that the total surplus maximizing two-part tariff sets the price of each unit consumed, p, equal to c and the fixed charge for each customer equal to F/N, where N is the number of customers served by the public utility.

If consumers differ in their willingness to pay for the product, then the surplus accruing to some consumers can be increased by the commission setting multi-part tariffs that charge different marginal prices for different ranges of monthly consumption. If the level of the monthly fixed charge necessary to recover total monthly fixed costs causes some consumers not to purchase the product, then a multi-part tariff can increase total consumer surplus. Assuming the marginal cost of a minute of telephone service is two cents per minute, setting a low monthly fixed charge and charging two cents per minute for the first 200•minutes of phone calls in the month and four cents per minute for all minutes above 200• minutes per month can allow the phone company to increase the number of consumers that benefit from having a telephone service without violating the promise of cost recovery. In this way, those consumers with the highest willingness will select through their consumption choice the higher marginal price, while those with the lowest willingness will select the lower marginal price, and virtually all consumers will pay a monthly fixed charge that does not cause them to disconnect from the telephone network. Brown and Sibley (1986, ch. 4) discuss consumer and producer welfare properties of multi-part tariffs.

The nature of the goods and services sold by public utilities often allows them to segment customers and to charge different prices for the same product. In addition, virtually all public utilities are multi-product firms, which implies that the regulatory process involves setting prices for all goods sold by the firm. Both of these circumstances provide opportunities for regulatory commissions to pursue objectives beyond the promise of cost recovery.

Consider the case of a homogenous product with increasing returns to scale in production that is sold to M different sets of consumers and a regulatory commission that can set a single price for each set of customers. Deriving the total surplus maximizing prices for all customer types subject to the constraint on cost recovery fits into the framework considered by Ramsey (1927). Let $CS_i(p_i)$ equal the consumer surplus accruing to consumers of type i and when they face price p_i, and $PS_i(p_i)$ equal the producer surplus from serving consumers of type i. Ramsey prices maximize the objective function $TS = \Sigma_{i=1}^{M} [CS_i(p_i) + PS_i(p_i)]$ subject to the cost recovery constraint, $F \leq \Sigma_{i=1}^{M} PS_i(p_i)$, where F is the firm's fixed cost. Let c equal the marginal cost of production and $\varepsilon_i(p_i)$ the own-price elasticity of the demand by customers of type i at price p_i. The solution to this constrained optimization problem yields the inverse elasticity pricing rule:

$$\frac{(p_i^* - c)}{p_i^*} = -\frac{k}{\varepsilon_i(p_i^*)}, \; i = 1, 2, \ldots, M$$

where k is some positive constant and the p_i^*, i=1,2,…,M, are called Ramsey prices. These prices raise the revenue necessary to achieve full cost recovery with the smallest total surplus loss. Those consumer

types with relatively more inelastic demands for the goods pay higher markups above marginal cost than other consumer types.

This same Ramsey-pricing logic can be applied to the case of multi-product public utilities. However, the simplicity of inverse elasticity pricing rule is complicated by the fact that customers can substitute across the products that the public utility offers. For example, in the pricing of postal delivery services, a business mailer has the option to use different US postal service products to communicate with its customers. To set Ramsey prices, the regulatory commission must know the multi-product cost function for postal products and the consumers' surplus associated with each postal product, which depends on the price charged for other postal products that the consumer can use as a substitute. As shown in Brown and Sibley (1986, ch. 3), both own-price and cross-price elasticities of demand now determine the total surplus maximizing markups that solve the Ramsey-pricing problem.

The properties of the regulatory pricing mechanisms described in this section rely on the assumption that the public utility's cost function is the minimum cost way to produce each vector of outputs. Specifically, let T equal the public utility's technology set, the pairs of vectors of input quantities, x, and output quantities, q, such that q can be produced using x. If w is the vector of input prices, then the minimum cost function is equal to

$$C(q) = \min_{x \geq 0} w'x \quad \text{subject to} \quad (x, q) \in T.$$

Although a price-taking profit-maximizing firm would like to produce along its minimum cost function, the structure of the regulatory process could cause a privately owned profit-maximizing regulated utility not to produce along its minimum cost function. Averch and Johnson (1962) present an example of a regulatory mechanism that causes a profit-maximizing firm not to produce in a least-cost manner. This work led to a massive theoretical and empirical literature exploring the distortions from minimum cost behaviour caused by regulatory price-setting processes.

State-owned utilities are likely to have even less incentive to produce in a least-cost manner. Besides the incentives provided by the regulatory price-setting process, earning revenues in excess of operating costs is just one of the many objectives that managers of state-owned companies must balance. As Waterson (1988, ch. 4) notes, state-owned utilities are often asked to pursue political or social goals that conflict with maximizing profits and therefore minimizing production costs.

Economists have begun to recognize the distinction between a public utility's observed cost function and its minimum cost function. The observed cost function, $CO(q)$, gives the firm's actual cost of producing output vector q given the incentives provided by the regulatory process. For example, in the case of a state-owned utility, political constraints could require a firm's management to hire a certain number of workers for each level of output despite the fact that it is technologically feasible to produce using fewer workers at each output level. In general, the value of the firm's observed cost function is greater than the value of its minimum cost function for the same level of output, and this difference can be substantial.

There is a vast empirical literature documenting violations of the assumption of cost-minimizing behaviour by public utilities. Christensen and Greene (1976) is a well-known example in the electricity

supply industry and Evans and Heckman (1984) is one for the telecommunications industry. There are few empirical studies of regulated utility behaviour where the assumption of cost-minimizing behaviour is not rejected.

Optimal pricing of public utility services with asymmetric information

Public utility regulation has long been recognized as an example of agency relationship where the regulator (the principal) attempts to provide incentives for the public utility (the agent) to serve its demand in a least-cost manner at a price that only recovers observed costs. This link between the utility's production costs and the price it is allowed to charge creates the opportunity for the public utility to exploit its superior information about the production process and the demand it faces. This recognition has led researchers and regulatory policymakers to consider ways to either break this link between the price charged and the firm's observed cost or to design incentive mechanisms that exploit the link between the regulated price and the firm's observed cost.

Price-cap regulation attempts to break this link by committing to change the price the utility is allowed to charge according to a formula that cannot be altered for a sustained period of time. For a profit-maximizing firm, the price-cap mechanism creates the same incentive to minimize cost as the assumption of price-taking behaviour. In the United Kingdom (UK) this form of regulation is also called RPI minus X price-cap regulation because the rate at which prices are allowed to change on an annual basis is equal to rate of change in the retail price index (RPI) minus an X-factor chosen by the regulatory commission. Armstrong, Cowan and Vickers (1994, ch. 3) describe the details of choosing an X-factor for a specific firm and extensions of this basic regulatory framework.

The major challenge associated with price-cap regulation is balancing the desire to commit to a pre-specified pattern for the X-factors for as long as possible against the fact that the longer duration of the commitment to this pattern of X-factors the greater the likelihood that the commitment will run afoul of the regulatory contract or political concerns. The experience of many public utilities in the UK privatized during the 1990s provides an instructive example of this phenomenon. According to Armstrong, Cowan and Vickers (1994), the regulator for the water industry and the regulator for the electricity distribution sector initially committed to a five-year initial duration for values of the X-factors. However, in both these industries the regulator was forced to abandon these commitments before the end of the five-year period because of what was argued to by the water utilities to be inadequate revenues and what was argued by many consumers as excessive revenues in the case of the electricity distribution sector. Wolak (1998) discusses practical problems associated with implementing price-cap mechanisms and why they often evolve into an extremely ineffective form of cost-of-service regulation.

The development of game theoretical models of private information environments has led economists to derive optimal prices in this ‘second-best’ world. Baron and Myerson (1982) derive the total surplus maximizing prices when the public utility has private information about its production process not observable by the regulatory commission. The Baron–Myerson model assumes that the commission offers the firm a menu of prices and fixed fees that depend on the public utility's report of its private information. This report determines what price and fixed fee the firm is able to charge to its customers and therefore the revenues it is allowed to earn.

Recognition that informational asymmetries between the regulatory commission and utility can lead to significant distortions from minimum cost production and regulated prices that recover revenues substantially higher than these minimum costs has led to an explosion of theoretical work on the design of optimal regulated prices in this private information environment. Laffont and Tirole (1993) provide a comprehensive presentation of this literature.

Wolak (1994) attempts to quantify the cost of this private information in the regulatory process. Using a sample of California water distribution utilities, he estimates a full or symmetric information solution to the utility and regulatory commission interaction and a private or asymmetric information model of this interaction based on the Baron–Myerson (1982) model. Wolak (1994) finds that the asymmetric information model of the regulatory process provides a superior fit to observed data relative to the symmetric information model in addition to computing various estimates of the cost of the informational asymmetry between the firm and regulatory commission.

Concluding comments

The explicit recognition of the impact of the private information possessed by the firm on the price set by the regulatory commission has increased the realism of the assumptions underlying models of the public utility price-setting process. Unfortunately, the form of the optimal price-setting mechanism derived from these models depends crucially on the source of informational asymmetries between the firm and the regulatory commission, as well as many other details of the economic environment. This implies that much more empirical work on actual regulatory price-setting processes is needed to implement these theoretical advances in actual regulatory practice.

See Also

- Averch–Johnson effect
- mechanism design
- principal and agent (i)
- principal and agent (ii)
- price discrimination (theory)
- Ramsey pricing

Bibliography

Armstrong, M., Cowan, S. and Vickers, J. 1994. *Regulatory Reform: Economic Analysis of the British Experience*. Cambridge, MA: MIT Press.

Averch, H. and Johnson, L. 1962. Behavior of the firm under regulatory constraint. *American Economic Review* 52, 1052–69.

Baron, D.P. and De Bondt, R.R. 1979. Fuel adjustment mechanisms and economic efficiency. *The*

Journal of Industrial Economics 27, 243–61.

Baron, D.P. and Myerson, R. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50, 911–30.

Brown, S.J. and Sibley, D.S. 1986. *The Theory of Public Utility Pricing*. Cambridge: Cambridge University Press.

Christensen, L. and Greene, W.H. 1976. Economies of scale in U.S. electric power generation. *Journal of Political Economy* 84, 655–76.

Coase, R.H. 1946. The marginal cost controversy. *Economia* 13, 169–89.

Evans, D. and Heckman, J. 1984. A test of subadditivity of the cost function with an application to the Bell system. *American Economic Review* 74, 615–23.

Gilbert, R.J. and Newbery, D.M. 1994. The dynamic efficiency of regulatory constitutions. *RAND Journal of Economics* 25, 538–54.

Joskow, P. 1973. Pricing decisions of regulated firms: a behavioral approach. *Bell Journal of Economics and Management Science* 4, 118–40.

Joskow, P. 1974. Inflation and environmental concern: structural change in the public utility price regulation. *Journal of Law and Economics* 17, 291–327.

Laffont, J.-J. and Tirole, J. 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

Lyon, T.P. and Mayo, J.W. 2005. Regulatory opportunism and investment behavior: evidence from the U.S. electric utility industry. *RAND Journal of Economics* 36, 628–44.

Phillips, C.F 1993. *The Regulation of Public Utilites*. Arlington, VA: Public Utilities Reports.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Waterson, M. 1988. *Regulation of the Firm and Natural Monopoly*. Cambridge, MA: Basil Blackwell.

Wolak, F.A. 1994. An econometric analysis of the asymmetric information regulator-utility interaction. *Annales d'Economie et de Statistique* 34, 13–69.

Wolak, F.A. 1998. Price cap regulation in newly privatized industries. *Oxford Energy Forum* 34, 12–14.

How to cite this article

Wolak, Frank A. "public utility pricing and finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000358> doi:10.1057/9780230226203.1367

The New Palgrave Dictionary of Economics Online

public works

P. Bridel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balanced budget multiplier; Beveridge, W. H.; crowding out; effective demand; general gluts; Hawtrey, R.G.; hoarding; Kahn, R. F.; Keynesianism; multiplier analysis; neoclassical synthesis; new classical macroeconomics; Pigou, A. C.; primary and secondary labour markets; propensity to consume; public works; Ricardo, D.; saving equals investment; Treasury View; underconsumptionism; unemployment; Webb, B. and S.

Article

Public works to relieve unemployment is an idea as old as the pyramids. From the cathedrals of the Middle Ages to the dams of the Tennessee Valley Authority, recent centuries abound with examples of such government-supported schemes. However, it is only fairly late in the history of economic theory that economists started devoting some analytical thought to this problem. In fact, it was around the turn of the 20th century, with the appearance on the Continent of the first systematic work on business cycles, that economists stopped looking at unemployment as a ‘pathological’ problem to be tackled primarily as one of charity and relief. Even down to the British 1909 *Report of the Poor Law Commission* a majority still considered unemployment as no challenge to economic theory: if properly applied, in the long run, the normal laws of value and distribution would see to a solution to this problem. A notable exception to this general neglect is, of course, the heated discussions that took place between early 19th-century classical economists during the so-called ‘general glut controversy’. In the course of that well-known debate on the self-adjusting capacities of the economic system, public works as a means to combat unemployment came many times to the forefront of the argument. Ricardo, James Mill and Say fairly easily won the day against public works. However, though dissenters like Malthus, Lauderdale, Bentham and Sismondi rejected Ricardo's doctrine on the stability of the economic system and Say's rigmarole about the equality of supply and demand, they involved themselves in a logically very damaging acceptance of the dominant Turgot–Smith saving-is-investment doctrine, together with the aggravating assumption that hoarding cannot take place. Trapped in such contradictory statements, these dissenting authors never managed to put their case convincingly either against Ricardo's self-

adjustment doctrine or, of course, in favour of contra-cyclical public works. Showing very little interest in ‘immediate and temporary effects [and fixing his] whole attention on the permanent state of things’ (1952, vol. 7, p. 120), Ricardo was left in a very strong theoretical position to dismiss arguments in favour of credit expansion and/or public works to reduce unemployment. Furthermore, his lack of a theory of output (that is, his full-capacity utilization assumption) allowed Ricardo to close in an even tighter way a line of argument that was to dominate economic theory for nearly a century, and economic policy for an even longer period. Anticipating nearly word for word Churchill's 1929 Budget Speech and what was to be known in the 1920s and 1930s as the ‘Treasury View’, in 1821 Ricardo already observed in the Commons:

> When [I] heard the honourable members talk of employing capital in the formation of roads and canals, they appeared to overlook the fact that the capital thus employed must be withdrawn from some other quarter. (1952, vol. 5, p. 32)

This ‘Ricardian view’ on public investment was to be perpetuated down the 19th century; after having been defended and illustrated by most classical economists, and notably by Mill, it even survived the marginalist revolution: Marshall never advanced much beyond it and Böhm-Bawerk subscribed substantially to its conclusions.

However, during the 1880s, Foxwell and Hobson were among the first to take up the challenge again. More preoccupied with the social effects of the ‘irregularity of employment’ linked with price fluctuations, than with strictly theoretical questions, they both called for abandoning the charity and relief approach and for turning the fight against unemployment into a major objective of economic policy. Foxwell concludes his analysis by favouring (unlike most leading economists) a slightly rising price level ‘for more regular employment’. Hobson, for his part, suggests solutions to prevent under-consumption crises at the root of unemployment by a redistribution of income to encourage ‘high consumption’. Despite a few attempts by local authorities to provide relief work to the unemployed, nothing systematic surfaced in Britain in the realm of economic policy despite growing concern about unemployment.

In the history of public works doctrine, the *Minority Report of the Poor Law Commission* (Royal Commission on the Poor Law and the Unemployed, 1909) clearly marks a watershed: for the first time ever the minority commissioners advocated a systematic counter-cyclical policy of public works and investment to smooth out cyclical fluctuations and to stabilize employment and the level of economic activity. However, this fundamental turnabout was not confined to Sidney and Beatrice Webb, largely responsible (with A.L. Bowley for the statistical material) for drafting the *Minority Report*. In the same year as the *Report*, Beveridge in *Unemployment, a Problem of Industry* broadly supported its main conclusion, and a few years later the Webbs in their volume *The Prevention of Destitution* (1911) offered a more elaborate version of their original argument.

However, the first modern refutation of the ‘Ricardian view’ by a professional economist is due to Pigou in his 1908 inaugural lecture as Marshall's successor at Cambridge. Worked out again in his *Wealth and Welfare* (1912) and *Unemployment* (1913), Pigou's argument became the standard pre-Keynesian case for public works: without having to resort to the notion of budget deficit, Pigou is able to demonstrate that public spending can increase aggregate employment and does not simply divert it from the private

to the public sector:

> … it is probable that only part of the extra taxes people pay would be taken from funds they would otherwise have devoted at that time directly or indirectly to wage payment. Hence, the true result of relief works and so on is not to leave the aggregate amount of unemployment in the country unaltered, but to diminish that amount. (1908, pp. 27–8)

In other words, and in the modern parlance of the balanced budget multiplier, the taxpayer's marginal propensity to consume is smaller than 1, while the government's marginal propensity to consume is unity: the net effect of such a tax increase is clearly expansionary. Pigou's failure to assess his argument quantitatively (in a multiplier-like fashion), his scepticism about the degree of labour mobility between the private and the public sectors and the resulting weak impact of his argument on policymakers do not detract however from his originality. Public works, even without any budget deficit, can lessen unemployment.

This became a standard argument for most economists, well before the First World War. Robertson gave his ‘cordial support’ to Pigou's analysis and, for the first time brought forward in his *Industrial Fluctuation* the symmetrical idea that, through public works, governments would ‘in times of depression [use] savings [that] are *not* otherwise so applied’ (1915, p. 253).

With the notable exception of Hawtrey in the 1920s, most British economists followed Pigou's lead even if they were bitterly arguing about the best way to close the gap between saving and investment. Hawtrey for his part never departed from his early critical position against the *Minority Report* outlined in *Good and Bad Trade* (1913, p. 260). He reiterated many times, most notably in his 1925 article, that the public works doctrine ‘overlook[s] the fact that the Government by the very fact of borrowing for its expenditure is withdrawing from the investment market savings which would otherwise be applied to the creation of capital’ (1925, p. 104). As an adviser to the Chancellor of the Exchequer, and despite the broad professional consensus in favour of public works, Hawtrey won the day in the British Treasury. Under successive Conservative and Labour Governments, the ‘Treasury View’ remained official wisdom for years (for a detailed discussion of that evolution see Hutchison, 1953, pp. 409–23 and Winch, 1969, pp. 94–113). Similarly, during the 1932 US presidential election Roosevelt kept attacking the outgoing Hoover administration for not balancing the budget. It is only with the rearmament in Germany and with the Second World War in other Western countries that the ‘Treasury View’ all but disappeared from the politicians’ standard set of arguments.

In his famous 1931 multiplier article, Kahn managed once and for all to dispose of this ‘Treasury View’. Kahn's intentions were, however, empirical, not theoretical. He set out not so much to point out the then generally accepted argument that an increase in government investment would generate ‘secondary employments’ but to provide ‘a stronger case’ for public works than that which ‘had always been recognized’ by giving a precise estimate of this multiplier effect. The ratio of secondary employment to primary employment was thus given its first formal expression (1931, p. 12)

However, and to dispel a very common confusion, with a multiplier equation strictly in the Pigovian tradition, Kahn did not anticipate the theoretical core of Keynes's *General Theory*, that is, the principle of effective demand. In Kahn's model, even if additional public spending has a multiplier effect on output and employment, such public works do not affect the discrepancy between saving and investment: as for Pigou and Robertson, previously, ‘unused’ savings are simply brought back into

circulation. In short, for Kahn ‘the whole point of a policy of public works is that it enables an increase in the rate of home investment … without that *fall* in the rate of interest that would be necessary if we were relying on private enterprise’ (1931, p. 26).

In the *General Theory*, with his principle of effective demand, Keynes added a new theoretical dimension to Kahn's multiplier approach to the public works doctrine. An increase in public investment not only leads to an increase in employment and output but *generates* an excess of income over that required for consumption (via a marginal propensity to consume smaller than 1) so that the volume of savings will increase until saving once again equals investment. Furthermore, it is clear that since saving and investment are brought into equality by variations in the level of income, and not by changes in interest rate, there need not be full employment. In Keynes's theoretical framework, public works and government investment are thus no longer seen as temporary remedies to cyclical fluctuations in private investment, but as a necessary component of an aggregate demand the deficiencies of which are no longer automatically cared for, even in the long run, by an interest-rate mechanism.

This proposition (and the income-adjustment principle underlying it) could of course find no place in the traditional analytical framework. However, systematic short-term counter-cyclical public works policies were soon to be integrated by mainstream economists within the so-called neoclassical synthesis and played in most countries a major role in post-war economic policies. Victim of its success during the 1950s and 1960s, this Keynesian (as opposed to Keynes’s) public works doctrine ran into serious practical and theoretical problems in the early 1970s. The growing size of public sectors, the inflationary wave resulting largely from an excessive use of expansionary fiscal policies and a rapidly growing theoretical dissatisfaction of the economic profession with the neoclassical synthesis brought back the ‘Treasury View’ onto the theoretical agenda. However even if with the successive advent of the monetarist school and the ‘governmental impotence’ theorem of the New Classical School, the ‘crowding out’ hypothesis is currently enjoying a new lease of life, few economists and even fewer politicians would dispute today the importance of modulating government spending in the course of the cycle.

See Also

- Hawtrey, Ralph George

Bibliography

Beveridge, W. 1909. *Unemployment: A Problem of Industry*. London: Longmans, Green.

Hawtrey, R.G. 1913. *Good and Bad Trade*. London: Constable.

Hawtrey, R.G. 1925. Public expenditure and the demand for labour. *Economica* 5(March), 38–48. Repr. in R.G. Hawtrey, *Trade and Credit*, London: Longmans, 1928.

Hutchison, T.W. 1953. *A Review of Economic Doctrines. 1870–1929*. Oxford: Clarendon Press.

Kahn, R.F. 1931. The relation of home investment to unemployment. *Economic Journal* 41(June), 173–98. Repr. in R.F. Kahn, *Selected Essays on Employment and Growth*, Cambridge: Cambridge University Press, 1972.

Pigou, A.C. 1908. *Economic Science in Relation to Practice*. Inaugural Lecture given at Cambridge. Cambridge: Cambridge University Press.

Pigou, A.C. 1912. *Wealth and Welfare*. London: Macmillan.

Pigou, A.C. 1913. *Unemployment*. London: Williams & Norgate.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*. 11 vols, ed. P. Sraffa. Cambridge: Cambridge University Press.

Robertson, D.H. 1915. *A Study of Industrial Fluctuation*. London: LSE Reprint, 1948 (with a new Introduction).

Royal Commission on the Poor Law and the Unemployed. 1909. *The Minority Report of the Poor Law Commission*. 2 vols. London: Fabian Society.

Webb, S. and B. 1911. *The Prevention of Destitution*. London: Longmans, Green.

Winch, D. 1969. *Economics and Policy. A Historical Study*. London: Hodder & Stoughton.

How to cite this article

Bridel, P. "public works." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000253> doi:10.1057/9780230226203.1368

The New Palgrave Dictionary of Economics Online

Pufendorf, Samuel von (1632 – 1694)

Arild Sæther

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Pufendorf studied in Leipzig and Jena. His first work, *Elementorum Jurisprudentiae Universalis* (1660), earned him a professorship at Heidelberg. In 1668 he moved to Lund. His works *De Naturae et Gentium* (1672) and *De Officio Hominis et Civis* (1673) were translated, spread all over Europe, and entered the curricula at most Protestant universities. Pufendorf's natural law writings include ethics, jurisprudence, government and political economy. A society in which individuals exchange to satisfy their needs brings with it growth, commerce, markets, prices and money. This theory laid the foundation for the progress of economics as a science.

Keywords

commercial society; demand and supply; externalities; Grotius, H.; Hobbes, T.; Hutcheson, F; international law; Locke, J; money; natural law; price; property; Pufendorf, S. von; self-interest; Smith, A; Snob effect; taxation; Veblen effect; voting

Article

Pufendorf was born in Dorfchemitz, Saxony, Germany on 8 January 1632. He matriculated at Leipzig University in 1650. First he studied theology, but found it dogmatic and turned to philosophy, philology and history. After two years he moved to Jena, concentrating on mathematics, the Cartesian demonstrative method and the natural law writings of Grotius and Hobbes.
After completing his Magister degree in 1658 he secured an engagement as a tutor in the family of the Swedish ambassador in Copenhagen. Shortly thereafter hostilities broke out between the two Nordic rivals. Disregarding diplomatic privileges, the Danes seized the Swedish retinue and accused Pufendorf of espionage. During eight months of harsh captivity, with no access to learned books, he reflected on his university studies and wrote down a system of jurisprudence. After his liberation, he journeyed with his pupils to the Netherlands, where his work was published in 1660 as *Elementorum Jurisprudentiae*

Universalis (Elements of Universal Jurisprudence). This work is considered the first useful textbook on natural law and it earned Pufendorf an enviable reputation and a professorship at the University of Heidelberg.

Here he published, anonymously, his historical and political work *De Statu Imperii Germanici* (On the Constitution of the German Empire). It contains a devastating criticism of the condition of public law within the Empire resulting from the Thirty Years' War, and suggested a path to its regeneration through a European commonwealth of sovereign states based on natural and international law. The imperial censor banned the book, but it was reprinted time and again, translated into several languages, and distributed across Europe. By 1710, some 300,000 copies had been printed in Germany alone. Pufendorf's reputation was now extended to non-academic circles. He achieved both fame and criticism.

In 1668 Pufendorf moved to the newly established university in Lund, Sweden. In 1672 he published his magnum opus *De Naturae et Gentium* (On the Law of Nature and Nations) in eight books, and the year after an abridged popularized version, *De Officio Hominis et Civis* (On the Duty of Man and Citizen), in two books.

In all, 44 editions of his major work have been published. It has been translated into English, French, German and Italian. His popularized version became, in modern parlance, an international bestseller. It has been translated into nine European languages, published in more than 150 editions in tens of thousands of copies. For more than 100 years they were among the most read academic books. The classicists Locke, Montesquieu, Rousseau, Hutcheson, Hume, Smith and countless more all studied and built on his works. Due to Pufendorf's works and reputation, natural law became part of university studies in jurisprudence, philosophy and ethics at most Protestant universities.

A new war resulted in the closing of Lund University in 1677. Pufendorf became royal historiographer in Stockholm. In the following years, he introduced empirical studies of the archives, and published 33 volumes of historical studies. He is regarded as a progenitor of 19th-century historicism.

In 1688 Pufendorf moved to Berlin as historiographer and judicial councillor at the court of Prussia. He continued his works on historical and theological issues. He died on 16 October 1694 of blood poisoning, contracted on a return journey from Stockholm, where he had been elevated to the nobility as a baron. He is buried in St. Nikolai Church in Berlin.

Pufendorf attempted in his works on natural law to mediate and unify Hobbes's natural law doctrine of 'egoism' and 'a war of all against all' with Grotius's natural law doctrine of 'man's inclination towards society'. His writings include ethics, jurisprudence, government, and political economics. These are seen as integral parts of a totality.

The foundation for his treatment is his theory of human behaviour, where the driving force is the interaction between man's self-interest and his existence as a social being. Man seeks society with his fellow man for the fulfilment of his own needs and desires. Man's sociable inclination is not innate; it must be cultivated. He also used his theory of the social man to create his historical account of the rise of property when society changes from hunting and gathering through agriculture to a commercial society. Individuals in a commercial society will need goods and services produced by others, because their own time and resources will fail to give them many necessary goods. On the other hand, individual men can contribute many things to the use of others. However, this will come to nothing if these individuals could not exchange and barter their different goods and services. When a society based on private property grew, it therefore brought with it commerce, the growth of markets, the creation of prices and the introduction of money. The theoretical foundation of a commercial society, in which all individuals

attempt to satisfy their own needs and thereby satisfy the need of others, is therefore the cornerstone in his natural law theory.

Price is divided into *ordinary* and *eminent*. The former is found in the properties of goods and services in so far as they afford service and pleasure for man. The latter is found in money as a common standard for their measurement. The price of a good or service is determined by the interaction between 'the aptitude' (utility) of it and the scarcity of it: in modern parlance, demand and supply. The price will rise towards a level where it covers the normal costs that accrue during production and transport. Lack of need (demand) lowers the price, but price will also be lowered if the number of suppliers increases. Pufendorf therefore comes very close to a Marshallian demand-and-supply analysis. In addition, the price will change if the quantity of money changes. Pufendorf seems to recognise the Snob and Veblen effects, externalities and differences in the elasticity of demand.

Pufendorf also presents his views concerning the state and the distribution of power, the state's right to tax, and the principles of taxation. Here he discusses weighted voting, qualified majorities and what has been known as single-peaked preferences.

It was the diffusion of these theories through popularization which laid the foundation for the progress of economics as a science.

See Also

- Hutcheson, Francis
- Smith, Adam

Selected works

1660. *Elementorum Jurisprudentiæ Universalis Libri Duo* [The Elements of Universal Jurisprudence], Oxford: Clarendon Press, 1931.

1672. *De Jure Naturae et Gentium Libri Octo* [On the Law of Nature and Nations], New York: Oceana Publications Inc.; London: Wiley & Sons Ltd., 1933; reprinted 1964.

1673. *De Officio Hominis et Civis*. Trans. M. Silverthorne as *On the Duty of Man and Citizen According to Natural Law*, ed. J. Tully. Cambridge: Cambridge University Press, 1991.

Bibliography

Gaertner, W. 2005. *De Jure Naturae et Gentium*: Samuel von Pufendorf's contribution to social choice theory and economics. *Social Choice and Welfare* 25, 231–41.

Hont, I. 1986. The language of sociability and commerce: Samuel Pufendorf and the theoretical foundations of the 'four-stages theory'. In *The Languages of Political Economy in Early Modern Europe*, ed. A. Pagden. Cambridge: Cambridge University Press.

Hutchison, T. 1988. *Before Adam Smith: The Emergence of Political Economy 1662–1776*. Oxford: Basil Blackwell.

Luig, K. 1972. *Zur Verbreitung des Naturrechts in Europa. Tijdschrift voor Rechtsgeschiedenis*, vol. 40, Groningen: Wolters-Noordhoff N.V.

Othmer, S.C. 1970. *Berlin und die Verbreitung des Naturrechts in Europa*. Berlin: Walter de Gruyter & Co.

Sæther, A. 1996. Pufendorf: The Grandfather of Modern Economics. In *Samuel Pufendorf und die europäische Frühaufklärung*, ed. F. Palladini and G. Hartung. Berlin: Akademie Verlag.

Sæther, A. 2000. Self-interest as an acceptable mode of human behaviour. In *The Canon in the History of Economics: Critical Essays*, ed. M. Psalidopoulos. London and New York: Routledge.

How to cite this article

Sæther, Arild. "Pufendorf, Samuel von (1632–1694)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000296> doi:10.1057/9780230226203.1369

The New Palgrave Dictionary of Economics Online

purchasing power parity

Lucio Sarno
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article expounds the purchasing power parity (PPP) hypothesis as a theory of exchange rate determination. The long history of PPP and its contribution to thinking in international finance is discussed, with reference to implications both for open economy theory and for economic policy. The large literature on empirical testing of the validity of PPP is reviewed, with particular reference to work carried out since 1990.

Keywords

aggregate demand; Balassa, B.; beggar-thy-neighbour; Bretton Woods system; Cassel, G.; co-integration; common factors; consumer price indices; crawling peg exchange-rate regime; de Malynes, G.; deflation; exchange rate depreciation; exchange rate dynamics; exchange rate regimes (empirics); external competitiveness; fixed exchange rates; flexible exchange rates; gold standard; gross domestic price deflators; Haberler, G.; Harrod, R.; hyperinflation; imperfect competition; indexation; inflation; inflationary expectations; information costs; international arbitrage; international capital flows; international finance; international portfolio diversification; international trade barriers; Johnson, H.; Keynes, J. M.; Kravis, I.; labour market contracts; law of one price; microfoundations; monetary policy; Monte Carlo methods; Mundell, R.; Mundell–Fleming model; nominal exchange rates; nominal interest rates; Nurkse, R.; open economy macroeconomics; price stability; productivity differentials; purchasing power disparities; purchasing power parity (PPP); quantity theory of money; rational expectations; real exchange rates; real interest rates; real price ratios; Ricardo, D.; Rueff, J.; Salamanca School; Samuelson, P.; spatial arbitrage; spatial price differentiation; sticky prices; Strong (absolute) purchasing power parity; terms of trade; trade barriers; unit roots; Viner, J.; weak (relative) purchasing power parity; white noise; wholesale price indices; Yeager, L.

Article

Purchasing power parity (PPP) is a theory of exchange rate determination. It asserts (in the most common form) that the exchange rate change between two currencies over any period of time is determined by the change in the two countries' relative price levels. Because the theory singles out price level changes as the overriding determinant of exchange rate movements, it has also been called the 'inflation theory of exchange rates'.

The PPP theory of exchange rates has somewhat the same status in the history of economic thought and in economic policy as the quantity theory of money: by different authors and at different points in time it has been considered an identity, a truism, an empirical regularity or a grossly misleading simplification. The theory remains controversial, as does the quantity theory of money, because strict versions are demonstrably wrong while soft versions deprive it of any useful content. In between there is room for theory and empirical evidence to specify the circumstances under which and the extent to which PPP provides a useful, though not exact, description of exchange rate behaviour.

The analogy with the quantity theory of money holds particularly in the effects of monetary disturbances. The latter theory fails to hold exactly when disturbances are primarily monetary, for instance in the course of hyperinflations, because changes in the expected rate of inflation generate systematic movements in velocity that break the one-to-one link between money and prices. In the same way, monetary disturbances cause exchange rate movements that at least temporarily deviate from PPP, implying changes in the exchange rate-adjusted relative price levels or 'real' exchange rates. It is true that when the economy, following a major monetary disturbance, has settled down again the cumulative changes in

money, prices and the exchange rate will tend to be close to each other. In that sense PPP holds. The same is decidedly not true, however, in the course of the disturbance.
And in the long run, just as changes in real income or financial innovation bring about trend changes in velocity that destroy the one-to-one relationship between the money supply and prices, there also are trend deviations from PPP: productivity growth differentials between countries, for example, lead to trend changes in real exchange rates.

Statement of the theory

Let p_i and p_i^* represent the price of the *i*th commodity at home and abroad, stated in home and foreign currency respectively, and *e* the exchange rate. The exchange rate is quoted as the number of units of domestic currency per unit of foreign money. Further, let *P* and P^* be the price level at home and abroad quoted in the respective currencies.
The strong or absolute version of PPP relies on the 'law of one price' in an integrated, competitive market. If we abstract from all frictions, the price of a given good will be the same in all locations when quoted in the same currency, say dollars: $p_1 = ep_i^*$. Consider now a domestic price index $P = f(p_1, \ldots, p_i, \ldots p_n)$ and a foreign price index $P^* = g(p_1^*, \ldots, p_i^*, \ldots, p_n^*)$. If the prices of each good, in dollars, are equalized across countries, and, if the same goods enter each country's market basket with the same weights (that is, the homogeneous-of-degree-one $g(\cdot)$ and $f(\cdot)$ functions are the same), then by definition *absolute* PPP prevails. The law of one price in this special case extends not only to individual goods but also to aggregate price levels. Spatial arbitrage then takes the form of the *strong* (or absolute) version of PPP:

$$e = P / P^* = \frac{\$ \text{ price of a standard market basket of goods}}{£ \text{ price of the same standard basket}} \tag{1}$$

where the RHS is the common multiple of the price of each good in one currency and in the other. Specifically, if $p_i / p_i^* = k$ for all *I*, we then have $e = P / P^* = k$. Note now the implication of absolute PPP. Whatever the monetary or real disturbances in the economy, because of instantaneous, costless arbitrage the prices of a common market basket of goods in the two countries, measured in a common currency, will be the same or $P / eP^* = 1$ at all times.
There can be no objection to (1) as a theoretical statement. Objections arise, however, when it is interpreted as an empirical proposition. In fact the (spot) prices of a given commodity will not necessarily be equal in different locations at a given time. Transport costs and other obstacles to trade, especially tariffs and quotas, do exist and hence the location of delivery does matter. Therefore we would not expect the price even of an ounce of gold of a specified fineness always to be the same in New York and in Calcutta. The fact that prices of the perfectly homogeneous commodity are not equalized across space at every point in time does not suggest market failure; it may simply reflect the inability to shift commodities costlessly and instantaneously from one location to the other. Information costs and impediments to trade stand in the way of strictest spatial equalization of price. But these impediments to trade do not preclude the possibility that common currency prices of any given good in different locations should be closely related and, indeed, arbitraged. They just will not be literally equalized. Impediments to trade and imperfection of competition, of course, also make possible spatial price differentiation, thus further limiting strong PPP.
The *weak* (or relative) version of PPP therefore restates the theory in terms of changes in *relative* price levels and the exchange rate: $e = \theta P / P^*$ where θ is a constant reflecting the given obstacles to trade. Given these obstacles an increase in the home price level relative to that abroad implies an equi-proportionate depreciation of the home currency:

$$\hat{e} = \hat{P} - \hat{P}^* \tag{2}$$

where $\hat{a}$ denotes a percentage change.
Equation (2) is the statement of PPP as it was applied by Gustav Cassel to an analysis of exchange rate changes during the First World War.

The general inflation which has taken place during the war has lowered this purchasing power in all countries, though in a different degree, and the rates of exchange

> should accordingly be expected to deviate from their old parities in proportion to the inflation of each country. At every moment the real parity is represented by this quotient between the purchasing power of the money in the one country and the other. I propose to call this parity '*purchasing power parity*'. As long as anything like free movement of merchandise and a somewhat comprehensive trade between the two countries takes place, the actual rate of exchange cannot deviate very much from this purchasing power parity. (Cassel, 1918, p. 413)

Absolute PPP in (1) was stated in terms of the relative prices in different currencies and locations of a *given* and common basket of identical goods. Going from there to relative PPP as in (2) may merely be a way of circumventing the qualifications arising from transport costs or obstacles to trade. But often more is involved because the shift, in practice, leads to a use of PPP in terms of particular price indices such as consumer price indices (CPIs), wholesale price indices (WPIs), or gross domestic price (GDP) deflators. Once that is done we go beyond the law of one price because the shares of various goods in the different national indices may not be the same and the goods that enter the respective indices may not be strictly identical as is clearly the case for non-tradable goods.

Once shares in the indices differ or commodities are not strictly identical, the appeal to the law of one price can no longer serve as support for PPP. Now PPP can hold, even in the weak form, only if the disturbances satisfy the conditions of the homogeneity postulate of monetary theory. The homogeneity postulate asserts that a purely monetary disturbance, with all equilibrium relative prices left unchanged, will lead to an equi-proportionate change in money and all prices, including the price of foreign exchange. In this very special experiment PPP holds even if the law of one price does not apply. The constancy of real variables under the assumption of a purely monetary disturbance (that is, an unanticipated, non-recurrent increase in money) assures that once the economy has adjusted the exchange rate depreciation matches the inflation of any individual price or the price of any market basket so that (2) applies. To appreciate the difference of this experiment from absolute PPP, note that under these conditions (2) could even be stated in terms of indices of non-tradable goods prices.

PPP theory as a theory of equilibrium must be supplemented by an adjustment mechanism. In the case of identical commodities the theory is simply that of spatial arbitrage. But when the goods are not strictly identical more is required. A high degree of substitution in world trade is generally assumed to be the mechanism through which exchange rate-adjusted prices are kept in line internationally. A further point concerns causation. In much of the literature, especially in the writings of Cassel, exchange rates adjust to prices. But there is an important alternative tradition that singles out exchange rate depreciation as an independent source of inflation.

Criticism of PPP focuses on systematic ways in which relative price changes destroy the strict validity of PPP. Keynes, although strongly supporting the idea of PPP as a broad guide, recognized these possible departures from purely monetary disturbances:

> If on the other hand these assumptions are not fulfilled and changes are taking place in the 'equation of exchange', as economists call it, between the services and products of one country and those of another, either on account of movements of capital, or reparation payments, or changes in the relative efficiency of labour, or changes in the urgency of the world's demand for that country's special products, or the like, then the equilibrium point between purchasing power parity and the rate of exchange may be modified permanently. (Keynes, 1923, p. 80)

This limitation of PPP led Samuelson to argue: 'Unless very sophisticated indeed, PPP is a misleading, pretentious doctrine, promising what is rare in economics, detailed numerical prediction' (1964, p. 153).

History

Versions of the PPP theory have been traced to the Salamanca School in 16th-century Spain and to the writings of Gerard de Malynes appearing in 1601 in England. The Swedish, French and English bullionists in the second part of the 18th century and in the early 19th century present further statements of PPP. Particularly noteworthy is the Bullion Report in England:

> Whether this $13\frac{1}{2}$ per cent, which stands against this country by the present exchange on Lisbon, is a real difference of exchange, occasioned by the course of trade and by the remittances to Portugal on account of government, or a nominal and apparent exchange occasioned by something in the state of our currency, or is partly real and partly nominal, may perhaps be determined by what your committee have yet to state. (Great Britain, 1810, p. ccxxii)

During the 19th century classical economists, including in particular Ricardo, Mill, Goschen and Marshall, endorsed and developed more or less qualified PPP views. This history is reviewed and discussed in Viner (1937), Schumpeter (1954), Holmes (1967) and Officer (1984).

Even though PPP theory was well established by the time of the First World War, the forceful use and development of the theory by the Swedish economist Gustav Cassel has made

him the outstanding protagonist of the theory. He turned the theory into a paradigm, with all the necessary trappings: an alleged challenge to gold standard orthodoxy, a catchy name, a formula, and the claim of empirical support for the new view.

Cassel's first contributions to the subject were published in 1916 in the *Economic Journal*. He presented the inflation theory of exchange rates and proceeded to a demonstration using price level and exchange rate data for the belligerent countries, the United States, and Sweden. J.M. Keynes as the editor appended a footnote drawing attention to the contribution and noting his surprise that, war disturbances notwithstanding, PPP should hold. A further challenge was the implication of PPP that the pre-war par with gold might not be re-established or, more guardedly, might require a powerful deflation in a country like Britain.

Cassel never abandoned an uncompromising PPP view of exchange rates even though he already in 1918 started recognizing the possibility that exchange rates might transitorily diverge from PPP. A decade later he made a clear statement of his final position:

> The fact that the rate of exchange corresponding to Purchasing Power Parity possesses such a remarkable stability is a sufficient reason for regarding Purchasing Power Parity as the fundamental factor determining the rate of exchange and for classifying all other factors that may influence the rate and perhaps make it deviate from the Purchasing Power Parity as factors of secondary importance, most suitably grouped under the head of ‘disturbances’. (Cassel, 1928a, p. 16)

He identified three groups of disturbances: actual and expected inflation or deflation, new hindrances to international trade, and shifts in international movements of capital. Although these disturbances are recognized, their quantitative effect on deviations from PPP is invariably seen as ‘confined within rather narrow limits’ (Cassel, 1928a, pp. 28–9). In insisting on the proposition that deviations from PPP are limited and transitory, Cassel neglected to pay close attention to the determinants of purchasing power *dis*parities. Even though he recognized that inflation first leads to undervaluation, and stabilization leads later to an overvaluation (Cassel, 1928b, p. 26), he never took these ideas further. His emphasis was on PPP. But he pointed out with some merit (Cassel, 1928b) that, without some quantifiable concept of PPP, a sensible discussion of over and undervaluation could hardly begin.

Keynes (1923, ch. 3) took up PPP, crediting Ricardo with the invention and Cassel with the name. Keynes recognized PPP as an important empirical possibility. Giving it all the right qualifications he still endorsed it for all practical purposes:

> This theory does not provide a simple or ready-made measure of the ‘true’ value of the exchanges. When it is restricted to foreign-trade goods, it is little better than a truism. When it is not so restricted, the conception of purchasing power parity becomes much more interesting, but is no longer an accurate forecaster of the course of the foreign exchanges. Thus defined ‘purchasing power parity’ deserves attention, even though it is not always an accurate forecaster of the foreign exchanges. The practical importance of our qualifications must not be exaggerated. (Keynes, 1923, pp. 77–8)

Cassel received support for PPP from the monetary disturbances of the 1913–28 period. Extensive PPP studies were conducted for the US government (see Young, 1925) and for the League of Nations. PPP emerged in the discussion of the resumption of the pre-war gold par in Britain in 1925, and Jacques Rueff used wage-based PPP to calculate an appropriate par for France's stabilization under Poincaré in 1926–8. But while it became a regular tool of applied macroeconomics, there was also plenty of controversy. Viner (1937) challenged the doctrinal view that classical economists had a concept of PPP, arguing that without the notion of a price level PPP could not be conceived. In fact Viner had little patience with PPP. The opposition is easily recognized today: Viner and other critics always reacted to the overstated claim that PPP must hold as a matter of fact or of theory, pointing out that only a purely monetary disturbance provided the theoretical or practical experiment in which PPP would apply. For them PPP as a theory was simply misstated and as a practical proposition overrated.

A new wave of interest in PPP emerged at the end of the Second World War, when once again exchange rates had to be set following the wartime suspension of trade and convertibility. Renewed interest in PPP followed in the late 1950s and early 1960s. Yeager (1958) and Haberler (1961) emphasized the practical usefulness of PPP and highlighted the role of high price elasticities in international trade as the factor supporting PPP. High elasticities in world trade would ensure that real disturbances had only small effects on relative prices, thus establishing more nearly the conditions under which exchange rate movements predominantly reflect differences in monetary experiences.

In the late 1930s Harrod had drawn attention to the fact that divergent international productivity levels could, via their effect on wages and home goods prices, lead to permanent deviations from Cassel's absolute version of PPP. This idea had already been developed by Ricardo and has now become central to work on international real income comparisons. Balassa (1964) and Samuelson (1964) elaborated similar ideas to argue that there are systematic trend deviations from PPP. This ‘productivity bias’ to PPP is discussed in more detail below.

PPP had yet another intellectual upturn with the move to flexible exchange rates in the early 1970s. The then fashionable ‘monetary approach to the balance of payments’ developed by Robert Mundell (1968; 1971), Harry Johnson and their students readily adapted to become a PPP-based monetary approach to the exchange rate (see Frenkel and Johnson, 1975; 1978; Mussa, 1979). The exchange rate under strict PPP conditions was interpreted as a monetary phenomenon. The absolute version of PPP in (1) above combined with the quantity theory of money for each country ($MV = PY$ and $M^*V^* = P^*Y^*$) yielded the key equation determining exchange rates by relative money supplies, velocities and real incomes:

$$e = (M / M^*)(V / V^*)(Y^* / Y). \tag{3}$$

Empirical research on the 1920s and on the very early data of the 1970s initially seemed to lend support to PPP and the monetary approach.
But large movements in real exchange rates of the 1970s led to the currently dominant PPP scepticism. The new direction following the Mundell–Fleming model (Mundell, 1968; Fleming, 1962) of the 1960s emphasized fluctuations in real exchange rates or the terms of trade (import prices relative to export prices) arising from the discrepancies between flexible, forward-looking asset markets and asset prices, and short-run sticky prices and wages. Work on exchange rate dynamics (Dornbusch, 1976) developed these ideas about transitory deviations from PPP in a rational expectations context.
Concern with PPP continued to be very active in the late 1970s and the early 1980s. The real exchange rates of the main currencies underwent large, persistent fluctuations with important effects on trade flows and resource allocation. At the same time currency experiments in Latin America involved dramatic real appreciations with ruinous consequences for several countries. Sometimes in history there was bafflement as to how, all things considered, PPP could work so closely. This time, however, the surprise was on the other side: how could real exchange rates get that far out of line? We now review in more detail the theory and the evidence for deviations from PPP.

Purchasing power disparities

Qualifications to PPP take one of several forms. Departures from PPP can be ‘structural’ in the sense that they arise systematically in response to new and lasting changes in equilibrium relative prices. Alternatively, they occur in a ‘transitory’ fashion as a result of disturbances to which the economy adjusts with differential speeds in goods and assets markets. These qualifications imply that even the weak or relative form of PPP cannot be expected to hold closely.
These disparities arise primarily for the following reasons. First, the terms of trade may change as a consequence of changes in trade patterns. Second, economic growth systematically affects the relative price of home and tradable goods. Third, monetary and exchange rate changes bring about transitory deviations in real price ratios and in PPP as a consequence of imperfectly flexible wages and prices.

Structural departures

The literature is replete with qualifications to PPP singling out particular real disturbances that change equilibrium relative prices. Thus it has been recognized since Ricardo that real prices of home goods are high ‘in countries where manufactures flourish’. It also has been argued that the ‘price level is high in borrowing countries’. The Ricardo–Harrod–Balassa–Samuelson theory provides a framework for these ideas.
Consider a Ricardian model where the law of one price applies to tradable goods and where there is also a home good. With perfect competition and constant returns, prices are given by unit labour costs. We define as R the relative consumer price levels of two countries measured in a common currency:

$$R \equiv P / eP^*. \tag{4a}$$

With identical homothetic tastes and the law of one price the international component of price indices is the same in both countries and hence cancels out in (4a). The relative price level is then determined by the relative prices of home goods in the two countries, measured in a common currency. Let h and h^* be the levels of productivity in tradable goods (at the competitive margin) relative to home goods in each country. It is easily shown (see Dornbusch, Fischer and Samuelson, 1977) that the relative price level then reduces to:

$$R = R(h / h^*),\ R' > 0. \tag{4b}$$

A uniform rise in tradable goods productivity at home would bring about a rise in the relative price level of the home country or a real appreciation. The mechanism is the following: with the law of one price applying to tradable goods, increased productivity in the tradable goods sector increases wages in that industry and hence raises economy-wide wages. But, without accompanying productivity gains in the home goods sector, costs and prices there must rise and hence the growing country's relative price level increases as shown by (4b).
In (4b) above the national productivity relatives h and h^* are measured in the tradable goods sector at the competitive margin. Shifts in technology, tastes, commercial policies or labour force growth will all change the equilibrium competitive margin and hence will change the real exchange rate. Thus real factors, as the literature since Ricardo has recognized, will introduce systematic departures from PPP. For example, a shift in world demand towards the home country's goods would raise the relative wage and reduce the range of goods produced by the home country. The rise in the relative wage, given productivity, raises the relative price level of the home country. Likewise an increase in spending relative to income (that is, borrowing or a current account deficit) will lead to a rise in the relative price level of the spending country.
A variant of the Ricardian productivity differential model as an explanation for the relatively low price of non-tradables in poor countries has been advanced by Kravis and Lipsey (1983) and Bhagwati (1984). They rely on differences in factor endowments and factor rewards rather than differences in production functions. In the poor labour-abundant country, the labour-using non-tradable services can be produced at a lower cost than in the rich, capital-abundant country. Whichever is the model, this effect, as we discuss below, has found ample support in empirical research on international real income and price comparisons.

Transitory deviations

There is no difficulty in accepting that prices of close substitutes or even identical goods could diverge across space at any point in time. This would be the case because, in the shortest time period, transportation and information costs make arbitrage difficult or even impossible. These difficulties would explain that PPP holds up to a constant and white noise error (see Aizenman, 1986). But in fact we have to explain relatively *persistent* and often *large* deviations from PPP. These can arise from divergent speeds of adjustment of the exchange rate compared with wages and prices. Particularly when flexible exchange rates behave like asset prices while wages are determined by long-term contracts, there is room for relative prices to show relatively persistent deviations from PPP.
Theoretical approaches to support the relative stickiness of prices can rely on the presence of long-term labour contracts combined with oligopolistic pricing in goods markets. A model of imperfect competition is essential because the less-than-perfect degree of substitution is a key ingredient in PPP deviations. Less-than-perfect substitution means that we are not dealing with the law of one price and arbitrage but with firms' decisions to set relative prices. A suggestive framework is the Dixit–Stiglitz (1977) model of product diversification with imperfect competition. Given constant returns and labour as the only factor each firm will set prices as a fixed and common markup over wages. In the world market for the products of a particular industry the relative price of domestic and foreign variants of the product is determined by relative unit labour costs measured in a common currency:

$$p / ep^* = w / w^* e \tag{5}$$

where w and w^* denote unit labour costs at home and abroad in the respective currencies. Given sluggish wages, for contract reasons or otherwise, exchange rate movements will be one-for-one reflected in changes in the real exchange rate.
The assumption that firms base their pricing entirely on home cost, as appears in this model, leaves no room for the alternative of spatial price differentiation. There is as yet no definitive or even large body of literature that develops the industrial organization aspects of pricing under flexible and volatile exchange rates (see Dornbusch, 1987).

Early empirical evidence

There is little doubt that the prices of primary commodities traded on major organized exchanges in different locations are fully arbitraged when literally all adjustments for contracts (maturity, delivery terms and location, and so on) are made. But much available evidence suggests that PPP in the strong or weak version does not apply in the same fashion to manufactured goods. This lack of close conformity with PPP is as true for individual commodity prices as it is for aggregate price indices. Moreover, the absence of a very tight PPP relation appears to hold especially during major monetary dislocations.

Studies of high-inflation episodes appear to offer support for PPP in that they show close *cumulative* movements of internal prices and the exchange rate. But even here the evidence is deceptive, as becomes clear when one looks at relative prices, which do show large variations. Indeed, particularly during high inflation the differing frequencies of adjustments of wages, prices and the exchange rate introduce considerable variability in relative prices which disappears only in the most intense stages of hyperinflation where all pricing comes to be based on the exchange rate. Kravis and Lipsey (1978) and Isard (1977) have shown tests of the law of one price at the level of narrowly defined manufactured goods. These studies established that for the same good (or highly substitutable goods) there are quite definitely persistent price discrepancies between domestic and export prices, between domestic and import prices, and between export prices to different markets.

Empirical studies on time series PPP relationships for aggregate price indices since the mid-1960s also show evidence of persistent deviations. Once relative prices are not strictly constant PPP will perform differently depending on the particular price index chosen for comparison. Commonly the choice is among CPIs, WPIs, and GDP deflators. WPIs are often ruled out on the argument that conceptually they are poorly defined, being neither producer nor consumer price indices. The preference is most often given to CPIs and GDP deflators, which have a clear methodological definition.

As a measure of the departure from PPP, research shows that, for the post-Bretton Woods period since 1971 or so, bilateral comparisons of exchange-rate-adjusted inflation rates (that is, comparisons of inflation rates measured in a common currency) reveal that the correlation coefficients are much lower than unity, which is the theoretical value implied by the weak form of the PPP hypothesis.

The strong deviations from PPP can likewise be found by looking at relative prices, in which case one would compare the variability of relative price indices (the standard deviation expressed as a fraction of the mean), measured in a common currency and using the United States as the numeraire country. The data for these relative price variability measures show a large increase in variability in the shift from fixed (Bretton Woods period) to flexible (post-Bretton Woods) exchange rates, suggesting that real exchange rates are approximately as volatile as nominal exchange rates (Baxter and Stockman, 1989).

The evidence on deviations from PPP leaves little doubt that they have been large and persistent. To pin down the major sources of these movements, however, is significantly more difficult. Among the chief explanations are capital flows induced by internationally divergent monetary–fiscal mixes interacting with sluggish wages and prices. Thus it would appear that a country that shifts in the direction of tight money and easy fiscal policy, for example, will experience real appreciation.

Besides these dominant macro-shocks there is, of course, a host of other factors. Jacob Frenkel has observed in this context:

> The experience during the 1970s illustrates the extent to which real shocks (oil embargo, supply shocks, commodity booms and shortages, shifts in the demand for money, differential productivity growth) result in systematic deviations from PPP … It should be noted, however, that to some extent the overall poor performance of the purchasing power parities doctrine is specific to the 1970s. During the floating rate period of the 1920s, the doctrine seems to have been much more reliable. (Frenkel, 1981, pp. 694–5)

The lack of solid empirical evidence in support of PPP extends to the assumption that divergent price developments ‘cause’ exchange rate depreciation. From the study of experiences of high inflation it is clear that in some instances capital flight and exchange depreciation precipitated increases in inflation. In fact Nurkse (1944) makes much of the point that expectations acting via capital flight on the exchange rate, and not actual money and prices, often initiate an inflationary episode.

With respect to structural PPP deviations, there is some evidence that establishes that over time real exchange rates, rather than showing constancy or a tendency to fluctuate around a constant level, in fact exhibit a distinct trend. Productivity levels or real incomes influence systematically the relative prices of tradable and non-tradable goods within a country and hence international relative price levels across countries and across time.

In the context of an international income comparison project, Kravis and associates have constructed indices of relative national price levels using an absolute price comparison approach. Drawing on a detailed sample of prices they construct matched sets of the price of individual commodity groups in a particular country relative to a reference country. For commodity *i* the relative price is p_i / p_i^*, where the *p*'s are measured in the respective countries’ currencies with an asterisk denoting the reference country. Using an arithmetic average with weights a_i, given by final expenditure shares, a PPP index is defined:

$$\mathrm{PPP} \equiv \sum a_i (p_i / p_i^*). \tag{6}$$

The expenditure shares a_i used in the weighting may be those of either one of the countries or some other appropriate weighting scheme. The Kravis real price level of a country (relative to the reference country) is defined as the PPP index in (6) divided by the actual exchange rate:

$$\text{Kravis real price level} \equiv \text{PPP} / e. \tag{7}$$

This real price level definition represents a measure of the deviation from the law of one price at the aggregate level.
Kravis and Lipsey (1983, p. 21) report the results of a cross-section study of 34 countries where the 1975 real price level as defined in (7) of the sample of countries (relative to the United States) is explained by the countries real income compared with that of the United States. The evidence shows that the higher a country's relative income is, the higher is its relative price level. Work by Hsieh (1982) using a time series approach further supports the extensive evidence on divergent productivity trends as a source of structural PPP deviations. It must be noted, though, that the evidence on structural deviations continues to be challenged by Officer (1984).

Implications of purchasing power disparities

The possibility that exchange rate movements do not conform to tight PPP patterns poses important issues for macroeconomic measurement, linkages, and policy. We review here several implications.

Real income comparisons

With strict PPP based on the law of one price, the purchasing power of a given income in one country and currency can be compared with the purchasing power of the income of any other country by simply measuring incomes in a common currency. If one income is 20 times larger than the other, measured in the same currency at actual exchange rates, then its command over goods and services is 20 times larger. But the fact that PPP does not hold leads to systematic biases in the comparisons. Specifically, as the work of Kravis and associates (1978; 1982; 1983) has shown, the real income of poor countries is severely underestimated when actual exchange rates are used to make the comparison. The low relative price of non-tradables in poor countries (due to the productivity differential discussed earlier) yields for poor countries true purchasing power of income significantly above what exchange rate-converted income suggests.
Note that the biases are particularly large for countries whose incomes are only a small fraction of the US levels, so that productivity differential effects play a maximal role. The poorer a country, the lower is the real price level. An interesting point is that these real price level differences apply both to goods and to services. One reason they also apply to goods is that these always have a local retail component which, on account of labour costs (though perhaps not transport), will tend to be low in poor countries. For low-income countries actual real income is two to three times what exchange rate-converted incomes suggest. These structural deviations from PPP, of course, would be invariant under a purely monetary disturbance so that the weak form of PPP still applies.

Interest rate linkages and PPP

Under perfect international mobility of capital and risk-neutral speculation there is a linkage between nominal interest rates and the anticipated rate of depreciation, which is given by the open economy Fisher equation:

$$i = i^* + x \tag{8a}$$

where i and i^* are the nominal interest rates at home and abroad and x is the anticipated rate of depreciation of the home currency. Adding and subtracting anticipated inflation rates on both sides yields an equation in terms of inflation-adjusted or real interest rates:

$$r^* = r - \dot{R} / R. \tag{8b}$$

Real interest parity, according to (8b), prevails when the real interest differential equals the expected rate of real appreciation, $\dot{R} / R$. From the real interest parity condition it is clear that under exact PPP the real exchange rate is constant. In the absence of restrictions on capital flows, real interest rates must therefore be strictly equalized across countries.
The real interest parity equation has two interesting implications. A first one is the linkages between the level of real exchange rates and monetary policy. Suppose that in a medium-term macroeconomic context, following a disturbance, the actual real exchange rate adjusts only gradually to the trend level R' according to the process: $\dot{R} / R = (1 / \lambda)(R' - R)$. Here $1/\lambda$ is the speed of adjustment, which depends among other things on the extent to which wages and prices are sticky. Combining this process with (8b) yields an equation for the equilibrium real exchange rate:

$$R = R' + \lambda(r - r^*). \tag{9}$$

The result shown here is that, when real interest rates at home exceed those abroad, the real exchange rate will be low or appreciated relative to its trend value. A tightening of monetary policy, by raising real interest rates, would thus bring about a (transitory) real appreciation. Equation (9) emerges from the dynamic Mundell–Fleming models and is often thought to explain real exchange rate movements and their tendency to return only gradually to their long-run value.
A second way to look at (8b) draws on the fact that the tradable–non–tradable goods distinction has implications for real exchange rates. Suppose the law of one price holds for tradable goods and that shares in the two countries' price indices are the same. Then, as argued before, the real exchange rate is equal to the relative price of non-tradable goods (in a common currency) in the two countries. Structural disturbances such as differential productivity growth or changes in aggregate demand will now have a systematic impact on relative non-tradable goods prices and hence real interest-rate differentials. Specifically, the country with the higher growth rate of productivity has a rising relative price of home goods and thus has a lower real rate of interest. As another example consider a country where aggregate demand is transitorily high. The real price of home goods will be high, but falling. Accordingly, the real interest rate will be higher than that abroad. Deviations from PPP, trend or short-run, thus introduce an equilibrium international interest-rate differential.
PPP deviations affect interest differentials another way. In (8a) above we assumed risk neutrality. But, once risk-averse speculators are admitted, the possibility that exchange rate movements could deviate from a strict PPP pattern introduces portfolio risk associated with the currency composition of the portfolio. PPP deviations are thus one basic motive for international portfolio diversification. A risk premium will appear and among the determinants of this premium is the variability of the real exchange rate. The risk premium will be an increasing function of real exchange rate uncertainty.

Exchange rate policy

In Cassel's view even small deviations from PPP would bring about large changes in trade flows and hence a rapid discipline to move prices back into line inter-nationally. But the reversion towards PPP has often not been quick and deviations from PPP have taken more nearly the pattern of persistent swings in a country's external competitiveness. The changes in competitiveness in turn have implied large swings in external balances, in output and in employment in the tradable goods sector. Changes in exchange rates that deviate from PPP at the same time influence the path of a country's inflation: real depreciation increases inflation and real appreciation dampens inflation. These effects of purchasing power disparities make the exchange rate an important issue in macroeconomic policy.
Countries with high inflation cannot afford a fixed exchange rate since the loss in external competitiveness would soon lead to excessive and growing external deficits and high

unemployment. If freely fluctuating rates are deemed too unstable the policy answer is often a crawling peg. In a crawling peg regime the rate of depreciation follows a PPP path such that over time the real exchange rate remains constant (see Williamson, 1965; 1982). Such a policy is an important advance over a system of occasional devaluations (too little, too late), but it is not without risks, for two reasons. First, freezing the real exchange rate may be a bad policy when disturbances in fact call for a path of, say, real depreciation. Second, there is a trade-off between stability of the real exchange rate and price stability. A policy of fully accommodating any and all price or cost disturbances by an offsetting depreciation may in fact remove price stability altogether (see Dornbusch, 1982).

PPP issues enter exchange rate policy also when a country seeks to gain macroeconomic advantages by a deliberate policy of driving the exchange rate away from PPP. A real depreciation serves to gain competitiveness and shift employment toward the depreciating country. In the 1930s this was called a ‘beggar-thy-neighbour’ policy, and in post-Second World War Europe it became ‘export-led growth’. A policy of real appreciation, by contrast, serves to reduce inflationary pressure as the rate of increase of tradable goods prices is pushed below the prevailing rate of inflation. These macroeconomic effects of purchasing power disparities are not difficult to bring about: easy money, in the short and medium term, serves to depreciate the exchange rate and thus create employment. This policy is more effective and more lasting the more sticky wages are and the smaller the connection between wages, prices and the exchange rate is. By contrast, in an economy that is strongly indexed and in particular with exchange rate influences on indexation, an attempt at creating employment via easy money would be frustrated as exchange depreciation precipitates offsetting wage and price inflation.

Deviations from PPP have also been used as a disinflation policy. Deliberate fixing of the exchange rate or pre-announced rates of depreciation below the prevailing rates of inflation have been adopted in various countries to break inflation. The experience has been almost uniformly disappointing and worse. The resulting overvaluation very often has led to excessive external deficits, borrowing and capital flight, and ultimately only moderate success at disinflation. The cases of Chile and Argentina in the late 1970s were particularly extreme. Exchange rate policies led to extreme overvaluation. But these economies had been opened to unrestricted trade or free capital flows. The public therefore could speculate against the overvalued currency by massive imports or capital flight while the governments financed the resulting deficits by external borrowing. In the end the scheme collapsed, leaving the private sector with foreign goods or foreign assets and the governments with huge foreign debts.

PPP disparities are relevant for the exchange rate choice between flexible and fixed or managed rates. In a world where exchange rate movements conform strictly to PPP and monetary policy governs prices there is no issue. Flexible rates then allow a country to choose its preferred rate of inflation. But once disparities are possible both as a result of structural trends and perhaps as a consequence of short-term capital movements the fixed versus flexible rate choice becomes more difficult. Flexible rates are preferable because there is no risk that the government pegs a rate that no longer corresponds to equilibrium. But flexible rates suffer the handicap that disequilibrating capital flows can drive the real exchange rate away from the level warranted by the fundamentals of the goods market. In particular, if exchange rates respond more to asset markets than price levels, persistent real appreciation or depreciation become a possibility. When this occurs there is invariably a call for PPP-based foreign exchange market intervention to bring rates back to ‘fundamentals’. Explicit target zones have been proposed as a means of maintaining the advantages of flexible rates within limits to maintain approximate PPP (see Williamson, 1983).

Flexible rates are also a concern because disequilibrating capital flows can provoke large changes in the rate of inflation. A loss of confidence, whether warranted or not, induces a capital outflow and a real exchange rate depreciation, as the experience of many East Asian countries in the late 1990s has demonstrated. If domestic financial policies are linked via the budget or indexation to the exchange rate, the real depreciation can initiate a sharp increase in inflation. Much of the discussion of the merits of flexible rates has concentrated on the question of whether speculative capital flows ‘cause’ the inflation or whether they merely respond to an inflationary situation, bringing about exchange depreciation in line with prevailing inflation. The Graham–Nurkse–Robinson view asserts, contrary to Milton Friedman, that destabilizing capital flows are the central element in the outbreak of major inflation experiences. Exchange stabilization, similarly, is seen as an essential step in stopping a runaway inflation and initiating a stabilization programme.

PPP is also relevant in the context of devaluation of a fixed rate. In the monetary approach to the balance of payments a firm tenet is the proposition that a devaluation cannot exert a lasting effect on relative prices or the balance of trade. Exchange rate depreciation raises the prices of all tradable goods in the same proportion and any effect then must be limited to a temporary depression of home goods prices due to reduced absorption. As money responds to the external surplus, real absorption rises and the initial real equilibrium is restored. This approach has the disturbing implication that devaluation does not appear to be an effective means of coping with trade or employment problems. In practice devaluation will work well when it is designed to speed up the adjustment from an initial disequilibrium in a situation where wages and prices are less than fully flexible downward. But a devaluation is likely to be ineffective if it is accompanied by a monetary expansion and wage increases, thus eliminating any real effects.

More recent developments

During the 1990s PPP attracted an enormous amount of interest. Presumably driven by the disbelief that such an intuitively appealing proposition about exchange rate behaviour had found little support in the data, researchers embarked in a ‘search’ for PPP using increasingly sophisticated time series methods. The early 1990s saw a proliferation of studies testing for PPP over the long run either by testing whether nominal exchange rates and relative prices move together (co-integrate) or by testing whether the real exchange rate has a tendency to revert to a stable equilibrium level over time (is stationary). The latter approach is motivated by the fact that the real exchange rate may be defined as the nominal exchange rate adjusted for relative national price levels and is, therefore, a measure of the deviation from PPP.

Regardless of the great interest in this area of research, the validity of long-run PPP and the properties of PPP deviations have remained the subject of an ongoing controversy. Unit root and co-integration studies generally report the absence of significant mean reversion of the real exchange rate for the recent floating experience (recent surveys include Taylor and Taylor, 2004; Sarno, 2005). However, the literature has been able to identify an important pitfall in studies of the long-run behaviour of PPP deviations. Specifically, one well-documented explanation for the inability to find clear-cut evidence of PPP is the low power of conventional tests to reject a false hypothesis of a unit root in (non-stationarity of) the real exchange rate (that is, the hypothesis that PPP is invalid) with a sample span corresponding to the length of the recent float (Frankel, 1990). Put simply, conventional time series methods would not be able to detect the reversion of exchange rates towards PPP even if PPP were indeed valid unless very long samples of data were made available.

Researchers have sought to overcome the ‘power’ problem in testing for PPP in various ways. One logical reaction to tackle this problem was to test for mean reversion in the real exchange rate using long spans of data. Lothian and Taylor (1996), for example, use two centuries of data on dollar–sterling and franc–sterling real exchange rates and provide evidence supporting PPP in the recent floating period. This evidence is ‘indirect’ in the sense that PPP was found to hold over the full sample, which includes the recent float. Lothian and Taylor could not find any significant evidence of a structural break between the pre- and post-Bretton Woods period, and argue that the widespread failure to detect mean reversion in real exchange rates during the recent float may simply be due to the shortness of the sample.

Long-span studies have, however, been subject to some criticism in the literature. One criticism relates to the fact that, because of the very long data spans involved, various exchange rate regimes are typically spanned. Also, real shocks may have generated structural breaks or shifts in the equilibrium real exchange rate. This is a necessary evil with long-span studies of which researchers are generally aware. The long samples required to generate a reasonable level of power with standard univariate unit root tests may be unavailable for many currencies (perhaps thereby generating a ‘survivorship bias’ in tests on the available data) and, in any case, may potentially be inappropriate because of differences in real exchange rate behaviour both across different historical periods and across different nominal exchange rate regimes (for example, Baxter and Stockman, 1989).

In light of the evidence provided by this literature, there remain several unresolved puzzles, among which two are prominent. First, it is still controversial whether long-run PPP is valid during the recent floating exchange rate regime from 1973 or so. Second, it is puzzling why the majority of studies which favour long-run PPP, such as the long-span studies, find empirical estimates of the persistence of PPP deviations that are too high – the half-life of shocks, that is, the time it takes for a shock to the real exchange rate to dissipate by one half, ranges between three and five years – to be explained in light of conventional nominal rigidities and to be reconciled with the large short-term volatility of real exchange rates (Rogoff, 1996).

A source of potentially important bias in estimates of the half-life is caused by cross-sectional aggregation in moving from the law of one price for individual goods to PPP deviations based on price indices. Imbs et al. (2005) demonstrate how such bias is bound to be present in estimates of the real exchange half-life and provide empirical evidence that the bias is substantial. Crucini, Telmer and Zachariadis (2005) adopt a similar approach to understanding the behaviour of deviations from the law of one price and PPP by examining micro-data on absolute prices of goods. They study good-by-good deviations from the law of one price for over 5,000 goods and services between European Union countries for the years 1975, 1980, 1985 and 1990, and report that between most countries there are roughly as many overpriced goods as there are underpriced goods so that PPP holds to a good approximation, particularly after wealth differences are controlled for.

It is instructive to graph the real exchange rate and its components over a long span of time to speculate on its low-frequency properties. The top panel of Figure 1 plots the time series for prices in the UK and the United States as well as the nominal dollar–sterling exchange rate over the sample period 1791–2000, with all time series expressed in logarithms.

Figure 1

200 years of prices and exchange rates. *Sources*: 1791–1991: Lothian and Taylor (1996); 1992–2000: International Financial Statistics database of the International Monetary Fund.

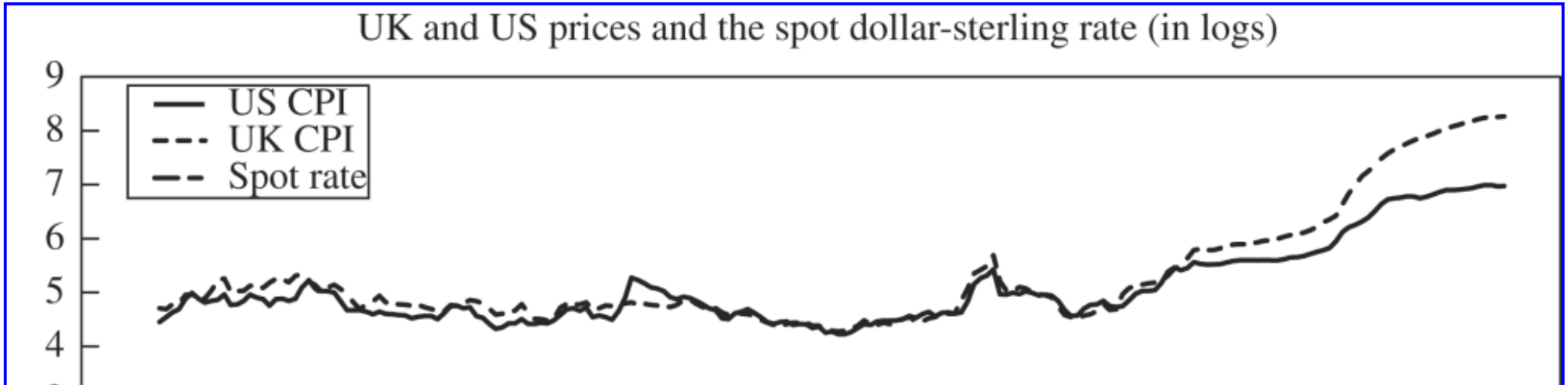

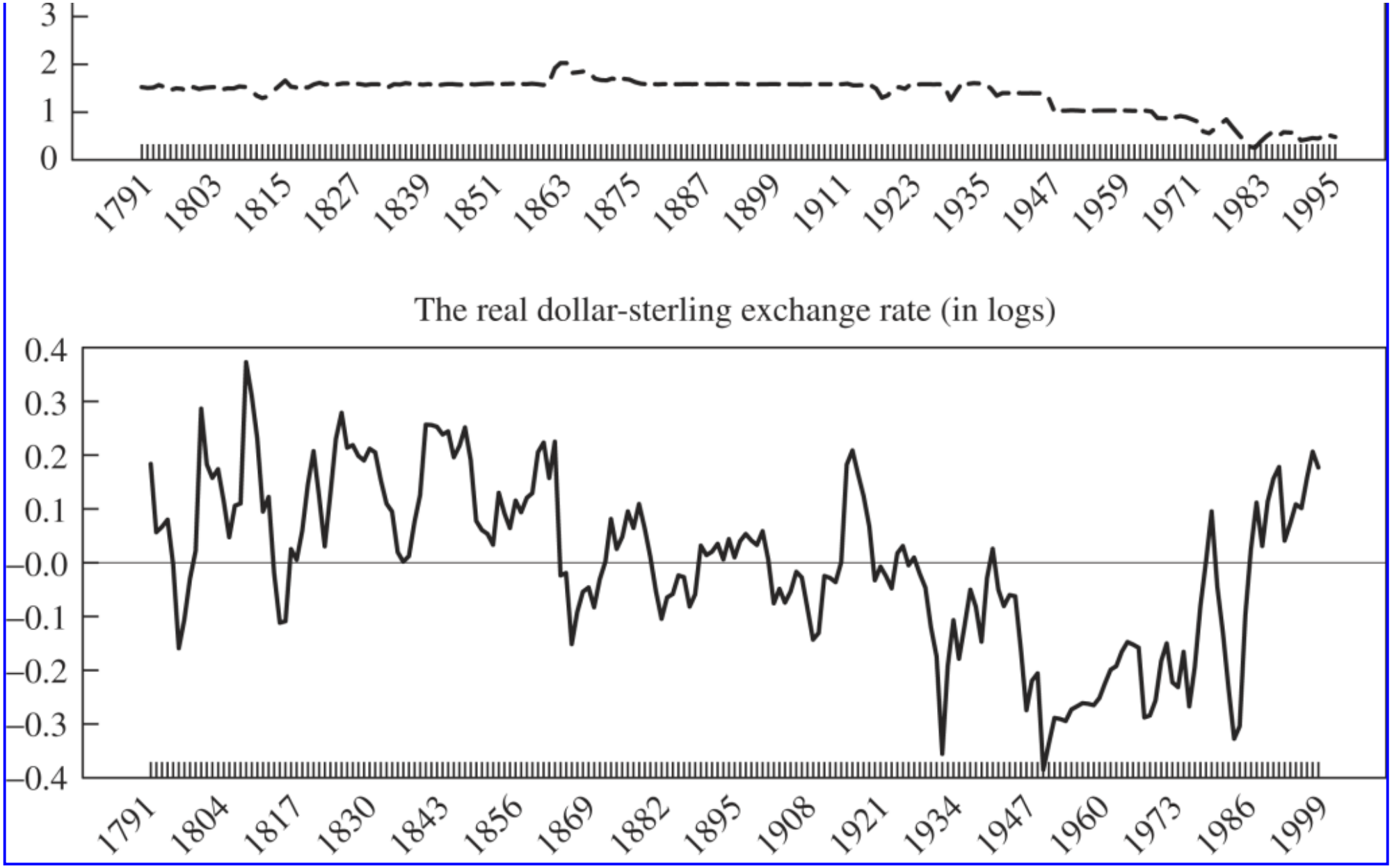

It is quite interesting how the price series move together – even without being adjusted by the exchange rate to express prices in a common currency – over such a long period. It is also apparent how the bigger and more persistent wedge between the two prices seems to occur in the post-Bretton Woods period, essentially from the 1970s onwards. This wedge also coincided with the beginning of a corresponding trend in the nominal exchange rate, exactly as one would expect under PPP. The bottom panel of Figure 1 then graphs the real exchange rate constructed from these time series (in deviation from the mean). The real exchange rate appears to have a tendency to return to its long-run mean (consistent with the PPP hypothesis), although the mean is crossed only 20 times in more than 200 years of data, indicating a remarkable degree of persistence in departures from PPP. Furthermore, the real exchange rate appears to be more persistent when it is in the proximity of the long-run mean, whereas reversion towards the mean happens more rapidly when the absolute size of the PPP deviation is large. This eyeball analysis of 200 years of real dollar–sterling exchange rate therefore suggests that this real exchange rate may be stationary, albeit persistent, and that it is very persistent in the neighbourhood of PPP, while being mean-reverting at a faster speed when the deviation from PPP gets larger. This is consistent with the existence of nonlinear dynamics in the real exchange rate, implying that the speed of mean reversion is state dependent.

In fact, the empirical literature on PPP has pursued formally the idea of nonlinearities in real exchange rate dynamics since the second half of the 1990s, providing several insights. In essence, the procedures conventionally applied by researchers to examine long-run PPP assume that the real exchange rate follows a linear process and depends on its own past values. In turn, this means that adjustment to the PPP equilibrium is assumed to be both continuous and of constant speed, regardless of the *size* of the past deviation from PPP. However, the presence of transactions costs may imply a nonlinear process, which has important implications for the conventional tests of long-run PPP. A number of authors have developed theoretical models of nonlinear real exchange rate adjustment arising from transactions costs in international arbitrage (for example, Dumas, 1992). In most of these models proportional or 'iceberg' transport costs ('iceberg' because a fraction of goods is presumed to 'melt' when shipped) create a band for the real exchange rate within which the marginal cost of arbitrage exceeds the marginal benefit. Assuming instantaneous goods arbitrage at the edges of the band then typically implies that the thresholds become reflecting barriers.

Drawing on recent work on the theory of investment under uncertainty, some of these studies show that the thresholds should be interpreted more broadly than as simply reflecting shipping costs and trade barriers per se, but also as resulting from the sunk costs of international arbitrage and the resulting tendency for traders to wait for sufficiently large arbitrage opportunities to open up before entering the market.

Taylor (2001) has shown that empirical estimates of the half-life of shocks to the real exchange rate may be biased upwards because of two empirical pitfalls. The first pitfall identified by Taylor relates to temporal aggregation in the data. Using a model in which the real exchange rate follows an AR(1) process at a higher frequency than that at which the data is sampled, Taylor shows analytically that the degree of upward bias in the estimated half-life rises as the degree of temporal aggregation increases, that is, as the length of time between observed data points increases. The second pitfall highlighted by Taylor concerns the possibility of nonlinear adjustment of real exchange rates. On the basis of Monte Carlo experiments with a nonlinear artificial data generating process, Taylor shows that there can also be substantial upward bias in the estimated half-life of adjustment from assuming linear adjustment when in fact the true adjustment process is nonlinear.

Overall, the theoretical models based on non-zero transactions costs of arbitrage in international goods markets suggest that the exchange rate will become increasingly mean-reverting with the size of the past deviation from the PPP equilibrium level. In other words, the speed at which the real exchange rate reverts to PPP depends on the size of the past deviation from PPP itself. When the real exchange rate is arbitrarily close to PPP, the real exchange rates may move randomly next period since agents have no arbitrage opportunities available in international goods markets. At the other extreme, when the real exchange rate deviates from PPP by a very large extent, it is likely that arbitrage forces will imply movements of goods and changes in prices (expressed in a common currency) that induce fast reversion to PPP.

Note that these arguments rationalize mean reversion in the real exchange rates based on ideas that relate to the law of one price in the sense that refers to tradable goods only. However, we argue that this is reasonable given that Engel (1999), in a study that measures the proportion of dollar real exchange rate movements that can be accounted for by movements in the relative prices of non-tradable goods, finds that relative prices of non-tradable goods appear to account for essentially none of the movement of dollar real exchange rates. Hence, much of the explanation for the time series properties of PPP deviations is likely to reside in the behaviour of deviations from the law of one price, that is, movements in the relative prices of tradable goods.

To turn to the empirics of nonlinear reversion to PPP, models that capture the nonlinear behaviour described above have shed light on several aspects of the behaviour of the real exchange rate. For example, Michael, Nobay and Peel (1997) apply a nonlinear model to monthly interwar data for the French franc–US dollar, French franc–UK sterling and UK sterling–US dollar as well as for the Lothian and Taylor (1996) long-span data-set described in Figure 1. Their results clearly reject the linear framework in favour of a nonlinear process. The systematic pattern in the estimates of the nonlinear models provides strong evidence of mean-reverting behaviour for PPP deviations, and helps explain the mixed results of previous studies.

Using data for the recent float alone, Taylor, Peel and Sarno (2001) provide strong confirmation that the four major real bilateral dollar exchange rates obtaining among the G5 economies are well characterized by nonlinearly mean-reverting processes. Their estimated model implies an equilibrium level of the real exchange rate in the neighbourhood of which the behaviour of the real exchange rate is close to a random walk, becoming increasingly mean-reverting with the absolute size of the deviation from equilibrium, consistent with the theoretical literature on the nature of real exchange rate dynamics in the presence of international arbitrage costs.

Impulse response functions based on these nonlinear real exchange rate models suggest that the speed of real exchange rate adjustment is typically much faster than the very slow speeds of real exchange rate adjustment often recorded in the literature. For example, the estimated half-lives (in months) for dollar–sterling and dollar–yen are the following

Shock (%):	*40*	*30*	*20*	*10*	*5*	*1*
Dollar–sterling	10	20	22	26	29	32
Dollar–yen	14	18	24	32	38	42

where the first row reports the size of the shock (in percentage terms) to the real exchange rate. The estimated half-lives of these major real dollar exchange rates illustrate the nonlinear nature of the response to shocks, with larger shocks mean-reverting much faster than smaller shocks. The dollar–sterling rate displays quite fast mean reversion, ranging from a half-life of less than one year for the largest shocks of 40 per cent to just under three years for small shocks of one percent; for shocks of five to ten per cent, the half-lives are just over two years. The dollar–yen displays higher persistence, with half-lives ranging from 14 to 42 months. These results therefore seem to shed some light on the PPP puzzles.

Only for small shocks occurring when the real exchange rate is near its equilibrium do nonlinear models consistently yield half-lives in the range of three to five years, which Rogoff (1996) terms ‘glacial’. For dollar–sterling, even small shocks of one to five per cent have a half-life less than three years; for larger shocks, the speed of mean reversion is even faster. An interesting experiment in terms of gauging the extent to which market integration and the reduction of trade costs impacts on the degree of mean reversion in real exchange rates is provided by the advent of the euro in 1999. Koedijk, Tims and van Dijk (2004) provide empirical evidence that the introduction of the euro and, more generally, the process of economic integration in Europe have accelerated convergence to PPP, consistent with a transactions-costs goods-market arbitrage view of the mean reversion properties of the real exchange rate.

The vast empirical literature briefly reviewed here suggests that there are at least three features that are potentially important in designing a suitable model for the deviations from PPP. The first feature is that the model needs to allow for the fact that adjustment towards PPP is likely to occur at different speeds via nominal exchange rates and prices. The majority of empirical studies on PPP are based on univariate representations of the real exchange rate. This approach is valid only if certain (common factor) restrictions in the process linking exchange rates and prices are satisfied. Employing a model which does not impose these restrictions increases the power of the econometrics methods employed, while allowing us to shed light on the relative importance of nominal exchange rates and prices in restoring the PPP equilibrium. The second desirable feature is that the model allows explicitly for the possibility that different monetary and exchange rate regimes generate regime shifts in the structural dynamics of PPP deviations, especially when one uses long spans of data. The third feature is that the model might be nonlinear, in accordance with the growing evidence that exchange rate dynamics displays statistically and economically important nonlinearities.

To account for these three features, Sarno and Valente (2006) extend the long-span data used by Obstfeld and Taylor (2004) and apply a general modeling methodology in which regime changes and nonlinearities in the dynamic relationship between exchange rates and prices are explicitly allowed for, without imposing common factor restrictions. They examine the G5 countries across different exchange rate regimes, including the gold standard, the Bretton Woods period, and the floating regime since the 1970s. Over the sample period examined, the economic history of the countries involved has seen a number of fundamental changes in monetary and exchange rate regimes, institutional structure and policy targets which, in addition to the continuous evolution of the financial system and various nominal and real shocks, represent serious potential pitfalls to researchers attempting to find an empirical model of the deviations from PPP that is stable over the full sample.

Sarno and Valente's results are supportive of long-run PPP for each of the four major exchange rates examined and of a simple basic conjecture: under fixed exchange rate regimes relative prices adjust to restore deviations from long-run equilibrium, while nominal exchange rates bear most of the burden of adjustment under flexible exchange rate regimes. This is consistent with the general notion that the relative importance of exchange rates and relative prices in restoring the long-run equilibrium level of the exchange rate varies over time and is affected by the nominal exchange rate arrangement in operation. Further, the estimated half-lives of the nonlinear exchange rate models are sensibly different for fixed and floating regimes. Under fixed exchange rate regimes, shocks to the PPP equilibrium relationship may be very persistent, implying half lives – on average across the exchange rates considered – from over five years for large real exchange rate shocks of 20 per cent to almost ten years for small shocks of 1 per cent. However, the corresponding half-lives during floating exchange rate regimes are drastically shorter, since the nominal exchange rate is allowed to operate and contribute to restoring PPP. In fact, shocks will last for less than one year on average for 20 per cent shocks. The properties of PPP deviations under floating exchange regimes implied by their model appear to be fairly consistent with standard models of open economy macroeconomics and with their dynamic properties under conventional nominal rigidities (for example, Chari, Kehoe and McGrattan, 2002). It is only under fixed exchange rate regimes, when the burden of adjustment towards PPP relies exclusively on relative prices, that we observe remarkably long half-lives of PPP shocks.

Concluding remarks

PPP remains an essential element of open economy macroeconomics for at least two reasons. First, it is a benchmark by which to judge the level of an exchange rate. Cassel argued that without PPP there would be no meaningful way of discussing overvaluation or undervaluation. That recognition has found a very concrete expression in the real exchange rate series now routinely calculated and reported by governments, international organizations and financial institutions. These series show exchange-rate-adjusted price relatives for a country relative to its trading partners. The series are constructed on the basis of GDP deflators, unit labour costs, manufacturing prices and wholesale prices for all major industrialized countries and increasingly for developing countries, too. They are used to judge changes in a country's external competitiveness, thus implicitly assuming, as Cassel did, that movements in *equilibrium* relative prices are negligible. Changes in real exchange rates then (and only then) unambiguously translate into changes in competitiveness from which to expect changes in trade flows and net exports. There is no question that these data provide a useful benchmark or starting point for policy discussion.

The second use of PPP is to serve as a simple prediction model for exchange rates at medium and long horizons. Under perfectly flexible wages and prices a monetary expansion would lead to equi-proportionate increases in wages, prices and the exchange rate, leaving all real variables unchanged. This combination of the quantity theory and PPP is an important insight in guiding policy. Expansionary monetary policy can be effective only if wages and prices are less than fully flexible and will be more effective the more flexible the exchange rate is. The essential channel is the real depreciation of the exchange rate that served to create employment, at least for a while. Similarly, exchange depreciation can be effective only if money wages and prices are unresponsive. Policy can be effective only if PPP fails to hold. Macroeconomic theory goes increasingly in the direction of sound microfoundations, information, contracting, and pricing models under transactions costs to explore what the basis of PPP failure is in the short run and to determine the resulting

extent and persistence of policy effects.

See Also

- international finance
- real exchange rates

Bibliography

Aizenman, J. 1986. Testing deviations from purchasing power parity (PPP). *Journal of International Money and Finance* 5, 25–35.

Balassa, B. 1964. The purchasing power parity doctrine: a reappraisal. *Journal of Political Economy* 72, 584–96.

Baxter, M. and Stockman, A. 1989. Business cycles and the exchange rate regime: some international evidence. *Journal of Monetary Economics* 23, 377–400.

Bhagwati, J. 1984. Why services are cheaper in poor countries. *Economic Journal* 94, 279–86.

Cassel, G. 1916. The present situation of the foreign exchanges I. *Economic Journal* 26, 62–5.

Cassel, G. 1918. Abnormal deviations in international exchanges. *Economic Journal* 28, 413–15.

Cassel, G. 1922. *Money and Foreign Exchange After 1914*. London: Macmillan, 1930.

Cassel, G. 1928a. *Foreign Investments*. Lectures of the Harris Foundation. Chicago: University of Chicago Press.

Cassel, G. 1928b. *Post-war Monetary Stabilization*. New York: Columbia University Press.

Chari, V., Kehoe, P. and McGrattan, E. 2002. Can sticky price models generate persistent and volatile real exchange rates? *Review of Economic Studies* 69, 533–63.

Crucini, M., Telmer, C. and Zachariadis, M. 2005. Understanding European real exchange rates. *American Economic Review* 95, 724–38.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84, 1161–76.

Dornbusch, R. 1982. PPP exchange rate rules and macroeconomic stability. *Journal of Political Economy* 90, 158–65.

Dornbusch, R. 1987. Exchange rates and prices. *American Economic Review* 77, 93–106.

Dornbusch, R., Fischer, S. and Samuelson, P. 1977. Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67, 823–39.

Dumas, B. 1992. Dynamic equilibrium and the real exchange rate in spatially separated world. *Review of Financial Studies* 5, 153–80.

Engel, C. 1999. Accounting for US real exchange rate changes. *Journal of Political Economy* 107, 507–38.

Fleming, M. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9, 369–79.

Frankel, J. 1990. Zen and the art of modern macroeconomics: the search for perfect nothingness. In *Monetary Policy for a Volatile Global Economy*, ed. W. Haraf and T. Willett. Washington DC: American Enterprise Institute.

Frenkel, J. 1981. Flexible exchange rates, prices and the role of ‘news’: lessons from the 1970s. *Journal of Political Economy* 89, 665–705.

Frenkel, J. and Johnson, H., ed. 1975. *The Monetary Approach to the Balance of Payments*. London: George Allen and Unwin.

Frenkel, J. and Johnson, H., ed. 1978. *The Economics of Flexible Exchange Rates*. Reading, MA: Addison Wesley.

Great Britain. 1810. *Report from the Select Committee on the High Price of Gold Bullion: Ordered by the House of Commons to be printed, 8 June 1810*. Great Britain: Parliamentary Debates.

Haberler, G. 1961. *A Survey of International Trade Theory: Special Papers in International Economics*. Princeton: Princeton University Press.

Harrod, R. 1939. *International Economics*. Chicago: University of Chicago Press, 1957.

Holmes, J. 1967. The purchasing power parity theory: in defense of Gustav Cassel as a modern theorist. *Journal of Political Economy* 75, 686–95.

Hsieh, D. 1982. The determinants of the real exchange rate. *Journal of International Economics* 12, 355–62.

Imbs, J., Mumtaz, H., Ravn, M. and Rey, H. 2005. PPP strikes back: aggregation and the real exchange rate. *Quarterly Journal of Economics* 120, 1–43.

Isard, P. 1977. How far can we push the ‘law of one price’. *American Economic Review* 67, 942–8.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London and New York: Macmillan and St. Martin's Press for the Royal Economic Society, 1971.

Koedijk, K., Tims, B. and van Dijk, M. 2004. Purchasing power parity and the euro area. *Journal of International Money and Finance* 23, 1081–107.

Kravis, I., Heston, A. and Summers, R. 1978. Real GDP per capita for more than one hundred countries. *Economic Journal* 88, 215–42.

Kravis, I., Heston, A. and Summers, R. 1982. *World Product and Income: International Comparisons of Real Gross Product*. Baltimore, MD: Johns Hopkins University Press.

Kravis, I. and Lipsey, R. 1978. Price behavior in the light of balance of payments theories. *Journal of International Economics* 8, 193–246.

Kravis, I. and Lipsey, R. 1983. *Toward an Explanation of National Price Levels*. Princeton Studies in International Finance No. 52. Princeton: Princeton University Press.

Lothian, J. and Taylor, M. 1996. Real exchange rate behavior: the recent float from the perspective of the past two centuries. *Journal of Political Economy* 104, 488–510.

Michael, P., Nobay, A. and Peel, D. 1997. Transactions costs and nonlinear adjustment in real exchange rates: an empirical investigation. *Journal of Political Economy* 105, 862–79.

Mundell, R. 1968. *International Economics*. London and New York: Macmillan.

Mundell, R. 1971. *Monetary Theory*. Pacific Palisades, CA: Goodyear.

Mussa, M. 1979. Empirical regularities in the behavior of exchange rates and theories of the foreign exchange market. In *Policies for Employment, Prices, and Exchange Rates*, vol. 11, ed. K. Brunner and A. Meltzer, Carnegie-Rochester Conference Series. Amsterdam: North-Holland.

Nurkse, R. 1944. *International Currency Experience: Lessons of the Interwar Period*. Geneva: League of Nations.

Obstfeld, M. and Taylor, A. 2004. *Global Capital Markets: Integration, Crisis, and Growth*. Cambridge and New York: Cambridge University Press.

Officer, L. 1984. *Purchasing Power Parity and Exchange Rates*. Greenwich, CT: JAI Press.

Rogoff, K. 1996. The purchasing power parity puzzle. *Journal of Economic Literature* 34, 647–68.

Samuelson, P.A. 1964. Theoretical notes on trade problems. *Review of Economics and Statistics* 46, 145–54.

Sarno, L. 2005. Towards a solution to the puzzles in exchange rate economics: where do we stand? *Canadian Journal of Economics* 38, 673–708.

Sarno, L. and Valente, G. 2006. Deviations from purchasing power parity under different exchange rate regimes: do they revert and, if so, how? *Journal of Banking and Finance* 30, 3147–69.

Schumpeter, J. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Taylor, A. 2001. Potential pitfalls for the purchasing power parity puzzle? Sampling and specification biases in mean reversion tests of the law of one price. *Econometrica* 69, 473–98.

Taylor, A. and Taylor, M. 2004. The purchasing power parity debate. *Journal of Economic Perspectives* 18(4), 135–58.

Taylor, M., Peel, D. and Sarno, L. 2001. Nonlinear mean-reversion in real exchange rates: toward a solution to the purchasing power parity puzzles. *International Economic Review* 42, 1015–42.

Viner, J. 1937. *Studies in the Theory of International Trade*. London: George Allen and Unwin.

Williamson, J. 1965. *The Crawling Peg*. Essays in International Finance No. 50. Princeton: Princeton University Press.

Williamson, J., ed. 1982. *The Crawling Peg: Past Performance and Future Prospects*. London: Macmillan.

Williamson, J. 1983. *The Exchange Rate System*. Institute for International Economics, Policy Analyses in International Economics. Cambridge, MA: MIT Press.

Yeager, L. 1958. A rehabilitation of purchasing power parity. *Journal of Political Economy* 66, 516–30.

Young, J. 1925. *European Currency and Finance*. Commission of Gold and Silver Enquiry, United States Senate, Serial 9. New York and London: Garland, 1983.

How to cite this article

Sarno, Lucio. "purchasing power parity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000260> doi:10.1057/9780230226203.1370

The New Palgrave Dictionary of Economics Online

purification

Stephen Morris
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Many complete information games have equilibria only in mixed strategies, where players are required to randomize over pure strategies among which they are indifferent according to a fixed probability distribution. Harsanyi showed that, if such a game were perturbed – with each player observing an idiosyncratic independent payoff shock with continuous support – then the perturbed games will have pure strategy equilibria which will converge to the mixed strategy equilibrium as the perturbations become small. We review the strong conditions on perturbations under which Harsanyi's result holds and discuss weaker conditions on perturbations which ensure the existence of pure strategy equilibria without approximating all mixed equilibria of the unperturbed game.

Keywords

approachability; complete information games; correlation; extensive form games; finite dynamic games; Harsanyi, J. C.; incomplete information games; mixed strategy equilibria; Nash equilibrium; normal form games; overlapping generations model; pure strategy equilibria; purification; regular Nash equilibrium; subgame perfection

Article

In a mixed strategy equilibrium of a complete information game, players randomize between their actions according to a particular probability distribution, even though they are indifferent between their actions. Two criticisms of such mixed strategy equilibria are (*a*) that players do not seem to randomize in practice, and (*b*), if a player were to randomize, why would he or she choose to do so according to probabilities that make other players indifferent between their strategies?

Since many games have no pure strategy equilibria, it is important to provide a more compelling rationale for the play of mixed strategy equilibria.

Harsanyi (1973) gave a ‘purification’ interpretation of mixed strategy equilibrium that resolves these criticisms. The complete information-game payoffs are intended as an approximate description of the strategic situation, but surely do not capture every consideration in the minds of the players. In particular, suppose that a player has some small private inclination to choose one action or another independent of the specified payoffs, but this information is not known to the other players. Then the behaviour of such players will look – to their opponents and to outside observers – as if they are randomizing between their actions, even though they do not experience the choice as randomization. Because of the private payoff perturbation, they will not in fact be indifferent between their actions, but will almost always be choosing a strict best response. Harsanyi's remarkable purification theorem showed that all equilibria (pure or mixed) of almost all complete information games are the limit of pure strategy equilibria of perturbed games where players have independent small shocks to payoffs.

There are other inpts of mixed strategy play: Reny and Robson (2004) present an analysis that unifies the purification interpretation with the ‘classical’ interpretation that players randomize because they think that there is a small chance that their mixed strategy may be observed in advance by other players. But Harsanyi's purification theorem justly provides the leading interpretation of mixed strategy equilibria among game theorists today.

I first review Harsanyi's theorem. Harsanyi's result applies to regular equilibria of complete information games with independent payoff shocks; since many equilibria of interest (especially in dynamic games) are not regular, Harsanyi's result cannot be relied upon in many economic settings of interest; I therefore briefly review what little is known about such extensions.

Harsanyi's theorem has two parts: (*a*) pure strategy equilibria always exist in suitably perturbed versions of a complete information game; and (*b*) for any regular equilibrium of a complete information game and any sequence of such perturbed games converging to the complete information game, there is a sequence of pure strategy equilibria converging to the regular equilibrium. An important literature has ignored the latter approachability question and focused on the former pure strategy existence qst, identifying conditions on an information structure – much weaker than Harsanyi's – to establish the existence of pure strategy equilibria. I conclude by reviewing these papers.

Harsanyi's th

Consider two players engaging in the symmetric coordination game below.

	A	B
A	2,2	0,0
B	0,0	1,1

As well as the pure strategy Nash equilibria (A,A) and (B,B), this game has a symmetric mixed strategy Nash equilibrium where each player chooses A with probability $\frac{1}{3}$ and B with probability $\frac{2}{3}$.

But suppose that, in addition to these common knowledge payoffs, each player i observes a payoff shock depending on the action he or she chooses. Thus,

	A	B
A	$2+\varepsilon.\eta_{1A}, 2+\varepsilon.\eta_{2A}$	$\varepsilon.\eta_{1A}, \varepsilon.\eta_{2B}$
B	$\varepsilon.\eta_{1B}, \varepsilon.\eta_{2A}$	$1+\varepsilon.\eta_{1B}, 1+\varepsilon.\eta_{2B}$

where $\varepsilon > 0$ is a commonly known parameter measuring the size of payoff shocks and (η_{1A}, η_{1B}) and (η_{2A}, η_{2B}) are distributed independently of each other, and player *i* observes only (η_{iA}, η_{iB}). Finally, assume that, for each player *i*, $\eta_i = \eta_{iA} - \eta_{iB}$ is distributed according to a continuous density *f* on the real line with corresponding c.d.f. *F*.

This perturbed game is one with incomplete information, where a player's strategy is a function $s_i: \mathbb{R} \to \{A, B\}$. In equilibrium, each player will follow a threshold strategy of the form

$$s_i(\eta_i) = \begin{cases} A, & \text{if } \eta_i \geq z_i \\ B, & \text{if } \eta_i < z_i \end{cases}.$$

Under such a strategy, the *ex ante* probability that player *i* will choose action *B* is $\pi_i = F(z_i)$, and the probability he or she will choose *A* is $1 - \pi_i$. Thus we can re-parameterize the strategy as

$$s_i(\eta_i) = \begin{cases} A, & \text{if } \eta_i \geq F^{-1}(\pi_i) \\ B, & \text{if } \eta_i < F^{-1}(\pi_i) \end{cases}.$$

Let us look for a strategy profile (s_1,s_2) of the incomplete information game, parameterized by (π_1, π_2), that forms an equilibrium of the incomplete information game. Since player 1 thinks that player 2 will choose action *A* with probability $1-\pi_2$ and action *B* with probability π_2, player 1's expected payoff gain from choosing action *A* over action *B* is then

$$2(1-\pi_2) + \varepsilon.\eta_1 - \pi_2.$$

Thus player 1's best response must be to follow a threshold strategy with threshold

$$F^{-1}(\pi_1) = \frac{3\pi_2 - 2}{\varepsilon}$$

or

$$\varepsilon F^{-1}(\pi_1) = 3\pi_2 - 2.$$

Symmetrically, we have

$$\varepsilon F^{-1}(\pi_2) = 3\pi_1 - 2.$$

Thus, there will be a symmetric equilibrium where both players choose action *B* with probability π if and only if

$$\varepsilon F^{-1}(\pi) = 3\pi - 2.$$

For small ε, this equation has three solutions tending 0, $\frac{2}{3}$ and 1 as $\varepsilon \to 0$. These solutions correspond to the three symmetric Nash equilibria of the above complete information game, respectively: (*a*) both always choose *A*, (*b*) both choose *B* with probability $\frac{2}{3}$, and (*c*) both always choose *B*.

Harsanyi's purification theorem generalizes the logic of this example. If we add small independent noise to each player's payoffs, then each player will almost always have a unique best response and thus the perturbed game will have a pure strategy equilibrium. There is a system of equations that describes equilibria of the unperturbed game. If that system of equations is regular, then a small perturbation of the system of equations will result in a nearby equilibrium.

I will report a statement of Harsanyi's result due to Govindan, Reny and Robson (2003), which weakens a number of the technical conditions in the original theorem.

Consider an *I* player complete information game where each player *i* has a finite set of possible actions A_i and a payoff function $g_i: A \to \mathbb{R}$ where $A = A_1 \times \cdots \times A_I$. An equilibrium $\alpha \in \Delta(A_1) \times \cdots \times \Delta(A_I)$ is a *regular Nash equilibrium* of the complete information game if the Jacobian determinant of a continuously differentiable map characterizing equilibrium is non-zero at α (see van Damme, 1991, Definition 1.5.1, p. 39).

The μ-perturbed game is then an incomplete information game where each player *i* privately observes a vector $\eta_i \in \mathbb{R}^{|A|}$. Player *i*'s payoff in the incomplete information game if action profile *a* is chosen is then $g_i(a) + \eta_{ia}$; thus η_i is a private value shock. Each η_i is independently drawn according to a measure μ_i, where each μ_i assigns probability 0 to *i*'s expectation of η_i being equal under any pair of *i*'s pure strategies a_i and a_i', given any mixed strategy profile of the other players; Govindan, Reny and Robson (2003) note that this weak condition is implied by μ_i being absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{|A|}$. A pure strategy for player *i* in the μ-perturbation is a function $s_i: \mathbb{R}^{|A|} \to A_i$. A

pure strategy profile *s* induces a probability distribution over actions $v_s \in \Delta(A)$, where

$$v_s(a) = \Pr_{\mu}\{\eta : s_i(\eta_i) = a_i \text{ for each } i\}$$

Theorem: (Harsanyi, 1973; Govindan, Reny and Robson, 2003). Suppose that α is a regular Nash equilibrium of the complete information game and that, for each *i*, μ_i^n converges to a point mass at $0 \in \mathbb{R}^{|A|}$. Then for all $\varepsilon > 0$ and all large enough *n*, the μ-perturbed game has a pure strategy equilibrium inducing a distribution on *A* that is within ε of α, that is,

$$\left| v_s(a) - \prod_{i=1}^{I} \alpha_i(a_i) \right| \leq \varepsilon$$

for all $a \in A$.

The pure strategy equilibria are ‘essentially strict’, that is, almost all types have a strict best response. An elegant proof in Govindan, Reny and Robson (2003) simplifies Harsanyi's original proof.

Dynamic games

Harsanyi's theorem applies only to regular equilibria of a complete information game. Harsanyi noted that all equilibria of almost all finite complete information games are regular, where ‘almost all’ means with probability one under Lebesgue measure on the set of payoffs. Of course, normal form games derived from general extensive form games are not generic in this sense. This raises the question of whether mixed strategy equilibria of extensive form games are purifiable in Harsanyi's sense.

Here is an economic example suggesting why this is an important qst. Consider an infinite overlapping generations economy where agents live for two periods; the young are endowed with two units of an indivisible and perishable consumption good, and the old have no endowment. Each young agent has the option of transferring one unit of consumption to the current old agent. Each agent's utility function is strictly increasing in own consumption when young and old, and values smoothed consumption (one when young, one when old) strictly more highly than consuming the endowment (two when young, none when old). Under perfect information, this game has a ‘social security’ subgame perfect Nash equilibrium where each young agent transfers one unit to the old agent if and only if no young agent failed to do so in the past. But suppose instead that each young agent observes only whether the previous young agent made a transfer, and restrict attention to subgame perfect Nash equilibria. Then Bhaskar (1998) has shown that there is no pure strategy equilibrium with a positive probability of transfers (in fact, this conclusion remains true if all agents only observe history of any commonly known finite length). To see why, suppose there was such an equilibrium: if the young agent at date *t* does not transfer, then the young agent at date $t+1$ must punish by not making a transfer; but the young agent at date $t+2$ did not observe the date *t* outcome, and so will think that the young agent at date $t+1$ deviated, and will therefore not make transfers; so the young agent at date $t+1$ would have an incentive to make transfers, and not to punish as required by the equilibrium strategy.

However, Bhaskar shows that there are mixed strategy equilibria with positive transfers. In particular, there is an equilibrium where the young always transfers in the first period or if he or she observed transfers in the previous period, and randomizes between making transfers or not if he or she did not observe transfers. This strategy profile attains the efficient outcome and involves mixing off the equilibrium path only. It is natural to ask whether this equilibrium can be ‘purified’: suppose that each young agent obtains a small ‘altruism’ payoff shock that makes transfers to the old slightly attractive. The mixed strategy might then be the limit of pure strategy equilibria where the more altruistic agents make the transfers and the less altruistic agents do not. However, Bhaskar shows that the mixed strategy equilibria cannot be purified. The logic of Harsanyi's purification result breaks down because the equilibrium is not regular.

Very little is known in general about purifiability of mixed strategy equilibria in extensive form games. Results will presumably depend on the regularity of the equations characterizing equilibria and the modelling of payoff choices in the extensive form (for example, do shocks occur at the beginning of the game or at each decision node?). The best hope of a positive purification result would presumably be in finite dynamic games, where Harsanyi's regularity techniques might be applied. But Bhaskar (2000) gives an example of a simple finite extensive form game where mixed strategy equilibria are not purifiable because of the non-regularity of equilibria even for generic assignment of payoffs to terminal

nodes. Mixed strategy equilibria play an important role in recent developments of the theory of repeated games. Bhaskar, Mailath and Morris (2006) report some positive and negative purification results in that context.

Purification without approachability

Harsanyi's purification theorem has two parts. First, the ‘purification’ part: all equilibria of the perturbed game are essentially pure; second, the ‘approachability’ part: every equilibrium of a complete information game is the limit of equilibria of such perturbed games. For the first part, Harsanyi's theorem uses the assumption of sufficiently diffuse independent payoff shocks. Only the second part required the strong regularity properties of the complete information game equilibria.

Radner and Rosenthal (1982) addressed a weaker version of the purification part of Harsanyi's theorem, asking what conditions on the information system of an incomplete information game will ensure that for every equilibrium (perhaps mixed) there exists an outcome equivalent pure strategy equilibrium. Thus they did not ask that every equilibrium be essentially pure and they did not seek to approximate mixed strategy equilibria of any unperturbed game. Each agent observing a signal with an atomless independent distribution is clearly sufficient for such a ‘purification existence’ result (whether or not the signal is payoff relevant). But what if there is correlation?

A simple example from Radner and Rosenthal (1982) illustrates the difficulty. Suppose that two players are playing matching pennies and each player *i* observes a payoff-irrelevant signal x_i, where (x_1,x_2) are uniformly distributed on $\{(x_1, x_2) \in \mathbb{R}^2_+ | 0 \leq x_1 \leq x_2 \leq 1\}$. In any equilibrium, almost all types of each player must assign probability $\frac{1}{2}$ to his or her opponent choosing each action (otherwise, that player would be able to obtain a payoff greater than his or her value in the zero sum game). Yet it is impossible to generate pure strategies of the players that make this property hold true. Another illustration of the importance of correlation for purification occurs in Carlsson and van Damme (1993), where it is shown that, while small independent noise leads to Harsanyi's purification result, small highly correlated noise leads to the selection of a unique equilibrium (the comparison is made explicitly in their Appendix B).

Radner and Rosenthal (1982) show the existence of a pure strategy equilibrium if each player observes a payoff-irrelevant signal with an atomless distribution and each player *i*'s payoff-irrelevant signal and payoff-relevant information (which may be correlated) are independent of each other player's payoff-irrelevant signal. This result extends if players observe additional finite private signals which are also uncorrelated with others’ atomless payoff-irrelevant signals. Their method of proof builds on the argument of Schmeidler (1973) showing the existence of a pure strategy equilibrium in a game with a continuum of players. Radner and Rosenthal (1982) also present a number of counter-exs – in addition to the matching pennies example above – with non-existence of pure strategy equilibrium. Milgrom and Weber (1985) show the existence of a pure strategy equilibrium if type spaces are atomless and independent conditional on a finite valued common state variable with payoff interdependence occurring only via the common state variable. Their result – which neither implies nor is implied by the Radner and Rosenthal (1982) conditions – has been used in many applications. Aumann et al. (1983) show that, if every player has a conditionally atomless distribution over others’ types (that is, his or her conditional distribution has no atoms for almost every type), there exists a pure strategy ε-equilibrium. Their theorem thus covers the matching pennies example described above.

The existence of such purifications deals with one of the two criticisms of mixed strategy equilibria raised above: people do not appear to randomize. In particular, in any such purification the ‘randomization’ represents the uncertainty in a player's mind about how other players will act, rather than deliberate randomization. This interpretation of mixed strategies was originally emphasized by Aumann (1974).

See Also

- global games
- mixed strategy equilibrium

Bibliography

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96.

Aumann, R., Katznelson, Y., Radner, R., Rosenthal, R. and Weiss, B. 1983. Approximate purification of mixed strategies. *Mathematics of Operations Research* 8, 327–41.

Bhaskar, V. 1998. Informational constraints and overlapping generations model: folk and anti-folk theorems. *Review of Economic Studies* 65, 135–49.

Bhaskar, V. 2000. The robustness of repeated game equilibria to incomplete payoff information. Mimeo, University College London.

Bhaskar, V., Mailath, G. and Morris, S. 2006. Purification in the infinitely-repeated Prisoners' Dilemma. Discussion Paper No. 1571, Cowles Foundation.

Carlsson, H. and van Damme, E. 1993. Global games and equilibrium selection. *Econometrica* 61, 989–1018.

Govindan, S., Reny, P. and Robson, A. 2003. A short proof of Harsanyi's purification th. *Games and Economic Behavior* 45, 369–74.

Harsanyi, J. 1973. Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory* 2, 1–23.

Milgrom, P. and Weber, R. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10, 619–32.

Radner, R. and Rosenthal, R. 1982. Private information and pure-strategy equilibria. *Mathematics of Operations Research* 7, 401–9.

Reny, P. and Robson, A. 2004. Reinterpreting mixed strategy equilibria: a unification of the classical and Bayesian views. *Games and Economic Behavior* 48, 355–84.

Schmeidler, D. 1973. Equilibrium points of non-atomic games. *Journal of Statistical Physics* 7, 295–301.

van Damme, E. 1991. *The Stability and Perfection of Nash Equilibria*. New York: Springer-Verlag.

How to cite this article

Morris, Stephen. "purification." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000320>
doi:10.1057/9780230226203.1371

The New Palgrave Dictionary of Economics Online

quantal response equilibria

Jacob K. Goeree, Charles A. Holt and Thomas R. Palfrey
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A quantal response specifies choice probabilities that are smooth, increasing functions of expected payoffs. A quantal response equilibrium has the property that the choice distributions match the belief distributions used to calculate expected payoffs. This stochastic generalization of the Nash equilibrium provides strong empirical restrictions that are generally consistent with data from laboratory experiments with human subjects. We define the concept of regular quantal response equilibrium and discuss several applications from the recent literature.

Keywords

coordination; extensive form games; fixed-point theorems; incomplete information; interchangeability; learning; Nash equilibrium; probabilistic choice models; quantal response equilibrium; sequential equilibria; Traveller's Dilemma

Article

Economic theory relies extensively on the assumption of perfect rationality, which makes it possible to construct general models with strong (and sometimes surprising) predictions. The evaluation of these models using field data requires the incorporation of random errors representing unobserved and omitted elements, measurement error, and so on. Evaluation of these models using data from laboratory experiments also requires an error structure, since choice behaviour in the laboratory is also noisy, showing clear mistakes and inconsistencies over time.

Probabilistic choice models (for example, logit, probit) have long been used to incorporate stochastic elements in to the analysis of individual decisions, and the quantal response equilibrium (QRE) is the analogous way to model games with noisy players. These probabilistic choice models are based on quantal response functions, which have the intuitive feature that deviations from optimal decisions are negatively correlated with the associated costs. That is, individuals are more likely to select better choices than worse choices, but do not necessarily succeed in selecting the very best choice. Formally, a quantal response function maps the vector of expected payoffs from available choices into a vector of choice probabilities that is monotone in the expected payoffs.

In a strategic game environment, a player's expected payoffs from different strategies are determined by beliefs about other players' actions, so beliefs determine expected payoffs, which in turn, generate choice probabilities according to some quantal response function. A QRE imposes the requirement that the beliefs match the equilibrium choice probabilities. Thus, QRE requires solving for a fixed point in the choice probabilities, analogous to the Nash equilibrium.

In fact, QRE is a generalization of Nash equilibrium, which converges to the Nash equilibrium as the quantal response functions become very steep, and approximate best response functions. This approach provides a useful theoretical framework for looking at comparative statics effects of parameter changes that may not alter Nash predictions. The incorporation of random elements also provides a foundation for standard statistical analysis of field and laboratory data in game theoretic applications.

A motivating example: generalized matching pennies

Before providing general definitions, it is useful to begin with a simple two-person matching pennies game in which the Row player chooses Top (T) or Bottom (B) and the Column player chooses left (L) or right (R). Row wins a penny (and Column loses a penny) if the outcome is (Top, Right) or (Bottom, Left) and Column wins a penny otherwise. Thus Row's expected payoff for Top (U_T) is a function of Column's probability of choosing Right (p_R), which is easily calculated as $U_T(p_R)=p_R-(1-p_R)=2p_R-1$, and similarly, $U_B(p_R)=1-2p_R$, so the optimal decision is to choose Top if Column is more likely to choose Right, that is, if $p_R > \frac{1}{2}$. Column's expected payoffs are computed analogously, as a function of Row's probability of choosing Top (p_T).

Figure 1 illustrates the best response functions in the unit square of mixed strategies in the game, with the y-axis representing the row player's Top choice probability and the x-axis representing the column player's Right choice. The best response for Row is indicated by the dark step function that jumps from 0 to 1 at $p_R = \frac{1}{2}$. Similarly, the Column player's best response line is the step function, shown in light grey, which crosses over from left to right at a height of $\frac{1}{2}$.

Figure 1

Players' best responses and quantal responses for a generalized matching pennies game

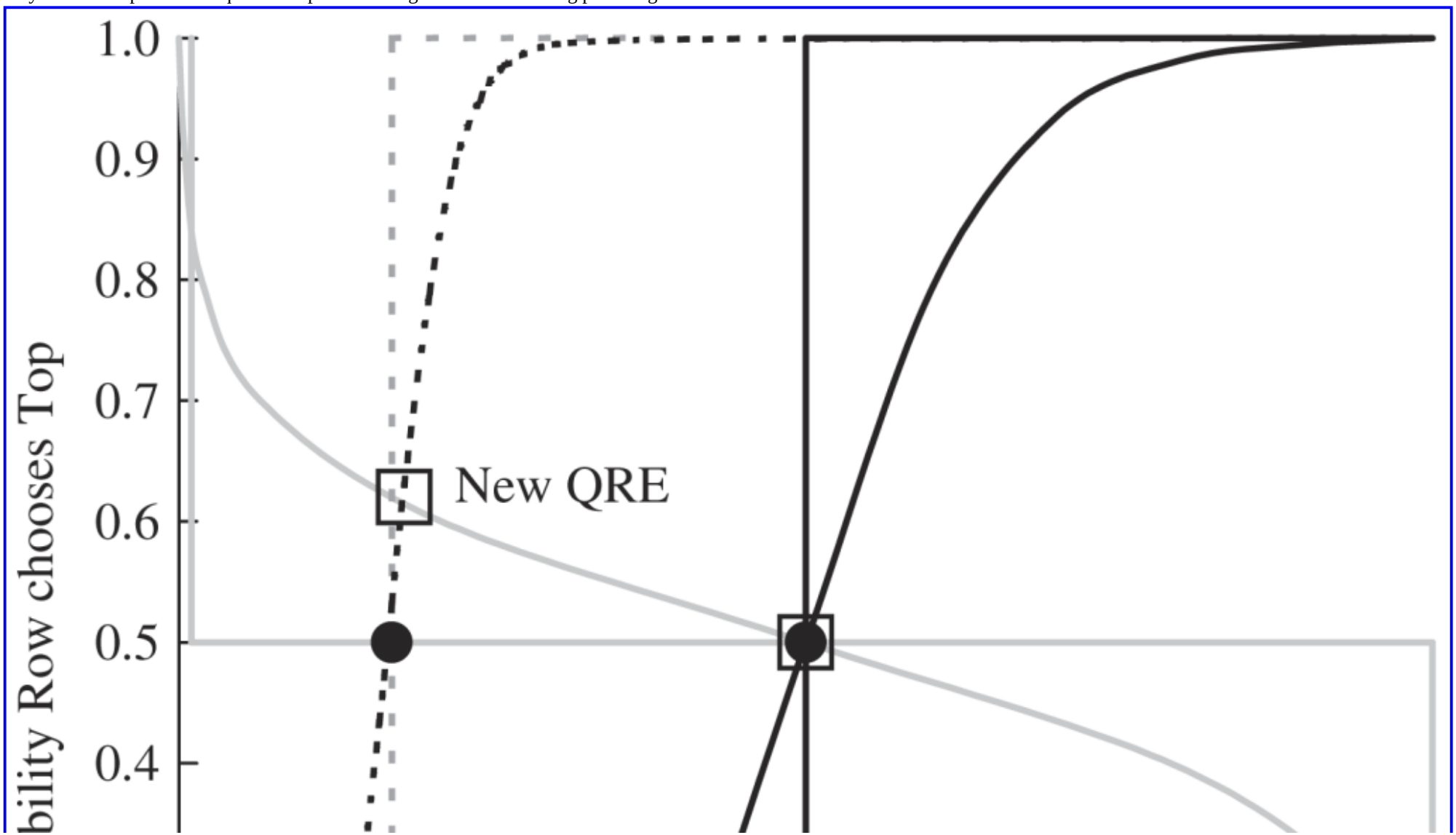

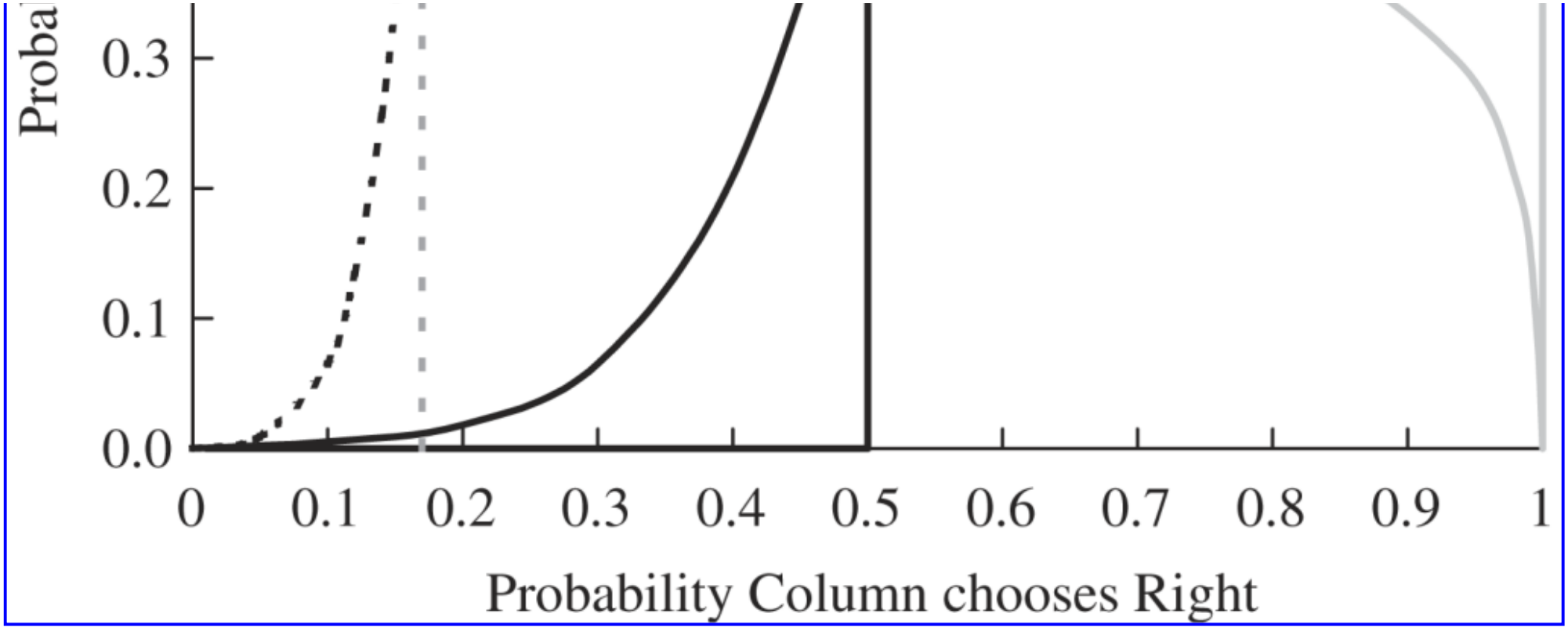

Using the same figure, we can represent a quantal response function, which smooths out the discontinuous best response function, reflecting the monotone and continuous choice probability as a function of payoffs. Such a quantal response function is illustrated by Row's dark curved line that rises smoothly from the bottom-left corner to the top-right corner. The probabilistic choice equals $\frac{1}{2}$ exactly at the point where row player is indifferent between Top and Bottom. A quantal response function is also drawn for Column. The intersection of these two quantal response functions occurs in the centre of the figure, and is the quantal response equilibrium, just as the intersection of the sharp best response function at the same point is the Nash equilibrium in mixed strategies ($\frac{1}{2}$ for each decision).

Now suppose that all payoffs stay the same except for the Top-Right outcome, which gives Row a higher payoff of 9 and Column a payoff of –1 as before. The increase in Row's Top-Right payoff shifts Row's best response line leftward, as indicated by the dashed line step function in the figure, and it also shifts Row's quantal response (smooth dashed line). The new Nash equilibrium (dot at the intersection of the step functions) is still at p_T=0.5, whereas the new QRE is at a higher level p_T=0.62. This intuitive 'own payoff' effect contrasts with the Nash equilibrium prediction of no change in Row's choice probabilities (since they are determined by the requirement that Column is indifferent). The own-payoff effect predicted by regular QRE accords with data from laboratory experiments that employ an asymmetric matching-pennies structure – for example, Ochs (1995), McKelvey, Palfrey and Weber (2000), Goeree and Holt (2001), and Goeree, Holt, and Palfrey (2003).

Definitions

Let $G=(N, S_1,\ldots,S_n, \pi_1,\ldots, \pi_n)$ be a normal-form game, where $N=\{1,\ldots,n\}$ is the set of players, $S_i=\{s_{i1},\ldots,s_{iJ(i)}\}$ is player *i*'s set of strategies and $S=S_1\times\cdots\times S_N$ is the set of strategy profiles, and $\pi_i: S_i\rightarrow R$ is player *i*'s payoff function. Furthermore, let $\Sigma_i=\Delta^{J(i)}$ be the set of probability distributions over S_i. An element $\sigma_i \in \Sigma_i$ is a mixed-strategy, which is a mapping from S_i to Σ_i, where $\sigma_i(s_i)$ is the probability that player *i* chooses pure-strategy s_i. Let $\Sigma=\Sigma_1\times\cdots\times\Sigma_N$ be the set of mixed-strategy profiles. Given a mixed-strategy profile $\sigma \in \Sigma$, player *i*'s expected payoff is $\pi_i(\sigma)=\Sigma_{s \in S} p(s)\ \pi_i(s)$, where $p(s)=\Pi_{i \in N}\ \sigma_i(s_i)$ is the probability distribution over pure-strategy profiles induced by σ.

Let P_{ij} denote the probability that player *i* selects strategy *j*. Recall that the main idea behind QRE is that strategies with higher expected payoffs are more likely to be chosen, although the best strategy is not necessarily chosen with probability 1. In other words, QRE replaces players' strict rational choice best-responses by smoothed best responses or

quantal responses.

Definition 1: P_i: $R^{J(i)} \rightarrow \Delta^{J(i)}$ *is a* ***regular quantal-response*** *function if it satisfies the following four axioms.*

- – **Interiority**: $P_{ij}(\pi_i)>0$ for all $j=1,\ldots,J(i)$ and for all $\pi_i \in R^{J(i)}$.
- – **Continuity**: $P_{ij}(\pi_i)$ is a continuously differentiable function for all $\pi_i \in R^{J(i)}$.
- – **Responsiveness**: $\partial P_{ij}(\pi_i)/\partial \pi_{ij}>0$ for all $j=1,\ldots,J(i)$ and for all $\pi_i \in R^{J(i)}$.
- – **Monotonicity**: $\pi_{ij}>\pi_{ik}$ implies that $P_{ij}(\pi_i)>P_{ik}(\pi_i)$ for all $j, k=1,\ldots, J(i)$.

These axioms are economically and intuitively compelling. Interiority ensures the model has full domain, that is, it is logically consistent with all possible data-sets. This is important for empirical applications of the model. Continuity is a technical restriction, which ensures that P_i is non-empty and single-valued. Furthermore, it seems a natural assumption since arbitrarily small changes in expected payoffs should not lead to jumps in choice probabilities. Responsiveness requires that if the expected payoff of an action increases, *ceteris paribus*, the choice probability must also increase. Monotonicity is a weak form of rational choice that involves binary comparisons of actions: an action with higher expected payoff is chosen more frequently than an action with a lower expected payoff.

Define $P(\pi)=(P_1(\pi_1),\ldots,P_n(\pi_n))$ to be regular if each P_i satisfies the above regularity axioms. Since $P(\pi) \in \Sigma$ and $\pi=\pi(\sigma)$ is defined for any $\sigma \in \Sigma$, $P \bigcirc \sigma$ defines a mapping from Σ into itself.

Definition 2: Let P be regular. A ***regular quantal response equilibrium*** *of the normal-form game G is a mixed-strategy profile* σ^* *such that* $\sigma^*=P(\sigma^*)$.

Since regularity of P includes continuity, $P \bigcirc \sigma$ is a continuous mapping. Existence of a regular QRE therefore follows directly from Brouwer's fixed-point theorem.

Theorem: There exists a regular quantal response equilibrium of G for any regular P.

Empirical implications of regular QRE

The axioms underlying regular QRE collectively have strong empirical implications, even without any parametric assumptions on P. To illustrate the nature of these restrictions, consider again the generalized matching-pennies game, where Row's payoff is X when the outcome is (top, right). If $X>1$, it is readily verified that Row's expected payoff of choosing 'top' is higher than of choosing 'bottom' when $p_R<2/(X+3)$ ($p_R>2/(X+3)$). Monotonicity therefore implies that, if (p_R^*, p_T^*) defines a regular QRE, it must satisfy the inequalities: $p_T^* \geq \frac{1}{2}$ if $p_R \geq 2/(X+3)$ and vice versa. Likewise, Column's expected payoff of choosing 'right' is higher (lower) than of choosing 'left' when $p_T < \frac{1}{2}$ ($p_T > \frac{1}{2}$). Thus (p_R^*, p_T^*) must satisfy $p_R^* \leq \frac{1}{2}$ if $p_T \geq \frac{1}{2}$, and vice versa. The region defined by these inequalities defines the set of possible regular QRE. For the specific case of $X=9$, this area is shown by the dark gray shaded area in Figure 2. The three black dots show the Nash equilibria for $X=9$ (left), $X=1$ (centre) and $X=0$ (right).

Figure 2

QRE Sets for generalized matching pennies with $X=9$ (dark) and $X=0$ (light)

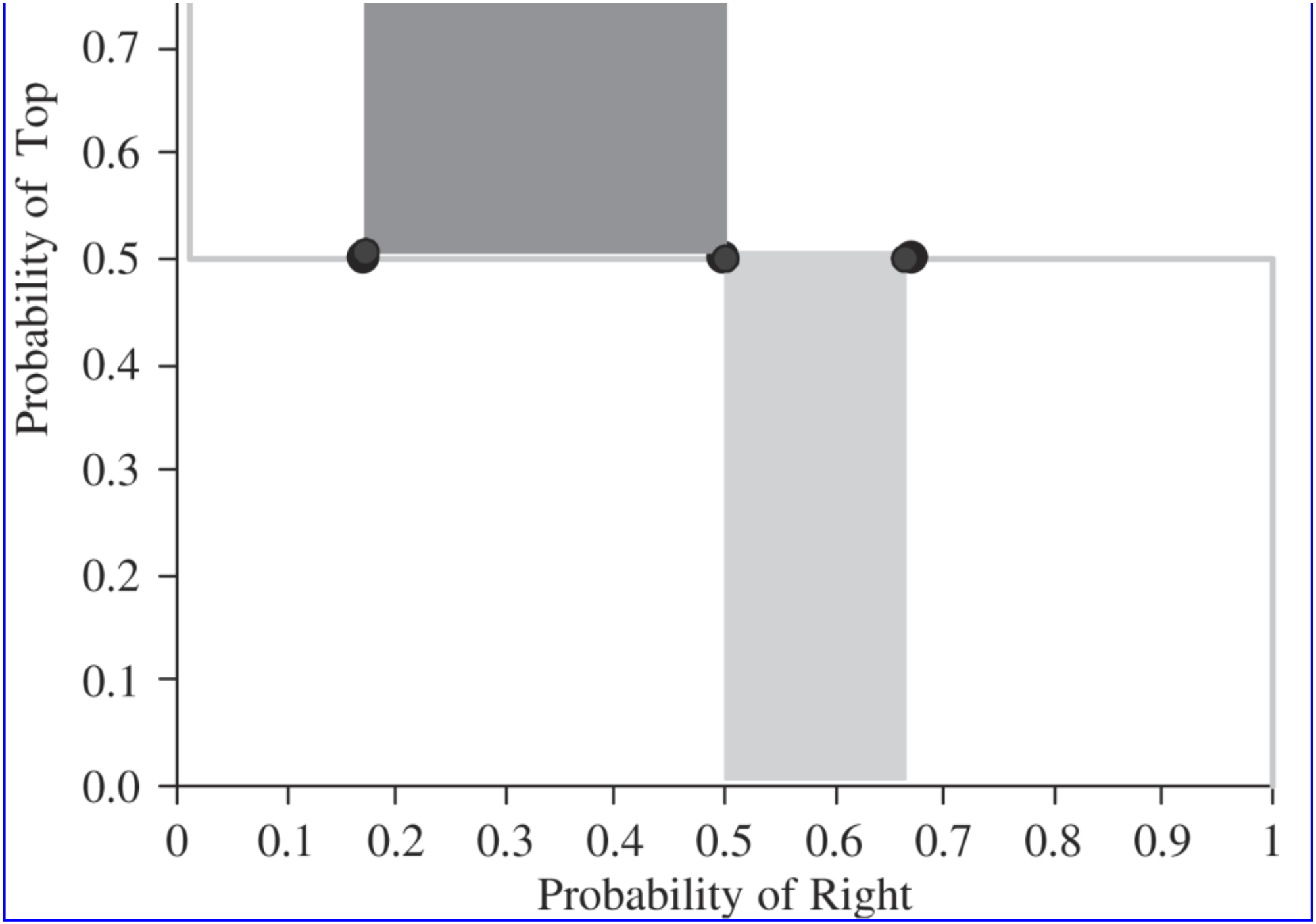

The case $-1<X<1$ can be analysed in a similar way. The set of regular QRE for $X=0$, for instance, is given by the light shaded area in Figure 2. Note that the Row player is predicted to choose ‘top’ more often than ‘bottom’ in any regular QRE when $X>1$, while the reverse is true for $X<1$. In fact, the responsiveness axiom can be used to show that if Row's payoff of the (top, right) outcome rises, Row's probability of choosing ‘top’ increases.

Proposition: **(Goeree, Holt and Palfrey, 2005)**. *In any regular QRE of the asymmetric matching pennies game, Row's probability of choosing Top is strictly increasing in X and*

Column's probability of choosing Right is strictly decreasing in X.

Quantal response equilibrium: a structural definition

The original definition of QRE (McKelvey and Palfrey, 1995) adopts an approach in the spirit of Harsanyi (1973) and McFadden (1974) whereby the choice probabilities are rationalized by privately observed, mean zero random disturbances to the expected payoffs. These disturbances are assumed to be private information to the players, thereby converting the original game into special kind of game of incomplete information. Any Bayesian equilibrium of this disturbed game is a QRE of the underlying game. The quantal response function generating the QRE is determined by the probability distribution of the random payoff disturbances.

Thus a smoothed response line can be interpreted to be the (inverse) distribution function of the differences between the disturbances, which has a value of $\frac{1}{2}$ when the expected payoffs are equal. For example, if the disturbances are i.i.d. and normally distributed, then the quantal response functions will take the shape of a 'probit' curve, while if the i.i.d. disturbances are distributed according to an extreme value distribution, the quantal response functions will have a logistic form. For example, the logit QRE for the generalized matching pennies game is a pair of probabilities that solve:

$$p_T = \frac{\exp(\lambda[(X+1)\,p_R - 1])}{\exp(\lambda[(X+1)\,p_R - 1]) + \exp(\lambda[1 - 2\,p_R])},\quad p_R = \frac{\exp(\lambda[1 - 2\,p_T])}{\exp(\lambda[1 - 2\,p_T]) + \exp(\lambda[2\,p_T - 1])}$$

where the numerators are exponential functions of the expected payoffs for the corresponding decision (*T* or *R*), and the denominators are normalizing factors that force probabilities to sum to 1. As the logit precision parameter λ increases, the response functions become more responsive to payoff differences, and the logit response functions converge to the sharp step functions shown in Figure 1.

The disturbances in the structural approach to QRE can be interpreted in several ways. One possibility is to interpret them literally. That is, one views the underlying game as simply a model of the average game being played, with each actual game player being a mean zero perturbation of the basic game. With this view, one can think of the payoff disturbances as reflecting the effects of unobservable components such as a player's mood or perceptual variations. A second possibility is to think of the players as statisticians, whose objective is to estimate the payoff of each strategy using some unknown set of instruments to perform the estimation. For general abstract games, a reasonable first cut is to suppose that their estimation errors are unbiased. The players then choose the strategy with the highest estimated expected payoffs, implicitly taking into account the fact that the other players are also estimating payoffs with some error, with the resulting equilibrium corresponding to QRE.

One can show that the quantal response function generated by i.i.d disturbances will always have the continuity and monotonicity properties of regular quantal response functions, and therefore will lead to regular QRE. In particular, the comparative static result of the previous section holds for the logit QRE (McKelvey and Palfrey, 1996). If disturbances are not i.i.d., however, non-monotonicities are possible. Haile, Hortacsu and Kosenok (2006) use this observation to show that, without restrictions on the disturbances, structural QRE can explain *any* data. One way to avoid this problem is to make the i.i.d. assumption or impose the weaker notion of *interchangeability* (Goeree, Holt and Palfrey, 2005). A second way to generate testable restrictions is to constrain the same structural assumptions to hold across different data-sets, thereby generating comparative static predictions.

Another solution is to simply work directly with the regularity axioms of Definition 1 – the resulting quantal response functions do impose empirical restrictions on the data and do not inherit the unintuitive features of the structural approach such as symmetry and strong substitutability. *Symmetry* requires that the effect of an increase in strategy *k*'s payoff on the probability of choosing strategy *j* is the same as the effect of an increase in strategy *j*'s payoff on the probability of choosing strategy *k*. *Strong substitutability* implies, among other things, that, if the payoff of strategy *k* rises, the probability of choosing *any* of the other strategies $j \neq k$ falls.

Applications: quantal response equilibrium in normal-form games

In an individual choice problem, the addition of 'noise' spreads out the distribution of decisions around the expected-payoff-maximizing decision. In contrast, expected payoffs in a game depend on other players' choice probabilities, and this interactive element can magnify the effects of noise via feedback effects. One notable example is a coordination game where each person's payoff is the minimum of all player's efforts, minus a cost parameter, *c*, times a player's own effort. If $c<1$, any common effort level is a Nash equilibrium, since a unilateral decrease below a common effort reduces the minimum and saves the cost, for a net loss of $1-c$ for each unit reduction in effort. Conversely, a unilateral effort increase does not alter the minimum, so the loss is $-c$. Therefore, *c* affects the downward slopes of the expected payoff function in each direction and, if there is any uncertainty about others'

decisions, low values of *c* make effort increases less risky. It is not surprising that reductions in the effort cost *c* tend to increase average efforts, both in laboratory experiments with human subjects and in a quantal response equilibrium with noisy behaviour (Anderson, Goeree and Holt, 2001). For the two-person coordination game experiments reported in Goeree and Holt (2005a), a reduction in the effort cost parameter from 0.75 to 0.25 raised average effort levels from 126 to 159 in the final five rounds. With payoffs in pennies and a precision parameter of 0.1, the logit QRE predictions for this game are 126 and 154 for the high and low effort costs. Thus the QRE tracks the strong behavioural response to the treatment change, whereas the range of Nash equilibria – from 110 to 170 – is unaffected by this change.

The Traveller's Dilemma is another game where small amounts of noise can have large effects. As in the coordination game, the payoffs depend on the minimum of all decisions (‘claims’). This is a two-person ‘lost luggage’ problem, where the airline representative interprets unequal damage claims as evidence that the high claimant is inflated unjustly. Each player earns the minimum of the two claims, with a penalty of *R* subtracted from the payoff for the high claimant and added to that of the low claimant. As with the coordination game, claims must be in a specified interval, but in the Traveller's Dilemma the unique Nash equilibrium is the lowest possible claim in this interval, irrespective of the magnitude of *R*, as long as the benefit *R* from a reduction below a common claim is greater than the smallest permitted claim reduction. In contrast, intuition suggests that claims will be high when the penalty from having the higher claim is low. In the Capra et al. (1999) experiment, reductions in the penalty parameter, *R*, induce dramatic increases in claims, moving the average from near Nash levels for high values of *R* to the opposite side of the range of feasible claims for low values of *R*. This strong treatment effect is tracked well by the quantal response equilibrium with the same precision parameter that tracks other coordination game data.

In addition to these applications, the QRE has been used to explain ‘anomalous’ behaviour in a wide variety of games, including signalling games, centipede games, two-stage bargaining, and overbidding in auctions (Goeree, Holt and Palfrey, 2002). In addition, the quantal response equilibrium has proven to be quite useful in the analysis of data from political science experiments: jury voting (Guarnaschelli, McKelvey and Palfrey, 2000), voter turnout (Levine and Palfrey, 2007), and behaviour in participation games (Goeree and Holt, 2005a; Cason and Van Lam, 2005).

Applications: quantal response equilibrium in extensive form games

The QRE approach has also been developed for extensive form games (McKelvey and Palfrey 1998), where the analysis in done using behavioural strategies. In the extensive form QRE, players follow Bayes' rule and calculate expected continuation payoffs based on the QRE strategies of the other players. Interiority implies that beliefs are uniquely defined at any information set and for any QRE strategy profile. Therefore issues related to belief-based refinements do not arise, and a quantal response version of sequential rationality follows immediately. When quantal response functions approach best response functions, then the limiting QRE of the extensive form game will select a subset of the *sequential equilibria* of the underlying game.

QRE in extensive form games will typically have different implied choice probabilities than would obtain if the same quantal response function were applied to the same game in its reduced normal form. This occurs for two reasons. First, QRE is not immune to ‘reduction’ of equivalent strategies, since duplicate strategies will generally change the quantal response choice probabilities, for much the same reason as the ‘red bus – blue bus’ example in discrete choice econometrics. Second, expected payoff differences are different when one collapses an extensive form game into its normal form: with behaviour strategies, expected payoffs are computed at the interim stage, conditioning on previous actions in the game; in contrast, normal form mixed strategies are calculated *ex ante*.

Summary

The quantal response equilibrium approach to the analysis of games has proven to be a useful generalization of the Nash equilibrium, especially when dealing with ‘noisy decisions’ made by boundedly rational players and by subjects in experiments. It can be extended to allow for learning and cognitive belief formation in one-shot games where learning is not possible. This approach provides a coherent framework for analysing an otherwise bewildering array of ‘biases’ and anomalies in economics.

We acknowledge financial support from the Alfred P. Sloan Foundation, the National Science Foundation (SBR 0094800 and 0551014; SES 0450712 and 0214013), and the Dutch National Science Foundation (VICI 453.03.606).

Bibliography

Anderson, S.K., Goeree, J.K. and Holt, C.A. 2001. Minimum effort coordination games: stochastic potential and the logit equilibrium. *Games and Economic Behavior* 34, 177–99.

Capra, C.M, Goeree, J.K., Gomez, R. and Holt, C.A. 1999. Anomalous behavior in a traveler's dilemma? *American Economic Review* 89, 678–90.

Cason, T.N. and Van Lam, M. 2005. Uncertainty and resistance to reform in laboratory participation games. *European Journal of Political Economy* 21, 708–37.

Goeree, J.K. and Holt, C.A. 2001. Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review* 91, 1402–22.

Goeree, J.K. and Holt, C.A. 2005a. An experimental study of costly coordination. *Games and Economic Behavior* 46, 281–94.

Goeree, J.K. and Holt, C.A. 2005b. An explanation of anomalous behavior in models of political participation. *American Political Science Review* 99, 201–13.

Goeree, J.K., Holt, C.A. and Palfrey, T.R. 2002. Quantal response equilibrium and overbidding in private-value auctions. *Journal of Economic Theory* 104, 247–72.

Goeree, J.K., Holt, C.A. and Palfrey, T.R. 2003. Risk averse behavior in asymmetric matching pennies games. *Games and Economic Behavior* 45, 97–113.

Goeree, J.K., Holt, C.A. and Palfrey, T.R. 2005. Regular quantal response equilibrium. *Experimental Economics* 8, 347–67.

Guarnaschelli, S., McKelvey, R.D. and Palfrey, T.R. 2000. An experimental study of jury decision rules. *American Political Science Review* 94, 407–23.

Haile, P., Hortacsu, A. and Kosenok, G. 2006. On the empirical content of quantal response equilibrium. Working paper, Yale School of Management, Yale University.

Harsanyi, J. 1973. Games with randomly disturbed payoffs: a new rationale for mixed strategy equilibrium. *International Journal of Game Theory* 2, 1–23.

Levine, D. and Palfrey, T.R. 2007. The paradox of voter participation: an experimental study. *American Political Science Review* 101, 143–58.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McKelvey, R.D. and Palfrey, T.R. 1995. Quantal response equilibrium for normal form games. *Games and Economic Behavior* 10, 6–38.

McKelvey, R.D. and Palfrey, T.R. 1996. A statistical theory of equilibrium in games. *Japanese Economic Review* 47, 186–209.

McKelvey, R.D. and Palfrey, T.R. 1998. Quantal response equilibrium for extensive form games. *Experimental Economics* 1, 9–41.

McKelvey, R.D., Palfrey, T.R. and Weber, R. 2000. The effects of payoff magnitude and heterogeneity on behavior in 2×2 games with a unique mixed-strategy equilibrium. *Journal of Economic Behavior and Organization* 42, 523–48.

Ochs, J. 1995. Games with unique, mixed strategy equilibria: an experimental study. *Games and Economic Behavior* 10, 202–17.

How to cite this article

Goeree, Jacob K., Charles A. Holt and Thomas R. Palfrey. "quantal response equilibria." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_Q000016> doi:10.1057/9780230226203.1372

The New Palgrave Dictionary of Economics Online

quantile regression

Moshe Buchinsky
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A semiparametric technique that has been gaining considerable popularity in economics, the quantile regression model has a number of attractive features. For example, it can be used to characterize the entire conditional distribution of a dependent variable given a set of regressors; it has a linear programming representation which makes estimation easy; and it gives a robust measure of location. Concentrating on cross-section applications, this article presents the basic structure of the quantile regression model, highlights the most important features, and provides the elementary tools for using quantile regressions in empirical applications.

Keywords

asymptotic covariance matrix; boostrap; censored quantile regression model; design matrix bootstrapping; equivariance; generalized method of moments; heteroskedasticity; kernel estimators in econometrics; least absolute deviation; linear programming; optimal minimum distance estimator; quantile regression; semiparametric estimation; Tobit model

Article

1 Introduction

The quantile regression is a semiparametric technique that has been gaining considerable popularity in economics (for example, Buchinsky, 1994). It was introduced by Koenker and Bassett (1978b) as an extension to ordinary quantiles in a location model. In this model, the conditional quantiles have linear forms. A well-known special case of quantile regression is the least absolute deviation (LAD) estimator of Koenker and Bassett (1978a), which fits medians to a linear function of covariates. In an important generalization of the quantile regression model, Powell (1984; 1986) introduced the censored quantile regression model. This model is an extension of the 'Tobit' model and is designed to handle situations in which some of the observations on the dependent variable are censored.
The quantile regression model has some very attractive features: (*a*) it can be used to characterize the entire conditional distribution of a dependent variable given a set of regressors; (*b*) it has a linear programming representation which makes estimation easy; (*c*) it gives a robust measure of location; (*d*) typically, the quantile regression estimator is more efficient than the least squares estimator when the error term is non-normal; and (*e*) *L*-estimators, based on a linear combination of quantile estimators (for example, Portnoy and Koenker, 1989) are, in general, more efficient than least squares estimators.
This article presents the basic structure of the quantile regression model. It highlights the most important features and provides the elementary tools for using quantile regressions in empirical applications. The article concentrates on cross-section applications, where the observations are assumed to be independently and identically distributed (i.i.d.).

2 The Model

2.1 Definitions and estimator

Any real-valued random variable *z* is completely characterized by its (right continuous) conditional distribution function $F(z) = \Pr(z \le a)$. For any $0 < \theta < 1$, the quantity $Q_\theta(z) \equiv F^{-1}(\theta) = \inf\left\{a: F(a) \ge \theta\right\}$ is called the θ th quantile of *z*. This quantile is obtained as a solution to a minimization problem of a particular objective function, the *check function*, given by $\rho_\theta(\lambda) = \lambda(\theta - I(\lambda < 0))$, where $I(\cdot)$ denotes the usual indicator function. That is,

$$Q_\theta(z) \equiv \underset{a}{\operatorname{argmin}} E[\rho_\theta(z-a)].$$

An estimate for the θ th quantile of z can be obtained from i.i.d. data z_i, $i = 1, \ldots n$, by minimizing the sample analogue of the population objective function defined above. That is,

$$\hat{Q}_\theta(z) \equiv \underset{a}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(z_i - a),$$

or alternatively

$$\hat{Q}_\theta(z) \equiv \underset{a}{\operatorname{argmin}} \left\{ \sum_{i:z_i \geq b} \theta|z_i - a| + \sum_{i:z_i < b} (1-\theta)|z_i - a| \right\}.$$

The last equation provides a clear intuition for the quantile estimates. The θ th quantile estimate is obtained by weighting the positive residuals by θ, while the negative residuals are weighted by the complement of θ, namely, $1-\theta$.
The extension of this idea to the case of a conditional quantile is straightforward. Suppose that the θ th conditional quantile of y, conditional on a $K\times 1$ vector of regressors $x = (1, x_2, \ldots, x_K)$, is

$$Q_\theta(y|x) = x'\beta_\theta.$$

This implies that the model can be written as

$$y = x'\beta_\theta + u_\theta, \quad (1)$$

and, by construction, it follows that $Q_\theta(u_\theta|x) = 0$.
This model, which was first introduced by Koenker and Bassett (1978b), can be viewed as a location model. That is,

$$\Pr(y \leq \tau|x) = F_{u_\theta}(\tau - x'\beta_\theta|x),$$

where u_θ has the (right continuous) conditional distribution function $F_{u_\theta}(\cdot|x)$, satisfying $Q_\theta(u_\theta|x) = 0$.
Similar to the unconditional case presented above, the population parameter vector β_θ is defined by

$$\beta_\theta \equiv \underset{\beta}{\operatorname{argmin}} E[\rho_\theta(y - x'\beta)|x].$$

The sample analogue for the θ th quantile conditional quantile is defined in a similar manner. Let (y_i, x_i), $i = 1, \ldots, n$, be an i.i.d. sample from the population. Then, $\hat{\beta}_\theta$, the estimator for β_θ, is defined by

$$\hat{\beta}_\theta = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\theta(u_{\theta i}), \text{ or alternatively} \tag{2}$$

$$\hat{\beta}_\theta = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \left\{ \sum_{i: y_i \geq x_i'\beta} \theta \left| y_i - x_i'\beta \right| + \sum_{i: y_i < x_i'\beta} (1 - \theta) \left| y_i - x_i'\beta \right| \right\}.$$

The θ th quantile regression problem in (2) can be also be rewritten as

$$\hat{\beta}_\theta = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(\theta - 1/2 + 1/2 \operatorname{sgn}(y_i - x_i'b) \right)(y_i - x_i'b),$$

where $\operatorname{sgn}(\lambda) = I(\lambda \geq 0) - I(\lambda < 0)$. The last equation gives, in turn, the K×1 vector of first-order conditions (F.O.C.):

$$\frac{1}{n} \sum_{i=1}^{n} \left(\theta - 1/2 + 1/2 \operatorname{sgn}(y_i - x_i'\hat{\beta}) \right) x_i = \frac{1}{n} \sum_{i=1}^{n} \psi(x_i, y_i, \beta) = 0, \tag{3}$$

where $\psi(x, y, \beta) = (\theta - 1/2 + 1/2 \operatorname{sgn}(y - x'\beta))x$. It is straightforward to show that under the quantile restriction $Q_\theta(u_{\theta i}|x_i) = 0$ the moment function $\psi(\cdot)$ satisfies $E[\psi(x_i, y_i, \beta_\theta)] \equiv E[\psi(x_i, y_i, \beta)]|_{\beta = \beta_\theta} = 0$. In the jargon of the generalized method of moments (GMM) framework, this establishes the validity of $\psi(\cdot)$ as a moment function. Consequently, using the methodology of Huber (1967), one can establish consistency and asymptotic normality of $\hat{\beta}_\theta$.

For illustration and discussion below, it is convenient to define the following: Let y denote the stacked vector of y_i, $i = 1, \ldots n$, and let X denote the stacked matrix of the row vectors x_i', $i = 1, \ldots, n$.

2.2 Linear programming and quantile regression

The problem in (2) can be shown to have a linear programming (LP) representation. This feature has some important consequences from both theoretical and practical standpoints.

Let the K×1 vector β be written as a difference of two non-negative vectors β^+ and β^-, that is, $\beta = \beta^+ - \beta^-$, for $\beta^+, \beta^- \geq 0$. Similarly let the n×1 residuals vector u be written as a difference of two

non-negative vectors u^+ and u^-, that is, $u = u^+ - u^-$, for $u^+, u^- \geq 0$. Furthermore, define the following quantities: $A \equiv (X, -X, I_n, -I_n)$, where I_n is an n dimensional identity matrix, $z = (\beta^{+\prime}, \beta^{-\prime}, u^{+\prime}, u^{-\prime})'$, $c = (0', 0', \theta \cdot l', (1-\theta) \cdot l')'$, **0** is a k×1 vector of zeros, and l is an n×1 vector of ones.
When written in matrix notation the problem in (2) takes the familiar *primal problem* of LP:

$$\min_z c'z$$
$$\text{subject to } Az = y,\ z \geq 0.$$

Furthermore, the dual problem, of LP is (approximately) the same as the F.O.C. given above, namely

$$\max_w w'y$$
$$\text{subject to } w'A \leq c'.$$

The *duality th* of LP implies that feasible solutions exist for both the primal and the dual problems, if the design matrix X is of full column rank, that is, $\text{rank}\ (X) = K$. The *equilibrium th* of LP guarantees then that this solution is optimal.
The LP representation of the quantile regression problem has several important implications from both computational and conceptual standpoints. First, it is guaranteed that an estimate will be obtained in a finite number of simplex iterations. Second, the parameter estimate is robust to outliers. That is, for all $y_i - x_i'\hat{\beta}_\theta > 0$, y_i can be increased toward ∞, and for all $y_i - x_i'\hat{\beta}_\theta < 0$, y_i can be decreased toward –∞, without altering the solution $\hat{\beta}_\theta$. In other words, the only thing that matters is not the exact value of y, but rather on which side of the estimated hyperplane it lies. This is important for many economic applications in which y_i might be censored, at say y_i^0. For example, for the right-censored model $\hat{\beta}_\theta$ will not be affected as long as for all i we have $y_i^0 - x_i'\hat{\beta}_\theta > 0$.

2.3 Equivariance properties

The quantile regression estimator has several important equivariance properties which help facilitate the computation procedure. That is, data-sets that are based on certain transformations of the original data set lead to estimators which are simple transformations of the original estimator. Denote the set of feasible solutions to the problem defined in (2) by $B(\theta, y, X)$. Then for every $\hat{\beta}_\theta \equiv \hat{\beta}(\theta, y, X) \in B(\theta, y, X)$ we have (see Koenker and Bassett, 1978b: Theorem 3.2):

$$\hat{\beta}(\theta, \lambda y, X) = \lambda\hat{\beta}(\theta, y, X), \text{ for } \lambda \in [0, \infty), \hat{\beta}(1-\theta, \lambda y, X) = \lambda\hat{\beta}(\theta, y, X), \text{ for } \lambda \in (-\infty, 0], \hat{\beta}(\theta, y + X\gamma, X) = \hat{\beta}(\theta, y, X) + \gamma, \text{ for } \gamma \in \Re^K, \hat{\beta}(\theta, y, XA) = A^{-1}\hat{\beta}(\theta, y, X), \text{ for nonsingular } k \times k \text{ matrix } A$$

These properties help in reducing the number of simplex iterations (of any LP algorithm) required for obtaining $\hat{\beta}_\theta$. For example, suppose that $\hat{\beta}_\theta^0$ is a good starting value for $\hat{\beta}_\theta$ (for example, the least-squares estimate from the regression of y on x, or an estimate obtained from only a small subset of the data available). Let $\hat{\beta}_\theta^R$ denote the quantile regression estimate from the θ th quantile regression of $y^R = y - X\hat{\beta}_\theta^0$ on x. Then $\hat{\beta}_\theta \equiv \hat{\beta}_\theta^R + \hat{\beta}_\theta^0$. In many cases it is faster to obtain the two estimates $\hat{\beta}_\theta^R$ and $\hat{\beta}_\theta^0$ than to estimate $\hat{\beta}_\theta$ directly.

2.4 Efficient estimation

The quantile regression estimator described above is not the efficient estimator for β_θ. An efficient estimator can be obtained by solving

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} f_{u_\theta}(0|x)(\theta - 1/2 + 1/2\ \mathrm{sgn}(y_i - x_i'\beta))(y_i - x_i'\beta).$$

That is, each observation is weighted by the conditional density of its error evaluated at zero. This estimation procedure requires the use of an estimate for the unknown density $f_{u_\theta}(0|x)$. Below we provide details about the estimation of the asymptotic covariance matrix, which, in turn, also provides information about possible estimates for $f_{u_\theta}(0|x)$. (For a more complete discussion of this estimator, see Newey and Powell, 1990.)

2.5 Interpretation of the quantile regression

How can the quantile's coefficients be interpreted? Consider the partial derivative of the conditional quantile of *y* with respect to one of the regressors, say *j*, that is, $\partial Q_\theta(y|x) / \partial x_j$. This derivative may be interpreted as the marginal change in the θ th conditional quantile due to a marginal change in the *j*th element of *x*. If *x* contains *K* distinct variables, then this derivative is given simply by $\beta_{\theta j}$, the coefficient on the *j*th variable. It is important to note that one should be cautious not to confuse this result with the location of an individual in the conditional distribution. In general, it need not be the case that an observation that happened to be in the θ th quantile of one conditional distribution will also be at the same quantile if *x* had changed. The above derivative reflects changes in the conditional distribution but has nothing to say about the location of an observation within the conditional distribution.

Note that an estimate for the θ th conditional quantile of *y* given *x* is given by $Q_\theta(y|x) = x'\hat{\beta}_\theta$. Hence, if one were to vary θ between 0 and 1 and estimate a different quantile regression estimate for each θ, one can trace the entire conditional distribution of *y*, conditional on *x*.

3 Large sample properties of $\hat{\beta}_\theta$

We denote the conditional distribution function of u_θ by $F_{u_\theta}(\cdot|x)$ and the corresponding density function by $f_{u_\theta}(\cdot|x)$.

Assumption A.1: The distribution functions $\{F_{u_{\theta i}}(\cdot|x_i)\}$ are absolutely continuous, with continuous density functions $f_{u_{\theta i}}(\cdot|x_i)$ uniformly bounded away from 0 and ∞ at the point 0, for $i = 1, 2, \ldots$

Assumption A.2: There exit positive definite matrices Δ_θ and Λ_0 such that

1. (i) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i' = \Lambda_0$;
2. (ii) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_{u_{\theta i}}(0|x_i) x_i x_i' = \Delta_\theta$; and
3. (iii) $\max_{i=1,\ldots,n} \|x\| / \sqrt{n} \to 0$.

Assumption A.3: The parameter vector β_θ is in the interior of the parameter space $\mathcal{B}_\theta$.

Assumption A.1 requires that the conditional density of $u_{\theta i}$, conditional on x_i, be bounded and that there be no mass point at the conditional θ th quantile at which β_θ is estimated.

Assumptions A.2 and A.3 provide regularity conditions very similar to those used for the usual least-squares estimator. Assumptions A.1 and A.2 are sufficient for establishing that $\hat{\beta}_\theta \to \beta_\theta$ as $n \to \infty$, while Assumption A.3 is needed in addition for establishing the asymptotic normality of $\hat{\beta}_\theta$ in the following theorem.

Theorem 1: Under Assumptions A.1–A.3

1. (i) $\sqrt{n}(\hat{\beta}_\theta - \beta_\theta) \xrightarrow{\mathcal{L}} N(0, \theta(1-\theta)\Delta_\theta^{-1}\Lambda_0\Delta_\theta^{-1})$;
2. (ii) if in addition $f_{u_\theta}(0|x) = f_{u_\theta}(0)$ with probability 1, then $\sqrt{n}(\hat{\beta}_\theta - \beta_\theta) \xrightarrow{\mathcal{L}} N(0, \omega_\theta^2 \Lambda_0^{-1})$, where $\omega_\theta^2 = \theta(1-\theta) f_{u_\theta}^2(0)$.

The result in (i) uses the fact that the (y_i, x_i) are independent, but need not be identically distributed. This is the case when $f_{u_\theta}(\cdot|x)$ depends on *x*, as is the case, for example, with heteroskedasticity. The result in (ii) simplifies the result in (i) when (y_i, x_i') are i.i.d.

3.1 Estimation of the asymptotic covariance matrix

Several estimators for the asymptotic covariance matrix are readily available. Some of the estimators are valid under Theorem 1(i), while others are valid only under the independence assumption of

Theorem 1(ii). In the following, we refer to the former as the *general case*, while the latter is referred to as the *i.i.d. case*. Note that under either cases Λ_0 can be easily estimated by its sample analogue, namely, $\widehat{\Lambda} = n^{-1}\Sigma_{i=1}^{n} x_i x_i'$.

3.1.1 The i.i.d. case

In this case the problem centers around estimating ω_θ^2, or more specifically around estimating $1/f_{u_\theta}(0)$. Let $\hat{u}_{\theta(1)}, \ldots, \hat{u}_{\theta(n)}$ be the ordered residuals from the θ th quantile regression.

Order estimator: Following Siddiqui (1960), an estimator for $1/f_{u_\theta}^2(0)$ is provided by

$$\frac{1}{\hat{f}_{u_\theta}^2(0)} = \frac{\left(\hat{u}_{\theta([n(\theta+h_n)])} - \hat{u}_{\theta([n(\theta+h_n)])}\right)^2}{4h_n},$$

for some bandwidth $h_n = o_p(1)$. Bofinger (1975) provides an optimal choice for the bandwidth, that minimizes the mean squared error, based on the normal approximation for the true $f_{u_\theta}(\cdot)$:

$$h_n = n^{-1/5}\left\{\frac{4.5\,\varphi^4\left(\Phi^{-1}(\theta)\right)}{\left[\left(2\Phi^{-1}(\theta)\right)^2 + 1\right]^2}\right\}^{1/5},$$

where Φ and ϕ denote the distribution function and density function of a standard normal variable, respectively.

Kernel estimator: The density $f_{u_\theta}(0)$ can be estimated directly by

$$\hat{f}_{u_\theta}(0) = (c_n n)^{-1}\sum_{i=1}^{n} \kappa(\hat{u}_{\theta i}),$$

where $\kappa(\cdot)$ is some kernel function $\kappa(\cdot)$ and $c_n = o_p(1)$ is the kernel bandwidth. It can be optimally chosen using a variety of cross-validation methods (for example, least-squares, log likelihood, and so on).

Bootstrap estimator for ω_θ^2: This estimator relies on bootstrapping the residual series $\hat{u}_{\theta i}, i = 1, \ldots, n$. Specifically, one can obtain *B* bootstrap estimates for q_θ, the θ 's quantile of u_θ, say $\hat{q}_{\theta 1}^*, \ldots, \hat{q}_{\theta B}^*$, from *B* bootstrap samples drawn from the empirical distribution $\hat{F}_{u_\theta}$. An estimator for ω_θ^2 is obtained then by

$$\widetilde{\omega}_\theta^2 = \frac{n}{B}\sum_{j=1}^{B} (q_{\theta j}^* - \bar{q}_\theta^*)^2,$$

where $\bar{q}_\theta^* = \frac{1}{B}\Sigma_{j=1}^{B} \hat{q}_{\theta j}^*$.

3.1.2 The general case

There are several alternative estimators for the general case. Here we provide two possible estimators that have been proven accurate in a variety of Monte Carlo studies (for example, Buchinsky, 1995).
Kernel estimator: Powell (1986) considered the following kernel estimator for Δ_θ

$$\hat{\Delta}_\theta = (c_n n)^{-1} \sum_{i=1}^{n} \kappa(\hat{u}_{\theta i}) x_i x_i',$$

where $\kappa(\cdot)$ is some kernel function and $c_n = o_p(1)$ is the kernel bandwidth. Note that the top left-hand element of the matrix $\hat{\Delta}_\theta$ is an estimate of the density $f_{u_\theta}(0)$. Hence, the same cross-validation methods discussed before can be used to optimally choose c_n.
Design matrix bootstrap estimators: There are several alternative ways for employing the bootstrap method of Efron (1979). The most general method is what is termed the *design matrix bootstrapping*, whereby one re-samples from the joint distribution of (y, x). Specifically, let (y_i^*, x_i^*), $i = 1, \ldots, n$, be a randomly drawn sample from the empirical distribution of (x,y), denoted $\hat{F}_{xy}$. Let $\hat{\beta}_\theta^*$ denote the quantile regression estimate based on the bootstrap sample. If we repeat this process B times, then an estimate for $V_\theta \equiv \theta(1-\theta)\Delta_\theta^{-1}\Lambda_0\Delta_\theta^{-1}$ is given by

$$\hat{V}_\theta = \frac{n}{B}\sum_{j=1}^{B} (\hat{\beta}_{\theta j}^* - \bar{\beta}_\theta^*)(\hat{\beta}_{\theta j}^* - \bar{\beta}_\theta^*)',$$

where $\bar{\beta}_\theta^* = \frac{1}{B}\sum_{j=1}^{B}\hat{\beta}_{\theta j}^*$. The estimate $\hat{V}_\theta$ is a consistent estimator for V_θ in the sense that the conditional distribution of $\sqrt{n}(\hat{\beta}_\theta^* - \hat{\beta}_\theta)$, conditional on the data, weakly converges to the unconditional distribution of $\sqrt{n}(\hat{\beta}_\theta - \beta_\theta)$.
One important caveat about bootstrapping is in order. If one already uses the bootstrap method, it can be used more efficiently and effectively, taking advantage of the higher-order refinement properties of the method. For example, one can directly construct confidence intervals, test statistics, and so forth, based on the bootstrap estimates without having to first compute an estimate for V_θ. The number of bootstrap repetitions required for the particular application may be different. Nevertheless, the exact number of repetitions can be computed using the method proposed by Andrews and Buchinsky (2000).

4 Set of quantile regressions

The model presented in (1) considered only the estimation for a single quantile θ. In practice one would like to estimate several quantile regressions at distinct points of the conditional distribution of the dependent variable. This section outlines the estimation of a finite sequence of quantile regressions and provides its asymptotic distribution.

4.1 Estimation and large sample properties

Consider the model given in (1) (dropping the i subscript) for p alternative θ's:

$$y = x'\beta_{\theta_j} + u_{\theta_j} \text{where}$$
$$Q_{\theta_j}(u_{\theta_j}|x) = 0,$$

for $j = 1, \ldots, p$. Without loss of generality assume that $0 < \theta_1 < \theta_2 < \cdots < \theta_p < 1$. Estimating the p quantile regressions amounts to running p separate regressions for θ_1 through θ_p. Let the stacked

vector of β's $\beta_\theta = (\beta'_{\theta_1}, \ldots, \beta'_{\theta_p})'$ denote the population's true parameter vector and let $\hat{\beta}_\theta = (\hat{\beta}'_{\theta_1}, \ldots, \hat{\beta}'_{\theta_p})'$ denote its corresponding estimate.
Theorem 2: under Assumptions A.1–A.3

1. (i) $\sqrt{n}(\hat{\beta}_\theta - \beta_\theta) \xrightarrow{\mathcal{L}} N(0, \Lambda_\theta)$, where $\Lambda_\theta = \{\Lambda_{\theta_{jk}}\}_{j,k=1,\ldots,p}$, and
2. (ii) $\Lambda_{\theta_{jk}} = (\min\{\theta_j, \theta_k\} - \theta_j\theta_k)\Delta_{\theta_j}^{-1}\Lambda_0\Delta_{\theta_k}^{-1}$; if in addition $f_{u_\theta}(0|x) = f_{u_\theta}(0)$ for $j = 1, \ldots, p$ with probability 1, then $\sqrt{n}(\hat{\beta}_\theta - \beta_\theta) \xrightarrow{\mathcal{L}} N(0, \Lambda_\theta)$, where $\Lambda_\theta = \Omega_\theta \otimes \Lambda_0^{-1}$ and $\Omega_{\theta_{jk}} = [\min\{\theta_j, \theta_k\} - \theta_j\theta_k] / \left[f_{u_{\theta_j}}(0) f_{u_{\theta_k}}(0)\right]$

4.2 Crossing of quantiles

Note that the estimated conditional quantiles, conditional on *x*, are given by $x'\hat{\beta}_{\theta_1}, \ldots, x'\hat{\beta}_{\theta_p}$. Since the estimates $\hat{\beta}_{\theta_j}$ $(j = 1, \ldots, p)$ for the *p* quantiles are obtained from separate quantile regressions, it is possible that for some vector x_0, $x_0'\hat{\beta}_{\theta_j} > x_0'\hat{\beta}_{\theta_k}$ even though $\theta_j < \theta_k$, that is, conditional quantiles may cross each other. This may not be of any practical consequence, since there may not be such a vector within the relevant range of plausible *x*'s. Nevertheless, in any empirical application these potential crossing need to be examined.

4.3 Testing for equality of slope coefficients

Under the i.i.d. assumption the *p* coefficient vectors $\beta_{\theta_1}, \ldots, \beta_{\theta_p}$ should be the same, except for the intercept coefficients. There are a number of ways for testing the null hypothesis of i.i.d. errors. Only two testing procedures are provided here. For other alternative methods see Koenker (2005).

4.3.1 Wald-type testing

This testing procedure is based on the optimal minimum distance (MD) estimator under the null hypothesis. Denote the parameter vector under the null by β_θ^R and note that $\beta_\theta^R = (\beta_{\theta_1 1}, \ldots, \beta_{\theta_p 1}, \beta_2, \ldots, \beta_k)'$ is a $(p + K - 1) \times 1$, with *p* distinct intercepts $\beta_{\theta_1 1}, \ldots, \beta_{\theta_p 1}$, and $k - 1$ common slope parameters $\beta_2, \ldots, \beta_k$.
An optimal estimate for the restricted coefficient vector β_θ^R is defined by

$$\hat{\beta}_\theta^R = \arg\min_{\beta^R} (\hat{\beta}_\theta - R\beta^R)'\hat{V}_\theta^{-1}(\hat{\beta}_\theta - R\beta^R),$$

where $\hat{V}_\theta$ is a consistent estimate for the covariance matrix of $\hat{\beta}_\theta$, the unrestricted parameter estimate from the *p* quantile regressions, estimated under the null (that is, under Theorem 2(ii). The matrix *R* is simply a $(p + K - 1) \times p \cdot k$ restriction matrix which imposes the restrictions implied by the i.i.d. assumption. A test statistic for equality of the slope coefficients is provided then by

$$W_n = n(\hat{\beta}_\theta - R\hat{\beta}_\theta^R)'\hat{V}_\theta^{-1}(\hat{\beta}_\theta - R\hat{\beta}_\theta^R).$$

Under the null hypothesis $W_n \xrightarrow{D} \chi^2(pK - p - K + 1)$ as $n \to \infty$. So, the null hypothesis is rejected if $W_n > \chi^2_{1-\alpha}(pK - p - K + 1)$, where $\chi^2_{1-\alpha}(m)$ denotes the $1-\alpha$ quantile of a χ^2-distribution with *m* degrees of freedom.

4.3.2 GMM-type testing

An alternative testing procedure can be applied using Hansen's (1982) GMM method. Define a moment function $\psi(x, y, \beta)$ by stacking the *p* individual moment functions as defined in (3). While this moment function is a *pk*×1 vector, under the null there are only $p + K - 1$ parameters to be estimated. Hansen's GMM framework provides an estimator for β_θ^R, say $\hat{\beta}_\theta^R$, defined by

$$\hat{\beta}_\theta^R = \arg\min_b \left(\frac{1}{n}\sum_{i=1}^n \psi(x_i, y_i, b)\right)' A^{-1} \left(\frac{1}{n}\sum_{i=1}^n \psi(x_i, y_i, b)\right), \tag{4}$$

An efficient estimator can be obtained if *A* is chosen so that $A \xrightarrow{p} E[\psi(x, y, \beta_\theta)\psi(x, y, \beta_\theta)']$ as $n \to \infty$. This framework provides us with a straightforward testing procedure. Under the null hypothesis

$$n\left(\frac{1}{n}\sum_{i=1}^n \psi\left(x_i, y_i, \hat{\beta}_\theta^R\right)\right)' A^{-1} \left(\frac{1}{n}\sum_{i=1}^n \psi\left(x_i, y_i, \hat{\beta}_\theta^R\right)\right) \xrightarrow{D} \chi^2(pK - p - K + 1),$$

as $n \to \infty$.

Note that, because of the linearity of the conditional quantiles, the GMM testing provides a test statistics which is (asymptotically) equivalent to that provided by the MD testing.

5 Censored quantile regression

An important extension to the quantile regression model was suggested by Powell (1984; 1986). This extension considers the case in which some of the observations are censored. This model is essentially a semiparametric extension to the well known 'Tobit' model and can be written as

$$y_i = \min\left\{y_i^0, x_i'\beta_\theta + u_{\theta i}\right\},$$

for $i = 1, \ldots, n$, where y_i^0 is the (known) top coding value of y_i in the sample, for $i = 1, \ldots, n$. (For simplicity of presentation it will be assumed that $y_i^0 = y^0$ for all $i = 1, \ldots, n$.)

This model can be written as a latent variable model. That is, we have $y_i^* = x_i'\beta_\theta + u_{\theta i}$, where $Q_\theta(u_{\theta i}|x_i) = 0$ and $y_i = y_i^* I(y_i^* \le y^0)$. It is easy to see that the observed conditional θ th quantile of y_i, conditional on x_i, is given by $Q_\theta(y_i|x_i) = \min\left\{y^0, x_i'\beta_\theta\right\}$.

Hence, Powell suggested the following estimator for β_θ

$$\hat{\beta}_\theta = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n \rho_\theta\left(y_i - \min\left\{y^0, x_i'\beta\right\}\right), \tag{5}$$

where $\rho_\theta(\lambda)$ is the same check function as defined above. Note that in order to obtain a consistent estimator of β_θ it is necessary that $x_i'\beta_\theta < y^0$ for at least a fraction of the sample. Intuitively, the larger the fraction, the more precise the estimator will be.

Powell (1986) showed that, under certain regularity conditions, similar to those established by Huber (1967), the estimator is asymptotically normal. That is, $\sqrt{n}(\hat{\beta}_\theta - \beta_\theta) \xrightarrow{D} N(0, V_\theta^C)$ as $n \to \infty$, where

$$V_\theta^C = \theta(1-\theta)\Delta_{C\theta}^{-1}\Lambda_{C\theta}\Delta_{C\theta}^{-1}, \Delta_{C\theta} = E\left[f_{u_\theta}(0|x)I(x'\beta_\theta \le y^0)xx'\right], \text{and} \Lambda_{C\theta} = E\left[I(x'\beta_\theta < y^0)xx'\right].$$

As in the basic quantile regression model, if $f_{u_\theta}(0|x) = f_{u_\theta}(0)$ with probability 1, then V_θ^C simplifies to $V_\theta^C = \omega_\theta^2\Lambda_{C\theta}^{-1}$, where $\omega_\theta^2 = \theta(1-\theta)f_{u_\theta}^2(0)$.
It is important to note that if $x_i'\hat{\beta}_\theta \le y^0$ for all observations, then the censored quantile regression estimate coincides with the basic quantile regression.
The simple intuition for this estimation procedure is that β_θ can be estimated only from that part of the sample for which it is observed, that is, for that fraction of the sample for which $y = y^* = x'\beta_\theta + u_\theta \le y^0$. As a result, we note that the asymptotic covariance is 'adjusted' for that fact. That is, the term $I(x'\beta_\theta \le y^0)$ is included in both $\Delta_{C\theta}$ and $\Delta_{C\theta}$.
A considerable drawback of the censored quantile regression model is that it does not have the attractive LP representation and the objective function is not globally convex in β.

6 Concluding remarks

The main goal of this article is to provide the basic structure of the quantile regression model. Versions of this model have been widely used in the empirical literature in a variety of situations not covered by this article. Furthermore, there have been substantial advancements in the theoretical literature as well. This literature includes quantile regression for nonlinear models, time-series models, and others. There are also a number of empirical studies that have used quantile regression extensively, in a variety of data configurations and economic contexts. For a brilliant in-depth exposition of a wide variety of topics related to quantile regression, interested readers should refer to Koenker (2005).

See Also

- linear programming
- semiparametric estimation

Bibliography

Andrews, D. and Buchinsky, M. 2000. A three-step method for choosing the number of bootstrap repetitions. *Econometrica* 68, 23–51.

Bofinger, E. 1975. Estimation of density function using order statistics. *Australian Journal of Statistics* 17, 1–7.

Buchinsky, M. 1994. Changes in the U.S. wage structure 1963–1987: application of quantile regression. *Econometrica* 62, 405–58.

Buchinsky, M. 1995. Estimating the asymptotic covariance matrix for quantile regression models: a Monte Carlo study. *Journal of Econometrics* 68, 303–38.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1–26.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Huber, P. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press.

Koenker, R. 2005. *Quantile Regression*. Econometric Society Monograph. New York: Cambridge University Press.

Koenker, R. and Bassett, G. 1978a. The asymptotic distribution of the least absolute error estimator. *Journal of the American Statistical Association* 73, 618–22.

Koenker, R. and Bassett, G. 1978b. Regression quantiles. *Econometrica* 46, 33–50.

Newey, W. and Powell, J. 1990. Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory* 6, 295–317.

Portnoy, S. and Koenker, R. 1989. Adaptive *L*-estimation for linear models. *Annals of Statistics* 17, 362–81.

Powell, J. 1984. Least absolute deviation estimation for the censored regression model. *Journal of Econometrics* 25, 303–25.

Powell, J. 1986. Censored regression quantiles. *Journal of Econometrics* 32, 143–55.

Siddiqui, M. 1960. Distribution of quantile from a bivariate population. *Journal of Research of the National Bureau of Standards* 64, 145–50.

How to cite this article

Buchinsky, Moshe. "quantile regression." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_Q000015> doi:10.1057/9780230226203.1373

The New Palgrave Dictionary of Economics Online

quantity theory of money

Milton Friedman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

After formally setting out the quantity theory of money, including the distinction between the *nominal* quantity of money and the *real* quantity of money, and various quantity equations, this article considers the Keynesian challenge to the theory (which seemed vindicated during the economically successful 1950s and 1960s) and the revival of belief in the quantity theory in the 1970s as rapid monetary growth was accompanied by stagflation and rising interest rates. It deals with the natural rate hypothesis and the theory of rational expectations, surveys the empirical evidence, and ends with a consideration of policy implications.

Keywords

American Economic Association; anticipated inflation; arbitrage; Bretton Woods System; buffer stocks; cash-balance version of the quantity equation; commodity standards; credit; deflation; demand for money; demand–supply analysis; deposit–currency ratio; deposit–reserve ratio; easy money; excess balances; expectations; export subsidies; fiat money; fiscal policy; Fisher, I.; foreign exchange controls; free banking; Friedman, M.; full employment; Great Depression; high-powered money; human vs. non-human wealth; Hume, D.; hyperinflation; import controls; income velocity; interest rate; Keynes, J. M.; law of one price; liquidity; liquidity preference; Lucas, R.; managed exchange-rate floating; Mill, J.S.; Modigliani, F.; monetary growth targeting; monetary illusion; monetary policy rules; monetary transmission mechanism; money; money multiplier; money supply; moneyness; Muth, J. F.; natural rate hypothesis; natural rate of unemployment; neutrality hypothesis about stabilization policy; nominal interest rate; nominal quantity of money; Phillips curve; price control; price indices; price revolution; price rigidity; quantity equations; quantity theory of money; rate of return on money; rational expectations; real interest rate; real quantity of money; Ricardo, D.; Sargent, T. J.; specie-flow mechanism; speculative motive; stagflation; static vs. long-run equilibrium; structural unemployment; Thornton, H.; Tobin, J.; transactions and precautionary motives; transactions velocity; transactions

version of the quantity equation; variability of inflation; velocity of circulation; wage rigidity

Article

Lowness of interest is generally ascribed to plenty of money. But … augmentation [in the quantity of money] has no other effect than to heighten the price of labour and commodities … In the progress toward these changes, the augmentation may have some influence, by exciting industry, but after the prices are settled … it has no manner of influence.

> [T]hough the high price of commodities be a necessary consequence of the increase of gold and silver, yet it follows not immediately upon that increase; but some time is required before the money circulates through the whole state.… In my opinion, it is only in this interval of intermediate situation, between the acquisition of money and rise of prices, that the increasing quantity of gold and silver is favourable to industry.… [W]e may conclude that it is of no manner of consequence, with regard to the domestic happiness of a state, whether money be in greater or less quantity. The good policy of the magistrate consists only in keeping it, if possible, still increasing …
> (David Hume, 1752).

In this survey, we shall first present a formal statement of the quantity theory, then consider the Keynesian challenge to the quantity theory, recent developments, and some empirical evidence. We shall conclude with a discussion of policy implications, giving special attention to the likely implications of the worldwide fiat money standard that has prevailed since 1971.

1 The formal theory

(a) Nominal versus real quantity of money

Implicit in the quotation from Hume, and central to all later versions of the quantity theory, is a distinction between the *nominal* quantity of money and the *real* quantity of money. The nominal quantity of money is the quantity expressed in whatever units are used to designate money – talents, shekels, pounds, francs, lira, drachmas, dollars, and so on. The real quantity of money is the quantity expressed in terms of the volume of goods and services the money will purchase.

There is no unique way to express either the nominal or the real quantity of money. With respect to the nominal quantity of money, the issue is what assets to include – whether only currency and coins, or also claims on financial institutions; and, if such claims are included, which ones should be, only deposits transferable by cheque, or also other categories of claims which in practice are close substitutes for deposits transferable by cheque. More recently, economists have been experimenting with the theoretically attractive idea of defining money not as the simple sum of various categories of claims but as a weighted aggregate of such claims, the weights being determined by one or another concept of the ‘moneyness’ of the various claims.

Despite continual controversy over the definition of ‘money’, and the lack of unanimity about relevant

theoretical criteria, in practice, monetary economists have generally displayed wide agreement about the most useful counterpart, or set of counterparts, to the concept of ‘money’ at a particular times and places (Friedman and Schwartz, 1970, pp. 89–197; Barnett, Offenbacher and Spindt, 1984; Spindt, 1985).

The real quantity of money obviously depends on the particular definition chosen for the nominal quantity. In addition, for each such definition, it can vary according to the set of goods and services in terms of which it is expressed. One way to calculate the real quantity of money is by dividing the nominal quantity of money by a price index. The real quantity is then expressed in terms of the standard basket whose components are used as weights in computing the price index – generally, the basket purchased by some representative group in a base year.

A different way to express the real quantity of money is in terms of the time duration of the flow of goods and services the money could purchase. For a household, for example, the real quantity of money can be expressed in terms of the number of weeks of the household's average level of consumption its money balances could finance or, alternatively, in terms of the number of weeks of its average income to which its money balances are equal. For a business enterprise, the real quantity of money it holds can be expressed in terms of the number of weeks of its average purchases, or of its average sales, or of its average expenditures on final productive services (net value added) to which its money balances are equal. For the community as a whole, the real quantity of money can be expressed in terms of the number of weeks of aggregate transactions of the community, or aggregate net output of the community, to which its money balances are equal.

The reciprocal of any of this latter class of measures of the real quantity of money is a velocity of circulation for the corresponding unit or group of units. For example, the ratio of the annual transactions of the community to its stock of money is the ‘transactions velocity of circulation of money’, since it gives the number of times the stock of money would have to ‘turn over’ in a year to accomplish all transactions. Similarly, the ratio of annual income to the stock of money is termed ‘income velocity’. In every case, the real quantity of money is calculated at the set of prices prevailing at the date to which the calculation refers. These prices are the bridge between the nominal and the real quantity of money.

The quantity theory of money takes for granted, first, that the real quantity rather than the nominal quantity of money is what ultimately matters to holders of money and, second, that in any given circumstances people wish to hold a fairly definite real quantity of money. Starting from a situation in which the nominal quantity that people hold at a particular moment of time happens to correspond at current prices to the real quantity that they wish to hold, suppose that the quantity of money unexpectedly increases so that individuals have larger cash balances than they wish to hold. They will then seek to dispose of what they regard as their excess money balances by paying out a larger sum for the purchase of securities, goods, and services, for the repayment of debts, and as gifts, than they are receiving from the corresponding sources. However, they cannot as a group succeed. One man's spending is another man's receipts. One man can reduce his nominal money balances only by persuading someone else to increase his. The community as a whole cannot in general spend more than it receives; it is playing a game of musical chairs.

The attempt to dispose of excess balances will nonetheless have important effects. If prices and incomes are free to change, the attempt to spend more will raise total spending and receipts, expressed in nominal units, which will lead to a bidding up of prices and perhaps also to an increase in output. If prices are fixed by custom or by government edict, the attempt to spend more will either be matched by an increase in goods and services or produce ‘shortages’ and ‘queues’. These in turn will raise the effective price

and are likely sooner or later to force changes in customary or official prices.

The initial excess of nominal balances will therefore tend to be eliminated, even though there is no change in the nominal quantity of money, by either a reduction in the real quantity available to hold through price rises or an increase in the real quantity desired through output increases. And conversely for an initial deficiency of nominal balances.

Changes in prices and nominal income can be produced either by changes in the real balances that people wish to hold or by changes in the nominal balances available for them to hold. Indeed, it is a tautology, summarized in the famous quantity equations, that all changes in nominal income can be attributed to one or the other – just as a change in the price of any good can always be attributed to a change in either demand or supply. The quantity theory is not, however, this tautology. On an analytical level, it has long been an analysis of the factors determining the quantity of money that the community wishes to hold; on an empirical level, it has increasingly become the generalization that changes in desired real balances (in the demand for money) tend to proceed slowly and gradually or to be the result of events set in train by prior changes in supply, whereas, in contrast, substantial changes in the supply of nominal balances can and frequently do occur independently of any changes in demand. The conclusion is that substantial changes in prices or nominal income are almost always the result of changes in the nominal supply of money.

(b) Quantity equations

Attempts to formulate mathematically the relations just presented verbally date back several centuries (Humphrey, 1984). They consist of creating identities equating a flow of money payments to a flow of exchanges of goods or services. The resulting quantity equations have proved a useful analytical device and have taken different forms as quantity theorists have stressed different variables.

The transactions form of the quantity equation

The most famous version of the quantity equation is doubtless the transactions version formulated by Simon Newcomb (1885) and popularized by Irving Fisher (1911):

$$MV = PT, \tag{1}$$

or

$$MV + M'V' = PT. \tag{2}$$

In this version the elementary event is a transaction – an exchange in which one economic actor transfers goods or services or securities to another actor and receives a transfer of money in return. The right-hand side of the equations corresponds to the transfer of goods, services, or securities; the left-hand side, to the matching transfer of money.

Each transfer of goods, services or securities is regarded as the product of a price and quantity; wage per week times number of weeks, price of a good times number of units of the good, dividend per share times number of shares, price per share times number of shares, and so on. The right-hand side of equations (1) and (2) is the aggregate of such payments during some interval, with P a suitably chosen *average* of the prices and T a suitably chosen *aggregate* of the quantities during that interval, so that PT is the total nominal value of the payments during the interval in question. The units of P are dollars (or other monetary unit) per unit of quantity; the units of T are number of unit quantities per period of time.

We can convert the equation from an expression applying to an *interval* of time to one applying to a *point* in time by the usual limiting process of letting the interval for which we aggregate payments approach zero, and expressing T not as an aggregate but as a rate of flow. The magnitude T then has the dimension of quantity per unit time; the product of P and T, of dollars (or other monetary unit) per unit time.

T is clearly a rather special index of quantities: it includes service flows (man-hours, dwelling-years, kilowatt-hours) and also physical capital items yielding such flows (houses, electric-generating plants) and securities representing both physical capital items and such intangible capital items as ‘goodwill’. Since each capital item or security is treated as if it disappeared from economic circulation once it is transferred, any such item that is transferred more than once in the period in question is implicitly weighted by the number of times it enters into transactions (its ‘velocity of circulation’, in strict analogy with the ‘velocity of circulation’ of money). Similarly, P is a rather special price index.

The monetary transfer analysed on the left-hand side of equations (1) and (2) is treated very differently. The money that changes hands is treated as retaining its identity, and all money, whether used in transactions during the time interval in question or not, is explicitly accounted for. Money is treated as a stock, not as a flow or a mixture of a flow and a stock. For a single transaction, the breakdown into M and V is trivial: the cash that is transferred is turned over once, or $V = 1$. For all transactions during an interval of time, we can, in principle, classify the existing stock of monetary units according as each monetary unit entered into 0, 1, 2,… transactions – that is, according as the monetary unit ‘turned over’ 0, 1, 2,… times. The weighted average of these numbers of turnover, weighted by the number of dollars that turned over that number of times, is the conceptual equivalent of V. The dimensions of M are dollars (or other monetary unit); of V, number of turnovers per unit time; so, of the product, dollars per unit time.

Equation (2) differs from equation (1) by dividing payments into two categories: those effected by the transfer of hand-to-hand currency (including coin) and those effected by the transfer of deposits. In equation (2) M stands for the volume of currency and V for the velocity of currency, M' for the volume of deposits, and V' for the velocity of deposits.

One reason for the emphasis on this particular division was the persistent dispute about whether the term *money* should include only currency or deposits as well. Another reason was the direct availability of data on $M'V'$ from bank records of clearings or of debits to deposit accounts. These data make it possible to calculate V' in a way that is not possible for V.

Equations (1) and (2), like the other quantity equations we shall discuss, are intended to be identities – a

special application of double-entry bookkeeping, with each transaction simultaneously recorded on both sides of the equation. However, as with the national income identities with which we are all familiar, when the two sides, or the separate elements on the two sides, are estimated from independent sources of data, many differences between them emerge. This statistical defect has been less obvious for the quantity equations than for the national income identities – with their standard entry ‘statistical discrepancy’ – because of the difficulty of calculating V directly. As a result, V in equation (1) and V and V' in equation (2) have generally been calculated as the numbers having the property that they render the equations correct. These calculated numbers therefore embody the whole of the counterpart to the ‘statistical discrepancy’.

Just as the left-hand side of equation (1) can be divided into several components, as in equation (2), so also can the right-hand side. The emphasis on transactions reflected in this version of the quantity equation suggests dividing total transactions into categories of payments for which payment periods or practices differ: for example, into capital transactions, purchases of final goods and services, purchases of intermediate goods, and payments for the use of resources, perhaps separated into wage and salary payments and other payments. The observed value of V might well depend on the distribution of total payments among categories. Alternatively, if the quantity equation is interpreted not as an identity but as a functional relation expressing desired velocity as a function of other variables, the distribution of payments may well be an important set of variables.

The income form of the quantity equation

Despite the large amount of empirical work done on the transactions equations, notably by Irving Fisher (1911, pp. 280–318; 1919, pp. 407–9) and Carl Snyder (1934, pp. 278–91), the ambiguities of the concepts of ‘transactions’ and the ‘general price level’ – particularly those arising from the mixture of current and capital transactions – have never been satisfactorily resolved. More recently, national or social accounting has stressed income transactions rather than gross transactions and has explicitly if not wholly satisfactorily dealt with the conceptual and statistical problems involved in distinguishing between changes in prices and changes in quantities. As a result, since at least the work of James Angell (1936), monetary economists have tended to express the quantity equation in terms of income transactions rather than gross transactions. Let Y = nominal income, P=the price index implicit in estimating national income at constant prices, N=the number of persons in the population, y=per capita national income in constant prices, and $y'=Ny$=national income at constant prices, so that

$$Y = PNy = Py'. \tag{3}$$

Let M represent, as before, the stock of money; but define V as the average number of times per unit time that the money stock is used in making *income* transactions (that is, payment for final productive services or, alternatively, for final goods and services) rather than all transactions. We can then write the

quantity equation in income form as

$$MV = PNy = Py'. \tag{4}$$

or, if we desire to distinguish currency from deposit transactions, as

$$MV + M'V' = PNy. \tag{5}$$

Although the symbols P, V, and V' are used both in equations (4) and (5) and in equations (1) and (2), they stand for different concepts in each pair of equations. (In practice, gross national product often replaces national income in calculating velocity even though the logic underlying the equation calls for national income. The reason is the widespread belief that estimates of GNP are subject to less statistical error than estimates of national income.)

In the transactions version of the quantity equation, each intermediate transaction – that is, purchase by one enterprise from another – is included at the total value of the transaction, so that the value of wheat, for example, is included once when it is sold by the farmer to the mill, a second time when the mill sells flour to the baker, a third time when the baker sells bread to the grocer, a fourth time when the grocer sells bread to the consumer. In the income version, only the net value added by each of these transactions is included. To put it differently in the transactions version, the elementary event is an isolated exchange of a physical item for money – an actual, clearly observable event. In the income version, the elementary event is a hypothetical event that can be inferred but is not directly observable. It is a complete series of transactions involving the exchange of productive services for final goods, via a sequence of money payments, with all the intermediate transactions in this income circuit netted out. The total value of all transactions is therefore a multiple of the value of income transactions only.

For a given flow of productive services or, alternatively, of final products (two of the multiple faces of income), the volume of transactions will be affected by vertical integration or disintegration of enterprises, which reduces or increases the number of transactions involved in a single income circuit, and by technological changes that lengthen or shorten the process of transforming productive services into final products. The volume of income will not be thus affected.

Similarly, the transactions version includes the purchase of an existing asset – a house or a piece of land or a share of equity stock – precisely on a par with an intermediate or final transaction. The income version excludes such transactions completely.

Are these differences an advantage or disadvantage of the income version? That clearly depends on what it is that determines the amount of money people want to hold. Do changes of the kind considered in the preceding paragraphs, changes that alter the ratio of intermediate and capital transactions to income, also

alter in the same direction and by the same proportion the amount of money people want to hold? Or do they tend to leave this amount unaltered? Or do they have a more complex effect?

The transactions and income versions of the quantity theory involve very different conceptions of the role of money. For the transactions version, the most important thing about money is that it is transferred. For the income version, the most important thing is that it is held. This difference is even more obvious from the Cambridge cash-balance version of the quantity equation (Pigou, 1917). Indeed, the income version can perhaps best be regarded as a way station between the Fisher and the Cambridge version.

Cambridge cash-balance approach

The essential feature of a money economy is that an individual who has something to exchange need not seek out the double coincidence – someone who both wants what he has and offers in exchange what he wants. He need only find someone who wants what he has, sell it to him for general purchasing power, and then find someone who has what he wants and buy it with general purchasing power.

For the act of purchase to be separated from the act of sale, there must be something that everybody will accept in exchange as ‘general purchasing power’ – this aspect of money is emphasized in the transactions approach. But also there must be something that can serve as a temporary abode of purchasing power in the interim between sale and purchase. This aspect of money is emphasized in the cash-balance approach.

How much money will people or enterprises want to hold on the average as a temporary abode of purchasing power? As a first approximation, it has generally been supposed that the amount bears some relation to income, on the assumption that income affects the volume of potential purchases for which the individual or enterprise wishes to hold cash balances. We can therefore write

$$M = kPNy = kPy', \tag{6}$$

where M, N, P, y, and y' are defined as in equation (4) and k is the ratio of money stock to income – either the observed ratio so calculated as to make equation (6) an identity or the ‘desired’ ratio so that M is the ‘desired’ amount of money, which need not be equal to the actual amount. In either case, k is numerically equal to the reciprocal of the V in equation (4), the V being interpreted in one case as measured velocity and in the other as desired velocity.

Although equation (6) is simply a mathematical transformation of equation (4), it brings out sharply the difference between the aspect of money stressed by the transactions approach and that stressed by the cash-balance approach. This difference makes different definitions of money seem natural and leads to placing emphasis on different variables and analytical techniques.

The transactions approach makes it natural to define money in terms of whatever serves as the medium of exchange in discharging obligations. The cash-balance approach makes it seem entirely appropriate to

include in addition such temporary abodes of purchasing power as demand and time deposits not transferable by check, although it clearly does not require their inclusion (Friedman and Schwartz, 1970, ch. 3).

Similarly, the transactions approach leads to emphasis on the mechanical aspect of the payments process; payments practices, financial and economic arrangements for effecting transactions, the speed of communication and transportation, and so on (Baumol, 1952; Tobin, 1956; Miller and Orr, 1966, 1968). The cash-balance approach, on the other hand, leads to emphasis on variables affecting the usefulness of money as an asset: the costs and returns from holding money instead of other assets, the uncertainty of the future, and so on (Friedman, 1956; Tobin, 1958).

Of course, neither approach enforces the exclusion of the variables stressed by the other. Portfolio considerations enter into the costs of effecting transactions and hence affect the most efficient payment arrangements; mechanical considerations enter into the returns from holding cash and hence affect the usefulness of cash in a portfolio.

Finally, with regard to analytical techniques, the cash-balance approach fits in much more readily with the general Marshallian demand-supply apparatus than does the transactions approach. Equation (6) can be regarded as a demand function for money, with *P*, *N*, and *y* on the right-hand side being three of the variables on which the quantity of money demanded depends and *k* symbolizing all the other variables, so that *k* is to be regarded not as a numerical constant but as itself a function of still other variables. For completion, the analysis requires another equation showing the supply of money as a function of these and other variables. The price level or the level of nominal income is then the resultant of the interaction of the demand and supply functions.

Levels versus rates of change

The several versions of the quantity equations have all been stated in terms of the levels of the variables involved. For the analysis of monetary change it is often more useful to express them in terms of rates of change. For example, take the logarithm of both sides of equation (4) and differentiate with respect to time. The result is

$$\frac{1}{M}\frac{dM}{dt} + \frac{1}{V}\frac{dV}{dt} = \frac{1}{P}\frac{dP}{dt} + \frac{1}{y'}\frac{dy'}{dt} \tag{7}$$

or, in simpler notation,

$$g_M + g_V = g_P + g_{y'} = g_{Y'}, \tag{8}$$

where g stands for the percentage rate of change (continuously compounded) of the variable denoted by its subscript. The same equation is implied by equation (6), with g_V replaced by $-g_k$.

The rate of change equations serve two very different purposes. First, they make explicit an important difference between a once-for-all change in the level of the quantity of money and a change in the rate of change of the quantity of money. The former is equivalent simply to a change of units – to substituting cents for dollars or pence for pounds – and hence, as is implicit in equations (4) and (6), would not be presumed to have any effect on real quantities, on neither V (nor k) nor y', but simply an offsetting effect on the price level, P. A change in the rate of change of money is a very different thing. It will tend, according to equations (7) and (8), to be accompanied by a change in the rate of inflation (g_P) which, as pointed out in section d below, affects the cost of holding money, and hence the desired real quantity of money. Such a change will therefore affect real quantities, V and g_V, y' and $g_{y'}$, as well as nominal and real interest rates.

The second purpose served by the rate of change equations is to make explicit the role of time, and thereby to facilitate the study of the effect of monetary change on the temporal pattern of response of the several variables involved. In recent decades, economists have devoted increasing attention to the short-term pattern of economic change, which has enhanced the importance of the rate of change versions of the quantity equations.

(c) The supply of money

The quantity theory in its cash-balance version suggests organizing an analysis of monetary phenomena in terms of (1) the conditions determining supply (this section); (2) the conditions determining demand (section d below); and (3) the reconciliation of demand with supply (section e below).

The factors determining the nominal supply of money available to be held depend critically on the monetary system. For systems like those that have prevailed in most major countries during the past two centuries, they can usefully be analysed under three main headings termed the proximate determinants of the quantity of money: (1) the amount of high-powered money – specie plus notes or deposit liabilities issued by the monetary authorities and used either as currency or as reserves by banks; (2) the ratio of bank deposits to bank holdings of high-powered money; and (3) the ratio of the public's deposits to its currency holdings (Friedman and Schwartz, 1963b, pp. 776–98; Cagan, 1965; Burger, 1971; Black, 1975).

It is an identity that

$$M = H \cdot \frac{\frac{D}{R}\left(1 + \frac{D}{C}\right)}{\frac{D}{R} + \frac{D}{C}}, \tag{9}$$

where H=high-powered money; D=deposits; R=bank reserves; C=currency in the hands of the public so that (D/R) is the deposit–reserve ratio; and (D/C) is the deposit–currency ratio. The fraction on the right-hand side of (9), i.e., the ratio of M to H, is termed the money multiplier, often a convenient summary of the effect of the two deposit ratios. The determinants are called proximate because their values are in turn determined by much more basic variables. Moreover, the same labels can refer to very different contents.

High-powered money is the clearest example. Until some time in the 18th or 19th century, the exact date varying from country to country, it consisted only of specie or its equivalent: gold, or silver, or cowrie shells, or any of a wide variety of commodities. Thereafter, until 1971, with some significant if temporary exceptions, it consisted of a mixture of specie and of government notes or deposit liabilities. The government notes and liabilities generally were themselves promises to pay specified amounts of specie on demand, though this promise weakened after World War I, when many countries promised to pay either specie or foreign currency. During the Bretton Woods periods after World War II, only the USA was obligated to pay gold, and only to foreign monetary agencies, not to individuals or other non-governmental entities; other countries obligated themselves to pay dollars.

Since 1971, the situation has been radically different. In every major country, high-powered money consists solely of fiat money – pieces of paper issued by the government and inscribed with the legend ‘one dollar’ or ‘one pound’ and the message ‘legal tender for all debts public and private’; or book entries, labelled deposits, consisting of promises to pay such pieces of paper. Such a worldwide fiat (or irredeemable paper) standard has no precedent in history. The ‘gold’ central banks still record as an asset on their books is simply the grin of a Cheshire cat that has disappeared.

Under an international commodity standard, the total quantity of high-powered money in any one country – so long as it remains on the standard – is determined by the balance of payments. The division of high-powered money between physical specie and the fiduciary component of government-issued promises to pay is determined by the policies of the monetary authorities. For the world as a whole, the total quantity of high-powered money is determined both by the policies of the various monetary authorities and the physical conditions of supply of specie. The latter provide a physical anchor for the quantity of money and hence ultimately for the price level.

Under the current international fiat standard, the quantity of high-powered money is determined solely by the monetary authorities, consisting in most countries of a central bank plus the fiscal authorities. What happens to the quantity of high-powered money depends on their objectives, on the institutional and political arrangements under which they operate, and the operating procedures they adopt. These are likely to vary considerably from country to country. Some countries (e.g., Hong Kong, Panama) have chosen to link their currencies rigidly to some other currency by pegging the exchange rate. For them, the amount of high-powered money is determined in the same way as under an international commodity standard – by the balance of payments.

The current system is so new that it must be regarded as in a state of transition. Some substitute is almost sure to emerge to replace the supply of specie as a long-term anchor for the price level, but it is not yet clear what that substitute will be (see section 5 below).

The deposit–reserve ratio is determined by the banking system subject to any requirements that are imposed by law or the monetary authorities. In addition to any such requirements, it depends on such factors as the risk of calls for conversion of bank deposits to high-powered money; the cost of acquiring

additional high-powered money in case of need; and the returns from loans and investments, that is, the structure of interest rates.

The deposit–currency ratio is determined by the public. It depends on the relative usefulness to holders of money of deposits and currency and the relative cost of holding the one or the other. The relative cost in turn depends on the rates of interest received on deposits, which may be subject to controls imposed by law or the monetary authorities.

These factors determine the *nominal*, but not the *real*, quantity of money. The real quantity of money is determined by the interaction between the *nominal* quantity supplied and the *real* quantity demanded. In the process, changes in demand for real balances have feedback effects on the variables determining the nominal quantity supplied, and changes in nominal supply have feedback effects on the variables determining the real quantity demanded. Quantity theorists have generally concluded that these feedback effects are relatively minor, so that the *nominal* supply can generally be regarded as determined by a set of variables distinct from those that affect the *real* quantity demanded. In this sense, the nominal quantity can be regarded as determined primarily by supply, the real quantity, primarily demand.

Instead of expressing the nominal supply in terms of the identity (9), it can also be expressed as a function of the variables that are regarded as affecting *H*, *D/R*, and *D/C*, such as the rate of inflation, interest rates, nominal income, the extent of uncertainty, perhaps also the variables that are regarded as determining the decisions of the monetary authorities. Such a supply function is frequently written as

$$M^S = h(R, Y, \ldots), \tag{10}$$

where *R* is an interest rate or set of interest rates, *Y* is nominal income, and the dots stand for other variables that are regarded as relevant.

(d) The demand for money

The cash-balance version of the quantity theory, by stressing the role of money as an asset, suggests treating the demand for money as part of capital or wealth theory, concerned with the composition of the balance sheet or portfolio of assets.

From this point of view, it is important to distinguish between ultimate wealth holders, to whom money is one form in which they choose to hold their wealth, and enterprises, to whom money is a producer's good like machinery or inventories (Friedman, 1956; Laidler, 1985; Friedman and Schwartz, 1982).

Demand by ultimate wealth holders

For ultimate wealth holders the demand for money, in real terms, may be expected to be a function primarily of the following variables:

1. 1. *Total wealth*

This is the analogue of the budget constraint in the usual theory of consumer choice. It is the total that must be divided among various forms of assets. In practice, estimates of total wealth are seldom available. Instead, income may serve as an index of wealth. However, it should be recognized that income as measured by statisticians may be a defective index of wealth because it is subject to erratic year-to-year fluctuations, and a longer-term concept, like the concept of permanent income developed in connection with the theory of consumption, may be more useful (Friedman, 1957, 1959).

The emphasis on income as a surrogate for wealth, rather than as a measure of the 'work' to be done by money, is perhaps the basic conceptual difference between the more recent analyses of the demand for money and the earlier versions of the quantity theory.

2. 2. *The division of wealth between human and non-human forms*
   The major asset of most wealth holders is personal earning capacity. However, the conversion of human into non-human wealth or the reverse is subject to narrow limits because of institutional constraints. It can be done by using current earnings to purchase non-human wealth or by using non-human wealth to finance the acquisition of skills, but not by purchase or sale of human wealth and to only a limited extent by borrowing on the collateral of earning power. Hence, the fraction of total wealth that is in the form of non-human wealth may be an additional important variable.
3. 3. *The expected rates of return on money and other assets*
   These rates of return are the counterparts to the prices of a commodity and its substitutes and complements in the usual theory of consumer demand. The nominal rate of return on money may be zero, as it generally is on currency, or negative, as it sometimes is on demand deposits subject to net service charges, or positive, as it sometimes is on demand deposits on which interest is paid and generally is on time deposits. The nominal rate of return on other assets consists of two parts: first, any currently paid yield, such as interest on bonds, dividends on equities, or cost, such as storage costs on physical assets, and, second, a change in the nominal price of the asset. The second part is especially important under conditions of inflation or deflation.
4. 4. *Other variables determining the utility attached to the services rendered by money relative to those rendered by other assets – in Keynesian terminology, determining the value attached to liquidity proper*
   One such variable may be one already considered – namely, real wealth or income, since the services rendered by money may, in principle, be regarded by wealth holders as a 'necessity', like bread, the consumption of which increases less than in proportion to any increase in income, or as a 'luxury', like recreation, the consumption of which increases more than in proportion.

Another variable that is important empirically is the degree of economic stability expected to prevail, since instability enhances the value wealth-holders attach to liquidity. This variable has proved difficult to express quantitatively although qualitative information often indicates the direction of change. For example, the outbreak of war clearly produces expectations of greater instability. That is one reason why a notable increase in real balances – that is, a notable decline in velocity – often accompanies the outbreak of war. Such a decline in velocity produced an initial *decline* in sensitive prices at the outset of both World War I and World War II – not the rise that later inflation would have justified.

The rate of inflation enters under item 3 as a factor affecting the cost of holding various assets,

particularly currency. The variability of inflation enters here, as a major factor affecting the usefulness of money balances. Empirically, variability of inflation tends to increase with the level of inflation, reinforcing the negative effect of higher inflation on the quantity of money demanded.

Still another relevant variable may be the volume of trading in existing capital goods by ultimate wealth holders. The higher the turnover of capital assets, the larger the fraction of total assets people may find it useful to hold as cash. This variable corresponds to the class of transactions omitted in going from the transactions version of the quantity equation to the income version.

We can express this analysis in terms of the following demand function for money for an individual wealth holder:

$$M^D = P \cdot f\left(y, w, R_M^*, R_B^*, R_E^*; u\right), \tag{11}$$

where *M*, *P*, and *y* have the same meaning as in equation (6) except that they relate to a single wealth-holder (for whom $y=y'$); *w* is the fraction of wealth in non-human form (or, alternatively, the fraction of income derived from property); an asterisk denotes an expected value, so R_M^* is the expected nominal rate of return on money; R_B^* is the expected nominal rate of return on fixed-value securities, including expected changes in their prices; R_E^* is the expected nominal rate of return on physical assets, including expected changes in their prices; and *u* is a portmanteau symbol standing for other variables affecting the utility attached to the services of money. Though the expected rate of inflation is not explicit in equation (11), it is implicit because it affects the expected nominal returns on the various classes of assets, and is sometimes used as a proxy for R_E^*. For some purposes it may be important to classify assets still more finely – for example, to distinguish currency from deposits, long-term from short-term fixed-value securities, risky from relatively safe equities, and one kind of physical assets from another. Furthermore, the several rates of return are not independent. Arbitrage tends to eliminate differences among them that do not correspond to differences in perceived risk or other nonpecuniary characteristics of the assets, such as liquidity. In particular, as Irving Fisher pointed out in 1896, arbitrage between real and nominal assets introduces an allowance for anticipated inflation into the nominal interest rate (Fischer, 1896; Friedman, 1956).

The usual problems of aggregation arise in passing from equation (11) to a corresponding equation for the economy as a whole – particular, from the possibility that the amount of money demanded may depend on the distribution among individuals of such variables as *y* and *w* and not merely on their aggregate or average value. If we neglect these distributional effects, equation (11) can be regarded as applying to the community as a whole, with *M* and *y* referring to per capital money holdings and per capital real income, respectively, and *w* to the fraction of aggregate wealth in non-human form. Although the mathematical equation may be the same, its significance is very different for the individual wealth-holder and the community as a whole. For the individual, all the variables in the equation other

than his own income and the disposition of his portfolio are outside his control. He takes them, as well as the structure of monetary institutions, as given, and adjusts his nominal balances accordingly. For the community as a whole, the situation is very different. In general, the nominal quantity of money available to be held is fixed and what adjusts are the variables on the right-hand side of the equation, including an implicit underlying variable, the structure of monetary institutions, which, in the longer run, at least, adjusts itself to the tastes and preferences of the holders of money. A dramatic example is provided by the restructuring of the financial system in the US in the 1970s and 1980s.

In practice, the major problems that arise in applying equation (11) are the precise definitions of y and w, the estimation of *expected* rates of return as contrasted with actual rates of return, and the quantitative specification of the variables designated by u.

Demand for business enterprises

Business enterprises are not subject to a constraint comparable to that imposed by total wealth of the ultimate wealth-holder. They can determine the total amount of capital embodied in productive assets, including money, to maximize returns, since they can acquire additional capital through the capital market.

A similar variable defining the ‘scale’ of the enterprise may, however, be relevant as an index of the productive value of different quantities of money to the enterprise. Lack of data has meant that much less empirical work has been done on the business demand for money than on the aggregate demand enterprises. As a result, there are as yet only faint indications about the best variable to use: whether total transactions, net value added, net income, total capital in nonmoney form, or net worth.

The division of wealth between human and non-human form has no special relevance to business enterprises, since they are likely to buy the services of both forms on the market.

Rates of return on money and on alternative assets are, of course, highly relevant to business enterprises. These rates determine the net cost of holding money balances. However, the particular rates that are relevant may differ from those that are relevant for ultimate wealth-holders. For example, the rates banks charge on loans are of minor importance for wealth-holders yet may be extremely important for businesses, since bank loans may be a way in which they can acquire the capital embodied in money balances.

The counterpart for business enterprises of the variable u in equation (11) is the set of variables other than scale affecting the productivity of money balances. At least one subset of such variables – namely, expectations about economic stability and the variability of inflation – is likely to be common to business enterprises and ultimate wealth-holders.

With these interpretations of the variables, equation (11), with w excluded, can be regarded as symbolizing the business demand for money and, as it stands, symbolizing aggregate demand for money, although with even more serious qualifications about the ambiguities introduced by aggregation.

Buffer stock effects

In serving its basic function as a temporary abode of purchasing power, cash balances necessarily fluctuate, absorbing temporary discrepancies between the purchases and sales they mediate.

Though always recognized, this `buffer stock’ role of money has seldom been explicitly modelled.

Recently, more explicit attention has been paid to the buffer stock notion in an attempt to explain anomalies that have arisen in econometric estimates of the short-run demand for money (Judd and Scadding, 1982; Laidler, 1984; Knoester, 1984).

(e) The reconciliation of demand with supply

Multiply equation (11) by *N* to convert it from a per capita to an aggregate demand function, and equate it to equation (10), omitting for simplicity the asterisks designating expected values, and letting *R* stand for a vector of interest rates:

$$M^S = h(R, Y, \ldots) = P \cdot N \cdot f(y, w, R, g_p, u). \quad (12)$$

The result is quantity equation (6) in an expanded form. In principle, a change in any of the underlying variables that produces a change in M^S and disturbs a pre-existing equilibrium can produce offsetting changes in any of the other variables. In practice, as already noted earlier, the initial impact is likely to be on *y* and *R*, the ultimate impact predominantly on *P*.

A frequent criticism of the quantity theory is that its proponents do not specify the transmission mechanism between a change in M^S and the offsetting changes in other variables, that they rely on a black box connecting the input – the nominal quantity of money – and the output – effects on prices and quantities.

This criticism is not justified insofar as it implies that the transmission mechanism for the quantity equation is fundamentally different from that for a demand–supply analysis of a particular product – shoes, or copper, or haircuts. In both cases the demand function for the community as a whole is the sum of demand functions for individual consumer or producer units, and the separate demand functions are determined by the tastes and opportunities of the units. In both cases, the supply function depends on production possibilities, institutional arrangements for organizing production, and the conditions of supply of resources. In both cases a shift in supply or in demand introduces a discrepancy between the amounts demanded and supplied *at the pre-existing price*. In both cases any discrepancy can be eliminated only by either a price change or some alternative rationing mechanism, explicit or implicit.

Two features of the demand–supply adjustment for money have concealed this parallelism. One is that demand–supply analysis for particular products typically deals with flows – number of pairs of shoes or number of haircuts per year – whereas the quantity equations deal with the stock of money at a point in time. In this respect the correct analogy is with the demand for, say, land, which, like money, derives its value from the flow of services it renders but has a purchase price and not merely a rental value. The second is the widespread tendency to confuse ‘money’ and ‘credit’, which has produced misunderstanding about the relevant price variable. The ‘price’ of money is the quantity of goods and services that must be given up to acquire a unit of money – the inverse of the price level. This is the price that is analogous to the price of land or of copper or of haircuts. The ‘price’ of money is not the

interest rate, which is the ‘price’ of credit. The interest rate connects stocks with flows – the rental value of land with the price of land, the value of the service flow from a unit of money with the price of money. Of course, the interest rate may affect the quantity of money demanded – just as it may affect the quantity of land demanded – but so may a host of other variables.

The interest rate has received special attention in monetary analysis because, without quite realizing it, fractional reserve banks have created part of the stock of money in the course of serving as an intermediary between borrowers and lenders. Hence changes in the quantity of money have frequently occurred through the credit markets, in the process producing important transitory effects on interest rates.

On a more sophisticated level, the criticism about the transmission mechanism applies equally to money and to other goods and services. In all cases it is desirable to go beyond equality of demand and supply as defining a stationary equilibrium position and examine the variables that affect the quantities demanded and supplied and the dynamic temporal process whereby actual or potential discrepancies are eliminated. Examination of the variables affecting demand and supply has been carried farther for money than for most other goods or services. But for both, there is as yet no satisfactory and widely accepted description, in precise quantifiable terms, of the dynamic temporal process of adjustment. Much research has been devoted to this question in recent decades; yet it remains a challenging subject for research. (For surveys of some of the literature, see Laidler, 1985; Judd and Scadding, 1982.)

(f) First-round effects

Another frequent criticism of the quantity equations is that they neglect any effect on the outcome of the source of change in the quantity of money. In Tobin's words, the question is whether ‘the genesis of new money makes a difference’, in particular, whether ‘an increase in the quantity of money has the same effect whether it is issued to purchase goods or to purchase bonds’ (1974, p. 87).

Or, as John Stuart Mill put a very similar view in 1844, ‘The issues of a *Government* paper, even when not permanent, will raise prices; because Governments usually issue their paper in purchases for consumption. If issued to pay off a portion of the national debt, we believe they would have no effect’ (1844, p. 589).

Tobin and Mill are right that the way the quantity of money is increased affects the outcome in some measure or other. If one group of individuals receives the money on the first round, they will likely use it for different purposes than another group of individuals. If the newly printed money is spent on the first round for goods and services, it adds directly at that point to the demand for such goods and services, whereas if it is spent on purchasing debt, or simply held temporarily as a buffer stock, it has no immediate effect on the demand for goods and services. Such effects come later as the initial recipients of the ‘new’ money dispose of it. However, as the ‘new’ money spreads through the economy, any first-round effects tend to be dissipated. The ‘new’ money is merged with the old and is distributed in much the same way.

One way to characterize the Keynesian approach (see below) is that it gives almost exclusive importance to the first-round effect by putting primary emphasis on flows of spending rather than on stocks of assets. Similarly, one way to characterize the quantity-theory approach is to say that it gives almost no importance to first-round effects.

The empirical question is how important the first-round effects are compared with the ultimate effects.

Theory cannot answer that question. The answer depends on how different are the reactions of the recipients of cash via alternative routes, on how rapidly a larger money stock is distributed through the economy, on how long it stays at each point in the economy, on how much the demand for money depends on the structure of government liabilities, and so on. Casual empiricism yields no decisive answer. Maybe the first-round effect is so strong that it dominates later effects; maybe it is highly transitory.

Despite repeated assertions by various authors that the first-round effect is significant, none, so far as I know, has presented any systematic empirical evidence to support that assertion. The apparently similar response of spending to changes in the quantity of money at widely separated dates in different countries and under diverse monetary systems establishes something of a presumption that the first-round effect is not highly significant. This presumption is also supported by several empirical studies designed to test the importance of the first-round effect (Cagan, 1972).

(g) The international transmission mechanism

From its very earliest days, the quantity theory was intimately connected with the analysis of the adjustment mechanism in international trade. A commodity standard, in which money is specie or its equivalent, was taken as the norm. Under such a standard, the supply of money in any one country is determined by the links between that country and other countries that use the same commodity as money. Under such a standard, the same theory explains links among money, prices, and nominal income in various parts of a single country – money, prices, and nominal income in Illinois and money, prices, and nominal income in the rest of the United States – and the corresponding links among various countries. The differences between interregional adjustment and international adjustment are empirical: greater mobility of people, goods, and capital among regions than among countries, and hence more rapid adjustment.

According to the specie-flow mechanism developed by Hume and elaborated by Henry Thornton, David Ricardo and their successors, ‘too’ high a money stock in country A tends to make prices in A high relative to prices in the rest of the world, encouraging imports and discouraging exports. The resulting deficit in the balance of trade is financed by shipment of specie, which reduces the quantity of money in country A and increases it in the rest of the world. These changes in the quantity of money tend to lower prices in country A and raise them in the rest of the world, correcting the original disequilibrium. The process continues until price levels in all countries are at a level at which balances of payments are in equilibrium (which may be consistent with a continuing movement of specie, for example, from gold- or silver-producing countries to non-gold- or silver-producing countries, or between countries growing at different secular rates).

Another strand of the classical analysis has recently been revived under the title ‘the monetary theory of the balance of payments’. The specie-flow mechanism implicitly assumes that prices adjust only in response to changes in the quantity of money produced by specie flows. However, if markets are efficient and transportation costs are neglected, there can be only a single price expressed in a common currency for goods traded internationally. Speculation tends to assure this result. Internally, competition between traded and nontraded goods tends to keep their relative price in line with relative costs. If these adjustments are rapid, ‘the law of one price’ holds among countries. If the money stock is not distributed among countries in such a way as to be consistent with the equilibrium prices, excess demands and

supplies of money will lead to specie flows. Domestic nominal demand in a country with ‘too’ high a quantity of money will exceed the value of domestic output and the excess will be met by imports, producing a balance of payments deficit financed by the export of specie; and conversely in a country with too ‘low’ a quantity of money. Specie flows are still the adjusting mechanism, but they are produced by differences between demand for output in nominal terms and the supply of output at world prices rather than by discrepancies in prices. Putative rather than actual price differences are the spur to adjustment. This description is highly oversimplified, primarily because it omits the important role assigned to short- and long-term capital flows by all theorists – those who stress the specie-flow mechanism and even more those who stress the single-price mechanism (Frenkel, 1976; Frenkel and Johnson, 1976).

In practice, few countries have had pure commodity standards. Most have had a mixture of commodity and fiduciary standards. Changes in the fiduciary component of the stock of money can replace specie flows as a means of adjusting the quantity of money.

The situation is still different for countries that do not share a unified currency, that is, a currency in which only the name assigned to a unit of currency differs among countries. Changes in the rates of exchange between national currencies then serve to keep prices in various countries in the appropriate relation when expressed in a common currency. Exchange rate adjustments replace specie flows or changes in the quantity of domestically created money. And exchange rate changes too may be produced by actual or putative price differences or by short- or long-term capital flows. Moreover, especially during the Bretton Woods period (1945–71), but more recently as well, governments have often tried to avoid changes in exchange rates by seeking adjustment through subsidies to exports, obstacles to imports, and direct controls over foreign exchange transactions. These measures involved either implicit or explicit multiple rate systems and were accompanied by government borrowing to finance balance-of-payments deficits, or governmental lending to offset surpluses. They sometimes led to severe financial crises and major exchange rate adjustments – one reason the Bretton Woods system finally broke down in 1971. Since then, exchange rates have supposedly been free to float and to be determined in private markets. In practice, however, governments still intervene in an attempt to affect the exchange rates of their currencies, either directly by buying or selling their currency on the market, or indirectly, by adopting monetary or fiscal or trade policies designed to alter the market exchange rate. However, most governments no longer announce fixed parities for their currencies.

2 Keynesian challenge to the quantity theory

The depression of the 1930s produced a wave of scepticism about the relevance and validity of the quantity theory of money. The central banks of the world – the Federal Reserve in the forefront – proclaimed that, despite the teachings of the quantity theory, ‘easy money’ was proving to be ineffective in stemming the depression. They pointed to the low level of short-term interest rates as evidence of how ‘easy’ monetary policy was. Their claims seemed credible not only because of the confusion between ‘lowness of interest’ and ‘plenty of money’ pointed out by Hume but also because of the absence of readily available evidence on what was happening to the quantity of money. Most observers at the time did not know, as we do now, that the Federal Reserve permitted the quantity of money in the United States to decline by one-third between 1929 and 1933, and hence that the accompanying contraction in economic activity and deflation of prices was entirely consistent with the quantity theory. Monetary

policy was incredibly ‘tight’ not ‘easy’.

The scepticism about the quantity theory was further heightened by the publication of John Maynard Keynes's *The General Theory of Employment, Interest and Money* (Keynes, 1936) which offered an alternative interpretation of economic fluctuations in general and the depression in particular. Keynes emphasized spending on investment and the stability of the consumption function rather than the stock of money and the stability of the demand function for money. He relegated the forces embodied in the quantity theory to a minor role, and treated fiscal rather than monetary policy as the chief instrument for influencing the course of events. Received wisdom both inside and outside the economics profession became ‘money does not matter’.

Keynes did not deny the validity of the quantity equation, in any of its forms – after all, he had been a major contributor to the quantity theory (Keynes, 1923). What he did was something very different. He argued that the demand for money, which he termed the liquidity-preference function, had a special form such that *under conditions of underemployment* the V in equation (4) and the k in equation (6) would be highly unstable and would passively adapt to whatever changes independently occurred in money income or the stock of money. Under such conditions, these equations, though entirely valid, were largely useless for policy or prediction. Moreover, he regarded such conditions as prevailing much, if not most of the time.

That possibility rested on two other key propositions. First, that, contrary to the teachings of classical and neoclassical economists, the *long-run equilibrium* position of an economy need not be characterized by ‘full employment’ of resources even if all prices are flexible. In his view, unemployment could be a deep-seated characteristic of an economy rather than simply a reflection of price and wage rigidity or transitory disturbances. This proposition has played an important role in promoting the acceptance of Keynesianism, especially by non-economists, even though, by now, it is widely accepted that, as a *theoretical* matter, the proposition is false. Keynes's error consisted in neglecting the role of wealth in the consumption function. There is no fundamental ‘flaw in the price system’ that makes persistent structural unemployment a possible or probable natural outcome of a fully operative market system (Haberler, 1941, pp. 242, 389, 403, 491–503; Pigou, 1947; Tobin, 1947; Patinkin, 1948; Johnson, 1961). The concept of ‘underemployment equilibrium’ has been replaced by the concept of a ‘natural rate of unemployment’ (see section 3 below).

Keynes's final key proposition was that, as an *empirical* matter, prices, especially wages, can be regarded as rigid – an institutional datum – for *short-run economic fluctuations*; in which case, the distinction between real and nominal magnitudes that is at the heart of the quantity theory is irrelevant for such fluctuations. This proposition, unlike the other two, did not conflict with the teachings of the quantity theory. Classical and neoclassical economists had long recognized that price and wage rigidity existed and contributed to unemployment during cyclical contractions, and to labour scarcity during cyclical booms. But to them, wage rigidity was a defect of the market; to Keynes, it was a rational response to the possibility of underemployment equilibrium (Keynes, 1936, pp. 269–71).

In his analysis of the demand for money (i.e., the form of equation (6) or (11)), Keynes treated the stock of money as if it were divided into two parts, one part, M_1, ‘held to satisfy the transactions- and precautionary-motives’, the other, M_2, ‘held to satisfy the speculative-motive’ (Keynes, 1936, p. 199). He regarded M_1 as a roughly constant fraction of income. He regarded the demand for M_2 as arising from ‘*uncertainty* as to the future course of the rate of interest’ (Keynes, 1936, p. 168) and the amount

demanded as depending on the relation between current rates of interest and the rates of interest expected to prevail in the future. Keynes, of course, recognized the existence of a whole complex of interest rates. However, for simplicity, he spoke in terms of ‘the rate of interest’, usually meaning by that the rate on long-term securities that were fixed in nominal value and that involved minimal risks of default – for example, government bonds. In a ‘given state of expectations’, the higher the current rate of interest, the lower would be the (real) amount of money that people would want to hold for speculative motives for two reasons: first, the greater would be the cost in terms of current earnings sacrificed by holding money instead of securities, and, second, the more likely it would be that interest rates would fall, and hence bond prices rise, and so the greater would be the cost in terms of capital gains sacrificed by holding money instead of securities.

To formalize Keynes's analysis in terms of the symbols we have used so far, we can write his demand (liquidity-preference) function as

$$M/P = M_1/P + M_2/P = k_1 y' + f\left(R - R^*, R^*\right) \tag{13}$$

where R is the current rate of interest, R^* is the rate of interest expected to prevail, and k_1, the analogue to the inverse of the income velocity of circulation of money, is treated as determined by payment practices and hence as a constant at least in the short run. Later writers in this tradition have argued that k_1 too should be regarded as a function of interest rates (Baumol, 1952; Tobin, 1956).

Although expectations are given great prominence in developing the liquidity function expressing the demand for M_2, Keynes and his followers generally did not explicitly introduce an expected interest rate into that function as is done in equation (13). For the most part, in practice, they treated the amount of M_2 demanded as a function simply of the current interest rate, the emphasis on expectations serving only as a reason for attributing instability to the liquidity function. Moreover, for the most part, they omitted P (and replaced y' by Y) because of their assumption that prices were rigid.

Except for somewhat different language, the analysis up to this point differs from that of earlier quantity theorists, such as Fisher, only by its subtle analysis of the role of expectations about future interest rates, its greater emphasis on current interest rates, and its narrower restriction of the variables explicitly considered as affecting the amount of money demanded.

Keynes's special twist concerned the empirical form of the liquidity-preference function at the low interest rates that he believed would prevail under conditions of underemployment equilibrium. Let the interest rate fall sufficiently low, he argued, and money and bonds would become perfect substitutes for one another; liquidity preference, as he put it, would become absolute. The liquidity-preference function, expressing the quantity of M_2 demanded as a function of the rate of interest, would become horizontal at some low but finite rate of interest. Under such circumstances, an increase in the quantity of money by whatever means would lead holders of money to seek to convert their additional cash balances into bonds, which would tend to lower the rate of interest on bonds. Even the slightest lowering would lead

speculators with firm expectations to absorb the additional money balances by selling any bonds demanded by the initial holders of the additional money. The result would simply be that the community as a whole would hold the increased quantity of money without any change in the interest rate; *k* would be higher and *V* lower. Conversely, a decrease in the quantity of money would lead holders of bonds to seek to restore their money balances by selling bonds, but this would tend to raise the rate of interest, and even the slightest rise would induce the speculators to absorb the bonds offered.

Or, again, suppose nominal income increases or decreases for whatever reason. That will require an increase or decrease in M_1, which can come out of or be transferred to M_2 without any further effects.

The conclusion is that, *under circumstances of absolute liquidity preference*, income can change without a change in *M* and *M* can change without a change in income. The holders of money are in metastable equilibrium, like a tumbler on its side on a flat surface; they will be satisfied with whatever the quantity of money happens to be.

Keynes regarded absolute liquidity preference as a strictly ‘limiting case’ of which, though it ‘might become practically important in future’, he knew ‘of no example … hitherto’ (1936, p. 207). However, he treated velocity as if in practice its behaviour frequently approximated that which would prevail in this limiting case.

Keynes's disciples went much farther than Keynes himself. They were readier than he was to accept absolute liquidity preference as the actual state of affairs. More important, many argued that when liquidity preference was not absolute, changes in the quantity of money would affect only the interest rate on bonds and that changes in this interest rate in turn would have little further effect. They argued that both consumption expenditures and investment expenditures were nearly completely insensitive to changes in interest rates, so that a change in *M* would merely be offset by an opposite and compensatory change in *V* (or a change in the same direction in *k*), leaving *P* and *y* almost completely unaffected. In essence their argument consists in asserting that only paper securities are substitutes for money balances – that real assets never are (see Hansen, 1957, p. 50; Tobin, 1961).

The apparent success during the 1950s and 1960s of governments committed to a Keynesian full-employment policy in achieving rapid economic growth, a high degree of economic stability, and relatively stable prices and interest rates, for a time strongly reinforced belief in the initial Keynesian views about the unimportance of variations in the nominal quantity of money.

The 1970s administered a decisive blow to these views and fostered a revival of belief in the quantity theory. Rapid monetary growth was accompanied not only by accelerated inflation but also by rising, not falling, average levels of unemployment (Friedman, 1977), and by rising, not declining, interest rates.

As Robert Lucas put it in 1981,

> Keynesian orthodoxy … appears to be giving seriously wrong answers to the most basic questions of macroeconomic policy. Proponents of a class of models which promised $3\frac{1}{2}$ to $4\frac{1}{2}$ percent unemployment to a society willing to tolerate annual inflation rates of 4 to 5 percent have some explaining to do after a decade [i.e., the 1970s] such as we have just come through. A forecast error of this magnitude and central importance to policy has consequences (pp. 559–60).

This experience undermined the belief that the price level could be regarded as rigid – or at any rate as determined by forces unrelated to the quantity of money; that the nominal quantity of money demanded could be regarded as a function primarily of the nominal interest rate, and that absolute liquidity preference was the normal state of affairs. No teacher of elementary economics since the late 1970s can, as so many did in the 1940s, 1950s, and 1960s, draw on the blackboard a downward sloping liquidity-preference diagram with the nominal quantity of money on the horizontal axis and a nominal interest rate on the vertical axis and confidently proclaim that the only important effect of an increase in the nominal quantity of money would be to lower the rate of interest. The distinction between the nominal interest rate and the real interest rate introduced by Irving Fisher in 1896 has entered – or re-entered – received wisdom (Fisher, 1896).

Despite its subsidence, the Keynesian attack on the quantity theory has left its mark. It has reinforced the tendency, already present in the Cambridge approach, to stress the role of money as an asset and hence to regard the analysis of the demand for money as part of capital or wealth theory, concerned with the composition of the balance sheet or portfolio of assets. The Keynesian stress on autonomous spending and hence on fiscal policy remains important in its own right but also has led to greater emphasis on the effect of government fiscal policies on the demand for money. Keynes's stress on expectations has contributed to the rapid growth in the analysis of the role and formation of expectations in a variety of economic contexts. Conversely, the revival of the quantity theory has led Keynesian economists to treat changes in the quantity of money as an essential element in the analysis of short-term change.

Finally, the controversy between Keynesians and quantity theorists has led both groups to distinguish more sharply between long-run and short-run effects of monetary changes; between ‘static’ or ‘long-run equilibrium’ theory and the dynamics of economic change.

As Franco Modigliani put it in his 1976 presidential address to the American Economic Association, there are currently ‘no serious analytical disagreements between leading monetarists [i.e., quantity theorists] and leading nonmonetarists [i.e., Keynesians]’ (1977, p. 1).

However, there still remain important differences on an empirical level. These all centre on the dynamics of short-run change – the process whereby a change in the quantity of money affects aggregate spending and the role of fiscal variables in the process.

The Keynesians regard a change in the quantity of money as affecting in the first instance ‘the’ interest rate, interpreted as a market rate on a fairly narrow class of financial liabilities. They regard spending as affected only ‘indirectly’ as the changed interest rate alters the profitability and amount of investment spending, again interpreted fairly narrowly, and as investment spending, through the multiplier, affects total spending. Hence the emphasis they give in their analysis to the interest elasticities of the demand for money and of investment spending.

The quantity theorists, on the other hand, stress a much broader and more ‘direct’ impact of spending, saying, as in section 1a above, that individuals will seek ‘to dispose of what they regard as their excess money balances by paying out a larger sum for the purchase of securities, goods, and services, for the repayment of debts, and as gifts than they are receiving from the corresponding sources’.

The two approaches can be readily reconciled on a formal level. Quantity theorists can describe the transmission mechanism as operating ‘through’ the balance sheet and ‘through’ changes in interest rates. The attempt by holders of money to restore or attain a desired balance sheet after an unexpected increase in the quantity of money tends initially to raise the prices of assets and reduce interest rates, which encourages spending to produce new assets and also spending on current services rather than on

purchasing existing assets. This is how an initial effect on balance sheets gets translated into an effect on income and spending. The resulting increase in spending tends to raise prices of goods and services which, in turn, by lowering the real value of the quantity of money and of nominal assets, tends to eliminate the initial decline in interest rates, even overshooting in the process.

The difference between the quantity theorists and the Keynesians is less in the nature of the process than in the range of assets considered. The Keynesians tend to concentrate on a narrow range of marketable assets and recorded interest rates. The quantity theorists insist that a far wider range of assets and interest rates must be taken into account – such assets as durable and semi-durable consumer goods, structures, and other real property. As a result, the quantity theorists regard the market rates stressed by the Keynesians as only a small part of the total spectrum of rates that are relevant.

This difference in the assumed transmission mechanism is largely a by-product of the different assumptions about price. The rejection of absolute liquidity preference forced Keynes's followers to let the interest rate be flexible. This chink in the key assumption that prices are an institutional datum was minimized by interpreting the ‘interest rate’ narrowly, and market institutions made it easy to do so. After all, it is most unusual to quote the ‘interest rate’ implicit in the sales and rental prices of houses and automobiles, let alone furniture, household appliances, clothes, and so on. Hence the prices of these items continued to be regarded as an institutional datum, which forced the transmission process to go through an extremely narrow channel. On the side of the quantity theorists there was no such inhibition. Since they regard prices as flexible, though not ‘perfectly’ flexible, it was natural for them to interpret the transmission mechanism in terms of relative price adjustments over a broad area rather than in terms of narrowly defined interest rates.

Less important differences are the tendency for Keynesians to stress the short-run as opposed to the long-run impact of changes to a far greater extent than the quantity theorists; and, a related difference, to give greater scope to the first-round effect of changes in the quantity of money.

3 The Phillips curve and the natural rate hypothesis

A major postwar development that contributed greatly to the revival of the quantity theory grew out of criticism by quantity theorists of the ‘Phillips curve’ – an allegedly stable inverse relation between unemployment and the rate of change of nominal wages such that a high level of unemployment was accompanied by declining wages, a low level by rising wages. Though not formally linked to the Keynesian theoretical system, the Phillips curve was widely welcomed by Keynesians as helping to fill a gap in the system created by the assumption of rigid wages. In addition, it appeared to offer an attractive trade-off possibility for economic policy: a permanent reduction in the level of unemployment at the cost of a moderate sustained increase in the rate of inflation. The Keynesian assumption that prices and wages could be regarded as institutionally determined made it easy for them to accept a relation between a nominal magnitude (the rate of change of wages) and a real magnitude (unemployment).

By contrast, the quantity theory distinction between real and nominal magnitudes implies that the Phillips curve is theoretically flawed. The quantity of labour demanded is a function of real not nominal wages; and so is the quantity supplied. Under any given set of circumstances, there is an equilibrium level of unemployment corresponding to an equilibrium structure of *real* wage rates. A higher level of unemployment will put downward pressure on real wage rates; a lower level will put upward pressure on real wage rates. The level of unemployment consistent with the equilibrium structure of real wage rates

has been termed the ‘natural rate of unemployment’ and defined as

> the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labour and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labour availabilities, the costs of mobility, and so on (Friedman, 1968, p. 8).

The nominal wage rate that corresponds to any given real wage rate depends on the level of prices. Whether that nominal wage rate is rising or falling depends on whether prices are rising or falling. If wages and prices change at the same rate, the real wage rate remains the same. Hence, in the long run, there need be no relation between the rate of change of *nominal* wages and the rate of change of *real* wages, and hence between the rate of change of nominal wages and the level of unemployment. In the long run, therefore, the Phillips curve will tend to be vertical at the natural rate of unemployment – a proposition that came to be termed the Natural Rate Hypothesis.

Over short periods, an *unanticipated* increase in inflation reduces real wages as viewed by employers, inducing them to offer higher nominal wages, which workers erroneously view as higher real wages. This discrepancy simultaneously encourages employers to offer more employment and workers to accept more employment, thereby reducing unemployment, which produces the inverse relation encapsulated in the Phillips curve. However, if the higher rate of inflation continues, the anticipations of workers and employers will converge and the decline in unemployment will be reversed. A negatively sloping Phillips curve is therefore a short-run phenomenon. Moreover, it will not be stable over time, since what matters is not the nominal rates of change of wages and prices but the difference between the actual and the *anticipated* rates of change. The emergence of stagflation in the 1970s quickly confirmed this analysis, leading to the widespread replacement of the original Phillips curve by an expectations-adjusted Phillips curve (Friedman, 1977).

Acceptance of the natural rate hypothesis has had far-reaching effects not only on received wisdom among economists but also on economic policy. It became widely recognized that expansionary monetary and fiscal policies at best gave only a temporary stimulus to output and employment and if long continued would be reflected primarily in inflation.

4 The theory of rational expectations

A subsequent theoretical development was the belated flowering of a seed planted in 1961 by John F. Muth, in a long-neglected article on ‘Rational expectations and the theory of price movements’ (Muth, 1961). The theory of rational expectations offers no special insight into stationary-state or long-run equilibrium analysis. Its contribution is to dynamics – short-run change, and hence potentially to stabilization policy.

It has long been recognized by writers of all persuasions that, as Abraham Lincoln put it over a century ago, ‘you can't fool all of the people all of the time.’ The tendency for the public to learn from experience and to adjust to it underlies David Hume's view that monetary expansion ‘is favourable to industry’ only in its initial stages, but that if it continues, it will come to be anticipated and will affect

prices and nominal interest rates but not real magnitudes. It also underlies the companion view associated with the natural rate hypothesis that a ‘full employment’ policy in which monetary, or for that matter fiscal, measures are used to counteract any increase in unemployment will almost inevitably lead not simply to uneven inflation but to uneven inflation around a rising trend – a conclusion often illustrated by analogizing inflation to a drug of which the addict must take larger and larger doses to get the same kick.

Nonetheless, the importance of anticipations and how they are formed in determining the dynamic response to changes in money and other magnitudes remained largely implicit until Lucas and Sargent applied the Muth rational expectations idea explicitly to the reliability of econometric models of the economy and to stabilization policies (Fischer, 1980; Lucas, 1976; Lucas and Sargent, 1981).

The theory of rational expectations asserts that economic agents should be treated as if their anticipations fully incorporate both currently available information about the state of the world and a correct theory of the interrelationships among the variables. Anticipations formed in this way will on the average tend to be correct (a statement whose simplicity conceals fundamental problems of interpretation, Friedman and Schwartz, 1982, pp. 556–7).

The rational expectations hypothesis has far-reaching implications for the validity of econometric models. Suppose a statistician were able to construct a model that predicted highly accurately for a past period all relevant variables; also, that a monetary rule could be devised that if used during the past period with that model could have achieved a particular objective – say keeping unemployment between 4 and 5 per cent. Suppose now that that policy rule were adopted for the future. It would be nearly certain that the model for which the rule was developed would no longer work. The economic equivalent of the Heisenberg indeterminacy principle would take over. The model was for an economy without that monetary rule. Put the rule into effect and it will alter rational expectations and hence behaviour. Even without putting the rule into effect, the model would very likely continue to work only so long as its existence could be kept secret because if market participants learned about it they would use it in forming their rational expectations and thereby falsify it to a greater or lesser extent. Little wonder that every major econometric model is always being sent back to the drawing board as experience confounds it, or that their producers have reacted so strongly to the theory of rational expectations.

The implication of one variant of the theory that has received the most attention and generated the most controversy is the so-called neutrality hypothesis about stabilization policy – in particular, about discretionary monetary policy directed at promoting economic stability. Correct rational expectations of economic agents will include correct anticipation of any systematic monetary policy; hence such policy will be allowed for by economic agents in determining their behaviour. Given further the natural rate hypothesis, it follows that any systematic monetary policy will affect the behaviour only of nominal magnitudes and not of such real magnitudes as output and employment. The authorities can affect the course of events only by ‘fooling’ the participants, that is, by acting in an unpredictable, ad hoc way. But, in general, such strictly ad hoc intervention will destabilize the economy, not stabilize it, serving simply to introduce another series of random shocks into the economy to which participants must adapt and which reduce their ability to form precise and accurate expectations.

This is a highly oversimplified account of the rational expectations hypothesis and its implications. All otherwise valid models of the economy will not be falsified by being known. All real effects of systematic and announced governmental policies will not be rendered nugatory. Serious problems have arisen in formulating the hypothesis in a logically satisfactory way, and in giving it empirical content,

especially in incorporating multi-valued rather than single-valued expectations and allowing for non-independence of events over time. Research in this area is exploding; rapid progress and many changes in received opinion can confidently be anticipated before the rational expectations revolution is fully domesticated.

5 Empirical evidence

There is perhaps no empirical regularity among economic phenomena that is based on so much evidence for so wide a range of circumstances as the connection between substantial changes in the quantity of money and in the level of prices. There are few if any instances in which a substantial change in the quantity of money per unit of output has occurred without a substantial change in the level of prices in the same direction. Conversely, there are few if any instances in which a substantial change in the level of prices has occurred without a substantial change in the quantity of money per unit of output in the same direction. And instances in which prices and the quantity of money have moved together are recorded for many centuries of history, for countries in every part of the globe, and for a wide diversity of monetary arrangements.

The statistical connection itself, however, tells nothing about direction of influence, and this is the question about which there has been the most controversy. A rise or fall in prices, occurring for whatever reason, could produce a corresponding rise or fall in the quantity of money, so that the monetary changes are a passive consequence. Alternatively, changes in the quantity of money could produce changes in prices in the same direction, so that control of the quantity of money implies control of prices. The second interpretation – that substantial changes in the quantity of money are both a necessary and a sufficient condition for substantial changes in the general level of prices – is strongly supported by the variety of monetary arrangements for which a connection between monetary and price movements has been observed. But of course this interpretation does not exclude a reflex influence of changes in prices on the quantity of money. The reflex influence is often important, almost always complex, and, depending on the monetary arrangements, may be in either direction.

Evidence from specie standards

Until modern times, money was mostly metallic – copper, brass, silver, gold. The most notable changes in its nominal quantity were produced by sweating and clipping, by governmental edicts changing the nominal values attached to specified physical quantities of the metal, or by discoveries of new sources of specie. Economic history is replete with examples of the first two and their coincidence with corresponding changes in nominal prices (Cipolla, 1956; Feavearyear, 1931). The specie discoveries in the New World in the 16th century are the most important example of the third. The association between the resulting increase in the quantity of money and the price revolution of the 16th and 17th centuries has been well documented (Hamilton, 1934).

Despite the much greater development of deposit money and paper money, the gold discoveries in Australia and the United States in the 1840s were followed by substantial price rises in the 1850s (Cairnes, 1873; Jevons, 1863). When growth of the gold stock slowed, and especially when country after country shifted from silver to gold (Germany in 1871–3, the Latin Monetary Union in 1873, the

Netherlands in 1875–6) or returned to gold (the United States in 1879), world prices in terms of gold fell slowly but fairly steadily for about three decades. New gold discoveries in the 1880s and 1890s, powerfully reinforced by improved methods of mining and refining, particularly commercially feasible methods of using the cyanide process to extract gold from low-grade ore, led to much more rapid growth of the world gold stock. Further, no additional important countries shifted to gold. As a result, world prices in terms of gold rose by 25 to 50 per cent from the mid-1890s to 1914 (Bordo and Schwartz, 1984).

Evidence from great inflations

Periods of great monetary disturbances provide the most dramatic evidence on the role of the quantity of money. The most striking such periods are the hyperinflations after World War I in Germany, Austria, and Russia, and after World War II in Hungary and Greece, and the rapid price rises, if not hyperinflations, in many South American and some other countries both before and after World War II. These 20th-century episodes have been studied more systematically than earlier ones. The studies demonstrate almost conclusively the critical role of changes in the quantity of money (Cagan, 1965; Meiselman, 1970; Sargent, 1982).

Substantial inflations following a period of relatively stable prices have often had their start in wartime, though recently they have become common under other circumstances. What is important is that something, generally the financing of extraordinary governmental expenditures, produces a more rapid growth of the quantity of money. Prices start to rise, but at a slower pace than the quantity of money, so that for a time the real quantity of money increases. The reason is twofold: first, it takes time for people to readjust their money balances; second, initially there is a general expectation that the rise in prices is temporary and will be followed by a decline. Such expectations make money a desirable form in which to hold assets, and therefore lead to an increase in desired money balances in real terms.

As prices continue to rise, expectations are revised. Holders of money come to expect prices to continue to rise, and reduce desired balances. They also take more active measures to eliminate the discrepancy between actual and desired balances. The result is that prices start to rise faster than the stock of money, and real balances start to decline (that is, velocity starts to rise). How far this process continues depends on the rate of rise in the quantity of money. If it remains fairly stable, real balances settle down at a level that is lower than the initial level but roughly constant – a constant expected rate of inflation implies a roughly constant level of desired real balances; in this case, prices ultimately rise at the same rate as the quantity of money. If the rate of money growth declines, inflation will follow suit, which will in turn lead to an increase in actual and desired real balances as people readjust their expectations; and conversely. Once the process is in full swing, changes in real balances follow with a lag changes in the rate of change of the stock of money. The lag reflects the fact that people apparently base their expectations of future rates of price change partly on an average of experience over the preceding several years, the period of averaging being shorter the more rapid the inflation.

In the extreme cases, those that have degenerated into hyperinflation and a complete breakdown of the medium of exchange, rates of price change have been so high and real balances have been driven down so low as to lead to the widespread introduction of substitute moneys, usually foreign currencies. At that point completely new monetary systems have had to be introduced.

A similar phenomenon has occurred when inflation has been effectively suppressed by price controls, so that there is a substantial gap between the prices that would prevail in the absence of controls and the legally permitted prices. This gap prevents money from functioning as an effective medium of exchange and also leads to the introduction of substitute moneys, sometimes rather bizarre ones like the cigarettes and cognac used in post-World War II Germany.

Other evidence

The past two decades have witnessed a literal flood of literature dealing with monetary phenomena. Expressed in broad terms, the literature has been of two overlapping types – qualitative and econometric – and has dealt with two overlapping sets of issues – static or long-term effects of monetary change and dynamic or cyclical effects.
Some broad findings are:
(1). For both long and short periods there is a consistent though not precise relation between the rate of growth of the quantity of money and the rate of growth of nominal income. If the quantity of money grows rapidly, so will nominal income, and conversely. This relation is much closer for long than for short periods.
Two recent econometric studies have tested the long-run effects using comparisons among countries for the post-World War II period. Lothian concludes his study for 20 countries for the period 1956–80:

> In this paper I have examined three sets of hypotheses associated with the quantity theory of money: the classical neutrality proposition [i.e., changes in the nominal quantity of money do not affect real magnitudes in the long run], the monetary approach to exchange rates [i.e., changes in exchange rates between countries reflect primarily changes in money per unit of output in the several countries], and the Fisher equation [i.e., differences in sustained rates of inflation produce corresponding differences in nominal interest rates]. The data are completely consistent with the first two and moderately supportive of the last (1985, p. 835).

Duck concludes his study for 33 countries and the period 1962 to 1982 – which uses overlapping data but substantially different methods:

> Its [the study's] findings suggest that (i) the real demand for money is reasonably well explained by a small number of variables, principally real income and interest rates; (ii) nominal income is closely related to the quantity of money, but is also related to the behaviour of other variables, principally interest rates; (iii) most changes in nominal income or its determinants are absorbed by price increases; (iv) even over a 20-year period some nominal income growth is to a significant degree absorbed by real output growth; (v) the evidence that expectations are rational is weak (1985, p. 33).

(2). These findings for the long run reflect a long-run real demand function for money involving, as Duck notes, a small number of variables, that is highly stable and very similar for different countries.

The elasticity of this function with respect to real income is close to unity, occasionally lower, generally higher, especially for countries that are growing rapidly and in which the scope of the money economy is expanding. The elasticity with respect to interest rates is, as expected, negative but relatively low in absolute value. The real quantity demanded is not affected by the price level (i.e., there is no ‘monetary illusion’) (Friedman and Schwartz, 1982; Laidler, 1985).

(3). Over short periods, the relation between growth in money and in nominal income is often concealed from the naked eye partly because the relation is less close for short than long periods but mostly because it takes time for changes in monetary growth to affect income, and how long it takes is itself variable. Today's income growth is not closely related to today's monetary growth; it depends on what has been happening to money in the past. What happens to money today affects what is going to happen to income in the future.

(4). For most major Western countries, a change in the rate of monetary growth produces a change in the rate of growth of nominal income about six to nine months later. This is an average that does not hold in every individual case. Sometimes the delay is longer, sometimes shorter. In particular, it tends to be shorter under conditions of high and highly variable rates of monetary growth and of inflation.

(5). In cyclical episodes the response of nominal income, allowing for the time delay, is greater in amplitude than the change in monetary growth, so that velocity tends to rise during the expansion phase of a business cycle and to fall during the contraction phase. This reaction appears to be partly a response to the pro-cyclical pattern of interest rates; partly to the linkage of desired cash balances to permanent rather than measured income.

(6). The changed rate of growth of nominal income typically shows up first in output and hardly at all in prices. If the rate of monetary growth increases or decreases, the rate of growth of nominal income and also of physical output tends to increase or decrease about six to nine months later, but the rate of price rise is affected very little.

(7). The effect on prices, like that on income and output, is distributed over time, but comes some 12 to 18 months later, so that the total delay between a change in monetary growth and a change in the rate of inflation averages something like two years. That is why it is a long row to hoe to stop an inflation that has been allowed to start. It cannot be stopped overnight.

(8). Even after allowance for the delayed effect of monetary growth, the relation is far from perfect. There's many a slip over short periods ‘twixt the monetary change and the income change.

(9). In the short run, which may be as long as three to ten years, monetary changes affect primarily output. Over decades, on the other hand, as already noted, the rate of monetary growth affects primarily prices. What happens to output depends on real factors: the enterprise, ingenuity and industry of the people; the extent of thrift; the structure of industry and government; the relations among nations, and so on. (In re points 3 to 9, Friedman and Schwartz, 1963a, 1963b; Friedman, 1961, 1977, 1984; Judd and Scadding, 1982.)

(10). One major finding has to do with severe depressions. There is strong evidence that a monetary crisis, involving a substantial decline in the quantity of money, is a necessary and sufficient condition for a major depression. Fluctuations in monetary growth are also systematically related to minor ups and downs in the economy, but do not play as dominant a role compared to other forces. As Friedman and Schwartz put it,

> Changes in the money stock are … a consequence as well as an independent source of

> change in money income and prices, though, once they occur, they produce in their turn still further effects on income and prices. Mutual interaction, but with money rather clearly the senior partner in longer-run movements and in major cyclical movements, and more nearly an equal partner with money income and prices in shorter-run and milder movements – this is the generalization suggested by our evidence (1963b, p. 695; Friedman and Schwartz, 1963a; Cagan, 1965, pp. 296–8).

(11). A major unsettled issue is the short-run division of a change in nominal income between output and price. The division has varied widely over space and time and there exists no satisfactory theory that isolates the factors responsible for the variability (Gordon, 1980, 1981, 1982; Friedman and Schwartz, 1982, pp. 59–62).

(12). It follows from these propositions that *inflation is always and everywhere a monetary phenomenon* in the sense that it is and can be produced only by a more rapid increase in the quantity of money than in output. Many phenomena can produce temporary fluctuations in the rate of inflation, but they can have lasting effects only insofar as they affect the rate of monetary growth. However, there are many different possible reasons for monetary growth, including gold discoveries, financing of government spending, and financing of private spending. Hence, these propositions are only the beginning of an answer to the causes and cures for inflation. The deeper question is why excessive monetary growth occurs.

(13). Government spending may or may not be inflationary. It clearly will be inflationary if it is financed by creating money, that is, by printing currency or creating bank deposits. If it is financed by taxes or by borrowing from the public, the main effect is that the government spends the funds instead of the taxpayer or instead of the lender or instead of the person who would otherwise have borrowed the funds. Fiscal policy is extremely important in determining what fraction of total national income is spent by government and who bears the burden of that expenditure. It is also extremely important in determining monetary policy and, via that route, inflation. Essentially all major inflations, especially hyperinflations, have resulted from resort by governments to the printing press to finance their expenditures under conditions of great stress such as defeat in war or internal revolution, circumstances that have limited the ability of governments to acquire resources through explicit taxation.

(14). A change in monetary growth affects interest rates in one direction at first but in the opposite direction later on. More rapid monetary growth at first tends to lower interest rates. But later on, the resulting acceleration in spending and still later in inflation produces a rise in the demand for loans which tends to raise interest rates. In addition, higher inflation widens the difference between real and nominal interest rates. As both lenders and borrowers come to anticipate inflation, lenders demand, and borrowers are willing to offer, higher nominal rates to offset the anticipated inflation. That is why interest rates are highest in countries that have had the most rapid growth in the quantity of money and also in prices – countries like Brazil, Chile, Israel, South Korea. In the opposite direction, a slower rate of monetary growth at first raises interest rates but later on, as it decelerates spending and inflation, lowers interest rates. That is why interest rates are lowest in countries that *have had* the slowest rate of growth in the quantity of money – countries like Switzerland, Germany, and Japan.

(15). In the major Western countries, the link to gold and the resultant long-term predictability of the price level meant that until some time after World War II, interest rates behaved as if prices were expected to be stable and both inflation and deflation were unanticipated; the so-called Fisher effect was almost completely absent. Nominal returns on nominal assets were relatively stable; real returns

unstable, absorbing almost fully inflation and deflation.
(16). Beginning in the 1960s, and especially after the end of Bretton Woods in 1971, interest rates started to parallel rates of inflation. Nominal returns on nominal assets became more variable; real returns on nominal assets, less variable (Friedman and Schwartz, 1982, pp. 10–11).

6 Policy implications

On a very general level the implications of the quantity theory for economic policy are straightforward and clear. On a more precise and detailed level they are not.
Acceptance of the quantity theory means that the quantity of money is a key variable in policies directed at controlling the level of prices or of nominal income. Inflation can be prevented if and only if the quantity of money per unit of output can be kept from increasing appreciably. Deflation can be prevented if and only if the quantity of money per unit of output can be kept from decreasing appreciably. This implication is by no means trivial. Monetary authorities have more frequently than not taken conditions in the credit market – rates of interest, availability of loans, and so on – as criteria of policy and have paid little or no attention to the quantity of money per se. The emphasis on credit as opposed to the quantity of money accounts both for the great contraction in the United States from 1929 to 1933, when the Federal Reserve System allowed the stock of money to decline by one-third, and for many of the post-World War II inflations.
The quantity theory has no such clear implication, even on this general level, about policies concerned with the growth of real income. Both inflation and deflation have proved consistent with growth, stagnation, or decline.
Passing from these general and vague statements to specific prescriptions for policy is difficult. It is tempting to conclude from the close average relation between changes in the quantity of money and in money income that control over the quantity of money can be used as a precision instrument for offsetting other forces making for instability in money income. Unfortunately the loose relation between money and income over short periods, the long and variable lag between changes in the quantity of money and other variables, and the often conflicting objectives of policy-makers precludes precise offsetting control.
An international specie standard leaves only limited scope for an independent monetary policy. Over any substantial period, the quantity of money is determined by the balance of payments. Capital movements plus time delays in the transmission of monetary and other impulses leave some leeway, which may be more or less extensive, depending on the importance of foreign transactions for a country and the sluggishness of response. As a result, monetary policy under an effective international specie standard has consisted primarily of banking policy, directed towards avoiding or relieving banking and liquidity crises (Bagehot, 1873).
Until 1971, departures from an international specie standard, at least by major countries, took place infrequently and only at times of crisis. Surveying such episodes, Fisher concluded in 1911 that ‘irredeemable paper money has almost invariably proved a curse to the country employing it’ (1911, p. 131), a generalization that has applied equally to most of the period since, certainly up to 1971, and that explains why such episodes were generally transitory.
The declining importance of the international specie standard and its final termination in 1971 have changed the situation drastically. ‘Irredeemable paper money’ is no longer an expedient grasped at in

times of crisis; it is the normal state of affairs in countries at peace, facing no domestic crises, political or economic, and with governments fully capable of obtaining massive resources through explicit taxes. This is an unprecedented situation. We are in unexplored terrain.

As Keynes pointed out in 1923, monetary authorities cannot serve two masters: as he put it, ‘we cannot keep *both* our own price level *and* our exchanges stable. And we are compelled to choose’ (p. 126). Experience since has converted his dilemma into a trilemma. In principle, monetary authorities can achieve any two of the following three objectives: control of exchange rates, control of the price level, freedom from exchange controls. In practice, it has in fact proved impossible to achieve the first two by accepting exchange controls. Such controls have proved extremely costly and ultimately ineffective. The Bretton Woods system was ultimately wrecked on this trilemma. The attempts by many countries to pursue an independent monetary policy came into conflict with the attempt to maintain pegged exchange rates, leading to the imposition of exchange controls, repeated monetary crises, accompanied by large, discontinuous changes in exchange rates, and ultimately to the abandonment of the system in 1971. Since then, most countries have had no formal commitment about exchange rates, which have been free to fluctuate and have fluctuated widely. Nonetheless, Keynes's dilemma is still alive and well. Monetary authorities have tried to influence the exchange rates of their currency and, at the same time, achieve internal objectives. The result has been what has been described as a system of managed floating.

One recent strand of policy discussions has consisted of attempts to devise a substitute for the Bretton Woods arrangements that would somehow combine the virtues of exchange rate stability with internal monetary stability. For example, one proposal, by McKinnon (1984), is for the USA, Germany, and Japan to fix exchange rates among their currencies and set a joint target for the rate of increase of the total quantity of money (or high-powered money) issued by the three countries together. So far, no such proposal has gained wide support among either economists or a wider public.

A different strand of policy discussions has been concerned with the instruments, targets, and objectives of monetary authorities. One element of the quantity theory approach that has had considerable influence is emphasis on the quantity of money as the appropriate intermediate target for monetary policy. Most major countries now (1985) follow the practice of announcing in advance their targets for monetary growth. That is so for the USA, Great Britain, Germany, Japan, Switzerland, and many others. The record of achievement of the announced targets varies greatly – from excellent to terrible. Recently, a considerable number of economists have favoured the use of nominal income (usually nominal gross national product) as the intermediate target. The common feature is the quantity theory emphasis on nominal magnitudes.

A more abstract strand of policy discussions has been concerned with the optimum quantity of money: what rate or pattern of monetary growth would in principle promote most effectively the long-run efficiency of the economic system – meaning by that a Pareto welfare optimum. This issue turns out to be closely related to a number of others, in particular the optimum behaviour of the price level; the optimum rate of interest; the optimum stock of capital, and the optimum structure of capital (Friedman, 1969, pp. 1–50).

One widely accepted answer is based on the observation that no real resource cost need be incurred in increasing the real quantity of money since that can be done by reducing the price level. The implication is that the optimum quantity of money is that at which the marginal benefit from increasing the real quantity is also zero. Various arrangements are possible that will achieve such an objective, of which perhaps the simplest, if money pays no interest, is a pattern of monetary growth involving a decline in

the price level at a rate equal to the real interest rate (Mussa, 1977; Ihori, 1985).

This answer, despite its great theoretical interest, has had little practical consequence. Short-run considerations have understandably been given precedence to such a highly abstract long-run proposition.

Finally, there has been a literal explosion of discussion of the basic structure of the monetary system. One component derives from the belief that Fisher's generalization about irredeemable paper money will continue to hold for the present world fiat money system and that we are headed for a world monetary collapse ending in hyperinflation unless a specie (gold) standard is promptly restored. In the United States, this monetary belief was powerful enough to lead Congress to establish a Commission on the Role of Gold. In its final report, ‘the Commission concludes that, under present circumstances, restoring a gold standard does not appear to be a fruitful method for dealing with the continuing problem of inflation. … We favour no change in the flexible exchange rate system’ (Commission, 1982, vol. 1, pp. 17, 20). The testimony before the Commission revealed that agreement on a ‘gold standard’ concealed wide differences in the precise meaning of the phrase, varying from a system in which money consisted of full-bodied gold or warehouse receipts for gold to one in which the monetary authorities were instructed to regard the price of gold as one factor affecting their policy.

A very different component of the discussion has to do with possible alternatives to gold as a long-term anchor to the price level. This include proposals for subjecting monetary authorities to more specific legislative or constitutional guidelines, varying from guidelines dealing with their objectives (price stability, rate of growth of nominal income, real interest rate, etc.) to guidelines specifying a specific rate of growth in money or high-powered money. Perhaps the most widely discussed proposal along this line is the proposal for imposing on the authorities the obligation to achieve a constant rate of growth in a specified monetary aggregate (Friedman, 1960, pp. 92–5; Commission, 1982, vol. 1, p. 17). Other proposals include freezing the stock of base money and eliminating discretionary monetary policy, and denationalizing money entirely, leaving it to the private market and a free banking system (Friedman, 1984; Friedman and Schwartz, 1986; Hayek, 1976; White, 1984a).

Finally, a still more radical series of proposals is that the unit of account be separated from the medium of exchange function, in the belief that financial innovation will establish an efficient payment system dispensing entirely with the use of cash. The specific proposals are highly sophisticated and complex, and have been sharply criticized. So far, their value has been primarily as a stimulus to a deeper analysis of the meaning and role of money. (For the proposals, see Black, 1970; Fama, 1980; Hall, 1982a, 1982b; Greenfield and Yeager, 1983; for the criticisms, see White, 1984b; McCallum, 1985).

One thing is certain: the quantity theory of money will continue to generate agreement, controversy, repudiation, and scientific analysis, and will continue to play a role in government policy during the next century as it has for the past three.

Bibliography

Angell, J.W. 1936. *The Behavior of Money*. New York: McGraw-Hill.

Bagehot, W. 1873. *Lambard Street*. London: Henry S. King.

Barnett, W.A., Offenbacher, E.K. and Spindt, P.A. 1984. The new Divisia monetary aggregates. *Journal of Political Economy* 92, 1049–85.

Baumol, W.J. 1952. The transactions demand for cash: an inventory theoretic approach. *Quarterly Journal of Economics* 66, 545–56.

Black, F. 1970. Banking and interest rates in a world without money: the effects of uncontrolled banking, *Journal of Bank Research* 1(3), 2–20.

Black, H. 1975. The relative importance of determinants of the money supply: the British case. *Journal of Monetary Economics* 1(2), 25–64.

Bordo, M.D. and Schwartz, A.J., eds. 1984. *A Retrospective on the Classical Gold Standard, 1821–1931*. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Burger, A.E. 1971. *The Money Supply Process*. Belmont: Wadsworth.

Cagan, P. 1965. *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. New York: Columbia University Press for the National Bureau of Economic Research.

Cagan, P. 1972. *The Channels of Monetary Effects on Interest Rates*. New York: National Bureau of Economic Research.

Cairnes, J.E. 1873. Essays on the gold question. In J.E. Cairnes, *Essays in Political Economy*, London: Macmillan.

Cipolla, C.M. 1956. *Money, Prices, and Civilization in the Mediterranean World, Fifth to Seventeenth Century*. Princeton: Princeton University Press.

Commission on the Role of Gold in the Domestic and International Monetary Systems. 1982. *Report to the Congress*, March. Washington, DC: The Commission.

Duck, N.W. 1985. Money, output and prices: an empirical study using long-term cross country data. Working Paper, University of Bristol, September.

Fama, E.F. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6(1), 39–57.

Feavearyear, A.E. 1931. *The Pound Sterling: a History of English Money*. 2nd edn., Oxford: Clarendon Press, 1963.

Fischer, S., ed. 1980. *Rational Expectations and Economic Policy*. Chicago: University of Chicago Press

for the National Bureau of Economic Research.

Fischer, I. 1896. *Appreciation and Interest*. New York: American Economic Association.

Fisher, I. 1911. *The Purchasing Power of Money*. 2nd revised edn, 1926; reprinted New York: Kelley, 1963.

Fisher, I. 1919. Money, prices, credit and banking. *American Economic Review* 9, 407–9.

Frenkel, J.A. 1976. Adjustment mechanisms and the monetary approach to the balance of payments. In *Recent Issues in International Monetary Economics*, ed. E. Claassen and P. Salin. Amsterdam: North-Holland.

Frenkel, J.A. and Johnson, H.G. 1976. The monetary approach to the balance of payments: essential concepts and historical origins. In *The Monetary Approach to the Balance of Payments*, ed. J.A. Frenkel and H.G. Johnson. Toronto: University of Toronto Press.

Friedman, M. 1956. The quantity theory of money – a restatement. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press for the National Bureau of Economic Research.

Friedman, M. 1959. The demand for money: some theoretical and empirical results. *Journal of Political Economy* 67, August, 327–51. Reprinted as Occasional Paper No. 68, New York: National Bureau of Economic Research, and in Friedman (1969).

Friedman, M. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, M. 1961. The lag in effect of monetary policy. *Journal of Political Economy* 69, 447–66. Reprinted in Friedman (1969).

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17. Reprinted in Friedman (1969).

Friedman, M. 1969. *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

Friedman, M. 1977. Inflation and unemployment (Nobel lecture). *Journal of Political Economy* 85, 451–72.

Friedman, M. 1984. Monetary policy for the 1980s. In *To Promote Prosperity: U.S. domestic policy in*

the mid-1980s, ed. J.H. Moore. Stanford: Hoover Institution Press.

Friedman, M. and Schwartz, A.J. 1963a. Money and business cycles. *Review of Economics and Statistics* 45(1), Supplement, February, 32–64. Reprinted in Friedman (1969).

Friedman, M. and Schwartz, A.J. 1963b. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press for the National Bureau of Economic Research.

Friedman, M. and Schwartz, A.J. 1970. *Monetary Statistics of the United States*. New York: Columbia University Press for the National Bureau of Economic Research.

Friedman, M. and Schwartz, A.J. 1982. *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867–1975*. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Friedman, M. and Schwartz, A.J. 1986. Has government any role in money? *Journal of Monetary Economics* 17(1), 37–62.

Gordon, R.J. 1980. A consistent characterization of a near-century of price behavior. *American Economic Review* 70, 243–49.

Gordon, R.J. 1981. Output fluctuations and gradual price adjustment. *Journal of Economic Literature* 19, 493–530.

Gordon, R.J. 1982. Price inertia and policy ineffectiveness in the United States, 1890–1980. *Journal of Political Economy* 90, 1087–117.

Greenfield, R.L. and Yeager, L.B. 1983. A laissez-faire approach to monetary stability. *Journal of Money, Credit, and Banking* 15, 302–15.

Haberler, G. 1941. *Prosperity and Depression*. 3rd edn., Geneva: League of Nations.

Hall, R.E. 1982a. Explorations in the gold standard and related policies for stabilizing the dollar. In *Inflation: Causes and Effects*, ed. R.E. Hall. Chicago: University of Chicago Press.

Hall, R.E. 1982b. ‘Monetary trends in the United States and the United Kingdom’: a review from the perspective of new developments in monetary economics. *Journal of Economic Literature* 20, 1552–6.

Hamilton, E.J. 1934. *American Treasure and the Price Revolution in Spain, 1501–1650*. Harvard Economic Studies, vol. 43, New York: Octagon, 1965.

Hansen, A. 1957. *The American Economy*. New York: McGraw-Hill.

Hayek, F.A. 1976. *Denationalization of Money*. 2nd extended edn., London: Institute of Economic Affairs, 1978.

Hume, D. 1752. Of interest; of money. In *Essays, Moral, Political and Literary*, vol. 1 of *Essays and Treatises*, a new edn, Edinburgh: Bell and Bradfute, Cadell and Davies, 1804.

Humphrey, T.M. 1984. Algebraic quantity equations before Fisher and Pigou. *Economic Review*, Federal Reserve Bank of Richmond 70(5), September–October, 13–22.

Ihori, T. 1985. On the welfare cost of permanent inflation. *Journal of Money, Credit, and Banking* 17(2), 220–31.

Jevons, W.S. 1863. A serious fall in the value of gold. In *Investigations in Currency and Finance*, 2nd edn. London: Macmillan, 1909.

Johnson, H.G. 1961. The General Theory after twenty-five years. *American Economic Association, Papers and Proceedings* 51, 1–17.

Judd, J.P. and Scadding, J.L. 1982. The search for a stable money demand function. *Journal of Economic Literature* 20, 993–1023.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. Reprinted London: Macmillan for the Royal Economy Society, 1971.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. Reprinted London: Macmillan for the Royal Economic Society, 1973.

Knoester, A. 1984. Pigou and buffer effects in monetary economics. Discussion Paper 8406 G/M, Institute for Economic Research, Erasmus University, Rotterdam.

Laidler, D. 1984. The ‘buffer stock’ notion in monetary economics. *Economic Journal* 94, Supplement, 17–34.

Laidler, D. 1985. *The Demand for Money: theories, evidence, and problems*. 3rd edn., New York: Harper & Row.

Lothian, J.R. 1985. Equilibrium relationships between money and other economic variables. *American Economic Review* 75, 828–35.

Lucas, R.E., Jr. 1976. Econometric policy evaluation: a critique. *Journal of Monetary Economics* supplementary series 1, 19–46.

Lucas, R.E., Jr. 1981. Tobin and monetarism: a review article. *Journal of Economic Literature* 19, 558–67.

Lucas, R.E., Jr. and Sargent, T.J., eds. 1981. *Rational Expectations and Economic Practice*. 2 vols, Minneapolis: University of Minnesota Press.

McCallum, B. 1985. Bank deregulation, accounting systems of exchange and the unit of account: a critical review. *Carnegie-Rochester Conference Series on Public Policy* 23, Autumn.

McKinnon, R. 1984. *An International Standard for Monetary Stabilization*. Cambridge, MA: MIT Press.

Meiselman, D., ed. 1970. *Varieties of Monetary Experience*. Chicago: University of Chicago Press.

Mill, J.S. 1844. Review of books by Thomas Tooke and R. Torrens. *Westminster Review*, June.

Miller, M.H. and Orr, D. 1966. A model of the demand for money by firms. *Quarterly Journal of Economics* 80, 413–35.

Miller, M.H. and Orr, D. 1968. The demand for money by firms: extensions of analytical results. *Journal of Finance* 23, 735–59.

Modigliani, F. 1977. The monetarist controversy, or should we forsake stabilization policies? *American Economic Review* 67, 1–19.

Mussa, M. 1977. The welfare cost of inflation and the role of money as a unit of account. *Journal of Money, Credit, and Banking* 9(2), 276–86.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, July, 315–35. Reprinted in Lucas and Sargent (1981).

Newcomb, S. 1885. *Principles of Political Economy*. New York: Harper & Brothers.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38, September, 543–64. Revised and reprinted in F.A. Lutz and L.W. Mints (American Economic Association), *Readings in Monetary Theory*, Homewood, IL: Irwin, 1951.

Phelps, E.S. 1967. Phillips curves, expectations of inflation, and optimal unemployment over time. *Economica* 34(135), 254–81.

Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32, November, 38–65. Reprinted in F.A. Lutz and L.W. Mints (American Economic Association), *Readings in Monetary Theory*, Homewood, IL: Irwin, 1951.

Pigou, A.C. 1947. Economic progress in a stable environment. *Economica* 14(55), 180–88.

Sargent, T.J. 1982. The ends of four big inflations. In *Inflation: Causes and Effects*, ed. R.E. Hall. Chicago: University of Chicago Press.

Snyder, C. 1934. On the statistical relation of trade, credit, and prices. *Revue de l'Institut International de Statistique* 2, 278–91.

Spindt, P.A. 1985. Money is what money does: monetary aggregation and the equation of exchange. *Journal of Political Economy* 93, 1975–2204.

Tobin, J. 1947. Money wage rates and employment. In *The New Economics*, ed. S. Harris. New York: Knopf.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38, 241–7.

Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25, 65–86.

Tobin, J. 1961. Money, capital and other stores of value. *American Economic Review, Papers and Proceedings* 51, 26–37.

Tobin, J. 1974. Friedman's theoretical framework. In *Milton Friedman's Monetary Framework: a Debate with His Critics*, ed. R.J. Gordon. Chicago: University of Chicago Press.

White, L.H. 1984a. *Free Banking in Britain: Theory, Experience and Debate, 1800–1845*. New York: Cambridge University Press.

White, L.H. 1984b. Competitive payments systems and the unit of account. *American Economic Review* 74, 699–712.

How to cite this article

Friedman, Milton. "quantity theory of money." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_Q000006> doi:10.1057/9780230226203.1374

The New Palgrave Dictionary of Economics Online

quasi-concavity

J.-P. Crouzeix

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Definition

A real function *f* defined on a convex subset *C* of a linear space *E* is said to be quasi-concave if

$$x, y \in C, t \in [0, 1] \Rightarrow f(tx + (1-t)y) \geq \mathrm{Min}[f(x), f(y)]. \quad (1)$$

A function *g* is said to be quasi-convex if – *g* is quasi-concave. Concave functions are quasi-concave, convex functions are quasi-convex.
For all $\lambda \in R$, let $S(\lambda) = \{x : x \in C, f(x) \geq \lambda\}$. These sets are termed as the upper level sets of *f*. Upper level sets play an essential role in quasi-concavity. In particular, an alternative and useful way of characterizing quasi-concavity is to say that a function *f* is quasi-concave if all its upper level sets are convex.
Like concave functions, quasi-concave functions enjoy nice properties: the set of maximizing points is convex, the infimum of a family of quasi-concave functions is quasi-concave, if *f* is quasi-concave and *k* is a real non-decreasing function on *R* then *k* O *f* is quasi-concave. But, in contrast, a quasi-concave function is not necessarily continuous on the interior of its domain and the sum of quasi-concave functions is not quasi-concave in general.
Quasi-concavity was pioneered by De Finetti (1949), Fenchel (1953), Arrow and Enthoven (1961), Mangasarian (1965). It occurs in consumer theory where, under reasonable assumptions, a consumer's preferences can be represented by a quasi-concave utility function. In producer theory, production functions can also be reasonably assumed to be quasi-concave.

Characterizations of quasi-concave differentiable functions

For the sake of simplicity, we assume hereafter that E is the n-dimensional space R^n and C is an open convex set of R^n. Assume that f is differentiable on C, then f is quasi-concave on C iff

$$x, y \in C,\ f(x) \leq f(y) \Rightarrow (y - x)\nabla f(x) \geq 0. \tag{2}$$

Because quasi-concavity is often concerned with maximization problems, it would be suitable for f to achieve its maximum at each x so that $\nabla f(x) = 0$, but this is not so. A slight change of (2) leads to the following definition: a differentiable function f defined on the open convex set C is said to be pseudoconcave if

$$x, y \in C,\ f(x) < f(y) \Rightarrow (y - x)\nabla f(x) > 0. \tag{3}$$

Pseudoconcave functions are quasi-concave and concave functions are psuedoconcave. Actually, pseudoconcave functions are those quasi-concave functions which achieve their maximum at each point so that $\nabla f(x) = 0$.
Now, assume that f is twice differentiable on C. Then f is pseudoconcave iff

$$\text{(a)}\quad x \in C,\ v\nabla f(x) = 0 \Rightarrow v\nabla^2 f(x)v \leq 0,$$

and

$$\text{(b)}\quad \text{if } x \in C,\ \nabla f(x) = 0.$$

then f achieves its maximum at x. Similarly, f is quasi-concave iff (a) holds and (b′) if $\nabla f(x) = 0$ then for all $h \in R^n$ the function $t \to f(x + th)$ is quasi-concave.
Condition (a) can be formulated alternatively in terms of the bordered hessian of the function.

Duality

Assume that f is continuous and quasi-concave on the closed convex subset C *of* R^n, define

$$F(p, \lambda) = \mathrm{Inf}\,[px : x \in C,\ f(x) \geq \lambda],$$

and

$$G(p, r) = \mathrm{Sup}\,[f(x) : x \in C,\ px \leq r].$$

Then

$$f(x) = \mathrm{Inf}_p\,[G(p, px)] = \mathrm{Inf}_p \mathrm{Sup}_\lambda\,[\lambda : F(p, \lambda) \leq px].$$

Thus, f can be generated from either F or G. The functions F and G each have a useful economic interpretation. Assume that C is the set of available commodities, p is the vector of commodity prices and the consumer's preferences are given by the utility functions f. Then $F(p, \lambda)$ is the minimal cost to be paid by the consumer to get a value of utility greater than or equal to λ and $G(p, r)$ is the maximal value of utility that he can get for expenditure r. The function $p \rightarrow G(p,1)$ is sometimes called the indirect utility function (Diewert, 1981, Crouzeix, 1983).

Concavifiability

If a consumer's preferences are represented by a continuous utility function u, then $f=k \circ u$ when k is a real continuous increasing function on R is also a continuous utility function that represents these preferences; and all utility functions are of this kind. In many economic problems, it is extremely useful to know if there exists a concave utility function, but this is not true in general. The problem was initiated by De Finetti (1949) and Fenchel (1953), a recent reference is Kannai (1981). Dual functions are extremely useful here, thanks to the following characterization to concave functions: a continuous quasi-concave function f is concave iff for all p the function $\lambda \rightarrow F(p, \lambda)$ is convex or alternatively iff for all p the function $r \rightarrow G(p, r)$ is concave.
If preferences can be represented by a concave utility function, then they admit a concave utility function u such that for any other concave utility function v there exists a real increasing concave function k such that $v=k \circ u$. This function is called (Debreu, 1976) a least concave utility. Least concave utility functions are useful in the context of decision making under uncertainty.

Additivity

The sum of quasi-concave functions is not quasi-concave, even when each are separable. Assume that the set *C* of commodities and a consumer's utility function have the following form

$$C = C_1 \times C_2 \times \cdots \times C_p, u(x_1, x_2, \ldots, x_p) = u_1(x_1) + u_2(x_2) + \cdots + u_p(x_p) x_i C_i.$$

Assume that *u* is quasi-concave on *C* and that the functions u_i, are not, constant. Then all functions u_i are concave except perhaps one which is concavifiable (Debreu and Koopmans, 1982; Crouzeix and Lindberg, 1986). Quasi-concavity and additivity cannot be brought together without problems.

Bibliography

An important and up to date discussion of quasiconcavity and related topics with their applications for economics as well as for mathematical programming can be found in *Generalized Concavity in Optimization and Economics*, a collection of papers by several authors edited by S. Schaible and W.T. Ziemba (New York: Academic Press, 1981).

Arrow, K.J. and Enthoven, A.C. 1961. Quasi-concave programming. *Econometrica* 29(4), October, 779–800.

Crouzeix, J.P. 1983. Duality between direct and indirect utility functions. *Journal of Mathematical Economics* 12(2), 149–65.

Crouzeix, J.P. and Lindberg, P.O. 1986. Additively decomposed quasi-convex functions. *Mathematical Programming* 35(l), 42–57.

Debreu, G. 1976. Least concave utility functions. *Journal of Mathematical Economics* 3(2), 121–29.

Debreu, G. and Koopmans, T.C. 1982. Additively decomposed quasi-convex functions. *Mathematical Programming* 24(l), 1–38.

De Finetti, B. 1949. Sulle stratificazioni convesse. *Annali di matematica pura ed applicata*, Series IV, 30, 173–83.

Diewert, W.E. 1981. Generalized concavity in economics. In *Generalized Concavity in Optimization and Economics*, ed. S. Schaible and W.T. Ziemba. New York: Academic Press.

Fenchel, W. 1953. Convex cones, sets and functions. Mimeo, Princeton University.

Kannai, Y. 1981. Concave utility functions. In *Generalized Concavity in Optimization and Economics*, ed. S. Schaible and W.T. Ziemba. New York: Academic Press.

Mangasarian, O.L. 1965. Pseudo-convex functions. *SIAM Journal on Control* 3(2), 281–90.

How to cite this article

Crouzeix, J.- P. "quasi-concavity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_Q000008> doi:10.1057/9780230226203.1375

The New Palgrave Dictionary of Economics Online

Quesnay, François (1694 – 1774)

G. Vaggi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

avances ; Baudeau, N.; capital accumulation; circulation of commodities; Du Pont de Nemours, P. S.; effective demand; *Ephémérides du citoyen* ; free trade; Mercier de La Rivière, P.-P.; Mirabeau, Marquis de; net product; Petty, W.; Physiocracy; price determination; price theory; *prix fundamental* ; Quesnay, F.; single tax; sterile class; surplus; *Tableau économique* ; Turgot, A. R. J.

Article

Quesnay was born at Mère, Seine-et-Oise. He came from a family of humble origin, the eighth of 13 children. His father Nicholas was a small merchant, and the family also had a piece of land; thanks to these two activities they were comfortably off. François Quesnay had no systematic education; at ten he could not even read, but early on he developed an interest in medicine. In 1711 he went to Paris for formal training in medicine and surgery. There he read Descartes and Malebranche, and the latter's *Recherche de la verité* had a profound impact on the young Quesnay. In 1717 he married Jeanne-Cathérine Dauphin, who gave him four children, two of whom survived. He began his career at Mantes, a small town not far from Paris, and in the 1720s and 1730s he made his reputation as a surgeon, particularly with respect to bleeding techniques. In 1736 he published *l'Essai physique sur l'oeconomie animale*, his first major work. Quesnay was deeply involved in the polemic between surgeons and physicians which took place in the 1740s. At that time he was also physician to the Duke of Villeroy and through him and the Comtesse d'Estrades he met Madame de Pompadour, Louis XV's favourite. Quesnay became her private physician and established himself at Versailles. In 1752 he saved the Dauphin from smallpox, and in gratitude the King granted Quesnay a noble title and a sum of money which he used to buy an estate at Beauvoir in the Minervois for his son Blaise-Guillaume. In 1750 and 1751 Quesnay published the last of his medical works and became a member of the French Académie des Sciences and of the Royal Society in London.

In the early 1750s Quesnay became interested in economics and in particular in agricultural matters. He was in contact with many important thinkers including D'Alembert and Diderot, Buffon, Helvetius and Condillac. He was induced to contribute to the *Encyclopédie*, for which he wrote two articles: *Evidence*, which appeared in Volume VI in January 1756, and *Fonction de 1'âme* which was never published there since the *Encyclopédie* had been condemned by the government. Quesnay preferred to publish his articles anonymously. In 1756 he also published his article *Fermiers*, while in 1757 he wrote *Grains* for the seventh volume of the *Encyclopédie*. These two articles are Quesnay's first economic writings. He wrote three more pieces for the *Encyclopédie: Hommes*, *Impôts* and *Intérêt de l'argent*. But after the attack on the King's life at Damiens, the enemies of the *encyclopédistes* managed to obtain the repeal of its royal privilege and Quesnay, like Turgot, withdrew his three articles from publication. The third appeared 1766 in the *Journal de l'agriculture; Hommes* was published in 1908 by Etienne Bauer and *Impôts* appeared in 1902 thanks to Gustave Schelle.

In 1757 Quesnay met Victor Riquetti, Marquis de Mirabeau (see Hecht, 1958, p. 256). Mirabeau came from an old noble family and supported the view that the main cause of national wealth was the number of people. Quesnay convinced him that agriculture was much more important than population, because it produced the commodities which were necessary for men's subsistence, Mirabeau became the most faithful propagator of Quesnay's ideas, and this episode marks the beginning of the Physiocratic school.

TABLEAU ECONOMIQUE

Objects to considered: 1 three kinds of expenditure; (2) their source; (3) their advances;

(4) their distribution; (5) their effects; (6) their reproduction; (7) their relations with one another; (8) their relations with the population; (9) with agriculture; (10) with industry; (11) with trade; (12) with the total wealth of a nation.

PRODUCTIVE EXPENDITURE relative to agriculture, etc.	*EXPENDITURE OF THE REVENUE* after production of taxes, is divided between productive expenditure and sterile expenditure	*STERILE EXPENDITURE* relative to industry, etc.
Annual advances required to produce a revenue of 600^l are 600^l	**Annual** revenue	**Annual advances** for the works of sterile expenditure are

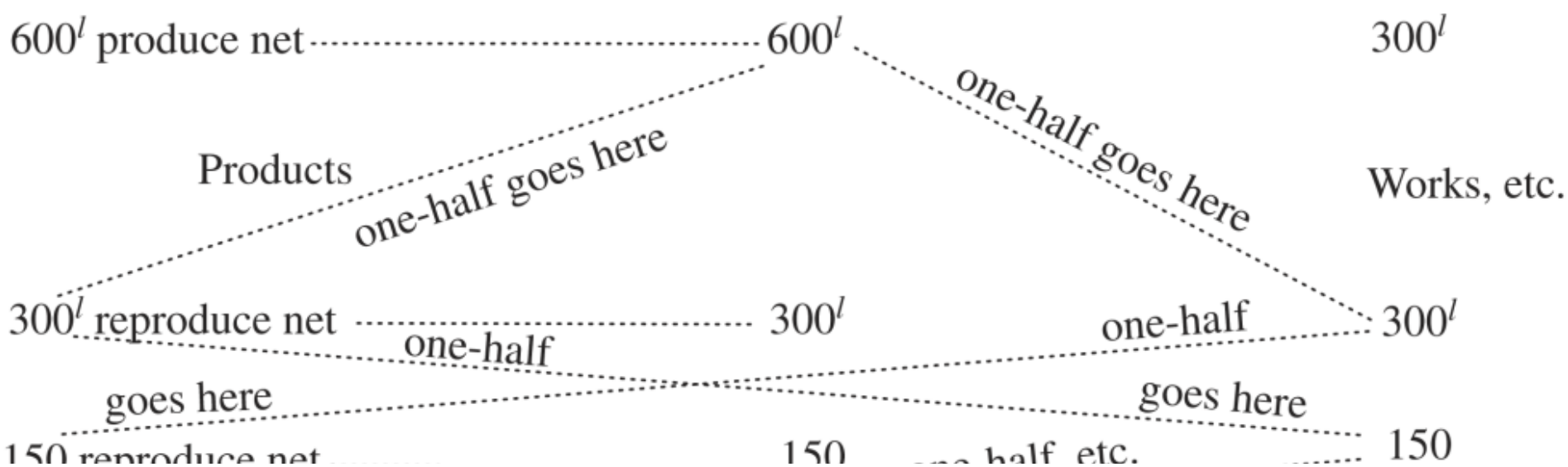

150 reproduce net one-half, etc.	150 one-half, etc.	
75 reproduce net	75	75
37..10^{s} reproduce net	37..10	37..10
18..15 reproduce net	18..15	18..15
9...7...6^{d} reproduce net	9...7...6^{d}	9...7...6^{d}
4...13...9 reproduce net	4...13...9	4...13...9
2... 6..10 reproduce net	2... 6..10	2... 6..10
1... 3... 5 reproduce net	1... 3... 5	1... 3... 5
0...11... 8 reproduce net	0...11... 8	0...11... 8
0... 5... 10 reproduce net	0... 5... 10	0...5... 10

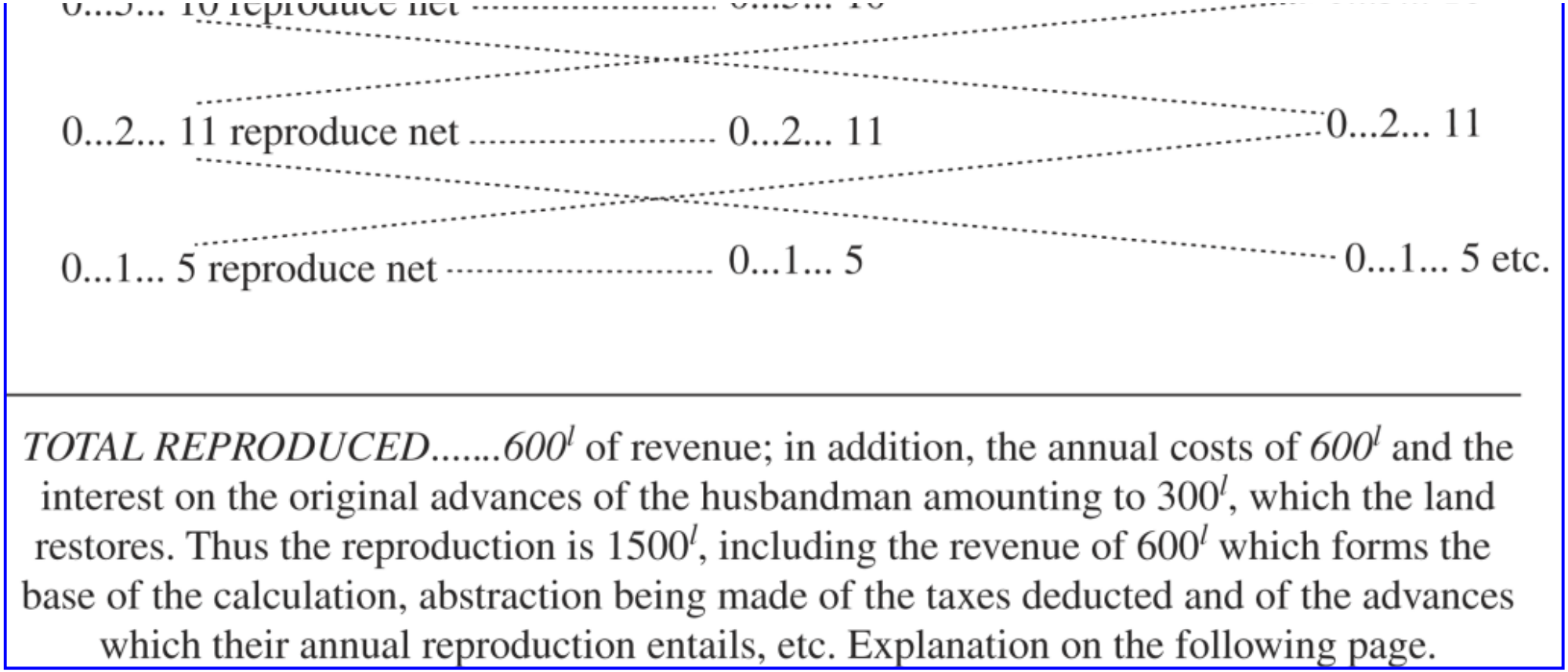

In 1758 Quesnay published the *Questions intéressantes sur la population, l'agriculture et le commerce* and at the end of the same year he wrote the first edition of the *Tableau économique*. In the first half of 1759 this was followed by two other editions. The *Tableau*, printed in a limited edition at Versailles, was presented to Louis XV, who was apparently greatly impressed by Quesnay's strange schemes. Quesnay was also engaged in supervising Mirabeau's writings, which were designed to illustrate the rather obscure *Tableau*, and, more generally, to spread Physiocratic doctrine. Mirabeau's *Theorie de l'impôt*, the result of this close collaboration, appeared at the end of 1760. The reactions to this work and Mirabeau's consequent imprisonment convinced the two that it was better to work silently. Almost three years elapsed before Mirabeau published the *Philosophie rurale* in three volumes, in November 1763. There can be no doubt that Quesnay revised the entire work and wrote some chapters (Meek, 1962, p. 38).

The year 1763 marked the beginning of a period of active intervention by the Physiocrats in economic debates. Quesnay found new followers: Du Pont de Nemours, Mercier de La Rivière, Baudeau and Turgot, a good friend of the Physiocrats. Quesnay contributed to the development of Physiocratic ideas with articles which appeared in the *Journal de l'agriculture* and after 1767 in the *Ephémérides du citoyen*. Among other works Quesnay wrote *Le droit naturel*, the *Mémoires sur les avantages de l'industrie et du commerce* and the *Dialogue sur les travaux des artisans*. In these latter works he discusses the sterility of industry and trade and the productivity of agriculture. The most famous of Quesnay's articles appeared in the *Journal de l'agriculture* of 1766 with the title *Analyse de la formule arithmétique du Tableau économique*, in which he presented a simplified version of the *Tableau*. This new *Tableau* was also used by Quesnay in the *Premier problèmes économique*, which appeared in the same periodical. Between 1767 and 1768 Quesnay published several articles in the *Ephémérides: Despotisme de la Chine, Analyse du gouvernment des Incas du Pérou*. In 1767 he also wrote the *Second problème économique* for the *Physiocratie*, a collection of his main writings prepared by Du Pont de Nemours. Between 1764 and 1767 Quesnay was the true master of the Physiocrats; new disciples joined the group and his ideas found some application in French economic policy. By 1768 the cultural and political impact of Physiocracy began to decline and Quesnay's theory was bitterly criticized. He spent his last years studying geometry, notwithstanding the advice of his friends and the fact that he was ridiculed by some of his enemies (Hecht, 1958, pp. 278–9). Quesnay died in December 1774 at Grand-Commun, a place not far away from Versailles, a few months after the death of Louis XV.

Quesnay was directly responsible for all the main aspects of Physiocratic thought and, in particular, for its economic analysis. Some Physiocrats contributed to the development and the explanation of particular aspects of the doctrine, but Quesnay was the one who put forward the most innovative concepts and the general framework in which they were inserted. The analysis of Quesnay's economic writings presents a peculiar problem; he did not write a single major text as a summary of his entire thought, but instead wrote small pieces,

articles for the *Encyclopédie* and for the periodicals which were controlled by the Physiocrats (Vaggi, 1987, ch. 2, part 2). Thus usually he discussed only one economic issue at the time. This methodology is quite clear in his first works, where he presents the major features of Physiocratic economics, such as the role of farmers and capital in agriculture; the importance of commercial policies and free trade, the question of the fiscal system. But the same method of analysis characterizes the writings of the period 1766–8, where Quesnay had to defend his theory from the criticisms of many important thinkers of the time.

Correct interpretation of Quesnay's economics must take all his writings into account and cannot be limited to the analysis of the most famous. There are clear logical links between Quesnay's different works. On several occasions he himself indicates where these connections are to be found. Indeed, despite his method of writing on one specific issue at a time, the analytical structure of Quesnay's economics is clearly characterized by systematic cause-and-effect relationships. It is an important feature of Quesnay's thought that he explained economic facts and individual actions on the basis of a view of society as a general interrelated system. Therefore, he believed that all aspects of social life were linked together and that it was possible to find the underlying causal relationships, which were nothing else than the outcome of natural laws.

The role and significance of the *Tableau économique* within Quesnay's economics needs further clarification since this is the most important and famous work of Physiocracy and has often been regarded as a summary of the entire corpus of Physiocratic economics (Herlitz, 1961, p. 134; Fox-Genovese, 1976, p. 258). The *Tableau* has also been regarded as the analytical synthesis of the logical structure of Quesnay's economics, or at least as its most relevant aspect. In any case, all too often knowledge of Physiocracy is limited to the *Tableau économique*. However, the various types of *Tableau* elaborated by Quesnay between 1758 and 1766 cannot be analysed in isolation; by themselves they do not provide an exhaustive presentation of Physiocratic economics. An accurate understanding of Quesnay's theory and of the *Tableau* itself requires the study of his other writings. Some of these are particularly significant because they supplement the *Tableau*. For instance, Quesnay's first economic articles, *Hommes*, *Fermiers* and *Grains*, were written just before the *Tableau* and can be regarded as the basis on which the analysis of the *Tableau* is carried out. Quesnay's economics must be regarded as a mosaic, where all the inlays are necessary, though the *Tableau* is the central part of the picture.

The *Tableau économique* is one of those works in the history of economics which have often been regarded as an anticipation of modern theories. The *Tableau* has been considered a first rough presentation of Keynes's multiplier and as a sort of general equilibrium system of a Walrasian type (Schumpeter, 1954, p. 242). For others, the *Tableau* is an input–output table (Phillips, 1955, pp. 137–8). Because of the *Tableau*, Quesnay has been regarded as an early econometrician. The *Tableau* has also been interpreted as the first classical system of price determination, thus anticipating Marx's reproduction schemes and Sraffa's price system (Cartelier, 1976, p. 57).

There are two main reasons why the *Tableau* impressed Quesnay's contemporaries and later interpreters so much. The first is the ‘obscurity’ of the schemes, the second is the fact that there is not just one *Tableau* but many. The history of the *Tableau économique* begins at the end of 1758 with the first manuscript edition of the work, which included a table and a few pages of comments entitled *Remarques sur les variations de la distribution des revenus annuels d'une nation* comprising 22 remarks. The second edition, the first to be printed, followed a few months later (Kuczynski and Meek, 1972, pp. xvi–xviii). This *Tableau* is similar to the earlier one, but is now followed by 23 remarks which are very similar to the *Maximes générales* at the end of the 1757 article *Grains* and are entitled *Extrait des oeconomies royales de M. de Sully*. The third edition appeared in 1759, but then it disappeared for more than two centuries. It was rediscovered and published in 1965 by Marguerite Kuczynski (Kuczynski and Meek, 1972, pp. xxv ff.). The third edition of the *Tableau* is made up of the table itself plus an enlarged version of the *Extrait* with 24 remarks and long footnotes and a new text called *Explications du Tableau économique*. Clearly, Quesnay must have felt that more explanations of the scheme were required. All three editions present the original type of *Tableau*, which is characterized by three columns whose figures are related to each other by means of descending lines crossing the table. This is the so called zig-zag version of the *Tableau*.

A similar *Tableau* can also be found in the sixth part of Mirabeau's, *L'ami des hommes*, published in 1760 (Mirabeau, 1758–60, vol. 2, pp. 118 ff.). The two side-columns give the annual advances of the productive sector – agriculture – and of the sterile sector – industry. The central column presents the revenue of the landlords and the way in which it is spent. In the first row of each column Quesnay writes the value of each of the three magnitudes at the beginning of the process of circulation of commodities and revenue. These figures are characterized by some peculiar ratios; the revenue and the annual advances of the productive class have the same value; these annual advances are twice as large as those of the sterile class. All the zig-zag *Tableaux* and most of the following ones have these ratios. Among the figures in the first row, particularly important is the value assumed by revenue, which is usually called the ‘basis’ of the *Tableau* because it determines all the other figures. The first zig-zag started with a revenue of 400 *livres*, while the next two editions had a revenue of 600 *livres*. The *Tableau* shows the effects of the expenditure of revenue on the other two classes. Quesnay generally assumes that landlords spend half of their revenue in the purchase of agricultural products and half in the purchase of manufactures (Eltis, 1984, pp. 20 ff.). According to Quesnay, all classes comply their pattern of consumption to this 50–50 division. From the second row onwards, the *Tableau* describes the exchanges which take place between the two sectors of the economy, agriculture and industry. The workers of each of the two sectors spends half of the money received by the landlords in purchasing goods of the other class, while the other half is spent inside the same class. The *Tableau* registers the exchanges of money and commodities which take place between the classes, but it abstracts from the circulation of commodities between people of the same class.

According to Quesnay there is a major difference between the productive and the sterile class, only the former giving rise to a net product over costs. In the *Tableau* this is expressed as horizontal lines which connect the first column to the revenue column. Therefore the *Tableau* is a synthetic way of describing the circulation of money and commodities in relation to both the expenditure of revenue and the technical and social relationships between the two main sectors of the economy. In the *Tableau* there is a concise description of the way in which the process of circulation of revenue must guarantee the reproduction of the annual advances which have been consumed during the previous year by the two sectors. The *Tableau's* iterative process is completed when the peasants and the artisans have no money left to spend, in particular when the value of their receipts is equal to that of the annual

advances which have been used up. In fact, as a result of the sequence of exchanges, all three classes receive 600 *livres*, which is the sum of the figures from the second row to the bottom of each column. The productive and sterile classes have recovered their annual advances, and the landlords have got back their revenue. Hence the process of circulation of commodities depends upon the technical requirements of production, which are described by the reconstruction of the means of production in each sector, and by a particular rule in the distribution of income, which states that the revenue accrues entirely to the owners of land. At the end of the process of circulation there are all the conditions needed to start a new productive cycle.

The *Tableau* also serves as a representation, and perhaps to Quesnay's mind also as a proof, that a surplus exists only in agriculture. The first two columns from the right show that agriculture reproduces its means of production plus a revenue for the landlords, and, for this reason, is regarded as a productive sector. Some problems arise with respect to the third column. The overall production of the sterile class is 600 *livres* and its annual advances are only 300 *livres*. However, Quesnay says that industrial activities are unproductive because they do not contribute to the landlord's revenue. But it is clear from the figures of the Tableau that at the end of the process of circulation the sterile class has 300 *livres* in excess of its advances. This derives from the assumption that each class spends half of its revenue in the purchase of the products of the other sector. In order to reconcile this result with the opinion that only agriculture yields a surplus, it is necessary to resort to some further considerations. Some commentators have pointed out that the *Tableau* does not include one particular act of exchange: the purchase by the sterile class of an amount of agricultural products whose value is 300 *livres* (Meek, 1962, pp. 275–7).

If this act of exchange is included in the picture the sterile class has no surplus left. Alternatively it has been said that in the *Tableau* the sterile class sells part of its output abroad in exchange for agricultural goods (Meek, 1962, p. 283; Gilibert, 1977, pp. 42–5). In any case industry does not seem to be as sterile as the Physiocrats maintain.

A few years after the first version, Quesnay slightly modified the *Tableau économique*. In the 1764 *Philosophie rurale* there are several *tableaux*, some of them are still of the zig-zag type but others have evolved into a new scheme which is called *Précis des résultats de la distribution representée dans le Tableau*, that is to say a summary of the *Tableau* itself (Mirabeau, 1764, vol. 1, p. 327). In the *Précis* Quesnay changed the ‘basis’ of the *Tableau:* the revenue and the annual advances of the productive class are now 2,000 *livres*, the advances of the sterile class are still half this value. Above all there is no iterative process with the descending lines. Instead, Quesnay gives the final results of the circulation of commodities and of revenue. Only a few exchanges are indicated in the scheme and the first and last row are almost exactly alike; they show that the process of reproduction of revenue and both sectors' advances has been completed. The only difference between the first and last row is the annual output of industry, 2,000 *livres*, and its annual advances, 1,000 *livres*. Thus at first sight it would seem that the value of industrial output still exceeds that of its inputs. But there is an important difference between the zig-zag and the *Précis*; in this latter type of *Tableau* the advances of industry are entirely made up of agricultural commodities and do not include manufactures. Moreover, even though this does not appear in the diagram Quesnay now quite clearly states in the accompanying text that the artisans purchase 1,000 *livres* of raw materials to transform them into manufactured goods. Therefore industry now uses 2,000 *livres* of primary products to produce an output of equal value. There is no net product left.

Because of this new act of exchange between industry and agriculture, the output of the primary sector must rise in order to account for these 1,000 *livres* of raw materials which go to industry. At the same time, farmers must buy 1,000 *livres* of manufactured goods as a repayment for the raw materials which they sell to the artisans. In the *Précis* the farmers receive this 1,000 *livres* of industrial products as ‘interests’ on their original advances (Meek, 1962, pp. 278–9). The third edition of the zig-zag already included these interests, but it was not at all clear whether these products came from industry, or, more probably from agriculture itself (Eltis, 1984, p. 26). The existence of these 1,000 *livres* of industrial products which are purchased by the farmers is another reason why the *Tableau économique* cannot be regarded as a proof of the sterility of industry. The cultivation of land requires industrial goods, and this prevents agriculture from being considered as a self-sufficient sector. Industry produces part of the means of production of the primary sector, and therefore contributes, albeit indirectly, to the production of the net product.

The *Précis* is an intermediate step in the evolution of the *Tableau économique*, whose final version is the so called *Analyse de la formule arithmétique du Tableau économique* of 1766. This is also the most well-known version of the *Tableau* and it has been often confused with the original 1758 zig-zag. There are many similarities between the *Précis* and the *Analyse*. Both schemes give a concise representation of all transactions; and the advances of the sterile class are entirely made up of agricultural goods. But in the *Analyse* Quesnay explicitly includes the purchase of manufactured goods by farmers, as their ‘interests’. Moreover, the *Analyse* is a very well-written article, which condenses both the scheme and the texts needed to explain it in a few pages, while in the *Philosophie rurale*, the *Précis* and its explanations run into many pages. In the *Analyse*, Quesnay gives the final explanation of the technicalities of the *Tableau*. The formula has become the true *Tableau économique*; this is the version which has been converted into an input–output table (Phillips, 1955), and which has been extensively analysed by modern economists (Tsuru, 1942).

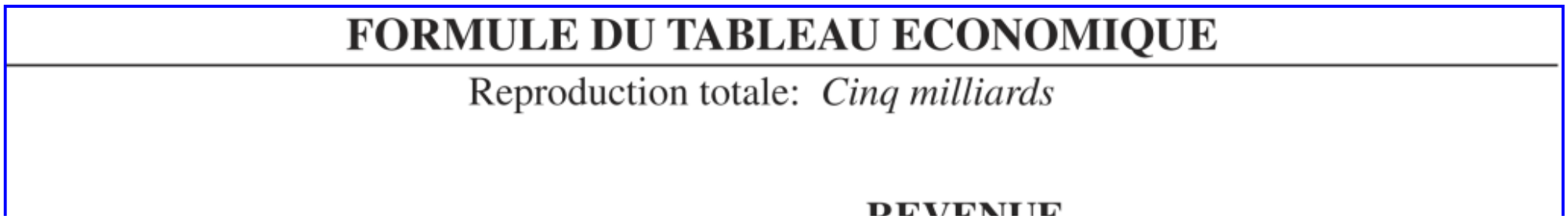

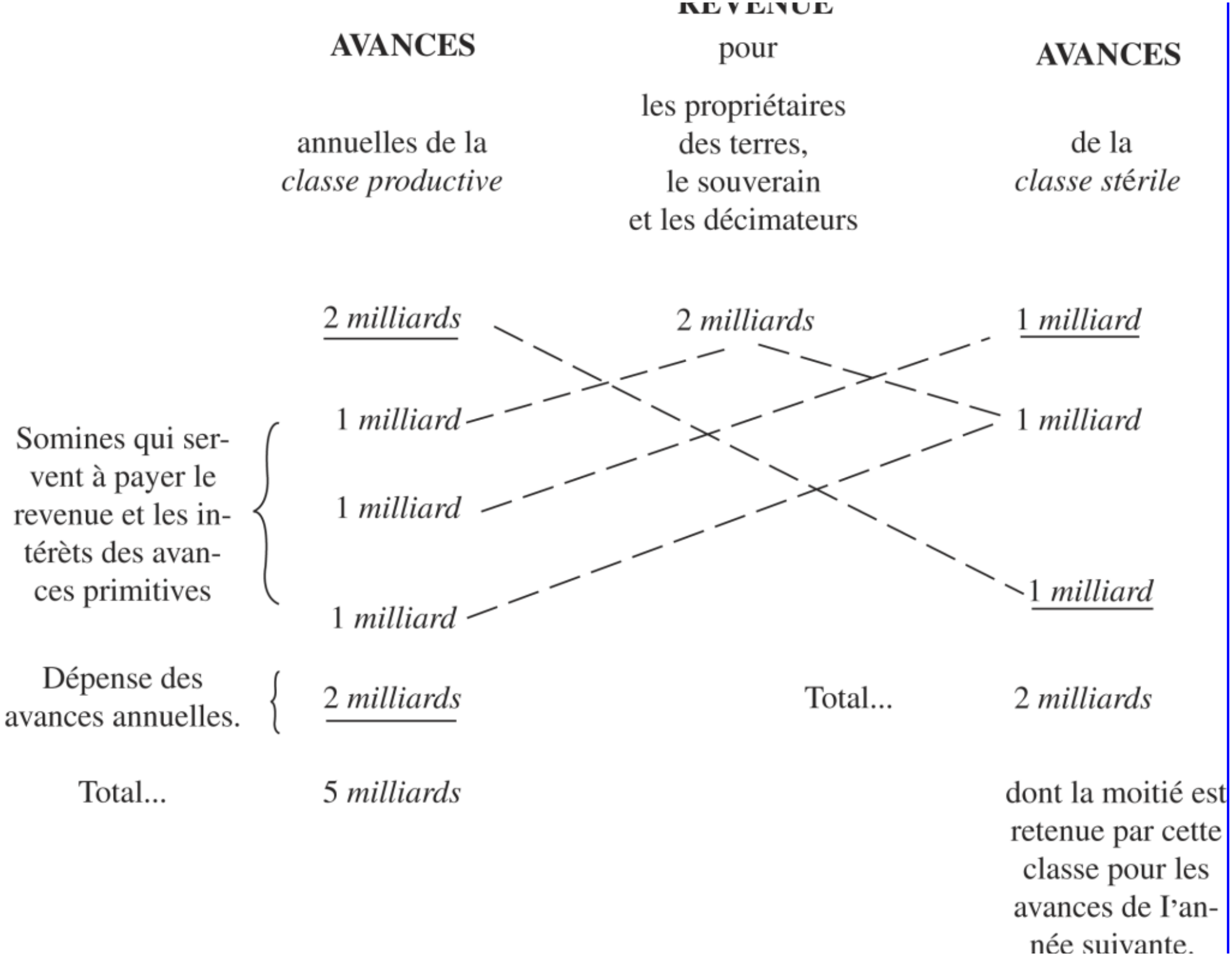
REVENUE
pour
les propriétaires
des terres,
le souverain
et les décimateurs
AVANCES
annuelles de la
classe productive
AVANCES
de la
classe stérile
2 milliards
2 milliards
1 milliard
1 milliard
1 milliard
1 milliard
1 milliard
1 milliard
Somines qui servent à payer le revenue et les intérèts des avances primitives
Dépense des avances annuelles.
2 milliards
Total...
2 milliards
Total...
5 milliards
dont la moitié est retenue par cette classe pour les avances de l'année suivante.

Si les propriétaires dépensaient plus à la *classe productive* qu'à la *classe stérile,* pour améliorer leurs terres et accroitre leurs revenus, ce surcroit.

The *Tableau* of the formula type works in the following way. Landlords receive two 'milliards' of *livres* from the farmers as their annual rent and spend this revenue half in the purchase of foodstuffs and half in that of manufactured goods. Artisans buy raw materials and necessaries of life from agriculture. Now the cultivators have the two milliards, one of which is used to buy the manufactured products which are necessary to maintain the initial value of the fixed capital of cultivation. The artisans need the necessities of life and one milliard goes back to the productive class. In the end one sees that industry has produced two milliards of products using two milliards of primary commodities as input. The total reproduction of agriculture is five milliards because the cultivators sell three milliards, one to the landlords and two to the artisans, but they also keep two milliards of output for themselves, to be used as raw materials and necessaries for future production. For Quesnay five 'milliards' is also the level of output in the whole economy, because the two milliards of manufactured goods are nothing more than reshaping of the same primary commodities which have been used as inputs. Of course this is a flaw in Quesnay's analysis; following his criterion, agricultural output should only be two milliards, because the overall means of production employed in agriculture amount to three milliards. Alternatively, the gross national product should include industrial output too, and its value would be seven milliards. This inconsistency in Quesnay's economics derives from his belief that industry cannot produce a surplus, a view by no means backed up by the figures of the formula-type *Tableau*, where it is quite clear that industrial products are used as inputs in agriculture (Meek, 1962, p. 154).

The *Analyse* has one further merit; from the very beginning Quesnay explicitly states all the hypotheses which characterize the economic system depicted in the *Tableau* (Meek, 1962, p. 298). The same assumptions can be found in Quesnay's comments and explanations in the other *Tableaux*, but in the *Analyse* they are grouped together in a few pages (Meek, 1962, pp. 151–3).

The main features of the economy described in the *Tableau* are as follows. First, in agriculture there are the best methods of cultivation, with large capital stock and high productivity, so that the annual advances can produce a surplus of the same size. A second assumption relates to the fiscal system; all duties and excises which exist in France ought to be substituted by a single tax on agricultural surplus. Third, free competition rules both in domestic and foreign trade in agricultural products, thus there is a *bon prix* for primary commodities and cultivation is a profitable activity. A fourth hypothesis relates to the landlords, who have made all the necessary ground advances, such as drainage, transport facilities etc. We could, of course, add other assumptions which clearly appear in the articles which precede the 1758 zig-zag. The State guarantees the ownership rights for all citizens and not only for landlords. In particular, the State protects the capital invested by cultivators as original advances (Kuczynski and Meek, 1972, pp. 7–8). Another assumption states that the landlords spend their revenue half in the purchase of foodstuffs and half in manufactures (p. 12). This hypothesis implies that in the *Tableau* the revenue is entirely consumed and is neither hoarded nor used to make financial investments, which are considered a sterile form of activity (pp. 4, 13).

It goes without saying that the society examined in the *Tableau* has little in common with the France of Louis XV. It is an ideal country where all reforms and economic measures advocated by the Physiocrats have already been implemented. This economic system is quite similar to the natural order of society. According to Quesnay, England is the country which most resembles this ideal society, and this explains why England is so prosperous and wealthy (see INED, 1958, vol. 2, pp. 474, 533). The *Tableau* is a normative benchmark for the government; it describes the circulation and distribution of the social product and surplus, and it shows the final effects of these exchanges on future production. The first zig-zag version of the *Tableau* (Table 1 above) and the *Analyse* are the most coherent descriptions of this ideal situation. In these works Quesnay highlights the normal and regular working conditions of an economic system which is a mirror of the natural order of society. From this point of view one could describe these *Tableaux* as types of equilibrium conditions, but no further similarity can be found with general equilibrium analysis (Meek, 1962, p. 292). But Quesnay is not interested in the conditions of logical consistency between *all* economic variables. He pinpoints *some specific* causal relationships which are regarded as particularly important for the development of the economy. Thus, even though the *Tableau* presents the mutual relationships between several aspects of the economy, Quesnay stresses the relationships of cause and effect affecting the increase in national wealth. In several of his writings Quesnay uses the *Tableau* to study what happens when government regulations or the behaviour of landlords does not conform to this natural order. For instance, in the 1767 article *Second problème économique* he analyses the consequences of several forms of indirect taxation, and shows that these taxes are ultimately damaging (Meek, 1962, pp. 186 ff.). In an earlier article, a formula *Tableau* had been used to illustrate the beneficial effects of an increase in the prices of agricultural products (Meek, 1962, p. 168). The *Tableau* is also used to examine what happens when the landlords modify their pattern of consumption. The proprietors can modify the level of surplus through their decisions to spend more or less revenue on the purchase of agricultural products (Mirabeau, 1764, vol. 3, pp. 33–53). Quesnay uses the zig-zag version and the *Précis* to show that the higher the proportion of revenue which is spent in the consumption of primary goods the higher will be the surplus. From these articles, it emerges that Quesnay was always primarily concerned with the effects of different economic measures on gross and net output. Far from regarding the *Tableau* as a static scheme, Quesnay uses it to show the government ways of speeding up economic growth.

The *Tableau économique* is the most original aspect of Quesnay's economics, but it does not exhaust his economic theory. On the contrary, the assumptions on which the *Tableau* is

built and, in general, the comments which supplement it show that the ideas of the Physiocrats are also to be found outside the *Tableau*. Many aspects of Quesnay's economics which are complementary to the *Tableau* are discussed in other writings. A peculiarity of the *Tableau* is the fact that it immediately conveys the view of the economy as a single system which must reproduce itself. The analysis of the conditions which have to be satisfied to guarantee this reproducibility is the main object of economic science. This analysis indicates the factors which lead to an increase of national wealth. The *Tableau* provides a schematic description of reproduction, which is based on the distinction between economic activities in a few major sectors. Moreover, social classes are distinguished according to their role in the process of production and expenditure of output and revenue. Thus there are two major sectors, agriculture and industry, and three main classes of people; two of them are identified with the above sector while the third is characterized by its ownership of the soil. In the *Tableau* Quesnay does not spend much time analysing another important social group, merchants, but this is due to the fact that in the *Tableau* natural order is assumed to rule. Professional traders should disappear in the ideal society where free competition rules in all markets (INED, 1958, vol. 2, pp. 941–2; Spengler, 1958, p. 62).

The power of professional traders is one of the issues which is not discussed in the *Tableau*, but is investigated in other writings. In the *Tableau* merchants are hidden inside the sterile class. Another limitation is that in the *Tableau* Quesnay does not separate salaried workers from master entrepreneurs. This problem is discussed in other works, in particular in his article *Fermiers*, where the cultivators are clearly described as capitalist entrepreneurs (INED, 1958, vol. 2, pp. 427, 483). But in Quesnay's economics there is no complete analysis of the capitalist relationship of production.

It must be noticed that the *Tableau*, and the entire economic theory of François Quesnay, is based on two main concepts, capital and net product. Quesnay uses the term *avances* to indicate all the commodities which exist at the beginning of the productive process and are the necessary inputs for production. These commodities are ‘advanced’ with respect to the output, but are part of the social product of the previous year. In order to satisfy the conditions of self-reproduction, the same commodities that are used up in production must also appear in the social product; otherwise production might come to a stop, or at the very least the level of activity would lower.

Given his belief that only the primary sector yields a net product, Quesnay concentrates his analysis on the advances of agriculture. In the *Analyse* Quesnay says that the advances of cultivation are made up of two milliards of annual advances and ten milliards of original advances. The former are entirely consumed during each productive process, and are a sort of circulating capital. Original advances last for more than one productive period and are subject to wear and tear, which is assumed to be worth one tenth of the initial stock. Therefore in the *Analyse* the value of the part of the social product needed to replace the capital goods which have worn out in production is equal to three milliards of livres. Quesnay calls this sum the annual returns, or *réprises*. Although the *Tableau* gives a concise and accurate description of all the types of advances, this issue is much more thoroughly dealt with in the articles written for the *Encyclopédie* just before the *Tableau*. In the article *Fermiers*, Quesnay examines all the different types of commodities that are required as inputs in the production of several primary commodities. These long lists of goods testify to Quesnay's attention to detail vis-à-vis the technical aspects of cultivation. He also gives numerical examples of the relationships between expenses and output with different methods of cultivation. Quesnay compares small-scale cultivation (which is mostly adopted in share-cropping and which uses oxen) with large-scale cultivation (where wealthy farmers employ horses). Horses are more powerful than oxen, but they are more expensive to feed, so that only rich farmers can use large-scale farming.

For Quesnay, productivity increases are strictly dependent on the amount of means of production available and hence on the original advances of the cultivators. The main way of increasing national wealth was by securing the existence of a large agricultural sector with wealthy farmers.

While the notion of surplus can be detected in the economic analysis of some authors who preceded Physiocracy – for instance, Petty's *A Treatise of Taxes and Contributions* (1662, pp. 30–31). But it is only with Quesnay that net product is defined precisely as the difference between the social product and its means of production. Here too the 1766 *Analyse* provides the most satisfactory description of the relationships between these magnitudes.

Quesnay believes that only agriculture can generate a net product; this view is taken for granted in the various *Tableaux* but is widely discussed in the *Dialogue sur les travaux des artisans*. This is highly significant; it testifies that Quesnay was not satisfied simply with the accounting definition of surplus, but also attempted to explain its origin and then the factors which might lead to its increase. The net product is the new wealth produced in each productive cycle. The main problem in Quesnay's analysis is of the nature of wealth and the ways in which it can be created. From this point of view Quesnay clearly set out the main economic issues which were subsequently examined by Adam Smith in the *Wealth of Nations*.

The concept of net product characterizes Quesnay's entire economic analysis and not just the *Tableau*. For instance in the 1757 article *Grains* he analyses the distribution of surplus between social classes; the Physiocratic theory of taxation is also built on the distinction between surplus and means of production. The notion of net product was the main analytical tool used by Quesnay to examine all the other important economic features in society. The idea of a single tax on rent is closely related to the determination of surplus. Rent is landlords' revenue, the landlords receiving the highest share of the net product of cultivation.

Agriculture is the only productive sector, and taxation of surplus is the only way of avoiding damage to future production. The stock of productive capital must not be reduced by taxes, otherwise it would be impossible to maintain the same level of activity as the previous year. Hence rent is the only taxable magnitude.

Quesnay's theory of circulation and distribution of income is also founded on the distinction between gross and net product. Part of the social product must circulate in such a way to replenish the means of production which have been consumed. As the *Tableau* clearly states the three milliards of returns to the cultivators depend on the methods of production. Hence the destination of output depends on the technical conditions which must be fulfilled to guarantee the self-sustaining character of the economic system. The other share of the social product is surplus; its distribution is not linked to technological factors, but customs and laws ascribe it to the landowners, including the Sovereign and the Church. This

appropriation of the net product is a social rule which is part of the reproduction of the social and economic system of the *ancien régime*. Therefore the Physiocratic theory of the distribution of income is based on technical and social conditions.

The identification of agricultural surplus with rent is a simplification of Quesnay's theory of income distribution, which has led many interpreters of Physiocracy to deny any relevant role to the concept of profits (see Meek, 1962, pp. 279–80, 384). In *Grains* Quesnay describes the way in which the net product is distributed; the highest share accrues to the landlords and to the King, but part of the surplus also goes to the farmers (INED, 1958, vol. 2, p. 475). Profits are regarded as a share of the net product also in other writings (pp. 601, 566). Quesnay puts less emphasis on farmers' profits in his later works, but the main problem arises because in the different types of *Tableau* there is no clear mention of their existence.

This awkward aspect of Physiocratic theory has been interpreted in the following way. Land leases between landlords and farmers are renewed every nine years (Weulersse, 1910, vol. 1, p. 405). Before this renewal the cultivators keep the surplus of cultivation; therefore they receive the entire benefits due either to technical progress (which diminishes production costs) or to an increase in the price of primary goods. But when lease contracts come up for renewal the farmers compete with each other for the right to cultivate the soil, thus rents rise and eventually absorb the whole net product. Farmers' profits are only a temporary share of surplus, and they finally ‘crystallize out’ into rent (Meek, 1962, pp. 279–80). The *Tableau* describes the final situation, thus profits are not included in it because they are not a regular part of the net product. The farmers' gains are sometimes regarded as salaries of superintendence to remunerate their work as entrepreneurs. According to Meek, this solution has the advantage of being consistent with Quesnay's fiscal theory. Farmers must be exempted from taxation because their incomes are not part of the surplus, and are not fully disposable; only rent has this requisite. Other authors consider profits as part of the net product, which is then made up of two parts, profits and rent, but only the latter is disposable for taxation because profits must be reinvested in production (Woog, 1950, pp. 21–2).

Whatever interpretation one decides to accept, it must be recognized that Quesnay was very ambiguous about farmers' profits. The fact that he played down their importance in his later writings is not necessarily due to the need for logical coherence, but can also be explained by his need to convince the nobles that cultivators would not become too rich and powerful (Vaggi, 1987, ch. 5, part 2). In any case it would be wrong to accuse Quesnay of failing to deal with profits. The true limitations of his analysis is the particular concept of profit he uses, namely ‘profit upon alienation’, a sort of windfall gain which exists simply because the cultivators are able to sell their products at a price higher than the cost of production. This notion can be traced back to mercantilist literature but was to be overturned by Turgot and Smith, who regarded profit as a rate on the capital invested (Groenewegen, 1971, pp. 333–4). In Quesnay's economics, profits have a precise economic role as the source of capital accumulation, because only the farmers can handle the increase in the original advances of cultivation. By means of a sustained process of investment in agriculture, cultivators transform part of the surplus of one productive cycle into the means of production of the next one. Even if farmers' profits were only a temporary share of net product, Quesnay's concept of profit is the notion which relates the three concepts of surplus, investments and capital accumulation. Moreover, Quesnay considers profits as the necessary stimulus to induce the cultivators to increase their advances; they act in consideration of profitability (INED, 1958, vol. 2, p. 807).

Another aspect of Quesnay's economics that is almost entirely ignored is his analysis of price determination. This interpretation too seems to derive from the fact that in the *Tableau économique* Quesnay does not openly discuss the question of value in exchange. The figures in the *Tableau are* in monetary terms, but they are currently interpreted as a proxy for physical magnitudes.

Many commentators have interpreted Quesnay's economics as a purely physical model: in agriculture the same goods are both the inputs and the outputs of the same process of production and the surplus can be measured simply as the difference between two physical quantities of agricultural product. The entire economic system can be described as a ‘corn model’ where the product and its means of production are homogeneous commodities; thus there is no need for a price theory. The manufactured products used in agriculture can simply be regarded as primary commodities which have changed shape thanks to the work of artisans (INED, 1958, vol. 2, p. 865).

Other authors maintain that prices exist in Physiocratic economics and that the agricultural output and its means of production are heterogeneous goods. However, they believe that Quesnay took the exchange ratios between the products of industry and those of agriculture as given, and fixed.

A careful reading of Quesnay's writings shows that he was very concerned about prices and changes in prices. This problem is not dealt with in the *Tableau*, but price theory is a necessary element in Quesnay's theory of the growth of national wealth. *Grains* and *Hommes* are the articles where Quesnay gave the clearest exposition of his price theory. Like many other economists before him, Quesnay separates the use value of commodities from their exchange value. In *Grains* he refers to a precise physical characteristic of a good which makes it suitable for the satisfaction of specific needs and desires. This is a sort of prerequisite which a good must have in order to be exchanged; there must be somebody willing to buy it. But the price of a commodity is independent of the utility derived from its consumption; use and exchange values are regulated by different laws (INED, 1958, vol. 2, p. 526). Quesnay singles out many concepts of price, which he uses to analyse the characteristics of the domestic and foreign trade of agricultural products. There is the *prix du vendeur*, which is the price paid by the merchant to the farmer. Besides this wholesale price, there is also the *prix de l'acheteur*, which governs the exchange between the merchant and the consumer. The retail price is always higher than the *prix du vendeur* which can also be called the ‘current price’ (INED, 1958, vol. 2, p. 752). The difference between the two prices is the merchant's gain, whose activity damages both the consumers and the producers, because it depresses the ‘current price’ and raises the retail one (p. 947). National wealth must be measured by current prices, when a product leaves the first producer there is no further increase in wealth, hence the merchants' gains are a burden for the whole economy. The merchants dominate the process of circulation of commodities because they are wealthy enough to stock the products of agriculture and wait for the best moments for their

purchases and their sales. On the contrary the farmers must always sell the entire output, because they are too poor to wait (INED, 1958, vol. 2, p. 985). Moreover, merchants usually enjoy exclusive rights, granted by the government, which give them a monopolistic position on certain markets. Free competition in domestic and foreign trade reduces merchants' profits and the difference between the retail and current price greatly diminishes; as was happening in England, where laissez-faire had already been implemented. The main purpose of free trade is that of raising the profitability of cultivation, without damaging the standard of living of consumers. This can be achieved thanks to the squeeze on merchants' profits. The merchants do not appear in the *Tableau*, even though they are mentioned in the accompanying texts. But here too one must remember that the *Tableau* describes an economy working according to natural laws, and free competition is one of the main features of natural order.

Quesnay was particularly interested in securing the free exportation of French corn. He believed that one of the main reasons for the backwardness of agriculture was the low level of the prices of its products. This fact made cultivation unprofitable, but domestic demand was not sufficient to raise these prices, because, on the whole, French consumers were too poor. In France there was lack of effective consumption (INED, 1958, vol. 2, p. 824), and Quesnay clearly separates the concept of effective demand, which is the demand of those who actually pay for a product, from the generic desire to consume more. A similar distinction was later to be found in Adam Smith's economics. Because of the lack of purchasing power at home it is necessary to allow the exportation of corn, so that the ‘current price’ can rise. Quesnay's arguments in favour of free trade are that it will strictly sustain the effective demand and the price of primary commodities, in order to achieve a higher profitability for cultivation. The ‘current price’ of corn was a *bon prix* in England, where the farmers were stimulated to reinvest their profits. It is the existence of a *bon prix* which guarantees profits to farmers and new investments in agriculture. Therefore the main aim of commercial policy is to restrict the gains made by merchants and increase the profits made by farmers. Landlords will also benefit from these measures, because they will receive higher rents when their leases are renewed and, above all, rent will rise because capital accumulation increases the productivity of agriculture and its surplus (Mirabeau, 1764, vol. 2, pp. 366 ff.).

Another important concept of price in Physiocratic economics is that of *prix fondamental*. This price is the sum of the overall expenses incurred by the farmer in the production of one unit of corn, it includes wages, raw materials and the repayment of wear and tear on fixed capital. These three items make up the technical cost of cultivation but they do not exhaust the expenses of production. To go on farming, the farmer must also fulfill his obligations towards the landlords by paying them the rent. This payment was a necessary social condition of production in 18th-century France, and was the way in which the king and the aristocracy received part of the fruits of their land.

The fundamental price is the price level below which the farmer makes a loss and stops farming the soil (INED, 1958, vol. 2, p. 529; Du Pont de Nemours, 1764, p. 18). This notion defines the lower limit of the current price, whose variations must be above the unit cost of production. If free competition rules in the process of circulation of commodities there is a positive difference between the current price and the fundamental one, which is the profit of the cultivator. In Quesnay's economics, theories of distribution and of value are closely related. Profits depend upon the technical conditions of cultivation, which influence the fundamental price, but are also affected by changes in the current price, which depend upon the state of the market for agricultural products and, in particular, on their effective demand.

Price theory is the final piece required to complete the picture of Quesnay's economic analysis of the ways of achieving growth and development. There are several steps to reach the well-ordered economy described in the *Tableau*. The implementation of free trade stimulates the effective demand for primary commodities and raises their current prices. This leads to higher profits for the farmers. If the fiscal system is based on the single tax on rent profits can be entirely reinvested in agriculture, thus raising the advances of cultivation. Capital accumulation in the productive sector of the economy leads to productivity increases. The surplus rises both in absolute terms and as a share of the social product; more and more resources can be reinvested in production, and national wealth grows. Quesnay develops a theory of growth which is based on the notion of surplus, but he used other concepts to complete it. In this mosaic the *Tableau* was also designed to single out the ultimate effects of this process of growth, to show the nobles that they would receive the benefits of Physiocratic economic policy.

Quesnay's main contributions to economic science are certainly the concept of surplus and the *Tableau économique*. But one must also stress that for the first time there is a complete and relatively coherent theory designed to answer the most important problems of economic systems on the basis of a general analysis of all their main features. All economic issues are examined with the help of precise concepts and on the basis of a theoretical approach. Economic policy proposals derive from theoretical speculation, and are part of a single general model of the growth of national wealth. Finally all the aspects of Quesnay's economics are linked, in one way or the other, to his notion of surplus, which provides a sort of unifying thread to his thought. Quesnay can quite appropriately be regarded as the founder of that approach to the analysis of economic events which is called theory of surplus.

See Also

- physiocracy
- surplus
- *Tableau économique*

Selected works

1756. Evidence. Fermiers. Grains. In *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers*, vols 6 and 7. Paris: Briasson. *Fermiers* trans. in Groenewegen (1983).

1759. *Tableau économique*. 3rd edn, Paris. Ed. M. Kuczynski and R. Meek, London: Macmillan, 1972.

1765. Hommes. Impôts. In *Encyclopédie*, vol. 8. Repr. in INED (1958).

Bibliography

Cartelier, J. 1976. *Surproduit et reproduction – la formation de l'économie politique classique*. Paris: Maspero.

Daire, E. 1846. *Physiocrates*. Paris: Librairie de Guillaumin.

Du Pont de Nemours, P.S. 1764. De l'importation et de l'exportation des grains. In *Collection des économistes et des réformateurs sociaux de la France*. Paris: Librairie Paul Geuthner, 1911.

Eltis, W.A. 1975. François Quesnay: a reinterpretation. 2. The theory of economic growth. *Oxford Economic Papers* 27, 327–51.

Eltis, W.A. 1984. *The Classical Theory of Economic Growth*. London: Macmillan.

Fox-Genovese, E. 1976. *The Origins of Physiocracy. Economic Revolution and Social Order in Eighteenth Century France*. Ithaca and London: Cornell University Press.

Gilibert, G. 1977. *Quesnay, la costruzione della macchina della prosperità*. Milan: Etas Libri.

Groenewegen, P.D. 1971. A re-interpretation of Turgot's theory of capital and interest. *Economic Journal* 83, 327–40.

Groenewegen, P.D., ed. 1983. *Quesnay. Farmers (1756) and Turgot. Sur la grande et la petite culture (1766)*. Reprints of Economic Classics, Series 2, No. 2. Sydney: University of Sydney.

Hecht, J. 1958. La vie de François Quesnay. In INED (1958), vol. 1.

Herlitz, L. 1961. The *Tableau économique* and the doctrine of sterility. *Scandinavian Economic History Review* 9(1), 3–51.

INED (Institut Nationale d'Etudes Démographiques). 1958. *François Quesnay et la Physiocratie*, 2 vols, ed. L. Salleron. Paris: INED.

Kuczynski, M. and Meek, R.L. 1972. *Quesnay's Tableau Economique*. London: Macmillan.

Marx, K. 1963. *Theories of Surplus Value*. London: Lawrence & Wishart.

Meek, R.L. 1962. *The Economics of Physiocracy: Essays and Translations*. London: Allen & Unwin.

Mirabeau, V.R. 1758–60. *L'ami des hommes ou traité de la population*. Repr. Aalen: Scientia Verlag, 1970.

Mirabeau, V.R. 1764. *Philosophie rurale, ou économie générale et politique de l'agriculture*. Amsterdam, chez les Libraires Associés. Repr. Aalen: Scientia Verlag, 1972.

Oncken, A., ed. 1888. *Oeuvres économiques et philosophiques de François Quesnay*. Paris: Jules Peelman.

Petty, W. 1662. A treatise of taxes and contributions. In *The Economic Writings of Sir William Petty*, vol. 1, ed. C.H. Hull. Cambridge: Cambridge University Press, 1899.

Phillips, A. 1955. The Tableau Economique as a simple Leontief model. *Quarterly Journal of Economics* 69(1), 137–44.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen & Unwin.

Spengler, J.J. 1958. Quesnay philosophe, empiriste, économiste. In INED (1958), vol. 1.

Tsuru, S. 1942. On reproduction schemes. In *The Theory of Capitalist Development*, ed. P. Sweezy. New York: Monthly Review Press.

Vaggi, G. 1987. *The Economics of François Quesnay*. London: Macmillan.

Weulersse, G. 1910. *Le mouvement physiocratique en France (de 1756 à 1770)*. Paris: Felix Alcan.

Woog, H. 1950. *The Tableau Economique of François Quesnay*. Berne: Franckle Verlag.

How to cite this article

Vaggi, G. "Quesnay, François (1694–1774)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_Q000011> doi:10.1057/9780230226203.1376

The New Palgrave Dictionary of Economics Online

R&D races

Jennifer F. Reinganum

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes recent theoretical and empirical research on R&D races. Two canonical models of an R&D race are described, and their implications for the investment behaviour of incumbent leaders and potential entrants are discussed. Empirical studies that attempt to verify or refute the implied patterns of investment are also discussed.

Keywords

R&D races; innovation; invention; patents

Article

Models of R&D races, in which the winning firm receives all (or most) of the reward, rely on two basic paradigms. In a deterministic race (Barzel, 1968), invention requires a known investment; the value of the patent grows over time and a potential inventor decides when to invest. Equivalently, the investment required to invent by time T is an increasing function $C(T)$. Let Πe^{-rT} denote the present value of a patent on an invention completed at T (assume that Π is also the social value). Among non-cooperative firms, the potential for pre-emption causes the date of invention to advance until $\Pi e^{-rT^0} - C(T^0) = 0$. Since T^0 precedes the socially optimal invention time, racing results in over-investment in R&D. This model assumes that only the winning firm actually invests its ‘bid’.

The first fully game-theoretic model of a stochastic race is presented by Loury (1979); for a decision-theoretic antecedent, see Kamien and Schwartz (1974). Firm i's lump-sum investment x_i yields a random invention date τ_i with $\Pr\{\tau_i \leq t\} = 1 - \exp\{-h(x_i)t\}$; thus, firm i's hazard rate is $h(x_i)$. The firm is also uncertain about the success date of its rival, which invests simultaneously. If firm i succeeds at t, the chance that it is the first inventor is $1 - \exp\{-h(x_j)t\}$. Firm i's expected profit is $\Pi h(x_i) / [r + h(x_j) + h(x_i)] - x_i$. Lee and Wilde (1980) re-specified investment as a *rate* per unit of

time. The resulting payoff for firm *i* is $[\Pi h(x_i) - x_i] / [r + h(x_j) + h(x_i)]$. In both versions, racing leads to over-investment. Malueg and Tsutsui (1997) incorporate uncertainty about the hazard rate; as time passes without success, firms become increasingly pessimistic and reduce their rates of investment. Weeds (2002) combines aspects of the Barzel and the Loury models with the theory of real options. Two firms monitor the stochastic growth of the patent value and decide when to invest. The required investment is exogenous and lump-sum, and in exchange the firm receives an exponentially distributed time until success. Firms do not invest until certain threshold values of the patent are reached. In one type of equilibrium, a firm invests only because it fears pre-emption by its rival; the rival invests strictly later, and firms' profits are equal (and low). In the other type, firms engage in mutual forbearance, and invest only at the optimal joint-investment time. Profits are equal (and high); this equilibrium involves strategic *delay* in investment relative to the social optimum.

Action-reaction or increasing dominance within a race

Although the exponential specification is gratifyingly simple, conditional on no success to date, the race looks exactly as it did at the beginning. If a firm could accumulate a ‘lead’, would a firm that is ‘ahead’ invest more than its rival? If so, then the race exhibits ‘increasing dominance’; if not, then the lagging firm tends to catch up and the race exhibits ‘action-reaction’ (Vickers, 1986). Of course, a lagging firm can invest more than the leader and still be less likely to win the race.

Harris and Vickers (1985) provide a deterministic racing model in which two firms alternate in making investments until one reaches the finish line. If one firm is sufficiently close to winning (in a region called its ‘safety zone’), the other drops out and the leader proceeds unchallenged. In the limit (as the rate of alternation goes to infinity), the winning firm invests just enough to reach its safety zone, whereupon its rival drops out and the firm continues, at a more leisurely pace, to the finish line. If the firms are otherwise symmetric, the firm that begins the race closer to the finish line will win.

Doraszelski (2003) provides a stochastic model wherein a firm's hazard rate depends on its accumulated stock of knowledge; its investment rate can vary with its own progress and that of its rival. Firm *i*'s hazard rate is $h^i = \lambda x_i + \gamma(z_i)^{\psi}$, where λ, γ, ψ are positive constants, x_i is firm *i*'s rate of investment and z_i is firm *i*'s accumulated stock of knowledge. Since x_i and z_i are substitutes in the hazard rate, a firm's equilibrium investment rate is decreasing in its knowledge stock, and the leading firm invests less than its rival; in this sense, the model exhibits action-reaction. Nevertheless, if one of the firms begins with a larger stock of knowledge, it remains *ex ante* more likely to win the race.

Progress can also be modelled by assuming that multiple stages must be completed in order to invent. Grossman and Shapiro (1987) provide a two-firm two-stage model; each stage involves a race of the Lee and Wilde form (see also Harris and Vickers, 1987, for a somewhat different model). They find that the leader invests more than the follower, but both increase their investments should the follower catch up. Thus, this literature suggests that a firm that is ‘ahead’ in the race is more likely to win.

Action-reaction or increasing dominance across multiple races

With a sequence of innovations, we ask whether a firm that is ‘ahead’ in the *market* (the one with a

larger market share) would invest more in the next race than its rival. That is, will the industry's evolution exhibit increasing dominance (persistence of the market leader) or action-reaction (turnover of the market leader)? Consider an incumbent monopolist, with current flow profits of *R*, racing with a potential entrant for a cost-reducing invention. If the incumbent wins, it receives the present value of monopoly profits with the new technology, Π. If the entrant wins, the two firms compete in Cournot–Nash fashion, receiving the present value of profits π_I and π_E for the incumbent and entrant, respectively. A drastic invention is one for which π_I=0 and $\pi_E = \Pi$ (that is, the incumbent can no longer compete when the entrant invents); if the innovation is non-drastic, then $\Pi > \pi_I + \pi_E$.
Gilbert and Newbery (1982) show that an incumbent monopolist would bid strictly more than a potential entrant for a non-drastic invention (and an equal amount for a drastic one); this is referred to as the ‘efficiency effect’. Suppose an entrant would invent at time T^E, where $C(T^E) = \pi_E e^{-rT^E}$. The incumbent can permit this and receive $(R/r)(1 - e^{-rT^E}) + \pi_I e^{-rT^E}$, or it can pre-empt by bidding just over $C(T^E)$, and receive essentially $(R/r)(1 - e^{-rT^E}) + \Pi e^{-rT^E} - C(T^E)$. Pre-emption is strictly preferred if $\Pi > \pi_I + \pi_E$; that is, if the invention is non-drastic. Vickers (1986) analyses a sequence of such races and finds that a sufficient condition for increasing dominance (action-reaction) is that each invention is sufficiently drastic (incremental).
Reinganum (1983) adapts the Lee and Wilde model to this scenario; the firms' payoffs are now $[\Pi h(x_I) + \pi_I h(x_E) + R - x_I] / [r + h(x_E) + h(x_I)]$ for the incumbent and $[\pi_E h(x_E) - x_E] / [r + h(x_E) + h(x_I)]$ for the entrant. In equilibrium, the incumbent invests at a higher rate than the entrant, at least for innovations that are sufficiently drastic. This is a consequence of Arrow's (1962) ‘replacement effect’ (originally identified for a single inventor): the incumbent has a lower incentive to advance the invention date (and replace himself) than does the entrant. Reinganum (1985) extends this result to a sequence of drastic inventions. Thus, the stochastic model suggests action-reaction, at least for sufficiently drastic inventions; however, increasing dominance can occur for sufficiently incremental inventions.

Empirical studies

Apparently, both models admit both investment patterns, but for opposite types (drastic versus incremental) of inventions. Since the nature of the inventive process (deterministic or stochastic) and that of the invention itself (drastic or incremental) are likely to vary across inventions and industries, the empirical pattern is also likely to vary. For a single race, the deterministic model has stark empirical implications: no real racing occurs beyond a possible initial burst. While the stochastic model implies investment by multiple firms, it is difficult to determine whether strategic effects play a significant role. The key indicator of strategic behaviour – the best response function – is an out-of-equilibrium phenomenon. Including rival investment in a regression explaining own investment should add no further information beyond that provided by the other explanatory variables. Using program-level data on pharmaceuticals R&D, Cockburn and Henderson (1994) detect no evidence of racing, but do find significant spillovers in output, calling into question the ‘winner-take-all’ assumption. They conclude that investment is driven by heterogeneous firm ability, adjustment costs, and technological opportunity.

In a laboratory experiment based on Harris and Vickers's (1987) multi-stage model, Zizzo (2002) finds that (contrary to predictions) the investments of leaders and followers are not significantly different. Lerner (1997) considers a sequence of races in the disk drive industry, with each invention serving to increase storage capacity. Firms that follow the technological leaders are most likely to introduce improved drives and to make the greatest technological progress. His data also show that leaders in a given year had about a 40 per cent chance of remaining leaders the next year. Czarnitzki and Kraft (2004) employ data from a survey of German manufacturing firms, which also asks firms to state their motives in conducting R&D; thus they distinguish between (self-designated) potential entrants and incumbents. R&D expenditures per dollar of sales are significantly higher for potential entrants (but the sales normalization confounds effects). The results of the latter two studies seem consistent with the stochastic model for relatively drastic inventions, but also with the deterministic model for relatively incremental inventions.

See Also

- patents

Bibliography

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Princeton: Princeton University Press for the NBER.

Barzel, Y. 1968. Optimal timing of innovations. *Review of Economics and Statistics* 50, 348–55.

Czarnitzki, D. and Kraft, K. 2004. An empirical test of the asymmetric models on innovative activity: who invests more into R&D, the incumbent or the challenger? *Journal of Economic Behavior and Organization* 54, 153–73.

Cockburn, I. and Henderson, R. 1994. Racing to invest? The dynamics of competition in ethical drug discovery. *Journal of Economics and Management Strategy* 3, 481–520.

Doraszelski, U. 2003. An R&D race with knowledge accumulation. *RAND Journal of Economics* 34, 24–42.

Grossman, G. and Shapiro, C. 1987. Dynamic R&D competition. *Economic Journal* 97, 372–87.

Gilbert, R. and Newbery, D. 1982. Pre-emptive patenting and the persistence of monopoly. *American Economic Review* 72, 514–26.

Harris, C. and Vickers, J. 1985. Perfect equilibrium in a model of a race. *Review of Economic Studies*

52, 193–209.

Harris, C. and Vickers, J. 1987. Racing with uncertainty. *Review of Economic Studies* 54, 1–21.

Kamien, M. and Schwartz, N. 1974. Patent life and R&D rivalry. *American Economic Review* 64, 183–7.

Lee, T. and Wilde, L. 1980. Market structure and innovation: a reformulation. *Quarterly Journal of Economics* 94, 429–36.

Lerner, J. 1997. An empirical exploration of a technology race. *RAND Journal of Economics* 28, 228–47.

Loury, G. 1979. Market structure and innovation. *Quarterly Journal of Economics* 93, 395–436.

Malueg, D. and Tsutsui, S. 1997. Dynamic R&D competition with learning. *RAND Journal of Economics* 28, 751–72.

Reinganum, J. 1983. Uncertain innovation and the persistence of monopoly. *American Economic Review* 73, 741–48.

Reinganum, J. 1985. Innovation and industry evolution. *Quarterly Journal of Economics* 50, 81–99.

Vickers, J. 1986. The evolution of market structure when there is a sequence of innovations. *Journal of Industrial Economics* 35, 1–12.

Weeds, H. 2002. Strategic delay in a real options model of R&D competition. *Review of Economic Studies* 69, 729–47.

Zizzo, D. 2002. Racing with uncertainty: a patent race experiment. *International Journal of Industrial Organization* 20, 877–902.

How to cite this article

Reinganum, Jennifer F. "R&D races." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000236> doi:10.1057/9780230226203.1377

The New Palgrave Dictionary of Economics Online

racial profiling

Petra E. Todd

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the recent theoretical and empirical literature in economics that aims to establish empirically whether police engage in racially biased law enforcement practices. It considers different objective functions that might be posited for police officers and the tests that can be derived under these objectives. Assuming a hit rate objective function leads to a simple, empirically implementable outcomes-based test that can potentially explain an observed empirical regularity in many police data-sets whereby disparities in hit rates tend to be very small despite large disparities in search rates.

Keywords

deterrence; omitted-variable bias; optimal auditing; racial profiling; rational expectations; statistical discrimination

Article

In recent years, numerous lawsuits have been brought against US city police departments alleging racially biased law enforcement practices. (Many of these lawsuits were initiated by the American Civil Liberties Union or the US Department of Justice.) As a result of closer scrutiny, many police departments now routinely collect administrative data on the characteristics of the individuals that they subject to stops and searches and on the outcomes of these encounters. It is a common finding in these data that African–Americans and Hispanic drivers are searched at a higher rate than white drivers. For example, African Americans represented 63 per cent of motorists searched by Maryland state police on the I-95 highway between 1995 and 1999, but only 18 per cent of motorists on the road. A refined version of this benchmark test for discrimination estimates the probability of being searched as a function of race and other observable characteristics thought to be related to criminal propensity. If race has explanatory power in the regression, then this is taken as evidence of discrimination (see, for example, Donohue, 1999).

One drawback of benchmark tests is that they require data on the full set of characteristics that a police officer uses in deciding whether to search a motorist. (A training manual issued by the Illinois State Police highlights some indicators of criminal activity, such as tinted windows and leased vehicles.) If some characteristics are missing, then race could have explanatory power due to omitted-variable bias. Alternatively, if race is found to be insignificant, it is still possible that police target individuals with certain characteristics, because of their correlation with race and not because of their use in predicting criminality. Additionally, benchmark tests can reveal only whether a disparity exists and not the motivation for the disparity. In many racial profiling investigations, it is clear that a disparity exists and the key concern is whether the higher rates of stop and search among certain groups can be justified as an optimal monitoring response to higher rates of criminality. The judicial standpoint on racial profiling is not clear-cut. The dominant view is that race or ethnicity can be used as a factor in determining the likelihood that a person has committed a crime, so long as its use relates to law enforcement and is not a pretext for racial harassment. However, a significant dissenting view argues that race should not be used as a criterion, except in very limited cases, as when the race of a perpetrator of a particular crime is known; see Kennedy, 1997. For detailed discussions of the legality of racial profiling practices, see, for example, Kennedy, 1997; Harcourt, 2004; Gross and Barnes, 2002; Persico and Castleman, 2005.

This article reviews the recent theoretical and empirical literature in economics that aims to empirically establish from data on stops and searches of motor vehicles whether police behaviour is indicative of racial bias. The early literature on crime (for example, Becker, 1968; Stigler, 1970) examined citizens' incentives to misbehave under an exogenous probability of being monitored. The more recent literature assumes that police and citizens behave strategically, with police deciding on optimal search strategies and citizens deciding whether to break the law, given police search strategies. Unlike the criminology literature, where it is sometimes assumed that police can make citizens believe that the monitoring probabilities are higher than in reality (for example, Sherman, 1990), the economics literature typically assumes rational expectations. (The recent economic literature is also related to the literature on optimal auditing, which mainly deals with income reporting and tax evasion; see Reinganum and Wilde, 1986; Border and Sobel, 1987; Scotchmer, 1987.) An advance in the literature is a better understanding of the assumptions on police and motorist behaviour required to justify alternative tests for discrimination.

We next describe the frameworks that have been developed in the recent economic literature and the tests that have been derived from using these frameworks. Subsequently, we present a brief summary of some of the empirical evidence from police stop/search data-sets.

Theoretical models of police – motorist behaviour

Two leading paradigms are put forward in the economic literature. One assumes that police officers operate in a decentralized way, allocating their search activities so as to catch as many criminals as possible. In the context of motor vehicle searches, the goal is to maximize the number of successful searches given a cost of search, where a successful search is defined as one that uncovers some contraband. As noted in Persico (2002), an objective function that maximizes successful searches, or so-called hit rates, will in general not minimize the aggregate crime rate, because it does not give enough weight to the deterrent effects of policing: that is, it does not reward preventing a crime from being committed.

Nevertheless, in light of principal–agent problems in policing, a hit rate objective may still be a reasonable approximation to police behaviour. It is likely difficult for a police chief to verify that individual officers engage in search activities that deter crime, because the amount of crime deterred is usually not observed. How many criminals an officer apprehends is observed, providing a rationale for rewarding officers on that basis. A model where police act as independent agents trying to catch criminals can be viewed as a second-best objective that a police chief might reasonably adopt.

The other modelling framework examined in the literature is one in which a centralized police chief allocates resources so as to minimize the overall crime rate. We describe the theoretical and empirical results derived using these two different modelling frameworks, with particular emphasis on devising tests for racial bias.

Models of hit rate maximization

The model of KPT (2001)

Knowles, Persico and Todd (2001) (KPT) develop a model of police–motorist behaviour that they use to study the implications of racial bias for equilibrium search outcomes. In the model, police officers decide which vehicles to subject to searches, and motorists decide whether to break the law by carrying contraband, such as drugs or illegal weapons, taking into account the probability of being searched.

In the absence of racial bias, each officer pursues a monitoring strategy that maximizes the number of successful search outcomes. Racial bias is introduced as a preference parameter that reduces the perceived cost of searching vehicles of black or Hispanic drivers relative to white drivers, which can lead to over-searching of these groups. An equilibrium implication of racially biased monitoring, shown in KPT and discussed further below, is that the hit rate (the rate at which contraband is seized) should be lower for the groups subject to bias. (The general idea that tastes for discrimination lead to lower profits for discriminators originated with Becker, 1957. For further discussion of such tests in policing contexts, see Ayres, 2002.)

Let $r \in \{A, W\}$ denote the race of the motorist (African-American or white), assumed to be observable by the police officer. (We assume two groups here, but the analysis extends straightforwardly to more groups.) Let c denote all characteristics other than race that are potentially used by the officer in the decision to search cars, which may be unobserved or only partially observed by the econometrician. For expositional ease, treat c as a one-dimensional variable (results extend to the multidimensional case), and denote the distribution of c in the white and African-American populations by $F(c|W)$ and $F(c|A)$.

It is assumed that an individual police officer allocates search efforts so as to maximize the number of convictions minus a cost of searching cars. Each officer can choose to search motorists of any type (c,r) at a marginal cost of t_r. Normalize the benefit of each arrest to equal 1, so that the cost is scaled as a fraction of the benefit (assume $t_W, t_A \in (0, 1)$). In deciding whether to carry contraband, motorists consider the probability of being searched and the penalty if they were to be caught. If they do not carry, their payoff is zero whether or not the car is searched. If they carry, their payoff is $v(c, r) + x$ if not searched and $-j(c, r)$ if searched, where both $j(c,r)$ and $v(c,r)$ are positive. x represents private information of the motorists about their own benefit from carrying contraband. $j(c,r)$ can be interpreted as the cost of being convicted. (If there were discrimination in the court system leading to higher

penalties for minority drivers, this could be thought of as operating through $j(c,r)$. KPT do not test for this type of discrimination.)

Denote by $\gamma(c,r)$ the probability that the police officer searches a motorist of type c,r. The expected payoff to a motorist from carrying contraband is

$$\gamma(c,r)[-j(c,r)] + [1-\gamma(c,r)][v(c,r)+x]. \tag{1}$$

Given $\gamma(c,r)$, the motorist chooses to carry contraband if this expression is greater than zero. Motorists with a high realized value of x strictly prefer to carry drugs and those with small values strictly prefer not to carry. However, police search strategies can be conditioned only on c and r, because x is not directly observed by the police. Let G denote the event that the motorist searched is found guilty of carrying contraband, and denote the probability that a motorist of type c,r carries contraband by $P(G|c,r)$. (We do not allow for the possibility of false accusation by police or planting of evidence, as considered in Donohue and Levitt, 2001.)

Assume that the police officer decides on the search probability $\gamma(c,r)$ (the probability of searching each motorist of type c,r) to maximize the number of successful searches, net of costs. He or she solves

$$\max_{\gamma(c,W),\,\gamma(c,A)} \sum_{r=W,A} \int [P(G|c,r) - t_r]\,\gamma(c,r)\,f(c|r)\,dc,$$

taking as given $P(G|c,r)$. The term $P(G|c,r) - t_r$ represents the expected profit from searching a motorist of type c,r. If $P(G|c,r) - t_r > 0$ then optimizing behaviour implies $\gamma(c,r) = 1$, that is, always search motorists of type c and r. If $P(G|c,r) = t_r$ then the police officer is willing to randomize over whether or not to search type c,r.

KPT introduce racial bias into this framework as a difference in the perceived cost of searching motorists of different races. That is, a police officer is said to be biased against race A (or to have a taste for discrimination) if $t_A < t_W$. If a police officer has no taste for discrimination and yet chooses search probabilities that differ by race, then the equilibrium is said to exhibit statistical discrimination. Statistical discrimination is motivated out of efficiency considerations and not out of racial bias. (Statistical discrimination is used here in the same sense as in Arrow, 1973.)

KPT (2001) study the equilibrium implications of this model for the case where officers are homogeneous in their costs of search and motorists are heterogeneous in the benefits they derive from carrying contraband. The model implies the following equilibrium conditions, for all c

$$P^*(G|c, A) = t_A,\ P^*(G|c, W) = t_W \gamma^*(c, A) = \frac{v(c, A)}{[v(c, A) + j(c, A)]},\ \gamma^*(c, W) = \frac{v(c, W)}{[v(c, W) + j(c, W)]},$$
(2)

where * denotes equilibrium values.
Suppose that $t_A = t_W = t$ (that is, police officers are not biased). Then, for all *c*, guilt probabilities at equilibrium must be equal across races:

$$P^*(G|c, A) = t = P^*(G|c, W).$$
(3)

If guilt probabilities were not equalized, a police officer could do better by reallocating searches towards the group with the higher hit rate.
An important observation is that equalization of hit rates does not imply equalization of search rates. The equilibrium search intensity may be higher for African-Americans even in the absence of racial bias. This happens if $v(c, W) / [v(c, W) + j(c, W)] < v(c, A) / [v(c, A) + j(c, A)]$. That is, if the expected value of carrying drugs is higher or the cost of being convicted lower for black motorists, then the search rate on that group would have to be higher in order to equalize the guilt probability to that of whites.
Equation (3) is the basis for the outcomes-based test proposed in KPT as a test for racial bias (a test for $t_W \neq t_A$). An advantage of the test is that it is implementable even in the absence of complete data on *c* and on γ^*. It suffices to have data on the frequency of guilt by race conditional on being searched,

$$D(r) = \int P^*(G|c, r) \frac{\gamma^*(c, r)\, f(c|r)}{\int \gamma^*(s, r)\, f(s|r)\, ds} dc.$$

Using (3) to substitute for $P(G|c,r)$ we get the implication

$$D(W) = t = D(A),$$
(4)

which KPT empirically test. In the model, there is nothing special about the characteristic ‘race’. The analogue of (4) should hold for any observed characteristic on which police can condition their search

decision. Thus, the model has the strong implication that the guilty rates should be equal for any set of observed conditioning variables, such as age, gender, or type of car.

The assumption that motorists respond to the probability of being searched is key to obtaining a test for bias that is applicable even without data on all the characteristics that police use in the search decision. If motorists did not react to the probability of being searched, testing for prejudice would require data on c. To see why, consider a model where the probability that a motorist with characteristic c and race r carries drugs does not depend on the actions of police, and the only optimizing agents are the police. Let $\pi(c,r)$ denote the probability that a type c,r carries drugs and suppose that $\pi(c,r)$ is increasing in c. Then, it is optimal for police officers to choose two cut-offs k_W and k_A and to search any motorist of race r with a c greater than k_r. In the absence of prejudice, police will choose k_W and k_A so that the probability that types k_W,W and k_A,A are guilty equals the marginal cost t of searching motorists. Without data on c, one cannot empirically identify the marginal motorists and so cannot test the equilibrium implications of this model, in the absence of strong assumptions on the shape of $\pi(c,r)$ and on the distribution of the unobservables. When $\pi(c,r)$ is determined endogenously, the only equilibrium is for $\pi(c,r)$ to equal t_r for all c. Thus, allowing for endogenous response of motorists to the probability of being searched eliminates the problem of having to identify the marginal motorists.

A number of papers have explored extensions or variations of the KPT model. For example, Antonovics and Knight (2004) raise the concern that police heterogeneity is a potential threat to the validity of the outcomes-based tests, and they present evidence that police are more likely to search the vehicles of drivers of a different race. Persico and Todd (2006) generalize the KPT model to allow for police heterogeneity in costs of search and for the possibility that drivers can adapt some of their characteristics to reduce the probability of being monitored. (For example, if drivers with sports cars are subject to high monitoring rates, an individual might choose to drive a different type of car.) They show that the hit rate test is still valid under these extensions. Persico and Todd (2005) further extend the model to allow for imperfections in the monitoring technology, namely, that searches do not always uncover contraband. Even with varying detectability rates across groups, the hit rate test can still be justified as a test for racial bias. (The extensions are developed in an application of the model to monitoring of passengers at airports.)

Hernandez-Murillo and Knowles (2004) consider how to test for racial bias with aggregated data that is contaminated by observations on non-discretionary stops. The KPT model assumes that police searches are discretionary, whereas Hernandez-Murillo and Knowles analyse a data-set from Missouri that mixes discretionary and non-discretionary stops. They derive tests for racial bias (inspired by the nonparametric bounding approach of Horowitz and Manski, 1995) that are robust to the contamination.

Dharmapala and Ross (2004) extend the KPT model by relaxing the assumption that police can search any motorist. They impose a technological limitation on the search capacities of police, whereby police observe motorists with probability less than 1. In this case, there can be some motorists for whom the constraint leads them to carry contraband all the time. Police would like to search this group harder to equalize guilty rates, but cannot. In equilibrium, such motorists will be searched whenever police observe them. Dharmapala and Ross (2004) demonstrate that, if this type of motorist is distributed differently among racial/ethnic groups, then the hit rate test breaks down. They also consider the set of equilibria in a modified version of the KPT model in which there are offences of varying levels severity

and motorists sort over the level of severity.

The model of Anwar and Fang (2006)

A limitation of the KPT model is that it assumes that police officers first see some motorist characteristics and then decide whether to search them. A more realistic assumption is that police see some information prior to the stop decision and then acquire more information from interacting with the motorist. Anwar and Fang (2006) develop a framework in which the officers' search decisions can depend on the additional information they acquire after the initial stop. They also allow for the possibility that police behaviour varies with the race of the police officer. For example, white police may be biased against minority drivers and minority police biased against white drivers. (Persico and Todd, 2004, also allow for police heterogeneity in the bias, but do not allow for the sign of the bias to differ for individual officers. That is, they do not allow for the possibility that some police may be biased while others may exhibit favouritism, which is the case considered in Anwar and Fang, 2006.)

The model of Anwar and Fang (2006) is in the spirit of statistical discrimination models (see, for example, Coate and Loury, 1993). It assumes that during the stop and prior to the search decision the police officer observes a noisy but informative signal about whether the driver carries contraband. The signal is informative in the sense that guilty drivers (those carrying illegal contraband) are more likely than innocent drivers to generate suspicious signals, such as nervousness in answering the police officer's questions. (It is assumed that the drivers themselves do not know at the time of deciding whether to carry contraband whether they will generate a signal, only the probability that they will generate one.) As in the KPT model, police officers are considered racially biased if their cost of search depends on the race of the motorist and the objective of officers is to maximize the number of successful searches.

Anwar and Fang (2006) develop two tests. The first is a test for whether police officers of different races use different search criteria when dealing with motorists of the same race, which would indicate police heterogeneity. The test is based on the observation that, if officers do not differ in search costs, then the search rates and success rates of different groups of officers should on average be the same. The second test they develop is a test for racial prejudice that can uncover whether at least one of the groups of officers (for example, white or minority officers) is searching in a racially biased manner, although it cannot distinguish which group is biased.

Models of crime minimization

The previous class of models assumed that individual officers adopt search strategies that maximize successful search rates, or so-called hit rates. As noted above, a hit rate objective function would be a reasonable approximation to police behaviour if officers are rewarded on the basis of criminals apprehended, something that is easily observed. As demonstrated in Persico (2002), however, the hit rate objective function does not minimize the aggregate crime rate.

An alternative modelling framework assumes that there is a centralized authority, a police chief, say, who can direct officers to focus their searches on particular subgroups. In such a model, the hit rate test fails as a test of the unbiasedness of the police chief, because, in the equilibrium of such a model, an

unbiased police chief will allocate searches to equate the deterrence effect and not the hit rates across groups. Crime deterred is unobserved, making it difficult to test for whether the deterrence effect is being equalized across groups. While one could conceivably introduce racial bias in a crime minimization model in the same way as in the hit rate maximization models – as a difference in the costs of searching different types of motorists – there is currently no empirically implementable test for racial bias in such a model. (Eeckhout, Persico and Todd, 2003, study optimal monitoring strategies for police assuming that the objective is to minimize crime. They show that in some cases it can be optimal to randomly subject even identical motorists to different levels of monitoring. This could be considered random profiling, in that motorists are randomly divided into different groups and are subjected to different levels of monitoring.)

Imposing a race-blind constraint on police behaviour

Persico (2002) studies the effects of constraints on police behaviours, within a model where police maximize hit rates, but the assumed socially efficient objective is to minimize the aggregate crime rate. He shows that imposing a ‘race-blind’ constraint on police search behaviour does not necessarily entail any loss in efficiency and can sometimes increase efficiency. That is, not allowing police to condition their search probabilities on race can sometimes lead to a lower-cost way of achieving a given crime rate. This somewhat surprising result follows because search strategies that aim to maximize arrests do not take into account the deterrence value that arrests have on different groups. The incentive scheme that minimizes the crime rate would place a higher value on arresting motorists of the race that is more likely to be deterred by the prospect of being arrested.

To see how a fairness constraint can increase efficiency, suppose, for example, that whites were more numerous in the population and were also less likely than blacks to carry drugs at a given search rate. In the absence of any constraints on search behaviour, police would search blacks at a higher rate so as to equalize the hit rates across groups. Under a fairness constraint, however, the two groups are pooled and experience the same probability of being searched. In equilibrium, the overall carrying rate will remain the same as in the unconstrained equilibrium and will equal the cost of search. However, equalizing search rates by race leads to an increase in the black carrying rate, and an offsetting decrease in the white rate. If whites are deterred by a relatively small increase in the probability of search and they are more numerous in the population, then it is possible to achieve the same overall carrying rate at a lower search cost in the constrained equilibrium than in the unconstrained equilibrium. Persico (2002) finds that whether imposing a fairness constraint leads to an increase in efficiency (defined as a decrease in the crime rate for the same cost) crucially depends on the proportion of blacks in the population relative to the cost of search.

For further consideration of how to incorporate efficiency and equity considerations into an assessment of racial profiling as a public policy, see Durlauf (2005) and Risse and Zeckhauser (2004). Also, see Dominitz (2003) for discussion of the statistical relationship between various outcomes that could be considered when formulating public policy, such as search rates, find rates, thoroughness of search, rates of detention of the innocent, and rates of apprehension of the guilty.

Empirical evidence

As noted above, models that assume that police maximize hit rates and that motorists take into account the probability of being caught when deciding whether to break the law lead to simple procedures to test for racial bias. The KPT test compares hit rates across different groups of motorists, which can be performed using the type of data that is conventionally available.

Table 1 (reproduced from Persico and Todd, 2006) summarizes findings from 16 different city-level and state-level racial profiling studies/reports, in which the hit rates by race/ethnicity are reported. The table displays what appears to be an empirical regularity: there is not a large disparity in hit rates for black and white drivers, especially when compared with the disparity in search/stop rates. This regularity is puzzling in the context of a crime-minimizing police chief, but not in light of a simple hit rate maximization model, which offers a rationale for the equalization of hit rates across races, namely, (*a*) that police are allocating searches in a way that maximizes efficiency in catching criminals and (*b*) that police departments, on average, are not afflicted by widespread bias against African Americans. In Table 1, the hit rates for Hispanics are in many cases notably lower than that of whites or blacks, which is suggestive of bias against Hispanics. Whether in fact this is really the case can be ascertained only with more work on the police data-sets that are becoming available.

Summary of hit rate findings for racial profiling studies

Location	Whites	Blacks	Hispanics	Source
Maryland	22.7	22.0	18.9	Knowles, Persico and Todd (2001)
Florida	32.0	34.0	11.0	Anwar and Fang (2006)
Tennessee	25.1	20.9	11.5	Cohen-Vogel and Doss (2002)
New Jersey	20.1	19.2	10.3	Verniero and Zoubak (1999)
Rhode Island	10.5	13.5	n/r	Farrell et. al. (2003)
New York (pedestrian)	23.5	17.8*	17.8*	Spitzer (1999)
Charlotte, NC	13.0	11.0	n/r	Smith et. al. (2004)
Lansing, MI	30.9	24.2	n/r	Carter, Katz-Bannister and Schafter (2002)
Missouri	6.8	8.7	n/r	Nixon (2003)
San Antonio, TX	23.2	17.5	14.7	Lamberth (2003)
Denver, CO	17.2	14.6	14.9	Thomas and Hansen (2004)
Denver, CO (pedestrian)	16.5	19.7	11.3	Thomas and Hansen (2004)
Los Angeles, CA	18.7	20.6	14.6	Tabulations provided by the LAPD
Sacramento, CA	23.8	18.2	17.2	Greenwald (2003)
San Diego, CA	26.5	22.4	28.0	Cordner, Williams and Velasco (2002)
Washington State	11.0	12.0	5.0	Lovrich et. al. (2003)
Wichita, KS	32.0	21.0	n/r	Persico and Todd (2006)

*A single hit rate is reported for all minorities.

Anwar and Fang (2006) apply their alternative test (not described here, for the sake of brevity) to a data-set on highway stops and searches collected by the Florida Highway Patrol. The data reveal search patterns that differ significantly by race of the officer. Despite the differences in search behaviour, however, the test does not reject the null hypothesis of no relative racial prejudice between black and white officers. (The authors advise caution in interpreting the results, as the test is informative only about relative racial bias and cannot rule out the possibility that all police – of both races – might be biased.)

Summary and conclusions

Recent advances in the economic literature have led to a better appreciation of the assumptions that underlie different approaches to testing for racial/ethnic bias in policing. Simple benchmark tests for discrimination only uncover whether a disparity exists; they do not reveal the motivation for the disparity, which plays a key role in racial profiling investigations. Assuming a hit rate objective function and strategic behaviour on the part of both police officers and motorists leads to a simple, empirically implementable outcomes-based test for racial bias. Such a model can potentially explain an observed empirical regularity in many police data-sets to the effect that there is little disparity in hit rates for black and white drivers, despite large disparities in search/stop rates. An alternative modelling framework is one in which a police chief allocates resources so as to minimize the aggregate crime rate. An implication of unbiased policing in this type of model is that the deterrence effect be equated across different groups of motorists, which is difficult to test empirically.

Even if an outcomes-based test concludes that racial disparities in search rates do not reflect racial bias, there is still the question of whether statistical discrimination is justified. Statistical discrimination may be considered unfair, because drivers experience different probabilities of being searched, depending on their race. An intriguing aspect of Persico's (2002) findings is that, when police are maximizing a second-best objective (hit rates), imposing a race-blind constraint can bring them closer to a first-best objective (minimizing overall crime). In practice, though, it may be difficult to ask police officers to simply ignore race in their decision-making. More race-neutral policing might instead be achieved by giving police differential rewards for hit rates on white and minority drivers. Designing optimal incentive schemes for achieving a particular objective is a current area of research.

Finally, the economic literature on racial profiling is relatively nascent, and there are many ways of extending existing models. For example, none of the existing models specifies how a police chief allocates police officers to patrol particular areas. For this reason, existing tests are usually applied to data on highway searches where selective allocation of officers to monitor certain populations is less of an issue. Also, existing theory has mainly been developed for discretionary stops, but a major proportion of police stops and searches are triggered by events that make a search of the vehicle mandatory.

See Also

- Arrow, Kenneth Joseph
- Becker, Gary S.

I am grateful to Steven Durlauf, Hanming Fang and Nicola Persico for helpful discussions

Bibliography

Antonovics, K. and Knight, B. 2004. A new look at racial profiling: evidence from the Boston Police Department. Working Paper No. W10634. Cambridge, MA: NBER.

Anwar, S. and Fang, H. 2006. An alternative test of racial prejudice in motor vehicle searches: theory and evidence. *American Economic Review* 96, 127–51.

Arrow, K. 1973. The theory of discrimination. In *Discrimination in Labor Markets*, ed. O. Ashenfelter and A. Rees. Princeton: Princeton University Press.

Ayres, I. 2002. Outcomes tests of racial disparities in police practices. *Justice Research and Policy* 4, 131–42.

Becker, G. 1957. *The Economics of Discrimination*. Chicago: Chicago University Press.

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Border, K. and Sobel, J. 1987. Samurai accountant: a theory of auditing and plunder. *Review of Economic Studies* 54, 525–40.

Carter, D., Katz-Bannister, A. and Schafter, J. 2002. Analysis of the lansing police department MATS data: an eighteen month status report. Report submitted to Chief Mark Alley, Lansing Police Department.

Coate, S. and Loury, G.C. 1993. Will affirmative action policies eliminate negative stereotypes? *American Economic Review* 83, 1220–40.

Cohen-Vogel, D. and Doss, B. 2002. *Vehicle Stops and Race: A Study and Report in Response to Public Chapter 910 of 2000*. Report to Comptroller of the Treasury. Nashville: Comptroller of the Treasury, State of Tennessee.

Cordner, G., Williams, B. and Velasco, A. 2002. *Vehicle Stops in San Diego: 2001*. San Diego, CA: San Diego Police Department.

Dharmapala, D. and Ross, S. 2004. Racial bias in motor vehicle searches: additional theory and evidence. *Contributions to Economic Analysis & Policy* 3(1), Article 12. Online. Available at http://www.bepress.com/bejeap/contributions/vol3/iss1/art12, accessed 2 September 2006.

Dominitz, J. 2003. How do the laws of probability constrain legislative and judicial efforts to stop racial profiling? *American Law and Economics Review* 5, 412–32.

Donohue, J. 1999. Expert witness testimony in the case. *Chavez v. Illinois State Police*.

Donohue, J. and Levitt, S. 2001. The impact of race on policing, arrest patterns, and crime. *Journal of Law and Economics* 44, 367–94.

Donohue, J. 2005. The law and economics of antidiscrimination law. Public Law Working Paper No. 101, Yale Law School.

Durlauf, S. 2005. Racial profiling as a public policy question: efficiency, equity, and ambiguity. *American Economic Review* 95, 132–6.

Eeckhout, J., Persico, N. and Todd, P. 2003. A theory of random crackdowns. Mimeo, University of Pennsylvania.

Epstein, R. 1992. *Forbidden Grounds: The Case Against Employment Discrimination Laws*. Cambridge, MA: Harvard University Press.

Farrell, A., McDevitt, D., Cronin, S. and Pierce, E. 2003. *Rhode Island Traffic Stop Statistics Act Final Report*. Boston, MA: Institute on Race and Justice, Northeastern University.

Greenwald, H. 2003. *Police Vehicle Stops in Sacramento, CA: Second of Three Reports*. School of Policy, Planning and Development, University of Southern California.

Gross, S. and Barnes, K. 2002. Road work: racial profiling and drug interdiction on the highway. *Michigan Law Review* 101, 653–754.

Harcourt, B. 2004. Rethinking racial profiling: a critique of the economics, civil liberties, and constitutional literature and of criminal profiling more generally. *University of Chicago Law Review* 71, 1276–380.

Hernandez-Murillo, R. and Knowles, J. 2004. Racial profiling or racist policing? Bounds tests in aggregate data. *International Economic Review* 45, 959–89.

Horowitz, J.L. and Manski, C.F. 1995. Identification and robustness with contaminated and corrupted data. *Econometrica* 63, 281–302.

Kennedy, R. 1997. *Race, Crime and the Law*. New York: Pantheon Books.

Knowles, J., Persico, N. and Todd, P. 2001. Racial bias in motor vehicle searches: theory and evidence. *Journal of Political Economy* 109, 203–29.

Lamberth, J. 2003. *Racial Profiling Data Analysis Study: Final Report for the San Antonio Police Department*. Online. Available at http://www.sanantonio.gov/sapd/pdf/LamberthSanAntonioRpt_2003.pdf, accessed 3 October 2006.

Lovrich, N., Gaffney, M., Mosher, C., Pickerill, M. and Smith, M. 2003. ‘WSP traffic stop data analysis project: data analysis project report. Division of Governmental Studies and Services, Washington State University.

Nixon, J. 2003. Executive Summary on 2003 Missouri traffic stops. Jefferson City: Missouri attorney general's office. Online. Available at http://www.ago.mo.gov/racialprofiling/2003/racialprofiling2003.htm, accessed 2 September 2006.

Persico, N. 2002. Racial profiling, fairness and effectiveness of policing. *American Economic Review* 92, 1472–97.

Persico, N. and Castleman, D. 2005. Detecting bias: using statistical evidence to establish intentional discrimination in racial profiling cases. *2005 University of Chicago Legal Forum*, 217–35.

Persico, N. and Todd, P. 2005. Passenger profiling, imperfect screening and airport security. *American Economic Review* 95, 127–31.

Persico, N. and Todd, P. 2006. Generalizing the Hit Rates Test for Racial Bias in Law Enforcement, with an Application to Vehicle Searches in Wichita. *Economic Journal* 116November, F351–67.

Reinganum, J. and Wilde, L. 1986. Equilibrium verification and reporting policies in a model of tax compliance. *International Economic Review* 27, 739–60.

Risse, M. and Zeckhauser, R. 2004. Racial profiling. *Philosophy and Public Affairs* 32(2), 31–70.

Ross, S. 1996. Mortgage lending discrimination and racial differences in loan default. *Journal of Housing Research* 7, 117–26.

Scotchmer, S. 1987. Audit classes and tax enforcement policy. *American Economic Review* 77, 229–33.

Sherman, L. 1990. Police crackdowns initial and residual deterrence. In *Crime and Justice: A Review of Research*, vol. 12, ed. M. Tonry and N. Morris. Chicago: University of Chicago Press.

Smith, W., Davison, E., Zingraff, M., Rice, K. and Bissler, D. 2004. An empirical investigation of the

possible presence and extent of arbitrary profiling in the Charlotte-Mecklenburg police department. Final Report to Charlotte-Mecklenburg Police Department. Department of Sociology and Anthropology, North Carolina State University.

Spitzer, E. 1999. *A Report to the People of the State of New York from the Office of the Attorney General*. New York: Civil Rights Bureau.

Stigler, G. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78, 526–36.

Thomas, D. and Hansen, R. 2004. Second annual report, Denver police department, contact card data analysis. Department of Geography, University of Colorado at Denver.

US General Accounting Office. 2000. *Customs Service: Better Targeting of Airline Passengers For Personal Searches Could Produce Better Results*. GAO/GGD-00-38. Washington DC: US Government Printing Office.

Verniero, P. and Zoubek, P. 1999. *Interim Report of the State Police Review Team Regarding Allegations of Racial Profiling*. Trenton, NJ: Attorney General's Office.

How to cite this article

Todd, Petra E. "racial profiling." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000229> doi:10.1057/9780230226203.1378

The New Palgrave Dictionary of Economics Online

radical economics

Diane Flaherty

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Contemporary radical economics comprises a broad set of methodological approaches, including Marxian political economy, institutionalism, Post Keynesianism, analytical political economy, radical feminism and postmodernism. Unlike radical economics in the mid-1980s, radical thought today emphasizes conflict other than class conflict, policy-relevant analysis and incorporation of more mainstream methods into radical research. Nonetheless, despite substantial evolution, radical economics remains faithful to its original vision. Uniting the various approaches is a set of unchanged core principles, the three most salient of which are the importance of history, embeddedness of individual choice in an institutional environment, and the centrality of conflict to understanding capitalism.

Keywords

agency; aggregate demand; alienation; altruism; analytical political economy; capital controls; capitalism; chaos theory; class; collective action; complexity; conflict; convergence; credit controls; distribution; endogeneity of preferences; equilibrium; experimental economics; exploitation; game theory; globalization; heterodox economics; *homo economicus* ; inequality; intrahousehold welfare; Keynesianism; labour theory of value; Marxian political economy; new institutional economics; optimization; path dependence; Post Keynesianism; postmodernism; profit; race; radical economics; radical feminism; simultaneous equation models; social construction of norms; surplus; trade unions; Washington consensus

Article

Contemporary radical economics comprises a broad set of methodological approaches, including Marxian political economy, institutionalism, Post Keynesianism, analytical political economy, radical feminism and postmodernism (see for example Pietrykowski, 2000; Colander, Holt and Rossiter, 2004; Dutt, 2005). This inclusive definition of radical economics differs in many respects from the radical

economics of the mid-1980s. Today radical thought emphasizes conflict outside of class conflict, policy-relevant analysis and incorporation of more mainstream methods into radical research.

Nonetheless, despite substantial evolution, radical economics (or, as it is often called now, heterodox economics) remains at the core faithful to its original vision. All approaches are grounded in a stable set of uniting principles, the three most salient of which are the importance of history, embeddedness of individual choice in an institutional environment and the centrality of conflict to understanding capitalism.

Roots of radical economics

Radical economics has always identified its fundamental project as the construction of realistic representations of the capitalist system, the better to identify and redress exploitation, alienation and inequality. Specifically, the shared goal from the beginning has been to incorporate a degree of reality not available in neoclassical models based on assumptions of atomistic individuals optimizing under conditions of complete information and perfect foresight.

The first unifying proposition of radical thought is the importance of history. The past shapes the present through inherited initial conditions: all choices in the current period are made within constraints imposed by history. Moreover, projections into the future must eschew assumptions of perfect foreknowledge or even, in some strands of radical thought, a known risk embodied in a fixed distribution of outcome probabilities. Path dependence renders the future unpredictable and unknowable.

Social construction of norms and endogeneity of preferences to institutional constraints constitute a second set of shared radical concepts. Individuals make choices on the basis of ‘background’ criteria derived from social norms, which in turn arise from stable institutions and rules of the game (Searle, 1995).

The radical critique of exogenous preferences does not simply replace individual agency with structure. Rather, most strands of radical thought assume interaction between individual and social structure. The dominant view posits that intentional human agency exerts its most powerful effect on institutions in contentious historical conjunctures (Setterfield, 2005; Searle, 1995). Conflictual historical periods give rise to re-examination, rejection and perhaps supersession of existing institutions because individuals challenge established norms.

Conflict in turn is seen by radicals as endemic to capitalism and a cause of chronic inefficiency. Conflict at the workplace makes capitalism operate at less than maximum achievable output (Bowles and Jayadev, 2005). At the macroeconomic level, conflict over distribution is implicated in lack of stability or consistent growth (Bhaduri, 2006; Pollin, 1997). Chronic conflict periodically and unavoidably rises to an acute level, crises emerge and institutions once supportive of growth fail to resolve the crisis.

Impetus to the evolution of radical thought

While most current practitioners of radical economics are, as noted above, still working broadly within the original radical vision, since the mid-1980s internal criticisms of radical research have precipitated significant change within the field. Citing lack of coherence or realism, critics urge theoretical overhaul to achieve a truly integrated analysis of history, institutions and conflict.

Immanent criticism argues that radical theory must explain collective action, both to establish the centrality of class and to sustain a critique of the neoclassical behavioural model predicated on atomistic, self-interested individuals. Radical theory in its original incarnation provided little analysis of how individuals decide to engage in collective action. Thus, radical analysis fell short in explaining how or when the working class decides to act in its own interest. Most radical economics rejected early on any explanation derived from an overly simplified materialist assertion that individuals are simply bearers of their class roles. Nonetheless, lacking a clear alternative, radical explanations still tended to fall back into an unexamined and unacknowledged use of functionalist arguments, in which behaviours derive from the structure and requirements of capitalism.

Exploitation, also a central proposition of radical theory, similarly required further elaboration. The labour theory of value as the theoretical foundation of the role of labour as the sole source of surplus and profit sustained damaging attack from analytical Marxists, losing pride of place as the radical theory of profit and prices. Without a labour theory of value, much of crisis theory inherited from Marx in turn became unsustainable.

Finally, despite theoretical commitment to a path-dependent theory of historical change, radical economics tended to fall back into an overly deterministic theory of history, in which capitalism moves according to knowable and immutable laws. The inconsistency of a deterministic theory of history with the notion of path dependence weakened claims to realism and superior understanding of capitalist development.

In addition to identifying such theoretical vulnerabilities of radical theory, critics argued for recognition of new empirical realities in capitalism. Most important has been the growing belief that class is simply too crude a tool with which to analyse conflict. Inequality along many dimensions, particularly race and gender, has emerged as a topic of radical investigation, further undercutting a classical Marxian analysis of capitalism. Radical feminists, for example, question the relevance of exploitation of labour to understanding inequality within the household. Resilience of capitalism in the face of crisis, in contrast to fragility of socialist countries, confronted radicals with another dilemma, since radical theory based on historical materialism had predicted opposite results.

Absence of radical analysis from policy debates fuelled the criticism that radicals neglected empirical work. Complaints have been widespread, coming even from the more Marxist of the radicals. Howard Sherman, for example, has argued that radicals are obliged ‘to always start from actual problems, not from ideal models, universal laws or any rigid rules of research’ (Sherman, 1995, p. 262). Empirical Marxists bemoaned radicals’ insistence on the purity of Marxian categories, which led radicals to avoid empirical analysis, resulting in exclusion from policy discussions (Dunne, 1991). Prominent non-Marxist radicals too have argued that radicals must confront reality through policy-relevant research if they are to claim superior reality of their analysis (Reich, 1995).

Contours of change

Three shifts in the body of radical research most clearly embody these critiques of early radical theory. First, characterization of the core injustice of capitalism has moved from a narrowly defined concept of exploitation at the point of production to broader inequality beyond both production and class. Second, the role of individual choice relative to structure has become more prominent in explanations of the nature and development of capitalism. The third shift marks movement away from a still ‘virtual’ radical

representation of capitalism to ever more realism.

The prominence given now to inequality signals a further loosening of bonds to Marxian theory. Inequality encompasses race and gender relations as sites of conflict independent of class. As a result, class is now one of several fault lines along which capitalism is both unequal and fragile, and cannot be said to be determinate of relations in other spheres of conflict. Attention to agency reflects that radicals now take seriously the need to demonstrate the inapplicability of *homo economicus* either at the point of production or at other sites such as the household. Thus, the new radical economics unpacks the ‘black boxes’ of the Marxian theory from which it originated. Gone are the stylized facts of homogenous workers confronting homogeneous capitalists or of a unitary household providing a sanctuary from, as well as valueless inputs to, capitalist production. On the agenda now are, for example, the effects of altruism on intra-household allocation decisions and the role of race in worker decisions to participate in strikes.

Similarly, a key point of the shift to the ‘real’ is to examine actually existing capitalist countries rather than construct an ideal capitalist type in relation to which actual economies are to be interpreted. The implication for policy is clear. Capitalist countries differ; some provide considerable room for redress of inequality and reduction of conflict, hence improved efficiency. Rather than condemning capitalism as an abstract, necessarily exploitative system, radicals increasingly focus on specific sites within capitalism where potential for improvement may be found. Again, even radicals more firmly within a Marxian tradition call for and applaud a new determination to ‘come to grips with … the realities of contemporary capitalism as opposed to the creation of a “virtual world”•’ (Fine, 2002, p. 2062).

Contested evolution

These three evolutionary shifts, while evident throughout radical research, are nonetheless both contested and incomplete. Movement towards fulfilling the promise of more realistic representations of capitalism has raised thorny issues of subject and method. Radicals now contend over the appropriate definition of reality and the ability of competing theories and methods to represent reality.

Radical analysis of globalization provides one example of competing visions of what is real and of methods appropriate to represent reality. On this topic there is still general agreement on the core principles of a radical theory and on the inadequacies of mainstream theory. Radicals reject the so-called Washington consensus, a set of liberalizing policies derived from the claim that opening of markets will lead to improved efficiency and economic growth. Radicals counter both efficiency and growth propositions of the liberalizing story, while focusing instead on distributional consequences of globalization. (Arestis and Sawyer, 2002, and Baker, Epstein and Pollin 1998, provide overviews of critiques of liberalization.)

Beyond unity in opposition to the free market position, however, radicals disagree profoundly on critical features of globalization. The appropriate definition and indicators of globalization are in dispute, meaning that radicals cannot agree on such basic issues as the degree of globalization of the world economy and role of nation states versus transnational corporations (Glynn and Sutcliffe, 1999).

Differences also are apparent in competing theoretical understandings of how globalization affects distribution and growth.

The Asian Tigers' perspective on globalization focuses on aggregate demand as the engine of growth and

maintains that aggregate demand must be sustained domestically, requiring both controls on capital flows and avoidance of excessive competition (Crotty, 2001; Baker, Epstein and Pollin, 1998). In contrast to the Washington consensus perspective, unemployment in this view is caused by insufficient aggregate demand rather than inflexible labour markets. The role of the state in a globalized economy, often modelled on recent Korean experience, is not only to manage domestic demand but also to develop non-traditional export industries through central control of credit. Class conflict plays at most a peripheral role. Instead, the main locus of conflict is between developed capitalist and developing countries.

This nation-state version of the Keynesian approach calls for controls on capital flows in part to sanitize domestic economics from the effects of ‘hot’ money. Within the same general aggregate demand framework, however, others deny that nation states have the capacity to control capital flows. A ‘one-world’ position, for example, calls instead for an international financial authority to regulate financial flows. This authority further must intervene even into the shaping of firm-level decisions to achieve consistency between firm incentives and international financial stability (Eatwell and Taylor, 2002).

An alternative radical vision disputes the relevance of a Keynesian model in the current historical conjuncture because history is not reversible. If history is indeed path-dependent, a theory grounded in history and institutions must conclude that it is not possible to revive the golden age of capitalism through a return to Keynesian policy. The current leaden age cannot be regilded to reproduce the growth of the post-war period exactly because the institutions that once supported growth through demand management policy are not sustainable in the globalized economy. In the new economic environment, it is not possible to reconstruct the institutions that once fuelled growth, namely, concentrated industrial sectors, strong trade unions (at least in manufacturing sectors), growing export markets and a world trade system denominated in dollars.

The golden age is also questioned for lack of attention to intra-country conflict of class, gender or race. The nation state is not a homogeneous unit in which aggregate growth benefits all. Moreover, this position contends that a new world order has emerged in which transnational corporations have the character and goals of a supra-national capitalist class. If capitalism has moved to a fundamentally new structure in which nation states have lost the power to determine policy for good, the restorative power attributed to Keynesian macro policy is a delusion.

A third radical vision takes a more positive view of globalization by examining intra-country class relations. This approach sees globalization as offering new space to national policies that can simultaneously increase growth and improve distribution. Focusing on capitalist inefficiency due to conflict, Bardhan, Bowles and Wallerstein (2006) conclude that pressures of globalization can force national governments to implement policies that both improve efficiency and reduce inequality. While the Bardhan et al. story rests on microeconomic behaviour, at the macro level, too, national policy and internal class dynamics are linked. The main contention is that poor countries too often merely use protection from the world economy to redistribute income and assets from the poor to the rich (see for example Griffin, 1998). Globalization can undercut the entrenched power of exploitative elites and hence enhance equality as well as growth.

The last divide among radicals to note here is disagreement about the role of foreign investment. The dominant position for some time has been that foreign investment is detrimental to developing countries because of concentration in low-wage industries and competition with existing domestic production. The stylized facts of this view are that foreign investment is footloose and relentlessly pursues ever-lower

wages. Therefore, on balance, foreign investment creates negative rather than positive spillover effects. Recently, however, several radical research projects have challenged the accuracy of this picture, pointing to a different set of realities. Nations are seen to have leverage over the terms and flow of foreign investment. Leverage in turn comes from political stability and skilled workers, which, rather than low wages, are the major determinants of investment patterns. Moreover, open countries have succeeded and closed countries have failed, depending upon the role of the domestic government in exploiting opportunities of openness (Chang, 1998).

Nihilism or high theory?

The globalization debate highlights the point that as radical theory has attempted to move towards less ‘virtual’ representations of capitalism, hitherto concealed splits have surfaced. Much of the purported greater realism comes from analytical methods based in mathematical or formal models, forcing radicals to confront underlying divisions about the proper role of mathematics and analytical models. Competing radical positions now contend over fundamental issues of method, particularly over whether radical core principles are sustained or abandoned by use of formal analytical techniques.

Debate over use of game theory and experimental economics is one illuminating example of current controversy. If the radical project includes understanding endogeneity of preferences and providing a more realistic concept of individual rationality than neoclassical economics, game theory would seem to provide a powerful tool. Feminist economics in particular seeks to move beyond a *homo economicus* model of behaviour to demonstrate that ‘many alternatives exist to the traditional self-interested model, with motivations responding, for instance, to notions of altruism, fairness, and reciprocity’ (Beneria, 1999, p. 71; see also, Folbre and Goodin, 2004). Further, policy relevance is the explicit goal of much radical experimental economics. Cross-country experiments are seen to yield insights to improve ‘the design of institutions and contracts, the allocation of property rights, the conditions for successful collective action …’, all considerations dear to radical economists (Henrich et al., 2001, p. 76).

Game theory from this perspective achieves greater realism by identifying parameter values or relationships from which a range of outcomes, here varying preferences or levels of collective action, may arise. To its proponents game theory offers a mechanism for demonstrating that there are no immutable laws or behaviours applicable to all times and all places. Norms and preferences are endogenous to and, therefore, vary with institutional arrangements, whether across countries or households.

Critics vehemently oppose game-theoretic models of preferences or norms exactly on grounds of insufficient realism. In this view, any reduction of complex reality to a model privileging at most a few variables and relationships is antithetical to the radical claim to realism. Models which use any form of optimization are considered to be ahistorical and lacking in institutional specificity, including work by new institutionalists like Bowles and Gintis (1998). Ben Fine, among many others, asserts that such work ‘relies on utility, production, inputs and informational asymmetries, timeless and rootless optimizing of individuals …’ (Fine, 2002, p. 2060).

Equilibrium, too, is a matter of much dispute. The strongest repudiation of equilibrium comes from the recent temporal single system interpretation of Marx, which maintains that incorporating history into theory requires neither equilibrium nor disequilibrium models, but non-equilibrium (Kliman and

McGlone, 1999). To many radicals the complexity of history and the necessity of non-equilibrium render inadmissible use of techniques such as optimization or simultaneous equation models (Lawson, 2006).

Defenders of the use of analytical models and mathematics respond that all analysis, mathematical or not, requires the same process of model construction in simplifying reality to a small set of main variables and relationships. The narrow neoclassical concept of equilibrium, with implausible assumptions about foresight and information, not equilibrium itself, must be abandoned. Without some notion of equilibrium, theory is simply nihilistic: ‘the social world is complex and determinate but it is impossible to say anything systematic about it’ (Foley, 2003, p. 3). An alternative equilibrium concept, compatible with heterodox theory and goals, can be defined ‘in the general sense of the balancing of forces within a particular model…’ without any market clearing or settlement to ‘tranquil states’ (Dutt, 1994, p. 3).

Adding fuel to the debate over the requirements of radical theory is the proposition that theories are converging and radicals are all post-Walrasians now. Convergence contends that mainstream economics is no longer bound to what Colander, Holt and Rossiter (2004) and Colander (2005) call the Walrasian unholy trinity of rationality, equilibrium and greed. Method and message are no longer linked because new methods explicitly are not ahistorical and asocial and hence are appropriate for radical inquiry (Gibson, 2005). Michael Reich has argued further that the liberal wing of neoclassical economics now can accommodate analyses of ‘disequilibrium economics, non-market-clearing equilibria, multiple equilibria and the new institutional economics, which have brought radical economics ‘out of the ghetto’ and into the liberal mainstream’ (Reich, 1995, p. 50). Duncan Foley expands this point, asserting that complexity and chaos theory finally can liberate radicals from dependence on the concept of ‘determination in the last instance’ or functionalist arguments to close the system (Foley, 2003).

The convergence contention poses a sharp choice for radicals. If new methods like complexity and chaos theory are not just consistent with but necessary for preservation of core radical principles, radicals who repudiate new methods are abandoning history, institutions and conflict as central concerns. The other pole of the dilemma emerges starkly from Colander's otherwise sympathetic assessment of post-Walrasian theory. While supporting Foley's call for more sophistication in radical analysis, Colander's contrasting conclusion is that, in the face of increasing complexity, recent theoretical innovations cannot provide a guide to policy but only an ‘aid to one's intuition’ (Colander, 2005, p. 23). With more realistic representations of capitalism the very complexity of the analytical tools means that no precise policy can be devised and we can only, as Colander says, muddle through.

Conclusion

Radical economics remains, as it was in the mid 1980s, a body of thought defined by common core principles while divided on method. What has changed is the depth of division across the several strands of radical economics. Attempts to develop more realistic and nuanced analyses of capitalism, together with the emergence of new methods of analysis, have generated sharp conflict over both method and object of radical inquiry. To the extent that the choice facing radicals is indeed nihilism or high theory, the centre of the paradigm is eroding and common ground is being lost. Nonetheless, the shifts in radical thought since the mid-1980s have yielded significant positive results. Radicals are indeed more involved

in policy discussions and more engaged with data, achieving successful policy interventions as exemplified by living wage legislation (Pollin, 2002). Self-criticism also has opened space for re-examination of basic tenets of radical theory and energized debate, which bodes well for continued dynamism and evolution of the broad radical paradigm.

See Also

- economy as a complex system
- family economics
- feminist economics
- institutional economics
- Keynesianism
- Marx's analysis of capitalist production
- methodology of economics
- Post Keynesian economics
- postmodernism
- social norms

Bibliography

Arestis, P. and Sawyer, M. 2002. *Neo-Liberal Economic Policy*. Cheltenham: Edward Elgar.

Baker, D., Epstein, G. and Pollin, R. 1998. *Globalization and Progressive Economic Policy*. Cambridge: Cambridge University Press.

Bardhan, P., Bowles, S. and Wallerstein, M. 2006. *Globalization and Egalitarian Redistribution*. New York: Russell Sage Foundation.

Beneria, L. 1999. Globalization, gender and the Davos man. *Feminist Economics* 5, 61–83.

Bhaduri, A. 2006. Endogenous economic growth: a new approach. *Cambridge Journal of Economics* 30, 69–84.

Bowles, S. and Gintis, H. 1998. *The Evolution of Strong Reciprocity*. Amherst, MA: Department of Economics, University of Massachusetts.

Bowles, S. and Jayadev, A. 2005. Guard labor. Working paper, Department of Economics, University of Massachusetts.

Chang, H.J. 1998. Globalization, transnational corporations and economic development: can the

developing countries pursue strategic industrial policy in a globalizing world economy? In *Globalization and Progressive Economic Policy*, ed. D. Baker, G. Epstein and R. Pollin. Cambridge: Cambridge University Press.

Colander, D. 2005. Post-Walrasian macro policy and the economics of muddling through. In *Interactions in Analytical Political Economy: Theory, Policy and Applications*, ed. M. Setterfield. Armonk, NY: M.E. Sharpe.

Colander, D., Holt, R.P.F. and Rossiter, J.B., Jr. 2004. The changing face of mainstream economics. *Review of Political Economy* 16, 485–99.

Crotty, J. 2001. Core industries, coercive competition and the structural contradictions of global neo-liberalism. Working paper, Department of Economics, University of Massachusetts.

Dunne, P. 1991. *Quantitative Marxism*. Cambridge: Polity Press.

Dutt, A. 1994. Analytical political economy: an introduction. In *New Directions in Analytical Political Economy*, ed. A. Dutt. Aldershot: Edward Elgar.

Dutt, A. 2005. On Post-Walrasian economics, macroeconomic policy and heterodox economics. In *Interactions in Analytical Political Economy: Theory, Policy and Applications*, ed. M. Setterfield. Armonk, NY: M.E. Sharpe.

Eatwell, J. and Taylor, L. 2002. *International Capital Markets*. Oxford: Oxford University Press.

Fine, B. 2002. Economics imperialism and the new development as Kuhnian paradigm shift? *World Development* 30, 2057–70.

Folbre, N. and Goodin, R. 2004. Revealing altruism. *Review of Social Economy* 62, 1–25.

Foley, D. 2003. *Unholy Trinity: Labor, Capital and Land in the New Economy*. London: Routledge.

Gibson, B. 2005. Thinking outside the Walrasian box. In *Interactions in Analytical Political Economy: Theory, Policy and Applications*, ed. M. Setterfield. Armonk, NY: M.E. Sharpe.

Glynn, A. and Sutcliffe, R. 1999. Still underwhelmed: indicators of globalization and their misinterpretation. *Review of Radical Political Economics* 31, 111–31.

Griffin, K. 1998. Comment on Harris and Mitchie ‘The Effects of Globalization on Policy Formation in South Africa’. In *Globalization and Progressive Economic Policy*, ed. D. Baker, G. Epstein and R. Pollin. Cambridge: Cambridge University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. and McElreath, R. 2001. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economics Association Papers and Proceedings* 91, 73–8.

Kliman, A. and McGlone, T. 1999. A temporal single-system interpretation of Marx's value theory. *Review of Political Economy* 11, 33–52.

Lawson, T. 2006. The nature of heterodox economics. *Cambridge Journal of Economics* 30, 483–506.

Pietrykowski, B. 2000. A primer in political economy. In *Political Economy and Contemporary Capitalism*, ed. R.Baiman, H. Boushey and D. Saunders. Armonk, NY: M.E. Sharpe.

Pollin, R. 1997. *The Macroeconomics of Saving, Finance and Investment*. Washington, DC: University of Michigan Press.

Pollin, R. 2002. What is a living wage? Considerations from Santa Monica, CA. *Review of Radical Political Economics* 34, 267–73.

Reich, M. 1995. Radical economics: successes and failures. In *Heterodox Economic Theories*, ed. F. Moseley. Aldershot: Edward Elgar.

Searle, J.R. 1995. *The Construction of Social Reality*. London: Penguin.

Setterfield, M. 2005. Interactions in political economy: an introduction. In *Interactions in Analytical Political Economy: Theory, Policy and Applications*, ed. M. Setterfield. Armonk, NY: M.E. Sharpe.

Sherman, H. 1995. *Reinventing Marxism*. Baltimore, MD: Johns Hopkins University Press.

How to cite this article

Flaherty, Diane. "radical economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000004> doi:10.1057/9780230226203.1379

The New Palgrave Dictionary of Economics Online

Rae, John (1796 – 1872)

K.H. Hennings
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bentham, J.; Böhm-Bawerk, E. von; capital accumulation; capital goods; capital theory; conspicuous consumption; division of labour; free trade; infant-industry protection; internal rate of return; inventions; Lauderdale, Eighth Earl of; liquidity preference; Rae, J. (1798–1872); Schumpeter, P. A.; Smith, A.; social preferences; social rate of return; technical progress; time preference; Veblen, T.; Wicksell effects

Article

John Rae was born in Aberdeen on 1 June 1796 into a merchant and shipping family. He graduated from the University of Aberdeen in 1815 and read medicine in the University of Edinburgh, but had to abandon his studies when his father's business failed in 1817. He emigrated to Canada in 1822 and turned to medical practice (whence ‘Dr’ Rae) and school teaching. He also participated in public affairs, but his career was shattered in 1848 when he was dismissed on spurious grounds after becoming embroiled in controversies about church control of education. Rae set out to start a new life first in California, and then the Hawaiian island of Maui. After another 20 years of farming, teaching, providing medical services to the natives, and serving as district judge and notary public, Rae went to live with a former pupil in New York, where he died on 12 July 1872.

None of Rae's many misfortunes and distractions could quell his scientific curiosity. He reported scientific experiments and inventions, lectured on scientific subjects, and wrote on public affairs, geology, and Polynesian language and customs (James, 1965). The only book he ever managed to get published, his *Statement of Some New Principles on the Subject of Political Economy* (1834), originally intended as an appendix to a larger work on the natural history and statistics of Canada, is one of the highlights of classical economic theory.

Rae's economics is rooted in a natural history of man which he had conceived in the tradition of Montesquieu, Turgot and the Scottish Enlightenment, but never came to execute. Political power and economic progress are seen to result not from the pursuit of self-interest, but to require ‘social instincts’ which create ‘an intelligent and moral community’ that furthers both the ‘effective desire of accumulation’ and the ‘rational spirit of invention’. Charging Adam Smith with building his system

exclusively on the pursuit of self-interest, and neglecting the role of inventions, Rae contended that economic activity is based primarily on an unselfish regard for the future. In consequence Rae emphasized the temporal aspect of economic activity, and developed a theory of capital accumulation and technical progress which goes far beyond what can be found in Adam Smith or other classical writers.

In language which Fisher was to take up, Rae argued that ‘provident forethought’ leads man to create ‘instruments’, that is, capital goods, in order to change the course of events. The sum total of such instruments constitutes the wealth of a society. All instruments are formed, directly or indirectly, by labour; all have the capacity to provide, directly or indirectly, for future wants; and they need time before they are finally exhausted (land being a special case). Rae assumes that the cost of production, and capacity, of any instrument can be measured, in a given society, in exogenously given wage units. All instruments whose capacity exceeds their cost of production can ‘be arranged in … a series, of which the orders are determined, by the proportions existing between the labour expended in the formation of instruments, the capacity given to them, and the time elapsing from the period of formation to that of exhaustion’ (1834, p. 100). Rae expresses this ‘order’ by the time which elapses before the instrument has yielded twice its cost of production, that is, by n in the expression $(1+r)^n=2$ where r is the internal rate of return of the quasi-rents associated with the instrument. Rae rejected working with the latter because it leads, in his view, to the assumption that the stock of all instruments is ‘an homogeneous quantity’ which he ‘found to be the foundation of much of the contradictions, in which the reasonings on these subjects are involved’ (1834, p. 197). His calculation rests on the assumption that every instrument can be associated with a unique rate of return. This need not be the case, but the possible multiplicity of internal rates of return does not affect his argument.

Rae argued that with knowledge stationary, both capital widening and capital deepening (that is, increasing the durability of instruments) necessarily lower the internal rate of return. Nevertheless, capital goods will be created as long as their internal rate of return is higher than the ‘effective desire of accumulation’, or rate of time preference, which Rae also expresses in time periods, that is, by m in the expression $(1+s)^m=2$ where s is the rate of time preference. Such time preference exists because life is finite and its end uncertain, and because ‘passion’ is often stronger than ‘reason’. But it is counteracted by the concern for future generations, or what Rae called ‘social and benevolent affections’ (1834, p. 122), which depend on a healthy climate that increases life expectancy, or on social circumstances such as internal and external security, good government, and so on. Hence the strength of the ‘effective desire of accumulation’, which Rae considers as much a social habit as an individual inclination, varies from one society to another. Variations from one person to another, Rae shows in an almost neoclassical manner, will be equalized by the exchange of instruments among them, so that a social rate of time preference can be juxtaposed to an internal rate of return which is equalized across different ‘employments’ by profit-seeking ‘merchants’.

Rae defines the equality of the social rate of return with the social rate of time preference as a stationary state in which accumulation ceases. ‘Gravitation’ towards it is slow. In a comparative static analysis Rae shows that the division of labour – which he views as a consequence of the accumulation of capital rather than its cause, as Adam Smith did – reduces the time for which instruments lie idle, and consequently increases their quasi-rents; hence more instruments can be created before the stationary state is reached, and wealth is increased. Similarly, foreign trade is said to increase the productivity of instruments, while conspicuous consumption (his term) will lower the effective desire of accumulation.

Rae also argues that as accumulation proceeds, more and more wealth will be tied up in instruments of increasing durability; hence the value of cash balances, and thus liquidity preferences, will increase. But far and away the most important factor making for changes in the progress of accumulation was in Rae's view the progress of inventions. Apart from raising quasi-rents, and hence the internal rate of return, and thus providing scope for more accumulation, inventions also raise the value of existing capital goods. Obviously assuming that these Wicksell effects were positive, Rae placed such capital ‘augmentation’ alongside capital accumulation as a factor in creating wealth. Indeed, Rae ascribes to inventions a more important role for economic progress (and thus the creation of political power) than capital accumulation, and criticizes Adam Smith for emphasizing savings too much, and neglecting technical progress.

The policy conclusions Rae draws from his analysis are also used to controvert Adam Smith. Instead of pursuing a policy of non-intervention, the ‘legislator’ should stimulate foreign trade and technical progress, encourage the transfer of knowledge, tax luxuries, and use tariffs to protect infant industries. It was in this sense that Rae tried to expose ‘the fallacies of the system of free trade, and of some other doctrines maintained in the *Wealth of Nations*’, as he announced on his title page. But, issued in the midst of a protectionist campaign, Rae's book was mistaken as a heavy-going anti-free-trade tract, and ignored. It did find a champion in Nassau Senior (Bowley, 1937, ch. 4) and through him in J.S. Mill, who quoted from it copiously in his *Principles* (1848), comparing Rae on accumulation to Malthus on population. But there the matter rested, except that it seems to have had a strong influence on Hearn's *Plutology* (1863). Rae was re-discovered by Mixter (1897) as a forerunner of Böhm-Bawerk, who acknowledged him as such (1900, ch. XI) despite some criticism. Together with a (botched, because re-arranged) reprint of Rae's book by Mixter (Rae, 1905), this brought Rae's work to the attention of capital theorists such as Irving Fisher (1907; 1930) who dedicated one of his main works to Rae, as well as Wicksell and Åkerman. It also influenced Schumpeter's (1911) concept of economic development, and Veblen's (1899) notion of conspicuous consumption.

In his criticism of Adam Smith, Rae did not go beyond Bentham (1787) and Lauderdale (1804). But he added poignancy because he derived it from a theory of economic development which was altogether novel. Based on a materialist conception of capital and a vintage-type approach complete with the distinction between capital goods and their value, Rae clearly separated the supply of from the demand for capital goods, and investigated their determinants. He saw but dimly the equality between discounted marginal returns and marginal costs, but he was clear about the equality of opportunities to invest to the ‘inclination … to yield up a present good’. He was quite clear, too, about the equality between the rate of return on capital and on money, and about what brought about such equalities: and also about the effects technical progress and the growth of knowledge have upon both demand and supply of capital. All this adds up to a remarkably original and creative performance which was, like that of Gossen, Cournot or Thünen, ahead of its time.

Selected works

1834. *Statement of Some New Principles on the Subject of Political Economy, Exposing the Fallacies of the System of Free Trade, and of some Other Doctrines Maintained in the ‘Wealth of Nations’*. Boston: Hilliard Gray & Co.

1905. *The Sociological Theory of Capital, being a Complete Reprint of the New Principles of Political Economy 1834*, ed. C.W. Mixter. New York: Macmillan.

Bibliography

Åkerman, G. 1923–4. *Realkapital und Kapitalzins*. Stockholm: Centraltryckeriet.

Bentham, J. 1787. *Defence of Usury*. In *Jeremy Bentham's Economic Writings*, vol. 1, ed. W. Stark. London: George Allen and Unwin, 1952.

Birchler, U.W. 1980. *John Rae (1796–1872). Seine Theorie der wirtschaftlichen Entwicklung*. Bern: Lang.

Böhm-Bawerk, E. von. 1900. *Kapital und Kapitalzins*, vol. 1, *Geschichte und Kritik der Kapitalzines-Theorien*. 2nd edn, Innsbruck: Wagner. Trans. as: *Capital and Interest, vol. 1, History and Critique of Interest Theories*. South Holland, IL: Libertarian Press, 1959.

Bowley, M. 1937. *Nassau Senior and Classical Economics*. London: George Allen & Unwin.

Edmonson, N. 1970. John Rae and liquidity preference. *History of Political Economy* 2, 432–40.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Fisher, I. 1930. *The Theory of Interest*. New York: Macmillan.

Hearn, W.E. 1863. *Plutology, or the Theory of the Efforts to Satisfy Human Wants*. Melbourne: Robertson.

James, R.W. 1965. *John Rae, Political Economist. An Account of his Life and a Compilation of his Main Writings*. 2 vols. Toronto: University of Toronto Press.

Lauderdale, [J. Maitland] Earl of. 1804. *An Inquiry Into the Nature and Origin of Public Wealth*. Edinburgh: Constable.

Lehmann, H. 1937. *John Raes Werk, seine philosophischen und methodologischen Grundlagen*. Dresden: Dittert.

Mill, J.S. 1848. *Principles of Political Economy*. Ed. W.J. Ashley. London: Longmans, 1909.

Mixter, C.W. 1897. A forerunner of Böhm-Bawerk. *Quarterly Journal of Economics* 11, 161–90.

Mixter, C.W. 1902. Böhm-Bawerk on Rae. *Quarterly Journal of Economics* 16, 385–412.

Robbins, L. 1968. *The Theory of Economic Development in the History of Economic Thought*. London: Macmillan.

Schumpeter, J.A. 1911. *Theorie der wirtschaftlichen Entwicklung*. Leipzig: Duncker & Humblot. Trans. as *The Theory of Economic Development*, Cambridge, MA: Harvard University Press, 1934.

Spengler, J.J. 1959. John Rae on economic development: a note. *Quarterly Journal of Economics* 73, 393–406.

Veblen, T.B. 1899. *The Theory of the Leisure Class*. New York: Macmillan.

How to cite this article

Hennings, K. H. "Rae, John (1796–1872)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000005> doi:10.1057/9780230226203.1380

The New Palgrave Dictionary of Economics Online

Rae, John (1845 – 1915)

Andrew Skinner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Keywords

Rae, J. (1845–1915); eight-hour movement

Article

John Rae was born in Wick, Caithness, Scotland on 26 May 1845, the eldest son of William Rae who was for some years Provost of the town. Rae received his early education at Edinburgh Academy, before proceeding to university in the city. He graduated in 1866 with first class honours in philosophy. Rae was awarded an honorary doctorate in 1897 by his alma mater. He died on 19 April 1915, having spent the last 15 years of his life in London, and was buried in Wick.

John Rae has been variously described as ‘author and journalist’ (1953, p. 582) and as ‘economist, writer on socialism’ (1966, p. 1057). He was certainly all of these, publishing numerous articles, notably in the *Fortnightly Review* (1885), the *Temple Bar* (1882, 1883, 1897), *MacMillan's Magazine* (1893) and the *National Review* (1889). The great bulk of his considerable output is to be found in the *Contemporary Review* (from 1880) of which he was assistant editor.

Rae's contributions to the *Contemporary Review* disclose an interest in at least five major areas. These include a review of contemporary literature on social philosophy (in seven parts), and a number of articles on the Socialism of Karl Marx and the Hegelians, Christian Socialism in Germany, and State Socialism and Social Reform. Rae also contributed articles on the crofting problem in the Highlands, supplementing these with pieces on the Highland Shealing (*Temple Bar*, 1883) and on the Scotch Village Community (*Fortnightly Review*, 1885). Rae wrote a number of articles on taxation and a review of recent economic literature. Finally, he addressed questions of industrial relations, in considering the implications of the eight-hour day in the context of unemployment and of foreign competition.

Rae's journalistic interests resulted in three major books. The first of these was entitled *Contemporary Socialism* (1884). This was followed by *Eight Hours for Work* (1894), a book which consisted largely of his articles on labour questions, supplemented by chapters on the connection between hours and wages,

the eight-hour movement of 1833, and current legislative proposals.
John Rae is now best known for his admirable *Life of Adam Smith* (1895) which was favourably reviewed in *The Times* for 8 March 1895 as presenting a ‘vivid picture’ of his subject. The review also drew attention to the point that the book's real merit lay ‘not in the originality of the matter, but in the patient industry, with which Mr Rae has collected his materials, old and new, and in the skill and judgement with which he has presented them to the reader’.
While more critical of Rae's scholarship (1965, p. 12), Jacob Viner has noted that Rae was a trained writer who made his *Life* ‘an interesting and highly readable book’ (p. 13). Viner also drew attention to the remarkable fact that ‘As a comprehensive biography, it had no substantial predecessor. Seventy years after its publication, it still has no substantial successor’ (1965, p. 5). These judgements are still valid.

See Also

- Smith, Adam

Selected works

1884. *Contemporary Socialism*. London and Wick: W. Isbister. 2nd edn, enlarged, London: Sonnenschein, 1891; 3rd edn (with additional chapter), London: Swan Sonnenschein, 1901; 4th edn, London: Swan Sonnenschein, 1908.

1894. *Eight Hours for Work*. London: Macmillan.

1895. *Life of Adam Smith*. London: Macmillan.

Bibliography

Viner, J. 1965. Guide to John Rae's *Life of Adam Smith*. In Rae's *Life of Adam Smith*. New York: Kelley.

The Wellesley Index to Victorian Periodicals, 1824–1900. 1966. vol. 1, ed. W.E. Houghton. Toronto: Toronto University Press.

Who was Who 1897–1915. 1953. Vol. 1, 4th edn. London: Adam & Charles Black.

How to cite this article

Skinner, Andrew. "Rae, John (1845–1915)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000006> doi:10.1057/9780230226203.1381

close print preview

The New Palgrave Dictionary of Economics Online

Ramsey model

David M. Newbery
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

arbitrage; Bellman's dynamic programming; bliss; calculus of variations; capital gains; capital theory; commodity taxation; dual economies; exhaustible resources; golden rule; Hamiltonians; heterogeneous capital goods; intertemporal competitive equilibrium; intertemporal welfare maximization; investment; Keynes–Ramsey rule; life-cycle savings; marginal utility of consumption; neoclassical growth theory; optimal accumulation; optimal growth; optimal monetary policy; optimal saving rate; optimal taxation; planning; Pontryagin maximum principle; population growth; production functions; public investment criteria; Ramsey model; Ramsey, F. P.; rate of interest; social welfare function; stationary state; taxation of capital

Article

Frank Plumpton Ramsey died at the age of 26 after making brilliant contributions to philosophy mathematical logic, and, of course, economics. His two contributions to economics both appeared in the *Economic Journal*, then edited by J.M. Keynes. The first, ‘A Contribution to the Theory of Taxation’, published in March, 1927, laid the foundation for the modern theory of commodity taxation. The second, the subject of this entry, was ‘A Mathematical Theory of Saving’, published in December, 1928. Keynes, in his obituary notice published two months after Ramsey's death, in the *Economic Journal* of March, 1930, described the latter as ‘one of the most remarkable contributions to mathematical economics ever made, both in respect of the intrinsic importance and difficulty of its subject, the power and elegance of the technical methods employed, and the clear purity of illumination with which the writer's mind is felt by the reader to play about its subject’.

Ramsey asked how much of its income should a nation save and derived a remarkably simple rule, usually known as the Keynes–Ramsey rule, as Keynes provided a non-technical argument for the result. The rule states that the rate of saving, multiplied by the marginal utility of consumption, should always be equal to the amount by which the total net rate of enjoyment of utility falls short of the maximum possible rate.

Ramsey's formulation of the problem served as a model for almost all subsequent studies of optimal

economic growth, and, with the critical addition of a growing population, might have created neoclassical growth theory about 30 years before Solow's (1956) contribution. He assumed a one-good world, in which labour with a stock of capital would produce a flow of output, part of which was consumed, and the balance was saved and thereby added to the stock of capital. The objective, or criterion, was to achieve the maximum level of enjoyment, summing over all time, where enjoyment was the utility of consumption, $U(C)$, less the disutility of working, $V(L)$. Ramsey made three crucial assumptions which together allowed him to solve explicitly an otherwise intractable problem. He assumed that there was no population growth, no technical progress, and no discounting of utility, ‘a practice which is ethically indefensible and arises merely from the weakness of the imagination’ (Ramsey, 1928, p. 543). He further supposed that there was a ‘maximum *obtainable* rate of enjoyment’ called *Bliss*, *B*, either because of capital or consumption saturation. As neither population grows nor future utilities are discounted, Ramsey then argues, rather informally, that it must be desirable to save enough to eventually reach bliss, or approximate to it indefinitely. To stop short means forgoing a finite amount of utility, which, summed over an infinite time horizon, is infinitely costly. Formally, Ramsey deals with this problem of a potentially unbounded integral of utility (summed without discounting over infinite time) by *minimizing* the amount by which enjoyment falls short of bliss integrated throughout time:

$$\min \int_0^\infty [B - U(C) + V(L)]\,dt \tag{1}$$

subject to

$$\frac{dK}{dt} + C = F(K, L). \tag{2}$$

Ramsey attacks the problem from two directions: economic and mathematical. His economic argument first solves for the relationship between consumption and the effort by equating the marginal disutility of labour to the product of the marginal product of labour and the marginal disutility of consumption. He then solves the basic arbitrage relationship equating the marginal utility of consuming a unit now with the marginal utility of consuming the product of investing the unit until the next instant of time. This key relationship implies that the marginal utility of consumption, $U'(C)$, must fall at the rate of interest, equal to the marginal product of capital, $\partial F/\partial K$. These two conditions, together with (2), the initial stock of capital, and a terminal condition as $t \to \infty$, produce a differential equation which can be integrated to give the result.

The mathematical approach observes that the calculus of variations gives the first two conditions directly, but also observes that the variable of integration in (1) can be changed from *t* to *K* by using (2) to give

$$\min \int_{K_0}^{\infty} \frac{B - U(C) + V(L)}{F(K, L) - C} dK \tag{3}$$

and since *C* and *L* are arbitrary functions of *K* all that is needed to minimize the integrand is to set its partial derivations to zero. Differentiating with respect to *C* gives

$$F(K, L) - C = \frac{B - [U(C) - V(L)]}{U'(C)}. \tag{4}$$

The left-hand side of (4) is the rate of investment or saving, while the right-hand side is equal to bliss minus the additional rate of enjoyment, divided by the marginal utility of consumption, and the whole is the Keynes–Ramsey rule.

Ramsey concluded from this rule that the optimal rate of saving should be ‘greatly in excess of that which anyone would normally suggest’ and gave an illustration in which the savings rate should be 60 per cent of income. One of the main themes explored by later writers was whether this was a robust conclusion, or whether the optimal rate of saving was very sensitive to the simplifying assumptions – a theme which is discussed below. Ramsey recognized that discounting utility would destroy the simple reasoning which led to (4), and was thus anxious to have an ethical reason for rejecting it. He believed that population growth would argue for higher rates of saving whilst technical progress would have ambiguous effects – as proved to be the case in later formal models.

Ramsey drew attention to two remarkable features of the rule. The first is that the level of saving does not depend on the production function. The second is that it does not depend on the rate of interest, unless this is actually zero. In fact, the first feature is only apparently the case, for in (4), *C* will depend on the level of output, *F*, and since savings, given by the right-hand side, also depends on *C*, it will depend on *F*. In his Section III, Ramsey clearly pointed out that the level of saving was motivated by the demand for future consumption, while the rate of interest was determined by the current stock of capital (in this one-sector model). In a concluding remark to this section he notes that ‘in the accounting of a Socialist State the function of the rate of interest would be to ensure the wisest use of existing capital, not to serve in any direct way as a guide to the proportion of income which should be saved’. The second result does not survive in more general models which allow for utility discounting. Nevertheless, the arbitrage relationship does suggest a way in which the rate of interest can guide the rate of saving. If

the rate of decline of the marginal utility of consumption is less than the rate of interest, taken to be the rate of return on investment, then the rate of saving is too low, and vice versa.

The main contribution of the paper was to pose a fruitful question – what should the rate of savings be – and propose a method of analysis – that of intertemporal welfare maximization using the techniques of dynamic optimization, in this case the calculus of variations. The main result was striking – the rate of saving should apparently be rather high. In addition to this contribution, the paper also contained various remarkable extensions. It considers the choice of savings rate for an individual facing constant factor prices, who wishes to optimize his lifetime consumption pattern, and as such provides a positive theory of life-cycle savings. It shows that if utility is to be discounted, then it must be discounted at a constant rate if one is to escape the contradiction ‘that successive generations are motivated by the same system of preferences’. Later, Strotz (1956) would return to this issue and the related problem of dynamic consistency. Finally, Ramsey shows that if a society consists of individuals who differ in their rate of discount, and if it is in steady state, then the equilibrium would be attained by a division of society into two classes, the thrifty enjoying bliss and the improvident at the subsistence level. In short, he characterizes the long-run general equilibrium of a society of heterogeneous individuals.

Ramsey thus laid the foundations for the study of optimal accumulation and optimal growth, as well as the positive theory of savings and the rate of interest. Space precludes a full assessment of the subsequent work his paper stimulated, though Burmeister and Dobell's (1970) textbook lists 107 references in their chapter on optimal economic growth, and much has happened since that date. Instead we shall briefly mention some of the themes of this subsequent work.

Ramsey's model represented a significant advance on the classical analysis of stationary states, since it made possible the analysis of non-stationary time paths of capital accumulation, but ultimately his model would tend towards a stationary state. With the development of growth theory the profession acquired a more appealing concept of long-run equilibrium – that of steady growth. In due course this suggested the obvious extension to Ramsey's model of incorporating these dynamic features – population growth at the steady rate n and Harrod-neutral technical progress at a steady rate g. The instantaneous level of national welfare was variously taken as $U(C_t)$, $U(C_t/L_t)$, or, most satisfactorily, $L_t u(C_t/L_t)$, where L_t was the total population or workforce, and C_t was total consumption. Since welfare now depended on time, it made no drastic difference to include a utility discount rate, δ, and to propose a more general objective such as

$$W = \int_0^\infty U(C_t, t)\mathrm{d}t = \int_0^\infty L_t u(C_t / L_t)\mathrm{e}^{-\delta t}\mathrm{d}t. \tag{5}$$

Steady growth now raised the question of the *existence* of an optimal savings policy in an acute form, for the integral in (5) might diverge unless δ was sufficiently large. Ramsey had faced a similar problem and avoided it by minimizing the shortfall from a reference path (or bliss). Similar devices were invoked to deal with divergent integrals, and much effort was expended on devising criteria of optimality and categorizing conditions under which an optimal savings plan existed, though many apparently

reasonable problems nevertheless failed to possess an optimal savings plan, as Hammond and Mirrlees (1973) demonstrated. (They also give references to earlier discussions of the problem of non-existence.) They observe that no restrictions on the class of utility function will ensure existence, nor will any realistic restrictions on the production assumptions by themselves be enough to avoid the problem. They then argue that if we could specify date after which events are of no significance, then the problem reduces to a finite horizon model, for which the utility integral would converge. Different people might disagree on the horizon date, but if the initial T_0 years of the plan were relatively insensitive to any horizon date later than some date T_1, then everyone would agree with the T year plan, and, in their language, the plan would be *agreeable*. Hammond and Mirrlees show that in the one good model with a general instantaneous utility function $U(C_t, t)$ the agreeable path is unique and locally optimal, and that if an optimal path exists it is agreeable. Establishing the existence of agreeable paths is, however, considerably easier than establishing the existence of optimal paths.

While existence problems are important and raise intriguing philosophical problems (what if optimal growth paths do not exist?), they are not central to the economics of the problem. One of the key issues that has engaged the attention of subsequent researchers is whether the optimal savings rate is indeed as high as Ramsey argued (though, as Samuelson, 1969, pointed out, Ramsey's conclusion depended on a particular choice of utility function). Certainly, Tinbergen (1956) was inclined to agree, but Mirrlees (1967) argued that Ramsey's model was seriously misleading, and that once population growth, technical progress and utility discounting were admitted, the initial value of the optimal rate of saving was typically very different from that implied by the Keynes–Ramsey rule. Once time enters the production function, it is no longer possible to obtain explicit solutions and an alternative solution strategy is required. Mirrlees argued that it was preferable to find the asymptotic form of the optimally developing economy in which output, consumption and consumption per head all grow at steady rates along a ‘modified Golden Rule’, and in which the savings rate is constant. The initial value of the savings rate could then be estimated by expanding around this asymptotic solution.

Mirrlees, in common with a large number of other optimal growth theorists, used a particular utility function – the iso-elastic form

$$u(c) = -c^{1-v},\ v > 1,\ = \log c,\ v = 1 \tag{6}$$

for which Ramsey's rule gives a savings rate of $1/v$ (providing an optimum exists). Mirrlees was impressed that for plausible values of the parameters of his model, the optimum savings rate was very different from the Ramsey value, and might be quite low. He also pointed out that the asymptotic solution, or the ‘modified Golden Rule’, would differ from the Golden Rule (according to which the rate of savings should equal the share of profit), if utilities were discounted – for the obvious reason that one would expect optimum policies to reflect the values regarding the distribution between generations.

Ramsey's model made skilful use of the classical calculus of variations, and in that vein Samuelson and Solow (1956) extended the model to deal with heterogeneous capital goods. In so doing they made

possible two notable contributions to capital theory. The first was to argue that on the optimum path it was not too misleading to think in terms of an abstract quantity of capital – heterogeneity did not significantly alter the Ramsey theory. Second, the Hahn–Samuelson problem of the indeterminacy of equilibrium with capital heterogeneity disappeared on the optimal path, though the significance of this did not emerge until the paper by Hahn (1966).

As Samuelson and Solow pointed out, the classical calculus of variations could be replaced by Hamiltonian methods which would be able to deal with inequality constraints. The powerful techniques of the Pontryagin Maximum Principle and Bellman's Dynamic Programming were in due course applied to various extensions of the Ramsey problem to good effect, and their advantages and interrelationships are well discussed in the textbook of Intriligator (1971). In both approaches shadow prices or co-state variables play an important role both in characterizing the solution and in demonstrating the relationships between optimality, intertemporal efficiency, and a set of intertemporal (shadow) prices (prices on futures markets) which might be used to decentralize the optimum. These shadow prices have a natural interpretation, for they value the capital stock in terms of the objective function, that is social welfare or the utility of consumption. The price guides the instantaneous allocation of output between consumption and investment, for consumption should be increased, if possible, until its value (the marginal utility of consumption) falls to the value of investment, that is, of the capital stock. The evolution of the price over time then satisfies the fundamental arbitrage relationship, so that asset holders obtain a return (including capital gains) on the asset equal to the return on other assets and to the return from delaying consumption.

The strengths of these alternative approaches are best appreciated in multisector models when there are constraints on reallocating resources. If investment goods are physically different from consumption goods, and capital is immobile between sectors, then savings will be constrained by the feasible output of the investment goods sector, and the planners' problem is primarily one of choosing the allocation of investment between the two sectors. In such a model the rate of return on capital will depend on the level of investment, and Ramsey's observation that in his model the two are independent is shown to be a feature of the one-good assumption. With two sectors corner solutions are quite likely (in the early stages) and the inequality constraints require the extra power of the new approaches.

The shadow prices are arguably most useful for cost–benefit analysis, rather than the more ambitious planning problems which so engaged the attentions of optimal growth theorists in the 1960s. Little and Mirrlees (1969; 1974) and Newbery (1972) and Stern (1972) were concerned to develop methods for calculating shadow or accounting prices in dual economy models of developing countries in which the level of aggregate savings was constrained. The two key accounting prices on which optimal growth models can shed light are the wage rate and the rate of discount to use in investment projects. The former emerges from the constraints on the allocation of labour and on the level of wages which must be paid, while the latter is again given by an arbitrage relation, or the rate of change of the shadow price of capital itself. The arbitrage equation gives a differential equation for the shadow price which, together with the equation for saving and the accumulation of capital, can be numerically integrated backwards from the asymptotic solution. Modern computers allow this to be done quickly, as illustrated in Newbery (1972).

The arbitrage equation comes into its own in exhaustible resource models where the return to the exhaustible resource must, while it remains in the ground, take the form of a capital gain equal to the

return on other assets. This rule, due originally to Hotelling (1931), and nicely exposited by Solow (1974), has achieved prominence since the dramatic rise in the oil price of 1973–4.

Although the revival of interest in the Ramsey model in the 1960s was initially motivated by the post-war popularity of national economic planning, a popularity which waned rapidly in the 1970s, the model and its successors remain useful for the more modest aims of characterizing intertemporal competitive equilibrium in asset markets, especially for exhaustible resources like oil and gas, and for providing a more satisfactory neoclassical theory of equilibrium growth with individually rational savers. The common feature of Ramsey's two contributions to economics was that they were normative, and postulated an additive (utilitarian) social welfare function as the objective to be maximized. Several writers have taken the natural step of combining both of Ramsey's two interests and enquiring what optimal tax (and monetary) policy should be in an intertemporal model in which savings and investment are affected by these policies. Arrow and Kurz (1969) were the first to explore these issues and the closely related issues of the problem of public investment criteria systematically in a growth model in which full optimality is not achieved.

Diamond (1973) extended their work to a model with many goods, and demonstrated the desirability (under constant returns) of equal efficiency, on average, between public and private production, even though aggregate efficiency was not desired. In particular the public and private sectors should use the same discount rates. Later work (surveyed, for example, by Kotlikoff, 1984) has explored the efficiency losses involved in an economy of intertemporal optimizing individuals in the presence of distortionary taxes on capital, and have used these estimates to rank alternative capital tax reform programmes – a compromise between the optimal tax approach of Diamond and the need to incorporate more of the complex features of particular economies.

In short, if the central question which Ramsey addressed of the right level of saving and investment has fallen from favour recently, nevertheless the spirit of the Ramsey model with its emphasis on intertemporal optimization lives on strongly, whether it be in the study of the oil market, the derivation of public investment rules, or the reform of the corporate tax system.

See Also

- calculus of variations
- neoclassical growth theory

Bibliography

Arrow, K. and Kurz, M. 1969. Optimal public investment policy and controllability with fixed private savings ratio. *Journal of Economic Theory* 1, 141–77.

Burmeister, E. and Dobell, A.R. 1970. *Mathematical Theories of Economic Growth*. New York: Macmillan.

Chakravarty, S. 1969. *Capital and Development Planning*. Cambridge, MA: MIT Press.

Diamond, P.A. 1973. Taxation and public production in a growth setting. Ch. 10 in Mirrlees and Stern (1973).

Hahn, F.H. 1966. Equilibrium dynamics with heterogeneous capital goods. *Quarterly Journal of Economics* 80, 633–46.

Hammond, P.J. and Mirrlees, J.A. 1973. Agreeable plans. Ch. 13 in Mirrlees and Stern (1973).

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Intriligator, M.D. 1971. *Mathematical Optimization and Economic Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Kotlikoff, L.J. 1984. Taxation and savings: a neoclassical perspective. *Journal of Economic Literature* 22, 1576–629.

Little, I.M.D. and Mirrlees, J.A. 1969. *Manual of Industrial Project Analysis for Developing Countries. Vol. 2: Social Cost Benefit Analysis*. Paris: OECD Development Centre.

Little, I.M.D. and Mirrlees, J.A. 1974. *Project Appraisal and Planning for Developing Countries*. London: Heinemann.

Mirrlees, J.A. 1967. Optimum growth when technology is changing. *Review of Economic Studies* 34, 95–124.

Mirrlees, J.A. and Stem, N.H., eds. 1973. *Models of Economic Growth*. London: Macmillan.

Newbery, D.M.G. 1972. Public policy in the dual economy. *Economic Journal* 82, 567–90.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–59.

Samuelson, P.A. 1969. Foreword to Chakravarty (1969).

Samuelson, P.A. and Solow, R. 1956. A complete capital model involving heterogeneous capital goods. *Quarterly Journal of Economics* 70, 537–62.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Solow, R.M. 1974. The economics of resources or the resources of economics. *American Economic Review* 64(2), 1–14.

Stern, N.H. 1972. Optimum development in a dual economy. *Review of Economic Studies* 39(2), 171–84.

Strotz, R.H. 1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23(3), 165–80.

Tinbergen, J. 1956. The optimal rate of savings. *Economic Journal* 66, 603–9.

How to cite this article

Newbery, David M. "Ramsey model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000009> doi:10.1057/9780230226203.1383

The New Palgrave Dictionary of Economics Online

Ramsey pricing

William J. Baumol

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

contestable markets; hyperplanes; inverse elasticity; lump-sum taxes; marginal cost pricing; optimal taxation; Pareto efficiency; Ramsey pricing; Ramsey, F. R.; second best

Article

Ramsey prices are prices that are Pareto optimal subject to a constraint on the total profits of a single supplier or group of suppliers. In particular, because a firm whose activities are characterized by scale economies will lose money if it sets the prices of its products equal to their marginal costs, Ramsey prices become for that firm the prices that are optimal (economically efficient) given the financial feasibility requirement that the firm's profits be non-negative. The same Ramsey prices can also be shown to be those necessary for maximization of the sum of consumers' and producers' surpluses.

The concept is named after Frank Ramsey, its discover, whose 1927 paper on the subject was one of several revolutionary contributions to economics, mathematics and philosophy this extraordinary man made before he died at the age of 26. Since then and until the 1970s, the principle was largely forgotten even though it was rediscovered and expanded upon by Pigou, Boiteux and Samuelson. In 1970 it was publicized and its history explored in an article by Baumol and Bradford, and the principle has since been widely recognized and accepted by economists and practitioners. As an illustration, in 1983 the Interstate Commerce Commission adopted Ramsey pricing as the underlying principle it would follow in the regulation of railroad rates.

Ramsey prices are an outstanding example of the use of pure economic theory to derive an operational solution to a difficult set of practical problems. It may also be as definitive as any available second-best theorem. The extraordinary achievement of the theorem lies in the very explicit formulae it is able to derive from so weak a premise – the Pareto optimality requirement that the prices be those which elicit such a set of outputs and purchase quantities that it is impossible to increase the welfare of any one individual without harming anyone else. Aside from the apparent weakness of this assumption, the definitive character of the Ramsey theorem is surprising in light of the conclusion suggested by much of the second-best literature, that where additional constraints are superimposed on the usual requirements

of optimality, one can expect no simple and straightforward results to emerge.

The Ramsey theorem and its interpretation

The Ramsey theorem is expressed in a variety of formulae all of which are essentially equivalent. Perhaps its simplest form asserts that when a producer supplies *n* commodities then Pareto optimality subject to a profit constraint requires the prices, p_j of these goods to satisfy

$$\frac{p_j - mc_j}{p_n - mc_n} = \frac{mr_j - mc_j}{mr_n - mc_n}, \ (j = 1, \ \ldots, \ n - 1),$$

$$\sum_{j=1}^{n} p_j y_j = c(y_1, \ \ldots, \ y_n) + k$$

(1)

where mc_j and mr_j are, respectively, the marginal cost and marginal revenue of output *j*, $c(\cdot)$ is the supplier's total cost function and *k* is any constant.
In the special case in which none of the seller's goods is either a complement or a substitute in demand, the preceding relationship is easily shown to take the special form which is widely known as ‘the inverse elasticity formula’:

$$\frac{(p_j - mc_j) \ / \ p_j}{(p_n - mc_n) \ / \ p_n} = \frac{E_n}{E_j}, \ (j = 1, \ \ldots, \ n - 1),$$

$$\sum p_j y_j = c(\cdot) + k,$$

(2)

where E_j is the price elasticity of demand for product j.

In the particular case where an optimum satisfies locally the requirements of constant returns to scale, so that marginal cost pricing yields zero economic profits exactly, then (for k=0) conditions (1) and (2) are automatically transformed into the marginal cost pricing conditions

$$p_j = mc_j, \ (j = 1, \ldots, n). \tag{3}$$

It is easy to show that no prices can be Pareto optimal subject to the profit constraint indicated unless they satisfy (1). Moreover, as long as the proper concavity–convexity conditions hold, any prices which satisfy (1) will be consistent with Pareto optimality so constrained.

One can suggest in rough intuitive terms why constrained Pareto optimality requires prices which satisfy (1), or (2) – in the case of demand independence. The latter is perhaps the most illuminating case, and so it is useful to summarize the argument briefly.

As a starting point, one should recall that the reason marginal cost pricing is necessary for a ‘first best’ (unconstrained) optimum is that such prices equate the pecuniary cost to the consumer of purchasing an additional unit of the item and the economic cost of producing it, that is, its marginal cost . Thus, when the consumer selects his purchases so as to maximize the utility he derives from a given outlay of money, he thereby automatically maximizes the utility derivable from a bundle of economic resources.

However, where returns to scale are not constant at the vector of purchases elicited by the prices p_i=mc_i then the requirement $\sum p_i y_i = c(\cdot)$ will be violated by those marginal cost prices. Consequently, prices will have to deviate from marginal costs in some pattern that satisfies the profit constraint. Of course, every such deviation will affect consumer purchases, and so the quantities produced, making them depart in different degrees from the optimal quantities that would have been selected under marginal cost pricing. The objective is to cause the p_i to deviate from the mc_i in a manner that satisfies the profit constraint and yet distorts consumer purchases from their optimal levels as little as possible.

For this purpose, consider two of the pertinent commodities, i and j, with i's demand highly elastic and j's very inelastic. Start with p_i=mc_i and p_j=mc_j and assume that at those prices profits are negative. Because of the high demand elasticity of i a small rise in p_i above mc_i will cause a relatively large ‘distortion’ in consumer demand from its Pareto optimal quantity. Moreover, also because of the high elasticity, the rise in p_i will yield a relatively small increase in revenue to help eliminate losses. In contrast, a similar percentage increase in p_j will cause a smaller percentage change in quantity of j demanded and a larger gain in revenue. Clearly less damage will be done to welfare if a larger share of the task of meeting the shortfall of total revenue relative to total cost is carried out via a rise in p_j, the price of the commodity with the more inelastic demand. This is, in essence, the logic of the inverse elasticity formula.

Informal derivation of the theorem

A simplified and rather informal derivation of the formulae is straightforward. For brevity only a single consumer and a single input, labour, is used in the following, but the proofs in the *k* consumer – *m* input cases are virtually identical. Let

- y_i=the supplier's output of i (i=1, …, n)
- x=the vector of outputs of all other goods
- R=the available quantity of resource
- r=unused resource (leisure)
- p_i=the price of i
- w=wage (price of leisure)
- $U(y_1, \ldots, y_n, x, r)$=the consumer's utility function
- $C(y_1, \ldots, y_n)$=the firm's input requirement function
- $K(x)$=the input requirement for production of x. Then, optimality requires maximization of

$$U(y_1, \ldots, y_n, x, r)$$

subject to the resource constraint

$$C(y_1, \ldots, y_n) + K(x) + r = R$$

and the budget constraint

$$\sum p_i y_i = wC(y_1, \ldots, y_n).$$

This yields the Lagrangian

$$L = U(\cdot) + \alpha[R - C(\cdot) - K(\cdot) - r] + \beta[\sum p_i y_i - wC(\cdot)]$$

Using the notation U_i for $\partial U / \partial V_i$, $U_r = \partial U / \partial r$, and so on, we have the first order conditions

$$U_i - \alpha C_i + \beta(mr_i - wC_i) = 0, \tag{4}$$

where $mr_i = \partial \Sigma p_i V_i / \partial V_i$ is the marginal revenue of *i*, and

$$U_r - \alpha = 0. \tag{5}$$

Since consumer equilibrium requires equality between price ratios and marginal rates of substitution (the ratios of marginal utilities) we have

$$\frac{U_i}{p_i} = \frac{U_r}{w} = k, \ (i = 1, \ \ldots, \ n)$$

so that (5) yields $U_i = kw = \alpha$, and therefore (4) becomes

$$p_i - wC_i + (\beta / k)(mr_i - wC_i) = 0, \tag{6}$$

which, writing $mc_i = wC_i$, yields the general Ramsey formula (1). To obtain the inverse elasticity formula (2) we simply use a standard relationship for the case of independent demands,

$$mr_i = p_i(1 - 1) / E_i),$$

substituting this into (6) we have

$$p_i - mc_i = (\beta / k)(p_i - mc_i - p_i / E_i),$$

or

$$(1 - \beta / k)(p_i - mc_i) = -(\beta / k) p_i / E_i$$

which immediately yields (2).

Applications

Aside from its obvious connection with pricing by the firm, the theorem also has applications to the principles of taxation and to the general equilibrium analysis of the economy. Indeed, Frank Ramsey presented his result as a theorem on taxation rather than pricing. The point is that, the theoretical concept of lump-sum taxes aside, any tax must be a levy on some sort of economic activity. Even if the price of that activity's product is equal to its marginal cost, the tax will in general drive a wedge between the two, particularly if the total tax revenue is required to meet some particular target. Thus the problem of determining the vector of tax rates on the economy's activities that will meet the overall revenue target with minimum social welfare loss is equivalent to determining the optimal vector of deviations between prices and marginal costs that will satisfy that revenue (budget) constraint. In sum, the search for the optimal (budget constrained) prices and the optimal (revenue constrained) tax rates are formally equivalent.

The analysis also has direct implications for general equilibrium theory, for it tells us that if lump-sum taxes are impossible, then a vector of (first-best) marginal cost prices may also be ruled out for the economy as a whole. Indeed, such a first-best parametric price solution is possible only if, at the corresponding vector of activity levels (outputs), the production frontier happens to be locally linear and homogeneous (meaning, in the differentiable case, that it must be tangent to a hyperplane through the origin in input–output space).

For suppose this is not so – say, that there are increasing returns to scale at any such point. Then marginal cost pricing will yield negative profits for the economy, and suppliers as a class will be able to survive financially with such prices only if they receive subsidies. But subsidies must be paid for by taxes, and any such taxes on activities whose pre-tax prices equal their marginal costs must yield after-tax prices which do not. In sum, one cannot escape the problem of finding the deviations of prices from their ‘first best’ magnitudes which meet the budget requirement that every subsidy must be covered by tax revenues. This, then, is the inescapable Ramsey problem for the entire economy if prices are parametric and no optimal output vector is a point of (at least local) linear homogeneity.

The case of diminishing returns poses corresponding problems, even though it is often thought to be compatible with competitive equilibrium and marginal cost pricing. As long as input quantities (including the input of entrepreneurship) can be expanded, marginal cost pricing will be incompatible

with equilibrium at an optimal point because marginal cost pricing will then yield positive economic profits and the number of firms will therefore increase. There can be *no* finite number of firms at which this manifestation of disequilibrium ceases unless marginal cost pricing is abandoned. But then the best equilibrium prices in terms of Pareto optimality must again be the Ramsey prices.

In sum, Ramsey pricing is no mere artifact of regulation of industry or tax policy. It is deeply embedded in the logic of the general equilibrium mechanism.

History of Ramsey analysis

The basic theorem apparently first appeared in Frank Ramsey's classic article (1927). While the article has sometimes attracted the attention it deserved, it did not effectively convey to the profession the wider implications of its second-best pricing analysis. In 1928 A.C. Pigou, who had apparently posed the original issue to Ramsey, published a restatement of the theorem. Here, too, it was presented as a result on the principles of taxation and not related to pricing. Ursula Hicks (1947) independently provided a similar discussion.

Perhaps the first work on Ramsey theory that was expressed in terms of pricing issues occurred in the aftermath of Hotelling's (1938) classic paper on marginal cost pricing. There the author had advocated a system of subsidies to firms subject to scale economies, but he himself came to recognize the tax implications and the consequences for the overall optimality of the solution. He and J.R. Hicks discussed the problem, and Hicks emerged with an independently discovered Ramsey theorem, which was never published.

Early after the Second World War, two major contributions were made to the literature. Paul Samuelson (1951) prepared a memorandum for the US Treasury pointing out the logic of the Ramsey approach to taxation. As is to be expected, Samuelson's contribution was highly sophisticated and offered substantial original insights, but, although widely circulated in public finance circles, it was never published. After having published a less sophisticated version of the theorem in 1951, Marcel Boiteux, Directeur-Général of Électricité de France 1967–87, published a major article on the subject in 1956. It explicitly dealt with the topic as an issue in pricing policy for nationalized or regulated firms and derived its results directly from a Pareto optimality model. Moreover, it provided a result more general than the inverse elasticity form of the theorem on which Ramsey and Pigou had focused.

An even deeper exploration of the subject was provided by Diamond and Mirrlees (1971) as part of their continuing work on the theory of optimal taxation. Their papers are important not only because of their careful analysis but also because they played a major role in bringing the subject to the profession's attention. Within a year or two of the appearance of their articles and that of Baumol and Bradford (1970), ‘everyone’ in the profession was fully aware of the notion of Ramsey pricing and its logic. Since then there has been an explosion of writings on the subject and it occurs centrally or peripherally in a wide variety of fields.

An illustrative and perhaps surprising application which suggests the unexpected places in which the construct can turn up, is the ‘weak invisible hand theorem’, that occurs in the contestable markets literature (see Baumol, Bailey and Willig, 1977). That theorem states that if a monopolist who is constrained by a regulatory (or other) profit ceiling chooses to adopt the Ramsey price vector rather than some other set of prices that enable him to earn his allowed return, then under a fairly attractive set of

assumptions the monopolist will be rewarded for his virtuous decision by being protected from entry by those prices. In other words, self-interest may impel a monopolist to adopt Ramsey prices because those prices are *sustainable* against entry, meaning that at those prices the monopolist will earn the profits that the constraint allows to him, but any rival firm that undertakes to enter the field will be predestined to lose money even if the incumbent undertakes no strategic (retaliatory) response.

Today Ramsey pricing is accepted as a basic proposition of microanalysis and appears with great frequency in new writings on the theory of the firm, industrial organization and public finance; it recurs regularly in the pricing discussions of American regulatory agencies.

See Also

- optimal taxation

Bibliography

Baumol, W.J. and Bradford, D.F. 1970. Optimal departures from marginal cost pricing. *American Economic Review* 60, 265–83.

Baumol, W.J., Bailey, E.E. and Willig, R.D. 1977. Weak invisible hand theorems on the sustainability of prices in a multiproduct monopoly. *American Economic Review* 67, 350–65.

Boiteux, M. 1956. Sur la géstion des monopoles publics astreints à l’équilibre budgétaire. *Econometrica* 24, 22–40.

Diamond, P.A. and Mirrlees, J.A. 1971. Optimal taxation and public production: II. *American Economic Review* 61, 261–78.

Hicks, U. 1947. *Public Finance*. London: Nisbet.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6, 242–69.

Pigou, A.C. 1928. *A Study in Public Finance*. London: Macmillan.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

Samuelson, P.A. 1951. Theory of optimal taxation. Unpublished memorandum for the US Treasury.

How to cite this article

Baumol, William J. "Ramsey pricing." The New Palgrave Dictionary of Economics. Second Edition.

Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000010> doi:10.1057/9780230226203.1384

The New Palgrave Dictionary of Economics Online

Ramsey, Frank Plumpton (1903 – 1930)

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bayes, T.; Bernoulli, D.; bootstrap; choice under uncertainty; De Finetti, B.; Dutch books; growth theory; optimal taxation; probability; Ramsey, F.P.; Sraffa, P.; subjective probability; von Neumann, J.; von Neumann–Morgenstern; Wittgenstein, L.

Article

There are interesting parallels in the careers of Frank Ramsey and John von Neumann. Each was born in 1903, one the product of the ‘High Intelligentsia of England’ (Keynes, 1933, p. vii) and the other the son of a wealthy banker in Budapest (Ulam, 1976, p. 79). Each was a creative mathematician of high order but each also made major contributions to at least two other disciplines. Each wrote just three papers in economic theory, all six of which were of fundamental importance. Moreover, with one exception every one of these seminal papers had to wait many years for its proper recognition; even the exception – the utility theory set out in the Appendix to von Neumann and Morgenstern (1947) – at first encountered serious misunderstanding within the profession. Indeed, considering them purely as economists, one wonders how these two geniuses would fare today, when promotion and tenure so often depend on a good immediate showing in citation indexes and the like.

The three papers of Ramsey are in subjective probability and utility (1926), optimal taxation (1927), and optimal one-sector growth (1928), while those of von Neumann are in game theory (1928), optimal multi-sector growth (1937; 1945–6), and objective probability and utility (1947). It is quite striking that their work both on growth theory and on choice under uncertainty should be so complementary, especially since there is no evidence that von Neumann knew of Ramsey's work in either field.

Another and grievous similarity was that both men died early, Ramsey on 19 January 1930 of complications associated with jaundice, and von Neumann (twice Ramsey's age) on 8 February 1957 of cancer. Both losses were tragic, especially that of the 26-year-old Frank Ramsey, whose ‘death at the height of his powers deprives Cambridge of one of its intellectual glories and contemporary philosophy of one of its profoundest thinkers’ (Braithwaite's Introduction to Ramsey, 1931, p. ix).

Life

Frank Plumpton Ramsey was born in Cambridge on 22 February 1903. His father was a mathematician, Fellow and later President of Magdalene College (Harrod, 1951, pp. 141, 320), and his brother Michael became Archbishop of Canterbury. He was educated at Winchester and at Trinity College Cambridge, and was a Scholar of both those ancient foundations. In the autumn of 1924 he became Fellow of King's College and University Lecturer in Mathematics and soon afterwards married Lettice Baker, who had been a student in the Moral Sciences Tripos. After his death she became a founder of Ramsey and Muspratt, a firm of portrait photographers that has long been an Oxbridge institution. She survived into the 1980s, in vigorous old age.

In physical appearance Ramsey was tall and portly, the latter a feature he shared with von Neumann; ‘I take no credit for weighing nearly 17 stone [238 pounds]’ (1931, p. 291). All accounts agree as to his simplicity and modesty, qualities which are happily reflected in his engaging literary style. ‘Ramsey reminds one of Hume more than of anyone else, particularly in his common sense and a sort of hard-headed practicality towards the whole business’ (Keynes, 1933, p. 301). But his unfailing cheerfulness did not disguise ‘the amazing, easy efficiency of the intellectual machine which ground away behind his wide temples and broad, smiling face’ (Keynes, 1933, p. 296). ‘He comes down to earth, however, with a satisfying bump, and earth is certainly the natural element of my old friend Lettice’ (Partridge, 1981, p. 129).

Ramsey and Wittgenstein

For many years it was thought that while still an undergraduate Ramsey assisted in the translation of the German text of Wittgenstein's *Tractatus Logico-Philosophicus* (1922). It now appears that ‘the first draft of the translation was produced by F.P. Ramsey alone’ (von Wright, 1982, p. 102). Just 19, he dictated it directly to a stenographer in the University Typing Office in Cambridge in the winter of 1921–2 (reminiscent, on a smaller scale, of the 19-year-old ‘John S. Mill’ beginning in 1825 to edit Bentham's massive Rationale of Judicial Evidence). Wittgenstein seems to have been pleased with Ramsey's translation (1973, p. 77), and a fast friendship was thereby established between the two philosophers that lasted for the rest of Ramsey's short life.

In September 1923 the Tractatus had been published for almost a year. Not only had Ramsey been its main translator but he had also written a long and penetrating review of it for Mind (reprinted in 1931, pp. 270–86). But still there were many passages which remained unclear to him. To remedy this he made a special journey to Austria, where Wittgenstein was teaching in the local school of a small village and living in spartan conditions. The eccentric philosopher and the brilliant undergraduate hit it off immediately. Ramsey stayed two weeks, spending every afternoon from 2 to 7 elucidating the great man's work: ‘we get on about a page an hour’ (Wittgenstein, 1973, p. 79).

In the several letters that Ramsey afterwards wrote to Wittgenstein we can glimpse what Keynes meant in referring to ‘the simplicity of his feelings and reactions, half-alarming sometimes and occasionally almost cruel in their directness and literalness’ (Keynes, 1933, p. 296). Consider for example these passages from his letters of 12 November and 20 December 1923 (Wittgenstein, 1973, pp. 81–3):

> I have not been doing much towards reconstructing mathematics; partly because I have been reading miscellaneous things, a little Relativity and a little Kant, and Frege … But I am awfully idle; and most of my energy has been absorbed since January by an unhappy passion for a married woman, which produced such psychological disorder, that I nearly resorted to psychoanalysis, and should probably have gone at Christmas to live in Vienna for nine months and be analysed, had not I suddenly got better a fortnight ago, since when I have been happy and done a fair amount of work.
>
> I think I have solved all problems about finite integers, except such as are connected with the axiom of infinity, but I may well be wrong.
>
> [December 20th] I was silly to think I had solved those problems. I'm always doing that and finding it a mare's nest … *I have been trying to prove a proposition in the Mengenlehre* either $2^{\aleph_0} = \aleph_1$ *or* $2^{\aleph_0} \neq \aleph_1$, which it is no one knows but I have had no success. (His italics)

In 1924 Ramsey actually did spend six months in Vienna in psychoanalysis (rarer then than now), after which ‘I feel that people know far less about themselves than they imagine, and am not nearly so anxious to talk about myself as I used to be, having had enough of it to get bored’ (1931, p. 290). The mathematical problem referred to in his second letter was of course the famous Continuum Hypothesis. His lack of success in this is scarcely surprising, since in the 1960s Paul Cohen showed the Hypothesis to be an undecidable proposition within Zermelo–Fraenkel set theory (see, for example, Cohen, 1966). It was, incidentally, a continual disappointment to von Neumann that it was not he but his hero Kurt Gödel who made the startling discovery, in 1930–31, of the necessary existence of such undecidable propositions (Ulam, 1976, pp. 76, 80).

Wittgenstein returned to Cambridge early in 1929 and began those ‘innumerable conversations’ with Ramsey that are acknowledged in the Preface/Foreword (dated January 1945) to his *Philosophical Investigations* (1953, p. x). Unfortunately, these were cut short by Ramsey's tragic death, a moving account of which may be found in Frances Partridge's *Memories* (1981, pp. 169–82); the grieving Wittgenstein was at Ramsey's bedside in the hospital until a few hours before he died.

The only other person acknowledged by name in the Preface to the *Investigations*, and for even greater help than Ramsey gave, was Piero Sraffa. The trio of Ramsey, Sraffa and Wittgenstein must have been a formidable discussion group indeed; a treasured piece of Cambridge folklore is a lunch at which the three of them discussed Keynes's theory of probability with its author. The odd pattern of belated recognition of intellectual indebtedness was continued in Sraffa's acknowledgement (1960, pp. vi–vii) of Ramsey's help, a mere 30 years after the fact.

Works

Ramsey's early work in philosophy was a continuation of the methods of Russell and Whitehead's *Principia*, but it is clear that the influence of Wittgenstein and the evolution of his own thinking were

moving him towards the end of his life in a quite different, more pragmatic direction. These later contributions were left fragmentary and incomplete at his death, but a very brief account of them and their relations to modern philosophy may be gleaned from the first two Introductions to the revised edition (1978) of (1931).

In mathematics proper, as distinct from the foundations of mathematics, his main contribution is a fundamental theorem which appeared actually as a byproduct of a paper of 1928 on formal logic (reprinted in 1931, pp. 82–111). It reads (1931, p. 82):

Theorem A: Let Γ be an infinite class, and and r positive integers; and let all those sub-classes of Γ which have exactly r members, or, as we may say, let all r-combinations of the members of Γ be divided into μ mutually exclusive classes $C_i(i=1, 2,\ldots, \mu)$, so that every r-combination is a member of one and only one C_i; then, assuming the Axiom of Selections [i.e. the Axiom of Choice], Γ must contain an infinite sub-class Δ such that all the r-combinations of the members of Δ belong to the same C_i.

This beautiful result was ignored until 1935, when it was essentially rediscovered by Paul Erdös and Esther Szekeres. Gradually, it led to the formation of a subdiscipline of combinatorial analysis known as *Ramsey Theory*, which already contains many hundreds of papers and is growing at a remarkable rate (see the survey by Graham, Rothschild and Spencer, 1980).

Ramsey's pioneering paper on optimal taxation seems to have been written in response to a request by Pigou to look into the problem (see Pigou, 1928, pp. 126–8) but his work on the theory of growth was apparently his alone, although greatly admired by and discussed with Keynes.

Mathematical expectation, probability and utility

The present discussion of Ramsey's great Chapter VII (1931, pp. 156–98) will consider it quite narrowly, as a contribution only to the theory of choice under uncertainty, and thus neglect the important question of its relation to traditional theories of probability. Ramsey himself adopted throughout a modest and peaceable tone towards probability theory, stressing that ‘the meaning of probability in logic’ may be quite different from ‘its meaning in physics’ (p. 157).

The chapter is entitled ‘Truth and Probability’ and dated 1926; presumably most of it was written then, in spite of a reference which bears the date 1927. It contains almost all of what he has to say on the subject, although further on in Chapter VIII and pages 256–7 there are a few unsystematic comments and glosses on the earlier work. The first ten pages form a critique of Keynes's theory of probability (1921), which may well have stimulated his own interest in the whole subject, so it is not until Section 3 that Ramsey begins his ‘inquiry … [into] … the logic of partial belief’.

Ignoring here all his careful qualifications, the theory outlined in that Section begins as follows (pp. 172–4):

> The old-established way of measuring a person's belief is to propose a bet, and see what are the lowest odds which he will accept. This method I regard as fundamentally sound; … I propose to take as a basis a general psychological theory … that we act in the way we think most likely to realize the objects of our desires … The question then arises how … to take account of varying degrees of certainty in his beliefs. I suggest that we introduce as

> a law of psychology that his behaviour is governed by what is called the mathematical expectation; … We thus define degree of belief in a way which *presupposes* the use of the mathematical expectation. (Italics added).

Ramsey was fully aware of the crucial dependence of his approach on mathematical expectation. Later in the *Foundations* he asks: ‘The question … why just this law of mathematical expectation. The answer to this is that if we use probability to measure utility, as explained in my paper, then consistency [for which see below] requires just this law’ (p. 251).

Putting the matter in its crudest (and so necessarily inaccurate) form, mathematical expectation as a principle of choice involves the use for any risky line of action α of a ‘probability’ π_i and a ‘valuation’ ν_i attached to each of the possible outcomes α_i, that constitute α, in such a way that: (i) the expected valuation $E(\alpha)$ of α is $\Sigma \pi_i \nu_i$ (or an appropriate integral if α has infinitely many members, an alternative which Ramsey expressly rejects: pp. 183–4); and (ii) α is chosen rather than another risky line of action β if and only if $E(\alpha)>E(\beta)$.

Implicit in this crude form is a conflation between events and outcomes. Outcomes depend upon decisions and events, and it is in events and not outcomes that the randomness present is usually held to reside, so that given the occurrence of an event the relevant outcome on which it depends follows deterministically. Nevertheless, the randomness that inheres in the events may be transferred to the outcomes that are conditional upon those events. In the words of Arrow (1951; 1971, p. 26): ‘no matter how complicated the structure of a game of chance is, we can always describe it by a single probability distribution of the final outcomes.’

Notice that because mathematical expectation depends linearly both on the probabilities and on the valuations, choice that follows this principle is made according to a *bilinear* form; there is however no necessity for the valuations of the possible outcome themselves to depend linearly upon those outcomes. Essentially, given any two of the three concepts, mathematical expectation, probabilities and valuations, the remaining one follows more or less naturally. For example, in Daniel Bernoulli's account of the theory of risk (1738), the π_i are apparently given ‘objectively’, for example by the tosses of a coin. Wishing to preserve the principle of mathematical expectation, and citing the St Petersburg Paradox as evidence for the inappropriateness of using money itself as valuation, Bernoulli was thus led to a specific *utility function* to compute the correct valuations, this being a nonlinear (actually, concave) function of wealth. This did not in fact resolve the basic difficulty of the Paradox (which resides in unboundedness of the mathematical expectation) but it was a novel and important idea that was very influential.

A quite different approach was used by Bayes (1763; 1958), who actually *defined* probability in terms of mathematical expectation: ‘The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening’ (1958, p. 298; Jeffreys, 1961, pp. 30–4, stresses the similarity here between Bayes and Ramsey). Possibly in ignorance of the earlier contribution, Bayes retained monetary valuations rather than replace them by Bernoullian utilities.

Both authors regarded the maximization of the mathematical expectation of gain as the appropriate principle of choice in an uncertain situation. But whereas Bernoulli accepted probabilities from the outside and altered the meaning of valuations so as to achieve consonance between the maximization of

mathematical expectation and rational choice, Bayes started with the outside monetary valuations and thence determined probabilities so as to square rational choice with mathematical expectation.
Ramsey was more subtle. He effectively ‘bootstrapped’ both the valuations *and* the probabilities from mathematical expectation, at the small cost of: (*a*) a very general assumption about preferences; (*b*) an assumed existence of a certain kind of event; and (*c*) a further principle, original with him, that no agent's subjective probabilities should be *inconsistent*. To be inconsistent means that ‘He could have a book made against him by a cunning better and would then stand to lose in any event’ (1931; p. 182); this no-win situation is now usually called a *Dutch book*.

Sketch of a proof

Ramsey provided sufficient detail for a formal proof of the existence of valuations and probabilities to be constructed from his system of axioms, but he did not construct one himself. Such proofs, for varying circumstances, have been given by Davidson and Suppes (1956) and Vickers (1962), while more informal discussions may be found in Jeffrey (1965; 1983, ch. 3) and Luce and Suppes (1965, pp. 291–4). Only the merest sketch is attempted here, and its mild technical detail follows Davidson and Suppes (1956) rather than Ramsey's original treatment, which was couched mainly in the concepts of Wittgenstein's *Tractatus* and the language of Russell and Whitehead's *Principia*, both long since unfamiliar.
Ramsey begins by considering the case where the agent has ‘certain [that is, sure] beliefs about everything’. He then adopts assumption (*a*) above, which expressed in modern language says that the agent has a complete preference preordering over ‘all possible courses of the world … [though] … we … have no definite way of representing them by numbers’ (1931, p. 176). Vickers points out that if different preferences can themselves be parts of different ‘courses of the world’ then the argument is ambiguous, and if not then the question is begged (1962, pp. 6–11); however, he shows how to resolve these problems by suitable amendment of Ramsey's definitions.
When ‘the subject is capable of doubt’ (p. 177), the theory proceeds by offering options. Suppose that the agent has two options: the first is α, in which he receives *x* if an event *e* occurs and a preferentially different outcome *y* if it does not; and the other is β, in which he receives *r* if *e* occurs and another outcome s if it does not. Assuming that probabilities $\pi(e)$ and $\pi(e')$ can be attached to the events *e* and to e' (the complement of *e*), respectively, and that valuations $\nu(x)$, $\nu(y)$, and so on, can be placed on the outcomes *x*, *y*, *r* and *s*, then the principle of mathematical expectation says that α is better than, indifferent to, or worse than β, according as

$$\pi(e)v(x) + \pi(e')v(y) > ; = ; < \pi(e)v(r) + \pi(e')v(s). \quad (1)$$

Ramsey's next assumption is (*b*) above, to the effect that there exists some event, say e^*, such that for *every* pair (*m*, *n*) of preferentially distinct outcomes the subject is indifferent between the option γ

consisting of m if e^* and n if not e^*, and another option δ consisting of n if e^* and m if not e^*. According to the principle of mathematical expectation, this implies

$$\pi(e^*)v(m) + \pi(e^{*\prime})v(n) = \pi(e^*)v(n) + \pi(e^{*\prime})v(m). \tag{2}$$

Since m and n are preferentially distinct, their valuations must be such that $\nu(m) \neq \nu(n)$. Then from this and (2) it follows that necessarily

$$\pi(e^*) = \pi(e^{*\prime}). \tag{3}$$

Although quantitative probabilities have not yet been defined, (3) shows that there is a clear qualitative sense in which event e^* has a (subjective) probability of 1/2, *provided* that the subjective probabilities of an event and its complement sum to unity. Ramsey terms *ethically neutral* any event (in his language, prpt) that has the properties of e^*; the force of the word ‘ethically’ is not explained (p. 177). The assumption that such events (prpts) exist is perhaps the weakest part of his theory of choice under uncertainty, although before it is rejected out of hand the careful philosophical discussion of it by Vickers (1962) and the equally careful empirical applications of it discussed by Davidson and Suppes (1956) should be consulted.

Now take the case of (1) where e is an ethically neutral event e^* and the option α is indifferent to the option β. Then from (1) and (3),

$$v(x) - v(r) = v(s) - v(y). \tag{4}$$

This says that differences in valuations can be equated, so that the latter are measurable by an *interval scale*; or what comes to the same thing, that they are measurable up to choice of unit and origin, so that for any other such scale μ, $\mu(\cdot) = \alpha + bv(\cdot)$, where $b>0$.

A valuation having been obtained in this fashion for each outcome, and assuming again that for any event e, $\pi(e) + \pi(e') = 1$, it follows from the case of equality in (1) that

$$[\pi(e)]^{-1} = 1 + [v(x) - v(r)] / [v(s) - v(y)] . \quad (5)$$

This gives a way of calculating the subjective probability $\pi(e)$ of any event, ethically neutral or not, in a way compatible simultaneously with the principle of mathematical expectation and with the valuations $v(\cdot)$ of the possible outcomes. Thus both valuations ('utilities') and subjective probabilities have been bootstrapped, in that order, from the simple assumptions (*a*) and (*b*), plus the assumption that any event and its negation have subjective probabilities that add up to 1. Ramsey dispenses with this last, auxiliary assumption by means of his principle (*c*) of consistency, which in effect insists upon the impossibility of Dutch books.

Dutch books

Although his paper is crystal clear that consistency means that the subjective probabilities of any set of disjoint and exhaustive events must sum to 1, and it is twice stated explicitly (pp. 182–3) that anyone who is not consistent in this sense can have a Dutch book made against him, Ramsey provided no formal proof of equivalence between these two ideas. Hence this result is usually attributed to de Finetti (1937; 1964), who gave a very neat proof. Not having read Ramsey's paper, de Finetti like Bayes worked with monetary valuations in his account of personal probability, though he admitted later (1964, p. 102 fn(a)) that 'Such a formulation could better, like Ramsey's, deal with expected *utilities*; I did not know of Ramsey's work before 1937, but I was aware of the difficulty of money bets.' What follows is a free adaptation of de Finetti's proof to Ramsey's problem.

Let there be *n* mutually incompatible and together exhaustive events e_i, for example, the faces of a die. Suppose then that I, knowing your subjective probabilities π_i, offer you the following wager: If e_i occurs I pay you σ_i. In return, you pay me an initial stake of $\Sigma \pi_i \sigma_i$ valuation units, where the sum is taken over the *n* events. If you behave according to Ramsey's theory of choice under uncertainty, then you should be on the margin of accepting this wager, since for you to attach probability π_i to e_i is to say that you would be indifferent between the following offers: receive σ_i valuation units contingent on the occurrence of e_i, and the amount $\pi_i \sigma_i$ for sure. Since by hypothesis the e_i are exclusive events, the separate amounts $\pi_i \sigma_i$ may be added together.

If event e_h, occurs, your gain is

$$Y_h = \sigma_h - \sum \pi_i \sigma_i \quad h = 1, 2, \ldots, n. \quad (6)$$

These are *n* linear equations, which can be put into matrix–vector notation. Writing *g* and *s* for the

vectors of the γ_i and σ_i, respectively, I for the $n\times n$ identity matrix, and P for the matrix whose (i, j)th element is π_j, the equations (6) become

$$g = (I - P)s. \tag{7}$$

Computation shows that $\det(I–P)=1–\Sigma\,\pi_i$. So if $\Sigma\,\pi_i\neq 1$, then for any desired vector of gains g stakes $s=(I–P)^{-1}g$ can be computed that will guarantee me the vector $–g$. In particular, I can specify g to be strictly negative, thus ensuring that you will lose whatever event occurs.
Conversely, suppose that your subjective probabilities are what de Finetti called coherent (and Ramsey, consistent), so that by definition $\Sigma\,\pi_i=1$. Then, multiplying each equation in (6) by π_h and adding over all events,

$$\sum \pi_h \gamma_h = \sum \pi_h \sigma_h - \sum \pi_h \sum \pi_i \sigma_i = 0. \tag{8}$$

Since each $\pi_h \geq 0$ and their sum is non-zero, it follows from (8) that not all the γ_h can be negative. Hence the condition that your subjective probabilities π_i, sum to 1 for all complete sets of incompatible events e_i, that is, that you obey the rules of probability calculus, is necessary and sufficient in order that no Dutch book can be made against you.

The reception of ‘Truth and Probability’

Ramsey's theory of choice under uncertainty was deeply original. Emile Borel, in his review (1924; 1964) of Keynes's theory of probability, had earlier sketched an interesting theory of subjective probability in terms of bets (note in particular his remark that ‘the method of betting permits us in the majority of cases a numerical evaluation of probabilities that has exactly the same characteristics as the evaluation of prices by the methods of exchange’; 1964, p. 57), but nobody had come close to the depth and comprehensiveness of Ramsey's theory. He was characteristically modest about its range of application: ‘I only claim for what follows approximate truth … like Newtonian mechanics … [it] can, I think, still be profitably used even though it is known to be false’ (p. 173).
Perhaps because the theory was too original, such modesty did not help its author, any more than his high reputation as a philosopher. I can find no evidence that anyone, let alone any economist, took any serious notice of Ramsey's work until after von Neumann and Morgenstern's quite separate utility theory had appeared in 1947. The latter theory was very much in the Bernoullian tradition, in which the

probabilities are given from outside, ‘objectively’. Coupling these with a complete preference preordering for such alternatives, suitable continuity, and the principle of mathematical expectation in the form of the independence axiom, the authors were able to deduce the existence of a utility function, unique up to positive affine transformations, which gave valuations compatible both with the outside probabilities and that principle.

The first published reference to Ramsey's theory known to me appears in Little (1950, p. 29, fn.1), who considered it ‘essentially the same’ as that of von Neumann and Morgenstern. Little's reference was soon followed by one in Arrow (1951), who acknowledged that Ramsey was brought to his attention by Norman Dalkey. Though complaining that ‘Ramsey's work was none too clear’ (1971, p. 26), Arrow did see that it originated ‘a new stage’ in decision theory, ‘in which a priori probabilities are derived from behavior postulates’ (1971, p. 22). Thereafter there was a gradual increase in the appreciation of Ramsey's contribution, although even as late as 1954 an excellent collection of papers on decision theory (Thrall, Coombs and Davis, 1954) contained not one reference to his work.

It is a common mistake to suppose that the line of descent in the theory of personal probability is direct from Ramsey to de Finetti (1937) to Savage (1954). We have seen that de Finetti did not know of Ramsey's work, his own remarkable contribution being very much in the Bayesian tradition which takes the valuations from outside and thence derives the probabilities. Moreover, a careful reading of Savage's fine book shows that Ramsey's influence was at best peripheral, the axiomatization of probabilities and valuations proceeding far more along the lines developed by de Finetti.

There have in fact been relatively few explicit exponents of Ramsey's approach. The most notable are probably Davidson and Suppes (for example, 1956) and Anscombe and Aumann (1963), who used an interesting bootstrapping argument to go from assumed probabilities for what they called ‘roulette’ lotteries to valuations, and thence to subjective probabilities for the much wider class of ‘horse’ lotteries, all very much in the Ramsey manner.

The direct heirs to Ramsey's work have been few but there is no doubt that its influence has been pervasive, to such extent that chairs in decision theory at US business schools have been named after him (though with what warrant is hard to say). Arrow (1965, p. 57) claimed that all arguments involving the expected-utility hypothesis ‘are only variations of Ramsey's’, while Savage (1962, p. 10) wrote that the ‘more thorough-going … formulation of Ramsey (1931) … is in no way obsolete’. Even now, not to experience that ‘clear purity of illumination with which the writer's mind is felt by the reader to play about its subject’ (Keynes on Ramsey, 1928) is a sad loss for the modern student.

See Also

- expected utility hypothesis
- von Neumann, John

Selected works

1926. Truth and probability. Chapter VII, pages 156–98 in (1931). Reprinted in (1978) and in Kyburg and Smokler (1964).

1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61. Reprinted in (1978).

1928. A mathematical theory of saving. *Economic Journal* 38, 543–9. Reprinted in (1978).

1931. *The Foundations of Mathematics*, ed. R.B. Braithwaite, with a Preface by G.E. Moore. London: Routledge & Kegan Paul.

1978. *Foundations*, ed. D.H. Mellor, with Introductions by the editor, L. Mirsky, T.J. Smiley and J.R.N. Stone. London: Routledge & Kegan Paul.

Bibliography

Anscombe, F.J. and Aumann, R.J. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34, 199–205.

Arrow, K.J. 1951. Alternative approaches to the theory of choice in risk-taking situations. *Econometrica* 19, 404–37.

Arrow, K.J. 1965. *Aspects of the Theory of Risk-Bearing*. Helsinki: Yrjö Jahnsson Foundation.

Arrow, K.J. 1971. *Essays in the Theory of Risk-Bearing*. Chicago: Markham.

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53, 370–418. Reprinted in *Biometrika* 45 (1958), 293–315.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, Vol. V. Trans. as ‘Exposition of a new theory on the measurement of risk’, *Econometrica* 22 (1954), 23–36.

Borel, E. 1924. A propos d'un traité de probabilité. *Revue Philosophique* 98, 321–36. Trans. in Kyburg and Smokler (1964).

Cohen, P.J. 1966. *Set Theory and the Continuum Hypothesis*. New York: W.A. Benjamin.

Davidson, D. and Suppes, P. 1956. A finitistic axiomatization of subjective probability and utility. *Econometrica* 24, 264–75.

De Finetti, B. 1937. La prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* 7, 1–68. Trans. in Kyburg and Smokler (1964).

Graham, R.L., Rothschild, B.L. and Spencer, J.H. 1980. *Ramsey Theory*. New York: Wiley.

Harrod, R.F. 1951. *The Life of John Maynard Keynes*. London: Macmillan.

Jeffrey, R.L. 1965. *The Logic of Decision*. New York: McGraw-Hill. 2nd edn, Chicago: University of Chicago Press, 1983.

Jeffreys, H. 1961. *Theory of Probability*, 3rd edn. Oxford: Clarendon Press.

Keynes, J.M. 1921. *A Treatise on Probability*. London: Macmillan.

Keynes, J.M. 1933. *Essays in Biography*. London: Macmillan.

Kyburg, H.E. and Smokler, H.E., eds. 1964. *Studies in Subjective Probability*. New York: Wiley.

Little, I.M.D. 1950. *A Critique of Welfare Economics*. Oxford: Clarendon Press.

Luce, R.D. and Suppes, P. 1965. Preference, utility, and subjective probability. In *Handbook of Mathematical Psychology*, vol. 3, ed. R.D. Luce, R.R. Bush and E. Galanter. New York: Wiley.

Neumann, J.von. 1928. Zur Theorie der Gesellschaftspiele. *Mathematische Annalen* 100, 295–320.

Neumann, J.von. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunksatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, 78–83. Trans. as ‘A model of general economic equilibrium’, *Review of Economic Studies* 13, 1945–6, 1–9.

Neumann, J.von. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Neumann, J.von. and Morgenstern, O. 1947. 2nd edn of (1944).

Partridge, F. 1981. *Memories*. London: Victor Gollancz.

Pigou, A.C. 1928. *A Study in Public Finance*. London: Macmillan.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Savage, L.J. et al. 1962. *The Foundations of Statistical Inference: A discussion*. London: Methuen.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Thrall, R.M., Coombs, C.H. and Davis, R.L., eds. 1954. *Decision Processes*. New York: Wiley.

Ulam, S.M. 1976. *Adventures of a Mathematician*. New York: Charles Scribner's Sons.

Vickers, J.M. 1962. A critical investigation of Frank Ramsey's theory of value and belief. Ph.D. thesis, Department of Philosophy, Stanford University.

Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus* , with an Introduction by Bertrand Russell. London: Kegan Paul, Trench, Trubner & Co.

Wittgenstein, L. 1953. *Philosophical Investigations*. Trans. G.E. Anscombe, Oxford: Basil Blackwell.

Wittgenstein, L. 1973. *Letters to C.K. Ogden*, ed. G.H. von Wright. Oxford: Basil Blackwell.

Wright, G.H.von. 1982. *Wittgenstein*. Minneapolis: University of Minnesota Press.

How to cite this article

Newman, Peter. "Ramsey, Frank Plumpton (1903–1930)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000011> doi:10.1057/9780230226203.1382

The New Palgrave Dictionary of Economics Online

rational behaviour

Amartya Sen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A clear distinction must be drawn between (*a*) the type of behaviour that might be described as *rational*, and (*b*) rational behaviour models that might be useful in making predictions about *actual* behaviour. Neither of the two standard views of rational behaviour – as ‘consistent choice’ or as ‘self-interest maximization’ – has emerged as an adequate representation of rationality or of actuality. The difficulties that these views encounter carry over to rational behaviour models accommodating uncertainty.

Keywords

adaptive expectations; binariness; bounded rationality; cardinal utility; choice; completeness; enforced maximization; Exchange; expectation formation; expected utility; externalities; instrumental rationality; natural selection; Pareto efficiency; preferences; Prisoner's Dilemma; probability; rational behaviour; rational expectations; revealed preference theory; risk; satisficing; self-interest; Sen, A. K.; social norms; strong independence; substantive rationality; sure thing principle; transitivity; uncertainty; utilitarianism

Article

The concept of rational behaviour is frequently used in economic theory. The interest in this concept springs from two quite distinct motivations. First, in so far as economic exercises often take a prescriptive form, it is interesting to know how one could behave rationally in a given situation. This may be called the ‘prescriptive motivation’. It should be warned that the prescription need not be necessarily of an ethical kind. Indeed, the prescriptive motivation is sometimes described in clearly non-ethical terms, involving the pursuit of self-interest only. In a classic presentation of this position, Harsanyi (1977) describes ‘perfectly rational behaviour’ in the context of game theory in the following terms:

> … our theory is a *normative* (prescriptive) theory rather than a *positive* (descriptive)

> theory. At least formally and explicitly it deals with the question of how each player *should* act in order to promote his own interests most effectively in the game and not with the question of how he (or persons like him) *will* actually act in a game of this particular type. (Harsanyi, 1977, p. 16)

The second motivation concerns the possible use of models of rational behaviour in explaining and predicting *actual* behaviour. This exercise is done, as it were, in two steps. The first step consists in characterizing rational behaviour and the second, following that, bases actual behaviour on rational behaviour. In this way the characterization of rational behaviour may end up specifying the predicted actual behaviour as well. This motivation underlies much of the theory of general equilibrium (see, for example, Edgeworth, 1881; Arrow, 1951; Debreu, 1959; Arrow and Hahn, 1971). The argument is that while actual behaviour can, in principle, take any form, it is reasonable to assume that much of the time it will, in fact, be of the kind that can be described as ‘rational’.

In reviewing the theory of rational behaviour, this duality of motivations has to be borne in mind. Even though the primary concern of this essay is with the way rational behaviour has been characterized, the nature of the second motivation makes it imperative that the possible use of rational behaviour models for explaining and predicting actual behaviour must not be overlooked.

Rationalizability, binariness and self-interest

In the presence of uncertainty, rational behaviour requires an appreciation of possible variations in the outcome of any chosen action, and such behaviour must, therefore, be based on systematic reading of uncertainties regarding the outcome and ways of dealing with them. Rational behaviour under uncertainty will be presently taken up, but before that the more elementary case when there is no uncertainty has to be dealt with. In fact, behaviour under certainty can be formally seen as an extreme case of behaviour under uncertainty when the uncertainty in question is not only small but simply absent. In this sense, rational behaviour under certainty must be subsumed by any theory that deals with rational behaviour in the presence of uncertainty.

Although there are many different approaches to rational behaviour under certainty, it is fair to say that there are two *main* approaches to this question. The first emphasizes *internal consistency*: rationality of behaviour is identified with a requirement that choices from different subsets should correspond to each other in a cogent and systematic way. Various conditions of internal consistency have been proposed in the literature, but the one which seems to command most attention in formal economic theory is *binariness*, which requires that the choices from different subsets can be seen as maximizing solutions from the respective subsets according to some binary relation R (often interpreted as ‘preference’, for example, xRy standing for ‘x being preferred or indifferent to y’). Or, to put it another way, rational behaviour, in this interpretation, amounts to our ability to find a binary relation R over the universal set of alternatives such that the choice from any particular subset of that universal set consists of exactly the R-maximal elements of that subset. Richter (1971) calls this ‘rationalizability’.

In other formulations – still within the general approach of internal consistency – the condition of rationalizability has been relaxed, demanding only a part of the kind of consistency that binary maximization must entail. On the other hand, in some other formulations, the demands have been made

stronger than that of maximization according to a binary relation by requiring further that the binary relation in question be an ordering, satisfying both completeness and transitivity.

An enormous variety of conditions of internal consistency have been proposed in the literature, but it can be shown that many of them are equivalent to each other, and indeed altogether they fall into a number of classes, with each class containing different, but essentially equivalent, demands. Such reductionist analyses can be found, for example, in Houthakker (1956), Uzawa (1956), Arrow (1959), Richter (1971), Sen (1971), Herzberger (1973), Suzumura (1983). For critiques (and arguments for the rejection of) the binary approach to rationality, see Kanger (1976), Gauthier (1985), Sen (1985a; 1986b) and Sugden (1985).

The second common approach to rational behaviour under certainty sees it in terms of reasoned pursuit of self-interest. The origins of this approach are often traced to Adam Smith, and it is frequently asserted that the father of modern economics saw human beings as tirelessly fostering their respective self-interests. As a piece of history of economic thought, this is, to say the least, dubious, since Adam Smith's (1776; 1790) belief in the hold of self-interest in some spheres of activity (for example, exchange) was qualified by his conviction that many other motivations are important in human behaviour in general (on this see Winch, 1978; Brennan and Lomasky, 1985; and Sen, 1987). But it is certainly true that the assumption of the ‘economic man’ relentlessly pursuing self-interest in a fairly narrowly defined form has played a major part in the characterization of individual behaviour in economics for a very long time.

Self-interest and consistency

Rational behaviour in the form of maximization in pursuit of self-interest makes the analysis of individual behaviour a good deal more tractable than a less structured assumption would permit. This is certainly one of its appeals. In addition this behavioural assumption is also quite crucial for the derivation of certain central results in traditional and modern economic theory, for example, Pareto optimality of competitive equilibria and vice versa (Arrow, 1951; Debreu, 1959; Arrow and Hahn, 1971). This is sometimes called the ‘Fundamental Theorem of Welfare Economics’. Roughly stated, it claims, first, that every perfectly competitive equilibrium (with each person maximizing utility, given the prices) under certain assumptions (such as no externalities) achieves Pareto optimality, and second, under a slightly different set of assumptions (including the requirement of no externalities, but also some additional requirements, such as the absence of increasing returns to scale), every Pareto optimal state is a perfectly competitive equilibrium with respect to some set of prices and some initial distribution of resources. This correspondence between Pareto optimality and competitive equilibria works neatly given individual self-interested behaviour precisely because Pareto optimality is one characteristic of self-interest maximization of a group, in the sense that in such a situation no one's self-interest can be further enhanced without hurting the self-interest of somebody else. It is the assumption of rational behaviour in the form of the pursuit of self-interest that established the close relationship between competitive equilibria and Pareto optimality (with price-taking behaviour and absence of externalities preventing people from getting in each other's way in their respective pursuit of self-interest). In this result and in many other similar ones, the particular characterization of rational behaviour chosen plays a strategically crucial role.

It can be argued that rational behaviour under the self-interest approach is a special case of that under the consistency approach. If a person does pursue self-interest, it may follow that his or her behaviour will have the consistency needed for maximization of a cogent function. On the other hand, a person can be consistent without necessarily maximizing self-interest, since the maximizing function may have a different interpretation altogether (for example, the pursuit of some moral values or political goals). Thus internal consistency of choice may be taken to be necessary but not sufficient for self-interested behaviour. There is undoubtedly something in this way of seeing the correspondence between the two common approaches to rational behaviour.

However, that alleged correspondence is also somewhat misleading, since the nature of self-interest need not necessarily take the uncomplicated form of being binary in character. Strictly speaking, neither does the self-interest thesis entail the consistency thesis, nor of course the other way round. While this must, in general, be correct, nevertheless the way self-interest has been actually viewed in standard economic theory has made it clearly binary and more typically an ordering (and often seen as being numerically representable). If self-interest must take this form, then it would indeed be the case that the self-interest approach is just a special case of the consistency approach.

In some treatises on rational behaviour, the distance between the self-interest approach and the consistency approach is bridged by some careful definitions. For example, in the ‘revealed preference theory’, pioneered by Samuelson (1938), consistency is demanded in the form of the ‘Weak Axiom of Revealed Preference’, to wit: if x is chosen from a set containing y, then y will not be chosen from any set containing x. This type of consistency is, on its own, without a particular substantive interpretation, except that it corresponds generally to some kind of maximization. However, the term ‘revealed preference’ might indicate that the chosen alternative is always also the preferred one. In so far as preference reflects self-interest (as is typically assumed to be the case), this established, through the terminology of ‘revealed preference’, what looks like a congruence of choice and self-interest.

The consistency entailed by the Weak Axiom of Revealed Preference does not, in general, entail transitivity, which is a property that might be thought to be a natural one to impose on the relation of self-interest. But that hole can be plugged *either* by demanding stronger conditions (such as Houthakker's, 1950, ‘Strong Axiom of Revealed Preference’) *or* by demanding that the consistency of the Weak Axiom be satisfied over all finite subsets, which makes the strong axiom equivalent to the weak (on this see Arrow, 1959; Sen, 1971). One way or another, the consistency imposed by revealed preference axioms can lead to a ‘preference’ relation that has the regularity properties normally associated with the concept of self-interest, and then the gap between the two could be seen as fully bridged.

However, that entire bridging exercise is based on *defining* the relation of choice as a relation of ‘preference’ which happens to be ‘revealed’ by the act of choice. But that terminology is arbitrarily imposed, and it is possible that the binary relation of choice, even when fully transitive and complete, may in fact reflect neither the person's preference, nor his or her self-interest. There is, obviously, scope for methodological arguments on this point, and these issues have often been joined.

In the philosophical literature, it is common to distinguish between ‘instrumental rationality’ and ‘substantive rationality’ (see Latsis, 1976). It is clear that the self-interest view of rational behaviour is one of substantive rationality requiring that rational behaviour must take the form of pursuing some independently defined self-interest. Obviously, this characteristic of substantiveness is not satisfied by the theory of revealed preference, since there the identification of choice with preference or self-interest

takes the form of *defining* the relation of choice as a relation of preference, which is not an independent way of characterizing preference or self-interest. But in other theories, the substantive exercise is carefully done, for example, in the typical general equilibrium theory (see Arrow, 1951; Debreu, 1959; Arrow and Hahn, 1971). The starting point of individual behaviour is, then, not a choice function but a utility function, representing the self-interest of the person in question. Choices follow from constrained maximization of that utility function. In this form, the substantive nature of the characterized rationality is strongly asserted, in the shape of pursuit of self-interest.

A number of criticisms have been recently made about the special nature of the assumption of self-interest maximization. Human beings may well have other motivations, and self-interest is just one of various things that a person might wish to pursue. Different types of criticisms of this substantive assumption have been made by such authors as Nagel (1970), Kornai (1971), Sen (1973; 1977; 1987), Scitovsky (1976), Leibenstein (1976), Schelling (1978), Wong (1978). Elster (1979; 1983), Hirschman (1982; 1983), McPherson (1982), Margolis (1982), Akerlof (1984), Schick (1984) and others.

If the assumption of self-interest maximization is seen as too narrow, it can be argued that merely requiring internal consistency is much too permissive. Indeed, it is tempting to think of the consistency approach as belonging to the ‘instrumental’ view of rationality. But this is not quite so, since the instrumental view requires that the person pursues some independently defined objective (even though the objective need not be based on self-interest only). In the consistency view there is no such independently defined function at all, and the binary relation that is precipitated by the choice function is a *reflection* of choice rather than a *determinant* of it. It is rather that the consistency approach opens the way to some instrumental view of rationality, involving the maximization of some objective function. Indeed, in this sense, the consistency approach can be seen as permissively admitting the approach of instrumental rationality implicit in the self-interest approach, where the objective function maximized happens to be the self-interest of the person in question.

The consistency approach can be criticized on grounds of inadequacy in characterizing rationality of behaviour. A person's choice function may be internally consistent in the sense that the different things chosen from different subsets correspond to each other in an apparently cogent and coherent way, but this does not in itself indicate that the person's behaviour is consistent with his or her aims or objectives. Indeed, a person who systematically does exactly the *opposite* of what has to be done for the pursuit of his or her objective function may end up producing a consistent choice behaviour, but the binary relation that will be revealed by the choices – the ‘opposite’ of the person's objective function – will be, clearly, at war with the goals and aims of that person. To describe such a person as behaving rationally would, obviously, lead to some interesting methodological difficulties.

Maximizing, satisficing and bounded rationality

These problems with the standard views of rationality tend to undermine the very foundations of these approaches. Some other approaches have involved more qualified use of the standard presumptions. For example, Herbert Simon (1957; 1979) has argued powerfully that individuals may not actually *maximize* any function at all, and their behaviour may take the form of what has been called ‘satisficing’. There are various ways of characterizing satisficing, but it can be thought of in terms of a person having a certain target level of achievement, which he or she will try to reach, but beyond which he or she may

not try to improve the achievement any further.

There is a genuine problem of interpretation involved in analysing satisficing, and it can be argued that satisficing behaviour really is maximization according to an effectively incomplete relation, such that the states satisfying the target level of achievement are all put in a non-comparable class as far as choice behaviour is concerned. Maximization can indeed be defined in terms of such incomplete relations (see, for example, Debreu's 1959 analysis of ‘maximal’ sets based on ‘pre-orderings’), and, if it is seen in these terms, the gap between satisficing and maximizing may be, at least formally, reduced. However, the content of the claim of satisficing is that the person in question *can* tell between the different levels of achievement which are all beyond the target level required, and, despite this discernibility, choice behaviour departs from relentless maximization of the level of achievement. In this version of the story, a substantial difference is indeed made by the notion of satisficing, and the implications of satisficing behaviour may, in this interpretation, be quite different from those of maximization.

Variations of the maximization assumption and the related consistency conditions can be justified by seeing the use of reason in human affairs in terms of what has been called ‘bounded rationality’. In this structure human choice is seen not in terms of grand maximizing behaviour, but as a series of particular decisions, not fully integrated with each other, taken in situations of partial information and based on limited reflection. This approach has been developed by Herbert Simon (1957; 1979; 1983) both at a theoretical level and in the context of specific empirical applications. The results differ quite substantially from that of rational behaviour seen in terms of consistency, or in terms of optimization according to self-interest. As Simon puts it:

> Rationality of the sort described by the behavioural model [of bounded rationality] doesn't optimize, of course. Nor does it even guarantee that our decisions will be consistent. As a matter of fact, it is very easy to show that choices made by an organism having these characteristics will often depend on the order in which alternatives are presented. (Simon, 1983, p. 23)

Natural selection and motives

Supporters of optimizing models have typically used two different types of arguments to defend the practice, against models of the kind characterized by ‘bounded rationality’ and other behavioural departures. One argument takes the direct form of arguing that human beings do optimize and take care to do so. The second argument suggests that natural selection will lead in this result: those who optimize do better, and those who do not get eliminated by natural selection. For example, non-profit-maximizing firms may go to the wall, so that only the profit-maximizing ones may survive (see Friedman, 1953).

This type of indirect justification of what has been called ‘enforced maximization’ has many pitfalls, since the analogy with natural selection in biology is at best tenuous (see Helm, 1984; Matthews, 1984), and the biological story itself is far from straightforward (Dawkins, 1982; Maynard Smith, 1982).

It is by no means clear that individual self-interest-maximizers will typically do relatively better in a group of people with diverse motivations. More importantly, when it comes to comparisons of survival of different *groups*, it can easily be the case that groups that emphasize values other than pure self-interest maximization might actually do better (see Sen, 1973; 1974; 1985b; Akerlof, 1984). It has been

argued that economic success has often come more plentifully in cultures that emphasize norms of conduct quite different from that of persistent maximization of individual self-interest, focusing on other values (for example, what Morishima, 1982, calls ‘the Japanese ethos’; see also Dore, 1983). The relation between social norms and individual conduct is an enormously complex field, and the simple assumptions of self-interest maximization, of straightforward models of apparent ‘consistency’, may overlook important aspects of the individual–society relationships (see, for example, Hirschman, 1970; 1982). This is not to argue that ‘natural selection’ arguments are worthless in economics – they may be far from that – but the results of the selection may lack the simplicity demanded by supporters of simple optimization and may take a more complex form (see Hirshleifer, 1977; Helm, 1984; Matthews, 1984).

In assessing the overall value of standard models of rational behaviour, it is important to pay attention to the distinction made earlier between the value of these structures as representations of *rationality* and their usefulness in terms of predicting *actual* behaviour. Some of the deficiencies of the optimizing ‘structure apply specifically to the latter. For example, models of ‘bounded rationality’ are often defended by claims of greater plausibility in explaining actual human conduct.

In fact, the entire enterprise of getting to actual behaviour via models of rationality may itself be seen us methodologically quite dubious. There is scope for argument here on both sides, since the unrealism of rational behaviour may be large, but the unrealism of any *specific kind* of ‘irrational’ behaviour could be larger still. Whether ‘bounded rationality’ is the right kind of compromise in getting a grip on actuality via limited use of rationality remains an interesting question.

Reason and rationality

As far as the other objective of rational behaviour models is concerned, that is, the ability of these models to capture the essence of rationality (no matter how people do actually behave), there are a number of complex philosophical issues underlying the question. It is easy enough to argue that mere internal consistency of choice cannot be adequate for rationality, nor can self-interest maximization be seen as uniquely rational in a way that pursuing other kinds of objectives (such as altruism, public spirit, class consciousness, group solidarity) must fail to be. What is much harder to do is to develop an alternative structure for rationality that would be regarded as satisfactory for the purpose of capturing what can be demanded of reason in human choice (whether or not it also serves the second purpose of giving us a good guess regarding actual behaviour). This question remains, to a great extent, an open one, which has been as yet rather inadequately explored.

Two difficulties, in particular, may be worth mentioning in this context. First, while ‘instrumental rationality’ must have some place in economics, and the role of reasoned choice of means for serving *given* ends cannot be dismissed, it is hard to believe that any kind of objectives no matter how bizarre – must be seen as okay, that is, not compromising the rationality of the person pursuing it. The need for rational assessment of objectives and preferences have been analysed by John Broome (1978), Derek Parfit (1984) and others, and both the procedural and substantive features of this type of assessment do deserve serious attention.

Second, even when goals are clearly given, the translation of these into actions depends on the pattern of social interdependence assumed in group behaviour, with members having partly divergent goals. As the discussions on the so-called ‘Newcomb's problem’ and other complex cases have brought out, the

correct individual decision may not be entirely unproblematic even when there appears to exist a strictly dominant strategy (see Nozick, 1969; Brams, 1975; Levi, 1975, Gibbard and Harper, 1978; Jeffrey, 1983, among others). The nature of beliefs permits alternative interpretations of the nature of the decision problem, and this philosophical question is of relevance to decision problems in economics as much as it is in other fields of human choice.

The Prisoner's Dilemma has been frequently used in economic arguments to illustrate the nature of inefficiencies of atomistic non-cooperative behaviour when the interdependence incorporates both congruence and conflict of interests in such a way that the combination of each person's dominant strategies produces an outcome that is inferior in terms of the goals of everyone in the group (see Luce and Raiffa, 1957). Attempts to resolve the problem by assuming temporal repetition of the game have not been easy, since it can be demonstrated that with complete knowledge and standard optimizing behaviour, a finitely repeated Prisoner's Dilemma will continue to produce the inferior outcome throughout (Luce and Raiffa, 1957, pp. 97–101).

Such non-cooperative behaviour is, however, violated in many experimental games as well as in the usual readings of many real-life situations. The apparent dissonace between received theory and observed behaviour has been explained in a variety of ways in the large literature that has developed on the Prisoner's Dilemma. The ‘ways out’ have included relaxing the assumption of mutual knowledge, for example, introducing uncertainty about the number of times for which the game will be played, admitting ignorance of the players about other people's knowledge and motivation, limiting the range of alternative strategies that can be considered, and other relaxations (see Howard, 1971; Basu, 1977, Davis, 1977; Radner, 1980; Smale, 1980; Kreps et al., 1982; Axelrod, 1984). Other analyses have emphasized more complex features of ‘practical reasoning’ involving various types of action ethics, sensitive beliefs, behavioural commitments, and instrumental use of reciprocity; see Sen (1974; 1985b), Watkins (1974; 1985). Levi (1975), Gauthier (1985), McClennen (1985). If it has done nothing else, the literature has at least brought out sharply the complexity of the nature of rationality in situations of interdependence as well as various conceptual and logistic difficulties in using models of rationality to understand the nature of actual behaviour.

It stems easy to accept that rationality involves many features that cannot be summarized in terms of some straightforward formula, such as binary consistency. But this recognition does not immediately lead to alternative characterizations that might be regarded as satisfactory, even though the inadequacies of the traditional assumptions of rational behaviour standardly used in economic theory have become hard to deny. It will not be an easy task to find replacements for the standard assumptions of rational behaviour – and related to it of actual behaviour – that can be found in the traditional economic literature, both because the identified deficiencies have been seen as calling for rather divergent remedies, and also because there is little hope of finding an alternative assumption structure that will be as simple and usable as the traditional assumptions of self-interest maximization, or of consistency of choice.

Uncertainty and expected utility

The extension of the modelling of rational behaviour from certainly to uncertainty involves both (*a*) the characterization of uncertainty, and (*b*) taking note of uncertainty thus characterized in making actual

decisions over alternative courses of actions. The model that has been most extensively used in this context is that of ‘expected utility’. This takes the form of weighing the value of each of the outcomes by the respective probabilities of the different outcomes. The probability-weighted overall ‘expected value’, thus derived, is then maximized in this approach to rational choice under uncertainty.

The use of probability calculus involves interpretational problems as to what the probabilities stand for. While the view of probability as a measure of relative frequency is a natural one to consider, there is clearly much cogency in interpreting probability as a measure of the degree of belief (as argued by Fisher, 1921, and Keynes, 1921).

Actual decision-taking operations involve a reading of the likelihood of different outcomes and an assessment of the different outcomes in the light of the respective likelihoods. In a pioneering contribution in axiomatizing conjointly characterized probabilities and utilities, Frank Ramsey (1931) provided the structure (and a possible derivation) of the expected utility calculus. Another major contribution in this area came from von Neumann and Morgenstern (1947). Given the probabilities of different outcomes, consistent and complete rankings of the possible lotteries over the outcomes (including lotteries of lotteries and so forth) permit the construction of cardinal utility functions for the respective rankings associated with the outcomes, provided the rankings in question satisfy certain regularity properties which were specified by von Neumann and Morgenstern (see also Marschak, 1946). The assigned cardinal utility numbers of the respective outcomes, weighted by the respective probabilities, when summed together, yield the expected values of the lotteries, and provide numerical representations of the overall goodness of the respective lotteries. Rational behaviour under expected utility maximization takes the form of choosing that lottery which has the highest overall value, thus calculated. The expected utility approach can be and has been used extensively both in economic theory and in applied economics (see, for example, Friedman and Savage, 1948; Arrow, 1971).

Independence and consistency

The axioms underlying the derivation of expected utility maximization have been subjected to a good deal of examination and scrutiny. There is scope for disputation about both the exact content and the plausibility of the expected utility axioms (for a very helpful introduction see Luce and Raiffa, 1957; see also Fishburn, 1970; 1981).

The axiom that has perhaps attracted the most criticism is the so-called ‘strong independence’. This independence condition can be stated in several different ways, but a rather immediate one is the following. If in a combined lottery over, say, lotteries L^1 and L^2, the latter L^2 is replaced by another lottery L^3 which is preferred to L^2 (leaving the probabilities and L^1 unchanged), then the modified combined lottery (over L^1 and L^3) would be preferred to the original one (over L^1 and L^2). And vice versa.

Another axiom, related to this one, is sometimes called ‘the sure thing principle’, which, in one version, requires that anything that raises the probability of the preferred component in a two-alternative lottery would improve the lottery. These axioms are implicit in expected utility maximization, even though the ‘independence’ condition can be dispensed with in a more limited (‘locally’ valid) version of expected utility behaviour (as has been shown by Mark Machina, 1982).

Various ‘counter-examples’ to expected utility maximization have been proposed in the literature, often

on the basis of considering interesting ‘hypothetical’ cases, but sometimes on the basis of experimental observations as well. In assessing these objections, we must distinguish, once again, between the claims *to* rationality of this model, and the claims of the model to explain actual behaviour *via* rationality.

It is certainly clear that very often people do act in a way that cannot be made consistent with expected utility maximization. (An early critique, with an alternative framework for choice behaviour, came from Shackle, 1938; 1952.) Observations of behaviour and articulated judgements under uncertainty have indicated different types of violations of expected utility behaviour (see, for example, Kahneman, Slovik and Tversky, 1982). There seem to be problems both in risk perception as well as in the utilization of probability information in making actual decisions. These departures from rational behaviour in the form of expected utility maximization have considerable implications on the way economic models may have to be constructed involving uncertainty (on this see Arrow, 1982; 1983). As a framework for understanding actual behaviour, the merits and demerits of the expected utility model are certainly becoming clearer on the basis of recent work. But the ‘bottom line’ of overall judgement continues to vary. While some have been extremely sceptical, others (such as Harsanyi) continue to emphasize, with some justice, the usefulness of this model in ‘explaining or predicting real-life human behaviour’ (1977, p. 16).

The need for departures – small or great – from the expected utility model in explaining *actual* behaviour does not, of course, settle the question of the rationality or irrationality of maximization of expected utility. However, a number of telling and powerful arguments have also been presented in the literature giving reasons for departing from ‘consistency’ of the kind demanded by the expected utility model (for arguments on both sides, see the collection of papers in Daboni, Montesano and Lines, 1986). Allais (1953) has followed up his empirical critique of expected utility model as representation of actual behaviour by arguments in favour of the reasonableness of the departures, and more arguments on this have been outlined in recent years (see Allais and Hugen, 1979; Stigum and Wenstop, 1983; and Daboni, Montesano and Lines, 1986). Also, the possibility of ‘state-dependent utilities’ has raised questions of a different sort, requiring reformulation of the original model (see Drèze, 1974).

One of the important considerations that the expected utility model may leave out consists of ‘counterfactual’ information. One's ‘disappointment’, ‘regret’, and so on may well depend on what one anticipated and what did not occur. Earlier discussions of such criteria as ‘minimax regret’ (see Savage, 1954) have been followed in recent years by various models of disappointment and regret (see, for example, Bell, 1982; Loomes and Sugden, 1982).

It is arguable that something which has not happened, but could have, should not really affect one's decision, and in particular, it is irrational to regret and sigh about what could have happened. But while it is indeed possible to argue that it is irrational to regret a past decision on the ground of what could have happened in the light of later information, nevertheless, if it is the case that one would willy-nilly regret the past decision if it turns out to he unfortunate, then it is not in any sense obviously irrational to *recognize* that fact and take that inescapable feeling into account. Clarity of analysis requires that we distinguish between (*a*) the rationality of what psychology we ought to have, and (*b*) the rationality of decisions, taking note of what psychology we might not be able to escape. Many counter-examples to expected utility behaviour presented in the literature relate – directly or indirectly – to mental-state considerations, for example, Allais (1953), MacCrimmon (1968), Bernard (1974), Drèze (1974),

Tversky (1975), Machina (1981), McClennen (1983) and others.

One reason why the inclusion of mental states among the influences on choice is resisted is the idea that mental state is a particular interpretation of *utility* of which another – alternative – interpretation is given by the numerical representation of choice, with which the expected utility model is concerned. In the context of utilitarianism, the mental-state utility and the numerical representation of choice can indeed be seen as *alternatives*, as they have been viewed in the ethical literature. However, in terms of the description of the world, both mental states and choices are distinct parts of the reality, and the acknowledgement of the existence of one does not deny the existence of the other. Indeed, it is not unreasonable to ask how each might relate to the other. The states of affairs over which choices may be considered (including choices over lotteries of those states) may, quite importantly, include the mental states of the parties involved.

On the other hand, including such mental states in the description of states of affairs makes the scope of such conditions as ‘strong independence’ rather limited. Varying an *alternative* lottery (for example, L^3 vis-à-vis L^2) might affect the description of the ‘prize’ of a given lottery (L^1) through variations of mental states (now included in the outcome of L^1) related to considering and reflecting on the nature of the alternative (L^1 vis-à-vis L^2) and the corresponding disappointment, regret, and so on. If L^1 is no longer ‘the same’ in the two cases, then ‘strong independence’ would make no demand. Thus ‘strong independence’ may be saved only at the cost of making it often trivially fulfilled (see Sen, 1985a). The same difficulty applies if strong independence is ‘rescued’ by including counterfactual information in describing states of affairs.

The basis of rationality implicit in expected utility calculation does, however, require descriptions of states of affairs in sufficient detail such that choices can be made taking all the relevant considerations into account. It can be argued, as indeed Peter Hammond (1986) has, that ‘consequential’ reasoning, taking into account all the relevant considerations, will push us in the direction of expected utility maximization. The important question is whether the relevant considerations would include either counterfactuals or mental states, and, if they do so, whether enough scope for the use of such conditions as ‘strong independence’ can be found to build up utility numbering in a way that would make the expected utility model work in practice. This is not a matter, obviously, of pure theory only, and much depends on the nature of people's psychology and what considerations might be regarded as rational, in taking note of the complexities of our psychology.

Concluding remarks

Attempts at constructing models of rational behaviour have certainly played a creative part in reducing the intractability of unstructured assessment of (*a*) the demands of rationality, and (*b*) facts of actual behaviour. On the other hand, models of rational behaviour actually presented have tended to ignore some of the complexities that have to be faced. This problem arises even when no uncertainty is introduced into the picture.

Neither of the two standard views of rational behaviour – as ‘consistent choice’ or as ‘self-interest maximization’ – has emerged as being really adequate as representations of rationality or of actuality. Various suggestions as to the directions in which we might go were reviewed earlier. Although none of the suggestions are unproblematic, many fruitful avenues of investigation have certainly been identified

in the critical literature.

These difficulties carry over to rational behaviour models accommodating uncertainty. The limitations of characterizing rational behaviour in terms of just internal consistency, as discussed in the context of choice under certainty, obviously would apply to the modelling of choice under uncertainty as well. Similarly, pursuit of self-interest cannot be seen us being *uniquely* rational in models of uncertainty, any more than they can be so seen when everything is certain. However, it is not really necessary that expected utility models be seen in terms of self-interest maximization, and indeed some writers, for example, Ramsey (1931), have explicitly repudiated that interpretation. In fact, what the expected utility models do concentrate on is ‘consistency’ in a very demanding sense, and in this context objections similar to the ones raised in models of choice *without* uncertainty can be raised a fortiori with uncertainty.

Rationality may be seen as demanding something other than just consistency of choices from different subsets. It must, at least, demand cogent relations between aims and objectives actually entertained by the person and the choices that the person makes. This problem is not eliminated by the terminological procedure of describing the cardinal representation of choices as the ‘utility’ of the person, since this does not give any independent evidence on what the person is aiming to do or trying to achieve.

A more difficult issue, as discussed in the context of certainty, concerns the *assessment* of aims and objectives pursued by a person, even if they are fully reflected in the choices actually made. As Patrick Suppes has put it, the standard normative model of expected utility ‘can be satisfied by cognitive and moral idiots … Put another way, the consistency of computations required by the expected-utility model does not guarantee the exercise of judgement and wisdom in the traditional sense’ (1984, pp. 207–8). Suppes argues in favour of moving to the Aristotelian view that the rational person acts ‘in accordance with good reasons’, and is not embarrassed by the fact that this leaves a certain amount of ‘pluralism’ in the possible approach to rationality.

In addition to those problems of rationality that are shared by models of certainty as well as uncertainty, there are some special problems that apply particularly to considerations of uncertain outcomes. The status of counterfactuals, and their influences on mental states, raise interesting and important questions as to what may or may not be relevant to take into account in rationally assessing alternative courses action.

While these problems were addressed earlier on in this paper, one issue that has not yet received much attention here concerns the nature of uncertainty itself. Reference was made earlier to the distinction between interpreting probabilities as degrees of belief, and interpreting them as frequencies. There are also other issues (see, for example, Levi, 1982; 1987). Even the very idea of having beliefs about possible outcomes in the form of probabilities in a situation of partial ignorance raises some interesting philosophical questions. At the very least, it is possible to make a distinction that was made by Frank Knight (1921) between ‘risk’ and ‘uncertainty’, with probability distributions being specified in the case of the former but not in the latter case. Whether arguments such as ‘insufficient reason’ can permit one to *construct* probability distributions even when we do not start with them remains a hard question to settle.

The area of expectation formation is also one in which the demands of rationality are not easy to specify. In some models of rational behaviour, no requirements of rationality are imposed on expectations at all, and the problem of rationality arises only in taking note of the actual expectations in arriving at decisions regarding action. In models of ‘adaptive expectations’ a step is taken in the direction of

making expectations responsive – in an intelligent way – on experience. What goes very much further than this is the assumption of ‘rational expectation’ by which each person anticipates what can, in some sense, be described as objective probabilities; see Muth (1961) and Lucas and Sargent (1982).

This approach not only raises the question as to what the philosophical status of objective probabilities might be, but also whether it is really a matter of *rationality* as such whether one is successful in guessing what the objective probabilities are. It is fair to say that the assessment of models of ‘rational expectation’ cannot be bused on the idea of rationality alone, since the demands of such a theory go well beyond the requirements of the use of reason, especially in a situation of ignorance. It is sensible enough to think that there are problems in models of behaviour in which people's expectations are systematically wrong, but to try to move from that recognition to one in which everyone manages to take note of objective probabilities fully is quite a dramatic step. Whether that step is worth taking in predicting actual behaviour might well be discussed and assessed in the light of the ability of such a theory to explain actual behaviour, but that, as we have already discussed, is a rather different problem from assessing the *rationality* us such of that behaviour.

In addition to the issue of the role of rationality involved in ‘rational expectation’ models, even the basic rational behaviour model (without such expectational assumptions), widely used in economics, raises, as we have seen, difficult – sometimes perplexing – questions. It is not hard to see the merit of trying to reduce a complex reality by characterizing rationality in rather narrow terms, but nor is it hard to fathom that such a narrowing might do grave injustice to the notion of rationality, which is, after all, one of the central concerns of human life.

We have to make a clear distinction between (*a*) what type of behaviour might be described as *rational*, and (*b*) what rational behaviour models might be useful in making predictions about *actual* behaviour. These different questions are not, of course, independent of each other. But the first step in pursuing their interrelations is to recognize the distinction between the two questions. What issues respectively arise in facing these distinct questions, and how they might possibly be related, were discussed earlier on in this article in the light of the existing literature. There was, however, no escape from noting the fact that the existing literature is indeed deeply incomplete in that real difficulties have been identified without providing an adequate structure for solutions. The need to go beyond the existing literature is apparent enough, but where to go is less clear.

See Also

- philosophy and economics
- social choice
- welfare economics

Bibliography

Akerlof, G.A. 1984. *An Economic Theorist's Book of Tales*. Cambridge: Cambridge University Press.

Allais, M. 1953. Le comportment de l'homme rational deviant to risque: critique de postulates et axiomes de l’école Américaine. *Econometrica* 21, 503–46.

Allais, M. and Hugen, O., eds. 1979. *Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of Decisions under Uncertainty with Allais' Rejoinder*. Dordrecht: Reidel.

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the Second Berkeley Symposium of Mathematical Statistics*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J. 1959. Rational choice functions and orderings. *Economics* 26, 121–7.

Arrow, K.J. 1971. *Essays in the Theory of Risk-Bearing*. Amsterdam: North-Holland.

Arrow, K.J. 1982. Risk perception in psychology and economics. *Economic Inquiry* 20, 1–9.

Arrow, K.J. 1983. Behaviour under uncertainty and its implientions for policy. In Stigum and Wenstop (1983).

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day. Republished, Amsterdam: North-Holland, 1979.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Academic Press.

Basu, K. 1977. Information and strategy in iterated Prisoners' Dilemma. *Theory and Decision* 8, 293–8.

Bell, D.E. 1982. Regret in decision making under uncertainty. *Operations Research* 30, 961–81.

Bernard, G. 1974. On utility functions. *Theory and Decision* 5, 205–42.

Brams, S.J. 1975. *Game Theory and Politics*. New York: Free Press.

Brennan, G. and Lomasky, L. 1985. The impartial spectator goes to Washington: toward a Smithian theory of economic behavior. *Economics and Philosophy* 1, 189–211.

Broome, J. 1978. Choice and value in economics. *Oxford Economic Papers* 30, 313–33.

Broome, J. 1984. Uncertainty and fairness. *Economic Journal* 94, 624–32.

Campbell, R. and Sowden, L., eds. 1985. *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press.

Chipman, J.S., Hurwicz, L., Richter, M.K. and Sonnenschein, H.F., eds. 1971. *Preferences, Utility and*

Demand. New York: Harcourt, Brace.

Daboni, L., Montesano, A. and Lines, M., eds. 1986. *Recent Developments in the Foundations of Utility Theory and Risk*. Dordrecht: Reidel.

Davidson, D., Suppes, P. and Siegel, S. 1957. *Decision Making: An Experimental Approach*. Stanford: Stanford University Press.

Davis, L.M. 1977. Prisoners, paradox and rationality. *American Philosophical Quarterly* 14, 319–27. Also in Campbell and Sowden (1985).

Dawkins, R. 1982. *The Extended Phenotype*. Oxford: Clarendon Press.

Debreu, G. 1959. *A Theory of Value*. New York: Wiley.

Dore, R. 1983. Goodwill and the spirit of market capitalism. *British Journal of Sociology* 34, 459–82.

Drèze, J.H. 1974. Axiomatic theories of choice, cardinal utility and subjective probability: a review. In *Allocation under Uncertainty: Equilibrium and Optimality*, ed. J.H. Drèze. London: Macmillan.

Edgeworth, F. 1881. *Mathematical Psychics*. London: Kegun Paul.

Elster, J. 1979. *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Elster, J. 1983. *Sour Grapes*. Cambridge: Cambridge University Press.

Fishburn, P.C. 1970. *Utility Theory and Decision Making*. New York: Wiley.

Fishburn, P.C. 1981. Subjective expected utility: a review of normative theories. *Theory and Decision* 31, 139–99.

Fisher, R.A. 1921. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, Series A 222, 309–68.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: Chicago University Press.

Friedman, M. and Savage, L.J. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56, 279–304.

Gauthier, D. 1985. Maximization constrained: the rationality of cooperation. In Campbell and Sowden (1985).

Gibbard, A. and Harper, W.L. 1978. Counterfactual and two kinds of expected utility. In Hooker, Leach and McClennen (1978).

Hammond, P.J. 1976. Changing basics and coherent dynamic choice. *Review of Economic Studies* 43, 159–73.

Hammond, P.J. 1986. Consequentialism and rationality in dynamic choice under uncertainty. In *Social Choice and Public Decision Making: Essays in Honor of K.J. Arrow*, vol. 1, ed. W. Heller, D. Starrell and R. Starr. Cambridge: Cambridge University Press.

Harsanyi, J.C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Helm, D. 1984. Predictions and causes: a comparison of Friedman and Hicks on method. *Oxford Economic Papers* 36(Supplement), 118–34.

Herzberger, H. 1973. Ordinal preference and rational choice. *Econometrica* 41, 187–237.

Hirschman, A.O. 1970. *Exit. Voice, and Loyalty*. Cambridge, MA: Harvard University Press.

Hirschman, A.O. 1982. *Shifting Involvements*. Princeton: Princeton University Press.

Hirschman, A.O. 1983. Against parsimony: three easy ways of complicating some categories of economic discourse. *American Economic Review* 74(May), 89–96.

Hirshleifer, J. 1977. Economics from a biological viewpoint. *Journal of Law and Economics* 20, 1–52.

Hooker, C.A., Leach, J.J. and McClennen, E.F., eds. 1978. *Foundations and Applications of Decision Theory*. Dordrechi: Reidel.

Houthakker, H.S. 1950. Revealed preference and the utility function. *Economica* 15, 159–74.

Houthakker, H.S. 1956. On the logic of preference and choice. In *Contributions to Logic and Methodology in Honor of J.J. Bochensk*, ed. A. Tymieniecke. Amsterdam: North-Holland.

Howard, N. 1971. *Paradoxes of Rationality*. Cambridge, MA: MIT Press.

Jeffrey, R.C. 1965. *The Logic of Decision*. New York: McGraw Hill.

Jeffrey, R.C. 1983. *The Logic of Decision*, 2nd edn. Chicago: University of Chicago Press.

Kahneman, D. and Tversky, A. 1979. Prospect theory; an analysis of decisions under risk. *Econometrica* 47, 263–91.

Kahneman, D., Slovik, P. and Tversky, A. 1982. *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kanger, S. 1976. Preference based on choice. Mimeo, Uppsala University.

Keynes, J.M. 1921. *A Treatise on Probability*. London: Macmillan.

Knight, F. 1921. *Risk, Uncertainty and Profit*. New York: Houghton Mifflin.

Kornai, J. 1971. *Anti-Equilibrium*. Amsterdam: North-Holland.

Kreps, D.M., Milgrom, P., Roberts, J. and Wilson, R. 1982. Rational cooperation in the finitely repeated Prisoner's Dilemma. *Journal of Economic Theory* 27, 245–52.

Latsis, S.J., ed. 1976. *Method and Appraisal in Economics*. Cambridge: Cambridge University Press.

Leibenstein, H. 1976. *Beyond Economic Man: A New Foundation for Microeconomics*. Cambridge, MA: Harvard University Press.

Levi, J. 1975. Newcomb's many problems. *Theory and Decision* 6, 161–75.

Levi, J. 1982. Ignorance, probability and rational choice. *Synthesis* 53, 287–417.

Levi, J. 1987. *Hard Choices*. Cambridge: Cambridge University Press.

Loomes, G. and Sugden, R. 1982. Regret theory: an alternative theory of rational choice. *Economic Journal* 92, 805–24.

Lucas, R.E. and Sargent, T.J. 1982. *Rational Expectation and Econometric Practice*. London: Allen & Unwin.

Luce, R.D. and Raiffa, H. 1957. *Games and Decisions*. New York: Wiley.

MacCrimmon, K.R. 1968. Descriptive and normative implications of decision theory postulates. In *Risk and Uncertainty*, ed. K. Borch and J. Mossin. London: Macmillan.

Machina, M. 1981. ‘Rational’ decision making vs. ‘rational’ decision modeling? *Journal of*

Mathematical Psychology 24, 163–75.

Machina, M. 1982. ‘Expected utility’ analysis without the independence axiom. *Econometrica* 50, 277–323.

McClennen, E.F. 1983. Some-thing doubts. In Stigum and Wenstop (1983).

McClennen, E.F. 1985. Prisoner's dilemma and resolute choice. In Campbell and Sowden (1985).

McPherson, M.S. 1982. Mill's moral theory and the problem of preference change. *Ethics* 92, 252–73.

Margolis, H. 1982. *Selfishness, Altruism and Rationality*. Cambridge: Cambridge University Press.

Marschak, J. 1946. Von Neumann's and Morgenstern's new approach to static economics. *Journal of Political Economy* 54, 91–115.

Matthews, R.C.O. 1984. Darwinism and economic change. *Oxford Economic Papers* 36(Supplement), 91–117.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Morishima, M. 1982. *Why Has Japan ‘Succeeded’? Western Technology and Japanese Ethos*. Cambridge: Cambridge University Press.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Clarendon Press.

Nozick, R. 1969. Newcomb's problem and two principles of choice. In *Essays in Honor of Carl G. Hempel*, ed. N. Rescher. Dordrecht: Reidel.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Radner, R. 1980. Collusive behaviour in non-cooperative epsilon-equilibrium of oligopolies with long but finite lives. *Journal of Economic Theory* 22, 136–54.

Ramsey, P.P. 1931. Truth and probability. In F.P. Ramsey, *The Foundations of Mathematics and other Logical Essays*. London: Kegan Paul.

Richter, M.K. 1971. Rational choice. In Chipman et al. (1971).

Samuelson, P.A. 1938. A note on the pure theory of consumers' behaviour. *Economica* 5, 61–71.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Schelling, T.G. 1978. *Micromotives and Macrobehavior*. New York: Norton.

Schelling, T.C. 1984. Self-command in practice, in policy, and in a theory of rational choice. *American Economic Review* 74, 1–11.

Schick, F. 1984. *Having Reasons: An Essay and Rationality and Sociality*. Princeton: Princeton University Pess.

Scitovsky, T. 1976. *The Joyless Economy*. London: Oxford University Press.

Sen, A.K. 1971. Choice functions and revealed preference. *Review of Economic Studies* 38, 307–17.

Sen, A.K. 1973. Behaviour and the concept of preference. *Economica* 40, 241–59. Repr. in Sen (1982).

Sen, A.K. 1974. Choice ordering and morality. In *Practical Reason*, ed. S. Körner. Oxford: Blackwell. Repr. in Sen (1982).

Sen, A.K. 1977. Rational fools: a critique of the behavioural foundations of economic theory. *Philosophy and Public Affairs* 6, 317–44. Repr. in Sen (1982).

Sen, A.K. 1982. *Choice, Welfare and Measurement*. Oxford: Blackwell; Cambridge, MA: MIT Press.

Sen, A.K. 1985a. Rationality and uncertainty. *Theory and Decision* 18, 109–27. Repr. in Daboni, Montesano and Lines (1986).

Sen, A.K. 1985b. Goals, commitment and identity. *Journal of Law Economics and Organization* 1, 341–55.

Sen, A.K. 1987. *On Ethics and Economics*. Oxford: Blackwell.

Shackle, G.L.S. 1938. *Expectations. Investment and Income*. Cambridge: Cambridge University Press.

Shackle, G.L.S. 1952. *Expectations in Economics*, 2nd edn. Cambridge: Cambridge University Press.

Simon, H.A. 1957. *Models of Man*. New York: Wiley.

Simon, H.A. 1979. *Models of Thought*. New Haven: Yale University Press.

Simon, H.A. 1983. *Reason in Human Affairs*. Oxford: Blackwell.

Smale, S. 1980. The Prisoner's Dilemma and dynamic systems associated to non-cooperative games. *Econometrica* 48, 1617–34.

Smith, A. 1776. In *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Clarendon Press, 1976.

Smith, A. 1790. In *The Theory of Moral Sentiments*, ed. D.D. Raphael and A.L. Macfie. Oxford: Clarendon Press, 1974.

Stigum, B.P. and Wenstop, F., eds. 1983. *Foundations of Utility and Risk Theory with Applications*. Dordecht: Reidel.

Sugden, R. 1985. Why be consistent? A critical analysis of consistency requirements in choice theory. *Economica* 52, 167–83.

Suppes, P. 1984. *Probabilistic Metaphysics*. Oxford: Blackwell.

Suzumura, K. 1983. *Rational Choice, Collective Decisions and Social Welfare*. Cambridge: Cambridge University Press.

Tversky, A. 1975. A critique of expected utility theory: descriptive and normative considerations. *Erkenntnis* 9, 163–73.

Uzawa, H. 1956. A note on preference and axioms of choice. *Annals of the Institute of Statistical Mathematics* 8, 35–40.

von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Watkins, J. 1974. Comment: self-interest and morality. In *Practical Reason*, ed. S. Körner. Oxford: Blackwell.

Watkins, J. 1985. Second thoughts on self-interest and morality. In Campbell and Sowden (1985).

Winch, D. 1978. *Adam Smith's Politics*. Cambridge: Cambridge University Press.

Wong, S. 1978. *Foundations of Paul Samuelson's Revealed Preference Theory*. London: Routledge.

How to cite this article

Sen, Amartya. "rational behaviour." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000022> doi:10.1057/9780230226203.1385

The New Palgrave Dictionary of Economics Online

rational choice and political science

Susanne Lohmann

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Rational choice in political science’ stands for the application of the economics approach in the study of political phenomena. The research program is to *rationalize* collective behaviour that comes across as stupid or counterproductive. In its high-brow (esoteric) variant, rational choice is on the way out in political science. In its low-brow (sensible) variant, rational choice is here to stay, not as the dominant approach, but as one of three equal, and complementary, approaches: the rationalist approach, which focuses on individual agency; the culturalist approach, which centres on collective identities; and the structuralist approach, which emphasizes historical institutionalism.

Keywords

Arrow, K. J.; asymmetric information; behavioural economics; Buchanan, James; Tullock, Gordon; Olson, Mancur; Hayek, F. A.; Aristotle; Hobbes, T.; Coase Theorem; collective action; complex systems approach in social science; complexity; Congressional committees; culturalist approach in political science; historical institutionalism; collective identities; individual agency; institutionalism; agency; public choice; externalities; government intervention; politics of monetary policy; logic of collective action; supply of and demand for collective action; greed, rationality, and equilibrium; Keynesian approach; Hayekian approach; developing countries; social complexity; post-autistic economics movement; political science; deadweight loss; development economics; experimental economics; free-rider problem; game theory; government failure; independence of irrelevant alternatives; market failure; political economy; positive political theory; preference aggregation; rational choice; rational choice in political science; representative agent; self-interest; social choice; social welfare function; special interests; structuralist approach in political science

Article

Chemical plants are vulnerable to terrorist attacks. Two of the most dangerous facilities are located in Dallas, in the US state of Texas, right next to Joe Barton's Congressional district. They constitute a risk for more than one million people. In 2005, Barton used his clout as chair of the House Energy and Commerce Committee to block chemical plant security legislation. A Republican, Barton served as a consultant for an oil and gas company before he was elected to Congress in 1984, and in the following decades he received more than $1.8 million in campaign contributions from the energy and chemical industries. Barton routinely sides with the energy industry at the expense of his constituents, and you can forget about the welfare of the people represented by his colleagues in Congress (Cohen, 2005).

For Barton to be in the position to benefit a special interest at the expense of a multitude, he must enjoy majority support in both his district and Congress. What is the logic whereby he gains such support? ‘Rational choice in political science’ stands for the import of the economics paradigm into the political science discipline, or the application of the economics approach in the study of political phenomena. The research programme is to *rationalize* collective behaviour that comes across as stupid or counterproductive. My purpose here is to spell out how this research programme is playing out in political science – or rather how it played out, for the research programme has recently lost much of its vitality (in its extreme form it is dead).

Economists and political scientists often use the same labels to denote different things and different labels to denote the same thing. For this reason, it is useful to start with some definitions: what do social choice, public choice, political economy, and positive political theory stand for, and how do they relate to ‘rational choice in political science’?

I shall illustrate these labels in the context of the scientific life cycle of rational choice in political science, which describes the usual arc of fringe, vibrancy, maturity, ossification, and renewal. Over time, the research programme cycled in its emphasis on external and internal scientific progress, as it moved from solving real-world puzzles to theoretically refining the solutions and back again; the cycle includes a forward movement.

In the economics discipline, behavioural and experimental economics have recently relaxed some of the more extreme greed-and-rationality assumptions. In political science, rational choice took a different turn. Today, rational choice in its high-brow (esoteric) variant is on the way out, in part because leading rational choice theorists are ‘holier than the Pope’ in their refusal to join the rather more relaxed approach to economics. In its low-brow (sensible) variant, rational choice is here to stay, but it has largely shed its imperialist ambitions. Instead of emerging as the dominant approach, rational choice coexists more or less peacefully with one of three complementary approaches: the rationalist approach, which focuses on individual agency; the culturalist approach, which centers on collective identities; and the structuralist approach, which emphasizes historical institutionalism. (My description of rational choice scholarship as ‘high-brow’ and ‘low-brow,’ or ‘esoteric’ and ‘sensible,’ is not meant to express approval or disapproval; the two types of scholarship complement each other as they contribute to internal and external scientific progress.)

The scientific life cycle of social choice, public choice, and political economy in the economics discipline

Because the scientific life cycle of rational choice in political science is an offshoot of the scientific life cycle of social choice, public choice, and political economy in the economics discipline, it is useful to start with an account of the latter.

Social choice was research-active from the 1950s through the 1970s; public choice, from the 1970s through the 1990s; political economy, from the 1980s through the 2000s. Dennis Mueller's textbook *Public Choice III* (2003) covers social choice and public choice. (*Public Choice III* is the third, and most comprehensive, edition of *Public Choice*, which was published in 1979.) Torsten Persson and Guido Tabellini's textbook *Political Economics* (2000) lays out the political economy programme.

When social choice started out, the reigning practice in economics was to derive normative statements about economic policy by maximizing a social welfare function subject to a set of economic constraints. At the time, it was taken for granted that people's preferences could be summarized by a social welfare function. Kenneth Arrow demonstrated, to the contrary, that it is impossible to represent people's preferences with a social welfare function that fulfills plausible criteria such as independence of irrelevant alternatives; this impossibility result holds if people's preferences are diverse. (The independence-of-irrelevant-alternatives criterion prohibits the social preference over two alternatives from switching places if the individuals' preferences over the two alternatives stay the same even as a third alternative is added to, or dropped from, the set of alternatives under consideration.) Social welfare functions subsequently went out of fashion among economists. (Actually, in many subfields of economics they returned through the side door; for example, in macroeconomics the illegitimacy of assuming a social welfare function was elegantly circumvented by assuming that the macroeconomy can be summarized by a representative agent.)

Social choice also demonstrated that voting rules affect voting outcomes. If people's preferences are sufficiently diverse, there exists no such thing as a neutral voting rule that will ‘simply’ aggregate people's preferences. At first blush, this result seems rather worrisome because of its potential to undercut the legitimacy of outcomes arrived at by democratic means. But we shall see how this insight would get picked up productively – after all, if institutions can warp democratic decision-making, this raises the possibility of designing political institutions to serve a corrective function.

Social choice consisted largely of mathematical exercises with little economic content; its concern with preference aggregation does not relate all that well to the standard economic concern with scarcity and constraint. Public choice, in comparison, employed microeconomic theory, and it was geared towards extending economic assumptions of self-interest and rationality to the political arena, with the idea of treating political and economic actors symmetrically. Before public choice entered the fray, economists were in the habit of spelling out what actions a benevolent dictator should take when he or she (or it?) encounters *market failure* due to externalities, information asymmetries and the like.

One early argument against government intervention consisted of the Coase Theorem, which implies that the system of economic actors will endogeneously adjust to internalize externalities (assuming that there exists a system of well-defined property rights and negligible transaction costs). The Coase argument can be – has been – exported to the political sector. For example, if the underlying problem in the political market consists of an information asymmetry between policymakers and voters, then information providers will have incentives to enter the political market, and voters will have incentives to take information cues from them (Wittman, 1989).

Public choice theorists, most prominently among them James Buchanan and Gordon Tullock, refused to see market failure behind every bush – in economic markets, that is. At the same time, their minds would surely boggle at the idea of political markets being self-correcting. In their eyes, *government failure* loomed large. The public choice argument against government intervention is that government consists of self-serving politicians, bureaucrats, and special interests who are poorly held in check by

ignorant voters. Public policy is thus riddled with biases and loopholes benefiting special interests at the expense of consumers and taxpayers.

Public choice theorists also proposed institutional solutions to government failure. Examples are the flat tax, the gold standard, and constitutional limits on government borrowing. The typical proposal seeks to tie the hands of politicians and bureaucrats and to create transparency vis-à-vis voters or other audiences. It forgoes the benefits of efficiency, equity and flexibility. That is, it does not offer the best solution for some idealized world lorded over by benevolent experts; it does not go out of its way to do good for the poor; and it disallows Keynesian-style economic stabilization and micro-intervention. It does, however, come with the potential to prevent self-serving politicians and bureaucrats from doling out goodies to special interests precisely because voters, or other audiences, can easily monitor slippages, make a public fuss, and ‘vote the bastards out’. The simplicity and transparency of the proposed institutions create a political cost of defecting from them.

It is worthwhile appreciating public choice for driving home this important point: as we compare different institutional solutions, we must take into account their relative political corruptibility. Public choice spelled out how the policy process is warped by collective action and political institutions. For example, if small groups have an easier time solving the free-rider problem of collective action than do large groups, then policy will be biased in favour of special interests (Olson, 1965); for example, too, the power of special interests is the result of Congressional committees being captured by high demanders, that is, members of Congress who represent constituencies (voters and campaign contributors) with a high demand for certain kinds of government handouts (Shepsle and Weingast, 1987).

From the outset public choice thus stood on two legs: one leg was about inserting politics into apolitical models of economic policy, as in, ‘the politics of monetary policy’; the other was about applying rational choice to political behaviour and institutions, as in, ‘the logic of collective action’.

Where public choice applied microeconomic theory, as in, ‘the supply of and demand for collective action’, political economy made use of game theory, as in, ‘greed, rationality, equilibrium’. The result was a higher standard of spelling out the rationality of political actors, including their informational states, and of making sure that all of their strategies and beliefs are consistent with each other so that the strategies and beliefs constitute an equilibrium.

For example, the story that policymakers pander to special interests at the expense of voters does not necessarily make sense if voters follow a voting strategy by which they vote for the incumbent when they are well off and for the challenger when they are hurting. To make this story fly, one has to specify how a policymaker can increase her re-election chances by taking something of value from the large mass of voters; losing a little of bit of it along the way (this is the deadweight loss created by redistribution, which generally distorts people's economic choices); and giving the remainder to special interests: why wouldn't the policymaker lose more votes among the large mass of voters than she would gain among the special interests (Lohmann, 1998)? And if special interests are powerful because of campaign contributions, why wouldn't voters reject a policymaker who is loaded with campaign contributions – after all, campaign contributions are a sign that the policymaker is pandering to special interests at their expense? In the same vein, if special interest handouts are the result of high demanders hogging Congressional committees, why would a majority in Congress go along with bills that benefit the committee members' constituents at the expense of *their* constituents? And why does a Congressional majority allow high demanders to self-select onto Congressional committees in the first place?

Political economy also differed from public choice by taking a balanced view of market and government failures. For example, when economic markets fail to aggregate distributed information about the mapping of economic policy into policy outcomes, then special interests or high demanders on Congressional committees may well supply the requisite information, with the result that the quality of economic policy improves (because policymakers are well-informed) even though policy outcomes are biased (because policymakers pander to special interests or high demanders) (Gilligan and Krehbiel, 1987). The implication is that we should not automatically assume that special interests and high demanders on Congressional committees are a Bad Thing; we need to consider the workings of the economic and political system as a whole, in which case a Political Bad might cancel out an Economic Bad, and the net effect is a Good Thing. Because political economy traded off the gains and losses of imperfect markets *and* imperfect politics, it came up with more complex and more flexible institutions than did public choice.

By the mid-1990s, the newly prominent subfield of experimental economics had amassed enough evidence to challenge the assumptions and predictions of standard economics models, and behavioural economics proceeded to explain the anomalies by relaxing the assumptions of greed and rationality (less so the assumption of equilibrium) in favour of a richer set of psychological motivations and cognitive limitations. All of this activity served to undercut the political economy programme of producing ever more refined rationality-and-equilibrium explanations of market-cum-government failure. Today, the extreme application of the game theory paradigm to political phenomena is passé. The cutting edge lies in employing richer models of human behaviour to understand the various forms of collective action we observe in reality, that is, in laboratory experiments and in the field.

Even as the one kind of political economy (the kind that applies old-style game theory to political phenomena) is intellectually stagnant, the other kind of political economy (the kind that inserts politics into models of economic policy) has been busy expanding into the political economy of development. Whereas public choice was largely focused on the developed countries, or the rich capitalist democracies, political economy increasingly included the developing countries, many of which were governed (some still are) by tin-pot dictators, military cliques, and the like. Whereas public choice was concerned about the discrepancy between economic theory and practice in developed democracies, the political economy of development worried about the disparities in economic performance across countries and sought to explain why and how some countries grew rich (why *these* countries, why *now*?) even as others remained poor.

Development economists who pushed this story, or variants of it, naturally appreciated the fact that well-functioning market economies rely on well-functioning governments, just as they naturally appreciated the fact that government in developed countries is functioning extremely well, both in historical and cross-country comparison. Such appreciation is a 180-degree reversal of the anti-government bias that permeated the public choice programme.

Whereas public choice had a Hayekian flavour to it (put into place simple institutions and let the economy do the rest), political economy took a rather more Keynesian approach (derive optimal institutions that will surgically correct the political economy). We are experiencing another reversal right about now, that is, a revival of the Hayekian approach. William Easterly's *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good* (2006) stands for the new bottom-up thinking (government is ‘governance by the local people’), though it is clear from

the popularity of Jeffrey Sachs's *The End of Poverty* (2005) that the old top-down thinking (government is ‘management by benevolent experts’) is not quite dead yet.

The emerging new approach – let us call it social complexity – stands in a tension with behavioural economics. The latter likes to make complicated assumptions about what is going on in people's heads even as it preserves the assumption of equilibrium, which implies that people have a complete and shared understanding of their environment (people might suffer from cognitive biases, but they best-respond to each other cognitive biases, and it all comes together very neatly). In contrast, social complexity likes to make simple assumptions about what going on in people's heads: people are relatively fixed in their behaviours, and their ‘ways of seeing’ the world are incomplete and diverse and partially inconsistent with each other (Hayek, 1945). Social complexity is actually closer to what the economics discipline used to stand for historically:

> If social phenomena showed no order except insofar as they were consciously designed, there would indeed be no room for theoretical sciences of society and there would be, as is often argued, only problems of psychology. It is only insofar as some sort of order arises as a result of individual action but without being designed by any individual that a problem is raised which demands theoretical explanation … (Hayek, 1955, p. 39)

The scientific life cycle of rational choice in political science

Now that I have reviewed how rational choice evolved in the economics discipline, let me examine how it spilled over into the political science discipline.

In the political science discipline, social choice and public choice led a peripheral existence for decades; they still do. It was only in the late 1980s and early 1990s that political economy and positive political theory exploded onto the stage. Indeed, rational choice briefly looked as if it would take over political science, only to lose influence in the early 2000s, especially in its high-brow variant; the low-brow variant has been folded into political science for the long run. Low-brow rational choice theorists, including the political economy crowd, like to use the Persson and Tabellini textbook. The high-brow crowd prefers the two-volume effort by David Austen-Smith and Jeffrey Banks, *Positive Political Theory I: Collective Preference* (1999) and *Positive Political Theory II: Strategy and Structure* (2005).

The two legs of public choice and political economy – inserting politics into models of economic policy and using the economics paradigm to model political phenomena – can be found in political science, albeit with different labels assigned to them. The label ‘political economy’ has come to stand for inserting politics into models of economic policy, as in ‘political economy of international trade’; this is typically done in a rational choice fashion, though there are some Marxist leftovers who call themselves political economists. The label ‘positive political theory’ denotes the use of rational choice to model political phenomena, as in ‘positive political theory of Congressional committees’. Why positive? To distinguish positive political theory from political theory, a subfield of the political science discipline that corresponds to the subfield political philosophy in the philosophy discipline, which is concerned with, for example, interpreting Aristotle's *Politics* or Hobbes's *Leviathan* – and also with expanding the culturalist approach, which I shall describe later.

Why did political economy and positive political theory succeed in gaining significant market share in

the political science market even as social choice and public choice were reduced to eking out a peripheral existence? Social choice modelled political behaviour and institutions in a way that was quite simply too abstract relative to the thick knowledge of political behaviour and institutions held by practising political scientists: at the time there was no research tradition in political science that could latch onto the idea that it might be interesting to prove the impossibility of a social welfare function; and as for the possibility of manipulating voting outcomes by fiddling with voting rules, what else is new? Public choice was rather more practical in its orientation, but it was for the longest time rejected as a right-wing enterprise (Lowi, 1992).

Political economy, with its more balanced take on market versus government failure, turned out to be politically more palatable. Perhaps more importantly, political economy – sailing under the flag of positive political theory – supported more refined models of political behaviour and institutions: from the perspective of a political scientist, it is not terribly interesting for a benevolent dictator to be replaced by a unitary-actor self-serving politician; it is exciting to explain why a majority in Congress would rationally constrain itself by applying closed rule to votes on proposals coming out of Congressional committees and to spell out the conditions under which the majority would allow for closed rule versus open rule (Gilligan and Krehbiel, 1987). (Under closed rule, a committee proposal must be voted up or down, with amendments prohibited. Open rule allows for amendments.)

The mid-1990s saw the first stirrings of a counter-reaction to rational choice. Donald Green and Ian Shapiro's *Pathologies of Rational Choice Theory* (1994) took potshots at rational choice. Leading rational choice theorists fought back, their responses were collected in a special issue of *Critical Review* (Friedman (1995; 1996). The two sides pretty much talked past each other, with Green and Shapiro emphasizing the empirical silliness of many rational choice models (and getting some of the models wrong) and the rational choice elite celebrating the theoretical rigour of rational choice models (and rejecting Green and Shapiro for not understanding their models).

In the early 2000s, in parallel to the post-autistic economics movement in the economics discipline, the *Perestroika* movement (Monroe, 2005) emerged seeking to liberate political science from rational choice. But if rational choice lost steam in political science, it was for a different reason than in the economics discipline. The behavioural and experimental economics revolution is not happening in political science because the leading rational choice theorists are all too heavily invested in the hyper-rational variant of the research programme. There exist political scientists who combine psychology and politics, but they are coming out of a different research tradition, one that is oblivious to economics: they rely on surveys to study mass opinion, and they are uninterested in running laboratory experiments on games that relax greed and rationality while preserving equilibrium for the simple reason that it never occurred to them that greed-rationality-equilibrium is an interesting benchmark in the first place.

Instead, people got fed up with rational choice for its esotericism. Its practitioners increasingly scored points by refining each others' theoretical models rather than by relating their models to urgent substantive problems in the real world.

Political science, as compared to the economics discipline, has a tendency to let a thousand flowers bloom: it has always supported a greater diversity of approaches in the leading doctoral programmes. Consistent with this diversity, rational choice is not actually going out of business. But it is the low-brow variant of rational choice that is surviving by combining ‘sensible’ rational choice arguments (such as the idea that collective action is subject to a free-rider problem) with an in-depth substantive

understanding of the issues. In parallel to the emergence of the political economy of development in the economics discipline, the focus of attention in political science has shifted away from the kind of positive political theory that mostly made its living in the subfields of American politics (especially in the subfield of Congressional studies) and international relations, and towards applying rational choice in the field of comparative politics, with an emphasis on developing countries.

Rational choice in its extreme form came with imperialist ambitions. Today, there is a firm understanding, at least in political science if not in the economics discipline, that rational choice is just one approach, with strengths in some domains and weaknesses in other domains, which is where alternative approaches come to life. Complementing the rationalist approach are the culturalist and structuralist approaches (Lichbach and Zuckerman, 1997; Lichbach, 2003).

By way of illustrating the culturalist approach, consider the massive social change that occurred in the United States over the second half of the 20th century – consider the civil rights movement, the women's movement, and assorted sexual liberation movements (gay, lesbian, bisexual, transgender), and contemplate the prevailing attitudes ‘then and now’ towards assorted ethnic minorities (African-American, Jewish, Native-American, Asian-American, Hispanic). Economists will glibly talk about herding effects and information cascades. But a discipline that is (variously) defined as being about scarcity or individual agency or greed-rationality-equilibrium simply does not carry much purchase when it comes to explaining such dramatic changes in collective identities.

Today, society is mired in religious conflict, both domestically and internationally. Think of the divide between the red states and blue states in the United States, that is, the states located in the vast middle of the country, whose voters predominantly back the Republican Party, and the states located on the East and West Coasts, whose voters for the most part support the Democratic Party. Think also of the divide between Islam and the West. We can talk about ‘the supply of and demand for religion’ or ‘the rational choice of religion’, but the intellectual action clearly lies someplace else, where a culturalist approach gives us a better purchase on reality.

By way of illustrating the structuralist approach, let us take a look at the current popularity of exporting democracy and building democratic institutions. In the West, the emergence of democracy took a couple of centuries (more if you include Greece and Rome). Why so long? It turns out that institutions are not designed by experts and plunked down by bureaucrats in the same way that, say, bridges are conceived by engineers and built by construction workers. Institutions are the product of social conflict and social movements, and they come into existence by spreading across the minds of a people. Social movements resolve social conflicts by locking in structures which subsequently are taken for granted, as if they exist ‘naturally’. In a well-functioning democracy, not everything is up for grabs all the time. There are huge swathes of society in which people simply play out the roles assigned to them by the structures they are embedded in. In these domains, ‘rational choice’ and ‘individual agency’ are grossly inadequate concepts. (Indeed some of the more disastrous interventions of academic economists in the real world – I have in mind the post-Communist transition to capitalism and democracy – derive from an impoverished understanding of historically evolved structures.)

We are coming full circle here, for the structuralist approach has a Hayekian touch to it. There are numerous indicators suggesting that a complex systems approach is on the rise in the social sciences. For example, as of 2006 UCLA supported a new undergraduate interdepartmental degree programme on Human Complex Systems, George Mason University, a new doctoral programme on Social Complexity.

Let me conclude. The strengths and weaknesses of rational choice in political science correspond to

those of the economics approach. Rational choice promises to make political science more scientific, as in: universally applicable, theoretically rigorous, cumulatively progressive. Thanks to the scholarship in social choice and public choice and political economy, which spilled over into political science, our understanding of collective action and political institutions is light years ahead of where it was in the 1950s.

Just like economics, however, rational choice in its extreme form (greed, rationality, equilibrium *über Alles*) is a problem. It is blind to thick and local knowledge; it disdains culture and history; and it has a tendency to degenerate into internal scientific progress rather than producing external scientific progress. Rational choice deserves to survive in political science, but it is just as well that it no longer overshadows complementary ‘ways of seeing’ the political world.

See Also

- collective action
- political institutions, economic approaches to
- public choice
- social choice

Bibliography

Austen-Smith, D. and Banks, J. 1999. *Positive Political Theory I: Collective Preference*. Ann Arbor: University of Michigan Press.

Austen-Smith, D. and Banks, J. 2005. *Positive Political Theory II: Strategy and Structure*. Ann Arbor: University of Michigan Press.

Cohen, A. 2005. A lawmaker works, oddly enough, to keep his voters' backyards dangerous: ‘Smokey Joe’ and the politics of corporate profit. *The New York Times*, 26 May, A28.

Easterly, W. 2006. *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York: Penguin.

Friedman, J., ed. 1995. Rational choice theory and politics. Special issue of *Critical Review* (vol. 9, nos. 1–2).

Friedman, J., ed. 1996. *The Rational Choice Controversy: Economic Models of Politics Reconsidered*. New Haven, CT: Yale University Press.

Gilligan, T. and Krehbiel, K. 1987. Collective decision-making and standing committees: an informational rationale for restrictive amendment procedures. *Journal of Law, Economics and Organization* 3, 287–335.

Green, D. and Shapiro, I. 1994. *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven, CT: Yale University Press.

Hayek, F.A. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

Hayek, F.A. 1955. *The Counter-Revolution of Science: Studies on the Abuse of Reason*. New York: Glencoe Free Press, 1964.

Lichbach, M.I. 2003. *Is Rational Choice Theory All of Social Science?* Ann Arbor: University of Michigan Press.

Lichbach, M.I. and Zuckerman, A.S., eds. 1997. *Comparative Politics: Rationality, Culture, and Structure*. New York: Cambridge University Press.

Lohmann, S. 1998. An information rationale for the power of special interests. *American Political Science Review* 92, 809–27.

Lowi, T.J. 1992. The state in political science: how we become what we study. *American Political Science Review* 86, 1–7.

Monroe, K.R. 2005. *Perestroika! The Raucous Rebellion in Political Science*. New Haven, CT: Yale University Press.

Mueller, D. 2003. *Public Choice III*, 3rd edn. Cambridge, MA: Cambridge University Press.

Olson, M. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Persson, T. and Tabellini, G. 2000. *Political Economics: Explaining Economic Policy*. Cambridge, MA: MIT Press.

Post-Autistic Economics Network. Home Page. Available online at. http://www.paecon.net, accessed 11 February 2007.

Sachs, J. 2005. *The End of Poverty*. New York: Penguin.

Shepsle, K. and Weingast, B. 1987. The institutional foundations of committee power. *American Political Science Review* 81, 86–108.

Wittman, D. 1989. Why democracies produce efficient results. *Journal of Political Economy* 97, 1395–424.

How to cite this article

Lohmann, Susanne. "rational choice and political science." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000253> doi:10.1057/9780230226203.1386

The New Palgrave Dictionary of Economics Online

rational choice and sociology

Peter Hedström and Charlotta Stern
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Rational-choice theorizing has a long tradition within sociology, but has always been controversial and contested. Yet it has influenced the theoretical vocabulary of the discipline at large and has made deep inroads into some important sociological areas such as social movements, social mobility, and religion. Most sociological rational-choice theories assume that actors act rationally in a broad sense, and focus on the aggregate outcomes that individual actors in interaction with one another are likely to bring about. This article reviews the most important contributions to the rational-choice tradition in sociology, and briefly discusses its historical past and its likely future.

Keywords

assumptions; behavioural economics; Blau, P.; Boudon, R.; class; Coleman, J.; collective action; economic sociology; Elster, J.; expected utility hypothesis; experimental economics; explanation; free rider problem; general equilibrium; Homans, G.; implicit contracts; methodological individualism; Olson, M.; Pareto, V.; Parsons, T.; preferences; public choice; public goods; rational choice and political science; rational-choice sociology; reciprocity; religion, sociology of; social interactions; social mobility; social networks, sociology of; social norms; transaction costs; trust; Weber, M.

Article

Rational-choice sociology is the branch of sociology which is most thoroughly influenced by economic theory. Yet it is not simply an application of economic theory to the explanation of social phenomena. Rational-choice sociology consists of a diverse set of theories only some of which can be said to have been imported from economics. The common denominator of rational-choice sociologists is that they use explanatory models in which actors are assumed to act rationally, in a wide sense of that term. Unlike in many other sociological theories, actors are not assumed to be governed by causal factors operating behind their backs, but are seen as conscious decision makers whose actions are significantly influenced by the costs and benefits of different action alternatives.

Most rational-choice sociologists do not seek to explain the actions of single individuals. The focus instead is on explaining macro-level or aggregate outcomes such as the emergence of norms, segregation patterns, or various forms of collective action. To make sense of outcomes like these, however, rational-choice sociologists focus on the actions and interactions that brought them about.

The emergence of rational-choice sociology

Rational choice-inspired theorizing has a long tradition within sociology. Max Weber, one of the founders of sociology, argued for the importance of basing sociological explanations on clearly articulated ideas about rational action (Weber, 1922). Only since the 1980s, however, have we seen the emergence of a more clearly defined rational-choice approach within sociology. Given the constraints imposed by the format of this article, we are not able to give due attention to the range of work produced by rational-choice sociologists. We instead single out a few contributions that have been particularly important for the development of the approach.

Some of the contributions that proved important for the development of rational-choice sociology were not themselves based on rational-choice assumptions. One case in point is the

work of George Homans (for example, 1958; 1964). At the height of his career Homans was a highly visible and influential sociologist who made many substantive and theoretical contributions to the discipline. Unlike many of his contemporaries he argued that sociological explanations should take the form of deductive arguments based on clearly explicated micro assumptions. In this respect he had much in common with current-day rational-choice sociologists. But unlike them he did not base his analyses on assumptions about rational actors. Instead he maintained that sociological theories should be based on assumptions derived from behavioural psychology: ‘the principles of behavioral psychology are the general propositions we use, whether implicitly or explicitly, in explaining all social phenomena’ (Homans, 1969, p. 204). Despite these differences between Homans's type of sociology and contemporary rational-choice sociology, Homans's emphasis on precise and deductive actor-based explanations meant that he paved the way for what later was to become rational-choice sociology (see Coleman, 1990a).

Another early work which was important for the emergence of rational-choice sociology was Peter Blau's *Power and Exchange in Social Life* (1964). The book covers a range of topics, but Blau was particularly interested in what we today would call implicit contract theory, and he focused in particular on the role of reciprocity in explaining the patterns of social interactions that are likely to emerge within a group of individuals. He also was interested in how differences in power and status emerge over time as the result of such exchanges (see also Emerson, 1962; Cook and Emerson, 1978).

Also of considerable importance was the economist Mancur Olson's (1965) analysis of the logic of collective action. In the pre-Olson era, most sociological theories of social movements and collective action did not problematize the distinction between individual and collective interests. Using standard microeconomic theory to analyse individuals' decisions whether or not to join an organization for collective action, Olson showed that one often should expect rational individuals to be free riders even when they would have been better off had they all joined the organization. In the light of Olson's contributions, social movement researchers started to pay much more attention to the role of individual incentives, and as a consequence, rational-choice ideas came to have a great deal of influence. Hechter's (1987) influential book on the principles of group solidarity exemplifies this trend.

In European sociology, one of the key contributors to the rational-choice tradition is Raymond Boudon (for example, 1981; 2000; 2003). In numerous publications he argued for the importance of explanations which assume that individuals act rationally. Boudon always has emphasized the importance of basing explanations on realistic theories of action, however. According to Boudon, it is important to recognize the cognitive limitations of real individuals. Individuals often act rationally in the sense of having good reasons for doing what they do, even if these actions may not necessarily be those prescribed by expected utility theory.

Other European sociologists who were important for the development of rational-choice sociology include Lindenberg (for example, 1985; 1990) and Opp (for example, 1986; 1989; see also Raub and Weesie, 1990; Abell, 1991). Lindenberg (a student of Homans) was one of the founders of the Interuniversity Center for Social Science Theory and Methodology (ICS), a Dutch graduate school built on the foundations of rational-choice theory, and he was also a driving force behind the establishment of rational-choice sections within the International Sociological Association (ISA) and the American Sociological Association (ASA).

Jon Elster is another social scientist who has been of considerable importance for rational-choice sociology. Elster's relation to rational-choice theory always has been somewhat ambivalent, however. On the one hand, he always has considered rational-choice theory to be the best available general theory of action (for example, Elster, 1986); on the other hand, most of his writings have been concerned with the limitations of rational-choice explanations. Much of his work since around 1980 has been concerned with the relationship between rationality, social norms, and emotions (for example, Elster, 1979; 1983; 1989; 1999). His writings in these areas have been widely read by sociologists and have established important links between sociological theory, the philosophy of action, and behavioural economics.

The single most important person to influence rational-choice sociology has been James Coleman. Coleman did early work on public choice theory (1966) and on the mathematics of collective action (1973), but his *Foundations of Social Theory* (1990b) is by far his most important contribution (see also Coleman, 1986, which is an important programmatic statement of his rational-choice position). This treatise of nearly 2,000 pages summarizes and extends much of the work he did during the preceding two decades. In *Foundations*, he shows how a range of traditional sociological concerns such as norms, authority systems, trust, and collective action can be addressed from a rational-choice perspective. In the final third of the book he uses a slightly modified general equilibrium model borrowed from economics to formalize many of the ideas discussed in earlier parts of the book. It is often said that *Foundations* is a book admired by many but read by few, but to judge from Marsden's analyses of citation statistics we may not yet have seen its full impact: ‘As of late 2004, more than 1850 indexed works have referenced it, the trend generally increasing over time’ (Marsden, 2005, p. 18).

Empirical research

Sociology is an empirically oriented discipline in which the success of a theoretical approach ultimately depends upon its ability to inspire new empirical research and/or to explain important empirical observations. There is a long tradition of implicit use of rationality-like assumption in empirical research, but in some areas, most notably in those concerned with social movements, social mobility, and religion, explicit rational-choice theorizing is closely allied with empirical research, and in these areas rational-choice has become an important part of the intellectual agenda.

As mentioned above, sociological research on social movements was much influenced by the work of Olson (1965), and this is clearly the area of sociology in which rational-choice

theories have made the deepest inroads. As a consequence, empirical research has paid a great deal of attention to the costs and benefits of participation when trying to explain the emergence and growth of social movements (see Udehn, 1993, for an overview). In sociology, such costs are often understood as being social in the sense that they depend upon the actions of those with whom individuals interact (for example, Opp and Gern, 1993; Hedström, 1994; Sandell and Stern, 1998).

An empirical regularity that has inspired a great deal of sociological research is the persistent influence of class background on educational choice. Boudon (1974) was an early attempt to use rational choice-inspired ideas to understand why this is so. Similarly, the educational choices of Italian youth was studied by Gambetta (1987) seeking to distinguish between the importance of choice-related factors and factors operating behind the back of the individuals.

Goldthorpe, one of the leading social mobility researchers of the last few decades, in an influential article argued for the importance of establishing closer ties between rational-choice theory and the type of statistical analyses that most social-mobility researchers were engaged in (Goldthorpe, 1996; see also Goldthorpe, 1998, and Blossfeld and Prein, 1998). Many others have followed in his path and rational-choice theory is now fairly central to this research community (for example, Breen, 1999; Jonsson, 1999; Morgan, 2002). Breen and Goldthorpe (1997), for example, developed a formal model aimed at explaining the class differential in educational attainment which assumed that families from different classes develop strategies which seek to minimize the risk of downwards mobility. This model has generated a great deal of empirical research (for example, Becker, 2003; Davies, Heinesen and Holm, 2002; Need and de Jong, 2001).

Perhaps somewhat surprisingly, sociology of religion is another area of sociology in which rational-choice theory has had a great deal of influence. For years, it was believed that modern, ‘rational’ thinking and exposure to alternative religious views would lead people to question the validity of religious belief systems and that religion would lose its foothold (for example, Berger, 1969). The situation in Europe was cited as evidence. Rational-choice sociologists, however, pointed to the United States and suggested that pluralism of religious alternatives instead is likely to increase the appeal of religion. They assumed that there exists a market for religious goods which is similar to any other market in that competition can be expected to breed efficiency and entrepreneurial activity which in turn is likely to lead to a more attractive range of religious goods and to higher consumption levels. Rational choice theorists suggested that the European situation with low religious participation was due to state regulation and ‘lazy’ religious monopolists running the churches, and these ideas have inspired a great deal of empirical research (for example, Iannaccone, Finke and Stark, 1997; Finke and Stark, 1992; Stark and Finke, 2000).

Economic sociology is another increasingly important sociological area in which rational-choice theory plays a significant, although not dominant, role. The best work in this tradition has a strong empirical grounding and explains social and economic outcomes in terms of actions constrained by the normative, institutional, and structural contexts in which the actors are embedded (see, in particular, Granovetter, 1985; Brinton and Nee, 1998).

The standing of rational-choice sociology within the discipline

Although many well-known sociologists work within the rational-choice tradition, rational-choice sociology remains controversial. In part this is because rational choice raises important questions about the very identity of sociology as an academic discipline. Classic sociologists such as Pareto (1915–16), Weber (1922), and Parsons (1937) sought to define the core identity of the discipline by contrasting it with economic theory in general, and with the micro-level assumptions of economic theory in particular. From such a perspective rational-choice sociology may appear more like an example of economic imperialism than as ‘real’ sociology, and as a consequence many contemporary sociologists consider the use of rational-choice assumptions to be a violation of a ‘disciplinary taboo’ (Baron and Hannan, 1994). The title of a recent book edited by Archer and Tritter, the former being an influential social theorist and past president of the International Sociological Association, describes the situation in a nutshell: *Rational Choice Theory: Resisting Colonisation* (Archer and Tritter, 2000). These concerns about the discipline being ‘colonized’ by rational-choice theorists appear unfounded, however. Currently there are only about 200 members in the rational-choice sections of the ASA and the ISA. Although rational-choice sociology has attracted many visible and productive sociologists, these numbers suggest that rational-choice sociology is more of an endangered species than a species likely to invade the discipline at large. This is in sharp contrast to the situation in political science, where the reception of rational choice has been positive and this approach is now widespread especially in the United States.

A recurrent theme in the criticisms advanced against rational-choice sociology concerns the realism of its assumptions. Concerns for realism are also present among many of those close to the rational-choice tradition. As mentioned above, Boudon (for example, 2003) has always emphasized the importance of realistic assumptions about the individual's social situation, incentives, and cognitive abilities. Similarly, Hedström (2005) has argued that knowingly accepting false assumptions because they lead to better predictions or to more elegant models threatens the explanatory value of the rational-choice approach because it gives incorrect answers to why we observe what we observe. Far from all sociologists are concerned about this, however. Some rational-choice sociologists take a similar position to that of Friedman (1953) and argue that the realism of the assumptions are rather irrelevant (for example, Jasso, 1988), and others argue that deviations from rationality can be ignored because they tend to be like random error terms that cancel out in the aggregate (for example, Hernes, 1992; Goldthorpe, 1998).

Sociological and economic versions of rational-choice theory

In an often-cited paper, Duesenberry (1960, p. 233) described the difference between sociology and economics as follows: ‘Economics is all about how people make choices. Sociology is all about why they don't have any choices to make.’ Although this is an obvious exaggeration of the differences between the disciplines, and particularly the differences between economists and rational-choice sociologists, it captures an important difference between the disciplines. This difference can be described using Coleman's (1986) so-called micro-macro graph (see Figure 1).

Figure 1
Coleman's (1986) micro–macro graph

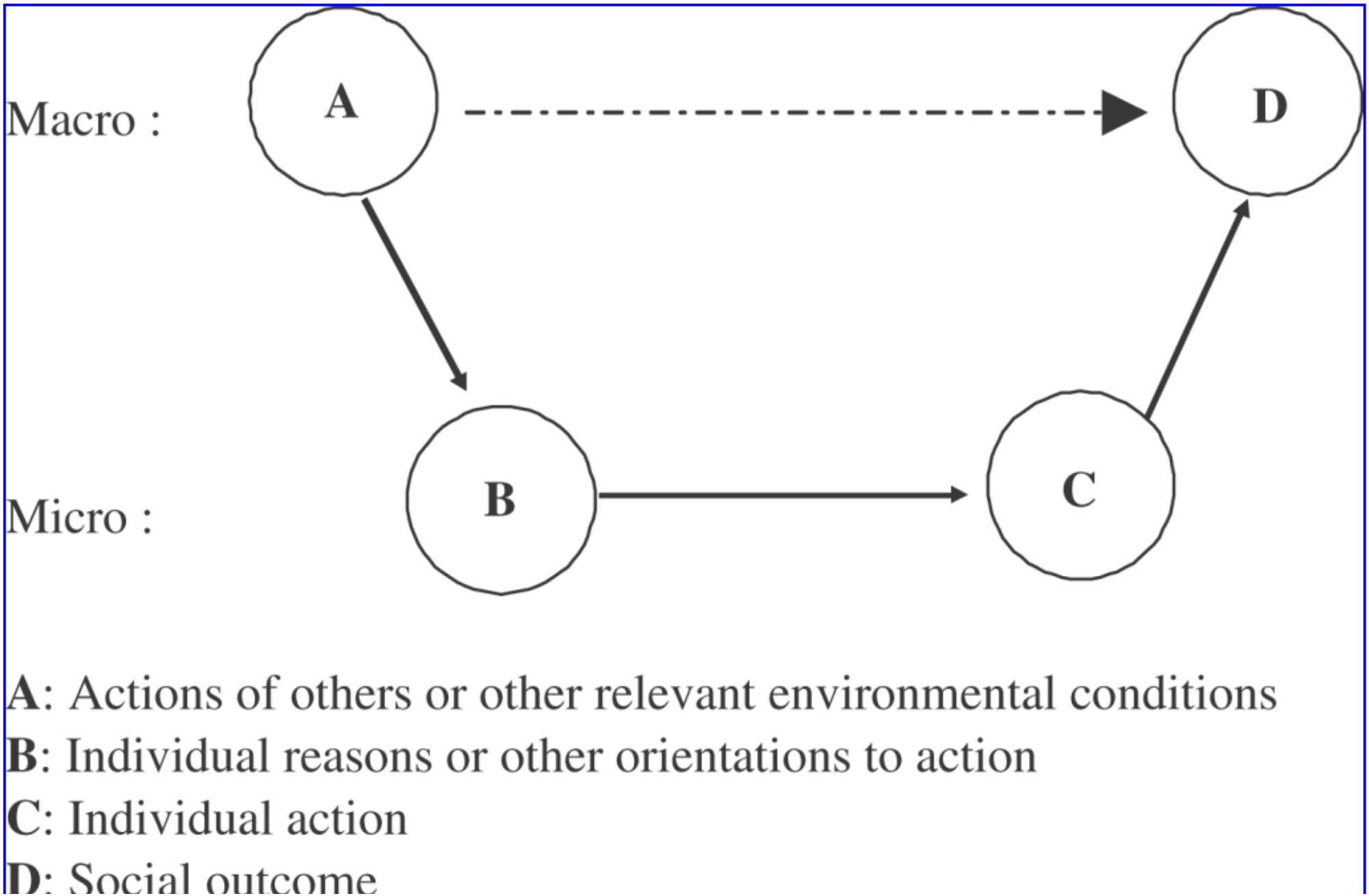

As mentioned above, rational-choice sociologists are macro-oriented but they are methodological individualists in the same sense as economists are, that is, they seek to explain macro outcomes and correlations, such as outcome **A** or the relationship between **A** and **D** in Figure 1, in terms of the intended and unintended outcomes of individuals' actions. Typically this entails explicating three causal links: (*a*) how individuals' orientations to action – their beliefs, preferences, and so on – are influenced by the social environments in which they are embedded (**A** → **B**); (*b*) how these orientations to action influence how they act (**B** → **C**); and (*c*) how these actions bring about the social outcomes to be explained (**C** → **D**).

As suggested by Duesenberry, sociologists tend to pay more attention to the macro-to-micro link (**A** → **B**) than to the latter two links. Sociologists tend to focus on how networks, social norms, socialization processes, and so on influence how individuals act by shaping their preferences, beliefs, opportunities, and so on (for example, Boudon, 1988; Burt, 1992; Coleman, 1990b; Granovetter, 1985; Hedström, 2005; Raub and Weesie, 1990). This choice of focus does not mean that sociologists believe that choices are unimportant, however; it simply is the result of an analytical focus on those aspects of the choice process which are closest to the intellectual heritage of the discipline, and therefore are perceived to be of particular sociological interest.

Another important difference between the disciplines concerns the ways in which one typically goes about analysing the type of processes described in Figure 1. While economic theory is highly mathematized, sociological theory, including sociological rational-choice theory, tends to be much more inductive and empirically oriented. For example, while most economists would specify some mathematical model in order to analyse these types of processes, most rational-choice sociologists rather would take their point of departure in the results of an empirical study. In the sociological analysis, the role of the rational-choice assumption would not be that of an assumption or a postulate of a formal model, but it would be a guide to the type of narrative used for interpreting the empirical results (see Goldthorpe, 1996, for a further discussion of this strategy).

These differences between the disciplines mean that rational-choice sociologists often use ‘broader’ notions of rational choice than economists typically do. As suggested by Camerer and Fehr (2006), the rationality assumption underlying most economic analyses consists of two components: (*a*) individuals are assumed to form, on average, correct beliefs about the world in which they are embedded, and (*b*) individuals are assumed to choose those actions that best satisfy their preferences, given these beliefs. In addition, it is typically assumed that the preferences are self-regarding, exogenously given, and stable through time. Given the sociological interest in how individuals' orientations to action – their beliefs, preferences, and so on – are influenced by the social environments in which they are embedded (the **A** → **B** link in Figure 1), assumptions about stable preferences and non-biased beliefs appear empirically problematic and they would seem to remove from the analysis some of the most interesting and intriguing aspects of the social sciences.

From the economic side of the fence, the more empirically and verbally oriented sociological approach may appear lacking in rigour, while from the sociological side of the fence there is considerable scepticism about analytical results derived from models which, at least in part, are based on assumptions that lack firm behavioural foundations. It seems likely that these disciplinary differences will become less important in the years to come because of converging trends within each discipline. Within economics, there is a growing interest in traditional sociological concerns such as norms, social interactions, and social networks, and experimental approaches are becoming increasingly more important. And within sociology there is a growing recognition of the importance of the type of formal deductive modelling that currently characterizes so much of economic theory.

Concluding remarks

At this point in time it is difficult to tell whether rational-choice sociology is destined to become an influential force within sociology. It has established itself within the discipline, and more so in Europe than in the United States, but there are no indications, to judge by the size of the rational-choice sections of the ASA and the ISA, that the number of rational-choice sociologists is increasing. Nevertheless, rational-choice theory has had and continues to have an important influence on the discipline, at least in forcing dissenters to clarify better their theoretical tools. One indication of this is that the mainstream sociological vocabulary now includes a range of concepts originating in rational-choice theory, such as free-riders, transaction costs, and collective goods. In addition, largely because of the influence of rational-choice theory, empirically oriented sociologists increasingly acknowledge the need for solid micro theories, and sociologists in general are increasingly concerned with the role of incentives in explaining actions and the collective outcomes that these actions bring about.

See Also

- collective action
- public choice
- public goods
- religion, economics of

We wish to thank Peter Abell, Filippo Barbera, Larry Blume, Richard Breen, Diego Gambetta, John Goldthorpe, Siegwart Lindenberg, Lars Udehn, Peyton Young, and Victor Nee for their useful comments.

Bibliography

Abell, P., ed. 1991. *Rational Choice Theory*. Aldershot: Edward Elgar.

Archer, M. and Tritter, J., eds. 2000. *Rational Choice Theory: Resisting Colonisation*. Florence, KY: Routledge.

Baron, J. and Hannan, M. 1994. The impact of economics on contemporary sociology. *Journal of Economic Literature* 32, 1111–46.

Becker, R. 2003. Educational expansion and persistent inequalities in Germany. *European Sociological Review* 19, 1–24.

Berger, P. 1969. *The Sacred Canopy: Elements of a Sociological Theory of Religion*. New York: Anchor Books.

Blau, P. 1964. *Exchange and Power in Social Life*. New York: John Wiley and Sons.

Blossfeld, H.-P. and Prein, G., eds. 1998. *Rational Choice Theory and Large-Scale Data Analysis*. Boulder, CO: Westview.

Boudon, R. 1974. *Education, Opportunity and Social Inequality*. New York: Wiley.

Boudon, R. 1981. *The Logic of Social Action*. London: Routledge.

Boudon, R. 1988. The logic of relative frustration. In *Rationality and Revolution*, ed. M. Taylor. Cambridge: Cambridge University Press.

Boudon, R. 2000. *The Origin of Values: Sociology and Philosophy of Beliefs*. New Brunswick: Transaction.

Boudon, R. 2003. Beyond rational choice theory. *Annual Review of Sociology* 29, 1–21.

Breen, R. 1999. Beliefs, rational choice and Bayesian learning. *Rationality and Society* 11, 463–80.

Breen, R. and Goldthorpe, J. 1997. Explaining educational differentials: towards a formal rational action theory. *Rationality and Society* 9, 275–305.

Brinton, M. and Nee, V., eds. 1998. *The New Institutionalism in Sociology*. Russel Sage Foundation: New York.

Burt, R. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Camerer, C. and Fehr, E. 2006. When does ‘economic man’ dominate social behavior? *Science* 311, 7–52.

Coleman, J. 1966. Foundations for a theory of collective decisions. *American Journal of Sociology* 71, 615–27.

Coleman, J. 1973. *The Mathematics of Collective Action*. Chicago: Aldine.

Coleman, J. 1986. Social theory, social research, and a theory of action. *American Journal of Sociology* 91, 309–1335.

Coleman, J. 1990a. Interview with James S. Coleman. In *Economics and Sociology*, ed. R. Swedberg. Princeton: Princeton University Press.

Coleman, J. 1990b. *Foundations of Social Theory*. Cambridge, MA: Belknap Press of Harvard University Press.

Cook, K. and Emerson, R. 1978. Power, equity, and commitment in exchange networks. *American Sociological Review* 43, 721–39.

Davies, R., Heinesen, E. and Holm, A. 2002. The relative risk aversion hypothesis of educational choice. *Journal of Population Economics* 15, 683–714.

Duesenberry, J. 1960. Comment on 'An Economic Analysis of Fertility'. In *Demographic and Economic Change in Developed Countries*, ed. NBER. Princeton: Princeton University Press.

Elster, J. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Elster, J. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Elster, J. 1986. Introduction. In *Rational Choice: Readings in Social and Political Theory*, ed. J. Elster. New York: New York University Press.

Elster, J. 1989. *The Cement of Society: A Study of Social Order*. Cambridge: Cambridge University Press.

Elster, J. 1999. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge: Cambridge University Press.

Emerson, R. 1962. Power-dependence relations. *American Sociological Review* 27, 31–41.

Finke, R. and Stark, R. 1992. *The Churching of America, 1776–1990: Winners and Losers in Our Religious Economy*. New Brunswick, NJ: Rutgers University Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*, ed. M. Friedman. Chicago: Chicago University Press.

Gambetta, D. 1987. *Were They Pushed or Did They Jump? Individual Decision Mechanisms in Education*. Cambridge: Cambridge University Press.

Goldthorpe, J. 1996. The quantitative analysis of large-scale data-sets and rational action theory: for a sociological alliance. *European Sociological Review* 12, 109–26.

Goldthorpe, J. 1998. Rational action theory for sociology. *British Journal of Sociology* 49, 167–92.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* 91, 481–510.

Hechter, M. 1987. *Principles of Group Solidarity*. Berkeley: University of California Press.

Hedström, P. 1994. Contagious collectivities: on the spatial diffusion of Swedish trade unions, 1890–1940. *American Journal of Sociology* 99, 1157–79.

Hedström, P. 2005. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.

Hernes, G. 1992. We are smarter than we think: a rejoinder to Smelser. *Rationality and Society* 4, 421–36.

Homans, G. 1958. Social behavior as exchange. *American Journal of Sociology* 63, 597–606.

Homans, G. 1964. Bringing men back in. *American Sociological Review* 29, 809–18.

Homans, G. 1969. The sociological relevance of behaviorism. In *Certainties and Doubts: Collected Papers, 1962–1985*. New Brunswick: Transaction Books, 1987.

Iannaccone, L., Finke, R. and Stark, R. 1997. Deregulating religion: the economics of church and state. *Economic Inquiry* 35, 350–64.

Jasso, G. 1988. Principles of theoretical analysis. *Sociological Theory* 6, 1–20.

Jonsson, J. 1999. Explaining sex differences in educational choice: an empirical assessment of a rational choice model. *European Sociological Review* 15, 391–404.

Lindenberg, S. 1985. Rational choice and sociological theory: new pressures on economics as a social science. *Journal of Institutional and Theoretical Economics* 141, 244–55.

Lindenberg, S. 1990. Homo socio-oeconomicus: the emergence of a general model of man in the social sciences. *Journal of Institutional and Theoretical Economics* 146, 727–48.

Marsden, P. 2005. The sociology of James S. Coleman. *Annual Review of Sociology* 31, 1–24.

Morgan, S. 2002. Modelling preparatory commitment and non-repeatable decisions. Information-processing, preference formation and educational attainment. *Rationality and Society* 14, 387–429.

Need, A. and de Jong, U. 2001. Educational differentials in the Netherlands: testing rational action theory. *Rationality and Society* 13, 71–98.

Olson, M. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Opp, K.-D. 1986. Soft incentives and collective action: participation in the anti–nuclear movement. *British Journal of Political Science* 16, 87–112.

Opp, K.-D. 1989. *The Rationality of Political Protest: A Comparative Analysis of Rational Choice Theory*. Boulder, CO: Westview press.

Opp, K.-D. and Gern, C. 1993. Dissident groups, personal networks, and spontaneous cooperation: the East German revolution of 1989. *American Sociological Review* 58, 659–80.

Pareto, V. 1915–16. *Mind and Society: A Treatise on General Sociology*. New York: Dover Publications, 1935.

Parsons, T. 1937. *The Structure of Social Action*. New York: Free Press.

Raub, W. and Weesie, J. 1990. Reputation and efficiency in social interactions: an example of network effects. *American Journal of Sociology* 96, 626–54.

Sandell, R. and Stern, C. 1998. Group size and the logic of collective action: a network analysis of a Swedish temperance movement 1896–1937. *Rationality and Society* 10, 327–45.

Simon, H. 1987. Behavioral economics. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Smelser, N. 1962. *Theory of Collective Behavior*. New York: Free Press.

Stark, R. and Finke, R. 2000. *Acts of Faith. Explaining the Human Side of Religion*. Berkeley: University of California Press.

Swedberg, R. 1990. *Economics and Sociology*. Princeton: Princeton University Press.

Udehn, L. 1993. 25 Years with the Logic of Collective Action. *Acta Sociologica* 36, 239–61.

Weber, M. 1922. *Economy and Society*. Berkeley: University of California Press, 1978.

How to cite this article

Hedström, Peter and Charlotta Stern. "rational choice and sociology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000249> doi:10.1057/9780230226203.1387

The New Palgrave Dictionary of Economics Online

rational expectations models, estimation of

Monika Piazzesi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Rational expectations impose cross-equation restrictions that have important implications for the estimation of models. These implications have lead to the development of new estimation and testing techniques. More recently, this development has generated techniques that handle models that cannot be solved analytically. Together with the rapid increase in computing power, these methods offer insights in to the working of these models and thereby enable their refinement.

Keywords

Bayesian methods on macroeconometrics; cross-equation restrictions; distributed lags; estimation; Euler equations; generalized method of moments; maximum likelihood; rational expectations; simulation-based estimation; term structure of interest rates; testing; vector autoregressions

Article

Most dynamic models in economics assume that agents form expectations rationally. An equilibrium of a dynamic model can typically be described by a probability distribution over sequences of data. The rational expectations assumption says that every agent's subjective belief about the data is a conditional of this equilibrium probability distribution, where the conditioning is on the agent's information set. Expectations are thus consistent with outcomes generated by the model. They are also optimal, in the sense that they correctly use all information available to the agent.
The rational expectations assumption was first proposed by John F. Muth in the early 1960s in his analysis of linear macroeconomic models. Prior to Muth's work, expectations in those models had been parametrized distributed lags. In the early 1970s, Robert E. Lucas Jr. studied the rational expectations equilibrium of a model with optimizing agents who have different information sets. It was recognized early on that taking rational expectations models to data required new techniques. Building on the early work on tests of the natural rate hypothesis by Sargent (1971), there has been much progress in rational

expectations econometrics since the mid-1970s (for example, see Hansen and Sargent, 1980; 1991; Lucas and Sargent 1981). In the meantime, the rational expectations assumption has come to be used in many fields of economics, including finance, labour economics and industrial organization.
Rational expectations impose cross-equation restrictions that have important implications for the estimation of models, which I will describe below. These implications have lead to the development of new estimation and testing techniques. More recently, this development has generated techniques that handle models that cannot be solved analytically. Together with the rapid increase in computing power, these methods offer insights in to the working of these models and thereby enable their refinement.

Cross-equation restrictions

The rational expectations assumption implies cross-equation restrictions that constrain parameters and shocks different places of the model. There are (at least) three reasons for why these restrictions have important implications for estimation. First, cross-equation restrictions constrain the parameters associated with agents' expectations to be consistent with the parameters from the equilibrium probability distribution. These restrictions reduce the overall number of parameters that have to be estimated. In particular, they eliminate any free parameters associated with expectations. To see why, consider a dynamic model with an agent who maximizes some objective function subject to constraints. To solve this optimization program, the agent needs to form expectations about future variables such as growth rates. In a model without rational expectations, these expectations might be based on some subjective belief about the future. This belief introduces free parameters that need to be estimated in addition to other model parameters, such as preference parameters.
Take, for example, an endowment economy populated by a representative agent with time separable power utility. The agent may be optimistic and believe in high mean growth rates for the endowment. This optimistic belief will have an affect on equilibrium outcomes. For example, the agent's Euler equation will only hold for a high short real rate, because the high mean growth rate implies a strong consumption smoothing motive. However, the actual mean growth rate in this economy may be lower than what the agent believes (so that the agent will be disappointed by the endowment realizations.)
The estimation of the model with an optimistic agent involves two parameters, the subjective mean of endowment growth and its true mean, which is the mean of the data generating process of endowment growth. The assumption of rational expectations reduces the number of parameters to estimate, because the two mean parameters collapse: the agent's subjective belief is equal to the true data-generating process. In this simple example, the cross-equation restrictions only eliminated one parameter. In more realistic examples, the agent's subjective belief may involve many parameters (for example, because it is described by a vector autoregression in many variables and with many lags), so that the restrictions are important for keeping the estimation tractable.
The second important implication of cross-equation restrictions is that the processes for different endogenous variables often involve the same parameters and shocks. As a consequence, different data series are informative about the same set of parameters. This implication can be used to increase the efficiency of the estimation. Going back to the example of a representative agent endowment economy, the equation describing the equilibrium process of an interest rate on a bond with m-period maturity is intimately related to the equation describing the process of an n-period interest rate for some $m \neq n$. The relationship between different interest-rate equations, or restrictions across equations, consists of

parameters that enter both equations (for example, expected growth) and also shock processes that affect both equations (for example, surprises in growth). These restrictions help in the estimation and can be tested empirically with data on interest rates with different maturities.

Some of the earliest tests of cross-equation restrictions were indeed tests of the implications of rational expectations for the term structure of interest rates. Sargent (1979) specifies a vector autoregression (VAR) for short and long rates. Assuming Gaussian disturbances, Sargent estimates this VAR using maximum likelihood and performs likelihood ratio tests to see whether the restrictions imposed by the expectations hypothesis are satisfied. Subsequently, these tests were further refined, and the expectations hypothesis (which is a stronger assumption than rational expectations) was rejected in many empirical studies. The lessons from these statistical rejections have resulted in refined models with rational expectations but time-varying risk premia (for example, Ang and Piazzesi, 2003).

The third important implication of rational expectations is that the data-generating process that underlies agent beliefs is equal to the true data-generating process. This enables the estimation of rational expectations models using the generalized method of moments based on moment conditions derived from Euler equations (see Hansen, 1982; Hansen and Singleton, 1982; generalized method of moments estimation). Using the law of iterated expectations, such a GMM estimation also allows for the case that agents in the model have more information than the econometrician.

Estimation methods

Estimation methods for rational expectations models can be distinguished by the amount of information they require. Generally speaking, there are full information methods and limited information methods. The goal of full information methods is to estimate the entire model by exploiting all its cross-equation restrictions. This estimation method is efficient and produces estimates for all the parameters in the model. These methods are maximum likelihood and its Bayesian counterparts (see Bayesian methods in macroeconometrics). To apply these methods, the econometrician needs to specify the entire structure of the model, including the distribution of shocks.

Limited information methods require less structure. The goal of these methods is to exploit only some of the restrictions imposed by the model and to obtain estimates for only some of the model parameters. These methods lose some of the efficiency of the full information methods, but they help the researcher to avoid contaminating the estimation results by model misspecification in parts of the model that are not of interest. For example, Hall (1978) and Hansen and Singleton (1982) use the Euler equations from a single agent model as moment conditions for GMM and measure the empirical counterparts of these moments using data on consumption and financial returns. This procedure gives estimates for preference parameters and does not depend on any specific assumption on the distribution of shocks in the model.

Faced with the difficulty that many models do not have analytical solutions and have to be solved numerically, there has been progress regarding simulation-based estimation methods. These methods compare moments of data simulated from the model using some parameter values with their empirical counterparts. For a textbook treatment of these methods, see Gourieroux and Monfort (1996), Gourieroux and Jasiak (2001), and Singleton (2006).

See Also

- Bayesian methods in macroeconometrics
- generalized method of moments estimation
- rational expectations

Bibliography

Ang, A. and Piazzesi, M. 2003. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics* 50, 745–87.

Gourieroux, C. and Jasiak, J. 2001. *Financial Econometrics*. Princeton: Princeton University Press.

Gourieroux, C. and Monfort, A. 1996. *Simulation-Based Econometric Methods*. Oxford: Oxford University Press.

Hall, R.E. 1978. Stochastic implications of the life cycle permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86, 971–87.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Hansen, L.P. and Sargent, T.J. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2, 7–46.

Hansen, L.P. and Sargent, T.J. 1991. *Rational Expectations Econometrics*. Boulder, CO: Westview Press.

Hansen, L.P. and Singleton, K. 1982. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249–65.

Lucas, R.E. and Sargent, T.J. 1981. *Rational Expectations and Econometric Practice*. Minneapolis: University of Minnesota Press.

Sargent, T.J. 1971. A note on the accelerationist controversy. *Journal of Money, Credit, and Banking* 3, 721–5.

Sargent, T.J. 1979. A note on maximum likelihood estimation of the rational expectations model of the term structure. *Journal of Monetary Economics* 5, 133–43.

Singleton, K. 2006. *Empirical Dynamic Asset Pricing*. Princeton: Princeton University Press.

How to cite this article

Piazzesi, Monika. "rational expectations models, estimation of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000269> doi:10.1057/9780230226203.1389

The New Palgrave Dictionary of Economics Online

rational expectations

Thomas J. Sargent

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Rational expectations is an equilibrium concept that attributes a common model (a joint probability distribution over exogenous variables and outcomes) to nature and to all agents in the model. The rational expectations equilibrium concept makes parameters describing agents' belief disappear as components of a model, giving rise to the cross-equation restrictions that offer rational expectations models their empirical power.

Keywords

adaptive expectations; cobweb theorem; computatuion of equolibria; consumer surplus; corn-hog cycles; curse of dimensionality; distributed lags; Euler equations; forecasting; generalized method of moments; identification; Lagrange multipliers; linear prediction theory; Markov chain Monte Carlo methods; Markov processes; maximum likelihood; Muth, J. F.; ommitment; ordinary differential equations; producer surplus; rational expectations; rational expectations equilibrium; real business cycles; reputation; self-confirming equilibria; shadow pricing; stochastic approximation; stochastic optimal growth models; time consistency

Article

‘Rational expectations’ is an equilibrium concept that can be applied to dynamic economic models that have elements of ‘self-reference’, that is, models in which the endogenous variables are influenced by the expectations about future values of those variables held by the agents in the model. The concept was introduced and applied by John F. Muth (1960; 1961) in two articles that interpreted econometric distributed lag models. Muth used explicitly stochastic dynamic models and brought to bear his extensive knowledge of classical linear prediction theory to interpret distributed lags in terms of economic parameters. For Muth, an econometric model with rational expectations possesses the defining property that the forecasts made by agents within the model are no worse than the forecasts that can be

made by the economist who has the model.

Muth's first concrete application of rational expectations was to find restrictions on a stochastic process for income that would render Milton Friedman's (1957) geometric distributed lag formula for permanent income an optimal predictor for income. Muth showed that, if the first difference of income is a first-order moving average process, then Friedman's formula is optimal for forecasting income over any horizon. The independence of this formula from the horizon makes precise the sense in which Friedman's formula extracts from past income an estimator of ‘permanent’ income. In working backwards from Friedman's formula to a process for income in this way, Muth touched Lucas's critique (1976). Given any distributed lag for forecasting income, one can work backwards as Muth did and discover a stochastic process for income that makes that distributed lag an optimal predictor for income over some horizon. Similarly, Sargent (1977) reverse engineered a joint inflation-money creation process that makes Cagan's (1956) adaptive expectations scheme for forecasting inflation a linear least squares forecast.

Solving a few such inverse-optimal prediction problems in the fashion of Muth and Sargent quickly reveals the dependence of a distributed lag for forecasting the future on the form of the stochastic process that is being forecast. In 1963, Peter Whittle published a book that conveniently summarized and made more accessible to economists the classical linear prediction theory that Muth had used. That book repeatedly applies the Wiener–Kolmogorov formula for the optimal *j*-step ahead predictor of a covariance stationary stochastic process x_t with moving average representation $x_t=c(L)\ \varepsilon_t$. The Wiener–Kolmogorov formula displays the dependence of the optimal distributed lag for predicting future x on the form of $c(L)$. That dependence underlies Lucas's critique of econometric policy evaluation procedures that were common when Lucas composed his critique in 1973. Those procedures had assumed that distributed lags in behavioural relations would remain invariant with respect to alterations in government policy rules, alterations that took the form of changes in $c(L)$ for government policy instruments. Although the formulas in Whittle's book were used extensively by Nerlove (1967) to work out additional examples along the lines of Muth, it was not until the writing of Lucas's critique in 1973 and its publication in 1976 that the implications for econometric practice of Muth's ideas and the prediction formulas in Whittle began to be widely appreciated.

Lucas and Prescott (1971) clarified and extended rational expectations as an equilibrium concept and also pointed the way to connecting theory with observations. They described the partial equilibrium of an industry in which there exists a fixed number of identical firms, each subject to costs of adjustment for a single factor of production, capital. The industry faces a downward sloping demand curve for its output that shifts randomly due to a demand shock that follows a Markov process. The representative firm maximizes the expected present value of its profits by choosing a contingency plan for investment. To state the firm's optimum problem, it is necessary to describe what the firm believes about the motion of variables that influence its future returns even though they are beyond the firm's control. The price of output is such an uncontrollable variable, but the demand curve for output and the hypothesis of market clearing make price a function of the capital stock in the industry as a whole. It follows that to state the firm's decision problem requires the firm's view about the law of motion of the industry-wide capital stock be stated. The representative firm's optimum problem can then be solved, yielding a law of motion for the capital stock of the representative firm in which both the individual firm's capital stock and the market-wide capital stock are both state variables. Multiplying this law of motion by the number of

firms then gives the actual law of motion for capital in the industry. In this way, the firm's optimization problem and the hypothesis of market clearing induce a mapping from a perceived law of motion to an actual law of motion for the industry's capital stock. A rational expectations equilibrium is a fixed point of this mapping. By studying an artificial planning problem that maximizes consumer plus producer surplus, Lucas and Prescott pursued an indirect approach to describing conditions under which a unique fixed point exists. In this way, they formulated a recursive competitive equilibrium.

From a practical perspective, an important property of a rational expectations model is that it imposes a communism of models and expectations. If we define a model as a probability distribution over a sequence of outcomes, possibly indexed by a parameter vector, a rational expectations equilibrium asserts that the same model is shared by (1) all of the agents within the model, (2) the econometrician estimating the model, and (3) nature, also known as the data generating mechanism. Different agents might have different information, but they form forecasts by computing conditional expectations with respect to a common joint density, that is, a common model. Communism of models gives rational expectations much of its empirical power and underlies the cross-equation restrictions that are used by rational expectations econometrics to identify and estimate parameters. A related perspective is that, within models that have unique rational expectations equilibria, the hypothesis of rational expectations makes agents' expectations disappear as objects to be specified by the model-builder or to be estimated by the econometrician. Instead, they are equilibrium outcomes.

The equilibrium law of motion for capital induces a stochastic process for capital that assumes the form of a Markov process. Lucas and Prescott showed that this Markov process converges in distribution to a unique invariant distribution. That justifies an asymptotic distribution theory adequate for doing time series econometrics, in particular, a mean ergodic theorem that guarantees that sample moments converge to the corresponding population moments. Lucas and Prescott's notion of a recursive competitive equilibrium thus takes a big step towards integrating dynamic theory and econometrics because it supplies an explicit mapping from economic parameters describing preferences, technology, and information sets to the population moments of observable sequences of economic time series. The task of econometrics under rational expectations is to ‘invert’ this mapping by using time series data to make inferences about economic parameters.

Hansen and Sargent (1980) used linear versions of Lucas–Prescott and Brock and Mirman (1972) models as laboratories for working out econometric techniques for estimating rational expectations models. They studied both generalized method of moments (GMM) and maximum likelihood approaches. They described how desirable statistical properties including consistency and asymptotic efficiency for estimators of the model's economic parameters induce a metric for measuring distance between the sample moments and the theoretical population moments implied by the equilibrium of the model at given parameter values. Typical metrics are those associated with the generalized method of moments, a special case of which is associated with the first-order conditions for maximizing a Gaussian likelihood function. Parameter estimates are obtained by minimizing the metric with respect to the parameter values, a nonlinear minimization problem.

Econometric identification of parameters means uniqueness of the minimizer of distance between the theory and the observations. Identification is partially achieved by the rich set of cross-equation restrictions that the hypothesis of rational expectations imposes (the same parameters appear in many equations, in highly nonlinear ways). These cross-equation restrictions achieve identification in a different manner from the Cowles Commission's ‘rank and order’ conditions, which explicitly excluded

cross-equation restrictions. Dynamic rational expectations models subvert such ‘exclusion restrictions’, and thereby destroy the neat division between ‘supply’ and ‘demand’ curves that underlay the ‘exclusion’ approach to identification.

Minimum distance estimation of a rational expectations model requires recomputing an equilibrium for each set of parameter values used during a descent with respect to the data-fitting metric. Except for linear models, Bellman's ‘curse of dimensionality’ makes it challenging to compute an equilibrium, so developing improved computational methods has become an important research area. Judd (1998) describes a variety of numerical approaches. Methods for computing equilibria are required not only for parameter estimation, but also for quantitatively evaluating the effects of proposed interventions, for example, new policies for setting government instruments. A new government policy implies, via the cross-equation restrictions, new laws of motion for all the endogenous variables in the models. It is no coincidence that full information estimation methods require calculations closely connected to those needed to evaluate policy.

Good computer programmes for solving and estimating complete rational expectations models have recently become available. A suite of Matlab programmes called Dynare was written by Michele Juilliard and colleagues and is available on the Internet. Dynare solves linear models as systems of expectational difference equations using methods originally described by Sargent (1979), Blanchard and Khan (1980), and Whiteman (1983). Dynare estimates models by either maximum likelihood or a Markov chain Monte Carlo procedure to construct a Bayesian posterior density over free parameters. Dynare also knows how to compute and estimate various linear and log-linear approximations to nonlinear models.

Hansen and Singleton (1982) suggested a short-cut estimation method capable of estimating the parameters of a subset of preference and technologies without computing or estimating a complete equilibrium. Their idea was to use back out parameter estimates from conditional moment restrictions implied by the first-order necessary conditions (Euler equations) for an agent's dynamic optimization problem. Hansen and Singleton pointed out that their GMM method requires special restrictions on the stochastic process of disturbances to the function being estimated, and that it typically fails to estimate enough parameters to permit evaluating many kinds of interventions. Nevertheless, its ease of use and presumed robustness to features of the environment that a researcher prefers not to specify have made it a very popular and fruitful approach.

As already mentioned, a rational expectations equilibrium is a fixed point from a perceived to an actual law of motion. It is tempting to hope that iterations on that mapping converge to a fixed point. But that is asking for too much because the mapping is not a contraction and it is easy to construct examples in which iterations diverge. Nevertheless, the mapping from a perceived to an actual law of motion plays an important role in studying how a rational expectations equilibrium can emerge as the limit point of a system of adaptive agents who use least squares on historical data to forecast the future, rather than the population moments from the equilibrium that are handed to them within a rational expectations equilibrium. By applying the theory of stochastic approximation, Marcet and Sargent (1989) and Woodford (1990) derived an ordinary differential equation (ODE) for beliefs that describe the limiting behaviour of such an adaptive system. That ODE expresses how the gap between the perceived and implied actual law of motion governs a limiting rate of change of beliefs. Necessary and sufficient conditions for convergence to a rational expectations equilibrium are stated in terms of the stability of

the associated ode. These conditions have been dubbed the E-stability conditions by Evans and Honkapohja (2001) and are useful for constructing algorithms for computing rational expectations equilibria via least squares learning algorithms or direct attacks on the ordinary differential equation governing E-stability. This is in effect what Krusell and Smith (1998) do, though they do not connect their method to the learning literature.

The literature on least squares learning and adaptive learning in games (for example, Marcet and Sargent, 1989; Woodford, 1990; Fudenberg and Levine, 1998) began partly as a response to a widespread scepticism about the plausibility of the communism of expectations imposed by rational expectations. How could people possibly come to learn to share a common model with each other, the econometrician, and nature? The learning literature offers an explanation. But the learning literature falls short of implying a communism of models as extensive as the one typically imposed in macroeconomics. A meta-theorem is that, if a system of least squares agents converges, it converges to a self-confirming equilibrium (see Fudenberg and Levine, 1998; Sargent, 1999). In a self-confirming equilibrium, agents' models agree about events that occur frequently enough (infinitely often) within the equilibrium. But agents can have different subjective distributions about events that occur infrequently because they are off the equilibrium path. For those events, a law of numbers just doesn't have enough observations to work on. In a macro model, it is typically irrelevant that private agents' beliefs can be wrong off an equilibrium path because, being atomistic, all that matters for them are their conditional forecasts along an equilibrium path. But for the government, its beliefs about off-equilibrium paths events influence its choices in important ways: designing government policy is all about evaluating the effects of alternative hypothetical outcome paths, most of which will not be observed. Kreps (1998) defends the concept of self-confirming equilibrium.

Lucas and Prescott's model can be used to study aspects of the theory of policy. Their model generates a stochastic process for output, price and industry capital that exhibits recurrent but aperiodic ‘cycles’, as realizations of stochastic difference equations do. Thus, Lucas and Prescott's model is an alternative to the ‘cobweb’ mechanism for generating fluctuations in commodity markets. Two-industry versions of the model can readily be constructed to model ‘corn-hog’ cycles. Models along the lines of Lucas and Prescott's reveal a different perspective on these cycles than do cobweb models. Lucas and Prescott show that, despite cyclical fluctuations, the equilibrium of their model is optimal in the sense that it maximizes the expected present value of consumer surplus net of producer surplus. Therefore, unlike cobweb models, in which cycles partly reflect erroneous and readily improved upon perceptions of private agents, matters cannot be improved by government interventions designed to smooth out the cycles. Models of this kind have been calibrated to price and quantity data from markets for cattle, housing, and engineers by Rosen, Murphy and Scheinkman (1994), Topel and Rosen (1988), and Ryoo and Rosen (2004).

For studying a variety of macroeconomic questions, researchers have used what can be interpreted as a version of Lucas and Prescott's model, suitably modified and reinterpreted to apply to an aggregative economy. Brock and Mirman (1972) analysed a centralized version of such an economy that took the form of a stochastic version of a one-sector optimal growth model. The planner in their model seeks to maximize the expected discounted value of utility of consumption subject to a technology for transforming consumption over time via investment in physical capital. Brock and Mirman gave conditions under which the optimal plan for capital and consumption induces a stochastic process that

converges in distribution, so that, like Lucas and Prescott's model, theirs is prepared for rigorous treatment econometrically. It is possible to decentralize Brock and Mirman's model into an equivalent economy consisting of competitive firms and households who interact in markets for labour and capital and who have rational expectations about the evolution of the wages and interest rates that they face. Decentralized versions of Brock–Mirman models have been used to construct equilibrium theories of stock prices and interest rates, typically by computing particular shadow prices associated with the planning problem (Lucas, 1978; Brock, 1982). Decentralized versions of the Brock–Mirman model form the backbone of the modern version of ‘real business cycle theory’ that was initiated by Kydland and Prescott (1982). Since the stochastic optimal growth model has a stochastic difference equation for capital as its equilibrium, it shares with the Lucas–Prescott model the property that it readily generates realizations for capital, output and consumption that display recurrent but aperiodic fluctuations of the kind observed in aggregate time series data. Kydland and Prescott embarked on the task of taking seriously the possibility that the preferences and technology of a small stochastic optimal growth model could be specified so that it would approximate closely the moments of a list of important aggregate economic time series for the United States. Kydland and Prescott have constructed several such models, each driven by a single unobserved shock, which they interpret as a disturbance to technology. This research strategy is charged with meaning, since it undertakes to explain aggregate time series data with a model whose equilibrium is optimal, and in which there is no government. The government is neither a contributing source to economic fluctuations nor a potential modifier of those fluctuations. Real business cycle models of this kind are capable of determining a long list of real variables, while remaining silent about all nominal variables.

But central banks are supposed to determine nominal variables, which has created an interest in adapting real business cycle models to include interactions among nominal and real variables. By directly imposing parameterized versions of wage and price inertia, Smets and Wouter (2003) and Woodford (2003) have formulated rational expectations models with enough shocks and rigidities to fit macro data well enough to be useful to research departments of leading central banks. These models can be estimated and simulated with Dynare.

The idea of rational expectations was essential for formulating the problem of time inconsistency in macroeconomics. Three ideas underlie the time consistency problem in multi-agent dynamic games and macroeconomic models: (1) the communism of models brought by rational expectations, (2) backward induction by all agents, and (3) the observation that different timing protocols generally imply different outcomes. The time inconsistency ‘problem’ was recognized in macroeconomics by Kydland and Prescott (1977) and Calvo (1978), who studied macro models in which a competitive economy with a representative agent confronts a benevolent government. These papers compare outcomes under two timing protocols. In one timing protocol, private agents choose sequentially but the government has a commitment technology that allows it once and for all at time zero to choose an entire history contingent sequence of actions (for example, tax rates or money supplies). In the other, the government, or a sequence of government administrations if you prefer, must choose sequentially, that is, anew each period. Outcomes under these two timing protocols typically differ, with outcomes being better under the timing protocol that allows the government to choose once and for all at time 0. The difference in outcomes shows the value of being able to commit at time 0. In the problem under commitment, among the constraints that the government faces at time 0 are a sequence of private agents' Euler equations that

involve their (rational) expectations of future government actions. The equilibrium time t values of the Lagrange multipliers on these ‘implementability constraints’ encode the costs in terms of the government's time t continuation value of confirming the time t expectations that the government's time 0 plan had induced private agents to expect. The presence of those implementability conditions in the government's constraint set gives rise to a conflict between the preference orderings of the government and the representative agent over outcomes. That conflict is the ultimate source of the timing inconsistency problem. Recursive methods for computing the optimal plan under commitment were first suggested by Kydland and Prescott (1980) and are surveyed in Ljungqivst and Sargent (2004). These methods are used extensively in the literature on rational expectations monetary models with ad hoc inertial wages and prices that Woodford (2003) catalogues and extends.

An important literature studies whether reputation can overcome the time inconsistency problem. The finding of this literature is that, by allowing history-dependent strategies, reputation can substitute for the ability to commit if the discount factor is sufficiently close to one. This literature, which is surveyed critically in Ljungqivst and Sargent (2004), exploits the communism of expectations inherent in rational expectations.

See Also

- business cycle measurement
- inflation expectations
- new classical macroeconomics
- self-confirming equilibria

Bibliography

Blanchard, O.J. and Kahn, C.M. 1980. The solution of linear difference models under rational expectations. *Econometrica* 48, 1305–11.

Brock, W.A. 1982. Asset prices in a production economy. In *Economics of Information and Uncertainty*, ed. J.J. McCall. Chicago: University of Chicago Press.

Brock, W.A. and Mirman, L. 1972. Optimal economic growth under uncertainty: the discounted case. *Journal of Economic Theory* 4, 479–513.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the Quantity Theory of Money*, ed. M. Friedman. Chicago: University of Chicago Press.

Calvo, G.A. 1978. On the time consistency of optimal policy in a monetary economy. *Econometrica* 46, 1411–28.

Evans, G.W. and Honkapohja, S. 2001. *Learning and Expectations in Macroeconomics*. Princeton:

Princeton University Press.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Fudenberg, D. and Levine, D.K. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Hansen, L.P. and Sargent, T.J. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2, 7–46.

Hansen, L.P. and Singleton, K.J. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–86.

Judd, K. 1998. *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Kreps, D.M. 1998. Anticipated utility and dynamic choice. In *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures, 1983–1997*, ed. D.P. Jacobs, E. Kalai and M.I. Kamien. Cambridge: Cambridge University Press.

Krusell, P. and Smith, A. 1998. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy* 106, 867–96.

Kydland, F.E. and Prescott, E.C. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

Kydland, F.E. and Prescott, E.C. 1980. Dynamic optimal taxation, rational expectations and optimal control. *Journal of Economic Dynamics and Control* 2, 79–91.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lucas, R.E., Jr. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and the Labor Market*, Vol. 1 of Carnegie-Rochester Conference in Public Policy. ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

Lucas, R.E., Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

Lucas, R.E., Jr. and Prescott, E.C. 1971. Investment under uncertainty. *Econometrica* 39, 659–81.

Ljungqivst, L. and Sargent, T.J. 2004. *Recursive Macroeconomic Theory*, 2nd edn. Cambridge, MA: MIT Press.

Marcet, A. and Sargent, T.J. 1989. Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory* 48, 337–68.

Muth, J.F. 1960. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55, 299–306.

Muth, J.F. 1961. Rational expectations and the theory of the price movements. *Econometrica* 29, 315–35.

Nerlove, M. 1967. Distributed lags and unobserved components in economic time series. In *Ten Economic Studies in the Tradition of Irving Fisher*, ed. W. Fellner. New York: Wiley.

Rosen, S. and Topel, R.H. 1988. Housing investment in the United States. *Journal of Political Economy* 96, 718–40.

Rosen, S., Murphy, K.M. and Scheinkman, J.A. 1994. Cattle cycles. *Journal of Political Economy* 102, 468–92.

Ryoo, J. and Rosen, S. 2004. The engineering labor market. *Journal of Political Economy* 112, S110–S139.

Sargent, T.J. 1977. The demand for money during hyperinflations under rational expectations: I. *International Economic Review* 18, 59–82.

Sargent, T.J. 1979. *Macroeconomic Theory*. New York: Academic Press.

Sargent, T.J. 1984. Autoregressions, expectations, and advice. *American Economic Review* 74, 408–15.

Sargent, T.J. 1999. *The Conquest of American Inflation*. Princeton: Princeton University Press.

Sargent, T.J. and Wallace, N. 1976. Rational expectations and the theory of economic policy. *Journal of Monetary Economics* 2, 169–83.

Smets, F. and Wouter, R. 2003. An estimated dynamic stochastic general equilibrium model of the Euro area. *Journal of the European Economic Association* 1, 1123–75.

Whiteman, C.H. 1983. *Linear Rational Expectations Models: A Users Guide*. Minneapolis: University of Minnesota Press.

Whittle, P. 1983. *Prediction and Regulation by Linear Least-Square Methods*, 2nd edn. Minneapolis: University of Minnesota Press.

Woodford, M. 1990. Learning to believe in sunspots. *Econometrica* 58, 277–307.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

Sargent, Thomas J. "rational expectations." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000025> doi:10.1057/9780230226203.1388

The New Palgrave Dictionary of Economics Online

rationality, bounded

Herbert A. Simon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Bounded rationality’ refers to rational choice that takes into account the cognitive limitations of the decision-maker – limitations of both knowledge and computational capacity. It is a central theme in the behavioural approach to economics. Theories of bounded rationality can be generated by relaxing one or more of the assumptions of subjective utility theory underlying neoclassical economics. They insist that the model of human rationality must be derived from detailed and systematic empirical study of human decision-making behaviour in laboratory and real-world situations. For example, a satisficing strategy may be postulated instead of the maximization of a utility function.

Keywords

behavioural approach to economics; bounded rationality; choice; cognitive limits; expected utility; money illusion; neoclassical economics; optimal search; probability; rational expectations; satisficing; Simon, H. A.; subjective utility theory; substantive vs. procedural rationality; uncertainty

Article

The term ‘bounded rationality’ is used to designate rational choice that takes into account the cognitive limitations of the decision-maker – limitations of both knowledge and computational capacity. Bounded rationality is a central theme in the behavioural approach to economics, which is deeply concerned with the ways in which the actual decision-making process influences the decisions that are reached.
The theory of subjective utility (SEU theory) underlying neo-classical economics postulates that choices are made: (1) among a given, fixed set of alternatives; (2) with (subjectively) known probability distributions of outcomes for each; and (3) in such a way as to maximize the expected value of a given utility function (Savage, 1954). These are convenient assumptions, providing the basis for a very rich and elegant body of theory, but they are assumptions that may not fit empirically the situations of economic choice in which we are interested.

Theories of bounded rationality can be generated by relaxing one or more of the assumptions of SEU theory. Instead of assuming a fixed set of alternatives among which the decision-maker chooses, we may postulate a process for generating alternatives. Instead of assuming known probability distributions of outcomes, we may introduce estimating procedures for them, or we may look for strategies for dealing with uncertainty that do not assume knowledge of probabilities. Instead of assuming the maximization of a utility function, we may postulate a satisficing strategy. The particular deviations from the SEU assumptions of global maximization introduced by behaviourally oriented economists are derived from what is known, empirically, about human thought and choice processes, and especially what is known about the limits of human cognitive capacity for discovering alternatives, computing their consequences under certainty or uncertainty, and making comparisons among them.

Generation of alternatives

Modern cognitive psychology has studied in considerable depth not only the processes that human subjects use to choose among given alternatives, but also the processes (problem-solving processes) they use to find possible courses of action (i.e., actions that will solve a problem) (Newell and Simon, 1972).

If we look at the time allocations of economic actors, say business executives, we find that perhaps the largest fraction of decision-making time is spent in searching for possible courses of action and evaluating them (i.e., estimating their consequences). Much less time and effort is spent in making final choices, once the alternatives have been generated and their consequences examined. The lengthy and crucial processes of generating alternatives, which include all the processes that we ordinarily designate by the word ‘design’, are left out of the SEU account of economic choice.

Study of the processes for generating alternatives quickly reveals that under most circumstances it is not reasonable to talk about finding ‘all the alternatives’. The generation of alternatives is a lengthy and costly process, and one where, in real-world situations, even minimal completeness can seldom be guaranteed. Theories of optimal search can cast some light on such processes, but, because of limits on complexity, human alternative-generating behaviour observed in the laboratory is usually best described as heuristic search aimed at finding satisfactory alternatives, or alternatives that represent an improvement over those previously available (Hogarth, 1980).

Evaluation of consequences

Cognitive limits, in this case lack of knowledge and limits of ability to forecast the future, also play a central role in the evaluation of alternatives. These cognitive difficulties are seen clearly in decisions that are taken on a national scale: whether to go ahead with the construction of a supersonic transport; the measures to be taken to deal with acid rain; Federal Reserve policies on interest rates; and, of course, the supremely fateful decisions of war and peace.

The cognitive limits are not simply limits on specific information. They are almost always also limits on the adequacy of the scientific theories that can be used to predict the relevant phenomena. For example, available theories of atmospheric chemistry and meteorology leave very wide bands of uncertainty in estimating the environmental or health consequences of given quantities and distributions of air pollutants. Similarly, the accuracy of predictions of the economy by computer models is severely limited

by lack of knowledge about fundamental economic mechanisms represented in the models' equations.

Criteria of choice

The assumption of a utility function postulates a consistency of human choice that is not always evidenced in reality. The assumption of maximization may also place a heavy (often unbearable) computational burden on the decision maker. A theory of bounded rationality seeks to identify, in theory and in actual behaviour, procedures for choosing that are computationally simpler, and that can account for observed inconsistencies in human choice patterns.

Substantive and procedural rationality

Theories of bounded rationality, then, are theories of decision making and choice that assume that the decision maker wishes to attain goals, and uses his or her mind as well as possible to that end; but theories that take into account in describing the decision process the actual capacities of the human mind.
The standard SEU theory is presumably not intended as an account of the process that human beings use to make a decision. Rather, it is an apparatus for predicting choice, assuming it to be an objectively optimal response to the situation presented. Its claim is that people choose as if they were maximizing subjective expected utility. And a strong *a priori* case can be made for the SEU theory when the decision making takes place in situations so transparent that the optimum can be reasonably approximated by an ordinary human mind.
Theories of bounded rationality are more ambitious, in trying to capture the actual process of decision as well as the substance of the final decision itself. A veridical theory of this kind can only be erected on the basis of empirical knowledge of the capabilities and limitations of the human mind; that is to say, on the basis of psychological research.
The distinction between substantive theories of rationality (like the SEU theory) and behavioural theories is closely analogous to a distinction that has been made in linguistics between theories of linguistic competence and theories of linguistic performance. A theory of competence would characterize the grammar of a language in terms of a system of rules without claiming that persons who speak the language grammatically do so by applying these rules. Performance theories seek to capture the actual processes of speech production and understanding.
The question of the desirability and usefulness of a procedural theory of decision involves at least two separate issues. First, which kind of theory, substantive or procedural, can better predict and explain what decisions are actually reached. Does SEU theory predict, to the desired degree of accuracy, the market decisions of consumers and businessmen, or does such prediction require us to take into account the cognitive limits of the economic actors?
Second, are we interested only in the decisions that are reached, or is the human decision making process itself one of the objects of our scientific curiosity? In the latter case, a substantive theory of decision cannot meet our needs; only a veridical theory of a procedural kind can satisfy our curiosity.

Bounded rationality in neoclassical economics

It should not be supposed that mainstream economic theory has been completely oblivious to human cognitive limits. In fact, some of the most important disputes in macroeconomic theory can be traced to disagreements as to just where the bounds of human rationality are located. For example, one of the two basic mechanisms that accounts for under employment and business cycles in Keynesian theory is the money illusion suffered by the labour force – a clear case of bounded rationality. In Lucas's rational expectationist theory of the cycle, the corresponding cognitive limitation is the inability of businessmen to discriminate between movements of industry prices and movements of the general price level – another variant of the money illusion. Thus the fundamental differences between these theories do not derive from different inferences drawn from the assumptions of rationality, but from different views as to where and when these assumptions cease to hold – that is, upon differences in their theories of bounded rationality.

What distinguishes contemporary theories of bounded rationality from these ad hoc and casual departures from the SEU model is that the former insist that the model of human rationality must be derived from detailed and systematic empirical study of human decision making behaviour in laboratory and real-world situations.

See Also

- rational behaviour
- satisficing

Bibliography

Cyert, R.M. and March, J.G. 1963. *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall; 2nd edn, 1975.

Hogarth, R.M. 1980. *Judgment and choice: The Psychology of Decision*. New York: Wiley.

Nelson, R.R. and Winter, S.G. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA: Harvard University Press.

Newell, A. and Simon, H.A. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Simon, H.A. 1982. *Models of Bounded Rationality*. 2 vols, Cambridge, MA: MIT Press(especially Sections VII and VIII).

Williamson, O. 1975. *Markets and Hierarchies*. New York: Free Press.

How to cite this article

Simon, Herbert A. "rationality, bounded." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_B000176> doi:10.1057/9780230226203.1391

The New Palgrave Dictionary of Economics Online

rationality, history of the concept

Esther-Mirjam Sent

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article offers a historical and methodological perspective on the concept of rationality. It gives an overview of the various interpretations of the notion, from self-interest to rational choice and expected utility to strategic rationality and rational expectations. It pays special attention to the ethical dimensions of the concept. The article further places rationality within a long-ranging discussion concerning the status of assumptions within economics. It explicitly considers efforts to test rationality directly. The article concludes with an evaluation of recent efforts to replace rationality with the notion of bounded rationality.

Keywords

adaptive control learning; adaptive expectations; Arrow's theorem; assumptions; Bayes' rule; Bayesian learning; behavioural economics; bounded rationality; Cairnes, J. E.; centipede game; chain store paradox; commitment; completeness of preferences; computation of equilibria; confirmation; error-term justification; ethics and economics; expected utility; falsificationism; folk theorem; Friedman, M.; fundamental theorems of welfare economics; game theory; Hutchison, T.; instrumental rationality; Introspection; Keynes, J. N.; learning; Machlup, F.; Microfoundations; multiple equilibria; Muth, J. F.; Nash equilibrium; Nash program; neoclassical economics; neural network learning; new behavioural economics; no-trade theorems; perfect equilibrium; pluralism in economics; power; preference satisfaction; preferences; Prisoner's Dilemma; procedural and substantive rationality; rational behaviour; rational expectations; rationality; regret; repeated games; Robbins, L. C.; satisficing; self-interest; Senior, N. W.; Simon, H.; social norms; strategic behaviour; subjective probability; sure-thing axioms; sympathy; testing; transitivity; verification; well-being

Article

Economics has always relied on some notion of rationality. Unlike philosophers, economists are not concerned with the rationality of beliefs, which are taken as data (Tisdell, 1975). In the 18th century, economics was integrated into the great scheme of the natural law and a rationalistic world view (Daston, 1983; Weber, 1904). During this time, the moral sciences aimed to reveal the rational grounds for action and belief. Overall, their focus was individualistic, psychological, and prescriptive. During the 19th century, a transition took place from a psychological framework to a sociological one. At the same time, the search for inexorable social laws replaced the computation of rational self-interest. However, economics continued to cling to rationality. Throughout much of the 20th century, many economists would separate economics from sociology upon the basis of rational or irrational behaviour (Samuelson, 1947).

Rationality is usually combined with a variety of other concepts (Arrow, 1987; Sen, 1987). Indeed, the force of the hypothesis comes from the addition of supplementary hypotheses. What has changed over time, then, is the interpretation of rationality. While it was initially associated with self-interest, in later readings, such as rational choice and expected utility, it became linked with ideas such as consistency and indifference. Recent appeals to it include strategic aspects of behaviour. Within macroeconomics, rational expectations economists have taken rationality to its extreme. Interpretations of rationality cover a wide range that includes it having the status of axiom, a priori truth, self-evident proposition, useful fiction, utopia, ideal type, analytical construct, heuristic construct, indisputable fact of experience, and typical behavioural pattern under capitalism.

Rationality is ubiquitous in modern economics, with the result that economists frequently make the assumption that it has the same meaning in all the contexts in which it is used. However, this is not the case. The assumption of rationality may be motivated by an appeal to the notion of self-interest, with due allowance made for the fact that preferences may extend to the welfare of others, but its use in expected utility theory, the analysis of strategic behaviour, rational expectations, and so on raise issues that are sufficiently profound that the meaning of the concept of rationality is fundamentally changed. Rationality may further be interpreted as either a positive or a normative notion. Efforts to test rationality interpret the notion in a descriptive manner. That is, rationality is presumed to characterize how people actually go about the business of reasoning. In response, one may investigate the psychological mechanisms and processes that underlie the patterns of reasoning that are observed. By contrast, a normative interpretation of rationality is concerned not so much with how people actually reason as with how they should reason (Suppes, 1961). The goal is to discover rules or principles that specify what it is to reason rationally – to specify standards against which the quality of human reasoning can be measured.

In the remainder of this article we first take a closer look at historical debates concerning the overall status of the rationality assumption. We then consider methodological concerns associated with the various historical interpretations of rationality, and subsequently address efforts to test rationality. After this, we take the historical debates up to the present, where we find that more and more economists are moving away from rationality towards the notion of bounded rationality.

The rationality assumption

Since rationality is such a central notion within economics, many of the debates about the status of

assumptions within economics are related to it. In the early 19th century, John Stuart Mill (1836) argued that economics is an abstract science because it reasons from assumed premises, such as rationality. As a result, its conclusions are true in the abstract. Moreover, Mill continued, that which is true in the abstract is always true in the concrete with proper allowances. This view found support among Nassau Senior, John Elliott Cairnes, and John Neville Keynes. In a similar vein, in the early 20th century, Lionel Robbins (1935) argued that the basic postulates of economics, such as rationality, are simple and indisputable facts of experience. In the opinion of Robbins, the propositions of economic theory are deductions from a series of such self-evident postulates. These insights came under serious attack by Terence Hutchison (1956). First, Hutchison claimed, the propositions of pure theory are empty. Second, maximization and equilibrium require perfect expectations. Third, economics needs more extensive use of empirical techniques. Finally, economists can use the psychological method of a priori facts, the method of *Verstehen*, and the method of introspection only for suggesting hypotheses, in Hutchison's opinion, and not for establishing them. He concluded that the rationality postulate was treated as analytic by economists, meaning that it is a priori true yet with empirical content. Instead, he claimed that it must be synthetic, meaning that it must be stated in testable form.

Hutchison's arguments about the status of the rationality assumption found a serious critic in Fritz Machlup (1956). First, Machlup argued, rationality is a theoretical construct. Second, empirical studies judge applicability. That is, they confirm rather than offer complete verification. Third, economists need to focus on ‘realistic’ assumptions embedded in a system of interrelated hypotheses. Finally, they need a suitable replacement in case of ‘rejection’. Machlup agreed with Hutchison that testing is important. However, whereas Hutchison wanted to test all statements, Machlup restricted this to specific assumptions and low-level hypotheses. And whereas Hutchison was after verification, Machlup sought confirmation.

A contrasting perspective on the status of the rationality assumption in economics came from Milton Friedman (1953), who argued that assumptions are largely irrelevant to the validation of theories. Instead, the latter should be judged, in Friedman's opinion, almost solely in terms of their instrumental value in generating accurate predictions. He did consider other criteria besides valid and meaningful predictions, but these were subsidiary. They included logical consistency, categories with meaningful empirical counterparts, advancing a substantive hypothesis capable of being tested, and simplicity and fruitfulness. Friedman argued that the standard theory in economics is successful due to its countless applications. He further claimed that the positive record follows from the dynamics of competition over time. In his opinion, the role of assumptions such as rationality is limited. They specify the conditions of validity, but do not determine these. They offer an economical mode of describing or presenting a theory. And indirect evidence may follow if assumptions are the implications of related hypotheses. In a similar vein, Armen Alchian (1950) had argued that individuals who act in a rational fashion will be successful and ‘selected’ for survival by the economic system.

Herbert Simon (1963) endeavored to rescue interest in the rationality assumption in economics by criticizing Friedman's so-called principle of unreality. According to Simon, one cannot use the validity of the market level to support the actor level. Instead, economists need to explain the market level through the use of the actor level. In Simon's opinion, valid theories about the market level follow from empirically valid assumptions about actors together with empirically valid composition laws. He therefore suggested the so-called principle of continuity of approximation instead. This holds that, if the

conditions of the real world approximate sufficiently well the assumptions of an ideal type, the derivations from these assumptions will be approximately correct. Simon argued that the unreality of premises is not a virtue but a necessary evil – a concession to the finite computing agency of the scientist that is made tolerable by the principle of continuity of approximation.

We return to Simon's alternative when we follow the historical narrative to the present towards the end of this article. We now look at methodological concerns associated with the various historical interpretations of rationality.

Rationality as self-interest

If rationality is interpreted in terms of self-interest, one of the questions that arises concerns the status of norms (Elster, 1989). Are norms rationalizations of self-interest? No, because some norms override self-interest. And norms need to have some kind of grip to be manipulated. Are norms followed out of self-interest? No, because norms do not need external sanctions. Moreover, some sanctions are performed for other motives. Do norms exist to promote self-interest? No, because followers of norms abide by them even when it is not in their interest to do so. Do norms exist to promote common interests? No, because not all norms are Pareto-improvements. In addition, some norms that would be Pareto-improvements are not observed. Do norms exist to promote genetic fitness? No, because self-interest and fear of sanctions do not provide the full explanation for adherence to norms. And we need to study emotions, envy, honour, and conformism. Additional questions arise with the later interpretations of rationality.

Rational choice and expected utility

Within expected utility theory, rationality was associated with (*a*) subjective probability, (*b*) Bayesian learning, and (*c*) maximization of expected utility (Sugden, 1991). In this interpretation, preferences are revealed by choice, and choices are supported by reasons. Efforts have been made to develop philosophical foundations of expected utility theory by appealing to Hume's instrumental rationality. According to the latter, actions can be motivated only by desires, and no desire can be brought into existence by reason alone. That is, reason is an instrument for achieving ends that are not themselves given by reason. However, there are two problems when it comes to linking expected utility theory with Hume's instrumental rationality. First, determinacy is not implied by Hume's theory of motivation. Second, consistency also does not follow from it. That is, the axioms associated with expected utility are much stronger than instrumental rationality. First, there is no justification for the completeness of preferences presumed within expected utility theory. Evidence for this can be found in framing effects, according to which the alternative framing of information in positive or negative terms affects judgments and decisions. Second, there is no justification for transitivity and sure-thing axioms due to the restricted interpretation of consequences. Evidence for this can be found in regret theory, according to which people take anticipated regret into account when they decide, which probably makes them loss averse.

Some have criticized the rational choice interpretation of rationality for presuming economic agents to be rational fools due to the severe constraints on the nature of the models that can be admitted into analysis (Sen, 1977). On the one hand, rational choice presumes too little. This is because choice may

reflect a compromise among a variety of considerations of which personal welfare may be just one. Rational choice has further come under attack for circularity because behaviour is explained in terms of preferences defined by behaviour. And it has been criticized for having too little structure; that is, it does not consider sympathy and commitment. With sympathy, which is egoistic, concern for others directly affects one's own welfare, which can be seen as an externality that would upset some standard results. With commitment, which is non-egoistic, concern for others does not affect one's own welfare but does cause action. Such action can be seen as involving counter-preferential choice requiring reformulation of the economic models, since personal choice can not longer be equated with personal welfare. In response, it has been argued that commitment needs to be accommodated as a part of behaviour by considering meta-rankings of preference rankings to express our moral judgments.

Similar concerns arise as a result of a wide range of impossibility results, such as Arrow's impossibility theorem, according to which supra-individual entities such as societies and nations cannot be said to have well-behaved preferences of the sort attributed to individual agents in rational choice approaches, under fairly general circumstances (Arrow, 1987). This devastated the hope that statements about collectivities could have solid microfoundations in individual rationality. We will return to ethical and justice matters after taking a closer look at the appeals to rationality within game theory.

Strategic rationality

Nash equilibrium, the basis for much game theory, goes further in assuming not only rationality but also common knowledge of rationality (Sugden, 1991). That is, it presumes that there is common knowledge of the mathematical description of the game, of the rationality of the players, and of the logical or mathematical theorems. However, common knowledge of rationality is not sufficient for Nash equilibrium. In addition, it is incoherent since it requires subjective probabilities to be formed, which may not be possible. Moreover, it is circular, though it establishes internal consistency of the infinite chain of reasoning, since it cannot explain choice because the outcome is not determinate. As a result, common knowledge of rationality should be seen as an equilibrium concept, where equilibrium may not exist.

Arguments have been made that the rationality associated with Nash equilibrium is self-defeating (Sent, 2004a). That is, all kind of frictions have been encountered within the Nash program. First, the folk theorem illustrates the (very real) possibility of encountering multiple equilibria in repeated games. The folk theorem states that in infinitely repeated games, for a range of discount factors that are high enough – though less than 1 – any payoff vector that is feasible in the set of payoffs between two players who are simultaneously individually rational is a Nash equilibrium payoff. Second, intuitively unreasonable equilibria may be selected in the finitely repeated Prisoner's Dilemma game, the chain store paradox, and the centipede game. As a result, the standard game-theoretic solutions yield results that are considered quite unintuitive. Finally, under certain conditions, theorems concerning the non-existence of trade and the impossibility of ‘agreeing to disagree’ about an event have been proved for Nash equilibria. Moreover, speculative trade cannot be explained as an outcome of different information structures. One possible resolution is to disconnect Nash equilibrium and common knowledge of rationality. It has been shown that common knowledge could generate many non-Nash equilibria. Likewise, it has been shown that even with common knowledge of rationality there may be no Nash equilibrium. Overall, however,

the foundations of theories associated with rationality are not secure.

Rational expectations

In recent economics, there has been much concern with how expectations are formed. The dominant approach argues that the rules of thumb, such as adaptive expectations, which fail to use available information optimally are hard to reconcile with the idea of rationality that was the foundation of most economic analysis (Muth, 1961). Instead, it is argued that, since agents are claimed to be optimizers, it is only natural to presume that they will also form their expectations rationally. Hence, some argue that the rational expectations hypothesis is nothing but a direct application of the rationality principle to the problem of expectations of future events. In particular, optimizing over perceptions implies that agents do the best they can and form their views of the future by taking account of all available information, including their understanding of how the economy works. If perceptions were not optimally chosen, there would exist unexploited utility or profit-generating possibilities within the system. The implication is that all such unexploited possibilities must disappear. When applied to macroeconomics, this appeared to be in sharp contrast with Keynesian theories, which modelled firms and consumers in ways that were seen as being ad hoc and inconsistent with the idea of rational behaviour. The typical Keynesian assumptions that markets did not clear and that economic agents did not always pursue optimizing strategies could be criticized on similar grounds, as implying ad hoc departures from the axiom of rational behaviour. From this perspective, to adopt rational expectations is thus to replace earlier ad hoc treatments with an approach squarely based on the microfoundations of incentives, information, and optimization.

A variety of problems have arisen within rational expectations economics as a result of its rationality assumption (Sent, 1997). First, how can there be trade among economic agents who are all rational? One suggestion, following a line of research started by Robert Lucas, is that equilibrium probability beliefs differ and that agents actually trade on the basis of different information. However, a whole series of no-trade theorems overrule this common-sense intuition (Varian, 1987). The second obstacle encountered by rational expectations economists involved error-term justification. In particular, close scrutiny of the justification of error terms revealed that the econometrician needed to be outwitted by the agents (Sargent, 1981). Finally, how can policy recommendations be made when agents, economists, and governments are put on an equal footing based on rational expectations? When policy recommendations are possible, symmetry is impossible. For making recommendations for improving policy amounts to assuming that in the historical period the system was not really in a rational equilibrium. When symmetry is possible, policy recommendations are impossible. For making the assumption that in the historical period the system was in a rational equilibrium raises the question of why we study a system that we cannot influence (Sargent, 1984).

Having considered methodological concerns associated with the various historical interpretations of rationality, we now take up ethical concerns explicitly, since they bear upon rationality in general.

Ethics

Some have criticized economists for focusing narrowly on rationality while ignoring ethics (Hausman

and McPherson, 1984). They claim that ethics is relevant to economists for a variety of reasons. First, economists need to know some morality to know what questions to ask. Second, economists evaluate moral commitments while describing them. Third, economists affect what they see by how they describe it. Fourth, economists are influenced by their moral values and their attitudes towards the values of the agents they study. Hence, ethics is part and parcel of economics, even when economists fail to acknowledge this in their focus on rationality. An additional concern is that they oftentimes mistakenly identify well-being with preference satisfaction. This is problematic for a variety of reasons. First, what people prefer may not be good for them. Second, people make mistakes. Third, people may prefer to sacrifice their own well-being in pursuit of some other end. As a result, appraisals of economic institutions and outcomes must consider moral concerns, such as freedom, rights, and justice, besides rationality.

Rationality, which is central to the notion of a competitive equilibrium, is a key element in the two fundamental theorems of welfare economics. The first states that any competitive equilibrium leads to an efficient allocation of resources. The second asserts the converse, that any efficient allocation can be sustainable by a competitive equilibrium. The first theorem appears to make a case for non-intervention: let the markets do the work and the outcome will be desirable. The second theorem states that out of all possible efficient outcomes (of which there may be many) one can achieve any particular efficient outcome by enacting a lump-sum wealth redistribution and then letting the market take over. It has been argued that perfectly competitive markets with individual factor endowments and private goods, free market activity and mutual concern, and the absence of rationality are morally free zones (Gauthier, 1991). Because of the free activity, there is liberty. Due to the absence of externalities, there is impartiality. And as a result of the first welfare theorem, there is optimality. This morally free zone, the argument goes, arises in a deeper moral framework, according to which moral constraint is compatible with mutual unconcern and rationally required. In addition, morality as a system of rationally required constraints makes possible the realization of one's interests and the fulfilment of one's preferences. This perspective has been criticized on four accounts. First, there are market failures. Second, the initial distribution is relevant. Third, there may be multiple equilibria. Fourth, no account of social policy is given. Instead, it has been claimed that markets are political, cultural and economic. First, they support a well-defined structure of power. Second, they shape our culture. Third, they foster or thwart desirable forms of human development. Fourth, they allocate resources and distribute income. Since rationality is only one element in these arguments, we shall not dwell on these concerns further.

Having considered ethical concerns explicitly, we now turn to efforts to test rationality, since they also bear upon rationality in general.

Testing rationality

As cautioned in the introduction of this article, efforts to test rationality interpret the notion in a descriptive manner. With this in mind, much energy has also been put into trying to test rationality directly (Blaug, 1992). According to some, this involves stating the problem situation, testing the predictions, assessing the evidence, considering the nature of explanation, considering alternative theories, and stating the hard core and heuristics. However, these steps are more difficult than appears at first sight. As philosophers of science have argued, immunizing stratagems are sometimes defensible,

verisimilitude is difficult to implement, and there is no metric of corroboration. Economics poses additional problems, because narrow falsificationism is too restrictive and broad falsificationism has no prescriptive force. Moreover, economics is characterized by many initial conditions and no general laws. Testing its models is not the same as testing theories. Its data do not correspond to the concepts. Finally, falsificationism is hardly ever practiced in economics. Indeed, in response to these problems with the Popperian position, Popper himself accorded a special status to the rationality principle within his situational logic as a ‘zero principle’. This situational logic is used to explain actions and events in social science. According to Popper, the rationality principle is an integral part of every, or nearly every, testable social theory. At the same time, he believed there to be good reasons for the rationality principle to be false, while a good approximation to truth. Still, he felt that social scientists should retain it despite the fact that it is false. This is because he believed that we learn more if we blame our situational model. Indeed, he saw the policy of upholding the rationality principle as part of our methodology.

After this methodological ‘detour’, we now return to the historical narrative by evaluating recent efforts to replace rationality with the notion of bounded rationality.

Bounded rationality

Recently, especially game theorists and rational expectations economists have embraced the notion of Bounded rationality. Game theorists have looked towards Bounded rationality in their efforts to save the rationality of the Nash equilibrium. This was needed because of frictions within the Nash program (Aumann, 1997; Rubinstein, 1998). First, the folk theorem illustrates the (very real) possibility of encountering multiple equilibria in repeated games. Second, intuitively unreasonable equilibria may be selected in the finitely repeated Prisoner's Dilemma game, the chain store paradox, and the centipede game. Finally, under certain conditions, theorems concerning the non-existence of trade and the impossibility of ‘agreeing to disagree’ about an event have been proved for Nash equilibria. It could be argued that one response of game theorists to these problems has been to incorporate Bounded rationality (Sent, 2004a). First, Bounded rationality functioned as a dynamic for selection among multiple equilibria by promising to ‘refine’ equilibria. Moreover, the evolutionary stable strategy concept of evolutionary game theory may be viewed as a further refinement of perfect equilibrium, one of the most common notions used to refine the Nash equilibrium. Second, Bounded rationality has been used to rule out unintuitive equilibria in the Prisoner's Dilemma game, the chain store paradox, and the centipede game. Third, absence of a fully rational treatment of knowledge may circumvent the no-trade theorems by allowing speculative trade. These attempts to strengthen Nash, then, lead to the paradoxical observation that rationality in games depends critically on irrationality.

Likewise, rational expectations economists have sought to reinforce the rational expectations hypothesis by focusing on convergence to this equilibrium through boundedly rational ‘learning’. They have also used Bounded rationality to deal with some of the problems associated with rational expectations such as multiple equilibria and the computation of equilibria (Sargent, 1993).

Economists have tried to capture Bounded rationality by replacing rational players with computing devices such as Turing machines, finite automata, or neural network algorithms. Players' rationality is bounded in the sense that they cannot consider strategies other than those that can be played by these computing devices. Rationality is bounded in Turing machines because these machines will sometimes compute for ever to give correct answers. In order to come up with a solution, the output follows an

arbitrary guessing rule after the machine has been stopped. Bounded rationality in finite automata is captured by imposing constraints on the number of states of the automata or assuming that states are costly. Neural networks, finally, increase the computational capability of a finite automaton by increasing the states of the machine.

Both game theorists and macroeconomists have developed models of boundedly rational learning (for example, Bray and Kreps, 1987). The basic idea in these models is that boundedly rational agents utilize one of three procedures for making and changing their choices on the basis of past outcomes. First, Bayesian learning assumes that players update their subjective probabilities in the face of inconsistencies through the use of Bayes' rule until consistency is achieved. This technique has been used in the problem of equilibrium selection. It has been criticized for not accurately representing ‘true’ learning. Second, least squares or adaptive control learning assumes that players use standard statistical or econometric procedures for estimation. These procedures are boundedly rational in that economic agents use models that are misspecified and forecasting procedures that are not part of the optimal decision making of these individuals. This approach has been used in the context of rational expectations models. It has been criticized for requiring the agents to still be quite smart. Third, neural network learning assumes that individuals construct explicitly approximate models of their environment which are updated as their information improves. In contrast to least squares or adaptive control learning, agents here know they hold misspecified models of reality. This technique has been used to explore how boundedly rational players can achieve consistent beliefs and, possibly, a Nash equilibrium.

Behavioural economics

Developments within game theory and rational expectations economics combined with the rise of behavioural economics suggest that economics as a whole is moving away from rationality and towards Bounded rationality (Sent, 2004b). In the 1960s, appeals to Bounded rationality on the part of behavioural economists were designed to develop an alternative to the mainstream model (for example, Simon, 1955). We could label these endeavours as old behavioural economics. At that time, few economists exhibited any interest in these efforts. In the 1970s, cognitive psychologists suggested ways to incorporate behavioural insights in ways that provided less of a threat to the standard model (for example, Kahneman and Tversky, 1974). We could label these efforts as new behavioural economics. At the same time, the mathematical foundations of the mainstream started showing some flaws. In the 1980s, disagreements emerged between old and new behavioural economists with the latter emerging as the victors in the 1990s, partly because Herbert Simon abandoned his efforts and partly because new behavioural economists suggested ways in which their insights may help rebuild the mainstream stronghold.

Bounded rationality is not a field in itself, but rather an approach to doing economic research (Simon, 1976). There are many interpretations of Bounded rationality and these are not always consistent. Herbert Simon, the father of Bounded rationality, used the term ‘Bounded rationality’ to highlight limitations of both knowledge and computational capacity. These bounds affect human cognitive capacity for discovering alternatives, computing their consequences under certainty and uncertainty, and making comparisons among them. Theories of Bounded rationality, then, are generated by analysing processes for generating alternatives, procedures such as heuristics for evaluating consequences, and

strategies such as satisficing for making choices. Decision making is characterized as a selective search in which heuristics are used to determine what paths should be taken, and the search halts when a satisfactory solution has been found. In this process, aspiration levels are adapted in response to success or failure.

To distinguish behavioural economics from neoclassical economics, Simon introduced a distinction between procedural and substantive rationality (Simon, 1976). He argued that psychologists have considered the former concept while economists have focused on the latter. Whereas the former involves decision makers following specific rules or procedures, the latter has decision makers consider their own total preference ordering. That is, procedural rationality is about the rationality of the procedure used to reach a decision, while substantive rationality is about the rationality of the decision itself. While procedural rationality is interested in how individuals make decisions, substantive rationality focuses attention on why they do so. That is, the former focuses on the methods individuals employ, whereas the latter considers the outcomes that follow. On the one hand, procedural rationality helps decision makers decide how to get there, but not where to go. On the other hand, substantive rationality tells them where to go, but not how to get there. Simon used the concept of Bounded rationality to explain why substantive rationality is often inappropriate as well as impossible.

While models of Bounded rationality have not always appeared as attractive as the axiomatized certainties of neoclassical economics, more and more economists are embracing one form or another of Bounded rationality. Models of Bounded rationality owe their revival partly to attempts to develop a viable alternative to neoclassical economics and partly to attempts to strengthen neoclassical economics. Whereas the first reason is in the spirit of Simon's contributions, Simon certainly would have opposed the second.

Conclusion

In adopting Bounded rationality, rational expectations economists and game theorists found themselves in the paradoxical position of using Bounded rationality to define rationality. In addition, they were confronted with the question how much Bounded rationality is admitted and how it is clarified. In fact, the dependence of the definition of rationality on irrationality is reminiscent of debates in philosophy concerning the definition of concepts in terms of their opposites, which has led to efforts to destabilize dichotomies. In addition, there is the practical challenge of overcoming the Bounded rationality of economists in modelling the Bounded rationality of economic agents. Ironically, when economists made the agents in their models more bounded in their rationality, they had to be smarter because these models became larger and more demanding econometrically. Bounded rationality researchers face innumerable decisions about how to represent decision-making processes and the ways that they are updated, which requires a large amount of rationality on their part.

Overall, the efforts of rational expectations economists and game theorists to embrace one form or another of Bounded rationality have more to do with intellectual puzzles than empirical anomalies (Aumann, 1997; Rubinstein, 1998; Sargent, 1993). Hence, in effect, economists have replaced one set of puzzles, concerning, for instance, the non-existence of trade, with another paradox, concerning the dependence of rationality on irrationality (Aumann and Sorin, 1989).

The move from rationality to Bounded rationality is part of the observation that the present situation in economics could be characterized as one of moderate pluralism. That is, recent years have witnessed not

only efforts to incorporate Bounded rationality approaches and behavioural insights, but also chaos theory, complexity approaches, evolutionary insights, experimental methods, and neuroeconomics. Therefore, the benchmark from which new behavioural economics considers deviations may itself be evolving – none of which alone or in combination appears yet to have established a new orthodoxy, leading to debate over the direction and future content of economics. Elaborating these, however, interesting, would take us beyond the scope and word limit of this article.

See Also

- behavioural economics and game theory
- ethics and economics
- rational expectations
- rationality, bounded
- Simon, Herbert A.

Bibliography

Alchian, A.A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58, 211–22.

Arrow, K.J. 1987. Economic theory and the hypothesis of rationality. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. New York: W.W. Norton.

Aumann, R.J. 1997. Rationality and bounded rationality. *Games and Economic Behavior* 21, 2–14.

Aumann, R.J. and Sorin, S. 1989. Cooperation and bounded recall. *Games and Economic Behavior* 1, 5–39.

Becker, G.S. 1962. Irrational behavior and economic theory. *Journal of Political Economy* 70, 1–13.

Blaug, M. 1992. *The Methodology of Economics: Or, How Economists Explain*. Cambridge: Cambridge University Press.

Bowles, S. 1991. What markets can – and cannot – do. *Challenge* 34(4), 11–16.

Bray, M. and Kreps, D. 1987. Rational learning and rational expectations. In *Arrow and the Ascent of Modern Economic Theory*, ed. G. Feiwel. New York: New York University Press.

Brunswik, E. 1955. In defense of probabilistic functionalism: a reply. *Psychological Review* 62, 236–42.

Caldwell, B.J. 1994. *Beyond Positivism: Economic Methodology in the Twentieth Century*. London: Routledge.

Daston, L. 1983. Rational individuals versus laws of society. In *Probability since 1800*, ed. M. Heidelberger, L. Krüger and R. Rheinwald. Bielefeld: Universität Bielefeld.

Elster, J. 1989. Social norms and economic theory. *Journal of Economic Perspectives* 3(4), 99–117.

Friedman, M. 1953. The methodology of positive economics. Reprinted in Hausman (1984).

Gauthier, D. 1991. *Morals by Agreement*. New York: Cambridge University Press.

Gigerenzer, G. and Todd, P.M. 1999. Fast and frugal heuristics: the adaptive toolbox. In *Simple Heuristics That Make Us Smart*, ed. G. Gigerenzer, P.M. Todd and A R. Group. New York: Oxford University Press.

Hausman, D. (ed.) 1984. *The Philosophy of Economics*. Cambridge: Cambridge University Press.

Hausman, D. and McPherson, M. 1984. Economics, rationality, and ethics. In Hausman (1984).

Hutchison, T.W. 1956. On verification in economics. Reprinted in Hausman (1984).

Kahneman, D. and Tversky, A. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–31.

Kirzner, I.M. 1962. Rational action and economic theory. *Journal of Political Economy* 70, 380–85.

Knight, F.H. 1925. Fact and metaphysics in economic psychology. *American Economic Review* 15, 247–66.

Machlup, F. 1956. On indirect verification. Reprinted in Hausman (1984).

Mill, J.S. 1836. On the definition and method of political economy. Reprinted in Hausman (1984).

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–55.

Neumann, J. von. and Morgenstern, O. 1944, 1955. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Peterson, C.R. and Beach, L.R. 1967. Man as an intuitive statistician. *Psychological Bulletin* 68, 29–46.

Robbins, L. 1935. The nature and significance of economic science. Reprinted in Hausman (1984).

Rubinstein, A. 1998. *Modeling Bounded Rationality*. Cambridge, MA: MIT Press.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sargent, T.J. 1981. Interpreting economic time series. *Journal of Political Economy* 89, 213–48.

Sargent, T.J. 1984. Autoregressions, expectations, and advice (with discussion). *American Economic Review* 74, 408–21.

Sargent, T.J. 1993. *Bounded Rationality in Macroeconomics*. Oxford: Oxford University Press.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: John Wiley & Sons.

Sen, A. 1977. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6, 317–44.

Sen, A. 1987. Rational behavior. In *The New Palgrave: A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. New York: W.W. Norton.

Sent, E.M. 1997. *The Evolving Rationality of Rational Expectations: An Assessment of Thomas Sargent's Achievements*. Cambridge: Cambridge University Press.

Sent, E.M. 2004a. The legacy of Herbert Simon in game theory. *Journal of Economic Behavior and Organization* 53, 303–17.

Sent, E.M. 2004b. Behavioral economics: how psychology made its (limited) way back into economics. *History of Political Economy* 36, 735–60.

Simon, H.A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69, 99–118.

Simon, H.A. 1976. From substantive to procedural rationality. In *Method and Appraisal in Economics*, ed. S. Latsis. Cambridge: Cambridge University Press.

Simon, H.A. 1963. Testability and approximation. Reprinted in Hausman (1984).

Smith, V. 1991. Rational choice. *Journal of Political Economy* 99, 877–897.

Smith, V. 1992. Game theory and experimental economics. In *Toward a History of Game Theory*, ed. E. R. Weintraub. Durham, NC: Duke University Press.

Sugden, R. 1991. Rational choice: a survey. *Economic Journal* 101, 751–83.

Suppes, P. 1961. The philosophical relevance of decision theory. *Journal of Philosophy* 58, 605–14.

Tisdell, C. 1975. Concepts of rationality in economics. *Philosophy of Social Science* 5, 259–72.

Tversky, A. and Kahneman, D. 1971. Belief in the law of small numbers. *Psychological Bulletin* 76, 105–10.

Varian, H.R. 1987. Differences of opinion in financial markets. Financial Risk: Theory, Evidence, and Implications, Proceedings of the 11th Annual Economic Policy Conference of the Federal Reserve Bank of St. Louis.

Weber, M. 1904. The area of economics, economic theory, and the ideal type. In *Essays in Economic Sociology*, ed. M. Weber. Princeton: Princeton University Press, 1999.

Winter, S.G. 1964. Economic ‘natural selection’ and the theory of the firm. *Yale Economic Essays* 4, 225–72.

How to cite this article

Sent, Esther-Mirjam. "rationality, history of the concept." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000256> doi:10.1057/9780230226203.1392

The New Palgrave Dictionary of Economics Online

rationality

Lawrence E. Blume and David Easley

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic theory takes the individual consumer and firm as a primitive unit of analysis, and so a theory of individual agency is required to derive hypotheses about the behaviour of markets and other systems of economic interest. One such theory is the principle of rationality, whereby agents act in their perceived best interest. This article surveys the implementation of this principle in economic models, and discusses the critiques of the rationality principle and some proposed alternatives from the perspective of the economic modeller.

Keywords

altruism; behavioural economics; Bentham, J.; bounded rationality; choice under uncertainty; cognitive models; consistency; cooperation; decision theory; economic laws; economic man; evolutionary models; expected utility; general choice theory; hyperplanes; interpersonal utility comparisons; marginal utility; mechanisms; menu choice models; methodological individualism; neuroeconomics; non-cooperative game theory; non-expected utility decision theory; preference formation; preference relation; prospect theory; rational choice; rational expectations; rational expectations equilibrium; rationality; rationality principle; reciprocity; revealed preference theory; Savage's subjective expected utility model; self-interest; social norms; social preferences; statistical decision theory; transitivity; trust; utilitarianism; utility

Article

> I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, …
> (Justice Potter Stewart, 378 U.S. 184, 197)

Rationality is for economists as pornography was to the US Supreme Court, undefinable but nonetheless easily identified; and yet, like the Justices of the Court, no two economists share a common definition. This article details some of the common meanings of individual (as opposed to social) rationality and discusses their uses. Our point of view is that of working economists rather than that of psychologists. Economics is committed to methodological individualism, the claim that social phenomena must be explained in terms of individual actions which in turn must be explained through individuals' motivations. This commitment requires a theory of human action. The *rationality principle*, that individuals act in their best interest as they perceive it, provides such a theory. In this article we evaluate the rationality hypothesis and its alternatives from the perspective of how they explain social phenomena such as the behaviour of a market. Our interest is in social life rather than in the psychology of an individual.

History and description

The use of the rationality principle in economics certainly predates the utilitarianism with which it is so often conflated. Adam Smith (1789, p. 19) describes, in his discussion of the division of labour, a tribe of hunters in which one person is particularly deft at making bows and arrows. ‘He frequently exchanges them for cattle or for venison with his companions; and he finds at last that he can in this manner get more cattle and venison, than if he himself went to the field to catch them. From a regard to his own interest, therefore, the making of bows and arrows grows to be his chief business …’ Moving from intuition to analysis, however, requires a sharp understanding of what it means to regard one's own interest, and this has become a source of endless debate for rational-actor social scientists.
The utility-maximization version of rationality springs from the utilitarianism of Bentham and Mill. According to Bentham (1789, p. i),

> Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne.

Although many thinkers toyed with utilitarian approaches to economic analysis, it was not until the 1870s, through the work of Jevons, Menger and Walras, that utility maximization began to assume the important role in economic analysis it has since held. For this trio, utility was a short cut to a theory of value. Perhaps this is why they were not overly concerned with the issues of measurable utility and the possibility of interpersonal utility comparisons which so exercised their successors. Utility for Bentham, on the other hand, was a physical measure of pain and pleasure which could be computed according to his ‘felicific calculus’. Although utility as a merely hedonic measure was rejected even by Mill, only in the 1930s, and after a half century's work beginning with Fisher (1892) and Pareto (1895) was it generally recognized that properties of demand derived from the shape of indifference curves, and so utility could admit a purely ordinal interpretation. This ‘shift in emphasis away from the physiological and psychological hedonistic, introspective aspects of utility’, as Samuelson (1947, p. 90–1) put it, led to the ‘purging out of objectionable, and sometimes unnecessary, connotations … of the Bentham …

variety'. The ultimate expression of this non-psychological view is the theory of revealed preference, whose purpose is '…to develop the theory of consumer's behavior freed from any vestigial traces of the utility concept' (Samuelson, 1938a, p. 71). The result is a mathematical structure that Edgeworth would have understood, interpreted in a manner completely foreign to his way of thinking.

Expected utility in the theory of choice under uncertainty is older than Benthamite utilitarianism. Both an expectation argument and a dominance (admissibility) argument for the existence of God were carefully laid out by Pascal (1672, p. 233). These remarkable few paragraphs touch on many important issues in contemporary decision theory, including the principle of insufficient reason, the problem of infinite utility payoffs, and incomplete preferences: 'Yes; but you must wager. It is not optional. You are embarked. Which will you choose then?' Even the concept of marginal utility predates Bentham, in Gabriel Cramer's and Daniel Bernoulli's famous near-resolutions of the St. Petersburg paradox. But despite this early progress, the formalization of the modern theory of choice under uncertainty begins only with Wald (1939), who at one go describes the key structures of statistical decision theory: loss functions, a priori distributions, and Bayes, admissible, and minimax decision rules. Interest quickly coalesced, however, around the expected utility models described in the two great testaments of decision science, von Neumann and Morgenstern (1947) and Savage (1956). Expected utility (EU) quickly became such a dominant paradigm for choice under uncertainty that research into alternatives was a backwater for 20 years. But criticisms of the expected utility models emerged almost before the ink was dry on the two manuscripts, in Allais' (1953) experiments and Cyert, Simon and Trow's (1956) empirical studies of firm behaviour, and by the late 1970s behavioural economics and non-EU decision theory were active areas of research.

Psychological utilitarianism and decision theory are the two traditions which most inform the modern economist's thinking about 'rationality', and yet, despite the long intellectual history of these ideas, no single vision of what it means to be a 'rational actor' has emerged. In the remainder of this article we single out several sources of confusion and disagreement. We discuss five models of rationality.

General choice theory (GCT)

A set *A* of alternatives is given, along with a collection $\mathcal{B}$ of non-empty subsets of *A*. The set *A* is the set of possible alternatives and any member *B* of $\mathcal{B}$ is a set of feasible alternatives, a set from which the decision-maker must choose. A *choice function C* assigns to each $B \in \mathcal{B}$ a nonempty subset of $\mathcal{B}$, the objects chosen by the decision-maker from the feasible set. In the theory of demand, for instance, *A* is the consumption set, *B* is the collection of possible budget sets and the choice function is the demand function. A textbook treatment of the rational decision-maker requires that she have a *preference relation* $\succsim$ on *A*, and we understand $a \succsim b$ to mean that she finds *a* to be 'at least as good as' *b*. By 'preference relation' we mean a binary relation which is complete, all alternatives can be compared, and transitive. Transitivity means that if *a* is at least as good as *b* and *b* is at least as good as *c*, then *a* is at least as good as *c*. Chosen objects in a set *B* of feasible alternatives are those maximal with respect to the preference relation; *b* is chosen from *B*, that is, $b \in C(B)$, if and only if $b \succsim a$ for all $a \in B$. Preference is the primitive expression of rationality. The role of utility is to provide a convenient *representation* of preference. A utility function u is a real-valued function on *A*, and to say that *u represents* $\succsim$ means that $u(a) \geq u(b)$ if and only if $a \succsim b$. While the decision theory toolkit of the working economist mostly

specializes this model, much contemporary economic theory does not require this much of rationality. In particular, the completeness and transitivity assumptions can be done away with in general equilibrium theory, and numerical representations for incomplete and non-transitive preferences are available. See, for instance, Aumann (1962), Chipman et al. (1971), and Gale and Mas-Colell (1975).

Expected utility theory (EU)

Expected utility is a specialization of GCT in which the set *A* and the preference relation have a specific structure. In EU theory, *X* is a finite set of prizes or outcomes, and the alternative set *A* is the set of all probability distributions on *X*. Preferences have the following representation: A *payoff function v* is a real-valued function on *X*, and any two probability distributions *p* and *q* in *A* are compared according to their expected values of *v*; that is $p \succsim q$ iff $E_p v \geq E_q v$. The content of this theory is that, geometrically speaking, indifference curves are parallel straight lines (hyperplanes). The first characterization of EU preferences was provided by von Neumann and Morgenstern (1947); today's standard axiomatic characterization of EU preference orders is due to Herstein and Milnor (1953).

Subjective expected utility theory (SEU)

When we choose whether to play roulette or a slot machine, we are choosing among probability distributions. When we bet on the outcome of a horse or political race, we are betting on the realization of uncertain outcomes, but not objects to which probabilities are necessarily attached. Savage's (1956) contribution was to provide a theory of what he called ‘personal probability’, a specialization of GCT, here interpreted as a decision-maker's degree of belief in the occurrence of some event. He characterized those preference relations which could be represented by the expectation of some payoff function with respect to a personal probability. In Savage's subjective expected utility (SEU) theory, *S* is a set of states, such as the possible outcomes of the election. There is also a set *X* of outcomes. A bet on the election is a function which assigns an outcome to every state. Savage called such functions $f: S \to X$ *acts*, and the set of acts is the alternative set *A*. A preference relation $\succsim$ on the set *A* has an SEU representation if there is a payoff function *v* on outcomes *X* and a probability distribution *p* on states *S* such that $f \succsim g$ if and only if $E_p\{v(f(s))\} \geq E_p\{v(g(s))\}$.
Methodological individualism requires the analysis of social phenomena to be ‘bottom-up’, that is, to begin with individuals. It is a stronger statement to claim, however, that the description of the individual is entirely pre-social; that in economic models, for instance, that individuals come to the market with preferences and beliefs already formed. Most modern economists do not make this claim, and instead work with models in which the description of the individual is an equilibrium outcome. The two most prominent examples of this method are rational expectations equilibrium and non-cooperative game theory.

Rational expectations equilibrium (REE)

The rational expectations hypothesis supposes a population of individuals solving decision problems which have a common state space, and furthermore that the state will be chosen according to the ‘true

distribution' μ, which is determined by the individuals' choices. The payoff $v(c,s)$ to a choice c depends on the state realization s, and preferences over choices are EU: $U_i(c) = E_\mu v(c, s)$. The hypothesis asserts that all beliefs will be correct; that is, that all SEU decision-makers have preference representations in which the beliefs are in fact the probability distribution μ, and μ in turn is the distribution of states which is determined by their actions. Rational expectations is a misuse of the adjective. Unfortunately it is probably too late to abandon the term. There is no connection between the rationality principle, which claims that individuals act in their perceived best interest, and the rational expectations hypothesis, which claims that those perceptions meet some *ex ante* standard of correctness. But so labelling a theory is certainly a nice rhetorical move for how it structures subsequent debate.

Non-cooperative game theory (NGT)

A population of individuals chooses actions. Individual i's payoff to action c_i, $v(c_i, c_{-i})$ depends upon the choices c_{-i} of the others. He holds probabilistic beliefs about the actions of others, and evaluates a choice according to EU. The social construction of the individual is seen in the determination of beliefs. *Undominated strategies* are those which can be rationalized by *some* choices of beliefs. *Rationalizable strategies* are those which can be justified by some beliefs satisfying a belief restriction, that it be *common knowledge* that all members of the population are EU-rational with some beliefs, and that payoffs be common knowledge (see epistemic game theory: an overview). Nash equilibrium requires, like REE, that everyone's beliefs are correct. Various Nash equilibrium refinements also have belief interpretations (see Nash equilibrium, refinements of).

Rationality and mind

The merits of the rational choice foundation of economics have been much discussed, both by its practitioners and by its critics. This discussion is often confused, in part because economists are not consistent in how they understand the contents of the rationality hypothesis. Economic theory holds two views of rationality. One is that rationality is consistency of choice, that the tools of choice theory are just an alternative encoding of certain choice functions; the other is that rationality is a theory of intentional behaviour, in which beliefs and desires are meaningful constructs.

Revealed preference theory is the sharpest formulation of the consistency view. It takes demand as primitive and asks if it is *consistent* with the maximization of a preference order. It recovers desires from choice, and only to the extent that choices are different can two desires, preference orders, be distinguished. This view permeates the foundations of decision theory. For Savage (1956, p. 17), 'It is possible that the person **prefers f** to **g**. Loosely speaking, this means that, if he were required to decide between **f** and **g**, no other acts being available, he would decide on **f**'. In this account, preference is defined by choice. This means specifically that if the choice function C on a collection $\mathcal{B}$ of choice sets satisfies certain conditions, then there is a complete and transitive binary relation such that for every $B \in \mathcal{B}$, $C(B)$ contains exactly the elements of B which are maximal in B with respect to the relation. The binary relation is nothing more than an alternative description of C on $\mathcal{B}$. Suppose a new choice set $B' \notin \mathcal{B}$ is considered. What can we guess about the contents of $C(B)$? Knowing that the decision-maker is consistent on B allows the observer to infer nothing at all about $C(B')$.

If revealed preference represents at all a psychology of choice, that psychology is a form of *radical behaviourism*. Radical behaviourism asserts that two mental states are distinguishable only to the extent that some observable behaviour distinguishes them. Behaviours are all that one can theorize about. Samuelson (1938b, p. 344) writes ‘of a steady tendency toward the removal of moral, utilitarian, welfare connotations …’ and of ‘the rejection of hedonistic, introspective, psychological elements’. Although the behaviourist position seems extreme, the leading graduate microeconomics textbook writes approvingly of revealed preference: ‘Perhaps most importantly, it makes clear that the theory of individual decision making need not be based on a process of introspection but can be given an entirely behavioral foundation’ (Mas-Colell, Whinston and Green, 1995, p. 5). Consistency is often justified as discipline. It requires a minimum of assumptions about the beliefs and desires of individuals, and minimizes the possibility of researchers' values and beliefs slipping unbidden into their analyses. It allows the data maximal scope to speak for itself.

Although received economics talks approvingly of rationality as mere consistency, this is not in fact what most economists do. Much of economics involves invisible-hand explanations; aggregate market behaviour emerges from the decisions of many agents. Whether the invisible hand lifts the cup aloft or knocks it over, economic explanation entails explaining how it coordinates for good or ill the motives and interests of diverse individual actors. These kinds of question call for explanations based on the motivations of economic actors, which purely behaviouralist explanations cannot provide. So economists in practice take an intentional view.

The intentional view holds that rational choice theory is a common-sense or ‘folk’ (as opposed to ‘scientific’) psychology. Just as in our everyday transactions we use the language of beliefs and desires to interpret and forecast the behaviour of others, so do economists interpret choice behaviour. The investor *believes* that the asset price will be higher tomorrow. She *wants* greater wealth tomorrow. So she *acts* by purchasing the asset. In this view belief and desire are in fact mental states that are connected to action. The folk psychology is a theory of mind which is presumed by economists to be both adequate for a descriptive psychology of decision and accurate enough in its predictions of individual behaviours for the uses to which it is put. Although utility does not exist as a psychophysical quantity, rational choice models provide a representation of the mental states involved in judgment and decision. (The stronger claim that mental process is a more or less efficient utility maximization algorithm is a view held only by the straw man regularly beaten up by rationality's critics.)

The economist's folk psychology goes further than everyday folk psychology by specifying analytic representations of beliefs, desires, and how they interact. No matter what representation is ultimately chosen by the textbook economist, his folk psychology rests on two points. (1) Rationality is instrumental. Its concern is the efficient pursuing of ends by available means, not the sensibility of the ends. (2) Desire is not anchored by any other aspect of the decision problem, whether the feasible set or the context of choice. Formally, desires are captured by a preference ordering on possible objects of choice whose existence is independent of the feasible set or the context of choice. This is the content of GCT.

The tension between the demands for a parsimonious behavioural theory and the need for an intentional theory of choice is often resolved by holding that, of course beliefs and desires exist, but we economists have access to them only as they are revealed in observed choice behaviour. In a recent critique of neuroeconomics, two well-known theorists write, ‘In standard economics, the testable implications of a theory are its content; once they are identified, the non-choice evidence that motivated a novel theory

becomes irrelevant' (Gul and Pesendorfer, 2005, p. 6). This view has a long history, perhaps with origins in the defence of marginal analysis against its early critics. Machlup (1946, p. 537) writes, 'Psychologists will readily confirm that statements by interviewed individuals about the motives and reasons for their actions are unreliable or at least incomplete', and also raises the oft-heard incentive problem of eliciting survey data, namely, that survey respondents may choose answers to meet their own goals, which may not include truth.

One source of confusion in evaluating claims for and against the economist's psychology is that the theory has both positive and normative components. According to Marshak (1950, p. 111), 'The theory of rational behavior is a set of propositions that can be regarded either as idealized approximations to the actual behavior of men or as recommendations to be followed'. Savage's early work with Milton Friedman (1948; 1952) was explicitly descriptive, but Savage (1956) is just as explicitly normative. It is not surprising that a description of decision in terms of beliefs and desires should have a normative component which evaluates how well goals are achieved. Confusion arises, however, when the descriptive and prescriptive positions are inappropriately conflated to justify the rationality assumptions as a statement of fact. Many undergraduate microeconomics texts justify transitivity assumptions by a money pump argument as a prelude to demand theory. The Dutch book is used to defend probabilistic descriptions of belief. But both of these arguments are, at their source, explicitly normative (see Davidson, McKinsey and Suppes, 1955, p. 146; Ramsey, 1931).

A descriptive theory of choice which is grounded not in empirical reality but in logical deductions from normative principles, like Dutch books and money pumps, is not science, but metaphysics. Furthermore, normative justifications are implicitly introspective. A money pump argument really says, 'you wouldn't fall into this trap, would you?' Significant empirical work in psychology (Nisbett and Wilson, 1977), however, indicates that introspective evidence is simply unreliable. When individuals turn to review and justify their decisions, they may have no access to the mental states which guided their choice. On the other hand, it seems to us quite reasonable to build models of financial asset pricing which assume that traders are probabilistically sophisticated, on the supposition that traders who are not will either not long survive in the market or not, as a group, be large enough to have a significant effect on prices. Financial markets, unlike Dutch books, actually exist, and the claim that individuals with probabilistically incoherent beliefs do not fare well is a claim of fact, to be tested against market data.

The conflation of positive and normative concerns in decision theory is more fundamental than simple carelessness in an argument. In his criticism of the fact/value dichotomy, Putnam (2002) asks us to consider the word 'cruel'. He observes that the word often has both descriptive and normative content, and in most uses they cannot be separated. The same could be said of the adjective 'rational' in economists' usage. Marshak (1950, p. 111) illustrates this perfectly when he writes that the purpose of EU is '… to describe the behavior of men who, it is believed, cannot be "all fools all the time"….' When the word 'rational' is used to describe a system in which all agents hold accurate probabilistic beliefs, the implication is that someone holding inaccurate beliefs gets it wrong. REE is often informally defended by the assertion that, if an economic actor's beliefs were incorrect, he would observe this and form new ones. The assertion is either a positive assertion, that actors do indeed have such beliefs, or a normative assertion, that they should hold such beliefs. The normative assertion is a metaphysical defence of the validity of the rational expectations hypothesis. The positive assertion is a claim of fact whose validity could in principle be put to test, but testing the claim would in fact require so rich a set of

ancillary maintained hypotheses that practically it is infeasible.

Given all the problems of the two views of rationality, one might wonder why economics needs a rational actor. Dennett (1971, p. 92) provides perhaps the best defence of belief/desire explanations. He contrasts what he calls the *design stance*, predicting behaviour from an understanding of how an agent is designed, or built, with the *intentional stance*, attributing to the agent beliefs and desires, and predicting from them. The intentional stance is useful, he writes, ‘Whenever we have reason to suppose the assumption of optimal design is warranted, and doubt the practicality of prediction from the design … stance’. Warranting the optimal design assumption means for Dennett not that the system actually be designed to achieve a fixed set of goals, but that this assumption is a useful first approximation. ‘Not surprisingly’, he observes,

> as we discover more and more imperfections …, our efforts at intentional prediction become more and more cumbersome and undecidable, for we can no longer count on the beliefs, desires, and actions going together that ought to go together. Eventually we end up, following this process, by predicting from the design stance; we end up, that is, dropping the assumption of rationality. (p. 95)

This movement, from rationality to realism, is the motivation for taking behaviour more seriously.

Rationality and behaviours

Game theory and general equilibrium theory are ‘system frameworks’. They imagine a collection of individual agents interacting in some systematic way, strategically in game theory, as described by the normal or extensive form of the game, and through markets in general equilibrium theory. In each case, the model produces an ‘equilibrium’ of the system. The first stage in the development of a system framework involves determining its consistency and internal coherence, that is, conditions which guarantee the existence of equilibrium. This analysis will be as abstract and general as possible, to encompass as large a repertory of behaviours as possible. The second stage is the application of the framework to derive useful statements about the world. This requires explicit behavioural assumptions about agent behaviour and describing the resulting equilibrium. These statements – predictions about market or game behaviour – can be examined empirically.

There are two difficulties with the received models of decision theory such as expected utility and dynamic programming in this kind of research program. First, as these models are formulated, behaviours are not accessible. For example, using expected utility to derive home bias in financial asset markets – that is, investors tend not to take positions in foreign assets – requires complicated assumptions about traders' beliefs. Second, these models are insufficiently rich to capture all the behaviours one might want to examine. For instance, the additively separable intertemporal expected utility model conflates time preference and risk aversion because the model is too thinly parametrized.

Behavioural economics is a research program which will, its proponents argue, replace rational actor models with a more psychologically informed view of human decision making. Much of behavioural economics, however, is less ambitious (and thus, perhaps, more useful). This work can be described as reformulating or extending rational actor models so as to make those observable behaviours whose

implications we wish to examine more accessible. While much of this work is at the core of behavioural economics, many who do this work eschew the label; not only behavioural economists are interested in behaviour. Here we discuss four categories of research which cover much work on behaviours, both by behavioural economics and by its critics.

Recontextualizing decision

GCT is a very parsimonious simplification of a decision problem. In modelling there is a trade-off between behavioural accuracy and parsimony in the description of decision problems. In general equilibrium models, for example, behavioural accuracy may improve descriptive and explanatory power, but parsimony is required because individual decisions are only one piece of the analysis, and complicated models of individual behaviour may generate only intractable market models.

One implication of GCT is that preferences are not choice-set dependent. Even in the early days of decision theory, important models such as minimax regret (Savage, 1951) violated the requirement of a single preference order on a universal space of potential choices. Furthermore, many choice-set effects appear to be perfectly rational. Consider the behaviour of a well-mannered but very hungry person at a dinner party. A plate is passed to him with three pieces of the main course, ordered in size such that $a \prec b \prec c$. Being both well-mannered and hungry, he chooses the second largest piece, *b*. Suppose now that the plate had been passed around the table in the other direction, so that when it comes to him there remains only *a* and *b*. Now according to his rule he chooses *a*. Is he called irrational by the GCT theorists at the table?

Kahneman and Tversky's (1979) prospect theory illustrates another way in which decision problems can be recontextualized. Here additional context, a status quo, is added to the description of the decision problems. Gambles are viewed as probability distributions over gains and losses relative to the status quo. Given a status quo, a preference order over all possible final outcomes exists, but that preference order varies with the status quo. There is, however, a stable preference order over the universe of all possible gains and losses; more context is added by redefining the objects of choice. A similar transformation is accomplished in Gul and Pesendorfer's (2004) model of choice with self-control problems. In the conventional infinite-horizon optimal consumption problem, the objects of choice are consumption paths. Gul and Pesendorfer, on the other hand, take the objects of choice to be pair consisting of a current period consumption and a decision problem to be solved tomorrow. Gul and Pesendorfer's model is an example of a *menu choice model*. Although used somewhat earlier, the first formal development of such models was by Kreps (1979) to describe preferences for flexibility.

Constructing rationality

The economist's conventional view of market interaction posits a collection of individuals with well-formed preferences meeting in a marketplace. The preferences, along with endowments and technologies, are exogenous to the system. On the other hand, some attendees at a large outdoor concert are there because they like the music, while others are there because of the crowd. Teenagers' evaluation of clothing style has perhaps as much to do with who wears such clothes as with their cut and pattern. These are all examples of socially constructed preferences.

Socially constructed preferences are a part of conventional economic theory. Both NGT and REE are models of socially constructed preferences. In each case desires are fixed, but beliefs adjust. However, neither of these equilibrium concepts is particularly well-supported by belief adjustment (learning) processes. The literature on learning Nash equilibrium is huge, and the state of the art is that, while one can construct learning dynamics that will find a Nash equilibrium, many intuitive learning processes will often fail. Blume and Easley (1982) show that rational equilibrium can easily fail to be reached by any reasonable learning process.

Restricting the socially determined component of preferences only to beliefs is an artificial constraint, and to limit social influence on preference formation to learning is to miss most of the interplay between the individual and the group. Manski (2000) observes that the implications of social interactions through learning and through tastes are distinct, and the difference is significant for policy analysis. Any theory of the interaction of desires requires a new set of primitives which describe the preference formation mechanism. One popular approach has been to model the evolution and workings of pro-social norms of cooperation and trust. Bowles (1998) is an engaging survey of this work. Much less has been done on the evolution and workings of anti-social norms, such as discrimination and stigmatization. Others have turned to biological metaphors. Here one might look at the population dynamics of rules or preferences on a game form or market where game or market outcomes (not utilities) determine the composition in the next round of the population's decision rules or preference orders (Güth and Kliemt, 1998; Blume and Easley, 1992; 2006). Pro-social behaviour such as reciprocity and altruism has also been investigated from the biological standpoint (Bergstrom, 2002; Sethi and Somanathan, 2003). One lesson of this literature is that the nature of the interaction between agents is at least as important as the model of choice in determining system outcomes. About the embeddedness of economic action in social life, Granovetter (1985, p. 506) writes,

> The notion that rational choice is derailed by social influences has long discouraged detailed sociological analysis of economic life and led revisionist economists to reform economic theory by focusing on its naive psychology. My claim here is that however naive that psychology may be, this is not where the main difficulty lies – it is rather in the neglect of social structure.

The content of preferences and beliefs

It has been conventional in economic analysis to construe self-interest very narrowly. No ‘other-regarding’ values are expressed in preferences, and conventionally to do otherwise is frowned upon. For instance, it is hard to explain why an individual votes in an election by her effect on the outcome, without referring to the psychic rewards of the act of voting. Yet the claim that people vote because of norms of citizenship and the like is often regarded as ‘nearly tautological’ (Ordeshook, 1986, p. 50.). On the other hand, critics of economic man often incorrectly assert that rational actors are excessively self-interested; incorrectly, because the existence of preferences and the content of preferences are distinct issues. The rationality hypothesis does not preclude other-regarding desires. Interest in those externalities that arise from ethical concerns, social norms, and other social constructions has increased enormously since the mid-1990s. Much of the literature on social interactions is a study of the

consequences of other-regarding preferences. Not surprisingly, other-regarding preferences usefully model both altruism and racism. This is not a fix for those critics who see the selfishness of traditional neoclassical models as a moral failing rather than a behavioural one.

A distinct problem which, unfortunately, has not been much addressed by behavioural economics is the use of individual preferences in the economist's version of moral philosophy. The same preferences which are revealed through shopping behaviour at the grocery store are supposed to be informative for the ethical questions posed by welfare economics. One could, in fact, distinguish ‘ethical preferences’ from ‘subjective preferences’ as Harsanyi (1955) has done, and it would be interesting to know if social psychology has anything to say about the relationship between the two types of decision problems, individual and social, which economists address.

Different psychologies

Some economists look to replace the folk psychology of beliefs and wants with something altogether different. Neuroeconomics is one such attempt, although the neuroeconomics literature seems to eschew drawing economic conclusions from imaging data. Unfortunately, the link between brain and mind is elusive. Eliminative materialism is a position taken by some cognitive scientists, which claims that beliefs and desires do not exist as mental states, and will have no place in an accurate account of the mind. Theoretical and methodological arguments in its support can be found, for instance, in Churchland (1981). An economics which takes its microfoundations entirely from cognitive science could look extremely different than the economics of today. But even if one is more hopeful than Churchland for the utility of the economist's folk psychology, the goal is far off. As one leading neuroimaging specialist puts it, ‘Despite fantastic technical developments, lingering methodological and conceptual limitations hinder progress in understanding how mental processes (wrapped up in folk psychology) reduce to or emerge from neural processes’ (Schall, 2004, p. 44). Savoy (2001, p. 36) has a bleaker view:

> Do the new discoveries about human brain function based on neuroimaging experiments really teach us things that are relevant for the study and understanding of behaviour? That is a question which you must answer. My own impression is that, at present, the overwhelming thrust of these data are toward understanding brain organisation, rather than human behaviour. Of course, we assume that when brain organisation is sufficiently well understood, it will lead to increases in our understanding of behaviour. But I do not think, as yet, there is a great deal of progress in that direction.

Neuroeconomics hopes to replace the belief/desire folk psychology that informs most of modern analytical economics with a more accurate scientific psychology. Alternatively, one could construct a different folk psychology which, like utility maximization, has no scientific pretensions, but is more descriptively accurate. Models of *intrapersonal* conflict are the most familiar example of this kind of framework. Strotz (1955) demonstrated the possibility of time-inconsistent planning in intertemporal utility maximization problems, and Pollak (1968) subsequently displayed the essentially strategic nature of the planning problem as a problem of competition between the selves choosing at different dates. Schelling (1984) described a variety of decision problems with aspects of intrapersonal conflicts, and

discussed them from a game-theoretic perspective. He, for instance, wrote about the competition between that part of him which desires nicotine and that part which wants to give it up. This is a contest for self-control. The literature today contains a number of intertemporal models which, following Pollak (1968) distinguish two kinds of behaviour. Sophisticated behaviour chooses today with full knowledge that her future selves may try to undo her decision. A choice is a subgame perfect equilibrium of a game played by all her selves. Naive behaviour chooses today assuming, perhaps incorrectly, that her future selves will stick with her decisions. These two models are intrinsically no more realistic than GCT, just different.

For the working economist, the ultimate test of a psychologically more accurate theory of individual choice is how it performs in explaining market and other social outcomes rather than how well it predicts the behaviour of an individual. Could theory *A*, more informed with insights from psychology and cognitive science, possibly be less useful for economists? Here are three possibilities: (1) Theory *A* might be extremely complex. Its application to a heterogeneous-agent financial market model, for instance, is simply impossible to work with, and no conclusions can be derived. (2) Theory *A* might require for its application data that we can observe in a controlled and heavily instrumented setting but simply cannot collect in the field. (3) Theory *A* may not be posed with concepts which are useful for the economist's interpretation of social outcomes. For instance, theory *A* might be couched in terms of chemical states of the brain, and not speak at all about agents' intentions, beliefs or desires. While it may be possible to construct a biochemical model of the invisible hand, it would not be useful for welfare economics.

Evidence on the question of whether these models lead to better market analyses is sparse, and mixed, and there is no evidence on how these models perform relative to menu choice models, which address the same questions from a rational choice perspective. More generally, more work needs to be done in evaluating behavioural models with respect to their economic performance. How useful are they for deriving implications about the performance of aggregate economic variables such as prices? This kind of research is already under way. Two examples are Kocherlakota (2001) and Laibson (1997).

An instance of point (3) can be seen in the time-inconsistency literature. Pollak (1968) and his followers (O'Donoghue and Rabin, 1999) see choice not as the expression of a single desire, but as the outcome of conflict, perhaps inefficient and destructive, between competing desires. Now the Pareto ranking of alternatives in a social interaction either becomes dependent on which of the many competing preference orders we modellers choose for each individual or it becomes empty if we try to respect them all. The advantage of menu choice models, the rational-choice approach to modelling problems of self-control, is that there is a well-defined notion of preference for each agent, from which a Pareto ranking can be constructed. To be fair, rational choice modelling also poses problems for welfare economics. If individuals make consistent errors in a class of choice problems, what can revealed preference say about intentions? In the presence of systematic error, a welfare economics built from revealed choice is at best misleading.

Conclusion

The purpose of decision models in economics is to explain the behaviour not of a single individual but of aggregates of individuals. Sometimes economists explain by appeal to ‘Laws’, such as ‘the Law of

Supply and Demand'. But this mode of explanation is mostly an intermediate product; useful, perhaps, for generating back-of-the-envelope predictions about the effects of a tax on market price, but not a source of understanding. There are few natural laws in the social sciences, and the domains of the few we can identify are very limited.

More often, economists appeal to 'mechanism'. We try to understand economic phenomena, such as the determination of prices in different kinds of markets, in terms of the mechanisms which generate them. Given our commitment to methodological individualism, this requires an explanation of how individual economic actors interact with one another. This is where rational actor theories are employed, and it is with respect to how these models do in this discussion rather than how they do in other domains, such as explaining individual behaviour, that the rationality principle should be evaluated.

Unfortunately, perhaps, at this point there are no serious alternatives to the rationality principle. For all of its buzz, proponents of *bounded rationality*, by which we mean models of behaviour that consider beliefs and desires but that do not optimize, have so far failed to deliver decision models which are robust and not tightly tied to a small class of decision problems.

It is perhaps too early in its intellectual history to ask for as much from *cognitive models*. We are sceptical about the value for social and economic systems analysis of unpacking the black box of consumer behaviour by deploying a rich and sophisticated model of cognitive process within a general equilibrium or game theoretic model. There is a point to reductionism. On the other hand, we are enthusiastic about the possibility that cognitive science will contribute to sharpening the rationality principle. The focus of much modern decision theory, such as Kahneman and Tversky (1979), Gilboa and Schmeidler (1989) and Gul and Pesendorfer (2004), has been to make the black-box model better by looking for formulations of rational choice models that better conform to the data. A better understanding of decision mechanism will doubtless suggest constraints on black-box behaviour which can be captured in reduced-form decision models, and perhaps it will uncover constraints that cannot be observed from behaviour alone.

Evolutionary models have also been proposed as an alternative framework to rational choice decision-making. Market forces, or a combination of markets and biology, favour some decision rules over others. In the long run, the market will be populated mostly by those decision rules that are 'most fit', rational or not. One can indeed ask if the forces of market selection favour rational decision rules (Blume and Easley, 1992; Sandroni, 2000), but the study of market population dynamics is complementary to rather than a substitute for rational choice models. Blume and Easley (2006), for instance, demonstrate how market forces select within the class of rational decision rules, favouring some kinds of preferences and beliefs over others.

Although there appear no be no serious alternatives to the rational choice paradigm on the near horizon, there is much to regret in how the rationality principle is discussed. The following statements should be self-evident, but clearly are not, judging by our reading of the literature: (1) Rationality does not mean complete or symmetric information. In fact, much of rational actor social science attempts to understand social outcomes when these conditions do not obtain. (2) Rationality does not require individuals to be entirely selfish. While much effort has been made to understand social norms from the point of view of entirely individualistic preferences, the insistence on relying on self-regarding rather than pro-social preferences is a matter of the content of preferences, rather than an axiom of rationality per se. (3) Rationality does not mean expected utility. Expected utility is one small class of decision models for choice under uncertainty. Its dominance in application was understandable in the 1970s, when few

alternatives were on the table. Since then decision theorists have been creative in developing better-behaved alternatives, and equilibrium and game theorists have been clever in applying them. (4) Rationality does not mean ‘rational expectations’. For a belief restriction to be a requirement of rationality, it must be clear that all those who are not ‘all fools all the time’ must have correct beliefs. No research into learning in economics suggests this is the case in any kind of complex environment.
There is also much to regret in how the rationality principle has been deployed in economic analysis. Given the explosion of decision-theoretic research since the 1970s, it is surprising how little this research has affected market and game theoretic analysis. The norm still seems to be self-interested preferences, expected utility and rational expectations (or Nash equilibrium). At this point the question of whether contemporary decision models such as Choquet expected utility and cumulative prospect theory have anything new to say about, say, asset pricing, is open. The value to economists of new decision theories, rational choice or not, is not in how they perform in a laboratory but how they perform in the analysis of markets and other social systems. Too rarely have modern decision theories been exposed to this test.
Rational actor social science is a broader tent than both its supporters and its critics make it out to be. We expect the rational choice framework to be as dominant when the next edition of the *New Palgrave* goes to press as it is today. But we also expect the set of decision-theoretic models deployed in the analysis of social systems will be quite different, and probably more diverse, than it is now.

See Also

- expected utility hypothesis
- methodological individualism
- rationality, history of the concept
- Savage's subjective expected utility model
- uncertainty
- utilitarianism and economic theory
- utility

Bibliography

Allais, M. 1953. Le comportement de l'homme rationel devant le risque: Critique des postulats et axiomes de l'Ecole Américaine. *Econometrica* 21, 503–46.

Aumann, R.J. 1962. Utility theory without the completeness axiom. *Econometrica* 32, 445–62.

Bentham, J. 1789. *An Introduction to the Principle of Morals and Legislations*. Based on information from English Short Title Catalogue. Eighteenth Century Collections Online. Farmington Hills, MI: Gale Group.

Bergstrom, T.C. 2002. Evolution of social behavior: individual and group selection. *Journal of*

Economic Perspectives 16(2), 67–88.

Blume, L. and Easley, D. 1992. Evolution and market behavior. *Journal of Economic Theory* 58, 9–40.

Blume, L.E. and Easley, D. 1982. Learning to be rational. *Journal of Economic Theory* 26, 340–51.

Blume, L.E. and Easley, D. 2006. If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets. *Econometrica* 74, 929–66.

Bowles, S. 1998. Endogenous preferences: the cultural consequences of markets and other economic institutions. *Journal of Economic Literature* 36, 75–111.

Chipman, J.S., Hurwicz, L., Richter, M.K. and Sonnenschein, H.F. 1971. *Preferences, Utility, and Demand*. New York: Harcourt Brace Jovanovich.

Churchland, P.M. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78(2), 67–90.

Cyert, R.M., Simon, H.A. and Trow, D.B. 1956. Observation of a business decision. *Journal of Business* 29, 237–48.

Davidson, D., McKinsey, J.C.C. and Suppes, P. 1955. Outlines of a formal theory of value, I. *Philosophy of Science* 22, 140–60.

Dennett, D.C. 1971. Intentional systems. *Journal of Philosophy* 68(4), 87–106.

Fishburn, P.C. 1991. Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty* 4, 113–34.

Fisher, I. 1892. *Mathematical Investigations in the Theory of Value and Prices*. New Haven: Connecticut Academy of Arts and Sciences.

Friedman, M. and Savage, L.J. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56, 279–304.

Friedman, M. and Savage, L.J. 1952. The expected-utility hypothesis and the measurability of utility. *Journal of Political Economy* 60, 463–74.

Gale, D. and Mas-Colell, A. 1975. An equilibrium existence theorem without ordered preference. *Journal of Mathematical Economics* 2, 9–15.

Gilboa, I. and Schmeidler, D. 1989. Maximin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* 91, 481–510.

Gul, F. and Pesendorfer, W. 2004. Self-control and the theory of consumption. *Econometrica* 72, 119–58.

Gul, F. and Pesendorfer, W. 2005. The case for mindless economics. Working paper, Princeton University.

Güth, W. and Kliemt, H. 1998. The indirect evolutionary approach: bridging the gap between rationality and adaptation. *Rationality and Society* 10, 377–99.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Herstein, I. and Milnor, J. 1953. An axiomatic approach to measurable utility. *Econometrica* 21, 291–7.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Kocherlakota, N. 2001. Looking for evidence of time-inconsistent preferences in asset market data. *Federal Reserve Bank of Minneapolis Quarterly Review* 25(3), 13–34.

Kreps, D.M. 1979. A representation theorem for ‘preference for flexibility’. *Econometrica* 47, 565–78.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112, 443–77.

Machlup, F. 1946. Marginal analysis and empirical research. *American Economic Review* 36, 519–54.

Manski, C.F. 2000. Economic analysis of social interactions. *Journal of Economic Perspectives* 14(3), 115–36.

Marshak, J. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18, 111–41.

Mas-Colell, A., Whinston, M.D. and Green, J.R. 1995. *Microeconomic Theory*. Oxford: Oxford University Press.

Nisbett, R.E. and Wilson, T.D. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 83, 231–59.

O'Donoghue, T. and Rabin, M. 1999. Doing it now or later. *American Economic Review* 89, 103–24.

Ordeshook, P.C. 1986. *Game Theory and Political Theory: An Introduction*. Cambridge: Cambridge University Press.

Pareto, V. 1895. Considerazioni sui Principi Fondamentali dell'Economia Politica Pura, part V. *Giornale degli Economisti, ser*. 2(5), 119–57.

Pascal, B. 1672. *Pascal's Pensées*. New York: E.P. Dutton and Co, 1958.

Pollak, R.A. 1968. Consistent planning. *Review of Economic Studies* 35, 201–8.

Putnam, H. 2002. The entanglement of fact and value. In *The Collapse of the Fact/Value Dichotomy and Other Essays*. Cambridge, MA: Harvard University Press.

Ramsey, F.P. 1931. Truth and probability. In *The Foundations of Mathematics and other Logical Essays*, ed. R.B. Braithwaite. London: K. Paul, Trench, Trubner and Co.

Samuelson, P.A. 1938a. A note on the pure theory of consumer's behavior. *Economica* 5(17), 61–71.

Samuelson, P.A. 1938b. The empirical implications of utility analysis. *Econometrica* 6, 344–56.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sandroni, A. 2000. Do markets favor agents able to make accurate predictions? *Econometrica* 68, 1303–42.

Savage, L.J. 1951. The theory of statistical decision. *Journal of the American Statistical Association* 46 (253), 55–67.

Savage, L.J. 1956. *The Foundations of Statistics*, 2nd edn. New York: Dover, 1972.

Savoy, R.L. 2001. History and future directions of human brain mapping and functional neuroimaging. *Acta Psychologica* 107(1–3), 9–42.

Schall, J.D. 2004. On building a bridge between brain and behavior. *Annual Review of Psychology* 55, 23–50.

Schelling, T. 1984. The intimate contest for self-command. In *Choice and Consequence: Perspectives of an Errant Economist*. Cambridge, MA: Harvard University Press.

Sethi, R. and Somanathan, E. 2003. Understanding reciprocity. *Journal of Economic Behavior and Organization* 50, 1–27.

Smith, A. 1789. *An Inquiry into the Nature and Causes of The Wealth of Nations*, 5th edn. Chicago: University of Chicago Press, 1976.

Strotz, R.H. 1955. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23(3), 165–80.

von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd edn. Princeton: Princeton University Press.

Wald, A. 1939. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics* 10, 299–326.

How to cite this article

Blume, Lawrence E. and David Easley. "rationality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000277> doi:10.1057/9780230226203.1390

The New Palgrave Dictionary of Economics Online

rationing

J. Peter Neary
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Rationing occurs whenever economic agents face quantity constraints on their demand for or supply of particular commodities. This article reviews the main results of rationing theory: a tightening of a ration constraint raises the demand for unrationed substitutes and reduces the price responsiveness of all unrationed goods (the Le Chatelier effect). It shows how the technique of virtual prices can be used to generalize these results to the case of strictly binding rations, and briefly reviews some applications, empirical and theoretical, of rationing theory to public and environmental economics, fix-price macroeconomics, and the effects of quotas on international trade.

Keywords

compensated demand; demand price; fixprice macroeconomics; income effects; international trade theory; labour supply; Le Chatelier principle; nonlinear budget constraints; nonlinear commodity taxation; public goods; quotas; rationing; reservation price; separability; shadow price; substitution effect; uncompensated demand; unemployment; virtual price

Article

Rationing refers to any situation in which economic agents face quantity constraints on their demand for or supply of particular commodities, unlike the standard situation in which they are free to purchase unlimited quantities subject only to fixed prices and a linear budget constraint.

Quantity constraints may impose consumption levels either below or above those that would be freely chosen: goods rationing in wartime illustrates the former, while examples of the latter include precommitted expenditures and unemployment (which may be viewed as ‘forced consumption’ of leisure). From an analytic point of view, the two cases are identical and may be described by the general term ‘rationing’. This article outlines the principal results of the microeconomic theory of consumer rationing and then notes some of its applications. Similar results apply to producer rationing, with the added simplification that income effects do not arise for a profit-maximizing firm. I concentrate throughout on ‘simple’ rationing (that is, exogenous restrictions on the consumption of particular commodities); some work has also been done on ‘points’ rationing (where the consumer has a number of ration ‘points’ or ‘coupons’ to be allocated between a group of commodities), and there is an extensive literature on the general case of nonlinear budget constraints, from both theoretical and empirical perspectives (see Hausman, 1985).

Consider first the case where only two commodities are consumed. This misses many important aspects of rationing, but allows most of the basic ideas to be introduced using a simple diagram. In Figure 1, the unconstrained optimal consumption bundle (x_0, y_0) is represented by point *A*, the point of tangency between the budget constraint *BC* and the highest attainable indifference curve, *II*. Suppose now that the consumer is faced with an additional constraint which stipulates that consumption of commodity *y* cannot exceed the level $\bar{y}$. The consumer is therefore forced to adjust consumption to the point *D*. Here, the budget constraint is still satisfied, consumption of *y* is constrained to equal $\bar{y}$, and expenditure has spilled over onto the unrationed commodity *x*, leading to a new higher consumption level $\tilde{x}$ (where a tilde ‘~‘ denotes a demand schedule for unrationed commodities in the presence of the ration constraint $\bar{y}$). The consumer is also at a lower indifference curve $I'I'$, so rationing reduces real income, or increases the true cost of living, even though prices and nominal income are unchanged.

Figure 1

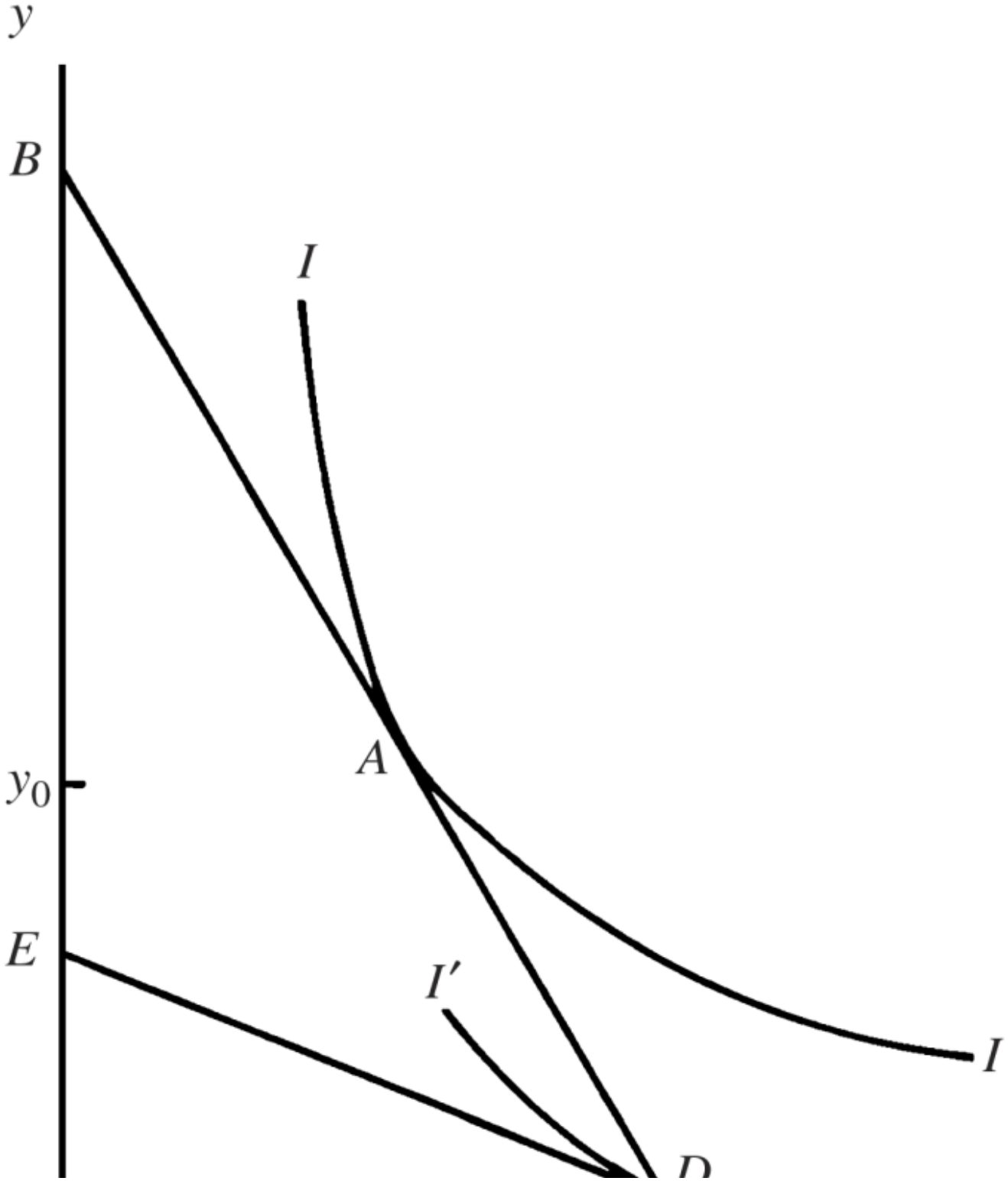
y
B
I
A
y_0
E
I′
I
D

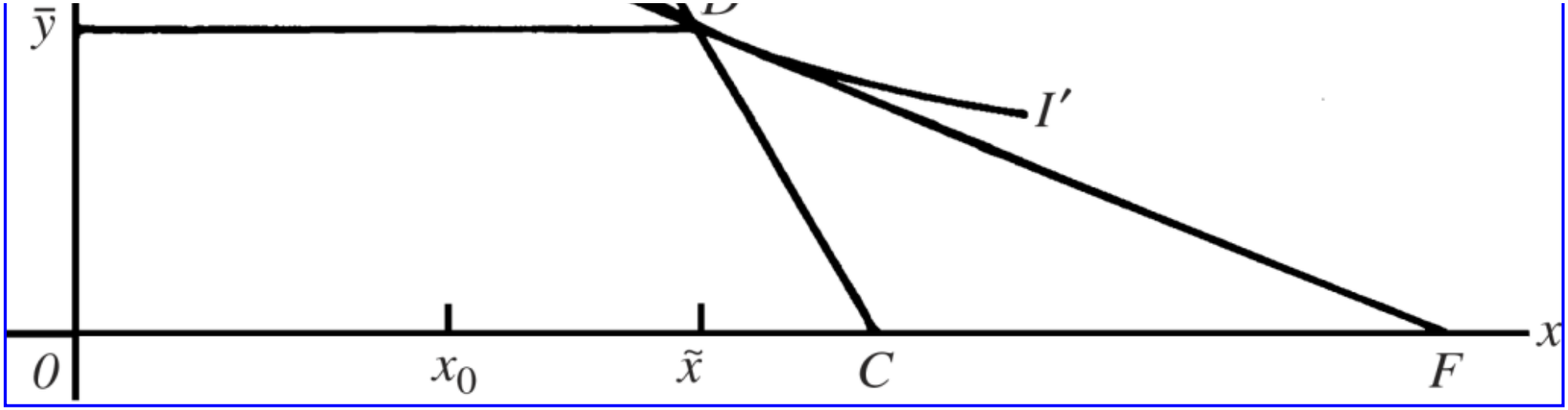

The case illustrated in Figure 1 is of course extremely special because the number of commodities is the same as the number of independent binding constraints, so the optimal consumption bundle is uniquely determined. In the more usual case, where there are fewer constraints than commodities, the consumer is free to allocate her uncommitted expenditure between a number of unrationed commodities, and a major focus of rationing theory has been on how this allocation, and its responsiveness to changes in exogenous variables, is affected by the presence of rationing. One important special case where rationing has very simple effects is when preferences are weakly separable between the rationed and unrationed commodities. This implies that the direct utility function $v(x, y)$, can be written in the form $U[f(x), y]$ where f is a scalar sub-utility function defined over a vector of unrationed commodities x. Weak separability implies that the demand for each good depends only on the prices of goods within the separable group and on the expenditure allocated to it. Hence the ration constraint has an income effect only and the constrained demand functions for the unrationed goods take the special form $\tilde{x}(p, I - q\bar{y})$. This specification is plausible in the case of some public goods (for example, increased spending on national defence is unlikely to affect the pattern of demand for private goods). Unfortunately, it is less satisfactory in other applications. For example, if leisure is the rationed commodity, weak separability implies that all other goods must be substitutes for it, irrespective of the extent of unemployment.

When preferences are unrestricted, a useful starting point to understanding the effects of rationing is to note that an unrationed consumer would choose to consume at point D under certain circumstances. Specifically, this would occur if the consumer were faced with a relative price ratio equal to the tangent to the indifference curve $I'I'$ at D, and were given an adjusted level of income such that that point represented the unconstrained utility-maximizing consumption bundle. The hypothetical relative price ratio required is given by the slope of the line EF. Following Rothbarth (1941) and Neary and Roberts (1980), the price of the rationed commodity underlying EF is called its *virtual price*: the price which would induce the consumer to purchase the ration level voluntarily. The advantage of this approach is that the effect of any exogenous shock on a rationed consumer may be decomposed into the sum of two effects on an orthodox unrationed consumer: the direct effect of the shock itself and the indirect effect arising from the induced change in the virtual price of the rationed good. (Note in passing that the terms ‘virtual price’, ‘demand price’ or, if the ration is set at a zero level, ‘reservation price’ are preferable to ‘shadow price’, since the latter risks confusion with the shadow price of the ration constraint which emerges from the consumer's maximization problem.)

It is clear that, for non-zero virtual prices to be unique and well defined, the indifference curve at D must be convex and differentiable. (Further technical details may be found in Neary and Roberts, 1980.) Consider then the general case where unrationed commodities are represented by a vector x and their prices by a vector p, while the commodity subject to a binding ration constraint is represented by a scalar y and its market price by q. (The algebra that follows applies equally to the general case with more than one rationed commodity, but it is easier to give intuition for the case of a single rationed good.) It is then straightforward to relate the Hicksian or compensated demand schedules for x in the presence of rationing to the corresponding unrationed schedules, since both are evaluated at the same utility level, that corresponding to the indifference curve $I'I'$:

$$\tilde{x}^c(p, \bar{y}, u) = x^c(p, \bar{q}, u). \tag{1}$$

Here, a superscript '*c*' denotes compensated demands. Crucially, the virtual price $\bar{q}$ is not a parameter but is defined implicitly by the condition that it equate the unconstrained demand for *y* to the ration level $\bar{y}$:

$$\bar{y} = y^c(p, \bar{q}, u). \tag{2}$$

Differentiating (1) and (2) now yields two important comparative statics results. Consider first the effect of a change in the ration level $\bar{y}$ on the demand for unrationed commodities *x*:

$$\tilde{x}^c_{\bar{y}} = x^c_{\bar{q}}(y^c_{\bar{q}})^{-1} \tag{3}$$

where subscripts indicate partial derivatives (for example, $\tilde{x}^c_{\bar{y}}$ is the vector whose *i*'th element gives the partial derivative of the rationed compensated demand function for x_i with respect to the level of the ration constraint). Since the compensated own-price derivative $y^c_{\bar{q}}$ is negative, the sign of (3) depends on the sign of $x^c_{\bar{q}}$. Thus, a tightening of the ration constraint (a reduction in $\bar{y}$) raises the compensated demand for unrationed commodities which are net substitutes for *y* and reduces it for commodities which are net complements for *y*.

Next, consider the effects on the demand for *x* of changes in their own prices. Differentiating (1) and (2) and rearranging yields:

$$\tilde{x}^c_p - x^c_p = -x^c_{\bar{q}}(y^c_{\bar{q}})^{-1}x^c_{\bar{q}}. \tag{4}$$

Since the substitution effect $y^c_{\bar{q}}$ is negative, this equation shows that the difference between the matrices of own-price responses of the unrationed commodities with and without rationing is a positive definite matrix. For any particular unrationed commodity, this implies that rationing reduces its responsiveness to its own price. This result is often referred to as the Le Chatelier principle, and was first introduced into economics by Samuelson (1947, pp. 36–9). Strictly speaking, the principle relates only to a comparison of compensated demands (compare (10) below). Moreover, it is a local result only, since it requires that the derivatives of both rationed and unrationed demand schedules be evaluated at the same consumption bundle, point *D* in Figure 1 (though Roberts, 1999, shows that it applies globally in an average sense). Despite these qualifications, the principle is often interpreted as implying in general that the imposition of restrictions on some aspects of behaviour makes individuals less responsive to exogenous changes in their environment.

Equations (3) and (4) are two of the most important results in rationing theory. However, their simplicity depends crucially on the fact that they refer to the properties of compensated demand schedules. There is one special case where equation (3) holds exactly for uncompensated (Marshallian) as well as compensated (Hicksian) demands, namely, where the ration 'just' binds, in the sense that the ration constraint coincides exactly with the amount of *y* that would be demanded by an unrationed consumer (so that points *A* and *D* in Figure 1 coincide). This was the case for which Tobin and Houthakker (1950–1) derived their results in a classic paper. For strictly binding ration constraints, any exogenous change has additional income effects, whose implications were first derived by Neary and Roberts (1980).

To illustrate the additional income effects which strictly binding ration constraints introduce, refer again to Figure 1. The distance *OC* measures the consumer's actual income in terms of *x*, *I*/*p* or $\tilde{x} + (q/p)\bar{y}$. However, this income would not be sufficient to induce an unrationed consumer faced with prices *p* and $\bar{q}$ to consume voluntarily at *D*; to do this they would need an income equal (in terms of good *x*) to the distance *OF*. Simple geometry shows that this distance equals $\tilde{x} + (\bar{q}/p)\bar{y}$, or $[I + (\bar{q} - q)\bar{y}]/p$. Hence, the uncompensated

demands of the rationed consumer may be equated to the uncompensated demands of an unrationed consumer, provided the latter are evaluated at the virtual price $\bar{q}$ and at a 'virtual income' $I + (\bar{q} - q)\bar{y}$:

$$\tilde{x}(p, q, \bar{y}, I) = x[p, \bar{q}, I + (\bar{q} - q)\bar{y}]. \tag{5}$$

In addition, the virtual price and income must be such that they induce an uncompensated demand for the rationed good equal to the ration constraint, so that (2) must be replaced by:

$$\bar{y} = y[p, \bar{q}, I + (\bar{q} - q)\bar{y}]. \tag{6}$$

Differentiating (5) and (6) now yields the full effects of exogenous changes on the demand for unrationed commodities. Consider first the effect of a change in income:

$$\tilde{x}_I = x_I - x_y^c y_I. \tag{7}$$

Thus, an increase in income affects demands for unrationed goods in two ways: first, it has a direct effect identical to the effect of an income increase in the absence of rationing (though evaluated at the virtual prices and income, of course); and second, by raising demand for the rationed good (on the assumption that y is normal so that y_I is positive), it is equivalent to a tightening of the ration constraint, and so has an indirect effect given by eq. (3).

Differentiating (5) and (6) also gives the effect of a change in the ration constraint:

$$\tilde{x}_y = \tilde{x}_y^c + \tilde{x}_I(\bar{q} - q). \tag{8}$$

This shows that a tightening of the ration constraint has a compensated or substitution effect given by (3) and an additional income effect, given by the last term in (8). This term vanishes if the virtual and actual prices of the rationed good coincide, which corresponds to the case where the ration constraint 'just' binds. In the case illustrated in Figure 1, where the consumer would like to consume more of the rationed good, $\bar{q}$ exceeds q, and so a tightening of the ration constraint, by lowering real income, tends to reduce the demand for all normal unrationed goods. (Of course, as already noted in discussing the diagram, the total effect must be an increase in spending on the unrationed goods as a group.)

Finally, the effect of changes in prices may be obtained in a similar manner. First, for an increase in the prices of the unrationed goods themselves:

$$\tilde{x}_p = x_p - \tilde{x}_y^c y_p. \tag{9}$$

This shows that an increase in the price of an unrationed good has a direct effect, equal to its effect in the absence of rationing, and an indirect effect: by changing the demand for the rationed good it is equivalent to a tightening or relaxation of the ration constraint and so has the usual effect given by (3). Equation (9) may be rewritten in a form which, by comparison with (4), shows how income effects may counteract the Le Chatelier principle:

$$\tilde{x}_p - x_p = (\tilde{x}_p^c - x_p^c) + \tilde{x}_y^c y_I x. \tag{10}$$

By contrast, the effect of a change in the price of the rationed good is much simpler:

$$\tilde{x}_q = -\tilde{x}_I \bar{y}. \tag{11}$$

This price change has no substitution effect, which explains why q is not an argument in the compensated rationed demand schedules (1). Its only effect is to lower real income by requiring the consumer to pay more for the rationed good, and so it reduces the demand for normal unrationed goods.

Before we leave the basic comparative statics of rationing, a problem which is peculiar to this area should be mentioned. All the results which have been derived assume that the values of the exogenous variables are such that the ration constraint is a binding one. However, it is quite possible for a finite change in an exogenous variable to render the constraint non-binding. For example, in Figure 1 this would occur if the ration constraint $\bar{y}$ rose above the unconstrained demand $y(p, q, I)$. If this happens, the ration constraint ceases to be binding and the ordinary unconstrained demand functions become applicable. Shifts of ‘regime’ such as this dictate great care in applying rationing theory in cases where large finite changes in exogenous variables occur; and in applications such as fixprice macroeconomics, where interest focuses on the interaction between constraints which impinge on different agents.

In empirical applications of rationing theory, attention has focused on deriving explicit forms for rationed demand functions which are tractable but not too restrictive. As in the case of consumer theory in the absence of rationing, progress in this direction comes most easily not by specifying functional forms for the direct utility function but by adopting a dual approach, which takes the expenditure function as its starting point. In the presence of rationing the constrained expenditure function gives the minimum cost of attaining a given utility level when consumption of the rationed commodity is predetermined:

$$\tilde{E}(\bar{y}, p, q, u) = Min[\, p'x + q\bar{y} : v(x, y) \geq u] = p'\tilde{x}^c(p, \bar{y}, u) + q\bar{y}. \tag{12}$$

Substituting from (1) and (2) yields, after some manipulation, the fundamental relationship between constrained and unconstrained expenditure functions:

$$\tilde{E}(\bar{y}, p, q, u) = E(p, \bar{q}, u) + (\bar{q} - q)\bar{y}. \tag{13}$$

In principle, this identity permits the derivation of a matched pair of rationed and unrationed demand functions, characterizing the behaviour of the same consumer in both environments. Two interesting specifications of the expenditure function which permit this are investigated in the labour supply context by Deaton and Muellbauer (1981).

Unfortunately, the derivation of such matched pairs of demand functions is not possible in general. An alternative approach is to specify a general functional form for the rationed expenditure function which imposes fewer restrictions on demand responses though at the cost of an inability to write the unrationed demand functions in closed form. This approach has been pursued by Deaton (1981), who derives a system of rationed demand functions which express budget shares as a linear function of the ration level and the logarithms of prices and real expenditure on unrationed goods. He shows that treating expenditure on housing as predetermined in this framework leads to more plausible results than when it is assumed to be unconstrained. (Specifically, the rationed system goes much of the way towards avoiding the implausible rejection of homogeneity in nominal variables, which has been found in many empirical studies of demand.)

Insights derived from rationing have proved useful in many branches of economic theory other than consumer economics. In public economics and environmental economics, the study of optimal public policy has been extended to public goods and bads (the consumption of which is predetermined from an individual consumer's point of view) and nonlinear commodity taxation (of which government-imposed consumption quotas are a special but empirically important case). In macroeconomics, rationing theory has been used to model both current and expected future quantity constraints on households and firms in formalizations of Keynes's contribution as the economics of ‘general disequilibrium’, in which the failure of prices to adjust rapidly in the short run faces agents with quantity constraints which ‘spill over’ to influence their behaviour in other markets (see Neary and Stiglitz, 1983, and further references given there). In international trade theory, it has been shown that the behaviour of an economy subject to quotas (quantitative restraints on imports) can be characterized in the same way as that of a household subject to rationing, with the added benefit that the virtual prices can be interpreted as the domestic market-clearing prices (see Anderson and Neary, 1992).

Finally, within a utility-maximizing framework, it may be noted that rationing necessarily imposes a welfare loss. This consideration underlies the instinctive preference by most economists for the use of the price system as an allocation mechanism rather than direct controls, a preference which is supported by the two fundamental theorems of welfare economics. Nevertheless, in situations where the conditions for these theorems do not obtain, it may be possible to give a second-best justification for rationing. While work along these lines pertains more to public economics than to rationing theory per se, mention may be made of two especially interesting contributions. One is a paper by Weitzman (1977), who develops a model where the just distribution of a particular commodity on the basis of need alone is considered a socially desirable end in itself. He shows that rationing the commodity is preferable to allocating it via the price system if tastes are homogeneous but income is unevenly distributed. The other is a paper by Guesnerie and Roberts (1984), who show that, in a second-best world with given commodity taxes (so that consumer prices diverge from marginal social valuations), rationing is likely to be welfare improving.

See Also

- demand theory
- labour supply
- Le Chatelier principle

Bibliography

Anderson, J.E. and Neary, J.P. 1992. Trade reform with quotas, partial rent retention and tariffs. *Econometrica* 60, 57–76.

Deaton, A.S. 1981. Theoretical and empirical approaches to consumer demand under rationing. In *Essays in the Theory and Measurement of Consumer Behaviour in Honour of Sir Richard Stone*, ed. A. Deaton. Cambridge: Cambridge University Press.

Deaton, A. and Muellbauer, J. 1981. Functional forms for labour supply and commodity demands with and without quantity constraints. *Econometrica* 49, 1521–32.

Guesnerie, R. and Roberts, K.W.S. 1984. Effective policy tools and quantity controls. *Econometrica* 52, 59–86.

Hausman, J.A. 1985. The econometrics of nonlinear budget sets. *Econometrica* 53, 1255–82.

Neary, J.P. and Roberts, K.W.S 1980. The theory of household behaviour under rationing. *European Economic Review* 13, 25–42.

Neary, J.P. and Stiglitz, J.E. 1983. Towards a reconstruction of Keynesian economics: expectations and constrained equilibria. *Quarterly Journal of Economics* 98(Supplement), 199–228.

Roberts, K. 1999. Rationality and the Le Chatelier Principle. *Journal of Economic Theory* 87, 416–28.

Rothbarth, E. 1941. The measurement of changes in real income under conditions of rationing. *Review of Economic Studies* 8, 100–7.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Tobin, J. and Houthakker, H.S. 1950–51. The effects of rationing on demand elasticities. *Review of Economic Studies* 18, 140–53.

Weitzman, M. 1977. Is the price system or rationing more effective in getting a commodity to those who need it most? *Bell Journal of Economics* 8, 517–24.

How to cite this article

Neary, J. Peter. "rationing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000034> doi:10.1057/9780230226203.1393

The New Palgrave Dictionary of Economics Online

Rau, Karl Heinrich (1792 – 1870)

H.C. Recktenwald

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cameralism; Freiburg School; Historical School, German; public finance; Rau, K. H.; Smith, A.

Article

Rau was born in Erlangen and was a lecturer (*Privatdozent*) and professor (1816). In 1822 Rau was appointed to a chair of economics at the University of Freiburg. Involved in political affairs, as were many German professors in the 19th century, he was appointed a member of the upper Chamber of Baden and in 1848 was elected to the Frankfurt Assembly.

At first influenced by Cameralist ideas, Rau was one of the main mediators and defenders of Smith's ‘system of natural liberty’, whose central principles, abstractly exposed, he embodied in a rich supply of illustrative facts in his famous *Lehrbuch* (1826–37) yet without attempting to test his hypotheses empirically, that is, to use factual materials as confirmation instead of pure description. To that extent he was not an original thinker. Yet he was a great teacher. Similar to Samuelson's *Economics* in our time, his best-selling textbook, published in eight editions (1862–9), was an authoritative work for the majority of economists teaching at German universities for several generations. Based on classical ideas, it thus shaped the economic and political *Weltbild* of future civil servants and lawyers.

Rau's tripartite division of economics, which was obviously influenced by Smith, was divided into three volumes, theory (economic laws), policy (*Polizeiwissenschaft*) and public finance; this division became the established tradition in the teaching of political economy at German universities and is divisive up to the present day. With the rise and the establishment of the German Historical School and its stress on both the ethical aspect of economic issues, that is, of the distribution of income and property, and on the historical character of economic principles, Rau's star faded, although his work on public finance became the foundation of Wagner's famous treatise.

Viewed in a historical continuum, the Freiburg School (Eucken, Röpke, von Hayek), Erhard's liberal economic policy and, more recently, a group of German economists who are attempting to revive Smith's tripartite theory of order (ethics, economics and politics as an entity) all indirectly resume that thread of Rau's concept, although on a different analytical level (Recktenwald, 1973, 1985). In the light

of a worldwide Smith renaissance in our epoch, Rau's editing function seems to merit secular attention.

Selected works

For a complete list, see C. Meitzel, *Handwörterbuch der Staatswissenschaft*, 4th edn, vol. 6. Jena: G. Fischer, 1925.

1825. *Über die Kameralwissenschaft*. Heidelberg: Groos.

1826–37. *Lehrbuch der politischen Ökonomie*. 3 vols. 8th edn, Heidelberg: C.F. Winter.

Bibliography

Recktenwald, H.C., ed. 1973. *Political Economy: A Historical Perspective*. London: Collier-Macmillan particularly the introduction.

Recktenwald, H.C., ed. 1985. *Ethik, Wirtschaft und Staat: Adam Smiths politische Ökonomie heute*. Darmstadt: Wissenschaftliche Buchgesellschaft particularly, 110–20 and 345–90 and G.J. Stigler's and P. A. Samuelson's contributions.

How to cite this article

Recktenwald, H. C. "Rau, Karl Heinrich (1792–1870)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000036> doi:10.1057/9780230226203.1394

The New Palgrave Dictionary of Economics Online

Raymond, Daniel (1786 – 1849)

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Lauderdale, Eighth Earl of; List, F.; national system; protectionism; Raymond, D.

Article

The first American to publish a treatise on economic topics, Raymond was born in Connecticut but made his home in Baltimore, where he practised law. *Thoughts on Political Economy* (1820) was written to while away the time as the young attorney waited for clients. The book constituted a challenge to classical orthodoxy and as such was warmly received by the protectionists. To make his voice more resounding they tried (without success) to secure Raymond a professorship at the University of Maryland that they were willing to underwrite. Raymond was an original thinker, whose ideas reverberated in the later writings of Frederick List, the historical economists and the 20th-century literature on economic development.

Raymond's principal concern was national economic development and, unlike the classics, he placed the nation rather than the individual in the centre of his analysis. Following Lauderdale, he distinguished between national and individual wealth, but unlike Lauderdale, to whom usefulness was the characteristic feature of public wealth and scarcity that of private wealth, Raymond interpreted national wealth in terms of its ‘capacity’ to produce goods. This view opens up to government a central position in promoting economic development by means of tariff protection. Raymond also underlines the conflicts of interest among different groups in the economy and again calls on government for their resolution.

While Raymond's basic ideas reflect the influence of Alexander Hamilton, his distrust of paper money and bank credit echoes the related views of Jefferson. Raymond also was highly critical of corporations. These incongruities were bound to affect the impact of his work.

See Also

- national system

Selected works

1819. *The Missouri Question*. Baltimore: Schaeffer & Maund.

1820. *Thoughts on Political Economy*. Baltimore: F. Lucas, Jr.

1828. *The American System*. Baltimore: Lucas & Deaver.

n.d. *The Elements of Constitutional Law*. 2nd edn, Baltimore: F. Lucas, Jr. and E.J. Cooder, 1832.

Bibliography

Conkin, P.K. 1980. *Prophets of Prosperity: America's First Political Economists*. Bloomington: Indiana University Press.

Dorfman, J. 1946. *The Economic Mind in American Civilization 1606–1865*. New York: Viking.

How to cite this article

Spiegel, Henry W. "Raymond, Daniel (1786–1849)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000038> doi:10.1057/9780230226203.1395

The New Palgrave Dictionary of Economics Online

real balances

Don Patinkin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

By the term ‘real balances’ is meant the real value of the money balances held by an individual or by the economy as a whole, as the case may be. The emphasis on real, as distinct from nominal, reflects the basic assumption that individuals are free of ‘money illusion’. It is a corresponding property of any well-specified demand function for money that its dependent variable is real balances. Indeed, Keynes in his *Treatise on Money* (1930, vol. 1, p. 222) designated the variation on the Cambridge equation that he had presented in his *A Tract on Monetary Reform* (1923, ch. 3: 1) as ‘The “Real-Balances” Quantity Equation’.

Implicit – and sometimes explicit – in the quantity-theory analysis of the effect of (say) an increase in the quantity of money is the assumption that the mechanism by which such an increase ultimately causes a proportionate increase in prices is through its initial effect in increasing the real value of money balances held by individuals and consequently increasing their respective demands for goods: that is, through what is now known as the ‘real-balance effect’. This effect, however, was not assigned a role in the general-equilibrium system of equations with which writers of the interwar period attempted to describe the workings of a money economy. In particular, these writers mistakenly assumed that in order for their commodity demand functions to be free of money illusion, they had to fulfil the so-called ‘homogeneity postulate’, which stated that these functions depended only on relative prices, and so were not affected by a change in the absolute price level generated by an equi-proportionate change in all money prices (Leontief, 1936, p. 192). Thus they failed to take account of the effect of such a change on the real value of money balances and hence on commodity demands. This in turn led them to contend that there existed a dichotomy of the pricing process, with equilibrium relative prices being determined in the ‘real sector’ of the economy (as represented by the excess-demand equations for commodities), while the equilibrium absolute price level was determined in the ‘monetary sector’ (as represented by the excess-demand equation for money): (Modigliani, 1944, sec. 13). This, however, is an invalid dichotomy, for it leads to contradictory implications about the determinacy or, alternatively, stability of the absolute price level (Patinkin, 1965, ch. 8).

Nor was the real-balance effect taken account of in Keynes's *General Theory* and in the subsequent Hicks (1937)–Modigliani (1944) IS–LM exposition of this theory, which rapidly became the standard

one of macroeconomic textbooks. According to this exposition, the only way in which a decline in wages and prices can increase employment is by its effect in increasing the real value of money balances, hence reducing the rate of interest, and hence (through its stimulating effect on investment) increasing the aggregate demand for goods and hence employment. A further and basic tenet of this exposition was that there was a minimum below which the rate of interest could not fall. So if the wage decline were to bring about a lowering of the rate of interest to this minimum before full employment were reached, any further decline in the wage rate would be to no avail. In brief, the economy would then be caught in the ‘liquidity trap’. And even though Keynes had stated in the *General Theory*, ‘whilst this limiting case might become practically important in the future, I know of no example of it hitherto’ (p. 207), the Keynesian theory of employment was for many years interpreted in terms of this ‘trap’.

It was against this background that Pigou (1943, 1947) pointed out that the increase in the real value of money holdings generated by the wage and price decline increased the aggregate demand for goods directly, and not only indirectly through its downward effect on the rate of interest. Pigou's rationale was that individuals saved in order to accumulate a certain amount of wealth relative to their income, and that indeed the savings function depended inversely on the ratio of wealth to income. Correspondingly, as wages and prices declined, the real value of the monetary component of wealth increased and with it the ratio of wealth to income, causing a decrease in savings, which means an increase in the aggregate demand for consumption goods. Pigou's argument (which was formulated for a stationary state) thus had the far-reaching theoretical implication that even if the economy were caught in the ‘liquidity trap’, there existed a low enough wage rate that would generate a full-employment level of aggregate demand. In this way Pigou (1943, p. 351) reaffirmed the ‘essential thesis of the classicals’ that ‘if wage-earners follow a competitive wage policy, the economic system must move ultimately to a full-employment stationary state’.

In his exposition and elaboration of Pigou's argument (which inter alia brought out the significance of the argument for dynamic stability analysis), Patinkin (1948) labelled the direct effect on consumption of an increase in the real value of money balances as the ‘Pigou effect’. However, in subsequent recognition of the fact that this effect is actually an integral part of the quantity theory – as well as the fact that Pigou had been anticipated in drawing the implications of this effect for the Keynesian system by Haberler (1941, pp. 242, 389, and 403) and Scitovsky (1941, pp. 71–2) – Patinkin (1956, 1965) relabelled it the ‘real-balance effect’ and presented it as a component of the wealth effect.

In an immediate comment on Pigou's article, Kalecki (1944) pointed out that the definition of ‘money’ relevant for the real-balance effect is not the usual one of currency *plus* demand-deposits: for example, in the case of a price decline, the increase in the real value of the demand deposits has an offset in the corresponding increase in the real burden on borrowers of the loans they had received from the banking system. Thus (emphasized Kalecki) the monetary concept relevant for the real-balance effect in a gold-standard economy is only the gold reserve of the monetary system.

More generally, the relevant concept is ‘outside money’ (equivalent to the monetary base, sometimes also referred to as ‘high-powered money’), which is part of the net wealth of the economy, as distinct from ‘inside money’, which consists of the demand deposits created by the banking system as a result of its lending operations and which accordingly is not part of net wealth (Gurley and Shaw, 1960). This distinction was subsequently challenged by Pesek and Saving (1967), who contended that banks regard

only a small fraction of their deposits as debt, so that these deposits too should be included in net wealth. In criticism of this view, Patinkin (1969, 1972a) showed that if perfect competition prevails in the banking system, the present value of the costs of maintaining its demand deposits equals the value of these deposits, so that the latter cannot be considered as a component of net wealth. This is also the case if imperfect competition with free entry prevails in the system. On the other hand, if – because of restricted entry – the banking sector enjoys abnormal profits, then the present value of these profits should be included in net wealth for the purpose of measuring the real-balance effect.

There remains the question of whether – for the purpose of measuring the real-balance effect – one should include government interest-bearing debt, as contrasted with the non-interest-bearing debt (viz., government fiat money) which is a component of the monetary base. Clearly, in a world of infinitely lived individuals with perfect foresight, the former does not constitute net wealth and hence is not a component of the real-balance effect: for the discounted value of the tax payments which the representative individual must make in order to service and repay the debt obviously equals the discounted value of the payments on account of interest and principal that he will receive. Nor is the assumption of infinitely lived individuals an operationally meaningless one: for as Barro (1974) has elegantly shown, if in making his own consumption plans, the representative individual with perfect foresight is sufficiently concerned with the welfare of the next generation to the extent of leaving a bequest for it, he is acting as if he were infinitely lived.

More specifically, Barro's argument is as follows: assume that an individual of the present generation achieves his optimum position by consuming C_o during his lifetime and leaving a positive bequest of B_o for the next generation. Clearly, such an individual could have increased his consumption to $C_0 + \Delta C_0$ and reduced his bequest to $B_0 - \Delta C_0$ – but preferred not to do so. Assume now that the individual also holds government bonds payable by the next generation, and let the real value of these bonds increase as the result of a decline in the price level, expected to be permanent. The revealed preference of the present generation for the consumption-bequest combination C_o, B_o implies that this increase in the real value of its holdings of government interest-bearing debt will not cause it to increase its consumption at the expense of the next generation. In brief, government debt in this case is effectively not a component of wealth and hence of the real-balance effect.

Needless to say, the absence of perfect foresight, and the fact that individuals might not leave bequests (as is indeed assumed by the life-cycle theory of consumption) means that government interest-bearing debt should to a certain extent be taken account of in measuring the real-balance effect – or what in this context is more appropriately labelled the ‘net-real-financial-asset effect’ (Patinkin, 1965, pp. 288–94).

If we assume consumption to be a function of permanent income, and if we assume that the rate of interest which the individual uses to compute the permanent income flowing from his wealth is 10 per cent and the marginal propensity to consume out of permanent income before 0.80, then the marginal propensity to consume out of wealth (and out of real balances in particular) is the product of these two figures, or 0.08. However, in the case of consumers' durables (in the very broad sense that includes – besides household appliances – automobiles, housing, and the like), the operation of the acceleration principle implies an additional real-balance effect in the short run. In particular, assume that when the individual decides on the optimum composition of the portfolio of assets in which to hold his real wealth, W, he also considers the proportion, q, of these assets that he wishes to hold in the form of consumers durables, K_d, so that his demand for the *stock* of consumer-durable goods is $K_d = qW$. Assume

now that wealth increases solely as a result of an increase in real balances, M/p. This leaves the representative individual with more money balances in relation to his other assets than he considers optimal. As a result he will attempt to shift out of money and into these other assets until he once again achieves an optimum portfolio. In the case of consumers' durables, this means that in addition to his preceding demand for new consumer-durable goods, he has a demand for

$$C_d = \Delta K_d = q[\Delta(M/p)] = q[(M/p)_t - (M/p)_{t-1}]$$

units, where $(M/p)_t$ represents real balances at time t. In general, the individual will plan to spread this additional demand over a few periods. In any event, once an optimally composed portfolio is again achieved, this additional effect disappears, so that the demand for new consumers' durables (which in the case of a stationary state is solely a replacement demand) will once again depend only on the ordinary real-balance effect as described at the beginning of this paragraph (Patinkin, 1967, pp. 156–62).

It is, of course, true that the process of portfolio adjustment generated by the monetary increase will cause a reduction in the respective rates of return on the other assets in the portfolio, so that the initial wealth effect of the monetary increase will be followed by substitution effects. Now, Keynes limited his analysis in the *General Theory* to portfolios consisting only of money and securities; hence (as indicated above) an increase in the quantity of money could increase the demand for goods only indirectly through the substitution effect created by the downward pressure on the rate of interest. But once one takes account of the broader spectrum of assets held by individuals, one must also take account of the direct real-balance effect on the purchase of these other assets as well.

Various empirical studies have shown that the real-balance effect as here defined (viz., as part of the wealth effect) is statistically significant (Patinkin, 1965, note M; Tanner, 1970). Other studies have demonstrated the statistical significance of yet another definition of this effect: namely, as the effect on the demand for commodities of an excess supply of money, defined as the excess of the existing stock of money over its ‘desired’ or ‘long-run’ level (Jonson, 1976; Laidler and Bentley, 1983; cf. also Mishan, 1958). It seems to me, however, that such a demand function is improperly specified: for though (as indicated above) the excess supply of money has a role to play in the consumption function (and particularly in that for consumers' durables), the complete exclusion of the real-balance effect *cum* wealth effect from the aforementioned demand function implies that in equilibrium there is no real-balance effect – an implication that is contradicted by the form of demand functions as derived from utility maximization subject to the budget constraint (Patinkin, 1965, pp. 433–8, 457–60; Fischer, 1981).

Granted the statistical significance of the real-balance effect, the question remains as to whether it is strong enough to offset the adverse expectations generated by a price decline – including those generated by the wave of bankruptcies that might well be caused by a severe decline. In brief, the question remains as to whether the real-balance effect is strong enough to assure the stability of the system: to assure that automatic market forces will restore the economy to a full-employment equilibrium position after an initial shock of a decrease in aggregate demand (Patinkin, 1948, part II; 1965, ch. 14: 1). On the assumption of adaptive expectations, Tobin (1975) has presented a Keynesian model with the real-

balance effect which under certain circumstances is unstable. On the other hand, McCallum (1983) has shown that under the assumption of rational expectations, the model is generally stable.

In any event, no one has ever advocated dealing with the problem of unemployment by waiting for wages and prices to decline and thereby generate a positive real-balance effect that will increase aggregate demand. In particular, Pigou himself concluded his 1947 article with the statement that such a proposal had ‘very little chance of ever being posed on the chequer board of actual life’. Thus the significance of the real-balance effect is in the realm of macroeconomic theory and not policy.

Correspondingly, recognition of the real-balance effect in no way controverts the central message of Keynes's *General Theory*. For this message – as expressed in the climax of that book, Chapter 19 – is that the only way a general decline in money wages can increase employment is through its effect in increasing the real quantity of money, hence reducing the rate of interest, and hence stimulating investment expenditures; but that even if wages were downwardly flexible in the face of unemployment, this effect would be largely offset by the adverse expectations and bankruptcies generated by declining money wages and prices, so that the level of aggregate expenditures and hence employment would not increase within an acceptable period of time. In Keynes's words: ‘the economic system cannot be made self-adjusting along these lines’ (ibid., p. 267). And there is no reason to believe that Keynes would have modified this conclusion if he had also taken account of the real-balance effect of a price decline (Patinkin, 1948, part III; 1976, pp. 110–11).

The above discussion has considered only the real-balance effect on the demand for goods. In principle, this effect also operates on the supply of labour: for the greater the real balances and hence wealth of the individual, the greater his demand for leisure as well, which means the smaller his supply of labour. This influence, however, has received relatively little attention in the literature (but see Patinkin, 1965, p. 204; Phelps, 1972; Barro and Grossman, 1976, pp. 14–16).

Another limitation of the discussion is that it deals only with a closed economy. In the analysis of an open economy, the real-balance effect plays an important role in some of the formulations of the monetary approach to the balance of payments.

See Also

- money illusion
- quantity theory of money

Bibliography

American Economic Association. 1951. *Readings in Monetary Theory*. Philadelphia: Blakiston.

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(November–December), 1095–117.

Barro, R.J. and Grossman, H.I. 1976. *Money, Employment and Inflation*. Cambridge: Cambridge University Press.

Fischer, S. 1981. Is there a real-balance effect in equilibrium? *Journal of Monetary Economics* 8(July), 25–39.

Gurley, J.G. and Shaw, E.S. 1960. *Money in a Theory of Finance*. Washington, DC: Brookings Institution.

Haberler, G. von. 1941. *Prosperity and Depression: A Theoretical Analysis of Cyclical Movements*, 3rd edn. Geneva: League of Nations.

Hicks, J.R. 1937. Mr Keynes and the ‘classics’: a suggested interpretation. *Econometrica* 5(April), 147–59. Reprinted in *Readings in the Theory of Income Distribution*, Philadelphia: Blakiston for the American Economic Association, 1946, 461–76.

Jonson, P.D. 1976. Money and economic activity in the open economy: the United Kingdom, 1880–1970. *Journal of Political Economy* 84(October), 979–1012.

Kalecki, M. 1944. Professor Pigou on ‘The classical stationary state’: a comment. *Economic Journal* 54 (April), 131–2.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London: Macmillan.

Keynes, J.M. 1930. *A Treatise on Money*. Vol. I: *The Pure Theory of Money*. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Laidler, D. and Bentley, B. 1983. A small macro-model of the post-war United States. *Manchester School* 51(December), 317–40.

Leontief, W. 1936. The fundamental assumption of Mr Keynes' monetary theory of unemployment. *Quarterly Journal of Economics* 51(November), 192–7.

McCallum, B.T. 1983. The liquidity trap and the Pigou Effect: a dynamic analysis with rational expectations. *Economica* 50(November), 395–405.

Mishan, E.J. 1958. A fallacy in the interpretation of the cash balance effect. *Economica* 25(May), 106–18.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12 (January), 45–88. Reprinted in American Economic Association (1951), 186–240.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38(September),

543–64. Reprinted with revisions and additions in American Economic Association (1951), 252–83.

Patinkin, D. 1956. *Money, Interest, and Prices*. Evanston, Ill.: Row, Peterson.

Patinkin, D. 1965. *Money, Interest, and Prices*, 2nd edn. New York: Harper & Row.

Patinkin, D. 1967. *On the Nature of the Monetary Mechanism*. Stockholm: Almqvist and Wicksell. Reprinted in Patinkin (1972b), 143–67.

Patinkin, D. 1969. Money and wealth: a review article. *Journal of Economic Literature* 7(December), 1140–60.

Patinkin, D. 1972a. Money and wealth. In Patinkin (1972b), 168–94.

Patinkin, D. 1972b. *Studies in Monetary Economics*. New York: Harper & Row.

Patinkin, D. 1976. *Keynes' Monetary Thought: A Study of Its Development*. Durham, North Carolina: Duke University Press.

Pesek, B.P. and Saving, T.R. 1967. *Money, Wealth and Economic Theory*. New York: Macmillan.

Phelps, E.S. 1972. Money, public expenditure and labor supply. *Journal of Economic Theory* 5(August), 69–78.

Pigou, A.C. 1943. The classical stationary state. *Economic Journal* 53(December), 343–51.

Pigou, A.C. 1947. Economic progress in a stable environment. *Economic* 14(August), 180–88. Reprinted in American Economic Association (1951), 241–51.

Scitovsky, T. 1941. Capital accumulation, employment and price rigidity. *Review of Economic Studies* 8 (February), 69–88.

Tanner, J.E. 1970. Empirical evidence on the short-run real balance effect in Canada. *Journal of Money, Credit and Banking* 2(November), 473–85.

Tobin, J. 1975. Keynesian models of recession and depression. *American Economic Review* 65(May), 195–202.

How to cite this article

Patinkin, Don. "real balances." The New Palgrave Dictionary of Economics. Second Edition. Eds.

Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000041> doi:10.1057/9780230226203.1396

The New Palgrave Dictionary of Economics Online

real bills doctrine versus the quantity theory

Timothy S. Fuerst

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The real bills doctrine and the quantity theory of money represent distinct theoretical models of price-level determination and consequently imply different prescriptions for the conduct of monetary policy. The real bills doctrine takes the price level as exogenous and recommends money supply movements that passively respond to the economy. In sharp contrast, the quantity theory insists that the only way to ensure price level stability is by constraining the money supply and not allowing the money supply to move passively in response to economic conditions.

Keywords

central banks; credit; equation of exchange; Federal Reserve System; fiat money; Fisher, I.; gold standard; Great Depression; Hume, D.; inflation; interest rate policy; Law, J.; monetary policy; money supply; neutrality of money; nominal interest rates; quantity theory of money; real bills doctrine; real bills doctrine versus the quantity theory; Ricardo, D.; Smith, A.; Thornton, H.; velocity of circulation

Article

Drawing on two very different hypotheses about the link between nominal money and economic activity, the real bills doctrine and the quantity theory of money represent sharply divergent advice on the conduct of monetary policy. The quantity theory has many prominent advocates, but the real bills doctrine has had a dominant influence in the history and practice of central banking. Further, the real bills doctrine was at the core of the Congressional act creating the US Federal Reserve System so that its importance echoes down to the current day.

The real bills doctrine views money as playing a decidedly passive role, calling for monetary expansion in line with economic activity. According to this view, economic activity is linked to business trade credit and the issuance of short-term debt instruments. Banks should freely purchase these ‘real bills’ with banknote issue, where the modifier ‘real’ refers to short-term debt instruments used to finance

productive activity as opposed to speculation. The doctrine dates to at least 1705 with the publication of *Money and Trade Considered* by John Law, who suggested that banknote issue should be secured by and thus linked to the nominal value of land. The most famous statement of the doctrine is by Adam Smith, whose linkage of note issue to bills of exchange gave the doctrine its name:

> When a bank discounts to a merchant a real bill of exchange drawn by a real creditor upon a real debtor, and which, as soon as it becomes due, is really paid by that debtor; it only advances to him a part of the value which he would otherwise be obliged to keep by him unemployed, and in ready money for answering occasional demands. The payment of the bill, when it becomes due, replaces to the bank the value of what it had advanced, together with the interest. The coffers of the bank, so far as its dealings are confined to such customers, resemble a water pond, from which, though a stream is continually running out, yet another is continually running in, fully equal to that which runs out; so that, without any further care or attention, the pond keeps equally, or very nearly full. (1776, p. 304)

Smith's water-pond metaphor illustrates the real-bills view that note issue would be self-regulating when tied to economic activity, that is, money issue could never be excessive when issued against short-term commercial bills.

The fundamental criticism of the real bills doctrine is that the value of commercial bills (or, in Law's case, the value of land) is tied proportionately to the price level. A commercial bill necessarily includes the dollar value of the goods or services to which it is linked. Thus, under the real bills doctrine, nominal note issue is tied to the nominal price level. If the price level is influenced by the money supply, then we have a circularity problem: nominal prices determine note issue, and note issue affects prices. Henry Thornton first noted the danger of this inflationary circle in his 1802 *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. (David Ricardo was also a prominent opponent of the doctrine.) The thrust of Thornton's criticism was that the real bills doctrine provided no limit on banknote issue. Smith seems to have avoided Thornton's criticism because in Smith's system the gold standard provided an overall restraint on note issue. An excessive banknote issue would result in a bank losing its gold holdings, and see a drain on its ‘coffers’. (See Laidler, 1981; 1984, for a defence of Smith.) But in a world with an inconvertible paper currency Thornton's inflationary critique is devastating.

Humphrey (2001) provides an algebraic description of the real bills doctrine. Suppose that the needs for trade credit are proportional to nominal production, *PY*, where *P* denotes the price level and *Y* denotes real production. The real bills doctrine would imply that banknote issue and thus the money supply (*M*) should be proportionally linked to the needs of trade credit so that we have:

$$M = kPY$$

where *k* is the constant of proportionality between trade credit and nominal production. The Thornton

inflationary critique is now obvious: even with an exogenous level of output (*Y*), there is no way of determining the two endogenous variables, the money supply (*M*) and price level (*P*). A real bills counter-argument would be that the price level is exogenous to money, that is, the money supply has no direct effect on prices. As discussed below, the quantity theory makes the exact opposite claim.

The real bills doctrine and the Great Depression

Remarkably, the real bills doctrine survived Thornton and Ricardo's withering 19th century criticism to find a central place in 20th century US monetary history. In a fascinating account, Meltzer (2003) and Humphrey (2001) trace the flowering of the real bills doctrine into the US Federal Reserve Act of 1913. US Federal Reserve Banks existed for the purpose of ‘accommodating commerce and business’ and were supposed to discount only ‘eligible paper’, which the Act defined as ‘notes, drafts, and bills of exchange arising out of actual commercial transactions’. Although, like Adam Smith, the Act presumed the existence of the gold standard, the real bills doctrine was deemed sufficient even in the absence of a specie constraint. For example, in the *Tenth Annual Report* (1924) of the Board of Governors of the Federal Reserve System, it is noted that ‘there is little danger that the credit created and distributed by the Federal Reserve Banks will be in excessive volume if restricted to productive issues’ (1924, p. 28). The Report further suggested no link between money and prices: ‘The interrelationship of prices and credit is too complex to admit of any simple statement’ (1924, p. 32). Adolph Miller, a founding member of the Federal Reserve Board and co-author of the Report, rejected the notion that ‘changes in the level of prices are caused by changes in the volume of credit and currency…or that changes in the volume of credit and currency are caused by Federal Reserve policy’ (quoted in Meltzer, 2003, pp. 187–8).

Meltzer (2003) convincingly argues that it was this belief in the self-regulating nature of the real bills doctrine that led the Federal Reserve to stand idly by as the US economy spiralled into the Great Depression in the early 1930s. From a real-bills perspective, monetary policy was very loose during these years because Reserve Banks stood ready to discount bills at historically low nominal rates of interest. Meltzer (2003, p. 321) concludes that

> the real bills doctrine implied that the correct policy was a passive one. Most [Federal Reserve] governors had always held these views … The economies of the United States and much of the rest of the world became victims of the Federal Reserve's adherence to an inappropriate theory and the absence of basic economic understanding such as that developed by [Henry] Thornton and [Irving] Fisher.

The quantity theory

In sharp contrast to the real bills doctrine, the quantity theory held as its fundamental principle that the quantity of nominal money (*M*) is largely exogenous and is the principal force determining the endogenous price level (*P*). This argument was first articulated by David Hume (1752). An immediate corollary is that changes in the price level, that is, inflation, are primarily determined by movements in

the supply of money. In the words of the celebrated quantity theorist Milton Friedman (1956, pp. 20–1):

> there is perhaps no other empirical relation in economics that has been observed to recur so uniformly under so wide a variety of circumstances as the relation between substantial changes over short periods in the stock of money and in prices; the one is invariably linked with the other and is in the same direction; this uniformity is, I suspect, of the same order as many of the uniformities that form the basis of the physical sciences.

The quantity theory's causal link between *M* and *P* included the concept of long-run monetary neutrality: exogenous changes in *M* would eventually be exactly matched by proportional changes in *P*. This inference is grounded on the stability of real money demand. In the words of Friedman: ‘The quantity theory is in the first instance a theory of the demand for money’ (1956, p. 4); ‘The quantity theorist accepts the empirical hypothesis that the demand for money is highly stable – more stable than functions such as the consumption function that are offered as alternative key relations’ (1956, p. 16). If we let $L(R, Y)$ denote real money demand as a function of the nominal interest rate (R) and the level of real production (Y), we have a money market equilibrium condition given by:

$$L(R, Y) = \frac{M}{P}.$$

The proportionality hypothesis is then quite clear: for a stable level of L, exogenous changes in M must be matched by changes in P of the exact same magnitude.
The quantity theory also included the concept of short-run non-neutrality. In the words of Hume (1752, p. 38):

> When any quantity of money is imported into a nation, it is not at first disposed into many hands but is confined to the coffers of a few persons, who immediately seek to employ it to advantage … It is easy to trace the money in its progress through the whole commonwealth, where we shall find that it must first quicken the diligence of every individual before it increase the price of labour.

‘There is always an interval before matters be adjusted to their new situation’ (1752, p. 40). Quantity theorists would argue that increases in M are initially met by increases in production (Y) and declines in interest rates (R), but that in the long run R and Y would return to their original levels and that P would thus fully reflect the new higher level of M.
The quantity theory is closely associated with the quantity equation which can be derived as follows. The previous money demand relationship can be re-written as

$$M\frac{Y}{L(R,\ Y)} = PY.$$

If we define the velocity of money as

$$V \equiv \frac{Y}{L(R,\ Y)}$$

then we can write this relationship as the celebrated quantity equation:

$$MV = PY.$$

This is Pigou's (1927) variant of Irving Fisher's (1922) classic equation of exchange. The quantity equation is a useful device for expositing the two central tenets of the quantity theory of money: (*a*) in the long run, output (*Y*) and velocity (*V*) are exogenous to money, so that exogenous movements in the money supply (*M*) are met by proportional movements in prices (*P*), and (*b*) in the short run, movements in the money supply are met by some combination of movements in velocity, prices and output, so that changes in *M* have non-neutral effects on output. The quantity equation can also be used to illustrate Thornton's inflationary critique of the real bills doctrine. For a given level of the nominal rate and an exogenous level of production, velocity is determined by the money demand function, but there is no restriction on the size of *M* or the size of *P*.

The contemporary policy debate

From the vantage point of the outset of the 21st century, there is a sense in which the quantity theory has won numerous intellectual battles but lost the war. Most economists subscribe to the principles of long-run monetary neutrality and short-run non-neutrality. Most would also agree that the quantity equation can be a useful intellectual organizing device. Finally, a standard result in any monetary theory course is the nominal indeterminacy that arises under an exogenous interest-rate operating procedure (for example, Sargent, 1987, ch. 4). This result is just the modern statement of Thornton's 1802 criticism of the real bills doctrine. Hence, it would appear that the quantity theory is in the ascendant.
But remnants of the real bills doctrine are pervasive in both monetary policy implementation and theoretical work. In terms of policy, essentially all central banks in the industrialized world typically ignore or downplay movements in monetary aggregates and instead conduct monetary policy according to an interest rate operating procedure, a close descendant of a real-bills policy. The rationale for such a policy choice is the assertion that the demand for money and thus velocity are unstable. Such a policy

implies seasonal movements in monetary aggregates to accommodate movements in real activity, a passive money supply movement that is directly out of a real-bills playbook.

From a theoretical perspective, there have been two prominent recent contributions in favour of interest rate policy. First, Sargent and Wallace (1982) provide something of a rehabilitation of the real bills doctrine by developing a model in which fluctuating nominal interest rates are harmful, and in which a policy of pegging the nominal interest rate at zero is Pareto efficient. Second, Woodford (2003) has pioneered an effort to conduct monetary policy analysis in ‘cashless’ models – models in which the price level is well defined even though there is no money in the model and the central bank follows an interest-rate operating procedure. We review each of these contributions in turn.

Sargent and Wallace (1982) consider a two-period-lived overlapping-generations model in which fiat money is held even though nominal interest rates are positive because of a legal restriction on private real lending. There are three types of agents: poor savers, rich savers, and borrowers. Using their logarithmic preference specification, the two classes of savers have a constant desired level of savings, say, S^P for the poor and $S^R \gg S^P$ for the rich. The borrowers have a demand for loans given by

$$D^L = \frac{D}{1+r}$$

where r is the real interest rate, and $D > S^R$. (Sargent and Wallace, 1982, consider the case in which the demand for loans fluctuates deterministically, but this is unimportant for their basic result.) The legal restriction is that borrowers cannot issue small-denomination notes. Hence, poor savers cannot lend directly to the borrowers, but can only save by accumulating fiat money. The equilibrium conditions for the money and credit markets are given by:

$$\text{Money market: } S^P = \frac{M_t}{P_t}$$

$$\text{Credit market: } S^R = \frac{D}{1+r_t}$$

where M_t and P_t denote the time-t money supply and price level, and r_t is the real rate of interest. Under what Sargent–Wallace call a ‘quantity-theory’ regime, the central bank keeps the money supply fixed at

some $M_t = M$. In this case, the price level and the real interest rate are constant and calculated from the above equilibrium conditions. This equilibrium is clearly not Pareto optimal as agents do not face the same inter-temporal rate of return – that is, rich savers earn a return of $r > 0$, while poor savers earn a zero real return on currency holdings.

Under a ‘real-bills’ regime the central bank stands ready to lend cash at a zero nominal rate of interest so that

$$(1 + r_t) = \frac{P_t}{P_{t+1}}.$$

In particular, the central bank purchases the ‘real bills’ issued by the borrowers. To finance these purchases the central bank creates the new fiat money denoted by N_t. The borrowers can then use this cash to purchase goods from the poor savers. By purchasing the borrowers' bonds with fiat money, the central bank is effectively opening up an avenue by which poor savers can lend to borrowers. Without this central bank intervention, the positive nominal rates in the credit market are symptoms of a problem – the inability of a fixed money stock to promote proper credit allocation. The real bills equilibrium conditions are given by:

$$\text{Money market: } S^P = \frac{M_t + N_t}{P_t}$$

$$\text{Credit market: } \frac{N_t}{P_t} + S^R = \frac{D}{(P_t / P_{t+1})}.$$

Combining, we have that an equilibrium under the real-bills regime is defined by a price sequence that satisfies:

$$S^P + S^R = \frac{M_t}{P_t} + \frac{D}{(P_t / P_{t+1})}.$$

Solving, we have:

$$P_t = \left(\frac{D}{S^P + S^R}\right)P_{t+1} + \left(\frac{1}{S^P + S^R}\right)M_t.$$

Assuming $D < (S^P + S^R)$, the set of stationary equilibria are given by

$$P_t = \left(\frac{1}{S^P + S^R}\right)\sum_{j=0}^{\infty}\left(\frac{D}{S^P + S^R}\right)^j M_{t+j}$$

where the path of the money supply is free. In the special case in which the money supply grows at a constant rate *g* we have

$$P_t = \left(\frac{1}{S^P + S^R - D(1+g)}\right)M_t.$$

Note that, if *g* becomes large enough, the monetary equilibrium disappears.
Sargent and Wallace restrict the analysis to a particular equilibrium in which the beginning-of-period money supply is held fixed, $M_t = M$ for $t = 0, 1, 2, 3\ldots$. However, the money supply grows and contracts *within* each period as the central bank accommodates the supply of one-period bonds issued by the borrowers (‘real bills’) with the passive expansion of N_t. In this equilibrium the price level is constant and the real return on savings is zero. This equilibrium is Pareto efficient, in contrast to the Pareto inefficiency of the quantity-theory regime. This is an argument in favour of the real bills doctrine and represents Sargent and Wallace's rehabilitation of the doctrine.
There are difficulties with this conclusion. First, the real-bills equilibrium selected by Sargent and Wallace does not Pareto-dominate the quantity-theory regime (rich savers are worse off under the real-bills regime). Second, there is an infinite number of other real-bills equilibria, all defined by the behaviour of the money stock, and not all of these are Pareto efficient. For example, if the money supply grows at a constant rate $g > 0$ the real-bills equilibrium is not Pareto efficient. Finally, Thornton's inflationary critique of the real-bills regime endures: since the money supply is entirely free, there are no restrictions on the short-term and long-term price level.

The second body of recent theoretical work that has a real-bills flavour is provided by Woodford (2003). The title of Woodford's treatise is *Interest and Prices*, a title that makes clear a principal assertion in the work: the money supply is largely irrelevant to price-level determination. The key relationship in the work is the Fisher equation linking nominal rates (i_t) to inflation rates and real rates (r_t):

$$i_t = r_t + p_{t+1} - p_t$$

where p_t is the log of the price level. For simplicity let us suppose that the real rate is exogenous. If the central bank conducts policy according to an exogenous nominal interest rate policy, then the Fisher equation uniquely determines the growth rate of prices (the inflation rate), but not the level of prices. This is, again, the Thornton critique of the real bills doctrine. But Woodford assumes that the central bank follows an endogenous interest rate policy in which the nominal rate responds to movements in prices:

$$i_t = \alpha p_t.$$

Assuming that $\alpha > 0$, the unique stationary equilibrium is given by:

$$p_t = \sum_{j=0}^{\infty} \left(\frac{1}{1+\alpha} \right)^{j+1} r_{t+j}.$$

From a quantity-theory perspective this is a remarkable conclusion: the price level is determined without any mention being made of the money supply. Where is the money demand curve? Either it does not matter (as the money supply moves passively to hit the interest rate target) or it does not even exist (a ‘cashless’ world). Woodford's (2003) analysis thus rejects the quantity theory as a useful guide for policy, and at the same time provides a 21st-century response to Thornton's 19th-century critique of the real bills doctrine: the money supply should be adjusted passively to hit the interest-rate target (as under a real-bills policy), but the interest-rate target should be moved endogenously to ensure price-level stability.

In the intellectual clash of ideas there are typically no clear winners or losers, but instead a synthesis of the combatants. This is surely true of the debate between the real-bills doctrine and the quantity theory of money. Current monetary policy practice and theory has a notable real-bills flavour in the near-universal use of interest rates as the operating target. To repeat, the advantage of such a policy is that it

allows the money supply to respond automatically to and thus accommodate natural movements in real economic activity. But Thornton and the quantity theorists provide a cautionary critique: under an exogenous interest rate policy, there is no way of limiting the inflationary circle between note issue and the price level. To respond to this quantity-theory critique, Woodford (2003) and others have proposed an endogenous interest-rate policy of the form outlined above. This is just one manifestation of the synthesis of the two combatants in this intellectual debate.

See Also

- monetarism
- quantity theory of money
- real bills doctrine

Bibliography

Board of Governors of the Federal Reserve System. 1924. *Tenth Annual Report of the Federal Reserve Board: Covering Operations for the Year 1923*. Washington: Government Printing Office.

Fisher, I. 1922. *The Purchasing Power of Money*, Reprinted, 2nd edn. New York: August M. Kelley, 1963.

Friedman, M. 1956. The quantity theory of money: a restatement. In *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Hume, D. 1752. Of interest; of money. In *Writings on Economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1970.

Humphrey, T.M. 1974. The quantity theory of money: its historical evolution and role in policy debates. *Federal Reserve Bank of Richmond Economic Review* 1974(May/June), 2–19.

Humphrey, T.M. 1982. The real bills doctrine. *Federal Reserve Bank of Richmond Economic Review* 1982(September/October), 3–13.

Humphrey, T.M. 2001. Monetary policy frameworks and indicators for the federal reserve in the 1920s. *Federal Reserve Bank of Richmond Economic Quarterly* 87(1), 65–92.

Laidler, D. 1981. Adam Smith as a monetary economist. *Canadian Journal of Economics* 14, 185–200.

Laidler, D. 1984. Misconceptions about the real bills doctrine: a comment on Sargent and Wallace. *Journal of Political Economy* 92, 149–55.

Law, J. 1705. *Money and Trade Considered*. Edinburgh: Anderson.

Meltzer, A.H. 2003. *A History of the Federal Reserve, Volume I: 1913–1951*. Chicago: University of Chicago Press.

Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32(November), 38–65.

Pigou, A.C. 1927. *Industrial Fluctuations*. London: Macmillan.

Sargent, T.J. 1987. *Macroeconomic Theory*, 2nd edn. Orlando, FL: Academic Press.

Sargent, T.J. and Wallace, N. 1982. The real-bills doctrine vs. the quantity theory: a reconsideration. *Journal of Political Economy* 90, 1212–36.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Indianapolis: Liberty Press, 1976.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. London: LSE Reprint Series, 1939.

Woodford, M. 2003. *Interest and Prices*. Princeton: Princeton University Press.

How to cite this article

Fuerst, Timothy S. "real bills doctrine versus the quantity theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000273> doi:10.1057/9780230226203.1398

The New Palgrave Dictionary of Economics Online

real bills doctrine

Roy Green
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Banking School,; Bullionist controversy; Cantillon, R.; convertibility; Currency School; fractional reserve banking; Fullarton, J.; land bank; law of circulation; law of reflux; Law, J.; mercantilism; overissue; Physiocracy; real bills doctrine; Ricardo, D.; Smith, A.; Steuart, Sir J.; Thornton, H.; Tooke, T.

Article

The ‘real bills doctrine’ has its origin in banking developments of the 17th and 18th centuries. It received its first authoritative exposition in Adam Smith's *Wealth of Nations*, was then repudiated by Thornton and Ricardo in the famous Bullionist Controversy, and was finally rehabilitated as the ‘law of reflux’ by Tooke and Fullarton in the currency–banking debate of the mid-19th century. Even now, echoes of the real bills doctrine reverberate in modern monetary theory.

The central proposition is that bank notes which are lent in exchange for ‘real bills’, that is, titles to real value or value in the process of creation, cannot be issued in excess; and that, since the requirements of the non-bank public are given and finite, any superfluous notes would return automatically to the issuer, at least in the long run. The grounds for rejecting the real bills doctrine have been many and varied. The main counter-argument is that overissue is not merely possible but inevitable in the absence of any external principle of limitation; in this view, commercial wants are insatiable and excess notes would not return to the issuer but undergo depreciation in the exact proportion to their excess.

By the time the real bills doctrine appeared in the economic literature, fractional reserve banking was already well established, releasing unproductive hoards for trade and investment. This did not satisfy John Law, that ‘reckless, and unbalanced but most fascinating genius’ (Marshall, 1923, p. 41n.). He outlined a primitive real bills doctrine in the course of his proposal for a land bank, which would issue paper money on ‘good security’. He imagined, however, that the need for a metallic reserve was superseded by the abolition of legal convertibility, and that *economic* convertibility would *always* be maintained by conformity with the real bills doctrine (Law, 1705, p. 89).

The problem was that, as a mercantilist, Law identified money with capital; he believed that creating paper money was equivalent to increasing wealth. It was his attempt to ‘break through’ the metallic barrier that gave him ‘the pleasant character mixture of swindler and prophet’ (Marx, 1894, p. 441). The spectacular collapse of Law's ‘System’ set off a negative reaction against financial innovation, which was reflected in Cantillon's ‘anti-System’ (Rist, 1940, p. 73) and in Hume's opposition to what he called ‘counterfeit money’ (1752, p. 168). A more positive effect was a shift in the focus of political economy itself to the production process. This shift was led by the Physiocrats and by Adam Smith, whose ‘original and profound’ (Marx, 1859, p. 168) analysis of money and banking was developed in the context of classical value theory.

A decade before the *Wealth of Nations*, Sir James Steuart had attempted to revive Law's ideas from a ‘neo-mercantilist’ viewpoint (1767, book IV, pt. 2). For Smith, by contrast, the role of bank credit was to increase not the quantity of capital but its *turnover* (1776, pp. 245–6; also Ricardo, *Works*, III, pp. 286–7). Output was fixed by the level of accumulation, which for all the classical economists included the speed of its turnover. Credit had the effect both of reducing the magnitude of reserve funds which economic agents needed to hold and of allowing the money material itself – treated as an element of circulating capital and an unproductive portion of the social wealth – to be displaced by paper, thus providing ‘a sort of wagon-way through the air’.

Smith followed Law and Steuart, however, in arguing that an overissue of bank notes could not take place if they were advanced upon ‘real’ bills of exchange, that is, those ‘drawn by a real creditor upon a real debtor’, as opposed to ‘fictitious’ bills, that is, those ‘for which there was properly no real creditor but the bank which discounted it, nor any real debtor but the projector who made use of the money’ (1776, p. 239; also p. 231). When a banker discounted fictitious bills, the borrowers were clearly ‘trading, not with any capital of their own, but with the capital which he advances to them’. When, on the other hand, real bills were discounted, bank notes were merely substituted for a substantial proportion of the gold and silver which would otherwise have been idle, and therefore available for circulation (p. 231). The quantity of notes was thus equivalent to the maximum value of the monetary metals that would circulate in their absence at a given level of economic activity (p. 227).

This development of the classical law of circulation applied to credit and fiduciary money alike, with the difference that in the latter case overissue in the ‘short run’ might result in a permanent depreciation of the paper. By contrast, credit-money, that is, banknotes which were exchanged for real bills, could never be in long-run excess:

> The coffers of the bank … resemble a water-pond, from which, though a stream is continually running out, yet another is continually running in, fully equal to that which runs out; so that, without any further care or attention, the pond keeps always equally, or very near equally full. (p. 231)

Only what Smith called ‘over-trading’ would upset this balance, by promoting excessive credit expansion and an accompanying drain of bullion.

Although the real bills doctrine was accepted by the Bank of England Directors as a guide to monetary management, it was challenged in the bullion controversy following the suspension of cash payments in 1797 as ‘the source of all the errors of these practical men’ (Ricardo, *Works*, III, p. 362; also Thornton,

1802, p. 244 and *passim*). In the view of the ‘bullionists’,

> The refusal to discount any bills but those for *bona fide* transactions would be as little effectual in limiting the circulation; because, though the Directors should have the means of distinguishing such bills, which can by no means be allowed, a greater proportion of paper currency might be called into circulation, not than the wants of commerce could employ but greater than what could remain in the channel of currency without depreciation. (Ricardo, *Works*, III, p. 219)

Indeed, there was no other limit to the depreciation, and corresponding rise in the price level, ‘than the will of the issuers’ (*Works*, III, p. 226).
Nevertheless, the bullionist argument itself was open to challenge, because it confused money with credit. The inconvertible paper of the Bank Restriction was issued not as forced currency but on loan; it was therefore responsible not for increasing the money supply but simply altering its *composition*, by substituting one financial asset for another in the hands of the public. Only when cash payments were restored, however, was any further attempt made to rehabilitate the real bills doctrine, this time as the ‘law of reflux’: provided notes were lent on sufficient security, ‘the reflux and the issue will, in the long run, always balance each other’ (Fullarton, 1844, p. 64; Tooke, 1844, p. 60). The ‘Banking School’ called this law ‘the great regulating principle of the internal currency’ (Fullarton, 1844, p. 68). Their opponents, the ‘Currency School’ orthodoxy, ‘never achieved better than this average measure of security’; and, after all, the average ‘is not to be despised’ (Marx, 1973, p. 131). The real bills doctrine made its next appearance in the Federal Reserve Act of 1913. In banking at least, discretion has always been the better part of valour.

See Also

- Banking School, Currency School, Free Banking School

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en general*. trans. H. Higgs. London: Macmillan, 1931.

Fullarton, J. 1844. *On the Regulation of Currencies*. London: John Murray.

Hume, D. 1752. *Essays, Literary, Moral and Political*. London: Ward, Lock & Co., n.d.

Law, J. 1705. *Money and Trade Considered*. Edinburgh: Anderson.

Marshall, A. 1923. *Money, Credit and Commerce*. London: Macmillan.

Marx, K. 1859. *A Contribution to the Critique of Political Economy*. Moscow: Progress Publishers, 1970.

Marx, K. 1894. *Capital*. vol. 3, Moscow: Progress Publishers, 1971.

Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin.

Ricardo, D. 1951–73. *The Works and Correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.

Rist, C. 1940. *History of Monetary and Credit Theory from John Law to the Present Day*. London: Allen & Unwin.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Routledge, 1890.

Steuart, J. 1767. *An Inquiry into the Principles of Political Oeconomy*. Edinburgh: Oliver & Boyd, 1966.

Thornton, H. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. London: LSE Reprint Series, 1939.

Tooke, T. 1844. *An Inquiry into the Currency Principle*. London: LSE Reprint Series, 1959.

How to cite this article

Green, Roy. "real bills doctrine." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000043> doi:10.1057/9780230226203.1397

The New Palgrave Dictionary of Economics Online

real business cycles

Ellen R. McGrattan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Real business cycles are recurrent fluctuations in an economy's incomes, products, and factor inputs – especially labour – that are due to non-monetary sources. These sources include changes in technology, tax rates and government spending, tastes, government regulation, terms of trade and energy prices. Most real business cycle (RBC) models are variants or extensions of a neoclassical growth model. One such prototype is introduced. It is then shown how RBC theorists, applying the methodology of Kydland and Prescott (*Econometrica* 1982), use theory to make predictions about actual time series. Extensions of the prototype model, current issues and open questions are also discussed.

Keywords

competitive equilibrium; depreciation; Great Depression; home production; household budget constraint; international business cycles; labour supply; labour-market search; Markov processes; monetary shocks; productivity shocks; real business cycles; real exchange rates; regulation; research and development; stabilization policies; stochastic growth models; technical change; technology shocks; terms of trade; total factor productivity; unemployment

Article

Real business cycles are recurrent fluctuations in an economy's incomes, products, and factor inputs – especially labour – that are due to non-monetary sources. Long and Plosser (1983) coined the term ‘real business cycles’ and used it to describe cycles generated by random changes in technology. Other real sources of fluctuations that have been studied include changes in tax rates and government spending, tastes, government regulation, terms of trade, and energy prices.
Kydland and Prescott (1982), who studied the *quantitative* predictions of a stochastic growth model with shocks to technology, found that covariances between model series and autocorrelations of model output were consistent with corresponding statistics for US data. These findings were viewed as surprising, for

two reasons. First, the findings ran counter to the idea that monetary shocks are the driving force behind business cycle fluctuations. Second, the policy implication for Kydland and Prescott's model was that stabilization policies are counterproductive. Fluctuations arise when households optimally respond to changes in technology.

The methodology that Kydland and Prescott (1982) used in their study of business cycles transformed the way in which applied research in macroeconomics is done. For this reason, the term ‘real business cycles’ is often associated with a methodology rather than Kydland and Prescott's original findings. Indeed, the methods of their 1982 paper have been used to study many different sources of business cycles, including monetary shocks.

Most real business cycle (RBC) models are variants or extensions of a neoclassical growth model. One such prototype is introduced. It is then shown how RBC theorists, following Kydland and Prescott (1982), use theory to make predictions about actual time series. Extensions of the prototype model are discussed. Current issues and open questions follow.

Prototype real business cycle model

Households choose sequences of consumption and leisure to maximize expected discounted utility. When aggregated, preferences are defined for a stand-in household that maximizes the expected value of

$$\sum \beta^t u(c_t, 1 - h_t) N_t, \tag{1}$$

where u is the utility function, c_t is per capita consumption at date t, $1–h_t$ is per capita leisure at date t, N_t is the population at date t which grows at rate η, and β is a discount factor.

The technology available in period t is $z_tF_t(K_t, H_t)$, where z_tF_t is the output produced at date t with K_t units of capital and H_t hours. The function F_t has constant returns to scale so that doubling the inputs doubles the output. The variable z_t is a stochastic technology shock assumed to follow a Markov process. The variation in z modelled here is variation in the effectiveness of factor inputs, capital and labour, to produce final goods and services or *total factor productivity* (TFP). Fluctuations in TFP arise from many possible sources. For example, improvements in TFP can arise from new inventions or innovations in existing production processes. Reductions in TFP can arise from increased regulation on producers.

Households are endowed with time each period, normalized without loss of generality to 1, which they can allocate to work or to leisure. They can invest x_t (per capita) in new capital goods. Doing so yields

$$N_{t+1}k_{t+1} = N_t[(1 - \delta)k_t + x_t], \tag{2}$$

where k_t is per capita beginning-of-period t capital, k_{t+1} is per capita end-of-period t capital, and δ is the rate of per period depreciation.

Households face taxes on purchases of consumption and investment and on incomes to capital and labour. With taxation, the household budget constraint in period t is

$$(1 + \tau_{ct})c_t + (1 + \tau_{xt})x_t = r_t k_t - \tau_{kt}(r_t - \delta)k_t + (1 - \tau_{ht})w_t h_t + \psi_t. \quad (3)$$

Variables r_t and w_t are pre-tax payments to capital and labour, respectively. Variables τ_{ct}, τ_{xt}, τ_{kt}, and τ_{ht} are tax rates on consumption, investment, capital, and labour, respectively. These tax rates are assumed to be stochastic and follow a Markov process. Variable ψ_t is the per capita transfer payment at date t made by the government to each household. Total transfer payments are equal to tax revenues less total spending by the government. The per capita spending of the government at date t is g_t.

To derive explicit predictions about the behaviour of these households, it is necessary to first define and then compute an equilibrium for the economy. In doing so, it is convenient to de-trend any variables that grow over time and deal only with stationary processes. To be precise, assume that there is a constant rate of improvement in production processes over time so that $F_t(K_t, H_t) \equiv F(K_t, (1 + \gamma)^t H_t)$ with F homogeneous of degree 1. If the per capita capital stock grows at rate γ and z_t and h_t are stationary, then output grows at rate γ. Certain assumptions on utility and the process for government spending also ensure that components of output grow at rate γ. Denote by $\tilde{v}_t$ the de-trended level of variable v_t, that is, $\tilde{v}_t = v_t / (1 + \gamma)^t$.

A competitive equilibrium is defined as household policy functions for consumption $c(\tilde{k}, \tilde{K}, s)$, investment $x(\tilde{k}, \tilde{K}, s)$, and hours $h(\tilde{k}, \tilde{K}, s)$, where $\tilde{k}$ is the (de-trended) stock of capital for the household, $\tilde{K}$ is the (de-trended) aggregate stock of capital, and $s = (\log z, \tau_c, \tau_x, \tau_k, \tau_h, \log \tilde{g})$; pricing functions $w(\tilde{K}, s)$ and $r(\tilde{K}, s)$; a function governing the evolution of the aggregate capital stock $\tilde{K}' = \Psi(\tilde{K}, s)$ that maps the current state into the capital stock next period $(\tilde{K}')$, and a function $\Phi(s', s)$ governing the transition of the stochastic shocks from s to s' such that (*a*) households maximize the expected value of (1) subject to (2) and (3) with the initial capital stock $\tilde{k}_0$ and functions for prices, aggregate capital, and the transition of s taken as given; (*b*) productive factors are paid their marginal products; (*c*) expectations are rational so that $\tilde{k} = \tilde{K}$ and

$$\Psi(\tilde{k}, s) = [(1 - \delta)\tilde{k} + x(\tilde{k}, s)] / [(1 + n)(1 + \gamma)];$$

and (*d*) markets clear:

$$c(\tilde{k}, \tilde{k}, s) + x(\tilde{k}, \tilde{k}, s) + g(s) = z(s)F(\tilde{k}, h).$$

Note that, in forming expectations about the future, households take processes for prices, tax rates and transfers as given. If households behave competitively, they assume that their own choice of capital next period does not affect the economy-wide level of capital. Therefore, in computing optimal decision functions for the household, it is necessary to distinguish the household's holdings of capital and the aggregate holdings of capital.

Comparing model predictions with data

Given equilibrium functions, properties of the model time series can be compared with data in a straightforward way. Starting with initial conditions on the state, the evolution of the state is determined by functions Ψ and Φ, resulting in sequences $\{\tilde{k}, s\}_{t=0}^{\infty}$ for the state. Equilibrium price and decision functions are then used with these sequences for the state to determine sequences of all prices and allocations.

A standard assumption for the transition $\Phi(s', s)$ is the vector autoregression

$$s_{t+1} = P_0 + Ps_t + Q\varepsilon_{t+1},$$

where each element of ε_t is a normally distributed random variable, independent of the other elements of ε and across time, with mean equal to zero and variance equal to 1. Allowing non-zero off-diagonals in the matrices *P* and *Q* allows for correlations in the elements of the vector *s*. For example, a standard assumption is that tax rates and spending are positively correlated.

If the elements of the matrix QQ' are not large, the equilibrium evolution of the capital stock is well approximated by the following function:

$$\log\tilde{k}_{t+1} = A_0 + A_k\log\tilde{k}_t + B_k s_t,$$

which is linear in the log of the de-trended, per capita capital stock and the stochastic states. Similarly, the logarithms of consumption, investment, output and hours of work can be well approximated as linear

functions of $\log \tilde{k}_t$ and s_t. (See Marimon and Scott, 1999, for an introduction to log-linear methods and nonlinear methods.)

Stacking the results in matrix form yields a system of equations

$$X_{t+1} = AX_t + B\varepsilon_{t+1}$$

$$Y_t = CX_t + \omega_t,$$

where X contains all variables of interest, some of which may not be observable, and Y is a vector of observables. This system can be easily simulated and lends itself nicely to standard methods of estimating model parameters. (See Anderson et al., 1996.)

An important feature of the analysis in Kydland and Prescott (1982) was the construction of the same statistics for the model and for the US data. Employing this methodology requires two necessary steps. The first concerns measurement: data series must be consistent with model series. For example, consumer durable expenditures are investments much like expenditures on new housing. National accountants treat expenditures on durables and housing differently, but the prototype model does not. Thus, revising the national accounts to include services, rents and depreciation of durables is necessary for data and model series to be consistent. The second step of Kydland and Prescott's (1982) methodology concerns reporting: the same statistics should be computed for the model and the revised data. Such comparisons are useful in highlighting similarities and deviations, which are both necessary ingredients to further the development of good theory.

Applying the two methodological tenets to the prototype model and US data yields a number of interesting results. Both the theory and the US data display pro-cyclical movements in consumption and investment, with the movements in investment being far greater in percentage terms. With tax rates and government spending fixed at mean US levels, the theory predicts fluctuations in per capita hours that are too smooth relative to US hours, and a correlation between hours worked and productivity that is too high relative to the correlation in US data. When fiscal shocks consistent with US policy are introduced, the theory predicts movements in per capita hours and a correlation between hours worked and productivity that are in line with the data.

Extensions of the prototype

During the 1980s and 1990s, business cycle research was exploratory but methodologically rooted. Researchers investigated the effects of many different shocks, the mechanisms that propagate them, and the welfare implications – in a consistent way that made clear what factors were important and why. A

brief history is provided here, but interested readers are referred to the volume edited by Cooley (1995) and to a summary of more recent work in King and Rebelo (1999) and Rebelo (2005).

Kydland and Prescott (1982) and Long and Plosser (1983) emphasize technology shocks as an important source of fluctuations. Greenwood, Hercowitz and Huffman (1988) also explore the role of technology shocks for the business cycle but restrict attention to technological changes affecting the productivity of new capital goods and allow for accelerated depreciation of old capital. Mendoza (1995) includes shocks to the terms of trade in an international business cycle model and shows that responses of real exchange rates to productivity shocks and terms-of-trade shocks are quite different, both qualitatively and quantitatively. Braun (1994), Christiano and Eichenbaum (1992), and McGrattan (1994) add fiscal shocks which are important for movement in hours and labour productivity, as noted above. Kim and Loungani (1992) add shocks to energy prices and show that the addition has only a modest impact on the variability of output and hours. Cooley and Hansen (1989) include monetary shocks and a cash-in-advance constraint and show that these additions have negligible effects on business cycle predictions.

The original technology-driven business cycle models under-predicted fluctuations in observed hours and over-predicted the correlation between hours and productivity, leading to further investigations of the model of the labour market and alternative mechanisms for propagating shocks. High – possibly infinite – elasticities were required in the original RBC models to generate fluctuations in aggregate hours comparable to the data. Rogerson (1988) motivates an infinite aggregate elasticity of labour supply in a world with variation in the fraction of people working: individuals work a standard workweek or not at all. This idea is implemented in an RBC model by Hansen (1985), who finds a significant increase in hours fluctuations relative to Kydland and Prescott (1982).

Another factor affecting the labour market is explored by Benhabib, Rogerson and Wright (1991) and Greenwood and Hercowitz (1991) who introduce home production. These researchers show that business cycle predictions depend crucially on the willingness and opportunity of households to substitute time in home work and market work. Under plausible parameterizations, the models do in fact generate greater variability of hours and lower correlations between hours and productivity.

The empirical performance of the RBC model is also improved when labour-market search frictions are introduced, as in Andolfatto (1996) and Merz (1995). Labour-market search models have also been used to study movements in unemployment and vacancies.

Current research and open questions

RBC research has evolved beyond the study of business cycles. The methodology that Kydland and Prescott (1982) introduced is now being applied to central questions in labour, finance, public finance, history, industrial organization, international macroeconomics and trade.

Within business cycle research, some open questions remain. What is the source of large cyclical movements in TFP? This question is especially interesting in the case of the US Great Depression, when TFP declined significantly (Cole and Ohanian, 2004). Are movements in TFP primarily due to new inventions and processes that are, by the nature of research and development, stochastically discovered? Or are movements in TFP primarily due to changing government regulations that may alter the efficiency of production? Are they due to unmeasured investments that fluctuate over time? The answers

matter for policymakers, and they matter for economists who calculate the welfare costs or gains of changing policies.

See Also

- business cycle measurement
- international real business cycles
- monetary business cycle models (sticky prices and wages)
- political business cycles
- welfare costs of business cycles

I wish to thank Gary Hansen, Lee Ohanian and Ed Prescott for their comments on an earlier draft. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

Bibliography

Anderson, E.W., Hansen, L.P., McGrattan, E.R. and Sargent, T.J. 1996. Mechanics of forming and estimating dynamic linear economies. In *Handbook of Computational Economics*, vol. 1, ed. H. Amman, D. Kendrick, and J. Rust. Amsterdam: North-Holland.

Andolfatto, D. 1996. Business cycles and labor-market search. *American Economic Review* 86, 112–32.

Benhabib, J., Rogerson, R. and Wright, R. 1991. Homework in macroeconomics: household production and aggregate fluctuations. *Journal of Political Economy* 99, 1166–87.

Braun, R.A. 1994. Tax disturbances and real economic activity in the postwar United States. *Journal of Monetary Economics* 33, 441–62.

Christiano, L. and Eichenbaum, M. 1992. Current real-business-cycle theories and aggregate labor-market fluctuations. *American Economic Review* 82, 430–50.

Cole, H.L. and Ohanian, L.E. 2004. New Deal policies and the persistence of the Great Depression: a general equilibrium analysis. *Journal of Political Economy* 112, 779–816.

Cooley, T.F. 1995. *Frontiers of Business Cycle Research*. Princeton: Princeton University Press.

Cooley, T.F. and Hansen, G.D. 1989. The inflation tax in a real business cycle model. *American Economic Review* 79, 733–48.

Greenwood, J. and Hercowitz, Z. 1991. The allocation of capital and time over the business cycle. *Journal of Political Economy* 99, 1188–214.

Greenwood, J., Hercowitz, Z. and Huffman, G.W. 1988. Investment, capacity utilization, and the real business cycle. *American Economic Review* 78, 402–17.

Hansen, G.D. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16, 309–27.

Kim, I.-M. and Loungani, P. 1992. The role of energy in real business cycle models. *Journal of Monetary Economics* 29, 173–89.

King, R.G. and Rebelo, S. 1999. Resuscitating real business cycles. In *Handbook of Macroeconomics*, vol. 1B, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Long, J.B., Jr. and Plosser, C.I. 1983. Real business cycles. *Journal of Political Economy* 91, 39–69.

Marimon, R. and Scott, A., eds. 1999. *Computational Methods for the Study of Dynamic Economies*. New York: Oxford University Press.

McGrattan, E.R. 1994. The macroeconomic effects of distortionary taxation. *Journal of Monetary Economics* 33, 573–601.

Mendoza, E.G. 1995. The terms of trade, the real exchange rate, and economic fluctuations. *International Economic Review* 36, 101–37.

Merz, M. 1995. Search in the labor market and the real business cycle. *Journal of Monetary Economics* 36, 269–300.

Rebelo, S. 2005. Real business cycle models: past, present, and future. *Scandinavian Journal of Economics* 107, 217–38.

Rogerson, R. 1988. Indivisible labor, lotteries and equilibrium. *Journal of Monetary Economics* 21, 3–16.

How to cite this article

McGrattan, Ellen R. "real business cycles." The New Palgrave Dictionary of Economics. Second

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000047> doi:10.1057/9780230226203.1399

The New Palgrave Dictionary of Economics Online

real cost doctrine

J. Maloney
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

alternative cost doctrine; Austrian economics; Edgeworth, F. Y.; Jevons, W. S.; labour supply; marginal utility theory; Marshall, A.; real cost doctrine; subjective theory of value

Article

Real cost doctrine is the doctrine that the supply price of a good is the price required to overcome the disutility involved in producing it. The worker, in other words, produces output up to the point at which his (decreasing) marginal utility of income equals his (increasing) marginal disutility of labour. The real cost doctrine can be seen as a half-way house inhabited by economists who had adopted a subjective theory of value but stopped short of the ‘alternative cost’ doctrine whereby the supply price of a resource is equal to its potential earning in its next most productive use. Much of the discussion which took place between English and Austrian economists concerned whether, and to what extent, the two doctrines logically came to the same thing.

Jevons (1871) formulated the real cost doctrine in terms of the diagram in Figure 1. Jevons assumes here (no such assumption is strictly necessary) that the worker at the start of the day not only enjoys his work but that, for a while, his enjoyment increases as he warms up to it. But, as the hours pass, the fatigue and boredom come to predominate over pleasure at an ever-increasing rate. The worker will maximize his surplus of utility over disutility by stopping at point *X* (*ab*=*bc*.)

Figure 1

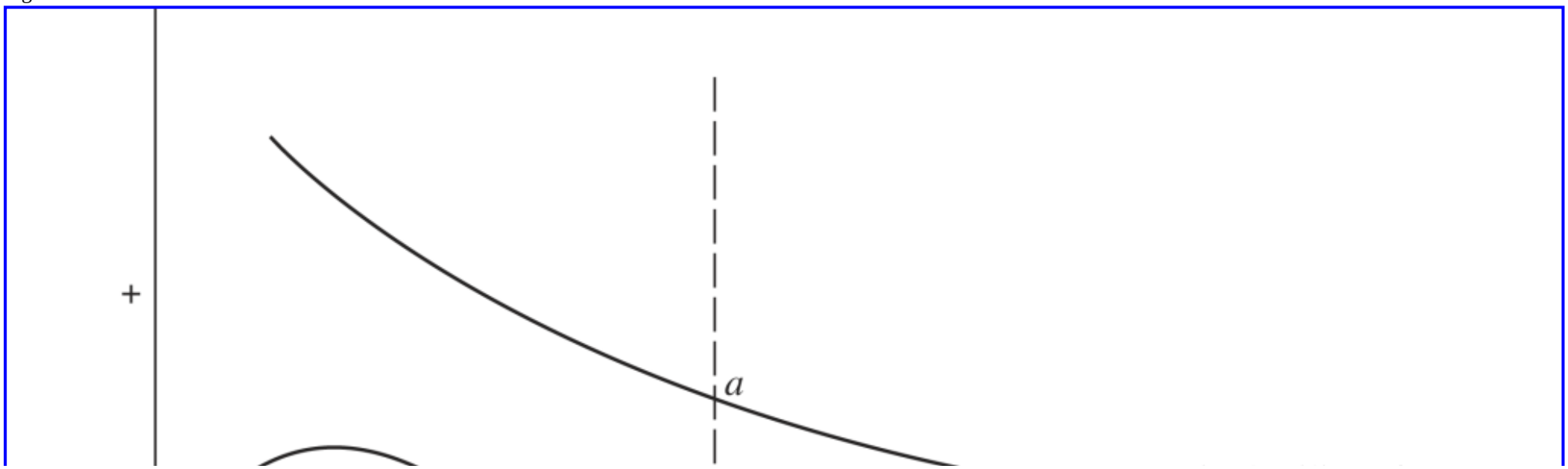

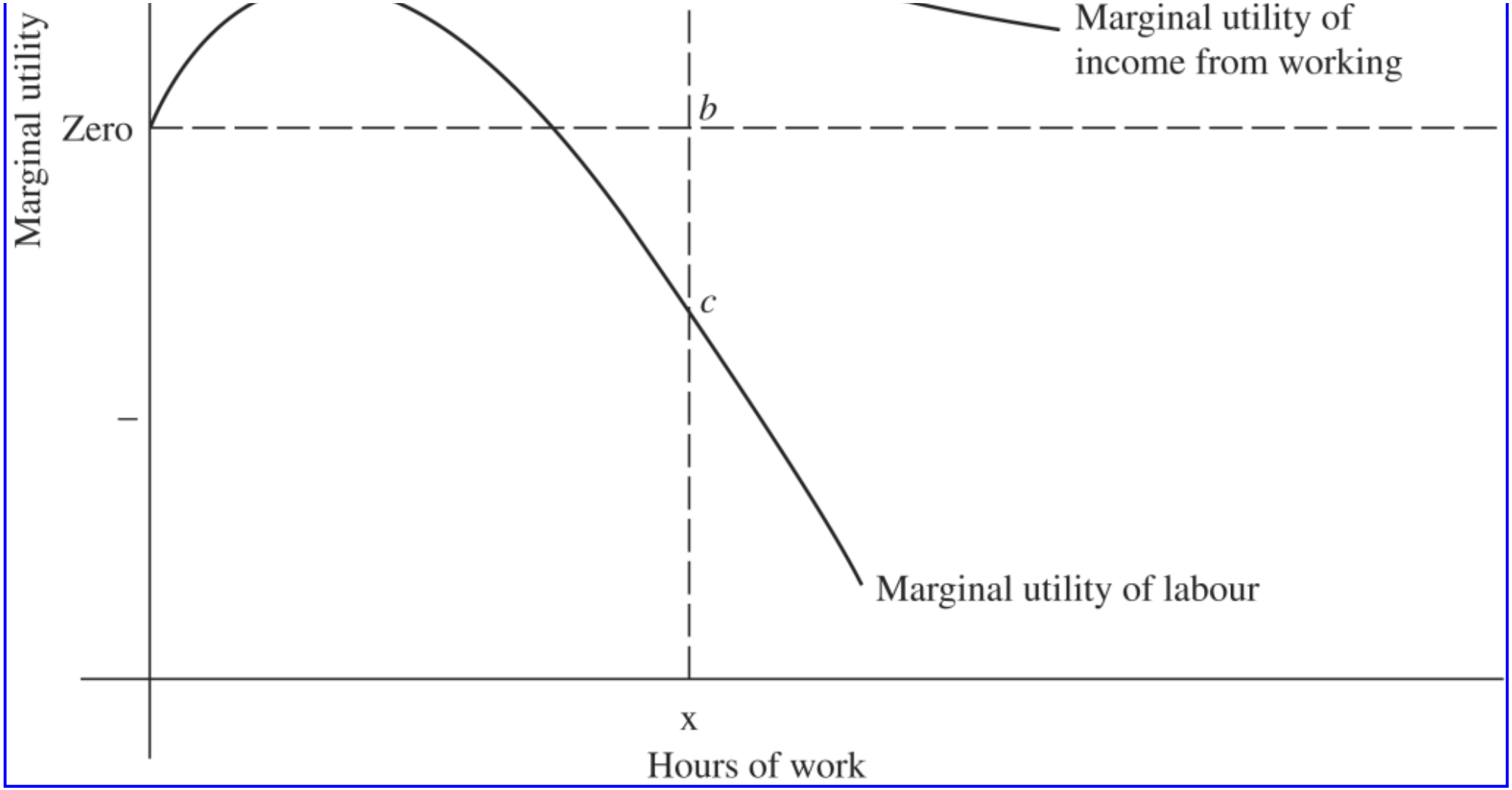

The idea that subjective disutility of labour is central in determining output and price is, perhaps, Jevons's most unquestionably original idea. Not only is it absent from the work of Walras and Menger, but its prefigurations in the classical period are rare and rudimentary when compared with the pre-1871 analyses of marginal utility theory. (Jennings, 1855, points out that marginal disutility of labour increases as the working day progresses but fails to build anything upon it.)

Marshall's theory of price determination, unveiled in his *Principles of Economics* (1890), differs little from Jevons's. Yet what looked radical in Jevons appears almost backward-looking in Marshall. This has something to do with the extension and dissemination of neoclassical principles in the intervening 20 years. But it also stems from a difference of presentation grounded in the contrast between Jevons's impatience with and Marshall's deference towards the Ricardian tradition. Much of Marshall's frequent praise for the English classical economists deftly sidesteps the question of how far they had actually been right. In the *Principles*, however, not only are cost and utility considerations given equal importance when determining price, but the fact that Marshall's conception of cost is ultimately a Jevonian ‘subjective disutility’ one is played down. It receives the strongest emphasis when Marshall argues that the capitalist as well as the worker undergoes real costs in the productive process, the capitalist's cost being that of ‘waiting’ rather than consuming his wealth immediately. (Nassau Senior had invoked Marx's sarcasm by speaking of capitalist ‘abstinence’ in the same context: Marshall tried both to circumvent the ridicule by renaming abstinence ‘waiting’ and to defend Senior from a neoclassical perspective, pointing out that *at the margin* of *aggregate* saving, considerable immediate sacrifice was undoubtedly involved.)

The rival doctrine, that of alternative cost, was espoused principally by the Austrians Wieser and Böhm-Bawerk and advertised in Britain by Wicksteed. All three denied the existence of any such thing as a supply curve, ‘supply’ simply being reverse demand. Böhm-Bawerk cited a horse fair: the buyer's utility from acquiring a horse and the seller's utility from keeping his horse played not just an equal but an identical role in determining price. Hence only a demand curve need be drawn; at the equilibrium price, it crosses the vertical line

representing the fixed stock of horses. Both Marshallian and Austrian analysis predict the same price.

But, of course, the fixed stock of horses makes this a very simple case: we are ignoring the cost of producing them. Such considerations, however, were no problem to the Austrians, who proclaimed that the costs of factors of production and raw materials ultimately depended on utilities from alternative uses forgone. Thus, as regards the labour market, the wage in a particular industry was governed by the demand for labour in other industries. Each worker had to be paid enough to keep him out of his next best paid available job. The Jevonian notion of disutility of labour dropped out of the picture, Böhm-Bawerk (1894) arguing against it on the *empirical* ground that few workers had the chance to make fine adjustments to the length of their working day. To this Edgeworth retorted that the Austrian doctrine implied that individuals made the choice to work or not to work once and for all at the beginning of their careers – it could not handle variations in labour supply due to variations in the wage rate.

The debate as a whole thus seemed to imply that the choice between real cost and alternative cost depended on whether *flexible* labour supply at the *individual* level (assumed by Jevons) or *inflexible* labour supply at the *aggregate* level (implied by the Austrians) was the more objectionable violation of reality. Yet logically the two theories come to exactly the same thing, and are seen to do so as long as the two ‘sides’ make one clarification apiece.

Austrians must make it clear that ‘forgone utility’ includes not only forgone income but also forgone leisure (when you work at all) and forgone non-pecuniary benefits (when you choose a less pleasant but better-paid job in preference to a more pleasant but worse-paid one). Böhm-Bawerk (1894) did spell this out.

Real cost theorists must make it clear that when a baker ponders whether to work another hour, what matters is not the disutility of the work as compared with doing nothing, but the disutility of work as compared with what he would choose to do (it might still be nothing!) if he were not baking. Equally it is not the ‘gross’ marginal utility of income which matters but the marginal utility of the *additional* income gained from spending another hour at the bakery rather than doing something else (other paid work, some leisure activity, or nothing). Edgeworth (1894) *failed* to spell this out; had he done so, a number of economists might have realized sooner than they actually did that both theories ultimately come to the same thing. (See Hobson, 1926, for an example of confusion persisting well into the 20th century.)

See Also

- Marshall, Alfred
- opportunity cost

Bibliography

Blaug, M. 1985. *Economic Theory in Retrospect*. 4th edn. Cambridge: Cambridge University Press.

Böhm-Bawerk, E. von. 1894. One word more on the ultimate standard of value. *Economic Journal* 4, 719–24.

Edgeworth, F.Y. 1894. Professor Böhm-Bawerk on the ultimate standard of value. *Economic Journal* 4, 518–21.

Hobson, J.A. 1926. *Free Thought in the Social Sciences*. London: G. Allen & Unwin.

Jennings, R. 1855. *Natural Elements of Political Economy*. London: Longman, Brown, Green, Longmans.

Jevons, W.S. 1871. *Theory of Political Economy*, ed. and with an introduction by R.D. Collison Black. Harmondsworth: Penguin. 1970.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Wicksteed, P.H. 1910. *The Common Sense of Political Economy*. London: Macmillan.

How to cite this article

Maloney, J. "real cost doctrine." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000048>
doi:10.1057/9780230226203.1400

The New Palgrave Dictionary of Economics Online

real exchange rates

Menzie D. Chinn

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The real exchange rate plays a central role in the open economy. This article describes the various ways in which the real exchange rate has been defined in the literature. It also examines the theoretical and empirical determinants of this variable.

Keywords

exchange rate dynamics; nominal exchange rates; productivity; purchasing power parity; real exchange rates; sticky prices; terms of trade; tradable and non-tradable goods; unit roots

Article

The real exchange rate plays a crucial role in models of the open economy. How the real exchange rate should be defined, how it behaves over time, and what determines it at various time horizons are all questions that have been posed over the years. They have taken on heightened importance in recent years, as the scope of international transactions has expanded and more and more economic activity is either directly or indirectly affected by economic activity in other countries.
The most common definition of the real rate is the nominal exchange rate adjusted by price levels,

$$q_t \equiv s_t - p_t + p_t^* \tag{1}$$

where s is the log exchange rate defined in units of home currency per unit of foreign, and p and p^* are log price levels. If purchasing power parity (PPP) holds, then q is always unity (or a constant if price

indices are used). One should expect PPP to hold in a world where transportation and transactions costs were negligible, consumption baskets were identical, and no arbitrage profits existed. Absent these conditions, the real exchange rate will vary.

One way of thinking about the determinants of movements in the real exchange rate is to appeal to a decomposition. Suppose the price index is a geometric average of traded and non-traded good prices:

$$p_t = \alpha p_t^N + (1 - \alpha) p_t^T \tag{2}$$

where the lower-case letters denote logged values. Then substituting (2) into (1) yields:

$$q_t \equiv (s_t - p_t^T + p_t^{T*}) + [-\alpha(p_t^N - p_t^T) + \alpha^*(p_t^{N*} - p_t^{T*})] \tag{3}$$

$$q_t \equiv q_t^T + [\omega_t] \tag{3'}$$

Equation (3) indicates that the real exchange rate can be expressed as the sum of two components: (i) the relative price of tradables q^T, (ii) the intercountry relative price of non-tradables in terms of tradables in the home country ω.

The determinants of the real exchange rate

If PPP holds only for tradable goods, then only the second term in eq. (3′) can be non-zero, and the relative tradables–non-tradables price is the determining factor in the value of the real exchange rate. Another possibility is that all goods are tradable but not perfectly substitutable; then the imperfect substitutes model results, and q^T is equivalent to q. More generally, both terms on the right hand side of eq. (3′) can take on non-zero values. In either of these cases, there are a large number of variables that could influence each relative price. And of course, there is nothing to rule out both relative price channels as being operative. In popular discussion, all three definitions of ‘the real exchange rate’ are used, sometimes leading to considerable confusion.

Most models of the real exchange rate can be categorized according to which specific relative price serves as the object of focus. If the relative price of non-tradables is key, then the resulting models – in a small country context – have been termed ‘dependent economy’ (Salter, 1959; Swan, 1960) or ‘Scandinavian’ model. In the former case, demand-side factors drive shifts in the relative price of non-tradables. In the latter, productivity levels and the nominal exchange rate determine the nominal wage rate and hence the price level, and thence the relative price of non-tradables. In this latter context, the real exchange rate is a function of productivity (Krueger, 1983, p. 157). Consequently, the two sets of models both focus on the relative non-tradables price but differ in their focus on the source of shifts in this relative price. Since the home economy is small relative to the world economy (hence, one is working with a one-country model), the tradable price is pinned down by the rest-of-the-world supply of traded goods. Hence, the ‘real exchange rate’ in this case is ($p^N – p^T$).

The relative price of tradables definition is most appropriate when considering the relative price that achieves external balance in trade in goods and services. This variable is also what macroeconomic policymakers refer to as ‘price competitiveness’; hence, anything that affects the markup of price over cost – including both the level of demand, input costs, and market structure – can determine the real exchange rate.

Notice the dichotomy between the relative price of tradables and the relative price of non-tradables breaks down when countries specialize in the production of goods. Then the real exchange rate is the same as the terms of trade; purchasing power parity would occur only if the two goods were perfect substitutes (see Lucas, 1982; Stockman, 1980).

Empirical modelling and results

Real exchange rate dynamics

In one special case, there is no need to model the real exchange rate. If relative PPP is *assumed* to hold, then q is a constant. Empirically, this is clearly not true in the short run but could be in the long run. Consequently, tremendous effort has been invested in investigating whether q is trend stationary, even though trend stationarity is not the same as purchasing power parity holding (the stronger condition of mean stationarity is required). Numerous studies have evaluated the trend stationarity of q directly by application of unit root tests, or indirectly by assessing whether the component series of q exhibit common long-term trends. Broadly speaking, the conclusions in this literature are mixed. Generally, panel methods, long time samples, and the use of producer or wholesale price indices provide more evidence in favour of a trend stationarity q than do pure time series methods, short samples, and the use of consumer price indices (see Rogoff, 1996; Taylor and Taylor, 2004). These results leave open the possibility that economic variables affect the movement of exchange rates over the short as well as the long run.

Modelling real exchange rate movements as a function of economic variables

The modelling of the real exchange rate determinants can be divided into two main categories. The first category includes models of the nominal exchange rate which, by virtue of the assumption of sticky

prices, become models of the real exchange rate. First and foremost among these are sticky price monetary models that incorporate exchange rate overshooting, such as Dornbusch (1976) and Frankel (1979). In the long run, purchasing power parity holds, so that these models are only short-run models. The second category includes models that focus on the determinants of the long-run real exchange rate. By far dominant in this category are those that centre on the relative price of non-tradables. These include the specifications based on the approaches of Balassa (1964) and Samuelson (1964) that model the relative price of non-tradables as a function of sectoral productivity differentials, including Hsieh (1982), Canzoneri, Cumby and Diba (1999) and Chinn (1999; 2000). They also include those models that search more broadly and include demand-side determinants of the relative price, such as DeGregorio and Wolf (1994). Engel (1999) has cast doubt upon the relevance of the relative non-tradables price. He demonstrates that for the G-7 economies, the variability of q^T as proxied by the tradable components of the CPI is comparable to the variability of ω even at horizons of 15 years. More recently, some version of the portfolio balance model has been resurrected. Lane and Milesi-Ferretti (2002) have forwarded a model wherein the real rate depends upon net foreign assets. Early panel evidence in favour of the importance of this factor is to be found in Gagnon (1996).

Some methodological approaches do not fall neatly into one or the other category. The analysis by Mark and Choi (1997) is one instance. They compare the usefulness of monetary and real factors in predicting exchange rate changes over long horizons, and find – surprisingly – that monetary factors have persistent effects on the real exchange rate. Using a different methodology, namely, a structural vector autoregression, Clarida and Gali (1995) find that monetary and demand-side factors dominate in the determination of exchange rates. Also relying upon a structural (permanent-transitory) decomposition involving the real exchange rate and the current account, Lee and Chinn (2006) find that positive permanent shocks (interpreted as productivity innovations) tend to appreciate the currency and (at least for the United States) have an impact comparable in magnitude to those of temporary shocks.

See Also

- cointegration
- exchange rate dynamics
- monetary business cycle models (sticky prices and wages)
- nominal exchange rates
- purchasing power parity
- real exchange rates
- terms of trade
- tradable and non-tradable commodities
- unit roots

Bibliography

Balassa, B. 1964. The purchasing power parity doctrine: a reappraisal. *Journal of Political Economy* 72,

584–96.

Canzoneri, M.B., Cumby, R.E. and Diba, B. 1999. Relative labor productivity and the real exchange rate in the long run: evidence for a panel of OECD countries. *Journal of International Economics* 47, 245–66.

Chinn, M. 1999. Productivity, government spending and the real exchange rate: evidence for OECD countries. In *Equilibrium Exchange Rates*, ed. R. MacDonald and J. Stein. Boston: Kluwer.

Chinn, M. 2000. The usual suspects: productivity and demand shocks and Asia-Pacific real exchange rates. *Review of International Economics* 8, 20–43.

Clarida, R. and Gali, J. 1995. Sources of real exchange rate movements: how important are nominal shocks? *Carnegie-Rochester Conference Series on Public Policy* 41, 9–66.

DeGregorio, J. and Wolf, H. 1994. Terms of trade, productivity, and the real exchange rate. Working Paper No. 4807. Cambridge, MA: NBER.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84, 1161–76.

Engel, C. 1999. Accounting for US real exchange rate changes. *Journal of Political Economy* 107, 507–38.

Frankel, J. 1979. On the mark: a theory of floating exchange rates based on real interest differentials. *American Economic Review* 69, 610–22.

Gagnon, J. 1996. Net foreign assets and equilibrium exchange rates: panel evidence. International Finance Discussion Papers No. 574. Washington, DC: Board of Governors of the Federal Reserve System.

Hsieh, D. 1982. The determination of the real exchange rate: the productivity approach. *Journal of International Economics* 12, 355–62.

Krueger, A.O. 1983. *Exchange-Rate Determination*. Cambridge: Cambridge University Press.

Lane, P.R. and Milesi-Ferretti, G.M. 2002. External wealth, the trade balance, and the real exchange rate. *European Economic Review* 46, 1049–71.

Lee, J. and Chinn, M. 2006. Current account and real exchange rate dynamics in the G-7 countries. *Journal of International Money and Finance* 25, 257–74.

Lucas, R. 1982. Interest rates and currency prices in a two-country world. *Journal of Monetary Economics* 10, 335–59.

Mark, N. and Choi, D.Y. 1997. Real exchange rate prediction over long horizons. *Journal of International Economics* 43, 29–60.

Rogoff, K. 1996. The purchasing power parity puzzle. *Journal of Economic Literature* 34, 647–68.

Salter, W.A. 1959. Internal and external balance: the role of price and expenditure effects. *Economic Record* 35, 226–38.

Samuelson, P. 1964. Theoretical notes on trade problems. *Review of Economics and Statistics* 46, 145–54.

Stockman, A. 1980. A theory of exchange rate determination. *Journal of Political Economy* 88, 673–98.

Swan, T. 1960. Economic control in a dependent economy. *Economic Record* 36, 51–66.

Taylor, A.M. and Taylor, M.P. 2004. The purchasing power parity debate. *Journal of Economic Perspectives* 18(4), 135–58.

Williamson, J. 1994. *Estimating Equilibrium Exchange Rates*. Washington, DC: Institute for International Economics.

How to cite this article

Chinn, Menzie D. "real exchange rates." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000051> doi:10.1057/9780230226203.1401

The New Palgrave Dictionary of Economics Online

real rigidities

David Romer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Real rigidities are forces that reduce the responsiveness of firms' profit-maximizing prices to variations in aggregate output resulting from variations in aggregate demand. Real rigidities make firms less inclined to take actions that dampen movements in aggregate output, and so increase the responsiveness of output to disturbances. They appear essential to any successful explanation of short-run macroeconomic fluctuations. As a result, various forms of real rigidity pervade modern models of business cycles.

Keywords

adjustment costs; aggregate demand; business cycles; capital-market imperfections; cyclical markups; efficiency wages; elasticity of substitution; imperfect competition; input–output analysis; labour mobility; labour supply; menu costs; nominal rigidities; real business cycles; real rigidities; staggered price setting; sticky prices; strategic complementarity

Article

‘Real rigidities’ is the name given to a large class of business cycle propagation mechanisms. Real rigidities appear essential to any successful explanation of business cycles.

The definition of real rigidities

Although the term ‘real rigidities’ appears vague, it in fact refers to a precise concept. Consider an economy of symmetric price-setting firms that is at its flexible-price equilibrium, and suppose that the money supply increases with prices unchanged, so that aggregate output increases. Now ask by how much a representative firm would want to increase its price if it faced no barriers to nominal price adjustment. By definition, the smaller the amount the firm would want to increase its price in response to

a given increase in aggregate output, the greater the degree of real rigidity.

One can see the meaning of ‘real rigidity’ more formally by observing that, in the experiment described above, the profits of the firm that is considering changing its price, neglecting any costs of price adjustment, typically can be written in the form $V(p_i - p, y)$, where p_i is the firm's price, p is the aggregate price level, and y is the departure of output from its flexible-price level (all in logs). In most models of this type, $V(\cdot)$ is a smooth function. The first-order condition for the profit-maximizing price is $V(p_i^* - p, y) = 0$ (subscripts denote partial derivatives). At the flexible-price equilibrium, $p_i^* = p$ and y=0. Starting from that equilibrium, the derivative of the representative firm's desired relative price, $p_i^* - p$, with respect to y is thus

$$\left.\frac{d(p_i^* - p)}{dy}\right|_{p_i^* - p = 0,\ y = 0} = \frac{V_{12}(0, 0)}{-V_{11}(0, 0)} \equiv \varphi.$$

For the flexible-price equilibrium to be stable, ϕ must be positive. By definition, a lower value of ϕ corresponds to greater real rigidity. Note that real rigidity is defined entirely in terms of relations among real variables: it refers to the (lack of) responsiveness of desired real prices to aggregate real output.

The definition of real rigidity in models without symmetric price-setting firms is analogous: any force that reduces the amount that price setters would change their relative prices in response to movements in aggregate output that are the result of changes in aggregate demand is a real rigidity.

Real rigidities and business cycles

Real rigidities are crucial to business cycles. At a general level, real rigidities make firms less inclined to take actions that dampen movements in aggregate output. As a result, they increase the responsiveness of output to disturbances.

The importance of real rigidities is easiest to see in a static model where firms face fixed costs of changing prices. Consider the model sketched above, with two extensions. First, replace the profit function with a second-order approximation around the flexible-price equilibrium. This implies that the representative firm's loss in profits from failing to charge its profit-maximizing price (neglecting costs of price adjustment) is $K(p_i - p_i^*)^2$, where $K \equiv -V_{11}$. It also implies that the representative firm's profit-maximizing price is given by $p_i^* - p = \varphi y$, where ϕ is as defined before. Second, assume that each firm faces a fixed cost $Z > 0$ of changing its nominal price.

The economy begins at its flexible-price equilibrium. We want to know by how much output can change in response to a change in aggregate demand before firms change their prices. Non-adjustment is an equilibrium as long as the representative firm's losses from failing to adjust are less than Z. Prior to the shock, the representative firm's price equals the aggregate price level, p. If the firm adjusts its price, it sets it to the new profit-maximizing level, $p + \varphi y$. Thus the condition for nonadjustment to be an

equilibrium is $K[p-(p+\varphi y)]^2 < Z$, or $|y| < 1/(\varphi\sqrt{Z/K})$. Thus, when ϕ is lower – that is, when real rigidity is greater – the range over which aggregate demand shocks affect real activity is greater.
Real rigidities are not just important to models with nominal rigidity, however. Consider, for example, the following minimalist real business cycle model. The markets for labour and goods are perfectly competitive, and the representative firm's production function is $y_i=a+n_i$ (y is output, a is productivity, and n is labour input, again all in logs). Labour supply is $n = \gamma w$, $\gamma > 0$, where w is the (log) real wage. In this model, the elasticity of profit-maximizing relative prices to demand-driven output fluctuations (that is, to variations in y with a fixed, which must come from variations in n) equals the elasticity of the real wage with respect to y, which is $1/\gamma$. Thus a larger value of γ corresponds to greater real rigidity. The production function implies that labour demand is perfectly elastic at $w=a$. Labour-market equilibrium therefore requires that $n = \gamma a$. A larger value of γ therefore implies that productivity shocks have a larger impact on employment, and thus that the output effects of the shocks are magnified to a greater extent. Thus, even in this purely Walrasian model of fluctuations, real rigidity acts as a propagation mechanism.
Real rigidities also act as amplification mechanisms in dynamic models of price adjustment. Consider an economy with barriers to price adjustment where output is above its flexible-price level, and suppose that some firms have an opportunity to change their prices, and that their new prices will be in effect for more than one period. Greater real rigidity increases the persistence of the departure of output from its flexible-price level, for three reasons. First, as in the static model of price adjustment, it reduces the benefits of price adjustment, and so makes firms more inclined not to adjust at all. Second, the fact that only some firms have the opportunity to adjust means that output will continue to be above its flexible-price level. Greater real rigidity then implies that the firms that adjust will respond by less, thus drawing out the period of above-normal output. Third, the fact that other firms will be in the same situation when they adjust their prices means that they will adjust by less, which in turn dampens the adjustments of the firms that adjust immediately.
There is a close link between real rigidity and strategic complementarity in profit-maximizing prices. If we assume the stylized aggregate demand curve $y = m - p$ (where m reflects factors that shift aggregate demand), then the expression for the representative firm's profit-maximizing relative price, $p_i^* - p = \varphi y$, implies $p_i^* = \varphi m + (1-\varphi)p$. Thus greater real rigidity corresponds to greater strategic complementarity in desired prices: when ϕ is lower, each firm wants its price to move more closely with other prices.
Real rigidity and strategic complementarity in desired prices are not identical, however. To see this, suppose the aggregate demand equation is instead $y = \beta(m-p)$, $\beta > 0$. Then $p_i^* = \varphi\beta m + (1-\varphi\beta)p$. Thus β affects strategic complementarity but not real rigidity. And it is real rigidity that is key to cyclical fluctuations. Nonetheless, because of the close link between the two concepts, and because many business cycle models assume $y = m - p$, the terms real rigidity and strategic complementarity in prices are often used interchangeably.

Types of real rigidities

Since any force that reduces the responsiveness of profit-maximizing relative prices to demand-driven

output fluctuations is a real rigidity, there are many possible real rigidities. Some might not be commonly thought of as ‘rigidities’. For example, as the simple real business cycle model shows, more elastic labour supply is a type of real rigidity.

Such neoclassical sources of real rigidity, however, are almost surely not strong enough to generate output fluctuations of the size and nature that we observe. In Walrasian models, the real wage is likely to rise sharply with the quantity of labour. For this not to occur, either the long-run elasticity of labour supply must be high or the intertemporal elasticity of substitution in labour supply must be high and short-run aggregate fluctuations must have a large transitory component. Neither of these conditions appears to hold in practice. In models of nominal disturbances and barriers to nominal price adjustment, the result is large incentives for price changes, and thus little nominal rigidity. In productivity-driven real business cycle models, the result is small movements in labour input, so that the dynamics of aggregate output largely mimic the dynamics of the underlying productivity shocks. Researchers have therefore turned their attention to non-Walrasian sources of real rigidity.

It appears difficult to understand substantial employment fluctuations without non-Walrasian real rigidities in the labour market. At a general level, what is needed is for some force causing workers to be off their labour supply curves, at least in the short run, so that the cyclical behaviour of the real wage is not governed by the elasticity of labour supply. For example, if there is equilibrium unemployment because of efficiency wages, the cyclical behaviour of the real wage depends on how the efficiency wage varies with aggregate output. As a result, the real wage can be (though it need not be) less procyclical than in a Walrasian labour market, with the result that fluctuations in employment and output are greater.

A more subtle real rigidity in the labour market arises if labour is imperfectly mobile in the short run (because of search frictions, for example), so that each firm faces an upward-sloping labour supply curve rather than perfectly elastic supply at the economy-wide wage. Consider, for example, a firm contemplating cutting its price and increasing production in a recession. With imperfectly mobile labour, this requires paying a higher real wage. Thus the amount the firm wants to reduce its price is smaller – that is, real rigidity is greater.

There can also be important real rigidities in other markets. In the goods market, forces making desired markups countercyclical act as real rigidities. When desired markups are more countercyclical, then, for a given degree of procyclicality of real marginal costs, desired movements in relative prices are smaller. Countercyclical desired markups can stem from a variety of sources. One simple but potentially important possibility is that, when economic activity is greater, firms' incentives to disseminate information and consumers' incentives to acquire it are greater, and so demand is more elastic.

Another feature of goods markets that can act as a real rigidity is input-output links among firms. If the prices charged by intermediate suppliers are sticky, the costs that firms face for intermediate inputs tend to rise by less than the suppliers' costs in a boom, thereby reducing the amount that firms would raise their prices if they were free to do so.

Capital-market imperfections can also create real rigidities. Capital-market imperfections can cause financing costs to be countercyclical, as higher output increases cash flow (and hence firms' ability to use internal finance) and raises asset values (and hence increases collateral and reduces the cost of external finance). With one component of costs countercyclical, desired prices are less procyclical. To give another example, financial difficulties in recessions can increase the importance of short-term profits to firms relative to expanding their customer base, and so can make desired markups

countercyclical.

There is an important distinction between two broad categories of real rigidities. One category consists of forces, such as limited short-run labour mobility among firms, that increase real rigidity by affecting what happens when one firm changes its prices and others do not. The other consists of forces, such as factors that reduce the procyclicality of the real wage, that increase real rigidity by affecting what happens when all firms' output moves together. In terms of the definition of real rigidity as $V_{12}(0,0)/[-V_{11}(0,0)]$, the first category consists of forces raising $-V_{11}(0,0)$, and the second consists of forces reducing $V_{12}(0,0)$.

The distinction between these two categories is important for two reasons. First, real rigidities that result from forces that affect what happens when one firm changes its price with other firms' prices fixed are not relevant to the properties of business cycle models with identical firms and fully flexible prices. Second, the two types of real rigidities have different microeconomic implications. Most importantly, factors that increase real rigidity by affecting what happens when one firm changes its price and others do not increase the costs of departures from the profit-maximizing price; as a result, they typically predict smaller movements in firms' relative prices in response to many types of shocks.

Selected literature

Ball and Romer (1990) establish that in a static setting, imperfect competition and barriers to nominal adjustment alone are unlikely to generate substantial nominal rigidity. They show the general importance of real rigidities to static menu-cost models and stress that forces making desired real wages relatively unresponsive to output fluctuations are likely to be essential to generating substantial nominal rigidity (see also Blanchard and Fischer, 1989, ch. 8). Earlier work by Akerlof and Yellen (1985) incorporates substantial real rigidity in a model of price stickiness, although it does not explicitly analyse the importance of real rigidity to the results. Haltiwanger and Waldman (1989) show that strategic complementarity magnifies the impact of non-responders on equilibrium outcomes, a result that is closely related to the finding that real rigidities magnify the effects of barriers to price adjustment. Kimball (1995) establishes the central role of real rigidities to the persistence of output fluctuations in models of staggered price adjustment, and stresses the importance of the distinction between forces that affect what happens when all firms' outputs move together and forces that affect what happens when one firm changes its price with other firms' prices fixed (see also Blanchard, 1987). Klenow and Willis (2006) show the differing microeconomic implications of the two categories of real rigidities. Romer (2006, ch. 6) provides a general discussion of the importance of real rigidities to static and dynamic models of nominal rigidity, catalogues many specific real rigidities and provides numerous references. Real rigidities pervade modern business cycle models. In real business cycle models, for example, such common features as indivisible labour supply, variable capital utilization and labour hoarding, and learning-by-doing (see, for example, Rogerson, 1988; Burnside and Eichenbaum, 1996; and Chang, Gomes and Schorfheide, 2002) magnify the effects of disturbances precisely because they are real rigidities. In models with price stickiness, some important recent analyses where real rigidities play a central role include Mankiw and Reis (2002), Gertler and Leahy (2006) and Carvalho (2006).

See Also

- cyclical markups
- monetary business cycle models (sticky prices and wages)
- new Keynesian macroeconomics
- real business cycles
- sticky wages and staggered wage setting

Bibliography

Akerlof, G.A. and Yellen, J.L. 1985. A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100, 823–38.

Ball, L. and Romer, D. 1990. Real rigidities and the non-neutrality of money. *Review of Economic Studies* 57, 183–203.

Blanchard, O.J. 1987. Aggregate and individual price adjustment. *Brookings Papers on Economic Activity* 1987(1), 57–109.

Blanchard, O.J. and Fischer, S. 1989. *Lectures on Macroeconomics*. Cambridge. MA: MIT Press.

Burnside, C. and Eichenbaum, M. 1996. Factor-hoarding and the propagation of business-cycle shocks. *American Economic Review* 86, 1154–74.

Carvalho, C. 2006. Heterogeneity in price stickiness and the real effects of monetary shocks. *Frontiers of Macroeconomics* 2(1), Article 1.

Chang, Y., Gomes, J.F. and Schorfheide, F. 2002. Learning-by-doing as a propagation mechanism. *American Economic Review* 92, 1498–520.

Gertler, M. and Leahy, J. 2006. A Phillips curve with an Ss foundation. Working Paper No. 11971. Cambridge, MA: NBER.

Haltiwanger, J. and Waldman, M. 1989. Limited rationality and strategic complements: the implications for macroeconomics. *Quarterly Journal of Economics* 104, 463–83.

Kimball, M.S. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27, 1241–77.

Klenow, P.J. and Willis, J.L. 2006. Real rigidities and nominal price changes. Working Paper No. 06-03.

Kansas City, MO: Federal Reserve Bank of Kansas City.

Mankiw, N.G. and Reis, R. 2002. Sticky information versus sticky prices: a proposal to replace the new Keynesian Phillips curve. *Quarterly Journal of Economics* 117, 1295–328.

Rogerson, R. 1988. Indivisible labor, lotteries and equilibrium. *Journal of Monetary Economics* 21, 3–16.

Romer, D. 2006. *Advanced Macroeconomics*, 3rd edn. New York: McGraw-Hill.

How to cite this article

Romer, David. "real rigidities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000271> doi:10.1057/9780230226203.1402

The New Palgrave Dictionary of Economics Online

real wage rates (historical trends)

Robert C. Allen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Historical studies of the real wage allow us to track the divergence in the world of economics since the Middle Ages and changes in the distribution of income during the Industrial Revolution. Before the 19th century, the real wage moved inversely to the population. Since then it has increased dramatically.

Keywords

Black Death; capital accumulation; capitalism; consumer price index; economic development; factor prices; great divergence; index numbers; Industrial Revolution; inequality in wages; international migration; international trade; Malthus's theory of population; national income accounting; price histories; productivity; real wage; Solow, R.; standards of living; subsistence; surplus labour; technical change; wage ladder

Article

The real wage indicates the purchasing power of a worker's income. The real wage is the ratio of the nominal wage (what someone is actually paid) to a price or price index. Sometimes that price is the price of the product of a competitive firm, in which case the real wage is the marginal product of labour and has a productivity interpretation. In the more common case, however, and the one this article deals with, the price deflator is a consumer price index. In this case, the real wage measures the standard of living of the worker. Since that bears on central questions of economic growth and distribution, real wages have been an important tool for measuring and interpreting economic growth and stagnation over the last millennium.

Measuring real wages raises practical problems that are particularly acute in historical investigations. First, one needs information about wages, prices and spending shares to perform any calculations. Data sources for these have to be developed, which ultimately involves extensive archival work. There are conceptual problems as well since many people in the past received some income in kind as well as cash and many were also employed on piece rates that must be converted to earnings before an assessment of their purchasing power can be made. Second, an index number must be chosen to compute the consumer price index. While theorists have advanced many useful arguments about why some formulae are better than others, data limitations often force compromises. One of the most extreme was the once common practice of deflating wages by the price of grain. Other products were ignored, as well as the inconvenient fact that most people in the West ate bread, not grain. Most recent studies have avoided this practice. Third, new products and improvements in the quality of old products bedevil historical studies as they do modern ones. Although product innovation was less common in the past, the creation of the global economy led to the introduction into Europe or mass availability of maize, potatoes, tomatoes, chilli peppers, sugar, tobacco, cotton cloth, tea, coffee and porcelain. Also, comparisons of real wages between continents with radically different diets raise the question of new products in a cross-sectional context. How do you compare the standard of living of an English worker eating bread, beef and beer with a Chinese worker eating fish and rice?

Real wages and economic growth in developed countries

Economic theorists have divided the history of the world into two phases. Before the onset of modern economic growth around 1800, income per head grew very slowly, if at all.

Increases in productivity simply resulted in more people. The real wage moved inversely with the population and remained constant in the long run. This was the Malthusian phase of history.

The second phase began in about 1800. Technology improved steadily raising income per head. Population growth was restrained, so an increase in the labour force did not swamp the increase in labour demand. Consequently, the real wage rose in step with productivity. This has been called the Solow phase in view of Solow's (1956) growth model. While these models can be nuanced, as we will see, they provide a starting point for real wage history: is it consistent with these models?

We can measure the real wage over the past 800 years thanks to the accumulated research of historians who have written ‘price histories’ of cities since the mid-19th century. The price historian finds an institution like a college or hospital that has existed for centuries and examines its accounts to abstract the prices of the things it bought and the wages it paid. Oxford and Cambridge colleges were the first to be studied (Rogers, 1866–92) and since then many European cities have been investigated. Phelps-Brown and Hopkins (1955; 1956) were the first to take advantage of this material and construct a real wage index for English building workers from 1264 to 1954. More recently, Allen (2001) and Clark (2005) have reworked this material and added new evidence to compute new real wage series. While there are differences among these authors on issues like real wage change in the Industrial Revolution, the broad outlines of the real wage story are the same (Figure 1).

Figure 1

Real wage in London, 1200–2000. *Sources*: Data before 1914: Allen (2001). Later wage and price indices: Phelps Brown and Hopkins (1955), Mitchell (1998), ILO (various years).

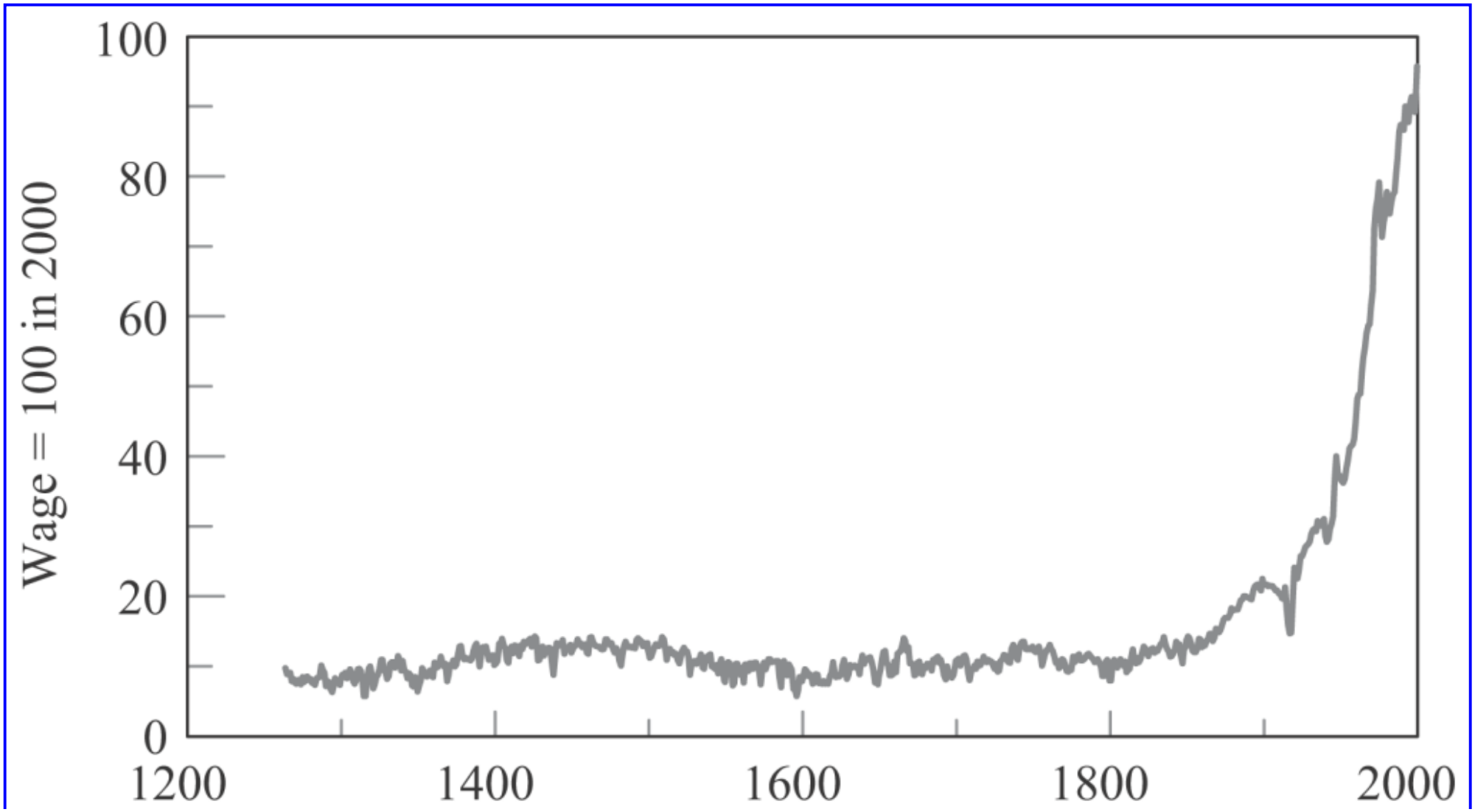

The Malthusian and Solow phases stand out in this figure. Before the 19th century, England followed the Malthusian pattern. There was no long-run trend in the real wage, although there were fluctuations. These coincided with population swings; in particular, the real wage rose after the Black Death in 1348–9, which killed about one-third of the population, and fell in the 16th century as the population started to rebound. Real wages only rose above the pre-industrial peak once the Industrial Revolution was well under way or completed. The rise since then has been spectacular by comparison, and today the real wage is ten times its level in the pre-industrial world. This is the Solow phase of economic history.

The period of the Industrial Revolution, roughly 1770–1860, is something of a problem. Were real wages rising then or falling? The classical economists, who were writing in the early 19th century, were pessimistic about the prospects of workers. While they agreed that capitalism was likely to cause output per capita to rise, they also believed that real wages would remain constant at ‘subsistence’. For Malthus, Ricardo and other mainstream economists, the reason was demographic: wages were the income of the bulk of the population, and a higher wage would lead them to have more children and live longer. The result would be an increase in the workforce that would push wages back to subsistence. While radicals like Marx and Engels rejected the demographic model, they agreed that wages would not rise under capitalism. Their explanation, however, turned on the demand for labour rather than its supply. Marx and Engels believed that technological progress would be so rapid and so labour-saving that the demand for labour would always fall short of the supply – again forcing wages back to subsistence. Only collective action or state interference would prevent this. None of the classical economists, in other words, expected the ‘Malthusian economy’ to transmute into the ‘Solow economy’.

By the 20th century, it was clear that these arguments were wrong, for living standards were far higher than they had been 100 years before, as Figure 1 shows. Kuznets (1955) raised the possibility that inequality went through an ‘invented U’ patterned during economic development. In his model, this worked through the wage structure itself. At the outset, workers were in low-productivity, low-wage sectors. As the modern sector grew, it attracted workers by offering higher wages. Inequality increased as workers moved to that sector since those employed there were earning more than their counterparts in agriculture. Inequality in wages declined as the process of labour reallocation was completed, for all workers were then earning the higher wage paid in the modern sector.

The problem of economic development in poor countries provoked Lewis (1954) to propose a model of growth and distribution that was more classical in spirit and that emphasized the differential movements of output per head and the real wage. Lewis divided the economy into two sectors. In the traditional sector, consisting of peasant agriculture and the urban ‘informal’ economy, the main inputs were land and labour, and the latter was in surplus. In the modern, industrial sector, output was produced with capital and labour, and the former was scarce. Growth occurred as capital was accumulated and the modern sector expanded. Its labour force was drawn from the traditional sector. Surplus labour in that sector meant that the marginal product of labour was low, perhaps even zero, and income was shared and at a subsistence level. An elastic supply of labour kept the real wage in the modern sector at subsistence even though output per worker was rising. This process of rising inequality would continue, in Lewis's view, until the modern sector had expanded to absorb all the surplus labour. Only then would the real wage rise in step with output per worker.

How well do these theories fare in practice? The question has been extensively researched and vigorously debated in the case of the British Industrial Revolution. Lindert and Williamson (1983) were the first to apply modern economic methods to the question. They computed economy-wide average earnings and a consumer price index founded on budget surveys and corresponding prices. Their conclusion was guardedly ‘optimistic’ in that they found the average real wage rose sharply after 1815. This conclusion was not universally accepted. Feinstein (1998) computed an alternative price index that significantly reduced the rate of real wage growth leading to his title ‘pessimism perpetuated’. Clark (2005), on the other hand, proposed yet another price index that tilted the conclusions back in a Lindert–Williamson direction. Most recently, Allen (2007a) has plumped for ‘pessimism preserved’.

These disagreements reflect the limitations of the data, which are only a poor reflection of the ideal information discussed above. There were no comprehensive and representative samples of consumer spending, and even the annual series of individual prices are problematic. Quality change, in particular the growing use of cotton rather than wool in clothing, is only imperfectly grasped with the available information. There is considerable scope for contradictory – yet plausible – readings of the evidence.

The impact of economic development on wage rates has been pursued for many other countries with mixed results. Over the long term, real wage change in Western Europe has followed a pattern like that for England shown in Figure 1 (Scholliers and Zamagni, 1995). The United States has been repeatedly studied, and revisions to price and wage indices have been as thoroughgoing for the USA, as they have been for Britain. For instance, Douglas's (1930) conclusion that real wages only rose by eight per cent during the boom from 1890 to 1914 was overturned by Rees (1961), who found that real wages increased by 40 per cent. Over the long term, of course, real wages have risen dramatically in America, but the real wage lagged behind GDP per head during early industrialization, according to Margo's (2000) study of the period 1800–60.

Outside the advanced Organisation for Economic Co-operation and Development (OECD) countries, the link between economic growth and real wage advance is much weaker. The economic boom experience by Tsarist Russia, for instance, was not reflected in urban or rural real wages (Allen, 2003a). Latin America enjoyed a substantial rise in GDP per head in the 20th century, with only an elusive impact on real wages. In Mexico, which has been studied more than most countries, there were periods when real wages surged and others when real wages collapsed. The declines look about as big as the gains, but the uncertainty arising from the introduction of new goods makes definitive conclusions hazardous (Bortz and Aguila, 2006).

Real wages and the great divergence

The difference in real wages can be computed between two places as well as between two times, and the geographical dimension has allowed real wages to play an important role in the ‘great divergence’ debate. Since 1800, incomes have grown most rapidly in the most prosperous countries, so their lead over the poor countries has increased. Charting this divergence is the first step in explaining it. Real GDP per head may be the best indicator, and economists have tried to extrapolate it far into the past. Errors, however, accumulate, and the estimates become increasingly problematic the further back one goes. Real wages provide an alternative, and simpler, approach to the problem. The real wage is individual in its focus – what could a particular worker buy with his or her income? – and so avoids the economy-wide assumptions of national income accounting. The real wage also requires fewer data, and it can be computed directly for dates in the past without the need to extrapolate backwards.

Indeed, the classical economists, who established a long-standing view on the subject, expressed Europe's lead in terms of real wages. Adam Smith (1776, pp. 74–5, 91, 187, 206) saw the world in terms of a wage ladder: workers in England and the Netherlands had a higher standard of living than workers in France or elsewhere on the European continent. Workers in Asia lagged behind Europeans. ‘The real price of labour, the real quantity of the necessaries of life which is given to the labourer…is lower both in China and Indostand than it is through the greater part of Europe.’

This view has recently been challenged by scholars of Asia who have argued that Asia was as prosperous as Europe at the time Smith wrote. According to Pomeranz (2000, p. 49), ‘It seems likely that average incomes in Japan, China, and parts of South-East Asia were comparable to (or higher than) those in western Europe even in the late eighteenth century.’

How does Smith's wage ladder stand up in terms of modern evidence? The price histories of European cities provide a start, for they allow us to compute real wage differences across Europe from the late Middle Ages to the 19th century (van Zanden, 1999; Allen, 2001). While today real wages are similar across Western Europe, the calculations show that the last time this was even approximately true was in the late 15th century. Between 1500 and 1750, real wages in Amsterdam and London, the booming maritime cities of north-western Europe, were trendless, while they fell sharply in other parts of Europe under the impact of population growth not offset by economic expansion (Allen, 2003b). Incomes had diverged in Europe, in the manner described by Smith, before the Industrial Revolution. Indeed, it was decades, if not a century, before modern economic growth was perceptible in the real wage data. So far as Europe was concerned, the great divergence preceded the Industrial Revolution rather than being its sequel.

What about Europe and Asia? It is only very recently that comparisons have been made across the continents. Parthasarathi's (1998) study of England and India supported the revision view, but a broader collection of data supports Smith's assessment (Allen, 2007b). Comparisons with Japan and China also show that real wages there were like those of the backward parts of Europe. Even the Yangtze Delta, the most advanced region in China, had real wages on a par with those in Milan, not London or Amsterdam (Allen et al., 2007). While the Ottoman Empire has not received as much attention as east Asia in the revisionist historiography, Özmucur and Pamuk (2002) have shown that the real wage in Istanbul was also like that in Italy.

The worldwide conclusions require comparisons across regions with radically different diets. The comparisons are made by computing the cost of Smith's ‘quantity of necessaries of life which is given to the labourer’. Figure 2 shows full-time, full-year earnings for a labourer deflated by the cost of maintaining a family on a mainly carbohydrate diet yielding 1,920 calories per adult male equivalent. In each region, the cheapest available carbohydrate is used for the calculation. A value of 1 indicates that the labourer's wage equalled this ‘bare bones’ subsistence, and, indeed, that was the case in the 18th century in much of Europe and Asia. Living standards were higher, however, in Amsterdam and London.

Figure 2

Subsistence ratio for labourers, various world cities, 1375–1875. Income/cost of subsistence basket. *Sources*: Allen (2001; 2007b); Allen et al. (2007).

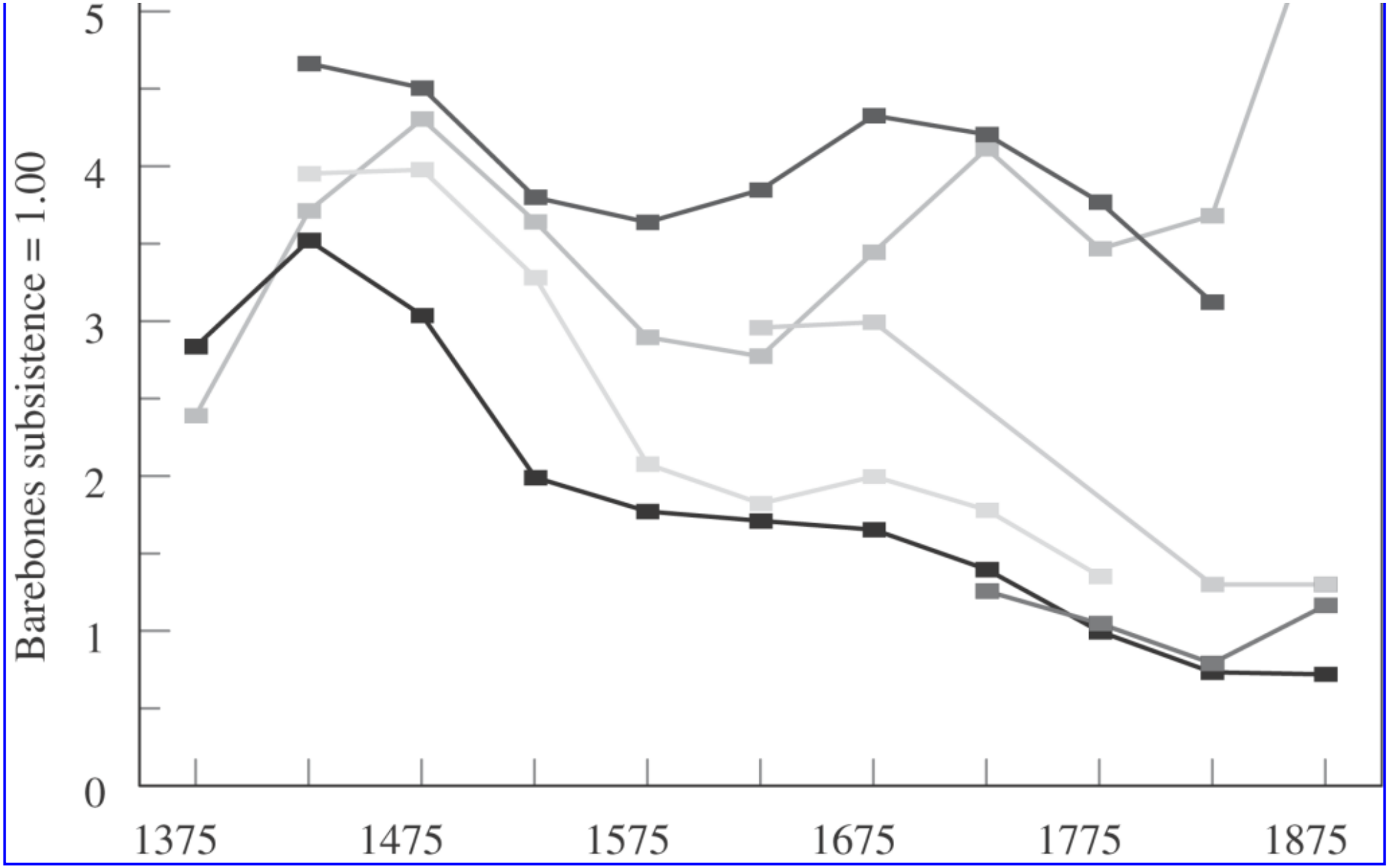

Real wages and globalization

The history of the global economy has attracted attention and real wages have played an important role in exposing its properties. Research on the 19th and 20th centuries has aimed to establish trends in real wages across countries as well as over time (Allen, 1994; Williamson, 1995). O'Rourke and Williamson (1999) argued that international trade and migration tightly bound economies and determined their relative factor prices. Trends in real wages, in other words, were determined by the evolution of the global economy rather than by the internal forces of capital accumulation and technical change that most previous theories have emphasized. In a study of the British economy, O'Rourke and Williamson (2005) argued that international factors determined factor prices from 1850 onwards and perhaps from as early as 1750. The relative importance of internal and external factors in determining the

real wage is a lively area of current research.

See Also

- economic growth in the very long run
- Industrial Revolution

Bibliography

Allen, R.C. 1994. Real incomes in the English-speaking world, 1879–1913. In *Labour Market Evolution: The Economic History of Market Integration, Wage Flexibility and the Employment Relation*, ed. G. Grantham and M. MacKinnon. London: Routledge.

Allen, R.C. 2001. The great divergence in European wages and prices from the middle ages to the First World War. *Explorations in Economic History* 38, 411–47.

Allen, R.C. 2003a. *Farm to Factory: A Reassessment of the Soviet Industrial Revolution*. Princeton: Princeton University Press.

Allen, R.C. 2003b. Poverty and progress in early modern Europe. *Economic History Review* 56, 403–43.

Allen, R.C. 2007a. Pessimism preserved: real wages in the British industrial revolution. Working Paper No. 314. Department of Economics, Oxford University.

Allen, R.C. 2007b. India in the great divergence. In *The New Comparative Economic History: Essays in Honor of Jeffery G. Williamson*, ed. T.J. Hatton, K.H. O'Rourke and A.M. Taylor. Cambridge, MA: MIT Press.

Allen, R.C., Bassino, J.-P., Ma, D., Moll-Murata, C. and van Zanden, J.-L. 2007. Wages, prices, and living standards in China, Japan, and Europe, 1738–1925. Working Paper No. 316. Department of Economics, Oxford University.

Beveridge, L. 1965. *Prices and Wages in England from the Twelfth to the Nineteenth Century*. London: Frank Cass.

Bortz, J. and Aguila, M. 2006. Earning a living: a history of real wage studies in twentieth-century Mexico. *Latin America History Review* 41, 112–38.

Boskin, M.J., Dulberger, E., Gordon, R., Griliches, Z. and Jorgenson, D. 1996. *Toward a More Accurate Measure of the Cost of Living*. Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index. Washington, DC: Government Printing Office.

Clark, G. 2005. The condition of the working class in England, 1209–2004. *Journal of Political Economy* 113, 1307–40.

Diewert, W.E. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4, 115–45.

Douglas, P.H. 1930. *Real Wages in the United States: 1890–1926*. Boston: Houghton Mifflin.

Feinstein, C.H. 1998. Pessimism perpetuated: real wages and the standard of living in Britain during and after the Industrial Revolution. *Journal of Economic History* 58, 625–58.

ILO (International Labour Organization) Various years. *Yearbook of Labour Statistics*. Geneva: ILO.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45, 1–28.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economics and Social Studies* 22, 139–91.

Lindert, P.H. and Williamson, J.G. 1983. English workers' living standard during the Industrial Revolution: a new look. *Economic History Review* 36, 1–25.

Margo, R.A. 2000. *Wages and Labor Markets in the United States 1820–1860*. Chicago: University of Chicago Press.

Mitchell, B.R. 1998. *International Historical Statistics: Europe, 1750–1993*. London: Macmillan Reference.

O'Rourke, K.H. and Williamson, J.G. 1999. *Globalization and History*. Cambridge, MA: MIT Press.

O'Rourke, K.H. and Williamson, J.G. 2005. From Malthus to Ohlin: trade, industrialisation and distribution since 1500. *Journal of Economic Growth* 10, 5–34.

Özmucur, S. and Pamuk, S. 2002. Real wages and standards of living in the Ottoman Empire, 1489–1914. *Journal of Economic History* 62, 293–321.

Parthasarathi, P. 1998. Rethinking wages and competitiveness in the eighteenth century: Britain and south India. *Past & Present* 158, 79–109.

Phelps Brown, E.H. and Hopkins, S.V. 1955. Seven centuries of building wages. *Economica* NS 22, 195–206.

Phelps Brown, E.H. and Hopkins, S.V. 1956. Seven centuries of the prices of consumables, compared with builders' wage rates. *Economica* NS 23, 296–314.

Pomeranz, K. 2000. *The Great Divergence: China, Europe, and the Making of the Modern World*. Princeton: Princeton University Press.

Rees, A. 1961. *Real Wages in Manufacturing, 1890–1914*. Princeton: Princeton University Press.

Rogers, J.E.T. 1866–92. *A History of Agriculture and Prices in England*, vols. 7. Oxford: Clarendon Press.

Scholliers, P. and Zamagni, V. 1995. *Labour's Reward: Real Wages and Economic Change in 19th- and 20th-Century Europe*. Aldershot: Edward Elgar.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. E. Cannan. New York: The Modern Library, 1937.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

van Zanden, J.L. 1999. Wages and the standard of living in Europe, 1500–1800. *European Review of Economic History* 3, 175–97.

Williamson, J.G. 1995. The evolution of global labor markets since 1830: background evidence and hypotheses. *Explorations in Economic History* 32, 141–96.

How to cite this article

Allen, Robert C. "real wage rates (historical trends)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000275> doi:10.1057/9780230226203.1403

The New Palgrave Dictionary of Economics Online

realized volatility

Torben G. Andersen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Realized volatility is a fully nonparametric approach to *ex post* measurement of the actual realized return variation over a specific trading period. It encompasses specific empirical procedures and an associated continuous-record asymptotic theory for arbitrage-free jump diffusions. It provides the ideal model-free benchmark for volatility model performance evaluation, and it has numerous natural areas of application within financial economics.

Keywords

affine models of the term structure; conditional return variance; derivative prices; econophysics; forecasting; historical volatility; integrated variance; jumps; martingales; microstructure noise; parametric time series models; realized volatility; return volatility; stochastic volatility models; yield curve

Article

Return volatility is critical for a range of issues in financial economics. In theory, an asset price reflects its return covariation with economy-wide risk factors, often captured through its covariance with returns of factor-replicating financial portfolios, including the broad (stock) market. Hence, assessments of asset pricing, fund performance and portfolio allocation are all directly linked to expectations of the future volatility and covariability of financial assets. Likewise, individual asset volatilities are key inputs to derivatives pricing and risk management. Finally, in recent years volatility *realizations* have become the object of direct contracting as the payoff on so-called volatility swaps is determined by the future value attained by a specified measure of return volatility over the contract horizon.
As a consequence, return volatility has been studied extensively in the literature. Until recently, there were two dominant paradigms. One uses parametric time series models within the GARCH (Engle, 1982; Bollerslev, 1986) or (genuine) stochastic volatility (Shephard, 2005) class to obtain conditional

return variance estimates and forecasts. A second exploits market prices of volatility sensitive contracts, such as options, to back out the (implied) expected future volatility. Although the latter approach also conditions on a specific pricing model, it embodies a wider information set as market prices reflect the views of market participants, not just historical returns. However, derivative prices carry premiums for bearing volatility risk and thus provide a less direct measure of expected (physical) volatility. Hence, these measures are complementary and each likely provides independent information. More importantly in this context, they both focus on an a priori concept of return volatility, largely synonymous with the conditional variance. This is appropriate for many purposes as financial decisions are made subject only to current expectations regarding the future market environment. However, such measures are identified only through specific parametric representations. Moreover, the *ex ante* expectations differ from the subsequent (random) volatility realizations. The latter may be assessed only from *ex post* model-free measurements of the actual return variation. Such measures are obviously useful for assessing (volatility) model performance. In addition, if accurate measures of realized volatility are available it is natural to exploit these directly for modelling and forecasting. With increasing availability of intra-day tick-by-tick trade and quote data, this perspective has gained in popularity and a voluminous literature is evolving on the approach. This article presents a brief overview of these developments and associated empirical applications.

Historical volatility

The concept of realized volatility refines and extends the *historical volatility* measure which has a fairly long precedent in the literature. To make the argument transparent, we initially consider an extremely simplified environment. Assume we are given the daily closing logarithmic asset price, denoted p_t. The associated daily continuously compounded return is then $r_t=p_t-p_{t-1}$. Moreover, assume the returns are conditionally mean zero with i.i.d. standardized residuals, that is,

$$r_t = \sigma_t \cdot z_t, \text{ with } z_t \sim \text{iid}(0, 1) \text{ and } Var(z_t^2) = \omega. \tag{1}$$

The goal of realized volatility measurement is to provide a model-free estimator for the return variation or volatility, σ_t^2, given only the concurrent return observations. Obviously, if volatility is time-varying, this is problematic. We have, at the daily level, conditional on current volatility,

$$E[r_t^2] = Var(r_t) = \sigma_t^2, \quad Var(r_t^2) = \omega \cdot \sigma_t^4, \quad \text{and} \quad E[r_t^2] \,/\, [Var(r_t^2)]^{1/2} = \omega^{-1/2}. \tag{2}$$

Hence, the concurrent squared return is an unbiased estimator of the underlying return variance. Unfortunately, it is extraordinarily noisy. The signal-to-noise ratio, defined as the mean of the estimator relative to the standard deviation, equals $\omega^{-1/2}$. Invariably, $\omega > 1$ at daily (and lower) frequencies, so the standard deviation of (estimated) realized volatility exceeds the expected value. Matters improve if we assume constant volatility, at σ_t^2, over the month representing, say, *K* daily returns. Letting $r_{t:K} = r_t + \ldots + r_{t+K}$ denote the monthly return, we may exploit the historical volatility indicator, $[r_t^2 + \ldots + (r_{t+K})^2]$ rather than simply $(r_{t:K})^2$. We then have,

$$Var(r_{t:K}) = E[r_t^2 + \ldots + (r_{t+K})^2] = K \cdot \sigma_t^2, \text{ and } Var[r_t^2 + \ldots + (r_{t+K})^2] = K \cdot \omega \cdot \sigma_t^4. \quad (3)$$

The signal-to-noise ratio for the monthly realized volatility is $(K/\omega)^{1/2}$ or a factor $K^{1/2}$ larger than for eq. (2). Equation (3) may also readily be converted into an estimator for daily volatility based on the sample mean of the daily squared return over the month. This estimator is consistent with convergence rate $K^{1/2}$. However, given the simplifying assumptions, this measure is best viewed as an informal gauge of the underlying level of volatility. It was applied for computation of annual realized volatility from monthly data by Officer (1973) and monthly volatility series from daily data by, amongst others, Merton (1980), and French, Schwert and Stambaugh (1987).

The two equivalent interpretations of eq. (3) have distinct properties within a more realistic time-varying volatility setting. Since it is untenable to assume constant volatility for a month, or even one day, estimation of daily volatility from a surrounding set of daily returns covering, say, one month is inherently problematic. Nonetheless, the monthly realized volatility measure is robust to variation in volatility as explained more formally below. Of course, given the rapidly evolving markets, we would often need to assess the time variation in realized volatility at a daily level. The above reasoning suggests this will require access to high-frequency intra-day data. This is the starting point for the modern realized volatility literature.

Realized volatility as an ex post return variability measure

Given the round-the-clock activity on financial markets, volatility is naturally seen as evolving stochastically in continuous time. A complete characterization of the volatility realization then consists of a full specification of the actual sample path. However, at most one price is observed at each point in time so *instantaneous* volatility cannot be assessed without relying on adjacent observations, which is justified only under auxiliary assumptions – just as daily volatility cannot be estimated from a single return. In contrast, realized volatility seeks to measure the temporally cumulated (instantaneous) volatility so the target is the (average) realization of volatility over a non-negligible interval, allowing for a feasible and robust estimator with desirable properties.

As a benchmark for analysis, let the logarithmic asset price, $p(t)$, be a continuous time stochastic process observed in a frictionless market. For brevity and clarity, the formal exposition is cast in a univariate setting, but all results for realized volatility generalize readily to the multivariate case. To avoid arbitrage, and subject only to weak auxiliary conditions, the price process constitutes a (special) semi-martingale (Harrison and Kreps, 1978; Back, 1991). The price process may then quite generally be represented as follows,

$$dp(t) = \mu(t)dt + \sigma(t)dB(t) + j(t)dJ(t),\ t \in [0, T], \tag{4}$$

where $\mu(t)$ is a predictable, continuous process with bounded variation, the volatility process $\sigma(t)$ is strictly positive, $B(t)$ denotes a standard Brownian motion, $J(t)$ is a jump indicator taking the values zero (no jump) or unity (jump) and, finally, the $j(t)$ indicates the jump in the return process if a jump occurs at time t and $j(t) \equiv 0$ otherwise. We assume the jump intensity, denoted $\lambda(t)$, to be bounded so there is a finite number of jump in the price path per time period. This is standard in the asset pricing literature although it does rule out some valid Lévy representations.

Equation (4) implies an instantaneous expected return of $\{\mu(t) + \lambda(t)E[j(t)]\}dt$, which is an order smaller than the instantaneous innovation, $\sigma(t)(dt)^{1/2}+j(t)$ namely order dt versus $dt^{1/2}$. Hence, for short horizons, the return variability is dominated by the unpredictable martingale component and the mean return is negligible. These features are captured formally by the notion of return quadratic variation defined below.

We denote the discretely observed continuously compounded return at time t, based on price observations at times t and $t–h$, for $h>0$, by $r(t, h) = p(t) - p(t-h)$. From equation (4), the h-period return then has the representation,

$$r(t, h) = p(t) - p(t-h) = \int_{t-h}^{t} \mu(\tau)d\tau + \int_{t-h}^{t} \sigma(\tau)dB(\tau) + \sum_{t-h \le \tau < t} j(\tau). \tag{5}$$

Formally, the sample path variation of the return process over $[t–h, t]$ is given by the *quadratic variation* of the logarithmic price process,

$$QV(t, h) \equiv \int_{t-h}^{t} \sigma^2(s)ds + \sum_{t-h < s \le t} j^2(s) \equiv IV(t, h) + \sum_{t-h < s \le t} j^2(s). \tag{6}$$

Equation (6) attributes the return variation to the diffusive volatility and the cumulative squared jumps. The first term is denoted the *integrated variance*. This quantity is often the focus of the broader realized volatility literature as many studies ignore jumps. The integrated variance is also of direct relevance for option pricing under stochastic volatility (Hull and White, 1987). This is in part due to the following result for the special case with neither jumps nor correlation between the return and volatility processes, that is, $B(t)$ is independent of $\sigma(s)$ for all $0 \leq s, t \leq T$. Conditional on the mean component and integrated variance, we then have,

$$r(t, h) \mid \{\mu(t, h), IV(t, h)\} \sim N(\mu(t, h), IV(t, h)), \tag{7}$$

where $\mu(t, h) = \int_{t-h}^{t} \mu(\tau)\, d\tau$. Since (innovations to) the mean component is of smaller order than the integrated variance for low values of h, the dominant feature is the time-varying second moment given by the realizations of integrated variance. Hence, the return distribution is a normal mixture governed by the integrated variance process. If return and volatility innovations are correlated, the distribution is no longer mixed normal, but the interpretation of the integrated variance as a return variability measure is maintained.

Of course, short of having a continuum of price observations available, the relevant quantities in eqs (6) or (7) are not directly observable. However, in theory, the quadratic variation can be approximated closely by the corresponding cumulative squared return process, motivating the following definition. Let $[t-h,t]$ be split into $M=h/\Delta$ sub-intervals of length Δ, with $0 < \Delta \ll h$, and define the *realized volatility* constructed from the equally spaced Δ-period returns as

$$RV(t, h; \Delta) = \sum_{m=1}^{M} r^2(t - h + m \cdot \Delta, \Delta)\,. \tag{8}$$

The basic theory for semi-martingales ensures that realized volatility is consistent for quadratic variation in the sense that, for finer and finer sampling of intra-day returns, eq. (8) will, in the limit, provide a perfect measure of the realizations of the latent quadratic variation, that is,

$$RV(t, h; \Delta) \to QV(t, h), \text{ as } \Delta \to 0 \,(\text{and } M = h/\Delta \to \infty). \tag{9}$$

This provides a formal basis for *ex post* measurement of realized volatility without parametric assumptions. It is a model-free measure of actual realizations while standard approaches provide parametric model forecasts of future volatility. We have the following approximate relationship between parametric forecasts and realized volatility, for small *h*,

$$Var[r(t,h)|\mathcal{F}_{t-h};\theta] \simeq E[RV(t,h;\Delta)|\mathcal{F}_{t-h}] \simeq E[QV(t,h)|\mathcal{F}_{t-h}], \tag{10}$$

where the left most expression denotes the conditional variance over [*t*–*h*, *t*] conditional on the available information at time *t*–*h*, $\mathcal{F}_{t-h}$, and the (true) model parameter vector, θ. Andersen, Bollerslev and Diebold (ABD) (2008) provide an in-depth discussion of the approximation behind eq. (10). The relation shows that realized volatility is *the* natural benchmark for assessing volatility forecast performance. Realized volatility can in principle be used to estimate the instantaneous volatility of a pure diffusion consistently. Ruling out jumps in the price process, but allowing volatility to be caglad (left-continuous, right limit sample path) and thus having potential discontinuities, we have

$$QV(t,h) \to \sigma^2(t), \text{ as } h \to 0. \tag{11}$$

This insight is certainly not new. Merton (1976; 1980) discusses the result explicitly, and Foster and Nelson (1996) develop asymptotic results. However, this limiting operation is impractical. Equation (10) merely requires that $M=h/\Delta \to \infty$. For eq. (11) to hold, with quadratic variation replaced by a feasible realized volatility estimator, Δ must converge to zero at a faster rate than *h*. This requires a double limiting procedure with ever more data sampled within an ever shrinking neighbourhood of *t*. Intensive sampling over short intervals magnifies the microstructure effects stemming from price discreteness, bid-ask bounce, temporary order-driven dependencies and other institutional features affecting returns at the highest frequencies. The issue of how to deal with such complications has inspired an extensive literature, summarized succinctly in Hansen and Lunde (2006). The practical complications induced by microstructure noise can be illustrated through the asymptotic theory for the realized volatility estimator. For a purely diffusive price process, it follows from Jacod and Protter (1998) and Barndorff-Nielsen and Shephard (BNS) (2002a; 2002b; 2004a), that asymptotically, as $\Delta \to 0$,

$$\sqrt{h/\Delta}\,[RV(t,h,\Delta) - IV(t,h)] \sim N(0, 2\cdot IQ(t,h)), \tag{12}$$

so the ratio of the integrated quarticity, $IQ(t,h) = \int_{t-h}^{t}\sigma^4(s)\,ds$, to the number of intra-day returns, $M=h/\Delta$, determines the precision of the realized volatility estimator, generalizing and improving the result discussed below eq. (3). A simple way to convey the implications of microstructure noise is to impose an exogenous bound, say Δ^*, below which no useful information from sampling is available as the semi-martingale assumption is blatantly violated at the highest frequencies. Hence, over an interval of length h, only $M^*=h/\Delta^*$ observations can be exploited for inference, and h must be of a certain size for realized volatility measures to possess meaningful precision. Equation (11) requires Δ to vanish at a rapid rate, so the bound Δ^* is quickly binding. If only a handful of returns is available within a few minutes of t, the sampling scheme behind eq. (11) breaks down. In contrast, for Δ^* fixed at one or five minutes, it is feasible to estimate the quadratic variation with reasonable accuracy for h equal to one trading day.

Alternative return variation measures

For a pure diffusion in a frictionless market, the basic realized volatility estimator exploiting all available observations is optimal for the quadratic variation. However, once price jumps and microstructure noise are introduced, matters become more complex. A number of issues may be addressed through alternative return variation measures, including the ability to disentangle the effect of jumps from the diffusive volatility, to estimate quantities needed for feasible inference about realized volatility, and to develop more robust-to-noise measures of integrated variance.

First, absent jumps and under appropriate regularity, one may extend the theory for the integrated variance to include the integrated variation of arbitrary powers, that is, for $\Delta\to 0$ the *realized power variation of order p* consistently estimates the p'th order integrated power variation ($p \le 1/2$),

$$(E|Z|^p)^{-1}(h/\Delta)^{1-p/2}\sum_{m=1}^{M}|r(t-h+m\cdot\Delta,\Delta)|^p \to \int_{t-h}^{t}\sigma^p(s)\,ds \tag{13}$$

where Z denotes a standard normal variable. For p=4, this result provides a simple estimator for the integrated quarticity which may be plugged into eq. (12) to yield a feasible distribution theory for realized volatility. An asymptotic distribution theory akin to eq. (12) is available for realized power variation (see Barndorff-Nielsen et al., 2006b). Another extension, the *realized k-skip bipower variation*, is also consistent for the integrated variance,

$$BV(t, h; k; \Delta) \equiv \frac{\pi}{2} \sum_{m=k+1}^{M} |r(t-h+m\cdot\Delta, \Delta)||r(t-h+(m-k)\cdot\Delta, \Delta)| \rightarrow IV(t, h) \equiv \int_{t-h}^{t} \sigma^2(s)\,ds. \tag{14}$$

For $k=1$, $BV(t,h;1,\Delta) \equiv BV(t,h;\Delta)$, is termed the *realized bipower variation*. These measures have convenient robustness properties. First, eq. (14) remains valid even if the return process follows a jump-diffusion. Hence, the bipower measures annihilate the jumps asymptotically and thus provide simple consistent estimators for the integrated variance. This allows for separation of the jump and diffusive contributions to the realized return variation, as

$$RV(t, h; \Delta) - BV(t, h; \Delta) \rightarrow QV(t, h) - IV(t, h) = \sum_{t-h<s\leq t} j^2(s), \text{ as } \Delta \rightarrow 0. \tag{15}$$

Combined with the asymptotic theory for realized power and bipower variation, the result renders formal statistical tests for the presence and impact of jumps feasible. This is applied for separate analysis of the diffusive and jump components (see BNS, 2004b; 2006; Huang and Tauchen, 2005; ABD, 2007). Alternative jump tests have recently been developed by Jiang and Oomen (2005), Andersen, Bollerslev and Dobrev (2007), and Lee and Mykland (2006).

The bipower variation measures also display robustness against microstructure noise. To first order, when not sampling at the very highest frequencies, the impact of noise may be mimicked by adding an i. i.d. process to the ‘efficient’ prices to generate noisy observations. This noise component induces negative return correlation which inflates the realized volatility estimator, resulting in a potentially strong upward bias. This may be alleviated by sampling more sparsely although this uses less data and reduces efficiency, pointing towards a bias-variance trade-off: one should sample sparsely enough to avoid a significant bias but frequently enough that efficiency is not compromised. An informal bias diagnostic is to apply the realized volatility estimator for different underlying frequencies over the full sample, that is, h is on the order of years or a decade. The long horizon minimizes sampling variability so that, absent the microstructure bias, all the measures should centre closely on the sample realized return volatility. A *volatility signature plot* depicts those realized volatility measures against the underlying sampling frequency with Δ ranging from seconds to a full day. For liquid financial markets, signature plots typically indicate inflated values at the highest frequencies which then decay quite smoothly to a stable level for sampling between 5 and 40•minutes. Andersen, Bollerslev, Diebold and Labys (ABDL) (2000a) suggest that the shortest sampling intervals not displaying a significant bias may be desirable choices. Of course, this criterion rewards unbiasedness over efficiency. Bandi and Russell

(2005a) explicitly trade off the microstructure bias with the efficiency gains from more data. An alternative is to apply skip-*k* bipower variation measures. Huang and Tauchen (2005) document that these tend to work well when applied to noisy observations from jump diffusions. Andersen, Bollerslev, Frederiksen and Nielsen (ABFN) (2006a) extend the volatility signature plots to include both power variation and skip-k bipower variation measures. Such generalized plots provide insights into the robustness of the realized quarticity and other quantities used for jump tests and may guide the choice of an adequate sampling frequency for analysis of a range of different issues in the presence of market frictions.

Finally, procedures have been developed to correct for microstructure bias while utilizing more of the available data. The proposals include the subsampling idea alluded to in Zhou (1996), and extended and formalized by Zhang, Mykland and Aït-Sahalia (2005) and Aït-Sahalia, Mykland and Zhang (2006) as well as the kernel based methods of Barndorff-Nielsen, Hansen, Lunde and Shephard (2006), and numerous filtering approaches discussed in Hansen and Lunde (2006).

Empirical applications

Hsieh (1991) is perhaps the first to apply intra-day returns for historical volatility measurement of the daily return variation. Closely related work appears in publications by the Olsen & Associates group. This is surveyed in Dacorogna et al. (2001). Zhou (1996) offers the first systematic study of the realized volatility estimator combining theoretical and empirical issues. Interestingly, he discusses contamination by market microstructure noise as well as ideas for a variety of feasible corrections. Comte and Renault (1998) also comment on estimating diffusive spot volatility through the empirical counterpart to quadratic variation.

In parallel work, Andersen and Bollerslev (AB) (1997; 1998b) explore the dynamics of high-frequency return volatility, documenting the striking effectiveness of cumulative absolute and squared returns as daily volatility measures. These findings inspired theoretical inquiries and are followed by the statement of consistency of realized volatility for the quadratic variation for a general multivariate jump-diffusion setting in ABDL (2001). In concurrent work, BNS (2001; 2002a) and Meddahi (2002) provide initial asymptotic theory for the realized volatility estimator, with the diffusive multivariate case treated in BNS (2004a).

Empirical work almost invariably operates with *h* equal to one trading day (or more). This is due to the pronounced intra-day volatility pattern which induces systematic shifts in the quadratic return variation over different segments of the trading day. As noted in AB (1997; 1998b) this type of largely deterministic effects are alleviated at the daily frequency, rendering this the natural basis for analysis.

The list of empirical applications is growing rapidly. A brief overview of the topics explored along with selective, but not exhaustive, references are provided below.

The most common use of realized volatility is as a basis for volatility forecasting and evaluation. AB (1998a) document the potential of realized volatility for assessment of standard volatility forecasts, a theme further explored in Andersen, Bollerslev and Lange (1999) and rationalized more formally in Andersen, Bollerslev and Meddahi (2004) using powerful analytical techniques developed in Meddahi

(2001). An integrated approach to measurement, modelling and forecasting of realized volatility is developed in ABDL (2003). Ghysels, Santa-Clara and Valkanov (2006) show that a combination of realized volatility measures for different frequencies and horizons may enhance forecast performance. Engle and Gallo (2006) follow a similar strategy but with a different modelling approach. ABD (2007) improve performance by separating the jump and diffusive volatility components in the forecasting procedure. Other studies on the presence and importance of price jumps using realized volatility related jump statistics are ABFN (2006b), Tauchen and Zhou (2006), and Fleming and Paye (2006), Andersen, Bollerslev and Dobrev (2007), Jiang and Oomen (2005), and Lee and Mykland (2006). In the same spirit, Liu and Maheu (2005) find that more jump-robust power variation measures are preferred to realized volatility for forecasting purposes. Earlier studies of forecast performance include Blair, Poon and Taylor (2001) and Martens (2002). Some initial studies of the role of microstructure noise and discretization error for forecasting and forecast evaluation is provided by Aït-Sahalia and Mancini (2006), ABM (2005; 2006), and Ghysels and Sinko (2006). The issue of how to include (noisy) overnight return information into the volatility measures and forecasts is addressed by Fleming, Kirby and Ostdiek (2003) and Hansen and Lunde (2005).

The evidence for long-range persistence in volatility is particularly striking when analysed via realized volatility rather than via daily return observations. AB (1998a) demonstrate that return series spanning only a couple of years are sufficient to identify a distinct hyperbolic decay in the realized power variation series. Moreover, the implied degree of fractional integration appears to be stable across subsamples, at around 0.35–0.45, implying a stationary volatility series. This finding is confirmed by virtually all subsequent studies of realized volatility exploring the issue, including ABDL (2001, 2003), Areal and Taylor (2002), Martens (2002), Zumbach (2004), Deo, Hsieh and Hurvich (2005), and Deo, Hurvich and Lu (2006). Moreover, related early work by the Olsen & Associates group also note the presence of scaling laws in volatility measures obtained from high frequency returns, see, for example, Müller et al. (1990) and the review in Dacorogna et al. (2001). This inspired an extensive amount of empirical work in the ‘econophysics’ area on volatility scaling laws which is summarized in Mantegna and Stanley (2000). These robust empirical results suggest that the long-memory property is not driven by occasional structural breaks but is present in the data generating process at high frequencies. As such, it sheds new light on a contentious issue which is not readily resolved without an effective volatility measure that improves the signal-to-noise ratio, allowing for more effective inference. Explicit estimation of the long memory features may be circumvented through a model with multiple volatility components, each governed by an autoregressive process, as such structures approximate the long-memory dependencies very well. This approach has been applied for realized volatility series by ABD (2007), BNS (2001), and Corsi (2003), among others.

The no-arbitrage implications for the return dynamics expressed in eqs (4) and (5), are necessarily quite weak and general, but do nonetheless have distributional implications which are potentially testable. As highlighted by eq. (7), auxiliary assumptions produce a mixture of normals result akin to the mixture-of-distributions theory, originating from work by Clark (1973) and Tauchen and Pitts (1983). The novelty of eq. (7) is the potential *ex post* observability of the mixing variable, the integrated variance, which enables direct inference regarding the distributional implications without any parametric assumptions. In

practice there will be some discretization error and microstructure distortions, but the properties of the realized volatility estimator should facilitate powerful tests. ABDL (2000b) confirm that returns standardized by realized volatility are much closer to normal than the standardized residuals from the usual volatility models based on daily data, even if the normality is not exact. Thomakos and Wang (2003) reach the identical conclusion for a different set of assets. Of course, eq. (7) is not valid in the presence of price jumps or dependence between the volatility and return innovation processes. The former issue may be addressed through jump identification procedures which seek to annihilate the impact of the jumps. The latter issue is accommodated by sampling the return process in ‘financial time’, consisting of calendar periods representing equal increments to the integrated variance process, as noted by Peters and de Vilder (2006) for a diffusive return process. This approach is extended to a jump-diffusive setting by Andersen, Bollerslev and Dobrev (2007) who also explore practical implementation issues for the associated distributional tests in detail. More extensive data sets and alternative jump identification techniques are considered in ABFN (2006b). They quite generally obtain jump-adjusted, financial-time sampled returns that are indistinguishable from i.i.d. standard Gaussian variates through realized volatility based empirical procedures. In sum, the results corroborate the general framework, and the tools developed for jump identification and measurement of the quadratic variation deliver empirically meaningful series of jumps and quadratic variation which are fully consistent with the theoretical underpinnings. In the process, direct evidence of the importance of jumps and the asymmetric return-volatility relationship is provided. Jumps are present for all asset classes and constitute a non-negligible fraction of overall return variation. For equities, negative returns tend to induce higher volatility than corresponding positive returns, a feature broadly recognized in the prior literature. Interestingly, there are also signs of significant asymmetric relationships for other asset classes, although both magnitude and sign may change over time. Similar issues are studied by Fleming and Paye (2006) and closely related topics are explored by Maheu and McCurdy (2002).

Multivariate applications of the realized volatility estimator are in principle straightforward. They are used to study the broad correlation patterns among individual stocks in ABDE (2001), for volatility timing in portfolio allocation in Fleming, Kirby and Ostdiek (2003), for estimation of systematic market risk exposure (time-varying market betas) in Andersen et al. (2005), for assessment of a broader set of risk loading coefficients in Bollerslev and Zhang (2003), and for dynamic portfolio choice in Bandi, Russell and Zhou (2008). In spite of these initial explorations, it is clear that the multivariate setting introduces additional practical complications as there is evidence of significant delays in the reaction of one security price to movements in another related asset. Sheppard (2006) explores these issues in some depth even if general prescriptions for practice do not directly follow. The favoured approach in current work is to include temporal cross-correlation patterns through measurement of the relations between the return of one asset with lead and lag returns for the other asset – in the spirit of the corrections for non-trading effects in the estimation of betas from daily data in Scholes and Williams (1977). Hayashi and Yoshida (2005; 2006) and Griffin and Oomen (2006) study such realized covariance and correlation estimators. Bandi and Russell (2005b) and Zhang (2006) seek to trade off bias and efficiency optimally. An alternative approach is proposed in Bauer and Vorkink (2006).

A natural comparison for realized volatility based forecasts is with forecasts implied by traded financial contracts such as options. In fact, an intriguing parallel exists between expected future realized volatility

and the pricing of volatility swaps, with the latter reflecting the expected future integrated variance under the risk-neutral (pricing) measure if jumps in the price path are absent. Hence, systematic differences between realized volatility and implied volatility measures reflect the market prices of volatility risk (see, for example, Britten-Jones and Neuberger, 2000; Carr and Madan, 1998; Carr and Wu, 2005). Moreover, Bondarenko (2004) shows that the implied volatility result remains approximately valid for jump diffusions and returns sampled at discrete intervals only. Recent empirical papers on the performance of realized volatility forecasts versus implied volatility measures include, for example, Andersen, Frederiksen and Staal (2006), Bollerslev, Gibson and Zhou (2005), Bollerslev and Zhou (2006), Busch, Christensen and Nielsen (2006), Chan, Kalimipulli and Sivakumar (2006) and Pong et al. (2004). The findings confirm that realized volatility forecasts contain information for future return variability over-and-beyond implied volatility forecasts, while standard volatility forecasts obtained from models utilizing only daily data are fully encompassed by the market based measures.

Future directions for research

Most of the empirical work associated with the realized volatility concept has focused directly on the measurement precision and forecast performance. In order for the approach to enter routinely in more mainstream applications within asset pricing, risk management and portfolio allocation the empirical studies must broaden in scope. As reviewed above, this has begun to happen, but much work remains. One recent example of using the concept for model specification testing is Andersen and Benzoni (2005). The study documents a serious deficiency in the ability of affine term structure models to accommodate the observed dynamics of realized yield volatility for the US Treasury market. Theoretically, the concurrent yield curve should span yield volatility, both *ex post* and *ex ante*, but this property is systematically violated as the yield variation at every maturity displays genuine stochastic features not associated with simultaneous directional shifts in the yield curve. The result extends earlier findings based on *ex ante* volatility measures at the monthly frequency. For generalizations of term structure models operating within the popular and tractable affine setting, the realized volatility measures promise to be valuable diagnostic tools.

The applications of realized volatility measures will surely continue to grow and broaden as the advantages of the enhanced precision in the measurement of the return variability are much too large to be ignored and many important questions await thorough analysis from this perspective. Of particular interest is the development of practical approaches to generate reliable measures for the high-dimensional case involving a large set of assets. On the theoretical front, work remains in terms of understanding the relative advantages of the different robust alternatives to the basic realized volatility measure. One potentially promising avenue is to further develop the locally constant volatility technique developed by Mykland (2006), seeking to provide a formal, yet simple and powerful, tool for asymptotic theory while allowing for time-varying price dynamics.

See Also

- ARCH models

- continuous and discrete time models
- kernel estimators in econometrics
- law(s) of large numbers
- long memory models
- martingales
- mean-variance analysis
- measurement error models
- mixture models
- options
- stochastic volatility models
- Wiener process

Bibliography

Aït-Sahalia, Y. and Mancini, L. 2006. Out of sample forecasts of quadratic variation. Manuscript, Princeton University and University of Zürich.

Aït-Sahalia, Y., Mykland, P.A. and Zhang, L. 2006. Ultra high frequency volatility estimation with dependent microstructure noise. Manuscript, Princeton University.

Andersen, T.G. and Benzoni, L. 2005. Can bonds hedge volatility risk in the U.S. Treasury market? A specification test for affine term structure models. Manuscript, Kellogg School, Northwestern University and University of Minnesota.

Andersen, T.G. and Bollerslev, T. 1997. Heterogeneous information arrivals and return volatility dynamics: uncovering the long-run in high frequency returns. *Journal of Finance* 52, 975–1005.

Andersen, T.G. and Bollerslev, T. 1998a. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39, 885–905.

Andersen, T.G. and Bollerslev, T. 1998b. Deutschemark-Dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance* 53, 219–65.

Andersen, T.G., Bollerslev, T. and Diebold, F.X. 2007. Roughing it up: including jump components in the measurement, modeling and forecasting of return volatility. *Review of Economics and Statistics* 89.

Andersen, T.G., Bollerslev, T. and Diebold, F.X. 2008. Nonparametric volatility measurement. In *Handbook of Financial Econometrics*, ed. L.P. Hansen and Y. Aït-Sahalia. Amsterdam: North-Holland.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. 2001. The distribution of realized stock

return volatility. *Journal of Financial Economics* 61, 43–76.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2000a. Great realizations. *Risk* 13, 105–8.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2000b. Exchange rate returns standardized by realized volatility are (nearly) Gaussian. *Multinational Finance Journal* 4, 159–79.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2001. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96, 42–55.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2003. Modeling and forecasting realized volatility. *Econometrica* 71, 579–625.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Vega, C. 2003. Micro effects of macro announcements: real-time price discovery in foreign exchange. *American Economic Review* 93, 38–62.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Wu, G. 2005. A framework for exploring the macroeconomic determinants of systematic risk. *American Economic Review* 95, 398–404.

Andersen, T.G., Bollerslev, T. and Dobrev, D. 2007. No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: theory and testable distributional implications. *Journal of Econometrics* 138, 125–80.

Andersen, T.G., Bollerslev, T., Frederiksen, P.H. and Nielsen, M.Ø. 2006a. Comment on realized variance and market microstructure noise. *Journal of Business & Economic Statistics* 24, 173–9.

Andersen, T.G., Bollerslev, T., Frederiksen, P.H. and Nielsen, M.Ø. 2006b. Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns. Manuscript, Kellogg School, Northwestern University.

Andersen, T.G., Bollerslev, T. and Lange, S. 1999. Forecasting financial market volatility: sample frequency vis-à-vis forecast horizon. *Journal of Empirical Finance* 6, 457–77.

Andersen, T.G., Bollerslev, T. and Meddahi, N. 2004. Analytic evaluation of volatility forecasts. *International Economic Review* 45, 1079–110.

Andersen, T.G., Bollerslev, T. and Meddahi, N. 2005. Correcting the errors: volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica* 73, 279–96.

Andersen, T.G., Bollerslev, T. and Meddahi, N. 2006. Market microstructure noise and realized volatility forecasting. Manuscript, Kellogg School, Northwestern University.

Andersen, T.G., Frederiksen, P.H. and Staal, A.D. 2006. The information content of realized volatility forecasts. Manuscript, Kellogg School, Northwestern University.

Areal, N.M.P.C. and Taylor, S.J. 2002. The realized volatility of FTSE-100 futures prices. *Journal of Futures Markets* 22, 627–48.

Back, K. 1991. Asset prices for general processes. *Journal of Mathematical Economics* 20, 371–95.

Bandi, F. and Russell, J.R. 2005a. Microstructure noise, realized volatility, and optimal sampling. Manuscript, Graduate School of Business, University of Chicago.

Bandi, F. and Russell, J.R. 2005b. Realized covariation, realized beta and microstructure noise. Manuscript, Graduate School of Business, University of Chicago.

Bandi, F., Russell, J.R. and Zhou, Y. 2008. Using high-frequency data in dynamic portfolio choice. *Econometric Reviews*, forthcoming.

Barndorff-Nielsen, O.E., Graversen, S.E., Jacod, J. and Shephard, N. 2006a. Limit theorems for bipower variation in econometrics. *Econometric Theory* 22, 677–720.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N. 2006b. Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. Manuscript, Nuffield College, Oxford.

Barndorff-Nielsen, O.E. and Shephard, N. 2001. Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society, Series B* 63, 167–241.

Barndorff-Nielsen, O.E. and Shephard, N. 2002a. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* 64, 253–80.

Barndorff-Nielsen, O.E. and Shephard, N. 2002b. Estimating quadratic variation using realised variance. *Journal of Applied Econometrics* 17, 457–77.

Barndorff-Nielsen, O.E. and Shephard, N. 2003. Realized power variation and stochastic volatility models. *Bernoulli* 9, 243–65.

Barndorff-Nielsen, O.E. and Shephard, N. 2004a. Econometric analysis of realised covariation: high-frequency covariance, regression and correlation in financial econometrics. *Econometrica* 72, 885–925.

Barndorff-Nielsen, O.E. and Shephard, N. 2004b. Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* 2, 1–37.

Barndorff-Nielsen, O.E. and Shephard, N. 2006. Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* 4, 1–30.

Bauer, G.H. and Vorkink, K. 2006. Multivariate realized stock market volatility. Manuscript, Bank of Canada and Sloan School, Massachusetts Institute of Technology.

Blair, B.J., Poon, S.-H. and Taylor, S.J. 2001. Forecasting S&P 100 volatility: the incremental information content of implied volatility and high-frequency index returns. *Journal of Econometrics* 105, 5–26.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–27.

Bollerslev, T., Gibson, M. and Zhou, H. 2005. Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities. Manuscript, Department of Economics, Duke University.

Bollerslev, T. and Mikkelsen, H.O. 1996. Modeling and pricing long memory in stock market volatility. *Journal of Econometrics* 73, 151–84.

Bollerslev, T. and Zhang, B.Y.B. 2003. Measuring and modeling systematic risk in factor pricing models using high-frequency data. *Journal of Empirical Finance* 10, 533–58.

Bollerslev, T. and Zhou, H. 2006. Volatility puzzles: a simple framework for gauging return-volatility regressions. *Journal of Econometrics* 131, 123–50.

Bondarenko, O. 2004. Market price of variance risk and performance of hedge funds. Manuscript, University of Illinois at Chicago.

Britten-Jones, M. and Neuberger, A. 2000. Option prices, implied price processes, and stochastic volatility. *Journal of Finance* 55, 839–66.

Busch, T., Christensen, B.J. and Nielsen, M.Ø. 2006. The role of implied volatility in forecasting realized volatility and jumps in foreign exchange, stock and bond markets. Manuscript, University of Aarhus and Cornell University.

Carr, P. and Madan, D. 1998. Towards a theory of volatility trading. In *Volatility: New Estimation Techniques for Pricing Derivatives*, ed. R. Jarrow. London: Risk Books.

Carr, P. and Wu, L. 2005. Variance risk premia. Manuscript, Courant Institute, New York University.

Chan, W.H., Kalimipulli, M. and Sivakumar, R. 2006. The economic value of using realized volatility in the index options market. Manuscript, Wilfrid Laurier University and University of Waterloo.

Clark, P.K. 1973. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* 41, 135–55.

Comte, F. and Renault, E. 1998. Long memory in continuous time stochastic volatility models. *Mathematical Finance* 8, 291–323.

Corsi, F. 2003. A simple long-memory model of realized volatility. Manuscript, University of Southern Switzerland.

Dacorogna, M.M., Gencay, R., Müller, U., Olsen, R.B. and Pictet, O.V. 2001. *An Introduction to High-Frequency Finance*. San Diego, CA: Academic Press.

Deo, R., Hsieh, M. and Hurvich, C.M. 2005. Tracing the source of long memory in volatility. Manuscript, Stern School, New York University.

Deo, R., Hurvich, C.M. and Lu, Y. 2006. Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment. *Journal of Econometrics* 131, 29–58.

Engle, R.F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987–1008.

Engle, R.F. and Gallo, G.M. 2006. A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* 131, 3–27.

Fleming, J., Kirby, C. and Ostdiek, B. 2003. The economic value of volatility timing using realized volatility. *Journal of Financial Economics* 67, 473–509.

Fleming, J. and Paye, B.S. 2006. High-frequency returns, jumps and the mixture-of-normals hypothesis. Manuscript, Jones School, Rice University.

Foster, D. and Nelson, D.B. 1996. Continuous record asymptotics for rolling sample estimators. *Econometrica* 64, 139–74.

French, K.R., Schwert, G.W. and Stambaugh, R.F. 1987. Expected stock returns and volatility. *Journal of Financial Economics* 19, 3–29.

Ghysels, E., Santa-Clara, P. and Valkanov, R. 2006. Predicting volatility: how to get the most out of returns data sampled at different frequencies. *Journal of Econometrics* 131, 59–95.

Ghysels, E. and Sinko, A. 2006. Volatility forecasting and microstructure noise. Manuscript, University of North Carolina, Chapel Hill.

Griffin, J.E. and Oomen, R.C.A. 2006. Covariance measurement in the presence of non-synchronous trading and market microstructure noise. Manuscript, University of Warwick.

Hansen, P.R. and Lunde, A. 2005. A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics* 3, 525–54.

Hansen, P.R. and Lunde, A. 2006. Realized variance and market microstructure noise. *Journal of Business & Economic Statistics* 24, 127–61.

Harrison, J.M. and Kreps, D. 1978. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20, 381–408.

Hayashi, T. and Yoshida, N. 2005. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* 11, 359–79.

Hayashi, T. and Yoshida, N. 2006. Estimating correlations with nonsynchronous observations in continuous diffusion models. Manuscript, Columbia University and University of Tokyo.

Hsieh, D.A. 1991. Chaos and nonlinear dynamics: application to financial markets. *Journal of Finance* 46, 1839–77.

Huang, X. and Tauchen, G. 2005. The relative contribution of jumps to total price variation. *Journal of Financial Econometrics* 3, 456–99.

Hull, J. and White, A. 1987. The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42, 281–300.

Jacod, J. and Protter, P. 1998. Asymptotic error distributions for the Euler method for stochastic differential equations. *Annals of Probability* 26, 267–307.

Jiang, G.J and Oomen, R.C.C. 2005. A new test for jumps in asset prices. Manuscript, Eller College of Management, University of Arizona and Warwick Business School.

Lee, S.S and Mykland, P.A. 2006. Jumps in real-time financial markets: a new nonparametric test and jump dynamics. Manuscript, University of Chicago.

Liu, C. and Maheu, J.M. 2005. Modeling and forecasting realized volatility: the role of power variation. Manuscript, University of Toronto.

Maheu, J.M. and McCurdy, T.H. 2002. Nonlinear features of realized FX volatility. *Review of Economics and Statistics* 84, 668–81.

Mantegna, R.N. and Stanley, H.E. 2000. *An Introduction to Econophysics – Correlations and Complexity in Finance*. Cambridge, MA: Cambridge University Press.

Martens, M. 2002. Measuring and forecasting S&P500 index-futures volatility using high-frequency data. *Journal of Futures Market* 22, 497–518.

Meddahi, N. 2001. An Eigen function approach for volatility modeling. Manuscript, Department of Economics, University of Montreal and CIRANO.

Meddahi, N. 2002. A theoretical comparison between integrated and realized volatilities. *Journal of Applied Econometrics* 17, 479–508.

Merton, R.C. 1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–44.

Merton, R.C. 1980. On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics* 8, 323–61.

Müller, U.A., Dacarogna, M.M., Olsen, R.B., Pictet, O.V., Schwarz, M. and Morgenegg, C. 1990. Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis. *Journal of Banking and Finance* 14, 1189–208.

Mykland, P.A. 2006. A Gaussian calculus for inference from high frequency data. Manuscript, University of Chicago.

Officer, R.R. 1973. The variability of the market factor of New York stock exchange. *Journal of Business* 46, 434–53.

Peters, R.T. and de Vilder, R.G. 2006. Testing the continuous semimartingale hypothesis for the S&P 500. *Journal of Business & Economic Statistics* 24, 444–54.

Pong, S., Shackleton, M.B. Taylor, S.J. and Xu, X. 2004. Forecasting currency volatility: a comparison of implied volatilities and ARFIMA models. *Journal of Banking & Finance* 28, 2541–63.

Protter, P. 1992. *Stochastic Integration and Differential Equations: A New Approach*, 2nd edn. New York: Springer.

Scholes, M. and Williams, J. 1977. Estimating betas from nonsynchronous data. *Journal of Financial Economics* 5, 309–27.

Shephard, N. 2005. *Stochastic Volatility. Selected Readings*. Oxford: Oxford University Press.

Sheppard, K. 2006. Realized covariance and scrambling. Manuscript, University of Oxford.

Tauchen, G.E. and Pitts, M. 1983. The price variability-volume relationship on speculative markets. *Econometrica* 51, 485–505.

Tauchen, G.E. and Zhou, H. 2006. Identifying realized jumps on financial markets. Manuscript, Duke University and Federal Reserve Board of Governors.

Taylor, S.J. 1986. *Modeling Financial Time Series*. Chichester, UK: Wiley.

Thomakos, D.D. and Wang, T. 2003. Realized volatility in the futures markets. *Journal of Empirical Finance* 10, 321–53.

Zhang, L. 2006. Estimating covariation: Epps effect, microstructure noise. Manuscript, University of Chicago at Illinois.

Zhang, L., Mykland, P.A. and Aït-Sahalia, Y. 2005. A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100, 1394–411.

Zhou, B. 1996. High-frequency data and volatility in foreign exchange rates. *Journal of Business and Economic Statistics* 14, 45–52.

Zumbach, G. 2004. Volatility processes and volatility forecasts with long memory. *Quantitative Finance* 4, 70–86.

How to cite this article

Andersen, Torben G. "realized volatility." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000264> doi:10.1057/9780230226203.1404

The New Palgrave Dictionary of Economics Online

reciprocity and collective action

Rajiv Sethi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A collective action problem arises when the private incentives faced by individual members of a group are not properly aligned with their shared goals. Such problems can be overcome if opportunistic behaviour is restrained by explicit sanctions or internalized social norms. In particular, collective action is facilitated by norms of reciprocity that induce individuals to undertake pro-social actions whenever they expect others to do the same. From this perspective, collective action requires coordinated expectations and effective communication. Experimental evidence suggests that reciprocity norms are widespread in human populations, and evolutionary mechanisms that can account for their prevalence have been identified.

Keywords

assortative matching; cheap talk; collective action; cooperation; coordination problems and communication; free-rider problem; public goods; reciprocity; social norms; subgame perfection; tragedy of the commons

Article

Advancing the common interest of a group sometimes requires its members to sacrifice their private interests. Such situations, in which individual incentives are not properly aligned with shared goals, are called collective action problems. They arise frequently in economic and social life, for instance in the context of political mobilization, electoral turnout, pollution abatement, common property management and the provision of public goods. They can involve relatively small groups such as families, teams, or business partnerships, or very large groups that cut across national boundaries.
In his classic work on collective action, Mancur Olson (1965) conjectured that individuals would be unable to overcome such problems unless their behaviour was constrained by rules that were externally imposed and enforced. Along similar lines, Garret Hardin (1968) argued in an influential paper that, left

to their own devices, individuals would face a ‘tragedy of the commons’ which could be overcome only by ‘mutual coercion, mutually agreed upon’. This view continues to have considerable currency in economics in the form of the *free-rider hypothesis*, which maintains that voluntary contributions that are socially beneficial but privately costly will not generally be observed (Bergstrom, Blume and Varian, 1986).

Despite the compelling logic underlying the free-rider hypothesis, there are numerous instances of groups having overcome collective action problems without external pressure, sometimes by designing and abiding by their own set of rules, and sometimes on the basis of less formal arrangements codified in social norms. The success of OPEC in constraining production to maintain price levels is based on a mutually beneficial agreement among member countries that has been sustained despite strong incentives for some producers to free-ride on the restraint practised by others. On a smaller scale, many examples of successful collective action in the management of local fisheries, forests, and other renewable resources have been documented (Bromley, 1992; Ostrom, 1990). Such resources are often held as common property, and the maintenance of sustainable stocks requires restraint in individual extraction levels. Restraint is typically enforced by formal or informal sanctions, and participation in such punishment mechanisms is itself a form of collective action. There also exist examples of collective action in the absence of any sanctioning mechanism. For instance, voter turnout is often substantial in large elections, contrary to the predictions of the free-rider hypothesis.

It has been argued that many instances of successful collective action arise in small and stable groups whose members interact with each other repeatedly. Under such circumstances, pro-social behaviour can be fully consistent with the standard economic hypotheses of rationality and self-interest. When interactions are repeated, self-interested cooperation can arise if one believes that non-cooperative actions will be punished in future periods. Moreover, such threats of punishment can be credible if abstaining from punishment is itself punished. Formally, cooperative behaviour can be sustained in subgame perfect equilibrium if interactions are infinitely (or indefinitely) repeated (Fudenberg and Maskin, 1986). Hence the tension between individual and common interest is less severe and collective action more likely to arise in small and stable groups.

While the threat of future punishment or the promise of future reward might motivate collective action in some instances, there are many situations in which individual actions are unobservable or repetition too infrequent for such considerations to be decisive. Voter turnout, for instance, or private donations to charity are not easily explained as self-interested responses to material incentives. Similarly, sacrifices involving risks to life and limb, as in the case of battlefield heroism or spontaneous collective violence, are unlikely to be driven by a calculated response to future costs and benefits. What, then, could account for such phenomena?

There is now a considerable body of experimental evidence to suggest that many individuals are willing to take actions that further the common interest provided that they are reasonably sure that other group members will also take such actions. Furthermore, they are willing to sanction the opportunistic behaviour of others even at some cost to themselves (Fehr and Gächter, 2000). The widespread prevalence of such preferences for reciprocity suggests that collective action can sometimes be viewed as a coordination problem: if the members of a group confidently expect others to further the common good, such expectations can be self-fulfilling. On the other hand, expectations of widespread free-riding can also be self-fulfilling, so building confidence in the behaviour of others is a critical ingredient of

successful collective action. Communication among group members can help coordinate expectations, and it is therefore not surprising that allowing for communication among experimental subjects can result in dramatically increased levels of cooperation. This is the case even if communication takes the form of ‘cheap talk’, with neither threats nor promises being enforceable (Ostrom, Walker and Gardner, 1992).

If preferences for reciprocity are indeed part of the explanation for successful collective action, this raises the question of how such preferences have come to be widespread in human populations in the first place. The existence of a willingness to sacrifice one's own material interest for the common good poses an evolutionary puzzle. In order to survive and spread in human populations, the possession of such preferences must confer on an individual some advantage relative to those who are entirely self-interested. One intriguing possibility is that, despite being disadvantageous to individuals within groups, traits that are advantageous for the group itself may survive because of competition among groups:

> There can be no doubt that a tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to give aid to each other and to sacrifice themselves for the common good, would be victorious over other tribes; and this would be natural selection. (Darwin, 1871, p. 166)

In order to be effective, however, this mechanism requires variability across groups to be sustained while variability within groups is suppressed (Sober and Wilson, 1998). Whether or not the conditions for this are empirically plausible remains an open question.

There exist other channels through which a preference for reciprocity can be materially advantageous to individuals. One is assortative interaction: if individuals with preferences for reciprocity are more likely to interact with each other than with opportunists, the former can end up with higher material payoffs than the latter. Such assortation arises naturally in structured populations with local interaction. Even in unstructured populations with random matching, a propensity to reciprocate or to sanction opportunistic behaviour can confer an advantage provided that such preferences are observable to others. The visible possession of such propensities can alter the behaviour of those with whom one is interacting in such a manner as to be materially rewarding. Even opportunistic individuals might be induced to behave cooperatively in interactions with those who have a credible reputation for reciprocity. Such considerations can provide the basis for an evolutionary theory of reciprocity (Sethi and Somanathan, 2001).

Reciprocity is a key feature of successful collective action, both in repeated interactions and in more spontaneous settings. The willingness to further the common good even at considerable personal cost is widespread in human populations, but is often contingent on the willingness of others to do the same. This perspective suggests that collective action problems are not insurmountable, but that communication and coordination are critical in overcoming them.

See Also

- collective action
- common property resources
- cooperation
- coordination problems and communication
- public goods
- social norms
- social preferences

Bibliography

Bergstrom, T.C., Blume, L. and Varian, H. 1986. On the private provision of public goods. *Journal of Public Economics* 29, 25–49.

Bromley, D.W. 1992. *Making the Commons Work: Theory, Practice and Policy*. San Francisco: Institute for Contemporary Studies.

Darwin, C. 1871. *The Descent of Man and Selection in Relation to Sex*. London: Murray.

Fehr, E. and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90, 980–94.

Fudenberg, D. and Maskin, E. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54, 533–54.

Hardin, G. 1968. The tragedy of the commons. *Science* 162, 1243–8.

Olson, M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Ostrom, E., Walker, J. and Gardner, R. 1992. Covenants with and without a sword: self-governance is possible. *American Political Science Review* 86, 404–17.

Sethi, R. and Somanathan, E. 2001. Preference evolution and reciprocity. *Journal of Economic Theory* 97, 273–97.

Sober, E. and Wilson, D.S. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

How to cite this article

Sethi, Rajiv. "reciprocity and collective action." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000274> doi:10.1057/9780230226203.1405

The New Palgrave Dictionary of Economics Online

recursive competitive equilibrium

Rajnish Mehra
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In this article we define a recursive competitive equilibrium, provide an example and review the related literature.

Keywords

Arrow–Debreu model of general equilibrium; Bellman's principle of optimality; business cycles; dynamic decentralization; Markov processes; neoclassical stochastic growth model; optimality; recursive competitive equilibrium; time-invariant equilibrium decision rules; total factor productivity; valuation equilibrium

Article

The underlying structure of most dynamic business-cycle and consumption-based asset-pricing models is a variant of the neoclassical stochastic growth model. Such models have been analysed by, among others, Cass (1965), Brock and Mirman (1972), and Donaldson and Mehra (1983). They focus on how an omniscient central planner seeking to maximize the present value of expected utility of a representative agent optimally allocates resources over the infinite time horizon.

Production is limited by an aggregate production function subject to technological (total factor productivity) shocks. The solution to the planning problem is characterized by time-invariant decision rules, which determine optimal consumption and investment each period. These decision rules have as arguments the economy's period aggregate capital stock and the shock to technology.

Business cycles, however, are not predicated on the actions of a central planner, but arise from interactions among economic agents in competitive markets. Given the desirable features of the stochastic growth paradigm – the solution methods are well known and the model generates well-defined proxies for all the major macro aggregates: consumption, investment, output, and so on – it is natural to ask if the allocations arising in that model can be viewed as competitive equilibria. That is, do price sequences exist such that economic agents, optimizing at these prices and interacting through competitive markets, achieve the allocations in question as competitive equilibria? This is the essential question of dynamic-decentralization theory.

Alternative approaches to dynamic decentralization: valuation equilibrium

One way of modelling uncertain dynamic economic phenomena is to use Arrow–Debreu general equilibrium structures and to search for optimal actions conditional on the sequence of realizations of all past and present random variables or shocks. The commodities traded are contingent claim contracts. These contracts deliver goods (for example, consumption and capital goods) at a future date, contingent on a particular sequential realization of uncertainty. Markets are assumed

to be complete, so that, for any possible future realization of uncertainty (sequence of technology shocks) up to and including some future period, a market exists for contracts that will deliver each good at that date contingent on that realization (event). This requires a very rich set of markets. All trading occurs in the first period: consumers contract to receive consumption and investment goods and to deliver capital goods in all future periods contingent on future states so as to maximize the expected present value of their utility of consumption over their infinite lifetimes. Firms choose their production plans so as to maximize the present value of discounted profits. Given current prices, they contract to deliver consumption and investment goods to, and to receive capital goods from, the consumer-investors. Under standard preference structures, these contingent choices never need to be revised. That is, if markets reopen, no new trades will occur.

In its most general formulation, a valuation equilibrium is characterized simply as a continuous linear functional that assigns a value to each bundle of contingent commodities. Only under more restrictive assumptions can this function be represented as a price sequence (Bewley, 1972; Prescott and Lucas, 1972; Mehra, 1988). The basic result is that for any solution to the planner's problem – that is, sequences of consumption, investment and capital goods – a set of state-contingent prices exists such that these sequences coincide with the contracted quantities in the valuation equilibrium. This decentralization concept is quite broad and applies to central-planning formulations much more general than the neoclassical growth paradigm. It reminds us that the financial structure underlying the stochastic growth paradigm is fundamentally one of complete contingent commodity markets. Nevertheless, it is a somewhat unnatural perspective for macroeconomists (all macro policies must be announced at time zero), and it presumes a set of markets much richer than any observed. These shortcomings led to the development of the concept of a recursive competitive equilibrium.

Recursive competitive theory

An alternative approach that has proved very useful in developing testable theories is to replace the attempt to locate equilibrium sequences of contingent functions with the search for time-invariant equilibrium decision rules. These decision rules specify current actions as a function of a limited number of ‘state variables’ which fully summarize the effects of past decisions and current information. Knowledge of these state variables provides the economic agents with a full description of the economy's current state. Their actions, together with the realization of the exogenous uncertainty, determines the values of the state variables in the next sequential time period. This is what is meant by a recursive structure. In order to apply standard time-series methods to any testable implications, these equilibrium decision rules must be time-invariant.

Recursive competitive theory was first developed by Mehra and Prescott (1977) and further refined in Prescott and Mehra (1980). These papers also establish the existence of a recursive competitive equilibrium and the supportability of the Pareto optimal through the recursive price functions. Excellent textbook treatments are contained in Harris (1987), Stokey, Lucas and Prescott (1989) and Ljungqvist and Sargent (2004). Since its introduction, it has been widely used in exploring a vide variety of economic issues including business-cycle fluctuations, monetary and fiscal policy, trade-related phenomena, and regularities in asset price co-movements. (See, for example, Kydland and Prescott, 1982; Long and Plosser, 1983; Mehra and Prescott, 1985.)

The recursive equilibrium abstraction postulates a continuum of identical economic agents indexed on the unit interval (again with preferences identical to those of the representative agent in the planning formulation), and a finite number of firms. As in the valuation equilibrium approach, consumers undertake all consumption and saving decisions. Firms, which have equal access to a single constant-returns-to-scale technology, maximize their profits each period, and are assumed to produce two goods: a consumption good and a capital good. Unlike in the valuation equilibrium approach, trading between agents and firms occurs every period. (This is in contrast to markets in an Arrow–Debreu setting where, as mentioned earlier, no trade would occur if markets were to reopen.) At the start of each period, firms observe the technological shock to productivity and purchase capital and labor services, which are supplied inelastically at competitive prices. The capital and labour are used to produce the capital and consumption goods. At the close of the period, individuals, acting competitively, use their wages and the proceeds from the sale of capital to buy the consumption and capital goods produced by the firms. Consumers then retain the capital good into the next period, when it again becomes available to firms and the process repeats itself. Note that firms are liquidated at the end of each period

(retaining no capital assets while technology is freely available), and that no trades between firms and consumer-investors extend over more than one time period. Capital goods carried over from one period to the next are the only link between periods, and period prices depend only on the state variables in that period.

Formally, a *recursive competitive equilibrium* (RCE) is characterized by *time invariant* functions of a limited number of ‘state variables’, which summarize the effects of past decisions and current information. These functions (decision rules) include (*a*) a pricing function, (*b*) a value function, (*c*) a period allocation policy specifying the individual's decision, (*d*) a period allocation policy specifying the decision of each firm and (*e*) a function specifying the law of motion of the capital stock.

While the restrictive structure of markets and trades makes this concept less general than the valuation equilibrium approach, it provides an interpretation of decentralization that is better suited to macro-analysis. More recently, the recursive equilibrium concept has been generalized to admit an infinitely lived firm which maximizes its value. When an RCE is Pareto optimal, its allocation coincides with that of the associated planning problem. The solution to the central-planning stochastic-growth problem may then be regarded as the aggregate investment and consumption functions that would arise from a decentralized, recursive homogeneous consumer economy. We illustrate this with the help of an example below, which considers an economy with a single capital good. The reader is referred to Prescott and Mehra (1980) for the more general case with multiple capital types.

An example

Consider the simplest central planning stochastic growth paradigm

$$w(k_0, \lambda_0) = \max E\left\{\sum_{t=0}^{\infty} \beta^t u(c_t)\right\} \qquad \text{(P1)}$$

subject to

$$c_t + k_{t+1} \leq \lambda_t f(k_t, l_t), \lambda_0, k_0 \text{ given}, \quad l_t = 1 \ \forall t.$$

In this formulation, $u(\cdot)$ is the period utility function of a representative consumer defined over his period t consumption c_t; k_t denotes capital available for production in period t and l_t denotes period t labour supply which is inelastically supplied by the consumer-investor at $l_t = 1$, for all t. The expression $f(k_t, \cdot l_t)$ represents the period technology (production function) which is shocked by the bounded stationary stochastic factor λ_t. (It is assumed that λ_t is subject to a stationary Markov process with a bounded ergodic set.) E denotes the expectations operator and the central planner is assumed to have rational expectations; that is, he uses all available information to rationally anticipate future variables. In particular he knows the conditional distribution of future technology shocks $F(\lambda_{t+1}; \lambda_t)$. For the purposes of this example we restrict preferences to be logarithmic and assume a Cobb–Douglas technology (to the best of my knowledge, this parameterization is the simplest example known to result in closed form solutions): $u(c_t) = \ln c_t$ and $f(k_t, l_t) = k_t^{a} l_t^{1-a}$. We also assume that $\alpha, \beta < 1$ and that capital fully depreciates each period.

These conditions are sufficient to guarantee a closed form solution to the planning problem:

$$c_t = (1 - \alpha\beta)k_t^{\alpha}\lambda_t, \text{ and}$$

$$k_{t+1} = i_t = \alpha\beta k_t^{\alpha}\lambda_t$$

where we identify as investment, i_t, the capital stock held over for production in period $t + 1$. These allocations are Pareto optimal.

We will show that the investment and consumption policy functions arising as a solution to this problem may be regarded as the aggregate investment and consumption functions arising from a *decentralized* homogenous consumer economy. We first qualitatively describe the RCE underlying this model, and then demonstrate the relevant equilibrium price and quantity functions explicitly. The one capital good is assumed to produce two goods – a consumer good and an investment (capital) good. At the beginning of each period, firms observe the shock to productivity (λ_t) and purchase capital and labour from individuals at competitively determined rates. Both capital and labour are used to produce the two output goods. Individuals use their proceeds from the sale of capital and labour services to buy the consumption good (c_t) and the investment good (i_t) at the end of the period. This investment good is used as capital (k_{t+1}) available for sale to the firm next period and the process continues recursively.

To cast this problem formally as a recursive competitive equilibrium, we introduce some additional notation. Let k_t denote the capital holdings of a particular (measure zero) individual at time t, and $\underline{k}_t$ the distribution of capital amongst other individuals in the economy. This latter distinction allows us to make formal the competitive assumption: all the economic participants will assume that $\underline{k}_t$ is exogenous to them and that the price functions depend solely on this aggregate (in addition to the technology shock). Clearly, in equilibrium, $k_t = \underline{k}_t$ for our homogeneous consumer economy. In addition, let p_i, p_c and p_k be the price of the investment, consumption and capital goods respectively and p_l be the wage rate. These prices are presumed to be functions of the economy-wide state variables exclusively and all participants take these prices as given for their own decision making purposes. The 'state variables' characterizing the economy are $(\underline{k}, \lambda)$ and the individual are $(k, \underline{k}, \lambda)$.

We use the symbols (c, i, k, l) to denote points in the 'commodity space' for the firm and the consumer. The c in the commodity point of the firm is a function specifying the consumption good supplied by the firm and is written as $c^s(\underline{k}_t, \lambda_t)$. Similarly, the c in the commodity point of the individual is the amount of the consumption good demanded by the individual and is written as $c^d(k_t, \underline{k}_t, \lambda_t)$. In equilibrium (as mentioned earlier, in equilibrium $k_t = \underline{k}_t$), since the market clears, of course $c^s = c^d$. The same comments apply to the other elements of the commodity point.

In the decentralized version of this economy, the problem facing a typical household is

$$v(k_0, \underline{k}_0, \lambda_0) = \max E\left\{\sum_{t=0}^{\infty} \beta^t \ln c^d(k_t, \underline{k}_t, \lambda_t)\right\} \quad \text{(P2)}$$

subject to

$$p_c(\underline{k}_t, \lambda_t)\, c^d(k_t, \underline{k}_t, \lambda_t) + p_i(\underline{k}_t, \lambda_t)\, i^d(k_t, \underline{k}_t, \lambda_t) \leq p_k(\underline{k}_t, \lambda_t)\, k^s(k_t, \underline{k}_t, \lambda_t) + p_l(\underline{k}_t, \lambda_t)\, l^s(k_t, \underline{k}_t, \lambda_t)$$

$$k_{t+1} \equiv k^s(k_{t+1}, \underline{k}_{t+1}, \lambda_{t+1}) = i^d(k_t, \underline{k}_t, \lambda_t), l^s(k_t, \underline{k}_t, \lambda_t) \leq 1 \text{and} \underline{k}_{t+1} = \psi(\underline{k}_t, \lambda_t)$$

is the law of motion of the aggregate capital stock.
With capital and labour priced competitively each period, the firm's objective function is especially simple – maximize period profits. The firm's problem then is

$$\max \left\{ p_c(\underline{k}_t, \lambda_t)\ c^s(\underline{k}_t, \lambda_t) + p_i(\underline{k}_t, \lambda_t)\ i^s(\underline{k}_t, \lambda_t) - p_k(\underline{k}_t, \lambda_t)\ k^d(\underline{k}_t, \lambda_t) - p_l(\underline{k}_t, \lambda_t)\ l^d(\underline{k}_t, \lambda_t) \right\}$$

subject to

$$c_t^s + i_t^s \leq \lambda_t (k_t^d)^{\alpha} (l_t^d)^{1-\alpha}.$$

Via Bellman's principle of optimality, the recursive representation of the individual's problem P2 is

$$v(k_t, \underline{k}_t, \lambda_t) = \max_{\{c^d, i^d, l^s, k^d\}} \left\{ \ln\left(c^d(k_t, \underline{k}_t, \lambda_t)\right) + \beta \int v(i^d(k_t, \underline{k}_t, \lambda_t), \psi(\underline{k}_t, \lambda_t), \lambda_{t+1})\, dF(\lambda_{t+1} | \lambda_t) \right\}$$

subject to

$$p_c(\underline{k}_t, \lambda_t)\, c^d(k_t, \underline{k}_t, \lambda_t) + p_i(\underline{k}_t, \lambda_t)\, i^d(k_t, \underline{k}_t, \lambda_t) \leq p_k(\underline{k}_t, \lambda_t)\, k^s(k_t, \underline{k}_t, \lambda_t) + p_l(\underline{k}_t, \lambda_t)\, l^s(k_t, \underline{k}_t, \lambda_t)$$

$$k_{t+1} \equiv k^s(k_{t+1}, \underline{k}_{t+1}, \lambda_{t+1}) = i^d(k_t, \underline{k}_t, \lambda_t), l^s(k_t, \underline{k}_t, \lambda_t) \leq 1 \text{and} \underline{k}_{t+1} = \psi(\underline{k}_t, \lambda_t)$$

is the law of motion of the aggregate capital stock.
The firm of course, simply maximizes its period profits and hence does not have a multiperiod problem.
The following functions that are a solution to the individual and firm maximization problem above satisfy the definition of recursive competitive equilibrium:

1. 1. A value function $v(k_0, \underline{k}_0, \lambda_0) = E\left\{\Sigma_{t=0}^{\infty}\beta^t \ln\left[(1-\alpha\beta)\lambda_t \underline{k}_t^{\alpha-1}\left\{\alpha(k_t - \underline{k}_t) + \underline{k}_t\right\}\right]\right\}$. It can be shown that $v(\underline{k}_0, \underline{k}_0, \lambda_0) = A + B\ln\underline{k}_0 + C\ln\lambda_0$ where *A*, *B* and *C* are constants which are functions of the preference and technology parameters.
2. 2. A continuous pricing function $p(\underline{k}_t, \lambda_t) = \{p_c(\underline{k}_t, \lambda_t),\ \ p_i(\underline{k}_t, \lambda_t),\ \ p_k(\underline{k}_t, \lambda_t),\ \ p_l(\underline{k}_t, \lambda_t)\}$ that has the same dimensionality as the commodity point, where

$$p_c(\underline{k}_t, \lambda_t) = p_i(\underline{k}_t, \lambda_t) = 1$$

(We have chosen the consumption good to be the numeraire.)

$$p_k(k_t, \lambda_t) = \alpha\lambda_t \underline{k}_t^{\alpha-1}$$

$$p_l(k_t, \lambda_t) = (1-\alpha)\lambda_t \underline{k}_t^{\alpha-1}.$$

3. 3. Consumption and investment functions for the individual that are a function of the current state of the individual $(k, \underline{k}, \lambda)$

$$c^d(k_t, \underline{k}_t, \lambda_t) = (1-\alpha\beta)\lambda_t \underline{k}_t^{\alpha-1}\left\{\alpha(k_t - \underline{k}_t) + \underline{k}_t\right\}$$

$$l^s(k_t, \underline{k}_t, \lambda_t) = 1$$

$$i^d(k_t, \underline{k}_t, \lambda_t) = \alpha\beta\lambda_t\underline{k}_t^{\alpha-1}\left\{\alpha(k_t - \underline{k}_t) + \underline{k}_t\right\}$$

$$k^s(k_{t+1}, \underline{k}_{t+1}, \lambda_{t+1}) = i^d(k_t, \underline{k}_t, \lambda_t).$$

4. 4. Decision rules for the firm that are contingent on the state of the economy $(\underline{k}, \lambda)$

$$c^s(\underline{k}_t, \lambda_t) = (1 - \alpha\beta)\lambda_t\underline{k}_t^{\alpha},$$

$$l^d(\underline{k}_t, \lambda_t) = 1,$$

$$i^s(\underline{k}_t, \lambda_t) = \alpha\beta\lambda_t\underline{k}_t^{\alpha},$$

$$k^d(\underline{k}_{t+1}, \lambda_{t+1}) = i^s(\underline{k}_t, \lambda_t).$$

5. 5. The law of motion for the capital stock specifying the next period capital stock as a function of the current state of the economy $(\underline{k}_t, \lambda_t)$

$$\underline{k}_{t+1} = \psi(\underline{k}_t, \lambda_t) = \alpha\beta\lambda_t\underline{k}_t^{\alpha}.$$

6. 6. The consumption and investment decisions of the individual $c^s(k, \underline{k}, \lambda)$, $l^s(k, \underline{k}, \lambda)$ and $i^s(k, \underline{k}, \lambda)$ maximize the expected utility subject to the budget constraint. So that

$$v(k_t, \underline{k}_t, \lambda_t) = \ln((1-\alpha\beta)\lambda_t \underline{k}_t^{\alpha-1}(\alpha(k_t - \underline{k}_t) + \underline{k}_t)) + \beta \int v(\alpha\beta\lambda_t \underline{k}_t^{\alpha-1}(\alpha(k_t - \underline{k}_t) + \underline{k}_t), \alpha\beta\lambda_t \underline{k}_t^{\alpha}) dF(\lambda_{t+1}|\lambda_t).$$

7. 7. The decision rules of the firm $c^d(\underline{k}_t, \lambda_t)$, $l^d(\underline{k}_t, \lambda_t)$, $i^d(\underline{k}_t, \lambda_t)$ maximize firm profit.

Demand equals supply

$$c^d(k_{t+1}, \underline{k}_{t+1}, \lambda_{t+1}) = c^s(\underline{k}_t, \lambda_t), \quad l^s(k_{t+1}, \underline{k}_{t+1}, \lambda_{t+1}) = l^d(\underline{k}_t, \lambda_t) \text{and } i^s(k_{t+1}, \underline{k}_{t+1}, \lambda_{t+1}) = i^d(\underline{k}_t, \lambda_t).$$

The law of motion of the representative consumers capital stock is consistent with the maximizing behaviour of agents $\psi(\underline{k}_t, \lambda_t) = i^d(\underline{k}_t, \underline{k}_t, \lambda_t)$. It is readily demonstrated that since $v(\underline{k}_0, \underline{k}_0, \lambda_0) = w(\underline{k}_0, \lambda_0)$, the competitive allocation is Pareto optimal. See eqs (P1) and (P2).

Having formulated expressions for the prices of the various assets and their laws of motion, it is a relatively simple matter to calculate rates of return (price ratios) and study their dynamics. For an application to risk premia, see Donaldson and Mehra (1984).

Some researchers have formulated models that can be cast in this same recursive setting, yet whose equilibria are not Pareto-optimal. As a consequence, the model's equilibrium can no longer be obtained as the solution to a central-planning-optimum formulation. These models incorporate various features of monetary phenomena, distortionary taxes, non-competitive labour market arrangements, externalities, or borrowing-lending constraints. Besides increasing general model realism, such features enable the models not only to better replicate the stylized facts of the business cycle, but also to provide a rationale for interventionist government policies. Monetary models of this class include those of Lucas and Stokey (1987, a monetary exchange model) and Coleman (1996, a monetary production model). Bizer and Judd (1989) and Coleman (1991) present models in which non-optimality is induced by tax distortions, while Danthine and Donaldson (1990) present a model in which non-optimality results from efficiency-wage considerations. In these models, equilibrium is characterized as an aggregate-consumption and an aggregate-investment function which jointly solves a system of first-order optimality equations on which market-clearing conditions have been imposed. Coleman (1991) provides a widely applicable set of conditions under which these suboptimal equilibrium functions exist. As already noted, however, these optimality conditions cannot, in general, characterize the solution to an optimum problem.

See Also

- Arrow–Debreu model of general equilibrium
- decentralization
- neoclassical growth theory
- real business cycles

Bibliography

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4, 514–40.

Bizer, D. and Judd, K. 1989. Taxation and uncertainty. *American Economic Review Papers and Proceedings* 19, 331–6.

Brock, W.A. and Mirman, L.J. 1972. Optimal economic growth and uncertainty: the discounted case. *Journal of Economic Theory* 4, 497–513.

Cass, D. 1965. Optimal growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32, 233–40.

Coleman, W.J. 1991. Equilibrium in a production economy with an income tax. *Econometrica* 59, 1091–104.

Coleman, W.J. 1996. Money and output: a test of reverse causation. *American Economic Review* 86, 90–111.

Danthine, J.P. and Donaldson, J.B. 1990. Efficiency wages and the business cycle puzzle. *European Economic Review* 34, 1275–301.

Donaldson, J.B. and Mehra, R. 1983. Stochastic growth with correlated production shock. *Journal of Economic Theory* 29, 282–312.

Donaldson, J.B. and Mehra, R. 1984. Comparative dynamics of an equilibrium intertemporal asset pricing model. *Review of Economic Studies* 51, 491–508.

Harris, M. 1987. *Dynamic Economic Analysis*. New York: Oxford University Press.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–71.

Ljungqvist, L. and Sargent, T.J. 2004. *Recursive Macroeconomic Theory*, 2nd edn. Cambridge, MA: MIT Press.

Long, J.B., Jr. and Plosser, C.I. 1983. Real business cycles. *Journal of Political Economy* 91, 39–69.

Lucas, R.E., Jr. and Stokey, N. 1987. Money and interest in a cash advance economy. *Econometrica* 55, 491–513.

Mehra, R. 1988. On the existence and representation of equilibrium in an economy with growth and non-stationary consumption. *International Economic Review* 29, 131–5.

Mehra, R. and Prescott, E.C. 1977. Recursive competitive equilibria and capital asset pricing. In R. Mehra, Essays in financial economics. Doctoral dissertation, Carnegie Mellon University.

Mehra, R. and Prescott, E.C. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–62.

Prescott, E.C. and Lucas, R.E., Jr. 1972. A note on price systems in infinite dimensional space. *International Economic Review* 13, 416–22.

Prescott, E.C. and Mehra, R. 1980. Recursive competitive equilibria: the case of homogeneous households. *Econometrica* 48, 1365–79.

Stokey, N., Lucas, R.E. and Prescott, E.C. 1989. *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

How to cite this article

Mehra, Rajnish. "recursive competitive equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000252> doi:10.1057/9780230226203.1406

The New Palgrave Dictionary of Economics Online

recursive contracts

Albert Marcet

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A number of dynamic models in economics are formulated with forward-looking elements in the constraints – for example, models of risk-sharing with participation constraints and models of optimal policy. Here, standard dynamic programming does not apply. Recent contributions show how to reformulate these models by either rewriting the forward-looking constraints (promised utility approach) or by using a Lagrangean formulation (recursive Lagrangean). Both make it possible to obtain a recursive formulation that allows for easier computation and analytical results. A number of applications can be found to optimal fiscal or monetary policy, risk sharing or investment with various financial constraints, and employment decisions.

Keywords

Bellman equation; commitment; contract theory; debt constraints; dynamic programming; incentive constraints; international capital flows; Lagrange multipliers; optimal fiscal policy; optimal monetary policy; optimal taxation; participation constraints; principal and agent; private information; Ramsey equilibria; recursive contracts; risk sharing; saddle point functional equations; time consistency; unemployment insurance

Article

In contract theory it is standard to introduce a participation constraint (PC) insuring that the contract offered to the agent delivers a utility higher than the best outside option. In a dynamic set-up agents may abandon the contract at any point in time, even after the contract has been in place for a while. For example, workers can leave a labour contract at almost no cost, or a borrower can stop repaying the loan if he or she declares bankruptcy. The possibility that the agent does not continue with the plan of the contract is usually called ‘default’. Hence, in a dynamic context, it is natural to require that the PC is satisfied in all periods, in order to avoid default.

It turns out that, if a PC in all periods and realizations is introduced in the design of the optimal contract, standard dynamic programming does not apply, the Bellman equation does not hold, and the solution is not guaranteed to be a time-invariant function of the usual state variables. This complicates enormously the solution of these models.

To discuss this in a simple risk-sharing model, consider two agents i=1,2 with utility function $E_0 \Sigma_{t=0}^{\infty} \beta^t u(c_t^i)$, where $\beta \in (0,1)$ is the discount factor and u the instantaneous utility. Each agent receives a stochastic endowment w_t^i and the realization of endowments is known both to the agents and the principal. The principal has full commitment, and will stick to his announced plan. Endowments provide the only supply of consumption good so that the following feasibility condition holds

$$c_t^1 + c_t^2 = w_t^1 + w_t^2 \tag{1}$$

A Pareto-optimal risk-sharing contract (implemented by a competitive equilibrium under complete markets) would set $\frac{u'(c_t^1)}{u'(c_t^2)}$ constant for all periods, so that agents would share all idiosyncratic risks. This allocation would be chosen as the optimal contract if agents would commit to never leave the risk-sharing arrangement. We refer to this allocation as the first best. The optimum satisfies the usual recursive structure in dynamic models, namely, that $c_t=F(w_t)$ where F is a time-invariant function and $w_t = (w_t^1, w_t^2)$.

Assume now agents cannot commit to staying in the contract for ever. An agent can leave the contract and consume for ever his individual endowment, so that a contract can only be implement if it satisfies

$$E_t \sum_{j=0}^{\infty} \beta^j u(c_{t+j}^i) \geq V_i^a(w_t)$$

at all periods and realizations, where $V_i^a(w_t) \equiv E_t \Sigma_{j=0}^{\infty} \beta^j u(w_{t+j}^i)$ is the utility of consuming in autarchy for ever after t.

It is clear that the above PC is likely to be violated by the first best allocation. In periods when w_t^i is high, the right side of the PC is high, but the agent has to surrender a large part of his endowment in the first best and the left side of the PC is too low. Therefore, PCs are often binding and they make the first best unfeasible.

A Pareto-optimal risk-sharing contract with PCs can now be found by maximizing the weighted utility

of the two agents $E_0 \Sigma_{t=0}^{\infty} \beta^t [\lambda u(c_t^1 c) + (1-\lambda) u(c_t^2)]$ subject to the above PC for all periods and realizations and for both agents. The parameter λ indexes all such Pareto-optimal allocations. The result is an optimal contract under full commitment by the principal and partial commitment by the agents.

The Bellman equation does not give the solution to this problem. A key feature of standard dynamic programming is that the set of feasible actions must depend only on variables that were determined last period and the current shock. But it is possible to evaluate if a certain consumption level $\bar{c}_t^i$ satisfies the PC at time *t* only if future plans for consumption are known.

Intuitively, a promise of higher consumption in the future makes a lower consumption today compatible with the PC. But in order to implement this plan the principal has to ‘remember’ all the promises for higher consumption that were made in the past. Therefore, the optimal solution is unlikely to be a function of only today's endowment, the principal also needs to recall if, say, ten periods ago, the PC of one of the agents was binding.

As argued by Kydland and Prescott (1977), the same problem arises in models of optimal policy. The future restricts today's actions through the first order conditions of optimality of the agents, this causes the Bellman equation to fail and, in their language, the solution was time inconsistent. We find the same difficulty in contracting models of private information with incentive constraints, where some relevant piece of information is hidden from the principal, and more generally, in game theoretical models where an agent optimizes subject to the plans for the future of another agent.

If the Bellman equation fails, the solution could depend on all past shocks, and solving for the variables as a function of all past shocks would be very difficult. Too many variables would appear as arguments of the decision function. To overcome this difficulty the ‘recursive contracts’ literature provides several alternatives. The general idea is to recover a recursive formulation by adding a co-state variable.

One approach builds on the paper of Abreu, Pierce and Stachetti (1990; hereafter APS). To show how this can be applied in the above risk-sharing model with PCs, consider the case where w_t is i.i.d. and has two possible realizations $\bar{w}$ and $\bar{\bar{w}}$ with probabilities π and $(1-\pi)$. Denote the utility of agent *i* for the whole future at *t* if $w_t = \bar{w}$ by $\bar{V}_t^i \equiv E_t(\Sigma_{j=0}^{\infty} \beta^j u(c_{t+j}^i) | w_t = \bar{w})$, and let $\bar{\bar{V}}_t^i$ be the analogue for realization $\bar{\bar{w}}$. The above PC can be reformulated as

$$V_t^i = u(c_t^i) + \beta\left(\pi \bar{V}_{t+1}^i + (1-\pi)\bar{\bar{V}}_{t+1}^i\right)$$

(2)

$$\bar{V}_{t+1}^i \geq V_i^a(\bar{w}), \quad \bar{\bar{V}}_{t+1}^i \geq V_i^a(\bar{\bar{w}}) \text{ for all } t > 0,$$

where V_t^i is the actual realized utility. The first equation insures that V_t^i is the expected discounted utility, the second guarantees that the PC holds.

We can view the planner's choice at *t* as choosing the promised utilities $\bar{V}_{t+1}^i$, $\bar{\bar{V}}_{t+1}^i$ and consumption c_t^i, while V_t^i is given by past choices. It is clear that, in the APS approach, today's choice variables $x_t = (\bar{V}_{t+1}^i, \bar{\bar{V}}_{t+1}^i, c_t^i)$ are restricted by yesterday's promised utilities only, and the Bellman equation delivers the optimal contract after the realized V_t^i is included in the list of state variables. The promised utility V_t^i plays the same role as capital in a standard growth model, and (2) plays the role of the transition equation. Therefore, the optimal solution for the choices can be described recursively by a time-invariant function $x_t=F(w_t, V_t)$ for all $t>0$.

A crucial caveat is that (2) is not sufficient to insure that the PCs are satisfied. The principal could choose arbitrarily high consumption and have ever higher *V*s to satisfy (2), in a sort of Ponzi scheme for utility. The promised utilities have to be further restricted to belong to a feasible set. Let us call $\bar{S} \subset R$ the feasible set of utilities such that, for each element $v \in \bar{S}$, there is a sequence of consumptions $\{c_{t+j}^i\}$ that satisfy (1) and the PCs such that $v = E_t(\Sigma_{j=0}^{\infty} \beta^j u(c_{t+j}^i) | w_t = \bar{w})$. Results in APS insure that this set is convex. Since in this case $\bar{S} \subset R$, this set is an interval and there exist bounds $\bar{V}_L^i$ and $\bar{V}_U^i$ such that adding the constraints

$$\bar{V}_L^i \leq \bar{V}_{t+1}^i \leq \bar{V}_U$$

(and similarly for $\bar{\bar{V}}_t^i$) to (2) is enough to insure feasibility. These bounds can be easily introduced in the Bellman equation and this guarantees that the chosen consumption sequences satisfies the PC. The only complication is that upper bound $\bar{V}_U$ needs to be computed separately, as it is not a datum of the problem ($\bar{V}_L^i$ is trivially equal to $V_i^a(\bar{w})$).

Another difference with standard dynamic programming is that the initial utility V_0^1 is an outcome of the solution and it is not fixed beforehand. This feature shows how time inconsistency arises in this model, since the choice for *V* in period zero is not given, but in future periods it is given from the past.

Promised utilities as co-states have been used extensively in models with incentive or participation constraints. Among others, Phelan and Townsend (1991) studied a model of risk-sharing with incentive constraints, Kocherlakota (1996) analysed the risk-sharing model with the PC described above, Hopenhayn and Nicolini (1997) a model of unemployment insurance and Alvarez and Jermann (2000) a

decentralized version of the above risk-sharing model with debt constraints. In models of Ramsey equilibria it has been used by Golosov, Kocherlakota and Tsyvinski (2003) to study optimal taxation under private information and Chang (1998) in a model of optimal monetary policy.

The main problem with this approach is the computation of the set of feasible utilities $\bar{S}$. In the specific model described above this is not too costly, because it involves finding only two numbers, namely, the upper bounds $\bar{V}_U$, $\bar{\bar{V}}_U$. But the difficulties multiply when more than one co-state variable is needed. For example, if a third agent is included in the above risk-sharing model, the co-state variables would be (V_t^1, V_t^2). Results in APS guarantee that the set of feasible utilities $\bar{S} \subset R^2$ is convex, but now it is a generic set, not an interval. Computing a set is much harder than computing two numbers. Some papers overcome these difficulties; for example, Abraham and Pavoni (2005), who show how to find such a set in a model of saving under private information, or the paper of Judd, Yeltekin and Conklin (2003). But the difficulties increase very fast with the dimensionality of the promised utilities.

Furthermore, in some models, the set of feasible promised utilities changes every period. If a 'traditional' state variable (say, capital stock) appears in the problem, the set of feasible utilities is different depending on the level of capital, so that the feasible set is now given by a correspondence $\bar{S}(k)$. The researcher now needs to solve for a mapping from capital stock to sets. Phelan and Stacchetti (2001) compute in this way the optimal fiscal policy in a model with capital.

An alternative to APS is the Lagrangean approach described in Marcet and Marimon (1998). The Lagrangean for the optimal risk-sharing problem with PC is

$$L = E_0 \sum_{t=0}^{\infty} \beta^t \left[\lambda u(c_t^1) + (1-\lambda) u(c_t^2) + \sum_{i=1,2} \gamma_t^i \left(E_t \sum_{j=0}^{\infty} \beta^j u(c_{t+j}^i) - V_i^a(w_t) \right) \right]$$

where $\gamma_t^i \geq 0$ is the Lagrange multiplier of the PC. This can be rewritten as

$$L = E_0 \sum_{t=0}^{\infty} \beta^t [(\lambda + \mu_t^1) u(c_t^1) + (1 - \lambda + \mu_t^2) u(c_t^2)] \text{ s.t. } \mu_t^i = \mu_{t-1}^i + \gamma_t^i, \gamma_t^i \geq 0 \ \mu_{-1}^i = 0$$

In this formulation, only current and past variables enter in the objective and in the constraints of this Lagrangean, and a proper initial condition for μ is given. In this approach, μ_t plays the role of the co-state variable instead of the promised utility in the APS approach. A saddle point functional equation (analogous but not equal to the Bellman equation) is satisfied, insuring that the optimal solution satisfies $(c_t, \gamma_t) = G(\mu_{t-1}, w_t)$ with $\mu_{-1}^i = 0$ for a time invariant function G.

The equilibrium satisfies $\frac{u'(c_t^1)}{u'(c_t^2)} = \frac{1-\lambda+\mu_t^2}{\lambda+\mu_t^1}$. If the PC for agent *i* is binding, the corresponding γ_t^i is strictly positive, the weight μ_t^i goes up and so does c_t^i. The increase in μ is permanent (at least until another PC is binding). In this way the principal avoids default by spreading the reward over time in order to enhance smoothing of consumption.

Note that the initial value of μ is given and equal to zero, while in future periods μ_{t-1} needs to be set according to past Lagrange multipliers. Therefore, the initial value of the co-state does not need to be found separately as in APS. It is clear that, if the principal could re-optimize ignoring past commitments at sometime *t*, he or she would ignore the past co-state and reset $\mu=0$. This is how time inconsistency is reflected in this formulation.

In the Lagrangean approach there is no need to find the set of feasible utilities. The only constraint on the co-states is the non-negativity constraint on γ s. Application to models with capital accumulation and several co-states is much easier; for example, Marcet and Marimon (1992) solve a risk-sharing growth model with PC as described above and capital accumulation, Aiyagari et al. (2002) in a Ramsey equilibrium for fiscal policy under incomplete markets, where debt is a state variable, Attanasio and Ríos-Rull (2000) risk-sharing in small villages, Scott (2007) a model of optimal taxes with capital, Kehoe and Perri (2002) international capital flows with capital accumulation under PC, King, Kahn and Wolman (2003) optimal monetary policy, Cooley, Marimon and Quadrini (2004) a model of investment under private information, Abraham and Carceles-Poveda (2006) discuss how to decentralize a model with participation constraints, and Ferrero and Marcet (2004) and Scholl (2004) a model of temporary exclusion in the case of default. The drawback of the Lagrangean approach is that, at this writing, the theory for the non-convex case and for the private information case is still incomplete.

See Also

- agency problems
- Bellman equation
- dynamic programming
- income taxation and optimal policies
- optimal fiscal and monetary policy (with commitment)
- optimal taxation
- risk sharing
- time consistency of monetary and fiscal policy

Bibliography

Abraham, A. and Carceles-Poveda, E. 2006. Endogenous incomplete markets, enforcement constraints, and intermediation. *Theoretical Economics* 1, 439–59.

Abraham, A. and Pavoni, N. 2005. The efficient allocation of consumption under moral hazard and hidden access to the credit market. *Journal of the European Economic Association* 3, 370–81.

Abreu, D., Pierce, D. and Stachetti, E. 1990. Towards a theory of discounted repeated games with imperfect monitoring. *Econometrica* 58, 1041–63.

Aiyagari, R., Marcet, A., Sargent, T.J., Seppälä, J. 2002. Optimal Taxation without State-Contingent Debt, *Journal of Political Economy* 110, 1220–54.

Alvarez, F. and Jermann, U.J. 2000. Efficiency, equilibrium, and asset pricing with risk of default. *Econometrica* 68, 775–98.

Attanasio, O. and Ríos-Rull, J.V. 2000. Consumption smoothing in island economies: can public insurance reduce welfare? *European Economic Review* 44, 1225–58.

Chang, R. 1998. Credible monetary policy with long-lived agents: recursive approaches. *Journal of Economic Theory* 81, 431–61.

Cooley, T.F., Marimon, R. and Quadrini, V. 2004. Aggregate consequences of limited contract enforceability. *Journal of Political Economy* 112, 817–47.

Ferrero, G. and Marcet, A. 2004. Limited commitment and temporary exclusion. Mimeo, Institut d'Analisi Economica, CSIC.

Golosov, M., Kocherlakota, N.R. and Tsyvinski, A. 2003. Optimal indirect and capital taxation. *Review of Economic Studies* 70, 569–87.

Hopenhayn, H.A. and Nicolini, J.P. 1997. Optimal unemployment insurance. *Journal of Political Economy* 105, 412–38.

Judd, K.L., Yeltekin, S. and Conklin, J. 2003. Computing supergame equilibria. *Econometrica* 71, 1239–54.

Kehoe, P.J. and Perri, F. 2002. International business cycles with endogenous incomplete markets. *Econometrica* 70, 907–28.

King, R.G., Kahn, A. and Wolman, A.L. 2003. Optimal monetary policy. *Review of Economic Studies* 70, 825–60.

Kocherlakota, N.R. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63, 595–609.

Kydland, F.E. and Prescott, E.C. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–92.

Marcet, A. and Marimon, R. 1992. Communication, commitment and growth. *Journal of Economic Theory* 58, 219–49.

Marcet, A. and Marimon, R. 1998. Recursive contracts. Working paper, Universitat Pompeu Fabra.

Phelan, C. and Stacchetti, E., 2001. Sequential equilibria in a Ramsey tax model. *Econometrica* 69, 1491–518.

Phelan, C. and Townsend, R.M. 1991. Computing multi-period, information-constrained optima. *Review of Economic Studies* 58, 853–81.

Scholl, A. 2004. Do endogenous incomplete markets explain cross-country consumption correlations and the dynamics of the terms of trade? Mimeo, Humboldt University, Berlin.

Scott, A. 2007. Optimal taxation and OECD labor taxes. *Journal of Monetary Economics* 54, 925–44.

How to cite this article

Marcet, Albert. "recursive contracts." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000251> doi:10.1057/9780230226203.1407

The New Palgrave Dictionary of Economics Online

recursive preferences

David K. Backus, Bryan R. Routledge, Stanley E. Zin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Recursive preferences characterize the trade-offs between current and future consumption by summarizing the future with a single index, the certainty equivalent of next period's utility. Recursive utility functions are built from two components. A risk aggregator encodes trade-offs across the outcomes of a static gamble and, hence, defines the certainty equivalent of future utility. A time aggregator encodes trade-offs between current consumption and the certainty equivalent of future utility. We suggest functional forms for time and risk aggregators with desirable properties for applications in economics and finance, such as the standard intertemporal consumption/portfolio problem, which we solve using dynamic programming.

Keywords

Bellman equation; certainty equivalent; disappointment aversion; dynamic optimization; elasticity of intertemporal substitution; expected utility; impatience; infinite horizons; preferences; rational expectations; recursive preferences; risk aggregator; risk aversion; stochastic dynamic models; time aggregator; time preference; utility functions; weighted utility

Article

1 Introduction

Recursive methods have become a standard tool for studying economic behaviour in dynamic stochastic environments. In this chapter, we characterize the class of preferences that is the natural complement to this framework, namely *recursive preferences*.

Why model preferences rather than behaviour? Preferences play two critical roles in economic models. First, preferences provide, in principle, an unchanging feature of a model in which agents can be confronted with a wide range of different environments, institutions, or policies. For each environment,

we derive behaviour (decision rules) from the same preferences. If we modelled behaviour directly, we would also have to model how it adjusted to changing circumstances. The second role played by preferences is to allow us to evaluate the welfare effects of changing policies or circumstances. Without the ranking of opportunities that a model of preferences provides, it's not clear how we should distinguish good policies from bad.

Why recursive preferences? Recursive preferences focus on the trade-off between current-period utility and the utility to be derived from all future periods. Since an agent's actions today can affect the evolution of opportunities in the future, summarizing the future consequences of these actions with a single index, that is, future utility, allows multi-period decision problems to be reduced to a series of two-period problems, and in the case of a stationary infinite-horizon problem, a single, time-invariant two-period decision problem. As we will see, this logic applies equally well to environments in which current actions affect the values of random events for all future periods. In this case, the two-period trade-off is between current utility and a *certainty equivalent* of random future utility. This recursive approach not only allows complicated dynamic optimization problems to be characterized as much simpler and more intuitive two-period problems, it also lends itself to straightforward computational methods. Since many computational algorithms for solving stochastic dynamic models themselves rely on recursive methods, numerical versions of recursive utility models can be solved and simulated using standard algorithms.

2 The stationary recursive utility function

Assume time is discrete, with dates t=0,1,2,…. At each t>0, an event z_t is drawn from a finite set Z, following an initial event z_0. The t-period history of events is denoted by $z^t=(z_0, z_1, \ldots, z_t)$ and the set of possible t-histories by Z^t. Environments like this, involving time and uncertainty, are the starting point for much of modern economics. A typical agent in such a setting has preferences over payoffs $c(z^t)$ for each possible history. A general set of preferences might be represented by a utility function $U(\{c(z^t)\})$. In what follows, we will think of consumption as a scalar. This is purely for exposition since the extension to a vector of consumption at each point in time is straightforward.

Consider the structure of preferences in this dynamic stochastic environment. We define the class of stationary recursive preferences by

$$U_t = V[c_t, \mu_t(U_{t+1})], \tag{1}$$

where U_t is short-hand for utility starting at some date-t history z^t, U_{t+1} refers to utilities for histories $z^{t+1}=(z^t, z_{t+1})$ stemming from z^t, V is a time aggregator, and μ_t is a certainty-equivalent function based on the conditional probabilities $p(z_{t+1}|z^t)$. As with other utility functions, increasing functions of U, with suitable adjustment of μ, imply the same preferences. This structure of preferences leads naturally to

recursive solutions of economic problems, with (1) providing the core of a Bellman equation.

In general, the properties of U_t depend on both the properties of the time aggregator and the certainty equivalent. Since the certainty equivalent will be scaled such that $\mu(x)=x$ when x is a perfect certainty, the time aggregator V is all that matters in deterministic settings. Similarly, for a purely static problem with uncertainty, the certainty-equivalent function μ is all that matters. We consider the specification of each of these components in turn.

It is important to note that the utility functions presented in this article are not ad hoc but rather have clear axiomatic foundations, and can be derived from more primitive assumptions on preference orderings. Since utility functions are the typical starting point for applied research, we skip this step and refer the interested reader to the axiomatic characterizations of recursive preferences in the papers cited at the end of this article.

3 The time aggregator

Time preference is a natural starting point. Suppose there is no risk and c_t is one-dimensional. Preferences might then be characterized by a general utility function $U(\{c_t\})$. A common measure of time preference in this setting is the marginal rate of substitution between consumption at two consecutive dates (c_t and c_{t+1}, say) along a constant consumption path ($c_t=c$ for all t). If the marginal rate of substitution is

$$MRS_{t,t+1} = \frac{\partial U / \partial c_{t+1}}{\partial U / \partial c_t},$$

then time preference is captured by the discount factor

$$\beta(c) \equiv MRS_{t,t+1}(c).$$

(Picture the slope, $-1/\beta$, of an indifference curve along the '45-degree line'.) If $\beta(c)$ is less than one, the agent is said to be impatient: along a constant consumption path (that is, in the absence of diminishing marginal utility considerations), the agent requires more than one unit of consumption at $t+1$ to induce a sacrifice of one unit at t.

For the traditional time-additive utility function,

$$U(\{c_t\}) = \sum_{t=0}^{\infty} \beta^t u(c_t),$$

(2)

$\beta(c)=\beta<1$ regardless of the value of c, so impatience is built in and constant. A popular and useful special case of this utility function implies a constant elasticity of intertemporal substitution by assuming $u(c)=c^{\rho}/\rho$ for $\rho<1$. Note that we can define the utility function in (2) recursively:

$$U_t = u(c_t) + \beta U_{t+1},$$
(3)

for t=1,2,…. The constant elasticity version can be expressed

$$U_t = [(1-\beta)c_t^{\rho} + \beta U_{t+1}^{\rho}]^{1/\rho},$$
(4)

where $\rho<1$ and $\sigma=1/(1-\rho)$ is the intertemporal elasticity of substitution. (To put this in additive form, use the transformation $\hat{U} = U^{\rho}/\rho$.) Note that U_t is homothetic and that the scaling we have chosen measures utility on the same scale as consumption:

$$U(c, c, c, \ldots) = c.$$

More generally, impatience summarized by the discount factor, $\beta(c)$, could vary with the level of consumption. Koopmans (1960) derives a class of stationary recursive preferences by imposing conditions on a general utility function U for a multi-dimensional consumption vector c. In the Koopmans class of preferences, time preference is a property of the time aggregator V. Consider our measure of time preference:

$$U_t = V(c_t, U_{t+1}) = V[c_t, V(c_{t+1}, U_{t+2})].$$

The marginal rate of substitution between c_t and c_{t+1} is therefore

$$\mathrm{MRS}_{t,t+1} = \frac{V_2(c_t, U_{t+1})V_1(c_{t+1}, U_{t+2})}{V_1(c_t, U_{t+1})}.$$

A constant consumption path at c is defined by $U=V(c, U)$, implying $U=g(c)=V[c, g(c)]$ for some function g.

In modern applications, we typically work in reverse order: we specify a time aggregator V and use it to characterize the overall utility function U. Any U constructed this way defines preferences that are stationary and dynamically consistent. In contrast to time-additive preferences, discounting depends on the level of consumption c.

The most common example of Koopmans's structure in applications is a generalization of eq. (3):

$$V(c, U) = u(c) + \beta(c)U,$$

where there is no particular relationship between the functions u and β. For this example, the intertemporal trade-off is given by

$$\mathrm{MRS}_{t,t+1} = \beta(c_t)\left[\frac{u'(c_{t+1}) + \beta'(c_{t+1})U_{t+2}}{u'(c_t) + \beta'(c_t)U_{t+1}}\right].$$

When $\beta'(c) \neq 0$, optimal consumption plans will depend on the level of future utility. And along a constant consumption path, discounting is decreasing (increasing) in consumption when $\beta'(c) < 0$ ($\beta'(c) > 0$). Also note that U_t in this example is not homothetic.

4 The risk aggregator

Turn now to the specification of risk preferences, which we consider initially in a static setting. Choices have risky consequences or payoffs, and agents have preferences defined over those consequences and their probabilities. To be specific, let us say that the state z is drawn with probability $p(z)$ from the finite set $\mathcal{Z}=\{1,2,\ldots, Z\}$. Consequences ($c$, say) depend on the state and the agent's preferences are represented

by a utility function of state-contingent consequences (‘consumption’):

$$U(\{c(z)\}) = U[c(1), c(2), \ldots, c(Z)].$$

At this level of generality there is no mention of probabilities, although we can well imagine that the probabilities of the various states will show up somehow in *U*. We regard the probabilities as known, which you might think of as an assumption of ‘rational expectations’.

We prefer to work with a different (but equivalent) representation of preferences. Suppose, for the time being, that *c* is a scalar; very little of the theory depends on this, but it streamlines the presentation. We define the *certainty equivalent* of a set of consequences as a certain consequence μ that gives the same level of utility:

$$U(\mu, \mu, \ldots, \mu) = U[c(1), c(2), \ldots, c(Z)].$$

If *U* is increasing in all its arguments, we can solve this for the certainty-equivalent function $\mu(\{c(z)\})$. Clearly μ represents the same preferences as *U*, but we find its form particularly useful. For one thing, it expresses utility in payoff (‘consumption’) units. For another, it summarizes behaviour towards risk directly: since the certainty equivalent of a sure thing is itself, the impact of risk is simply the difference between the certainty equivalent and expected consumption.

The traditional approach to preferences in this setting is expected utility, which takes the form

$$U(\{c(z)\}) = \sum_{Z} p(s)u[c(z)] = Eu(c),$$

or

$$\mu(\{c(z)\}) = u^{-1}\left(\sum_{Z} p(z)u[c(z)]\right) = u^{-1}[Eu(c)].$$

Preferences of this form have been used in virtually all economic theory. The utility function of Kreps and Porteus employs a general time aggregator and an expected utility certainty equivalent. Following Epstein and Zin, many recent applications, particularly in dynamic asset pricing models, use the

homothetic version of this utility function which combines the constant elasticity time aggregator in (4) with a linear homogeneous (constant relative risk aversion) expected utility certainty equivalent.
Empirical research both in the laboratory and in the field has documented a variety of difficulties with the predictions of expected utility models. In particular, people seem more averse to bad outcomes than implied by expected utility. In response to this evidence, there is a growing body of work that looks at decision making under uncertainty outside of the traditional expected utility framework. Without surveying all of these extensions, we demonstrate the basic mechanics of recursive utility with non-expected utility certainty equivalents by studying one particular analytically convenient class of preferences in detail, the Chew–Dekel class. Notable among the alternatives to this class are recursive and dynamic extensions of the Gilboa and Schmeidler ‘max-min’ preferences.
The Chew–Dekel certainty equivalent function μ for a set of payoffs and probabilities $\{c(z), p(z)\}$ is defined implicitly by a *risk aggregator M* satisfying

$$\mu = \sum_z p(z) M[c(z), \mu]. \tag{5}$$

Such preferences satisfy a weaker condition than the notorious independence axiom that underlies expected utility, yet like expected utility, they lead to first-order conditions in decision problems that are linear in probabilities, hence easily solved and amenable to econometric analysis. We assume M has the following properties: (i) $M(m, m)=m$ (sure things are their own certainty equivalents), (ii) M is increasing in its first argument (first-order stochastic dominance), (iii) M is concave in its first argument (risk aversion), and (iv) $M(kc, km)=kM(c, m)$ for $k>0$ (linear homogeneity). Most of the analytical convenience of the Chew–Dekel class follows from the linearity of eq. (5) in probabilities. (Note that this implies that indifference curves on the probability simplex are linear, but not necessarily parallel.) Examples of tractable members of the Chew–Dekel class include the following:

1. 1. *Expected utility*. A version with constant relative risk aversion (that is, linear homogeneity) is implied by

$$M(c, m) = c^{\alpha} m^{1-\alpha} / \alpha + m(1 - 1/\alpha).$$

If $\alpha \leq 1$, M satisfies the conditions outlined above. Applying (5), we find

$$\mu = \left(\sum_z p(z) c(z)^{\alpha}\right)^{1/\alpha},$$

the usual expected utility with a power utility function.

2. 2. *Weighted utility*. A relatively easy way to generalize expected utility given (5): weight the probabilities by a function of outcomes. A constant-elasticity version follows from

$$M(c, m) = (c/m)^{\gamma} c^{\alpha} m^{1-\alpha}/\alpha + m[1 - (c/m)^{\gamma}/\alpha].$$

For *M* to be increasing and concave in *c* in a neighbourhood of *m*, the parameters must satisfy either (a) $0 < \gamma < 1$ and $\alpha + \gamma < 0$ or (b) $\gamma < 0$ and $0 < \alpha + \gamma < 1$. Note that (a) implies $\alpha < 0$, (b) implies $\alpha > 0$, and both imply $\alpha + 2\gamma < 1$. The associated certainty equivalent function is

$$\mu^{\alpha} = \frac{\Sigma_z p(z) c(z)^{\gamma+\alpha}}{\Sigma_x p(x) c(x)^{\gamma}} = \sum_z \hat{p}(z) c(z)^{\alpha},$$

where

$$\hat{p}(z) = \frac{p(z) c(z)^{\gamma}}{\Sigma_x p(x) c(x)^{\gamma}}.$$

This version highlights the impact of bad outcomes: they get greater weight than with expected utility if $\gamma < 0$, less weight otherwise.

3. 3. *Disappointment aversion*. Another model that increases sensitivity to bad events ‘disappointments’) is defined by the risk aggregator

$$M(c, m) = \begin{cases} c^{\alpha} m^{1-\alpha}/\alpha + m(1 - 1/\alpha) & c \geq m \\ c^{\alpha} m^{1-\alpha}/\alpha + m(1 - 1/\alpha) + \delta(c^{\alpha} m^{1-\alpha} - m)/\alpha & c < m \end{cases}$$

with $\delta \geq 0$. When $\delta = 0$ this reduces to expected utility. Otherwise, disappointment aversion places additional weight on outcomes worse than the certainty equivalent. The certainty equivalent function satisfies

$$\mu^{\alpha} = \sum_{z} p(z)c(z)^{\alpha} + \delta \sum_{z} p(z)I[c(z) < \mu][c(z)^{\alpha} - \mu^{\alpha}] = \sum_{z} \hat{p}(z)c(z)^{\alpha},$$

where $I(x)$ is an indicator function that equals one if x is true and zero otherwise, and

$$\hat{p}(z) = \left(\frac{1 + \delta I[c(z) < \mu]}{1 + \delta \Sigma_{x} p(x) I[c(x) < \mu]}\right) p(z).$$

It differs from weighted utility in scaling up the probabilities of all bad events by the same factor, and scaling down the probabilities of good events by a complementary factor, with good and bad defined as better and worse than the certainty equivalent. (This implies a ‘kink’ in state-space indifference curves at certainty, which is referred to as ‘first-order’ risk aversion.) All three expressions highlight the *recursive* nature of the risk aggregator M: we need to know the certainty equivalent to know which states are bad so that we can compute the certainty equivalent (and so on).

5 Optimization and the Bellman equation

For an illustrative application of recursive utility, we turn to the classic Merton–Samuelson consumption/portfolio-choice problem. Consider a stationary Markov environment with states z and conditional probabilities $p(z'|z)$. Preferences are represented by a constant-discounting/constant-elasticity aggregator and a general linear homogeneous certainty equivalent. A dynamic consumption/portfolio problem for this environment is characterized by the Bellman equation which implicitly defines the value function:

$$J(a, z) = \max_{c, w} \left\{(1 - \beta)c^{\rho} + \beta\mu[J(a', z')]^{\rho}\right\}^{1/\rho},$$

subject to the wealth constraint, $a' = (a - c)\Sigma_{i} w_i r_i(z, z') = (a - c)\Sigma_{i} w_i r_i' = (a - c)r_p'$, where a

denotes wealth, r_p is the return on the portfolio $(w_1, w_2, \ldots, w_{N-1}, 1 - \Sigma_{i=1}^{N-1} w_i)$, of assets with risky returns $(r_1, r_2, \ldots, r_N)$. The budget constraint and linear homogeneity of the time and risk aggregators imply linear homogeneity of the value function: $J(a, z) = aL(z)$ for some scaled value function L. The scaled Bellman equation is

$$L(z) = \max_{b,w} \left\{ (1-\beta) b^{\rho} + \beta (1-b)^{\rho} \mu [L(z') r_p(z, z')]^{\rho} \right\}^{1/\rho},$$

where $b \equiv c / a$. Note that $L(z)$ is the marginal utility of wealth in state z.
This problem divides into separate portfolio and consumption decisions. The portfolio decision solves: choose $\{w_i\}$ to maximize $\mu[L(z') r_p(z, z')]$. The portfolio first-order conditions are

$$\sum_{z'} p(z'|z) M_1 [L(z') r_p(z, z'), \mu] L(z') [r_i(z, z') - r_j(z, z')] = 0 \tag{6}$$

for any two assets i and j.
Given a maximized μ, the consumption decision solves: choose b to maximize L. The intertemporal first-order condition is

$$(1-\beta) b^{\rho-1} = \beta (1-b)^{\rho-1} \mu^{\rho}. \tag{7}$$

If we solve for μ and substitute into the (scaled) Bellman equation, we find

$$\mu = [(1-\beta)/\beta]^{1/\rho} [b/(1-b)]^{(\rho-1)/\rho}$$

$$L = (1-\beta)^{1/\rho} b^{(\rho-1)/\rho}. \tag{8}$$

The first-order condition (7) and value function (8) allow us to express the relation between consumption and returns in a familiar form. Since μ is linear homogeneous, the first-order condition implies $\mu(x' r_p') = 1$ for

$$x' = L' / \mu = \left[\beta (c'/c)^{\rho-1} (r_p')^{1-\rho}\right]^{1/\rho}.$$

The last equality follows from $(c'/c) = (b'/b)(1-b) r_p'$, a consequence of the budget constraint and the definition of *b*. The intertemporal first-order condition can therefore be expressed

$$\mu(x' r_p') = \mu\left(\left[\beta (c'/c)^{\rho-1} r_p'\right]^{1/\rho}\right) = 1, \tag{9}$$

a generalization of the tangency condition for an optimum (set the marginal rate of substitution equal to the price ratio). Similar logic leads us to express the portfolio first-order conditions (6) as

$$E\left[M_1(x' r_p', 1) x' (r_i' - r_j')\right] = 0.$$

If we multiply by the portfolio weight w_j and sum over *j* we find

$$E\left[M_1(x' r_p', 1) x' r_i'\right] = E\left[M_1(x' r_p', 1) x' r_p'\right]. \tag{10}$$

Euler's theorem for homogeneous functions allows us to express the right side as

$$E\left[M_1(x'r'_p, 1)x'r'_p\right] = 1 - EM_2(x'r'_p, 1).$$

Whether this expression is helpful depends on the precise form of *M*. For example, with disappointment aversion, (10) is

$$E\left[z^{\alpha-1}(1+\delta I[z<1])\frac{r'_i}{r'_p}\right] = 1 + \delta EI[z<1],$$

where $z = [\beta(c'/c)^{\rho-1}r'_p]^{1/\rho}$. This reduces to the Kreps–Porteus model when $\delta = 0$, and to the time-additive expected utility model when, in addition, $\rho = \alpha$.

6 Conclusion

A recursive utility function can be constructed from two components: (*a*) a time aggregator that completely characterizes preferences in the absence of uncertainty and (*b*) a risk aggregator that defines the certainty equivalent function that characterizes preferences over static gambles and is used to aggregate the risk associated with future utility. We looked at natural candidates for each of these components and gave an example of how Bellman's equation can be used to characterize optimal plans in a dynamic stochastic environment when agents have recursive preferences.

7 Further reading

For more on this subject, see Backus, Routledge and Zin (2004) and the references cited there. Much of the material in this chapter builds from Epstein and Zin (1989), who extend the preferences in Kreps and Porteus (1978) to allow for a stationary infinite-horizon model and for non-expected utility certainty equivalents. They also derive the consumption/portfolio-choice results of Section (5). For more on time aggregators, see Koopmans (1960), Uzawa (1968), Epstein and Hynes (1983), Lucas and Stokey (1984), and Shi (1994). Common departures from expected utility are documented in Kreps (1988, ch. 14) and Starmer (2000). Epstein and Schneider (2003) and Hansen and Sargent (2004) propose different dynamic and recursive extensions of the max-min risk preference of Gilboa and Schmeidler (1993). The Chew–Dekel risk aggregator was proposed by Chew (1983; 1989) and Dekel (1986). Examples within this class: weighted utility (Chew, 1983), disappointment aversion (Gul, 1991), semi-weighted utility

(Epstein and Zin, 2001), and generalized disappointment aversion (Routledge and Zin, 2003).

See Also

- Bellman equation
- time preference

Bibliography

Backus, D.K., Routledge, B.R. and Zin, S.E. 2004. Exotic preferences for macroeconomists. In *NBER Macroeconomics Annual 2004*, vol. 19, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Chew, S.H. 1983. A generalization of the quasi-linear mean with applications to the measurement of inequality and decision theory resolving the Allais paradox. *Econometrica* 51, 1065–92.

Chew, S.H. 1989. Axiomatic utility theories with the betweenness property. *Annals of Operations Research* 19, 273–98.

Dekel, E. 1986. An axiomatic characterization of preferences under uncertainty: weakening the independence axiom. *Journal of Economic Theory* 40, 304–18.

Epstein, L.G. and Hynes, J.A. 1983. The rate of time preference and dynamic economic analysis. *Journal of Political Economy* 91, 611–35.

Epstein, L.G. and Schneider, M. 2003. Recursive multiple-priors. *Journal of Economic Theory* 113, 1–31.

Epstein, L.G. and Zin, S.E. 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. *Econometrica* 57, 937–69.

Epstein, L.G. and Zin, S.E. 2001. The independence axiom and asset returns. *Journal of Empirical Finance* 8, 537–72.

Gilboa, I. and Schmeidler, D. 1993. Updating ambiguous beliefs. *Journal of Economic Theory* 59, 33–49.

Gul, F. 1991. A theory of disappointment aversion. *Econometrica* 59, 667–86.

Hansen, L.P. and Sargent, T.J. 2004. Misspecification in recursive macroeconomic theory. Manuscript, University of Chicago, January.

Koopmans, T.C. 1960. Stationary ordinal utility and impatience. *Econometrica* 28, 287–309.

Kreps, D.M. 1988. *Notes on the Theory of Choice*. Boulder, CO: Westview Press.

Kreps, D.M. and Porteus, E.L. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46, 185–200.

Lucas, R.E. and Stokey, N.L. 1984. Optimal growth with many consumers. *Journal of Economic Theory* 32, 139–71.

Routledge, B.R. and Zin, S.E. 2003. Generalized disappointment aversion and asset prices. Working Paper No. 10107. Cambridge, MA: NBER.

Shi, S. 1994. Weakly nonseparable preferences and distortionary taxes in a small open economy. *International Economic Review* 35, 411–28.

Starmer, C. 2000. Developments in non-expected utility theory. *Journal of Economic Literature* 38, 332–82.

Uzawa, H. 1968. Time preference, the consumption function, and optimum asset holdings. In *Value, Capital, and Growth: Papers in Honour of Sir John Hicks*, ed. J.N. Wolfe. Chicago: Aldine.

How to cite this article

Backus, David K., Bryan R. Routledge and Stanley E. Zin. "recursive preferences." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000260> doi:10.1057/9780230226203.1408

The New Palgrave Dictionary of Economics Online

redistribution of income and wealth

F.A. Cowell

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

education; efficiency; equity; inequality; multiple equilibria; pensions; price control; progressive and regressive taxation; redistribution of income and wealth; social insurance

Article

The topic of redistribution is sometimes interpreted narrowly in rather dry terms: as the description and quantification of the simple fact of change in an income or wealth distribution. This can apply both to an actual change that takes place through time and also to the apparent alteration of the distribution at a point in time by taxes and transfers, and principally involves problems of measurement that are common to other fields of applied economics. However, redistribution can also be seen as a specific goal for economic policymakers: as such it is a subject of special interest in its own right. Sections 1–4 below concentrate primarily on this second interpretation; some issues arising under the first interpretation are considered in Section 5.

1 The reason for wanting to redistribute

Perhaps the simplest and most direct reason for wishing to see a redistribution of income, consumption or wealth in the community is simple fellow feeling on the part of the citizens of the community. This can be incorporated into the utilitarian approach to welfare judgements within the tradition of Bentham and Mill, in two ways. One might suppose that judgements about distribution are made in a state of primordial ignorance about one's own position in the distribution: social aversion to inequality is thus rationalized as individual aversion to risk (Harsanyi, 1955). Secondly, it might be supposed that the poor are made to feel worse off in their plight by the very knowledge that the well-to-do are well-to-do, and the rich are made to feel uncomfortable by the low living standards of the poor – see Hochman and Rodgers (1969). Thus the problems of inequality are rationalized within individual utilities as ‘externalities’ in a manner similar to health hazards from pollution. A weakness of this approach is that

it puts a heavy burden on the particular configuration of individual preferences that happen to be present within a given community at a given moment: should one *really* only redistribute if enough citizens happen to feel upset by it? And what if some citizens *like* knowing that the very poor are very poor? An alternative approach is to take the motivation for redistribution as a direct moral imperative – see Tawney (1965), Rawls (1971); improvement in the well-being of the disadvantaged is perceived as a social objective in its own right, along with other apparently desirable goals such as civil liberties and growth in national income.

2 The objectives of redistribution

Whatever the precise reasons for wishing to redistribute income or wealth may be, in *broad* terms the principal goals of redistribution policy can be stated very simply: the primary objectives are usually some goal of greater equity and of 'social insurance'; and as a secondary, though important, desideratum, one is usually also concerned with economic efficiency.

In order to examine these objectives in more detail two concepts need to be carefully distinguished: redistribution '*ex ante*' – the rearrangement of the structure of *income-earning opportunities* – and redistribution '*ex post*' – the reallocation of income or wealth that results from the economic processes of production and exchange, whatever individual opportunities may have been. In practice the two concepts may be difficult to disentangle since a policy measure that apparently just rearranges the prizes (such as an income-tax scheme) may also have repercussions on some people's *ex ante* opportunities (by, for example, affecting market wages); but both are relevant to a discussion of the relationship between equity and other goals.

In a very simplified model of the distribution of income, 'equity' can be expressed fairly easily: if one considers that the cake has been cut very unequally, then one sets about trying to even up the slices. But in a dynamic view of the economy where people make economic choices which affect their future incomes, the slices-of-a-fixed-size-cake analogy can be misleading, and the position may be further complicated when those choices have to be made in the face of uncertainty. Obviously the size of the national cake to be 'shared out' is not, in practice, fixed: individual incomes (and hence the total income in the community) are determined by the choices people make as to how much they work, save, and take entrepreneurial risks, and again the total stock of wealth obviously also depends on the rate at which people save. So the elementary equity question of who ought to get what cannot be divorced in practice from the issue of how individual incomes and wealth holdings are generated: efficiency considerations have to be taken into account in the pursuit of greater equity. There is a second, more subtle, difficulty: because of incompetence, ignorance or plain 'bad luck' people who may have looked alike in terms of their original economic opportunities turn out to be very dissimilar in terms of outcomes once a few rounds have been played of the great economic game that determines how much everybody actually gets. Hence there is a good case for a government concerned with distributional equity to pay attention to both the *ex ante* and the *ex post* concepts of redistribution (Hammond, 1981).

For this reason an interest in social insurance is often taken to be a natural counterpart of a concern for equity. The public provision of protection against the slings and arrows of outrageous fortune is particularly important for those events for which conventional insurance markets are likely to give inadequate coverage, such as unemployment or ill health, for example – see Atkinson (1987). By filling

such gaps social insurance may actually improve the efficient working of the economy. Besides this, social insurance can also apply to *ex post* redistribution that is intended to circumvent the otherwise unsatisfactory workings of some markets. For example, the markets for private insurance and savings might, under ideal circumstances, allow people to look after themselves effectively; but in practice problems such as imperfect information and the consequent rationing of insurance of credit to those people who are perceived to be good risks will mean that coverage is far from complete (Arrow, 1985). Hence the provision of state pensions as a means of cushioning the possibly unfortunate effects of restrictions on savings by people of modest means.

3 What should be redistributed?

Whether it is *income* (the flow of spending power during a given period) or *wealth* (the command over resources that a person may possess at a given point in time) that is to be redistributed depends to some extent on the precise definition of these terms (in particular the relevant period over which income is measured and the range of assets to be counted in as personal wealth) and also on the degree of importance that one attaches to *ex ante* or *ex post* concepts of redistribution. For example, some components of wealth (land, financial assets) may be regarded as part of the range of economic opportunities which results in the flow of spendable income. Again weekly money income might be more relevant than broader concepts of wealth or long-term income if one's primary concern is for redistribution to alleviate short-term need rather than to alter the structure of economic opportunities (Atkinson, 1983, ch. 3).

However, the issue of what one ought to use in order to achieve the objectives of redistribution cited above raises further questions. One of the most important of these is whether one ought to redistribute income itself (which yields purchasing power over consumption goods) or rather the consumption goods directly. The standard answer provided by economists is that cash is unquestionably more effective, since it allows individuals to be the judges of what is best for their welfare and to make substitutions between different goods under varying market conditions in pursuit of that welfare: money to buy soup is supposedly more effective than the provision of soup kitchens. However, this conclusion is strictly relevant only if one imposes a number of stringent conditions, for example, the assumption that everyone has access to perfect market opportunities and accurate information on which to base his judgement in the market. It is invalid in the presence of multiple market equilibria (Foldes, 1967). It ignores pressing requirements of crises such as war and famine: extreme circumstances may require direct intervention to act more swiftly and reliably to maintain living standards than the often capricious and sluggish movements of the ‘invisible hand’.

4 The available instruments

Among the more obvious instruments available for *ex post* redistribution are taxes on income, wealth and the transfer of wealth via gifts and bequests, and transfer payments such as pensions and social-security benefits. However, it is not easy to draw a hard-and-fast line around the range of instruments that might be taken to be redistributive tools, particularly if one is concerned with description rather than prescription. There appears to be a good case in practice for including ‘indirect’ taxes (such as value

added tax), subsidies and also those benefits ‘in kind’ which are bestowed on *particular* households or persons, since the impact of these items on personal spending power is usually fairly clear. This may, for example, be extended to include such goods as state-provided education. However the precise distributional impact of publicly provided goods that are really consumed *jointly* by the community (in which category we might include items such as public sanitation, the police services, or even national defence) is less easy to determine, but should not be assumed to be negligible.

As an alternative to raising taxes and the public provision of goods and services, a government wishing to redistribute real spending power may choose to intervene directly in the market mechanism. The most obvious example of this policy is price control. This term applies not only to rationing and the regulation of prices paid by consumers for goods – which can be an effective method of intervention to achieve redistribution – in emergencies such as wartime, but also to the control of prices that individuals receive for services that they may supply (for example, minimum wage legislation) or assets that they possess (control of house rents).

The instruments available for the purposes of *ex ante* redistribution (that is, the means of reorganizing the *opportunities* for creating income and accumulating wealth) are more disparate. One has the immediate problem that the range of policies considered to be available is strongly influenced by the economic philosophy which one considers to be relevant to the analysis and by the political and social system within the community. Take a prime example of this: education. There are many opinions on the potential for using this as a redistributional tool, some of which may be crudely summarized by the following three views: (*a*) it is a passport to higher positions on a ladder of economic opportunity whose rungs are pretty rigidly fixed, so that greater equality can be achieved simply by changing the method of issuing the passports; (*b*) it forms part of a complex of personal or family investment decisions, whereby intervention in the provision of education might upset the efficient allocation of the market mechanism without doing anything to alleviate the inequality of economic opportunity; (*c*) even if effective redistribution *could* be achieved in principle, substantial reorganization of educational opportunities is bound to be limited by what are seen as fundamental freedoms of choice. Note that the divergence of view concerns both economic role of education and the extent to which one is free to use it as an instrument of public policy (Le Grand, 1982, ch. 4).

5 The effectiveness of the policy

Any attempt to quantify the effectiveness of redistribution policy has to surmount a number of extremely troublesome obstacles.

In the first place one has to confront the problem of ‘unequal inequalities’, which essentially arises from an attempt to compare intrinsically complex social states. Even if one puts this in elementary terms, whereby every person's welfare is accurately measured by his or her income, a fundamental difficulty remains: apart from special circumstances – for example, a comparison involving a hypothetical state of perfect equality – the question of which of two distributions is the more unequal does not generally have a clear-cut answer. In practice, even a very successful redistribution policy will have diminished rather than completely eliminated real income differences, so that an assessment of the policy's impact necessarily involves a comparison of the apparent change in inequality that has been achieved relative to the degree of inequality that would have obtained otherwise. There is no single method for measuring such inequality changes that commands universal support, and hence no generally agreed measuring rod

to ascertain the extent of redistribution under all circumstances (Cowell, 1977; Foster, 1985). One of the practical difficulties to which this gives rise is that it is difficult to be dogmatic about labelling policy instruments in terms of degrees of ‘progressivity’ (Lambert, 1985). Moreover, in many cases redistribution may involve not just a narrowing (or indeed expansion) of income differentials, but also a *re-ranking* of income receivers within the pecking order so that, to quantify redistribution effectively, more is required than a simple measurement of the change in overall dispersion (Cowell, 1985).

The second problem follows directly from this: who is to say what *would* have happened otherwise and, therefore, what change in inequality has actually been achieved? If one is merely concerned with the documentation of trends in the perceived inequities of income distribution through time, this may not be too difficult. But if at any moment of history one attempts to draw the inference that ‘according to our chosen inequality index, the inequality of disposable income would have been 20 per cent higher than it is now but for the high marginal tax rates on upper income groups’, then one is making a much bolder assertion about how the underlying economic mechanisms are supposed to work. For the very presence of the instruments of redistribution policy will have influenced the choices people make about their jobs, business enterprises and savings, which in turn, can be expected to affect the resulting income distribution. The ‘distribution before tax’ – obtainable from a statistical office's published figures – cannot automatically be taken to be the same thing as the distribution *without* the tax – the income distribution that one might expect to see if the relevant redistribution instrument were to be abolished. Some allowance for this problem is usually possible in the case of *ex post* redistribution instruments – for example, it is possible to estimate the likely repercussion on the supply of different types of labour that will arise because of the supplementation of some people's incomes by public transfers and the reduction in other people's incomes through taxation (Hausman, 1985; Killingsworth, 1983), or the impact on private savings of the presence of state-provided pensions and social insurance schemes (Danziger, Haveman and Plotnick, 1981; Kotlikoff, 1984). However, the allowance to be made for these feedback effects is usually quite sensitive to the particular model of household behaviour that is applied.

Despite these reservations, some broad conclusions are possible. Very narrowly based measures run the danger of the ‘demarcation trap’: for example, subsidizing particular commodities or taxing only certain forms of income or wealth may present some people with an incentive to change their behaviour, or even misrepresent their true status, so as to profit by the artificial distinctions drawn by the selective tax or subsidy scheme. The effectiveness of the measure may thereby be reduced and, even if this does not happen, the discrimination of the scheme may itself create substantial inequities by treating essentially similar people in different ways. On the other hand, very broadly based measures may scatter their shot so widely that much of it misses the target: blanket allowances or exceptions within income- or wealth-tax laws, and some broadly defined educational subsidies are often found to be regressive in their actual *ex post* impact on income and wealth. Finally it is usually the case that *taxes*, taken as a whole, turn out not to be very progressive in terms of their *ex post* impact whereas *transfers* usually are.

See Also

- progressive and regressive taxation
- social insurance

Bibliography

Arrow, K.J. 1985. The economics of agency. In *Principals and Agents: The Structure of Business*, ed. J. Pratt and R. Zeckhauser. Cambridge, MA: Harvard Business School Press.

Atkinson, A.B. 1983. *The Economics of Inequality*, 2nd edn. Oxford: Oxford University Press.

Atkinson, A.B. 1987. Income maintenance and social insurance: a survey. In *Handbook of Public Economics*, vol. 2, ed. A.J. Auerbach and M.S. Feldstein. Amsterdam: North-Holland.

Cowell, F.A. 1977. *Measuring Inequality*. Oxford: Philip Allan.

Cowell, F.A. 1985. Measures of distributional change: an axiomatic approach. *Review of Economic Studies* 52, 135–51.

Danziger, S., Haveman, R.H. and Plotnick, R. 1981. How income transfer programs affect work, savings and the income distribution: a critical review. *Journal of Economic Literature* 19, 975–1028.

Foldes, L.P. 1967. Income redistribution in money and in kind. *Economica* 34, 30–41.

Foster, J. 1985. Inequality measurement. In *Fair Allocation*, ed. H.P. Young. Providence, RI: American Mathematical Society.

Hammond, P.J. 1981. *Ex-ante* and *ex-post* welfare economics. *Economica* 48, 235–50.

Harsanyi, J.C. 1955. Cardinal welfare, individualist ethics and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Hausman, J.A. 1985. Taxation and labour supply. In *Handbook of Public Economics*, vol. 1, ed. A.J. Auerbach and M.S. Feldstein. Amsterdam: North-Holland.

Hochman, J.M. and Rodgers, J.D. 1969. Pareto optimal redistribution. *American Economic Review* 59, 542–57.

Killingsworth, M.R. 1983. *Labour Supply*. Cambridge: Cambridge University Press.

Kotlikoff, L.J. 1984. Taxation and savings: a neoclassical perspective. *Journal of Economic Literature* 22, 1576–629.

Lambert, P.J. 1985. Tax-progressivity: a survey of the literature. Working Paper. London: Institute for Fiscal Studies.

Le Grand, J. 1982. *The Strategy of Equality*. London: Allen & Unwin.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Tawney, R.H. 1965. *Equality*. London: Allen & Unwin.

How to cite this article

Cowell, F. A. "redistribution of income and wealth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000067> doi:10.1057/9780230226203.1409

The New Palgrave Dictionary of Economics Online

reduced rank regression

Søren Johansen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The reduced rank regression model is a multivariate regression model with a coefficient matrix with reduced rank. The reduced rank regression algorithm is an estimation procedure which estimates the reduced rank regression model. It is related to canonical correlations and involves calculating eigenvalues and eigenvectors. We give a number of different applications to regression and time series analysis, and show how the reduced rank regression estimator can be derived as a Gaussian maximum likelihood estimator. We briefly mention asymptotic results.

Keywords

instrumental variable estimation; limited information maximum likelihood; maximum likelihood; reduced rank regression

Article

Reduced rank regression is an explicit estimation method in multivariate regression that takes into account the reduced rank restriction on the coefficient matrix.

Reduced rank regression model: We consider the multivariate regression of *Y* on *X* and *Z* of dimension *p*, *q*, and *k*, respectively: $Y_t = \Pi X_t + \Gamma Z_t + \varepsilon_t,\ t = 1, \ldots, T$. The hypothesis that Π has reduced rank less than or equal to *r* is expressed as $\Pi = \alpha\beta'$, where α is $p \times r$, and β is $q \times r$, where $r < \min(p,q)$, and gives the reduced rank model

$$Y_t = \alpha\beta' X_t + \Gamma Z_t + \varepsilon_t,\ t = 1, \ldots, T. \quad (1)$$

Reduced rank regression algorithm: In order to describe the algorithm, which we call *RRR*(*Y*, *X*|*Z*), we introduce the notation for product moments $S_{yx} = T^{-1}\Sigma_{t=1}^{T} Y_t X_t'$, $S_{yx.z} = S_{yx} - S_{yz}S_{zz}^{-1}S_{zx}$, and so on. The algorithm consists of the following steps:

1. 1. First, regress *Y* and *X* on *Z* and form residuals $(Y|Z)_t = Y_t - S_{yz}S_{zz}^{-1}Z_t$, $(X|Z)_t = X_t - S_{xz}S_{zz}^{-1}Z_t$ and product moments

$$S_{yx.z} = T^{-1}\sum_{t=1}^{T}(Y|Z)_t(X|Z)_t' = S_{yx} - S_{yz}S_{zz}^{-1}S_{zx},$$

   and so on.
2. 2. Next, solve the eigenvalue problem

$$\left|\lambda S_{xx.z} - S_{xy.z} S_{yy.z}^{-1} S_{yx.z}\right| = 0, \tag{2}$$

where |.| denotes determinant. The ordered eigenvalues are $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ and the eigenvectors are $V=(v_1,\ldots,v_q)$, so that $S_{xx.z} V \Lambda = S_{xy.z} S_{yy.z}^{-1} S_{yx.z} V$, and V is normalized so that $V' S_{xx.z} V = I_p$ and $V' S_{yx.z} S_{xx.z}^{-1} S_{xy.z} V = \Lambda$. The singular value decomposition provides an efficient way of implementing this procedure; see Doornik and O'Brien (2002).

3. 3. Finally, define the estimators

$$\hat{\beta} = (v_1, \ldots, v_r)$$

together with $\hat{\alpha} = S_{yx.z}\hat{\beta}$, and $\hat{\Omega} = S_{yy.z} - S_{yx.z}\hat{\beta}(\hat{\beta}' S_{xx.z}\hat{\beta})^{-1}\hat{\beta}' S_{xy.z}$. Equivalently, once $\hat{\beta}$ has been determined, $\hat{\alpha}$ and $\hat{\Gamma}$ are determined by regression.

The technique of reduced rank regression was introduced by Anderson and Rubin (1949) in connection with the analysis of limited information maximum likelihood and generalized to the reduced rank regression model (1) by Anderson (1951). An excellent source of information is the monograph by Reinsel and Velu (1998), which contains a comprehensive survey of the theory and history of reduced rank regression and its many applications.

Note the difference between the unrestricted estimate $\hat{\Pi}_{OLS} = S_{yx.z} S_{xx.z}^{-1}$ and the reduced rank regression estimate $\hat{\Pi}_{RRR} = S_{yx.z}\hat{\beta}(\hat{\beta}' S_{xx.z}\hat{\beta})^{-1}\hat{\beta}'$ of the coefficient matrix to X.

Applications of the reduced rank model and algorithm

The reduced rank model (1) has many interpretations depending on the context. It is obviously a way of achieving fewer parameters in the possibly large $p \times q$ coefficient matrix Π. Another interpretation is that, although X is needed to explain the variation of Y, in practice only a few, r, factors are needed as given by the linear combinations $\beta' X$ in (1).

Restrictions on Π: Anderson (1951) formulated the problem of estimating Π under $p-r$ unknown restrictions $\mathcal{C}'\Pi = 0$. In (1) these are given by the matrix $\mathcal{C} = \alpha_\perp$, that is, a $p \times (p-r)$ matrix of full rank for which $\alpha'_\perp \alpha = 0$. The matrix $\alpha_\perp$ is estimated by solving the dual eigenvalue problem $|\lambda S_{yy.z} - S_{yx.z} S_{xx.z}^{-1} S_{xy.z}| = 0$, which has eigenvalues Λ and eigenvectors W, and the estimate is $\hat{\alpha}_\perp = (w_{r+1}, \ldots, w_p)$. If $p=q$, we can choose $W = S_{yy.z}^{-1} S_{yx.z} V \Lambda^{-1/2}$.

Canonical correlations: Reduced rank regression is related to canonical correlations (Hotelling, 1936). This is most easily expressed if $p=q$, where we find

$$\begin{pmatrix} W & 0 \\ 0 & V \end{pmatrix}' \begin{pmatrix} S_{yy.z} & S_{yx.z} \\ S_{xy.z} & S_{xx.z} \end{pmatrix} \begin{pmatrix} W & 0 \\ 0 & V \end{pmatrix} = \begin{pmatrix} I_p & \Lambda^{1/2} \\ \Lambda^{1/2} & I_q \end{pmatrix}.$$

This shows that the variables $W'Y$ and $V'X$ are the empirical canonical variates.

Instrumental variable estimation: Let the variables U, V, and X be of dimension p, q, and k respectively with $k \geq q$. Assume that they

are jointly Gaussian with mean zero and variance Σ, and that $E(U - \gamma' V)X' = 0$, so that X is an instrument for estimating γ. This means that $\Sigma_{ux} = \gamma' \Sigma_{vx}$, so that

$$E\left(\begin{pmatrix} U \\ V \end{pmatrix} | X\right) = \begin{pmatrix} \gamma' \Sigma_{vx} \\ \Sigma_{vx} \end{pmatrix} \Sigma_{xx}^{-1} X = \begin{pmatrix} \gamma' \\ I_q \end{pmatrix} \Sigma_{vx} \Sigma_{xx}^{-1} X = \alpha\beta'.$$

It follows that the $(p+q) \times k$ coefficient matrix in a regression of $Y = (U', V')'$ on X has rank q. Thus a reduced rank regression of Y on X is an algorithm for estimating the parameter of interest γ using the instruments X. This is the idea in Anderson and Rubin (1949) for the limited information maximum likelihood estimation.

Non-stationary time series: The model

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \Gamma \Delta Y_{t-1} + \varepsilon_t, \; t = 1, \ldots, T \tag{3}$$

determines a multivariate times series Y_t, and the reduced rank of $\alpha\beta'$ implies non-stationarity of the time series. Under suitable conditions (see Johansen, 1996) Y_t is non-stationary and ΔY_t and $\beta' Y_t$ are stationary. Thus Y_t is a cointegrated time series; see Engle and Granger (1987).

Common features: Engle and Kozicki (1993) used model (3) and assumed reduced rank of the matrix $(\alpha, \Gamma) = \xi\eta'$, so that $\Delta Y_t = \xi\eta'(Y_{t-1}'\beta, \Delta Y_{t-1}')' + \varepsilon_t$. In this case $\xi_\perp' \Delta Y_t = \xi_\perp' \varepsilon_t$ determines a random walk, where the common cyclic features have been eliminated.

Prediction: Box and Tiao (1977) analysed the model $Y_t = \Pi Y_{t-1} + \Gamma Y_{t-2} + \varepsilon_t$, and asked which linear combinations of the current values, $v' Y_t$, are best predicted by a linear combination of the past (Y_{t-1},Y_{t-2}), and hence introduced the analysis of canonical variates in the context of prediction of times series.

The Gaussian likelihood analysis

If the errors ε_t in model (1) are i.i.d. Gaussian $N_p(0, \Omega)$, and independent of $\{X_s, Z_s, s \le t\}$, the (conditional or partial) Gaussian likelihood is

$$-\frac{T}{2}\log|\Omega| - \frac{1}{2}\sum_{t=1}^{T} (Y_t - \alpha\beta' X_t - \Gamma Z_t)' \Omega^{-1} (Y_t - \alpha\beta' X_t - \Gamma Z_t).$$

Anderson (1951) introduced the *RRR* algorithm as a calculation of the maximum likelihood estimator of $\alpha\beta'$. The Frisch–Waugh theorem shows that one can partial out the parameter Γ by regression, as in the first step of the algorithm. We next regress (Y|Z) on $(\beta' X | Z)$ and find estimates of α and Ω, and the maximized likelihood function as functions of β:

$$\hat{\alpha}(\beta) = S_{yx.z}\beta(\beta' S_{xx.z}\beta)^{-1}, \tag{4}$$

$$\widehat{\Omega}(\beta) = S_{yy.z} - S_{yx.z}\beta(\beta' S_{xx.z}\beta)^{-1}\beta' S_{xy.z},$$

$$L_{\max}^{-2/T}(\beta) = |\widehat{\Omega}(\beta)|.$$

The identity

$$\left|\begin{pmatrix} S_{yy.z} & S_{yx.z}\beta \\ \beta' S_{xy.z} & \beta' S_{xx.z}\beta \end{pmatrix}\right| = |S_{yy.z}||\beta' S_{xx.z}\beta - \beta' S_{xy.z}S_{yy.z}^{-1}S_{yx.z}\beta| = |\beta' S_{xx.z}\beta||S_{yy.z} - S_{yx.z}\beta(\beta' S_{xx.z}\beta)^{-1}\beta' S_{xy.z}| = |\beta' S_{xx.z}\beta||\widehat{\Omega}(\beta)|$$

shows that $L_{\max}^{-2/T}(\beta) = |S_{yy.z}||\beta'(S_{xx.z} - S_{xy.z}S_{yy.z}^{-1}S_{yx.z})\beta| / |\beta' S_{xx.z}\beta|$ so that β has to be chosen to minimize this. Differentiating with respect to β we find that β has to satisfy the relation $S_{xx.z}\beta = S_{xy.z}S_{yy.z}^{-1}S_{yx.z}\beta\xi$ for some $r \times r$ matrix ξ. This shows (see Johansen, 1996) that the space spanned by the columns of β, is spanned by *r* of the eigenvectors of (2), and hence, that choosing the largest λ_i gives the smallest value of $L_{\max}^{-1/T}(\beta)$, so that

$$\hat{\beta} = (v_1, \ldots, v_r),\ L_{\max}^{-2/T}(\hat{\beta}) = |S_{yy.z}|\prod_{i=1}^{r}(1 - \lambda_i).$$

Hypothesis testing: The test statistic for rank of Π can be calculated from the eigenvalues because the eigenvalue problem solves the maximization of the likelihood for all values of *r* simultaneously. The likelihood ratio test statistic for the hypothesis $\mathrm{rank}(\Pi) \le r$, as derived by Anderson (1951), is

$$-2\log LR(\mathrm{rank}(\Pi) \le r) = T\sum_{i=r+1}^{\min(p,q)} \log(1 - \lambda_i). \tag{5}$$

Bartlett (1938) suggested using this statistic to test that *r* canonical correlations between *Y* and *X* were zero and hence that Π had reduced rank.

The simplest hypothesis to test on β is $\beta = H\phi$. We can estimate β under this restriction by $RRR(Y, H'X|Z)$ and therefore calculate the likelihood ratio statistic using reduced rank regression. If, on the other hand, we have restrictions on the individual vectors

$\beta = (H_1\varphi_1, \ldots, H_r\varphi_r)$, then reduced rank regression does not provide a maximum likelihood estimator, but we can switch between reduced rank regressions as follows. For fixed $\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_r$, we can find an estimator for ϕ_i and hence $\beta_i = H_i\varphi_i$ by

$$RRR(Y, H_i' X|Z, (\beta_1, \ldots \beta_{i-1}, \beta_{i+1}, \ldots, \beta_r)' X).$$

By switching between the vectors in β, we have an algorithm which is useful in practice and which maximizes the likelihood in each step.

A switching algorithm: Another algorithm, that is useful for this model, is to consider the first order condition for β, when Γ has been eliminated, which has solution

$$\hat{\beta}(\alpha, \Omega) = S_{xx.z}^{-1} S_{xy.z} \Omega^{-1} \alpha \left(\alpha' \Omega^{-1} \alpha\right)^{-1}.$$

Combining this with (4) suggests a switching algorithm, as follows.

First choose some initial estimator $\hat{\beta}_0$, then switch between estimating α and Ω for fixed β by least squares, and estimating β for fixed α and Ω by generalized least squares.

This switching algorithm maximizes the likelihood function in each step and any limit point will be a stationary point. It seems to work well in practice. There are natural hypotheses one can test in the reduced rank model, like general linear restrictions on β, which are not solved by the reduced rank regression algorithm, whereas the above algorithm can be modified to give a solution.

Asymptotic distributions in the stationary case: The asymptotic distributions of the estimators and test statistics can be described under the assumption that the process (Y_t, X_t, Z_t) is stationary with finite second moments. It can be shown that estimators are asymptotically Gaussian and test statistics for hypotheses both for rank and for β are asymptotically χ^2; see Robinson (1973).

Asymptotic distributions in the non-stationary case: If the processes are non-stationary a different type of asymptotics is needed. As an example consider model (3) for $I(1)$ variables. When discussing the asymptotic distribution of the estimators, the normalization by $\hat{\beta}' S_{xx.z} \hat{\beta} = I_r$ is not convenient, and it is necessary to identify the vectors differently.

One can then prove (see Johansen, 1996), that the estimates of α, Γ and Ω are asymptotically Gaussian and have the same limit distribution as if β were known: that is, the asymptotic distribution they have in the regression of ΔY_t on the stationary variables $\beta' Y_{t-1}$ and ΔY_{t-1}.

The asymptotic distribution of $\hat{\beta}$ is mixed Gaussian, where the mixing parameter is the (random) limit of the observed information. Therefore, by normalizing on the observed information, we obtain asymptotic χ^2 inference for hypotheses on β.

The limit distribution of the likelihood ratio test statistic for rank, see (5), is given by a generalization of the Dickey–Fuller distribution:

$$DF_{p-r} = \mathrm{tr}\left\{\int_0^1 (\mathrm{d}W) W' \left(\int_0^1 WW' \mathrm{d}u\right)^{-1} \int_0^1 W(\mathrm{d}W)'\right\},$$

where W is a standard Brownian motion in p–r dimensions. The quantiles of this distribution can at present only be calculated by simulation if $p - r > 1$. The limit distribution has to be modified if deterministic terms are included in the model.

See Also

- instrumental variables

- maximum likelihood

Bibliography

Anderson, T.W. 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* 22, 327–51.

Anderson, T.W. and Rubin, H. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20, 46–63.

Bartlett, M.S. 1938. Further aspects of the theory of multiple regression. *Proceedings of the Cambridge Philosophical Society* 34, 33–40.

Box, G.E.P. and Tiao, G.C. 1977. A canonical analysis of multiple time series. *Biometrika* 64, 355–65.

Doornik, J.A. and O'Brien, R.J. 2002. Numerically stable cointegration analysis. *Computational Statistics and Data Analysis* 41, 185–93.

Engle, R.F. and Granger, C.W.J. 1987. Co-integration and error correction: representation, estimation and testing. *Econometrica* 55, 251–76.

Engle, R.F. and Kozicki, S. 1993. Testing for common factors (with comments). *Journal of Business Economics and Statistics* 11, 369–78.

Hotelling, H. 1936. Relations between two sets of variables. *Biometrika* 28, 321–77.

Johansen, S. 1996. *Likelihood Based Inference on Cointegration in the Vector Autoregressive Model*. Oxford: Oxford University Press.

Reinsel, G.C. and Velu, R.P. 1998. *Multivariate Reduced Rank Regression*. Lecture Notes in Statistics. New York: Springer.

Robinson, P.M. 1973. Generalized canonical analysis for time series. *Journal of Multivariate Analysis* 3, 141–60.

How to cite this article

Johansen, Søren. "reduced rank regression." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000265> doi:10.1057/9780230226203.1410

The New Palgrave Dictionary of Economics Online

regime switching models

James D. Hamilton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

If the parameters of a time-series process are subject to change over time, then a full description of the data-generating process must include a specification of the probability law governing these changes, for example, postulating that the parameters evolve according to the realization of an unobserved Markov chain. This article describes classical and Bayesian algorithms for estimation and inference in such models and discusses some of the issues that arise in particular cases such as GARCH and state-space models.

Keywords

ARMA models; asset prices; econometrics; GARCH models; Gaussian densities; Gibbs sampler; Kalman filter; Markov chain Monte Carlo methods; Markov processes; maximum likelihood; numerical optimization methods in economics; regime-switching models; state-space models; vector autoregressions

Article

Many economic time series occasionally exhibit dramatic breaks in their behaviour, associated with events such as financial crises (Jeanne and Masson, 2000; Cerra and Saxena, 2005; Hamilton, 2005) or abrupt changes in government policy (Hamilton, 1988; Sims and Zha, 2006; Davig, 2004). Of particular interest to economists is the apparent tendency of many economic variables to behave quite differently during economic downturns, when underutilization of factors of production rather than their long-run tendency to grow governs economic dynamics (Hamilton, 1989; Chauvet and Hamilton, 2006). Abrupt changes are also a prevalent feature of financial data, and the approach described below is quite amenable to theoretical calculations for how such abrupt changes in fundamentals should show up in asset prices (Ang and Bekaert, 2002a; 2000b; Garcia, Luger and Renault, 2003; Dai, Singleton and Yang, 2003).

Consider how we might describe the consequences of a dramatic change in the behaviour of a single variable y_t. Suppose that the typical historical behaviour could be described with a first-order autoregression,

$$y_t = c_1 + \varphi y_{t-1} + \varepsilon_t, \tag{1}$$

with $\varepsilon_t \sim N(0, \sigma^2)$, which seemed to adequately describe the observed data for $t = 1, 2, \ldots, t_0$. Suppose that at date t_0 there was a significant change in the average level of the series, so that we would instead wish to describe the data according to

$$y_t = c_2 + \varphi y_{t-1} + \varepsilon_t \tag{2}$$

for $t = t_0 + 1, t_0 + 2, \ldots$ This fix of changing the value of the intercept from c_1 to c_2 might help the model to get back on track with better forecasts, but it is rather unsatisfactory as a probability law that could have generated the data. We surely would not want to maintain that the change from c_1 to c_2 at date t_0 was a deterministic event that anyone would have been able to predict with certainty looking ahead from date $t = 1$. Instead, there must have been some imperfectly predictable forces that produced the change. Hence, rather than claim that expression (1) governed the data up to date t_0 and (2) after that date, what we must have in mind is that there is some larger model encompassing them both,

$$y_t = c_{s_t} + \varphi y_{t-1} + \varepsilon_t, \tag{3}$$

where s_t is a random variable that, as a result of institutional changes, happened in our sample to assume the value $s_t = 1$ for $t = 1, 2, \ldots, t_0$ and $s_t = 2$ for $t = t_0 + 1, t_0 + 2, \ldots$. A complete description of the probability law governing the observed data would then require a probabilistic model of what caused the change from $s_t = 1$ to $s_t = 2$. The simplest such specification is that s_t is the realization of a two-state Markov chain with

$$\Pr(s_t = j | s_{t-1} = i, s_{t-2} = k, \ldots, y_{t-1}, y_{t-2}, \ldots) = \Pr(s_t = j | s_{t-1} = i) = p_{ij}.$$

(4)

On the assumption that we do not observe s_t directly, but only infer its operation through the observed behavior of y_t, the parameters necessary to fully describe the probability law governing y_t are then the variance of the Gaussian innovation σ^2, the autoregressive coefficient ϕ, the two intercepts c_1 and c_2, and the two state transition probabilities, p_{11} and p_{22}.

The specification in (4) assumes that the probability of a change in regime depends on the past only through the value of the most recent regime, though, as noted below, nothing in the approach described below precludes looking at more general probabilistic specifications. But the simple time-invariant Markov chain (4) seems the natural starting point and is clearly preferable to acting as if the shift from c_1 to c_2 was a deterministic event. Permanence of the shift would be represented by $p_{22} = 1$, though the Markov formulation invites the more general possibility that $p_{22} < 1$. Certainly in the case of business cycles or financial crises, we know that the situation, though dramatic, is not permanent. Furthermore, if the regime change reflects a fundamental change in monetary or fiscal policy, the prudent assumption would seem to be to allow the possibility of it changing back again, suggesting that $p_{22} < 1$ is often a more natural formulation for thinking about changes in regime than $p_{22} = 1$.

A model of the form of (3)–(4) with no autoregressive elements ($\omega = 0$) appears to have been first analysed by Lindgren (1978) and Baum et al. (1980). Specifications that incorporate autoregressive elements date back in the speech recognition literature to Poritz (1982), Juang and Rabiner (1985), and Rabiner (1989), who described such processes as ‘hidden Markov models’. Markov-switching regressions were introduced in econometrics by Goldfeld and Quandt (1973), the likelihood function for which was first correctly calculated by Cosslett and Lee (1985). The formulation of the problem described here, in which all objects of interest are calculated as a by-product of an iterative algorithm similar in spirit to a Kalman filter, is due to Hamilton (1989; 1994). General characterizations of moment and stationarity conditions for such processes can be found in Tjøstheim (1986), Yang (2000), Timmermann (2000), and Francq and Zakoïan (2001).

Econometric inference

Suppose that the econometrician observes y_t directly but can only make an inference about the value of s_t based on what we see happening with y_t. This inference will take the form of two probabilities

$$\xi_{jt} = \Pr(s_t = j|\Omega_t; \boldsymbol{\theta})$$

(5)

for $j = 1, 2$, where these two probabilities sum to unity by construction. Here $\Omega_t = \{y_t, y_{t-1}, \ldots, y_1, y_0\}$ denotes the set of observations obtained as of date *t*, and θ is a vector of population parameters, which for the above example would be $\boldsymbol{\theta} = (\sigma, \varphi, c_1, c_2, p_{11}, p_{22})'$, and which for now we presume to be known with certainty. The inference is performed iteratively for $t = 1, 2, \ldots, T$, with step *t* accepting as input the values

$$\xi_{i,t-1} = \Pr(s_{t-1} = i|\Omega_{t-1}; \boldsymbol{\theta})$$
(6)

for $i = 1, 2$ and producing as output (5). The key magnitudes one needs in order to perform this iteration are the densities under the two regimes,

$$\eta_{jt} = f(y_t|s_t = j, \Omega_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y_t - c_j - \varphi y_{t-1})^2}{2\sigma^2}\right],$$
(7)

for $j = 1, 2$. Specifically, given the input (6) we can calculate the conditional density of the *t*th observation from

$$f(y_t|\Omega_{t-1}; \boldsymbol{\theta}) = \sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij}\xi_{i,t-1}\eta_{jt}$$
(8)

and the desired output is then

$$\xi_{jt} = \frac{\sum_{i=1}^{2} p_{ij}\xi_{i,t-1}\eta_{jt}}{f(y_t|\Omega_{t-1}; \boldsymbol{\theta})}.$$
(9)

As a result of executing this iteration, we will have succeeded in evaluating the sample conditional log likelihood of the observed data

$$\log f(y_1, y_2, \ldots, y_T | y_0; \boldsymbol{\theta}) = \sum_{t=1}^{T} \log f(y_t | \Omega_{t-1}; \boldsymbol{\theta}) \tag{10}$$

for the specified value of θ. An estimate of the value of θ can then be obtained by maximizing (10) by numerical optimization.

Several options are available for the value ξ_{i0} to use to start these iterations. If the Markov chain is presumed to be ergodic, one can use the unconditional probabilities

$$\xi_{i0} = \Pr(s_0 = i) = \frac{1 - p_{jj}}{2 - p_{ii} - p_{jj}}.$$

Other alternatives are simply to set $\xi_{i0} = 1/2$ or estimate ξ_{i0} itself by maximum likelihood.

The calculations do not increase in complexity if we consider an $(r \times 1)$ vector of observations $\mathbf{y}_t$ whose density depends on N separate regimes. Let $\Omega_t = \{\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1\}$ be the observations through date t, $\mathbf{P}$ be an $(N \times N)$ matrix whose row j, column i element is the transition probability p_{ij}, η_t be an $(N \times 1)$ vector whose jth element $f(\mathbf{y}_t | s_t = j, \Omega_{t-1}; \boldsymbol{\theta})$ is the density in regime j, and $\hat{\boldsymbol{\xi}}_{t|t}$ an $(N \times 1)$ vector whose jth element is $\Pr(s_t = j | \Omega_t, \boldsymbol{\theta})$ Then (8) and (9) generalize to

$$f(\mathbf{y}_t | \Omega_{t-1}; \boldsymbol{\theta}) = \mathbf{1}'(\mathbf{P}\hat{\boldsymbol{\xi}}_{t-1|t-1} \odot \boldsymbol{\eta}_t) \tag{11}$$

$$\hat{\boldsymbol{\xi}}_{t|t} = \frac{\mathbf{P}\hat{\boldsymbol{\xi}}_{t-1|t-1} \odot \boldsymbol{\eta}_t}{f(\mathbf{y}_t | \Omega_{t-1}; \boldsymbol{\theta})} \tag{12}$$

where **1** denotes an $(N \times 1)$ vector all of whose elements are unity and $\odot$ denotes element-by-element multiplication. Markov-switching vector autoregressions are discussed in detail in Krolzig (1997). Vector applications include describing the co-movements between stock prices and economic output (Hamilton and Lin, 1996) and the tendency for some series to move into recession before others (Hamilton and Perez-Quiros, 1996). There further is no requirement that the elements of η_t be Gaussian densities or even from the same family of densities. For example, Dueker (1997) studied a model in which the degrees of freedom of a Student t distribution change depending on the economic regime. One is also often interested in forming an inference about what regime the economy was in at date *t* based on observations obtained through a later date *T*, denoted $\hat{\xi}_{t|T}$. These are referred to as 'smoothed' probabilities, an efficient algorithm for whose calculation was developed by Kim (1994).

Extensions

The calculations in (11) and (12) remain valid when the probabilities in **P** depend on lagged values of $\mathbf{y}_t$ or strictly exogenous explanatory variables, as in Diebold, Lee and Weinbach (1994), Filardo (1994) and Peria (2002). However, often there are relatively few transitions among regimes, making it difficult to estimate such parameters accurately, and most applications have assumed a time-invariant Markov chain. For the same reason, most applications assume only $N = 2$ or 3 different regimes, though there is considerable promise in models with a much larger number of regimes, either by tightly parameterizing the relation between the regimes (Calvet and Fisher, 2004), or with prior Bayesian information (Sims and Zha, 2006).

In the Bayesian approach, both the parameters θ and the values of the states $\mathbf{s} = (s_1, s_2, \ldots, s_T)'$ are viewed as random variables. Bayesian inference turns out to be greatly facilitated by Monte Carlo Markov chain methods, specifically, the Gibbs sampler. This is achieved by sequentially (for $k = 1, 2, \ldots$) generating a realization $\theta^{(k)}$ from the distribution of $\boldsymbol{\theta}|\mathbf{s}^{(k-1)}, \Omega_T$ followed by a realization of $\mathbf{s}^{(k)}$ from the distribution of $\mathbf{s}|\boldsymbol{\theta}^{(k)}, \Omega_T$. The first distribution, $\boldsymbol{\theta}|\mathbf{s}^{(k-1)}, \Omega_T$, treats the historical regimes generated at the previous iteration, $s_1^{(k-1)}, s_2^{(k-1)}, \ldots, s_T^{(k-1)}$, as if fixed known numbers. Often this conditional distribution takes the form of a standard Bayesian inference problem whose solution is known analytically using natural conjugate priors. For example, the posterior distribution of ϕ given other parameters is a known function of easily calculated OLS coefficients. An algorithm for generating a draw from the second distribution, $\mathbf{s}|\boldsymbol{\theta}^{(k)}, \Omega_T$, was developed by Albert and Chib (1993). The Gibbs sampler turns out also to be a natural device for handling transition probabilities that are functions of observable variables, as in Filardo and Gordon (1998).

It is natural to want to test the null hypothesis that there are *N* regimes against the alternative of $N + 1$, for example when $N = 1$, to test whether there are any changes in regime at all. Unfortunately, the likelihood ratio test of this hypothesis fails to satisfy the usual regularity conditions because, under the

null hypothesis, some of the parameters of the model would be unidentified. For example, if there is really only one regime, the maximum likelihood estimate $\hat{p}_{11}$ does not converge to a well-defined population magnitude, meaning that the likelihood ratio test does not have the usual χ^2 limiting distribution. To interpret a likelihood ratio statistic, one instead needs to appeal to the methods of Hansen (1992) or Garcia (1998). An alternative is to rely on generic tests of the hypothesis that an *N*-regime model accurately describes the data (Hamilton, 1996), though these tests are not designed for optimal power against the specific alternative hypothesis of $N + 1$ regimes. A test recently proposed by Carrasco, Hu and Ploberger (2004) that is easy to compute but not based on the likelihood ratio statistic seems particularly promising. Other alternatives are to use Bayesian methods to calculate the value of *N* implying the largest value for the marginal likelihood (Chib, 1998) or the highest Bayes factor (Koop and Potter, 1999), or to compare models on the basis of their ability to forecast (Hamilton and Susmel, 1994).

A specification where the density depends on a finite number of previous regimes, $f(\mathbf{y}_t|s_t, s_{t-1}, \ldots, s_{t-m}, \Omega_{t-1}; \boldsymbol{\theta})$, can be recast in the above form by a suitable redefinition of regime. For example, if s_t follows a 2-state Markov chain with transition probabilities $\Pr(s_t = j|s_{t-1} = i)$ and $m = 1$, one can define a new regime variable s_t^* such that $f(\mathbf{y}_t|s_t^*, \Omega_{t-1}; \boldsymbol{\theta}) = f(\mathbf{y}_t|s_t, s_{t-1}, \ldots, s_{t-m}, \Omega_{t-1}; \boldsymbol{\theta})$ as follows:

$$s_t^* = \begin{cases} 1 & \text{when } s_t = 1 \text{ and } s_{t-1} = 1 \\ 2 & \text{when } s_t = 2 \text{ and } s_{t-1} = 1 \\ 3 & \text{when } s_t = 1 \text{ and } s_{t-1} = 2 \\ 4 & \text{when } s_t = 2 \text{ and } s_{t-1} = 2 \end{cases}.$$

Then s_t^* itself follows a 4-state Markov chain with transition matrix

$$\mathbf{P}^* = \begin{bmatrix} p_{11} & 0 & p_{11} & 0 \\ p_{12} & 0 & p_{12} & 0 \\ 0 & p_{21} & 0 & p_{21} \\ 0 & p_{22} & 0 & p_{22} \end{bmatrix}.$$

More problematic are cases in which the order of dependence *m* grows with the date of the observation *t*. Such a situation often arises in models whose recursive structure causes the density of y_t given Ω_{t-1} to

depend on the entire history $y_{t-1}, y_{t-2}, \ldots, y_1$ as is the case in ARMA, GARCH or state-space models. Consider for illustration a GARCH(1,1) specification in which the coefficients are subject to changes in regime, $y_t = h_t v_t$, where $v_t \sim N(0, 1)$ and

$$h_t^2 = \gamma_{s_t} + \alpha_{s_t} y_{t-1}^2 + \beta_{s_t} h_{t-1}^2. \tag{13}$$

Solving (13) recursively reveals that the conditional standard deviation h_t depends on the full history $\{y_{t-1}, y_{t-2}, \ldots, y_0, s_t, s_{t-1}, \ldots, s_1\}$. One way to avoid this problem was proposed by Gray (1996), who postulated that, instead of being generated by (13), the conditional variance is characterized by

$$h_t^2 = \gamma_{s_t} + \alpha_{s_t} y_{t-1}^2 + \beta_{s_t} \tilde{h}_{t-1}^2 \tag{14}$$

where

$$\tilde{h}_{t-1}^2 = \sum_{i=1}^{N} \hat{\xi}_{i,t-1|t-2} (\gamma_i + \alpha_i y_{t-2}^2 + \beta_i \tilde{h}_{t-2}^2).$$

In Gray's model, h_t in (14) depends only on s_t since $\tilde{h}_{t-1}^2$ is a function of data Ω_{t-1} only. An alternative solution, due to Haas, Mittnik and Paolella (2004), is to hypothesize N separate GARCH processes whose values h_{it} all exist as latent variables at date t,

$$h_{it}^2 = \gamma_i + \alpha_i y_{t-1}^2 + \beta_i h_{i,t-1}^2 \tag{15}$$

and then simply pose the model as $y_t = h_{s_t} v_t$. Again, the feature that makes this work is the fact that h_{it} in (15) is a function solely of the data Ω_{t-1} rather than the states $\{s_{t-1}, s_{t-2}, \ldots, s_1\}$.

A related problem arises in Markov-switching state-space models, which posit an unobserved state vector $\mathbf{z}_t$ characterized by

$$\mathbf{z}_t = \mathbf{F}_{s_t}\mathbf{z}_{t-1} + \mathbf{Q}_{s_t}\mathbf{v}_t$$

with $\mathbf{v}_t \sim N(0, \mathbf{I}_n)$, with observed vectors $\mathbf{y}_t$ and $\mathbf{x}_t$ governed by

$$\mathbf{y}_t = \mathbf{H}'_{s_t}\mathbf{z}_t + \mathbf{A}'_{s_t}\mathbf{x}_t + \mathbf{R}_{s_t}\mathbf{w}_t$$

for $\mathbf{w}_t \sim N(0, \mathbf{I}_r)$. Again, the model as formulated implies that the density of $\mathbf{y}_t$ depends on the full history $\{s_t, s_{t-1}, \ldots, s_1\}$. Kim (1994) proposed a modification of the Kalman filter equations similar in spirit to the modification in (14) that can be used to approximate the log likelihood. A more common practice recently has been to estimate such models with numerical Bayesian methods, as in Kim and Nelson (1999).

See Also

- Markov chain Monte Carlo methods
- Markov processes
- maximum likelihood
- mixture models
- nonlinear time series analysis
- numerical optimization methods in economics
- structural change

Bibliography

Albert, J. and Chib, 1993. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. J*ournal of Business and Economic Statistics* 11, 1–15.

Ang, A. and Bekaert, G. 2002a. International asset allocation with regime shifts. *Review of Financial Studies* 15, 1137–87.

Ang, A. and Bekaert, G. 2002b. Regime switches in interest rates. *Journal of Business and Economic*

Statistics 20, 163–82.

Baum, L., Petrie, E., Soules, G. and Weiss, N. 1980. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* 41, 164–71.

Calvet, L. and Fisher, A. 2004. How to forecast long-run volatility: regime-switching and the estimation of multifractal processes. *Journal of Financial Econometrics* 2, 49–83.

Carrasco, M., Hu, L. and Ploberger, W. 2004. Optimal test for Markov switching. Working paper. University of Rochester.

Cerra, V. and Saxena, S. 2005. Did output recover from the Asian crisis? *IMF Staff Papers* 52, 1–23.

Chauvet, M. and Hamilton, J. 2006. Dating business cycle turning points. In *Nonlinear Time Series Analysis of Business Cycles*, ed. C. Milas, P. Rothman and D. van Dijk. Amsterdam: Elsevier.

Chib, S. 1998. Estimation and comparison of multiple change-point models. *Journal of Econometrics* 86, 221–41.

Cosslett, S. and Lee, L.-F. 1985. Serial correlation in discrete variable models. *Journal of Econometrics* 27, 79–97.

Dai, Q., Singleton, K. and Yang, W. 2003. Regime shifts in a dynamic term structure model of U.S. Treasury bonds. Working paper, Stanford University.

Davig, T. 2004. Regime-switching debt and taxation. *Journal of Monetary Economics* 51, 837–59.

Diebold, F., Lee, J.-H. and Weinbach, G. 1994. Regime switching with time-varying transition probabilities. In *Nonstationary Time Series Analysis and Cointegration*, ed. C. Hargreaves. Oxford: Oxford University Press.

Dueker, M. 1997. Markov switching in GARCH processes and mean-reverting stock-market volatility. *Journal of Business and Economic Statistics* 15, 26–34.

Filardo, A. 1994. Business cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics* 12, 299–308.

Filardo, A. and Gordon, S. 1998. Business cycle durations. *Journal of Econometrics* 85, 99–123.

Francq, C. and Zakoïan, J.-M. 2001. Stationarity of multivariate Markov-switching ARMA models.

Journal of Econometrics 102, 339–64.

Garcia, R. 1998. Asymptotic null distribution of the likelihood ratio test in Markov switching models. *International Economic Review* 39, 763–88.

Garcia, R., Luger, R. and Renault, E. 2003. Empirical assessment of an intertemporal option pricing model with latent variables. *Journal of Econometrics* 116, 49–83.

Goldfeld, S. and Quandt, R. 1973. A Markov model for switching regressions. *Journal of Econometrics* 1, 3–16.

Gray, S. 1996. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* 42, 27–62.

Haas, M., Mittnik, S. and Paolella, M. 2004. A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics* 2, 493–530.

Hamilton, J. 1988. Rational-expectations econometric analysis of changes in regime: an investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control* 12, 385–423.

Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–84.

Hamilton, J. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hamilton, J. 1996. Specification testing in Markov-switching time-series models. *Journal of Econometrics* 70, 127–57.

Hamilton, J. 2005. What's real about the business cycle? *Federal Reserve Bank of St. Louis Review* 87, 435–52.

Hamilton, J. and Lin, G. 1996. Stock market volatility and the business cycle. *Journal of Applied Econometrics* 11, 573–93.

Hamilton, J. and Perez-Quiros, G. 1996. What do the leading indicators lead? *Journal of Business* 69, 27–49.

Hamilton, J. and Susmel, R. 1994. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics* 64, 307–33.

Hansen, B. 1992. The likelihood ratio test under non-standard conditions. *Journal of Applied*

Econometrics 7, S61–S82. Erratum, 11(1996), 195–8.

Jeanne, O. and Masson, P. 2000. Currency crises, sunspots, and Markov-switching regimes. *Journal of International Economics* 50, 327–50.

Juang, B.-H. and Rabiner, L. 1985. Mixture autoregressive hidden Markov models for speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 30, 1404–13.

Kim, C. 1994. Dynamic linear models with Markov-switching. *Journal of Econometrics* 60, 1–22.

Kim, C. and Nelson, C. 1999. *State-Space Models with Regime Switching*. Cambridge, MA: MIT Press.

Koop, G. and Potter, S. 1999. Bayes factors and nonlinearity: evidence from economic time series. *Journal of Econometrics* 88, 251–81.

Krolzig, H.-M. 1997. *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Berlin: Springer.

Lindgren, G. 1978. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics* 5, 81–91.

Peria, M. 2002. A regime-switching approach to the study of speculative attacks: a focus on EMS crises. In *Advances in Markov-Switching Models*, ed. J. Hamilton and B. Raj. Heidelberg: Physica Verlag.

Poritz, A. 1982. Linear predictive hidden Markov models and the speech signal. *Acoustics, Speech and Signal Processing, IEEE Conference on ICASSP '82*, vol. 7, 1291–4.

Sims, C. and Zha, T. 2006. Were there switches in U.S. monetary policy? *American Economic Review* 96, 54–81.

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77, 257–86.

Timmermann, A. 2000. Moments of Markov switching models. *Journal of Econometrics* 96, 75–111.

Tjøstheim, D. 1986. Some doubly stochastic time series models. *Journal of Time Series Analysis* 7, 51–72.

Yang, M. 2000. Some properties of vector autoregressive processes with Markov-switching coefficients. *Econometric Theory* 16, 23–43.

How to cite this article

Hamilton, James D. "regime switching models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000269> doi:10.1057/9780230226203.1411

The New Palgrave Dictionary of Economics Online

regional and preferential trade agreements

Pravin Krishna

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses analytical developments in the literature on the economics and politics of preferential trade agreements (PTAs). It describes results obtained in the traditional theory that demonstrate the ambiguous welfare outcomes of preferential trade liberalization. Theoretical approaches to designing necessarily welfare-improving PTAs are also discussed. Finally, this article sets out recent analyses in the literature concerning the dynamic expansion of trade blocs, the endogenous determination of policy (relating to preferences within a PTA and to extra-union trade), and the effects of preferential agreements on the multilateral trade system.

Keywords

ASEAN Free Trade Area (AFTA); bargaining; coalitions; comparative advantage; customs unions; European Economic Community (EEC); European Free Trade Association (EFTA); free trade areas; General Agreement on Tariffs and Trade (GATT); Great Depression; income distribution; international trade; Meade, J. E.; Mercosur; most-favoured-nation (MFN) principle; multilateral free trade; Mundell, R.; natural trading partners hypothesis; North American Free Trade Agreement (NAFTA); preferential trade agreements; redistribution of income; regional and preferential trade agreements; sunk costs; tariffs; terms of trade; trade creation; trade diversion; Viner, J.

Article

Strongly influenced by the perception that restricted commerce and preferences in trade relations had contributed to the Great Depression of the 1930s and the subsequent outbreak of war, the discussions leading to the General Agreement on Tariffs and Trade (GATT) in 1947 were driven by the desire to create an international economic order based on a liberal and non-discriminatory multilateral trade system. Enshrined in Article I of the GATT, the principle of non-discrimination (commonly referred to as the most-favoured-nation or MFN clause) precludes member countries from discriminating against

imports based upon the country of origin. However, in an important exception this central prescript, the GATT, through its Article XXIV, permits its members to enter into preferential trade agreements (PTAs), provided these preferences are complete. In so doing, it sanctions the formation of free trade areas (FTAs), whose members are obligated to eliminate internal import barriers, and customs unions (CUs), whose members additionally agree on a common external tariff against imports from non-members. Additional derogations to the principle of non-discrimination now include the Enabling Clause, which allows tariff preferences to be granted to developing countries (in accordance with the Generalized System of Preferences) and permits preferential trade agreements among developing countries in goods trade. Among the more prominent existing PTAs are the North American Free Trade Agreement (NAFTA), the European Economic Community (EEC) and the European Free Trade Association (EFTA), all formed under Article XXIV, and the Mercosur (the CU between the Argentine Republic, Brazil, Paraguay, and Uruguay) and the ASEAN (Association of South East Asian Nations) Free Trade Area (AFTA), both formed under the Enabling Clause.

Static welfare analysis

Motivated by ongoing discussions concerning optimal trade arrangements in the post-war period, especially over the possibility of a European customs union, Jacob Viner (1950, pp. 41–50) developed a seminal analysis of the economics of preferential trade. Viner's analysis undermines the presumption that cutting tariffs is necessarily welfare improving. On the one hand, because of discriminatory liberalization, there will be commodities that a member country may ‘newly import from the other but which it formerly did not import at all because the price of the protected domestic good was lower than the price of any foreign source plus the duty’. Viner calls this shift from a high-cost to a lower-cost point ‘trade creation’ and associates it with welfare-improvement for the importing country. He also argues that, on the other hand, ‘there may be other commodities, which one of the members will now newly import from the other’, whereas before the PTA it ‘imported them from a third country, because that was the cheapest possible source of supply even after the payment of duty’. He calls this shift in imports from a low-cost third country to a higher-cost member country ‘trade diversion,’ associating it with an increase in the cost of imports and, thus, welfare losses for the importing country.

The demonstration that preferential trade liberalization may be welfare decreasing stimulated a substantial theoretical literature on the ‘static’ welfare effects of PTAs. Post-Vinerian analysis of the welfare effects of preferential trade include Meade's (1955) more explicit and comprehensive formulation of the problem in a three-country three-good setting. Meade argues that not only the magnitudes of trade creation and trade diversion but also the extent of cost reductions (in the former) and increases (in the latter) were necessary to arrive at a welfare evaluation. Subsequent analysis also developed examples of both welfare improving trade diversion and welfare-decreasing trade creation in general equilibrium contexts broader than those considered by Viner (see, for instance, Gehrels, 1956–7, Lipsey, 1957, and Bhagwati, 1971). However, the intuitive appeal of the concepts of trade creation and trade diversion has ensured their continued use in the economic analysis of preferential trade agreements, especially in policy analysis (see Panagariya, 2000, for a comprehensive survey).

The effects of preferences on intra-union and extra-union terms-of-trade are analysed by Mundell (1964), who argues that a country granting tariff preferences moves intra-union terms of trade against it

and in favour of its partner by increasing its demand for imports from its partner. Extra-union terms of trade are improved for the partner, but change ambiguously for the preference-granting country. Thus, tariff preferences have asymmetric effects on the preference-granting and preference-receiving country. More sharply, Panagariya (1997a) shows how, even with fixed extra-union terms of trade, if a preference-granting country continues to import from the rest of the world, intra-union terms-of-trade losses (manifesting themselves as intra-union tariff revenue transfers as also seen in Berglas, 1979, and Riezman, 1979) unambiguously worsen its welfare, while its preference-receiving partner unambiguously gains.

Wonnacott and Lutz (1987), Krugman (1991) and Summers (1991) propose geographic proximity between partner countries as important in ensuring that preferential liberalization improves welfare. Specifically, they suggest that countries entering into preferential arrangements with geographically proximate countries are likely to do better than in agreements with distant countries, because the former are more likely than the latter to be trade creating, leading to a larger improvement in welfare. Proximate countries are thus argued to be ‘natural’ partners for preferential trade. Bhagwati and Panagariya (1996) and Panagariya (1997b), however, provide a number of examples in which, between two otherwise identical potential partners, a country achieves a superior outcome by granting trade preferences to the distant partner. For instance, it may be that a preference granted to a distant partner leads to a smaller transfer (loss) of tariff revenue with a closer country, since, with an initial non-discriminatory tariff, the liberalizing country imports less from the more distant partner. Thus, the ‘natural trading partners' hypothesis is shown to lack general theoretical validity.

Numerous studies have attempted to evaluate quantitatively the trade creation and trade diversion effects of PTAs. Recently, focusing on the effects of PTAs on excluded countries, Chang and Winters (2002) show how Mercosur was associated with significant declines in the prices of non-members' exports to the region. Yeats (1998) shows how under Mercosur the greatest increases of intra-union trade flows were in goods in which the member countries had the least comparative advantage, confirming the trade diversionary effects of preferential liberalization.

Srinivasan, Whalley and Wooton (1993) note that the econometric frameworks used in most *ex post* studies of trade flows generally lack microeconomic underpinnings, making an evaluation of the associated welfare consequences difficult. Krishna (2003) develops an econometric framework for the analysis of PTAs with a strong welfare-theoretic foundation, so that the estimated parameters relating to trade creation and trade diversion effects fit directly into theoretically derived welfare expressions. His application of this framework to the evaluation of the natural trading partners hypothesis does not find any support in US data.

Necessarily welfare-improving preferential trade areas

The generally ambiguous welfare results provided by the theoretical literature raised an important question relating to the *design* of necessarily welfare-improving PTAs. A classic result stated independently by Kemp (1964) and Vanek (1965) and proved subsequently by Ohyama (1972) and Kemp and Wan (1976) provides a welfare-improving solution for the case of CUs. Starting from a situation with an arbitrary structure of trade barriers, if two or more countries freeze their net external

trade vector with the rest of the world through a set of common external tariffs and eliminate the barriers to internal trade (which implies the formation of a CU), the welfare of the union as a whole necessarily improves (weakly) and that of the rest of the world does not fall. A Pareto-improving preferential trade agreement may thus be achieved. The logic behind the Kemp–Wan theorem is as follows. By fixing the combined, net extra-union trade vector of member countries at its pre-union level, non-member countries are guaranteed their original level of welfare. Moreover, if we take the extra-union trade vector as an endowment, the joint welfare of the union is maximized by allowing free trade of goods internally (thus equating the marginal rate of substitution and marginal rate of transformation for each pair of commodities to each other and across all agents in the union). The PTA thus constructed has a common internal price vector, implying further a common external tariff for member countries. This customs union is (weakly) welfare improving; the rest of the world is no worse off and the welfare of member countries is jointly improved (weakly). Welfare improvement is achieved even if additional 'non-economic' objectives (such as maintaining the output of a sector or its employment of a factor) are introduced, as Krishna and Bhagwati (1997) show. The Kemp–Wan–Ohyama design, by freezing the external trade vector and thus eliminating trade diversion, offers a way to sidestep the complexities and ambiguities inherent in the analysis of PTAs. It has played an important role in shaping the way that economists think about issues relating to the design and implementation of PTAs.

The Kemp–Wan–Ohyama analysis of welfare improving CUs does not extend easily to FTAs, however. In the case of an FTA, member-specific tariff vectors imply that the domestic-price vectors differ across member countries and the FTA generally fails to equalize marginal rates of substitution across its members. Without a common internal price vector, however, the Kemp–Wan–Ohyama methodology lacks application. Nevertheless, Panagariya and Krishna (2002) have provided a corresponding construction of necessarily welfare-improving FTAs. The Panagariya–Krishna FTA, in complete analogy with the Kemp–Wan CU, freezes the external trade vector of the area, with the essential difference that the trade vector of each member country with the rest of the world is frozen at the pre-FTA level. Since, in FTAs, different member countries impose different external tariffs, it is necessary to specify a set of rules of origin to prevent a subversion of FTA tariffs by importing through the lower-tariff member country and directly trans-shipping goods to the higher-tariff country (which, if allowed, would bring the FTA arbitrarily close to a CU). The Panagariya–Krishna solution requires that all goods for which *any* value is added within the FTA are to be traded freely. Importantly, the proportion of domestic value added in final goods does not enter as a criterion in the rules of origin.

Theory thus suggests that ensuring welfare improvement requires that, along with elimination of internal barriers, external tariff vectors should eliminate trade diversion – member countries should continue to import the same amounts from the rest of the world as they did initially. There have, however, been significant departures in practice. While Article XXIV of the GATT stipulates that internal restrictions be eliminated on 'substantially all trade', the qualifier 'substantially' is vulnerable to abuse. Numerous goods are typically exempt from internal liberalization by member countries. Furthermore, restrictive rules of origin also serve to ensure a level of protection from both intra- and extra-union imports, as Krueger (1999) notes. On external tariffs, Article XXIV requires that external barriers not be more restrictive than initially. For FTAs, since countries retain individual tariff vectors, this could be taken to imply that no tariff is to rise. For CUs, since a common external tariff is to be chosen and initial tariffs on the same good likely vary across countries, the tariff vector would necessarily change for each country. The expectation is that that the 'general incidence' of trade barriers should not be higher or

more restrictive than before. As Bhagwati (1993) notes, it is clear that Article XXIV's ambiguity in this regard leaves plenty of room for protectionist behaviour by member countries. The 1994 ‘Understanding on the Interpretation’ of Article XXIV issued by the GATT provides greater clarity on the issue of measurement and choice of the common external tariff – indicating that the GATT secretariat would compute weighted average tariff rates and duties collected in accordance with the methodology used in the assessment of tariff offers in the Uruguay Round of trade negotiations and examine trade flow and other data to arrive at suitable measures of non-tariff barriers. Nevertheless, it may be observed that leaving external barriers at their initial level and removing internal barriers do not eliminate trade diversion. Indeed, with this configuration, some trade diversion is practically guaranteed.

Preferential trade agreements and multilateral free trade

Recent analysis in the literature has focused on issues concerning the expansion of trade blocs, the endogenous determination of policy (relating to trade preferences within a PTA and extra-union trade), and the effects of preferential agreements on the multilateral trade system (that is, whether trade blocs will serve as ‘building blocs’ or ‘stumbling blocs’ in the path to multilateral free trade, in Bhagwati's, 1993, phrasing).

Krugman (1993) analyses the welfare consequences of exogenously formed and expanded trade blocs. Considering a fully symmetric structure of countries, each specialized in production in a differentiated product variety, Krugman asks how world welfare is affected by the expansion of trade blocs if member countries liberalize fully their mutual trade but apply optimal tariffs against non-members. As the (symmetric) trade blocs increase in size, their market power increases and so do the (optimal) tariffs they impose on non-members. On the other hand, increasing the number of countries within a bloc increases the volume of goods that is traded freely. Krugman finds that the net effect on world welfare is non-monotonic in bloc size. Specifically, world welfare (which is maximized with global free trade) falls as the world is divided up into trade blocs but rises again as bloc sizes decrease and the trade diversion losses (relative to trade creation gains) fall. Bond and Syropoulos (1996) show how generalizing the assumptions of Krugman's model relating to consumption preferences and the pattern of production and trade may alter the relationship between optimal tariffs and market size so that optimal tariffs fall as bloc size increases. More severely, Deardorff and Stern (1994) and Srinivasan (1993) question the robustness of Krugman's conclusions concerning non-monotonicity of welfare itself, demonstrating a substantial divergence in results when Krugman's assumptions regarding the structure of endowments and comparative advantage are changed.

A different strand of the literature has examined endogenously determined trade blocs and the internal political and economic incentives (if any) for their successive expansion. Taking the ‘interest-group’ approach to trade policy determination, Grossman and Helpman (1995) and Krishna (1998) both model the influence of powerful producers in considering entry into a PTA. While the models and analytic frameworks differ in detail, they come to a similar and striking conclusion, namely, that PTAs that divert trade are more likely to win internal political support. This is so because governments must respond to conflicting pressures from their exporting sectors, which gain from lower trade barriers in the partner, and from their import-competing sectors, which suffer from lower trade barriers at home, when deciding on whether to form or enter a PTA.

As Krishna (1998) argues, trade diversion effectively shifts the burden of the gain to member-country exporters from member-country import-competing sectors and onto non-member producers, who have little political clout inside the member countries. Krishna (1998) also argues that such PTAs will lower the within-union incentives for any subsequent multilateral liberalization – producers in trade-diverting PTAs may oppose multilateral reform since this would take away the gains from benefits of preferential access that they enjoyed in the PTA that diverted trade to them. Under some circumstances, the within-union incentives for further multilateral liberalization are completely eliminated.

Levy (1997) models trade policy as being determined by majority voting. Countries are assumed to differ in their endowments of factors (labour and capital). Countries are also assumed to produce different varieties of goods – so that trade reform will result in gains to individuals due to the greater number of varieties that are available for consumption. However, it should also be clear that any changes in trade policy result in changes in the distribution of income (by altering the relative rewards to the different factors of production). The arguments that emerge out of this framework are as follows. First, preferential trade integration with partners with similar relative factor endowments (that is, with similar capital–labour ratios) is more likely to receive majority support – since this results in minimal income redistribution and still provides variety gains from trade. Second, bilateral agreements could render infeasible multilateral liberalization (which, even if it brings greater variety gains, would involve trade with countries with quite different relative endowments of capital and labour and could therefore result in much more drastic income redistribution).

McLaren (2002) provides an analysis of the role of sunk costs and trade policy determination. He argues, roughly speaking, that the expectation of a preferential trading agreement could induce agents in the economy to undertake costly and irreversible investments that makes the members within the bloc more specialized towards each other and less so towards the rest of the world. In other words, they increase dependence on each other and lower it towards the rest of the world, and thus reduce their desire to liberalize trade with other countries. Exploring the first-mover advantage that member countries gain having invested in sunk costs, Freund (2000a) finds that with preferential trade member countries gain and that non-members lose relative to multilateral outcomes, with the former dominating the latter in magnitude.

A parallel literature has raised the question of what external tariffs will be chosen by member countries, examining, in particular, whether external tariffs can be expected to rise or fall following a PTA. No clear answers to this question emerge. Panagariya and Findlay (1996) finds that external tariffs rise after tariff preferences are granted, as political lobbying for protection is directed away from imports from the partner country to imports from the rest of the world. Emphasizing tariff revenue competition between FTA members, Richardson (1995) finds that external tariff may fall in an FTA as welfare-minded member countries competitively reduce tariffs (so as to retain the source of extra-union imports and earn tariff revenue). In a general equilibrium context, with political lobbying over tariffs, Cadot, de Melo and Olarreaga (1999) reach a similar conclusion for FTAs, while finding that CUs are likely to raise their external tariffs. Cadot, de Melo and Olarreaga (2002) confirm these results for the case of quantity restrictions, where the protective effect that a quantity restriction imposed by a member country has in partner country markets proves central to the analysis. However, when collective action problems over lobbying for external trade policy dominate, FTAs may choose higher external tariffs than CUs, as Richardson (1994) argues. Finally, in a symmetric three-country oligopoly model, Ornelas (2005) finds

that a PTA's endogenously determined external tariff may be lower than the pre-union MFN tariff. Cho and Krishna (2006) demonstrate that the opposite may obtain with asymmetric costs across partner countries, the likelihood of the external tariff being higher than the pre-union MFN tariff increasing with the inefficiency of the partner country relative to non-members.

Empirical analysis has offered mixed results as well. Bohara, Gawande and Sanguinetti (2004) report evidence of lowered external tariffs following Mercosur, while Cho (2006) finds tariffs in Mexico to be higher on average following NAFTA and systematically higher in goods in which its trading partners were inefficient suppliers (as proxied by pre-FTA export levels relative to the rest of the world). Limão (2006) examines data on US trade barriers and finds that those imported goods on which any partners received preferences were subject to smaller (subsequent) multilateral liberalization than others, suggesting a negative effect of preferences on multilateral reforms.

The economic incentives of non-member countries are considered by Baldwin (1995), who argues that PTA expansion could have ‘domino’ effects – increasing the size of a bloc increases the incentive for others to join it (as they then gain preferential access to increasingly large markets). On the assumption of open-membership rules (that is, insiders do not oppose the entry of new members who abide by the same rules as the members), the successive expansion of the PTA could then lead to multilateral free trade – a conclusion that is also reached by the work of Yi (1996), which develops a model of endogenous coalition formation to addresses this question.

Aghion, Antras and Helpman (2004) analyse the links between bilateral and multilateral negotiations over trade policy as a dynamic bargaining game in which a leading country endogenously decides whether to sequentially negotiate free trade agreements with subsets of countries or engage in simultaneous multilateral bargaining with all countries at once. They show that, if a coalition formed between the leading country and a follower generates a negative effect on outsiders (that is, there are negative coalitional externalities), the leader prefers sequential bargaining to multilateral bargaining. Conversely, positive coalition externalities imply that multilateral bargaining is preferred. Importantly, while political economy pressures may cause bilateral agreements to impede multilateral agreement, as in Levy (1997) and Krishna (1998), examples where bilateral agreements enable multilateral agreement are also found.

Self-enforcing trade agreements (which work by balancing any benefits that member countries may achieve by deviating from the agreement with the future losses they suffer due to punishments for the deviation) have been variously analysed in the international trade literature. Since bilateral (multilateral) agreements may alter both the benefits of deviating from an existing multilateral (bilateral) agreement and the future punishment costs of this deviation, self-enforcing agreements provide a context in which the links between preferential trade agreements and multilateralism may be studied. Bagwell and Staiger (1997a; 1997b) consider the impact of FTAs and CUs on multilateral tariff cooperation during a transition period when the exogenously agreed-upon lowering of tariffs within FTAs and CUs is implemented. Saggi (2006) shows how exogenously specified FTAs and CUs may undermine self-enforced multilateral tariff cooperation, the former by lowering the cooperation incentives of non-member countries and the latter by lowering the cooperation incentives of members. Freund (2000b) finds that exogenous multilateral liberalization may encourage and help sustain self-enforcing PTAs, thus explaining the recent trend towards bilateralism as a causal response to multilateralism. A similar

causal link is explored by Cadot, de Melo and Olarreaga (2001), which views bilateral agreements as an endogenous (protective) response to the competitive pressures that domestic producers face with multilateral liberalization.

Conclusions

A half-century of research has advanced significantly our understanding of the implications of trade discrimination even if the frequently equivocal theoretical and empirical results have established among economists and policymakers an ambivalent attitude towards preferential trade agreements. However, concerns regarding the fragmentation of the world trade system have grown with the rapid proliferation of preferential trade in recent years. Several hundred PTAs are currently in existence. Indeed, many countries belong to multiple PTAs – resulting in a confusing criss-crossing of trade preferences that Bhagwati (1995) has aptly described as ‘spaghetti-bowl’ regionalism. Several more preferential agreements are in process. With this inexorable erosion of non-discriminatory disciplines within the trade system, research on preferential trade is certain to remain central to the field of international trade policy for many years to come.

See Also

- Mercosur
- North American Free Trade Agreement (NAFTA)
- trade policy, political economy of

Bibliography

Aghion, P., Antras, P. and Helpman, E. 2004. Negotiating free trade. Working Paper No. 10721. Cambridge, MA: NBER.

Bagwell, K. and Staiger, R. 1997a. Multilateral cooperation during the formation of free trade areas. *International Economic Review* 38, 291–319.

Bagwell, K. and Staiger, R. 1997b. Multilateral cooperation during the formation of customs unions. *Journal of International Economics* 42, 91–123.

Baldwin, R. 1995. A domino theory of regionalism. In *Expanding European Regionalism: The EU's New Members*, ed. R. Baldwin, P. Haaparanta and J. Kiander. Cambridge: Cambridge University Press.

Berglas, E. 1979. Preferential trading: the n commodity case. *Journal of Political Economy* 87, 315–31.

Bhagwati, J. 1971. Trade-diverting customs unions and welfare improvement: a clarification. *Economic Journal* 81, 580–7.

Bhagwati, J. 1993. Regionalism and multilateralism: an overview. In *New Dimensions in Regional Integration*, ed. J. de Melo and A. Panagariya. Cambridge: Cambridge University Press.

Bhagwati, J. 1995. U.S. trade policy: the infatuation with free trade areas. In *The Dangerous Drift to Preferential Trade Agreements*, ed. J. Bhagwati and A. Krueger. Washington, DC: AEI Press.

Bhagwati, J. and Panagariya, A. 1996. *Free Trade Areas or Free Trade? The Economics of Preferential Trade Areas*. Washington, DC: AEI Press.

Bohara, A., Gawande, K. and Sanguinetti, P. 2004. Trade diversion and declining tariffs: evidence from Mercosur. *Journal of International Economics* 64, 65–88.

Bond, E. and Syropoulos, C. 1996. The size of trading blocs, market power and world welfare effects. *Journal of International Economics* 40, 411–37.

Cadot, O., de Melo, J. and Olarreaga, M. 1999. Regional integration and lobbying for tarriffs against non-members. *International Economic Review* 40, 635–58.

Cadot, O., de Melo, J. and Olarreaga, M. 2001. Can bilateralism ease the pains of multilateral trade liberalization? *European Economic Review* 45, 27–44.

Cadot, O., de Melo, J. and Olarreaga, M. 2002. Harmonizing external quotas in an FTA: a step backward? *Economics and Politics* 14, 259–82.

Chang, W. and Winters, A. 2002. How regional blocs affect excluded countries: the price effects of Mercosur. *American Economic Review* 92, 889–904.

Cho, M.-J. 2006. On the external trade barriers of preferential trade agreements: empirical evidence from Mexican manufacturing industries. Mimeo, Brown University.

Cho, M.-J. and Krishna, P. 2006. On the external trade barriers of preferential trade agreements. Mimeo, Brown University.

Deardorff, A. and Stern, R. 1994. Multilateral trade negotiations and preferential trading arrangements. In *Analytical and Negotiating Issues in the Global Trading System*, ed. A. Deardorff and R. Stern. Ann Arbor: University of Michigan Press.

Freund, C. 2000a. Different paths to free trade: the gains from regionalism. *Quarterly Journal of Economics* 115, 1317–41.

Freund, C. 2000b. Multilateralism and the endogenous formation of preferential trade agreements. *Journal of International Economics* 52, 359–76.

Gehrels, F. 1956–57. Customs union from a single–country viewpoint. *Review of Economic Studies* 24, 61–4.

Grossman, G. and Helpman, E. 1995. The politics of free trade arrangements. *American Economic Review* 85, 667–90.

Kemp, M. 1964. *The Pure Theory of International Trade*. Englewood Cliffs, NJ: Prentice-Hall.

Kemp, M. and Wan, H. 1976. An elementary proposition concerning the formation of customs unions. *Journal of International Economics* 6, 95–8.

Krishna, P. 1998. Regionalism and multilateralism: a political economy approach. *Quarterly Journal of Economics* 113, 227–50.

Krishna, P. 2003. Are regional trading partners ‘natural’? *Journal of Political Economy* 111, 202–31.

Krishna, P. and Bhagwati, J. 1997. Necessarily welfare-improving customs unions with industrialization constraints: the Cooper–Massell–Johnson–Bhagwati proposition. *Japan and the World Economy* 154, 169–87.

Krueger, A. 1999. Free trade agreements as protectionist devices: rules of origin. In *Trade, Theory and Econometrics: Essays in Honor of John S. Chipman*, ed. J. Melvin, R. Moore and R. Riezman. Studies in the Modern World Economy, vol. 15. London: Routledge.

Krugman, P. 1991. The move to free trade zones. In *Policy Implications of Trade and Currency Zones*, Federal Reserve Bank of Kansas City. Kansas City: Federal Reserve Bank of Kansas City.

Krugman, P. 1993. Regionalism versus multilateralism: analytical notes. In *New Dimensions in Regional Integration*, ed. J. de Melo and A. Panagariya. Cambridge: Cambridge University Press.

Levy, P. 1997. A political economic analysis of free trade arrangements. *American Economic Review* 87, 506–19.

Limão, N. 2006. Preferential trade agreements as stumbling blocks for multilateral trade liberalization: evidence for the United States. *American Economic Review* 96, 896–914.

Lipsey, R. 1957. The theory of customs unions: trade diversion and welfare. *Economica* 24, 40–6.

McLaren, J. 2002. A theory of insidious regionalism. *Quarterly Journal of Economics* 117, 571–608.

Meade, J. 1955. *The Theory of Customs Unions*. Amsterdam: North-Holland.

Mundell, R. 1964. Tariff preferences and the terms of trade. *Manchester School of Economic Studies* 32, 1–13.

Ohyama, M. 1972. Trade and welfare in general equilibrium. *Keio Economic Studies* 9, 37–73.

Ornelas, E. 2005. Endogenous Free trade agreements and the multilateral trading system. *Journal of International Economics* 67, 471–97.

Panagariya, A. 1997a. The Meade model of preferential trading: history, analytics and policy implications. In *International Trade and Finance: New Frontiers for Research: Essays in Honor of Peter Kenen*, ed. B. Cohen. Cambridge: Cambridge University Press.

Panagariya, A. 1997b. Preferential trading and the myth of natural trading partners. *Japan and the World Economy* 9, 471–89.

Panagariya, A. 2000. Preferential trade liberalization: the traditional theory and new developments. *Journal of Economic Literature* 161, 316–60.

Panagariya, A. and Findlay, R. 1996. A political economy analysis of free trade areas and customs unions. In *The Political Economy of Trade Reform: Essays in Honor of Jagdish Bhagwati*, ed. R. Feenstra, D. Irwin and G. Grossman. Cambridge, MA: MIT Press.

Panagariya, A. and Krishna, P. 2002. On necessarily welfare-enhancing free trade areas. *Journal of International Economics* 57, 353–67.

Richardson, M. 1994. Why a free trade area? The tariff also rises. *Economics & Politics* 6, 79–95.

Richardson, M. 1995. Tariff revenue competition in a free trade area. *European Economic Review* 39, 1429–37.

Riezman, R. 1979. A 3×3 model of customs unions. *Journal of International Economics* 9, 341–54.

Saggi, K. 2006. Preferential trade agreements and multilateral tariff cooperation. *International Economic Review* 47, 29–57.

Srinivasan, T. 1993. Discussion on regionalism vs multilateralism: analytical notes. In *New Dimensions in Regional Integration*, ed. A. Panagariya and J. de Melo. Washington, DC: World Bank.

Srinivasan, T., Whalley, J. and Wooton, I. 1993. Measuring the effects of regionalism on trade and welfare. In *Regional Integration and the Global Trading System*, ed. K. Anderson and R. Blackhurst. New York: St Martin's Press.

Summers, L. 1991. Regionalism and the world trading system. In *Policy Implications of Trade and Currency Zones*, Federal Reserve Bank of Kansas City. Kansas City: Federal Reserve Bank of Kansas City.

Vanek, J. 1965. *General Equilibrium of International Discrimination*. Cambridge, MA: Harvard University Press.

Viner, J. 1950. *The Customs Unions Issue*. New York: Carnegie Endowment for International Peace.

Wonnacott, P. and Lutz, M. 1987. Is there a case for free trade areas? In *Free Trade Areas and US Trade Policy*, ed. J. Schott. Washington, DC: Institute for International Economics.

Yi, S.-S. 1996. Endogenous formation of customs unions under imperfect competition: open regionalism is good. *Journal of International Economics* 41, 153–77.

Yeats, A. 1998. Does Mercosur's trade performance raise concerns about the effects of regional trade arrangements? *World Bank Economic Review* 12, 1–28.

How to cite this article

Krishna, Pravin. "regional and preferential trade agreements." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000246> doi:10.1057/9780230226203.1412

The New Palgrave Dictionary of Economics Online

regional development, geography of

Jeffrey D. Sachs and Gordon McCord

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

New theoretical work on spatial concentration of industry – particularly the ‘new economic geography’ – has significantly helped our understanding why some regions develop more than others, why cities arise and where they are located. However, this work rarely incorporates Adam Smith's observation that spatial differences in economic activity also reflect variations in physical geography, which make some places more productive than others at particular times; nor has it accommodated regional development policy – the use of economic incentives to attract industry to particular locations. A full theory of regional development would integrate theories of agglomeration economies with physical geography and with public economics.

Keywords

agglomeration; agricultural productivity; backward and forward linkages; cities; climate; clusters; coastal proximity; comparative advantage; competitive advantage; complementarities; core–periphery model; demand smoothing; foreign direct investment; geography and economic development; growth poles; human capital accumulation; industrial policy; knowledge spillovers; learning region; linkages; location choice; monopolistic competition; new economic geography; nutrition and development; preferential attachment; regional development; returns to scale; search costs; spatial concentration; Special Economic Zones (China); transport costs

Article

Differences in economic activity across regions have interested economists since Adam Smith, who argued that high overland transport costs in the interior of Africa and Asia ‘seem in all ages’ to have had hindered economic development. However, economists' attraction to the study of spatial variations in economic activity has fluctuated over time. Standard trade theory based on comparative advantage helps to explain how the location of economic activity is affected by the spatial distribution of primary

resources (such as land, labour and water), but standard trade theory says little about the interdependence of location decisions by economic agents, nor does it consider in any depth the more detailed aspects of physical geography (climate, soils, topography, disease epidemiology).
Neoclassical growth models focus on the accumulation of physical, human and technological capital, which individually or together complement raw labour and land as factors of production, but only recent theory (particularly in the work dubbed the ‘new economic geography’) has begun to grapple with location choices and the spatial concentration of industry (Henderson, 1988; Krugman, 1991; Fujita, Krugman and Venables, 1999). While these newer theories have contributed importantly to our understanding of why some regions develop more than others, and why cities arise and where they are located, they rarely incorporate Smith's observation that spatial differences in economic activity are also related to variations in physical geography, which intrinsically make some places more productive than others at particular points in time. Nor do they yet go into depth on regional development policy, that is, the use of economic incentives to attract industry to one location or another. A full theory of regional development will integrate theories of agglomeration economies with physical geography and with public economics.

Theories of agglomeration

Economic activity and population around the globe are concentrated in highly dense metropolitan areas, which suggest that there is an important economic benefit of economic agglomeration (spatial co-location of economic agents). Alfred Marshall (1920) suggested that spatial concentration happens because of knowledge spillovers, larger markets for specialized skills, and backward and forward linkages associated with large local markets.
The initial literature to tackle the intractability of modelling economic geography grew from the von Thünen model (1826), which begins with the existence of a city and derives characteristics about land rents and land use surrounding the city; the resulting unplanned, efficient outcome is a concentric ring pattern of production referred to as ‘von Thünen cones’. The model doesn't, however, attempt to explain the *raison d’être* of the city itself.
Later models aimed to explain why population and economic activity tend to agglomerate in the first place. Spatial concentration occurs because production is cheaper due to the large amount of nearby economic activity in agglomeration economies. These increasing returns to scale exist for several reasons: larger markets support more highly specialized products; efficiency increases as a large number of producers and consumers allows for less idle time (a source of increasing returns called demand smoothing); economies of scale of intermediate inputs make production cheaper even for sectors without increasing returns; externalities diffuse learning and expertise, as people can see each others' products and work methods; and search costs are lowered when the search process is spatially concentrated. Florida (1995) pioneered the concept of the ‘learning region’: to minimize transport costs and maximize learning, firms benefit from spatially concentrating their activities, and thus firms looking to augment their capabilities have strong incentive to locate in these learning regions.

New economic geography

The ‘new economic geography’ of recent decades grew from the Dixit and Stiglitz (1977) model of monopolistic competition under increasing returns to scale. Though admittedly a special case, this model became a workhorse in many fields, and a foundation for the new economic geography. The theoretical backbone of new economic geography is the core–periphery model in Krugman (1991), which looks at three effects: the ‘market-access effect’ (monopolistic firms locate in big markets and export to small markets), the ‘cost-of-living effect’ (cost of living is cheaper where there are more firms, due to low transport costs), and the ‘market-crowding effect’ (imperfectly competitive firms look to locate in regions with few competitors). The model was an important step forward in understanding spatial dynamics, but has the downside of being difficult to manipulate analytically and requires numerical simulations (instead of explicit expressions) to derive results.

Another important concept in the location of economic activity is that of clusters, especially in the work of Porter (1990; 1995; 1998a; 1998b). A cluster is a group of interconnected companies and institutions in a particular location (perhaps a city, or a state, or even a group of neighbouring countries). Companies in a cluster benefit from important complementarities, spillovers and a relationship with public institutions, which improve productivity and productivity growth, and stimulate new business formation. The important contribution of this literature is that a firm's comparative advantage (or ‘competitive advantage’ in the business phrase) can include characteristics outside the firm itself; often geography and location have important implications on how firms or industries can compete in the market.

One of the striking implications of the new economic geography is that spatial concentration arises in a homogeneous region, where there is no fundamental geographical advantage to locating in one place or another. The precise location of firms is accidental. Early advantages in agglomeration can lead to a snowball effect. First movers in regional development can achieve a lasting competitive advantage by attracting other mobile workers and investors. Growth proceeds with ‘preferential attachment’ to the places that get an early start.

The role of physical geography

In addition to the new economic geography models of agglomeration, a second basic approach seeking to shed light on growth poles and regional development is based on intrinsic geographical advantages. The assumption of homogeneous space is abandoned, and the role of coasts, hinterlands, rivers, mountains and a vast array of other geographical variables is brought to the fore. Adam Smith himself asserted that the division of labour is limited by the extent of the market, so that coastal regions, by virtue of their ability to engage in sea-based trade, enjoy a wider scope of the market than interior regions. More recently, climatic conditions have been found to have pervasive effects on regional development through disease ecology, agricultural productivity, transport costs, vulnerability to natural hazards, water stress and other factors that may affect economic performance.

Several studies (Gallup, Sachs and Mellinger, 1999; Bloom and Sachs, 1998) have noted that tropical areas are consistently poorer than temperate-zone areas, and hypothesize that this may be related to the effects of tropical ecology on human health and agricultural productivity. Tropical infectious diseases, for example, impose very high burdens on human health that in turn may lead to shortfalls in economic performance much larger than their direct short-run effects on health. Another study (Gallup and Sachs, 2000) found that, after purchased inputs such as capital, labour and fertilizers are controlled for, the

average productivity of tropical food production falls short of the productivity of temperate-zone food production. In the course of economic development, this poor performance in food productivity may have had serious adverse effects on nutrition levels, with adverse consequences for human capital accumulation, labour productivity and susceptibility to infectious disease. These geographical penalties can often be compensated by other kinds of interventions (such as malaria control or improved agronomic practices), but, since those interventions require added resources, affected regions may persistently lag behind more fortuitously located regions.

Geographical advantages can trigger subsequent agglomeration based on increasing returns to scale. The agglomeration is then self-reinforcing, even after the initial spatial advantage loses some of its importance. For example, Chicago's port is not as important as when it was the main driver of the city's growth in the middle of the 19th century. Glaeser (2005) illustrates that New York's rise in the 19th century was due to a technological change that moved ocean shipping from a point-to-point system to a hub and spoke system, and the city's geography made it the natural hub. Today, however, New York's pre-eminence is based not mainly on the port, but on the legacies of the earlier success: finance, business, remarkable infrastructure and the benefits of agglomeration.

Changing dimensions of geography

It is important to stress the changing nature of a region's geographic advantage as technology changes. In early civilizations, when transport and communications were too costly to support much interregional and international trade, regional advantage came from agricultural productivity and local transport rather than from access to oceans. As a result, early civilizations almost invariably emerged in highly fertile river valleys such as those around the Nile, Indus, Tigris, Euphrates, Yellow and Yangtze rivers. These civilizations produced high-density populations that in later eras were often disadvantaged by their remoteness from international trade. As the advantages of overland trade between Europe and Asia gave way to oceanic commerce in the 16th century and later, and as the trade routes to the Americas were discovered, economic advantage shifted from the Middle East and eastern Mediterranean to the North Atlantic. In the 19th century, the high costs of transporting coal for steam power meant that industrialization almost invariably depended on proximity to coal fields.

In the late 20th century, air transport and telecommunications have reduced the advantages of coastlines relative to hinterlands. The telecommunications sector, in particular, is deeply affecting the global division of labour and the nature of agglomeration economies. The disadvantages of interior and distant regions may well be eased or eliminated by the advances in telecommunications which allow for more disbursed production and new growth poles far from traditional trade routes. It is notable that Bangalore has become a booming centre of global information technology, despite being an inland city in southern India, and despite the weakness of India's roads and ports at the time of Bangalore's ascendancy. The examples of Bangalore and of course California's Silicon Valley show that today's competitive advantage has to do much more with the location of excellent universities and an attractive living environment for highly skilled and mobile information workers, much like the ‘learning regions’ described by Florida (1995).

Regional policy design

The presence of agglomeration economies, increasing returns and clusters suggests that countries can identify areas of potential growth poles and use policy tools and public investment to trigger these processes. Special policy instruments such as export-processing zones and special tax promotion schemes have helped developing countries to establish clusters in textiles and apparel, electronics, consumer appliances, software and automotive components, to name just a few industries where active industrial policy has played a hand. In the case of growth poles in the knowledge economy (such as Silicon Valley and Bangalore), the importance of government support for higher education and R&D and for the creation of science parks is especially apparent. Spillovers from military technology may play a role as well.

It is clear, however, that the successful development strategies of some countries cannot produce the same salubrious results when implemented in very different settings. When China opened some coastal pockets for foreign direct investment, these Special Economic Zones quickly blossomed into vibrant export platforms and created backward linkages with the immediate hinterland. When landlocked Mongolia turned the entire country into a free trade and investment zone in the late 1990s, however, the inflow of foreign capital was a trickle compared with China's experience, and was based mainly on primary commodities (such as copper). Even within China, the coastal provinces in the east have boomed relative to the interior provinces of western China. Physical geography therefore continues to condition economic development. Geographical determinism should be avoided, however; special geographical hindrances may well call for special compensating investments (in roads, disease control, telecommunications, and so on), or for promotion of a judicious choice of industries (those that can be sustained in the face of high transport costs, for example).

Empirical studies

Empirical evidence supports the idea that economies of scale, agglomeration forces (Davis and Weinstein, 1998; 1999; Midelfart-Knarvik et al., 2000; Overman and Puga, 2002; Hanson, 2005), and backward and forward linkages (Midelfart-Knarvik and Steen, 1999) help explain why economic activity clusters together, and that the von Thünen model helps explain economic dynamics near cities (Fafchamps and Shilpi, 2003). The traditional core–periphery model has considerable empirical support, given that the core regions of the global economy (particularly North America, Western Europe, and Japan), enjoy ever-increasing levels in productivity. At a smaller scale, studies of wages in the United States and in developing countries show that *ceteris paribus* workers earn much more in urban areas than rural areas, reflecting their higher productivity (Glaeser and Mare, 1994; Bairoch, 1988).

While looking for the presence of increasing returns to scale yields insights, it does not address the constraints physical geography may place upon economic growth. For example, Adam Smith's observations on the role of access to navigable water still hold. Cross-country empirical research affirms that the level and growth rate of per capita income continue to be strongly positively correlated with geographic variables such as climate and coastal proximity (Gallup, Sachs and Mellinger, 1999; Mellinger, Sachs and Gallup, 2000), while within-country differences in growth rates in India and China are clearly related to geography as well (Demurger et al., 2002; Sachs, Bajpai and Ramiah, 2002).

Smith's observations also implicitly underscore the highly favourable economic geography enjoyed by the nations of western Europe. Extensive ocean shorelines host a succession of natural harbours, and

numerous navigable rivers penetrate deep into the interior. In addition, despite the large landmass of the United States, 57 per cent of income was generated in counties within 80•km from the coast, though these counties account for only 13 per cent of land mass (Rappaport and Sachs, 2003).

Future theoretical and empirical work in understanding regional development should aim to disentangle the forces of differential geography and self-organizing agglomeration economies. Policy studies should examine in depth how regional development policy has been used in the past, and which instruments are particularly important. Economists and business specialists should aim to provide new tools to help specific regions identify appropriate instruments for regional development, including which kinds of industries are likely to flourish in which kinds of spatial settings.

See Also

- location theory
- Marshall, Alfred
- new economic geography
- spatial economics
- Thünen, Johann Heinrich von
- urban agglomeration

Bibliography

Bairoch, P. 1988. *Cities and Economics Development: From the Dawn of History to the Present*. Chicago: University of Chicago Press.

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G. and Robert-Nicoud, F. 2003. *Economic Geography and Public Policy*. Princeton: Princeton University Press.

Bloom, D. and Sachs, J.D. 1998. Geography, demography, and economic growth in Africa. *Brookings Papers on Economic Activity* 1998(2), 207–95.

Clark, G., Feldman, M.P. and Gertler, M. 2000. *The Oxford Handbook of Economic Geography*. Oxford: Oxford University Press.

Davis, D.R. and Weinstein, D.E. 1998. Market access, economic geography, and comparative advantage: an empirical assessment. Working Paper No. 6787. Cambridge, MA: NBER.

Davis, D. and Weinstein, D.E. 1999. Economic geography and regional production structure: an empirical investigation. *European Economic Review* 43, 379–407.

Demurger, S., Sachs, J.D., Wing, T.-W., Bao, S., Chang, G. and Mellinger, A. 2002. Geography, economic policy, and regional development in China. *Asian Economic Papers* 1, 146–97.

Dixit, A.K. and Stiglitz, J.E. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Dunning, J.H., ed. 2002. *Regions, Globalization, and the Knowledge-Based Economy*. New York: Oxford University Press.

Fafchamps, M. and Shilpi, F. 2003. The spatial division of labor in Nepal. *Journal of Development Studies* 39(6), 23–66.

Florida, R. 1995. Towards the learning region. *Futures* 27, 527–36.

Fujita, M., Krugman, P. and Venables, A.J. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Gallup, J.L. and Sachs, J.D. 2000. Agriculture, climate, and technology: why are the tropics falling behind? *American Journal of Agricultural Economics* 82, 731–77.

Gallup, J.L., Sachs, J.D. and Mellinger, A.D. 1999. Geography and economic development. *International Regional Science Review* 22, 179–232.

Glaeser, E.L. 2005. Urban colossus: Why is New York America's largest city? Working Paper No. 11398. Washington, DC: NBER.

Glaeser, E. and Mare, D. 1994. Cities and skills. Working Paper No. E94-11, Hoover Institution.

Hanson, G.H. 2005. Market potential, increasing returns, and geographic concentration. *Journal of International Economics* 67, 1–24.

Henderson, J.V. 1988. *Urban development: Theory, Fact, and Illusion*. Oxford: Oxford University Press.

Henderson, J.V. and Thisse, J.-F., eds. 2004. *Handbook of Regional and Urban Economics*, vol. 4. Amsterdam: North-Holland.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99, 483–99.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Mellinger, A.D., Sachs, J.D. and Gallup, J.L. 2000. Climate, coastal proximity, and development. In *Oxford Handbook of Economic Geography*, ed. G.L. Clark, M.P. Feldman and M.S. Gertler. Oxford:

Oxford University Press.

Midelfart-Knarvik, K.H., Overnman, H.G., Redding, S.J. and Venables, A.J. 2000. The location of European industry. Economic Papers No. 142. European Commission Directorate-General for Economic and Financial Affairs.

Midelfart-Knarvik, K. and Steen, F. 1999. Self-reinforcing agglomerations? An empirical industry study. *Scandinavian Journal of Economics* 101, 515–32.

O'Flaherty, B. 2005. *City Economics*. Cambridge, MA: Harvard University Press.

Overman, H.G. and Puga, D. 2002. Unemployment clusters across European regions and countries. *Economic Policy* 34, 115–47.

Porter, M. 1990. *The Competitive Advantage of Nations*. New York: Free Press.

Porter, M. 1995. The competitive advantage of the inner city. *Harvard Business Review* 73, 55–71.

Porter, M. 1998a. Clusters and competition: new agendas for companies, governments and institutions. In *On Competition*. Boston: Harvard Business School Press.

Porter, M. 1998b. The microeconomic foundations of economic development. In World Economic Forum, *Global Competitiveness Report 1998*. Geneva: WEF.

Rappaport, J. and Sachs, J.D. 2003. The United States as a coastal nation. *Journal of Economic Growth* 8, 5–46.

Sachs, J.D., Bajpai, N. and Ramiah, A. 2002. Understanding regional economic growth in India. *Asian Economic Papers* 1(3), 32–62.

Saxenian, A.L. 2006. *The New Argonauts: Regional Advantages in a Global Economy*. Cambridge, MA: Harvard University Press.

von Thünen, J.H. 1826. *Der Isolierte Staat in Beziehung auf Landtschaft und Nationalökonomie*. Hamburg. Trans V.M. Wartenberg as *Von Thünen's Isolated State*. Oxford: Pergamon Press, 1966.

How to cite this article

Sachs, Jeffrey D. and Gordon C. McCord. "regional development, geography of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January

2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000077>
doi:10.1057/9780230226203.1413

The New Palgrave Dictionary of Economics Online

regression-discontinuity analysis

Wilbert van der Klaauw

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In recent years regression discontinuity analysis has grown into a popular approach for evaluating causal relationships in empirical economics. The method takes advantage of a discontinuity in the probability of treatment as a function of a continuous variable to identify a meaningful average treatment effect. This article summarizes the regression discontinuity approach to identifying and estimating causal effects and describes several validity tests.

Keywords

assignment mechanisms; control function approach; identification; instrumental variables; kernel estimators; propensity score; regression discontinuity analysis; selection bias; semiparametric estimation; treatment effect

Article

The regression discontinuity (RD) data design is a quasi-experimental evaluation design first introduced by Thistlethwaite and Campbell (1960) as an alternative approach to evaluating social programmes. The design is characterized by a treatment assignment or selection rule which involves the use of a known *cut-off* point with respect to a continuous variable, generating a discontinuity in the probability of treatment receipt at that point. Under certain comparability conditions, a comparison of average outcomes for observations just left and right of the cut-off can be used to estimate a meaningful causal impact. While interest in the design had previously been mainly limited to evaluation research methodologists (Cook and Campbell, 1979; Trochim, 1984), the design is currently experiencing a renaissance among econometricians and empirical economists (Hahn, Todd and van der Klaauw, 1999; 2001; Angrist and Krueger, 1999; Porter, 2003). Among the main econometric contributions have been the formal derivation of identification conditions for causal inference and the introduction of semiparametric estimation procedures for the design. At the same time, a large and rapidly growing

number of empirical applications are providing new insights into the applicability of the design, which have led to the development of several sensitivity and validity tests.

The popularity of the RD design in applied economic research can be linked to several of its features. First, the assignment rules in many existing programmes and procedures for allocating social resources, frequently lend themselves to RD evaluations. In many cases, programme resources are allocated based on some type of formula that has a cut-off structure. One area of economic research where the design has proven especially fruitful in recent years has been the evaluation of educational interventions. Education programmes are frequently assigned to schools or students who score below a cut-off on some scale (student performance, poverty), and school and programme funding decisions are often based on allocation formulas containing discontinuities. Similarly, the design has proven useful in evaluating the socio-economic impacts of a diverse set of government programmes and laws, many of which use eligibility cutoffs or funding formulas involving thresholds in allocating scarce resources to those potential recipients who need or deserve them most (see van der Klaauw, 2007a). A second attractive feature of the design is that it is intuitive and its results can be easily communicated, often with a visual portrayal of sharp changes in both treatment assignment and average outcomes around the cut-off value of the assignment variable (Bloom et al., 2005). Third, a researcher can choose from among several different methods to estimate effects that have credible causal interpretations (Hahn, Todd and van der Klaauw, 2001).

Consider the general problem of evaluating the impact of a binary treatment on an outcome variable, using a random sample of individuals where for each individual i we observe an outcome measure y_i and a binary treatment indicator t_i, equal to one if treatment was received and zero otherwise. The evaluation problem that arises in determining the effect of t on y, is due to the fact that each individual either receives or does not receive treatment and is never observed in both states. Let $y_i(1)$ be the outcome given treatment, and $y_i(0)$ the outcome in absence of treatment. Then the actual outcome we observe equals $y_i = t_i y_i(1) + (1 - t_i) y_i(0)$. A common regression model representation for the observed outcome can then be written as

$$y_i = \beta + \alpha_i t_i + u_i \tag{1}$$

where $\alpha_i = y_i(1) - y_i(0)$ and $y_i(0) = E[y_i(0)] + u_i = \beta + u_i$. Non-random assignment or selection into treatment implies that a comparison of average outcomes of treatment recipients and non-recipients ($E[y_i(1)|t_i=1]$ and $E[y_i(0)|t_i=0]$) would generally not provide us with a valid treatment effect estimate.

Hahn, Todd and van der Klaauw (HTV) (2001) analysed the conditions under which a discontinuity in the treatment assignment or selection rule can be exploited to solve the selection bias problem and to identify a meaningful causal effect. Following Trochim (1984) they distinguish between two different forms of the design, depending on whether the treatment assignment is related to the assignment variable by a deterministic function (*sharp design*) or a stochastic one (*fuzzy design*). In the case of a sharp RD

design, individuals are assigned to or selected for treatment solely on the basis of a cut-off score on an observed continuous variable *x*. This variable, alternatively called the assignment, selection, running or ratings variable, could represent a single characteristic or a composite variable constructed using several characteristics. Those who fall below some distinct cutoff point $\bar{x}$ are placed in the control group (t_i=0), while those on or above that point are placed in the treatment group (t_i=1) (or vice versa). Thus, assignment occurs through a known and measured deterministic decision rule: $t_i = t(x_i) = 1\{x_i \geq \bar{x}\}$ where 1{.} is the indicator function. As the assignment variable itself may be correlated with the outcome variable, the assignment mechanism is clearly not random.

However, if we have reason to believe that persons close to the threshold with very similar *x* values are comparable, then we may view the design as almost experimental near $\bar{x}$, suggesting that we could evaluate the causal impact of treatment by comparing the average outcome for those with ratings just above to those with ratings just below the cutoff. More formally, consider the following *local continuity* (*LC*) *assumption*:

> $E[u_i|x]$ and $E[\alpha_i|x]$ are continuous in *x* at $\bar{x}$, or equivalently, $E[y(1)|x]$ and $E[y(0)|x]$ are continuous at $\bar{x}$,

then on the assumption that the density of *x* is positive in a neighbourhood containing $\bar{x}$,

$$\lim_{x\downarrow\bar{x}} E[y_i|x] - \lim_{x\uparrow\bar{x}} E[y_i|x] = \lim_{x\downarrow\bar{x}} E[\alpha_i t_i|x] - \lim_{x\uparrow\bar{x}} E[\alpha_i t_i|x] + \lim_{x\downarrow\bar{x}} E[u_i|x] - \lim_{x\uparrow\bar{x}} E[u_i|x] = E[\alpha_i|x = \bar{x}]. \quad (2)$$

The RD approach therefore identifies the average treatment effect for individuals close to the discontinuity point. Note that the continuity assumption formalizes the idea that individuals just above and below the cut-off need to be ‘comparable’, requiring them to have similar average potential outcomes when receiving and when not receiving treatment. While in the absence of additional assumptions (such as a common effect assumption) one could learn about treatment effects only for a sub-population of persons near the discontinuity point, as pointed out by HTV this local effect is highly relevant to policymakers who are contemplating less restrictive eligibility rules and marginal expansions of programmes via a change in the cut-off.

The continuity assumption required for identification is not innocuous. Even if treatment receipt is determined solely on the basis of a cut-off score on the assignment variable, this is not a sufficient condition for the identification of a meaningful causal effect. The continuity assumption rules out coincidental functional discontinuities in the *x* – *y* relationship such as those caused by other programmes employing assignment mechanisms based on the exact same assignment variable and cut-off. In addition, the continuity restriction generally rules out certain types of behaviour both on the part of potential treatment recipients who exercise control over their value of *x* and programme administrators in choosing the assignment variable and cut-off point. Lee (2007) analyses the conditions under which an ability to manipulate the assignment variable may invalidate the RD identification

assumptions. He shows in the context of a sharp RD design that as long as individuals do not have *perfect* control over the position of the assignment variable relative to the cut-off score, the continuity assumption will be satisfied.

While in the sharp RD design treatment assignment is known to depend on the selection variable x in a deterministic way, in the case of a fuzzy design (Campbell, 1969), treatment assignment depends on x in a stochastic manner but in such a way that the propensity score function $\Pr(t=1|x)$ is again known to have a discontinuity at $\bar{x}$. Instead of a 0 – 1 step function, the selection probability as a function of x would now contain a jump smaller than 1 at $\bar{x}$. The fuzzy design can occur in case of misassignment relative to the cut-off value in a sharp design, with values of x near the cut-off appearing in both treatment and control groups. This situation is analogous to having no-shows (treatment group members who do not receive treatment) and/or crossovers (control group member who do receive the treatment) in a randomized experiment. This could occur if in addition to the position of the individual's score relative to the cut-off value, assignment is based on additional variables observed by the administrator, but unobserved by the evaluator.

A comparison of average outcomes of recipients and non-recipients, even if near the cut-off, would not generally lead to correct inferences regarding an average treatment effect. However, as shown by HTV, one can again exploit the discontinuity in the selection rule to identify a causal impact of interest by noting that under the LC assumption and with a locally constant treatment effect ($\alpha_i = \alpha$ in a neighbourhood around $\bar{x}$), the treatment effect α is identified by

$$\frac{\lim_{x \downarrow \bar{x}} E[y_i|x] - \lim_{x \uparrow \bar{x}} E[y_i|x]}{\lim_{x \downarrow \bar{x}} E[t_i|x] - \lim_{x \uparrow \bar{x}} E[t_i|x]}, \tag{3}$$

where the denominator is always non-zero because of the known discontinuity of $E[t|x]$ at $\bar{x}$.

In the case of varying treatment effects, HTV show that under the local continuity assumption, and a local conditional independence assumption requiring t_i to be independent of α_i conditional on x near $\bar{x}$, the ratio above identifies $E[\alpha_i|x = \bar{x}]$, the average treatment effect for cases with values of x close to $\bar{x}$. The conditional independence assumption is a strong assumption which may be violated if individuals self-select into or are selected for treatment on the basis of expected gains from treatment. HTV show that, under a weaker local monotonicity assumption similar to that assumed by Imbens and Angrist (1994), the ratio (3) will instead identify a local average treatment effect (LATE) at the cut-off point, which represents the average treatment effect of the ‘compliers’, that is, the subgroup of individuals whose treatment status would switch from non-recipient to recipient if their score x crossed the cut-off.

More recently Battistin and Rettore (2003) considered the special case where an eligibility rule divides the population into eligibles and non-eligibles according to a sharp RD design, and with eligible individuals self-selecting into treatment. In this case the LC assumption alone is sufficient for the ratio to identify $E[\alpha_i|t_i = 1, x = \bar{x}]$, the average treatment effect on the treated, for those near the cut-off.

As indicated by these identification results, estimation of treatment effects in an RD design involves

estimating boundary points of conditional expectation functions. The most common empirical strategy in the literature has been to adopt parametric specifications for the conditional expectations functions. Consider the following alternative representation of outcome eq. (1) in case of a sharp RD design:

$$y_i = m(x_i) + \delta t_i + e_i, \tag{4}$$

where $e_i = y_i - E[y_i|t_i, x_i]$, $t_i = 1\{x_i \geq \bar{x}\}$, $m(x) = E[u_i|x] + (E[\alpha_i|x] - E[\alpha_i|\bar{x}])1\{x \geq \bar{x}\}$. Then under the local continuity assumption $m(x)$ will be a continuous function of x at $\bar{x}$, and $\delta = E[\alpha_i|\bar{x}]$ (the average treatment effect at $\bar{x}$) will measure the discontinuity in the average outcome at the cut-off. This suggests that if the correct specification of $m(x)$ were known, and was included in the regression, we could consistently estimate the treatment effect for the sharp RD design. This idea of including a specification of $m(x)$ in the regression of y on t in order to correct for selection bias caused by selection on observables, is in the econometrics literature known as the *control function approach* (Heckman and Robb, 1985). A popular choice among empirical researchers has been to use global polynomials or to use splines (piecewise polynomials) which, even though globally continuous, have a knot at the cut-off (Trochim, 1984; van der Klaauw, 2002; McCrary, 2007).

In the case of a fuzzy RD design, when assuming local independence of t_i and α_i conditional on x, then in a neighbourhood of $\bar{x}$,

$$y_i = m(x_i) + \delta E[t_i|x_i] + w_i, \tag{5}$$

where $w_i = y_i - E[y_i|x_i]$ and $m(x) = E[u_i|x] + (E[\alpha_i|x] - E[\alpha_i|\bar{x}])E[t|x]$. With the local continuity assumption again implying that $m(x)$ will be continuous at the cut-off, and with $E[t_i|x_i]$ being discontinuous at $\bar{x}$, δ in this regression will measure the ratio in (3), which in this case equals the average local treatment effect $E[\alpha_i|\bar{x}]$. Similarly, δ can be interpreted as a local average treatment effect if we replaced the local independence assumption with the local monotonicity condition of Imbens and Angrist (1994).

This naturally leads to the two-stage procedure adopted by van der Klaauw (2002), where in the first stage we estimate the propensity score function specified as $t_i = E[t_i|x_i] + v_i = f(x_i) + \gamma 1\{x_i \geq \bar{x}\} + v_i$ where $f(\cdot)$ is continuous in x at $\bar{x}$ and γ measures the discontinuity in the propensity score function at $\bar{x}$. In the second stage the control function-augmented outcome equation is then estimated with t_i replaced by the first-stage estimate of $E[t_i|x_i] = \Pr[t_i = 1|x_i]$ as in Maddala and Lee (1976). With correctly specified $f(x)$ and $m(x)$ functions, this two-stage procedure yields a consistent estimate of the treatment effect. The approach is similar in spirit to those proposed earlier in the RD evaluation literature by

Spiegelman (1979) and Trochim and Spiegelman (1980). Note that in case of a parametric approach, if we assume the same functional form for $m(x)$ and $f(x)$, then the two-stage estimation procedure described here will be equivalent to two-stage least squares (in case of linear-in-parameter specifications) with $1\{x_i \geq \bar{x}\}$ and the terms in $m(x)$ serving as instruments. Because of the popularity of this particular parametrization, the RD approach is often interpreted as being equivalent to an instrumental variable approach, as it implicitly imposes an exclusion restriction by excluding $1\{x_i \geq \bar{x}\}$ as a variable in the outcome equation.

Valid parametric inference for the estimation of the treatment effect requires a correct specification of the control function $m(x)$ and of $f(x)$ in the treatment equation. To mitigate the potential for misspecification bias, several semiparametric estimation procedures have been proposed for estimating $m(x)$ and $f(x)$, or equivalently for estimating the limits $\lim_{x \downarrow \bar{x}} E[z|x]$ and $\lim_{x \uparrow \bar{x}} E[z|x]$ in (3) semiparametrically. These methods rely on less-restrictive smoothness conditions away from the discontinuity, with estimates based mainly on data in a neighbourhood on either side of the cut-off point. Asymptotically this neighbourhood needs to shrink, as with usual nonparametric estimation, implying that we should expect a slower than parametric rate of convergence in estimating treatment impacts. HTV considered the use of kernel and local linear regression estimators, while Porter (2003) proposed estimating the limits using local polynomial regression and partially linear model estimation. RD estimators based on local polynomial regression and partially linear model estimation have better boundary behaviour than the kernel-based estimator and as shown by Porter, achieve the optimal rate of convergence. This result is based on a known degree of smoothness of the conditional expectation functions. Sun (2005) proposed an adaptive estimator to first estimate the degree of smoothness in the data prior to implementing either estimator.

The internal validity of the RD approach relies on the local continuity of conditional expectations of potential outcomes around the discontinuity point. While this assumption is fundamentally untestable, a number of validity tests have been developed to bolster the credibility of the RD design. First, economic behaviour may lead to sorting of individuals around the cut-off point, where those below the cut-off may differ on average from those just above the cut-off. Such precise sorting around the cut-off would generally be accompanied by a discontinuous jump in the density of the assignment variable at the cutoff. Several approaches have been used for assessing this possibility (McCrary, 2007; Lee, 2007; Chen and van der Klaauw, 2007; Lemieux and Milligan, 2004). Second, one can test for evidence that individuals on either side of the cut-off are observationally similar by directly comparing average characteristics (McEwan and Urquiola, 2005) or by repeating the RD analysis treating the characteristics as outcome variables (van der Klaauw, 2007b). Alternatively, one can test for an imbalance of relevant characteristics by assessing the sensitivity of RD estimates to the inclusion of observed characteristics as controls (van der Klaauw, 2002; Lee, 2007). Third, in some applications data are available from a baseline period in which the programme did not yet exist, or for a group of individuals that was not eligible for treatment. In such a case the credibility of the design can be significantly enhanced by repeating the RD analysis with such data. Finding a zero treatment effect in such a falsification test would suggest that a non-zero post-programme effect was not an artifact of the specific RD model specification, estimation approach chosen or caused by another programme using the same cut-off and assignment variable.

Finally, while this exposition has focused on the binary treatment case with a selection rule containing a

single discontinuity at a known cut-off, the approach can be readily extended to one where there are multiple treatment dose levels and multiple cut-offs or ‘cut-off ranges’ within which the treatment dose varies continuously (van der Klaauw, 2007a). Similarly, the approach can be modified to cover cases where the assignment or selection variable is discrete instead of continuous (Lee and Card, 2006).

See Also

- causality in economics and econometrics
- natural experiments and quasi-natural experiments
- propensity score
- selection bias and self-selection
- semiparametric estimation
- treatment effect

Bibliography

Angrist, J.D. and Krueger, A.B. 1999. Empirical strategies in labor economics. In *Handbook of Labor Economics*, vol. 3, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Battistin, E. and Rettore, E. 2003. Another look at the regression discontinuity design. Working Paper No. CWP01/03, CeMMAP, Institute for Fiscal Studies.

Bloom, H.S., Kemple, J., Gamse, B. and Jacob, R. 2005. Using regression discontinuity analysis to measure the impacts of reading first. Paper presented at the Annual Conference of the American Educational Research Association, Montreal, Canada, April.

Campbell, D.T. 1969. Reforms as experiments. *American Psychologist* 24, 409–29.

Chen, S. and van der Klaauw, W. 2007. The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics*.

Cook, T.D. and Campbell, D.T. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton-Mifflin.

Hahn, J., Todd, P. and van der Klaauw, W. 1999. Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. Working Paper No. 7131. Cambridge, MA: NBER.

Hahn, J., Todd, P. and van der Klaauw, W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69, 201–09.

Heckman, J.J. and Robb, R. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman and B. Singer. New York: Cambridge University Press.

Imbens, G.W. and Angrist, J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 467–76.

Lee, D.S. 2007. Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*. (forthcoming).

Lee, D.S. and Card, D. 2006. Regression discontinuity inference with specification error. Technical Working Paper No. 322. Cambridge, MA: NBER.

Lemieux, T. and Milligan, K. 2004. Incentive effects of social assistance: a regression discontinuity approach. Working Paper No. 10541. Cambridge, MA: NBER.

Maddala, G.S. and Lee, L. 1976. Recursive models with qualitative endogenous variables. *Annals of Economic and Social Measurement* 5, 525–45.

McCrary, J. 2007. Testing for manipulation of the running variable in the regression discontinuity design. *Journal of Econometrics*.

McEwan, P.J. and Urquiola, M. 2005. Economic behavior and the regression-discontinuity design: evidence from class size reduction. Working paper, Columbia University.

Porter, J. 2003. Estimation in the regression discontinuity model. Unpublished manuscript, Harvard University.

Spiegelman, C.H. 1979. Estimating the effect of a large scale pretest posttest social program. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 370–3.

Sun, Y. 2005. Adaptive estimation of the regression discontinuity model. Working paper, University of California, San Diego.

Thistlethwaite, D. and Campbell, D. 1960. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *Journal of Educational Psychology* 51, 309–17.

Trochim, W.K. 1984. *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Beverly Hills, CA: Sage.

Trochim, W. and Spiegelman, C.H. 1980. The relative assignment variable approach to selection bias in

pretest-posttest group designs. *Proceedings of the Survey Research Section, American Statistical Association*, pp. 376–80.

van der Klaauw, W. 2002. Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. *International Economic Review* 43, 1249–87.

van der Klaauw, W. 2007a. Regression-discontinuity analysis: a survey of recent developments in economics. Unpublished manuscript, Federal Reserve Bank of New York.

van der Klaauw, W. 2007b. Breaking the link between poverty and low student achievement: an evaluation of title I. In *Journal of Econometrics*.

How to cite this article

van der Klaauw, Wilbert. "regression-discontinuity analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000243> doi:10.1057/9780230226203.1414

The New Palgrave Dictionary of Economics Online

regular economies

Egbert Dierker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

General equilibrium theory describes those states of an economy in which the individual plans of many agents with partially conflicting interests are compatible with each other. Such a state is called an equilibrium. The concept of an equilibrium simply being based on a consistency requirement lends itself to the study of specific questions of quite different character. Indeed, equilibrium theory provides a unifying framework for the analysis of questions arising in various branches of economic theory. In our opinion it is fruitful to view equilibrium theory as a method of thinking applicable to a variety of problems of different origin.

Ideally one would like to have general principles which ensure that equilibria exist, that they are unique, and that, therefore, the equilibria resulting from different policy measures can unequivocally be compared. Moreover, one would like to know whether equilibria have some desirable properties when no single agent can exert an essential influence on the global outcome to his personal advantage. These welfare questions are particularly interesting because the concept of an equilibrium itself is not based on the well-being of the economic agents. Finally, although the concept of an equilibrium as described above is static in nature, one would like to have a dynamic theory according to which some equilibrium is approached in the course of time.

These and related questions such as the computability of equilibria have been studied in the past with different degrees of success. There are general principles which yield the existence of an equilibrium in an astonishingly large variety of cases. Furthermore, the welfare properties of equilibria are well understood. However, it is easy to construct examples of economies with an infinite number of equilibria and it appears to be very difficult to provide conditions which lead, without being artificial or and hoc, to the uniqueness of an equilibrium. As a consequence, comparative statics does not have a basis which makes it generally a well-defined problem. Also, the difficulties encountered when studying the uniqueness issue present severe obstacles for the development of a dynamic theory.

The theory of regular economies may be viewed as an effort to advance general equilibrium theory in the absence of a satisfactory uniqueness result. The seminal paper is Debreu (1970). Debreu explicitly allows for the multiplicity of equilibria. However, he requires each equilibrium to be *locally* unique. Each equilibrium is well determined and robust in the sense that it is not destroyed by a small change in the parameters.

A regular economy is an economy with a certain, finite number of equilibria, all of which respond continuously to small parameter changes. Hence each of these equilibria can be traced for some while during a parameter change. Thus there is a basis for doing comparative statics locally, that is to say as long as the equilibrium under consideration stays robust. If, at a certain point, it ceases to be robust, a drastic change is to be expected, the size and direction of which are probably hardly predictable. The focus of the theory of regular equilibria is more on the continuous behaviour of robust equilibria than on drastic changes.

It is most remarkable that Debreu (1970), by using concepts and techniques developed in the mathematical field of differential topology, has introduced a new kind of thought into economic analysis. In the meantime this way of thinking has penetrated many areas of economic theory at different levels. One of the first applications has occurred in the technically advanced area of core theory, where the continuous dependence of the set of price equilibria on the characteristics of the agents, which is guaranteed in a regular economy, plays an important role. An application on a purely conceptual level in oligopoly theory is incorporated in the notion of a demand function which an oligopolist faces in the Cournot–Nash context. The graph of this function is considered as given by the equilibria of an exchange economy with initial endowments as varying parameters.

The dependence of the equilibria on initial endowments will be discussed in detail in the next section because this case is particularly suited to illustrate basic ideas of the theory of regular economies.

Debreu's theorem on regular equilibria

The purpose of this section is to describe the kind of reasoning typical for the theory of regular economies in a prototypical situation. It is desirable to deal with parameter variations taking place in some Euclidean space because the mathematical structures to be used are most familiar in this case. We shall study exchange economies which differ by the allocation of initial endowments.

There are *l* commodities and *m* consumers. Individual initial endowments are supposed to be positive in each component. If we denote the strictly positive orthant in $\mathbb{R}^l$ by *P*, then an initial allocation is a vector $(e_1, \ldots, e_m) \in P^m$. Since the demand function f_i of each consumer *i* is considered as fixed, an economy *E* is fully specified by $(e_1, \ldots, e_m)$. The space of all economies under consideration can thus be identified with P^m, an extremely simple subset of a Euclidean space. We want to examine how the exchange equilibria of an economy–there may be several such equilibria–depend on the particular economy $E \in P^m$.

We assume that all goods are desired so that attention may be restricted to strictly positive relative prices. Price systems are normalized; to be specific we consider price systems in

$$S = \left\{ p = (p_1 \ldots p_2) \square 0 \| p \| = \left(\sum_{k-1}^{l} p_k^2 \right)^{1/2} = 1 \right\}.$$

If consumer *i* initially possesses the commodity bundle e_i, his demand at the price system *p* is $f_i(p, p \cdot e_i) \in \mathbb{R}^l_+$, where $p \cdot e_i = w_i > 0$ is *i*'s wealth. Hence the aggregate excess demand of the economy *E*, given by the initial allocation $(e_1, \ldots, e_m) \in P^m$, at *p* is

$$Z_E(p) = \sum_{i=1}^{m} [f_i(p, p \cdot e_i) - e_i].$$

We assume Walras's Law which states that the value $p \cdot Z_E(p)$ of the excess demand is identically equal to zero. Furthermore, every f_i is supposed to be continuous.

The desirability of all commodities will be captured in the following condition, which is always satisfied when consumers have strictly monotone preferences.

(D) If the price of at least one good approaches zero and the wealth $w_i > 0$ of every agent stays away from zero, then

$$\sum_{i=1}^{m} \| f_i(p, w_i) \|$$

tends to infinity.

An *equilibrium price system* of *E* is a price system $p \in S$ at which the consumption plans $f_i(p, p \cdot e_i)$ of all agents are consistent, i.e. a zero of the excess demand function Z_E. It is not difficult to show the following consequence of the desirability assumption (D) by a fixed point argument:

Every economy $E \in P^m$ *has at least one equilibrium if* (D) *holds*.

We would like to know how the equilibrium prices vary when the initial allocation is modified. Therefore we look at the graph Γ of the correspondence ('multi-valued function') Π which assigns to every economy $E \in P^m$ its set $\{ p \in S + Z_E(p) = 0 \}$ of equilibrium price systems. Defining $Z: P^m \times S \to \mathbb{R}^l$ by $Z(E, p) = Z_E(p)$ we get

$$\text{graph } (\Pi) = \Gamma = Z^{-1}(0).$$

Since Z is a continuous function, Γ is a closed set. It is well known that, in the case of a (single-valued) function, the closedness of the graph is intimately related to the continuity of the function. Here, where Π is a correspondence rather than a function, we obtain the following continuity result: *the graph* Γ *of the equilibrium price correspondence* Π *is upper hemi-continuous and compact-valued, if* (D) *holds*.

This is tantamount to the following explicit statement. If (E_n) is a sequence of economies in P^m converging to $E \in P^m$ and if $p_n \in \Pi(E_n)$ is an equilibrium price system to E_n for all n, then the sequence (p_n) has a subsequence which converges to an equilibrium price system of the limit economy E, provided (D) holds.

To improve our understanding of Γ, we assume that the demand functions f_i are continuously differentiable (C^1 for short) and we invoke the implicit function theorem in the following manner. Walras's Law allows us to disregard one market, say the lth, and to concentrate on

$$\hat{Z}: P^m \times S \to \mathbb{R}^{l-1}$$

which is obtained from Z by deleting the last component. Let p be an equilibrium price system of E, i.e. $\hat{Z}(E, p) = 0$. A simple calculation yields that the derivative $d\hat{Z}(E, p)$ has maximal rank at (E,p). Therefore, the graph Γ is given by a smooth surface of dimension lm. That is to say each point in Γ has a neighbourhood in Γ which can be mapped onto an open subset of $\mathbb{R}^{lm}$ by a C^1 diffeomorphism, i.e. a C^1 map with a C^1 inverse. Such a locally Euclidean space is called a C^1 manifold: see Figure 1.

Figure 1

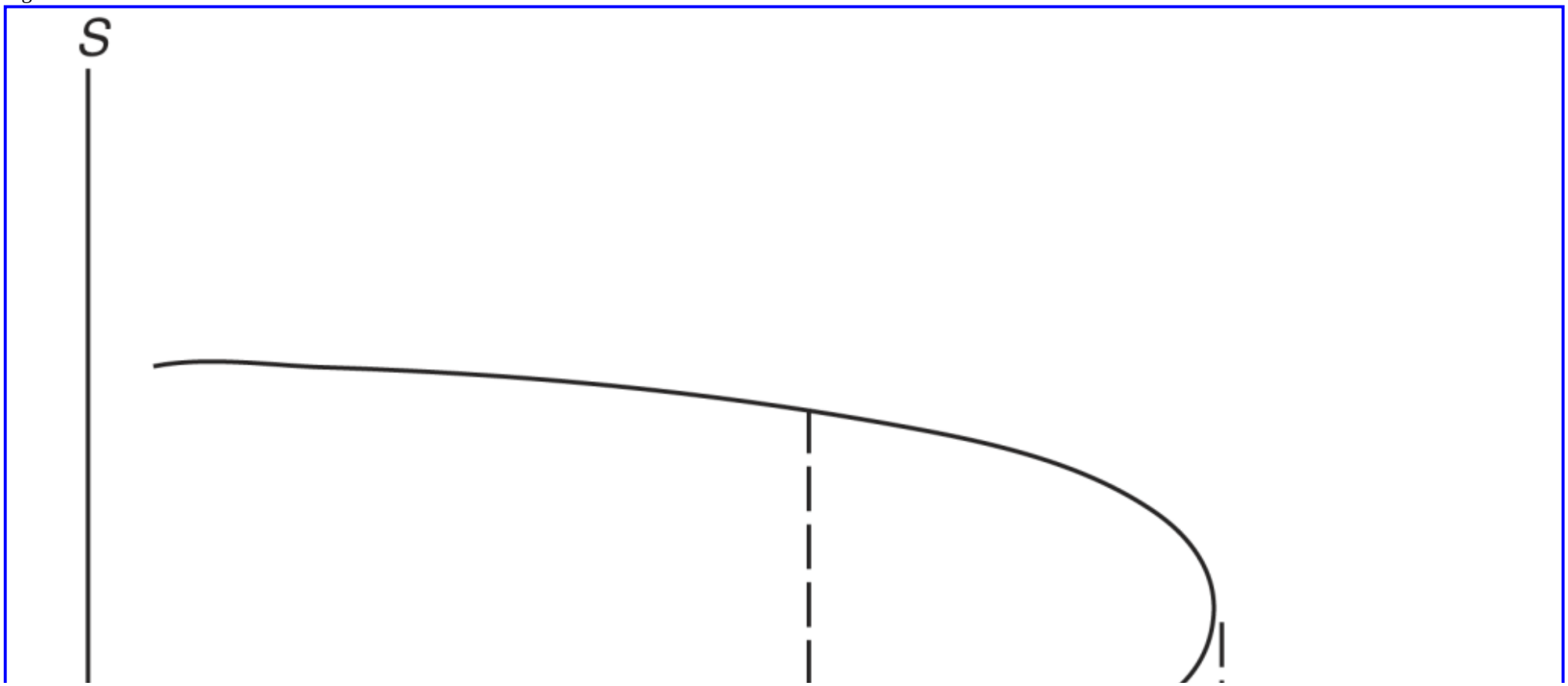

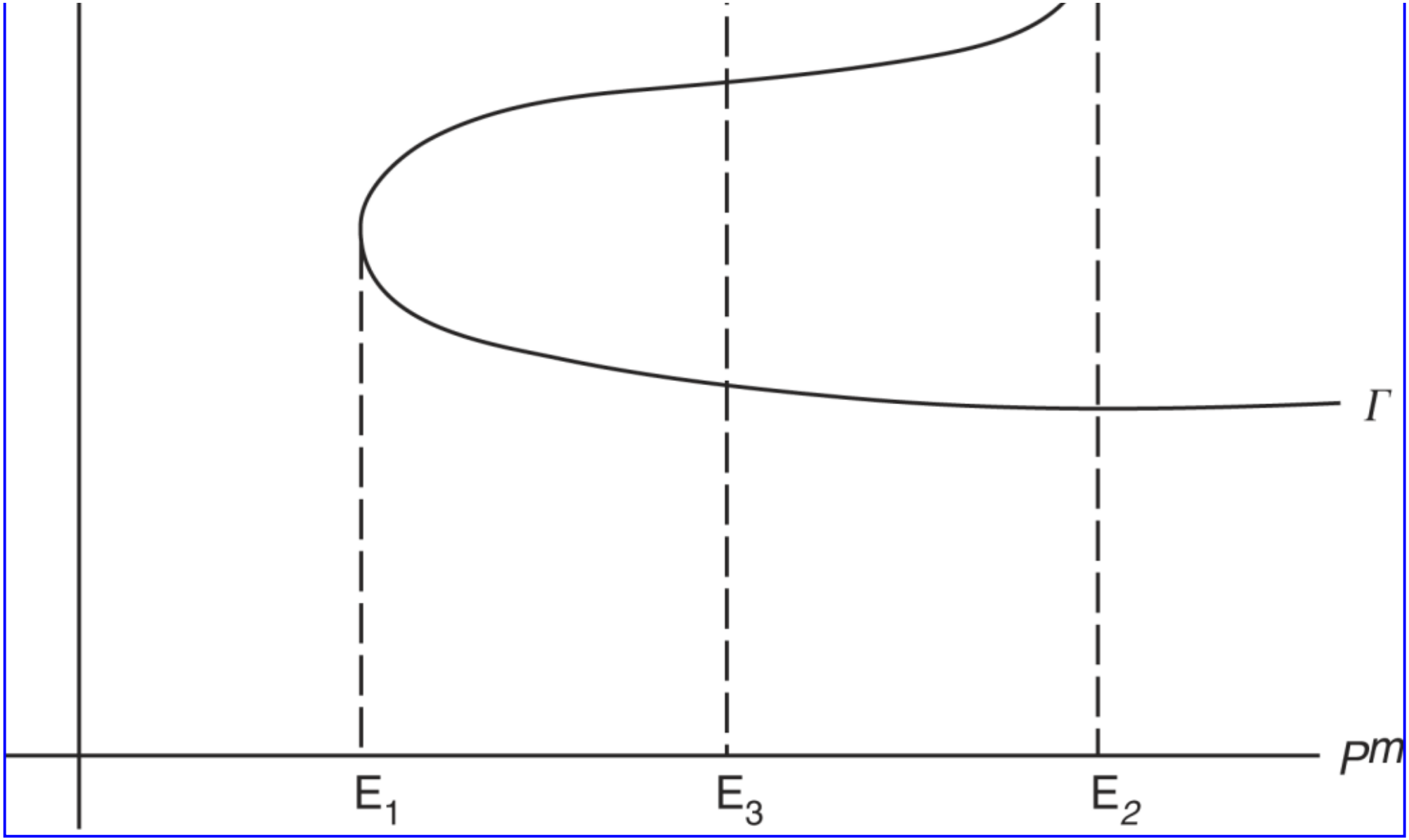

We have seen that *the graph* Γ *of the equilibrium price correspondence* Π *is a* C^1 *manifold*, but Figure 1 suggests more. In Figure 1, Γ is not only locally Euclidean, there is even a global diffeomorphism between Γ and $\mathbb{R}^{lm}$. Indeed, one can show that this global equivalence holds (see Balasko, 1975).

The equilibrium price correspondence is continuous except at two points, E_1 and E_2. If a parameter variation leads through E_1 (or E_2) the equilibrium may be forced to jump. The equilibrium reached after the jump, however, is robust in the sense that no sudden change must occur when one passes through E_1 (or E_2) again. One can imagine a situation such as in E_1 takes place when a slight reduction in the supply of an important raw material leads to a drastic increase in its price. If later on the supply begins to increase again prices perhaps vary but stay at their high level. A reversion of this phenomenon may occur when the supply has reached the much higher level corresponding to E_2.

In Figure 2 we have drawn a two-dimensional parameter space. The following remarkable phenomenon may happen here.

Figure 2

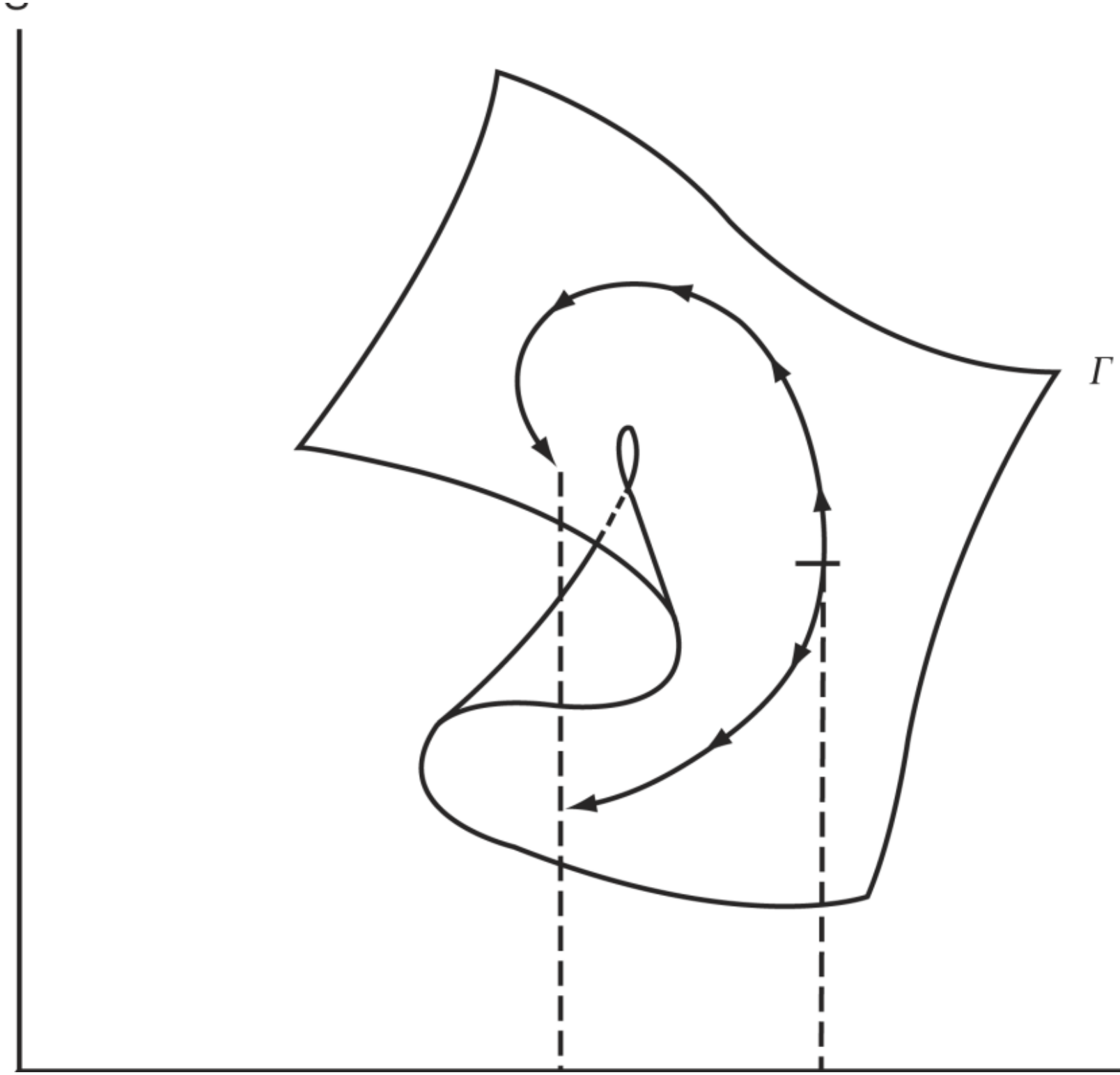
Γ

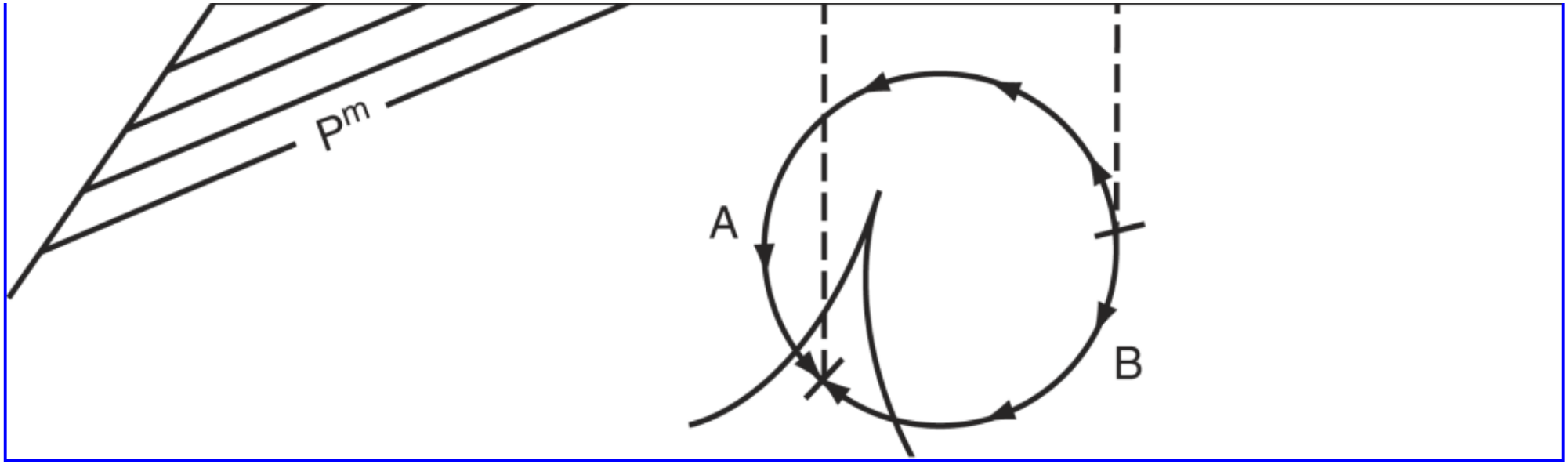

There are two paths, A and B, in the parameter space which have their starting point and endpoint in common. Following either path there is no need for the equilibrium to jump. However, the two equilibria reached at the end are quite different equilibria of the same economy. In other words, if two or more policy variables are at one's disposal one must be aware of the possibility that the final outcome depends very well on the order in which the variables are utilized.

The economies E_1 and E_2 in Figure 1 are characterized by the fact that the graph Γ has a vertical tangent above E_1 and above E_2. Similarly, in Figure 2, Γ has vertical tangents above all points on the cusp drawn in the bottom plane, which represents P^m. Apparently qualitative changes of the equilibrium price set at an economy E are associated with vertical tangents of Γ above E. This motivates the following definitions. A *critical point* of the projection pr: $\Gamma \to P^m$ is a point in Γ at which the derivative of pr has rank less than dim $P^m = lm$. A *critical value* of pr: $\Gamma \to P^m$ is the image of a critical point. A *regular value* of pr: $\Gamma \to P^m$ is a point in P^m which is not a critical value. Figures 1 and 2 suggest that almost all points in P^m are regular values. Indeed, the concepts introduced above are defined in differential topology in a quite general context and Sard's theorem, an analytical tool of great importance, asserts that the critical values of a sufficiently differentiable mapping are rare. More precisely, *Sard's theorem* applied to our particular problem yields that the set of critical values of pr: $\Gamma \to P^m$ is a (Lebesgue) null set.

Null sets are small in a probabilistic sense. At this point we make essential use of the space of economies P^m being part of a Euclidean space. If, for instance, consumers' demand functions or preferences are allowed to vary instead of consumers' endowments, it is not clear how null sets are to be defined. However, one can express quite easily when two demand functions or preference orderings are close to each other. That is to say metric structures are very often naturally given when there is no obvious way to define null sets. A set can then be defined to be small in a topological sense if its closure is nowhere dense.

Furthermore, if the concepts of smallness in the probabilistic and in the topological sense are both well-defined, as they are in the case of variable initial endowments, one has to be aware of the fact that the two variants of the intuitive notion of smallness apply to quite different sets. Defining a *critical economy* $E \in P^m$ as a critical value of pr: $\Gamma \to P^m$ and *regular economy* as a regular value of pr we ask ourselves whether the null set of critical economies has a null closure. We know already that the desirability assumption (D) implies that the equilibrium price correspondence is upper hemi-continuous and compact-valued or, in more intuitive terms, that Γ has only finitely many layers above some compact ball B of economies in P^m. Hence the points in Γ which lie above B and have a vertical tangent form a compact set. Projecting this set down to B yields a compact set, the set of critical economies in B. Since this set is also null by Sard's theorem, it is nowhere dense. We obtain:

The set of critical economies in P^m is a closed null set if (D) holds.

Let $E \in P^m$ be a regular economy. Then E has a finite number of equilibria and this number is locally constant. If E has r equilibria, then there is a neighbourhood U of E and there are rC^1 functions $g_1, \ldots, g_r$ such that the set $\Pi(E')$ of equilibrium price systems of any economy $E' \in U$ is given by $\{g_1(E'), \ldots, g_r(E')\}$. In particular, *the equilibrium price correspondence Π is continuous in a neighbourhood of a regular economy*.

These results, with minor differences, have been obtained by G. Debreu (1970), whose proof, however, differs from the exposition given here.

Extensions

When one wants to extend the theory of regular equilibria to more general spaces of economies, it is often useful to employ a definition of a regular economy which focuses on the given economy and does not refer to the graph Γ or to the parameter space. To motivate the following definition we contrast Figure 3, in which the excess demand of a critical economy such as E_1 or E_2 in Figure 1 is drawn, with Figure 4, which shows the graph of a regular economy such as E_3. It is assumed that there are two goods so that it suffices, according to Walras's Law, to look at the excess demand ζ_1 for good 1.

Figure 3

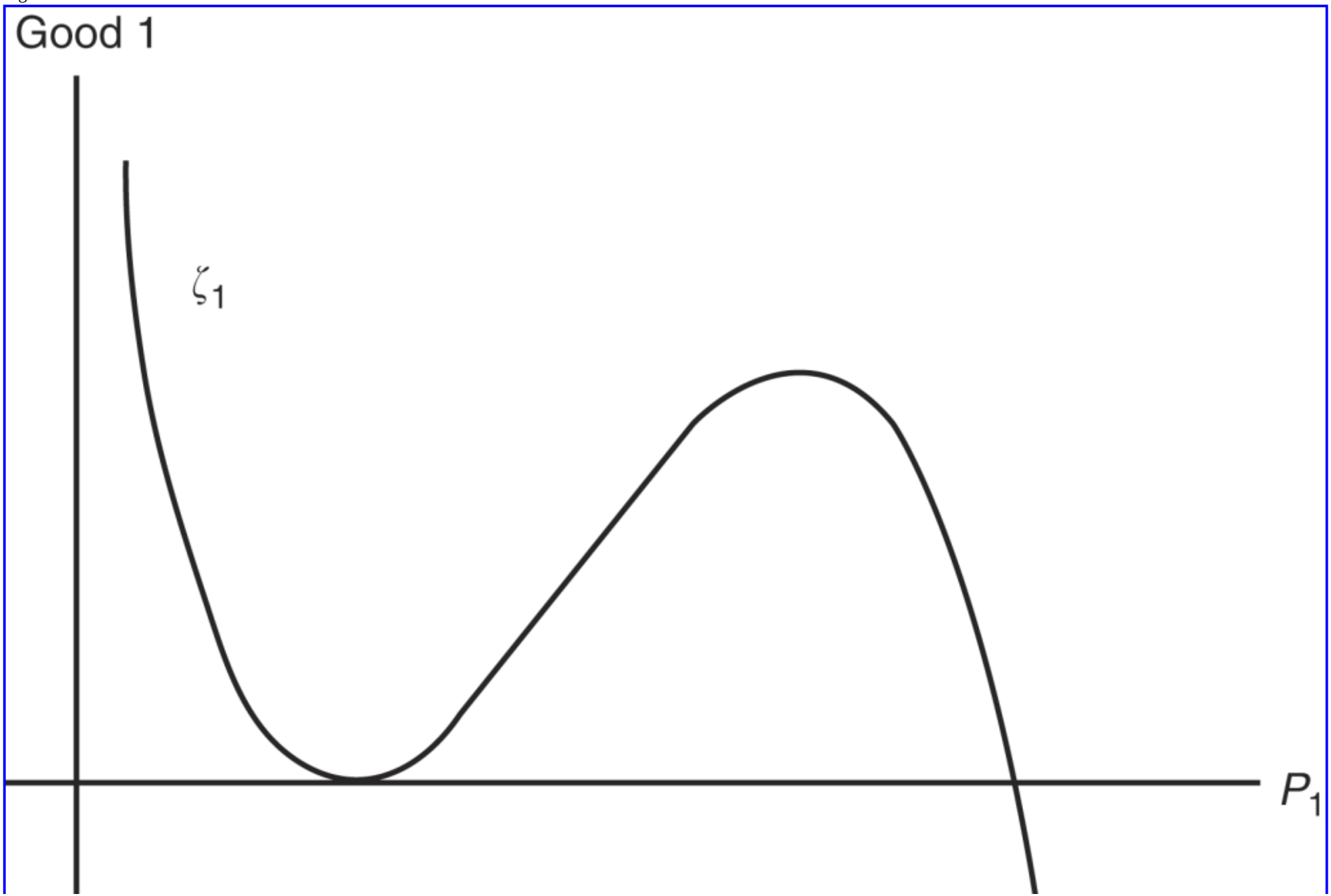

Figure 4

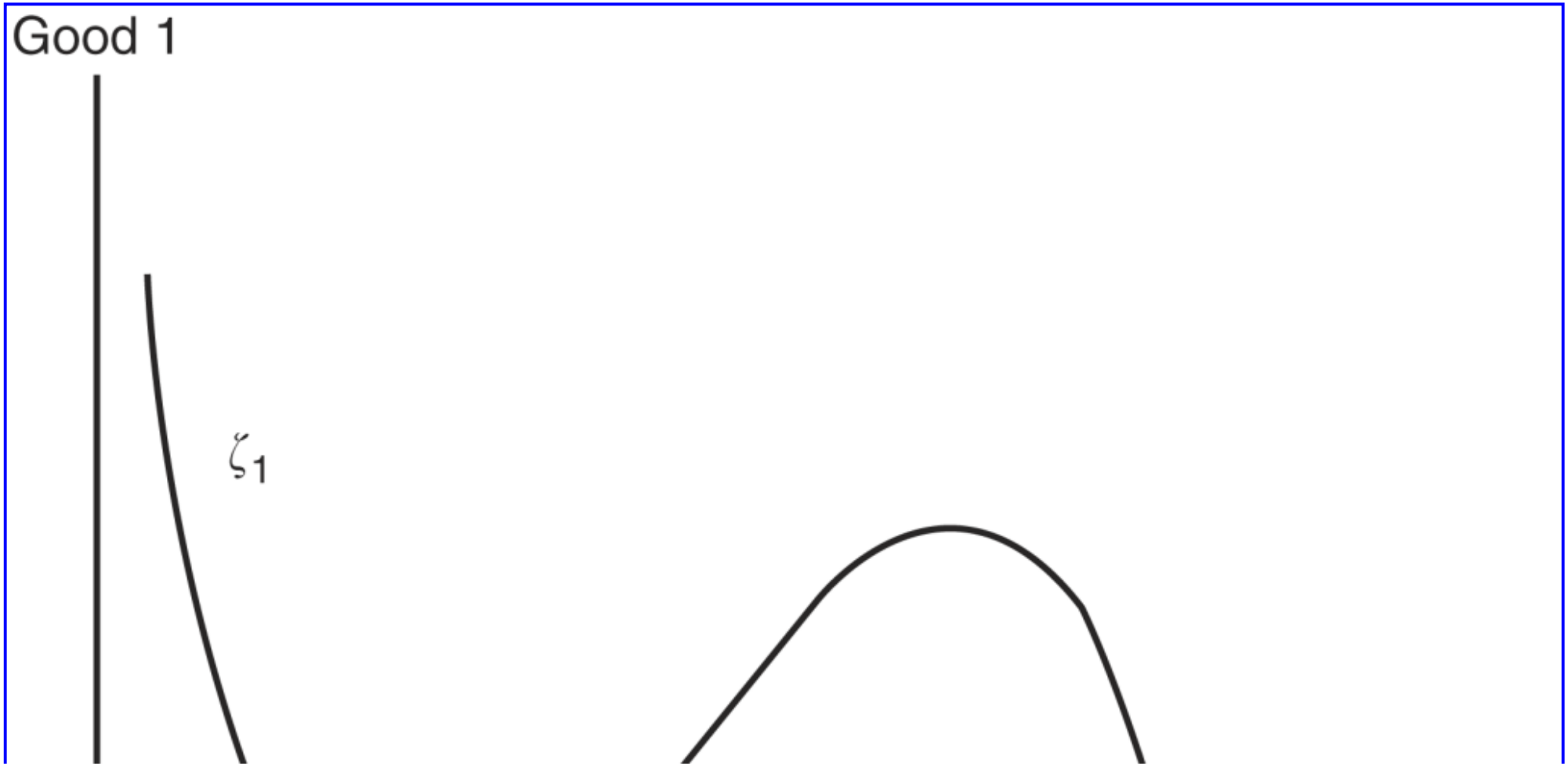

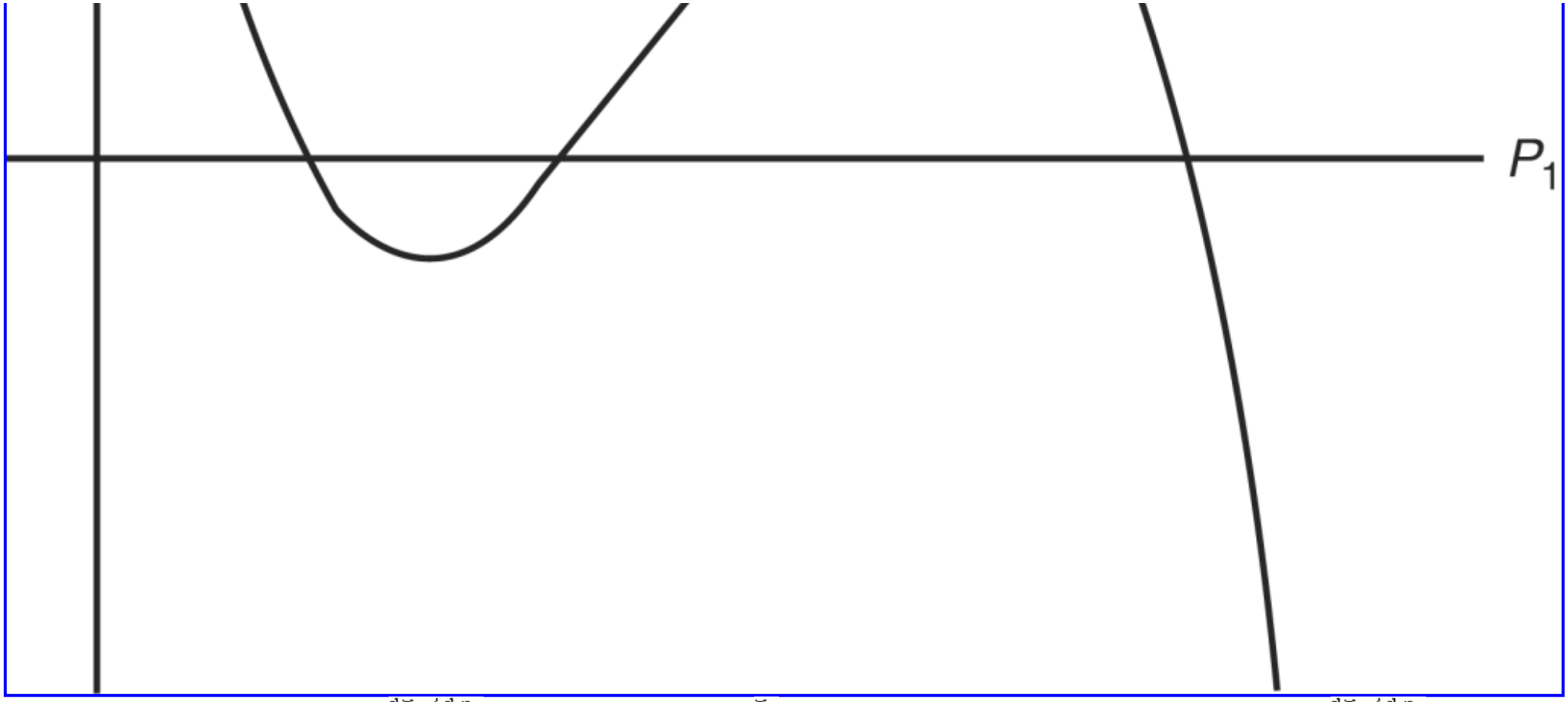

In Figure 3 there is one equilibrium at which $d\zeta_1/dp_1$ vanishes. Shifting the graph of ζ_1 a little upwards destroys this equilibrium. In Figure 4, however, $d\zeta_1/dp_1$ does not vanish at any equilibrium and all equilibria are robust.

Let the excess demand function $\zeta: S \to \mathbb{R}^l$ of an economy *E* be C^1. A price system $p \in S$ is called a *regular equilibrium price system* if $\zeta(p) = 0$ and the matrix

$$\left[\frac{\partial \zeta_h}{\partial p_k}(p)\right]_{h,k=1,\ldots,l-1}$$

is regular. A *regular economy* is an economy all equilibrium price systems of which are regular. One can show that this definition, introduced by E. and H. Dierker (1972), is independent of the way in which goods are indexed and that it is consistent with the definition given above.

The results on regular economies obtained in various frameworks are quite similar to those established in the previous section. It is shown that almost all economies, in an appropriate sense, are regular. Every regular equilibrium is locally unique and can be traced along its path when the economy varies gradually, as long as it stays regular. Economic models in which results of this kind have been precisely formulated and verified deal with variations in consumption and production (see, in particular, Smale, 1974). Also the case of many consumers, that is to say of consumption sectors described by the distribution of consumers' characteristics, has been treated. The basic mathematical tool is always some variant of Sard's Theorem. References can be found in my survey article (Dierker, 1982).

The study of regular equilibria has led to a revival of the differentiable viewpoint in general equilibrium theory and related areas. Readers interested in this modern development are referred to the excellent book by A. Mas-Colell (1985), which also contains an extensive, systematic presentation of the theory of regular equilibria.

See Also

- existence of general equilibrium
- global analysis in economic theory
- mathematical economics

Bibliography

Balasko, Y. 1975. Some results on uniqueness and on stability of equilibrium in general equilibrium theory. *Journal of Mathematical Economics* 2, 95–118.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

Dierker, E. 1982. Regular economies. In *Handbook of Mathematical Economics*, ed. K. Arrow and M. Intriligator. Amsterdam: North-Holland, ch. 17, 795–830.

Dierker, E. and Dierker, H. 1972. The local uniqueness of equilibria. *Econometrica* 40, 867–81.

Mas-Colell, A. 1985. *The Theory of General Economic Equilibrium: A Differentiable Approach*. Econometric Society Monographs, Cambridge: Cambridge University Press.

Smale, S. 1974. Global analysis and economics IV: finiteness and stability of equilibria with general consumption sets and production. *Journal of Mathematical Economics* 1, 119–27.

How to cite this article

Dierker, Egbert. "regular economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000084> doi:10.1057/9780230226203.1415

The New Palgrave Dictionary of Economics Online

regulation

Robert Boyer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Régulation’ theory analyses the long-term transformation in capitalist economies and their consequences for growth patterns and cyclical adjustments. The degree of coherence of a specific configuration of the major institutional forms – monetary regime, wage-labour relation, form of competition, state–citizen institutionalized compromise and mode of support of the international regime – defines various accumulation regimes and ‘régulation’ modes. Over one century, several regimes have been observed along with a succession of changing patterns for the related structural crises. The demise of the post-Second World War Fordist regime has been associated with an uncertain process of institutional restructuring and the coexistence of various brands of capitalism.

Keywords

adjustment; accumulation regimes; capitalism; coevolution; complementarity; credit; diffusion of technology; fine tuning; Fordism; hierarchy; historicism; information technology; institutionalism; isomorphism; Kalecki, M.; Keynesianism; Marxism; mass production; *régulation* ; stabilization policies; stagflation; welfare state

Article

Since the 1980s, the term ‘*régulation*’ has suggested state intervention in the name of economic management though its opposite, ‘dérégulation’, has been more widely used. In the area of economic policy and in accordance with Keynesian precepts, regulation indicates the adjustment of macroeconomic activity by means of budgetary or monetary contra-cyclical interventions. In the area of public management, a complete body of literature, under the name of regulation theory, has investigated the methods for organizing the decentralization of the supply of various public utilities.
This term is also used in physics and biology. In mechanics, a regulator is a means to stabilize the rotary speed of a machine. In biology, regulation corresponds to the reproduction of substances such as DNA.

In general terms, the theory of systems involves the study of the role of a set of negative and positive feedback loops in relation to the stability of a complex network of interactions.

Here, a third and different, but not totally unrelated, meaning of the term will be developed. Theories of régulation constitute an area of research which has focused on analysing long-term transformations in capitalist economies. Initially, it focused on American and French capitalisms (Aglietta, 1982; Benassy, Boyer and Gelpi, 1979) but it was progressively extended first to major OECD economies (Mazier, Basle and Vidal, 1999) then to Latin American countries (Hausmann, 1981; Ominami, 1985) and ultimately Asian countries (Bertoldi, 1989; Boyer, 1994). A general presentation of the present state of the theory is to be found in Boyer and Saillard (2002) and a large sample of national case studies in Jessop (2001). Basically, the theory of régulation combines Marxian intuitions and Kaleckian macroeconomics with institutionalist and historicist studies, mobilizing most of the tools of modern economic analysis.

At a primary level, a form of régulation denotes any dynamic process of adaptation of production and social demand resulting from a conjunction of economic adjustments linked to a given configuration of social relations, forms of organization and productive structures (Boyer, 1990).

Most economic theories emphasize the general invariables of eminently abstract systems, in which history serves merely as a confirmation or, failing that, as a perturbation. Neoclassical theory studies the shift of a stable equilibrium after an external shock, Keynesian economists stress the role of effective demand and fine tuning whatever the context and the period. Even Marxists tend to extrapolate, as general laws, the quite specific evolutions observed in the early phases of capitalism. In contrast, the régulation approach seeks a broader interaction between history and theory, social structures, institutions and economic regularities (de Vroey, 1984).

The starting point is the hypothesis that accumulation has a central role and is the driving force of capitalist societies. This necessitates a clarification of factors that reduce or delay the conflicts and disequilibria inherent in the formation of capital, and which allow for an understanding of the possibility of periods of sustained growth (Boyer and Mistral, 1978). These factors are associated with particular regimes of accumulation, namely, the form of articulation between the dynamics of the productive system and social demand, the distribution of income between wages and profits on the one hand, and on the other hand the division between consumption and investment. It is then useful to explain the organizational principles which allow for mediation between such contradictions as the extension of productive capacity under the stimulus of competition on product markets, for labour and finance. The notion of institutional form – defined as a set of fundamental social relations (Aglietta, 1982) – enables the transition between constraints associated with an accumulation regime and collective strategies; between economic dynamics and individual behaviour. A small number of key institutional forms, which are the result of past social struggles and the imperatives of the material reproduction of society, frame and channel a multitude of partial strategies which are decentralized and limited in terms of their temporal horizon. Five main institutional forms do shape accumulation regimes.

The forms of competition describe by what mechanisms the compatibility of a set of decentralized decisions by firms and individuals is ensured. They are competitive while the *ex post* adjustment of prices and quantities ensures a balance; they are monopolist if the *ex ante* socialization of revenue is such that production and social demand evolve together (Lipietz, 1979). The type of monetary constraint explains the interrelations between credit and money creation: credit is narrowly limited in terms of

movement of reserves when money is predominantly metallic; the causality is reversed when on the contrary the dynamics of credit conditions the money supply in systems where the external parity represents the only constraint weighing upon the national monetary system (Benassy, Boyer and Gelpi, 1979). The nature of institutionalized compromises defines different configurations of relations between the state and the economy (André and Delorme, 1983; Jessop, 1990): the state-as-referee when only general conditions of commercial exchange are guaranteed; as the interfering state when a network of régulation and budgetary interventions codifies the rights of different social groups. Modes of support for the international regime are also derived from a set of rules which organize relations between the nation state and the rest of the world in terms of commodity exchange, migration, capital movements and monetary settlements. History goes beyond the traditional contrast between an open and a closed economy, free trade and protectionism; it makes apparent a variety of configurations (Mistral, 1986; Lipietz, 1986a). Finally, forms of wage relations indicate different historical configurations of the relationship between capital and labour, that is, the organization of the means of production, the nature of the social division of labour and work techniques, type of employment and the system of determination of wages, and finally, workers' way of life including the welfare state. If, in the first stages of industrialization, wage-earners are defined first of all as producers, during the second stage they are simultaneously producers and consumers.

At this point appears the notion of régulation, as a conjunction of mechanisms and principles of adjustment associated with a configuration of wage relations, competition, state interventions and hierarchization of the international economy. Finally, a distinction between ‘small’ and ‘big’ crises is called for (Boyer, 1990). The former, which are of a rather cyclical nature, are the very expression of régulation in reaction to the recurrent imbalances of accumulation. The latter are of a structural nature: the very process of accumulation throws into doubt the stability of institutional forms and the régulation which sustains it because the profit does not recover by contrast with conventional business cycles.

Thus, in long-term dynamics as well as in short-term development, institutions are important. Historical research confirms that sometimes institutional forms make an impression on the system in operation; at other times they register major changes in direction. At the end of a period which can be counted in decades, the very mode of development – that is, the conjunction of the mode of régulation and the accumulation regime – is affected: there will be changes in the tendencies of long-term growth and eventually in inflation, specificities of cyclical processes (Mazier, Basle and Vidal, 1999).

So a periodization of advanced capitalist economies emerges which is not part of the traditional Marxist theory. Despite the rise in monopoly, the interwar period is still marked by competitive regulation. After the Second World War an accumulation regime without precedent is instituted – that of intensive accumulation centered on mass consumption (Bertrand, 1983) – known as Fordist and channelled through monopolist-type regulation.

In fact, the alteration in wage relations – in particular the transition to Fordism, that is, the synchronization of mass production and wage-earners' access to the ‘American way of life’ – and in monetary management, that is, transition to internally accepted credit money – seems to have played a greater role than the change in modes of competition or conjunctural fine tuning à la Keynes (Aglietta, 1982; Aglietta and Orlean, 1982; Boyer, 1988).

Since the 1960s, many economies have been experiencing a big crisis without historical precedent: stagflation, absence of cumulative depression, breaking-down of most previous economic regularities,

length of the period of technological and institutional restructuring (Boyer and Mistral, 1978; Lipietz, 1985). In consequence, it is logical that former economic policies lose their efficacy (Boyer, 1990). First, because the crisis is not cyclical but structural; this invalidates the policy of fine-tuning; second, because the structural changes which permitted the 1929 crisis to be overcome have become blocked and cannot be repeated (Lipietz, 1986b).

Since the formative years, the research programme has been developing both extensively and intensively. The collapse of the Soviet bloc economies has pointed to the need to investigate the necessary and sufficient institutions required for a viable capitalist economy (Emergo, 1995; Hollingsworth and Boyer, 1997): economic viability depends on the compatibility of a complete set of institutional forms. In the epoch of financialization (Aglietta, 1998), information and communication technologies diffusion (Boyer, 2004), rise of services (Petit, 1986) and strengthening of foreign competition (Lipietz, 1986a), no clear follower to Fordism has yet emerged and diffused. Nevertheless, since the mid-1970s a series of trials and errors concerning the reform of the monetary regime, the tax and welfare system, competition and wage relations has finally delineated a new institutional architecture, quite complex to analyse. Conversion, layering and recomposition of existing institutional forms have replaced the strong synchronization associated with major crises and world wars (Boyer, 2005b).

The large number of international comparisons has systematically exhibited the persisting diversity of various brands of capitalism. Within industrialized countries: market dominated, corporate-led, state governed and social democratic versions, with some possible sub-variants, coexist (Amable, 2004). An equivalent but different variety is observed for Latin American countries (Quémia, 2001). Consequently, the financial crises experienced by Mexico, Brazil and Argentina are quite different, even if they all point out the destabilizing role of global finance upon contrasted domestic accumulation regimes (Boyer and Neffa, 2004).

These numerous structural changes call for new directions for the research agenda of régulation theory. Can the concepts of complementarity, hierarchy, isomorphism and coevolution explain how various mixes of institutions can cohere and define a coherent accumulation regime (Boyer, 2005a; Socio-Economic Review, 2005)? What kind of political economy analysis can explain the emergence and restructuring of institutional forms, especially the choice of monetary regime, the configuration of the welfare state or the nature of insertion into the world economy (Palombarini, 1999)? How to analyse multilevel régulation modes, especially in order to understand the complex process of European integration (Boyer and Dehove, 2001)? Finally, is not the anthropogenetic model, based on the production of humankind by education, health care and culture, a possible follower of the Fordist regime (Boyer, 2004)?

See Also

- Beveridge, William Henry
- Braudel, Fernand
- capitalism

- collective bargaining
- competition
- contemporary capitalism
- development economics
- growth and institutions
- institutional economics
- Kalecki, Michal
- Marx, Karl Heinrich
- profit and profit theory

Bibliography

Aglietta, M. 1982. *Regulation and Crisis of Capitalism*. New York: Monthly Review Press.

Aglietta, M. 1998. Le capitalisme de demain. Note de la fondation Saint-Simon (101), November.

Aglietta, M. and Orlean, A. 1982. *La violence de la monnaie*. Paris: PUF.

Amable, B. 2004. *The Diversity of Modern Capitalisms*. Oxford: Oxford University Press.

André, Ch. and Delorme, R. 1983. *L'état et l'économie*. Paris: Seuil.

Benassy, J.P., Boyer, R. and Gelpi, R.M. 1979. Régulation des économies capitalistes et inflation. *Revue économique* 30(3), 397–441.

Bertoldi, M. 1989. The growth of Taiwanese economy 1949–1989: success and open problems of a model of growth. *Review of Currency Law and International Economics* 39, 245–88.

Bertrand, H. 1983. Accumulation, régulation, crise: un modèle sectionnel théorique et appliqué. *Revue économique* 34, 305–43.

Boyer, R. 1979. Wage formation in historical perspective: the French experience. *Cambridge Journal of Economics* 3, 99–118.

Boyer, R. 1988. Technical change and the theory of 'regulation'. In *Technical Change and Economic Theory: The Global Process of Development*, ed. G. Dosi et al. London: Pinter Publishers.

Boyer, R. 1990. *The Regulation School*. New York: Columbia University Press.

Boyer, R. 1994. Do labour institutions matter for economic development? In *Workers, Institutions and*

Economic Growth in Asia, ed. G. Rodgers. Geneva: ILO/ILLS.

Boyer, R. 2004. *The Future of Economic Growth*. Cheltenham: Edward Elgar.

Boyer, R. 2005a. Coherence, diversity and the evolution of capitalisms. *Evolutionary and Institutional Economic Review* 2(1), 43–80.

Boyer, R. 2005b. How and why capitalisms differ. *Economy and Society* 34, 509–57.

Boyer, R. and Dehove, M. 2001. Théories de l'intégration européenne. *La Lettre de la Régulation*, No. 38, 1–3.

Boyer, R. and Mistral, J. 1978. *Accumulation, inflation, crises*. Paris: PUF.

Boyer, R. and Neffa, J., eds. 2004. *La crisis Argentina (1976–2001)*. Madrid, Buenos Aires: Editorial Mino y Davila.

Boyer, R. and Saillard, Y., eds. 2002. *Regulation Theory: The State of the Art*. London: Routledge.

de Vroey, M. 1984. A regulation approach interpretation of the contemporary crisis. *Capital and Class* 23 (Summer), 45–66.

Emergo: Journal of Transforming Economies and Societies. 1995. Special issue, vol. 2(4).

Hausmann, R. 1981. State landed property, oil rent and accumulation in Venezuela. Ph.D. thesis, Cornell University.

Hollingsworth, R.J. and Boyer, R., eds. 1997. *Contemporary Capitalism*. Cambridge: Cambridge University Press.

Jessop, B. 1990. *State Theory. Putting Capitalist States in their Places*. Oxford: Polity Press.

Jessop, B., ed. 2001. *Regulation Theory and the Crisis of Capitalism*, 5 vols. Cheltenham: Edward Elgar.

Lipietz, A. 1979. *Crise et inflation, pourquoi?* Paris: Maspéro.

Lipietz, A. 1985. *The Magic World: From Value to Inflation*. London: Verso.

Lipietz, A. 1986a. New tendencies in the international division of labor. In *Production, Work, Territory*, ed. A.J. Scott and M. Storper. London: Allen & Unwin.

Lipietz, A. 1986b. Behind the crisis: the exhaustion of a regime of accumulation. *Review of Radical Political Economics* 18(1–2), 13–32.

Mazier, J., Basle, M. and Vidal, J.F. 1999. *When Economic Crises Endure*. London: M.E. Sharpe.

Mistral, J. 1986. Régime international et trajectoires nationales. In *Capitalisme fin de siècle*, ed. R. Boyer. Paris: PUF.

Ominami, C. 1985. *Les transformations dans la crise des rapports nord–sud*. Paris: La Découverte.

Palombarini, S. 1999. *Vers une théorie régulationniste de la politique économique. L'Année de la Régulation 1999*, vol. 3, Paris: La Découverte.

Petit, P. 1986. *Slow Growth and the Service Economy*. London: Frances Pinter.

Quémia, M. 2001. *Théorie de la régulation et développement: trajectoires latino-américaines. L'Année de la régulation*, vol. 5, Paris: Presses de Sciences-Po.

Socio-Economic Review. 2005. A dialogue on institutional complementarity. Vol. 2(1), 43–80.

More information on régulation theory is available from the Association Recherche et Régulation. Online. Available at http://web.upmf-grenoble.fr/regulation, accessed 19 October 2006.

How to cite this article

Boyer, Robert. "regulation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000085> doi:10.1057/9780230226203.1416

The New Palgrave Dictionary of Economics Online

Reid, Margaret Gilpin (1896 – 1991)

Mary Jean Bowman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Friedman, M.; household production; life stages; permanent income; permanent income hypothesis; Reid, M. G.; transitory income

Article

Margaret Reid, a leading scholar in analysis of the economics of consumer behaviour, was made a distinguished fellow of the American Economic Association in 1980. She was Professor of Economics at Iowa State College (1930–43), the University of Illinois (1948–51), and the University of Chicago (1951–61).

A realistic theorist, Reid always looked behind data to processes that generate structural relationships. Her 1934 book on household production anticipated by three decades analyses built on the allocation of time, and she was the first (1947) to use wage-equivalent time measures of household work.

Already in Iowa she had questioned attempts to improve resource allocations by farm women that disregarded the nature of income effects. She went on to criticize assessments of the war-time cost-of-living index that neglected effects of changing incomes on the quality of goods traded, and she became the ‘directing’ member of the technical committee responsible for a report to the President's Commission on the Cost of Living (1945). Later on she challenged conventional treatments of income elasticities of consumption in general and of housing expenditures in particular (1952; 1962).

The concepts of ‘permanent’ and ‘transitory’ income were early a part of Reid's thinking (1952; 1953). Friedman drew on Reid in his 1957 application of the permanent income hypothesis to short-term shifts in consumption and saving, and Modigliani built on her work in his treatment of ‘life stages’ (Modigliani and Ando, 1960 and subsequently). In Reid's hands the concepts of ‘permanent’ and ‘transitory’ income evolved subtly and progressively in multiple facets of the analysis of consumer behaviour. After her retirement she probed interactions between health and income both over life cycles and across cohorts.

Selected works

1934. *Economics of Household Production*. New York: Wiley.

1943. *Food for People*. New York: Wiley.

1945. (With associates.) Appendix IV. Prices and the cost of living in wartime – an appraisal of the Bureau of Labor Statistics Index of the Cost of Living in 1941–44. In *Report of the President's Committee on the Cost of Living*. Washington, DC: Government Printing Office.

1947. The economic contribution of homemakers. In *Women's Opportunities and Responsibilities*, ed. L. M. Young. Philadelphia: Annals of the American Academy of Political and Social Science.

1952. Effect of income concept upon expenditure curves of farm families. In *Conference on Research in Income and Wealth, Studies*, vol. 15, part IV. Philadelphia: NBER.

1953. Savings by family units in consecutive periods. In *Savings in the Modern Economy*, ed. W.W. Heller, F.M. Boddy and C.L. Nelson. Minneapolis: University of Minnesota Press.

1955. Food, liquor and tobacco. In *America's Needs and Resources: A New Survey*, ed. J.F. Dewhurst. New York: Twentieth Century Fund.

1960. Comments on J. Crockett and I. Friend, ‘A Complete Set of Consumer Demand Relationships’. In *Proceedings of the Conference on Consumption and Savings*, vol. I, ed. I. Friend and R. Jones. Philadelphia: Wharton School of Finance and Commerce, University of Pennsylvania.

1962. *Housing and Income*. Chicago: University of Chicago Press.

Bibliography

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Modigliani, F. and Ando, A. 1960. The ‘permanent income’ and the ‘life cycle’ hypothesis of saving behavior: comparisons and tests. In *Proceedings of the Conference on Consumption and Savings*, ed. I. Friend and R. Jones. Philadelphia: Wharton School of Finance and Commerce, University of Pennsylvania.

How to cite this article

Bowman, Mary Jean. "Reid, Margaret Gilpin (1896–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://

www.dictionaryofeconomics.com/article?id=pde2008_R000092> doi:10.1057/9780230226203.1417

The New Palgrave Dictionary of Economics Online

religion and economic development

Sriya Iyer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The role of religion in economic development warrants a nuanced perspective that integrates economic theory with an understanding of socio-political structures, appreciating the econometric issues that arise in quantifying religious processes. Existing research focuses on religious structures and organizations, state religions, faith-based welfare programmes, the regulation of religion, and the impact of religion on measures of well-being such as income and education. Viewing religion as spiritual capital, with the attendant role played by religious network externalities in fostering economic development, is vital for development policy. Contemporary research in religion and economic development is flourishing, encompassing all these diverse concerns.

Keywords

Aquinas, St Thomas; club goods; economic demography; education; fertility; health care; human capital; human capital; insurance; Islamic economic institutions; poverty alleviation; Protestantism; religion and economic development; religion, economics of; religious capital; religious networks; Schumpeter, J.; self-selection; Smith, A.; social capital; social interaction (empirics); social norms; Tawney, R.; technology; terrorism; trust; Weber, M.; well-being; work ethic

Article

> The number of micro-level social anthropological studies is continually growing. Many of these concentrate on what to the economist may appear odd aspects of society such as ritual and religion … and to which he pays little or no attention. For instance, an understanding of the complex of Hindu religious beliefs as they operate at village level … is directly relevant to the problem of developing India's economy. This is but one of numerous examples which can be quoted to support the claim that development economists work in the dark unless they acquaint themselves with the relevant socio-

political literature. (Epstein, 1973, p. 6)

How times have changed since Scarlett Epstein first lamented economists' general neglect of the role of religion in the study of economic development. She need not have been quite so fearful: contemporary economics has seen the light, as it were, increasingly demanding a perspective on religion in order better to understand how it interacts with economic decision making. The increasing resilience of religion in both developed and developing countries, influencing globally both political will and popular debate, has been observed by scholars investigating the economics of religion (Iannaccone, 1998; Stark and Finke, 2001; Glaeser, 2005). Recent studies have investigated how religion affects growth (Guiso, Sapienza and Zingales, 2003; North and Gwin, 2004; Noland, 2005; Barro and McCleary, 2003; Glahe and Vorhies, 1989) with emphasis on particular religious traditions such as Islam, Hinduism or Catholicism (Kuran, 2004; Sen, 2004; Fields, 2003). Other studies have focused on the impact of religion on fertility (Lehrer, 2004; McQuillan, 2004). Still others examine the impact of religion on political outcomes (Glaeser, Ponzetto and Shapiro, 2005) and the role of religious organizations as insurance (Dehejia, DeLeire and Luttmer, 2005). Other studies examine how the causality may run the other way, from economic development to religion (Berman, 2000; Botticini and Eckstein, 2005; Goody, 2003).

Several theories have been advanced to account for the links between religion and development. First, there are theories that typify the ‘rational choice’ approach to religion and development. This approach considers the resilience of religion as a rational economic response to changes in the political, ecological and economic environments in which religions operate. In addition, a range of other structural theories encompass family socialization, social networks and a belief in other-worldly or supernatural elements. However, regardless of the scholastic tradition from which one approaches the study of religion, examining the interactions between religion and development poses significant challenges: first, to understand the endogenous interactions between religion and economic growth; second, to examine the techniques and methods needed to quantify these interactions; and third, to evaluate the impact of religion on development policy more widely.

Early writings

The economic concern with religion and development is not new, nor is it restricted to scholars of the 21st century. The writings of Thomas Aquinas, notably the *De Regno (De Regimine Principum) ad Regem Cypri*, written in 1267, dealt extensively with religion and public finance. Indeed, some scholars have considered the ideas in this work, as in Aquinas's *Summa Theologica* (1265–72), strikingly relevant for poverty reduction today; their themes of the ‘universal common good’ and ‘global civil society’ have implications for current debates about globalization and human development (Linden, 2003). The links between religion and development also feature in Joseph Schumpeter's *History of Economic Analysis* (1954). Jacques Le Goff authored *La Naissance du Purgatoire* (1981), which argued that purgatory was a necessary religious innovation for medieval capitalist development. However, it was in 1904 that Max Weber put forward his famous theory of the Protestant ethic and the spirit of capitalism, arguing that economic development in northern Europe could be explained by developments that were associated

with Protestantism – the concern with savings, entrepreneurial activity, the frugality which Puritanism demanded, and the literacy needed to read the scriptures. The essence of Weber's thesis was that nascent capitalism emerged in the 16th century in Europe on account of the Protestant ethic which arose from the Reformation. Ascetic Protestantism encouraged diligence, discipline, self-denial and thrift. Both Lutheran and Calvinist doctrines urged adherents robustly to undertake their ‘calling’. Spiritual grace from religion was attained by demonstrating temporal success in one's calling. The Protestant ethic thus involved the diligent undertaking of one's calling as a religious obligation, which promoted a work ethic that increased savings, capital accumulation, entrepreneurial activity, and investment, all of which in turn fostered economic development. Many scholars have criticized Weber's thesis, typified in the writings of Tawney (1926) and Gorski (2005). Tawney was concerned with reverse causality: how religion affected development, and in turn how economic and social changes themselves acted on religious beliefs. In his words, ‘“The capitalist spirit” is as old as history, and was not, as has sometimes been said, the offspring of Puritanism’ (1926, p. 225). Tawney argued that Puritanism both helped mould the social order and in turn was moulded by it. Gorski (2005) focuses more on whether Weber's thesis stands up to closer historical scrutiny, highlighting other aspects of the Reformation that contributed to economic development such as Protestant migration, reforms to landholding, fewer religious holidays, and insurgencies, all of which influenced labour supply and the actions of government in Protestant countries.

The economic view of religion

Against this backdrop, recent academic interest linking religion and development has centred on the economics of religion. Studies in the economics of religion have focused on applying the tools of modern economic analysis to the analysis of religious institutions, faith-based welfare programmes and the economic regulation of the church (Oslington, 2003). Three principal themes emerge: first, identifying what determines religion and religiosity; second, examining how religion and religiosity may be described as social capital; and third, understanding the micro and macro consequences of religiosity. Adam Smith (1776) made reference to the church in the *Wealth of Nations*; and recent work by economists such as Becker and Iannaccone has been very important for the development of this field. The broadly socio-economic view of religion, which expounds the rational choice approach, is set out in the work of Azzi and Ehrenberg (1975), Iannaccone (1998), Stark, Iannaccone and Finke (1996), and Stark and Finke (2000). The focus here has been both on the supply side (the structures of religious organizations) and on the demand side (the preferences of consumers in religious economies). The micro view explains religious activity as the outcome of rational choice, with utility derived both in the individual's lifetime and in the afterlife. For example, if we think of religion as a club good, then many practices are used by religions to screen potential free riders and to ensure better monitoring of the existing faithful (Iannaccone, 1992). Religion also influences individual welfare through the externalities occasioned by social behaviour (Becker and Murphy, 2000). Religious forces are important as they change the environment in which individuals operate, directly affecting individuals' choices and behaviour by changing the utilities of goods. Moreover, greater trust fostered by the religious environment can encourage repeated interactions, leading to more cooperative behaviour within networks.

It is in this way that the second theme – religion as social capital – becomes important. Three aspects are emphasized here: social networks, social norms, and sanctions to penalize deviations from norms. Corresponding to this emphasis, economists of religion have been examining ‘spiritual capital’ – or religious capital – which embodies the norms, networks and sanctions exercised by groups that are organized on the basis of religion and religious networks.

Finally, the macro and micro consequences of religiosity have been examined. For example, there are a number of channels through which religious capital might affect economic growth. Religious capital affects output by changing the manner in which technology and human capital are used. Religious capital exerts a positive impact on human capital by increasing education. For example, particularly in many less developed countries, religious networks are important not only for the religious services they provide but also for their non-religious services, specifically with respect to health and education. Moreover, as religious institutions provide this insurance function, these networks determine the extent to which education is taken up (Borooah and Iyer, 2005). In developed countries, too, this would have implications for religious market structure and the growth of residential neighbourhoods that may be based upon faith-based activities (Gruber, 2005). So understanding the economic consequences of religion is of central concern.

The empirics of religion and development

Most empirical economic studies of religion and development attempt to solve classic decompositions of the form $Y_i - Y_j = \Sigma \beta (X_i - X_j)$ where the idea is to examine the various factors (*X*) that affect measures of religious attendance or behaviour (*Y*) across individuals (*i*, *j*), or more widely across countries, or alternatively in varied historical time periods, thence to arrive at conclusions based on the effects suggested by the parameters (β) estimated.

Empirical studies of religion and development across countries have investigated religious movements, examining particularly sect behaviour, with an emphasis on contrasting the ‘European experience of religious monopoly’ with the ‘American case of religious cacophony’ (Warner, 1993, p. 1081), drawing implications for the issue of whether regulation of religious organizations is necessary. This concern manifests itself in a plethora of research projects, especially on religion in the United States (Marty, 1986–96; Finke and Stark, 1988; Warner, 1993). In cross-country studies, economists have also revisited Weber's hypothesis. Barro and McCleary (2003) assess the effect of religious participation and beliefs on a country's rate of economic progress. Using international survey data for 59 countries drawn from the World Values Survey and the International Social Sciences Program conducted between 1981 and 1999, these authors find that greater diversity of religions is associated with higher church attendance and stronger religious beliefs. For a given level of church attendance, increases in some religious beliefs – notably belief in heaven, hell and an afterlife – tends to increase economic growth.

Other studies have focused more on particular religions in varied historical time periods. For example, very useful insights have been gained by focusing on Islam and on Judaism. For Islam, there have been detailed investigations into financial systems in the Middle East including *zakat* (alms for charity) and the manner in which Islamic banks have been using a financing method equivalent to the rate of interest to overcome adverse selection and information problems. There has also been more detailed investigation into Islamic law and financial activity historically with implications for poverty reduction

in the Middle East (Kuran, 2004). There is research that has examined Jewish occupational selection using historical data from the eighth and ninth centuries onward to explain the selection of Jews into urban, skilled occupations prompted by educational and religious reform in earlier centuries (Botticini and Eckstein, 2005). Data are also being used to elucidate the role of religion in explaining historical differences in education among Hindus and Muslims in India (Borooah and Iyer, 2005).

A primary focus of current studies of religion and development is on explaining differences across individuals. For example, using data from the General Social Survey and the US Census, Gruber (2005) investigates religious market structure by estimating the effects of religious participation on economic measures of well-being, and concluded that residing in an area with more co-religionists improves well-being through the impact of increased religious participation. This particular study is also valuable from the methodological point of view, as it addresses a common problem in empirical studies of religion and development – the persistent endogeneity of religion to economic measures of well-being – and consequently the common econometric problem of how best to identify religion effects. While this particular study successfully uses ethnic heritage to provide an exogenous source of variation, and is thereby able to draw out cleanly the effects of religious participation on the variables of interest, econometrically the potential endogeneity of most religion variables is possibly the single most significant limitation of incorporating religion into empirical work in economics. This is mirrored in the many efforts to identify the effects of religion which generally have not been able to deal with self-selection issues easily.

To this end, fields such as economic demography have much to offer the study of religion and development. For example, recent research in economics has made a start towards examining the religious and economic reasons behind fertility differences between religious groups, especially in developing countries (Iyer, 2002). The economics of religion has also elucidated the study of politics, both local and international: Glaeser (2005) presents an economic model of religious group behaviour and the so-called ‘political economy of hatred’. The economic approach to religion has been evaluating whether religion and politics are mutually exclusive. Glaeser, Ponzetto and Shapiro (2005) link religion with strategic extremism – the issues and platforms espoused by political parties, and the manner in which private information matters for this. Other studies have focused on terrorism and display a more general preoccupation with understanding views and attitudes in the Muslim world (Gentzkow and Shapiro, 2004).

Drawing a perspective from all these classes of studies, it strikes one that emerging economies are experiencing appreciable modern economic growth, yet this is coterminous with the increasing resilience of religious institutions. And it is this dichotomy between the sacred and the secular which epitomises the puzzle of the relationship between religion and economic development. It seems reasonable to address this puzzle by combining quantitative analysis of sample data with nuanced qualitative evaluations of the textual theology of religion, linking these to the manner in which individuals and institutions interpret religion at a local level. As well, an appreciation of the approach of the interdisciplinary economist would permit a more informed understanding of all these concerns. Economists will enthusiastically study religion and economic development in the future, and they will do so with ascetic assiduity – researching data with all the intensity of religious fervour in order to provide thoughtful prophecy for development policy.

See Also

- Islamic economic institutions
- religion, economics of
- social capital
- social interactions (empirics)
- Weber, Max

Bibliography

Aquinas, St Thomas. 1265–72. *Summa Theologica*. Online. Available at http://www.ccel.org/ccel/aquinas/summa.html, accessed 22 June 2006.

Aquinas, St Thomas. 1267. *De Regno (De Regimine Principum) ad Regem Cypri*. In *Aquinas: Political Writings*, ed. and trans. R.W. Dyson, Cambridge: Cambridge University Press, 2002.

Azzi, C. and Ehrenberg, R. 1975. Household allocation of time and church attendance. *Journal of Political Economy* 83, 27–56.

Barro, R.J. and McCleary, R. 2003. Religion and economic growth across countries. *American Sociological Review* 68, 760–81.

Becker, G. and Murphy, K. 2000. *Social Economics: Market Behavior in a Social Environment*. Cambridge, MA: Harvard University Press.

Berman, E. 2000. Sect, subsidy and sacrifice: an economist's view of ultra-orthodox Jews. *Quarterly Journal of Economics* 115, 905–53.

Borooah, V. and Iyer, S. 2005. *Vidya, veda* and *varna*: the influence of religion and caste on education in rural India. *Journal of Development Studies* 41, 1369–404.

Botticini, M. and Eckstein, Z. 2005. Jewish occupational selection: education, restrictions, or minorities? *Journal of Economic History* 65, 922–48.

Dehejia, R., DeLeire, T. and Luttmer, E. 2005. Insuring consumption and happiness through religious organizations. Working Paper No. 11576. Cambridge, MA: NBER.

Epstein, T. 1973. *South India: Yesterday, Today and Tomorrow*. London: Macmillan.

Fields, B. 2003. *The Catholic Ethic and Global Capitalism*. Aldershot, UK and Burlington, VT: Ashgate.

Finke, R. and Stark, R. 1988. Religious economies and sacred canopies: religious mobilization in American cities, 1906. *American Sociological Review* 53, 41–9.

Gentzkow, M. and Shapiro, J. 2004. Media, education and anti-Americanism in the Muslim world. *Journal of Economic Perspectives* 18(3), 117–33.

Glaeser, E. 2005. The political economy of hatred. *Quarterly Journal of Economics* 120, 45–86.

Glaeser, E., Ponzetto, G. and Shapiro, J. 2005. Strategic extremism: why Republicans and Democrats divide on religious values. *Quarterly Journal of Economics* 120, 1283–330.

Glahe, F. and Vorhies, F. 1989. Religion, liberty and economic development: an empirical investigation. *Public Choice* 62, 201–15.

Goody, J. 2003. Religion and development: some comparative considerations. *Development* 46(4), 64–7.

Gorski, P. 2005. The little divergence: the Protestant Reformation and economic hegemony in early modern Europe. In *The Protestant Ethic Turns 100: Essays on the Centenary of the Weber Thesis*, ed. W. Swatos and L. Kaelber. Boulder and London: Paradigm Publishers.

Gruber, J. 2005. Religious market structure, religious participation, and outcomes: is religion good for you? *Advances in Economic Analysis & Policy* 5(1), article 5.

Guiso, L., Sapienza, P. and Zingales, L. 2003. People's opium? Religion and economic attitudes. *Journal of Monetary Economics* 50, 225–82.

Iannaccone, L. 1992. Sacrifice and stigma: reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy* 100, 271–91.

Iannaccone, L. 1998. Introduction to the economics of religion. *Journal of Economic Literature* 36, 1465–95.

Iannaccone, L., Finke, R. and Stark, R. 1997. Deregulating religion: the economics of church and state. *Economic Inquiry* 35, 350–64.

Iyer, S. 2002. *Demography and Religion in India*. Delhi: Oxford University Press.

Kuran, T. 2004. Why the Middle East is economically underdeveloped: historical mechanisms of institutional stagnation. *Journal of Economic Perspectives* 18(3), 71–90.

Lehrer, E. 2004. Religion as a determinant of economic and demographic behaviour in the United States.

Population and Development Review 30, 707–26.

Le Goff, J. 1981. *La Naissance du Purgatoire*. Paris: Gallimard.

Linden, I. 2003. *A New Map of the World*. London: Darton, Longman and Todd.

Marty, M. 1986–96. *Modern American Religion*, 3 vols. Chicago: University of Chicago Press.

McQuillan, K. 2004. When does religion influence fertility? *Population and Development Review* 30, 25–56.

Noland, M. 2005. Religion and economic performance. *World Development* 33, 1215–32.

North, C. and Gwin, C. 2004. Religious freedom and the unintended consequences of state religion. *Southern Economic Journal* 71, 103–17.

Oslington, P., ed. 2003. *Economics and Religion*, vols. 1 and 2. The International Library of Critical Writings in Economics 167. Cheltenham: Edward Elgar.

Schumpeter, J. 1954. *History of Economic Analysis*. London: Oxford University Press.

Sen, A. 2004. Democracy and secularism in India. In *India's Emerging Economy: Performance and Prospects in the 1990s and Beyond*, ed. K. Basu. Cambridge, MA and London: MIT Press.

Smith, A. 1776. *An Inquiry Into the Nature and Causes of the Wealth of Nations*, 5th edn, ed. E. Cannan. London: Methuen and Co. Ltd., 1904.

Stark, R. and Finke, R. 2000. *Acts of Faith: Explaining the Human Side of Religion*. Berkeley: University of California Press.

Stark, R. and Finke, R. 2001. Beyond church and sect: dynamics and stability in religious economies. In *Sacred Markets and Sacred Canopies: Essays on Religious Markets and Religious Pluralism*, ed. T. Jelen. Lanham: Rowman and Littlefield.

Stark, R., Iannaccone, L. and Finke, R. 1996. Religion, science, and rationality. *American Economic Review* 86, 433–7.

Tawney, R. 1926. *Religion and the Rise of Capitalism*. London: Penguin Books, 1990.

Warner, R. 1993. Work in progress toward a new paradigm for the sociological study of religion in the United States. *American Journal of Sociology* 98, 1044–93.

Weber, M. 1904. *The Protestant Ethic and the Spirit of Capitalism*, trans. T. Parsons. London: Routledge, 1992.

How to cite this article

Iyer, Sriya. "religion and economic development." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000245> doi:10.1057/9780230226203.1418

The New Palgrave Dictionary of Economics Online

religion, economics of

Laurence R. Iannaccone and Eli Berman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Adam Smith invented the economics of religion, famously arguing for church-state separation on efficiency grounds since state religions become inefficient monopoly providers of religious services and because competition for monopoly status is often violent. Smith also developed theories of religious sects and sectarian violence. Modern applications of theory and data generally support Adam Smith's conjectures. Recent work also explores: religious activity as a consumer choice, the demand for spiritual services, religious human capital and religious social capital, club models of sects – benign and violent – and the macroeconomic consequences of beliefs and religiosity.

Keywords

addiction; Boulding, K.; capitalism; charitable donations; clubs; fertility; firm, theory of; free-rider problem; human capital; intertemporal utility; marriage and divorce; mutual aid; rational choice; rationality; religion, economics of; religious capital; religious economics; rent seeking; sect; social capital; social cohesion; social norms; stable preferences; terrorism; Weber, M.; women's work and wages

Article

Adam Smith laid the foundations for the economic study of religion in *The Wealth of Nations* (1776, pp. 788–814). He argued that self-interest motivates the clergy; that market forces constrain churches just as they constrain secular firms; that competition improves the quality of religious services provided; and that government regulation distorts the provision of religion, reducing quality and promoting conflict. He also outlined a theory of sectarianism, a theory of religious violence and civility, and a general theory of Church and State.

After this inspired start the economics of religion lay dormant and nearly dead for two centuries. It is now enjoying a rebirth, animated by new data, methods and theory. Economists and other social scientists have harnessed rational choice models and modern empirical tools to study secularization, pluralism, church growth, religious extremism, conversion, fertility, Church–State relations, and more. The field now claims hundreds of papers, scores of contributors, an annual conference and international association (the Association for the Study of Religion, Economics, and Culture), university research centres, and even an AEA subject code (Z12). (New university centres are at Harvard, George Mason University, the University of Southern California and in Canberra, Australia.)

Current research on religion and economics falls into three related subfields: economic theories of religion, studies of religion's economic consequences, and religious assessments of economic policy. Adam Smith's critique of state-supported religion in the *Wealth of Nations* exemplifies the first subfield; Max Weber's ‘Protestant ethic’ conjecture the second. Together these two subfields constitute *the economics of religion* – the subject of this article. Our goal is to introduce readers to the distinctive economic ideas and models that have enhanced the social-scientific study of religious beliefs, behaviour, and institutions. (For a more complete review of the literature prior to 1998, see Iannaccone, 1998.)

This article makes no attempt to survey the field of religious economics, both because the latter tends to be religion-specific and because it is far from the mainstream of economic research. Religious economics seeks to evaluate economic behaviour and institutions in the light of sacred precepts. Mahmoud El-Gamal's recent book, *Islamic Finance* (2006), is a good example, examining whether current practices in banks that follow Islamic law actually serve the objectives of those laws. The literature on religious economics is large, diverse, and as old as religion itself – including, for example, the many biblical injunctions concerning property, slavery, wages, tithing, interest, wealth and poverty. With the help of

economists and philosophers, contemporary clerics continue to debate the merits of income inequality, tax laws, private property, deficit spending, monetary policy, income redistribution, workers rights, interest rates, banking laws, entrepreneurship, government regulation, international trade, debt relief, unionization, entitlement programmes and much more. For representative readings in religious economics, see Oslington (2003) and the *Journal of Markets and Morality* (published by the Acton Institute for the Study of Religion and Liberty).

Social scientists once viewed religion as a dying vestige of our primitive and pre-scientific past. Modern research and contemporary events have destroyed this simplistic view of human history. The rise of radical Islam, the revival of religious practice in much of the former Soviet Union, the explosive growth of Protestantism in Latin America and sub-Saharan Africa, and the contribution of religion to identity (and often to conflict) all over the world testify to the continuing vitality of religion. And although religious belief and activity have declined in many economically advanced countries since the 1960s, the corresponding US data display remarkable stability, whether one focuses on rates of attendance, contributions, membership, or belief. Indeed, religiosity has become one of the strongest predictors of voting patterns and political orientation in America (Glaeser et al., 2006).

We cannot say why economics ignored religion for so long. The brilliant and iconoclastic economist Kenneth Boulding discussed economic features of religion long before the modern revival, but his insights seem to have gone largely unnoticed by economists or sociologists. (Boulding's essays on religion and economics from the 1950s appear in Boulding, 1970.) The other social sciences have subfields dedicated to the study of religion and most have sought to understand the connection between religious and economic trends – the most famous and influential generalization being Max Weber's ‘Protestant ethic’ thesis. (Weber studied economic history and was well-acquainted with Smith's *Wealth of Nations*. His essay, ‘The Protestant Sects and the Spirit of Capitalism’, describes how denominational membership enhanced the reputation and business prospects of Americans around 1900: see Weber, 1920. The essay appears to have been inspired by Smith's theory about the ways in which sect membership benefits a poor person: see Smith, 1776.) It seems likely, however, that most economists saw religion as too far removed from the realm of rational choice and market behaviour. We encourage the reader to revisit the issue of religion and rationality after reading this essay.

1 Economics, sociology, and rational choice

Nearly all economic theories rely on the twin assumptions of *rational choice* and *stable preferences*. In the realm of religion, this means choosing which religion, if any, to accept and how extensively to participate in it. These optimal choices need not be permanent. Economic models do a good job explaining differences in religious activity, both over time and across individuals. In keeping with the assumption of stable preference, however, these explanations rarely invoke varied norms, tastes or beliefs. A good economic story explains behaviour in terms of optimal responses to varying circumstances – such as prices, incomes, skills, experiences, technologies or resource constraints.

Although the previous paragraph merely extends modern economic orthodoxy to the realm of religion, it borders on sociological heresy. The commitment of economists to rational choice and stable preferences must be understood as relative, not absolute. Since the late 1970s, economists have devoted a great deal of attention to modelling preference formation. Formal models of religious capital formation (Iannaccone, 1984; 1990) are, in fact, directly linked to Becker's (1996) subsequent work on rational addiction and taste change. Recent work in the fields of behavioural, experimental and evolutionary economics underscores the extent to which choice systematically deviates from rationality; and social norms, social networks and imperfect information constrain choices further still. But it would be wrong to conclude that economists and sociologists now embrace a common ‘world view’ – as is readily apparent when one contrasts the papers presented by economists and sociologists at the joint annual meetings of the *Association for the Study of Religion, Economics, and Culture* and the *Society for the [Social] Scientific Study of Religion*. Most sociologists remain very sceptical about the value of formal models, rational choice theory and methodological individualism – a legacy passed down from the founders of the field, who promoted sociology as a corrective to errors and omissions of economics (Swedberg, 1990). Add the influence of Weber (1920; 1963), who made ‘rationality’ central to his analysis of religion while using the word in ways foreign to most contemporary economists, and ‘doctrinal’ debate is unavoidable. But the overall response to economic forays into religion has been surprisingly ecumenical, with several leading sociologists of religion going so far as to characterize economic theory and market models as ‘the new paradigm’ for religious research (Stark and Finke, 2000; Warner, 1993; Young, 1997).

2 Households and consumer choice

Economists finally returned to the study of religion in the 1970s, inspired by Gary Becker's path-breaking work on economics of the family. The first papers modelled church attendance and religious contributions as a special form of household production – one that involved trade-offs between time and money inputs, secular versus religious outputs, and present versus afterlife utility (Azzi and Ehrenberg, 1975; Ehrenberg, 1977). Formally, households maximize an intertemporal utility function $U = U(Z_1, \ldots, Z_n, A)$, where Z_t denotes secular consumption activities in period *t*, and *A* is consumption activity in an afterlife (of possibly infinite duration). In each period (of this life) households can spend their time, *T*, and goods, *X*, on either secular consumption or religious activities, $Z_t = Z(T_{Z_t}, X_{Z_t})$, $R_t = R(T_{R_t}, X_{R_t})$. Religious activities over a lifetime create afterlife consumption, $A = A(R_1, \ldots, R_n)$. Combined with a standard lifetime budget constraint, and on the assumption that the marginal product of religious activity does not decrease with age, the Azzi–Ehrenberg model predicts that religious activity increases with age, *ceteris paribus*. The model also predicts that households with high value of time (high wages) will substitute

goods for time in producing religious activity.

As Azzi and Ehrenberg (and many others) have shown, religious activity does tend to increase with age. But it is not at all clear that the Azzi–Ehrenberg model captures the principal cause of this age effect. Ulbrich and Wallace (1983) found that activity increases with age even among those who do *not* believe in the afterlife. And Iannaccone (1984) showed that even in the absence of afterlife expectations, the rational accumulation of religious human capital (that is, rational religious ‘addiction’) could simultaneously account for the observed age effect as well as observed patterns of religious conversion, intermarriage, and marital stability (see also Lehrer and Chiswick, 1993; Neuman, 1986).

Predictions concerning religious substitution are on much stronger ground. Substitution of goods for time is observed across individuals, households and denominations. Although we cannot directly observe most religious commodities, we can observe the inputs used to produce them. The principal time and money inputs – attendance and contributions – are routinely measured in surveys. More specialized studies provide detailed information on time (such as time devoted to religious services, private prayer and worship, religious charity, and many other religious activities) and money (such as expenditures for special attire, transport, religious books and paraphernalia, sacrificial offerings, and contributions used to finance staff, services and charitable activities of religious organizations). Several studies, including Ehrenberg (1977), Iannaccone (1990), Hungerman (2005) and Gruber (2004), have found that attendance and donations are substitutes – and the recent work demonstrates that substitution remains strong even after one controls for endogeneity bias.

Both in theory and in fact, substitution induces different methods of religious organization and worship across different socio-economic strata. High-income congregations tend to hold shorter services, make heavy use of professional staff and inhabit more elaborate facilities. Longer services, volunteer workers, rented meeting halls and pot-luck dinners are typical of poorer congregations. We observe these differences within denominations and even within congregations (as members improve their socio-economic status), but the differences are especially stark *across* the denominations of a religious tradition, such as Reform Judaism versus Orthodox Judaism or Episcopalians versus Southern Baptists. Many Episcopalian or Presbyterian congregations have plenty of money to cover salaries and operating expenses but remain hard-pressed to recruit volunteers for their choirs, youth programmes, committees and other traditional programmes. For such denominations prosperity has proved a mixed blessing.

Economic trends forced adaptation and none more so than the growth of women's wages and workforce participation. As women have moved into the labour force and overall family earnings have grown, congregations have had to purchase many services formerly supplied by volunteers. The pattern is illustrated by Luidens and Nemeth's (1994) study of expenditure trends in Presbyterian denominations, which found that their (fourfold) increase in real per-capita giving from the 1940s to the 1990s was spent primarily on local congregational services previously supplied by volunteers.

3 Religion, magic and uncertainty

Contemporary theories of rational religious belief begin with just a few assumptions about human nature and the human condition – in essence scarcity, rationality, and the capacity to conceive of supernatural beings or forces (Iannacone and Berman, 2006; cf. Stark and Bainbridge, 1987). From these, they derive a universal demand for supernaturalism and a universal distinction between magic (emphasizing control of impersonal supernatural forces) and religion (emphasizing interaction with supernatural beings). Specialized suppliers arise naturally in both realms, but markets for magic and religion operate quite differently. It is relatively easy to test (and disprove) a magician's ability to control supernatural forces, but much harder to falsify a priest's claims concerning God. In practice, only religion can sustain long-term relationships, high levels of commitment, and moral communities. As Emile Durkheim (1915, p. 42) famously observed, ‘there is no church of magic’.

As we shall see in Section 6, a strong religion can induce its members to foreswear all other suppliers of supernatural goods and services. But exclusivity is not a ‘natural’ outcome. Given the tremendous uncertainty that surrounds the supernatural, rational consumers are inclined to patronize many different suppliers – investing, so to speak, in diversified *portfolios* of supernatural commodities (Iannaccone, 1995). Diversification over different supernatural products and suppliers is pervasive in the (non-communal) market for magic, including the so-called ‘New Age’ movement. It also prevails in most polytheistic settings, including in the Greco-Roman world, and it remains common in Asian religious traditions. Judaism, Christianity and Islam display a much greater capacity to sustain exclusivity, but (as we shall discuss in Section 6) only within communal settings that promote collective action, strong social ties, and large investments in religious capital.

4 Religious capital

James Coleman's (1988, p. 97) concept of ‘social capital’ helps connect rational choice theory to sociological analysis. Iannaccone's (1984; 1990) concept of ‘religious capital’ offers an analogous bridge from rational choice theory to the sociology of religion. Both concepts are inspired by human capital theory (Becker, 1964; Schultz, 1961), and both emphasize relationships rather than purely individual capacities.

Let S_{R_t} denote the stock of relationships, sensitivities and skills that alter a person's real or perceived benefits from religious activity at time *t*. Religious commodity production thus depends on current inputs of time and money and the current stock of religious capital, $R_t = R(T_{R_t}, X_{R_t}, S_{R_t})$. The SR variable can encompass a range of concepts, including religious habits, spiritual capital and social capital. Indeed, the mathematical models and empirical analyses remain essentially the same whether one frames the model in terms of the

formation of religious ‘preferences’ or the accumulation of (unobservable) religious ‘capital’. In either case, however, religious experience has two key features. First, past experience alters the value of current religious activities and thereby affects rates of religious participation: $S_{R_t} = F(T_{R_t}, X_{R_t}, S_{R_{t-1}})$. Second, most religious experience is ‘context specific’, yielding maximal benefits within the context of specific relationships, congregations, denominations and traditions. Religious capital remains a distinctive form of social and human capital because religions claim to promote relationships with supernatural beings. This enables religious institutions to maintain exceptionally high levels of commitment, but not without collective production, exclusivity and sacrifice. The bundle appeals less to people with better secular opportunities; hence we observe a ‘church-to-sect’ *spectrum* of denominations within most religious traditions. For details, see Iannaccone (1995) and Iannaccone and Berman (2006).

Capital models yield predictions that are well supported by evidence, including: (*a*) children tend to choose the same or similar religious denominations as did their parents; (*b*) conversion (like career choice) tends to occur early in adulthood, leaving time to accumulate religion-specific capital; (*c*) interfaith marriage is less likely when religious capital accumulation is high; (*d*) shared-faith marriages lead to higher rates of religious participation (due to complementarities in household production, and not mere sorting of more religious partners into shared-faith marriages); and (*e*) shared-faith marriages have lower rates of divorce and higher rates of fertility (Iannaccone, 1990; Lehrer and Chiswick, 1993; Waite and Lehrer, 2003).

Religion also contributes to extended relationships, social networks and shared norms. Indeed, Coleman's (1988) seminal article on social capital concerned the impact of (Catholic) religious schools. Empirical studies find that nearly half of all associational memberships, personal philanthropy and volunteering in the United States is church-related, leading Putnam (2000) to conclude that ‘[f]aith communities … are arguably the single most important repository of social capital in America’. Yet social capital research has yet to give much attention to religion – see, for example, the literature review by Sobel (2002). There remain tremendous opportunities for policy-relevant research on religion's contribution to cooperation (Sosis and Ruffle, 2003), social multipliers (Becker and Murphy, 2000), threshold effects (Granovetter, 1978), public preferences (Kuran, 1995), and much more.

5 Measuring the effects of religious capital

Numerous empirical studies suggest that religious belief and participation yield a wide range of benefits, including mental and physical health, longevity, reduced substance abuse and marital stability (see Koenig, McCullough and Larson, 2001, for an extensive review of the relevant research). The statistical results must, however, be viewed with caution. We lack good instruments for religion on both the supply and demand sides, and most research examines only contemporary American data. Problems of spurious correlation and unobserved heterogeneity may afflict many published studies, as Heaton (2006) notes in his re-analysis of data on religion and crime. On the other hand, the positive association between religion and health has held up despite many different efforts to root out spuriousness, and Freeman's (1986) careful data analysis provides compelling evidence that church attendance really does lead to higher employment rates, higher school attendance, less crime and lower alcohol consumption and drug use among Black males in the United States.

There are many plausible reasons why religiosity might promote beneficial outcomes. As Adam Smith emphasized in his *Theory of Moral Sentiments* (1759), faith in an omniscient deity can solve otherwise intractable problems of self- and social control. Since religion is the quintessential credence good, religious institutions tend to be relatively efficient producers of moral restraint. And there can be no doubt that communities of faith do provide many concrete services while seeking to instil faith in the young, maintain faith among adults and constrain deviant behaviour. The potential benefits from these mechanisms are underscored by the (not undisputed) evidence that religious constraints on sexual conduct have reduced or limited AIDS among Muslims in central Africa and Christians in Uganda (Green, 2003).

6 Club models of religion

Club models have made major contributions to our understanding of ‘cults’, ‘sects’ and religious extremism. They also account for characteristics of religion that seem inconsistent with rational choice and risk-aversion – including the success of groups that demand exclusivity, sacrifice and stigma.

Club models start with the fundamental fact that religious ‘commodities’ are more compelling and gratifying when they are produced and consumed *in groups*. Effective congregations require highly committed members, not mere customers. In this respect, effective congregations are more like families than firms. This suggests that models in which the *i*th member's religious satisfaction has the form $R_i = F(T_{R_i}, X_{R_i}, Q)$, where *Q* is an index of the religious inputs of all the *other* group members.

As economists well know, shirking and free-riding constantly threaten collective action, especially in large groups. Paying people to attend church, accept church doctrine or support fellow members fails to solve the problem because a member's commitment and inputs to the group are difficult to observe, and payment rewards the wrong motivations. But the problems *can* be mitigated by seemingly gratuitous costs – the sacrifice and stigma characteristic of deviant religious. Sacrifice and stigmas enhance utility by screening out people who lack commitment and boosting involvement among those who remain in the group. Such groups manifest many distinctive characteristics that empirical researchers have long associated with ‘sectarian’ religions, including distinctive diet, dress or sexual conduct; physical separation from mainstream society; painful or costly rites; rules that limit social contact with non-members; and prohibitions restricting normal economic or recreational activities. (For more on the modern theory of church and sect, see Iannaccone, 1988; 1991.)

Sect theory also accounts for people's willingness to forgo religious diversification despite the obvious risk associated with most religious assurances. Sectarian religions can maintain levels of commitment and involvement that compensate for the increased risk associated with exclusivity. Corresponding constraints can almost never be sustained in standard, secular markets (nor in the impersonal market for magic) because exclusivity does not enhance the production non-collective goods and services (Iannaccone, 1995).

The club model has received wide acceptance, in part because it fits the data so well. Both cross-sectional surveys and case studies find substantially higher levels of mutual aid and social cohesion in more sectarian religious communities. Iannaccone's work on (mostly Christian) sects has been extended to radical religious Jews (Berman, 2000) and Muslims (Berman and Stepanyan, 2003; Chen, 2004). Despite some lingering debate over the extent of free-rider problems in mainline churches or the actual level of costs imposed by contemporary conservative churches, the basic model remains the natural starting point for studies of high-cost groups. The club model works well not only for religious groups routinely called sects, cults and fundamentalists, but also for communes, gangs, radical militias and terrorist organizations. The basic insight is that an organization designed to exclude free-riders and limit free-riding will be well equipped to exclude defectors and limit defection, the Achilles' heel of militias and terrorists. Thus religious sects prove to be especially effective at terrorism, militia activity and suicide terrorism (Berman, 2003; Berman and Laitin, 2005; Berman and Stepanyan, 2003).

7 Churches as firms

Many religious organizations are legally designated as firms, and many more look surprisingly firm-like. Around the time economists became interested in religious households, several sociologists of religion began thinking of churches as firms, re-examining old data sources with new theories of rational exchange, entrepreneurship and market competition (Finke and Stark, 1988; Stark and Bainbridge, 1985). Finke and Stark (1992) trace the explosive growth of Methodist and Baptist churches in 19th century America to superior marketing, organization and clergy incentives. By the 1990s, these economic and sociological streams of scholarship together included studies of sectarianism, denominational vitality, 'franchising' of religious brands, religious extremism, doctrinal innovation, Church and State, religious markets, non-Western faiths, religious history, and more. Ekelund et al. (1996) analyse numerous features of medieval Catholicism in terms of its monopoly status. Drawing upon standard theories of monopoly, rent seeking and transaction costs, they offer economic explanations for interest rate restrictions, marriage laws, the Crusades, the organization of monasteries, indulgences, and the doctrines of heaven, hell and purgatory (see also Ekleund, Hébert, and Tollison, 2006). Work on churches as firms continues to grow rapidly, in part because firms are easier to model than clubs, but also because the theory of the firm is so rich in predictions and data.

8 Religious markets and government intervention

Whether we think of them as clubs or as firms, individual denominations collectively constitute a *religious market* as long as they provide services that are substitutes. The theories of religion described above predict the existence of different market segments: exclusive 'sects' that operate like clubs, inclusive 'churches' sustained by a core of professionals which are more firm-like, and markets for 'magic' organized around simple exchanges between practitioners and clients.

Almost all economists and sociologists of religion accept the notion that religion in America constitutes a vast competitive market, overflowing with 'products' that range from New Age paraphernalia to orthodox liturgies. Scholars likewise accept that market success requires entrepreneurship, innovation, and sensitivity to the demands of consumers. As a result, themes that rarely surfaced prior to Finke and Stark's *Churching of America* (1992) now parade as common sense. Even the harshest critics of rational choice theory (such as Bruce, 1999), emphasize the centrality of religious choice in today's world.

The most informative studies closely study how markets actually work. Market-oriented research must carefully address numerous issues, including product attributes, marketing strategies, incentive structures, exchange relationships, consumer characteristics, and Church–State relationships. Andrew Chesnut's (2003) study of rapidly growing religious movements in Latin America illustrates this point by showing how specific religions offer distinctive products that directly address the health- and family-oriented concerns of poor and middle-class women. Anthony Gill (1998) shows that Catholic bishops are much more likely to side with the poor in Latin American countries where Protestant growth threatens the Church's historic monopoly.

Adam Smith (1776, pp. 788–814) argued that established religions face the same incentive problems that plague other state-sponsored monopolies: lack of competition generates a low quality product.

> The teachers of [religion] …, in the same manner as other teachers, may either depend altogether for their subsistence upon the voluntary contributions of their hearers; or they may derive it from some other fund to which the law of their country many entitle them…. Their exertion, their zeal and industry, are likely to be much greater in the former situation than the latter. In this respect the teachers of new religions have always had a considerable advantage in attacking those ancient and established systems of which the clergy, reposing themselves upon their benefices, had neglected to keep up the fervour of the faith and devotion in the great body of the people.

Iannaccone tested Smith's conjecture with modern data. Figure 1 illustrates that within predominantly Protestant countries, church attendance declines sharply as the religious market becomes more concentrated. (The Herfindahl-style ‘Protestant Concentration Index’ proxies state support for particular religions and has the form $H = \Sigma_i S_i^2$, where S_i is the population share of the *i*th Protestant denomination.) All other surveyed measures of religiosity, including belief in God, fall with concentration as well. The data, and Smith's theory, strongly suggest that America's ‘religious exceptionalism’ is largely a product of religious laissez-faire. North and Gwin (2004) report similar results using a much larger number of countries and more direct measures of Church–State relationships.

Figure 1

Market concentration and church attendance. *Source*: Gallup polls. See Iannaccone (1991) for details.

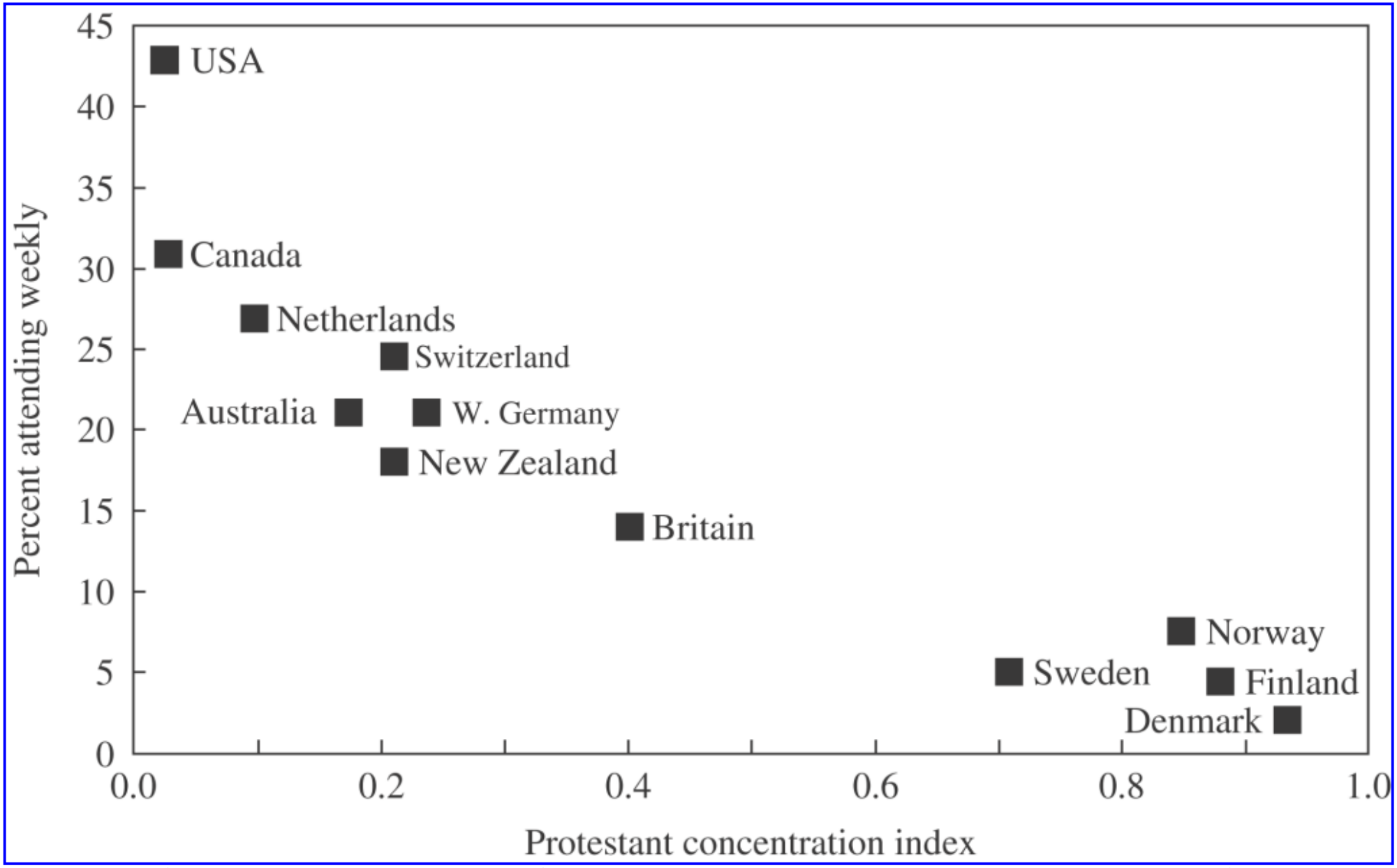

Several studies have found positive correlations between local levels of religious diversity and religious activity *within* the USA, including an especially well-crafted study by Finke,

Guest, and Stark (1996). But much of this work suffers from specification problems that inevitably arise when rates of religious *membership* (as opposed to more direct measures of belief and participation) are regressed onto (membership-based) measures of religious diversity (Voas, Olson and Crockett, 2002). Nor is it clear that concentration *should* signal the presence of inefficient religious ‘monopoly’ across cities or states given the nation's minimal barriers to religious entry and innovation. There is, however, strong historical evidence that religious competition raises religious participation (note especially the work of Finke and Stark, 1992, and Olds, 1994, who show that post-colonial disestablishment led to the rapid growth in overall church membership rates, clergy demand, and the – primarily Baptist and Methodist – non-established denominations, while the major denominations that had enjoyed the support of colonial governments – primarily Episcopal, Presbyterian, and Congregational) rapidly *lost* market share).

9 Macroeconomic consequences of religion

Weber's *The Protestant Ethic and the Spirit of Capitalism* (1920) famously claimed that the Calvinist doctrine of predestination triggered a mental revolution within Protestantism that gave rise to modern capitalism. This remains the most influential single conjecture on the macroeconomic effects of religion, which is unfortunate since nearly all subsequent empirical research has shown it to be *false*.

Almost all the capitalist institutions that Weber emphasized actually preceded the Reformation (Stark, 2005; Tawney, 1998). Across and within European countries economic development was uncorrelated with religion (Samuelsson, 1993; Delacroix and Nielsen, 2001). The second country to industrialize was Belgium, which is Catholic. Although Germany and the Netherlands were early developers and majority Protestant, the fastest growth within those countries was among Catholic families of the Rhineland and Amsterdam, which were majority Catholic.

Despite much work by historians and sociologists, there is no consensus concerning the macroeconomic impact of Protestantism, Christianity, monotheism, or religiosity in general. Economists have recently entered this field of enquiry with cross-national studies of survey and census data. Barro and McCleary's (2003) cross-national analysis of survey and census data suggests that belief in hell boosts economic growth whereas frequency of church attendance retards growth (perhaps because the former induces honest and industry whereas the latter reduced time spent working). Using cross-national data from the World Value Surveys, Guiso, Sapienza and Zingales (2003) find that religious beliefs in general, and Christian beliefs in particular, are positively associated with economic attitudes conducive to higher per capita income and growth. Many other economists have begun doing similar studies, but data problems abound. In addition to standard econometric difficulties, there are scarcely any cross-national religious surveys that predate the 1980s; we cannot validate most responses concerning religion; and the meaning of religious participation and belief varies dramatically across cultures. There is better evidence of links between religious and socio-economic variables at the level of individuals and groups than for countries or cultures. For example, average family size and socio-economic status differ quite substantially across different religious groups in both rich and poor nations (Chiswick, 1983; Iyer, 2002). Historical studies do suggest strong relationships between religious and economic institutions, most notably those of medieval Europe. Ekelund, Hébert and Tollison (1996) interpret many distinctive features of medieval Catholicism as forms of rent seeking, and there is no doubt that the Church was by far the most important economic institution in medieval Europe. Richardson (2005) offers strong evidence that the doctrine of purgatory gained rapid acceptance because it served to link religious and economic activities (within guilds) in a way that solved commitment problems that arose because of the social disruptions induced by the Black Death. Timur Kuran (2004) makes a compelling case that specific Islamic legal institutions contributed significantly to the economic decline in Muslim countries relative to those of Europe over the past 500 years.

Recent attempts to promote development in poor and post-Communist countries affirm the importance of ethical norms and moral precepts, many of which have religious foundations (Hayek, 1988, pp. 135–40). Communism may be the most striking example of an economically and socially destructive religion, albeit a religion without traditional deities. In this sense, the strongest evidence for Weberian-style theory may be negative: some powerful systems of belief *do* retard economic progress.

10 Religious militancy

American economists have tended to ignore religion as a subject of public policy, in part because the ‘establishment’ and ‘free exercise’ clauses of the First Amendment radically limit the religious role of government. Within this constitutionally mandated environment of religious laissez-faire (which initially constrained the federal government, but later extended to the states), Americans have maintained extraordinarily high rates of religious activity, diversity and tolerance. But elsewhere, religion remains a major factor in wars, civil unrest and ethnic conflict.

Adam Smith recognized that a detached and lazy clergy was just one cost associated with the marriage of Church and State. When government favours a particular religion in return for its support of the state, the favoured group inevitably demands the suppression of its competitors, and all other groups resist suppression and fight to capture favoured status. It is no coincidence that the USA has remained remarkably free of religious partisanship and militancy while other nations burn with religious conflict. Policies analogous to those embodied in the First Amendment's free exercise and establishment clauses may be key components of the so-called ‘war on terror’ (Iannaccone and Berman, 2006).

Conclusion

The economics of religion has animated research on secularization, pluralism, church growth, religious extremism, religious markets, the consequences of religion, and more. Forecasting the future of the field is a task best left to prophets. Yet promising areas include the study of non-Western religions, religious militancy, religion and demography, the relationship between religious decline and the growth of the welfare state, and the role of religion in the formation of preferences and social capital. Insights from experimental economics, behavioural economics, game theory, industrial organization, and the economics of information and uncertainty have scarcely been explored. And if the past is any indication of the future, economists still have much to learn from religious historians, sociologists, anthropologists, and other scholars after 200 years of wandering in the secular wilderness.

See Also

- Smith, Adam
- social capital

Bibliography

Azzi, C. and Ehrenberg, R. 1975. Household allocation of time and church attendance. *Journal of Political Economy* 84, 27–56.

Barro, R.J. and McCleary, R.M. 2003. Religion and economic growth across countries. *American Sociological Review* 68, 760.

Becker, G.S. 1964. *Human Capital: A Theoretical and Empirical Analysis*, 1st edn. New York: Columbia University Press for the NBER.

Becker, G.S. 1996. *Accounting for Tastes*. Cambridge, MA: Harvard University Press.

Becker, G.S. and Murphy, K.M. 2000. *Social Economics: Market Behavior in a Social Environment*. Cambridge, MA; London: Belknap Press of Harvard University Press.

Berman, E. 2000. Sect, subsidy and sacrifice: an economist's view of Ultra-Orthodox Jews. *Quarterly Journal of Economics* 115, 905–53.

Berman, E. 2003. Hamas, Taliban and the Jewish underground: an economist's view of radical religious militias. Working paper No. 10004, Cambridge, MA: NBER.

Berman, E. and Laitin, D.D. 2005. Hard targets: theory and evidence on suicide attacks. Working Paper No. 11740. Cambridge, MA: NBER.

Berman, E. and Stepanyan, A. 2003. Fertility and education in radical Islamic sects: evidence from Asia and Africa. Mimeo, UC San Diego.

Boulding, K.E. 1970. *Beyond Economics: Essays on Society, Religion, and Economics*. Ann Arbor, MI: Ann Arbor Paperbacks.

Bruce, S. 1999. *Choice and Religion: A Critique of Rational Choice Theory*. Oxford and New York: Oxford University Press.

Chen, D. 2004. Club goods and group identity: evidence from Islamic resurgence during the Indonesian financial crisis. Working paper, University of Chicago.

Chesnut, R.A. 2003. *Competitive Spirits: Latin America's New Religious Economy*. New York: Oxford University Press.

Chiswick, B.R. 1983. The earnings and human capital of American Jews. *Journal of Human Resources* 18, 313–35.

Coleman, J.S. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 94(Supplement), S95–S120.

Delacroix, J. and Nielsen, F. 2001. The beloved myth: Protestantism and the rise of capitalism in nineteenth century Europe. *Social Forces* 80, 509–53.

Durkheim, E. 1915. *The Elementary Forms of the Religious Life*. Trans. K.E. Fields. New York: The Free Press, 1995.

Ehrenberg, R.G. 1977. Household allocation of time and religiosity: replication and extension. *Journal of Political Economy* 85, 415–23.

Ekelund, R.B., Hébert, R.F. and Tollison, R.D. 2006. *The Marketplace of Christianity*. Cambridge, MA: MIT Press.

Ekelund, R.B., Hébert, R.F., Tollison, R.D., Anderson, G.M. and Davidson, A.B. 1996. *Sacred Trust: The Medieval Church as an Economic Firm*. New York: Oxford University Press.

El-Gamal, M.A. 2006. *Islamic Finance: Law, Economics and Practice*. Cambridge: Cambridge University Press.

Finke, R., Guest, A.M. and Stark, R. 1996. Mobilizing religious markets: pluralism and religious participation in the empire state, 1850–1865. *American Sociological Review* 61, 203–18.

Finke, R. and Stark, R. 1988. Religious economies and sacred canopies: religious mobilization in American cities, 1906. *American Sociological Review* 53, 41–9.

Finke, R. and Stark, R. 1992. *The Churching of America, 1776–1990: Winners and Losers in Our Religious Economy*. New Brunswick, NJ: Rutgers University Press.

Freeman, R.B. 1986. Who escapes? The relation of church-going and other background factors to the socio-economic performance of black male youths from inner-city poverty tracts. In *The Black Youth Employment Crisis*, ed. R.B. Freeman and H.J. Holzer. Chicago: University of Chicago Press.

Gill, A. 1998. *Rendering Unto Caesar: The Catholic Church and the State in Latin America*. Chicago: University of Chicago Press.

Glaeser, E.L., Ward, B.A., Glaeser, E.L. and Ward, B.A. 2006. Myths and realities of American political geography. *Journal of Economic Perspectives* 20(2), 119–44.

Granovetter, M. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83, 1420–43.

Green, E.C. 2003. *Rethinking Aids Prevention: Learning from Successes in Developing Countries*. Westport, CT: Praeger.

Gruber, J. 2004. Pay or pray? The impact of charitable subsidies on religiosity. Working Paper No. 10374. Cambridge, MA: NBER.

Guiso, L., Sapienza, P. and Zingales, L. 2003. People's opium? Religion and economic attitudes. *Journal of Monetary Economics* 50, 225–82.

Hayek, F.A. 1988. *The Fatal Conceit: The Errors of Socialism*. Chicago: University of Chicago Press.

Heaton, P. 2006. Does religion really reduce crime? *Journal of Law and Economics* 49, 147–72.

Hungerman, D. 2005. Are church and state substitutes? Evidence from the 1996 Welfare Reform. *Journal of Public Economics* 89, 2245–67.

Iannaccone, L.R. 1984. Consumption capital and habit formation with an application to religious participation. Ph.D. thesis, University of Chicago.

Iannaccone, L.R. 1988. A formal model of church and sect. *American Journal of Sociology* 9 (Supplement), 241–68.

Iannaccone, L.R. 1990. Religious participation: a human capital approach. *Journal for the Scientific Study of Religion* 29, 297–314.

Iannaccone, L.R. 1991. The consequences of religious market regulation: Adam Smith and the economics of religion. *Rationality and Society* 3, 156–77.

Iannaccone, L.R. 1992. Sacrifice and stigma: reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy* 100, 271–92.

Iannaccone, L.R. 1995. Risk, rationality, and religious portfolios. *Economic Inquiry* 38, 285–95.

Iannaccone, L.R. 1998. An introduction to the economics of religion. *Journal of Economic Literature* 36, 1465–95.

Iannaccone, L.R. and Berman, E. 2006. Religious extremism: the good, the bad, and the deadly. *Public Choice* 128, 109–29.

Iyer, S. 2002. *Demography and Religion in India*. Oxford and New York: Oxford University Press.

Koenig, H.G., McCullough, M.E. and Larson, D.B. 2001. *Handbook of Religion and Health*. Oxford and New York: Oxford University Press.

Kuran, T. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.

Kuran, T. 2004. *Islam and Mammon: The Economic Predicaments of Islamism*. Princeton: Princeton University Press.

Lehrer, E.L. and Chiswick, C.U. 1993. Religion as a determinant of marital stability. *Demography* 30, 385–404.

Luidens, D. and Nemeth, R. 1994. Congregational vs. denominational giving: an analysis of giving patterns in the Presbyterian church in the United States and the reformed church in America. *Review of Religious Research* 36, 111–22.

Neuman, S. 1986. Religious observance within a human capital framework: theory and application. *Applied Economics* 18, 1193–202.

North, C.M. and Gwin, C.R. 2004. Religious freedom and the unintended consequences of the establishment of religion. *Southern Economic Journal* 71, 103–17.

Olds, K. 1994. Privatizing the church: disestablishment in Connecticut and Massachusetts. *Journal of Political Economy* 102, 277–97.

Oslington, P., ed. 2003. *Economics and Religion*. Cheltenham: Edward Elgar.

Putnam, R.D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Richardson, G. 2005. Craft guilds and Christianity in late-medieval England: a rational-choice analysis. *Rationality and Society* 17, 139–89.

Samuelsson, K. 1993. *Religion and Economic Action: The Protestant Ethic, the Rise of Capitalism, and the Abuses of Scholarship*. Toronto: University of Toronto Press.

Schultz, T.W. 1961. Investment in human capital. *American Economic Review* 51, 1–17.

Smith, A. 1759. *The Theory of Moral Sentiments*. Ed. D.D. Raphael and A.L. Macfie, Indianapolis: Liberty Fund, 1984.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Indianapolis: Liberty Classics, 1981.

Sobel, J. 2002. Can we trust social capital? *Journal of Economic Literature* 40, 139–54.

Sosis, R.H. and Ruffle, B.J. 2003. Religious ritual and cooperation: testing for a relationship on Israeli religious and secular kibbutzim. *Current Anthropology* 44, 713–22.

Stark, R. 2005. *The Victory of Reason: How Christianity Led to Freedom, Capitalism, and Western Success*. New York: Random House.

Stark, R. and Bainbridge, W.S. 1985. *The Future of Religion*. Berkeley and Los Angeles: University of California Press.

Stark, R. and Bainbridge, W.S. 1987. *A Theory of Religion*. Bern: Peter Lang Publishing.

Stark, R. and Finke, R. 2000. *Acts of Faith: Explaining the Human Side of Religion*. Berkeley and Los Angeles: University of California Press.

Swedberg, R. 1990. *Economics and Sociology: Redefining their Boundaries: Conversations with Economists and Sociologists*. Princeton: Princeton University Press.

Tawney, R.H. 1926. *Religion and the Rise of Capitalism*. Repr., with new introduction by Adam B. Seligman. New Brunswick, NJ, and London: Transaction, 1998.

Ulbrich, H. and Wallace, M. 1983. Church attendance, age, and belief in the afterlife: some additional evidence. *Atlantic Economic Journal* 11, 44–51.

Voas, D., Olson, D.V.A. and Crockett, A. 2002. Religious pluralism and participation: why previous research is wrong. *American Sociological Review* 67, 212–30.

Waite, L.J. and Lehrer, E.L. 2003. The benefits from marriage and religion in the United States: a comparative analysis. *Population and Development Review* 29, 255–75.

Warner, R.S. 1993. Work in progress toward a new paradigm in the sociology of religion. *American Journal of Sociology* 98, 1044–93.

Weber, M. 1920. *The Protestant Ethic and the Spirit of Capitalism*. Trans. S. Kalberg, 3rd edn. Los Angeles: Roxbury, 2002.

Weber, M. 1963. *The Sociology of Religion*, trans. E. Fischoff. Boston: Beacon Press, 2002.

Young, L.A. 1997. *Rational Choice Theory and Religion: Summary and Assessment*. New York: Routledge.

How to cite this article

Iannaccone, Laurence R. and Eli Berman. "religion, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000231> doi:10.1057/9780230226203.1419

The New Palgrave Dictionary of Economics Online

rent control

Stephen Malpezzi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Rent control generically describes a range of regulations governing rents, as well as related contract features such as security of tenure and required maintenance. There is debate in the literature about the efficacy of controls based on (1) whether the housing market is best modelled as a competitive market, or one where landlords have market power; and (2) whether regulators have sufficient information and appropriate mechanisms to improve imperfect market outcomes. Many empirical studies find that rent controls score badly as redistributive systems. Many basic questions, especially regarding dynamic effects on the supply of housing, have yet to be credibly answered.

Keywords

asymmetrical information; housing markets; housing supply; loss aversion; market power; property taxation; rent control; tax incidence

Article

Rent controls of one kind or another affect roughly 40 per cent of the world's urban dwellers. Rent control is usually thought of as a policy applied to private markets, but publicly provided housing (for example, much urban housing in Russia and in China) is also subject to controls. In addition to regulations governing rents, controls often address additional contract features such as security of tenure and required maintenance. Actual rent control regimes vary enormously in their design and in their effects.

History

Rent controls are often instituted in response to a major economic or political shock which limits the responsiveness of the housing market. Controls were introduced in the Second World War in Europe,

North America, and, under European colonial influence, much of the developing world as well. Most jurisdictions in the United States and Canada removed controls in the post-war years; however, controls of varying degrees of stringency were maintained in much of Europe and the developing world. Poorer countries tend to have more stringent regimes, though enforcement patterns vary at least as much as *de jure* codes.

Exactly why controls exist, or at least are retained after wartime or similar emergencies are clearly over, is still debated. An obvious point of political economy is that there are more tenants than landlords; but there is little correlation between the fraction of a country's population renting and the stringency of controls, according to Malpezzi and Ball (1991). On the other hand the relatively small number of US cities with rent control tend to have large renter populations, notably New York. Fischel (2001) presents several interesting conjectures about the political economy of controls, notably that homeowners might ally with landlords to oppose controls because they fear negative spillovers from reduced maintenance of stringently controlled buildings, as well as shifting property tax burdens. The strong opposition to relaxation of controls in New York, while nearby uncontrolled jurisdictions see little agitation for imposition, might be analysed in Kahneman and Tversky's (1979) loss aversion framework. A clear understanding of the political economy of controls awaits future research.

Features

One key feature is whether regulations set the level of rents, or control increases in rent. Others include how controlled rents are adjusted for changes in costs (with cost pass-through provisions, or adjustments for inflation); how close the adjustment is to changes in market conditions; how it is applied to different classes of units; or whether rents are effectively frozen over time. Other key provisions which vary from place to place include breadth of coverage, how initial rent levels are set, treatment of new construction, whether rents are reset for new tenants, and tenure security provisions. Rent control's effects can vary markedly depending on these specifics, and on market conditions, as well as enforcement practices.

Theory

Rent control can be analysed as an implicit tax on housing capital. In the simplest case, where imposition of controls reduces the price of an existing stock of rental housing, the tax is borne by landlords for the benefit of tenants. Over time, as the market adjusts to controls, the incidence of the ‘tax’ becomes more complicated.

Much of the debate in the literature about the efficacy of controls stems from maintained assumptions about the nature of the housing market, and the regulator, in turn. The first question is: is the housing market best modelled as a competitive market, or one where landlords have market power, for example from information asymmetries? If the former, then clearly rent control reduces the efficiency of the rental market, although the magnitude of such effects can be debated, and distributional arguments remain. If the latter, a second question readily follows: does the regulator have sufficient knowledge, and an appropriately designed set of regulations, to improve on the market outcome? Arnott (1995) ably reviews the contrast between competitive and ‘market power’ theoretical approaches, and also discusses why it is so difficult to resolve these issues empirically.

Whatever one's priors about market power, there are many alternative adjustment mechanisms which can arise in a notionally controlled market. Four of the adjustments can be embodied in rent control laws: (*a*) indexing (keeping real rents constant), (*b*) reassessment for new tenants, (*c*) differential pricing of new and existing units, and (*d*) differential pricing for upgraded units. Three are market responses which many would generally consider undesirable outcomes, namely, (*e*) outright evasion, (*f*) side payments such as key money, (*g*) adoption by tenants of maintenance expenditures, and (*h*) accelerated depreciation and abandonment, (*i*) distortions in consumption, not only in the composite housing services but also crowding, length of stay, mobility and tenure choice.
Key questions are: What are the efficiency losses from controls? Are the benefits to some tenants worth the costs? Do they redistribute income as intended? Several broad approaches have been taken in the empirical literature to answer these questions.

Static analysis

One of the first published studies of the costs and benefits of rent control is Olsen's. Using data from New York City in 1968, Olsen (1972) found the average controlled rent for an apartment was $999 a year (for comparison, the average income was $6,229). Olsen first estimated how much the controlled units would rent for in the absence of controls. The average estimated uncontrolled rent for controlled units was $1,405, implying a subsidy (static cost to landlords) of $406. Olsen next estimated how much households in controlled units would spend in the uncontrolled market, given their income and family size. The average estimated market expenditure for the controlled households was $l,470, indicating that they consumed slightly less housing than they would have in the free market. Olsen then computed the economic benefit of rent control to each surveyed controlled tenant using a simple consumer surplus model. Olsen's estimate of the average net benefit is $213, little more than half the gross subsidy of $406. Examining the distribution of these benefits among controlled households, Olsen found the annual benefit decreased by about one cent for every dollar of additional income, $9 a year of head's age, and $69 per additional household member. Rent control in New York City in 1968 appears to redistribute income, but very weakly, and in no way proportional to its cost.
A number of other studies have been carried out along these lines (Malpezzi and Ball, 1991, review several). For example, in Cairo, Egypt, monthly rents for a typical unit are less than 40 per cent of estimated market rents. But ‘key money’ (illegal upfront payments to landlords) and other side payments make up about a third of the difference.

Dynamic analysis

Murray et al. (1991) is an early study of rent control dynamics. A simulation model was used to predict the time path of rents and the quantity of housing services given alternative control regimes. The magnitude of the effects varied substantially with details of the regime. In general, Murray et al. find that dynamic losses can be substantial; in fact they outweigh static consumer's surplus losses by as much as a factor of 18. Generally tenant benefits are were substantially less than landlord costs; the transfer efficiency in three representative cases ranged from 65 per cent to 83 per cent.
Another potential dynamic effect of controls, with possible spillovers to labour markets, is reduced

household mobility. Several studies, such as Munch and Svarer (2002) and Simmons-Moseley and Malpezzi (2006), find that household mobility is inversely related to the estimated net benefits received from a control regime.

Given their potential importance, dynamic effects of controls are understudied. For example, no one has yet credibly analysed the effects of controls on the aggregate supply of housing. Reviews of the theoretical literature by Arnott (1995), and of the empirical literature by Turner and Malpezzi (2003), point out that empirical work lags theory in this area. Malpezzi and Ball (1991) did find that countries with stricter rent control regimes invested less in housing, in the aggregate; but while the analysis accounted for income and demographics on the demand side, other potential constraints on housing supply (for example, land use constraints, financial constraints) were not well specified. Since these may well be correlated with the strength of controls, these results cannot be viewed as the final word. Given the myriad ways real world regimes work, and the variety of possible ways around controls (legal and illegal) the size of the net aggregate effect on supply remains unknown.

Distributional issues

Such evidence as exists casts doubt on controls' effectiveness as income transfer mechanisms. In Kumasi and Rio, benefits were found to be somewhat ‘progressive’ in the common sense of the term (larger benefits to poorer households). On the other hand, in Cairo and Bangalore, no relationship was found between the benefits gained from reduced rent and household income, because rent control is not well targeted to low-income groups (Malpezzi and Ball, 1991). In fact, research on New York controls by Glaeser and Luttmer (2003) suggests that previous research largely *underestimates* the misallocation of housing under controls, and that, because of excess demand for controlled units, benefits are more or less randomly distributed.

Another questionable assumption behind redistribution as a rationale for controls is the notion that landlords are rich and tenants are poor. In Cairo, Kumasi and Bangalore, the income of tenants and landlords was compared; and, while the landlords' median income was higher in all three, there was significant overlap. In Cairo, for example, about 25 per cent of tenants had incomes that were higher than the landlord median, and about 25 per cent of landlords had incomes lower than the tenant median. There is no guarantee the transfers will occur only from high-income landlords to low-income tenants. Most careful empirical studies find that at least some tenants are, on balance, worse off under controls because of constraints on housing consumption. And in markets with significant uncontrolled sectors, rent controls can drive up the price of uncontrolled housing, an important unintended consequence further complicating the incidence of its costs.

See Also

- housing policy in the United States
- housing supply

Bibliography

Arnott, R. 1995. Time for revisionism on rent control. *Journal of Economic Perspectives* 9(1), 99–120.

Fischel, W.A. 2001. *The Homevoter Hypothesis: How Home Values Influence Local Government Taxation, School Finance, and Land-Use Policies*. Cambridge, MA: Harvard University Press.

Glaeser, E.L. and Luttmer, E.F.P. 2003. The misallocation of housing under rent control. *American Economic Review* 93, 1027–46.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Malpezzi, S. and Ball, G. 1991. Rent control in developing countries. Discussion Paper No. 129. Washington, DC: World Bank.

Munch, J.R. and Svarer, M. 2002. Rent control and tenancy duration. *Journal of Urban Economics* 52, 542–60.

Murray, M.P., Rydell, C.P., Barnett, C.L., Hillstead, C.E. and Neels, K. 1991. Analyzing rent control: the case of Los Angeles. *Economic Inquiry* 29, 601–25.

Olsen, E.O. 1972. An econometric analysis of rent control. *Journal of Political Economy* 80, 1081–100.

Simmons-Mosley, T.X. and Malpezzi, S. 2006. Household mobility in New York City's regulated rental housing market. *Journal of Housing Economics* 15, 38–62.

Turner, B. and Malpezzi, S. 2003. A review of empirical evidence on the costs and benefits of rent control. *Swedish Economic Policy Review* 10, 11–56.

How to cite this article

Malpezzi, Stephen. "rent control." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000099> doi:10.1057/9780230226203.1421

The New Palgrave Dictionary of Economics Online

rent seeking

Gordon Tullock
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Rent seeking ’ refers to the investment of resources in efforts to create monopolies. Such investments impose a social cost (which may outweigh the benefit to the monopolist) because they are unproductive. That cost is greater than the mere cost of lobbying by special interests for privilege when the privilege is conferred in a way that is economically inefficient but politically feasible (which is often true of regulations). Research on rent seeking has demonstrated that the true social costs of promoting special interests thus greatly exceed the deadweight costs of the distortions introduced into the economy.

Keywords

consumer surplus; corruption; deadweight loss; Krueger, A. O.; lobbying; monopoly; Morgan, J. P.; political influence; pseudo-equilibrium; rent; rent seeking; social cost; special interests; Tullock, G.

Article

The term ‘rent-seeking’ was introduced by Ann O. Krueger (1974), but the relevant theory had already been developed by Gordon Tullock (1967). The basic and very simple idea is best explained by reference to Figure 1. On the horizontal axis we have as usual the quantity of some commodity sold, on the vertical axis its price. Under competitive conditions the cost would be the line labelled PP and that would also be its price. Given a demand curve, DD, quantity Q would be sold at that price. If a monopoly were organized, it would sell Q’ units at a price of P’.

Figure 1

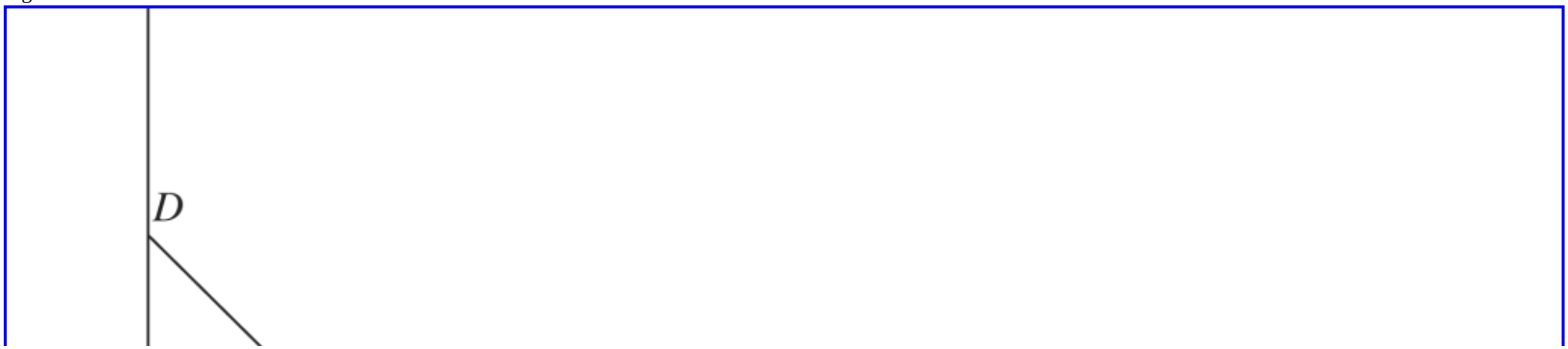

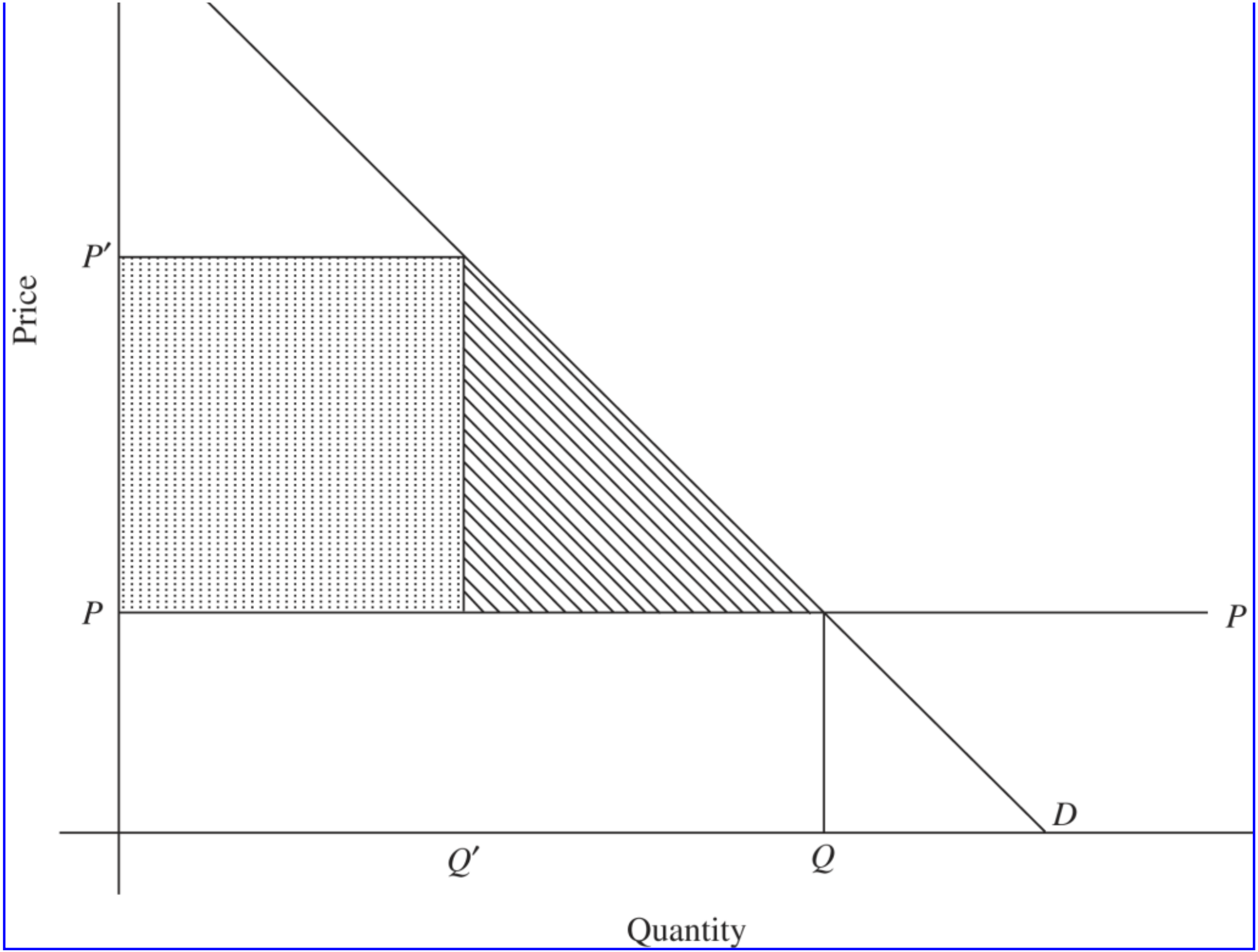

The traditional theory of monopoly argued that the net loss to society is shown by the shaded triangle, which represents the consumer surplus that would have been derived from the

purchase of those units between Q’ and Q, that are now neither purchased nor produced. The dotted rectangle, on the other hand, has traditionally been regarded simply as a transfer from the consumers to the monopolist. Since they are all members of the same society, there is no net social loss from this transfer.

This argument tends to annoy students of elementary economics (because they don't like monopolists), but until the development of the work on rent seeking it was nevertheless thought to be correct by most economists. Its basic problem, however, is that it assumes that the monopoly is created in a costless manner, perhaps by an act of God, whereas in fact real resources are used to create monopolies.

Most discussion of rent seeking has tended to concentrate on those monopolies that are government created or protected, probably because these are observed to be the commonest and strongest. It should be kept in mind, however, that purely private monopolies are possible – indeed, some actually exist. Concentration on government-created monopolies (or restrictions of various sorts that increase certain peoples' income) is probably reasonable, granted the contemporary frequency of such activities. Nevertheless, as we point out below there are certain significant areas where private rent seeking causes net social loss.

In the initial work both of Tullock and Krueger it was assumed that profit-seeking businessmen would be willing to use resources in an effort to obtain a monopoly, whether it was privately or government sponsored, up to the point where the last dollar so invested exactly counterbalanced the improved probability of obtaining the monopoly. From this it was deduced that the entire dotted rectangle (Figure 1) would be exhausted. Although this assumption is open to question (see Tullock, 1980), for the time being we will continue to assume that in effect there is no transfer from purchasers to the monopolist, but simply a social loss which comes from the fact that resources have been invested in unproductive activity, i.e. the negatively productive activity of creating a trade restriction of some sort. Theoretical reasons exist for believing that this assumption probably does not fit perfectly anywhere, but it is just as likely to overestimate as to underestimate the social cost; it will be discussed more thoroughly below.

To quote an aphorism frequently used in rent seeking: ‘the activity of creating monopolies is a competitive industry.’ For this reason it is anticipated that quite a number of people at any given time are putting at least some resources into an effort to secure a monopoly, only some of whom are successful. The situation is like a lottery, in which many people buy lottery tickets, a few win a very large amount of money and the rest lose, perhaps large or small amounts, depending on how much they have committed. In almost all existing lotteries, of course, the total investment of resources by the gamblers is considerably greater than the total payoff, whereas here it is still assumed that total resources committed to rent-seeking equal the total monopoly profits.

Thus the activity of creating monopolies could both absorb very large resources, particularly those resources that take the form of exceptionally talented individuals who devote their attention to this difficult and highly rewarded activity, and lead to considerable redistribution of wealth in the community. Suppose that ten different lobbyists go to Washington representing ten different associations, and each spends one million dollars over the course of a couple of years in the hope of influencing Congress to provide them with a monopoly. Only one of the lobbyists is successful and the monopoly turns out to have a present discounted value of ten million dollars. There is a substantial redistribution of resources from the unsuccessful lobbyists to the successful.

This substantial redistribution has occurred simultaneously with a considerable waste of resources in general, both because these highly intelligent people could otherwise be doing something of higher productivity and because the economy's use of resources has been further distorted by the creation of the monopoly. Further, although so far the discussion has been primarily about monopoly, actually very many possible interventions in the market process raise the same problem. A simple maximum or minimum price may have very large redistributive effects and the people who thus benefit may put considerable resources into receiving them. Of course there are many situations in which one lobbyist is pushing for a particular restriction and another lobbyist is pushing against it. The second activity is sometimes called ‘rent avoidance’, but it is costly and of course would not exist if there were not also rent seeking activity.

Another area is simple direct transfers. A tax on A for the purpose of paying B will lead to lobbying activity for the tax on the part of B and against it on the part of A. The total of these two lobbying activities could very well equal the total amount transferred (or prevented from being transferred), although one or other of these entrepreneurs will of course gain if his lobby is successful. Assume that A puts in $50 for lobbying to get $100 from B and B puts in $50 lobbying against that. Regardless of the outcome, one party will gain $50 from his lobbying. Society has lost $100.

Of course it is not true that everyone in society is in an equally good position to seek rents. Some kinds of interest are more readily organized than others and we would anticipate that they would win. There are however very many such interests and anyone who spends any time in Washington quickly realizes that there is a major industry engaged in just this kind of activity.

Actual social cost however is clearly very much greater than the mere cost of the various lobbying organizations in Washington. In particular it is normally necessary for the rent-seeking group to undertake directly productive activities in a way that is markedly inefficient, because it is necessary to introduce a certain element of deception into the process. In 1937, when the US Civil Aeronautics Board was organized, it would not have been politically feasible to put a direct tax on purchasers of airline tickets and use it to pay off the stockholders of the airline companies. Regulation, which has a similar effect but at a very much higher cost to the users of airlines per dollar of profit to the owners, was however, politically possible. The necessity of using inefficient methods of transferring funds to the potential beneficiary, because the efficient methods would be just too open and above board, is often one of the major costs of rent seeking. The rent avoidance lobbyist would have had too easy a time if the proposal had been a tax on uses of airlines for the benefit of the stockholders.

Note that in this case the argument against rent seeking turns out also to be an argument against political corruption. Suppose you are in a society which has an exchange control

system and that it is possible to buy foreign currency by bribing an official in the exchange control office. This is the kind of situation dealt with by Krueger (1974), who was able to obtain a measure of the total social cost in Turkey and India where the amounts of the necessary bribes were well known; the cost varied from 7–15 per cent of the total volume of transactions.

Traditionally economists have tended to view this kind of bribery as in itself desirable, because it gets around an undesirable regulation. However, it leads to rent seeking. In this case the rent seeking does not come from the users of the permits but from the competition to get into the position where you can receive the bribe. Throughout the underdeveloped world, large numbers of people take fairly elaborate educational programmes which have no real practical value for their future life and engage in long periods of complicated political manoeuvring in hope that they will be appointed, let us say, a customs inspect in Bombay. Since these young men have a free career choice presumably the expected returns from this career are the same as in any other. The difference is that a doctor, say, begins earning money immediately on completing medical school whereas the young man who has studied economics and is now trying to obtain appointment as customs inspector will have a considerable period of time in which he is not appointed at all. Indeed, there will probably be enough such candidates that he has only perhaps one chance in five of being so appointed. The total cost of the rent seeking is the inappropriate education and the political manoeuvring of the five people of whom only one is appointed.

So far we have assumed that the total cost of rent seeking is the present discounted value of the income stream represented by the dotted rectangle in Figure 1. This assumes a special form for the function which ‘produces’ the monopoly or other privilege. It must be linear, with each dollar invested having exactly the same payoff in probability of achieving the monopoly as the previous dollar (Tullock, 1980). Most functions do not have this form, instead they are either increasing or decreasing cost functions.

If the organizing of private monopolies, or of influencing the government into giving you public monopolies, is subject to diseconomies of scale, then total investment in rent seeking will be less than the total value of the rents derived even if we assumed a completely competitive market with completely free entry. When there are economies of scale the situation is even more unusual. Either there is no equilibrium at all or there is a pseudo-equilibrium, in which total investment to obtain the rents is greater than the rents themselves. This is called a pseudo-equilibrium, because although it meets all the mathematical requirements for an equilibrium, it is obviously absurd to assume that people would, to take a single example, pay $75.00 for a 50–50 chance of $100.

Obviously, what is needed is empirical research, and an effort to measure the production functions appropriate to rent seeking. So far, however, no one has been able to develop a very good way of making such measurements. It seems likely that it would be easier to measure the costs of generating political influence than of private monopolies, if only because many of the expenditures used to influence the government appear in accounts in various places. The costs of private monopolies on the other hand, tend to be much more readily concealed. This does not mean that they do not exist.

The reader has no doubt been wondering what is wrong with rents and why we concern ourselves deeply with rent seeking. The answer to this is that the term itself is an unfortunate one. Obviously, we have nothing against rents when they are generated by, let us say, discovering a cure for cancer and then patenting it. Nor do we object to popular entertainers like Michael Jackson earning immense rents on a rather unusual collection of natural attributes together with a lot of effort on his part to build up his human capital. On the other hand, we do object to the manufacturer of automobiles increasing the rent on his property, and his employees increasing the rent on their union memberships, by organizing a quota against imported cars. All of these things are economic rents, but strictly speaking the term ‘rent seeking’ applies only to the latter. Its meaning might be expanded to seeking rents from activities which are themselves detrimental. The man seeking a cure for cancer is engaged in an activity which clearly is not detrimental to society. Thus we may observe immediately that activities aimed at deriving rents cover a continuum, but that the term ‘rent seeking’ is only used for part of that continuum.

The analysis of ‘rent seeking’ has been one of the most stimulating fields of economic theory in recent years. The realization that the explanation of the social cost of monopoly which was contained in almost every elementary text in economics was wrong, or at the very least seriously incomplete, came as quite a surprise. Revision of a very large part of economic theory in order to take this error into account is necessary. And history also needs to be revised. That J.P. Morgan was an organizer of cartels and monopolies during most of his life is well known, as is the fact that he received very large fees for this, fees which were part of the rent seeking cost of generating these monopolies. It is possible to argue that as a stabilizing factor in the banking system, Morgan more than repaid to the United States the social cost of his monopolistic activities in industry. But that there was a very large rent seeking cost is obvious. This cost is in addition to the deadweight cost of the monopolies.

To date, research on rent seeking has to a considerable extent changed our way of looking at things. We now talk of a great deal of government activity as rent seeking on the part of somebody or other. It was known that special interest existed, but we have traditionally tended to underestimate its cost greatly because we looked only at the deadweight costs of the distortion introduced into the economy. The realization that the actual cost is much greater socially, that the large-scale lobbying industry is truthfully a major social cost, is new although presumably, at all times, anyone who thought about the matter must have realized that these highly talented people could produce more in some other activity.

See Also

- bribery
- directly unproductive profit-seeking (DUP) activities

Bibliography

Buchanan, J., Tollison, R. and Tullock, G., eds. 1980. *Toward a Theory of the Rent-Seeking Society*. College Station, Texas: Texas A and M University Press.

Krueger, A.O. 1974. The political economy of the rent-seeking society. *American Economic Review* 64, 291–303.

Tullock, G. 1967. The welfare cost of tariffs, monopolies, and theft. *Western Economic Journal (now Economic Inquiry)* 5, 224–32.

Tullock, G. 1980. Efficient rent seeking. In Buchanan, Tollison and Tullock, eds., 91–112.

How to cite this article

Tullock, Gordon. "rent seeking." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000100>
doi:10.1057/9780230226203.1422

The New Palgrave Dictionary of Economics Online

rent

Armen A. Alchian
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Rent’ is the payment for use of a resource, whether it be land, labour, equipment, ideas, or even money. The term is often restricted to payment for use of land or equipment. ‘Economic rent’ is payment for use of any resource whose supply is fixed. Rent serves a social purpose because market levels of rent indicate which uses of fixed resources are the highest valued, and direct such resources to those uses. ‘Monopoly rent’ is paid to producers in markets that are artificially restricted; it may be dissipated by ‘rent seekers’ who compete for monopoly status.

Keywords

composite quasi-rent; differential rents; economic rent; land; land tax; Marshall, A.; monopoly rents; producer surplus; quasi-rent; rent; rent seeking; Ricardian rents; Ricardo, D.

Article

‘Rent’ is the payment for use of a resource, whether it be land, labour, equipment, ideas, or even money. Typically the rent for labour is called ‘wages’; the payment for land and equipment is often called ‘rent’; the payment for use of an idea is called a ‘royalty’; and the payment for use of money is called ‘interest’. In economic theory, the payment for a resource where the availability of the resource is insensitive to the size of the payment received for its use is named ‘economic rent’ or ‘quasi-rent’ depending on whether the insensitivity to price is permanent or temporary.
To early economists, ‘rent’ meant payments for use of land; Ricardo, in particular, called it the payment for the ‘uses of the original and indestructible powers of the soil’ (Ricardo, 1821, p. 33). Subsequently, in recognition that a distinctive feature of what was called ‘land’ was its presumed indestructibility (i.e. insensitivity of amount supplied to its price), the adjective ‘economic’ was applied to the word ‘rent’ for any resource the supply of which is indestructible (maintainable for ever at no cost) and non-augmentable, and hence invariant to its price. In the jargon of economics, the quantity of present and future available supply is completely inelastic with respect to price, a situation graphically represented by a vertical supply line in the usual ‘Marshallian’ price-quantity graphs.

Economic rent

The concept of ‘economic rent’ is graphically depicted by the standard demand and supply lines in Figure 1 with a vertical supply curve (quantity supplied invariant to price) at the amount X_r. At all prices the supply is constant. The entire return to the resource is an ‘economic rent’. If the aggregate quantity of such resources may in the future be increased by production of more indestructible units of the resource in response to a higher price (but the amount available at any moment is fixed regardless of the rent for its services), the supply line at the current moment is vertical. The supply curve for future amounts slopes upward from the existing amount, as depicted by the line FF in Figure 1. The long run rent would be P_r and the equilibrium stock would be X_r: at that equilibrium stock the ‘market supply’ (in Marshall's terminology) would be a vertical line. Thus, the supply of indestructible units would have depended on past anticipated prices about the present prices, but the supply of current units would be insensitive to the current price or rent. The return could be called

'economic rent', except that no convention has been developed with respect to the terminology for this situation of indestructible but augmentable resources.

Figure 1

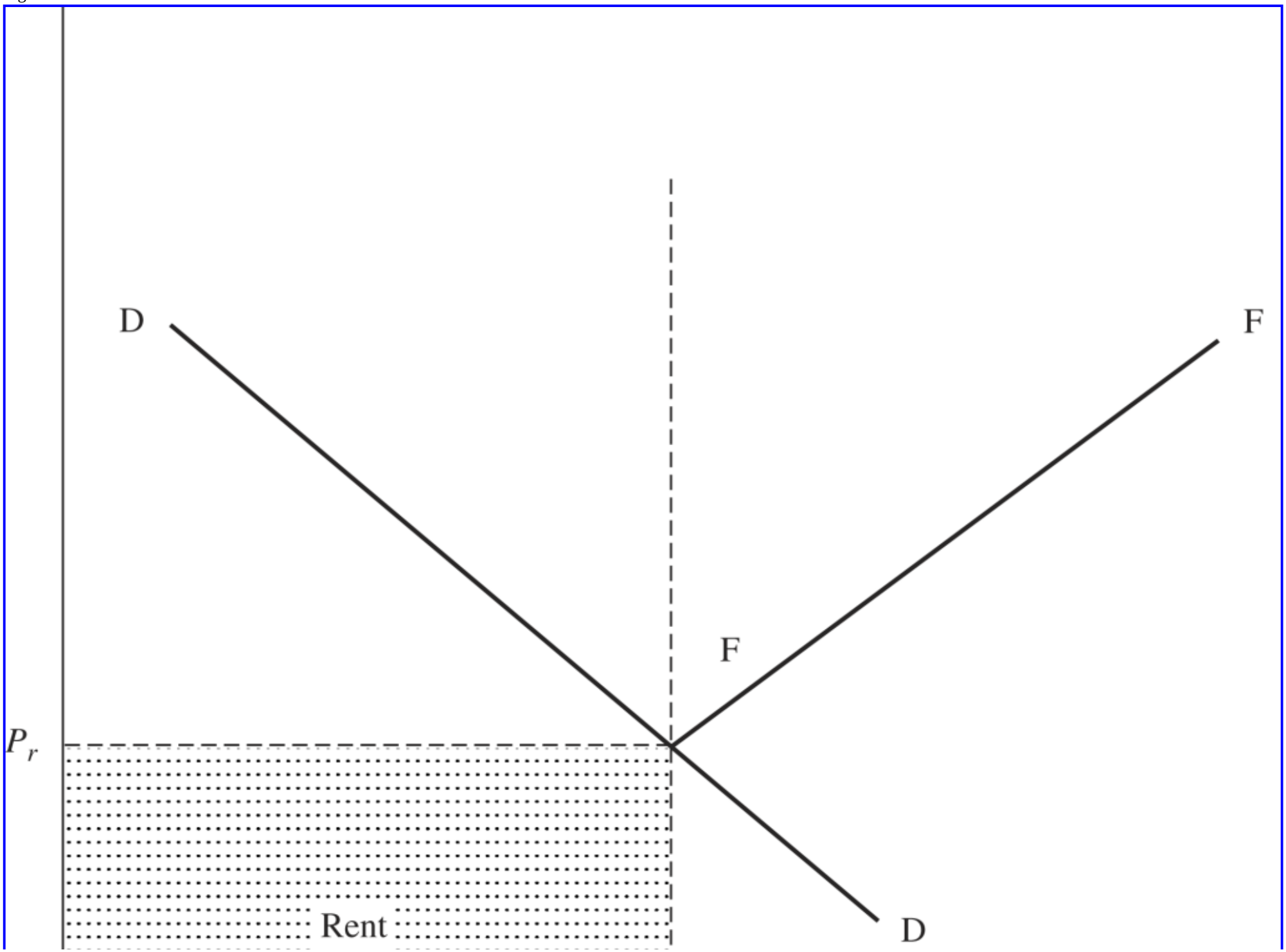

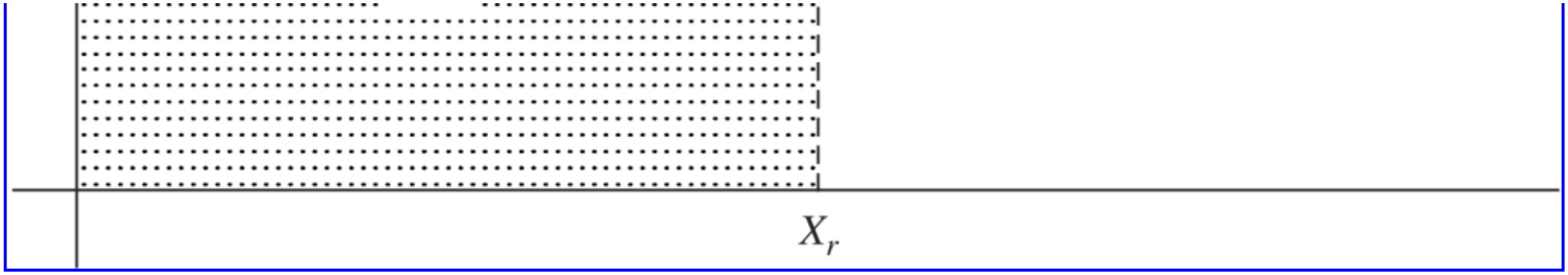

Quasi-rent

Closely related to ‘economic rent’ is ‘quasi-rent’, a term apparently initiated by Alfred Marshall (Marshall, 1920, pp. 74, 424–6). Because virtually every existing resource is unresponsive to a change in price for at least some very small length of time, the return to every resource is like an ‘economic rent’ for at least a short interval of time. In time, the supplied amount will be altered, either by production or non-replacement of current items. Yet, the fact that the amount available is not instantly affected by price led to the term ‘quasi-rent’, which denotes a return, variations in which do not affect the current amount supplied to the demander but do affect the supply in the future.

If a rental (payments) stream to an existing resource is not sufficient to recover the costs incurred in its production the durability of that existing resource will nevertheless enable the resource to continue to provide services, at least for some limited time. In other words, because of the resource's durability it will continue for some interval to yield services even at a rent insufficient to recover its cost of production, but sufficient for current costs of use including interest on its salvage value (which is its highest value in some other use). Any excess over those current costs is a ‘quasi-rent’.

Quasi-rent resembles an ‘economic rent’ in that it exceeds the amount required for its current use, albeit temporarily – except that a flow of rents that did not cover all ‘quasi-rent’ would preserve it for only a finite future interval, after which the resource would be diminished until not worth more than its salvage value. If the resource received a payment exceeding all the initially anticipated and the realized costs of production and operation, it will have achieved a profit, that is, more than pure interest on the resource's investment cost. The question exists as to whether ‘quasi-rent’ means just that portion of the rent in excess of the minimum operating costs over the remaining life of the asset, or all the excess, including profits, if any. Convention seems still to be missing. Marshall seems to have excluded interest on the investment as well as any profits from what he called quasi-rents, so that any excess over variable costs of operation were partitioned into quasi-rents, interest on investment and profits (Marshall, 1920, pp. 412, 421, 622).

Composite quasi-rent

‘Composite quasi-rent’ was another important, but subsequently ignored, concept coined by Marshall (Marshall, 1920, p. 626). When two separately owned resources are so specific to each other that their joint rent exceeds the sum of what each could receive if not used together, then that joint rent to the pair was called ‘composite quasi-rent’. The two resources presumably already had been made specific to each other (worth more together than separately) by some specializing interrelated investments. Marshall cited the example of a mill and a water power site, presumably a mill built next to a dam to serve the mill, each possibly separately owned. One or both of the parties could attempt to hold up or extract a portion of the other party's expropriable quasi-rent. It is interesting to quote Marshall about this situation:

> The mill would probably not be put up till an agreement had been made for the supply of water power for a term of years; but at the end of that term similar difficulties would arise as to the division of the aggregate producer's surplus afforded by the water power and the site with the mill on it. For instance, at Pittsburg when manufacturers had just put up furnaces to be worked by natural gas instead of coal, the price of the gas was suddenly doubled. And the history of mines affords many instances of difficulties of this kind with neighbouring landowners as to rights of way, etc., and with the owners of neighbouring cottages, railways and docks (Marshall, 1920, p. 454).

A reason for attributing importance to the concept of ‘composite quasi-rent’ is now apparent. If it arises with resources that have been made specific to each other in the sense that the service value of each depends on the other's presence, the joint value of composite quasi-rent might become the object of attempted expropriation by one of the parties, especially by the one owning the resource with controllable flow of high alternative use value. To avoid or reduce the possibility of this behaviour, a variety of preventative arrangements,

contractual or otherwise, can be used prior to making the investments in resources of which at least one will become specific to the other. These include, among a host of possibilities: joint ownership, creation of a firm to own both, hostages and bonding, reciprocal dealing, governmental regulation, and use of insurers to monitor uses of interspecific assets. This is not the place to discuss these arrangements, beyond asserting that without the concept of ‘quasi-rent’ and especially ‘expropriable quasi-rent’ – which Marshall called ‘composite quasi-rent’ – a vast variety of institutional arrangements would otherwise be inexplicable as a means of increasing the effectiveness of economic activity.

Though Marshall briefly mentioned similar problems between employers and employees, I have not found any subsequent exposition by him about the precautionary contractual arrangements and institutions that attempt to avoid this problem, which has become a focus of substantial important research on what is called, variously, ‘opportunism, shirking, expropriable quasi-rents, principal–agent conflicts, monitoring, problems of measuring performance, asymmetric information, etc.’.

Ricardian rent

The rents accruing to different units of some otherwise homogeneous resource may differ and result in differences of rent over the next most valued use, differences that are called ‘Ricardian rents’. This occurs where the individual units, all regarded as of the same ‘type’ in other uses, are actually different with respect to some significant factor for its use *here*, though this factor, which is pertinent *here*, is irrelevant in any other uses. Examples of such factors can be location, special fertility, or talent that is disregarded in the other potential uses. For some questions, the inaccurate ‘homogenization’ can be a convenient simplification, but for explaining each unit's actual rents, it can lead to confusion and misunderstanding. The service value, hence rents, for the use of the services here may differ, though equal in every relevant respect elsewhere. Whether the specific use uniqueness is created by natural talent or sheer accident, the special differences in use value here imply differences in payments, often called ‘Ricardian rents’ to distinguish them from differences in rents (prices) obtained because of monopolizing or unnatural restrictions on any potential competitors, which may lead to higher rents, called ‘monopoly rents’ for the protected resources.

Differential rents

‘Differential rents’ are another category representing rent differences in a sort of reverse homogeneity. Units of resource that are equal with respect to their value in use *here* differ among themselves in their values of use elsewhere. This can be represented graphically as in Figure 2. The differential rents of successive units are represented by the differences between the price line and the curve RR, which arrays the units from those with the lowest alternative use values to the highest, a curve labelled RR. The arrayed units are not homogeneous for uses elsewhere, so even if identical for use here, calling them successive units of the same good is misleading. They are not totally homogeneous; if they were, each unit would have the same as any other unit's use value and rent elsewhere. A curve like RR is equivalent to Marshall's particular expenses curve, which arrayed units according to each individual unit's cost of production, or use value elsewhere, from lowest to highest (Marshall, 1920, p. 810n). The difference between price or rent here and the value on the RR curve is called ‘producers' surplus’ or ‘differential rent’. In sum, ‘Ricardian rents’ indicate differences in rents to units that are equal in their best alternative use values, but different in their rent value here, while ‘differential rents’ are the premia to units that are the same value here but different in their best alternative use values.

Figure 2

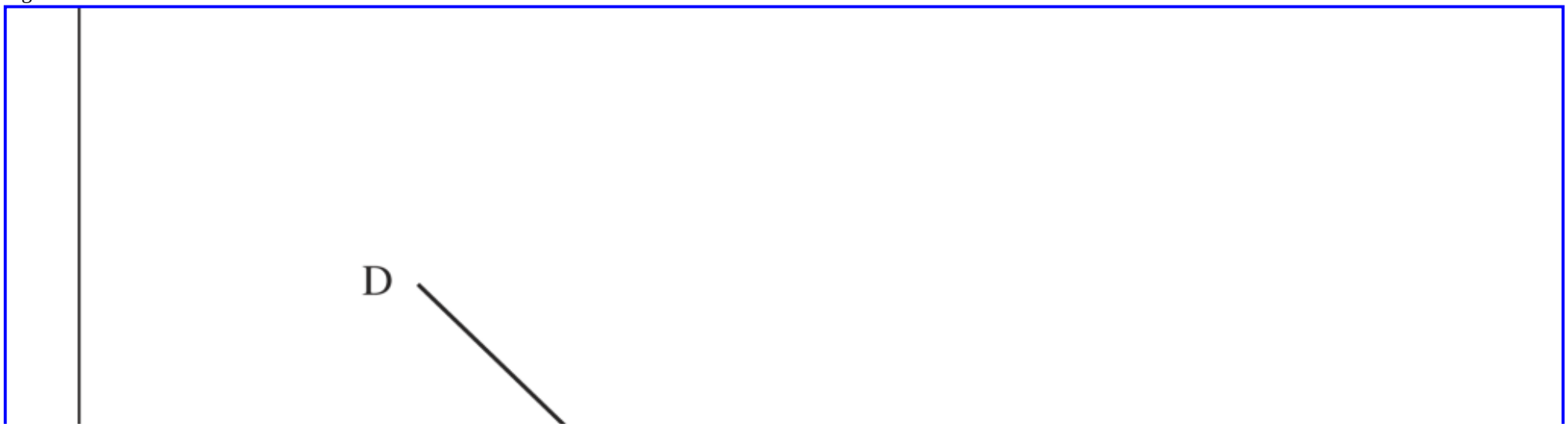

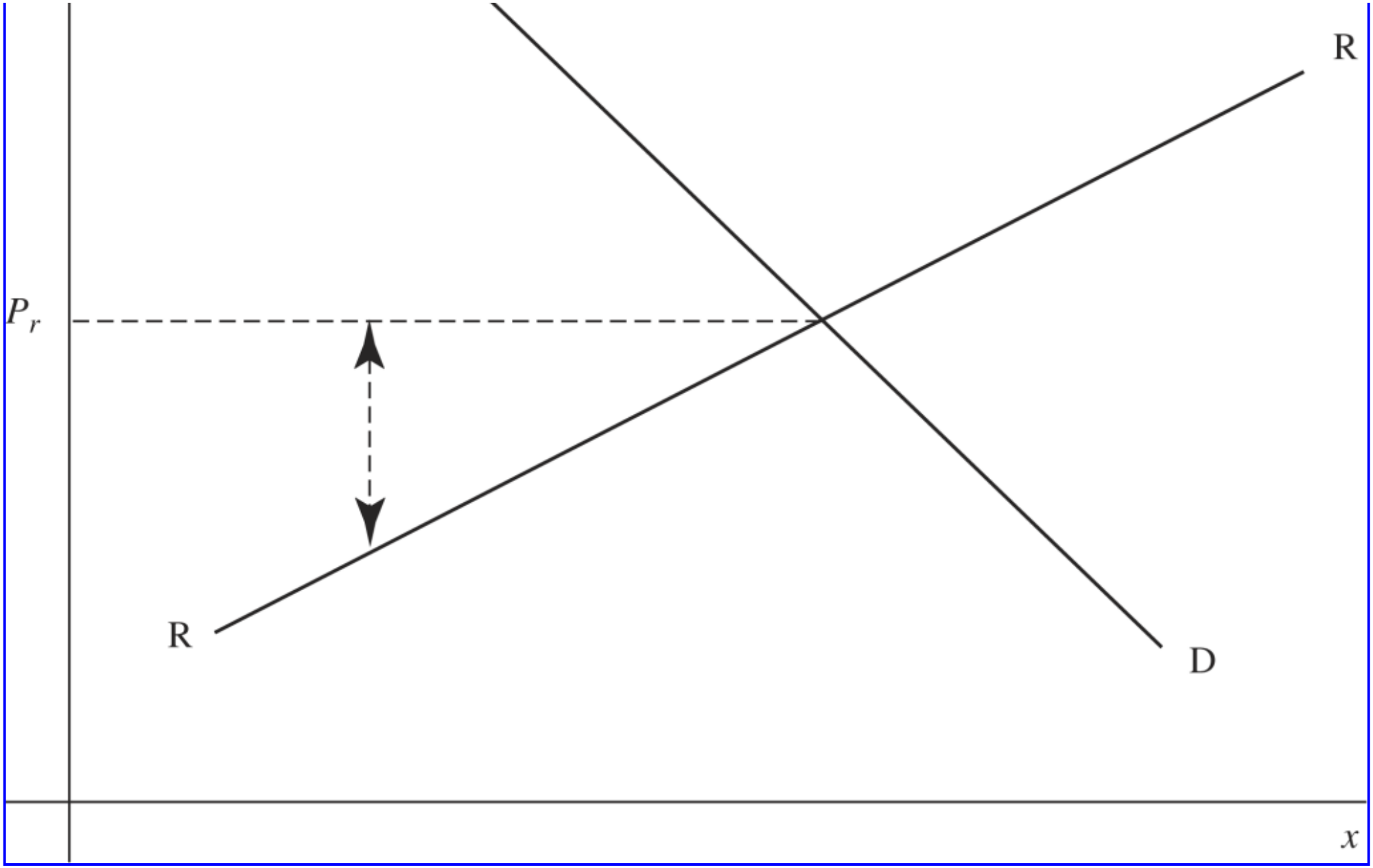

It is worth digressing to note that an upward rising true supply curve, which reflects increasing marginal costs of production, is different from the RR curve. In the true supply curve the area between the supply curve and the price line does not represent any of the above mentioned rents nor ‘producers' surplus’ (as it does with the RR curve). It is the portion of earnings of the supplier that exceed the variable costs and are applicable to cover the costs (possibly past investment costs) that are invariant to the rate of output. That area does not represent any excess of rental or sale value of units produced over their full costs, since only the variable costs are under the marginal cost curve. It represents the classic distribution of income to capital, if, for example, labour is presumed to be a variable input and capital a fixed input.

High rents a result, not cause, of high prices

An earlier unfortunate analytic confusion occurred in the common misimpression that high rents of land made its products more expensive. Thus the high rent of land in New York was and is still often believed to make the cost of living, or the cost of doing business, higher in New York. Or higher rent for some agricultural land is believed to increase the cost of growing corn on that land. Proper attention to the meaning of ‘demand’ and ‘costs’ would have helped avoid that confusion. Demand *here* for some unit for resource is the highest value use of that resource if used *here*. The cost of using it *here* is the highest valued forsaken alternative act elsewhere. For any resource the cost of its use *here* is its best value elsewhere, that is, its demand elsewhere. Land rent is high for ‘this’ use because the land's value in some other use is high. The reason the rent is high *here* and can be paid is that its use value *here* is bid by competitors for its use *here* into the offered rent and exceeds the value in some other use. The product of the land can get higher price *here*; that is why the rent is bid up so high, even though the particular winning bidder then believes a high price of the products must be obtained because the rent was high, rather than the reverse. As with every marketable resource, its highest value use *here* determines its rent, rather than the reverse. It was the implication of this kind of analysis that Marshall attempted to summarize in the famous aphorism, which he attributed to Ricardo (1817): ‘Rent does not enter into [Money] Cost of production’ (Marshall, 1890, p. 482).

Probably the source of the confusion in believing that high rents of land caused high prices for products produced on expensive land is that an individual user of that expensive resource has to be able to charge a higher price for the product, if the rent is to be covered. Bidders for that land compete for the right to the land that can yield a service worth so much – though to any individual successful bidder that rent has to be paid regardless of how well the successful bidder may be at actually achieving the highest valued use of the land. Hence it may appear to an individual bidder that the rent determines the price that must be charged, rather than, as is the correct interpretation, the achievable high valued use enables the high bid for the land for the person best able to detect and achieve that highest valued use.

Function of rent

Some people were aware of this bidding for the ‘land’ and concluded that the rent served no social purpose, since the land would exist anyway. But the high receipt resulting from competitive bidding for its uses serves a useful purpose. It reveals which uses are the highest valued and directs the land to that use. In principle, a 100 per cent tax on the land rent would not alter its supply (assuming initially that ‘land’ is the name of whatever has a fixed indestructible supply). This would be correct if in this case the ‘owner’ of the land had any incentive left to heed the highest bidder where the highest bid determines the rent. The assertion assumes that somehow the highest valued use can be known and that amount of tax be levied without genuine bona fide competitive bids for its use, a dubious if not plainly false proposition.

Monopoly rent

Let the word ‘monopoly’ denote any seller whose wealth potential is increased by restrictions on other potential competitors, restrictions that are artificial or contrived in not being naturally inevitable. Laws prohibiting others from selling white wine, or opening restaurants, or engaging in legal practice are examples. It should be immediately emphasized that this does not imply nor is it to be inferred that all such restrictions are demonstrably undesirable. Nevertheless, the increased wealth potential is a ‘monopoly rent’. Whether it is realized by the monopolist as an increase in wealth depends upon the costs of competing for the imposition of such restrictions. Competition for ‘monopoly rents’ may transfer them to, for example, politicians who impose the restrictions, and in turn may be dissipated by competition among politicians seeking to be in a position to grant such favours. The ‘monopoly rents’ may be dissipated (by what is often called ‘rent-seeking’ competition for such monopoly status of rights to grant it) into competitive payments for resources that enable people to achieve status to grant such restrictions. Those who initially successfully and cheaply obtained such ‘monopoly’ status may obtain a wealth increase, just as successful innovators obtain a profit stream before it is eliminated by competition from would-be imitators.

See Also

- Malthus, Thomas Robert
- Marshall, Alfred
- Ricardo, David
- Thünen, Johann Heinrich von
- West, Edward

Bibliography

Marshall, A. 1890. *Principles of Economics*. 1st edn, London: Macmillan.

Marshall, A. 1920. *Principles of Economics*. 8th edn, London: Macmillan; reprinted, 1946.

Ricardo, D. 1821. *Principles of Political Economy and Taxation*. 3rd edn, London: Dent Dutton, 1965.

How to cite this article

Alchian, Armen A. "rent." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000098>
doi:10.1057/9780230226203.1420

The New Palgrave Dictionary of Economics Online

repeated games

Kandori Michihiro
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article shows why self-interested agents manage to cooperate in a long-term relationship. When agents interact only once, they often have an incentive to deviate from cooperation. In a repeated interaction, however, any mutually beneficial outcome can be sustained in an equilibrium. This fact, known as the folk theorem, is explained under various information structures. This article also compares repeated games with other means to achieve efficiency, and briefly discusses the scope for potential applications.

Keywords

antitrust enforcement; bargaining; cartels; collusion; contract theory; cooperation and its evolution; correlated equilibrium; equilibrium selection; finite horizons; folk th; imperfect monitoring; infinite horizons; informal contracts; international policy coordination; long-term relationships; mechanism design; multiple equilibria; perfect monitoring; private monitoring; public monitoring; repeated games; relational contracts; subgame perfect equilibrium; trigger strategy; uniqueness of equilibrium

Article

Repeated games provide a formal and quite general framework to examine why self-interested agents manage to cooperate in a long-term relationship.
Formally, repeated games refer to a class of models where the same set of agents repeatedly play the same game, called the ‘stage game’, over a long (typically, infinite) time horizon. In contrast to the situation where agents interact only once, *any* mutually beneficial outcome can be sustained as an equilibrium when agents interact repeatedly and frequently. A formal statement of this fact is known as the folk theorem.

Repeated games and the general theories of efficiency

Thanks to the developments since the mid-1970s, economics now recognizes three general ways to achieve efficiency: (*a*) competition; (*b*) contracts; and (*c*) long-term relationships. For standardized goods and services, with a large number of potential buyers and sellers, promoting market competition is an effective way to achieve efficiency. This is formulated as the classic First and Second Welfare Theorems in general equilibrium theory. There are, however, other important resource allocation problems which do not involve standardized goods and services. Resource allocation within a firm or an organization is a prime example, as pointed out by Ronald Coase (1937), and examples abound in social and political interactions. In such cases, aligning individual incentives with social goals is essential for efficiency, and this can be achieved by means of *incentive schemes* (penalties or rewards). The incentive schemes, in turn, can be provided in two distinct ways: by a formal contract or by a long-term relationship. The penalties and rewards specified by a formal contract are enforced by the court, while in a long-term relationship the value of future interaction serves as the reward and penalty to discipline the agents' current behaviour. The theory of contracts and mechanism design concern the former case, and the theory of repeated games deals with the latter. These theories provide general methods to achieve efficiency, and have become important building blocks of modern economic theory.

An example: collusion of gas stations and the trigger strategy

Consider two gas stations located right next to each other. They have identical and constant marginal cost *c* (the wholesale price of gasoline) and compete by publicly posting their prices. Suppose their joint profit is maximized when they both charge $p = 10$, whereby each receives a large profit π. Although this is the best outcome for them, they have an incentive to deviate. By slightly undercutting its price, each can steal all the customers from its opponent, and its profit (almost) doubles. The only price free from such profitable deviation is $p = c$, where their profit is equal to zero. In other words, the only Nash equilibrium in the price competition game is an *inefficient* (for the gas stations) outcome where both charge $p = c$. This situation is the rule rather than the exception: the Nash equilibrium in the stage game, the only outcome that agents can credibly achieve in a one-shot interaction, is quite often inefficient for them. This is because agents seek only their private benefits, ignoring the benefits or costs of their actions for their rivals.

In reality, however, gas stations enjoy positive profits, even when there is another station nearby. An important reason may well be that their interaction is not one-shot. Formally, the situation is captured by a *repeated game*, where the two gas stations play the price competition game (the stage game) over an infinite time horizon $t = 0, 1, 2, \ldots$•. Consider the following repeated game strategy:

1. 1. Start with the optimal price $p = 10$.
2. 2. Stick to $p = 10$ as long as no player (including oneself) has ever deviated from $p = 10$.
3. 3. Once anyone (including oneself) deviated, charge $p = c$ for ever.

This can be interpreted as an explicit or implicit agreement of the gas stations: charge the monopoly price $p = 10$, and any deviation triggers cut-throat price competition ($p = c$ with zero profit). Let us now check whether each player has any incentive to deviate from this strategy. Note that, if neither station deviates, each enjoys profit π every day. As we saw above, a player can (almost) double its stage payoff by slightly undercutting the agreed price $p = 10$. Hence the short-term gain from deviation is at most π. If one deviates, however, its future payoff is reduced from π to zero in each and every period in the future. Now assume that the players discount future profits by the *discount factor* $\delta \in (0, 1)$. The number δ measures the value of a dollar in the next period. The discounted future loss is $\delta\pi + \delta^2\pi + \cdots = \frac{\delta}{1-\delta}\pi$. If this is larger than the short-term gain from defection (π), no one wants to deviate from the collusive price $p = 10$. The condition is $\pi \leq \delta / (1 - \delta)\pi$, or equivalently, $1 / 2 \leq \delta$.

Next let us check whether the players have an incentive to carry out the threat (the cut-throat price competition $p = c$). Since $p = c$ is the Nash equilibrium of the stage game, charging $p = c$ in each period is a best reply if the opponent always does so. Hence, the players are choosing mutual best replies. In this sense, the threat of $p = c$ is credible or self-enforcing.

In summary, under the strategy defined above, players are choosing mutual best replies *after any history*, as long as $1 / 2 \leq \delta$. In other words, the strategy constitutes a *subgame perfect equilibrium* in the repeated game. Similarly, in a general game, any outcome which Pareto dominates the Nash equilibrium can be sustained by a strategy which reverts to the Nash equilibrium after a deviation. Such a strategy is called a *trigger strategy*.

Three remarks: multiple equilibria, credibility of threat and renegotiation, and finite versus infinite horizon

A couple of remarks are in order about the example. First, the trigger strategy profile is not the only equilibrium of the repeated game. The repetition of the stage game Nash equilibrium ($p = c$ for ever) is also a subgame perfect equilibrium. Are there any other equilibria? Can we characterize *all* equilibria in a repeated game? The latter question appears to be formidable at first sight, because there are an infinite number of repeated game strategies, and they can potentially be quite complex. We do have, however, some complete characterizations of all equilibria of a repeated game, such as folk theorems and self-generation conditions as will be discussed subsequently.

Second, one may question the credibility of the threat ($p = c$ for ever). In the above example, credibility was formalized as the subgame perfect equilibrium condition. According to this criterion, the threat $p = c$ is credible because a *unilateral* deviation by a *single* player is never profitable. The threat $p = c$, however, may be upset by *renegotiation*. When players are called upon to carry out this grim threat after a deviation, they may well get together and agree to ‘let bygones be bygones’. After all, when there is a better equilibrium in the repeated game (for example, the trigger strategy equilibrium), why do we expect the players to stick to the inefficient one ($p = c$)? This is the problem of *renegotiation proofness* in repeated games. The problem is trickier than it appears, however, and economists have not yet agreed on what is the right notion of renegotiation proofness for repeated games. The reader may get a sense of difficulty from the following observation. Suppose the players have successfully renegotiated away $p = c$ to play the trigger strategy equilibrium again. This is self-defeating, however, because the players now have an incentive to deviate, as they may well anticipate that the threat $p = c$ will be again subject to renegotiation and will not be carried out. For a comprehensive discussion of this topic (and also of a number of major technical results on repeated games), see an excellent survey by D. Pearce (1990).

Third, let me comment on the assumption of an *infinite* time horizon. Suppose that the gas stations are to be closed by the end of next year (due to a new zoning plan, for ex). This situation can be formulated as a *finitely* repeated game. On the last day of their business, the gas stations just play the stage game, and therefore they have no other choice but to play

the stage game equilibrium $p = c$. In the penultimate day, they rationally anticipate that they will play $p = c$ *irrespective of* their current action. Hence they are effectively playing the stage game in the penultimate day, and again they choose $p = c$. By induction, the *only* equilibrium of the finitely repeated price competition is to charge $p = c$ in *every* period. The impossibility of cooperation holds no matter how long the time horizon is, and it is in sharp contrast to the infinite horizon case.

Although one may argue that players do not really live infinitely long (so that the finite horizon case is more realistic), there are some good reasons to consider the infinite horizon models. First, even though the time horizon is finite, if players do not know in advance exactly *when* the game ends, the situation can be formulated as an infinitely repeated game. Suppose that, with probability $r > 0$, the game ends at the end of any given period. This implies that, *with probability 1*, the game ends in a finite horizon. Note, however, that the expected discounted profit is equal to $\pi(0) + (1-r)\delta\pi(1) + (1-r)^2\delta^2\pi(2) + \cdots$, where $\pi(t)$ is the stage payoff in period *t*. This is identical to the payoff in an infinitely repeated game with discount factor $\delta' = (1-r)\delta$. Second, the drastic 'discontinuity' between the finite and infinite horizon cases in the price competition example hinges on the uniqueness of equilibrium in the stage game. Benoit and Krishna (1985) show that, if each player has multiple equilibrium payoffs in the stage game, the long but finite horizon case enjoys the same scope for cooperation as the infinite horizon case (the folk theorem, discussed below, approximately holds for *T*-period repeated game, when $T \to \infty$).

The repeated game model

Now let me present a general formulation of a repeated game. Consider an infinitely repeated game, where players $i = 1, 2, \ldots, N$ repeatedly play the same stage game over an infinite time horizon $t = 0, 1, 2, \ldots$. In each period, player i takes some action $a_i \in A_i$, and her payoff in that period is given by a stage game payoff function $g_i(a)$, where $a = (a_1, \ldots, a_N)$ is the action profile in that period. The repeated game payoff is given by

$$\Pi_i = \sum_{t=0}^{\infty} g_i(a(t))\delta^t,$$

where $a(t)$ denotes the action profile in period *t* and $\delta \in (0, 1)$ is the discount factor. It is often quite useful to look at the *average payoff* of the repeated game, which is defined to be $(1-\delta)\Pi_i$. Note that, if one receives the same payoff *x* in each period, the repeated game payoff is $\Pi_i = x + \delta x + \delta^2 x + \cdots = x/(1-\delta)$. This example helps to understand the definition of average payoff: in this case $(1-\delta)\Pi_i$ is indeed equal to *x*, the payoff per period.

A *history* up to time *t* is the sequence of realized action profiles before t: $h^t = (a(0), a(1), \ldots, a(t-1))$. A *repeated game strategy* for player *i*, denoted by s_i, is a complete contingent action plan, which specifies a current action after any history: $a_i(t) = s_i(h^t)$ (a minor note: to determine $a_i(0)$, we introduce a dummy history h^0 such that $a_i(0) = s_i(h^0)$). A repeated game strategy profile $s = (s_1, \ldots, s_N)$ is a *subgame perfect equilibrium* if it specifies mutual best replies after any history.

The folk th

Despite the fact that a repeated game has an infinite number of strategies, which can be arbitrarily complicated, we do have a *complete* characterization of equilibrium payoffs. The folk theorem shows exactly which payoff points can be achieved in a repeated game.

Before stating the theorem, we need to introduce a couple of concepts. First, let us determine the set of physically achievable average payoffs in a repeated game. Note that, by alternating between two pure strategy outcomes, say *u* and *v*, one may achieve any point between *u* and *v* as the average payoff profile. Hence, an average payoff profile can be a weighted average (in other words, a convex combination) of pure strategy payoff profiles in the stage game. Let us denote the set of all such points by *V*. Formally, the set of *feasible average payoff profiles V* is the smallest convex set that contains the pure strategy payoff profiles of the stage game.

Second, let us determine the points in *V* that cannot possibly be an equilibrium outcome. For example, if a player has an option to stay out to enjoy zero profit in each period, it is a priori clear that her equilibrium average payoff cannot be less than zero. In general, there is a payoff level that a player can guarantee herself in any equilibrium, and this is formulated as the *minimax* payoff. Formally, the minimax payoff for player *i* is defined as $\underline{v}_i = \min_{\alpha_{-i}} \max_{\alpha_i} g_i(\alpha)$, where $\alpha = (\alpha_1, \ldots, \alpha_N)$ is a mixed action profile (α_i is a probability distribution over player *i*'s pure actions) and $g_i(\alpha)$ is the associated expected payoff. To understand why min and max are taken in that particular order, consider the situation where player *i*

always *correctly anticipates what others do*. If player *i* knows that others choose $\alpha_{-i} = (\alpha_1, \ldots \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_N)$, he can play a best reply against α_{-i} to obtain $\max_{\alpha_i} g_i(\alpha)$. Note well that $\max_{\alpha_i} g_i(\alpha)$ is a function of α_{-i}. In the worst case, where others take the most damaging actions α_{-i}, player *i* obtains the minimax payoff (this is exactly what the definition says). From this definition it is clear that, in any equilibrium of the repeated game, *the average payoff to each player is at least her minimax payoff*. In any equilibrium, each player correctly anticipates what others do, and simply by playing the stage game best reply in each period, any player can make sure that her average payoff is more than her minimax payoff. (A comment: we consider mixed strategies in the definition of the minimax payoff because in many games the minimax payoff is smaller when we consider mixed strategies.)

From what we saw, now it is clear that the set of equilibrium average payoff profiles of a repeated game is *at most* $V^* = \{v \in V \mid \forall i \; v_i > \underline{v}_i\}$. (The points with $v_i = \underline{v}_i$ are excluded to avoid minor technical complications.) The set V^* is called the *feasible and individually rational payoff set*. This is the set of physically achievable average payoff profiles in the repeated game where each player receives more than her minimax payoff. The folk theorem shows that any point in this 'maximum possible region' can indeed be an equilibrium outcome of the repeated game. (Throughout this article, I maintain a minor technical assumption that each player has a finite number of actions in the stage game.)

Folk th: In an *N*-player infinitely repeated game, any feasible and individually rational payoff profile $v \in V^*$ can be achieved as the average payoff profile of a subgame perfect equilibrium when the discount factor δ is close enough to 1, provided that either $N = 2$, or $N \geq 3$ and no two players have identical interests.

Formally, no two players have identical interests if there are *no* players *i* and *j* ($i \neq j$) whose payoffs satisfy $g_i(a) = bg_j(a) + c$, $b > 0$ (that is, no two players have the same preferences over the stage game outcomes). This is a 'generic' condition that is almost always satisfied: the case where players have identical interests is very special in the sense that the equality $g_i(a) = bg_j(a) + c$ fails by even a slight change of the payoff functions. Hence, the folk theorem provides a general theory of efficiency: it shows that, for virtually any game, any mutually beneficial outcome can be achieved in a long term relationship, if the discount factor is close to 1. *Although game-theoretic predictions quite often depend on the fine details of the model, this result is a notable exception for its generality*.

The crucial condition in the folk theorem is a high discount factor. The discount factor δ may measure the (subjective) patience of a player, or, it may be equal to $1/(1+r)$, where *r* is the interest rate per period. Although the discount factor may not be directly observable (in particular, in the former case), it should be high when one period is short. Hence, an empirically testable implication is that players who have daily interaction (such as the gas stations in our ex) have a better scope for cooperation than those who interact only once a year. An important message of the folk theorem is that a high *frequency of interaction* is essential for the success of a long term relationship.

The name 'folk th' comes from the fact that game theorists had anticipated that something like it should be true long before it was precisely formulated and proved. In this sense, the assertion had been folklore in the game theorist community. The proof is, however, by no means obvious, and there is a body of literature to prove the theorem in various degrees of generality. Early contributions include Aumann (1959), Friedman (1971) and Rubinstein (1979). The statement above is based on Fudenberg and Maskin (1986) and its generalization by Abreu, Dutta and Smith (1994). The proof is constructive: a clever strategy, which has a rather simple structure, is constructed to support any point in V^*.

Repeated games versus formal contracts

To discuss the scope of applications, I now compare a long-term relationship (repeated game) and a formal contract as a means to enforce efficient outcomes. As our gas station example shows, quite often an agent has an incentive to deviate from an efficient outcome, because it increases her private returns at the expense of the social benefit. Such a deviation can be deterred if we impose a sufficiently high penalty so that the *incentive constraint*

$$\text{gain from deviation} \leq \text{penalty}$$

is satisfied. This is the basic and common feature of repeated games and contracts. A formal contract explicitly specifies the penalty and it is enforced by the court. In repeated games, the penalty is indirectly imposed through future interaction. In this sense the theory of repeated games can be regarded as the theory of *informal or relational contracts*.

When is a long-term relationship a better way to achieve cooperation than a formal contract? First, a long-term relationship is useful when a formal contract is too costly or impractical. For example, it is often quite costly for a third party (the court) to verify whether there was any deviation from an agreement, while defections may be directly observed by the players themselves. In practice, what constitutes 'cooperation' is often so fuzzy or complicated that it is hard to write it down explicitly, although the players have a common and good understanding about what it is. 'Pulling enough weight' in a joint research project may be a good example. In those situations, a long-term relationship is a more practical way to achieve cooperation than a formal contract. In fact, a classic study by Macaulay (1963) indicates that the vast majority of business transactions are executed without writing

formal contracts. Second, there are some cases where a court powerful enough to enforce formal contracts simply does not exist. For example, in many problems in development economics and economic history, the legal system is highly imperfect. Even for developed countries in the modern age, there are no legal institutions which have enough binding power to enforce international agreements. Hence, repeated games provide a useful framework to address such problems as the organization of medieval trade, informal mutual insurance in developing countries, international policy coordination, and measures against global warming. Lastly, there is no legal system to enforce cartels or collusion, because the existing legal system refuses to enforce any contract that violates antitrust laws. Hence a long-term relationship is the only way to enforce a cartel or collusive agreement.

Is the folk theorem a negative result?

The theory of repeated games based on the folk theorem is often criticized because it does not, as the criticism goes, have any predictive power. The folk theorem basically says that anything can be an equilibrium in a repeated game. One could argue, however, that this criticism is misplaced if we regard the theory of repeated games as a theory of informal contracts. Just as anything can be enforced when the party agrees to sign a binding contract, in repeated games any (feasible and individually rational) outcome is sustained if the players agree on an equilibrium. Enforceability of a wide range of outcomes is the essential property of effective contracts, formal or informal. The folk theorem correctly captures this essential feature.

This criticism is valid, however, in the sense that the theory of repeated games does not provide a widely accepted criterion for equilibrium selection. When we regard a repeated game as an informal contract, where the players explicitly try to agree on which equilibrium to play, the problem of equilibrium selection boils down to the problem of bargaining. In such a context, it is natural to assume that an efficient point (in the set of equilibria) is played. In the vast majority of applied works of repeated games with symmetric stage games (such as the gas stations ex), it is common to look at the best symmetric equilibrium. In contrast, when players try to find an equilibrium through trial and error, the theory of repeated games is rather silent about which equilibrium is likely to be selected. A large body of computer simulation literature on the evolution of cooperation, pioneered by Axelrod (1984), may be regarded as an attempt to address this issue.

Imperfect monitoring

So far we assumed that players can perfectly observe each other's actions. In reality, however, long term relationships are often plagued by *imperfect monitoring*. For example, a country may not verify exactly how much CO_2 is emitted by neighbouring countries. Workers in a joint project may not directly observe each others' effort. Electronic appliance shops often offer secret discounts for their customers, and each shop may not know exactly how much is charged by its rivals. In such situations, however, there are usually some pieces of information, or *signals*, which imperfectly reveal what actions have been taken. Published meteorological data indicates the amount of CO_2 emission, the success of the project is more likely with higher effort, and a shop's sales level is related (although not perfectly) to its rivals' prices.

According to the nature of the signals, repeated games with imperfect monitoring are classified into two categories: the case of *public monitoring*, where players commonly observe a public signal, and the case of *private monitoring*, where each player observes a signal that is not observable to others. Hence, the CO_2 emission game and the joint-project game are examples with imperfect public monitoring (published meteorological data and the success of the project are publicly observed), while the secret price-cutting game by electronic shops is a good example with imperfect private monitoring (one's sales level is private information).

This difference may appear to be a minor one, but, somewhat surprisingly, it is not. The imperfect *public* monitoring case shares many features with the *perfect* monitoring case, and we now have a good understanding of how it works. In contrast, the imperfect private monitoring case is not fully understood, and we have only some partial characterizations of equilibria. In what follows, I sketch the main results in the imperfect public and private monitoring cases.

Imperfect public monitoring

At first sight, this case might look much more complicated than the perfect monitoring case, but those two cases are similar in the sense that they share a *recursive structure*. Consider the set W^* of all average payoff profiles associated with the subgame perfect equilibria of a perfect monitoring repeated game. Any point $w \in W^*$ is a weighted average of the current payoff g and the continuation payoff $w' : (1-\delta)g + \delta w'$. The continuation payoff typically changes when a player deviates from g, in such a way that the short-term gain from deviation is wiped out. Subgame perfection requires that all continuation payoffs are chosen from the equilibrium set W^*. In this sense, W^* is generated by itself, and this stationary or recursive structure turns out to be quite useful in characterizing the set of equilibria.

The set of equilibria in an imperfect public monitoring game also shares the same structure. Consider the equilibria where the public signal determines which continuation equilibrium to play. When a player deviates from the current equilibrium action, it affects both her current payoff and (through the public signal) her continuation payoff. The equilibrium action should be enforceable in the sense that any gain in the former should be wiped out in the latter, and this is easier when the continuation payoff admits large variations. Formally, given

the range of continuation payoffs W, we can determine the set $B(W)$ of enforceable average payoffs. The larger the set W is, the more actions can be enforced in the current period (and therefore the larger the set $B(W)$ is). As in the perfect monitoring case, the equilibrium payoff set $W = W^*$ generates itself: it satisfies the *self-generation condition* of Abreu, Pearce and Stacchetti (1990) $W \subseteq B(W)$. W^* is the largest (bounded) set satisfying this condition, and the condition is in fact satisfied with equality. Conversely, it is easy to show that any (bounded) set satisfying the self-generation condition is contained in the equilibrium payoff set W^*.

This provides a simple and powerful characterization of equilibria, which is an essential tool to prove the folk theorem in the imperfect public monitoring case. The folk theorem shows that, despite the imperfection of monitoring, we can achieve any feasible and individually rational payoff profile under a certain set of conditions.

Before presenting a formal statement, let me sketch the basic ideas behind the folk theorem. When monitoring is imperfect, players have to be punished when a ‘bad’ signal outcome ω is observed, and this may happen with a positive probability even if no one defects. For example, in the joint project game, the project may fail even though everyone works hard. A crucial difference between the perfect and imperfect monitoring cases is that, in the latter, punishment occurs *on the equilibrium path*. The resulting welfare loss, however, can be negligible under certain conditions.

Consider a two-player game, where the probability distribution of the signal $\omega \in \Omega = \{\omega^1, \ldots, \omega^K\}$, when no one defects, is given by $P^* = (p^*(\omega^1), \ldots, p^*(\omega^K))$ in Figure 1. Suppose that each player's defection changes the probability distribution to exactly the same point P'. Then, there is absolutely no way to tell which player deviates, so that the only way to deter a defection is to *punish all players simultaneously*, when a ‘bad’ outcome emerges. This means that surplus is thrown away, and we are bound to have substantial welfare loss. Now consider a case where different players' actions affect the signal asymmetrically: player 1's defection leads to point P', while the defection by player 2 leads to P''. In this asymmetric case, one can *transfer* future payoff from player 1 to 2 when player 1's defection is suspected. Under such a transfer, surplus is never thrown away, and this enables us to achieve efficiency.

Figure 1

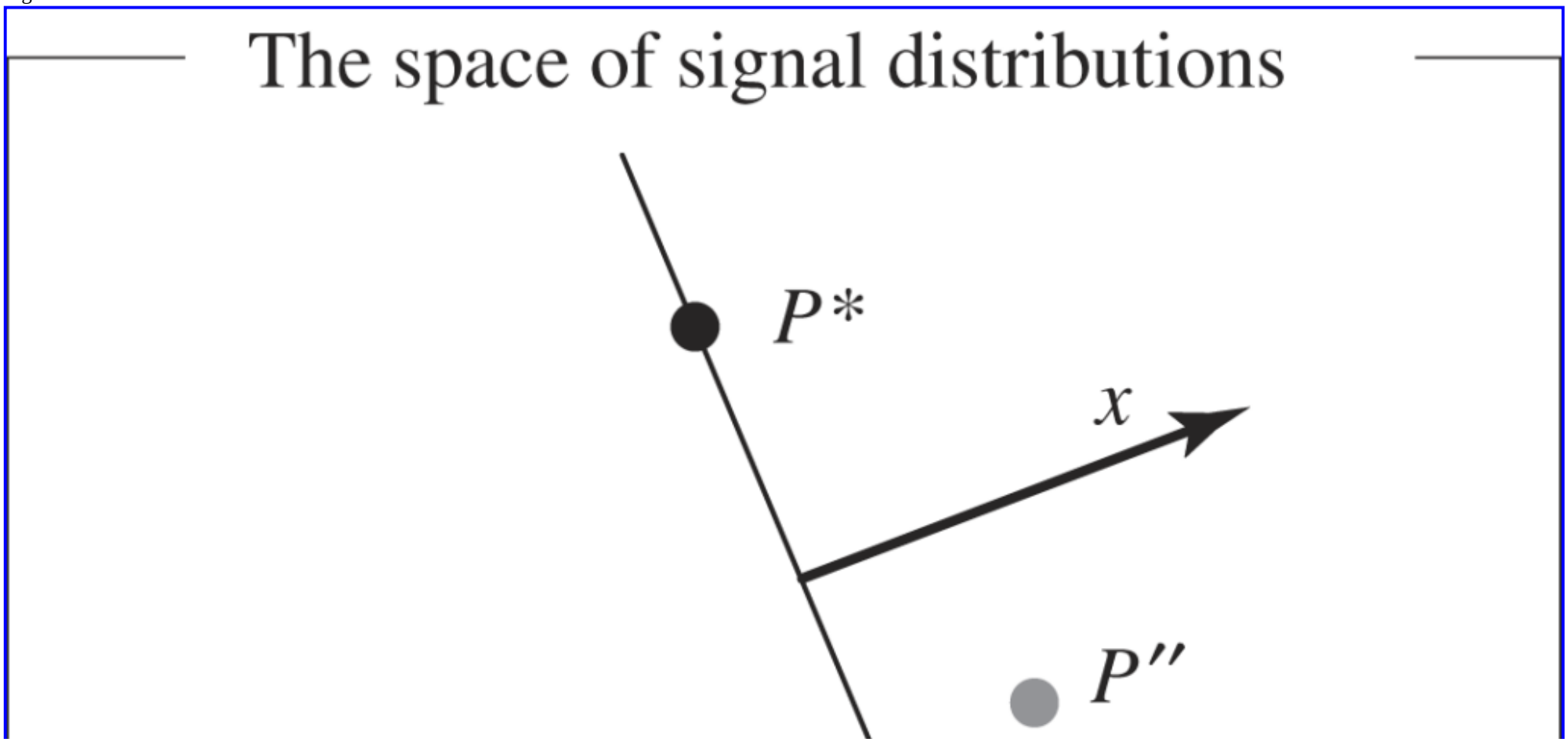

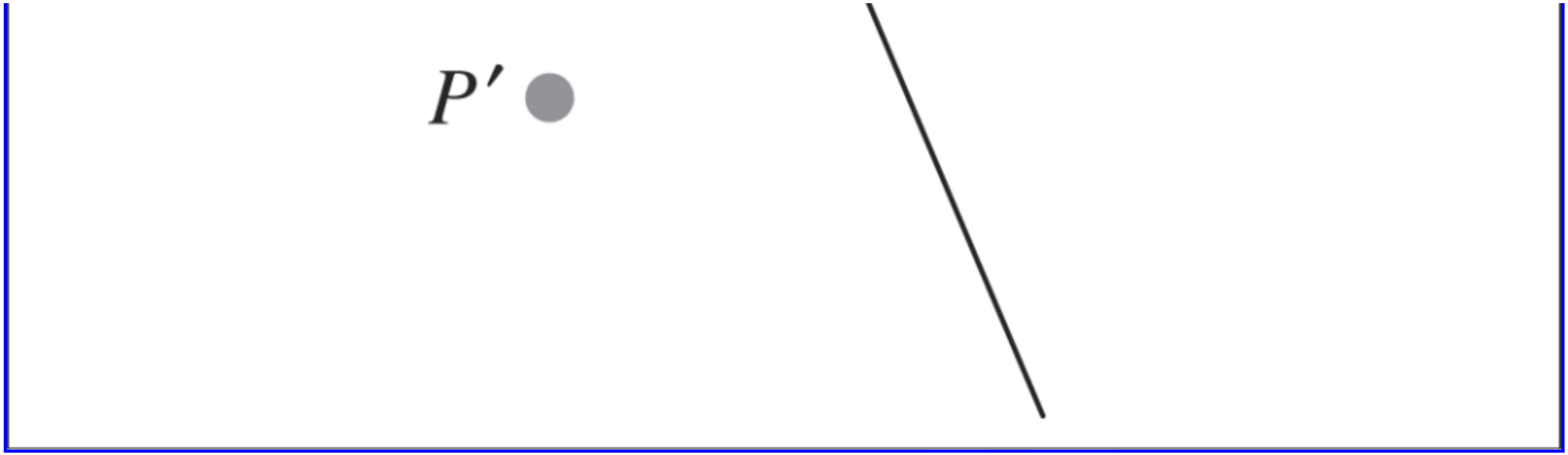

More precisely, consider the normal vector x of the hyperplane separating P' and P'' in the figure, and let $w_1 = x$ and $w_2 = -x$ be the continuation payoffs of player 1 and player 2 respectively. Figure 1 indicates that player 1's expected continuation payoff $P \cdot w_1 = P \cdot x$ is reduced by her own defection $(P' \cdot x < P^* \cdot x)$. Similarly, player 2's defection reduces her expected continuation payoffs $(P^* \cdot (-x) > P''(-x))$. Note that this asymmetric punishment scheme does not reduce the joint payoff, because by construction $w_1 + w_2$ is identically equal to 0. This is an essential idea behind the folk theorem under imperfect public monitoring: *When different players' deviations are statistically distinguished, asymmetric punishment deters defections without welfare loss*.

When can we say that different players' deviations are statistically *distinguished*? Note well that the above construction is impossible when P'' is exactly in between P^* and P' (that is, when P'' is a convex combination of P^* and P'). Such a case can be avoided if P^*, P' and P'' are linearly independent. The linear independence of the equilibrium signal distribution (P^*) and the distributions associated with the players' unilateral deviations (P' and P''), is a precise formulation of what it means when the signal 'statistically distinguishes different players' deviations'.

Let us now generalize this observation. Given an action profile (for simplicity of exposition, assume it is pure) to be sustained, there is an associated signal distribution P^*. Consider any pair of players i and j, and let $|A_k|$ be the number of player k's actions $(k = i, j)$ in the stage game. Since each player $k = i, j$ has $|A_k| - 1$ ways to deviate, we have $|A_i| + |A_j| - 2$ signal distributions associated with their unilateral deviations. If those distributions and the equilibrium distribution P^*, altogether $|A_i| + |A_j| - 1$ vectors, are linearly independent, we say that the signal can discriminate between deviations by i and deviations by j. This is called the *pairwise full rank condition*. This holds only when the dimension of the signal space ($|\Omega|$, the number of signal outcomes) is larger than the number of those vectors (that is, $|\Omega| \geq |A_i| + |A_j| - 1$). Conversely, if this inequality is satisfied, the pairwise full rank condition holds 'generically' (that is, it holds unless the signal distributions have a very special structure, such as exact symmetry). This leads us to the folk theorem under imperfect public monitoring (this is a restatement of Fudenberg, Levine and Maskin, 1994, in terms of genericity):

Folk theorem under imperfect public monitoring: Suppose that the signal space is large enough in the sense that $|\Omega| \geq |A_i| + |A_j| - 1$ holds for each pair of players i and j. Then, for a generic choice of the signal distributions and the stage game, any feasible and individually rational payoff profile $v \in V^*$ can be asymptotically achieved by a sequential equilibrium as the discount factor δ tends to 1.

In contrast to the perfect monitoring case, the proof is non-constructive. Rather than explicitly constructing equilibrium strategies, the theorem is proved by showing that any smooth subset of V^* is self-generating. In fact, the exact structure of the equilibrium strategy profile to sustain, for example, an efficient point is not so well understood. Sannikov (2005) shows that detailed structure of equilibrium strategies can be obtained if the model is formulated in continuous time.

Imperfect private monitoring

Now consider the case where all players receive a private signal about their opponents' actions. Although this has a number of important applications (a leading example is the secret

price cutting model), this part of research is still in its infancy. Hence, rather than just summarizing definitive results as in the previous subsections, I explain in somewhat more technical detail the source of difficulties and the nature of existing approaches.

The difficulties come from a subtle but crucial difference from the perfect or public monitoring case. I explain below the difference from three viewpoints, in the increasing order of technicality.

1. 1. In the perfect or public monitoring case, players share a mutual understanding about when and whom to punish. They can cooperate to implement a specific punishment, and, more importantly, they can mutually provide the incentives to carry out the punishment. This convenient feature is lost when players have diverse private information about each others' actions.
2. 2. In the perfect or public monitoring case, public information directly tells the opponents' future action plans. In the private monitoring case, however, each player has to draw statistical inferences about the history of the opponents' private signals to estimate what they are going to do. The inferences quickly become complicated over time, even if players adopt relatively simple strategies.
3. 3. In the perfect or public monitoring case, the set of equilibria has a recursive structure, in the sense that a Nash equilibrium of the repeated game is always played after any history. Now consider a Nash equilibrium of, for example, the repeated Prisoner's Dilemma with imperfect private monitoring. After the equilibrium actions in the first period, say (C,C), players condition their action plans on their private signals ω_1 and ω_2. Hence the continuation play is a *correlated equilibrium*, where it is common knowledge that the probability distribution of the correlation device (ω_1,ω_2) is given by $p(\omega_1,\omega_2|C,C)$. When player 1 deviates to D in the first period, however, the distribution of correlation device is *not* common knowledge: player 1 knows that it is $p(\omega_1,\omega_2|D,C)$, while player 2 keeps the equilibrium expectation $p(\omega_1,\omega_2|C,C)$. Hence, after a deviation, the continuation play is no longer a correlated equilibrium in the usual sense. In addition, the space of the correlation device (the history of private signals) becomes increasingly rich over time. Therefore, the equilibria in the private monitoring case do not have a compact recursive structure; a continuation play is chosen from a different set, depending on the history.

One way to get around these problems is to allow communication (Compte, 1998; Kandori and Matsushima, 1998). In their equilibrium, players truthfully communicate their private signal outcomes in each period. The equilibrium is constructed in such a way that each player's report of her signal is utilized to discipline *other* players and does *not affect one's own continuation payoff*. This implies that each player is indifferent about what to report, and therefore truth telling is *a* best reply. Such an equilibrium, which depends on the history of publicly observable messages, works in much the same way as the equilibria in the public monitoring case. Hence, with communication, the folk theorem is obtained in the private monitoring case.

The remaining issue is to characterize the equilibria in the private monitoring case without communication. From the viewpoint of potential applications, this is important, because collusion or cartel enforcement is a major applied area of repeated games, where communication is explicitly prohibited by the antitrust law.

One may expect that, when players' private information admits sufficient positive correlation, an equilibrium can be constructed in a similar way to the public monitoring case. Sekiguchi (1997) is the first to construct a non-trivial (and nearly efficient) equilibrium in the private monitoring game without communication, and his construction is basically built on such an idea. Strong correlation of private information is, however, not assumed in his model but is derived endogenously. He assumes that private signals provide nearly perfect observability and considered *mixed* strategies. In such a situation, the privately observed random variables, the action-signal pairs, are strongly correlated (because a player's random action is strongly correlated with another player's signal under nearly perfect observability). Mailath and Morris (2002) show that, in general, there is ‘continuity’ between the public and private but sufficiently correlated monitoring cases, in the sense that any strategy with a *finite memory* works in either case.

Those papers are examples of the *belief-based approach*, which directly addresses the statistical inference problem (see point 2. above). Some other papers follow this approach, and they provide judiciously constructed strategies in rather specific examples, where the inference problem becomes tractable. Aside from the case with near perfect correlation, however, we are yet to have generally applicable results or techniques from this approach.

More successful has been the *belief-free approach*, where an equilibrium is constructed in such a way that the inference problem becomes *irrelevant*. As a leading example, here I explain Ely and Valimaki's work (2002) on the repeated Prisoner's Dilemma with imperfect private monitoring. Each player's strategy is a Markov chain with two states, *R* (reward) and *P* (punishment). A specific action is played in each state (*C* in *R*, and *D* in *P*), and the transition probabilities between the states depend on the realization of the player's private signal. Choose those transition probabilities in such a way that the *opponent* is always indifferent between C and D *no matter which state the player is in*. This requirement can be expressed as a simple system of dynamic programming equations, which has a solution when the discount factor is close to 1 and the private signal is not too uninformative. By construction, any action choice is optimal against this strategy after any history, and in particular this strategy is a best reply to itself (so that it constitutes an equilibrium). Note that one's incentives do not depend on the opponent's state, and therefore one does not have to draw the statistical inferences about the history of the opponent's private signals.

There are certain difficulties, however, in obtaining the folk theorem with such a class of equilibria. First, players may be punished simultaneously in this construction, and our discussion about the public monitoring case shows that some welfare loss is inevitable (unless monitoring is nearly perfect). Second, even if we restrict our attention to the nearly perfect monitoring case, there is a certain set of restrictions imposed on the action profiles that can be sustained by such a belief-free equilibrium.

Those difficulties can be resolved when we consider *block strategies*. Block strategies treat the stage games in T consecutive periods as if they were a single stage game, or a block stage game, and applies the belief-free approach with respect to those block stage games. It is now known that, by using the block strategies, the folk theorem under private monitoring holds in the nearly perfect monitoring case (Hörner and Olszewski, 2006) and for some two-player games where monitoring is far from perfect (Matsushima, 2004). In the former, the block structure of the stage game helps to satisfy the restrictions imposed on the sustainable actions in belief-free equilibria. In the latter, an equilibrium is constructed where players choose constant actions in each block. This means that players have T samples of private signals for the constant actions, so that the observability practically becomes nearly perfect when T is large. With this increased observability and some restrictions on payoff functions, the folk theorem is obtained. For this construction to be feasible, the signals have to satisfy certain strong conditions, such as independence (across players).

The general folk theorem, or a general characterization of equilibria, for the private monitoring case is yet to be obtained, and it remains an important open question in economic theory. A comprehensive technical exposition of the perfect monitoring, imperfect public monitoring, and private monitoring cases can be found in Mailath and Samuelson (2006).

See Also

- cartels
- cooperation
- reputation
- social norms

Bibliography

Abreu, D., Dutta, P. and Smith, L. 1994. The folk theorem for repeated games: a NEU condition. *Econometrica* 62, 939–48.

Abreu, D., Pearce, D. and Stacchetti, E. 1990. Towards a theory of discounted repeated games with imperfect monitoring. *Econometrica* 58, 1041–64.

Aumann, R. 1959. Acceptable points in general cooperative N-person games. In *Contributions to the Theory of Games*, vol. 4, ed. R.D. Luce and A.W. Tucker. Princeton: Princeton University Press.

Axelrod, R. 1984. *Evolution of Cooperation*. New York: Basic Books.

Benoit, J.P. and Krishna, V. 1985. Finitely repeated games. *Econometrica* 53, 905–22.

Coase, R. 1937. The nature of the firm. *Economica* n.s. 4, 386–405.

Compte, O. 1998. Communication in repeated games with imperfect private monitoring. *Econometrica* 66, 597–626.

Ely, J. and Valimaki, J. 2002. A robust folk theorem for the Prisoner's Dilemma. *Journal of Economic Theory* 102, 84–105.

Friedman, J. 1971. A non-cooperative equilibrium for supergames. *Review of Economic Studies* 38, 1–12.

Fudenberg, D., Levine, D. and Maskin, E. 1994. The folk theorem with imperfect public information. *Econometrica* 62, 997–1040.

Fudenberg, D. and Maskin, E. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54, 533–54.

Hörner, J. and Olszewski, W. 2006. The folk theorem for games with private almost-perfect monitoring. *Econometrica* 74, 1499–544.

Kandori, M. and Matsushima, H. 1998. Private observation, communication and collusion. *Econometrica* 66, 627–52.

Macaulay, S. 1963. Non-contractual relations in business: a preliminary study. *American Sociological Review* 28, 55–67.

Mailath, G. and Morris, S. 2002. Repeated games with imperfect private monitoring: notes on a coordination perspective. *Journal of Economic Theory* 102, 189–228.

Mailath, G. and Samuelson, L. 2006. *Repeated Games and Reputations: Long-Run Relationships*. Oxford: Oxford University Press.

Matsushima, H. 2004. Repeated games with private monitoring: two players, *Econometrica* 72, 823–52.

Pearce, D. 1990. Repeated games: cooperation and rationality. In *Advances in Economic Theory*, ed. J. Laffont. Cambridge: Cambridge University Press.

Rubinstein, A. 1979. Equilibrium in supergames with overtaking criterion. *Journal of Economic Theory* 21, 1–9.

Sannikov, Y. 2005. Games with imperfectly observable actions in continuous time. Mimeo, University of California, Berkeley.

Sekiguchi, T. 1997. Efficiency in repeated Prisoner's Dilemma with private monitoring. *Journal of Economic Theory* 76, 345–61.

How to cite this article

Michihiro, Kandori. "repeated games." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000104>
doi:10.1057/9780230226203.1423

The New Palgrave Dictionary of Economics Online

reputation

Martin W. Cripps
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We explain what reputation effects are, how they arise and the factors that limit or strengthen them.

Keywords

collusion; complete information games; conflicting interest games; extensive form games; imperfect monitoring; incomplete information games; industrial organization; Jensen's inequality; Markov equilibria; martingales; prisoner's dilemma; repeated games; reputation; signalling; tit-for-tat; uncertainty

Article

In a dynamic setting signals sent now may affect the current and future behaviour of other players; thus, signals can have effects unrelated to their current costs and benefits. It is the interplay between signals and their long-run consequences that is studied in the literature on reputation.
The literature on reputation has two main themes. The first is that introducing a small amount of incomplete information in a dynamic game can dramatically change the set of equilibrium payoffs: introducing something to signal can have big implications in a dynamic model. These kinds of result can also be interpreted as providing a robustness check. Dynamic and repeated games typically have many equilibria, and reputation results allow us to determine which equilibria continue to be played when a game is ‘close’ to complete information. The second theme of the literature on reputations is that introducing incomplete information in a dynamic game may introduce new and important signalling dynamics in the players' strategies. Thus reputation effects tell us something about behaviour. This theme is particularly important in applications to macroeconomics and to industrial organization, for example. For either of these themes to be relevant it is necessary to have a dynamic game with incomplete information, so work on reputation has been influenced by, and influences, the larger literature on repeated and dynamic games of incomplete information. An excellent detailed treatment of reputation can be found in Mailath and Samuelson (2006).

An example

Most of the results below will be described in the context of a simple infinitely repeated trading game. The row player is a seller who can produce high or low quality. The column player is a buyer. Producing high quality is always expensive for the seller, so she would rather produce low quality; the buyer, however, wants to buy only a high-quality product. The only non-standard element is that the buyer regrets not buying a high-quality product. The trading game (Figure 1) has a unique equilibrium (L, N).
Figure 1
A trading game

	Buy (B)	Not Buy (N)
High Quality (H)	(1, 1)	(–1, –1)
Low Quality (L)	(2, –1)	(0, 0)

Let us record some facts about this game. The set

$$V \equiv \{(x, y): x > 0,\ y > -1/3,\ y \leq x \text{ and } y \leq 3 - 2x\} \subset \mathbb{R}^2,$$

illustrated in Figure 2, is the set of feasible and strictly individually rational payoffs for the trading game. The axes are drawn through the minmax payoffs to make *V* clear. If the seller could commit to a pure strategy, she would prefer to choose *H* as the buyer's best response to this is *B*. However, she could do even better by committing to a mixed strategy; playing (3/4,1/4) for example would also ensure the buyer played *B* and give the seller a bigger payoff. Reputation arguments can provide ways for these commitment payoffs to be achieved by sellers who are not actually committed to anything.

Figure 2

Sets of equilibrium payoffs and reputation bounds

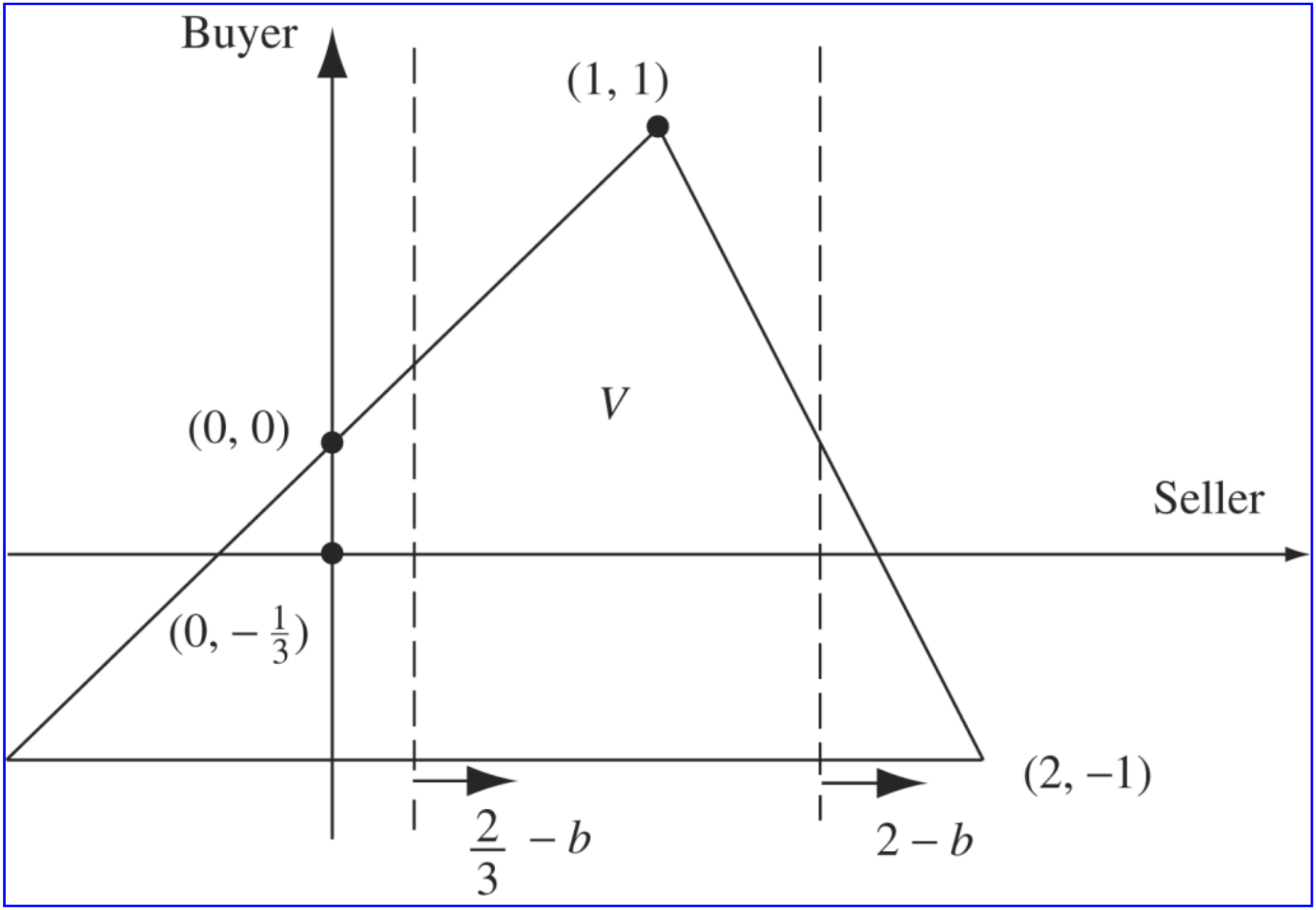

The trading game is played in each of the periods t=1,2,… with perfect monitoring; at the end of the period the players get to observe all payoffs and the pure action taken by their opponent. If both players' discount factors, δ <1, were sufficiently large, any point in V could be sustained as an equilibrium payoff. If the seller is long lived but faces an infinite sequence of buyers who each live one period, then any point on the line segment joining (0,0) to (1,1) is an equilibrium payoff. (No seller payoff above 1 is achievable if mixed actions are not observable; see Fudenberg, Kreps and Maskin, 1990.)

The stage is now set. To understand how reputation works we will need to introduce something for the seller to signal. Its commitment to high quality? Its low cost of high quality? Its commitment to always ripping off customers…? At this stage it is unnecessary to be specific, and we will concentrate on the general issues of learning. There are two types of sellers, ‘strong’ and ‘normal’, that the buyer may face in a game. The seller is told their type by nature at time t=0. The buyer, however, is unaware of nature's selection and spends the rest of the game looking at the seller's behaviour and trying to figure out what type she is. The normal seller plays action $a \in \{H, L\}$ with probability $\tilde{\sigma}^t(a)$ at time t, and the strong seller plays a with probability $\hat{\sigma}^t(a)$ at time t. Everything we say in the section below applies to the case where normal and strong sellers follow history-dependent strategies. (These behaviour strategies do depend on the – public – history of play before time t, but let us keep this out of our notation.) An initially uninformed buyer attaches probability p^t to the strong type and $1 - p^t$ to the normal type at time t; again this depends on the observed history. Our buyer expects the seller to play a with probability $\bar{\sigma}^t(a) = p^t\hat{\sigma}^t(a) + (1 - p^t)\tilde{\sigma}^t(a)$, and as time passes the buyers observe the outcomes of this strategy and revise their prior accordingly.

Tricks with Bayes's rule and martingales

Now we generate three properties of learning that are extensively used in the reputations literature. We will call them the ‘merging’ property, the ‘right ballpark’ property and the ‘finite surprises’ property. These properties are based on some simple facts about how Bayesian agents revise their beliefs, that is, how uncertainty about the seller's type is processed by the buyers or any other observer of its behaviour. A more advanced treatment of these results can be found in Sorin (1999). We defer any derivation of reputation results to the next section, so a reader could skip this section.

How does the buyer revise his or her beliefs in the light of an observed action a^t? A plain application of Bayes' rule tells us

$$p^{t+1} = \frac{Pr(a^t \cap \text{Strong})}{Pr(a^t)} = \frac{p^t\hat{\sigma}^t(a^t)}{\bar{\sigma}^t(a^t)}.$$

Or, in terms of the change in the beliefs

$$p^{t+1} - p^t = \frac{p^t[\hat{\sigma}^t(a^t) - \bar{\sigma}^t(a^t)]}{\bar{\sigma}^t(a^t)} = \frac{p^t(1 - p^t)[\hat{\sigma}^t(a^t) - \tilde{\sigma}^t(a^t)]}{\bar{\sigma}^t(a^t)}.$$

These equalities are powerful tools when combined with the properties of the priors.

Merging property. This tells us exactly how the long-run behaviour of the sellers is related to the buyer's long run beliefs. Either $p^t(1 - p^t) \to 0$ and the buyer eventually learns the type of the seller and can perfectly predict their actions, or all types of the seller end up behaving in the same way $\hat{\sigma}^t(a^t) - \tilde{\sigma}^t(a^t) \to 0$ and again the buyer can perfectly predict their actions. Nothing else can happen!

The stochastic process $\{p^t\}$ is a martingale on [0,1] with respect to public histories. To see this there is a simple calculation we can do.

$$E(p^{t+1}|h_t) = \sum_{a^t} Pr(a^t)\, p^{t+1} = \sum_{a^t} \bar{\sigma}^t(a^t) \frac{p^t\hat{\sigma}^t(a^t)}{\bar{\sigma}^t(a^t)} = p^t.$$

(The expectation $E(\cdot)$ is taken with respect to the buyer's beliefs about future play.) Bounded martingales converge almost surely (see Williams, 1991, for example), which implies $|p^{t+1} - p^t| \to 0$ almost surely. Applying this to the second equality above (noting that $|\bar{\sigma}^t(a^t)| \leq 1$), we get

$$p^t(1 - p^t)|\hat{\sigma}^t(a^t) - \tilde{\sigma}^t(a^t)| \to 0, \text{ (Merging)}$$

almost surely. This kind of result is extensively used in Hart (1985) and the literature that stems from his work.

Right ballpark property. The strong seller knows that the future will evolve according to the strategy $\hat{\sigma}$ (we use $\hat{Pr}(\cdot)$ and $\hat{E}(\cdot)$ to denote her probability measure and its expectation). This seller might ask, as she plays out an equilibrium, how little probability the buyers can attach to the strong seller, or how low p^t could get when she plays $\hat{\sigma}$. Of course, when the seller is in fact the strong type it is very unlikely that p^t becomes low – beliefs must stay in the right ballpark. (For example, if $\hat{\sigma}$ was actually a pure strategy the strong seller cannot ever believe p^t will decrease. As she plays $\hat{\sigma}$ there will be periods in which the normal type of seller could have done something different, so observing the actions of $\hat{\sigma}$ will cause buyers to revise p^t upwards.)

From the perspective of the strong seller, the likelihood ratio is a martingale:

$$\hat{E}\left(\frac{1 - p^{t+1}}{p^{t+1}}\middle| h_t\right) = \frac{1 - p^t}{p^t}.$$

(The calculation is just like the earlier one for p^t, where we use $\hat{Pr}(a^t) = \hat{\sigma}(a^t)$.) Let τ be the first time, s say, that $p_s \leq v$ and let C^t be the event that $\tau \leq t$. That is, sometime in the first t periods $p_s < v$. Then the martingale property combined with the optional stopping theorem (for example, Williams, 1991) implies

$$\frac{1 - p^0}{p^0} = \hat{E}\left(\frac{1 - p^{t+1}}{p^{t+1}}\right) \geq \hat{Pr}(C^t)E\left(\frac{1 - p^\tau}{p^\tau}|C^t\right) \geq \hat{Pr}(C^t)\frac{1 - v}{v}.$$

The above gives an upper bound on $\hat{Pr}(C^t)$ that is independent of t. Thus it also bounds the probability that p_t is ever below v:

$$\hat{Pr}(\exists t \text{ s.t. } p_t < v) \leq \frac{v}{p^0}. \text{ (Right Ballpark)}$$

Hence, the strong seller knows that it is very unlikely that the buyer's posterior will ever be close to certain she is actually the normal seller.

Finite surprises property. The strong seller might also ask how many times (as she plays $\hat{\sigma}$) the uninformed buyers will make a big mistake in predicting her strategy, that is, how

many periods does $\|\hat{\sigma}^t - \bar{\sigma}^t\| > v$ occur when the seller actually plays $\hat{\sigma}$. Here we are helped by the fact that our seller has only two actions, so the variation distance between the mixed actions is just twice the difference in probability of the realized action $\|\hat{\sigma}^t - \bar{\sigma}^t\| = 2|\hat{\sigma}^t(a^t) - \bar{\sigma}^t(a^t)|$. Let M_N be the event that there are more than *N mistakes*, $\|\hat{\sigma}^t - \bar{\sigma}^t\| > v$, before time *T*. The finite surprises property is that independently of the equilibrium $\hat{P}r(M_N) \to 0$ as $T, N \to \infty$. Thus, it is very unlikely that there are many periods in which the buyers do not think the seller will play as the strong type if the seller is indeed this type.

Jensen's inequality applied to the likelihood ratio above implies that the prior is a submartingale, that is, $\hat{E}(p^{t+1}|h_t) \geq p^t$. There is a second property of martingales we can now use: they cannot move around very much: $\Sigma_{t=1}^{T} \hat{E}((p^{t+1} - p^t)^2) \leq 1$. (A proof of this fact follows from $\hat{E}(p^{t+1} - p^t)^2 \leq \hat{E}((p^{t+1})^2 - (p^t)^2)$.) A substitution from the first Bayes' rule equality above then tells us

$$1 \geq \sum_{t=1}^{T} \hat{E}((p^t[\hat{\sigma}^t(a^t) - \bar{\sigma}^t(a^t)])^2).$$

It is obvious that only a few of the (non-negative) terms in the sum above can be much above zero, otherwise the upper bound will be violated. The right ballpark property tells us it is very unlikely that $p^t < v$. On the event $\left\{p^t \geq v \forall t\right\} \cap M_N$, the p^t in the above expectation is greater than ν and there are at least *N* differences that are bigger than $\nu/2$, so the sum is at least $N\nu(\nu/2)^2$, hence

$$1 \geq \sum_{t=1}^{T} \hat{E}(p^t[\hat{\sigma}^t(a^t) - \bar{\sigma}^t(a^t)])^2 \geq \hat{P}r(\left\{p^t \geq v \forall t\right\} \cap M_N)\frac{Nv^3}{4}.$$

Using the fact that $Pr(A \cap B) \geq Pr(B) - Pr(A^c)$ we now have an upper bound on $\hat{P}r(M_N)$.

$$\frac{4}{Nv^3} + \hat{P}r(\exists t \text{ s.t. } p_t < v) \geq \hat{P}r(M_N).$$

The right ballpark property gives us

$$\hat{P}r(M_N) \leq \frac{v}{p^0} + \frac{4}{Nv^3}. \text{ (Finite Surprises)}$$

As the size of the surprises becomes small $v \to 0$ and the number of surprises becomes large $Nv^3 \to \infty$, the strong seller must attach smaller and smaller probability to M_N. Fudenberg and Levine (1989; 1992), for example, invoke this property.

Basic reputation results: behaviour

The three tools above are sufficient to establish most well-known reputation results. The arguments below are entirely general, and are widely applied, but we use them only in the trading game. To make things simple, suppose that for some reason the strong seller is committed to playing (b, $1-b$), that is, in every period t the strong buyer provides high quality with probability b. We reserve the discussion of more complicated types of reputations for a later section.

From the perspective of the buyer, any equilibrium will consist of two phases: an initial phase when there is learning and signalling about the seller's type (this is sometimes called reputation building, although often reputation destruction is what occurs), and a terminal phase when the learning has virtually settled down. It is the merging property that tells us there must be this latter phase. The play in the game moves into this second phase either because the buyer is almost sure he knows the type of the seller (reputation considerations have vanished) or because the sellers are playing in the same way. Thus the equilibria of dynamic signalling games are inherently non-stationary, which is in contrast to much of the work on repeated games. Of course, Markovian equilibria can be calculated but these too will exhibit the two phases of play. The initial learning, when reputation builds or is destroyed, depends on the particular equilibrium and the game being studied. This phase may last only one period (if a once and for all revealing action is taken by a seller) but frequently it is long and has a random duration (if both types of seller randomize, for example).

Let us first examine reputation destruction in the case where $b \approx 1$, so the strong seller is committed to high quality and only very occasionally slips up. There is an equilibrium of this game where the normal type of seller will offer low quality more often than the strong type, and thereby gradually reveal her type (destroy her reputation for being good). Nevertheless, as this occurs she will enjoy heightened payoffs. The trade-offs our normal seller experiences in this game are what drive the reputation destruction. A seller offering low quality today enjoys the benefit of a higher payoff now, but the observation of low quality typically leads the buyers to revise downwards their probability of the strong seller and buy less in the future, whereas a seller offering high quality will lead the buyer's posterior on the strong seller to be revised upwards and an increased likelihood of buying in the future. Exactly how the normal seller chooses to trade off long-run benefits and short-run costs is unclear. It is possible that pooling dominates and that future buying is so strong that the normal seller prefers to offer high quality today even if it costs something in the short run. However, in this equilibrium the normal seller perceives the long-run benefits to be relatively small and prefers to offer low quality today. The normal seller can be thought of as exploiting, or cashing in, the value of her accumulated reputation. We also know, from the finite surprises property, that there will be finite opportunities for the normal seller to do this. Relatively soon there will come a time where the buyers know the seller is normal and purchase accordingly.

Reputation building (as opposed to destruction) is more likely in a world where there is the possibility that one is thought to be bad, for example, if the strong type is committed to ripping customers off and only occasionally produces a good product $(b \approx 0)$. In such a world the normal seller wants to tell buyers she is not this type, because by playing as the strong type she is doomed to never trade. She is building a reputation for *not* being the strong type. To do this the normal type will have to incur the cost of repeatedly offering high quality, even if the buyer is not buying. This is expensive and will drag down the normal seller's equilibrium payoff. But, as above, it will increase the likelihood of future buying by decreasing the likelihood of a strong seller. In contrast to the reputation destruction case, there are short-run costs borne by the normal type to achieve long-run gains. Again, the nature of these costs and benefits rely on the buyers' uncertainty about the seller's type.

Basic reputation results: payoffs

Reputation issues can have an extreme effect on payoffs, and this is what first came to the attention of economists. The general question of how the presence of something to signal in the repeated game affects the equilibrium payoffs could be answered in a number of ways. One way would be to calculate equilibria explicitly. This is usually difficult and would not establish results that hold for *all* equilibria.

Instead, a different approach is taken that is described in the following recipe:

1. 1. If the seller is strong, then in finite time the buyers will believe they face a seller who plays arbitrarily close to (b, $1-b$) for ever.
2. 2. Figure out what the buyers will do when the seller is strong.
3. 3. Use step 2 to evaluate the normal seller's payoff if she pretends to be strong for ever.
4. 4. At a Nash equilibrium the answer to step 3 is a lower bound on the normal seller's equilibrium payoff.

Step 1 is independent of the model and is a result of our earlier calculations. The right ballpark property tells us that p^t does not tend to zero when the seller is strong. The merging property then implies either $p^t \to 1$, or eventually all remaining normal types of buyer are also playing arbitrarily close to (b, $1-b$). In either of these cases, at a large but finite time the buyers believe that they face a seller who will always play (b, $1-b$).

Before proceeding to apply this recipe, we illustrate its power with the remarkable results we expect to get. Let us first consider a world where buyers are short run. We will show that

introducing an arbitrarily small probability that there is a strong seller places a lower bound on the normal seller's equilibrium payoffs of 2–*b* (when $b > 1/3$). Thus for *b* close to 1/3 the equilibrium payoffs in the complete information game (the segment joining (0,0) and (1,1)) and the incomplete information game are disjoint! Moreover, the normal seller can get almost his maximum feasible payoff at every equilibrium. In the second case, where buyers are also long run, we will get less strong conclusions; nevertheless, we will show that the normal type of seller must get at least 2/3–*b* when $b > 1/3$. These payoffs are illustrated in Figure 2.

The really difficult part of our recipe is step 2, because we have to understand how the buyers will behave in equilibrium. We therefore need to consider as separate cases what happens if buyers are short run or long run. Also, the amount of discounting that the sellers do affects the answer to step 3, so we need to consider different arguments for different amounts of discounting. The following catalog moves from simple to more elaborate arguments and from stronger to weaker reputation effects.

Reputation without discounting: short-term buyers

When a buyer lives only one period he plays a best response to the seller's current action. By step 1 in the very long run this will be *B* if $b > 1/3$ and *N* if $b > 1/3$. Step 3 is simple; by playing $\hat{\sigma}$ for ever the normal seller knows that in a large but finite time she can ensure the buyer will behave as above and so she will receive a stage game payoff approximately $R^*(b)$, where

$$R^*(b) := \begin{cases} 2-b & b > 1/3, \\ -b & b < 1/3. \end{cases}$$

If there is no discounting, and limits are correctly taken, $R^*(b)$ will equal the normal type's payoff from playing $\hat{\sigma}$ for ever. Thus, at step 4, at any Nash equilibrium the normal type must get at least $R^*(b)$.

In a general game R^* is equal to the seller's payoff from playing the strong type's stage game strategy when the buyer plays his or her unique best response. (If the best response is not unique this is not correct.)

Reputation with discounting: short-term buyers

Step 2 is as above – we still have short-term buyers. When the normal seller discounts payoffs, however, playing $\hat{\sigma}$ and eventually getting $R^*(b)$ every period does not tell us what her payoff discounted to time zero will be. There is an order of limits issue; as the discounting of the seller becomes weaker $(\delta \to 1)$, it could be that the equilibria change and there are more and more periods where the seller is not getting $R^*(b)$. It is now that the finite surprises property plays an important role. First notice that, when ν is chosen appropriately and $\|\bar{\sigma} - \hat{\sigma}\| < \nu$, then playing a best response to $\bar{\sigma}$ is the same as playing a best response to $\hat{\sigma}$. Hence, it is only when a surprise occurs that the normal seller is not getting $R^*(b)$ from playing $\hat{\sigma}$. But the probability of more than *N* surprises can be made very small *independently of the discounting*. So, as the discounting becomes weak and *N* periods have a small effect on total discounted payoff, there is a small probability of the normal seller of getting anything less than $R^*(b)$ when she plays $\hat{\sigma}$. Any Nash equilibrium, therefore, gives the normal seller at least $R^*(b)$. This is the kind of argument first made in specific cases by Kreps and Wilson (1982) and Milgrom and Roberts (1982), and generalized in Fudenberg and Levine (1989; 1992).

Reputation without discounting: one long-run buyer

If the buyer lives for many periods, he will not necessarily play a short-run best response to (*b*, 1–*b*) even if he expects it to be played for ever. We can, however, use some weaker information. At an equilibrium the buyer must on average get at least –1/3 (his minmax payoff) against (*b*,1–*b*). This implies that the buyer has to buy with at least probability 1/3 when $b > 1/3$ and buy with at most probability 1/3 when $b < 1/3$. There are, consequently, some bounds on the normal seller's payoff when she has played $\hat{\sigma}$ for a sufficiently long time. While playing (*b*,1–*b*) she gets 2–*b* when the buyer buys and – *b* if not; thus, if the buyer buys with probability greater than 1/3, she expects to receive a payoff of at least 2/3–*b*. If the buyer buys with at most probability 1/3, she expects to get at least – *b*. The seller is not discounting, so what she gets in the long run from playing $\hat{\sigma}$ is also what she expects to

get at time zero. Our answer to step 3, therefore, is

$$R^{\dagger}(b) := \begin{cases} (2/3) - b & b > 1/3, \\ -b & b < 1/3; \end{cases}$$

and we have a weaker lower bound on the normal type's payoff.

In an arbitrary game $R^{\dagger}$ is equal to the seller's worst payoff from playing as the strong type when the buyer plays a response that gives him more than his minmax payoff. In certain cases this can be a very strong restriction – for example, if the seller has a pure strategy that minmaxed the buyer and there is a unique response for the buyer that ensured he received his minmax payoff. Certain games, known as games of conflicting interests, have the property that the best action for the seller to commit to is pure and minmaxes the buyer. $R^{\dagger}$ is a very tight bound for such games.

Reputation with discounting: one long-run buyer

This final case combines most of the above issues. If the seller discounts the future much less than the buyer, then in the long run the seller must get $R^{\dagger}(b)$ from playing $\hat{\sigma}$. If a normal seller pretends to be strong, the buyers think there are at most N periods when the strong strategy is not played. Imagine now we have a buyer who cares only about what happens in the next t' periods. Such a buyer can think there are at most $t' N$ periods in which $\hat{\sigma}$ is not played for the next t' periods. (This kind of argument is due to Schmidt, 1993.) As the seller becomes very patient Nt' periods become of vanishing importance and the normal seller's payoff is bounded below by $R^{\dagger}(b)$. If the seller and buyer discount equally, however, reputation effects cannot be found except in some very special cases.

Imperfect monitoring: temporary and bad reputation

The analysis of reputation given above presupposes perfect monitoring by the buyers and sellers of each others' actions. In many dynamic and repeated games this is not likely. To what extent do the above results continue to hold when the players are not able to see exactly what their opponent did in any one period? Perhaps reputations are harder to establish if the observed behaviour is noisy? On the other hand, perhaps deviations from the strong type's action are harder to detect and so reputations last longer and are more valuable.

The merging, right ballpark and finite surprises properties all hold true under imperfect monitoring, with a suitable redefinition, provided there is enough statistical information for the buyer to eventually identify the seller's behaviour. (This is a full-rank condition on the players' signals.) As a result, the bounds on payoffs given in the previous section continue to hold.

Under imperfect monitoring with adequate statistical information there is one new behavioural feature of these games – reputation is almost always temporary, that is, the buyer will eventually get to know the seller's type. To see why this is so, let us amend the game in Figure 1 by restricting the buyer to imperfectly observe the seller's action. With probability $1 - \varepsilon$ the buyer observes the seller's true action in the current period, but with probability ε he observes the reverse action. (We must also assume the buyer does not see his own payoffs, otherwise he can deduce the seller's action from his payoff.) Consider a game where the seller always provides high quality (b=0) and suppose that reputation is permanent in such a game. Then p would, at least some of the time, converge to a number that is not zero or one. (Remember beliefs have to converge.) The merging property tells us that, in this case where the limit of beliefs is between zero and one, the buyer will be certain the normal seller is always providing high quality. Such buyers will ignore the occasional low-quality product as just unlucky outcomes, and there will be no loss of seller reputation if the buyer ever receives low quality. The normal type of seller can, therefore, deviate from always providing high quality, gain one unit of profit, and not face any costs in terms of loss of reputation. This cannot be an equilibrium. The initial claim that reputation is permanent has to be false as a result of this contradiction. The details of this argument can be found in Cripps, Mailath and Samuelson (2004).

When the monitoring is not statistically informative, ‘bad reputation’, due to Ely and Valimaki (2003), is a possibility. Uninformative monitoring is a particular problem in repeated extensive form games, because players do not get to see the actions their opponent would have taken on other branches of the game tree. Bad reputation may arise in our example if the buyer could take an action (such as not to buy) that stopped the seller being able to signal her type. Then, the normal seller might find herself permanently stuck in a situation where she cannot sell. This is not particularly surprising if the buyers were strongly convinced they faced a strong seller that almost always provided low quality. However, in certain circumstances this problem is much more severe: even if the buyers were almost certain the seller were normal, every equilibrium has trade ending in a bounded and finite time. Thus, it is possible that introducing something for the seller to signal has huge negative costs for her equilibrium payoffs. To illustrate this, suppose the seller were a restaurant with

imperfect control over quality, although it does have a strategy (for example, doubling the butter and salt content!) that makes it more likely the buyer will think the meal he received is good – but is actually damaging to the buyer. When play has reached the position where just one more bad meal will lead the buyer to permanently avoid the restaurant, then the restaurant will choose to use this unhealthy strategy. Knowing this, the buyer will choose to go elsewhere for his last but one meal too, and there is an unravelling of the putative equilibrium. Buyers eat at the restaurant only if they get very few bad meals, because they know they are in for clogged arteries and high blood pressure after that. Bad reputation arises because the seller cannot resist the temptation of taking actions that are actually unfavourable to the buyer in an effort to regain his good opinion. They actually have the reverse effect of ultimately driving the buyers away.

Reputation for what?

In our discussion we consider a strong type of seller who is committed to playing a particular fixed (random) action in each period. Is this form of uncertainty the only relevant one, or are there other potential types of strong seller that may do even better for our normal seller? There are two alternatives to consider: the strong seller is committed to playing a history-dependent strategy, or the strong player is equipped with a payoff function and her strategy is determined by an equilibrium.

If the seller faces a sequence of short-term buyers, then committing to a fixed stage game action is the best she could ever do, because each buyer's optimization focuses on what the seller does in the current period – the future is irrelevant. Even when the buyers are long lived, there are circumstances where committing to play a fixed action imparts a strategic advantage in repeated play, for example in most coordination or common interest games. However, there are other repeated games, such as the Prisoner's Dilemma, and dynamic games where committing to a fixed stage action is worthless. What the seller would like to do is to commit to a strategy, such as tit-for-tat, which would persuade a sufficiently patient buyer to cooperate with the strong type. Provided some rather strong conditions are satisfied, this is possible.

Our recipe for reputation results will break down when we consider strong sellers with payoffs rather than actions; nevertheless, reputation results are possible. For example, if the strong seller had payoffs of 2 for high quality and zero for low quality he would be strategically identical to a seller who always provided high quality.

Many players: social reputation and other considerations

Thus far we have resolutely stuck to a model of two players, but it is clear that reputation is a pervasive social and competitive phenomenon. Here we sketch some of the issues in many-player reputation. The literature on this area is in its infancy; very little can be said with much certainty now.

The easiest case to deal with is what happens as the number of uninformed players (the buyers in our example) increases. Here the benefit to the seller of building a reputation for high quality increases, as providing a good product today means the seller is more likely to trade with many buyers tomorrow. In a way, increasing the number of buyers is like making the seller more patient, and so we would expect the seller to be more inclined to build a reputation in this case.

A second case would be where there are very large numbers of informed buyers trying to acquire reputations for individual or group characteristics. Models of career concerns are similar to reputation models and have many workers trying to acquire reputations for individual characteristics. Also, there are models of group reputation, such as Tirole (1996), where a particular class of individuals behaves in a particular way to perpetuate the ‘group's’ reputation. In both these types of model the large numbers assumption allows one individual's reputation decision to be treated as virtually independent of others. Thus they can be analysed using quite simple tools.

A final case is where a few informed agents are in competition or collusion with each other. Collusion in team reputation obviously introduces a public goods issue. If one player contributes to the good name of the group, he or she does not get to enjoy the full benefits of the contribution. Typically, therefore, reputations for such teams are harder to establish. One might conjecture that competition appears to drive a player towards excessive investment in reputation, but there are many effects at work that we do not completely understand. For example, competitors may also act to undermine their rival's reputation and to interfere with its development. This is a fertile region for applied and theoretical investigations.

See Also

- repeated games
- signalling and screening

Bibliography

Cripps, M.W., Mailath, G.J. and Samuelson, L. 2004. Imperfect monitoring and impermanent reputations. *Econometrica* 72, 407–32.

Ely, J. and Valimaki, J. 2003. Bad reputation. *Quarterly Journal of Economics* 118, 785–814.

Fudenberg, D., Kreps, D. and Maskin, E. 1990. Repeated games with long-run and short-run players. *Review of Economic Studies* 57, 555–74.

Fudenberg, D. and Levine, D.K. 1989. Reputation and equilibrium selection in games with a patient player. *Econometrica* 57, 759–78.

Fudenberg, D. and Levine, D.K. 1992. Maintaining a reputation when strategies are imperfectly observed. *Review of Economic Studies* 59, 561–79.

Hart, S. 1985. Nonzero-sum two-person repeated games with incomplete information. *Mathematics of Operations Research* 10, 117–53.

Kreps, D. and Wilson, R. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27, 253–79.

Mailath, G.J. and Samuelson, L. 2006. *Repeated Games and Reputations: Long-Run Relationships*. Oxford: Oxford University Press.

Milgrom, P. and Roberts, J. 1982. Predation, reputation and entry deterrence. *Journal of Economic Theory* 27, 280–312.

Schmidt, K. 1993. Reputation and equilibrium characterization in repeated games of conflicting interests. *Econometrica* 61, 325–51.

Sorin, S. 1999. Merging, reputation, and repeated games with incomplete information. *Games and Economic Behavior* 29, 274–308.

Tirole, J. 1996. A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies* 63, 1–22.

Williams, D. 1991. *Probability with Martingales*. Cambridge: Cambridge University Press.

How to cite this article

Cripps, Martin W. "reputation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000110>
doi:10.1057/9780230226203.1424

The New Palgrave Dictionary of Economics Online

resale markets

Philippe Jehiel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Resale markets are necessary to correct misallocations of assets, but do they always ensure that goods end up in the hands of those who value them most? This article reviews theoretical arguments as to why this need not necessarily be so and when inefficiencies might be expected despite the presence of resale markets. Policy implications are also suggested.

Keywords

Akerlof, G.; allocative externalities; asymmetric information; auctions; bargaining; Chicago School; Coase Theorem; commitment; efficient allocation; incentive constraints; interdependent values; limited liability; market failure; market imperfections; mechanism design; participation constraints; private information; property rights allocation; resale markets; reservation utility; revelation principle; risk aversion; transferability

Article

Resale markets seem necessary to correct misallocations of assets, where misallocations may be the result of mistakes in initial purchasing decisions, or more generally of changes in the state of the economy. For the sake of illustration, a car owner may after a while find it desirable to buy a new car, and he may be willing to resell his old car on the second-hand market. A manager of a firm holding a Universal Mobile Telephone System (UMTS) licence may be willing to resell her licence to another firm if she realizes that the firm is unable to cover its cost (generated by the licence acquisition). A homeowner may need to resell his house if he has to move to another country or jurisdiction.
A question of primary interest is whether such resale markets are good for the economy. Or, to put it differently, whether, when and how should such resale markets be regulated? This article starts with the laissez-faire viewpoint on this issue; it then proceeds to show how asymmetric information and commitment issues mitigate that viewpoint.

The laissez-faire viewpoint

The classical neoliberal viewpoint as represented by the Chicago School would favour laissez-faire. Within the present context, this would imply that resale markets should not be regulated. The premise of this line of thought is that resale markets give the right flexibility so that assets can be allocated to the right agents at any point in time. This view has important consequences for the theory of mechanism and market design. Indeed, it implies that the initial allocation of property rights is irrelevant, as resale markets should be able to correct any misallocations (this is one version of the so-called Coase Theorem – Coase, 1960). Thus, according to this view, a government interested in maximizing economic efficiency should worry neither about the method of privatization nor about how to allocate licences for operating public services. It should simply allow for well-functioning resale markets.

Of course, very few economists truly believe that real resale markets can achieve such a fantastic job of always allocating assets to the right agents at the right time. On the academic side, Akerlof (1970) provides an early theoretical example of market failure in the context of the market for used cars (more on this below). Coase himself argues that transaction costs which are numerous are likely to invalidate the above angelic view about resale markets. On the ‘real world’ side, it seems implausible that the method of privatization or the allocation of licences for the use of public services is irrelevant for economic efficiency. In fact, recent years have seen a rapid growth of auction methods to allocate licences or privatize publicly owned firms, suggesting an interest on the part of practitioners in market design. It is worth pointing out that, in the case of licence auctions, most governments have chosen not to allow for resale markets, suggesting some distrust towards their functioning.

In the tradition of Coase, the words ‘transaction costs’ will be interpreted to mean any reason why inefficiencies may arise in transactions. Of course, some of the reasons need not be related to the intuitive notion of transaction costs, and one could alternatively use the more neutral terminology of ‘market imperfections’. The rest of this article will review how theoretical insights from the mechanism design literature and the bargaining literature help identify significant sources of transaction costs. The review will abstract from transferability issues, which is a legitimate idealization for transactions that are not too big for the financial capabilities of the parties. The theoretical insights will then be used to shed some light on whether and how to regulate resale markets.

The role of private information

It is relatively intuitive to see why private information may be a source of inefficiency in transactions. A seller who privately knows her valuation for the object for sale has an incentive to pretend that she values the object more than she really does, in the hope that this will lead the buyer to increase his purchasing price. Similarly, a buyer has an incentive to pretend that he values the good less than he really does, in the hope that he will obtain a lower selling price. But such distortions inevitably induce inefficiencies whenever the gains of trade are not large enough. This intuition has been formalized in the work of Myerson and Satterthwaite (1983), who show that, if the distributions of valuations are independently distributed between a seller and a buyer, and if it is not known who values the good more, inefficiencies must arise in any bargaining game in which no outside money is given to the bargaining parties. One of the strengths of Myerson and Satterthwaite's work is that it applies to any bargaining

game, including protocols in which a broker could help improve the bargaining outcome and protocols allowing for several stages of bargaining. The result is obtained by relying on the so-called revelation principle, which allows for the derivation of constraints that should be satisfied in any Nash–Bayes equilibrium of any game (whether static or dynamic): these constraints are the so-called incentive constraints – an agent with valuation ν should find his own strategy no worse than the strategy of the same agent with valuation ν' – and the participation constraints – an agent should get at least what he could get by staying outside the game. Myerson and Satterthwaite then proceed to show that these constraints together with the constraint that the bargaining parties receive no outside money cannot be simultaneously satisfied unless there are inefficiencies (see Milgrom, 2004, for an exposition of this and other impossibility results).

The above buyer–seller set-up assumes that agents know how valuable the good is to them. This is referred to as a ‘private values set-up’. Akerlof (1970) identifies another source of bargaining inefficiency in set-ups in which the value to the buyer is a function of the information held by the seller – this is sometimes called an informational externality and referred to as an ‘interdependent values set-up’. For example, a seller of a used car may know the quality of his or her car, and the quality of the car obviously affects the valuation of both the seller and the buyer. In an elegant example, in which the buyer is known to value the good α times as much as the seller with $2 > \alpha > 1$ and the quality (identified here with the valuation of the seller) is distributed uniformly on [0,1], Akerlof shows that there can be no trade. The no-trade result arises because a selling price of *p* would be acceptable to the seller only if the quality is below *p*, resulting in an average quality of *p*/2. But such an average quality does not justify buying the good at price *p* for the buyer, as $\alpha p/2 - p < 0$. One of the beauties in Akerlof's example is that it illustrates that, even in situations in which it is common knowledge that the buyer values the good more than the seller, there is no trade in equilibrium. Even though Akerlof restricts his analysis to special trading mechanisms, the inefficiency he identifies can be shown to arise in any equilibrium of any bargaining game, with the use of the same mechanism design techniques as those of Myerson and Satterthwaite. It also extends (even though not in the extreme form of no trade) to other classes of problems with interdependent values (see Samuelson, 1984).

In the above bargaining set-ups, a specific form of property rights was assumed. Within the same examples, other efficiency conclusions would arise with alternative property right structures, thereby illustrating how the initial allocation of property rights may affect efficiency in the presence of informational asymmetries. Obviously, in Akerlof's interdependent values example, if the person valuing the good more is initially the owner of the good there is no inefficiency, which thereby offers a simple illustration of this idea. (See Jehiel and Pauzner, 2006, for further elaboration.) In the private values situation considered by Myerson and Satterthwaite, if the two parties are *ex ante* symmetric and initially own 50 per cent shares of the object, a double auction (in which the party quoting the highest price would buy the 50 per cent shares of the other party at a selling price in between the two quoted prices) would result in an efficient allocation of property rights. Cramton, Gibbons and Klemperer (1987) generalize the latter insight by showing that mixed ownership is economically superior in partnership dissolution problems with private values.

The above bargaining inefficiencies implicitly assume that no outside money can be introduced on to the bargaining table. Otherwise, with large enough subsidies, efficiency could be obtained in the above bargaining set-ups, thereby suggesting that an appropriate public intervention may eliminate the

inefficiency due to asymmetric information. However, in interdependent values situations in which agents hold multidimensional signals that are independently distributed across agents, Jehiel and Moldovanu (2001) show that the sole incentive constraints make it generically impossible to achieve the first-best allocation no matter how much money is introduced on to the bargaining table. This result is especially relevant in transactions involving several items because then private information is naturally multidimensional. The result then implies that no public intervention can eliminate the bargaining inefficiencies. (A similar conclusion arises even with one-dimensional private information if the single crossing condition is violated; see Maskin, 1992.)

The above results assume that there is no correlation in the private information held by the various agents. Whenever there are correlations, incentive constraints are less severe because the report made by agent *i* can be used to deter misreports by agent *j*. The works of Crémer and McLean (1985; 1988) and Johnson, Pratt and Zeckhauser (1990) (see also Myerson, 1981) suggest that inefficiencies can be totally eliminated even under moderate correlations if agents are risk neutral and transfers can be arbitrarily large. However, limited liability and risk aversion (which seem plausible, especially if very large transfers are involved) ensure that the qualitative insights obtained for the case without correlation continue to hold with moderate correlation (see Robert, 1991). Hence, inefficiencies due to asymmetric information continue to hold even in the correlated case, as long as correlation is not too large. (See also Compte and Jehiel, 2006, who argue within Myerson and Satterthwaite's private values set-up that inefficiencies may arise even with large correlation whenever agents have the option to leave the bargaining table at any time, thereby obtaining their reservation utility.)

As already mentioned, the above inefficiencies hold even if multiple stages of bargaining are allowed, as long as the only inferences of the players come from the equilibrium play of the other parties and not from the release of new hard information (either in an exogenous manner or through endogenous information acquisition). If new information becomes available, the situation is different. Obviously, if the private information held by the various agents become public at some stage, then at this stage bargaining parties with full commitment abilities should be able to implement an efficient agreement. This is because, if inefficiencies were to arise at that stage, a party could propose a Pareto improvement with no further move, keeping the generated surplus for herself: this can be viewed as an application of the Coase Theorem. But, even if one adopts the view that eventually private information becomes publicly available, a critical issue is about how long this takes. If it takes very long, inefficiencies are still likely to be significant because the transitory phase is long. If it does not take long and full commitments are possible, efficiency can be expected.

The role of commitment

The above reported results assume full commitment abilities on the part of the bargaining parties. Another major source of inefficiencies is the limited commitment abilities of the agents. From the viewpoint of mechanism design, the relaxation of commitment abilities of the proposing party (sometimes called Principal) is generally thought of as a bad thing. But one should be cautious here about the criterion used to assess what ‘good’ or ‘bad’ means. Clearly, from the viewpoint of the Principal limited commitment ability is a bad thing because it puts additional constraints on the Principal's maximization exercise. However, from the viewpoint of society (as measured by social

welfare), the conclusion is far from clear. For example, Coase's conjecture suggests that a monopolist with no commitment ability may end up pricing his good efficiently if consumers are forward-looking (they anticipate the distribution of future prices correctly) and patient enough. In a similar vein, the commitment ability of an auctioneer may allow him to use inefficient reserve prices, which he might be unable to exploit under weaker commitment scenarios. (See McAfee and Vincent, 1997, for a formal approach, and Zheng, 2002, for an optimal auction model in which, even though the seller can commit not to lower his reservation price if there is no interested buyer, buyers can resell the object if they wish.) Clearly, more work is required to understand the pros and the cons of commitment from a mechanism design perspective with non-benevolent principals.

In a number of transactions, the transacting parties impose a cost or benefit on third parties: think of the sale of pollution rights or the sale of technologies through patents in imperfectly competitive markets. From the viewpoint of the transaction, this corresponds to an externality in the sense that the trade between a subset of agents affects the payoffs of other agents (see Jehiel, Moldovanu and Stacchetti, 1996). Abstracting from informational asymmetries, Jehiel and Moldovanu (1999) in a one-object environment and Gomes and Jehiel (2005) in a general multi-object environment study resale markets in such set-ups with allocative externalities. They establish that the lack of commitment ability may induce long-run inefficiencies in resale markets whenever there are allocative externalities and agents are patient and forward-looking. Furthermore, if we take as given the legal constraints governing how goods can be exchanged, the initial allocation of property rights is shown to have no effect on the long-run properties of the equilibrium pattern of sales in such markets, as long as parties are forward-looking and patient enough. Thus, in such a complete information world, the lack of commitment ability induces inefficiencies in the presence of allocative externalities and at the same time makes it irrelevant how the initial property rights are allocated.

Practical implications

What are the lessons to be drawn from these theoretical observations? What do these results imply for the desirability of resale markets?

A first category of problems concerns those situations in which private information is persistent. Then the above inefficiency results show that in most scenarios, no matter how exchanges are organized, no matter whether or not resales are permitted, and no matter how well resale markets work, inefficiencies are inevitable. In interdependent value situations with multidimensional signals, even subsidies may not be enough to eliminate the inefficiencies.

Full commitments including controls over resales would seem desirable from a mechanism design viewpoint, as long as the proposing parties seek to maximize total welfare. However, with non-benevolent agents there is no reason in general to expect the full commitment scenario to be preferable to weaker commitment scenarios whenever private information is persistent.

A second category of problems concerns those situations with vanishing private information that will be identified with complete information. Then resale markets permit an efficient allocation of goods whenever agents care solely about their own allocation (that is, when there are no externalities). However, when there are allocative externalities in the sense that the allocation of agent *i* directly influences the well-being of agent *j*, resale markets do not allow parties with limited commitment

abilities to reach an efficient state of the economy. Yet, even when there are allocative externalities, the efficiency of the economy is unaffected by the initial allocation of property rights, suggesting that in such situations the only role for government interventions is through the legal framework, not the allocation of property rights. For example, it may be desirable from this perspective to require by law that the transacting parties compensate those agents suffering from the transaction.

In complete information situations, it would seem that full commitments including controls over resales should improve efficiency. However, that view ignores the reality of a changing environment, which is one of the basic rationales for the existence of resale markets. Because the economy is changing, resale markets are necessary. The complete contracting scenario implicitly assumed by the full commitment idea is impractical in that it might involve agents that are not even present in the economy (think of a future homeowner who may not yet be born and whose future possession already exists). From a practical viewpoint, the main issue is about understanding the effect of the legal framework that governs resale markets on the overall efficiency of the economy. Some insights about how the legal framework might improve the economic performance of resale markets have been suggested above (see the idea of compensating those agents who suffer from the transaction). Admittedly, more work on both the theoretical and empirical sides is required to understand this as well as the additional effect of persistent private information on resale markets.

See Also

- bargaining
- Coase theorem
- efficient allocation
- incentive compatibility
- market failure
- mechanism design

Bibliography

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Compte, O. and Jehiel, P. 2006. Veto constraint in mechanism design: inefficiency with correlated types. Mimeo. Paris-Jourdan Sciences Economiques and University College London.

Cramton, P., Gibbons, R. and Klemperer, P. 1987. Dissolving a partnership efficiently. *Econometrica* 55, 615–32.

Crémer, J. and McLean, R. 1985. Optimal selling strategies under uncertainty for a discriminating

monopolist when demands are interdependent. *Econometrica* 53, 345–62.

Crémer, J. and McLean, R. 1988. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica* 56, 1247–57.

Gomes, A. and Jehiel, P. 2005. Dynamic processes of social and economic interactions: on the persistence of inefficiencies. *Journal of Political Economy* 113, 626–67.

Jehiel, P. and Moldovanu, B. 1999. Resale markets and the assignment of property rights. *Review of Economic Studies* 66, 971–91.

Jehiel, P. and Moldovanu, B. 2001. Efficient design with interdependent valuations. *Econometrica* 69, 1237–59.

Jehiel, P., Moldovanu, B. and Stacchetti, E. 1996. How (not) to sell nuclear weapons. *American Economic Review* 86, 814–29.

Jehiel, P. and Pauzner, A. 2006. Partnership dissolution with interdependent values. *RAND Journal of Economics* 37 1–22.

Johnson, S., Pratt, J. and Zeckhauser, R. 1990. Efficiency despite mutually payoff–relevant private information: the finite case. *Econometrica* 58, 873–900.

McAfee, P. and Vincent, D. 1997. Sequentially optimal auctions. *Games and Economic Behavior* 18, 246–76.

Maskin, E. 1992. Auctions and privatization. In *Privatization*, ed. H. Siebert. Kiel: Institut für Weltwirtschaften der Universität Kiel.

Milgrom, P. 2004. *Putting Auction Theory to Work*. Cambridge: Cambridge University Press.

Myerson, R. 1981. Optimal auction design. *Mathematics of Operations Research* 6, 58–73.

Myerson, R. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 28, 265–81.

Robert, J. 1991. Continuity in auction design. *Journal of Economic Theory* 55, 169–79.

Samuelson, W. 1984. Bargaining under asymmetric information. *Econometrica* 52, 995–1005.

Zheng, C. 2002. Optimal auction with resale. *Econometrica* 70, 2197–224.

How to cite this article

Jehiel, Philippe. "resale markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000244> doi:10.1057/9780230226203.1425

The New Palgrave Dictionary of Economics Online

research joint ventures

Bruno Cassiman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A research joint venture (RJV) is an agreement between two or more partners to perform research and development (R&D). RJVs provide a mechanism to bridge the divide between the optimal public R&D policy – free dissemination of knowledge – and private incentives to invest in R&D – appropriation of returns to investments. Three important issues related to appropriation of returns to R&D condition the private incentives to form a RJV: coordination of R&D investments between RJV partners, free-riding inside and outside the RJV, and information sharing between RJV partners.

Keywords

adverse selection; antitrust; cartels; collusion; competition policy; externalities; free riding; incentive compatibility; industrial organization; information sharing among firms; innovation; innovation policy; mergers; public goods; R&D contracting; R&D cooperation; repeated games; research and development (R&D); research joint ventures; risk sharing; spillovers; technology licensing; transaction costs; transfer of technology

Article

A *research joint venture* (RJV) is an agreement between two or more partners to perform research and development (R&D), where each partner has an active role in the generation of new knowledge and technology. As such, a RJV is distinct from the *ex ante* or *ex post* agreement to acquire knowledge or technology as in *R&D contracting* or the *licensing* of technology respectively. Many times RJV and *R&D cooperation* are used as synonyms in the literature.
Two features distinguish R&D from ordinary capital investments. First, R&D is a public good (Arrow, 1962). The use by one firm of the information produced by its R&D investments does not diminish the amount of information available to other firms. Second, and related to its public good nature, R&D investment is plagued by an *externality* problem. Firms investing in R&D typically cannot fully

appropriate the returns to their own R&D investments. This tends to reduce the incentive to invest in R&D when firms act non-cooperatively (Spence, 1984; d'Aspremont and Jacquemin, 1988).

Both of these characteristics of R&D investments have a profound impact on the optimal way of organizing R&D, as they affect the incentives to invest in R&D.

From a welfare perspective the optimal economy-wide organization would involve the free distribution of the knowledge produced by these R&D investments. However, such a policy would provide little incentive for private investment in R&D in the first place. RJVs provide a mechanism to bridge this divide between public policy and private incentive.

Incentives to form RJVs

Given the public-good nature of R&D, firms do have an incentive to jointly develop technology and share the costs and risk of these projects. Mariti and Smiley (1983) provide evidence for the importance of cost and risk-sharing for the success of R&D cooperation. Developing new technology from scratch implies incurring a high (fixed) cost. Transferring and sharing knowledge that is already developed has a low (marginal) cost. Therefore, firms with complementary products (Röller, Tombak and Siebert, 1997) or complementary knowledge (Sakakibara, 1997) have an incentive to form RJVs to share knowledge for the development of new products. Furthermore, from a *transaction costs* perspective R&D collaboration allows access to specialized and complementary know-how, while at the same time allowing for a transfer of technology at lower transaction costs than with arm's length arrangements. As a result the total cost of developing new knowledge through a RJV is reduced (Pisano, 1990; Oxley, 1997).

While knowledge transfer and cost sharing provide the most common and trivial incentive for the formation of RJVs, the industrial organization literature emphasizes competitive motives for engaging in R&D cooperation and RJVs. R&D is imperfectly appropriable and R&D results, therefore, leak out involuntarily to rival firms. These models concentrate on horizontal R&D cooperation among rival companies as a mechanism to internalize these *spillovers*. The R&D process is represented as a two-stage, non-tournament model where in a first stage firms make R&D investments that (strategically) affect second-stage output market decisions through either a cost-reducing or a demand-enhancing effect. Firms can cooperate – form an RJV – in the R&D stage, but may continue to compete in the product market (for example, Katz, 1986; d'Aspremont and Jacquemin, 1988; Kamien, Müller and Zang, 1992; Suzumura, 1992; Leahy and Neary, 1997). From this literature we discern three important issues conditioning the interrelation between the profitability of RJVs and spillovers: *coordination*, *free-riding* and *information sharing*.

Coordination

Cooperation in these models is typically industry-wide and takes the form of firms coordinating R&D choices in order to maximize joint profits. As a result investment in R&D in an RJV is increasing in the level of the spillover as the firms internalize the positive effect these spillovers have on their partners. In addition, when spillovers are high enough – that is, above a critical level – coordination in R&D will result in higher R&D investment than in non-coordinating firms. At the critical spillover level, the

profitability of cooperative and non-cooperative R&D strategies coincides. (When goods are substitutes, the level of product differentiation and the number of rivals are important parameters that determine the critical spillover level; de Bondt, Slaets and Cassiman, 1992.)

Coordination through joint profit maximization without incurring any explicit costs to R&D cooperation increases the firms' profitability in these models. But, more importantly, spillovers increase the profitability of cooperation in R&D. Furthermore, for spillovers above the critical level, firms have an increasing incentive to engage in R&D coordination (De Bondt and Veugelers, 1991). This means that, when spillovers are high enough, firms have an increasing incentive to engage in R&D coordination. Such cooperation would furthermore enhance welfare as R&D investment and market output increase.

Free riding

Most models focus on the welfare and profitability of R&D cooperation, ignoring the stability of such cooperation. The stability of RJVs can be threatened by free riding of non-participating companies on the output of the venture, or by free riding by partners who may conceal their technological expertise while trying to absorb as much as possible of the partner's knowledge (Shapiro and Willig, 1990). Kesteloot and Veugelers (1994) find that cooperative agreements that are profitable, and at the same time also stable, require involuntary – outgoing – spillover levels that are not too high. (Using a repeated game, cheating can be prevented by grim-trigger strategies specifying an eternal dissolution of an industry-wide venture. An alternative approach to solve the internal stability problem is through the organizational design of the venture. Perez-Castrillo and Sandonis, 1996, characterize incentive compatible and individually rational contracts that lead to disclosure of knowledge and, hence, the formation of profitable research joint ventures.) Hence, although higher spillover levels increase the profits from cooperation through coordination, they also increase the profits from cheating by a partner and from free riding by an outsider to the cooperative agreement. Therefore, cooperative ventures become more profitable the more able firms are to restrict outgoing spillovers by protecting their information while selectively sharing information with partners.

Information sharing

Some models take into account the fact that firms can indeed manage spillovers by voluntarily increasing the spillovers among cooperating partners. Such information sharing is found to further increase the profitability of cooperation in R&D. In addition, information sharing not only increases the profitability of R&D cooperation; it also makes such agreements more stable. Eaton and Eswaran (1997) show that, when technology trading cartels are not necessarily industry-wide, information sharing is an even stronger stabilizing force. In this case a much stronger punishment can be specified, namely, the ejection of the cheating firm from a technology-trading coalition, followed by the continuation of information sharing by the non-cheating members. Similarly, De Bondt and Wu (1997) find that information sharing produces larger coalition sizes that are both internally and externally stable. Katsoulacos and Ulph (1998) explicitly model the choice of spillovers by cooperating and non-cooperating firms, and find that RJVs will always share at least as much information as non-cooperating firms because the former maximize joint profits. When firms act non-cooperatively, however, one would

expect that the aim is to minimize the creation of spillovers – the *outgoing* spillovers – through the use of effective legal and strategic protection measures while at the same time to maximize the *incoming* spillovers. Kamien and Zang (2000) show that firms that coordinate their R&D expenditures maximize information flows – their *incoming* spillovers – through the choice of very broad research directions for the RJV. If the firms cannot coordinate their R&D expenditures, they are more concerned about managing their *outgoing* spillovers by choosing a more narrow research approach. This result emphasizes a potential dual role of spillovers: outgoing spillovers which might jeopardize the cooperative agreement, and incoming spillovers which increase the attractiveness of the cooperative agreement. In an empirical paper Cassiman and Veugelers (2002) indeed show that *incoming* spillovers and appropriability have important and separately identifiable effects: firms with higher incoming spillovers and better appropriation have a higher probability of cooperating in R&D.

RJVs and social welfare

When firms are allowed to form RJVs, R&D investments increase with the level of spillovers, exceeding the non-cooperative investment level when the spillovers are substantial (d'Aspremont and Jacquemin, 1988). Competing firms that cooperate in R&D might thus increase not only profits but also welfare when the spillovers are substantial. Policywise, a case can then be made for allowing RJVs to form when spillovers are high. However, when spillovers are low firms acting non-cooperatively with respect to R&D bring about higher welfare than when allowed to form an RJV (Suzumura, 1992). The only effect of a RJV in this case is to reduce R&D competition, which in turn decreases welfare (Katz, 1986). (It has often been suggested that RJVs might also facilitate collusion in the output market. A necessary condition for a RJV to be welfare improving in this case is that total R&D investments increase. Martin, 1997, analyses the increased potential for tacit collusion in RJVs, while Yi, 1995, looks at the welfare effects of product market collusion by an industry-wide RJV. Greenlee and Cassiman, 1999, discuss the effects of collusion in the output market on RJV formation.) This theoretical finding has fuelled the debate on the issue of relaxing antitrust regulation with respect to RJVs. In evaluating cooperative R&D, regulators often use the same ‘rule of reason’ as in the case of mergers. Given the dynamic nature of R&D, insensitive application of static merger guidelines may lead to undesirable outcomes (Ordover and Willig, 1985). Appropriate standards for evaluating RJVs should be developed. Jorde and Teece (1990) propose the creation of an administrative procedure for evaluating and possibly certifying cooperative R&D agreements in order to establish a safe harbour from antitrust litigation. But Shapiro and Willig (1990) argue that this would provide too much protection to RJVs, especially because the regulator needs a great deal of information to evaluate a RJV, and much of this information might be proprietary. Policymakers have attempted to address these issues. In the USA firms can register their RJVs under the National Cooperative Research Act (NCRA). By registering under the NCRA, firms become exempt from treble damages under antitrust regulation. However, cooperative R&D ventures need not register under the NCRA. In that case, they are liable under the usual antitrust regulation. Scott (1988) actually notes that cooperative research registered under the NCRA predominantly falls into industries without severe appropriability problems, while supposedly welfare-enhancing RJVs do not seem to register, leading to a suspicion of adverse selection of RJVs under the NCRA (Cassiman, 2000).

In Europe the 1986 Single European Act amendments to the Treaty of Rome gave the Community specific responsibility for strengthening ‘the scientific and technological basis of European industry’. In addition to the EEC block exemption of Article 85(1) of the EC treaty for cooperative ventures in R&D, a variety of programmes were initiated, many of which explicitly fostered inter-firm cooperation tied to Community funding for part of the R&D costs of the proposed projects (Martin, 1996). Nevertheless, the debate on the exact implementation of these policies is still ongoing and has initiated a broader debate on the interaction between innovation policy and competition policy.

Conclusion

While the industrial organization models of RJVs have focused on imperfect appropriation among competitors, several areas for research on RJVs remain thoroughly unexplored. First, empirical work has indicated that most RJVs are formed with customers, suppliers or research organizations rather than with competitors. Recent empirical work has started to tackle the issue of different types of partners for the RJVs, but little theoretical work has followed (Fritsch and Lukas, 2001; Belderbos, Carree and Lokshin, 2004; Veugelers and Cassiman, 2005).

Second, and related, we still know very little about the actual effect of engaging in RJVs on firm (innovation) performance. Brandstetter and Sakakibara (1998) find some evidence of the formation of RJVs on research productivity, and Belderbos, Carree and Lokshin (2004) show that cooperation in R&D leads firms to generate more sales from products that are new to the market. But most empirical studies interpret R&D cooperation as an indirect indication of RJV's profitability. To really uncover the incentives to engage in RJVs, we need to understand how RJVs improve the innovation performance of firms relative to alternative organizational forms.

Finally, little progress has been made yet in understanding the organization of RJVs from a theory of the firm perspective. Why would firms make joint investments in R&D and share property rights and decision rights over the outcome of future research outcomes? When is this efficient or when does it enhance the competitiveness of firms?

See Also

- externalities
- information sharing among firms

Bibliography

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, ed. R. Nelson. Princeton: Princeton University Press.

Belderbos, R., Carree, M. and Lokshin, B. 2004. Cooperative R&D and firm performance. *Research Policy* 33, 1477–92.

Brandstetter, L. and Sakakibara, M. 1998. Japanese research consortia: a microeconomic analysis of industrial policy. *Journal of Industrial Economics* 46, 207–33.

Cassiman, B. 2000. Research joint ventures and optimal R&D policy with asymmetric information. *International Journal of Industrial Organization* 18, 283–314.

Cassiman, B. and Veugelers, R. 2002. R&D cooperation and spillovers: some empirical evidence from Belgium. *American Economic Review* 92, 1169–84.

d'Aspremont, C. and Jacquemin, A. 1988. Cooperative and non-cooperative R&D in duopoly with spillovers. *American Economic Review* 78, 1133–7.

De Bondt, R., Slaets, P. and Cassiman, B. 1992. Spillovers and the number of rivals for maximum effective R&D. *International Journal of Industrial Organization* 10, 35–54.

De Bondt, R. and Veugelers, R. 1991. Strategic investment with spillovers. *European Journal of Political Economy* 7, 345–66.

De Bondt, R. and Wu, C. 1997. Research joint venture cartels and welfare. In *R&D Cooperation: Theory and Practice*, ed. J. Poyago-Theotoky. London: Macmillan.

Eaton, B. and Eswaran, M. 1997. Technology trading coalitions in supergames. *RAND Journal of Economics* 28, 135–49.

Fritsch, M. and Lukas, R. 2001. Who cooperates on R&D? *Research Policy* 30, 297–312.

Greenlee, P. and Cassiman, B. 1999. Product market objectives and the formation of research joint ventures. *Managerial and Decision Economics* 20, 115–30.

Jorde, T. and Teece, D. 1990. Innovation and cooperation: implications for competition and antitrust. *Journal of Economic Perspectives* 4(3), 75–96.

Kamien, Müller, E. and Zang, I. 1992. Research joint ventures and R&D cartels. *American Economic Review* 82, 1293–306.

Kamien, M. and Zang, I. 2000. Meet me halfway: research joint ventures and absorptive capacity. *International Journal of Industrial Organization* 18, 995–1012.

Katsoulacos, Y. and Ulph, D. 1998. Endogenous spillovers and the performance of research joint ventures. *Journal of Industrial Economics* 46, 333–58.

Katz, M. 1986. An analysis of co-operative research and development. *RAND Journal of Economics* 17, 527–43.

Kesteloot, K. and Veugelers, R. 1994. Stable R&D co-operation with spillovers. *Journal of Economics and Management Strategy* 4, 651–72.

Leahy, D. and Neary, P. 1997. Public policy towards R&D in oligopolistic industries. *American Economic Review* 87, 642–62.

Mariti, P. and Smiley, R. 1983. Co-operative agreements and the organisation of industry. *Journal of Industrial Economics* 38, 183–98.

Martin, S. 1997. Public policy toward cooperation in research and development the European Community Japan the United States. In *Competition Policy in the Global Economy*, ed. L. Waverman, W. Comanor and A. Goto. London: Routledge.

Martin, S. 1996. Protection, promotion and cooperation in the European semiconductor industry. *Review of Industrial Organization* 11, 721–35.

Ordover, J. and Willig, R. 1985. Antitrust for high-technology industries: assessing research joint ventures and mergers. *Journal of Law and Economics* 28, 311–33.

Oxley, J.E. 1997. Appropriability hazards and governance in strategic alliances: a transaction costs approach. *Journal of Law, Economics and Organization* 13, 387–409.

Perez-Castillo, D. and Sandonis, J. 1996. Disclosure of know-how in research joint ventures. *International Journal of Industrial Organization* 15, 51–75.

Pisano, G. 1990. The R&D boundaries of the firm: an empirical analysis. *Administrative Science Quarterly* 35, 153–76.

Röller, L., Tombak, M. and Siebert, R. 1997. Why firms form research joint ventures: theory and evidence. Discussion Paper No. 1654, CEPR.

Sakakibara, M. 1997. Heterogeneity of firm capabilities and cooperative research and development: an empirical examination of motives. *Strategic Management Journal* 18, 134–64.

Scott, J.T. 1988. Diversification versus cooperation in R&D investment. *Managerial and Decision Economics* 9, 173–86.

Shapiro, C. and Willig, R. 1990. On the antitrust treatment of production joint ventures. *Journal of*

Economic Perspectives 4(3), 113–30.

Spence, M. 1984. Cost reduction, competition and industry performance. *Econometrica* 52, 101–21.

Suzumura, K. 1992. Cooperative and noncooperative R&D in an oligopoly with strategic commitments. *American Economic Review* 82, 1307–20.

Veugelers, R. and Cassiman, B. 2005. R&D cooperation between firms and universities: some empirical evidence from Belgian manufacturing. *International Journal of Industrial Organization* 23, 355–79.

Yi, S. 1995. R&D cooperation, product-market collusion and welfare. Working paper, Department of Economics, Dartmouth College.

How to cite this article

Cassiman, Bruno. "research joint ventures." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000237> doi:10.1057/9780230226203.1426

The New Palgrave Dictionary of Economics Online

reservation price and reservation demand

Ian Steedman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

factor supply; free goods; labour supply; opportunity cost doctrine; real cost doctrine; reservation demand; reservation price; Wicksteed, P. H

Article

The simplest example of a reservation price is that price below which an owner will refuse to sell a particular object in an auction. Since the owner could always, in principle, enforce such a price by outbidding everyone else, this leads immediately to the more general concept of a reservation price as that price at which the owner of a fixed stock will choose to *retain* some given amount from that stock, rather than supply more, and of the amount retained as the owner's ‘reservation demand’ at the price in question. Considering alternative hypothetical prices, one sees that the owner's supply curve of the commodity can equally well be described as an ‘own (reservation) demand’ curve, where ‘supply’ and ‘own demand’ sum identically to the given stock. The same is naturally true of the market supply curve. Thus consider the standard example of the determination of the price of first-edition copies of a certain old book. A demand curve may be drawn up for those who at present own no copies. Taking account of each present owner's reservation price (or prices for those who possess more than one copy), we may also draw up a supply curve. (Of course ‘supply’ by present owners may be negative at low prices.) Confrontation of the demand and supply curves will then show the market-clearing price. Equally, however, we could have drawn up the ‘reservation demand’ curve of present owners, summed it with the demand curve of non-owners and then confronted the ‘total’ demand curve with the given stock. Since ‘supply’ and ‘reservation demand’ sum identically to total stock, at every price, the alternative diagram inevitably shows the same market-clearing price as does the first; it does not show the number of books traded, however.

It will be clear that an agent's reservation price for any type of commodity can be expected to depend on one or more of the following considerations: the scope for direct ‘own use’ of the commodity; the agent's present need for liquidity; the agent's other resources; the perishability of the commodity and thus the various elements of storage costs (including interest costs); expectations about future prices,

there being always a speculative element in the reservation price of any commodity which is not immediately perishable. These considerations all emerge in theories of ‘factor supply’, for example in the theory of household labour supply. Since ‘labour time’ is instantly perishable, there is no strictly speculative element to take into account (although someone seeking work may refuse a particular job offer because the wage offered is below a ‘reservation wage’ based on expectations as to the wage that can be obtained after further job searching). The conventional theory is, however, firmly based on viewing labour supply in terms of the ‘reservation demand’ for time not spent in market employment, and it is this that leads to the familiar argument that the income effect of a ‘wage’ change can both be large and contrary to the substitution effect, with the result that labour supply may be either positively or negatively related to the level of the ‘wage’. Analogous arguments bear on the supply of land services by landowners who have an ‘own use’ for their land, on the supply of agricultural products, and so on. The reservation price concept is also useful in the context of privately owned natural resources, a context which introduces two further determinants of reservation price. The lowest price at which a natural resource owner will be prepared to extract the resource will naturally depend on extraction costs, both the present extraction costs and those expected in the future; it will also depend on the expected growth rate, if any, of the resource. It is to be noted that the ‘neoclassical rule of free goods’ would never have to be applied to primary inputs for which (*a*) there was a positive price below which supply would be zero, and (*b*) demand at a zero price would be positive (both conditions holding for all prices of other commodities).

It was noted above, in connection with the market for first-edition copies of a book, that the ‘total’ demand curve diagram gives the same information with respect to price, and less information with respect to quantity, than does the more conventional supply and demand diagram. How then could P.H. Wicksteed – whose name is so strongly associated with the concept of a supply curve being merely a ‘reversed demand curve’ – have been so insistent that the former diagram is actually *superior* to the conventional one? (See Wicksteed, 1910, Book II, Ch. IV, and 1914.) Because the ‘total’ demand curve diagram emphasizes the idea that essentially the *same* kind of forces underlie the conventional supply curve as underlie the usual demand curve, thus breaking down the idea that there is an asymmetry in market forces, with subjective factors being dominant on the ‘demand side’ and objective ones on the ‘supply side’. The diagram in which a single demand curve (inclusive of reservation demand) confronts a fixed supply is at once congenial to any author both seeking to stress the subjective elements of the economic process and upholding the opportunity cost doctrine as against the real cost doctrine. While acknowledging that the demand and supply curves diagram illuminates the process through which the market clearing price is *discovered*, therefore, Wicksteed insisted that the other diagram brings out far more clearly the fundamental *determinants* of that price, namely, subjective marginal valuations and given supplies. With reference to continuously produced commodities, as opposed to first-edition copies, maintenance of this viewpoint would presumably require that the ‘given stocks’ referred to should be those of primary inputs. Here it may be noted that, even in the course of denouncing the conventional supply curve, Wicksteed admitted that ‘as we recede from the market and deal with long periods … cases may arise in which something like a “supply curve” seems legitimate’ and that nature does not have ‘reserve prices in which she expresses her own demand!’ (1914, p. 16, n.1).

See Also

- Wicksteed, Philip Henry

Bibliography

Wicksteed, P.H. 1910. *The Common Sense of Political Economy, Including a Study of the Human Basis of Economic Law*. London: Macmillan.

Wicksteed, P.H. 1914. The scope and method of political economy in the light of the ‘marginal’ theory of value and distribution. *Economic Journal* 24(March), 1–23.

How to cite this article

Steedman, Ian. "reservation price and reservation demand." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000113> doi:10.1057/9780230226203.1427

The New Palgrave Dictionary of Economics Online

residential real estate and finance

Peter Englund
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Residential real estate is a major asset for most households. This article focuses on three issues relating to housing as an investment. (*a*) Are returns to housing investment predictable? (*b*) What is the optimal fraction of real estate in an investment portfolio? (*c*) How important are borrowing constraints, and how do they influence housing prices? It concludes that housing risks are difficult to hedge in practice and that developing suitable derivative markets would fulfil an important function.

Keywords

asset pricing; Case–Shiller indexes; hedonic indexes; household portfolios; housing and mortgage markets; options; overlapping generations; portfolio analysis; residential real estate and finance; tenure choice

Article

Residential real estate is in any definition a major asset class. The average Swedish household invests three-quarters of its net wealth in its own home. Yet it was not until after 1990 that central questions in finance were asked about real estate. Is the market for real estate informationally efficient? What is the optimal fraction of real estate in a household portfolio? What role do financial constraints play in the pricing of real estate? These are particularly challenging questions in view of the special nature of residential real estate assets: properties are heterogeneous, transactions are infrequent, the trading parties are typically amateurs, and the market is best characterized as a search market where identical properties may trade at quite different prices. For all these reasons the data problems are of a different order of magnitude from those in the core areas of finance. Naturally, progress has been slow and we should not expect answers ever to be as sharp as for assets like stocks and bonds.
This article is organized around the questions posed above. Other areas, in particular the important field of mortgages and mortgage-backed securities, are not discussed.

Market efficiency

Standard theories of portfolio choice and asset pricing presume that markets are informationally efficient in the sense that it is impossible to make profits from trading strategies based on publicly available information, such as past returns. There is ample evidence indicating that real estate markets are not efficient in this sense. Time series studies of real estate returns typically find a strong pattern of positive autocorrelation on quarterly or yearly data (Case and Shiller, 1989; Englund and Ioannides, 1996). Such a pattern could in principle reflect time-varying risk premia, but this interpretation appears implausible. A problem with most studies of housing returns is that they measure only the time variation in the capital-gains part of returns and ignore the value of housing services (the implicit rent). An exception is Meese and Wallace (1994), which is based on micro evidence on unregulated rents. They confirm, for the San Francisco Bay Area, that returns on owner-occupied homes are indeed predictable based on past returns, but they also show that the profits involved are not sufficient to cover realistic transaction costs for a round-trip trade. There is no money to be made by shifting between renting and owning, with housing consumption fixed, but it may be profitable to time moves according to predicted returns. A general conclusion is that transaction costs in a broad sense are important in understanding real estate markets.

Portfolio choice

Research on portfolio choice has been hampered by a lack of reliable high-frequency data. Much recent research has been stimulated by the repeat-sales indexes for US metropolitan areas developed and analysed by Case and Shiller (1989). Goetzmann (1993) uses the Case–Shiller indexes to compute optimal portfolios (efficient frontiers) in mean-variance space, taking into account the idiosyncratic component of housing return, that is, the added risk of an individual home above the general return risk captured by a price index. He finds optimum housing shares to be on the order of 10–50 per cent of household net wealth depending on risk attitudes. It is well known from portfolio analysis of other assets that calculated portfolio shares are quite sensitive to input data, particularly expected return, and hence should be treated with caution. Nevertheless, later studies using data for European countries (using hedonic indexes not available in the United States) have obtained similar results (see, for example, Englund, Hwang and Quigley, 2002, for Stockholm; le Blanc and Lagarenne, 2004, for Paris; and Iacoviello and Ortalo-Magné, 2004, for London). The discrepancy between computed optimal portfolio shares and real world numbers, often in the order of several hundred per cent, is striking and has provided a challenge for further research.

The standard mean-variance analysis is obviously oversimplified in several ways. First, it is static. Grossman and Laroque (1990) consider lifetime portfolio choice when utility is derived from a durable good (housing), which can only be traded at a cost (proportional to house value). Housing trades are determined in analogy with Ss-models from inventory theory, and optimal portfolios are shown to be mean-variance efficient like in the static case.

Second, the standard analysis does not account for housing services as a consumption good separate from non-durable goods. Flavin and Yamashita (2002) analyse a two-good version of the Grossman and Laroque model with a stochastic relative price of housing. Based on correlations calculated from Case–

Shiller indexes, their model indicates that the optimal fraction of financial assets going into stocks is inversely related to the fraction invested in housing, and hence should increase with age, consistent with empirical observations. More recently some authors have analysed models with finite lifetimes, using numerical solution techniques. A key factor in determining the attractiveness of investing in housing is the correlation between labour income and the returns to housing: the stronger the correlation, the smaller is the optimal housing portfolio share.

Third, we have so far assumed housing to be consumed by owning, disregarding the alternative of renting. The issue of tenure choice is a classic one in the housing literature; see Henderson and Ioannides (1983) for a two-period model that brings out some of the basic features. Only rarely have issues of risk been included in the analysis. Among the exceptions are Rosen, Rosen and Holtz-Eakin (1984) and Turner (2003), who find that volatile house prices deter young households from entering into owner-occupancy. These studies do not explicitly measure the relative risks of owning compared with renting. More recently, Sinai and Souleles (2005) have emphasized that owning one's home is a way of hedging the risk associated with stochastic variations in the cost of renting. This is a particularly important aspect for households with a long expected stay in the same dwelling or the same housing market. Empirically, Sinai and Souleles confirm, for US households, that the probability of homeownership is indeed an increasing function of rent risk.

For most households, net wealth falls far short of the value of the house they demand for consumption purposes. Hence, any portfolio study that includes housing has to take a stand on the availability of borrowing. In fact, most households are constrained in their access to borrowing, at least when young, and financial constraints exert an important influence on savings and housing choices over the life cycle; see King (1980) for an early study emphasizing borrowing constraints. Integrating down-payment constraints into models of dynamic portfolio and tenure choice remains an important topic for future study.

Asset pricing and financing constraints

The standard approach to real estate price determination (as in Poterba, 1984) is explicitly couched in asset pricing terms: the price is the discounted value of the housing services generated by the property net of operating and maintenance costs. In principle, housing services could be valued based on market rents for comparable dwellings. In applying this approach, lip service is often paid to risk-adjusting the discount rate. It is fair to say, however, that there is no established theory or pragmatic consensus on the choice of discount rate. In recent years there has been a surge of interest in integrating housing into the standard asset-pricing paradigm. So far, however, interest has focused on the impact on financial asset prices of introducing housing collateral rather than on pricing real estate assets.

More attention has been paid to the direct impact of financial constraints on pricing. If the representative homebuyer is constrained by borrowing opportunities rather than by lifetime resources, then wealth shocks have a direct impact on housing demand. This implies that a shock to the demand and supply of housing services will be reinforced through its impact on financing constraints. An income shock, for example, will increase demand and housing prices, thereby releasing borrowing constraints. This will in turn give an extra boost to demand and prices. There will be a ‘financial multiplier’: the more important financial constraints are, the more sensitive prices will be to shocks to underlying fundamentals. This

view of real estate pricing was formulated by Stein (1995) and has been inserted into an overlapping-generations framework with demographic fluctuations by Ortalo-Magné and Rady (2006). Its empirical validity has been investigated in some studies. As an example, Lamont and Stein (1999) show that variations in the sensitivity of house prices to income shocks across US states can be explained by differences in loan-to-value ratios. Financial constraints may also explain the strong impact of variations in house prices on consumption observed in many studies; see, for example, Case, Quigley and Shiller (2005) for the United States and internationally.

Historically, mortgage lending has been further restricted by regulations in virtually all countries. Dismantling these regulations has in many cases caused price booms. But borrowing constraints remain important facts of life even in unregulated market environments, and there are large differences across countries even today, reflecting history and legal institutions. Chiuri and Jappelli (2003) show that average downpayment ratios vary from close to 50 per cent in Italy to a little above 10 per cent in Sweden and United Kingdom. They find that these differences, which they largely ascribe to legal tradition – relating to foreclosure, for example – explain differences in homeownership rates across countries, in particular the age when young households enter into owner occupancy.

The future

Not only has the area of real estate economics been lagging in its adoption of new analytical frameworks from finance, markets have also been slow in adopting new financial instruments and contracts to handle better the important risks many household confront in relation to their housing investment. While households have access to a wide variety of mortgage instruments, markets remain seriously incomplete and fail to offer flexible and liquid contracts related to housing price risks. As Robert Shiller (2003) has forcefully argued, this is one of the macro risks in society that remain uninsurable despite their fundamental importance for individual welfare. Options or futures on relevant housing price indexes could go a long way towards providing such insurance. It remains to be seen how long it will take to develop liquid markets in such instruments.

See Also

- capital asset pricing model
- efficient markets hypothesis
- household portfolios
- housing supply

Bibliography

Case, K., Quigley, J. and Shiller, R. 2005. Comparing wealth effects: the stock market versus the housing market. *Advances in Macroeconomics* 5, 1–32.

Case, K. and Shiller, R. 1989. The efficiency of the market for single-family homes. *American*

Economic Review 79, 125–37.

Chiuri, M. and Jappelli, T. 2003. Financial market imperfections and homeownership: a comparative study. *European Economic Review* 47, 857–75.

Englund, P., Hwang, M. and Quigley, J. 2002. Hedging housing risk. *Journal of Real Estate Finance and Economics* 24, 167–200.

Englund, P. and Ioannides, Y. 1996. House price dynamics: an international empirical perspective. *Journal of Housing Economics* 6, 119–36.

Flavin, M. and Yamashita, T. 2002. Owner-occupied housing and the composition of the household portfolio. *American Economic Review* 92, 345–62.

Goetzmann, W. 1993. The single family home in the investment portfolio. *Journal of Real Estate Finance and Economics* 6, 201–22.

Grossman, S. and Laroque, G. 1990. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. *Econometrica* 58, 25–51.

Henderson, J. and Ioannides, Y. 1983. A model of housing tenure choice. *American Economic Review* 73, 98–113.

Iacoviello, M. and Ortalo-Magné, F. 2004. Hedging housing risk in London. *Journal of Real Estate Finance and Economics* 27, 191–209.

King, M. 1980. An econometric model of tenure choice and demand for housing as a joint decision. *Journal of Public Economics* 53, 137–59.

Lamont, O. and Stein, J. 1999. Leverage and house-price dynamics in U.S. cities. *RAND Journal of Economics* 30, 498–514.

le Blanc, D. and Lagarenne, C. 2004. Owner-Occupied housing and the composition of the household portfolio: the case of France. *Journal of Real Estate Finance and Economics* 29, 259–75.

Meese, R. and Wallace, N. 1994. Testing the present value relation for housing prices: should I leave my house in San Francisco. *Journal of Urban Economics* 35, 245–66.

Ortalo-Magné, F. and Rady, S. 2006. Housing market dynamics: on the contribution of income shocks and credit constraints. *Review of Economic Studies* 73, 459–85.

Poterba, J. 1984. Tax subsidies to owner-occupied housing: an asset-market approach. *Quarterly Journal of Economics* 99, 729–52.

Rosen, H., Rosen, K. and Holtz-Eakin, D. 1984. Housing Tenure, uncertainty, and taxation. *Review of Economics and Statistics* 66, 405–16.

Shiller, R. 2003. *The New Financial Order: Risk in the 21st Century*. Princeton: Princeton University Press.

Sinai, T. and Souleles, N. 2005. Owner-occupied housing as a hedge against rent risk. *Quarterly Journal of Economics* 120, 763–89.

Stein, J. 1995. Prices and trading volume in the housing market: a model with downpayment constraints. *Quarterly Journal of Economics* 110, 379–406.

Turner, T. 2003. Does investment risk affect the housing decisions of families? *Economic Inquiry* 41, 675–91.

How to cite this article

Englund, Peter. "residential real estate and finance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000233> doi:10.1057/9780230226203.1428

The New Palgrave Dictionary of Economics Online

residential segregation

Jacob L. Vigdor
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Housing market equilibria display residential segregation when there are systematic disparities in the physical location of households belonging to different racial, ethnic, socio-economic, or other social groups. Historically, segregation has often been enforced through non-market processes such as legal restrictions. Modern segregation, by contrast, is largely driven by cross-group differences in willingness to pay for housing in group enclaves. Segregation often generates social concern, particularly when the segregated group is of low socio-economic status. Empirical studies, including a few based on randomized mobility experiments, suggest that there are negative consequences of growing up in an enclave neighbourhood.

Keywords

Census data; dissimilarity index; ethnic identity; ghettoes; housing markets; immigration; inequality; internal migration; racial segregation; residential integration; residential segregation; socio-economic segregation; spatial mismatch hypothesis; spectral segregation index; tipping; zoning

Article

The term ‘residential segregation’ describes a housing market equilibrium marked by systematic disparities in the physical location of households belonging to different racial, ethnic, socio-economic, or other social groups.

While history is replete with examples of groups forced to live in complete isolation from the remainder of society, residential segregation is not inherently a dichotomous phenomenon. Rather, housing markets may exhibit varying degrees of segregation; social scientists have endeavoured to quantify this variation for the better part of a century. The term ‘ghetto’ is often ascribed to social groups experiencing segregation that exceeds a loosely defined threshold.

Residential segregation may be the outcome of a past residential sorting process wherein centralized authorities restricted some agents' location choices. Very simple economic theory, and an increasing amount of empirical evidence, however, point to the conclusion that modern-day residential segregation is driven primarily by the operation of decentralized market forces.

Even if residential segregation is a pure market phenomenon, many observers harbour concerns that segregated housing market equilibria are suboptimal from a social welfare perspective. Some debate exists as to whether segregated housing markets are inefficient. Arguments hinge on whether households are fully informed at the time they make location decisions, or whether they face borrowing constraints. It is a less controversial observation that residential segregation has important implications for distributional equity. In segregated equilibria, for example, wealthy households have the opportunity to avoid subsidizing the local public good consumption of poorer households. Over the past several decades, there have been many attempts to estimate the relationship between residential segregation and inequality between groups, in both the short term and the long term.

Here, basic evidence is provided on the existence and magnitude of contemporary residential segregation. This evidence draws heavily on the experience of racial and immigrant groups in the United States; the measurement of segregation in other nations is limited in scope and often confined to very recent observations. As discussed below, this is more a reflection of data limitations than any genuine lack of interest. The basic economic theory of why segregation exists is then outlined, and empirical evidence that has been brought to bear on the issue discussed. The concluding discussion considers the potential implications of segregation on socio-economic outcomes and human capital investment.

Measuring segregation

There are many ways to measure segregation (Massey and Denton, 1988). The metrics most commonly used in sociology and economics require the existence of neighbourhood-level data on the distribution of groups in a city or region. Some measures, including the spectral segregation index (Echenique and Fryer, 2005), require additional data on the physical location of these neighbourhoods and, in some cases, their land area. A central challenge to the systematic measurement of segregation is the lack of comparable neighbourhood-level data across nations, or even cities within nations, and over time. The United States, for example, has collected data on the race of its inhabitants since 1790, but did not report race at a consistently defined neighbourhood level until 1940. The United Kingdom did not systematically collect information on the ethnic identity of its inhabitants until the 1991 Census. Given the existence of required data, the most commonly used segregation indices classify the residential separation of any particular group between two extremes: perfect segregation, where group members never share a neighbourhood with individuals not belonging to the group, and perfect integration, where group members form an equal share of the population in all neighbourhoods. The dissimilarity index (Duncan and Duncan, 1955), records groups on the scale from 0 (perfectly integrated) to 1 (perfectly segregated) using the following formula:

$$D = \frac{1}{2}\sum_i \left| \frac{A_i}{A} - \frac{B_i}{B} \right| \quad (1)$$

where i indexes neighborhoods, A_i and B_i represent the number of group members and others in neighborhood i, respectively, and A and B represent the total population of group members and others in the city or region. The dissimilarity index has a relatively intuitive interpretation: it is the share of group members, or others, who would have to switch neighbourhoods in order to achieve perfect integration. While many demographers state a preference for other indices based on various criteria, the dissimilarity index is most commonly used in existing literature.

Stylized facts

The absence of neighbourhood-level data makes it difficult to gauge the contemporary level of segregation in many cities, let alone historical levels. The most comprehensive historical data pertain to American cities. Cutler, Glaeser and Vigdor (1999; 2005) use these data to compute long-term trends in dissimilarity for African-Americans and foreign-born individuals, respectively. Figure 1 plots weighted averages of these measures across cities, with weights equal to the population of the group in question in each city. Immigrant segregation is computed separately for each country-of-origin group in each city; the immigrant time series represents the weighted average of these data.

Figure 1
Dissimilarity of African-Americans and immigrants in the United States, 1890–2000

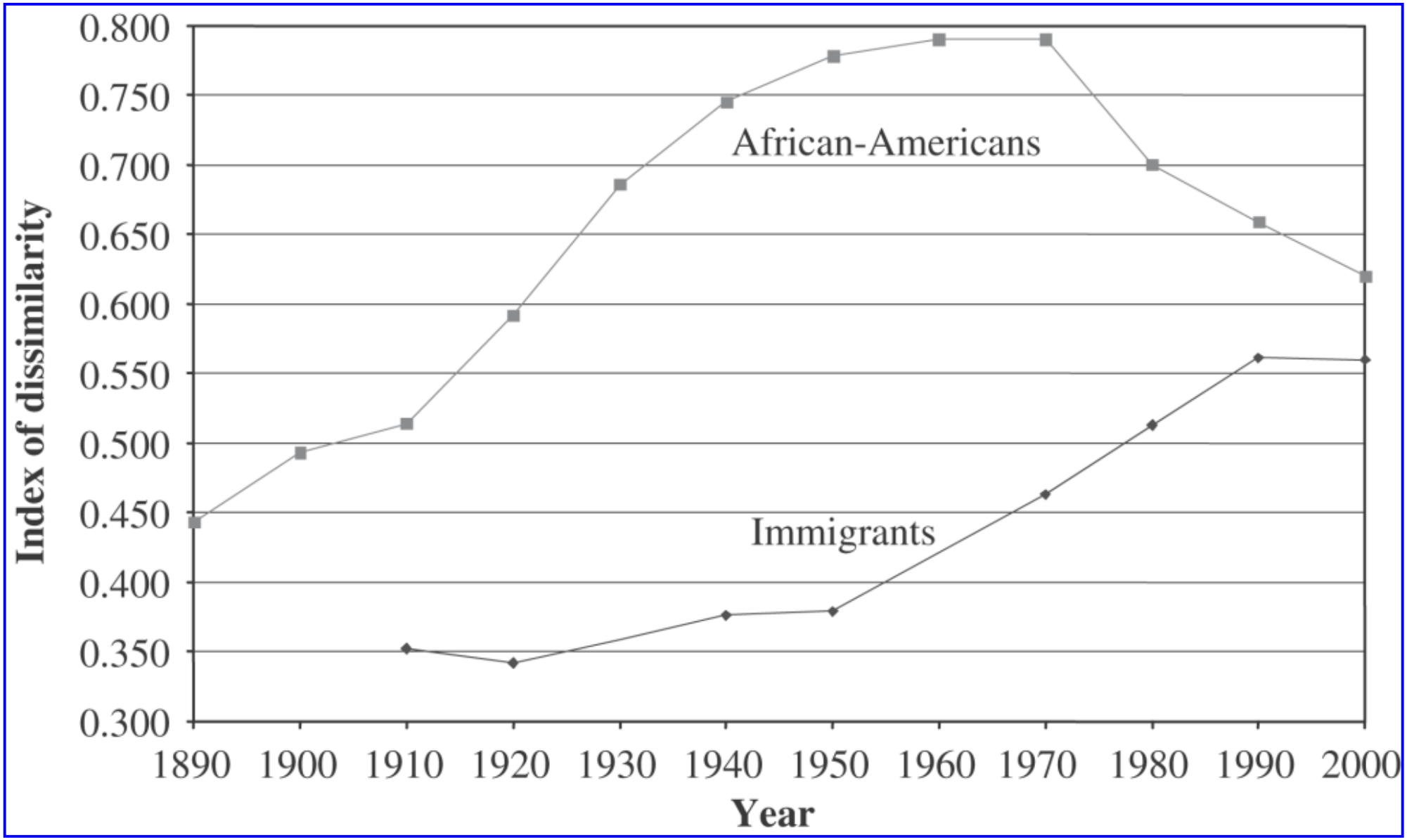

As the relatively low initial levels of black–white dissimilarity indicate, urban ghetto neighbourhoods were relatively uncommon in the United States at the turn of the 20th century. The birth of the African-American ghetto coincided with the so-called Great Migration of blacks from the southern part of the United States to northern cities between the First World War and 1965. Segregation reached its peak at the end of this period of migration. In 1970 dissimilarity levels in some areas, chiefly large industrial cities of the north-eastern and mid-western United States, were at or near 0.90. Since that time, segregation has fallen pervasively throughout the nation but most acutely in rapidly growing cities in the southern and western parts of the country.

Immigrant segregation, quite strikingly, displays the opposite trend to racial segregation in the United States. Immigrant dissimilarity remained stable at relatively low levels through the first half of the century, then rose steadily. Thus, even as racial ghettos have declined over the past few decades, immigrant segregation has risen. Cutler, Glaeser and Vigdor (2005) present further data indicating that the rise in average segregation can be attributed primarily to the growth of groups that have always experienced high segregation, rather than to the increasing segregation of individual groups. The growing, highly segregated groups generally originate in less developed countries and gravitate toward the largest cities in the United States. The limited amount of data available from other nations supports the general trend found in the United States: individual racial and ethnic group are experiencing

stable or declining segregation in most parts of the world.

Socio-economic segregation, or the degree to which households in poverty tend to cluster together in neighbourhoods, increased in the United States in the 1970s and 1980s, but showed some evidence of lessening in the 1990s.

Table 1 presents some representative dissimilarity index values for groups in some of the world's largest cities, using recently available data.

Recent dissimilarity indices for various groups in major world cities

City	Group	Year	Dissimilarity
Barcelona	Latin American immigrants	2001	0.290
Cape Town	Blacks*	1996	0.928
Chicago	African-Americans	2000	0.778
Cologne	Turkish immigrants**	1994	0.337
Lima	High SES households	1993	0.440
London	Blacks (Caribbean, African and Other)*	2001	0.468
London	South Asians*	2001	0.544
Los Angeles	Mexican immigrants	2000	0.446
Mexico City	High SES households	2000	0.380
New York	African-Americans	2000	0.670
Santiago	High SES households	1992	0.490
Tokyo	Individuals over 65	1995	0.147

Note: dissimilarity indices measure the separation of each group from the remainder of the population, except as indicated.

*Dissimilarity from whites•**Dissimilarity from Germans•SES: socio-economic status

Sources: Barcelona: Martori i Cañas and Hoberg (2004); Cape Town: Rospabe and Selod (2003); Chicago and New York: Glaeser and Vigdor (2002); Cologne: Friedrichs (1998); Lima, Mexico City and Santiago: Arriagada Luco and Vignoli (2003); London: Burgess, Wilson and Lupton (2005); Los Angeles: Cutler, Glaeser and Vigdor (2005); Tokyo: Nakagawa (2003).

Why are groups segregated?

Theories of racial segregation can be classified into two types. The first type permits some form of discrimination in housing markets. The second models segregation as the equilibrium outcome of a fully competitive market. A potential third class of models explains one form of segregation as the direct consequence of a second form – for example, it explains racial segregation as a consequence of economic segregation. This third class is of less interest to attempts to explain segregation more generally.

In a discrimination-based model, location choices are constrained for members of one group, defined by race, ethnicity, or other observable characteristic. The constraints on location choice might include explicit legal barriers or implicit patterns of ‘steering’ households towards certain locations. Historical examples of explicit legal barriers abound. In a few cases, governments have attempted to restrict location choices as a matter of public law, have enforced contracts between private parties restricting racial ownership or occupancy of property, or have adopted policies that had the effect of limiting the residential options of certain groups. In the United States, federal legislation had made most explicit forms of housing market discrimination illegal by the end of the 1960s.

While few observers would argue that explicit racial or ethnic barriers to location choice persist in the developed world, the existence and prevalence of implicit discriminatory patterns is a subject of continuing debate. Government policies such as zoning laws, which local governments use to regulate the density and nature of residential development within their borders, may implicitly perpetuate segregation. Housing audit studies provide evidence of discriminatory behaviour among real estate agents, mortgage brokers, or landlords. In these studies, auditors of different races present carefully matched, fictionalized credentials to housing market agents. The behaviour of these agents is then analysed to uncover any systematic differences in treatment by race. Recent studies, such as Ondrich, Ross and Yinger (2003), find evidence of significant racial disparities in treatment in the United States.

While disparities in treatment of housing market auditors can be interpreted as evidence of continued racism, such behaviour can also be consistent with unbiased, profit-maximizing motives. As in models of statistical discrimination in labour markets or other settings, agent behaviour could be motivated by accurate perceptions of differences in average preferences across racial groups.

Such an interpretation is consistent with the second type of racial segregation theory, which posits that segregated housing market equilibria are fully consistent with decentralized, unconstrained household choices. Preference-based theories of segregation owe some intellectual debt to Tiebout's (1956) vision of residential sorting, but evolve most clearly from Schelling's (1978) simulation of residential sorting in the presence of very slight preferences for neighbours of one's own group. Schelling's simulations show that a small initial concentration of same-group neighbours can rapidly evolve into a vast enclave community. This process of ‘tipping’ is driven by group members' heightened willingness to pay for locations in close proximity to the initial cluster. As the enclave grows in size, it becomes disproportionately more attractive to group members than to others. So long as the segregated group in question maintains a steady population share in the entire market, the enclave is very unlikely to dissipate.

Much anecdotal evidence supports the Schelling model. The neighbourhood integration that has taken place in the United States since 1970, for example, has left most African-American enclaves untouched. Rather, integration has occurred either in newly developed neighbourhoods on the fringe of urban areas or in locations marked by significant demolition and redevelopment.

While intuitively appealing and supported by anecdotal observation, true empirical tests of preference-based theories are rendered difficult by the unobservability of household preferences. Econometric models associated with the measurement of willingness to pay, such as discrete choice models, often assume away the existence of housing market discrimination (for example, Bayer, McMillan and Rueben, 2004). Survey-based methods of eliciting preferences are valid only to the extent that respondents can accurately separate their valuation of neighbourhood racial composition from all other attributes, and truthfully reveal this valuation. What survey evidence that exists supports the notion that groups harbour preferences for same-group neighbours (Vigdor, 2003).

Why might individuals care about the racial or ethnic composition of their neighbourhoods? Group members may prefer to congregate in enclaves in order to take advantage of scale economies enabling the supply of group-specific community institutions or consumer goods. Individuals may also seek to limit exposure to other groups on the basis of stereotyped perceptions of inferiority, greater criminality, or other characteristics. It is also possible that individuals care, not about the race of their neighbors directly, but about characteristics correlated with race, such as socio-economic status. These varying hypotheses have dramatically different implications for the social value of segregation. Unfortunately, these various explanations are observationally equivalent. Each predicts that segregation occurs in equilibrium because willingness to pay for housing in a group enclave is relatively higher among group members.

While there is currently no consensus on the importance of housing market discrimination in perpetuating segregation, Cutler, Glaeser and Vigdor (1999) present evidence that any such importance has declined. In 1940, at a time when many forms of housing market discrimination were legal – and in some cases practised by government itself – restrictions on African-American location choice had the impact of increasing equilibrium prices in segregated areas. By 1970 that premium had disappeared, suggesting that these artificial barriers to mobility had been removed.

Does segregation influence economic outcomes?

A number of hypothesized causal mechanisms link segregation to socio-economic outcomes. The ‘spatial mismatch’ hypothesis contends that segregation reduces the average income of certain groups to the extent that their residential enclaves are located at some distance from growing employment centres (Kain, 1968). Segregation may also lead to differences in education quality across racial or ethnic groups, to the extent that schooling is tied to residential location. Finally, there may be other localized factors that differ across neighbourhoods and have the net impact of leading to different human capital investment trajectories. For example, children growing up in different neighbourhoods may develop different consumption or investment preferences by being exposed to different types of role models.

Numerous attempts have been made to empirically estimate the impact of segregation on outcomes, whether operating immediately through spatial mismatch-type mechanisms or developmentally. Much of this empirical literature is plagued by a fundamental endogeneity problem: since individuals choose their own neighbourhoods, any correlation between neighbourhood characteristics and individual outcomes might reflect selection rather than any causal effect of the former on the latter. Researchers have implemented three strategies for circumventing these selection problems. The first is to focus on the outcomes of young adults, whose location choices are presumably determined by their parents rather than themselves. Vigdor (2002) points out that the strategy of studying young adults is suspect in the presence of inter-generational transmission of economic outcomes.

A second basic strategy for identifying the impact of segregation on outcomes in the presence of selective migration is to model location choice and socio-economic outcomes simultaneously. Some research in this vein makes use of individual data-sets with detailed geographic identification, recently made available by the US Census Bureau. A simultaneous equation model can uncover the true causal impact of segregation on outcomes if it employs an instrumental variable – a factor than affects location choice but otherwise bears no correlation to individual outcomes. In practice, identifying a valid instrumental variable is very difficult.

Recently, researchers have addressed selective migration concerns by turning their attention to randomized mobility experiments, in which a ‘treatment’ group is offered a voucher redeemable for housing only in certain neighbourhoods, while a ‘control’ group is offered no such aid. While these experiments generally do not permit examination of the causal impact of segregation per se, they do allow a more general study of the potential importance of neighbourhood characteristics in determining outcomes. In general, studies find little impact of neighbourhood factors on the socio-economic outcomes of adults. There is more evidence in favour of developmental impacts on youth. Orr et al. (2003) present an overview of research results stemming from one such randomized mobility experiment, the Moving to Opportunity demonstration programme.

See Also

- finance
- ghettoes
- housing policy in the United States
- immigration and the city
- neighbours and neighbourhoods
- spatial mismatch hypothesis
- symmetry breaking
- urban economics
- urban housing demand

Bibliography

Arriagada Luco, C. and Vignoli, J. 2003. Segregación residencial en áreas metropolitanas de América Latina: magnitud, caracteristicas, evolución e implicaciones de politica. Series Población y Desarrollo No. 47. Santiago de Chile: UN Comisión Económica para América Latina y El Caribe.

Bayer, P., McMillan, R. and Rueben, K. 2004. Residential segregation in general equilibrium. Working Paper no. 11095. Cambridge. MA: NBER.

Burgess, S., Wilson, D. and Lupton, R. 2005. Parallel Lives? Ethnic segregation across schools and neighborhoods. *Urban Studies* 42, 1027–56.

Cutler, D., Glaeser, E. and Vigdor, J. 1999. The rise and decline of the American ghetto. *Journal of Political Economy* 107, 455–506.

Cutler, D., Glaeser, E. and Vigdor, J. 2005. Is the melting pot still hot? Explaining the resurgence of immigrant segregation. Working Paper No. 11295. Cambridge, MA: NBER.

Duncan, O. and Duncan, B. 1955. A methodological analysis of segregation indexes. *American Sociological Review* 20, 210–17.

Echenique, F. and Fryer, R.G., Jr. 2005. On the measurement of segregation. Working Paper No. 11258. Cambridge, MA: NBER.

Friedrichs, J. 1998. Ethnic segregation in Cologne, Germany, 1984–1994. *Urban Studies* 35, 1745–63.

Glaeser, E. and Vigdor, J. 2002. Residential segregation: promising news. In *Redefining Urban & Suburban America: Evidence from Census 2000*, vol. 1, ed. B. Katz and R. Lang. Washington, DC: Brookings Institution Press.

Kain, J. 1968. Housing segregation, negro employment, and metropolitan decentralization. *Quarterly Journal of Economics* 82, 175–97.

Martori i Cañas, J. and Hoberg, K. 2004. Indicadores cuantativos de segregación residencial. El caso de la población inmigrante en Barcelona. *Scripta Nova* 8 (169). Online. Available at http://www.ub.es/geocrit/sn/sn-169.htm, accessed 1 August 2005.

Massey, D. and Denton, N. 1988. The dimensions of racial segregation. *Social Forces* 67, 281–315.

Nakagawa, M. 2003. Analysis of age discrimination in the rental housing Market in Japan: an approach using a fair housing audit. Discussion paper No. 577. Osaka: Institute of

Social and Economic Research, Osaka University.

Ondrich, J., Ross, S. and Yinger, J. 2003. Now you see it, now you don't: why do real estate agents withhold available houses from black customers? *Review of Economics and Statistics* 85, 854–73.

Orr, L., Feins, J., Jacob, R., Beecroft, E., Sanbonmatsu, L., Katz, L., Liebman, J. and Kling, J. 2003. *Moving to Opportunity: Interim Impacts Evaluation*. Washington, DC: Office of Policy Development and Research, US Department of Housing and Urban Development.

Rospabe, S. and Selod, H. 2003. Does city structure cause unemployment? The case study of Cape Town. Unpublished manuscript. Paris: Laboratoire d'Economie Appliquée.

Schelling, T. 1978. *Micromotives and Macrobehavior*. New York: W.W. Norton.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Vigdor, J. 2002. Locations, outcomes, and selective migration. *Review of Economics and Statistics* 84, 751–55.

Vigdor, J. 2003. Residential segregation and preference misalignment. *Journal of Urban Economics* 54, 587–609.

How to cite this article

Vigdor, Jacob L. "residential segregation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000228> doi:10.1057/9780230226203.1429

The New Palgrave Dictionary of Economics Online

reswitching of technique

Roberto Scazzieri
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Reswitching of technique is the property whereby, when multiple production techniques are available in a wage–profit economy, the same technique may be optimal at different levels of the rate of interest. This means that a virtual movement of the rate of interest in a given direction might make it rational to use techniques that had been previously excluded, so that the rate of interest cannot provide an unambiguous ranking of techniques. This possibility is rooted in the complex interactions (movements of relative prices) that occur in a production economy with different proportions between labour and intermediate inputs.

Keywords

capital deepening; capital intensity; factor-price frontiers; intermediate products; recurrence of technique; reswitching of technique; technical choice

Article

Reswitching of technique refers to the virtual adoption of production techniques, either by the individual producer or by the economic system as a whole. Standard economic theory treats technical adoption on the assumption that there is a multiplicity of techniques for producing any given good, and that the producer, as a rational decision maker, will switch from one technique to another according to a certain hypothetical sequence as the prices of productive factors are changed. This sequence would depend on the ranking of techniques in terms of capital per man or ‘capital intensity’, so that a lower rate of interest (which is equal to the rate of profit in equilibrium) would be associated with the ‘adoption’ of a technique characterized by higher capital per man. This process is known as capital deepening.

The development of discrete production models in the 1950s led to the discovery that this view of ‘rational’ technical adoption is not necessarily well founded. David Champernowne (1953) and Joan Robinson (1956) pointed out that a movement of the rate of interest in a given direction might make it optimal once again to use techniques that had been previously excluded. This phenomenon is known as *reswitching of technique*.

The original discovery was associated with the belief that reswitching was nothing more than a ‘curiosum’, which could not be left out on grounds of pure logic but was nevertheless unlikely to happen. The discussion of this phenomenon by Piero Sraffa (1960) showed that reswitching is the normal outcome of a situation in which the various production processes are characterized by different proportions between ‘direct’ labour and the quantity of ‘past’ labour. (This latter is the quantity of labour that is indirectly required in a production process, being required in producing its intermediate inputs.) Sraffa's analysis also provides a clear insight into the reasons for technical reswitching along the hypothetical sequence associated with changes in the rate of profit. It is worthwhile considering his example in some detail.

The ‘pure products’ case

It is useful to start with the consideration of a special category of commodities, which we might call of the *pure product* type. These are commodities that are never used as productive inputs, so that their price reflects production cost, but cost is never influenced by the variation of their particular prices.

Let *a* and *b* be commodities of that type, and let them be produced with different proportions of direct labour to past labour. (This structure of labour requirements is representative of

the differences in the proportions between labour and intermediate inputs in the production processes of the two commodities.)
Let *a* require more labour than *b* if we consider labour applied eight years before the year in which the product is ready, whereas *b* requires more labour than *a* in the cases of labour applied in the current year and 25 years earlier. This situation may be represented as follows (*n* is the date at which labour is applied):

1. (i) $n = 8$, $l_{a(8)} = y + 20$, $l_{b(8)} = y$
2. (ii) $n = 0$, $l_{a(0)} = x$, $l_{b(0)} = x + 19$
3. (iii) $n = 25$, $l_{a(25)} = y$, $l_{b(25)} = y + 1$

We are now in a position to examine in which way the cost difference between the two products may vary if the rate of profit is raised from 0 to a maximum value of 25 per cent. (An increase of the rate of profit is equivalent to a change in the weight of the different labour terms in each cost equation.)
The cost difference is expressed by the following equation:

$$p_a - p_b = 20w(1 + r)^8 - \left[19w + w(1 + r)^{25}\right]. \quad (1)$$

On the assumption that the wage rate (*w*) is inversely related to the rate of profit according to the following expression:

$$w = 1 - \frac{r}{25\%},$$

the cost difference equation will be represented by the curve in Figure 1.
Figure 1

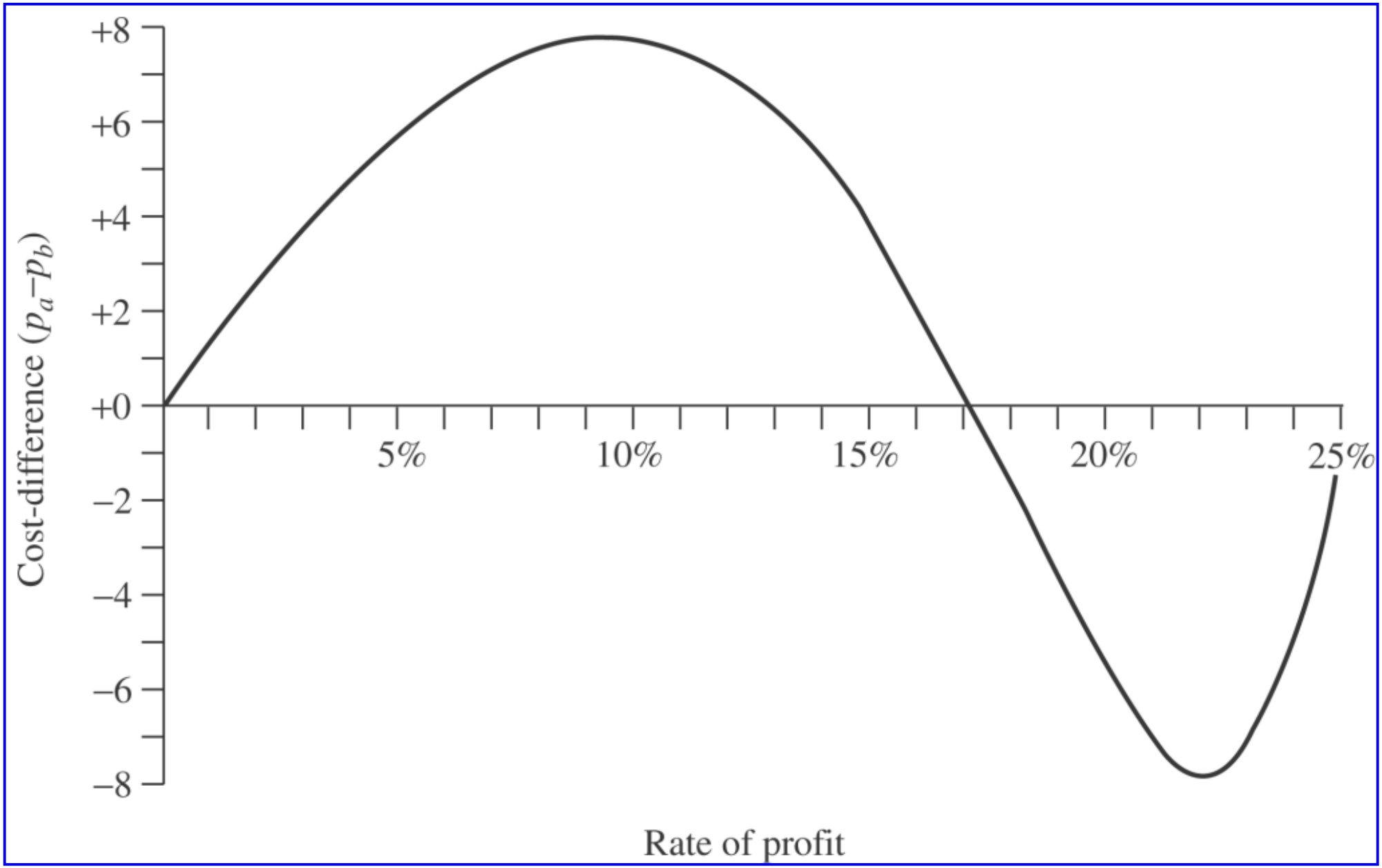

The cost of *a* rises relatively to *b* as *r* increases between zero and nine per cent. The reason for this is that the change of *r* leaves the value of current labour unaffected, whereas the 'excess labour' of date 8 is much greater than the excess labour of date 25. The increase in the value of $l_{b(25)}$ is more than offset by the increase in the value of $l_{a(8)}$ and the compound effect of these two variations is an increase in the cost difference. Beyond $r = 9\%$, the increasing weight of remote labour terms brings the cost difference down. This reduction stops at $r = 22\%$, since at this particular level of the rate of profit the decline of the wage rate starts offsetting the increase in the value of remote labour terms due to a higher *r*.

The above argument has straightforward implications for technical choice in the case of commodities of the pure product type. For in this case we can take for granted that the price of each commodity reflects its cost of production, whereas this price has no influence at all on the cost. Under such conditions, eq. (1) permits us to examine in which way the relative profitability of two techniques is varied as *r* goes from 0 to *r* (max). In fact, we may take eq. (1) to illustrate the difference between the unit costs of production of the same commodity produced with two alternative techniques. (For reasons of symmetry with the previous argument we call such alternative techniques *a* and *b* respectively.) Figure 1 can be

applied to this particular case. An immediate shortcoming would be that a change in the price of direct to 'dated' labour, as reflected in an increasing *r*, is associated with a positive excess of unit cost p_a over unit cost p_b until the curve intersects the horizontal axis for the first time. This involves that, over this interval, technique *b* is more profitable than technique *a*. A further increase of *r* (until *r* (max)) is associated with a negative difference $(p_a - p_b)$, so that technique *a* is more profitable than technique *b*. However, the same figure shows that the *reduction* of the cost difference stops at $r = 22\%$. For any *r* such that $22\% < r < r(\max)$, the cost difference is increasing once again. This increase stops at $r = 25\%$, when techniques *a* and *b* become equally profitable.

The movement of the cost difference when *r* is increasing shows that the relative profitability of techniques *a* and *b* is subject to fluctuations which depend on the particular interval within which *r* is changed. The relative profitability of technique *a* with respect to technique *b* is initially decreasing, then increasing, finally decreasing again. These fluctuations show that the 'unevenness' of the input structure may bring about multiple switches between the two techniques as we consider a steadily increasing *r*: the same technique might be adopted at low and high rates of profit, with the alternative technique being adopted at intermediate levels of *r*.

The 'intermediate products' case

It might appear that the above picture gets greatly complicated when we consider the more general case of products that are used as productive inputs either of themselves or of other commodities. For in this new situation the price of a commodity reflects its production cost, but this cost might itself be influenced by that price. (Directly in the case of a product used in its own production, indirectly in the case of a product that is, at some stage, a necessary means of production for at least one of its inputs.)

An immediate consequence of the consideration of interdependence between production processes is that inspection of the cost difference equation is no longer sufficient in order to assess the relative profitability of alternative techniques. The mutual influence between prices and production costs brings about the need of comparing systems of interrelated techniques (production technologies) rather than individual techniques. This requires consideration of the price system that will be associated with each technology at any given distribution of income between wages and profits.

The analysis of the 'intermediate products' case can be carried out by examining a simple model with two alternative two-good technologies A and B, in which all products are used as inputs of themselves and of the other commodity. We shall also assume that the two technologies differ only in the technique used to produce commodity 1.

The two price systems may be written as follows:

$$(a_{11}p_1 + a_{21}p_2)(1 + r) + l_1(a)w = p_1$$

$$(a_{12}p_1 + a_{22}p_2)(1 + r) + l_2(a)w = p_2, \tag{2.1}$$

$$(b_{11}p_1 + b_{21}p_2)(1 + r) + l_1(b)w = p_1$$

$$(b_{12}p_1 + b_{22}p_2)(1 + r) + l_2(b)w = p_2, \tag{2.2}$$

where $a_{ij}(i, j = 1, 2)$ and $b_{ij}(i, j = 1, 2)$ are the quantities of commodity *i* required to produce one unit of commodity *j* with technologies A and B respectively, $l_i(a)$ and $l_i(b)$ are the quantities of labour entering one unit of commodity *i* with technologies A and B respectively, $p_i(i = 1, 2)$ is the price of product *i*, *w* is the unit wage and *r* the rate of profit. The quantities a_{ij}, b_{ij}, $l_i(a)$ and $l_i(b)$ are known, whereas *r*, *w*, p_i are unknown.

Either product is common to both systems. We may thus choose either commodity 1 or 2 as the common standard of prices (*numéraire*) in both systems. If we put the price of commodity 1 equal to unity, commodity 1 becomes the common *numéraire* of price systems (2.1) and (2.2). At this stage, it is found convenient to assess the relative profitability of alternative technologies by considering the functional relationship between *r* and *w* for each technology.

The systems of eqs. (2.1) and (2.2) would each be associated with a particular relation between the rate of profit and the unit wage. The wage–profit relationships for the two systems would respectively be given by the following expressions:

$$w_A = \frac{1 - (a_{22} + a_{11})(1 + r) + (a_{11}a_{22} - a_{21}a_{12})(1 + r)^2}{(1 + r)[a_{21}l_2(a) - a_{22}l_1(a)] + l_1(a)} \tag{3.1}$$

$$w_B = \frac{1 - (b_{22} + b_{11})(1 + r) + (b_{11}b_{22} - b_{21}b_{12})(1 + r)^2}{(1 + r)[b_{21}l_2(b) - b_{22}l_1(b)] + l_1(b)} \tag{3.2}$$

It may be immediately noted that *w* is always a decreasing function of *r*, independently of the sign of the second order derivative (see also Morishima, 1966, p. 521). We may also note that the unit wage is expressed in terms of the same numéraire in (3.1) and (3.2). This suggests that the relationships between *r* and *w* (also known as factor-price frontiers) can be plotted as negatively sloped curves on the same diagram.

The intersections between the two curves occur at those levels of the rate of profit which are associated with the same unit wage in both technologies. The number of intersections can be obtained by equating *w* in eqs. (3.1) and (3.2) and solving for *r*. The resulting equation will generally have more than one positive solution (Bruno, Burmeister and Sheshinski, 1966, p. 534). In the case of technologies such that each product is a necessary input for all commodities including itself (all products are *basic commodities*), the maximum number of intersections is given by the number of distinct commodities in the two alternative systems of production (Bharadwaj, 1970). This implies that, in the two-good technologies of our example, there will be at most two intersections. Figure 2 represents a case in which there are two intersections in the positive quadrant.

Figure 2

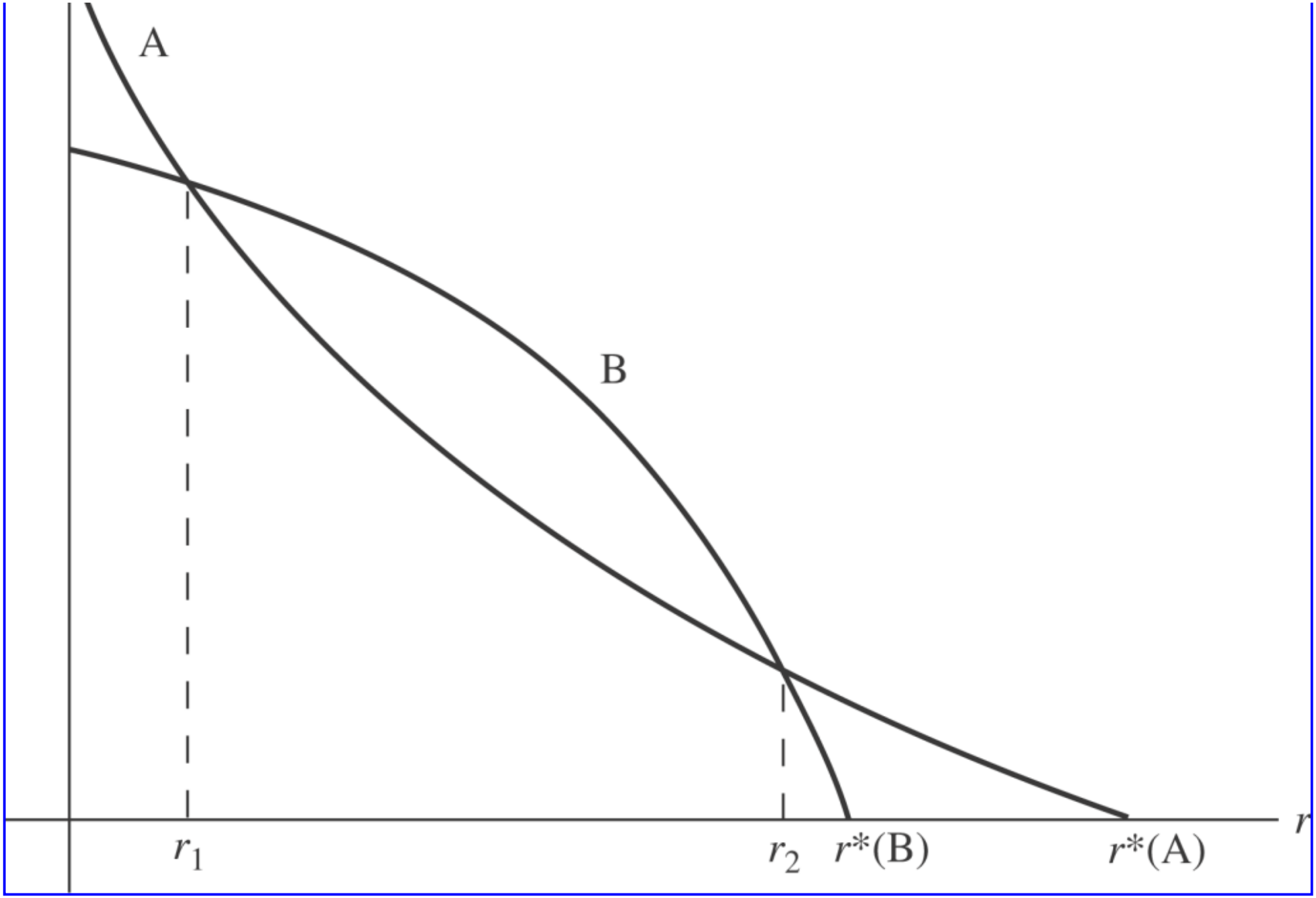

Technologies A and B can now be compared, on grounds of profitability, by considering which technology yields the higher rate of profit for any given wage. (Or, alternatively, which technology yields the higher wage rate for any given rate of profit.)

Figure 2 makes clear that the relative profitability of the two technologies is subject to fluctuation as r increases from 0 to r^*(B) (the maximum rate of profit with technology B). At a low level of the rate of profit $(r < r_1)$, technology A is more profitable ('cheaper') than B. At $r = r_1 = r_2$, A and B are equally profitable. At levels of r between r_1 and r_2, B is more profitable than A. But at any rate of profit higher than r_2, A is again more profitable than B.

Reswitching of technique may be shown to be possible between complete production systems as well as between individual techniques. Shortly after the identification of the reswitching possibility by Champernowne (1953) and Robinson (1956), and its subsequent analysis by Sraffa (1960), Morishima (1964) and Hicks (1965), David Levhari (1966) proposed the argument that reswitching between production systems is possible only in the case of a ‘reducible’ or ‘decomposable’ technology matrix, so that reswitching would not occur with technologies producing only basic commodities (‘irreducible’ or ‘indecomposable’ technologies). Levhari's argument was disproved by Pasinetti and others (Pasinetti, 1966; Morishima, 1966; Garegnani, 1966). It was also acknowledged to be false by Levhari and Samuelson (Levhari and Samuelson, 1966; Samuelson, 1966). Conditions excluding reswitching were then discovered by Bruno, Burmeister and Sheshinski (1966) and other authors (Starrett, 1969). Their outstanding feature is the introduction of technological assumptions that eliminate those ‘complicated patterns of price-movement with several ups and downs’ (Sraffa, 1960, p. 37) on which the very possibility of reswitching is founded.
As shown above, the possibility of reswitching in the comparison between alternative states of the economy is associated with differences in the proportions between labour and intermediate inputs for any given pair of production techniques.
This implies that reswitching can be observed only if the economic system is represented in such a way as to bring in view the ‘ups and downs’ of relative prices. This property was implicitly recognized by John Hicks (1973), when he noted that the possibility of reswitching arises when techniques are ‘no longer capable of being distinguished by a single parameter’, so that any switch along the technological frontier ‘will be a matter of balance between advantages and disadvantages, a balance which itself is affected by prices’ (Hicks, 1973, pp. 44–5). This consideration is at the basis of Hicks's ‘simple profile’, in which any given technique is described by a single parameter (the ratio of construction labour to utilization labour) and reswitching is excluded (Hicks, 1973, pp. 44–6). Joseph Stiglitz took a different view when he distinguished between reswitching as a possibility relative to the comparison among steady states, and ‘recurrence of techniques’ as an outcome for an economy ‘on its optimal development trajectory’ (Stiglitz, 1973, p. 138). In particular, Stiglitz noted that ‘recurrence of techniques may occur in technologies which do not allow reswitching’, and that ‘in technologies in which there is reswitching there may be no recurrences’ (Stiglitz, 1973, p. 139). John Wright (1975, p. 22) examined a related issue showing that reswitching can be avoided if one assumes an appropriate ‘rate of fall of discount rate through time’. A few years later, Edwin Burmeister and Peter Hammond (1977) explored a related issue, and suggested that reswitching can be excluded as soon as we allow economies to ‘jump’ over intermediate states (techniques) along a given optimal adjustment path. Recent literature has examined the likelihood of reswitching from the computational or the empirical point of view. In particular, Stefano Zambelli (2004) has shown that a discrete production model is significantly likely to (computationally) generate a reswitching economy, whereas Zonghie Han and Bertram Schefold (2006) have found the empirical likelihood of reswitching to be significant but not very high. Another strand of literature investigated the relationship between the possibility of reswitching and the stability of optimal paths. In this connection, John Barkley Rosser further explored the problem set-up examined in Burmeister and Hammond (1977) and noted the existence of a trade off between the observability of reswitching and the smoothness of optimal adjustment paths (Rosser, 1983; 2000). More recently, Michael Mandler (2005) and Bertram Schefold (2005) have discussed alternative conditions under which reswitching may or may not be associated with unstable economic dynamics.

Synthesis and appraisal

The capital controversy of the 1960s has conclusively shown that the logical possibility of reswitching is of a general nature. Disagreement about the implications of reswitching for economic theory as a whole does not conceal the fact that a crucial discovery in the theory of technical choice was made. In particular it was shown that choice of technique is related to income distribution in a much more complex way than it was once thought to be, and that the rate of interest (or the rate of profit) cannot provide an unambiguous ranking of different technical alternatives as the distribution of income is varied.
The discussion of reswitching called attention to a paradox that had long been overlooked. This is that rational choice, in its classical formulation, presupposes not only agents capable to rank alternatives in a consistent way, but also objective states of the world making such a consistent ranking feasible (see Urmson, 1950, pp. 154–9; Scazzieri, 1982). The reswitching debate has shown that a ‘granular’ representation of production techniques leads to a complex pattern of interaction such that any given technique may be associated with two or more different positions on the profitability ranking of techniques (see above). This discovery was made possible by the consideration of price movements in a capital-using economy (see above). Its most immediate implication has been to cast doubt upon the representation of capital structure in terms of simple aggregate parables. However, reswitching also called attention to another, perhaps more fundamental, feature of technical choice. This is the dual nature of the grading procedure associated with choice. For grading situations express not only the agent's ability to rank states of the world in a consistent way, but also the possibility to rank those states in terms of ‘objective’ characteristics independent of the agent's preferences and choices. The reswitching debate has proved that the latter prerequisite may be a will-o’-the-wisp as soon as we consider the complex interactions that take place in a production economy.

See Also

- capital theory
- capital theory (paradoxes)
- preference reversals
- reverse capital deepening
- technical change

Bibliography

Bharadwaj, K. 1970. On the maximum number of switches between two production systems. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 106, 409–29.

Bruno, M., Burmeister, E. and Sheshinski, E. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80, 526–53.

Burmeister, E. and Hammond, P. 1977. Maximin paths of heterogeneous capital accumulation and the instability of paradoxical steady states. *Econometrica* 45, 853–70.

Champernowne, D. 1953. The production function and the theory of capital: a comment. *Review of Economic Studies* 21, 112–35.

Garegnani, P. 1966. Switching of technique. *Quarterly Journal of Economics* 80, 554–67.

Han, Z. and Schefold, B. 2006. An empirical investigation of paradoxes: reswitching and reverse capital deepening in capital theory. *Cambridge Journal of Economics* 30, 737–65.

Hicks, J. 1965. *Capital and Growth*. Oxford: Clarendon Press.

Hicks, J. 1973. *Capital and Time: A Neo-Austrian Theory*. Oxford: Clarendon Press.

Levhari, D. 1966. A nonsubstitution theorem and switching of techniques. *Quarterly Journal of Economics* 79, 98–105.

Levhari, D. and Samuelson, P. 1966. The nonswitching theorem is false. *Quarterly Journal of Economics* 80, 518–19.

Mandler, M.A. 2005. Well-behaved production economies. *Metroeconomica* 56, 477–94.

Morishima, M. 1964. *Equilibrium, Stability and Growth: A Multi-Sectoral Analysis*. Oxford: Clarendon Press.

Morishima, M. 1966. Refutation of the nonswitching theorem. *Quarterly Journal of Economics* 80, 520–25.

Pasinetti, L. 1966. Changes in the rate of profit and switches of techniques. *Quarterly Journal of Economics* 80, 503–17.

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Rosser, J.B., Jr. 1983. Reswitching as a cusp catastrophe. *Journal of Economic Theory* 31, 182–93.

Rosser, J.B., Jr. 2000. *From Catastrophe to Chaos: A General Theory of Economic Discontinuities*. 2nd edn, vol. 1: *Mathematics, Microeconomics, Macroeconomics, and Finance*. Boston, Dordrecht and London: Kluwer Academic Publishers.

Samuelson, P.A. 1966. A summing up. *Quarterly Journal of Economics* 80, 568–83.

Scazzieri, R. 1982. Scale and efficiency in models of production. In *Advances in Economic Theory*, ed. M. Baranzini. Oxford: Basil Blackwell.

Schefold, B. 2005. Reswitching as a case of instability of intertemporal equilibria. *Metroeconomica* 56, 438–76.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Starrett, D. 1969. Switching and reswitching in a general production model. *Quarterly Journal of Economics* 80, 673–87.

Stiglitz, J.E. 1973. Recurrence of techniques in a dynamic economy. In *Models of Economic Growth*, ed. J.A. Mirrlees and N.H. Stern. London and Basingstoke: Macmillan.

Urmson, J.O. 1950. On grading. *Mind* 59, 145–69.

Wright, J.F. 1975. The dynamics of reswitching. *Oxford Economic Papers*, n.s. 27, 21–46.

Zambelli, S. 2004. The 40% neoclassical aggregate theory of production. *Cambridge Journal of Economics* 28, 99–120.

How to cite this article

Scazzieri, Roberto. "reswitching of technique." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000126> doi:10.1057/9780230226203.1430

The New Palgrave Dictionary of Economics Online

retirement

Michael Hurd
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article uses a simple life-cycle economic model of retirement to characterize the optimal retirement age and the effects of the wage rate, wealth, and the time horizon on that age. The model is then extended to include pensions, both public and private, which can produce non-convexities in the lifetime budget constraint. The model is further extended to include health effects on retirement, uncertainty and joint retirement (the coordination of retirement dates by husband and wife). The chapter concludes with a discussion of retirement in the context of behavioural economics.

Keywords

age discrimination; annuities; bequest motive; collective models of household; dynamic programming; health care expenditure; health insurance; labour supply; leisure; life expectancy; pensions; precautionary savings; probability distributions; retirement; retirement hazard; risk aversion; Social Security in the United States; subjective probability; uncertainty

Article

The common-sense definition of retirement is leaving employment of a substantial nature by a worker in his or her fifties, sixties or older with no intention of returning to work. However, this definition has no empirical counterpart because we do not observe intentions in the data. Rather, empirical work typically measures retirement in one of two ways. First, a worker is said to retire when he or she leaves the labour force in his or her fifties or older for a ‘considerable’ period of time. The ‘considerable’ period may be limited by the length of the observation period in panel data, but it is meant to distinguish retirement from normal job change by workers in their fifties or sixties. The second definition is an affirmation by the worker that he is retired. This definition aims to address right-censoring in panel data by using the individual's own assessment of retirement status. Because many workers state that they are retired after they have left a career job yet continue to work, this definition often adds the requirement of departure from the labour force. Which definition should be preferred will depend on the empirical analysis and the objective of the research. For some research questions, the definition can make a substantial difference, for example in the study of ‘unretirement’. In this article I think of retirement as the transition from being in the labour force to not being in the labour force by people in their fifties or older.

Historical trends in labour force participation

In 1957 the labour force participation rate of men aged 60–64 in the United States was about 83 per cent; by 1987 it had fallen to 55 per cent, and since then has risen to about 58 per cent. The participation rate of women aged 60–64 rose over this time period because of the historical increase in the participation rate of women of younger ages: an increasing rate of retirement by older women was offset by an increasing number of women reaching age 60 and still in the labour force. Although the levels and rates of decline are somewhat different, participation rates of older men fell sharply in nine European countries and Canada. What caused these very large declines? In the United States and in many European countries the generosity of the public pension system increased sharply in the late 1960s and 1970s. For example, a good measure of the generosity of the system in the United States is the monthly Social Security benefit for men were they to retire at age 65. The average of those Social Security benefits was $307 in 1957 and $649 in 1987 (both in 1987 dollars),

for an annual growth rate of 2.5 per cent. Since 1987 the real growth rate has been just 1.1 per cent and that growth has been due to wage growth, not to changes in the programme rules which have been stable. The coincidence of the decline in labour force participation with the increase in Social Security benefits suggests that Social Security was at least partly responsible for the decline, but there were changes in other determinants of labour supply as well. The private pension system expanded, and real household income increased both because of a rise in earnings and an increase in dual earner households. One objective of research on retirement has been to quantify the contributions of these and other sources to the decline in labour force participation, to predict the future course of labour force participation of the older population, and to understand the response to policy change such as alterations in the structure and generosity of Social Security.

The leading edge of the baby-boom generation will begin to retire in substantial numbers in about 2008, leading to a worsening of the financial health of the Social Security and Medicare systems in the United States. For example, the ratio of the population 65 or over to the working age population (ages 20–64) is a commonly used measure of demographic aging. In 2000 this ratio was 0.21; it is forecast to increase to 0.36 by 2030, an increase of 72 per cent. The retirement of the baby-boom generation will affect the Social Security and Medicare trust funds, requiring adjustments to those programmes. What will be the effect of those changes on retirement? In particular could policy delay retirement without unduly harming workers while improving fiscal balance? To make a good assessment of the effects of policy requires a model of retirement behaviour.

Data

Since the 1970s the most important advance in our ability to study retirement behaviour has been the development of the Health and Retirement Study (HRS). The HRS is a longitudinal data collection on about 20,000 people aged 51 or over in the United States. The HRS was fielded in 1992 with the express purpose of providing data with which to study retirement and health, and their interactions. As such it contains data on all the relevant economic variables that affect retirement, many health variables and many other non-economic variables that have additional effects. The HRS is a biennial longitudinal survey, and as of 2006 it had fielded eight waves. The original cohort was initially 51–61, so that by wave 8 it was 65–75 and had mostly retired. New cohorts aged 51–61 were added in 1998 and in 2004, and they were re-interviewed in successive waves. Based on the success of the HRS, the English Longitudinal Study of Ageing was modelled on HRS and was fielded in England in 2002. It is also a biennial panel. The Survey of Health, Ageing and Retirement in Europe was fielded in 2004 in 11 European countries, and a second wave with an expanded roster of countries followed in 2006. It is modelled on the HRS and ELSA with the aim of providing data that will permit international comparative studies.

Economic models of retirement

Retirement is an aspect of labour supply, and so the same general framework applies. However, in a number of ways it is easier to study retirement than hours worked: most of the retirement incentives are well measured; typically (although not always) retirement can be freely chosen whereas the choice of hours may be constrained by the demands of employers. As a consequence the response of retirement to incentives is substantial whereas the response of hours to the wage rate, at least among males, is small. Although some models of retirement are very complex, many of the ideas can be illustrated with a simplified version of a retirement model which, nonetheless, incorporates most of the important aspects of economic model of retirement.

Retirement must be placed in a life-cycle context because the gain from additional work is an addition to lifetime economic resources, and its value depends on life expectancy. Consider a worker who will live another N years and who is contemplating whether to retire. Should he work another year he would lose a year of leisure which has utility of U and which initially I assume is constant no matter what the age of the worker. He would gain a year's income which he could add to his stock of wealth. The increase in utility from the income is $V' \times wage$: the marginal utility of wealth multiplied by the annual wage. To maximize utility the worker should not work when $U > V' \times wage$. Under the universal assumption that the marginal utility of consumption declines in consumption, V' will be smaller at older ages with wealth held constant: at older ages the fixed amount of wealth would have to be consumed over fewer years so that per period consumption would be greater than at younger ages. Greater consumption would cause the marginal utility of consumption to be lower and therefore the marginal utility of wealth to be lower. At some age V' declines enough that $V' \times wage < U$, and at that age the worker would leave the labour force.

In a complete life-cycle model, consumption and, therefore, saving would be chosen by the worker as well as retirement. Yet we would like to think of a ‘wealth’ effect on retirement. Variation in wealth across workers and the accompanying variation in retirement ages can be generated by variation in wages to which the worker reacts both in the choice of consumption and the retirement age. In this example, wealth is endogenous; but we might think of some variation in wealth that is exogenous to the model. Examples would be variation in initial wealth or through inheritances, variation in rates of return on assets or variation in required expenditures during the working life such as the number of children. Having in mind some exogenous variation in wealth across individuals, I will speak of a ‘wealth effect’ on retirement, but it should be understood that its estimation is difficult because it is endogenous in a complete life-cycle model of retirement behaviour.

These ideas are illustrated in a model of retirement choice. In this model the worker's problem is to choose the retirement age R and consumption level c to maximize lifetime utility

$$\max_{R,c}\left\{\int_0^R (u(c)+L)\,dt+\int_R^N (u(c)+L+U)\,dt\right\}$$

where *c* is consumption which is assumed to be constant. (In this model with fixed lifespan, consumption will be constant if the interest rate and the subjective time rate of discount are the same.) *L* is baseline utility from leisure and *U* is the additional utility from leisure that someone gets when retired. The lifetime budget constraint is *Nc*=*Rw* where *w* is the fixed wage. A corner solution is possible when someone places little value on leisure: he will work his entire life. But in the more usual case of retirement before age *N*, the solution satisfies the lifetime budget constraint and the first-order condition

$$u'(c) = U / w$$

where u' is marginal utility. Then some manipulation will show that $dR / dU < 0$, and $dc / dU < 0$: an increased value of leisure in retirement will reduce the retirement age and the budget constraint will require a reduction in consumption. Also, $dR / dN > 0$: increases in life expectancy will increase the retirement age. Because those in good health have greater life expectancy, healthy people will work longer, independent of any health effect on productivity or on the disutility of work.

An increase in the wage will increase consumption: $dc / dw > 0$. The effect of *w* on *R* is indeterminate because of the income and substitution effects whose relative magnitudes depend on the utility function. For example, if utility is constant relative risk aversion so that $u'(c) = c^{-\gamma}$, then $dR / dw > 0$ if $\gamma < 1$ and $dR / dw < 0$ if $\gamma > 1$.

In the context of this model and other models that allow consumption to be chosen, wealth will be an object of choice. We can observe co-variation in wealth and *R* across individuals due to variation in *w*, but that co-variation will not show how *R* would change were we to add additional wealth to someone's wealth holdings. In this model such an addition would reduce *R* whereas the observed variation in assets at retirement associated with variation in *R* could either be positive or negative depending on the details of the utility function.

Because u' is constant in age, the model predicts that once retired, no one will ‘unretire’.

The retirement hazard

A common object of study in retirement research is the retirement hazard: the probability of retirement at age *t* given working at *t*. The retirement hazard can be found from the simple retirement model by considering only the part of the population still working at age *t*. Among those workers find those who will chose $R = t + 1$; the ratio of the number of those workers to the number of workers at *t* is the retirement hazard.

Estimation of a hazard model requires panel data where the hazard would be expressed in discrete time: the probability of retirement between *t* and *t*+1 conditional on working at *t*. Retirement hazard models are a rather natural way to think about the retirement process particularly in the context of time-varying covariates such as the wage rate or health – just as in the simple model, an increase in wealth will increase the retirement hazard because of the reduction in V'. The model does not make a prediction about the variation in the retirement hazard with age, which will depend on the rate at which V' declines with wealth.

Other predictions depend on whether we are thinking of long-run comparisons across individuals, or short-run reactions by an individual to a change in the environment. For example, if we compare the retirement behaviour of two individuals, one who has a high wage rate and one who has a low wage rate, we cannot predict who will retire earlier because their saving rates would have been different and so their wealth levels would be different: in the comparison of *U* with $u'w$, u' and *w* move in opposite directions.

A good deal of the work on retirement comes from extensions of this simple model to take account of complexities in the budget set, changes in *U* with age and uncertainty about the future. A leading example is the study of the effects of private pensions on retirement. Private pensions are either defined contribution (DC) pensions or defined benefit (DB) pensions. In a DC plan, the employer and/or the employee puts money as specified by the plan into an investment account usually at each pay period. The amount is a small fraction of pay (say, six per cent) which implicitly increases the wage by a small per cent. The account grows at the rate determined by the portfolio held in the account. At retirement the funds are available to the retired worker typically to spend as he wishes. Thus the plan is defined by the contribution rules (hence, DC). What the worker actually receives will depend on the performance of the portfolio.

In a DB plan the worker will receive a pension or annuity at retirement which is based on the years of service with the employer, on the age at retirement and on a measure of earnings in the last few years of employment. Thus the pension plan is defined by the benefit that a worker will receive on retirement. Most DB plans have the curious feature that the benefit

will depend on the age of retirement in a highly nonlinear way. If PV_a is the expected present value of lifetime pension benefits (pension wealth) conditional on retiring at *a*, $PV_{a+1} - PV_a$ is the addition to pension wealth (additional compensation) from working from *a* to *a*+1. DB plans often have a critical age, say *A*, at which a full or unreduced pension benefit is paid to a worker who retires at that age. Workers who retire before *A* may have their pension benefit reduced substantially. Then the apparent compensation from working from *A*–1 to *A* is the wage plus $PV_A - PV_{A-1}$. It is not hard to find examples where the total compensation is more than twice the wage. Said differently, the pension is reduced sufficiently for early retirement so that the gain in pension wealth from a year's work exceeds the wage. Furthermore, often pensions are not adjusted upward if a worker retires past *A* even though for a given pension level the expected present value declines with age. For example, for a single male aged 60 under the assumption that the pension is not indexed and that the nominal interest rate is five per cent, the pension should increase by about eight per cent per year of delayed retirement after age 60 to keep pension wealth constant. If the pension is not adjusted upward at all, the implicit wealth loss from delaying retirement for a year would be about eight per cent of the expected present value of the pension. Assuming the pension replaces 50 per cent of the pre-retirement wage and assessing pension wealth at 12.2 times annual pension income (which is PV_A in this example) means that the loss in pension wealth from delaying retirement is about 50 per cent of the wage. Said differently, the worker would be working for just half of the apparent wage.

The large gain before age 60 in pension wealth creates a large gain in compensation for working from 59 to 60, and the large loss reduces compensation substantially should the worker not retire at 60. These changes in DB pension wealth modify the money wage to produce net compensation. Net compensation for working from age 59 to 60 would be large and net compensation for working from 60 to 61 would be small. It is likely that a worker aged 59 would calculate that $U<V' \times wage$ because *wage*, which is understood to be the net wage, would be large. A worker aged 60 would calculate $U>V' \times wage$ because the net wage would be small. Thus we would observe many retirements at age 60. More generally spikes in compensation induced by DB plans cause correspondingly large spikes in the retirement hazard at critical ages such as *A*. An important part of research on retirement is to obtain data on the details of DB plans so as to relate retirement to these spikes.

DC plans matter, but mainly as an addition to wealth. Typically DC plans do not have special ages at which the implicit compensation is very large or small: rather they add a small percentage to the implicit wage. (However, DC plans can have early withdrawal penalties which may affect retirement among those who have no private savings.)

This simple model of retirement has a number of advantages: estimation would show the effect of changing the net wage or changing wealth on the retirement hazard, and the estimations follow in a straightforward way from what is directly observable in data. It is clear what variation in the data produces the results. The model has considerable flexibility: the retirement hazards can be age specific so that the wage and wealth effects are different for each age. Because the estimation only requires the net wage and wealth at *t* and the retirement outcome at *t*+1, just two waves of panel data are needed. Data far out of sample are not needed: for example, to study the retirement hazard of 59 year-olds, one does not need data on what their wage would be should they work until, say, 65.

However, the simple model has a number of disadvantages. Sometimes DB pensions can induce non-convexities in the lifetime budget constraint as shown in Figure 1. In that figure the vertical axis is lifetime earnings on an arbitrary scale, and the horizontal axis is age at retirement, inverted to show increasing years of retirement. The lifetime budget constraint has a slope equal to the net wage. In our example the large implicit net wage from working while 59 causes the slope of the budget constraint to steepen, and the small implicit net wage from working while 60 causes it to flatten. We would expect that normal shaped indifference curves would cause large numbers to retire at age 60, and, indeed, this is what is observed in data when we have good information about DB plans. The simple model would replicate this clustering at 60. But the simple model would not replicate the prediction that very few would retire at 58: the apparent gain from working while 57 is about the same as the gain from working at 55, 56 and 58, so the simple model would predict that about the same number would retire at each of those ages. But workers can foresee that if they work until they are 59 they will have the option of working from age 59 to age 60, resulting in considerable financial gain. Of course the reason the simple model would not replicate the data is that it makes only a two period comparison: retirement at *t* compared with retirement at *t*+1. It does not make global utility comparisons.

Figure 1

Lifetime earnings and stylized indifference curve

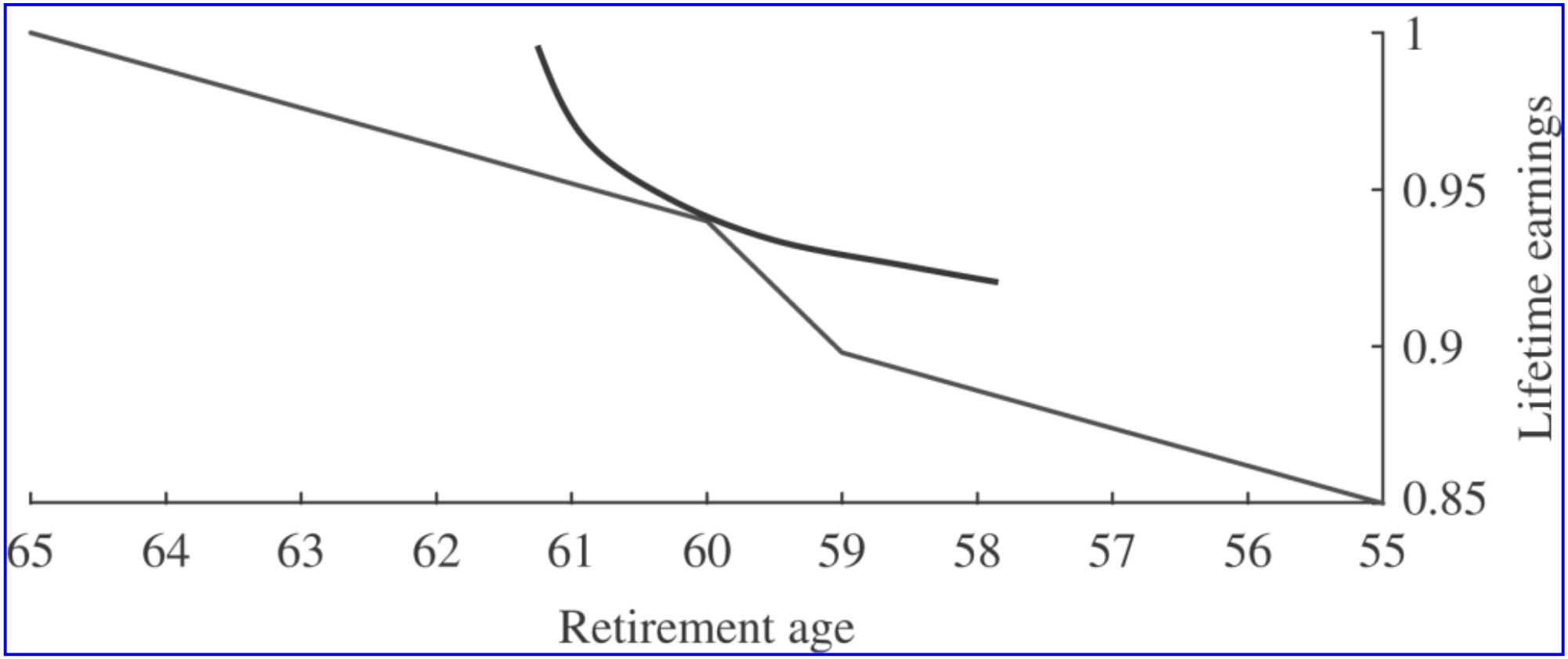

The simple model does not take account of uncertainty. For example, if a possible decline into bad health will require considerable health care expenditures at some future date, a worker may consider retiring later to build up precautionary saving. In this simple model any such tendency would show up in other estimated parameters.

A second type of model is designed to handle non-convexities in the budget set. It is the option value model. In a simplified form it specifies a utility function in which utility depends on years of leisure and on lifetime earnings (Stock and Wise, 1990). A worker will chose the retirement that maximizes utility which is shown at age 60 in the figure. A worker aged 58 would observe that the gain from working another year would be relatively modest, but that the gain from working two more years would be substantial. He would work another year so as to have the option of working the year after that.

The main advantage of the option value model over the simple model is that it can account for non-convexities in the budget set. If properly specified and estimated, it can simulate a greater range of policy options than the simple model. For example, an expansion of the budget set at age 63 by the introduction of a work bonus could only locally affect predicted retirement in the simple model: a worker would have to remain in employment until age 63 in the absence of the alteration. But in the option value model workers who had been contemplating retirement at 60, 61 or 62 could be affected.

A disadvantage of the option value model is that it is dependent on the specification of the utility function. Also it requires the construction of the budget set even at ages where the worker is not observed to work. In the extreme, it requires the construction of the budget set for all future ages. For example, if a worker continues to work at age 57, it may be that he has strong tastes for work or it may be that he has a DB plan with a large incentive to work until age 60. To study his retirement behaviour at age 57 we have to construct the budget set that he perceives at age 57.

In the same way as the simple model, the option value model does not account for uncertainty. This creates some tension because the model assumes that at age t the worker uses information about the environment at t and has expectations about what the environment will be at t+1, at t+2, at t+3 and so forth. Based on this information he may decide to retire at, say, t+4. At t+1 he will use information about the environment at t+1 which will usually be different from the information that was used at t about the environment at t+1. This new information along with new expectations about the future environment may cause an alteration in the intended retirement age. Yet the model does not allow the decision at t to be influenced by the knowledge that new information will be arriving and that the (tentative) decision could be changed.

Social Security and retirement

Social Security, the public pension system in the United States, is a DB pension but it differs from private (employer provided) pensions in at least two ways. First, it is almost universal so that its empirical effects on retirement are difficult to study due to the lack of programme variation across individuals. Second, at critical retirement ages Social Security is approximately actuarially fair; that is, $PV_{a+1} - PV_a \approx 0$ for most workers so that it does not generate the strong retirement incentives of private DB plans. Nonetheless, it is clear that Social Security has an important influence on retirement. First, the retirement hazard is much greater at 62 than at 61 or at 63, as illustrated in Figure 2. Age 62 is the age at which a worker can first claim Social Security benefits, and there is no other explanation for the elevated hazard. Until year 2003, 65 was the normal retirement age under Social Security, and, in addition, the age of entitlement to Medicare, the health care insurance plan for the elderly. Second, this pattern is found in international comparisons where there is programme variation that can help identify programme effects (Gruber and Wise, 1999).

Figure 2
Stylized labour force participation and retirement hazard rates

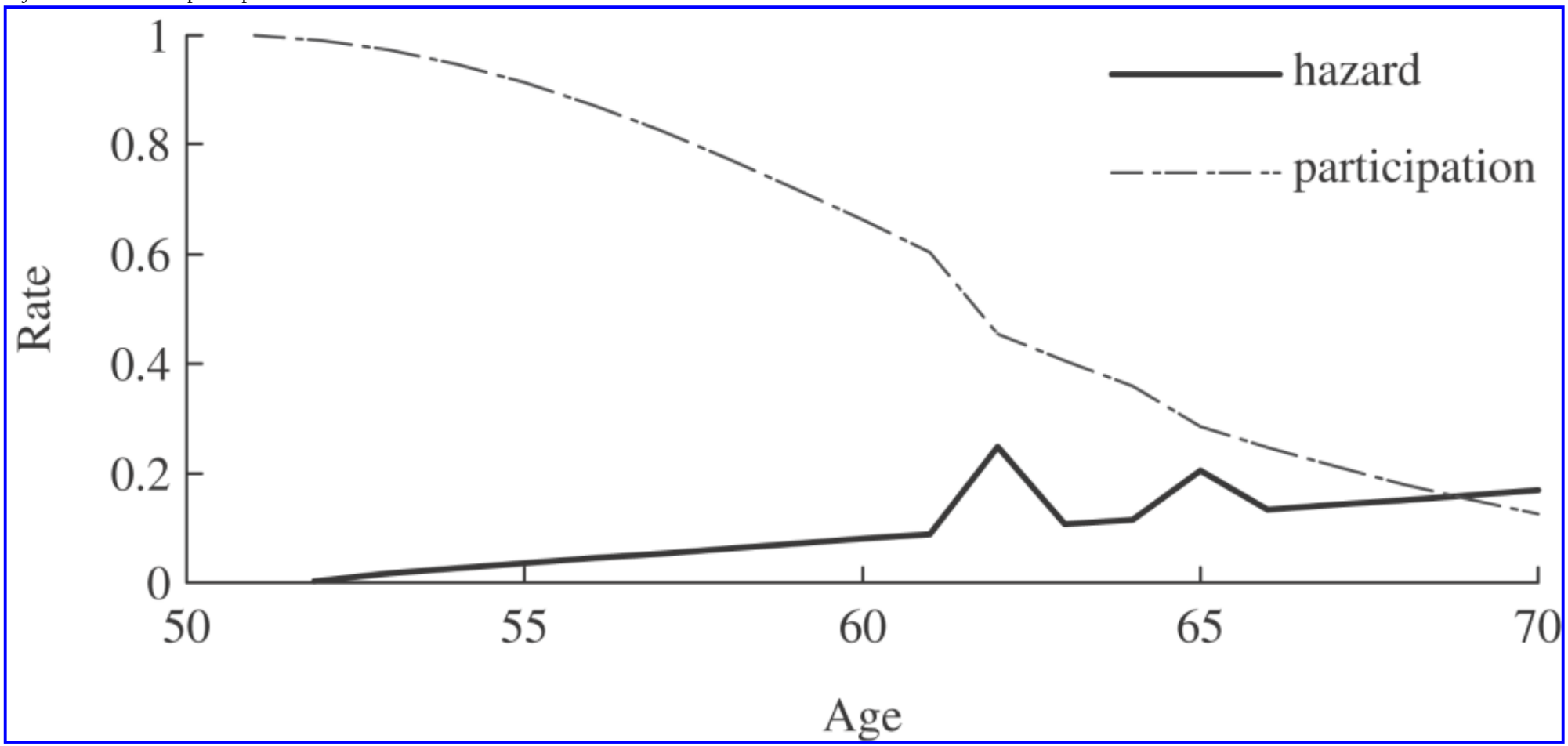

Despite the empirical evidence of the effect of Social Security on retirement, its influence is difficult to explain in an economic model: why should a worker retire at 62 and claim Social Security benefits rather than at some other age, when there are no economic gains from doing so? One possible explanation is based on a liquidity constraint: low-wage workers have been forced by the Social Security system to save more than they would desire and so they do not engage in any private saving. Thus they reach their early sixties with greater-than-desired retirement resources but access to those resources is conditional on retirement. As a consequence, they retire at 62 when they are first able to access them. A difficulty with this explanation is that many workers who have wealth (demonstrating that they were not forced to over-save by the Social Security system) also retire at 62. A second possible explanation concerns the rate of return (about three per cent real) used in the calculation of $PV_{a+1} - PV_a$. If some individuals believe they can obtain a higher rate on their investments, they can increase their lifetime resources by taking Social Security payments and investing them at a higher rate of return than the implicit rate associated with delayed

claiming. It is difficult to evaluate this explanation because we do not know the rates of return people expect. A third explanation is that most people believe their life expectancies to be less than average; in that case they are better off taking Social Security benefits early because they may die before they have received substantial benefits. While individual variation in subjective survival may explain the desire of some to retire and claim Social Security benefits at 62, the average subjective survival as measured by subjective survival in the HRS is close to life-table survival rates. Thus, these factors may explain some (small) part of the excessive retirements at age 62, but they are inadequate for explaining the major part of the excess.

We expect that greater wealth will lead to earlier retirement, and, indeed a contributory factor to the decline in labour force participation in the older population is probably increases in wealth. However, a wealth effect is difficult to show empirically, for several reasons. Wealth is measured with substantial error, and this tends to reduce any estimated effects. Taste variation for retirement can mean that observed wealth is not causative for retirement but rather the result of a desired retirement age. For example, those who place little value on leisure will want to retire late in life and so will save at a low rate. Thus when they reach normal retirement age they will be observed to have low wealth and not to retire. But the delay in retirement is not the result of the lack of wealth: rather the lack of wealth is the result of wanting to retire late. In this model it is necessary to find an instrumental variable to correct for the endogeneity and observation error of wealth. As always this is difficult.

The wage rate measures the price of leisure. In the simple model which says that retirement will occur when $U>V' \times wage$ an increase in the wage would lead to delayed retirement. It is, of course, necessary to control for wealth in an estimation aimed at finding a wage effect: those with high wages in the past will have accumulated more wealth, reducing V', the marginal utility of wealth. If we do not account for wealth, we may observe little relationship between the wage and retirement in data.

Health

The HRS as well as the international data gathering efforts collected many non-economic variables that are likely to influence retirement. The leading class of additional variables measures health. In the evaluation of whether $U>V' \times wage$, U is understood as the utility of leisure relative to the utility of working. It is likely that worsening health increases U because it reduces the utility of working more than it reduces the utility from leisure. Thus a first-order effect of poor health is earlier retirement, both in cross-person comparisons (comparing those in poor health with those in good health) and within person comparisons (comparing those suffering a decline in health with those having stable health). If health declines on average with age, the retirement hazard will tend to increase in age. However, health also influences V', the marginal utility of wealth, although the direction of that influence is not solidly established. If the institutional setting exposes individuals to considerable health spending risk, V' will be higher among those in poor health because of the high productivity of (private) spending on health. If individuals are fully insured against health spending risk, V' may be lower among those in worse health: health could prevent individuals from fully enjoying what money can purchase. Of course, in simple cross-person comparisons other influences on retirement vary systematically with health and they must be controlled. For example, those in worse health have less wealth, causing V' to be larger; they have reduced life expectancy, which reduces V'; and they have lower wages. Thus the relationship between health and $V' \times wage$ is ambiguous. However, as an empirical matter, we observe in panel data that those in worse health retire earlier than those in better health.

The response to an unexpected decline in health (a health shock) is easier to understand because some factors that vary across persons are constant. If exposure to health care expenditure risk is relatively small, V' would decline because of reduced life expectancy and (possibly) because of a reduced ability to spend wealth. If, in addition, the health shock caused a decline in the wage, $V' \times wage$ would decline leading to an increased likelihood of retirement. Indeed, empirically health shocks such as a heart attack are associated with elevated retirement hazard rates.

The availability of health insurance on the job and of employer-sponsored retiree health insurance should affect retirement before the age of 65 because it will change both the expected value of out-of-pocket health care costs and the variance in those costs (Blau and Gilleskie, 2001; 2006). For couples the situation is more complex because in retirement one spouse can be covered by the health insurance of the other. This variation in the provision of health spending insurance provides opportunities for the identification of an insurance effect on retirement.

Accounting for uncertainty

The effects of uncertainty are put in sharpest perspective under the assumption that it is costly to return to work once retired. If it is not costly, an individual can simply return to work as new information arrives in the future, buffering the effects of any negative shocks. This means the decision to ‘retire’ has less consequence. It is undoubtedly true that it is costly to return to work once retired, although the magnitude of the cost varies substantially across persons because of differences in specific human capital by occupation.

A worker contemplating retirement should be thinking about uncertainties that he would face should he continue to work and uncertainties should he retire. The first type would include wage growth, the likelihood of job displacement, the likelihood of a health event that would limit work, the evolution of his pension entitlement and other job characteristics. Remaining on the job gives an option to experience these outcomes both positive and negative. Retiring means both forgoing these options and forgoing the option of continuing to work both in the coming year and in subsequent years. The second type of uncertainty, that associated with the state of retirement, includes the rates of return on assets, health

expenditures in the health insurance environment associated with retirement, survival or life expectancy, uncertainty about one's own utility function especially about one's ability to enjoy wealth, and the utility associated with full-time, uninterrupted leisure.

It is obvious that decision-making under these kinds of uncertainty is difficult. For example, economic resources have to last for many years on average; yet a typical survival curve shows that there are significant chances of dying shortly following retirement and significant chances of dying many years after retirement. In 2003 a 65-year-old man had a life expectancy of 16.8 years, or, stated differently, he could expect to die at age 81.8. However, he had an 11 per cent chance of dying before age 70 and an 11 per cent chance of surviving to age 93. To find the optimal or even satisfactory consumption path is difficult: on the one side he would need to guard against running out of resources should he survive to 93; on the other side excessively low consumption is likely to lead to his dying with considerable wealth, which, if he has no bequest motive, is wasted.

The market solution to this problem is annuities. However, the private purchase of annuities is minimal: among 65–69-year-olds just three per cent receive any income from privately purchased annuities; among those aged 70–74, six per cent receive income from annuities. There are a number of possible explanations for the lack of privately purchased annuities. The rate of return is low because of the profit of the sellers. The price of annuities is actuarially unfair to most people because the typical purchaser lives longer than average, which increases the seller's break-even price. Finally, some people may have a bequest motive: complete annuitization would eliminate any bequests.

In my view, none of these explanations is adequate to explain the lack of annuitization. Profits are not unusually high compared with profit margins in other financial products. While annuities are actuarially unfair for many people, they are not unfair for people who expect to be long lived; yet, annuity purchase is low among such people. Even people with a bequest motive should find it advantageous to annuitize partially. A possible explanation that has been little explored concerns the actual insurance that annuities provide. Privately purchased annuities are not indexed, so people will be concerned about the real value of consumption an annuity would be able to finance at advanced old age. Even a fairly moderate level of inflation will reduce substantially the real value of an annuity over 25 years. Also, the appropriate time horizon is 25–30 years: what is the probability the annuity provider will still be in business?

Estimation of the effects of uncertainty on retirement is conducted in the context of a dynamic programming (DP) model (Rust and Phelan, 1997). The model will specify all possible future states of the world and assign utilities to them conditional on economic resources. Then by well-established backward solution methods the algorithms will find the expected utility associated with continued work and associated with retiring, where the expectation is taken with respect to the joint distribution of stochastic elements in the model. Thus, the analyst will supply the probability distribution of the age of death, the probability distributions of rates of return on assets, the probability distribution of health shocks and associated spending, and so forth. The model predicts continued work when the expected utility from work is greater than the expected utility from retiring, and it predicts retirement at the age when the reverse becomes true. For the reasons discussed in the context of the certainty retirement model, expected utility associated with continued work declines with age and expected utility associated with retirement increases with age, so a worker will eventually be predicted to retire. The model is adjusted with respect to parameters and specification until the predicted retirement ages match most closely those observed in the data.

The data requirements of such a DP model are immense: the analyst needs to assign probabilities to all future exogenous outcomes such as mortality, asset returns and so forth, but the probabilities are not observable. The appropriate probabilities are subjective probabilities: those used by the respondent when making the retirement decision under uncertainty. In particular, the probabilities need not be the same as any observable probabilities of the corresponding events. For example, a population life table displays an estimate of the population mortality risk at each age. Even if the population subjective survival probabilities match those in the life table, individuals should have subjective survival probabilities that deviate from the life table because each person has risk factors that will alter the objective survival probabilities. People with above-average health will survive longer than people with below-average health; people with less education die earlier than people with more education. However, the subjective survival probabilities of people with those characteristics need not correspond even to the objective survival probabilities conditional on those observable characteristics. First, people undoubtedly have private information about their true survival probabilities. Second, they may well have biased subjective survival probabilities in the sense that the probabilities on which people base their retirement decisions are not good predictors of their actual survival. In a similar manner, people have subjective probability distributions over rates of return on assets that may not correspond at all to historic market rates of return or to rates of return predicted by any model based on rational expectations. In this situation, there are no objective data from which the analyst can find the probabilities of the stochastic events required by the dynamic programming model.

Because subjective probabilities are so important in the study of intertemporal decision making, including retirement, the HRS asks respondents to state their probabilistic beliefs about important stochastic events such as survival. Although the model requires survival probabilities to each future age, knowing even the subjective survival probability to a single age is a considerable improvement over using life tables: based on a model of the relationship between subjective survival, actual survival and a life table, one can estimate individualized subjective survival curves (Gan, Hurd and McFadden, 2005).

The advantages of DP models are that they incorporate uncertainty in a formal manner and in principle they can provide an estimate of the effects of uncertainty on retirement. For example, they could predict the response to a mean-preserving spread in survival such as a 20 per cent change of dying before age 70 and a 20 per cent change of surviving to age 93. DP models produce estimates of utility function parameters and so they are capable of out-of-sample prediction. For example, they could predict retirement patterns were the normal retirement age under Social Security increased to age 70.

The disadvantages of dynamic programming models of retirement include the data requirements. Because of the complexity and data requirements, dynamic programming models of retirement are able to account for only a limited number of stochastic events. A significant problem is that the data are not subject to validation: thus a model failure could be due to an incorrectly specified model or to invalid data, particularly the probability distributions of the stochastic events. A second disadvantage is that it is difficult to understand what in the

data is causing observed model outcomes because of the complexity of the model.

Joint retirement

Most people are married when they reach retirement age, and, because of increases in the labour force participation of wives, often both spouses work. Among working husbands aged 55–59 in 2004, 74 per cent of their wives also work. On average, husbands are about three years older than their wives, so that, when the husbands reach age 62, their wives will be about 59. A husband may be influenced by his Social Security benefits to retire but it may be disadvantageous for the wife to retire at 59. Nonetheless, we observe in data some coordination of retirement dates; that is, the probability a wife will retire given the retirement of the husband is greater than the unconditional probability that a wife will retire, and similarly for the conditional probability a husband will retire (Blau, 1998; Gustman and Steinmeier, 2000). A way to model this is to assume a household utility function in which the value of leisure of one spouse is increased by the leisure of the other spouse. That is, their leisures are complements. In a reduced form, the retirement of the husband will be influenced by the incentives he faces such as his wage rate and pension provisions, but also by the incentives his wife faces. Notice that these effects are in addition to any operating through the lifetime budget constraint: if their leisures are not complements, we should observe the early retirement of the husband balanced by the late retirement of the wife in compensation for the loss of earnings of the husband.

Joint retirement offers an arena for the study of household decision-making when there may be conflict between husband and wife as in a collective model of household utility. A typical empirical implementation of the collective model studies the demand for various purchased goods; but it is difficult to know which spouse benefits from a particular purchase. For example, suppose we observe that in households where the husband earns substantially more than the wife the household spends relatively little on clothing. Is this evidence that the power allocation in the household is related to relative earnings? The answer would depend on the assumption that wives benefit more from clothing purchases than husbands. In the case of early retirement, however, we observe who the primary beneficiary of the leisure is, providing sharper identification of the collective model versus the unitary model.

We should anticipate that models of joint retirement will become more important and useful. First, an increasing number of wives reach traditional retirement ages while working, so the quantitative importance of joint retirement will increase. Second, the strong influence of DB pension plans on retirement incentives has made the cost of coordinating retirement high when there are conflicts between optimal retirement dates induced by the plans. With the shift to DC plans, these conflicts will become quantitatively less important in the population because DC plans do not have the sharp retirement incentives of DB plans. We should expect an increase in the amount of coordinated retirements.

Behavioural economics and retirement

The discussion of the determination of retirement has assumed rational decision-making in the context of the life-cycle model: individuals and couples are assumed to maximize expected lifetime utility conditional on beliefs about the probabilities of future events. In that set-up we are still very far from testing important aspects of the model because of our limited measures of those beliefs: an apparent failure of the model could be due to its being an incorrect characterization of decision-making but it could also be due to our lack of valid measures of those beliefs. Nonetheless, there is evidence at least in saving behaviour that the forward-looking model does not apply to all people. Some people strongly prefer the present to the future; they lack the ability to process relevant information; they apparently use rules of thumb; they are heavily influenced by defaults; they do not take actions that would result in fairly large financial gains even though the cost of those actions seems to be small. These examples are about saving behaviour and portfolio choice. It is more difficult to find evidence of non-optimizing behaviour in retirement choice, possibly because it is more difficult to understand what the optimal decision is. For example, job characteristics including distaste for work, perceived discrimination, health and how it interacts with job characteristics could all have large influences on retirement; yet they are mostly unobserved.

In view of these difficulties it is worthwhile asking what would constitute evidence against the rational retirement model. One type of evidence would be an empirically important ‘normative’ retirement age: a high rate of retirement at an age that may be economically disadvantageous (or at least not economically advantageous) where that age is determined by social norms or convention. Thus, a substantial number of individuals who retire at that age would have better economic outcomes had they retired at some other age. However, to find this empirically faces considerable difficulty. At the population level in the United States excessive retirements occur at 62 and 65, ages of importance under Social Security and Medicare. To be certain that these retirements are due to convention rather than to rational economic choice, we need a great deal of information about expectations and personal tastes, and we need to have confidence that our models are complete. In my view we are not in a position to assert that individuals are making a mistake when they retire at those ages.

A somewhat different category of evidence concerns economic preparation for retirement. Do we observe large numbers of workers retiring with inadequate resources to finance their consumption in retirement? If the answer is ‘yes’, retirement is suboptimal in the sense that the marginal utility of wealth is too high for the retirement age chosen. While there is controversy in the literature about the empirical facts concerning preparation for retirement, the main assertion in the literature about a lack of maximizing behaviour has been that saving is suboptimal, not that retirement is suboptimal. Abstracting from spending on health shocks, this argument has considerable validity because saving is under the control of the individual whereas the individual has somewhat less control over retirement. For example, unemployment often leads to retirement because of the difficulty for older workers to find re-employment. Or unmeasured job characteristics such as discrimination against older workers may make continued employment uncomfortable.

Future course of research

The economic models of retirement that have been discussed here and estimated in the literature assume that the decision to work at a given wage is made by the worker. Yet, there is a large literature on the desire of employers to shed older workers; furthermore, there is unemployment at older ages, implying that sometimes retirement is not completely voluntary. Although there are laws against age discrimination in employment, it is likely that employers are able to put some pressure on older workers to retire. However, this observation about the demand side of the labour market for older workers has been formed during an era of increasing number of workers in their forties and fifties when firms may, indeed, have felt they had too many older workers. But as the baby-boom generation begins to retire, the attitude of firms towards older workers should change: firms will want to retain them. Thus, an important research question concerns the evolution of the demand side. How will employers accommodate older workers who may not want to work full time or with the intensity of younger workers? Connected to this question is the long-run productivity of older workers. Apparently firms have wanted to shed older workers because their productivity relative to their costs (including the cost of health care) declined with age. With changing technological requirements on the job, will this unfavourable age-related decline in productivity worsen or improve over time?

We have witnessed a long-run improvement in the health of the older population both in terms of life expectancy and in terms of disability. According to economic theory, increased longevity should cause an increase in the retirement age because of the necessity of financing increased years in retirement. A decline in disabilities will allow more workers to remain in the workforce, and better health is likely to lead to greater productivity at any given age. An important research objective is to quantify these effects and to forecast how changes in retirement will affect important policy concerns. For example, in the United States working past age 65 reduces Medicare expenditures because employer-provided health insurance pays for health care before Medicare. Should enough workers remain in the labour force, the financing difficulties with Medicare will be partially solved, requiring less vigorous policy intervention.

Research on the interaction between health and retirement will likely increase in importance. For example, the sanguine scenario of later retirement because of a reduction in the rate of disability depends on continuing improvements in health. Yet there is considerable uncertainty in forecasts of health, particularly because of the high levels of obesity in the working-age population. We do not have a good understanding of how obesity leads to disability.

Although we have made some progress in our understanding of intertemporal decision-making, a great deal remains to be done. We need to understand how people make such decisions, what information they use, what expectations or probability distributions they have and how they form them, how those expectations evolve as they approach retirement, and so forth. These investigations would be helped were we to have methods of estimating at the individual level preference parameters such as risk aversion and the subjective time rate of discount independently of actual choices. A good example is portfolio choice, where we would like to estimate risk aversion. Lacking data on expected rates of return, we have to make assumptions about those expectations so that risk aversion is conditional on those assumptions. If we knew something about risk aversion we could estimate beliefs about expected rates of return.

These objectives will likely lead to a greater use of subjective data combined with objective data. For example, rather than just studying the determinants of actual retirement, we can study the determinants of the subjective probability of retirement at some target age, say, 62. In panel data, this method can control for unobserved heterogeneity in a straightforward way because we can observe the change in the subjective probability of retirement as the environment changes (Chan and Stevens, 2004).

Because of the continued evolution in survey methods and the ongoing data collection by the HRS, we will have greater sample sizes with which to estimate retirement models. We should be able to observe the effects of natural experiments that can help identify our models. For example, the HRS was in the field in 2002 for the beginning of the natural experiment of the increase in the normal retirement age under Social Security. We will be able to find directly any movement in the retirement spike associated with that age. The continued data collection, natural experiments, innovations in survey design, and the greater use of subjective data should lead to considerable progress in modelling and retirement decision and in quantifying the determinants of retirement.

See Also

- labour supply
- pensions
- Social Security in the United States

Financial support from the National Institute on Ageing is gratefully acknowledged. Many thanks to Susann Rohwedder for her valuable advice and suggestions.

Bibliography

Blau, D.M. 1998. Labor force dynamics of older married couples. *Journal of Labor Economics* 16, 595–629.

Blau, D. and Gilleskie, D. 2001. Retiree health insurance and the labor force behavior of older men in the 1990's. *Review of Economics and Statistics* 83, 64–80.

Blau, D. and Gilleskie, D. 2006. Health insurance and retirement of married couples. *Journal of Applied Econometrics* 21, 935–53.

Chan, S. and Stevens, A.H. 2004. Do changes in pension incentives affect retirement? A longitudinal study of subjective retirement expectations. *Journal of Public Economics* 88, 1307–33.

Gan, L., Hurd, M. and McFadden, D. 2005. Individual subjective survival curves. In *Analyses in the Economics of Aging*, ed. D. Wise. Chicago: University of Chicago Press.

Gruber, J. and Wise, D. 1999. *Social Security and Retirement Around the World*. Chicago: University of Chicago Press.

Gustman, A.L. and Steinmeier, T.L. 2000. Retirement in a family context: a structural model for husbands and wives. *Journal of Labor Economics* 18, 503–45.

Rust, J. and Phelan, C. 1997. How Social Security and Medicare affect retirement behavior in a world of incomplete markets. *Econometrica* 65, 781–831.

Stock, J. and Wise, D. 1990. Pensions, the option value of work and retirement. *Econometrica* 58, 1151–80.

How to cite this article

Hurd, Michael. "retirement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_A000225> doi:10.1057/9780230226203.1431

The New Palgrave Dictionary of Economics Online

returns to scale measurement

Susanto Basu
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

If output grows faster than inputs, holding technology constant, the production function exhibits increasing returns to scale. Increasing returns in the aggregate production function may be due to overhead (fixed) costs, diminishing marginal cost, positive spillovers from aggregate activity, the entry of new varieties of inputs or changes in the distribution of inputs across heterogeneous firms. Each channel has significant implications for models of growth, trade and business cycles. Returns to scale are hard to estimate and even difficult to define, since the definition may depend on the degree of aggregation and the time horizon under study.

Keywords

capital utilization; Cobb–Douglas function; cost functions; imperfect competition; markups; measurement error; production functions; productivity growth; research and development; returns to scale; returns to scale measurement; scale economies; shadow pricing; spillovers; sunspot equilibrium; technology; transmission problem

Article

Knowing the degree of returns to scale (RTS) in a firm, or its average in an industry or economy, is important for a variety of economic questions. First, it is important for assessing the plausibility of models of endogenous growth, which typically require at least constant returns to *reproducible* inputs, and thus increasing returns overall. Second, the size of scale economies is an important determinant of the gains from trade. Third, knowing the RTS is important for assessing the plausibility of certain business-cycle models, which often rely on the existence of substantial increasing returns to scale (IRS). Fourth, the RTS is a lower bound on the size of the markup of price over marginal cost, which is a quantity of great interest in industrial organization. Finally, a basic tenet of productive efficiency requires that the value marginal product of an input be equalized across the uses to which that input could be devoted. Knowing the RTS in each use is important for checking that this condition holds. Assume that firms have a production function for gross output:

$$Y = F(\ddot{K}, \ddot{L}, M, T). \quad (1)$$

Firms use capital services $\ddot{K}$, labour services $\ddot{L}$, and intermediate inputs of materials and energy, *M*. *T* indexes ‘technology’, not directly observed, which is defined to include any inputs that affect firm-level production but are not compensated by the firm (including, for example, Marshallian externalities as well as exogenous technical

change). All variables are functions of time.

Assume $\dot{L} = EHN$, where the number of employees, *N*, and hours worked per employee, *H*, are observed, but the effort of each employee, *E*, is unobserved. Capital services are the product of the observed capital stock, *K*, and its unobserved utilization rate, *Z* (for example, the number of shifts the machine is operated): $\dot{K} = ZK$. The capital stock and the number of employees may be quasi-fixed (costly to adjust). The adjustment cost can be modelled explicitly in the production function (see Berndt, 1986).

F is (locally) homogeneous of arbitrary degree γ in the priced inputs. Constant returns implies $\gamma = 1$. RTS equals the sum of output elasticities:

$$\gamma = \frac{F_1\dot{K}}{Y} + \frac{F_2\dot{L}}{Y} + \frac{F_3 M}{Y}, \tag{2}$$

where F_J is the marginal product of input *J*. Assuming firms minimize cost, we can denote the firm's cost function by $C(Y)$. γ also equals the inverse of the elasticity of cost with respect to output:

$$\gamma(Y) = \frac{C(Y)}{YC'(Y)} = \frac{C(Y)/Y}{C'(Y)} = \frac{AC}{MC}, \tag{3}$$

where *AC* equals average cost and *MC* equals marginal cost. IRS may reflect overhead (fixed) costs or decreasing marginal cost; both imply that average cost exceeds marginal cost. If increasing returns take the form of overhead costs, then γ (*Y*) is not a constant structural parameter, but depends on the level of output the firm produces. To make this point more clear, consider a special case of (1):

$$Y = G(K, L, M, T) - \Phi, \tag{4}$$

where Φ is a *flow* (per-period) fixed cost and *G* is homogeneous of degree ρ in *K*, *L* and *M*. In this case, $\gamma = \rho(Y + \Phi)/Y$. Thus, RTS, γ, may strictly exceed ρ. Some papers use empirical estimates of γ to calibrate ρ. Since this procedure is not generally correct, some of the results in these papers (for example, the existence of sunspot equilibria), do not follow from the existence of IRS per se. Indeed, IRS is compatible with increasing marginal costs (ρ <1), as in the standard Chamberlinian model of imperfect competition.

Even if firms are identical, the RTS of the *aggregate* production function (either of an industry or an economy) is not necessarily the γ of every firm; it also depends on the dynamics of firm entry and exit. Suppose that in the long run all changes in aggregate output are accommodated by changes in the number of firms, with firm-level output remaining constant. Then the aggregate production function has constant returns to scale in the long run, but increasing returns when the number of firms is fixed in the short run. (However, if the new firms produce new varieties of goods, then the aggregate function may exhibit a form of increasing returns through a ‘love of variety’ in production, as in Ethier, 1982.)

Firms may charge a price *P* with a markup, μ, where $\mu \equiv P/MC$. RTS is a technical property of the production function, whereas the markup is a behavioural parameter.

However, from (3), the two are linked:

$$\gamma = \frac{C(Y)}{YC'(Y)} = \frac{P}{C'(Y)}\frac{C(Y)}{PY} = \mu(1 - s_{\pi}), \tag{5}$$

where s_{π} is the share of pure economic profit in gross revenue. As long as pure economic profits are small, as most estimates suggest, eq. (5) shows that μ approximately equals γ. Large markups thus require large increasing returns. Since most studies estimate low profit rates, and since $\mu \geq 1$, eq. (5) shows that firm-level RTS must either be approximately constant or increasing. Internal IRS also *requires* that firms charge a markup, to avoid losses.

One can estimate RTS from either the production function or the cost function, using the implications of eq. (2) or eq. (3) (see Berndt, 1986, for an exposition of the cost approach). The two literatures have developed to have different aims. The cost-function literature typically takes second-order approximations to the underlying production function, which allows it to estimate elasticities of substitution between inputs, but pays little attention to the issue that observed factor prices may not be allocative, especially at high frequencies. The production-function literature takes first-order approximations, but devotes more attention to correcting biases from unobserved right-hand side variables (the quantity analogue of unobserved true factor prices). Neither literature has found a good solution for dealing with issues of endogenous regressors (for example, the presence of output in the firm cost function when one allows for non-constant returns to scale).

Taking logs of both sides of (1) and differentiating with respect to time gives:

$$dy = \frac{F_1 K}{Y} d\tilde{k} + \frac{F_2 L}{Y} d\tilde{l} + \frac{F_3 M}{Y} dm + dt. \tag{6}$$

Small letters denote growth rates (so *dy*, for example, equals $\dot{Y}/Y$); the output elasticity with respect to technology is normalized to one.

Cost minimization puts additional structure on (6). (The advantage of the cost minimization framework is that it is unnecessary to specify the potentially very complicated, dynamic profit maximization problem that gives rise to *P* or μ.) Suppose that firms take the price of all *J* inputs, P_J, as given by competitive markets. The first-order conditions for cost-minimization then imply that

$$PF_J = \mu P_J. \tag{7}$$

If firms make pure economic profits, these appear in the data as factor payments (most often to capital, sometimes to labour). In order for (7) to hold, the prices of capital and labour must be defined as the rental price (or shadow rental price) of capital and the competitive wage rate for labour. The relationship still holds if some factors are quasi-fixed (costly to adjust), as long as we define the input price of the quasi-fixed factors as the appropriate *shadow* prices, or implicit rental rates.

Using eqs (5) and (7), we can write each output elasticity as the product of RTS multiplied by total expenditure on each input divided by total *cost* (not revenue). Thus, for example,

$$\frac{F_1 ZK}{Y} = Y\frac{P_K K}{\sum P_J J} \equiv Y c_K.$$
(8)

c_J are the *cost shares* of each type of input, and sum to 1.

Substitute these output elasticities into (6) and use the definition of input services:

$$dy = Y\left[c_K d\tilde{k} + c_L d\tilde{l} + c_M dm\right] + dt = Y[c_K(dk + dz) + c_L(dn + dh + de) + c_M dm] + dt = Y[c_K dk + c_L(dn + dh) + c_M dm] + Y[c_K dz + c_L de] + dt \equiv Ydx + Ydu + dt$$
(9)

Defining *dx* as a share-weighted average of conventional (observed) input growth, and *du* as a weighted average of unobserved variation in capital utilization and effort, we obtain our basic estimating equation for γ, the last line of (9). Note that to create the cost shares c_J one needs to construct an estimate of pure profits, as in Hall (1990). Alternatively, one can assume zero economic profit on average and use the observed revenue shares.

Regarding (9) as an estimating equation, one immediately faces three issues.

First, the econometrician usually does not observe utilization *du* directly. In this case, the regression suffers from measurement error. Unlike classical measurement error, variations in utilization *du* are likely to be (positively) correlated with changes in the measured inputs *dx*, leading to an upward bias in the estimated γ.

Second, should one take the output elasticities as constant (appropriate for a Cobb–Douglas production function or for a first-order log-linear approximation), or time-varying? That is, should one allow γ and the share-weights in (9) to change over time? If the elasticities are not truly constant over time, then treating them as constant may introduce bias.

Third, even if the output elasticities are constant and all inputs are observable, one faces the ‘transmission problem’: The technical change term, *dt*, is likely to be correlated with a firm's input choices, leading to biased OLS estimates of γ. In principle, one can solve this problem by instrumenting the right-hand-side variables, or by using a proxy for *dt*, following Olley and Pakes (1996).

Approaches to controlling for *du* also involve the use of proxies. One method builds on the intuition that firms view all inputs (whether observed by the econometrician or not) identically. For example, a firm should equate the marginal cost of obtaining more services from the observed intensive margin (for example, working current workers longer hours) and from the unobserved intensive margin (working them harder each hour). If the costs of increasing hours and effort are convex, firms will choose to use both margins. Thus, changes in an observed input – for example, hours per worker – provide a measure of unobserved changes in the intensity of work. This suggests a regression of the form:

$$dy = \gamma dx + \kappa dh + dt,$$
(10)

where *dh* is the growth rate of hours per worker. Basu and Fernald (2001) summarize research showing that regression (10) controls for variable effort. In addition, if the cost of varying the workweek of capital takes the form of a shift premium – for example, one needs to pay workers more to work at night – then this regression corrects for variations in utilization of capital as well as labour. (If the cost of varying capital's workweek is ‘wear and tear’ – that is, capital depreciates in use – then the regression is somewhat more complicated, but theory still suggests appropriate proxies.)

In principle, allowing for time-varying factor shares in an estimating equation like (10) is always preferable to having constant shares, since using time-varying shares approximates the true function to a second order. However, attempting to estimate the time-varying shares requires observing (or estimating) the true shadow cost of inputs at each point in time. If observed factor payments at each point in time do not correspond to the factor's true cost each period – for example, if firms smooth wage payments by offering workers insurance through an implicit contract – then treating the observed prices as allocative may introduce larger biases (see also Carlton, 1983, on intermediate goods prices).

Since one is unlikely to observe allocative factor prices period-by-period, one probably should take a first-order approximation and assume constant, not time-varying, elasticities. For estimating the RTS a first-order approximation may suffice, and it will be accurate as long as the true average factor price is the mean of the observed prices over the sample period.

So far, the discussion has concerned the estimation of internal returns to scale. However, a number of interesting models assume the existence of spillovers between competitive firms with internal constant returns, leading to *external* increasing returns in the aggregate production function. The empirical literature searching for such spillovers follows two sharply divergent tracks. The search for high-frequency spillovers is usually atheoretical, and amounts to augmenting disaggregated estimating equations like (10) with measures of aggregate activity. However, since most such exercises do not attempt to control for unobserved changes in utilization (omitting, for example, the $\kappa\, dh$ term in (10)), they are vulnerable to the charge that the putative externalities are actually proxies for unobserved changes in internal inputs. Furthermore, Basu (1995) presents a model where apparent external effects are actually driven by a different economic mechanism, and shows that his model can be distinguished from true technological spillovers by examining gross-output data, as opposed to the commonly used value-added data. Performing the test, apparent externalities are found in value-added but not in gross output, suggesting they are not true spillovers.

However, the search for long-run external effects is based firmly on the economic insight that knowledge creation has built-in increasing returns, since knowledge is non-rival. Thus, there is a long tradition of searching for externalities to R&D, summarized by Griliches (1998). R&D spillovers appear to be a fact, but their exact magnitude is still an issue subject to debate. And there is no consensus at all on whether the magnitude of the spillover is large enough to permit fully endogenous long-run growth.

So far, the discussion has been couched in terms of firm-level output, or aggregation over identical firms. For some applications, one wants to know the RTS for an industry or a sector but allow – plausibly – for the possibility that firms have heterogeneous characteristics, including different γ's. It turns out that, in this realistic scenario, there is not even an unambiguous definition of increasing returns to scale. Basu and Fernald (1997) show that industry output growth equals:

$$d\bar{y} = \bar{\gamma} d\bar{x} + d\bar{u} + R + d\bar{t}. \tag{11}$$

$\bar{\gamma}$ is the average RTS across firms; $d\bar{y}$, $d\bar{x}$ and $d\bar{u}$ are appropriately – weighted averages of firm-level output and input growths; *R* represents various reallocation (or aggregation) effects; and $d\bar{t}$ is an appropriately – weighted average of firm-level technology.

The intuition for '*R*' is that γ need not be the same across firms within an industry (or the economy). Output growth therefore depends on the *distribution* of input growth as well as on its mean: if inputs grow faster in firms where they have above-average marginal products (γ is higher), industry output grows more rapidly as well. Thus, aggregate productivity growth is not just firm-level productivity growth writ large; comparing eqs (9) and (11) shows that there are qualitatively new effects at the aggregate level. Is the RTS of an industry just the average of firm-level RTS, $\bar{\gamma}$, or does it include the aggregation effects, *R*, which are also the result of deviations from constant returns and perfect competition? The answer will depend on the economic question being asked (see Basu and Fernald, 1997, section V), but empirically the magnitudes are often quite different.

See Also

- capital utilization
- cyclical markups

- external economies
- multiple equilibria in macroeconomics
- production functions
- returns to scale
- technical change

Bibliography

Basu, S. 1995. Intermediate goods and business cycles: implications for productivity and welfare. *American Economic Review* 85, 512–31.

Basu, S. and Fernald, J.G. 1997. Returns to scale in U.S. production: estimates and implications. *Journal of Political Economy* 105, 249–83.

Basu, S. and Fernald, J.G. 2001. Why is productivity procyclical? Why do we care? In *New Developments in Productivity Analysis*, ed. C. Hulten, E. Dean and M. Harper. Cambridge, MA: NBER.

Berndt, E. 1986. *The Practice of Econometrics: Classic and Contemporary*. New York: Addison-Wesley.

Carlton, D.W. 1983. Equilibrium fluctuations when price and delivery lags clear the market. *Bell Journal of Economics* 14, 562–72.

Ethier, W.J. 1982. National and international returns to scale in the modern theory of international trade. *American Economic Review* 72, 389–405.

Griliches, Z. 1998. *R&D and Productivity*. Chicago: University of Chicago Press.

Hall, R.E. 1990. Invariance properties of Solow's productivity residual. In *Growth/Productivity/Unemployment: Essays to Celebrate Bob Solow's Birthday*, ed. P. Diamond. Cambridge, MA: MIT Press.

Olley, G.S. and Pakes, A. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64, 1263–97.

How to cite this article

Basu, Susanto. "returns to scale measurement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_I000297> doi:10.1057/9780230226203.1433

The New Palgrave Dictionary of Economics Online

returns to scale

John Eatwell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

constant returns; decreasing returns; dynamic increasing returns; increasing returns; indivisibilities; returns to scale; substitution; technical change

Article

The technique of production of a commodity *y* may be characterized as a function of the required inputs x_i:

$$y = f(x_1, x_2, \ldots x_n)$$

If all inputs are multiplied by a positive scalar, *t*, and the consequent output represented as $t^s y$, then the value of *s* may be said to indicate the magnitude of returns to scale.
If $s=1$, then there are constant returns to scale: any proportionate change in all input results in an equiproportionate change in output. If $s>1$, there are increasing returns to scale. If $s< 1$ (though not less than zero, given the possibility of free disposal) then there are decreasing returns to scale.
These mathematical definitions suggest a symmetry between the three classifications of returns to scale. This appearance of symmetry is entirely spurious.
The original arguments from which is derived the economic rationale underlying the various categories of returns to scale are to be found in the works of the classical economists. Yet there, as Sraffa (1925) pointed out, each category is derived from quite different economic phenomena. Increasing returns derived from the process of accumulation and technological change, associated as they were with the division of labour attendant upon the extension of the market. Decreasing returns were held to derive from the limited availability of land, and were an important component of the theory of income distribution, being the foundation of the theory of rent.

Yet it was from these disparate origins that Marshall (1890) attempted to formulate a unified, symmetric, analysis of returns to scale which would provide the rationale for the construction of the supply curve of a competitive industry, derived in turn from the equilibria of the firms within the industry. Marshall himself recognized the incompatibility of the assumption of competition and presence of increasing returns (1890, Appendix H). Piero Sraffa (1925; 1926) exposed the entire exercise as ill-founded by demonstrating that neither increasing nor decreasing returns to scale are compatible with the assumption of perfect competition in the theory of the firm or of the partial-equilibrium industry supply curve – a result which, although prominently published and debated, has apparently escaped the notice of those who still draw that bogus U-shaped cost curve whilst purporting to analyse the equilibrium of the competitive firm.

The difficulties identified by Sraffa rest upon the *economic* rationales for variable returns to scale.

The idea of constant returns to scale derives essentially from the proposition that a given set of production conditions may be replicated so long as all the requisite inputs may be varied in the same proportion. Indivisibilities in the production process may limit exact replication to particular levels of output. But the concept, though less precise, is not in any way diminished by the presence of indivisibilities, particularly if the optimal scale of operation of a given technique is small relative to the overall level of output.

The presence of decreasing returns *to scale* would suggest that replication is, for some reason, impossible. Yet if all inputs are correctly enumerated and all increased in the same proportion, then, barring indivisibilities, there can be no barrier to replication. Decreasing returns can derive only from a fixed input (or an input which cannot be increased in the same proportion as others) which prevents replication. In other words, there is no such thing as decreasing returns *to scale*. Decreasing returns derives from *substitution*, from the necessity of changing input proportions.

Whilst decreasing returns to scale do not exist, increasing returns to scale are typically based on propositions so general as to defy precise clarification.

There are some examples in which outputs are an increasing function of inputs for purely technical reasons. The capacity of a pipeline, for example, is defined by the area of its cross-section, πr^2 whereas the circumference of that cross-section is equal to $2\pi r$. If it were possible to increase capacity merely by increasing the circumference (if the walls of the pipe did not require strengthening), then a quadrupling of capacity could be achieved simply by doubling the material inputs.

There is one odd symmetry in this ‘technical’ case of increasing returns. Whereas decreasing returns can derive only from substitution and not from scale, increasing returns can derive only from scale, not from substitution! Choice of optimal proportions of inputs (with free disposal and no indivisibilities) will always ensure *at least* constant returns.

Such technical examples are not, however, the examples which typically come to mind in the discussion of increasing returns to scale. More typical are examples of mass production, of production lines, or, today, of production integrated by means of sophisticated information systems. Yet these examples, which are akin to Adam Smith's analysis of increasing returns, are associated more with technological change, and with the possibilities for change inherent in a larger, or more rapidly growing, market, than with a simple increase in the scale of identical inputs. Generalization of the concept to ‘dynamic increasing returns’ (Young, 1928; Kaldor, 1966) increasing returns associated with growth of output further distances the idea of increasing returns from the formal characteristics of scale.

These arguments suggest that the concept of ‘returns to scale’ is not merely a very limited means of

characterizing technology, but it is also a very *limiting* concept. None of the interesting characteristics of the relationship between scale of production and method of production are captured by the idea of returns to scale. Indeed, the only really satisfactory formal characterization of returns to scale is that of constant returns – and this only because replication is formally a precise notion, however empty empirically.

Bibliography

Kaldor, N. 1966. *Causes of the Slow Rate of Economic Growth in the United Kingdom*. Cambridge: Cambridge University Press.

Marshall, A. 1890. *Principles of Economics*. 9th (Variorum) edn, London: Macmillan, 1961.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Methuen, 1961.

Sraffa, P. 1925. Sulla relazioni fra costo e quantità prodotta. *Annali di Economia* 2, 277–328.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36, 535–50.

Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

How to cite this article

Eatwell, John. "returns to scale." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000134> doi:10.1057/9780230226203.1432

The New Palgrave Dictionary of Economics Online

returns to schooling

David Card
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The returns to schooling represent the incremental increase in earnings associated with an increase in schooling. Under assumptions first spelled out by Jacob Mincer in 1958, each additional year of schooling will lead to a percentage gain in earnings that is equal to the interest rate. More recent research has treated the return to schooling as a causal parameter that can vary across people, and by the level of education.

Keywords

ability bias; control functions; degree effects; equalizing differences; human capital earnings function; instrumental variables; Mincer, J.; publication bias; returns to schooling; self-selection bias; treatment effect; wage differentials

Article

1 Introduction

The return to schooling is the internal rate of return on an additional year of schooling: the discount rate at which the present value of the gains associated with the investment equals the costs. The notion of treating education as a capital investment – and calculating the return accordingly – was proposed by Walsh (1935) in an aptly titled article, ‘Capital Concept Applied to Man’. Subsequent contributions (Mincer, 1958; 1974; Becker, 1962; 1964; 1967) have elaborated the theoretical underpinnings of this exercise, while advances in data availability and econometric methods have led to refinements in the empirical procedures used to calculate the return to schooling (see Griliches, 1977; Card, 2001; Harmon, Oosterbeek and Walker, 2003, for surveys).
Following Mincer (1974), the term return to schooling also refers to the coefficient of years of schooling in a linear regression of log earnings on years of schooling and controls for labour market experience.

Under certain simplifying assumptions this coefficient is approximately equal to the internal rate of return to an additional year of schooling (see Section 2.1 below). More generally, however, applied economists use the term return to schooling to denote the causal effect of additional schooling on log earnings, holding constant experience (or in some cases age). In this sense, which I will adopt below, the return to schooling is a structural parameter that may vary with the level of schooling, personal characteristics, and the economic environment. Moreover, the observed *ex post* returns to schooling can differ from the *ex ante* returns that were anticipated when the schooling decision was made (Cunha and Heckman, 2006).

2 Theoretical framework

2.1 The internal rate of return and equalizing differences

The internal rate of return is an accounting concept that can be implemented without reference to a particular theory of wages and schooling. As was recognized by Walsh (1935), however, if there is free entry into different schooling options, and if increases in the supply of workers with a given schooling level reduce relative wages for the group, internal rates of return to different choices will be driven down to a common level.

(Walsh, 1935, p. 284, wrote: ‘Investment in training … tends to be made as long as the returns promise to cover the cost of that training with an ordinary commercial profit. And this of course is the fundamental characteristic of the competitive, equalizing market …’)

Using this ‘equalizing differences’ framework, Mincer (1958) showed that the equilibrium wage differential between two occupations requiring differing amounts of schooling will equal the difference in years of schooling multiplied by the discount rate.

Willis (1986) considers the choice of an optimal schooling level S (measured in units of time) under four assumptions: (1) individuals maximize the discounted present value of earnings using a common interest rate r; (2) earnings are zero while in school, and equal to $f(S)g(t-S)$ at age t (where age is measured in units of time since the completion of compulsory schooling); (3) the duration of work life is independent of S; (4) the only cost of schooling is the opportunity cost of forgone earnings. Under these assumptions, the internal rate of return for a marginal increase in schooling from an initial level S_0 is $f'(S_0)/f(S_0)$ – that is, the proportional earnings differential per year of education between people with schooling S_0 and those with a little more (or less), holding constant work experience. (Under the assumptions specified, the internal rate of return r equates $V(S_0,r)$ and $V(S_0+\varepsilon, r)$, where $V(S,r) = \int_S^{S+n} f(S)g(t-S)e^{-rt}dt = f(S)e^{-rS}\int_0^n g(x)e^{-rx}dx$. Equality implies that $e^{r\varepsilon} = f(S_0+\varepsilon)/f(S_0) \approx 1+\varepsilon f'(S_0)/f(S_0)$ and taking the limit as $\varepsilon \to 0$ gives $r = f'(S_0)/f(S_0)$.)

If people can choose freely between schooling opportunities, in equilibrium log earnings will be a linear function of years of schooling (with slope r), and the internal rate of return for any schooling choice will equal r. Consistent with this insight, one of the most important regularities in labour economics is that a regression of log earnings on years of schooling and controls for experience yields a coefficient that is

comparable to a discount rate for a risky investment – of the order of 5–15 per cent per year. Though the precise magnitude of such a coefficient varies over time and across labour markets, the predictability of the magnitude of the estimated return to schooling is unmatched in any other area of empirical microeconomics.

2.2 An extended model

While a simple equalizing differences framework provides a useful starting point for understanding the relationship between earnings and schooling, it does not explain why different people choose different levels of schooling. In fact, children's education choices are strongly correlated with their parent's schooling and socio-economic status, and with their own test scores in early grades. (See Card, 1999, and Solon, 1999.)

These correlations raise a fundamental question: to what extent do people with more education have other attributes – like ability or privileged family background – that would cause them to earn more even in the absence of extra schooling? In the literature this possibility is known as ability bias. A closely related question is whether people who acquire additional schooling have higher returns than those who do not – a sorting or self-selection bias of the type identified by Roy (1951).

Becker (1967) presented a simple model of earnings and schooling determination that can be used to address these issues. In this model, an individual faces a market opportunity locus $y(S)$ that gives the level of earnings y associated with different schooling choices S, and chooses a level of schooling by equating the marginal benefit of schooling with the marginal cost. Following Card (1995a), it is convenient to assume the individual chooses S to maximize a utility function $U(S, y)=\log y-h(s)$, where h is an increasing convex function. An optimal schooling choice satisfies the first-order condition

$$h'(S) = y'(S) / y(S).$$

Note that, because the objective function is linear in $\log y$, the optimal choice of schooling is independent of factors that generate a parallel shift in the $\log y(S)$ function. Griliches (1977) presented a more general model of preferences with the feature that a uniform upward shift in log earnings for all levels of schooling leads to a lower schooling choice.

Individual heterogeneity in the optimal schooling outcome arises from two sources: differences in the costs of (or tastes for) schooling, represented by heterogeneity in $h(S)$; and differences in the economic benefits of schooling, represented by heterogeneity in the marginal return $y'(S)/y(S)$. A tractable assumption is that both functions are linear in S, with additive heterogeneity components:

$$y'(s) / y(S) = b_i - k_1 s, h'(S) = r_i + k_2 S.$$

Here b_i and r_i are random variables with means $\bar{b}$ and $\bar{r}$ and some joint distribution across individuals (indexed by i), and k_1 and k_2 are non-negative constants. This specification implies that the optimal schooling choice is *linear* in the individual-specific heterogeneity terms:

$$S_i^* = (b_i - r_i) / k,$$

where $k=k_1+k_2$.
The assumed model for the marginal returns to schooling implies that log earnings are generated by a model of the form

$$\log y_i = \alpha_i + b_i S_i - \frac{1}{2} k_1 S_i^2,$$

where α_i is a person-specific constant of integration. This is a generalization of the semi-logarithmic functional form adopted in Mincer (1974) and hundreds of subsequent studies. In particular, individual heterogeneity potentially affects both the *intercept* of the earnings equation (via α_i) and the *slope* of the earnings-schooling relation (via b_i). In general the optimal schooling choice will be positively correlated with b_i, leading to a 'self-selection bias' that arises because people with higher returns to schooling acquire more schooling. If α_i is also positively correlated with S_i (via a positive correlation with b_i or a negative correlation with r_i) the relationship between earnings and schooling will also include an 'ability bias', that is, a bias that arises because people with a higher level of earnings for each level of schooling have characteristics that lead them to acquire more schooling.
A particularly simple version of this model has only two schooling choices (Willis and Rosen, 1979). In this case the model reduces to a discrete choice model for the longer schooling option, and an earnings equation with a random intercept and random coefficient on a dummy representing the longer schooling option. A more general version arises if one relaxes the linearity assumptions for the marginal costs and marginal returns, but maintains additive heterogeneity: that is, $y'(S)/y(S)=b_i+\lambda(S)$, $h'(S)=r_i+\mu(S)$. In this case, Rau-Binder (2006) shows the optimal schooling is $S=\theta^{-1}(b_i - r_i)$, where $\theta(S)=\mu(S) - \lambda(S)$, and log earnings are generated by a model of the form $\log y_i=\alpha_i+\phi(S)+b_i S_i$, where $\phi'(S)=\lambda(S)$.
What does this class of models imply about the return to schooling? For individual i, the marginal return to the last unit of schooling is:

$$\beta_i = b_i - k_1 S_i^* = b_i(1 - k_1 / k) + r_i k_1 / k,$$

which varies across people unless one of two conditions is satisfied: either $b_i = \bar{b}$ for all i and k_1=0 (so each additional unit of schooling has the same proportional effect on earnings for everyone); or $r_i = \bar{r}$ for all i and k_2=0 (so everyone uses the same discount rate and invests in schooling until the return on their last unit of schooling is driven down to $\bar{r}$). Even if one of these conditions is satisfied and β_i is constant across the population, it is not necessarily true that one can obtain an unbiased estimate of the average marginal return to schooling $\bar{\beta} = E[\beta_i]$ from observational data on earnings and schooling. In the first case (homogeneous returns) the implied earnings model is

$$\log y_i = \alpha_i + \bar{b}S_i.$$

Only if α_i and S_i are uncorrelated will an ordinary least squares (OLS) regression yield a consistent estimate of $\bar{b}$. In the second case (homogeneous interest rates) the implied earnings model is:

$$\log y_i = \alpha_i + \bar{r}S_i + \frac{1}{2}k_1 S_i^2.$$

Since people with higher values of b_i invest in more schooling, the implied relationship between earnings and schooling is convex, leading to an upward bias in the OLS estimator relative to the true marginal return to schooling, $\bar{r}$ (Mincer, 1997). Any correlation between α_i and S_i will confound the situation even further.

For the general case where marginal returns vary across the population, Card (1999) shows that an OLS regression of earnings on schooling yields a coefficient b_{ols} that has probability limit

$$\text{plim } b_{ols} = \bar{\beta} + \lambda + \psi\bar{S},$$

where λ=cov[α_i, S_i]/var[S_i] represents an ability bias term and ψ=cov[b_i, S_i]/var[S_i] represents a self-selection or sorting bias term. (This expression assumes that the heterogeneity terms have symmetric distributions – see Card, 1999.) Since people with higher returns at each level of schooling will tend to acquire more schooling, the sorting bias term should be positive, although the magnitude may be small. The sign of the ability bias term is less clear: several studies – including the seminal paper by Willis and

Rosen (1979) – have obtained negative estimates of λ. In any case, observed pay differences between people with different levels of education may imply rates of return that are above or below $\bar{\beta}$, the average marginal return to education in the population.

The Mincer–Willis equalizing differences model is a long-run general equilibrium model in which wage differentials across education groups are determined by a free entry condition on the supply side. Becker's (1967) model, in contrast, is a partial equilibrium model describing the schooling decisions made by different individuals in a given cohort, taking the earnings generating function as given. Once these decisions are made, shifts in the demand and supply for different education groups can lead to realized returns that are higher or lower than were originally anticipated *ex ante*. Moreover, the fraction of a cohort that acquires higher education can affect their *ex post* returns – a general equilibrium effect. In the mid-1970s for example, the college-high school premium in the United States was relatively low, and analysts described an ‘oversupply’ of college-educated labour (Freeman, 1976). Within 15 years, however, the premium bounced back, and it now appears that cohorts born in the 1950s have enjoyed higher returns to education than they expected *ex ante*.

2.3 Dynamic models of schooling

A more realistic alternative to the static Becker (1967) model is one in which young people make a series of decisions about whether to enrol in school (for example, Keane and Wolpin, 1997; 2001). If they do, their education increments by an amount which may depend on effort and ability, and they then become eligible to enter a higher level of schooling the next period. Individuals also choose a level of savings or borrowing which can depend on tuition costs, earnings, family transfers, and access to loans and grants. This class of models sheds light on a number of features that are inconsistent with (or simply ignored by) a static framework. For example, a dynamic model can be used to formally address the question of how students learn about their potential returns to different levels of schooling (Arcidiacono, 2004), and how schooling choices are affected by risk aversion and access to credit markets (Keane and Wolpin, 2001).

A dynamic framework is also helpful for understanding the distribution of observed education choices in the presence of ‘sheepskin’ or ‘degree’ effects that create non-concavities in the earnings–schooling relationship. In the United States, for example, people with three years of college education have about the same earnings as those with only two years of college (Park, 1994). (Likewise, people with three years of high school earn about the same as those with only two years of high school; see Hungerford and Solon, 1987.) From a static modelling perspective it is unclear why anyone would ever plan to leave college after three years. From a dynamic perspective, however, the outcome of three years of college can be explained by noting that the true return to the third year of college is the option value of entering the fourth year (Altonji, 1993). Students begin their third year of college knowing it is a necessary step to graduation, but may receive some information – for example, about their ability to complete the programme – that causes them to re-evaluate the costs and benefits of enrolment and drop out without graduating.

A dynamic perspective suggests that one should calculate the distribution of final education outcomes conditional on starting a specific education programme, and use this distribution, in combination with

the estimated costs and earnings for each outcome, to measure the *ex ante* return to programme entry. (In fact, such a calculation is explicitly built into dynamic optimization models like the ones estimated by Keane and Wolpin, 1997 and Eckstein and Wolpin, 1999.) An interesting case in point is entry to a junior college, which has three main outcome possibilities: early dropout, completion of an Associates (AA) degree, or entry (with two years of college credit) to a four-year college programme. The third node creates an option value to entering an academic programme at junior college that is ignored in simple *ex post* comparisons of earnings between those who are observed holding an AA degree and those with only high school education.

3 Evidence on the returns to schooling

3.1 Mincerian studies

Most of the existing evidence on the returns to schooling is based on Mincer's (1974) ‘human capital earnings function’: an OLS regression of log earnings on years of completed schooling and a polynomial of post-schooling experience (that is, current age minus an estimate of age at the completion of schooling). As noted in Section 2.1, under certain simplifying assumptions the coefficient of schooling can be interpreted as an estimate of the internal rate of return to alternative education choices. Though the empirical validity of these assumptions varies from application to application, Mincer's model has fitted in hundreds of studies of earnings determination around the world.

Several issues arise in the specification of the human capital earnings function (HCEF) that affect the magnitude of the estimated returns to schooling. One is the choice of earnings measure. Since better-educated people tend to work more hours per week and weeks per year than those with less education, the estimated returns to schooling are usually larger for annual earnings than for weekly or hourly earnings (Card, 1999).

Arguably, earnings should also include the cash value of work-related benefits like health insurance and pensions, leading to an additional source of ‘returns’ to schooling. A related issue is the treatment of taxes and transfer income during periods of nonwork, for example, from unemployment insurance and welfare programmes. From the perspective of an individual investor, the return to a given schooling choice presumably depends on the expected net incomes associated with the choice (that is, taking into account expected taxes and transfers). Interestingly, the earnings measures available in conventional surveys for many European countries (for example, France and Spain) are net of social security and income taxes, whereas the earnings measures available for other countries (in particular the United States) exclude taxes. Thus, there is some adjustment for taxes built-in to conventional human capital earnings functions estimated for many European countries, but not for the United States. Finally, schooling may affect longevity or health, leading to another indirect effect on earnings.

A second issue is functional form. Mincer's equalizing differences framework implies that log earnings are related to the opportunity cost of a given schooling choice, measured in years of forgone earnings, plus an additive term in years of post-schooling experience. Mincer (1974) assumed a linear path for on-the-job investments in human capital after the completion of schooling and showed that earnings would then depend on a *quadratic* function of years of post-schooling experience. Unless the assumptions underlying this derivation are correct, however, the conditional expectation of earnings, given education

and age, will differ from this highly restrictive functional form. Empirically, the model adopted by Mincer (1974) is probably too restrictive (Lemieux, 2006). For example, Murphy and Welch (1990) conclude that a model with a third- or fourth-order polynomial in experience provides a significant improvement in fit.

Researchers have generalized the HCEF by including dummies for degrees (or a complete set of dummies for all possible schooling choices), by including interactions between schooling and experience (or cohort), and by including interactions between schooling and characteristics like gender, cognitive ability, family background, and school quality. Estimation results from such models can be used to calculate 'returns' to schooling that vary by the level or type of schooling and by individual characteristics. Although the resulting estimates cannot be strictly interpreted as internal rates of return, it is conventional to refer to the implied marginal effects as *returns to schooling*.

Related to the issue of functional form is the question of whether post schooling choices – like occupation or industry – should be added as controls to HCEF. From the perspective of calculating the returns to alternative schooling choices, the answer is 'no', since some of the return to additional schooling is the increased chance of working in a more highly paid occupation or industry (Becker, 1964). A more subtle issue is region or urban location, since some part of the wage differential associated with these choices is caused by differences in the cost of living (which presumably should be subtracted from earnings to calculate the return to schooling).

Recent surveys of the returns to schooling based on the Mincerian HCEF (Psacharopoulos, 1994; Psacharopoulos and Patrinos, 2004; Harmon, Walker and Westergaard-Nielson, 2001; 2003) suggest that returns are in the range of 5–15 per cent for most OECD countries, and somewhat higher in developing countries, on average. In Europe, returns appear to be relatively low in Scandinavia (around 5 per cent) and relatively high in the United Kingdom and Ireland (10 per cent or more). Estimated returns in the United States are comparable to those in the United Kingdom, with evidence of a positive trend in both countries over the 1980–2000 period (Katz and Murphy, 1992; Card and Lemieux, 2001; Gosling and Lemieux, 2003). Using meta-analytic techniques, Harmon, Walker and Westergaard-Nielson (2001) conclude that estimated returns are on average 1–2 points lower when the sample is limited to the public sector, when the earnings model includes controls for occupation or 'ability' measures, and when allowances are made for taxes. They also conclude that returns are slightly higher for women than men. In the United States, estimated returns for the mid-1990s from a conventional HCEF based on hourly earnings were about 10 per cent for men and 11 per cent for women (Card, 1999, Table 1).

3.2 Causal studies

In his pioneering study Walsh noted: 'No doubt the students who go on from high school to college are, on average, richer in natural endowments than those who are left behind. They are a selected lot …' (Walsh, 1935, pp. 272–3). Two main methods have been developed to control for the potential selection biases that confound simple earnings comparisons between people with different levels of schooling: (1) comparisons of siblings or twins; (2) comparisons based on interventions or exogenous factors that affected the education choices of one group relative to another. Detailed discussions of these methods are presented in Card (1995a; 1999; 2001), Krueger and Lindahl (2001), Harmon, Oosterbeek

and Walker (2003), and Blundell, Dearden and Sianesi (2004). This section presents a brief overview of some of the main methodological issues – and some of the associated findings – without attempting a comprehensive review.

Gorseline (1932) first proposed the use of sibling comparisons to control for selection biases between different education groups. The basic idea can be illustrated using a variant of the 'homogeneous returns' model discussed in Section 2.2. Letting y_{ij} and S_i denote the earnings and schooling of sibling j (j=1,2) from family i, the homogeneous returns model posits:

$$\log y_{ij} = \alpha_{ij} + \bar{b} S_{ij},$$

where α_{ij} represents the level of earnings that sibling j would receive in the absence of schooling. One possible assumption is that $\alpha_{i1} = \alpha_{i2}$: that is, that the two siblings have equal 'ability'. In this case, one can obtain an unbiased estimate of the true return to schooling from a within-family regression, since

$$\log y_{i1} - \log y_{i2} = \bar{b}(S_{i1} - S_{i2}).$$

Chamberlain and Griliches (1975) re-analysed Gorseline's sample of Indiana brothers and obtained a within-family estimate of $\bar{b}$ equal to 0.080 – only slightly below the estimate of 0.082 obtained from a conventional earnings model estimated by OLS on the same data. (Chamberlain and Griliches, 1975, also included the sibling's differences in age and age-squared as added regressors.) Of course siblings may not have identical abilities, and if they don't, a within-family estimator b_w can be worse (that is, more biased) than the corresponding OLS estimator b_{ols}. Assuming a homogenous returns model is correct, the bias in the OLS estimate is

$$\text{plim } b_{ols} - \bar{b} = \text{cov}[\alpha_{ij}, S_{ij}] \,/\, \text{var}[S_{ij}],$$

while the bias in the within-family estimator is

$$\text{plim } b_w - \bar{b} = \text{cov}[\alpha_{ij} - \alpha_{i2}, S_{i1} - S_{i2}] \,/\, \text{var}[S_{i1} - S_{i2}].$$

Although differencing eliminates the shared component of α_{i1} and α_{i2}, it is possible that the remaining within-family difference in ability is large relative to the within-family variance of schooling, implying a larger bias in b_w than b_{ols}. (A similar analysis can be conducted when both the slope and intercept of the earnings function have person-specific components; see Card, 1999.)

One approach to the concern over ability differences between siblings is to focus on identical (monozygotic) twins. Unfortunately, schooling differences are small among identical twins, and even a little measurement error in reported education can lead to large attenuation bias in the within-twin estimate of the return to schooling. Ashenfelter and Krueger (1994) proposed an innovative solution based on asking each twin about its own and its sibling's education. Their method has been widely adapted in the literature (for example, Ashenfelter and Rouse, 1998; Miller, Mulvey and Martin, 1995; Bonjour et al., 2003) and leads to estimated returns within twins that are comparable to the corresponding OLS estimates, or only slightly smaller. (As noted by Bound and Solon, 1999, if the measurement error in schooling is mean-reverting, the Ashenfelter–Krueger approach ‘over-corrects’ and leads to an upward bias in the resulting estimator.)

Despite the intuitive appeal of identical twins to some researchers, others (for example, Bound and Solon, 1999) have questioned whether twins who choose different schooling levels are really ‘identical’ or whether the small differences in upbringing and experience that lead them to choose different schooling also contribute to their different earnings. Fundamentally, the problem is that the source of the differential schooling choices is unobserved, so different observers can argue that the choice was driven by factors that are either correlated or uncorrelated with earnings. A similar problem arises in ‘matching’ estimates of the return to schooling (see Blundell, Dearden and Sianesi, 2004), which attempt to compare earnings between people who are very similar in all dimensions except their choice of schooling. Indeed, a perfect matching algorithm applied to a sample of twins would presumably match twins to each other, leading to a within-family estimate of the return to schooling.

A second approach to the issue of selection bias is the use of instrumental variables (IV) methods. Specifically, the researcher posits the existence of a variable Z that is exogenous to individuals but affects their schooling choices. As shown by Heckman (1978) in a different context, the equation relating *S* to *Z* need not represent a well-specified model, only a linear projection. For example, assume:

$$S_i = Z_i\pi + \xi_i.$$

If earnings are generated by the homogeneous returns model:

$$\log y_i = \alpha_i + \bar{b}S_i,$$

and Z_i is orthogonal to α_i, then a consistent estimate of $\bar{b}$ can be obtained by IV, using Z_i as an instrument for S_i. Individual-level instruments that have been proposed include quarter of birth (Angrist and Krueger, 1991), the sex composition of one's siblings (Butcher and Case, 1994), and distance to the nearest college (Card, 1995b). Other IV studies use school system reforms such as changes in the minimum school-leaving age (Harmon and Walker, 1995; Oreopoulos, 2006; Meghir and Palme, 2005), changes in tuition at local state colleges (Kane and Rouse, 1993; Fortin, 2006), and expansions in local infrastructure (Duflo, 2001).

Many IV studies yield estimated returns to schooling that are as large as or slightly larger than the corresponding IV estimates (see for example, Card, 2001; Harmon, Oosterbeek and Walker, 2003). Since the IV approach was motivated by the concern that OLS leads to an *overestimate* of the returns to schooling, this is potentially puzzling, and three explanations have been offered. First, OLS estimates are downward-biased by measurement error in education, and the measurement error bias may offset any upward selectivity bias (Griliches, 1977). Second, the search for IV designs that yield statistically significant estimates may create a ‘publication bias’ in favour of samples and specifications with relatively large IV coefficients (Ashenfelter, Harmon and Oosterbeek, 1999). Third, if the returns to education vary across the population, certain instruments may identify returns for subgroups with relatively high marginal returns to schooling (Card, 1995a).

The third explanation can be most easily understood in the context of a social experiment with a randomly assigned intervention (indexed by Z_i=1). Let (S_{i0}, y_{i0}) represent the schooling and earnings outcomes for person i if he or she were assigned to the control group, and (S_{i1}, y_{i1}) denote the outcomes if he or she was assigned to treatment (note that only one of these pairs is observed). The treatment effect on schooling for person i is $\Delta S_i = S_{i1} - S_{i0}$, while the effect on log earnings is $\Delta \log y_i = \log y_{i1} - \log y_{i0}$. Assuming that individual i's marginal return to schooling in the absence of the intervention is β_i, and that the intervention only affects earnings through its effect on schooling, $\Delta \log y_i = \beta_i \Delta S_i$. An IV estimate of the return to schooling based on assignment status is numerically equal to the difference in mean log earnings between the treatment and control groups, divided by the corresponding difference in their average schooling, and has probability limit

$$\text{plim } b_{IV} = \frac{E[\log y_i | Z_i = 1] - E[\log y_i | Z_i = 0]}{E[S_i | Z_i = 1] - E[S_i | Z_i = 0]} = \frac{E[\beta_i \Delta S_i]}{E[\Delta S_i]}.$$

If $E[\beta_i \Delta S_i] = E[\beta_i] E[\Delta S_i]$, then the IV estimator gives a consistent estimate of the average marginal return to education $\bar{\beta} = E[\beta_i]$. This will be true if the intervention induces the same change in schooling for everyone, or more generally if $E[\Delta S_i | \beta_i]$ is independent of β_i. Otherwise, provided that $\Delta S_i \geq 0$ for all i (that is, no one reduces schooling because of the intervention) the IV estimate is a weighted average of the β_i's, with the weight for person i equal to $\Delta S_i / E[\Delta S_i]$.

An intervention that induces larger gains in schooling for people with high values of β_i can lead to an

IV estimate that overstates $\bar{\beta}$. Card (1995a) argued that this might be true for interventions – like the increases in the minimum school leaving age studied by Harmon and Walker (1995) and Oreopoulos (2006) – that mainly affect children from disadvantaged family backgrounds who stop school early because of high marginal costs rather than because of low marginal benefits. An alternative explanation for the finding that an IV estimate exceeds the corresponding OLS estimate is that the assumptions underlying the particular instrumental variable are invalid. In particular, in the absence of a true experiment, one can never 'prove' that the instrument is as good as randomly assigned. Even in an experimental setting, it is also possible that an intervention has an independent causal effect on earnings, confounding the interpretation of the IV estimate. (For example, Willis and Rosen, 1979, and Heckman and Li, 2004, use parental education as instruments for schooling, though others have argued that parental education has an independent effect on earnings.)

A generalization of IV that is useful when a researcher believes there may be random payoffs to schooling is a control function approach, first used in the schooling context by Garen (1984). (Other recent applications include Conneely and Uusitalo, 1997, Blundell, Dearden and Sianesi, 2004, and Rau-Binder, 2006.) This method relies on assumptions about the relationship between the error component in the equation relating schooling to the instrument(s) *Z*, and the random slope and intercept in the earnings equation. Assuming these are satisfied, a control function approach can recover unbiased estimates of the average marginal return to schooling, as well as useful information on how the returns to education vary with the unobserved factors driving the choice of schooling (Rau-Binder, 2006).

4 Summary

The idea of treating schooling as an investment that yields internal rates of return comparable to other investments in the economy has proven extremely useful, and has led to an unusually coherent body of research that combines theoretical modelling and detailed empirical analysis. Much of the existing empirical work is conducted in the framework of Mincer's (1974) human capital earnings function, which relates the logarithm of earnings to completed schooling – measured in years to reflect the opportunity cost of the investment – and a control for post-schooling experience. In a strict equalizing differences framework the coefficient of schooling is the internal rate of return to schooling. In a more general framework that recognizes the endogenous nature of the schooling decision, and the importance of ability differences that partially determine the choice of schooling, observed differences in earnings across different education groups will not necessarily reveal the rate of return to schooling for any one person, or for the population as a whole. Nevertheless, existing evidence from studies of siblings and twins, and from studies that focus on arguably exogenous sources of variation in education choices, suggest that the return to schooling is in the range of 5–15 per cent, and not too different from the value implied by the simple Mincerian approach.

See Also

- Becker, Gary S.
- control functions

- Griliches, Zvi
- human capital
- Mincer, Jacob
- Rosen, Sherwin
- Schultz, T.W.
- selection bias and self-selection

Bibliography

Altonji, J.J. 1993. The demand for and return to education when education outcomes are uncertain. *Journal of Labor Economics* 11, 48–83.

Angrist, J.D. and Krueger, A.B. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979–1014.

Arcidiacono, P. 2004. Ability sorting and the returns to college major. *Journal of Econometrics* 121, 343–75.

Ashenfelter, O., Harmon, C. and Oosterbeek, H. 1999. A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour Economics* 6, 453–70.

Ashenfelter, O. and Krueger, A.B. 1994. Estimates of the economic return to schooling from a new sample of twins. *American Economic Review* 84, 1157–73.

Ashenfelter, O. and Rouse, C. 1998. Income, schooling, and ability: evidence from a new sample of twins. *Quarterly Journal of Economics* 113, 869–95.

Becker, G.S. 1962. Investment in human capital: a theoretical analysis. *Journal of Political Economy* 70, 9–49.

Becker, G.S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: Columbia University Press.

Becker, G.S. 1967. *Human Capital and the Personal Distribution of Income*. Ann Arbor, MI: University of Michigan Press.

Blundell, R., Dearden, L. and Sianesi, B. 2004. Evaluating the impact of education on earnings in the U. K.: models, methods, and results from the NCDS. Working Paper No. 03/20, Institute for Fiscal Studies.

Bonjour, D., Cherkas, L., Haskel, J., Hawkes, D. and Spector, T. 2003. Returns to education: evidence from UK twins. *American Economic Review* 93, 1799–812.

Bound, J. and Solon, G. 1999. Double trouble: on the value of twins-based estimation of the returns to education. *Economics of Education Review* 18, 169–82.

Butcher, K.F. and Case, A. 1994. The effect of sibling composition on women's education and earnings. *Quarterly Journal of Economics* 109, 531–63.

Card, D. 1995a. Earnings, schooling, and ability revisited. In *Research in Labor Economics*, vol. 14, ed. S. Polachek. Greenwich, CT: JAI Press.

Card, D. 1995b. Using geographic variation in college proximity to estimate the return to schooling. In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. L.N. Christofides, E. K. Grant and R. Swidinsky. Toronto: University of Toronto Press.

Card, D. 1999. The causal effect of education on earnings. In *Handbook of Labor Economics*, vol. 3A, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Card, D. 2001. Estimating the return to schooling: progress on some persistent econometric problems. *Econometrica* 69, 1127–60.

Card, D. and Lemieux, T. 2001. Can falling supply explain the rising return to college for younger men? A cohort-based analysis. *Quarterly Journal of Economics* 116, 705–46.

Chamberlain, G. and Griliches, Z. 1975. Unobservables with a variance–covariance structure: ability, schooling, and the economic success of brothers. *International Economic Review* 16, 422–49.

Conneely, K. and Uusitalo, R. 1997. Estimating heterogeneous treatment effects in the Becker schooling model. Discussion Paper, Industrial Relations Section, Princeton University.

Cunha, F. and Heckman, J.J. 2006. Identifying and estimating the distributions of *ex post* and *ex ante* returns to schooling: a survey of recent developments. Working Paper, University of Chicago.

Duflo, E. 2001. Schooling and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *American Economic Review* 91, 795–813.

Eckstein, Z. and Wolpin, K.I. 1999. Why youths drop out of high school: the impact of preferences, opportunities and abilities. *Econometrica* 67, 1295–339.

Fortin, N. 2006. Higher education policies and the college premium: cross-state evidence from the 1990s. *American Economic Review* 96, 959–87.

Freeman, R.B. 1976. *The Over-Educated American*. New York: Academic Press.

Garen, J. 1984. The returns to schooling: a selectivity bias approach with a continuous choice variable. *Econometrica* 52, 1199–218.

Gorseline, D.E. 1932. *The Effect of Schooling Upon Income*. Bloomington: University of Indiana Press.

Gosling, A. and Lemieux, T. 2003. Labor market reforms and changes in wage inequality in the United Kingdom and the United States. In *Seeking a Premier Economy: The Economic Effects of British Economic Reforms, 1980–2000*, ed. D. Card, R. Blundell and R.B. Freeman. Chicago: University of Chicago Press.

Griliches, Z. 1977. Estimating the returns to schooling: some econometric problems. *Econometrica* 45, 1–22.

Harmon, C. and Walker, I. 1995. Estimates of the economic return to schooling for the United Kingdom. *American Economic Review* 85, 1278–86.

Harmon, C., Oosterbeek, H. and Walker, I. 2003. The returns to education: microeconomics. *Journal of Economic Surveys* 17, 115–56.

Harmon, C., Walker, I. and Westergaard-Nielson, N. 2001. Introduction. In *Education and Earnings in Europe*, ed. C. Harmon, I. Walker and N. Westergaard-Nielson. Aldershot: Edward Elgar.

Harmon, C., Walker, I. and Westergaard-Nielson, N. 2003. The returns to education: microeconomics. *Journal of Economic Surveys* 17, 115–56.

Heckman, J.J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46, 931–59.

Heckman, J.J. and Li, X. 2004. Selection bias, comparative advantage, and heterogeneous returns to education: evidence from China in 2000. *Pacific Economic Review* 9, 155–71.

Hungerford, T. and Solon, G. 1987. Sheepskin effects in the return to education. *Review of Economics and Statistics* 69, 175–7.

Kane, T.J. and Rouse, C.E. 1993. Labor market returns to two- and four-year colleges: is a credit a credit and do degrees matter? Working Paper No. 4268. Cambridge, MA: NBER.

Katz, L.F. and Murphy, K.M. 1992. Changes in relative wages, 1963–1987: supply and demand factors. *Quarterly Journal of Economics* 107, 35–78.

Keane, M.P. and Wolpin, K.I. 1997. The career decisions of young men. *Journal of Political Economy* 105, 473–521.

Keane, M.P. and Wolpin, K.I. 2001. The effect of parental transfers and borrowing constraints on educational attainment. *International Economic Review* 42, 1051–103.

Krueger, A.B. and Lindahl, M. 2001. Education for growth: why and for whom? *Journal of Economic Literature* 39, 1101–36.

Lemieux, T. 2006. The ‘Mincer equation’ thirty years after *Schooling, Experience, and Earnings*. In *Jacob Mincer: A Pioneer of Modern Labor Economics*, ed. S. Grossbard-Shechtman. Heidelberg: Springer.

Meghir, C. and Palme, M. 2005. Educational reform, ability and parental background. *American Economic Review* 95, 414–24.

Miller, P.W., Mulvey, C. and Martin, N. 1995. What do twins studies reveal about the economic returns to education? A comparison of Australian and US findings. *American Economic Review* 85, 586–99.

Mincer, J. 1958. Investment in human capital and personal income distribution. *Journal of Political Economy* 66, 281–302.

Mincer, J. 1974. *Schooling, Experience, and Earnings*. New York: NBER (distributed by Columbia University Press).

Mincer, J. 1997. Changes in wage inequality, 1970–1990. *Research in Labor Economics* 16, 1–18.

Murphy, K.M. and Welch, F. 1990. Empirical age–earnings profiles. *Journal of Labor Economics* 8, 202–29.

Oreopoulos, P. 2006. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review* 96, 152–75.

Park, J.H. 1994. Returns to schooling: a peculiar deviation from linearity. Working Paper No. 335, Industrial Relations Section, Princeton University.

Psacharopoulos, G. 1994. Returns to investment in education: a global update. *World Development* 22, 1325–43.

Psacharopoulos, G. and Patrinos, H.A. 2004. Returns to investment in education: a further update. *Education Economics* 12, 111–34.

Rau-Binder, T. 2006. Semiparametric estimation of microeconometric models with endogenous regressors and sorting. Manuscript, Department of Economics, University of California Berkeley.

Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.

Solon, G. 1999. Intergenerational mobility in the labor market. In *Handbook of Labor Economics*, vol. 3A, ed. O.C. Ashenfelter, D. Card and R. Layard. Amsterdam: North-Holland.

Walsh, J.R. 1935. Capital concept applied to man. *Quarterly Journal of Economics* 49, 255–85.

Willis, R. 1986. Wage determination: a survey and reinterpretation of human capital earnings functions. In *Handbook of Labor Economics*, vol. 1, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

Willis, R. and Rosen, S. 1979. Education and self-selection. *Journal of Political Economy* 87, S7–S36.

How to cite this article

Card, David. "returns to schooling." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000240> doi:10.1057/9780230226203.1434

The New Palgrave Dictionary of Economics Online

revealed preference theory

Marcel K. Richter

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Economists do not observe preferences. They may, however, observe demand behaviour – the choices made by consumers. Is there a way for economists to tell whether the observed behaviour is generated through the maximization of a preference relation or utility function? Since most economic theories are ultimately based on a consumer who maximizes a preference or utility, the question is clearly important for developing and testing theories.

Revealed preference theory answers this question by characterizing choice behaviour that is generated by preference or utility maximization. Relating choice behaviour and preference maximization is also a goal of integrability theory. What distinguishes the theories from each other, and from the other parts of rationality theory, is the special nature of their tools: integrability theory uses mathematical integration in its proofs, and usually states its hypotheses in differential form; revealed preference theory uses a variety of mathematical tools for its proofs, and its hypotheses are usually in a discrete ‘revelation’ form. The distinctions are not always sharp, however, and we shall see areas in which the theories overlap.

Samuelson invented revealed preference theory in 1938. The basic idea, much of the terminology, and some of the axioms are due to him. In the following outline, a useful paradigm is the one that guided the first three decades: a consumer with a finite-dimensional euclidean commodity space, facing ‘competitive’ budgets determined by fixed positive prices, and satisfying a budget equality constraint.

1 The problem of rationality

From the economist's point of view, unobservable preferences generate observable choices. Since many preference relations may generate the same choice correspondence, the map from preferences to choice correspondences is many-one: We cannot hope to find *the* preference generating choices, but only *some* preferences – a set of ‘equivalent’ preferences. For example:

It is well known in preference theory that a lexicographic preference on the plane does not admit a real-valued utility function (Debreu, 1954). A hasty conclusion might be that there is no hope of representing a lexicographic-maximizing consumer as a utility-maximizing consumer. Too hasty! For her behaviour clearly maximizes this function g on the non-negative plane (for positive prices):

$$g(x_1, x_2) = x_1$$
(1)

Even if her ‘intention’ is to maximize a lexicographic preference, she acts *as if* her intention were to maximize *g*. In fact, even if the choices were made by a committee, a machine, or any other mindless decision maker, we can still say the actions are *as if* the intent were *g*-maximization.

This example shows the distinction between a typical question in utility and preference theory (‘Does *this* preference have a utility function?’), and the basic question in revealed preference and integrability theories (‘Is this demand generated by *some* preference?’). It also demonstrates the need for precise definitions. (Our notation will follow the glossaries of Richter, 1966, 1971.)

To describe choices, the theory requires an underlying set *X* and a family $\mathcal{B}$ of subsets $B \subset X$. (Often *X* is the non-negative orthant of *n*-space and each *B* is a ‘competitive’ budget determined by positive prices and income.) We call any $B \in \mathcal{B}$ a *budget*. A *choice* or *demand correspondence h* is a function assigning to each $B \in \mathcal{B}$ a subset $h(B) \subset B$ interpreted as the set of elements chosen from *B*. And any binary relation on *X* is called a *preference*. Rationality theory relates choices *h* on (*X*, $\mathcal{B}$) to preferences *R* on *X* in two ways.

(i) If we start with a preference relation R we can ask what kind of choice it generates. There are two obvious senses in which it could generate a choice *h*. First, we might have, for all $B \in \mathcal{B}$

$$h(B) = \{x \in B: \forall y_{y \in B} xRy\},$$
(2)

i.e., the set of elements chosen from *B* is the set of *R*-most preferred elements in *B*. Then we say that *R rationalizes h* (Richter, 1971).

Alternatively, we might have, for all $B \in \mathcal{B}$

$$h(B) = \{x \in B: \forall y_{y \in B} yRx\},$$
(3)

i.e., the set of elements chosen from *B* is the set of elements in *B* for which nothing in *B* is *R*-more preferred. Then we say that *R motivates h* (Kim and Richter, 1986).

Definition (2): is appropriate if we think of *R* as a ‘weak’ (i.e. reflexive) relation, while (3) is appropriate if we think of *R* as a ‘strict’ (asymmetric) relation.

(ii) Conversely, if we start with a choice *h* we can ask whether any preference *R* generates, or ‘explains’

h. If there exists some *R* generating *h* in the sense of (2) (Richter, 1966, 1971), then we say that *h* is *rational*. Often we are interested in *reflexive*-rationality (rationalization by a reflexive preference), *transitive*-rationality (rationalization by a transitive preference), *regular*-rationality (rationalization by a reflexive, transitive, and total preference), etc. For example, *utility–rationality* requires the existence of a function $f: X \to R^1$ satisfying

$$h(B) = \{x \in B: \forall y_{y \in B} f(x) \geqq f(y)\} \tag{4}$$

for all $B \in \mathcal{B}$ – i.e., the set of elements chosen from *B* is the set of those elements in *B* with the highest utility.

If there exists some *R* generating *h* in the sense of (3), then we say that *h* is *motivated* (Kim and Richter, 1986). Again, we are often interested in *asymmetric*-motivation (motivation by an asymmetric preference), etc. In fact, *h* is rational if and only if it is motivated (Clark, 1985; Kim and Richter, 1986). Of course, the example (1) makes it clear that such a rationalizing or motivating *R* will not usually be unique: there will be a whole equivalence class of such relations generating the same choice (Kim and Richter, 1986).

It is important to note that rationality and motivation have been defined as properties of demand, not of preference. We do not say, for example, that a particular preference is rational or irrational. Instead, the definitions relate demand and preferences.

An economist who derives comparative statics results from preference maximization is answering qsts of type (i). The issue arises – for both theoretical development and empirical testing – whether any further results can be derived, or whether all the (independent) consequences of preference maximization have been found. This is usually a much more difficult problem. A major task of both revealed preference and integrability theory is to address this issue, by answering qsts of type (ii). The two qsts are parts, then, of the fundamental Problem of Rationality: give necessary (i) and sufficient (ii) conditions for a demand to be rational (of a particular type), or motivated (of a particular type). Revealed preference theory solves the problem through axioms with a unique flavour.

2 Revealed Preference Solutions

It is important to distinguish revealed preference definitions from revealed preference axioms, and these in turn from revealed preference theorems.

2.1 Revelation Definitions

If consumer (i.e. choice) *h* selects alternative $x \in B$ from *B* – i.e., if $x \in h(B)$ – when alternative *y* could have been selected – i.e., if $y \in B$ – then we write *xVy*. And it is natural to say that *x* is *revealed as good as y*. If also $x \neq y$ then we write *xSy*, and it is natural to say that *x* is *revealed preferred to y*. This

terminology of Samuelson's is very suggestive, because if $\succcurlyeq$ is any rationalization, then *xVy* implies $x \succcurlyeq y$ as does *xSy*. In fact, if $x \in h(B)$ & $y \in (B)h(B)$ and if $\succcurlyeq$ is regular, then its asymmetric part $\succ$ also satisfies $x \succ y$. So an observer of *h* can deduce properties common to all rationalizations. But beware: *xSy* is a statement about *choice*, not about a particular preference.

Unlike the psychologist, who may be able to present an individual with binary choices, and thereby uncover a total ordering, the economist will typically observe *S* as only a partial ordering. This is one of the challenging features of revealed preference theory. It is why, mathematically, revealed preference theory is a study of partial orders, in contrast to the classical theory of preference, which is a theory of total orders. It is also why there is generally more than one preference in the equivalence class of preferences that rationalize or motivate a given choice.

2.2 Revelation axioms

We describe four revealed preference axioms. Samuelson proposed the *asymmetry* of *S* as a basic axiom of consumer theory: for all $x, y \in X$

$$xSy \Rightarrow ySx. \tag{5}$$

In other words, if *x* is revealed preferred to *y* (under some budget), then *y* is never (under any budget) revealed preferred to *x*. As Samuelson noted, this is a property of any single-valued demand function maximizing a regular preference. This is now called the *Weak Axiom of Revealed Preference*.

Houthakker noted other necessary consequences of regular-rationality, for single-valued demand functions: there can be no cycles of the form

$$xSv_1Sv_2S\ldots Sv_kSx. \tag{6}$$

In other words, *x* is never, even indirectly, revealed preferred to itself. Houthakker proposed this as a new axiom, now called the *Strong Axiom of Revealed Preference*. If we define *xHy* to mean that *xSy* or $xSv_1Sv_2S\ldots Sv_kSy$, then we can rephrase Houthakker's axiom as saying that *H* is asymmetric. In other words, if *x* is (even indirectly) revealed preferred to *y*, then *y* is never (even indirectly) revealed preferred to *x*.

Richter noted still another consequence of regular-rationality. For this it is convenient to define *xWy* to mean either *xVy* or $xVu_1V\ldots Vu_kVy$. Clearly regular-rationality implies: for all $x, y \in X$ & $B \in \mathcal{B}$

$$x \in h(B) \ \& \ y \in B \ \& \ yWx \Rightarrow y \in h(B). \tag{7}$$

In other words, if *x* is chosen from *B*, and if *y* is also available in *B* and is revealed (even directly) as good as *x*, then *y* is also chosen from *B*. This is the *Congruence Axiom of Revealed Preference*. He also noted a behavioural consequence of any rationality: for all $x, y \in X$ & $B \in$

$$x \in B \ \forall y_{y \in B} xVy \Rightarrow x \in h(B). \tag{8}$$

In other words, if *x* is in *B* and is revealed as good as everything in *B*, then *x* is chosen from *B*. This is the *V-Axiom*.
We will use these axioms to discuss the main solutions to the Problem of Rationality.

2.3 Revelation theorems

(a) *Weak Axiom*. Samuelson proposed the Weak Axiom in 1938, as a foundation for all consumer theory (1938 a,b). He did not name it, but he suggested that (for single-valued demand functions) it followed from maximizing a utility function (cf. also Samuelson (1955), pp. 110–11). In the opposite direction, his idea of founding consumer theory on it was implicitly a conjecture that it implied utility-rationality, or at least regular-rationality. Indeed, after preliminary work by I.M.D. Little, Samuelson succeeded in showing that, for two commodities and Lipschitz-continuous demand functions, the Weak Axiom implied regular-rationality (Samuelson, 1948).
(b) *Strong Axiom*. Then in 1950 Houthakker (1950) proposed the Strong Axiom (by a different name) as a basis for consumer theory, and showed that, for any number of commodities, it implied utility-rationality for Lipschitz-continuous demand functions. Samuelson (1950) then gave the Weak and Strong Axioms their modern names.
In 1959, Uzawa (1960, 1971) developed a more precise analogue of Houthakker's result, showing that the Strong Axiom and a Lipschitzian hypothesis on the demand implied irreflexive-transitive-monotone-convex-lower semi continuous-motivation. His proof was along the lines of the Samuelson–Little–Samuelson–Houthakker analytic methods.
Although the Strong Axiom implied the Weak, it was still not clear whether the Weak implied the Strong. Indeed, Rose (1958) showed that the Weak Axiom does imply the Strong Axiom, when there are only two commodities and prices are positive (needed!). Then Gale (1960) constructed an example with three commodities, showing that the Weak Axiom did not imply the Strong. And Kihlstrom, Mas-Colell and Sonnenschein (1976) showed how to obtain very easily many examples, for any number of commodities greater than two. And Shafer (1977b), affirming a conjecture of Samuelson (1953), showed

that the full strength of the Strong Axiom is needed: even for three goods, there is no upper bound on the length of *S*-cycles that must be ruled out. In the opposite direction, several authors have discussed special conditions under which the Weak Axiom does imply the Strong (Arrow, 1959; Uzawa, 1960, 1971).

Richter (1966) used set-theoretic methods – very different from the analytic methods of Samuelson, Little, Houthakker and Uzawa – to simplify the proofs, eliminate extraneous assumptions, and strengthen the rationality results. In a framework of abstract budget spaces, and without the technical assumptions required by the earlier analytical approaches, he showed that the Strong Axiom is equivalent to regular-rationality for demand functions. Thus the Strong Axiom completely exhausts the theory of demand functions maximizing a regular preference.

Richter (1966) also showed that, if a competitive demand satisfies the Strong Axiom, then it is utility-rational if its range is well behaved, but it may not be utility-rational otherwise (Richter, 1971).

Extensions

There have been many extensions. Richter (1966) showed that the V-Axiom characterized rationality, and the Congruence Axiom characterized regular-rationality, for demand correspondences. (Hansson (1968) gave an alternative criterion for regular-rationality.) Other extensions have obtained stronger properties of the rationalization under special hypotheses (Hurwicz and Richter, 1971; Mas-Colell, 1978, Theorem 1; Richter, 1986; Matzkin and Richter, 1986); uniqueness of the rationalization within certain classes (Mas-Colell, 1977); revealed preference axioms characterizing more general rationality types (Richter, 1971; Kim and Richter, 1986; Kim, 1987); dual axioms (Sakai, 1977; Richter, 1979); and axioms for stochastic rationality (McFadden and Richter, 1970).

Applications

Several applications have supported Samuelson's original idea that revealed preference could provide an alternative to preference theory as a foundation for consumer theory. Revealed preference techniques have been applied to prove the existence of competitive equilibrium (Wald, 1936, 1951); to prove the stability of competitive equilibrium (Arrow and Hurwicz, 1958, 1960); to prove the Hicks Composite Commodity Theorem (Richter, 1970; Calsamiglia, 1978); to analyse and characterize aggregate excess demand functions (Debreu, 1974; McFadden et al., 1974); to prove aggregation properties for correspondences (Shafer, 1977a); to prove properties of measurable demand correspondences (Yamazaki, 1984); to prove theorems about social choice functions (Plott, 1973); etc.

3 Revealed preference and integrability

With the same rationality goal as revealed preference theory, integrability theory uses axioms on the Slutsky or Antonelli matrices to characterize rational choice (cf. Hurwicz, 1971; see also integrability of demand). Under some smoothness assumptions on the demand function, the basic theorems state that

symmetry and negative semidefiniteness of these matrices is necessary and sufficient for (upper-semicontinuous-) regular-rationality.

Samuelson established a link between revealed preference theory and integrability theory by showing that his Weak Axiom implied negative semidefiniteness of the matrices (Samuelson, 1938b, 1955, pp. 111–14). Later Kihlstrom, Mass-Colell, and Sonnenschein (1976) demonstrated that negative semidefiniteness was equivalent to a Weak Weak Axiom.

This left open the question of finding a revealed preference axiom equivalent to the symmetry. The Strong Axiom was clearly too strong, since it already implied regular-rationality, and therefore both symmetry and negative-semidefiniteness. Then Hurwicz and Richter (1979a,b) showed that a differential axiom of Ville (1946, 1951) provided the exact strength needed. Although it does not even imply the Weak Axiom, it is similar in spirit to the Strong Axiom and can be given a revealed preference inpt. It thus serves, like Kihlstrom, Mas-Colell and Sonnenschein's Weak Weak Axiom, as a bridge between the Revealed Preference and Integrability approaches to consumer rationality. Richter (1979) discussed these bridges from the viewpoint of duality.

4 Other notions of rationality

Many economists have used notions of rationality different from Richter's notion (2).

Sometimes the term ‘rational’ has been applied to preference, rather than demand. (In such applications it is often a synonym for ‘transitive’.) In Uzawa (1957) and Arrow (1959), on the other hand, it was applied to demand, but only in terms of axioms on demand behaviour. By contrast, (2) is applied to demand, but in terms that relate both demand and preferences.

Some economists have used weaker notions of rationality than (2), requiring only: for all $x, y \in X$ & $B \in \mathcal{B}$

$$h(B) \subset \{x \in B: \forall y_{y \in B} xRy\}. \tag{9}$$

In other words, every element chosen from *B* is *R*-most preferred in *B*, but *B* may contain other *R*-most preferred elements that are not chosen. We will call this *subsemi-rationality*, although it has often been referred to as rationality.

A drawback of this concept is its loose linkage of preference and demand. Any constant function, for example, satisfies (9). On the other hand, if one interprets *h*(*B*) as a set of incomplete observations, then one might wonder whether, with more observations of choices from *B*, the set *h*(*B*) of chosen elements might grow. Then one might want to find a preference *R* satisfying just (9), rather than insisting (as does (2)) that *R* explain *precisely* the observed set *h*(*B*).

Afriat (1967) gave conditions on a demand function, over a finite set of budgets, that are necessary and sufficient for it to be subsemi-rationalized by a continuous monotone concave function. His work was clarified by Diewert (1973) who gave a criterion for continuous-monotone-concave-subsemirationality

in terms of a linear programming problem. Varian (1983) restated Afriat's finite-budgets result in terms of a Generalized Axiom of Revealed Preference – weaker than the Strong Axiom.
Matzkin and Richter (1986) obtained full rationality by replacing the Generalized Axiom with the Strong Axiom, which they proved was necessary and sufficient for continuous-monotone-strictly concave-utility-rationality in the finite case. No revealed preference criterion for concave-regular-rationality is known for the not-necessarily finite case.

See Also

- demand theory
- integrability of demand
- Samuelson, Paul Anthony

Bibliography

Afriat, S.N. 1967. The construction of utility functions from expenditure data. *International Economic Review* 8, 67–77.

Arrow, K.J. and Hurwicz, L. 1958. On the stability of competitive equilibrium, I. *Econometrica* 26, 522–52.

Arrow, K.J. 1959. Rational choice functions and orderings. *Economica* N.S. 26, 121–7.

Arrow, K.J. and Hurwicz, L. 1960. Some remarks on the equilibria of economic systems. *Econometrica* 28, 640–46.

Calsamiglia, X. 1978. Composite goods and revealed preference. *International Economic Review* 19, 395–404.

Clark, S.A. 1985. A complementary approach to the strong and weak axioms of revealed preference. *Econometrica* 53, 1459–63.

Debreu, G. 1954. Representation of a preference ordering by a numbering function. In *Decision Processes*, ed. R.M. Thrall, C.H. Coombs and R.L. Davis. New York: Wiley, 159–65.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–21.

Diewert, W.E. 1973. Afriat and revealed preference theory. *Review of Economic Studies* 40, 419–25.

Gale, D. 1960. A note on revealed preference. *Economica* . NS 27, 348–54.

Hansson, B. 1968. Choice structures and preference relations. *Synthese* 18, 443–58.

Houthakker, H.S. 1950. Revealed preference and the utility function. *Economica* NS 17, 159–74.

Hurwicz, L. 1971. On the problem of integrability of demand functions. In *Preferences, Utility and Demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt, Brace, Jovanovich, ch. 9.

Hurwicz, L. and Richter, M.K. 1971. Revealed preference without demand continuity assumptions. In *Preferences, Utility and Demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt, Brace, Jovanovich, ch. 3.

Hurwicz, L. and Richter, M.K. 1979a. An integrability condition with applications to utility theory and thermodynamics. *Journal of Mathematical Economics* 6, 7–14.

Hurwicz, L. and Richter, M.K. 1979b. Ville axioms and consumer theory. *Econometrica* 47, 603–19.

Kihlstrom, R., Mas-Colell, A. and Sonnenschein, H. 1976. The demand theory of the weak axiom of revealed preference. *Econometrica* 44, 971–8.

Kim, T. and Richter, M. 1986. Nontransitive-nontotal consumer theory. *Journal of Economic Theory* 38, 324–63.

Kim, T. 1987. Intransitive indifference and revealed preference. *Econometrica*.

Mas-Colell, A. 1977. The recoverability of consumers' preferences from market demand behavior. *Econometrica* 45, 1409–30.

Mas-Colell, A. 1978. On revealed preference analysis. *Review of Economic Studies* 45, 121–31.

Matzkin, R. and Richter, M.K. 1986. Testing concave rationality. Department of Economics, University of Minnesota, Minneapolis.

McFadden, D. and Richter, M.K. 1970. Stochastic rationality and revealed stochastic preference. Presented to the 1970 Winter Meetings of the Econometric Society.

McFadden, D. and Richter, M.K. 1988. Stochastic rationality and revealed stochastic preference. In *Uncertainty, Preferences and Optimality. Essays in Honor of Leonid Hurwicz*, ed. J.S. Chipman, D. McFadden and M.K. Richter, New York.

McFadden, D., Mas-Colell, A., Mantel, R. and Richter, M.K. 1974. A characterization of community

excess demand functions. *Journal of Economic Theory* 9, 361–74.

Plott, C.R. 1973. Path independence, rationality, and social choice. *Econometrica* 41, 1075–91.

Richter, M.K. 1966. Revealed preference theory. *Econometrica* 34, 635–45.

Richter, M.K. 1971. Rational Choice. In *Preferences, Utility, and Demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter and H.F. Sonnenschein. New York: Harcourt, Brace, Jovanovich, ch. 2.

Richter, M.K. 1979. Duality and rationality. *Journal of Economic Theory* 20, 131–81.

Richter, M.K. 1986. Continuous demand functions. Department of Economics, University of Minnesota, Minneapolis.

Rose, H. 1958. Consistency of preference: the two-commodity case. *Review of Economic Studies* 25, 124–5.

Sakai, Y. 1977. Revealed favorability, indirect utility, and direct utility. *Journal of Economic Theory* 14, 113–29.

Samuelson, P.A. 1938a. A note on the pure theory of consumer's behaviour. *Economica* NS 5, 61–71.

Samuelson, P.A. 1938b. A note on the pure theory of consumer's behaviour: an addendum. *Economica* NS 5, 353–4.

Samuelson, P.A. 1948. Consumption theory in terms of revealed preference. *Economica* NS 15, 243–53.

Samuelson, P.A. 1950. The problem of integrability in utility theory. *Economica* NS 17, 355–85.

Samuelson, P.A. 1953. Consumption theorems in terms of overcompensation rather than indifference comparisons. *Economica* NS 20, 1–9.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Shafer, W.J. 1977a. Revealed preference and aggregation. *Econometrica* 45, 1173–82.

Shafer, W.J. 1977b. Revealed preference cycles and the Slutsky matrix. *Journal of Economic Theory* 16, 293–309.

Uzawa, H. 1957. Note on preference and axioms of choice. *Annals of the Institute of Statistical Mathematics* 8, 35–40.

Uzawa, H. 1960. Preference and rational choice in the theory of consumption. In *Mathematical Methods in the Social Sciences, 1959*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press, ch. 9.

Uzawa, H. 1971. Preference and rational choice in the theory of consumption. In *Preferences, Utility, and Demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter and H.F. Sonnenschein. New York: Harcourt, Brace, Jovanovich, ch. 1.

Varian, H.R. 1983. Non-parametric tests of consumer behaviour. *Review of Economic Studies* 50, 99–110.

Ville, J. 1946. Sur les conditions d'existence d'une ophélimité totale et d'un indice du niveau des prix. *Annales de l'Université de Lyon* 9, Sec. A(3), 32–9.

Ville, J. 1951. The existence conditions of a total utility function. *Review of Economic Studies* 19, 123–8.

Wald, A. 1936. Über einige Gleichungssyteme der mathematischen ökonomie. *Zeitschrift für Nationalökonomie* 7, 637–70.

Wald, A. 1951. On some systems of equations of mathematical economics. *Econometrica* 19, 368–403.

Yamazaki, A. 1984. The critical set of a demand correspondence in the price space and the weak axiom of revealed preference. *Hitotsubashi Journal of Economics* 25, 137–44.

How to cite this article

Richter, Marcel K. "revealed preference theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000135> doi:10.1057/9780230226203.1435

The New Palgrave Dictionary of Economics Online

revelation principle

Roger B. Myerson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In any economic institution, individuals must be given appropriate incentives to share private information or to exert unobserved efforts. The revelation principle is a technical insight that allows us to make general statements about what allocation rules are feasible, subject to incentive constraints, in economic problems with adverse selection and moral hazard. The revelation principle tells us that, for any general coordination mechanism, any equilibrium of rational communication strategies for the economic agents can be simulated by an equivalent incentive-compatible direct-revelation mechanism, where a trustworthy mediator maximally centralizes communication and makes honesty and obedience rational equilibrium strategies for the agents.

Keywords

adverse selection; Aumann, R.; Bayesian games; capitalism; centralization; communication; correlated equilibrium; decentralization; direct-revelation mechanisms; Hayek, F. von; honesty; incentive compatibility; incentive constraints; moral hazard; Nash, J.; obedience; principal and agent; private information; revelation principle; sequential equilibrium; socialism; strategic-form games; trust

Article

Communication is central to the economic problem (Hayek, 1945). Opportunities for mutually beneficial transactions cannot be found unless individuals share information about their preferences and endowments. Markets and other economic institutions should be understood as mechanisms for facilitating communication. However, people cannot be expected to reveal information when it is against their interests; for example, a seller may conceal his willingness to sell at a lower price. Rational behaviour in any specific communication mechanism can be analysed using game-theoretic equilibrium concepts, but efficient institutions can be identified only by comparison with all possible communication mechanisms. The revelation principle is a technical insight that allows us, in any given economic

situation, to make general statements about all possible communication mechanisms.

The problem of making statements about all possible communication systems might seem intractably complex. Reports and messages may be expressed in rich languages with unbounded vocabulary. Communication systems can include both public announcements and private communication among smaller groups. Communication channels can have noise that randomly distorts messages. A communication mechanism may also specify how contractually enforceable transactions will depend on agents' reports and messages. So a general communication mechanism for any given set of agents may specify (*a*) a set of possible reports that each agent can send, (*b*) a set of possible messages that each agent can receive from the communication system, and (*c*) a probabilistic rule for determining the messages received and the enforceable transactions as a function of the reports sent by the agents.

However, the revelation principle tells us that, for many economic purposes, it is sufficient for us to consider only a special class of mechanisms, called ‘incentive-compatible direct-revelation mechanisms’. In these mechanisms, every economic agent is assumed to communicate only with a central mediator. This mediator may be thought of as a trustworthy person or as a computer at the centre of a telephone network. In a direct-revelation mechanism, each individual is asked to report all of his private information confidentially to the mediator. After receiving these reports, the mediator then specifies all contractually enforceable transactions, as a function of these reports. If any individual controls private actions that are not contractually enforceable (such as efforts that others cannot observe), then the mediator also confidentially recommends an action to the individual. A direct-revelation mechanism is any rule for specifying how the mediator determines these contractual transactions and privately recommended actions, as a function of the private-information reports that the mediator receives.

A direct-revelation mechanism is said to be ‘incentive compatible’ if, when each individual expects that the others will be honest and obedient to the mediator, then no individual could ever expect to do better (given the information available to him) by reporting dishonestly to the mediator or by disobeying the mediator's recommendations. That is, the mechanism is incentive compatible if honesty and obedience is an equilibrium of the resulting communication game. The set of incentive-compatible direct-revelation mechanisms has good mathematical properties that often make it easy to analyse because it can be defined by a collection of linear inequalities, called ‘incentive constraints’. Each of these incentive constraints expresses a requirement that an individual's expected utility from using a dishonest or disobedient strategy should not be greater than the individual's expected utility from being honest and obedient, when it is anticipated that everyone else will be honest and obedient.

The analysis of such incentive-compatible direct-revelation mechanisms might seem to be of limited interest, because real institutions rarely use such fully centralized mediation and often generate incentives for dishonesty or disobedience. For any equilibrium of any general communication mechanism, however, there exists an incentive-compatible direct-revelation mechanism that is essentially equivalent. This proposition is the revelation principle. Thus, the revelation principle tells us that, by analysing the set of incentive-compatible direct-revelation mechanisms, we can derive general properties of all equilibria of all coordination mechanisms.

The terms ‘honesty’ and ‘obedience’ here indicate two fundamental aspects of the general economic problem of communication. In a general communication system, an individual may send out messages or reports to share information that he knows privately, and he may also receive messages or recommendations to guide actions that he controls privately. The problem of motivating individuals to report their private information honestly is called ‘adverse selection’, and the problem of motivating

individuals to implement their recommended actions obediently is called ‘moral hazard’. To describe the intuition behind the revelation principle, let us consider first the special cases where only one or the other of these problems exists.

Pure adverse selection

First, let us formulate the revelation principle for the case of pure adverse selection, as developed in Bayesian social choice theory. In this case we are given a set of individuals, each of whom has some initial private information that may be called the individual's ‘type’, and there is a planning question of how a social allocation of resources should depend on the individuals' types. Each individual's payoff can depend on the resource allocation and on the types of all individuals according to some given utility function, and each type of each individual has some given probabilistic beliefs about the types of all other individuals. A general communication system would allow each individual i to send a message m_i in some rich language, and then the chosen resource allocation would depend on all these messages according to some rule $\gamma(m_1, \ldots m_n)$. In any equilibrium of the game defined by this communication system, each individual i must have some strategy σ_i for choosing his message as a function of his type t_i, so that $m_i = \sigma_i(t_i)$.

For the given equilibrium $(\sigma_1, \ldots, \sigma_n)$ of the given social-choice rule γ, the revelation principle is satisfied by a mediation plan in which each individual is asked to confidentially report his type t_i to a central mediator, who then implements the social choice

$$\mu(t_1, \ldots t_n) = \gamma(\sigma_1(t_1), \ldots, \sigma_n(t_n)).$$

So the mediator computes what message would be sent by the reported type of each individual i under his or her strategy σ_i, and then the mediator implements the resource allocation that would result from these messages under the rule γ. It is easy to see that honesty is an equilibrium under this mediation plan μ. If any individual could gain by lying to this mediator, when all others are expected to be honest, then this individual could have also gained by lying to himself when implementing his equilibrium strategy σ_i under the given mechanism γ, which would contradict the optimality condition that defines an equilibrium. So μ is an incentive-compatible direct-revelation mechanism that is equivalent to the given general mechanism γ with the given equilibrium $(\sigma_1, \ldots, \sigma_n)$.

In this case of pure adverse selection, the revelation principle was introduced by Gibbard (1973), but for a narrower solution concept (dominant strategies, instead of Bayesian equilibrium). The revelation principle for the broader solution concept of Bayesian equilibrium was recognized by Dasgupta, Hammond and Maskin (1979), Harris and Townsend (1981), Holmstrom (1977), Myerson (1979), and Rosenthal (1978).

Pure moral hazard

Next let us formulate the revelation principle for the case of pure moral hazard, as developed in Aumann's (1974) theory of correlated equilibrium. In this case we are given a set of individuals, each of whom controls some actions, and each individual's payoff can depend on the actions $(c_1, \ldots, c_n)$ that are chosen by all individuals, according to some given utility function $u_i(c_1, \ldots c_n)$. That is, we are given a game in strategic form. In this case of pure moral hazard, nobody has any private information initially, but a communication process could give individuals different information before they choose their actions. In a general communication system, each individual *i* could get some message m_i in some rich language, with these messages $(m_1, \ldots, m_n)$ being randomly drawn from some joint probability distribution ρ. In any equilibrium of the game generated by adding this communication system, each individual *i* has some strategy σ_i for choosing his action c_i as a function of his message m_i, so that $c_i = \sigma_i(m_i)$.

For the given equilibrium $(\sigma_1, \ldots, \sigma_n)$ of the game with the given communication system ρ, the revelation principle is satisfied by a mediation plan in which the mediator randomly generates recommended actions in such a way that the probability of recommending actions $(c_1, \ldots, c_n)$ is the same as the probability of the given communication system ρ yielding messages $(m_1, \ldots, m_n)$ that would induce the players to choose $(c_1, \ldots, c_n)$ in the σ equilibrium. That is, the probability $\mu(c_1, \ldots, c_n)$ of the mediator recommending $(c_1, \ldots, c_n)$ is

$$\mu(c_1, \ldots, c_n) = \rho(\{(m_1, \ldots, m_n) | \sigma_1(m_1) = c_1, \ldots, \sigma_n(m_n) = c_n\}).$$

Then the mediator confidentially tells each individual i only which action c_i is recommended for him. Obedience is an equilibrium under this mediation plan μ because, if any individual could gain by disobeying this mediator when all others are expected to be obedient, then this individual could have also gained by disobeying himself in implementing his equilibrium strategy σ_i in the given game with communication system ρ. So μ is an incentive-compatible direct-revelation mechanism that is equivalent to the given mechanism ρ with the given equilibrium $(\sigma_1, \ldots, \sigma_n)$.

General formulations

Problems of adverse selection and moral hazard can be combined in the framework of Harsanyi's (1967) Bayesian games, where players have both types and actions. The revelation principle for general Bayesian games was formulated by Myerson (1982; 1985). A further generalization of the revelation principle to multistage games was formulated by Myerson (1986). In each case, the basic idea is that any equilibrium of any general communication system can be simulated by a maximally centralized communication system in which, at every stage, each individual confidentially reports all his private information to a central mediator, and then the mediator confidentially recommends an action to each individual, and the mediator's rule for generating recommendations from reports is designed so that honesty and obedience form an equilibrium of the mediated communication game.

The basic assumption here is that, although the motivations of all economic agents are problematic, we can find a mediator who is completely trustworthy and has no costs of processing information. Asking agents to reveal all relevant information to the trustworthy mediator maximizes the mediator's ability to implement any coordination plan. But telling any other agent more than is necessary to guide his choice of action would only increase the agent's ability to find ways of profitably deviating from the coordination plan.

For honesty and obedience to be an equilibrium, the mediation plan must satisfy incentive constraints which say that no individual could ever expect to gain by deviating to a strategy that involves lying to the mediator or disobeying a recommendation from the mediator. In a dynamic context, we must consider that an individual's most profitable deviation from honesty and obedience could be followed by further deviations in the future. So, to verify that an individual could never gain by lying, we must consider all possible deviation strategies in which the individual may thereafter choose actions that can depend disobediently on the mediator's recommendations (which may convey information about others' types and actions).

When we use sequential equilibrium as the solution concept for dynamic games with communication, the set of actions that can be recommended in a sequentially incentive-compatible mechanism must be restricted somewhat. In a Bayesian game, if some action d_i could never be optimal for individual i to use when his type is t_i, no matter what information he obtained about others' types and actions, then obedience could not be sequentially rational in any mechanism where the mediator might ever recommend this action d_i to i after he reports type t_i. Myerson (1986) identified a larger set of *co-dominated actions* that can never be recommended in any sequentially incentive-compatible mechanism. Suppose that, if any individual observed a zero-probability event, then he could attribute this surprise to a mistake by the trembling hand of the mediator. Under this assumption, Myerson (1986) showed that the effect of requiring sequential rationality in games with communication is completely characterized by the requirement that no individuals should ever be expected to choose any co-dominated actions. (See Gerardi and Myerson, 2007.)

Limitations

The revelation principle says that each equilibrium of any communication mechanism is equivalent to the honest-obedient equilibrium of an incentive-compatible direct-revelation mechanism. But this direct-revelation mechanism may have other dishonest equilibria, which might not correspond to equilibria of the original mechanism. So the revelation principle cannot help us when we are concerned about the whole set of equilibria of a communication mechanism. Similarly, a given communication mechanism may have equilibria that change in some desirable way as we change the players' given beliefs about each others' types, but these different equilibria would correspond to different incentive-compatible mechanisms, and so this desirable property of the given mechanism could not be recognized with the revelation principle.

The assumption that perfectly trustworthy mediators are available is essential to the mathematical simplicity of the incentive-compatible set. Otherwise, if individuals can communicate only by making public statements that are immediately heard by everybody, then the set of equilibria may be smaller and harder to compute.

In principal-agent analysis we often apply the revelation principle to find the incentive-compatible mechanism that is optimal for the principal. If the principal would be tempted to use revealed information opportunistically, then there could be loss of generality in assuming that the agents reveal all their private information to the principal. But we should not confuse the principal with the mediator. The revelation principle can still be applied if the principal can get a trustworthy mediator to take the agents' reports and use them according to any specified mechanism.

There are often questions about whether the allocation selected by a mechanism could be modified by subsequent exchanges among the individuals. An individual's right to offer his possessions for sale at some future date could be accommodated in mechanism design by additional moral-hazard constraints. For example, suppose the principal can sell an object each day, on days 1 and 2. The only buyer's value for such objects is either low $1 or high $3, low having probability 0.25. To maximize the principal's expected revenue with the buyer participating honestly, an optimal mechanism would sell both objects for $3 if the buyer's type is high, but would sell neither if the buyer is low. But if no sale is recommended then the principal could infer that the buyer is low and would prefer to sell for $1. Suppose now that the principal cannot be prevented from offering to sell for $1 on either day. With these additional moral-hazard constraints, an optimal mechanism uses randomization by the mediator to conceal information from the principal. If the buyer reports low then the mediator recommends no sale on day 1 and selling for $1 on day 2. If the buyer reports high, then with probability 1/3 the mediator recommends no sale on day 1 and selling for $3 on day 2, but with probability 2/3 recommends selling for $1.50 on both days. A no-sale recommendation on day 1 implies probability 0.5 of low, so that obedience yields the same expected revenue $0.5 \times (0 + 1) + 0.5 \times (0 + 3)$ as deviating to sell for $1 on both days.

A proliferation of such moral-hazard constraints may greatly complicate the analysis, however. So in practice we often apply the revelation principle with an understanding that we may be overestimating the size of the feasible set, by assuming away some problems of mediator imperfection or moral hazard. When we use the revelation principle to show that a seemingly wasteful mechanism is actually efficient when incentive constraints are recognized, such overestimation of the incentive-feasible set would not weaken the impact of our results (as long as this mechanism remains feasible).

Centralized mediation is emphasized by the revelation principle as a convenient way of characterizing what people can achieve with communication, but this analytical convenience does not imply that centralization is necessarily the best way to coordinate an economy. For fundamental questions about socialist centralization versus free-market decentralization, we should be sceptical about an assumption that centralized control over national resources could not corrupt any mediator. The power of the revelation principle for such questions is instead its ability to provide a common analytical framework that applies equally to socialism and capitalism. For example, a standard result of revelation-principle analysis is that, if only one producer knows the production cost of a good, then efficient incentive-compatible mechanisms must allow this monopolistic producer to take positive informational rents or profits (Baron and Myerson, 1982). Thus the revelation principle can actually be used to support arguments for decentralized multi-source production, by showing that problems of profit-taking by an informational monopolist can be just as serious under socialism as under capitalism.

Nash (1951) advocated a different methodology for analysing communication in games. In Nash's approach, all opportunities for communication should be represented by moves in our extensive model of the dynamic game. Adding such communication moves may greatly increase the number of possible

strategies for a player, because each strategy is a complete plan for choosing the player's moves throughout the game. But if all communication will occur in the implementation of these strategies, then the players' initial choices of their strategies must be independent. Thus, Nash argued, any dynamic game can be normalized to a static strategic-form game, where players choose strategies simultaneously and independently, and Nash equilibrium is the general solution for such games.

With the revelation principle, however, communication opportunities are omitted from the game model and are instead taken into account by using incentive-compatible mechanisms as our solution concept. Characterizing the set of all incentive-compatible mechanisms is often easier than computing the Nash equilibria of a game with communication. Thus, by applying the revelation principle, we can get both a simpler model and a simpler solution concept for games with communication. But, when we use the revelation principle, strategic-form games are no longer sufficient for representing general dynamic games, because normalizing a game model to strategic form would suppress implicit opportunities for communicating during the game (see Myerson, 1986). So the revelation principle should be understood as a methodological alternative to Nash's strategic-form analysis.

See Also

- mechanism design

Bibliography

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96.

Baron, D. and Myerson, R. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50, 911–30.

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules: some results on incentive compatibility. *Review of Economic Studies* 46, 185–216.

Gerardi, D. and Myerson, R. 2007. Sequential equilibrium in Bayesian games with communication. *Games and Economic Behavior* 60, 104–34.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Harris, M. and Townsend, R. 1981. Resource allocation under asymmetric information. *Econometrica* 49, 1477–99.

Harsanyi, J. 1967. Games with incomplete information played by Bayesian players. *Management Science* 14, 159–82, 320–34, 486–502.

Hayek, F. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

Holmstrom, B. 1977. On incentives and control in organizations. Ph.D. thesis, Stanford University.

Myerson, R. 1979. Incentive-compatibility and the bargaining problem. *Econometrica* 47, 61–73.

Myerson, R. 1982. Optimal coordination mechanisms in generalized principal-agent problems. *Journal of Mathematical Economics* 10, 67–81.

Myerson, R. 1985. Bayesian equilibrium and incentive compatibility: an introduction. In *Social Goals and Social Organization*, ed. L. Hurwicz, D. Schmeidler and H. Sonnenschein. Cambridge: Cambridge University Press.

Myerson, R. 1986. Multistage games with communication. *Econometrica* 54, 323–58.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

Rosenthal, R. 1978. Arbitration of two-party disputes under uncertainty. *Review of Economic Studies* 45, 595–604.

How to cite this article

Myerson, Roger B. "revelation principle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000137> doi:10.1057/9780230226203.1436

The New Palgrave Dictionary of Economics Online

reverse capital deepening

Roberto Scazzieri
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Reverse capital deepening is the property whereby it may be efficient to associate a lower (higher) rate of interest with a lower (higher) capital per worker. This property is inconsistent with the traditional belief that, by virtue of the substitution principle, production techniques that are more ‘capital intensive’ will become optimal as the rate of interest is lowered. Reverse capital deepening is an important instance of the apparent paradoxes associated with indirect effects in a production economy. It entails that technical choice cannot be considered a monotonic function of the rate of interest, and questions the widespread policy implications of the traditional view.

Keywords

Böhm-Bawerk, E. von; Clark, J. B.; capital accumulation; capital deepening; capital theory; Hayek, F. A.; reswitching of technique; reverse capital deepening; technical choice; wage–profit relationship

Article

It has long been taken for granted that there is an inverse monotonic relationship between the rate of interest (or the rate of profit) and the quantity of capital per worker. This belief was founded on the principle of substitution, whereby ‘cheaper’ is substituted for ‘more expensive’ as the relative price of two inputs is changed.

In the field of capital theory, the principle of substitution persuaded many economists, such as E. von Böhm-Bawerk (1889), J.B. Clark (1899) and F.A. von Hayek (1941), that a lower rate of interest (which is equal to the rate of profit in equilibrium) is associated with the use of more ‘capital intensive’ techniques, and thus with the substitution of capital for other productive factors, such as labour or land. This process is called capital deepening.

Recent discussions have shown that this is not necessarily true, since a lower rate of interest might be associated with *lower*, rather than higher, capital per worker. This phenomenon is called *reverse capital deepening*.

This discovery was made at the same time as it was realized that it is not generally possible to order ‘efficient’ techniques in such a way that technical choice becomes a monotonic function of the rate of interest (and of the rate of profit).

It can be shown that both reverse capital deepening and reswitching of technique are related to the same fundamental property of the economic system: the possibility (in fact, the near generality) of nonlinear wage–profit relationships. To illustrate this proposition, it is useful to begin by considering the hypothetical case of linear wage–profit relationships (see Figure 1).

Figure 1

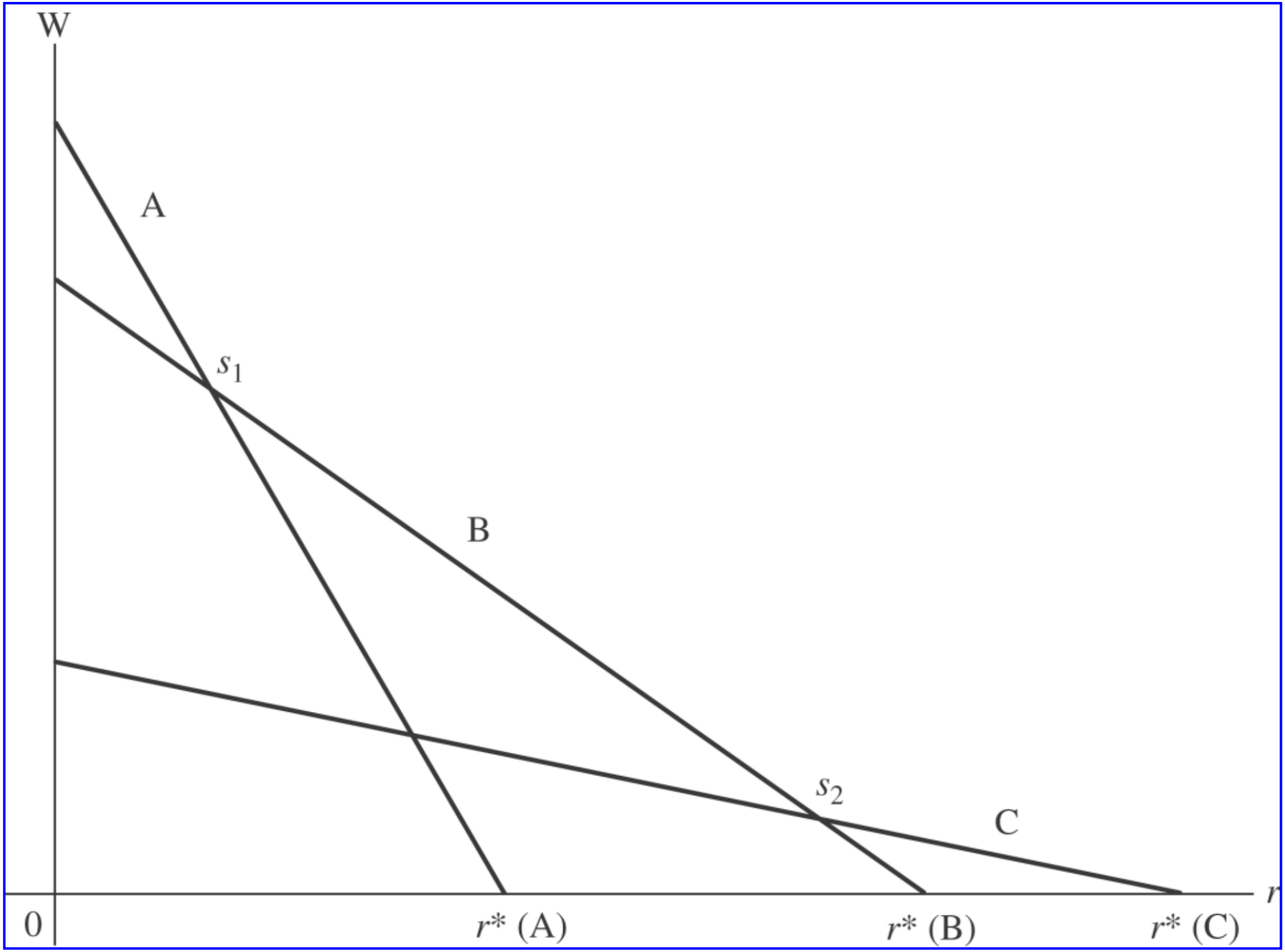

The linearity of the three wage–profit relationships makes reswitching impossible as r increases between 0 and r^*(C) (which is the maximum rate of profit with technology C). The

reason is that no wage–profit line can ever be crossed more than once by another wage–profit line. In this special case, there is an inverse monotonic relationship between the rate of profit and the quantity of capital per worker. This may be shown as follows. We can read the net final output per worker on the *w*-axis of Figure 1 at the point at which r=0. (At that point the net final output per worker coincides with the maximum wage.) The net final output per worker associated with technology A is higher than the net final output per worker associated with technology B. The net final output per worker associated with technology B is higher than the net final output per worker associated with technology C. At switchpoints s_1 and s_2 the wage is the same for both technologies between which substitution takes place. It follows that, at switchpoint s_1, profit per worker is higher with technology A than with technology B. Similarly, at switchpoint s_2, profit per worker is higher with technology B than with technology C. Assuming that the rate of profit is uniform across technologies, we find that, at s_1, A is associated with higher capital per worker than B. A higher rate of profit (or rate of interest) is thus associated with substitution of ‘less capital’ for ‘more capital’. In this particular case, the traditional approach to capital theory would seem to be well founded.

However, these properties disappear altogether once we drop the assumption of linear wage–profit relationships. (It might be interesting to inquire into the economic meaning of straight wage–profit relationships, which are possible only in the case of a technology characterized by a uniform proportion between labour and intermediate inputs in all production processes: only in this case a change in the rate of profit leaves relative prices unaffected.)

But in general wage–profit relationships are of the nonlinear type, which means that the proportion between labour and intermediate inputs is generally different from one production process to another. This feature of the wage–profit frontier makes it possible for wage–profit curves to intersect more than once, thus bringing about the possibility of multiple switching. Under the same circumstances it can be shown that the relationship between the rate of profit and capital per worker is no longer of the inverse monotonic type. This can be seen in the reswitching case (Figure 2), but it can also be seen in the case in which the wage–profit curves never intersect more than once on the efficiency frontier (Figure 3). (See also Pasinetti, 1966.)

Figure 2

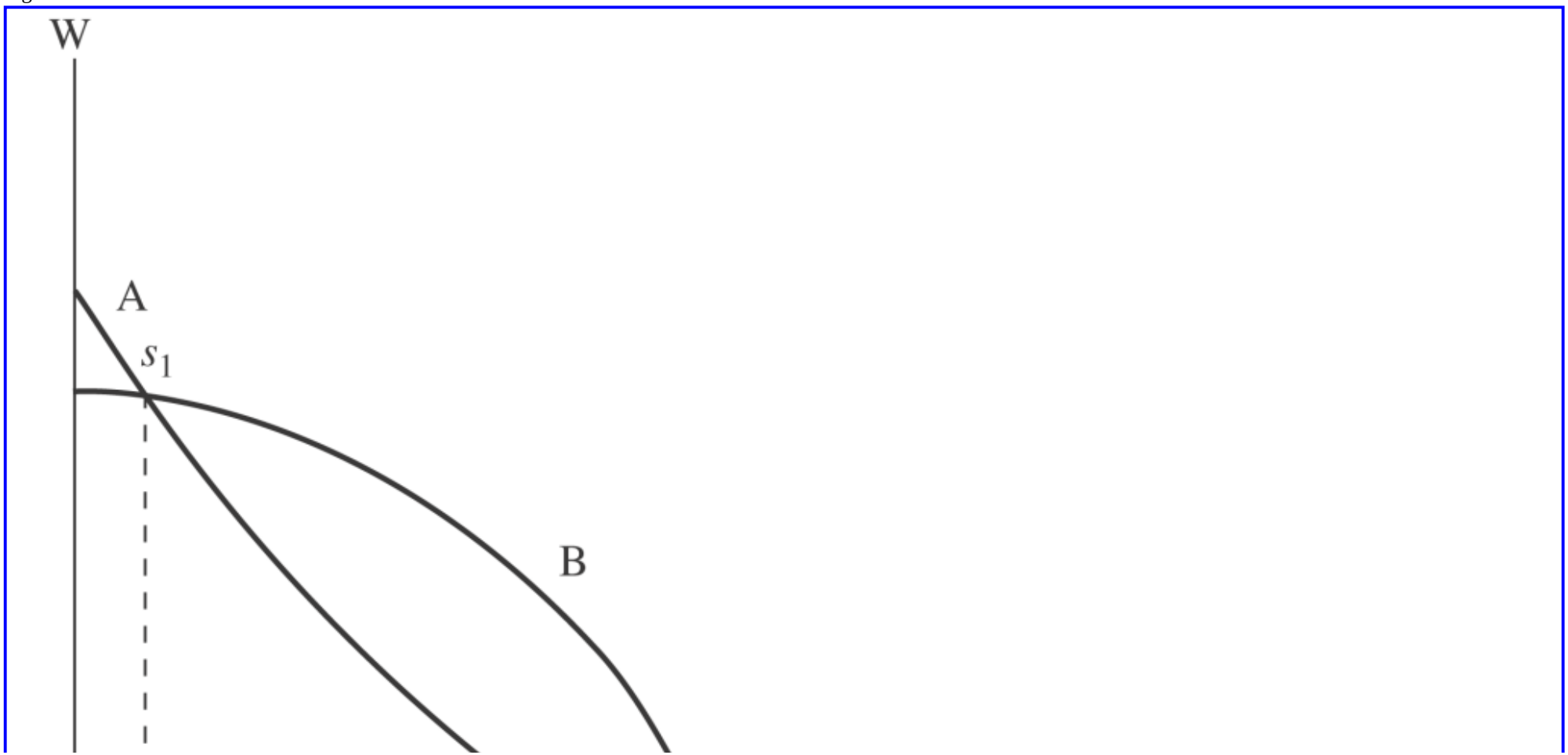

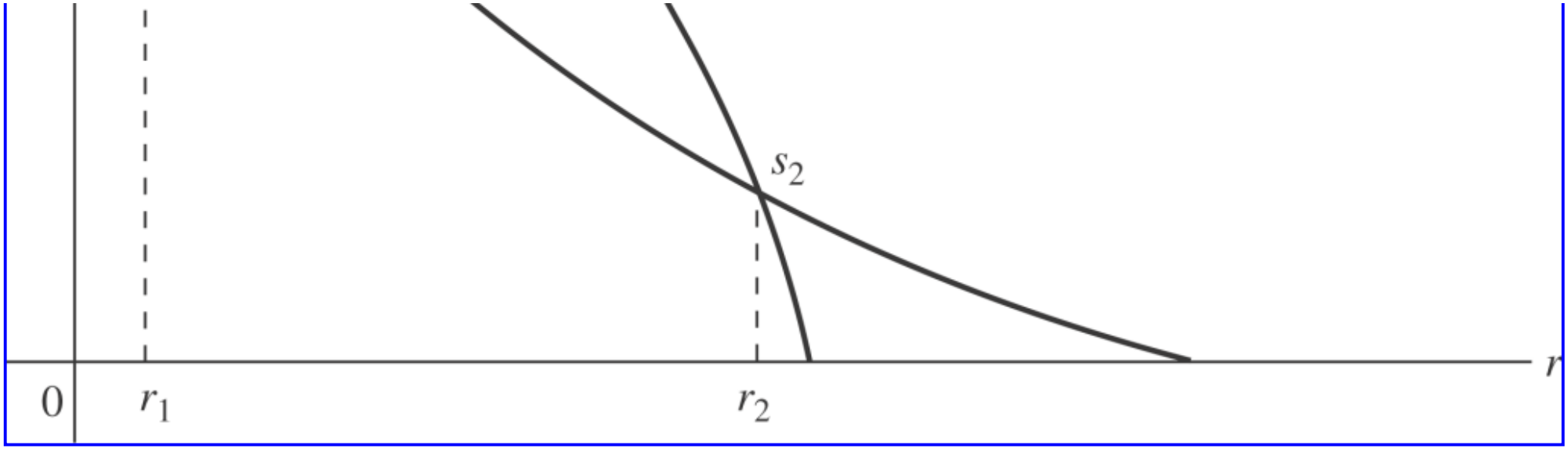

Figure 3

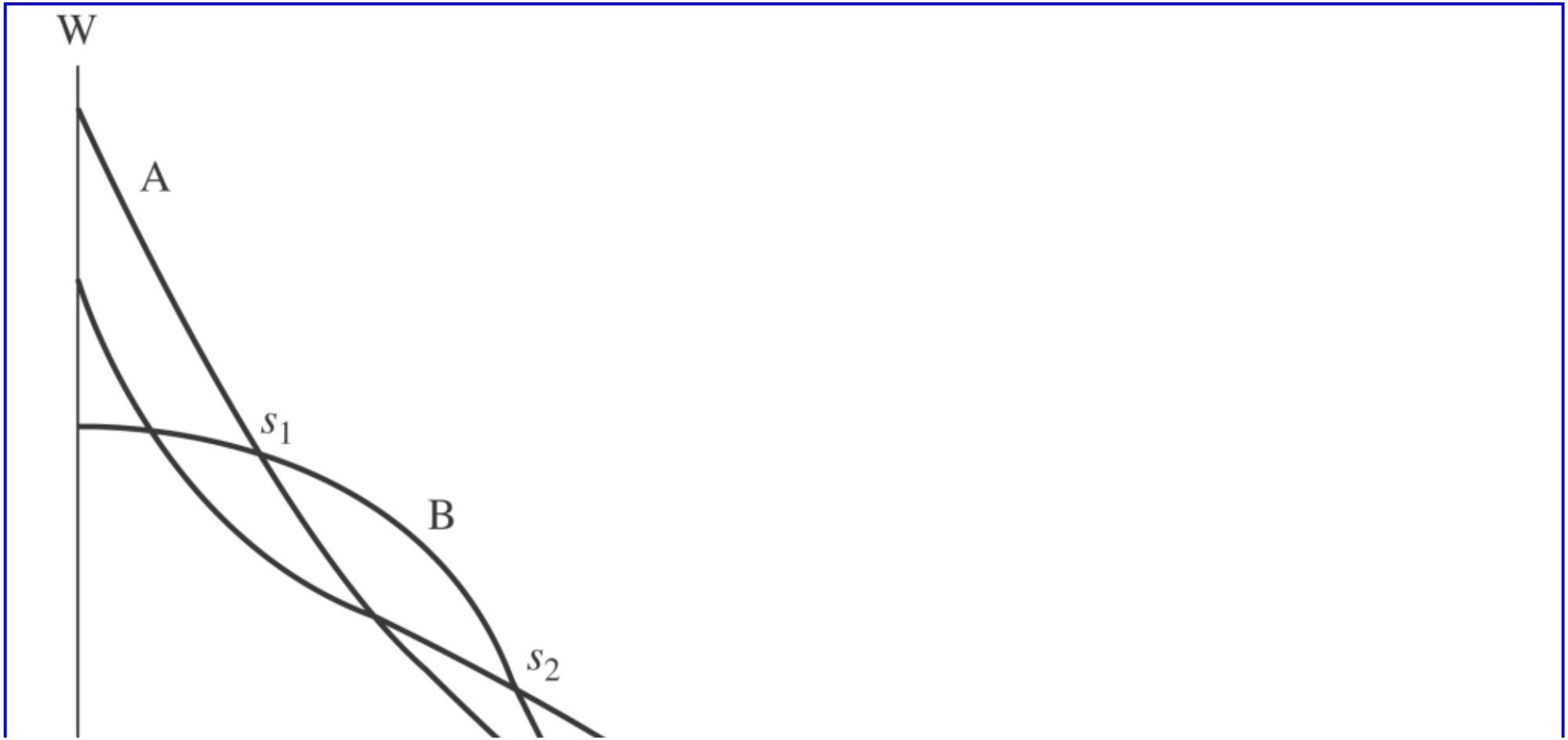

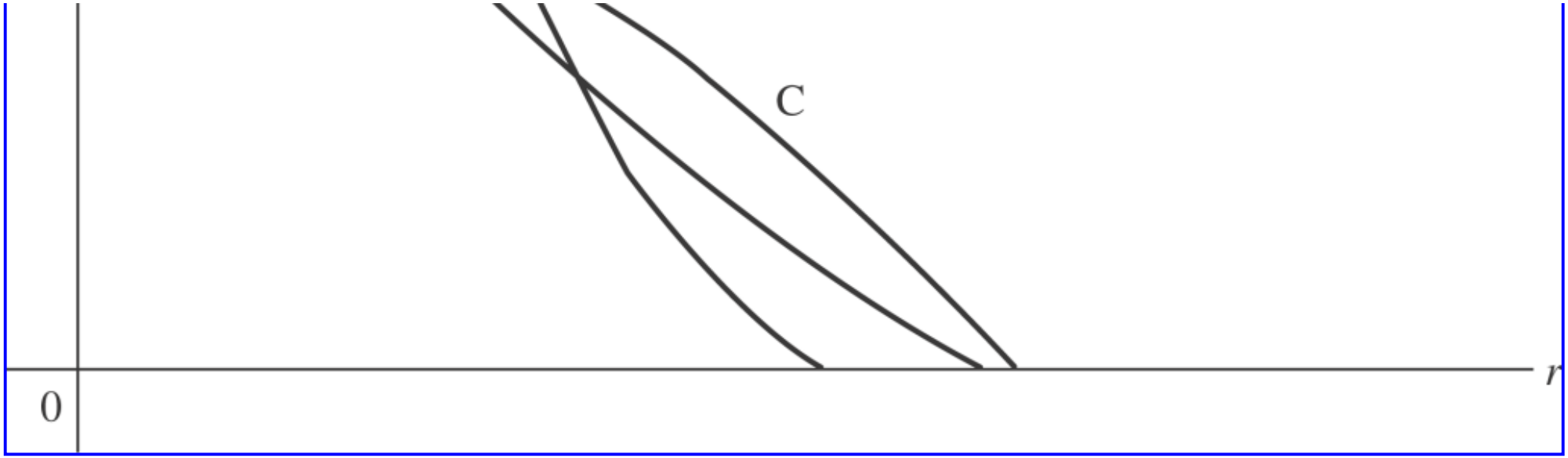

In Figure 2, reswitching is associated with reverse capital deepening. Technology A is the more profitable at levels of the rate of profit lower than r_1, it is ‹overtaken› by technology B at rates of profit between r_1 and r_2, it becomes again the more profitable at rates of profit higher than r_2. At the same time, switchpoint s_1 is associated with the substitution of the technology with lower value of capital per worker (B) for the technology with the higher value of capital per worker (A), whereas at switchpoint s_2 the opposite happens: the technology with higher capital per worker (A) is substituted for the technology with lower capital per worker (B), in spite of the fact that the rate of profit is higher (reverse capital deepening).

In Figure 3, there is no reswitching but we still have a reverse capital deepening. For no wage–profit curves cross one another more than once on the efficiency frontier, but at switchpoint s_2 an increasing rate of profit is associated with the substitution of a technology with higher capital per worker (C) for a technology with lower capital per worker (B).

Complementarity in production is often at the root of apparently perverse price behaviour (see Broome, 1978). One reason for this had been noted by John Hicks, when he wrote that the ‘net effect’ of a change in the price of productive factor *x* upon the price of complementary factor *y* ‘is … compounded out of two contrary tendencies, a direct effect tending to raise it, and indirect effect tending to reduce it; either may be dominant’ (Hicks, 1946, p. 107). Reverse capital deepening is an especially important instance of a widespread phenomenon associated with indirect effects in a production economy. This possibility is associated with other phenomena which are not compatible with traditional beliefs about capital and capital accumulation. Simple inspection of Figure 2 or 3 shows that at a switchpoint associated with reverse capital deepening (s_2 in either figure) a technology with higher capital per worker and higher net final product per worker is substituted for a technology with lower capital per worker and lower net final product per worker. At such switchpoints a higher rate of profit (and rate of interest) could be associated with a higher ratio of capital per worker to net final product per worker, that is, with a higher capital–output ratio.

Figures 2 and 3 also alert us as to the possibility that a technology adopted at a high rate of interest is associated with higher maximum consumption per head than a technology adopted at a lower rate of interest. In addition, transition to a lower rate of interest may involve the switch to a lower maximum consumption per head. (This can be seen at switchpoint s_2 in either figure, where maximum consumption per head can be read on the *w*-axis at point r=0.) This behaviour of consumption per head in relation to the rate of interest is clearly incompatible with the view that a higher rate of interest brings about a special type of exchange, in which less consumption in the current period is substituted against higher consumption in the future. Reverse capital deepening alerts us as to the possibility that a higher rate of interest might be associated with greater current consumption per head than the consumption per head feasible with the technology adopted at a lower rate of interest (see Bruno, Burmeister and Sheshinski, 1966; Samuelson, 1966).

The relevance of reverse capital deepening is that the foundation of the traditional view that technical choice is a monotonic function of the rate of interest is seriously questioned. Similarly, the widespread policy implications of the traditional view are also questioned. However, there is as yet no full agreement as to the main consequences of this result. For example, Christopher Bliss (1975, p. 279) has noted that reverse capital deepening makes it impossible to see the accumulation of capital as a process associated with ‘a continuous increase in consumption per capita,…a continuous decline in the rate of interest and … a continuous increase in the real wage rate’. He also called attention to the fact that the ‘extended accumulation history’ of an economic system moving through real time is normally different from the hypothetical history we can tell by ‘travelling’ across steady states (1975, pp. 194, 280–1). In particular, he emphasized that the statement that the rate of interest may be expected to fall as capital deepening takes place ‘cannot be interpreted’ in the case of extended accumulation history, as we would have, in that case, ‘a whole structure of interest rates … not a single rate of interest’ (1975, p. 294). A different point of view has

been expressed by Pierangelo Garegnani, who has maintained that capital paradoxes in general, and reverse capital deepening in particular, by making traditional beliefs untenable, suggest a ‘correction’ to traditional theory, which would make it reasonable to expect ‘instabilities or tendencies to zero of wages, or of net returns on capital’ (Garegnani, 1990, p. 76). This author's view is that, rather than introducing such a correction, one should drop any idea of a causal connection from marginal products to the distribution of the social product, and further develop the conjecture that distribution is brought about by ‘more complex economic and social forces like those envisaged by the old classical economists’ (Garegnani, 1990, pp. 76–7). As it is common in theoretical sciences (see Kuhn, 1970; 2000), the discovery of an apparent anomaly has induced economists to look for a more general theory, or to switch to an altogether different framework.

See Also

- capital theory
- capital theory (paradoxes)
- production functions
- reswitching of technique

Bibliography

Bliss, C.J. 1975. *Capital Theory and the Distribution of Income*. Amsterdam, Oxford, New York: North-Holland Publishing Company and American Elsevier Publishing Company.

Böhm-Bawerk, E. von. 1889. *Positive Theorie des Kapitales*. Zweite Ableitung: *Kapital und Kapitalzins*. Trans. as *The Positive Theory of Capital*, London: Macmillan, 1891.

Broome, J. 1978. Perverse prices. *Economic Journal* 88, 778–87.

Bruno, M., Burmeister, E. and Sheshinski, E. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80, 526–53.

Clark, J.B. 1899. *The Distribution of Wealth*. New York: Macmillan.

Garegnani, P. 1990. Quantity of capital. In *The New Palgrave: Capital Theory*, ed. J. Eatwell et al. London and Basingstoke: Macmillan.

Hayek, F.A. von. 1941. *The Pure Theory of Capital*. London: Routledge.

Hicks, J. 1946. *Value and Capital*, 2nd edn. Oxford: Clarendon Press.

Kuhn, T.S. 1970. *The Structure of Scientific Revolutions*, 2nd edn enlarged. Chicago: University of Chicago Press.

Kuhn, T.S. 2000. *The Road Since Structure: Philosophical Essays, 1970–1993*, with an autobiographical interview, ed. J. Conant and J. Haugeland. Chicago and London: Chicago University Press.

Pasinetti, L.L. 1966. Changes in the rate of profit and switches of techniques. *Quarterly Journal of Economics* 80, 503–17.

Samuelson, P. 1966. A summing up. *Quarterly Journal of Economics* 80, 568–83.

How to cite this article

Scazzieri, Roberto. "reverse capital deepening." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000139> doi:10.1057/9780230226203.1437

The New Palgrave Dictionary of Economics Online

rhetoric of economics

Donald N. McCloskey

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

rhetoric; rhetoric of economics

Article

Rhetoric is the study and practice of persuasive expression, an alternative since the Greeks to the philosophical programme of epistemology. The rhetoric of economics examines how economists persuade – not how they say they do, or how their official methodologies say they do, but how in fact they persuade colleagues and politicians and students to accept one economic assertion and reject another.

Some of their devices arise from bad motives, and bad rhetoric is what most people have in mind when they call a piece of writing ‘rhetorical’. An irrelevant and inaccurate attack on Milton Friedman's politics while criticizing his economics would be an example, as would a pointless and confusing use of mathematics while arguing a point in labour economics. The badness does not reside in the techniques themselves (political commentary or mathematical argument) but in the person using them, since all techniques can be abused. Aristotle noted that ‘if it be objected that one who uses such power of speech unjustly might do great harm, *that* is a charge which may be made in common against all good things except virtue itself’. Cato the Elder demanded that the user of analogy (or in our time the user of regression) be *vir bonus dicendi peritus*, the *good* man skilled at speaking. The protection against bad science is good scientists, not good methodology.

Rhetoric, then, can be good, offering good reasons for believing that the elasticity of substitution between capital and labour in American manufacturing, say, is about 1.0. The good reasons are not confined by syllogism and number. They include good analogy (production is *just like* a mathematical function), good authority (Knut Wicksell and Paul Douglas thought this way, too), good symmetry (if mining can be treated as a production function, so should manufacturing). Furthermore, the reasonings of syllogism and number are themselves rhetorical, that is, persuasive acts of human speech. An econometric test will depend on how apt is an analogy of the error term with drawings from an urn. A mathematical proof will depend on how convincing is an appeal to the authority of the Bourbaki style.

‘The facts’ and ‘the logic’ matter, of course; but they are part of the rhetoric, depending themselves on the giving of good reasons.

Consider, for example, the sentence in economics, ‘The demand curve slopes down’. The official rhetoric says that economists believe this because of statistical evidence – negative coefficients in demand curves for pig iron or negative diagonal items in matrices of complete systems of demand – accumulating steadily in journal articles. These are the tests ‘consistent with the hypothesis’. Yet most belief in the hypothesis comes from other sources: from introspection (what would I do?); from thought experiments (what would they do?); from uncontrolled cases in point (such as the oil crisis); from authority (Alfred Marshall believed it); from symmetry (a law of demand if there is a law of supply); from definition (a higher price leaves less for expenditure, including this one); and above all, from analogy (if the demand curve slopes down for chewing gum, why not for housing and love too?). As may be seen in the classroom and seminar, the range of argument in economics is wider than the official rhetoric allows.

The rhetoric of economics brings the traditions of rhetoric to the study of economic texts, whether mathematical or verbal texts. It is a literary criticism of economics, or a jurisprudence, and from literary critics like Wayne Booth (1974) and lawyers such as Chaim Perelman (Perelman and Olbrechts-Tyteca, 1958) much can be learned. Although its precursors in economics are methodological criticisms of the field (such as Frank Knight, 1940), censorious joking (such as Stigler, 1977), and finger-wagging presidential addresses (such as Leontief, 1971, or Mayer, 1980), the main focus of the work has been the analysis of how economists seek to persuade, whether good or bad (Klamer, 1984; Henderson 1982; Kornai, 1983; McCloskey, 1986). Econometrics has its own rhetorical prehistory, more self-conscious than the rest (Leamer, 1978), reaching back to the founders of decision theory and Bayesian statistics. The movement has parallels in other fields. Imre Lakatos (1976), Davis and Hersh (1981), and others have uncovered a rhetoric in mathematics; Rorty (1982), Toulmin (1958), and Rosen (1980) in technical philosophy; and numbers of scientists in their own fields (Polanyi, 1962; Medawar, 1964). Historians and sociologists of science have since the 1960s accumulated much evidence that science is a conversation rather than a mechanical procedure (Kuhn, 1977; Collins, 1985). The analysis of conversation from scholars in communication and literary studies (Scott, 1967) has provided ways of rereading various fields (a sampling of these is contained in Nelson, Megill and McCloskey, 1987).

A rhetoric of economics questions the division between scientific and humanistic reasoning, not to attack quantification or to introduce irrationality into science, but to make the scientific conversation more aware of itself. It is a programme of greater, not less, rigour and relevance, of higher, not lower, standards in the conversations of mankind.

See Also

- philosophy and economics

Bibliography

Booth, W. 1974. *Modern Dogma and the Rhetoric of Assent*. Chicago: University of Chicago Press.

Collins, H.M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. London: Sage.

Davis, P.J. and Hersh, R. 1981. *The Mathematical Experience*. Boston: Houghton Mifflin.

Henderson, W. 1982. Metaphors in economics. *Economics* 18(4), 147–53.

Klamer, A. 1984. *Conversations with Economists: New Classical Economists and Opponents Speak Out on the Current Controversy in Macroeconomics*. Totowa, NJ: Rowman and Allanheld.

Knight, F. 1940. ‘What is truth’ in economics? *Journal of Political Economy* 48, 1–32.

Kornai, J. 1983. The health of nations: reflections on the analogy between medical science and economics. *Kyklos* 36(2), 191–212.

Kuhn, T. 1977. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Lakatos, I. 1976. *Proofs and Refutations: the Logic of Mathematical Discovery*. Cambridge: Cambridge University Press.

Leamer, E. 1978. *Specification Searches: Ad Hoc Inferences with Nonexperimental Data*. New York: Wiley.

Leontief, W. 1971. Theoretical assumptions and nonobserved facts. *American Economic Review* 61, 1–7.

Mayer, T. 1980. Economics as a hard science: realistic goal or wishful thinking? *Economic Inquiry* 18, 165–78.

Medawar, P. 1964. Is the scientific paper fraudulent? *Saturday Review*, 1 August.

McCloskey, D.N. 1986. *The Rhetoric of Economics*. Madison: University of Wisconsin Press.

Nelson, J., Megill, A. and McCloskey, D.N. eds. 1987. *The Rhetoric of the Human Sciences: Papers and Proceedings of the Iowa Conference*. Madison: University of Wisconsin Press.

Perelman, C. and Olbrechts-Tyteca, L. 1958. *The New Rhetoric: A Treatise on Argumentation*. Notre Dame, IN: University of Notre Dame Press.

Polanyi, M. 1962. *Personal Knowledge: Towards a Post-critical Philosophy*. Chicago: University of Chicago Press.

Rorty, R. 1982. *The Consequences of Pragmatism: Essays*. Minneapolis: University of Minnesota Press.

Rosen, S. 1980. *The Limits of Analysis*. New York: Basic Books.

Scott, R. 1967. On viewing rhetoric as epistemic. *Central States Speech Journal* 18(1), 9–17.

Stigler, G.J. 1977. The conference handbook. *Journal of Political Economy* 85, 441–3.

Toulmin, S. 1958. *The Uses of Argument*. Cambridge: Cambridge University Press.

How to cite this article

McCloskey, Donald N. "rhetoric of economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000142> doi:10.1057/9780230226203.1438

The New Palgrave Dictionary of Economics Online

Ricardian equivalence theorem

Andrew B. Abel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Ricardian equivalence theorem states that government bonds and lump-sum taxes are equivalent means to finance government spending. Thus, a lump-sum tax cut financed by the issuance of government bonds would not affect consumption. Consumers could hold the newly issued bonds, and use them to pay the higher taxes when the government increases taxes to repay the principal and interest on the bonds. Intergenerational altruism implies that Ricardian equivalence holds even if the recipients of a tax cut die before future taxes are increased to fully repay the bonds. This article explores situations where Ricardian equivalence does or does not hold.

Keywords

adverse selection; altruism; annuities; bequest motive; bequests; bonds; capital accumulation; consumption function; gift motive; insurance markets; intergenerational altruism; life insurance; liquidity constraints; lump-sum taxes; mortality; national income determination; pensions; precautionary savings; public debt; Ricardian equivalence theorem; taxation; wealth

Article

The Ricardian equivalence theorem is the proposition that the method of financing any particular path of government expenditure is irrelevant. More precisely, whether government purchases are financed by levying lump-sum taxes or by issuing government bonds does not affect the consumption of any household, nor does it affect capital formation. In this sense, financing government purchases by lump-sum taxes is *equivalent* to financing these purchases by issuing bonds. The fundamental logic underlying this proposition was presented by David Ricardo in Chapter XVII (‘Taxes on Other Commodities than Raw Produce’) of *The Principles of Political Economy and Taxation*. Although Ricardo clearly explained why government borrowing and taxes could be equivalent, he warned against accepting the argument on its face: ‘From what I have said, it must not be inferred that I consider the system of

borrowing as the best calculated to defray the extraordinary expenses of the state. It is a system which tends to make us less thrifty – to blind us to our real situation' (1821, pp. 162–3).

The question of whether lump-sum taxes and government debt are equivalent arises in the specification of the consumption function. The aggregate consumption function plays an important role in models of national income determination, and aggregate consumption is often specified to depend on contemporaneous aggregate disposable income and on aggregate wealth. The question is whether the public's holding of bonds issued by the government should be treated as part of aggregate net wealth. Indeed, this is the eponymous question of Barro's (1974) classic article on Ricardian equivalence. If consumers recognize that government bonds, in the aggregate, represent future tax liabilities, then these bonds would not be part of aggregate wealth. If, on the other hand, consumers do not recognize, or for some reason do not care about, the implied future tax liabilities associated with these bonds, then they should be counted as part of aggregate wealth in an aggregate consumption function. Patinkin (1965, p. 289), citing Carl Christ and Christ's discussions with Milton Friedman, recognized this question and specified that a fraction k of the stock of outstanding government bonds is to be treated as wealth. Under the Ricardian equivalence view, k would equal zero; under the view that consumers ignore all future tax liabilities, k would equal 1. Bailey (1971) also examined the question of whether future tax liabilities affect aggregate consumption in a model of national income determination, though his formulation of the aggregate consumption function does not explicitly include aggregate wealth.

The question of whether government bonds are net wealth and the question of the effects of alternative means of financing a given amount of government expenditure are, in many contexts, basically the same question. For purposes of exposition, it is clearest to focus on one particular formulation of the question. The discussion here will focus on the question of the choice between current taxation and debt finance.

The underlying logic of the Ricardian equivalence theorem is quite simple and can be displayed by considering a reduction in current lump-sum taxes of 100 dollars per capita. This reduction in government tax revenue is financed by the sale of government bonds on the open market in the amount of 100 dollars per capita. For simplicity, suppose that the bonds are one-year bonds with an interest rate of five per cent per year. In addition, suppose that the population of taxpayers is constant over time. In the year following the tax cut, the bonds are redeemed by the government. In order to pay the principal and interest on the bonds, taxes must be increased by 105 dollars per taxpayer in the second year.

Now consider the response of households to this intertemporal rearrangement of their tax liabilities. Households can afford to maintain their originally planned current and future consumption by increasing their current saving by 100 dollars. In fact, the additional 100 dollars of private saving could be held in the form of newly issued government bonds. In the second year, when the government increases taxes by 105 dollars to redeem the bonds, households pay the extra tax using the principal and interest on the bonds. Thus, the originally planned path of consumption continues to be feasible after the tax change. In addition, since the originally planned path of consumption was chosen by the consumer before the tax change, it would continue to be chosen after the tax change since all relative prices remain unchanged. Therefore, household behaviour is invariant to the switch between tax finance and debt finance for a given amount of government spending.

In the basic example, the tax cut in the current year is financed by the issuance of one-year government bonds. However, the invariance result continues to hold if the current tax cut is financed by the issuance of N-year bonds. The argument is that once again each consumer uses the extra 100 dollars of disposable income in the first year to purchase 100 dollars of newly issued government bonds. If these government

bonds pay interest in years before the bond is redeemed, then the government must increase lump-sum taxes in those years to service the bonds. Consumers who are holding the bonds and receive interest use the interest on their bonds to pay the increased taxes. Then, when the bonds mature after *N* years, each consumer uses the principal and final interest on these bonds to pay the higher taxes that are levied to redeem the debt. Once again, consumers can afford to maintain the originally planned path of current and future consumption and find it optimal to do so.

Having seen that the Ricardian equivalence theorem holds even if long-term bonds are issued to pay for the current tax cut, it is natural to ask whether the invariance result continues to hold even if some or all of the currently living consumers die before the bonds are redeemed. The first answer to this question would appear to be that consumers who are alive during the tax cut, but who die before the newly issued bonds are retired, would have a reduction in the present value of their taxes and thus an increase in the present value of their disposable income. Equivalently, such consumers could afford to increase their current and future consumption. It is not necessary for these consumers to hold the extra bond that is issued in the first year because they will not have to use the bonds to pay for the future tax increase needed to redeem the bonds. Therefore, these consumers would tend to increase their current and future consumption, *ceteris paribus*.

A self-interested consumer who receives a tax cut financed by government bonds will increase his consumption if he knows with certainty that he will die before future taxes are collected to fully repay the newly issued bonds. But if the consumer is uncertain about when he will die, the situation involves some additional considerations. I begin by ignoring survival-contingent assets such as annuities and life insurance, and I will assume that all consumers have positive net financial assets so that I can put aside issues related to borrowing costs for consumers who may die before repaying their loans. To keep the argument simple, suppose that lump-sum taxes are reduced in the current year by 100 dollars per taxpayer, and the government finances the tax cut by issuing 20-year zero-coupon bonds. Twenty years in the future the government will increase lump-sum taxes to pay off these bonds. The present value of the future tax increase is 100 dollars per current taxpayer. If the number of taxpayers 20 years in the future is the same as in the current year, then tax increase in the future will be $100(1+r)^{20}$ per taxpayer, where *r* is the annual interest rate on the government bonds. In this case, the current tax cut will not affect the current consumption of any taxpayer. A current taxpayer could use the 100 dollars from the current tax cut to buy 100 dollars of government bonds, and simply plan to hold on to the bonds for 20 years. In the event that the consumer is still alive and consuming 20 years in the future, he can use his bonds, which will have grown in value to $100(1+r)^{20}$ to pay the additional lump-sum taxes in that year, without changing consumption in that year (or in the current year). In the event that the consumer dies before 20 years elapse, he will, of course, consume the same level, namely zero, as in the absence of the tax cut. Thus, buying and holding 100 dollars of government bonds in the current year just allows the consumer to maintain consumption unchanged at all ages and in all states (that is, the state in which the consumer is alive in 20 years and the state in which he is not alive in 20 years). Therefore, Ricardian equivalence holds in this case.

The example in the preceding paragraph illustrates that Ricardian equivalence can hold even when selfish consumers receive a bond-financed tax cut and, with some unpredictability, die before the taxes are levied to fully repay the bonds. A crucial step in the argument is the assumption that the number of future taxpayers is the same as the number of current taxpayers. But with some taxpayers dying over

time, the only way to maintain a constant number of taxpayers is for new taxpayers to arrive – through birth or immigration – at the same rate at which taxpayers are dying. In an economy with a growing population, that is, in an economy in which the sum of the birth rate and net immigration rate exceeds the death rate, the increase in future lump-sum taxes *per taxpayer* will be smaller than $100(1+r)^{20}$, because the cost of paying off government bonds is spread among a larger number of taxpayers. Therefore, a consumer who receives a tax cut of 100 dollars in the current year will face a future tax increase that has a present value smaller than 100 dollars, and so will increase consumption in the current period (and in the future period, if he is alive). Alternatively, in an economy with a shrinking population, a lump-sum tax increase will have the opposite effect and will reduce current consumption. (An analytic version of this example is in Abel, 1989.)

Ricardian equivalence is often illustrated in the context of perfect markets. If consumers face uncertainty about the length of their lives, and if they do not have bequest motives of any sort, they will want to hold annuities, which are assets that pay off if the owner of the annuity is alive, but pay zero if the owner is not alive. If all consumers face the same publicly known probability, *p*, of dying each year, then the actuarially fair annual gross rate of return on annuities will be $(1+r)/(1-p)$. If all consumers invest one dollar in an annuity that pays a lump sum in 20 years, then in 20 years each survivor will receive $[(1+r)/(1-p)]^{20}$ dollars. Whether consumers who receive a 100 dollar lump-sum tax cut in the current year will change their current consumption depends on the amount of the tax increase *per taxpayer* 20 years in the future when the bonds used to finance the tax cut are paid off. If the birth rate and the net immigration rate are both zero, then the population of taxpayers in 20 years will be a fraction $(1-p)^{20}$ of the population in the current year. Thus, to repay the principal and interest on the 100 dollars of bonds issued per current taxpayer, the lump-sum tax will have to increase by $[(1+r)/(1-p)]^{20}$ dollars per current taxpayer. Thus, a current taxpayer could use the 100 dollar tax cut in the current period to purchase 100 dollars of annuities in the current period. If the consumer survives for 20 years, the payoff of the annuity, $[(1+r)/(1-p)]^{20}$, will be just sufficient to pay the increased lump-sum tax in that year. Thus, Ricardian equivalence holds in this case with perfect annuities and a zero birth rate and zero net immigration rate. Ricardian equivalence will fail to hold, however, if the birth rate is positive, because the tax burden in 20 years will be spread among a group of taxpayers consisting of surviving taxpayers from the current period plus additional taxpayers. In this case, the tax increase per future taxpayer will be smaller than $[(1+r)/(1-p)]^{20}$. Thus, recipients of the tax cut in the current period would be able to increase consumption in the current period somewhat and use the remainder of the tax cut to buy enough annuities to pay the increased lump-sum tax and to increase consumption in 20 years.

The examples with uncertain longevity illustrate that, as emphasized by Weil (1989), the departure from Ricardian equivalence does not result solely from the chance of dying before the future tax increase. In the examples in which the tax cut in the current year induces an increase in current consumption, the effect results from the fact that future increase in taxes per future taxpayer is smaller than the current tax cut per taxpayer. In the case of perfect annuities, this effect is made possible by a positive birth rate, and in the case without annuities the effect was made possible by growing population resulting from a death rate lower than the birth rate plus the net immigration rate.

Altruistic consumers

If consumers are entirely self-interested, then escaping future taxes through death can lead to departures from the Ricardian equivalence theorem, as discussed above. However, Robert Barro (1974) presented an ingenious argument that extends the Ricardian equivalence theorem to situations in which consumers die before future taxes are increased to repay the bonds that are issued to finance the current tax cut. Before discussing the substantive content of Barro's argument, it is interesting to observe that the term ‘Ricardian equivalence theorem’ apparently was first used by James Buchanan (1976) in a published comment on Barro's paper. Buchanan's comment begins by pointing out Barro's failure to credit Ricardo with the idea that debt and taxes may be equivalent and, indeed, the comment is titled, ‘Barro on the Ricardian Equivalence Theorem’. Previously, Buchanan had referred to this result as the ‘equivalence hypothesis’ (1958, p. 118).

Barro postulated that consumers have bequest motives of a particular form that has been labelled ‘altruistic’. An altruistic consumer obtains utility from his own consumption as well as from the utility of his children. Therefore, a consumer who is altruistic toward all of his children cares not only about his own consumption but also indirectly about the consumption of all his children. Furthermore, if all of the altruistic consumer's children are also altruistic and care about the utility of all of their children, then the altruistic consumer cares indirectly about the consumption of all of his grandchildren. Provided that all consumers are altruistic, the argument can be extended ad infinitum with the important implication that an altruistic consumer cares, at least indirectly, about the entire path of current and future consumption of himself and all of his descendants.

Barro's insight that an intergenerationally altruistic consumer cares about the entire path of his family's consumption defuses the argument that consumers who know they will escape future taxes through death will increase consumption in response to a current tax cut. For altruistic consumers, it does not matter whether they themselves or their descendants pay the higher taxes necessary to pay the principal and interest on the newly issued bonds. In response to a 100 dollar tax cut in the current year, an altruistic consumer will not change his consumption but will hold an additional 100 dollars of government bonds. If the bonds are not redeemed until after the consumer dies, he will bequeath them to his children who can then use the bonds to pay the higher taxes in the year in which the bonds are redeemed, or else bequeath the bonds to their children if the bonds are not redeemed during their lifetimes.

The fact that a consumer leaves a bequest is not prima facie evidence that he is altruistic in the sense defined above. Bequests may arise as the accidental outcome of an untimely death or they may arise for motives other than altruism. For instance, if the utility that a consumer obtains from leaving a bequest depends only on the size of the bequest, then he will not care about tax increases that may be levied on his children or his children's children. In this case Ricardian equivalence would not hold.

The argument that each current and future consumer in a family of intergenerationally altruistic consumers cares about his own consumption as well as the consumption of all of his descendants for ever raises the question of whether the government must ever pay off the newly issued government bonds. If the government could roll over the principal and interest on this debt for ever, so that it would never be necessary to increase future taxes, it would seem that a current tax cut financed by issuance of government bonds would reduce the present value of the taxes paid by the current and future members of the family and hence would lead to an increase in the family's consumption. If the government

attempted to roll over its debt each year by issuing new bonds, the quantity of these bonds would grow in perpetuity at the rate of interest. If the rate of interest exceeds the economy's growth rate, then these bonds would not willingly be held in private portfolios. Alternatively, if the rate of interest falls short of the economy's growth rate – a condition that signals an inefficient over-accumulation to capital – then, as pointed out by Feldstein (1976), it is possible for the government to roll over the debt permanently.

Carmichael (1982) has shown that in this case the altruistic bequest motive will not be operative (that is, the non-negativity constraint on bequests will bind) but that an altruistic gift motive from children to parents (which specifies that a consumer's utility depends on his own consumption and the utility of his parents) may be operative. If the gift motive is operative, then Carmichael argues that Ricardian equivalence will hold, despite the fact that government bonds may be regarded as net wealth.

Departures from Ricardian equivalence

Now that we have described a fairly general set of conditions under which Ricardian equivalence holds, it is useful to discuss several of the conditions that might lead to a violation of Ricardian equivalence. A clear overview of reasons why the Ricardian equivalence theorem may not provide an accurate description of the actual effects of debt finance vs. tax finance is provided by Tobin (1980).

The basic argument underlying the Ricardian equivalence theorem is that it makes no difference whether the government issues debt in the amount of 100 dollars per capita or whether it collects taxes of 100 dollars per capita since in the latter case consumers can borrow 100 dollars per capita to pay the higher taxes. In the former case, public borrowing is increased by 100 dollars per capita, and in the latter case private borrowing is increased by 100 dollars per capita. Under the appropriate conditions it makes no difference whether the borrowing is by the public sector or by the private sector. In order for the choice between debt finance and tax finance to have an effect, it must be the case that any changes in government borrowing cannot be fully offset by changes in private sector behaviour. Equivalently, there must be something that the government can do in credit markets that the private sector cannot do.

The government can borrow by issuing bonds, but in some situations consumers may not be able to borrow. For instance, a young consumer with a high prospective income might like to borrow to increase his consumption when young with the intention of repaying the loan when his income is higher in the future. However, for a variety of reasons it may simply not be possible for the young consumer to borrow the desired amount; if this is the case, the consumer is described as ‘liquidity-constrained’. A liquidity-constrained consumer who receives a tax cut in the current period may choose to consume some, or even all, of the tax cut rather than save the entire tax cut. In effect, the current tax cut allows the consumer to borrow in order to increase current consumption, which is what the liquidity-constrained consumer wanted to do anyway. The current tax cut financed by an issue of government bonds can be viewed as the government borrowing on behalf of the consumer. Although this example makes it seem clear that a liquidity-constrained consumer would increase his current consumption in response to a current tax cut, some caution is required in interpreting this result. Unless the reason for the liquidity constraint is specified, one cannot determine what will be the effect of the tax cut. For example, suppose that a consumer is able to borrow some funds, but is liquidity-constrained in the sense that he would like to borrow even more funds. If his creditors determine how much they are willing to lend by looking at his ability to repay the loan, then, in response to the prospective tax increase accompanying the current tax cut, his lenders may reduce the amount they are willing to lend by the

amount of the tax cut. In this case, Ricardian equivalence would hold.

The Ricardian equivalence theorem requires not only that consumers be intergenerationally altruistic, but that their bequest motives be operative in the sense that consumers can bequeath whatever amount they choose subject to their budget constraint. To be more precise, it is possible that an altruistic consumer may like to leave a negative bequest to his children, but he is constrained from leaving a bequest less than zero. The fact that a consumer may want to leave a negative bequest does not necessarily violate the assumption that the consumer is altruistic. It may be that the consumer's children will all be so much wealthier than the consumer that, even though the consumer cares about the utility of his children, he could achieve higher utility by taking some of his children's resources and consuming them himself. Formal conditions that imply that altruistic consumers would like to leave negative bequests have been presented by Drazen (1978) and Weil (1987). Under these conditions, if the consumer is constrained from leaving a negative bequest, he will instead leave a zero bequest. In such cases, a tax cut that is followed by a tax increase after the consumer's death will reduce the present value of the taxes paid by the consumer and he will increase his consumption. In effect, the current tax cut helps the consumer achieve the desired negative bequest by giving him current resources and taking resources away from his descendants.

Another reason for departure from the Ricardian equivalence theorem is that policy may redistribute resources among families that have different marginal propensities to consume. For instance, suppose that one half of the consumers receive a 200 dollar tax cut in the current year and the other half of the consumers have unchanged taxes in the current year. The government finances the tax cut by issuing bonds in the amount of 100 dollars per capita, and in the following period it redeems the bonds and pays the interest. For simplicity, suppose that the population is constant and that the interest rate on government bonds is five per cent per year. Then in the year following the tax cut there is a tax increase of 105 dollars per consumer. Finally, suppose that this tax increase is levied on all consumers equally. In this case, the tax cut in the current year redistributes resources from the consumers whose taxes are unaffected to the consumers whose taxes are reduced in the current year. The recipients of the transfer will increase their consumption and the other consumers will reduce their consumption. The reallocation of consumption across consumers may be viewed as a violation of Ricardian equivalence. Whether aggregate consumption rises or falls depends on the marginal propensities to consume of the recipients of the transfer compared with the marginal propensities to consume of the other consumers. If all consumers have equal marginal propensities to consume, then there will be no effect on aggregate consumption or capital accumulation. However, if, for instance, the recipients of the transfers have a higher marginal propensity to consume than the other consumers, then aggregate consumption would increase. It should be pointed out that, in some sense, this example does not represent a violation of the Ricardian equivalence theorem, because it ignores the possibility that there might exist an insurance market for individual tax liabilities. If there were such a market, then consumers could have insured themselves against the redistribution of taxes. Such markets do not generally exist, but whether the Ricardian equivalence theorem holds may depend on the reason why these markets do not exist.

To see the role of insurance markets in a different context, consider consumers who each contribute 100 dollars to a social security fund during their working life. Suppose that at the end of their working lives some of the consumers die and the others survive and live in retirement. Although the number of consumers who die at retirement may be predictable, the identities of those who will die are not predictable. The surviving retired consumers each receive an equal share of the social security fund

(with accrued interest) to which they contributed while they were working. Each survivor's social security income is greater than the 100 dollars (plus interest) which he contributed, because the fund contains the contributions plus interest of his peers who died at the end of the working life.

Does the introduction of this type of social security system affect consumption and capital accumulation or does Ricardian equivalence imply that consumption and capital accumulation will be unaffected? To answer this question, it is useful to observe that this stylized social security system has the characteristics of an actuarially fair annuity. That is, consumers pay a premium when young (the social security tax) and receive a payment if, and only if, they survive to old age. Furthermore, if all consumers face the same probability of dying, the rate of return to the survivors is equal to the actuarially fair rate of return. If there were a competitive annuity market offering the actuarially fair rate of return, the social security system would have no effect on consumption or capital accumulation. Workers who are taxed 100 dollars are essentially forced to hold 100 dollars of the publicly provided actuarially fair annuity called social security; however, these consumers can afford, and will choose, to maintain their originally planned consumption and bequests by reducing their holdings of privately supplied annuities by 100 dollars. This reduction in the holding of private annuities allows consumers to re-establish their initial portfolios of annuities and other assets while maintaining consumption unchanged. Thus, the Ricardian equivalence theorem holds in this example, provided that consumers each originally planned to hold at least 100 dollars of private annuities.

If the probability of surviving until retirement differs across consumers, and if individual consumers are better informed about their own survival probabilities than are insurance companies, then the funded social security system described above will affect consumption. The reason is that, if an insurance company offered annuities at a price that would be actuarially fair to the average consumer, it would suffer from what is known as ‘adverse selection’. As a simple example, suppose that insurance companies know the average mortality probability but have no additional information about the mortality probabilities of individual consumers. If an insurance company offered annuities at a price that would be actuarially fair to the average consumer, then consumers who believe they are healthier than average would view these annuities as a bargain; consumers who believe they are less healthy (or engage in more dangerous activities) than average would view these annuities as overpriced because these consumers have a smaller chance of living to reap the rewards. As the healthy consumers would buy a disproportionately large share of annuities, they would, on average, inflict losses on the sellers of these annuities and would induce these sellers to charge a higher price for annuities. However, the social security system can supply its annuities at the actuarially fair price for the average consumer because a compulsory social security system is immune to adverse selection. That is, because the government can determine the amount of the publicly provided annuity held by each individual, it does not have to worry that a disproportionately large share of annuities are held by healthy consumers. Therefore, as shown in Abel (1986), the annuity offered by the social security system would yield a higher rate of return than private annuities, or, equivalently, would be made available at a lower price to consumers. Because of the difference in the prices of the publicly provided and privately supplied annuities, consumers could not exactly offset the effects of social security by transacting in private annuity markets.

The example in which adverse selection leads to violation of the Ricardian equivalence theorem was constructed to obey the strict set of rules demanded by strong adherents to the view that the choice between debt finance and tax finance is irrelevant. In particular, the following assumptions were maintained: (*a*) consumers are forward-looking and understand that a bond-financed tax cut implies an

increase in future taxes; (*b*) consumers have operative altruistic bequest motives so that they care about taxes after their death; (*c*) there is a complete set of competitive markets; and (*d*) only lump-sum taxes are changed. However, actual economies display several important departures from each of these assumptions. Violations of these assumptions are discussed below.

First, despite the widespread appeal of rational expectations in modern economics, it may simply be that consumers do not fully appreciate the link between a current tax cut and an increase in future taxes. If consumers did not understand this link at all, then a current tax cut would tend to increase current consumption.

Second, consumers may not have a bequest motive, either because they have no children or because they do not care about the welfare of anyone else. Even if consumers do have a bequest motive, it may not be operative as discussed above. Even if the bequest motive is operative, it may not be of the appropriate form for the Ricardian equivalence theorem to hold. If a consumer's utility depends directly on the size of the bequest he leaves rather than on the utility of his heirs, then a current tax cut followed by a tax increase on his heirs, would tend to raise the current consumption of the consumer. The reason is that he does not care about his heirs' utility per se. His bequest yields utility directly just as any other consumption good. As a result of the decrease in taxes he must pay over his lifetime, the consumer will have a higher level of lifetime income and can increase his own consumption and the bequest he leaves. If his own consumption and the bequest are both normal goods in his utility function, then he will choose to increase both.

Even if all consumers have operative altruistic bequest motives, a tax cut may increase current consumption. If all consumers have several children, but if each consumer cares about the utility of only one of his children, then there will be consumers in future generations whose utility is ignored by all current consumers. To the extent that future taxes are levied on these consumers, some part of future tax liabilities associated with a current tax cut will be ignored by current consumers. In this case, a tax cut would increase contemporaneous aggregate consumption.

Bernheim and Bagwell (1988) have challenged the plausibility of the assumption of intergenerational altruism by showing that this assumption leads to some untenable conclusions. If consumers A and B are unrelated to each other and both are altruistic toward consumer C, then A and B are effectively linked to each other, if both consumers A and B both plan to give positive transfers (bequests) to consumer C. For example, unrelated grandparents (consumers A and B) who plan to make positive transfers to their common grandchildren (consumers C) are effectively linked to each other. If the government transfers a dollar from consumer A to consumer B, these consumers can, and will choose to, undo this transfer and maintain their originally chosen patterns of consumption. The mechanism for undoing the government transfer is for consumer A to reduce his transfer to consumer C by one dollar and for consumer B to increase his transfer to consumer C by one dollar. Bernheim and Bagwell have argued that, if one takes intergenerational altruism seriously, such linkages are so widespread that all consumers are effectively linked to each other. In this case, all government transfers among consumers, including a transfer from future taxpayers to current taxpayers in the form of a bond-financed tax cut, would have no effect. Bernheim and Bagwell go on to show that even non-lump-sum taxes, and indeed prices themselves, would not affect consumption. Rather than conclude that all taxes and prices are irrelevant, Bernheim and Bagwell conclude that their findings cast doubt on the policy conclusions, including Ricardian equivalence, that are based on the assumption of intergenerational altruism.

Third, various types of insurance markets may be absent or, as described above, may suffer from adverse

selection. Chan (1983) and Barsky, Mankiw and Zeldes (1986) have argued that, if there are no markets for insuring against unpredictable fluctuations in after-tax income, then a current tax cut could increase current consumption. The argument, which was outlined by Barro (1974, p. 1115) and Tobin (1980, p. 60), is that to the extent that individual tax liabilities are proportional to income the tax system provides partial insurance against fluctuations in individual disposable income. Therefore, the increase in tax rates that follows a lump-sum tax cut in the current year will reduce the variability of future disposable income. The reduction in the riskiness of future disposable income reduces current precautionary saving that consumers undertake to guard against low future consumption. The counterpart of the reduction in precautionary saving is an increase in current consumption.

Fourth, most taxes are not lump-sum taxes. Generally, taxes are levied on economic activities, and changes in these taxes provide incentives to alter the levels of these activities. Although the existence of distortionary taxes does not in all cases imply that Ricardian equivalence is violated when applied to lump-sum tax changes, it does strain the interpretation of empirical tests of Ricardian equivalence that examine historical data on deficits and consumption.

As discussed above, there are many potential sources of departure from the Ricardian equivalence theorem, and ultimately the importance of these departures is an empirical question. The existing literature that attempts to test empirically whether Ricardian equivalence holds has produced mixed results, some claiming to show that it holds, and others the opposite. In judging the empirical relevance of the Ricardian equivalance theorem, however, the important question from the viewpoint of fiscal policy formulation is not whether the theorem holds exactly but whether there are departures from it that are quantitatively substantial. Existing empirical work has not yet produced a consensus on this question.

See Also

- government budget constraint
- public debt
- public finance

Bibliography

Abel, A. 1986. Capital accumulation and uncertain lifetimes with adverse selection. *Econometrica* 54, 1079–97.

Abel, A. 1989. Birth, death and taxes. *Journal of Public Economics* 39, 1–15.

Bailey, M. 1971. *National Income and the Price Level*, 2nd edn. New York: McGraw-Hill.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82, 1095–117.

Barsky, R., Mankiw, G. and Zeldes, S. 1986. Ricardian consumers with Keynesian propensities. *American Economic Review* 76, 676–91.

Bernheim, B. and Bagwell, K. 1988. Is everything neutral? *Journal of Political Economy* 96, 308–38.

Buchanan, J. 1958. *Public Principles of Public Debt*. Homewood, IL: Richard D. Irwin.

Buchanan, J. 1976. Barro on the Ricardian equivalence theorem. *Journal of Political Economy* 84, 337–42.

Carmichael, J. 1982. On Barro's theorem and debt neutrality: the irrelevance of net wealth. *American Economic Review* 72, 202–13.

Chan, L. 1983. Uncertainty and the neutrality of government financing policy. *Journal of Monetary Economics* 11, 351–72.

Drazen, A. 1978. Government debt, human capital and bequests in a lifecycle model. *Journal of Political Economy* 86, 337–42.

Feldstein, M. 1976. Perceived wealth in bonds and social security: a comment. *Journal of Political Economy* 84, 331–6.

Patinkin, D. 1965. *Money, Interest and Price*, 2nd edn. New York: Harper and Row.

Ricardo, D. 1821. *The Principles of Political Economy and Taxation*. London: M. Dent and Sons, 1911.

Tobin, J. 1980. *Asset Accumulation and Economic Activity*. Chicago: University of Chicago Press.

Weil, P. 1987. ‘Love thy children’: reflections on the Barro debt neutrality theorem. *Journal of Monetary Economics* 19, 377–91.

Weil, P. 1989. Overlapping families of infinitely-lived agents. *Journal of Public Economics* 38, 183–98.

How to cite this article

Abel, Andrew B. "Ricardian equivalence theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000146> doi:10.1057/9780230226203.1440

The New Palgrave Dictionary of Economics Online

Ricardian trade theory

Kiminori Matsuyama
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Ricardian trade theory takes cross-country technology differences as the basis of trade. By abstracting from the roles of factor endowment and factor intensity differences, which are the primary concerns of factor proportions theory, Ricardian trade theory offers a simple and yet powerful framework within which to examine the effects of country sizes, of technology changes and transfers, and of income distributions. Moreover, its simple production structure makes it relatively easy to allow for many goods and many countries, and hence capable of generating valuable insights which are lost in the standard two-country, two-sector model of international trade.

Keywords

absolute advantage; Cobb–Douglas functions; comparative advantage; constant returns to scale; diminishing returns to scale; export-based and import-biased technology; export subsidies; factor endowment; factor intensity; factor proportions; gains from trade; globalization; iceberg costs; immiserizing growth; increasing returns to scale; international trade; North–South trade; product cycles; productivity growth; Ricardian trade theory; size of nations; specific factor models; structural change; technical change; terms of trade; trade costs; trade, technology diffusion and growth; transfer of technology; transfer problem

Article

Ricardian trade theory takes cross-country technology differences as the basis of trade. By abstracting from the roles of cross-country factor endowment differences and cross-industry factor intensity differences, which are the primary concerns of factor proportions theory (such as Heckscher–Ohlin and Specific Factor models), Ricardian trade theory offers a simple and yet powerful framework within which to address many positive and normative issues of international trade. It is particularly well-equipped to examine the effects of country sizes, of technology changes and transfers, and income distributions. Furthermore, its simple production structure makes it relatively easy to allow for many tradable goods and many countries, and hence capable of generating valuable insights, which are lost in the standard two-country, two-goods model of international trade.

Let us start with the Ricardian model with a continuum of tradable goods, adopted from Dornbusch, Fischer and Samuelson (DFS) (1977). The world consists of two countries, Home and Foreign. There is a continuum of competitive industries, indexed by $z \in [0, 1]$, each producing a homogenous tradable good, also indexed by z. There is only one non-tradable factor of production, called labour. (Or, if there are many non-tradable factors, they can be aggregated into a single composite factor.) Let $a(z)$ and $a^*(z)$ be the Home and Foreign unit labour requirements of good z, that is, labour input required to produce one unit of output z at Home and Foreign. Without loss of generality, we can index z so that Home's relative efficiency, $A(z) \equiv a^*(z) / a(z)$, is non-increasing in z. In Figure 1, it is strictly decreasing. In short, Home (Foreign) has a *comparative advantage* in low-indexed (high-indexed) goods.

Figure 1
The equilibrium factor terms of trade and patterns of specialization

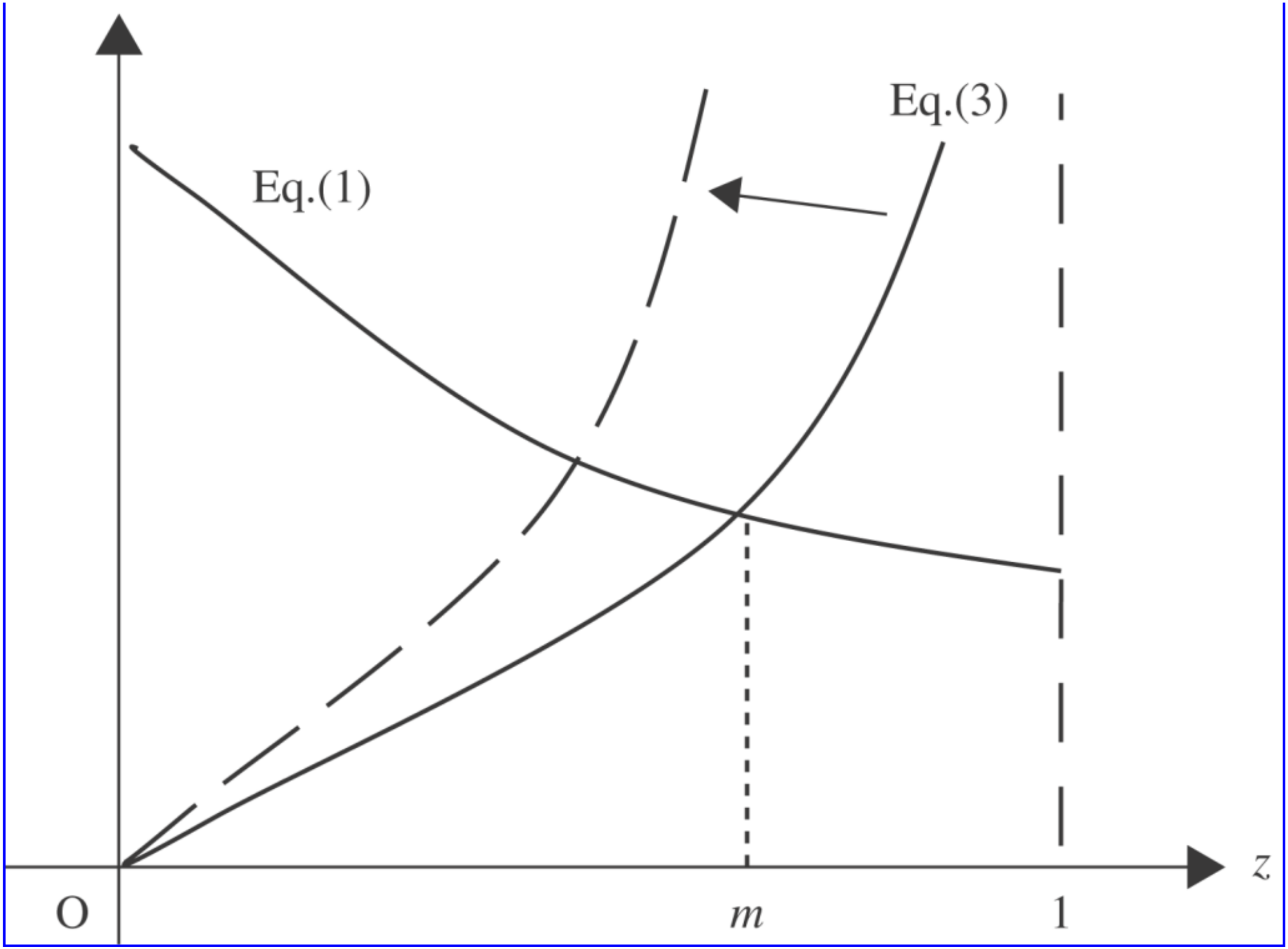

Let w and w^* denote the wage rates at Home and Foreign. Then, the prices in autarky are given by $p(z)=wa(z)$ at Home and $p^*(z)=w^*a^*(z)$ at Foreign. Under free trade (and in the absence of any trade costs), the price of each good is equalized across the two countries and is given by $p(z)=p^*(z)=\min\{wa(z), w^*a^*(z)\}$. Then, for a given relative wage rate or a

given level of the *factoral terms of trade*, w/w^*, there is a marginal good, m, defined by

$$\frac{w}{w^*} = A(m), \tag{1}$$

such that Home produces only goods in $[0,m]$, and Foreign produces only goods in $[m,1]$, and the prices become

$$p(z) = p^*(z) = wa(z), \quad z \in [0, m]; p(z) = p^*(z) = w^* a^*(z), \quad z \in [m, 1]. \tag{2}$$

To pin down the relative wage rate, we must specify the demand conditions. To keep it simple, let us assume that there are $L(L^*)$ households at Home (Foreign), each supplying one unit of labour, and that every household shares the symmetric Cobb–Douglas preferences defined over $z \in [0, 1]$, as $U = \int_0^1 \log[c(z)]\,dz$ and $U^* = \int_0^1 \log[c^*(z)]\,dz$. Then, the world income (and the world total expenditure), $wL+w^*L^*$, is also equal to the world expenditure on each good. Since Home produces the goods in $[0,m]$, the total expenditure on the Home goods is $m(wL+w^*L^*)$, which must be equal to the Home income, wL, in equilibrium. This condition yields

$$\frac{w}{w^*} = \frac{m}{1-m}\left[\frac{L^*}{L}\right] \tag{3}$$

which is depicted by the upward sloping curve in Figure 1. It is upward-sloping because a higher m means that a larger fraction of the world expenditure goes to the goods produced by the Home labour, hence its relative wage goes up. As shown by the intersection of the two curves in Figure 1, eq. (1) and eq. (3) jointly determine the equilibrium relative wage rate, w/w^*, and the equilibrium patterns of trade and specialization, m, as Home exports and Foreign imports goods in $[0,m)$ and Home imports and Foreign exports goods in $(m,1]$. In short, *the patterns of trade follow the patterns of comparative advantage*.

The standard two-country, two-goods Ricardian model, found in many college textbooks, may be recovered as a special case of this model, where $A(z)=A_1$ for $z \in [0, \alpha]$ and $A(z)=A_2$ for $z \in (\alpha, 1]$, with $A_1>A_2$, as shown in Figure 2. By aggregating all the goods in $[0, \alpha]$ as a composite good, called Good 1, and all the goods in $(\alpha, 1]$ as another composite good, called Good 2, the model becomes a two-sector model, where the households have the preferences, $U=\alpha\log(C_1)+(1-\alpha)\log(C_2)$ and $U^* = \alpha\log(C_1^*) + (1-\alpha)\log(C_2^*)$. Viewed this way, the model highlights the restrictive feature of the two-good assumption in the textbook Ricardian model.

Figure 2

The two-goods case

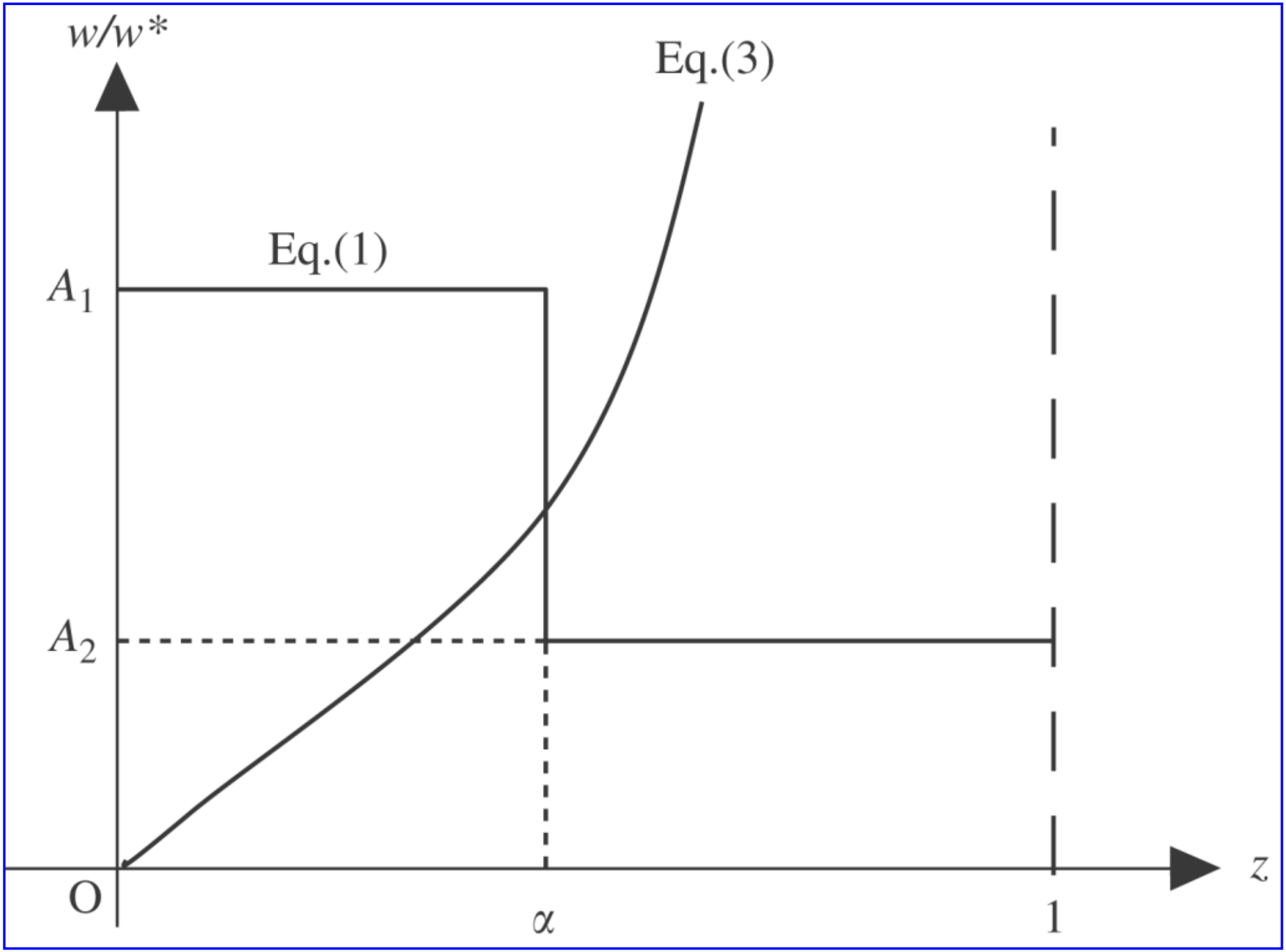
w/w*
Eq.(3)
Eq.(1)
A1
A2
O
α
1
z

Gains from trade and country size effects

The Home and Foreign welfares are measured by $U = \int_0^1 \log[w/p(z)]\,dz$ and $U^* = \int_0^1 \log[w^*/p^*(z)]\,dz$, respectively. In autarky, they are equal to

$$U_A = -\int_0^1 \log[a(z)]\,dz;\ U_A^* = -\int_0^1 \log[a^*(z)]\,dz; \tag{4}$$

and, under free trade, they are equal to

$$U_T = -\int_0^m \log[a(z)]\,dz + \int_m^1 \log\left[\frac{w}{w^* a^*(z)}\right]dz;\ U_T^* = \int_0^m \log\left[\frac{w^*}{wa(z)}\right]dz - \int_m^1 \log[a^*(z)]\,dz. \tag{5}$$

Subtracting (4) from (5) and using (1) show that the welfare changes from autarky to free trade by:

$$\Delta U = U_T - U_A = \int_m^1 \log\left[\frac{A(m)}{A(z)}\right]dz;\ \Delta U^* = U_T^* - U_A^* = \int_0^m \log\left[\frac{A(z)}{A(m)}\right]dz, \tag{6}$$

both of which are strictly positive in the case depicted in Figure 1. More generally, both countries gain from trade, as long as $0<m<1$ and $A(0)>A(m)>A(1)$. Note that this condition could hold even when one country, say Home, has *absolute advantage* over the other, say Foreign, that is, $A(z)>1$ for $z \in [0, 1]$. Clearly, such absolute advantage allows Home households to enjoy higher wage income and hence a higher standard of living than Foreign households, $w>w^*$, and $U_T - U_T^* = \log(w) - \log(w^*) > 0$. Yet both countries gain from trade as long as trade allows them to specialize in the goods that they are relatively good at producing.

In the two-goods case, shown in Figure 2, both countries gain from trade only when $m=\alpha$ and $A_1>w/w^*>A_2$, which requires that $A_2(1-\alpha)/\alpha<L^*/L<A_1(1-\alpha)/\alpha$. Home does not gain from trade if $w/w^*=A_2$, which occurs when $L^*/L \leq A_2(1-\alpha)/\alpha$, and Foreign does not gain from trade if $w/w^*=A_1$, which occurs when $L^*/L \geq A_1(1-\alpha)/\alpha$. This result of the two-sector model is often interpreted as saying that a large country cannot gain from trade, as it remains *incompletely* specialized, or that a country must *fully* specialize in order to gain from trade, but this is due to an artificial feature of the two-goods model which restricts the cross-country differences in technology.

What is generally true is that, as one country becomes large (small) relative to the rest of the world, its gains from trade become smaller (large). As shown in Figure 1, an increase in L^*/L, which shifts the upward-sloping curve to the left, leads to a higher w/w^* and a lower m, which implies a higher ΔU and a lower ΔU^*, as seen from (6). For example, a faster population growth in the South (Foreign) allows the North (Home) to specialize further, which improves its factoral terms of trade and its standard of living at the expense of the South. (The same phenomenon might be described by the protectionist in the North as saying, ‘because of the cheap labour in the South, the North loses its competitive advantage and industries move from the North to the South’.)

This also suggests that a country with a small population could enjoy higher per capita income even with limited technological superiority. In Figure 3, Home's technologies are inferior in almost all the goods, yet, thanks to its relatively small population, Home enjoys higher per capita income. This may explain why countries like Norway and Switzerland

enjoy a high standard of living even though their geography and climates are not particularly suitable to most economic activities. With smaller populations, they can maintain a high standard of living by specializing in a narrower range of activities that they are particularly good at. This effect is difficult to see within a two-goods model.

Figure 3

Gains from trade and country size effects

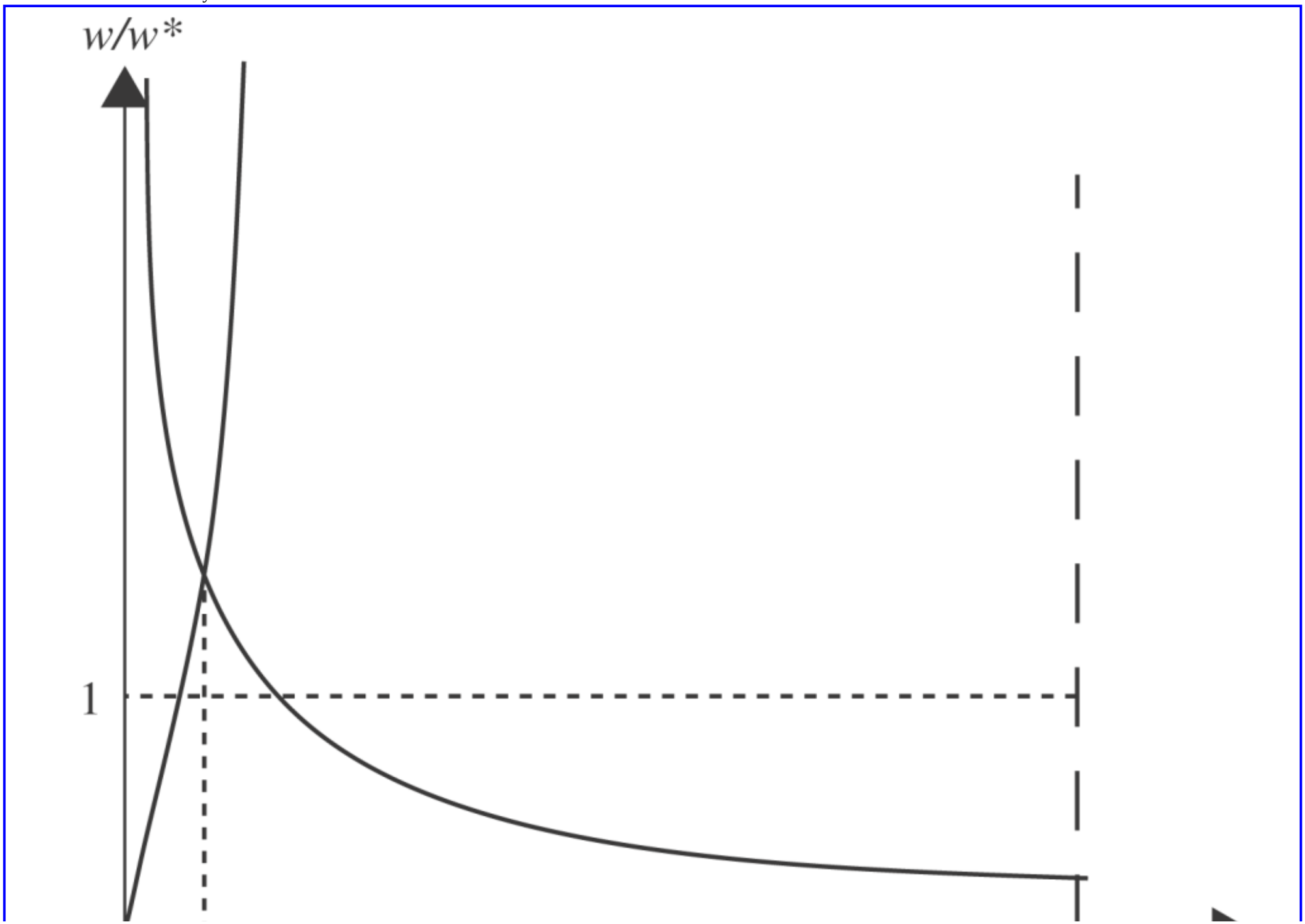

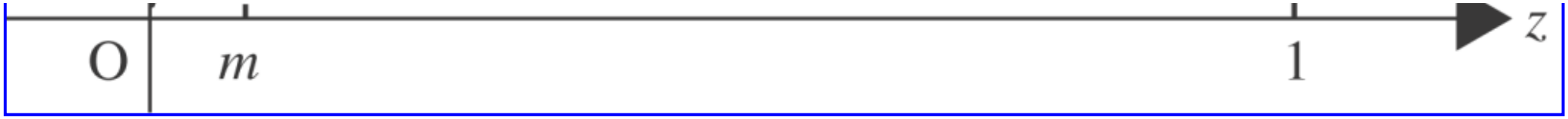

These results suggest diminishing returns to scale (DRS) at the country level, even though technologies satisfy the constant returns to scale (CRS) property. This is because the endogeneity of the terms of trade, w/w^*, introduces *de facto* diminishing returns. To see some macroeconomic implications, let us reinterpret the model in the following way. Home and Foreign produce their GDPs, Y and Y^*, by the CRS aggregate production function, $\log(Y) = \int_0^1 \log[c(z)]\,dz$ and $\log(Y^*) = \int_0^1 \log[c^*(z)]\,dz$, where $c(z)$ and $c^*(z)$ denote the inputs of the tradeable intermediate goods, $z \in [0, 1]$. The representative household at Home and at Foreign supplies L and L^* units of the composite of the primary factors, which may include not only labour but also physical and human capital. Then, the expressions analogous to (5) become

$$\log\left[\frac{Y}{L}\right] = \int_m^1 \log\left[\frac{w}{w^* a^*(z)}\right] dz - \int_0^m \log[a(z)]\,dz;$$

$$\log\left[\frac{Y^*}{L^*}\right] = \int_0^m \log\left[\frac{w^*}{wa(z)}\right] dz - \int_m^1 \log[a^*(z)]\,dz.$$

The effect of a change in L and L^* can be seen by totally differentiating the above expressions, which yields the following growth accounting:

$$\frac{dY}{Y} = \frac{dL}{L} + (1-m)\left[\frac{dw}{w} - \frac{dw^*}{w^*}\right]; \frac{dY^*}{Y^*} = \frac{dL^*}{L^*} + m\left[\frac{dw^*}{w^*} - \frac{dw}{w}\right].$$

If L and L^* grow at the same rate, w/w^* remains constant, and hence both Y and Y^* also grow at the same rate. However, if L grows faster than L^*, then Y grows slower than L and Y^* grows faster than L^* through the terms of trade effect. This example also suggests that, even when there are increasing returns to scale (IRS) in the aggregate production technologies, naive cross-country growth regression exercises which do not take into account interdependence among countries might fail to uncover economies of scale. See also Acemoglu and Ventura (2002), which studies how such a terms-of-trade mechanism generates stable cross-country distribution of income in the world even when different countries accumulate factors at different rates.

Technology changes and transfers

Because it takes cross-country technology differences as the basis of trade, the Ricardian model is well-suited to study the effects of technology changes. Let $g(z) \equiv -d\log[a(z)]$ and $g^*(z) \equiv -d\log[a^*(z)]$ denote the rate of productivity change in industry z at Home and Foreign. By totally differentiating (5), the Home welfare changes can be expressed as

$$dU = \int_0^m g(z)\,dz + \int_m^1 g^*(z)\,dz + \left[\frac{1-m}{1+\xi(m)m(1-m)}\right]\left[g(m) - g^*(m)\right]$$

where $\xi(z) \equiv -d\log[A(z)]/dz$. For example, it is easy to see that Home always gains from productivity growth in its export sectors, $g(z)>0$ for $z \in [0, m]$. Does Home gain from productivity growth abroad, as well? If Foreign experiences a *uniform* productivity growth in its export sectors (that is, $g^*(z)=g^*>0$ for all $z \in [m, 1]$), then the answer is yes because

$$dU = \left[\frac{m(1-m)^2\xi(m)}{1+\xi(m)m(1-m)}\right]g^* > 0,$$

unless $\xi(m)=0$, in which case productivity gains in the Foreign export sectors are entirely offset by the terms of trade change.
On the other hand, Home may lose from Foreign productivity growth if it is concentrated around the marginal sector, m. The following example, taken from Jones (1979), illustrates this possibility. Let $A(z) = A_1 = a_1^*/a_1$ for $z \in [0, \alpha_1]$; $A(z) = A_2 = a_2^*/a_2$ for $z \in (\alpha_1, \alpha_1+\alpha_2]$, and $A(z) = A_3 = a_3^*/a_3$ for $z \in (\alpha_1+\alpha_2, 1]$, with $A_1>A_2>A_3$. Again, we may view this example as a three-sector model, by aggregating all the goods within each segment, as Goods 1, 2, and 3. When the upward-sloping curve intersects with $A(z)$ at its middle segment, $w/w^*=A_2$. Then, the Home and Foreign welfares are given by

$$U_T = (1-\alpha_1+\alpha_2)\log(a_2^*/a_3^*); U_T^* = -(\alpha_1+\alpha_2)\log(a_2^*) - (1-\alpha_1+\alpha_2)\log(a_3^*),$$

where we normalize $a_1=a_2=a_3=1$ without loss of generality to simplify the expression. The Home welfare declines (and the Foreign welfare improves) unambiguously when Foreign productivity growth takes the form of a reduction in a_2^*, depicted by the arrow in Figure 4. The reason for the Home loss is that the Home purchasing power measured in Good 2 remains unchanged as Foreign productivity growth is completely offset by the increase in the Foreign wage, while the Home purchasing power measured in Good 3 declines, as the increase in the Foreign wage makes it more expensive.
Figure 4
The three-goods case

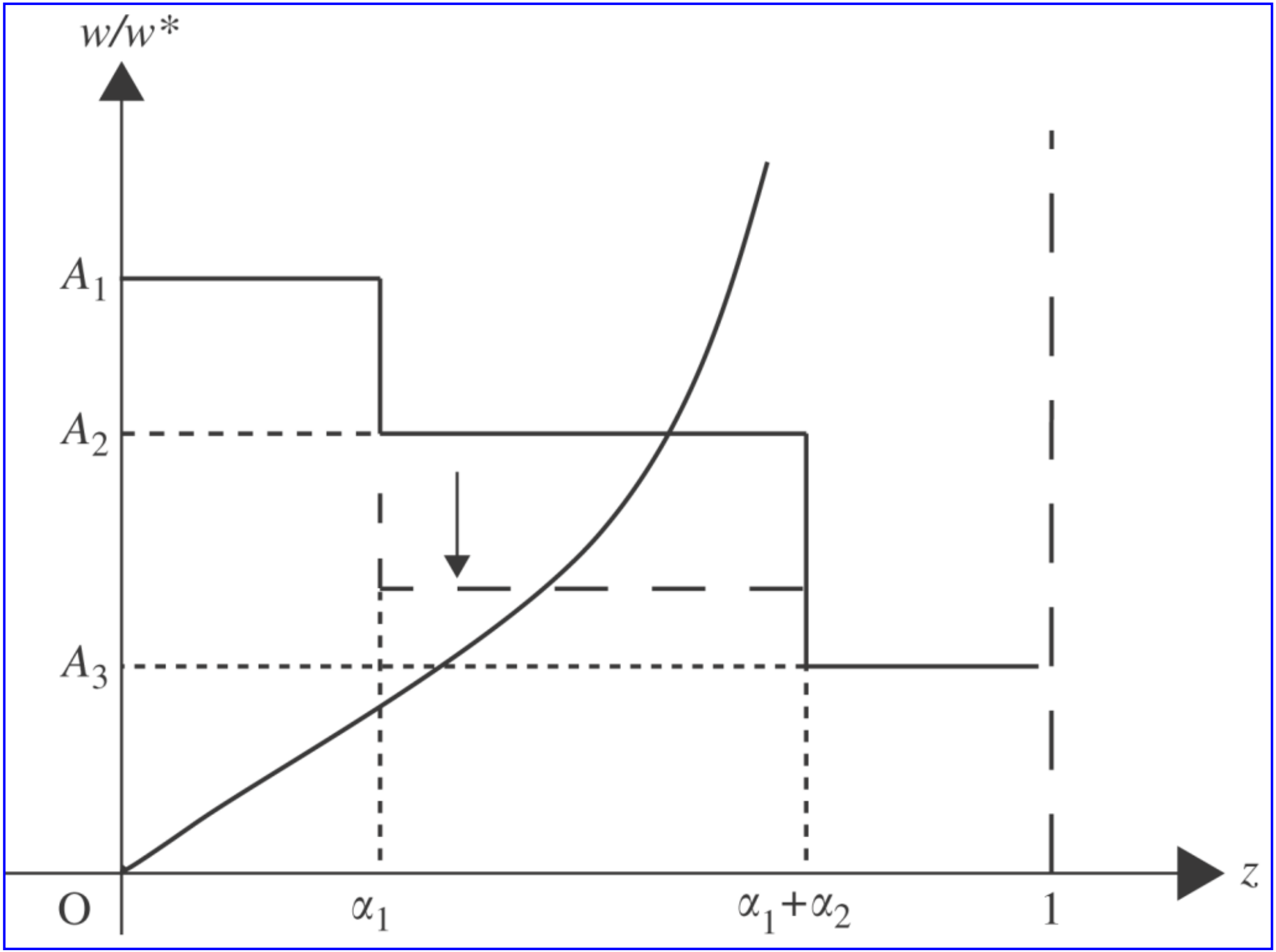

The above example again demonstrates the restrictiveness of the two-goods assumption. In particular, it suggests that the widely used distinction between ‘export-biased’ and ‘import-

biased' technology changes in two-sector models, first introduced by Hicks (1953), is of very limited value, as it can be applied only when these changes are uniform across all the export (or import) sectors. Indeed, it can be misleading because the effects of technology changes that take place in *some* export sectors may be very different from those of 'export-biased technological changes' in the two-sector model. (A similar point can be made for the analysis of trade policies. In the standard competitive two-sector model of trade, the government cannot improve its national welfare by providing export subsides to its export sector. This should be interpreted as saying that export subsidies provided *uniformly* to all of its export sectors cannot improve its national welfare. Indeed, using the Ricardian models similar to the one above, Itoh and Kiyono, 1987, showed that *selective* export subsidies that *target* sectors around the marginal sector can improve the national welfare.)

The same mechanism could operate when the technologically lagging country succeeds in catching up with the technologically leading country. Suppose $a(z)=1$ for all $z\in[0,1]$ and $A(z)=a^*(z)=\Lambda^{1-z}$, where $\Lambda>1$ is a parameter, representing the extent to which Foreign 'lags behind' Home technologically. For example, each tradable good is produced by performing two tasks by x_1 and x_2, with the Cobb–Douglas production function, $F^z(x_1, x_2)=[x_1/z]^z[x_2/(1-z)]^{1-z}$, and that the unit labour requirement for task 1 is equal to one everywhere, while the unit labour requirement for task 2 is one at Home and $\Lambda>1$ at Foreign. One may think that Λ reflects the technology gap, which affects when performing task 2, but not task 1. Then, Home has absolute advantage in all the goods, but Foreign has comparative advantage in the high-indexed goods, which can be produced mostly by performing the simple task 1. (Krugman, 1986, offered another story behind a similar parameterization of $A(z)$.) With this parameterization, we have

$$U_T = \frac{(1-m)^2}{2}\log\Lambda,\ U_T^* = -\frac{1-m^2}{2}\log\Lambda,$$

where m is determined by the condition, $mL^*/(1-m)L=\Lambda^{1-m}$. Home could lose if Foreign succeeds in narrowing the gap (that is, a reduction in Λ). The reason is easy to understand. As the gap narrows, Foreign becomes more similar to Home, and Home gains little from trading with a country similar to itself. Indeed, if Foreign catches up completely, $\Lambda=1$, Home loses all the gains from trading with Foreign because the two countries become identical. Note that the underlying mechanism in this example is the same as in the three-goods example. When Foreign narrows the gap, their productivity growth is not uniform across its export sectors. It is larger in the sectors in which Home has bigger absolute advantages. However, it is false to say that Home suffers because Foreign productivity growth is 'import-biased'. The Home loss is caused by Foreign productivity growth around the marginal sector, not at the lower end of the spectrum.

Nontraded goods, trade costs, and effects of globalization

We have been examining the effects of trade by comparing the two extreme cases: autarky, where no goods are traded, and 'free' trade, where all goods are costlessly tradeable. Let us now introduce some trade costs and examine the effects of (partial) trade liberalization by reducing the trade costs. Matsuyama (2007c) conducts such exercises by following DFS (1977), which proposed two alternative ways of introducing trade costs in their model: traded–nontraded dichotomy and uniform iceberg costs.

Traded – nontraded dichotomy

Suppose that only the goods in $[0,k]$ are tradable at zero cost and $A(z)\equiv a^*(z)/a(z)$ is continuous, and strictly decreasing in z, within this range. On the other hand, trade costs are so high for the goods in $(k,1]$ that they need to be produced locally. At this point, we do not have to specify the production technologies for these nontradables.

Given the marginal good, $m\in[0,k]$, defined by $A(m)=w/w^*$, Home produces all the goods in $[0,m]$ for both countries and Foreign produces all the goods in $(m, k]$. In addition, each country produces all the goods in $(k,1]$ locally. Therefore, the total expenditure on the goods produced at Home is equal to $m(wL+w^*L^*)+(1-k)wL$, which must be equal to the Home income, wL, in equilibrium. This condition yields

$$\frac{w}{w^*} = \frac{m}{k-m}\left[\frac{L^*}{L}\right]. \tag{7}$$

The equilibrium is determined jointly by eq. (1) and eq. (7), as shown in Figure 5. DFS (1977) used this extension to study the classical transfer problem. In the presence of the nontraded goods, the German households spend a larger share of their income on the goods produced in Germany than the households abroad. Because of this ‘home bias’ in demand, an exogenous income transfer from Germany to the Allies (the war reparations after the Treaty of Versailles of 1919) shifts demand away from the German goods, which leads to a deterioration of the German terms of trade, imposing the additional burden on the German economy.

Figure 5

Non-uniform globalization

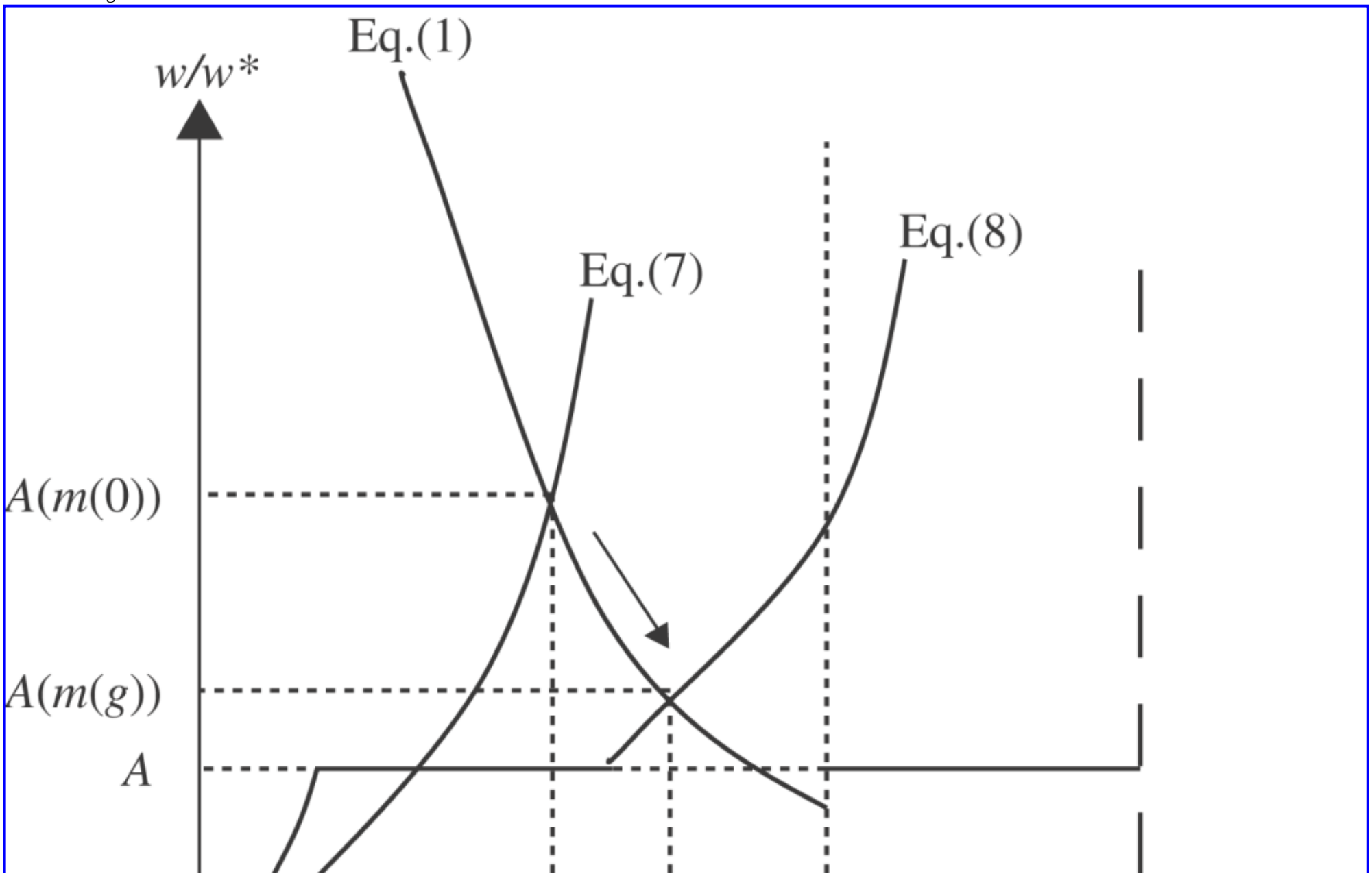

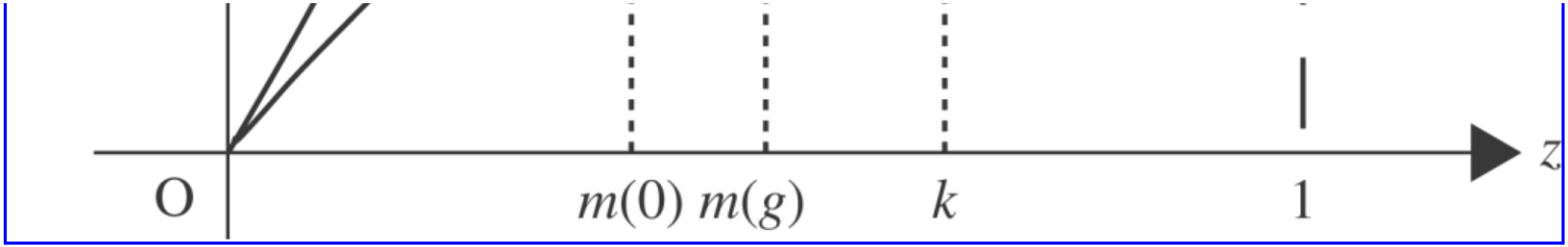

Let us use this model to study the effects of a globalization. Imagine that some nontradables become tradable. For example, the governments might decide to lift the bans on trading some goods that can be traded costlessly. Or advances in information and communication technologies might open up the possibility of trading some labour services at zero cost. The effects, of course, depend on the relative efficiency of the two countries in producing these newly tradables. Consider the case where $A(z)=A$ for all $z\in(k, 1]$, and that a fraction g of these goods becomes newly tradable at zero cost. Then, if $w/w^*>A$, all of the newly tradeables are produced at Foreign. Therefore, given the marginal good, $m\in[0, k]$, Home produces all the goods in $[0,m]$ for both countries and $(1-g)(1-k)$ fraction of the goods (those which remain nontradable) locally. Hence, in equilibrium, $wL=m(wL+w^*L^*)+(1-g)(1-k)wL$. On the other hand, if $w/w^*<A$, all of the newly tradables are produced at Home. Therefore, Home produces $m+g(1-k)$ fraction of the goods for both countries and $(1-g)(1-k)$ fraction of the goods locally. Hence, in equilibrium, $wL=[m+g(1-k)](wL+w^*L^*)+(1-g)(1-k)wL$. Thus, we have, instead of (7),

$$\frac{w}{w^*} = \frac{m+g(1-k)}{k-m}\left[\frac{L^*}{L}\right] \text{ for } \frac{w}{w^*} < A; \frac{w}{w^*} = \frac{m}{k+g(1-k)-m}\left[\frac{L^*}{L}\right] \text{ for } \frac{w}{w^*} > A; \tag{8}$$

and $w/w^*=A$, otherwise. Note that setting $g=0$ in eq. (8) recovers eq. (7). A higher g shifts the graph to the right above $w/w^*=A$ and to the left below $w/w^*=A$. For each value of g, eq. (1) and eq. (8) jointly determines the marginal good and the Home relative wage, which we denote by $m(g)$ and $A(m(g))$.

Suppose that, before globalization, $g=0$, the equilibrium Home relative wage, $A(m(0))$, is higher than A, as shown in Figure 5. The arrow indicates the shift caused by an increase in g, which is small enough to keep the Home relative wage higher than A. When some nontraded sectors are opened up, Home abandons the production of these new tradeables, and instead starts producing and exporting the goods in $(m(0), m(g))$, which Home imported previously. The Home relative wage declines as a result, from $A(m(0))$ to $A(m(g))$. The Home and Foreign welfares may be evaluated by

$$U(g) = \int_{m(g)}^{k} \log\left[\frac{A(m(g))}{A(z)}\right]dz + g(1-k)\log\left[\frac{A(m(g))}{A}\right];$$

$$U^*(g) = m(g)\log\left[\frac{A}{A(m(g))}\right] + \int_{m(g)}^{k} \log\frac{A}{A(z)}dz - \log A$$

where we use the normalization, $A(z)=a^*(z)/a(z)=a^*(z)$ for all $z\in[0,1]$, to simplify the expressions. A globalization (an increase in g) affects the Home welfare through two effects

that operate in the opposite directions. On one hand, it allows Home to reallocate its labour to the sectors where they have higher relative efficiency, that is, from A to $A(m(g))$ or higher. On the other hand, its relative wage rate, or the factoral terms of trade, $w/w^*=A(m(g))$, deteriorates. The overall effect is generally ambiguous. However, if an increase in g brings down its relative wage rate $A(m(g))$ sufficiently close to A, the positive reallocation effect is dominated by the negative terms of trade effect, so that a further globalization harms the Home welfare. In contrast, a globalization unambiguously improves the Foreign welfare, because both effects operate positively.
The possibility that Home could lose when a globalization takes this form should not be too surprising, because it can be viewed as a form of non-uniform technological changes. Indeed, this mechanism may capture some of the widely held concerns that high-wage countries might lose from 'outsourcing' simple tasks to low-wage countries.

Uniform iceberg cost

Suppose now that all the goods, $z\in[0,1]$, are tradeable but subject to the iceberg cost. Each good, when shipped abroad, melts away in transit and only a fraction $g<1$ arrives at the destination. Thus, in order to supply one unit of each good, the exporter must produce $1/g>1$ units of the good. Then, Home exports to Foreign only when $a(z)w/g<a^*(z)w^*$, or $w/w^*g<A(z)$, and Foreign exports to Home only when $a(z)w>a^*(z)w^*/g$, or $wg/w^*>A(z)$. Thus, there are two marginal goods, defined by

$$\frac{w}{w^* g} = A(m^-) > A(m^+) = \frac{wg}{w^*}, \tag{9}$$

such that Home produces all the goods in $[0, m^-)$ for both countries; Foreign produces all the goods in $(m^+, 1]$ for both countries, and each produces the goods in $[m^-, m^+]$, which becomes (endogenously) nontraded goods. The demand condition now becomes

$$\frac{w}{w^*} = \frac{m^-}{1-m^+}\left[\frac{L^*}{L}\right]. \tag{10}$$

Eqs (9) and (10) jointly determine three endogenous variables, m^-, m^+, and w/w^*.
One could proceed to examine the effects of a reduction in the trade cost, by increasing g. This is left as an exercise for interested readess.

Multiple countries

The two-country assumption is clearly restrictive for certain purposes, such as analysing the income distribution across countries, studying the patterns of bilateral trade flows, let alone the issues related to the regional integration, such as NAFTA. It is relatively straightforward to extend the two-goods Ricardian model for an arbitrary number of countries; see, for example, Becker (1952) for a finite number of countries and Matsuyama (1996) and Yanagawa (1996) for a continuum of countries. It has been a challenge to allow for an arbitrary number of goods *and* countries in a tractable way. For example, Wilson (1980) extended the DFS model in many dimensions, including a finite number of countries, but it does not permit more than a local perturbation analysis. Acemoglu and Ventura (2002), in their analysis of the cross-country income distribution, assumed the extreme form of technological heterogeneity by adopting the Armington assumption, which prevents the patterns of specialization from changing endogenously.
Eaton and Kortum (2002) developed a parsimonious representation of the Ricardian model with a continuum of goods, which allows for an arbitrary number of countries with the iceberg costs that are uniform across sectors but vary across country pairs. Their key idea is to view the technology heterogeneity across countries as a realization from the Frechet distributions, instead of trying to index the goods in a particular order. This yields simple expressions relating the bilateral trade volumes to technology and geographical barriers, and they use these expressions to estimate the parameters needed to quantify the effects of various policy experiments. For further development, see Alvarez and Lucas (2004).

Multi-stage trade and vertical specialization

Sanyal (1983) proposed a reinterpretation of the DFS model, according to which the final good is produced through many stages of production, $z \in [0,1]$, in order to analyse trade in intermediate inputs and vertical specialization. If the order in which these inputs need to be produced in the vertical chain of production perfectly coincides with the pattern of comparative advantage, as Sanyal assumed, these inputs are traded only once, as one country specializes in the earlier stages of production and the other specializes in the later stages. Under more general patterns of comparative advantage, however, these inputs may be traded back and forth many times. In such a setting, even a small reduction in trade costs could cause a large and nonlinear increase in the volume of trade, as documented by Hummels, Ishii and Yi (2001).

More general preferences

With the Cobb–Douglas specification, each good receives a constant share of the expenditure regardless of the prices. Its homotheticity implies that the rich and the poor consume all the goods in the same proportions (when they all face the identical prices), so that the demand compositions are independent of the income distribution within each country. These features, while greatly simplifying the analysis, are too restrictive for addressing many important issues related to growth and development.

Consider, for example, the Fisher–Clark–Kuznets thesis, that is, the changing patterns of sectoral compositions, with the decline of agriculture, the rise and the fall of manufacturing, and the rise of the service sectors. To understand such patterns of structural change in the context of a global economy, Matsuyama (2007b) relaxed the Cobb-Douglas assumption to allow for non-unitary price and non-unitary income elasticities in the three-goods (two tradable and one nontradable) Ricardian model.

Non-homothetic preferences also play the key roles in many models of North–South trade. Flam and Helpman (1987), Stokey (1991), and Matsuyama (2000) all built two-country (North and South) Ricardian models with a continuum of goods, with the open-ended goods space, $z \in [0,\infty)$, and considered non-homothetic preferences with the property that, as the household's income goes up, its demand compositions shift towards higher-indexed goods. When the South, the poorer country, has comparative advantages in lower-indexed goods, the demand has home biases (in spite of the absence of any trade costs). Furthermore, the asymmetry of demand generates many comparative statics results that are absent in the standard Ricardian model. For example, in Matsuyama (2000), immiserizing growth might occur; uniform productivity growth in the South might make the South worse off, as all the benefits go to the North. Or, as the South's population grows, some industries migrate from the North to the South, and new industries are born in the North, generating the patterns of product cycles. Flam and Helpman (1987) and Matsuyama (2000) also looked at the roles of income distributions within each country by endowing different households with different amounts of labour.

Endogenous technologies and increasing returns

So far, we have taken the cross-country differences in technology as exogenous and examined their effects on patterns of specialization and trade. However, the patterns of trade and specialization may also affect technologies. Many Ricardian models with endogenous technologies have been developed to examine such two-way causality between technology and trade. Endogenous technologies have also been used as a natural way of introducing increasing returns in production. Due to the space constraint, we cannot do justice to this vast literature, which contains many alternative approaches to endogenize technologies (static external economies of scale, dynamic increasing returns due to learning-by-doing with or without inter-industry spillovers, agglomeration economies with endogenous product varieties, R&D activities, and so on) with a wide range of results with different policy implications. The interested reader should start with a survey by Grossman and Helpman (1995).

Beyond technologies: policy-induced and institution-based comparative advantage

The Ricardian set-up has also been used to explain how the differences in national policies and institutions give rise to the patterns of comparative advantage, even in the absence of any inherent technology differences. In Copeland and Taylor (1994), the clean environment is a normal good, so that the rich North chooses a higher pollution tax than the poor South. As a result, the North (South) ends up having comparative advantages in less (more) polluting industries. In Matsuyama (2005), Costinot (2006), and Acemoglu, Antras and Helpman (2007), industries differ according to the severity of agency or contractual problems, and the country with a better (worse) institutional set-up to deal with these problems has comparative advantages in industries that are more (less) subject to these problems. One may view this line of research as an attempt to endogenize technology differences. Unlike the literature surveyed by Grossman and Helpman (1995), however, the main objective here is to look at the deeper or more fundamental causes of technology differences, rather than looking at the two-way causality between technology and trade.

Finally, because the Ricardian trade theory abstracts from the roles of factor endowment differences across countries and factor intensity differences across industries as the basis of

trade, it is an ideal set-up in which to isolate the roles of factor endowments and intensity differences that are unrelated to the basis of trade. For example, Matsuyama (2007a) uses a two-country Ricardian model to examine how factor intensity affects the extent of globalization and how globalization affects factor prices when certain factors are used more intensively in international trade than in domestic trade. The model is Ricardian in the sense that the patterns of comparative advantage are determined entirely by the exogenous technological differences. The factor proportions matter, however, because they determine the extent of globalization, as the effective trade costs vary with the relative endowments of the factor used intensively in international trade.

See Also

- comparative advantage
- globalization
- international trade theory
- terms of trade

Bibliography

Acemoglu, D., Antras, P. and Helpman, E. 2007. Contracts and technology adoption. *American Economic Review* 97, 916–43.

Acemoglu, D. and Ventura, J. 2002. The world income distribution. *Quarterly Journal of Economics* 117, 659–94.

Alvarez, F. and Lucas, R.E. Jr. 2004. General equilibrium analysis of the Eaton–Kortum model of international trade. Working Paper, University of Chicago.

Becker, G. 1952. A note on multi-country trade. *American Economic Review* 42, 558–68.

Copeland, B. and Taylor, S. 1994. North–South trade and the environment. *Quarterly Journal of Economics* 109, 755–87.

Costinot, A. 2006. On the origins of comparative advantage. Working paper, UCSD.

Dornbusch, R., Fischer, S. and Samuelson, P.A. 1977. Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67, 823–39.

Eaton, J. and Kortum, S. 2002. Technology, geography, and trade. *Econometrica* 70, 1741–79.

Flam, H. and Helpman, E. 1987. Vertical product differentiation and North–South trade. *American Economic Review* 77, 810–22.

Grossman, G. and Helpman, E. 1995. Technology and trade. In *Handbook of International Economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: North-Holland.

Hicks, J.R. 1953. An inaugural lecture. *Oxford Economic Papers* 5, 117–35.

Hummels, D., Ishii, J. and Yi, K.-M. 2001. The nature and growth of vertical integration in world trade. *Journal of International Economics* 54, 75–96.

Itoh, M. and Kiyono, K. 1987. Welfare enhancing export subsidies. *Journal of Political Economy* 95, 115–37.

Jones, R.W. 1979. Technical progress and real incomes in a Ricardian trade model. In *International Trade: Essays in Theory*, ed. R.W. Jones. Amsterdam: North-Holland.

Krugman, P. 1986. A ‘technology-gap’ model of international trade. In *Structural Adjustment in Developed Open Economies*, ed. K. Jungenfelt and D. Hague. London: Macmillan.

Matsuyama, K. 1996. Why are there rich and poor countries?: symmetry-breaking in the world economy. *Journal of the Japanese and International Economies* 10, 419–39.

Matsuyama, K. 2000. A Ricardian model with a continuum of goods under nonhomothetic preferences: demand complementarities, income distribution and north-south trade. *Journal of Political Economy* 108, 1093–120.

Matsuyama, K. 2005. Credit market imperfections and patterns of international trade and capital flows. *Journal of the European Economic Association* 3, 714–23.

Matsuyama, K. 2007a. Beyond icebergs: towards a theory of biased globalization. *Review of Economic Studies* 74, 237–53.

Matsuyama, K. 2007b. Productivity-based theory of manufacturing employment declines: a global perspective. Working paper, Department of Economics, Northwestern University.

Matsuyama, K. 2007c. Uniform versus non-uniform globalization. Working paper, Department of Economics, Northwestern University.

Sanyal, K.K. 1983. Vertical specialization in a Ricardian model with a continuum of stages of production. *Economica* 50, 71–8.

Stokey, N. 1991. The volume and composition of trade between rich and poor countries. *Review of Economic Studies* 58, 63–80.

Wilson, C.A. 1980. On the general structure of Ricardian models with a continuum of goods. *Econometrica* 48, 1675–702.

Yanagawa, N. 1996. Economic development in a Ricardian world with many countries. *Journal of Development Economics* 49, 271–88.

How to cite this article

Matsuyama, Kiminori. "Ricardian trade theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000276> doi:10.1057/9780230226203.1441

The New Palgrave Dictionary of Economics Online

Ricardo, David (1772 – 1823)

Terry Peach

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses the life and work of David Ricardo. The first section provides a comprehensive overview of his life, his contributions to political economy and his political activities. This is followed by more detailed consideration of his monetary writings (including the ‘law of markets’ and ‘comparative advantage’), his early writings on profits and the ‘corn model’ interpretation, the labour theory of value, the ‘new view’ and ‘neoclassical’ interpretations of his work, and his Sraffa-inspired interpretation as a ‘classical’ economist.

Keywords

absolute advantage; absolute and exchangeable value; Bank of England; barter; Barton, J.; Bentham, J.; Bullion Committee; Bullion Controversy; classical distribution theories; classical economics; classical theory of money; comparative advantage; competition of capitals; convertibility; Corn Laws; corn model; credit; De Quincey, T.; deflation; depreciation; differential rent; diminishing returns; direct and indirect labour; fixed and circulating capital; general equilibrium; inflation; ingot plan; invariable measure of value; King, P., Lord; labour theory of value; law of markets; Malthus, T. R.; Malthus's population theory; market wage; McCullough, J. R.; Marshall, A.; Mill, J.; Mill, J. S.; money supply; money wage; national debt; natural price; natural wage; neoclassical economics; paper money; political economy; profit and profit theory; public debt; pure labour theory of value; rate of interest; rate of profit; real wage; rent; Ricardian equivalence theorem; Ricardo, D.; Say, J.-B.; Say's equality; Say's identity; seigniorage; shared incidence principle; Smith, A.; specialization; specie-flow mechanism; Sraffa, P.; Sraffian economics; stationary state; subsistence and subsistence wages; supply and demand; surplus; Thornton, H.; Torrens, R.; utilitarianism; velocity of circulation

Article

David Ricardo was one of the most outstanding political economists of the 19th century and one of the most influential of all time. Born on 18 April 1772 at 36 Broad Street Buildings in the City of London, he was the third of 15 surviving children of Abraham Israel Ricardo (1733–1812) and his wife, Abigail Delvalle (1753–1801). Abraham's family were Sephardic Jews who had emigrated from Portugal to Holland at the end of the 16th century. His father (David's grandfather) is described as ‘a non-official broker in funds and stocks’ on the Amsterdam exchange (Heertje, 2004, p. 283). Abraham also became a stockbroker, first in Amsterdam and then in London, where he moved in 1760. He married Abigail (from an established London Sephardic family of tobacco and snuff merchants) in 1769 and was granted citizenship in 1771. As related by David's brother, Moses Ricardo, their father

> was a man of good intellect, but uncultivated. His prejudices were exceedingly strong; and they induced him to take the opinions of his forefathers in points of religion, politics, education &c., upon faith, and without investigation. Not only did he adopt this rule for himself, but he insisted on its being followed by his children; his son [David], however, never yielded his assent on any important subject, until after he had thoroughly investigated it. It was perhaps in opposing these strong prejudices that he was first led to that freedom and independence of thought for which he was so remarkable. (Ricardo, 1951–55, *Works*, X, p. 5; hereafter ‘*Works*’)

According to Moses Ricardo, the young David was allotted a ‘common-school education’, typical of ‘those who are destined for a mercantile line of life’ (*Works* X, p. 3), the emphasis being on reading, writing and arithmetic. Less typically, at the age of 11 David was sent to Amsterdam for two years. Details of the visit are sketchy. It has been suggested that he attended the famous Talmud Tora, attached to the Portuguese Synagogue in Amsterdam, although recent scholarship has favoured a more mundane account in which he was privately tutored in the ‘common-school’ subjects with the addition of French and Spanish (Heertje, 2004). Following his return to London, David's full-time education continued only until he reached the age of 14, when he began working for his father as a clerk and messenger on the London Stock Exchange, although he was allowed ‘any masters for private instruction whom he chose to have’ during his spare time (Moses Ricardo, *Works* X, p. 3). He was later to complain bitterly of years of neglected education ‘at the most essential period of life’ (*Works* VII, p. 305), to which he frequently attributed his difficulties in written composition.

In 1792 the Ricardo family moved to Bow, close to the house of Edward Wilkinson, a Quaker and surgeon. Before long, David was romantically involved with Wilkinson's daughter, Priscilla Ann (1768–1829), whom he married on 20 December 1793. The young couple was promptly disowned by both sets of parents. Although the breach with the Wilkinson's was short-lived, it is said that David neither spoke to nor saw his mother again. He was also disinherited and removed from his father's business (he was reconciled with his father after his mother's death). The marriage was the occasion of his breach with the Jewish faith, which, according to Sraffa, was ‘the culmination of a gradual estrangement [with Judaism] … in progress for some time before’ (Sraffa, *Works* X, pp. 38–9). He was subsequently to attend the meetings of the non-conformist Unitarians.

The break with his father was not to prove an insurmountable obstacle to David's financial prospects. With the assistance of City friends, he embarked on his spectacularly successful career as a jobber on the

stock market and a loan contractor for government stock. Before long he had amassed a considerable fortune, and in 1815, the year in which he made his single most profitable transaction, he began a gradual retirement from business. The total value of his estate at death has been estimated at between £675,000 and £775,000, roughly equivalent to more than £500 million ($950,000,000) at 2006 prices. As Ricardo's wealth grew, so too did his family and social standing. Eight children were born between 1795 and 1810. From 1812, the family's prestigious London address was 56 Upper Brook Street, Grosvenor Square. To this was added in 1814 Ricardo's country seat of Gatcombe Park near the small village of Minchinhampton, where he is reported to have financed almshouses, endowed a school, provided an infirmary and started a savings bank. His petition for his own coat of arms was also granted in 1814. Having entered the squirearchy (he was High Sheriff of Gloucestershire in 1818), acquired his reputation as a leading intellectual and political economist and become a prominent Member of Parliament (he took his seat in 1818), his company was sought increasingly by luminaries of the aristocracy, the political classes and the intelligentsia. Ricardo was highly gratified by his success, no more so than as a recognized authority on his favourite subject, namely, as he described it, ‘political economy’.

It is said that Ricardo's interest in political economy was stimulated by chancing upon a copy of Adam Smith's *Wealth of Nations* in a travelling library while on a visit to Bath. Prior to this time, we are told by his brother that a predilection for subjects of an abstract and general nature had led to a leisurely interest in science, including mathematics, chemistry, geology and mineralogy (in 1807 he was a founding member of the London Institution for the Advancement of Literature and the Diffusion of Useful Knowledge, which was charged with promoting science, as well as literature and the arts, and in 1808 he joined the Geological Society of London). Yet, although his interest was awakened, to the point where he become an avid reader of the early articles on political economy in the *Edinburgh Review*, he was for several years too preoccupied with furthering his financial career to treat political economy as anything more that ‘an agreeable subject for half an hour's chat’ (as he later reminisced to his old friend Hutches Trower, *Works* VII, p. 246). The turning point came in 1809.

The free convertibility of paper currency into gold had been suspended under the Bank Restriction Act of 1797, following the run on the Bank that had been provoked by fears of a French invasion (in the context of the French wars of 1792–1815). In the aftermath of restriction, the market price of gold (in terms of the now unconvertible paper currency) had risen above the (fixed) mint price and the ‘exchanges’ had become ‘unfavourable’, so that premiums were now to be paid on bills of exchange drawn on overseas banks for the purposes of settling international debts. This gave rise to the first phase of the ‘Bullion Controversy’ (*c*.1797–1801), with contributions from writers including Henry Thornton and Lord Peter King. The controversy was concerned with the reasons for the depreciation of ‘paper’ relative to gold and the deterioration in the exchanges, with the ‘bullionists’ (as represented by King) arguing that the fault lay with the Bank of England for overissuing paper, while the ‘anti-bullionists’ (including Thornton) emphasized the role of special government payments overseas and poor domestic harvests, independently of the Bank's issues (this was the debate that dominated the early entries on political economy in the *Edinburgh Review*). After 1801, however, the price of gold fell back towards the mint price, the exchanges improved and the controversy duly subsided, to be revived in 1808 by a further marked depreciation of paper relative to gold (of around 20 per cent) and an accompanying fall in the exchanges. It was to this second phase of the controversy that Ricardo's first publication was directed, taking the form of an anonymous letter to the *Morning Chronicle* newspaper,

published 29 August 1809.

With uncharacteristic rhetorical excess (Ricardo intoned gravely about the ‘present evil’ of the depreciation of paper and the ‘disastrous consequences’ and ‘future ruin’ that might follow) the argument in the letter followed the standard bullionist position that the root cause of the ‘ills’ was the ‘over-issues of the Bank [of England]’, the remedy being (in a remark that anticipates his later ‘Ingot Plan’) that ‘the Bank be enjoined by Parliament gradually to withdraw … notes from circulation, without obliging them, in the first instance, to pay in specie’ (*Works* III, p. 21). The argument was developed in further letters to the *Morning Chronicle*, and in his (signed) pamphlets, *The High Price of Bullion* (1810–11) and the *Reply to Mr. Bosenquet* (1811). The *Bullion* essay was a straightforward elaboration of Ricardo's position (in which he acknowledged that he ‘can add but little to the arguments which have been so ably urged by Lord King’, *Works* III, pp. 51–2), with an appendix to the fourth edition in which he developed his plan to resume convertibility by requiring the Bank of England to pay on demand in bullion ingots (not specie) bank notes of the value of at least £20, the alleged advantages being that this would reduce the supply of domestic paper, prevent an excessive demand on the Bank for gold (since the demand for bullion ingots would be less than the demand for specie), and prevent the withdrawal of low face-value bank notes (although Ricardo was later to acknowledge that a secondary market in bullion could facilitate the exchange of small notes pro rata). The second pamphlet was a reply to criticisms of the parliamentary Bullion Committee Report (1810), with which Ricardo broadly agreed, and of Ricardo's own currency writings.

The contributions to the Bullion Controversy brought Ricardo to the attention of political and intellectual figures including Thomas Robert Malthus and James Mill (Ricardo was also in correspondence with Spencer Perceval, the Tory Prime Minister, and the opposition Whig leader, George Tierney). Both Malthus and Mill were to play critical roles in the development of Ricardo's subsequent career, although their influences were profoundly different. At the time of Ricardo's entry on the public stage, Malthus was a seasoned writer, the author of the *Essay on Population* and, arguably, the leading political economist of the day. Although he and Ricardo became, and remained, close friends, their relationship was marked by disagreement over many areas within the new ‘science’ of political economy. While Ricardo borrowed from some of Malthus's writings (including population theory and the theory of differential rent) other aspects of his work evolved dialectically from epistolary skirmishes with his contemporary.

Malthus introduced himself to Ricardo by letter in June 1811, by which time it seems that Mill was already considered a close friend and an ally on the bullion question. Mill had been an early contributor to the *Edinburgh Review*, but by this time his attention had turned to writing his *A History of British India* and it seems unlikely that he had much influence over the content of Ricardo's political economy with the exception of the ‘law of markets’ (in short, the doctrine that ‘supply creates its own demand’). But that is not to detract from Mill's influence on Ricardo in other respects. Mill advised, encouraged, cajoled and even (only semi-humorously) bullied the ever-reticent Ricardo, who almost certainly would not have completed his major work without Mill's incessant prodding (see J.S. Mill, 1873, p. 42). It was also James Mill, as associate and disciple of Jeremy Bentham (with whom Ricardo also became personally acquainted), who was to coach the initially sceptical Ricardo in political utilitarianism and persuade him to enter parliament.

Currency issues dominated the early Ricardo–Malthus correspondence, but by late summer of 1813 their attention began to turn to a new subject, the forces governing movements in the general rate of profit.

Up to this point Ricardo had robustly endorsed Adam Smith's ‘competition of capitals thesis’, according to which the general rate of profit is regulated by the intensity of competition in the labour market (determining movements in wage rates) and in the output market (determining prices). Exactly how Ricardo himself interpreted Smith's doctrine is not clear, nor is it possible to pinpoint the reason for the change of focus in the Ricardo–Malthus correspondence. However, it is certain that Ricardo was newly emphasizing the conditions of producing food as an influence on profits. His position was developed in lost ‘papers on the profits of Capital’, following which, in response to a clamour by the landed aristocracy for a revision of the old Corn Law, and the report in May 1814 of the Parliamentary Committee on the Corn Trade, his deliberations become more narrowly centred on the effects on profits of restrictions on the free importation of corn. The outcome was his *Essay on The Influence of a Low Price of Corn on the Profits of Stock; Shewing the Inexpediency of Restrictions on Importation* (1815; reprinted in *Works* IV).

The central argument of the *Essay* may be given as follows. On the assumption of an economy closed to the importation of foreign corn (the principal subsistence commodity and wage good), the increasing demand for corn from a growing population must be met either by the more intensive cultivation of land or by cultivating land that is less fertile or more disadvantageously situated relative to the final market. Either way, the expansion of output will encounter diminishing returns which in turn lead to a higher corn price, higher money wages and, therefore, a lower agricultural rate of profit. Only the landlords benefit, because they receive more *differential* rent: following Malthus, rent is the difference in return from the ‘best’ and the ‘worst’ land on the assumption that the return from the ‘worst’ land is sufficient only to give farmers the general rate of profit; ergo, landlords benefit as that difference increases. To complete the argument, the reduction in profitability is transmitted to capitalists generally by means of higher money wages. As for labourers, the argument appears to be that they might also suffer in consequence of a (‘temporary’) fall in labour demand, itself the result of lower profitability. Hence Ricardo's provocative conclusion that ‘the interest of the landlord is always opposed to the interest of every other class in the community’ (*Works* IV, p. 21).

The *Essay* was a transitional work in which Ricardo repudiated some of the fundamental tenets of the prevailing orthodoxy, as derived from Adam Smith and upheld by Malthus, but failed to supply a fully convincing logical alternative. It was James Mill who persuaded Ricardo to develop his ideas in the form of a major treatise. Two years later Mill's exhortations were rewarded with the publication of *On the Principles of Political Economy and Taxation* (1817; reprinted in *Works* I). However, before Ricardo could begin serious work on the *Principles* he had to fulfil his commitment to Pascoe Grenfell M.P., who had enlisted Ricardo's support for an assault on the Bank of England. Ricardo was more than happy to oblige (‘I always enjoy any attack upon the Bank’, *Works* VI, pp. 268–9), and the result was his *Proposals for an Economical and Secure Currency; with Observations on the Profits of the Bank of England, as they regard the Public and the Proprietors of Bank Stock* (1816; reprinted in *Works* IV). In language suggested by James Mill, Ricardo lamented that ‘a great and opulent body like the Bank of England’ should ‘wish to augment their hoards by undue gains wrested from the hands of an overburthened people’ (*Works* IV, p. 93). It was intolerable that a mere ‘company of merchants’ should make vast profits by overcharging on the management of the public debt and other public business, through their ‘seignorage’ on the issue of paper money and by reducing their unprofitable stock of bullion (as they were enabled to do by the Restriction Act). For the longer term (after the expiry of the Bank Charter in 1833), Ricardo's preferred solution was to strip the Bank of its management of the

money supply, which he would entrust to ‘commissioners responsible to parliament only, the state’ (*Works* IV, p. 114) (this plan was developed in the *Plan for the Establishment of a National Bank* [1824], drafted by Ricardo in 1823, reprinted in *Works* IV). For the shorter term, he suggested that the government should seek more favourable terms for the management of the debt. Above all, however, he again advocated a swift return to a fully convertible paper currency, to be achieved by his Ingot Plan. The result would indeed be an ‘economical and secure currency’ which, along with its other advantages (of cheapness in comparison with a fully metallic currency and of stability by constraining movements in the market price of gold) would facilitate short-term, compensating changes in the money supply in response to fluctuations in the availability of credit.

With the *Proposals* dispatched to the printers, the way was open for Ricardo to commence work on his *Principles*; or, be more precise, it was *almost* so, for he still had to contend with a hectic social life, recurring bouts of lethargy and defeatism, continuing business interests, the demands of a large family and a ‘temptation of being out in the air in fine weather’ (*Works* VI, p. 263). Fortunately for posterity, the summer of 1816 offered very few outdoor temptations and Ricardo dedicated himself to his task. The *Principles* was published on 19 April, 1817. It was the result of little more than six or seven months' sustained activity on Ricardo's part.

The ‘principal problem in Political Economy’ is defined in the *Principles* as the determination of the ‘laws’ which regulate ‘the natural course of rent, profit, and wages’ over time. These issues had been addressed in the *Essay* and, indeed, the *Principles* was initially conceived as an *Essay* writ large. In the process of writing the later work, however, its scope was enlarged in previously unforeseen ways as Ricardo developed his ideas. The result was a volume comprising 31 chapters, covering not only the ‘laws’ governing rent, profit and wages, but also a labour theory of value, a theory of international comparative advantage, monetary theory, several chapters devoted to ‘the influence of taxation on different classes of the community’, and strictures on the writings of predecessors and contemporaries.

The ‘core’ theoretical analysis as it relates to ‘the natural course’ may be summarized as follows.

In terms of the newly adopted ‘pure’ labour theory of value, (changes in) the exchangeable value of competitively produced, freely reproducible commodities are determined exclusively by (changes in) the *quantities* of labour expended on their production, where the relevant quantity of labour is the *greatest* quantity expended per unit of output sold. The theory applies only when commodities exchange at their *natural prices*, defined by uniform wage and profit rates (rent is excluded as a component of price, as explained shortly). In addition, it is assumed that one domestically produced commodity, gold (not to be confused with its real-world namesake), serves as the ‘invariable standard’ (the *numéraire*) in terms of which all prices are expressed, its ‘invariability’ defined in terms of a given and unchanging labour input per unit of its output. It follows that any change in a commodity's gold-denominated natural price is an exact reflection of a corresponding change in the labour expended on *its* production. This theory of value was used by Ricardo beyond the first chapter of the *Principles*.

Next, there is the theory of differential rent, derived from Malthus. As Ricardo explained, the relevance of the theory is not confined to agriculture but applies whenever units of the same (homogeneous) class of commodity are produced by different quantities of labour. If all units sell at the same natural price, determined by the *greatest* labour input per unit; and, if the rate of profit from the sale of the unit requiring the greatest labour input is equal to the general (uniform) rate; then, an additional surplus revenue will be earned on units requiring a lower labour input, and it is this additional surplus that constitutes (differential) *rent*. (If we assume with Ricardo that capitalist producers, in agriculture and

elsewhere, are profit-maximizers, they will *always* extend production to the point where revenue from the sale of the incremental output, requiring the greatest labour input, is sufficient *only* to yield the general rate of profit; moreover, this greatest labour input *must* determine price, otherwise the general rate of profit could not be received and the output would never be produced.)

On wages, Ricardo introduces the distinction between the market price of labour (or market wage) and the natural price (or natural wage), the latter defined (in money terms) as ‘that price which is necessary to enable the labourers, one with another, to subsist and to perpetuate their race, without either increase or diminution’ (*Works* I, p. 93). The price that *is* necessary depends on the ‘real’ natural wage: on ‘the quantity of food, necessaries, and conveniences [which] become essential … from habit’ (*Works* I, p. 93). Habits may change over time (perhaps under the influence of education, as Ricardo hoped) but, for analytical purposes, the natural (real) wage is a datum. In the event that the market wage is above or below the natural wage, a Malthusian-style population mechanism is triggered: population expands (or contracts) and the market wage returns to the natural level.

To turn to profits, the eponymous chapter in the *Principles* is chiefly a revision of the central argument from the *Essay*, although it does contain the ingredients for a more general theory of the rate of profit. Now in terms of the labour theory of value, the attempt to expand the output of corn in a closed economy encounters diminishing returns in the form of a greater labour input per unit of output; hence, the (natural) price of corn rises proportionately. This in turn increases money wages, because corn enters the given real wage. The rate of profit (calculated with reference to the output produced by the greatest labour input) must therefore fall (since the rise in the natural price of corn is proportionate only to the increase in the *quantity* of labour expended on its production and does not reflect the increased *cost* of that labour) and, by the reasoning explained above, differential rent increases. Natural prices outside agriculture are either unchanged or, if corn is required as a material input, rise only to reflect the increase in the labour expended on their production; hence, the fall in the agricultural rate is communicated to other sectors by an increase in money wages. Perforce, the ‘natural course or rent, profit, and wages’ is for rent to increase, the rate of profit to fall and (money) wages to rise, although it must be stressed that this ‘prediction’ is *entirely* contingent on a host of assumptions and should not be taken as evidence of a gloomy or pessimistic attitude by Ricardo to Britain's economic prospects (such an inference, although commonly made, could not be further from the truth).

Finally, the more general theory from Ricardo's analysis is that (changes in) the rate of profit depend exclusively on (changes in) the labour expended on the production of the given real wage where, to borrow J.S. Mill's distinction, the ‘labour expended’ covers the ‘direct’ labour and the ‘indirect’ labour expended on the production of the non-labour inputs to the production of wage-goods. Provided one grants him his assumptions (of a labour theory of value, a given real wage and known labour conditions of production), Ricardo had thus produced a strong candidate for the first logically coherent theory of the determination of the general rate of profit in the history of economic thought.

There was a great deal riding on the success of the *Principles*, not just Ricardo's growing reputation as a political economist. Mill had suggested in 1815 that Ricardo should enter Parliament, a suggestion from which the latter had recoiled with horror. One year later he was becoming more amenable to Mill's plan, writing to his friend: ‘If my book succeeds … perhaps my ambition may be awakened, and I may aspire to rank with senators’ (*Works* VII, p. 113). Much to Ricardo's relief, the book did succeed to an extent far surpassing his self-deprecatory expectations.

Ricardo entered Parliament on 26 February 1819 as the independent member for the rotten borough of

Portarlington in Ireland: a constituency which he never visited, with 12 or so electors in the ‘pocket’ of Lord Portarlington to whom Ricardo had advanced £25,000 as a loan on the mortgage of his estates. Ricardo availed himself of every opportunity to educate the House of Commons in the ‘true principles of political economy’. These principles dictated the gradual repeal of trade restrictions generally and the of the Corn Law in particular; the gradual repeal of the Poor Laws; the repayment of the National Debt (his heroic proposal to replay the debt over two or three years by the imposition of a property tax was met with widespread incredulity); minimal taxation and a balanced budget; and a return to a convertible currency. With the signal exception of convertibility (Peel's Bill of 1819 for the Resumption of Cash Payments owed much to his proposals), Ricardo mostly found himself on the losing side, but that did nothing to shake his convictions. His parliamentary contributions are testimony to his belief in political economy as a subject of direct empirical relevance (the view of Ricardo as a pure theorist is a travesty). They also mark him as a zealous advocate of a free-market capitalist system with minimal government interference, who believed that Great Britain ‘would be the happiest country in the world, and its progress in prosperity would be beyond the power of imagination to conceive, if we got rid of two great evils – the national debt and the corn laws’ (*Works* V, p. 55). Additionally, he spoke out on a range of ‘liberal’ issues including religious tolerance, slavery, freedom of speech and the right to petition. He also aligned himself with the ‘radical’ cause for the reform of parliament.

The contention that ‘good government’ would not be achieved without a reform of parliament had been put to Ricardo by James Mill in 1815 but was at that time rejected on the grounds that Mill exaggerated the ‘sinister interest’ of politicians in pursuing their own selfish interests and undervalued the corrective influence of enlightened public opinion. Three years later Ricardo's position had changed. Partly as a result of Mill's bombardment of Ricardo with ‘radical’ messages, partly because of his growing conviction that the Tory government was failing to pursue ‘right measures’, and after reading Jeremy Bentham's *Plan of Parliamentary Reform*, Ricardo was won over to the ‘radical’ cause. As he came to argue, ‘good government’ – government ‘administered for the happiness of the *many*, and not for the benefit of the *few*’ (*Works* VII, p. 299) – required that politicians should ‘legislate for the public benefit only, and not … attend to the interests of any particular class’ (*Works* VIII, p. 275); yet, under present arrangements, politicians fell prey to the interests of particular classes, particularly the landed class; hence the necessity for reform. However, Ricardo's proposals fell some way short of those of his ‘radical’ contemporaries. The introduction of the secret ballot was, for him, an almost sufficient basis for securing good government under existing circumstances, although he did make a case for triennial parliaments and a modest extension of the franchise to include householders. He might therefore be described as a moderate reformer in the utilitarian tradition of Bentham and Mill.

The infamous proposal for the speedy repayment of the national debt was also presented to the public in an invited article on the *Funding System* (1820), written in autumn 1819 for publication in the *Supplement to the Encyclopaedia Britannica*. This article is noteworthy for its exposition of what has come to be known, misleadingly, as the ‘Ricardian equivalence theorem’. To take Ricardo's own argument, suppose that a war involves the expenditure of £20 million. This can be financed either by raising £20•m in taxes or by borrowing £20•m and repaying by taxes £1•m per annum in perpetuity (at an assumed annual interest rate of five per cent), or by borrowing the £20•m and (for example) repaying by taxes £1.2•m per annum, which would clear the interest (at five per cent) and the initial £20•m over a period of 45 years. ‘In point of economy’, as Ricardo stated, ‘there is no real difference in either of the modes’, because the present value of £1•m per annum in perpetuity or £1.2•m over 45 years, both at five

per cent annually, *is* £20•m, hence the idea of *equivalence*. But, he continued, ‘the people who pay the taxes never so estimate them, and therefore do not manage their private affairs accordingly’: the different modes are *not* equivalent because, according to him, individuals are prone to undervalue the true cost of repaying a loan over time. That being so, Ricardo's proposal was for the pay-as-you-spend mode of financing which, he believed, would make people ‘less disposed wantonly to engage in an expensive contest [namely war], and if engaged in it … be sooner disposed to get out of it’ (*Works* IV, p. 186).

While Ricardo was writing his *Funding System*, his friend Malthus was putting the finishing touches to his own *Principles*, published in April 1820, which contained an unsparing critique of Ricardo's central doctrines (Malthus's *Principles* together with Ricardo's comments are reprinted in *Works* II). Of all Malthus's criticisms, those levelled at Ricardo's treatment of value were the most acute, thus prompting Ricardo to a major revision of his first chapter for the third edition of his *Principles* (1821) (a lightly revised second edition of the *Principles* had been published in 1819). In addition to the defence against Malthus, the third edition is distinguished by a new chapter ‘On Machinery’ in which Ricardo, stimulated by the work of John Barton, famously declared that ‘the opinion entertained by the labouring class, that the employment of machinery is frequently detrimental to their interests, is not founded on prejudice and error, but is conformable to the correct principles of political economy’ (*Works* I, p. 392). To avoid misunderstanding, although there may be a very distant family resemblance between Ricardo's analysis and the standard ‘neoclassical’ case of factor substitution in response to changes in factor prices within a timeless framework, a principal difference is that Ricardo was describing a process *over time*, in which the substitution of machinery for labour was likely only to apply to new ventures. Nor did his analysis end there, since the capitalists were expected to expand accumulation in consequence of their higher profits, so tending to reverse the fall in the demand for labour (and wages).

Following the third edition of the *Principles*, Ricardo's next and last publication within his own lifetime was *On Protection to Agriculture* (1822; reprinted in *Works* IV): a veritable tour de force, written in little more than three weeks.

A new Corn Law had been passed in 1815 which prohibited the importation of foreign corn until the price had been at least 80 shillings per quarter for six weeks, by which time the ports could be opened to duty-free importation. Imports had been triggered in 1817–19, with prices first rising too 111 shillings in June 1817 and then (under the pressure of importation, followed by good domestic harvests) falling steadily to 55 shillings in the second half of 1820. After more bumper harvests, 1822 then witnessed the lowest average corn prices since 1792, with a fall to 34 shillings in November. The ‘agricultural distress’ was widespread and severe, and the powerful landed interest turned to Parliament for assistance. A parliamentary committee was established to investigate the causes and possible remedies for the distress, with Ricardo as one of its members.

Ricardo was not optimistic that Parliament would, or could, shift its position towards a free trade in corn; as he wrote in correspondence, ‘I have no hope of good measures being adopted, the landlords are too powerful in the House of Commons to give us any hope that they will relinquish the tax which they have in fact contrived to impose on the rest of the community’ (*Works* IX, p. 158). He was proved right. *On Protection* was his response to the protectionist report of the committee, in which he maintained his position that free trade was the only long-term solution while proposing a revised version of the 1815 Act (with measures for dampening price fluctuations). The pamphlet is also distinguished by sharp restatements of Ricardo's central doctrines, a wealth of detailed empirical analysis and a pungent defence

of his own position with regard to Peel's Bill of 1819 for the resumption of cash payments. *On Protection* shows him at the peak of his career, a true master of his subject and a political economist in the most rounded sense.

In the summer of 1822 Ricardo embarked with his family on a four-month grand tour of Continental Europe. Upon his return he resumed his hectic life as an active parliamentarian, attended meetings of the Political Economy Club (which he had co-founded in 1821), drafted his plan for an independent national bank, and continued with his deliberations on ‘value’. He was also looking forward to hosting a visit to Gatcombe from his old friend Hutches Trower, to whom he wrote: ‘we shall walk and ride, we will converse on politics, on Political Economy, and on Moral Philosophy, and neither of us will be the worse for the exercise of our colloquial powers (*Works* IX, p. 377). But it was not to be. On 11 September 1823 Ricardo died from the effects of an abscess in the middle ear. He was buried at Hardenhuish Park, Wiltshire, on the estate of his daughter, Henrietta, and her husband, Thomas Clutterbuck.

The newspaper obituaries of the time were lavish in their praise of Ricardo's achievements, both as a political economist and as a ‘Senator’ (see Peach, 2003). By his friends, he was applauded for having virtually revolutionized economic theory, not merely for its own sake but as means of guiding government policy and thus promoting the ‘general happiness’ of society. Of course, his critics were to asses his contributions less kindly, but in producing what was arguably the first coherent supply-side analysis of value, distribution and growth – never mind anything else – his place in doctrinal history was assured.

The following sections consider in more details various aspects of Ricardo's work and a selection of the main interpretative disputes that continue to surround it.

Monetary contributions, the law of markets and comparative advantage

As Peake (1978, p. 31) rightly observed, ‘Ricardo's total productive output was dominated by monetary questions’, and it was in this area that he had the greatest practical influence in his own lifetime.

A simple approximation to Ricardo's ‘model’ includes the following assumptions: the domestic currency consists entirely of paper money (‘paper’) issued by a central bank; money prices are a function of the supply of paper (*ceteris paribus*); the bank allows the free convertibility of paper into gold on demand at a permanently fixed mint price, initially equal to the globally determined market price in terms of paper; and all profit-seeking economic agents have virtually perfect market information. Now suppose that the bank increases its supply of paper. Domestic commodity prices rise, as does the market price of gold in terms of paper. (Ricardo treated gold as just another commodity, a view that later ensnared him in the position that the exchangeable value of gold is determined by the labour expended on its production even though its value is *incessantly* fluctuating in response to changes in the volume and pattern of world trade.) Gold has become relatively cheaper (or ‘redundant’) because, by assumption, it may be purchased at the lower mint price. Profit-seeking agents therefore exchange paper for gold which, because of its new relative cheapness compared with other domestic commodities, is now exported in preference to those commodities in exchange for foreign produce. Hence, the supply of domestic paper contracts, domestic prices fall, the market price of gold (in paper) returns to the mint price and the *status quo ante* is restored.

Now suppose that paper is no longer freely convertible into gold at the mint price. As before, the supply

of paper is increased, but the ‘stimulus which a redundant currency gives to the exportation of the coin [namely gold] … cannot, as formerly, relieve itself’ (*Works* III, p. 78), so the market price of gold *remains* above the mint price. In addition, the paper cost of bills of exchange drawn on foreign banks will increase to reflect the fall in paper relative to gold (the foreign exchange becomes ‘unfavourable’). Hence (following Lord King) Ricardo's ‘two unerring tests’ of a depreciation in Bank-notes, ‘the rate of exchange and the price of bullion’ (*Works* III, p. 75). (Ricardo flatly rejected the measurement of depreciation either in terms of changes in the exchangeable value of gold for domestic commodities – because commodities ‘are continually varying in value’ among themselves – or in terms of subjectively perceived ‘enjoyment’, ‘because two persons may derive very different degrees of enjoyment from the possession of the same commodity’; *Works* IV, pp. 59, 61.)

As to the consequences of ‘depreciation’, the picture is mixed. Ricardo stressed that the rate of interest is determined by the rate of profit in the ‘real’ economy (by the ‘competition of capitals’ or by the conditions of producing wage goods, in the earlier and later writings respectively); and that the ‘trifling’ effect on the rate of profit (hence on output) of an increase in paper is confined to an interval ‘of momentary duration’ before money wages adjust to restore the (assumed) given real wage (*Works* III, pp. 91–2, 318–19, 329; *Works* V, p. 446; *Works* VI, pp. 16–17). This dominant position supports a (mostly) neutral money interpretation, but it also raises the question of why Ricardo became so exercised by depreciation if its real effects were insignificant. The answer is possibly to be found in his concern with the effects of rising prices on recipients of fixed money incomes, especially ‘monied men’ (see *Works* III, pp. 21, 95–6; *Works* VI, p. 68), regarded by Sayers (1953, p. 65) as a ‘shattering’ inconsistency with the view ‘that long-run effects come quickly and easily’. In addition, the later Ricardo was also to remark on the danger of an ‘easy’ inconvertible paper-money regime in facilitating speculation and over-trading (*Works* V, pp. 397, 446).

Whatever the economic costs of depreciation, Ricardo campaigned tirelessly for a resumption of convertibility at the pre-restriction mint price of gold. In his evidence before the Parliamentary Committees on Cash Payments (1819) he argued, with heroic simplicity, that, in the prevailing circumstances of a four per cent premium in the market over the mint price of gold, a reduction in paper currency of about four per cent would be sufficient to restore parity, with a consequent fall of domestic prices generally also of around four per cent (*Works* V, pp. 416–17). This objective could be achieved ‘in a few months’ (*Works* V, p. 396). However, ‘by a consideration of the fears which I think many people very unreasonably entertained’, he was ‘reconciled’ to a plan for the phased return to the old mint price over one or two years (*Works* V, p. 451). The logic, if not the detail, of Ricardo's argument was accepted by the committees, leading to Peel's Bill (1819) with its provision for a staged return to convertibility at the old par over a period from February 1820 to May 1823. Payments were to be made only in bullion ingots in the first two stages, in line with Ricardo's recommendation. Contrary to his proposals, however, the third stage gave the Bank the *option* of making payments in specie, while the fourth and final stage saw a return to *full* convertibility at par.

Ricardo regretted that Parliament had not adopted his plan in its totality, but was otherwise supportive of the bill. Certainly, he did not foresee the events that were to follow which led, on his 1822 estimation, to a ten per cent depreciation of paper, thus making ‘the reverting to a fixed currency as difficult a task to the country as possible’ (*Works* IX, pp. 140, 152). The fault lay not with his analysis, however, but with the Bank of England, who had (in his opinion) needlessly purchased large quantities of gold in anticipation of resumption, thus raising its market price *independently* of note issues.

Ricardo may have been justified in blaming the Bank, but he must also stand accused of reasoning *as if* his simple model had captured all relevant aspects of reality. He was aware, on one level, that nominal inflation or deflation was *not* determined exclusively by changes in the Bank's supply of paper, and that the market price of gold could differ from the mint price *independently* of changes in the domestic money supply. He had noted in his early monetary writings that the ‘regulator of prices’ must include not only the quantity of money, but also ‘the rapidity of its circulation’ and ‘the mass of commodities’ (*Works* III, p. 311); later (1815–16), in response to post-war economic conditions, he allowed that the quantity of money was also determined by the independent behaviour (in context, the bankruptcy) of the country banks, and he conceded that ‘bullionists’, himself included, had underrated the effects on the market price of gold from changes in world demand for the metal (*Works* VI, p. 344; *Works* IV, p. 62); finally, under hostile questioning from some members of the Parliamentary Committees on Resumption, he admitted the further qualification that changes in the general state of confidence could affect domestic prices by influencing the availability of credit, itself a substitute for currency (*Works* V, p. 419). Yet, for the most part (as with his ‘four per cent’ calculation, noted above), he argued as if these counteracting influences were nugatory to the point that they could be ignored completely. This was a treacherous foundation on which to build economy policy.

It was also Ricardo's habitual presumption that real-world economic actors behaved ‘rationally’ and it is for this reason that he was highly critical of the Bank for purchasing gold when (on his analysis) it was not in their interest to do so. The same presumption was at the heart of his version of the law of markets, described by Keynes as the (flawed) doctrine that ‘supply creates its own demand in the sense that the aggregate demand price is equal to the aggregate supply price for all levels of output and employment’ (Keynes, 1936, 21–2).

The ‘law’ is commonly attributed to Jean-Baptiste Say although its roots extend back to Adam Smith's *Wealth of Nations*. It seems likely, however, that it was derived by Ricardo from James Mill, who had sketched out the argument in his 1808 review of William Spence's *Britain Independent of Commerce* in the *Edinburgh Review*. It was first used by Ricardo in his early monetary writings to argue that foreign markets will never be so ‘glutted’ by British produce as to constrain further British exports when money becomes comparatively dearer (that is, the opposite of ‘redundancy’). Later, it was used to bolster the argument that the *only* cause of a permanent reduction in general profitability is an increase in the labour expended on the production of wage goods. It was also invoked by Ricardo in the distressed aftermath of the French wars to support his unshakably optimistic view that recovery was *always* imminent.

Ricardo's version of the ‘law’ may be reduced to the following propositions: first, commodities will continue to be produced only if they return at least the going general rate of profit; second, capitalist producers (and *only* capitalist producers) are not ‘for any length of time … ill-informed of the commodities which [they] can most advantageously produce’ (*Works* I, p. 290); third, the desire to consume *something* is ‘implanted in every man's breast; nothing is required but the means’ (*Works* I, p. 292); fourth, all money income is spent, either by the direct recipients or by those to whom the recipients lend (all) their unspent money income (there is no hoarding). If, to take the extreme case, commodities *always* exchange at natural prices (which implies that the producers earn precisely the going general rate of profit) and all income is *always* spent (on the same output), we have ‘Say's identity’ version of the law, defined by an excess demand for money of zero at all times. This is the version that Keynes attributed to Ricardo. Yet, although Ricardo's emphasis was always on equilibrium or long-period tendencies, he did (as he was forced to by external events) allow for strictly ‘temporary’ periods of

capital misallocation in which capitalists produce the ‘wrong’ commodities and demand money (to satisfy creditors) in excess of revenue. The ‘Say's equality’ version of the ‘law’, allowing for ‘temporary’ disequilibria of excess demand for money, would therefore better describe his position. Above all, however, what is most striking is his belief that the ‘law’ encapsulated real-world tendencies, to the extent that he condemned out of hand all proposals for relief works (because only the capitalists knew best how to allocate capital) and, ultimately, was left totally bemused by the scale and duration of the post-war distress (see *Works* VIII, p. 277).

Economists have been considerably more impressed by his statement of comparative advantage in the chapter ‘On Foreign Trade’ in the *Principles*. Following Ricardo's own example, assume two commodity bundles of cloth (x_1) and wine (x_2), both of which could in principle be produced in England or Portugal ([x_1, x_2] and $[x_1^*, x_2^*]$ respectively). To produce the bundles in England would require 100 labourers for cloth (a_1) and 120 labourers for wine (a_2); and to produce them in Portugal would require 90 labourers for cloth (a_1^*) and 80 labourers for wine (a_2^*). Portugal therefore has an absolute advantage in the production of *both* bundles. Labour (alias ‘capital’) is immobile internationally, trade initially takes place by way of real barter and, implicitly, bundles are produced under constant returns to scale in both countries.

As Ricardo states, it will be advantageous for England to specialize in making cloth and Portugal to specialize in wine, because both countries thereby obtain more of the other commodity per unit of their domestic labour than if they attempted to make it themselves. For example, if Portugal used 80 labourers to make cloth she could obtain only $0.89x_1$, but if she can exchange $1x_2^*$ (also the produce of 80 labourers) for $1x_1$, as Ricardo supposes, then it would be ‘advantageous for her to export wine in exchange for cloth’ (*Works* I, p. 135). Similarly, if England used 100 labourers to produce wine she could obtain only $0.833x_2$, so she also benefits by exchanging $1x_1$ (the produce of 100 labourers) for $1x_2^*$. Ricardo's example implies that the pattern of specialization is dictated by the ‘four magic numbers’ (Samuelson, 1972, p. 378), namely, $a_1 / a_2 < a_1^* / a_2^*$. However, the principal purpose of the analysis was not so much to illustrate comparative advantage per se, but to show that the ‘same [labour theory] rule which regulates the relative value of commodities in one country, does not regulate the relative value of the commodities exchanged between two or more countries’ (*Works* I, p. 133). If, as Ricardo supposes, there is a rate of exchange of $1x_1$ for $1x_2$, ‘England would give the produce of the labour of 100 men, for the produce of the labour of 80’: something that ‘could not take place between the individuals of the same country’ (*Works* I, p. 135).

It was also Ricardo's purpose to show that the introduction of money (gold) would leave the analysis unaffected: gold will be ‘distributed in such proportions amongst the different countries … as to accommodate [itself] to the natural traffic which would take place if no such metal existed, and the trade between countries were purely a trade of barter’ (*Works* I, p. 137). To give the flavour of the argument, suppose England and Portugal each produce both commodities and that the initial gold prices are $px_1 = px_1^*$ which, given the ‘magic numbers’, implies $p_2 > p_2^*$. Wine is therefore exported from Portugal to England and is paid for by gold. But, on Ricardo's quantity theory reasoning, the influx of gold to Portugal, and its efflux from England, will raise prices in the former country and reduce them in

the latter. Hence, the specie-flow mechanism ensures that $px_1 = px_1^*$ is unsustainable (the same would apply to $px_2 = px_2^*$, by similar reasoning) and that the price of Portuguese cloth must exceed the price of English cloth (just as the price of wine must be higher in England than in Portugal), thus leading to complete specialization. Contrary to Ricardo, however, 'the natural traffic which would take place if no such metal existed', hence relative world prices, are not unique (as he implied), with the range of possible outcomes defined by the condition of $a_1 / a_2 \leq p_1 / p_2 \leq a_1^* / a_2^*$.

Debate continues as to whether Ricardo was the true originator of the comparative advantage doctrine, or whether that accolade should be awarded to his contemporary, Robert Torrens (see Ruffin, 2002; 2005, for a recent case in Ricardo's favour). But, regardless of who may have crossed the line first, it is with Ricardo's name that comparative advantage has become indelibly linked.

Early writings on profit (1813 – 15) and the 'corn model' interpretation

In the introduction to his masterful edition of Ricardo's *Collected Works* it was suggested by Piero Sraffa that the early Ricardo had devised a model in which corn is the sole agricultural input and output, thus supplying a 'rational foundation' for the 'principle of the determining role of the profits of agriculture', putatively articulated by Ricardo in 1814 with the words 'it is the profits of the farmer which regulate the profits of all other trades' (*Works* VI, p. 104). By implication, when Ricardo wrote that agricultural profits 'regulate' other profits, he had intended a statement of *unique determination* in full awareness of the logically required assumptions. Sraffa revealed, however, that the corn model (or 'corn ratio theory of profits') was 'never stated by Ricardo in any of his extant letters and papers' although, on the basis of indirect textual evidence, he claimed that Ricardo 'must have formulated it' either in lost papers or conversation (*Works* I, p. xxxi). Later, with the publication of Sraffa (1960), it transpired that the corn model had additional significance as a simple precursor of Sraffa's own 'Standard system' in which corn is the sole 'basic commodity'; and Sraffa also disclosed that the interpretation was the outcome of his *own* theoretical work: 'it was only when the Standard system and the distinction between basics and non-basics had emerged in the course of the present investigation that the ['corn model'] interpretation of Ricardo's theory suggested itself as a natural consequence' (Sraffa, 1960, p. 93).

The corn model interpretation was widely embraced. With a beguiling pedagogical simplicity, it could 'explain' Ricardo's regulatory statements and his later development of the pure labour theory, with Malthus entering the story to remind Ricardo that agricultural capital does *not* consist entirely of corn, thus necessitating a new (labour) theory of value. However, beginning in the early 1970s with the work Samuel Hollander, doubts have increasingly been aired about the textual basis for the interpretation. What follows is the view of one such critic. (For a sample of critical interpretations, see Faccarello, 1982; Hollander, 1973; 1975; 1979; and Peach, 1984; 1993; 2001. The case for the defence has been made by Eatwell, 1975, and Garegnani 1982, among others.)

If Ricardo's writings are sifted for confirmation for the interpretation – in other words, if the corn model is *presumed* – then it is easy enough to find 'evidence' in its favour. Without that presumption, the picture is rather different. Thus, Ricardo's assertion in correspondence that the 'rate of profits and of interest must depend on the proportion of production to the consumption necessary to such production' (*Works* VI, p. 108) is said by Sraffa to be the 'nearest that Ricardo comes to an explicit

statement on these [corn model] lines' (*Works* I, p. xxxii). Yet, although it is *possible* to conceive of such a 'proportion' in material terms, the expression itself provides no evidence of the way it was conceived by Ricardo; moreover, very similar expressions had been used by him in his earlier monetary writings in which there is no question of him having adopted corn model assumptions. As for the regulatory statements, while it is *possible* to impose a corn model rationalization, the problem is in establishing that the same rationalization was applied by Ricardo. Here, too, the evidence is disobliging. The *Essay* (1815) is replete with such statements (for example, 'The general profits of stock depend wholly on the profits of the last portion of capital employed on the land'; *Works* IV, p. 21), but we can be sure they were not thought *by Ricardo* to depend on the corn model because he *explicitly* assumed heterogeneous inputs to agriculture (including 'buildings, implements, &c.'; *Works* IV, p. 10). Indeed, an arresting feature of the *Essay* is that its specious corn model appearance derives from the use of corn (alias wheat) to *value* the physically heterogeneous agricultural capital. It was Ricardo's 'failure' to *revalue* this capital as corn became more difficult to produce (in the initial agricultural phase of the argument) that drew Malthus's criticism and led, ultimately, to Ricardo's adoption of the labour theory, not an assumption that agricultural capital *comprises* of corn alone.

A question for those who reject the corn model is how the pre-*Essay* Ricardo could arrive at his 'regulatory' position if, as he believed at the time, a rise in the price of corn would be followed by a rise in prices generally. Samuel Hollander has conjectured that Ricardo may have invoked a monetary constraint, so that an agriculturally induced rise in money wages would not be passed on in higher prices. Alternatively, it may have been that his view of pricing was *integral* to the analysis: with price rises common to output *and* the heterogeneous inputs to agriculture, Ricardo might have reasoned that an increase in the capital–output ratio *must* reduce profitability. Admittedly, these alternative interpretations do not have the simple elegance and logical consistency of the corn model but, then again, the period 1813–15 was one in which Ricardo was struggling to establish new ideas within an inherited theoretical framework, much of which was later to be discarded. The existence of contradictions and unresolved theoretical issues during this period is unsurprising.

The labour theory of value

The unmodified or 'pure' labour theory of value (PLTV) was adopted by Ricardo in early 1816, on the basis of which he drafted material that would form the first seven chapters of the *Principles* (up to and including the chapter 'On Foreign Trade'). But then he discovered a source of modification to the PLTV resulting from differences in capital structure between production processes. At first, the discovery impeded his progress, but then it seems he had the inspiration to turn it to his advantage (so he thought) in the form of the 'curious effect': the iconoclastic demonstration that prices *fall* following a general rise in wages and consequent reduction in the rate of profit. What he did not do, however, was provide any justification for using the PLTV in the light of the 'curious effect' analysis.

A simple 'dated labour' framework may serve to illustrate Ricardo's position. Assume three commodities (x_1, x_2, x_3), each produced by ten units of homogeneous labour (L) applied over two discrete production periods (t–1, t), with the following conditions of production:
$10L_t \rightarrow x_1;\ 5L_t + 5L_{t-1} \rightarrow x_2;\ 10L_{t-1} \rightarrow x_3$. If we denote the uniform wage and profit rates as w and r, each taking period t values, the natural price equations for the commodities are:

$$px_1 = 10L_t \cdot w(1+r) \quad (1)$$

$$px_2 = 5L_t \cdot w(1+r) + 5L_{t-1} \cdot w(1+r)^2 \quad (2)$$

$$px_3 = 10L_{t-1}(1+r)^2. \quad (3)$$

A PLTV requires $px_1 = px_2 = px_3$, because each commodity is produced with the same quantity of labour. However, it is evident (with $r > 0$) that $px_3 > px_2 > px_1$. Moreover, if distribution (between *w* and *r*) changes, there will be price fluctuations *even though* labour inputs are unchanged. Thus, assume that x_1 is the *numéraire* commodity (so that $px_1 \equiv 1$); in principle, on the basis of (1) a new (lower) *r* can be calculated for a given rise in *w*, and these numbers may be entered into (2) and (3) to obtain the new natural prices of x_2 and x_3. With the ‘compounding’ (or magnification) of the effect of the lower *r* on px_2 and, even more so, on px_3, the result will be a fall of both prices (expressed in terms of x_1) with px_3 falling to the greater extent.

To relate the above to the chapter ‘On Value’ in the first two editions of the *Principles*, the differences in production conditions (or capital structure) are, in Ricardo's terms, a reflection of differences in (*a*) the durability of fixed capital; (*b*) the ratios of fixed to circulating capital; (*c*) the durability of circulating capital (added in the second edition), where the fixed–circulating capital distinction depends, essentially, on the time required to repay a capital expenditure (the longer the time, the more ‘fixed’ the expenditure). In the case given by Ricardo, the two extremes (corresponding to x_1 and x_3) are a commodity produced by unassisted labour in one year and a commodity produced by unassisted machinery that lasts 100 years. Then, taking the former commodity as his ‘invariable standard’ (alias *numéraire*), he calculates that a fall of seven per cent in the rate of profit would reduce the price of the latter by 68 per cent: a vivid illustration of the ‘curious effect’ (*Works* I, p. 60).

The effect implies, by Ricardo's own testimony, that the PLTV is subject to a (truly) ‘considerable modification’ from differences in capital structure, but the really curious feature of the first two editions

of the *Principles* is that the PLTV (used beyond the first chapter) had been undermined by its own author. This does, indeed, deserve the obloquy of a shattering inconsistency, and one not lost on Malthus, who, in his *Principles* (1820, *Works* II), employed the ingenious tactic of using Ricardo's own analysis to demonstrate the untenability of the PLTV. His criticisms hit home.

Ricardo comprehensively rewrote the chapter ‘On Value’ for the third edition of the *Principles*, newly adopting two strategies for the defence of the PLTV. First, he ruthlessly extirpated all the numerical examples that *had* suggested a ‘considerable modification’ to the PLTV and replaced them by others, according to which the ‘greatest effects which could be produced on the relative prices of … goods from a rise of wages, could not exceed 6 or 7 per cent’ (*Works* I, p. 36). Second, he introduced a new section ‘On an invariable measure of value’, where he indicated his desire to find a ‘perfect measure of value’ in terms of which prices would change *only* to reflect changes in the quantities of labour expended on production. This was tantamount to claiming that the discovery of the ‘perfect’ standard would itself sanction a PLTV, his problem being, however, that any commodity standard must be produced with *some* capital structure and, as he had demonstrated, the ‘unwanted’ price fluctuations are inescapable if capital structures differ. Hence his second-best solution of assuming that the standard is produced using an (unweighted) ‘average’ capital structure (cf. x_2, above), allegedly characteristic of ‘most commodities’: at least for *them* a PLTV would apply. There would still be a ‘curious effect’ with a fall in profitability, just as some commodities (such as our x_1) would now *rise* in price, but this was announced *sotto voce* (*Works* I, p. 46) and the effect was nowhere near as ‘curious’, at least in its magnitude, as it had been before.

Through this process of ‘double indemnification’ Ricardo had, *for the first time*, justified his use of the PLTV in explicit acknowledgement of the problems caused by differences in capital structures: either the differences are small and can be ignored, or (really a variation on the same theme) all the relevant commodities, including the standard, are part of a ‘general mass’ with the same capital structure. The ‘exceptions’ to the PLTV may therefore be ignored.

If we leave aside the dubious merit of the defence, its very inclusion is evidence that Ricardo was not retreating in his advocacy of the PLTV, contrary to claims by earlier commentators including J. Hollander (1904) and Cannan (1929, p. 177). That view was laid to rest by Sraffa, who opined that ‘the theory of edition 3 appears to be the same, in essence and in emphasis, as that of edition 1’ (*Works* I, p. xxxviii): a view that may itself be criticized for undervaluing the scale and significance of the changes (cf. Hollander, 1979, p. 217). But why was the theory so important to him?

One attraction is that it provided him with (in its own terms) a logically coherent framework for establishing his central theoretical propositions, particularly of the dependency of the general rate of profit on the conditions of producing wage goods.

A second possibility is that the theory appealed because of its (supposed) empirical relevance; hence Stigler's attribution to Ricardo of ‘an *empirical* labour theory of value, that is, a theory that the relative quantities of labour … are the dominant determinants of relative value’ (Stigler, 1958, p. 60; emphasis in original). However, although Ricardo did make empirical claims on behalf of theory (for which, it must be said, no evidence was adduced), those claims were arguably more a reflection of his commitment than its basis.

There was also an increasing tendency on Ricardo's part to identify the very *essence* of value with expended labour time. This ‘value’, referred to him at different times as ‘natural value’, ‘positive value’

and ‘real value’, was conceived as an attribute of *individual* commodities; hence the criticism, *fully accepted by Ricardo*, that he had moved beyond a purely relative usage of ‘value’ (as in Stigler's interpretation) and had turned it into something absolute (*Works* IX, p. 38). From this perspective, the role of the ‘perfect’ standard was to harmonize the ‘labour values’ with (relative) cost-of-production ‘values’ (or natural prices), since (changes in) the latter would become an exclusive reflection of (changes in) the former. As Ricardo forlornly conceded, however, ‘perfection’ is ruled out by unequal capital structures: the point he develops in *Absolute Value and Exchangeable Value* (*Works* IV, pp. 361–412), poignantly truncated by his final illness.

The analytical and ‘philosophical’ attractions are therefore central to understanding Ricardo's PLTV commitment. Of course, with the benefit of nearly two centuries of hindsight, it could be (and has been) argued that the labour theory can be jettisoned, to be replaced (say) with Sraffa's physically specified input–output equations. That argument may be formally correct, but we would no longer have *Ricardo's* theoretical and conceptual system. For him, the labour theory of value was both fundamental and indispensable.

The ‘new view’

The inappropriately styled ‘new view’ (anticipated by Cannan, 1893, pp. 247–53, 350, with modern restatements by Casarosa, 1978, Hicks and Hollander, 1977, and Hollander, 1990; 2001; 2002, among others) can be treated either as a stand-alone interpretation of Ricardo's treatment of wages or as part of a more far-reaching attempt to assimilate Ricardo's work to ‘neoclassical’ economics.

In the second great ‘rehabilitation’ of Ricardo (the first was J.S. Mill's attempt to have him reinstated as ‘the greatest political economist’: Mill, 1848, p. 397), Alfred Marshall applied his principle of ‘generous interpretation’ to distance Ricardo from the labour theory of value (by that time with its Marxian connotations) and absorb him within the mainstream intellectual tradition. Thus he averred that Ricardo had been ‘feeling his way’ towards a subjective utility analysis and that, despite appearances, he had attributed coordinate importance to supply *and demand* in the determination of natural prices (Marshall, 1920, Appendix I). Interestingly, however, Marshall's generosity deserted him when it came to Ricardo's treatment of wages, which he regarded as indefensible (Pigou, 1925, p. 413).

Most subsequent commentators have considered Marshall's' interpretation as far too generous, the prominent exception being Samuel Hollander, who goes even further in claiming to find a ‘fundamentally important core of general-equilibrium economics’ in Ricardo's work, implying a ‘strong continuity of doctrine’ between Ricardo's and later ‘neoclassical’ analysis (Hollander, 1987, pp. 6–7; cf. Morishima, 1989). As part of this ‘general equilibrium’ analysis, Ricardo had (allegedly) treated the wage rate as an *endogenous variable*, and it is this feature that is emphasized by the ‘new view’ interpretation.

The ‘non-wage’ aspects of the ‘neoclassical’ Ricardo may be dealt with briefly. First, with regard to utility (in the sense of subjective satisfaction), there is no question that it was treated as a *precondition* for exchangeable value and, in circumstances of fixed supplies, it was also conjectured by Ricardo that prices would be *proportionate* to ‘utilities’ (*Works* II, pp. 24–5; *Works* VIII, 276–7). However, there was no attempt by him to develop an analysis of *diminishing* marginal utility and, for the purpose of explaining exchangeable values (at natural prices), his emphasis was on *objective* determination by

quantities of labour time. As to the ‘coordinate’ influence of supply and demand, this confuses the *process* by which market prices tend to their natural levels (which does involve output variations and, therefore, a pre-‘neoclassical’ species of supply-and-demand reasoning) with the *determination* of the natural price levels: for Ricardo, the latter is *independent* of supply and demand, thus effectively denying any *theoretical* relationship between output and (labour) conditions of production (see, for example, *Works* VIII, p. 207). Finally, on ‘general equilibrium’, there is not a single developed instance of such an analysis in the entire corpus of Ricardo's writings. It is an interpretation obtained only by reconstructing his work and, in the process, obliterating his own hallmark emphasis on unidirectional relationships.

With the ‘new view’, however, there is at least a textual basis. The most compelling evidence is from three paragraphs in the chapter ‘On Wages’. The first paragraph opens thus: ‘In the natural advance of society, the wages of labour will have a tendency to fall, as far as they are regulated by supply and demand’, the assumption being that ‘the supply of labourers will continue to increase as the same rate, while the demand for them will increase at a slower rate’ (*Works* I, p. 101). Wages are therefore falling continuously in ‘the natural advance’ and only reach their ‘natural’ level (defined for a stationary population) in the terminal stationary state. However, ‘we must not forget, that wages are also regulated by the prices of the commodities on which they are expended’ (*Works* I, p. 101) and, particularly, by the rising price of corn (from diminishing returns on the land). Money wages therefore *rise* in the ‘natural advance’ but not by so much as to fully compensate the labourers for the rising corn price, so that real wages secularly decline as before. The effect of diminishing agricultural returns is in this way ‘shared’ between capitalists and labourers and has come to be known as the ‘shared incidence principle’.

There is no doubt that the new view passages exist, and there are also muted refrains of the analysis elsewhere in the *Principles* (*Works* I, pp. 215, 220). At the same time, the natural wage analysis – with the natural wage, defined for a stationary population, as the active centre of gravity for market wages in *all* stages of society – is by far the dominant analysis in the *Principles*; and, unlike the new view, it is the only one consistent with the repeated claim that real-wage variations are of only ‘temporary’ significance, particularly with regard to movements in the general rate of profit. Based solely on the *Principles*, the proposition that the new view represents Ricardo's *true* position is difficult to sustain. Malthus, for one, did not recognize Ricardo as a (kindred) new view theorist; hence the trenchant criticisms of Ricardo's natural wage analysis in his own *Principles* (Malthus, 1820. *Works* II, pp. 256–64). As Samuel Hollander (2007) has emphasized, however, Ricardo protested that he maintained ‘no other doctrine than that which has been well explained by Mr. Malthus' (*Works* II, p. 288). Yet he also reaffirmed his *own* definition of the natural wage (*Works* II, pp. 227–8), which is inexplicable if he truly agreed with Malthus (for whom Ricardo's natural wage would be irrelevant outside the stationary state). While it cannot be denied, then, that Ricardo was on some level sympathetic to the new view analysis, he was at no time an unequivocal exponent of that doctrine. Even in his later writings (such as *On Protection to Agriculture*, *Works* IV), in which the natural wage is not mentioned explicitly, the real wage is treated as a given and fixed entity, without a trace of the new view. It is also significant that Ricardo was never to criticize the writings of contemporaries, including their avowed representations of his own position, for the (universal) failure to include the new view (Peach, 2007). His own credentials as a new view theorist must therefore remain in considerable doubt.

Conclusion: Ricardo as a ‘classical’ economist?

Ricardo was to achieve great fame as a political economist during the tragically short period in which he wrote on the subject, although his ideas, especially his policy proposals, were often bitterly contested by critics of differing political and theoretical persuasions (see Peach, 2003). Following his death, his name, if not always his own doctrines, lived on through the writings of his ‘New School’ disciples, notably James Mill, Thomas De Quincey and the indefatigable J.R. McCulloch, and through the efforts of J.S. Mill. With the advent of ‘neoclassical’ economics, however, Ricardo's stock began to plummet, with Marshall's attempted rehabilitation to no avail. By the time Sraffa's edition of the *Collected Works* appeared in the 1950s, Ricardo's positive contribution was not uncommonly reduced to anaemic generalities such as the development of a ‘professional frame of mind’ or an ‘abstract deductive approach’, the onward progress of economic science having established the ‘inadequacy’ of much of his substantive work.

The *Collected Works* prompted a flurry of new scholarly interest in Ricardo, but it was only after the publication of Sraffa (1960) that he was subjected to his third major ‘rehabilitation’, this time not as a ‘mainstream’ economist (as with J.S. Mill and Marshall) but as a precursor of *Sraffa's* economics or, as related by Sraffa's followers, as a founder of the ‘classical’ (or ‘surplus’) tradition that Sraffa (1960) had revived.

The defining characteristics of ‘classical’ economics are alluded to in the Preface of Sraffa (1960) and amount to the assumption of *given* outputs and methods of production. The distribution between wages and profits may then be ‘solved’ by taking one distributive variable as given and calculating the other as a residual (or ‘surplus’). But how well does this apply to Ricardo's approach? At one level – the calculation of profit as a ‘surplus’ – it does so well enough. Where the problems arise is with the other attribution of given outputs.

According to Sraffa, ‘The “principal problem in Political Economy” was in [Ricardo's] view the division of the national product between classes and in the course of that investigation he was troubled by the fact that the size of this product appears to change when the division changes’ (*Works* I, p. xlviii). Hence, ‘the problem of value which interested Ricardo was how to find a measure of value which would be invariant to changes in the division of the product’ (*Works* I, p. xlviii); and, as Sraffa remarks parenthetically, Ricardo may have come close to solving his ‘problem’ with the ‘average’ standard adopted in the third edition of the *Principles*: ‘If measured in such a standard, the average price of all commodities, and their aggregate value, would remain unaffected by a rise or fall of wages’ (*Works* I, pp. xliv–xlv).

Several objections can be made against Sraffa's interpretation. First and foremost, it implies that Ricardo's ‘principal problem’ was with purely ‘notional’ redistributions of a *given national product* (that is, with given and unchanging outputs). However, Ricardo's own ‘principal problem’, as he defined it himself, was with the ‘natural course of rent, profit, and wages’ *over time*, and for the purpose of *his* investigation there will be at least one output, that of corn, that *cannot* be treated as given and unchanging in terms of its conditions of production. Second, as Ricardo clarified, *his* analysis of distribution was to be framed at the level of the individual firm, or farm, not in terms of social or ‘national’ aggregates. Third, there is no evidence that *he* envisaged a ‘price-balancing’ function for his ‘average’ standard (that is, to ensure constancy in the total value of national output); indeed, Ricardo's opinion was that *all* distributions-induced price changes are evidence of a ‘defect’ in the standard.

Ricardo was not a full-fledged ‘classical’ economist in the Sraffa mould. To describe him more loosely

as a ‘surplus theorist’ is unexceptionable, although by focusing on only one (albeit important) area of Ricardo's writings it is also a ‘thin’ characterization. Ricardo was a towering intellectual force whose work ranged over all the main areas of political economy. Forcing him into classificatory boxes of a later construction is a disservice to him and a hindrance to those who would seek to understand the full richness and extent of his historical significance.

See Also

- Sraffian economics

Bibliography

Cannan, E. 1893. *A History of the Theories of Production and Distribution in English Political Economy from 1776 to 1848*. London: Percival.

Cannan, E. 1929. *A Review of Economic Theory*. London: P.S. King & Son.

Casarosa, C. 1978. A new formulation of the Ricardian system. *Oxford Economic Papers* 1, 38–63.

Eatwell, J.L. 1975. The interpretation of Ricardo's *Essay on Profits*. *Economica* 42, 182–7.

Faccarello, G. 1982. Sraffa versus Ricardo: the historical irrelevance of the ‘corn profit’ model. *Economy and Society* 11, 122–37.

Garegnani, P. 1982. On Hollander's interpretation of Ricardo's early theory of profits. *Cambridge Journal of Economics* 6, 65–77.

Heertje, A. 2004. The Dutch and Portuguese background of David Ricardo. *European Journal of the History of Economic Thought* 11, 281–94.

Hicks, J. and Hollander, S. 1977. Mr. Ricardo and the moderns. *Quarterly Journal of Economics* 91, 351–69.

Hollander, J.H. 1904. The development of Ricardo's theory of value. *Quarterly Journal of Economics* 19, 455–91.

Hollander, S. 1973. Ricardo's analysis of the profit rate, 1813–15. *Economica* 40, 260–82.

Hollander, S. 1975. Ricardo and the Corn Profit Model: Reply to Eatwell. *Economica* 52, 188–202.

Hollander, S. 1979. *The Economics of David Ricardo*. London: Heinemann.

Hollander, S. 1987. *Classical Economics*. London: Blackwell.

Hollander, S. 1990. Ricardian growth theory: a resolution of some problems in textual interpretation. *Oxford Economic Papers* 42, 730–50.

Hollander, S. 2001. Classical economics: a reification wrapped in an anachronism? In *Reflections on the Classical Canon in Economics*, ed. E.L. Forget and S. Peart. London: Routledge.

Hollander, S. 2002. The canonical classical growth model: content, adherence and priority. In *Competing Economic Theories*, ed. S. Nisticò and D. Tosato. London: Routledge.

Hollander, S. 2007. Ricardo as a ‘classical economist’: the ‘new view re-examined: A reply to Dr Peach. *History of Political Economy* 39, 307–12.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Marshall, A. 1920. *Principles of Economics*. 8th edn. London: Macmillan, 1979.

Mill, J. 1808. Review of Britain Independent of Commerce, by W. Spence. *Edinburgh Review* 11, 429–49.

Mill, J.S. 1848. *Principles of Political Economy*. New York: Kelley, 1987.

Mill, J.S. 1873. *Autobiography*. London: Penguin, 1989.

Morishima, M. 1989. *Ricardo's Economics: A General Equilibrium Theory of Distribution and Growth*. Cambridge: Cambridge University Press.

Peach, T. 1984. David Ricardo's early treatment of profitability: a new interpretation. *Economic Journal* 94, 733–51.

Peach, T. 1993. *Interpreting Ricardo*. Cambridge: Cambridge University Press.

Peach, T. 2001. Hollander de Vivo and the ‘further evidence’ for the corn model interpretation of Ricardo: a conspiracy of silence? *Cambridge Journal of Economics* 25, 685–692.

Peach, T. 2003. *David Ricardo: Critical Responses*, 4 vols. London: Routledge.

Peach, T. 2007. Ricardo as a ‘classical economist’: the ‘new view’ re-examined. *History of Political Economy* 39, 293–306.

Peake, C.F. 1978. Henry Thornton and the development of Ricardo's economic thought. *History of Political Economy* 10, 193–212. In *David Ricardo: Critical Assessments* (second series), vol. 5, ed. J. Cunningham Wood. London: Routledge, 1994.

Pigou, A.C. 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Ricardo, D. 1951–55. *The Works and Correspondence of David Ricardo*, ed. P. Sraffa with the collaboration of M.H. Dobb, vols. I–X, Cambridge: Cambridge University Press.

Ruffin, R.J. 2002. David Ricardo's discovery of comparative advantage. *History of Political Economy* 34, 725–48.

Ruffin, R.J. 2005. Debunking a myth: Torrens on comparative advantage. *History of Political Economy* 37, 711–22.

Samuelson, P.A. 1972. The way of an economist. In *The Collected Papers of Paul A. Samuelson*, ed. R. C. Merton. Cambridge, MA: MIT Press.

Sayers, R.S. 1953. Ricardo's views on monetary questions. *Quarterly Journal of Economics* 67, 30–49. In *David Ricardo: Critical Assessments*, vol. 4, ed. J. Cunningham Wood. Beckenham: Croom Helm, 1985.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Stigler, G.J. 1958. Ricardo and the 93% labour theory of value. *American Economic Review* 48, 357–67. In *David Ricardo: Critical Assessments*, vol. 2, ed. J. Cunningham Wood. Beckenham: Croom Helm, 1985.

How to cite this article

Peach, Terry. "Ricardo, David (1772–1823)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000259> doi:10.1057/9780230226203.1439

The New Palgrave Dictionary of Economics Online

risk aversion

Jan Werner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An agent, perhaps an individual or a firm, is said to be *risk averse* if the agent prefers a deterministic outcome equal to the expectation of a risky outcome over that risky outcome. Risk aversion seems to be a common characteristic; introspection suggests as much. More importantly, it gives qualitative explanation to economic behaviour in many instances where risk is present. If individuals and firms were not risk averse, insurance markets would not exist. Needless to say, there are activities which are inconsistent with agents being risk averse. Gambling is perhaps the best example of such an activity.

Keywords

Arrow–Pratt theory of risk aversion; Arrow–Pratt measure of absolute risk aversion; Arrow–Pratt measure of relative risk aversion; certainty equivalent; concavity; dual utility; equity premium puzzle; expected utility theory; insurance markets; maxmin expected utility; multivariate risk; probability measures; random variables; rank-dependent expected utility; relative risk; representation of preferences; risk aversion; risk compensation; risk–return trade-off; Schur concave functions; state-dependent utility; stochastic dominance; von Neumann–Morgenstern utility function; Wald's criterion

Article

Arrow – Pratt theory of risk aversion

The classical theory of risk aversion, due to Pratt (1964) and Arrow (1965), is rooted in the expected utility theory of decision making. An agent's preferences are assumed to have an expected utility representation. The objects of choice are real valued random variables defined either on a finite or infinite set of states of the world with probabilities of states that may be either objective or subjective. The intended interpretation of a random variable is as an agent's risky wealth.
An agent whose expected utility representation of preferences is written $E[u(\tilde{x})]$, where u is the von

Neumann–Morgenstern utility function and *E* denotes the expectation (expected value), is *risk averse* if

$$E[u(\tilde{x})] \leq u(E(\tilde{x})) \quad (1)$$

for every risky wealth $\tilde{x}$. If (1) holds with strict inequality for every non-deterministic $\tilde{x}$, the agent is *strictly risk averse*. Jensen's inequality implies that, if utility function *u* is concave, the agent is risk averse. The converse is also true. Thus, the concavity of *u* is a necessary and sufficient condition for risk aversion. Moreover, strict concavity of *u* is a necessary and sufficient condition for strict risk aversion. Examples of strictly concave von Neumann–Morgenstern utility functions, commonly used in applied work, include the negative exponential utility $u(w) = e^{-\alpha w}$ with $\alpha > 0$, the logarithmic utility $u(w) = \ln(w)$, and the power utility $u(w) = \frac{1}{1-\alpha} w^{1-\alpha}$ with $\alpha > 0$, $\alpha \neq 1$.
It is useful to have a measure of the intensity of risk aversion. The most natural measure is *risk compensation*. It is by definition the amount $\rho(w, \tilde{z})$ of deterministic wealth one could extract from an agent in exchange for relieving her of zero-expectation risk $\tilde{z}$ at an initial deterministic wealth *w*,

$$E[u(w + \tilde{z})] = u(w - \rho(w, \tilde{z})). \quad (2)$$

A risk-averse agent has non-negative risk compensation for every zero-expectation risk, at every level of initial wealth. Risk compensation makes possible interpersonal comparisons of risk aversion and, for any agent, comparisons of risk aversion at different levels of her initial wealth. If risk compensation $\rho_1(w, \tilde{z})$ of an agent with von Neumann–Morgenstern utility function u_1 is greater than or equal to risk compensation $\rho_2(w, \tilde{z})$ of another agent with utility function u_2, for every deterministic wealth *w* and risk $\tilde{z}$ with $E(\tilde{z}) = 0$, then the agent with u_1 is said to be *more risk averse* than the one with u_2. An agent has increasing, decreasing or constant risk aversion if, for every zero-expectation risk, her risk compensation is increasing, decreasing or constant in *w*, for every $\tilde{z}$ with $E(\tilde{z}) = 0$.
Another measure of risk aversion is *certainty equivalent*. It is by definition the amount $c(\tilde{x})$ of deterministic wealth such that an agent is indifferent between this deterministic wealth and risky wealth $\tilde{x}$,

$$E[u(\tilde{x})] = u(c(\tilde{x})). \quad (3)$$

For a risk-averse agent, certainty equivalent is lower than the expectation of risky wealth. Since certainty equivalent and risk compensation are related by $c(w+\tilde{x}) = w - \rho(w, \tilde{z})$, these two measures can be interchangeably used in the Arrow–Pratt theory of risk aversion.
Although with considerable intuitive appeal, risk compensation is not all that practical. It is only implicitly defined in (2). The basic insight of Arrow and Pratt is that there is a simple measure of risk aversion which is in a certain sense equivalent to risk compensation, namely, the *Arrow–Pratt measure of absolute risk aversion*

$$A(w) = -\frac{u''(w)}{u'(w)}. \tag{4}$$

The negative of the second derivative $u''(w)$ is a mathematical measure of the degree of concavity of *u*. It is rescaled by the first derivative (assumed non-zero) which makes the measure invariant under any affine transformation of *u*. For 'small' risk $\tilde{z}$ with $E(\tilde{z}) = 0$, risk compensation $\rho(w, \tilde{z})$ equals approximately half the product of the variance $\sigma^2(\tilde{z})$ and the Arrow–Pratt measure at *w*,

$$\rho(w, \tilde{z}) \cong \frac{1}{2} A(w) \sigma^2(\tilde{z}) \tag{5}$$

as can be demonstrated using quadratic approximation of expected utility $E[u(w+\tilde{z})]$.
The important theorem of Pratt establishes an equivalence of the two measures as criteria for comparative risk aversion. For that theorem, let u_1 and u_2 be two strictly increasing von Neumann–Morgenstern utility functions, twice differentiable with continuous second derivatives.
Theorem: **(Pratt, 1964)**: The following conditions are equivalent:

1. (*a*) $A_1(w) \geq A_2(w)$ *for every w*.
2. (*b*) $\rho_1(w, \tilde{z}) \geq \rho_2(w, \tilde{z})$ *for every w and every random variable* $\tilde{z}$ *with* $E(\tilde{z}) = 0$,
3. (*c*) u_1 *is a concave transformation of* u_2*; that is,* $u_1 = f \circ u_2$ *for f concave and strictly increasing*.

There is a version of the Pratt's theorem which has equalities in (*a*) and (*b*) and an affine transformation in (*c*). This version implies that the Arrow–Pratt measure identifies the von Neumann–Morgenstern utility function up to an affine transformation. For example, the negative exponential utility is (up to an

affine transformation) the only strictly concave utility with constant absolute risk aversion. There is also a strict version of the Pratt's theorem with strict inequalities in (*a*) and (*b*) and a strictly concave transformation in (*c*).

A corollary to the Pratt's Theorem says that an agent has increasing, decreasing or constant risk aversion if and only if her Arrow–Pratt measure of risk aversion is increasing, decreasing or constant. One needs only to consider utility functions $u_1(w) = u(w)$ and $u_2(w) = u(w + \Delta w)$ for arbitrary $\Delta w > 0$.

Arrow (1965) and Pratt (1964) extended their theory of measurement of risk aversion to relative risk, that is, risk per dollar of an agent's wealth. Risk compensation $\rho_r(w, \tilde{\zeta})$ for relative risk $\tilde{\zeta}$ with $E(\tilde{\zeta}) = 0$ at initial wealth *w* is defined by

$$E[u(w + w\tilde{\zeta})] = u(w - w\rho_r(w, \tilde{\zeta})). \tag{6}$$

The Arrow–Pratt measure of relative risk aversion is $R(w) = -\frac{u''(w)}{u'(w)}w$. The measures ρ_r and *R* are related in the same way that their counterparts for absolute risk are related. Power and logarithmic utility functions have constant relative risk aversion.

Risk compensation is defined when the agent's initial position is risk-free. The approximation (5) of risk compensation by the Arrow–Pratt measure holds at a risk-free position, too. Measures of risk aversion that can be used when the initial position is risky have been developed by Ross (1981) and Machina and Neilson (1987).

When random variables are vector-valued (multivariate), the Arrow–Pratt theory can be applied to risk in one coordinate (for example, consumption of one good) when values of other coordinates are deterministic. Alternatively, multivariate risk aversion can be defined by requiring condition (1) to hold for every multivariate random variable $\tilde{x}$. Multivariate random variables arise when objects of choice are consumption plans of multiple goods or consumption plans over multiple time periods. Multivariate risk aversion is equivalent to concavity of the von Neumann–Morgenstern utility function and implies that the induced ordinal preferences over multiple goods under certainty are convex. The theory of comparative risk aversion has been extended to the multivariate case by Kihlstrom and Mirman (1974) under the restriction that utility functions induce the same ordinal preferences.

Rabin and Thaler (2001) have pointed out a peculiar feature of risk aversion under expected utility. If an agent rejects a small actuarially favourable gamble at every level of wealth (or at a big enough range of wealth), then she will reject a gamble with a modest loss and an arbitrarily large gain. They presented a calibration exercise which shows that any risk-averse agent who rejects an even-chance gamble of losing $10 or winning $11 will turn down an even-chance gamble of losing $1,000 or winning any sum of money. The significance of Rabin and Thaler's observation is a subject of current debate.

Risk aversion without expected utility

A representation of preferences which is closely related to but more general than the expected utility is the state-dependent expected utility. For a finite set S of states of the world, it is written $\sum_{s \in S} \pi_s u_s(x_s)$, where π_s is the probability of state s and u_s is the state-dependent utility. Werner (2005a) has shown that an agent with state-dependent expected utility is risk averse in the sense of having a preference for deterministic outcomes over risky outcomes with equal expectations if and only if the utility functions u_s are state independent and concave. When utility functions u_s are state dependent, a risk-free wealth may have risky utility and, more importantly, risky marginal utility. Karni (1985) has developed a theory of aversion to risk in marginal utility defined by an agent being unwilling to take an actuarially fair gamble when starting from a position of risk-free marginal utility of wealth. Measures of risk aversion analogous to the measures introduced by Arrow and Pratt can be defined, and an extension of the Pratt's theorem obtains for utility functions that have the same set of wealth profiles with risk-free marginal utility. State-dependent utilities arise in instances of behaviour under health risk.

The Arrow–Pratt theory of risk aversion is based on the simple notion that every risky outcome is more risky than the deterministic outcome with equal expectation. For preferences that do not have an expected utility representation (state-independent, or not), this concept of ‘more risky’ is too restrictive to deliver a meaningful notion of risk aversion. A weaker concept of ‘more risky’ has been introduced by Rothschild and Stiglitz (1970). It is a partial ordering of random variables according to the integrals of their cumulative distribution functions. For two random variables $\tilde{x}$ and $\tilde{y}$ with the same expectation, $\tilde{x}$ is *more risky* than $\tilde{y}$ if

$$\int_{-\infty}^{w} F_y(t)\, dt \leq \int_{-\infty}^{w} F_x(t)\, dt, \quad \forall w, \tag{7}$$

where F_x is the cumulative distribution function assigning to each w the probability $Prob\{\tilde{x} \leq w\}$.

An agent whose utility function on random variables is monotone decreasing with respect to the ordering of more risky is said to be *strongly risk averse* (see Cohen, 1995). It follows that a strongly risk-averse agent, when starting from a risky position $\tilde{x}$, is unwilling to take a gamble $\tilde{z}$ with zero expectation conditional on each possible realization x of $\tilde{x}$, that is, $E(\tilde{z}|\tilde{x} = x) = 0$ for every x. The ordering (7) has been known in mathematics as the second-order stochastic dominance and the strongly risk-averse functions have been known as the Schur concave functions (see Marshall and Olkin, 1979). Chew, Karni and Safra (1987) derived necessary and sufficient conditions for strong risk aversion of two types of utility functions: the rank-dependent expected utility of Quiggin (1982), and the dual utility of Yaari (1987). Characterization results for general utility functions can be found in Machina (1982), Chew and Mao (1995), and Dana (2005). For an expected utility, strong risk aversion and risk aversion in the sense of (1) are equivalent. Concavity of the von Neumann–Morgenstern utility function is necessary and sufficient for strong risk aversion of expected utility (see Rothschild and Stiglitz, 1970).

An important representation of preferences under uncertainty, more general than the expected utility, is

the maxmin (or multiple-prior) expected utility (see Gilboa and Schmeidler, 1989). Under the maxmin expected utility representation an agent has a set Π of probability measures (priors) on states instead of a single probability measure, and a von Neumann–Morgenstern utility function *u*. The set Π is assumed to be closed and convex. An agent's utility of risky wealth $\tilde{x}$ is

$$\min_{P \in \Pi} E_P[u(\tilde{x})], \tag{8}$$

where $E_P[u(\tilde{x})]$ is the expectation of $u(\tilde{x})$ with respect to probability measure *P*. Multiplicity of probability measures reflects the agent's ambiguous information about states of the world. Taking the minimum reflects the concern with the ‘worst case’ scenario. If the set Π consists of all probability measures, then the maxmin expected utility (8) equals $\min_s u(x_s)$ meaning that the agent follows the Wald's criterion of choice. Maxmin expected utility may exhibit risk aversion with respect to some probability measure in the set of priors. If the von Neumann–Morgenstern utility function *u* is concave, the agent prefers deterministic wealth in the amount of $E_P(\tilde{x})$ over risky wealth $\tilde{x}$ for every probability measure *P* in her set of priors. Thus the agent is risk averse in the Arrow–Pratt sense with respect to every *P* in the set Π. Wald's minimum utility is also strongly risk averse with respect to every probability measure. Many maxmin expected utility functions are not distribution invariant under any probability measure on states, rendering the question of strong risk aversion meaningless for these functions. Werner (2005b) proposes a concept of more risky, stronger than (7), such that adding a gamble with zero conditional expectation makes an initial risky position more risky, but without identifying random variables with their probability distributions. For many (but not all) sets of priors, maxmin expected utility with concave von Neumann–Morgenstern utility function is risk averse in that sense under some probability measure from the set of priors.

Some implications of risk aversion

The choice of insurance coverage provides a good illustration of implications of risk aversion on agents' behaviour. Suppose that an expected-utility maximizing individual with initial wealth *w* faces a risk of losing *L* with probability π or not losing it with probability $1 - \pi$. She is offered insurance against the loss at price *p* per dollar of coverage. A strictly risk-averse individual will choose full coverage giving her risk-free wealth $w - pL$, if the insurance is priced actuarially fair, that is $p = \pi$. If it is priced above the fair value, that is $p > \pi$, then the optimal coverage will be partial. Schlesinger (1997) shows that these results continue to hold under risk aversion without expected utility.

A risk-averse investor who invests her initial risk-free wealth among many risky assets and a risk-free asset will choose an optimal portfolio with risky payoff only if the expected return on that portfolio exceeds the risk-free return. For a strictly risk-averse investor, the expected return on the optimal portfolio must strictly exceed the risk-free return if the payoff is risky. This is the fundamental *risk–return trade-off* in asset markets and it is a consequence of risk aversion. It continues to hold when the

investor's initial position includes an endowment portfolio of assets. In an equilibrium in competitive asset markets where many strictly risk-averse investors trade their endowment portfolios, the market portfolio (that is, the outstanding supply of assets) must have expected return that exceeds the risk-free return. This is so because the return on the market portfolio is a weighted sum of the returns on investors' optimal portfolios, with weights equal to the respective shares of total wealth. Expected returns on optimal portfolios exceed the risk-free return, with some exceeding it strictly, if the payoff of the market portfolio is risky. Thus, risk aversion provides a qualitative explanation of the expected return in equity markets exceeding the risk-free return. Attempts to give a quantitative explanation of the observed high excess return on equities over risk-free bonds have led to the *equity premium puzzle* (see Mehra and Prescott, 1985).

See Also

- expected utility hypothesis
- non-expected utility theory
- risk
- stochastic dominance

Bibliography

Arrow, K. 1965. *Aspect of the Theory of Risk Bearing*. Helsinki: Yrjö Jahnsson Foundation.

Chew, S., Karni, E. and Safra, Z. 1987. Risk aversion in the theory of expected utility with rank dependent preferences. *Journal of Economic Theory* 42, 370–81.

Chew, S. and Mao, M. 1995. A Schur concave characterization of risk aversion for non-expected utility preferences. *Journal of Economic Theory* 67, 402–35.

Cohen, M. 1995. Risk-aversion concepts in expected- and non-expected-utility models. *Geneva Papers on Risk and Insurance Theory* 20, 73–91.

Dana, R.-A. 2005. A representation result for concave Schur-concave functions. *Mathematical Finance* 15, 613–34.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with nonunique prior. *Journal of Mathematical Economics* 18, 141–53.

Karni, E. 1985. *Decision Making under Uncertainty*. Cambridge, MA: Harvard University Press.

Kihlstrom, R. and Mirman, L. 1974. Risk aversion with many commodities. *Journal of Economic Theory* 8, 361–88.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Machina, M. 1982. Expected utility analysis without the independence axiom. *Econometrica* 50, 277–323.

Machina, M. and Neilson, W. 1987. The Ross characterization of risk aversion: strengthening and extension. *Econometrica* 55, 1139–49.

Marshall, A. and Olkin, I. 1979. *Inequalities: Theory of Majorization and its Applications*. New York: Academic Press.

Mehra, R. and Prescott, E. 1985. The equity premium: a puzzle. *Journal of Monetary Economics* 15, 145–61.

Pratt, J. 1964. Risk aversion in the small and in the large. *Econometrica* 32, 132–6.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3, 323–43.

Rabin, M. and Thaler, R. 2001. Anomalies: risk aversion. *Journal of Economic Perspectives* 15(1), 219–232.

Ross, S. 1981. Some stronger measures of risk aversion in the small and in the large with applications. *Econometrica* 49, 621–38.

Rothschild, M. and Stiglitz, J. 1970. Increasing risk. I: a definition. *Journal of Economic Theory* 2, 225–43.

Schlesinger, H. 1997. Insurance demand without the expected-utility paradigm. *Journal of Risk and Insurance* 64(1), 19–39.

Werner, J. 2005a. A simple axiomatization of risk-averse expected utility. *Economics Letters* 88, 73–7.

Werner, J. 2005b. Risk and risk aversion when states of nature matter. Mimeo.

Yaari, M. 1987. The dual theory of choice under risk. *Econometrica* 55, 95–115.

How to cite this article

Werner, Jan. "risk aversion." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

article?id=pde2008_R000155> doi:10.1057/9780230226203.1443

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

risk sharing

Ethan Ligon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Agents increase their expected utility by using state-contingent transfers to share risk; many institutions seem to play an important role in permitting such transfers. If agents are suitably risk-averse, then in the absence of any frictions the benchmark Arrow–Debreu model predicts that *all* risk will be shared, so that idiosyncratic shocks will have no effect on individuals; we call this full risk sharing. Real-world tests of full risk sharing tend to reject it; accordingly, researchers have devised models incorporating various frictions to try to explain the partial risk sharing evident in the data.

Keywords

Arrow–Debreu economy; commitment; credit; Euler equations; financial markets; full risk sharing; idiosyncratic risk; insurance; Lagrange multipliers; Pareto efficiency; partial risk sharing; permanent-income hypothesis; risk aversion; risk sharing; separability; sharecropping; state-contingent transfers; von Neumann–Morgenstern preferences; von Neumann–Morgenstern utility functions

Article

Any two agents may be said to share risk if they employ state-contingent transfers to increase the expected utility of both by reducing the risk of at least one. A very wide variety of human institutions seem to play an important role in risk sharing, including insurance, credit, financial markets, and sharecropping in developing countries.

To be precise, consider a set of agents indexed by $i = 1, \ldots, n$ each with von Neumann–Morgenstern utility function U_i and a finite set of possible states of the world $s = 1, \ldots, S$, each of which occurs with probability $p(s)$. For simplicity, suppose that each agent i receives a quantity of a single consumption good $x_i(s)$ in state s, thus receiving expected utility

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$EU_i(x_i) = \sum_{s=1}^{S} p(s)U_i(x_i(s)),$$

where x_i denotes the random variable, $\{x_i(s)\}$ denotes its realizations, and E is the expectation operator. We assume that U_i is strictly increasing, weakly concave, and continuously differentiable for all $i = 1, \ldots, n$, so that all agents are at least weakly risk averse. Define the risk faced by agent i to be a quantity

$$R_i(x_i) = EU_i(x_i) - EU(x_i).$$

This cardinal measure orders probability distributions in the same manner as Rothschild and Stiglitz (1970). We say that i faces *idiosyncratic risk* if $R_i(x_i) > 0$ and $\mathrm{corr}(U_i'(x_i), U_j'(x_j)) < 1$ for some j, where U_j' denotes j's marginal utility. If any agent i bears idiosyncratic risk, then there exists a set of state-contingent transfers of the consumption good between i and j, $\left\{\tau_i^j(s)\right\}$ which will strictly increase the expected utility of each, while strictly decreasing the risk of at least one of i and j. Implementing such transfers is risk sharing.

Full risk sharing

What might be termed *full risk sharing* (Allen and Gale, 1988; Rosenzweig, 1988) is a situation in which all idiosyncratic risk is eliminated. While agents may still face risk, this risk is shared, so that marginal utilities of consumption are perfectly correlated across all agents. Full risk sharing is a hallmark of any Pareto-efficient allocation in an Arrow–Debreu economy, provided that agents have von Neumann–Morgenstern preferences, no one is risk-seeking, and at least one agent is strictly risk averse.
Let us establish the necessity of full risk sharing for any interior Pareto-efficient allocation in a simple multi-period endowment economy. The environment is similar to that described above, but agents consume in several periods indexed by $t = 1, \ldots, T$, with agent i discounting future expected utility using a discount factor β_i. Different states of the world are realized in each period, with the probability of state $s_t \in \{1, \ldots, S\}$ being realized in period t allowed to depend on the period, and so given by $p_t(s_t)$. Then consider the problem facing a social planner, who assigns state-contingent consumption allocations to solve

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\max_{\{(c_{it}(s))\}} \sum_{i=1}^{n} \lambda_i \sum_{t=1}^{T} \beta_i^{t-1} \sum_{s_t=1}^{S} p_t(s_t) U_i(c_{it}(s_t))$$

subject to the resource constraints

$$\sum_{i=1}^{n} c_{it}(s_t) \leq \sum_{i=1}^{n} x_{it}(s_t),$$

which must be satisfied at every period t and state s_t; the planner takes as given the initial state s_0 and a set of positive weights $\{\lambda_i\}$. By varying these weights one can compute the entire set of interior Pareto-efficient allocations (Townsend, 1987).

If we let $\mu_t(s_t)$ denote the Lagrange multiplier associated with the resource constraint for period t in state s_t, then the first order conditions for the social planner's problem are

$$\lambda_i \beta_i^{t-1} p_t(s_t) U_i'(c_{it}(s_t)) = \mu_t(s_t). \quad (1)$$

Since this condition must be satisfied in all periods and states for every agent, it follows that

$$U_i'(c_{it}(s_t)) = \frac{\lambda_j}{\lambda_i} \left(\frac{\beta_j}{\beta_i} \right)^{t-1} U_j'(c_{it}(s_t))$$

for any period t, any pair of agents (i, j) and any state s_t, so that $\mathrm{corr}(U_i'(c_{it}), U_j'(c_{jt})) = 1$, and we have full risk sharing.

Thus far, we've considered risk sharing only in the context of an endowment economy. However, the thrust of the claims advanced above holds much more generally. If we were, for example, to add production and some kind of intertemporal technology (for example, storage), the first order conditions of the planner's problem with respect to state-contingent consumptions (1) would remain unchanged – the effect of these changes would be that the Lagrange multipliers $\{\mu_t(s_t)\}$ would change. This is an

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

illustration of what is sometimes called 'separability' between production and consumption, which typically prevails only when there is full risk sharing (see, for example, Benjamin, 1992).

Risk sharing can also be thought of as a means to smooth consumption across possible states of the world. This suggests a connection to the permanent-income hypothesis, which at its core involves agents smoothing consumption across periods. And indeed, it's easy to show that full risk sharing in every period implies the kind of smoothing across periods implied by the consumption Euler equation. However, the consumption Euler equation doesn't imply full risk sharing.

Tests of full risk sharing

The insight that Pareto-efficient allocation among risk-averse agents implies full risk sharing has led to tests of versions of (1). The usual strategy involves adopting a convenient parameterization of U_i, and then calculating the logarithm of both sides of (1). For example, if $U_i(c) = \frac{c^{1-\gamma}}{1-\gamma}$, with $\gamma > 0$, then this yields the relationship

$$\gamma \log c_{it}(s_t) = \log \frac{\mu_t(s_t)}{p_t(s_t)} - \log \frac{\lambda_i}{\beta_i} - t \log \beta_i. \quad (2)$$

This is a simple consumption function, which we would expect to be consistent with any efficient allocation. The quantity $\frac{\mu_t(s_t)}{p_t(s_t)}$ is related to the aggregate supply of the consumption good. Note that this is the only determinant of consumption which depends on the random state. This reflects the fact that the only risk borne by agents in an efficient allocation will be aggregate risk. The second term varies with neither the state nor the date, and can be thought of as depending on the levels of consumption that agent *i* can expect (in a decentralization of this endowment economy, λ_i could be interpreted as a measure of *i*'s time zero wealth). The final term has to do with differences in agents' patience.

Now, suppose one has panel data on realized consumption for a sample of agents over some period of time. If we let $\tilde{c}_{it}$ denote observed consumption for agent *i* in period *t*, (2) implies the estimating equation

$$\log \tilde{c}_{it} = \eta_t + \alpha_i + \delta_i t + \varepsilon_{it}, \quad (3)$$

where $\eta_t = \log \frac{\mu_t(s_t)}{\gamma p_t(s_t)}$, $\alpha_i = -\log \frac{\lambda_i}{\gamma \beta_i}$, $\delta_i = -\log \beta_i$, and ε_{it} is some disturbance term. Because this final disturbance term isn't implied by the model, it's typically motivated by assuming that it's related

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
either to measurement error in consumption or to some preference shock.
The reduced form consumption eq. (3) can be straightforwardly estimated by using ordinary least squares, but this doesn't constitute a test of full risk sharing. To construct such a test, one typically uses data on other time-varying, idiosyncratic variables which would plausibly influence consumption under some alternative model which predicts less than full risk sharing. Perhaps the most obvious candidate for such a variable is some measure of income, for example the observed endowment realizations $\tilde{x}_{it}$ referred to in the model above. Then one can add (the logarithm of) this variable to reduced form as an additional regressor, yielding an estimating equation of the form

$$\log \tilde{c}_{it} = \eta_t + \alpha_i + \delta_i t + \varphi \log \tilde{x}_{it} + \varepsilon_{it} \quad (4)$$

(Mace, 1991; Cochrane, 1991; Deaton, 1992; Townsend, 1994). Then full risk sharing and an auxiliary assumption that ε_{it} is mean independent of the regressors implies the exclusion restriction $\varphi = 0$, which can be easily tested.

Partial risk sharing

Restrictions along the lines of (4) have been used to test for full risk sharing in a wide variety of settings, including within-dynasty risk sharing (Hayashi, Altonji and Kotlikoff, 1996) in the United States, risk sharing across countries (Obstfeld, 1994), risk sharing within networks in the Philippines (Fafchamps and Lund, 2003), and risk sharing across households in India (Townsend, 1994), Africa, or the United States (Mace, 1991). A typical finding is that the estimated response of consumption to income shocks is small but significant, leading one to reject the null hypothesis of full risk sharing. In this case it is tempting to interpret the estimated relationship as determining the response of consumption to income. However, this is generally a mistake. By rejecting the hypothesis of full risk sharing one also rejects the model which generated the hypothesis, so that theory no longer supports the interpretation of (4) as a consumption function.
Given this kind of evidence against full risk sharing, scholars have been led to devise and test alternative models in which some kind of friction leads to agents bearing some idiosyncratic income risk. Two promising frictions are private information and limited commitment. In the case of private information, realized or announced incomes may provide a useful signal regarding hidden actions or information, and thus an agent's consumption will optimally depend on this signal, leading to a balance between risk sharing and incentives (Holmström, 1979); Ligon (1998) tests this weaker risk-sharing hypothesis in three Indian villages, and is unable to reject it. In the case of limited commitment, an agent who receives an unusually large endowment realization may be tempted to renege on a pre-existing risk-sharing arrangement unless she receives a larger share of resources (Kocherlakota, 1996); a test of this model in the same three Indian villages by Ligon, Thomas and Worrall (2002) finds that this model predicts a response of consumption to income of just the right magnitude. Still, the construction, estimation, and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
testing of well-specified models which predict only partial risk sharing remains in its infancy.

See Also

- Euler equations
- permanent-income hypothesis

Bibliography

Allen, F. and Gale, D. 1988. Optimal security design. *Review of Financial Studies* 1, 229–63.

Benjamin, D. 1992. Household composition, labor markets, and labor demand: testing for separation in agricultural household models. *Econometrica* 1, 287–322.

Cochrane, J. 1991. A simple test of consumption insurance. *Journal of Political Economy* 99, 957–76.

Deaton, A. 1992. *Understanding Consumption*. Oxford: Clarendon Press.

Fafchamps, M. and Lund, S. 2003. Risk-sharing networks in rural Philippines. *Journal of Development Economics* 71, 261–87.

Hayashi, F., Altonji, J. and Kotlikoff, L. 1996. Risk-sharing between and within families. *Econometrica* 64, 261–94.

Holmström, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10, 74–91.

Kocherlakota, N. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63, 595–610.

Ligon, E. 1998. Risk-sharing and information in village economies. *Review of Economic Studies* 65, 847–64.

Ligon, E., Thomas, J. and Worrall, T. 2002. Informal insurance arrangements with limited commitment: theory and evidence from village economies. *Review of Economic Studies* 69, 209–44.

Mace, B. 1991. Full insurance in the presence of aggregate uncertainty. *Journal of Political Economy* 99, 928–56.

Obstfeld, M. 1994. Risk-taking, global diversification, and growth. *American Economic Review* 84, 1310–29.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Rosenzweig, M. 1988. Risk, implicit contracts and the family in rural areas of low-income countries. *Economic Journal* 98, 1148–170.

Rothschild, M. and Stiglitz, J. 1970. Increasing risk: I. A definition. *Journal of Economic Theory* 2, 225–43.

Townsend, R. 1987. Microfoundations of macroeconomics. In *Advances in Economic Theory: Fifth World Congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

Townsend, R. 1994. Risk and insurance in village India. *Econometrica* 62, 539–91.

How to cite this article

Ligon, Ethan. "risk sharing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000231> doi:10.1057/9780230226203.1445

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

risk

Mark J. Machina and Michael Rothschild
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The phenomenon of *risk* plays a pervasive role in economics. Without it, financial and capital markets would consist of the exchange of a single instrument each period, the communications industry would cease to exist, and the profession of investment banking would reduce to simple accounting. One need but consult the contents of any recent economics journal to see how the recognition of risk has influenced current research in the discipline. This article presents an overview of the modern economic theory of the characterization of risk and the modelling of economic agents' responses to it.

Keywords

capital asset pricing model; expected utility hypothesis; mean-standard deviation; portfolio theory; probabilistic sophistication hypothesis; probability distribution; Riemann–Stieltjes integral; risk; risk aversion; risk preference; stochastic dominance; subjective probability; uncertainty; von Neumann and Morgenstern

Article

The phenomenon of *risk* is one of the key determining factors in the formation of investment decisions, the operation of financial markets, and several other aspects of economic activity.

Risk versus uncertainty

The most fundamental distinction in this branch of economic theory, due to Knight (1921), is that of ‘risk’ versus ‘uncertainty’. A situation is said to involve *risk* if the randomness facing an economic agent presents itself in the form of exogenously specified or scientifically calculable *objective probabilities*, as with gambles based on a roulette wheel or a pair of dice. A situation is said to involve *uncertainty* if the randomness presents itself in the form of alternative possible *events*, as with bets on a horse race, or decisions involving whether or not to buy earthquake insurance.
The standard approach to the modelling of preferences under uncertainty (as opposed to risk) has been the *state-preference approach* (for example, Arrow, 1964; Debreu, 1959, ch. 7; Hirshleifer, 1965; 1966; Karni, 1985; Yaari, 1969). Given the absence of exogenously specified objective probabilities, this approach represents the randomness facing the individual by a set of mutually exclusive and exhaustive *states of nature* or *states of the world* $S = \{s_1, \ldots, s_n\}$. Depending upon the particular application, this partition of all conceivable futures may either be very coarse, as with the pair of states (it snows here tomorrow, it doesn't snow here tomorrow) or else very fine, so that the description of a single state might read ‘it snows more than three inches here tomorrow *and* the temperature in Paris at noon is 73° *and* the price of gold in New York is over $900.00/ounce’. The objects of choice in this framework consist of *state-payoff bundles* of the form $(c_1,\ldots, c_n)$, which specify the payoff that the individual will receive in each of the respective states. As with regular commodity bundles, individuals are assumed to have preferences over state-payoff bundles which can be represented by indifference curves in the *state-payoff space* $\{(c_1,\ldots, c_n)\}$. Even though the state-preference approach has led to important advances in the analysis of choice under uncertainty (see, for example, the above citations), the advantages of being

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

able to draw on modern probability theory has led economists to hypothesize that an individual's *beliefs* in such settings can nevertheless still be represented by so-called *personal probabilities* or *subjective probabilities*, which take the form of an additive *subjective probability measure* $\mu(\cdot)$ over the state space $\mathcal{S}$. In such a case, a given state-payoff bundle (c_1, …, c_n) will be viewed as yielding outcome c_i with probability $\mu(s_i)$, so that the individual would evaluate the bundle ($c_1,\ldots,c_n$) in the same manner as he or she would evaluate a casino gamble which yielded the payoffs ($c_1,\ldots, c_n$) with respective objective probabilities ($\mu(s_1),\ldots, \mu(s_n)$). The hypothesis that individuals have such probabilistic beliefs and evaluate state-payoff bundles in such a manner is termed the *hypothesis of probabilistic sophistication*, and permits a unified application of probability theory to the analysis of decisions under both objective risk and subjective uncertainty. The joint hypothesis of probabilistic sophistication and expected utility risk preferences has been axiomatized by Ramsey (1926), Savage (1954), Anscombe and Aumann (1963), Pratt, Raiffa and Schlaifer (1964) and Raiffa (1968, ch.5), and probabilistic sophistication without expected utility has been axiomatized by Machina and Schmeidler (1992).

Choice under risk: the expected utility model

For reasons of expositional ease, we consider a world with a single commodity (for example, wealth). An agent making a decision under either risk or probabilistic uncertainty can therefore be thought of as facing a choice set of alternative univariate probability distributions. In order to consider both discrete (for example, finite outcome) distributions as well as distributions with density functions, we represent each such probability distribution by means of its cumulative distribution function $F(\cdot)$, where $F(x) \equiv \text{prob}(\tilde{x} \succsim x)$ for the random variable $\tilde{x}$.

In such a case we can model the agent's preferences over alternative probability distributions in a manner completely analogous to the approach of standard (that is, non-stochastic) consumer theory: he or she is assumed to possess a ranking $\succsim$ over distributions which is complete, transitive and continuous (in an appropriate sense), and hence representable by a real-valued *preference* function $V(\cdot)$ over cumulative distribution functions, in the sense that $F^*(\cdot) \succsim F(\cdot)$ (that is, the distribution $F^*(\cdot)$ is weakly preferred to $F(\cdot)$) if and only if $V(F^*) \geq V(F)$.

Of course, as in the non-stochastic case, the above set of assumptions implies nothing about the functional form of the preference functional $V(\cdot)$. For reasons of both normative appeal and analytic convenience, economists typically assume that $V(\cdot)$ is a *linear functional* of the distribution $F(\cdot)$, and hence takes the form

$$V(F) \equiv \int U(x)\, dF(x) \tag{1}$$

for some function $U(\cdot)$ over wealth levels x, where $U(\cdot)$ is referred to as the individual's *von Neumann–Morgenstern utility function*. (For readers unfamiliar with the *Riemann–Stieltjes integral* $\int U(x)dF(x)$, it represents nothing more than the expected value of $U(\tilde{x})$ when $\tilde{x}$ possesses the cumulative distribution function $F(\cdot)$. Thus if $\tilde{x}$ took the values $x_1,\ldots, x_n$ with probabilities $p_1,\ldots, p_n$ then $\int U(x)dF(x)$ would equal $\sum U(x_i)p_i$, and if $\tilde{x}$ possessed the density function $f(\cdot) = F'(\cdot)$ then $\int U(x)dF(x)$ would equal $\int U(x)f(x)dx$.)

Since the right side of (1) may be thought of as the mathematical expectation of $U(\tilde{x})$, this specification is known as the expected utility model of preferences over random prospects (for a more complete statement of this model, see expected utility hypothesis). Within this framework, an individual's attitudes towards risk are reflected in the shape of his or her utility function $U(\tilde{x})$. Thus, for example, an individual would always prefer shifting probability mass from lower to higher outcome levels if and only if $U(x)$ were an increasing function of x, a condition which we shall henceforth always assume. Such a shift of probability mass is known as a *first order stochastically dominating shift*.

Risk aversion

The representation of an individual's preferences over distributions by the shape of his or her von Neumann–Morgenstern utility function provides the first step in the modern economic characterization of risk. After all, whatever the notion of ‘riskier’ means, it is clear that bearing a random wealth $\tilde{x}$ is riskier than receiving a certain payment of $\bar{x} = E[\tilde{x}]$ (the expected value of the random variable $\tilde{x}$). We therefore have from Jensen's inequality that an individual would be *risk averse*, that is, would always prefer a payment of $E[\tilde{x}]$ (and obtaining utility $U(E[\tilde{x}])$) to bearing the risk $\tilde{x}$ (and obtaining expected utility $E[U(\tilde{x})]$) if and only if his or her utility function were concave. This condition is illustrated in Figure 1, where the random variable $\tilde{x}$ is assumed to take on the values x' and x'' with respective probabilities 2/3 and 1/3.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Figure 1
Von Neumann–Morgenstern utility function of a risk-averse individual

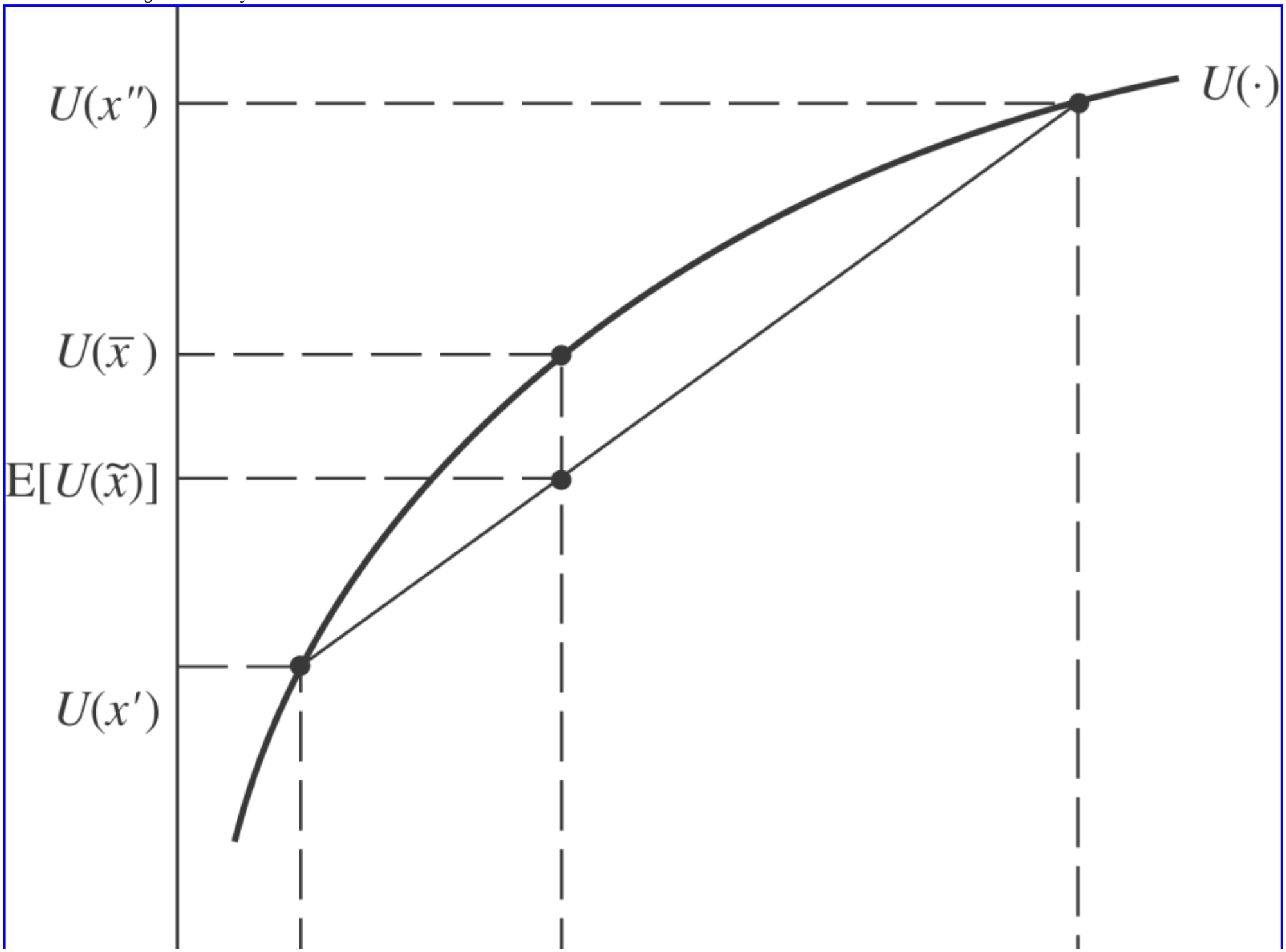

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

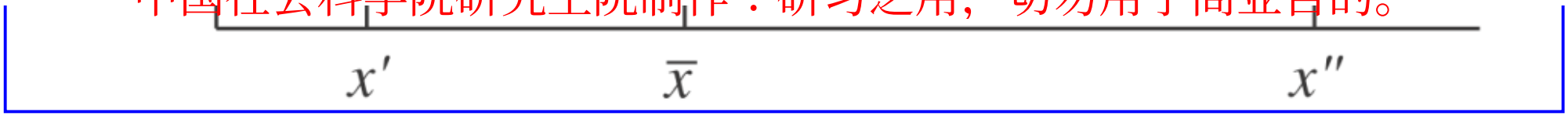

Of course, not all individuals need be risk averse in the sense of the previous paragraph. Another type of individual is a *risk lover*. Such an individual would have a *convex* utility function, and would accordingly prefer receiving a random wealth $\tilde{x}$ to receiving its mean $E[\tilde{x}]$ with certainty. An example of such a utility function is given in Figure 2.

Figure 2

Von Neumann–Morgenstern utility function of a risk-loving individual

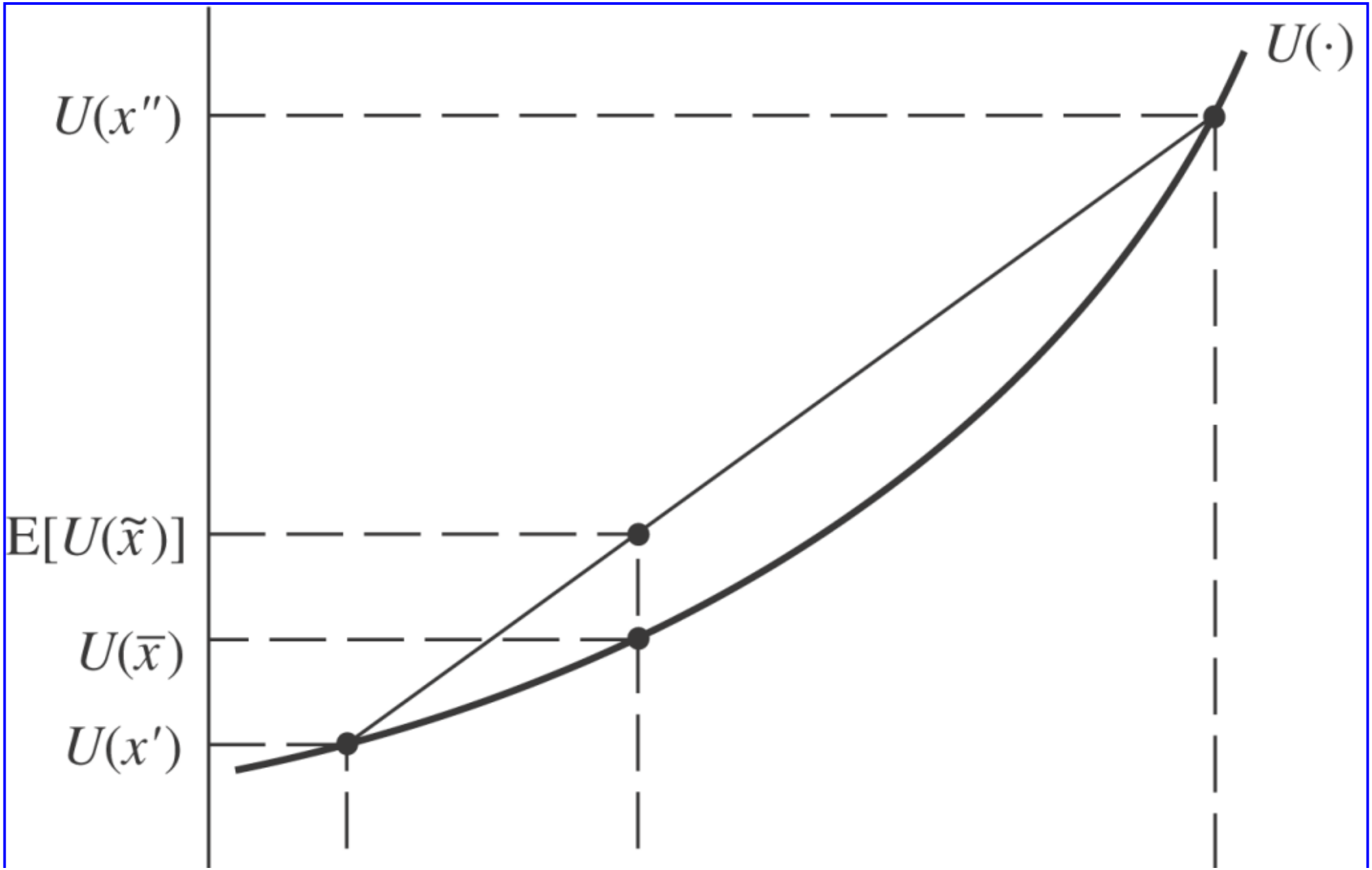

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

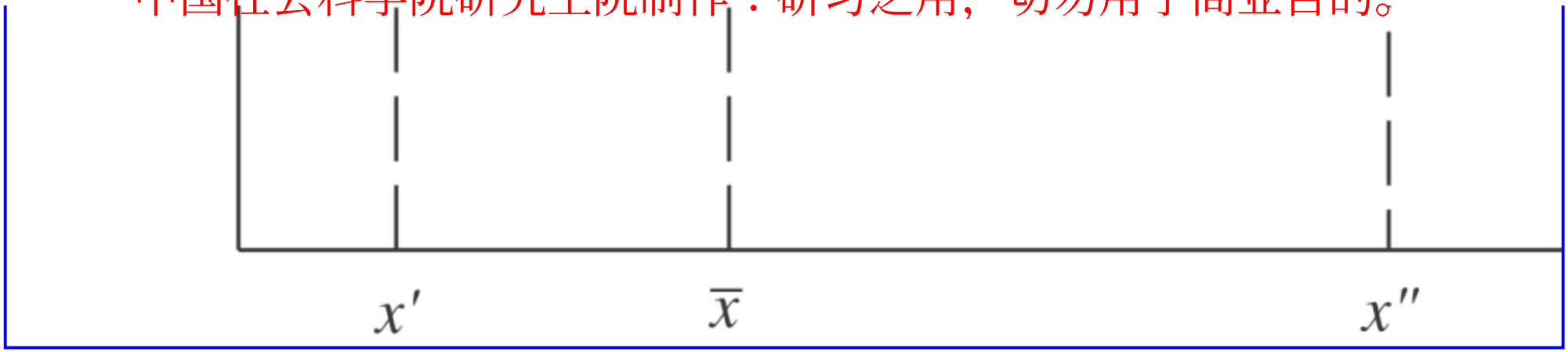

Standard deviation as a measure of risk

While the above characterizations of risk aversion and risk preference allow for the derivation of many results in the theory of choice under risk, they say nothing about which of a pair of non-degenerate random variables $\tilde{x}$ and $\tilde{y}$ is the more risky. Since real-world choices are almost never between risky and riskless situations but rather over alternative risky situations, such a means of comparison is necessary.

The earliest and best-known univariate measure of the riskiness of a random variable $\tilde{x}$ is its *variance* $\sigma^2 = E[(\tilde{x} - E[\tilde{x}])^2]$ or alternatively its *standard deviation* $\sigma = E[(\tilde{x} - E[\tilde{x}])^2]^{1/2}$. The tractability of these measures, as well as their well-known statistical properties, led to the widespread use of mean-standard deviation analysis in the 1950s and 1960s, and in particular to the development of modern portfolio theory by Markowitz (1952; 1959), Tobin (1958) and others. As an example of this, consider Figure 3. Points *A* and *B* correspond to the distributions of a riskless asset with (per dollar) gross return r_0 and a risky asset with random return $\tilde{r}$ with mean $\mu_{\tilde{r}}$ and standard deviation $\sigma_{\tilde{r}}$. An investor dividing a dollar between the two assets in proportions $\alpha:(1-\alpha)$ will possess a portfolio whose return has a mean of $\alpha \cdot r_0 + (1-\alpha) \cdot \mu_{\tilde{r}}$ and standard deviation $(1-\alpha) \cdot \sigma_{\tilde{r}}$, so that the set of attainable (μ, σ) combinations consists of the line segment connecting the points *A* and *B* in the figure. It is straightforward to show that, if the individual were also allowed to *borrow* at rate r_0 in order to finance purchase of the risky asset (that is, could sell the riskless asset short), then the set of attainable (μ, σ) combinations would be the ray emanating from *A* and passing through *B* and beyond.

Figure 3

Portfolio analysis in the mean-standard deviation model

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

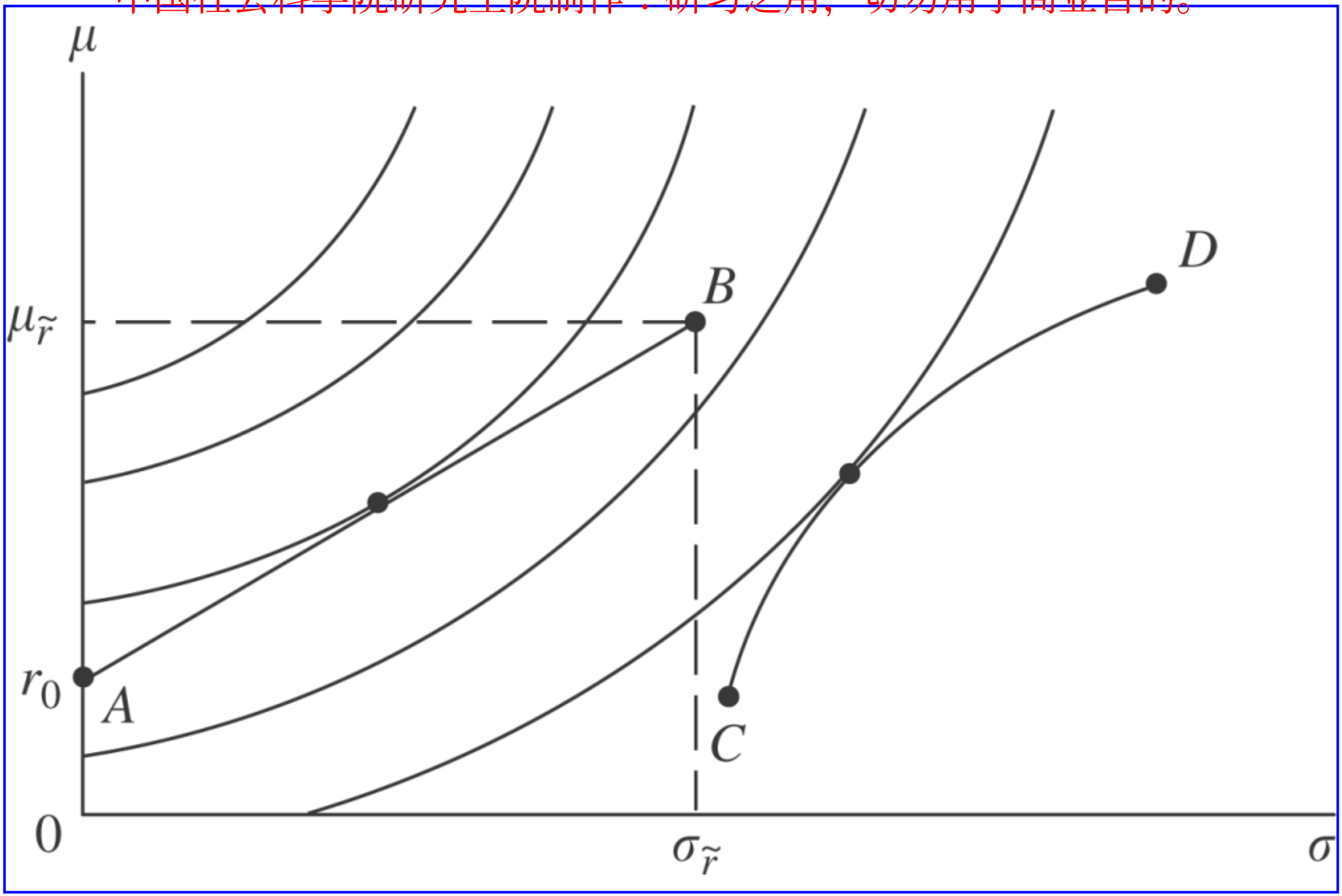

If we then represent the individual's risk preferences by means of indifference curves in this diagram, we obtain his or her optimal portfolio (the example in the figure implies an equal division of funds between the two assets). In the more general case of choice between a pair of risky assets, the set of (μ , σ) combinations generated by alternative divisions of wealth between them will trace out a possibly nonlinear locus such as the one between points *C* and *D* in the diagram, with the curvature of this locus determined by the degree of statistical dependence (that is, covariance) between the two random returns.

As mentioned, the representation and analysis of risk and risk-taking by means of the variance or standard deviation of a distribution proved tremendously useful in the theory of finance, culminating in the mean-standard deviation-based *capital asset pricing model* of Sharpe (1964), Lintner (1965), Mossin (1966) and Treynor (1999). However, by the late

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1960s the mean-standard deviation approach was under attack for two reasons.

The first reason (known since the 1950s) was the fact that an expected utility maximizer would evaluate all distributions solely on the basis of his or her means and standard deviations if and only if their von Neumann–Morgenstern utility function took the quadratic form $U(x) \equiv ax + bx^2$ for $b \neq 0$. The sufficiency of this condition is established by noting that $E[U(\tilde{x})] = E[a\tilde{x} + b\tilde{x}^2] = a \cdot E[\tilde{x}] + b \cdot (E[\tilde{x}]^2 + \sigma^2)$. To prove necessity, note that the distributions that yield a 2/3:1/3 chance of the outcomes $(x - \delta) : (x + 2\delta)$ and a 1/3:2/3 chance of the outcomes $(x - 2\delta) : (x + \delta)$ both possess the same mean and variance for each *x* and δ, so that $(2/3) \cdot U(x - \delta) + (1/3) \cdot U(x + 2\delta) \equiv (1/3) \cdot U(x - 2\delta) + (1/3) \cdot U(x + \delta)$ for all *x* and δ. Differentiating with respect to δ and simplifying yields $U'(x + 2\delta) + U'(x - 2\delta) \equiv U'(x + \delta) + U'(x - \delta)$ for all *x* and δ. This implies that $U'(\cdot)$ must be linear and hence that $U(\cdot)$ must be quadratic.

The assumption of quadratic utility is objectionable. If an individual with such a utility function is risk averse (that is, if $b<0$), then (*a*) utility will decrease as wealth increases beyond $1/(2b)$, and (*b*) the individual will be more averse to constant additive risks about high wealth levels than about low wealth levels – in contrast to the observation that those with greater wealth take greater risks (see for example Hicks, 1962, or Pratt, 1964).

Borch (1969) struck the second and strongest blow to the mean-standard deviation approach. He showed that, for any two points (μ_1, σ_1) and (μ_2, σ_2) in the (μ, σ) plane which a mean-standard deviation preference ordering would rank as indifferent, it is possible to find random variables $\tilde{x}_1$ and $\tilde{x}_2$ which possess these respective (μ, σ) values and where $\tilde{x}_2$ first order stochastically dominates $\tilde{x}_1$. However, *any* person with an increasing von Neumann–Morgenstern utility function would strictly prefer $\tilde{x}_2$ to $\tilde{x}_1$. In response to these arguments and the additional criticisms of Feldstein (1969), Samuelson (1967) and others, the use of mean-standard deviation analysis in economic theory waned. See, however, the work of Meyer (1987) for a partial rehabilitation of such two-moment models of preferences.

Besides the variance or standard deviation of a distribution, several other univariate measures of risk have been proposed. Examples include the *mean absolute deviation* $E[|\tilde{x} - E[\tilde{x}]|]$, the *interquartile range* $F^{-1}(.75)–F^{-1}(.25)$, and the classical statistical measures of *entropy* $\sum \ln(p_i) \cdot p_i$ or $\int \ln(f(x)) \cdot f(x) \cdot dx$. Although they provide the convenience of a single numerical index, each of these measures is subject to problems of the sort encountered with the variance or standard deviation. In particular, the entropy measure is based exclusively on the *probability levels* of a random variable, and is particularly unresponsive to its *outcome values* – for example, the 50:50 gambles over the values \$49:\$51 and \$0:\$100 possess identical entropy levels.

Increasing risk

By the late 1960s, the failure to find a satisfactory univariate measure of risk led to another approach to this problem. Working independently, several researchers (Hadar and Russell, 1969; Hanoch and Levy, 1969; and Rothschild and Stiglitz, 1970; 1971) developed an alternative characterization of increasing risk. The appeal of this approach is twofold. First, it formalizes three different intuitive notions of increasing risk. Second, it allows for the straightforward derivation of comparative statics results in a wide variety of economic situations. Unlike the univariate measures described above, however, this approach provides only a partial ordering of random variables. In other words, not all pairs of random variables can be compared with respect to their riskiness.

We now state three alternative formalizations of the notion that a cumulative distribution function $F^*(\cdot)$ is riskier than another distribution $F(\cdot)$ with the same mean. In the following, all distributions are assumed to be over the outcome interval $[0, M]$ unless otherwise indicated.

The first definition of increasing risk captures the notion that 'risk is what all risk averters hate'. Thus an increase in risk must lower the expected utility of all risk averters. Formally:

- (A) *$F^*(\cdot)$ and $F(\cdot)$ have the same mean and $\int U(x)\,dF^*(x) \leq \int U(x)\,dF(x)$ for every concave utility function $U(\cdot)$.*

Note that this relationship will *not* be satisfied by every pair of distributions with the same mean. That is to say, there exist pairs $F(\cdot)$ and $F^*(\cdot)$, with the same mean, but such that some risk-averse utility functions prefer $F(\cdot)$ to $F^*(\cdot)$ but other risk-averse utility functions prefer $F^*(\cdot)$ to $F(\cdot)$. This reflects the above-stated fact that comparative risk is a *partial* rather than a *complete* order over the family of probability distributions, even over families of distributions with a common mean. (Although comparative risk is not a complete order, it is a *transitive* order, in the sense that, if the pair $F^*(\cdot)$ and $F(\cdot)$ satisfy condition (A), and the pair $F^{**}(\cdot)$ and $F^*(\cdot)$ satisfy condition (A), then the pair $F^{**}(\cdot)$ and $F(\cdot)$ will also satisfy condition (A).)

The second characterization of the notion that a random variable $\tilde{y}$ with distribution $F^*(\cdot)$ is riskier than a variable $\tilde{x}$ with distribution $F(\cdot)$ is that $\tilde{y}$ consists of the variable $\tilde{x}$ plus an additional zero-mean noise term $\tilde{\varepsilon}$. One possible specification of this is that $\tilde{\varepsilon}$ statistically independent of $\tilde{x}$. However, this condition is too strong in the sense that it does not allow the variance of $\tilde{\varepsilon}$ to depend upon the magnitude of $\tilde{x}$, as in the case of heteroskedastic noise. Instead, Rothschild and Stiglitz (1970) modelled the addition of noise by the condition:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- (B) *$F(\cdot)$ and $F^*(\cdot)$ are the respective cumulative distribution functions of the random variables $\tilde{x}$ and $\tilde{x}+\tilde{\varepsilon}$, where $E[\tilde{\varepsilon}|x] \equiv 0$ for all values of x.*

The third notion of increasing risk involves the concept, due to Rothschild and Stiglitz (1970), of a *mean preserving spread*. Intuitively, such a spread consists of moving probability mass from some region in the centre of a probability distribution out to its tails in a manner that preserves the expected value of the distribution, as seen in the top panels of Figures 4 and 5. In the discrete case of Figure 4, probability mass is moved from the pair of outcome values *b* and *c* out to the outcome values *a* and *d*. In the continuous density case of Figure 5, probability mass is moved from the interval (*b*, *c*) out to the intervals (*a*, *b*) and (*c*, *d*). We can unify, generalize and formalize this condition by saying that $F^*(\cdot)$ differs from $F(\cdot)$ by a 'mean preserving spread' if they have the same mean and there exists a single crossing point x_0 such that $F^*(x) \geq F(x)$ for all $x \leq x_0$ and $F^*(x) \leq F(x)$ for all $x \geq x_0$ (see the middle panels of Figures 4 and 5). Since it is clear that *sequences* of such spreads will also lead to riskier distributions, the third characterization of increasing risk is:

- (C) *$F^*(\cdot)$ may be obtained from $F(\cdot)$ by a finite sequence, or as the limit of an infinite sequence, of mean preserving spread.*

Figure 4
Mean preserving spread of a discrete distribution

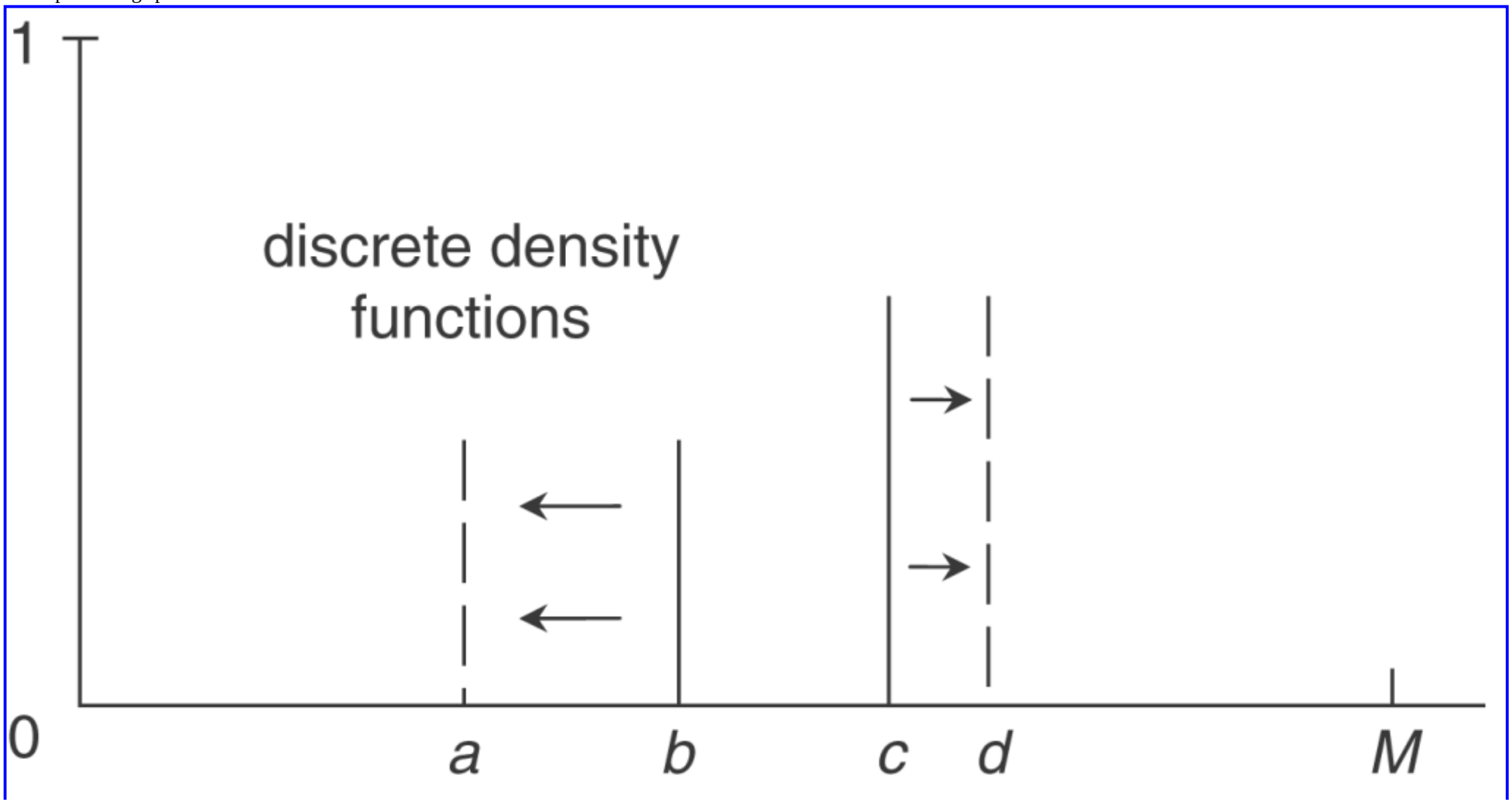

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

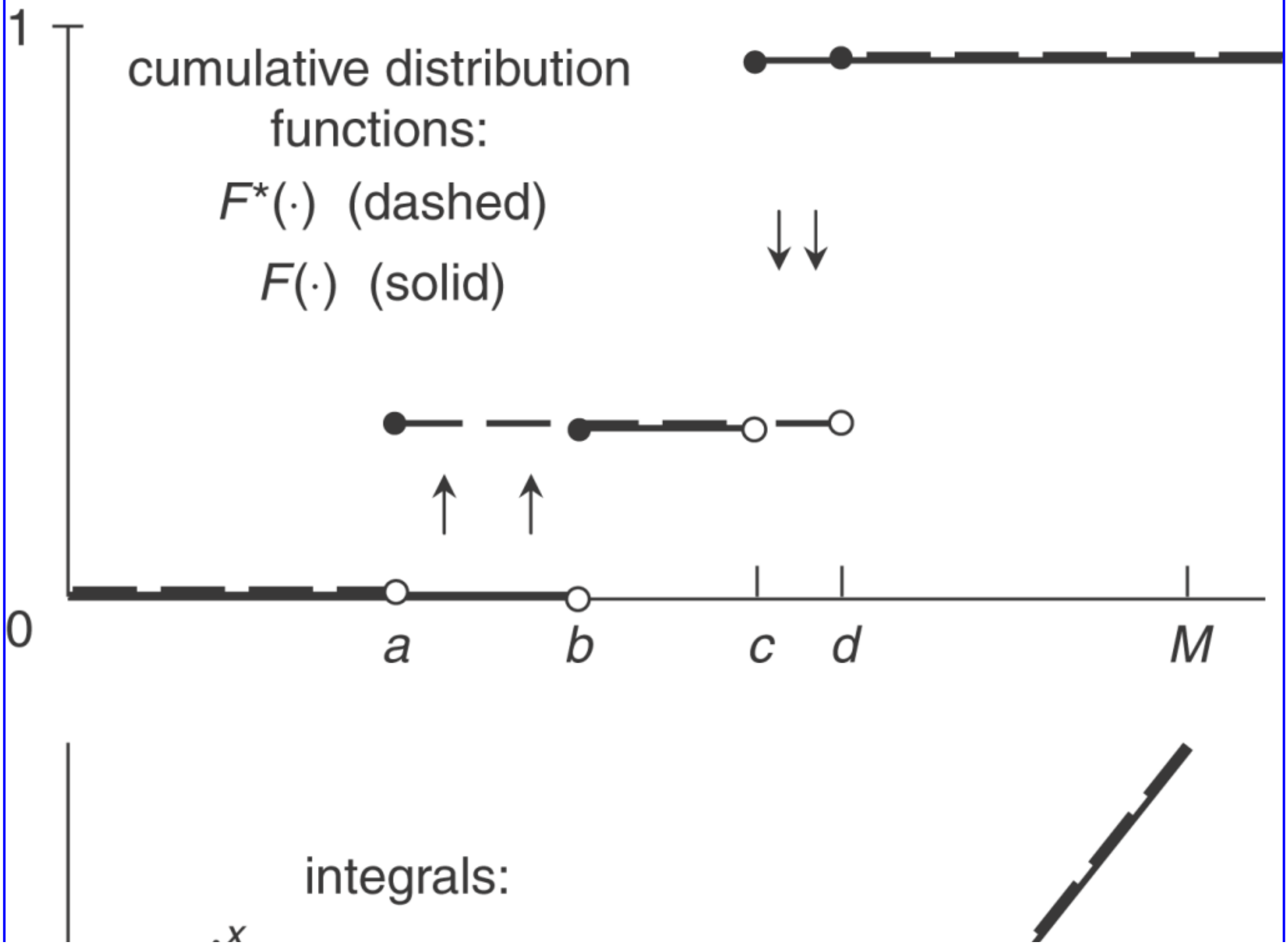

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

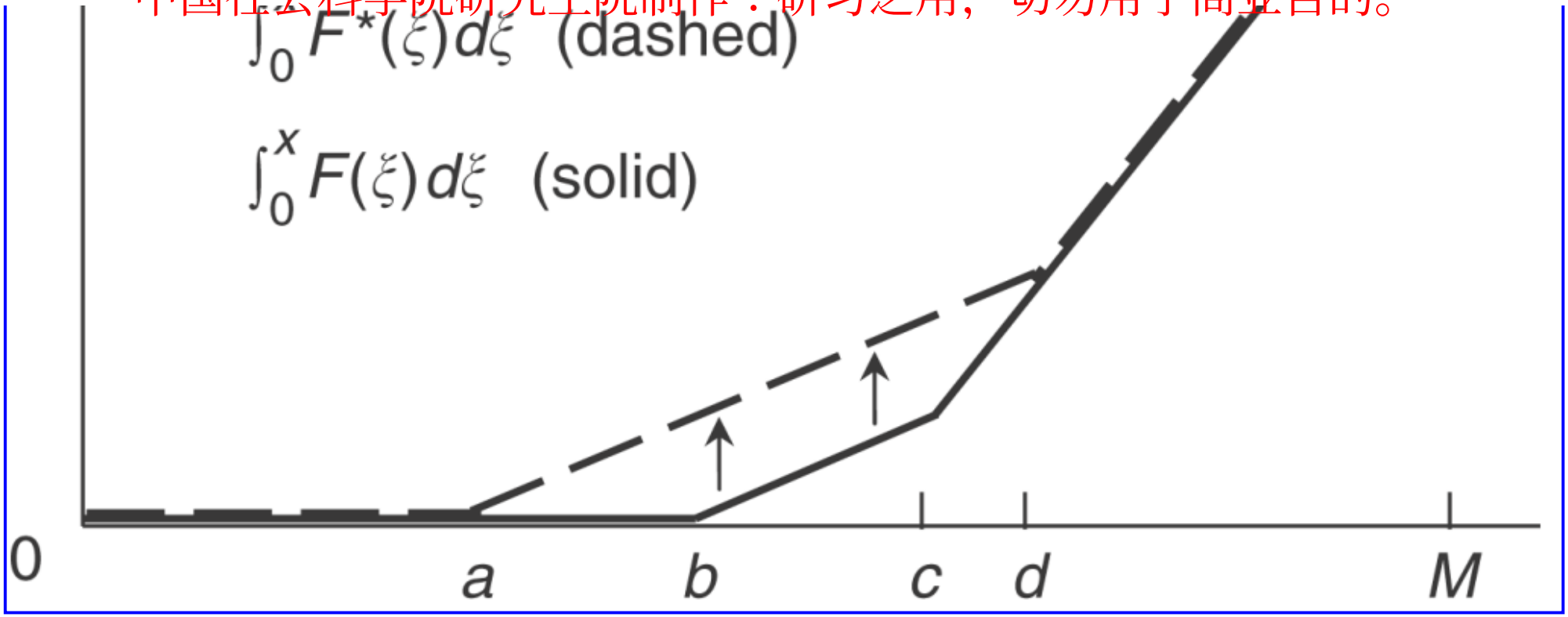

Figure 5
Mean preserving spread of a density function

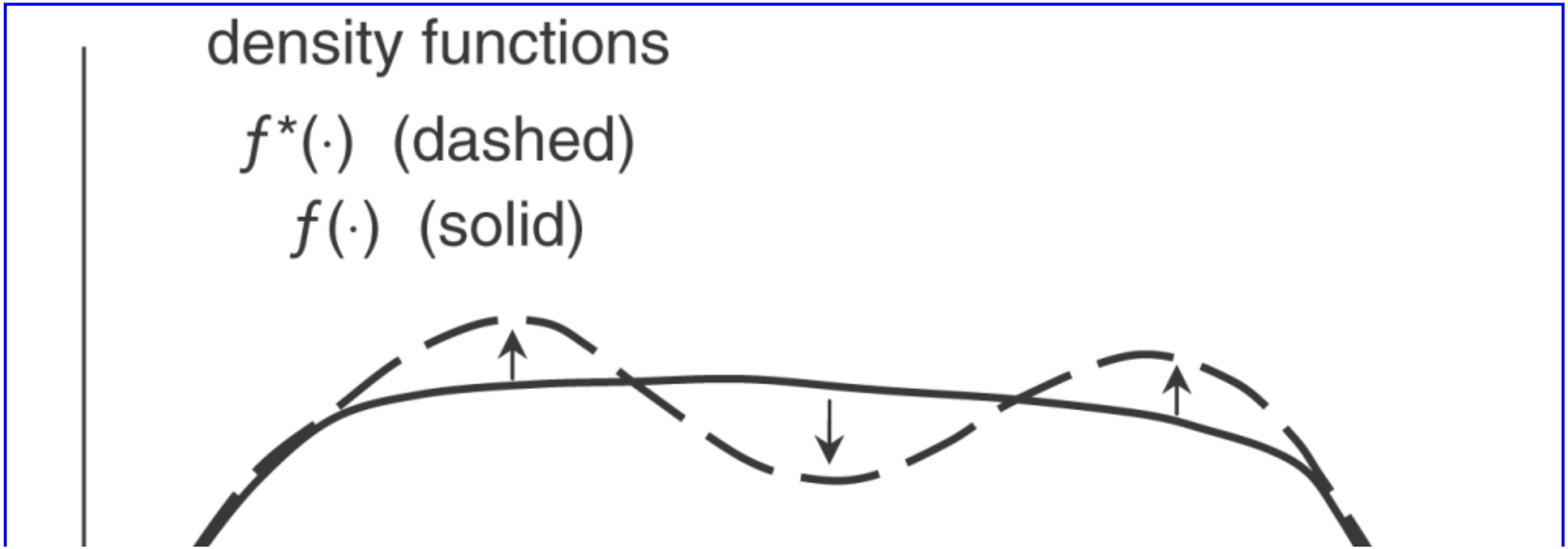

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

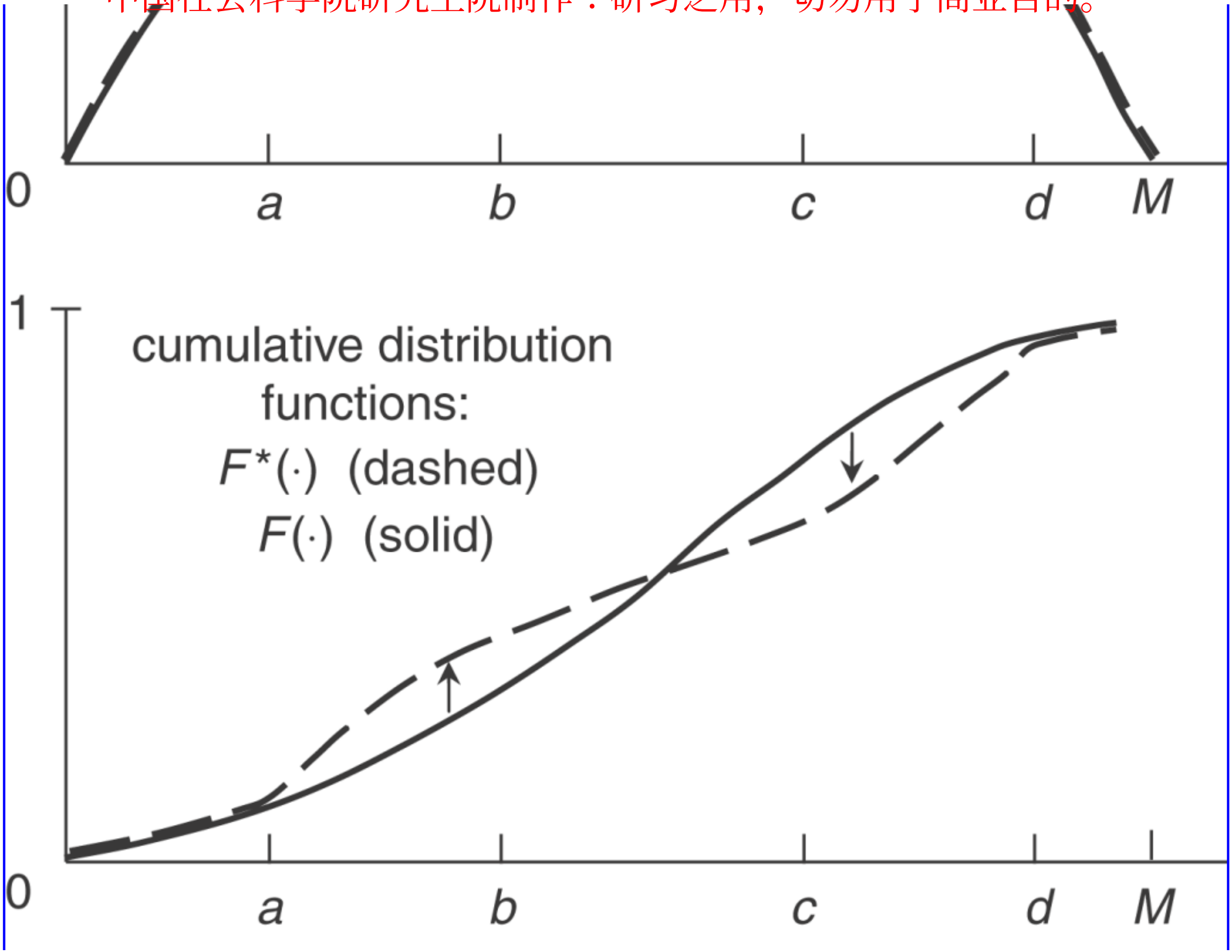

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

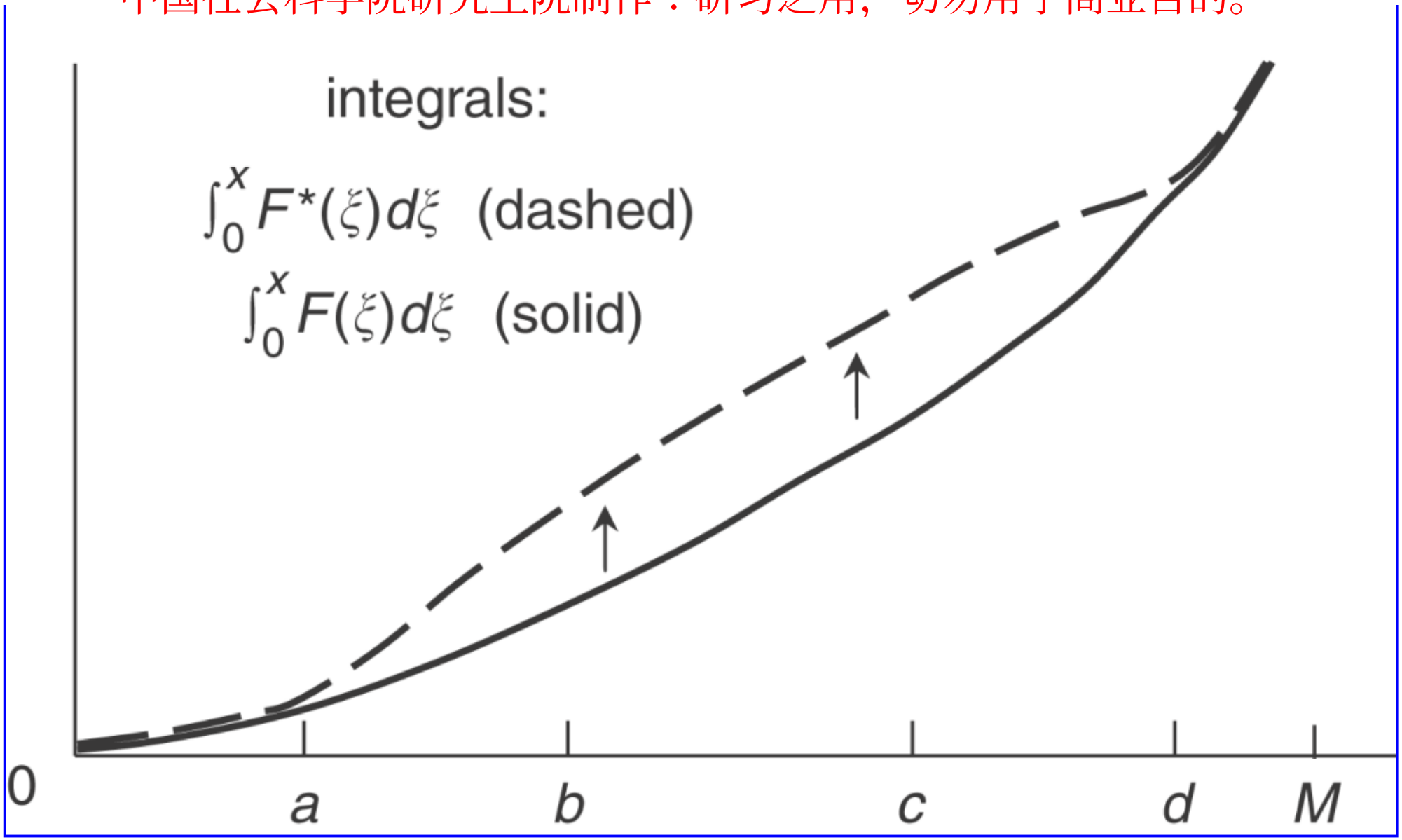

Although the single crossing property of the previous paragraph serves to characterize cumulative distribution functions that differ by a single mean preserving spread, distributions that differ by a sequence of such spreads will typically not satisfy the single crossing condition. However, if we consider the integrals of these cumulative distribution functions, we see from the bottom panels of Figures 4 and 5 that a mean preserving spread will always serve to raise or preserve the value of this integral for each x, and (since $F^*(\cdot)$ and $F(\cdot)$ have the same mean) will exactly preserve it for $x=M$. In contrast to the single crossing property, this so-called 'integral condition' *will* continue to be satisfied by distributions which differ by a sequence of one or more mean preserving spreads. Accordingly, we may rewrite condition (C) above by the analytically more convenient:

- (C′) *The integral* $\int_0^x [F^*(\xi) - F(\xi)] \cdot d\xi$ *is non-negative for all* $x>0$, *and is equal to* 0 *at* $x=M$.

Rothschild and Stiglitz (1970) showed that these three concepts of increasing risk are the same by proving that conditions (A), (B) and (C/C′) are equivalent. Thus, a single partial

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

ordering of distribution functions corresponds simultaneously to the notion that risk is what risk averters hate, to the notion that adding noise to a random variable increases its risk, and to the notion that moving probability mass from the centre of a probability distribution to its tails increases the riskiness of the distribution. The original Rothschild–Stiglitz formulation and proofs have since been further strengthened and extended by Machina and Pratt (1997).

This characterization permits the derivation of general and powerful comparative statics theorems concerning economic agents' responses to increases in risk. The general framework for these results is that of an individual with a von Neumann–Morgenstern utility function $U(x, \alpha)$ which depends upon both the outcome of some random variable $\tilde{x}$ as well as a *control variable* α which the individual chooses so as to maximize expected utility $\int U(x, \alpha)dF(x; r)$, where the distribution function $F(\cdot; r)$ depends upon some exogenous parameter r (x for example might be the return on a risky asset, and α the amount invested in it). For convenience, we assume that $F(0; r) \equiv prob(\tilde{x} \leq 0) \geq 0$ for all r. The first order condition for this problem is then:

$$\int U_\alpha(x, \alpha)dF(x; r) = 0 \tag{2}$$

where $U_\alpha(x, \alpha)=\partial U(x, \alpha)/\partial \alpha$, and we assume that the second derivative $U_{\alpha\alpha}(x, \alpha)=\partial^2 U(x, \alpha)/\partial \alpha^2$ is always negative to insure we have a maximum. Implicit differentiation of (2) then yields the comparative statics derivative:

$$d\alpha / dr = -\int U_\alpha(x, \alpha)dF_r(x; r) / \int U_{\alpha\alpha}(x, \alpha)dF(x; r) \tag{3}$$

where $F_r(x;r)=\partial F(x;r)/\partial r$. Since the denominator of this expression is negative by assumption, the sign of $d\alpha/dr$ is given by the sign of the numerator $\int U_\alpha(x, \alpha)dF_r(x;r)$. Integrating by parts twice yields:

$$\int U_\alpha(x, \alpha)dF_r(x; r) = \int U_{xx\alpha}(x, \alpha) \cdot \left[\int_0^x F_r(\xi; r)d\xi\right]dx = \int U_{xx\alpha}(x, \alpha) \cdot \left[\frac{d}{dr}\int_0^x F(\xi; r)d\xi\right]dx \tag{4}$$

Thus, if increases in the parameter r imply increases in the riskiness of the distribution $F(\cdot,r)$, it follows from condition (C′) that the signs of the square-bracketed terms in (4) will be non-negative, so that the effect of r upon α depends upon the sign of $U_{xx\alpha}(x, \alpha)=\partial^3 U(x, \alpha)/\partial x^2 \partial \alpha$. Thus, if $U_{xx\alpha}(x, \alpha)$ is uniformly negative a mean preserving increase in risk in the distribution of x will lead to a fall in the optimal value of the control variable α, and vice versa. Another way to see this is to note that if $U_\alpha(x, \alpha)$ is concave in x then a mean preserving increase in risk will lower the left side of the first order condition (2), which (since $U_{\alpha\alpha}(x, \alpha) \leq 0$) will require a drop in α to re-establish the equality. Economists, mathematicians and scientists routinely use this technique when analysing models involving risk; see for example Rothschild and Stiglitz (1971), Dionne, Eeckhoudt and Gollier (1993), Eeckhoudt, Gollier and Schlesinger (1996), Jewitt (1987), Ormiston (1992), Tzeng (2001), Nowak (2004), Chateauneuf, Cohen and Meilijson (2004), Baker (2006), and Beladi, de la Vina and Firoozi (2006).

Related topics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The characterization of risk outlined in the previous section has been extended along several lines. Diamond and Stiglitz (1974), for example, have replaced the notion of a mean preserving spread with that of a mean *utility* preserving spread to obtain a general characterization of a *compensated increase in risk*. They relate this notion to the well known Arrow–Pratt characterization of comparative risk aversion (see expected utility hypothesis).

In addition, researchers such as Ekern (1980), Fishburn (1982), Fishburn and Vickson (1978), Hansen, Holt, and Peled (1978), Tesfatsion (1976), and Whitmore (1970) have extended the above work to the development of a general theory of *stochastic dominance*, which provides a whole sequence of similarly characterized partial orders on distributions, each presenting a corresponding set of equivalent conditions involving algebraic conditions on the distributions, types of spreads, and classes of utility functions which prefer (or are averse to) such spreads. The comparative statics analysis presented above may be similarly extended to such characterizations. An extensive bibliography of the stochastic dominance literature is given in Bawa (1982). Finally, various extensions of the notions of increasing risk and stochastic dominance to the case of multivariate distributions may be found in Epstein and Tanny (1980), Fishburn and Vickson (1978), Huang, Kira and Vertinsky (1978), Lehmann (1955), Levhari, Parousch and Peleg (1975), Levy and Parousch (1974), Russell and Seo (1978), Sherman (1951), and Strassen (1965); see also the mathematical results in Marshall and Okun (1979).

See Also

- expected utility hypothesis
- risk aversion
- uncertainty

Bibliography

Anscombe, F. and Aumann, R. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34, 199–205.

Arrow, K. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31, 91–6.

Baker, E. 2006. Increasing risk and increasing informativeness: equivalence theorems. *Operations Research* 54, 26–36.

Bawa, V. 1982. Stochastic dominance: a research bibliography. *Management Science* 28, 698–712.

Beladi, H., de la Vina, L. and Firoozi, F. 2006. On information value and mean-preserving transformations. *Applied Mathematics Letters* 19, 843–8.

Borch, K. 1969. A note on uncertainty and indifference curves. *Review of Economic Studies* 36, 1–4.

Chateauneuf, A., Cohen, M. and Meilijson, I. 2004. Four notions of mean-preserving increase in risk, risk attitudes and applications to the rank-dependent expected utility model. *Journal of Mathematical Economics* 40, 547–71.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of General Equilibrium*. New Haven: Yale University Press.

Diamond, P. and Rothschild, M., eds. 1989. *Uncertainty in Economics: Readings and Exercises*, 2nd edn. New York: Academic Press.

Diamond, P. and Stiglitz, J. 1974. Increases in risk and in risk aversion. *Journal of Economic Theory* 8, 337–60.

Dionne, G., Eeckhoudt, L. and Gollier, C. 1993. Increases in risk and linear payoffs. *International Economic Review* 34, 309–19.

Eeckhoudt, L., Gollier, C. and Schlesinger, H. 1996. Changes in background risk and risk taking behavior. *Econometrica* 64, 683–9.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ekern, S. 1980. Increasing nth degree risk. *Economics Letters* 6, 329–33.

Epstein, L. and Tanny, S. 1980. Increasing generalized correlation: a definition and some economic consequences. *Canadian Journal of Economics* 13, 16–34.

Feldstein, M. 1969. Mean-variance analysis in the theory of liquidity preference and portfolio selection. *Review of Economic Studies* 36, 5–12.

Fishburn, P. 1982. Simplest cases of *n*'th degree stochastic dominance. *Operations Research Letters* 1, 89–90.

Fishburn, P. and Vickson, R. 1978. Theoretical foundations of stochastic dominance, in Whitmore and Findlay (1978).

Gärdenfors, P. and Sahlin, N.-E., eds. 1988. *Decision, Probability, and Utility: Selected Readings*. Cambridge: Cambridge University Press.

Hadar, J. and Russell, W. 1969. Rules for ordering uncertain prospects. *American Economic Review* 59, 25–34.

Hanoch, G. and Levy, H. 1969. The efficiency analysis of choices involving risk. *Review of Economic Studies* 36, 335–46.

Hansen, L., Holt, C. and Peled, D. 1978. A note on first degree stochastic dominance. *Economics Letters* 1, 315–9.

Hey, J., ed. 1997. *The Economics of Uncertainty, Vol. II: Uncertainty and Dynamics*. Cheltenham: Edward Elgar.

Hicks, J. 1962. Liquidity. *Economic Journal* 72, 787–802.

Hirshleifer, J. 1965. Investment decision under uncertainty: choice-theoretic approaches. *Quarterly Journal of Economics* 79, 509–36.

Hirshleifer, J. 1966. Investment decision under uncertainty: applications of the state-preference approach. *Quarterly Journal of Economics* 80, 252–77.

Huang, C., Kira, D. and Vertinsky, I. 1978. Stochastic dominance for multi-attribute utility functions. *Review of Economic Studies* 45, 611–16.

Jewitt, I. 1987. Risk aversion and the choice between risky prospects: the preservation of comparative statics results. *Review of Economic Studies* 54, 73–85.

Karni, E. 1985. *Decision Making Under Uncertainty: The Case of State Dependent Preferences*. Cambridge, MA: Harvard University Press.

Knight, F. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.

Kyberg, H. and Smokler, H., eds. 1964. *Studies in Subjective Probability*. New York: Wiley.

Lehmann, E. 1955. Ordered families of distributions. *Annals of Mathematical Statistics* 26, 399–419.

Levhari, D., Paroush, J. and Peleg, B. 1975. Efficiency analysis for multivariate distributions. *Review of Economic Studies* 42, 87–91.

Levy, H. and Paroush, J. 1974. Toward multivariate efficiency criteria. *Journal of Economic Theory* 7, 129–42.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 44, 243–69.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Machina, M. and Pratt, J. 1997. Increasing risk: some direct constructions. *Journal of Risk and Uncertainty* 14, 103–27.

Machina, M. and Schmeidler, D. 1992. A more robust definition of subjective probability. *Econometrica* 60, 745–80. Reprinted in Hey (1997).

Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7, 77–91.

Markowitz, H. 1959. *Portfolio Selection: Efficient Diversification of Investment*. New Haven, CT: Yale University Press.

Marshall, A. and Okun, I. 1979. *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press.

Meyer, J. 1987. Two-moment decision models and expected utility maximization. *American Economic Review* 77, 421–30.

Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34, 768–83.

Nowak, M. 2004. Interactive approach in multicriteria analysis based on stochastic dominance. *Control and Cybernetics* 33, 463–76.

Ormiston, M. 1992. First and second degree transformations and comparative statics under uncertainty. *International Economic Review* 33, 33–44.

Pratt, J. 1964. Risk aversion in the small and in the large. *Econometrica* 32, 122–36. Reprinted in Diamond and Rothschild (1989).

Pratt, J., Raiffa, H. and Schlaifer, R. 1964. The foundations of decision under uncertainty: an elementary exposition. *Journal of the American Statistical Association* 59, 353–75.

Raiffa, H. 1968. *Decision Analysis: Introductory Lectures on Choice Under Uncertainty*. Reading, MA: Addison Wesley.

Ramsey, F. 1926. Truth and probability. In *The Foundations of Mathematics and other Logical Essays*, ed. R.B. Braithwaite. London: Routledge & Kegan Paul, 1931. Reprinted in Kyberg and Smokler (1964), Ramsey (1978), and Gärdenfors and Sahlin (1988).

Ramsey, F. 1978. In *Foundations: Essays in Philosophy, Mathematics and Economics*, ed. D.H. Mellor. London: Routledge & Kegan Paul.

Rothschild, M. and Stiglitz, J. 1970. Increasing risk: I. A definition. *Journal of Economic Theory* 2, 225–43. Reprinted in Diamond and Rothschild (1989).

Rothschild, M. and Stiglitz, J. 1971. Increasing risk: II. its economic consequences. *Journal of Economic Theory* 3, 66–84.

Rothschild, M. and Stiglitz, J. 1972. Addendum to ‘increasing risk: I. a definition’. *Journal of Economic Theory* 5, 306.

Russell, W. and Seo, T. 1978. Ordering uncertain prospects: the multivariate utility functions case. *Review of Economic Studies* 45, 605–11.

Samuelson, P. 1967. General proof that diversification pays. *Journal of Financial and Quantitative Analysis* 2, 1–13.

Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley and Sons. Rev. edn, New York: Dover Publications, 1972.

Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Sherman, S. 1951. On a theorem of Hardy, Littlewood, Polya, and Blackwell. *Proceedings of the National Academy of Science* 37, 826–31. (Errata. *Proc. Nat. Ac. Sci*. 38, 382.)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Strassen, V. 1965. The existence of probability measures with given marginals. *Annals of Mathematical Statistics* 36, 423–39.

Tesfatsion, L. 1976. Stochastic dominance and the maximization of expected utility. *Review of Economic Studies* 43, 301–15.

Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25, 65–86.

Treynor, J. 1999. Toward a theory of market value of risky assets. In *Asset Pricing and Portfolio Performance*, ed. R. Korajczyk. London: Risk Books. In Korajczyk (1999).

Tzeng, L. 2001. Increase in risk and weaker marginal-payoff-weighted risk dominance. *Journal of Risk and Insurance* 68, 329–37.

Whitmore, G. 1970. Third-degree stochastic dominance. *American Economic Review* 60, 457–9.

Whitmore, G. and Findlay, M. 1978. *Stochastic Dominance: An Approach to Decision Making Under Risk*. Lexington, MA: Heath.

Yaari, M. 1969. Some remarks on measures of risk aversion and on their uses. *Journal of Economic Theory* 1, 315–29. Reprinted in Diamond and Rothschild (1989).

How to cite this article

Machina, Mark J. and Michael Rothschild. "risk." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000152> doi:10.1057/9780230226203.1442

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

risk-coping strategies

Stefan Dercon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Risk is part of life in developing countries. Despite generally imperfect credit and lacking insurance markets, households use a variety of strategies to manage and cope with risk, including savings and informal credit transactions, mutual support networks, and income and asset diversification. Most evidence suggests that these strategies achieve only partial consumption smoothing and risk-sharing, while they are not without long-term costs in terms of investment and poverty. This article discusses the nature and evidence on the typical strategies used, and explores its implications. It also highlights some outstanding questions.

Keywords

asymmetric information; buffer stocks; credit markets in developing countries; health insurance; household portfolios; income diversification; income shocks; income smoothing; insurance markets; intertemporal consumption smoothing; labour supply; micro-insurance; optimal portfolio mix; precautionary savings; risk; risk aversion; risk coping strategies; risk diversification; risk sharing; risk-sharing contracts; risk-sharing networks; self-insurance

Article

Risk is pervasive, not least in developing countries. A high dependence on agriculture for income and employment means that people's livelihoods are strongly affected by climatic vagaries. Poor health care and immunization means that illness is common, affecting labour supply. Poor infrastructure and market institutions result in limited market integration, leading to a high sensitivity of prices to local shocks. Economic instability, conflict and political instability add further to a high-risk environment in many developing countries.

While risk is widespread, in developing countries insurance markets are typically missing or incomplete. This causes potentially serious welfare losses, especially since government-led alternatives such as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
social security are largely missing, compounded by imperfections in credit markets that limit the extent to which credit can be used as to substitute for insurance. Economic agents are typically risk averse and, even in the poorest settings, the evidence suggests that they do not just passively undergo the consequences of risk. Instead, given risk aversion, they try to make activity and asset portfolio choices to balance their need to make a living, but without exposing themselves to too much risk. The strategies used to achieve this are often referred to as 'risk management and coping strategies'. The economic analysis of these strategies has been one of the areas in which research based on some of the poorest high-risk rural settings in developing countries has made a substantial contribution to the economic literature in general. Our focus in this overview is on the empirical literature, often drawing on the evidence generated from the sample based on six Indian villages for which the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) collected exceptionally detailed panel data over ten years in the high-risk environment of the semi-arid tropics of India (Morduch, 2004).

Self-insurance via savings

It is possible to distinguish a number of commonly observed different risk strategies. First, we can consider strategies that aim to cope with the consequences of shocks. Risk aversion is sufficient to induce households to try to smooth consumption or nutrition, and indeed standard models of consumption smoothing when insurance and credit markets are imperfect can shed light on this type of strategy (Deaton, 1992). When shocks occur, households may decide to curb their consumption loss through the sacrifice of existing assets. Households may pre-empt this by trying to self-insure against risk through precautionary savings. A precautionary motive for savings would be sufficient for savings to increase in response to increased risk, so that a buffer stock is built to deplete when shocks occur. Even though formal credit markets in high-risk settings are typically underdeveloped, informal credit transactions may also be used for smoothing purposes.

There is a large literature testing whether households in developing countries smooth consumption, building on standard models of permanent income and often using shocks to identify the relevant effects (for example, Paxson, 1992). Furthermore, which assets tend to be used for this purpose in the face of shocks has occasionally been investigated. (Finding positive evidence of using assets to smooth is not sufficient to show that any savings were built up for precautionary motives to start with. This would require evidence that greater risk indeed increased savings, which is harder to test.) Nevertheless, the evidence suggests that in some settings productive assets are sold off for smoothing, for example in work by Rosenzweig and Wolpin (1993) using the ICRISAT data, while in other settings (for example, in Burkina Faso, as in Fafchamps, Udry and Czukas, 1998) livestock were not sold off despite serious income shocks. Furthermore, there is much anecdotal evidence that, during famines in mixed farming environments, livestock is not being sold despite serious human nutritional losses and risks.

Squaring these findings with basic theoretical models remains a challenge, and different suggestions can be made. For example, different technologies may underlie the pattern of returns to different assets, so that optimal portfolio mixes would suggest that assets are depleted at different rates when shocks occur. Another possibility is that asset returns and prices are risky as well, so that the reliability of the buffer is limited. For example, if incomes, asset prices and returns have a high positive covariance, then selling assets when incomes are low may not be the optimal strategy, since the future gains from holding on to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
assets is actually high. Alternatively, behavioural theories based on experiments, such as that risk-loving rather than risk-averse behaviour may be displayed in the face of losses that are potentially very large (as in Kahneman and Tversky, 1979), may provide some insight to these puzzles.

A related strategy to self-insurance observed in poor settings involves the key asset available to the poor, namely, labour. In response to shocks, labour supply is adjusted to increase involvement in productive activities, including off-farm or temporary migration (for example, Kochar, 1995). Furthermore, children may be taken out of school and into work in response to income shocks (Jacoby and Skoufias, 1997).

Risk-sharing through mutual support

A second common strategy to cope with the consequences of risk involves non-market institutions based on risk sharing, whereby households or other economic agents use transfers to smooth outcomes across a group of people when shocks occur. Conceptually, this is the cross-sectional equivalent of standard permanent income models: it is concerned with smoothing over space rather than over time. Unless risk preferences differ between economic agents, unlike self-insurance strategies this strategy is relevant only for idiosyncratic shocks, not covariate risk. Townsend (1994) is the seminal paper on high-risk developing-country environments. The basic prediction of Townsend's model, under specific assumptions, is that household consumption is dependent on average village resources but not on individual income realizations. This provides a clear basis for empirical testing of the perfect risk-sharing hypothesis: do idiosyncratic shocks to income affect consumption or nutrition outcomes within a well-defined setting? Using the ICRISAT data from India, he finds that perfect risk-sharing is rejected within the village, even though substantial risk-pooling takes place. Other studies (including some on the same data) suggest that some risk-sharing typically occurs in villages, but the evidence is typically not consistent with perfect risk-sharing.

The failure of perfect risk-sharing in village settings has attracted much attention in terms of theoretical work. Work has focused on accounting for information asymmetries, private savings and the role of enforceability problems related to these informal risk-sharing contracts *ex post* in repeated game contexts (for example, Ligon, Thomas and Worrall, 2002). The work on enforceability has found that constrained efficient contracts will contain updating rules offering higher weights in the risk-sharing contracts in particular states of the world to those facing stronger incentives to leave, and that over time, the weights have memory. The consequence is that these risk-sharing contracts take on features more common in credit contracts, leading to their description as ‘quasi-credit’ arrangements. There is some empirical evidence consistent with these models.

The lack of perfect risk-sharing within villages has led to further work investigating whether risk-sharing occurs in other settings, for example within households or extended families, or in other social groupings. Partial risk-sharing has been documented across ethnic groups as well as within families. More recently, risk-sharing across networks has been explored as well. It is relatively straightforward to analyse risk-sharing within networks beyond specifically exogenously defined groupings, such as caste, but for most group or network links network membership has to be endogenously modelled. Theoretical work has emerged analysing the formation and stability of insurance networks, including in the face of group deviations (for example, Genicot and Ray, 2003). Better data-sets focusing on linkages between

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
households in communities has also allowed further evidence to emerge of the extent and nature of risk-sharing within networks and groups. Integrating the endogeneity of network formation for insurance purposes in empirical analysis nevertheless remains a challenge.

The literatures on intertemporal consumption smoothing and risk-sharing in developing countries appear to converge at least in terms of diagnosis: consumption is relatively smooth in the face of income shocks, but not perfectly. Nevertheless, it is often not clear in the tests whether consumption smoothing in practice occurs through transfers or through self-insurance or credit. Even in the ICRISAT data on India, as used in the Townsend data, this has remained an issue of contention (Townsend, 1995; Morduch, 2004). One strategy is to specifically study the actual responses to shocks (such as dissavings, credit transactions or transfers) and their contribution to smoothing. Deaton and Paxson (1994) provide an alternative test to distinguish whether insurance or credit is responsible for observed smoothing by looking at the changing distribution over time of consumption for a particular cohort in a number of countries, including Taiwan. If consumption smoothing is present due to formal or informal insurance, then inequality can be expected to remain constant over time. If smoothing consumption occurs through credit market transactions, then inequality can be expected to increase over time due to changes in permanent income. Their results suggest that credit markets are more important than insurance for the observed patterns of smoothing.

Income smoothing

Households do not just use strategies that aim to cope with the consequences of shocks; they may try to reduce or mitigate the risk they face, not least given the limits to risk-sharing and self-insurance. To put it more directly, they may aim to smooth income (Morduch, 1995). In rural settings, this strategy has been a central force in shaping farming systems and institutions. The most straightforward strategy is income diversification, whereby income sources are combined and the resulting portfolio faces reduced risk even if the underlying income processes are equally risky when taken separately (as long as they are not perfectly covariate). In some cases, risk management may imply diversifying into (or even specializing in) specific low-risk technologies or activities, such as growing drought-resistant crops or gathering firewood for sale. Social institutions have also developed to include means of reducing exposure to risk. Geographically dispersed marriage patterns, such as those observed by Rosenzweig and Stark (1989) in villages in southern India, can be interpreted as linked to risk diversification within extended families. Local institutions to manage commons and natural resources or land tenure arrangements appear to include risk management as part of their rationale.

Testing the specific role of risk in observed diversification patterns in activities and assets is nevertheless not straightforward. Given the multiple market imperfections faced by the poor, for example in labour markets, the fact of observed multiple activities is not sufficient to sustain the conclusion that risk is its cause. Furthermore, the opportunity to shape risk faced by households also implies econometric problems in standard tests of consumption smoothing and risk sharing, requiring exogenous sources of variation in risk faced. While rainfall variation may provide a useful instrument for common risk, finding sources of exogenous variation for the identification of idiosyncratic shocks is more difficult. These problems suggest that further exploration of risk management strategies will remain methodologically challenging. Furthermore, most investigations of risk strategies in developing

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

countries have been in rural settings, with a focus on agricultural households in relatively stationary environments. This was mainly due to the lack of suitable panel data-sets in urban settings. With more long-term data-sets becoming available, more attention can be paid to the changes in risk strategies following increased integration of local economies and the rise in migration opportunities.

Risk strategies and persistent poverty

All the above hints at an important consequence. While risk strategies contribute to avoiding serious consumption fluctuations, they are not without consequences for welfare, investment and poverty. More specifically, households tend to trade risk and smooth consumption in the short run for lower mean welfare outcomes in the long run. Precautionary motives for saving and credit constraints may induce asset portfolios to focus more on liquidity than on returns. Sales of productive assets for smoothing or taking children out of school to increase labour supply will reduce permanent income. Income smoothing strategies will involve leaving aside profitable opportunities for activity and asset portfolios with a lower mean return. Evidence from villages in the ICRISAT sample in India suggests that these effects may well be substantial, with those households with limited protection against risk (identified by rainfall variability) opting for portfolios with lower returns (Rosenzweig and Binswanger, 1993). More specifically, they find that an increase in rainfall variability by one standard deviation would reduce returns to assets by 35 per cent for the poorest wealth tercile of farmers but have no effect on the richest tercile, which is likely to be better protected against risk through its assets or access to credit. If anything, this type of evidence suggests that risk and the lack of appropriate insurance or protection may well be one of the factors that keep poor people poor.

Risk strategies and policy responses

The perceived failure to keep consumption smooth in the face of risk and the long-run costs attached to existing risk strategies has also stimulated an increasing interest in finding appropriate policy responses, not least since insured risk appears to affect the ability of many poor people to grow out of poverty. Standard transfer schemes, such as food aid or cash transfers, may all provide protection against shocks. However, how existing risk strategies could be strengthened remains less explored (Dercon, 2004). For example, it is clear that self-insurance through savings could offer substantial benefits in terms of protecting against the consequences of risk. Even if insurance and credit markets were not functioning well, improving the availability of better savings product could assist poor people to improve their risk management. Similarly, informal insurance schemes could be strengthened, for example by linking micro-insurance initiatives to indigenous group-based systems. A number of insurance-related initiatives have been taken in this respect by governments and NGOs. For example, weather-indexed insurance contracts that trigger payments if rainfall falls below a predetermined level are being piloted in a number of countries. Health insurance schemes, often based at local health facilities, have also become more widespread. Much work is still needed on developing these initiatives, not least in terms of evaluating their effectiveness with appropriate techniques, for example randomized interventions.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- famines
- Kahneman, Daniel
- micro-credit
- risk
- risk aversion
- risk sharing

Bibliography

Deaton, A. 1992. *Understanding Consumption*. Oxford: Clarendon Press.

Deaton, A. and Paxson, C. 1994. Intertemporal Choice and inequality. *Journal of Political Economy* 102, 437–67.

Dercon, S., ed. 2004. *Insurance against Poverty*. Oxford: Clarendon Press.

Fafchamps, M. and Lund, S. 2003. Risk-sharing networks in rural Philippines. *Journal of Development Economics* 71, 261–87.

Fafchamps, M., Udry, C. and Czukas, K. 1998. Drought and saving in West Africa: are livestock a buffer stock? *Journal of Development Economics* 55, 273–305.

Genicot, G. and Ray, D. 2003. Group formation in risk-sharing arrangements. *Review of Economic Studies* 70, 87–113.

Jacoby, H. and Skoufias, E. 1997. Risk, financial markets and human capital in a developing country. *Review of Economic Studies* 64, 311–35.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–92.

Kochar, A. 1995. Explaining household vulnerability to idiosyncratic income shocks. *American Economic Review* 85, 159–64.

Ligon, E., Thomas, J. and Worrall, T. 2002. Informal insurance with limited commitment: theory and evidence from village economies. *Review of Economic Studies* 69, 209–44.

Morduch, J. 1995. Income smoothing and consumption smoothing. *Journal of Economic Perspectives* 9 (3), 103–14.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Morduch, J. 2004. Consumption smoothing across space. In *Insurance against Poverty*, ed. S. Dercon. Oxford: Clarendon Press.

Paxson, C. 1992. Using weather variability to estimate the responses of savings to transitory income in Thailand. *American Economic Review* 82, 15–33.

Rosenzweig, M. and Binswanger, H. 1993. Wealth, weather risk and the composition and profitability of agricultural investments. *Economic Journal* 103, 56–78.

Rosenzweig, M. and Stark, O. 1989. Consumption smoothing, migration, and marriage: evidence from rural India. *Journal of Political Economy* 97, 905–26.

Rosenzweig, M. and Wolpin, K. 1993. Credit market constraints, consumption smoothing, and the accumulation of durable production assets in low-income countries: investment in bullocks in India. *Journal of Political Economy* 101, 223–44.

Skees, J., Varangis, P., Larson, D. and Siegel, P. 2004. Can financial markets be tapped to help poor people cope with weather risks. In *Insurance against Poverty*, ed. S. Dercon. Oxford: Clarendon Press.

Townsend, R. 1994. Risk and insurance in village India. *Econometrica* 62, 539–91.

Townsend, R. 1995. Consumption insurance: an evaluation of risk-bearing systems in low-income economies. *Journal of Economic Perspectives* 9(3), 83–102.

How to cite this article

Dercon, Stefan. "risk-coping strategies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000230> doi:10.1057/9780230226203.1444

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rist, Charles (1874 – 1955)

Roger Dehem
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

monetary history; monetary theory; Rist, Charles

Article

Born at Prilly, Switzerland, 1874; died at Versailles, 1955. Professor at Montpellier (1899–1912) and Paris (1913–33), Rist was the most notable and influential thinker and actor in the field of money in France in the first half of the 20th century. As a member of the *Comité des experts* (1926) and as a vice-governor of the Bank of France (1926–8), he took an active part in monetary reconstruction in the 1920s. He supported the novel idea of stabilization with devaluation (1926–8). He was also involved as an expert in monetary reforms in Romania (1928), Austria, Turkey and Spain. He was France's delegate at the London Economic Conference (1933).

Although Rist is most widely known for his *History of Economic Doctrines,* written in cooperation with Charles Gide, his lasting claim to fame rests on his profound and consistent interpretation of monetary history and thought as demonstrated in his masterwork, *History of Monetary and Credit Theory*. Based on his first-hand experience in times of great instability, Rist's critical analysis of monetary thought from a long-run viewpoint provides an impressive perspective on the evolution of money. By emphasizing the ‘store of value’ function of money, and by postulating the inability of the state to safeguard it, Rist is critical of authors who supported some form of non-metallic currency, such as John Law, Smith, Ricardo, Wicksell, Knapp and Keynes. He is in sympathy with Cantillon, Galiani, Turgot, Thornton, Tooke and Walras. What he describes as the confusion between money and credit is to be dispelled by drawing a distinction between money proper (gold), credit instruments (convertible banknotes and deposits) and inconvertible paper money. In strong opposition to Keynesianism, Rist is a sceptic in regard of managed currencies and international agreements of the Bretton Wood type. Rist provides the key to the understanding of the French position in monetary matters as opposed to the typical Anglo-American stance in the past 60 years.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1915. (With Ch. Gide.) *A History of Economic Doctrines*. London: G. Harrap. 2nd edn, 1948.

1924. *La déflation en pratique*. Paris: Giard.

1933. *Essais sur quelques problèmes économiques et monétaires*. Paris: Sirey.

1940. *History of Monetary and Credit Theory from John Law to the Present Day*. London: Allen and Unwin.

1961. *The Triumph of Gold*. New York: Philosophical Library.

How to cite this article

Dehem, Roger. "Rist, Charles (1874–1955)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000160> doi:10.1057/9780230226203.1865

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Robbins, Lionel Charles (1898 – 1984)

B.A. Corry
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Beveridge, W. H.; British classical economics; Cambridge School; economic liberalism; economic science; Hayek, F. A. von; history of economic thought; interpersonal utility comparisons; Keynes, J. M.; labour supply; laissez-faire; methodology of economics; philosophy and economics; representative firm; Robbins, L. C.; scarcity; Torrens, R.; value judgements

Article

Lionel Robbins, who in 1961 became Baron Robbins of Clare Market, was one of the major academic economists of the interwar period. He remained active after the Second World War but never really regained the centre of the stage that he had occupied. He was also a great public servant for his country, serving it well and loyally in many aspects of social, political and cultural life. He was truly a ‘Renaissance man’.

Robbins was born in 1898 in Middlesex, the son of Rowland Richard Robbins – for many years President of the National Farmers' Union – and Rosa Marion Robbins. He spent one year reading for an Arts degree at University College London and then volunteered for war service with the Royal Artillery. He saw active service on the Western Front, was wounded and invalided back to England in 1918. He was an undergraduate at the London School of Economics and Political Science from 1920 to 1923, from which he graduated with a BSc (Econ.) degree, choosing political ideas as his major field of study, and having had the left-wing Harold Laski as his tutor. Beveridge employed him as a research assistant for a year, after which Robbins was a tutor in economics at New College, Oxford. He returned to teach economics at LSE from 1925 to 1927, then back to New College as Fellow (1928–9) and finally, at the incredibly young age of 31, back to the Senior Professorship in Economics at LSE to succeed Allyn Young.

Apart from government service during the Second World War, Robbins remained at LSE as Head of the Economics Department until 1960 when, on accepting the Chairmanship of the *Financial Times*, the University of London forced him to resign his professorship – a move than brought Robbins great personal distress, although he retained his connection with LSE and taught courses there until a year or

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
so of his death in 1984.

Outside academic and government advisory activity, Robbins had a distinguished record in arts administration, being connected with both the National Gallery and the Royal Opera House, but he may perhaps be best remembered, in such ‘outside’ activities, for his contribution to the structure of higher education in the United Kingdom. He chaired the committee – commonly referred to as the Robbins Committee on Higher Education – that proposed the criterion that all qualified applicants should receive a place, and financial support, to read for a degree. The acceptance of the ‘Robbins Principle’ led to a vast expansion of degree courses, especially in the social sciences in the 1960s and early 1970s in the UK.

Robbins's contributions to economics may be considered under four headings; economic theory, methodology and philosophy of economics, the theory of economic policy, and the history of economic thought.

Those who only knew Robbins later in his life often forget that he made his initial mark in economics as a theorist. Three contributions here are worth noting; he launched a sustained attack on Marshall's concept of the Representative Firm which was apparently so successful that it drove the concept out of the pages of microeconomic texts. Robbins basically argued that the understanding of the equilibrium neither of the firm nor of the industry was aided by introducing the Representative Firm, hence it should be eliminated from analysis. More recent work has shown a greater sympathy towards Marshall's construct and it seems clear now that Robbins failed to understand the exact dynamic problem that Marshall was trying to cope with and why the Representative Firm was an important contribution to this problem.

Robbins also pioneered the micro-analysis of the labour supply function. Although he did not explicitly use the division of a wage change into an income and substitution effect, he showed clearly why the sign on the response of hours to a real wage rate change would be ambiguous.

In macroeconomics Robbins was a firm exponent of the Austrian theory of the trade cycle and here he was greatly influenced by Frederick von Hayek, whom he brought to LSE from Vienna in 1928. The central feature of the Austrian analysis was that depression was due primarily to under-saving (or excess consumption) and these views, which Robbins expounded as an explanation of the 1930s depression in his book *The Great Depression*, led to a head-on collision between the senior LSE economists and the Cambridge School centred around Keynes. This rift was not finally healed until the wartime collaboration in Whitehall between Robbins and Keynes. After the war in the Marshall Lectures for 1946, published as *The Economic Problem in Peace and War*, Robbins announced his conversion to full employment policies via control of aggregate demand, although it is not clear that he became a Keynesian.

The second area where Robbins made a major contribution and where he wrote what is probably his best known work in economics was that of the methodology and philosophy of economics. His *Nature and Significance of Economic Science* was one of the most cited, if not most read, books on the subject in the period 1932–60, and it influenced greatly economists' views about the nature of their discipline. There are several strands to the book, none original in themselves, but Robbins put them together in beautifully clear prose and in a very persuasive manner. The major themes were; first, that economic science could be clearly demarcated from those discussions of economic issues that involved value judgements – by which latter term Robbins meant evaluative statements of the form ‘better or worse’ where interpersonal comparisons of utility were involved. He also argued that there was a clear demarcation between

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic science and other branches of social enquiry such as social psychology, sociology, politics and so on.

The second major theme was that the subject matter of economic science was not a particular activity (for example, Cannan's view that economics was the science of wealth), but rather an aspect of all human conduct. This aspect was the ‘fact’ of economic scarcity – a manifestation of unlimited ends on the part of individuals and society and means of satisfying those ends that were limited in supply. In words so often quoted in economics texts Robbins defined economic science as ‘that science that studies the relationship between ends and means that have alternative uses’ – a definition that is more than reminiscent of Menger's exposition of the economizing process.

These two aspects of the *Nature and Significance* were widely accepted by the world of academic economists and are still propagated. But they have always had their critics; in particular, the view that there is a body of scientific economics ‘free from value’ is much disputed.

The third aspect of the book – Robbins's views on the procedures for checking the validity of economic theory – was less fortunate in its effect on the development of the subject. Robbins appeared to argue that the central propositions of economics were derived from very basic, and obvious, assumptions and a process of logical deduction from these assumptions. Moreover, these deductions gave essentially qualitative predictions. Robbins expressed great scepticism about the feasibility and meaningfulness of quantitative work in economics, and by the implication of his message inhibited the development of econometric testing in economics.

Robbins's contributions to discussions of economic policy were basically consistent throughout his career, although the purity of his earlier thoughts was muddled as he grew older. His major policy theme was his advocacy of, what may be loosely termed, economic liberalism. Robbins decreasingly argued this on the grounds of some alleged theoretical or a priori superiority of market solutions over collectivist or interventionist plans, but rather as an empirical point that the liberal solution seemed best to combine liberty and efficiency. In his earlier writings, for example *The Economic Causes of War* (1939a) and *The Economic Basis of Class Conflict* (1939b) he adopted an extreme free trade position and it was this stance as much as macro-theory debate that led to his conflict with Keynes in the 1930s. His later work revealed a much greater readiness to allow ad hoc exceptions to strict economic liberalism – he espoused, among other measures, the Beveridge plan, grants for higher education, subsidies for the arts, control of the exports of works of art, overall macro-control for full employment. Probably the most rounded statement of his policy beliefs is to be found in his *Economic Problem in Peace and War*.

Finally, mention must be made of Lionel Robbins's contribution to the teaching and study of the history of economic thought. He, together with one or two other scholars of his generation – like his great friend, Jacob Viner – kept interest in the subject alive and flourishing when many economists regarded it, as they still do, as irrelevant to their studies. Much of his influence came via his masterly teaching of the subject and via the important theses that were produced under his supervision, as much as from his own specific contributions. He also aided the production of important series in the history of economic thought such as the LSE reprints and the collected works of Bentham and J.S. Mill.

Of his specific contributions, two are minor classics, his *Theory of Economic Policy in Classical Political Economy* (1952) and *Robert Torrens and the Evolution of Classical Economics* (1958). In the former work, Robbins argued very persuasively, if not entirely convincingly, that the British classical economists did not adhere to the Continental laissez-faire dogma but rather argued for freedom in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

economic relationships as a general principle with many ad hoc exceptions. He further tried to clear them of any anti-working class bias.

The book on Torrens is a perfect example of how to survey the collected works of a writer who, though not of the first rank of classical economists, is nonetheless a useful writer by whom to assess the general achievement of the classical school.

See Also

- value judgements

Selected works

1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

1934. *The Great Depression*. London: Macmillan.

1937. *Economic Planning and International Order*. London: Macmillan.

1939a. *The Economic Causes of War*. London: Jonathan Cape.

1939b. *The Economic Basis of Class Conflict*. London: Macmillan.

1947. *The Economic Problem in Peace and War*. London: Macmillan.

1952. *The Theory of Economic Policy in English Classical Political Economy*. London: Macmillan.

1958. *Robert Torrens and the Evolution of Classical Economics*. London: Macmillan.

1971. *Autobiography of an Economist*. London: Macmillan.

How to cite this article

Corry, B. A. "Robbins, Lionel Charles (1898–1984)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000161> doi:10.1057/9780230226203.1446

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Robertson, Dennis (1890 – 1963)

M. Anyadike Danes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cooperative monetary economy; cooperative non-monetary economy; cyclical fluctuations; Keynes, J. M.; liquidity preference; rate of interest; Robertson, D.

Article

Dennis Robertson was born in 1890, the son of a clergyman and schoolmaster, and was educated at Eton and Trinity College, Cambridge. After taking a Part I in Classics and Part II in Economics he was elected a Fellow of Trinity College in 1914 and in 1930 became a Reader in the University of Cambridge. He left Cambridge in 1938 to become a Professor in the University of London but during most of his time in that post he was seconded to the Treasury on war-related work. Elected in 1944 to succeed Pigou in the Chair of Political Economy, he returned to the University of Cambridge, holding that position until his retirement in 1957. He died in Cambridge in 1963.
Economics in Cambridge when Robertson commenced working at it was dominated by Marshall. Not by the man himself (although still alive he had retired in 1908) but by his analytical methods and by his *Principles of Economics*. It was quite natural that the topic selected by Robertson for his fellowship dissertation should involve a ‘Marshallian’ approach to a subject on which Marshall himself had written relatively little: the nature and causes of fluctuations in the general level of economic activity. As Robertson recorded in the introduction to the published version of this dissertation:

> In some of the more abstract portions of this essay I shall make use, without further explanation or apology, of the processes and terminology in common use among the school of economic thought associated in this country chiefly with the name of Dr Marshall. My reason is that after a study of many facts and theories I am deliberately of the opinion that one cause of the obscurity which still surrounds this problem is that in the attack upon it full and systematic use has never hitherto been made of the weapons supplied by this particular intellectual armoury. (1915, p. 11)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Although Robertson did not suspect it then, the refinement and further development of the ideas about cycles and growth in economic activity presented in this study were to occupy him for the next 20 years. Two different sorts of factors led him in this direction. The first was the need to develop a framework for designing an organized policy response to the large-scale dislocation of economic life which had followed the end of the First World War, while the second was a more specific, personal, influence. In the early 1920s Keynes commissioned him to write an introductory textbook (in the Cambridge Economic Handbook series) to be entitled, simply, ‘Money’. The difficulties he encountered in attempting to provide an elementary account of monetary theory made Robertson particularly aware that, even in its more sophisticated variants, existing theoretical work provided an inadequate basis for dealing with the economic problems of the 1920s. The combined influence of these two resulted in a prolonged period of reflection and research, yielding a series of loosely related publications which recorded the development of a fairly comprehensive analytical scheme interrelating the problems of money, the trade cycle, economic growth, and the role of the state in promoting economic progress.
Robertson's approach to this analysis involved the development of successively more complicated, more ‘realistic’ models of economies each of which constituted a different, abstract, ‘type’. All ‘types’ shared the characteristics that production was undertaken on the basis of ‘rational’ decision-making by competing producer ‘groups’, each making different products with a fixed labour force and a productive process involving fixed capital. Now although each type of economy was both a *production* and an *exchange* economy it was the possibility that these activities could be ‘organized’ in different ways that distinguished the different types. Production could be organized in two ways, cooperatively or non-cooperatively, while exchange could also be organized in two ways, direct exchange or monetary exchange. In total there were, then, four types of economies. The distinction between the two types of productive organization turned on the decision-making functions of the members of the groups: in a cooperative group decisions were *made* and *carried out* by the group members acting together, while in a non-cooperative group ‘entrepreneurs’ made decisions and ‘workers’ carried them out. In respect of the organization of exchange it was on the existence and use of money that the distinction rested, in one case exchange was carried out by ‘direct barter’, while in the other, money supplied through a (potentially) government-controlled banking system provided the means of exchange.
Robertson's basic analytical building block was the ‘cooperative non-monetary economy’, an economy where each competing industrial group made its employment and, thus, output decision cooperatively, and exchanged its output without the use of money. Although in such an economy no distinction was made between the members of the group, a distinction was made between two different categories of producer groups, those providing consumer goods and those producing capital goods. The first group, consumer goods producers exchanged some of their output with the second group for capital goods, thereby providing consumption goods for capital goods producers. Now an economy of this type, Robertson argued, would experience cyclical fluctuations in aggregate output deriving from the effect of gestation lags on the time pattern of the supply of capital goods and of the durability of capital goods on the time-pattern of demand for their replacement.
A non-cooperative non-monetary economy would, though, experience fluctuations of even greater severity than those felt in an otherwise identical cooperative economy. This proposition derived directly from the fact that in a non-cooperative economy production decisions were taken by entrepreneurs who hired workers to carry them out, and workers and entrepreneurs had differing ‘interests’. These divergent interests were reflected most importantly in the different utility attached to leisure by the two

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
classes. An entrepreneur, for example, would wish to expand output further in the boom and contract it further in the slump than the workers in his group; and since entrepreneurs were in control, their interests prevailed. Although the degree of fluctuation in the non-cooperative economy was more pronounced than in the cooperative, Robertson adopted it as the benchmark which defined the 'appropriate' degree of fluctuation to be aimed at by policymakers concerned with stabilization. He did so because he maintained that the failure to recognize that production was, in practice, organized non-cooperatively could lead to an attempt to reduce fluctuations too much. Such attempts, by altering the structure of incentives, could damage the longer-run growth possibilities of the economy.

The cooperative monetary economy construct was built directly on to the foundations provided by the cooperative non-monetary economy and this type of economy exhibited, therefore, a cyclical pattern in the production of fixed capital which generated cyclical fluctuations in output as a whole. Now the introduction of money also required a slight change of focus, since in the monetary case Robertson concentrated not on fixed capital but on the demand for circulating capital, essentially on the demand for consumption goods which were consumed by those engaged in the process of production. This concern with circulating capital was necessarily associated with the analysis of monetary economies because Robertson made the assumption (reflecting British banking practice) that it was with the finance of the acquisition of circulating capital that the banking system was concerned. His analysis then described the policies which, if adopted by the banking system, would lead to fluctuations being of no greater amplitude than in the corresponding non-monetary economy. A failure to implement such policies would lead to fluctuations in the price level, and thus in output, of greater magnitude than was 'appropriate'.

The difference made by the substitution of cooperation in the monetary type turned principally on the effect on decision-making of changes in income distribution. In particular, it was assumed that only entrepreneurial incomes adjusted quickly to changes in the price level, so that variations in the price level over the cycle were an additional source of influence on production decisions. The nature of this influence led entrepreneurs to expand their activities further in the boom (as rising prices increased their profits) and contract them further in the slump (as falling prices reduced their profits) than would have been the case in the corresponding cooperative economy. But these changes in income distribution were not permanent. In the course of the boom workers managed to restore real wages to pre-recovery levels, the expansion of output would be slowed, and in the slump, as entrepreneurs restored profits to their pre-depression levels, the contraction of output would be slowed. The end of the boom and the slump, though, if an 'appropriate' monetary policy were adopted, would be dictated by the behaviour of the underlying non-cooperative non-monetary economy. So non-cooperation in the monetary case had additional effects only on the amplitude of cyclical fluctuations.

Robertson also developed a set of tools to analyse the process of cyclical change in monetary economies. He divided time up into a sequence of market periods (during each of which the supply of goods was fixed) and then focused on the dynamics of the transfer of resources from current consumption by those already in employment to those newly employed to increase output during the expansion phase of the cycle. The mechanism generating this transfer was a price-level increase as the newly employed (whose wages had been borrowed from the banking system) outbid the existing employed on the goods market. Robertson's aim was to show how the magnitude of the price-level increase was determined and the nature of the monetary policies which could be adopted in order to minimize it. The rate of inflation was shown to depend upon the relationship between the rate at which the banking system made new loans to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
producers and the rate at which this new money was absorbed into the money-holdings of the existing employed. The faster the new money was absorbed, that is, the faster that the existing employed gave up their claims on current output, the smaller the rise in the price-level accompanying the transfer of resources from the public to the expanding producers. To the extent that this money was not immediately absorbed, the existing employed were *forced* to share current output with those producers by price-level changes. By minimizing these changes, then, the monetary authority through its control of the banking system would also be able to minimize the amount of ‘forced’ saving which accompanied the recovery. A similar approach was also applied to the non-cooperative case, but here policy design was more difficult because the inflation led to changes in the distribution of income between workers and entrepreneurs. Even so, monetary policy could play a useful role in reducing fluctuations to their ‘irreducible’ non-monetary amplitude.

The central concern of Robertson's analytical work was to provide an explanation of fluctuations in aggregate activity which was closely linked to a broader concern, that of remedying the adverse effects of such fluctuations. The identification of the use of capitalistic (though not necessarily capitalist) production methods as the source of fluctuations, though, left with a rather ambivalent attitude to possible remedies: capitalistic production methods always produced cycles, but also brought with them the possibility of economic progress. And he thought that there was a trade-off between these two, greater stability being associated with slower growth, less stability with faster growth:

> From some points of view the whole cycle of industrial change presents the appearance of a perpetual immolation of the present upon the altar of the future. During the boom sacrifices are made out of all proportion to the enjoyment over which they will ultimately give command: during the depression enjoyment is denied lest it should debar the possibility of making fresh sacrifices. Out of the welter of industrial dislocation the great permanent riches of the future are generated. (1926, p. 254)

He concluded that the choice between these two conflicting goals was ultimately a question of: ‘ethics, rather than economics’.

The theoretical framework sketched above had emerged by the early 1930s. But its further development was interrupted by the publication in 1936 of Keynes's *General Theory of Employment, Interest and Money*. Robertson's response to this book was to examine how the *General Theory* might affect his vision of how the world worked. The central issue for Robertson was whether Keynes had provided a more satisfactory explanation than he had himself of the forces which determined the behaviour of the trend rate of growth of economic activity. The distinguishing feature of Keynes's approach identified by Robertson was in the treatment of the theory of the rate of interest. He interpreted as Keynes's central proposition the contention that there was an inherent tendency for the rate of interest to remain above the level consistent with the maintenance of full employment. And although Robertson was prepared to accept that an argument could, in principle, be made out along such lines he did not accept that Keynes had succeeded in doing so. In particular he maintained that while ‘liquidity preference’ might make the interest rate ‘sticky’ in the short period, with its downward movement resistant to monetary expansion, he rejected such an approach to the long-period theory of interest rate determination, summarizing the argument in the following way:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> … the rate of interest is what it is because it is expected to become other than it is; if it is not expected to become other than it is, there is nothing left to tell us why it is what it is. The organ which secretes it has been amputated, and yet it somehow still exists – a grin without a cat. (‘Mr Keynes and the Rate of Interest’ in *Essays in Monetary Theory*, 1940, p. 36)

Keynes's theoretical argument was, therefore, flawed. And the associated case for stabilizing the economy at a level other than that identified in Robertson's own analysis as ‘appropriate’ was consequently not proven.

The first repercussion of this reaction to the *General Theory* was an estrangement between Robertson and Keynes, virtually ending a close friendship which had lasted for more than 20 years (Robertson having been a student of Keynes, then a fellow teacher and collaborator in research). It then motivated Robertson's decision to leave Cambridge for London in 1938. Moreover, even after Keynes's death in 1946, strained and difficult relations with Keynes's disciples in Cambridge left him a somewhat isolated figure. The impact of Keynes's *General Theory* on Robertson's professional life was no less significant, the whole terrain of the area in which he worked was changed: from being on the creative frontier of the subject he felt himself forced into the role of commentator and critic. In the years after 1936 he wrote almost nothing new in what had been his specialist field. An explanation was provided in a letter to a friendly reviewer of one of his collections of essays who had called upon Robertson to prepare a monograph combining and extending his earlier analytical work, and to whom he wrote:

> … I'm afraid there is no chance of my responding to your challenge and trying to produce a full length synthetic Theory of Money or Fluctuations or What-you-will. I'm too old and too lazy! But even if I were younger and less lazy, I think history had made it impossible. I believe that once Keynes had made up his mind to go the way he did it was my particular function to … [elucidate and criticise the details of his work] … and to go on pegging away at them (as is still necessary). It will not be easy for *anyone* for another twenty years to produce a positive and constructive work which is not in large measure a commentary on Keynes, – that is the measure of his triumph. For me, it would now be psychologically impossible, and the attempt is not worth making. (Private letter of D.H. Robertson to T.J. Wilson, 31 October 1953)

Selected works

1915. *A Study of Industrial Fluctuation*. London: P.S. King & Son. Reprinted with a new introduction, in *Reprints of Scarce Works on Political Economy*, London: London School of Economics and Political Science, 1948.

1922. *Money*. Cambridge Economic Handbook. London: Nisbet & Co. Revised edn, 1924; new edn, 1948.

1923. *The Control of Industry*. Cambridge Economic Handbook. London: Nisbet & Co. Rev. edn, 1928;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

new edn (with S.R. Dennison), 1960.

1926. *Banking Policy and the Price Level*. London: P.S. King & Son. Reprinted, 1926; reprinted with revisions, 1932; reprinted in the USA with a new preface, New York: Augustus M. Kelley, 1949.

1931a. *Economic Fragments*. London: P.S. King & Son.

1931b. (With A.C. Pigou.) *Economic Essays and Addresses*. London: P.S. King & Son.

1940. *Essays in Monetary Theory*. London: P.S. King & Son.

1950. *Utility and All That*. London: George Allen & Unwin.

1956. *Economic Commentaries*. London: Staples Press.

1957–9. *Lectures on Economic Principles*. 3 vols. London: Staples Press. Paperback edn in one volume, London: Fontana, 1963.

1960. *Growth, Wages, Money*. The Marshall Lectures at the University of Cambridge. London: Cambridge University Press.

Bibliographic addendum

Fletcher, G. 2007. *Dennis Robertson: Essays on his Life and Work*. Basingstoke: Palgrave Macmillan.

How to cite this article

Danes, M. Anyadike. "Robertson, Dennis (1890–1963)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000162> doi:10.1057/9780230226203.1447

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Robinson Crusoe

M.V. White
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capital; commodity exchange; consumer surplus; individualism; marginalist economics; producer surplus; rational behaviour; Robinson Crusoe; supply and demand

Article

Written by Daniel Defoe, *Robinson Crusoe* was first published in 1719–20. By the end of the 19th century there were many references made to a Crusoe economy to illustrate the principles of supply and demand economic theory. Crusoe thus became a representative rational economic individual, allocating his available resources to obtain maximum satisfaction in the present or future.

The figure of Crusoe as the personification of supply and demand economic theory can be found in W.S. Jevons's *Theory* (1871), C. Menger's *Principles* (1871), P. Wicksteed's *Alphabet of Economic Science* (1888), E. Böhm-Bawerk's *Theory of Capital* (1890), A. Marshall's *Principles* (1891) and K. Wicksell's *Value, Capital and Rent* (1893). The principal uses of the device were to show how an isolated individual would allocate consumption items so as to maximize utility in a marginalist fashion and distribute labour effort between producing items for consumption or investment (creating ‘capital’). Calculations were made according to the relative amounts of pleasure and pain immediately or ultimately involved in the various activities. Marshall also used Crusoe to illustrate producer and consumer surplus, while F.Y. Edgeworth's *Mathematical Psychics* (1881) introduced ‘the black’, Friday, when discussing issues in the theory of commodity exchange.

The role of a Crusoe economy was not simply to illustrate various components of supply and demand theory. It was also utilized to support the claim that the principles of rational behaviour, as defined by that theory, could be applied to any type of economy – from the isolated individual to ‘modern civilization’. This point was made particularly clear in J.B. Clark's *The Distribution of Wealth* (1899). Similar references to a Crusoe economy can be found in textbooks today.

Two general characteristics of an economic Crusoe's actions are important to note. First, he must be able to calculate in a precise fashion making fine decisions between whether to work or rest, to consume or save/invest. Second, he has no resources other than those available in the island environment. Both

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
characteristics mean the economic Crusoe bears no relation to the Crusoe in Defoe's novel. Defoe's Crusoe wastes time because he cannot calculate in a marginalist fashion; he cannot rationally allocate labour time because labour is as useful in one pursuit as another; and he would not have survived without items salvaged from the shipwreck. Other decisions, such as whether to consume or save, also preclude calculation on the basis of relative pleasure and pain (White, 1982). Moreover, the relation between Crusoe and Friday is not based on voluntary reciprocal exchanges, as in the supply and demand parable, but rather on power and violence (Hymer, 1980). The economic Crusoe could not, therefore, have been produced by relying on the letter of Defoe's text.

It is possible to find some references to Crusoe by English political economists during the 1830s, but these were sporadic and no attempts were made to utilize Crusoe in a systematic fashion. An economic Crusoe thus appears only after mid-century with references in F. Bastiat's *Economic Harmonies* (1850) and H. Gossen's *Entwickelung* (Gossen, 1854). These references owed a good deal to the rewriting of Defoe's text within the literary genre of the ‘Robinsonade’.

The Robinsonade literature dates from the early 18th century (Gove, 1941) and includes voyage or shipwreck narratives, imaginary voyages to ‘isolated lands' and more general discussions of colonial settlements which depicted various stages of societal development. The last group of Robinsonade texts bears some resemblance to the four-stage theory of societies produced during the Scottish Enlightenment, remnants of which can be found in the work of the classical political economists (Meek, 1976). One such remnant was the illustrative device, used by A. Smith and D. Ricardo, of hunters exchanging commodities according to the labour embodied in them. While Marx was critical of this device, he noted it made sense in the context of the previous century's Robinsonades. However he considered the later discussion of Crusoe by Bastiat for example, was ‘twaddle’ because it depicted an individual ‘outside society’ (Marx, 1857–8, pp. 83–5).

Bastiat's Crusoe relied on a different type of Robinsonade literature, particularly J.H. Campe's *Robinson the Younger* (1779/80). Campe rewrote Defoe's tale to show Crusoe's survival on the island was not dependent on the shipwreck items. Gossen also appealed to Campe's novel to illustrate his marginalist explanation of work and consumption decisions. By the mid-19th century, then, the ‘individualist’ Robinsonade was utilized by those theorists who conceptualized the economy as a series of voluntary exchanges, where the principles of economic activity were those of the individual writ large.

English supply and demand economists could also draw upon a discernible shift in the readings of Defoe's text by literary commentators after 1850. Earlier commentary had stressed the novel was useful for showing, especially to children and the ‘working classes’, the virtue of work and the need to accept the given social organization ordained by Divine Providence. Commentary after mid-century represented Crusoe more as an individual calculating costs and benefits in the manner of an English shop keeper. It was even argued Crusoe represented a ‘national ideology’ in that regard. The remarkable similarity between this Crusoe and the illustrative device of English supply and demand economic theory suggests the latter was able to appropriate the former as a recognizable referent.

The economic Crusoe served, in effect, as a useful defensive device against ‘historical’ critics of economic theory such as T.E. Cliffe Leslie and J.K. Ingram. Writing between the mid-1860s and early 1880s, the critics argued that there were no universal laws of economic behaviour since behaviour could change according to the type of society being considered. Supply and demand theorists, such as Jevons, rejected that criticism, claiming historical studies could only confirm the ‘universal’ laws of behaviour

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

assumed in the theory (Jevons, 1876, pp. 196–7). In this context, the economic Crusoe provided an apparently tangible reference point when supply and demand theory began its analysis with the actions of an ‘isolated’ or representative individual. Indeed, Gossen had used Campe's Crusoe in precisely that fashion when criticizing the German ‘National Economists’ in 1854 (Gossen, 1854, pp. 45–7). The role of an economic Crusoe, as both illustrative and defensive device for supply and demand theory, was thus inscribed from its inception.

See Also

- economic man
- rational behaviour

Bibliography

Gossen, H.H. 1854. *Entwickelung der Gesetze des menschlichen Verkehrs und der daraus fliessenden Reglen für menschliches Handeln*. Brunswick: Vieweg.

Gove, P.B. 1941. *The Imaginary Voyage in Prose Fiction*. New York: Columbia University Press.

Hymer, S. 1980. Robinson Crusoe and the secret of primitive accumulation. In *Growth, Profits and Property*, ed. E.J. Nell. New York: Cambridge University Press.

Jevons, W.S. 1876. The future of political economy. In *The Principles of Economics and Other Papers*, ed. H. Higgs. London: Macmillan. 1905.

Marx, K. 1857–8. *Grundrisse*. Harmondsworth: Penguin, 1973.

Meek, R.L. 1976. *Social Science and the Ignoble Savage*. Cambridge: Cambridge University Press.

White, M.V. 1982. Reading and rewriting. The production of an economic Robinson Crusoe. *Southern Review* 15(2), 115–42.

How to cite this article

White, M. V. "Robinson Crusoe." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000163> doi:10.1057/9780230226203.1448

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Robinson, Edward Austin Gossage (1897 – 1993)

Z.A. Silberston
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

monopoly; optimum firm size; Robinson, E. A. G.; Robinson, J. V.

Article

Austin Robinson was educated at Marlborough College and Christ's College, Cambridge. During the First World War he served as a pilot in the RNAS and the RAF. After finishing his studies at Cambridge he became a Fellow of Corpus Christi College, from 1923 to 1926. In 1926 he married Joan, daughter of Major-General Sir Frederick Maurice and later to become the eminent economist. From 1926 to 1928 Robinson was tutor to the Maharaja of Gwalior. He returned to Cambridge as a university lecturer in economics in 1929, and from then on was an important figure on the Cambridge economics scene. He became Professor of Economics in 1950. He retired in 1965 (and it so happened that Joan Robinson was appointed to his chair). After his retirement, he continued to play a prominent role in Cambridge economics, as well as on the national and international scene.

Austin Robinson's first book, *The Structure of Competitive Industry* (1931), established his reputation as an economist. This seminal work drew on Alfred Marshall's writings on industry, and considered in detail the problems involved in determining the optimum size of firm. But although it emphasized the importance of scale, and inspired much of the later empirical work in this area, it also recognized that low British productivity in manufacturing industry was not primarily the consequence of scale, but of attitudes towards work and competition. All subsequent writing on this subject owed a considerable debt to Robinson. He followed up his work on competitive industry with a book on *Monopoly* (1941), as well as with a number of articles, including work on Africa. He was a member of the group surrounding Keynes when he was formulating the *General Theory*, and wrote a review of it in *The Economist*, insisting on signing it (against the traditions of the paper) because of the exceptionally controversial nature of the subject.

Robinson's long association with the *Economic Journal* began in 1934, as Assistant Editor to Keynes, and was later to be followed by much editorial work. Robinson did distinguished service during the war. He was a member of the Economic Section, War Cabinet Office, from 1939 to 1942, and from 1942 to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1945 was Economic Adviser and Head of Programmes Division, Ministry of Production. This was followed by a period as Economic Adviser to the Board of Trade. He returned to Cambridge in 1946, but served a further period in government on the Economic Planning Staff from 1945 to 1947. He was joint editor of the *Economic Journal* from 1944 to 1970, and played a leading role in the profession in other ways, holding a number of important posts, including that of managing editor of the Royal Economic Society's edition of Keynes's works. He was much involved in the work of the new International Economic Association: he was President from 1959 to 1962 and editor of its publications for many years. A good deal of his subsequent writing and editorial work, much of it on the problems of developing countries, was carried out in the context of the work of the IEA.

Austin Robinson's career was a remarkable one. He combined writing, teaching, editorial work and administration with advising governments in both the developed and developing world. He played a leading role in the economics profession for an exceptionally long period, internationally as well as in Britain, and did so throughout with much distinction.

See Also

- Royal Economic Society

Selected works

1931. *The Structure of Competitive Industry*. London: Nisbet & Co.; Cambridge: Cambridge University Press.

1941. *Monopoly*. London: Nisbet & Co.; Cambridge: Cambridge University Press.

How to cite this article

Silberston, Z. A. "Robinson, Edward Austin Gossage (1897–1993)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000165> doi:10.1057/9780230226203.1449

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Robinson, Joan Violet (1903 – 1983)

Luigi L. Pasinetti
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

animal spirits; beggar-thy-neighbour; Cambridge Circus; capital accumulation; capital theory; disguised unemployment; effective demand; employment theory; Euler's Theorem; firm, theory of; Harrod, R. F.; historical time and logical time; imperfect competition; investment; Kahn, R. F.; Kalecki, M.; Keynes, J. M.; logical time; Luxemburg, R.; Marshall, A.; Marshall–Lerner conditions; Marxist economics; multiplier; perfect competition; price formation; production function; reswitching of techniques; Robinson, E. A. G.; Robinson, J. V.; savings; Sraffa, P.; two-sector models; unorthodox economics; Wicksell, J. G. K.

Article

Joan Robinson (née Maurice) was born at Camberley, Surrey, on 31 October 1903. She died in Cambridge on 5 August 1983.
She is the only woman (with the possible, but controversial, exception of Rosa Luxemburg) among the great economists. In 1975, which was proclaimed Woman's Year, most economists in the United States expected that she would naturally be chosen for the Nobel Memorial Prize in Economics for that year. She had received triumphant acclaim, as a Special Ely Lecturer, at the American Economic Association annual meeting three years earlier, in spite of the harsh hostility that her theories had always met in the United States. The American magazine *Business Week*, after sounding out the American economics profession, felt so sure of the choice as to anticipate the event by publishing a long article on her, presenting her explicitly as being ‘on everyone's list for this year's Nobel Prize in Economics’. But the Swedish Royal Academy missed that opportunity (and alas, never regained it). Ever since, in shop-talk among economists all over the world, Joan Robinson has become the greatest Nobel Prize winner that never was.

Basic biography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Joan Robinson was the daughter of Major General Sir Frederick Maurice and of Helen Marsh (who was herself the daughter of a Professor of Surgery and Master of Downing College, Cambridge). Sir Frederick pursued a brilliant career in the British Army, but in 1918 he found himself at the centre of a public debate, and he gave up his army career on a point of principle. This was very much in the family tradition. Sir Frederick's grandfather – Joan Robinson's great-grandfather – was Frederick Denison Maurice, the Christian Socialist who lost his chair of theology at King's College London, for his refusal to believe in eternal damnation.

Joan Robinson certainly had many of these traits: toughness and endurance of character, nonconformism and unorthodoxy of views, the absence of any reverential feeling or timidity, even in the face of the world's celebrities, a passionate longing for the new and the unknown.

She was educated at St Paul's Girls' School in London. (Curiously enough, Richard Kahn was educated in the boys' section of the same school.) In October 1922, she was admitted to the University of Cambridge, going up to Girton College, where she read economics at a time when the dominant figures in Cambridge were Marshall and Pigou. Marshall had retired (he died in 1924) but was extremely influential not only in Cambridge but in the whole of the British Isles. Pigou, his favourite pupil and chosen successor, was the Professor of Political Economy, at whose lectures Cambridge students absorbed the official *verbum* of Marshallian economics. Keynes was a sort of outsider, part-time in Cambridge and part-time in London, always involved with government policies, either at the Treasury or in public opposition. In those days he lectured on strictly orthodox monetary theory and policies. His lectures were not given regularly but were well attended.

The intellectual environment must have appeared solidly traditional. Joan graduated in 1925, as a good girl would: with second class honours.

In the following year (1926), she married E.A.G. Robinson (later Professor Sir Austin Robinson), who was six years her senior and at the time a junior Fellow of Corpus Christi College. Together they left Cambridge and set off for India, where they stayed for two years. Austin Robinson served as tutor of the Maharajah of Gwalior. Joan was there as Austin's wife but did some teaching at the local school. When they returned, after their two-year Indian engagement, Austin Robinson took a permanent post as Lecturer in Economics at Cambridge, where they settled for life. They had two daughters.

It was on the return to Cambridge from India (summer 1928) that Joan Robinson began to do some College supervision of undergraduates, and then to do economics research in earnest. The Cambridge intellectual environment had changed dramatically. After Edgeworth's death (1926), Keynes became the sole editor of the *Economic Journal* and was engaged on his *Treatise on Money* (Keynes, 1930). Most of all, he had brought to Cambridge Piero Sraffa, the young Italian economist who had dared to launch a scathing attack on Marshallian economics (Sraffa, 1926). Moreover, some new stars were rising in the firmament of Keynes's entourage – Frank Ramsey, the brilliant mathematician; Ludwig Wittgenstein, the Austrian philosopher whom Keynes persuaded to come to Cambridge; and Richard Kahn, Keynes's favourite pupil. It was with Richard Kahn that Joan Robinson began an intense intellectual partnership that lasted for her whole life.

On a strictly academic level, Joan Robinson slowly ascended the academic ladder: Junior Assistant Lecturer in 1931, Full Lecturer in 1937, Reader in 1949. It was suggested in Cambridge that the fact that her husband was in the same faculty kept her back at all stages of her academic career. She became full professor only on Austin Robinson's retirement, in 1965. Her association with the Cambridge colleges was more irregular. But she was, in succession, a Fellow of Girton College and of Newnham College.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Yet whatever the formal position in the faculty or in the Cambridge colleges, she was for years one of the major attractions in Cambridge for many generations of undergraduates, not only in economics. In the post-war period, she was certainly the best-known member of the Cambridge economics faculty abroad. An indefatigable traveller, she did not limit her foreign visits to universities; she also wanted to know local customs and local conditions of life, even far away from urban centres. Her strong constitution and temperamental toughness helped her enormously. A friend from Makerere University, who took her, when she was already in her seventies, on a month's travel in tribal Africa was amazed at how much she could endure in terms of living in most primitive conditions with raw food, lack of facilities and exposure to harsh tropical weather, day and night.

It would be impossible to list here all the places she visited or the talks, seminars and public lectures she gave all over the world. She rarely stayed in Cambridge during the summer or term vacations or during her sabbatical years, though punctually and punctiliously returning there on the eve of the terms of her teaching. Asia was her favourite continent (especially India and China). But hundreds of students in North and South America, Australia, Africa and Europe also knew her at first hand.

In Cambridge she rarely missed her classes, lectures and seminars and she was a regular attendant of other people's seminars, especially visitors', never avoiding discussion and confrontation. Professor Pigou – a well-known misogynist – had included her in his category of ‘honorary men’.

She was extremely popular with the students – a clear, brilliant, stimulating teacher. She was a person who inspired strong feelings – of love and hate. Her opponents were frightened by her, and her friends really admired, almost worshipped her. Her nonconformism in everyday life and even in her clothing (most of which she bought in India) was renowned.

She retired from her professorship in Cambridge on 30 September 1971. On retirement she did not agree to continue lecturing in Cambridge. (Later on, in the late 1970s, she gave in partially, giving a course of lectures on ‘the Cambridge tradition’.) But her writing and lecturing abroad, at the invitation of economics faculties and students all over the world, continued unabated.

When, in the late 1970s, King's College (Keynes's college) finally dropped the traditional anachronistic ban on women and became co-educational, Joan Robinson, upon an enthusiastic and unanimous proposal by all economists of the college, became the first woman to be made an Honorary Fellow of King's College. (She had earlier become an Honorary Fellow of Girton College and of Newnham College.)

Towards the end of her life, she became very concerned and disappointed with the direction in which economic theory had turned and with the ease with which the younger economists could bend their elegant models to suit the new conservative moods and the selfish economic policies of politicians and governments. Her friends also noticed a sort of stiffening rigidity in her views that had not appeared before. This was unfortunate, as it contributed to increasing the hostility of her opponents towards her. She suffered a stroke in early February 1983, from which she never recovered. She lay for a few months in a Cambridge hospital, and died peacefully six months later.

Distinctive traits of her intellectual personality

In order to understand better the nature of Joan Robinson's contributions to economic theory, it may be helpful to begin by considering explicitly a few characteristic traits of her intellectual personality.

Joan Robinson had a remarkable analytical ability. Since she did not normally use mathematics, this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
remarkable intellectual ability was of a nature that defies conventional description. In her early works she made use of geometrical representations, backed up by calculus (normally provided by Richard Kahn). In her mature works, her way of reasoning took up a more personal feature. Her style is difficult to imitate (as when she invites the readers to follow her in the construction of economic exercises) but very effective. The results are always impressive. Those who used to argue with her knew that she could grasp and keep in the back of her mind (to be brought out at the appropriate moment) a whole series of chain effects and interdependences which her interlocutors could hardly imagine.

She was not the type of person who could go on thinking in isolation. The way she could best express herself was by having somebody in constant confrontation. She could put her views best either in opposition or in support of somebody else's position. This made her extraordinarily open to concepts and contributions coming from the people she encountered. The accurate historian of economic ideas will probably find in her works traces of almost every person she met. It is therefore important, in considering Joan Robinson's contributions, to keep in mind at least the most important economists who influenced her. These include her teachers (Marshall through Pigou, Keynes, Shove), her contemporaries (Sraffa, Kaldor, and Kalecki, through whom she went back to Marx, but especially Richard Kahn, who read, criticized and improved every single one of her works) and also a whole series of other (younger) people – pupils and students.

This raises the question of her originality. The prefaces to her books are packed with acknowledgements, sometimes heavy acknowledgements – consider, for example, the following excerpt from the *Economics of Imperfect Competition*:

> … this book contains some matter which I believe to be new. Of not all the new ideas, however, can I definitely say that ‘this is my own invention’. I particularly have had the constant assistance of Mr R.F. Kahn … many of the major problems … were solved as much by him as by me. (Robinson, 1933, p. v)

But one must remember what has been said above. In fact, Joan Robinson was a highly original thinker, but of a particular type. Besides the contributions to economic theory that are distinctly hers she had her own highly original way, even in small details, of presenting other authors' views, which she always did through a distinctly personal re-elaboration. Sometimes the re-elaboration is so personal as to sound parochial. But this trait is not exclusive to Joan Robinson. Cambridge parochialism is shared by almost all purely Cambridge-bred economists since Marshall (Keynes included). It sometimes creates unnecessary difficulties of communication with economists outside Cambridge (that is, with the overwhelming majority!) or introduces a few odd notes into an otherwise impeccable performance.

One can clearly detect an evolution in Joan Robinson's approach to economics that with age strengthened her innovative tendencies. It looks as if she was very cautious in her early years, preoccupied at first with building up solid analytical foundations. But as soon as she felt sure of her analytical equipment, she began to venture more and more into the exciting field of innovation. In her mature works her typical style became established. A sort of mixture of educational, temperamental and intellectual factors made her one of the leading unorthodox economists of the 20th century. Always impatient with dogmas, constantly fighting for new unorthodox ideas, relentlessly attacking established beliefs, she acquired a sort of vocation to economic heresies (see Robinson, 1971). Her attitude reminds one of a dictum by Pietro Pomponazzi, the Italian Renaissance philosopher: ‘It is better to be a heretic if

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
one wishes to find the truth.'
Strongly related to this attitude is the social message that comes from her writings. Her 'box of tools' and her logical chain of arguments were not proposed for their own sake; they were always aimed at practical action, with a view to the world's most pressing problems – unemployment before the war, underdevelopment and the struggle of ex-colonial nations after the war (very noticeable is her special concern for Asia and her enthusiasm, at points rather naive, for Communist China). Consistently, she has been among the strongest assertors – second perhaps only to Gunnar Myrdal – of the non-neutrality of economic science and of the necessity of stating explicitly one's convictions and beliefs.
And yet, in spite of her bold attacks and her satirical mood, her literary style is surprisingly feminine – rich with fable-like parables, with down-to-earth examples from everyday life ('the price of a cup of tea …') and with similes from scenes and examples taken from nature (the *Accumulation of Capital* begins with the economic life of the robin). Her sparkling prose and her entertaining asides make Joan Robinson one of the most brilliant writers among economists and certainly one of the most enjoyable and delightful to read.

Her scientific achievements

Joan Robinson wrote numerous books and an enormous number of articles, most of which have been collected in her *Collected Economic Papers* (1951–79).
They fall neatly into three broad groups, corresponding to the three basic phases of her intellectual development. A first group belongs to the phase of her by now classic *Economics of Imperfect Competition* (1933). A second group belongs to the phase of explanation, propagation and defence of Keynes's *General Theory*. Finally, a third group of writings grew around the major work of her maturity, *The Accumulation of Capital* (1956). Other books and articles have originated from miscellaneous or wider interests or from the desire to provide students with economics exercises or with a non-orthodox economics textbook (Robinson and Eatwell, 1973c). Altogether, they make an impressive list. Even neglecting her articles (most of which are reprinted in the books), her bibliography contains no fewer than 24 books.
The most widely known of Joan Robinson's works is still the first, *The Economics of Imperfect Competition* (1933). It was the book of her youth, which placed her immediately in the forefront of the development of economic theory. It is a work conceived in Cambridge, at the end of a decade characterized by an intense controversy on cost curves and the laws of returns (see Sraffa, 1926, and the Symposium on the 'laws of returns' by Robertson, Sraffa and Shove, 1930). With this controversy in the background, Joan Robinson's book emerges in 1933 as a masterpiece in the traditional sense of the word. The restrictive conditions of perfect competition on which Marshall's theory was constructed are abandoned, and perfect competition is shown to be a very special case of what in general is a monopolistic situation. A whole new analysis of market behaviour is carried out on new, more general, assumptions; and yet the whole method of analysis, the whole approach – though refined and perfected – is still the traditional Marshallian one. Sraffa's criticism of the master is accepted, but is incorporated into the traditional fold by a generalization of Marshall's own theoretical framework. The outcome is extremely elegant and impressive. The whole matter of market competition is clarified. Marshall's ambiguities are eliminated, the various market conditions are rigorously defined, a whole technical apparatus (a 'box of analytical tools') is developed to deal with market situations in the general case

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(from demand and supply curves to marginal cost and marginal revenue curves). In a sense, therefore, rather than a radical critique, the *Economics of Imperfect Competition* might well be regarded as the completion and coronation of Marshallian analysis. This may help to explain why Joan Robinson herself came to like that book less and less, as her thought later developed on different lines. In 1969 she came to the point of writing a harsh eight-page criticism of it. Very courageously she published it, on the occasion of a reprint of the book, as a Preface to the second edition!

The book had appeared almost simultaneously with the *Theory of Monopolistic Competition* by Edward Chamberlin (1933); and the two books are normally bracketed together as indicating the decisive breakaway of economic theory from the assumptions of perfect competition. Chamberlin always complained about this association. For although the two books represent the simultaneous discovery of basically the same thing, made quite independently by two different authors, they are in fact substantively different.

It may also be added that, looked at in retrospect, these two books do not appear so conclusive a contribution to the theory of the firm as they appeared to be in the 1930s. The behaviour of firms in oligopolistic markets and the policies of the large corporations have turned out to require more complicated analysis. At the same time, the assumption of perfect competition, far from being completely dead, has recently come back in different guises in the works of many theoretical economists. Yet there is no doubt that the two books remain there to represent a definite turning-point in the development of the theory of the firm – so much so as to be referred to as representing the ‘monopolistic competition revolution’ (Samuelson, 1967). Very characteristically, Edward Chamberlin, after writing the *Theory of Monopolistic Competition*, spent the whole of his life in refining, completing and adding appendices to his masterpiece (no fewer than eight editions!). For Joan Robinson, the *Economics of Imperfect Competition* was only the first step on a very long way to a series of works in quite different and varied fields of economic theory.

It should be added that the *Economics of Imperfect Competition* was not Joan Robinson's only contribution to microeconomic theory in the 1930s. Her name appears again and again on the pages of the avant-garde economic journals of the time. From among her papers, explicit mention must be made at least of her remarkably lucid article on ‘rising supply price’ and of her contribution to clarifying the meaning of Euler's theorem as applied to marginal productivities, in the traditional theory of production (see her *Collected Papers*, vol. 1).

But something of extraordinary importance was happening in Cambridge in the 1930s. Keynes was in the process of producing his revolutionary work (Keynes, 1936). Joan Robinson abandoned the theory of the firm and threw herself selflessly and entirely into the new paths opened up by him. This was a really brave decision, if one thinks that her first book had gained her great reputation in the economic profession. Very rarely do we find someone who, after striking success and becoming a leading figure in a certain field, pulls out of it and puts himself or herself into the shadow of someone else, be this someone else even of the stature of Keynes. Joan Robinson did precisely that. She was one of the members – actually an important member, as is revealed by the recent publication of her correspondence with Keynes (see Keynes, 1973; 1979) – of that group of young economists known as the ‘Cambridge Circus’ (and including Kahn, Sraffa, Harrod, Meade, besides Austin and Joan Robinson) who regularly met for discussion, and played a crucial role in the evolving drafts of Keynes's *General Theory*.

It must be said that the new Keynes's ways were more congenial to her temperament. They were a break with tradition and this suited her nonconformist attitude; they dealt with the deep social problems of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
unemployment and this appealed to her social conscience. It is in this vein that she published her *Essays in the Theory of Employment* (1937a) and her *Introduction to the Theory of Employment* (1937b). These twin books were simply meant to be a help to the readers of Keynes's *General Theory*. In fact, they turned out to be much more than that. In particular, Joan Robinson contributes to the clarification of a major piece of Keynesian theory – the process through which investments determine savings – which had remained rather obscure from the *General Theory*. For her, this appeared important because it broke a crucial link in traditional theory, which presented the rate of interest as a compensation for the 'sacrifice' of supplying capital (that is, for saving). Joan Robinson stresses the role of investment as an independent variable, while total saving is shown as being determined passively by investment through the operation of the multiplier; the conclusion being that the rate of interest cannot be remunerating anybody's 'sacrifice'. Even more so in depression times, when thrift – a 'private virtue' – becomes a 'public vice'. Other concepts, introduced by Joan Robinson at the time, that were to remain permanently in the following economic literature on the theory of employment are those concerning what she called 'beggar-my-neighbour' policies, 'disguised unemployment' and the generalization of the Marshall–Lerner conditions on international trade, in terms of 'the four elasticities'.

Towards the end of the 1930s, Joan Robinson met Kalecki, and discovered that quite independently of, and in fact earlier than, Keynes he had come to the same conclusions. Kalecki had started from a Marxist background, against which Keynes was prejudiced. This led her to re-reading Marx and to re-thinking her own position vis-à-vis Marxian theory (Robinson, 1942).

Joan Robinson's flirtation with Marx is very curious. It has all the charm of a meeting and all the clamour of a clash. She is no doubt attracted by Marx's general conception of society. She finds in Marx much that she approves of. But she finds his scientific nucleus embedded in, and in need of being liberated from, ideology. To obtain this, she says, one must work hard. Her writings on Marx are specifically aimed at 'separating the wheat of science from the chaff of ideology'. Needless to say, this has caused her a lot of trouble with the Marxists. It should be kept in mind that in Continental Europe discussions on Marx have a long and complex tradition of philological heaviness and ideological passion. Joan Robinson's discussion is short and simple. She is always looking at Marx as 'a serious economist'. Accordingly, she always tries to go straight to what she thinks is his economic analysis. Her insistence on the necessity of rescuing Marx, as a scholar and a first-rate analytical mind, has recently been vindicated, especially after the publication of Sraffa's book (1960; see also, for example, Samuelson, 1971).

But the post-war period was opening up new vistas. With Keynes's *General Theory* in the background, Joan Robinson saw a formidable task ahead, consisting in nothing less than a reconstruction of economic theory. This led, after a decade of intense work, to the publication of her second major contribution to economic theory – *The Accumulation of Capital* (1956), the work of her maturity and the one that expresses Joan Robinson's genius at her best. Here she has chosen to move on new and controversial ground. While in her first book the direction – once established – was clear and she had to fill in the details, here the direction itself is not entirely clear and has to be continually adjusted. The details acquire less importance and may well be abandoned altogether and replaced with others at a second attempt. As a consequence, a lot of re-writing had to be done.

The Rate of Interest and Other Essays (1952), with its central essay devoted to a 'Generalization of the *General Theory*', turned out to be a sort of preparation. *The Accumulation of Capital* represents the central nucleus of what she perceived as a new framework for economic theory. Then the *Exercises in*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Economic Analysis (1960a), the *Essays in the Theory of Economic Growth* (1962a) and a series of other articles fill in the gaps, clarify obscurities, and take the arguments further.

The 'Generalization of the *General Theory*' represents Joan Robinson's response to an interchange with Harrod, following Harrod's *Towards a Dynamic Economics* (1948) and also his earlier review of her *Essays in the Theory of Employment* (1937a). Joan Robinson breaks away from the limitations of the short run, but has not yet defined clearly her direction. Yet, once the process of 'generalization', that is, 'dynamization' of the *General Theory* is started, the author is compelled to recast the Keynesian arguments in terms of the more fundamental categories of capital accumulation, labour supply, technical progress and natural resources. Through this recasting, it became inevitable that she should go to the earlier methodological approach (common to Ricardo and Marx) of stating the problems in terms of social aggregates. The evidence of her intense searching may be found at the end of the book in a chapter of 'acknowledgements and disclaimers', where she describes in succession the way she has been influenced by, or has reacted to, Marx, Marshall, Rosa Luxemburg, Kalecki and Harrod.

The years of transition from the *Rate of Interest and Other Essays* (1952) to the *Accumulation of Capital* (1956) had been marked by a series of intense discussions in Cambridge, especially with Kahn, Sraffa, Kaldor and Champernowne. In the end, Joan Robinson emerged centring her attention on the problem of capital accumulation as the basic process in the development of a capitalist economy. She began with a scathing attack on the traditional concept of 'production function' (in a well-known article, now in her *Collected Papers*, vol. 2, which elicited a chain of angry responses: see, for example, Solow, 1955–6, and Swan, 1956). Then she patiently proceeded to a reconstruction. A crucial step was her own way of rediscovering the Swedish economist Knut Wicksell.

The Accumulation of Capital (1956) bears the same title as Rosa Luxemburg's book, to whose translation into English Joan Robinson wrote an introduction (Luxemburg, 1951). This was a great tribute to another woman economist. But we should not be misled. Joan Robinson's book belongs to an entirely different age and takes an entirely different approach. Set into a Keynesian framework extended to the long run, it takes its origin from a welding together of Harrod's economic dynamics and of Wicksell's capital theory. The main question Joan Robinson poses to herself is by now a typically classical one: what are the conditions for the achievement of a cumulative long-term growth of income and capital (what she characteristically christened a 'golden age'); and what is the outcome of this process, in terms of growth of gross and net output and of the distribution of income between wages and profits, given a certain evolution through time of the labour force and of technology? To answer these question Joan Robinson builds up a two-sector dynamic model with a finite number of techniques; and goes on to show the interactions of the relations between wages and profits, the stock of capital and the techniques of production, entrepreneurial expectations and the degree of competition in the economy, bringing in the effects of higher degrees of mechanization and both 'neutral' and 'biased' technical progress. The basic model and the basic answers are all worked out very quickly in the book. The rest is then devoted to relaxing the simplifying assumptions. The whole analysis is carried out *without* the use of mathematics. This is remarkable. Joan Robinson squeezes out of the model, one by one, all the answers that are needed. The non-use of mathematics has certain obvious disadvantages. Though the analysis need not necessarily be any less rigorous, in many passages it is not so easy to follow. It has, however, some advantages, which Joan Robinson is very ready and able to exploit. She succeeds, for example, in freeing herself from the symmetry that a mathematically formulated model normally imposes. In Joan Robinson's model, certain results are always more likely to happen than their

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
symmetrical counterpart. Symmetry and formal elegance play no part; only relevance does, or at least it does in the way perceived by the author.

The overall result is, again, impressive. The oversimplified dynamic model of Harrod is enormously enriched by the introduction of the choice among a finite number of alternative techniques. At the same time the Wicksellian analysis of accumulation at a given technology is completed by the new analysis of a constant flow of inventions of various types. And this marriage of Harrod's model to Wicksellian analysis is made to fructify in a number of directions. So many and so rich are in fact these directions that Joan Robinson herself did not pursue all of them, as became evident from the abundant literature that followed.

To this literature, Joan Robinson contributed a whole series of essays and books (see her *Collected Papers*, vols 2–5, and J. Robinson, 1960b; 1962a), which represent clarifications and further elaborations. They also represent her way of recasting and adjusting her arguments in response to opposition from her critics and to comments, remarks and stimuli of any sort from her friends, as well as her way of coming to grips with results – not always or not entirely compatible with hers – coming from the works of other scholars, colleagues and pupils, who were broadly working on similar problems and with the same aims.

Meanwhile, proceeding on parallel lines, many other separate strands of thinking were emerging from her remarkable intellectual activity. At least a few must briefly be mentioned here.

First, a whole series of concepts and ideas were coming to fruition, which – though not belonging to her major fields of interest – came to complete her overall coverage of economic theory: writings on the theory of international trade (including her professorial inaugural lecture at Cambridge on *The New Mercantilism*, 1966a), on Marxian economics (at various stages in her career), and on the theory of economic development and planning, reproducing her lectures delivered during her world travels or coming from calm reflection, once she had returned home (see her *Collected Papers*, and also J. Robinson, 1970b; 1979b).

Second, her deeply felt concern with economics students and economics teaching in general gave origin to books, such as Joan Robinson (1966b; 1971) and especially (with Eatwell) (1973c), which contributed to giving substance to, and disseminating all over the world, her strongly felt conviction that an overall approach to economic reality, alternative to that of traditional economics, does exist and is viable.

Third, the ideas, reflections, rationalizations, accumulated in the course of her life took the form of books such as *Economic Philosophy* (1962b) and *Freedom and Necessity* (1970a), which were concerned with wider issues than economics itself, attempting to give an overall conception of the world and a whole philosophy of life. These writings contribute, not marginally, to place Joan Robinson among the influential thinkers of this century. At the same time, they may well be enjoyed, by the general reader, even more than her masterpieces. From a purely literary point of view, they make delightful reading.

It should be added that there are, moreover, many themes which, while not being exclusively connected with any specific work of Joan Robinson's, recur time and again in her writings, so as to have become characteristically associated with her approach. Here are a few: (*a*) the concept of ‘entrepreneurs' animal spirits’ – an expression picked up from Keynes and developed as an important element contributing to explain investment in capitalist economies; (*b*) the conviction that Marshall's notions of prices and rate of profit, with reference to industry, are much more akin to Ricardo's notions than to Walras's; (*c*) a sharp distinction between ‘logical’ time and ‘historical’ time, both of which have a place in economic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
analysis but with different roles. On this point Joan Robinson's characterization of the evolution of an economy in historical time as concerning decisions to be taken between ‘an irrevocable past and an uncertain future’ is well known; (*d*) an equally sharp distinction between *comparisons* of equilibrium–growth positions and *movements* from one equilibrium–growth position to another, in dynamic analysis; (*e*) a tendency, especially in the later part of her life, to shift nearer and nearer to the positions of Kalecki, as opposed to those of Keynes, in interpreting the overall working of the institutions of capitalist economies, especially with reference to what she found as a more satisfactory integration in Kalecki of the concept of effective demand with the process of price formation.

Finally, one must mention specifically an issue which may well continue to give rise to controversial evaluations. This concerns the role that may be assigned to Joan Robinson in the well-known controversy on capital theory that flared up between the two Cambridges in the 1960s (see Pasinetti et al., 1966). One view on this issue is that Joan Robinson had the merit of anticipating the controversy by her (already mentioned) attacks on the neoclassical production function in the mid-1950s (see Harcourt, 1972). Another view is that Joan Robinson, herself a victim of her emotional temperament, started her attacks on the traditional concepts too early and misplaced the whole criticism, by neglecting the really basic point (the phenomenon of reswitching of techniques; see Sraffa, 1960) that in the end delivered the fatal blow to the neoclassical notion of production function. What one can say for certain is that a hint at the reswitching phenomenon does appear in the *Accumulation of Capital*, but is relegated to the role of a curiosum, in an entirely secondary section. Perhaps the phenomenon had been pointed out to her but she grossly underestimated its importance. What is curious is that she continued to underestimate it even after it was brought to the foreground (see her ‘Unimportance of Re-switching’ in *Collected Papers*, vol. 5).

But at this point the works of Joan Robinson merge into those of that remarkable group of Cambridge economists – notably, Piero Sraffa, Nicholas Kaldor and Richard Kahn, among others, besides Joan Robinson herself (on this, see the Preface to Pasinetti, 1981) – who happened to be concentrated in Cambridge in the post-war period and who took up, continued and expanded the challenge that Keynes had launched on orthodox economic theory. This remarkable group of economists started a stream of economic thought which is obviously far from complete. Its basic features, however, are clear enough; they embody a determined effort to shift the whole focus of economic theorizing away from the problems of optimum allocation of given resources, where it had remained for almost a century, and move it towards the fundamental factors responsible for the dynamics of industrial societies. This shift of focus inevitably brings into the foreground the once central themes of capital accumulation, population growth, production expansion, income distribution, and thus technical progress and structural change.

It is perhaps too early to try to evaluate the relative role played by Joan Robinson as a member of this remarkable group of economists. The single components of the group have made contributions which are sometimes complementary, at other times overlapping, and at yet other times even partly contradictory. To mention only one major problem, Piero Sraffa's book appeared too late for Joan Robinson to be able to incorporate it into her theoretical framework; and the brave efforts she later made to this effect are not always convincing. They actually reveal here and there a sort of ambivalent attitude. At the same time, her *Accumulation of Capital* ventures into fields of economic dynamics which Sraffa does not touch at all. Quite obviously, the common fundamental thrust behind post-Keynesian analysis

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
does not presuppose complete identity of views or complete harmony of approach.
Future developments will clarify issues and will reveal which of the lines of approach proposed are the most useful, fruitful or fecund. There can be little doubt, however, that if this theoretical movement is going to prove successful, quite a lot of rewriting will have to be done in economic theory. If, and when, this rewriting occurs, Joan Robinson's contributions are going to take a major place.

Selected works

1933. *The Economics of Imperfect Competition*. London: Macmillan. 2nd edn, 1969.

1937a. *Essays in the Theory of Employment*. London: Macmillan.

1937b. *Introduction to the Theory of Employment*. London: Macmillan.

1942. *An Essay on Marxian Economics*. London: Macmillan.

1951. *Collected Economic Papers*, vol. 1. Oxford: Basil Blackwell. (Vol. 2, 1960b; vol. 3, 1965; vol. 4, 1973a; vol. 5, 1979a.)

1952. *The Rate of Interest and Other Essays*. London: Macmillan.

1956. *The Accumulation of Capital*. London: Macmillan.

1960a. *Exercises in Economic Analysis*. London: Macmillan.

1960b. *Collected Economic Papers*, vol. 2. Oxford: Basil Blackwell.

1962a. *Essays in the Theory of Economic Growth*. London: Macmillan.

1962b. *Economic Philosophy*. London: C.A. Watts.

1965. *Collected Economic Papers*, vol. 3, Oxford: Basil Blackwell.

1966a. *The New Mercantilism – An Inaugural Lecture*. Cambridge: Cambridge University Press.

1966b. *Economics – An Awkward Corner*. London: Allen & Unwin.

1970a. *Freedom and Necessity*. London: Allen & Unwin.

1970b. *The Cultural Revolution in China*. London: Penguin Books.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1971. *Economic Heresies: Some Old-fashioned Questions in Economic Theory*. London: Macmillan.

1973a. Collected Economic Papers, vol. 4. Oxford: Basil Blackwell.

1973b ed. *After Keynes*. Papers presented to Section F (economics) of the 1972 annual meeting of the British Association for Advancement of Science. Oxford: Basil Blackwell.

1973c. (With J. Eatwell.) *An Introduction to Modern Economics*. New York: McGraw-Hill.

1978. *Contributions to Modern Economics*. Oxford: Basil Blackwell.

1979a. *Collected Economic Papers*, vol. 5. Oxford: Basil Blackwell.

1979b. *Aspects of Development and Underdevelopment*. Cambridge: Cambridge University Press.

1980. *Further Contributions to Modern Economics*. Oxford: Basil Blackwell.

Bibliography

Chamberlin, E. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Harcourt, G. 1972. *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.

Harrod, R.F. 1948. *Towards a Dynamic Economics*. London: Macmillan.

Keynes, J.M. 1930. *A Treatise on Money*. 2 vols. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1973, 1979. *The Collected Writings of John Maynard Keynes*, vols 13 and 14 (1973) and 29 (1979), ed. D.C. Moggridge. London: Macmillan for the Royal Economic Society.

Luxemburg, R. 1913. *The Accumulation of Capital*. Trans. A. Schwarzschild, with an Introduction by J. Robinson, London: Routledge & Kegan Paul, 1951.

Pasinetti, L.L. 1974. *Growth and Income Distribution: Essays in Economic Theory*. Cambridge: Cambridge University Press.

Pasinetti, L.L. 1981. *Structural Change and Economic Growth: A Theoretical Essay on the Dynamics of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the Wealth of Nations. Cambridge: Cambridge University Press.

Pasinetti, L.L., Levhari, D., Samuelson, P.A., Bruno, M., Burmeister, E., Sheshinski, E., Morishima, M. and Garegnani, P. 1966. Contributions to 'Paradoxes in Capital Theory – A Symposium'. *Quarterly Journal of Economics* 80, 503–83.

Robertson, D.H., Sraffa, P. and Shove, G. 1930. Increasing returns and the representative firm: a symposium. *Economic Journal* 40, 76–116.

Samuelson, P.A. 1967. The monopolistic competition revolution. In *Monopolistic Competition: Studies in Impact*, ed. R.M. Kuenne. New York: Wiley & Sons.

Samuelson, P.A. 1971. Understanding the Marxian notion of exploitation: a summary of the so-called transformation problem between Marxian values and competitive prices. *Journal of Economic Literature* 9, 339–431.

Solow, R.M. 1955–6. The production function and the theory of capital. *Review of Economic Studies* 23 (2), 101–8.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36, 535–50.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Wicksell, K. 1901. *Lectures on Political Economy*. Vol. 1, ed. L. Robbins, London: Routledge & Kegan Paul, 1934.

Bibliographic addendum

Essay collections on Robinson's contributions to economics include B. Gibson, ed., *Joan Robinson's Economics: A Centennial Celebration*, Cheltenham: Edward Elgar, 2005, and G.C. Harcourt, ed., *Joan Robinson: Critical Assessments of Leading Economists*, London: Routledge, 2002. A combination of biography and bibliography of Robinson's work is J. Cicarelli and J. Cicarelli, *Joan Robinson: A Bio-Bibliography*, Westport, CT: Greenwood Press, 1996.

How to cite this article

Pasinetti, Luigi L. "Robinson, Joan Violet (1903–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000166> doi:10.1057/9780230226203.1450

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

robust control

Noah Williams
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robust control is an approach for confronting model uncertainty in decision making, aiming at finding decision rules which perform well across a range of alternative models. This typically leads to a minimax approach, where the robust decision rule minimizes the worst-case outcome from the possible set. This article discusses the rationale for robust decisions, the background literature in control theory, and different approaches which have been used in economics, including the most prominent approach due to Hansen and Sargent.

Keywords

ambiguity; ambiguity aversion; control theory; error modelling; Kalman filtering; Lagrange multipliers; linear quadratic control; max-min expected utility; minimax; model uncertainty; optimal control; perturbation; probability distribution; risk aversion; robust control; uncertainty aversion; unstructured uncertainty

Article

Robust control considers the design of decision or control rules that fare well across a range of alternative models. Thus robust control is inherently about model uncertainty, particularly focusing on the implications of model uncertainty for decisions. Robust control originated in the 1980s in the control theory branch of the engineering and applied mathematics literature, and it is now perhaps the dominant approach in control theory. Robust control gained a foothold in economics in the late 1990s and has seen increasing numbers of economic applications in the past few years. (For related surveys see Hansen and Sargent, 2001; and Backus, Routledge and Zin, 2005. For a more comprehensive view of the leading approach to robust control in economics, see Hansen and Sargent, 2008.)
The basic issues in robust control arise from adding more detail to the opening sentence above – that a decision rule performs well across alternative models. To begin, define a model as a specification of a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
probability distribution over outcomes of interest to the decision maker, which is influenced by a decision or control variable. Then model uncertainty simply means that the decision maker faces subjective uncertainty about the specification of this probability distribution. A first key issue in robust control, then, is to specify the class of alternative models which the decision maker entertains. As we discuss below, there are many approaches to doing so, with the most common cases taking a benchmark *nominal model* as a starting point and considering perturbations of this model. How to specify and measure the magnitude of the perturbations are key practical considerations.

With the model set specified, the next issue is how to choose a decision rule and thus what it means for a rule to 'perform well' across models. In Bayesian analysis, the decision maker forms a prior over models and proceeds as usual to maximize expected utility (or minimize expected loss). Just as we defined a model as a probability distribution, a Bayesian views model uncertainty as simply a hierarchical probability distribution with one layer consisting of shocks and variables to be integrated over, and another layer averaging over models. In contrast, most robust control applications focus on minimizing the worst case loss over the set of possible models (a minimax problem in terms of losses, or max-min expected utility). Stochastic robust control problems thus distinguish sharply between shocks which are averaged over and models which are not. The robust control approach thus presumes that decision makers are either unable or unwilling to form a prior over the forms of model misspecification. Of course, decision makers must be able to specify the set of models as discussed above, but typically this involves bounding the set of possibilities in some way rather than fully specifying each alternative.

Finally, there are some approaches which seek a middle ground between the average case and the worst case, for example by maximizing expected utility subject to a bound on the worst case loss. These have been less prominent both in control theory (Limebeer, Anderson and Hendel, 1994, is one example) and in economics (Tornell, 2003, is one exception), and thus will not be discussed further. For the remainder of the article robust control will mean a minimax approach.

Robustness and worst case analysis

Broadly speaking, the control theory literature has adopted the worst-case philosophy out of concerns for stability. A basic desideratum for robust control in practice is that the system remain stable in the face of perturbations, and since instability may be equated with infinite loss, minimizing the worst case outcomes will insure stability (when possible). Moreover, many engineering applications have specific performance objectives which must be maintained, and a cost function penalizing deviations is not clearly specified. However, in dealing with economic agents rather than controlled machines, decision theoretic criteria naturally come into play. In this sphere, robust control is closely related to the notions of Knightian uncertainty, ambiguity and uncertainty aversion, which are all roughly equivalent (although sometimes differing in formalization).

Starting with the observations of the classic Ellsberg (1961) paradox – that (some) decision makers prefer environments with known odds to those with uncertain probabilities – there has been a broad literature in decision theory which has weakened the Savage axioms to incorporate preferences which display such aversion to uncertainty or ambiguity. The most widely used characterization is due to Gilboa and Schmeidler (1989), who axiomatized ambiguity preferences with multiple priors. Decision-making with multiple priors can be represented as max-min expected utility: maximizing the utility with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
respect to the least favourable prior from a convex set of priors. More recently, Epstein and Schneider (2003) have extended the static environment of Gilboa and Schmeidler to a dynamic context, where the set of priors is updated over time. Hansen et al. (2006) formally established the links between robust control and ambiguity aversion, showing that the model set of robust control as discussed above can be thought of as a particular specification of Gilboa and Schmeidler's set of priors. Moreover, although the ambiguity preferences are characterized by posing particular counterfactuals which require multiple priors, once the least favourable prior is chosen, behaviour could be rationalized as Bayesian with that prior. Thus from a Bayesian viewpoint Sims (2001) views robust control as a means of generating priors, which then naturally leads to questioning whether the worst case prior accurately reflects actual beliefs and preferences. (See also Svensson, 2001. Hansen et al., 2006, show how to back out the Bayesian prior which rationalizes robust decision-making.)

Finally, in many cases robust or ambiguity-averse preferences are similar to enhanced risk aversion, and in some cases they are observationally equivalent. This insight dates to Jacobson (1973) and Whittle (1981) in the control theory literature, and the relations between robust control and a particular specification of Kreps and Porteus (1978), Epstein and Zin (1989) and Duffie and Epstein (1992) recursive utility with enhanced risk aversion have been shown by Anderson, Hansen and Sargent (2003), Hansen et al. (2006) and Skiadas (2003).

Control theory background

Since many of the ideas and inspiration for robust control in economics come from control theory, we give here just a broad outline of its development. More detail and different perspectives can be found in the books by Zhou, Doyle and Glover (1996), Ba•ar and Bernhard (1995), and Burl (1999). Throughout the late 1960s and early 1970s optimal control came into its own, largely through the work of Kalman on linear quadratic (LQ) control and filtering. While this approach remains widely used today throughout economics, starting in the late 1970s and early 1980s the control theory literature started to change as theory and practice showed some of the limitations of the LQ approach. Although LQ control with full observation (the so-called linear quadratic regulator or LQR) was known to be robust to some types of model perturbations, Doyle (1978) showed that there are no such assurances in the case of partial observation (the so-called linear-quadratic-Gaussian or LQG case, which is an LQR control matched with a Kalman filter). Doyle's paper title and abstract are classic in the literature – title: ‘Guaranteed Margins for LQG Regulators’, abstract: ‘There are none’.

Spurred by this and related work, control theorists started to move away from LQ control to look for a more robust approach. Zames (1981) was influential in the development of H_∞ control as a more robust alternative to LQ control. Loosely speaking, in LQ control the quadratic cost means that performance is measured with a 2-norm across frequencies. By contrast, H_∞ uses an ∞ – norm that looks at the peak of the losses across frequencies. It is also interpretable as the maximal magnification of the disturbances to outputs of interest. While the early robust control literature used a frequency domain approach, in the late 1980s Doyle and others developed state space formulations (see Doyle et al., 1989, for example) which gave explicit solutions and allowed for alterative formalizations. For example, the H_∞ approach was given alternative justifications in terms of penalizing disturbances from the nominal model, which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

can be implemented as a dynamic game between a decision maker seeking to minimize losses and a malevolent agent seeking to maximize loss. (See Ba•er and Bernhard, 1995, for a development of this approach.) Finally, the uncertainty sets in the H_∞ approach are unstructured – they represent perturbations of the model which are bounded but have no particular form. The implications of structured perturbations have been studied more recently. Some examples include parametric perturbations, unmodelled dynamics, or uncertainty only about particular channels or connections in a model. Applications with structured uncertainty use the structured singular value (also known as μ) rather than the H_∞ norm as a measure of performance. Although there are some important stability and performance criteria, in general constructing control rules is a more daunting task, and the theory is not as fully developed as the unstructured case.

The Hansen – Sargent approach

In the economics literature, the most prominent and influential approach to robust control is due to Hansen and Sargent (and their co-authors), which is summarized in their monograph Hansen and Sargent (2008). This approach starts with a nominal model and uses entropy as a distance measure to calibrate the model uncertainty set. More specifically, the model set consists of those models whose relative entropy or Kullback–Leibler distance from the nominal model is bounded by a specified value. Note that this puts no structure on the uncertainty, but only restricts the alternative models to those which are difficult to distinguish statistically from the nominal model. In practice, a Lagrange multiplier theorem is typically used to convert the entropy constraint into a penalty on perturbations from the model. Then the solution of the control problem is found via a dynamic game implementation: the agent maximizes utility by his choice of control, while an evil agent minimizes utility by his choice of perturbation, while being penalized by the entropy of the deviations. Relative to the control theory literature such as Ba•ar and Bernhard (1995), the main differences are that all models are stochastic, while control theory largely uses deterministic models. One exception is Petersen, James, and Dupuis (2000) who use a similar approach to consider uncertain stochastic systems. In addition, discounting is not typically considered in control theory, while it is natural in economics. In full information problems discounting has relatively little effect, but it raises important issues in problems with partial information (see Hansen and Sargent, 2005a; 2005b). Finally, the Hansen–Sargent approach naturally extends beyond the LQ setting laid out in Hansen, Sargent and Tallarini (1999), with some examples in Anderson, Hansen and Sargent (2003), Cagetti et al. (2002) and Maenhout (2004).

To be more concrete, consider an LQ example where x_t is the state, i_t is the agent's control, and ε_t is an i.i.d. Gaussian shock. The nominal model is:

$$x_{t+1} = Ax_t + Bi_t + C\varepsilon_{t+1}, \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and the agent's intertemporal preferences are:

$$E_0 \sum_{t=0}^{\infty} \beta^t \left(x_t' Q x_t + i_t' R i_t \right)$$
(2)

where 0< β <1 and *Q* and *R* are negative definite matrices. The approach of Hansen and Sargent perturbs the nominal model with an additional 'misspecfication shock' w_{t+1} which is allowed to be correlated with the state x_t:

$$x_{t+1} = Ax_t + Bi_t + C(\varepsilon_{t+1} + w_{t+1}).$$
(3)

The shock w_{t+1} is used to represent alternative models. These models are made to be close to the nominal model in an entropy sense by imposing the bound:

$$E_0 \sum_{t=0}^{\infty} \beta^t w_{t+1}' w_{t+1} \leq \eta$$
(4)

for some constant η ≥ 0. The agent then maximizes (2) with respect to the worst case perturbed model (3) from the set (4). Using a Lagrange multiplier theorem, the constraint set can be converted to a penalty and the decision problem can by solved recursively by solving the Bellman equation for a two-player zero sum game:

$$V(x) = \max_{i} \min_{w} \left\{ \left\{ x' Qx + i' Ri + \beta \theta w' w + \beta E[V(Ax + Bi + C(\varepsilon + w)) | x] \right\} \right.$$
(5)

where θ >0 is a Lagrange multiplier on the constraint (4) and the expectation is over the Gaussian shock

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
ε . Often this multiplier formulation is taken as the starting point, for example Maccheroni, Marinacci and Rustichini (2006) characterize preferences of this form, with θ governing the degree of robustness. As $\theta \to \infty$ the penalization becomes so great that only the nominal model remains (thus $\eta \to 0$), and the decision rule is less robust. Conversely, there is typically a minimal value of θ beyond which the value is $V(x)=-\infty$. This gives the most robust decision rules, allowing for the largest uncertainty set.

Adding structure to the uncertainty set

The approach discussed above uses unstructured uncertainty, and has been well developed and extended in different dimensions. We now discuss some alterative approaches which put more structure on the uncertainty set. There are many reasons to do so. It may be that some of the models that are close to the nominal model in a statistical sense may not be plausible economically. Alternatively, the decision makers may have a discrete set of models in mind, and bounding them all in one uncertainty set may include extraneous implausible models. Perhaps most substantively, the decision maker may be more confident some aspects of the model relative to others. Some examples of this include knowing the model up to the values of parameters, or being more certain about the dynamics of certain variables in the model. Not taking into account the particular structure may give a misleading impression of the actual uncertainty the decision makers face.

There are many ways of building in structured uncertainty, and the distinctions between cases are not always clear. For example, consider the same nominal model (1) as above, but suppose that instead of the unstructured perturbations (3) the uncertainty is instead solely in the values of the parameters *A* and *B*. Thus we can represent the parametric perturbed models as:

$$x_{t+1} = (A + \hat{A})x_t + (B + \hat{B})i_t + C\varepsilon_{t+1} \tag{6}$$

for some matrices $\hat{A}$ and $\hat{B}$. Of course it's possible to rewrite (6) as a version of (3) with:

$$w_{t+1} = \hat{A}x_t + \hat{B}i_t, \tag{7}$$

so in principle parametric perturbations are just a special case of the unstructured uncertainty. However what makes a substantive difference is how uncertainty is measured, that is whether we restrict w_{t+1} as in (4) or whether we restrict the parameters $\hat{A}$ and $\hat{B}$, say by bounding them in a confidence ellipsoid around the nominal model. Moreover, as (7) makes clear the differences between the uncertainty

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
measurements will depend on the actual control rule i_t in place. Onatski and Williams (2003) provide an example of a simple estimated model where the uncertainty specifications matter dramatically for outcomes. In particular, the optimal policy for the largest possible unstructured uncertainty set (that is for the minimal value of θ) leads to instability for relatively small parametric perturbations. Thus the particular structure and measurement of uncertainty can have important implications for decisions. (Peterson, James and Dupuis, 2000, modify the unstructured approach described above to deal with structured uncertainty by separating the entropy penalty for unstructured perturbations from a different penalization for structured perturbations.)
Some economic applications with structured uncertainty include the following:

1. 1. The simplest cases are uncertainty sets with discrete possible models. Some examples include: Levin and Williams (2003), who consider both Bayesian and minimax approaches; Cogley and Sargent (2005); and Svensson and Williams (2006) and who focus on a Bayesian approach, and the recent work of Hansen and Sargent (2006), who have built this type of structure into their robust approach.
2. 2. Another common form is parameter uncertainty within a fully specified model. Brainard (1967) is the classic reference from a Bayesian perspective with many references in this line, while Giannoni (2002) and Chamberlain (2000) consider minimax approaches.
3. 3. Somewhat more broad are cases with different parametric model specifications. For example this includes uncertainty about dynamics (lags and leads), variables which may enter, uncertainty about data quality, and other features which are built into parametric extensions of the nominal model. Examples include the model error modelling approach of Onatski and Williams (2003) and the empirical specifications of Brock, Durlauf and West (2003).
4. 4. Finally, the model sets may be nonparametric but structured in particular ways. For example, Onatski and Stock (2002) consider different structured types of uncertainty such as linear time-invariant perturbations, nonlinear time-varying perturbations, and perturbations which only enter particular parts of the model. Other examples include nonparametric specifications of uncertainty which differs across frequencies as in Onatski and Williams (2003) and Brock and Durlauf (2005).

See Also

- ambiguity and ambiguity aversion
- model uncertainty
- stochastic optimal control
- uncertainty

Bibliography

Anderson, E., Hansen, L.P. and Sargent, T. 2003. A quartet of semi-groups for model specification,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
robustness, prices of risk, and model detection. *Journal of the European Economic Association* 1, 68–123.

Backus, D.K., Routledge, B.R. and Zin, S.E. 2005. Exotic preferences for macroeconomists. In *NBER Macroeconomics Annual 2004*, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Ba•ar, T. and Bernhard, P. 1995. *H_∞-Optimal Control and Related Minimax Design Problems*. Boston: Birkhauser.

Brainard, W. 1967. Uncertainty and the effectiveness of policy. *American Economic Review* 57, 411–25.

Brock, W.A. and Durlauf, S.N. 2005. Local robustness analysis: theory and application. *Journal of Economic Dynamics and Control* 29, 2067–92.

Brock, W.A., Durlauf, S.N. and West, K.D. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 2003(1), 235–301.

Burl, J.B. 1999. *Linear Optimal Control: $\mathcal{H}_2$ and $\mathcal{H}_\infty$ Methods*. Menlo Park, CA: Addison-Wesley.

Cagetti, M., Hansen, L.P., Sargent, T.J. and Williams, N. 2002. Robustness and pricing with uncertain growth. *Review of Financial Studies* 15, 363–404.

Chamberlain, G. 2000. Econometric applications of maxmin expected utility theory. *Journal of Applied Econometrics* 15, 625–44.

Cogley, T. and Sargent, T.J. 2005. The conquest of U.S. inflation: learning and robustness to model uncertainty. *Review of Economic Dynamics* 8, 528–63.

Doyle, J.C. 1978. Guaranteed margins for LQG regulators. *IEEE Transactions on Automatic Control* 23, 756–7.

Doyle, J.C., Glover, K., Khargonekar, P.P. and Francis, B.A. 1989. State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems. *IEEE Transactions on Automatic Control* 34, 831–47.

Duffie, D. and Epstein, L.G. 1992. Stochastic differential utility. *Econometrica* 60, 353–94.

Ellsberg, D. 1961. Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Epstein, L. and Schneider, M. 2003. Recursive multiple-priors. *Journal of Economic Theory* 113, 1–31.

Epstein, L. and Zin, S. 1989. Substitution, risk aversion and the temporal behavior of consumption and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
asset returns: a theoretical framework. *Econometrica* 57, 937–69.

Giannoni, M. 2002. Does model uncertainty justify caution? Robust optimal monetary policy in a forward-looking model. *Macroeconomic Dynamics* 6, 111–44.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

Hansen, L.P. and Sargent, T.J. 2001. Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics* 4, 519–35.

Hansen, L.P. and Sargent, T.J. 2005a. Robust estimation and control under commitment. *Journal of Economic Theory* 124, 258–301.

Hansen, L.P. and Sargent, T.J. 2005b. Robust estimation and control without commitment. Working paper, Department of Economics, University of Chicago.

Hansen, L.P. and Sargent, T.J. 2006. Fragile beliefs and the price of model uncertainty. Working paper, Department of Economics, University of Chicago.

Hansen, L.P. and Sargent, T.J. 2008. *Robustness*. Princeton, NJ: Princeton University Press.

Hansen, L.P., Sargent, T. and Tallarini, T. 1999. Robust permanent income and pricing. *Review of Economic Studies* 66, 873–907.

Hansen, L.P., Sargent, T.J., Turmuhambetova, G.A. and Williams, N. 2006. Robust control and model misspecification. *Journal of Economic Theory* 128, 45–90.

Jacobson, D.H. 1973. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control AC* 18, 1124–31.

Kreps, D.M. and Porteus, E.L. 1978. Temporal resolution of uncertainty and dynamic choice. *Econometrica* 46, 185–200.

Levin, A.T. and Williams, J. 2003. Robust monetary policy with competing reference models. *Journal of Monetary Economics* 50, 945–75.

Limebeer, D.J.N., Anderson, B.D.O. and Hendel, B. 1994. A Nash game approach to mixed H_2/H_∞ control. *IEEE Transactions on Automatic Control* 39, 69–82.

Maccheroni, F., Marinacci, M. and Rustichini, A. 2006. Dynamic variational preferences. *Journal of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Economic Theory 128, 4–44.

Maenhout, P.J. 2004. Robust portfolio rules and asset pricing. *Review of Financial Studies* 17, 951–83.

Onatski, A. and Stock, J.H. 2002. Robust monetary policy under model uncertainty in a small model of the US economy. *Macroeconomic Dynamics* 6, 85–110.

Onatski, A. and Williams, N. 2003. Modeling model uncertainty. *Journal of the European Economic Association* 1, 1087–122.

Petersen, I.R., James, M.R. and Dupuis, P. 2000. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control* 45, 398–412.

Sims, C.A. 2001. Pitfalls of a minimax approach to model uncertainty. *American Economic Review* 91, 51–4.

Skiadas, C. 2003. Robust control and recursive utility. *Finance and Stochastics* 7, 475–89.

Svensson, L.E. 2001. Robust control made simple. Working paper, Department of Economics, Princeton University.

Svensson, L.E.O. and Williams, N. 2006. Monetary policy with model uncertainty: distribution forecast targeting. Working paper, Department of Economics, Princeton University.

Tornell, A. 2003. Exchange rate anomalies under model misspecification: a mixed optimal/robust approach. Working paper, Department of Economics, UCLA.

Whittle, P. 1981. Risk sensitive linear quadratic Gaussian control. *Advances in Applied Probability* 13, 764–77.

Zames, G. 1981. Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control* 26, 301–20.

Zhou, K., Doyle, J. and Glover, K. 1996. *Robust and Optimal Control*. Upper Saddle River, New Jersey: Prentice Hall.

How to cite this article

Williams, Noah. "robust control." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
article?id=pde2008_R000250> doi:10.1057/9780230226203.1451

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

robust estimators in econometrics

P. •ížek and W. Härdle
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Econometric data are often obtained under conditions that cannot be well controlled, and so partial departures from the model assumptions in use (data contamination) occur relatively frequently. To address this, we first introduce concepts of robust statistics for qualifying and quantifying sensitivity of estimation methods to data contamination as well as important approaches to robust estimation. Later, we discuss how robust estimation methods have been adapted to various areas of econometrics, including time series analysis and general GMM-based estimation.

Keywords

breakdown point; data contamination; generalized method of moments; heteroscedasticity; influence function; instrumental variables; least squares; linear regression; maximum likelihood; maximum-bias curve; measurement errors; M-estimation; Newton–Raphson procedure; nonlinear models; probability distribution; qualitative and quantitative robustness; quantile regression; random variables; robust econometrics; robust statistics; S-estimation; simultaneous equations models; time-series analysis

Article

Econometrics often deals with data under, from the statistical point of view, non-standard conditions such as heteroscedasticity or measurement errors, and the estimation methods thus need either to be adapted to such conditions or to be at least insensitive to them. Methods insensitive to violation of certain assumptions – for example, insensitive to the presence of heteroscedasticity – are in a broad sense referred to as ‘robust’ (for example, robust to heteroscedasticity). On the other hand, there is also a more specific meaning of the word ‘robust’, which stems from the field of robust statistics. This latter notion defines robustness rigorously in terms of the behaviour of an estimator both at the assumed (parametric) model and in its neighbourhood in the space of probability distributions. Even though the methods of robust statistics have been used only in the simplest settings, such as estimation of location, scale or linear regression for a long time, they have motivated a range of new econometric methods, which we focus on in this article.

The concepts and measures of robustness are introduced first (Section 1), followed by the most common types of estimation methods and their properties (Section 2). Various econometric methods based on these common estimators are discussed in Section 3, covering tasks from time-series regression over

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

GMM estimation to simulation-based methods.

1 Measures of robustness

Robustness properties can be formulated within two frameworks: qualitative and quantitative robustness. *Qualitative robustness* is concerned with the situation in which the shape of the underlying (true) data distribution deviates slightly from the assumed model. It focuses on questions like stability and performance loss over a family of such slightly deviating distributions. *Quantitative robustness* is involved when the sensitivity of estimators to a proportion of aberrant observations is studied.

A simple example can make this clear. Suppose one has collected a sample on an individual's income (after say ten years of schooling) and one is interested in estimating the mean income. If $\{x_i\}_{i=1}^n$ denotes the logarithm of this data and we suppose that they have a cumulative distribution function (cdf) F, assumed to be $N(\mu, \sigma^2)$, the maximum likelihood estimator (MLE) is $\bar{x} = \int u dF_n(u) = T(F_n)$, where $F_n(u) = n^{-1}\sum_{i=1}^n I(x_i \le u)$, and $\mu = \int u dF(u) = T(F)$. Qualitative robustness asks here the question: how well will μ be estimated if the true distribution is in some neighbourhood of F•? Quantitative robustness would concentrate on the question: will $T(F_n)$ be bounded if some observations $x_i \to \infty$? In fact, the latter question is easy to answer: if $x_i \to \infty$ for some i, $T(F_n) = \bar{x} \to \infty$ as well. So we can say here in a loose sense that $\bar{x}$ is not quantitatively robust.

1.1 Formalities

In the following we present a mathematical set-up that allows us to formalize our thoughts on robustness.

The notion of the sensitivity of an estimator T is put into theory by considering a model characterized by a cdf F and its neighbourhood $\mathcal{F}_{\varepsilon, G}$: distributions $(1-\varepsilon)F + \varepsilon G$, where $\varepsilon \in (0,1/2)$ and G is an arbitrary probability distribution, which represents data contamination. Hence, not all data necessarily follow the pre-specified distribution, but the ε-part of data can come from a different distribution G. If $H \in \mathcal{F}_{\varepsilon, G}$, the estimation method T is then judged by how sensitive or robust the estimates $T(H)$ are to the size of $\mathcal{F}_{\varepsilon, G}$, or alternatively, to the distance from the assumed cdf F. Two main concepts for robust measures analyse the sensitivity of an estimator to infinitesimal deviations, $\varepsilon \to 0$, and to finite (large) deviations, $\varepsilon > 0$, respectively. Despite generality of the concept, easy interpretation and technical difficulties often limit our choice to point-mass distributions (Dirac measures) $G = \delta_x$, $x \in \mathbb{R}$, which simply represents an (erroneous) observation at point $x \in \mathbb{R}$. This simplification is also used in the following text.

1.2 Qualitative robustness

The influence of infinitesimal contamination on an estimator is characterized by the *influence function*, which measures the relative change in estimates caused by an infinitesimally small amount ε of contamination at x (Hampel et al., 1986). More formally,

$$IF(x; T, F) = \lim_{\varepsilon \to 0} \frac{T\{(1-\varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon}. \quad (1)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

For each point x, the influence function reveals the rate at which the estimator T changes if a wrong observation appears at x. In the case of sample mean $\bar{x} = T(F_n)$ for $\{x_i\}_{i=1}^{n}$, we obtain

$$IF(x;\, T,\, F_n) = \lim_{\varepsilon \to 0}\left[(1-\varepsilon)\int u\,dF_n(u) + \varepsilon\int u\,d\delta_x(u) - \int u\,dF_n(u)\right] / \varepsilon = \lim_{\varepsilon \to 0}\left[-\int u\,dF_n(u) + \int u\,d\delta_x(u)\right] = x - \bar{x}.$$

The influence function allows us to define various desirable properties of an estimation method. First, the largest influence of contamination on estimates can be formalized by the *gross-error sensitivity*,

$$\gamma(T,\, F) = \sup_{x \in \mathbb{R}} IF(x;\, T,\, F), \tag{2}$$

which under robustness considerations is finite and small. Even though such a measure can depend on F in general, the qualitative results (for example, $\gamma(T,F)$ being bounded) are typically independent of F. Second, the sensitivity to small changes in data, for example moving an observation from x to $y \in \mathbb{R}$, can be measured by the *local-shift sensitivity*

$$\lambda(T,\, F) = \sup_{x \neq y} \frac{\|IF(x;\, T,\, F) - IF(y,\, T,\, F)\|}{\|x - y\|}. \tag{3}$$

Also, this quantity should be relatively small since we generally do not expect that small changes in data cause extreme changes in values or sensitivity of estimates. Third, as unlikely large or distant observations may represent data errors, their influence on estimates should become zero. Such a property is characterized by the *rejection point*,

$$\rho(T,\, F) = \inf_{r > 0}\left\{r : IF(x;\, T,\, F) = 0,\ \|x\| \geq r\right\}, \tag{4}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which indicates the non-influence of large observations.

1.3 Quantitative robustness

Alternatively, the behaviour of the estimator *T* can be studied for any finite amount ε of contamination. The most common property looked at in this context is the estimator's bias $b(T, H) = E_H\{T(H)\} - E_F\{T(F)\}$, which measures a distance between the estimates for clean data, *T*(*F*), and contaminated data, $T(H), H \in \mathcal{F}_{\varepsilon, G}$. The corresponding *maximum-bias curve* measures the maximum bias of *T* on $\mathcal{F}_{\varepsilon, G}$ at any ε :

$$B(\varepsilon, T) = \sup_{x \in \mathbb{R}} b\left\{T, (1-\varepsilon)F + \varepsilon\delta_x\right\}. \quad (5)$$

Although the computation of this curve is rather complex, Berrendero and Zamar (2001) provide general methodology for its computation in the context of linear regression.
The maximum-bias curve is not only useful on its own, but allows us to define further scalar measures of robustness. The most prominent is the *breakdown point* (Hampel, 1971), which is defined as the smallest amount ε of contamination that can cause an infinite bias:

$$\varepsilon^*(T) = \inf_{\varepsilon \geq 0} \left\{\varepsilon : B(\varepsilon, T) = \infty\right\}. \quad (6)$$

The intuitive aim of this definition specifies the breakdown point $\varepsilon^*(T)$ as the smallest amount of contamination that makes the estimator *T* useless. Note that in most cases $\varepsilon^*(T) \leq 0.5$ (He and Simpson, 1993). This definition and the upper bound, however, apply only in simple cases, such as location or linear regression estimation (Davies and Gather, 2005). The most general definition of breakdown point formalizes the idea of ‘useless’ estimates in the following way: an estimator is said to break down if, under contamination, it is not random anymore, or, more precisely, it can achieve only a finite set of values (Genton and Lucas, 2003). This definition is based on the fact that estimates are functions of observed random samples and are thus random quantities themselves unless they fail. Although the latter definition includes the first one, the latter one may generally depend on the underlying model *F*, for example in time-series context.

2 Estimation approaches

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Denote by F_n an empirical distribution function (edf) corresponding to a sample $\{x_i\}_{i=1}^n \in \mathbb{R}$ drawn from a model based on probability distribution F. Most estimation methods can be defined as an extremum problem, minimizing a contrast $\int h(z, \theta) dF(z)$ over θ in a parameter space, or as a solution of an equation, $\int g(z, \theta) dF(z) = 0$ in θ. The estimation for a given sample utilizes finite-sample equivalents of these integrals, $\int h(z, \theta) dF_n(z)$ and $\int g(z, \theta) dF_n(z)$, respectively.

Consider the pure location model $X_i = \mu + \sigma\varepsilon_i$, $i = 1, \ldots, n$, with a known scale σ and $\varepsilon \sim F$. The cdf of X is then $F\{(x - \mu) / \sigma\}$. With a quadratic contrast function $h(x, \theta) = (x - \theta)^2$, the estimation problem is to minimize $\int (x - \theta)^2 dF\{(x - \mu) / \sigma\}$ with respect to θ. For known F, this leads to $\theta = \mu$ and one sees that, without loss of generality, one can assume $\mu = 0$ and $\sigma = 1$. For the sample $\{x_i\}_{i=1}^n$ characterized by edf F_n, the location parameter μ is estimated by

$$\hat{\mu} = \arg\min_{\theta} \int (x - \theta)^2 dF_n(x) = n^{-1} \sum_{i=1}^{n} x_i = \bar{x}.$$

Note that for $g(x, \theta) = x - \theta$, the parameter μ is the solution to $\int g(x, \theta) dF(x) = 0$. The estimator may therefore be alternatively defined through $\mu = T(F) = \int u dF(u)$.

As indicated in the introduction, this standard estimator of location performs unfortunately rather poorly under the sketched contamination model. Estimating a population mean by the least squares (LS) or sample mean $\bar{x} = T(F_n)$ has the following properties. First, the influence function (1)

$$IF(x; T, F) = \lim_{\varepsilon \to 0} \frac{T\{(1 - \varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{\{(1 - \varepsilon)\int u dF(u) + \varepsilon x\} - \int u dF(u)}{\varepsilon} = \lim_{\varepsilon \to 0} \varepsilon^{-1}\left\{-\varepsilon \int u dF(u) + \varepsilon x\right\} = x - \int u dF(u) = x - T(F).$$

Hence, the gross-error sensitivity (2) $\gamma(T, F) = \infty$, the local-shift sensitivity (3) $\lambda(T, F) = 0$, and the rejection point (4) $\rho(T, F) = \infty$. Second, the maximum-bias (5) is infinite for any $\varepsilon > 0$ since

$$\sup_{x \in \mathbb{R}} \left\| T\left\{(1 - \varepsilon)F + \varepsilon\delta_x\right\} - T(F) \right\| = \sup_{x \in \mathbb{R}} \left\| -\varepsilon T(F) + \varepsilon x \right\| = \infty.$$

Consequently, the breakdown point (6) of the sample mean $\bar{x} = T(F_n)$ is zero, $\varepsilon^*(T) = 0$.

Thus, none of robustness measures characterizing the change of T under contamination of data (even infinitesimally small) is finite. This behaviour,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

typical for LS-based methods, motivated alternative estimators that have the desirable robust properties. In this section, the *M*-estimators, *S*-estimators and τ-estimators are discussed as well as some extensions and combination of these approaches. Even though there is a much wider range of robust estimation principles, we focus on those already studied and adopted in various areas of econometrics.

2.1 M-estimators

To achieve more flexibility in accommodating requirements on robustness, Huber (1964) proposed the *M-estimator* by considering a general extremum estimator based on $\int \rho(z, \theta) dF(z)$, thus minimizing $\int \rho(z, \theta) dF_n(z)$ in finite samples. Providing that the first derivative $\psi(z, \theta) = \partial \rho(z, \theta) / \partial \theta$ exists, an *M*-estimator can be also defined by an implicit equation $\int \psi(z, \theta) dF_n(z) = 0$.

This extremely general definition is usually adapted to a specific estimation problem such as location, scale or regression estimation. In a univariate location model, $F(z)$ can be parameterized as $F(z - \theta)$ and hence one limits $\rho(z, \theta)$ and $\psi(z, \theta)$ to $\rho(z - \theta)$ and $\psi(z - \theta)$. In the case of scale estimation, $F(z) = F(z / \theta)$ and consequently $\rho(z, \theta) = \rho(z / \theta)$ and $\psi(z, \theta) = \psi(z / \theta)$. In linear regression, $z = (x, y)$ and a zero-mean error term $\varepsilon = y - x^T \theta$. Analogously to the location case, one can then consider $\rho(z, \theta) = \rho(y - x^T \theta)$ and $\psi(z, \theta) = \psi(y - x^T \theta) x$, or more generally, $\rho(z, \theta) = \rho(y - x^T \theta, x)$ and $\psi(z, \theta) = \psi(y - x^T \theta, x)$ (*GM*-estimators). Generally, we can express $\rho(z, \theta)$ as $\rho\{\eta(z, \theta)\}$, $\psi\{\eta(z, \theta)\}$, where $\eta(z, \theta) \sim F$.

Some well-known choices of univariate objective functions ρ and ψ are given in Table 1; functions $\rho(t)$ are usually assumed to be non-constant, non-negative, even and continuously increasing in $|t|$. This documents flexibility of the concept of *M*-estimators, which include LS and quantile regression as special cases.

Examples of ρ and ψ functions used with *M*-estimators

	$\rho(t)$	$\psi(t)$
Least squares	t_2	$2t$
Least absolute deviation	$\|t\|$	Sign (t)
Quantile estimation	$\{\tau - I(x < 0)\} x$	$\tau - I(x < 0)$
Huber: for $\|t\| \le c$	t^2	$2t$
••• for $c < \|t\|$	$c\|t\|$	c sign(t)
Hample: for $\|t\| \le c$	t^2	$2t$
••• for $a < \|t\| \le b$	$a\|t\|$	a sign(t)
••• for $b < \|t\| \le c$	$\frac{ac}{c-b} t - \frac{a}{c-b} t^2 \text{sign}(t)$	$a(c - \|t\|)/(c - b)$
••• for $c < \|t\|$	$a\|t\|$	0
Biweight (Tukey)	$-(c^2 - t^2)^3 I(\|t\| \le c) / 6$	$t(c^2 - t^2)^2 I(\|t\| \le c) / 6$
Sine (Andrews)	$-c \cos(x / c) I(\|t\| \le \pi c)$	$\sin(x / c) I(\|t\| \le \pi c)$

On the other hand, many of the ρ and ψ functions in Table 1 depend on one or more constants $a, b, c \in \mathbb{R}$. If an estimator T is to be invariant to the scale of data, one can apply the estimator to rescaled data, that is, to minimize $\int \rho\{(z - \theta) / s\} dF_n(z)$ or to solve $\int \psi\{(z - \theta) / s\} dF_n(z) = 0$ for a scale estimate s like the median absolute deviation (MAD). Alternatively, one may also estimate parameters θ and scale s simultaneously by considering

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$\rho(z, \{\theta, s\}) = \rho((z-\theta)/s)$ or

$$\psi(z, \{\theta, s\}) = \{\psi_1(z, \{\theta, s\}), \psi_s(z, \{\theta, s\})\}.$$

Let us now turn to the question how the choice of functions ρ and ψ determines the robust properties of *M*-estimators. First, the influence function of an *M*-estimator can generally depend on several quantities such as its asymptotic variance or the position of explanatory variables in the regression case, but the influence function is always proportional to function $\psi(z,b)$. Thus, the finite gross-error sensitivity, $\gamma(T, F) < \infty$, requires bounded $\psi(t)$ (which is not the case with LS). Similarly, the finite rejection point, $\rho(T, F) < \infty$, leads to $\psi(t)$ being zero for all sufficiently large *t* (the *M*-estimators defined by such a ψ-function are called redescending). Hampel et al. (1986) shows how, for a given bound on $\gamma(T,F)$, one can determine the most efficient choice of ψ function (for example, the skipped median, $\psi(t) = sign(t)I(|t| < K), K > 0$, in the location case).
More formally, the optimality of *M*-estimators in the context of qualitative robustness can be studied by the *asymptotic relative efficiency* (ARE) of an estimator $\hat{\theta}^1$ relative to another estimator $\hat{\theta}^2$:

$$ARE(\hat{\theta}^1, \hat{\theta}^2) = \frac{\text{as. var}(\hat{\theta}^1)}{\text{as. var}(\hat{\theta}^2)}. \quad (7)$$

For example, at the normal distribution with $\hat{\theta}^1$ and $\hat{\theta}^2$ being the least absolute deviation (LAD) and LS estimators, ARE equals $2/\pi \approx 0.64$. Under the Student cdf t_5, the ARE of the two estimators climbs up to ≈ 0.96. For Huber's M-estimator, we see that its limit cases are the median for $c \to 0$ and the mean for $c \to \infty$. At the normal distribution and for $c = 1.345$, we have ARE of about 0.95. This means that this *M*-estimator is almost as efficient as MLE, but does not lose so drastically in performance as the standard mean under contamination because of the bounded influence function.
Whereas the influence function of M-estimators is closely related to the choice of its objective function, the global robustness of M-estimators is in a certain sense independent of this choice. Maronna, Bustos and Yohai (1979) showed in linear regression that the breakdown point of M-estimators is bounded by 1/p, where p is the number of estimated parameters. As a remedy, several authors proposed one-step M-estimators that are defined, for example, as the first step of the iterative Newton–Raphson procedure, used to minimize $\int \rho(z, \theta)\, dF(z)$, starting from initial robust estimators $\hat{\theta}^0$ of parameters and $\hat{s}^0$ of scale (see Welsh and Ronchetti, 2002, for an overview). Possible initial estimators can be those discussed in subsections 2.2 and 3.3. For example for an M-estimator of location $\hat{\theta}$ defined by a function $\psi(x, \theta) = \psi(x - \theta)$, its one-step counterpart can be defined at sample $\{x_i\}_{i=1}^n$ by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\hat{\theta} = \hat{\theta}^0 + \hat{s}^0 \sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}^0}{\hat{s}^0}\right) / \sum_{i=1}^{n} \psi'\left(\frac{x_i - \hat{\theta}^0}{\hat{s}^0}\right),$$

where $\hat{\theta}^0$ and $\hat{s}^0$ represent initial robust estimators of location and scale like the median and MAD, respectively. Such one-step estimators, under certain conditions on the initial estimators, preserve the breakdown point of the initial estimators, and at the same time have the same first-order asymptotic distribution as the original M-estimator (Simpson, Ruppert and Carroll, 1992; Welsh and Ronchetti, 2002). Further development of such ideas includes an adaptive choice of parameters of function ψ in the iterative step (Gervini and Yohai, 2002).

2.2 S-estimators

An alternative approach to *M*-estimators achieving high breakdown point (HBP) was proposed by Rousseeuw and Yohai (1984). The S-estimators are defined by minimization of a scale statistics $s^2(z, b) = s\{\eta(z, b)\}$ defined as the *M*-estimate of scale,

$$\int \rho[\eta(z, b) / s\{\eta(z, b)\}]\, dF_n(z) = K = \int \rho(t)\, dF(t),$$

at the model distribution *F*; the functions ρ and η are those defining *M*-estimators in subsection 2.1. More generally, one can define *S*-estimators by means of any scale-equivariant statistics s^2, that is, $s\{c\eta(z, b)\} = |c| s\{\eta(z, b)\}$. Under this more general definition, *S*-estimators include as special cases LS and LAD estimators. Further, they encompass several well-known robust methods including least median of squares (LMS) and least trimmed squares (LTS): whereas the first defines the scale statistics $s^2\{\eta(z,b)\}$ as the median of squared residuals $\eta(z,b)$, the latter uses the scale defined by the sum of *h* smallest residuals $\eta(z,b)$. In order to appreciate the difference to *M*-estimators, it is worth pausing for a moment and to present LMS, the most prominent representative of *S*-estimators, in the location case:

$$\arg\min_{\theta} \operatorname{med}\left\{(x_1 - \theta)^2, \ldots, (x_n - \theta)^2\right\}.$$

Due to its definition, the *S*-estimators have the same influence function as the *M*-estimator constructed from the same function ρ. Contrary to *M*-estimators, they can achieve the highest possible breakdown point $\varepsilon^* = 0.5$. For example, this is the case of LMS and LTS. For Gaussian data, the most efficient (in the sense of ARE (7)) among the *S*-estimators with $\varepsilon^* = 0.5$ is, however, the one corresponding to $K = 1.548$ and ρ being the Tukey

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

biweight function (see Table 1). Given the HBP of *S*-estimators, their maximum-bias behaviour is of interest too. Although it depends on the function ρ and constant *K* (Berrendero and Zamar, 2001), Yohai and Zamar (1993) proved that LMS minimizes maximum bias among a large class of (residual admissible) estimators, which includes most robust methods.

An important shortcoming of HBP *S*-estimation is, however, its low ARE: under Gaussian data, efficiency relative to LS varies from zero per cent to 27 per cent. Thus, *S*-estimators are often used as initial estimators for other, more efficient methods. Nevertheless, if an *S*-estimator is not applied directly to sample observations, but rather to the set of all pairwise differences of sample observations, the resulting generalized *S*-estimator exhibits higher relative efficiency for Gaussian data, while preserving its robust properties (Croux, Rousseeuw and Hossjer, 1994; Stromberg, Hossjer and Hawkins, 2000).

2.3 τ -estimators

The *S*-estimators improve upon *M*-estimators in terms of their breakdown-point properties, but at the cost of low Gaussian efficiency. Although one-step *M*-estimators based on an initial *S*-estimate can remedy this deficiency to a large extent, their exact breakdown properties are not known. One alternative approach, proposed by Yohai and Zamar (1988), extends the principle of *S*-estimation in the following way. Assuming that ρ_1 and ρ_2 are non-negative, even, and continuous functions, the *M*-estimate $s^2(z, \theta) = s^2\{\eta(z, \theta)\}$ of scale can be defined as in the case of *S*-estimation,

$$\int \rho_1[\eta(z, \theta) / s\{\eta(z, \theta)\}]\, dF_n(z) = K = \int \rho_1(t)\, dF(t).$$

Next, the τ -estimate of scale is defined by

$$\tau^2(z, \theta) = s^2\{\eta(z, \theta)\} \int \rho_2[\eta(z, \theta) / s\{\eta(z, \theta)\}]\, dF_n(z)$$

and the corresponding τ *-estimator* of parameters θ is then defined by minimizing the τ -estimate of scale, $\tau^2(z, \theta)$.

As a generalization of *S*-estimation, the τ -estimators include *S*-estimators as a special case for $\rho_1 = \rho_2$ because then $\tau^2(x, \theta) = \theta s^2(z, \theta)$. On the other hand, if $\rho_2(t) = t^2$, $\tau^2(z, \theta) = \int \eta^2(z, \theta)\, dF_n(z)$ is just the standard deviation of model residuals. Compared with *S*-estimators, the class of τ estimators can improve in terms of relative Gaussian efficiency because its breakdown point depends only on function ρ_1, whereas its asymptotic variance is a function of both ρ_1 and ρ_2. Thus, ρ_1 can be defined to achieve the breakdown point equal to 0.5 and ρ_2 consequently adjusted to reach a pre-specified relative efficiency for Gaussian data (for example, 95 per cent).

3 Methods of robust econometrics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The concepts and methods of robust estimation discussed in Sections 1 and 2 are typically proposed in the context of a simple location or linear regression models, on the assumption of independent, continuous and identically distributed random variables. This, however, rarely corresponds to assumptions typical for most econometric models. In this section, we therefore present an overview of developments and extensions of robust methods to various econometric models. As the *M*-estimators are closest to the commonly used LS and MLE, most of the extensions employ *M*-estimation. The HBP techniques are not that frequently found in the economics literature (Zaman, Rousseeuw and Orhan, 2001; Sapra, 2003) and are mostly applied only as a diagnostic tool.

In the rest of this section, robust estimation is first discussed in the context of models with discrete explanatory variables, models with time-dependent observations, and models involving multiple equations. Later, robust alternatives to general estimation principles, such as MLE and generalized method of moments (GMM), are discussed. Before doing so, let us mention that dangers of data contamination are not studied only from the theoretical point of view. There is a number of studies that check the presence of outliers in real data and their influence on estimation methods. For example, there is evidence of data contamination and its adverse effects on LS and MLE in the case of macroeconomic time series (Balke and Fomby, 1994; Atkinson, Koopman and Shephard, 1997), in financial time series (Sakata and White, 1998; Franses, van Dijk and Lucas, 2004), marketing data (Franses, Kloek and Lucas, 1999), and many other areas. These adverse effects include biased estimates, masking of structural changes, and creating seemingly nonlinear structures, for instance.

3.1 Discrete variables

To achieve a HBP, many robust methods such as LMS often eliminate a large portion of observations from the calculation of their objective function. This can cause non-identification of parameters associated with categorical variables. For example, having data on income $\{y_i\}_{i=1}^n$ of men and women, where gender is indicated by $\{d_i\}_{i=1}^n \in \{0, 1\}$, one can estimate the mean income of men and women by a simple regression model $y_i = a + bd_i$. If a HBP method such as LMS or LTS is used to estimate the model and it eliminates a large portion of observations from the calculation (for example, one half of them), the remaining data could easily contain only income of men or only income of women, and consequently the mean income of one of the groups could not be then identified. Even though this seems unlikely in our simple example, it becomes more pronounced as the number of discrete variables grows (see Hubert and Rousseeuw, 1997, for an example).

A common strategy employs a robust estimator with a HBP for a model with only continuous variables, and using this initial estimate, the model with all variables is estimated by an *M*-estimator. Such a combined procedure preserves the breakdown point of the HBP estimator: even though a misclassified values of categorical explanatory variables can bias the estimates, this bias will be bounded in common models as the categorical variables are bounded as well. See Hubert and Rousseeuw (1997) and Maronna and Yohai (2000), who combine an initial *S*-estimator with an *M*-estimator.

3.2 Time series

In time series, there are several issues not addressed by the standard theory of robust estimation because of time-dependency of observations. First, the asymptotic behaviour of various robust methods has to be established; see Koenker and Machado (1999) and Koenker and Xiao (2002) for L_1 regression; Künsch (1984); and Bai and Wu (1997) for *M*-estimators and Sakata and White (2001), Zinde-Walsh (2002) and •íz•ek (2006) for various *S*-type estimators. In these cases, the results are usually established for general nonlinear models.

Second, the effects of data contamination are more complex and widespread due to time dependency: an error in one observation is transferred, by means of a model, to others close in time. The possible effects of outliers in time series are elaborated by Chen and Liu (1993) and Tsay, Pena and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Pankratz (2000), for instance. The first work also offers a sequential identification of outliers (an alternative procedure based on τ -estimators is offered by Bianco et al., 2001). Consequently, the robust properties in time series differ from those experienced in cross-sectional data. For example, the breakdown point is asymptotically zero in the case of *M*-estimators (Sakata and White, 1995) and can be much below 0.5 for various *S*-estimators (Genton and Lucas, 2003).

A further issue specific to time series is testing for stationarity of a series. Effects of outliers are in this respect similar to those of neglected structural changes. To differentiate between random outliers and real structural changes, robust tests for change-point detection have been proposed by Gagliardini, Trojani and Urga (2005) and Fiteni (2002; 2004); the last of these papers uses τ -estimation. The asymptotics of *M*-estimators under unit-root assumption and the corresponding tests have been established, for example, by Lucas (1995), Koenker and Xiao (2004), and Haldrup, Montans and Sanso (2005). An early reference is Franke, Härdle and Martin (1984).

3.3 Multivariate regression

An important application of robust methods in economics concerns the multivariate regression case. This is relatively straightforward with exogenous explanatory variables only, see Koenker and Portnoy (1990), Bilodeau and Duchesne (2000), and Lopuhaä (1992) for the *M*-, *S*- and τ -estimation, respectively. Estimating general simultaneous equations models has to mimic either three-stage LS or full information MLE (Marrona and Yohai, 1997). Whereas Koenker and Portnoy (1990) follow with the weighted LAD the first approach, Krishnakumar and Ronchetti (1997) use *M*-estimation together with the second strategy.

3.4 General estimation principles

There are naturally many more model classes for which one can construct robust estimation procedures. Since most econometric models can be estimated by means of MLE or GMM, it is however easier to concentrate on robust counterparts of these two estimation principles. There are other estimation concepts, such as nonparametric smoothing, that can employ robust estimation (Härdle, 1982), but they go beyond the scope of this article.

First, recent contributions to robust MLE can be split to two groups. One simply defines a weighted maximum likelihood, where weights are computed from an initial robust fit (Windham, 1995; Markartou, Basu and Lindsay, 1997). Alternatively, some erroneous observations can be excluded completely from the likelihood function (Clarke, 2000; Marazzi and Yohai, 2004). This approach requires existence of an initial robust estimate, and thus it is not useful for models for which there are no robust methods available. The second approach is motivated by the *S*-estimation, namely, LTS, and defines the maximum trimmed likelihood as an estimator maximizing the product of the *h* largest likelihood contribution, that is, those corresponding only to *h* most likely observations (Hadi and Luceño, 1997). This estimator has been studied mainly in the context of generalized linear models (Müller and Neykov, 2003), but its consistency is established in a much wider class of models (•ížek, 2004).

Second, more widely used GMM has also attracted attention from its robustness point of view. A special case – instrumental variable estimation – has been studied, for example, by Wagenvoor and Waldman (2002) and Kim and Muller (2007). See also Chernozhukov and Hansen (2006) for instrumental variable quantile regression. More generally, Ronchetti and Trojani (2001) have proposed an *M*-estimation-based generalization of GMM, studied its robust properties, and designed corresponding tests. This work became a starting point for others, who have extended the methodology of Ronchetti and Trojani (2001) to robustify simulation-based methods of moments (Genton and Ronchetti, 2003; Ortelli and Trojani, 2005).

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- adaptive estimation
- categorical data
- computational methods in econometrics
- generalized method of moments estimation
- maximum likelihood
- measurement error models
- time series analysis
- two-stage least squares and the *k*-class estimator

This work was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

Bibliography

Atkinson, A., Koopman, S. and Shephard, N. 1997. Detecting shocks: outliers and breaks in time series. *Journal of Econometrics* 80, 387–422.

Bai, Z.D. and Wu, Y. 1997. General M-Estimation. *Journal of Multivariate Analysis* 63, 119–35.

Balke, N. and Fomby, T. 1994. Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. *Journal of Applied Econometrics* 9, 181–200.

Berrendero, J. and Zamar, R. 2001. Maximum bias curves for robust regression with non-elliptical regressors. *Annals of Statistics* 29, 224–51.

Bianco, A., Ben, M., Martínez, E. and Yohai, V. 2001. Regression models with ARIMA errors. *Journal of Forecasting* 20, 565–79.

Bilodeau, M. and Duchesne, P. 2000. Robust estimation of the SUR model. *Canadian Journal of Statistics* 28, 277–88.

Chen, C. and Liu, L.-M. 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association* 88, 284–97.

Chernozhukov, V. and Hansen, C. 2006. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics* 132, 491–525.

•ížek, P. 2004. General trimmed estimation: robust approach to nonlinear and limited dependent variable models. Discussion Paper No. 2004/130, CentER ,Tilburg University.

•ížek, P. 2006. Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference* 136, 3967–88.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Clarke, B. 2000. An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes. *Probability and Statistics* 20, 25–50.

Croux, C., Rousseeuw, P. and Hossjer, O. 1994. Generalized S-estimators. *Journal of the American Statistical Association* 89, 1271–81.

Davies, L. and Gather, U. 2005. Breakdown and groups. *Annals of Statistics* 33, 988–93.

Fiteni, I. 2002. Robust estimation of structural break points. *Econometric Theory* 18, 349–86.

Fiteni, I. 2004. τ -estimators of regression models with structural change of unknown location. *Journal of Econometrics* 119, 19–44.

Franses, P., Kloek, T. and Lucas, A. 1999. Outlier robust analysis of longrun marketing effects for weekly scanning data. *Journal of Econometrics* 89, 293–315.

Franses, P., van Dijk, D. and Lucas, A. 2004. Short patches of outliers, ARCH and volatility modelling. *Applied Financial Economics* 14, 221–31.

Franke, J., Härdle, W. and Martin, R. 1984. *Robust and Nonlinear Time Series Analysis*. Berlin: Springer.

Gagliardini, P., Trojani, F. and Urga, G. 2005. Robust GMM tests for structural breaks. *Journal of Econometrics* 129, 139–82.

Genton, M. and Lucas, A. 2003. Comprehensive definitions of breakdown-points for independent and dependent observations. *Journal of the Royal Statistical Society*, Series B 65, 81–94.

Genton, M. and Ronchetti, E. 2003. Robust indirect inference. *Journal of the American Statistical Association* 98, 67–76.

Gervini, D. and Yohai, V. 2002. A class of robust and fully efficient regression estimators. *Annals of Statistics* 30, 583–616.

Hadi, A. and Luceño, A. 1997. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis* 25, 251–72.

Haldrup, N., Montans, A. and Sanso, A. 2005. Measurement errors and outliers in seasonal unit root testing. *Journal of Econometrics* 127, 103–28.

Hampel, F. 1971. A general qualitative definition of robustness. *Annals of Mathematical Statistics* 42, 1887–96.

Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. 1986. *Robust Statistics: The Approach Based on Influence Function*. New York: Wiley.

Härdle, W. 1982. Robust regression function estimation. *Journal of Multivariate Analysis* 14, 169–80.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

He, X. and Simpson, D. 1993. Lower bounds for contamination bias: globally minimax versus locally linear estimation. *Annals of Statistics* 21, 314–37.

Huber, P. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73–101.

Hubert, M. and Rousseeuw, P. 1997. Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference* 57, 153–63.

Kim, T.-H. and Muller, C. 2007. Two-stage Huber estimation. *Journal of Statistical Planning and Inference* 137, 405–18.

Koenker, R. and Machado, J. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* 94, 1296–310.

Koenker, R. and Portnoy, S. 1990. M-estimation of multivariate regressions. *Journal of the American Statistical Association* 85, 1060–8.

Koenker, R. and Xiao, Z. 2002. Inference on the quantile regression process. *Econometrica* 70, 1583–612.

Koenker, R. and Xiao, Z. 2004. Unit root quantile autoregression inference. *Journal of the American Statistical Association* 99, 775–87.

Krishnakumar, J. and Ronchetti, E. 1997. Robust estimators for simultaneous equations models. *Journal of Econometrics* 78, 295–314.

Kunsch, H. 1984. Infinitesimal robustness for autoregressive processes. *Annals of Statistics* 12, 843–63.

Lopuhaä, H. 1992. Multivariate τ-estimators. *Canadian Journal of Statistics* 19, 307–21.

Lucas, A. 1995. An outlier robust unit root test with an application to the extended Nelson–Plosser data. *Journal of Econometrics* 66, 153–73.

Marazzi, A. and Yohai, V. 2004. Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of Statistical Planning and Inference* 122, 271–91.

Markatou, M., Basu, A. and Lindsay, B. 1997. Weighted likelihood estimating equations: the discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference* 57, 215–32.

Maronna, R., Bustos, O. and Yohai, V. 1979. Bias- and efficiency-robustness of general *M*-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation*, ed. T. Gasser and M. Rossenblatt. Berlin: Springer.

Maronna, R. and Yohai, V. 1997. Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference* 57, 233–44.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Maronna, R. and Yohai, V. 2000. Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference* 89, 197–214.

Müller, C. and Neykov, N. 2003. Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* 116, 503–19.

Ortelli, C. and Trojani, F. 2005. Robust efficient method of moments. *Journal of Econometrics* 128, 69–97.

Ronchetti, E. and Trojani, F. 2001. Robust inference with GMM estimators. *Journal of Econometrics* 101, 37–69.

Rousseeuw, P. and Yohai, V. 1984. Robust regression by means of *S*-estimators. In *Robust and Nonlinear Time Series Analysis*, ed. J. Franke, W. Härdle and R. Martin. Heidelberg: Springer.

Sakata, S. and White, H. 1995. An alternative definition of finite-sample breakdown point with application to regression model estimators. *Journal of the American Statistical Association* 90, 1099–106.

Sakata, S. and White, H. 1998. High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. *Econometrica* 66, 529–67.

Sakata, S. and White, H. 2001. S-estimation of nonlinear regression models with dependent and heterogeneous observations. *Journal of Econometrics* 103, 5–72.

Sapra, S. 2003. High-breakdown point estimation of some regression models. *Applied Economics Letters* 10, 875–78.

Simpson, D., Ruppert, D. and Carroll, R. 1992. On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* 87, 439–50.

Stromberg, A., Hossjer, O. and Hawkins, D. 2000. The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association* 95, 853–64.

Tsay, R., Pena, D. and Pankratz, A. 2000. Outliers in multivariate time series. *Biometrika* 87, 789–804.

Wagenvoor, R. and Waldman, R. 2002. On B-robust instrumental variable estimation of the linear model with panel data. *Journal of Econometrics* 106, 297–324.

Welsh, A. and Ronchetti, E. 2002. A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference* 103, 287–310.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Windham, M. 1995. Robustifying model fitting. *Journal of the Royal Statistical Society*, Series B 57, 599–609.

Yohai, V. and Zamar, R. 1988. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* 83, 406–13.

Yohai, V. and Zamar, R. 1993. A minimax-bias property of the least a-quantile estimates. *Annals of Statistics* 21, 1824–42.

Zaman, A., Rousseeuw, P. and Orhan, M. 2001. Econometric applications of high-breakdown robust regression techniques. *Economics Letters* 71, 1–8.

Zinde-Walsh, V. 2002. Asymptotic theory for some high breakdown point estimators. *Econometric Theory* 18, 1172–96.

How to cite this article

•ížek, P. and W. Härdle. "robust estimators in econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000254> doi:10.1057/9780230226203.1452

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rogers, James Edwin Thorold (1823 – 1890)

O. Kurer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic history; English Historical School; Rogers, J. E. T.

Article

Rogers was educated at King's College London, and Magdalen College, Oxford. From 1859 until his death he held the first Tooke Professorship of Statistics and Economic Science at King's College London. In 1862 he was elected Drummond Professor of Political Economy in the University of Oxford, a post he lost in 1868 largely because of his outspoken radical views, but to which he was re-elected in 1888. He was ordained but abandoned the clerical profession. From 1880 to 1886 he served as a rather inconspicuous member of the House of Commons.

His chief work is his monumental *History of Agriculture and Prices*, where he did much to turn economic history into the field of distribution and attempted to use more exact methods in economic historical investigations on a large scale. His work is marred by his casual deductions. He argued for a high standard of living of the English labourer during the Middle Ages and explained the subsequent deterioration by legislative interference by the landowners controlling the government.

Politically, he was greatly influenced by his friend and brother-in-law Richard Cobden. He was firmly opposed to extensive government intervention. He did however support trade unions as providing the remedy for nearly all social ills. His advocacy of laissez-faire separates him from the rest of the English Historical School, his allies in attacking theoretical economics in looking to economic history as a realistic foundation for the proper understanding and solution of contemporary social and economic problems.

Selected works

1884. *Six Centuries of Work and Wages: The History of English Labour*. London: Swan Sonnenschein.

1886–1902. *A History of Agriculture and Prices in England. From the Year After the Oxford Parliament*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(*1259*) *to the Commencement of the Continental War* (*1793*), *7 vols*. Oxford: Clarendon.

1888. *The Economic Interpretation of History*. New York: Putnam.

1892. *The Industrial and Commercial History of England*. Ed. A.G.L. Rogers. New York: Putnam. Published posthumously.

Bibliography

Ashley, W.J. 1889. James E. Thorold Rogers. *Political Science Quarterly* 4, 381–407.

De Marchi, N.B. 1976. On the early dangers of being too political an economist: Thorold Rogers and the 1868 election to the Drummond Professorship. *Oxford Economic Papers* 28, 364–80.

Hewins, W.A.S. 1897. James Edwin Thorold Rogers. In *Dictionary of National Biography*. Oxford: Oxford University Press.

Wood, J.C. 1983. *British Economists and the Empire, 1860–1914*. Beckenham, Kent: Croom Helm.

How to cite this article

Kurer, O. "Rogers, James Edwin Thorold (1823–1890)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000169> doi:10.1057/9780230226203.1453

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Roos, Charles Frederick (1901 – 1958)

Karl A. Fox

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Econometric Society; econometrics; mathematical economics; Roos, C. F.

Article

Born on 18 May 1901, in New Orleans, Roos completed his Ph.D. in mathematics at Rice Institute in 1926. Influenced directly by his supervisor Evans (1922; 1924; 1930) and indirectly by Volterra, his main interests in graduate work were the calculus of variations, integral equations, and applications of those areas of mathematics to problems in dynamic economics.

Although he published several brilliant articles (Roos, 1925; 1927a; 1927b; 1927c; 1928; 1930), Roos found no journal which would readily accept manuscripts in which he combined economics, mathematics and sometimes statistics at suitably advanced levels (cf. Roos, 1934, p. xiii). Spurred by similar frustrations, Frisch and Roos jointly took the initiative which led to creation of the Econometric Society in 1930 (of which Roos became President in 1948) and publication of its journal, *Econometrica*, from 1933 on.

In 1930 Roos set out to write a treatise on dynamic economics; he published an important book under that title in 1934. It was reviewed enthusiastically by Tintner (1936) and uncomprehendingly by Freeman (1935). *Dynamic Economics* (1934) is a brilliant combination of mathematical economic theory and applied econometrics. Roos's mathematical approach inspired Tintner to write a dozen articles on dynamic economic theory (for example, Tintner, 1938).

Roos held a series of administrative positions during 1931–7 and published a major book on *NRA Economic Planning* (1937). In 1938 he founded an econometric consulting firm in New York and directed it until his death. Examples of his later work are Roos and von Szeliski (1939a; 1939b) and Roos (1955; 1957). He died in Milwaukee on 7 January 1958.

Hotelling (1958) describes Roos as ‘a unique and outstanding figure’, while Davis (1958) presents a complete list of his writings.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Selected works

1925. A mathematical theory of competition. *American Journal of Mathematics* 47(July), 163–75.

1927a. Dynamical economics. *Proceedings of the National Academy of Sciences* 13, 145–50.

1927b. A dynamical theory of economic equilibrium. *Proceedings of the National Academy of Sciences* 13, 280–85.

1927c. A dynamical theory of economics. *Journal of Political Economy* 35, 632–56.

1928. A mathematical theory of depreciation and replacement. *American Journal of Mathematics* 50, 147–57.

1930. A mathematical theory of price and production fluctuations and economic crises. *Journal of Political Economy* 38, 501–22.

1934. *Dynamic Economics: theoretical and statistical studies of demand, production and prices*. Cowles Commission Monograph No. 1. Bloomington, IN: Principia Press.

1937. *NRA Economic Planning*. Cowles Commission Monograph No. 2. Bloomington, IN: Principia Press.

1939a. (With V. von Szeliski.) *The Dynamics of Automobile Demand*. Detroit: General Motors Corporation.

1939b. (With V. von Szeliski.) The concept of demand and price elasticity; the dynamics of automobile demand. *Journal of the American Statistical Association* 34, 652–66.

1955. Survey of economic forecasting techniques. *Econometrica* 23, 363–95.

1957. *Dynamics of Economic Growth: the American Economy, 1957–1975*. New York: Econometric Institute.

Bibliography

Davis, H.T. 1958. Charles Frederick Roos. *Econometrica* 26(4), 580–89. Contains a complete bibliography of Roos's writings (91 items).

Evans, G.C. 1922. A simple theory of competition. *American Mathematical Monthly* 29(10), 371–80.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Evans, G.C. 1924. The dynamics of monopoly. *American Mathematical Monthly* 31(February), 77–83.

Evans, G.C. 1930. *Mathematical Introduction to Economics*. New York: McGraw-Hill.

Freeman, H.A. 1935. Review of C.F. Roos, *Dynamic Economics*. *American Economic Review* 25, 520.

Hotelling, H. 1958. C.F. Roos, econometrician and mathematician. *Science* 128, 1194–5.

Tintner, G. 1936. Review of Dynamic Economics. *Journal of Political Economy* 44, 404–9.

Tintner, G. 1938. The theoretical derivation of dynamic demand curves. *Econometrica* 6, 375–80.

How to cite this article

Fox, Karl A. "Roos, Charles Frederick (1901–1958)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000171> doi:10.1057/9780230226203.1454

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Roscher, Wilhelm Georg Friedrich (1817 – 1894)

B. Schefold

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cameralism; German Historical School; history of economic thought; methodology of economics; self-interest; stages of economic development

Article

Roscher was born in Hannover into a well-established civil service family. He studied history and political science in Göttingen and Berlin. In 1840 he became lecturer in both subjects at Göttingen, in 1843 he was appointed extraordinary professor of political economy, and in the next year was promoted professor. In 1848 he transferred to Leipzig, where he taught for the rest of his life. Roscher had a Protestant background and was deeply religious, adhering to a rather ‘primitive form of religious belief’ (Max Weber, 1903–6).

Roscher may be considered as one of the most important German economists of his time. He was one of the founders and the leading exponent of the German ‘older’ Historical School. He did not develop any new theory: his main contribution to political economy lay in the field of method. He adhered to what he called the ‘historical-physiological method’, as opposed to the ‘idealistic method’ (1842; 1854–94, vol. 1, pp. 43–56). This inductive method intended to provide a description of the actual course of economic development and of real economic life. Thus, Roscher tried to analyse laws of economic development by comparing the history of different people and nations and showing analogies in stages of their development. The emphasis was on historical relativism: economic behaviour depended to a large extent on the specific national and historic conditions of the different people and nations. This implied that a nation had to be regarded as a whole, as an ‘organic unity’, and not as the mere sum of individuals. This was opposed to what Roscher called the ‘idealistic method’, which intended to provide an ideal picture, logically derived from abstract principles, of the functioning of the economic system. An example of this was the classical economists' deduction of economic laws from a system of hypotheses. Although Roscher emphasized that in economic analysis there existed generally no definite causal relationships but reciprocal ones, he did not reject the existence of ‘laws of motion’ within economic life. However, these laws were distinct from laws of natural science in that they dealt with free human

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
beings gifted with reason and hence with changing motives for action (1854–94, vol. 1, pp. 26–9). Roscher was closer to the theoretical system of the classics than the exponents of the ‘younger’ historical school. He tended to regard it as the appropriate system of analysis of the current stage of economic development. He only modified and supplemented it with a careful historical analysis, but he may still be regarded as being in the classical tradition.

The first volume of Roscher's main work, *System der Volkswirtschaft* (1854–94: 1854) still looked very much like a traditional textbook. It analysed essentially the same topics as the classical economists – production, distribution and prices. Roscher was already strongly influenced by supply and demand approaches, but still determined the exchange value of a commodity by its cost of production. His theory of rent was Ricardian and his thinking about population development followed Malthusian patterns. Differing from classical textbooks, Roscher supplemented the theoretical analysis with a historical description – the reader finds the history of rent, interest and wages, of population development, of the prices of necessary and luxury commodities, and of luxury in general. Roscher accepted the classical notion of individual self-interest as a central axiom of modern economic behaviour, but he did not follow the classical patterns in deriving his economic principles from this assumption. As a result of his religious beliefs, he included human conscience as a regulating mechanism into his analysis of the role of self-interest (1854–94, vol. 1, pp. 20–3).

The other four volumes of the *System der Volkswirtschaft* (1859; 1881; 1886; 1894), which may be perceived as his main contribution to applied economics, were even more historically oriented and focused on agriculture, trade and industry, public finance, social policy and poor relief.

Roscher classified economic development into stages of maturity. The economic factors that govern the development of nations were land, labour and capital which subsequently dominated the different stages (1861, ch. 1). Later, Roscher presented a more detailed analysis of stages of political and societal development (1892) on the basis of a classification of the different forms of government during history: early patriarchal kingdom, aristocracy of knights and priests, absolute monarchy, democracy. The latter then degenerated into a plutocracy, which is followed by a military dictatorship Roscher called ‘Caesarismus’. Roscher did not systematically attempt an integration of his theory of political development and the stages of economic evolution.

He wrote several contributions on the history of economic thought. His compendium on the history of political economy in Germany (1874) was his most outstanding work and has remained important. Roscher may be regarded as the most eminent historian of cameralism and early German political economy. His treatise on economic problems of the location of large towns (1871) was an original contribution to economic theory.

Roscher supported German imperialism. In order to secure raw materials and markets for German goods, as well as to relieve the national labour market and prevent social unrest, he advocated an expansive German colonial policy, especially towards Eastern Asia, where he saw Germany's colonial future (1885). He was a conservative but he remained all his life unaffiliated to any political party or group.

See Also

- Historical School, German

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Selected works

1842. *Leben, Werk und Zeitalter des Thukydides*. Göttingen.

1854–94. *System der Volkswirtschaft*. Stuttgart: Cotta. Vol. 1: *Die Grundlagen der Nationalökonomie*, 1854. Trans. from 13th edn by J.J. Lalor as *Principles of Political Economy*, 2 vols, New York, 1878. Vol. 2: *Nationalökonomik des Ackerbaues und der verwandten Urproduktionen*, 1859. Vol. 3: *Nationalökonomik des Handels und Gewerbeflei β es*, 1881. Vol. 4: *System der Finanzwissenschaft*, 1886. Vol. 5: *System der Armenpflege und der Armenpolitik*, 1894.

1861. *Ansichten der Volkswirtschaft aus dem geschichtlichen Standpunkt*. Leipzig and Heidelberg: Winter.

1871. *Betrachtungen über die geographische Lage der grossen Städte*. Leipzig.

1874. *Geschichte der Nationalökonomik in Deutschland*. Munich: Oldenbourg.

1885. *Kolonien, Kolonialpolitik und Auswanderung*. Leipzig: Winter Part II, ch. 1. Translated from 3rd edn by E.H. Baldwin and E.G. Bourne as *The Spanish Colonial System*, New York, 1904.

1892. *Politik: Geschichtliche Naturlehre der Monarchie, Aristokratie und Demokratie*. Stuttgart.

Bibliography

Cunningham, W. 1894–5. Why had Roscher so little influence in England? *Annals of the American Academy of Political and Social Sciences* 5, 317–34.

Weber, M. 1903–6. *Roscher und Knies und die logischen Probleme der Historischen Nationalökonomie*. In *Gesammelte Aufsätze zur Wissenschaftslehre*. Tübingen: Mohr, 1922.

How to cite this article

Schefold, B. "Roscher, Wilhelm Georg Friedrich (1817–1894)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000173> doi:10.1057/9780230226203.1455

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rosen, Sherwin (1938 – 2001)

Robert Tamura
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Sherwin Rosen made fundamental contributions in equilibrium theory, human capital theory, income distribution theory and investment theory. One characteristic feature of Rosen's work is the minimal use of heterogeneity of individuals. His work explains price disparities, differential earnings, and investment cycles. Underlying differences in characteristics produce price differences. Human capital theory is enriched by characterization of accumulation beyond schooling. Skewed income distributions arise from outcomes of tournaments, superstars from economies of scale or hierarchical complementarities. Rational investment cycles occur when the capital stock is large relative to investment, and when the breeding stock is large relative to the overall stock.

Keywords

agglomeration economy; American Academy of Arts and Sciences; American Economics Association; baby boom; compensating differentials; differentials; division of labour; earnings generating function; Econometric Society; Economies of scale; endogenous comparative advantage; equilibrium theory; fertility; game theory; Giffen goods; hedonic models; human capital; income and consumption inequality; income distribution theory; information aggregation; investment cycles; life-cycle earnings; marginal rate of substitution; Midwest Economic Association; National Academy of Sciences; non-convexity; occupational choice; optimal knowledge accumulation; price differences; Rosen, S.; Roy model; Society of Labor Economists; superstars; symmetric ignorance; value of life

Article

Rosen was born in Chicago on 29 September 1938. He died in Chicago on 17 March 2001. He earned his BS in economics from Purdue University in 1960. He obtained his graduate economics degrees from the University of Chicago: his MA in 1962 and his Ph.D. in 1966. His first appointment was as assistant professor of economics at the University of Rochester in 1964. Promoted to associate professor in 1968

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and full professor in 1970, he became the Kenan Professor of Economics in 1975. Rosen returned to the University of Chicago in 1977, and became the Bergman Professor of Economics in 1983. From 1992 until his death he served as the Edwin A. and Betty L. Bergman Distinguished Service Professor. In addition, he served as department chairman during 1988–94. He was a Senior Research Associate of the National Bureau of Economic Research from 1968 and a Senior Research Fellow of the Hoover Institution during 1983–96 before becoming a Senior Fellow in 1997.

Rosen was elected a fellow of the Econometric Society in 1976, and a fellow of the American Academy of Arts and Sciences in 1984. He became a member of the National Academy of Sciences in 1998. He was President of the Midwest Economic Association during 1996–7, President of the Society of Labor Economists in 2000, and President of the American Economics Association for 2001.

Rosen was a prolific scholar and one of the leading economists of his generation. His contributions spanned many fields, including equilibrium theory, human capital theory, income distribution theory and investment theory. His research provided the theoretical underpinnings of labour economics, urban economics and health economics. A unifying aim of his research is to explain differential market outcomes. Price differences of goods can be explained by their differential amounts of characteristics. These price differences could be driven by differences in preferences arising from wealth differences or differences in technologies available to firms. For example, cars sell for different prices because they contain different attributes, and workers earn different amounts across jobs because the jobs have different characteristics. Earnings may differ if workers differ in their human capital. Life-cycle earnings are explained from the characterization of human capital accumulation beyond formal schooling. Returns to higher education are best modelled as arising from revealed preference of workers, both college-educated and non-college educated. Earnings may differ between identical workers because they are in different job classifications. Small differences in worker productivity can manifest themselves in large earnings differences if there are production scale economies (creating superstars, for example), strong complementarities, or increasing returns in skill use. Finally, differences in returns and investments can arise from predictable future demand shifts or unpredicted contemporaneous demand shocks. Rational investment cycles are likely if investment is small relative to the stock of capital, and if the seed capital is a large proportion of the stock of capital.

Rosen was the author of two books, *A Disequilibrium Model of Demand for Factors of Production* (Nadiri and Rosen, 1973) and Markets and Diversity (2005), and editor of three collections: *Studies in Labor Markets* (1981), *Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure* (Rosen and Winship, 1988), and *Implicit Contract Theory* (1994).

Equilibrium theory

Rosen's 1974 article ‘Hedonic prices and implicit markets: product differentiation in pure competition’ is the quintessential example of his work in equilibrium theory. Consider the following labour market application. Rosen's analysis allows for a job to be characterized by N dimensions, but for clarity we focus on only two, its wage and its dirtiness. Some jobs are dirtier than others. They provide meaner working conditions including unheated and/or non-air-conditioned workplaces, less pleasant coworkers, few or no promotion possibilities, high unemployment risk, large variability in hours demanded by the employer, inflexible hours of work, fewer vacations, worse fringe benefits like poor or no health insurance or disability insurance, poor pensions, and so on. Consider aggregating all of these features

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
into a single measure called ‘dirt’. A worker likes wages and dislikes dirt. Employers can offer any combination of wages and dirt as a package to prospective workers. Assume that (*a*) worker preferences are convex; hence a worker's dislike for dirt increases with the level of dirt on the job, and he or she requires ever larger increases in wages to accept an additional unit of dirt as the level of dirt increases; and (*b*) firm production technologies are convex; firms require greater wage reductions for a unit reduction in dirt as the job becomes less dirty. There are three extreme cases. First, if all workers are identical in wealth and preferences, and all employers have access to the same technology of production, then there is a single equilibrium point. This occurs at the tangency of the representative worker's iso-utility curve and the representative employer's iso-profit curve. With free entry, competition drives the equilibrium to the zero profit iso-profit curve. In the second case, suppose all workers are identical in wealth and preferences, but employers have different technologies. For example, mining firms find it more costly than software design firms to provide cleaner work environments. In equilibrium the economist will observe a locus of points, which traces out the representative worker's iso-utility curve, and each observed increase in dirt is associated exactly with the worker's compensating differential to accept the increased dirtiness. Finally, suppose workers have different preferences, say because of wealth differences. Assume that dirt is an inferior good. Let all firms have access to a single technology. In equilibrium the economist observes a locus of points, which traces out the representative firm's zero iso-profit curve. The second example identifies preferences, the third example identifies technology. Of course, the world is not so stark or clean for an economist. Preferences are heterogeneous, workers have differing skill levels, firms have different technologies. Thus, econometrically the problem is one of finding controls that allow for identification (see Ekeland, Heckman and Nasheim, 2004). One important application to the labour market is Murphy (a Rosen student) and Topel (1987).

Rosen's (1974) paper serves as the benchmark for thinking about how markets link customers of multiple characteristic goods and services with the suppliers of these complex goods and services. One important application of this model is by Roback (1982), a Rosen student. Her model examines the compensating differentials in worker wages and land rents arising from differences in location-specific amenities, say, climate or population density. Another application of this hedonic approach is the examination of the increased wages that firms pay to workers in order to induce them to accept greater risks to their health or, in particular, their lives. Rosen and Thaler (1976) allow variation in earnings due to variation in on-the-job risks to life, controlling for productivity (schooling and experience) and in other job characteristics to identify the reservation price of mortality risk for the typical worker. This allows for the calculation of the economic value of a life. Rosen (1988) revisits this arena by examining the valuation placed on increasing longevity. These two papers served as inspiration for an entire sub-field of health economics, highlighted by Murphy and Topel (2003).

Rosen (1978) examines the assignment solution of workers to tasks within an organization, in a world with a fixed number of inputs and many worker types. Rosen shows that the division of labour corresponding to the optimum assignment determines the marginal rates of substitution between worker types or between job categories. Thus the division of labour determines the extent of product and factor market substitutions in the economy. This paper provides an application of economics to the optimal determination of job types, or the efficient bundling of activities into a job. Rosen (1982b) extends the analysis. It is further generalized in Tamura (1992); with a continuum of intermediate tasks, and N different worker types, each of measure 1, output can be shown to come from the following reduced form:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$Y = \left\{ \sum_{i=1}^{N} h_i^{\rho} \right\}^{\frac{1}{\rho}}$$
(1)

where type *i* workers have h_i units of human capital and $0 < \rho < 1$. With each individual a set of measure zero, each worker is paid the marginal product of his or her human capital and, given the constant returns to scale in the distribution of human capital, output is completely exhausted. However, since $\rho < 1$, there is an agglomeration economy in participation. Earnings for an individual of type *j* are the product of the marginal product of human capital of type *j* workers, w_j, and the amount of human capital of type *j* workers, h_j, or:

$$y_j = w_j h_j = \left\{ \sum_{i=1}^{N} h_i^{\rho} \right\}^{\frac{1-\rho}{\rho}} h_j^{\rho-1} h_j = \left\{ \sum_{i=1}^{N} h_i^{\rho} \right\}^{\frac{1-\rho}{\rho}} h_j^{\rho}$$
(2)

Assume that the human capital of worker type *j* grows at rate λ_j. Suppose workers of type *j* have more human capital than workers of type *i*. If the growth rates of human capital differ across type, then the relative earnings of these two worker types will change. In particular, notice:

$$\frac{y_{jt+1}}{y_{it+1}} = \left(\frac{\lambda_j h_{jt}}{\lambda_i h_{it}} \right)^{\rho} = \left(\frac{\lambda_j}{\lambda_i} \right)^{\rho} \frac{y_{jt}}{y_{it}}$$
(3)

Thus earnings become more (less) unequal if $\lambda_j > (<) \lambda_i$. Nothing about differences in firm investments in technical change is required, merely differences in the abilities of workers to continue to accumulate human capital. As Rosen (1972a; 1972b) notes, higher education can help individuals become better learners. Thus rising wage inequality can be the result of rising task specialization of the more skilled. Hence the works of Acemoglu (1998; 2002) can be thought of as arising from underlying primitives of differential worker abilities to learn.

Rosen also made fundamental contributions with Li, Mussa, and Suen. Mussa and Rosen (1978) provide

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

an equilibrium analysis of the product quality choice of monopolists. Li and Rosen (1998) examine the effect of breaches of contracts, unravelling, on optimal assignments of workers to firms when worker quality is uncertain. Li, Rosen and Suen (2001) examine the role of committees in information aggregation. If individuals have idiosyncratic information, committees help to aggregate the information. However, if committee members have conflicting preferences, the only equilibrium truth-telling rules are binary: yes or no, promote or do not promote, hire or not hire, keep or fire.

Human capital theory

Rosen applied his characteristics approach in order to make fundamental contributions to human capital theory. Rosen (1972a) models jobs as producing both output and learning opportunities for workers. Jobs differ in their learning opportunities. These opportunities are costly to firms; they produce less market output in return for producing more skills for workers in the future. With identical workers and many firms in equilibrium, young workers seek out the firms with the best learning opportunities. Workers accept lower earnings to pay the firm for the learning opportunities associated with their job. As they gain experience and skill, but have fewer years of work remaining, they switch to jobs offering less rapid learning and greater production. The theory produces occupational switching and the typical age-earnings profile. With heterogeneity in ability, the model is capable of producing a distribution of outcomes by age. The most able learners choose jobs with the most rapid learning possibilities, while less able learners choose to forgo those jobs entirely. Rosen (1972b) displays an early grasp of dynamic programming. He formulates the optimal accumulation of knowledge from learning by producing by analysing the excess valuation of production over and above current profits for the acquisition of higher future profits. He formulates a model of optimal knowledge accumulation as an explanation for technological progress. Curiously, Rosen notes that a stationary solution to an infinite horizon problem is not possible. However modern endogenous growth models in fact do take his first functional form in the paper. As long as output grows at a constant rate, knowledge growth will continue at a constant rate. Rosen presages Romer (1986) and Lucas (1988) by arguing that knowledge creation is likely to have important spillovers across workers and industries.

In a contribution to a Feschrift volume for his advisor H. Gregg Lewis, on the occasion of his retirement from the University of Chicago, Rosen (1976) applies a novel twist on the problem of life-cycle earnings. He considers the standard formulation of time *t* wealth value of human capital:

$$W(t) = \int_t^N y(s)\, e^{-r(s-t)}\, ds \qquad (4)$$

where *N* is an exogenously specified retirement age, *y*(*s*) is the earnings at age *s*, and the individual faces a constant interest rate, *r*. Differentiating (4) with respect to *t* and rearranging produces:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$rW(t) - \dot{W}(t) = y(t) \tag{5}$$

The standard interpretation is to consider the first term as the potential earnings, the second term as the dollar cost of human capital investment and the right-hand side as observed earnings. In his words (1976, p. S46), ‘The method adopted here is to go behind the scenes of (5) and use the theory to parameterize *y*(*t*) directly in the form of restrictions on the unobservable *W*(*t*).’

Rosen considers the accumulation of human capital as a self-directed process as in Ben-Porath (1967); *y*=*Rk*(1–*s*), where *s* is the proportion of knowledge spent on learning. Rosen considers two possible tractable formulations on the earnings generating function: (*a*) the learning function depends on total capital resources spent on accumulation, $\dot{k} = h(sk)$, which produces $y = Rk - h^{-1}(\dot{k})$, and (*b*) the learning function is linear in the stock of knowledge, $\dot{k} = h(s)k$, which produces $y = Rk[1 - h^{-1}(\frac{\dot{k}}{k})]$. Rosen chooses the latter functional form. The reader familiar with endogenous growth models will immediately see that his preferred specification is the *Ak* model of Jones and Manuelli (1990), Rebelo (1991) and Lucas (1988). Rosen also assumes that children are born with a fixed fraction of their parents' capital, such that at age 0, $k^{t+1}(0) = \gamma k^{t}(0) > k^{t}(0)$; thus he formulates Lucas (1988) without the human capital externalities, but with perpetual growth! Despite the difficulty imposed by finite time horizon models, Rosen derives closed form solutions for the optimal rate of human capital accumulation, *s*(*t*), as well as for *k*(*t*) and *y*(*t*). Unfortunately, economists appear to have misgivings about working with hyperbolic sines, cosines and cotangents! Rosen (1976) produces the standard life-cycle shapes of observed earnings, potential earnings and human capital investments. From his explicit analytic solutions, Rosen is able to estimate the structural parameters of this model using census data. His estimates are broadly consistent with empirical results on rates of returns to schooling. More interestingly, he conducts counterfactual experiments about the nature of college. Schooling could be purely vocational in substance, increasing the knowledge of the future worker. It could also make the individual permanently more productive at future learning. Rosen posits different pairs of learning efficiencies and initial knowledge immediately after college completion that make the college graduate indifferent to college or work after high school graduation. He conducts the same counterfactual for high school graduates. Clearly, this thought experiment is one that foreshadows his seminal work with Robert Willis.

Willis and Rosen (1979) present a version of the Roy (1951) model for educational choice. Individuals can choose between stopping after high school graduation and continuing on to college. In their model there exists comparative advantage. Revealed preference implies that those workers that stopped after high school chose optimally to ignore college because their own rate of return to college education would be less than their cost of funds. Revealed preferences of college graduates imply the opposite. Now, if in addition high school graduates have an absolute advantage in high school occupations relative to what college graduates could earn in those jobs as high school graduates, estimated rates of return to college would be biased. The estimated rate of return to college would be below the true return to the college graduate, but more than the prospective return to the high school graduate. This revealed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
preference of educational-occupational selection indicates that there might not exist much ability bias in estimated rates of return to college. After nearly 30 years of empirical work, this is the dominant view in the economics profession (see Ashenfelter and Krueger, 1994; Ashenfelter and Zimmerman, 1997; Ashenfelter and Rouse, 1998).

Rosen (1983) identifies an increasing returns feature to human capital. The key point is that the marginal cost of creating human capital is independent of the intensity of use of human capital. That is to say, human capital investment is like a sunk cost. Once acquired, the marginal cost of using human capital is zero. The more intensely an individual uses his or her human capital, the greater the return to the human capital. Identical workers have an incentive to specialize their human capital or to endogenously choose their comparative advantage. This endogenous comparative advantage is a further extension of Willis and Rosen (1979). Thus, more specialized workers spend their careers in large markets that more fully utilize their skills. The largest metropolitan areas will be home to the most specialized human capital, in any field. Medical specialists will agglomerate in large metro areas and not smaller cities or rural areas; see Baumgardner (1988a; 1988b), a Rosen student. The increasing returns to utilization and endogenous comparative advantage model he envisions are explored in Tamura (1992; 1996; 2002; 2006).

Income distribution theory

Rosen made seminal contributions to understanding the functional distribution of earnings in the economy. Underlying his work is a search for the answer to the fundamental question: ‘Why are earnings so skewed?’ Furthermore, his work operates under the constraint that the answer should arise from a minimal amount of heterogeneity in underlying individual talent. Ideally, *ex ante* identical individuals would produce the observed skewed earnings distribution. One can view Rosen's work in human capital theory specifically as producing answers to this question with close to this ideal assumption of identical initial human capital endowments. In addition to his human capital programme, his research in this category includes Lazear and Rosen (1981) and his solo authored works (1981; 1982a; 1986a; 1997a). In Lazear and Rosen (1981) and Rosen (1986a), workers are paid as a result of internal relative comparisons. Assume that worker effort is not observable. If individual worker productivity is measured with noise, but a large proportion of that noise is common for all workers of the firm or for workers at similar levels within the firm, the use of relative productivity in order to determine compensation is efficient. This is because, by using relative comparisons, the effects of the unmeasured noise tend to be eliminated or greatly mitigated. For all workers, the wage bill must equal the value produced by the workers. However, workers are paid in relation to their place in the tournament, and hence paid in line with their job title. Increasing the spread between job levels or ranks raises the effort level of workers in the tournament. The larger the total wage bill, with the number of workers at the firm held constant, the greater is the average effort level, and the greater the average ability of workers. In noisier environments, the spread between winning and losing workers must be greater than the earnings spread in more predictable environments. This larger spread is required because increasing noise dissipates the return to worker effort. Thus, in order to elicit the same level of effort, noisier industries must have greater earnings disparity.

Lazear and Rosen (1981) deals with a single contest. However, internal hierarchies are tournaments with many rounds. As a worker successfully progresses up the job ladder, there are fewer and fewer rounds

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
left to play. In order to maintain the efficient level of effort, the prize gap must increase. Hence the gap between the CEO and the second in command of the firm must be larger than the gap between the second in command and his or her direct subordinates. Even if the CEO is only marginally better than the second in command, this larger prize must be given in order to provide the correct incentives throughout the organization. The pay gap serves to motivate not only the CEO but all workers in the internal hierarchy, and especially those close to the CEO in rank. As Lazear (2003, p. 13) notes, ‘(t)he theory helps explain why there is a larger spread in earnings between the top and bottom in new industries than in old ones.’ As a consequence newer industries often pay workers in stocks or stock options in order to enlist greater effort levels. When the winners of these new companies in new industries are anointed, the stock options induce huge pay differentials within and across firms in this industry. Rosen (1986a) also shows that the single elimination tournament among players with heterogeneous talents is more likely to be the efficient tournament design than a round robin format. It promotes survival of the fittest at a more rapid rate. Rosen (1986a) shows this in an environment of ‘symmetric ignorance’, or the ‘veil of ignorance’, in which all players know only the common distribution from which all players' talents are drawn, including their own. Through Bayesian updating, survival in each round provides information about the ability of the contestant. These papers were among the first to apply game theory to labour economics. Furthermore, in the conclusion, Rosen (1986a) identifies the interesting area of further research, namely, the effects of player optimism or pessimism. His conjectures again presage the seminal works of Benabou and Tirole (2002; 2003; 2004) on micro models of behavioural attitudes.

On the question of skewed income distribution from small initial differences, Rosen (1981) shows that markets where costs of reproduction are trivial overwhelmingly choose to reward the individual who is perceived to be the best, even when the best is only trivially better than the second-best performer. Hence the entertainment industry with low-cost reproduction of movie prints greatly increases the skewness of the earnings distributions of actors, producers and directors in comparison with the earnings distributions of these same labour inputs in the days of the travelling show, or the Broadway theatre. Adding books, LPs, video tapes, CDs, DVDs, and so on continues to lower the cost of ‘owning’ a performance. Hence, an individual perceived by the market to be the best will harvest the overwhelming bulk of the demand. The individual performance is captured or recorded once, and then can be replicated at near zero marginal cost. An additional example is the falling costs of journal publication, producing rising skewness in the earnings distribution in academics.

This research leads directly to Rosen (1982a). The CEO can supply the same effort level working for a family firm with $1 million of revenue or a publicly traded firm with $100 billion of revenue. The marginal return to talent, however, greatly varies between the two. Hence, those workers with the lowest disutility of effort, or the greatest productivity of effort, will be more valuable working for organizations with greater sales. Essentially, managers are distributing their efforts across a greater scope of inputs, just as superstar performers spread their efforts to ever larger groups of customers. Once again, marginally better managers will earn significantly more than slightly less able managers because they work with a much larger scale of complementary inputs.

Rosen's (1997a) presidential address to the Society of Labor Economists shows that there is an endogenous reason for income inequality among *ex ante* identical workers. His model relies on non-convexity of preferences. These can arise from a variety of primitives: Friedman (1953) provides one; Bergstrom (1986) utilizes state dependent utility functions; Becker, Murphy and Werning (2005) use

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
status; Becker, Murphy and Tamura (1990) and Tamura (1994) produce one with non-convexities in human capital. With these assumptions, Rosen demonstrates that the equilibrium and efficient outcome includes occupation lotteries or specialized investment in order to convexify the non-convex portion of utility. The winners get to enjoy higher utility, and the losers enter a lower level of utility. *Ex ante*, individuals are better off for entering into the lottery.

Investment theory

Another area that receives considerable attention from Rosen is investment theory. Rosen applies his customary analytical insights to understanding the dynamics of investment, particularly in areas where the 'time to build' aspect is significant (see, for example, Kydland and Prescott, 1982). This occurs in Rosen (1983), where the costs of acquiring human capital are separable from the rate of intensity of use. Furthermore, Rosen (1987) focuses on the role that investment in anticipation of future demand plays in price and quantity dynamics. These issues are explored in more detail in Rosen and Topel (1988), who produce a rational explanation for boom–bust cycles, observed in the hog market and elsewhere. With a rising supply price, rational individuals build in anticipation of demand. When investment is a small fraction of the stock of the durable good, anticipated future demand shocks produce contemporaneous price changes. If an anticipated large permanent increase in demand will occur five years into the future, then investment will occur today. The immediate rise in investment and the slow shifting out of supply leads to a reduction in current rentals to the service flow. Investment continues rising because the value of the durable good continues to rise as the number of periods before the permanent demand shift shrinks. Until the demand shock appears, rental rates continue to fall; this is the bust phase of the cycle. When the demand shift arrives, rental rates jump, but less than they would have with no anticipatory investment; this is the boom phase of the cycle. Rosen and Topel produce a bust-boom cycle that is created because of the anticipated nature of the demand shock. Unanticipated shocks would produce even more dramatic changes in the rental rate of the durable, but no boom–bust characteristics. These insights are evident in Rosen (1992) and, from two dissertations he supervised, Siow (1984) and Zarkin (1985).

One might ask when known future demand shocks would arise. Two examples are the baby boom and Disney World. The baby boom, starting in 1946 and continuing through 1964, produced above-trend rates of fertility in American women. It is known that by age six a child must be enrolled in school. Hence, with generally little uncertainty, college students in the mid-1940s would have foreseen an increase in the demand for primary school teachers starting in 1951, for secondary school teachers in 1959, and for college faculty in the 1960s. In the second example, Walt Disney announced the construction of Disney World in Orlando, Florida, in the mid-1960s. It opened to the public only in 1970; but the model predicts increased construction of hotels, housing, schools, shopping areas, and so on in anticipation of the future increase in population. Examples of unexpected shocks would be the space race induced by Sputnik, the Soviet satellite, in the late 1950s, the space-science bust of the late 1960s, and the unexpected end of the baby boom.

Murphy, Rosen and Scheinkman (1994) apply the dynamic model of investment to the cattle industry. The long gestation cycle of cows (eight months) and their relatively short reproductive life (eight to ten years) implies that the breeding stock is likely to be a very large portion of the overall cattle herd. Thus,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
demand shocks are likely to greatly affect the breeding stock and hence the industry's ability to respond to future demand shocks. The authors show that their model does an excellent job of fitting the data from 1875 to 1990, despite the change in technology arising from corn feeding as opposed to range feeding, introduced in the 1930s and 1940s, which halved the time of the beef production cycle.

Rosen (1999) re-examines the Irish potato famine. He disproves the idea that potatoes were a Giffen good. As in the cattle industry, seed potatoes are a large proportion of the crop. Rosen argues that rational expectations of Irish potato farmers, who assumed that the potato blight was a temporary and not a permanent productivity shock, sealed their doom, since they did not consume their seed stock. When the blight turned out to be permanent, their exposure to imminent starvation was *ex post* predictable and tragic.

Conclusion

A measure of Rosen's influence on the economics profession can be seen by the number of published academic tributes to him (see, for example, Hartog, 2002; Lazear, 2003; Sanderson, 2001). In addition to his fertile research, Rosen possessed a great talent for synthesis, not only in his own work, as testified by Lazear (2003), but in entire fields. This is evident by his seminal contributions in this regard for human capital theory in ‘Human capital: a survey of empirical research’ (1977), ‘Implicit contracts: a survey’ (1985), ‘The theory of equalizing differences’ (1986b), ‘Public employment taxes and the welfare state in Sweden’ (1997b) and ‘Theories of the distribution of earnings’ (Neal and Rosen, 2000). Rosen was influential in much of Lazear's work (1995) on personnel economics.

Sherwin Rosen married Sharon Girsburg from Chicago. They were the embodiment of the marriage covenant, a beacon to all who knew them. They shared their love for 40 years, and had two daughters, Jennifer and Adria. Sherwin Rosen was a beloved professor at the University of Rochester and the University of Chicago. He was treasured by his colleagues and affectionately admired by graduate students. His concern for the success of his junior colleagues and of graduate students was legendary. His keen insight lit the seminars and classes, and his infectious laughter filled the hearts of his colleagues and graduate students. His work continues to illuminate the way for the economics profession, and his memory inspires and warms his former colleagues and students.

See Also

- compensating differentials
- hedonic prices
- human capital
- personnel economics
- Roy model
- superstars, economics of
- value of life

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1972a. Learning and experience in the labour market. *Journal of Human Resources* 7, 326–42.

1972b. Learning by experience as joint production. *Quarterly Journal of Economics* 86, 366–82.

1973. (With M. Nadiri.) *A Disequilibrium Model of Demand for Factors of Production*. New York: NBER, Columbia University Press.

1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* 82, 34–55.

1976. A theory of life earnings. *Journal of Political Economy* 84, S45–68.

1976. (With R. Thaler.) The value of saving a life: evidence from the labor market. In *Household Production and Consumption: NBER Studies in Income and Wealth*, vol. 40, ed. N. Terleckyj. New York: Columbia University Press.

1977. Human capital: a survey of empirical research. In *Research in Labor Economics*, vol. 1, ed. R. Ehrenberg. Amsterdam: North-Holland.

1978. Substitution and the division of labor. *Economica* 45, 235–50.

1978. (With M. Mussa.) Monopoly and product quality. *Journal of Economic Theory* 18, 301–17.

1979. (With R. Willis.) Education and self selection. *Journal of Political Economy* 87, S7–36.

1981. The economics of superstars. *American Economic Review* 71, 845–58.

1981. ed. *Studies in Labor Markets*. Chicago: NBER, University of Chicago Press.

1981. (With E. Lazear.) Rank-order tournaments as optimal labour contracts. *Journal of Political Economy* 89, 841–64.

1982a. Authority, control and the distribution of earnings. *Bell Journal of Economics* 13, 311–23.

1982b. Further notes on the division of labour and the extent of the market. Working paper. Chicago: University of Chicago.

1983. Specialization and human capital. *Journal of Labor Economics* 1, 43–9.

1985. Implicit contracts: a survey. *Journal of Economic Literature* 23, 1144–75.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1986a. Prizes and incentives in elimination tournaments. *American Economic Review* 76, 701–15.

1986b. The theory of equalizing differences. In *Handbook of Labor Economics*, vol. 1, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

1987. Dynamic animal economics. *American Journal of Agricultural Economics* 69, 547–57.

1988. The value of changes in life expectancy. *Journal of Risk and Uncertainty* 1, 285–304.

1988. (With R. Topel.) Housing investment in the United States. *Journal of Political Economy* 96, 718–40.

1988. (With C. Winship, eds.) Organizations and institutions: sociological and economic approaches to the analysis of social structure. Special issue of *American Journal of Sociology* 94.

1992. The market for lawyers. *Journal of Law and Economics* 35, 215–46.

1994. ed. *Implicit Contract Theory*. London: Edward Elgar.

1994. (With K. Murphy and J. Scheinkman.) Cattle cycles. *Journal of Political Economy* 102, 468–92.

1997a. Manufactured inequality. *Journal of Labor Economics* 15, 189–96.

1997b. Public employment taxes and the welfare state in Sweden. In *The Welfare State in Transition*, ed. R. Freeman, B. Swedenborg and R. Topel. Chicago: NBER, University of Chicago Press.

1998. (With H. Li.) Unraveling in matching markets. *American Economic Review* 88, 371–87.

1999. Potato paradoxes. *Journal of Political Economy* 107, S294–313.

2000. (With D. Neal.) Theories of the distribution of earnings. In *Handbook of Income Distribution*, vol. 1, ed. A. Atkinson, and F. Bourgignon. Amsterdam: North-Holland.

2001. (With H. Li and W. Suen.) Conflicts and common interests in committees. *American Economic Review* 91, 1478–97.

2002. Markets and diversity. *American Economic Review* 92, 1–15.

2005. *Markets and Diversity*. Cambridge, MA: Harvard University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Acemoglu, D. 1998. Why do new technologies complement skills? Directed technical change and wage inequality. *Quarterly Journal of Economics* 113, 1055–89.

Acemoglu, D. 2002. Directed technical change. *Review of Economic Studies* 69, 781–810.

Ashenfelter, O. and Krueger, A. 1994. Estimates of the economic return to schooling from a new sample of twins. *American Economic Review* 84, 1157–73.

Ashenfelter, O. and Rouse, C. 1998. Income, schooling and ability: evidence from a new sample of identical twins. *Quarterly Journal of Economics* 113, 253–84.

Ashenfelter, O. and Zimmerman, D. 1997. Estimates of the returns to schooling from sibling data: fathers, sons, brothers. *Review of Economic Statistics* 79, 1–9.

Baumgardner, J. 1988a. The division of labor, local markets, and worker organizations. *Journal of Political Economy* 96, 509–27.

Baumgardner, J. 1988b. Physicians' services and the division of labor across local markets. *Journal of Political Economy* 96, 948–82.

Becker, G., Murphy, K. and Tamura, R. 1990. Human capital, fertility and economic growth. *Journal of Political Economy* 98, S12–37.

Becker, G., Murphy, K. and Werning, I. 2005. The equilibrium distribution of income and the market for status. *Journal of Political Economy* 113, 282–310.

Benabou, R. and Tirole, J. 2002. Self-confidence and personal motivation. *Quarterly Journal of Economics* 117, 871–915.

Benabou, R. and Tirole, J. 2003. Intrinsic and extrinsic motivation. *Review of Economic Studies* 70, 489–520.

Benabou, R. and Tirole, J. 2004. Willpower and personal rules. *Journal of Political Economy* 112, 848–87.

Ben-Porath, D. 1967. The production of human capital and the life cycle in earnings. *Journal of Political Economy* 75, 352–65.

Bergstrom, T. 1986. Soldiers of fortune. In *Essays in Honor of Kenneth J. Arrow*, vol. 2, ed. W. Heller,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
R. Starr and D. Starrett. New York: Cambridge University Press.

Ekeland, I., Heckman, J. and Nesheim, L. 2004. Identification and estimation of hedonic models. *Journal of Political Economy* 112, S60–109.

Friedman, M. 1953. Choice, chance and the personal distribution of income. *Journal of Political Economy* 61, 277–90.

Hartog, J. 2002. Desperately seeking structure: Sherwin Rosen (1938–2001). *Economic Journal* 112, 519–31.

Jones, L. and Manuelli, R. 1990. A convex model of equilibrium growth: theory and policy implications. *Journal of Political Economy* 98, 1008–38.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Lazear, E. 1995. *Personnel Economics*. Cambridge: MIT Press.

Lazear, E. 2003. Sherwin Rosen. In *Biographical Memoirs of the National Academy of Sciences*, vol. 83. Washington, DC: National Academies Press. Online. Available at http://www.nap.edu/readingroom/books/biomems/srosen.pdf, accessed 10 August 2005.

Lucas, R., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Murphy, K. and Topel, R. 1987. Unemployment, risk, and earnings: testing for equalizing wage differences in the labor market. In *Unemployment and the Structure of Labor Markets*, ed. K. Lang and J. Leonard. New York: Basil Blackwell.

Murphy, K. and Topel, R., eds. 2003. *Measuring the Gains from Medical Research: An Economic Approach*. Chicago: University of Chicago Press.

Rebelo, S. 1991. Long-run policy analysis and long-run growth. *Journal of Political Economy* 99, 500–21.

Roback, J. 1982. Wages, rent and the quality of life. *Journal of Political Economy* 90, 1257–78.

Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Roy, A. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sanderson, A. 2001. Sherwin Rosen, 1938–2001. *Journal of Sports Economics* 2, 211–12.

Siow, A. 1984. Occupational choice under uncertainty. *Econometrica* 52, 631–45.

Tamura, R. 1992. Efficient equilibrium convergence: heterogeneity and growth. *Journal of Economic Theory* 58, 355–76.

Tamura, R. 1994. Fertility, human capital and the ‘wealth of families’. *Economic Theory* 4, 593–603.

Tamura, R. 1996. Regional economies and market integration. *Journal of Economic Dynamics and Control* 20, 825–45.

Tamura, R. 2002. Human capital and the switch from agriculture to industry. *Journal of Economic Dynamics and Control* 27, 207–42.

Tamura, R. 2006. Human capital and economic development. *Journal of Development Economics* 79, 26–72.

Zarkin, G. 1985. Occupational choice: an application to the market for public school teachers. *Quarterly Journal of Economics* 100, 409–46.

How to cite this article

Tamura, Robert. "Rosen, Sherwin (1938–2001)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000235> doi:10.1057/9780230226203.1456

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rosenstein-Rodan, Paul Narcyz (1902 – 1985)

Richard S. Eckaus
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

development economics; disguised unemployment; Rosenstein-Rodan, P.N.

Article

Rodan was one of the founders and first leaders of the field of development economics. His formative intellectual years were in the Austrian School of economics at the University of Vienna. He moved to the Department of Political Economy at University College London, in 1931.
Rodan's early essays in economics show a preoccupation with themes which reappeared throughout his professional career: the interaction and complementarity of economic processes (1933) and their temporal patterns (1934). Rodan's seminal article on developing countries (1943) argued that complementarities and externalities in demand and production created a need for the programming of investment. The arguments were subsequently extended to justify the need for an across-the-board ‘big push’ for a successful start to the development process (1963). He was among the first to apply the concept of ‘disguised unemployment’, described by Joan Robinson (1936), to developing countries as a persisting rather than cyclical problem.
Rodan first became actively engaged in development policy during his tenure at the World Bank from 1947 to 1954. In 1954 he moved to the Department of Economics at the Massachusetts Institute of Technology, where he produced an influential article (1961) which demonstrated that feasible levels of assistance to developing countries would substantially improve their growth performance. After retirement from MIT in 1968 he moved to the University of Texas and then to Boston University in 1972, where he established and worked in the Center for Latin American Development Studies until his death. Rodan was an active policy adviser to international agencies and governments of many countries and served on the Panel of Experts, the ‘Nine Wise Men’ of the Alliance for Progress, from 1961 to 1966.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1933. La complementarita, prime delle tre fase del progresso della teoria economica pura. *Riforma Sociale* 44, 257–308.

1934. The role of time in economic theory. *Economica* NS 1, 77–97.

1943. Problems of industrialization of eastern and south-eastern Europe. *Economic Journal* 53, 202–11.

1957. Disguised unemployment and underemployment in agriculture. *Monthly Bulletin of Agricultural Economics and Statistics* 6(7–8), 1–6.

1961. International aid for underdeveloped countries. *Review of Economics and Statistics* 43, 107–38.

1963. Notes on the theory of the ‘big push’ in economic development. In *Proceedings of a Conference of the International Economics Association*, ed. H.S. Ellis. London: Macmillan.

Bibliography

Robinson, J. 1936. Disguised unemployment. *Economic Journal* 46, 225–37.

How to cite this article

Eckaus, Richard S. "Rosenstein-Rodan, Paul Narcyz (1902–1985)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000174> doi:10.1057/9780230226203.1457

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rosenthal, Robert W. (1944 – 2002)

Barton L. Lipman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

backward induction; centipede game; iterated dominance; mixed strategy equilibria; multi-unit auctions; quantal response equilibrium; rationality; repeated games; revelation principle; Rosenthal, R. W.; social norms; sovereign debt

Article

Robert W. Rosenthal (1944–2002) was an economic theorist whose thoughtful papers inspired a wide range of new ideas. As Radner and Ray (2003) point out, Rosenthal (1978) gives one of the first formal statements of the revelation principle, a result noted in Myerson's first paper (1979) on the subject. Rosenthal (1979) initiated the study of repeated games with varying opponents, a modelling device used by Milgrom, North and Weingast (1990), Kandori (1992), and others to study social norms and other issues. He also wrote influential papers on pricing (Rosenthal, 1980; 1982), multi-unit auctions (Krishna and Rosenthal, 1996), purification of mixed strategy equilibria (Radner and Rosenthal, 1982; Aumann et al., 1983), sovereign debt (Fernandez and Rosenthal, 1990), analysis of experimental data (Brown and Rosenthal, 1990), and many other topics.

He is arguably best-known for his 1981 *Journal of Economic Theory* paper in which he discussed what Binmore (1987) named the ‘centipede game’. Like its older cousin, the Prisoner's Dilemma, the centipede game beautifully summarizes a fundamental and intriguing strategic problem. Like the game which inspired but was overshadowed by it, Selten's (1978) chain store paradox, the centipede calls into question one of the most basic principles of game theory, namely, backward induction.

Consider the game shown in Figure 1. In this game, backward induction predicts that 1 plays *A* and 2 plays *D*. The reasoning seems very compelling. If 2 is rational, then, faced with a choice between a payoff of 3 and a payoff of 4, he obviously chooses 4. Hence 2 will play *D*. If 1 knows that 2 is rational, 1 knows that 2 will play *D*. Hence, if 1 is rational, he will choose *A* to get 4 instead of playing *B* which would yield 3. Thus the hypothesis that each player is rational and knows the other is rational seems to predict the backward induction solution. In longer games, there will be longer chains of reasoning, of course. However, the reasoning above has led many to conclude that backward induction is the implication of rationality and common knowledge of rationality.

Figure 1

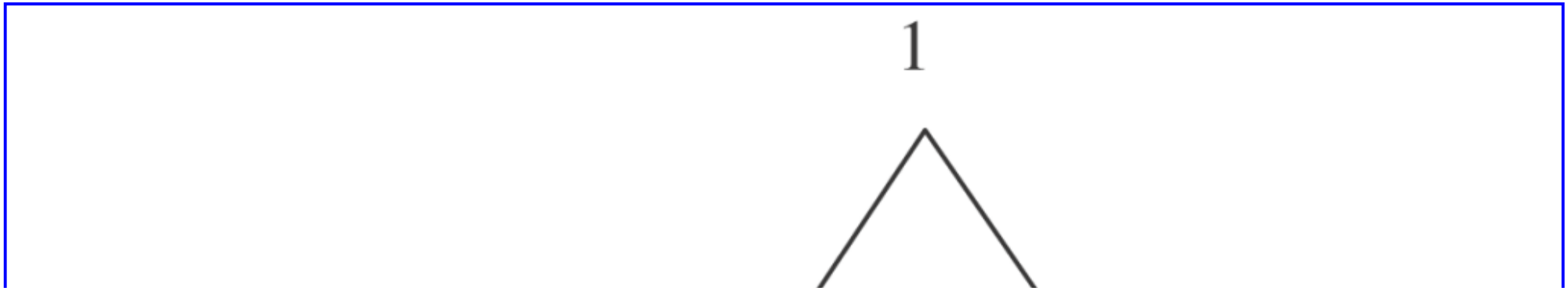

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

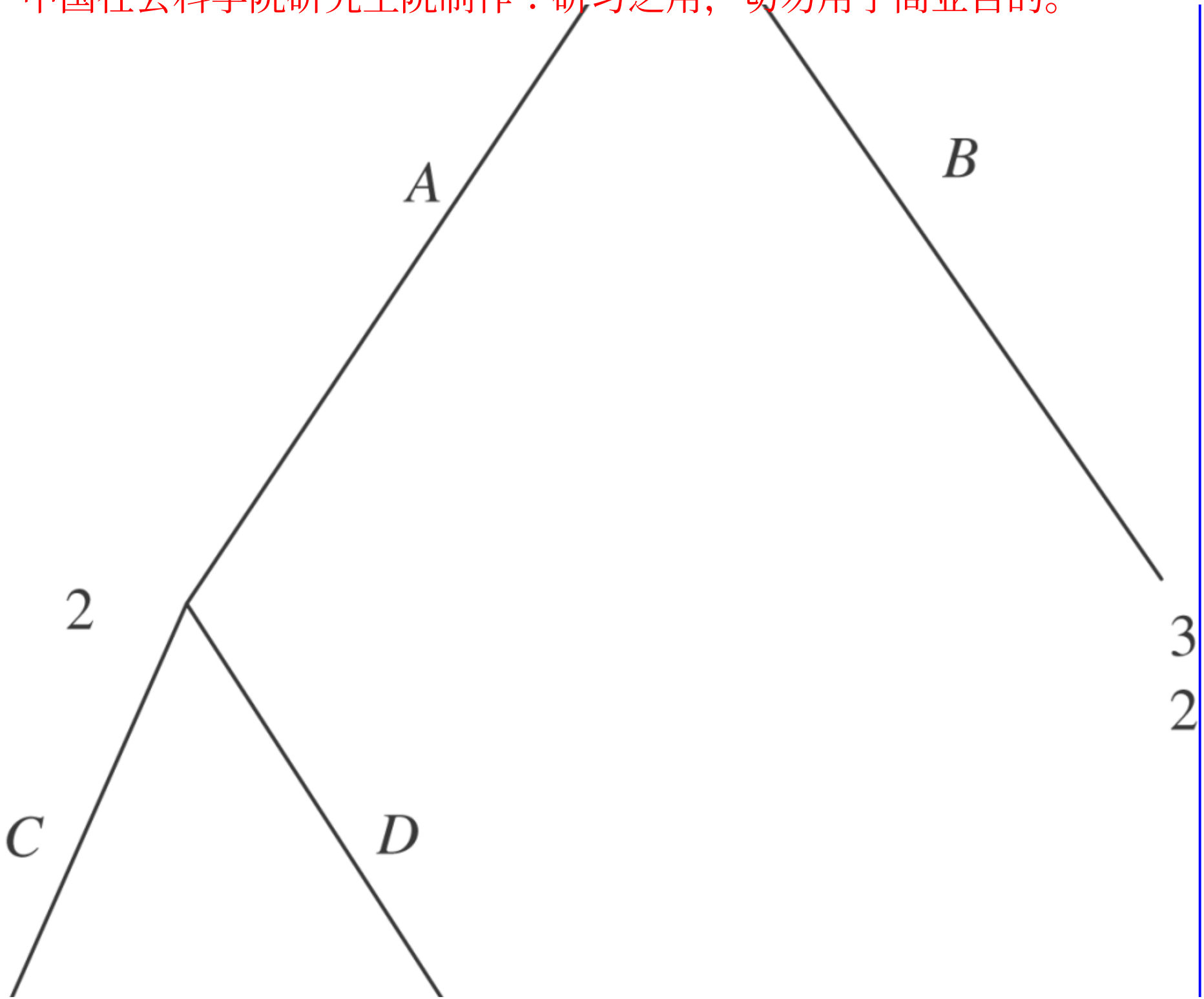

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

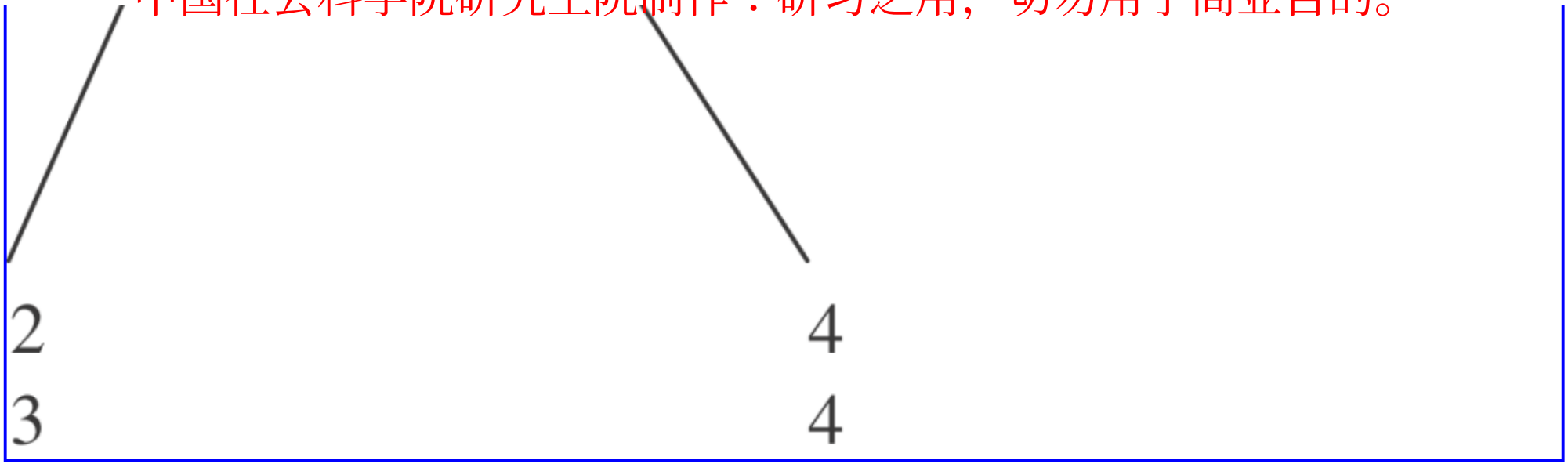

A version of Rosenthal's centipede is shown in Figure 2. Here backward induction predicts that 1 chooses *d* at his first choice node, ending the game right away. Now the reasoning seems more suspect. If 2 is rational, he should choose *D* at the end rather than *A*. If 1 anticipates this, he should choose *d* at his last decision node. Similar reasoning shows that 2 should choose *D* at his first decision node and that 1 should choose *d* at his first node. Yet it is clear that each player must be virtually certain about his opponent's choice at the next move to justify choosing down rather than across, a certainty that seems extremely implausible in practice.

Figure 2

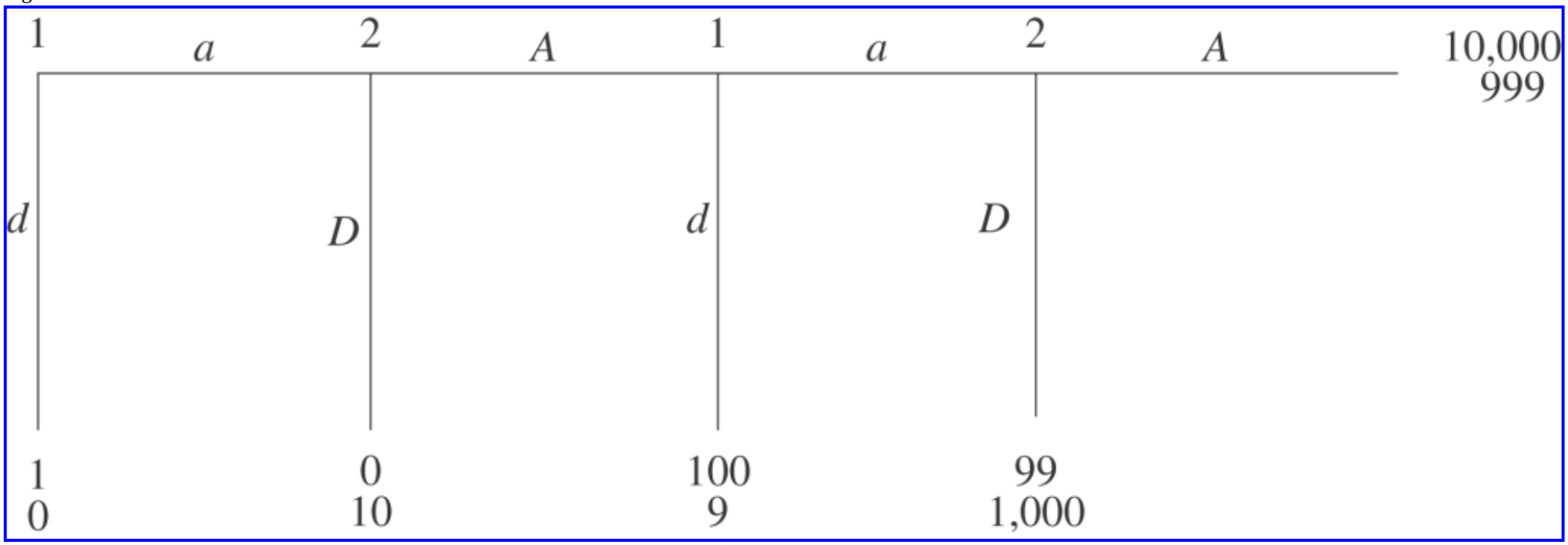

Many writers have argued that Rosenthal's centipede shows the paradoxical nature of backward induction. Consider player 1 at his second decision node. Here he is supposed to be certain that player 2 will choose *D* at the following node, justifying his own choice of *d*. Yet he also knows that 2 should have chosen *D* at the previous node and did not. If 2 failed to be rational in the past, why should 1 remain confident that he will be rational in the future? If 1 does have doubts, perhaps he should play *a* — a move which would make 2 glad to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

have deviated from 'rationality' at the preceding move!

Rosenthal's work led to a major debate on the question of backward induction. See, for example, Aumann (1995; 1998), Binmore (1987; 1996) and Reny (1993). See also Glazer and Rosenthal (1992) for a conceptually related critique of the use of iterated dominance in implementation theory.

Interestingly, McKelvey and Palfrey's (1992) experiments with the centipede game led them to develop the notion of quantal response equilibrium (McKelvey and Palfrey, 1995), an idea which echoes Rosenthal's own analysis. Rosenthal suggested that players might make 'mistakes' where these mistake probabilities would be decreasing in the cost of the mistakes, an idea he explored further in Rosenthal (1989). Quantal response equilibrium is another formulation of this idea.

See Also

- behavioural economics and game theory
- epistemic game theory: an overview
- quantal response equilibria
- rationality, bounded

I thank Hseuh Ling Huynh, Andy Postlewaite and Debraj Ray for comments.

Selected works

1978. Arbitration of two-party disputes under uncertainty. *Review of Economic Studies* 45, 595–604.

1979. Sequences of games with varying opponents. *Econometrica* 47, 1353–66.

1980. A model in which an increase in the number of sellers leads to a higher price. *Econometrica* 48, 1575–9.

1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* 25, 92–100.

1982. A dynamic model of duopoly with customer loyalties. *Journal of Economic Theory* 27, 69–76.

1982. (With R. Radner.) Private information and pure-strategy equilibria. *Mathematics of Operations Research* 7, 401–9.

1983. (With R. Aumann, Y. Katznelson and R. Radner.) Approximate purification of mixed strategies. *Mathematics of Operations Research* 8, 327–41.

1989. A bounded-rationality approach to the study of noncooperative games. *International Journal of Game Theory* 18, 273–92.

1990. (With J. Brown.) Testing the minimax hypothesis: a re-examination of O'Neill's game experiment. *Econometrica* 58, 1065–81.

1990. (With R. Fernandez.) Strategic models of sovereign-debt renegotiations. *Review of Economic Studies* 57, 331–49.

1992. (With J. Glazer.) A note on Abreu–Matsushima mechanisms. *Econometrica* 60, 1435–8.

1996. (With V. Krishna.) Simultaneous auctions with synergies. *Games and Economic Behavior* 17, 1–31.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Aumann, R. 1995. Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8, 6–19.

Aumann, R. 1998. On the centipede game. *Games and Economic Behavior* 23, 97–105.

Binmore, K. 1987. Modeling rationality players, I. *Economics and Philosophy* 3, 9–55.

Binmore, K. 1996. A note on backward induction. *Games and Economic Behavior* 17, 135–7.

Kandori, M. 1992. Social norms and community enforcement. *Review of Economic Studies* 59, 81–92.

McKelvey, R. and Palfrey, T. 1992. An experimental study of the centipede game. *Econometrica* 60, 803–36.

McKelvey, R. and Palfrey, T. 1995. Quantal response equilibria in normal form games. *Games and Economic Behavior* 10, 6–38.

Milgrom, P., North, D. and Weingast, B. 1990. The role of institutions in the revival of trade: the law merchant, private judges, and the champagne fairs. *Economics and Politics* 2, 1–23.

Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47, 61–73.

Radner, R. and Ray, D. 2003. Robert W. Rosenthal. *Journal of Economic Theory* 112, 365–8.

Reny, P. 1993. Common belief and the theory of games with perfect information. *Journal of Economic Theory* 59, 257–74.

Selten, R. 1978. The chain-store paradox. *Theory and Decision* 9, 127–59.

How to cite this article

Lipman, Barton L. "Rosenthal, Robert W. (1944–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000258> doi:10.1057/9780230226203.1458

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rostow, Walt Whitman (1916 – 2003)

Douglas Calvin Dacy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

development economics; Harrod–Domar growth model; International Economic Association; Rostow, W.

Article

Walt Whitman Rostow, economic historian, historian of economic thought, pioneer of modern development economics, and social scientist with interests in demography, politics, sociology and cultural aspects of development, was born in 1916. A professor of economics and history at the University of Oxford in 1946–7, Cambridge University, 1949–50, Massachusetts Institute of Technology 1950–61 and University of Texas at Austin 1968–2003, he is best known for his *Stages of Economic Growth: A Noncommunist Manifesto* (1960) and for his service as National Security Advisor to US President Lyndon B. Johnson during the Vietnam War. He led an active intellectual life engaged in public policy issues up to his death in 2003.

Several themes developed in his first publication, ‘Investment and the Great Depression’ (1938), recur in his first book, *Essays on the British Economy in the Nineteenth Century* (1948), and his *Process of Economic Development* (1953). His book co-authored with A.D. Gayer and Anna J. Schwartz, *The Growth and Fluctuation of the British Economy* (1952), was considered a classic study, and his work co-authored with Max Millikan, *A Proposal: Key to an Effective Foreign Policy* (1957), made his reputation in the field of foreign policy. These books established Rostow as one of the world's foremost economic historians of his age.

His *Stages of Economic Growth* was a blockbuster. It stepped on many toes, assuring his reputation as the one of the most controversial economists of the last half of the 20th century. At the time, his model clashed with that of Harrod–Domar. They modelled steady-state (equilibrium) growth, with no historical context, and focused on two variables: saving and output–capital ratios. Naturally, an economic historian would ask how an economy got there in the first place. Rostow though he saw a pattern in how countries got there. Development proceeded through five stages: traditional society, preconditions for take-off, take-off to sustained growth, drive to maturity and age of high mass consumption. His critics saw these

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
stages as ‘empty boxes’, not empirically verifiable and devoid of predictive power. Most were especially critical of the take-off stage. He had not demonstrated empirically the necessity of a significant rise in the saving and output–capital ratios. His critics were not convinced about his dating of stages for the seven countries he studied. Besides, he had not heeded Marshall's dictum, *Natura non facit saltum* – nature does nothing in jumps. Controversy swirled over his discontinuous, disequilibrium approach to economic growth.

His work was so upsetting to many of the world's most distinguished researchers in the field of development that the International Economic Association convened a conference in Konstanz in 1960 devoted exclusively to Rostow's work. This exclusivity was a first for the Association, and is indicative of the importance placed on his work. If Rostow did not convince his critics, or they him, the conference gave him worldwide notoriety, and his ideas were embraced by many economists in developing countries. Twenty years later the controversy continued.

Seizing on earlier criticism, Rostow published a massive volume, *The World Economy: History and Prospect* (1978). It examines world economic history from 1790 to 1976 in terms of population dynamics, long-term trends, cyclical fluctuations in production, prices and international trade. It extends the work of *Stages* with later data, and expands coverage to 20 countries.

In 1982 Charles P. Kindleberger and Guido Di Tella edited a three-volume Festschrift in Rostow's honour. A reviewer, Mancur Olson (1985), noted a paradox: many of the contributors were critics, and in a Festschrift! He pondered over an interesting question: how can so many distinguished critics also be admirers? Henry Rosovsky (1965) probably had the right explanation in an earlier comment: ‘I invariably learn more by disagreeing with Professor Rostow than I do by agreeing with most other writers.’

Among economists with roots in the 1960s, Rostow's visible positions in the US government made him the most influential. He helped to form the Alliance for Progress and was President John F. Kennedy's representative on it. As an architect of the Vietnam War and President Johnson's National Security Advisor, he became controversial in the political arena. He knew many of the world's leaders and was known by most of them. Through his public service, he became the Keynes of his day.

He continued to write books on important issues such as East–West relations, verification of nuclear arsenals, foreign aid and world population problems. He died in 2003 at age 87 just before his final book, *Concept and Controversy* (2003), was published. For many years, Rostow's ideas energized the field of economic development. That, even alone, is a major contribution to economics.

See Also

- growth and cycles
- Kondratieff cycles

Selected works

1938. Investment and the Great Depression. *Economic History Review* 8, 136–58.

1948. *Essays on the British Economy of the Nineteenth Century*. London: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1952. (With A. Gayer and A. Schwartz.) *The Growth and Fluctuation of the British Economy*. Oxford: Oxford University Press.

1953. *The Process of Economic Development*. New York: W.W. Norton.

1957. (With M. Millikan.) *A Proposal: Key to an Effective Foreign Policy*. New York: Harper.

1960. *The Stages of Economic Growth: A Noncommunist Manifesto*. Cambridge: Cambridge University Press.

1960. *The United States in the World Arena*. New York: Harper.

1963. *The Economics of Take-Off into Sustained Growth*. London: Macmillan.

1975. *How It All Began: Origins of the Modern Economy*. New York: McGraw-Hill.

1978. *The World Economy: History and Prospect*. Austin: University of Texas Press.

1990. *Theorists of Economic Growth from David Hume to the Present*. New York: Oxford University Press.

1998. *The Great Population Spike and After*. New York: Oxford University Press.

2003. *Concept and Controversy: Sixty Years of Taking Ideas to Market*. Austin: University of Texas Press.

Bibliography

Kindleberger, C. and di Tella, G. 1982. *Economics in the Long View: Essays in Honour of W.W. Rostow*. London: Macmillan.

Olson, M. 1985. Review of C. Kindleberger and G. di Tella, eds., *Economics in the Long View: Essays in Honor of W.W. Rostow*. *Journal of Economic Literature* 23, 622–5.

Rosovsky, H. 1965. The take-off into sustained controversy. *Journal of Economic History* 25, 271–5.

How to cite this article

Dacy, Douglas Calvin. "Rostow, Walt Whitman (1916–2003)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000270> doi:10.1057/9780230226203.1459

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

rotating saving and credit associations (ROSCAs)

Jean-Marie Baland
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

rotating saving and credit associations (roscas); saving; social sanctions

Article

Rotating saving and credit associations (roscas) are the simplest form of collective financial institution. A rosca is a group of individuals who meet at regular intervals, each of whom contributes at each meeting a pre-determined amount to a collective ‘pot’ which is then given to one member. The latter is then excluded from receiving the pot in future meetings, while still being obliged to contribute to the pot. The meeting process repeats itself until each member has received the pot, thereby completing a cycle. Then the rosca can start a new cycle. From this description, the main virtues of roscas are clear: they do not require storage of funds, accounting and durations of obligations are transparent, and there are no complicated interest payments or debt management. Roscas are very popular in developing countries. For instance, average membership rates in Indonesia have been estimated at 40 per cent of the population (Armendariz de Aghion and Morduch, 2005), 20 per cent in Taiwan (Levenson and Besley, 1996) and 40 per cent in a Kenyan slum (Anderson and Baland, 2002). Although roscas do exist alongside more formal financial institutions, they are often the sole saving and credit institution in many rural areas.

Roscas vary widely in terms of the size of the contributions, the number of members and the frequency of meetings. Also, the process by which the pot is allocated can be a lottery (random roscas), or follow a fixed order imposed, for instance, by the leaders in the group (fixed roscas), or be determined by a bidding process (bidding roscas).

The literature identified four differing motives for individuals to save through roscas. First, roscas allow individuals to purchase indivisible goods earlier in expected terms than through the accumulation of individual savings (Besley, Coate and Loury, 1993). Roscas thus provide an implicit positive interest rate to those receiving the pot early. Second, as emphasized by Anderson and Baland (2002), roscas may be used by married women as a way to commit the household to higher saving rates than what can be done at home. Given the presence of social sanctions, husbands are then forced to comply with the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

saving rate imposed by the rosca. Relatedly, Gugerty (2006) argues that people facing intertemporal inconsistency join roscas to bind themselves to a particular saving pattern (see also Ardener, 1964; Ambec and Treich, 2007). Lastly, bidding roscas provide some insurance to their members against short-term income shocks by providing implicit short-run credit to those willing to pay the highest bid.

The most common problem of roscas has to do with enforcement. Indeed, the first members to receive the pot are de facto borrowers from the other members and, absent social sanctions, are better off not repaying their debts. Given that the size of the pot is fixed, they can always replicate (and also do better than) the best that the rosca can offer them by saving on their own. Social sanctions are then necessary to discipline members. Also, as argued in Anderson, Baland and Moene (2003), the rule for allocating the pot can be chosen to partially address this issue. Selection of members is another issue faced by roscas, particularly bidding roscas, where higher bidders may be exposed to more intrinsic risks than others (see Eeckhoudt and Munshi, 2005; Klonner and Rai, 2006).

See Also

- micro-credit

Bibliography

Ambec, S. and Treich, N. 2007. Roscas as financial agreements to cope with self-control problems. *Journal of Development Economics* 82, 120–37.

Anderson, S. and Baland, J.-M. 2002. The economics of ROSCAs and intrahousehold allocation. *Quarterly Journal of Economics* 117, 963–95.

Anderson, S., Baland, J.-M. and Moene, K. 2003. Sustainability and organizational design in Roscas: some evidence from Kenya. Mimeo, Center for Research in the Economics of Development (CRED), University of Namur.

Ardener, S. 1964. The comparative study of rotating credit associations. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 44, 202–29.

Armendariz de Aghion, B. and Morduch, J. 2005. *The Economics of Microfinance*. Cambridge, MA: MIT Press.

Besley, T., Coate, S. and Loury, G. 1993. The economics of rotating savings and credit associations. *American Economic Review* 83, 792–810.

Eeckhoudt, J. and Munshi, K. 2005. Mitigating regulatory inefficiency: the non-market response to financial regulation in India. Mimeo, Brown University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gugerty, M. 2006. You can't save alone: commitment and rotating savings and credit associations in Kenya. Working Paper No. 2006-08, Evans School of Public Affairs, University of Washington.

Klonner, S. and Rai, A. 2006. Adverse selection in credit markets: evidence from bidding roscas. Mimeo, Cornell University.

Levenson, A. and Besley, T. 1996. The anatomy of an informal financial market: rosca participation in Taiwan. *Journal of Development Economics* 51, 45–68.

How to cite this article

Baland, Jean-Marie. "rotating saving and credit associations (ROSCAs)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000234> doi:10.1057/9780230226203.1460

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rothbard, Murray N. (1926 – 1995)

Benjamin Powell and Edward Stringham
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

anarchism; antitrust enforcement; apriorism; Austrian economics; cartels; economic history; economics, definition of; Great Depression; Hayek, F. A. von; Hazlitt, H.; interpersonal utility comparisons; libertarianism; market failure; Mises, L. E. von; monetary policy; monopolies; perfect competition; private enforcement of law; public goods; Rand, A.; Rothbard, M. N.; welfare economics

Article

Murray Rothbard was influential in continuing the tradition of the Austrian school of economics in America. In more than two dozen books and hundreds of articles, his work spanned economics, history, philosophy and political science. He earned his Ph.D. from Columbia University, but was influenced mainly by Ludwig Von Mises' seminar at New York University. Rothbard was a strong believer in *apriorism*, the idea that economic laws could be discovered using logical reasoning (as opposed to empirical testing), and he attempted to build on and extend the economic logic of Mises and others in that tradition. Rothbard's treatise, *Man, Economy, and State* (1962), analysed the economics of market exchange, while his follow-up volume, *Power and Market* (1970), analysed the economics of government intervention. An underlying theme of his work is that the market is the realm of mutually beneficial exchange, whereas the government is the realm of coercion where some gain at the expense of others.

Rothbard considered economics to be a value-free science, but he believed economic reasoning can be used to determine whether normative views are internally consistent. He was strongly critical of government intervention in the economy, arguing against those who believe that government policies can be Pareto-superior and make all people better off. For example, Rothbard was one of the only economists writing in the 1950s and 1960s to argue against all antitrust laws. He thought that perfect competition was an unattainable ideal, and he said that monopolies or cartels do not pose problems on the free market. He believed that the only monopolies that warrant concern are those sanctioned by government. Rothbard was critical of arguments about market failure in general, insisting that mainstream notions of economic efficiency were unrealistic. He criticized the welfare economics of his

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
day on the grounds that it rests on unscientific interpersonal comparisons of utility.
In addition, Rothbard wrote a great deal on economic history, often documenting government getting in the way of markets. For example, his 1963 book *America's Great Depression* argued that government caused and lengthened the Great Depression through distortionary monetary and regulatory policies. Rothbard also devoted much of his writing to political philosophy, and here too he was unabashedly libertarian. Rothbard's contribution is particularly noteworthy because he was one of the first economists to argue that markets do not depend on the existence of government. Before him, even the most free-market theorists, such as Ludwig von Mises, Henry Hazlitt, Ayn Rand, and Friedrich Hayek, had simply assumed that services like law enforcement must be provided collectively by the state. But in *Power and Market* and *For a New Liberty* (1973) Rothbard maintained that public goods such as law enforcement must be analysed in terms of marginal units and, as with other goods, those marginal units can be provided privately. He pointed to historical examples of private law enforcement and speculated how a purely private system might function. Rothbard's ideas advancing private property anarchism were radical, but they influenced many economists who now write about alternatives to government law (Stringham, 2006). Rothbard's thorough libertarian views pushed free-market thinking to become more free market.

See Also

- Austrian economics
- Hayek, Friedrich August von
- libertarianism
- Menger, Carl
- Mises, Ludwig Edler von

Selected works

1962. *Man, Economy, and State: A Treatise on Economic Principles*. Princeton: Van Nostrand.

1963. *America's Great Depression*. Princeton: Van Nostrand.

1970. *Power and Market*. Kansas City: Sheed Andrews and McMeel.

1973. *For a New Liberty: The Libertarian Manifesto*. New York: Macmillan.

Many of Rothbard's writings are freely available online at http://www.rothbard.org, accessed 26 September 2006.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Raimondo, J. 2000. *An Enemy of the State: The Life of Murray N. Rothbard*. Amherst, NY: Prometheus Books.

Stringham, E., ed. 2006. *Anarchy and the Law: The Political Economy of Choice*. Somerset, NJ: Transaction Publishers.

How to cite this article

Powell, Benjamin and Edward Stringham. "Rothbard, Murray N. (1926–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000257> doi:10.1057/9780230226203.1461

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

rotten kid theorem

Theodore C. Bergstrom
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The rotten kid theorem states that, if a household head is sufficiently rich and benevolent towards other household members, then it is in the self-interest of other household members to take those actions that maximize the total income of the household, even at a cost to their own private income. This theorem holds under certain restrictive assumptions, but the assumptions needed for it to be true are not satisfied in many common family decision-making environments.

Keywords

Becker, G.; conditional transferable utility; Dawkins, R.; family economics; kin selection; leisure; normal goods; rotten kid theorem

Article

$u_i(x_i) = x_i$, and the utility function of the household head, $u(x_0, \ldots, x_n)$, is strictly increasing in all the x_i's. Every household member earns some personal income, the amount of which depends on her own actions a_i, but possibly also on the actions of other household members. Let a be the vector of actions chosen by household members, let $m_i(a)$ be i's personal income, and let $m(a) = \Sigma m_i(a)$. Feasible allocations must satisfy the household budget constraint, $\Sigma_i x_i = m(a)$. For any income y, define $(x_0(y), \ldots x_n(y))$ as the allocation that maximizes $u(x_0, \ldots, x_n)$ subject to $\Sigma_i x_i = y$. Assume that consumption for each i is a *normal good* so that $x_i(y)$ is a strictly increasing function of y. Finally, assume that the household head has personal income large enough so that in equilibrium he chooses to donate money to all other persons in the household. This means that, for all feasible a and for each kid, i, $x_i(m(a)) > m_i(a)$. Consider the following two-stage game. In the first stage, household members choose their actions and thus determine total family income $m(a)$. In the second stage the household head finds the allocation $x(m(a))$ that maximizes $u(x_1, \ldots, x_n)$ subject to $\Sigma_i x_i = m(a)$ and donates

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
$x_i(m(a)) - m_i(a)$ to kid *i*. In the first stage of the game, each kid realizes that, after the head has redistributed income, her own consumption will be $x_i(m(a))$. The normal goods assumption implies that $x_i(m(a))$ is an increasing function of *m*(*a*). Therefore, the self-interest of each kid coincides with maximizing total family income, *m*(*a*). (To ensure that a maximum exists, assume that each m_i is continuous and that each a_i must be chosen from a closed bounded set.)

The trouble with the rotten kid theorem is that it fails to hold in models that make slight concessions toward realism. Bergstrom (1989) shows that, in general, the rotten kid theorem fails if kids care about their activities as well as about consumption. For example, if leisure is a complement to consumption, a child can manipulate the parents' transfer in his or her favour by taking too much leisure. Lindbeck and Weibull (1988) and Bruce and Waldman (1990) show that the rotten kid theorem fails when individuals can choose between current and future consumption. Lundberg and Pollak (2003) show a dramatic failure of the rotten kid theorem when families choose between discrete options like whether to move house or whether to have a child.

Bergstrom (1989) explored the most general conditions under which a rotten kid theorem can be proved. He showed that, in general, a necessary and sufficient condition for the conclusion of the rotten kid theorem to be satisfied is that there is ‘conditional transferable utility’. This means that the utility possibility sets corresponding to all possible activity choices are nested and are bounded above by parallel straight line segments. For example, there is conditional transferable utility if kids care only about their consumption, so that $u_i(x_i, a) = x_i$, and if total family income is *m*(*a*). Then the utility possibility frontier conditional on *a* is the simplex $\{(u_1, \ldots, u_n) | \Sigma_1^n u_i \leq m(a) \text{ and } u_i \geq 0 \text{ for all } i\}$. In general, however, if the kids' utilities depend on their actions, kids will be able to influence the ‘slope’ of the utility possibility frontier by their choice of actions, *a*. For example, a selfish kid may benefit by choosing an action that reduces family income but makes it ‘cheaper’ for the parent to invest in her utility rather than that of her sibling. Bergstrom shows that the most general class of environments for which there is conditional transferable utility requires that each kid *i* has a utility function of the form $u(x_i, a) = A(a)x_i + B_i(a)$ where x_i is *i*'s expenditure on consumer goods and *a* is the vector of family members' activities. This allows the possibility that activities a_i generate externalities in consumption as well as in income-earning. (Bergstrom and Cornes, 1983, show that in a public goods economy the efficient quantity of public goods is independent of income distribution if and only if preferences can be represented in this form, which is dual to the Gorman polar form for public goods.) Then, for any *a*, the upper boundary of the utility possibility set is $\{u | \Sigma u_i = A(a)m(a) + \Sigma B_i(a)\}$. If utilities of kids are normal goods for the head, then each kid will maximize her utility by maximizing $F(a) = A(a)m(a) + \Sigma B_i(a)$. Thus selfish kids would act in the family interest, as the rotten kid theorem asserts.

An interesting debate in evolutionary biology parallels the economists' rotten kid theorem. Alexander (1974) maintained that natural selection favours genetic lines in which offspring act so as to maximize family reproductive success. Dawkins (1976) disputed Alexander's argument, citing Hamilton's theory of kin selection (1964), which implies that in sexual diploid species offspring value the reproductive success of their siblings at only half of their own. Alexander (1979) conceded Dawkins's point, but offered an additional reason that offspring would act in the interest of their parents, namely, that ‘the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
parent is bigger and stronger than the offspring, hence in a better position to pose its will'. Bergstrom and Bergstrom (1999) propose an evolutionary model that could support the Becker–Alexander conclusion that children will act in the family interest. They construct a two-locus genetic model, where a gene at one locus controls an animal's behaviour when the animal is a juvenile and a gene at the other controls its behaviour when it is a parent. Then the frequency of recombination between genes at these two loci determines the evolutionary outcome of parent–offspring conflict. If recombination between these genes is rare, offspring will tend to act in the genetic interest of their parent. If recombination is frequent, there can be an equilibrium where some offspring successfully 'blackmail' their parents into giving them more resources than is optimal for the family's reproduction.

See Also

- Becker, Gary S.
- family economics

Bibliography

Alexander, R. 1979. *Darwinism and Human Affairs*. Seattle: University of Washington Press.

Alexander, R. 1974. The evolution of social behavior. *Annual Review of Ecological Systems* 5, 25–83.

Becker, G. 1974. A theory of social interactions. *Journal of Political Economy* 82, 1063–93.

Becker, G. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Bergstrom, T. 1989. A fresh look at the rotten kid theorem. *Journal of Political Economy* 97, 1138–59.

Bergstrom, C. and Bergstrom, T. 1999. Does Mother Nature punish rotten kids? *Journal of Bioeconomics* 1, 47–72.

Bergstrom, T. and Cornes, R. 1983. Independence of allocative efficiency from distribution in the theory of public goods. *Econometrica* 51, 1753–65.

Bruce, N. and Waldman, M. 1990. The rotten-kid theorem meets the Samaritan's Dilemma. *Quarterly Journal of Economics* 105, 155–65.

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Hamilton, W. 1964. The genetical evolution of social behavior, I and II. *Journal of Theoretical Biology* 7, 1–52.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lindbeck, A. and Weibull, J. 1988. Altruism and efficiency: the economics of fait accompli. *Journal of Political Economy* 96, 1165–82.

Lundberg, S. and Pollak, R. 2003. Efficiency in marriage. *Review of the Economics of the Household* 1, 153–67.

How to cite this article

Bergstrom, Theodore C. "rotten kid theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000239> doi:10.1057/9780230226203.1462

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Roy model

James J. Heckman and Christopher Taber
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Roy (1951) model of self-selection on outcomes is one of the most important models in economics. It is a framework for analysing comparative advantage. The original model analysed occupational choice with heterogeneous skill levels and has subsequently been applied in many other contexts. This article presents the model, discusses its identification, and describes some empirical applications based on the model.

Keywords

comparative advantage; discrete choice; estimation; identification; occupational choice; Roy model; selection bias; self-selection; separability

Article

The Roy (1951) model of self-selection on outcomes is one of the most important models in economics. It is a framework for analysing comparative advantage. The original model analysed occupational choice with heterogeneous skill levels and has subsequently been applied in many other contexts. We first discuss the model. We then summarize what is known about identification of the model. We end by describing some applications based on the model and its extensions.

Basic models

In the original Roy (1951) model, agents can pursue one of two possible occupations: hunting and fishing. They cannot pursue both at the same time. There is no interaction among agents so the choice of one agent does not affect the choice of another agent either through prices or through external effects. Let π_f and π_r be the price of fish and rabbits respectively in the village. Let F_i denote the number of fish that individual i would catch if he chooses to fish. Similarly let R_i denote the number rabbits he would catch. Then individual i's wage is

$$w_{fi} = \pi_f F_i$$

if he fishes and

$$w_{ri} = \pi_r R_i$$

if he hunts. The income that worker i receives for working in sector j is thus proportional to π_j (where $j \in \{r, f\}$). If workers are pure income maximizers, they will choose the occupation with higher income. Thus a worker chooses to fish if $w_{fi} > w_{ri}$. If F_i and R_i are continuous random variables, $\Pr(\pi_r R_i = \pi_f F_i) = 0$, so the indifference set is negligible. A fundamental aspect of the Roy model is that it allows for heterogeneity in (F_i, R_i). This heterogeneity can arise from inherent ability differences or human capital investment.
An important issue is self-selection. Under what conditions will the best workers self-select into an occupation? Will people who self-select be above average? For example, for fishing, under what conditions is the average productivity of people working in the fishing sector above the population mean productivity:

$$E[\log(F_i) | \pi_f F_i \geq \pi_r R_i] > E[\log(F_i)] \ ?$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Assume, as did Roy (1951), that log skills are jointly normally distributed

$$\begin{bmatrix} \log(F_i) \\ \log(R_i) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_f \\ \mu_r \end{bmatrix}, \begin{bmatrix} \sigma_{ff} & \sigma_{fr} \\ \sigma_{fr} & \sigma_{rr} \end{bmatrix} \right).$$

Then it is straightforward to show that

$$E[\log(F_i) | \pi_f F_i \geq \pi_r R_i] = \mu_f + \underbrace{\frac{(\sigma_{ff} - \sigma_{fr})}{\sigma} \lambda\left(\frac{\log(\pi_f) - \log(\pi_r) + \mu_f - \mu_r}{\sigma} \right)}_{\textit{selection effect}}$$

where σ^2 is the variance of $\log(F_i/R_i)$ and $\lambda(\cdot)$ is the inverse Mills ratio. (See selection bias and self-selection.)

The function λ is positive but decreasing in its arguments with $\lim_{\pi_f \to \infty} \lambda(\cdot) = 0$. The selection effect is the second term on the right-hand side of this expression. There is a parallel expression for $E(\log R_i | \pi_F F_i \leq \pi_R R_i)$ with the subscripts *f* and *r* interchanged.

Recall that $E[\log(F)] = \mu_f$ and that λ and σ must both be positive. Therefore, the question of whether there is positive selection into fishing depends only upon the sign of $\sigma_{ff} - \sigma_{fr}$. It does not depend on skill prices. Moreover, since

$$\sigma^2 = (\sigma_{ff} - \sigma_{fr}) + (\sigma_{rr} - \sigma_{fr}) > 0,$$

at least one of $(\sigma_{ff} - \sigma_{fr})$ and $(\sigma_{rr} - \sigma_{fr})$ must be positive. Thus, there must be positive selection into one of the occupations, and there can be positive selection into both.

If, however, there is positive selection into only one occupation, the question arises as to which occupation is most likely to have positive selection. Roy argues that relatively simple tasks (setting traps for rabbits in his case) can be described by a small standard deviation of skill. For more difficult skills (fishing in his example) the standard deviation will be relatively higher as there is a bigger difference between the most skilled and the least skilled. Thus, if fishing is the more difficult task, $\sigma_{ff} > \sigma_{rr}$, there must be positive selection into fishing (that is, $E(\log(F_i) | \pi_f F_i \geq \pi_r R_i) > E(\log(F_i))$).

Whether there is positive selection into hunting depends on the value of σ_{fr} relative to σ_{rr}. When $\sigma_{fr}<0$, we will see positive selection into hunting. At the other extreme, if hunting and fishing are perfectly correlated, then σ_{fr} must be larger than σ_{rr}, and there is negative selection into hunting. Intuitively, since *F* and *R* are perfectly positively correlated, and *F* is more dispersed, persons with low values of *F* can avoid low incomes by using their value of *R*. Persons with high values of *F* (and *R*) should fish because the upper tail of *F* is more dispersed. For cases in between, either positive or negative selection is possible depending on the sign of $\sigma_{rr} - \sigma_{fr}$. Heckman and Honoré (1990) generalize this result to a broader class of distribution functions.

This model has been generalized in a number of ways. There can be more than two occupational choices. Following Heckman and Sedlacek (1985), one can assume that individuals possess a vector of skills S_i and that different tasks use the different skills according to the function $T_j(S_i)$. We still let π_j denote task prices so that we can write an individual's wage at task *j* as

$$w_{ji} = \pi_j T_j(S_i).$$

Another extension of the model allows individuals to care about aspects of the job other than just their wages (see Heckman and Sedlacek, 1985). Let $U_{ji}(w)$ be the utility that individual *i* would receive from performing task *j* under wage level *w*. This allows for some tasks (such as playing basketball) to be generally preferred to more unpleasant tasks (such as cleaning bathrooms). Individuals then choose the occupation that yields the highest level of utility for them $U_{ji}(w_{ij})$. This is the generalized Roy model in which the generalization comes in the agent decision rules.

The generalized Roy model can be trivially extended to a model of labour force participation by allowing non-market work to be one of the tasks. To see this, let *j*=0 denote the home sector as in Gronau (1974) and Heckman (1974). Of course, in general, there will not be a market price for home-produced goods, but one can interpret $T_0(S_i)$ as the value of goods produced at home. One could also assume that staying at home is pure leisure in which $T_0(S_i) = 0$, but people enjoy staying at home $U_{0i}(0) > U_{ji}(0)$ for *j*>0. The Roy model has been generalized to allow for uncertainty in agent decision making in Cunha, Heckman and Navarro (2005). See the reviews in Heckman, Lochner and Todd (2006) and Cunha and Heckman (2007).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Identification

The economics of these models is simple, but identification and estimation are considerably more difficult. Heckman and Honoré (1990) consider identification of the basic Roy model with two occupations and income maximization. They consider two different cases: (*a*) the standard Roy model, in which the two occupations represent two different sectors of the economy and the econometrician has data on wages in both sectors; and (*b*) a case motivated by labour supply in which the econometrician has wage data from one sector (the market sector) but not from the other (the home sector). It is important to keep in mind that the comparative advantage decision at the heart of the Roy Model is just one factor that can lead to selection bias. selection bias and self-selection discusses the more general framework for thinking about sample selection and also discusses in some detail how the Roy model fits into this framework.

Heckman and Honoré (1990) consider identification from a single cross section. When one can observe wages in both sectors, under log normality, the Roy model is identified even without any regressors in the model. However, when one relaxes the log normality assumption, without regressors in the outcome equation the model is no longer identified. This is true despite the strong assumption of agent income maximization.

Heckman and Honoré (1990) provide conditions under which one can identify these models using variation across markets, or by using variation in observables within a market. To see the intuition behind the latter case, consider the model in which

$$\log(F_i) = g_f(Z_{fi}, X_i) + \varepsilon_{fi} \log(R_i) = g_r(Z_{ri}, X_i) + \varepsilon_{ri},$$

and prices are normalized to 1. In this context, it is helpful for identification to have an exclusion restriction – that is, a variable Z_{fi} that varies separately from (X_i, Z_{ri}) and a variable Z_{ri} that varies separately from (X_i, Z_{fi}). As long as there is sufficient variation in the excluded variables, Heckman and Honoré (1990) show that with a location normalization the full model is identified provided that $(\varepsilon_{fi}, \varepsilon_{ri})$ are independent of (Z_{fi}, Z_{ri}, X_i), that is, they identify g_f, g_r and the joint distribution of $(\varepsilon_{fi}, \varepsilon_{ri})$. (They also establish identification when only one sector's output is observed.)

To see the intuition for why the model is identified, consider an 'identification at infinity' argument. For convenience, take the location normalization to be

$$E(\varepsilon_{fi}) = E(\varepsilon_{ri}) = 0.$$

Suppose that g_r is such that for any x, say x_0,

$$\lim_{z_r \to -\infty} g_r(z_r, x_0) = -\infty.$$

Let $J_i \in \{f, r\}$ be an indicator of the occupation that was chosen by individual i. Then

$$\lim_{z_r \to -\infty} E[\log(F_i) | J_i = f, X_i = x, Z_{fi} = z_f, Z_{ri} = z_r] = g_f(z_f, x) + \lim_{z_r \to -\infty} E(\varepsilon_{fi} | g_f(z_f, x) + \varepsilon_{fi} > g_r(z_r, x) + \varepsilon_{ri}, X_i = x, Z_{fi} = z_f, Z_{ri} = z_r) = g_f(z_f, x).$$

By varying (z_f, x) one can trace out g_f. This occurs because conditioning on the event

$$g_f(z_f, x) + \varepsilon_f > g_r(z_r, x) + \varepsilon_r$$

becomes irrelevant as z_r becomes arbitrarily small. Identification of g_r is analogous using variation in z_f.

To identify the joint distribution of $(\varepsilon_{fi}, \varepsilon_{ri})$ note that from the data one can observe

$$\Pr(J_i = f, \log(F_i) < s | X_i = x, Z_{fi} = z_f, Z_{ri} = z_r) = \Pr(g_f(z_f, x) + \varepsilon_{fi} > g_r(z_r, x) + \varepsilon_{ri}, g_f(z_f, x) + \varepsilon_{fi} < s | X_i = x, Z_{fi} = z_f, Z_{ri} = z_r) = \Pr(\varepsilon_{fi} - \varepsilon_{ri} < g_f(z_f, x) - g_r(z_r, x), \varepsilon_{fi} < s - g_f(z_f, x) | X_i = x, Z_{fi} = z_f, Z_{ri} = z_r)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which is the cumulative distribution function of $(\varepsilon_{fi} - \varepsilon_{ri}, \varepsilon_{fi})$ evaluated at the point $(g_f(z_f, x) - g_r(z_r, x), s - g_f(z_f, x))$. By varying the point of evaluation one can identify the joint distribution of $(\varepsilon_{fi} - \varepsilon_{ri}, \varepsilon_{fi})$ from which one can derive the joint distribution of $(\varepsilon_{fi}, \varepsilon_{ri})$. Thus the model is identified. Heckman and Honoré (1989) also present conditions for identification of a competing risk version of a Roy model when there are no exclusion restrictions ($Z_r = X = Z_f$) but g_f and g_r can be independently varied. Buera (2006) makes stronger differentiability assumptions and relaxes the separability assumption in the choice equation. He also identifies a Roy model without exclusion restrictions.

Identifying the more general model where individuals choose fishing when

$$U_{fi}(w_{fi}) > U_{ri}(w_{ri})$$

is possible under a variety of assumptions. Consider the separable case in which

$$U_{fi}(w_{fi}) - U_{ri}(w_{ri}) = h(Q_i, Z_{fi}, Z_{ri}, X_i) + v_i$$

where Q_i is an additional variable that might affect the relative utilities of the two options. The function h is identified up to a normalization (see, for example, Matzkin, 1992).

Identification of parts of the model follows from the preceding reasoning. If there is a variable that affects sectoral choice, but not wages as a fisherman, we can identify g_r. Note that this exclusion restriction could be in the form of either Q_i or Z_{ri}. We can then identify the joint distribution of (v_i, ε_{ri}) using an argument analogous to the above. Using the same argument we can identify g_f and the joint distribution of (v_i, ε_{fi}). A formalization of this argument can be found in Heckman (1990) for the case in which h is linear and is extended in Heckman and Smith (1998), Carneiro, Hansen and Heckman (2003) and Heckman and Navarro (2007). One cannot, without further assumptions, identify the joint distribution of $(v_i, \varepsilon_{ri}, \varepsilon_{fi})$. (Abbring and Heckman, 2007, present conditions for identification of the joint distribution by restricting dependence relations. See also Aakvik, Heckman and Vytlacil, 2005.)

If one is interested in evaluating policies in which wages can change, this reduced form model is not enough since there is no separation of wage effects from non-wage effects in the choice model. Assume further that we can write

$$h(Q_i, Z_{fi}, Z_{ri}, X_i) + v_i = \alpha_1 F_i - \alpha_2 R_i + h^*(Q_i, Z_{fi}, Z_{ri}, X_i) + v_i^* = \alpha_1 g_f(Z_{fi}, X_i) - \alpha_2 g_r(Z_{ri}, X_i) + h^*(Q_i, Z_{fi}, Z_{ri}, X_i) + \alpha_1 \varepsilon_{fi} - \alpha_2 \varepsilon_{ri} + v_i^*.$$

Identification of this model is possible if there are exclusion restrictions in Z_{fi} and Z_{ri}, that is, if there are components of Z_{fi} and Z_{ri} that do not affect h^*. Under sufficient variation of these variables and imposing a normalization, the model is identified. An interesting special case of the model is when $\alpha_1 = \alpha_2$. In this case one needs a somewhat weaker exclusion restriction in that one could use variation in X_i. That is, we could use a variable that affects labour market outcomes, but not sectoral choice directly.

Empirical models

There are many examples that build on the Roy model, but in labour economics three stand out. The earliest empirical application of this model is to the labour supply decision (Heckman, 1974; Gronau, 1974). We refer interested readers to labour supply rather than discuss these models explicitly. The second application is to occupational choice, which is most closely linked to the original Roy model. The third, and perhaps most well known, application is to education.

We start by describing the empirical applications of the model to education. Willis and Rosen (1979) consider a model in which students decide whether to attend college. Students may have a comparative advantage in either the college sector or the high school sector. Their model assumes that decisions about schooling are made in an environment of perfect certainty on the principle of income maximization. They assume access to outcome measures in two periods. The decision to attend college depends on interest rates which are not observable to the econometrician. (One could reinterpret their model as a generalized Roy model if one interprets the interest rate as representing utility differences rather than interest rates.) Semiparametric identification requires two types of exclusion restrictions: a variable that influences the decision to attend college but not directly wages, and a variable that influences wages but not the decision to attend college directly. For the former type of exclusion restriction, Willis and Rosen (1979) use family background variables, arguing that they will be correlated with interest rates but uncorrelated with wages. For the latter type they use test scores, arguing that they are related to skill as in the Roy model, but unrelated to the interest rate.

Although they discuss comparative advantage in the labour market, as did Roy, they do not present direct empirical evidence on this question because they cannot estimate the joint distribution of schooling outcomes across both choices. They present some indirect evidence on the importance of comparative advantage in the labour market because they can identify the counterfactual means of what college students would earn if they had been high school students and what high schools students would earn had they been college students.

There are many extensions of this model, including Taber (2000), Cameron and Taber (2004) and Heckman, Lochner and Todd (2006). Cunha, Heckman and Navarro (2005) and Cunha and Heckman (2007) extend the model to allow for uncertainty, to identify agent information and to directly test for comparative advantage in the labour market by identifying the joint distribution of outcomes for the two counterfactual states (college and high school).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In a series of papers, Heckman, Lochner and Taber (1998a; 1998b; 1998c) estimate a general equilibrium version of this model. That is, they allow the skill prices π_j to be endogenous. They show that accounting for equilibrium effects is essential for estimating the impact of policy on earnings inequality. In particular Heckman, Lochner and Taber (1998b) show that ignoring equilibrium effects overstates the impact of a tuition subsidy on college enrolment by an order of magnitude. They also decompose the policy effect on earnings inequality into its various components.

Other papers estimate a Roy model of occupational choice. Most notably, Heckman and Sedlacek (1985; 1990) estimate models in which workers choose between industrial sectors. In some cases they allow for non-market work. They show how to estimate the model, but reject a pure Roy model. They show instead that a more general model with utility maximization and non-participation can fit the data well. Gould (2002) extends this framework to address the changing wage structure. He shows that workers choose sectors to maximize their comparative advantage and that this activity tends to decrease earnings inequality. However, he shows that the importance of this effect decreases over time as sectors increasingly value more similar skill sets.

Keane and Wolpin (1997) and Eckstein and Wolpin (1999) estimate dynamic discrete choice models of occupational and educational choice that extends the Roy model to a dynamic setting with uncertainty with serially independent shocks. Agents in their model make labour supply, education and occupational choice simultaneously. Heckman and Navarro (2007) present a nonparametric identification analysis of a dynamic discrete choice model with serially correlated shocks. Abbring and Heckman (2007) survey the dynamic discrete choice literature, including these papers.

See Also

- labour economics
- labour supply
- returns to schooling
- selection bias and self-selection

Bibliography

Aakvik, A., Heckman, J.J. and Vytlacil, E.J. 2005. Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics* 125, 15–51.

Abbring, J.H. and Heckman, J.J. 2007. Econometric evaluation of social programs, part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In *Handbook of Econometrics*, vol. 6, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Buera, F.J. 2006. Non-parametric identification and testable implications of the Roy model. Unpublished manuscript, Department of Economics, Northwestern University.

Cameron, S.V. and Taber, C. 2004. Estimation of educational borrowing constraints using returns to schooling. *Journal of Political Economy* 112, 132–82.

Carneiro, P., Hansen, K. and Heckman, J.J. 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. 2001 Lawrence R. Klein Lecture. *International Economic Review* 44, 361–422.

Cunha, F. and Heckman, J.J. 2007. Identifying and estimating the distributions of ex post and ex ante returns to schooling: a survey of recent developments. *Labour Economics* 14, 870–93.

Cunha, F., Heckman, J.J. and Navarro, S. 2005. Separating uncertainty from heterogeneity in life cycle earnings. The 2004 Hicks Lecture. *Oxford Economic Papers* 57, 191–261.

Eckstein, Z. and Wolpin, K.I. 1999. Why youths drop out of high school: the impact of preferences, opportunities, and abilities. *Econometrica* 67, 1295–339.

Gould, E.D. 2002. Rising wage inequality, comparative advantage, and the growing importance of general skills in the United States. *Journal of Labor Economics* 20, 105–47.

Gronau, R. 1974. Wage comparisons – a selectivity bias. *Journal of Political Economy* 82, 1119–43.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–94.

Heckman, J.J. 1990. Varieties of selection bias. *American Economic Review* 80, 313–8.

Heckman, J.J. and Honoré, B.E. 1989. The identifiability of the competing risks model. *Biometrika* 76, 325–30.

Heckman, J.J. and Honoré, B.E. 1990. The empirical content of the Roy model. *Econometrica* 58, 1121–49.

Heckman, J.J., Lochner, L.J. and Taber, C. 1998a. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics* 1, 1–58.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Heckman, J.J., Lochner, L.J. and Taber, C. 1998b. General-equilibrium treatment effects: a study of tuition policy. *American Economic Review* 88, 381–6.

Heckman, J.J., Lochner, L.J. and Taber, C. 1998c. Tax policy and human-capital formation. *American Economic Review* 88, 293–7.

Heckman, J.J., Lochner, L.J. and Todd, P.E. 2006. Earnings equations and rates of return: the Mincer equation and beyond. In *Handbook of the Economics of Education*, ed. E.A. Hanushek and F. Welch. Amsterdam: North-Holland.

Heckman, J.J. and Navarro, S. 2007. Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics* 136, 341–96.

Heckman, J.J. and Sedlacek, G.L. 1985. Heterogeneity, aggregation, and market wage functions: an empirical model of self-selection in the labor market. *Journal of Political Economy* 93, 1077–125.

Heckman, J.J. and Sedlacek, G.L. 1990. Self-selection and the distribution of hourly wages. *Journal of Labor Economics* 8(1, Part 2), S329–63.

Heckman, J.J. and Smith, J.A. 1998. Evaluating the welfare state. In *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, ed. S. Strom. New York: Cambridge University Press.

Keane, M.P. and Wolpin, K.I. 1997. The career decisions of young men. *Journal of Political Economy* 105, 473–522.

Matzkin, R.L. 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica* 60, 239–70.

Roy, A. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.

Taber, C.R. 2000. Semiparametric identification and heterogeneity in discrete choice dynamic programming models. *Journal of Econometrics* 96, 201–29.

Willis, R.J. and Rosen, S. 1979. Education and self-selection. *Journal of Political Economy* 87(5, Part 2), S7–S36.

How to cite this article

Heckman, James J. and Christopher Taber. "Roy model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000272> doi:10.1057/9780230226203.1463

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Roy, René François Joseph (1894 – 1977)

Maurice Allais
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

demand functions; index numbers; Roy, R. F. J.

Article

René Roy was born in Paris on 21 May 1894. He entered the Ecole Polytechnique in 1914, and joined the army on 15 August 1914. He was seriously wounded on 14 April 1917 at the Chemin des Dames, as a result of which he was blinded at the early age of 23. This tragedy, which meant the collapse of all his youthful hope and dreams, brought him to the slough of despond, and exceptional spiritual strength alone enabled him eventually to accept the unacceptable with serenity and to undertake a double career as an engineer and economist that was to last 60 years.

He studied at the Ecole Polytechnique (from 1918 to 1920) graduating first in his year, and then at the Ecole Nationale des Ponts et Chaussées (1920–2). He entered the Ministry of Public Works and Transport as a state engineer in 1922, specializing in problems of local railway networks and urban transport until his retirement in 1964. He died in Paris in 1977.

In parallel with this activity, he became Professor of General Political Economy and Social Economy at the Ecole des Ponts et Chaussées in 1929, and Professor of Econometrics at the Statistical Institute of the University of Paris in 1931. In 1949 he taught econometrics at the Ecole d'Application de l'Institut National de la Statistique et des Etudes Economiques (School of Instruction of the National Institute of Statistics and Economic Studies). From 1947 he was in charge of an Econometrics Seminar at the National Centre of Scientific Research.

He was elected President of the Paris Statistical Society in 1949 and of the International Econometrics Society in 1953. He was also a fellow of the International Statistical Society (1949), a member of the Academy of Moral and Political Science (1951), and an honorary fellow of the Royal Statistical Society (1957). He received the degree of Doctor Honoris Causa from the University of Geneva in 1964.

René Roy's research was focused mainly on transport, demand functions, economic indices, fields of choice and their respective relationships. His main published works are *Le régime économique des voies ferrées d'intérêt local* (1925), his doctoral thesis; *La demande de biens de consommation directe* (1935);

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
De l'utilité – contribution à la théorie des choix (1942); ‘Les nombres indices’ (*Journal de la Société de Statistique de Paris*, 1949); and *Eléments d’économétrie* (1970). In addition, in collaboration with François Divisia and Jean Dupin, he published in 1953–4 *A la recherche du franc perdu*, whose three volumes cover the movement of prices, production and wealth respectively in France from 1914 to 1950. Roy's analysis of the basic relationships of demand functions and price and quantity index numbers are contained in his 1949 publication.

René Roy's ability to analyse very difficult questions and constantly stay abreast of the main publications of his era was a truly remarkable achievement for a totally sightless person. He showed that accomplishment is possible in the face of an irremediable adversity by dint of unremitting energy associated with remarkable intelligence. His book *Vers la lumière* (1930) gives us his message as a blind man.

See Also

- indirect utility function

Selected works

1925. *Le régime économique des voies ferrées d'intérêt local*. Paris.

1930. *Vers la lumière*. Paris: Bibliotheque-Charpentier.

1935. *Le demande de biens de consommation directe*. Paris: Hermann.

1942. *De l'utilité – contribution à la théorie des choix*. Paris: Hermann.

1949. Les nombres indices. *Journal de la Société de Statistique de Paris* 90(1–2), 15–34.

1953–4. (With F. Divisia and J. Dupin.) *A la recherche du franc perdu*. Paris: Société d'Editions Hommes et Mondes.

1970. *Eléments d’économétrie*. Paris: Presses Universitaires de France.

How to cite this article

Allais, Maurice. "Roy, René François Joseph (1894–1977)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000185> doi:10.1057/9780230226203.1464

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Royal Economic Society

Roger Middleton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; British Association for the Advancement of Science; British Economic Association; Edgeworth, F. Y.; Harrod, R. F.; Keynes. J. M.; Marshall, A.; Royal Economic Society; Royal Statistical Society

Article

Originally founded in 1890 as the British Economic Association (BEA), the Royal Economic Society (RES) assumed its current title in 1902 when it obtained a Privy Council charter and royal patronage. The RES is now unquestionably the leading organization of professional economists in Britain, with its flagship publication the *Economic Journal* (*EJ*), a world-class general journal for theoretical and applied research (having in 2004 an International Statistical Institute, ISI, journal citation ranking of 15/172 and 1.723 impact factor). Such dominance, however, has not always been the case and was not easily achieved, with the RES's fortunes, like those of many other long-established economics associations, subject to a changing complex of pressures, including at times competitors.

The RES was the eventual institutional result of the long process whereby political economy was transformed into economics as the ‘sciences of the social’ were dissolved and reconstructed into the modern social sciences in the second half of the 19th century. The establishment of the BEA followed a long period of consultations over whether a new learned society was required to propagate what has come to be known in the modern literature as Marshall's mission for the professionalization of economics (Middleton, 1998), or whether the Royal Statistical Society (established 1834) and/or Section F (Economics and Statistics, established 1835) of the British Association for the Advancement of Science, were sufficient vehicles.

The year 1890 witnessed also the publication of Marshall's *Principles of Economics* and the completion of the first instalment of the first Palgrave dictionary, which was published the following year, as was the first issue of the *EJ*. For Keynes (1940, p. 409), these happy concurrences made this the beginnings of the ‘modern age of British economics’. This now looks somewhat questionable: the BEA and *EJ* were part of a pre-professionalization trend towards ‘clearer demarcation and definition of the field of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
scholarly endeavour' (Kadish and Freeman, 1990, p. 23) and the achievements of the professionalization agenda were as yet limited. That this was the case is apparent from what was one of the central issues of the debate preceding the BEA's formation: would it be a closed society of professional economists (at this time, it was still not true that a majority of these consisted of academics) or would it be open to all, however imperfect their claim to be called an economist? Mindful of the contemporaneous example of the American Economic Association (AEA), where many of its leading figures had been deeply involved in methodological and policy disputations which had been unhelpful to the professionalization agenda, the BEA's founders resolved to follow a more cautious and restricted policy than its transatlantic counterpart: membership was not open but dependent upon a candidate's approval by Council (with the chosen designated as Fellows from 1902 to 1964 under the Royal Charter); the BEA was to refrain from organizing discussions and conferences, partly for fear of exposing the substantial differences of opinion within their ranks; and, in its early years, the Council routinely chose a prominent public figure rather than an academic as its president, beginning with the then Chancellor of the Exchequer, G. J. Goschen.
While the early years of the BEA were often difficult, the *EJ* was an undoubted success, and this notwithstanding the rival Oxford publication, the *Economic Review*, which had been launched a few months earlier (and survived until 1914). The *EJ* was not conceived as a specialist publication for an exclusively academic or quasi-academic economics audience, but instead as a 'means of disseminating economic truth amongst readers from all walks of life, while setting new standards of economic investigation' (Kadish and Freeman, 1990, pp. 36–7). This general informational role was to endure strongly until the 1970s and was still be present in the 1990s when, as the journal became increasingly internationalized, it almost exclusively focused on scholarly papers and its policy forum, with less emphasis on book reviews and with reports on its learned society activities hived off to the website (http://www.res.org.uk).
The *EJ's* initial editor was F. Y. Edgeworth, followed by Keynes (singularly and jointly, 1912–45) and then R. F. Harrod (1945–61), during which time the *EJ* consolidated its status as the leading British economic journal, despite the appearance of several rivals. Cambridge, Oxford and London economists initially dominated, but from the 1970s onwards, as the balance of professional influence and authority shifted away from the older centres, the Council had to respond to pressures to make the RES a more democratic organization. Concurrently, the *EJ* editors found that the ever-increasing scale, scope and technical nature of the discipline necessitated increased personnel to provide expert opinion on journal submissions. An editorial board was established in 1971 and has since evolved; most importantly, it now has more foreign (largely, but not exclusively, US) than British economists.
RES membership is now over 3,300 individuals, of whom 60 per cent are not British residents, and there are a further 2,400 institutional subscribers to the *EJ* and the *Econometrics Journal* (established 1998). Notwithstanding the internationalization of the *EJ*, the RES mission statement remains essentially British: to be the 'professional association which promotes the encouragement of the study of economic science in academic life, government service, banking, industry and public affairs'. Increasingly an umbrella organization for a large number of activities, mainly but not exclusively to do with university economics, the RES maintains also its activities as a major publisher of scholarly editions, of which the 30-volume collected writings of Keynes is one of its major achievements, even if it did nearly bankrupt the Society. It has also finally resolved one of the issues of disagreement at its creation: since 1990 the RES has operated an annual conference, selected papers from which appear in a special issue of the *EJ*, itself now enlarged for its second century from a quarterly to a bimonthly publication.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- Keynes, John Maynard
- Marshall, Alfred

Bibliography

Coats, A.W. 1968. The origins and early development of the Royal Economic Society. *Economic Journal* 78, 349–71.

Kadish, A. and Freeman, R.D. 1990. Foundation and early years. In *A Century of Economics: 100 Years of the Royal Economic Society and the Economic Journal*, ed. J.D. Hey and D. Winch. Oxford: Basil Blackwell.

Keynes, J.M. 1940. The Society's jubilee, 1890–1940. *Economic Journal* 50, 401–9.

Middleton, R. 1998. *Charlatans or Saviours? Economists and the British Economy from Marshall to Meade*. Cheltenham: Edward Elgar.

How to cite this article

Middleton, Roger. "Royal Economic Society." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000186> doi:10.1057/9780230226203.1465

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Rubin causal model

Guido W. Imbens and Donald B. Rubin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Rubin Causal Model (RCM), a framework for causal inference, has three distinctive features. First, it uses ‘potential outcomes’ to define causal effects at the unit level, first introduced by Neyman in the context of randomized experiments and randomization-based inference, but not used formally in non-randomized studies or with other modes of inference until Rubin (1974; 1975). Second is its formal use of a probabilistic assignment mechanism, which mathematically describes how treatments are given to units, with possible dependence on background variables and the potential outcomes themselves. Third is an optional probability distribution on all variables, including the potential outcomes, which thereby unifies frequentist and model-based forms of statistical inference for causal effects within one framework.

Keywords

assignment mechanism; assignment-based inference; Bayesian inference; causal inference; Fisher, R. A.; Haavelmo, T.; Hurwicz, L.; instrumental variables; interval estimates; Markov chain Monte Carlo methods; matching; multiple imputation; Neyman, J.; posterior predictive distribution; potential outcomes; principal stratification; probability; program evaluation; propensity scores; randomization-based inference; randomized experiment; regression coefficients; Roy model; Rubin causal model; simultaneous equations models; Tinbergen, J.; treatments; units

Article

The Rubin Causal Model (RCM) is a formal mathematical framework for causal inference, first given that name by Holland (1986) for a series of previous articles developing the perspective (Rubin, 1974; 1975; 1976; 1977; 1978; 1979; 1980). There are two essential parts to the RCM, and a third optional one. The first part is the use of ‘potential outcomes’ to define causal effects in all situations – this part defines ‘the science’, which is the object of inference, and it requires the explicit consideration of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
manipulations that define the treatments whose causal effects we wish to estimate. The second part is an explicit probabilistic model for the assignment of 'treatments' to 'units' as a function of all quantities that could be observed, including all potential outcomes; this model is called the 'assignment mechanism', and defines the structure of experiments designed to learn about the science from observed data or the acts of nature that lead to the observed data. The third possible part of the RCM framework is an optional distribution on the quantities being conditioned on in the assignment mechanism, including the potential outcomes, thereby allowing model-based Bayesian 'posterior predictive' (causal) inference. This part of the RCM focuses on the model-based analysis of observed data to draw inferences for causal effects, where the observed data are revealed by applying the assignment mechanism to the science. A full-length text that discusses estimation and inference for causal effects from this perspective is Imbens and Rubin (2006).

Implications of the RCM for research design

Before defining each of these three parts of the RCM, it is helpful to consider the implications of this structure for applied research about causal effects. The first part implies that we should always start by carefully defining all causal estimands (quantities to be estimated) in terms of potential outcomes, which are all values that could be observed in some real or hypothetical experiment that compares the results under an active treatment with the results under a control treatment. That is, causal effects are defined by a comparison of (*a*) the values that would be observed if the active treatment were applied and (*b*) the values that would be observed if, instead, the control treatment were applied. This step contrasts with the common practice of defining causal effects in terms of parameters in some model, where the manipulations defining the active versus control treatments are often left implicit and ill-defined, with the resulting causal inferences correspondingly weak and ill-defined. This first part can be completely abstract and can take place before any data are observed or even collected. In the RCM, however, there is 'no causation without manipulation' (Rubin, 1975, p. 238), where the manipulation (that is, the treatment) could be real or hypothetical. The collection of potential outcomes with and without this manipulation defines the scientific objective of causal inference in all studies, whether randomized, observational or entirely hypothetical.

The second part of the RCM, the assignment mechanism, implies that, given the defined science, we should continue by explicating the design of the real or hypothetical study being used to estimate that science. The assignment mechanism describes why some study units will be (or were) exposed to the active treatment and why other study units will be (or were) exposed to the control treatment, and the reasons are formalized by the mathematical statement of the assignment mechanism. When the study is a true experiment, the assignment mechanism may involve the consideration of background (that is, pretreatment assignment) variables for the purpose of creating strata of similar units to be randomized into treatment and control, thereby improving the balance of treatment and control groups with respect to these background variables (that is, covariates). A true experiment automatically cannot use any outcome (post-treatment) variables to influence design because they are not yet observed. If the observed data were not generated by a true experiment, but rather by non-randomized observational data, there still should be an explicit design phase. That is, in an observational study, the same guidelines as in an experiment should be followed.

More explicitly, the design step in the analysis of an observational data set for causal inference should

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
structure the data to approximate (or reconstruct or replicate) a true randomized experiment as closely as possible. In this design step, the researcher never uses or even examines any outcome data but rather identifies subsets of units such that the treatments can be thought of as being randomly assigned within the subsets. This assumed randomness of treatment assignment is assessed by examining, within these subsets of units, the similarity of the distributions of the covariates in the treatment group and in the control group. Because this design step is focused on creating these subsets of units with balanced distributions of covariates between treatment and control groups, and never uses outcome data, the researcher cannot select a design to produce a desired answer, even unconsciously.

The third part of the RCM is optional; it derives inferences for causal effects from the observed data by conceptualizing the problem as one of imputing the missing potential outcomes. That is, once outcome data are available (that is, observations of the potential outcomes corresponding to the treatments actually received by the various units), then the modelling of the outcome data given the covariates should be structured to derive predictions of those potential outcomes that would have been observed if the treatment assignments had been different. This modelling will generate stochastic predictions (that is, imputations) for all missing potential outcomes in the study, which, when combined with the actually observed potential outcomes, will allow the calculation of any causal-effect estimand. Because the imputations of the missing potential outcomes are stochastic, repeating the process results in different values for the causal-effect estimand. This variation across the multiple imputations (Rubin, 1987; 2004a) generates interval estimates and tests for the causal estimands. Typically, in practice this third part is implemented using simulation-based methods, such as Markov chain Monte Carlo computation applied to Bayesian models.

The conceptual clarity in the first two steps of the RCM often allows previously difficult causal inference situations to be easily formulated and handled. The optional third part often extends this success by relying on modern computational power to handle analytically intractable problems. With this overview in place, we consider features of the RCM in more detail.

Potential outcomes and causal effects

For defining causal effects, there are three basic primitives – concepts that are fundamental and on which we must build: units, treatments and potential outcomes. A unit is a physical object, for example a person, at a particular point in time. A treatment is an action that can be applied or withheld from a unit. We focus on the case of two treatments, although the extension to more than two treatments is simple in principle although not necessarily so with real data. Associated with each unit are two potential outcomes: the value of an outcome variable Y at a future point in time if the active treatment is applied, and the value of Y at the *same* future point in time if instead the control treatment is applied. The objective is to learn about the causal effect of the application of the active treatment relative to the control on Y, where, by definition, the causal effect is a comparison of the two potential outcomes. For example, the unit could be a person ‘now’ without a job, the active treatment could be participating in a job training programme, and the control could be not participating. The outcome Y could be the total income over the next three years, with the two potential outcomes being the total income with and without job training; the causal effect of being trained versus not being trained is the comparison of the person's three-year total income with and without the training.

Notationally, let W indicate which treatment the unit receives: $W = 1$ the active treatment, and $W = 0$ the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
control treatment. Also let $Y(1)$ be the value of the potential outcome if the unit received the active version, and $Y(0)$ the value if the unit received the control version. The causal effect of the active treatment relative to the control is the comparison of $Y(1)$ and $Y(0)$ – typically the difference, $Y(1) - Y(0)$, or perhaps the difference in logs, $\log[Y(1)] - \log[Y(0)]$, or some other comparison, possibly the ratio. We can observe only one or the other of $Y(1)$ and $Y(0)$ as indicated by W: $Y_{obs} = WY(1) + (1 - W)Y(0)$. The 'fundamental problem facing inference for causal effects' (Rubin, 1978, p. 38) is that, for any individual unit, we observe the value of the potential outcome for this unit under only one of the possible treatments, namely, the treatment actually assigned, and the potential outcome under the other treatment is missing. Thus, inference for causal effects is a missing-data problem – the 'other' value is missing, so the nature of causal inference is that at least 50 per cent of the values of the potential outcomes are missing. Covariates have values that are unaffected by the treatments, such as age or sex of the unit in the job training example, and are denoted by X. Even when X represents a lagged Y, such as total income last year, $Y(1) - X$ is not the causal effect of training unless $Y(0) = X$, but rather a change of income across time.

To clarify the RCM set-up with potential outcomes, consider a specific difficult case: what is the causal effect of race on hiring practices? To consider this explicitly causal question in the RCM, we must consider the manipulations that define the active and control treatments. Literally changing one's race is presumably impossible given current medical technology, but one can conceptualize experiments that can plausibly capture what researchers want to know, that is, how employers react to race when all else is constant. For example, suppose that résumés are submitted by mail to groups of employers, where the treatment to be applied to each résumé (that is, each unit) is the name attached to it (see, for example, Bertrand and Mullainathan, 2004). Here, the active treatment is the use of a distinctive African-American name on the résumé, and the control treatment is the use of a traditional name. In this case, the explication of what is meant by 'the causal effect race' is through the description of the manipulations, and the causal effect to be estimated is thereby well-defined: the causal effect of having a résumé with an African-American name compared with a traditional name on the resultant hiring outcome. Whether that effect corresponds to what the investigator wants to estimate or to what others believe is relevant to policy is another issue, but the causal nature of the comparison is clear. If it is not the desired quantity estimand or is deemed not relevant, then other more appropriate manipulations should be described.

Suppose, now that there are N units rather than only one. To make the representation with only two potential outcomes for each unit adequate, must accept an assumption, the stable unit treatment value assumption (SUTVA; Rubin, 1980), which rules out interference between units (Cox, 1958) and rules out different versions of the treatments for the units (for example, no 'technical errors'; Neyman, 1935; Rubin, 1990b). SUTVA can be weakened, but still some such assumption regarding the full set of potential outcomes is required. Often, in practice, SUTVA is made more plausible by aggregating the units. For example, training some of the unemployed in a local labour market may affect job opportunities for others in that local market. Therefore changing the unit of analysis to be the local labour market in a study with many geographically separated local labour markets may make it more plausible that there is no effect of the exposure of one unit to the treatment on other units.

Under SUTVA, all causal estimands (quantities to be estimated) can be defined from the matrix of values with i^{th} row: $(X_i, Y_i(0), Y_i(1))$, $i = 1, \ldots, N$. A causal estimand involves a comparison of $Y_i(0)$ and $Y_i(1)$ on all N units, or on a common subset of units; for example, the average causal effect across all

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
units that are female as indicated by their X_i, or the median $Y_i(1)$ minus the median $Y_i(0)$ for the set of units with X_i indicating male and $Y_i(0)$ indicating no income. By definition, all relevant scientific information that is recorded is encoded in this matrix, and so the labelling of its rows is a random permutation of 1,…, N; that is, the N-row matrix $\{X, Y(0), Y(1)\}$ is row exchangeable. For convenience, we refer to this array of values as the 'science', functions of which we wish to estimate.

Brief history of potential outcomes to define causal effects

The basic idea that causal effects are the comparisons of potential outcomes on a common set of units seems so direct that it must have ancient roots, and we can find elements of this definition of causal effects among both philosophers (for example, Mill, 1843, p. 327) and experimenters (for example, Fisher, 1918, p. 214). But apparently there was no formal notation for potential outcomes until Neyman (1923), which appears to have been the first place where a mathematical analysis is written for a randomized experiment. This notation became standard for work in randomized experiments with randomization-based inference, and was a major advance. Independently and nearly simultaneously, Fisher (1925) recommended physically randomizing treatments to units in experiments, as well as a different, but compatible, method of randomization-based inference, although Fisher apparently never used the potential outcomes notation. But despite the almost immediate acceptance in the late 1920s of Fisher's proposal for randomized experiments, and of Neyman's notation for potential outcomes in randomized experiments, and of both men's proposals for randomization-based inference, this potential outcome notation was not used for causal inference more generally for a half century thereafter, apparently not until introduced by Rubin (1974). As a result, the insights into causal inference that accompanied the use of the potential outcomes notation were entirely limited to the relatively simple setting of randomization-based inference in randomized experiments.

The approach used in nonrandomized settings, during the half-century following the introduction of Neyman's seminal notation for randomized experiments, was based on mathematical models (for example, regression models) relating the observed value of the outcome variable $Y_{\text{obs},i}$ to X_i and W_i, and then defining causal effects as parameters (for example, regression coefficients) of these models. This was the standard approach in medical and social science, including economics, and led to substantial confusion – the role of randomization cannot even be directly stated mathematically using the observed outcome notation. Of course, there were seeds of this first part of the RCM in social science before 1974, in particular in economics, in Tinbergen (1930), Haavelmo (1944) and Hurwicz (1962), but we can find no previous use of explicit notation like Neyman's to define causal effects. The use of the idea of potential outcomes certainly did appear in discussions in economic theory, for example, in the context of supply and demand functions (for example, Haavelmo, 1944) or the Roy (1951) model, but these discussions did not lead to inference in terms of potential outcomes. Instead, inference took place in terms of the specification of simultaneous equations using observed quantities and distributional properties of error terms (for example, Heckman and Robb, 1984, in the context of program evaluation models).

Nevertheless, the potential outcome part of the RCM framework for defining causal effects, namely, a generalization of Neyman's notation to allow non-randomized data, seems to have been basically

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
accepted and adopted by most researchers by the end of the 20th century; compare, for example, Imbens and Angrist (1994) and Heckman, Ichimura and Todd (1998) with the earlier formulation in Heckman and Robb (1984). An article exploring whether the full potential outcomes framework can be avoided when conducting causal inference is Dawid (2000), which included discussion by others that was largely supportive of the propriety of potential outcomes for causal inference.

The assignment mechanism and assignment-based causal inference

The second part of the RCM framework is the specification of an ‘assignment mechanism’: a probabilistic model for how some units received the active treatment and how other units received the control – how we conceptualize the design for how some potential outcomes were revealed and others remained hidden (that is, missing). The assignment mechanism is fundamental to causal inference. It specifies the conditional probability of each vector of assignments $W = (W_1, \ldots, W_N)^T$ given the matrix of all covariates and potential outcomes:

$$\Pr(W|X, Y(0), Y(1)). \tag{1}$$

It appears that Rubin (1975) was the first place that expressed the possible dependence of the assignment vector on the potential outcomes in this direct way, which allows the statement of what makes randomized experiments special, and more generally, generates a classification of assignment mechanisms. Again, economic theory sometimes implied a specific assignment mechanism, but this theory was never explicitly stated as in the general form of (1). For example, individuals may choose the occupation that maximizes their earnings, as in the Roy model, which would lead to $W_i = \operatorname{argmax}_w(Y_i(W))$, or more generally individuals may optimize an objective function that involves expectations over the unknown components of the potential outcomes. Imbens and Rubin (2006) provide details of such examples.
Randomized experiments are special in that they have ‘unconfounded’ and ‘probabilistic’ assignment mechanisms. Unconfounded assignment mechanisms (Rubin, 1990a) are free of dependence on either $Y(0)$ or $Y(1)$:

$$\Pr(W|X, Y(0), Y(1)) = \Pr(W|X). \tag{2}$$

Assignment mechanisms are ‘probabilistic’ (or ‘probability’ as in Rubin, 1990a) if each unit has a positive probability of receiving either treatment:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$0 < \Pr(W_i = 1|X, Y(0), Y(1)) < 1. \quad (3)$$

'Strongly ignorable' assignment mechanisms (Rosenbaum and Rubin, 1983a) satisfy (2) and (3), and thus have unit level probabilities, or 'propensity scores', $\Pr(W_i = 1|X_i)$, that are strictly between 0 and 1, and are free of all potential outcomes.
Ignorable assignment mechanisms (Rubin, 1978), are free from dependence on missing potential outcomes but may depend on observed potential outcomes $Y_{obs} = \{Y_{obs,i}\}$

$$\Pr(W|X, Y(0), Y(1)) = \Pr(W|X, Y_{obs}). \quad (4)$$

Ignorable but confounded assignment mechanisms arise in practice, especially in sequential experiments. All strongly ignorable assignment mechanisms are unconfounded, and all unconfounded assignment mechanisms are ignorable, but not the other way. Strongly ignorable assignment mechanisms allow particularly straightforward estimation of causal effects, and are the basic template for the analysis of observational studies. More generally, observational studies have possibly confounded, non-ignorable, assignment mechanisms. A confounded assignment mechanism is one that depends on the potential outcomes, and so does not satisfy (2); a non-ignorable assignment mechanism does not even satisfy (4), and thus allows treatment assignment (or, to use common economics terminology, 'selection') to depend on unobserved values, that is, the missing potential outcomes, $Y_{mis} = \{Y_{mis,i}\}$, $Y = \{Y_{obs}, Y_{mis}\}$.
When the assignment is strongly ignorable, it can generally be represented as a 'regular' assignment mechanism, which is proportional to the product of the propensity scores:

$$\Pr(W|X, Y(0), Y(1)) \propto \prod_{i=1}^{N} \Pr(W_i = 1|X_i). \quad (5)$$

Regular assignment mechanisms are the basic template in the RCM for the analysis of observational data, because two units with the same propensity score but different treatments are essentially randomized into the two treatment conditions. Therefore, with regular assignment mechanisms, matching on the propensity score (for example, as in Rosenbaum and Rubin, 1984), or subclassifying on

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
it (for example, as in Rosenbaum and Rubin, 1985), restores the assumed underlying experimental design, and inference is straightforward based only on the assignment mechanism. These assignment-based methods of inference are due to Neyman (1923) and Fisher (1925), and they involve the calculation of large-sample confidence intervals and exact significance tests for null hypotheses, respectively; both are discussed in Rubin (1990a; 1990b; 1991). For the validity of either Fisher's or Neyman's approach, the analysis must formally be defined a priori, as part of the design. But the existence of these assignment-based modes of inference helps justify the view in the RCM that the model for the assignment mechanism is more fundamental for causal inference than the model for the science, which is not needed for randomization-based inference.

Thus, in the RCM an observational study should be designed as if its data arose from a ‘broken’ randomized experiment, where the unknown propensity scores must be reconstructed on the basis of the covariates X prior to ever observing any potential outcomes. In such settings, it is often quite advantageous to use estimated propensity scores (for example, as in Rosenbaum and Rubin, 1984; Rubin and Thomas, 1992a; 1992b; 1996; 2000; Hirano, Imbens and Ridder, 2003). When estimated propensity scores for some units are so low that they have essentially no chance of being treated, then those units should be discarded from further consideration when estimating the treatment effect in the treated (see, for example, Peters, 1941; Belson, 1956; Cochran and Rubin, 1973; Rubin, 1973a; 1973b; 1977; Rosenbaum and Rubin, 1985; Dehijia and Wahba, 1999; Crump et al., 2005). The result of the design phase should be treatment and control groups with very similar distributions of observed Xs, because of either matching or subclassification. If a data-set does not permit similar X distributions to be constructed in treatment and control groups, it cannot be used to support causal inferences without extraneous assumptions justifying extrapolations. Rubin (2002) offers an example of such matching and subclassification in the context of the US tobacco litigation, and Rubin (2006a) is a book devoted to matched sampling.

A striking example of the applied success of the above approach to inference in observational studies is Dehijia and Wahba (1999), which reanalysed the classic Lalonde (1986) data on job-training experiments but using the assignment-based approach of the RCM. In contrast to the wild variety of contradictory, but highly significant, answers found by the traditional econometric methods, Dehijia and Wahba used matching on the propensity score to arrive at inferences that tracked those from the underlying randomized experiment in the overall sample and in a variety of subsamples (see also Abadie and Imbens, 2006).

Posterior predictive, or model-based, causal inference

The third part of the RCM involves an optional distribution on the N-row array of science, $\Pr(X,Y(0),Y(1))$, thereby allowing Bayesian, or model-based, inference as well as assignment-based inference. An important virtue of the RCM framework is that it distinctly separates the science – its definition in the first part (and a possible model for it in the third part) from the design of what is revealed about the science – the assignment mechanism in the second part, which can also involve some scientific insights as when it is assumed to be generated by equilibrium conditions, as in supply and demand models, or by optimizing behaviour, and so on.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bayesian inference for causal effects directly and explicitly confronts the missing potential outcomes, Y_{mis}, by using the specification for the assignment mechanism and the specification for the underlying data to derive the posterior predictive distribution of Y_{mis}, that is, the distribution of Y_{mis} given all observed values:

$$\Pr(Y_{mis}|X, Y_{obs}, W).$$

This distribution is posterior because it conditions on all observed values (X,Y_{obs},W) and is predictive because it predicts (stochastically) the missing potential outcomes. From this distribution and all of the observed values (the observed potential outcomes, Y_{obs}; the observed assignments, W; and observed covariates, X), the posterior distribution of any causal effect can, in principle, be calculated. This conclusion is immediate if we view the posterior predictive distribution as specifying how to take a random draw of Y_{mis}. Once a value of Y_{mis} is drawn, any causal effect can be directly calculated from the drawn value of Y_{mis} and the observed values of X and Y_{obs}, for example, the median causal effect for males: med{Y_i (1)−Y_i (0)|X_i indicate males}. Repeatedly drawing values of Y_{mis} and calculating the causal effect for each draw generates the posterior distribution of the desired causal effect. Thus, we can view causal inference entirely as a missing data problem, where we multiply-impute (Rubin, 1987; 2004a) the missing potential outcomes to generate a posterior distribution for the causal effects.
For example, the treated units have $Y_i(1)$ observed and $Y_i(0)$ missing. Under ignorability, the regression of $Y_i(0)$ on X_i among treated units, for which there is no direct evidence, can be shown to be the same as the regression of $Y_i(0)$ on X_i among controls, for which we have data. Thus, this third part of the RCM tells us to build a realistic model of $Y_i(0)$ given X_i among control subjects, and use it to impute the missing $Y_i(0)$ among the treated from their X_i values, while being wary of issues of extrapolation beyond the observed range of X_i control values. Analogously, build a model of $Y_i(1)$ given X_i among the treated, and use it to impute the missing $Y_i(1)$ among controls. The general structure is outlined in Rubin (1978), and is developed in detail in Imbens and Rubin (2006); a chapter-length summary appears in Rubin (2007).

Advantages of the RCM

Because of the flexibility in the RCM for (*a*) formulating causal estimands, and (*b*) positing assignment mechanisms, it can handle difficult cases in principled ways. With observational studies, estimated propensity scores play a key role, because the initial analysis proceeds as if the assignment mechanism were unconfounded. To assess the consequences of this assumption, sensitivity analyses can be conducted under various hypothetical situations, typically with fully missing covariates, U, such that treatment assignment is unconfounded given U but not given the observed data. Assumed relationships (given X) between U and W, and between U, $Y(0)$ and $Y(1)$, are then varied, for example, as in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Rosenbaum and Rubin (1983b), utilizing the third part of the RCM. Ideally, this speculation occurs at the design stage. Extreme versions of sensitivity analyses lead to large-sample bounds (for example, Manski, 2003).

A complication, common when the units are people, is non-compliance with assigned treatment. Early work related to this issue can be found in economics using the terminology of instrumental variables, and the bridge from this terminology to the basic RCM is developed in Imbens and Angrist (1994) and in Angrist, Imbens and Rubin (1996), and the connection to the full RCM approach is presented in Imbens and Rubin (1997) and in Hirano et al. (2000). Another complication is censoring due to death, where units may 'die' before the final outcome can be measured. This problem is formulated from the RCM perspective in Rubin (2006b), with bounds given in Zhang and Rubin (2003); see Zhang, Rubin and Mealli (2007) for application to the evaluation of job-training programmes. This topic is also related to 'direct' and 'indirect' causal effects (Rubin, 2004b; 2005). Combinations of such complications are considered in Barnard et al. (2003) in the context of a school choice example, as well as in Mealli and Rubin (2002; 2003), Jin and Rubin (2007) and Frangakis and Rubin (1999; 2001) in other contexts. The above examples can all be viewed as special cases of 'principal stratification' (Frangakis and Rubin, 2002).

The references in the preceding paragraph are clearly idiosyncratic in the sense of their being specific applications of the RCM in which the authors of this article have been participants, and are not representative, but we hope they provide indications of the breadth of recent applications of the RCM.

See Also

- Bayesian econometrics
- Bayesian statistics
- econometrics
- matching
- matching estimators
- treatment effect

Bibliography

Abadie, A. and Imbens, G.W. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74, 235–67.

Angrist, J.D., Imbens, G.W. and Rubin, D.B. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91, 444–72 (an Applications Invited Discussion Article with discussion and rejoinder).

Barnard, J., Hill, J., Frangakis, C. and Rubin, D. 2003. A principal stratification approach to broken randomized experiments: a case study of vouchers in New York City. *Journal of the American*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Statistical Association 98, 299–323 (with discussion and rejoinder).

Bertrand, M. and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94, 991–1013.

Belson, W.A. 1956. A technique for studying the effect of a television broadcast. *Applied Statistics* 5, 195–202.

Cochran, W.G. and Rubin, D.B. 1973. Controlling bias in observational studies: a review. *Sankhya*, A 35, 417–46.

Cox, D.R. 1958. *The Planning of Experiments*. New York: Wiley.

Crump, R., Hotz, J., Imbens, G. and Mitnik, O. 2005. Moving the goalposts: addressing limited overlap in estimation of average treatment effects by changing the estimand. Unpublished manuscript, Department of Economics, University of California, Berkeley.

Dawid, A.P. 2000. Causal inference without counterfactuals. *Journal of the American Statistical Association* 95, 407–24 (with discussion).

Dehijia, R. and Wahba, S. 1999. Causal effects in non-experimental studies: re-evaluating the evaluations of training programs. *Journal of the American Statistical Association* 94, 1053–62.

Fisher, R.A. 1918. The causes of human variability. *Eugenics Review* 10, 213–20.

Fisher, R.A. 1925. *Statistical Methods for Research Workers*, 1st edn. Edinburgh: Oliver and Boyd.

Frangakis, C. and Rubin, D.B. 1999. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 86, 366–79.

Frangakis, C.E. and Rubin, D.B. 2001. Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. *Biometrics* 57, 333–42 (with discussion and rejoinder, 343–53).

Frangakis, C.E. and Rubin, D.B. 2002. Principal stratification in causal inference. *Biometrics* 58, 21–9.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 15, 413–19.

Heckman, J., Ichimura, H. and Todd, P. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65, 261–94.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Heckman, J. and Robb, R. 1984. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Hirano, K., Imbens, G. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–89.

Hirano, K., Imbens, G., Rubin, D.B. and Zhou, X. 2000. Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics* 1, 69–88.

Holland, P.W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–70.

Hurwicz, L. 1962. On the structural form of interdependent systems. In *Logic, Methodology, and Philosophy of Science, Proceedings of the 1960 International Congress*, ed. E. Nagel, P. Suppes and A. Tarski. Stanford, CA: Stanford University Press.

Imbens, G.W. and Angrist, J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 467–76.

Imbens, G.W. and Rubin, D.B. 1997. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 25, 305–27.

Imbens, G.W. and Rubin, D.B. 2006. *Causal Inference in Statistics and the Medical and Social Sciences*. Cambridge: Cambridge University Press.

Jin, H. and Rubin, D.B. 2007. Principal stratification for causal inference with extended partial compliance: application to Efron–Feldman data. *Journal of the American Statistical Association*.

Lalonde, R. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* 76, 604–20.

Manski, C.F. 2003. *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Mealli, F. and Rubin, D.B. 2002. Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services Outcome Research Methodology* 3, 225–32.

Mealli, F. and Rubin, D.B. 2003. Assumptions allowing the estimation of direct causal effects: discussion of ‘Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status’ by Adams et al. *Journal of Econometrics* 112, 79–87.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Mill, J.S. 1843. A System of Logic. In *Collected Works of John Stuart Mill*, vol. 7, ed. J.M. Robson. Toronto: University of Toronto Press, 1973.

Neyman, J. 1923. On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in *Statistical Science* 5 (1990), 465–80.

Neyman, J. 1935. Statistical problems in agricultural experimentation. Supplement to *Journal of the Royal Statistical Society, B* 2, 107–8 (with discussion). (With cooperation of K. Kwaskiewicz and St. Kolodziejczyk.)

Peters, C.C. 1941. A method of matching groups for experiments with no loss of population. *Journal of Educational Research* 34, 606–12.

Rosenbaum, P.R. and Rubin, D.B. 1983a. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Rosenbaum, P.R. and Rubin, D.B. 1983b. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, B* 45, 212–18.

Rosenbaum, P.R. and Rubin, D.B. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–24.

Rosenbaum, P.R. and Rubin, D.B. 1985. Constructing a control group using multivariate matched sampling incorporating the propensity score. *American Statistician* 39, 33–8.

Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.

Rubin, D.B. 1973a. Matching to remove bias in observational studies. *Biometrics* 29, 159–83. Correction note: 1974. *Biometrics* 30, 728.

Rubin, D.B. 1973b. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 29, 185–203.

Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.

Rubin, D.B. 1975. Bayesian inference for causality: the importance of randomization. *Proceedings of the Social Statistics Section of the American Statistical Association*, 233–9.

Rubin, D.B. 1976. Inference and missing data. *Biometrika* 63, 581–92.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Rubin, D.B. 1977. Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics* 2, 1–26.

Rubin, D.B. 1978. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 6, 34–58.

Rubin, D.B. 1979. Discussion of ‘Conditional independence in statistical theory’ by A.P. Dawid. *Journal of the Royal Statistical Society*, Series B 41, 27–8.

Rubin, D.B. 1980. Discussion of ‘Randomization analysis of experimental data in the Fisher randomization test’ by Basu. *Journal of the American Statistical Association* 75, 591–3.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B. 1990a. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25, 279–92.

Rubin, D.B. 1990b. Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 5, 472–80.

Rubin, D.B. 1991. Practical implications of modes of statistical inference for causal effects. *Biometrics* 47, 1213–34.

Rubin, D.B. 2000. The utility of counterfactuals for causal inference. Comment on A.P. Dawid, ‘Causal inference without counterfactuals’. *Journal of the American Statistical Association* 95, 435–8.

Rubin, D.B. 2002. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2, 169–88.

Rubin, D.B. 2004a. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. Reprinted with new appendices as a Wiley Classic.

Rubin, D.B. 2004b. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* 31, 161–70 (with discussion and reply, 196–8).

Rubin, D.B. 2005. Causal inference using potential outcomes: design, modeling, decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association* 100, 322–31.

Rubin, D.B. 2006a. *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

Rubin, D.B. 2006b. Causal inference through potential outcomes and principal stratification:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
applications to studies with ‘censoring’ due to death. *Statistical Science* 21, 299–321.

Rubin, D.B. 2007. Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In *Handbook of Statistics: Epidemiology and Medical Statistics*, ed. C.R. Rao, J. P. Miller and D.C. Rao. Amsterdam: North-Holland.

Rubin, D.B. and Thomas, N. 1992a. Affinely invariant matching methods with ellipsoidal distributions. *Annals of Statistics* 20, 1079–93.

Rubin, D.B. and Thomas, N. 1992b. Characterizing the effect of matching using linear propensity score methods with normal covariates. *Biometrika* 79, 797–809.

Rubin, D.B. and Thomas, N. 1996. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 52, 249–64.

Rubin, D.B. and Thomas, N. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95, 573–85.

Tinbergen, J. 1930. Determination and interpretation of supply curves: an example. *Zeitschrift fur Nationalokonomie*. Reprinted in *The Foundations of Econometric Analysis*, ed. D.F. Hendry and M.S. Morgan. Cambridge: Cambridge University Press, 1997.

Zhang, J. and Rubin, D.B. 2003. Estimation of causal effects via principal stratification when some outcomes are truncated by ‘death’. *Journal of Educational and Behavioral Statistics* 28, 353–68.

Zhang, J., Rubin, D. and Mealli, F. 2007. Evaluating the effects of job training programs on wages through principal stratification. *Advances in Econometrics* 21.

How to cite this article

Imbens, Guido W. and Donald B. Rubin. "Rubin causal model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000247> doi:10.1057/9780230226203.1466

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Ruggles, Richard (1916 – 2001)

Edward N. Wolff
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

household surveys; Kuznets, S.; life-cycle saving model; national economic accounting; national income accounting; Ruggles, R.; saving and investment; Stone, J. R. N.

Article

Richard Ruggles and his wife, Nancy Ruggles, with whom he co-authored almost all of his work, did pioneering work in the field of national economic accounting. Mr Ruggles attended Harvard for both undergraduate and graduate study, earning his BA in 1939, an MA in 1941 and his Ph.D. in 1942. After earning his doctorate, Richard Ruggles joined the Office of Strategic Services as an economist during the Second World War. He worked for the office in London, where he estimated the production rates of tanks at German factories using photographs of the serial numbers from captured or destroyed tanks. In 1945–6 he was with the US Strategic Bombing Survey in Tokyo and Washington. Mr Ruggles returned briefly to Harvard as an instructor in 1946 before joining the Yale faculty a year later as an assistant professor of economics. He was named an associate professor in 1949 and a full professor in 1954. He was appointed the Stanley Resor Professor of Economics in 1954. He chaired the department of economics from 1959 to 1962. He also conducted research for numerous government agencies and bodies, including the United Nations, the Organization of American States, the Federal Reserve Board, the Bureau of the Census and the National Bureau of Economic Research, as well as the Ford Foundation.

Three principal themes emerge in the work of the Ruggles. The first is the reconciliation of macrodata with microdata. National accounts were developed during the 1930s and 1940s by Simon Kuznets, Richard Stone and Richard Ruggles, among others. The 1950s saw the development of microdata such as the Current Population Survey (CPS) in the United States. In principle, the data contained in microdata sources such as income should be consistent with the corresponding entries in the national accounts. In practice, however, this was seldom the case. Several requirements are put forward by Nancy and Richard Ruggles to fully integrate the two sources. First, the definition of sectors should be the same. For example, while household microdata include only households, the macro ‘household’

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
accounts often include non-profits and group quarter residents, such as those living in college dormitories, nursing homes and prisons. Second, definitions and imputations, such as the treatment of pensions and imputed rent, should be consistent between the two sources. Third, alignment of macro and microdata should not rely exclusively on macro totals. For example, in national accounts personal interest is computed as a residual whereas in microdata the household provides a direct estimate.

The second theme is the synthesis of microdata from several sources. Since household surveys can ask only a limited number of questions, different surveys concentrate on different characteristics. The CPS focuses on demographics and income, while the Consumer Expenditure Survey is very strong on expenditures and the Survey of Consumer Finances concentrates on assets and liabilities. Another problem is that different microdata focus on different parts of the income distribution. The CPS focuses mainly on the middle classes but its income data are weak for the lower and upper tails while the Internal Revenue Service Tax Model, a sample of tax returns, contains detailed income data on the upper tail but limited information on the bottom tail since these families do not file tax returns.

The solution proposed by the Ruggles is a statistical match of microdata. The idea is to merge microdata files which are complementary in terms of the variables they contain or the parts of the distribution that they cover. One such successful match described in their work was between the 1970 Census of Population and the 1969 Tax Model.

A third theme is the importance of institutional sectoring for the analysis of economic behaviour. In several papers, the Ruggles focus on the measurement of savings. Though most theories of savings, such as the life-cycle model, implicitly assume that all savings is done by households, Nancy and Richard Ruggles argue that savings is done by different institutions. In their accounting scheme, they develop separate current and capital accounts for the household, enterprise, and government sectors. They find that the household and the enterprise sectors are each self-financing. On net, the household sector channels almost no financial savings to the enterprise sector, and almost all investment done by enterprises is financed through enterprise savings. These results have wide-ranging implications for theories of savings and investment.

Selected works

1947. *An Introduction to National Income and Income Analysis*. New York: McGraw-Hill.

1993. Accounting for saving and capital formation in the United States, 1947–1991. *Journal of Economic Perspectives* 7(2), 3–17.

1970. (With N.D. Ruggles.) *The Design of Economic Accounts*. New York: NBER.

1975. (With N.D. Ruggles.) The role of microdata in the national economic accounts. *Review of Income and Wealth* 21, 203–16.

1986. (With N.D. Ruggles.) The integration of macro- and micro data for the household sector. *Review of Income and Wealth* 32, 245–76.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1992. (With N.D. Ruggles.) Household and enterprise saving and capital formation in the United States, 1947–1991: market transactions view. *Review of Income and Wealth* 38, 119–64.

1977. (With N.D. Ruggles and E.N. Wolff.) Merging microdata: rationale, practice and testing. *Annals of Economic and Social Measurement* 6, 407–28.

How to cite this article

Wolff, Edward N. "Ruggles, Richard (1916–2001)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000267> doi:10.1057/9780230226203.1467

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Saint-Simon, Claude-Henri (1760 – 1825)

K. Tribe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

industrial society; Saint-Simon, C.-H. de Rouvroy; science of production

Article

Born in 1760 into a noble family, Saint-Simon spent the first 40 years of his life as a soldier and speculator before devoting himself to the study of science and society. Commissioned in 1778, he served with the French forces in the Caribbean and in America, taking part in the Battle of Yorktown (1781). In 1787 he left the army and became associated with a Spanish project for a canal linking Madrid to the Atlantic, in which project he intended to direct the workforce of 6,000 men. The outbreak of the French Revolution prompted his return to France, where he became President of the Municipal Assembly in Falvy, near Péronne. His ambitions for social improvement led him into the purchase of aristocratic and church property from the government, and a financial partnership he formed to this end met with great success. A period of imprisonment in 1793–4 ended with the fall of Robespierre, and in the ensuing period his business interests expanded rapidly. On the proceeds of his financial successes he founded a salon and became a patron of the sciences. During the peace of 1801–2 he travelled to England, and then to Geneva to visit Madame de Stäel. While in Geneva he published his first text of any significance on the reform of society, *Lettre d'un habitant de Genève à l'humanité* (1802). Returning to Paris via Germany, he published further pieces on social reorganization, though his writing was interrupted by the collapse of his personal fortune in 1806. With the support of a former servant, Saint-Simon found time for full-time study and in 1807 was able to publish an introduction to the scientific tasks of the 19th century. In 1814 he was joined in his work by the historian Augustin Thierry, and with his aid assumed the role of a leading publicist for liberal interests. This found its most direct expression in the editing of a series of journals: *L'Industrie* (1816–18); *Le Politique* (1819); and *L'Organisateur* (1819–20). This last brought him some recognition in France and abroad, and led him to the publication of a series of pieces in 1821 under the title *Du Système industriel*, one of his most important works. Despite this success, he felt that his work failed to gain appropriate recognition, and he attempted suicide in 1823. Nursed back to health by a loyal band of followers, he continued writing and studying, his final years being marked

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
increasingly by his interest in religious sentiment as a means of social change. He died in May 1825.
The work of Saint-Simon has been variously described as corporatist, totalitarian and even anarchist. Some care is needed in such characterizations, for the work of Saint-Simon must be distinguished from that of his followers like Comte or Halévy who contributed to the formation of ‘Saint-Simonianism’; and it is also necessary to treat with care the accounts of Saint-Simon to be found in the work of those heavily influenced by him, such as Proudhon or Durkheim. Saint-Simon's own writing was not presented in a systematic manner, nor did he pay much regard to scholarly niceties when developing his ideas. Nevertheless, his name is associated with a number of important ideas which have marked the development of social thought.
Saint-Simon believed that the study of society should be conducted on a scientific basis; that a positive, empirical science of society was both necessary and possible. Society, he argued, was like an organism governed by natural laws; and a ‘healthy’ society was one which is well-organized. Proper recognition of this fact would make possible the reconstruction of society on sound foundations – utopia would become constructible through the application of science to society.
Future society would be *industrial* society, in which ‘general directors’ would ensure that useful work was unhindered and government would therefore administer things, not people. Politics would become the ‘science of production’ – the link to 19th-century socialist thought is here quite evident. ‘Industry’ embraced all kinds of productive activity, and so ‘industrial society’ is one of productive activity in general, not a vision of a technological or manufacturing future. Accordingly, society is primarily arranged into ‘industrious’ and ‘idle’ classes. The future society was not to be a classless one: differences between groups would continue to exist, but would not be a source of social antagonism.
In later life the teachings of Saint-Simon assumed an ever more spiritual cast, a development that was promoted after his death by the group of followers that he had gathered. In the early 1830s the sect had a large number of influential adherents, although a formal organization in the shape of a ‘Church’ soon collapsed. Napoleon III was an open admirer of Saint-Simon’'ss ideas, and the creation of the Crédit Mobilier investment bank, a model commercial bank involved in economic and financial developments, owed much to Saint-Simon's ideas.

Selected works

1964. *Social Organization, The Science of Man and other writings*, ed. F.M.H. Markham. New York: Harper & Row.

1966. *Oeuvres de Claude-Henri de Saint-Simon*. 6 vols, Paris: Editions Anthropos.

1975. *Henri Saint-Simon (1760–1825), Selected Writings on Science, Industry and Social Organisation*, ed. K. Taylor, London: Croom Helm.

How to cite this article

Tribe, K. "Saint-Simon, Claude-Henri (1760–1825)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000002> doi:10.1057/9780230226203.1468

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Samuelson, Paul Anthony (born 1915)

Stanley Fischer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Paul Anthony Samuelson (born in Gary, Indiana, in 1915) has made fundamental contributions to nearly all branches of economic theory. Besides the specific analytic contributions, Samuelson more than anyone else brought economics from its pre-1930s verbal and diagrammatic mode of analysis to the quantitative mathematical style and methods of reasoning that have dominated for the last three decades. Beyond that, his *Economics* (McGraw Hill, first edition, 1948, now in its twelfth edition, the first with a co-author, William D. Nordhaus) has educated millions of students, teaching that economics, however dismal, need not be dull.

Ten eminent economists describe and evaluate his work in their respective fields in Brown and Solow (1983). Arrow (1967) and Lindbeck (1970) provide useful overall reviews. (See also the papers in Feiwel, 1982.)

Samuelson's work consists of *Foundations of Economic Analysis* (1947, reprinted in an enlarged edition in 1983), *Economics, Linear Programming and Economic Analysis* (1958, joint with Robert Dorfman and Robert M. Solow) and his *Collected Scientific Papers* (Volumes I and II, 1966; Volume III, 1972, Volume IV, 1977, and Volume V, 1986). The five volumes of the *Collected Papers* include 388 articles, most of them indeed scientific.

Bliss in his 1967 review of the first two volumes of the *Collected Scientific Papers* comments on the impossibility for anyone other than Samuelson of reviewing his work. The task has not been made any easier by the publication of another two volumes of collected papers, and by the 144-page summary of developments in economic theory since the *Foundations* in the 1983 enlarged edition. Rather than try to be comprehensive, I will describe the major analytic contributions in several areas, ending with macroeconomics where I also discuss Samuelson's views and advice on economic policy. I conclude with a description of his role at and through MIT.

Although the topic-by-topic approach is unavoidable, the man and the economist is more than the sum of his contributions in several areas. The verve and sparkle of his style, the breadth of his economic and general knowledge, the mastery of the historical setting and the generosity of his hyphenated freight-train allusions to predecessors, are unique. Samuelson's presidential address to the American Economic Association (1961: II, ch. 113) is a good sampler. (References to the *Collected Scientific Papers* (*CSP*)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
will give year of publication of the original article where needed, followed by volume number, and chapter and/or page number as needed.)

I Background

Samuelson has provided fragments of his autobiography in ‘Economics in a Golden Age: A Personal Memoir’ (1972: IV, ch. 278), and in biographical articles on contemporaries and teachers. He attended 14 schools, in Gary, Indiana, on the North Side of Chicago, in Florida, and then at Hyde Park High in Chicago. From Hyde Park High he entered the University of Chicago in January 1932, taking his first economics course from Aaron Director. ‘It was as if I was made for economics’ (1972: IV, p. 885). Milton Friedman and George Stigler were Chicago graduate students at the time. Jacob Viner's famous course in economic theory provided the sound non-mathematical microeconomics that any economist needs to truly understand the field (1972: IV, ch. 282; see also Bronfenbrenner, 1982).

In 1935 he moved to graduate school at Harvard, propelled by a fellowship that required him to leave Chicago and attracted, he claims, by the ivy and the monopolistic competition revolution. Samuelson spent five years at Harvard, the last three as a Junior Fellow. It was the time of both the Keynesian and monopolistic competition revolutions, and ‘Harvard was precisely the right place to be’ (1972: IV, p. 889). The teachers he mentions most are Hansen, Leontief, Schumpeter and E.B. Wilson, the mathematical physicist and mathematical economist.

His fellow students make up the larger part of the honour roll of early post-Second World War United States economics (1972: IV, p. 889). Among them was his wife of 40 years, Marion Crawford (1915–78), author of a well-known 1939 article on the tariff. Abram Bergson (particularly his 1938 article on the social welfare function) and Lloyd Metzler are most mentioned among his other fellow students. Samuelson was the dominant presence among the students: Cary Brown in conversation describes the excitement as his papers were analysed and absorbed by the graduate students; James Tobin in correspondence says that the students loved the seminars, where Samuelson could be counted on to put down his seniors with brash brilliance.

The Keynesian revolution and Alvin Hansen had a greater impact on Samuelson's work and attitudes than the monopolistic competition revolution and Chamberlin. Chamberlin is barely mentioned in his reminiscences of Harvard and his only monopolistic competition article appeared in 1967 in the Chamberlin *festschrift* (III, ch. 131). Much of Samuelson's work assumes perfect competition, but none of his macroeconomics or his policy advice gives any credence to the view that the macro-economy is better left alone than treated by active policy (except perhaps his views on flexible exchange rates).

His first published article ‘A Note on the Measurement of Utility’ (1937: I, ch. 20) appeared when he was a 21-year-old graduate student. By 1938 the flow was up to five articles a year, a rate of production that has been maintained with perturbations for half a century. And of course, since 1948 he has produced a new edition of *Economics* almost every three years.

Samuelson moved to MIT as an Assistant Professor in 1940 and has remained there since. Harvard's failure to match MIT's offer at that time has been the subject of much speculation. Samuelson himself has been eager to find excuses for Harvard (1972: IV, ch. 278, footnote 11, p. 896). His is not the best position from which to judge or to write freely; he has noted that academic life, and by implication the chairman of the Economics Department, Burbank, one of the few of whom Samuelson speaks harshly in print, were not innocent of anti-Semitism in that pre-Second World War era. Burbank was a political

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
power in the Department and University. His attitude to mathematical economics can be gauged by the fact that indifference curves were outlawed in the introductory course he supervised.
It is hard to believe that even the Harvard of 1940 would have been unable to find room for an economist of Samuelson's already recognized stature unless a non-academic reason or reasons stood in the way. Among those reasons were anti-Semitism, his then brashness, and his brilliance: indeed Schumpeter is rumoured to have told his colleagues that it would have been easier to forgive their vote if it had been based on anti-Semitism rather than on the fact that Samuelson was smarter than they were.
Samuelson has been at MIT since 1940, virtually without a break. Except for a few months away on a Guggenheim, he has taken time off only in Cambridge, Mass. He proudly claims that he has never been in Washington for as long as a week – though he was a major adviser to President Kennedy. His only departure from academic economics came in 1944–5, when he worked at MIT's Radiation Laboratory. He became one of 12 MIT Institute Professors in 1966.
He has gathered all the honours the profession can offer: the first John Bates Clark medal (1947); the second Nobel Memorial Prize in Economics (1970); he has been President of the American Economic Association (1961), the Econometric Society (1951), and the International Economic Association (1965–8); and he has been awarded numerous other prizes and honorary degrees.
Although many graduate students have passed through his classes and been profoundly affected by him, there is no Samuelson school of economics, no overarching grand design for either economics or the world that is uniquely his. It is for that reason that his contributions have to be discussed field by field. The nearest that he has come to proclaiming a vision is in the *Foundations*.

II Foundations of Economic Analysis

Foundations, published in 1947, is based on Samuelson's 1941 David Wells prize-winning dissertation, *Foundations of Analytic Economics*, subtitled ‘The Observational Significance of Economic Theory’. Its themes are partially described by the subtitle and by the motto from J. Willard Gibbs, ‘Mathematics is a language’. The thesis, dated 1940, is very close in content to the *Foundations*.
The *Foundations* in places claims to be an attempt to derive empirically meaningful comparative equilibrium results from two general principles, that of maximization, and Samuelson's *correspondence principle*. The correspondence principle states that the hypothesis of dynamic stability of a system yields restrictions that make it possible to answer comparative equilibrium questions.
The maximizing theme recurs in Samuelson's 1970 Nobel Prize lecture, ‘Maximum Principles in Analytic Economics’ (III, ch. 130). The point is not the now common view that only models in which everyone is relentlessly maximizing are worth considering. Rather it is that the properties of the maximum (for instance, second-order conditions) usually imply the comparative static properties of the system. Samuelson also invokes the generalized Le Chatelier principle, which loosely interpreted states that elasticities are larger the fewer constraints are imposed on a system. Analogies from physics (and biology) figure prominently in Samuelson's analytic methods and explanations of his results.
The correspondence principle was intended to do for market or macroeconomic comparative statics what maximization did for the comparative statics of the individual or firm. The principle can be useful when the analyst knows something about the dynamic behaviour of a system, but as noted by Tobin (1983), is ambiguous in that different dynamics may be consistent with the same steady-state behaviour.
The simplest example of the ambiguity can be seen in a demand–supply diagram where the supply curve

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
is negatively sloped. Whether a tax on the good will increase or reduce price depends on which curve is more steeply sloped. Whether the market is stable or not depends on the same fact and whether quantity or price rises in response to excess demand – that is, whether dynamics are Marshallian or Walrasian.
In the introduction to the 1983 enlarged edition, Samuelson records correctly that the *Foundations* was better off for not sticking to its narrow themes. Substance keeps breaking in on the methodology. The treatment of the theory of the consumer and firm, developed in detail, does not differ in essence from that of Hicks in *Value and Capital*. But where Hicks hides the mathematics in appendices, Samuelson flaunts his in the text. Nonetheless Samuelson takes pains to provide economic insight, including interpretations of Lagrange multipliers as shadow prices. These portions of the *Foundations* apparently existed in 1937–8 and were written independently of *Value and Capital* (Bronfenbrenner, 1982, p. 349), though not of course of Hicks and Allen (I, ch. 1, p. 4).
The theory of revealed preference (see below) receives prominence, as do two chapters on welfare economics, and in Part II chapters on the stability of general equilibrium. A few pages on money in the utility function (pp. 117–24) remain authoritative. The mathematical appendices on maximization and difference equations have been useful despite an elliptical style that leaves many steps to be filled in by the user.
Samuelson's thesis is dated 1940: *Foundations* is the work of a 25-year-old. There are signs of youth in the eagerness to proselytize for the new mathematical faith and its overreaching in trying to impose an entirely coherent theme on the material. But the book bears the unmistakable mark of the master, in command of the economics of his material, at home with technique, and most remarkably for a young man in a hurry, thoroughly familiar and patient with the literature. It is, as Schumpeter no doubt remarked, a remarkable performance.

III Consumer theory and welfare economics

Samuelson's first published paper (1937: I, ch. 20) set up a finite horizon continuous time intertemporal optimization model of a consumer with additively separable utility function and exponential discounting, and derived the result that the profile of consumption is determined by the relation between the interest rate and rate of time preference. The focus is however the measurability of utility.
The theory of revealed preference, his major achievement in consumer theory, made its unnamed appearance in 1938 in ‘A note on the pure theory of consumer's behaviour’ in *Economica* (I, ch. 1; see also Houthakker, 1983, and Mas-Colell, 1982, for exceptionally lucid accounts). The purpose was to develop the entire theory of the consumer free of ‘any vestigial traces of the utility concept’ (I, p. 13). Rather than postulate a utility function or, as Hicks and Allen had done, a preference ordering, Samuelson imposed conditions directly on the choices made by individuals – their preferences as revealed by their choices. The key condition was the weak axiom of revealed preference, applying to choices made in two situations, say zero and one. With prices and quantities of goods j, j=1, …, n in situation i given by p^i_j and x^i_j, the axiom is

$$\sum_j p_j^0\left(x_j^1 - x_j^0\right) \leq 0 \text{ implies } \sum_j p_j^1\left(x_j^1 - x_j^0\right) \leq 0$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In words, if the individual chooses consumption bundle zero when he could have chosen the bundle one, he will not choose one when zero is available.

This minimal condition of consistency is shown to imply most of the conditions on demand implied by utility theory. But the symmetry and negative definitiveness of the Slutsky matrix could not be established using the weak axiom. Equivalently, the issue was the so-called integrability of demand functions, with the question being whether the preference map could be recovered given enough observations on the individual's choices. Houthakker (1950) solved the problem, by proposing the strong axiom of revealed preference, namely that for any finite string of choices in which B is revealed preferred to A, C is revealed preferred to B, …, and Z is revealed preferred to Y, then A is *not* revealed preferred to Z. In this case, and given appropriate continuity conditions on demand, the demand functions are integrable and an entire preference map, satisfying the Slutsky conditions, can be recovered from the individual's choices.

The full equivalence between the properties of the demand functions of an individual and the preference ordering is the leading example of Samuelson's definition of the operational or observational significance of economic theory. Samuelson regards a theory as meaningful if it is potentially refutable by data. A single consumer could make a succession of choices that contradict the strong axiom. But the theory is not operational in the sense that a modern econometrician would want it to be: it does not apply to aggregate data, nor, in the form in which Samuelson left it, does it apply to choices that are made in chronological time.

Revealed preference links the theory of demand, index numbers, and parts of welfare economics. The link between demand and index number theory comes in the *Foundations*' (pp. 147–8) recognition that the fundamental index number problem is to deduce from price and quantity information alone whether an individual is better off. Using the weak axiom, Samuelson demonstrates the conditions under which, in a comparison of two situations, it is possible to say whether an individual is better off in one (*Foundations*, pp. 156–63). He argues that index numbers add no information on the essential question and indeed may be positively misleading in tempting the observer to attach significance to the numerical scale of measurement.

A similar concern no doubt motivates Samuelson's long-standing hostility to the use of consumer surplus measures. He has frequently argued that there is no need for the concept. He asserts in *Foundations* (p. 197) that there is no need for consumer's surplus in answering, for example, the question of whether Robinson Crusoe, a socialist state, or a capitalist one, should build a particular bridge. That view may have been moderated over the decades: the 1985 Samuelson–Nordhaus *Economics* (p. 418) states that the concept 'is extremely useful in making many decisions about public goods – it has been employed in decisions about airports, roads, dams, subways, and parks' (bridges are conspicuously absent).

The revealed preference axiom comes into play too in Samuelson's 'Evaluation of Real National Income' (1950: II, ch. 77), a largely negative report on the then new welfare economics that attempted to deduce from aggregate data criteria that would make it possible to say whether society was better off in one situation than another. Taking, as he has since 1938, the viewpoint that a Bergsonian social welfare function is the best way of understanding social welfare issues, Samuelson showed that no index-number type national income comparison between situations A and B could reveal whether society's feasible utility possibility frontier (a useful Samuelson innovation, apparently simultaneously invented by Allais)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in A lies uniformly outside that of B. And, he argued, we could claim situation A is better than B only if that is the case.

In the *Foundations* (chapter 8) Samuelson draws extensively on the Bergsonian social welfare function to elucidate definitively the notion of Pareto optimality and the 'germ of truth in Adam Smith's doctrine of the Invisible Hand' (*Foundations*, 1983 edn, p. xxiv). Arrow (1967) is critical of Samuelson's failure to look behind the social welfare function, and of his failure to link it to actual policy decisions. Similar sentiments are conveyed along with a more complete evaluation of Samuelson's welfare economics in Arrow (1983). Samuelson (1967: III, ch. 167) asserts that the Bergson Social Welfare Function and the Arrow Constitution Function are distinct concepts, though the argument is difficult to follow.

The expected utility theorem shows Samuelson wrestling for decades with his doubts over the independence axiom (I: ch. 12, 1950; ch. 13, 1952; ch. 14, 1952; *Foundations*, 1983, pp. 503–18). Despite his tentative 1983 acceptance of the expected utility formulation, he notes with approval Machina's (1982) development of expected utility without the independence axiom. Of course, these doubts have not kept him from making creative use of the expected utility approach in models of portfolio choice and finance.

IV Capital theory

The theory of capital and growth sections of the first four volumes of *CSP* account for 38 papers, the largest single category. Although capital theory is the branch of economics most vulnerable to Samuelson's comparative technical advantage, and although both his earliest papers are placed in that category in *CSP* (1937: I, ch. 17, ch. 20) the output in this area is concentrated in *CSP* III, covering the years from 1965 to 1970. Solow (1983) provides a fine review of this part of Samuelson's research, some of which he co-authored.

Among the early papers, the 1943 Schumpeter *festschrift* contribution 'Dynamics, Statics, and the Stationary State' (I, ch. 19) discusses the economics of the steady state and the possibility of a zero interest rate. Samuelson argues that a steady state with a zero real interest rate is possible if the rate of time preference of the infinitely lived individuals is zero; he has in mind a situation in which the marginal product of capital can be driven to zero. In this article (I, p. 210), as in his first paper (I, p. 216), Samuelson makes highly favourable reference to Ramsey, in contrast to the famous unflattering 1946 remark (II, p. 1528). The well-known argument that a zero rate of interest is impossible because income generating assets would have an infinite value is rejected, on the ground that an infinite value is not a problem since assets could trade against each other at finite price ratios. Some second thoughts are presented in a 1971 paper (IV, ch. 217); curiously, Samuelson discusses the Schumpeter issue entirely in a model with infinite horizon maximizers rather than in an overlapping generations framework.

The modern contributions in *CSP* I include the famous 1958 consumption loans model, which will be examined in the macroeconomics section, and the surrogate production function (1962: I, ch. 28). As Solow (1983) notes, much of the capital theory in *CSP* is related to developments in Dorfman, Samuelson and Solow (1958), which itself grew out of a 1949 Samuelson three-part memorandum for the Rand Corporation.

Notable among the contributions is a variety of turnpike theorems. A turnpike theorem is conjectured in the 1949 Samuelson memorandum, and fully worked out in the 1958 volume. The theorem states that for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
any accumulation programme, starting from an initial vector of capital goods, and with specified terminal conditions, as the horizon lengthens the optimal programme spends an increasing proportion of its time near the von Neumann ray; more generally in problems with intermediate consumption, the economy spends time near the modified golden rule. Several of the papers in the capital and growth section of *CSP* III contain turnpike theorems. A periodic turnpike result is reported in 1976 (IV, ch. 224).
The surrogate production function was an attempt to justify the aggregate production function as being consistent with an underlying model with heterogeneous capital goods and production techniques, and one type of labour. The article names and uses the factor price frontier, noting that it had been used earlier by others, including himself (in 1957: I, ch. 29). Samuelson shows that a downward sloping factor price frontier is traced out in a competitive multi-capital goods multi-technique economy, with higher steady state wages accompanying a lower steady state interest rate. Further, this frontier has the same properties as in the one-sector model, with the slope of the factor price frontier equal to the capital labour ratio. The theorem is correct, but as noted by Solow (1983), the conditions for it to obtain are special.
Under more general conditions, the famous reswitching result may occur in which a given technique of production that had been used at a low interest rate comes back into use again at a high interest rate (see the November 1966 *Quarterly Journal of Economics*). Reswitching implies that the one-sector neoclassical production function cannot be viewed as a general ‘as if’ construct that describes the behaviour of economies with several techniques of production. Cambridge (England) critics of neoclassical capital theory viewed reswitching as a confirmation of the view that marginal productivity had nothing to do with distribution, since the same techniques of production might be used with two (or many) different distributions of income. Various criticisms are offered by Robinson (1975) and responded to with forebearance in *CSP* (1975: IV, ch. 216).
Samuelson started the surrogate production function article by denying the need for any concept of aggregate capital. That position would be strengthened by the reswitching result. However, as with so many useful constructs in economics, the concept of aggregate capital has survived the demonstration that its validity may be limited. Neither Samuelson nor other neoclassics have been constrained by reswitching from using one-sector production functions or marginal productivity factor-pricing conditions.
The property that the slope of the factor price frontier is equal to the capital–labour ratio is one example of the duality between price and quantity that Samuelson began to emphasize in the *Foundations* and has used repeatedly since. *Foundations* (p. 68) contains the Roy's Identity envelope condition that the derivative of the minimized cost function of the firm with respect to the wage of factor *i* is the demand for factor *i*. It also provides shadow price interpretations of Lagrange multipliers. Samuelson has used duality in optimal growth and linear programming problems (‘Market mechanisms and maximization’, 1949: I, ch. 33, is a gem) and in *CSP* (1965: III, ch. 134).

V Dynamics and general equilibrium

Chapters IX through XI of *Foundations* cover stability analysis and dynamics, in both individual markets and the economy at large. The basic assumption of this dynamics is the ‘law of supply and demand’ that price rises in response to excess demand.
The impetus for the multi-market analysis came partly from Hicks's *Value and Capital* (1939) discussion

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of stability, in which there is no explicit dynamical system. The Samuelson approach is general equilibrium, though it does not start from the primitives of endowments. As Hahn (1983) notes, the underlying microeconomics is not specified. Samuelson nonetheless set the agenda of the next 15 years for the study of dynamics in a more explicitly general equilibrium framework, and most important, in a framework in which the issue of stability is precisely posed.

Explicit use of the law of supply and demand in theoretical work has fallen out of favour, though the Phillips curve can be interpreted as using that approach. The monopolistic competition wing of macroeconomics prefers to model price setting by firms and workers explicitly rather than rely on an auctioneer, and the equilibrium approach assumes prices are continuously at market-clearing levels. The older approach is used in disequilibrium macroeconomics, but is typically regarded as suspect.

Samuelson has not been a general equilibrium theorist in the sense of one striving for maximum generality. He has been general equilibrium in the sense opposed to partial equilibrium: he frequently works with models of the whole economy, in growth and capital theory, in trade and macroeconomics, and in his excursions into the history of thought.

The most micro-oriented of these general equilibrium contributions are the non-substitution theorem (1951: I, ch. 36) and factor–price equalization. The non-substitution theorem was presented at a 1949 conference, and was obtained independently by Samuelson and Georgescu-Roegen (I, p. 521). Consider an economy where labour is the only primary factor, and where goods are used either for consumption or as input into the production of other goods. Suppose the production function for each good is neoclassical, permitting substitution among factors of production, but there is no joint production.

The theorem is that relative prices in this economy are independent of demand, that is, are determined on the supply side alone. There is a single least cost way of producing each good, where cost is determined by direct and indirect labour requirements. Hahn (1983) provides a clear account of the theorem, and generalizations to dynamic systems with capital (1961: I, ch. 37). The question in the system with capital is whether, given the interest rate, the relative price structure is unique. Conditions for uniqueness are discussed in Hahn. The link with the surrogate production function, published at about the same time, is clear. The nonsubstitution theorem is used also in Samuelson's discussions of Ricardo (1959: I, chs 31, 32).

VI International trade

‘Our subject puts its best foot forward when it speaks out on international trade’ (1969: III, p. 683), and some of Samuelson's best-known contributions are undoubtedly in this field. Jone's 1983 article describes Samuelson's considerable impact on trade theory: on the gains from trade; the transfer problem; the Ricardian model; the Heckscher–Ohlin–Samuelson model; and the Viner–Ricardo model.

Earliest among the well-known contributions is the 1941 Stolper–Samuelson result (II, ch. 66), which uses the two-sector, two-country Heckscher–Ohlin model with identical production functions in the two countries to analyse the effects of the opening of trade, or the imposition of a tariff, on the wage. The result is that protection will benefit the factor that is relatively (to the other country) scarce. Or, the opening of trade benefits the relatively plentiful factor. But the paper contains more than that result. As Jones (1983) notes, it introduces the basic elements of Heckscher–Ohlin theory for small-scale trade models – and those models were the analytic core of real trade theory for decades.

Stolper–Samuelson flags the issue of factor–price equalization, the question of whether trade in goods

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
alone can produce the factor price equalization that would obtain if factors were freely mobile. Ohlin claimed that trade would cause a necessarily incomplete tendency to equalization. Samuelson (1948 and 1949: II, ch. 67; 68) showed in the Heckscher–Ohlin context conditions under which equalization would be complete: identical production functions in the two countries, no factor-intensity reversals, and similarity of the ratio of endowments (so that countries are not specialized in production). The paper was remarkable and surprising, and did not suffer from the happy coincidence that a 1933 Abba Lerner contribution rediscovered by Lionel Robbins had independently reached the same conclusions in a similar model.

Factor price equalization in more generality is considered in the famous 1953 paper ‘Prices of factors and goods in general equilibrium’ (II, ch. 70), which caused a substantial literature including Gale–Nikaido (1965). It is striking that many of Samuelson's famous papers led to prolonged discussion of the exact conditions needed for his particular results to obtain: he opened more doors in economics than he closed.

The transfer problem is an old issue in the literature that arose in the 1920s, after Second World War, and arises again in contemplation of the world debt crisis. Samuelson's 1952 and 1954 papers (II, chs 74, 75) are classics in this extensive literature, on the issue of whether a transfer from one country to another (such as German reparations) is likely also to worsen the terms of trade of the country making the transfer, which Samuelson describes as the orthodox presumption. In the modern context the orthodox presumption would be that the developing countries will have to suffer a terms of trade loss to run current account surpluses to reduce their indebtedness. Samuelson typically argues that there is no presumption about the terms of trade shift, though the orthodox presumption is more likely to hold where there are non-traded goods or impediments to trade (1971: III, ch. 163).

Samuelson's contributions to trade theory are classics: the contributions are basic, the models are tractable and fecund, the problems come from the real world as well as the literature, the articles continue to reward the reader. And they continue to be read.

VII Finance

Despite his long-time personal interest in capital markets, Samuelson's contributions to finance theory started only as he turned 50. These papers are concentrated in *CSP* III and IV; the earlier ones are self-reviewed in ‘Mathematics of speculative price’ (1972: IV, ch. 240). Merton (1983) describes and evaluates six of Samuelson's favourite papers in finance, broadly defined to include his 1952 paper on expected utility and the independence axiom (I, ch. 14).

The two most important papers are ‘Proof that properly anticipated prices fluctuate randomly’ (1965: III, ch. 198) and ‘Rational theory of warrant pricing’ (1965: III, ch. 199). ‘Proof …’ provides a first precise formulation of the consequences for speculative prices of market efficiency. The theorem describes the behaviour of the current price of a commodity for delivery at a given future date, for example June 1990 wheat. Assuming that speculators do not have to put up any money to enter the contract, the result is that the market price should be the expectation at each date of the June 1990 wheat price. Given rational expectations, there is no serial correlation in the changes in price. Hence ‘properly anticipated prices fluctuate randomly’.

Samuelson says of this theorem, which is now entirely basic: ‘This theorem is so general that I must

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
confess to having oscillated over the years between regarding it as trivially obvious (and almost trivially vacuous) and regarding it as remarkably sweeping. Such perhaps is characteristic of basic results' (III, p. 186).

Note what the theorem does not say, using the exchange rate as the example. The theorem is not that the exchange rate fluctuates randomly; predictable inflation or predictable business cycle fluctuations can cause predictable movements in the exchange rate. Rather it is the current price of foreign exchange at a *given* future date that fluctuates randomly. The notion that efficiency produces random motion is itself fascinating. But far more important is the restriction on empirical behaviour implied by efficiency that Samuelson derives in a well-defined context. Testing for efficiency of speculative markets has become a major industry.

'Rational theory of warrant pricing' missed its target, but it is as Merton (1983) remarks, a near miss. Samuelson had pursued option pricing for well over a decade. He supervised Kruizenga's 1956 MIT dissertation on the topic, and was familiar with Bachelier's 1900 continuous time stochastic calculus calculation of rational option prices. Samuelson derived a partial differential equation for the option price that depends, among other variables, on the expected return on the stock and the required return on the option. The remarkable feature of the Black–Scholes solution to the problem is that the rational price of the warrant does not depend on the expected return on the stock, but rather on the risk-free rate. Nonetheless, the Samuelson differential equation can be specialized to the correct Black–Scholes equation.

Other contributions to finance theory include papers on diversification (1967: III, ch. 201), and on conditions under which mean-variance analysis can be justified (1970: III, ch. 203) – with continuous time models providing the best argument for the procedure.

VIII Macroeconomics

All the Samuelson contributions described to this point are firmly neoclassical. His work in macroeconomics presents a more mixed picture. I take up in turn the early multiplier–accelerator model, which is not at all price-theory oriented, the neoclassical synthesis, Samuelson the policy adviser and commentator, and the entirely neoclassical consumption loans model.

The multiplier – accelerator model

In a 1959 note (II, ch. 84) on the multiplier–accelerator model, Samuelson describes his contribution as being the algebraic generalization of a numerical example of Alvin Hansen's. The model (1939: II, ch 82, 83) is a simple one in which current consumption is proportional to lagged output and investment is determined by the difference between current and lagged consumption (the accelerator). This implies a second-order difference equation, which can generate asymptotic or oscillatory damped approaches to equilibrium, or oscillatory or non-oscillatory explosive paths for output. Although Frisch and Slutsky had already written on the ability of stochastic difference equations to mimic cycle-like behaviour, Samuelson does not – except for a quotation from J.M. Clark that receives little emphasis – link his second-order equation with a stochastic forcing term.

Samuelson (1939: II, p. 1111), while emphasizing the simplicity of the algebraic analysis, argues for the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
empirical importance of the accelerator. This judgment has held up over time as flexible accelerator effects continue to feature strongly in modern estimated investment functions. From the theoretical point of view, the multiplier–accelerator model is interesting for the lack of concern over microfoundations. Where a 1980s macroeconomist might agonize over the microfoundations of the consumption function, over the accelerator, or over the impact of rational expectations of future output on investment, Samuelson proceeds constructively with a simple implicitly fix-price model. The famous 45-degree diagram popularized in *Economics* – and for several editions on the cover – forcefully emphasizes Samuelson's view that aggregate demand is the key determinant of output.

> In the 1940 ‘Theory of pump-priming reexamined’ (II, ch. 85) he stipulates the basic features of the private economy forming the environment within which governmental action must take place … – (1) The economic system is not perfect and frictionless so that there exists the possibility of unemployment and under-utilization of productive resources …

This view pervades Samuelson's macroeconomics. Indeed, when asked recently his view of the causes of wage and price stickiness, he replied that he decided 40 years ago that wages and prices were sticky, that he could understand the behaviour of the economy and give policy advice on that basis, that he had seen nothing since then to lead him to change his view on the issue – and that he had not seen a payoff to researching the question.

He was of course aware of the issues. An abstract of a paper presented at the 1940 meetings of the Econometric Society (II, ch. 88) describes a totally modern discussion of the question of whether general involuntary unemployment is impossible in a world of price flexibility. His penetrating 1941 review of Pigou's *Employment and Equilibrium* (II, ch. 89) outlines a simple classical model in which price flexibility through its effects on aggregate demand produces full employment even with a constant real wage. This is not however Pigou's model; according to Samuelson, Pigou adopts a model in which money wage flexibility is an alternative to active monetary policy. Samuelson never regarded the Pigou effect as being of real world significance (1963: II, ch. 115).

The neoclassical synthesis

Tobin (1983, p. 197) describes the neoclassical synthesis as Samuelson's greatest contribution to macroeconomics. The synthesis is outlined in articles in the early 1950s (1951: II, ch. 98; 1953: II, ch. 99; 1955: II, ch. 100) and developed in successive editions of *Economics*. It argues that monetary and fiscal policy can be used to keep the economy close to full employment, and the monetary–fiscal mix can be used to determine the rate of investment.

The synthesis represents the views of mainstream macroeconomics in the 1950s and 1960s, and perhaps in the 1970s and even the 1980s. Its activist spirit was evident in the Kennedy administration. Its acceptance must have been helped by the widespread use of Samuelson's *Economics* and by the many clones that preached its message.

Perhaps the most notorious component of the neoclassical synthesis is the 1960 Samuelson–Solow ‘Analytic aspects of anti-inflation policy’ (II, ch. 102), which presents a United States Phillips curve. This article is frequently cited as containing the view that the Phillips curve presents society's long-run

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
tradeoffs between inflation and unemployment.
It does not. The paper starts by discussing the difficulties of distinguishing cost–push from demand–pull inflation. Samuelson and Solow then plot the scatter of percentage changes in average hourly earnings in manufacturing against the unemployment rate (the years plotted are not specified, but include the 1930s). The discussion that follows considers alternative points on the Phillips curve as policy choices for the next few years. But the authors warn explicitly that the discussion is short-term, and that it would be wrong to think that the menu of choices represented by the Phillips curve 'will maintain its same shape in the longer run. … [I]t might be that … low-pressure demand would so act upon wage and other expectations as to shift the curve downward in the longer run …' (II, p. 1352). This is though hardly a clear demonstration of the vertical long-run Phillips curve – for Samuelson–Solow suggest that low demand might also cause the Phillips curve to shift up (a notion that many in Europe now find entirely believable) – but it is clear evidence that the authors were not guilty of believing the Phillips curve would stay put no matter what. In conversation, Samuelson has said that he was always the Kennedy administration pessimist about the long-run Phillips curve tradeoff.

The policy adviser and commentator

Samuelson has long taken an active part in economic policy debates, through Congressional testimony, as consultant to the Treasury and the Fed, in his *Newsweek* column that ran every three weeks from 1966 to 1981, in other newspaper columns, public addresses, advice to candidates and Presidents, and in contributions at academic conferences and in symposia.
His views reflect the neoclassical synthesis, a disdain for rules rather than discretion in determining policy, and an almost shameless eclecticism. He knows the macroeconomic numbers and can speak the language of policy discussions. He is a cautious forecaster, rarely committing numbers to print, preferring to decide on which side of the consensus to place his bets. His 1941 consumption function remains his only econometric work (II, ch. 87); he has said that the major disappointment in economics in the last 40 years has been the failure of econometric evidence to settle disputes.
Macroeconomics is Samuelson's primary applied economics field, with finance the second. He keeps up with the current state of the macroeconomy, drawing on forecasts and empirical work of others. He is sceptical of individual forecasts though a law of averages permits him to put some trust in the mean or median forecast. His eclecticism makes his policy views less exciting than those of economists with a strong view of the way the world works – but he has never sought to be interesting rather than right. (This despite his 1962: II, ch. 113, p. 1509) comment on John Stuart Mill: 'It is almost fatal to be flexible, eclectic, and prolific if you want your name to go down in the history books ….')
Nonetheless, Samuelson's implied attitude to the applications of economics gives pause. As Arrow (1967) notes, his work reveals ambivalence about the relevance of neoclassical price theory. He shows no great faith that his microeconomics can be applied to the real world. No doubt comparative advantage plays a role in that attitude. But the theoretical sophistication he brings from microeconomics does not distinguish his macroeconomic policy advice and forecasting from that of the pack; his neoclassical training is not seriously used in Samuelson's applied macroeconomics. *Economics* may be evidence however that he values simple microeconomics.

The consumption loans model

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the classroom Samuelson has confessed that among his many offspring the consumption loans model (I, ch. 21) is his favourite. The affection is amply rewarded: within macroeconomics the two-period lived overlapping generations structure of the model has been used in countless papers in which a tractable framework with an explicit time structure is needed. The original consumption loans model examined the role of money or bonds as institutions for making Pareto-improving trades feasible; the structure has been used subsequently to examine the dynamics of capital accumulation, the burden of the debt, Ricardian equivalence, social security, the role of money, the effects of open market operations, intertemporal substitution of leisure, labour contracts, government financial intermediation, and more.
The set-up for the original model is one in which people live two periods, with utility functions defined over consumption in the two periods. Each young person receives an endowment of one nonstorable chocolate in period 1. In the absence of trade each person could consume only in period 1. Trades are possible in which the current young give part of their chocolate to the current old in return for chocolate to be received next period from the then young. But there is no double coincidence of wants, no direct way of making the bilateral trades.
Now comes the ostensible point of the model: the social contrivance of money makes trade possible, and its introduction is a Pareto-improving change given the pattern of endowments. The consumption loans model has been much criticized as a model of money, because it implies the velocity of circulation is one per generation. Equivalently, the criticism is that the model describes money as effectuating integenerational transactions whereas in practice other assets, such as bonds, serve that role. (Patinkin 1983 discusses the consumption loans model as a basis for monetary theory and also Samuelson's excursions into the history of monetary thought.)
This is certainly correct. But the significance of the consumption loans model is not its rationale for the existence of money. Rather the model has been so influential and popular because it provides a simple *tractable* general equilibrium structure for modelling intertemporal problems with life-cycle maximizing individuals. The earlier examples prove how easily the general structure can be adapted. It can also be adapted to more periods of life (in the original article Samuelson extended lifetimes to three periods), with 50 period lifetime models being easily solvable on computers. Its strength lies in the elegance and robustness with which it captures the essential point that finite lived individuals exist in an infinitely lived economy (we are each but not all dead in the long run).
Nearly 30 years after the consumption loans model was first published, Malinvaud (1987) drew attention to the little-noticed earlier discovery and extended development of the overlapping generations model by Allais (1947). No doubt Samuelson's eminence and location in the United States, as well as publication in *Journal of Political Economy* rather than a book, had much to do with his independent discovery providing the impetus for the exploitation of this extraordinarily useful model – though even so, it took several years before the overlapping generations structure found its way into common use.

IX Samuelson and MIT

MIT had famous economists before Samuelson: Francis A. Walker, third president of MIT (1881–99) and first president of the American Economic Association (1886–92), and Davis R. Dewey, president of the AEA (1909) and editor of the *American Economic Review* (1911–40). But the modern era, in which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the Department of Economics has risen to worldwide prominence within the profession begins with the arrival of Samuelson in 1940. Brown and Solow (1983) describe the MIT Department of the 1930s, and the transformation that nearly began in 1941 after Samuelson arrived and the first Ph.D. class, including Lawrence Klein and George P. Shultz, was about to get under way. Second World War intervened, and it was only in the late 1940s and early 1950s that the faculty and the Ph.D. programme reached full strength.

The MIT department and Ph.D. programme have been consistently among the best in the world since the early 1960s. The names of the faculty members are well known. Equally remarkable is the collection of eminent economists who are MIT Ph.Ds, whose names are legion but whom it would only be invidious to begin to list.

Samuelson's role in this success was pivotal but not domineering. His research habits (including sheer hard work), the open-door policy for students (a lesser burden for someone of whom the students were in awe than for others) and fellow faculty, his absolute refusal to use authority instead of reason in faculty meetings, his zest for conversation about economics, economists, and all else, made him a role model for a department where cooperation and friendliness have been extraordinary. He helped shape the department but he did not dictate its shape; he told one of his young co-authors that as a young man he decided that at age 40 he would stop taking initiatives in the department, at 50 he would venture an opinion only when asked, and at 60 would stop attending faculty meetings. Within the margin of error allowed to economists, he held to that resolution.

Samuelson the teacher played a lesser role. His world-wide fame (and that of other faculty members) doubtless was a major reason many of the outstanding students were there. But, at least in the last two decades, he supervised relatively few theses. His method of supervision was ideally suited to better students, for he would ask broad questions and give general guidance rather than involve himself in details.

His classroom lectures in the period 1966–9 when I heard them were not a model of organization. His advanced theory lectures were given in the first class of the day and it was always possible to tell whether the traffic had been bad that day by whether his hand-written mimeographed lecture notes were available at the beginning of the lecture or only later. The time until the notes arrived was taken up by stories setting the historical background for the problem, and anecdotes about the protagonists. The day he lectured en route to deliver his contribution to the Irving Fisher *festschrift* (1967: III, ch. 184) was especially memorable, though word filtered back from New Haven that his Yale audience was less than enchanted by the stories. His students were not surprised to find in his Nobel lecture (1970: III, ch. 130) both that he had been warned that the lecture was to be serious, and that he started a less than serious story with that warning.

His lectures were simply not designed for the novice. But they were superb for those with some background. He explained finer points, threw out open questions, made unexpected connections between topics, and communicated the zest with which he approaches economics.

X Concluding comments

Among the missing from this list of Samuelson's contributions are his work in the history of thought (where he has been more interested in clarifying analysis than in evaluating contributions), his methodological articles, the famous public goods theorem, the recent work on mathematical biology, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
informative and entertaining biographies of contemporaries, the frank self-evaluations, and *Economics*. The extraordinary success of *Economics* is something of a mystery, for the book is not easy – as witness the fact that simpler texts that follow Samuelson's structure have found a large market. *Economics* is a multi-level book that in its appendices, footnotes and allusions goes far beyond elementary economics. Depending on what students retain from their economics courses, *Economics* may have done much to raise the level of public discourse about economic policy.

Samuelson's self-evaluations, as in ‘Economics in a Golden Age’ (1972: IV, ch. 278), must have shocked many readers. The typical self-effacing scientist does not include stories of Newton and Gauss in his intellectual autobiography. Reflection leads to a different perspective: it would have been easy for Samuelson not to ‘tell the truth and shame the devil’ (1972: IV, p. 881). But how much more interesting it is to have the account of how Samuelson views his own achievements.

Samuelson was described in 1967 as ‘knocking on the door … of the pantheon of the greats …’ (Seligman, p. 160). He may have been let in by now. But the final word has to be left to Franco Modigliani, who, after the speeches at the 1983 party at which Samuelson was presented with the Brown–Solow *festschrift*, walked over to the seated Samuelson, wagged his finger at him, and said ‘You’, and after a pause ‘You have enriched our lives.’

Selected works

1947. *Foundations of Economic Analysis*. Cambridge, Mass: Harvard University Press. Enlarged edn, 1983.

1948. *Economics, an Introductory Analysis*. 12th edn, co-author William Nordhaus, New York: McGraw-Hill, 1985.

1958. (With R. Dorfman and R.M. Solow.) *Linear Programming and Economic Analysis*. New York: McGraw-Hill.

1966–86. *The Collected Scientific Papers of Paul A. Samuelson*, vols I–V. Vols I and II, ed. J.E. Stiglitz, Cambridge, Mass.: MIT Press, 1966; vol. III, ed. R.C. Merton, Cambridge, Mass. and London: MIT Press, 1970; vol. IV, ed. H. Nagatani and K. Crowley, Cambridge, Mass. and London: MIT Press, 1977; vol. V. ed. K. Crowley, Cambridge, Mass. and London: MIT Press, 1986.

Bibliography

Allais, M. 1947. *Economie et intérêt*. Paris: Imprimerie Nationale.

Arrow, K.J. 1967. Samuelson collected. *Journal of Political Economy* 75(5), October, 730–37.

Arrow, K.J. 1983. Contributions to welfare economics. In Brown and Solow (1983).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bachelier, L. 1900. Théorie de la speculation, trans. A.J. Boness and ed. P. Cootner, as *The Random Character of Stock Market-Prices*, Cambridge, Mass.: MIT Press, 1967.

Bliss, C.J. 1967. Review of *Collected Scientific Papers of Paul A. Samuelson*, 2 vols. *Economic Journal* 77(306), June, 338–45.

Bronfenbrenner, M. 1982. On the superlative in Samuelson. In Feiwel (1982).

Brown, E.C. and Solow, R.M., eds. 1983. *Paul Samuelson and Modern Economic Theory*. New York: McGraw-Hill.

Feiwel, G., ed. 1982. *Samuelson and Neoclassical Economics*. Boston: Kluwer-Nijhoff.

Gale, D. and Nikaido, H. 1965. The Jacobian matrix and the global univalence of mappings. *Mathematische Annalen* 159(2), 81–93.

Hahn, F.H. 1983. On general equilibrium and stability. In Brown and Solow (1983).

Hicks, J.R. 1939. *Value and Capital*. Oxford: Oxford University Press.

Houthakker, H. 1950. Revealed preference and the utility function. *Economica* 17, May, 159–74.

Houthakker, H. 1983. On consumption theory. In Brown and Solow (1983).

Jones, R.W. 1983. International trade theory. In Brown and Solow (1983).

Lindbeck, A. 1970. Paul Samuelson's contribution to economics. *Swedish Journal of Economics* 72, 341–54.

Machina, M.J. 1982. ‘Expected utility’ analysis without the independence axiom. *Econometrica* 50(2 March), 277–324.

Malinvaud, E. 1987. The overlapping generations model in 1947. *Journal of Economic Literature* 25, 103–5.

Mas-Colell, A. 1982. Revealed preference after Samuelson. In Feiwel (1982).

Merton, R.C. 1983. Financial economics. In Brown and Solow (1983).

Patinkin, D. 1983. Monetary economics. In Brown and Solow (1983).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Robinson, J. 1975. The unimportance of reswitching. *Quarterly Journal of Economics* 89(1), February, 32–9.

Samuelson, M.C. 1939. The Australian case for protection re-examined. *Quarterly Journal of Economics* 54, November, 143–9.

Seligman, B.B. 1967. On the question of operationalism. *American Economic Review* 57(1), March, 146–61.

Solow, R.M. 1983. Modern capital theory. In Brown and Solow (1983).

Tobin, J. 1983. Macroeconomics and fiscal policy. In Brown and Solow (1983).

How to cite this article

Fischer, Stanley. "Samuelson, Paul Anthony (born 1915)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000006> doi:10.1057/9780230226203.1469

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sargan, John Denis 1924 – 1996

David F. Hendry and Peter C.B. Phillips
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Denis Sargan was the leading British econometrician of his generation, playing a central role in establishing the technical basis for modern time-series econometric analysis. In a distinguished 40-year career as teacher, researcher, and practitioner Sargan transformed the role of econometrics in the analysis of macroeconomic time series and the teaching of econometrics. His research spanned exact finite-sample, approximate and asymptotic distribution theory for econometric estimators and tests, systems and single equation methods of modelling time-series and panel data, and problems of identification, together with applications, numerical analysis and simulation methods. Much of this foundational research remains relevant to current work.

Keywords

asymptotic theory; bootstrap; cointegration; computation; data mining; distribution of wealth; Durbin–Watson statistic; dynamic simplification; dynamic specification; Econometric Society; econometrics; Edgeworth expansions; equilibrium-correction mechanisms; estimation; finite-sample theory; full information maximum likelihood; Gaussian random walks; inference; inflationary expectations; instrumental variables; Leontieff, W.; limited information maximum likelihood; maximum likelihood; measurement error; moment approximations; Monte Carlo methods; Phillips curve; Phillips, A. W. H.; Sargan, J. D.; simplification; simulation-based methods; simultaneous equations; statistical methods; structural estimation; subjective probability; testing; three-stage least squares; unit root theory; weak instrumentation

Article

J.D. Sargan was born on 23 August 1924, in Doncaster, Yorkshire, where he spent his childhood. He was Emeritus Professor of Econometrics at the London School of Economics when he died at his home in Theydon Bois, Essex, on Saturday 13 April 1996. He received his secondary education at Doncaster

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Grammar School and, at the age of 17, gained a State Scholarship for entrance to St Johns College, Cambridge, where he took a first in mathematics and was Senior Wrangler.

Immediately after obtaining his degree, he was drafted into war work as a junior scientific officer attached to the RAF in Haverfordwest, where he provided basic statistical advice on the testing of new weapons systems. Like many of his generation Sargan's enthusiasm for economics was aroused by Keynes's *General Theory of Employment, Interest and Money*, and he decided to use his knowledge of mathematics and statistics to help tackle some of the pressing economic problems that faced society in the post-war years. Accordingly, in 1946 he returned to Cambridge to do more statistics, particularly time series, and to read economics, taking advantage of regulations that enabled him to complete a BA degree in economics in a year. More detailed biographical information is given in Hendry and Phillips (2003), on which much of the following discussion draws.

Starting his professional career as a lecturer in economics at Leeds University in 1948, Sargan went on to become the leading British econometrician of his generation, playing a central role in establishing the technical basis for modern time-series econometric analysis. In a distinguished career spanning more than 40 years as a teacher, researcher and practitioner, particularly during the period that he was Professor of Econometrics at the LSE, Sargan transformed both the role of econometrics in the analysis of macroeconomic time series and the teaching of econometrics. His influence on British econometrics was profound and continues today in the traditions he established.

Early research

Much of Sargan's research in the first decade of his career at Leeds University over 1948–58 was devoted to economic issues associated with the distribution of wealth, duopoly, production and growth. His paper (1957a) on the distribution of wealth is recognized to this day as the most general analytic treatment of the determination of the wealth distribution. His work (1958a; 1961a) on the instability of Leontieff's dynamic input–output model also attracted attention, showing that the Leontief model is not well adapted to explaining the behaviour of a decentralized economy. In addition to these researches on economic issues, he also published an early paper (1953a) on subjective probability and Bayesian thinking in economics, and another paper (1953b) on some of the statistical properties of the covariogram and periodogram.

Sargan's first foray into econometric methodology began with his paper (1957b) on ‘The Dangers of Oversimplification’, a discussion of the path-breaking analysis of the Oxford Savings Survey by Malcolm Fisher. Sargan's comments revealed a concern with three issues that recurred in his later research on econometric methodology: the abstract and constrained form of economic-theory models relative to the complexities of the data under analysis; the oversimplified nature of many estimated regression equations, excluding effects that were likely to be important in practice; and the problems of interpreting tests of large numbers of hypotheses. The first two concerns may have led to his subsequent interest in estimating relatively general and unrestricted models, and the third to his ideas about ‘data mining’ and model selection which became manifest in later research that was published posthumously in Sargan (2001a; 2001b). This discourse on oversimplification was closely followed by two major papers in econometrics that developed a theory of instrumental variable (IV) estimation, published in *Econometrica* in 1958 and the *Journal of the Royal Statistical Society* (JRSS) in 1959.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Instrumental variables

The two IV papers broke new ground, taking several large steps forward in the analytical treatment of simultaneous equations and the econometric methodology of estimation and inference. They quickly established Sargan as a technically accomplished new thinker in the econometrics arena, and they remain of lasting significance. The *Econometrica* paper laid out the methodology of IV estimation as we presently know it, provided asymptotics, related the approach to canonical correlation analysis and limited information maximum likelihood, gave tests for overidentification and under-identification, developed significance tests and confidence intervals, suggested instruments for use in practical work, and discussed the accuracy of the asymptotic theory. In the course of the latter discussion, Sargan pointed out that biases in estimation are likely to be large when the structural equation is almost unidentified, thereby foreshadowing some concerns over the effects of weak instrumentation that have occupied professional interest in recent years, following their explicit treatment in P. Phillips (1989), Nelson and Startz (1990) and Staiger and Stock (1997).

The *JRSS* paper advanced the analysis of IV estimation by considering the more general case where the structural coefficients $a = a(\theta)$ satisfied some analytic constraints and could be functionalized on a vector of fundamental parameters θ, applying the theory to the case of structural models with autoregressive errors, constructing statistical tests and confidence intervals, and again looking at overidentified and unidentified cases. The framework of this paper made it possible to consider problems of dynamic specification in a rigorous manner by means of statistical testing in a nonlinear in parameters context, thereby laying a foundation for much subsequent research in econometric methodology including Sargan's own later work (1980) on dynamic simplification.

Focus on econometric theory

Sargan took up a Fulbright scholarship in the United States for two years from 1958, spending the first academic year at the University of Minnesota, teaching summer school at Stanford University, moving to the University of Chicago for 1959–60 and visiting the Cowles Foundation at Yale in 1960. These visits firmly focused his growing interest on the econometric theory of estimating structural economic models from time-series data and, together with the publication of his two papers on IV estimation, brought his work in econometrics to the attention of the North American academic community. From this point forward, Sargan's career fell under the spell of a deep fascination with the design of statistical methods suitable for studying empirical economic problems and the intellectual problems involved in working out their finite sample and asymptotic properties.

In July 1960, Sargan returned to Leeds University to take up a readership, and his growing reputation for insightful, rigorous and powerful analyses led to his election to a fellowship of the Econometric Society in 1963. In the same year, he was recruited by the London School of Economics as a Reader in Statistics, in the same department as Jim Durbin, before joining A.W.H. (Bill) Phillips (already famous for the Phillips machine and the Phillips curve) in the economics department as Professor of Econometrics in 1965.

The period of the early 1960s saw the publication of some of Sargan's most influential papers and the formation of fundamental ideas that would play a major role in his later research. Two articles in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Econometrica (1961b; 1964b) studied maximum likelihood estimation of structural systems. The first set up a framework that enabled structural estimation in the presence of autoregressive errors, thereby accomplishing a marriage of two earlier theories. The second elegantly established the asymptotic equivalence of full information maximum likelihood (FIML) and three-stage least squares (3SLS), thereby confirming the asymptotic efficiency of the latter. A third paper, presented at the Copenhagen meetings of the Econometric Society in 1963 and later abstracted in *Econometrica* (1964b), conceived the notion of approximating the distributions of econometric estimators by means of Edgeworth expansions. This paper was never published, but it gradually evolved into a major research programme concerned with the theory and application of Edgeworth expansions, formally beginning nearly a decade later with the publication of Sargan and Mikhail (1971).

The ‘Colston paper’

A fourth paper was prepared for the Colston Society conference on National Economic Planning held at Bristol University in 1963 and was published in 1964. This ‘Colston paper’ (1964a), as it has become known, is possibly Sargan's most famous paper and is certainly his most important contribution to empirical econometric methodology. The paper laid out the conceptual basis of the so-called ‘LSE approach’ to econometric modelling, so Sargan is justly credited with the foundation of that approach. The main characteristics of the ‘LSE approach to econometric modelling’ (which in fact draws on work from many other institutions) are blending prior economic theory ideas with thorough data analysis to develop empirical models consistent with both sources of information, but with neither having precedence. In the context of time series, this led to an emphasis on commencing empirical modelling from relatively general dynamic equations capable of capturing the properties of the data while representing the relevant economic theories, rather than estimating stochastic implementations of theory models. Few papers can have contained so many novel ideas, each of which really deserved a separate article.

The paper is characteristically self-effacing and modest about its many practical contributions, though technically brilliant and economical in its execution. First, as a framework for constructing models, Sargan considered the use of ‘long-run’ economic analysis to specify the equilibrium of a model and introduced ‘equilibrium-correction’ mechanisms as a behavioural dynamic, following some earlier work (particularly A. Phillips, 1954) on trade cycle adjustment mechanisms. In doing so, Sargan established what is now perhaps the most widely used form of time-series econometric equation in empirical work. Second, Sargan viewed the presence of autoregressive errors in time-series models as a simplification (by virtue of the implicit factorization) of more general system dynamic reactions, and he constructed mis-specification tests that were valid even after estimating dynamic equations. This work translated into Sargan's later concern (1980) for direct tests of dynamic specification and simplification strategies in inference. To address another practical problem of empirical research, the Colston paper formulated a procedure for comparing linear against logarithmic specifications, and investigated the impact of data transformations on the selection of models. The paper further proposed a nonlinear in parameters IV estimator for models where the data were subject to measurement errors, devised and implemented operational computer programs for the new econometric methods, and included a proof that the required iterative computations would converge. Finally, the paper illustrated the methodology by matching the econometric theory to the specific, topical and difficult empirical problem of wage and price

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
determination in the United Kingdom. Previous models had related the changes in the variables, namely, wage inflation and price inflation. Such formulations precluded any relationship between the levels of wages and prices, which could therefore drift apart over time. Sargan argued that economic agents are concerned about the level of real wages, not just price inflation, so he formulated a model with a long-run equilibrium and incorporated real wages in the wage equation, thereby distinguishing the equation from many other models, including the Phillips curve. The paper also included a data-based proxy for ‘inflation expectations’, which was called ‘an extrapolation of past price movements into the future’ and the disequilibrium of real wages from its target depended on unemployment, productivity and political factors. In modern parlance, the levels variables were integrated whereas the differences and the equilibrium errors were not, so the equation implicitly required cointegration between the levels.
Sargan's analysis highlighted the role of real-wage resistance in wage bargains, interpreting the equilibrium correction – the deviation of real wages from a productivity trend – as a ‘catch-up’ mechanism for recouping losses incurred from unanticipated inflation. As the 1960s proceeded, this real-wage resistance proved to be an insuperable barrier to the successful implementation of incomes policy in the UK. The Colston paper also included a policy discussion in which permanent and transitory effects were distinguished to ascertain which changes would persist and which fade out (such as devaluations).
Prior to Sargan's Colston paper, it was common in empirical econometric practice to test for residual autocorrelation (for example, by the Durbin–Watson statistic), and if it were shown to be present, estimate a ‘generalized’ model that allowed for an autoregressive-error process. Sargan reversed this convention, interpreting autoregressive errors as a restriction on the dynamic specification of a model that, when valid, permitted the adoption of a more parsimonious representation. He also stressed that empirical specifications should be stringently evaluated, and formulated tests for the validity of the instrumental variables used in estimation and for higher-order autoregressive errors based on the residuals from the estimated equations. Despite the existence of this test, which was valid in dynamic equations, the Durbin–Watson test continued to be widely and invalidly used for many years in dynamic systems. Regarding computation, the paper carefully addressed the logic of the calculations both to embed all of the estimators in a common framework and to ensure as efficient an iterative procedure as possible, including good selections of the initial values and step lengths, and checks for multiple optima. Sargan's demonstration that the step-wise iterative computations converged to a local optimum was the first of its kind in econometrics and reflected his keen interest in numerical analysis.
Thus, in matters of econometric theory, empirical methodology, numerical computation, empirical application and the integration of economic ideas and econometric technique, the Colston paper was a watershed of new ideas and stands as one of the classic works of econometrics. Many new avenues of research were opened up, leading through equilibrium correction to cointegration analysis, encompassing, general-to-specific modelling, and a greater emphasis on model evaluation (Hendry, 1995, provides an overview).

Advanced theory

While the Colston paper constituted Sargan's most influential work from the perspective of empirical practice, the challenges that fired his intellectual passion principally lay elsewhere – in advanced theory. His greatest theoretical interest was in developing a finite sample distribution theory of estimation and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
inference, perhaps the most technically demanding field of econometric analysis. His main contributions in this area began with the publication of Sargan and Mikhail (1971), continued throughout the 1970s and 1980s, and covered all approaches – exact analytical derivation, asymptotic series approximations of both distributions and moments, and simulation-based methods. Despite the near-intractable nature of the manifold problems in this field, Sargan devoted a huge effort to produce solutions, pushing the frontiers of knowledge forward in remarkable ways in each of the main approaches.

Since economic systems are typically dynamic and/or simultaneous, the finite sample distributions of most econometric estimators and tests are extremely complicated, and the exact derivation of these distributions is a technically daunting task in all but the most trivial cases. Even when an exact theory is developed, the final results are often of limited applicability, rely on strong distributional assumptions and do not extend to dynamic settings because of formidable mathematical complexity. Sargan (1976a, Appendix A) provided the first general exact results for the distribution of the IV estimator in a structural equation, but he was able to resolve the distribution in closed form (as distinct from integral form) only in the just identified case. The general overidentified case was resolved subsequently in P. Phillips (1980).

Even in cases where the exact distribution itself is unattainable, certain interesting features of the distribution may be established, such as the existence or non-existence of moments. In this regard, Sargan (1970) gave an elegant demonstration of the fact that structural form FIML estimators, for instance, have no finite integral-order moments (mean, variance, and so on), thereby establishing that that these distributions typically have thick tails. By contrast, IV estimators generally have finite moments up to an order that is determined by the degree of overidentification of the structural equation and, on this topic, Sargan (1978) provided a definitive analysis of moment existence for the 3SLS estimator. In related work that was eventually published in his collected papers, Sargan (1975b) examined the tail behaviour of reduced form estimators and here showed that FIML estimators have finite moments to a certain order (determined by the sample size) whereas IV estimators like 3SLS typically have no finite reduced form moments in overidentified cases. These exact results reveal that FIML procedures can offer some advantages, in terms of reduced outlier activity, when the fitted reduced form is used, for example in prediction.

More general results can be obtained using series expansion and other approximations. Indeed, Sargan hoped that general approximation formulae using Edgeworth asymptotic series could be developed and incorporated into regression software, possibly with the use of computerized algebra, and then used to adjust critical values and improve inference. That goal has not yet been realized and, even with the advent of more recent bootstrap technology, continues to be elusive, partly because available approximations are rarely accurate enough and partly because major difficulties are encountered with all approaches in time-series models as the zone of non-stationarity is approached.

In terms of computer-intensive methods, Sargan helped at an early stage in the development and implementation of ideas (such as the use of antithetic and control variates) that made simulation methods viable and computationally efficient (Sargan, 1976a, Appendix D). He also made important headway in validating approaches based on moment approximations (Sargan, 1974a), and even considered the complex case where Monte Carlo estimates and moment approximations are developed in cases where the actual moments fail to exist (Sargan, 1982), so that the approximations characterize pseudo-moments (or moments of suitable approximating distributions). Pseudo-moment expansions of this type provide an intriguing way of interpreting the descriptive moment statistics conventionally reported in Monte

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Carlo experiments. When the moments of the underlying distribution are infinite, Sargan's results reveal that such simulation-based moment statistics can be validly interpreted as estimates of the actual moments of the Edgeworth approximating distributions up to a certain order, depending on the sample size and the number of replications. This work resolved a major potential concern in the reporting of simulation-based research, since many simulation experiments are conducted in settings where the existence of moments has not been established.

Sargan's work on asymptotic expansions of the finite sample distributions of econometric estimators and test statistics was extraordinary in its coverage and its generality, dealing with IV estimators (Sargan and Mikhail, 1971), full information maximum likelihood (Sargan, 1970), k-class estimators (Sargan, 1975a), asymptotic chi-squared criteria (Sargan, 1980), and the theory of validity of the expansions in econometric contexts (Sargan and Satchell, 1986) together with formulae and algorithms for implementation of the approach (Sargan, 1976a). The final reference in this list is Sargan's famous Walras–Bowley lecture, which was presented at the 1974 San Francisco meetings of the Econometric Society and contained a multifaceted analytical development of the subject complete with four long technical appendices dealing with different approaches and detailing formulae that must have been obtained over many years of research. In a lucid discussion of density expansions in a general setting, this paper gave explicit formulae for the components of the Edgeworth expansion for a general form of econometric statistic and revealed the dependence of the correction terms on the form of the statistic and the cumulants of the sample moments on which the statistic depended. Importantly, given subsequent research, the paper also supplemented the idea of analytic expansions with a simulation-based approach (originally due to George Barnard) that is now recognized as a version of the modern parametric bootstrap.

Sargan's Walras–Bowley lecture and several of his other papers in this demanding field are filled with technical innovations and show little sign of aging even after decades of subsequent research. Although asymptotic expansions have been found an unreliable means of improving inferential accuracy, Sargan's theoretical contributions helped blaze the trail of finite-sample theory in the 1970s and early 1980s, and they furnish a substantial body of results that have improved our understanding of the properties of econometric estimators and tests. Edgeworth expansions of the sort Sargan sought to validate and implement are now routinely used (for example, Hall, 1992; Horowitz, 2001) to validate the improvements delivered by bootstrap methods in practical econometric applications.

Other contributions

In addition to the main themes of his research outlined above, Sargan made several other intriguing contributions to econometric theory. His work (1975a) on ‘large models’, for instance, still stands as a lone pioneering piece of technical analysis of the consequences of having a system whose size is large relative to the available database, and was strangely unlike any of the other papers published in the symposium on large macroeconometric models in which it appeared. Instead, as Robinson (2003) has argued, Sargan's ideas on large simultaneous systems are more relevant to the semiparametric methods that are now commonplace in econometrics.

Sargan's (1974a) work on continuous time stochastic models represents another major contribution, again in a very different field. This paper provided the first formal asymptotic study of the effects of approximating open-loop linear differential equation systems with discrete time simultaneous equations.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Such discrete approximations were in use in practical work, and later have become more popular through the use of Euler approximations of differential equations in financial econometrics. Sargan's work built on an earlier study by Bergstrom (1966) and analysed the order of magnitude, in terms of the sampling interval, of the inconsistency in various IV and FIML econometric estimates of the coefficients in the continuous system. The paper also examined more applied issues such as the impact of this bias on forecasting.

Research on identification

In other important research, Sargan (1980) addressed the thorny issue of the effects of near non-identification on modelling and inference. Early researchers on simultaneous equations methodology had recognized the importance of, but practical difficulties in assessing, identification. Tests for under-identification (such as those in Sargan, 1958b) were a manifestation of this concern. In practical work, however, these tests are seldom used, and most empirical research proceeds by assuming an equation is identified by order conditions. Sargan recognized that, in the event of near lack of identification, the asymptotic properties of econometric estimators and tests would be affected – in fact, in an early discussion, Sargan (1959, section 6) hints at some of the possibilities, including slower rates of convergence than the usual $\sqrt{n}$ rate for a sample of size *n*. Subsequently, Sargan (1975b) explored the relationship between identification and consistent estimability in systems of simultaneous stochastic equations. Then, in his presidential address to the Econometric Society in 1980, he considered nonlinear-in-parameter models that were ‘nearly unidentified’, in the sense that the first-order rank condition for local identification failed, but higher-order defining shape conditions held so that there was still identification. In singular cases like these, which followed the earlier discussion in the 1959 paper, Sargan found that the conventional asymptotic theory for IV estimation broke down, with slower rates of convergence and a non-normal limit theory applying. Sargan (1983b) later showed that similar problems of singularity occur in dynamic models with autoregressive errors. This work on near lack of identification anticipated future research, and its arena of application has proved to be far wider than may originally have been envisaged. It is especially relevant, for instance, in micro-econometric applications where the relevance condition is weak and the IVs are sometimes barely correlated with the regressors. A prominent example in this field has been the study of the impact of schooling on earnings, where intrinsic ability affects both, is unmeasured and therefore contaminates the equation error. In such cases, the search for an instrumental variable that satisfies orthogonality with the error can lead to some arcane choices that end up being only weakly correlated with the regressors they service (Angrist and Krueger, 1991). The impact of such weak (or nearly irrelevant) instruments in applied econometric work is now an intensive area of research – see Andrews and Stock (2005) for an overview.

More time-series analysis

While Sargan retired before unit root and cointegration theory revolutionized time-series econometrics in the 1980s, he had studied Gaussian random walks, presenting an early paper on the subject at the UK econometric study group held at LSE in 1973, some results of which later appeared in joint work with Bhargava (1983c) in *Econometrica*. In further work, Sargan and Bhargava (1983d) showed that in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
regression models with moving average errors where there is a root on or near the unity circle, the likelihood function can have a local maximum at unity with reasonably high probability and that the limit theory is nonnormal in the unit root case, invalidating conventional tests. Accordingly, their paper argued against the empirical practice of checking for overdifferencing and in support of a most powerful invariant test of independence based on the Berenblut–Webb (1973) statistic. This approach has subsequently received much attention in the unit root testing context.

This brief summary of Sargan's theoretical contributions to econometrics shows the enormous range of his research interests. While almost all of the econometric theory he developed related to time series models fitted by time-domain methods, he also worked on adapting frequency-domain methods to simultaneous equations models in econometrics (Espasa and Sargan, 1977), missing data (Sargan and Drettakis, 1974b) and took some interest in panel data problems (Sargan and Bhargava, 1983e). By the time he retired in 1984, he had worked on most of the important problems and research areas in econometrics of his generation.

Overview

Sargan's appointment at the LSE in 1965 took its econometrics group to the technical forefront in research. He can be credited with the creation of a generation of econometricians in the UK who were trained to high technical levels in all aspects of quantitative economics but who were especially strong in econometric theory and methodology. His devotion to teaching and research training was exemplary. In total he supervised 36 successful doctorates in a host of fields covering much of the discipline of econometric method and many of its applications. Sargan held a ‘modern’ view of dissertation research as a process by which students learnt the practice of research by means of intimate involvement on the part of a supervisor. In this regard, his generosity to his students and colleagues was famous at the LSE and beyond, and undoubtedly played a major role in attracting doctoral students in econometrics. A full listing of his graduate students and their dissertation titles is contained in Sargan (1988a).

Sargan's contributions earned him international distinction and honours. In 1980 he served as President of the Econometric Society, presiding over the World Congress of the Society at Aix-en-Provence. He was made a Fellow of the British Academy in 1981 and assumed the Tooke Professorship of Economic Science and Statistics at the LSE in 1982. On retirement in 1984, he became Emeritus Professor of Economic Science and Statistics at the University of London, and an international conference was held in his honour at Oxford University. He became an honorary foreign member of the American Academy of Arts and Sciences in 1987, was awarded a fellowship of the LSE in 1990, and received an honorary doctorate from the University of Carlos III, Madrid in 1993, where a further celebratory conference was held for him.

The wide range of Sargan's research is celebrated by the topics addressed in the volume edited by Hendry and Wallis (1984), which commemorated his 60th birthday. He was interviewed for the journal *Econometric Theory* by P. Phillips (1985). Maasoumi edited his collected works, published as Sargan (1988a), which, together with his advanced econometrics lecture notes edited by Desai (Sargan, 1988b) well illustrate his analytic rigour and intellect. Three issues of econometrics journals have appeared in his memory: one in *Journal of Applied Econometrics*, 2001, which was on empirical macro-econometrics; another in *Econometric Reviews*, 2001, gave a biographical history of Sargan's career and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
printed several of his still unpublished papers; a third, in *Econometric Theory*, 2003, brought together two of Denis Sargan's essays on econometrics published for the first time, a laudation by Antoni Espasa, and three memorial essays offering an intellectual overview of his work.

Sargan had an enormous intellectual influence within the UK, both on the training of econometric theorists and on econometric practice. Outside the UK, his influence has not been as strong and, particularly in North America, different traditions and interests have prevailed. The Colston volume was an obscure source for economists and this undoubtedly limited the impact of his work on econometric methodology; and his choice of problems in econometric theory also did not always relate well to the immediate concerns of empirical researchers or other econometricians – he had his own vision of what the subject needed, and he pursued that vision with determination. Yet, when the history of econometrics in the second half of the 20th century is written, Sargan's place among the leaders of the econometric profession in that era is assured. The research agenda that he initiated has proved to be of tremendous scope, affecting almost every major area of the discipline. His scientific works show a remarkable durability, some of them (like the Colston paper and Walras–Bowley lecture) having the status of enduring classics. The world of econometric theory and its applications has moved on, but the themes of Sargan's research program persist in ongoing work, and his technical accomplishments remain part of the edifice of theory, technique and methodology that we collectively call econometrics.

See Also

- cointegration
- computational methods in econometrics
- data mining
- distributed lags
- econometrics
- Edgeworth, Francis Ysidro
- finite sample econometrics
- full and limited information methods
- inflation dynamics
- instrumental variables
- maximum likelihood
- measurement error models
- Phillips, Alban William Housego
- serial correlation and serial dependence
- simultaneous equations models
- specification problems in econometrics
- two-stage least squares and the *k*-class estimator
- vector autoregressions

We thank many individuals for their information and help in writing this biography. In particular, Mary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

*Sargan kindly provided details of Sargan's early life, and we have drawn on reviews, obituaries and memoirs written with, or by, Lord Meghnad Desai, Neil Ericsson, Toni Espasa, Esfandiar Maasoumi, Grayham Mizon, Hashem Pesaran, Peter Robinson and Kenneth Wallis, as well as our own memoir (*Hendry and Phillips, 2003*).*

Selected works

1953a. Subjective probability and the economist. *Yorkshire Bulletin of Economic and Social Research* 5, 53–64.

1953b. An approximate treatment of the properties of correlogram and periodogram. *Journal of the Royal Statistical Society, Series B* 15, 140–52.

1957a. The distribution of wealth. *Econometrica* 25, 568–90.

1957b. The dangers of oversimplification. *Bulletin of the Oxford Institute of Statistics* 19, 171–8.

1958a. The instability of the Leontief dynamic model. *Econometrica* 26, 381–92.

1958b. The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.

1959. The estimation of relationships with autocorrelated residuals by the use of the instrumental variables. *Journal of the Royal Statistical Society, Series B* 21, 91–105.

1961a. Lags and the stability of dynamic systems: a reply. *Econometrica* 29, 670–3.

1961b. The maximum likelihood estimation of economic relationships with autoregressive residuals. *Econometrica* 29, 414–26.

1964a. Wages and prices in the United Kingdom: a study in econometric methodology. In *Econometric Analysis for National Economic Planning*, vol. 16 of Colston Papers, ed. P. Hart, G. Mills and J. Whitaker. London: Butterworth Co.

1964b. 3SLS and FIML estimates. *Econometrica* 32, 77–81.

1970. The finite sample distribution of FIML estimators. In J. Sargan, *Contributions to Econometrics*, vol. 1, ed. E. Maasoumi. Cambridge: Cambridge University Press, 1988.

1971. (With W. Mikhail.) A general approximation to the distribution of instrumental variables estimates. *Econometrica* 39, 131–69.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1974a. Some discrete approximations to continuous time stochastic models. *Journal of the Royal Statistical Society, Series B* 36, 74–90.

1974b. (With E. Drettakis.) Missing data in an autoregressive model. *International Economic Review* 15, 39–58.

1975a. Asymptotic theory and large models. *International Economic Review* 16, 75–91.

1975b. The identification and estimation of sets of simultaneous stochastic equations. In J. Sargan, *Contributions to Econometrics*, vol. 1, ed. E. Maasoumi. Cambridge: Cambridge University Press, 1988.

1976a. Econometric estimators and the Edgeworth approximation. *Econometrica* 44, 421–48.

1976b. The existence of moments of estimated reduced form coefficients. In J. Sargan, *Contributions to Econometrics*, vol. 2, ed. E. Maasoumi. Cambridge: Cambridge University Press, 1988.

1977. (With A. Espasa.) The spectral estimation of simultaneous equation systems with lagged endogeneous variables. *International Economic Review* 18, 583–605.

1978. The existence of moments of 3SLS estimators. *Econometrica* 46, 1329–50.

1980. Some tests for dynamic specification for a single equation. *Econometrica* 48, 879–97.

1982. On Monte Carlo estimates of moments that are infinite. In *Advances in Econometrics*, ed. R. Basmann and G. Rhodes, Jr. Greenwich, CT: JAI Press.

1983a. Identification and lack of identification. *Econometrica* 51, 1605–33.

1983b. Identification of models with autoregressive errors. In *Studies in Econometrics, Time Series and Multivariate Statistics*, ed. S. Karlin, T. Amemiya and L. Goodman. New York: Academic Press.

1983c. (With A. Bhargava.) Testing residuals from least squares regression for being generated by the Gaussian random walk. *Econometrica* 51, 153–74.

1983d. (With A. Bhargava.) Maximum likelihood estimation of regression models with first order moving average errors when the root lies on the unit circle. *Econometrica* 51, 799–820.

1983e. (With A. Bhargava.) Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51, 1635–59.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1986. (With S. Satchell.) A theory of validity for Edgeworth expansions. *Econometrica*, 54, 189–213.

1988a. *Contributions to Econometrics*, 2 vols., ed. E. Maasoumi. Cambridge: Cambridge University Press.

1988b. *Lectures on Advanced Econometrics*. New York: Basil Blackwell.

2001a. Model building and data mining. *Econometric Reviews* 20, 159–70.

2001b. The choice between sets of regressors. *Econometric Reviews* 20, 171–86.

Bibliography

Andrews, D. and Stock, J. 2005. Inference with weak instruments. Discussion Paper No. 1530. Cowles Foundation, Yale University.

Angrist, J. and Krueger, A. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979–1014.

Berenblutt, I. and Webb, G. 1973. A new test for autocorrelated errors in the linear regression model. *Journal of the Royal Statistical Society, Series B* 35, 33–50.

Bergstrom, A. 1966. Nonrecursive models as discrete approximations to systems of stochastic differential equations. *Econometrica* 34, 173–82.

Hall, P. 1992. *The Bootstrap and Edgeworth Approximation*. New York: Springer Verlag.

Hendry, D. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D. and Phillips, P. 2003. John Denis Sargan 1924–1996. *Proceedings of the British Academy* 120, 385–409.

Hendry, D. and Wallis, K. 1984. *Econometrics and Quantitative Economics*. New York: Basil Blackwell.

Horowitz, J. 2001. The bootstrap. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Nelson, C. and Startz, R. 1990. Some further results on the small sample properties of the instrumental variable estimator. *Econometrica* 58, 967–76.

Phillips, A. 1954. Stabilization policy in a closed economy. *Economic Journal* 64, 290–313.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Phillips, A. 1959. The estimation of parameters in systems of stochastic differential equations. *Biometrika* 46, 67–76.

Phillips, P. 1980. The exact finite sample density of instrumental variable estimators in an equation with *n*+1 endogenous variables. *Econometrica* 48, 861–78.

Phillips, P. 1985. The ET interview: John Denis Sargan. *Econometric Theory* 1, 119–39.

Phillips, P. 1989. Partially identified econometric models. *Econometric Theory* 5, 181–240.

Robinson, P. 2003. Denis Sargan: some perspectives. *Econometric Theory* 19, 481–94.

Staiger, D. and Stock, J. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.

How to cite this article

Hendry, David F. and Peter C.B. Phillips. "Sargan, John Denis 1924–1996." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000010> doi:10.1057/9780230226203.1470

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

satisficing

Herbert A. Simon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Satisficing’ (choosing an option that meets or exceeds specified criteria but is not necessarily either unique or the best) is an alternative conception of rational behaviour to optimizing. It is an attractive alternative when genuine optima could be computed only with infeasible levels of effort, or when goals are incommensurable. The standards that determine ‘satisfactory’ may be determined by the adjustment of aspiration levels in response to experience. Satisficing may provide an improved representation of actual choice behaviour, but at the cost of less predictive power than optimizing theory.

Keywords

computation; convexity; incommensurability; linear programming; operations research; optimization; satisficing; search costs; Stigler, G.

Article

A decision maker who chooses the best available alternative according to some criterion is said to optimize; one who chooses an alternative that meets or exceeds specified criteria, but that is not guaranteed to be either unique or in any sense the best, is said to satisfice. The term ‘satisfice’, which appears in the *Oxford English Dictionary* as a Northumbrian synonym for ‘satisfy’, was borrowed for this new use by H.A. Simon (1956), in ‘Rational Choice and the Structure of the Environment’.

Optimization and its problems

In the literature of economics and statistical decision theory, rationality has usually been defined in such a way as to imply some form of optimization, for example, maximization of utility subject to budget constraints. In simple situations, like the illustrative examples used in economics textbooks, computing a maximum may be a simple process, requiring, perhaps, nothing more onerous than taking a first

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
derivative and setting it equal to zero. Even in much more complex situations, involving thousands of linear equalities and inequalities but also a linear criterion function, the powerful methods of linear programming often permit optimal choices to be found with tolerable amounts of computing effort.
In many (most?) real-world situations, however, genuine optima (maxima or minima) are simply not computable within feasible limits of effort (*see* rationality, bounded). This is especially true when decisions must be made without benefit of computer, but it is frequently true even when powerful computing facilities are available. The complexity of the world is not limited to thousands or even tens of thousands of variables and constraints, nor does it always preserve the linearities and convexities that facilitate computation.

The satisficing alternative

Faced with a choice situation where it is impossible to optimize, or where the computational cost of doing so seems burdensome, the decision maker may look for a satisfactory, rather than an optimal alternative. Frequently, a course of action satisfying a number of constraints, even a sizeable number, is far easier to discover than a course of action maximizing some function.
The example has been given of searching for a needle in a haystack. Given a probability density distribution of needles of varying degrees of sharpness throughout the haystack, searching for the sharpest needle may require effort proportional to the size of the haystack. The task of searching for a needle sharp enough to sew with requires an effort that depends only on the density of needles of the requisite sharpness, and not at all on the size of the stack. The attractiveness of the satisficing criterion derives from this independence of search cost from the size and complexity of the choice situation.
In a formal sense, a process of satisficing could always be converted into a process of optimizing by taking into account the cost of search, and only searching up to the point where the expected gain derivable from another minute of search is just equal to the opportunity cost of that minute (Simon, 1955; Stigler, 1961). However, this conversion imposes a new, possibly heavy, informational and computational burden upon the chooser: the burden of estimating the expected marginal return of search and the opportunity cost. Solving these estimation problems may be as difficult as making the original choice, or even more difficult. An alternative is to search until a satisfactory alternative is found.
Conversely, most of the so-called optimization models of operations research and management science can more profitably be viewed as satisficing models. In the application of OR optimization techniques, some highly simplified approximation to a real-world situation is reduced to a formal model (for example, a linear programming or integer programming model), and an optimum is then calculated for this approximation with respect to a similarly approximate criterion function. The resulting ‘optimal’ decision will often provide a satisfactory decision for the real-world situation, but without guarantee that it will be better than a decision arrived at by some alternative satisficing technique.
How may the satisficer set the level of the criteria that define ‘satisfactory’? Psychology proposes the mechanism of aspiration levels: if it turns out to be very easy to find alternatives that meet the criteria, the standards are gradually raised; if search continues for a long while without finding satisfactory alternatives, the standards are gradually lowered. Thus, by a kind of feedback mechanism, or ‘tâtonnement’, the decision maker converges toward a set of criteria that are attainable, but not without effort. The difference between the aspiration level mechanism and the optimization procedure is that the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
former calls for much simpler computations than the latter. It is somewhat analogous to the difference between adaptive and rational expectations, respectively, in the theory of choice under uncertainty.

Incommensurability of goals and outcomes

Satisficing can also provide another kind of computational advantage over optimizing. Human decision makers often find it very difficult to make trade-offs among aspects or dimensions of value that seem to them incommensurable. Of course, it is the function of the utility function to insure commensurability in all cases, but we may not wish to postulate in advance that such a function exists, or may not know how to characterize it. Three classes of situations into which incommensurability is especially likely to intrude are: (1) cases of uncertainty, where, for each alternative, a bad outcome under one contingency must be balanced against a good outcome under another; (2) cases of multiperson choice, where one person's gain is another's loss; and (3) cases where each choice involves gain along one dimension of value and loss along another very different one.

It has been observed empirically that in circumstances of these kinds, and especially when each outcome entails unpleasant as well as pleasant consequences, decision makers do not proceed promptly to a choice, but instead seek to avoid the necessity for comparison. One common reaction, for example, is to refuse to choose among the given set of alternatives, and instead, to initiate a search for a new alternative that: in case (1), will ensure at least a minimally satisfactory outcome under all contingencies; in case (2), will ensure all participants a satisfactory outcome; and in case (3), will ensure an outcome that is at least minimally satisfactory along all dimensions. The acceptance of such alternatives comes within our definition of satisficing (Hogarth, 1980).

Consequences for economic theory

It is easier to reconcile a satisficing than an optimizing theory of economic decision making with what is known empirically of actual choice behaviour and of the computational limits of the human mind. On the other hand, a great deal is given up by the substitution of the former for the latter – given up in terms of the strength and variety of theorems that can be derived from the postulate of rationality in the two cases. To make predictions about behaviour on the basis of a satisficing theory requires much more empirical data about, for example, aspiration levels and their adaptivity, than does prediction on the basis of the optimizing theory. The magnitude of the difference becomes less, however, when we recognize that the optimizing theory says nothing about the shape or content of the utility function. It simply postulates a consistency of behaviour over time that may not be found if the decision maker is satisficing instead of optimizing.

In the last analysis, a choice between the two kinds of postulates will have to be made in terms of their relative effectiveness and accuracy in predicting and explaining economic behaviour, at both micro and macro levels. There is still little consensus in the economics profession as to the circumstances under which one postulate or the other will be the more advantageous.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- behavioural economics and game theory
- economic man
- rationality, bounded

Bibliography

Hogarth, R.M. 1980. Judgment and Choice: The Psychology of Decision. New York: Wiley.

Radner, R. 1975. Satisficing. *Journal of Mathematical Economics* 2, 253–62.

Simon, H.A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69, 99–118. Reprinted in Simon (1982), ch. 7.2.

Simon, H.A. 1956. Rational choice and the structure of the environment. *Psychological Review* 63, 129–38. Reprinted in Simon (1982), ch. 7.3.

Simon, H.A. 1982. *Models of Bounded Rationality*. Cambridge, Mass.: MIT Press.

Stigler, G.J. 1961. The economics of information. *Journal of Political Economy* 69, 213–25.

Winter, S.G. 1971. Satisficing, selection and the innovating remnant. *Quarterly Journal of Economics* 85, 237–61.

How to cite this article

Simon, Herbert A. "satisficing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000013> doi:10.1057/9780230226203.1471

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sauvy, Alfred (1898 – 1990)

Ronald D. Lee
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic–demographic equilibrium; marginal product of labour; maximum population; minimum population; optimum population; Sauvy, A.

Article

Sauvy was born 31 October 1898, in Villeneuve-de-la-Raho, France. He graduated from the Ecole Polytechnique, and is known as a demographer, economist, statistician and sociologist. He was an adviser to Jean Monnet and Paul Reynaud 1938–1940, a member of the Population Commission of the United Nations from 1947 to 1974, and he occupied a chair in social demography at the College of France from 1959 to 1969. He founded the Institut National d'Etudes Démographiques, one of the world's leading centres of demographic research, which he directed from 1945 to 1962. He was president of the International Union for the Scientific Study of Population from 1961 to 1963. A prolific author, his works include 45 books and many articles, on a broad range of topics from French economic history since the First World War to the effect of technological change on employment and the history of thought in demography. But he is best known among demographers and economists for his two-volume treatise *Théorie générale de la population* (1966), published in English in 1969 as *The General Theory of Population*.

The General Theory of Population attempts both theoretical and substantive generality and is largely independent of the English-language literature on the subject. Thus one finds no references to or integration of the English literature on the economics of fertility or the consequences of population change or optimum population theory. Rather, the treatise presents a highly individual view of the subject. The book is full of briefly presented but penetrating insights on a wide variety of topics, often illustrated with descriptive data. For an economist, however, the main interest of the book is in its more rigorous development and extension of the concepts of economic–demographic equilibrium, and of maximum, minimum and optimum population. The analysis is entirely comparative statics, based on the assumption of first increasing, then decreasing returns to labour and population. At the minimum population and again at the maximum, the average product of labour equals the subsistence level; in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

between these population sizes there remains an economic surplus after subsistence needs are met. Per capita product is of course maximized when the marginal product equals the average; more interestingly, total surplus is maximized when the marginal product of labour equals subsistence, at what is termed the ‘power optimum’ population – with the notion that a costly collective social goal can best be met at this size. The military optimum will be at a point between the power optimum and the maximum, since it requires both soldiers and surplus output. The implications of inequalities in income distribution are also discussed; the labour intensity of the consumption goods demanded by the rich will influence the size of the equilibrium population. Sauvy's theories, particularly of the determinants of the equilibrium population, have had an important influence on the thinking of economic demographers and social historians on the subject of homeostasis in human populations.

Selected works

1966. *Théorie générale de la population*. Paris: PUF. Trans. C. Campos as *The General Theory of Population*. New York: Basic Books, 1969.

How to cite this article

Lee, Ronald D. "Sauvy, Alfred (1898–1990)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000014> doi:10.1057/9780230226203.1472

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Savage, Leonard J. (Jimmie) (1917 – 1971)

I. Richard Savage
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bayesian statistics; De Finetti, B.; expected utility; likelihood; probability; Ramsey, F. P.; Savage, L. J.; stable estimation

Article

L.J. (Jimmie) Savage, né Leonard Ogashevitz, was born in Detroit on 20 November 1917 and died in New Haven on 1 November 1971. His interests were encyclopedic: as a youth he immersed himself in the *Book of Knowledge*, and at the time of his death he was preparing for the Peabody Museum a demonstration-exhibit on animal odorants. The dominant theme of Savage's professional work was the mathematical analysis of normative behaviour.

He received a BS (1938) and Ph.D. (1941) from the University of Michigan. In the early 1940s he obtained a broad postdoctoral exposure to pure and applied mathematics at the Institute for Advanced Study in Princeton, at Cornell, Brown, the Statistical Research Group of Columbia, the Courant Institute at New York University, and at Woods Hole Marine Biological Laboratory. From 1946 to 1960 he was at the University of Chicago, where he was central to the development of the statistics programme. Subsequently, he held professorships at Michigan and at Yale. Always, he was intellectually generous with students, colleagues, visitors and correspondents.

Savage's basic views and results on normative behaviour appear in his *Foundations* (1954). His essential theme, still being elaborated, is the relation between a person's probability for an event and his utility for the event. In particular, his probabilities and utilities must be consistent with the principle of maximizing his expected utility. These results flow from compelling axioms of coherent behaviour and they recommend specific strategies for applied statistics, such as the use of Bayesian statistics and the likelihood principle.

Savage's axioms imply that all probabilities reflect individual experience so that there is no reason for two people to have the same probability for a particular event. His theory conflicts with traditional views that hold probabilities to be basic constants of nature. At first Savage thought this conflict would not be significant in applying statistical theory but he remarked in the preface to the second edition of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Foundations (1972) that he was not successful in bringing the theories of statistics together at the applied level. He recognized the long process from elegant theory to serious applications. His paper on elicitation of probabilities (1971) develops methods to implement his theory of personal probability. And Savage (1977) warns against holding theoretical foundations as adequate to cover all aspects of applied statistics.

Savage's work on the foundations of statistics had major antecedents in Frank Ramsey and B. De Finetti, whose work was developed, polished and taught to a generation of scholars by Savage himself. Hewitt and Savage (1955) is both elegant mathematics and an extension of a basic result of de Finetti in the foundations of statistics. Dubins and Savage (1965) stems from a normative problem and bears mathematical fruit. Exposition of the basic ideas of applied Bayesian statistics combined with the new theory of stable estimation appears in Edwards, Lindman and Savage (1963). Additional biographical and critical analysis as well as most of Savage's published papers appear in the selection prepared by Ericson et al. (1981).

See Also

- Bayesian statistics
- expected utility hypothesis

Selected works

1948. (With M. Friedman.) The utility analysis of choices involving risk. *Journal of Political Economy* 56, 279–304.

1954. *The Foundations of Statistics*. New York: John Wiley & Sons. 2nd rev. edn, New York: Dover Publications, 1972.

1955. (With E. Hewitt.) Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society* 80, 470–501.

1963. (With W. Edwards and H. Lindman.) Bayesian statistical inference for psychological research. *Psychological Review* 70, 192–242.

1965. (With L.E. Dubins.) *How to Gamble If You Must: Inequalities for Stochastic Processes*. New York: McGraw-Hill. Reprinted with a new Bibliographic Supplement and Preface by L.E. Dubins as *Inequalities for Stochastic Processes* (*How to Gamble If You Must*), New York: Dover Publications, 1976.

1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66, 783–801.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1977. The shifting foundations of statistics. In *Logic, Laws and Life: Some Philosophical Complications*, ed. R.G. Colodny. Pittsburgh: University of Pittsburgh Press.

1981. *The Writings of Leonard Jimmie Savage – A Memorial Selection*. Prepared by a Committee (W.H. DuMouchel, W.A. Ericson (chair), B. Margolin, R.A. Olshen, H.V. Roberts, I.R. Savage and A. Zellner) for the American Statistical Association and the Institute of Mathematical Statistics, Washington, DC.

How to cite this article

Savage, I. Richard. "Savage, Leonard J. (Jimmie) (1917–1971)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000015> doi:10.1057/9780230226203.1473

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Savage's subjective expected utility model

Edi Karni
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews Savage's subjective expected utility theory and takes a critical view of the corresponding concept of subjective probabilities as a representation of decision makers' beliefs.

Keywords

Bayesian statistics; Bernoulli, D.; de Finetti, B.; decisions under uncertainty; expected utility hypothesis; monotonicity; objective probability; probability; Ramsey, F. P.; Savage, L. J.; St Petersburg paradox; state-dependent preferences; subjective probability; sure thing principle; utility; utility-maximizing behaviour

Article

In his seminal book *The Foundations of Statistics*, Savage (1954) advanced a theory of decision making under uncertainty and used that theory to define choice-based subjective probabilities. He intended these probabilities to express the decision maker's beliefs, thereby furnishing Bayesian statistics with a behavioural foundations.

The interpretation of probability as a numerical expression of beliefs is as old as the idea of probability itself. According to Hacking (1984), the notion of probability emerged in the 1650s with a dual meaning: (*a*) the relative frequency of a random outcome in repeated trials and (*b*) a measure of a decision maker's degree of belief in the truth of propositions or the likely realization of events. Both the 'objective' and the 'subjective' probabilities, as these inpts are now called, played important roles in the developments that lead to the formulation of Savage's subjective utility model.

In the early stages of their respective evolutions, the notion of utility was predicated on the existence of objective probabilities, and the notion of subjective probabilities presumed the existence of some form of utility. The ideas of utility and expected utility-maximizing behaviour were originally introduced by Bernoulli (1738). Bernoulli's preoccupation with resolving the famous St Petersburg paradox justifies

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
his taking for granted the existence of probabilities in the sense of relative frequencies. In the same vein, von Neumann and Morgenstern's (1944) axiomatic characterization of expected utility-maximizing players facing opponents who may employ a randomizing device to determine the choice of a pure strategy assumes that probabilities of these strategies are relative frequencies.

In the late 1920s and 1930s, Ramsey and de Finetti formalized the concept of choice-based subjective probability assuming that individuals seek to maximize expected utility when betting on the truth of propositions. In the behaviourist tradition, they explored the possibility of inferring the degree of confidence a decision maker has in the truth of a proposition from his betting behaviour and quantifying the degree of confidence, or belief, by probability. Invoking the axiomatic approach and taking the existence of utilities as given, Ramsey (1931) sketched a proof of the existence of subjective probabilities. De Finetti (1937) proposed a definition of subjective probabilities assuming linear utility and no arbitrage opportunities.

These developments culminated in the work of Savage. While synthesizing the ideas of de Finetti and von Neumann and Morgenstern, Savage introduced a new analytical framework and conditions that are necessary and sufficient for the existence and joint uniqueness of utility and probability, and the characterization of individual choice as expected utility-maximizing behaviour.

Savage's analytical framework

Decision making under uncertainty pertains to situations in which a choice of a course of action, by itself, does not determine a unique outcome. To formalize this notion Savage (1954) introduced an analytical framework consisting of a set *S*, whose elements are *states of the world* (or states, for brevity); an arbitrary set *C*, of *consequences*; and the set *F*, of *acts* (that is, functions from the set of states to the set of consequences). Acts correspond to courses of action, consequences describe anything that may happen to a person, and states are possible resolutions of uncertainty, that is, ‘a description of the world so complete that, if true and known, the consequences of every action would be known’ (Arrow, 1971, p. 45). Implicit in this definition is the notion that there is a unique true state. Events are subsets of the set of states. An event is said to obtain if it includes the true state.

A decision maker is characterized by a preference relation, $\succcurlyeq$, on *F*. The statement $f \succcurlyeq f'$ has the interpretation ‘the course of action *f* is at least as desirable as the course of action *f*• ′ ’. Given $\succcurlyeq$, the strict preference relation $\succ$ and the indifference relation ~ are defined as follows: $f \succ f'$ if $f \succcurlyeq f'$ and not $f' \succcurlyeq f$; $f \sim f'$ if $f \succcurlyeq f'$ and $f' \succcurlyeq f$.

The preference structure

The evaluation of a course of action in the face of uncertainty involves the decision maker's taste for the possible consequences and his beliefs regarding their likely realization. Savage's subjective expected utility theory postulates a preference structure, depicted axiomatically, permitting the numerical expression of the decision maker's valuation of the consequences by a utility function, that of his beliefs by a (subjective) probability measure on the set of all events, and the evaluation of acts by the mathematical expectations of the utility with respect to the subjective probability.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To state Savage's postulates, I employ the following notation and definitions. Given an event E and acts f and h, let $f_E h$ be the act such that $(f_E h)(s) = f(s)$ if $s \in E$, and $(f_E h)(s) = h(s)$ otherwise. An event E is *null* if $f_E h \sim f'_E h$ for all acts f and f', otherwise it is *non-null*. A constant act is an act that assigns the same consequence to all the states. To simplify the exposition, I denote the constant acts by their values (that is, if $f(s) = x$ for all s, I denote the act f by x).

The first postulate asserts that the preference relation is transitive and that all acts are comparable.

P.1: A preference relation is a transitive and complete binary relation on F.

The second postulate, also known as the sure thing principle, requires that the preference between acts depend solely on the consequences in states in which the payoffs of the two acts being compared are distinct. This implies that the valuation of the consequences of an act in one event is independent of the payoffs of the same act in the complementary event.

P.2: For all acts, f, f', h, h' and every event E, $f_E h \succsim f'_E h$ if and only if $f_E h' \succsim f'_E h'$.

The sure thing principle makes it possible to define conditional preferences as follows. For every event E, $f \succsim_E f'$ if $f \succsim f'$ and for every s not in E, $f(s) = f'(s)$.

The third postulate asserts that the ordinal ranking of consequences is independent of the event and the act that yield them.

P.3: For every non-null event E and all constant acts, x and y, $x \succsim y$ if and only if $x_E f \succsim y_E f$ for every act f.

In view of P.3, it is natural to refer to an act that assigns to an event E a consequence that ranks higher than the consequence it assigns to the complement of E as a *bet* on E. Ramsey (1931) was the first to suggest that a decision maker's belief that an event E is at least as likely to obtain as another event E' should manifest itself through preference for a bet on E over the same bet on E'.

The fourth postulate, which requires that the betting preferences be independent of the specific consequences that define the bets, formalizes this idea.

P.4: For all events E and E' and constant acts x, y, x' and y' such that $x \succ y$ and $x' \succ y'$, $x_E y \succsim x_{E'} y$ if and only if $x'_E y' \succsim x'_{E'} y'$.

Postulates P.1–P.4 imply the existence of a transitive and complete relation on the set of events that has the interpretation ‘at least as likely to obtain as’ representing the decision maker's beliefs as qualitative probabilities. They also imply that the decision maker's risk attitudes are event-independent.

The fifth postulate renders the decision making problem and the qualitative probabilities non-trivial by ruling out that the decision maker is indifferent among all acts.

P.5: For some constant acts x and x', $x \succ x'$.

The sixth postulate introduces a form of continuity of the preference relation. It asserts that no consequence is either infinitely better or infinitely worse than any other consequence. Put differently, the next postulate requires that there be no consequence that, if it were to replace the payoff of an act on a non-null even, no matter how unlikely, will reverse a strict preference ordering of two acts.

P.6: For all acts f, g, and h satisfying $f \succ g$, there is a finite partition $(E_i)_{i=1}^{n}$ of the set of states such that, for all i, $f \succ g_{E_i} h$ and $h_{E_i} f \succ g$.

A probability measure is non-atomic if every non-null event may be partitioned into two non-null sub-events. Formally, π is a non-atomic probability measure on the set of states if for every event E and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

number $0 < \alpha < 1$ there is an event $E' \subset E$ such that $\pi(E') = \alpha\pi(E)$. Postulate P.6 implies that there are infinitely many states of the world and that, if there exists a probability measure representing the decision maker's beliefs, it must be non-atomic. Moreover, the probability measure is defined on the set of all events, hence it is finitely additive (that is, for every event E, $0 \leq \pi(E) \leq 1$, $\pi(S) = 1$ and for any two disjoint events, *E* and *E* ′ , $\pi(E \cup E') = \pi(E) + \pi(E')$.

The seventh postulate is a monotonicity requirement asserting that, if the decision maker considers an act strictly better (worse) than each of the payoffs of another act on a given non-null event, then the former act is conditionally strictly preferred (less preferred) than the latter.

P.7: For every event *E* and all acts *f* and *f*• ′ , if $f \succ_E f'(s)$ for all *s* in *E* then $f \succcurlyeq_E f'$ and if $f'(s) \succ_E f$ for all *s* in *E* then $f' \succcurlyeq_E f$.

Representation

Savage's theorem establishes an equivalence between a preference relation having the properties described by the seven postulates and a preference relation induced by the maximization of the expectations of a utility function on the set of consequences with respect to a probability measure on the set of all events. The utility function is unique up to a positive affine transformation and the probability measure is unique.

Savage's theorem: : Let $\succcurlyeq$ be a preference relation on *F*. Then the following two conditions are equivalent:

1. (a) $\succcurlyeq$ satisfies postulates P.1–P.7.

Interpretation and criticism

In Savage's theory, consequences are assigned utilities that are independent of the underlying state of the world, and events are assigned probabilities that are independent of acts. These assignments, however, are not implied by the postulates. This observation merits elaboration.

The structure of the preference relation, in particular postulates P.3 and P.4, implies that the preference relation is state independent. In other words, the ordinal rankings of both consequences and bets are independent of the underlying events. This implies event-independent risk attitudes but does not, by itself, rule out the possibility that the states affect the decision maker's well-being, or that the utility of the consequences is state dependent. Put differently, Savage's model implies state-independent preferences but not a state-independent utility function. The utility and probability that figure in the representation of the preferences in Savage's theorem are unique as a pair, that is, the probability is unique given the utility and the utility is unique (up to a positive affine transformation) given the probability. It is possible, therefore, to define new probability measures and state-dependent utility functions – and thereby to obtain a new subjective expected utility representation – without violating any of Savage's postulates. For instance, let γ be a bounded, positive, non-constant, real-valued function on *S*, and let $\Gamma = \int_S \gamma(s)\, d\pi(s)$. For every event *E*, define $\hat{\pi}(E) = \int_E \gamma(s)\, d\pi(s) / \Gamma$ and let $\hat{u}(x, s) = \Gamma u(x) / \gamma(s)$ for all *s* in *S* and *x* in *C*. Then, for every act *f*,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
$\int_S u(f(s))\,d\pi(s) = \int_S \hat{u}(f(s), s)\,d\hat{\pi}(s)$. Because γ is arbitrary and non-constant, $\pi \neq \hat{\pi}$. This shows that the uniqueness of the probability in Savage's theory is predicated on the convention that the utility function is state independent (that is, constant acts are constant utility acts). This convention is not implied by the postulates, it has no choice manifestation, and its validity is not subject to refutation in the context of Savage's analytical framework. Moreover, the employment of this convention renders the definition of probability in Savage's model arbitrary and the claim that it represents the decision maker's beliefs scientifically untenable. That said, it is noteworthy that, in so far as the theory of decision making under uncertainty is concerned, because all the representations obtained using the procedure outlined above are equivalent, the failure to correctly quantify the decision maker's beliefs is not critical. In so far as providing choice-based foundations of Bayesian statistics is concerned, however, this failure is fatal.
A somewhat related aspect of Savage's model that is similarly unsatisfactory concerns the interpretation of null events. Ideally, an event should be designated as null and be ascribed zero probability if and only if the decision maker believes it to be impossible. In Savage's model an event is null if the decision maker displays indifference among all acts that agree on the payoff on the complement of the said event. However, this definition does not make a distinction between an event that the decision maker perceives as impossible and one whose possible outcomes he perceives as equally desirable. It is possible, therefore, that events that the decision maker believes possible, or even likely, are defined as null and assigned zero probability. Consider this example: a passenger who is indifferent to the size of his estate in the event that he dies is about to board a flight. For such a passenger, a plane crash is a null event and is assigned zero probability, even though he may believe that the plane could crash. This problem renders the representation of beliefs by subjective probabilities dependent on the implicit and unverifiable assumption that in every event some outcomes are strictly more desirable than others. If this assumption is not warranted, the procedure may result in a misrepresentation of beliefs.
The requirement that the preferences be state independent imposes significant limitations on the range of applications of Savage's theory. Choosing a disability insurance policy, for example, is an act whose consequences – the indemnities – depend on the realization of the decision maker's state of health. In addition to affecting the decision maker's well-being, it is conceivable that alternative states of disability influence his risk attitudes. Disability may also alter the decision maker's ordinal ranking of the consequences, which is a violation of P.3. For instance, a leg injury may reverse a decision maker's preferences between going hiking and attending a concert. Similar observations apply to the choice of life and health insurance policies.
Savage presented his seven postulates as principles that a rational individual ought to follow rather than an hypothesis describing how individuals actually choose among courses of action in the face of uncertainty. Indeed, almost from the moment of it inception, the descriptive validity of Savage's model – in particular, the sure thing principle, which is responsible for the specific functional form of the representation and the separability and linearity in the probabilities – has been qsted. It has repeatedly been shown in experimental settings that the theory fails systematically to predict subjects' choice. The most severe and remarkable criticism in this regard is due to Ellsberg (1961), who demonstrated using simple thought experiments that individuals display choice patterns that are inconsistent with the existence of beliefs representable by a probability measure.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Bernoulli, Daniel
- de Finetti, Bruno
- Ramsey, Frank Plumpton
- Savage, Leonard J. (Jimmie)
- state-dependent preferences
- utility
- von Neumann, John

Bibliography

Arrow, K. 1971. *Essays in the Theory of Risk Bearing*. Chicago: Markham Publishing Co.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiatatum Imperalas Petropolitanae* 5, 175–92. Translated as ‘Exposition of a new theory on the measurement of risk’, *Econometrica* 22 (1954), 23–6.

de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Annals de l'Institute Henri Poincaré* 7, 1–68. Trans. H. Kyburg in H. Kyburg and H. Smokler, eds., *Studies in Subjective Probabilities*. New York: John Wiley and Sons, 1964.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75, 643–59.

Hacking, I. 1984. *The Emergence of Probabilities*. Cambridge: Cambridge University Press.

Ramsey, F. 1931. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*, ed. R. Braithwaite and F. Plumpton. London: K. Paul, Trench, Truber and Co.

Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

How to cite this article

Karni, Edi. "Savage's subjective expected utility model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000479> doi:10.1057/9780230226203.1474

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sax, Emil (1845 – 1927)

K. Schmidt

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Sax was born in Jauernig (then, Austrian-Silesia; today, Javornik in Czechoslovakia). He studied in Vienna, where he became an university lecturer. After some practical activity (among other things in the railway organization), he became, in 1879, professor of political economy at the University of Prague. From 1879 to 1885 he was a member of the Vienna Chamber of Deputies. In 1893 he abandoned his Prague professorship and retired to Abbazia, Istria (then, Italy; today, Opatija in Yugoslavia), where he died in 1927.

Sax holds a peculiar place within the older Austrian school. He shared its basic idea according to which ‘value’ virtually opens the way to the explanation of all economic problems. However, in methodological matters, in the conception of economics, and in the interpretation of the value phenomenon itself, he went his own way. For Sax ‘value’ is not the rationally perceived significance of a commodity for the welfare of a person, but an emotional relationship between the person and the world of goods. His conception of economics is based on the distinction between individualism and collectivism. Behaviour is individualistic if it results (self-determined) from the individual personality; behaviour is collectivistic if the individual is motivated only as a member of a (larger and stable) group and in relation to this group. According to Sax, these two fundamental forces shape all economic and social phenomena in a characteristic way: the simple feeling of value becomes, individualistically, the exchangeable value; collectivistically, the complicated determination of value within a group. In relation to the social environment, these two fundamental forces can appear egoistic, altruistic and (as a mixture of both) mutualistic. Sax believed he had overcome the psychological one-sidedness of the classical school, and therefore he considered the findings of his theoretical work as exact results of inductive research.

According to Sax, the distinction between individualism and collectivism corresponds to the (value-based) theories of private and public economy. The absolute tax level (today called tax–GNP ratio) is determined by evaluations of such individuals in whom the fundamental force of collectivism is effective; they take care of the fact that the levels of the private and the public sphere are balanced. The relative tax level (today called tax apportionment to individuals) is deduced by Sax from the ‘equivalence of value’ of the individual tax liabilities; this results in the application of the equal sacrifice theory. For Sax, compulsion in taxation merely substitutes for a lack of correct insight. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
‘statemindedness of those governing’ and the ‘resistance of those governed’ prevent the abuse of power. Besides fiscal theory problems, Sax paid particular attention to the then young science of transport economics. His statements on transport policy (1878–9) are founded on a theoretically analysed historical experience, although the theoretical sections do not reach up to the abstract (mathematical) level such as Launhardt's treatment of transport problems. Despite the fact that much of Sax's writings on transport economics has since become obsolete, the book remains an excellent piece of applied economics.

Selected works

1878–9. *Die Verkehrsmittel in Volks- und Staatswirthschaft*, 2nd edn. 3 vols, Berlin: Springer, 1918–22.

1884. *Das Wesen und die Aufgaben der Nationalökonomie, Ein Beitrag zu den Grundproblemen dieser Wissenschaft*. Vienna: Hölder.

1887. *Grundlegung der theoretischen Staatswirthschaft*. Vienna: Hölder.

1924. *Die Wertungstheorie der Steuer*. Zeitschrift für Volkswirtschaft und Sozialpolitik, *NS 4*, Vienna and Leipzig: Deuticke.

Bibliography

Beckerath, E.von. 1930. Emil Sax, ein Nachruf. *Zeitschrift für Nationalökonomie*. 1, Vienna: Springer.

Beckerath, E.von. 1956. Sax, Emil. *Handwörterbuch der Sozialwissenschaften*, Vol. 9. Stuttgart: Fischer; Tübingen: Mohr; Göttingen: Vandenhoeck & Ruprecht.

How to cite this article

Schmidt, K. "Sax, Emil (1845–1927)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000022> doi:10.1057/9780230226203.1866

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Say, (Jean-Baptiste) L é on (1826 – 1896)

R.F. Hébert
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

public works; Say, (J.-B.) L.; Say, J.-B.; Turgot, A. R. J.

Article

French statesman, financier and economist, born in Paris in 1826; died there on 22 April 1896. He was the son of Horace Emile Say, the grandson of Jean-Baptiste Say, the nephew of Louis Auguste Say and Charles Comte. Léon Say became one of the most prominent statesmen of the French Third Republic. He served as Finance Minister from 1872 to 1879, and again in 1882, overseeing the largest financial operation of the century – payment of war reparations in Germany. His financial policies were directed towards a decrease in public expenditures and the removal of barriers to internal trade. A brilliant speaker and debater, he railed against socialism from the left and protectionism from the right. With Gambetta and Freycinet, he launched the ambitious programme of public works that bears the latter's name. Upon leaving the Cabinet, Say returned to his seat in parliament, assuming the leadership of the free trade party. He was at one time considered for the presidency of the republic, but was gradually set apart from his constituency by a rising tide of radicalism.

As an economist, Say's talents fall somewhere between the modest gifts of his father and the more imposing skills of his grandfather. He left no large work nor did he create any school of thought. Like his father, he was faithful to the doctrines of his namesake, and was a competent editor of his grandfather's works. In his youth he associated briefly with Léon Walras in a scheme to promote cooperative associations of production. He later became a frequent contributer to the *Journal des économistes*, mostly on economic policy, and a lecturer at the Ecole des Sciences Politiques, which was the prototype of the London School of Economics and Political Science. Say had a broad knowledge of history and theory, and he was capable of sustained exposition at a high level. As an example, his *Solutions dómocratiques de la question des impôts'* (1886) was directed against the idea of using taxation as a means of social equalization. He argued, instead, that the basis of taxation should always be real (based on property), never personal.

A curious parallel exists in the careers of Say and Turgot, whose name Say declared he could not even

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

pronounce without emotion. They shared a body of ideas and a similar destiny. Both achieved eminence as finance ministers in the French government, only to be turned out upon losing public favour. Say however, helped to immortalize his predecessor by writing one of the earliest biographies of Turgot.

Selected works

1886. *Les solutions démocratiques de la question des impôts*. 2 vols, Paris: Guillaumin.

1888. *Turgot*, trans. M.B. Anderson. Chicago: A.C. McClurg & Co.

1891. *Economie sociale*. Paris: Guillaumin.

1891–2. *Nouveau dictionaire d'économie politique*. 2 vols. Paris: Guillaumin.

1896. *Finances publiques, liberté du commerce*. Paris: Guillaumin.

1898–1901. *Les Finances de la France sous la Troisième République*. 4 vols. Paris: C. Levy.

Bibliography

Michel, G. 1900. *Léon Say, sa vie, ses oeuvres*. 2nd edn. Paris: Calmann Levy.

Picot, G.M.R. 1907. Léon Say. In *Etudes d'histoire contemporaine. Notices historiques*, vol. 2. Paris: Hachette.

How to cite this article

Hébert, R. F. "Say, (Jean-Baptiste) Léon (1826–1896)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000023> doi:10.1057/9780230226203.1476

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Say, Jean-Baptiste (1767 – 1832)

Thomas Sowell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate supply and demand; classical economics; labour theory of value; Malthus, T. R.; methodology of economics; Mill, J.; Ricardo, D.; Say's Law; Sismondi, J. C. L. S. de; Smith, A.; utility theory of value

Article

Although Jean-Baptiste Say is remembered primarily for Say's Law, one of the cornerstones of classical economics, he was also an early proponent of the utility theory of value, and was therefore very much at odds with his classical contemporaries, to whom labour was the source of value. Say's best-known work, his *Traité d'économie politique* (published in five editions, from 1803 to 1826) was intended as a shorter and more systematic presentation of economics than Adam Smith's *Wealth of Nations*. The success of this book made Say the best-known expositor of Smith in Europe and America, and he became in 1815 France's first professor of political economy. Translations of the *Traité* were used as textbooks at universities on both sides of the Atlantic.

Say was not, however, a mere uncritical expositor of Smith. The central importance of labour in Smith's discussions of value was replaced by Say's concern to show utility as the ultimate foundation of value. Production itself was defined as the production of utility, not of physical output. He noted in the first chapter of his *Traité* that this was subjective utility, which the economist must take as given data, however much moralists might attempt to change people's valuations. Businessmen also played a much more important and honourable role in Say than in Smith – Say having been a businessman himself and descended from a mercantile family. The *Traité d'économie politique* also went beyond Smith in developing what Say called 'one of the most important truths of political economy' – that supply creates its own demand, the doctrine ultimately named Say's Law.

Much controversy has surrounded the question of Say's originality in developing this principle, or rather, related series of principles. Claims have been made for James Mill as the real author of Say's Law. However, Mill's earliest published discussion of issues involving aggregate supply and demand came in an 1804 review in *The Literary Journal* of a book by Lauderdale – one year after the first edition of Say's *Traité* was published. While the chapter ('Des Débouchés') in which Say's Law was first set forth

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
was very brief in the first edition, there was further discussion of the same principle in that same edition, notably in Chapter 5 of the second volume. Later editions brought these scattered discussions together in an enlarged chapter on aggregate supply and demand, but much of the substance was there from the beginning. Mill was only the first of many to reformulate and elaborate what Say had done.

Say was concerned with the methodology as well as the substantive propositions of economics. He advocated systematic analysis rather than naive empiricism, but was also highly critical of the abstract deductive method of Ricardo and his followers. According to Say, Ricardo ‘pushes his reasonings to their remotest consequences, without comparing their results with those of actual experience’. During a friendly correspondence with Ricardo, Say pointed out that facts ‘are the masters of us all’. In the introduction to his *Traité d’économie politique*, Say also expressed his fear of ‘our always being misled in political economy, whenever we have subjected its phenomena to mathematical calculation’.

Say was in touch with the leading economists of his day, by mail and in person. He never resolved his differences with Ricardo as to whether value was based on labour or utility, but in attempting to clarify his position in 1822, Say spoke of ‘the last quantity of useful things’ as being crucial – a suggestion of the missing *marginal* concept essential to the utility theory of value. In his correspondence with Sismondi and Malthus, he came ultimately to reconcile Say's Law with their theories of aggregate disequilibrium. The fifth edition of his *Traité d’économie politique* in 1826 incorporated some of Sismondi's reasoning at the end of his chapter on Say's Law of markets (unfortunately, the English translation is from the previous edition) and called this to Malthus's attention as an admitted ‘restriction’ on this doctrine. A later textbook by Say, *Cours complet d’économie politique*, published in 1828–9, followed the chapter on Say's Law with one entitled ‘Limits of Production’, a phrase from Sismondi along with Sismondian analysis in the chapter.

Say was a policy-oriented economist rather than a model-builder like Ricardo. In his introduction to the new restrictions added to his chapter on the law of markets, Say remarked: ‘Now, we are studying practical political economy here.’ To Malthus he wrote: ‘It is better to stick to facts and their consequences than to syllogisms.’

How to cite this article

Sowell, Thomas. "Say, Jean-Baptiste (1767–1832)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000025> doi:10.1057/9780230226203.1477

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Say's Law

Thomas Sowell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand; aggregate equilibrium; circular flow; demand for money; gluts; Keynes, J. M.; Lauderdale, Eighth Earl of; Malthus, T. R.; Mercier de la Rivière, P.-P.; Mill, J.; Mill, J. S.; Owen, R.; quantity theory of money; Rodbertus, J. K.; saving function; Say's equality; Say's identity; Say's Law; Sismondi, J. C. L. S. de; Smith, A.; stagnationism; supply and demand; Walras's Law; Wicksell, J. G. K.

Article

Say's Law, the apparently simple proposition that supply creates its own demand, has had many different meanings, and many sets of reasoning underlying each meaning – not all of these by Jean-Baptiste Say. Historically, Say's Law emerged in the wake of the Industrial Revolution, when the two striking new economic phenomena of vastly increased output and the economy's cyclical inability to maintain sales and employment led some to fear that there was some inherent limit to the growth of production – some point beyond which there would be no means of purchasing it all. At the very least, some feared, there would not naturally or automatically be generated sufficient purchasing power to absorb the ever-growing output of the industrial economy, unless special policy arrangements were made to insure that income would be large enough to purchase output. Robert Owen and Karl Rodbertus exemplified these views, which were not those of any school of economists.

Say's Law attempted to answer such concerns by pointing out that the production of output tends of itself to generate purchasing power equal to the value of that output: supply creates its own demand. But Say's Law did not spring forth, full blown, like Minerva from the head of Zeus. It emerged piecemeal over a span of years, enveloped in controversies that ultimately involved nearly every noted economist of the early 19th century, and as its elaboration proceeded its definitions shifted under polemical stress. Moreover, the basic terms of discourse in economics were themselves in a process of evolution. The words ‘supply’ and ‘demand’ had different meanings for those economists like Sismondi and Malthus, groping towards the schedule or functional meanings of today, from those in the writings of David Ricardo or John Stuart Mill, who rigidly defined the terms as quantity supplied and quantity demanded. They repeatedly argued past each other.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The central meaning of Say's Law was implied by J.B. Say's rhetorical question, ‘how could it be possible that there should now be bought and sold in France five or six times as many commodities as in the miserable reign of Charles VI?’ This dramatized his main point, that there was no long-run limit to the growth of output, or of the demand for it. This did not deny that a *short run* derangement of the economy could take place, but Say described this as ‘an evil which can only be passing’. Moreover, he attributed these short-run phenomena to a wrong mixture of output, compared to consumer demand, rather than to aggregate overproduction. This was an ad hoc addendum, not logically implied by the principle of supply creating its own demand. Say thus created a subsidiary meaning of Say's Law – a denial that there was such a thing, even in the short run, as aggregate overproduction – which is to say, that there was no such thing as an equilibrium aggregate output. There could be a ‘partial glut’ of particular commodities produced in excess of demand but there could be no ‘general glut’ of commodities in the aggregate. It was this vulnerable subsidiary argument which Sismondi and Malthus attacked in the 19th century and Keynes in the 20th.

The first edition of Say's *Traité d’économie politique* in 1803 contained the crucial propositions of Say's Law, though not all in his chapter on markets (‘Des Débouchés’). The quantity of products demanded was ‘without a doubt’ determined by the quantity of products created, according to Say. ‘The demand for products in general is therefore always equal to the sum of the products’, he said. The distinction between secular stagnation and short-run downturns, and between partial and general gluts, was also present from the first edition of Say's *Traité*. These ideas all reappeared in James Mill's writings shortly afterward, but Say's priority is clear, both from the dates of the publications and from Mill's citation of Say in his own early writings, depriving him of even subjective originality.

The elder Mill did, however, make a significant contribution to the evolution of Say's Law. Where Say had asserted the inherent sufficiency of demand to purchase supply in terms of half the goods being essentially bartered for the other half, or in terms of saved money being spent as investment, Mill added the behavioural theory that people produced only because of, and only to the extent of, their demand for other goods. Each individual's supply equalled his demand *ex ante;* therefore society's supply must equal society's demand *ex ante*. Unfortunately, Mill also cited the *ex post* identity of supply and demand as evidence, as did J.R. McCulloch, Robert Torrens and John Stuart Mill.

While Say's priority over James Mill is readily established, the notion that in the long run aggregate demand ‘has no known limits’ was stated by the Physiocrat Mercier de la Rivière in the year of Say's birth. Nor was this a passing remark. His book, *L'Ordre naturel et essential des sociétés politique* (1767), especially Chapter 36, contained both the concept of a circular flow of money and of goods, and discussions of the conditions under which the existing level of aggregate output would be reproduced in subsequent time periods, as well as the conditions in which it would fall because receipts failed to cover the supply price of inputs. Yet it was not through Mercier de la Rivière but through Say that Say's Law entered the mainstream of economics. Ironically, both Say and Mill attacked Mercier de la Rivière. While his statement that aggregate demand is unlimited anticipated both of them, his concept of aggregate equilibrium and disequilibrium was anathema to the subsidiary version of Say's Law that was an integral part of the doctrine during the early 19th century.

Adam Smith also anticipated Say when he asserted in *The Wealth of Nations* (p. 407) that ‘a particular merchant’ could have a glut of goods but that a whole nation could not. Moreover, Smith's doctrine that savings rather than consumption promoted growth provided yet another dimension to Say's Law and the controversies surrounding it. One possible meaning of Smith's statement was that a shift in the savings

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
function – a willingness to save and invest more at a given rate of return – would tend to increase future output. But another interpretation, equally permissible in the absence of functional concepts, was that an increased quantity saved would promote future growth. Lord Lauderdale interpreted Smith in the latter sense, and attacked this proposition on grounds that there was some equilibrium level of savings and investment which, if exceeded, would reduce rates of return to a point that would cause the existing levels of savings to decline in subsequent time periods. In short, Lauderdale argued that there could be a general glut of capital, just one step from saying that there could be a general glut of aggregate output.
The leading critics of Say's Law during the classical era – Sismondi, Malthus and Lauderdale – all asserted short-run disequilibrium, not long-run stagnation, but the long-run comparative-statics approach of the Ricardians made it especially difficult for them to understand what the critics were saying in short-run dynamic terms. While Say himself ultimately came to understand – and reproduce in his later writings – the aggregate equilibrium theories which he now reconciled with the central meaning of Say's Law, more than 20 years later John Stuart Mill was still representing Lauderdale, Sismondi and Malthus as stagnationists. However, after completely misinterpreting their positions, J.S. Mill also set forth the most sophisticated analysis of the issues in classical economics in the second of his *Essays on Some Unsettled Questions in Political Economy* (1844).
After denouncing the ‘mistakes’, the ‘completely erroneous’ ideas and ‘palpable absurdities’ of those who emphasized the need for adequate aggregate demand, J.S. Mill nevertheless conceded that there could be ‘general excess’ in the sense that when money was not immediately respent, a seller ‘does not therefore necessarily add to the *immediate* demand for one commodity when he adds to the supply of another’. Thus there may be ‘a superabundance of all commodities relative to money’. Both the previous classical economists and such critics as Sismondi and Malthus had analysed the issue of aggregate output in essentially barter terms, despite incidental references to money. The theory of equilibrium aggregate output in Sismondi and Malthus was based on a balance of the utility of output and the disutility of the efforts required to produce it – a balance which could be temporarily unbalanced in their short-run dynamic models, though not in the long-run comparative statics model of James Mill.
John Stuart Mill's model, in which the role of money was important, was a different dimension, though not unique – Robert Torrens having expressed similar ideas more than two decades earlier. However, on the crucial issue of an aggregate equilibrium output in a non-monetary model, J.S. Mill remained adamant that there could be only internal disproportionality, not aggregate overproduction. Output, ‘if distributed without miscalculation among all kinds of produce in the proportion which private interest would dictate, creates, or rather constitutes, its own demand’. The issue remained to J.S. Mill one of internal ‘proportions’, not aggregate amounts.
Discussions of Say's Law virtually disappeared from economics for at least a generation after John Stuart Mill wrote on it in the 1840s. Even the sweeping challenges of neoclassical economics to classical orthodoxy, beginning in the 1870s, largely bypassed the issue of Say's Law. Isolated criticisms came from beyond the pale – from Marx, Hobson and assorted cranks. Within the economics profession, Say's Law was one of those things simply assumed and ignored. Early in the 20th century, Knut Wicksell explored the relationship between the quantity theory of money and Say's Law. The classical assumption that money is demanded only for transactions during the current period is incompatible with the price level being determined by the quantity of money, for a change in the price level requires that, at some point in the process, there must be either an excess or deficient money demand for goods in the aggregate, causing the general price level to go up or down in response. However, Wicksell's own

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
belated recognition by the English-speaking world meant that Say's Law did not become a major concern again until the appearance of John Maynard Keynes's *General Theory of Employment, Interest and Money* in 1936.

Modern, and especially Post Keynesian, discussions of Say's Law have revealed it to be not one, but a number of related, propositions. The most general of these propositions is that the aggregate value of goods supplied (including money) equals the aggregate value of goods demanded (including money). Thus an excess supply of goods is the same as an excess demand for money. This proposition has been christened ‘Walras’ Law’. Where there is assumed to be no excess demand for money, as in James Mill, for example, then aggregate supply is identically equal to aggregate demand. This proposition has been christened ‘Say's Identity’. When the equality of aggregate supply and demand is stated as an equilibrium condition – a sense in which equilibrium output theorists like Sismondi and Malthus could subscribe to it – then it merely states that both equilibrium and disequilibrium levels of output may exist. This proposition has been christened ‘Say's Equality’. This Ricardo, James Mill, and initially Say, all denied.

The Keynesian Revolution not only produced a more sophisticated theory of aggregate equilibrium, but also contributed to the distortion of Say's Law, which Keynes reduced to Say's Identity. According to Keynes, Say's Law ‘is equivalent to the proposition that there is no obstacle to full employment’. Only the cruder statements of the Ricardians said that.

Say's Law has been an important proposition in many ways. By indicating that the possibility of purchasing output from the income generated during its production is ultimately not limited by the mere size of output, Say's Law exposed the fallacy of recurrent popular fears that economic growth must collide with some impassable limit. The modern Post Keynesian delineations of the different senses of Say's Law (Walras’ Law, Say's Identity, Say's Equality) more precisely specify the conditions of aggregate equilibrium and disequilibrium, and indicate the theories of economic behaviour behind them. Finally, the long history of controversies over Say's Law sheds light on the enormous difficulties involved when even intelligent thinkers with honesty and goodwill try to understand each other's theories without clearly defined terms and without a clear sense of the conceptual framework of the opposing views. In short, its implications reach beyond economics to intellectual history in general.

Bibliography

Becker, G. and Baumol, W.J. 1960. The classical monetary theory: the outcome of the discussion. In *Essays in Economic Thought*, ed. J.J. Spengler and W.R. Allen. Chicago: Rand McNally.

Malthus, T.R. 1820. *Principles of Political Economy*. New York: Augustus M. Kelley, 1951.

Mill, J. 1803. *Commerce Defended*, ch. 6. London: C. and R. Baldwin.

Say, J.-B. 1803. *Traité d’économie politique*, vol. 1, ch. 15. Paris: Chez Rapilly, 1826.

Sismondi, J.C.L. de. 1827. *Nouveaux principles d’économie politique*, vol. 1, books 1, 4; vol. 2, Appendix. Paris: Delaunay Libraire.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sowell, T. 1972. *Say's Law: An Historical Analysis*. Princeton: Princeton University Press.

Sowell, T. 1974. *Classical Economics Reconsidered*. Princeton: Princeton University Press.

Spengler, J.J. 1960. The Physiocrats and Say's law of markets. In *Essays in Economic Thought*, ed. J.J. Spengler and W.R. Allen. Chicago: Rand McNally.

How to cite this article

Sowell, Thomas. "Say's Law." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000028> doi:10.1057/9780230226203.1478

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Scandinavia, economics in

Niels Kærgård, Bo Sandelin and Arild Sæther
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Scandinavia includes in a narrow sense Denmark, Norway and Sweden, which have similar languages and have strongly influenced one another. Danish economists made early contributions to neoclassical distribution theory, econometric analysis and multiplier theory. Like most economists from small-language communities they understood the major European languages but wrote in their domestic languages, which delayed international knowledge about their contributions. In Norway Ragnar Frisch revolutionized economics in the 1930s, but met opposition from colleagues. Swedish economics flourished in the early 20th century with Knut Wicksell and Gustav Cassel and later with the Stockholm School. In recent decades national traits have largely disappeared.

Keywords

Aarum, P. T.; Agardh, C. A.; Aschehough, T. H.; balanced budget multiplier; benefit principle; Berch, A.; Böhm-Bawerk, E. von; Borch, K. H.; cameralism; capital asset pricing model; Cassel, G.; central bank independence; Christian-Social-Historical School; Chydenius, A.; Cowles Commission; Davidson, D.; econometrics; Enlightenment; Frisch, R.; Gelting, J.; German Historical School; Haavelmo, T.; Hammarskjöld, D.; Heckscher–Ohlin trade theory; Hecksher, E. F.; Hoffmeyer, E.; Holberg, L.; identification; Institute for International Economic Studies (Stockholm); Jæger, O.; Johansson, A.; Johansen, L.; Kock, K.; Kydland, F.; Lindahl, E. R.; Lundberg, E.; Malthus, T. R.; marginal revolution; mathematical maximization theory; Meidner, R.; monopolistic competition; Mossin, J.; multiplier analysis; Myrdal, G.; neoclassical growth theory; Norwegian School of Economics and Business Administration (NHH); Ohlin, B.; Oslo School; Prescott, E.; Pufendorf, S.; Rabenius, L.; Rehn, G.; Sandmo, A.; Scandinavia, economics in; Schweigaard, A.; Steffen, G.; Stockholm School; Swedish model; Walras, L.; Warming, J.; Wedervang, I.; Westergaard, H.; Wicksell, K.; Zeuthen, F.

Article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Denmark – Norway before 1814

Ludvig Holberg (1694–1754) was the first to forcefully advocate political economy as a science and academic discipline in the dual monarchy of Denmark–Norway. Today he is recognized as the Molière of northern Europe, but few know about his contribution to economics. Strongly influenced by the natural law philosopher Samuel Pufendorf (1632–94), but also by English and French Enlightenment philosophy, he studied economic phenomena and problems from the perspective of moral philosophy. His contributions cover a wide spectrum and include his literary authorship, his achievements at the University of Copenhagen, and his efforts in re-establishing Sorö Academy as a modern centre of higher education. The latter was made possible in 1747 when Holberg bequeathed his estates to Sorö Academy. The Academy placed modern sciences on in its curriculum, and quickly became an alternative to the University of Copenhagen, which was marked by a strong theological influence. At the Academy students were taught ‘Political Economy, Commerce and Cameral Sciences’. This proved a great success. In the second half of the 18th century the Academy functioned as the academic home for social scientists and social critics. At Sorö Jens Schelderup Sneedorff (1724–64) was appointed the first professor of political economy and public law in 1751. He and his successor Andreas Schytte (1726–77) were influenced both by German cameralism and, like Holberg, by English and French philosophers. Schytte wrote the first textbooks in political economy in the local language.

At the university Ole Stockfleth Pihl (1729–65) was appointed the first professor in political economy in 1761. Before that he had been the editor and publisher of the monthly *Oeconomisk Journal*. As the position had no salary he resigned after two years. The next professor, Johan Christian Fabricius (1745–1808), was appointed in 1772. His salary was so small that after four years he accepted a chair in natural history, political economy and cameral sciences at University of Kiel, the second university in the dual monarchy. Here he created ‘an economic garden’ and wrote a widely used textbook in political economics.

Another important factor in the development of political economy as a science was the establishment of *Danmark og Norges Oeconomiske Magazin* in 1756. Its initiator and editor, the bishop of Bergen Erik Pontoppidan (1698–1764), was an enlightened mercantilist, whom the king had called on to carry out reforms at the university. His reforms were not successful, but the *Economic Magazin* became a sanctuary and a workshop for those who were engaged with economic questions in the middle of the 18th century. Otto Diderich Lütken (1713–88), its most important contributor, is considered the most original of economic thinkers in Denmark–Norway in the 17th century. He published several essays discussing theoretical as well as practical economic issues. In one article he claims, as Malthus did 40 years later, that there is a connection between population and available food, and that population would increase until a shortage of food put an end to further growth.

The influence of the *Economic Magazin* was considerable. People connected with this periodical and with Sorö Academy gained influence from the mid-18th century and far into the 19th century. Their thinking found expression in the agricultural and social reforms carried out towards the end of the 19th century, and is credited with giving impetus to a translation of Adam Smith's *Wealth of Nations* in 1779, initiated by Norwegian tradesmen.

Denmark after 1814

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Characteristics of Danish development

Unlike in many other countries, social development in Denmark has been characterized by its continuity. No political changes have been sufficiently radical to create a break in academic traditions. In 1849 the country changed from being an absolute monarchy to being a democracy, but A.F. Bergsøe was professor in political economy during the whole period 1845–54. In 1901, after decades of political struggle, a right-wing government was replaced by parliamentarism and a left-wing government, but H. W. Scharling, Minister of Finance in the right-wing government, was professor in political economy without interruption from 1869 to 1911. Furthermore, the economic systems in Denmark have not been subjected to disturbances great enough to leave traces in the science of economics. The Danish economy has been marked by relatively stable growth for 200 years.

Apart from continuity, a dominant fact is the small size of the country. Until 1936, the University of Copenhagen was the only institution offering university-level teaching in economics. Courses in economics were started in 1848, and from that time there were two chairs in political economy, one of them a chair in cameralism and public economics dating back to 1762. In 1886, Harald Westergaard was given a personal chair, bringing the number of professors to three. This number did not increase again until after the First World War, and not dramatically until after the Second World War. In the 1960s and 1970s, the number of professors begins to shoot up, with five universities and 14 full professors in economics in 1960, and eight universities and 34 chairs in 1995. These figures actually underestimate the growth; before 1960, there were scarcely any teachers who were not full professors, whereas the number of assistant and associated professors today is much greater than the number of professors.

International contacts

Perhaps the tiny domestic research environment made international contacts even more necessary than would have been the case in larger countries. An example of this can be seen in the marginal revolution at the beginning of the 1870s. When this revolution was started by Jevons, Menger and Walras, they had no contact with each other. Jevons died in 1882 without having heard of Menger (Howey, 1972). In Denmark, the publication of Walras's first volume was reviewed by *Nationaløkonomisk Tidsskrift* in 1875, and the reviewer (an institutional economist, A. Petersen) found, ‘surprisingly’, that Walras did not know Jevons's work (Kærgård, 1996).

Danish economists corresponded in French with Walras and in English with Jevons at a time when Danish economics was largely German-oriented. Contact with other Scandinavian economists was particularly close. From 1863 onwards, there were regular Scandinavian conferences on political economy, and a ‘Marstrand Meeting’ for Scandinavian economic researchers was arranged for most years between 1936 and 1985.

All this has changed in recent decades, when Danish economists have become an ordinary part of the general international research community. At the same time, the trend has been to move from publishing in books towards publishing in international journals, and from publishing in Danish to publishing in English.

Danish contributions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

If we consider the contributions made by Danish economists, we find none who have come even close to an established position in the history of economics. Hutchinson's standard text *A Review of Economic Doctrine 1870–1929* mentions 359 names; these include six Swedes, one Norwegian and no Danes (Boserup, 1980).

Thus the history of Danish economics cannot until the most recent decades show internationally known names, but it is filled with overlooked precursors and with discoveries described in Danish, never known outside the domestic border. Mentioned can be Otto Ditlev Lütken, who wrote on population growth and scarcity of food before Malthus (Sæther, 1993); Bing and Julius Petersen, who wrote on neoclassical distribution theory in 1873 (Whitaker, 1982); Westergaard, who in the 1870s was the first to use mathematical maximization theory in economics (Creedy, 1980; Davidsen, 1986; and Kærgård and Davidsen, 1998); Mackeprang, whose thesis of 1906 was the first econometric analysis (Kærgård, 1984); Warming's description of the identification problem from 1906 (Kærgård, 1984, and Kærgård, Andersen and Topp, 1998); Wulf and Warming's development of the multiplier theory from 1896 to 1932 (Boserup, 1969; Topp, 1981); Frederik Zeuthen's discussion of monopolistic competition in the late 1920s (Brems, 1976); Jørgen Pedersen's description of fiscal policy in 1937 (Topp, 1988) and Gelting's derivation of the balanced budget multiplier in 1941 (Hansen, 1975). None of these discoveries were published in English, and were not made known internationally until after the theories had become widely known.

There are several possible explanations for the high number of unknown Danish contributions to economic theory. Brems (1986) suggests two barriers to the dissemination of economic theories: a linguistic barrier (Anglo-Saxons do not speak German and French) and a mathematical barrier. Unlike the economists in the larger European countries, those from the small-language communities such as Denmark understood all major languages, and were therefore better acquainted with all the international schools and could combine their ideas. However, economists from smaller countries often wrote in their own language, and consequently their work never became widely known. With so few professors of economics in small countries, it was furthermore necessary for them to be very versatile, and they therefore tended to move from one subject to another, and a more persistent approach is necessary to be established as a pioneer (one might recall Walras's battle over years to achieve recognition).

Danish economists. generalists not specialists

We can see from the above that the relationship between Danish economists and the various international schools has been like that between butterflies and flowers. They have generally fluttered from school to school, taking from each what they felt useable; very few Danish economists have been orthodox disciples of one of the recognized schools. A couple of examples can be mentioned. At the time of the great methodological battle in Germany between the neoclassical school and the Christian-Social-Historical School, Westergaard was in close contact with both schools. In the 1870s, he corresponded with Jevons concerning mathematical-economic problems, and at the same time he was the leading representative in Denmark of the Christian-Social-Historical school (Kærgård, 1995). During the debate among neoclassicists, monetarists and Keynesians in the 1960s and 1970s, Anders Ølgaard played a central role in economic debates in Denmark as Chairman of the Danish Board of Economic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Advisors, arguing from a typical Keynesian viewpoint, at the same time as he was writing a substantial treatise on neoclassical growth theory (Ølgaard, 1966).

Danish economists were almost always more than just theorists. Between 1870 and 1970 there were a total of 15 people who held chairs in political economy at the University of Copenhagen; of these, six were members of parliament, and of those six three held posts as members of the government. Another was member of the Copenhagen municipal government. Among the remaining eight were Bertil Ohlin, who was in Copenhagen only for a short time and later became politically active in Sweden; Erik Hoffmeyer, who was the first director of the Central Bank of Denmark for more than 30 years; Harald Westergaard, who was a leading church politician and social reformer; and Carl Iversen and Anders Ølgaard, two chairmen of the Board of Economic Advisors. All 15 were active in the public debate, writing numerous newspaper articles. Jointly, they held almost innumerable positions in commercial life, councils and commissions. Recent decades have revealed a completely different type of university economist, with purely academic and theoretical interests.

Norway after 1814

In 1814 Christen Smith (1785–1816) was appointed professor of botany and political economy at the first Norwegian university in Oslo. Botany and political economy would be considered a strange combination of subjects nowadays, but at that time the logic of such an arrangement was clear. The wealth of nature would create prosperity for the people. Unfortunately, before he could take up his position Smith died during a British-led botanic expedition to Congo.

Breakthrough of political economy

His successor, Gregers Fougner Lund (1786–1836), was not appointed until 1822. He wanted to move political economy away from mercantilism towards economic liberalism. His views were supported by Jacob Aall (1773–1844), ironmaster and member of parliament, who, with his essays on economic problems, became highly influential.

Anton Schweigaard (1808–70), who took over the chair in law, political economy and statistics in 1836, dominated economic thinking in Norway for almost half a century. He supported the liberal economic policy recommended by the classical economists. However, he did not follow them blindly, since in his opinion they sometimes carried their policies too far. In spite of being a spokesman for free trade he rejected the doctrine of laissez-faire. On many questions he was closer to the Continental economists, especially Say in France and Hermann and Rau in Germany. With Schweigaard political economy had gained a firm foothold as a science.

Tradition and renewal

When Schweigaard died, his former student, Torkel H. Aschehoug (1822–1909), professor of law, took over his teaching responsibilities in economics and statistics. Until his death he dominated political economy within the academic world and beyond. He was behind, or strongly supported, several important events: the creation of the Statistical Bureau of Census in 1876 and the establishment in 1883

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of *Statsøkonomisk forening*, an association for Norwegian economists, which he chaired for 20 years. The latter became a forum where the enlightened elite of bureaucrats, government ministers, parliamentarians and academics discussed economic issues. From 1887 it published an economic journal, *Statsøkonomisk Tidsskrift* (renamed *Norsk Økonomisk Tidsskrift* in 1997), in which economists discussed both theoretical and practical issues.

A second chair in ‘pure economics’ was created in 1877, and an independent study programme in ‘political economy’ was established in 1905. Aschehough wanted to give an account of economic science in the Norwegian language. The first edition of his *Socialøkonomik*, completed in 1891, dealt with the theories of the classics, and the moderns Böhm-Bauwerk, Jevons, Menger, Schmoller and Walras. Later editions, however, were strongly influenced by Marshall's *Principles of Economics* and in particular his theory of value.

Professional build-up

Oskar Jæger (1863–1933), Peder Thorvald Aarum (1867–1926) and Ingvar Wedervang (1891–1961) were the central persons in the Norwegian economic profession between Aschehoug and Ragnar Frisch (1895–1973).

Jæger's contributions span from treatises on methodology to thoughts on public finance, including an active, although disputed, participation in economic politics. His historical lectures in political economy were concerned with the development of ‘modern’ analysis from an Austrian point of view. He mentions Marshall, of course, but his mainstay is Böhm-Bawerk.

Aarum's university career was relatively short, but his textbooks in theoretical and practical economics gave him considerable influence. He followed Marshall, and claimed that the interactions of demand and supply in the market simultaneously determined price and quantity. Market equilibrium became a key concept. He also introduced the extensive use of diagrammatic exposition in his lectures and books. Aarum was regarded as ‘the modern’ among Norwegian economists.

Wedervang is considered one of the great profession builders in Norwegian economics. He was behind the parliament's decision to create a new chair in economics at the university in 1931 for Ragnar Frisch. Together with Frisch, he was behind the establishment of the Rockerfeller supported economic institute in 1932. He and Frisch were its first directors. Furthermore, he was in the forefront when a new five-year study programme in economics was adopted in 1934, and when the parliament decided to establish the Norwegian School of Economics and Business Administration in Bergen in 1936.

The Oslo School

In 1919 Ragnar Frisch graduated from the study programme in political economy. After studies in France, Germany, England, the USA and Italy he became Aarum's research assistant in 1925. After defending his doctoral dissertation *‘Sur un problème d’économie pure’* in 1926 he again went to the United States, but returned when the university made its offer in 1931. During the 1930s Frisch participated actively in international economic activities and conferences. He was among the small group of initiators who, in 1931, established the Econometric Society. In 1933 he became the first editor of *Econometrica*, a position he held for more than 20 years. When the Prize in Economic Sciences in Memory of Alfred Nobel was created in 1969, Frisch together with Jan Tinbergen received it for their

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
development and application of dynamic models for the analysis of economic problems.
In the beginning of the 1930s not only Norway's economy was at a low ebb, so too was the status of economics as a science. Frisch started his grand project of bringing economics as a science out of ‘the fog’. He believed that economic theory should be based on mathematical models and quantitative analysis. The new economics should be shaped in a precise mathematical language. It was only with mathematical models that it would be possible to carry out complicated analysis and reasoning. He promoted this with enthusiasm, genius and force. All opposition was brushed aside. The study programme in economics was changed into a programme with a strong emphasis on mathematical analysis and economic research. His best students were attached to the institute as research assistants. Frisch created a revolution, but change did not come without conflict. He was applauded, but also met with opposition from his colleagues. There was, however, no organized opposition against him. When Wedervang left in 1937 to become the first rector of the new business school, his and other positions were filled with Frisch's students. On the strength of the new study programme and his new staff, Frisch succeeded, in a short time, in creating his own school within economic research. This Oslo School, which to this day influences Norwegian economics, particularly at the University of Oslo, broke with tradition by introducing quantitative methods into economic research and teaching. The development of national accounts, national budgets and economic planning were given top priority. This work was strengthened by Leif Johansen (1930–82), who became Frisch's assistant in 1952 and took over Frisch's chair when he retired in 1965. Among Johansen's most important contribution was his doctoral dissertation, ‘A Multi-Sectoral Study of Economic Growth’, which became the basis for the long-term economic planning by the Ministry of Finance. Macroeconomic planning, research and policy became the alfa and omega in the Norwegian post-war economy.
Trygve Haavelmo (1911–99) joined Frisch as a research assistant in 1933. In 1938 he was visiting professor at University of Aarhus and in 1939 a research fellow at Harvard University. Caught in the United States by the war, he worked for Nortraship, an organization set up by the Norwegian government in exile to administer the war effort of the Norwegian merchant marine. After the war he stayed a year with Cowles Commission in Chicago, where, according to Schumpeter (1954, p. 1163), he ‘exerted an influence that would credit to the lifetime work of a professor’. On returning to Norway he was appointed professor of economics in 1948, a position he held until his retirement in 1979. With his research contributions, teaching, generosity and gentle personality, he had a decisive influence on the development of economics. He won the Nobel Prize in 1989 for his fundamental contributions to econometrics.
During the economic depression of the inter-war period Frisch developed a deep mistrust of the market economy and the working of the price mechanism. National economic planning administered and managed by well-trained economists was, in his opinion, clearly superior to the shifting bustles of the market. As a consequence Frisch, as well as Johansen, who was a member of the Communist Party, were great admirers of the Soviet economic planning system, and claimed it was superior to the market economies of the Western world. They were therefore not easily attuned to other ideas.

Challenges to the Oslo School

Karl H. Borch (1919–86) was in 1959 recruited to the Norwegian School of Economics and Business Administration (NHH), first as a university fellow and from 1963 as a professor of insurance. This

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
institution was at the time not so strongly focused on research. However, Borch stood out as an eminent researcher and a spiritual leader for the younger researchers. With his international network he strongly urged his students to pursue doctoral studies abroad and particularly in North America.

The new competence-building and international recognition achieved by Borch and his colleagues, together with the economic developments in the 1970s, slowly broke the monopoly and the influence of the Oslo School in Norwegian economics and politics. Economic planning in the Frisch–Johansen tradition was from the mid-1980s no longer central. More emphasis was placed on the functioning of competitive markets under uncertainty. Two of Borch's students should in particular be mentioned: Jan Mossin (1936–87) and Agnar Sandmo (b. 1938). Mossin was among a group of international researchers who independently contributed to the development of the modern theory for financial markets, the capital asset pricing model. Sandmo's research, to a large extent focused on the theory of taxation, is based on the assumption that we live in a world where we must deal with uncertainty and where there are limited opportunities for action. Markets and social institutions do not function in an ideal way. We must accept compromises and second-best solutions. Another prominent economist at NHH, Victor D. Norman (b. 1946), who earned his Ph.D. from MIT in 1972, has made significant contributions to international trade theory.

This work had a marked influence on Norwegian monetary and fiscal policies and also laid the basis for increased independence of the central bank. This line of research was also pursued by the Norwegian, Finn E. Kydland (b. 1943). from NHH, who, in 2004, together with Edward C. Prescott (b. 1940) was awarded the Nobel Prize for their contribution to dynamic macroeconomics, notably the time consistency of economic policy and the driving forces behind business cycles.

Sweden

Institutional evolution

As an academic discipline, political economy in Sweden can be dated back to 1741, when the first chair was created at Uppsala University. Official policy in the mid-1700s aimed at promoting economic growth, and economic debate was flourishing. The creation of three more chairs in political economy before 1760 was an element in this effort. However, because of changed priorities and loss of territories, a decline soon set in, and during most of the 19th century political economy at Swedish universities was quite weak.

At the end of the 19th century and beginning of the 20th century, culture and science in Sweden were especially influenced by the German-language area. Study tours were mainly directed to Germany. Doctoral theses in economics were written in either Swedish or German; the first in English was published in 1929, while virtually all have been in English in the 2000s. Half of the books on economics acquired by university libraries in 1903–7 were published in Germany or Austria, and only a fourth in the UK or the USA. Fifty years later the proportions were almost reversed. The two world wars and Nazi oppression help explain the transition from German to Anglo-Saxon influence, as do faster growth of American population and academic research, less importance of geographical proximity and possibly a lingering dominance of the Historical School in Germany (Sandelin, 2001).

In the absence of a better measure, the growth and specialization of academic economics may be described by the number and scope of chairs. At the beginning of the 20th century, there were only two

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
chairs of economics in Sweden, one in Uppsala (David Davidson) and one in Lund (occupied in 1901 by Knut Wicksell). Both were located in the faculty of law, and both also included fiscal law. In 1903 a chair in economics and sociology was created in Gothenburg (Gustaf Steffen), and in 1904 one in economics and public finance was created in Stockholm (Gustav Cassel). By comparison, in 1996, a few years before the principles for appointing professors were radically changed, there were 57 chairs in Sweden, of which 45 were directed towards a special field within economics, and only one formally included more than economics (Sandelin, 1998, p. 2; 2000; Sandelin, Sarafoglou and Veiderpass, 2000, p. 46).

Early Swedish economists such as Wicksell, Cassel, Heckscher, and the Stockholm School economists had a common forum in the journal *Ekonomisk Tidskrift*, founded by David Davidson in 1899. Its name was changed in 1965 to *The Swedish Journal of Economics* — then in 1976 to *The Scandinavian Journal of Economics*, when the circle of contributors and editors was widened. Those changes left room for a Swedish-language journal directed to a broader audience and dedicated to practical economic problems; so *Ekonomisk Debatt* was born in 1973, published by *Nationalekonomiska föreningen*, the economists' association, founded in 1877.

Before the neoclassical breakthrough

International currents are visible in early Swedish thought. Some early authors have been labelled mercantilists, the most influential probably Anders Berch, who became the first professor of political economy appointed at a Swedish university (Uppsala, in 1741), and who published the first textbook, *Inledningen til almänna hushålningen* (1747), which then enjoyed a monopoly in academic teaching for more than 80 years. Opponents of mercantilist ideas arose, among them the clergyman Anders Chydenius, who published liberal booklets in the mid-1760s that have resulted in him being called a Swedish physiocrat.

Despite this beginning, political economy at Swedish universities was not strong at the beginning of the 19th century, though there were a few representatives of classical economic thought. Lars Rabenius – who appreciated Adam Smith's ideas, though with reservations – published a textbook in 1829 which finally replaced Berch's old book. Carl Adolph Agardh, who had attended and been influenced by Say's lectures in Paris in the 1820s, thought that the classical economists gave the state too modest a role, so his ideas were more akin to the Historical School.

Gustaf Steffen, who was professor of economics and sociology in Gothenburg during 1903–29, was the last Swedish professor who can be classified with the Historical School. The majority of university economists during this period were turning towards neoclassical ideas (Lönnroth, 1991, 1998; Magnusson, 1987).

The early modern generation

David Davidson, Knut Wicksell and Gustav Cassel introduced modern economics into Sweden around 1900; Eli Heckscher may also be included in this group, although his main works were published later. David Davidson (1854–1942) was Wicksell's teacher and an important adviser on domestic monetary and fiscal policy, though he did not address an international audience. The editor of the *Ekonomisk*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Tidskrift for 40 years and one of its main contributors, he also influenced policy directly as a member of several government committees on taxation (Uhr, 1991).

Knut Wicksell (1851–1926), an ardent participant in public discussions on social matters of all kinds, was the most important introducer of neoclassical economic thought into Sweden. His book *Value, Capital and Rent* (1893) was permeated with derivatives and marginalist concepts. The original German edition has the subtitle *nach den neueren nationalökonomischen Theorien* ('according to the new economic theories') and it is the theories of Walras, Jevons, Menger, and — especially concerning capital — Böhm-Bawerk that he tries to bring together.

Wicksell's analysis of just taxation from the perspective of the benefit principle in his next book, *Finanztheoretische Untersuchungen* (1896), has become an unavoidable point of reference. Likewise, his idea of a cumulative process of inflation, expounded in *Interest and Prices* (1898), is still referred to. Wicksell's *Lectures on Political Economy* (vol. 1, 1901; vol. 2, 1906) is not simply a textbook version of the ideas developed in his earlier books, but contains refined and modified approaches to questions raised earlier.

Gustav Cassel (1866–1945) began as a mathematician but turned later to economics. He studied with Wagner and Schmoller in Berlin, and around 1900 evinced doubt about the benefit of unlimited competition; later he became more sceptical of government intervention. His basic economic thought, expounded in his *Theoretische Sozialökonomie* (1918), was evidently much influenced by Walras, although he did not give him proper credit in that book. During the 1920s Cassel worked in various positions with international monetary problems, and during many decades he was, like Wicksell, Heckscher, Ohlin, and others, a persistent participant in public discussions, publishing several hundred newspaper articles (Magnusson, 1991; Carlson, 1994).

Especially because of his book *Mercantilism* (1931), the youngest in the group, Eli F. Heckscher (1879–1952), is internationally known mainly as an economic historian. As such, he pleaded for the integration of historical and neoclassical analysis. His lasting contribution to economic theory is an article in the *Ekonomisk Tidskrift* in 1919, which provided the basis of the so-called Heckscher–Ohlin theory in international trade (Henriksson, 1991).

As noted, the early modern generation was extensively involved in public debate; Wicksell considered it his 'foremost duty to educate the Swedish people' (Jonung, Hedlund-Nyström and Jonung, 2001, p. 19). This attitude was taken over by several of the Stockholm School economists.

The Stockholm School

Both Cassel and Heckscher were advocates of traditional economic liberalism, and sceptical of major government intervention. Around 1930, when nobody could overlook the problem of unemployment, this scepticism was challenged by a group of young economists, some of whom were disciples of Cassel and Heckscher. Dag Hammarskjöld (1905–61), Alf Johansson (1901–81), Karin Kock (1891–1976), Erik Lindahl (1891–1960), Erik Lundberg (1907–87), Gunnar Myrdal (1898–1987), Bertil Ohlin (1899–1979), and Ingvar Svennilsson (1908–72) were members of this group, called 'the Stockholm School' by Ohlin in an article in the *Economic Journal* in 1937. Ohlin believed that the Stockholm School had developed a theory of employment and had demonstrated how employment can be stimulated by economic policy, before Keynes did so.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Although several individual members are well-known, sometimes for other contributions, such as Myrdal's institutional analyses and Ohlin's contributions to the theory of international trade, the Stockholm School did not live on in the way Keynes's thinking did; it was hardly more than a national phenomenon, for several reasons. The few university positions in Sweden could not absorb all of them. They wrote mainly in Swedish, often in the form of government reports. They emphasized the dynamics of economic problems, which were difficult to present pedagogically. They also tended to analyse special rather than general cases, in the belief that useful general conclusions were difficult to draw. And their approach in some ways conflicted with techniques coming into vogue after the Second World War (Siven, 1985; Jonung, 1987).

The first post-war decades

Although it can be considered as a national phenomenon in the sense that it had little influence outside Sweden, the Stockholm School itself was not devoid of influences from abroad. It was mainly a theoretical school, and theory is more international than empirical knowledge.

Swedish economics was hardly more internationalized in the 1940s and 1950s than it was during the preceding decades. As before, most economic research was performed when students wrote their *magnum opus*, the doctoral dissertation. (A new system of graduate education, similar to the American, was introduced in 1969.) And most of the dissertations were more empirical than theoretical, focusing on the Swedish economy.

Outside the university world, the trade union economists Rudolf Meidner and Gösta Rehn developed ideas on the relationship between inflation and employment, and recommended a general deflationary policy, combined with selective measures directed towards those parts of the economy that would suffer from it. The latter part of the recommendation – selective means – was politically accepted and characterized the ‘Swedish model’ of actual economic policy for many years.

Disappearance of national traits

The closer we come to our own time, the less reason there is to classify economic thought geographically. National traits have become less evident as communications have improved. In an evaluation of Swedish economic research, Dixit, Honkapohja and Solow (1992, p. 129) concluded that

> over the past three or four decades the literature of analytical economics has become almost completely homogenous worldwide. Mainstream economists in all countries now contribute to a single international literature as part of a single intellectual community. … One can easily imagine a new idea or technique arising anywhere in the world of mainstream economics, and being pursued at first by its originator and his or her graduate students, but one cannot easily imagine a distinctively national school arising within the mainstream. Good ideas circulate much too rapidly.

Nevertheless, we may point to a couple of Swedish characteristics. Persson, Stern and Gunnarsson (1992, p. 118) found that non-mainstream economists like neo-Ricardians and Post Keynesians were

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
much less cited by Swedish economists than by the world's economists on average, and this probably remains true. Similarly, as found by Dixit, Honkapohja and Solow (1992, p. 139), the application of advanced econometric techniques seems still to prevail over the creation of new ones.
Stockholm University's Institute for International Economic Studies has remained the most successful research unit in Sweden for several decades. As in other small European countries, many Swedish university economists have traditionally been involved in government committees and commissions. A change may have occurred in recent years, however, partly as a consequence of faster growth in the supply of qualified economists than in the demand for people willing to accept such side-commissions, and partly because young economists may be giving pure research higher priority than before.

See Also

- Cassel, Gustav
- Davidson, David
- Frisch, Ragnar Anton Kittel
- Haavelmo, Trygve
- Heckscher, Eli Filip
- Johansen, Leif
- Lindahl, Erik Robert
- Lundberg, Erik Filip
- Myrdal, Gunnar
- Ohlin, Bertil Gotthard
- Stockholm School
- Wicksell, Johan Gustav Knut
- Zeuthen, Frederik Ludvig Bang

Bibliography

Aarum, T. 1924–8. *Læren om Samfundets Økonomi I and II*. Kristiania (Oslo): Olaf Norlis Forlag.

Andersen, P. 1983. On rent of fishing grounds: a translation of Jens Warming's 1911 article, with an introduction. *History of Political Economy* 15, 391–6.

Aschehoug, T. 1908. *Socialøkonomik*. Kristiania (Oslo): Aschehoug og Nygaard forlag.

Bergh, T. and Hanisch, T. 1984. *Vitenskap og politikk. Linjer i norsk sosialøkonomi gjennom 150 år*. Oslo: H. Aschehoug & Co. (W. Nygaard).

Bisgaard, H. 1902. *Den danske nationaløkonomien idet 18. århundre* [Danish Economic Thought in the 18th century]. Copenhagen: Hagerups forlag.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Boserup, M. 1969. A note on the prehistory of the Kahn multiplier. *Economic Journal* 79, 667–9.

Boserup, M. 1980. The international transmission of ideas: a small-country case study. *History of Political Economy* 12, 420–33.

Brems, H. 1976. From the years of high theory: Frederik Zeuthen (1888–1959). *History of Political Economy* 8, 400–11.

Brems, H. 1986. *Pioneering Economic Theory 1630–1980: A Mathematical Restatement*. Baltimore: Johns Hopkins University Press.

Carlson, B. 1994. *The State as a Monster: Gustav Cassel and Eli Heckscher on the Role and Growth of the State*. Lanham, NY and London: University Press of America.

Cassel, G. 1918. *The Theory of Social Economy*. Revised English translation, New York: Kelley, 1967.

Creedy, J. 1980. The early use of Lagrange multipliers in economics. *Economic Journal* 90, 371–6.

Danmark og Norges Oeconomiske Magazin (1757–64), 1–8. Copenhagen.

Davidsen, T. 1986. Westergaard, Edgeworth and the use of Lagrange multipliers in economics. *Economic Journal* 96, 808–11.

Dixit, A., Honkapohja, S. and Solow, R. 1992. Swedish economics in the 1980s. In *Economics in Sweden: An Evaluation of Swedish Research in Economics*, ed. L. Engwall. London and New York: Routledge.

Frisch, R. 1947. *Notater til økonomisk teori*, 2nd, 3rd and 4th edns. Oslo: Socialøkonomisk Institutt, University of Oslo.

Hansen, B. 1975. Introduction to Jørgen Gelting's ‘Some alternations on the Financing of Public Activity’. *History of Political Economy* 7, 32–5.

Heckscher, E. 1919. Utrikeshandelns inverkan på inkomstfördelningen. *Ekonomisk Tidskrift* 21, del II, 1–32.

Heckscher, E. 1931. *Mercantilism*. English translation London: Allen and Unwin, 1935.

Henriksson, R. 1991. Eli F. Heckscher: the economic historian as economist. In *The History of Swedish Economic Thought*, ed. B. Sandelin. London and New York: Routledge.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Howey, R. 1972. The origins of marginalism. *History of Political Economy* 4, 281–302.

Jæger, O. 1922. *Teoretisk Socialøkonomik*. Kristiania (Oslo): Centrum Avskrivnings byraa.

Jonung, L. 1987. Stockholmsskolan – vart tog den vägen? *Ekonomisk Debatt* 15, 318–26.

Jonung, L., Hedlund-Nyström, T. and Jonung, C., eds. 2001. *Att uppfostra svenska folket. Knut Wicksells opublicerade manuskript*. Stockholm: SNS Förlag.

Kærgård, N. 1984. The earliest history of econometrics: some neglected Danish contributions. *History of Political Economy* 16, 437–44.

Kærgård, N. 1995. Cooperation not opposition: marginalism and socialism in Denmark 1871–1924. In *Socialism and Marginalism in Economics 1879–1930*, ed. I. Steedman. London and New York: Routledge.

Kærgård, N. 1996. Denmark and the marginal revolution. In *Research in the History of Economic Thought and Methodology*, vol. 14, ed. W. Samuels and J. Biddle. Greenwich and London: JAI Press Inc.

Kærgård, N., Andersen, P. and Topp, N.-H. 1998. Jens Warming – an odd genius. In *European Economists of the Early 20th Century*, vol. 1, ed. W. Samuels. Cheltenham: Edward Elgar.

Kærgård, N. and Davidsen, Th. 1998. Harald Westergaard: from young pioneer to established authority. In *European Economists of the Early 20th Century*, vol. 1, ed. W. Samuels. Cheltenham: Edward Elgar.

Lönnroth, J. 1991. Before economics. In *The History of Swedish Economic Thought*, ed. B. Sandelin. London and New York: Routledge.

Lönnroth, J. 1998. Gustaf Steffen. In *European Economists of the Early 20th Century. Volume 1: Studies of Neglected Thinkers of Belgium, France, The Netherlands and Scandinavia*, general ed. W. Samuels. Cheltenham and Northampton: Edward Elgar.

Magnusson, L. 1987. Mercantilism and ‘reform’ mercantilism: the rise of economic discourse in Sweden during the eighteenth century. *History of Political Economy* 19, 415–33.

Magnussson, L. 1991. Gustav Cassel, popularizer and enigmatic Walrasian. In *The History of Swedish Economic Thought*, ed. B. Sandelin. London and New York: Routledge.

Munthe, P. 1992. *Norske økonomer. Sveip og Portretter*. Oslo: Oslo University Press.

Nielsen, A. 1948. *Det Statsvidenskabelige Studium i Danmark før 1848* [Studies in Political Economy in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Denmark before 1848]. Copenhagen: Nordisk Forlag.

Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment. *Economic Journal* 47, 53–69, 221–40.

Ølgaard, A. 1966. *Growth, Productivity and Relative Prices*. Amsterdam: North-Holland.

Oxenbøll, E. 1977. *Dansk økonomisk tænkning 1700–1770* [Danish Economic Thought 1700–1770]. Copenhagen: Copenhagen University Press.

Persson, O., Stern, P. and Gunnarsson, E. 1992. Swedish economics on the international scene. In *Economics in Sweden: An Evaluation of Swedish Research in Economics*, ed. L. Engwall. London and New York: Routledge.

Sandelin, B. 1998. Introduction: Swedish economics 1900–1960. In *Swedish Economics, Volume I: Introduction and Selected Essays*, ed. B. Sandelin. London and New York: Routledge.

Sandelin, B. 2000. Nationalekonomin i Sverige under 100 år. *Ekonomisk Debatt* 28, 59–69.

Sandelin, B. 2001. The de-Germanization of Swedish economics. *History of Political Economy* 33, 517–39.

Sandelin, B., Sarafoglou, N. and Veiderpass, A. 2000. The post-1945 development of economics and economists in Sweden. In *The Development of Economics in Western Europe since 1945*, ed. A. Coats. London and New York: Routledge.

Sæther, A. 1981. *Den økonomiske tenkning i Danmark-Norge på 1700 tallet* [Economic Thought in Denmark–Norway in the 18th Century]. Kristiansand: Agder State College.

Sæther, A. 1993. Otto Diederich Lütken – 40 years before Malthus? *Population Studies* 47, 511–17.

Schumpeter, J. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Siven, C.-H. 1985. The end of the Stockholm School. *Scandinavian Journal of Economics* 87, 577–93.

Statsøkonomisk tidsskrift 1890–1936. Kristiania (Oslo): Aschehoug og Nygaard.

Topp, N.-H. 1981. A nineteenth-century multiplier and its fate: Julius Wulff and the multiplier theory in Denmark, 1896–1932. *History of Political Economy* 13, 824–45.

Topp, N.-H. 1988. Fiscal policy in Denmark 1930–45. *European Economic Review* 32, 512–18.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Uhr, C. 1991. David Davidson: the transition to neoclassical economics. In *The History of Swedish Economic Thought*, ed. B. Sandelin. London and New York: Routledge.

Whitaker, J. 1982. A neglected classic in the theory of distribution. *Journal of Political Economics* 90, 333–55.

Wicksell, K. 1893. *Value, Capital and Rent*. English translation, London: George Allen and Unwin, 1954.

Wicksell, K. 1896. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens*. Jena: Verlag von Gustav Fischer.

Wicksell, K. 1898. *Interest and Prices: A Study of the Causes Regulating the Value of Money*. English translation, London: Macmillan, 1936.

Wicksell, K. 1901. *Lectures on Political Economy. Vol. 1, General Theory*. English translation, London: Routledge, 1934.

Wicksell, K. 1906. *Lectures on Political Economy. Vol. 2, Money*. English translation, London: Routledge, 1935.

How to cite this article

Kærgård, Niels, Bo Sandelin and Arild Sæther. "Scandinavia, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000520> doi:10.1057/9780230226203.1479

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schelling, Thomas C. (born 1921)

Timur Kuran
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Thomas Schelling has contributed path-breaking works to the study of coordination problems, group behaviour, and self-control. Early in his career, he framed the Cold War as a game in which parties have a mutual interest in coordinating their actions through a ‘focal point’. Later he explained how, in the absence of racism, racial segregation may be triggered by a ‘tipping’ process through which residential homogenization feeds on itself. His latest major insight has been that addictions stem from an inability to reconcile conflicting inner drives.

Keywords

addiction; critical mass; focal point; game theory; multiple equilibria; network models; prisoner's dilemma; residential segregation; Schelling, T. C.; self-control; subgame perfection; tipping

Article

Thomas Schelling was born in 1921 in Oakland, California. He received a BA from the University of California at Berkeley in 1944 and a Ph.D. in economics from Harvard University in 1951. Between 1948 and 1953 he worked in the Paris Marshall Plan Headquarters, in the White House, and in the Executive Office of the president, on negotiations relating to foreign aid, the European Payments Union, and NATO. He taught at Yale University (1953–8), then at Harvard University (1958–90), and finally at the University of Maryland at College Park (1990–2003). He also had a long association with the RAND Corporation, with appointments as an adjunct fellow for most of his career (1956–2002) and as a full-time researcher in 1958. He was awarded the Nobel Prize in economics in 2005.
Schelling is known for works that use the tools of economics to illustrate major social phenomena while also making foundational theoretical advances. His publications illuminate patterns and paradoxes concerning military strategy and arms control, nuclear proliferation, conflict and bargaining, coordination and conventions, tipping points and critical mass, racial segregation and integration, addiction, health policy, and business ethics. In terms of style, he has avoided the formalization that now

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
characterizes most economic research. Using mathematics sparsely, he has managed to convey intricate theoretical arguments mainly through eloquent prose and penetrating examples.

In the 1950s, when Schelling's academic career took off, policymakers around the globe were consumed by the rivalry between the two nuclear-armed superpowers, the United States and the Soviet Union. His first classic book, *The Strategy of Conflict* (1960), framed the challenge facing the two sides as coordinating on a commonly expected and mutually acceptable outcome. Avoiding nuclear confrontation required the negotiators to focus on a particular set of concessions. The underlying logic, Schelling proceeded to show, applies to a very broad set of problems in which communication is incomplete, if not impossible. People who have a mutual interest in coordinating their behaviours will look for a ‘focal point’ capable of generating a common expectation as to what is feasible.

One of his famous examples involves two strangers who are instructed to meet each other on a particular day in New York City. Unable to communicate, they look for an obvious place to meet, so obvious that each will know that it is obvious to both of them. At the time that Schelling was writing, the information desk at Grand Central Station provided just such a focal point. In this case, people achieved coordination not by speculating on what the other would do but by identifying a common course of action with the understanding that the other party was trying to do the same. Typically, the solution entailed a set of actions that stood out among its numerous alternatives. Hence, there was no uniquely ‘correct’ answer. What made an alternative ‘correct’ was simply that enough people thought so.

In a class of contexts, the required focal point lies in the structure of the game being played. For instance, in certain games with multiple equilibria it consists of a Pareto-dominant equilibrium – an outcome that no one can improve upon without harming at least one other player. Schelling's key insight, which the Grand Central Station experiment encapsulates, is that in a wide range of other contexts the focal point emerges from factors not captured by a formal representation of the game. It may depend on such factors as analogy, precedent, accidental arrangement, symmetry, aesthetic configuration, even on what the parties know about each other.

The Strategy of Conflict laid the foundations for a version of applied game theory that focuses not on zero-sum games in which players have diametrically opposed interests but on positive-sum games in which the players have both common and conflicting interests. The nuclear arms race offered the paradigmatic case: each superpower wanted to avoid touching off a mutually destructive nuclear showdown but also to dominate the other. In identifying lessons for policymakers, Schelling played a pioneering role in the development of various concepts included in the basic toolkit of modern game theory: commitment, credibility, threats, and brinkmanship. He also introduced, albeit informally, the concept now known as subgame perfection, which is a generalization of backward induction. His insight that mutual nuclear deterrence requires the threats of adversaries to be credible was an early illustration of an idea now central to thinking about strategic interactions in economics, political science, sociology, and beyond.

Another of Schelling's classic books, *Micromotives and Macrobehavior* (1978), deals with settings in which a group's aggregate behaviour is more than the sum of the behaviours of its members. What unites the members is that in acting and reacting to their environment they fail to perceive, and usually do not care, how their own choices combine with those of others to produce unintended and unanticipated consequences for the whole group. One of his influential applications of this insight concerns racial segregation. A popular explanation for racial segregation in American cities was deep-seated racism. Schelling showed that racial segregation could arise even in places where racism was not particularly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
acute. It could emerge, in fact, even in a community whose members all wanted to live in a racially mixed neighbourhood.

To see the underlying logic, consider a population whose members are either white or black. Regardless of colour, everyone prefers to live in a racially diverse neighbourhood. By the same token, each member of the population is also averse to being in the minority, though to varying degrees. Suppose now that a certain neighbourhood is 52 per cent white and 48 per cent black. Because whites are in the majority, the neighbourhood attracts disproportionately many whites. The accentuated imbalance is acceptable to the white majority. Its members do not mind if their share rises, say, to 65 per cent. But black residents who are most sensitive to being in the minority begin to move out, which reduces the proportion of blacks even further. That reduction then triggers further exits. The upshot is that the neighbourhood becomes fully white even though that was not the intention of anyone in either the majority or the minority.

In developing this analysis, Schelling familiarized the social sciences with concepts that have since gained broad applications. One of these is ‘critical mass,’ a shorthand for the minimum level of activity, often defined as a number or ratio, required to make that activity self-sustaining. If a neighbourhood will experience black flight once the black proportion falls to 40 per cent, that percentage marks the critical mass of blacks required to keep the neighbourhood integrated. A related Schelling concept is ‘tipping’, which entails a cumulative process dependent on differences in critical mass across individuals. In the presence of such differences, a behaviour can feed on itself. A few black departures can induce further departures by making the share of blacks dip below more individual thresholds, and the process can repeat itself until none remain.

Schelling's analysis points to the dangers of inferring individual characteristics from observations of collective outcomes, and of jumping to conclusions about aggregate behaviour from what one knows about individual preferences. That individually rational behaviours may generate persistent inefficiencies was already understood. The ‘Prisoner's Dilemma’ offers a case in point. Schelling's work on the tensions between ‘micromotives’ and ‘macrobehaviour’ helped to show that such tensions are much more pervasive than had been appreciated. In the case of racial sorting, people who would rather live in a more or less balanced neighbourhood than in a racially homogeneous one end up with the less preferred outcome. Moreover, once racial segregation has run its course, it is difficult to reverse, because few will voluntarily move into a neighbourhood in which they would form a tiny minority.

Identifying analogous tensions in a rich variety of other settings, Schelling demonstrated that market activity can produce lasting inefficiencies when an expectation-driven interactive process shapes the choices of participants. His examples included the custom of sending holiday cards. People feel obliged to send cards, he observed, to people from whom they expect to receive them, even when they sense that they will receive them only because the senders expect to receive cards themselves. Accordingly, two acquaintances may send each other cards for years on end, even though both find the custom burdensome and each is capable of ending the habit unilaterally. Individuals may keep sending cards to people they have not seen for decades for no other reason than the suspicion that cessation could signal something undesirable. Society as a whole would be better off, Schelling infers, with a ‘bankruptcy proceeding’ through which all holiday-card lists are obliterated to allow people to start over, motivated only by the holiday spirit, without accumulated obligations.

Schelling's insight into the possibilities of disharmony between ‘micromotives’ and ‘macrobehaviour’ has stimulated a wide range of other studies based on the observation that people fail to account for the externalities of their decisions. Refined or extended versions of his framework have been used to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
explain, among other phenomena, the emergence and disappearance of clothing fashions, the unpredictability of revolutions, the dynamics of jury decisions, the significance of momentum in political campaigns, the broadening of ethnic cleavages, norms against price competition, and economically dysfunctional behaviours inimical to development. The micro–macro interactions characteristic of Schelling's works appear also in various network models in which local interactions can trigger unintended and disproportionate global consequences.

Where *The Strategy of Conflict* and *Micromotives and Macrobehavior* address problems involving interactions among people, *Choice and Consequence* (1984), Schelling's third classic book, focuses on conflicts within individuals. Economics has traditionally treated the individual as a unified and internally consistent utility maximizer. Yet difficulties with reconciling conflicting impulses are central to the human experience. The problem of addiction, Schelling suggests, entails a failure to manage inner conflicts. It arises from inadequate self-control. The key insight of *Choice and Consequence* is that problems of self-control are commonly dampened, if not resolved fully, through tricks of the mind, social institutions, and public policies.

Many American taxpayers understate the number of their dependants to the Internal Revenue Service, the tax collection agency of the American government, to have an excessive portion of their income withheld, thus ensuring a hefty refund at the end of the year. Because excess withholdings yield no interest, this practice amounts to an expensive method of saving. If people find it convenient, the reason is that they see themselves as lacking the self-control to allow savings accumulate in a bank. By placing savings in the custody of the IRS, a person's responsible and forward-looking self gains control over the impulsive self that coexists with it.

Addictions stem from self-control difficulties, which anti-addiction treatments are designed to address. Obese people who cannot lose weight on their own check into ‘fat farms’ that regulate their diet. Likewise, alcoholics enrol in programmes that either directly limit their access to alcohol or boost their self-control through support networks. In either case, the addict effectively conspires with society to favour his responsible self over his impulsive self. In studying various self-control problems for which social remedies have been devised, Schelling finds that society does not always side with the responsible and forward-looking self. With regard to terminally ill people who are in so much pain that they want to die immediately, yet cannot bring themselves to take the final step, in most countries both the law and public morality side with the self that wants to keep living. Thus, people are prohibited from writing legally binding contracts to get assistance with dying.

The fundamental contribution of Schelling's writings on personal self-control has been in spreading awareness of a broad category of human problems that call for social intervention informed by a combination of economics and psychology. Education campaigns aimed at making people understand the consequences of their behaviours will not be effective, at least not by themselves. Most alcoholics know well that their addiction makes them unproductive, unhealthy and socially disconnected. Their troubles stem not from ignorance but from an inability to mediate conflicting demands within themselves. The period since Schelling published these insights has only raised their significance. With obesity now an acute social problem in developed and underdeveloped countries alike, and other addictions spreading as a result of rising prosperity, problems of self-control now lie at the centre of vast research programmes.

Over and beyond his penetrating and often pioneering insights into specific economic, social and political problems, Thomas Schelling has made an enduring contribution to economics and its sister

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
disciplines by showing that a deep understanding of social systems requires attention to cultural context, social dynamics, strategic interactions, and human complexities. To make sense of why a person would deliberately place an alarm clock far away from his bed, one must realize that he may be tormented by warring impulses. Although his problem can be modelled as one of utility maximization, only by taking account of the competition within him can one understand why one possible solution is likely to work better than another. Only by considering the cultural context can one understand why strangers asked to meet each other are more likely to succeed in certain cities and times than in others. No approach limited to the formal properties of payoff matrices, or the strangers' formal decision problems, will suffice. And only by understanding the mechanics of strategic decision-making, and the factors that make threats credible, can one appreciate how the massive nuclear stockpiles of the United States and the Soviet Union allowed human civilization to survive the Cold War.

See Also

- behavioural economics and game theory
- preference reversals
- residential segregation
- self-confirming equilibria
- signalling and screening
- social norms
- threshold models

Selected works

1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

1966. *Arms and Influence*. New Haven, CT: Yale University Press.

1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1, 143–86.

1978. *Micromotives and Macrobehavior*. New York: W.W. Norton.

1984. *Choice and Consequence: Perspectives of an Errant Economist*. Cambridge, MA: Harvard University Press.

Bibliography

Ayson, R. 2004. *Thomas Schelling and the Nuclear Age*. London: Frank Cass.

Swedberg, R. 1990. Thomas C. Schelling. In *Economics and Sociology*. Princeton: Princeton University

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Press.

Zeckhauser, R. 1989. Distinguished Fellow: reflections on Thomas Schelling. *Journal of Economic Perspectives* 3(2), 153–64.

How to cite this article

Kuran, Timur. "Schelling, Thomas C. (born 1921)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000528> doi:10.1057/9780230226203.1480

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schlesinger, Karl (1889 – 1938)

G. Schwödiauer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate money demand function; cash reserves; demand for money; existence of general equilibrium; mathematical economics; purchasing power parity; Schlesinger, K.; Wald, A.; Walras, L.; Zeuthen, F. L. B.

Article

Karl Schlesinger was born in Budapest; in 1919, after Béla Kun's communist revolution in Hungary, he moved to Vienna, where he committed suicide when Hitler occupied Austria in March 1938. As early as 1914 he had published his important work on monetary theory, *Theorie der Geld- und Kreditwirtschaft*, which went, however, more or less unnoticed at that time because it used mathematical tools and was written in German – a forbidding combination at a time when the only German-speaking economists interested in theory, the Austrians, were rather averse to mathematical economics. Schlesinger was also an exceptional figure in so far as he was not a university teacher but a banker and influential member of the financial community. Nevertheless, he became a respected member of the Vienna Economic Society and, in the 1930s, one of the most active participants in Karl Menger's mathematical colloquium.

As an economic theorist, Schlesinger was a Walrasian, in fact the only Walrasian (with the exception, perhaps, of Wicksell) who significantly advanced Walras's theory of the demand for money balances and of equilibrium in the money market. In his 1914 book, Schlesinger clearly distinguished between payments the magnitudes and future dates of which are fixed, and those whose time profile is subject to uncertainty. While the first type of payment streams offers no choice but generates a money ‘demand’ equal to the maximum cumulative payments deficit for a given period (though Schlesinger correctly points to the possibility of modifying the payment stream by investing and disinvesting temporary cash surpluses), the second type, which lies at the centre of Schlesinger's analysis, gives rise to a choice between higher and lower cash reserves held as an insurance against illiquidity losses. Schlesinger determines the individual demand for these precautionary balances from the equality between the marginal utilities respectively of interest income lost due to holding a cash reserve and of the insurance service provided by this reserve. He also demonstrated the economies of scale from an increase in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
number (but not in the nominal magnitudes) of transactions. Finally, Schlesinger derives an aggregate money demand function, additively separable in transactions and precautionary demand, virtually identical to the one set up by Keynes much later, and determines the partial equilibrium money rate of interest as that which equalizes aggregate demand and stock supply of money. Schlesinger also addressed himself to problems of international monetary economics: in a publication of 1916 he advocated and gave a clear exposition of the purchasing power parity theory. In 1931, in the context of a book review, Schlesinger developed a rigorous and detailed mathematical analysis of money creation on the level of individual banks and for the financial system as a whole.

Apart from his writings on money, Schlesinger made another original and remarkable contribution to economic theory, viz. to the mathematical theory of Walrasian general economic equilibrium described by *n* zero-profit conditions (equating commodity prices which are given functions of quantities produced with the respective sums of products of factor input coefficients and factor prices) and *m* factor market equilibrium conditions (equating factor supplies with the respective sums of products of factor input coefficients and quantities of goods produced). At the beginning of the 1930s it came to be recognized that such a system of equations need not have a solution, at least not an economically meaningful solution (in non-negative output and factor prices). In 1934 Schlesinger suggested, in Menger's colloquium, to introduce non-negative slack variables on the demand side of the factor market equations and to enlarge the system of $n+m$ equations by additional m equations setting the respective products of slack variables and factor prices equal to zero. Schlesinger had arrived at this ingenious idea independently of an identical proposal made by Zeuthen in 1932. Going definitely beyond Zeuthen, however, he also raised the conjecture that this procedure would solve the existence problem. Schlesinger's conjecture was proved to hold true by A. Wald, with whom Schlesinger had taken lessons in mathematics.

Selected works

1914. *Theorie der Geld- und Kreditwirtschaft*. Munich: Duncker & Humblot. Partial English translation as ‘Basic principles of the money economy’, *International Economic Papers* 9 (1959), 20–38.

1916. *Die Veränderungen des Geldwertes im Kriege*. Vienna: Manz.

1931. Bankpolitik, von F. Somary [book review]. *Archiv für Sozialwissenschaft und Sozialpolitik* 66, 1–35.

1934. Über die Produktionsgleichungen der ökonomischen Wertlehre. *Ergebnisse eines mathematischen Kolloquiums* 6, 10–11.

Bibliography

Menger, K. 1973. Austrian marginalism and mathematical economics. In *Carl Menger and the Austrian School of Economics*, ed. J.R. Hicks and W. Weber. Oxford: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Morgenstern, O. 1968. Schlesinger, Karl. In *International Encyclopedia of the Social Sciences*, vol. 14. New York: Macmillan and Free Press.

Nagatani, K. 1978. *Monetary Theory*. Amsterdam: North-Holland.

Patinkin, D. 1965. *Money, Interest and Prices*. Supplementary Note D, 573–80. 2nd edn. New York: Harper & Row.

How to cite this article

Schwödiauer, G. "Schlesinger, Karl (1889–1938)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000034> doi:10.1057/9780230226203.1481

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schmoller, Gustav von (1838 – 1917)

B. Schefold

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

German Historical School; mercantilism; Methodenstreit; Schmoller, G. von

Article

Schmoller was born in Heilbronn, the son of a Württemberg civil servant. He studied Staatswissenschaften (a combination of economics, history and administrative science) in Tübingen. After a short period in the financial department of the Württemberg civil service, which he had to quit because of his pro-Prussian views, be became Professor in Halle (1864–72), Strassburg (1872–82), and Berlin (1882–1913).

Schmoller was the leading economist of Imperial Germany. He was the leader of the ‘Kathedersozialisten’ (socialists of the chair), and founder and long-time chairman of the Verein für Socialpolitik. He was editor or co-editor of several publications such as *Staats- und sozialwissenschaftliche Forschung* and *Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft im Deutschen Reich* – later known simply as *Schmollers Jahrbuch;* he was named official historian of Brandenburg and Prussia, and supervised the publication of the *Acta Borussica* and the *Forschungen zur brandenburgischen und preussischen Geschichte*. Thus Schmoller was one of the major organizers of research in the social sciences. He is said to have controlled almost every important academic appointment in economics in the German Reich.

As the outspoken leader of the ‘younger’ Historical School, Schmoller was against the abstract axiomatic–deductive approach of the classicals and neoclassicals (1893; 1900, pp. 1–124). When Menger, the Austrian marginal utility theorist, attacked Schmoller's point of view and asserted the necessity of applying the exact methods of natural sciences and abstract logical reasoning to political economy, the Methodenstreit (struggle over methods) began, which was by and large a dispute between the inductive and the deductive method. It occupied two generations of German-speaking economists, produced a vast literature and was perceived essentially as ‘a history of wasted energies’ (Schumpeter) by theoretical economists of the next generation. However, it may also be viewed as the expression of the endeavour to preserve seminal insights into the historical and changing nature of economic and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
social phenomena against simplified and mechanistic views of the laws of ‘rational’ behaviour, and as such it had important consequences for the development of neighbouring disciplines, especially sociology.

Although Schmoller put the emphasis on the inductive method, he was not excluding deduction from economic reasoning. In his opinion, it was of the utmost importance for the application of deductive methods and for economic theory formation in general to be based on the knowledge of sufficient historical facts and material. He advocated a somewhat interdisciplinary approach that would also take into account the psychological, sociological and philosophical aspects of the problems. Through detailed and monographic historical research he intended to free political economy from ‘false abstractions’ (1904, p. vi), and to put it on a solid empirical foundation. His most important historical studies were his works on the history of the weavers guild of Strassburg (1879), on the guilds in 17th- and 18th-century Brandenburg and Prussia, on the Prussian silk industry in the 18th century, on the history of Prussian financial policy (1898a) and on the history of German towns in general and Strassburg in particular (1922). He was also interested in the history and formation of social classes and the historical development of class struggle (1900, Book II; 1904, pp. 496–577; 1918). He further made some important contributions to the study of mercantilism (1898a, ch. 1; 1904, pp. 580–605), which he regarded essentially as the process of the formation of the national state and the national economy. The adoption of mercantilist policies was of special significance for Germany, whose backwardness in the 17th and 18th century Schmoller ascribed to the absence of a centralized national state and the consequent domination of particularist regional and local interests. Schmoller perceived the enlightened and despotic sovereigns – especially the Prussian kings – as the only power that would implement a policy aimed at the breaking-up of these particularist tendencies and at the establishment of large and unified economic territories. An important step in that direction was the abolition of town autonomy after 1713 under Friedrich Wilhelm I (1922, pp. 231–428).

This glorification of the Prussian state and its rulers was probably the most characteristic feature of Schmoller's work. He regarded the Prussian monarchy with its corps of loyal civil servants, which he perceived as standing above the social classes and their egoistic interests, as the central achievement of German history. Only this type of government had been able to overcome the earlier feudal corporate state and the class rule of the Junker (1898a, p. 302), and was at present capable of implementing social reforms.

Social reform and social justice were central to Schmoller's thinking. We may regard him as a conservative in the specific German or, better, Prussian sense of the word. He rejected Marxism, Manchester Liberalism and also reactionary, anti-reformist views such as those of the historian Heinrich von Treitschke, with whom he had a famous polemic on the notion of social reform (1874–5; Small, 1924–5).

Schmoller advocated a paternalistic social policy to raise the material and cultural standard of the working classes as the only means to prevent revolution, integrate the workers into the monarchic state, and keep the traditions of Prussia alive. He even envisaged an alliance between the monarchy and the working classes (1918, p. 648).

It is nowadays generally agreed that Schmoller's influence on the development of the economic sciences in Germany was rather unfortunate: it contributed to the neglect of economic theory in Germany for a full half century. Neither Schmoller nor his pupils achieved their goal of building a new theory based on the historical material they collected, however valuable it was. Schmoller's main work, the *Grundrisse*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(1900; 1904) remained rather traditional in its theoretical part – the treatment of value and prices was not too far away from mainstream neoclassical economics – and constituted all in all a rather incoherent analysis. Perhaps this was the main reason why Schmoller's work and with it the whole Historical School was to fall into oblivion in Germany soon after his death.

See Also

- Historical School, German

Selected works

1870. *Zur Geschichte der deutschen Kleingewerbe im 19. Jahrhundert: Statistische und nationalökonomische Untersuchungen*. Halle: Waisenhaus.

1879. *Die Strassburger Tucher- und Weberzunft; Urkunden und Darstellungen nebst Regesten und Glossar. Ein Beitrag zur Geschichte der deutschen Weberei und des deutschen Gewerberechts vom XIII.–XVII. Jahrhundert*. Strassburg: Trübner.

1888. *Zur Litteraturgeschichte der Staats- und Sozialwissenschaften*. Leipzig: Duncker & Humblot.

1898a. *Umrisse und Untersuchungen zur Verfassungs-, Verwaltungs-und Wirtschaftsgeschichte besonders des preussischen Staates im 17. und 18. Jahrhundert*. Leipzig: Duncker & Humblot.

1898b. *Über einige Grundfragen der Sozialpolitik und der Volkswirtschaftslehre*. Leipzig: Duncker & Humblot. It contains the works *Über einige Grundfragen des Rechts und der Volkswirtschaft* (1874–5), *Die Volkswirtschaft, die Volkswirtschaftslehre und ihre Methode* (1893), and *Wechselnde Theorien und feststehende Wahrheiten im Gebiete der Staats- und Sozialwissenschaften und die heutige deutsche Volkswirtschaftslehre* (1897).

1900–1904. *Grundriss der allgemeinen Volkswirtschaftslehre*. Vol. 1, 1900; vol. 2, 1904. Munich and Leipzig: Duncker & Humblot.

1918. *Die soziale Frage:* In *Klassenbildung, Arbeiterfrage, Klassenkampf*, ed. L. Schmoller. Munich und Leipzig: Duncker & Humblot.

1922. *Deutsches Städtewesen in älterer Zeit*. Bonner Staatswissenschaftliche Untersuchungen, Bonn and Leipzig: Schröder.

Bibliography

Brinkmann, C. 1937. *Gustav Schmoller und die Volkswirtschaftslehre*. Stuttgart: Kohlhammer.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Small, A.W. 1924–5. The Schmoller–Treitschke controversy. *American Journal of Sociology* 30, 49–86.

How to cite this article

Schefold, B. "Schmoller, Gustav von (1838–1917)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000035> doi:10.1057/9780230226203.1482

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schmookler, Jacob (1918 – 1967)

N. Rosenberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic growth; inventions; patents; Schmookler, J.; technical change

Article

Schmookler was born in Woodstown, New Jersey, in 1918 and died in Minneapolis, Minnesota, in 1967. In 1951 he received his Ph.D. in economics from the University of Pennsylvania, where he was a student of Simon Kuznets. He subsequently held teaching positions at Michigan State University and the University of Minnesota.

Schmookler's work helped to establish the importance of technological change as a contributor to economic growth. His article ‘The Changing Efficiency of the American Economy, 1869 to 1938’ appeared in 1952 (Schmookler, 1972, pp. 3–36), several years before the seminal papers of Abramovitz and Solow. However, in his later work he also analysed the specific economic mechanisms that determined the allocation of resources among different categories of invention. By an extremely careful and original use of patent data, Schmookler demonstrated the decisive role played by changes in demand in shaping the pattern of inventive activity. In his most important work, *Invention and Economic Growth* (1966), he showed how changes in demand have accounted for variations in inventive activity in a specific industry (such as railroad equipment, petroleum refining and building) over time, as well as different rates of inventive activity among different industries at a given moment of time.

Schmookler's writing showed that technological change need not be treated as an exogenous variable. On the contrary, he showed that the changing direction of inventive activity could be accounted for by readily identifiable economic variables, most especially changes in the pattern of demand that determine the size of the prospective market, and hence potential profitability, for particular classes of inventions. Thus, inventions are significant not only because they influence the growth rate of an economy but also because they constitute forms of economic activity in their own right.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1966. *Invention and Economic Growth*. Cambridge, MA: Harvard University Press.

1972. *Patents, Invention and Economic Change*, ed. Z. Griliches and L. Hurwicz. Cambridge, MA: Harvard University Press. (This volume contains the patent data which formed the basis for Schmookler's research. It also contains a complete bibliography of his writings.)

How to cite this article

Rosenberg, N. "Schmookler, Jacob (1918–1967)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000036> doi:10.1057/9780230226203.1483

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

scholastic economics

Odd Langholm
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

For approximately four centuries prior to the Reformation and the break-up of Christianity, major advances in economic ethics and analysis were made by representatives of medieval scholasticism. Theologians and philosophers, aided by canonists and Romanists, the scholastics aimed at moral instruction based on an understanding of economic phenomena and relationships. Condemning avarice, fraud and economic coercion, they dealt with trade and price, money and usury, labour and wages. They thus forged a link between ancient and early modern economic thought.

Keywords

Albert the Great; Aquinas, St. T.; Aristotle; barter; bullion; collusion; exchange; fraud; just price; metallist theory of money; monopoly; price discrimination; Salamanca School; scholastic economics; specie; truck; usury; utility

Article

The tradition in economic thought known as scholastic economics was part of the system of scholastic moral theology and philosophy taught in the schools and universities of western Europe in the Middle Ages. Its purpose was the construction of viable norms of commercial behaviour in a Christian world. The literary authorities on which it was based include the Bible and the writings of the Church fathers, and papal and conciliar decrees from the time of the early Church to the Carolingian period. In the 12th century the theologian Peter Lombard in his *Sentences* and the canonist Gratian in his *Decretum* codified much of this material along with later decrees. In the 13th century Gratian's work was augmented with the *Decretals* of Pope Gregory IX to form the *Corpus Juris Canonici*. The major names in scholastic economics were theologians writing with a leaning towards canon law. A majority of them were mendicant friars. The canonists observed a distinction between the norms that apply in the internal forum of conscience and the more liberal ones that found expression in the external forum of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
ecclesiastical courts. The latter were partly adapted from Roman law expressed in the *Code* and *Digest* of the *Corpus Juris Civilis*. The translations of Aristotle's *Ethics* and *Politics* gave access to some other bits of ancient texts that invited economic reasoning. The literary vehicles of scholastic economics were commentaries on the Bible and the *Sentences* and on Aristotle's works, theological *summae*, records of academic disputations, sermons, handbooks for confessors, as well as a few treatises dealing specifically with some given economic subject.

The schoolmen were primarily concerned with external economics, that is, with buying and selling and other forms of exchange of goods and services. This is so because of the ethical problem encountered in the exchange situation. The key word is consent. Mutual consent is required for an economic contract to be valid, but traders may be tempted by avarice to obtain terms of exchange to which the other party does not truly consent. In the case of buying and selling, the schoolmen envisaged a positive and a negative approach to consent. Because no one suffers injustice voluntarily, the former approach consisted in estimating the just price. The latter consisted in identifying factors incompatible with consent. A just price could be estimated on the basis of cost or market factors, the main thesis being that a competitive market price is a just price and that cost must adapt to the market. According to Aristotle, justice requires equivalence between goods given and received in exchange or between goods and money. It used to be argued that this is meaningless because there would be no motivation for anyone to exchange if what he got was of equal value to what he gave. This objection rests on a confusion of exchange value with utility, a confusion of which the schoolmen were not guilty. In their discussions of justice, equality referred to value in exchange, though not to an exact point; in the words of Thomas Aquinas (d. 1274), the just price was rather an estimate that was not completely precise. According to John Duns Scotus (d. 1308) it permitted of a certain latitude. Scotus also noted that the parties to a sale would often disagree about the just price. Both might then yield a little in order to reach an agreement, and, if they were then satisfied, an element of gift was involved in the contract. Consent, and hence justice, was still preserved.

Consent and justice are precluded in cases of fraud and coercion. In reply to a Roman law maxim stating that a thing is simply worth the amount at which it can be sold, a number of schoolmen from William of Auxerre (d. 1231) and Roland of Cremona (d. 1259) protested that this is not true if the thing can be sold at that amount only by means of fraud. Fraud is wilful misrepresentation of the object or conditions of exchange. The buyer must know what he buys. But knowledge also presumes a mental capacity to receive and digest information. Peter Olivi (d. 1298) states that terms of exchange are invalid if their acceptance by the buyer is due to his mental backwardness, ignorance or inexperience. A confusion regarding fraud arose because Roman law permitted the parties to a sale to outwit and ‘deceive’ one another in the course of the haggling and bargaining about price. The law accepted agreements thus obtained unless the result of the bargaining involved a deviation of more than one half of the just price. The theologians were inclined to reject this one-half idea and insisted on a just price within rather narrower limits. In principle, ‘deception’ as an element of the bargaining process must be distinguished from deliberate and blatant fraud, such as telling lies about the nature and quality of merchandise, hiding defects, using false weights and measures, soaking wool or diluting wine, and the like, in order to obtain an unjust price. No law condones this. Thomas Aquinas's catalogue of fraudulent practices in the *Summa Theologiae* is representative of the schoolmen's position.

Besides fraud and deception, consent is incompatible with practices considered by the schoolmen to be coercive. Monopoly is mentioned in Aristotle's *Politics* but merely in a context that did not invite

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
disapproval. Monopoly and collusion among sellers to set minimum prices are prohibited in the *Code* of Justinian. The issue was brought into the commentary tradition on the *Decretals* by the canonist Hostiensis (d. 1271) and through him to a number of authors of works for the internal forum. Among the more important ones, mention may be made of John of Freiburg (d. 1314), Astesanus of Asti (d. 1330), Raniero of Pisa (d. c. 1348), Antonino of Florence (d. 1459), and Angelo Carletti (d. 1495). According to these thinkers, merchants agree among themselves that only one of them shall trade in a certain necessary commodity or that all shall sell at the same excessive price. A related tradition can be traced back to Alexander of Hales (d. 1240). Alexander, in his *Summa Theologica* (1951–7, III, 490, p. 724) berates those who ‘take over the whole marketplace of commodities’ in order to raise prices. Later authors, from Astesanus to Angelo Carletti, paraphrase Alexander and rebuke those who disturb the run of the market by such means. John Duns Scotus has a name for them: they are ‘regraters’. Angelo vividly describes the merchants standing at the city gates buying up all the new grain, preventing its reaching the marketplace to be sold there. Battista Trovamala (d. c. 1495) explains how the price of victuals can be raised by impeding or detaining ships bringing home such goods or by giving earnest money for the purchase of all the spice in the city in order to sell it later at a profit.

From Carolingian price regulation, transmitted via canon law, the schoolmen received two ordinances known by their incipits (opening words) as *Placuit* and *Quicumque*. *Placuit* forbids price discrimination against travellers in local markets. Such people should pay the market price or (in an alternative version, which comes to much the same thing) pay what the residents pay. This injunction was cited by John of Freiburg and Astesanus and numerous followers in the Dominican and Franciscan penitential traditions. The great preacher Bernardino of Siena (d. 1444) in his vernacular sermons in the cities of Tuscany repeatedly impressed on his crowds of listeners the simple message that all should pay the same price. It was argued that residents ganging up against foreigners is not morally very different from merchants conspiring against local customers.

Quicumque lacks this conspiracy element. Originally a capitulary of Charlemagne included in the *Decretum*, it censures the malpractice of buying up hoards of victuals or wine cheaply at the time of harvest or vintage in order later to sell them at a much higher price (practices known as ‘engrossing’ or ‘forestalling’). Large profits on such operations are possible because ‘scarcity is created’– literally: ‘dearth is induced’ (*caristia inducatur*). This phrase originated in canon law literature; its earliest known appearance is in a gloss by Laurence of Spain (d. 1248). It proved to possess a remarkable literary appeal. Besides later canonists, it was copied by Albert the Great (d. 1280), Ulrich of Strasbourg (d. 1277), Thomas Aquinas and Antonino of Florence in theological works, by John of Freiburg, Bartolomeo of San Concordio (d. 1347), Durand of Champagne (d. c. 1350), Angelo Carletti and Battista Trovamala in their penitential summas, and by Henry of Langenstein (d. 1397), Matthew of Cracow (d. 1410), and John Gerson (d. 1429) in their treatises on economic contracts, as well as by numerous other schoolmen. Many of these authors state that buyers faced with ‘created scarcity’ are forced to pay excessive prices. A person paying more than usual because of scarcity due to crop failure or other natural misadventures can of course also be said to be forced to do so. Unlike much of modern economics, however, scholastic economics maintained a distinction between personal and impersonal compelling agents, because the former could be held morally responsible for their acts.

The reason why the schoolmen came down so hard on creating scarcity is that scarcity is one of the basic phenomena of all economic life. Economic activity is, in a word, a common struggle against scarcity. Using a terminology adopted from Peter Olivi by Bernardino of Siena, value rests on *raritas* (scarcity),

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
virtuositas (usefulness, in an objective sense), and *complacibilitas* (desirability, pleasingness). These factors vary from person to person, from time to time, and from place to place, and determine the outcome of all voluntary transactions. In a beautiful piece of analysis, Richard of Middleton (d. 1307) shows that, just as a profit can be made in long-distance trade because just prices differ with plenty and scarcity in different locations, thus, also, both parties can profit justly in local exchange because they value what they receive more highly than what they give – an argument that readily can be restated in Olivi's terms. The idea that both parties to an exchange are better off because they would not otherwise have exchanged is not a major insight recently discovered. Moreover, it is a truism. The schoolmen's point is that both parties profit by *just* exchange voluntarily agreed upon in the absence of fraud and duress.

The schoolmen held no comprehensive theory of wages. To some extent, arguments about the just price could be applied by analogy to the just wage. Possession of qualified labour skills and professional competence, when these are in demand, also represents economic power. The recognition of this fact by the *Ethics* commentator Gerald Odonis (d. 1349) and some of those who copied his cases should not be permitted to obscure the fact that power in medieval labour relations was heavily concentrated on the part of the employer rather than on the part of the employee. Thomas Aquinas emphasized that manual labourers should be paid a decent wage and that it should be paid promptly because they lived from hand to mouth. This admonition was repeated by numerous authors of penitential manuals, including some well-known names like Henry of Langenstein, John Gerson and Antonino of Florence. Because the wage was in large part fixed by custom or by institutions like the guilds, their emphasis was on paying promptly rather than on paying justly. Antonino's main contribution in this area was his attack on abuse of the truck system, whereby workers were paid in goods rather than in money, goods that they had to sell at a loss. Battista Trovamala notes that a double fraud was sometimes involved in truck practices. Merchants who barter goods rather than buy and sell them may do so justly but ostensibly at more than the just price. If a merchant has workers in his employ, he might claim that the goods paid them in lieu of money were bought, rather than bartered, at these inflated prices.

Scholastic monetary theory was largely Aristotelian. Scholars do not agree on whether Aristotle held a metallist theory of money. The schoolmen were definitely metallists. Money is a medium of exchange, a measure of value, and a store of wealth. In order to serve well in these capacities, certain properties are required. A number of commentaries on the *Ethics* and the *Politics* contain lists of such properties. That by Henry of Friemar (d. 1340) has five items: small size, to prevent subtraction (clipping and so forth); a stamp of the sovereign to prevent falsification; a determined weight so that prices can be fixed exactly; durability to ensure uncorrupted future validity; a content of precious material, like gold and silver, for easy and prompt valuation. John Buridan (d. c. 1360) adds a sixth requirement: coins of small denomination for the petty purchases of the poor. In exchange between merchants of different regions, money-changing (*campsoria*) was necessary. Henry of Ghent (d. 1293) points out that the changer may grant himself a fee for his labour spent counting and guarding money. It is true that money is non-vendible, says Giles of Lessines (d. 1304), but money-changing is not properly a sale but a form of barter. Guido Terreni (d. 1342) makes a different distinction. Sometimes two kinds of money are exchanged in a place where both are current, one being domestic currency and the other foreign currency permitted at a certain rate of exchange. If this rate underestimates the foreign currency, it may be bought and sold by weight. If exchange is made in a place where neither or only one kind of money is current, non-current money becomes a commodity. It can be bought or sold by weight at a just price and taken to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
a location where it is current and worth more.
Whereas tampering with coins for a quick profit might be tempting for the individual citizen, debasement of the currency in order to replenish the treasury proved irresistible to medieval monarchs. In early 14th-century France debasement had reached ruinous proportions. Some schoolmen spoke out against it. Drawing on Guido Terreni and others, John Buridan argues as follows. The primary measure of goods in exchange is human need. It is by that yardstick that money is a measure as well. Granted that we do not often need gold and silver, wealthy people desire them for display, and so we find that they are worth the same, or nearly the same, in specie as in bullion. Only if money is thus measured in proportion to need can goods be measured in proportion to money. For whatever proportion they have to need, the same proportion they will have to money proportioned to need. If a certain coin is in circulation and the king issues another coin, it is true that he can stipulate its value in relation to the former, for instance declaring that one new penny is to be given in exchange for three of the old. If it is not worth that much, however, or very near it, according to the relation of its material to human need, the king sins and profits unjustly at the cost of the common people. Much of the credit for the scholastic theory of money has been granted to Nicole Oresme (d. 1382), whose merit is a detailed and knowledgeable description of the various ways in which the currency can be debased and of the social consequences of debasement.
The gravest specific economic sin was usury, unjust profit on a loan, usually a loan of money. Usury could also occur in credit sales or other sales involving time. Such arrangements can be interpreted as combinations of a cash sale and a loan. The schoolmen considered usury to be condemned in Luke 6:35: ‘Lend, hoping for nothing again.’ In addition, several arguments from natural reason were put forward. According to Peter the Chanter (d. 1197), the usurer profits merely because time passes, but time belongs to God. Thomas Aquinas argued that money is consumed in use by being spent, just as natural fungibles like grain and wine are consumed when we eat and drink them; money, therefore, has no use separate from its substance, and to charge a profit for lending it is to make the borrower pay twice for the same thing or to pay for something that does not exist. Gerardo of Siena (d. 1336) tied in with Aquinas by declaring that money is sterile, not because it is ‘barren metal’ (an idea wrongly attributed to Aristotle) but because it is a fungible. Contrary to fruit-bearing things, all fungibles are barren because they are always of the same value in terms of themselves. This intriguing argument has proved difficult to refute. An argument from coercion was originally proposed by William of Auxerre. The borrower pays usury because he is forced by need to do so. Adopting a broad definition of need, Thomas accepted this argument and proceeded to demonstrate the analogy between exploiting a needy borrower and exploiting a needy buyer, thus combining the two main areas of scholastic economics by a common formula. Whereas usury was universally condemned, the schoolmen recognized the lender's right to compensation for all damage incurred owing to default in repayment. Some extrinsic titles to interest within the period of the loan were also hesitantly recognized. After all, in the words of Gerald Odonis, two persons cannot use the same money at the same time. The full development of extrinsic titles was mainly the work of later canonists.
The decline of scholastic economics is a complex process impossible to date. In one sense part of it is still present in the European moral consciousness. In another sense, of a continuously developing system, it did not last much longer than to the latest authors cited in this article, that is, to the end of the 15th century. The so-called ‘second scholasticism’ associated with the Counter-Reformation School of Salamanca broke with medieval scholasticism to an extent that makes treatment under the same heading

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
unadvisable. The change is well illustrated by the shift of focus from the attack on usury to the defence of titles to interest. In a larger perspective this and other changes are expressions of the fading-out of the natural law philosophy of Aquinas in favour of the natural rights philosophy on which most of modern economic thought is based.

See Also

- Albert the Great, Saint Albertus Magnus
- Aquinas, St Thomas
- just price

Bibliography

Alexander of Hales. 1951–7. *Summa Theologica*, 4 vols., Quaracchi: Collegii S. Bonaventurae.

Aquinas, St T. 1961–65. *Summa Theologiae*. 5 vols., Madrid: Biblioteca de autores Cristianos.

Corpus Iuris Canonici. 1879. 2 vols. Leipzig: Taucnitz.

Corpus Iuris Civilis Iustiniani cum commenttariis. 1627. 8 vols. Osnabrück: Zeller, 1965–6.

de Roover, R. 1967. *San Bernardino of Siena and Sant'Antonino of Florence: The Two Great Economic Thinkers of the Middle Ages*. Boston: Baker Library, Harvard Graduate School of Business Administration.

Friedberger, W. 1967. *Der Reichtumserwerb im Urteil des Hl. Thomas von Aquin und der Theologen im Zeitalter des Frühkapitalismus*. Passau: Passavia.

Gilchrist, J. 1969. *The Church and Economic Activity in the Middle Ages*. London: Macmillan.

Gordon, B. 1975. *Economic Analysis before Adam Smith*. London: Macmillan.

Langholm, O. 1983. *Wealth and Money in the Aristotelian Tradition*. Oslo: Universitetsforlaget.

Langholm, O. 1992. *Economics in the Medieval Schools*. Leiden: Brill.

Langholm, O. 1998. *The Legacy of Scholasticism in Economic Thought*. Cambridge: Cambridge University Press.

Langholm, O. 2003. *The Merchant in the Confessional*. Leiden: Brill.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Little, L.K. 1978. *Religious Poverty and the Profit Economy in Medieval Europe*. Ithaca: Cornell University Press.

Noonan, Jr. J.T., 1957. *The Scholastic Analysis of Usury*. Cambridge, MA: Harvard University Press.

Peter, Lombard. 1971, 1981. *Sententiae*, 2 vols., Rome: Collegii S. Bonaventurae.

Todeschini, G. 1994. *Il Prezzo della Salvezza: Lessici medievali del pensiero economico*. Rome: Studi superiori NIS.

Viner, J. 1983. Religious thought and economic society. *History of Political Economy* 10, 9–189.

Wood, D. 2002. *Medieval Economic Thought*. Cambridge: Cambridge University Press.

How to cite this article

Langholm, Odd. "scholastic economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000489> doi:10.1057/9780230226203.1484

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Scholes, Myron (born 1941)

Toni M. Whited
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Myron Scholes is best known for his contribution to the derivation of the widely used Black–Scholes option pricing formula. His contributions to financial economics are not, however, limited to his research on option pricing. His other notable work includes seminal empirical tests of the capital asset pricing model, both theoretical and empirical examinations of the effects of dividend taxation on stock returns, and studies of tax issues relating to the cost of capital, capital structure, real estate values, the optimal liquidation of assets, and employee compensation.

Keywords

Black, F.; Black–Scholes formula; capital asset pricing model; capital gains taxation; cost of capital; derivatives; dividend policy; employee stock ownership; human capital; liquidity premium; Merton, R.; option pricing; options; Scholes, M.; stock returns; tax clienteles; taxation of capital income

Article

Best known for his Nobel prize-winning work on derivatives pricing, Myron Scholes made significant contributions to a wide range of topics in financial economics, from asset pricing to dividend policy and tax incentives. Born 1 July 1941 in Timmins, Ontario, Canada, he earned a Bachelor's degree in Economics from McMaster University in Hamilton, Ontario in 1962. He then entered the MBA course at the University of Chicago, transferring to the Ph.D. course after his second year. He earned his MBA in 1964 and his Ph.D. in 1968, writing his dissertation on the effects of information and signalling on the shape of the demand curves for traded securities. Upon finishing graduate studies he took a position as Assistant Professor of Finance at the Sloan School of Management at the Massachusetts Institute of Technology. After five years he moved to the Graduate School of Business at the University of Chicago. He was first Visiting Professor, and then in 1974 he accepted a permanent position. He stayed there until 1981, when he became Visiting Professor at Stanford University for two years, becoming a permanent

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
faculty member of the university's Business School and Law Schools in 1983, remaining there until his retirement in 1996. He is currently the Frank E. Buck Emeritus Professor of Finance at Stanford, as well as a Managing Partner of Oak Hill Capital Management and the Chairman of Oak Hill Platinum Partners. He serves on the board of directors of several corporations.

It was during his time at MIT that Scholes worked with Fischer Black and Robert Merton to develop the Black–Scholes formula for option pricing. An option is the right but not the obligation to buy or sell a security on or before a certain date at a certain price. Options are members of a family of securities known as derivatives, because their value is derived from the value of another security. Scholes and Merton earned a Nobel Prize in economics in 1996 for ‘a new method to determine the value of derivatives’. Almost every MBA and many undergraduates across the world learn the famous Black–Scholes formula, which is widely used by derivatives traders; indeed, traders often quote option ‘prices’ not in dollars and cents but in terms of one of the inputs into the Black–Scholes formula – the volatility of the return on the stock on which the option is written, or ‘vol’. The formula has had ramifications far beyond the area of derivatives, offering new ways of thinking about related areas, such as corporate finance, capital budgeting, and financial markets and institutions.

Black (1989) and Bernstein (1992) provide detailed histories of the development of the formula, which can be summarized as follows. Between 1968 and 1969 Fischer Black derived a partial differential equation that described the value of an option as a function of the maturity of the option and the underlying stock price. He did so by applying the capital asset pricing model (CAPM) to options, rather than stocks, at each instant in time. Unable to solve the equation, Black teamed up with Scholes in 1969. Scholes's contribution to the formula was the logic used to solve the equation – logic that is still taught to students studying finance. Black and Scholes noted that the differential equation did not involve the expected return on the stock and that, therefore, any expected stock return is consistent with the option price, including the risk-free rate of interest. They then turned to a much earlier work, Sprenkle (1961), which contains a formula for the expected future value of an option as a function of the stock return. Using the risk-free rate as the stock return in this formula gave them the expected future value of the option. Using the risk-free rate to discount this future value back to its present value, Black and Scholes came up with what is now known as the Black–Scholes formula, which did indeed solve Black's differential equation. The published version of this work appeared in the *Journal of Political Economy* in 1973. It did not, however, contain Black's derivation of the formula, but one suggested by Robert Merton, who pointed out that it is possible to replicate option payoffs with those from a portfolio of stocks and bonds. Hedging the option with this portfolio produces a risk-free position, and equating the return on this risk-free position with the risk-free rate of interest produces Black's differential equation.

Scholes's contributions to financial economics are not limited to his research on option pricing. For example, Black, Jensen and Scholes (1972) performs one of the earliest and most prominent tests of the CAPM, finding that the systematic risk of a stock explains less of the stock's return than the amount predicted by the CAPM. In related work, Scholes and Williams (1977) devise a method for testing asset pricing models with high frequency data, even when stocks are traded asynchronously. This research was an outgrowth of Scholes's work to develop a database on daily stock prices for the Center for Research in Security Prices at the University of Chicago. Another notable paper is Bulow and Scholes (1983), which explores seeming anomalous features of pension plans such as early retirement benefits or ad hoc increases in benefits. These features allow employees to have claims, in excess of vested benefits,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
on the assets of defined-benefit pension plans. They explain this anomaly as a product of employee human capital, which ends up being compensated via the pension plan in the optimal contract with stockholders.

Although important, these papers were diversions from Scholes's main line of research: the effects of taxation on financial decisions and financial markets. His most prominent work in this area examines the effects of dividend taxation on security prices. Black and Scholes (1974) was the lead article in the first issue of the *Journal of Financial Economics*. The paper tests the idea in Modigliani and Miller (1961) that changes in dividend policy can affect firm value if capital gains are taxed at a lower rate than dividends, because the firm can easily increase its value by cutting dividends and instead distribute cash to shareholders via stock repurchases. Black and Scholes found no evidence that dividend policy affects returns, despite the existence of differential taxation. Their explanation of the result appealed to the idea of tax clienteles. Firms adjust the supply of shares at any particular dividend yield to meet the demands of different groups of investors with different tax preferences. Market equilibrium then implies that no firm can increase its value by changing its dividend policy. Perhaps motivated by these results, Miller and Scholes (1978) present sufficient conditions for dividend policy to be irrelevant, even in the presence of taxes. Miller and Scholes (1982a) again test the relationship between returns and dividends, taking care to purge any reaction of returns to dividend of signalling or information effects. In contrast to Black and Scholes (1974), they find significant tax effects.

Scholes's interest in taxes extended beyond dividend policy. He also worked on tax issues relating to the cost of capital, capital structure, property values and the optimal liquidation of assets. His research with Mark Wolfson examined corporate tax planning under uncertainty and information asymmetry. Much of this work is summarized in their 1992 book, *Taxes and Business Strategy: A Planning Approach*.

Scholes's interest in taxation also intersected with his interest in the use of stock to compensate employees. For example, Miller and Scholes (1982b) argue that deferred compensation programmes structured to optimize employee incentives are observationally equivalent to those structured to optimize taxes. In his 1990 presidential address to the American Finance Association (Scholes, 1991), he outlined the interplay between taxes and incentives in deferring employee compensation. When an employee accepts deferred compensation, he essentially agrees to allow the firm to invest in its own stock on his account. The tax advantage to deferred compensation therefore depends on the after-tax rates of return that can be earned by the corporation compared to the return an employee could earn investing on his own. The employee stock ownership implicit in deferred compensation, however, aligns incentives of employees with those of shareholders. Optimal compensation schemes trade off these benefits with any possible tax costs.

In the latter part of his career Scholes once again became interested in derivatives, this time from a more practical and institutional angle. He became the managing director and co-head of the fixed-income derivatives group at Salomon Brothers. In 1994 he and Robert Merton joined several colleagues from Salomon to start a firm called Long-Term Capital Management (LTCM). Their basic strategy was to take a long position in one instrument and to offset the long position by a short position in a similar instrument or its derivative. Most of LTCM's bets were variations on the same theme, convergence between short positions in liquid Treasuries and long positions in theoretically underpriced instruments that commanded a credit or liquidity premium. While profitable for several years, this strategy unravelled in 1998 when Russia defaulted on its rouble and domestic dollar debt. Money fled into high

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

quality instruments such as Treasuries, and LTCM's short positions in these instruments plummeted in value. Although LTCM failed, it presaged the enormous growth in hedge funds that we see today, in 2007.

See Also

- Black, Fischer
- Merton, Robert C.
- Miller, Merton
- options

Bibliography

Bernstein, P.L. 1992. *Capital Ideas: The Improbable Origins of Modern Wall Street*. New York: Free Press.

Black, F. 1989. How we came up with the option formula. *Journal of Portfolio Management* 15, 4–8.

Black, F., Jensen, M.C. and Scholes, M.S. 1972. The capital asset pricing model: some empirical tests. In *Studies in the Theory of Capital Markets*, ed. M.C. Jensen. New York: Praeger.

Black, F. and Scholes, M.S. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Black, F. and Scholes, M.S. 1974. The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics* 1, 1–22.

Bulow, J. and Scholes, M. 1983. Who owns the assets in a defined-benefit pension plan? In *Financial Aspects of the U.S. Pension System*, ed. Z. Bodie and J. Shoven. Chicago: University of Chicago Press.

Miller, M. and Scholes, M.S. 1978. Dividends and taxes. *Journal of Financial Economics* 6, 333–64.

Miller, M. and Scholes, M.S. 1982a. Dividends and taxes: some empirical results. *Journal of Political Economy* 90, 1118–41.

Miller, M. and Scholes, M.S. 1982b. Executive compensation taxes and incentives. In *Financial Economics: Essays in Honor of Paul Cootner*, ed. K. Cootner and W. Sharpe. Englewood Cliffs, NJ: Prentice-Hall.

Modigliani, F. and Miller, M.H. 1961. Dividend policy, growth and the valuation of shares. *Journal of Business* 34, 411–33.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Scholes, M.S. 1991. Stock and compensation. *Journal of Finance* 46, 803–23.

Scholes, M.S. and Williams, J. 1977. Estimating betas from nonsynchronous data. *Journal of Financial Economics* 5, 309–27.

Scholes, M.S. and Wolfson, M.A. 1992. *Taxes and Business Strategy: A Planning Approach*. Englewood Cliffs, NJ: Prentice-Hall.

Sprenkle, C.M. 1961. Warrant prices as indications of expectations. *Yale Economic Essays* 1, 179–232.

How to cite this article

Whited, Toni M. "Scholes, Myron (born 1941)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000456> doi:10.1057/9780230226203.1485

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

school choice and competition

Caroline Hoxby
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The two insights behind school choice are that externalities motivate government financing, but not provision, of schools, and that government-run schools will probably, owing to lack of competition, be x-inefficient. Under school choice, students choose among schools that compete for them on a truly level playing field. Governments play a financing, auditing, refereeing, and market-design role. Modern research focuses on evaluating the effects of school choice on the x-efficiency of schools, analysing how programme design affects the supply of schools, and investigating how information and matching mechanisms affect demand.

Keywords

charter schools; deferred-acceptance mechanism; educational externalities; Friedman, M.; magnet schools; school choice; school competition; school vouchers; strategy-proof mechanisms; Tiebout choice; top trading cycles; tuition tax credits; vouchers; x-efficiency

Article

The economic idea of school choice and competition was introduced by Milton Friedman in two related essays, both entitled ‘The Role of Government in Education’ (1955; 1962). His proposals remain the core of the school choice idea, but myriad economists and non-economists have refined it since. The essence of the idea is that students should be free to choose among schools that compete for them on a level playing field. Students should carry the resources for their education with them in the form of a tax-financed fee. Schools that attract students should be able to expand, possibly by creating branded spin-offs. Schools that fail to attract students should close. Families should play the lynchpin role by making choices that govern students’ experiences and schools’ sustainability. The government, in contrast, should create structures that make schools compete on the basis of the value they add to students, where parents are the ultimate judges of value. For instance, the government should raise taxes that support

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
fees, ensure that schools provide accurate information to parents, prevent financial malfeasance, enforce contracts and health and safety standards, and provide mechanisms that clear students' preference lists for schools and clear the market in school buildings and equipment.

The logic of school choice

Friedman's first insight was that externalities associated with primary and secondary education provide substantial motivation for the government to finance schools, but much less motivation for the government to *run* them. A more modern view would suggest that the government does not merely play a financing role but also plays the market referee role that it adopts in a variety of other industries with public goods dimensions. The government might also engage in limited regulation. For instance, if society decides that all students should learn civics in order to be responsible citizens, then a plausible regulation might impose a test of civics knowledge. So long as the government does not attempt to manage schools, but confines itself to setting parameters and auditing outcomes, the essence of the school choice idea is retained.

Friedman's second insight was that the typical state-run school has a virtual monopoly over certain students and is therefore likely to be x-inefficient, taking rents in forms such as overemployment and a quiet life for managers (for instance, overpaying for school supplies or failing to fire problem teachers). The effect of school choice on x-efficiency is paramount: most supporters envision it to be the main benefit of school choice. Some hypothesize that school choice might also deliver benefits by allowing students to match themselves to pedagogical methods or peers especially conducive to their learning. However, neither of these matching benefits is essential to the logic of school choice, and it is fairly easy to describe situations in which allowing students to choose their peers helps some to learn and hinders others (Epple and Romano, 1998).

School choice can be and has been implemented in numerous forms. The schools involved may be private ('school vouchers'), state schools that are independently managed ('charter schools'), or state schools run by fully autonomous local governments. In practice, it is the supply side that distinguishes true school choice. Beware of programmes that adopt the choice nomenclature but exhibit supply side inflexibility (schools do not readily open, expand, shrink or close) or that structure fees in such a way that schools do not compete on a level playing field. The essence of the school choice idea is that incentives flow from the consequences of competition among schools on the basis of their value-added (as judged by parents). If competition is inconsequential or biased in favour of certain schools, then the school choice idea is not being implemented. Most magnet school programmes, controlled choice programmes, and open enrolment programmes fail to be school choice – largely because competition is inconsequential. On the other hand, some phenomena not explicitly labelled as school choice fulfil the idea to some extent. Examples include tuition tax credits (see Lips and Feinberg, 2006) and competition among autonomous local school districts, a phenomenon called Tiebout choice (Tiebout, 1956).

Evaluations of school choice programmes

Modern work on school choice takes three main forms: empirical evaluations of school choice programmes, design work (theoretical analysis, simulation and experiments) on demand-side problems,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and design work on supply side problems.

The primary task for an evaluation is identifying the general equilibrium effects of school choice on schools' x-efficiency and on students' peer groups. (It is far more important to identify the effects on x-efficiency than the effects on sorting. This is because it appears – as explained below – that sorting is susceptible to management.) To identify the general equilibrium effects, a researcher needs to find fairly random circumstances that cause some geographic areas to have more pressure from school choice than other areas do. By comparing areas randomly subjected to more and less school choice, general equilibrium effects are revealed. So far, the most convincingly random circumstances have come from the prevalence of natural boundaries (more natural boundaries translate into more local jurisdictions and thus more Tiebout choice: see Hoxby, 2000) and laws that translate small differences in state schools' performance into large differences in the school choice they face. For instance, Chakrabarti (2006c), using regression discontinuity combined with the implementation of a school choice policy, compares Florida schools ranked just above and below the state's threshold for failure. Those ranked just below were exposed to school choice pressure starting in a certain year; those just above were exposed to none. Chakrabarti finds statistically significant improvements in x-efficiency in the traditional state schools exposed to pressure. Unfortunately for the sake of research, the amount of pressure on these schools was small and discontinued after a short time because the programme in question ended.

Some researchers have attempted to rely solely on the implementation of school choice programmes to identify the general equilibrium effects of school choice. However, simple before–after comparisons rarely generate convincing estimates because societies do not implement school choice randomly. They usually do so when concerned about current conditions and after policy debates that often result in the simultaneous introduction of other educational policies. A counterfactual constructed from pre-school choice data is therefore rarely convincing. Even worse, because the policy they are studying is often uniform across the state or country in question, authors of before–after studies are often tempted to use *endogenous* geographic variation in the take-up of the choice programme. This is highly problematic because students who take up the programme are far less likely to be satisfied with their traditional state school than are students who have the same opportunity to take it up but refuse to do so. Hsieh and Urquiola (2006) is one example of such problematic studies. Of course, more credence can be given to studies that combine before–after variation with plausibly exogenous geographic variation in the *availability* (not take-up) of the choice programme (Akabayashi, 2006; Ahlin, 2003; Chakrabarti, 2006a; Hoxby, 2004).

If researchers are to have a serious chance of identifying the general equilibrium effects of school choice, they must study an instance in which the supply of choice schools is fairly ample and can expand with demand. After all, x-efficiency effects cannot be expected if students cannot leave traditional state schools because there are no places available at choice schools. Also, the student sorting consequences of a tiny programme (such as one that affects only a few per cent of students) are unlikely to be distinguishable. Ironically, the very conditions that impede general equilibrium analysis facilitate analysis of the partial effect of particular choice schools on the students who attend them. If choice schools are inadequately supplied so that they enrol only a tiny share of students and are routinely oversubscribed, then researchers are in a fairly good position to identify the change in achievement caused by attending a particular choice school – not including the effects that general equilibrium might have on the school.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Why is this? Because students select to participate in choice programmes, selection biases plague studies that attempt to estimate the partial effect of voucher-taking schools, charter schools, and other choice schools on the students who attend them. Researchers have attacked the selection problem with a variety of methods, but the only one that enjoys widespread approbation is randomizing applicants so that some are offered a place and others are not. Randomization often occurs naturally when a choice school is oversubscribed. It allows researchers to compare lotteried-in and lotteried-out students, all of whom have the motivation and other qualities that made them self-select into applying. Formally, one estimates a local average treatment-on-the-treated effect by instrumenting for attendance at a choice school with an indicator for being lotteried into a choice school (see Angrist, Imbens, and Rubin, 1996 for a description of the method). Most randomization-based studies have found positive and significant effects of attending a choice school on outcomes such as test scores, educational attainment, parental satisfaction, altruism, and some other social outcomes (Angrist et al., 2001; Angrist, Bettinger and Kremer, 2006; Bettinger and Slonim, 2006; Howell and Peterson, 2002; Howell et al., 2002; Hoxby and Rockoff, 2005; Hoxby and Murarka, 2007a; 2007b; Peterson and Howell, 2004; Peterson et al., 1999; Rouse, 1998; West, Peterson and Campbell, 2001; Wolf, Howell and Peterson, 2000; Wolf, Peterson and West, 2001; Wolf and Hoople, 2006). Randomization-based studies have non-trivial limitations, however. Most importantly, it is not obvious that the partial effect matters much. If a choice school created incentives or sorting such that students' achievement rose equally in all schools (including the school itself), we would surely consider the school to be a success. Yet the school's partial effect would be nil. In addition, the results of randomized methods cannot be extrapolated to students who differ substantially from the applicants, and one cannot evaluate choice schools that are undersubscribed.

In addition to randomization, researchers have tried a variety of methods – instrumental variables, regression discontinuity, matching – to estimate an unbiased partial effect of choice schools. Although such methods could in theory be useful, they have in practice been substantially less convincing than randomization. Researchers have also often tried panel methods, but these are really never convincing owing to problems that are familiar to anyone who has studied the extensive programme evaluation literature on self-selection into training schemes. If a student attends a traditional state school for some years and then applies to, say, a charter school, his application is not a random event but deliberate selection into treatment. (Charter schools are state schools that families are fully free to choose and that have autonomous management and finances. The fee each child brings with him is tax-financed, applicants are admitted via a lottery, and schools can lose their charter if they do not produce positive results.) The student's pre-application achievement trajectory, no matter how well estimated, is an unreliable predictor of how he would perform post-application in the counterfactual world in which he was forced to stay in his traditional state school. His decision to apply to a choice school is an indication of concern about his current trajectory and of an intention to change it in some way. Therefore, in the counterfactual world, his trajectory might reasonably be expected to fall (if the choice school was the only treatment likely to help him) or rise (if he were to undertake other remedies in the absence of a choice school's being available): see Hoxby and Murarka (2007b) for further discussion of this point.

The final major problem that affects empirical evaluations is the substantial divergence between most actual school choice programmes and the school choice idea. For instance, many programmes have fees that are only a fraction of the per-pupil amounts enjoyed by traditional state schools, thereby creating a playing field that is obviously not level. Most school choice programmes have strict limits on the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
number of students and schools who can participate so that the programmes generate no pressure for x-efficiency once the limits are reached, as they often quickly are (see Lips and Feinberg, 2006). If such divergences between the idea and implementation of school choice were the inevitable result of applying an idea to a world of imperfect information, then researchers could interpret their estimates as the effects of *practical* school choice. The divergences are not inevitable, however; they result from legislators bowing to opponents of school choice. They thus leave researchers with an awkward choice between (i) abstaining from interpreting their results as a test of school choice and (ii) extrapolating their results beyond the data in order to form a meaningful test. Good studies do some of both, combining careful extrapolation with warnings not to interpret their results as definitive tests of the school choice idea (for example, Chakrabarti, 2006b). Unfortunately, numerous studies do neither.

Design work on demand-side problems in school choice

The second major form of modern work on school choice is design work (simulation, theoretical analysis and actual experiments) on demand-side problems. Prominent demand-side problems are how to ensure that families use accurate information about schools (especially schools' value-added when making decisions; how to ensure that students get their most preferred school (given the competing demands of other students) in a world where students submit a preference ordering; and how to incorporate social priorities such as desires to have schools that are socio-economically integrated or neighbourhood-focused. Progress on demand-side problems is occurring in part because such problems are susceptible to remedy, though perhaps not solution, with established tools. Also, districts that run school choice programmes have demonstrated a willingness to provide data for simulations and even to run experiments. For instance, although no information on schools' value-added is perfect or complete, econometrics and data systems have been focused on providing such information more often and more accurately. Districts like Charlotte-Mecklenberg, North Carolina, have engaged in experiments, giving more complete and digestible information to randomly selected families. Such experiments suggest that even modest improvements in the information received by families cause them to revise their choices (Hastings and Van Weelden, 2007).

Just as important, economists have proposed mechanisms for processing students' preference lists in order to guarantee that students get their most preferred choice, given other students' preferences. A good mechanism is one that is strategy-proof and that produces a stable match. A strategy-proof mechanism is one in which a student can do no better than submit his true preferences. A stable match is one in which there is no student and school pair where the student prefers the school over his assignment and the school gives the student a higher priority than a student who has been assigned to the school. For example, Random Serial Dictatorship is strategy-proof and stable. It takes the preference lists submitted by students, assigns each student a random number, assigns the first student his top choice, the second student his top choice among available schools, and so on. Not only theoretical proofs but also evidence based on actual experiments – that is, districts switching their mechanisms – have shown that such mechanisms successfully clear the market (Abdulkadiro•lu, Pathak and Roth, 2006; Abdulkadiro•lu et al., 2006; Pathak, 2006). For instance, after New York City switched to a stable, strategy-proof mechanism, families and schools engaged in substantially less strategic behaviour and families were more satisfied with the schools their children ended up attending.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Stable, strategy-proof mechanisms can incorporate some social priorities. In practice, priorities are most often given to siblings, economically disadvantaged children, children who attend schools classified as failing, disabled children, and children with limited English. (Race- and ethnicity-based priorities are being phased out by courts.) With social priorities, not only students but also schools submit ordered lists. Schools' lists, however, are based on the priorities, not on the personal preferences of school staff. The deferred-acceptance mechanism deals with the lists as follows.

> (Step 1) Each student 'proposes' to his first choice. Each school tentatively assigns its seats to its proposers one at a time following only their priority order. Any proposers who remain when all seats are assigned are rejected.
> (Each subsequent step) Each student who was rejected in the previous step 'proposes' to his or her next choice. Each school considers the students it has been holding together with its new proposers and tentatively assigns its seats to these students one at a time following their priority order. Any proposers who remain when all seats as assigned are rejected.
> (Termination) The mechanism stops when no student proposal is rejected. Each student is assigned to his or her final tentative assignment.

The above description is borrowed from Pathak (2006). Another mechanism that can be used in this situation is top trading cycles (Abdulkadiro•lu and Sönmez, 2003).
The availability of mechanisms that incorporate social priorities implies that the general equilibrium effects of school choice on student sorting can be *managed*. By *managed*, one means that unwelcome or extreme outcomes such as segregation can be avoided. One does not mean that a robust school choice programme would have no effects on student sorting: it is hard to imagine a programme in which students' choices are meaningful and yet so micro-managed that sorting looked the same before and after programme implementation.

Design work on supply side problems in school choice

The third major form of modern work on school choice is design work on supply side problems. This work attempts to address the questions about which programme designs are most likely to ensure a level playing field for schools in competition with one another. Most importantly, what should the structure of fees be? What should vouchers or charter school fees be? How much should they be topped-up for students who are disabled, limited-English proficient, economically disadvantaged or otherwise expensive to educate? There is also a series of questions about growing and shrinking schools. For instance, should a school that is very oversubscribed be given funds to expand its school or move to a larger one? Should shrinking schools be given incentives to cede their space to growing schools? Unlike the demand-side work, supply side work is limited and may remain so for some time. This is partly because actual experiments are rare: governments choose the supply side policies of their school choice programme after considerable political wrangling and then stick with them. Moreover, supply side work that uses theory or simulation has so far not produced general truths. Instead, it generates results that are sensitive to assumptions about peer effects and the effects of school spending on achievement (see Epple

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and Romano, 1998; Nechyba, 2003). Yet there is no way, at present, to ground such assumptions in agreed facts. Suppose one asks how much additional money is needed by a school if it is to achieve the same value-added with a child whose parents are poor secondary school drop-outs as it would with a child whose parents are affluent college graduates. Estimates vary enormously, are very imprecise, and are sometimes even 'wrong-signed' (Hanushek, 2002). Yet, an estimate is what one needs to model the effect of raising the amount by which the basic fee is topped up for disadvantaged students.

At present, as of 2007, evidence-based design work on the supply side is fairly indirect. Authors occasionally compare a programme before and after its design has been changed (Chakrabarti, 2006a; Hoxby, 2004). Comparing programmes across states has also been tried (Chakrabarti, 2006b; Hoxby, 2006), but this is inherently difficult because states differ along many lines, not just a single dimension of school choice design. Because, for a test of the school choice idea, it is crucial that schools compete on an even playing field, design work on supply side problems is important. It is most likely to be advanced by deliberate experimentation within school choice programmes – for instance, varying the fee structure in interesting ways across different metropolitan areas within a state.

See Also

- education production functions
- educational finance
- local public finance
- matching and market design
- mechanism design
- Tiebout hypothesis

Bibliography

Abdulkadiro•lu, A., Pathak, P. and Roth, A. 2006. Strategy-proofness versus efficiency: redesigning the New York City High School match. Working paper, New York University and Columbia University.

Abdulkadiro•lu, A., Pathak, P., Roth, A. and Sönmez, T. 2006. Changing the Boston school choice mechanism. Working Paper No. 11965. Cambridge, MA: NBER.

Abdulkadiro•lu, A. and Sönmez, T. 2003. School choice: a mechanism design approach. *American Economic Review* 93, 729–47.

Ahlin, Å. 2003. Does school competition matter? Effects of a large-scale school choice reform on student performance. Working paper, Nationalekonomiska institutionen, Uppsala University.

Akabayashi, H. 2006. Average effects of school choice on educational attainment: evidence from Japanese high school attendance zones. Working paper, Keio University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Angrist, J., Bettinger, E. and Kremer, M. 2006. Long-term consequences of secondary school vouchers: evidence from administrative records in Colombia. *American Economic Review* 96, 847–62.

Angrist, J., Bettinger, E., Bloom, E., King, E. and Kremer, M. 2001. Vouchers for private schooling in Colombia: evidence from a randomized natural experiment. Working Paper No. 8343. Cambridge, MA: NBER.

Angrist, J., Imbens, G. and Rubin, D. 1996. Identification of causal effects esing instrumental variables. *Journal of the American Statistical Association* 91, 444–55.

Bettinger, E. and Slonim, R. 2006. Using experimental economics to measure the effects of a natural educational experiment on altruism. *Journal of Public Economics* 90, 1625–48.

Chakrabarti, R. 2006a. Can increasing private school participation and monetary loss in a voucher program affect public school performance? Evidence from Milwaukee. Working paper, Federal Reserve Bank of New York.

Chakrabarti, R. 2006b. Impact of voucher design on public school performance: evidence from Florida and Milwaukee voucher programs. Working paper, Federal Reserve Bank of New York.

Chakrabarti, R. 2006c. Vouchers, public school response and the role of incentives: evidence from Florida. Working paper, Federal Reserve Bank of New York.

Epple, D. and Romano, R. 1998. Competition between private and public schools, vouchers, and peer-group effects. *American Economic Review* 88, 33–62.

Friedman, M. 1955. The role of government in education. In *Economics and the Public Interest*, ed. R. Solo. New Brunswick, NJ: Rutgers University Press.

Friedman, M. 1962. The role of government in education. In *Capitalism and Freedom*. Chicago: University of Chicago Press.

Hanushek, E. 2002. Publicly provided education. In *Handbook of Public Economics*, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Hastings, J. and Van Weelden, R. 2007. Preferences, information, and parental choice behavior in public school choice. Working Paper No. 12995. Cambridge, MA: NBER.

Howell, W. and Peterson, P. 2002. *The Education Gap: Vouchers and Urban Schools*. Washington, DC: Brookings Institution.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Howell, W., Wolf, P., Campbell, D. and Peterson, P. 2002. School vouchers and academic performance: results from three randomized field trials. *Journal of Policy Analysis and Management* 21, 191–218.

Hoxby, C. 2000. Does competition among public schools benefit students and taxpayers? *American Economic Review* 90, 1209–38.

Hoxby, C. 2004. School choice and school competition: evidence from the United States. *Swedish Economic Policy Review* 10, 9–65.

Hoxby, C. 2006. The supply of charter schools. In *Charter Schools against the Odds*, ed. P. Hill. Stanford: Hoover Institution Press.

Hoxby, C. and Murarka, S. 2007a. How New York City's charter schools affect student achievement: results of the first year of evaluation. Working Paper No. 13255. Cambridge, MA: NBER.

Hoxby, C. and Murarka, S. 2007b. Methods of assessing the achievement of students in charter schools. In *Charter Schools: Their Growth and Outcomes*, ed. M. Berends. Mahwah, NJ: Lawrence Erlbaum.

Hoxby, C. and Rockoff, J. 2005. The impact of charter schools on student achievement. Working paper, Harvard University.

Hsieh, C. and Urquiola, M. 2006. The effects of generalized school choice on achievement and stratification: evidence from Chile's voucher program. *Journal of Public Economics* 90, 1477–503.

Lips, D. and Feinberg, E. 2006. School choice: 2006 progress report. Washington, DC: Heritage Foundation. Online. Available at http://www.heritage.org/Research/Education/bg1970.cfm, accessed 26 June 2007.

Nechyba, T. 2003. What can be (and what has been) learned from general equilibrium simulation models of school finance? *National Tax Journal* 56, 387–414.

Pathak, P. 2006. Lotteries in student assignment. Working paper, Harvard University.

Peterson, P. and Howell, W. 2004. Voucher research controversy: new looks at the New York City evaluation. *Education Next* 4, 73–8.

Peterson, P., Myers, D., Howell, W. and Mayer, D. 1999. School choice in New York City. In *Earning and Learning: How Schools Matter*, ed. P. Peterson and S. Mayer. Washington, DC: Brookings Institution.

Rouse, C. 1998. Private school vouchers and student achievement: an evaluation of the Milwaukee

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

parental choice program. *Quarterly Journal of Economics* 113, 553–602.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

West, M., Peterson, P. and Campbell, D. 2001. School choice in Dayton, Ohio after two years: an evaluation of the parents advancing choice in education scholarship program. Working Paper No. RWP02-021. Cambridge, MA: Kennedy School of Government.

Wolf, P., Howell, W. and Peterson, P. 2000. School choice in Washington, D.C.: an evaluation after one year. Working Paper No. PEPG/00–08. Program on Education Policy and Governance, Harvard University.

Wolf, P. and Hoople, D. 2006. Looking inside the black box: what school factors explain voucher gains in Washington, DC? *Peabody Journal of Education* 81, 7–26.

Wolf, P., Peterson, P. and West, M. 2001. Results of a school voucher experiment: the Case of Washington, D.C. after two years. Working Paper No. PEPG/01–05. Program on Education Policy and Governance, Harvard University.

How to cite this article

Hoxby, Caroline. "school choice and competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_V000064> doi:10.1057/9780230226203.1486

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schultz, Henry (1893 – 1938)

Jacob L. Mosak

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

econometrics; exchange; mathematical economics; Moore, H. L.; Schultz, H.; statistics and economics

Article

Schultz was one of a small group of pioneering econometricians who, in the 1920s and 1930s, laid the groundwork for the phenomenal development of mathematical economics and econometrics that occurred after the Second World War. His graduate courses in mathematical economics and statistics inspired students to address themselves to economic problems in quantitative terms, to reformulate economic theory in empirically testable form, and to test theories by means of diligent search for relevant statistical information and careful application of appropriate statistical analysis. His own research, culminating in his magnum opus, *The Theory and Measurement of Demand* (1938), could well serve as a model for a proper approach to economic analysis today. Schultz devoted all his professional life to the integration of pure economic theory with empirical analysis. Unlike considerable econometric work today, which is often empirical without much grounding in economic theory, his statistical analysis is solidly based on mathematical economic theory, as well as on the statistical theory of correlation and curve-fitting. Elegant summaries of both fields, based in large part upon his lecture notes and his research work, appear in the book, along with the empirical studies of demand for a large number of agricultural commodities for which the theories served as the basic foundation. The student wishing to get a good introduction to mathematical economics as formulated by Cournot, Walras and Pareto, and to the fundamentals of Gaussian curve-fitting analysis, will find clear and lucid presentations of these subjects in Schultz's book. At the same time he will not fail to be impressed by the extraordinary concern for statistical accuracy and precision demonstrated in the empirical analysis throughout the book. Schultz, who was born in Russian Poland on 4 September 1893, was brought by his parents to the United States in 1907, as part of the large wave of migration of Russian Jews after the Russo–Japanese War of 1905. Despite the family's poverty, Schultz's drive and determination enabled him to enter college and even to pursue a graduate education – no small feat in those days for the oldest child of an immigrant family. After receiving his AB from City College in 1916, Schultz entered Columbia

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
University, where he came under the lasting influence of one of the world's leading econometricians, Professor Henry L. Moore. Schultz felt so indebted to him that in 1938, when he was himself internationally recognized as an outstanding authority in econometrics, he dedicated his major work to ‘Professor Henry Ludwell Moore, trail blazer in the statistical study of demand’.

Schultz's studies were interrupted by the First World War, during which he was wounded in the Meuse–Argonne offensive. In the spring and summer of 1919, an Army scholarship enabled him to study at the London School of Economics and at the Galton Laboratory of University College, London, under two leading statisticians, A.L. Bowley and Karl Pearson. After returning to the United States in 1920, he served with several agencies of the United States government and became Director of Statistical Research of the Children's Bureau of the Department of Labor. At the same time he continued his academic work, receiving the Ph.D. degree from Columbia University in 1925. His dissertation on ‘The Statistical Law of Demand as Illustrated by the Demand for Sugar’ was published the same year in the *Journal of Political Economy*.

The following year Schultz received his appointment at the University of Chicago. His courses in mathematical economics and statistics were highly organized and presented his students with a clear and systematic exposition both of the classic texts and of the most important up-to-date journal articles from the English, French, Italian and Russian literature. He was a voracious reader of economic and statistical literature, and enriched his courses by examination of related material in the fields of biology, physics and psychology.

It was the research laboratory, however, that absorbed most of his energies. Almost from the time he came to the University of Chicago, he embarked on an ambitious research project intended to harness theory and empirical analysis for the purpose of filling some of the ‘empty boxes’ in the theory of exchange. At first his interest focused on determining the coefficients of elasticity of demand as well as the magnitudes of the average shifts in demand over time. At a later stage be became interested in testing the consistency of demand coefficients for a set of commodities whose demands were interrelated. The work was so thoroughly organized, documented and proof-checked, and the calculations were so systematically set out, with automatic sum-checks at every appropriate point, that research assistants picking up a worksheet prepared by others even a decade earlier had no difficulty tracing the sources and checking the accuracy of every figure on the sheet.

Shortly after his book was published, Schultz was killed, together with his wife and two daughters, on 26 November 1938, in a car accident in California, where he had gone to teach on sabbatical leave from the University of Chicago. Paul Douglas wrote of Schultz:

> All in all, he was about the finest man it has ever been my privilege to know in academic life. The world of scholarship and of science (for I may so use the term in connection with him) was the richer for his fruitful life. We are much the poorer for his death at the full height of his powers.

Selected works

1925. The statistical law of demand as illustrated by the demand for sugar. *Journal of Political Economy* 33, Part I, 481–504; Part II, 577–637.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1938. *The Theory and Measurement of Demand*. Chicago: University of Chicago Press.

How to cite this article

Mosak, Jacob L. "Schultz, Henry (1893–1938)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000040> doi:10.1057/9780230226203.1487

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schultz, T.W. (1902 – 1998)

James R. Walker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Theodore W. Schultz transformed the study of agricultural economics and economic growth and development. He made seminal contributions advocating the importance of human capital and particularly education for understanding economic growth in developed and developing countries. In this brief sketch I review his intellectual contributions to economics.

Keywords

agricultural economics; economic demography; economic growth; education in developing countries; Green Revolution; human capital; Marshall, A.; measurement; poverty; Schultz, T. W.

Article

Theodore Schultz shared the Nobel Prize in 1979 with Sir Arthur Lewis ‘for pioneering research into economic development research with particular consideration of the problems of developing countries’. Early in his career, Schultz studied agricultural organization and production, which evolved into studies of economic growth and finally into studies of economic development. His training and early contributions were in agricultural economics, but it is too limiting to label him an agricultural economist (as did the press release announcing his Nobel Prize award). He made seminal contributions in agriculture but also was a 20th-century pioneer advocating the importance of human capital and particularly education for understanding economic growth in developed and developing countries. In this brief sketch I review his intellectual contributions to economics.

Biographical details

T.W. Schultz was born on 30 April 1902 in South Dakota and died on 26 February 1998. He received his undergraduate degree in 1927 from South Dakota State University. He was a student of John R.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Commons and received his Ph.D. in Agricultural Economics from the University Wisconsin in 1930. Besides the Nobel Prize he received the Francis A. Walker Medal from the American Economic Association (1972), and the Leonard Elmhurst Medal from the International Agricultural Economic Association (1976). He was President of the American Economic Association in 1960, a member of the National Academy of Science (1974), and a fellow of the American Farm Economic Association (1957), the American Academy of Arts and Sciences (1958), the American Philosophical Society (1962), a guest of the Soviet Academy of Science (1960) and a Founding member of the National Academy of Education (1965). Also, he received about half a dozen honorary degrees from universities in the United States and abroad. Nerlove (1999) offers a longer and more detailed appraisal of Schultz's contributions and provides a complete listing of Schultz's prolific writings; see Johnson (1998) and Nerlove (1999) for a commentary on Schultz's personality. They give strong testimony to Schultz's role as mentor and colleague.

Style

Three features of Schultz's style characterize his writings. First, his wit and humility are much on display. Schultz travelled the world and served as an advisor to many governments and international organizations but he retained his humble Midwestern perspective and values. His work is serious and deeply considered, but presented with an irreverence to self and others that makes for easy and enjoyable reading. (A hint of this irreverence is evident in his 1979 Nobel Banquet Speech; Schultz, 1980.) Second, his analysis is thoroughly modern; he thought in terms of supply and demand, and gained much analytical depth through careful assessment of opportunities, incentives and information. His analyses considered a broad sweep of potential factors such as political economy and institutional constraints, besides the more obvious economic factors all economists are trained to recognize. Indeed, one characterization of Schultz's approach is that he merged the analytical insight of Irving Fisher with the breath and style of argumentation of (his mentor) John R. Commons. Schultz applied these tools with equal force to problems in agriculture, labour, public economics, macroeconomics and development. He was not bound by traditional field–subfield definitions, but, rather, his writing is problem-focused; for example, on understanding the economics of being poor. In his Presidential Address to the American Economics Association (AEA), for instance, Schultz (1961) applied ideas of human capital to age–earnings profiles, population flows from rural to urban areas and deciphering the mysteries of economic growth for developed and developing countries. And, as another example, in *Transforming Traditional Agriculture* (1964), analyses of the US agricultural market are seamlessly interwoven with analyses of agricultural markets in developing countries. This uniformity of approach is all the more astonishing when compared to his contemporaries who viewed farmers as ‘different’ from non-farmers and especially saw differences between farmers in the United States and those in developing countries. Schultz had enormous faith in people's common-sense response to incentives they face. He believed people were knowledgeable about economic opportunities and responded to those opportunities. He tacitly assumed preferences were constant and sought explanation and prediction in responses to differences in opportunities or abilities. This perspective also made him critical of many government-sponsored agricultural and anti-poverty programmes. Schultz recognized that programmes could distort individual incentives in unanticipated ways or would fail because they did not remove barriers permitting individuals to act on existing economic incentives.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Schultz wrote frequently about ‘disequilibrium’ and its importance. However, I think the appropriate modern term is ‘dynamics’, not disequilibrium. Academic economists perceive a clinical Walrasian auctioneer at work to obtain equilibrium. Schultz saw a myriad of ‘equilibrium distortions’ that create profit opportunities, which individuals perceive and arbitrage. Schultz was interested in how farmers, students and entrepreneurs reallocate resources in response to new information on economic costs and returns. In ‘The Value of the Ability to Deal with Disequilibrium’ (1975) Schultz argues that the more educated are more able and thus quicker to process information on the economic environment. Their early arbitrage activity provides, Schultz argued, another return to education.

Third, even more impressive than Schultz's broad conceptual approach to economics was his tireless work in linking economic theory with economic measurement. In his empirical studies, understanding the economic phenomenon takes centre stage, not statistical technique or theoretical elegance. Empirical measures are fully described. He considered possible biases and other deficiencies in each. Finally, Schultz assesses the likely quantitative magnitude of each bias and the consequences for empirical analysis. Every effort is made to get the closest connection between available (generally aggregate or tabular) measures and the theoretical concepts. The honesty and the fullness of the presentation makes the analysis compelling.

Agricultural economics

In his 1979 Nobel Prize lecture Schultz summarized the motivation for his research thus: ‘Most of the people in the world are poor, so if we knew the economics of being poor, we would know much of the economics that really matters. Most of the world's poor people earn their living from agriculture, so if we knew the economics of agriculture, we would know much of the economics of being poor’ (Schultz, 1992). His seminal works on agriculture include *Agriculture in an Unstable Economy* (1945) and *Production and Welfare of Agriculture* (1949) and *Transforming Traditional Agriculture* (1964). According to Schultz (1993, p. 2) understanding the economics of being poor starts with the hypothesis that there are relatively few significant economic inefficiencies in established communities where most people are poor. In these traditional agricultural settings, farmers have used the same technology and the same factor inputs for generations. Consequently, they have acquired significant experience in their abilities and the means of production available to them. They live and operate within a stationary environment in which there has been no significant change in technology for generations. Thus, contrary to the claims of Schultz's contemporaries, it is not that these households saved and invested too little and did not respond to normal economic incentives. Rather, it was that they did not have profitable opportunities. Schultz pushed the frontier on technological change in agriculture as the key factor for transforming traditional agriculture by creating profit opportunities for investment. Schultz recognized that the Green Revolution created new high-yielding varieties, which were more responsive to fertilizers and other modern inputs that helped offset the diminishing marginal utility of land and created profit opportunities that led to investment and economic growth. Indeed, it is the availability of new technology and high return investment opportunities that characterizes modern agriculture. Schultz encouraged his student Zvi Griliches to investigate the diffusion of the new hybrids (Griliches, 1957). *Transforming Traditional Agriculture* (1964) is Schultz's forceful summary of the transition. Thomas Balogh's (1964) vitriolic review of *Transforming Traditional Agriculture* is the best evidence of its revolutionary nature.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Human capital

Schultz's research in agricultural production evolved into a study of economic growth. In the 1930s, Schultz began to see that new fertilizers expanded the productivity capacity of land. Yet he quickly realized that technological advances could not explain all the gains in productivity; a search was on for a more complete explanation. In the 1940s he came to see ‘acquired ability of labour’ as a major source of the unexplained gains in productivity. Scarce resources produced these augmented abilities; the stage was set for a formal study of the investment in man.

To study the investment in man, a new concept of capital was needed. Schultz recognized that many eminent economists before him (notably Adam Smith, Irving Fisher and Frank Knight) considered human abilities as capital. In his writings in the 1950s and 1960s, Schultz was critical of Alfred Marshall's perspective on capital that included only physical equipment. (Schultz, 1993, p. 4, presents a revised, more generous assessment of Marshall's perspective on human capital. From interactions with Marshall's defenders, Schultz accepted that analytically Marshall saw human investments as a form of capital but considered human capital as impractical because it was divorced from the marketplace.)

Schultz argued that Marshall's narrow definition of capital, among other deficiencies, led to the popular perception that economics studied only material things. And more perniciously the restricted definition led to the notion that productivity of labour is homogenous and independent of capital, so only the number of hours of work matter. (Marshall's view remained popular among leading theorists outside Chicago and Columbia. The role of physical capital was emphasized by the World Bank, for example, throughout the 1960s and 1970s. Human capital made its way into aggregate growth models only in the 1980s. It is telling that the 1979 Nobel Committee press release considers ‘Schultz on the Human Factor’ and makes only one reference to human capital, and then only in quotes. Some 13 years later human capital was more accepted and used without quote marks in the press release announcing Gary Becker's Nobel Award.)

By the mid-1950s, while a Fellow at the Center for Advanced Study, Schultz's research in the economics of education took shape:

> During the year at the Center, I began to see that the productive essences of what I was identifying as capital and labour were not constant but were being improved over time and that these improvements were being left out in what I was measuring as capital and labour. It became clear to me also that in the United States many people are investing in man are having a pervasive influence upon economic growth, and that the key investment in human capital is education. (Schultz, 1963, p. viii)

Schultz spent the last 40 years of his career understanding investments ‘in man’.

Human capital to Schultz was the acquisition of all useful skills and knowledge that is part of deliberate investment (Schultz, 1961, p. 1). Rather than offer formal definitions, Schultz defined human capital by example:

> Much of what we call consumption constitutes investment in human capital. Direct

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> expenditures on education, health, and internal migration to take advantage of better job opportunities are clear examples. Earnings foregone by mature students attending school and by workers acquiring on-the-job training are equally clear examples. Yet, nowhere do these enter our national accounts. The use of leisure time to improve skills and knowledge is widespread and it too is unrecorded. In these and similar ways the *quality* of human effort can be greatly improved and its productivity enhanced. I shall contend that such investments in human capital accounts for most of the impressive rise in real earnings per worker. (Schultz, 1961, p. 1)

Schultz's research on human capital sought to clarify the investment process and the incentives to invest in human capital. He studied mainly formal education and organized research. The application of the investment approach in the intervening 40 years expanded to consider an array of different forms in many vistas.

It was Schultz's overarching view of human capital that made him a 20th-century pioneer in human capital theory. Jacob Mincer and Gary Becker were pioneers as well, but they focused on the effect of human capital on the level and distribution of earnings: it was Schultz who pushed the profession to see human capital investments in their totality – education, training, work experience, migration and health. This totality of vision made Schultz a leader in the re-emergence of economic demography. He organized two conferences, in 1972 and 1973, whose proceedings were originally published in the March 1973 and March 1974 in issues of the *Journal of Political Economy* and subsequently published in book form as *Economics of the Family: Marriage, Children and Human Capital* (1974). The conferences generated several seminal papers, including Willis on fertility, Becker and Lewis on the quality–quantity dimensions of children, and Mincer and Polacheck on human capital and women's earnings. In ‘Fertility and Economic Values’ (1974) Schultz recognized the large advances published in these studies but pushed the field to extend the static models to consider richer models of life-cycle behaviour and to collect panel data necessary to support their estimation – this was a prophetic vision, as it neatly summarized work in the economics of fertility over the next 30 years.

Envoi

As I reread several of Schultz's papers, I lament the cost of the short half-life of ideas in economics; of what we miss by not reading some of the original formulations. Textbook treatments offer modern notation and language with a well-honed presentation of ideas that speeds initial learning. Yet the vibrancy and freshness (and, yes, sometimes confusion over issues that get sorted out later) of the original frequently is lost. We have lost sight of Schultz's original contributions and it is our loss. Researchers interested in migration, economic demography, development and economic growth would be well served to read some of T.W. Schultz classics. His broad view of economics continues to be fresh and makes for fruitful reading.

The National Science Foundation and the National Institute of Child Health and Human Development provided research support. I thank Glen Cain, T. Paul Schultz and Kenneth I. Wolpin for comments.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1945. *Agriculture in an Unstable Economy*. New York: McGraw-Hill.

1949. *Production and Welfare of Agriculture*. New York: Macmillan.

1953. *The Economic Organization of Agriculture*. New York: McGraw-Hill.

1960. Capital formation by education. *Journal of Political Economy* 68, 571–83.

1961. Investment in human capital. *American Economic Review* 51, 1–17.

1962. Reflections on investment in man. *Journal of Political Economy* 70, S1–S8.

1963. *The Economic Value of Education*. New York: Columbia University Press.

1964. *Transforming Traditional Agriculture*. New Haven, CT: Yale University Press.

1974. Fertility and economic values. In *Economics of the Family, Marriage, Children and Human Capital*. Chicago: University of Chicago Press.

1975. The value of the ability to deal with disequilibrium. *Journal of Economics Literature* 13, 827–46.

1980. Banquet speech, Nobel Banquet, 10 December 1979. In *Les Prix Nobel. The Nobel Prizes 1979*, ed. W. Odelberg. Stockholm: Nobel Foundation. Online Available at http://nobelprize.org/nobel_prizes/economics/laureates/1979/schultz-speech.html, accessed 4 June 2007.

1992. The economics of being poor. In *Nobel Lectures, Economics 1969–80*, ed. A. Lindbeck. Singapore: World Scientific Publishing Co. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1979/schultz-lecture.html, accessed 10 June 2007.

1993. *The Economics of Being Poor*. New York: Basil Blackwell.

See Also

- agricultural economics
- agriculture and economic development
- education in developing countries
- human capital
- poverty

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Balogh, T. 1964. Review of Transforming Traditional Agriculture. *Economic Journal* 70, 996–9.

Griliches, Z. 1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 4, 501–22.

Johnson, D.G. 1998. In memoriam: Theodore W. Schultz. *Economic Development and Cultural Change* 47, 29–213.

Nerlove, M. 1999. Transforming economics: Theodore W. Schultz, 1902–1998: in memoriam. *Economic Journal* 109, F726–48.

How to cite this article

Walker, James R. "Schultz, T.W. (1902–1998)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000041> doi:10.1057/9780230226203.1488

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schumpeter, Joseph Alois (1883 – 1950)

Arnold Heertje
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article on Joseph Alois Schumpeter depicts him as a visionary and as one of the greatest figures in the history of economics. Both his life and his work are described and discussed. His life has not been without disaster and turmoil, and his work on methodology, economic development, innovation, capitalism, socialism and democracy not without originality, antagonism and independence. Above all, his major work on the history of economics is a lasting monument for all students of the emergence and ups and downs of new ideas in a major field of social analysis.

Keywords

business cycles; capitalism; Cournot, A.A.; creative destruction; credit; democracy, economic theory of; diffusion of technology; dynamic efficiency; Econometric Society; economic development; endogeneity and exogeneity; endogenous growth; entrepreneurship; innovation; inventions; Juglar cycles; Keynes, J. M.; Kitchin cycles; Kondratieff cycles; markets; Marshall, A.; methodological individualism; money; monopoly; oligopoly; patents; perfect competition; public choice; Ricardo, D.; Schumpeter, J.A.; static efficiency; structural change; technical change; Walras, L.

Article

Economist and social scientist of Austrian origin, Schumpeter was born in Triesch, Moravia, on 8 February 1883, the son of the owner of a textile factory, and died in Taconic, Connecticut, on 8 January 1950. Schumpeter attended an academic high school in Vienna and studied law and economics at the University of Vienna. In 1908 he published an important book on the nature and content of economic theory, which established his fame as the ablest among the younger group of Austrian economists. Schumpeter had F. von Wieser and E. von Böhm-Bawerk as his teachers. After being nominated at the University of Czernowitz, he became professor of economics at the University of Graz in 1911. His famous book on *The Theory of Economic Development* was published in 1912. Much of his later work

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
on business cycles and the evolution of capitalism into socialism is to a certain extent an elaboration and improvement of the ideas and analysis presented in his book on economic development.

From 1925 to 1932 Schumpeter was a professor at the University of Bonn and in 1932 he became a permanent professor of economics at the Harvard University, a post he held until 1950. His impressive work on *Business Cycles* appeared in 1939, and in 1942 he published *Capitalism, Socialism and Democracy*, in which he predicted the gradual decay of capitalism. On the basis of his numerous works, one would be inclined to think that Schumpeter devoted his whole life to teaching, writing and theorizing. In fact his life has been even more colourful. From spring 1919 up to October of the same year he held the position of Austrian Minister of Finance in Renner's cabinet. During his term he was in favour of sound finance and a capital levy; he even started as a strong defender of massive socialization. He changed his mind, however, partly because he acknowledged the necessity to import capital and gradually his attitude caused tensions with his socialist colleagues in the cabinet, which finally led to his downfall. Between his professorships in Graz and Bonn, Schumpeter accepted the presidency of a Viennese private bank, the Biedermann Bank. Around 1926 the bank went bankrupt and Schumpeter was left with huge debts, probably due to speculations with borrowed money. Schumpeter was married three times, for the first time in 1907 to an Englishwoman, a marriage that ended in divorce in 1920. In 1925 he married a Viennese, 21 years his junior, who died in 1926 in childbirth. In 1937 Schumpeter married Elizabeth Boody, who, after his death in 1950, edited his monumental work, the *History of Economic Analysis* (Stolper, 1994).

In Schumpeter's interpretation of capitalism, the entrepreneur, who applies new combinations of factors of production, plays a central role. He is the innovator, and the agent of economic change and development. Centred around the role of the Schumpeterian entrepreneur is the rise and decay of capitalism. The gifted few, pioneering in the field of new technologies, new products and new markets, carry out innovations and, joined some time later by many imitators, they are at the heart of the short and long cycles observed in economic life. The importance Schumpeter assigns to the creation of money and overexpansion of credit in his early work foreshadows his later work. He argues that since in static analysis there is no room for (*a*) new combinations of factors of production; (*b*) the entrepreneur (in the Schumpeterian sense); and (*c*) profit, there is no need for the further creation of money. He also questions whether in a static context one can speak of economic development at all. Here he is not questioning that economic phenomena and magnitudes change, but he is suggesting that the causes of change may not lie in economic factors. In modern economic theory, this would be expressed by asking whether economic development is due to endogenous or exogenous factors. In the case of exogenous factors, Schumpeter would not speak of economic development at all, for he regards factors such as population growth, consumer preferences, technical development and social organization as non-economic factors. On the other hand, he argues that changes in human nature and social organization can in fact be attributed to economic causes, which can then be regarded as endogenous factors.

The somewhat ambiguous terminology Schumpeter uses in order to describe and explain the development of the economic process is repeated in his book on economic development. Saving is no longer regarded as a factor leading to economic development in the sense of entrepreneurial innovations. Capital formation and the increase of population determine the growth rate in the stationary economy. The terms ‘development’ and ‘economic development’ are not used to describe the actual course of economic events but to distinguish between changes caused in the economic process by endogenous factors and other changes.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Not all endogenous factors lead to economic development, and Schumpeter explicitly excludes continuous endogenous changes. His theory of economic development is reduced to the treatment of spontaneous and discontinuous changes in the economic cycle. The endogenous changes Schumpeter has in mind are not found on the demand side of the economic process, but on the supply side. Economic development consists of the discontinuous introduction of new combinations of products and means of production. The five examples mentioned by Schumpeter show that the term ‘new combination’ must be taken in a very broad sense; it comprises a new product, a new method of production, the opening-up of a new market, the utilization of new raw materials, and the reorganization of sectors of the economy.
He restricts the meaning of the word ‘enterprise’ to the creation of new combinations, and the meaning of the word ‘entrepreneurs’ to those economic figures who introduce new combinations. Schumpeter's entrepreneur operates in an uncertain world, has the courage to start up new ventures, and must be strong enough to swim against the tide of society. In Schumpeter's view, new combinations can be financed only if a successful appeal can be made to the banking system to create money.
According to Schumpeter, ups and downs in economic development can be explained quite simply by the fact that new combinations or innovations appear, if at all, discontinuously in groups or swarms. The appearance of entrepreneurs in ‘bursts’ is due exclusively to the fact that the appearance of one or a few entrepreneurs facilitates the appearance of others. This is the only reason for an upswing in the business cycle. The downturn sets in as a result of smaller profit margins due to imitation and a new equilibrium will be reached after the diffusion process is completed.
In his book on business cycles, Schumpeter sharpened his analytical tools. He introduced the concept of the production function, which – in his words – tells us all we need to know about the technological aspects of production. Schumpeter regards the setting up of a new production function as the introduction of new combinations. The changes caused by innovations are no longer regarded as economic development but as economic evolution. He uses the term ‘technical development’ only for innovations that involve the introduction of new methods of production. Innovations must be distinguished from inventions. The application of new combinations by entrepreneurs is possible without inventions, while inventions as such need not necessarily lead to innovations and need not have any economic consequences. Innovation itself is the independent endogenous factor that causes economic life to go through a number of cycles.
Innovations lead to cyclical fluctuations whose length is determined by both the character and the period of implementation of the innovations. Schumpeter applies this general explanation to the 40-month Kitchin, the ten-year Juglar and the 60-year Kondratieff cycles. The combination of the use of the innovations, overinvestment and of credit expansion going too far, brings the upswing to an end. The recession, which in Schumpeter's view is a healthy phase of restructuring, sets in, paving the way for a new burst of future innovations. Schumpeter's prediction of the decay of capitalism is based on the vision that it is not economic failure but rather the economic success of capitalism that causes the march into socialism. Social rather than economic factors are according to Schumpeter responsible for the structural change in the organization of society (Schumpeter, 1939).
Typical of this economic scene is the process of creative destruction. In Schumpeter's view it is the essential fact about capitalism. The process of creative destruction concerns the implementation of new combinations that incessantly modifies the economic structure from within. The competitive character of capitalism is mainly determined by creative destruction and far less by the case of textbook competition, in which prices play such a dominant role. This process of creative destruction must be judged by its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
long-term results. Schumpeter has the highest possible opinion of the dynamic character and productive capability of capitalism. In weighing up the static optimal allocation of resources in case of perfect competition, and dynamic efficiency of monopolistic structures, in particular with regard to innovative activities, he has an outspoken preference for monopoly and oligopoly and a disdain for free competition. Schumpeter did not adhere to the theory that vanishing investment opportunities and a slowdown of technical change would lead to stagnation and in the end to a breakdown of capitalism (Metcalfe, 1998).

The rate of growth of production is not reduced because the technical possibilities are exhausted, but because capitalism suffers from a change in the behaviour of entrepreneurs. On the one hand, it is now easier than before to do things that lie outside familiar routine – innovation itself is being reduced to routine. Technological progress is increasingly becoming the business of teams of trained specialists who make technical change a predictable process. On the other hand, characteristics such as personality, willpower and a dynamic attitude count less in environments which have become accustomed to economic and social change.

Economic development thus becomes more and more impersonal and mechanical. The success of the capitalist mode of production makes capitalism itself redundant: capitalism undermines the social institutions which protect it. These institutions are the remnants of the feudal system and the existence of many small businesses and farmers. The disappearance of these social forms weakens the political position of the bourgeoisie. The elimination of the socioeconomic function of the entrepreneur, especially in large corporations where technical change is a matter of routine and management is bureaucratized, reinforced by the growing influence of the public sector, further undermines the bourgeoisie. Above all, however, capitalism produces an army of critical and frustrated intellectuals who by their negative attitude contribute to the decline of capitalism and help to establish an atmosphere in which private property and bourgeois values are daily subjected to heavy criticism (for example, in newspapers). In short, capitalism's economic success leads to its political failure (Schumpeter, 1942).

Schumpeter's thoughts on the effects of monopolistic and oligopolistic markets on technical change and on the influence of technology on the emergence of big business have given a great impetus to the study of the relationship between technical change and market structure. His sharp distinction between innovations and inventions has triggered off several reactions, and both theoretical analysis and empirical research have been directed to the question whether innovations are really as independent of inventions as Schumpeter supposed. A critical evaluation of Schumpeter's theory has led to a discussion of the nature of technical change, the time-lag between invention and application, the significance of patents and the diffusion process of technical change. Schumpeter's predictions about the decay of capitalism have not come about (Heertje, 1981). He seems to have underestimated the dynamic character of capitalism, the importance of which he himself emphasized so eloquently. He did not realize that a new generation of entrepreneurs might come to the fore, prepared to apply new technology and start a new wave of innovations. There is much more room for Schumpeterian entrepreneurial activity, especially on a small-scale basis, than Schumpeter foresaw.

It can be admitted that in large firms in particular the emergence of new technical knowledge is to a certain extent mechanized and that the decision-making process about its application, in the sphere of both production and marketing, is often hampered by bureaucratic features, threatening the static and dynamic efficiency of the firm and therefore the level of output. But these decisions still take place in a world of uncertainty and financial risk. These latter aspects of the decision-making process come more

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to the fore as the size of the firm becomes smaller. Within these firms, individuals who take final and major decisions and bear the responsibility of profits and losses do still exist. If, due to a lack of such individuals, Schumpeterian developments arise, the result will often be a new dynamic leadership. Many small firms still exist or come into being in order to gain a place in the market. Their managers take initiatives and often have to overcome resistance of consumers to achieve their market goals. It may be true that nowadays nearly everybody is confronted with change and new developments at a very high rate, but this does not imply that everybody accepts this state of affairs. So, on the whole, there are traces of a Schumpeterian development in Western economies, but mechanization and routinization of the entrepreneurial function are by no means the general picture. Economic life is still a melting pot of conflicting tendencies, ups and downs, major risks and minor certainties. In short, it is the dynamic, ever-changing scene for entrepreneurs who have to be innovative and sensitive to new opportunities. Those entrepreneurs who follow the Schumpeterian line will be punished by the market, that is, by competitors who are prepared to take risks and to bear losses.

On balance, it seems that innovation as a process of development, application and diffusion of new technical possibilities has not been reduced to routine. Nor have people become so reconciled to change that personal characteristics such as willpower and perseverance are no longer needed to break traditional patterns. However, even this kind of modification of Schumpeter's views does not mean that over some much longer run he may still not turn out to have been right. It therefore seems appropriate to look more carefully at the empirical evidence about the entrepreneurial function and about the acceptability of the level of output as the sole yardstick for economic performance. This may enable us to look into the future of capitalism, as well (Heertje, 2006).

Is the essence of the entrepreneurial function really the exploitation of new technical possibilities and of new opportunities in general? On the basis of Schumpeter's distinction between inventions and innovations, it is natural to identify entrepreneurial activity with innovative activity. But in the no man's land between invention and innovation there is a missing link; a link which according to Kirzner comprises three essential entrepreneurial components in human action: (*a*) the alertness to information; (*b*) the awareness of new existing opportunities, waiting to be noticed; and (*c*) the response to possibilities offered by the market system (Kirzner, 1986). Although in many cases those who are on the lookout for new opportunities are the same men and women who exploit them, Kirzner's refinement of Schumpeter's characterization of entrepreneurship is of the utmost importance – particularly if this view is combined with the idea of the market as an ongoing process of creative discovery. It may be argued that certain stages of the application and execution of new ideas can be routinized. However, the property of being alert to marketable applications of what already exists, but is currently overlooked by others, is still an individualistic characteristic, not one for being automatized. Even if we comprehend in the definition of the entrepreneurial function the application and implementation of new combinations, it is still true that first of all one has to be alert to what may be applied. In this sense entrepreneurship never was and never will be a matter of routine (Arena and Romani, 2002).

With regard to the level of output as a yardstick for economic performance, it may be observed that other aspects of welfare are taken into account as well. The operational meaning of welfare, in the broad sense of the level of the satisfaction of wants in so far as this depends on the allocation of scarce resources, comes more and more to the fore. The level of output is no longer the only thing that matters; the quality of growth, job satisfaction and other immaterial aspects of welfare are becoming more and more important. They are often the source for new methods of production and products, which give new

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
impetus to capitalism. As an Austrian, Schumpeter would certainly welcome the broadening of the economic dimension to the formal and subjective concept of welfare (Hennipman, 1995).

Furthermore, socialism as an alternative to capitalism is on a worldwide scale less and less attractive as a system able to take care of aspirations and heterogeneous preferences of individuals and of technical change. Since the beginning of the 1980s the economic recovery inspired the reinforcement of the market. Against this background a revival of Schumpeter's ideas in economic theory took place. While Keynes dominated economic thinking during a large part of the 20th century, Schumpeter is a major source of inspiration ever since. His concern with permanent and endogenous changes in economic life show up the limited significance of static equilibrium theory and paved the way for a neo-Austrian emphasis on the importance of the market mechanism as a vehicle for the discovery of new products and new methods of production, in short of the market as a dynamic institution. The dynamics of the market mechanism combined with the economics of research and development, new methods of production and products and of the diffusion of knowledge and of the application of new technology have led in modern economic theory to a Schumpeterian-inspired approach as part of endogenous growth theory (Aghion and Howitt, 1998). It comes as no surprise that in 1986 the International J.A. Schumpeter Society was founded in Augsburg, Germany. Since then every two years an international conference is organized.

Within the realm of social theory, Schumpeter's distinction between political and methodological individualism should also be mentioned. In particular, the concept of methodological individualism has proved to be very important for the analysis of social phenomena, outside the sphere of the market mechanism. As a method of analysis, methodological individualism prescribes starting from the economic behaviour of the individual in order to build a theory about the structure and working of the political process and about the behaviour of groups. In this sense, Schumpeter's thinking is the opposite of Marx's analysis in terms of the class struggle. The modern theory of public choice, in which the maximization of individual welfare of politicians and bureaucrats plays an essential role, in order to describe their social behaviour as part of the government, is a direct application of methodological individualism. Related to this development is the economic theory of democracy, of which Schumpeter is a forerunner. In his view, the democratic method is that institutional arrangement for arriving at political decision in which individuals acquire the power to decide by means of a competitive struggle for the people's vote. In other words, Schumpeter introduces the idea that democracy is a type of horizontal coordination in the public sector that can be compared to the role of the market mechanism in the private sector of the economy. The political process is regarded as a market process in which the voters are the demanders and politicians and bureaucrats are the suppliers.

This idea appears to be very fruitful in both theory and practice and contributes to Schumpeter's fame as a social and economic thinker of lasting significance.

Schumpeter's view of society is based on the integration of historical facts, philosophical considerations and sociological visions. While Marx predicted the breakdown of capitalism as an inevitable consequence of the objective inner structure of the system, determined in its development by technical change, Schumpeter, although also pointing out structural changes from within, leaves room for the role of individuals, who by their behaviour can turn the tide. His message on the march into socialism is not meant to be defeatist; it would only be so if all differences between individuals disappeared and, in particular, leadership vanished (Shionoya, 1995).

Schumpeter will always be referred to for his impressive contributions to the history of economic thought. His *Economic Doctrine and Method*, originally published in German in 1914, is an early

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
expression of his interest. His book *Ten Great Economists* provides further evidence, and his monumental work, the *History of Economic Analysis*, is a lasting culmination. It illustrates his detailed knowledge of the vast literature on economic theory since the days of the Greeks and Romans.

On the whole his discussion of the theoretical contributions of numerous economists is fair, generous and well-balanced. There are, however, a few notable exceptions. He ranks Cournot higher than Ricardo, and although we find both economists in modern economic theory, it seems fair to conclude that Ricardo's contributions to economics leave a broader scope and impact. Furthermore, he considers Walras the greatest economist of all time, greater than, for example, Marshall. His great appreciation for two mathematical economists is in contrast with his own non-mathematical treatment of economics, although he even became a founder and first president of the Econometric Society.

Reading Schumpeter, one realizes that his lasting significance stems from historical description and non-mathematical theoretical analysis. His inability to put his ideas about the development of economic life into a mathematical form does change our assessment of him. But whatever the final evaluation of Schumpeter may be, it cannot be denied that he gave new direction to the development of economic science by posing some entirely new questions. Schumpeter's preoccupation with the dynamics of economic life broke the spell of the static approach to economic problems.

Throughout his life Schumpeter was an *enfant terrible*, who was always ready to take extreme positions for the sake of argument, and often seized the chance to irritate people. But he was also a giant on whose shoulders many later scholars contributing to economic science stood. As a political economist he is no longer in the shadow of Keynes, but in the centre of the economic scene, both in the theoretical and empirical sense.

See Also

- creative destruction
- creative destruction (Schumpeterian conception)
- endogenous growth theory

Selected works

1912. *The Theory of Economic Development*. Leipzig: Duncker & Humblot. Trans. R. Opie. Cambridge, MA: Harvard University Press, 1934. Reprinted, New York: Oxford University Press, 1961.

1914. *Economic Doctrine and Method*. Trans. R. Aris. London: George Allen & Unwin, 1954.

1939. *Business Cycles*. 2 vols. New York: McGraw-Hill.

1942. *Capitalism, Socialism and Democracy*. New York: Harper & Brothers. London: George Allen & Unwin, 1943. 5th edn, London: George Allen & Unwin, 1976.

1951. *Ten Great Economists*. New York: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1954. *History of Economic Analysis*, ed. E. Boody. New York: Oxford University Press; London: George Allen & Unwin. Repr. with a new introduction by M. Perlman, London: Routledge, 1994.

1986. *Aufsätze zur Wirtschaftspolitik*. Tübingen: J.C.B. Mohr.

Bibliography

Aghion, P. and Howitt, P. 1998. *Endogenous Growth Theory*. London: MIT Press.

Allen, R.L. 1991. *Opening Doors: The Life and Work of Joseph Schumpeter*. London: Transaction.

Arena, R. and Romani, P.-M. 2002. Schumpeter on entrepreneurship. In *The Contribution of Joseph Schumpeter to Economics*, ed. R. Arena and C. Dangel-Hagnauer. London: Routledge.

Bös, D. and Stolper, W., eds. 1984. *Schumpeter oder Keynes*. Berlin: Springe-Verlag.

Frisch, H., ed. 1982. *Schumpeterian Economics*. New York: Praeger.

Harris, S.E., ed. 1951. *Schumpeter, Social Scientist*. Cambridge, MA: Harvard University Press.

Heertje, A., ed. 1981. *Schumpeter's Vision*. New York: Praeger.

Heertje, A. 2006. *Schumpeter on the Economics of Innovation and the Development of Capitalism*. Cheltenham: Edward Elgar.

Hennipman, P. 1995. *Welfare Economics and the Theory of Economic Policy*, ed. D.A. Walker and J. van den Doel. Aldershot: Edward Elgar.

Kirzner, I.M. 1986. *Discovery and the Capitalist Process*. Chicago: University of Chicago Press.

Metcalfe, J.S. 1998. *Evolutionary Economics and Creative Destruction*. London: Routledge.

Reisman, D. 2004. *Schumpeter's Market*. Cheltenham: Edward Elgar.

Shionoya, Y. 1995. *Schumpeter and the Idea of Social Science*. Cambridge: Cambridge University Press.

Stolper, W.F. 1994. *Joseph Alois Schumpeter: The Public Life of a Private Man*. Princeton: Princeton University Press.

Swedberg, R. 1991. *J.A. Schumpeter, His Life and Work*. Cambridge: Polity Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Heertje, Arnold. "Schumpeter, Joseph Alois (1883–1950)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000043> doi:10.1057/9780230226203.1489

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Schumpeterian growth and growth policy design

Philippe Aghion
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Schumpeterian growth theory features quality-improving innovations that displace previous technologies, and are motivated by prospective monopoly rents. It predicts, first, that a higher rate of growth should be associated with a higher rate of firm entry and exit, and that exit can enhance productivity growth; second, that some countries may converge to the technological frontier whereas others may diverge; third, that a given policy will have contrasting effects on sectors or countries at different distances from the frontier, and therefore growth policy must be adapted accordingly. In particular, entry and delicensing enhance growth more the closer sectors or countries are to the technological frontier.

Keywords

capital accumulation; Cobb–Douglas functions; competition policy; creative destruction; endogenous growth; entrepreneurs; entry and exit; growth policy design; horizontal innovation model; implementation innovations; industrial organization; information technology and economic growth; innovation; instrumental variables; monopoly rent; product variety; Ramsey model; research and development; Schumpeterian growth model; technology frontier; total factor productivity; Washington consensus

Article

Three main ideas underlie the Schumpeterian growth paradigm: (*a*) growth is primarily driven by technological innovations; (*b*) innovations are produced by entrepreneurs who seek monopoly rents from them; (*c*) new technologies drive out old technologies.
The Schumpeterian growth model (Aghion and Howitt, 1992; 1998) grew out of modern industrial organization theory (Tirole, 1988). It focuses on quality-improving innovations that render old products obsolete, and hence involves the force that Schumpeter called ‘creative destruction’. In this article we

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
argue that the Schumpeterian paradigm holds the best promise of delivering a systematic, integrated, and yet operational framework for analysing and developing context-dependent growth policies.
Schumpeterian theory begins with a production function specified at the industry level:

$$Y_{it} = A_{it}^{1-\alpha} K_{it}^{\alpha}, \; 0 < \alpha < 1 \quad (1)$$

where A_{it} is a productivity parameter attached to the most recent technology used in industry i at time t. In this equation, K_{it} represents the flow of a unique intermediate product used in this sector, each unit of which is produced one-for-one by capital. Aggregate output is just the sum of the industry-specific outputs Y_{it}. (Although the theory focuses on individual industries and explicitly analyses the microeconomics of industrial competition, the assumption that all industries are *ex ante* identical gives it a simple aggregate structure. In particular, it is easily shown that aggregate output depends on the aggregate capital stock K_t according to the Cobb–Douglas aggregate per-worker production function: $Y_t = A_t^{1-\alpha} K_t^{\alpha}$ where the labour-augmenting productivity factor A_t is just the unweighted sum of the sector-specific A_{it}'s. As in neoclassical theory, the economy's long-run growth rate is given by the growth rate of A_t, which here depends endogenously on the economy-wide rate of innovation.)
Each intermediate product is produced and sold exclusively by the most recent innovator. A successful innovator in sector i improves the technology parameter A_{it} and is thus able to displace the previous innovator as the incumbent intermediate monopolist in that sector, until displaced by the next innovator.
First implication:: faster growth generally implies a higher rate of firm turnover, because the process of creative destruction generates entry of new innovators and exit of former innovators.
There are two main inputs to innovation, namely, the private expenditures made by the prospective innovator, and the stock of innovations that have already been made by past innovators. The latter input constitutes the publicly available stock of knowledge to which current innovators are hoping to add. The theory is quite flexible in modelling the contribution of past innovations. It encompasses the case of an innovation that leapfrogs the best technology available before the innovation, resulting in a new technology parameter A_{it} in the innovating sector i, which is some multiple γ of its pre-existing value. And it also encompasses the case of an innovation that catches up to a global technology frontier $\bar{A}_t$ which we typically take to represent the stock of global technological knowledge available to innovators in all sectors of all countries. In the former case the country is making a leading-edge innovation that builds on and improves the leading-edge technology in its industry. In the latter case the innovation is just implementing technologies that have been developed elsewhere.
For example, consider a country in which in any sector leading-edge innovations take place at the frequency μ_n and implementation innovations take place at the frequency μ_m. Then the change in the economy's aggregate productivity parameter A_t will be:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$A_{t+1} - A_t = \mu_n(\gamma - 1)A_t + \mu_m(\bar{A}_t - A_t)$$
(2)

and hence the growth rate will be:

$$g_t = \frac{A_{t+1} - A_t}{A_t} = \mu_n(\gamma - 1) + \mu_m\left(a_t^{-1} - 1\right)$$
(3)

where:

$$a_t = A_t / \bar{A}_t$$

is an inverse measure of ‘distance to the frontier’. We then obtain a second important implication of the paradigm.

Second implication:: by taking into account the fact that innovations can interact with each other in different ways in countries or sectors at various distances from the frontier, Schumpeterian theory provides a framework in which to analyse how a country's growth performance will vary with its proximity to the technological frontier a_t, to what extent the country will tend to converge to that frontier, and what kinds of policy changes are needed to sustain convergence as the country approaches the frontier.

We could take as given the critical innovation frequencies μ_m and μ_n that determine a country's growth path as given, just as neoclassical theory often takes the critical saving rate *s* as given. However, Schumpeterian theory goes deeper by deriving these innovation frequencies endogenously from the profit-maximization problem facing a prospective innovator, just as the Ramsey model endogenizes *s* by deriving it from household utility maximization. This maximization problem and its solution will typically depend upon institutional characteristics of the economy such as property rights protection and the financial system, and also upon government policy.

Equation (3) incorporates Gerschenkron's (1962) ‘advantage of backwardness’, in the sense that the further the country is behind the global technology frontier (that is, the smaller a_t is) the faster it will grow, given the frequency of implementation innovations. As in Gerschenkron's analysis, the advantage arises from the fact that implementation innovations allow the country to make larger quality improvements the further it has fallen behind the frontier. As we shall see below, this is just one of the ways in which distance to the frontier can affect a country's growth performance.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In addition, as stressed by Acemoglu, Aghion and Zilibotti (2006) (AAZ), growth equations like (3) make it quite natural to capture Gerschenkron's idea of ‘appropriate institutions’. Suppose indeed that the institutions that favour implementation innovations (that is, that lead to firms emphasizing μ_m at the expense of μ_n) are not the same as those that favour leading-edge innovations (that is, that encourage firms to focus on μ_n); we then obtain the following:

Third implication:: far from the frontier a country or sector will maximize growth by favouring institutions that facilitate implementation; however, as it catches up with the technological frontier, to sustain a high growth rate the country will have to shift from implementation-enhancing institutions to innovation-enhancing institutions as the relative importance of leading-edge innovations for growth is also increasing.

As formally shown in AAZ, failure to operate such a shift can prevent a country from catching up with the frontier level of per capita GDP, and Sapir et al. (2003) argued that this failure largely explains why Europe stopped catching up with US per capita GDP from the mid-1970s. More specifically, suppose that the global frontier (the United States) grows at some rate $\bar{g}$. Then eq. (3) implies that in the long run a country that engages in implementation investments (with $\mu_m > 0$) will ultimately converge to the same growth rate as the world technology frontier. That is, the relative gap a_t that separates this economy from the technology frontier will converge asymptotically to the steady-state value:

$$\hat{a} = \frac{\mu_m}{\bar{g} + \mu_m - \mu_n(\gamma - 1)} \tag{4}$$

which is an increasing function of the domestic innovation rates and a decreasing function of the global productivity growth rate. An insufficient emphasis on innovation (μ_n) in Europe will reduce $\hat{a}$, that is, the long-run level of European per capita GDP compared with that of the United States.

The model can also explain why, since the mid-1990s, the EU has been growing at a lower rate than the United States. A plausible story, which comes out naturally from the above discussion, is that the European economy caught up technologically to the United States following the Second World War, but then its growth began to slow down before the gap with the United States had been closed, because its policies and institutions were not designed to optimize growth when close to the frontier. That by itself would have resulted in a growth rate that fell to that of the United States but no further. But then what happened was that the information technology revolution resulted in a revival of $\bar{g}$ in the late 1980s and early 1990s. Since Europe was as not as well placed as the United States to benefit from this technological revolution, the result was a reversal of Europe's approach to the frontier, which accords with the Schumpeterian steady-state condition (4); and the fact that Europe is not adjusting its institutions and policies in order to produce the growth-maximizing innovation policy acts as a force delaying growth convergence with the United States. (Endogenizing μ_m can also generate divergence in growth rates. For example, human capital constraints as in Howitt and Mayer-Foulkes, 2005, or credit

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
constraints as in Aghion, Howitt and Mayer-Foulkes, 2005, make the equilibrium value of μ_m increasing in *a*, which turns the growth equation (3) into a nonlinear equation. That μ_m be increasing in *a* follows in turn from the assumption that the cost of innovating is proportional to the frontier technology level that is put in place by the innovation – Ha and Howitt, 2007, provide empirical support for this proportionality assumption – whereas the firm's investment is constrained to be proportional to current local productivity. Then, countries very far from the frontier and/or with very low degrees of financial development or of human capital will tend to grow in the long run at a rate which is strictly lower than the frontier growth rate $\bar{g}$. However, our empirical analysis in this paper shows that this source of divergence does not apply to EU countries.)
In the next section we concentrate on a particular policy area, namely, entry and exit.

Entry and exit

Is it always growth-enhancing to liberalize entry and to facilitate exit? That exit could be growth-enhancing follows immediately from the fact that creative destruction is about better technologies or inputs replacing old and increasing obsolete technologies. Now, what about entry? Is it unambiguously good for innovation and growth by incumbent firms? As above, the answer may depend upon firms' distance from the technological frontier. In particular, suppose that firms can improve their technologies only in a gradual (or step-by-step) fashion, and that new potential entrants are endowed with the current frontier technology. Then, incumbent firms that are initially close to the frontier can match or even leapfrog a potential entrant's technology, and therefore can deter entry by innovating. In contrast, firms that are initially far below the frontier cannot prevent entry by innovating as they can never match an entrant's technology. What does this imply for how these firms will react to increased entry threat? The answer is simple: a greater entry threat will induce firms that are close to the frontier to invest more in innovation in order to protect their monopoly rents, whereas it will discourage firms far below the frontier from investing in innovation as such investment is less likely to be of any use the greater is the probability of entry.
In short, exit can be growth-enhancing and, regarding the effects of entry, the closer a firm or sector is to the frontier, the more positively (or the less negatively) it will react to increased entry threat. These predictions have been corroborated by a variety of empirical findings. First, Aghion et al. (2005a) investigate the effects of entry threat on total factor productivity (TFP) growth of UK manufacturing establishments, using panel data with more than 32,000 annual observations of firms in 166 different four-digit industries over the 1980–93 period. They estimate the equation:

$$Y_{ijt} = \alpha + \beta E_{jt} + \gamma E_{jt} \cdot D_{jt} + \eta_i + \tau_t + \varepsilon_{ijt} \quad (5)$$

where Y_{ijt} is TFP growth in firm *i*, industry *j*, year *t*, η and τ are fixed establishment and year effects, and E_{jt} is the industry entry rate, measured by the change in the share of UK industry employment in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
foreign-owned plants. In order to verify that this effect of entry on incumbent productivity growth is a result of increased incumbent innovation rather than technology spillover from, or copying of, the superior technologies brought in by the entrants, Aghion et al. (2005a) also estimate eq. (5) using a patent count rather than productivity growth as the dependent variable. They provide direct evidence that the escape competition is stronger for industries that are closer to the frontier. Specifically, the interaction coefficient γ is highly significantly negative in all estimations. A one-standard deviation increase in the entry variable above its sample mean would reduce the estimated number of patents by 10.8 per cent in an industry far from the frontier (at the 90th percentile of D_{jt}) and would increase the estimated number by 42.6 per cent in an industry near the frontier (at the tenth percentile). Thus it seems that the positive effect of entry threat on incumbent productivity growth in Europe is indeed much larger now than it was immediately after the Second World War, and that the relative neglect of entry implications of competition policy is having an increasingly detrimental effect on European productivity growth.

On exit and growth, in ongoing research Aghion, Antras and Prantl combine UK establishment-level panel data with the input–output table to estimate the effect on TFP growth arising from growth in high-quality input in upstream industries, and also from exit of obsolete input-producing firms in upstream industries. Specifically, we take a panel of 23,886 annual observations of more than 5,000 plants in 180 four-digit industries between 1987 and 1993, together with the 1984 UK input–output table, to estimate an equation of the form:

$$g_{ijt} = \alpha + \beta \cdot q_{jt-1} + \gamma \cdot x_{jt-1} + \delta \cdot Z_{ijt-1} + \eta_i + \varphi_j + \tau_t + \varepsilon_{ijt} \quad (6)$$

where g_{ijt} is the TFP growth rate of firm i in industry j, q_{jt-1} is a measure of upstream quality improvement, and x_{jt-1} is a measure of exit of obsolete upstream input-producing firms. Establishment, industry and year effects are included, along with the other controls in Z_{ijt-1}, including a measure of the plant's market share.

The result of this estimation is a significant positive effect of both upstream quality improvement and upstream input-production exit. These results are robust to taking potential endogeneity into account by applying an instrumental variable approach, using instruments similar to those of Aghion et al. (2005a) described above. The effects are particularly strong for plants that use more intermediate inputs, that is, plants with a share of intermediate product use above the sample median. Altogether, the results we find are consistent with the view that quality-improving innovation is an important source of growth. The results are, however, not consistent with the horizontal innovation model, in which there should be nothing special about the entry of foreign firms, and according to which the exit of upstream firms should if anything reduce growth by reducing the variety of inputs being used in the industry.

Conclusion and comparison with alternative endogenous growth models

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
What have we learned from our discussion so far? First, that Schumpeterian growth theory features quality-improving innovations that displace previous technologies, and that such innovations are motivated by the prospect of monopoly rents. Second, that, due to the natural conflict between new and old technologies, a higher rate of growth is likely to be associated with a higher rate of firm turnover (entry and exit).

In particular, exit can have a positive effect on productivity growth in downstream industries because it replaces less efficient input producers by more efficient ones. Third, that quality improvements can be generated, either by imitating current frontier technologies or by innovating upon previous local technologies, and that the relative importance of either type of innovation depends upon a sector's or a country's initial distance from the corresponding technological frontier. Fourth, that the same policy will tend to have contrasting effects on sectors or countries at different distances from the frontier, and that therefore growth policy must be adapted to the particular context of a sector or country. Fifth, that entry and delicensing have a more positive effect on growth in sectors or countries that are closer to the technological frontier, but have a less positive effect on sectors or countries that lie far below the frontier. This suggests that, although disregarding entry was of no great concern during the 30 years immediately after the Second World War, when Europe was still far behind the United States and catching up with it, nevertheless, now that Europe has come close to the world technology frontier this relative neglect of entry considerations is having an increasingly depressing effect on European growth. It also suggests a role for complementary policies aimed at reallocating resources and workers from laggard to more frontier sectors and activities in order to maximize the positive effects of competition and entry on productivity growth.

Finally, one may want to contrast the Schumpeterian growth paradigm with the two alternative models of endogeneous growth. The first version of endogenous growth theory was the so-called AK theory (see Frankel, 1962; Romer, 1986; Lucas, 1988), whereby knowledge accumulation is a serendipitous by-product of capital accumulation by the various firms in the economy. Here thrift and the resulting capital accumulation are the keys to growth, not novelty and innovation. The second model of endogenous growth theory is by Romer (1990), according to which aggregate productivity is a function of the degree of product variety. Innovation causes productivity growth in the product-variety paradigm by creating new, but not necessarily improved, varieties of products. The driving force of long-run growth in the product-variety paradigm is innovation, as in the Schumpeterian paradigm. In this case, however, innovations do not generate better intermediate products, just more of them. Also as in the Schumpeterian model, the equilibrium R&D investment and innovation rate results from a research arbitrage equation that equates the expected marginal payoff from engaging in R&D with the marginal opportunity cost of R&D. But the fact that there is just one kind of innovation, which always results in the same kind of new product, means that the product-variety model is limited in its ability to generate context-dependent growth, and is therefore of limited use for policymakers.

In particular, the theory makes it very difficult to talk about the notion of technology frontier and of a country's distance from the frontier. Consequently, it has little to say about how the kinds of policy appropriate for promoting growth in countries near the world's technology frontier may differ from those appropriate for technological laggards, and thus little to say by way of explaining why Asia is growing fast with policies that depart from the Washington consensus, or why Europe grew faster than the United States during the first three decades after the Second World War but not thereafter. In addition, nothing in this model implies an important role for exit and turnover of firms and workers; indeed increased exit

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in this model can do nothing but reduce the economy's GDP, by reducing the variety variable that uniquely determines aggregate productivity. As we just argued above, these latter implications of the product-variety model are inconsistent with an increasing number of recent studies demonstrating that labour and product market mobility are key elements of a growth-enhancing policy near the technological frontier.

See Also

- creative destruction
- endogenous growth theory
- information technology and the world economy
- Schumpeter, Joseph Alois

Bibliography

Acemoglu, D., Aghion, P. and Zilibotti, F. 2006. Distance to frontier, selection, and economic growth. *Journal of the European Economic Association* 4, 37–74.

Aghion, P., Bloom, N., Blundell, R., Griffith, R. and Howitt, P. 2005b. Competition and innovation: an inverted-U relationship. *Quarterly Journal of Economics* 120, 701–28.

Aghion, P., Blundell, R., Griffith, R., Howitt, P. and Prantl, S. 2004. Entry and productivity growth: evidence from micro-level panel data. *Journal of the European Economic Association* 2, 265–76.

Aghion, P., Blundell, R., Griffith, R., Howitt, P. and Prantl, S. 2005a. The effects of entry on incumbent innovation and productivity. Working Paper No. 12027. Cambridge, MA: NBER.

Aghion, P., Burgess, R., Redding, S. and Zilibotti, F. 2005. Entry liberalization and inequality in industrial performance. *Journal of the European Economic Association* 3, 291–302.

Aghion, P. and Howitt, P. 1992. A model of growth through creative destruction. *Econometrica* 60, 323–51.

Aghion, P. and Howitt, P. 1998. *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Aghion, P., Howitt, P. and Mayer-Foulkes, D. 2005. The effect of financial development on convergence: theory and evidence. *Quarterly Journal of Economics* 120, 173–222.

Frankel, M. 1962. The production function in allocation and growth: a synthesis. *American Economic Review* 52, 995–1022.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gerschenkron, A. 1962. *Economic Backwardness in Historical Perspective*. Cambridge, MA: Harvard University Press.

Ha, J. and Howitt, P. 2006. Accounting for trends in productivity and R&D: a Schumpeterian critique of semi-endogenous growth theory. *Journal of Money, Credit, and Banking* 39, 733–74.

Helpman, E. 1993. Innovation, imitation, and intellectual property rights. *Econometrica* 61, 1247–80.

Howitt, P. and Mayer-Foulkes, D. 2005. R&D, implementation and stagnation: A Schumpeterian theory of convergence clubs. *Journal of Money, Credit and Banking* 37, 147–77.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Nickell, S. 1996. Competition and corporate performance. *Journal of Political Economy* 104, 724–46.

Nicoletti, G. and Scarpetta, S. 2003. Regulation, productivity and growth. *Economic Policy* 36, 11–72.

Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Romer, P. 1990. Endogenous technical change. *Journal of Political Economy* 98, 71–102.

Sapir, A. et al. 2003. *An Agenda for a Growing Europe*. Oxford: Oxford University Press.

Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

How to cite this article

Aghion, Philippe. "Schumpeterian growth and growth policy design." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_G000176> doi:10.1057/9780230226203.1490

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

science, economics of

Arthur M. Diamond, Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The importance of the economics of science is substantially due to the importance of science as a driver of technology, and technology as a driver of productivity and growth. Believing that science matters, economists have attempted to understand the behaviour of scientists and the operation of scientific institutions. One goal is to see how far science can be understood as a market, and how far the market for science and scientists can be understood as efficient. When inefficiency is found, a related goal is to propose changes in resource levels or incentives, to increase the speed of scientific advance.

Keywords

academic tenure; Arrow, K.; Babbage, C.; Becker, G.; cobweb model; compensating differentials; crowding out; economic growth; Friedman, M.; Griliches, Z.; human capital; implicit contracts; life-cycle investment models; Matthew effect; Merton, R.; mobility; Mokyr, J.; National Science Foundation; new growth theory; Peirce, C.; Polanyi. M.; productivity; public goods; rent seeking; research and development; research productivity; Romer, P.; Schumpeter, J.; Science funding; science, economics of; Smith, A.; social capital; Stigler, G.; Technology

Article

The economics of science aims to understand the impact of science on the advance of technology, to explain the behaviour of scientists, and to understand the efficiency or inefficiency of scientific institutions.

The first economics of science may have been Adam Smith's idealistic, but sadly untrue, discussion in the *Theory of Moral Sentiments* (1759, p. 124) of Newton having been motivated purely by curiosity rather than a desire to achieve fame and fortune. If Smith's account was the beginning of a positive economics of science, then Charles Babbage's argument (1830) for the reform of British scientific institutions may count as one of the earliest instances of a mainly normative economics of science. Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
usually mentioned as an early founder of the normative economics of science is the American philosopher C.S. Peirce (1839–1914), who advocated the application of economic tools of analysis to decide which scientific projects to adopt (see Wible, 1998).

The ‘modern’ economics of science grew out of three main topics The first topic addressed how the advance of science contributed to the advance of technology, and hence productivity and growth. The second topic, which overlaps with concerns in the history and philosophy of science, addressed how science advances. A third topic is the empirical data collection and econometric analysis of the supply, demand, compensation and productivity of scientists.

Diamond (1996) and Stephan (1996) both provide broad surveys of the economics of science, with Diamond perhaps devoting more space to interdisciplinary and policy issues. Special attention has been paid to the contributions to the economics of science of three of the field's founders: Mansfield, Griliches and Stigler (Diamond, 2003; 2004; 2005). Several of the more important papers in the economics of science through 1998 are included in the two-volume collection edited by Stephan and Audretsch (2000).

In what follows, we first consider the literature that most makes the case for why the economics of science should be a priority for our attention: the literature on science as a contributor to economic productivity and growth. We proceed to briefly summarize some economic discussion of some of the ‘deep’ issues in the economics of science, which sometimes overlap with issues in philosophy of science, such as the objectives of scientists (truth, fame, fortune?), and the constraints that are most relevant to their choices about which projects to pursue and which theories to adopt. Next we look at some of the studies that have attempted to model and measure a variety of aspects of the market for science and scientists. Finally, we give examples of some of the studies in normative economics of science that argue for changes in funding or for institutional reform.

Impact of science on technology, productivity, and growth

The importance of technology as a driver of economic growth and well-being has been appreciated since Adam Smith's *Wealth of Nations* (1776), and emphasized most notably by Schumpeter (1942). If technology is the main driver of economic growth, the next question is: what is the main driver of technology?

Rosenberg (1982) made the credible point that most economists, for most of the history of the profession, had viewed the process by which new technologies are developed and adopted as a ‘black box’. In the years since, partly led by Rosenberg himself, economists have increasingly attempted to say more about what goes on inside the box, especially concerning the role of science in advancing technology.

Several economic historians have examined the role of science in the advance of technology and economic growth over the broad sweep of history. Mokyr (1990; 2002), Rosenberg (Mowery and Rosenberg, 1989; Rosenberg and Birdzell, 1986), and Landes (1998), broadly agree that the advance of science is a necessary but not sufficient condition for substantial and rapid advance in technology and economic growth.

Nelson (1959) catalogued many examples of how science had contributed to the advance of technology. More recently, Rosenberg (Mowery and Rosenberg, 1989, pp. 11–14) has claimed that the distinction

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
between science and technology is often hard to make, providing several examples of how scientific advances have resulted from the pursuit of ‘practical’ results. Although most economists adopt Nelson's view that mainly new science enables the advance of new technology, it is not hard to find examples where the advance of science was enabled by new instruments provided by advanced technology (Ackermann, 1985).

Mokyr (1990), in his broad economic history of the advance of technology over the ages, generally finds the advance to be slow and fitful until the industrial revolution. Until the mid-1800s, the relationship between science and technology was loose (Mansfield, 1968; Mokyr, 1990, pp. 167–70). Those who advanced science and technology shared an attitude of optimism about the prospects of understanding and controlling nature (Landes, 1969). But, beginning in the mid-19th century, and especially with the development of commercial laboratories toward the end of the 19th century, the relationship between science and technology became closer, with advances in science more often and more clearly being a necessary condition for technological advances.

Beginning with Nelson's taxonomic paper (1959), evidence for this latter claim has been provided by economists in a variety of forms. Griliches's main contribution, in a pair of papers (1957; 1958), was to measure the return to scientific research on hybrid corn and to measure and explain varying rates of adoption (see Diamond, 2004). Surveys of research managers by Nelson (1986) and by Mansfield (1991; 1992) provided evidence that science is sometimes important for technical change, although the importance varies considerably with the industry and with the sub-field of science.

The development by Romer (1986; 1990) and others of the ‘new growth theory’ attracted further attention to science as a driver of technology, because such models include a more prominent role for knowledge (‘recipes’) than earlier models. Such models implied the possibility of increasing returns to investments in knowledge if various spillover effects were large enough. Stimulated partly by such models and partly by independent research by economists such as Griliches (see Diamond, 2004), considerable empirical work has been undertaken measuring the spillover effects of scientific knowledge; for example, Jaffe (1989), Adams (1990) and Jaffe, Trajtenberg and Henderson (1993).

Deep understanding of science

Stigler and Becker (1977) have argued that everyone has the same utility function. This contrasts with Adam Smith's suggestion in his *Theory of Moral Sentiments* (1759, p. 124) that scientists, or at least the best scientists, were more purely motivated by curiosity. Gordon Tullock (1966, pp. 34–6), and Kenneth Arrow (2004) suggest that scientists have a range of motives, from those who fit the Smith ideal to those motivated by fame and fortune (Levy, 1988). However, in their research on the behaviour of scientists, economists usually follow Stigler and Becker in assuming that scientists mainly value income and prestige. Scientists valuing both income and prestige might explain Stern's finding (2004) that industrial scientists are willing to give up some income in exchange for greater ability to publish their results.

The process of theory choice among scientists has been explained using economic tools (Diamond, 1988b; Hull, 1988; Goldman and Shaked, 1991). The explanations have been criticized by Hands (1997). Brock and Durlauf (1999) build on the work of Kitcher (1993) in their construction of a dynamic model of scientists’ adoption of new theories. A key assumption of their set-up is that one source of a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
scientist's utility is the ‘conformity’ of a scientist's views with those of other scientists. The model allows the possibility that science is progressive, even if social considerations have some weight in scientists’ utility functions, partially answering many of those in the social studies of science field who believe that the admission of social considerations undermines the special cognitive status of science.
Stigler authored several papers that presented evidence on hypotheses about the determinants of successful science. He provided evidence and arguments on questions such as: does a scientist's biography help us understand the scientist's contributions (1976), and how efficiently is error weeded out in science (1978)? Stigler generally framed his studies as seeking just to understand, not to reform, although the results sometimes stimulated thoughts of reform in others. A fuller account of Stigler's contributions to the economics of science appears in Diamond (2005).

The market for science and scientists

Michael Polanyi (1962) optimistically portrayed science as an efficient marketplace of ideas. Much research in the economics of science in the last few decades shares Polanyi's research programme of explaining the behaviour of scientists and scientific institutions on the basis of rational optimization within an efficient marketplace.
Many studies in the economics of science fall within the domain of labour economics, and assume that scientists are rational maximizers of income, and sometimes of prestige. Within labour economics, a significant theoretical and empirical literature has developed that examines the economics of higher education. Since many scientists, and especially most scientists who are credited with major scientific discoveries, have been associated with universities, this literature is relevant to the economics of science, even when the examples or data are not drawn explicitly from science.
Some of the earliest economics of science studies collected data to analyse the supply of and demand for scientists. Early examples of this genre were Blank and Stigler (1957) and Arrow and Capron (1959). Richard Freeman's ‘cobweb’ model (1975) of the labour market for physicists had an unstable equilibrium because students’ occupational choices in the model are based on systematically biased forecasts of the future demand for physicists. Siow later (1984) showed that professional labour markets are better characterized by assuming the students’ forecasts are based on rational expectations.
Human capital theory and measurement have been used to estimate earnings functions for scientists, including as independent variables measures of productivity, such as articles produced, citations received, and teaching evaluations. Many of the studies also include one or more variables intended to measure hypotheses of discrimination, such as gender and race variables (for example, McDowell, 1982). Yet other studies include variables intended to measure what has recently been called ‘social capital’ and has previously been identified with Robert K. Merton's (1968) ‘Matthew effect’ (the rich get richer) or with ‘old-boy’ networks.
One goal of many of the earnings regression studies has been to learn how much of the variation in academic salaries can be explained on the basis of variation in measures of academic productivity. Lovell's (1973) paper was one of the first to include measures of research productivity in an academic earnings regression. Early studies tended to focus on number of articles published as the measure of research productivity. A pair of papers by Stigler and Friedland (1975; 1979) helped establish the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
credibility of citations as a measure of academic productivity. One of the first to include citations in an earnings function as a measure of productivity may have been Laband (1986). Subsequent studies using the citation measure include Hamermesh, Johnson and Weisbrod (1982), Diamond (1986b), Sauer (1988), and Kenny and Studley (1995). A review of the literature on bibliometric measures of productivity can be found in Diamond (2000).

The simplest models of compensation assume that workers are paid the value of their current productive output (their 'marginal revenue product'). To account for observed anomalies with this hypothesis, especially in professional labour markets, a literature has developed supposing that there are long-term implicit labour contracts. For example, universities may provide scientists with insurance for variability of research output by paying the scientist more than the value of their output in low-output years and less than the value of output in high-output years. If scientists are uncertain at the beginning of their careers whether they will become high- or low-productivity scientists, they may also demand insurance against the possibility of their being low-productivity scientists (S. Freeman, 1977). This might explain the observed greater variability in measures of scientists' productivity than in measures of their salaries. An alternative explanation is to make use of a compensating differentials argument (Frank, 1984). The assumption is that scientists receive utility from being paid more than other scientists. So the top scientist would be paid less than the value of her productivity, because she is receiving a compensating differential in the form of being at the top of the pecking order.

Implicit contract models have also been developed to try to explain important scientific labour market practices, such as academic tenure. For example, Carmichael (1988) argues that tenure is an institutional device to reduce the costs to incumbent faculty of correctly identifying promising new faculty, while Waldman (1990) claims that faculty value tenure because it serves as a signal to outside institutions of the faculty members' quality, and hence increases outside higher salary offers. Siow (1998) claims that specializing is risky, since sub-fields of specialization may suddenly become obsolete; so without tenure as a form of insurance, faculty would under-specialize.

Implicit contract models are often clever and sometimes plausible. But as clever, plausible models multiply that explain the same stylized facts (for example, the existence of academic tenure), the credibility of the exercise may suffer. It may also be worth mentioning that, ceteris paribus, economists will be more popular with their peers if they create models justifying tenure, and other academic institutions, than if they create models showing tenure is inefficient.

Other mainly empirical studies have examined the mobility of academic scientists between university positions (Rees, 1993), and the mobility of industrial scientists between technical and managerial jobs (Biddle and Roberts, 1994). Another extensive, mainly empirical, literature makes use of standard theory on the optimal allocation of time over the life cycle to motivate analysis of scientific productivity over the life cycle. Life-cycle investment models (for example, Diamond, 1984) often suggest that it makes sense to invest in human capital early in the life cycle. These models often imply concave age–productivity profiles. Empirical evidence confirms this generalization (Diamond, 1986a; Stephan and Levin, 1992), but with very different peak productivity ages for different fields of science. Age-related differences in the rate of acceptance of new theories have also been examined (Hull, Tessner and Diamond, 1978; Diamond, 1980; 1988a; 1988b; Levin, Stephan and Walker, 1995).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Efficiency and reform of scientific institutions

In the previous section we discussed research that for the most part argues that the institutions for rewarding and allocating resources to scientists can be explained as efficient aspects of a well-functioning market of ideas. Bartley (1990) has argued that while a ‘marketplace of ideas’ is an appropriate goal and standard, it is an inaccurate description of the current institutions of science. In one of his headings (1990, p. ix), he claims that our current institutions for academic science are ‘where consumers do not buy, producers do not sell, and owners do not control’.

In an early study, Stigler (1965) provided evidence that, when economics underwent a transition from a science done by amateurs to one done by professionals, the discipline became much more theoretical and mathematical, and much less applied and policy-oriented. Stigler's friend and colleague, Milton Friedman, argued (1981) that the funding of the National Science Foundation (NSF) had had a similar effect, and argued further that this effect had slowed the advance of knowledge. The debate was renewed 13 years later (Friedman, 1994; Griliches, 1994). Edward Lazear (1997) has developed a model implying more modest advice for the NSF: the foundation should give fewer but larger awards.

Other economists have studied the funding of science. Arrow (1962), Johnson (1972) and others have argued that science is a public good that will be under-provided by the private sector.

Dasgupta and David (1994) accept the public goods argument of the ‘old’ economics of science of Arrow (1962), but want to add to it findings of some sociologists on the secrecy that sometimes results from the competition for priority, in order to develop a ‘new’ economics of science. Their ‘new’ economics of science argues for greater government funding of science, accompanied by increased incentives for scientists to share their findings sooner with other scientists and with those seeking to apply the findings to new technologies.

Romer (2001) argues that, if roughly half a million more scientists and engineers were supplied and appropriately deployed, the US economy could sustain a half a per cent greater rate of growth in GDP. He suggests that major changes would be required in academic institutions and government policies to achieve this goal, but he believes the resulting implications for the economy of success would be ‘staggering’ (2001, p. 227).

Kealey (1996) and Martino (1992) explicitly dispute the traditional public goods argument for government support of science on the grounds that private industry often has both the incentives and the ability to do substantial high-quality scientific research. Hanson (1995) supports an alternative private form of science funding when he argues that greater scientific innovation would occur if more of the funding for science came from a betting market, where those who predict accurately the ultimate outcome of currently debated scientific questions receive more resources.

Although not opposing the science-as-public-good theory as strongly as Kealey and Martino, Rosenberg (1990) has emphasized the incentives that private firms have to invest in science. He examined firms that hired Ph.D. scientists and that allowed the scientists considerable leeway in the allocation of their time and in the publication of their results. He argues that this was in the firm's interest because of the value of such scientists as a resource in keeping up with and explaining scientific advances relevant to the firm's product development efforts. Besides Rosenberg's paper, there is a considerable literature measuring returns to firm investment in R&D. Some of these studies might be considered part of the economics of science to the extent that they study ‘basic research’, a label that is sometimes used

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
interchangeably with ‘science’. Examples of this literature are surveyed in Audretsch et al. (2002).

Several scholars have attempted to measure the extent to which public expenditures on science add to the total funding of science, or the extent to which they simply crowd out private funding on science. Diamond (1999), for example, using highly aggregated time-series measures of government and industry investment in science, found no evidence of crowding out. An evaluative survey of this literature has been published by David, Hall and Toole (2000).

Some economists have explained the behaviour of some scientists and the structure of some scientific institutions in terms of rent-seeking behaviour. Rent seeking is a zero-sum process in which an agent invests resources to obtain an uncompensated transfer from another agent (Tullock, 1967). In one example, McKenzie (1979) suggested that there is a fixed fund for department salaries, and that department members can increase their share of the fund either by being more productive themselves or by sabotaging the productivity of others, perhaps, for example, by the calling of unnecessary meetings. Other rent-seeking accounts of academic institutions have been provided by Brennan and Tollison (1980) and Grubel and Boland (1986).

We mentioned in the previous section economic models of academic tenure that attempt to explain the institution as an efficient response to features of the academic labour market. Others (for example, Rogge and Goodrich, 1973) have followed Alchian (1959) in presenting a basically rent-seeking account of tenure as an inefficient institution that exists because it is in the interests of a sufficiently powerful special interest group.

In an account highly complementary to the rent-seeking hypothesis, Goolsbee (1998) has studied data on federal funding of science and found that it largely results in windfalls for scientists. Goolsbee's results call into question the extent to which federal funding actually increases the amount of science produced.

See Also

- Arrow, Kenneth Joseph
- crowding out
- Griliches, Zvi
- implicit contracts
- public goods
- rent seeking
- Stigler, George Joseph
- technology

Bibliography

Ackermann, R. 1985. *Data, Instruments and Theory*. Princeton, NJ: Princeton University Press.

Adams, J. 1990. Fundamental stocks of knowledge and productivity growth. *Journal of Political Economy* 98, 673–702.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Alchian, A. 1959. Private property and the relative cost of tenure. In *The Public Stake in Union Power*, ed. P. Bradley. Charlottesville, VA: University Press of Virginia.

Arrow, K. 1962. Economic welfare and the allocation of resources for inventions. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, ed. R. Nelson. Princeton: Princeton University Press.

Arrow, K. 2004. Foreword. In *Reflections of Eminent Economists*, ed. M. Szenberg, and L. Ramrattan. Cheltenham, UK: Edward Elgar.

Arrow, K. and Capron, W. 1959. Dynamic shortages and price rises: the engineer–scientist case. *Quarterly Journal of Economics* 73, 292–308.

Audretsch, D., Bozeman, B., Combs, K., Feldman, M., Link, A., Siegel, D., Stephan, P., Tassey, G. and Wessner, C. 2002. The economics of science and technology. *Journal of Technology Transfer* 27, 155–203.

Babbage, C. 1830. *Reflections on the Decline of Science in England, and on Some of Its Causes*. London: B. Fellowes.

Bartley, W., III 1990. *Unfathomed Knowledge, Unmeasured Wealth: On Universities and the Wealth of Nations*. LaSalle, IL: Open Court.

Biddle, J. and Roberts, K. 1994. Private sector scientists and engineers and the transition to management. *Journal of Human Resources* 29, 82–107.

Blank, D. and Stigler, G. 1957. *The Demand and Supply of Scientific Personnel*. New York: National Bureau of Economic Research.

Brennan, H. and Tollison, R. 1980. Rent seeking in academia. In *Toward a Theory of the Rent-Seeking Society*, ed. J. Buchanan, R. Tollison and G. Tullock. College Station, TX: Texas A & M University Press.

Brock, W. and Durlauf, S. 1999. A formal model of theory choice in science. *Economic Theory* 14, 113–30.

Carmichael, H. 1988. Incentives in academics: why is there tenure? *Journal of Political Economy* 96, 453–72.

Dasgupta, P. and David, P.A. 1994. Toward a new economics of science. *Research Policy* 23, 487–521.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

David, P., Hall, B. and Toole, A. 2000. Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy* 29, 497–529.

Diamond, A., Jr. 1980. Age and the acceptance of cliometrics. *Journal of Economic History* 40, 838–41.

Diamond, A., Jr. 1984. An economic model of the life-cycle research productivity of scientists. *Scientometrics* 6, 189–96.

Diamond, A., Jr. 1986a. The life-cycle research productivity of mathematicians and scientists. *Journal of Gerontology* 41, 520–25.

Diamond, A., Jr. 1986b. What is a citation worth? *Journal of Human Resources* 21, 200–15.

Diamond, A., Jr. 1988a. The polywater episode and the appraisal of theories. In *Scrutinizing Science: Empirical Studies of Scientific Change*, ed. A. Donovan, L. Laudan and R. Laudan. Dordrecht: Kluwer Academic Publishers.

Diamond, A., Jr. 1988b. Science as a rational enterprise. *Theory and Decision* 24, 147–67.

Diamond, A., Jr. 1996. The economics of science. *Knowledge and Policy* 9(2, 3), 6–49.

Diamond, A., Jr. 1999. Does federal funding ‘crowd in’ private funding of science? *Contemporary Economic Policy* 17, 423–31.

Diamond, A. Jr. 2000. The complementarity of scientometrics and economics. In *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*, ed. B. Cronin and H. Adkins. Medford, NJ: Information Today, Inc.

Diamond, A., Jr. 2003. Edwin Mansfield's contributions to the economics of technology. *Research Policy* 32, 1607–17.

Diamond, A., Jr. 2004. Zvi Griliches's contributions to the economics of technology and growth. *Economics of Innovation and New Technology* 13, 365–97.

Diamond, A., Jr. 2005. Measurement, incentives, and constraints in Stigler's economics of science. *European Journal of the History of Economic Thought* 12, 637–63.

Frank, R. 1984. Are workers paid their marginal products? *American Economic Review* 74, 549–71.

Freeman, R. 1975. Supply and salary adjustments to the changing science manpower market: Physics, 1948–1973. *American Economic Review* 65, 27–39.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Freeman, S. 1977. Wage trends as performance displays productive potential: a model and application to academic early retirement. *Bell Journal of Economics* 8, 419–43.

Friedman, M. 1981. An open letter on grants. *Newsweek*, 18 May, 99.

Friedman, M. 1994. National science foundation grants for economics: correspondence. *Journal of Economic Perspectives* 8(1), 199–200.

Goldman, A. and Shaked, M. 1991. An economic model of scientific activity and truth acquisition. *Philosophical Studies* 63, 31–55.

Goolsbee, A. 1998. Does government R&D policy mainly benefit scientists and engineers? *American Economic Review* 88, 298–302.

Griliches, Z. 1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25, 501–22.

Griliches, Z. 1958. Research cost and social returns: hybrid corn and related innovations. *Journal of Political Economy* 66, 419–31.

Griliches, Z. 1994. National science foundation grants for economics: response. *Journal of Economic Perspectives* 8(1), 203–5.

Grubel, H. and Boland, L. 1986. On the efficient use of mathematics in economics: some theory, facts and results of an opinion survey. *Kyklos* 39, 419–42.

Hamermesh, D., Johnson, G. and Weisbrod, B. 1982. Scholarship, citations and salaries: economic rewards in economics. *Southern Economic Journal* 49, 472–81.

Hands, D. 1997. Caveat emptor: economics and contemporary philosophy of science. *Philosophy of Science* 64(4), S107–S116.

Hanson, R. 1995. Could gambling save science? Encouraging an honest consensus. *Social Epistemology* 9, 3–33.

Hull, D. 1988. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago: University of Chicago Press.

Hull, D., Tessner, P. and Diamond, A., Jr. 1978. Planck's principle: do younger scientists accept new scientific ideas with greater alacrity than older scientists?. *Science* 202, 717–23.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Jaffe, A. 1989. Real effects of academic research. *American Economic Review* 79, 957–69.

Jaffe, A., Trajtenberg, M. and Henderson, R. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108, 577–98.

Johnson, H. 1972. Some economic aspects of science. *Minerva* 10, 10–18.

Kealey, T. 1996. *The Economic Laws of Scientific Research*. New York: St Martin's Press.

Kenny, L. and Studley, R. 1995. Economists' salaries and lifetime productivity. *Southern Economic Journal* 62, 382–93.

Kitcher, P. 1993. *The Advancement of Science*. Oxford: Oxford University Press.

Laband, D. 1986. Article popularity. *Economic Inquiry* 24, 173–80.

Landes, D. 1969. *The Unbound Prometheus: Technological Change 1750 to the Present*. Cambridge: Cambridge University Press.

Landes, D. 1998. *The Wealth and Poverty of Nations*. New York: W.W. Norton.

Lazear, E. 1997. Incentives in basic research. *Journal of Labor Economics* 15, S167–S197.

Levin, S., Stephan, P. and Walker, M. 1995. Planck's principle revisited: a note. *Social Studies of Science* 25, 275–83.

Levy, D. 1988. The market for fame and fortune. *History of Political Economy* 20, 615–25.

Lovell, M. 1973. The production of economic literature: an interpretation. *Journal of Economic Literature* 11, 27–55.

Mansfield, E. 1968. *The Economics of Technological Change*. New York: W.W. Norton.

Mansfield, E. 1991. Academic research and industrial innovation. *Research Policy* 20, 1–12.

Mansfield, E. 1992. Academic research and industrial innovation: a further note. *Research Policy* 21, 295–96.

Martino, J. 1992. *Science Funding*. New Brunswick, NJ: Transaction Publishers.

McDowell, J. 1982. Obsolescence of knowledge and career publication profiles: some evidence of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
differences among fields in costs of interrupted careers. *American Economic Review* 72, 752–68.

McKenzie, R. 1979. The economic basis of departmental discord in academe. *Social Science Quarterly* 59, 653–64.

Merton, R. 1968. The Matthew effect in science. *Science* 159(3810), 56–63.

Mokyr, J. 1990. *The Lever of Riches: Technological Creativity and Economic Progress*. Oxford: Oxford University Press.

Mokyr, J. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.

Mowery, D. and Rosenberg, N. 1989. *Technology and the Pursuit of Economic Growth*. Cambridge: Cambridge University Press.

Nelson, R. 1959. The simple economics of basic scientific research. *Journal of Political Economy* 67, 297–306.

Nelson, R. 1986. Institutions supporting technical advance in industry. *American Economic Review, Papers and Proceedings* 76, 186–9.

Polanyi, M. 1962. The republic of science: its political and economic theory. *Minerva* 1, 54–73.

Rees, A. 1993. The salaries of Ph.D.'s in academe and elsewhere. *Journal of Economic Perspectives* 7 (1), 151–8.

Rogge, B. and Goodrich, P. 1973. Education in a free society. In *Education in a Free Society*, ed. A. Burleigh. Indianapolis: Liberty Fund.

Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Romer, P. 1990. Endogenous technological change. *Journal of Political Economy* 98(5), S71–102.

Romer, P. 2001. Should the government subsidize supply or demand in the market for scientists and engineers? In *Innovation Policy and the Economy*, vol. 1, ed. A. Jaffe, J. Lerner and S. Stern. Cambridge, MA: MIT Press.

Rosenberg, N. 1982. *Inside the Black Box: Technology and Economics*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Rosenberg, N. 1990. Why do firms do basic research (with their own money)? *Research Policy* 19, 165–74.

Rosenberg, N. and Birdzell, L., Jr. 1986. *How the West Grew Rich*. New York: Basic Books.

Sauer, R. 1988. Estimates of the returns to quality and coauthorship in economic academia. *Journal of Political Economy* 96, 855–66.

Schumpeter, J. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper and Row.

Siow, A. 1984. Occupational choice under uncertainty. *Econometrica* 52, 631–45.

Siow, A. 1998. Tenure and other unusual personnel practices in academia. *Journal of Law, Economics and Organization* 14, 152–73.

Smith, A. 1759. *The Theory of Moral Sentiments*. Oxford: Clarendon Press, 1976.

Smith, A. 1776. *An Inquiry Into the Nature and Causes of the Wealth of Nations*. Indianapolis: Liberty Press, 1976.

Stephan, P. 1996. The economics of science. *Journal of Economic Literature* 34, 1199–235.

Stephan, P. and Audretsch, D., eds. 2000. *The Economics of Science and Innovation*. 2 vols. Cheltenham: Edward Elgar.

Stephan, P. and Levin, S. 1992. *Striking the Mother Lode in Science: The Importance of Age, Place, and Time*. Oxford: Oxford University Press.

Stern, S. 2004. Do scientists pay to be scientists? *Management Science* 50, 835–53.

Stigler, G. 1965. Statistical studies in the history of economic thought. In *Essays in the History of Economics*, Chicago: The University of Chicago Press.

Stigler, G. 1976. The scientific uses of scientific biography, with special reference to J.S. Mill. In *Papers of the Centenary Conference: James and John Stuart Mill*, ed. J. Robson and M. Laine. Toronto: University of Toronto Press.

Stigler, G. 1978. The literature of economics: the case of the kinked oligopoly demand curve. *Economic Inquiry* 16(2), 185–204.

Stigler, G. and Becker, G. 1977. De gustibus non est disputandum. *American Economic Review* 67(2),

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

76–90.

Stigler, G. and Friedland, C. 1975. The citation practices of doctorates in economics. *Journal of Political Economy* 83, 477–507.

Stigler, G. and Friedland, C. 1979. The pattern of citation practices in economics. *History of Political Economy* 11, 1–20.

Tullock, G. 1966. *The Organization of Inquiry*. Durham, NC: Duke University Press.

Tullock, G. 1967. The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal* 5, 224–32.

Waldman, M. 1990. Up-or-out contracts: a signaling perspective. *Journal of Labor Economics* 8, 230–50.

Wible, J. 1998. *The Economics of Science: Methodology and Epistemology as If Economics Really Mattered*. London: Routledge.

How to cite this article

Diamond, Arthur M., Jr. "science, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000222> doi:10.1057/9780230226203.1491

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

scientific realism and ontology

Uskali Mäki

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article defines the concept of realism and explores its implications for ontology, defending the idea of ontological realism and its relevance to economics, while rejecting the idea of some special ‘realist ontology’ that informs us about the ways of the real world. The main focus is on what it is for the world (its constituents, structure, and ways of functioning) to exist. Economics-relevant scientific realism suggests that much of the social world is characterized by non-causal science-independence. Implications of this are outlined for causation, social construction, economics-dependence of the economy, modelling, and truth in economics.

Keywords

causal realism; causality in economics; Granger-causality; Hume, D.; methodological individualism; microfoundations; mind-independence; models; neuroeconomics; normal science; ontological realism; ontology; rhetoric of economics; science-independence; scientific realism; social constructivism

Article

Economists customarily talk about the ‘realism’ of economic models and of their assumptions and make descriptive and prescriptive judgements about them: this model has more realism in it than that model, the realism of assumptions does not matter, and so on. This is not the way philosophers mostly use the term ‘realism’; thus there is a major terminological discontinuity between the two disciplines. The following remarks organize and critically elaborate some of the philosophical usages of the term and show some of the ways in which they relate to economists' concerns. In the philosophy of science, scientific realism is the mainstream position – or rather a heterogeneous collection of positions – that includes ideas about the nature of scientific theories and how they are related to the real world and about the goals and achievements of scientific inquiry. However, most of what philosophers have contributed around these ideas is not designed to deal with the peculiarities of economics, so some important

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
adjustments are needed to make scientific realism an interesting position for economists.
Economists and their critics, as well as the consumers of economic research, are often intrigued by two kinds of issue that connect with realism: Is this model about something real, something that really exists (or instead about some imaginary fiction or social construct perhaps)? Is this model true (partly true, approximately true) about something real (instead of just being useful or convenient or persuasive)? These are the general questions discussed here – not by answering them, but by way of clarifying the conceptual prerequisites of trying to answer them, and in particular by way of examining what it is to be a realist about them. Misuses of 'realism' are abundant enough to warrant this exercise. The focus will be on ontology, and on how realism relates to it, so we will be mainly talking about ontological realism.

What exists?

Ontological realism is a philosophical thesis that deals with two questions: What exists? What is existence? Consider the first question. Slightly different ways of putting it include:

- What is there in the world?
- What is the furniture of the world?
- What is the world made of?
- What is its structure?
- What is the case?
- What is the way the world is?
- What is the way the world works?

To such questions one expects answers of the form, 'X exists' or 'Z is the case' or 'WWW is the way the world works'. But isn't it the task of science to provide answers of this form to such questions? What is the role of distinct ontological reflection? The quick response is in two parts: the answers provided by ordinary scientific practice ('normal science') are often implicit and only presupposed, while ontological reflection seeks to make them explicit; and the explicit answers that daily scientific practice supplies are mostly more specific and concrete that those offered by focused ontological scrutiny. There is an overlap and continuity between 'normal' science and ontological reflection regarding the kinds of answers they seek to offer to those questions. Consistency between the two is desirable, even though occasional (and often fruitful) conflicts arise.
Scientists may claim that there are neural brain states and processes that cause human behaviour, or that there are beliefs and wants that causally produce such behaviour. Scientists may make claims about there being causal connections between certain aggregate variables, such as the money stock and inflation, or between inflation and unemployment. Such claims have ontological presuppositions, but practising scientists seldom engage themselves in explicating and elucidating them. Are there mental states such as beliefs and wants in addition to brain states? In referring to preferences and expectations in their explanations, are economists presupposing that they exist? Are there macroeconomic aggregates in addition to individuals and their attributes? Do economists commit themselves to their existence when invoking them in explaining economic phenomena? And what is causation all about? Is there just one kind of causation out there, or are there perhaps different kinds of causal connections and other causal facts, depending, among other things, on whether we talk about connections between brains, minds,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
individual actions, institutions, or economic aggregates?
Explicitly or implicitly, economists hold presuppositional beliefs about such ontological matters. They give, or imply, answers to the sorts of questions above, provoked by developments and debates in economics, such as those around neuroeconomics and microfoundations. What does this have to do with realism? Not much, as such. Various alternative answers to those questions are compatible with ontological realism. One can be a realist about brains or minds, or both, and one can be a realist about human individuals or social institutions, or both. There is no single privileged 'realism' that would determine the contents of our ontology – there is no such thing as 'the realist ontology'. It indicates a misunderstanding of 'realism' to suggest that there is one general 'realist ontology' that has specific contents concerning issues such as law and causation, or the relations between human individuals and social structures. Realism as such does not imply specific answers to questions 'What exists?' or 'What is the case?' and the like.
Ontological realism is always Realism-about-X, so we get a variety of realisms depending on the value of X. So one can choose to be, or not to be, a realist about electrons, molecules, cells, minds, human individuals, and social organizations, as well as more generally about relations and causal processes, natures and necessities, numbers and sets, parts and wholes, material states and moral values. One can coherently be a realist about some such things while being an anti-realist about others; one can be a realist about molecules without being a realist about morality, or vice versa. And, again, choosing to be a realist does not as such determine what you are a realist about. You choose Realism-about-X as a package, so it is not your choice of realism that implies your choice of some specific X.

What is existence?

It also works the other way around. Your choice of X as the kind of thing that you think exists does not yet make you a realist. In order to qualify as a realist it is not enough to hold that the hardware of the brain exists or that social structures are causally powerful, or what have you. The general form of the thesis of ontological realism is 'X exists' or 'Z is the case' or some such. It is not enough to be specific about the X and the Z; one also needs to say more about the meanings of 'exists' and 'is the case'. No answer to 'what exists' – no list of things that are claimed to exist or to be the case – is alone sufficient for ontological realism. We also need to carefully answer the question, 'What is existence?'

Irreducible existence

Two requirements are needed in order to come up with an appropriate idea of existence. First, realism requires that existence claims be understood literally. Thus, one is a realist about X if one takes X to exist, and by this one means, literally, that X exists rather than ultimately meaning that something else exists. One is a realist about ions and institutions just in case one takes ions and institutions to exist, period. On the other hand, one is not a realist about X if one holds that X exists in the sense that Y exists by virtue of the fact that X is ultimately Y. In such an anti-realist manoeuvre one substitutes a reductionist reading for a literal reading of existence claims. This table at which I write exists in the sense that a certain bundle of atoms exists – tables are ultimately nothing but bundles of atoms. Or to say that the table exists is another way of saying that a certain collection of sense data exists – middle-sized material objects are collections of sense data, literally speaking. Minds and mental states such as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
expectations exist, but their existence is a matter of human brains and neurons existing. A business corporation exists, but to say so is just a convenient way of speaking: it exists more precisely in the sense that an organized collective of human individuals exists – social collectives are sets of individual people after all. This way of using existence claims amounts to an ontologically reductionist strategy that allows for existence claims that are not supposed to be taken literally: an appropriately literal reading is a post-reduction reading. In contrast, an ontological realist about X reads ‘X exists’ literally and insists on its irreducibility to ‘Y exists’.

Consider causation. In order to qualify as a realist about causation it is not sufficient to hold the view that there are causal facts in the world. It is also required that causation be viewed as an irreducible notion, one that cannot be analysed fully in terms of other, non-causal notions. For example, an ontological empiricist about causation might say that causation exists, and then add that it exists in the sense that empirical regularities or constant conjunctions of observable events exist – simply because this is what causation is, in the end. Here causation will be reduced to, or analysed into, non-causal facts. In philosophers' jargon, ‘causal realism’ is a name for a position that denies such a reduction and instead requires claims about causation to be taken at face value. For a realist, causation is, literally, a matter of causing, producing, bringing about, propagating, enabling, inhibiting and so on. A causal realist analyses causation in causal terms, whereas a causal anti-realist analyses causation in non-causal terms and thereby analyses causation away.

So conceived, causal realism is an ontological position. This is compatible with epistemologies of causation that employ non-causal terms. David Hume seems to have been a causal realist – causation in the world is a matter of causes producing their effects – while at the same time his epistemological scepticism suggested that we only have epistemic access to constant conjunctions (see, for example, Strawson, 1989). Or consider Granger-causality, defined partly in terms of predictability of effects. Predictability is an epistemic notion, so a causal realist should not include it in his concept of causation in the world. But Granger-causality does not require such inclusion, so as such it does not rule out the possibility of causal realism. The general idea behind it, that of providing an ‘operational definition’ of causation, does not require Humean scepticism either: Granger-causality can be thought of as a (fallible) element in the economists' imperfect epistemic endeavours to discover irreducibly causal relations in the world.

Independent existence

The second requirement for a position to qualify as realist is that it must hold that whatever exists has to exist in some suitable independent way. On this issue the peculiar features of society and the social sciences impose special requirements on the appropriate conception of realism. Ontological realism about some X – in its answer to the question, ‘what is existence?’ – claims that X exists independently of some Z. Further versions of this idea depend on how ‘independent’ and ‘Z’ are specified.

The usual manner of defining ontological realism is in terms of mind-independence: some X exists independently of the human mind. But even though this idea would seem to apply to many sorts of things, it is problematic in the social sciences. While it may be plausible to claim that galaxies and quarks exist mind-independently, this does not seem a good idea in the case of, say, people's preferences and expectations or a society's institutions and organizations. The conventional existence test that tends to appeal to most people's intuitions is to imagine a situation without minds: take away human minds,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
will galaxies survive? Yes, they will. Take away human minds, will social institutions survive? No way. Social objects are mind-dependent. This does not require, of course, that whatever there is in society is intentionally created by people, as intended results of their purposeful action. Much of what there is in society is being produced as unintended consequences of the actions of people with minds.

Not only are social objects mind-dependent in general, they are also representation-dependent. They are dependent on representations in general for the trivial reason that people are animals active in producing and using representations as an integral feature of human action. Moreover, and more strongly, many – but surely not all – social things are dependent on representations *about them*: the existence of X is dependent on representations of X – on being represented *as* X. Many contracts between economic actors only exist when linguistically represented. Take away the representation, and the contract ceases to exist. Likewise, the euro, the currency in use in most members of the European Union, is partly constituted by representations, among them certain treaties of the European Council. In contrast, the existence of DNA molecules is not dependent on any particular representations of them: supposing they exist at all, they existed both before and after Watson and Crick's double helix representation, and independently of it.

Social reality is variously shaped by, and dependent on, people's beliefs and expectations, goals and wants, plans and impulses, emotions and reasonings, speech acts such as promises and persuasions, agreements and disagreements, collaborations and rivalries, meanings and their interpretations, customs and conventions, and so on. None of this is mind-independent in certain obvious senses.

Science-independence

Thus ontological realism about society and social sciences requires some other idea of independent existence. Here it is advisable to start being more precise about the issue: we are concerned with the ontology of *scientific* realism. This suggests that we need some notion of science-independence (an idea that has been largely ignored by philosophers debating scientific realism). A scientific realist takes galaxies and quarks to be independent of astronomy and physics, of the theories and explanations and procedures in these sciences. But consider social facts. Much of what there is in society is increasingly dependent on science, natural science included. This dependence works through various powerful influences of science on the world views and technologies prevalent in society. It is through these channels that people's beliefs and aspirations, and society's norms and institutions, are shaped by science. These constitute the stuff of which societies are made, so social matters are not science-independent in an obvious sense.

What about the economy and the science of economics, how do they relate to one another? We can repeat some of the things we said above, and we can add an idea of a stronger dependence. The Lucas critique provides one expression to the old and obvious idea that the economy is not fully economics-independent. Many economic facts seem to be dependent on theories and procedures in the science of economics. Among the more striking cases, just think of the dependence of certain practices of finance in the real world on certain theories of finance – such as the Black–Scholes–Merton formula for option pricing (see MacKenzie, 2006). Economic theories and research results shape people's beliefs and world views, and policy advice based on economic theories and research results shapes economic policies, and these in turn shape the economy. Economic theories, people's beliefs and economic facts are furthermore often connected through mechanisms of self-fulfilment and self-defeat. The economy is thus variously

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
dependent on economics. One might suspect that the notion of science-independence does not serve scientific realism well in the case of economics.

In order to see why all this does not undermine scientific realism about economics and the economy, we need to pay attention to the notion of independence itself in ‘exists independently of’. Of the many kinds of independence, it will be sufficient to distinguish two general categories (Mäki, 2002).

Consider again the idea above about society being science-dependent just because science shapes the world views and technologies in society, and these in turn shape people's behaviour and social institutions. What kind of claim is this? I take shaping (or whatever similar terms one may want to use) to be a matter of causal influence. Thus, a causal claim is being made: social matters are causally dependent on science. This is a relief to a realist about society and social sciences since such causal dependence can easily be accommodated by scientific realism.

Scientific realism does not need to deny or despair over the causal dependence of the economy on economics either. The causal influences of economics on the economy travel through obvious channels. Policymakers and others (such as students, investors, entrepreneurs, workers, consumers) learn, directly or indirectly, about economic theories, explanations and predictions, and are inspired by them enough to modify their beliefs. These modified beliefs make a difference for the behaviour of these actors, and this has consequences for the economy. The connections, and hence the dependencies, are causal.

The same holds for cases in which the connections from theory to the world change the latter so as to make the theory more closely correspond to the world – often characterized as the ‘performativity’ of theory. If it is the case that students of economics act more than other students in accordance with the conventional behavioural assumptions of economic theory, this might be because their image of appropriate human behaviour and thereby their actual behaviour is influenced by what they are taught in class. If certain practices in real world finance are in line with the Black–Scholes–Merton formula for option pricing, this may be because the theoretical formula has managed to travel from academic research to economic practice so as to shape the latter. In such cases, the connections are causal.

Rather than constituting a threat to realism, such causal connections between economic theory and the economy pose a constructive challenge to realism. The economy is organized into causal structures and processes, and investigated by economists. Adding economic theory to the picture in some cases means adding another causal set of connections to our image of social reality – science is not done outside of society, it is very much part of social reality, after all. This invites further scientific inquiry into the detailed features of these causal connections. Good social science illuminates the roles that ideas play in the social world. Good social science should also illuminate the roles that scientific ideas play in the social world.

A realist is comfortable with causal connections in general, and so is not disturbed by causal connections from theories to the world either. To see why, it is important to be precise about the nature of this dependence. Indeed, literally speaking, there is no causation flowing from theories to the world here. Economic theories do not shape the economy. People do. People, in various roles as economic actors, are inspired, directly or indirectly, by the contents of economic theories, this shapes their beliefs, their beliefs shape their motives, and those motives drive them in action that shapes the economy. Phrases such as ‘self-fulfilling’ and ‘self-defeating’ theories and predictions are therefore not to be taken literally: theories do not fulfil or defeat themselves (Mäki, 2002; 2005a). If they had the capacity to do so, scientific realism would be defeated. Likewise, the idea of ‘performativity’ of economic theories is not to be understood literally: theories or their utterances alone do not serve performative functions in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the sense of themselves giving rise to what those theories are about.

Social construction

So scientific realism is comfortable with a lot of science-dependence of society. What sort of science-dependence does scientific realism rule out? As an analogy, consider any of the numerous economist jokes (such as ‘No reality, please. We're economists.’ Blaug, 2002, p. 36). There is no way the joke could be funny without being perceived as such. The joke's being funny is constituted by some people – not necessarily economists – finding it funny. Being funny and being perceived as funny are inseparable, not because the funniness of the joke (causally) makes people laugh, but because their laughter (conceptually) makes it funny. The connection is not causal – as in funny jokes causing people to find them funny – but rather conceptual or constitutive – being perceived as funny is part of the definition of ‘funny,’ or partly constitutes what it is to be funny.

Economic facts in social reality would have the same status as funny jokes if they were facts only if perceived to be so by economists. The economy being in state S and some of its developments being governed by mechanism M are so just because economists hold theories and other beliefs that say so. This is what the realist will deny. A realist will grant there is a formal contract between two economic actors only provided these actors believe it is there, and they – perhaps together with third parties – have agreed to represent the contract. But a scientific realist about the economics of contracts takes those contracts to be science-independent in the sense of not being created by acts of economic theorizing. The contracts exist and have the properties they have independently of being believed or claimed to be so by economists. While facts about contracts are constituted by the beliefs and representations by contracting parties, they are not constituted by acts of scientific theorizing about them. Creating a theory of X is not a matter of creating X. Creating a theory of (the causation of) business cycles does not, just by that token, create (that very causation of) business cycles. Saying so does not make it so.

This idea of non-causal science-independence suggests how to identify some of the opponents of scientific realism. They are those who do not subscribe to the non-causal science-independence of matters of fact in the world, including social and economic facts. Of contemporary (both academic and broader cultural) relevance are forms of social constructivism (see Hacking, 1999). This is a doctrine that comes in an obscure variety of forms that are characteristically not distinguished from one another. Modest versions claim that beliefs about the world are outcomes of social construction (including education, negotiation, persuasion, testimony, imitation, indoctrination, herd behaviour, group pressure, cognitive path dependency, and other social processes). More radical versions claim that the world itself is socially constructed. More weakly, one may argue that beliefs and myths about gender, race and schizophrenia are socially constructed – or that scientific theories about atomic structure and evolution are so constructed. More strongly, one may argue that gender and race, as well as quarks and evolution, are products of social processes of cognitive construction.

Realism is comfortable with weaker forms of social constructivism. Generally held beliefs about the world are socially constructed – simply because cognition is essentially social activity, with numerous cognitive agents interacting with one another under changing institutional constraints. Moreover, parts of the world itself are socially constructed in obvious ways. Indeed, society itself is socially constructed: social objects, properties, states, and processes are outcomes of social processes. Since this is a general fact about social reality, ontological realism about society grants it.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
So what does realism rule out when it comes to society and social sciences? It is incompatible with positions that deny the non-causal science-independence of social matters. Claiming that social facts are socially constructed by social scientists just by way of constructing concepts and theories about society is beyond the scientific realist. A special version of this is the idea that social facts – certain ordinary kinds of facts about society outside of academia – emerge as an outcome of rhetorical persuasion by economists of their audiences: certain facts about, say, the causation of inflation emerge as soon as masses of economists and others are persuaded by monetarist economists about the causal link between money and inflation. What precisely is the problem with this thought that bothers a realist? First consider what realism accepts. What the realist will accept is that the beliefs held by economists about the causation of inflation are socially constructed by ordinary social processes of academic (and possibly non-academic) persuasion. The realist also accepts that the beliefs held by people other than academic economists – such as politicians, journalists, union leaders and consumers – are outcomes of persuasion. Finally, realism accepts that such beliefs may motivate behaviour that has consequences for the causal connection between money and inflation. Such a connection between beliefs and behaviour is part of the causal structure giving rise to actual inflation rates. So what does realism deny? It denies the radical social constructivist idea that, when an economist puts forth a theory about the causation of inflation and manages to get it generally accepted, this alone will give rise to the causal facts described by the theory. This is to deny that agreement without action is sufficient for making or changing the world.
For a realist, theory construction does not amount to world construction. Even theory dissemination alone is not sufficient for world construction (where ‘world’ refers to what the theory is about). Of course, it is a social fact that masses of people are persuaded about whatever, and this indeed is an outcome of rhetorical practices. What a realist should deny is that what those people are persuaded about is itself constructed just by those rhetorical practices. Again, it is a different idea that people so persuaded may take action and this action may make a difference for some social facts. This requires that the beliefs adopted in consequence of persuasion become causally efficacious in relation to actual behaviour. But this gives us what I called causal construction, and it is compatible with scientific realism.

Model worlds

Much of social constructivism may be inspired by the observation that scientists do not have direct access to the details and complexities of the real world. Scientists do not directly investigate the messy concreteness and complexity of the world, but rather engage in various manoeuvres that seek to ‘prepare’ that world for closer inspection. Economists will easily recognize this as part of their model building practices. Faced with the immense complexity of the world, economists are forced to simplify their images of it by way of using various procedures of omission, idealization and abstraction. The models they build and employ isolate small slices of the world for detailed scrutiny while leaving most details out of the picture. Such procedures of isolation are among those manoeuvres of preparation, and they result in models that appear to describe simple imaginary model worlds rather than the real world. One may then say that models and model worlds are constructed rather than, say, discovered, and that they are socially constructed just because scientific work is intrinsically a social activity. The realist grants this much, but then goes on to insist that, even if economic models of the social world are socially constructed by economists, the social world itself is not socially constructed by the modelling practices of economists.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

At this point it will be helpful to have an account of models that is able to deal with this set of issues (Mäki, 2005b). I take models to be what can be called substitute systems or surrogate systems that serve as direct subjects of inquiry. They are substitute systems in that they are examined instead of the real systems that they replace and stand for. The real systems 'out there' may be too complex (such as social systems), too far away in time or space (such as the origin of the species) to be capable of being directly investigated, or it might be unethical to examine the real system directly (such as using human subjects for medical experimentation). Much of scientists' activity is a matter of building and manipulating such substitute systems in order to learn about their properties and behaviour. A realist about models and the real world would add that the properties of models are directly examined by scientists in order to indirectly learn about the properties of the real world.

A radical social constructivist will not buy this image. In a radical constructivist framework, there is no distinction between model systems and real systems, nor is there room for an idea of indirect acquisition of information about the latter by way of examining the former. There are only socially constructed model systems, and it is a matter of power and persuasion which models will be taken to provide the facts. This is one way of illuminating the strong social constructivist ontology in a nutshell.

Truth

This takes us to the issue of truth. Economists appear to have great difficulties with applying the concept of truth. They characteristically believe that, since models do not reproduce the whole complexity of their subject matter, they are necessarily false, or that, because models involve false assumptions, models themselves are false. In order to see why these beliefs about the falsehood of models are false, we should take a realist perspective to the ontology of truth. This will help further clarify the very concept of realism. Here the focus is on what truth is (rather than what is true) (see Mäki, 2004).

Truth is a property of truth bearers. A truth bearer is true just in case there is a truth maker that makes the truth bearer true. The sentence, or statement, 'The cat is on a mat' is a truth bearer that is made true by its truth maker in the world, namely, the cat being on a mat. 'Education is a major cause of economic growth' is a truth bearer that is made true by its truth maker, namely, education being causally responsible for at least 15 per cent of the growth in the output of an economy (supposing this is how we have defined 'major'). Now all models of the connection between education and growth omit lots of things in the world and make false idealizing assumptions about others. But whether one takes a model to convey true information about the world depends on what one takes the relevant truth bearers to be. In the case that the intended truth bearer is, say, 'Education is responsible for about one quarter of economic growth (in some spatio-temporally specified location)', then its truth makers include those situations in the real world in which education is causally responsible for 23–27 per cent of growth (supposing this interval is what we meant by 'about one quarter'). In the case that the intended truth bearer is something different, namely, a description of the causal mechanism by way of which education contributes to growth, then its truth maker would be that mechanism in the real world (such as some skill mechanism or a signalling mechanism, depending on what we seek to describe with the model).

No model offers the whole truth of its domain, and no model, if spelled out in full, is devoid of false elements (false if considered as potential truth bearers). Understanding that models can be true, or can be used for making true claims, requires understanding the intended relevant truth bearers. These intuitions provide the basis for resisting the popular idea that models cannot possibly be true.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The realist ontology of truth so conceived is rather straightforward. Realism requires that the relevant truth makers be real, that they exist in a suitably independent manner. Truth makers are, after all, just those things the existence of which ontological realism deals with. Truth makers must exist without being non-causally science-dependent. Creating a model of a signalling mechanism does not create that signalling mechanism: the mechanism exists (if it does) independently of any economic models about it – it exists even though there were no models about it.

What does realism say about truth making, the way in which truth makers make truth bearers true? Again, a suitable independence requirement is imposed. The way in which truth makers make true statements or models true must be independent of our ways of coming to hold beliefs about their truth. In particular, epistemic matters do not enter truth by way of, say, making truth making dependent on evidential considerations or our capacity to recognize truth. So we do not make true models true by coming to recognize them as true in virtue of having collected lots of supportive evidence, or having become persuaded by the most prestigious economists, or the like. While we make truth bearers, we do not make them true. Facts do.

Possible existence and possible truth

Here is an important further qualification of what it takes to be a realist about the world and models about it. Above, I have said that it is not sufficient to hold that some entity X exists and that theory T is true about X in order to count as a realist about X and T (it is not enough because one has to add further restrictions in terms of irreducibility, non-causal science-independence, and independence of epistemic attitudes, for example). It is now time to weaken this by saying that holding those ideas is not necessary either. It is enough for a realist to hold the view that X might exist and T might be true about X, that it is possible for X to be the case and T to be true. Such a view will reveal a realist attitude: there is a fact of the matter concerning whether or not X exists and whether or not T is true. It is an attitude that will give real existence and objective truth a chance, but one that at the same time is prepared for concluding that X does not exist or T is not true, after all. This is an advisable attitude for epistemically insecure and self-critical scientific practice based on consistent fallibilism.

Ontological realism about X is primarily an account of what it is for X to exist (if it does). To this, one may add the weaker claim that X might exist, or the stronger claim that X does exist. Likewise with truth. Realism about truth is primarily an account of what it is for a truth bearer T to be true. One may then add the weaker claim that T might be true, or the stronger claim that T indeed is true. Naturally each such claim requires different sorts of supportive argument.

The bite of ontology

Does ontology make a difference for scientific practice and its evaluation? That ontology may have some bite is evident even in the case of the weak version of realism put in terms of possible existence. Some economists criticize the model of perfect competition on the grounds that perfect competition is not even conceivable: its perfection has been taken so far that such a competitive market has become an impossibility (Richardson, 1960). If this is correct, then ontological realism would be an inappropriate attitude regarding perfect competition as depicted by the model.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In general, economists hold, or entertain, ontological convictions, explicitly or implicitly, and these may have consequences regarding the preferred ways of theorizing and explaining. Whatever mismatch there may be between ontology and economic theory, sometimes it plays an important role in driving intellectual progress, sometimes not. When there is a clash between the ontological convictions of an economist and the apparent ontological implications of a theory held by this economist, there may be a healthy pressure to revise the theory so as to realign ontology and theory. Such a clash may, in suitable circumstances, serve as a driving tension that creates pressure to modify theories or invent new theories such that the economist will be able to take the theory with an ontologically realist attitude. An economist may, deep down in his convictions, view certain facts – such as increasing returns, bounded rationality, and institutional structure – as causally powerful features of the economy. At the same time, the modelling conventions and techniques of the discipline may be such that the economist will be unable to incorporate such things in his models. There is a tension that needs to be resolved, and the resolution will be sought in the spirit of ontological realism.

Yet, in general, one's favourite ontology does not determine the contents of theory or model used, nor is it determined by these. The impact of ontology is rather a matter of constraint. In many cases even this much impact is too much. An economist may correctly say things that are quite appropriate in their respective contexts even though those things may clash with deeper ontological convictions. An economist may appropriately say ‘this country is applying monetarist tenets in its economic policies’ even though the deeper conviction may be in line with ontological individualism claiming that no such collective entities as countries exist – only individuals do. Or an economist may say that ‘individuals have preferences’ while a philosopher of mind or a neuroeconomist may endorse the ontological claim to the effect that ‘preferences do not exist’ simply because she believes nothing mental does: preferences really are just configurations of neural states. To claim that individuals have preferences as part of economic discourse does not contradict the eliminativist materialism endorsed by the philosopher – just as the claim ‘there are 15 chairs in the seminar room’ may be correct even though one's philosophical ontology may imply that chairs do not exist, while bundles of atoms do. In such cases, there is no conflict between the two claims in each pair due to the idea that one of the claims, and only one of the claims, informs us, after all, how things are ‘at bottom’ or ‘ultimately’.

A parallel idea is that ontology does not determine methodology, but rather serves as a constraint on it. Thus ontological individualism (the belief that only individuals exist) does not imply a commitment to methodological individualism – the obligation to spell out the microfoundations of economic theories and explanations. The economist may legitimately decide that the things he or she wishes to know can be established using aggregative models not grounded in individual behaviour. For example, it may be that certain facts depend on distributions across individuals and not on individuals themselves. Similarly, the conventional procedure of theoretical isolation (isolating certain phenomena that are to be analysed) by building models of small and simple hypothetical worlds does not imply the ontological conviction that the actual systems in the real world are equally simple and isolated from other aspects of reality. What economists do believe is that, in order to study complex systems, simple models are needed, because the real world is too complicated and because the limited questions asked about it require focusing attention on a limited set of mechanisms (while ontological realism may be used to require that these mechanisms exist also in the complex real world).

Methodology is under-determined by ontology. It depends on ontology but it also depends on other things such as the cognitive interests of the inquirers and their limited cognitive capacities. But even this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
much is enough for concluding that ontology matters.

See Also

- assumptions controversy
- causality in economics and econometrics
- Lucas critique
- methodological individualism
- methodology of economics
- microfoundations
- models
- philosophy and economics
- rhetoric of economics

Bibliography

Blaug, M. 2002. Ugly currents in modern economics. In *Fact and Fiction in Economics. Realism, Models, and Social Construction*, ed. U. Mäki. Cambridge: Cambridge University Press.

Hacking, I. 1999. *The Social Construction of What?* Cambridge, MA: Harvard University Press.

MacKenzie, D. 2006. *An Engine, Not a Camera: How Financial Models Shape Markets*. Cambridge, MA: MIT Press.

Mäki, U. 2002. Some non-reasons for non-realism about economics. In *Fact and Fiction in Economics. Realism, Models, and Social Construction*, ed. U. Mäki. Cambridge: Cambridge University Press.

Mäki, U. 2004. Some truths about truth for economists, their critics and clients. In *Economic Policy-Making under Uncertainty: The Role of Truth and Accountability in Policy Advice*, ed. P. Mooslechner, H. Schuberth and M. Schurtz. Cheltenham: Edward Elgar.

Mäki, U. 2005a. Reglobalising realism by going local, or (how) should our formulations of scientific realism be informed about the sciences. *Erkenntnis* 63, 231–51.

Mäki, U. 2005b. Models are experiments, experiments are models. *Journal of Economic Methodology* 12, 303–15.

Richardson, G. 1960. *Information and Investment*. Oxford: Clarendon.

Strawson, G. 1989. *The Secret Connexion: Causation, Realism, and David Hume*. Oxford: Clarendon.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Mäki, Uskali. "scientific realism and ontology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_R000248> doi:10.1057/9780230226203.1492

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Scitovsky, Tibor (1910 – 2002)

Irma Adelman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

external economies; final distribution of income; free trade; initial and final distributions of income; pecuniary external economies; Scitovsky, T.; tariffs; technological external economies; value judgements

Article

Tibor Scitovsky was born in Budapest. He received a degree in law from the University of Budapest and a degree in economics from the London School of Economics. He migrated to the United States in 1939 and served in the US Army during the Second World War. He taught at Stanford, the University of California at Berkeley, Yale, Harvard, and the London School of Economics. He also worked at the OECD from 1966 to 1968.

His writings are brilliant, original, succinct, lucid and full of subtlety. They always enlighten and move the debate forward. He made fundamental and lasting contributions to a large number of subjects: welfare economics, international trade, economic development and microeconomics. One can discern a unifying theme to these varied contributions; it is to indicate ways in which neoclassical equilibrium analysis fails to capture important aspects of economic reality and, therefore, leads to misleading policy implications from efficiency, stability or welfare points of view. He stresses dynamics, and interdependence among the utilities of consumers and decision-outcomes of producers, as the major sources of divergence of social optima from market equilibria of perfectly competitive economies.

His work in welfare economics points to the impossibility of purging policy analysis from value judgements concerning the optimality of the initial and final distributions of income. This is true whether the initial and final situations are efficient or inefficient, in static terms, or whether or not the possibility for compensation of losers by winners exists. (Compensation restores the original distribution and therefore implies the judgement that the original distribution was optimal.) Rather than abandon the possibilities for policy recommendations entirely, economists should make explicit the value judgements that underlie their policy advice.

His contributions to economic development make the capturing of external economies the cornerstone of development strategy. His classic paper ‘Two Concepts of External Economies’ (1954) distinguishes

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
between technological and pecuniary external economies. In developing countries, the existence of pecuniary externalities argues for the planning of coordinated investment decisions since market prices provide imperfect signals when those decisions are interdependent. He also argues for economic integration and export-led growth in economies too small to secure the advantages of both economies of large-scale production and pecuniary economies of balanced growth.

In trade theory, his ‘Reconsideration of the Theory of Tariffs’ (1942) points to the parallelism between tariffs and the monopolist's markup (or monopsonist's markdown) for exploiting his trading partners and argues that market forces could never approximate free trade, which would have to be imposed and enforced by international agreement or by a dominant large power against each nation's selfish preferences.

His most controversial but most original book, *The Joyless Economy* (1976), tries to introduce into consumption theory the psychologist's classification of satisfactions into comfort, stimulation and pleasure, with emphasis on the psychological trade-off between them. The second part of the book explores the implications of the consumer's ignorance of that psychological trade-off on the rationality of his choice behavior, using American lifestyles as an illustration.

Reading Scitovsky's writings, one is made painfully aware of how much has been lost by the modern trend to mathematize and computerize. By comparison, modern economics appears mechanical and myopic, lacking in subtlety and sweep. Many of the themes raised in his writings appear as fresh today as they were when they were first written. And there are many points still worth following up decades after they were first made. His later work on the integration of microeconomics with macroeconomics, contained in his analysis of the real side of inflation, which builds on the price-maker price-taker distinction first introduced by him in his book on *Welfare and Competition* (1952), is a case in point.

Selected works

1942. Reconsideration of the theory of tariffs. *Review of Economic Studies* 9(2), 89–110.

1952. *Welfare and Competition: The Economics of a Fully Employed Society*. London: Allen & Unwin.

1954. Two concepts of external economies. *Journal of Political Economy* 62, 70–82.

1958. *Economic Theory and Western European Integration*. London: Allen & Unwin.

1964. *Papers on Welfare and Growth*. London: Allen & Unwin.

1969. *Money and the Balance of Payments*. Chicago: Rand McNally.

1970. (With I.M.D. Little and M.F. Scott.) *Industry and Trade in Some Developing Countries*. Oxford: Oxford University Press.

1976. *The Joyless Economy*. Oxford: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1986. *Human Desire and Economic Satisfaction*. Brighton: Wheatsheaf; New York: New York University Press.

How to cite this article

Adelman, Irma. "Scitovsky, Tibor (1910–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000047> doi:10.1057/9780230226203.1493

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Scott, William Robert (1868 – 1940)

Andrew Skinner

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Scott, W. R.; Smith, A.

Article

W.R. Scott (‘the chief’ as A.L. Macfie used to describe him) was born in Armagh on 31 August 1868, the eldest son of Charles and Margaret Scott. He was educated at Canon Stewart's Preparatory School and then St Columba's College, Rathfarnham. Scott went to Trinity College Dublin in 1885, graduating BA in 1889 and MA two years later. 1891 saw his first major publication, *An Introduction to Cudworth's Treatise Concerning Eternal and Immutable Morality*. Scott's philosophical interests were also marked by his *Simple History of Ancient Philosophy* (1894).

In 1896 Scott took up the post of assistant to the professor of moral philosophy in the University of St. Andrews, and three years later became the University's first lecturer in political economy. Scott was responsible for planning the teaching of economics until 1915 when he was translated to the Adam Smith Chair of Political Economy in Glasgow, in succession to William Smart. Scott died in Glasgow on 3 April 1940, after a brief illness.

Scott was extremely active as examiner, teacher, researcher, and adviser to government with a marked interest in contemporary, as well as historical, issues. There are three identifiable strands to his work. The first was through his interest in contemporary economic and social problems, encouraged by his chairmanship for many years of his family's firm of millers in Tyrone. It led to an active involvement in public affairs, in days when economists were less consulted than subsequently. Apart from his membership and chairmanship of committees, national and regional, especially in the 1920s, Scott's name became attached to several departmental and other reports, even when he was not the sole or main author.

He was appointed by the Secretary for Scotland to examine home industries in the North. The ensuing report was published in 1914. He subsequently addressed the *Economic Problems of Peace after War* (1917; 1918), and later worked (with James Cunnison) on *Industries of the Clyde Valley during the War* (1924a) as part of a Carnegie Series.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

A second stream is represented by his pioneering work in economic history. Scott published an edition of *The Records of a Scottish Cloth Manufactory at New Mills, Haddingtonshire 1681–1703* (1905). This was followed by one of his best known studies, the definitive, three volume, *Constitution and Finance of English, Scottish and Irish Joint Stock Companies to 1720* (1910–12). He was president of the Economic History Society from 1928 until his death.

A third contribution is represented by Scott's work as an historian of ideas. His book on *Francis Hutcheson* (1900) is the work of a philosopher well versed in economics and is still a classic. Scott also dramatically advanced contemporary knowledge of Hutcheson's most famous pupil with the publication of *Adam Smith as Student and Professor* (1937) which featured his discovery of important Smith papers in the Buccleuch MSS. Scott's ability to find records was impressive. Some criticism of his handling of them is possible but many of the difficulties which impeded his work have been reduced for later scholars by the discovery and cataloguing of many relevant records of economic and intellectual history which he helped to bring about.

At the time of his death Scott was working on a bibliography, which was published under the auspices of the British Academy, and edited by his successor in Glasgow, Alec Lawrence Macfie.

As his obituarist noted in the *Glasgow Herald* (4 April 1940),

> If one today were to seek the model of his famous and beloved predecessor (Adam Smith) it would be impossible to find a closer re-incarnation than William Robert Scott. He had the same temper, controlled, wide-eyed impartiality of mind allied with an absorbing fire of enthusiasm for reasoned practice and disinterested policy. This is the tribute he would most have appreciated, and it is one which all his friends and students would endorse as the most appropriate.

The British Academy published an appreciation by Sir John Clapham, *William Robert Scott 1868–1940* (London, 1940).

Selected works

1891. *An Introduction to Cudworth's Treatise concerning Eternal and Immutable Morality, with Life of Cudworth and a few Critical Notes*. London: Longmans.

1893. *Geography of Ptolemy Elucidated*, ed. T.G. Rylands. Dublin: Ponsonby & Weldrick.

1894. *A Simple History of Ancient Philosophy*. London: E. Stock.

1900. *Frances Hutcheson: His Life, Teaching, and Position in the History of Philosophy*. Cambridge: Cambridge University Press.

1903. *Free Trade in Relation to the future of Britain and the Colonies: A Plea for an Imperial Policy: a lecture*. St Andrews: W.C. Henderson & Son.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1903–4. The fiscal policy of Scotland before the Union. *Scottish Historical Review* 1, 173–90.

1903–5. Scottish industrial undertakings before the Union. *Scottish Historical Review* 1(1903–4), 407–15; 2 (1905), 53–60, 287–97, 406–11; 3 (1906), 71–6.

1905. *Records of a Scottish Cloth Manufactory at New Mills, Haddingtonshire, 1681–1703*. Edinburgh: Scottish History Society.

1907. Scottish industry before the Union; Scottish industry after the Union. In *The Union of 1707*, ed. P. Hume Brown. Glasgow.

1910–12. *The Constitution and Finance of English, Scottish and Irish Joint Stock Companies to 1720*. 3 vols. Cambridge: Cambridge University Press.

1911a. Is increasing utility possible? In *Celebration of the Five-Hundredth Anniversary of the Foundation of St. Andrews*. Glasgow.

1911b. *Scottish Economic Literature: A List of Authorities*. Glasgow: Scottish Exhibition of History, Art and Industry.

1913. The trade of Orkney at the end of the eighteenth century. *Scottish Historical Review* 10, 360–8.

1917. Mercantile shipping in the Napoleonic wars. *Scottish Historical Review* 14, 272–5.

1917–18. *Economic Problems of Peace after War*. The W. Stanley Jevons Lectures at University College London. Cambridge: Cambridge University Press.

1920. William Cunningham, 1849–1919. *Proceedings of the British Academy* 9, 465–74.

1921. ed. Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 2 vols. London: Bohn's Standard Library.

1923. Books as links of empire: *The Wealth of Nations*. *Empire Review*.

1924a. (With J. Cunnison) *The Industries of the Clyde Valley during the War*. Oxford: Clarendon Press.

1924b. Adam Smith. *Proceedings of the British Academy*. London.

1924c. Adam Smith and the City of Glasgow. *Proceedings of the Royal Philosophical Society of Glasgow* 52.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1926a. Alfred Marshall, 1842–1924. *Proceedings of the British Academy*. London.

1926b. Scottish land settlement. In *Scotland During the War, in Economic and Social History of the World War*, ed. R.D. Jones et al. London: Humphrey Milford.

1928. Joseph Shield Nicholson, 1850–1927. *Proceedings of the British Academy*. London.

1929–30. Economic resiliency. *Economic History Review* 2, 291–9.

1931. The manuscript of Adam Smith's Glasgow lectures. *Economic History Review* 3, 91–2.

1934. Adam Smith and the Glasgow merchants. *Economic Journal* 44, 506–8.

1935a. Adam Smith at Downing Street, 1766–67. *Economic History Review* 6, 79–89.

1935b. The manuscript of an early draft of part of *The Wealth of Nations. Economic Journal* 45, 427–38.

1936. New light on Adam Smith. *Economic Journal* 46, 401–11.

1937. *Adam Smith as Student and Professor, with Unpublished Documents*. Glasgow: Jackson.

1938a. A manuscript criticism of the *Wealth of Nations* in 1776 by Hugh Blair. *Economic History* 4, 47–53.

1938b. *Adam Smith: An Oration*. Glasgow: Jackson.

1939. *Greek Influence on Adam Smith*. Athens: Pyros Press.

1940. Studies relating to Adam Smith during the last fifty years. In *Proceedings of the British Academy*, ed. A.L. Macfie. London.

How to cite this article

Skinner, Andrew. "Scott, William Robert (1868–1940)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000048> doi:10.1057/9780230226203.1494

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Scrope, George Poulett (1797 – 1876)

Barry Gordon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

abstinence theory of interest; cost of production; gluts; Malthus's theory of population; quasi-rent; Ricardo, D.; risk–effort theory of profit; Scrope, G. P.

Article

George Poulett Scrope was one of the most prolific contributors to the literature of political economy in the mid-19th century. He was also one of the more able critics of features of the Ricardian economic orthodoxy which came to dominate that literature. Scrope is at his best as an economist in the series of articles he contributed when chief economics reviewer for the *Quarterly Review* (1831–3). His *Principles of Political Economy* (1833c) is disappointing by comparison.

After education at Harrow, Scrope entered Pembroke College, Oxford in 1815. During the following year he moved to St John's College, Cambridge where he graduated in 1821. He married the heiress Emma Phipps Scrope, altering his surname (which had been ‘Thomson’) to that of his wife, and establishing himself as one of the leading gentlemen of the county of Wiltshire. Scrope was appointed a magistrate of the county in 1823.

Research in geology was an early and enduring involvement. Scrope's distinguished work in this field led to his election to the Geological Society (1824). Two years later, he became a Fellow of the Royal Society, and he continued to publish papers on geological subjects until shortly before his death.

Scrope entered Parliament, and he remained Member for Stroud from 1833 to 1867. During his first 20 years in the Commons he spoke frequently and was a member of numerous parliamentary committees. His contributions in the legislature distinguish him as a ‘man with a philosophic and inquiring mind, trying to explain the upheavals in economic relations, and also to guide policy on behalf of interests that went beyond his own personal gain’ (Fetter, 1980). In debate, Scrope found himself in alliance at times with doctrinaire Ricardians such as Joseph Hume. Scrope supported repeal of the Corn Laws and the Navigation Acts. He was also in favour of parliamentary reform. On a variety of issues, however, he was decidedly ‘unorthodox’. Those issues included: the nature of the currency and the structure of the banking system; the public funding of education; the maintenance of outdoor relief for the unemployed;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and closer government regulation of working conditions in factories. The problems of Ireland were among his special concerns, and he was a leading advocate of the extension of poor law provisions to that country.

The interventions of Scrope within Parliament were supplemented by his publication of numerous pamphlets on current policy issues. In addition, he contributed periodical articles in which he assailed the economic theories of the followers of David Ricardo, and the population doctrine of Thomas Robert Malthus. Scrope rejects a theory of value based on cost of production and he argues for recognition of a relationship between value-in-use and value-in-exchange. He finds the reasoning of Richard Jones on rent much superior to that of Ricardo. Noting the absence of any account of a basis for profit in Ricardian theory, he constructs an abstinence theory of interest and allies this to a risk–effort theory of profit which incorporates the concept of quasi-rent. Scrope is also concerned with the possibilities of over-saving and of a general glut of markets. In this latter respect, his ideas are akin to those of Malthus. However, Scrope is a persistent and incisive critic of Malthusian population theory and the policy implications which his contemporaries drew from it.

Selected works

Scrope's bibliography (Sturges, 1984) contains 175 items on economic, geological, and local history topics. The following is a list of some of the more notable publications dealing with economic issues.

1830a. *The Currency Question Freed from Mystery*. London.

1830b. *On Credit – Currency and its Superiority to Coin*. London.

1831a. The political economists. *Quarterly Review* 44(January), 1–51.

1831b. Malthus and Sadler – population and emigration. *Quarterly Review* 45(April), 97–245.

1832. The rights of industry and the banking system. *Quarterly Review* 47(July), 407–55.

1833a. Martineau's monthly novels. *Quarterly Review* 49(April), 136–51.

1833b. *An Examination of the Bank Charter Question*. London: John Murray.

1833c. *Principles of Political Economy*. London: Longman. Reprinted, New York: Kelley, 1969. Italian trans., 1855. 2nd edn, published as *Political Economy for Plain People*, 1873.

1848. Irish clearances and improvement of waste lands. *Westminster and Foreign Quarterly Review* 50 (October), 163–87.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fetter, F.W. 1958. The economic articles in the *Quarterly Review* and their authors, 1809–1852. *Journal of Political Economy* 66, Pt I, 47–64; Pt II, 154–70.

Fetter, F.W. 1980. *The Economist in Parliament: 1780–1868*. Durham, NC: Duke University Press.

Gordon, B. 1965. Say's Law, effective demand and the contemporary British periodicals, 1820–1850. *Economica* 32, 438–46.

Gordon, B. 1969. Criticism of Ricardian views on value and distribution in the British periodicals, 1820–1850. *History of Political Economy* 1, 390–87.

Opie, R. 1928. A neglected English economist: George Poulett Scrope. *Quarterly Journal of Economics* 44, November, 101–37.

Rashid, S. 1981. Political economy and geology in the early nineteenth century: similarities and contrasts. *History of Political Economy* 13, 726–44.

Rudwick, M.S. 1974. Poulett Scrope on the volcanoes of Auvergne: Lyellian time and political economy. *British Journal for the History of Science* 7, 205–42.

Sturges, P. 1984. *A Bibliography of George Poulett Scrope: Geologist, Economist and Local Historian*. Kress Library Publication No. 24. Boston, MA: Harvard Business School.

How to cite this article

Gordon, Barry. "Scrope, George Poulett (1797–1876)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000054> doi:10.1057/9780230226203.1495

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

search models of unemployment

David Andolfatto
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Unemployment, as it is conventionally defined, is a measure of full-time job search. Individuals generally have the option of allocating their time across many competing uses. It follows that an economic interpretation of unemployment data requires a theory that explains the circumstances in which people may prefer to engage in job search at the expense of other activities. Search models of unemployment are designed to do just this.

Keywords

Beveridge curve; business cycles; efficient allocation; involuntary unemployment; leisure; matching; matching markets; natural rate of unemployment; reservation wage; search models of unemployment; search theory; unemployment; unemployment insurance; wage determination

Article

To the average person (perhaps even the average economist), unemployment is often equated with a state of involuntary idleness. This commonly held view, however, is inconsistent with the way in which unemployment is in fact defined and measured. According to International Labor Organization (ILO) conventions, which are followed by most national labour force surveys, unemployment relates to those individuals who are not employed but who are *actively searching* for work (over the reference period of the survey – for example, the previous four weeks). Non-employed (jobless) individuals who have not engaged in active job search are classified as non-participants. This latter category includes the set of so-called ‘discouraged workers’.
Since active job search consumes time and other resources, relating unemployment to a state of idleness seems wrong. Furthermore, since individuals are free to allocate their time across many competing activities, with time devoted to search yielding at least the prospect of a future payoff, the notion of unemployment as an involuntary state is potentially misleading. To understand unemployment – or, at

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
least, unemployment *data* (as opposed to some other pet notion of unemployment) – one must therefore entertain a theory that explains the circumstances under which individuals find it in their interest to remain non-employed while searching for work. Search models of unemployment are designed to do just this.

Search theory

A hallmark of conventional labour market theory (the neoclassical model and its sticky-wage variants) is the assumption that labour is exchanged in a centralized marketplace. Among other things, a centralized market embodies the idea that there is perfect information concerning the location of all jobs and workers. In any such environment, devoting precious time to an activity like search literally makes no sense (whether or not the wage is market-clearing); that is, individuals either have a worthwhile job opportunity to exploit or they do not. In the former case, they become employed; in the latter, they become non-employed (and would be labelled as non-participants by standard labour force surveys).
The starting point for any search model of unemployment then is to dispense with the notion of a centralized marketplace for labour. Instead, the labour market is viewed as a set of decentralized locations, where firms and workers can potentially meet to form mutually beneficial relationships. In a decentralized market, meetings are to some extent determined by search effort and to some extent by chance. In many ways, the labour market resembles a matching market for couples; that is, one is generally aware that the opposite side of the market consists of better and worse matches (we seldom take the view that there are *no* potential matches). The exact location of the better matches is unknown, but may be discovered with some effort. In the meantime, it may make sense to refrain from matching with ‘substandard’ opportunities that are currently available. But since search is costly, it will generally not be optimal to wait for one's ‘soulmate’ to come along. Furthermore, since relationships are not perfectly durable, there is no reason to expect the stock of singles to converge to zero over time. Nor is it clear, given the technology of match-formation, that having everyone matched at all times (irrespective of match quality) is in any way desirable – even if it is feasible; and even if people generally desire to be matched.
In the context of the labour market then, the key friction that potentially rationalizes job search (and, hence, unemployment) is imperfect information over the location of one's best job opportunity. In such an environment, job search constitutes a form of investment in the acquisition of information. The idea of job search as an information-gathering activity has been in place for some time; see, for example, Stigler (1962) as well as the several papers and references contained in Phelps (1970). Perhaps the most influential early formalization of the theory of job search is provided by McCall (1970); see also Sargent (1987, ch. 2). In the next section, I present the basic idea of this classic literature by way of a simple model.

A simple model

I begin by describing a simple ‘one-sided’ search model; for example, the case in which an individual searches for a job opportunity that pays a given wage depending on match quality. Each time period $t = 1, 2, \ldots, \infty$ is divided into two subperiods that I call stage 1 and stage 2. There will be no

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
intertemporal aspect to individual decision-making, so the model's dynamic equilibrium can be thought of as a sequence of static equilibria.

Each person enters stage 1 endowed with one unit of (indivisible) time, a job opportunity that pays a real wage $w \in [0, \bar{w}]$, and a leisure opportunity that generates a return $0 < b < \bar{w}$. One can interpret *b* as either an unemployment insurance benefit or, more generally, as the consumption available when non-employed (for example, home production). The real wage *w* can be interpreted as the quality of the job–worker match associated with any given job opportunity. Let us assume that individuals do not know the precise location of their best job match (that is, the match that would yield them $\bar{w}$). However, it is assumed that individuals know the distribution of match qualities associated with the given set of available jobs; let $F(w)$ denote the fraction of job matches that yield a wage no greater than *w*. The act of job search is modelled as a random draw from $F(w)$.

At the beginning of stage 1, a choice has to be made: should I search or not? The act of not searching means that time must be allocated during stage 1 across one of two activities: work or leisure. The utility payoff to work is $u(w)$, while the utility payoff to leisure is $u(b)$, where $u'' < 0 < u'$. On the assumption that I choose not to search, optimal behaviour entails choosing to work if $w \geq b$. Alternatively, I could choose to search instead. In this case, my time endowment is 'transported' to stage 2. Here, the cost of job search is given by the assumption of *no recall*; that is, I must abandon my current opportunity *w* to exercise my search option. Let $w' \sim F(w')$ denote the wage associated with a new job opportunity I find as a result of my search. At this stage, I now have the opportunity to either work or enjoy leisure; that is, I will only choose to work at this new job if $w' \geq b$. Hence, the expected utility payoff from search is given by:

$$S(b) \equiv F(b)u(b) + \int_b^{\bar{w}} u(w')\,dF(w').$$

Implicit in this discussion is the assumption that consumption levels during stage 1 and 2 are viewed as perfect substitutes. I am also assuming that consumption is non-storable and the absence of insurance/loan markets.

The problem has been set up here in such a way that it will never be optimal for a person to choose leisure during stage 1. This is because there is always a zero-cost option to consume leisure during stage 2. Hence, the relevant choice during stage 1 is whether to work or search. The optimal strategy is to set a *reservation wage* $w_R(b)$, such that one should choose work if $w \geq w_R(b)$ and choose search if $w < w_R(b)$. This reservation wage must satisfy:

$$u(w_R) = S(b). \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

That is, a person's reservation wage is defined to be the wage that would make a person just indifferent between working at that wage and abandoning it in favour of another activity (in this case, job search). In other words, the reservation wage is a measure of an individual's ‘choosiness’ over job opportunities.

Note that as $S'(b) = F(b)u'(b) > 0$, the reservation wage is an increasing function of the option value associated with leisure. Intuitively, a person with a good outside option (b) can afford to be more discriminating with respect to the quality of any given job opportunity w.

Thus far I have described the choice problem of an individual beginning with an endowed wage/leisure opportunity (w, b). I would like to now describe how this behaviour translates into the evolution of aggregates over time. To render static decision-making optimal in this environment, assume that all individuals begin each period with a match quality w drawn from the distribution $G(w)$. The interpretation here is that economy experiences a ‘structural’ shock every period that ‘shuffles’ individual match qualities, but otherwise leaves aggregate production possibilities unchanged. Following this structural shock, individuals behave in the manner described above. For the sake of simplicity, assume that all individuals share a common value for b that remains constant over time.

It is an easy matter now to characterize the equilibrium unemployment rate. At the beginning of each period, individuals are randomly assigned a match quality w, whose distribution is $G(w)$. The fraction $[1 - G(w_R)]$ will choose to exploit their employment opportunity and the fraction $G(w_R)$ will find it optimal to abandon their opportunity in favour of search. Job searchers find new job opportunities with associated match qualities w' drawn from the distribution $F(w')$. Those who find a job that pays $w' \geq b$ choose to work; those that find a job that pays $w' < b$ choose leisure instead. Thus, $F(b)$ denotes the fraction of job-searchers who are ‘unsuccessful’ in their search. On the assumption that the reference period for a labour force survey is the length of a period (and that the survey is performed at the end of the period), the equilibrium unemployment rate is given by:

$$U = G(w_R)F(b).$$

In other words, the unemployed are those who: (a) performed no work during the reference period; and (b) were actively searching for work during the reference period.

Associated with the ‘steady-state’ unemployment rate U are steady-state flows of workers making transitions from employment to unemployment (job destruction) and transitions from unemployment to employment (job creation); these flows are given by:

$$JC = [1 - G(w_R)F(b)]U; JD = G(w_R)F(b)(1 - U).$$

Any given individual may, of course, experience a variety of employment/unemployment histories. The model developed above constitutes an example of the modern approach to the theory of unemployment. Note that unemployment here is interpreted as an equilibrium phenomenon; in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
particular, it is not the product of irrational behaviour or markets that fail to clear. Instead, unemployment is the natural by-product of an economy subject to ongoing structural disturbances that depreciate the value of existing employer–employee matches, and where match formation takes time owing to the imperfect information pertaining to the location of new job opportunities.

One by-product of the modern theory of unemployment is that renders obsolete much of the traditional language that was used to describe the phenomenon (see Rogerson, 1997). Consider, for example, the traditional classification of unemployment into its 'voluntary' and 'involuntary' components. In the model developed above, there is a sense in which unemployment is both voluntary and involuntary. It is voluntary in the sense that the model people (as with real people) can choose not to search, and instead allocate time to inferior job opportunities or home production activities. On the other hand, it is also involuntary in the sense that the economic circumstances that compel individuals to become unemployed are often beyond their immediate control. By the same token, however, the same classification might be made for employed workers (for example, the existence of those who cannot afford not to work, or the so-called working poor). Whether these traditional labels have any substantive meaning, however, is questionable. In particular, note that the well-being of these model people (and presumably, people in reality) does not depend of how we (as theorists) choose to label them.

Welfare

A classic question in the theory of the labour market is whether an equilibrium level of unemployment corresponds in any way to an efficient allocation. Not surprisingly, the answer to this question depends critically on the nature of the economic environment; see Diamond (1982), Mortensen (1982), Hosios (1990) and Moen (1997). In the present context, the optimal level of unemployment can be characterized as the solution to the following planning problem:

$$\max_{c, w_R} \left\{ u(c) : c \leq \int_{w_R}^{\overline{w}} w \, dF(w) + F(w_R) \left[F(b) b + \int_b^{\overline{w}} w' \, dF(w') \right] \right\}.$$

The socially optimal reservation wage is given by:

$$w_R^* = F(b) b + \int_b^{\overline{w}} w' \, dF(w').$$

The optimal unemployment rate corresponds to the equilibrium unemployment rate if $w_R^* = w_R$, where w_R is the solution to (1). As it turns out, this will be the case if $u'' = 0$ (risk-neutral individuals).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

However, with risk-averse individuals, one can easily establish that $w_R < w_R^*$, which implies here that the equilibrium unemployment rate is too *low*. The suboptimality of equilibrium in this particular model has nothing to do with search per se; rather, it is due to the assumed lack of insurance. The intuition is straightforward: in the absence of a well-functioning insurance market, individuals are too willing to hold on to marginal jobs rather than risk an even worse outcome in the event of an unsuccessful job search.

One striking implication of the modern theory of unemployment is that the unemployment rate bears no obvious relation to any sensible measure of social welfare. Consider, for example, two economies *A* and *B*, identical in every respect except that $b_A > b_B$ (so that the residents of economy *A* are in some sense 'wealthier' than those of economy *B*). In this case, the unemployment rate in economy *A* will be higher than in economy *B*. Nevertheless, given the choice, individuals would rather live in the high-unemployment economy. On the other hand, suppose that the two economies differ only in terms of *F*, with F_A stochastically dominating F_B. In this case, economy *A* will have lower unemployment and higher welfare. The basic lesson here is that 'economic performance' should not be measured in terms of how individuals choose to allocate their time across competing activities, but rather should be measured in terms of the level and distribution (or stochastic properties) of broadly defined consumption.

Unemployment insurance

Governments in many countries operate a fiscal policy known as unemployment insurance (UI). UI systems are characterized by transfers of income to those who are unemployed (or otherwise meet certain eligibility requirements) that are typically financed via a payroll tax (or out of general tax revenue). Presumably, the motivation for such transfers rests on the belief that: (1) private insurance markets are unavailable (or work poorly) and (2) self-insurance (via precautionary saving) is either grossly inefficient, or perhaps beyond the means of many workers (a more cynical view interprets UI as one of many government transfer schemes designed to benefit various special interests at the expense of the general taxpayer).

Search models of unemployment have been applied in both positive and normative investigations concerning the effects of UI programmes. For example, in the context of the model developed above, let *b* denote a UI benefit that is financed via a payroll tax τ. For given programme parameters (b, τ), the reservation wage now satisfies:

$$u((1-\tau)w_R) = F(b)u(b) + \int_b^{\bar{w}} u((1-\tau)w')\,dF(w'), \tag{2}$$

which implicitly characterizes $w_R(b, \tau)$. In equilibrium, the tax and benefit level is related by the government budget constraint:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\tau\left[\int_{w_R}^{\overline{w}} w\,dF(w) + F(w_R)\int_{b}^{\overline{w}} w'\,dF(w')\right] = F(w_R)F(b)b. \quad (3)$$

For a given UI policy parameter (say, an exogenous level of *b*), eqs (2) and (3) characterize an equilibrium $(\hat{w}_R, \hat{\tau})$. The positive and normative implications of the theory can then be deduced by varying the policy parameter *b*. Andolfatto and Gomme (1996) consider a similar such experiment, albeit in a considerably richer environment that (among other things) models the UI system in greater detail (in particular, UI programmes are typically characterized by several programme parameters, including eligibility requirements, replacement rates and benefit duration parameters).

Several papers have recently examined the issue of how an optimal UI system should be designed. This question becomes particularly interesting when one reasonably assumes that workers have private information concerning the nature of their job opportunities and/or the intensity with which they search. In a classic paper, Shavell and Weiss (1979) demonstrate that an optimal UI programme should 'front load' UI payments, with benefit levels declining monotonically over the unemployment spell. The high initial benefit level provides the desired insurance, while the declining benefit profile mitigates adverse incentive effects by stimulating (unobserved) search intensity. Wang and Williamson (1996) and Hopenhayn and Nicolini (1997) flesh out other properties of optimal UI systems in the context of dynamic moral hazard environments. Among other things, these authors report that optimally designed programmes can deliver potentially large welfare benefits relative to existing systems.

Business cycles

Search models of unemployment have also been used to interpret various aspects of the business cycle. Naturally, a key set of questions deal with the cyclical properties of unemployment itself. However, researchers have also investigated the extent to which 'search frictions' in the labour market shape the pattern of the business cycle more generally (see Mortensen and Pissarides, 1994; Merz, 1995; Andolfatto, 1996).

An empirical relationship that has drawn considerable attention in the literature is the so-called Beveridge curve; that is, the tendency for unemployment and vacancies to move in opposite directions over the business cycle (for example, see Blanchard and Diamond, 1989). The level of job vacancies, as measured by the help-wanted index, is highly volatile and tends to lead unemployment over the cycle. A natural interpretation of these facts is that as job opportunities become more plentiful, unemployed workers are able to find jobs more easily, with search frictions preventing instantaneous adjustment. Vacancies themselves can be interpreted as the business sector equivalent of unemployed workers. Since finding suitable workers is costly and time-consuming, and since employment relationships once formed are durable, recruiting intensity constitutes a form of capital spending that presumably reacts to actual and expected shocks in much the same way as other forms of capital spending (for example, see Howitt, 1988).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

A seminal paper in this area is Pissarides (1985), who appeals to the concept of an aggregate matching technology to model (in a reduced-form manner) the outcome of uncoordinated search in the labour market (see also Pissarides, 2000; Petrongolo and Pissarides, 2001). According to this approach, match formation (job creation) is the product of the aggregate search intensity of workers (unemployment) and firms (vacancies), both of which serve as complementary inputs into an aggregate matching function, that is, $m_t = M(v_t, u_t)$. Ignoring individual differences, *m*/*v* and *m*/*u* can be interpreted as the probabilities that vacant jobs and unemployed workers contact each other. If *M* displays constant returns to scale, then these probabilities depend only on the labour-market tightness variable $\theta \equiv v / u$. In particular, $m / v = q(\theta)$ and $m / u = p(\theta)$, where $q(\theta) \equiv M(\theta, 1)$ and $p(\theta) \equiv M(1, \theta^{-1})$. Note that $q'(\theta) < 1$ and $p'(\theta) > 1$. In other words, firms find it more difficult to find workers in a tighter labour market; while the converse is true for workers.

With a fixed labour force *L*, the stock of employed workers at date *t* is given by $L–u_t$. Again note that the stock of employment constitutes a form of capital. Employment capital depreciates over time as matches dissolve owing to idiosyncratic shocks that affect the viability of individual relationships. Let $0 < \sigma < 1$ denote the fraction of employment relationships that are terminated at the end of each period. In this case, the job destruction flow is given by $\sigma(L - u_t)$. For simplicity, assume that σ is exogenous. At the same time, the stock of employment is replenished by the flow of job creation, $p(\theta_t)u_t$, so that the stock of unemployment evolves over time according to:

$$u_{t+1} = u_t - \sigma(L - u_t) + p(\theta_t)u_t. \tag{4}$$

In a steady state, the flow of job creation just offsets the flow of job destruction, so that the 'natural' rate of unemployment satisfies (normalizing *L*=1):

$$u = \left(\frac{\sigma}{\sigma + p(\theta)}\right). \tag{5}$$

Equations (4) and (5) relate the dynamics of the unemployment rate to the labour-market tightness variable θ, as well as to parameters describing the structure of the matching market. In the simple version considered here, the unemployed search passively (workers have one unit of time that they allocate either to work or search at zero utility cost). Consequently, the equilibrium $\theta_t = v_t / u_t$ is determined entirely by vacancy creation, which is endogenized as follows.

Assume that each firm has a single job that, together with a worker, produces $y_t > 0$ units of output.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Assume that y_t follows a first-order Markov process with $G(y', y) = \Pr[y_{t+1} \leq y' | y_t = y]$. For simplicity, assume that match quality is identical across all firm–worker pairs and that wages are determined by an exogenous bargaining process that divides output in some manner between firm and worker; that is, let $0 < \xi < 1$ denote the fraction of output that accrues to the firm (alternatively, one could model wage determination by imposing a Nash bargaining solution). Let J denote the capital value of a firm currently matched with a worker; this value must satisfy the following Bellman equation:

$$J(y) = \xi y + \beta(1 - \sigma) \int J(y') G(dy', y), \tag{6}$$

where $0 < \beta < 1$ denotes a discount parameter. Implicit in this formulation is that the value of the firm falls to zero in the event that the match is dissolved. Note that if match productivity exhibits positive persistence, then $\int J(y') G(dy', y)$ is increasing in y. In other words, a positive productivity shock (increase in y) is associated with information that leads firms to revise upward their estimate of the returns to match formation.

Finally, consider the cost–benefit analysis associated with vacancy creation. Assume that creating (and maintaining) a vacancy entails the flow cost $\kappa > 0$, measured in units of output. A vacant job potentially meets an unemployed worker with probability $q(\theta)$, with the match producing a flow of output beginning in the subsequent period. Let Q denote the capital value of a vacant job; this value must satisfy the following Bellman equation:

$$Q = -\kappa + \beta \left[q(\theta) \int J(y') G(dy', y) + (1 - q(\theta)) Q' \right]. \tag{7}$$

Assuming free entry into vacancy creation, the level of v will expand as long as $Q > 0$. But note that for a fixed level of unemployment u, any expansion in the number of vacancies increases labour-market tightness θ and therefore reduces the success probability $q(\theta)$. In equilibrium, θ adjusts to a point that renders further vacancy creation unprofitable; that is, $Q = Q' = 0$. Consequently, the equilibrium labour-market tightness variable $\theta(y)$ is determined by:

$$q(\theta) \beta \int J(y') G(dy', y) = \kappa. \tag{8}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The economic mechanism at work here can be described as follows. Consider a positive productivity shock. Since the shock is persistent, the return to job creation (contemporaneous recruiting activities that augment the future stock of employment) increases (see (6)). Naturally, in response to these bright prospects, firms increase their recruiting activities and create new vacancies. However, as the competition for new workers intensifies, the probability of success falls to the point where further expansion becomes uneconomical (see (8)). On impact then, the effect of the shock leaves current unemployment unchanged, but increases the demand for labour (supply of vacancies). The dynamic effects of the shock can then be traced out by appealing to eq. (4). In particular, note that since the job-finding probability for workers $p(\theta)$ is increasing in θ, the effect of the shock is to lower the future rate of unemployment. These effects continue to be propagated forwards in time through the search mechanism.

Current issues

Several authors have observed that the unemployment rate appears to exhibit a type of cyclical asymmetry, rising sharply at the onset of a recession, but declining only slowly over the course of a subsequent recovery (for example, see Neftçi, 1984; Hussey, 1992). In an influential study, Davis and Haltiwanger (1992) investigate US manufacturing data and report that job destruction appears to be significantly more volatile than job creation over the cycle. The natural conclusion that follows from this body of work is that recessions are attributable to shocks that lead to brief, but sharp, increases in job losses followed by relatively dampened, but prolonged, periods of job creation as the business sector slowly rebuilds its employment capital (see, for example, Hall, 1995).

Shimer (2005a) has recently cast doubt on this ‘conventional’ view of the cycle. In his detailed examination of Current Population Survey data, Shimer concludes that almost all the cyclical variability in the unemployment rate is attributable to fluctuations in the job-finding probability (and in particular, the job-finding rate associated with transitions from unemployment to employment, rather than from non-employment to employment). Surprisingly (relative to received wisdom), the separation rate appears to be very nearly acyclical and relatively stable. In addition to these patterns, Shimer reports a very high degree of correlation between the job-finding probability and the vacancy–unemployment ratio.

Taken together, this new body of evidence suggests that, at least for the purpose of understanding cyclical behaviour, one may reasonably begin by organizing ideas around a simple Pissarides-style search model along the lines described above; that is, with a constant labour force L, an exogenous separation probability σ, and a job-finding probability $p(\theta)$ that varies with labour-market tightness. Whether a suitably calibrated version of this model can account for observation, however, remains a topic of current debate (for example, see Shimer, 2005b; Hagedorn and Manovskii, 2005; Hall, 2005; Mortensen, 2005; as well as Hornstein, Krusell and Violante, 2005 for a nice summary).

Apart from cyclical phenomena, several outstanding issues remain unresolved in terms of understanding secular and cross-country measures of labour-market activity. At the forefront of these phenomena is the dramatic rise in European unemployment rates relative to the United States since the early 1970s (for example, Ljungqvist and Sargent, 1998; Blanchard and Wolfers, 2000). But unemployment is just one of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the three major classifications of time allocation, along with employment and non-participation (the latter two of which contain many more people than the former). Rogerson (2001) documents several interesting facts concerning the cross-country differences and low-frequency movements in employment-to-population ratios. Among other things, cross-country differences are large and persistent over time, with considerable movement within the distribution of employment-to-population ratios across countries. Furthermore, while employment and unemployment closely mirror each other at business cycle frequencies, the same is not true at lower frequencies. This latter observation suggests that the commonly invoked short-cut of abstracting from participation decisions for business cycle analysis may be inappropriate when investigating the causes of secular movements and cross-country differences in time allocation.

On the theoretical front, several issues remain the subject of ongoing research. Paramount among these are investigating the microeconomic foundations of the matching technology and the process of wage determination in decentralized markets. On these and related matters, the interested reader may refer to Rogerson, Shimer and Wright (2005) – a recent survey that contains 167 references while claiming to only ‘scratch the surface’ of this interesting and rapidly expanding body of research.

See Also

- involuntary unemployment
- matching
- matching and market design
- search theory
- unemployment insurance

Bibliography

Andolfatto, D. 1996. Business cycles and labor-market search. *American Economic Review* 86, 112–32.

Andolfatto, D. and Gomme, P. 1996. Unemployment insurance and labor market activity in Canada. *Carnegie-Rochester Series on Public Policy* 44, 47–82.

Blanchard, O. and Diamond, P. 1989. The Beveridge curve. *Brookings Papers on Economic Activity* 1, 1–60.

Blanchard, O. and Wolfers, J. 2000. The role of shocks and institutions in the rise of European unemployment. *Economic Journal* 110, 1–33.

Davis, S. and Haltiwanger, J. 1992. Gross job creation, gross job destruction, and employment reallocation. *Quarterly Journal of Economics* 107, 818–63.

Diamond, P. 1982. Wage determination and efficiency in search equilibrium. *Review of Economic*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Studies 49, 217–27.

Hagedorn, M. and Manovskii, I. 2005. The cyclical behavior of equilibrium unemployment and vacancies revisited. Working paper. University of Pennsylvania.

Hall, R.E. 1995. Lost jobs. *Brookings Papers on Economic Activity* 1, 221–56.

Hall, R.E. 2005. Employment fluctuations with equilibrium wage stickiness. *American Economic Review* 95, 50–65.

Hopenhayn, H. and Nicolini, J.P. 1997. Optimal unemployment insurance. *Journal of Political Economy* 105, 412–38.

Hornstein, A., Krusell, P. and Violante, G. 2005. Unemployment and vacancy fluctuations in the matching model: inspecting the mechanism. *Federal Reserve Bank of Richmond Economic Quarterly* 91, 19–51.

Hosios, A.J. 1990. On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies* 57, 79–98.

Howitt, P. 1988. Business cycles with costly search and recruiting. *Quarterly Journal of Economics* 103, 147–65.

Hussey, R. 1992. Nonparametric evidence on asymmetry in business cycles using aggregate employment time-series. *Journal of Econometrics* 51, 217–31.

Ljungqvist, L. and Sargent, T. 1998. The European unemployment dilemma. *Journal of Political Economy* 106, 514–50.

McCall, J.J. 1970. Economics of information and job search. *Quarterly Journal of Economics* 84, 113–26.

Merz, M. 1995. Search in the labor market and the real business cycle. *Journal of Monetary Economics* 36, 269–300.

Moen, E. 1997. Competitive search equilibrium. *Journal of Political Economy* 105, 385–411.

Mortensen, D.T. 1982. Property rights and efficiency in mating, racing, and related games. *American Economic Review* 72, 968–79.

Mortensen, D.T. 2005. More on unemployment and vacancy fluctuations. Working paper. Northwestern

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
University.

Mortensen, D.T. and Pissarides, C. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61, 397–415.

Neftçi, S. 1984. Are economic time-series asymmetric over the cycle? *Journal of Political Economy* 92, 307–28.

Petrongolo, B. and Pissarides, C. 2001. Looking into the black box: a survey of the matching function. *Journal of Economic Literature* 39, 390–431.

Phelps, E.S. 1970. *Microeconomic Foundations of Employment and Inflation Theory*. New York: W.W. Norton.

Pissarides. 1985. Short-run equilibrium dynamics of unemployment, vacancies, and real wages. *American Economic Review* 75, 676–90.

Pissarides. 2000. *Equilibrium Unemployment Theory*, 2nd edn. Cambridge, MA: MIT Press.

Rogerson, R. 1997. Theory ahead of language in the economics of unemployment. *Journal of Economic Perspectives* 11(1), 73–92.

Rogerson, R. 2001. The employment of nations: a primer. *Federal Reserve Bank of Cleveland Economic Review* 4, 27–50.

Rogerson, R., Shimer, R. and Wright, R. 2005. Search-theoretic models of the labor market: a survey. *Journal of Economic Literature* 43, 959–88.

Sargent, T.J. 1987. *Dynamic Macroeconomic Theory*. Cambridge, MA: Harvard University Press.

Shavell, S. and Weiss, L. 1979. The optimal payment of unemployment insurance benefits over time. *Journal of Political Economy* 87, 1347–62.

Shimer, R. 2005a. Reassessing the ins and outs of unemployment. Working paper. University of Chicago.

Shimer, R. 2005b. The cyclical behavior of equilibrium unemployment and vacancies. *American Economic Review* 95, 25–49.

Stigler, G.J. 1962. Information in the labor market. *Journal of Political Economy* 70, 94–105.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Wang, C. and Williamson, S. 1996. Unemployment insurance with moral hazard in a dynamic economy. *Carnegie-Rochester Conference on Public Policy* 44, 1–41.

How to cite this article

Andolfatto, David. "search models of unemployment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000521> doi:10.1057/9780230226203.1497

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

search theory (new perspectives)

Shouyong Shi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the recent development of search theory. We begin with a standard model of undirected (or random) search and illustrate that the equilibrium is generically inefficient. Then, we show that the market can restore efficiency by allowing individuals on one side of the market to use prices to direct other individuals' searches. Two mechanisms of directed search – price posting and auction – are analysed and a number of applications of the framework are discussed. In addition, we briefly describe search theory as a microfoundation for monetary economics. Finally, we conclude by speculating a few directions of future research.

Keywords

assortative matching; asymmetric information; auctions; bargaining; contract theory; directed search; efficient allocation; fiat money; first-price auctions; labour market search; matching; mechanism design; monetary search theory; optimal dynamic contracts; price dispersion; price posting; random-matching models; risk aversion; search models of unemployment; search theory; sequential search models; trading frictions; unemployment insurance; value theory of money; wage dispersion; wage posting

Article

Search theory is an analysis of resource allocation in economic environments with trading frictions. These frictions include the difficulty of bringing potential traders together, coordinating agents' decisions, informing agents of trading opportunities and keeping records of agents' trading histories. In the market, trading frictions appear in various forms of transactions cost and they generate important regularities in quantities and prices. For example, there are unemployed workers, underutilized capital, and unsold goods in inventory, which indicate that markets are unable to exhaust all potentially desirable trades. Also, the law of one price predicted for a frictionless economy is at odds with the dispersion of prices often observed for similar goods.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In his classic article on search theory in the 1987 edition of *The New Palgrave* (reproduced in this edition), Diamond surveyed the early development of the theory. He formulated two types of search models. One is sequential search models, in which agents on one side of the market post prices. Agents on the other side receive price quotes at an exogenous rate and decide sequentially whether to accept the quotes. The other type is random-matching models, in which agents determine prices through bargaining after they are matched. Diamond explained how these models can generate price dispersion in the equilibrium. He also showed that the market fails to allocate resources efficiently in these models. Since 1987, the literature has extensively applied the two models to analyse economic issues such as price (wage) inequality, unemployment, business cycles, marriages, and investments in physical and human capital.

Search theory has also experienced developments on the theoretical side. One development is the exploration of the mechanisms which can improve efficiency of the market. A particular mechanism is directed search or competitive search, as pioneered by Peters (1991). This exploration has led to the formulations of search as a strategic game. Another development uses search to construct a microfoundation for monetary theory. This article will focus on directed search, with only a brief description of the search theory of money. For reference, we call the models in Diamond's search theory random (or undirected) search models.

The main difference between directed and undirected search lies in whether individual agents are able to use pricing mechanisms to change their matching frequency. Random search models specify pricing and matching as two independent processes. In particular, the matching frequency is either a parameter or a function of aggregate variables only. Because prices play a limited role, the market cannot internalize the externalities in the search process. By contrast, directed search models allow individual agents to use pricing mechanisms to directly affect the matching frequency. The explicit trade-off between the matching frequency and price improves efficiency.

Because efficiency is a fundamental issue in economics, it is the primary motivation for studying directed search. Another motivation is simply the fact that directed search is a realistic description of many markets. Sellers often advertise prices and buyers know many price offers before they visit stores (firms). Incorporating directed search may lead to a robust explanation for phenomena such as persistent unemployment – however, we will not pursue this empirical agenda here.

Instead, we will start with random search models and illustrate the inefficiency of the equilibria. Then, we will describe three models of directed search and related issues. Our conclusion will follow the brief description of monetary search theory. To simplify the language, we will treat the market as a labour market and let the time horizon be one period. The models are applicable to the goods market with straightforward modifications and they can be extended to infinite horizon (see Cao and Shi, 2000).

Random search and inefficiency

Consider a labour market that contains a large number of workers and firms. The number of workers searching for jobs is a fixed number u. All workers are the same and they are risk neutral. When employed, a worker produces goods whose value is $y>0$. When unemployed, a worker enjoys leisure, the utility of which is normalized to 0. The number of vacancies is v, which is determined by competitive entry of firms. A potential firm can incur a cost c to create a vacancy, where $0<c<y$. The technology of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
production has constant returns to scale so that a firm treats each vacancy separately. Normalize the production cost to 0.

Let us use a matching function, $M(u,v)$, to describe the total number of matches in the period. Let $\theta = u / v$ denote the 'tightness' of the market. The matching probability for a worker is $p(\theta) = M(u, v) / u$ and the matching probability for a vacancy is $q(\theta) = M(u, v) / v$. Assume that M is increasing, concave and differentiable in each argument for all θ such that $p,q \in (0,1)$. Moreover, the function has constant returns to scale. Thus, $p(\theta)$ is decreasing in θ, $q(\theta)$ is increasing in θ, and $q(\theta) = \theta p(\theta)$. Moreover, assume that $q(\theta)$ is concave, $q(0) = 0$, and $q(\infty) = 1$. The matching share of workers is defined as

$$s(\theta) = \frac{u}{M} \frac{\partial M(u, v)}{\partial u}.$$

Then, $s(\theta) = 1 + \theta p'(\theta) / p$. The matching share of firms is $(1-s)$.

Once a worker and a firm are matched, the two choose the wage for the worker, w. Assume that this is done with Nash bargaining, which maximizes the geometrically weighted surplus of the two sides of the match: $w^{\sigma}(y-w)^{1-\sigma}$. Here, the worker's bargaining weight is $\sigma \in [0,1]$. The solution for the wage share is $w/y = \sigma$.

The value of a vacancy is $J = q(\theta)(y - w)$ and the value of a worker's search is $V = p(\theta)w$. With competitive entry of firms, a firm's net profit is zero; that is, $J=c$. This equilibrium condition can be rewritten as $(1 - \sigma)q(\theta) = c / y$. A unique solution for θ exists if $0 < c < (1 - \sigma)y$.

In the equilibrium, some workers are unemployed and some jobs are vacant. However, the existence of unemployment alone is not a sufficient indication of inefficiency. With the matching technology, not all resources can be fully utilized. The appropriate notion of efficiency must respect the constraint of this technology.

Let us measure efficiency with a social welfare function. Define social welfare as the weighted sum of expected values of agents in the economy, where all agents are given the same weight. This measure is also equal to the expected utility of an agent who is ignorant of whether he or she is a worker or a firm. Because firms earn zero net profit, the welfare level is equal to the sum of workers' values, uV. Using the condition of competitive entry to substitute w, we can express the welfare level as aggregate output minus the vacancy cost, that is, $u[p(\theta)y - c / \theta]$. Because u is exogenous, the level of θ that maximizes welfare satisfies $-\theta^2 p'(\theta) = c / y$. If we compare this efficient outcome with the equilibrium outcome, we can see that the equilibrium is efficient if and only if:

$$\sigma = s(\theta).$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This condition is the Hosios condition (see Hosios, 1990). It is required for efficiency for the following reason. The social value created by a marginal firm is $v[\partial M(u,v)/\partial v] = v(1-s)q$. In contrast, the firm's value in the equilibrium is $q(y-w)$. For the equilibrium to be efficient, the firm's value in the equilibrium must be equal to its social value. This requirement is met if and only if the wage share is equal to the matching share of workers, as the Hosios condition requires.

More specifically, a firm's entry into the market creates two externalities. One is positive – the presence of an additional firm increases the matching frequency of workers. The other externality is negative – an additional firm reduces the matching frequency of other firms. The Hosios condition ensures that the two externalities cancel out with each other. If $\sigma > s(\theta)$, a firm is under-compensated for its entry cost and the amount of entry is deficient; if $\sigma < s(\theta)$, a firm is over-compensated and the amount of entry is excessive.

With random matching, the equilibrium cannot satisfy the Hosios condition generically, because both sides of the condition involve exogenous elements of the model. In particular, when the matching function is Cobb–Douglas, the matching share s is a constant which is unrelated to the workers' wage share.

The inefficiency is not specific to the random-matching model. Instead, it is common to all undirected search models. For example, if a sequential search model is used instead of a random-matching model, then firms post wages rather than bargain over wages. There can be a non-degenerate distribution of wage shares in the equilibrium (see Burdett and Mortensen, 1998). However, the matching share will still be independent of the wage share and the independence leads to the violation of the Hosios condition.

Directed search and efficiency

Directed search links the wage share to the matching share by explicitly modelling an agent's trade-off between the wage and the matching frequency. To capture this trade-off, suppose that all agents expect each wage level to be associated with a market tightness by a function $\theta(w)$. Search is 'directed' in the sense that, by posting a particular wage, a firm expects to change the matching probability by affecting workers' applications. For a firm posting wage w, the matching probability is $q(\theta(w))$; for a worker who applies for wage w, the matching probability is $p(\theta(w))$. The functions $p(\theta)$ and $q(\theta)$ have the properties assumed above. Given the tightness function, each firm chooses a wage to post to maximize the expected value $J = q(\theta(w))(y-w)$, and each worker chooses to apply for a wage that maximizes the expected value $V = p(\theta(w))w$. The equilibrium tightness must be consistent with competitive entry and workers' application decisions.

Without restricting the function $\theta(\cdot)$, there can be many equilibria. For example, take an arbitrary wage $w_0 \in (0,y)$, and let θ_0 satisfy: $q(\theta_0)(y-w_0) = c$. Suppose that workers believe that all firms will post only wage w_0. With this belief, workers will apply only for wage w_0. But if no worker applies to other wages, then all firms will indeed post only wage w_0. That is, the pair (w_0, θ_0), together with the particular belief, is an equilibrium. In this equilibrium, $\theta(w)$ is not well-defined for $w \neq w_0$, because there is no firm or worker at such wages.

One way to avoid this problem is to introduce a small measure of non-optimizing firms that post every feasible wage and to analyse the limit of the equilibrium when this measure approaches zero. Another

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

way is to impose restrictions on the beliefs out of the equilibrium, as we do here. Let E be the set of equilibrium wages. For $w^* \in E$, denote the expected value of applying to w^* as $V^* = p(\theta(w^*))w^*$. We require that, for every $w^* \in E$, the function $\theta(\cdot)$ must satisfy $p(\theta(w))w = V^*$ for w in a neighbourhood of w^*. That is, a firm believes that workers will apply for a deviating wage to such an extent that they will be indifferent between the deviating wage and the equilibrium wage.

The restriction implies the following features of the trade-off between the wage level and the matching probability. First, because $p(\theta)$ is a decreasing function, a worker who applies for a wage higher than an equilibrium wage expects to face a tighter market and, hence, a lower matching probability. Similarly, a firm that posts a wage higher than an equilibrium wage expects to increase its matching probability. Second, because $p(\theta)$ is differentiable, the restriction implies that the function $\theta(\cdot)$ is differentiable. Thus, the trade-off between the wage level and the matching probability is smooth.

To characterize the equilibrium, suppose $w^* \in E$, with $V^* = p(\theta(w^*))w^*$. Each firm takes V^* as given and chooses w to solve the following problem:

$$\max\ q(\theta(w))(y - w) \text{ s.t. } p(\theta(w))w = V^*.$$

Under the earlier assumptions on the function $q(\theta)$, the problem above is a concave problem and the solution is interior for all $V^* \in (0, y)$. Using the relationship $q(\theta) = \theta p(\theta)$, we can derive the first-order condition of the problem as $w^* / y = s(\theta)$. The equilibrium satisfies the Hosios condition!

As before, we can determine the tightness in the equilibrium by the entry condition, $J=c$. Then, the worker's indifference condition recovers V^*. It is easy to see that the market tightness is identical to the efficient one. Thus, the equilibrium is efficient.

The reason why the equilibrium is efficient can be related to hedonic pricing. With directed search, the market functions as if there is a price (in terms of wage) for every level of tightness. The inverse of the function $\theta(\cdot)$ serves as such a pricing function. Given this function, each worker chooses to apply for a wage level that maximizes his or her expected utility and each firm posts a wage to maximize expected profit. In the equilibrium, the market prices the tightness efficiently. That is, the increase in wage that a firm is willing to give for a marginal increase in the tightness is equal to the increase in wage that a worker asks for to compensate for a tighter market. As a result, the equilibrium internalizes search externalities. Because of this link to hedonic pricing, directed search is also called competitive search (see Moen, 1997).

Directed search can also induce the efficient amount of investment. Suppose that each firm chooses the level of capital before entering the labour market. Anticipating that the equilibrium wage will divide the match surplus efficiently, firms will choose the efficient level of capital.

Strategic formulation of directed search

In the above analysis, the matching function is a black box – it is specified exogenously as in models of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
undirected search. Because the matching function is important for the analysis of efficiency, it is important to derive a matching function from agents' strategic behaviour. Peters (1991) and Burdett, Shi and Wright (2001) formulate such a strategic game of directed search. The formulation also justifies the restriction above on the beliefs out of equilibrium. Let us describe the game where both u and v are fixed numbers greater than or equal to two. Competitive entry can be introduced in the same way as above. The one-period game is as follows. First, all firms post wages simultaneously. Each worker observes all firms' posted wages. (The essence of the model is the same if each worker can observe only two wages that are randomly drawn from posted wages.) Then, all workers choose the firms to which they apply. Assume that a worker can apply for only one job in the period, but the worker can use mixed strategy in the application. After receiving applicants, a firm randomly chooses one to be employed. Production takes place immediately and an employed worker is paid the posted wage. Then, the game ends.

There are many equilibria of this game that are asymmetric in the sense that identical agents do not use the same strategy. When $u=v=2$, for example, one asymmetric equilibrium is that one worker applies only to one firm and the other worker applies only to the other firm, while the two firms post zero wage. In this equilibrium, there is no unemployment – unemployment is eliminated by implicit coordination between the two workers. That is, a worker believes that the other worker will not apply for the same job as he or she does. Other asymmetric equilibria involve trigger strategies that also feature implicit coordination. Such coordination is unlikely to be attainable when there are many agents in the market. To emphasize the lack of coordination, we focus on the symmetric equilibrium, where all (identical) workers use the same mixed strategy to apply for the jobs. In this equilibrium, it is probable that two or more workers will apply for the same job, in which case some workers will be unemployed.

To characterize the equilibrium, consider a particular firm, called firm A. Suppose that other firms post wage w, but firm A posts wage x. If x is close to w, some workers will apply to firm A: if no other worker applied to firm A, a lone applicant to firm A would be employed with certainty, which would generate higher expected utility than applying to w. In fact, workers will increase the probability of applying to firm A until the expected utility from this application is the same as that from applying to other firms. Let a be the probability with which a worker applies to firm A. Then, firm A will receive one or more workers with probability $[1-(1-a)^u]$, and the expected number of applicants received by the firm will be ua. A worker who applies to firm A will be employed with probability $[1-(1-a)^u]/(ua)$. Because a worker's application probabilities across the firms must add up to one, a worker applies to each firm other than firm A with probability $\pi(a)=(1-a)/(v-1)$. The probability of employment in such a firm is $[1-(1-\pi(a))^u]/(u\pi(a))$. For a worker to be indifferent between firm A and other firms, the expected payoff must be the same from these firms. That is,

$$\frac{1-(1-a)^u}{ua}x=\frac{1-[1-\pi(a)]^u}{u\pi(a)}w.$$

This equation defines a smooth function $a=f(x,w)$. This function serves the same role as the tightness function did in the above formulation of directed search – it describes how a firm's wage offer will affect

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

workers' application, given other firms' wage offers. Note that f is an increasing function of x. Taking other firms' wage offers as given, firm A chooses x to solve:

$$\max\ (v-x)[1-(1-a)^{u}]\ \text{s.t.}\ \ a=f(x,w).$$

Denote the solution to this problem as $x=g(w)$.
A symmetric equilibrium is a wage level w such that $w=g(w)$. In this equilibrium, $a=\pi(a)=1/v$. The first-order condition of the above maximization problem, evaluated in the equilibrium, yields:

$$w=v\left[\frac{(1-1/v)^{-u}-1}{u/v}-\frac{1}{v-1}\right]^{-1}.$$

The formulation above reveals two features of a market with a finite number of agents. First, the number of matches generated in the equilibrium is $v[1-(1-1/v)^{u}]$. This matching technology exhibits decreasing returns to scale. The reason is that when the number of agents increases, the coordination failure becomes more severe, and so the number of matches per agent falls. Second, when a firm chooses its wage offer, it cannot take as given the payoff which a potential applicant can get by applying elsewhere. We have made this interdependence explicit with the notation $\pi(a)$. That is, when firm A raises the offer, it will attract all workers to apply for it with a higher probability, which will increase the probability of employment at other firms. For any given offer by other firms, a worker's payoff of applying to those firms will increase as a result of the wage increase by firm A. These two features complicate the analysis.
Fortunately, the complexity disappears in the limit when the market becomes infinitely large. Suppose that u and v approach infinity, with a fixed ratio $\theta=u/v$. Then, the matching probability is $(1-e^{-\theta})$ for a firm and $(1-e^{-\theta})/\theta$ for a worker. These matching probabilities have all the properties assumed above and, in particular, they are independent of the scale. Moreover, $u\pi(a)\rightarrow\theta$, which is independent of an individual firm's offer, x. The payoff to a worker who applies to a firm other than firm A is $w(1-e^{-\theta})/\theta$, which is also independent of x. In the limit economy, the equilibrium satisfies the Hosios condition and it is efficient.

Other pricing mechanisms and price dispersion

Price-posting is not the only mechanism to direct search. There are other mechanisms of directed search that can generate efficiency, such as auction (for example, Julien, Kennes and King, 2000). In contrast to price-posting, auction induces price dispersion. Thus, efficiency does not necessarily require identical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

workers to be paid the same wage in an economy with risk-neutral agents.

Consider the following game with first-price auctions. Each firm announces a reserve wage and the following scheme. If two or more workers participate in the firm's auction, the participants bid on the wage and the lowest bidder is employed at the bid wage; if two or more workers have the lowest bid, one of them is chosen randomly by the firm; if only one worker participates, the worker is paid the reserve wage. After observing all firms' announcements, workers choose the auction in which they will participate. A worker can participate in only one firm's auction, although the choice can be a mixed strategy.

Choose an arbitrary firm and call it firm *A*. Let *x* be the reserve wage announced by firm *A* and *a* the probability with which a worker participates in this firm's auction. For a worker who participates in firm *A*'s auction, there are two possible outcomes. The first is that the worker is the only participating worker, in which case the worker gets wage *x*. This outcome occurs with probability $(1-a)^{u-1}$. The other possibility is that the firm receives one or more other participants. In this case, the participants bid the wage down to 0. Thus, by participating in firm *A*'s auction, a worker expects to obtain a value $(1-a)^{u-1}x$. For firm *A*, there are also two cases. If only one worker participates in the firm's auction, profit is $(y-x)$. This case occurs with probability $ua(1-a)^{u-1}$. If two or more workers participate, profit is *y*. The probability for this case is $[1-(1-a)^{u}-ua(1-a)^{u-1}]$. Thus, the expected value (profit) of firm *A* is:

$$ua(1-a)^{u-1}(y-x)+[1-(1-a)^{u}-ua(1-a)^{u-1}]y.$$

For a firm other than firm *A*, let *r* be the reserve wage announced by the firm, $\pi(a)$ the probability of a worker's participation in the firm's auction, and $V(r, a)$ the expected value for a worker from such participation.

In order for a worker to be indifferent between firm *A*'s auction and other firms' auctions, the expected value must be the same; that is, $(1-a)^{u-1}x=V(r,a)$. Taking this condition as a constraint and taking other firms' auctions as given, firm *A* chooses *x* to maximize the expected profit above. Let $x=g(r)$ be the optimal choice. Then, a symmetric equilibrium is a reserve wage *r* such that $r=g(r)$.

As in the case of wage-posting, the characterization of the equilibrium is simplified in the limit economy where $u\to\infty$ and $\theta=u/v\in(0,\infty)$. In such a limit, we have $\pi(a)v\to 1$. Hence, $\pi(a)$ and $V(r,a)$ are independent of *a*. Solving the above maximization yields $r=y$. Thus, in contrast to directed search with wage posting, auction generates a wage differential. Some employed workers are paid their productivity but others are paid their reservation wage, 0.

Despite the dispersion of wages, the equilibrium is efficient. With risk-neutral agents, it is expected wage, rather than the actual wage, that is important for efficiency. With auction, the expected payoff is $ye^{-\theta}$ to a worker and $y[1-(1+\theta)e^{-\theta}]$ to a firm. These expected payoffs are the same as those in directed search with wage posting.

Related issues

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Risk aversion and asymmetric information: when workers are risk averse, different mechanisms of directed search can differ in efficiency. For example, price-posting generates lower risks in workers' income than auction. If the insurance market is imperfect, then wage-posting may give higher expected utility to workers than auction. Moreover, unemployment insurance, financed by lump-sum taxes, can improve welfare in this case (see Acemoglu and Shimer, 1999). On the other hand, price-posting is unlikely to be efficient when there is asymmetric information about the quality of goods or workers' productivity. Auction can allocate resources more efficiently in the presence of private information.
Heterogeneity and assortative matching: directed search models can be extended to allow workers to be heterogeneous. To achieve efficiency in such an extension, firms must rank different types of workers in addition to announcing wages. Heterogeneity can also appear on both sides of the market. In this case, an interesting question is whether the matching pattern is assortative, that is, whether similar attributes are matched with each other. In a frictionless economy, the efficient matching pattern is positively assortative, provided that the attributes on the two sides of the market are complementary. Moreover, the competitive equilibrium can implement the efficient matching outcome. When search frictions are introduced through undirected search, the equilibrium pattern of matches is neither assortative nor efficient (for example, Sattinger, 1995; Shimer and Smith, 2000). Introducing directed search can restore efficiency. However, the efficient matching pattern may be non-assortative when utility is transferable. There is a trade-off between the matching quality and the matching rate (for example, Shi, 2001; 2002).
Multiple applications: most search models assume that an agent on one side of the market can visit only one agent on the other side of the market in a given period; for example, a worker can apply only for one job at a time. This assumption may not be realistic in some markets. When workers can apply for multiple jobs simultaneously, there is a new source of the failure of coordination among firms: two firms may select the same worker and one of them will fail to obtain the worker. If the left-out firm has no recourse to other applicants it received, then the equilibrium is inefficient even with directed search. However, there are rules of selection, such as the one described by Gale and Shapley (1962), that can eliminate this difficulty of coordination and restore efficiency.
On-the-job search: search on the job is rarely examined in directed search models; sequential (undirected) search models have dominated the analysis on this topic (see Mortensen, 2003; 2007; labour market search). These models are constructed typically in continuous time. They assume that each worker, whether employed or not, receives a wage offer according to a Poisson process. While an unemployed worker receives one offer at a time, an employed worker effectively receives two offers – his current wage and the new offer from another firm. In this environment, the equilibrium must have a continuous distribution of wages with no mass point in the interior of the support; otherwise, a firm's payoff function would be discontinuous on the right side of the mass point. These models yield strong predictions on the shape of the wage distribution, some of which are counter-factual. Allowing on-the-job search to be directed can make the predictions more realistic and generate limited wage mobility endogenously, as shown by Delacroix and Shi (2006).

Search as a microfoundation for monetary theory

A surprising development of search theory is its use in monetary theory. For monetary economics, a fundamental question is why intrinsically useless objects, such as fiat money, can have a positive value

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in the equilibrium (see Wallace, 2007: fiat money). A familiar but informal answer is that such objects relieve the difficulty of exchange by acting as media of exchange. To capture this role of money, traditional monetary theory has used shortcuts while keeping the assumption of frictionless (Walrasian) markets. Examples include the requirement that agents must hold cash in advance of purchases and the assumption that money yields direct utility which cannot be generated by other assets. These short cuts seem incompatible with the Walrasian markets in the model and they are unable to explain why different media of exchange can have different values. To formalize the difficulty of exchange, Kiyotaki and Wright (1993) abandoned the short cuts and replaced the Walrasian exchange with random bilateral matches. The resulting model is a value theory of money, which gives money a role in improving efficiency of the market. Shi (1995) and Trejos and Wright (1995) integrate sequential bargaining into monetary search models to determine prices: see Shi (2006), Wright (2007) and matching for a survey.
Monetary theory has gone one step further to analyse optimal trading mechanisms. Using the method of mechanism design, the theory characterizes the set of allocations that are compatible with agents' incentives in the presence of search frictions. Next, the theory examines the efficient allocations and asks whether the implementation of these allocations entails particular types of trade, such as the use of money, banking, or a payments system – for example, Green and Zhou (2005). This analysis has clarified the relationship between optimal trading mechanisms and different components of search frictions, such as the difficulty for agents to meet, the difficulty for the society to keep record of agents' transactions, and the difficulty of enforcing trades.

Conclusions

Search theory was initially formulated to understand price dispersion and unemployment. Recent research has shifted the focus to the pricing mechanism and efficiency in frictional economies. Directed search is formulated to allow agents to explicitly make a trade-off between prices and matching frequency. The main finding is that directed search can restore efficiency that failed in earlier search models. However, even the efficient allocation cannot fully utilize all resources, because of the constraint of the matching technology. Moreover, the efficient allocation may not have the assortative pattern that emerges in a frictionless economy. The literature has explored different pricing mechanisms of directed search and used search to develop a microfoundation for monetary theory.
By focusing on pricing mechanisms and efficiency, the research has brought search theory close to the task of analysing the interactions between trades inside economic organizations and outside in the market. These interactions are important for explaining the observed forms of contracts and trading institutions. Monetary search theory has already taken up this task by using the approach of mechanism design. Other fields can also benefit from incorporating search frictions. An example is the literature on optimal dynamic contracts. By introducing search, one can endogenize the duration in which an agent stays with a particular contract. In general, the integration of search theory and contract theory awaits future research.
Another direction of future research is to explore the empirical implications of directed search. Most empirical work on wage distribution and business cycles has employed random search models (see Mortensen, 2003; Andolfatto, 2007; search models of unemployment). It is known that wage dispersion in these models is sensitive to the assumption of undirected search. Moreover, these models need a large

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
amount of heterogeneity among workers and firms to explain the observed distribution of wages, which seems to diminish the spirit of search. Directed search may offer a useful alternative approach in the empirical investigation.

See Also

- fiat money
- labour market search
- matching
- search models of unemployment
- search theory

Bibliography

Acemoglu, D. and Shimer, R. 1999. Efficient unemployment insurance. *Journal of Political Economy* 107, 893–928.

Andolfatto, D. 2007. Search models of unemployment. In *The New Palgrave Dictionary of Economics*, 2nd edn. ed. L. Cornell and S. Durlauf. London: Macmillan.

Burdett, K. and Mortensen, D. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39, 257–73.

Burdett, K., Shi, S. and Wright, R. 2001. Pricing and matching with frictions. *Journal of Political Economy* 109, 1060–85.

Cao, M. and Shi, S. 2000. Coordination, matching, and wages. *Canadian Journal of Economics* 33, 1009–33.

Delacroix, A. and Shi, S. 2006. Directed search on the job and the wage ladder. *International Economic Review* 47, 651–99.

Diamond, P. 1987. Search theory. In *The New Palgrave Dictionary of Economics*, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Gale, D. and Shapley, L. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69, 9–15.

Green, E. and Zhou, R. 2005. Money as a mechanism in a Bewley economy. *International Economic Review* 46, 351–71.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hosios, A. 1990. On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies* 57, 279–98.

Julien, B., Kennes, J. and King, I. 2000. Bidding for labor. *Review of Economic Dynamics* 3, 619–49.

Kiyotaki, N. and Wright, R. 1993. A search-theoretic approach to monetary economics. *American Economic Review* 83, 63–77.

Moen, E.R. 1997. Competitive search equilibrium. *Journal of Political Economy* 105, 385–411.

Mortensen, D. 2003. *Wage Dispersion: Why Are Similar People Paid Differently*. Cambridge: MIT Press.

Mortensen, D. 2007. Labour market search. In *The New Palgrave Dictionary of Economics*, 2nd edn. ed. L. Cornell and S. Durlauf. London: Macmillan.

Peters, M. 1991. Ex ante price offers in matching games: non-steady state. *Econometrica* 59, 1425–54.

Sattinger, M. 1995. Search and the efficient assignment of workers to jobs. *International Economic Review* 36, 283–302.

Shi, S. 1995. Money and prices: a model of search and bargaining. *Journal of Economic Theory* 67, 467–96.

Shi, S. 2001. Frictional assignment. 1. efficiency. *Journal of Economic Theory* 98, 232–60.

Shi, S. 2002. A directed search model of inequality with heterogeneous skills and skill-biased technology. *Review of Economic Studies* 69, 467–91.

Shi, S. 2006. A microfoundation of monetary economics. *Canadian Journal of Economics* 39, 643–88.

Shimer, R. and Smith, L. 2000. Assortative matching and search. *Econometrica* 68, 343–70.

Trejos, A. and Wright, R. 1995. Search, bargaining, money and prices. *Journal of Political Economy* 103, 118–41.

Wallace, N. 2007. Fiat money. In *The New Palgrave Dictionary of Economics*, 2nd edn. ed. L. Cornell and S. Durlauf. London: Macmillan.

Wright, R. 2007. Matching models. In *The New Palgrave Dictionary of Economics*, 2nd edn. ed. L. Cornell and S. Durlauf. London: Macmillan.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Shi, Shouyong. "search theory (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000505> doi:10.1057/9780230226203.1499

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

search theory

P. Diamond
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Search theory analyses resource allocation with imperfect technologies for informing agents of trading opportunities and for bringing together potential traders. This article examines some search theory analyses using a variety of special allocation mechanisms together with very simple preferences and production technologies. It explores models where information is gathered from visiting stores, and then looks at the changes that come from additional devices for information spread. It examines price setting both on a take-it-or-leave-it basis and resulting from idealized negotiations. While the presentation is in terms of a retail market, a similar approach has been useful in labour economics.

Keywords

advertising; bargaining; competitive equilibrium; equilibrium; job search; law of one price; Poisson distribution law; price reputation; reservation price; search costs; search theory; Walrasian analysis

Article

Walrasian analysis presumes that resource allocation can be adequately modelled using the assumption of instantaneous and costless coordination of trade. In contrast, Search Theory is the analysis of resource allocation with specified, imperfect technologies for informing agents of their trading opportunities and for bringing together potential traders. The modelling advantages of assuming a frictionless coordination mechanism, plus years of hard work, permit Walrasian analysis to work with very general specifications of individual preferences and production technologies. In contrast, search theorists have explored a variety of special allocation mechanisms together with very simple preferences and production technologies. Lacking more general theories, we examine the catalogue of analyses that have been completed.

Paralleling the Walrasian framework, we first examine individual choice and then equilibrium. There are a large number of variations on the basic search-theoretic choice problem. We explore one set-up in detail, while mentioning some of the variations that have been developed. Coordination of trade involves two separate steps: information gathering about opportunities, and arrangement of individual trades. One simple case is where information gathering is limited to visiting stores sequentially, combining the costs of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
collecting goods and of gathering information. Alternatively, there can be an information gathering mechanism which is independent of the process of ordering and receiving the good. We begin with models where the only information gathering is associated with visiting stores, and then look at the changes that come from additional devices for information spread.

Once two potential traders have met there are several ways of determining whether they trade and the terms of trade if they do. Among these are price setting on a take-it-or-leave-it basis, idealized negotiations where any mutually advantageous trade occurs at a price satisfying some bargaining solution, and more realistic negotiation processes that recognize the time and cost of negotiation, the possibility of a negotiating impasse, and the possible arrival of alternatives for one or the other of the trading partners. We explore the first two mechanisms.

One final distinction in the literature is between one-time purchases of commodities and ongoing trade relations. Infrequently purchased consumer goods are the classic example of the former, while the employment relationship is the classic example of the latter. Introducing on-going relationships permits the exploration of delayed learning of the quality of a match and associated rearrangements through quits and firings. Intermediate between these two cases is a situation such as that of frequently purchased consumer goods, where past trades facilitate future trades but do not bring about the closeness of an employment relationship. We discuss mainly the one-shot purchase. The discussion of individual choice and partial equilibrium will be given in terms of a consumer purchase. The parallel discussion of labour markets is only briefly mentioned.

I Individual choice

Consider a consumer in a store who is deciding whether to make a purchase or to visit another store with an unknown price. Denote by $U(p, 1)$ the utility that the consumer receives (net of purchase costs) if the purchase is made in the first store at a price equal to p. This assumes an ability to purchase the optimal number of units at a constant per unit price of p. If the purchase is made at the second store at price p, utility is $U(p, 2)$. This utility is less than $U(p, 1)$ because of the cost and the time delay from visiting a second store. We assume that the entire purchase is made at a single store, that it is impossible to return to the first store, and that there are no other stores that can be visited. Ignoring the possibility of making no purchase and no further searches, the alternative to purchasing in store 1 at price p, is a single visit to store 2 where the price will be drawn from a (known) distribution which we denote $F(p)$. The purchase should be made in store 1 if the utility of purchase there is at least as large as the expected utility of purchase in store 2:

$$U(p, 1) \geq \int U(p, 2)\,dF(p)\,. \tag{I–1}$$

As long as the consumer views the distribution of prices in store 2 as independent of the price in the first store, the rule in (I–1) yields a cut-off price, p^*, given by (I–2):

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$U(p^*, 1) = \int U(p, 2)\,dF(p).$$

(I–2)

For prices above p^*, optimal behaviour calls for visiting the second store, while for prices below p^*, optimal behaviour calls for making a purchase in the first store. Thus p^* is the cut-off price. Implicit in this formulation is the assumption that it is not desirable to make some purchase in store 1 and the remaining purchase in store 2. While this assumption is true for many consumer goods it is certainly not true for all of them. Without this assumption the decision resembles portfolio choice and has not been explored in the literature. A similar analysis applies to the search for high quality.

If the consumer does not know with certainty the distribution of prices in the second store, the consumer's beliefs about those prices may depend upon the price observed in the first store. We write the subjective distribution of prices in the second store, conditional on an observed price of p_1 in the first store as $F(p; p_1)$. The purchase should be made in the first store if p_1 satisfies the inequality:

$$U(p_1, 1) \geq \int U(p, 2)\,dF(p; p_1).$$

(I–3)

With no restriction on the beliefs of the consumer as to the structure of prices found in both stores, the set of prices resulting in a purchase in store 1 does not necessarily satisfy a cut-off price rule. For example, if either a high or a low observed price implies the same price in both stores, while an intermediate price in store 1 implies a low price in store 2, then the consumer should purchase in store 1 at the high and low prices but not at an intermediate price. Thus, the intermediate price might signal a price war. If the information content of the price found in store 1 is a greater likelihood of similar prices in store 2, the optimality of a cut-off price rule is restored (Rothschild, 1974). For the remainder of this entry we restrict analysis to the case of known distributions. The caveats implicit in this counter example should be kept in mind.

Returning to the set-up with a known distribution, we can increase the options of the shopper by adding the possibility of returning to the first store after observing the price in the second store. Denote by $U(p, 3)$ the utility that is realized if this option is followed. The utility function $U(p, 3)$ is less than $U(p, 2)$, which, in turn, is less than $U(p, 1)$. Once in the second store, the choice is between buying there and returning to the first store with both prices known. Therefore it pays to purchase in the first store in the first period if the price there, p_1, satisfies the inequality:

$$U(p_1, 1) \geq \int \max[U(p, 2), U(p_1, 3)]\,dF(p).$$

(I–4)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

That is, the purchase should be made in store 1 if utility there is higher than expected utility with optimal behaviour in choosing between the second store and returning to the first store. This is a particularly simple example of the backwards induction that can be used to solve the finite horizon sequential shopping problem. Behaviour in the first store, (I–4), again satisfies a cut-off price rule if the utility function has constant search costs and discount rate.

We now specialize the example by assuming additive, constant search costs c and utility discounting with a discount factor R. That is, $U(p, 2)$ equals $RU(p, 1) - c$, with $R \leq 1$. Returning to the choice problem without a return to store 1, we denote by $V(p_1)$ expected utility on observing the price p_1 in store 1, given optimal behaviour:

$$V(p_1) = \max\left[U(p_1),\ -c + R\int U(p)\mathrm{d}F(p)\right]. \tag{I–5}$$

The value of being in a store that has price p_1 is the larger of (i) the utility from making the optimal purchase at that price, and (ii) the expected utility if the search cost c is paid and the purchase is made in the second store. Using this function V, we can describe choice in the first period of a new three-period search problem with no return to previous stores. The optimal rule is to purchase if

$$U(p_1) \geq -c + R\int V(p)\mathrm{d}F(p). \tag{I–6}$$

That is, purchase is made in the first period if the achievable utility there is at least as large as that achievable with optimal behaviour, beginning with a visit to a randomly selected second store. The latter utility is the discounted expected optimized utility minus the search costs of the visit, recognizing that the second period choice is again a choice between a purchase and a search in the following period. By having $F(p)$ independent of p_1, we are sampling with replacement rather than sampling without replacement from the known distribution of prices. The choice rule given in (I–6) again shows cut-off price behaviour for period one choice. However, the cut-off price is higher in the second period than in the first because of the reduction in options as the end of the search process comes closer. Denoting the cut-off prices in the two periods by p_1^* and p_2^*, they satisfy the two equations:

$$U(p_1^*) = -c + R\int V(p)\mathrm{d}F(p)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(I–7a)

$$U(p_2^*) = -c + R\int U(p)\mathrm{d}F(p)$$

(I–7b)

$V(p)$ is at least as large as $U(p)$ since it represents the choice between purchase and searching again. Thus $p_1^* \leq p_2^*$, with a strict inequality in problems where the search cost and discount rate are not so large as to always imply a purchase in the current store.

There are additional reasons for cut-off prices to rise over time or equivalently, in a job search setting, for reservation wages to fall over time. In many settings, search costs rise over time. The utility of a purchase or of finding a job can fall over time. In the job setting, these can arise from declining wealth being used to finance consumption while searching for a job and from the shortening period over which any job might be held.

A known finite horizon for the end of search is incorrect in many settings. In addition, with many periods, the backwards induction optimization process is a cumbersome description of individual choice. Fortunately, the infinite horizon stationary case is easy to analyse. In this setting, a parallel analysis to that in (I–7) is a straightforward application of dynamic programming principles. With the assumption of a stationary environment the cut-off price is the same period after period. Denote by p^* the cut-off price and by V the optimized expected value of utility after paying the search cost to enter a store but before observing its price. Then V equals the utility of purchase if a purchase is made plus the probability of not making a purchase times the discounted optimized utility from facing the same problem one period later after paying search cost c:

$$V = \int_0^{p^*} U(p)\mathrm{d}F(p) + [1 - F(p^*)][-c + RV].$$

(I–8)

Solving (I–8) for V we have:

$$V = \frac{\int_0^{p^*} U(p)\mathrm{d}F(p) - c[1 - F(p^*)]}{1 - R[1 - F(p^*)]}.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(I–9)

The optimal p^* maximizes V and can be calculated by differentiation. More intuitively, we note that a purchase just worth making will give the same utility as will waiting to search again:

$$U(p^*) = -c + RV.$$

(I–10)

Rearranging terms, we can write the implicit equation for p^*:

$$(1-R)U(p^*) = -c + R\int_0^{p^*}[U(p) - U(p^*)]\mathrm{d}F(p).$$

(I–11)

Using this first order condition we can analyse the comparative statics of optimal search behaviour. Naturally, the cut-off price increases if the search cost increases or if the discount factor becomes smaller. Interestingly, an increase in the riskiness of the distribution of prices (holding constant mean utility from a randomly selected price, $\int U(p)\mathrm{d}F$) makes search more valuable and so lowers the cut-off price. This result follows from the structure of optimal choice – decreases in low prices make search more attractive while increases in high prices are irrelevant since no purchase is made at high prices. Analysis of the relationship between the expected number of searches and the distribution of prices is complicated since it depends on the shape of that distribution.

Thus far we have assumed that all stores are *ex ante* identical; that is, that a choice to search is a choice to draw from the distribution $F(p)$. In many problems one can choose where to search. In that case, one is choosing which distribution $F(p)$ to sample from or, if there are limited draws allowed from a particular distribution, the sequence of distributions from which prices should be sampled. Interestingly, the reservation prices which tell whether to purchase or to sample again from a given distribution also serve to rank distributions.

In the choice problem analysed so far we have used discrete time, with the arrival of one offer in each time period. There are two straightforward generalizations. First, one might have the opportunity to receive more than one offer in any period, with the number of offers received being a function of the chosen level of search costs. In this way one can model the choice of search intensity. Second, the process of attempting to locate stores might have a stochastic rather than a determinate time structure. The simplest such model has the arrival of purchase opportunities satisfying the Poisson distribution law. That is, at any moment of time there is a constant flow probability of an offer arriving, any such offer being an independent draw from the distribution of available prices. Let us denote by a, the arrival rate of these offers; and by c, the constant

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
search cost from being available to receive these offers. Utility is discounted at the constant (exponential) rate *r*. One can derive the optimal cut-off price and the optimized level of expected utility by analysing the discrete time process as above and passing to the limit. As a more intuitive alternative, let us think of the opportunity to purchase as an asset, where *V* now represents the value of that asset. The utility discount rate plays the role of an interest rate in asset theory. The asset is priced properly when the rate of discount times the value of the asset equals the expected flow of benefits from holding that asset. The expected flow of benefits is the gain that will come from making a purchase at a price below the cut-off price rather than continuing to search, adjusted for the probability of such an event, less search costs. Thus asset value satisfies

$$rV = a\int_0^{p^*} [U(p) - V]\,dF(p) - c. \tag{I–12}$$

It is worthwhile to make any purchase with a higher utility than that from continued search. Thus the cut-off price satisfies

$$rU(p^*) = a\int_0^{p^*} [U(p) - U(p^*)]\,dF(p) - c. \tag{I–13}$$

Again, one can introduce search intensity by having the Poisson arrival rate be a function of the search cost. In the equilibrium discussions below we will use the choice problem in the form (I–13).

So far we have ignored events after a purchase. In the labour setting this is equivalent to the assumption that taking a job is the end of search. In practice individuals frequently shift from job to job with no intervening period of unemployment. One can model job choice recognizing the possibility of continued search while working. Such an analysis must consider the rules that cover compensation between the parties in the event of a quit or firing, with no compensation and compensatory and liquidated damages being the situations analysed in the literature. The search for a better job is only one aspect of turnover. Also, one can model learning about the quality of match in a particular job as a function of the time on the job and the stochastic realization of experience. With a shadow value for quitting to search for a new job, one then has a second aspect of the theory of turnover.

The formulation of job taking given above has been combined with data on individual experience to examine empirically the determinants of the distribution of spells of unemployment. Since this essay focuses on equilibrium and the empirical literature has not examined the determinants of the distribution of opportunities, we do not explore this sizeable and interesting literature, nor the estimates of the effect of unemployment compensation on the distribution of unemployment spells. For an example, see Kiefer and Neumann (1979).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the model above we have assumed that no additional information is received during the search process. In practice, individuals are simultaneously searching for many different consumer goods and often for jobs and investment opportunities as well. The relations among search processes coming from the arrival of information and the random positions with simultaneous search for many different goods have not been explored in the literature. Focusing on search for a single good, we have added several new factors to the theory of demand, particularly the cost of attempting to purchase elsewhere and the knowledge and beliefs of shoppers about opportunities elsewhere. In practice, these are important determinants of demand.

II Equilibrium with bargaining

The theory of choice above is a simple version of the complex problem people face when making decisions about information gathering and purchases over time. That simplicity yields a choice theory that can be embedded in an equilibrium model. To complete an equilibrium model, we need to model the determination of two endogenous variables: the arrival rate of purchase opportunities and the distribution of their prices. In this section, we consider prices that satisfy the bargaining condition of equal division of the gains from trade. In the next section, we consider take-it-or-leave-it prices set by suppliers. In both cases we assume that there are no reputations either of soft bargaining or low price setting that affect the arrival rate of potential customers. We begin by assuming that all buyers are identical and all sellers are identical. This case brings out the role of search in determining the level of prices. Below we consider determinants of the distribution of prices.

Axiomatic bargaining theory relates the terms of trade to the threat points of the two bargainers and the shapes of their utility functions. To avoid complications from the latter, we assume that a single unit is purchased and utility from purchase equals a constant, u_d, minus the price paid. We also assume that each seller has a single unit to sell. The utility from a sale is the price received less the cost of the good, u_s. One might think of this as a homogeneous used car market. To divide equally the gains from trade, the differences between the utility position with the trade and the utility position without it are equalized for the two parties. The value of purchasing at price p is $u_d - p$; expected utility without a trade is V_d, the optimized expected utility from continued search. We restrict ourselves to an economic environment where all trades take place at the same price. With a degenerate distribution of prices, we can rewrite the value equation (I–12) as

$$rV_d = a_d(u_d - p - V_d) - c_d. \quad \text{(II–1)}$$

For suppliers, the utility from a sale is $p - u_s$. The gain from selling now rather than later is $p - u_s$, less the value of having a car for sale, V_s. The carrying cost of having a car available for sale can be incorporated in the search cost. The value equation for suppliers is

$$rV_s = a_s(p - u_s - V_s) - c_s. \quad \text{(II–2)}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

We ignore the sufficient conditions for search to be worthwhile $(V_d, V_s \geq 0)$.
Equal division of the gains from trade implies

$$u_d - p - V_d = \frac{r(u_d - p) + c_d}{r + a_d} = \frac{r(p - u_s) + c_s}{r + a_s} = p - u_s - V_s.$$
(II–3)

We have assumed the same utility discount rate for both parties. Thus we have a relationship between the equilibrium price, the arrival rates of trading opportunities, the search costs, and the utility from ownership. Solving (II–3) for the equilibrium price, we have:

$$p = \frac{(r + a_s)(ru_d + c_d) + (r + a_d)(ru_s - c_s)}{r(2r + a_s + a_d)}.$$
(II–4)

Without direct search costs $(c_d = c_s = 0)$, the position of the price between the seller's reservation price of u_s and the demander's reservation price u_d depends on the relative ease of finding alternative trading partners. As it becomes very easy to find buyers (a_s becomes infinite), the price goes to u_d. Alternatively, as it becomes very easy to find suppliers (a_d becomes infinite), the price goes to u_s. Furthermore, an increase in one's search cost pushes the price in an unfavourable direction. In this extremely simplified setting, (II–4) brings out the new element that search theory brings to equilibrium analysis, namely, the dependence of equilibrium prices on the abilities of traders to find alternatives. Implicit in Walrasian theory is the idea that a perfectly substitutable trade can be found costlessly and instantaneously. In this restricted sense, there is no consumer surplus in a Walrasian equilibrium.
To complete the theory we need to determine the two endogenous arrival rates of trading partners. Assuming a search process without history, these depend on the underlying technology for bringing together buyers and sellers and the stocks of buyers (N_d) and sellers (N_s). We write the arrival rates as $a_d \bullet (N_d, N_s)$ and $a_s \bullet (N_d, N_s)$. The two arrival rate functions satisfy the accounting identity between the numbers of purchases and of sales:

$$a_d(N_d, N_s)N_d = a_s(N_d, N_s)N_s.$$
(II–5)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Next, we must examine the determinants of the stocks of buyers and sellers. This theory can be based on the stocks of traders or the flows of new traders. One extreme example is that the steady state stocks of buyers and sellers are exogenous. One then inserts the functions a_d and a_s in the price equation (II–4).

An alternative extreme to perfect inelasticity is the assumption of perfectly elastic supplies of buyers and sellers at given reservation values for search, $\bar{V}_s$ and $\bar{V}_d$. Assuming reservation values that are consistent with the existence of equilibrium with positive trade, the equality of gains from trade (II–3) implies

$$p = \frac{u_d + u_s + \bar{V}_s - \bar{V}_d}{2}. \qquad \text{(II–6)}$$

The numbers of traders actively searching adapts to give this simple formula. Substituting from (II–6) in (II–1) and (II–2) we have the necessary values of a_d and a_s and so two equations for N_s and N_d.

For a market with professional suppliers one can consider the case of inelastic demand $(\bar{N}_d)$ and a perfectly elastic supply $(\bar{V}_s)$. If we assume further that demanders visit suppliers at a rate and cost independent of the number of suppliers, then a_d and c_d are parameters. Solving (II–3) for p in terms of the exogenous variables we now have

$$p = \frac{(r + a_d)(u_s + \bar{V}_s) + ru_d + c_d}{2r + a_d}. \qquad \text{(II–7)}$$

In this case the response of price to an increase in the cost of the good or the reservation utility of suppliers is $(r + a_d) / (2r + a_d)$, which is less than one. The speed of the search process relative to the interest rate determines the extent to which search equilibrium is different from Walrasian equilibrium. In a labour setting, an analogue to (II–7) shows how unemployment compensation affects wages by changing search costs.

Efficiency

There are two decisions implicit in the model above – whether to enter the search market and whether to accept a particular trade opportunity. The decision to enter a search market, like the choice of search intensity, affects the ease of trade of others. There is nothing in the process that determines prices which reflects the externalities arising from the impact of changed numbers on the opportunities to trade. Thus, in general, equilibrium will not be efficient and one has the possibility of both too much entry and too little entry.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In order to explore the efficiency of the choice of acceptable trades, we need a reason for waiting for a better deal in the future. This can be done by introducing differences in traders or differences in matches between preferences of demanders and goods on sale. However formulated, we have the proposition that the marginally acceptable trade generates no surplus to the two agents making that trade, yet the marginal trade changes the search environment of others. This involves externalities of the same kind as the entry decision already discussed. Again, in general, equilibrium is not efficient.

Individual differences

There are many patterns of differences among demanders in their evaluations of different goods. We explore two simple cases which have been dubbed quality differences and variety differences. With quality differences, all demanders have the same utility evaluation of goods. One asks how the price of a good varies with the quality of the good. With variety differences, all demanders have the same distribution of utility evaluations of the set of goods in the market, but demanders disagree as to which is better. There is then an issue of ‘matching’ preferences with goods. One asks how the price in a transaction varies with the quality of the match.

We use q as the index of universally agreed on quality, and denote by $p(q)$ the price paid in a transaction for a good of quality q. By suppressing all other differences, we have the same price in all the purchases of a good of any quality. We denote by $V_s(q)$ the optimized net value to a supplier of having a unit of quality q for sale. Paralleling (II–2), we can calculate the net gain to a supplier of selling his unit. This gain, $p(q) - u_s(q) - V_s(q)$, satisfies

$$p(q) - u_s(q) - V_s(q) = \frac{r[p(q) - u_s(q)] + c_s}{r + a_s}. \tag{II–8}$$

For the demander, we denote by V_d the value of entering the search market to make a purchase, and by $u_d(q)$ the utility, gross of purchase price, of purchasing a unit of quality q. Paralleling (I–12), the utility discount rate times the value of being a demander is equal to the net flow of gains from search. The gross flow of gains equals the arrival rate of purchase opportunities times the expected gain from a purchase. The expected gain is the utility of buying the good less the price that has to be paid for the good less the shadow value of being a searcher. Denoting the distribution of qualities in a randomly selected trade encounter by $F(q)$, the value of being a demander satisfies

$$rV_d = a_d \int [u_d(q) - p(q) - V_d] dF(q) - c_d. \tag{II–9}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

A full equilibrium analysis of this model would require determination of the distribution $F(q)$ as well as the arrival rates a_d and a_s. $V_s(q)$ would play an important role in determining $F(q)$. Such a model could consider investment in human capital with a search labour market. We will not carry out such an analysis, but focus merely on the relative prices $p(q)$, given a non-degenerate distribution $F(q)$. This problem is kept simple by the uniformity of product evaluations, which results in consumers' purchasing the first unit encountered, just as in the homogeneous case above. In any trade, the gains are shared equally between buyer and seller. Using (II–8) and (II–9) to eliminate V_d and $V_s(q)$ in the equal gain condition (II–3), we have the equilibrium price function

$$(2r + a_s)\,p(q) = (r + a_s)u_d(q) + ru_s(q) - c_s + (r + a_s) \times \left\{c_d - a_d \int [u_d(z) - p(z)]\,dF(z)\right\} / (r + a_d). \tag{II–10}$$

This generalization of the homogeneous case, (II–4), shows a price that rises with quality assuming that cost does.

$$p'(q) = \frac{(r + a_s)u_d'(q) + ru_s'(q)}{2r + a_s}. \tag{II–11}$$

The speed of search relative to the interest rate determines the magnitude of deviation from the Walrasian result that with identical demanders all transactions give the same utility level $[p'(q) = u_d'(q)]$. With pure quality differences, all consumers have the same expected utility from search, while suppliers have expected utilities which vary with the quality of goods for sale. In a symmetric variety model, both demanders and suppliers have the same expected utility from search. The variable q now represents the quality determined by the particular match of demander and good. We view the distribution of these qualities, $F(q)$, as given and the same for all demanders and all goods. Implicitly we are assuming random matching between demanders and different goods. It is now the case that a sufficiently poor match will not result in a trade. We denote by u_d•(q) the utility evaluation, gross of purchase price, of buying a good, by $u_s(q)$ the cost of supplying a good, and by $p(q)$ the price when the quality of a match is q. The value of search for a supplier satisfies

$$rV_s = a_s \int_{q_1}^{q_2} [p(q) - u_s(q) - V_s]\,dF(q) - c_s \tag{II–12}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where q_1 is the lower bound of match qualities at which it is mutually advantageous to carry out a trade. At the lowest acceptable quality, q_1, $p(q_1)$ is equal to $u_s(q_1) + V_s$. The value of search for a demander continues to satisfy (II–9). The assumption that all mutually advantageous trades are taken implies that q_1 also equates the gain from a purchase $u_d(q_1) - p(q_1)$ with the utility from search V_d. Equating the gains from trade for buyer and seller and solving for the price we have

$$2\,p(q) = u_d(q) + u_s(q) - V_d + V_s. \qquad \text{(II–13)}$$

Price increases with match quality to reflect the changed cost of supply, $u_s'(q)$, plus half the change in surplus, $[u_d'(q) - u_s'(q)]\,/\,2$.

Recapitulating our analysis of search equilibrium with bargaining, we have seen two themes. The first is how the search for trading partners introduces an additional element in the determination of trading prices: namely, the relative ease of the two potential trading partners in finding alternative trades. Secondly, the presence of a costly trade coordination mechanism is naturally replete with externalities as the availability of traders affects the trading opportunities of others.

In the model used in this section negotiation is instantaneous while search is slow. A fascinating literature explores equilibrium in models where the negotiation process is an explicit game of exchanging bids that can be interrupted by the arrival of an alternative trading partner (cf. Rubinstein and Wolinsky, 1985).

III Equilibrium with price setting

In contrast to the bargaining theory used above, we now assume that prices are set on a take-it-or-leave-it basis by suppliers. This rule of (not) bargaining over prices gives the supplier a potential for monopoly power. The search for alternatives limits this monopoly power. The fundamental question is how much. We begin with the assumption that the only source of price information is visiting randomly chosen suppliers sequentially one at a time. We assume many identical suppliers, implying equal profitability of different pricing strategies used in equilibrium. If all buyers have identical positive search costs and identical demand curves that yield a unique profit maximizing price, then the unique equilibrium is the price that would be set by a monopolist. This result assumes a sufficient number of suppliers that buyers will not search for a single low price. This extreme result comes from the uniformity of trading opportunities. The best a buyer can do is wait to make exactly the same deal in the future. Therefore a buyer is always willing to pay a little bit more today than he has to pay in the future. Thus the demand curve for an individual seller coincides with the underlying demand curve in the neighbourhood of the equilibrium price. Even though this result is limited to unrealistic cases, it is interesting that the price is independent of the cost and speed of search, as long as search is not costless and instantaneous.

Given the pervasive reality of price distributions in retail markets, it has been natural for the literature to concentrate on generating equilibria without uniform prices. With differences in demanders, either from differences in underlying characteristics or from differences in their history of past purchases, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
equilibrium can involve a distribution of prices and the structure of that distribution will depend upon search costs. In this case, consumers care about the characteristics of other consumers since these characteristics affect price setting behaviour. Similarly, with differences among suppliers the equilibrium price distribution varies with search costs.

Information gathering

When visiting a store is the only way to learn its price, price quotations are gathered one at a time. Separating the gathering of price information from going to stores does not necessarily change the model. If price quotations are still gathered one at a time, the cost of going to purchase the good can be deducted from the utility of acquiring it, leaving the model unchanged. However, the separation of the gathering of information from the collection of goods opens up the possibility of sometimes receiving price quotations one at a time and sometimes two or more at a time. This possibility destroys the single price equilibrium in the model of identical buyers and sellers. To see this result, note that profit per sale is continuous in price but, with uniform prices, the number of sales is discontinuous in price since a slight decrease in price wins all sales when a firm's price is one of two that are learned simultaneously. With positive profit made on each sale it would always pay to decrease price slightly below the uniform price of all other suppliers. With constant costs the competitive price is not a possible equilibrium either since a price increase gains profits when one is the only price quote while losing zero profit sales when one is not the only price quote. Thus there is necessarily a distribution of prices in equilibrium. Without price reputations, a store can choose any price it wants without affecting the flow of information about that store. Therefore, with identical firms the equilibrium will satisfy an equal profit condition. There will be low-price high-volume stores and high-price low-volume stores. One way to complete this model is to allow purchasers a choice of intensity of search which stochastically generates varying numbers of price quotations per period. We examine three additional models – price guides, advertising, and word of mouth.

Price guide

In this extension of the model we continue to have consumers seek price information one price at a time. In addition, consumers can purchase a guide to lowest cost shopping, with the purchase cost varying across consumers. A consumer who purchases such a guide is directed to one of the lowest price stores; a consumer who does not follows the search procedure described above. Assuming free entry of identical firms with U-shaped costs and an equilibrium where some consumers purchase the price guide and some do not but otherwise consumers are identical, we have a two-price equilibrium. Some of the stores set the price at the competitive equilibrium level. These stores sell to all consumers who purchase price information and those sequential shoppers who are lucky enough to find one of these stores on their first shopping visit. The remaining stores have higher prices, equal to the cut-off price for searching consumers or the profit maximizing price for selling to such a consumer, whichever is lower. The fraction of stores of the two kinds and the aggregate quantity of stores per consumer are determined by the zero profit condition for the two pricing strategies. When more consumers purchase the price guide, there will be more stores setting the competitive price and a drop in the cut-off price of searching consumers. This external benefit to searching consumers implies the inefficiency of the original equilibrium. A very slight subsidization of the cost of the price guide involves a second-order efficiency cost to the purchase of guides, no effect on firms (which have zero profits), and a first-order gain to searching consumers.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Advertising

It is obviously counterfactual to have all the information flows resulting from actions by shoppers. Advertising is a pervasive modern phenomenon. We continue to assume that stores have no price reputations. If the form of advertising is direct communication of prices to individual consumers, we can construct a model that again results in a distribution of prices. Stochastic communication from stores to consumers naturally generates a distribution of the number of price quotes that consumers receive. Any specific model of the stochastic structure of attempted communication will generate a distribution of numbers of price quotes learned by consumers. Free entry then implies a particular equilibrium distribution of prices provided some consumers receive a single price quotation and others receive more than one.

Word of mouth

It is natural to model both the seeking of price information and the spreading of price information as costly activities. However, some price information passes between consumers as a costless activity, part of the pleasure of discussing life. The presence of word-of-mouth communication in addition to sequential shopping alters equilibrium. The natural way to model word-of-mouth price communication brings price reputations into the model, since the prices set in one period affect communications about stores in future periods when their prices might be different. In order to isolate the effect of word of mouth we consider a very artificial model. Stores set prices which must hold for two periods. Consumers shop in the first or second period but are otherwise identical. In the first period, there is only sequential search, visiting stores one at a time as modelled above. Between the first and second periods there are random contacts between first-period shoppers and second-period shoppers. In this way, each second-period shopper receives information about the price in some positive number of stores. We assume that some people hear of only one store, while others hear of at least two. Then there will be a distribution of prices, with the structure of the distribution depending on the details of the word-of-mouth process. This analysis can be extended by having shoppers tell not only of the prices they paid, but also of prices they have heard from others. Both types of communication require a model of memory. The density of stores has different effects on equilibrium prices for different models of memory. This approach has been used in a setting of search for quality rather than low price to argue that doctors' fees can be higher where there are more doctors per capita (Satterthwaite, 1979).

Recapitulating the analysis of search equilibrium with price setting but not price reputation, we have seen two themes. One is the tendency for even low-cost search to generate sizeable amounts of monopoly power because of similar incentives for all suppliers. The second is a tendency for equilibrium to have a distribution of prices. Since price distributions are a widespread phenomenon in decentralized economies, it is reassuring that the theory produces such distributions.

IV Additional issues

We have considered the search analogue to competitive equilibrium. It was assumed that there were many small firms, whose behaviour was adequately approximated while ignoring their impacts on certain aspects of equilibrium. Search theory has also examined equilibria with small numbers of firms. It may pay a monopolist to have a distribution of prices across his outlets rather than a single price as a method of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
discriminating among consumers with different search costs, even though the need to search for a low price adds to the cost of purchase of the good (Salop, 1977). In a duopoly or oligopoly setting, it is natural to consider randomized pricing strategies which again give rise to a distribution of prices (Shilony, 1977). This may be one of the many factors that go into the empirical fact of sales by retail outlets.
The technology of shopping in the models above is extremely simple. Little has been done to marry the underlying search issues with some of the realities of the geographic distributions of consumers and firms and the normal travels of shoppers. Similarly, little has been done to model the search basis for the role of intermediaries.

Price reputations

All the models mentioned above omit or severely limit the inter-temporal links in profitability that arise from price reputations. This is a major hole in the existing literature. Probably significant progress in this area will have to await the discrimination of cases in which optimal strategies (whether determinate or stochastic) are stationary, from those in which optimal strategies involve building up a reputation which is then run down. In such a setting analysis will be very sensitive to the assumptions made about consumer knowledge both of existing prices and of price strategies followed by firms. It would be nice to have both an empirical evaluation of the level of consumer ignorance about opportunities, and a theoretical structure capable of examining the relationship between equilibrium and the extent to which consumers are accurately informed.

Conclusion

Walrasian theory assumes that consumers are perfectly informed about the prices of all commodities in the economy. This assumption is central for the law of one price, that a homogeneous commodity sells at the same price in all transactions in a given market. This assumption is also central for a variety of inequalities on prices, limiting price differences to be less than transportation costs. These inequalities are consequences of the absence of opportunities for arbitrage profits. In order to make a rigorous arbitrage argument, there must be simultaneous purchase and sale of the same commodity at different prices net of transportation costs. If the purchase and sale are at different times, there is likely to be risk for the would-be arbitrageur. Similarly, a proper arbitrage argument requires homogeneous commodities. It is improper to apply arbitrage arguments to labour markets for example, although migration arguments may lead to similar conclusions. In search theory with a known distribution of prices, there is a cost to finding any trading partner and possibly a large cost to finding one willing to trade at some particular price. This idea captures one aspect of the limitations on the extent of arbitrage arguments.
Realistically, one must recognize that infrequent traders are often ill-informed about the distribution of prices in the market. This introduces two important changes in the basic theory. One is that gathering information changes beliefs about the distribution of prices, as well as revealing the location of possible transactions. The second is the incentive created for sellers to find consumers whose beliefs make them willing to transact at high prices. The differences between the search for suckers and the hunt for the highest value use of resources has not been clearly drawn in the literature, yet this distinction is valid and important for evaluating the functioning of some markets. Search-based theory and empirical work have a long way to go until we have satisfactory answers to a number of allocation questions that are totally ignored in a Walrasian setting. Nevertheless, the theory has already shown how informational realities can

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
seriously alter the conclusions of Walrasian theory.

It would have been highly duplicative to have reviewed search theory of the labour market as well as that of the retail market. For a survey of labour search theory and a partial guide to the literature, see Mortensen (1984). Individual patterns of unemployment spells are the key empirical fact requiring revision of the Walrasian paradigm.

The failure of the profession, thus far, to produce a satisfactory integration of micro and macroeconomics based on the Walrasian paradigm (with or without price stickiness) raises the thought that such an integration might come out of search theory. For a presentation of this view and discussion of some applications of search ideas to macro unemployment issues, see Diamond (1984).

See Also

- exchange
- search theory (new perspectives)

Bibliography

Diamond, P. 1984. *A Search-Equilibrium Approach to the Micro Foundations of Macroeconomics*. Cambridge, Mass.: MIT Press.

Kiefer, N. and Neumann, G. 1979. An empirical job search model with a test of the constant reservation wage hypothesis. *Journal of Political Economy* 87, 69–82.

Mortensen, D. 1984. Job search and labor market analysis. In *Handbook of Labour Economics*, ed. R. Layard and O. Ashenfelter. Amsterdam: North-Holland.

Rothschild, M. 1974. Searching for the lowest price when the distribution is not known. *Journal of Political Economy* 82, 689–711.

Rubinstein, A. and Wolinsky, A. 1985. Equilibrium in a market with sequential bargaining. *Econometrica* 53, 1133–50.

Salop, S. 1977. The noisy monopolist: imperfect information, price dispersion and price discrimination. *Review of Economic Studies* 44, 393–406.

Satterthwaite, M. 1979. Consumer information, equilibrium industry price, and the number of sellers. *Bell Journal of Economics* 10(2), 483–502.

Shilony, Y. 1977. Mixed pricing in oligopoly. *Journal of Economic Theory* 14, 373–88.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Diamond, P. "search theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000056> doi:10.1057/9780230226203.1498

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

search-and-matching models of monetary exchange

Randall Wright
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We review a class of equilibrium search (matching) models that can be used to study the trading process and develop a theory of money as a medium of exchange. Developing such a theory is one of the longest-standing issues in economics, and search-based models provide a natural framework in which to formalize venerable stories about money facilitating trade by avoiding the double coincidence of wants problem in pure barter. The approach provides a compelling microfoundation for monetary theory: it is based on sound economic thinking going back to the classical economists, brought up to date with modern and rigorous methods.

Keywords

bargaining; barter; Bellman equation; commodity money; fiat money; matching; microfoundations; money; multiple equilibria; overlapping generations models; search-and-matching models of monetary exchange; trigger strategies

Article

In this article we review a class of equilibrium search (matching) models that can be used to study the trading process, and in particular to develop a formal theory of *money as a medium of exchange*. Developing such a theory is one of the longest-standing issues in economics, but it met with at best limited success prior to the development of search-based models, which provide a natural framework in which to formalize venerable stories about money helping to facilitate exchange.
These stories, going back to Smith, Jevons, Menger, Wicksell, and others – many of which are reprinted in Starr (1990) – concern a *double coincidence of wants problem* in bilateral exchange, as discussed below. Overlapping generations models (for example, Wallace, 1980) provide an alternative approach. Ostroy and Starr (1990) survey earlier attempts to develop microfoundations for monetary theory, including Jones (1976), which is similar in spirit if not detail to modern search models. There is not the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

space here to discuss the pros and cons of the various approaches, but it seems fair to say search and matching models now dominate the area.

Background

Diamond (1982) introduced a framework that, although it cannot be used directly, can be extended naturally to build microfoundations for monetary economics. In his model, a [0,1] continuum of infinitely lived agents interact in an economy where activity takes place in two distinct sectors: one for production and one for exchange. In the first sector, agents encounter potential production opportunities randomly over time according to a Poisson process with arrival rate α. Each opportunity yields a unit of output at cost $c \geq 0$, where c is random with CDF $F(c)$. Since c is observed before a production decision is made, given an opportunity, there is a reservation cost k such that agents produce if $c \leq k$. For now, these goods are indivisible, and agents can store at most one at a time.

All goods yield utility of consumption $u > 0$, except by assumption agents cannot consume their own output; hence they must trade. Traders with goods meet bilaterally in the exchange sector according to a Poisson process with arrival rate γ. Upon meeting they trade, consume, and return to production. Since all goods are the same, and indivisible, every meeting yields trade, and every trade is a one-for-one swap. Generally, $\gamma = \gamma(N)$ depends on the measure of agents in the exchange sector N. This is based on a matching technology that gives the number of agents who meet a partner per unit time as $m(N)$, with $m'(N)>0$, implying $\gamma(N) = m(N)/N$ for all $N>0$.

Let V_0 and V_1 be the value functions for producers and traders. The flow Bellman equation for a producer is

$$rV_0 = \alpha E\max\{V_1 - V_0 - c, 0\} = \alpha\int_0^k (k - c)\,dF(c),$$

where $k=V_1-V_0$. Similarly, for a trader

$$rV_1 = \gamma(N)(u + V_0 - V_1) = \gamma(N)(u - k).$$

(We focus on steady states; for dynamics, see for example Diamond and Fudenberg, 1989.)

In words, the flow value rV_0 equals the arrival rate of opportunities times the expected option value of switching from production to exchange, while rV_1 equals the arrival rate of meetings times the gain from trading and switching back. Combining these equations,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$rk = \gamma(N)(u - k) - \alpha\int_0^k (k - c)\,dF(c).$$

Given N this has a unique solution for k. Given k, in steady state, the flow of agents from production to exchange must equal the flow back,

$$(1 - N)\alpha F(k) = m(N).$$

An equilibrium is a pair (N, k) satisfying these last two equations. It is simple to derive results concerning existence, comparative statics, and so on. As Diamond emphasizes, under increasing returns in $m(N)$, if any non-degenerate equilibrium (one with production) exists then multiple such equilibria exist. Under constant returns, a unique non-degenerate equilibrium exists if parameters fall in a certain range – for example, u is not too low, r not too high, and so on. To complete our review of this basic model, notice that exchange is trivial, even though it is restricted to bilateral trade, because there is only one good (or, all goods are the same). To make money interesting we need to generalize this. (Diamond, 1984, took a short cut to getting money into the model with a cash-in-advance constraint. By changing the environment as we do below, we see this is not only uninteresting, it is unnecessary.)

To ease the presentation we first simplify the production process. Assume everyone is *always* in the exchange sector, and, instead of carrying goods around, they can produce whenever they meet someone, at deterministic cost $C \geq 0$. Now, following Kiyotaki and Wright (1991; 1993), assume goods come in varieties, say colours. Each agent produces a particular colour, but different agents like to consume different colours. The simplest specification assumes agents get $u=U>C$ from any good in some set, and $u=0$ from other goods, and x is the probability output in the relevant set (that is, an agent wants what the other agent can produce) in any random meeting. Also, since agents can produce whenever they want, to simplify things we assume goods are non-storable.

When goods *are* storable, Kiyotaki and Wright (1989) determine endogenously which objects serve as media of exchange, potentially including commodity plus fiat money. That model illustrates the trade-off between fundamental properties like storability and equilibrium properties like acceptability. It has many implications – for example, there can be multiple equilibria with different monies, objects with bad fundamental properties may end up as money, and so on. Generalizations and applications of the model include Marimon, McGrattan and Sargent (1990); Aiyagari and Wallace (1991; 1992); Kehoe, Kiyotaki and Wright (1993); Wright (1995); and Duffy and Ochs (1999). Here, by making goods non-storable, we focus on determining how an economy operates when there is a single candidate medium of exchange, namely, fiat money.

On the assumption that exchange requires mutual agreement, which occurs when I want to consume your good and you want to consume mine, trade now occurs only in a meeting with probability x^2, at least if the event that I want your good is independent of the event that you want mine (see below). This captures nicely the famous *double coincidence problem* with direct barter: trade requires meeting

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

someone who produces something you like – which would be a coincidence – and also likes what you produce – a double coincidence. Payoffs are given by $rV_B = \gamma x^2(U - C)$, where the subscript on V_B stands for 'barter'. If x is small, which is the case if there is a lot of specialization, double coincidence meetings are rare and V_B is very low.

But is it really necessarily the case that trade occurs iff both parties want to consume what the other produces? Following ideas in Kocherlakota (1998), suppose agents get together at the start of time and discuss when to trade. Clearly, they agree that whenever *either* agent wants what the other produces he should get it, since this maximizes *ex ante* welfare

$$rV_C = \gamma\left[x^2(U - C) + x(1 - x)U - (1 - x)xC\right] = \gamma x(U - C),$$

where the subscript on V_C stands for 'cooperation' (or perhaps 'commitment' or 'credit'). As long as $x<1$, $V_C>V_B$. However, suppose agents cannot commit now to do things when they meet later that are not in their interest at that time. Then trades must satisfy incentive compatibility (IC), the binding condition being that you should be willing to produce in meetings where you do not consume.

If we can keep a public record of all agents' behaviour, we can try to use *trigger strategies* to support cooperative trade as follows: instruct agents to cooperate as long as everyone else does; but if anyone deviates, trigger to … 'something bad'. One can argue the worst trigger is 'autarky' which yields $V_A=0$; or it may be 'barter' which yields V_B. In the former case the relevant IC condition is $-C + V_C \geq V_A$, which simplifies to $rC \leq \gamma x(U - C)$; in the latter case it is $-C + V_C \geq V_B$, which simplifies to $rC \leq \gamma x(1 - x)(U - C)$. In either case, if r is small we can sustain cooperative trade. Moreover, one can prove formally that money has no role here (Kocherlakota, 1998; Wallace, 2001); instead of proving this here we move to models where money does have a role.

First-generation search models of money

Suppose it is difficult to use triggers because, say, there is incomplete monitoring or record keeping, or, to take the simplest situation, suppose agents have no memory – they just cannot recall what happened in previous meetings! Kocherlakota (1998), Kocherlakota and Wallace (1998), Wallace (2001), Corbae, Temzilides and Wright (2003), Araujo (2004), and Aliprantis, Camera and Puzzello (2007) explore less extreme variations, but our assumption allows us to make the point more easily. In our 'memoryless' world, your continuation payoff V_M cannot depend on what you do in a given meeting. Hence, the relevant constraint to get you to produce without consuming is $-C + V_M \geq V_M$, which is violated for any $C>0$. There is no scope for using threats to sustain cooperation without memory (generally, there is limited scope when memory is imperfect, which is what we need; we use the starkest case merely for tractability).

Suppose we introduce into this world a new object called fiat *money*. By definition, a medium of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
exchange is an object that is accepted in trade not to be used for consumption – or production – but to be traded again later for something else. When an object serving as a medium of exchange is for some people at some times a consumption good, it is called commodity money. When an object with no consumption value serves as a medium of exchange it is fiat money. At the start of time, we endow a fraction M of the population each with m=1 unit of money and the rest with m=0. Initially, those with m=0 can produce; after this, agents can produce after they consume but not before. This implies agents with money cannot produce, and at any point in time everyone either has m=1 or m=0. Now, even without memory, agents have an option other than pure barter: offer money for goods. Let Π be the probability a random producer accepts such an offer, and let π be your best response.
If V_m and V_p are the value functions of agents with and without money,

$$rV_p = \gamma(1-M)x^2(U-C) + \gamma Mx\pi(V_m - V_p - C) rV_m = \gamma(1-M)x\Pi(U + V_p - V_m).$$

For example, rV_p equals the arrival rate of agents with goods $\gamma(1-M)$, times the double coincidence probability x^2, times the gain from barter $U–C$; plus the arrival rate of agents with money γM, times the probability of trade $x\pi$, times the gain $V_m–V_p–C$. We restrict attention to pure strategies (mixed strategy equilibria are not robust here; see Shevchenko and Wright, 2004). Then the best response condition is π=0 if $V_m–V_p<C$ and π=1 if $V_m–V_p>C$. It is easy to see π=0 is always an equilibrium, and π=1 is an equilibrium iff

$$rC \le \gamma(1-M)x(1-x)(U-C).$$

Naturally, π=0 is an equilibrium – if no one else accepts money, why would you? It is more interesting that π=1 can be an equilibrium, since then intrinsically worthless money is valued, as a medium of exchange. Given M, one can check π=1 yields higher payoffs than π=0. Alternatively, if we choose M to maximize welfare, one can check M>0 iff x is not too big. Hence, introducing money can improve welfare, even given the assumption that money holders cannot produce. The convention of money as a medium of exchange is good because it eases trade. Now, π=1 is only an equilibrium when r is not too high, and one can check the cutoff for r here is more stringent (and payoffs lower) than when we had memory and triggers – that is, money is not a perfect mechanism. One reason money is not as good as memory is the *random* nature of matching. The problem is that you might, for example, have two meetings in a row where you want a good from someone who does not want your good, and in the second one you will have run out of money (this can also happens with positive probability when we relax the upper bound of unity on money holdings). However, in an *endogenous* (rather than random) matching model this never happens – when you have no money you do not go to someone whose good you like, but to someone who likes your good; see Corbae, Temzilides and Wright (2003). Still, money

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

can do pretty well here, and if we cannot use triggers it is the only way to improve on pure barter. The model is obviously crude, yet it gets at the essence of money. To recap, the results assume the following explicit frictions: (*a*) a double coincidence problem (generated here by random bilateral matching, although there are other devices in the related literature); (*b*) imperfect commitment; and (*c*) imperfect memory (or anything else that makes it difficult to use triggers). These frictions are severe – but no one said it was going to be easy to get money into economic theory in an interesting way. There are many extensions and applications of this model (some of which are surveyed in Rupert et al., 2000), but in the interest of space, we now move on to models where prices are endogenous. We mention one extension to endogenous specialization in Kiyotaki and Wright, 1993, based on ideas in Adam Smith. Consider the case where the probability that someone accepts your good x is a *choice variable*: if you want a large fraction of the population to like your output, you cannot specialize too much, which reduces productivity. Thus, the arrival rate in the production sector – to return to Diamond's two sector set-up – is $\alpha(x)$, with $\alpha' < 0$. When choosing x, you take the average X as given, and in equilibrium $x=X$. Two results follow. First, monetary equilibria have lower x than non-monetary equilibria, so the use of money enhances specialization and productivity. Second, $x \to 0$ as $\gamma \to \infty$, so when frictions vanish, agents specialize completely, and since the double coincidence probability is x^2, barter completely disappears.

Second-generation models

Suppose that goods are no longer indivisible, but can be consumed and produced in any amount $q \geq 0$, which yields utility $U(q)$ and disutility $-C(q)$, respectively. These functions have all the usual properties, plus $C(0)=U(0)=0$. We maintain for now the assumptions that money is indivisible, money holders cannot produce, and everyone holds $m \in \{0,1\}$. But we relax the assumption of independence in generating the double coincidence problem: the probability that I like your good is x, but now the probability that I like your good and you like mine is y, and not necessarily x^2, in general. (Consider N goods and N types, where type n produces good n, but likes good $n+1$ modulo N. If $N=2$ then $x=y=1/2$ (if I like your good you must like mine), while if $N \geq 3$ then $x=1/N$ and $y=0$ (if I like your good you cannot like mine). It is only under independence, which does not hold in these examples, that we necessarily have $y=x^2$.)
Conditional on money being accepted ($\pi=1$), we have

$$rV_p = \gamma(1-M)y[U(\hat{q}) - C\hat{q})] + \gamma Mx[V_m - V_p - C(\bar{q})] \quad rV_m = \gamma(1-M)x[U(\bar{q}) + V_p - V_m],$$

where $\hat{q}$ is the amount traded in barter and $\bar{q}$ the amount traded for money. It facilitates the presentation to start with the case $y=0$ and then give general results. Now, to determine the equilibrium value of money, as in Shi (1995) or Trejos and Wright (1995), we say the following: when I meet you and want your good, if I have $m=1$ while you have $m=0$, we *bargain* over the q you produce for my money, taking as given $\bar{q}$ in all other meetings. Equilibrium is a fixed point, $q = \bar{q}$, and the price level is $p=1/q$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

One can use any bargaining solution, including generalized Nash

$$\max\,[U(q) + V_p - V_m]^{\theta}[V_m - V_p - C(q)]^{1-\theta},$$

for any $\theta \in (0,1)$; we use $\theta = 1$ because it is so easy. (See Rupert et al., 2001, for $\theta \in (0,1)$ and other generalizations. Alternatives to bargaining studied in versions of this model include posting – for example, Curtis and Wright, 2004 – and auctions – Julien, Kennes and King, 2007. Or, instead of imposing a particular pricing mechanism, one can study the entire set of incentive-feasible trades – Wallace, 2001.) When $\theta = 1$, agents with m=0 get no gains from trade since they have no bargaining power (and y=0). Hence V_p=0, V_m=$C(q)$, and the Bellman equation for V_m reduces to

$$rC(q) = \gamma(1 - M)x[U(q) - C(q)].$$

This is 1 equation in q, with two solutions: 0 and a unique $q > 0$. Again, we get equilibrium where an intrinsically worthless object is valued as a medium of exchange. It is easy to do comparative statics, welfare analysis, and so on in this model (for example, it is immediate that q falls and p rises when M or r increase).

Once we reintroduce some barter – that is, once we allow $y > 0$ – one can show that, in addition to q=0, generically there either exist two equilibria with $q > 0$ or no equilibrium with $q > 0$. If y is small then equilibrium with q>0 always exists. It is not much harder to analyse the general case $\theta \in (0,1)$. This model has a large number of variations, extensions and applications – too many to review here (again see Rupert et al., 2000). Suffice it to say that the basic results of first-generation models more or less go through, with additional insights concerning prices.

Third-generation models

The approach sketched above provides a compelling microfoundation for monetary theory: it is based on sound economic thinking going back to some very famous economists, brought up to date with modern and rigorous methods and ideas. Still, obviously those first- and second-generation models are quite abstract and quite special. In particular, the assumption that agents hold $m \in \{0,1\}$ is severe, and precludes using the models for much quantitative and policy analysis. The difficult part of relaxing this and allowing any $m \geq 0$ is that we need to keep track of the distribution of m across agents, which is complicated by the random nature of matching and the endogenous amount of money spent in each match. There are several ways to deal with this problem. Some analytic results are available in Green and Zhou (1998) and Camera and Corbae (1999) for example, while computation methods are used by Molico (2006).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Another approach is to amend the environment to get around this problem while hopefully maintaining the spirit and essence of the matching models outlined above. There are two main ways to do this, following either Shi (1997) or Lagos and Wright (2005). The Shi model assumes the fundamental decision makers are not individuals, but families, each with a large number of members. If the individual members experience independent random meetings, when they return to the household at the end of each period the total amount of money in the family is pinned down by the law of large numbers. Hence, each household starts the next period with the same (deterministic) amount of money. There are many extensions and applications of this framework (see Shi, 2006, for some references).

The Lagos–Wright model alternatively assumes that at the end of each round of decentralized trade agents go to a centralized market where they can (among other things) rebalance their money holdings. On the assumption of quasi-linear utility, all agents choose the same *m* for next period, independent of the amount with which they start. Again, agents enter each round of decentralized trade with the same *m* here, just as in the family model (although there are several interesting differences between the approaches). Versions of either model are easily used for quantitative and policy analysis. These models are perhaps still special, since they use ‘tricks’ to harness the distribution of *m*, but this is merely for technical convenience in deriving analytic results. If one is willing to use a computer, most of the special assumptions can be avoided (see, for example, Chiu and Molico, 2006).

Conclusion

We have reviewed several generations of search-and-matching models of the exchange process that can be used to provide microfoundations for monetary economics. While the literature is big, and growing fast, it is to be hoped that this article conveys some of the main ideas and models in an accessible fashion.

See Also

- bargaining
- barter
- commodity money
- fiat money
- inside and outside money
- money
- money, classical theory of
- money and general equilibrium
- search theory
- search theory (new perspectives)

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Aiyagari, S.R. and Wallace, N. 1991. Existence of steady states with positive consumption in the Kiyotaki Wright model. *Review of Economic Studies* 58, 901–16.

Aiyagari, S.R. and Wallace, N. 1992. Fiat money in the Kiyotaki Wright model. *Economic Theory* 2, 447–64.

Aliprantis, C.D., Camera, G. and Puzzello, D. 2007. Anonymous markets and monetary trading. *Journal of Monetary Economics* 54, 1905–28.

Araujo, L. 2004. Social norms and money. *Journal of Monetary Economics* 51, 241–56.

Camera, G. and Corbae, D. 1999. Monetary patterns of exchange with search. *International Economic Review* 40, 985–1008.

Chiu, J. and Molico, M. 2006. Liquidity, redistribution and the welfare cost of inflation. Mimeo, Bank of Canada.

Corbae, D., Temzilides, T. and Wright, R. 2003. Directed matching and monetary exchange. *Econometrica* 71, 731–56.

Curtis, E. and Wright, R. 2004. Price setting, price dispersion and the value of money, or the law of two prices. *Journal of Monetary Economics* 51, 1599–621.

Diamond, P.A. 1982. Aggregate demand management in search equilibrium. *Journal of Political Economy* 90, 881–94.

Diamond, P.A. 1984. Money in search equilibrium. *Econometrica* 52, 1–20.

Diamond, P.A. and Fudenberg, D. 1989. Rational expectations business cycles in search equilibrium. *Journal of Political Economy* 97, 606–19.

Duffy, J. and Ochs, J. 1999. Emergence of money as a medium of exchange: an experimental study. *American Economic Review* 89, 847–77.

Green, E. and Zhou, R. 1998. A rudimentary model of search with divisible money and prices. *Journal of Economic Theory* 81, 252–71.

Jones, R. 1976. The origin and development of media exchange. *Journal of Political Economy* 84, 757–75.

Julien, B., Kennes, J. and King, I. 2007. Bidding for money. *Journal of Economic Theory*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kehoe, T.J., Kiyotaki, N. and Wright, R. 1993. More on money as a medium of exchange. *Economic Theory* 3, 297–314.

Kiyotaki, N. and Wright, R. 1989. On money as a medium of exchange. *Journal of Political Economy* 97, 927–54.

Kiyotaki, N. and Wright, R. 1991. A contribution to the pure theory of money. *Journal of Economic Theory* 53, 215–35.

Kiyotaki, N. and Wright, R. 1993. A search theoretic approach to monetary economics. *American Economic Review* 83, 63–77.

Kocherlakota, N. 1998. Money is memory. *Journal of Economic Theory* 81, 232–51.

Kocherlakota, N. and Wallace, N. 1998. Optimal allocations with incomplete record keeping and no commitment. *Journal of Economic Theory* 81, 272–89.

Lagos, R. and Wright, R. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 113, 463–84.

Marimon, R., McGrattan, E.R. and Sargent, T.J. 1990. Money as a medium of exchange in an economy with artificially intelligent agents. *Journal of Economic Dynamics and Control* 14, 329–73.

Molico, M. 2006. The distribution of money and prices in search equilibrium. *International Economic Review* 47, 701–22.

Ostroy, J.M. and Starr, R.M. 1990. The transaction role of money. In *Handbook of Monetary Economics*, ed. B. Friedman and F. Hahn. Amsterdam: North-Holland.

Rupert, P., Schindler, M., Shevchenko, A. and Wright, R. 2000. The search-theoretic approach to monetary economics: a primer. *Federal Reserve Bank of Cleveland Review* 36, 10–28.

Rupert, P., Schindler, M. and Wright, R. 2001. Generalized search theoretic model of monetary exchange. *Journal of Monetary Economics* 48, 605–22.

Shevchenko, A. and Wright, R. 2004. A simple model of money with heterogeneous agents and partial acceptability. *Economic Theory* 24, 877–85.

Shi, S. 1995. Money and prices: a model of search and bargaining. *Journal of Economic Theory* 67, 467–96.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Shi, S. 1997. A divisible search model of fiat money. *Econometrica* 65, 75–102.

Shi, S. 2006. Viewpoint: a microfoundation of monetary economics. *Canadian Journal of Economics* 39, 643–88.

Starr, R.M. 1990. *General Equilibrium Models of Monetary Economics*. New York: Academic Press.

Trejos, A. and Wright, R. 1995. Search, bargaining, money and prices. *Journal of Political Economy* 103, 118–41.

Wallace, N. 1980. The overlapping generations model of fiat money. In *Models of Monetary Economies*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Wallace, N. 2001. Whither monetary economics. *International Economic Review* 42, 847–70.

Wright, R. 1995. Search, evolution, and money. *Journal of Economic Dynamics and Control* 19, 181–206.

How to cite this article

Wright, Randall. "search-and-matching models of monetary exchange." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_M000397> doi:10.1057/9780230226203.1496

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

seasonal adjustment

Svend Hylleberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The main objective of seasonally adjusted time series is to provide easy access to a common time series data-set purged of what is considered seasonal noise. Although the application of officially seasonally adjusted data may save costs, it may also imply less efficient use of the information available, and data may be distorted. Hence, in many cases, seasonality may need to be treated as an integrated part of an econometric analysis. In this article we present several ways to integrate seasonal adjustment into econometric analysis in addition to applying data adjusted by the two most popular adjustment methods.

Keywords

ARIMA models; ARMA models; autoregressive models; band spectrum regression; band-pass filters; basic structural model (BSM); Box–Jenkins model; cointegration; common seasonal features; evolving seasonals model; habit persistence; HEGY test; Henderson moving averages; Jevons, W.; Kalman filter; long memory; maximum likelihood; noise models; ordinary least squares (OLS); periodic autoregressive model (PAR); periodic cointegration; production smoothing; real business cycles; seasonal adjustment; seasonal cointegration; seasonal difference filter; seasonal dummies; time series models; TRAMO/SEATS seasonal adjustment programme; transformations; univariate seasonal models; unobserved components (UC) models; vector autoregressions; X-11 seasonal adjustment programme

Article

Seasonal adjustment of economic time series dates back to the 19th century and it is based on an attitude properly expressed by Jevons, who wrote:

> Every kind of periodic fluctuation, whether daily, weekly, monthly, quarterly, or yearly, must be detected and exhibited not only as a subject of a study in itself, but because we must ascertain and eliminate such periodic variations before we can correctly exhibit those which are irregular or non-periodic, and probably of more interest and importance. (1884, p. 4)

The most popular model behind seasonal adjustment in the beginning of the 20th century was either the so-called additive unobserved components (UC) model

$$X_t = T_t + C_t + S_t + I_t, t = 1, 2, \ldots, n \quad (1)$$

where the observed series X_t is divided into a trend component, T_t, a business cycle component, C_t, a seasonal component, S_t, and an irregular component, I_t, or the multiplicative UC model

$$X_t = T_t * C_t * S_t * I_t, t = 1, 2, \ldots, n \quad (2)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which is applied if the series is positive and the oscillations increase with the level of the series.
The definitions of the individual components could vary, but Mills (1924, p. 357) defined the trend component, T_t, as the smoothed, regular, long-term movement of the series X_t, while the seasonal component, S_t, contains fluctuations that are definitely periodic in character with a period of one year, that is, 12 months or 4 quarters. The business cycle component, C_t, is less markedly periodic, but characterized by a considerable degree of regularity with a period of more than one year, while the irregular component, I_t, has no periodicity. A detailed description of the historical development is given in Hylleberg (1986).
The rationale behind seasonal adjustments is that the unobserved components model is useful, that the components are independent, and that the components of interest are the trend and cycle components.
The assumption of independence is highly questionable, as the actual economic time series is a result of economic agents' reaction to some exogenous seasonally varying explanatory factors such as the climate, the timing of religious festivals and business practices. For typical economic agents, decisions designed to smooth seasonal fluctuations will naturally interact with non-seasonal fluctuations, since the costs of such smoothing will necessarily be interrelated through budget constraints and so forth. Therefore, not only is the independence assumption economically unreasonable, but seasonal patterns may be expected to change if economic agents change their behavioural rules.
Hylleberg (1992, p. 4) defines seasonality as

> A systematic, although not necessarily regular, intra-year movement caused by the changes of the weather, the calendar, and timing of decisions, directly or indirectly through the production and consumption decisions made by agents of the economy. These decisions are influenced by endowments, the expectations and preferences of the agents, and the production techniques available in the economy.

Such a view of seasonality is somewhat different from the views expressed by most statistical data-producing agencies. The views of the statistical offices are well represented by the arguments put forward by OECD (1999, p. vii), where the implied definition of seasonality stresses the fixed timing of certain events during the year. Likewise, they indicate that the reason for changes in the seasonal pattern is 'the trading day effect', that is, the changing number of working days in a month, the changing number of Saturdays, and movable feasts such as Easter, Pentecost, Chinese New Year and Korean Full Moon Day. Obviously, such factors do influence the seasonal pattern in economic time series, but in the longer run technical progress and economic considerations based on these will imply changes in the seasonal pattern as well. In addition, the seasonal economic time series may constitute an invaluable and plentiful source of data for testing theories about economic behaviour, as the seasonal pattern is a recurrent though changing event where the pattern, despite the changes, is somewhat easier to forecast than many other economic phenomena. (For a general discussion of seasonality and the literature, see Hylleberg, 1986. For a presentation and discussion of the results since then, see Hylleberg, 1992; Franses, 1996; Ghysels and Osborn, 2001; Brendstrup et al., 2004.)
Seasonal adjustment and treatment of the seasonal components may in practice be undertaken in two ways: simply applying the seasonally adjusted data produced by the statistical agencies, or integrating the modelling and adjustment into the econometric analysis undertaken.

Officially applied seasonal adjustment programmes

Several different methods for seasonal adjustment are in actual use, but the most popular programme is the X-12-ARIMA seasonal adjustment programme (see Findley et al., 1998) which is a further development of the popular X-11 seasonal adjustment programme (see Shiskin, Young and Musgrave, 1967; Hylleberg, 1986). Another popular seasonal adjustment programme is the TRAMO/SEATS programme developed in Gomez and Maravall (1996).

X-12-ARIMA seasonal adjustment programme

The main characteristics of the X-11 seasonal adjustment method for the monthly multiplicative model (see (2)),

$$X_t = TC_t * S_t * TD_t * H_t * I_t, \tag{3}$$

where TC_t is the combined trend-cycle component, while TD_t is the trading day component, and H_t the holiday component, is the repeated application of selected moving averages such as a 12-month centred moving average to estimate TC_t followed by an actual extraction of the estimated trend-cycle component. The extraction by the moving average filters takes place after a prior

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

adjustment for trading days and certain holidays, and a varying seasonal pattern is taken care of by applying so-called Henderson moving averages with a 9, 13 or 23 number of terms. The Henderson trend filters are used in preference to simpler moving averages because they can reproduce polynomials of up to degree 3, thereby capturing trend-turning points.

In addition, treatment for so-called extreme observations was possible, and a refined asymmetric moving averages filter is used at the ends of the series.

In order to robustify the initial seasonally adjusted series against data revisions, the X-11 seasonal adjustment method was improved by extending the series by forecasts and backcasts from an ARIMA model before seasonally adjusting the series (see Dagum, 1980).

The X-12-ARIMA seasonal adjustment programme described in Findley et al. (1998) extends the facilities of X-11-ARIMA by adding a modelling module denoted RegARIMA, which not only facilitates modelling the processes in order to forecast and backcast the time series, but also facilitates modelling of trading day and holiday effects, detection of outlier effects, dealing with missing data, detection of sudden level changes, and detection of changes in the seasonal pattern, trading day effects and so forth. The second major improvement on the earlier programmes is the inclusion of a module for diagnostics which contains many helpful 'tests'. The third improvement is a user-friendly interface.

Although X-12 is a major improvement to X-11, it has its critics. For example, Wallis (1998) doubts that the trend estimation procedure taken over from X-11 is still the best available despite the results obtained since the mid-1970s, and he stresses the need to give the user of the adjusted numbers an indication of their susceptibility to revision.

TRAMO/SEATS seasonal adjustment programme

The main difference between the X-12 programme and the TRAMO/SEATS programme is that the former uses signal-to-noise ratios to choose between the different moving average filters available while SEATS uses signal extraction with filters derived from a time series (ARIMA) model.

The programme also contains a preadjustment programme, TRAMO, which basically performs tasks similar to RegARIMA in X-12.

The signal extraction is based on an additive model such as (1) or

$$Y_t = \mu_t + \gamma_t + \varepsilon_t, \tag{4}$$

where μ_t is the trend-cycle component, γ_t the seasonal component, and ε_t is the irregular component. It is then assumed that the μ_t and γ_t can be modelled as two distinct ARIMA processes

$$A_C(L)\,(1-L)^d \mu_t = B_C(L) v_t \text{ and } A_S(L)\,(1-L^s)^D \gamma_t = B_S(L) w_t \tag{5}$$

where the processes v_t, w_t and ε_t are independent, serially uncorrelated processes with zero means and variances σ_u^2, σ_w^2 and σ_ε^2, and d and D are integers, while L is the lag operator. This class of model is also called the unobserved components autoregressive integrated moving-average model (UCARIMA) by Engle (1978).

Hence, the TRAMO/SEATS programme requires the estimation of the UCARIMA parameters for each specific series, which in principle should allow computation of the correct number of degrees of freedom. This is not possible in X-12 due to the adjustments undertaken within the programme based on the characteristics of the individual series.

A discussion of the merits and drawbacks of X-12 and TRAMO/SEATS may be found in Ghysels and Osborn (2001), Hood, Ashley and Findley (2004), and several working papers from EUROSTAT (see Mazzi and Savio, 2005), which find that X-12 is slightly preferable to TRAMO/SEATS when applied to short time series – a result to be expected as the model-based approach requires more data. In fact, the main differences between the two leading competitors reflect the difference between the model-based approach of TRAMO/SEATS, which tailors a seasonal filter to each series, and the uniform filter applied by X-12 (see below). However, the model-based approach relies on a very restrictive set of models, and the uniform filter approach is not really applying the same filter, as individual characteristics like outliers, smoothness, and so on, have an influence on the filter.

Seasonal adjustment as an integrated part of the analysis

The main objective of the production of seasonally adjusted time series is to give the policy analyst or adviser easy access to a common time series data-set that has been purged of what is considered noise contaminating the series. Obviously, the application of the seasonally adjusted data may be more or less formal and meticulous, ranging from eyeball analysis to thorough econometric analysis.

However, although the application of officially seasonally adjusted data may have the advantage of saving costs, it also implies that the user runs a severe risk of not making the most effective use

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of the information available and – perhaps more serious – of applying a dataset distorted by the applied seasonal filter.
The possible reasons for these shortcomings are:

- the seasonal component is a noise component but
  - – the wrong seasonal adjustment filter has been applied, or
  - – the data have been seasonally adjusted individually without consideration being given to the fact that they are often used as input to a multivariate analysis;
- • the seasonal components of different time series may be closely connected and contain valuable information across series.

Filtering the data before they are applied may of course distort the outcome of the analysis if the wrong filter is used, but even if the filter is 'correct' as seen from the individual series, the filtering may produce biased estimates of the parameters in certain cases where, for instance, a regression model is applied (see Hylleberg, 1986, p. 3). However, this result is complicated by the application of other transformations to the original series. Which filter to apply may in fact depend on the order of the applied transformations, as shown by Ghysels (1997).
Hence, in order to optimally model many economic phenomena, seasonality may need to be treated as an integrated part of an econometric analysis based on unadjusted quarterly, monthly, weekly and daily time series or panel data observations. This may be done in many different ways depending on the specific context and the set of reasonable assumptions one can make within that context. As both X-12 and TRAMO/SEATS seasonal adjustment programmes are available to the individual researcher, they may both be applied as part of an integrated approach and their use somewhat adapted to the specific analysis. In what follows we discuss some alternative methods, of three kinds: pure noise models, time series models and economic models.

Pure noise models

The first group comprises seasonal adjustment methods which are based on the assumption that the seasonal component is noise. Thus, the group also contains the officially applied seasonal adjustment programmes presented earlier. The seasonal adjustment methods in this group are distinguished by their ability to take care of a changing seasonal component.

Seasonal dummies

The use of seasonal dummy variables to filter quarterly and monthly times series data is a very simple, straightforward and therefore popular method in econometric applications. The dummy variable method is designed to take care of a constant, stable seasonal component. The popularity of the seasonal dummy variable method is partly due to its simplicity and the flexible way it can be used either as a pre-filtering device whereby the series are regressed on a set of seasonal dummy variables and the residuals used in the final regression, or within the regression as an extension of the set of regressors with seasonal dummy variables (see Frisch and Waugh, 1933; Lovell, 1963).

Band spectrum regression and band pass filters

A natural and quite flexible way to analyse time series with a strong and somewhat varying periodic component is to perform the analysis in the frequency domain, where the time series is represented as a weighted sum of cosine and sine waves. Hence, the time series are Fourier-transformed and the seasonal filtering of the time series may take place by removing specific frequency components from the Fourier-transformed data series.
Application of such filters dates back a long time (see Hannan, 1960). Band spectrum regression is further developed and analysed by Engle (1974; 1980) and Hylleberg (1977; 1986). The so-called real business cycle literature has since named it 'band pass filtering' (see Baxter and King, 1999).
Let us assume that we have data series with *T* observations in a vector *y* and a matrix *X* related by $y = X\beta + \varepsilon$, where ε is the disturbance term and β a coefficient vector. Band spectrum regression is then performed as a regression in the transformed model $A\Psi y = A\Psi X\beta + A\Psi\varepsilon$, where the transformation by the matrix Ψ is a finite Fourier transformation of the data. The transformation by the diagonal matrix *A* with zeros and ones on the diagonal, symmetric around the south-west north-east diagonal, is a filtering which removes the frequency components corresponding to the elements with the zeros. Hence, by an appropriate choice of zeros in the main diagonal of *A* the exact seasonal frequencies and possibly a band around them may be filtered from the series.
An obvious advantage of the band spectrum regression representation is that the model $A\Psi y = A\Psi X\beta + A\Psi\varepsilon$ lends itself directly to a test for the appropriate filtering, as argued in Engle (1974). In fact the test is just the well-known so-called Chow test applied to a stacked model with the null hypothesis that the parameters are the same over the different frequencies. A drawback of band spectrum regression is that the temporal relations between series may be affected in a complicated way by the two-sided filter (see Engle, 1980; Bunzel and Hylleberg, 1982).

Seasonal integration and seasonal fractional integration

A simple filter often applied in empirical econometric work is the seasonal difference filter $(1 - L^s)^d$, where *s* is the number of observations per year and *d* the number of times the filter should be applied to render the series stationary at the long run and seasonal frequencies (see Box and Jenkins, 1970).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the unit root literature a time series is said to be integrated of order *d* if its *d*'th difference has a stationary and invertible ARMA representation. Hylleberg et al. (1990) generalize this to seasonal integration and denote, for instance, a quarterly series y_t, $t = 1, 2, \ldots, T$ represented by the model $(1 - L^4) y_t = \varepsilon_t$, $\varepsilon_t \sim i.i.d.$ $(0, \sigma^2)$ as integrated of order 1 at frequency θ since $(1 - L^4) = (1 - L)(1 + L)(1 + L^2)$ has real roots at the unit circle at the frequencies $\theta = \left\{0, \frac{1}{2}, [\frac{1}{4}, \frac{3}{4}]\right\}$, where θ is given as the share of a total circle of 2π.

Many empirical studies have applied the so-called Hylleberg, Engle, Granger, and Yoo (HEGY) test for seasonal unit roots developed by Hylleberg et al. (1990) and Engle et al. (1993) for quarterly data, extended to monthly data by Beaulieu and Miron (1993), and to daily data integrated at a period of one week by Kunst (1997). These tests are extensions of the well-known Dickey–Fuller test for a unit root at the long-run zero frequency (Dickey and Fuller, 1979) and at the seasonal frequencies (Dickey, Hasza and Fuller, 1984).

The HEGY test is the simplest and most easily applied test for seasonal unit roots. In the quarterly case the test is based on an autoregressive model $\varphi(L) y_t = \varepsilon_t$, $\varepsilon_t \sim \ddot{u}d$ $(0, \sigma^2)$ where $\phi(L)$ is a lag polynomial with possible unit roots at frequencies $\theta = \left\{0, \frac{1}{2}, [\frac{1}{4}, \frac{3}{4}]\right\}$. A rewritten linear regression model where the possible unit roots are isolated in specific terms is

$$\varphi^*(L) y_{4t} = \pi_1 y_{1t-1} + \pi_2 y_{2t-1} + \pi_3 y_{3t-2} + \pi_4 y_{3t-1} + \varepsilon_t y_{1t} = (1 + L + L^2 + L^3) y_t y_{2t} = -(1 - L + L^2 - L^3) y_t y_{3t} = -(1 - L^2) y_t y_{4t} = (1 - L^4) y_t. \quad (6)$$

The lag polynomial $\phi^*(L)$ is a stationary and finite poly-nomial by assumption. Denoting integration of order *d* at frequency θ by $I_\theta \bullet (d)$ we thus have $y_{1t} \sim I_0(1)$, $y_{2t} \sim I_{\frac{1}{2}}(1)$, and $y_{3t} \sim I_{[1/4,3/4]}(1)$, while $y_{1t} \sim I_{\frac{1}{2},[1/4,3/4]}(0)$, $y_{2t} \sim I_{0,[1/4,3/4]}(0)$, $y_{3t} \sim I_{0,\frac{1}{2}}(0)$, and $y_{4t} \sim I_{0,\frac{1}{2},[1/4,3/4]}(0)$, provided $y_t \sim I_{0,\frac{1}{2},[1/4,3/4]}(1)$.

The HEGY tests of the null hypothesis of a unit root are conducted by '*t*-value' tests on π_1 for the long-run unit root, π_2 for the semi-annual unit root, and '*F*-value tests' on π_3 and π_4 for the annual unit roots. In fact, the '*t*-value' tests on π_1 is just the unit root test of Dickey and Fuller with a special augmentation applied. As in the Dickey–Fuller cases the statistics are not *t* or *F* distributed but have non-standard distributions, which for the 't' are tabulated in Fuller (1976) while critical values for the 'F' test are tabulated in Hylleberg et al. (1990).

As in the Dickey–Fuller case the correct lag-augmentation in the auxiliary regression (6) is crucial. The errors need to be rendered white noise in order for the size to be close to the stipulated significance level, but the use of too many lag coefficients reduces the power of the tests.

Obviously, if the data-generating process (DGP) contains a moving average component, the augmentation of the autoregressive part may require long lags (see Hylleberg, 1995) and the HEGY test may be seriously affected by autocorrelation in the errors, moving average terms with roots close to the unit circle, so-called structural breaks, and noisy data with outliers.

The existence of seasonal unit roots in the DGP implies a varying seasonal pattern where 'summer may become winter'. In most cases such an extreme situation is not logically possible, and the findings of seasonal unit roots should be taken as an indication of a varying seasonal pattern and the unit root model as a parsimonious approximation to the DGP.

Another test where the null is no unit root at the zero frequency is suggested by Kwiatkowski et al. (1992) and extended to the seasonal frequencies by Canova and Hansen (1995), and further developed by Busetti and Harvey (2003). See Hylleberg (1995) for a comparison of the Canova–Hansen test and the HEGY test. See also Taylor (2005) for a variance ratio test.

Arteche (2000) and Arteche and Robinson (2000) have extended the analysis to include non-integer values of *d* in the definition of a seasonally integrated process. In case *d* is a number between 0 and 1 the process is called fractionally seasonally integrated. The fractionally integrated seasonal process is said to have strong dependence or long memory at a frequency ω since the autocorrelations at that frequency die out at a hyperbolic rate, in contrast to the much faster exponential rate in the weak dependence case where $d = 0$. In the integrated case where $d = 1$ the autocorrelations never die out.

The difficulty with the fractional model is estimation of the parameter *d*. Even in the quarterly case there are three possible *d* parameters, and the testing procedure may become very elaborate, requiring, for instance, a sequence of clustered tests as in Gil-Alana and Robinson (1997).

Time series models

The time series models may be univariate models such as the Box–Jenkins model, unobserved components models, time varying parameter models or evolving seasonal models, or multivariate models with seasonal cointegration or periodic cointegration, or models with seasonal common features.

Univariate seasonal models

The Box–Jenkins model. In the traditional analysis of Box and Jenkins (see Box and Jenkins, 1970), the time series where *s* is the number of quarters, months, and so on, in the year were made stationary by application of the filters $(1 - L)$ *and/or* $(1 - L^s) = (1 - L) S(L)$, where $S(L) = (1 + L + L^2 + L^3 + \ldots\ldots L^{s-1})$, as many times as was deemed necessary from the form of the resulting

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

autocorrelation function. After stationarity had been obtained, the filtered series were modelled as an autoregressive moving average model (ARMA) model. Both the AR and the MA part could be modelled as consisting of a non-seasonal and seasonal lag polynomial. Hence, the so-called seasonal ARIMA model has the form

$$\varphi(L)\varphi_S(L^S)(1-L^S)^D(1-L)^d y_t = \theta(L)\theta_S(L^S)\varepsilon_t \quad (7)$$

where $\phi(L)$ and $\theta(L)$ are invertible lag polynomials in L, while $\phi_s(L^s)$ and $\theta_s(L^s)$ are invertible lag polynomials in L^s, and D and d integers.

In light of the results mentioned in the section on seasonal unit roots, the modelling strategy of Box and Jenkins may easily be refined to allow for situations were the non-stationarity exists only at some of the seasonal frequencies.

The 'structural' or unobserved components model. When modelling processes with seasonal characteristics, one must apply complicated and high-ordered polynomials in the ARMA representation. As an alternative to this, the unobserved components model (UC) discussed earlier was proposed. It is easily seen that the UCARIMA model is a general ARIMA model with restrictions on the parameters. Alternatively, the UC model may be specified as a so-called structural model following Harvey (1993).

The structural model is based on a very simple and quite restrictive modelling of the components of interest such as trends, seasonals and cycles. The model is often specified as (4). The trend μ_t is normally assumed to be stationary only in first or second differences, whereas the seasonal component γ_t is stationary when multiplied by the seasonal summation operator $S(L)$. In the basic structural model (BSM) the trend is specified as

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \beta_t = \beta_{t-1} + \zeta_t \quad (8)$$

where each of the error terms is independently distributed. (If $\sigma_\xi^2 = 0$ this collapses to a random walk plus drift. If $\sigma_\eta^2 = 0$ as well it corresponds to a model with a linear trend.) The seasonal component is specified as

$$S(L)\gamma_t = \sum_{j=0}^{n-1} \gamma_{t-j} = \omega_t \quad (9)$$

where s is the number of periods per year and where $\omega_t \sim N(0, \sigma_\omega^2)$. (This specification is known as the dummy variable form, since it reduces to a standard deterministic seasonal component if $\sigma_\omega^2 = 0$. Specifying the seasonal component this way makes it slowly changing by a mechanism that ensures that the sum of the seasonal components over any s consecutive time periods has an expected value of zero and a variance that remains constant over time.) The BSM model can also be written as

$$y_t = \frac{\xi_t}{\Delta^2} + \frac{\omega_t}{S(L)} + \varepsilon_t, \quad (10)$$

where $\xi_t = \eta_t - \eta_{t-1} + \xi_{t-1}$ is equivalent to an MA(1) process. Expressing the model in the form (10) makes the connection to the UCARIMA model in (4) clear.

Estimation of the general UC model is treated in Hylleberg (1986) and estimation of the structural model is treated in Harvey, Koopman and Shephard (2004).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the structural approach the problems of specifying the ARMA models for the components is thus avoided by a priori restrictions. Harvey and Scott (1994) argue that the type of model above, which has a seasonal component evolving relatively slowly over time, can fit most economic time series, irrespective of the apparently strong assumptions of a trend component with a unit root and a seasonal component with all possible seasonal unit roots present.

Periodic models and other time varying parameter models. The periodic model extends the non-periodic time series models by allowing the parameters to vary with the seasons. The so-called periodic autoregressive model of order h (PAR(h)) assumes that the observations in each of the seasons can be described using different autoregressive models (see Franses, 1996).

Consider a quarterly times series y_t which is observed for N years. The stationary PAR(h) quarterly model can be written as

$$y_t = \sum_{s=1}^{4} \mu_s D_{s,t} + \sum_{s=1}^{4} \varphi_{1s} D_{s,t} y_{t-1} + \ldots + \sum_{s=1}^{4} \varphi_{hs} D_{s,t} y_{t-h} + \varepsilon_t \tag{11}$$

with $s = 1, 2, 3, 4$, $t = 1, 2, \ldots, T = 4N$, and where $D_{s,t}$ are seasonal dummies, or as $y_t = \mu_s + \varphi_{1s} y_{t-1} + \ldots + \varphi_{ps} y_{t-h} + \varepsilon_t$.

It has been shown that any PAR model can be described by a non-periodic ARMA model (Osborn, 1991). In general, however, the orders will be higher than in the PAR model. For example, a PAR (1) corresponds to a non-periodic ARMA(4,3) model. Furthermore, it has been shown that estimating a non-periodic model when the true DGP is a PAR can result in a lack of ability to reject the false non-periodic model (Franses, 1996). Fitting a PAR model does not prevent the finding of a non-periodic AR process, if the latter is in fact the DGP. In practice it is thus recommended that one starts by selecting a PAR(h) model and then tests whether the autoregressive parameters are periodically varying using the method described above.

A major weakness of the periodic model is that the available sample for estimation $N = n / s$ often is too small. Furthermore, the identification of a periodic time series model is not as easy as it is for non-periodic models.

Now, let us rewrite the series y_t, $t = 1, 2, 3, \ldots T$, as $y_{s,\tau}$, where $s = 1, 2, 3, 4$ indicating the quarter, and $\tau = 1, 2, \ldots, n$ indicating the year. The PAR(1) process can then be written as

$$y_{s,\tau} = \varphi_s y_{s-1,\tau} + \varepsilon_{s,\tau}, \quad s = 1, 2, 3, 4; \ \tau = 1, 2, \ldots, n \tag{12}$$

where $y_{0,\tau} = y_{4,\tau-1}$, or in vector notation

$$\Phi(L) Y_\tau = U_\tau \tag{13}$$

where

$$\Phi(L) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\varphi_2 & 1 & 0 & 0 \\ 0 & -\varphi_3 & 1 & 0 \\ 0 & 0 & -\varphi_4 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & -\varphi_1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} L$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$Y_\tau' = [y_{1,\tau}, y_{2,\tau}, y_{3,\tau}, y_{4,\tau}]\, U_\tau' = [\varepsilon_{1,\tau}, \varepsilon_{2,\tau}, \varepsilon_{3,\tau}, \varepsilon_{4,\tau}]$$

with L operating on the seasons, that is, L $y_{s,\tau} = y_{s-1,\tau}$ and especially L $y_{1,\tau} = y_{0,\tau} = y_{4,\tau-1}$. The PAR(1) process in (13) is stationary provided $|\Phi(z)| = 0$ has all its roots outside the unit circle, which is the case if and only if $|\varphi_1\varphi_2\varphi_3\varphi_4| < 1$.

The model may be estimated by maximum likelihood or OLS. Testing for periodicity in (11) amounts to testing the hypothesis $H_0: \varphi_{is} = \varphi_i$ for $s = 1, 2, 3, 4$ and $i = 1, 2, \ldots, p$, and this can be done with a likelihood ratio test, which is asymptotically χ^2_{3p} under the null, irrespective of any unit roots in y_t (see Boswijk and Franses, 1995).

The vector representation of the PAR model forms an effective vehicle for generating estimation and testing procedures directly from the general result for stationary vector autoregression (VAR) models, but it also creates an effective way to handle the non-stationary case and compare the periodic models with the models with seasonal integration.

In the non-stationary case, a periodically integrated process of order 1(PI(1)) is defined as a process, where there exists a quasi-difference

$$D_s y_{s,\tau} = 1 - \alpha_s y_{s-1,\tau}\alpha_1\alpha_2\alpha_3\alpha_4 = 1 \text{not all } \alpha_s = 1,\ s = 1, 2, 3, 4. \tag{14}$$

such that $D_s y_{s,\tau}$ has a stationary and invertible representation. Notice that the *PI(1)* process is neither an integrated $I_0(1)$ process nor a seasonally integrated $I_{0,\frac{1}{2}[1/4,3/4]}(1)$ process as shown in Ghysels and Osborn (2001).

The periodic models can be considered special cases of what are referred to as the time-varying parameter models (see Hylleberg, 1986). These are regression models of the form

$$Y_t = X_t'\beta_t + u_t B(L)\ (\beta_t - \bar{\beta}) = AY_t + \xi_t \tag{15}$$

which can be written in state-space form and estimated using the Kalman filter. However, the number of parameters is often greater than the number of observations, and in practice one may be forced to restrict the parameter space. Gersovitz and MacKinnon (1978), applying Bayesian techniques, adopt the sensible assumption that the parameters vary smoothly over the seasons.

The evolving seasonals model. The evolving seasonals model was promulgated by Hannan, Terrell and Tuckwell (1970). The model has been revitalized by Hylleberg and Pagan (1997), who show that the evolving seasonals model produces an excellent vehicle for analysing different commonly applied seasonal models as it nests many of them. The model has been used by Koop and Dijk (2000) to analyse seasonal models from a Bayesian perspective.

The evolving seasonals model for a quarterly time series is based on a representation like

$$y_t = \alpha_{1t}\cos(\lambda_1 t) + \alpha_{2t}\cos(\lambda_2 t) + 2\alpha_{3t}\cos(\lambda_3 t) + 2\alpha_{4t}\sin(\lambda_3 t), = \alpha_{1t} + \alpha_{2t}\cos(\pi t) + 2\alpha_{3t}\cos(\pi t/2) + 2\alpha_{4t}\sin(\pi t/2), = \alpha_{1t}(1)^t + \alpha_{2t}(-1)^t + \alpha_{3t}[i^t + (-i)^t] + \alpha_{4t}[i^{t-1} + (-i)^{t-1}], \tag{16}$$

where $\lambda_1 = 0, \lambda_2 = \pi, \lambda_3 = \pi/2, \cos(\pi t) = (-1)^t, 2\cos(\pi t/2) = [i^t + (-i)^t], 2\sin(\pi t/2) = [i^{t-1} + (-i)^{t-1}], i^2 = -1$, while $\alpha_{jt}, j = 1, 2, 3, 4$; is a linear function of its own past and a stochastic term e_{jt}, $j = 1, 2, 3, 4$: For instance,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\alpha_{1t} = \rho_1\alpha_{1,t-1} + e_{1t}, \alpha_{2t} = \rho_1\alpha_{2,t-1} + e_{2t}, \alpha_{3t} = \rho_3\alpha_{3,t-2} + e_{3t}, \alpha_{4t} = \rho_4\alpha_{4,t-3} + e_{4t}. \quad (17)$$

In such a model, $\alpha_{1t}(1)^t = \alpha_{1t}$ represents the trend component with the unit root at the zero frequency, $\alpha_{2t}(-1)^t$ represents the semi-annual component with the root –1, while $\alpha_{3t}[i^t + (-1)^t]\,\alpha_{4t}[i^{t-1} + (-1)^{t-1}]$ represents the annual component with the complex conjugate roots ±*i*. In Hylleberg and Pagan (1997) it is shown that the HEGY auxiliary regression in (6) has an evolving seasonals model representation, and also that the Canova–Hansen test and the PAR(h) model may be presented in the framework of the evolving seasonals model.

Multivariate seasonal time series models

The idea that the seasonal components of a set of economic time series are driven by a smaller set of common seasonal features seems a natural extension of the idea that the trend components of a set of economic time series are driven by common trends.

If the seasonal components are seasonally integrated, the idea immediately leads to the concept of seasonal cointegration, introduced in Engle, Granger and Hallman (1989), Hylleberg et al. (1990), and Engle et al. (1993). In case the seasonal components are stationary, the idea leads to the concept of seasonal common features (see Engle and Hylleberg, 1996), while so-called periodic cointegration considers cointegration season by season (see Birchenhal et al., 1989; Ghysels and Osborn, 2001).

Seasonal cointegration. Seasonal cointegration exists at a particular seasonal frequency if at least one linear combination of series which are seasonally integrated at the particular frequency is integrated of a lower order.

Consider the quarterly case where y_t and x_t are both integrated of order 1 at the zero and at the seasonal frequencies, that is, the transformations corresponding to 6 are $\{y_{1t}, x_{1t}\} \sim I_0(1)$, $\left\{y_{2t}, x_{2t}\right\} \sim I_{\frac{1}{2}}(1)$ and $\{y_{3t}, x_{3t}\} \sim I_{[1/4,3/4]}(1)$. Cointegration at the frequency $\theta = 0$ then exists if $y_{1t} - k_1x_{1t} \sim I_0(0)$ for some non-zero k_1, cointegration at the frequency $\theta = \frac{1}{2}$ exists if $y_{2t} - k_2x_{2t} \sim I_{\frac{1}{2}}(0)$ for some non-zero k_2, while cointegration at the frequency $\theta = [1/4, 3/4]$ exists if $y_{2t} - k_3x_{2t} - k_4x_{2,t-1} \sim I_{[1/4,3/4]}(0)$ for some non-zero pair $\{k_3, k_4\}$. The complex unit roots at the annual frequency [1/4,3/4] lead to the concept of polynomial cointegration, where cointegration exists if one can find at least one linear combination including a lag of the seasonally integrated series which is stationary.

In Hylleberg et al. (1990) and Engle et al. (1993), seasonal cointegration is analysed along the path set up in Engle and Granger (1987).

The well-known drawbacks of this method, especially when the number of variables included exceeds two, is partly overcome by Lee (1992), who extends the maximum likelihood (ML)-based methods of Johansen (1995) for cointegration at the long-run frequency, to cointegration at the semi-annual frequency $\theta = \frac{1}{2}$.

To adopt the ML-based cointegration analysis at the annual frequency $\theta = [1/4, 3/4]$ with the complex pair of unit roots ±*i* is somewhat more complicated, however. The general results may be found in Johansen and Schaumburg (1999), and Cubadda (2001) applies the results of Brillinger (1981) on the canonical correlation analysis of complex variables to obtain tests for cointegration at all the frequencies of interest, that is, at the frequencies 0 and π with the real unit roots ±1 and at the frequency $\theta = [1/4, 3/4]$ with the complex roots ±*i*.

Periodic cointegration. Periodic cointegration extends the notion of seasonal cointegration by allowing the coefficients in the cointegration relations to be periodic (see Ghysels and Osborn, 2001). Consider the example given above with two quarterly time series y_t and x_t, $t = 1, 2, \ldots, T$ which are integrated of order 1 at the zero and seasonal frequencies implying that a transformation by the fourth difference $1 - L^4$ will make the two series stationary. Such series are called seasonally integrated series. Let us rewrite the series as $y_{s,\tau}$ and $x_{s,\tau}$ with $s = 1, 2, 3, 4$ indicating the quarter, and $\tau = 1, 2, \ldots, n$ indicating the year. Hence, the eight yearly series $y_{s,\tau}$, $x_{s,\tau}$ $s = 1, 2, 3, 4$ are all integrated of order 1 at the zero frequency.

Hence, full periodic cointegration exists (see Boswijk and Franses, 1995) if $y_{\tau t} - k_s x_{\tau r} \sim I_0(0)$ for some non-zero k_s, $s = 1, 2, 3, 4$, $\tau = 1, 2, 3$. In case stationarity is only obtained for some $s = 1, 2, 3, 4$, partially periodic cointegration exists.

Several interesting and useful results reviewed in Ghysels and Osborn (2001) follow:

1. 1. Two seasonally integrated series may fully or partially periodically cointegrate.
2. 2. Two $I_0(1)$ processes cannot be periodically cointegrated. They are either non-periodically cointegrated or not cointegrated at all.
3. 3. If two *PI*(1) processes cointegrate in one quarter. they cointegrate in all four quarters.

Periodic cointegration is a promising but currently not fully exploited area of research, which has the inherent problem that it requires a large sample. It is therefore not surprising that the recent advances in this area happen when data are plentiful (daily) and it is possible to restrict the model appropriately (Haldrup et al., 2007).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Common seasonal features. Although economic time series often exhibit non-stationary behaviour, stationary economic variables exist as well, especially when conditioned on some deterministic pattern such as linear trends, seasonal dummies, breaks and so on. However, a set of stationary economic times series may also exhibit common behaviour and, for instance, share a common seasonal pattern. The technique for finding such patterns, known as common seasonal features (see Engle and Hylleberg, 1996; Cubadda, 1999) is based on earlier contributions defining common features by Engle and Kozicki (1993) and Vahid and Engle (1993).

Consider a multivariate autoregression written in error correction form as

$$\Delta Y_t = \sum_{j=1}^{p} B_j \Delta Y_{t-j} + \Pi v_{t-1} + \Gamma z_t + \varepsilon_t,\ t = 1, 2, \ldots, T, \tag{18}$$

where Y_t is $k \times 1$ vector of observations on the series of interest in period t and the error correction term is Πv_{t-1}. The vector v_t contains the cointegrating relations at the zero frequency, and the number of cointegrating relations is equal to the rank of Π. If Π has full rank equal to k the series are stationary. In the quarterly case the vector z_t is a vector of trigonometric seasonal dummies, such as $\{\cos(2\pi ht/4 + 2\pi j/T),\ h = 1, 2;\ j \in (-\delta T \square j \square \delta T),\ \sin(2\pi h4 + 2\pi j/T),\ h = 1, 2;\ j \in (-\delta T \square j \square \delta T),\ j \neq 0,$ when $h = 2\}$. The use of trigonometric dummy variables facilitates the 'modelling' of a varying seasonal pattern, since a proper choice of δ takes care of the neighboring frequencies to the exact seasonal frequencies. If δ is 0, the filter is equivalent to the usual seasonal dummy filter.

The implication of a full rank of the $k \times m$ matrix Γ, equal to min[k, m], is that different linear combinations of the seasonal dummies in z_t are needed in order to explain the seasonal behaviour of the variables in Y_t. However, if there are common seasonal features in these variables we do not need all the different linear combinations, and the rank of Π is not full. Thus, a test of the number of common seasonal features can be based on the rank of Π (see Engle and Kozicki, 1993).

The test is based on a reduced rank regression similar to the test for cointegration by Johansen (1995). Hence, the hypotheses are tested using a canonical correlation analysis between of z_t and ΔY_t, where both sets of variables are purged of the effect from the other variables in (18).

This kind of analysis has proved useful in some situations, but it is difficult to apply in cases where the number of variables is large, and the results are sensitive to the lag-augmentation as in the case of cointegration. In addition, the somewhat arbitrary nature of the choice of z_t poses difficulties.

Economic models of seasonality

Many economic time series have a strong seasonal component, and obviously economic agents must react to that. Hence, the seasonal variation in economic time series must be an integrated part of the optimizing behaviour of economic agents, and the seasonal variation in economic time series must be a result of the optimizing behaviour of economic agents, reacting to exogenous factors such as the weather, the timing of holidays, and so on.

The fact that economic agents react and adjust to seasonal movements on one hand and influence them on the other, implies that the application of seasonal data in economic analysis may widen the possibilities for testing theories about economic behaviour. The relative ease with which the agents may forecast at least some of the causes of the seasonality may be quite helpful in setting up testable models for production smoothing, for instance.

Apart from what is caused by the easiness of forecasting exogenous factors, the type of optimizing behaviour and the agents' reactions to a seasonal phenomenon may be expected not to differ fundamentally from what is happening in a non-seasonal context. However, the recurrent characteristic of seasonality may be exploited.

The economic treatment of seasonal fluctuation has been discussed in the real business cycle (RBC) literature (for example, Chatterjee and Ravikumar, 1992; Braun and Evans, 1995), working with a utility optimizing consumer faced with some feasibility constraint. However, in most of this RBC branch seasonality arises from deterministic shifts in tastes and technology. A few other papers incorporate seasonality through stochastic productivity shocks (see for example Wells, 1997; Cubadda, Savio and Zelli, 2002).

Another area is the production smoothing literature (for instance, Ghysels, 1988; Miron and Zeldes, 1988; Miron, 1996) and habit persistence (for example, Osborn, 1988), where a model for seasonality and habit persistence is presented in a life-cycle consumption model.

See Also

- data filters

The author is grateful for helpful comments from Niels Haldrup and Steven Durlauf.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Arteche, J. 2000. Gaussian semiparametric estimation in seasonal/cyclical long memory time series. *Kybernetika* 36, 279–310.

Arteche, J. and Robinson, P. 2000. Semiparametric inference in seasonal and cyclical long memory processes. *Journal of Time Series Analysis* 21, 1–25.

Baxter, M. and King, R. 1999. Measuring business cycles: approximate band-pass filters for economic time series. *Review of Economics and Statistics* 81, 575–93.

Beaulieu, J. and Miron, J. 1993. Seasonal unit roots in aggregate US data. *Journal of Econometrics* 55, 305–28.

Birchenhal, C., Bladen-Howell, R., Chui, A., Osborn, D. and Smith, J. 1989. A seasonal model of consumption. *Economic Journal* 99, 837–43.

Boswijk, H. and Franses, P. 1995. Periodic cointegration: representation and inference. *Review of Economics and Statistics* 77, 436–54.

Box, G. and Jenkins, G. 1970. *Time Series Analysis, Forecasting, and Control*. San Francisco: Holden-Day.

Braun, R. and Evans, C. 1995. Seasonality and equilibrium business cycle theories. *Journal of Economic Dynamics and Control* 19, 503–31.

Brendstrup, B., Hylleberg, S., Nielsen, M., Skipper, L. and Stentoft, L. 2004. Seasonality in economic models. *Macroeconomic Dynamics* 8, 362–94.

Brillinger, D. 1981. *Time Series: Data Analysis and Theory*. San Francisco: Holden Day.

Bunzel, H. and Hylleberg, S. 1982. Seasonality in dynamic regression models: a comparative study of finite sample properties of various regression estimators including band spectrum regression. *Journal of Econometrics* 19, 345–66.

Busetti, F. and Harvey, A. 2003. Seasonality tests. *Journal of Business and Economic Statistics* 21, 421–36.

Canova, F. and Hansen, B. 1995. Are seasonal patterns constant over time? A test for seasonal stability. *Journal of Business and Economic Statistics* 13, 237–52.

Chatterjee, S. and Ravikumar, B. 1992. A neoclassical model of seasonal fluctuations. *Journal of Monetary Economics* 29, 59–86.

Cubadda, G. 1999. Common cycles in seasonal non-stationary time series. *Journal of Applied Econometrics* 14, 273–91.

Cubadda, G. 2001. Complex reduced rank models for seasonally cointegrated time series. *Oxford Bulletin of Economics and Statistics* 63, 497–511.

Cubadda, G., Savio, G. and Zelli, R. 2002. Seasonality, productivity shocks, and sectoral comovements in a real business cycle model for Italy. *Macroeconomic Dynamics* 6, 337–56.

Dagum, E. 1980. The X-11-ARIMA seasonally adjustment method. Technical Report 12-564E. Ottawa: Statistics Canada.

Dickey, D. and Fuller, W. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.

Dickey, D., Hasza, D. and Fuller, W. 1984. Testing for unit roots in seasonal time series. *Journal of the American Statistical Association* 79, 355–67.

Engle, R. 1974. Band spectrum regression. *International Economic Review* 15, 1–11.

Engle, R. 1978. Estimating structural models of seasonality. In *Seasonal Analysis of Economic Time Series*, ed. A. Zellner. Washington, DC: US Census Bureau.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Engle, R. 1980. Exact maximum likelihood methods for dynamic regressions and band spectrum regressions. *International Economic Review* 21, 391–407.

Engle, R. and Granger, C. 1987. Co-integration and error correction: representation, estimation and testing. *Econometrica* 55, 251–76.

Engle, R., Granger, C. and Hallman, J. 1989. Merging short and long run forecasts: an application of seasonal cointegration to monthly electricity sales forecasting. *Journal of Econometrics* 40, 45–62.

Engle, R., Granger, C., Hylleberg, S. and Lee, H. 1993. Seasonal cointegration – the Japanese consumption function. *Journal of Econometrics* 55, 275–98.

Engle, R. and Hylleberg, S. 1996. Common seasonal features: global unemployment. *Oxford Bulletin of Economics and Statistics* 58, 615–30.

Engle, R. and Kozicki, S. 1993. Testing for common features. *Journal of Business and Economic Statistics* 11, 369–80.

Findley, D., Monsell, B., Bell, W., Otto, M. and Chen, B. 1998. New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics* 16, 127–76.

Franses, P. 1996. *Periodicity and Stochastic Trends in Economic Time Series*. Oxford: Oxford University Press.

Frisch, R. and Waugh, F. 1933. Partial time regressions as compared with individual trends. *Econometrica* 1, 387–401.

Fuller, W. 1976. *Introduction to Statistical Time Series*. New York: John Wiley and Sons.

Gersovitz, M. and MacKinnon, J. 1978. Seasonality in regression: an application of smoothness priors. *Journal of the American Statistical Association* 73, 264–73.

Ghysels, E. 1988. A study towards a dynamic theory of seasonality for economics time series. *Journal of the American Statistical Association* 83, 68–72.

Ghysels, E. 1997. Seasonal adjustments and other data transformations. *Journal of Business and Economic Statistics* 15, 410–18.

Ghysels, E. and Osborn, D. 2001. *The Econometric Analysis of Seasonal Time Series*. Cambridge: Cambridge University Press.

Gil-Alana, L. and Robinson, P. 1997. Testing of unit root and other non-stationary hypotheses in macroeconomic time series. *Journal of Econometrics* 80, 241–68.

Gomez, V. and Maravall, A. 1996. *Programs TRAMO and SEATS*. Madrid: Banco de Espana.

Haldrup, N., Hylleberg, S., Pons, G. and Sansó, A. 2007. Common periodic correlation features and the interaction of stocks and flows in daily airport data. *Journal of Business and Economic Statistics* 25, 21–32.

Hannan, E. 1960. *Time Series Analysis*. London: Methuen.

Hannan, E., Terrell, R. and Tuckwell, N. 1970. The seasonal adjustment of economic time series. *International Economic Review* 11, 24–52.

Harvey, A. 1993. *Time Series Models*. London: Prentice Hall/Harvester Wheatsheaf.

Harvey, A. and Scott, A. 1994. Seasonality in dynamic regression models. *Economic Journal* 104, 1324–45.

Harvey, A., Koopman, S. and Shephard, N., eds. 2004. *State Space and Unobserved Component Models: Theory and Applications*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hood, C., Ashley, J. and Findley, D. 2004. An empirical evaluation of the performance of TRAMo/SEATS on simulated series. Technical report. Washington, DC: US Census Bureau.

Hylleberg, S. 1977. A comparative study of finite sample properties of band spectrum regression estimators. *Journal of Econometrics* 5, 167–82.

Hylleberg, S. 1986. *Seasonality in Regression*. Orlando, FL: Academic Press.

Hylleberg, S., ed. 1992. *Modelling Seasonality*. Oxford: Oxford University Press.

Hylleberg, S. 1995. Tests for seasonal unit roots: general to specific or specific to general. *Journal of Econometrics* 69, 5–25.

Hylleberg, S., Engle, R., Granger, C. and Yoo, S. 1990. Seasonal integration and cointegration. *Journal of Econometrics* 44, 215–38.

Hylleberg, S. and Pagan, A. 1997. Seasonal integration and the evolving seasonals model. *International Journal of Forecasting* 13, 329–40.

Jevons, W. 1884. *Investigations in Currency and Finances*. London: Macmillan.

Johansen, S. 1995. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Johansen, S. and Schaumburg, E. 1999. Likelihood analysis of seasonal cointegration. *Journal of Econometrics* 88, 301–39.

Koop, G. and Dijk, H. 2000. Testing for integration using evolving trend and seasonals models: a Bayesian approach. *Journal of Econometrics* 97, 261–91.

Kunst, R. 1997. Testing for cyclical non-stationarity in autoregressive processes. *Journal of Time Series Analysis* 18, 123–35.

Kwiatkowski, D., Phillips, P., Schmidt, P. and Shin, Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root – how sure are we that economic time series have a unit root? *Journal of Econometrics* 54, 159–78.

Lee, H. 1992. Maximum likelihood inference on cointegration and seasonal cointegration. *Journal of Econometrics* 54, 1–47.

Lovell, M. 1963. Seasonal adjustment of economic time series. *Journal of the American Statistical Association* 58, 993–1010.

Mazzi, G.L. and Savio, G. 2005. The seasonal adjustment of short time series. Technical Report KS-DT-05-002. EUROSTAT.

Mills, F. 1924. *Statistical Methods*. London: Pitman.

Miron, J. 1996. *The Economics of Seasonal Cycles*. Cambridge, MA: MIT Press.

Miron, J. and Zeldes, S. 1988. Seasonality, cost shocks and the production smoothing model of inventories. *Econometrica* 56, 877–908.

OECD. 1999. Feature article: seasonal adjustment. *Main Economic Indicators*. November. Paris: OECD.

Osborn, D. 1988. Seasonality and habit persistence in a life cycle model of consumption. *Journal of Applied Econometrics* 3, 255–66.

Osborn, D. 1991. The implications of periodically varying coefficients for seasonal time-series processes. *Journal of Econometrics* 48, 373–84.

Shiskin, J., Young, A. and Musgrave, J. 1967. The X-11 variant of the census method II seasonal adjustment program. Technical Paper No. 15. Washington, DC: US Census Bureau.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Taylor, A. 2005. Variance ratio tests of the seasonal unit root hypothesis. *Journal of Econometrics* 124, 33–54.

Vahid, F. and Engle, R. 1993. Common trends and common cycles. *Journal of Applied Econometrics* 8, 341–60.

Wallis, K. 1998. Comment. *Journal of Business and Economic Statistics* 16, 164–5.

Wells, J. 1997. Business cycles, seasonal cycles, and common trends. *Journal of Macroeconomics* 19, 443–69.

How to cite this article

Hylleberg, Svend. "seasonal adjustment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000496> doi:10.1057/9780230226203.1500

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

second best

Peter Bohm
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Lagrange multipliers; marginal cost pricing; marginal rate of substitution; marginal rate of transformation; Pareto efficiency; public finance; second best; trade creation; trade diversion

Article

One of the passages most often quoted in the literature on economic policy is the following from a seminal paper by R.G. Lipsey and K. Lancaster (1956, p. 11):

> The general theorem for the second best optimum states that if there is introduced into a general equilibrium system a constraint which prevents the attainment of one of the Paretian conditions, the other Paretian conditions, although still attainable, are, in general, no longer desirable.

The implication of this theorem was that most of the simple and general guidelines for policy provided by welfare economics – for example, the ‘Paretian conditions’ stating that price should equal marginal cost – would not be relevant for real-world economies which are likely to be subject to constraints on policy. The Lipsey–Lancaster article seems to have come as a shock to economists in general and has since had a significant impact on the theory, and practice, of economic policy. Apparently, until the publication of this article, the conventional wisdom was that it was desirable to pursue a ‘piecemeal policy’, here and there fulfilling the ‘Paretian conditions’ – which, if applied everywhere, would lead to a Pareto optimum – regardless of whether these conditions actually were attained elsewhere.
This state of affairs in 1956 was somewhat puzzling considering that the Lipsey–Lancaster conclusion was not entirely novel. As early as 1909, V. Pareto himself had argued that free trade (which in modern terminology may be interpreted as fulfillment, as far as possible, of the ‘Paretian conditions’) may not be preferable to protection and that individuals may not end up in a better position if one of several distortions to resource allocation were eliminated. Both of these arguments are in line with the general theory of second best. Even closer to the Lipsey–Lancaster result was the statement by Paul Samuelson

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in his *Foundations* (1947, p. 252) that a 'given divergence in a subset of the optimum conditions necessitates alterations in the remaining ones'.

Second best reasoning had also been prevalent in various areas of applied welfare economics. (For reviews using different perspectives, see Lipsey and Lancaster, 1956; Negishi, 1972; and McKee and West, 1981.) Thus, concerning *optimal pricing* in the presence of monopolies and other forms of imperfect competition, Hicks (1940) had argued that price equal to marginal cost in an industry A is not compatible with efficiency if other industries B are not competitive. As marginal inputs in A at the expense of B are then worth more than is reflected by input prices, the marginal cost confronting A is less than the true social marginal cost and therefore unsuitable as a benchmark for the price in this industry.

Early 'second best results' had also been obtained in *public finance*. For example, given that leisure is untaxable, ordinary income taxation cannot be considered more efficient than indirect taxation of one commodity, as in both cases at least one 'Paretian condition' (that between commodities and leisure vs. that between the commodity subject to an indirect tax and the other commodities, respectively) is violated. Hence, it is an open question which tax system is the better of these two imperfect alternatives (Little, 1951). Somewhat later, it was shown that a 'second best optimal' way of raising a given amount of government revenue, barring the use of lump-sum taxes, was a set of unequal indirect taxes with low tax rates on commodities which are substitutes for leisure and high tax rates on commodities complementary to leisure (Corlett and Hague, 1953). In fact, a similar result had already been obtained by Ramsey in 1927. These cases clearly illustrate that when all 'Paretian conditions' cannot be met, it may not be efficient to fulfil some of them.

The field of *trade policy* had been especially rich in providing examples of second best reasoning. Viner (1950) showed that in a world of trade protection, a reduction of some trade barriers or introduction of a customs union for some of the trading countries – both of which constitute steps towards free trade and the fulfillment of some of the 'Paretian conditions' – will not necessarily increase efficiency in world production. In the customs union case, the explanation is that the positive welfare effect of trade creation within the union may be outweighed by the negative welfare effect of trade diversion between member countries and the rest of the world.

Meade (1955a; 1955b) dealt with a number of trade policy problems as well as some domestic policy problems where it would not be possible to reach a Pareto optimum – or Utopian optimum, as he tellingly called it. Assuming the existence of several market imperfections and efficiency-distorting policy interventions, he analysed the effect of reducing or eliminating one of them and tried to determine what policy rule would perform better. Meade coined the term 'theory of second best' for this type of policy analyses whose real-world relevance and basic similarities were elucidated especially in his *Trade and Welfare* (1955a).

In retrospect, it may be argued that the catalogue of separate but similar policy issues in distorted economies provided by Meade goes as far as the theory of second best has reached. But Lipsey and Lancaster were the ones who put second best problems on the map of the average economist. This was accomplished by their attempt to present a general theory of second best, covering the main characteristics of the particular cases dealt with by Meade and others up to that point. Although their 1956 article contained a number of comments on these particular cases as well as reservations on their general theory, it was their concise version of this theory that gained most of the attention.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The Lipsey–Lancaster general theorem of the second best departs from ‘the typical choice situation in economic analysis’ where an objective function $F(x_1, \ldots, x_n)$ is to be maximized or minimized subject to a constraint $\Phi(x_1, \ldots, x_n) = 0$. Lipsey and Lancaster called the solution to this problem the Paretian optimum. To make the problem explicitly harmonize with what is generally meant by a Paretian optimum, we may interpret $x_1,\ldots, x_n$ as the elements of the consumption vectors of all individuals in the economy in some given order (x_1 being Alpha's consumption of commodity I, x_2 Alpha's consumption of commodity II, and so on, up to x_n, the last individual's consumption of the last commodity). As in most of the literature on the subject, we assume for simplicity that the objective function reveals an interest in efficiency alone (that is, attaining any Pareto optimum) and not in distribution (that is, one particular Pareto optimum).

Optimizing $F(\cdot)$ subject to $\Phi(\cdot)=0$, where Φ can be seen as the transformation function specifying the constraint given by available technology and initial resources, we get the following necessary optimum conditions (assuming that all functions are well behaved):

$$F_i / F_n = \Phi_i / \Phi_n i = 1, \ldots, n - 1. \tag{1}$$

These ‘Paretian conditions’ – or first best Pareto optimum conditions as they are now commonly called – may be interpreted as requiring equality between the marginal rates of substitution and the marginal rates of transformation. The purpose of deriving these conditions in the present context, it should be stated explicitly, is (*a*) to check whether they are fulfilled in a particular situation and, if they are not, (*b*) to provide guidelines for policy.

Lipsey and Lancaster then tried to formulate an additional constraint which would cover most of what the literature had observed as obstacles to achieving a first best Pareto optimum. They attempted to accomplish this about as generally as when the function Φ is taken to represent the production constraint of the economy. They argued that, if for some reason monopoly elements, externalities or other so-called imperfections were ‘out of bounds’ for policy intervention, one of the conditions (1) could not be fulfilled due to a constraint

$$F_1 / F_n = k\Phi_1 / \Phi_n \tag{2}$$

with $k \neq 1$ and k – ‘for simplicity’ – assumed to be constant. The resulting problem amounts to the optimization of the Lagrangean function

$$F - \lambda\Phi - \mu(F_1 / F_n - k\Phi_1 / \Phi_n)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(3)

where λ and μ are Lagrangean multipliers. A solution to this problem, it should be noted, is also a Pareto optimum as it does not allow anyone to become better off without making someone else worse off, given the two constraints now in force. The necessary optimum conditions can be written

$$\frac{F_i}{F_n} = \frac{\Phi_i + (\mu / \lambda)(Q_i - kR_i)}{\Phi_n + (\mu / \lambda)(Q_n - kR_n)}, \; i = 2, \ldots, n-1$$

(4)

where

$$Q_i = (F_n F_{1i} - F_1 F_{ni}) / F_n^2 \quad \text{and} \quad R_i = (\Phi_n \Phi_{1i} - \Phi_1 \Phi_{ni}) / \Phi_n^2.$$

Aside from some special cases (see below), this means that, in second best optimum, $F_i / F_n \neq \Phi_i / \Phi_n$, from which follows the theorem quoted in the introduction.

These second best optimum conditions are obviously quite complicated – in fact, so complicated that in many cases a great deal of detailed information would be required even to determine the signs of the second derivatives F_{ni}, Φ_{ni}, etc. Hence, it would often be impossible to know whether in second best optimum $F_i / F_n > \Phi_i / \Phi_n$ or the opposite. Moreover, it is no longer possible to translate the second best optimum conditions into intuitively simple relationships between price and marginal cost, which was true for (some of) the first best optimum conditions (eq. 1).

The great impact of this result on economists in general must be attributed to the simplicity of the theorem itself, given that, in essence, the same thing had been said on earlier occasions. A large part of the ensuing debate concerned whether this simplicity was warranted by real-world conditions. In particular, (*a*) the origin and form of the additional constraint were questioned. A second dominating issue in the debate concerned (*b*) the complexity of the rules for second best policy and attempts at identifying important cases where simple first best optimum conditions are still valid in second best situations. We deal with these two issues in turn.

The nature of the additional constraint

In regard to the alleged generality of the formulation of the Lipsey–Lancaster theory, the question was asked: What exactly does this constraint (2) stand for? Clearly, the optimization problem in the ‘second best literature’ refers to a national government *or* an independent unit of government (such as a public

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
monopoly) which operates as if it tried to optimize function F. Obviously, the government-unit perspective (represented for example, by, the work of Davis and Whinston, 1965; 1967, and developed by McFadden, 1969) could allow a number of constraints like (2) concerning variables that are out of reach for this unit. But what about additional constraints imposed on the overall allocation problem confronting the *national government?* Here, two interpretations have been considered in the literature. *First*, for an initial state of the economy in which one of the conditions (1) is violated, it may be *technically* impossible for the government to have this condition fulfilled along with all the others. For example, markets for certain commodities such as specific kinds of insurance may not be possible to introduce or may be too costly to administer. Or it may be impossible or prohibitively expensive to correct for certain externalities or instances of imperfect competition. Likewise, when the government, in an attempt to attain a feasible Pareto optimum, needs to raise money for subsidies or for the production of public goods, there may not be any non-distortive taxation scheme available (note, for example, the ‘impossibility’ of taxing leisure assumed above).

It should be noted that constraints of this technical type are irremovable, in the same way that the constraint imposed by the transformation function is irremovable. Thus, if a market economy does not by itself reach a first best Pareto optimum, the government could not reach one either when policy constraints of this type are strictly binding. Then, a Lipsey–Lancaster Paretian optimum does not exist and the only optimum conceivable is in fact what has here been called a second best optimum (see, for example, McKee and West, 1981).

Second, the constraint (2) can be interpreted as a *behavioural* constraint on policy, reflecting the fact that certain measures, although technically speaking feasible and in principle capable of removing the restraint, are not at the government's disposal or just not believed to be so. For example, the law may prohibit or delimit the use of a specific policy instrument. Or the government may have other and hierarchically higher goals than Pareto optimality: it may for example, simply dislike nationalization of certain industries. Or the government may want to avoid the use of a policy for ‘political reasons’, believing, for example, that it would lose the next election if this policy was used. Traditions, idiosyncrasies, and so on could play a similar role. Economists who have paid attention to the origin of constraints like (2) seem to have adhered primarily to this ‘behavioural’ interpretation.

Obviously, the ‘behavioural’ type of constraint need not be such that all policies with the same effects on the objective function are restrained to the same extent. (For a different perspective which holds that constraints are or should be, in some narrow sense, ‘rational’, see Faith and Thompson, 1981.) For example, assume that there are two policy instruments, say, a tax and a regulation, each of which, if unconstrained, would have attained the first best, as k could then be made equal to 1, but that policy now is constrained so that just one of them is ruled out. If so, a constraint of type (2) would not exist. This means that, in general, it is not possible to presume what constraints on policy instruments imply in terms of the relations between endogenous variables such as the marginal rates of transformation and substitution in (2). (Actually, the literature has not been able to present any great number of cases where policy constraints yield an expression like (2) with k constant and not equal to one.) Instead, to obtain a solution to the allocation problem, it must be specified exactly what the actual constraints on policy instruments are, that is, to what domain or what combinations with other instruments or variables each policy instrument is restricted.

Specifying the policy constraints in this way has some important implications for a general theory of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
second best (McManus, 1959; 1967; Bohm, 1967).

First, it cannot be known beforehand whether policy constraints prevent the attainment of a first best optimum or not.

Second, the actual impact of policy constraints, including now also the technical ones mentioned earlier, will depend on the actual behavioural properties of the agents operating in the market economy. Thus, actual rules of market behaviour would, at least in principle, have to be added as constraints on the objective function along with the constraints on policy.

Third, the many possible forms of the policy constraints imply that there cannot be any *general* second best optimum conditions in the sense that there exist general first best optimum conditions (when no policy constraints exist) in terms of a specific relationship between marginal rates of substitution and marginal rates of transformation. In fact, constraints on policy instruments will require that the optimum-feasible solution be derived directly in terms of optimum-feasible values for the policy instruments. Thus, there would not even be any role for policy guidelines in the form of second best conditions such as (4). This, of course, does not preclude the existence of special cases, which *ex post* turn out to coincide with the Lipsey–Lancaster formulation. These are the cases where the constraint on policy happened to affect only the relationship between one marginal rate of substitution and one marginal rate of transformation exactly in the way specified by (2), with no impact whatsoever on the use of policy instruments that could influence other such rates in the economy.

Given that additional constraints on the allocation problem can have any shape – with (2) being a possible *ex post* formulation of one of many special cases – the use of the term ‘second best’ becomes somewhat unclear. Should a *second best problem* be defined as an allocation problem with constraints on policy regardless of whether a first best optimum will turn out to be impossible? Or should it be reserved for such problems where analysis will eventually show that a first best optimum cannot be reached? Although the second alternative is in line with the intended problem formulation in Lipsey–Lancaster, it is obviously inconvenient as it cannot be used until after the problem has been solved. The term *second best optimum*, on the other hand, is predominantly used for a constrained optimum not equal to a first best optimum and is not likely to cause much of a problem. Hence, a practical, and nowadays probably the dominant, terminology is to distinguish between first best and second best problems according to the first-mentioned definition, where a second best problem may have a first best *or* a second best optimum solution.

First best rules for second best problems

To deal with the second issue prominent in the literature on second best, we return to the problem as it was formulated by Lipsey and Lancaster. They had argued that second best conditions, in contrast to first best ones, were so complicated and required so much information that, on this account, the conditions could not be used for practical policy. This spurred a number of economists to undertake a rescue operation, in which they tried to show that in many instances the simple first best conditions would still be relevant for the controllable part of the economy. To the extent this was true, it would restore at least part of the relevance of piecemeal policy, that is, policy guided by principles which are unaffected by the exact nature of the circumstances in the uncontrollable part of the economy.

First, it has been pointed out that first best rules may be optimal even with the particular Lipsey–

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lancaster formulation of the second best problem, for example when $Q_i–kR_i=Q_j–kR_j=0$; $i, j=2,\ldots, n$; $i \neq j$ (see Santoni and Church, 1972; Dusansky and Walsh, 1976; Rapanos, 1980). Similar special cases may be found for other forms of additional constraints on the objective function (Mishan, 1962).

Second, and of more general interest for practical policy, it was pointed out that if the additional constraints affect only a limited set of markets in the economy, the relation between this sector and the remaining 'perfectly controllable' sector may be such that first best conditions are optimal in the latter sector even when they turn out to be unattainable in the former sector. This must be true, of course, if the two sectors are completely independent of each other with different primary inputs, no input deliveries between the sectors and different final consumers. Approximately the same results hold if interdependence between sectors is negligible, for example due to one of the sectors being relatively small (Mishan, 1962).

Other attempts at identifying similar cases of separability have been made without much success in terms of general and easily applicable principles for ascertaining when the first best optimum conditions that are still attainable should in fact be attained. Moreover, the very idea of identifying two sectors, one of which is imperfectly controllable, has appeared to be less and less attractive as a description of the real world. Instead, in most countries, income taxes, distorting institutional rules and regulations, and so on emerge as irremovable constraints affecting the economy as a whole. Moreover, as the typical allocation problems confronting real-world governments are beset with a multitude of policy constraints – technical as well as behavioural – it is only by pure chance that optimum conditions for the irrelevant first best problem can be found to be a priori relevant. This does not mean, of course, that first best optimum conditions never will *turn out* to be valid in second best optimum or that information may not be so inadequate that these conditions appear to be acceptable as a rule of thumb (see Ng, 1977).

Thus, the outcome of the literature on general second best theory up to this point is disillusioning in several respects. There do not seem to be any *general* second best problems of the type formalized by Lipsey and Lancaster, much less any general second best optimum *conditions*. Granted that there remains some disagreement on the purpose of second best theory, what has emerged from the literature by way of a general description of second best problems can be summarized as follows:

All economies – even real-world centrally planned economies – have at least some (most often, a very large number of) given behaviour functions which contribute to determining the outcome of any policy 'intervention'. Hence, these functions must be observed in the formulation of the allocation problem; or, which is the same thing, they must be included as constraints on the optimization of the objective function. 'At the same time, the authorities have at their command a set of policy instruments which enter these functions as arguments. The optimum is then found by [optimizing the objective function] subject to all the constraints over the domain of these instruments. Indeed the whole problem has little practical interest without some such explicit policy formulation' (McManus, 1967, p. 321). The fact that this is a highly demanding analytical task in actual practice requires simplifications and approximations of the models to be used, but it cannot justify an oversimplification of the actual problem to be tackled. This in effect may seem to take us back to the case-by-case approach of applied welfare economics that was used by Meade and others in the beginning of the 1950s (represented in later and technically more elaborate studies by, for example, Boiteux, 1956; Rees, 1968; and Guesnerie, 1975). However, matters have changed in one important respect since then. Much more empirical knowledge is now available concerning behaviour of individual markets which in itself improves the outlook for practical second

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
best policy.

This is not to say that attempts to construct a general theory of second best have not made a significant contribution to economic theory and policy. Should one result be highlighted, it may quite likely be the general theorem of second best as quoted in the introduction. If nothing else, this theorem has probably made economists more careful when providing governments with policy prescriptions.

See Also

- marginal and average cost pricing
- optimal tariffs
- Pareto efficiency

Bibliography

Allingham, M. and Archibald, G.C. 1975. Second best and decentralisation. *Journal of Economic Theory* 10, 157–73.

Boadway, T.J. and Harris, R. 1977. A characterisation of piecemeal second best policy. *Journal of Public Economics* 8(2), 169–90.

Bohm, P. 1967. On the theory of ‘second best’. *Review of Economic Studies* 34, 301–14.

Boiteux, M. 1956. Sur le gestion des monopoles publics astreints à l’équilibre budgetaire. *Econometric*a. Translated into English as: On the management of public monopolies subject to budgetary constraints, *Journal of Economic Theory* 3(3), (1971), 219–40.

Corlett, W.J. and Hague, D.C. 1953. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21, 21–30.

Davis, O.A. and Whinston, A.B. 1965. Welfare economics and the theory of second best. *Review of Economic Studies* 32, 1–14.

Davis, O.A. and Whinston, A.B. 1967. Piecemeal policy in the theory of second best. *Review of Economic Studies* 34, 323–31.

Dusansky, R. and Walsh, J. 1976. Separability, welfare economics, and the theory of second best. *Review of Economic Studies* 43, 49–51.

Faith, R. and Thompson, E. 1981. A paradox in the theory of second best. *Economic Enquiry* 19, 235–44.

Guesnerie, R. 1975. Production of the public sector and taxation in a simple second best model. *Journal*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Economic Theory 10, 127–56.

Hatta, T. 1977. A theory of piecemeal policy recommendations. *Review of Economic Studies* 44, 1–21.

Hicks, J.R. 1940. The rehabilitation of consumers' surplus. *Review of Economic Studies* 8, 108–16.

Kawamata, K. 1977. Price distortion and the second best optimum. *Review of Economic Studies* 44, 23–29.

Lipsey, R.G. and Lancaster, K. 1956. The general theory of second best. *Review of Economic Studies* 24 (1), 11–32.

Little, I.M.D. 1951. Direct versus indirect taxes. *Economic Journal* 61, 577–84.

McFadden, D. 1969. A simple remark on the second best Pareto optimality of market equilibria. *Journal of Economic Theory* 1, 26–38.

McKee, M. and West, E.G. 1981. The theory of second best: a solution in search of a problem. *Economic Inquiry* 19, 436–48.

McManus, M. 1959. Comments on the general theory of second best. *Review of Economic Studies* 26, 209–24.

McManus, M. 1967. Private and social costs in the theory of second best. *Review of Economic Studies* 34, 317–21.

Meade, J.E. 1955a. *Trade and Welfare*, [*including the*] *Mathematical Supplement*. London, New York: Oxford University Press.

Meade, J.E. 1955b. *The Theory of Customs Unions*. Amsterdam: North-Holland.

Mishan, E.J. 1962. Second thoughts on second best. *Oxford Economic Papers* 14, 205–17.

Negishi, T. 1972. *General Equilibrium and International Trade*. Amsterdam: North-Holland.

Ng, Y.K. 1977. Towards a theory of third-best. *Public Finance* 32(1), 1–15.

Pareto, V. 1909. *Manuel d'Economie Politique*. Paris: Girard et Brière.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37, 47–61.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Rapanos, V.T. 1980. A comment on the theory of second best. *Review of Economic Studies* 47, 817–19.

Rees, R. 1968. Second-best rules for public enterprise pricing. *Economica* 35, 260–73.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Santoni, G. and Church, A. 1972. A comment on the general theorem of second best. *Review of Economic Studies* 39, 527–30.

Sontheimer, K.C. 1971. An existence theorem for the second best. *Journal of Economic Theory* 3, 1–22.

Viner, J. 1950. *The Customs Union Issue*. New York: Carnegie Endowment for International Peace; London: Stevens & Sons.

How to cite this article

Bohm, Peter. "second best." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000060> doi:10.1057/9780230226203.1501

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

second economy (unofficial economy)

Michael Alexeev
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The second (shadow, unofficial) economy played a major role in Soviet-type economies (STEs) and served as a precursor of unofficial sectors in the transition economies. This article lists the main causes of the second economy, outlines approaches to its measurement, and examines its effect on the economic performance of the STEs. The main focus is on the unofficial economy during transition to markets. The article presents estimates of the unofficial economy, describes negative externalities that it generates, discusses reasons for its different size and dynamics in different countries, and outlines approaches to reducing it. Positive aspects of the unofficial economy are also noted.

Keywords

bribery; command economy; corruption; developing economies; entrepreneurship; informal economy; liberalization; path dependency; price control; public goods; rationing; repressed inflation; rule of law; second economy; shadow economy; Soviet economic reform; speculation; state-owned enterprises; tax evasion; transfer pricing; transition and institutions; trust; unofficial economy; wealth effects

Article

The second economy in the Soviet-type command economies (STEs) was defined by Grossman (1977) as all economic activities that are either undertaken directly for private gain or are knowingly illegal in some substantial way. Both legal and illegal economic activities fall within this definition. The second economy served as a precursor of the unofficial sector in the economies in transition. Accordingly, the institution of unofficial economy in transition is properly understood as an heir to the second economy in the STEs.

The official primary coordination mechanism in the STEs was central planning. The well-known difficulties with implementing this mechanism created both the incentives and the opportunity for the existence of a large second economy. The scope of the legal part of the second economy depended on

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the laws of a given STE, but it usually included private agriculture, small-scale construction, certain professional services (for example, lawyers working privately), and so on. While the legal second economy was important, it was usually smaller than the illegal sector, also referred to as unofficial, underground, shadow economy, or black or parallel markets. Moreover, the concept of legal second economy is not applicable to market economies. Accordingly, this article focuses on the unofficial economic activities both in the STEs and during transition to markets, with less attention devoted to the legal second economy in the STEs. For the economies in transition, the unofficial economy is defined here as all value-added market-based economic activities that either evade taxation or are not registered, or both.

The most common illegal economic activities performed by individuals in the former USSR and other STEs were the theft of state property, corruption, and 'speculation' defined as resale of goods by individuals for profit (Grossman, 1979). While the first two types of activities are present in all modern economies, the dominant role of the state and state-owned enterprises made these activities particularly widespread in the command economies. The general illegality of speculation was unique to the STEs. Other significant illegal economic activities included unofficial production by individuals and small teams, much of it taking place at state-owned enterprises on company time and using state-owned equipment.

The unofficial economic activities among individuals in the STEs were mainly caused by pervasive price controls, prohibition of many private economic activities that are common in market economies, high taxes on the permitted activities, ubiquitous and poorly protected public property, enormous discretionary power of the bureaucracy, and social attitudes that accommodated economic crimes that did not directly harm specific individuals. None of these causes was unique to the STEs, but their rarely found confluence provided a hospitable environment for the unofficial economy, so that the entire system could be called a quasi-market economy (Leitzel, 1995). In many STEs the second economy was a normal part of everyday life for a typical consumer (see Grossman, 1977; 1979; 1989, for fascinating descriptions related to the former USSR).

A large 'shadow economy' existed also among socialist enterprises. Much of it had to do with the informal or semi-formal exchanges of goods between state-owned enterprises performed without the prior direction by the planners. Such exchanges, even when legal by themselves, were often accompanied by illegal side payments and other informal inducements. In addition, enterprises sometimes hired labour or acquired other inputs from individuals in the unofficial economy.

Measuring the unofficial economy

The more or less precise dimensions of the illegal economy are all but impossible to ascertain. The approaches to measuring it in various countries are reviewed by Schneider and Enste (2000). With respect to STEs and the economies in transition, the two most commonly used methods have been the surveys of either consumers or firms, and the electricity consumption method. Both approaches have certain advantages and disadvantages. Unless a survey is conducted among emigrants, as was usually the case prior to the collapse of the Communist regimes, the survey questions about unofficial economy are often indirect. For example, the consumers may be asked about their expenditures and official incomes and savings. If reported expenditures exceed legitimate incomes, the difference can serve as a basis for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
estimating unofficial income. In surveys of firms, the managers may be asked about hidden sales in their industry rather than in their specific firm. The reliability of the survey method depends crucially on the sample selection and on the veracity of the respondents' answers.

The electricity consumption approach assumes a close connection between electricity consumption and the true size of economic activity in a country. In the simplest case, the elasticity of electricity consumption with respect to the combined official and unofficial GDP is taken to be one, implying that the dynamics of electricity output coincides with the dynamics of total GDP. Assuming some initial size of the unofficial economy and knowing the change of the official GDP would then permit calculating the change in the size of the unofficial sector. The two main disadvantages of this method are the need to know the initial size of the unofficial economy and the elasticity of electricity consumption by GDP. The latter information is particularly difficult to infer for the economies in transition that are undergoing structural reforms and experiencing large changes of relative prices. However, an important advantage of this approach is its ease of implementation in a uniform fashion across time and countries. Moreover, it can be used to extrapolate a point estimate of the unofficial economy both into the future and into the past.

Among the major STEs, survey-based estimates of the second economy have been published only for the former USSR. The estimates based on the surveys of Soviet emigrants relate to the second half of 1970s and range approximately between ten per cent and 30 per cent of incomes of urban households. (The lower estimate is due to Ofer and Vinokur, 1992. The higher estimate is from Grossman, 1991.) An alternative set of estimates based on the Soviet-era official family budget survey data puts the second economy at around 23 per cent of household income (Kim, 2003). Using these estimates to obtain the shares of second economy in GDP is difficult, however, particularly because no estimates of the shadow economy among the socialist enterprises have been published. The size of the second economy also varied greatly across the former Soviet republics. In the Caucuses and in central Asian republics both legal and illegal private economic activities were widespread, while in the Baltic republics the second economy was relatively insignificant (Grossman, 1991; Kim, 2003; Alexeev and Pyle, 2003).

Kaufman and Kaliberda (1996) cite a number of micro studies of the unofficial economy in east European countries in the early 1990s and extrapolate these estimates backwards to 1989 using the electricity consumption method. The results range from six per cent of GDP in Czechoslovakia to over 20 per cent of GDP in Bulgaria, Romania and Hungary.

The unofficial sectors in the economies in transition have been measured by the above methods as well as by currency demand method and the latent variable estimation. The former approach estimates a currency demand function for a country and assumes that all 'excess' demand for cash relative to this demand function is due to the growth of the unofficial economy. The latter approach uses latent variable estimation techniques to infer the size of the unofficial economy as an unobserved variable influenced by a number of different factors and affecting several observed economic indicators. Each of these methods has its strengths and weaknesses that are discussed in Schneider and Enste (2000) and Schneider (2005). The latter work combines the two approaches to estimate the unofficial economy as a share of GDP in 110 countries in year 2000 and some earlier years. The 2000 estimates for the economies in transition range from 13.1 per cent in China to over 60 per cent in Georgia and Azerbaijan.

The quantitative importance of the second economy raises the issue of the nature of its impact on welfare, on the 'first' economies, and on the attempts at partial reform. More important for the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economics of transition is the transformation of the relationship between the official and unofficial economies and the dynamics of the latter as a result of radical market-oriented reforms.

The second economy's impact on the STEs

The second economy appears to be necessary for the survival of an STE. The marginal effect of the second economy's expansion is less clear. A growing second economy increases the influence of market forces on a STE, obviating various types of non-price rationing and correcting some resource misallocations. For instance, rationing was undermined even for such a seemingly easy-to-control good as residential housing (Alexeev, 1988). Also, certain theoretical concepts applied to a STE such as ‘forced savings’ or ‘monetary overhang’ become questionable in the presence of a developed second economy (Alexeev, Gaddy and Leitzel, 1991). On the other hand, second economy growth may have significant negative effects, as is discussed later.

Ericson (1983) provides the first formal analysis of the second economy's impact on the official sector. In his model, enterprise manager *i*, *i*=1,…, *n*, receives a planned output target, $y^i \in R_+$, and the input allocations given by $\omega^i \in R_+^{m+1}$, where the first component, ω_0, represents the amount of official funds and the last m components denote the amounts of material inputs. The production function is described by $f^i: X^i \to R_+$, where $X^i \subset R_+^m$ is the set of all feasible combinations of material inputs. Given f^i, the plan is not necessarily feasible. The manager's preferences are represented by a utility function monotonic in $x_0^i \in R_+$, and $y^i = f^i(x^i)$. If the initial inputs are misallocated, the managers have incentives to trade legally at the official prices $p = (1, p_1, \ldots, p_m)$. However, these fixed prices are not likely to equilibrate the secondary market. The disequilibrium generates incentives for the managers to induce other managers to trade by offering informal side-payments from the manager's hidden slush fund of loose cash, the initial amount of which is c^i. The flexible prices in this informal market are denoted by $a = (0, a_1, \ldots, a_m) \in R^{m+1}$. It is also assumed that a fixed proportion, α, of this cash ‘sticks’ to the palms of individuals who handle the transactions. Given this leakage, the manager faces the following problem (index *i* is suppressed):

$$\mathrm{Max}_z\{u(x_0, f(x))\} \text{ s.t. } pz \le 0 \text{ and } az \le c - \alpha(a_+ z_- + a_- z_+),$$

where $z_j = x_j - \omega_j, \quad j = 0, 1, \ldots, m$, represent trades of inputs among managers and cash prices of material inputs may be either positive or negative, depending on whether payment is necessary to induce a purchase or sale. In addition, $z_+ \equiv \max(0, z)$ and $z_- \equiv \max(0, -z)$, with a_+ and a_- defined similarly. Note that a_0=0, reflecting the assumption that the official funds cannot buy cash and there is no need for financial inducements to transfer official funds. (As is discussed below, the wall between the official funds and loose cash had seriously eroded by the late 1980s as a result of partial reforms in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
USSR and other STEs. This erosion was in part responsible for the eventual collapse of the command economic system.)

Under certain technical conditions, this exchange economy has a suitably defined equilibrium, and any trades that occur with the inducement of loose cash represent a Pareto improvement. Given the managers' preferences, the second economy in this model always enhances plan fulfillment. (This equilibrium, however, is not in general constrained Pareto-optimal – where the 'constraint' is given by $pz \leq 0$ – because the leakage might stop trading before Pareto optimal allocation is reached.)

Of course, the first economy and the overall economic welfare may suffer if the official plans are inefficient. Also, the second economy always enhances plan fulfilment in Ericson's framework only because enterprise managers' preferences do not exhibit a trade-off between plan fulfillment and personal gain from unofficial dealings. Ericson's model takes no account of the potential negative externalities that the second economy transactions can create in an otherwise distorted economy. Some of these externalities are examined by Stahl and Alexeev (1985) in a model of a queue-rationed exchange economy. Here, N consumers face inelastic supply of goods $X \in R^n_{++}$ with a fixed official price $p_1 \in R^n_{++}$. If a shortage arises, the demand is rationed by queues with a deterministic waiting time $\tau \in R^n_+$. Consumers are endowed with money, M^i, and leisure, L^i. Consumer i has well-behaved preferences $u^i(x^i, l^i)$ with respect to consumption of goods, $x^i \in R^n_+$, and leisure, $l^i \in R_+$. In the absence of black markets, each consumer solves (superscript i is suppressed):

$$\text{Max}_{x,l}\, u(x, l) \text{ s.t. } p_1 x \leq M \text{ and } \tau x + 1 \leq L$$

An appropriately defined equilibrium that requires no excess demand always exists.

Black markets are introduced by allowing consumers to resell goods they have purchased in the official markets. Denote the black market trades by $y^j \in R^n$, with $y^j_j > 0$ if the consumer buys the corresponding good and $y^j_j < 0$ if he sells. Black market prices, $p_2 \in R^n_+$, are flexible. The new budget constraints become (superscript i is suppressed):

$$p_1(x - y) + p_2 y \leq M$$

$$\tau(x - y) + 1 \leq L$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

A necessary condition for no excess demand is $p_2 = p_1 + wT$, where *w* is interpreted as an implicit wage for standing in line. Accordingly, the additional requirements for an equilibrium include $p_2 = p_1 + w^* T^*$ and $\Sigma y^* = 0$. Equilibrium exists if $\Sigma M \leq p_1 X$. Given the aggregate queuing time and p_1, the equilibrium with black markets is efficient while this is not usually true in the absence of black markets.

The main insight, however, is that the introduction of black markets does not always result in a Pareto improvement over pure queue rationing. This happens because black markets create wealth effects that may result in longer queues. Nonetheless, in this framework the poor typically prefer queuing with black markets to pure queue rationing (Polterovich, 1993). The comparisons for the rich are ambiguous.

In addition to the above effects, the unofficial economic activities both in the consumption sector and among socialist enterprises reduce the feedback to the policymakers from their actions by covering up policy consequences and resource misallocations, and rendering the official statistics, including those on consumer incomes, savings, consumption and employment, less relevant. Also, the efficiency of unofficial transactions is reduced by the need for secrecy, which impedes information flows within black markets and limits the range of available contract enforcement mechanisms, resulting in greater uncertainty and a suboptimally small scale of operations, among other things. Moreover, the potential profitability of illegal transactions breeds corruption and weakens the official institutions.

On balance, the effect of the second economy on the official economy and on overall welfare is theoretically ambiguous on the margin.

The second economy's impact on the pre-reform crises and on reforms within an STE

As a command economy matures and becomes increasingly complex, the functioning of the traditional administrative coordination mechanism deteriorates. Many STEs responded to their worsening economic situation by introducing reforms aimed at liberalizing parts of the economy, particularly labour markets and consumer-oriented sectors. Such partial reforms had generally failed to improve overall economic performance, largely because of the second-best considerations. Black markets, however, may alleviate some of the adverse consequences of partial reforms. For example, Boycko (1992) and Osband (1992) demonstrate that wage increases caused by partial liberalization of labour markets in an STE lead to a cycle of greater shortages, longer queues, and, therefore, lower output if prices of consumer goods remain fixed. This ‘pre-reform crisis’ is moderated when black markets are brought into the picture, because the additional queuing is done mainly by lower-productivity agents who can then resell the goods in black markets while without the unofficial economy everybody has to queue up more (Alexeev and Sabyr, 2004). In another example, Leitzel (1998) shows beneficial effects of black markets by introducing them into a simple model of repressed inflation developed by Lipton and Sachs (1990). In Leitzel's model, retail trade employees divert goods in short supply to black markets, thereby reducing welfare losses due to rising repressed inflation and pre-empting the adverse distributional consequences of full price liberalization.

The presence of black markets may also worsen the effects of partial reforms. Consider, for example, the so-called ‘dual track’ partial reform mechanism (Lau, Qian and Roland, 2000). It attempts to conduct

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
reforms in a Pareto-improving fashion by preserving the output and supply quotas that exist in an STE while liberalizing the markets for the above quota output. If the government can enforce the pre-reform quotas, this mechanism avoids the potential severe misallocation of resources due to partial liberalization emphasized by Murphy, Shleifer and Vishny (1992) and, at the same time, prevents excessive redistribution that may result from full liberalization. However, the violation of pre-reform quotas via the unofficial trades can generate large rents. If such violations are allowed to take place, resource misallocation may easily occur.

The growth of the second economy by itself represents an implicit market reform of an STE, so that partial reforms simply provide an official permission for certain unofficial activities that have been already widespread (Leitzel, 1995). In part, the second economy growth pre-empts the potential negative effects of full liberalization, making radical market reforms politically more acceptable. In addition, the second economy familiarizes the general population with the workings of the market, develops trust and entrepreneurship, thereby facilitating popular acceptance of the market mechanism (Grossman, 1989). However, to the extent the second economy improves the functioning of the overall economy, it also reduces the potential benefits of full liberalization relative to the unreformed or partially reformed economy and, therefore, may postpone radical reforms.

Whatever the impact of the second economy on the performance of partial reforms, it is clear that historically such reforms spurred an explosive growth of the unofficial sector in the STEs. Leitzel (2003) argues that this growth was unavoidable given partial relaxation of state controls that made evasion harder to detect while leaving most of the formal distortions in place. In the USSR, for example, the permission of crypto-private firms and transactions between them and state-owned enterprises greatly reduced the ability of central planners to monitor state-owned enterprises. Enterprise managers then employed transfer pricing schemes and self-dealing to channel state-provided resources to private firms related to the managers. Tax collections plummeted while subsidies increased, undermining the state budget and the entire state sector.

The unofficial economy during transition to markets

Radical market reforms of the STEs seek to turn these economies, including most of their unofficial sectors, into well-functioning legal markets. Due to general economic liberalization, particularly the removal of such a major cause of black markets as price controls, the unofficial economies can be expected to diminish. This indeed seems to have occurred in some countries, but not in others. Using the electricity consumption approach, Johnson, Kaufmann and Shleifer (1997) estimate that between 1989 and 1995 the unofficial economy grew significantly in Russia, Ukraine and some other former Soviet republics while it remained approximately the same in Hungary, Slovakia and Estonia, and declined noticeably in Poland. (Note that the electricity consumption approach attributes the entire change – decline – in measured output relative to electricity consumption to the unofficial economy. Alternatively, the dynamics of GDP-to-electricity consumption ratios in these economies could have been influenced by relative prices and availability of electricity, government policies with respect to energy conservation, and the changing structure of the economy.)

Also, according to Johnson, McMillan and Woodruff (2000), private manufacturing firms in Russia and Ukraine tend to hide their activities from tax authorities on a dramatically greater scale than do their

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
counterparts in eastern Europe. The firms in Russia and Ukraine under-reported around 30–40 per cent of their sales and 25–40 per cent of their employee salaries, as compared with five to ten per cent in Poland, Slovakia and Romania.

On the assumption that the measurements are approximately correct, what may explain such different dynamics? One answer relies on path dependency that arises from the possibility of multiple equilibria in the relationship between the official and unofficial economies. Consider, for example, a labour allocation model by Johnson, Kaufmann and Shleifer (1997). The aggregate amount of labour normalized to one is allocated between the official economy (subscript F) and the unofficial sector (subscripted by I), $L_F + L_I = 1$. The agents in each sector $J=F, I$ maximize their utilities, $U_J=(1-t_J)Q_J$, where t_J is the tax rate and Q_J denotes the quantity of the output-enhancing public good in each sector by, respectively, the government and the mafia. (Q_J can be used only by sector J agents.) Let $Q_J=(T_J)^{\beta}$ where $T_J=t_JQ_JL_J$ is aggregate tax revenue collected in sector J. This model has an unstable interior equilibrium and two stable corner equilibria: in one $L_F=L$, $L_I=0$, and in the other $L_F=0$, $L_I=L$. Therefore, depending on the initial conditions, the economy may end up either entirely above or below ground. If parameter β is greater in the official economy than in the unofficial sector, the latter equilibrium is inefficient.

An important assumption in this model is that, while the government is hurt by labour escaping underground, it does not attempt to either stop the escape or fight the mafia that facilitates underground activities. Roland and Verdier (2003) show that one reason for government passivity may be the lack of resources for law-enforcement activities. Suppose economic agents choose between becoming producers and predators. Producers pay taxes to fund law enforcement to fight off predators. The more agents become producers, the greater are the resources for law enforcement. That is, similarly to Johnson, Kaufmann and Shleifer (1997), a fiscal externality is present. Assume also that law enforcement has fixed costs in order to become effective and that a ‘compliance externality’ exists, so that the probability of a predator being punished is inversely related to the number of predators. Under certain assumptions on the utilities of producers and predators, this model again has multiple equilibria, in some of which the government lacks the resources to fight the predators. (The government may also refrain from fighting the mafia that provides public goods underground because the presence of predatory mafia actually benefits predatory government. Alexeev, Janeba and Osborne (2004) demonstrate that, when public goods are expensive to provide, the state has higher revenues in the economy with the mafia than without it. This is because, when public goods provision is difficult and few of them are provided, the main effect of the mafia is to increases the costs of underground economic activities. This makes it possible for the state to increase its tax rate without having too many entrepreneurs escape underground.)

On the assumption that the above models reflect some essential features of the transition processes, an important policy issue is how a country can avoid a ‘bad’ equilibrium or escape one if it finds itself in it. Roland and Verdier suggest that the promise of EU accession can be one mechanism that solves coordination problems in their model described above. Obviously, this solution is not available to all economies in transition. Some argue that the collapse of government institutions and the ensuing flight of taxpayers underground can be avoided by conducting reforms gradually. China is often cited as a successful example of a gradual approach. However, such strategy is risky both because it may not avoid the deterioration of government institutions (for example, despite Ukraine's gradual reforms, the data

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
mentioned in the beginning of this section testify to the similarities between Ukraine's and Russia's unofficial economies) and because it may postpone genuine reforms indefinitely, as appears to have happened in Belarus and Uzbekistan. Friedman et al.'s (2000) findings based on a cross-section of countries suggest a rather pessimistic conclusion that the size of the unofficial economy depends largely on the quality of institutions, which in turn is determined by exogenous factors such as geography, religion, linguistic fractionalization, and the origin of the legal system, while the policy parameters that are more easily controlled by governments such as tax rates do not matter very much. These empirical results imply that over-regulation, corruption and a weak rule of law are associated with larger unofficial economy. (The positive relationship between corruption and unofficial economy is less intuitive than the other correlations. This is because the economic agents may either use the unofficial economy to escape bribe requests or bribe officials to evade taxation and regulations.)

The high tax burden can, of course, also serve as a reason for shifting resources into the unofficial economy. However, high tax rates can also generate revenues necessary to increase the output of productivity-enhancing public goods, making the official economy more attractive. The net effect of tax rates on the unofficial economy is ambiguous. We note also that the impact of taxation presumably depends on the effective tax rates that are often determined by how the tax system is administered rather than on the statutory tax rates as long as the latter are within a reasonable range.

Despite the above considerations, if the government of an economy in transition is truly intent on legalizing a significant part of the unofficial economy, the rationalization of the tax system can serve as a reasonable first step, as Alexeev, Conrad and Hay (2004) argue and as Russia's tax reforms of 2001–3 demonstrate. Ivanova, Keen and Klemm (2005) and Sinelnikov-Mourylev et al. (2003) show that a dramatic reduction of the highest marginal tax rates on personal income in Russia induced many taxpayers to legalize their incomes, leading to an increase in government revenues. Of course, tax reform alone is not sufficient. At the very least, it needs to be accompanied by reform of government regulations, administration and the courts. Unfortunately, these administrative reforms may be difficult to implement in those economies in transition that need them most.

While many economies in transition may find it difficult to reduce the unofficial sector, its size appears to be smaller than in a number of developing countries at a comparable level of development such as Mexico, South Korea and Chile. (See Schneider and Enste, 2000. Campos, 2000, argues, however, that the nature of the unofficial sectors in the economies in transition and the developing economies may be different. In the latter, the unofficial economy appears to consist mostly of small labour-intensive businesses trying to escape excessive government regulation and taxation. In the former, large firms hiding their activities account for a larger part of the unofficial economy. If true, these differences would testify to a greater degree of government capture by large firms in the economies in transition, implying that reforms aimed at reducing unofficial economies there may be more difficult to implement. This argument may be plausible, but it is hard to test given the available data and the structural differences between major economies in transition and developing economies. For example, greater capital intensity of the unofficial sector in the economies in transition can be simply a consequence of greater capital intensity of the overall economy.)

More important, the impact of the unofficial economy on welfare in any real-world institutional environment, and particularly in the rather distorted economies in transition, is ambiguous, even on the margin. Among other things, this is because the possibility of going underground often provides an effective check on the power of governments to over-regulate and overtax.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- bribery
- command economy
- informal economy
- rationing
- tax compliance and tax evasion
- transition and institutions

Bibliography

Alexeev, M. 1988. Market vs. rationing: the case of soviet housing. *Review of Economics and Statistics* 70, 414–20.

Alexeev, M., Conrad, R. and Hay, J. 2004. Nalogooblozhenie i pravovaia reforma v perekhodnoi ekonomike: predvaritel'nyi analiz [Taxation and legal reform in an economy in transition: a preliminary analysis]. In *Pravovye reformy i ekonomicheskii rost* [Legal Reform and Economic Growth], ed. E. Berglof and S. Shishkin. Moscow: CEFIR.

Alexeev, M., Gaddy, C. and Leitzel, J. 1991. An economic analysis of the Ruble overhang. *Communist Economies and Economic Transformation* 3, 467–79.

Alexeev, M., Janeba, E. and Osborne, S. 2004. Taxation and evasion in the presence of extortion by organized crime. *Journal of Comparative Economics* 32, 375–88.

Alexeev, M. and Pyle, W. 2003. A note on measuring the unofficial economy in the former Soviet republics. *Economics of Transition* 11, 153–75.

Alexeev, M. and Sabyr, L. 2004. Black markets and pre-reform crises in former socialist economies. *Economic Systems* 28, 1–12.

Boycko, M. 1992. When higher incomes reduce welfare: queues, labor supply, and macro equilibrium in socialist economies. *Quarterly Journal of Economics* 107, 907–20.

Campos, N. 2000. Never at noon: on the nature and causes of the transition shadow, Discussion paper. Prague: CERGE-EI.

Ericson, R. 1983. On an allocative role of the Soviet second economy. In *Marxism, Central Planning, and the Soviet Economy: Economic Essays in Honor of Alexander Erlich*, ed. P. Desai. Boston: MIT

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Press.

Friedman, E., Johnson, S., Kaufmann, D. and Zoido-Lobaton, P. 2000. Dodging the grabbing hand: the determinants of unofficial activity in 69 countries. *Journal of Public Economics* 76, 459–93.

Grossman, G. 1977. The ‘second economy’ of the USSR. *Problems of Communism* 26, 25–40.

Grossman, G. 1979. Notes on the illegal private economy and corruption. In *Soviet Economy in a Time of Change*, vol. 1, ed. Joint Economic Committee of Congress. Washington: US Government Printing Office.

Grossman, G. 1989. The second economy: boon or bane for the reform of the first economy. In *Economic Reforms in the Socialist World*, ed. S. Gomulka, Y. Ha and C. Kim. Armonk, NY: M.E. Sharpe.

Grossman, G. 1991. Wealth estimates based on the Berkeley–Duke émigré questionnaire: a statistical compilation. Berkeley–Duke Occasional Papers on the Second Economy in the USSR, No. 27.

Ivanova, A., Keen, M. and Klemm, A. 2005. The Russian flat tax reform. *Economic Policy* 20, 397–444.

Johnson, S., Kaufmann, D. and Shleifer, A. 1997. The unofficial economy in transition. *Brookings Papers on Economic Activity* 1997(2), 159–239.

Johnson, S., McMillan, J. and Woodruff, C. 2000. Why do firms hide? Bribes and unofficial activity after communism. *Journal of Public Economics* 76, 495–520.

Kaufmann, D. and Kaliberda, A. 1996. Integrating the unofficial economy into the dynamics of post-socialist economies: a framework for analysis and evidence. In *Economic Transition in Russia and the New States of Eurasia*, ed. B. Kaminski. London: M.E. Sharpe.

Kim, B. 2003. Informal economy activities of Soviet households: size and dynamics. *Journal of Comparative Economics* 31, 532–51.

Lau, L., Qian, Y. and Roland, G. 2000. Reform without losers: an interpretation of China's dual-track approach to reforms. *Journal of Political Economy* 108, 120–63.

Leitzel, J. 1995. *Russian Economic Reform*. London: Routledge.

Leitzel, J. 1998. Goods diversion and repressed inflation: notes on the political economy of price liberalization. *Public Choice* 94, 255–66.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Leitzel, J. 2003. *The Political Economy of Rule Evasion and Policy Reform*. London: Routledge.

Lipton, D. and Sachs, J. 1990. Creating market economy in eastern Europe: the case of Poland. *Brookings Papers on Economic Activity* 1990(1), 75–133.

Murphy, K., Shleifer, A. and Vishny, R. 1992. The transition to a market economy: pitfalls of partial reform. *Quarterly Journal of Economics* 107, 889–906.

Ofer, G. and Vinokur, A. 1992. *The Soviet Household Under the Old Regime: Economic Conditions and Behavior in the 1970s*. Cambridge, MA: Cambridge University Press.

Osband, K. 1992. Economic crisis in a shortage economy. *Journal of Political Economy* 100, 673–90.

Polterovich, V. 1993. Rationing, queues, and black markets. *Econometrica* 61, 1–28.

Roland, G. and Verdier, T. 2003. Law enforcement and transition. *European Economic Review* 47, 669–85.

Schneider, F. 2005. Shadow economies around the world: what do we really know? *European Journal of Political Economy* 21, 598–642.

Schneider, F. and Enste, D. 2000. Shadow economies: size, causes, and consequences. *Journal of Economic Literature* 38, 77–114.

Sinelnikov-Mourylev, S., Batkibekov, S., Kadochnikov, P. and Nekipelov, D. 2003. Otsenka rezul'tatov reformy podokhodnogo naloga v Rossiiskoi Federatsii [Evaluation of personal income tax reform in the Russian Federation]. Working Paper No. 52. Moscow: Institute for Economy in Transition (IET).

Stahl, D. and Alexeev, M. 1985. The influence of black markets on queue-rationed CPE. *Journal of Economic Theory* 35, 234–50.

How to cite this article

Alexeev, Michael. "second economy (unofficial economy)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000476> doi:10.1057/9780230226203.1502

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

seemingly unrelated regressions

Hyungsik Roger Moon and Benoit Perron
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article considers the seemingly unrelated regression (SUR) model first analysed by Zellner (1962). It describes estimators used in the basic model as well as recent extensions.

Keywords

bootstrap; feasible generalized least squares estimator; quasi-maximum likelihood estimator; generalized least squares estimator; heteroskedasticity; heteroskedasticity and autocorrelation; minimum distance estimator; ordinary least squares estimator; seemingly unrelated regressions; vector autoregressions; Zellner, A.

Article

A seemingly unrelated regression (SUR) system comprises several individual relationships that are linked by the fact that their disturbances are correlated. Such models have found many applications. For example, demand functions can be estimated for different households (or household types) for a given commodity. The correlation among the equation disturbances could come from several sources such as correlated shocks to household income. Alternatively, one could model the demand of a household for different commodities, but adding up constraints leads to restrictions on the parameters of different equations in this case. On the other hand, equations explaining some phenomenon in different cities, states, countries, firms or industries provide a natural application as these various entities are likely to be subject to spillovers from economy-wide or worldwide shocks. There are two main motivations for use of SUR. The first one is to gain efficiency in estimation by combining information on different equations. The second motivation is to impose and/or test restrictions that involve parameters in different equations. Zellner (1962) provided the seminal work in this area, and a thorough treatment is available in the book by Srivastava and Giles (1987). A recent survey can be found in Fiebig (2001). This article selectively overviews the SUR model, some of the estimators used in such systems and their properties, and several extensions of the basic SUR model. We adopt a classical perspective, although much Bayesian analysis has been done with this model (including Zellner's contributions).

Basic linear SUR model

Suppose that y_{it} is a dependent variable, $x_{it} = (1, x_{it,1}, x_{it,2}, \ldots, x_{it,K_i-1})'$ is a K_i-vector of explanatory variables for observational unit i, and u_{it} is an unobservable error term, where the double index it denotes the tth observation of the ith equation in the system. Often t denotes time and we will refer to this as the time dimension, but in some applications t could have other interpretations, for example as a location in space. A classical linear SUR model is a system of linear regression equations,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$y_{1t} = \beta_1' x_{1t} + u_{1t}$$
$$\vdots$$
$$y_{Nt} = \beta_N' x_{Nt} + u_{Nt}$$

where i=1,…,N, and t=1,…,T. Denote $L = K_1 + \cdots + K_N$. Further simplification in notation can be accomplished by stacking the observations either in the t dimension or for each i. For example, if we stack for each observation t, let $Y_t = [y_{1t}, \ldots, y_{Nt}]'$, $\tilde{X}_t = \mathrm{diag}(x_{1t}, x_{2t}, \ldots, x_{Nt})$, a block-diagonal matrix with x_{1t},…,x_{Nt} on its diagonal, $U_t = [u_{1t}, \ldots u_{Nt}]'$, and $\beta = [\beta_1', \ldots, \beta_N']'$. Then,

$$Y_t = \tilde{X}_t' \beta + U_t. \quad (1)$$

Another way to present the SUR model is to write it in a form of a multivariate regression with parameter restrictions. For this, define $X_t = [x_{1t}', x_{2t}', \ldots, x_{Nt}']'$ and $A(\beta) = \mathrm{diag}(\beta_1, \ldots, \beta_N)$ to be a (L×N) block diagonal coefficient matrix. Then, the SUR model in (1) can be rewritten as

$$Y_t = A(\beta)' X_t + U_t, \quad (2)$$

and the coefficient $A(\beta)$ satisfies

$$\mathrm{vec}(A(\beta)) = G\beta, \quad (3)$$

for some (NL×L) full rank matrix G. In the special case where $K_1 = \cdots = K_N = K$, we have $G = \mathrm{diag}(i_1, \ldots, i_N) \otimes I_K$, where i_j denotes the jth column of the N×N identity matrix I_N.

Assumption

In the classical linear SUR model, we assume that for each i=1,…,N, $x_i = [x_{i1}, \ldots, x_{iT}]'$ is of full rank K_i, and that conditional on all the regressors $X' = [X_1, \ldots, X_T]$, the errors U_t are i.i.d. over time with mean zero and homoskedastic variance $\Sigma = E(U_t U_t' | X)$. Furthermore, we assume that Σ is positive definite and denote by σ_{ij} the (i,j)th element of Σ: that is, $\sigma_{ij} = E(u_{it} u_{jt} | X)$.

Under this assumption, the covariance matrix of the entire vector of disturbances $U' = [U_1, \ldots, U_T]$ is given by $E[\mathrm{vec}(U)(\mathrm{vec}(U))'] = \Sigma \otimes I_T$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Estimation of β

In this section we summarize four estimators of β that have been widely used in applications of the classical linear SUR. Other estimators (such as Bayes, empirical Bayes or shrinkage estimators) have also been proposed. Interested readers should refer to Srivastava and Giles (1987) and Fiebig (2001).

1. Ordinary least squares (OLS) estimator. The first estimator of β is the ordinary least squares (OLS) estimator of Y_t on regressor $\tilde{X}_t$,

$$\hat{\beta}_{OLS} = \left(\sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t'\right)^{-1} \sum_{t=1}^{T} \tilde{X}_t' Y_t.$$

This is just the vector that stacks the equation-by-equation OLS estimators, $\hat{\beta}_{OLS} = (\hat{\beta}_{1,OLS}', \ldots, \hat{\beta}_{N,OLS}')'$, where $\hat{\beta}_{i,OLS} = (\Sigma_{t=1}^{T} x_{it} x_{it}')^{-1} \Sigma_{t=1}^{T} x_{it} y_{it}$.

2. Generalized least squares (GLS) and feasible GLS (FGLS) estimator. When the system covariance matrix Σ is known, the GLS estimator of β is

$$\hat{\beta}_{GLS} = \left(\sum_{t=1}^{T} \tilde{X}_t \Sigma^{-1} \tilde{X}_t'\right)^{-1} \sum_{t=1}^{T} \tilde{X}_t \Sigma^{-1} Y_t.$$

When the covariance matrix Σ is unknown, a feasible GLS (FGLS) estimator is defined by replacing the unknown Σ with a consistent estimate. A widely used estimator of Σ is

$$\hat{\Sigma} = \left(\hat{\sigma}_{ij}\right),$$

where $\hat{\sigma}_{ij} = \frac{1}{T} \Sigma_{t=1}^{T} \hat{e}_{it} \hat{e}_{jt}$ and $\hat{e}_{kt}$ is the OLS residuals of the *k*th equation: that is, $\hat{e}_{kt} = y_{kt} - \hat{\beta}_{k,OLS}' x_{kt}$, *k*=*i*, *j*. Then

$$\hat{\beta}_{FGLS} = \left(\sum_{t=1}^{T} \tilde{X}_t \hat{\Sigma}^{-1} \tilde{X}_t'\right)^{-1} \sum_{t=1}^{T} \tilde{X}_t \hat{\Sigma}^{-1} Y_t.$$

The FGLS estimator is a two-step estimator where OLS is used in the first step to obtain residuals $\hat{e}_{kt}$ and an estimator of Σ. The second step computes $\hat{\beta}_{FGLS}$ based on the estimated Σ in the first step. This estimator is sometimes referred to as the restricted estimator as opposed to the unrestricted estimator proposed by Zellner that uses the residuals from an OLS regression of (2) without imposing the coefficient restrictions (3), that is, from regressing each regressand on all distinct regressors in the system.

3. *Gaussian quasi-maximum likelihood estimator (QMLE)*. The Gaussian log-likelihood function is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$L(\beta, \Sigma) = \text{const} + \frac{T}{2}\det\Sigma - \frac{1}{2}\sum_{t=1}^{T}\left(Y_t - \tilde{X}_t'\beta\right)'\Sigma^{-1}\left(Y_t - \tilde{X}_t'\beta\right),$$

or equivalently,

$$L(\beta, \Sigma) = \text{const} + \frac{T}{2}\det\Sigma - \frac{1}{2}\sum_{t=1}^{T}\left(Y_t - A(\beta)'X_t\right)'\Sigma^{-1}\left(Y_t - A(\beta)'X_t\right),$$

where $A(\beta)$ denotes the coefficient A in (2) with the linear restriction of (3), and the QMLE $(\hat{\beta}_{QMLE}\ \hat{\Sigma}_{QMLE})$ maximizes $L(\beta, \Sigma)$. When the vector U_t has a normal distribution, this estimator is the maximum likelihood estimator.

4. Minimum distance (MD) estimator. The idea of the MD estimator is to obtain an estimator of the unrestricted coefficient A in (2), $\hat{A}$, and then, obtain an estimator of β by minimizing the distance between $\hat{A}$ and β in (3). For this, assume that $T > L$ and that the whole regressor matrix X has full rank L. When $\hat{A}$ is the OLS estimator of $A(\beta)$, that is $\hat{A} = (\Sigma_{t=1}^{T} X_t X_t')^{-1}\Sigma_{t=1}^{T} X_t Y_t'$, the optimal MD estimator $\hat{\beta}_{MD}$ minimizes the optimal MD objective function

$$Q_{MD}(\beta) = \left(\text{vec}(\hat{A}) - G\beta\right)'\left(\hat{\Sigma}^{-1} \otimes \sum_{t=1}^{T} X_t X_t'\right)\left(\text{vec}(\hat{A}) - G\beta\right).$$

In this case, we have

$$\hat{\beta}_{MD} = \left(G'\left(\hat{\Sigma}^{-1} \otimes \sum_{t=1}^{T} X_t X_t'\right)G\right)^{-1}\left(G'\left(\hat{\Sigma}^{-1} \otimes \sum_{t=1}^{T} X_t X_t'\right)\text{vec}(\hat{A})\right).$$

Relationship between the estimators

Some of the above estimators are tightly linked. For example, *if we use the same consistent estimator* $\hat{\Sigma}$, the FGLS and the MD estimators above are identical: that is, $\hat{\beta}_{FGLS} = \hat{\beta}_{MD}$. Also, if we use the QMLE estimator of Σ, $\hat{\Sigma}_{QMLE}$ in place of $\hat{\Sigma}$, $\hat{\beta}_{QMLE}$ is identical to $\hat{\beta}_{FGLS}$ (and to $\hat{\beta}_{MD}$). By the Gauss–Markov theorem, the GLS estimator $\hat{\beta}_{GLS}$ is more efficient than the OLS estimator $\hat{\beta}_{OLS}$ when the system errors are correlated across equations. However, this efficiency gain disappears in some special cases described in Kruskal's theorem (Kruskal, 1968). A well-known special case of this theorem is when the regressors in each equation are the same. For other cases, readers can refer to Greene (2003, ch. 14) and Davidson and MacKinnon (1993, pp. 294–5). The efficiency gain relative to OLS tends to be larger when the correlation across equations is larger and when the correlation among regressors in different equations is smaller.

Note also that efficient estimators propagate misspecification and inconsistencies across equations. For example, if any equation is misspecified (for example some relevant variable has been omitted), then the entire vector β will be inconsistently

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
estimated by the efficient methods. In this sense, equation-by-equation OLS provides some degree of robustness since it is not affected by misspecification in other equations in the system.

Distribution of the estimators

In the literature on the classical linear SUR, the FGLS estimator $\hat{\beta}_{FGLS}$ is often called the SUR estimator (SURE). The usual asymptotic analysis of the SURE is carried out when the dimension of index *t*, *T*, increases to infinity with the dimension of index *i*, *N*, kept fixed. For asymptotic theories for large *N*, *T*, one can refer to Phillips and Moon (1999). Under regularity conditions, the asymptotic distributions as $T \to \infty$ of the aforementioned estimators are

$$\sqrt{T}\left(\hat{\beta}_{OLS} - \beta\right) \Rightarrow N\left(0, \left[E\left(\tilde{X}_t \tilde{X}_t'\right)\right]^{-1} E\left(\tilde{X}_t \Sigma \tilde{X}_t'\right)\left[E\left(\tilde{X}_t \tilde{X}_t'\right)\right]^{-1}\right)$$

and

$$\sqrt{T}\left(\hat{\beta}_{GLS} - \beta\right), \quad \sqrt{T}\left(\hat{\beta}_{FGLS} - \beta\right), \sqrt{T}\left(\hat{\beta}_{MD} - \beta\right) \Rightarrow N\left(0, \left[E\left(\tilde{X}_t \Sigma^{-1} \tilde{X}_t'\right)\right]^{-1}\right) \qquad \equiv N\left(\left(G'\left(\Sigma^{-1} \otimes E\left(X_t X_t'\right)\right) G\right)^{-1}\right).$$

It is straightforward to show that the SUR estimator using the information in the system is more efficient (has a smaller variance) than the estimator of the individual equations. By using the above distributional results, it is straightforward to construct statistics to test general nonlinear hypotheses.

Finite sample properties of SURE have been studied extensively either analytically in some restrictive cases (Zellner, 1963; 1972; Kakwani, 1967), by asymptotic expansions (Phillips, 1977; Srivastava and Maekawa, 1995) or by simulation (Kmenta and Gilbert, 1968). Most work has focused on the two-equation case. The above approximations appear to be good descriptions of the finite-sample behaviour of the estimators analysed when the number of observations, *T*, is large relative to the number of equations, *N*. In particular, efficient methods provide an efficiency gain in cases where the correlation among disturbances across equations is high and when correlation among regressors across equations is low. Non-normality of disturbances has also been found to deteriorate the quality of the above approximations. Bootstrap methods have also been proposed to remedy these documented departures from normality and improve the size of tests.

Extensions

In this section we discuss several extensions of the classical linear SUR model where the assumption on the error terms is no longer satisfied.

Autocorrelation and heteroskedasticity

As in standard univariate models, non-spherical disturbances can be accommodated by either modelling the residuals or computing robust covariance matrices. In addition to standard dynamic effects, serial correlation can arise in this environment due to the presence of individual effects (see Baltagi, 1980). One could define the equivalent of White (in the case of heteroskedasticity) or HAC (in the case of serial correlation) standard errors to conduct inference with the OLS estimator as in the single-equation framework.

For efficiency in estimation some parametric assumption on the disturbance process is often imposed (see Greene, 2003). For example, in the case of heteroskedasticity, Hodgson, Linton, and Vorkink (2002) propose an adaptive estimator that is efficient under the assumption that the errors follow an elliptical symmetric distribution that includes the normal as a special case. An

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

intermediate approach is to use a restricted (or parametric) covariance matrix to try to capture some efficiency gains in estimation, and then use a nonparametric heteroskedasticity and autocorrelation (HAC) consistent estimator of the covariance matrix to do inference. This two-tier approach (dubbed quasi-FGLS) has been suggested by Creel and Farell (1996).

Endogenous regressors

When the regressor X_t in the SUR model is correlated with the error term U_t, one needs instrumental variables (IVs), say, $Z_t = [z_{1t}', \ldots, z_{Nt}']'$ to estimate β. We suppose that the IVs satisfy the usual rank condition. The generalized method of moments (GMM) estimator (or the IV estimator), then, utilizes the moment condition

$$E\left[\mathrm{vec}\left(Z_t U_t'\right)\right] = 0.$$

The optimal GMM estimator $\hat{\beta}_{GMM}$ is derived by minimizing the GMM objective function with the optimal choice of weighting matrix given by $\left(\hat{\Sigma} \otimes \left(\Sigma_{t=1}^{T} Z_t Z_t'\right)\right)^{-1}$,

$$Q_{GMM}(\beta) = \left[\sum_{t=1}^{T} \mathrm{vec}\left\{Z_t\left(Y_t - A(\beta)' X_t\right)'\right\}\right]' \left(\hat{\Sigma} \otimes \sum_{t=1}^{T} Z_t Z_t'\right)^{-1} \times \left[\sum_{t=1}^{T} \mathrm{vec}\left\{Z_t\left(Y_t - A(\beta)' X_t\right)'\right\}\right].$$

Then, we have

$$\hat{\beta}_{GMM} = \left\{G'\left(\hat{\Sigma} \otimes \left(\sum_{t=1}^{T} X_t Z_t'\right)\left(\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\sum_{t=1}^{T} Z_t X_t'\right)\right)^{-1} G\right\}^{-1} \times G'\left\{\hat{\Sigma}^{-1} \otimes \left(\sum_{t=1}^{T} X_t Z_t'\right)\left(\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\sum_{t=1}^{T} Z_t X_t'\right)\right\}\mathrm{vec}\left(\hat{A}_{2SLS}\right),$$

where

$$\hat{A}_{2SLS} = \left\{\left(\sum_{t=1}^{T} X_t Z_t'\right)\left(\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\sum_{t=1}^{T} Z_t X_t'\right)\right\}^{-1} \times \left(\sum_{t=1}^{T} X_t Z_t'\right)\left(\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\sum_{t=1}^{T} Z_t Y_t'\right)$$

is the two-stage least squares estimator of $A(\beta)$. When X_t is exogenous, so that $X_t = Z_t$, the GMM objective function $Q_{GMM}(\beta)$ and minimum distance objective function $Q_{MD}(\beta)$ are identical, and in this case $\hat{\beta}_{GMM} = \hat{\beta}_{MD}$.

Vector autoregressions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

When the index t in the SUR model denotes time and the regressors x_{it} include lagged dependent variables, the classical linear SUR model becomes a vector autoregression model (VAR) with exclusion restrictions. In this case, the regressors X are no longer strictly exogenous, and the assumption in the previous section is violated. A special case is when the order of the lagged dependent variables is one. In this case, for $\{y_{it}\}_t$ to be stationary, it is necessary that the absolute value of the coefficient of y_{it-1} is less than one. If the coefficient of y_{it-1} is one, $\{y_{it}\}_t$ is non-stationary. Non-stationary SUR VAR models have been used in developing tests for unit roots and cointegration in panels with cross-sectional dependence: see for example Chang (2004), Groen and Kleibergen (2003) and Larsson, Lyhagen, and Lothgren (2001).

Seemingly unrelated cointegration regressions

When the non-constant regressors in X_t are integrated non-stationary variables but the errors in U_t are stationary, we call model (1) (or equivalently (2)) a seemingly unrelated cointegration regression model; see Park and Ogaki (1991), Moon (1999), Mark, Ogaki and Sul (2005), and Moon and Perron (2004). These papers showed that for efficient estimation of β, an estimator of the long-run variance of U_t, not of the spontaneous covariance Σ as in the previous section, should be used in FGLS. In addition, some modification of the regression is necessary when the integrated regressors and the stationary errors are correlated. Empirical applications in the main references include tests for purchasing power parity, the relation between national saving and investment, and tests of the forward rate unbiasedness hypothesis.

Nonlinear SUR (NSUR)

An NSUR model assumes that the conditional mean of y_{it} given x_{it} is nonlinear, say $h_i(\beta, x_{it})$, that is, $y_{it} = h_i(\beta, x_{it}) + u_{it}$. Defining $H(\beta, X_t) = (h_1(\beta, x_{1t}), \ldots, h_N(\beta, x_{Nt}))$, we write the NSUR model in a multivariate nonlinear regression form,

$$Y_t = H(\beta, X_t) + U_t.$$

In this case, we may estimate β using (quasi) MLE assuming that Y_t are Gaussian conditioned on X_t or GMM utilizing the moment condition that $E[g(X_t)U_t'] = 0$ for any measurable transformation g of X_t.

See Also

- Bayesian econometrics
- bootstrap
- cointegration
- generalized method of moments estimation
- heteroskedasticity and autocorrelation corrections
- linear models
- serial correlation and serial dependence
- spatial econometrics
- statistical inference
- vector autoregressions

Bibliography

Baltagi, B. 1980. On seemingly unrelated regressions with error components. *Econometrica* 48, 1547–52.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Chang, Y. 2004. Bootstrap unit root tests in panels with cross-sectional dependency. *Journal of Econometrics* 120, 263–93.

Creel, M. and Farell, M. 1996. SUR estimation of multiple time-series models with heteroskedasticity and serial correlation of unknown form. *Economic Letters* 53, 239–45.

Davidson, R. and MacKinnon, J. 1993. *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

Fiebig, D.G. 2001. Seemingly unrelated regression. In *A Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Greene, W. 2003. *Econometric Analysis*, 5th edn. Englewood Cliffs, NJ: Prentice-Hall.

Groen, J. and Kleibergen, F. 2003. Likelihood-based cointegration analysis in panels of vector error correction models. *Journal of Business and Economic Statistics* 21, 295–318.

Hodgson, D., Linton, O. and Vorkink, K. 2002. Testing the capital asset pricing model efficiently under elliptical symmetry: a semiparametric approach. *Journal of Applied Econometrics* 17, 619–39.

Larsson, R., Lyhagen, J. and Lothgren, M. 2001. Likelihood-based cointegration tests in heterogeneous panels. *Econometrics Journal* 4, 109–42.

Kakwani, N.C. 1967. The unbiasedness of Zellner's seemingly unrelated regression equations estimator. *Journal of the American Statistical Association* 62, 141–2.

Kmenta, J. and Gilbert, R.F. 1968. Small sample properties of alternative estimators for seemingly unrelated regressions. *Journal of the American Statistical Association* 63, 1180–200.

Kruskal, W. 1968. When are Gauss–Markov and least squares estimators identical? *Annals of Mathematical Statistics* 39, 70–5.

Mark, N., Ogaki, M. and Sul, D. 2005. Dynamic seemingly unrelated cointegrating regressions. *Review of Economic Studies* 72, 797–820.

Moon, H.R. 1999. A note on fully-modified estimation of seemingly unrelated regression models with integrated regressors. *Economics Letters* 65, 25–31.

Moon, H.R. and Perron, P. 2004. Efficient estimation of SUR cointegration regression model and testing for purchasing power parity. *Econometric Reviews* 23, 293–323.

Park, J. and Ogaki, M. 1991. Seemingly unrelated canonical cointegrating regressions. Working Paper No. 280, University of Rochester.

Phillips, P.C.B. 1977. Finite sample distribution of Zellner's SURE. *Journal of Econometrics* 6, 147–64.

Phillips, P.C.B. and Moon, H.R. 1999. Linear regression limit theory for nonstationary panel data. *Econometrica* 67, 1057–111.

Srivastava, V.K. and Giles, D.E.A. 1987. *Seemingly Unrelated Regression Equations Models*. New York: Dekker.

Srivastava, V.K. and Maekawa, K. 1995. Efficiency properties of feasible generalized least squares estimators in SURE models under non-normal disturbances. *Journal of Econometrics* 66, 99–121.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regression equations and tests of aggregation bias. *Journal of the American Statistical Association* 57, 500–9.

Zellner, A. 1963. Estimators for seemingly unrelated regression equations: some finite sample results. *Journal of the American Statistical Association* 58, 977–92.

Zellner, A. 1972. Corrigenda. *Journal of the American Statistical Association* 67, 255.

How to cite this article

Moon, Hyungsik Roger and Benoit Perron. "seemingly unrelated regressions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000517> doi:10.1057/9780230226203.1503

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

selection bias and self-selection

James J. Heckman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This entry defines the problems of selection bias and presents the conditions required to solve them. It gives examples of common sampling frames, presents economic selection mechanisms, and discusses the assumptions required to use selected samples to determine features of the population distribution.
The analytical framework developed to understand selection bias problems is also fruitful in understanding the economics of self-selection. The prototypical model of choice theoretic self-selection is the Roy model, in which agents choose among a variety of discrete ‘occupational’ opportunities. The Roy model is presented and its fruitful extension to a variety of settings is demonstrated.

Keywords

censored regression model; choice-based sampling; general stratified sampling; length-biased sampling; random sampling; Roy model; selection bias and self-selection; size-biased sampling

Article

The problem of selection bias in economic and social statistics arises when a rule other than simple random sampling is used to sample the underlying population that is the object of interest. The distorted representation of a true population as a consequence of a sampling rule is the essence of the selection problem. Distorting selection rules may be the outcome of decisions of sample survey statisticians, self-selection decisions by the agents being studied, or both.
A random sample of a population produces a description of the population distribution of characteristics that has many desirable properties. One attractive feature of a random sample generated by the *known rule* that all individuals are equally likely to be sampled is that it produces a description of the population distribution of characteristics that becomes increasingly accurate as sample size expands.
A sample selected by any rule not equivalent to random sampling produces a description of the population distribution of characteristics that does not accurately describe the true population distribution of characteristics no matter how big the sample size. Unless the rule by which the sample is selected is known or can be recovered from the data, the selected sample cannot be used to produce an accurate description of the underlying population. For certain sampling rules, even knowledge of the rule generating the sample does not suffice to recover the population distribution from the sampled distribution.
This entry defines the problem of selection bias and presents conditions required to solve the problem. Examples of various types of commonly encountered sampling frames are given and specific economic selection mechanisms are presented. Assumptions required to use selected samples to determine features of the population distribution are discussed.
The analytical framework developed to understand the inferential problems raised by selection bias is also fruitful in understanding the economics of self-selection. The prototypical choice theoretic model of self-selection is that of Roy (1951). In his model, agents choose among a variety of discrete ‘occupational’ opportunities. Agents can pursue only one ‘occupation’ at a time. While every person can, in principle, do the work in each ‘occupation’, at least at some level of competence, self-interest drives individuals to choose that ‘occupation’ which produces the highest income (utility) for them. As in the statistical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
selection bias problem, there is a latent population (of skills). Observed (utilized) skill distributions are the outcome of a selection rule by agents. The relationship between observed and latent skill distributions is of considerable interest and underlies recent work on worker hierarchies (see Willis and Rosen, 1979). The 'occupations' can be: (a) market work or non-market work (b) unemployed and searching or working at the offered wage (c) working in one province or working in another, or (d) any choice among a set of mutually exclusive opportunities.

Because the insights in the Roy model underlie much recent research, we present a brief exposition of it and demonstrate how it can be or has been fruitfully extended to a variety of settings. An important issue, closely linked to the problem of identifying population parameters from selected sample distributions, is the empirical content of economic models of self-selection and worker hierarchies. Are they artefacts of distributional assumptions for unobservable skills or are they genuine behavioural hypotheses?

1 A definition and some examples of selection bias

Any selection bias model can be described by the following set-up. Let **Y** be a vector of outcomes of interest and let **X** be a vector of 'control' or 'explanatory' variables. The population distribution of **(Y,•X)** is *F***(y,•x)**. To simplify the exposition we assume that the density is well defined and write it as *f***(y,•x)**.

Any sampling rule can be interpreted as producing a non-negative weighting function $\omega(\mathbf{y}, \mathbf{x})$ that alters the population density. Let $(\mathbf{Y}^*,\bullet\mathbf{X}^*)$ denote the sampled random variables. The density of the sampled data $g(\mathbf{y}^*,\bullet\mathbf{x}^*)$ may be written as

$$g(\mathbf{y}^*, \mathbf{x}^*) = \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*) / \int \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*) d\mathbf{y}^* d\mathbf{x}^* \tag{1.1}$$

where the denominator of the expression is introduced to make the density $g(\mathbf{y}^*,\bullet\mathbf{x}^*)$ integrate to one as is required for proper densities.

Alternatively, the weight may be defined as

$$\omega^*(\mathbf{y}^*, \mathbf{x}^*) = \frac{\omega(\mathbf{y}^*, \mathbf{x}^*)}{\int \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*) d\mathbf{y}^* d\mathbf{x}^*}$$

so that

$$g(\mathbf{y}^*, \mathbf{x}^*) = \omega^*(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*). \tag{1.2}$$

Sampling schemes for which $\omega(\mathbf{y}, \mathbf{x}) = 0$ for some values of **(Y,•X)** create special problems. For such schemes, not all values of **(Y,•X)** are sampled. Let indicator variable $i(\mathbf{x}, \mathbf{y}) = 0$ if a potential observation at values **y,•x** cannot be sampled and let $i(\mathbf{y}, \mathbf{x}) = 1$ otherwise. Let $\Delta = 1$ record the occurrence of the event 'a potential observation is sampled, i.e. the value of **y**, **x** is observed' and let $\Delta = 0$ if it is not. In the population, the proportion that is sampled is

$$\Pr(\Delta = 1) = \int i(\mathbf{y}, \mathbf{x}) f(\mathbf{y}, \mathbf{x}) d\mathbf{y} d\mathbf{x}.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(1.3)

while

$$\Pr(\Delta = 0) = 1 - \Pr(\Delta = 1).$$

For samples in which $\omega(\mathbf{y}, \mathbf{x}) = 0$ for a non-negligible proportion of the population $(\Pr(\Delta = 0) > 0)$ it is clarifying to consider two cases. A *truncated sample* is one for which $\Pr(\Delta = 1)$ is not known and cannot be consistently estimated. For such a sample, (1.1) is the density of all of the sampled **Y** and **X** values. A *censored sample* is one for which $\Pr(\Delta = 1)$ is known or can be consistently estimated. The sampling rule in this case is such that values of **y**,•**x** for which $\omega(\mathbf{y}, \mathbf{x}) = 0$ are not known but it is known whether or not $i(\mathbf{y}, \mathbf{x}) = 0$ for all values of **Y**,•**X**. In this case it is notationally convenient to define $(\mathbf{Y}^*, \mathbf{X}^*) = (\mathbf{0}, \mathbf{0})$ for values of **y**,•**x** such that $\omega(\mathbf{y}, \mathbf{x}) = i(\mathbf{y}, \mathbf{x}) = 0$. Such a definition is innocuous provided that in the population there is no point mass (concentration of probability mass) at (**0**,•**0**). (Any value other than (**0**,•**0**) can be selected provided that there is no point mass at that value.) Given $\Delta = 0$ the distribution of $\mathbf{Y}^*$,•$\mathbf{X}^*$ is

$$G(\mathbf{y}^*, \mathbf{x}^*) = 1 \quad \text{for} \quad \Delta = 0$$

at

$$\mathbf{Y}^* = \mathbf{0} \quad \text{and} \quad \mathbf{X}^* = \mathbf{0}.$$

The joint density of $\mathbf{Y}^*$,•$\mathbf{X}^*$, Δ for the case of a censored sample is obtained by combining (1.1) and (1.3). Thus

$$g(\mathbf{y}^*, \mathbf{x}^*, \delta) = \left[\frac{\omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*)}{\int \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*) d\mathbf{y}^* d\mathbf{x}^*}\right]^{\delta} \times \left[\int i(\mathbf{y}, \mathbf{x}) f(\mathbf{y}, \mathbf{x}) d\mathbf{y} d\mathbf{x}\right]^{\delta} \times [1]^{1-\delta} \left[\int (1 - i(\mathbf{y}, \mathbf{x})) f(\mathbf{y}, \mathbf{x}) d\mathbf{y} d\mathbf{x}\right]^{1-\delta}.$$

(1.4)

The first term on the right-hand side of (1.4) is the conditional density of $\mathbf{Y}^*$,•$\mathbf{X}^*$ given $\Delta = 1$. The second term is the probability that $\Delta = 1$. The third term is the conditional density of $\mathbf{Y}^*$,•$\mathbf{X}^*$ given $\Delta = 0$. This density assigns unit mass to $\mathbf{Y}^* = 0, \mathbf{X}^* = \mathbf{0}$ when $\Delta = 0$. The fourth term is the probability that $\Delta = 0$. Notice that in the case in which $\omega(\mathbf{y}, \mathbf{x}) > 0$ for all **y**,•**x**, $\Delta = 1$ and (1.4) is identical to (1.1).

In a random sample $\omega(\mathbf{y}^*, \mathbf{x}^*) = 1$ (and so $\omega^*(\mathbf{y}^*, \mathbf{x}^*) = 1$). In a selected sample, the sampling rule weights the data differently. Values of (**Y**,•**X**) are over-sampled or under-sampled relative to their occurrence in the population. In the case of truncated samples, the weight is zero for certain values of the outcome.

In many problems in economics, attention focuses on $f(\mathbf{y}|\mathbf{x})$, the conditional density of **Y** given $\mathbf{X} = \mathbf{x}$. In such problems knowledge of the population distribution of **X** is of no direct interest. If samples are selected solely on the **x** variables (‘selection

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
on the exogenous variables'), $\omega(\mathbf{y}, \mathbf{x}) = \omega(\mathbf{x})$ and there is no problem about using selected samples to make valid inference about the population conditional density. This is so because in the case of selection on the exogenous variables

$$g(\mathbf{y}^*, \mathbf{x}^*) = f(\mathbf{y}^*|\mathbf{x}^*)\frac{\omega(\mathbf{x}^*)f(\mathbf{x}^*)}{\int\omega(\mathbf{x}^*)f(\mathbf{x}^*)d\mathbf{x}}$$

and

$$g(\mathbf{x}^*)\frac{\omega(\mathbf{x}^*)f(\mathbf{x}^*)}{\int\omega(\mathbf{x}^*)f(\mathbf{x}^*)d\mathbf{x}^*}.$$

Thus

$$g(\mathbf{y}^*|\mathbf{x}^*) = \frac{g(\mathbf{y}^*, \mathbf{x}^*)}{g(\mathbf{x}^*)} = f(\mathbf{y}^*|\mathbf{x}^*).$$

For such problems, sample selection distorts inference only if selection occurs on **y** (or **y** and **x**). Sampling on both **y** and **x** is termed *general stratified sampling*.

From a sample of data, it is not possible to recover the true density $f(\mathbf{y},\bullet\mathbf{x})$ without knowledge of the weighting rule. On the other hand, if the weighting rule is known $(\omega(\mathbf{y}^*, \mathbf{x}^*))$ the density of the sampled data is known ($g(\mathbf{y}^*,\bullet\mathbf{x}^*)$), the support of $(\mathbf{y},\bullet\mathbf{x})$ is known and $\omega(\mathbf{y},\bullet\mathbf{x})$ is non-zero, then $f(\mathbf{y},\bullet\mathbf{x})$ can always be recovered because

$$\frac{g(\mathbf{y}^*, \mathbf{x}^*)}{\omega(\mathbf{y}^*, \mathbf{x}^*)} = \frac{f(\mathbf{y}^*, \mathbf{x}^*)}{\int\omega(\mathbf{y}^*, \mathbf{x}^*)f(\mathbf{y}^*, \mathbf{x}^*)d\mathbf{y}^*d\mathbf{x}^*} \quad (1.5)$$

and by hypothesis both the numerator and denominator of the left-hand side are known. From the requirement that $(\mathbf{y}^*,\bullet\mathbf{x}^*)$ has a well defined density

$$\int f(\mathbf{y}^*, \mathbf{x}^*)d\mathbf{y}^*d\mathbf{x}^* = 1.$$

Integrating the left-hand side of (1.5) it is possible to determine $\int\omega(\mathbf{y}^*, \mathbf{x}^*)f(\mathbf{y}^*, \mathbf{x}^*)d\mathbf{y}^*d\mathbf{x}^*$ and hence to use (1.5) to recover the population density of the data.

The requirements that (a) the support of $(\mathbf{y},\bullet\mathbf{x})$ is known and (b) $\omega(\mathbf{y}, \mathbf{x})$ is nonzero are not innocuous. In many important problems in economics requirement (b) is not satisfied: the sampling rule excludes observations for certain values of $\mathbf{y},\bullet\mathbf{x}$ and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
hence it is impossible without invoking further assumptions to determine the population distribution of (**Y**,•**X**) at those values. If neither the support nor the weight is known, it is impossible, without invoking strong assumptions, to determine whether the fact that data are missing at certain **y,•x** values is due to the sampling plan or that the population density has no support at those values. We now turn to some specific sampling plans of interest in economics.

Example 1

Data are collected on incomes of individuals whose income *Y* exceeds a certain value *c* (for cut-off value). The rule is to observe *Y* if $Y > c$. Thus $\omega(y) = 1$ if $y > c$ and $\omega(y) = 0$ if $y \leq c$. Because the weight is zero for some values of *y*, we know that knowledge of the sampling rule does *not* suffice to recover the population distribution. From a random sample of the entire population, the social scientist knows or can consistently estimate (a) the sample distribution of *Y* above *c* and (b) the proportion of the original random sample with income below *c* (*F*(*c*) where *F* is the distribution function of *Y*). The social scientist does not observe values of *Y* below *c*.

In this example, observed income is a *truncated random variable*. The point of truncation is *c*. The *sample* of observed income is said to be *censored*. If the proportion of the original random sample with income below *c* is not known and cannot be consistently estimated, the *sample* is *truncated*. In a truncated sample, nothing is known about the proportion of the underlying population that can appear in the sample. A sample is truncated only if $\omega(\mathbf{y}) = 0$ for some intervals of **y** (for **y** continuous) or if $\omega(\mathbf{y}) = 0$ at values of **y** at which there is finite probability mass. In a censored sample, the proportion of the underlying population that can appear in the sample is known, at least to an arbitrarily high degree of approximation, as sample size increases.

Let $Y^* = Y$ if $Y > c$. Define $Y^* = 0$ otherwise (the choice of the value for Y^* when *Y* is not observed is inessential and any value can be used in place of 0 provided that the true distribution places no mass at the selected value). Define an indicator variable $\Delta = 1$ if $Y > c$. $\Delta = 0$ otherwise. Then the distribution of Y^* is

$$G(y^* | Y > 0) = F(y^* | Y > c) = F(y^* | \delta = 1) = \frac{F(y^*)}{1 - F(c)}, \; y^* > c. \tag{1.6a}$$

$$G(y^* | Y^* > 0) = 1 \quad \text{for} \quad Y^* = 0 (\Delta = 0). \tag{1.6b}$$

Observe that (1.6a) is obtained from (1.1) by setting $\omega(y^*) = 1$ if $y > c$, and $\omega(y^*) = 0$ otherwise, and integrating up with respect to y^*. The distribution of Δ is

$$\text{pr}(\Delta) = [1 - F(c)]^{\delta} [F(c)]^{1-\delta}.$$

The joint distribution of (Y^*,• Δ) is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$F(y^*, \delta) = F(y^*|\delta)\Pr(\delta) = \left\{\frac{F(y^*)}{(1-F(c))}\right\}^{\delta}[1-F(c)]^{\delta}(1)^{1-\delta}[F(c)]^{1-\delta} = [F(y^*)]^{\delta}[F(c)]^{1-\delta}.$$

(1.7)

Note that (1.7) is obtained from (1.4) by setting $\omega(y) = 0,\ y < c,\ \omega(y) = 1$ otherwise, by setting $i(y) = \omega(y)$ and by integrating up with respect to y^*. For normally distributed Y, (1.7) is the 'Tobit' distribution.

The difference between the information in a truncated sample and the information in a censored sample is encapsulated in the contrast between (1.6a) and (1.7). Clearly there is more information in a censored sample than in a truncated sample because one can obtain (1.6a) from (1.7) (by conditioning on $\Delta = 1$) but not vice versa.

Inferences about the population distribution based on assuming that $F(y^*|Y > c)$ closely approximates $F(y)$ are potentially very misleading. A description of population income inequality based on a subsample of high income people may convey no information about the true population distribution.

Without further information about F and its support, it is not possible to recover F from $G(y^*)$ from either a censored or a truncated sample. Access to a censored sample enables the analyst to recover $F(y)$ for $Y > c$ but obviously does not provide any information on the shape of the true distribution for values of $Y \leq c$.

This problem is routinely 'solved' by assuming that F is of a known functional form. This solution strategy does not always work. If F is normal, then it can be recovered from a censored or truncated sample (Pearson, 1901). If F is Pareto, F cannot be recovered from either a truncated or a censored sample (see Flinn and Heckman, 1982). If F is real analytic (i.e. possesses derivatives of all order) and the support of Y is known, then F can be recovered (Heckman and Singer, 1985).

Example 2

Expand the discussion in the previous example to a linear regression setting. Let

$$Y = \boldsymbol{X\beta} + U$$

(1.8)

be the population earnings function where Y is earnings, $\mathbf{X}$ is a regressor vector assumed to be distributed independently of mean zero disturbance U. 'β' is a suitably dimensioned parameter vector. Conventional assumptions are invoked to ensure that ordinary least squares applied to a random sample of earnings data consistently estimates β.

Data are collected on incomes of persons for whom Y exceeds c. Again the weight depends solely on y, i.e. $\omega(y, \mathbf{x}) = 0,\ y \leq c,\ \omega(y, \mathbf{x}) = 1,\ y > c$. The social scientist knows or can consistently estimate (a) the sample distribution of Y above c (b) the sample distribution of the $\mathbf{X}$ for Y above c and (c) the proportion of the original random sample with income below c. The social scientist does not observe values of Y below c.

As before, let $Y^* = Y$ if $Y > c$. Define $Y^* = 0$ otherwise. $\Delta = 1$ if $Y > c$, $\Delta = 0$ otherwise. The probability of the event $\Delta = 1$ given $\mathbf{X} = \mathbf{x}$ is

$$\Pr(\Delta = 1|\mathbf{X} = \mathbf{x}) = \Pr(Y > c|\mathbf{X} = \mathbf{x}) = \Pr(Y > c - \mathbf{x}\boldsymbol{\beta}|\mathbf{X} = \mathbf{x}).$$

Invoking independence between U and $\mathbf{X}$ and letting F_u denote the distribution of U,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\Pr(\Delta = 1|\mathbf{X} = \mathbf{x}) = 1 - F_U(c - \mathbf{x}\boldsymbol{\beta})$$
(1.9a)

and

$$\Pr(\Delta = 0|\mathbf{X} = \mathbf{x}) = F_U(c - \mathbf{x}\boldsymbol{\beta}).$$
(1.9b)

The distribution of Y^* conditional on $\mathbf{X}$ is

$$G(y^*|Y > 0, \mathbf{X} = \mathbf{x}) = F(y^*|\mathbf{X} = \mathbf{x}, Y > c) = F(y^*|\mathbf{X} = \mathbf{x}, \Delta = 1) = \frac{F_U(y^* - \mathbf{x}\boldsymbol{\beta})}{1 - F_U(c - \mathbf{x}\boldsymbol{\beta})}, \quad y^* > c.$$
(1.10a)

$$G(y^*|Y \le 0) = 1 \quad \text{for} \quad Y^* = 0(\Delta = 0).$$
(1.10b)

The joint distribution of $(Y^*, \bullet\, \Delta)$ given $\mathbf{X} = \mathbf{x}$ is

$$F(y^*, \delta|\mathbf{X} = \mathbf{x}) = F(y^*|\delta, \mathbf{x})\Pr(\delta|\mathbf{x}). = \left\{F_U(y^* - \mathbf{x}\boldsymbol{\beta})\right\}^{\delta}\{F_U(c - \mathbf{x}\boldsymbol{\beta})\}^{1-\delta}.$$
(1.11)

In particular,

$$E(Y^*|\mathbf{X} = \mathbf{x}, \Delta = 1) = \mathbf{x}\boldsymbol{\beta} + E(U|\mathbf{X} = \mathbf{x}, \delta = 1) = \mathbf{x}\boldsymbol{\beta} + \int_{c-\mathbf{x}\boldsymbol{\beta}}^{\infty} \frac{z\,dF_U(z)}{(1 - F_U(c - \mathbf{x}\boldsymbol{\beta}))}$$
(1.12)

where z is a dummy variable of integration. In contrast, the population mean regression function is

$$E(Y|\mathbf{X} = \mathbf{x}) = \mathbf{x}\boldsymbol{\beta}.$$
(1.13)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The contrast between (1.12) and (1.13) is illuminating. Many behavioural theories in social science produce empirical counterparts of (1.8) with population conditional expectations like (1.13). Such theories sometimes restrict the signs, permissible values and other relationships among the coefficients in β. When the theoretical model is estimated on a selected sample $(\Delta = 1)$ the true conditional expectation is (1.12) not (1.13). The conditional mean of *U* depends on **x**. In terms of conventional omitted variable analysis, $E(U|\mathbf{X} = \mathbf{x}, \Delta = 1)$ is omitted from the regression. Since this term is a function of **x** it is likely to be correlated with **x**. Least squares estimates of β obtained on selected samples which do not account for selection are biased and inconsistent.

To illustrate the nature of the bias, it is useful to draw on the work of Cain and Watts (1973). Suppose that *X* is a scalar random variable (e.g. education) and that its associated coefficient is positive $(\beta > 0)$. Under conventional assumptions about *U* (e.g. mean zero, independently and identically distributed and distributed independently of *X*), the population regression of *Y* on *X* is a straight line. The scatter about the regression line and the regression line are given in Figure 1. When $Y > c$ is imposed as a sample inclusion requirement, lower population values of *U* are excluded from the sample in a way that systematically depends on *x* $(Y > c \text{ or } U > c - x\beta)$. As *x* increases, the conditional mean of $U\,[E(U|X = x, \Delta = 1)]$ decreases. Regression estimates of β that do not correct for sample selection (that is, include $E(U|X = x, \Delta = 1)$ as a regressor) are downward biased because of the negative correlation between *x* and $E(U|X = x, \Delta = 1)$. See the flattened regression line for the selected sample in Figure 1.

Figure 1

Figure 1

In models with more than one regressor, no sharp result on the sign of the bias in the regression estimate that results from ignoring the selected nature of the sample is available except when the **X** variables are from certain distributions (e.g. normal, see Goldberger, 1983). None the less, the key result – that conventional least squares estimates of β obtained from selected samples are biased and inconsistent – remains true.

As in example 1, it is fruitful to distinguish between the case of a truncated sample and the case of a censored sample. In the truncated sample case, no information is available about the fraction of the population that would be allocated to the truncated sample $[\Pr(\Delta = 1)]$. In the censored sample case, this fraction is known or can be consistently estimated. In the censored sample case it is fruitful to distinguish two further cases: (*a*) the case in which **X** is not observed when $\Delta = 0$ and (*b*) the case in which it is. Case (*b*) is the one most fully developed in the literature (Heckman and MaCurdy, 1981).

Note that the conditional mean $E(U|\mathbf{X} = \mathbf{x}, \Delta = 1)$ is a function of $c - \mathbf{x}\boldsymbol{\beta}$ solely through $\Pr(\Delta = 1|\mathbf{x})$. Since $\Pr(\Delta = 1|\mathbf{x})$ is monotonic in $c - \mathbf{x}\boldsymbol{\beta}$ the conditional mean depends solely on $\Pr(\Delta = 1|\mathbf{x})$ and the parameters F_u i.e. since

$$F_u^{-1}(1 - \Pr(\Delta = 1|\mathbf{x})) = c - \mathbf{x}\boldsymbol{\beta}, E(U|\mathbf{X} = \mathbf{x}, \Delta = 1) = \int_{F_u^{-1}[1-\Pr(\Delta=1|\mathbf{x})]}^{\infty} \frac{z\,dF_u(z)}{\Pr(\Delta = 1|\mathbf{x})}.$$

This relationship demonstrates that the conditional mean is a function of the probability of selection. As the probability of selection goes to 1, the conditional mean goes to zero. For samples chosen so that the values of **x** are such that the observations are certain to be included in the sample, there is no problem in using ordinary least squares on selected samples to estimate β. Thus in Figure 1, ordinary least squares regressions fit on samples selected to have large **x** values closely approximate the true regression function and become arbitrarily close as **x** becomes large. The condition mean in (1.12) is a surrogate for $\Pr(\Delta = 1|\mathbf{x})$. As this probability goes to one, the problem of sample selection in regression analysis becomes negligibly small.

Heckman (1976) demonstrates that β and F_u are identified if *U* is normally distributed and standard conditions invoked in regression analysis are satisfied. Gallant and Nychka (1984) and Cosslett (1984) establish conditions for identification for non-normal *U*. In their analyses, F_u is consistently non-parametrically estimated.

Example 3

The next example considers *censored random variables*. This concept extends the notion of a truncated random variable by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
letting a more general rule than truncation on the outcome of interest generate the selected sample. Because the sample generating rule may be different from a simple truncation of the outcome being studied, the concept of a censored random variable in general requires at least two distinct random variables.

Let Y_1 be the outcome of interest. Let Y_2 be another random variable. Denote observed Y_1 by Y_1^*. If $Y_2 < c$, Y_1 is observed. Otherwise Y_1 is not observed and we can set $Y_1^* = 0$ or any other convenient value (assuming that Y_1 has no point mass at Y_1=0 or at the alternative convenient value). In terms of the weighting function ω, $\omega(y_1, y_2) = 0$ if $y_2 > c$, $\omega(y_1, y_2) = 1$ if $y_2 \leq c$. Selection rule $Y_2 < c$ does not necessarily restrict the range of Y_1. Thus Y_1^* is not in general a truncated random variable. Define $\Delta = 1$ if $Y_2 < c$; $\Delta = 0$ otherwise. If $F(y_1,\bullet y_2)$ is the population distribution of $(Y_1,\bullet Y_2)$, the distribution of Δ is

$$\Pr(\Delta = \delta) = [1 - F_2(c)]^{1-\delta}[F_2(c)]^{\delta}, \delta = 0, 1,$$

where F_2 is the marginal distribution of Y_2. The distribution of Y_1^* is

$$G(y_1^*) = F(y_1^* | \delta = 1) = \frac{F(y_1^*, c)}{F_2(c)}, \Delta = 1,$$
(1.14a)

$$G(y_1^* = 0) = 1, \Delta = 0.$$
(1.14b)

Note that (1.14a) is the distribution function corresponding to the density in (1.1) when $\omega(y_1, y_2) = 1$ if $y_2 \leq c$ and $\omega(y_1, y_2) = 0$ otherwise.

The joint distribution of (Y_1^*, Δ) is

$$G(y_1^*, \delta) = [F(y_1^*, c)]^{\delta}[1 - F_2(c)]^{1-\delta}.$$
(1.15)

This is the distribution function corresponding to density (1.4) for the special weighting rule of this example. In a censored sample, under general conditions it is possible to consistently estimate $\Pr(\Delta = \delta)$ and $G(y_1^*)$. In a truncated sample, only conditional distribution (1.14a) can be estimated. A degenerate version of this model has $Y_1 \equiv Y_2$. In that case, censored random variable Y_1 is also a truncated random variable. Note that a censored random variable may be defined for a truncated or censored sample.

Example 3 and variants of it have wide applicability in economics. Let Y_1 be the wage of a woman. Wages of women are observed only if women work. Let Y_2 be an index of a woman's propensity to work. In Gronau (1974) and Heckman (1974), Y_2

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
is postulated as the difference between reservation wages (the value of time at home determined from household preference functions) and potential market wages Y_1. Then if $Y_2 < 0$, the woman works. Otherwise, she does not. $Y_1^* = Y_1$ if $Y_2 < 0$ is the observed wage.

If Y_1 is the offered wage of an unemployed worker, and Y_2 is the difference between reservation wages (the return to searching) and offered market wages, $Y_1^* = Y_1$ if $Y_2 < 0$ is the accepted wage for an unemployed worker (see Flinn and Heckman, 1982). If Y_1 is the potential output of a firm and Y_2 is its profitability, $Y_1^* = Y_1$ if $Y_2 > 0$. If Y_1 is the potential income in occupation one and Y_2 is the potential income in occupation two, $Y_1^* = Y_1$ if $Y_1 - Y_2 < 0$ while $Y_2^* = Y_2$ if $Y_1 - Y_2 \geq 0$. We develop this example at length in section 2 where we consider explicit economic models of self-selection. There we discuss the identifiability of this model.

Example 4

This example builds on example 3 by introducing regressors. This produces the *censored regression model* (Heckman, 1976; 1979). In example 3 set

$$Y_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + U_1 \tag{1.16a}$$

$$Y_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + U_2 \tag{1.16b}$$

where $(\mathbf{X}_1, \mathbf{X}_2)$ are distributed independently of (U_1, U_2), a mean zero, finite variance random vector. Conventional assumptions are invoked to ensure that if Y_1 and Y_2 can be observed, least squares applied to a random sample of data on $(Y_1, Y_2, \mathbf{X}_1, \mathbf{X}_2)$ would consistently estimate β_1 and β_2. $Y_1^* = Y_1$ if $Y_2 < 0$. If $Y_2 < 0$, $\Delta = 1$. Then the regression function for the selected sample is

$$E(Y_1^*|\mathbf{X}_1 = \mathbf{x}_1, Y_2 < 0) = E(Y_1^*|\mathbf{X}_1 = \mathbf{x}_1, \Delta = 1) = \mathbf{X}_1\boldsymbol{\beta}_1 + E(U_1|\mathbf{X}_1 = \mathbf{x}_1, \Delta = 1) \tag{1.17}$$

and the regression function for the population is

$$E(Y_1|\mathbf{X}_1 = \mathbf{x}_1) = \mathbf{X}_1\boldsymbol{\beta}_1. \tag{1.18}$$

As in the regression analysis of truncated random variables, there is an illuminating contrast between the conditional

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
expectation for the selected sample (1.17) and the population regression function (1.18). The two functions differ by the conditional mean of $U_1[E(U_1|\mathbf{X}_1 = \mathbf{x}_1, \Delta = 1)]$. In the regression analysis of truncated random variables, ordinary least squares estimates of β (in equation (1.14)) are biased and inconsistent because the conditional mean is improperly omitted from the selected sample regression. The same analysis applies to the regression analysis of censored random variables. The conditional mean is a surrogate for the probability of selection $[\Pr(\Delta = 1|\mathbf{x}_2)]$. As $\Pr(\Delta = 1|\mathbf{x}_2)$ goes to one, the problem of sample selection bias becomes negligible. However, in the censored regression case, a new phenomenon appears. If there are variables in $\mathbf{X}_2$ not in $\mathbf{X}_1$, such variables may appear to be statistically important determinants of Y_1 when ordinary least squares is applied to data generated from censored samples.

As an example, suppose that survey statisticians use some extraneous (to $\mathbf{X}_1$) variables to determine sample enrolment. Such variables may appear to be important determinants of Y_1 when in fact they are not. They are important determinants of Y_1^*. In an analysis of self-selection, let Y_1 be the wage that a potential worker could earn were he to accept a market offer. Let Y_2 be the difference between the best non-market opportunity available to the potential worker and Y_1. If $Y_2 < 0$, the agent works. The conditional expectation of observed wages ($Y_1^* = Y$ if $Y_2 < 0$) given $\mathbf{x}_1$ and $\mathbf{x}_2$ will be a non-trivial function of $\mathbf{x}_2$. Thus variables determining non-market opportunities will determine Y_1^*, even though they do not determine Y_1. For example, the number of children less than six may appear to be significant determinants of Y_1 when inadequate account is taken of sample selection, even though the market does not place any value or penalty on small children in generating wage offers for potential workers.

Heckman (1976) develops the analysis of this model when $(U_1, \cdot U_2)$ is normally distributed. Gallant and Nychka (1984) and Cosslett (1984) demonstrate that under mild restrictions on $F(u_1, \cdot u_2)$, if there is one continuous valued variable in $\mathbf{X}_2$ not in $\mathbf{X}_1$ (so that there is no exact linear dependence between $\mathbf{X}_2$ and $\mathbf{X}_1$), $\beta_1, \cdot \beta_2$ and $F(u_1, \cdot u_2)$ can be consistently non-parametrically estimated. Heckman and MaCurdy (1986) develop this class of models at length.

Example 5

This example demonstrates how self-selection bias affects the interpretation placed on estimated consumer demand functions when there is self-selection. We postulate a population of consumers with a quasi-concave utility function $U(\mathbf{Z}, \cdot E)$ which depends on the consumption of goods and preference shock E which represents heterogeneity in preferences among consumers. The support of E is $\mathbf{E}$. For price vector $\mathbf{P}$ and endowment income M, the consumer's problem is to

$$\text{Max } U(\mathbf{Z}, E) \text{ subject to } \mathbf{P}'\mathbf{Z} \leq M.$$

In the population $\mathbf{P}$ and M are distributed independently of E. First-order conditions for this problem are

$$\frac{\partial U(\mathbf{Z}, E)}{\partial \mathbf{Z}} \leq \lambda \mathbf{P}, \tag{1.19}$$

where λ is the Lagrange multiplier associated with the budget constraint. Focusing on the demand for the first good, Z_1, none of it is purchased if at zero consumption of Z_1

$$\left.\frac{\partial U(\mathbf{Z}, E)}{\partial Z_1}\right|_{Z_1 = 0} \leq \lambda P_1. \tag{1.20}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

that is, marginal valuation is less than marginal cost in utility terms. Conventional interior solution demand functions for Z_1 are defined for a given **P**, M only for values of E such that

$$\left.\frac{\partial U(\mathbf{Z}, E)}{\partial Z_1}\right|_{Z_1=0} \geq \lambda P_1. \tag{1.21}$$

Let the set of E for which conventional interior solution consumer demand functions for Z_1 are defined be denoted by $\underline{\underline{E}}$. Then

$$\underline{\underline{E}} = \left\{E \text{ such that} \left.\frac{\partial U(\mathbf{Z}, E)}{\partial Z_1}\right|_{Z_1=0} \geq \lambda P_1 \text{ for given } \mathbf{P}, M\right\}.$$

Let $\Delta_1 = 0$ if the consumer does not purchase Z_1. Let $\Delta_1 = 1$ otherwise. If $F(\varepsilon)$ is the population distribution of E, the proportion purchasing none of good Z_1 given **P**,•M is

$$\Pr(\Delta_1 = 0|\mathbf{P}, M) = 1 - \int_{\underline{\underline{E}}} dF(\varepsilon).$$

Provided inequality (1.21) is satisfied, $\Delta_1 = 1$ and interior solution demand function

$$Z_1 = Z_1(\mathbf{P}, M, E) \tag{1.22}$$

is well defined and $Z_1 = Z_1^*$. When $\Delta_1 = 0$, observed $Z_1 = Z_1^* = 0$.
Equation (1.22) is the conventional object of interest in consumer theory. Partial derivatives of that function *holding E and the other arguments constant* have well defined economic interpretations. Suppose that some non-negligible proportion of the population buys none of good Z_1. Regression estimates of the parameters of (1.22) using Z_1^* approximate the conditional expectation

$$E(Z_1|\Delta_1 = 1, \mathbf{P}, M) = \int_{\underline{\underline{E}}} Z_1(\mathbf{P}, M, \varepsilon) dF(\varepsilon). \tag{1.23}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The derivatives of (1.23) are different from the derivatives of (1.22). In order to define these derivatives, it is helpful to define $I_{\underline{\underline{E}}}(E)$ as an indicator function for set $\underline{\underline{E}}$ which equals one if $E \in \underline{\underline{E}}$ and equals zero otherwise. When prices or income change, the set of values of *E* that satisfy inequality (1.21) changes. Let $\underline{\underline{E}} + \Delta\underline{\underline{E}}_{\mathbf{P}}$ be the set of *E* values that satisfy (1.21) when there is a finite price change $\Delta\mathbf{P}$. $I_{\underline{\underline{E}}+\Delta\underline{\underline{E}}_{\mathbf{P}}}(E)$ is an indicator function which equals one when $E \in \underline{\underline{E}} + \Delta\underline{\underline{E}}_{\mathbf{P}}$. Then the derivatives of (1.23) are, for the *j*th price

$$\frac{\partial E(Z_1|\Delta = 1, \mathbf{P}, M)}{\partial P_j} = \int_{\underline{\underline{E}}} \frac{\partial Z_1(\mathbf{P}, M, \varepsilon)}{\partial P_j} dF(\varepsilon) + \lim_{\Delta P_j \to 0} \int_{\underline{\underline{E}}} \frac{\left[\left(I_{\underline{\underline{E}}+\Delta\underline{\underline{E}}_{\mathbf{P}_j}}(\varepsilon) - I_{\underline{\underline{E}}}(\varepsilon)\right) Z(\mathbf{P}, M, \varepsilon)\right]}{\Delta P_j} dF(\varepsilon). \tag{1.24}$$

When the limit in the second term does not exist, the derivative does not exist. We assume for expositional convenience that the limit is well defined.

The first expression on the right-hand side of (1.24) is the average effect of price change on commodity demand. The second term on the right-hand side of (1.24) arises from the change in sample composition of *E* as the proportion of non-purchasers changes in response to price change. This term generates the selection bias.

Neither term is the same as the price derivative of (1.22) for an arbitrary value of $E = \varepsilon$ although the first term on the right-hand side of (1.24) approximates the price derivative of (1.22) for some value of $E = \varepsilon$.

A similar decomposition of the derivatives of the conditional demand function can be performed if it is defined solely for a sample of non-zero purchasers (see Heckman and MaCurdy, 1981; 1986).

Just as in the statistical sample selection bias problem, there is a population of interest. In this case, the population parameters of interest are the distribution of *E* and the parameters of $U(\mathbf{Z}, \cdot E)$. Those who buy Z_1 are a self-selected sample of the population. Estimates of population parameters estimated on self-selected samples are biased and inconsistent. There is a population distribution of $Z_1(\mathbf{P}, \cdot M, \cdot E)$ generated by the distribution of *E*. Observations of Z_1 are obtained only if $E \in \underline{\underline{E}}$ ($\omega(E) = 1$ if $E \in \underline{\underline{E}}$, $\omega(E) = 0$ otherwise). Alternatively one can express the inclusion criteria in terms of the latent population distribution of Z_1 induced by *E* (given **P** and *M*) *and write* $\omega(z_1) = 1$ if $z_1 > 0$, $\omega(z_1) = 0$ if $z_1 \leq 0$.

Heckman (1974) and Heckman and MaCurdy (1981) provide further discussion of this type of model which is widely used in applied economics and consider issues of identifiability for such models.

Example 6. Length biased sampling

Let *T* be the duration of an event such as a completed unemployment spell or a completed duration of a job with an employer. The population distribution of *T* is $F(t)$ with density $f(t)$. The sampling rule is such that *individuals* are sampled at random. Data are recorded on a completed spell *provided that at the time of the interview the individual is experiencing the event*. Such sampling rules are in wide use in many national surveys of employment and unemployment.

In order to have a sampled completed spell, a person must be in the state at the time of the interview. Let '0' be the date of the survey. Decompose any completed spell *T* into a component that occurs before the survey T_b and a component that occurs after the survey T_a. Then $T = T_a + T_b$. For a person to be sampled, $T_b > 0$. The density of *T* given $T_b = t_b$ is

$$f(t|t_b) = \frac{f(t)}{1 - F(t_b)}, \; t \geq t_b. \tag{1.25}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Suppose that the environment is stationary. The population entry rate into the state at each instant of time is *k*. From each vintage of entrants into the state distinguished by their distance from the survey date t_b, only $1 - F(t_b) = Pr(T > t_b)$ survive. Aggregating over all cohorts of entrants, the population proportion in the state at the date of the interview is *P* where

$$P = \int_0^{\infty} k(1 - F(t_b))\mathrm{d}t_b \tag{1.26}$$

which is assumed to exist. The density of T_b^*, sampled pre-survey duration, is

$$g(t_b^* | t_b^* > 0) = \frac{k(1 - F(t_b^*))}{P}. \tag{1.27}$$

The density of sampled completed durations is thus

$$g(t^*) = \int_0^{t^*} f(t^* | t_b^*) g(t_b^* | t_b^* > 0)\mathrm{d}t_b^* = k\frac{f(t^*)}{1 - F(t_b^*)}\frac{1 - F(t_b^*)}{P}\int_0^{t^*}\mathrm{d}t_b^* = k\frac{t^* f(t^*)}{P}.$$

Observe from (1.26) that by a standard integration by parts argument

$$P = k\int_0^{\infty}(1 - F(z))\mathrm{d}z = k\int_0^{\infty} z\mathrm{d}F(z) = kE(T).$$

Note that

$$g(t^*) = \frac{t^* f(t^*)}{E(T)}. \tag{1.28}$$

In this form (1.28) is equivalent to (1.1) with $\omega(t) = t$. Hence the term ‘length biased sampling’. Intuitively, longer spells are oversampled when the requirement is imposed that a spell be in progress at the time the survey is conducted $(T_b > 0)$. Suppose, instead, that individuals are randomly sampled and data are recorded on the *next* spell of the event (after the survey date). As long as successive spells are independent, such a sampling frame does not distort the sampled distribution because no requirement is imposed that the sampled spell be in progress at the date of the interview. It is important to notice that the source of the bias is the requirement that $T_b > 0$, not that only a fraction of the population experiences the event $(P < 1)$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The simple length weight $(\omega(t) = t)$ that produces (1.28) is an artefact of the stationarity assumption. Heckman and Singer (1985) consider the consequences of non-stationarity and un-observables when there is selection on the event that a person be in the state at the time of the interview. They also demonstrate the bias that results from estimating parametric models on samples generated by length biased sampling rules when inadequate account is taken of the sampling plan. Vardi (1983, 1985) and Gill and Wellner (1985) consider nonparametric identification and estimation of models with densities of the form (1.28).

It is unfortunate that the lessons of length biased sampling are not adequately appreciated in economics. Two widely cited studies by Clark and Summers (1979) and Hall (1982) use length biased data to prove, respectively, that unemployment and employment spells are 'surprisingly long'. Whether their findings are artefacts of sampling plans remains to be determined.

Example 7. Choice based sampling

Let D be a discrete valued random variable which assumes a finite number of values I. $D = i$, $i = 1, \ldots, I$ corresponds to the occurrence of state i. States are mutually exclusive. In the literature the states may be modes of transportation choice for commuters (Domencich and McFadden, 1975), occupations, migration destinations, financial solvency status of firms, schooling choices of students, etc. Interest centres on estimating a population choice model

$$\Pr(D = i|\mathbf{X} = \mathbf{x}),\ i = 1, \ldots, I. \tag{1.29}$$

The population density of $(D, \mathbf{X})$ is

$$f(d, \mathbf{x}) = \Pr(D = d|\mathbf{X} = \mathbf{x})\, h(\mathbf{x}) \tag{1.30}$$

where $h(\mathbf{x})$ is the density of the data.

In many problems, plentiful data are available on certain outcomes while data are scarce for other outcomes. For example, interviews about transportation preferences conducted at train stations tend to over-sample train riders and under-sample bus riders. Interviews about occupational choice preferences conducted at leading universities over-sample those who select professional occupations.

In choice based sampling, selection occurs solely on the D coordinate of $(D,\bullet\mathbf{X})$. In terms of (1.1) (extended to allow for discrete random variables), $\omega(d, \mathbf{X}) = \omega(d)$. Then sampled $(D^*,\bullet\mathbf{X}^*)$ has density

$$g(d^*, \mathbf{x}^*) = \frac{\omega(d^*)\, f(d^*, \mathbf{x}^*)}{\Sigma_{i=1}^{I} \int \omega(i)\, f(i, x^*)\, \mathrm{d}x^*} \tag{1.31}$$

Notice that the denominator can be simplified to

$$\sum_{i=1}^{I} \omega(i)\, f(i)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $f(d^*)$ is the marginal distribution of D^* so that

$$g(d^*, \mathbf{x}^*) = \frac{\omega(d^*) f(d^*, \mathbf{x}^*)}{\sum_{i=1}^{I} \omega(i) f(i)} \tag{1.32}$$

Also, integrating (1.31) with respect to **x** using (1.32) we obtain

$$g(d^*) = \frac{\omega(d^*) f(d^*)}{\sum_{i=1}^{I} \omega(i) f(i)} \tag{1.33}$$

which makes transparent how the sampling rule causes the sampled proportions to deviate from the population proportions. Note further that as a consequence of sampling only on *D*, the population conditional density

$$h(\mathbf{x}^* | d^*) = \frac{f(d^*, \mathbf{x}^*)}{f(d^*)} \tag{1.34}$$

can be recovered from the choice-based sample. The density of **x** in the sample is thus

$$g(\mathbf{x}^*) = \sum_{i=1}^{I} h(\mathbf{x}^* | i) g(i). \tag{1.35}$$

Then using (1.32)–(1.35) we reach

$$g(d^* | \mathbf{x}^*) = f(d^* | \mathbf{x}^*) \left\{ \left[\frac{\omega(d^*)}{\sum_{i=1}^{I} \omega(i) f(i)} \right] \left[\frac{1}{\sum_{i=1}^{I} f(i | \mathbf{x}^*) \frac{g(i)}{f(i)}} \right] \right\}. \tag{1.36}$$

The bias that results from using choice based samples to make inference about $f(d^* | \mathbf{x}^*)$ is a consequence of neglecting the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

terms in braces on the right-hand side of (1.36). Notice that if the data are generated by a random sampling rule, $\omega(d^*) = 1$, $g(d^*) = f(d^*)$ and the term in braces is one.
Manski and Lerman (1977), Manski and McFadden (1981) and Cosslett (1981) provide illuminating discussions of choice based sampling.

Example 8. Size biased sampling

Let N be the number of children in a family. $f(N)$ is the density of discrete random variable N. Suppose that family size is recorded only when at least one child is interviewed. Suppose further that each child has an independent and identical chance β of being interviewed. The probability of sampled family size of $N^* = n^*$ is

$$g(n^*) = \frac{\omega(n^*) f(n^*)}{E[\omega(N^*)]} \tag{1.37}$$

where $\omega(n^*) = 1 - (1-\beta)^{n^*}$ (the probability that at least one child from a family of size n^* will be sampled) and

$$E[\omega(N^*)] = \sum_{n^*} \left(1 - (1-\beta)^{n^*}\right) f(n^*)$$

is the probability of observing a family. In a large population $\beta \to 0$ with increasing population size. Using l'Hospital's rule, and assuming that passage to the limit under the summation sign is valid

$$\lim_{\beta \to 0} g(n^*) = \frac{n^* f(n^*)}{E(N^*)} \tag{1.38}$$

Thus the limit form of (1.37) is identical to (1.28). Larger families tend to be oversampled and hence a misleading estimate of family size will be produced from such samples. Since the model is formally equivalent to the length biased sampling model, all references and statements about identification given in example 6 apply with full force to this example. See the discussion in Rao (1965).

2 Economic models of self-selection

We begin our analysis by expositing the Roy model of self-selection for workers with heterogeneous skills. The statistical framework for this model has been outlined in examples 3 and 4. Following Roy, we assume that there are two market sectors in which income-maximizing agents can work. Agents are free to enter the sector that gives them the highest income. However, they can work in only one sector at a time.
Each sector requires a unique sector-specific task. Each agent has two skills, T_1 and T_2, which he cannot use simultaneously. The model is short run in that aggregate skill distributions are assumed to be given. There are no costs of changing sectors, and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
investment is ignored. Because of this assumption, the model presented here applies to environments with certain or uncertain prices for sector-specific tasks. For simplicity and without any loss of generality (given the preceding assumptions), we assume an environment of perfect certainty.

Let T_i be the amount of sector i specific task a worker can perform. The price of task i is π_i. An agent works in sector 1 if his income is higher there, that is

$$\pi_1 T_1 > \pi_2 T_2 \tag{2.1}$$

Indifference between sectors is a negligible probability event if the $T_i = 1, 2$ are assumed to be continuous nondegenerate random variables. Throughout we assume that prices are positive $(\pi_i > 0)$.

The log wage in task i of an individual with endowment T_i is

$$\ln W_i = \ln \pi_i + \ln T_i \tag{2.2}$$

The proportion of the population working at task i is the proportion of the population for whom

$$T_1 > \frac{\pi_2}{\pi_1} T_2.$$

Roy assumes that ($\ln T_1$, $\ln T_2$) is normally distributed with mean (μ_1, μ_2) and covariance matrix Σ. Letting (U_1, U_2) be a mean zero normal vector, agents in the Roy model choose between two possible wages:

$$\ln W_1 = \ln \pi_1 + \mu_1 + U_1$$

or

$$\ln W_2 = \ln \pi_2 + \mu_2 + U_2.$$

Workers enter sector 1 if $\ln W_1 > \ln W_2$. Otherwise they enter sector 2.

Letting

$$\sigma^* = \sqrt{\mathrm{var}(U_1 - U_2)}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and

$$c_i = (\ln(\pi_i / \pi_j) + \mu_i - \mu_j) / \sigma^*,\ i \neq j, \Pr(i) = P(\ln W_i > \ln W_j) = \Phi(c_i),\ i \neq j,\ i,\ j = 1, 2$$

where $\Phi(\bullet)$ is the cumulative distribution function of a standard normal variable. When standard sample selection bias formulae are used (see, e.g. Heckman, 1976), the mean of log wages observed in sector *i* is

$$E(\ln W_i | \ln W_i > \ln W_j) = \ln \pi_i + \mu_i + \frac{\sigma_{ii} - \sigma_{ij}}{\sigma^*}\lambda(c_i), i,\ j = 1, 2,\ \ i \neq j, \tag{2.3}$$

where

$$\lambda(c) = \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}c^2\right)}{\Phi(c)}$$

is a convex monotone decreasing function of *c* with $\lambda(c) \geq 0$ and

$$\lim_{c \to \infty} \lambda(c) = 0, \quad \lim_{c \to -\infty} \lambda(c) = \infty.$$

Convexity is proved in Heckman and Honoré (1986).
The variance of log wages observed in sector *i*

$$\mathrm{var}(\ln W_i | \ln W_i > \ln W_j) = \sigma_{ii}\left\{\rho_i^2 [1 - c_i\lambda(c_i) - \lambda^2(c_i)] + (1 - \rho_i^2)\right\},\ \ i \neq j \tag{2.4}$$

where $\rho_i = \mathrm{correl}(U_i, U_i - U_j),\ i \neq j = 1, 2$. The variance of the log of observed wages never exceeds σ_{ii}, the population variance, because the term in braces in (2.4) is never greater than unity. In general, sectoral variances decrease with increased selection. For example, if ρ_1 and ρ_2 do not equal zero, as π_1 increases with π_2 held fixed so that people shift from sector 2 to sector 1, the variance in the log of wages in sector 1 increases while the variance in the log of wages in sector 2 decreases. Using the fact that $W_i = \pi_i T_i$, we may use (2.3) to write

$$E(\ln T_1 | \ln W_1 > \ln W_2) = \mu_1 + \frac{\sigma_{11} - \sigma_{12}}{\sigma^*}\lambda(c_1),$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(2.5a)

$$E(\ln T_2 | \ln W_1 > \ln W_2) = \mu_2 + \frac{\sigma_{22} - \sigma_{12}}{\sigma^*}\lambda(c_2).$$

(2.5b)

Focusing on (2.5a) and noting that λ is positive for all values of c_1 (except $c_1 = \infty$), the mean of log task 1 used in sector 1 exceeds, equals, or falls short of the population mean endowment of log task 1 as $\sigma_{11} - \sigma_{12}$ is greater than, equal to, or less than zero. If endowments of tasks are uncorrelated ($\sigma_{12} = 0$) self-selection always causes the mean of ln T_1 employed in sector 1 to be above the population mean μ_1. The opposite case occurs when $\sigma_{11} - \sigma_{12}$ is negative. This case can arise only when values of ln T_1 and ln T_2 are sufficiently positively correlated. If this occurs, the mean of log task 1 used in sector 1 falls below the population mean μ_1. Since covariance matrices must be positive semi-definite, $\sigma_{11} + \sigma_{22} - 2\sigma_{12} \geq 0$. Thus if $\sigma_{11} - \sigma_{12} < 0$, $\sigma_{22} - \sigma_{12} > 0$ so the mean of log task 2 employed in sector 2 necessarily lies above the population mean μ_2.

In the Roy model the unusual case can arise in at most one sector. Notice from (2.5) that only if $\sigma_{11} - \sigma_{12} = 0$ (so $\rho_1^2 = 0$) is the variance of log task 1 employed in sector 1 identical to the variance of log task 1 in the population. Otherwise, the sectoral variance of observed log task 1 is less than the population variance of log task 1.

To gain further insight into the effect of self-selection on the distribution of earnings for workers in sector 1, it is helpful to draw on some results from normal regression theory. The regression equation for ln T_2 conditional on ln T_1 is

$$\ln T_2 = \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(\ln T_1 - \mu_1) + \varepsilon_2,$$

(2.6)

where $E(\varepsilon_2) = 0$ and $\mathrm{var}(\varepsilon_2) = \sigma_{22}[1 - (\sigma_{12}^2 / \sigma_{11}\sigma_{22})]$.

Figure 2 plots regression function (2.6) for the case $\sigma_{12} = \sigma_{11}$ and $\mu_2 > \mu_1 > 0$. For each value of ln T_1, the population values of ln T_2 are normally distributed around the regression line. Individuals with high values of ln T_1 also tend to have a high value of ln T_2. Assuming $\pi_1 = \pi_2$, individuals with (ln T_1, ln T_2) endowments above the 45° line of equal income shown in Figure 1 choose to work in sector 2, while those individuals with endowments below this line work in sector 1. Because $\sigma_{12} = \sigma_{11}$, the regression function is parallel to the line of equal income.

The distribution of ε_2 about the regression line is the same for all values of ln T_1. When individuals are classified on the basis of their ln T_1 values the same proportion of individuals work in sector 1 at all values of ln T_1. For this reason the distribution of ln T_1 employed in sector 1 is the same as the latent population distribution. If π_1 is raised (or π_2 is lowered) so that the 45° equal income line is shifted upward, the same proportion of people enter sector 1 at each value of $T_1 = t_1$. Figure 3 plots regression function (2.6) for the case $\sigma_{12} > \sigma_{11}$ and $\mu_2 > \mu_1 > 0$.

Figure 2

Figure 2

Figure 3

Figure 3

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As before we set $\pi_1 = \pi_2$. Individuals with endowments above the 45° line choose to work in sector 2, while those with endowments below this line work in sector 1. When individuals are classified on the basis of their T_1 values, the fraction of people working in sector 1 decreases the higher the value of T_1. Self-selection causes the mean of log task 1 employed in sector 1 to be less than the mean of log task 1 in the total population. People with high values of T_1 are under-represented in sector 1 and low T_1 values are over-represented. In the extreme, when ln T_1 and ln T_2 are perfectly positively correlated, all high-income individuals are in sector 2, while all the low-income individuals are in sector 1. The highest-paid sector 1 worker earns the same as the lowest-paid sector 2 worker (Roy, 1951; Willis and Rosen, 1979). In this case there is really only one skill dimension and individuals can be unambiguously ranked along this scale.

If π_1 is raised (or π_2 is lowered) so that the line of equal income is shifted upward, the mean of ln T_1 employed in sector 1 must rise. The only place left to get T_1 is from the high end of the T_1 distribution. Unlike the case of $\sigma_{12} = \sigma_{11}$ in which a 10 per cent increase in π_1 results in a 10 per cent increase in measured average earnings in sector 1, when $\sigma_{12} > \sigma_{11}$ a 10 per cent increase in π_1 results in a greater than 10 per cent increase in the measured average earnings in sector 1 as the average quality of the sector 1 workforce increases. The variance of log wages in sector 1 increases.

If $\sigma_{11} < \sigma_{12}$ then $\sigma_{12} < \sigma_{22}$ in order for Σ to be a covariance matrix. In the population, log task 2 must have greater variability than log task 1. Individuals with high T_1 values tend to have high T_2 values. But the population distribution of log task 2 has more mass in the tails. The higher an agent's value of T_1, the more likely it is that he will be able to get higher income in sector 2. At the lower end of the distribution, the process works in reverse: lower T_1 individuals on average have poor T_2 values. Self-selection causes the ln T_1 distribution in sector 1 to have an evacuated right tail, an exaggerated left tail, and a lower mean than the population mean of ln T_1.

If $\sigma_{12} < \sigma_{11}$ (a case not depicted graphically), the proportion of each T_1 group working in sector 1 increases, the higher the value of T_1. The mean of the log task employed in sector 1 exceeds μ_1. A 10 per cent increase in π_1 produces an increase of less than 10 per cent in the average earnings of workers in sector 1 as the mean of ln T_1 employed in sector 1 declines. In fact if $\sigma_{12} > \sigma_{22}$ it is possible for an increase in π_1 to cause measured sector 1 wages to decline. Thus through a selection phenomenon it is possible for the average wage of people working in sector 1 to decline even though the price per unit skill increases there.

How robust are these conclusions if the normality assumption is relaxed? Heckman and Sedlacek (1985) show that many propositions derived from assumed normality of skills do not hold up for more general distributions. For example, increasing selection need not decrease sectoral variances. The effects of selection on mean employed skill levels are ambiguous. Heckman and Honoré (1986) demonstrate that in a single cross-section of data, it is possible to identify all of the parameters of the model from the data if the normality assumption is invoked. However, in a single cross-section many other models can explain the data equally well. In particular, intuitive notions about the degree of correlation or dependence among skills have no empirical content and so models of skill ‘hierarchies’ based on the extent of such dependence have no content for single cross-sections of data with all individuals facing common prices.

To show this, write the density of skills as $f(t_1, t_2)$. Let

$$Z_1 = \begin{cases} T_1 \text{ if } T_1 > T_2 \\ 0 \quad \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} T_2 \text{ if } T_2 > T_1 \\ 0 \quad \text{otherwise} \end{cases}$$

Prices are normalized to unity $(\pi_1 = \pi_2 = 1)$. Then the density of Z_1 is

$$Q_1'(z_1) = \int_{\{t_2 | t_2 < z_1\}} f(z_1, t_2)\, dt_2 = \int_0^{z_1} f(z_1, t_2)\, dt_2.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The density of Z_2 is

$$Q_2'(z_2) = \int_0^{z_2} f(t_1, z_2)\, dt_1.$$

Note that $Q_1'(n)$ and $Q_2'(n)$ summarize all of the available data on observed earnings.
Now if T_1, T_2 are independent with cdf's F_1^* and F_2^* respectively

$$Q_1'(n) = f_1^*(n) F_2^*(n)$$
$$Q_2'(n) = F_1^*(n) f_2^*(n).$$

Define

$$\bar{Q}(n) = \int_0^n [Q_1'(l) + Q_2'(l)]\, dl = F_1^*(n) F_2^*(n).$$

Then

$$\int_\varphi^\infty \frac{Q_1'(n)}{\bar{Q}(n)} dn = \int_\varphi^\infty \frac{f_1^*(n)}{F_1^*(n)} dn = -\ln F_1^*(\varphi).$$

Thus we can write

$$F_i^*(\varphi) = \exp - \left(\int_\varphi^\infty \left[\frac{Q_i'(n)}{\bar{Q}(n)} \right] dn \right) \quad i = 1, 2$$

so that we can always rationalize the data on wages in a single cross-section by a model of skill independence, and economic models of skill hierarchies have no empirical content for a single cross-section of data.
Suppose, however, that the observing economist has access to data on skill distributions in different market settings i.e. settings in which relative skill prices vary. To take an extreme case, suppose that we observe a continuum of values of π_1 / π_2 ranging from zero to infinity. Then it is possible to identify $F(t_1, t_2)$ and it is possible to give empirical content to models based on the degrees of dependence among latent skills.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This point is made most simply in a situation in which Z is observed but the analyst does not know Z_1 or Z_2 (i.e. which occupation is chosen). When $\pi_1 / \pi_2 = 0$ everyone works in occupation two. Thus we can observe the marginal density of t_2. When $\pi_1 / \pi_2 = \infty$ everyone works in occupation one. As π_1 / π_2 pivots from zero to infinity it is thus possible to trace out the full joint distribution of (T_1, T_2).

To establish the general result, set $\sigma = \pi_2 / \pi_1$. Let $F(t_1, t_2)$ be the distribution function of T_1, T_2. Then

$$\Pr(Z \le n) = \Pr(\max(T_1, \sigma T_2) \le n) = \Pr\left(T_1 \le n, T_2 \le \frac{1}{\sigma} n\right) = F\left(n, \frac{n}{\sigma}\right).$$

As σ varies between 0 and ∞ the entire distribution can be recovered since N is observed for all values in $(0, \infty)$. Note that it is not necessary to know which sector the agent selects.

This proposition establishes the benefit of having access to data from more than one market. Heckman and Honoré (1986) show how access to data from various market settings and information about the choices of agents aids in the identification of the latent skill distributions.

The Roy model is the prototype for many models of self-selection in economics. If T_1 is potential market productivity and T_2 is non-market productivity (or the reservation wage) for housewives or unemployed individuals, precisely the same model can be used to explore the effects of self-selection on measured productivity. In such a model, T_2 is never observed. This creates certain problems of identification discussed in Heckman and Honoré (1986). The model has been extended to allow for more general choice mechanisms. In particular, selection may occur as a function of variables other than or in addition to T_1 and T_2. Applications of the Roy model include studies of the union–non-union wage differential (Lee, 1978), the returns to schooling (Willis and Rosen, 1979), and the returns to training (Bjorklund and Moffitt, 1986, and Heckman and Robb, 1985). Amemiya (1984) and Heckman and Honoré (1986) present comprehensive surveys of empirical studies based on the Roy model and its extensions.

Bibliography

Amemiya, T. 1984. Tobit models: a survey. *Journal of Econometrics* 24, 3–61.

Bjorklund, A. and Moffitt, R. 1986. Estimation of wage gains and welfare gains from self selection models. *Review of Economics and Statistics* 24, 1–63.

Cain, G. and Watts, H. 1973. Toward a summary and synthesis of the evidence. In *Income Maintenance and Labor Supply*, ed. G. Cain and H. Watts. Madison: University of Wisconsin Press.

Clark, K. and Summers, L. 1979. Labor market dynamics and unemployment: a reconsideration. *Brookings Papers on Economic Activity*, 13–60.

Cosslett, S. 1981. Maximum likelihood estimation from choice based samples. *Econometric*a.

Cosslett, S. 1984. Distribution free estimator of regression model with sample selectivity. Unpublished manuscript, University of Florida.

Domencich, T. and McFadden, D. 1975. *Urban Travel Demand*. Amsterdam: North-Holland.

Flinn, C. and Heckman, J. 1982. New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics* 18, 5–168.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Gallant, R. and Nychka, R. 1984. Consistent estimation of the censored regression model. Unpublished manuscript, North Carolina State.

Gill, R. and Wellner, J. 1985. Large sample theory of empirical distributions in biased sampling models. Unpublished manuscript, University of Washington.

Goldberger, A. 1983. Abnormal selection bias. In *Studies in Econometrics, Time Series and Multivariate Statistics*, ed. S. Karlin, T. Amemiya and L. Goodman. Wiley, NY. Gronau, R. 1974. Wage comparisons – a selectivity bias. *Journal of Political Economy* 82(6), 1119–1144.

Hall, R. 1982. The importance of lifetime jobs in the US economy. *American Economic Review* 72, September, 716–724.

Heckman, J. 1974. Shadow prices, market wages and labor supply. *Econometrica* 42(4), 679–94.

Heckman, J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5(4), 475–92.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47(1), 153–62.

Heckman, J. and Honoré, B. 1986. The empirical content of the Roy model. Unpublished manuscript, University of Chicago.

Heckman, J. and MaCurdy, T. 1981. New methods for estimating labor supply functions. In *Research in Labor Economics*, Vol. 4, ed. R. Ehrenberg. Greenwich, Conn.: JAI Press.

Heckman, J. and MaCurdy, T. 1986. Labor econometrics. In *Handbook of Econometrics*, vol. 3, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Heckman, J. and Robb, R. 1985. Alternative methods for evaluating the effect of training on earnings. In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Heckman, J. and Sedlacek, G. 1985. Heterogeneity, aggregation and market wage functions. *Journal of Political Economy* 93, December, 1077–125.

Heckman, J. and Singer, B. 1985. Econometric analysis of longitudinal data. In *Handbook of Econometrics*, Vol. III, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Lee, L.F. 1978. Unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables. *International Economic Review* 19, 415–33.

Manski, C. and Lerman, S. 1977. The estimation of choice probabilities from choice based samples. *Econometrica* 45, 1977–88.

Manski, C. and McFadden, D. 1981. Alternative estimates and sample designs for discrete choice analysis. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. Manski and D. McFadden. Cambridge: MIT Press.

Pearson, K. 1901. Mathematical contributions to the theory of evolution. *Philosophical Transactions*, 195, 1–47.

Rao, C.R. 1965. On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Distributions*, ed. G. Patil. Calcutta: Pergamon Press.

Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3, 135–46.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Vardi, Y. 1983. Nonparametric estimation in the presence of length bias. *Annals of Statistics* 10, 616–20.

Vardi, Y. 1985. Empirical distributions in selection bias models. *Annals of Statistics*, 13, 178–203.

Willis, R. and Rosen, S. 1979. Education and self selection. *Journal of Political Economy* 87, S7–S36.

How to cite this article

Heckman, James J. "selection bias and self-selection." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000084>
doi:10.1057/9780230226203.1504

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

workers' cooperation, mutual training, and tenure longevity is another old idea in economics. A recent neoclassical application, with abundant citations, is that of Oliver E. Williamson (1975), who argues that such institutional devices as implicit contracts, collective bargaining, internal promotion ladders, and seniority rights are economically efficient when jobs and workers are heterogeneous and idiosyncratic.

A fixed structure of wages for jobs, which is emphasized by segmentation economists, is descriptively accurate and useful for analysing short-run behaviour, but even in the short run a human capital model of supply-side productivity traits can explain the match of workers to a hierarchy of wage-fixed jobs. In the long run the human capital model can explain changes in workers' productivity traits, and neoclassical models generally would predict changes in the structure of both jobs and wages.

A discussion of empirical work and policy issues concerning segmented labour markets is beyond the scope of this entry (see the bibliography below). It should be stated, however, that the sometime claim that the neoclassical economists ignore the demand side of the market in policy discussions is unfounded.

That labour market outcomes and processes are complex and controversial is evident in the intellectual legacy of the above-listed five sources of inequality. The criticisms and empirical work of the segmented labour market economists have added to this legacy, but they, like the earlier dissenters, the Marxists and the Institutionalists, remain on the bank of the mainstream.

GLEN G. CAIN

BIBLIOGRAPHY

The literature on segmented labour markets is extensive and diversified, and there are disputes about who are the leading theorists and which are the landmark articles. These characteristics make it difficult to provide a brief bibliography. In addition to the items cited in the text, several survey articles and books contain lengthy bibliographies: Taubman and Wachter (1986); Gordon, Edwards and Reich (1982); Wilkinson (1981); Cain (1976). The application of segmented labour market theories to development economics is not, however, covered in these sources, and the author is unaware of any survey or bibliographic sources for this application.

Becker, G.S. 1964. *Human Capital*. New York: Columbia University Press for the National Bureau of Economic Research.

Berger, S. and Piore, M.J. 1980. *Dualism and Discontinuity in Industrial Societies*. Cambridge: Cambridge University Press.

Braverman, H. 1974. *Labor and Monopoly Capital*. New York: Monthly Review Press.

Cain, G. 1976. The challenge of segmented labor market theories to orthodox theory: a survey. *Journal of Economic Literature* 14(4), December, 1215–57.

Cairnes, J.E. 1874. *Some Leading Principles of Political Economy*. New York: Harper & Brothers.

Gordon, D.M., Edwards, R.C. and Reich, M.S. 1982. *Segmented Work, Divided Workers: The Historical Transformation of Labor in the United States*. Cambridge: Cambridge University Press.

Marshall, A. [1890] 1959. *Principles of Economics*. 8th edn, London: Macmillan.

Mill, J.S. [1848] 1900. *The Principles of Political Economy*, Vol. 1. Revised edn, The World's Greatest Classics, New York: Colonial Press.

Myrdal, G. 1944. *An American Dilemma*. New York: Harper & Row.

Taubman, P. and Wachter, M.L. 1986. Segmented labor markets. In *Handbook of Labor Economics*, ed. O. Ashenfelter and R. Layard, Amsterdam: Elsevier Science Publishers.

Werthheim, W.F. 1967. Economy, dual. In *International Encyclopedia of the Social Sciences*, Vol. 4, New York: Macmillan and Free Press, 495–500.

Wilkinson, F. (ed.) 1981. *The Dynamics of Labor Market Segmentation*. New York: Academic Press.

Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

seigniorage. Full-bodied monies such as gold coin contain metal approximately equal in value to the face value of the coin. Under the gold standard, metal could be brought to the mint and freely coined into gold, less a small *seigniorage* charge for the privilege. Subsidiary or token coin and paper money by contrast cost much less to produce than their face value. The excess of the face value over the cost of production of currency is also called *seigniorage*, because it accrued to the *seigneur* or ruler who issued the currency, in early times.

The use of paper money instead of full-bodied coin by modern governments generates a very large social saving in the use of the resources that would otherwise have to be expended in mining and smelting large quantities of metal. The value of this seigniorage can be measured by considering the aggregate demand curve for currency, as a function of the rate of interest. The area under this demand curve represents the aggregate flow of social benefits from holding currency, under certain assumptions. The social cost of holding currency is measured by the opportunity cost of the resources it takes to produce the currency. If gold were used for currency, its opportunity cost would be measured by the rate of interest that could be earned on those resources if transferred to some other use. Thus the area under the demand curve between the market rate of interest and the cost of providing paper currency represents the flow of seigniorage or social saving that accrues from the use of paper currency instead of gold.

In the international monetary system, gold remains a very large fraction of total holdings of international reserves (about 45 per cent of total reserves valued at market prices at the end of March 1985). Substitution of fiduciary reserve assets such as Special Drawing Rights created by the International Monetary Fund or United States dollars for gold would generate a substantial social gain in the form of seigniorage equal to the excess of the opportunity cost of capital over the costs of providing the fiduciary asset. If interest is paid to the holders of the reserve asset, the seigniorage is split between the issuer and the holder.

The existence of these large seigniorage gains is what led to the development of the gold exchange standard, under which first British sterling, before World War II, and since then United States dollars and other currencies have substituted for gold in international reserve holdings. As interest rates paid on these reserve assets have risen, more of the seigniorage has accrued to holders of reserve assets.

Further substitution of fiduciary reserve assets for gold in the international monetary system has frequently been suggested, and the Second Amendment to the Charter of the International Monetary Fund adopted in 1978 proposed such a goal. Little progress has been made, however, since the underlying issue is one of trust in the financial probity of the issuer and its continued political stability, as well as its continued willingness to convert reserve assets into usable currencies over long periods of time.

S. BLACK

selection. *See* COMPETITION AND SELECTION.

selection bias and self-selection. The problem of selection bias in economic and social statistics arises when a rule other than simple random sampling is used to sample the underlying

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

population that is the object of interest. The distorted representation of a true population as a consequence of a sampling rule is the essence of the selection problem. Distorting selection rules may be the outcome of decisions of sample survey statisticians, self-selection decisions by the agents being studied or both.

A random sample of a population produces a description of the population distribution of characteristics that has many desirable properties. One attractive feature of a random sample generated by the *known rule* that all individuals are equally likely to be sampled is that it produces a description of the population distribution of characteristics that becomes increasingly accurate as sample size expands.

A sample selected by any rule not equivalent to random sampling produces a description of the population distribution of characteristics that does not accurately describe the true population distribution of characteristics no matter how big the sample size. Unless the rule by which the sample is selected is known or can be recovered from the data, the selected sample cannot be used to produce an accurate description of the underlying population. For certain sampling rules, even knowledge of the rule generating the sample does not suffice to recover the population distribution from the sampled distribution.

This entry defines the problem of selection bias and presents conditions required to solve the problem. Examples of various types of commonly encountered sampling frames are given and specific economic selection mechanisms are presented. Assumptions required to use selected samples to determine features of the population distribution are discussed.

The analytical framework developed to understand the inferential problems raised by selection bias is also fruitful in understanding the economics of self-selection. The prototypical choice theoretic model of self-selection is that of Roy (1951). In his model, agents choose among a variety of discrete 'occupational' opportunities. Agents can pursue only one 'occupation' at a time. While every person can, in principle, do the work in each 'occupation', at least at some level of competence, self-interest drives individuals to choose that 'occupation' which produces the highest income (utility) for them. As in the statistical selection bias problem, there is a latent population (of skills). Observed (utilized) skill distributions are the outcome of a selection rule by agents. The relationship between observed and latent skill distributions is of considerable interest and underlies recent work on worker hierarchies (see Willis and Rosen, 1979). The 'occupations' can be: (a) market work or non-market work (b) unemployed and searching or working at the offered wage (c) working in one province or working in another, or (d) any choice among a set of mutually exclusive opportunities.

Because the insights in the Roy model underly much recent research, we present a brief exposition of it and demonstrate how it can be or has been fruitfully extended to a variety of settings. An important issue, closely linked to the problem of identifying population parameters from selected sample distributions, is the empirical content of economic models of self-selection and worker hierarchies. Are they artefacts of distributional assumptions for unobservable skills or are they genuine behavioural hypotheses?

1. A DEFINITION AND SOME EXAMPLES OF SELECTION BIAS

Any selection bias model can be described by the following set-up. Let $\mathbf{Y}$ be a vector of outcomes of interest and let $\mathbf{X}$ be a vector of 'control' or 'explanatory' variables. The population distribution of $(\mathbf{Y}, \mathbf{X})$ is $F(\mathbf{y}, \mathbf{x})$. To simplify the exposition we assume that the density is well defined and write it as $f(\mathbf{y}, \mathbf{x})$.

Any sampling rule can be interpreted as producing a non-negative weighting function $\omega(\mathbf{y}, \mathbf{x})$ that alters the population density. Let $(\mathbf{Y}^*, \mathbf{X}^*)$ denote the sampled random variables. The density of the sampled data $g(\mathbf{y}^*, \mathbf{x}^*)$ may be written as

$$g(\mathbf{y}^*, \mathbf{x}^*) = \omega(\mathbf{y}^*, \mathbf{x}^*)\mathbf{f}(\mathbf{y}^*, \mathbf{x}^*) \Big/ \int \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*)\, d\mathbf{y}^*\, d\mathbf{x}^* \quad (1.1)$$

where the denominator of the expression is introduced to make the density $g(\mathbf{y}^*, \mathbf{x}^*)$ integrate to one as is required for proper densities.

Alternatively, the weight may be defined as

$$\omega^*(\mathbf{y}^*, \mathbf{x}^*) = \frac{\omega(\mathbf{y}^*, \mathbf{x}^*)}{\int \omega(\mathbf{y}^*, \mathbf{x}^*)\mathbf{f}(\mathbf{y}^*, \mathbf{x}^*)\, d\mathbf{y}^*\, d\mathbf{x}^*}$$

so that

$$g(\mathbf{y}^*, \mathbf{x}^*) = \omega^*(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*). \quad (1.2)$$

Sampling schemes for which $\omega(\mathbf{y}, \mathbf{x}) = 0$ for some values of $(\mathbf{Y}, \mathbf{X})$ create special problems. For such schemes, not all values of $(\mathbf{Y}, \mathbf{X})$ are sampled. Let indicator variable $i(\mathbf{x}, \mathbf{y}) = 0$ if a potential observation at values $\mathbf{y}, \mathbf{x}$ cannot be sampled and let $i(\mathbf{y}, \mathbf{x}) = 1$ otherwise. Let $\Delta = 1$ record the occurrence of the event 'a potential observation is sampled, i.e. the value of $\mathbf{y}, \mathbf{x}$ is observed' and let $\Delta = 0$ if it is not. In the population, the proportion that is sampled is

$$\Pr(\Delta = 1) = \int i(\mathbf{y}, \mathbf{x}) f(\mathbf{y}, \mathbf{x})\, d\mathbf{y}\, d\mathbf{x}. \quad (1.3)$$

while

$$\Pr(\Delta = 0) = 1 - \Pr(\Delta = 1).$$

For samples in which $\omega(\mathbf{y}, \mathbf{x}) = 0$ for a non-negligible proportion of the population ($\Pr(\Delta = 0) > 0$), it is clarifying to consider two cases. A *truncated sample* is one for which $\Pr(\Delta = 1)$ is not known and cannot be consistently estimated. For such a sample, (1.1) is the density of all of the sampled $\mathbf{Y}$ and $\mathbf{X}$ values. A *censored sample* is one for which $\Pr(\Delta = 1)$ is known or can be consistently estimated. The sampling rule in this case is such that values of $\mathbf{y}, \mathbf{x}$ for which $\omega(\mathbf{y}, \mathbf{x}) = 0$ are not known but it is known whether or not $i(\mathbf{y}, \mathbf{x}) = 0$ for all values of $\mathbf{Y}, \mathbf{X}$. In this case it is notationally convenient to define $(\mathbf{Y}^*, \mathbf{X}^*) = (\mathbf{0}, \mathbf{0})$ for values of $\mathbf{y}, \mathbf{x}$ such that $\omega(\mathbf{y}, \mathbf{x}) = i(\mathbf{y}, \mathbf{x}) = 0$. Such a definition is innocuous provided that in the population there is no point mass (concentration of probability mass) at $(\mathbf{0}, \mathbf{0})$. (Any value other than $(\mathbf{0}, \mathbf{0})$ can be selected provided that there is no point mass at that value). Given $\Delta = 0$, the distribution of $\mathbf{Y}^*, \mathbf{X}^*$ is

$$G(\mathbf{y}^*, \mathbf{x}^*) = 1 \quad \text{for} \quad \Delta = 0$$

at

$$\mathbf{Y}^* = \mathbf{0} \quad \text{and} \quad \mathbf{X}^* = \mathbf{0}.$$

The joint density of $\mathbf{Y}^*, \mathbf{X}^*, \Delta$ for the case of a censored sample is obtained by combining (1.1) and (1.3). Thus

$$g(\mathbf{y}^*, \mathbf{x}^*, \delta) = \left[\frac{\omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*)}{\int \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*)\, d\mathbf{y}^*\, d\mathbf{x}^*}\right]^{\delta} \times \left[\int i(\mathbf{y}, \mathbf{x}) f(\mathbf{y}, \mathbf{x})\, d\mathbf{y}\, d\mathbf{x}\right]^{\delta} \times [1]^{1-\delta} \left[\int (1 - i(\mathbf{y}, \mathbf{x})) f(\mathbf{y}, \mathbf{x})\, d\mathbf{y}\, d\mathbf{x}\right]^{1-\delta}. \quad (1.4)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The first term on the right-hand side of (1.4) is the conditional density of $\mathbf{Y}^*, \mathbf{X}^*$ given $\Delta = 1$. The second term is the probability that $\Delta = 1$. The third term is the conditional density of $\mathbf{Y}^*, \mathbf{X}^*$ given $\Delta = 0$. This density assigns unit mass to $\mathbf{y}^* = 0, \mathbf{x}^* = \mathbf{0}$ when $\Delta = 0$. The fourth term is the probability that $\Delta = 0$. Notice that in the case in which $\omega(\mathbf{y}, \mathbf{x}) > 0$ for all $\mathbf{y}, \mathbf{x}, \Delta = 1$ and (1.4) is identical to (1.1).

In a random sample $\omega(\mathbf{y}^*, \mathbf{x}^*) = 1$ (and so $\omega^*(\mathbf{y}^*, \mathbf{x}^*) = 1$). In a selected sample, the sampling rule weights the data differently. Values of $(\mathbf{Y}, \mathbf{X})$ are over-sampled or under-sampled relative to their occurrence in the population. In the case of truncated samples, the weight is zero for certain values of the outcome.

In many problems in economics, attention focuses on $f(\mathbf{y}|\mathbf{x})$, the conditional density of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. In such problems knowledge of the population distribution of $\mathbf{X}$ is of no direct interest. If samples are selected solely on the $\mathbf{x}$ variables ('selection on the exogenous variables'), $\omega(\mathbf{y}, \mathbf{x}) = \omega(\mathbf{x})$ and there is no problem about using selected samples to make valid inference about the population conditional density. This is so because in the case of selection on the exogenous variables

$$g(\mathbf{y}^*, \mathbf{x}^*) = f(\mathbf{y}^*|\mathbf{x}^*) \frac{\omega(\mathbf{x}^*) f(\mathbf{x}^*)}{\int \omega(\mathbf{x}^*) f(\mathbf{x}^*)\, d\mathbf{x}}$$

and

$$g(\mathbf{x}^*) = \frac{\omega(\mathbf{x}^*) f(\mathbf{x}^*)}{\int \omega(\mathbf{x}^*) f(\mathbf{x}^*)\, d\mathbf{x}^*}$$

Thus

$$g(\mathbf{y}^*|\mathbf{x}^*) = \frac{g(\mathbf{y}^*, \mathbf{x}^*)}{g(\mathbf{x}^*)} = f(\mathbf{y}^*|\mathbf{x}^*).$$

For such problems, sample selection distorts inference only if selection occurs on $\mathbf{y}$ (or $\mathbf{y}$ and $\mathbf{x}$). Sampling on both $\mathbf{y}$ and $\mathbf{x}$ is termed *general stratified sampling*.

From a sample of data, it is not possible to recover the true density $f(\mathbf{y}, \mathbf{x})$ without knowledge of the weighting rule. On the other hand, if the weighting rule is known ($\omega(\mathbf{y}^*, \mathbf{x}^*)$), the density of the sampled data is known ($g(\mathbf{y}^*, \mathbf{x}^*)$), the support of $(\mathbf{y}, \mathbf{x})$ is known and $\omega(\mathbf{y}, \mathbf{x})$ is nonzero, then $f(\mathbf{y}, \mathbf{x})$ can always be recovered because

$$\frac{g(\mathbf{y}^*, \mathbf{x}^*)}{\omega(\mathbf{y}^*, \mathbf{x}^*)} \frac{f(\mathbf{y}^*, \mathbf{x}^*)}{\int \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*)\, d\mathbf{y}^*\, d\mathbf{x}^*} \tag{1.5}$$

and by hypothesis both the numerator and denominator of the left-hand side are known. From the requirement that $(\mathbf{y}^*, \mathbf{x}^*)$ has a well defined density

$$\int f(\mathbf{y}^*, \mathbf{x}^*)\, d\mathbf{y}^*\, d\mathbf{x}^* = 1.$$

Integrating the left-hand side of (1.5) it is possible to determine $\int \omega(\mathbf{y}^*, \mathbf{x}^*) f(\mathbf{y}^*, \mathbf{x}^*)\, d\mathbf{y}^*\, d\mathbf{x}^*$ and hence to use (1.5) to recover the population density of the data.

The requirements that (a) the support of $(\mathbf{y}, \mathbf{x})$ is known and (b) $\omega(\mathbf{y}, \mathbf{x})$ is nonzero are not innocuous. In many important problems in economics requirement (b) is not satisfied: the sampling rule excludes observations for certain values of $\mathbf{y}, \mathbf{x}$ and hence it is impossible without invoking further assumptions to determine the population distribution of $(\mathbf{Y}, \mathbf{X})$ at those values. If neither the support nor the weight is known, it is impossible, without invoking strong assumptions, to determine whether the fact that data are missing at certain $\mathbf{y}, \mathbf{x}$ values is due to the sampling plan or that the population density has no suport at those values. We now turn to some specific sampling plans of interest in economics.

Example 1. Data are collected on incomes of individuals whose income Y exceeds a certain value c (for cutoff value). The rule is to observe Y if $Y > c$. Thus $\omega(y) = 1$ if $y > c$ and $\omega(y) = 0$ if $y \leqslant c$. Because the weight is zero for some values of y, we know that knowledge of the sampling rule does *not* suffice to recover the population distribution. From a random sample of the entire population, the social scientist knows or can consistently estimate (a) the sample distribution of Y above c and (b) the proportion of the original random sample with income below c ($F(c)$ where F is the distribution function of Y). The social scientist does not observe values of Y below c.

In this example, observed income is a *truncated random variable*. The point of truncation is c. The *sample* of observed income is said to be *censored*. If the proportion of the original random sample with income below c is not known and cannot be consistently estimated, the *sample* is *truncated*. In a truncated sample, nothing is known about the proportion of the underlying population that can appear in the sample. A sample is truncated only if $\omega(\mathbf{y}) = 0$ for some intervals of $\mathbf{y}$ (for $\mathbf{y}$ continuous) or if $\omega(\mathbf{y}) = 0$ at values of $\mathbf{y}$ at which there is finite probability mass. In a censored sample, the proportion of the underlying population that can appear in the sample is known, at least to an arbitrarily high degree of approximation, as sample size increases.

Let $Y^* = Y$ if $Y > c$. Define $Y^* = 0$ otherwise (the choice of the value for Y^* when Y is not observed is inessential and any value can be used in place of 0 provided that the true distribution places no mass at the selected value). Define an indicator variable $\Delta = 1$ if $Y > c$. $\Delta = 0$ otherwise. Then the distribution of Y^* is

$$G(y^*|Y > 0) = F(y^*|Y \quad c) = F(y^*|\delta = 1) = \frac{F(y^*)}{1 - F(c)}, y^* > c. \tag{1.6a}$$

$$G(y^*|Y^* > 0) = 1 \qquad \text{for} \qquad Y^* = 0\,(\Delta = 0). \tag{1.6b}$$

Observe that (1.6a) is obtained from (1.1) by setting $\omega(y^*) = 1$ if $y > c$, and $\omega(y^*) = 0$ otherwise, and integrating up with respect to y^*. The distribution of Δ is

$$\mathrm{pr}(\Delta) = [1 - F(c)]^{\delta}[F(c)]^{1-\delta}.$$

The joint distribution of (Y^*, Δ) is

$$\begin{aligned} F(y^*, \delta) &= F(y^*|\delta)\Pr(\delta) \\ &= \left\{\frac{F(y^*)}{(1 - F(c))}\right\}^{\delta} [1 - F(c)]^{\delta}(1)^{1-\delta}[F(c)]^{1-\delta} \\ &= [F(y^*)]^{\delta}[F(c)]^{1-\delta}. \end{aligned} \tag{1.7}$$

Note that (1.7) is obtained from (1.4) by setting $\omega(y) = 0, y < c, \omega(y) = 1$ otherwise, by setting $i(y) = \omega(y)$, and by integrating up with respect to y^*. For normally distributed Y, (1.7) is the 'Tobit' distribution.

The difference between the information in a truncated sample and the information in a censored sample is encapsulated in the contrast between (1.6a) and (1.7). Clearly there is more information in a censored sample than in a truncated sample because one can obtain (1.6a) from (1.7) (by conditioning on $\Delta = 1$) but not vice versa.

Inferences about the population distribution based on assuming that $F(y^*|Y > c)$ closely approximates $F(y)$ are potentially very misleading. A description of population income inequality based on a subsample of high income people may convey no information about the true population distribution.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Without further information about F and its support, it is not possible to recover F from $G(y^*)$ from either a censored or a truncated sample. Access to a censored sample enables the analyst to recover $F(y)$ for $y > c$ but obviously does not provide any information on the shape of the true distribution for values of $y \leqslant c$.

This problem is routinely 'solved' by assuming that F is of a known functional form. This solution strategy does not always work. If F is normal, then it can be recovered from a censored or truncated sample (Pearson, 1901). If F is Pareto, F cannot be recovered from either a truncated or a censored sample (see Flinn and Heckman, 1982). If F is real analytic (i.e. possesses derivatives of all order) and the support of Y is known, then F can be recovered (Heckman and Singer, 1985).

Example 2. Expand the discussion in the previous example to a linear regression setting. Let

$$Y = \mathbf{X}\boldsymbol{\beta} + U \tag{1.8}$$

be the population earnings function where Y is earnings, $\mathbf{X}$ is a regressor vector assumed to be distributed independently of mean zero disturbance U. '$\boldsymbol{\beta}$' is a suitably dimensioned parameter vector. Conventional assumptions are invoked to ensure that ordinary least squares applied to a random sample of earnings data consistently estimates $\boldsymbol{\beta}$.

Data are collected on incomes of persons for whom Y exceeds c. Again the weight depends solely on y, i.e. $\omega(y, \mathbf{x}) = 0, y \leqslant c, \omega(y, \mathbf{x}) = 1, y > c$. The social scientist knows or can consistently estimate (a) the sample distribution of Y above c (b) the sample distribution of the $\mathbf{X}$ for Y above c and (c) the proportion of the original random sample with income below c. The social scientist does not observe values of Y below c.

As before, let $Y^* = Y$ if $Y > c$. Define $Y^* = 0$ otherwise. $\Delta = 1$ if $Y > c, \Delta = 0$ otherwise. The probability of the event $\Delta = 1$ given $\mathbf{X} = \mathbf{x}$ is

$$\begin{aligned}\Pr(\Delta = 1 \mid \mathbf{X} = \mathbf{x}) &= \Pr(Y > c \mid \mathbf{X} = \mathbf{x}) \\ &= \Pr(Y > c - \mathbf{x}\boldsymbol{\beta} \mid \mathbf{X} = \mathbf{x}).\end{aligned}$$

Invoking independence between U and $\mathbf{X}$ and letting F_u denote the distribution of U,

$$\Pr(\Delta = 1 \mid \mathbf{X} = \mathbf{x}) = 1 - F_u(c - \mathbf{x}\boldsymbol{\beta}) \tag{1.9a}$$

and

$$\Pr(\Delta = 0 \mid \mathbf{X} = \mathbf{x}) = F_u(c - \mathbf{x}\boldsymbol{\beta}). \tag{1.9b}$$

The distribution of Y^* conditional on $\mathbf{X}$ is

$$\begin{aligned}G(y^* \mid Y > 0, \mathbf{X} = \mathbf{x}) &= F(y^* \mid X = x, Y > c) \\ &= F(y^* \mid \mathbf{X} = \mathbf{x}, \Delta = 1) \\ &= \frac{F_u(y^* - \mathbf{x}\boldsymbol{\beta})}{1 - F_u(c - \mathbf{x}\boldsymbol{\beta})}, \qquad y^* > c. \end{aligned} \tag{1.10a}$$

$$G(y^* \mid Y \leqslant 0) = 1 \quad \text{for} \quad Y^* = 0\,(\Delta = 0). \tag{1.10b}$$

The joint distribution of (Y^*, Δ) given $\mathbf{X} = \mathbf{x}$ is

$$\begin{aligned}F(y^*, \delta \mid \mathbf{X} = \mathbf{x}) &= F(y^* \mid \delta, \mathbf{x}) \Pr(\delta \mid \mathbf{x}) \\ &= \{F_u(y^* - \mathbf{x}\boldsymbol{\beta})\}^{\delta}\{F_u(c - \mathbf{x}\boldsymbol{\beta})\}^{1-\delta}\end{aligned} \tag{1.11}$$

In particular,

$$\begin{aligned}E(Y^* \mid \mathbf{X} = \mathbf{x}, \Delta = 1) &= \mathbf{x}\boldsymbol{\beta} + E(U \mid \mathbf{X} = \mathbf{x}, \delta = 1) \\ &= \mathbf{x}\boldsymbol{\beta} + \int_{c - \mathbf{x}\boldsymbol{\beta}}^{\infty} \frac{z \, dF_u(z)}{(1 - F_u(c - \mathbf{x}\boldsymbol{\beta}))}\end{aligned} \tag{1.12}$$

where z is a dummy variable of integration. In contrast, the population mean regression function is

$$E(Y \mid \mathbf{X} = \mathbf{x}) = \mathbf{x}\boldsymbol{\beta}. \tag{1.13}$$

The contrast between (1.12) and (1.13) is illuminating. Many behavioural theories in social science produce empirical counterparts of (1.8) with population conditional expectations like (1.13). Such theories sometimes restrict the signs, permissible values and other relationships among the coefficients in $\boldsymbol{\beta}$. When the theoretical model is estimated on a selected sample ($\Delta = 1$), the true conditional expectation is (1.12) not (1.13). The conditional mean of U depends on $\mathbf{x}$. In terms of conventional omitted variable analysis, $E(U \mid \mathbf{X} = \mathbf{x}, \Delta = 1)$ is omitted from the regression. Since this term is a function of $\mathbf{x}$ it is likely to be correlated with $\mathbf{x}$. Least squares estimates of $\boldsymbol{\beta}$ obtained on selected samples which do not account for selection are biased and inconsistent.

To illustrate the nature of the bias, it is useful to draw on the work of Cain and Watts (1973). Suppose that X is a scalar random variable (e.g. education) and that its associated coefficient is positive ($\beta > 0$). Under conventional assumptions about U (e.g. mean zero, independently and identically distributed and distributed independently of X), the population regression of Y on X is a straight line. The scatter about the regression line and the regression line are given in Figure 1. When $Y > c$ is imposed as a sample inclusion requirement, lower population values of U are excluded from the sample in a way that systematically depends on x. ($Y > c$ or $U > c - x\beta$). As x increases, the conditional mean of $U[E(U \mid X = x, \Delta = 1)]$ decreases. Regression estimates of β that do not correct for sample selection (i.e. include $E(U \mid X = x, \Delta = 1)$ as a regressor) are downward biased because of the negative correlation between x and $E(U \mid X = x, \Delta = 1)$. See the flattened regression line for the selected sample in Figure 1.

In models with more than one regressor, no sharp result on the sign of the bias in the regression estimate that results from ignoring the selected nature of the sample is available except when the $\mathbf{X}$ variables are from certain distributions (e.g. normal, see Goldberger, 1983). None the less, the key result – that conventional least squares estimates of $\boldsymbol{\beta}$ obtained from selected samples are biased and inconsistent remains true.

As in example 1, it is fruitful to distinguish between the case of a truncated sample and the case of a censored sample. In the truncated sample case, no information is available about the fraction of the population that would be allocated to the truncated sample [$\Pr(\Delta = 1)$]. In the censored sample case, this

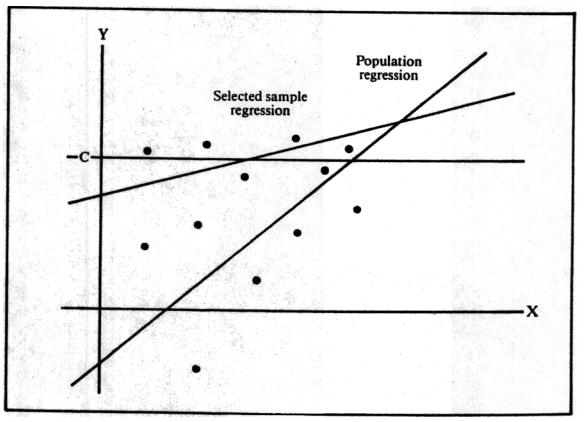

Figure 1

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

fraction is known or can be consistently estimated. In the censored sample case it is fruitful to distinguish two further cases: (a) the case in which $\mathbf{X}$ is not observed when $\Delta = 0$ and (b) the case in which it is. Case (b) is the one most fully developed in the literature (Heckman and MaCurdy, 1981).

Note that the conditional mean $E(U|\mathbf{X}=\mathbf{x}, \Delta = 1)$ is a function of $c - \mathbf{x}\boldsymbol{\beta}$ solely through $\Pr(\Delta = 1|\mathbf{x})$. Since $\Pr(\Delta = 1|\mathbf{x})$ is monotonic in $c - \mathbf{x}\boldsymbol{\beta}$, the conditional mean depends solely on $\Pr(\Delta = 1|\mathbf{x})$ and the parameters F_u i.e. since

$$F_u^{-1}(1 - \Pr(\Delta = 1|\mathbf{x})) = c - \mathbf{x}\boldsymbol{\beta},$$

$$E(U|\mathbf{X}=\mathbf{x}, \Delta = 1) = \int_{F_u^{-1}\{1-\Pr(\Delta=1|\mathbf{x})\}}^{\infty} \frac{z\, dF_u(z)}{\Pr(\Delta = 1|\mathbf{x})}.$$

This relationship demonstrates that the conditional mean is a function of the probability of selection. As the probability of selection goes to 1, the conditional mean goes to zero. For samples chosen so that the values of $\mathbf{x}$ are such that the observations are certain to be included in the sample, there is no problem in using ordinary least squares on selected samples to estimate $\boldsymbol{\beta}$. Thus in Figure 1, ordinary least squares regressions fit on samples selected to have large $\mathbf{x}$ values closely approximate the true regression function and become arbitrarily close as $\mathbf{x}$ becomes large. The condition mean in (1.12) is a surrogate for $\Pr(\Delta = 1|\mathbf{x})$. As this probability goes to one, the problem of sample selection in regression analysis becomes negligibly small.

Heckman (1976) demonstrates that $\boldsymbol{\beta}$ and F_u are identified if U is normally distributed and standard conditions invoked in regression analysis are satisfied. Gallant and Nychka (1984) and Cosslett (1984) establish conditions for identification for non-normal U. In their analyses, F_u is consistently non-parametrically estimated.

Example 3. The next example considers *censored random variables*. This concept extends the notion of a truncated random variable by letting a more general rule than truncation on the outcome of interest generate the selected sample. Because the sample generating rule may be different from a simple truncation of the outcome being studied, the concept of a censored random variable in general requires at least two distinct random variables.

Let Y_1 be the outcome of interest. Let Y_2 be another random variable. Denote observed Y_1 by Y_1^*. If $Y_2 < c$, Y_1 is observed. Otherwise Y_1 is not observed and we can set $Y_1^* = 0$ or any other convenient value (assuming that Y_1 has no point mass at $Y_1 = 0$ or at the alternative convenient value). In terms of the weighting function ω, $\omega(y_1, y_2) = 0$ if $y_2 > c$, $\omega(y_1, y_2) = 1$ if $y_2 \leqslant c$.

Selection rule $Y_2 < c$ does not necessarily restrict the range of Y_1. Thus Y_1^* is not in general a truncated random variable. Define $\Delta = 1$ if $Y_2 < c$; $\Delta = 0$ otherwise. If $F(y_1, y_2)$ is the population distribution of (Y_1, Y_2), the distribution of Δ is

$$\Pr(\Delta = \delta) = [1 - F_2(c)]^{1-\delta}[F_2(c)]^{\delta}, \quad \delta = 0, 1,$$

where F_2 is the marginal distribution of Y_2. The distribution of Y_1^* is

$$G(y_1^*) = F(y_1^*|\delta = 1) = \frac{F(y_1^*, c)}{F_2(c)}, \qquad \Delta = 1, \tag{1.14a}$$

$$G(y_1^* = 0) = 1, \qquad \Delta = 0. \tag{1.14b}$$

Note that (1.14a) is the distribution function corresponding to the density in (1.1) when $\omega(y_1, y_2) = 1$ if $y_2 \leqslant c$ and $\omega(y_1, y_2) = 0$ otherwise.

The joint distribution of (Y_1^*, Δ) is

$$G(y_1^*, \delta) = [F(y_1^*, c)]^{\delta}[1 - F_2(c)]^{1-\delta}. \tag{1.15}$$

This is the distribution function corresponding to density (1.4) for the special weighting rule of this example. In a censored sample, under general conditions it is possible to consistently estimate $\Pr(\Delta = \delta)$ and $G(y_1^*)$. In a truncated sample, only conditional distribution (1.14a) can be estimated. A degenerate version of this model has $Y_1 \equiv Y_2$. In that case, censored random variable Y_1 is also a truncated random variable. Note that a censored random variable may be defined for a truncated or censored sample.

Example 3 and variants of it have wide applicability in economics. Let Y_1 be the wage of a woman. Wages of women are observed only if women work. Let Y_2 be an index of a woman's propensity to work. In Gronau (1974) and Heckman (1974), Y_2 is postulated as the difference between reservation wages (the value of time at home determined from household preference functions) and potential market wages Y_1. Then if $Y_2 < 0$, the woman works. Otherwise, she does not. $Y_1^* = Y_1$ if $Y_2 < 0$ is the observed wage.

If Y_1 is the offered wage of an unemployed worker, and Y_2 is the difference between reservation wages (the return to searching) and offered market wages, $Y_1^* = Y_1$ if $Y_2 < 0$ is the accepted wage for an unemployed worker (see Flinn and Heckman, 1982). If Y_1 is the potential output of a firm and Y_2 is its profitability, $Y_1^* = Y_1$ if $Y_2 > 0$. If Y_1 is the potential income in occupation one and Y_2 is the potential income in occupation two, $Y_1^* = Y_1$ if $Y_1 - Y_2 < 0$ while $Y_2^* = Y_2$ if $Y_1 - Y_2 \geqslant 0$. We develop this example at length in section 2 where we consider explicit economic models of self-selection. There we discuss the identifiability of this model.

Example 4. This example builds on example 3 by introducing regressors. This produces the *censored regression model* (Heckman, 1976; 1979). In example 3 set

$$Y_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + U_1 \tag{1.16a}$$

$$Y_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + U_2 \tag{1.16b}$$

where $(\mathbf{X}_1, \mathbf{X}_2)$ are distributed independently of (U_1, U_2), a mean zero, finite variance random vector. Conventional assumptions are invoked to ensure that if Y_1 and Y_2 can be observed, least squares applied to a random sample of data on $(Y_1, Y_2, \mathbf{X}_1, \mathbf{X}_2)$ would consistently estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. $Y_1^* = Y_1$ if $Y_2 < 0$. If $Y_2 < 0$, $\Delta = 1$. Then the regression function for the selected sample is

$$\begin{aligned} E(Y_1^*|\mathbf{X}_1 = \mathbf{x}_1, Y_2 < 0) &= E(Y_1^*|\mathbf{X}_1 = \mathbf{x}_1, \Delta = 1) \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + E(U_1|\mathbf{X}_1 = \mathbf{x}_1, \Delta = 1) \end{aligned} \tag{1.17}$$

and the regression function for the population is

$$E(Y_1|\mathbf{X}_1 = \mathbf{x}_1) = \mathbf{X}_1\boldsymbol{\beta}_1. \tag{1.18}$$

As in the regression analysis of truncated random variables, there is an illuminating contrast between the conditional expectation for the selected sample (1.17) and the population regression function (1.18). The two functions differ by the conditional mean of $U_1[E(U_1|\mathbf{X}_1 = \mathbf{x}_1, \Delta = 1)]$. In the regression analysis of truncated random variables, ordinary least squares estimates of $\boldsymbol{\beta}$ (in equation (1.14)) are biased and inconsistent because the conditional mean is improperly omitted from the selected sample regression. The same analysis applies to the regression analysis of censored random variables. The conditional mean is a surrogate for the probability of selection $[\Pr(\Delta = 1|\mathbf{x}_2)]$. As $\Pr(\Delta = 1|\mathbf{x}_2)$ goes to one, the problem of sample selection bias becomes negligible. However, in the censored regression case, a new phenomenon appears. If there are variables in $\mathbf{X}_2$ not in $\mathbf{X}_1$, such variables may appear to be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

statistically important determinants of Y_1 when ordinary least squares is applied to data generated from censored samples.

As an example, suppose that survey statisticians use some extraneous (to $\mathbf{X}_1$) variables to determine sample enrolment. Such variables may appear to be important determinants of Y_1 when in fact they are not. They are important determinants of Y_1^*. In an analysis of self-selection, let Y_1 be the wage that a potential worker could earn were he to accept a market offer. Let Y_2 be the difference between the best non-market opportunity available to the potential worker and Y_1. If $Y_2 < 0$, the agent works. The conditional expectation of observed wages ($Y_1^* = Y$, if $Y_2 < 0$) given $\mathbf{x}_1$ and $\mathbf{x}_2$ will be a non-trivial function of $\mathbf{x}_2$. Thus variables determining non-market opportunities will determine Y_1^*, even though they do not determine Y_1. For example, the number of children less than six may appear to be significant determinants of Y_1 when inadequate account is taken of sample selection, even though the market does not place any value or penalty on small children in generating wage offers for potential workers.

Heckman (1976) develops the analysis of this model when (U_1, U_2) is normally distributed. Gallant and Nychka (1984) and Cosslett (1984) demonstrate that under mild restrictions on $F(u_1, u_2)$, if there is one continuous valued variable in $\mathbf{X}_2$ not in $\mathbf{X}_1$ (so that there is no exact linear dependence between $\mathbf{X}_2$ and $\mathbf{X}_1$), $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $F(u_1, u_2)$ can be consistently non-parametrically estimated. Heckman and MaCurdy (1986) develop this class of models at length.

Example 5. This example demonstrates how self-selection bias affects the interpretation placed on estimated consumer demand functions when there is self-selection. We postulate a population of consumers with a quasi-concave utility function $U(\mathbf{Z}, E)$ which depends on the consumption of goods and preference shock E which represents heterogeneity in preferences among consumers. The support of E is $\mathbf{E}$. For price vector $\mathbf{P}$ and endowment income M, the consumer's problem is to

$$\text{Max } U(\mathbf{Z}, E) \quad \text{subject to} \quad \mathbf{P}'\mathbf{Z} \leqslant M.$$

In the population $\mathbf{P}$ and M are distributed independently of E. First order conditions for this problem are

$$\frac{\partial U(\mathbf{Z}, E)}{\partial \mathbf{Z}} \leqslant \lambda \mathbf{P}, \tag{1.19}$$

where λ is the Lagrange multiplier associated with the budget constraint. Focusing on the demand for the first good, Z_1, none of it is purchased if at zero consumption of Z_1

$$\left.\frac{\partial U(\mathbf{Z}, E)}{\partial Z_1}\right|_{Z_1=0} \leqslant \lambda P_1 \tag{1.20}$$

i.e. marginal valuation is less than marginal cost in utility terms. Conventional interior solution demand functions for Z_1 are defined for a given $\mathbf{P}, M$ only for values of E such that

$$\left.\frac{\partial U(\mathbf{Z}, E)}{\partial Z_1}\right|_{Z_1=0} \geqslant \lambda P_1. \tag{1.21}$$

Let the set of E for which conventional interior solution consumer demand functions for Z_1 are defined be denoted by $\underline{\underline{\mathrm{E}}}$. Then

$$\underline{\underline{\mathrm{E}}} = \left\{ E \,\middle\|\, \left.\frac{\partial U(\mathbf{Z}, E)}{\partial Z_1}\right|_{Z_1=0} \geqslant \lambda P_1 \quad \text{for given} \quad \mathbf{P}, M \right\}.$$

Let $\Delta_1 = 0$ if the consumer does not purchase Z_1. Let $\Delta_1 = 1$ otherwise. If $F(\epsilon)$ is the population distribution of E, the proportion purchasing none of good Z_1 given $\mathbf{P}, M$ is

$$\Pr(\Delta_1 = 0 \mid \mathbf{P}, M) = 1 - \int_{\underline{\underline{\mathrm{E}}}} \mathrm{d}F(\epsilon).$$

Provided inequality (1.21) is satisfied, $\Delta_1 = 1$ and interior solution demand function

$$Z_1 = Z_1(\mathbf{P}, M, E) \tag{1.22}$$

is well defined and $Z_1 = Z_1^*$. When $\Delta_1 = 0$, observed $Z_1 = Z_1^* = 0$.

Equation (1.22) is the conventional object of interest in consumer theory. Partial derivatives of that function *holding E and the other arguments constant* have well defined economic interpretations. Suppose that some non-negligible proportion of the population buys none of good Z_1. Regression estimates of the parameters of (1.22) using Z_1^* approximate the conditional expectation

$$E(Z_1 \mid \Delta_1 = 1, \mathbf{P}, M) = \int_{\underline{\underline{\mathrm{E}}}} Z_1(\mathbf{P}, M, \epsilon)\, \mathrm{d}F(\epsilon). \tag{1.23}$$

The derivatives of (1.23) are different from the derivatives of (1.22). In order to define these derivatives, it is helpful to define $I_{\underline{\underline{\mathrm{E}}}}(E)$ as an indicator function for set $\underline{\underline{\mathrm{E}}}$ which equals one if $E \in \underline{\underline{\mathrm{E}}}$ and equals zero otherwise. When prices or income change, the set of values of E that satisfy inequality (1.21) changes. Let $\underline{\underline{\mathrm{E}}} + \Delta\underline{\underline{\mathrm{E}}}_{\mathbf{P}}$ be the set of E values that satisfy (1.21) when there is a finite price change $\Delta\mathbf{P}$. $I_{\underline{\underline{\mathrm{E}}}+\Delta\underline{\underline{\mathrm{E}}}_{\mathbf{P}}}(E)$ is an indicator function which equals one when $E \in \underline{\underline{\mathrm{E}}} + \Delta\underline{\underline{\mathrm{E}}}_{\mathbf{P}}$. Then the derivatives of (1.23) are, for the jth price

$$\frac{\partial E(Z_1 \mid \Delta = 1, \mathbf{P}, M)}{\partial P_j} = \int_{\underline{\underline{\mathrm{E}}}} \frac{\partial Z_1(\mathbf{P}, M, \epsilon)}{\partial P_j} \mathrm{d}F(\epsilon) + \lim_{\Delta P_j \to 0} \int_{\underline{\underline{\mathrm{E}}}} \frac{[(I_{\underline{\underline{\mathrm{E}}}+\Delta\underline{\underline{\mathrm{E}}}_{P_j}}(\epsilon) - I_{\underline{\underline{\mathrm{E}}}}(\epsilon)]Z(\mathbf{P}, M, \epsilon)}{\Delta P_j} \mathrm{d}F(\epsilon). \tag{1.24}$$

When the limit in the second term does not exist, the derivative does not exist. We assume for expositional convenience that the limit is well defined.

The first expression on the right-hand side of (1.24) is the *average effect* of price change on commodity demand. The second term on the right-hand side of (1.24) arises from the change in sample composition of E as the proportion of non-purchasers changes in response to price change. This term generates the selection bias.

Neither term is the same as the price derivative of (1.22) for an arbitrary value of $E = \epsilon$ although the first term on the right-hand side of (1.24) approximates the price derivative of (1.22) for some value of $E = \epsilon$.

A similar decomposition of the derivatives of the conditional demand function can be performed if it is defined solely for a sample of non-zero purchasers (see Heckman and MaCurdy, 1981, 1986).

Just as in the statistical sample selection bias problem, there is a population of interest. In this case, the population parameters of interest are the distribution of E and the parameters of $U(\mathbf{Z}, E)$. Those who buy Z_1 are a self-selected sample of the population. Estimates of population parameters estimated on self-selected samples are biased and inconsistent. There is a population distribution of $Z_1(\mathbf{P}, M, E)$ generated by the distribution of E. Observations of Z_1 are obtained only if $E \in \underline{\underline{\mathrm{E}}}(\omega(E) = 1$ if $E \in \underline{\underline{\mathrm{E}}}, \omega(E) = 0$ otherwise). Alternatively one can express the inclusion criteria in terms of the latent population distribution of Z_1 induced by E (given $\mathbf{P}$ and M) *and write* $\omega(z_1) = 1$ if $z_1 > 0, \omega(z_1) = 0$ if $z_1 \leqslant 0$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Heckman (1974) and Heckman and MaCurdy (1981) provide further discussion of this type of model which is widely used in applied economics and consider issues of identifiability for such models.

Example 6. Length biased sampling. Let T be the duration of an event such as a completed unemployment spell or a completed duration of a job with an employer. The population distribution of T is $F(t)$ with density $f(t)$. The sampling rule is such that *individuals* are sampled at random. Data are recorded on a completed spell *provided that at the time of the interview the individual is experiencing the event*. Such sampling rules are in wide use in many national surveys of employment and unemployment.

In order to have a sampled completed spell, a person must be in the state at the time of the interview. Let '0' be the date of the survey. Decompose any completed spell T into a component that occurs before the survey T_b and a component that occurs after the survey T_a. Then $T = T_a + T_b$. For a person to be sampled, $T_b > 0$. The density of T given $T_b = t_b$ is

$$f(t \mid t_b) = \frac{f(t)}{1 - F(t_b)}, \quad t \geqslant t_b. \tag{1.25}$$

Suppose that the environment is stationary. The population entry rate into the state at each instant of time is k. From each vintage of entrants into the state distinguished by their distance from the survey date t_b, only $1 - F(t_b) = \Pr(T > t_b)$ survive. Aggregating over all cohorts of entrants, the population proportion in the state at the date of the interview is P where

$$P = \int_0^\infty k(1 - F(t_b))\,\mathrm{d}t_b \tag{1.26}$$

which is assumed to exist. The density of T_b^*, sampled pre-survey duration, is

$$g(t_b^* \mid t_b^* > 0) = \frac{k(1 - F(t_b^*))}{P}. \tag{1.27}$$

The density of sampled completed durations is thus

$$\begin{aligned} g(t^*) &= \int_0^{t^*} f(t^* \mid t_b^*) g(t_b^* \mid t_b^* > 0)\,\mathrm{d}t_b^* \\ &= k \frac{f(t^*)}{1 - F(t_b^*)} \frac{1 - F(t_b^*)}{P} \int_0^{t^*} \mathrm{d}t_b^* \\ &= k \frac{t^* f(t^*)}{P}. \end{aligned}$$

Observe from (1.26) that by a standard integration by parts argument

$$P = k \int_0^\infty (1 - F(z))\,\mathrm{d}z = k \int_0^\infty z\,\mathrm{d}F(z) = kE(T).$$

Note that

$$g(t^*) = \frac{t^* f(t^*)}{E(T)}. \tag{1.28}$$

In this form (1.28) is equivalent to (1.1) with $\omega(t) = t$. Hence the term 'length biased sampling'. Intuitively, longer spells are oversampled when the requirement is imposed that a spell be in progress at the time the survey is conducted ($T_b > 0$). Suppose, instead, that individuals are randomly sampled and data are recorded on the *next* spell of the event (after the survey date). As long as successive spells are independent, such a sampling frame does not distort the sampled distribution because no requirement is imposed that the sampled spell be in progress at the date of the interview. It is important to notice that the source of the bias is the requirement that $T_b > 0$, not that only a fraction of the population experiences the event ($P < 1$).

The simple length weight ($\omega(t) = t$) that produces (1.28) is an artefact of the stationarity assumption. Heckman and Singer (1985) consider the consequences of non-stationarity and unobservables when there is selection on the event that a person be in the state at the time of the inverview. The also demonstrate the bias that results from estimating parametric models on samples generated by length biased sampling rules when inadequate account is taken of the sampling plan. Vardi (1983, 1985) and Gill and Wellner (1985) consider nonparametric identification and estimation of models with densities of the form (1.28).

It is unfortunate that the lessons of length biased sampling are not adequately appreciated in economics. Two widely cited studies by Clark and Summers (1979) and Hall (1982) use length biased data to prove, respectively, that unemployment and employment spells are 'surprisingly long'. Whether their findings are artefacts of sampling plans remains to be determined.

Example 7. Choice based sampling. Let D be a discrete valued random variable which assumes a finite number of values I. $D = i, i = 1, \ldots, I$ corresponds to the occurrence of state i. States are mutually exclusive. In the literature the states may be modes of transportation choice for communters (Domencich and McFadden, 1975), occupations, migration destinations, financial solvency status of firms, schooling choices of students, etc. Interest centres on estimating a population choice model

$$\Pr(D = i \mid \mathbf{X} = \mathbf{x}), \quad i = 1, \ldots, I. \tag{1.29}$$

The population density of $(D, \mathbf{X})$ is

$$f(d, \mathbf{x}) = \Pr(D = d \mid \mathbf{X} = \mathbf{x}) h(\mathbf{x}) \tag{1.30}$$

where $h(\mathbf{x})$ is the density of the data.

In many problems, plentiful data are available on certain outcomes while data are scarce for other outcomes. For example, interviews about transportation preferences conducted at train stations tend to over-sample train riders and under-sample bus riders. Interviews about occupational choice preferences conducted at leading universities over-sample those who select professional occupations.

In choice based sampling, selection occurs solely on the D coordinate of $(D, \mathbf{X})$. In terms of (1.1) (extended to allow for discrete random variables), $\omega(d, \mathbf{X}) = \omega(d)$. Then sampled $(D^*, \mathbf{X}^*)$ has density

$$g(d^*, \mathbf{x}^*) = \frac{\omega(d^*) f(d^*, \mathbf{x}^*)}{\sum_{i=1}^{I} \int \omega(i) f(i, x^*)\,\mathrm{d}x^*} \tag{1.31}$$

Notice that the denominator can be simplified to

$$\sum_{i=1}^{I} \omega(i) f(i)$$

where $f(d^*)$ is the marginal distribution of D^* so that

$$g(d^*, \mathbf{x}^*) = \frac{\omega(d^*) f(d^*, \mathbf{x}^*)}{\sum_{i=1}^{I} \omega(i) f(i)} \tag{1.32}$$

Also, integrating (1.31) with respect to $\mathbf{x}$ using (1.32) we obtain

$$g(d^*) = \frac{\omega(d^*) f(d^*)}{\sum_{i=1}^{I} \omega(i) f(i)} \tag{1.33}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which makes transparent how the sampling rule causes the sampled proportions to deviate from the population proportions. Note further that as a consequence of sampling only on D, the population conditional density

$$h(\mathbf{x}^*|d^*) = \frac{f(d^*, x^*)}{f(d^*)} \tag{1.34}$$

can be recovered from the choice based sample. The density of $\mathbf{x}$ in the sample is thus

$$g(\mathbf{x}^*) = \sum_{i=1}^{I} h(x^*|i)g(i). \tag{1.35}$$

Then using (1.32)–(1.35) we reach

$$g(d^*|\mathbf{x}^*) = f(d^*|\mathbf{x}^*) \left\{ \left[\frac{\omega(d^*)}{\sum_{i=1}^{I} \omega(i)f(i)} \right] \left[\frac{\sum_{i=1}^{I} f(i|\mathbf{x}^*)\frac{g(i)}{f(i)}}{} \right] \right\} \tag{1.36}$$

The bias that results from using choice based samples to make inference about $f(d^*|x^*)$ is a consequence of neglecting the terms in braces on the right-hand side of (1.36). Notice that if the data are generated by a random sampling rule, $\omega(d^*) = 1, g(d^*) = f(d^*)$ and the term in braces is one.

Manski and Lerman (1977), Manski and McFadden (1981) and Cosslett (1981) provide illuminating discussions of choice based sampling.

Example 8. Size biased sampling. Let N be the number of children in a family. $f(N)$ is the density of discrete random variable N. Suppose that family size is recorded only when at least one child is interviewed. Suppose further that each child has an independent and identical chance β of being interviewed. The probability of sampled family size of $N^* = n^*$ is

$$g(n^*) = \frac{\omega(n^*)f(n^*)}{E[\omega(N^*)]} \tag{1.37}$$

where $\omega(n^*) = 1 - (1-\beta)^{n^*}$ (the probability that at least one child from a family of size n^* will be sampled) and

$$E[\omega(N^*)] = \sum_{n^*} (1 - (1-\beta)^{n^*}) f(n^*)$$

is the probability of observing a family. In a large population $\beta \to 0$ with increasing population size. Using l'Hospital's rule, and assuming that passage to the limit under the summation sign is valid

$$\lim_{\beta \to 0} g(n^*) = \frac{n^* f(n^*)}{E(N^*)} \tag{1.38}$$

Thus the limit form of (1.37) is identical to (1.28). Larger families tend to be oversampled and hence a misleading estimate of family size will be produced from such samples. Since the model is formally equivalent to the length biased sampling model, all references and statements about identification given in example 6 apply with full force to this example. See the discussion in Rao (1965).

2. ECONOMIC MODELS OF SELF-SELECTION

We begin our analysis by expositing the Roy model of self-selection for workers with heterogeneous skills. The statistical framework for this model has been outlined in examples 3 and 4. Following Roy, we assume that there are two market sectors in which income-maximizing agents can work. Agents are free to enter the sector that gives them the highest income. However, they can work in only one sector at a time.

Each sector requires a unique sector-specific task. Each agent has two skills, T_1 and T_2 which he cannot use simultaneously. The model is short run in that aggregate skill distributions are assumed to be given. There are no costs of changing sectors, and investment is ignored. Because of this assumption, the model presented here applies to environments with certain or uncertain prices for sector-specific tasks. For simplicity and without any loss of generality (given the preceding assumptions), we assume an environment of perfect certainty.

Let T_i be the amount of sector i specific task a worker can perform. The price of task i is π_i. An agent works in sector 1 if his income is higher there, that is

$$\pi_1 T_1 > \pi_2 T_2 \tag{2.1}$$

Indifference between sectors is a negligible probability event if the $T_i = 1, 2$ are assumed to be continuous nondegenerate random variables. Throughout we assume that prices are positive ($\pi_i > 0$).

The log wage in task i of an individual with endowment T_i is

$$\ln W_i = \ln \pi_i + \ln \; T_i \tag{2.2}$$

The proportion of the population working at task i is the proportion of the population for whom

$$T_1 > \frac{\pi_2}{\pi_1} T_2.$$

Roy assumes that $(\ln T_1, \ln T_2)$ is normally distributed with mean (μ_1, μ_2) and covariance matrix Σ. Letting (U_1, U_2) be a mean zero normal vector, agents in the Roy model choose between two possible wages:

$$\ln W_1 = \ln \pi_1 + \mu_1 + U_1$$

or

$$\ln W_2 = \ln \pi_2 + \mu_2 + U_2.$$

Workers enter sector 1 if $\ln W_1 > \ln W_2$. Otherwise they enter sector 2.

Letting

$$\sigma^* = \sqrt{\text{var}(U_1 - U_2)}$$

and

$$c_i = (\ln(\pi_i/\pi_j) + \mu_i - \mu_j)/\sigma^*, \quad i \neq j,$$

$$\Pr(i) = P(\ln W_i > \ln W_j) = \Phi(c_i), \quad i \neq j, \quad i, j = 1, 2$$

where $\Phi(\;)$ is the cumulative distribution function of a standard normal variable. When standard sample selection bias formulae are used (see, e.g. Heckman 1976), the mean of log wages observed in sector i is

$$E(\ln W_i | \ln W_i > \ln W_j) = \ln \pi_i + \mu_i + \frac{\sigma_{ii} - \sigma_{ij}}{\sigma^*} \lambda(c_i), \quad i, j = 1, 2, \quad i \neq j, \tag{2.3}$$

where

$$\lambda(c) = \frac{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}c^2)}{\Phi(c)}$$

is a convex monotone decreasing function of c with $\lambda(c) \geqslant 0$, and

$$\lim_{c \to \infty} \lambda(c) = 0, \quad \lim_{c \to -\infty} \lambda(c) = \infty.$$

Convexity is proved in Heckman and Honoré (1986).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The variance of log wages observed in sector i

$$\mathrm{var}(\ln W_i | \ln W_i > \ln W_j) = \sigma_{ii}\{\rho_i^2[1 - c_i\lambda(c_i) - \lambda^2(c_i)] + (1-\rho_i^2)\}, \quad i \neq j \qquad (2.4)$$

where $\rho_i = \mathrm{correl}(U_i, U_i - U_j)$, $i \neq j = 1, 2$. The variance of the log of observed wages never exceeds σ_{ii}, the population variance, because the term in braces in (2.4) is never greater than unity. In general, sectoral variances decrease with increased selection. For example, if ρ_1 and ρ_2 do not equal zero, as π_1 increases with π_2 held fixed so that people shift from sector 2 to sector 1, the variance in the log of wages in sector 1 increases while the variance in the log of wages in sector 2 decreases.

Using the fact that $W_i = \pi_i T_i$, we may use (2.3) to write

$$E(\ln T_1 | \ln W_1 > \ln W_2) = \mu_1 + \frac{\sigma_{11} - \sigma_{12}}{\sigma^*}\lambda(c_1), \qquad (2.5a)$$

$$E(\ln T_2 | \ln W_1) > \ln W_2) = \mu_2 + \frac{\sigma_{22} - \sigma_{12}}{\sigma^*}\lambda(c_2). \qquad (2.5b)$$

Focusing on (2.5a) and noting that λ is positive for all values of c_1 (except $c_1 = \infty$), the mean of log task 1 used in sector 1 exceeds, equal, or falls short of the population mean endowment of log task 1 as $\sigma_{11} - \sigma_{12}$ is greater than, equal to, or less than zero. If endowments of tasks are uncorrelated ($\sigma_{12} = 0$), self-selection always causes the mean of $\ln T_1$ employed in sector 1 to be above the population mean μ_1. The opposite case occurs when $\sigma_{11} - \sigma_{12}$ is negative. This case can arise only when values of $\ln T_1$ and $\ln T_2$ are sufficiently positively correlated. If this occurs, the mean of log task 1 used in sector 1 falls below the population mean μ_1. Since covariance matrices must be positive semidefinite, $\sigma_{11} + \sigma_{22} - 2\sigma_{12} \geqslant 0$. Thus if $\sigma_{11} - \sigma_{12} < 0$, $\sigma_{22} - \sigma_{12} > 0$ so the mean of log task 2 employed in sector 2 necessarily lies above the population mean μ_2. In the Roy model the unusual case can arise in at most one sector. Notice from (2.5) that only if $\sigma_{11} - \sigma_{12} = 0$ (so $\rho_1^2 = 0$) is the variance of log task 1 employed in sector 1 identical to the variance of log task 1 in the population. Otherwise, the sectoral variance of observed log task 1 is less than the population variance of log task 1.

To gain further insight into the effect of self-selection on the distribution of earnings for workers in sector 1, it is helpful to draw on some results from normal regression theory. The regression equation for $\ln T_2$ conditional on $\ln T_1$ is

$$\ln T_2 = \mu_2 + \frac{\sigma_{12}}{\sigma_{12}}(\ln T_1 - \mu_1) + \epsilon_2, \qquad (2.6)$$

where $E(\epsilon_2) = 0$ and $\mathrm{var}(\epsilon_2) = \sigma_{22}[1 - (\sigma_{12}^2/\sigma_{11}\sigma_{22})]$.

Figure 2 plots regression function (2.6) for the case $\sigma_{12} = \sigma_{11}$ and $\mu_2 > \mu_1 > 0$. For each value of $\ln T_1$, the population values of $\ln T_2$ are normally distributed around the regression line. Individuals with high values of $\ln T_1$ also tend to have a high value of $\ln T_2$. Assuming $\pi_1 = \pi_2$, individuals with $(\ln T_1, \ln T_2)$ endowments above the 45° line of equal income shown in Figure 1 choose to work in sector 2, while those individuals with endowments below this line work in sector 1. Because $\sigma_{12} = \sigma_{11}$, the regression function is parallel to the line of equal income.

The distribution of ϵ_2 about the regression line is the same for all values of $\ln T_1$. When individuals are classified on the basis of their $\ln T_1$ values the same proportion of individuals work in sector 1 at all values of $\ln T_1$. For this reason the distribution of $\ln T_1$ employed in sector 1 is the same as the latent population distribution. If π_1 is raised (or π_2 is lowered) so that the 45° equal income line is shifted upward, the same proportion of people enter sector 1 at each value of $T_1 = t_1$. Figure 3 plots regression function (2.6) for the case $\sigma_{12} > \sigma_{11}$ and $\mu_2 > \mu_1 > 0$.

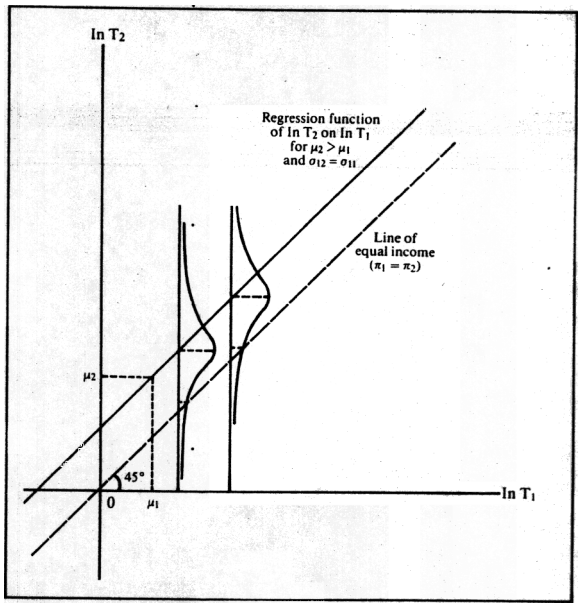

Figure 2

As before we set $\pi_1 = \pi_2$. Individuals with endowments above the 45° line choose to work in sector 2, while those with endowments below this line work in sector 1. When individuals are classified on the basis of their T_1 values, the fraction of people working in sector 1 decreases the higher the value of T_1. Self-selection causes the mean of log task 1 employed in sector 1 to be less than the mean of log task 1 in the total population. People with high values of T_1 are under-represented in sector 1 and low T_1 values are over-represented. In the extreme, when $\ln T_1$ and $\ln T_2$ are perfectly positively correlated, all high-income individuals are in sector 2, while all the low-income individuals are in sector 1. The highest-paid sector 1 worker earns the same as the lowest-paid sector 2 worker (Roy, 1951; Willis and Rosen, 1979). In this case there is really only one skill dimension and individuals can be unambiguously ranked along this scale.

If π_1 is raised (or π_2 is lowered) so that the line of equal income is shifted upward, the mean of $\ln T_1$ employed in sector 1 must rise. The only place left to get T_1 is from the high end of the T_1 distribution. Unlike the case of $\sigma_{12} = \sigma_{11}$, in which a 10 per cent increase in π_1 results in a 10 per cent increase in measured average earnings in sector 1, when $\sigma_{12} > \sigma_{11}$, a 10 per cent increase in π_1 results in a greater than 10 per cent increase in the measured average earnings in sector 1 as the average quality of the sector 1 work-force increases. The variance of log wages in sector 1 increases.

If $\sigma_{11} < \sigma_{12}$, than $\sigma_{12} < \sigma_{22}$ in order for Σ to be a covariance matrix. In the population, log task 2 must have greater variability than log task 1. Individuals with high T_1 values tend to have high T_2 values. But the population distribution of log task 2 has more mass in the tails. The higher an agent's value of T_1, the more likely it is that he will be able to get higher income in sector 2. At the lower end of the distribution, the process works in reverse: lower T_1 individuals on average have poor T_2 values. Self-selection causes the $\ln T_1$ distribution in sector 1 to have an evacuated right tail, an exaggerate left tail, and a lower mean than the population mean of $\ln T_1$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

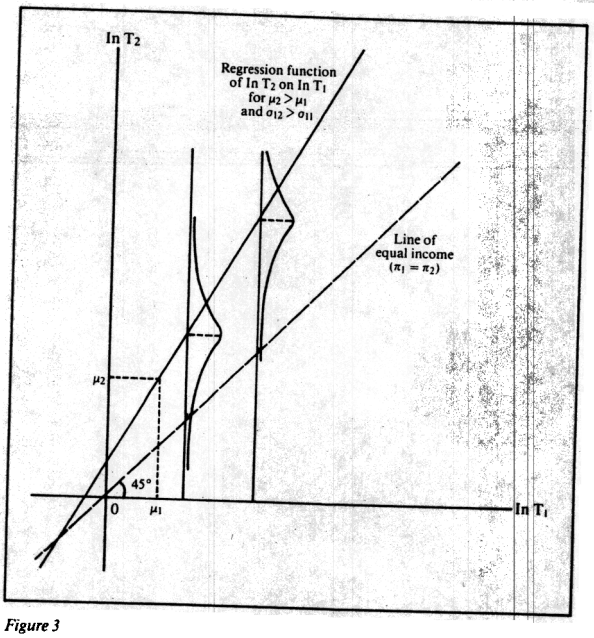

Figure 3

If $\sigma_{12} < \sigma_{11}$ (a case not depicted graphically), the proportion of each T_1 group working in sector 1 increases, the higher the value of T_1. The mean of the log task employed in sector 1 exceeds μ_1. A 10 per cent increase in π_1 produces an increase of less than 10 per cent in the average earnings of workers in sector 1 as the mean of $\ln T_1$ employed in sector 1 declines. In fact if $\sigma_{12} > \sigma_{22}$ it is possible for an increase in π_1 to cause measured sector 1 wages to decline. Thus through a selection phenomenon it is possible for the average wage of people working in sector 1 to decline even though the price per unit skill increases there.

How robust are these conclusions if the normality assumption is relaxed? Heckman and Sedlacek (1985) show that many propositions derived from assumed normality of skills do not hold up for more general distributions. For example, increasing selection need not decrease sectoral variances. The effects of selection on mean employed skill levels are ambiguous. Heckman and Honoré (1986) demonstrate that in a single cross-section of data, it is possible to identify all of the parameters of the model from the data if the normality asumption is invoked. However, in a single cross-section many other models can explain the data equally well. In particular, intuitive notions about the degree of correlation or dependence among skills have no empirical content and so models of skill 'hierarchies' based on the extent of such dependence have no content for single cross-sections of data with all individuals facing common prices.

To show this, write the density of skills as $f(t_1, t_2)$. Let

$$Z = \begin{cases} T_1 & \text{if} \quad T_1 > T_2 \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} T_2 & \text{if} \quad T_2 > T_1 \\ 0 & \text{otherwise} \end{cases}$$

Prices are normalized to unity ($\pi_1 = \pi_2 = 1$). Then the density of

$$Q_1'(z_1) = \int_{\{t_2 | t_2 < z_1\}} f(z_1, t_2)\, dt_2$$
$$= \int_0^{z_1} f(z_1, t_2)\, dt_2$$

The density of z_2 is

$$Q_2'(z_2) = \int_0^{z_2} f(t_1, z_2)\, dt_1.$$

Note that $Q_1'(n)$ and $Q_2'(n)$ summarize all of the available data on observed earnings.

Now if T_1, T_2 are independent with cdf's F_1^* and F_2^* respectively

$$Q_1'(n) = f_1^*(n) F_2^*(n)$$
$$Q_2'(n) = F_1^*(n) f_2^*(n).$$

Define

$$\bar{Q}(n) = \int_0^n [Q_1'(l) + Q_2'(l)]\, dl$$
$$= F_1^*(n) F_2^*(n).$$

$$\int_\phi^\infty \frac{Q_1'(n)}{\bar{Q}(n)}\, dn = \int_\phi^\infty \frac{f_1^*(n)}{F_1^*(n)}\, dn$$
$$= -\ln F_1^*(\phi).$$

Thus we can write

$$F_i^*(\phi) = \exp - \left(\int_\phi^\infty \left[\frac{Q_i'(n)}{\bar{Q}(n)} \right] dn \right) \qquad i = 1, 2$$

so that we can always rationalize the data on wages in a single cross-section by a model of skill independence, and economic models of skill hierarchies have no empirical content for a single cross-section of data.

Suppose, however, that the observing economist has access to data on skill distributions in different market settings i.e. settings in which relative skill prices vary. To take an extreme case, suppose that we observe a continuum of values of π_1/π_2 ranging from zero to infinity. Then it is possible to identify $F(t_1, t_2)$ and it is possible to give empirical content to models based on the degrees of dependence among latent skills.

This point is made most simply in a situation in which Z is observed but the analyst does not know Z_1 or Z_2 (i.e. which occupation is chosen). When $\pi_1/\pi_2 = 0$, everyone works in occupation two. Thus we can observe the marginal density of t_2. When $\pi_1/\pi_2 = \infty$, everyone works in occupation one. As π_1/π_2 pivots from zero to infinity it is thus possible to trace out the full joint distribution of (T_1, T_2).

To establish the general result, set $\sigma = \pi_2/\pi_1$. Let $F(t_1, t_2)$ be the distribution function of T_1, T_2. Then

$$\Pr(Z \leqslant n) = \Pr(\max(T_1, \sigma T_2) \leqslant n)$$
$$= \Pr\left(T_1 \leqslant n, T_2 \leqslant \frac{1}{\sigma} n \right)$$
$$= F\left(n, \frac{n}{\sigma} \right).$$

As σ varies between 0 and ∞, the entire distribution can be recovered since N is observed for all values in $(0, \infty)$. Note that it is not necessary to know which sector the agent selects.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This proposition establishes the benefit of having access to data from more than one market. Heckman and Honoré (1986) show how access to data from various market settings and information about the choices of agents aids in the identification of the latent skill distributions.

The Roy model is the prototype for many models of self-selection in economics. If T_1 is potential market productivity and T_2 is non-market productivity (or the reservation wage) for housewives or unemployed individuals, precisely the same model can be used to explore the effects of self-selection on measured productivity. In such a model, T_2 is never observed. This creates certain problems of identification discussed in Heckman and Honoré (1986). The model has been extended to allow for more general choice mechanisms. In particular, selection may occur as a function of variables other than or in addition to T_1 and T_2. Applications of the Roy model include studies of the union–non-union wage differential (Lee, 1978), the returns to schooling (Willis and Rosen, 1979), and the returns to training (Bjorklund and Moffitt, 1986) and Heckman and Robb (1985). Amemiya (1984) and Heckman and Honoré (1986) present comprehensive surveys of empirical studies based on the Roy model and its extensions.

JAMES J. HECKMAN

BIBLIOGRAPHY

Amemiya, T. 1984. Tobit models: a survey. *Journal of Econometrics* 24, 3–61.

Bjorklund, A. and Moffitt, R. 1986. Estimation of wage gains and welfare gains from self selection models. *Review of Economics and Statistics* 24, 1–63.

Cain, G. and Watts, H. 1973. Toward a summary and synthesis of the evidence. In *Income Maintenance and Labor Supply*, ed. G. Cain and H. Watts, Madison: University of Wisconsin Press.

Clark, K. and Summers, L. 1979. Labor market dynamics and unemployment: a reconsideration. *Brookings Papers on Economic Activity*, 13–60.

Cosslett, S. 1981. Maximum likelihood estimation from choice based samples. *Econometrica*.

Cosslett, S. 1984. Distribution free estimator of regression model with sample selectivity. Unpublished manuscript, University of Florida.

Domencich, T. and McFadden, D. 1975. *Urban Travel Demand*. Amsterdam: North-Holland.

Flinn, C. and Heckman, J. 1982. New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics* 18, 5–168.

Gallant, R. and Nychka, R. 1984. Consistent estimation of the censored regression model. Unpublished manuscript, North Carolina State.

Gill, R. and Wellner, J. 1985. Large sample theory of empirical distributions in biased sampling models. Unpublished manuscript, University of Washington.

Goldberger, A. 1983. Abnormal selection bias. In *Studies in Econometrics, Time Series and Multivariate Statistics*, ed. S. Karlin, T. Amemiya and L. Goodman, Wiley, NY

Gronau, R. 1974. Wage comparisons-a selectivity bias. *Journal of Political Economy* 82, (6), 1119–1144.

Hall, R. 1982. The importance of lifetime jobs in the U.S. economy. *American Economic Review* 72, September, 716–724.

Heckman, J. 1974. Shadow prices, market wages and labor supply. *Econometrica* 42(4), 679–94.

Heckman, J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5(4), 475–92.

Heckman, J. 1977. Sample selection bias as a specification error. *Econometrica* 47(1), 153–62.

Heckman, J. and Honoré, B. 1986. The empirical content of the Roy model. Unpublished manuscript, University of Chicago.

Heckman, J. and MaCurdy, T. 1981. New methods for estimating labor supply functions. In *Research in Labor Economics*, Vol. 4, ed. R. Ehrenberg, Greenwich, Conn.: JAI Press.

Heckman, J. and Robb, R. 1985. Alternative methods for evaluating the effect of training on earnings. In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Heckman, J. and Sedlacek, G. 1985. Heterogeneity, aggregation and market wage functions. *Journal of Political Economy* 93, December, 1077–125.

Heckman, J. and Singer, B. 1985. Econometric analysis of longitudinal data. In *Handbook of Econometrics*, Vol. III, ed. Z. Griliches and M. Intriligator, Amsterdam: North-Holland.

Lee, L. F. 1978. Unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables. *International Economic Review* 19, 415–33.

Manski, C. and Lerman, S. 1977. The estimation of choice probabilities from choice based samples. *Econometrica* 45, 1977–88.

Manski, C. and McFadden, D. 1981. Alternative estimates and sample designs for discrete choice analysis. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. Manski and D. McFadden, Cambridge: MIT Press.

Pearson, K. 1901. Mathematical contributions to the theory of evolution. *Philosophical Transactions*, 195, 1–47.

Rao, C. R. 1965. On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Distributions*, ed. G. Patil, Calcutta: Pergamon Press.

Roy, A. D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3, 135–46.

Vardi, Y. 1983. Nonparametric estimation in the presence of length bias. *Annals of Statistics* 10, 616–20.

Vardi, Y. 1985. Empirical distributions in selection bias models. *Annals of Statistics*, 13, 178–203.

Willis, R. and Rosen, S. 1979. Education and self selection. *Journal of Political Economy* 87, S7-S36.

self-interest. Two of the basic questions with which moral philosophers have been concerned are: (a) what are the fundamental principles of morality? (b) why should we obey them? One tempting answer to the second question is: because obeying them is in your own interest. Tempting, because any other answer simply invites a further 'why?'. For example, 'why bother about helping others to get what they want?' clearly demands an answer. But 'why bother about getting what *you* want?', though of course it can be *asked*, hardly makes sense.

Self-interest as the answer to the second question, however, implies a similar answer to the first. Self-interest can only be a reason for obeying moral principles if those principles do always benefit us as individuals, so that the fundamental one becomes: Do whatever will enable you to satisfy your own desires. And this seems perverse, since most moralists tell us to consider others rather than ourselves. Self-sacrifice, we are told, is noble, and self-seeking base.

Thomas Hobbes answers this objection by pointing out that, while human desires are diverse, so that there is no common end, there is a single means common to all ends. They all require the cooperation of other people, or at least their non-interference. Everyone has an interest in maintaining a peaceful and harmonious society. Moral principles are simply the rules which everyone must follow in order to obtain such a society. We should obey them because obeying them makes for peace and security, and without peace and security no one has much chance of satisfying *any* desires. If morality requires us to consider others and not ourselves, it is for our own sakes in the long run.

To suppose that men imposed moral restraints on themselves for this reason might suggest a far-sightedness greater than most of us are capable of. Bernard Mandeville suggested that men are motivated less by this consideration than by vanity. Morality, he conjectured, came about through the artifice of a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

self-confirming equilibria

In-Koo Cho and Thomas J. Sargent
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Self-confirming equilibria are limiting outcomes of purposeful interactions among a collection of adaptive agents, each of whom averages past data to approximate moments of conditional probability distributions. Self-confirming equilibria are powerful tools to investigate dynamic economic problems such as the limiting behaviour of learning systems, the selection of plausible equilibria in games and dynamic macroeconomic models, the incidence and distribution of rare events that occasionally arise as large deviations from self-confirming equilibria, and how agents respond to model uncertainty.

Keywords

adaptive learning; law of large numbers; learning rules; ordinary differential equations; perfect foresight; rational expectations equilibrium; recursive learning algorithms; robust control; self-confirming equilibrium; stochastic approximation algorithms

Article

A self-confirming equilibrium is the answer to the following question: what are the possible limiting outcomes of purposeful interactions among a collection of adaptive agents, each of whom averages past data to approximate moments of the conditional probability distributions of interest? If outcomes converge, a law of large numbers implies that agents' beliefs about conditional moments become correct on events that are observed sufficiently often. Beliefs are not necessarily correct about events that are infrequently observed. Where beliefs are correct, a self-confirming equilibrium is like a rational expectations equilibrium. But there can be interesting gaps between self-confirming and rational expectations equilibria where beliefs of some important decision makers are incorrect.
Self-confirming equilibria interest macroeconomists because they connect to an influential 1970s argument made by Christopher Sims that advocated rational expectations as a sensible equilibrium concept. This argument defended rational expectations equilibria against the criticism that they require

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that agents ‘know too much’ by showing that we do not have to assume that agents start out ‘knowing the model’. If agents simply average past data, perhaps conditioning by grouping observations, their forecasts eventually become unimprovable.

Research on adaptive learning has shown that that the glass is ‘half full’ and ‘half empty’ for this clever 1970s argument. On the one hand, the argument is correct when applied to competitive or infinitesimal agents: by using naive adaptive learning schemes (various versions of recursive least squares), agents can learn every conditional distribution that they require to play best responses within an equilibrium. On the other hand, large agents (for example, governments in macro models) who can influence the market outcome cannot expect to learn everything that they need to know to make good decisions: in a self-confirming equilibrium, large agents may base their decisions on conjectures about off-equilibrium-path behaviours that turn out to be incorrect. Thus, a rational expectations equilibrium is a self-confirming equilibrium, but not vice versa.

While agents’ beliefs can be incorrect off the equilibrium path, the self-confirming equilibrium path still restricts them in interesting ways. For macroeconomic applications, the government's model must be such that its off-equilibrium path beliefs rationalize the decisions (its Ramsey policy or Phelps policy, in the language of Sargent, 1999) that are revealed along the equilibrium path. The restrictions on government beliefs required to sustain self-confirming equilibria have only begun to be explored in macroeconomics, mainly in the context of some examples like those in Sargent (1999). Analogous restrictions have been more thoroughly analysed in the context of games (Fudenberg and Levine, 1993).

The freedom to specify beliefs off the equilibrium path makes the set of self-confirming equilibria generally be larger than the set of Nash equilibria, which often admit unintuitive outcomes in extensive form games (Kreps, 1998). A widely used idea of refining self-confirming equilibrium is to embed the decision making problem within a learning process in which decision makers estimate unknown parameters through repeated interactions, and then to identify a stable stationary point of the learning dynamics (for example, Fudenberg and Kreps, 1993; 1995; Evans and Honkapohja, 2001).

The gap between a self-confirming equilibrium and a rational expectations equilibrium can be vital for a government designing a Ramsey plan, for example, because its calculations necessarily involve projecting outcomes of counterfactual experiments. For macroeconomists, an especially interesting feature of self-confirming equilibria is that, because a government can have a model that is wrong off the equilibrium path, a policy that it thinks is optimal can very well be far from optimal. Even if a policy model fits the historical data correctly and is unimprovable, one cannot conclude that the policy is optimal. As a result, it requires an entirely a priori theoretical argument to diminish the influence of a good fitting macroeconomic model on public policy (Sargent, 1999).

Formal definitions

An agent i is endowed with strategy space A_i and state space X_i. Generic elements of A_i and X_i are called a strategy and a state, respectively. A probability distribution P_i over $A_i \times X_i$ describes how actions and states are related. A utility function is $u_i: A_i \times X_i \to \mathbb{R}$. Let $\mu_i(\cdot: a_i)$ be a probability distribution over X_i, which represents i's belief about the state conditioned on action a_i. Agent i's decision problem is to solve

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\max_{a_i \in A_i} \int_{X_i} u_i(a_i, x_i)\, d\mu_i(x_i : a_i).$$
(1)

Single person decision problems

Here a self-confirming equilibrium is a simply a pair (a_i^*, μ_i^*) satisfying

$$a_i^* \in \arg\max_{a_i \in A_i} \int_{X_i} u_i(a_i, x_i)\, d\mu_i^*(x_i : a_i)$$
(2)

$$\mu_i^*(x_i : a_i^*) = \mathcal{P}_i(x_i : a_i^*).$$
(3)

(2) implies that the choice must be optimal given his subjective belief μ_i^*, while (3) says that the belief must be confirmed, conditioned on his equilibrium action a_i^*. Self-confirming equilibrium has the two key ingredients of rational expectations equilibrium: optimization and self-fulfilling property. The key difference is (3), which imposes a self-confirming property conditioned only on equilibrium action a_i^*. The decision maker can entertain $\mu_i(\cdot : a_i) \neq \mathcal{P}(\cdot : a_i)$, conditioned on $a_i \neq a_i^*$. In this sense, the agent can have multiple beliefs about the state conditioned on his own action.
If we strengthen things to require

$$\mu_i(\cdot : a_i) = \mathcal{P}(\cdot : a_i)\ \forall a_i,$$
(4)

then we attain a rational expectations equilibrium. As will be shown later, (4) is called the unitary belief condition (Fudenberg and Levine, 1993), which is one of the three key features that distinguishes a self-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
confirming equilibrium from a rational expectations equilibrium or Nash equilibrium.

Multi-person decision problems

If we interpret the state space as the set of the strategies of the other players $X_i=A_{-i}$, we can naturally extend the basic definition to the situation where more than one person is making a decision. A self-confirming equilibrium is a profile of actions and beliefs, $\{(a_1^*, \mu_1^*), \ldots, (a_n^*, \mu_n^*)\}$ such that (2) and (3) hold for every i=1,…, n. As we move from single person to multi-person decision problems, however, (3) differs three ways from a Nash equilibrium, in addition to the unitary belief condition (4). (1) If there are more than two players, the belief of player $i \neq j$, k about player k's strategy can be different from player j's belief about player k's strategy (failure of consistency). (2) Similarly, player i can entertain the possibility that player j and player k correlate their strategies according to an un-modelled randomization mechanism, leading to correlated beliefs. (3) If we require that a self-confirming equilibrium should admit unitary and consistent beliefs, while excluding correlated beliefs, then the self-confirming equilibrium is a Nash equilibrium (Fudenberg and Levine, 1993).

Dynamic decision problems

Suppose that player i solves (1) repeatedly. The first step to embed self-confirming equilibria in dynamic contexts is to spell out learning rules that specify how beliefs respond to new observations. We define a learning rule as a mapping that updates belief μ_i into a new belief when new data arrive. Define $Z_i \subset X_i$ as a subspace of X_i that is observed by a decision maker. Let $\mathcal{M}_i$ be the set of probability distributions over $Z_i \times A_i$. These represent player i's belief about the state, that is, the model entertained by player i. A learning rule is defined as

$$T_i : \mathcal{M}_i \times Z_i \rightarrow \mathcal{M}_i.$$

A belief $\mu_i^* \in \mathcal{M}_i$ is a steady state of the learning dynamics if

$$a_i^* \in \arg\max_{a_i \in A_i} \int_{X_i} u_i(a_i, x_i)\, d\mu_i^*(x_i : a_i) \tag{5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\mu_i^* = \mathcal{T}_i(\mu_i^*, z_i) \quad (6)$$

for every z_i in the support of $\mathcal{P}_i(x_i; a_i^*)$.

The steady state of learning dynamics is a self-confirming equilibrium for a broad class of recursive learning dynamics including Bayesian (Fudenberg and Levine, 1993), and least square learning algorithms (Sargent, 1999; Evans and Honkapohja, 2001) are self-confirming equilibria.

Refinements

We can study the salience of the self-confirming equilibrium by examining the stability of the associated steady states. The stability property provides a useful foundation for selecting a sensible self-confirming equilibrium (Bray, 1982; Marcet and Sargent, 1989; Woodford, 1990; Evans and Honkapohja, 2001).

With a possible exception of the Bayesian learning algorithm, most ad hoc learning rules are motivated by the simplicity of some updating scheme as well as its ability to support sufficiently sophisticated behaviour in the limit. By exploiting the convergence properties of learning dynamics, we can often devise a recursive algorithm to *calculate* a self-confirming equilibrium, that is, a fixed point of the $\mathcal{T}_i$. This approach to computing equilibria has occasionally proved fruitful to compute equilibria in macroeconomics (for example, Aiyagari et al., 2002).

In principle, a player need not know the other player's payoff in order to play a self-confirming equilibrium. Self-confirming equilibria allow a player to entertain any belief conditioned on actions not used in the equilibrium. This is one of the main sources of multiplicity. In the game theoretic context, a player can delineate the set of possible actions of the other players, even if he does not have perfect foresight. If each player knows the payoff of the other players, and if it is common knowledge that every player is rational, then a player can eliminate the actions of the other players that cannot be rationalized. By exploiting the idea of sophisticated learning of Milgrom and Roberts, (1990; 1991), Dekel, Fudenberg and Levine (1999) restricted the set of possible beliefs off the equilibrium path to eliminate evidently unreasonable self-confirming equilibrium.

Applications

Self-confirming equilibria and recursive learning algorithms are powerful tools to investigate a number of important dynamic economic problems such as (1) the limiting behaviour of learning systems (Evans and Honkapohja, 2001; Fudenberg and Levine, 1998); (2) the selection of plausible equilibria in games and dynamic macroeconomic models (Marcet and Sargent, 1989; Kreps, 1998); (3) the incidence and distribution of rare events that occasionally arise as large deviations from self-confirming equilibria (Cho, Williams and Sargent, 2002; Sargent, 1999); and (4) formulating plausible models of how agents respond to model uncertainty (Cho and Kasa, 2006).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Remarkably, related mathematics tie together all of these applications. The mean dynamics that propel the learning algorithms to self-confirming equilibria (in item (1)) are described by ordinary differential equations (ODE) derived through an elegant stochastic approximation algorithm (for example, Fudenberg and Levine, 1998; Woodford, 1990; Marcet and Sargent, 1989). Because the stationary point of the ODE is a self-confirming equilibrium, the stability of the ODE determines the selection criterion used to make statements about item (2) (Fudenberg and Levine, 1998; Evans and Honkapohja, 2001). Remarkably, by adding an adverse deterministic shock to that same ODE, we obtain a key object that appears in a *deterministic* control problem that identifies the large-deviation excursions in item (3) *away* from a self-confirming equilibrium (Cho, Williams and Sargent, 2002). Finally, that same large deviations mathematics is associated with robust control ideas that use entropy to model how agents cope with model uncertainty (Pandit and Meyn, 2006).

Financial support from the National Science Foundation (ECS-0523620) and Deutsche Bank through the Institute for Advanced Study at Princeton is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Deutsche Bank, or the Institute for Advanced Study.

Bibliography

Aiyagari, R., Marcet, A., Sargent, T.J. and Seppälä, J. 2002. Optimal taxation without state-contingent debt. *Journal of Political Economy* 110, 1220–54.

Bray, M.M. 1982. Learning, estimation and the stability of rational expectations. *Journal of Economic Theory* 26, 318–39.

Cho, I.-K. and Kasa, K. 2006. Learning and model validation. Working paper, University of Illinois and Simon Fraser University.

Cho, I.-K., Williams, N. and Sargent, T.J. 2002. Escaping Nash inflation. *Review of Economic Studies* 69, 1–40.

Dekel, E., Fudenberg, D. and Levine, D.K. 1999. Payoff information and self-confirming equilibrium. *Journal of Economic Theory* 89, 165–85.

Evans, G.W. and Honkapohja, S. 2001. *Learning and Expectations in Macroeconomics*. Princeton: Princeton University Press.

Fudenberg, D. and Kreps, D.M. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5, 320–67.

Fudenberg, D. and Kreps, D.M. 1995. Learning in extensive games, I: self-confirming and Nash

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

equilibrium. *Games and Economic Behavior* 8, 20–55.

Fudenberg, D. and Levine, D.K. 1993. Self-confirming equilibrium. *Econometrica* 61, 523–45.

Fudenberg, D. and Levine, D. 1998. *Learning in Games*. Cambridge, MA: MIT Press.

Kreps, D.M. 1998. Anticipated utility and dynamic choice. In *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures, 1983–1997*, ed. D. P. Jacobs, E. Kalai and M.I. Kamien. Cambridge: Cambridge University Press.

Marcet, A. and Sargent, T.J. 1989. Convergence of least squares learning mechanisms in self referential linear stochastic models. *Journal of Economic Theory* 48, 337–68.

Milgrom, P. and Roberts, D.J. 1990. Rationalizability, learning and equilibrium in games with strategic complementarities. *Econometrica* 58, 1255–77.

Milgrom, P. and Roberts, D.J. 1991. Adaptive and sophisticated learning in normal form games. *Games and Economic Behavior* 3, 82–100.

Pandit, C. and Meyn, S.P. 2006. Worst-case large-deviations with application to queueing and information theory. *Stochastic Process and their Applications* 116, 724–56.

Sargent, T.J. 1999. *The Conquest of American Inflation*. Princeton: Princeton University Press.

Woodford, M.D. 1990. Learning to believe in sunspots. *Econometrica* 58, 277–307.

How to cite this article

Cho, In-Koo and Thomas J. Sargent. "self-confirming equilibria." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000534> doi:10.1057/9780230226203.1505

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Seligman, Edwin Robert Anderson (1861 – 1939)

A.W. Coats
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Adams, H. C.; inheritance tax; land tax; progressive taxation; proportional income taxation; public finance; Seligman, E. R. A.; taxation criteria.

Article

Seligman was born in New York City on 25 April 1861 and died on 18 July 1939 at Lake Placid, New York. An economist of unusual erudition, energy and wide-ranging interests, Seligman successfully combined a life of distinguished scholarship with philanthropy and active participation and leadership in a variety of reform causes. Raised in a talented and wealthy New York Jewish business family, Seligman studied privately under Horatio Alger Jr. and at Columbia Grammar School before graduating from Columbia in 1879. After three years' study in Berlin, Heidelberg (under Karl Knies), Geneva and Paris he returned to Columbia obtaining MA and LLB degrees in 1884, the Ph.D. cum laude in 1885 and a full professorship in political economy at age 30, a post he held until retirement in 1931. Dignified, wise and balanced in outlook, Seligman personified the best in late 19th-century efforts to blend orthodox classical and German historical economics. His original studies of neglected British and American economists, and his compilation of perhaps the world's greatest library of economic works, reveal his lifetime devotion to doctrinal history, while his widely read and durable *Economic Interpretation of History* (1902) testifies to the breadth and sensitivity of his historical knowledge. Like Henry Carter Adams, with whom he created the field of public finance in America, Seligman was influenced by Adolph Wagner. But he was more of a theorist than Adams, and his concepts of 'faculty' or ability to pay, and benefit, were the first systematic modern efforts to develop theoretical criteria of taxation. A severe critic of Henry George, Seligman nevertheless favoured taxes on land values and progressive inheritance taxes, and advocated proportional income taxes as early as 1894. Sympathetic to labour unions, federal railroad legislation, effective central banking measures and other moderate reform proposals, including deficit finance and public works during the depression of the 1930s, Seligman also advocated US aid to Europe after 1918 and the cancellation of their debts.
Seligman served on innumerable public bodies as a taxation and financial specialist, and as a social

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

reformer, for example as Chairman of the Bureau of Municipal Research and with the National Civic Federation. A founder, first Treasurer and later President (1902–4) of the American Economic Association, he was an outstanding champion of academic freedom and co-founder of the American Association of University Professors, of which he was President 1919–20. His success as fund-raiser and Editor in Chief of the *Encyclopaedia of the Social Sciences* 1927–35 was a fitting culmination of an outstanding career of scholarly and public service.

Selected works

1892. *On the Shifting and Incidence of Taxation*. Baltimore: American Economic Association. 5th edn, revised, New York: Columbia University Press, 1927.

1894. *Progressive Taxation in Theory and Practice*. Baltimore: American Economic Association. 2nd edn, revised and enlarged, Princeton: American Economic Association, 1908.

1895. *Essays in Taxation*. New York and London: Macmillan. 10th edn, 1925.

1902. *The Economic Interpretation of History*. New York: Columbia University Press. 2nd edn, revised, 1924.

1905. *The Principles of Economics, with Special Reference to American Conditions*. New York: Longmans Green & Co. 12th edn, revised, 1929.

1911. *The Income Tax: A Study of the History, Theory, and Practice of Income Taxation at Home and Abroad*. New York: Macmillan. 2nd edn, revised and enlarged, 1914.

1925a. *Essays in Economics*. New York: Macmillan.

1925b. *Studies in Public Finance*. New York: Macmillan.

How to cite this article

Coats, A. W. "Seligman, Edwin Robert Anderson (1861–1939)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000088> doi:10.1057/9780230226203.1506

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Selten, Reinhard (born 1930)

Eric van Damme
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes the main contributions to game theory and boundedly rational economic behaviour of Reinhard Selten, winner, together with John Nash and John Harsanyi, of the Nobel Memorial Prize in Economics in 1994.

Keywords

asymmetric games; bargaining; bounded rationality; cognitive bounds on rationality; evolutionarily stable strategies; evolutionary game theory; extensive form games; forward induction; game theory and biology; Harsanyi, J. C.; ideal rationality; imperfect competition; laboratory experimentation; learning; motivational bounds on rationality; Nash equilibrium; Nash, J.; non-cooperative games; oligopoly games; payoff dominance; Prisoner's Dilemma; rationality; risk dominance; selection theory; Selten, R.; sequential equilibrium; Simon, H.; stag hunt game; subgame consistency; subgame perfection; tracing procedure; trembling-hand perfection

Article

Reinhard Selten was born in Breslau, then in Germany, and now called Wroclaw, in Poland. His father, who was of Jewish origin, died of illness when he was 12 years old, and the family had a difficult time during the war. After having been refugees for a couple of years, they settled in Hessia, where Selten completed high school in 1951. In that same year, he had his first contact with game theory through a popular article in *Fortune*. From 1951 to 1957, Selten studied mathematics in Frankfurt, where, under the guidance of Ewald Burger, he wrote his Master's thesis on cooperative game theory, aiming to axiomatize a value for extensive form games. He continued to work for his Ph.D. in Frankfurt, in the institute of Professor Heinz Sauermann, where he soon became involved in laboratory experiments. After receiving his Ph.D. in mathematics in 1961, he remained in Frankfurt until 1969, when he moved to the Free University of Berlin. In 1972, he moved to Bielefeld, to the newly established Institute of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Mathematical Economics. In 1984, he moved to the University of Bonn, where he set up the Bonn Experimental Laboratory, to which he is still affiliated. In 1994, he received the Nobel Prize in Economics, which he shared with John Nash and John Harsanyi. For more personal details, see his autobiography for the prize (Selten, 1995).

Selten describes himself as a ‘methodic dualist’: throughout his career, he has investigated both the consequences of ideal rationality in game situations, as well as the limitations of this concept to describe actual observed behaviour, and he has proposed several more descriptive theories. For Selten's perspective on his own work and how it hangs together, see Selten (1993). A selection of his path-breaking contributions has been collected in Selten (1999).

Perfect equilibria

John Nash has proposed the fundamental solution concept for non-cooperative games, that is, games in which the players cannot make binding agreements outside of the formal rules. A *Nash equilibrium* has a natural stability property: once having arrived at it, either through introspection, learning or evolution, no player has an incentive to unilaterally deviate. Nash showed that each game has at least one equilibrium, and that a non-self-destroying, single-valued theory of rationality has to prescribe playing such an equilibrium. See strategic and extensive form games for more details for formal definitions of other concepts discussed in this article.

Applying John van Neumann's insight that any extensive form (that is, dynamic) game can be reduced to an equivalent one in which the players move just once and simultaneously, Nash focused mainly on the latter. In his early experiments, Selten was, however, working with dynamic oligopoly games. When calculating the Nash equilibria, to provide a benchmark to compare the actual outcomes with, he soon discovered that there typically were many, including those that could not be considered compatible with rationality. To eliminate these, he proposed the refinement of subgame perfection. The idea underlying this concept is that, once a subgame (a part that constitutes a game in itself) is reached, everything outside of it (the things that could have happened but did not) has become irrelevant, so that the logic underlying the equilibrium should be applied to the subgame as well.

By requiring subgame perfection, one eliminates equilibria that rely on non-credible threats or promises. As an example, suppose that two players have to divide $4. One player, the proposer, *P*, may offer 3, 2 or 1 to the responder, *R*, who can only accept or reject; if he accepts, the division is implemented, otherwise each player gets zero. Assume it is commonly known that each player cares only about his own monetary payoff and prefers more money to less. The game has a Nash equilibrium in which *P* offers 3 and *R* rejects any outcome in which he is offered less. The threat to reject any amount, however, is not credible: confronted with offer *m*, by assumption *R* prefers *m* to 0, hence, he should accept it. In the unique subgame perfect equilibrium, *P* offers 1, which *R* accepts.

Later, Selten discovered that subgame perfection is not sufficient to rule out all ‘non-rational’ equilibria. In Selten (1975), he introduced a further refinement: ‘trembling-hand perfection’, a concept that insists that equilibria be robust with respect to small perturbations of the strategies. Formally, it is assumed that, whenever a player has to move, with a small probability he makes a mistake, and an equilibrium is trembling-hand perfect if, under these circumstances, each player is still willing to play it. Ideal rationality is thus viewed as a limiting case of rationality with small errors.

Selten was the first to refine Nash's concept for analysing dynamic games. In the beginning of the 1970s,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
he was also one of the first to successfully apply these refined concepts to dynamic game models of imperfect competition, thus pioneering the use of game theory in industrial organization. Subgame perfection is now routinely applied in situations of strategic interaction involving a time element. The robustness approach proposed in Selten (1975) also turned out to be very fruitful. On the one hand, it gave rise to solution concepts, such as sequential equilibrium, that were easy to use in applications; on the other hand, it induced a search for further refinements and stability concepts that uncovered the deeper mathematical structure of ideal rationality; see Kohlberg and Mertens (1986). We refer to Nash equilibrium, refinements of for further details.

Equilibrium selection and the cooperation with John Harsanyi

In 1965, during a workshop in Jerusalem, Selten met John Harsanyi, and they started to cooperate. During the academic year 1967–8, Selten visited Harsanyi in Berkeley to work on two-person bargaining under incomplete information. They decided to take the rationality assumption in situations of strategic interaction to its logical limit, and started a project that would be finished only 20 years later with Harsanyi and Selten (1988). The motivation for this work is that the traditional justification of the Nash equilibrium concept is incomplete: it relies on the assumption that each game has a unique solution, but a game typically has multiple Nash equilibria. The natural question thus is, whether, by strengthening the rationality requirements, a unique selection can be obtained.

An essential ingredient of their selection theory is the notion of risk dominance. The concept tries to assess whether, in a situation in which players are a priori uncertain about which equilibrium should be played, they can still coordinate on a single equilibrium and, if so, which one. The stag hunt game may illustrate the concept and can show that there may be a conflict between payoff dominance and risk dominance. Assume two players that each can choose between a safe action, S, and a risky one, R. S yields 1 for sure, while R gives payoff X, with $X>1$, if the other player also chooses R, but it yields 0 otherwise. Both (R,R) and (S,S) are equilibria and the former Pareto dominates the latter. If X is not too large, however, one can gain only little by playing R and the downward risk is considerable. In Harsanyi and Selten (1988), risk dominance is formalized by means of the tracing procedure, a theoretical model of the thinking process that converts any mixed strategy profile into an equilibrium of the game. In our example, (S,S) is the risk dominant equilibrium if $X<2$ and there is a conflict between risk dominance and payoff dominance if $1<X<2$. The concept has found applications also in other domains of game theory. For example, in an evolutionary setting, where players repeatedly play the game in a myopic fashion, adjusting strategies through time, it was found that risk dominance is related to the long-run stochastic stability of the equilibrium (see learning and evolution in games: an overview).

For extensive form games, Harsanyi and Selten propose subgame consistency, a natural extension of subgame perfection, as an important selection principle: if g is a subgame of G, then in g the solution of G should prescribe the solution of g. Again the idea is that, once g is reached, everything else has become irrelevant. To illustrate, let g be the stag hunt game with $X=1.8$, and let G be the game in which player 1 first chooses whether to take up an outside option yielding both players 1.5 or to play g. A selection theory incorporating subgame consistency and risk dominance implies that player 1 should take up the outside option. On the other hand, the strategy ‘go to g and play S’ is a dominated strategy in the overall game, so that repeated elimination of dominated strategies, or forward induction, produces

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the outcome (*R*,*R*). The example shows that several desirable properties easily conflict. The conclusion is that it is possible to construct a single valued theory of rationality in non-cooperative games, but that an ideal theory does not exist.

Bounded rationality and experiments

When, after finishing his Master's thesis in 1957, Selten started to work in the institute of Professor Heinz Sauermann on a project on the ‘Theory of the Firm’, under the influence of Herbert Simon's ideas he very quickly became convinced that it was necessary to model economic behaviour as boundedly rational. (Also see rationality, bounded.) In Selten's view, bounded rationality (defined as the rationality exhibited by actual human behaviour) differs fundamentally from the ideal rationality that we have been discussing in this article thus far. Bounded rationality, hence, is not viewed as an approximation to full rationality; it simply is something different. Furthermore, since actual behaviour cannot be invented in the armchair, the development of theories of bounded rationality needs an empirical basis. Consequently, laboratory experimentation becomes an important source of empirical evidence, also because of the possibility of gathering data under controlled circumstances. See also experimental economics. Selten's first experimental paper, co-authored with Sauermann, had already appeared in 1959. Given the sharp distinction drawn between ideal rationality and bounded rationality, it is not too surprising that the work was not directed at testing a theory; rather, it was an explorative piece, trying to uncover regularities. Ever since that first paper, Selten's work in this area has tried to uncover the structure of boundedly rational decision making, with an emphasis on individual data.

Selten distinguishes between cognitive bounds on rationality and motivational bounds. The former arise from the limited human ability to think and compute. Motivational bounds are different: even when the standard rational solution appears obvious, and is fully understood, as in the finitely repeated Prisoner's Dilemma game, a player may not have the incentive to implement it. In his paper on the chain store game, Selten argued that we lack a ‘behavioural trust’ in abstract subgame perfection arguments, and he proposed a theory of decision making in which decisions can arise at three different levels: the routine level, using past experience and analogies; the imagination level, involving various scenarios; and the reasoning level in which the individual makes a conscious effort to analyse the situation in rational, logical way. Since different levels involve different decision costs, not all levels may be activated, and even if they are, then, in Selten's view, there is no reason why the decision at the higher level should be selected. This brief description may make clear why boundedly rational behaviour may indeed be very different from normative rationality.

An important tool in the uncovering of the structure of boundedly rational decision making has been the use of the strategy method: in experiments, subjects are asked not just to make decisions as the game goes along, but also to specify or programme strategies for the entire game. In this way more of the strategic reasoning of the subject is revealed. Selten game theory and biology. Originally, the fundamental solution concept in evolutionary game theory, that of evolutionarily stable strategies (ESS), had been defined only for symmetric static games; Selten extended it to dynamic games and to asymmetric games, thus expanding the range of applicability. It is important to note that evolutionary game theory is not normative but descriptive, with equilibrium being thought of as the result of a dynamic process. Even in games arising in economics and social situations, equilibrium is increasingly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
viewed in this way.
Selten has often worked in areas away from the mainstream. As he has written: ‘Since I am slow, I have to try to be early.’ In the areas that he pioneered in the first part of his career, many others have meanwhile joined him.

See Also

- experimental economics
- game theory and biology
- learning and evolution in games: an overview
- Nash equilibrium, refinements of
- rationality, bounded
- strategic and extensive form games

Selected works

1959. (With H. Sauermann.) Ein Oligopolexperiment. *Zeitschrift für die gesamte Staatwissenschaft* 115, 427–71.

1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatwissenschaft* 121, 301–24 and 667–89.

1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4, 25–55.

1978. The chain store paradox. *Theory and Decision* 9, 127–59.

1988. (With J.C. Harsanyi.) *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

1993. In search of a better understanding of economic behavior. In *Makers of Modern Economics*, ed. A. Heertje. New York: Simon and Schuster.

1995. Autobiography. *Les Prix Nobel. The Nobel Prizes* 1994, ed. T. Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1994/selten-autobio.html, accessed 23 March 2007.

1998. Features of experimentally observed bounded rationality. *European Economic Review* 42, 413–36.

1999. *Game Theory and Economic Behavior. Selected Essays*, 2. vols. Cheltenham-Northampton:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Edward Elgar.

Bibliography

Kohlberg, E. and Mertens, J.-F. 1986. On the strategic stability of equilibria, *Econometrica* 54, 1003–38.

How to cite this article

Damme, Eric van. "Selten, Reinhard (born 1930)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000445> doi:10.1057/9780230226203.1507

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

semiparametric estimation

James L. Powell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Semiparametric estimation methods are used for models which are partly parametric and partly nonparametric; typically the parametric part is an underlying regression function which is assumed to be linear in the observable explanatory variables, while the nonparametric component involves the distribution of the model's ‘error terms’. Semiparametric methods are particularly useful for limited dependent variable models (for example, the binary response or censored regression models), since fully parametric specifications for those models yield inconsistent estimators if the parametric distribution of the errors is misspecified.

Keywords

binary response models; censored regression (‘Tobit’) models; fixed effects; identification; kernel estimators; limited dependent variable models; linear regression models; maximum likelihood; maximum score estimation; nonparametric estimation; panel data models; propensity score; sample selection models; selectivity bias; semiparametric estimation; semiparametric regression models

Article

Introduction

Semiparametric estimation methods are used to obtain estimators of the parameters of interest – typically the coefficients of an underlying regression function – in an econometric model, without a complete parametric specification of the conditional distribution of the dependent variable given the explanatory variables (regressors). A structural econometric model relates an observable dependent variable y to some observable regressors x; some unknown parameters β, and some unobservable ‘error term’ ε, through some functional form $y=g(x, \beta, \varepsilon)$; in this context, a semiparametric estimation problem does not restrict the distribution of ε (given the regressors) to belong to a parametric family determined by a finite number of unknown parameters, but instead imposes only broad restrictions on the distribution of ε (for example, independence of ε and x, or symmetry of ε about zero given x) to obtain identification of β and construct consistent estimators of it.
Thus the term ‘semiparametric estimation’ is something of a misnomer; the same estimator can be considered a parametric, semiparametric or nonparametric estimator depending upon the restrictions imposed upon the economic model. For example, if a random sample of dependent variables $\{y_i\}$ and regressors $\{x_i\}$ are assumed to satisfy a linear regression model $y_i = x_i'\beta + \varepsilon_i$, the classical least squares estimator can be considered a ‘parametric’ estimator of the regression coefficient vector β if the error terms $\{\varepsilon_i\}$ are assumed to be normally distributed and independent of $\{x_i\}$. It could alternatively be considered a ‘nonparametric estimator’ of the best linear predictor coefficients $\beta = \left[E(x_i x_i')\right]^{-1} E(x_i, y_i)$ if only the weak condition $E(x_i \varepsilon_i)$ is imposed (implying that β is a unique function of the joint distribution of the observations). And the least squares estimator would be ‘semiparametric’ under the intermediate restriction $E(\varepsilon_i|x_i)=0$, which imposes a parametric (linear) form for the conditional mean $E(y_i|x_i)=x_i'\beta$ of the dependent variable but imposes no further restrictions on the conditional distribution. So the term

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
'semiparametric' is a more suitable adjective for models which are partly (but not completely) parametrically specified than it is for the estimators of those parameters.

Nevertheless, while most econometric estimation methods that do not explicitly specify the likelihood function of the observable data (for example, least squares, instrumental variables, and generalized method-of-moments estimators) could be considered semiparametric estimators, 'semiparametric' is sometimes used to refer to estimators of a finite number of parameters of interest (here, β) that involve explicit nonparametric estimators of unknown nuisance functions (for example, features of the distribution of the errors ε). Such 'semiparametric estimators' use nonparametric estimators of density or regression functions as inputs to second-stage estimators of regression coefficients or similar parameters. Occasionally terms like 'semi-nonparametric', 'distribution-free' and even 'nonparametric' have been used to describe such estimation methods, with the latter terms referring to the treatment of the error terms in an otherwise-parametric structural model.

The primary objective of semiparametric methods is to identify and consistently estimate the unknown parameter of interest β by determining which combinations of structural functions $g(x, \beta, \varepsilon)$ and weak restrictions on the distribution of the errors ε permit this. Given identification and consistent estimation, the next step in the statistical theory is determination of the speed with which the estimator $\hat{\beta}$ converges to its probability limit β. The rate of convergence for estimators for standard parametric problems is the square root of the sample size *n*, while nonparametric estimators of unknown density and regression functions (with continuously distributed regressors) generically converge at a slower rate; if a semiparametric estimator can be shown to converge at the parametric rate, that is, if it is 'root-*n* consistent', then its relative efficiency to a parametric estimator (for a correctly specified parametric model) will not tend to zero as n increases. For inference, it is also useful to demonstrate the asymptotic (that is, approximate) normality of the distribution of $\hat{\beta}$ in large samples, so that asymptotic confidence regions and hypothesis tests can be constructed using normal sampling theory. Finally, for problems where existence of root-*n* consistent, asymptotically normal semiparametric estimators can be shown, the question of efficient estimation arises. The solution to this question has two parts – determination of the efficiency bound for the semiparametric estimation problem and construction of a feasible estimator that attains that bound.

Econometric applications

In econometrics, most of the attention to semiparametric methods dates from the late 1970s and early 1980s, which saw the development of parametric models for discrete and *limited dependent variable* (*LDV*) models. Unlike the linear regression model, those models are not additive in the underlying error terms, so the validity (specifically, the consistency) of maximum likelihood and related estimation methods depends crucially on the assumed parametric form of the error distribution. As shown for particular examples by Arabmazar and Schmidt (1981; 1982) and Goldberger (1983), failure of the standard assumption of normally distributed error terms makes the corresponding likelihood-based estimators inconsistent. This is in contrast to the linear regression model, where the maximum likelihood (classical least squares) estimator is consistent under much weaker assumptions than normally (and identically) distributed errors.

Much of the early literature on semiparametric estimation concentrated on a particular limited dependent variable model, the *binary response model*, which arguably presents the most challenging setting for identification and estimation of the underlying regression coefficients. Early examples of semiparametric identification assumptions and estimation methods for this model give a flavour of the approaches used for other econometric models, among them the *censored regression* and *sample selection* models. The discussion here treats only selected assumptions and estimators for these models, and not their numerous variants; more complete surveys of semiparametric models and estimation methods are given by Manski (1989), Powell (1994), Newey (1994), and Pagan and Ullah (1999).

Semiparametric binary response models

The earliest semiparametric estimation methods in the econometrics literature on LDV models concerned the *binary response model*, in which the dependent variable *y* assumed the values zero or 1 depending upon the sign of some underlying latent (unobservable) dependent variable y^* which satisfies a linear regression model $y^* = x'\beta + \varepsilon$; that is,

$$y_i = 1\left\{x_i'\beta - \varepsilon_i > 0\right\},$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where '1{*A*}' denotes the indicator function of the event *A*; that is, it is 1 if *A* occurs and is zero otherwise. For a parametric model, in which the errors ε_i are assumed to be independent of x_i and distributed with a known marginal cumulative distribution function $F(\varepsilon)$, the average log-likelihood function takes the form

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln F(x_i'\beta) + (1 - y_i)\ln(1 - F(x_i'\beta))\right]$$

for a random sample of size *n*, and consistency of the corresponding the maximum likelihood estimator $\hat{\beta}_{ML}$ requires correct specification of *F* unless the regressors satisfy certain restrictions (as discussed by Ruud, 1986). When *F* is unknown, a scale normalization on β is required, and a constant (intercept) term will not be identified no normalization on the location of ε is imposed.

Manski (1975; 1985) proposed a semiparametric alternative, termed the 'maximum score' estimator, which defined the estimator to maximize the number of correct matches of the value of y_i with an indicator function $1\{x_i'\beta > 0\}$ of the positivity of the regression function. That is, the maximum score estimator $\hat{\beta}_{MS}$ maximizes the average 'score' function

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}[y_i \cdot 1\{x_i'\beta > 0\} + (1 - y_i)\cdot 1\{x_i'\beta \le 0\}]$$

over β. Unlike the maximum likelihood estimator $\hat{\beta}_{ML}$, consistency of $\hat{\beta}_{MS}$ requires only that the median of the error terms was zero given the regressors, that is, the conditional cumulative $F(\varepsilon|x)$ of given $x_i=x$ had $F(\lambda|x)>1/2$ when $\lambda > 0$, and $F(\lambda|x)<1/2$ when $\lambda<0$. However, the estimation approach is generally not root-n consistent (as shown by Chamberlain, 1986). A variant of the maximum score estimator, proposed by Horowitz (1992), essentially 'smoothed' the indicator functions for positivity of $x_i'\beta$ in the minimand $S_n(\beta)$ using a continuous approximation to it, similar to the smoothing used in nonparametric kernel estimators of regression and density functions. The rate of convergence of the resulting 'smoothed maximum score' estimator can be made arbitrarily close to the root-*n* rate if the distribution of the regressors is sufficiently smooth.

To obtain root-*n* consistent estimators of the unknown β, the assumption on the error term ε can be strengthened to independence of ε and *x*. Han (1987) proposed an alternative to the maximum score estimator, termed the 'maximum rank correlation' estimator, which compared the sign of the difference y_i-y_j of the dependent variable to the corresponding difference $(x_i-x_j)'\beta$ in the regression functions across all distinct pairs of observations *i* and *j*. The estimator $\hat{\beta}_{MRC}$ maximizes

$$M_n(\beta) = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sgn(y_i - y_j)\cdot sgn((x_i - x_j)'\beta)$$

$\Leftrightarrow$ over β, where $sgn(u)\equiv 1\{u>0\}-1\{u<0\}$. The rationale for this estimator is based upon the monotonicity of $\Pr\{y_i = 1|x\} = F(x_i'\beta)$ in $x_i'\beta$ so that, given $y_i\neq y_j$, $\Pr\{y_i>y_j|x_i, x_j\}$ exceeds 1/2 when $x_i'\beta > x_j'\beta$. Han's article gave conditions under which $\hat{\beta}_{MRC}$ was shown to be consistent, and Sherman (1993) showed that this estimator was root-*n* consistent and asymptotically normal.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

An alternative estimation approach for β under the assumption of independence of u and x combines estimation of the parameter vector β with nonparametric estimation of the unknown distribution function F. Cosslett (1983) proposed a 'nonparametric maximum likelihood' estimator $\hat{\beta}_{NPML}$ obtained by simultaneously maximizing the likelihood function $L_n(\beta)=L_n(\beta; F)$ over both β and F, where the latter function is restricted to be nondecreasing with values in the unit interval. While consistency of this estimator could be established, its rate of convergence could not. An alternative estimation method, proposed by Klein and Spady (1993), used kernel regression methods to estimate the unknown distribution function F in the likelihood function. The resulting estimator was shown to be root-n consistent and asymptotically normally distributed under additional regularity conditions; furthermore, the estimator was shown to achieve the semiparametric efficiency bound for this problem, that is, its asymptotic covariance matrix is the smallest possible among regular estimators of β which impose only the independence restriction between x and u.

Still other estimators for β when u and x are independent exploit the single index regression structure of this model, since the conditional expectation of y_i given x_i only depends upon the 'single index' $x_i'\beta$:

$$E[y_i|x_i] \equiv g(x_i) = F(x_i'\beta).$$

If the vector of regressors x_i is continuously distributed with joint density function $f_X(x)$ which is continuous for all x, Stoker (1986) noted that the vector of slope parameters β is proportional to the expectation of the derivative of $g(x)$,

$$E\left[\frac{\partial g(x_i)}{\partial x}\right] = E\left[F'(x_i'\beta)\right]\cdot\beta.$$

Using integration-by-parts, this 'average derivative' can in turn be expressed as the expected value of the product of $-y_i$ and the derivative of the logarithm of the density f_X of the regressors,

$$E\left[\frac{\partial g(x_i)}{\partial x}\right] = -E\left[y_i\frac{\partial \log[f_X(x_i)]}{\partial x}\right].$$

Härdle and Stoker (1989) proposed a semiparametric estimator of this representation of β (up to scale) using nonparametric (kernel) estimators of f_X and its gradient, while Powell, Stock and Stoker (1989) constructed a similar estimator of the 'density-weighted average derivative'

$$E\left[f_X(x_i)\frac{\partial g(x_i)}{\partial x}\right] = E[f_X(x_i)F'(x_i'\beta)]\cdot\beta = -2E\left[y_i\frac{\partial f_X(x_i)}{\partial x}\right],$$

which is also proportional to β under the single index restriction.

Though the motivation given here was based upon the binary response model under independence of the errors and regressors, the average derivative and weighted average derivative estimators apply to other models with a single index structure, for example, any *transformation model* with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$y_i = T(x_i'\beta + \varepsilon_i),$$

for T a nondegenerate function (possibly unknown) and with ε_i continuously distributed and independent of x_i. The same is true for the 'single index regression' estimator proposed by Ichimura (1993), defined to minimize

$$R_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{F}(x_i'\beta;\beta)\right)^2 t_n(x_i);$$

in this expression, $\hat{F}(u;\beta)$ represents a nonparametric regression estimator of $E[y_i|x_i'\beta = u]$ and $t_n(x_i)$ represents a 'trimming' term which is zero whenever x_i lies outside a set for which F is sufficiently precisely estimated. Unlike the average derivative $\hat{\beta}_{AD}$ and weighted average derivative $\hat{\beta}_{WAD}$ estimators, which require the regressors to be jointly continuously distributed, root-n consistency and asymptotic normality of the single index regression estimator $\hat{\beta}_{SIR}$ require only that $x_i'\beta$ has a continuous distribution, so that some of the regressors can be discrete. The criterion function $R_n(\beta)$ is the nonlinear least squares analogue of the maximand for the Klein and Spady (1993) estimator (which also involved a similar trimming term $t_n(x_i)$). The asymptotic covariance matrices for both estimators have the same general form as the corresponding nonlinear least squares and maximum likelihood estimators with F known, except for the replacement of the cross product of the regressors x_i with the cross product of $x_i - E[x_i|x_i'\beta]$, adjusting the asymptotic covariance matrices upward to account for the nonparametric estimation of the unknown function F.

The problem of consistent estimation of β in binary response models is compounded for *panel data* models with *fixed effects* (that is, individual-specific intercept terms), written as

$$y_{it} = 1\left\{x_{it}'\beta + \alpha_i - \varepsilon_{it} > 0\right\}$$

for individuals i ranging from 1 to n and time periods t from 1 to T. For this model, even if the distribution function F of the error terms ε_{it} is known, the maximum likelihood estimators of β and the fixed effects $\{\alpha_i\}$ will generally be inconsistent if the number of time periods T is fixed as N increases. A consistent semiparametric estimation strategy using a variant of the maximum rank correlation estimator was proposed by Manski (1987); for the special case T=2 (that is, two time periods), the estimator $\hat{\beta}_{BPD}$ can be defined as the maximizer of the criterion

$$P_{(n)}(\beta) = \frac{1}{n}\sum_{i=1}^{n} sgn(y_{i2} - y_{i1}) \cdot sgn\left((x_{i2} - x_{i1})'\beta\right),$$

which is analogous to $M_n(\beta)$, except that the differencing is across time periods rather than across individuals. While consistency of $\hat{\beta}_{BPD}$ was established under weak conditions on the error terms, it is not possible to obtain a root-n consistent estimator unless the errors are logistic (Chamberlain, 1993) or other restrictive assumptions (for example, independence of the fixed effect α_i and the regressors x_{it}, or the conditions in Honoré and Lewbel, 2002; Lee, 1999) are imposed.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Other semiparametric econometric models

Many of the identifying assumptions imposed on semiparametric binary response models give identification and yield consistent estimators for other limited dependent variable models, though these models can sometimes be identified and consistently estimated using assumptions that are uninformative for binary response. Consider, for example, the censored regression ('Tobit') model, in which the dependent variable y_i satisfies a linear regression model if it is nonnegative, and is zero otherwise:

$$y_i = \max\left\{0,\ x_i'\beta + \varepsilon_i\right\}.$$

For this model, as for the binary response model, the dependent variable is a monotonic function of the error term ε_i; since monotone transformations by definition preserve orderings, the median (or any other percentile) of this monotonic transformation of ε_i is the monotonic transformation evaluated at the median. Thus the assumption that the errors ε_i have conditional median zero given x_i implies that the conditional median of y_i given x_i takes the form $\max\left\{0,\ x_i'\beta\right\}$, depending only on the unknown coefficients β and not on the shape of the distribution of ε_i. Using this fact, and the characterization of medians as minimizers of a least absolute deviations criterion, Powell (1984) proposed estimation of the unknown β vector by the minimizer $\hat{\beta}_{CLAD}$ of

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - \max\left\{0,\ x_i'\beta\right\} \right|$$

for this model; it is analogous to the maximum score estimator $\hat{\beta}_{MS}$ for the binary response model, which can be defined as the minimizer of the sample average absolute deviation of y_i from its conditional median function $1\left\{x_i'\beta > 0\right\}$ for binary response with median zero errors. (The maximum rank correlation estimator $\hat{\beta}_{MRC}$ and binary panel data estimator $\hat{\beta}_{BPD}$ can also be expressed as solutions to least absolute deviations problems.) Unlike $\hat{\beta}_{MS}$, though, the censored median estimator $\hat{\beta}_{CLAD}$ is root-n consistent and asymptotically normally distributed under weak regularity conditions, without need for a scale normalization. An alternative estimator for this model, which involved a nonparametric estimator of the probability that y_i equals zero given x_i, was proposed by Buchinsky and Hahn (1998).

A stronger restriction on the error distribution is conditional symmetry about zero given the regressors; while this restriction is no more informative than the implied zero median restriction for binary response, it yields different identification approaches for censored regression. Specifically, the 'symmetrically censored' residual

$$u_i(\beta) \equiv \min\left\{y_i - x_i'\beta,\ x_i'\beta\right\} = \min\left\{\max\left\{-x_i'\beta,\ \varepsilon_i\right\},\ x_i'\beta\right\}$$

is an even function of ε_i when the regression function $x_i'\beta$ is positive, and thus is itself conditionally symmetric about zero. This implies a population moment restriction

$$0 = E\left[1\left\{x_i'\beta > 0\right\}\psi(\tilde{u}_i(\beta)) \cdot x_i\right],$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

for $\psi(u)=-\psi(-u)$ an odd function of its argument. Powell (1986) proposed a 'symmetrically censored least squares' estimator of β based upon this restriction with $\psi(u)=u$; like the censored median estimator $\hat{\beta}_{CLAD}$ – which exploits the same moment condition with $\psi(u)=sgn(u)$ – the estimator $\hat{\beta}_{SCLS}$ is root-n consistent and asymptotically normally distributed under weak assumptions. Neither estimator involves explicit nonparametric estimation of the error distribution, a feature shared by the maximum score estimator $\hat{\beta}_{MS}$ and its relatives $\hat{\beta}_{MRC}$ and $\hat{\beta}_{BPD}$ for binary response.

As for the binary response model or most limited dependent variable models, consistent estimation of slope coefficients using panel data with fixed effects is challenging, with maximum likelihood estimators for β being inconsistent when the number of time periods is fixed and the number estimated fixed effects increases. For the special case T=2, writing

$$y_{it} = \max\left\{0,\ x_{it}'\beta + \alpha_i + \varepsilon_{it}\right\},$$

Honoré (1992) noted that the difference in 'identically trimmed' residuals

$$\tilde{Q}_i(\beta) = \max\left\{-x_{i1}'\beta,\ y_{i2} - x_{i2}'\beta\right\} - \max\left\{-x_{i2}'\beta,\ y_{i1} - x_{i1}'\beta\right\} = \max\left\{-x_{i1}'\beta,\ -x_{i2}'\beta,\ \alpha_i + \varepsilon_{i2}\right\} - \max\left\{-x_{i1}'\beta,\ -x_{i2}'\beta,\ \alpha_i + \varepsilon_{i2}\right\}$$

would be symmetrically distributed about zero if the error terms ε_{i1} and ε_{i2} were identically distributed given x_{i1} and x_{i2} and value of the fixed effect α_i. This implies population moment conditions of the form

$$0 = E[\psi(\tilde{Q}_i(\beta)) \cdot (x_{2i} - x_{i1})],$$

again with $\psi(u)$ an odd function of its argument. Setting $\psi(u)=sgn(u)$ and $\psi(u)=u$ yields root-n consistent and asymptotically normal estimators which are similar to the censored least absolute deviations estimator $\hat{\beta}_{CLAD}$ and symmetrically-censored least squares estimator $\hat{\beta}_{SCLS}$, respectively.

Other estimation approaches for censored regression involve explicit nonparametric estimation of features of the distribution of the error terms, which is common for other semiparametric econometric models. One such model is the *semiparametric regression* (or *semilinear regression*) model, for which some regressors enter linearly while others enter nonparametrically. The model can be written algebraically as

$$y_i = x_i'\beta + \lambda(w_i) + \varepsilon_i \equiv x_i'\beta + u_i,$$

where the error terms ε_i are restricted to satisfy $E[\varepsilon_i|x_i,w_i]=0$, or, equivalently, $E[u_i|x_i,w_i]=E[u_i|w_i]\equiv\lambda(w_i)$; the regressors x_i and w_i thus enter parametrically (linearly) or nonparametrically in the conditional mean of y_i. Robinson (1988) exploited the fact that

$$y_i - E[y_i|w_i] = (x_i - E[x_i|w_i])'\beta + \varepsilon_i$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

to construct a root-*n* consistent, asymptotically-normal estimator of β by applying least squares estimation to this equation, replacing the unknown quantities $E[y_i|w_i]$ and $E[x_i|w_i]$ by nonparametric (kernel) estimators. For the parameters β to be identified for this model, the covariance matrix of the 'residual regressors' must be nonsingular, ruling out functional dependence of x_i on w_i.
Though the semilinear regression model is not itself a limited dependent variable model, it arises as a consequence of 'selectivity bias' in a bivariate limited dependent variable model, the *censored selection model*, in which a linear latent 'outcome' variable $y_i^* = x_i'\beta + \varepsilon_i$ is observed only if some related binary 'selection' variable d_i equals 1:

$$d_i = 1\left\{w_i'\delta - \eta_i > 0\right\},$$
$$y_i = d_i \cdot (x_i'\beta + \varepsilon_i),$$

where the regressors x_i and w_i are observed and the unobserved error terms η_i and ε_i need not be mutually independent.
Heckman (1979) showed that, for the uncensored (d_i=1) subsample from this model, the dependent variable satisfied a semilinear regression model, since

$$E[y_i|d_i = 1, x_i, w_i] = x_i'\beta + \lambda(w_i'\delta);$$

when the errors are jointly normal, as Heckman assumed, the function $\lambda(u)$ has a known parametric form, but is nonparametric if the error distribution is not in a parametric family. Cosslett (1991) developed a consistent two-step estimator for the regression parameters β in the outcome equation, computing a binary nonparametric maximum likelihood estimator $\hat{\delta}_{NPML}$ of δ in the first step and using a step-function approximation to $\lambda(u)$ in a least squares fit of the outcome equation for the uncensored observations. Ahn and Powell (1993) proposed a root-*n* consistent two-step estimator of β for the related semilinear model

$$E[y_i|d_i = 1, x_i, w_i] = x_i'\beta + \lambda^*(p(w_i)),$$

where the 'propensity score' $p(w_i) \equiv E[d_i|w_i]$ is first estimated by a nonparametric regression method; this semilinear model is implied by a generalization of the original censored selection model, replacing the linear form of the regression function $w_i'\delta$ in the selection equation with an unknown function of the regressors w_i.
Some variations of the censored selection model admit other semiparametric identification strategies. For example, if the selection equation is censored rather than binary, for example, if

$$d_i = \max\left\{0, w'\delta + \eta_i\right\},$$
$$y_i = 1\left\{d_i > 0\right\} \cdot (x_i'\beta + \varepsilon_i),$$

then Honoré, Kyriazidou and Udry (1997) construct a root-*n* consistent two-step estimator of β under the assumption that the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

errors η_i and ε_i are jointly symmetric about zero given the regressors w_i and x_i, using a symmetrically censored least squares estimator of δ in the first step and exploiting the symmetry of $y_i - x_i'\beta$ about zero given that $0 < d_i < 2w_i'\delta$ in the second step. In contrast, estimation of censored selection models for panel data with fixed effects is no less challenging than for binary panel data models; Kyriazidou (1997) proposes a consistent (but not root-*n* consistent) two-step estimator for the panel data selection model

$$d_{it} = 1\left\{w_{it}'\delta + v_i - \eta_{it} > 0\right\},$$
$$y_{it} = d_{it} \cdot (x_{it}'\beta + \alpha_i + \varepsilon_{it}),$$

using Manski's (1987) binary panel data estimator to estimate δ in the first step and a semilinear regression estimator similar to the Ahn and Powell (1993) approach in the second step.

See Also

- nonlinear panel data models
- nonparametric structural models
- quantile regression
- robust estimators in econometrics
- selection bias and self-selection
- Tobit model

Bibliography

Ahn, H. and Powell, J.L. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58, 3–29.

Arabmazar, A. and Schmidt, P. 1981. Further evidence on the robustness of the Tobit estimator to heteroscedasticity. *Journal of Econometrics* 17, 253–8.

Arabmazar, A. and Schmidt, P. 1982. An investigation of the robustness of the Tobit estimator to non-normality. *Econometrica* 50, 1055–63.

Buchinsky, M. and Hahn, J. 1998. An alternative estimator for the censored quantile regression model. *Econometrica* 66, 653–72.

Chamberlain, G. 1986. Asymptotic efficiency in semiparametric models with censoring. *Journal of Econometrics* 32, 189–218.

Chamberlain, G. 1993. Feedback in panel data models. Unpublished manuscript, Department of Economics, Harvard University.

Cosslett, S.R. 1983. Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 51, 765–82.

Cosslett, S.R. 1991. Distribution-free estimator of a regression model with sample selectivity. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. W.A. Barnett, J.L. Powell and G. Tauchen. Cambridge: Cambridge University Press.

Goldberger, A.S. 1983. Abnormal selection bias. In *Studies in Econometrics, Time Series, and Multivariate Statistics*, ed. S. Karlin, T. Amemiya and L. Goodman. New York: Academic Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Han, A.K. 1987. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* 35, 303–16.

Härdle, W. and Stoker, T.M. 1989. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84, 986–95.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–62.

Honoré, B.E. 1992. Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60, 533–65.

Honoré, B.E., Kyriazidou, E. and Udry, C. 1997. Estimation of type 3 Tobit models using symmetric trimming and pairwise comparisons. *Journal of Econometrics* 76, 107–28.

Honoré, B.E. and Lewbel, A. 2002. Semiparametric binary choice panel data models without strictly exogenous regressors. *Econometrica* 70, 2053–63.

Horowitz, J.L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–31.

Ichimura, H. 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *Journal of Econometrics* 58, 71–120.

Klein, R.W. and Spady, R.H. 1993. An efficient semiparametric estimator for discrete choice models. *Econometrica* 61, 387–421.

Kyriazidou, E. 1997. Estimation of a panel data sample selection model. *Econometrica* 65, 1335–64.

Lee, M-J. 1999. A root-*n* consistent semiparametric estimator for related effect binary response panel data. *Econometrica* 60, 533–65.

Manski, C.F. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C.F. 1985. Semiparametric analysis of discrete response, asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27, 205–28.

Manski, C.F. 1987. Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* 55, 357–62.

Manski, C.F. 1989. *Analog Estimation Methods in Econometrics*. New York: Chapman & Hall.

Newey, W.K. 1994. The asymptotic variance of semiparametric estimators. *Econometrica* 62, 1349–82.

Pagan, A.P. and Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Powell, J.L. 1984. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25, 303–25.

Powell, J.L. 1986. Symmetrically trimmed least squares estimation of Tobit models. *Econometrica* 54, 1435–60.

Powell, J.L. 1994. Estimation of semiparametric models. In *Handbook of Econometrics*, vol. 4, ed. R.F. Engle, and D.L. McFadden. Amsterdam: North-Holland.

Powell, J.L., Stock, J.H. and Stoker, T.M. 1989. Semiparametric estimation of weighted average derivatives. *Econometrica* 57, 1403–30.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Robinson, P. 1988. Root-N-consistent semiparametric regression. *Econometrica* 56, 931–54.

Ruud, P. 1986. Consistent estimation of limited dependent variable models despite misspecification of distribution. *Journal of Econometrics* 32, 157–87.

Sherman, R.P. 1993. The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61, 123–37.

Stoker, T.M. 1986. Consistent estimation of scaled coefficients. *Econometrica* 54, 1461–81.

How to cite this article

Powell, James L. "semiparametric estimation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000469> doi:10.1057/9780230226203.1508

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sen, Amartya (born 1933)

Sudhir Anand
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Amartya Sen has made fundamental contributions to social choice theory, welfare economics, economic measurement, axiomatic choice theory, rationality and economic behaviour, development economics, poverty and famines, gender inequalities and family economics, among many other areas. His contributions to the field of welfare economics were cited for his award of the 1998 Nobel Memorial Prize in Economics. Sen's work combines foundational and theoretical originality, and the willingness to reconsider basic assumptions. His approach is unusual for its breadth of concern coupled with an uncompromising rigour of analysis. His writings address some of the most important human and ethical issues of our time.

Keywords

acyclicity; Arrow's th; Atkinson's th; capabilities; collective choice; complete ordering; development economics; division of labour; economic measurement; family economics; famines; freedom; gender bias; Gini coefficient; household production; inequality measurement; interpersonal utility comparisons; labour surplus; leximin; liberal paradox; Lorenz curve; maximin; non-dictatorship; nutrition; oligarchy; Pareto principle; partial ordering; Pigou–Dalton condition; poverty lines; poverty measurement; quasi-concavity; quasi-ordering; quasi-transitivity; Sen, A.; social choice; social decision function; social preferences; social welfare function; transitivity; utilitarianism; voting paradoxes; weak equity axiom; welfare economics; welfarism; well-being

Article

Amartya Sen was born in Santiniketan, Bengal in 1933 on the campus of Rabindranath Tagore's Viswa Bharati (both a school and a college) where his maternal grandfather taught Sanskrit and where both his mother and he had been students. At Santiniketan he concentrated on Sanskrit, mathematics and physics, before studying economics (with mathematics minor) at Presidency College, Calcutta and then at Trinity College, Cambridge. His Ph.D. thesis at Cambridge University on the ‘choice of techniques’ was the basis of his election to a competitive Prize Fellowship at Trinity, and this allowed him time to pursue his interests in logic, epistemology, and moral and political philosophy – in addition to economics. Some of his later work in social choice theory and welfare economics would draw on the intersection of these interests, but he has also made major contributions to many of these fields separately.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Sen has taught at several universities: Trinity College, Cambridge; Jadavpur University in Calcutta; the Delhi School of Economics; the London School of Economics; Oxford, where he was Drummond Professor of Political Economy and Fellow of All Souls College; and Harvard, where he is Lamont University Professor, Professor of Economics and Philosophy, and Senior Fellow of the Society of Fellows. He has also held visiting appointments at MIT, Stanford, Berkeley and Cornell. During 1998–2003 Sen was back at Cambridge—as Master of Trinity College. (For further details of his personal and academic life, see Sen, 1999a.) He is widely regarded as one of the finest teachers anywhere, with his lectures and classes always overflowing or oversubscribed. Amartya Sen has been president of the Econometric Society, the American Economic Association, the Indian Economic Association, and the International Economic Association – and he holds more than 90 honorary doctorates. He has won innumerable awards including the Agnelli International Prize in Ethics, the Feinstein World Hunger Award, the Nobel Memorial Prize in Economics, the Bharat Ratna (highest civilian award in India), the Eisenhower Medal, Companion of Honour, UK, and the George C. Marshall Award. His research output is prodigious, and as of 2007 he had published more than 25 books (his works have been translated into more than 30 languages) and about 400 articles. Details of his publications (as well as honorary fellowships and awards) may be found in his CV, at http://www.economics.harvard.edu/faculty/sen/cv.pdf. A list of his publications up to 1998 is available in the *Scandinavian Journal of Economics* (1999).

Sen has made fundamental contributions to several areas in economics including social choice theory, economic measurement, and welfare economics. He has also worked in many other areas, such as axiomatic choice theory, rationality and economic behaviour, development economics, gender and feminist economics, famines and hunger, economic methodology, project evaluation and cost–benefit analysis, and so on. In this brief review of his contributions to economics, I will concentrate on the first three areas, and will also have to overlook his major work in philosophy, history, and on India's economy, society, culture and politics (his CV lists 20 separate areas, of which this review considers just *three*).

As one reviewer put it, Sen is as comfortable writing in the *Journal of Philosophy* or *Philosophy and Public Affairs* as he is in *Econometrica* or the *Economic Journal*. His writings outside economics address some of the most important issues and theories in ethics, justice, and legal and moral philosophy.

In the space available for this review, some areas within economics where Sen's research has had a major impact can only be flagged. These, in general, are more accessible and have been the subject of several other accounts of his work. Thus, in development economics Sen's research began with his Ph.D. thesis, published as *Choice of Techniques* (1968, 3rd edition). It made rigorous a series of arguments that had often been advanced casually concerning the need for ‘appropriate’ (labour-intensive) technology in poor and populous countries. Further work developed the idea of ‘labour surplus’ in a framework that brought the division of labour within the household to economists’ attention. The game-theoretic concepts used in this work were also applied to explain low rates of saving in developing countries. He has also studied important qsts of policy, with research on cost–benefit analysis and accounting prices, including ‘shadow wages’. A summary of his contributions to employment and technological choice in different institutional and labour market contexts may be found in his monograph *Employment, Technology and Development* (1975). A more extensive collection of his writings on development economics, covering many other topics, is available in *Resources, Values and Development* (1984).

Another major contribution was his work on the understanding of famine, drawn together in his book *Poverty and Famines: An Essay on Entitlement and Deprivation* (1981a). He introduced the notion that famines may be due to a decline not in the overall supply of food but in the purchasing power, more generally ‘entitlements’, of vulnerable people – in particular, the poor. The real incomes of a section of the population can decline radically for a variety of reasons (unemployment, relative price changes, production failure, and so on) which can lead to a collapse of their command over food even without any change in *overall* food production and supply. This analysis can provide both a better explanation of famines as well as a more effective approach to the remedying of starvation and hunger – through the regeneration of ‘food entitlements’ (see Drèze and Sen, 1989).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In the area of gender and family economics, Sen (1990) has highlighted the deprivation suffered by females both inside and outside the household. In traditionally unequal societies, gender bias in nutrition, health care and medical attention – as well as women often having to work harder than men – has led Sen to broaden the information base in assessing relative female disadvantage. He has used mortality and survival information to draw attention to the coarsest aspects of gender-related inequality, and has estimated the number of 'missing women' by comparing the female–male ratio with what would be expected in the absence of gender bias in mortality (including female foeticide) (see Sen, 1992b; 2003).

Sen has been instrumental in the recharacterization of development away from the metric of gross national product per capita and its growth. In his book *Development as Freedom* (1999b), he argues that development can be seen as a process of expanding the real freedoms that people enjoy. Freedoms depend not only on individual incomes but on other determinants such as social and economic arrangements, and political and civil rights. He argues that substantive freedoms are not only the primary ends of development but also among its principal means. His related work on 'capabilities' has directly influenced the assessment of development by the United Nations Development Programme in its *Human Development Report*s.

Sen's work in economics combines foundational and theoretical originality, and the willingness to reconsider basic assumptions. His approach is unusual for its breadth of concern coupled with an uncompromising rigour of analysis. It is important to stress the range of his scholarship in an age when the profession is becoming increasingly specialized. For all its diversity there is considerable harmony in his work, with research in one area informing that in another. In this academic biography I shall review his contributions in the areas cited by the Royal Swedish Academy of Sciences in its press release on 14 October 1998 for his award of the 1998 Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel. Under the general heading of 'welfare economics', the citation lists three specific areas: social choice, welfare distributions, and poverty.

Contributions to social choice theory

Sen's book *Collective Choice and Social Welfare* (1970a) is a classic work in the theory of social choice. It is a magisterial study that contains profound and original theorems, including the famous 'liberal paradox' that has spawned a secondary literature of hundreds of articles in learned journals. His analysis of the causes of the various paradoxes of collective choice, including voting and other decision-making procedures, has been central in the economics, philosophy and political theory literatures. His deep insights into Arrow's impossibility theorem, particularly concerning its informational base, have greatly enhanced our understanding of the entire subject of social choice theory. Sen's seminal papers in social choice, including many written after 1970, are reprinted in his book *Choice, Welfare and Measurement* (1982), and later papers are reprinted in his collection *Rationality and Freedom* (2002).

As Arrow (1999, p. 163) states,

> …we cannot do justice to Sen's work in social welfare on the basis of one or two seminal papers, although there have been several such, as I will point out. Rather it is the work as a whole and the way the various parts interplay that must be understood to see the importance of Sen's contribution. His exploration of the notions of social welfare takes place at every level of analysis, formal-mathematical, conceptual, and empirical. It is by far the most comprehensive study of its kind, drawing on profound understanding of both economics and moral philosophy.

Some formal notation is needed to help understand Sen's contributions in social choice theory. Let X be the set of social states, and n the number of individuals in the society. Each individual i (i=1,•2,•…,•n) has a weak preference relation R_i over the set X, where R_i is assumed to be an *ordering*, that is, reflexive, complete and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
transitive. P_i denotes the corresponding strict preference relation. The social weak preference relation is denoted by R, with P its asymmetric factor.
An Arrow *social welfare function* (SWF) is a functional relation that specifies one social ordering for any given n-tuple of individual orderings $\{R_i\}$, one ordering for each person:

$$R = f(\{R_i\}) = f(R_1, R_2, \ldots, R_n).$$

Arrow was concerned with the case in which the value of the function $f(\cdot)$, viz. R, is required to be an ordering, that is, reflexive, complete and transitive.
Four conditions are imposed on the function $f(\cdot)$.

- (U) (*unrestricted or universal domain*). The domain of $f(\cdot)$ includes all possible n-tuples of individual orderings of X.
- (P) (*weak Pareto principle*). For any pair of social states $\{x, y\}$, if for all i: $xP_i{\bullet}y$, then xPy.
- (I) (*pairwise independence of irrelevant alternatives*). The social preference between two alternatives x and y depends only on the individual preferences between x and y (and not on individual preferences over other, 'irrelevant', alternatives).
- (ND) (*non-dictatorship*). There is no individual i such that for all preference n-tuples in the domain of $f(\cdot)$, for each ordered pair (x, y),

$$xP_i y \rightarrow xPy.$$

Denoting the set of individuals in the society as H, Arrow's impossibility theorem can be stated as follows.
Theorem (Arrow): If H is finite and the number of alternatives is at least three, there is no SWF satisfying conditions (U), (P), (I) and (ND).
In other words, a SWF satisfying conditions (U), (P) and (I) must be dictatorial.
This result has turned out to be very robust, and nobody has done more to enhance our understanding of it than Sen. He suggested weakening the requirement of transitivity of the social preference relation R in Arrow's theorem as a way out of the dilemma (Sen, 1969, theorem V). If we demand only *quasi-transitivity* of R, that is, transitivity of the strict preference relation P, it turns out to generate not dictatorship but (still) an oligarchy. Combined with conditions (U), (P) and (I), a reflexive, complete and quasi-transitive social preference relation yields an oligarchy (Sen, 1970a, attributes this result to Allan Gibbard in an unpublished paper). An oligarchy is a group of individuals G who are together decisive (whenever $xP_i{\bullet}y$ for all i in G then xPy) and every member of G has a veto (for any i, $xP_i{\bullet}y$ implies *not* yPx). Thus the replacement of transitivity by quasi-transitivity translates the possibility of dictatorship to an oligarchy with veto powers (that is, a group that is decisive with each person in the group having a veto).
One extreme case of an oligarchy is a dictator (one-person oligarchy). The other extreme makes the oligarchy group include every individual in the society. In that case, the fact that all individuals taken together happen to be decisive is not remarkable; it follows immediately from the Pareto principle. But it also gives every member

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of the society 'veto' power in the sense that if anyone prefers any *x* to any *y*, this precludes the possibility of *y* being socially preferred to *x*.
The oligarchy group will consist of everyone in the society if the condition of anonymity is imposed, that is, the social ranking does not depend on who holds which preferences.

$$\text{(A) (anonymity). If } \{R_i\} \text{ is a permutation of } \{R_i'\}, \text{ then } f(\{R_i\}) = f(\{R_i'\}).$$

Anonymity rules out any proper subset of *H* being oligarchic (including dictatorship), and leaves the whole of *H* as the oligarchy group. Condition (A) together with (U), (P) and (I) are necessary and sufficient for a quasi-transitive *R* to be the 'weak Pareto-extension rule' where *x* is socially preferred to *y* if and only if everyone prefers *x* to *y* (Sen, 1970a, theorem 5*3). Hence *x* and *y* are socially indifferent if they are weak Pareto-incomparable, which covers a much wider variety of cases than their being Pareto-indifferent. The Pareto-extension rule thus gives everyone a 'veto'.
The consequences of further weakening of quasi-transitivity of the social preference relation *R* to *acyclicity*, admitting no cycle of strict preference, have also been explored as a way out of the Arrow problem. Acyclicity is sufficient if we are interested simply in choosing the best element(s) of a subset *S* of *X* – called the choice set *C* (*S*). Sen shows (1970a, lemma 1*1) that if *R* is reflexive and complete, and *X* is finite, then a necessary and sufficient condition for *C*(*S*) to be non-empty for every non-empty subset *S* of *X* is that *R* be *acyclical*.
A *social decision function* (SDF) is defined as a function $f(\cdot)$ that maps *n*-tuples of individual orderings $\{R_i\}$, one ordering for each person, into reflexive, complete and *acyclic* social preference relations *R*. It turns out that, even without quasi-transitivity, the 'veto' result continues to be obtained for an SDF – by supplementing the weaker demand of acyclicity with some other conditions (a variety of such results is discussed in Sen, 1977a). While the existence of vetoers may be less unattractive than that of a dictator, it is, according to Sen (1986, p. 1085), 'unappetizing enough not to provide a grand resolution of the Arrow problem'.
As Sen (1993, p. 507) states,

> … these and other weakenings [of the 'consistency' requirement of the social preference relation *R*] cannot avoid the 'spirit' of Arrow's impossibility theorem. The impossibility can be regenerated through balancing the weakening of 'social preference' by corresponding strengthenings, which are plausible enough, of other conditions, in particular, the non-dictatorship requirement (avoiding not just a dictator, but also an oligarchy, or a vetoer, or a partial vetoer, and so on).

There is a detailed review and assessment of such theorems in Sen (1986).
The discussion so far has imposed some sort of consistency or rationality condition on the social preference relation *R*, for example, transitivity, quasi-transitivity or acyclicity. However, as Sen (1993; 1995) notes, the idea of consistency or rationality applied to 'social preferences' is more difficult to defend than in the case of an individual's preferences. The question thus arises whether dropping the requirement that social choice be based on a (transitive) binary preference relation negates Arrow's theorem reformulated in 'choice-functional' terms. A *functional collective choice rule* (FCCR) makes the value of the function $f(\{R_i\})$ not a social preference relation *R* but a *choice function* *C*(*S*), which specifies for every non-empty set *S* of social states a non-empty subset *C*(*S*) of states chosen from *S*:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$C(S) = f(\{R_i\}).$$

For a FCCR the Arrow conditions (U), (P), (I) and (ND) can be translated in different ways, but the typical translation has taken the form of restricting choices over pairs only. The consistency conditions for the choice function $C(\cdot)$, which are analogous to rationality conditions for the social preference relation R, can be classified or factorized into requirements of two essentially different types – *contraction* consistency and *expansion* consistency. (Sen, 1971, has shown that standard contraction consistency – Property α – and standard expansion consistency – Property γ – are necessary and sufficient for the choice function to be binary – in that its informational content can be exactly captured by a binary relation defined on X.) A series of choice-functional impossibility theorems has been established using alternative forms of such consistency conditions, together with (U), (P), (I) and (ND) translated into choice-functional terms. As in the social-relational case, the conditions imposed on the FCCR (apart from non-dictatorship) can lead it to be dictatorial, oligarchic, or have a vetoer (the results are reviewed in Sen, 1977a; 1986).

In the choice-functional context the robustness of Arrow's impossibility is demonstrated in a remarkable theorem by Sen (1993). He shows (1993, theorem 3) that, without imposing *any* inter-menu consistency condition on the choice function, the Arrow result survives in the form of a dictator who is decisive in rejecting any dispreferred alternative from a given set S, irrespective of the preferences of others. The condition (P) is translated in the choice context as the rejection of Pareto-inferior states; (U) as unrestricted domain applying to the FCCR; (I) as the rejection decisiveness of a group of individuals over any ordered pair (x, y) in S not being affected by individual preferences over pairs other than (x, y), that is, over 'irrelevant' alternatives; and (ND) as there being no individual who is rejection-decisive over a given set S of social states.

Since Sen's proof invokes only one set (or menu) of social states S, no inter-menu consistency condition for social choice is considered. By the same token, the dictator that emerges from combining the modified conditions (U), (P) and (I) with the FCCR is a rejection dictator for the given set S – and not necessarily for another set of social states, which could have a different individual as rejection dictator. The non-dictatorship condition in Sen's theorem is thus stronger than requiring the absence of a single individual who is a rejection dictator across all sets of social states. His theorem identifies exactly how far it is possible to go down the route of weakening and finally eliminating any consistency requirement for the social choice mechanism without escaping Arrow's problem. With no consistency condition on social choice, it is still not possible to avoid an individual who dictates the rejection of every state in a given set S, no matter what the other individuals prefer. Sen's theorem demonstrates just how robust is Arrow's impossibility. The source of Arrow's problem seems to lie elsewhere than in the consistency conditions imposed on the choice function $C(\cdot)$ or the social preference relation R (see above).

Interpersonal comparisons and social welfare functionals

Sen has argued that the real source of the Arrow impossibility problem is 'the tension between the informational eschewal implicitly imposed by Arrow's set of axioms and the demands of discriminating social choice also entailed by the same axioms. The positive possibilities lie, therefore, in making more room for use of information (both utility and non-utility information) in social choice' (1993, p. 514, n. 38). Thus, in the original (social-relational) form of Arrow's theorem, the domain of the social welfare function (SWF) consists of n-tuples of individual *orderings* $\{R_i\}$ only, with no information on valuation of intensity of preferences, or interpersonal comparisons of utility. Sen (1970a) formulated a framework to enrich the informational base to allow different assumptions of measurability and comparability of individual utilities. This has opened up an entire new field

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
with rich and important results.

The informational base of the Arrovian approach can be improved by making the social preference relation R not a function of the n-tuple of individual orderings $\{R_i\}$, but the n-tuple of individual utility functions $\{u_i\bullet(\cdot)\}$. For example, the classical utilitarian characterization of social welfare is a special case of such a form. However, the difficulty with this way of formalizing the functional relations arises from the fact that, given the measurability and comparability assumptions of individual utilities, the utility function has to be represented not by *one* n-tuple of individual utilities, but by a *set* of n-tuples of individual utilities which are *informationally identical* (for the given assumptions of measurability and interpersonal comparability). For example, utility functions can be nominally varied through alternative representations without involving any ‘real’ change (for instance, doubling all the utility numbers), and such informationally equivalent representations should lead to the same social ordering R. This problem is met in Sen's (1970a) approach of *social welfare functionals* through imposing a class of invariance requirements, which demand the same outcome for each of the n-tuples of utility functions in the class that could reflect the same underlying reality – given the chosen characterization of measurability and comparability of utilities.

A social welfare functional (SWFL) specifies exactly one social ordering R over the set X of social states for any given n-tuple $\{u_i\bullet(\cdot)\}$ of individual utility functions defined over X, that is, $R=F(\{u_i\})$. The invariance requirement takes the general form of specifying that, for any two n-tuples of utility functions $\{u_i\}$, $\{u_i{}'\}$ that are related in a particular way (reflecting the assumptions that are made of measurability and interpersonal comparability of individual utilities), the social ordering generated must be the same, that is, $F(\{u_i\})=F(\{u_i{}'\})$. Thus Arrow's SWF corresponds to a SWFL where the invariance class permits monotonic increasing transformations of the individual utility functions $u_i\bullet$; these transformations can be *different* for each i – thus incorporating ordinal non-comparability of the utility functions $\{u_i\}$ which represent the orderings $\{R_i\}$ in Arrow's SWF.

Using this SWFL framework, Sen (1970a, theorem 7*1) showed that utilitarianism requires cardinal ‘unit’ but not ‘level’ comparability (that is, comparability of utility differences but not of levels). He also showed that cardinality of individual utilities without interpersonal comparability does not avoid Arrow's impossibility (1970a, theorem 8*2). He noted that Rawls's maximin criterion only required ordinal interpersonal comparisons ‘to discover who is the worst-off person’, and then the ‘minimal element in the set of individual welfares is maximized’ (Sen, 1970a, pp. 136–7). Many other results were presented by him using the SWFL framework, including several concerning ‘partial comparability’ (Sen, 1970c). The field he opened up has generated a great number of important results, including axiomatic characterizations of utilitarianism, and the maximin (or its lexicographic version, leximin) criterion as a positional dictatorship rule which requires only ‘level’ comparability of ordinal utility functions. Hammond (1976), d'Aspremont and Gevers (1977), Maskin (1978) and Roberts (1980a), among others, have derived many of the results in this area (individual references for other contributions can be found in Sen, 1986). Roberts (1980b) has presented a comprehensive characterization of the welfare functions that result from assuming different degrees of measurability and comparability of individual utilities, including cardinal full comparability and ratio-scale full comparability.

The impossibility of the Paretian liberal

Apart from Arrow's impossibility theorem, perhaps the most well-known result in social choice theory is Sen's (1970b) impossibility of a Paretian liberal. Sen (1979, p. 539) has argued that the Arrow impossibility can be ‘seen as resulting from combining a version of welfarism ruling out the use of non-utility information with making the utility information rmkably poor (particularly in ruling out interpersonal utility comparisons)’. Welfarism entails that two social states are ranked exclusively on the basis of the individual utilities in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
respective states, with no regard to the non-utility features of the states.
It turns out that the impossibility of the Paretian liberal is closely related to the difficulties with welfarism. Paretianism can be seen as essentially a weak form of welfarism, which makes non-utility information redundant in the special case where everyone's utility rankings of two social states coincide. Considerations of liberty require the use of non-utility information, viz. the specification of an individual's ‘protected sphere’ with the social ranking respecting the individual's ranking of states that fall within it. Sen (1970a, theorem 6*1) shows that this use of non-utility information goes not only against welfarism but it can go even against Paretianism. Sen introduces a condition called ‘minimal liberalism’ as follows.

- (L) (*minimal liberalism*). There are at least two individuals such that for each such individual *i* there is a personal domain with at least one pair of social states $\{x, y\}$ such that:

$$xP_iy \rightarrow xPy, \text{ and } yP_ix \rightarrow yPx.$$

Theorem (Sen): There is no social decision function $f(\cdot)$ that satisfies (U), (P) and (L).
Arrow (1999) was greatly impressed by this result, stating that it ‘brilliantly combines simplicity and depth’. He added that he found it

> very surprising [because] both the Pareto judgment and the idea that each individual has some private domain of choice, even if others would make different choices over that domain, are hard to deny; and independence, which on the whole is central to most variations of the Impossibility Theorem, is not assumed here. The paradox arises because ‘nosy’ preferences of others about choices that are in an individual's domain of private choice enter into the Pareto judgment. The result is not only surprising analytically but also addresses profound ethical qsts on the relation between even the vestigial remnant of utilitarianism contained in the Pareto principle and the existence of individual ‘rights’, a scope (however small) over which the individual has complete control. (Arrow, 1999, pp. 165–6)

Sen's ‘impossibility of the Paretian liberal’ is rightly seen as a seminal result in social choice theory. (The choice-functional version of the ‘impossibility of the Paretian liberal’ was originally presented in Sen, 1970a, pp. 81–2, with conditions (U), (P) and (L) translated for a FCCR and with *no* consistency condition imposed on the choice function. A formal proof is also given in Sen, 1993, theorem 2.)

Contributions to economic measurement

Inequality and welfare

The contributions of Sen to the literature on inequality and welfare measurement go back to his classic monograph *On Economic Inequality* (1973a) (*OEI-1973*). This book has been re-issued after a quarter century with a substantial annexe written jointly with James Foster. In the annexe there is a review of the themes in Sen (1973a) that have motivated a great deal of subsequent work in the area and also an acknowledgement that a ‘significant part of *OEI-1973* was, in fact, Atkinson-inspired’ (Sen and Foster, 1997, p. 114). Indeed, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
literature on inequality measurement has been much influenced by Atkinson's celebrated paper on the ranking of income distributions (see also Kolm, 1969). Atkinson (1970) demonstrated the equivalence of three different rankings of income distributions – according to Lorenz dominance, the principle of transfers, and social welfare for all additively separable concave welfare functions.

Formally, let the vector $\mathbf{y}=(y_1,y_2,\ldots,y_n)$ denote an income distribution among n individuals with person i receiving income y_i, for $i=1,2,\ldots,n$. Let the vector $\mathbf{x}=(x_1,x_2,\ldots,x_n)$ denote a different income distribution among the n individuals with the same total income. Atkinson's theorem then shows that the following statements are equivalent:

1. (i) **x** Lorenz dominates **y**;
2. (ii) **x** can be obtained from **y** by a sequence of income transfers from richer to poorer individuals;
3. (iii) $\Sigma_i U(x_i) \geq \Sigma_i U(y_i)$ for any non-decreasing concave function $U(\cdot)$, that is **x** yields as much welfare as **y** for any additively separable, symmetric, non-decreasing, concave social welfare function $(\Sigma_i U(y_i))$.

In fact, a slightly stronger theorem than this can be proved by adopting a weaker criterion for the welfare ranking (iii). The welfare function $\Sigma_i U(y_i)$ can be replaced by a symmetric, non-decreasing, quasi-concave function of individual incomes $W(y_1,y_2,\ldots,y_n)$ (Dasgupta, Sen and Starrett, 1973). The quasi-concavity restriction on the welfare function can be weakened still further. For the theorem to go through, it is clear that the *weakest* requirement on the function is that welfare does *not decrease* by a transfer of income from a richer to a poorer individual. Such a function may be called egalitarian or, as in the literature, S-concave – a property that is weaker than quasi-concavity (Rothschild and Stiglitz, 1973). Thus we consider two further welfare rankings of the income distributions **x** and **y**:

1. (iv) **x** yields as much welfare as **y** for all symmetric, non-decreasing, quasi-concave welfare functions;
2. (v) **x** yields as much welfare as **y** for all symmetric, non-decreasing, egalitarian or S-concave welfare functions.

As shown in Anand (1983, p. 338), the stronger theorem establishing equivalence of (iii) with (iv) and (v) follows as an immediate corollary to Atkinson's theorem. Since the class of additively separable concave functions is contained in the class of quasi-concave functions, which in turn is contained in the class of egalitarian or S-concave functions, it follows that (v) implies (iv), and (iv) implies (iii). But from Atkinson's theorem, (iii) implies (ii), and by the very definition of an egalitarian welfare function, (ii) implies (v); therefore, (iii) implies (v). Hence the chain of implications is complete, with (v)→(iv)→(iii)→(v), and there is no information loss in ranking by welfare functions from the more restrictive additively separable class. It is enough to check that distribution **x** yields as much welfare as distribution **y** for all members of the class of additively separable concave functions, and it will automatically do so for all members of the more general class of S-concave functions too. Hence, Atkinson's theorem in fact establishes the equivalence of (i), (ii) and (v).

Sen's own wide-ranging contributions to the measurement of inequality and welfare begin with *On Economic Inequality* (1973a) and are followed inter alia by Sen (1974; 1976a; 1978; 1992a). In *OEI-1973* he argues for inequality to be seen as a ‘quasi-ordering’ (or partial ordering) without insisting that it must be a complete ordering; he has in fact defended such assertive incompleteness in many different evaluative contexts. Sen (1973a, pp. 72–4) shows that when there are multiple criteria each of which yields a complete ordering (for example, welfare functions or inequality indices in a given class), then their intersection generates a partial ordering. Thus the Lorenz ranking of distributions **x** and **y** (with the same mean income) is their intersection quasi-ordering generated by all welfare functions in the class of additively separable concave functions. Sen

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
initially used the intersection (or unanimity) quasi-ordering approach to measure relative inequality in a positive or descriptive sense by employing different statistical measures of inequality, which avoids ‘exclusive reliance on any particular measure and on the complete ordering generated by it’ (1973a, p. 72). Consider the class of inequality indices that satisfy mean and population-size independence, and the Pigou–Dalton condition (viz. a transfer of income from a richer to a poorer person reduces the value of the inequality index). Then all indices from this class will show less inequality for a Lorenz dominant distribution (Anand, 1983, pp. 339–40), and the intersection quasi-ordering generated by all measures in this class is the Lorenz partial ordering (Sen and Foster, 1997, pp. 142–5).

When Lorenz curves cross, further structure has to be imposed on our social values (or relative inequality indices) – in terms of agreement about sub-classes of welfare functions (inequality indices) or choice of a specific one. Sen provides an axiomatic derivation for a welfare function by drawing on his definition of the Gini coefficient as an affine transformation of the rank-order weighted sum of individual income levels. With individuals labelled in non-descending order of income so that $y_1 \leq y_2 \leq \cdots \leq y_n$, and μ their mean income, Sen (1973a, p. 31) expresses the Gini coefficient as:

$$G = (n+1)/n - (2/n^2\mu)[ny_1 + (n-1)y_2 + \cdots + 2y_{n-1} + y_n] = (n+1)/n - (2/n^2\mu)\sum_i (n+1-i)y_i$$

so that the poorest person receives a weight of n, the ith poorest person a weight of $(n+1-i)$, and the richest (or nth poorest) person a weight of unity. Hence the weights on incomes are based on the *ranking* of individual welfare (assumed to be monotonic in income) levels, which leads to a social welfare ordering of distributions of the same total income by the negative of the Gini coefficient.

The social welfare function behind the Gini coefficient has been used by Sen in different contexts, and has been defended by him in that it ‘makes much use of [ordinal] level comparability without bringing in interpersonally comparable cardinal welfare functions, but at the same time shuns Rawlsian extremism’ (Sen, 1974, p. 398). The rank-order weighting scheme leads him to a distributionally-adjusted measure of real national income (Sen, 1976a), which is the country's mean income μ multiplied by $(1-G)$, where G is the Gini coefficient. The real income, or ‘welfare’ in the space of income, measure $\mu(1-G)$, was the motivation for the development of the generalized Lorenz curve, and Sen's measure is simply twice the area under the generalized Lorenz curve – just as $(1-G)$ is twice the area under the ordinary Lorenz curve.

Poverty measurement

Expressing dissatisfaction with the earlier literature on poverty measurement, Sen (1976b) proposed a new index of poverty which he derived axiomatically. Two types of indices have been used to measure the extent of poverty once the poverty line has been chosen. The most common index is the headcount ratio H, which simply counts the proportion of people below the poverty line. The other index is the income-gap ratio I, which measures the proportionate average income shortfall of the poor from the poverty line. The former index ignores the amounts by which the incomes of the poor fall below the poverty line, while the latter is independent of the number actually in poverty. Both, moreover, are insensitive to a transfer of income from the poor to the very poor. In other words, neither measure is sensitive to the distribution of income among the poor. Sen's measure of poverty incorporates all three of these concerns into a single index.

The index is derived axiomatically after the general form for the poverty measure is taken to be a ‘normalized weighted sum of the income gaps of the poor’. Two axioms then suffice to derive the index. The first specifies

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the weights on the income gaps of the poor, and the second the normalization of the index. Sen argues for weights based on the rank order of the poor below the poverty line, and normalizes his index such that when all the poor have the same income the index is simply $H{\cdot}I$. The weight on the income gap of the ith poorest person is the number of poor people with incomes at least as large as that of person i. From the definition of the Gini coefficient above, this weighting scheme will yield the Gini coefficient of the income distribution among the poor, G_p.

For an ordered income distribution **y** (with $y_1 \leq y_2 \leq \cdots \leq y_n$), let z be the poverty line, q the number of people with income less than or equal to z, and ν the mean income of the poor. Then $H=(q/n)$, $I=(z-\nu)/z$, and Sen's poverty measure

$$P = (q/n)\cdot[z-\nu+\{q/(q+1)\}\nu G_p]/z = (q/n)\cdot[z-\nu(1-G_p)]/z \text{ for large } q = H\cdot[I+(1-I)G_p].$$

The effect of the weighting scheme is to augment the average income gap $(z-\nu)$ by the Gini coefficient G_p times the mean income ν of the poor. Thus, the equivalent of an additional income loss arises when inequality among the poor is taken into account. The correction for this loss involves deflating the mean income ν of the poor by $(1-G_p)$, which yields the familiar equally distributed equivalent income corresponding to the rank-order welfare function. Hence the weighted income gap is calculated by taking the difference not between the poverty line and the mean income of the poor, but between the poverty line and the equally distributed equivalent income of the poor.

Weighting schemes based on welfare functions different from the rank-order function will produce different expressions for equally distributed equivalent income and, by the same token, different measures of inequality. Hence, Sen's approach will generate a different index of poverty for each different welfare function.

Sen's pioneering article has led to an enormous theoretical and empirical literature on poverty measurement. There have been different normalizations of the Sen index, and authors have used different weights on the income gaps of the poor, for example, those that depend on the *size* of a poor person's income gap rather than on the *number* of poor people with higher incomes (as with rank-order weights). The huge theoretical literature that has been generated by the original article is surveyed in Sen and Foster (1997).

Contributions to welfare economics

Utility, income and capability

Sen has criticized utilitarianism from many points of view – among them that it is welfarist and concerned solely with the *sum-total* of individual utilities, not with the interpersonal *distribution* of that sum. Consider the pure distribution problem for two people in which person 1 derives exactly twice as much utility as person 2 from any given level of income, say, because person 2 has some handicap (for example, being a cripple). In this case the utilitarian solution is to give person 1 a higher income than person 2. Even if income were equally divided, person 1 would have enjoyed more utility than person 2 (twice as much); but instead of reducing this inequality the utilitarian rule compounds it – by giving more income to person 1 who is already better off.

The perverse consequences of utilitarianism on inequality in both income and utility spaces led Sen (1973a) to formulate the Weak Equity Axiom (WEA) – which, according to Arrow (1999, p. 167), is another ‘ex of Sen's seminal role in the area of formal theories of social choice’. WEA states that, in distributing a given total

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
income, an individual getting less utility from any level of income should get a higher income. This axiom is in general inconsistent with utilitarianism, and its consequences have been much explored in the social choice literature; for example, similar equity axioms over the set of social states *X* have been used to characterize the leximin rule (the results are discussed in Sen, 1977b; 1986). WEA highlights interest in equality as a criterion independent of utility or income. But equality of what?

What is missing in the utility or income framework is some notion of ‘basic capabilities’: a person being able to do certain basic things. The ability to move about is relevant for a cripple, but there are others such as ‘the ability to meet one's nutritional requirements, the wherewithal to be clothed and sheltered, the power to participate in the social life of the community’ (Sen, 1980, p. 218).

Sen is critical of the use of both income (or commodity possession) and utility as measures of well-being, arguing that they constitute the wrong space in which to make such assessments. He argues that ‘well-being’ has to do with *being* well, which must take into account the capability to live long, be well-nourished, be literate, and so on. As he puts it, the ‘value of the living standard lies in the living, and not in the possessing of commodities, which has derivative and varying relevance’ (Sen, 1987, p. 25). What is valued intrinsically are people's achievements – their ‘beings’ and ‘doings’ – or their ‘capabilities to function’. Income can have importance as an instrument for expanding capabilities, while utility may provide evidence of achievement. But Sen's argument is that the space in which well-being should be assessed has to be more directly linked to what matters most – not its instrumental antecedents (income) nor its evidential correlates (utility).

Poverty as capability deprivation

The conversion of income into achieved well-being is subject to great variation. The mapping of income into basic capabilities is significantly affected by personal heterogeneities (such as age, disability and illness), gender and social roles, environmental and epidemiological conditions, among many other factors. There is thus an important need to go beyond income information in poverty analysis, in particular to see poverty as capability deprivation.

Apart from the basic capabilities involved in leading a minimally acceptable life, such as avoiding hunger or undernourishment or preventable morbidity, there are elementary social functionings, such as ‘appearing in public without shame’ or ‘taking part in the life of the community’. The commodity or income requirements of such capability fulfillment will vary between communities within a country, and between countries. Indeed, as Sen (1983, p. 161) has argued convincingly, ‘poverty is an absolute notion in the space of capabilities’ but it often takes ‘a relative form in the space of commodities’. For example, to avoid shame from the inability to meet the demands of custom or convention in a society, the relative income requirements will typically be higher in richer countries. These requirements vary with what *others* in the community standardly have (for example, in 18th century Europe, according to Adam Smith, ‘a creditable day-labourer would be ashamed to appear in public without a linen shirt’ – cited in Sen, 1983, p. 161). *Relative* deprivation in the space of incomes can thus lead to *absolute* deprivation in the space of capabilities (Sen, 1992a, p. 115).

As Atkinson (1999, p. 186) notes, ‘[i]n this way, Sen exposes the popular confusion that an absolute approach implies a constant real income standard when measuring poverty. Rising living standards in the community as a whole may lead to a poverty line that increases in real terms, even when our concern is limited to absolute deprivation in the space of capabilities.’

A concluding remark

The importance of the subject of welfare economics cannot be overemphasized. Welfare judgements lie at the heart of economic policy analysis and prescription. Yet there has been a tendency to avoid examining welfare

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

statements critically or the assumptions underlying them. Sen's work in welfare economics is of far-reaching significance. His critique of utilitarianism and the Pareto principle, and his insistence on embodying distributional judgements and non-utility information (for example, personal liberty, rights and capabilities), have dramatically shifted the focus of the subject. Sen has done more than anyone else to investigate, scrutinize and develop the foundations of welfare economics and social choice theory. He has used formal analysis, conceptual elucidation, measurement theory, philosophical reasoning, and empirical work to advance the subject – as no one has done before him. The range and reach of Sen's contributions to welfare economics are truly formidable. Their totality essentially defines the present subject matter of this critical branch of economics.

See Also

- Arrow, Kenneth Joseph
- Arrow's theorem
- development economics
- Gini ratio
- inequality (measurement)
- poverty
- social choice
- welfare economics

Selected works

1968. *Choice of Techniques*, 3rd edn. Oxford: Basil Blackwell and Delhi: Oxford University Press.

1969. Quasi-transitivity, rational choice and collective decisions. *Review of Economic Studies* 36, 381–93.

1970a. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

1970b. The impossibility of a Paretian liberal. *Journal of Political Economy* 78, 152–7.

1970c. Interpersonal aggregation and partial comparability. *Econometrica* 38, 393–409.

1971. Choice functions and revealed preference. *Review of Economic Studies* 38, 307–17.

1973a. *On Economic Inequality*. Oxford: Clarendon Press.

1973b. Behaviour and the concept of preference. *Economica* 40, 241–59.

1973. (With P. Dasgupta and D. Starrett.) Notes on the measurement of inequality. *Journal of Economic Theory* 6, 180–7.

1974. Informational basis of alternative welfare approaches: aggregation and income distribution. *Journal of Public Economics* 3, 387–403.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1975. *Employment, Technology and Development*. Oxford: Clarendon Press and New York: Oxford University Press.

1976a. Real national income. *Review of Economic Studies* 43, 19–39.

1976b. Poverty: an ordinal approach to measurement. *Econometrica* 44, 219–31.

1977a. Social choice theory: a re-examination. *Econometrica* 45, 53–89.

1977b. On weights and measures: informational constraints in social welfare analysis. *Econometrica* 45, 1539–72.

1977c. Rational fools: a critique of the behavioural foundations of economic theory. *Philosophy and Public Affairs* 6, 317–44.

1978. Ethical measurement of inequality: some difficulties. In *Personal Income Distribution*, ed. W. Krelle and A.F. Shorrocks. Amsterdam: North-Holland.

1979. Personal utilities and public judgements: or what's wrong with welfare economics? *Economic Journal* 89, 537–58.

1980. Equality of what? In *The Tanner Lectures on Human Values*, Vol. 1, ed. S.M. McMurrin. Salt Lake City: University of Utah Press and Cambridge: Cambridge University Press.

1981a. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Clarendon Press and New York: Oxford University Press.

1981b. Ingredients of famine analysis: availability and entitlements. *Quarterly Journal of Economics* 95, 433–64.

1982. *Choice, Welfare and Measurement*. Oxford: Basil Blackwell.

1983. Poor, relatively speaking. *Oxford Economic Papers* 35, 153–69.

1984. *Resources, Values and Development*. Oxford: Basil Blackwell and Cambridge, MA: Harvard University Press.

1985. *Commodities and Capabilities*. Amsterdam: North-Holland.

1986. Social choice theory. In *Handbook of Mathematical Economics*, Vol. 3, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

1987. *The Standard of Living*, Tanner Lectures with rejoinders by Bernard Williams and others, ed. G. Hawthorne. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1989. (With J. Drèze.) *Hunger and Public Action*. Oxford: Clarendon Press.

1990. Gender and cooperative conflicts. In *Persistent Inequalities*, ed. I. Tinker. New York: Oxford University Press.

1992a. *Inequality Reexamined*. Oxford: Clarendon Press, New York: Russell Sage Foundation and Cambridge, MA: Harvard University Press.

1992b. Missing women. *British Medical Journal* 304 (March), 587–8.

1993. Internal consistency of choice. *Econometrica* 61, 495–521.

1995. Rationality and social choice. *American Economic Review* 85, 1–24.

1997. Maximization and the act of choice. *Econometrica* 65, 745–79.

1997. (Annexe with J.E. Foster.) *On Economic Inequality*. Expanded edition with an annexe '*On Economic Inequality* after a Quarter Century'. Oxford: Clarendon Press and New York: Oxford University Press.

1999a. Autobiography. In *Les Prix Nobel. The Nobel Prizes 1998*, ed. T. Frängsmyr. Stockholm: Nobel Foundation, 1999. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1998/sen-autobio.html, accessed 3 June 2007.

1999b. *Development as Freedom*. New York: Alfred Knopf and Oxford: Oxford University Press.

1999c. The possibility of social choice. *American Economic Review* 89, 349–78.

2002. *Rationality and Freedom*. Cambridge, MA and London, England: Harvard University Press.

2003. Missing women—revisited. *British Medical Journal* 327 (December), 1297–8.

Bibliography

Anand, S. 1983. *Inequality and Poverty in Malaysia: Measurement and Decomposition*. New York: Oxford University Press.

Arrow, K.J. 1999. Amartya K. Sen's contributions to the study of social welfare. *Scandinavian Journal of Economics* 101, 163–72.

Atkinson, A.B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 244–63.

Atkinson, A.B. 1999. The contributions of Amartya Sen to welfare economics. *Scandinavian Journal of Economics* 101, 173–90.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
d'Aspremont, C. and Gevers, L. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44, 199–209.

Hammond, P.J. 1976. Equity, Arrow's conditions, and Rawls' difference principle. *Econometrica* 44, 793–804.

Kolm, S.-C. 1969. The optimal production of social justice. In *Public Economics*, ed. J. Margolis and H. Guitton. London: Macmillan.

Maskin, E.S. 1978. A theorem on utilitarianism. *Review of Economic Studies* 45, 93–6.

Roberts, K.W.S. 1980a. Possibility theorems with interpersonally comparable welfare levels. *Review of Economic Studies* 47, 409–20.

Roberts, K.W.S. 1980b. Interpersonal comparability and social choice theory. *Review of Economic Studies* 47, 421–39.

Rothschild, M. and Stiglitz, J.E. 1973. Some further results on the measurement of inequality. *Journal of Economic Theory* 6, 188–204.

Scandinavian Journal of Economics. 1999. Bibliography of A.K. Sen's publications, 1957–1998. Vol. 101, 191–203.

How to cite this article

Anand, Sudhir. "Sen, Amartya (born 1933)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000454> doi:10.1057/9780230226203.1509

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Senior, Nassau William (1790 – 1864)

N. De Marchi

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Early 19th-century English political economists divided on method. Some, Senior among them, favoured Smith's approach: observe behaviour and try to tie it to likely motives. Others, such as Ricardo and the Mills, preferred to build models based on strong assumptions, then qualify the conclusions while still insisting on their basic truth. On policy questions, and on the basic motives behind economic behaviour, they often agreed, but Smith's approach allowed for more than one answer to questions such as: what is value? Much ink was spilt and much muddle resulted, clarification (on value) being achieved only with Jevons and Marshall.

Keywords

abstinence; balance of trade; Cairnes, J. E.; distribution; efficient allocation; Jones, R.; labour's share of income; love of ease; marginal productivity theory; Marx, K. H.; mercantilism; microfoundations; Mill, J. S.; myopia; paper money; Political Economy Club; Poor Law; profit; rent; Ricardo, D.; Say, J.-B.; science of political economy; Senior, N. W.; Smith, A.; value; wages fund; Whewell, W.

Article

Born at Compton Beauchamp in Berkshire, the eldest son of John Raven Senior, Vicar of Durnford, Nassau Senior studied for the Bar in London, was the first Drummond Professor of Political Economy at Oxford, 1825–30, and was elected to a second term, 1847–52. In 1831 he was appointed Professor of Political Economy at King's College, London, but was soon forced to resign over his controversial recommendation that some of the revenues of the established Church in Ireland be turned over to the Roman Catholics. Senior became a respected adviser: he served on the Commission for inquiring into the Administration and Operation of the Poor Laws (1832–4), being mainly responsible for the writing of its report, and was consulted by Lord John Russell on Irish Poor Law Reform in 1836. In 1841 Senior drew up the Report of the Commission on the condition of the Unemployed Hand-loom Weavers (on

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
this see Stigler, 1949, Lecture 3). Two years after it was founded in 1821 Senior was elected to membership of the Political Economy Club and remained a member, except for the years 1848–53, until his death.

Senior first came to the notice of those conversant with political economy through an article on the Corn Laws in the *Quarterly Review* (1821), and he was a regular contributor to the *Edinburgh Review* from 1841 to 1855. He was at home in both literary and political circles in London and cultivated an interest in Continental affairs via frequent travels and the company of men of influence, among whom were Guizot and De Tocqueville. Conversations with such men in France and Italy and in Ireland were assiduously recorded (and checked for accuracy) and, together with a traveller's observations on these and other countries, filled many journals. Several of these were published, including conversations with De Tocqueville spanning the years 1834–59.

It was Senior's intention to publish a systematic account of political economy, collecting the ideas that were largely scattered in periodicals, lectures, official reports and pamphlets into a major treatise. The plan was not fulfilled and his main printed legacy is his 1836 *Outline of the Science of Political Economy*, plus a collection of pamphlets and public letters published by Augustus Kelley, entitled *Selected Writings on Economics* (1966) and his *Three Lectures on the Rate of Wages* (1830a). S. Leon Levy undertook the work Senior never brought to fruition, in a volume entitled *Industrial Efficiency and Social Economy* (1928). This is a composite, organized by a plan of the editor's own making, and comprising selections from periodical articles, reports and – mainly – manuscript lectures from Senior's second term as Drummond Professor. The work is supposed to represent Senior's ‘mature’ thoughts, but the manner of its composition makes it of very limited value from a scholarly point of view.

Senior belonged to the band of eminent political economists of the second quarter of the 19th century who may be called respectful dissenters from Ricardo's doctrines. He did not dissent on methodological grounds, as did Whewell and Richard Jones (De Marchi and Sturges, 1973; Hollander, 1985, vol. 1, chs 1–3), but on value and distribution he followed Smith and Say more closely than Ricardo. Thus ‘value’ to Senior meant value in exchange rather than cost of production or labour cost (1836, pp. 13–14). And distribution he preferred to treat as a question of factor incomes rather than of functional shares. The expression ‘high wages’, for example, Senior used to stand for high nominal or real wages rather than for a large share of labour's product actually received by labour (1830a, pp. 2–3). This is not to say that he anticipated the marginal productivity theory of distribution, although at least as far as labour and capital are concerned he liked to think of the incomes attributable to each as payment for services rendered or, even more especially, as a ‘reward … [for] conduct’ (1836, p. 89).

Senior's approach to distribution displays tensions that are unavoidable in trying to combine a Ricardian concern with macro-issues such as capital and population growth and the time-path of wages with a Smithian predilection to treat value and distribution as the outcome of voluntary exchanges entered into by individuals. To illustrate: Senior retained the Ricardian theory of rent, whereby rent is an intra-marginal surplus. This surplus accrues to ownership. Where, however, there is competitive access to the powers of nature, the price of the product equals the sum of wages, a reward for labour services, and profit, a return for waiting or ‘abstinence’ (1836, p. 89). This falls short of an integrated theory of distribution, reflecting as it does the institutional fact of appropriation, on the one hand, and economic conduct by free agents, on the other. Similarly, in discussing the time-path of wages, Senior reverted to a wages-fund approach: given labour productivity, ‘the rate of wages depends on the extent of the fund for the maintenance of labourers, compared with the number of labourers to be maintained’ (1830a, p. xii).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
This refers to the average wage; but we are not told how individual contracts struck between worker and employer upon the value of labour services relate to this average. The problem is familiar to modern theorists wanting to make explicit the microfoundations of macroeconomics; though the major difficulty, changing relative wages, Senior neatly sidestepped by following Smith in the conviction that, once established, relative wage scales remained fixed (1830b, p. 15).

Senior's Ricardianism is perhaps most evident in his views on trade and on the international aspects of money. In Three Lectures on the Transmission of the Precious Metals, delivered at Oxford in June 1827 and printed the following year, on an issue he felt to be ‘next to the Reformation, next to the question of free religion, the most momentous that has ever been submitted to human decision’ (1828, p. 88), Senior employed Hume's doctrine (and Ricardo's) to show that no country can have a permanently favourable or unfavourable balance of trade or exchange rate, then used this against the mercantilist view. His main concern is the efficient allocation of labour, and this leads him to the Ricardian conclusion that, if a domestic tax on one industry hurts its international competitiveness, then a ‘countervailing duty’ on the competing import is ‘not a departure from the principles of free trade but an application of them’ (1828, p. 70; compare Ricardo 1822, vol. 4, p. 217). In another set of three lectures, this time On the Cost of Obtaining Money (1830b), Senior addressed the question of international comparisons of wages and argued that the productivity of labour measured in the goods required to import precious metals (in a non-mining country) determines whether wages are high or low. In both sets of lectures Senior discusses paper money and reaches the basic Ricardian conclusion that variations in the amount of the currency, whether metal or paper, may cause sudden disturbances but these will be transitory.

It is worth stressing that the Stoic tradition so evident in Smith's writings – self-respect issuing from prudent behaviour, most notably self-restraint – also infuses Senior's discussions of distribution and related social issues. Senior takes it as given that men tend to be myopic and to prefer taking their ease to working (Senior, 1827, p. 8; 1836, pp. 26 ff.). He therefore considered the supply of goods to be a result of prudential exertions to overcome these obstacles (1836, pp. 15–16). There is no such exertion attached to mere ownership. Hence the income of a landlord is a transfer and categorically distinct from wages and profits, which are properly considered rewards. They are like the good which results from confronting and overcoming unavoidable evil (pain). Although Senior is properly extolled for having glimpsed marginal utility, his contribution on the side of supply and the overcoming of obstacles is at least as interesting.

Marx, it need hardly be added, saw Senior as ‘a mere apologist of the existing order’ (Marx, 1905–10, vol. 3, p. 353). This in part refers to Senior's view that profit is the reward of ‘abstinence’. Senior's sacrifice–reward approach to the sharing out of the price of product at the margin, however, was adopted fully by the later Ricardians, John Stuart Mill (1848, p. 400) and John Elliott Cairnes (1874, p. 74).

Senior's views on method too are indistinguishable in all but two details from John Stuart Mill's. One reasons in political economy from true premises known by introspection (the desire for wealth, subject to a preference for least effort and present enjoyment) or from observation (the laws of return and of population). Being true these premises will, upon correct reasoning, yield true principles. In the science of political economy, therefore, there resides as much certainty as in any science outside axiomatic logic (Senior, 1827, p. 11). Nonetheless, the psychological drives for wealth, leisure and present satisfaction produce counteracting conduct, so that it is difficult to assign motives (causes) to or predict behaviour (1827, p. 9).

Mill tended to bundle the three psychological motives together and make the wealth motive do most of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the work. He also chose to reason hypothetically, as if the desire for wealth were the sole motivation of an individual. This meant that his results could be true only in the abstract, or to the degree to which that assumption was in fact true. In later years Senior, despite his early warning that counteracting motives could not readily be disentangled (1827, p. 9), argued against Mill's simplifying approach. His Four Introductory Lectures from the second period of his tenure of the Drummond Professorship (1847–52) lamented the slow progress made in political economy. He found reason in the Millian manner of dealing with the science. Hypothetical reasoning, in Senior's view, rendered the subject unattractive, because unrealistic, and laid the reasoner open to error, either from forgetting some relevant additional cause or from forgetting that the reasoning itself was based on arbitrary assumptions and was not directly transferable to real world situations (Senior, 1852, pp. 63–5). In more recent discussion, T.W. Hutchison has kept Senior's objections alive, protesting in particular a tendency to confuse tautological with empirical propositions, and to leap straight from abstract models to policy conclusions (Hutchison, 1938). While Senior's concern is understandable, to the extent to which his earlier judgement was well-founded – ‘that we are liable to the greatest mistakes when we endeavour to assign motives [causes] to … conduct’ (1827, p. 9) – it is not clear that he gains anything by his later modified approach, and there is in that case less practical difference between his position and Mill's than some modern commentators have made out (Bowley, 1937, pp. 59–62).

Both Senior and Mill cautioned against applying the principles of political economy without the utmost care and a broad and detailed knowledge of the facts applying to any case in question. Senior, however, nonetheless confidently offered advice on the great issues of the 1830s and 1840s: Poor Law reform; the Factory Act (Ten Hours Bill); agricultural unrest, overpopulation and emigration; free trade and the role of banks in commercial crises. In few of the views he expressed is there anything very remarkable – an exception is noted below in connection with the Ten Hours Bills – but he had a striking ability to cut through to the heart of complex issues, and he had a persuasive pen. To illustrate, his overriding criterion for judging all schemes of relief to the able-bodied poor was whether they destroyed incentives by separating effort from reward (1830a, preface; 1834, pp. 126 ff.). In a particular argument against reducing factory hours, he held that if capital replenishment is, say, 11/12ths of gross turnover, then interest plus profits depend essentially on the last hour of a 12-hour day. A ten-hour day, therefore, would spell ruin (1837, pp. 12–13). This not only presupposed constant returns to hours worked, but confused stocks and flows in the calculation of returns (Johnson, 1969).

See Also

- Jevons, William Stanley
- Marshall, Alfred
- Mill, James
- Mill, John Stuart
- Ricardo, David
- Smith, Adam

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1821. Report on the state of agriculture. *Quarterly Review* 25(50), July.

1827. *An Introductory Lecture on Political Economy*. In Senior (1966).

1828. *Three Lectures on the Transmission of the Precious Metals*. In Senior (1966).

1830a. *Three Lectures on the Rate of Wages*. London: Murray.

1830b. *Three Lectures on the Cost of Obtaining Money*. In Senior (1966).

1834. *Report from his Majesty's Commissioners on the Administration and Practical Operation of the Poor Laws*. London: British Parliamentary Papers.

1836. *An Outline of the Science of Political Economy*. New York: Kelley Reprint, 1965.

1837. *Two Letters on the Factory Acts*. In Senior (1966).

1852. *Four Introductory Lectures on Political Economy*. In Senior (1966).

1872. *Correspondence and Conversations of Alexis De Tocqueville with Nassau William Senior from 1834 to 1859*. 2 vols, ed. M.C.M. Simpson, 2nd edn, London: Henry S. King.

1966. *Selected Writings on Economics. A Volume of Pamphlets 1827–1852*. New York: Kelley Reprint.

Bibliography

Bowley, M. 1937. *Nassau Senior and Classical Economics*. London: Allen & Unwin.

Cairnes, J.E. 1874. *Some Leading Principles of Political Economy Newly Expounded*. London: Macmillan.

De Marchi, N. and Sturges, R.P. 1973. Malthus’ and Ricardo's inductivist critics: four letters to William Whewell. *Economica* 40, 379–93.

Hollander, S. 1985. *The Economics of John Stuart Mill*, 2 vols. Oxford: Blackwell.

Hutchison, T.W. 1938. *The Significance and Basic Postulates of Economic Theory*. London: Macmillan.

Johnson, O. 1969. The ‘last hour’ of Senior and Marx. *History of Political Economy* 1, 359–69.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Levy, S.L., ed. 1928. *Industrial Efficiency and Social Economy* by Nassau W. Senior, 2 vols. New York: Henry Holt & Co.

Marx, K. 1905–10. *Theories of Surplus Values*, 3 vols. London: Lawrence & Wishart, 1969; 1972.

Mill, J.S. 1848. *Principles of Political Economy with Some of their Applications to Social Philosophy*. In *Collected Works of John Stuart Mill*, vols 2 and 3, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Ricardo, D. 1822. On protection to agriculture. In *The Works and Correspondence of David Ricardo*, vol. 4, ed. P. Sraffa with the collaboration of M.H. Dobb, Cambridge: Cambridge University Press for the Royal Economic Society, 1951.

Stigler, G.J. 1949. The classical economists: an alternative view. In *Five Lectures on Economic Problems*. London: Longmans, Green for the London School of Economics and Political Science.

How to cite this article

De Marchi, N. "Senior, Nassau William (1790–1864)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000093> doi:10.1057/9780230226203.1510

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

separability

Charles Blackorby, Daniel Primont and R. Robert Russell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Given a smooth increasing function, two variables are separable from a third if the marginal rate of substitution between the first two variables is independent of the third. In this case it is possible to construct an aggregator function over the first two variables that is independent of the third. This is much weaker than additive separability or additivity. We show that separability is equivalent to decentralization and that additive separability is closely related to two-stage budgeting.

Keywords

additivity; conditional indirect utility function; consumer expenditure; decentralization; demand theory; Gorman, W. M.; Gorman's overlapping th; indirect utility function; Leontief, W.; marginal rates of substitution; optimal income allocation; price aggregation; separability; social welfare function; two-stage budgeting; Sono independence; Strotz, R.

Article

1 Introduction

Separability, as discussed here, refers to certain restrictions on functional representations of consumer (or social) preferences or producer technologies. These restrictions add structure to the decision-making tasks undertaken by economic agents. They also allow the economic researcher to study the behaviour of these agents in a more effective manner.
To keep things simple we will focus on consumers. Consumers must make choices among a large number of goods – both consumption and leisure, both present and future. The apparent complexity of their decision-making problems may lead consumers to engage in simplified budgeting practices. Are these procedures consistent with rational behaviour? In modelling consumer behaviour, economists – theoretical and empirical – employ models that consider only a subset or several subsets of the complete

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
list of goods and services consumed. Can these practices be justified? And, if so, what kinds of restrictions do these rationalizations place on the preferences and the behaviour of the consumer? In what follows, we discuss several proposals that address this class of problems and present their solutions.
One way to think about reducing the complexity of the allocation decision is to imagine that the consumer receives a lump sum of money that he or she first allocates to broad classes of commodities, such as food, shelter and recreation. Detailed decisions about how to spend the money that has been allocated to the food budget are postponed until one is actually in the store buying specific food items. More formally, if the correct amount of money to spend on food commodities, for example, has been allocated, under what circumstances is the consumer able to dispense the food budget among food commodities knowing only the food prices? If the consumer can arrive at an optimal pattern of food expenditures in this way, preferences are said to be *decentralizable*. It is fairly obvious that, if food commodities were separable from all other commodities, decentralization would be possible. What is somewhat surprising, however, is that separability is necessary as well as sufficient for this practice to be rationalized. We first present more formally the concepts of decentralization and separability and then discuss their equivalence.
Consumption bundles are N-tuples, $x=(x_1,\ldots, x_N)$, that are elements of a consumption space, Ω. We take Ω to be a closed and convex subset of $\mathbb{R}^N_+$. Thus we begin with a consumer with a well-defined, neoclassical utility function, $U:\Omega \to \mathbb{R}$. Partition the set of variable indices, $I=\{1, 2,\ldots, N\}$, into two subsets $\{I^1, I^2\}$. This divides the set of goods into two groups, one and two, and we can write the consumption space as $\Omega=\Omega^1\times\Omega^2$ and the consumption vector as $x=(x^1,x^2)$. The consumer faces commodity prices given by $p=(p_1,\ldots, p_N)$ and allocates income y among the N goods. The price vector can also be written as $p=(p^1, p^2)$.
The consumer's utility maximization problem,

$$\max_x U(x) \text{ subject to } p\cdot x = y, \tag{1}$$

can be rewritten as

$$\max_{x^1,x^2} U(x^1, x^2) \text{ subject to } p^1\cdot x^1 + p^2\cdot x^2 = y. \tag{2}$$

The solution to (1) [or (2)] is denoted by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$x^* = (\Phi^1(y, p), \Phi^2(y, p)). \tag{3}$$

Expenditure on goods in group two is denoted by y_2. Group two expenditure is optimal if $y_2 = p^2 \cdot \Phi^2(y, p)$.

2 Decentralization and separability

The consumer's utility maximization problem may be decentralized if the consumer is able to optimally allocate expenditure on group-two commodities knowing only group-two optimal expenditure and group-two prices. More formally, the utility maximization problem is *decentralizable* (for group two) if there exists a function φ^2 such that

$$\varphi^2(y_2, p^2) = \Phi^2(y, p) \text{if } y_2 = p^2 \cdot \Phi^2(y, p).$$

Now the question is: what restriction on the consumer's utility function allows for decentralizability of the utility maximization problem? As it turns out, decentralizability for group-two goods is possible if and only if group-two goods are separable from group-one goods. We turn to the formal definition of separability.

According to the original Leontief–Sono definition of separability, the goods in group two are *separable* from the goods in group one if

$$\frac{\partial}{\partial x_k}\left(\frac{\partial U(x) / \partial x_i}{\partial U(x) / \partial x_j}\right) \equiv 0, \tag{4}$$

for all $i, j \in I^2$, $k \in I^1$ (see Leontief, 1947a; 1947b; Sono, 1945; 1961).

The condition (4) says that marginal rates of substitution between pairs of goods in group two are independent of quantities in group one and hence that an aggregator function exists for group-two goods. Under fairly mild conditions, this aggregator function may be defined by

$$U^2(x^2) := U(0^1, x^2)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

for any choice of a fixed vector, $O^1 \in \Omega^1$; that is, $(O^1, x^2) \in \Omega$. Different choices of O^1 give rise to different aggregator functions that are ordinally equivalent. Having accomplished this, we also define a macro function, $\mathbb{U}$, that is defined by

$$U(x^1, x^2) = \mathbb{U}(x^1, U^2(x^2)). \tag{5}$$

Example: Let I={1,2,3,4}, I^1={1,2}, and I^2={3,4}. Consider the utility function given by

$$U(x_1, x_2, x_3, x_4) = x_1^{1/3} x_3^{1/3} x_4^{1/3} + x_2^{1/2} x_3^{1/4} x_4^{1/4}.$$

By defining $u_2 = U^2(x_3, x_4) = x_3^{1/3} x_4^{1/3}$ and $\mathbb{U}(x_1, x_2, u_2) = x_1^{1/3} u_2 + x_2^{1/2} u_2^{3/4}$, we observe that

$$U(x_1, x_2, x_3, x_4) = \mathbb{U}(x_1, x_2, u_2),$$

and hence that group two is separable from group one. One may also confirm this fact using the Leontief–Sono conditions. The reader should notice that the choice of the aggregator and macro function is not unique. An alternative choice would be $u_2 = U^2(x_3, x_4) = x_3^{1/2} x_4^{1/2}$ and $\mathbb{U}(x_1, x_2, u_2) = x_1^{1/3} u_2^{2/3} + x_2^{1/2} u_2^{1/2}$. In this case, the aggregator function is chosen to be homogeneous of degree one. The reader should also notice that group one is *not* separable from group two; separability is not a symmetric concept.

The formal result on decentralizability can now be stated. The consumer's utility maximization problem is decentralizable for group two if and only if group two is separable from group one – that is, if and only the utility function can be written as

$$U(x^1, x^2) = \mathbb{U}(x^1, U^2(x^2)).$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The implication of this result is that the consumer utility maximization problem may be broken up into two parts:

1. 1. Solve

$$\max_{x^2} U^2(x^2) \text{subject to } p^2 x^2 \leq y_2 \tag{6}$$

to get the conditional demand function, $\phi^2(y_2, p^2)$, for group two.

2. 2. Solve

$$\max_{x^1, y_2} u(x^1, U^2(\varphi^2(y_2, p^2))) \text{subject to } p^1 x^1 + y_2 \leq y \tag{7}$$

to get the optimal demands for group one and the optimal income allocation for group two.

While separability is inherently an asymmetric concept, we may want to consider the symmetric case. We may also want to extend the analysis to consider R groups of goods. To this end, let $\{I^1,\ldots, I^R\}$ be a partition of the original set of variable indices I. The goods vector and the price vector can be written as $x=(x^1,\ldots, x^R)$ and $p=(p^1,\ldots, p^R)$, respectively. Then the utility function is separable in the partition $\{I^1, \ldots, I^R\}$ if and only if the utility function may be written as

$$U(x^1, \ldots, x^R) = u(U^1(x^1), \ldots, U^R(x^R)). \tag{8}$$

Expenditure on goods in group r is denoted by y_r, $r=1,\ldots,R$. Group r expenditure is optimal if $y_r = p^r \cdot \Phi^r(y, p)$.

The consumer's utility maximization problem may be decentralized if the consumer is able to optimally allocate expenditure on group r commodities knowing only group r optimal expenditure and group r prices. More formally, the utility maximization problem is *decentralizable* for the partition $\{I^1,\ldots, I^R\}$ if there exist functions Φ^r, $r=1,\ldots,R$, such that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\varphi^r(y_r, p^r) = \Phi^r(y, p) \text{ if } y_r = p^r \cdot \Phi^r(y, p),\ r = 1, \ldots, R.$$

A useful device for explicating the necessary and sufficient conditions for decentralizability of the utility maximization problem is the conditional indirect utility function. For the R group case it is defined as

$$H(y_1, \ldots, y_R, p) = \max_x \{U(x) : p^r x^r \leq y^r,\ r = 1, \ldots, R\}. \tag{9}$$

This function yields the maximum utility conditional on an income allocation, $(y_1,\ldots,y_R)$, among the R groups. In a second stage one can solve for the optimal income allocation by solving

$$\max_{y_1,\ldots,y_R} H(y_1, \ldots, y_R, p) \text{ subject to } \sum_{r=1}^{R} y^r \leq y. \tag{10}$$

Denote the solution to (10) by the expenditure-allocation functions

$$y_r = \theta^r(y, p),\ r = 1, \ldots, R. \tag{11}$$

A remarkable feature of the conditional indirect utility function is that

$$H(\theta^1(y, p), \ldots, \theta^R(y, p), p) = W(y, p) = \max_x \{U(x) : p \cdot x \leq y\}.$$

Thus, the overall consumer utility maximization may always be performed in the two steps given in (9) and (10) even in the absence of separability restrictions on the utility function.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The results for decentralizability in the R-group symmetric case can be stated in two parts. First, the utility function is separable in the partition $\{I^1,\ldots, I^R\}$ as in (8) if and only if the conditional indirect utility function may be written as

$$H(y_1, \ldots, y_R, p) = U(v^1(y_1, p^1), \ldots, v^R(y_R, p^R))$$

where

$$v^r(y_r, p^r) = \max_{x^r}\left\{U^r(x^r): p^r x^r \leq y_r\right\}, r = 1, \ldots, R, \tag{12}$$

are the conditional indirect utility functions for the aggregator functions, $U^r(x^r)$, r=1,…,R. Second, the consumer's utility maximization problem is decentralizable for the partition $\{I^1,\ldots, I^R\}$ if and only if the utility function is separable in the partition $\{I^1,\ldots,I^R\}$.

Separability in the partition $\{I^1,\ldots, I^R\}$ allows the following two-step budgeting procedure. In the first step the consumer solves the problem in (12), which yields conditional demand functions $x^r = \varphi^r(y_r, p^r), \ r = 1, \ldots, R$. In the second step, the consumer solves (10), which yields the optimal income allocation. While the first step economizes the informational requirements – knowledge of only in-group prices are needed – the second step requires the entire vector of prices. However, additional restrictions on the utility function will reduce both the computational and informational burden of the second step; these restrictions will be discussed shortly.

Decentralization may be possible in more than one partition of the set of commodities. Moreover, the separable sectors might well overlap. Suppose, for example, that U is a social welfare function and x^r is the consumption vector of consumers in group r. As some policies affect only a subset of the population, one might want to analyse separately the social welfare of, say, two subgroups of the population: (1) a particular ethnic group and (2) those in a particular geographical region. Clearly, the two groups could overlap. A deep result of Gorman (1968) indicates that the existence of such overlapping separable groups has powerful implications for the structure of the welfare function.

3 Additive structures

Let I^r and I^c be the two (separable) groups of interest and suppose that $I^1 := I^r - I^s \neq \varnothing$, $I^2 := I^r \cap I^s \neq \varnothing$, and $I^3 := I^s - I^r \neq \varnothing$. Let I^4 be the complement of $I^r \cup I^c$ in In Φ. *Gorman's overlapping th* indicates that I^1, I^2 and I^3 are also separable from their complements in I.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
One of the important aspects of Gorman's overlapping theorem is that this structure is equivalent to the following representation of *U*:

$$U(x) = U(U^1(x^1) + U^2(x^2) + U^3(x^3), x^4). \tag{13}$$

(Additional regularity conditions on the ‘essentiality’ of each group are required for this result; see Gorman, 1968; Blackorby, Primont and Russell, 1978; 1998.)

Note that what drives this result on (groupwise) additive structures is not simply separability of each of the sectors, I^1,I^2 and I^3, from their complements but also separability of arbitrary unions of these sectors. This observation suggests the general result on groupwise-additive structures (see Debreu, 1959; Gorman, 1968):

$$U(x) = \left(\sum_{r=1}^{R} U^r(x^r)\right), R > 2, \tag{14}$$

if and only if arbitrary unions of subsets of the partition $\{I^1,\ldots, I^R\}$ are separable from their complements. Note that a special case of (14) is when each element of the partition *I* is a singleton (*R*=*n*), in which case,

$$U(x) = \left(\sum_{i=1}^{N} U^i(x_i)\right), N > 2. \tag{15}$$

That is, *u* is additive in the variables themselves.

Another example in the context of social welfare functions illustrates the power of Gorman's overlapping theorem. Suppose that the social decision rule satisfies the anonymity condition: only the individuals’ utilities and not their names should matter in social evaluation. This standard social choice assumption is equivalent to symmetry of the social welfare function: permuting the names of individuals does not affect the value of the function. Let {1,2} be the subset of individuals who are affected by some set of policies, and suppose that these policies are judged solely by their effects on these two individuals. This implies that {1,2} is separable from its complement in the set of citizens and hence that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$W(u_1, \ldots, u_R) = J(F(u_1, u_2), u_3, \ldots, u_R). \tag{16}$$

The fact that *W* is symmetric means that it can be written as

$$J(F(u_1, u_2), u_3, \ldots, u_R) = J(F(u_1, u_3), u_2, u_4, \ldots, u_R). \tag{17}$$

This in turn implies that {1,3} is separable from its complement in the set of citizens. However, these sets have a non-empty intersection, {2}, and the overlapping theorem implies that the social welfare function can be written as

$$W(u_1, \ldots, u_R) = J(f^1(u_1) + f^2(u_2) + f^3(u_3), u_4, \ldots, u_R).$$

Proceeding by induction yields a strong implication: the social welfare function must be additive:

$$W(u_1, \ldots, u_R) = W(f^1(u_1) + \ldots f^R(u_R)).$$

The Gorman overlapping argument establishing the groupwise additive structure (13) requires the existence of at least two overlapping separable sets resulting in at least three separable sets. Put differently, separability of arbitrary unions of non-overlapping sets in a binary partition $\{I^1, I^2\}$ adds no restriction to separability of the two sets from their complements. But two-group additivity arises in many contexts. One example is the typical additive utility function in overlapping-generations models where each generation's finite lifetime is divided into two periods (often a ‘work’ period and a ‘retirement’ period). A special case of two-group additivity is the quasi-linear utility function that is so critical to the analysis of public goods and incentive compatibility. The stronger conditions required for two-group additivity are based on Sono's (1945; 1961) independence condition. A set of variables, I^1, is said to be *Sono independent* of I^2 if there exist functions ψ^{ij}, $i, j \in I^1$, such that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\frac{\partial}{\partial x_i}\left(\ln\frac{\partial U(x)/\partial x_k}{\partial U(x)/\partial x_j}\right) = \frac{\partial}{\partial x_i}\left(\ln\frac{\partial U(x)/\partial x_\ell}{\partial U(x)/\partial x_j}\right) = \psi^{jk}(x^1)$$

$$\forall i,\ j \in I^1,\ \forall k,\ \ell \in I^2$$

That is, the effect of changing the consumption of a commodity in sector 1 on the marginal rates of substitution between pairs of variables, one in sector I^1 and the other in sector I^2, depends only on the values of consumption levels in sector 1.

The necessary and sufficient condition for two-group additivity is as follows: U can be written as

$$U(x) = u(U^1(x^1) + U^2(x^2))$$

if and only if I^1 is independent of I^2 and is separable from I^2. (This version of the two-group additivity result is attributable to Blackorby, Primont and Russell, 1978. A somewhat weaker version was proved by Sono, 1945; 1961, who maintained separability of I^2 from I^1 as well.) Although this structure is commonly referred to as ‘separable’, it is clearly stronger than separability as the term has historically been used.

4 Two-stage budgeting

It was emphasized in Section 2 that the principal motivation for separability is that it rationalizes decentralizability of the (possibly complex) expenditure-constrained maximization problem. But, since separability only guarantees that sectoral expenditure is optimally allocated within the sector, full rationalization requires in addition that the optimal amount of money be allocated to each sector. This is accomplished by solving the allocation problem (10).

But solving this problem is tantamount to solving the overall optimization problem and does not seem to reduce the informational requirements. For this reason, Strotz (1957) and Gorman (1959) emphasized the existence of sectoral price aggregates (indexes) that can be used to simplify the first stage of the two-stage optimization problem. Formally, we say that price aggregation is possible if the allocation functions in (11) can be written as

$$y_r = \Theta^r(y, \Pi^1(p^1), \ldots, \Pi^R(p^R)),\quad r = 1, \ldots, R \tag{18}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(where the price-index functions, Π^r, r=1,…,R, are *not* assumed to be homogeneous of degree one or even homothetic).
Price aggregation is equivalent to the following structure for the direct utility function (in an appropriate permutation of the sectoral indices, 1,…, R):

$$U(x) = F\left(\sum_{r=1}^{D} f^r(x^r) + G(f^{D+1}(x^{D+1}), \ldots, f^R(x^R))\right), \tag{19}$$

where each f^r, $r = D+1, \ldots, R$, is homothetic and hence can be normalized to be homogeneous of degree one and the sectoral indirect utility functions dual to the first D aggregator functions, defined by

$$v^r(y_r, p^r) = \max_{x^r}\{U^r(x^r) : p^r x^r \le y_r\}, \; r = 1, \ldots, R, \tag{20}$$

have the structure,

$$v^r(y_r, p^r) = v^r\left(\frac{y_r}{\Pi^r(p^r)}\right) + w^r(p^r), \; r = 1, \ldots, D, \tag{21}$$

where each w^r is homogeneous of degree zero in p^r and Π^r is homogeneous of degree 1 in p^r. (Gorman, 1959, assuming away the troublesome two-group case, proved a restricted version of this result. Exploiting Sono independence and some newer results not available to Gorman in 1959, Blackorby and Russell, 1997, proved the general result, showing that the entire structure needed for two-stage budgeting is, in fact, imbedded in the two-group case.)
An interesting special case of (19) is obtained if D=0:

$$U(x) = F(f^1(x^1), \ldots, f^R(x^R)). \tag{22}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In this case, the first stage of the budgeting algorithm can be expressed as choosing $X^1,\ldots,X^R$ to maximize $F(X^1,\ldots, X^R)$ subject to the budget constraint, $\sum_r \Pi^r(p^r)X^r = y$. This structure allows the two-stage budgeting to be accomplished using only price and quantity aggregates in the first stage.

5 Closing remarks

We have chosen consumer demand theory as a way to illustrate the use of separability in economic analysis. However, separability assumptions, either explicit or implicit, are found in numerous areas in economics. For references to the literature, the reader can consult Blackorby, Primont and Russell (1978; 1998).

See Also

- duality
- Gorman, W.M. (Terence)
- indirect utility function
- Strotz, Robert H.

Bibliography

Blackorby, C., Primont, D. and Russell, R.R. 1978. *Duality, Separability, and Functional Structure: Theory and Economic Applications*. New York and Amsterdam: North-Holland.

Blackorby, C., Primont, D. and Russell, R.R. 1998. Separability: a survey. In *Handbook of Utility Theory*, ed. S. Barbara, P.J. Hammond and C. Seidl. Dordrecht: Kluwer.

Blackorby, C. and Russell, R.R. 1997. Two-stage budgeting: an extension of Gorman's th. *Economic Theory* 9, 185–93.

Debreu, G. 1959. Topological methods in cardinal utility theory. In *Mathematical Methods in the Social Sciences*, ed. K. Arrow, S. Karlin and P. Suppes. Stanford, CA: Stanford University Press.

Gorman, W.M. 1959. Separability and aggregation. *Econometrica* 27, 469–81.

Gorman, W.M. 1968. The structure of utility functions. *Review of Economic Studies* 32, 369–90.

Leontief, W.W. 1947a. A note on the interrelation of subsets of independent variables of a continuous function with continuous first derivatives. *Bulletin of the American Mathematical Society* 53, 343–50.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Leontief, W.W. 1947b. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15, 361–73.

Sono, M. 1945. The effect of price changes on the demand and supply of separable goods [in Japanese]. *Kokumin Keisai Zasshi* 74, 1–51.

Sono, M. 1961. The effect of price changes on the demand and supply of separable goods. *International Economic Review* 2, 239–71.

Strotz, R.H. 1957. The empirical implications of a utility tree. *Econometrica* 25, 269–80.

How to cite this article

Blackorby, Charles, Daniel Primont and R. Robert Russell. "separability." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000095> doi:10.1057/9780230226203.1511

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sequence economies

Ross M. Starr
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A ‘sequence economy’ is a general equilibrium model including markets at a sequence of dates, reopening over time. It is alternative to the Arrow–Debreu model with a full set of futures markets where all exchanges for current and future goods are transacted without transaction cost at a single market date. Sequence economy markets reopen and may be incomplete (some markets, particularly futures, may be inactive) because of transaction costs. The model can provide a microeconomic general equilibrium foundation for the store-of-value function of money, since markets reopening over time create an incentive to carry money and debt intertemporally.

Keywords

Arrow–Debreu model; budget constraint; fiat money; futures markets; general equilibrium; incomplete markets; infinite horizons; intertemporal transfers; overlapping generations models; sequence economies; spot markets; store-of-value function of money; sunspot equilibrium; temporary equilibrium; transaction costs

Article

A ‘sequence economy’ is a general equilibrium model in discrete time including specific provision for the availability of markets at a sequence of dates (Hicks, 1939; Radner, 1972). Markets reopen over time, and at each date firms and households act so that plans and prospects for actions on markets available in the future significantly affect their current actions.
This model is in contrast to the Arrow–Debreu model with a (complete) full set of futures markets (Debreu, 1959). There, all exchanges for current and future goods (including contingent commodities, futures contracts contingent on the realization of uncertain events) are transacted on a market at a single point in time. In the Arrow–Debreu model, there is no need for markets to reopen in the future; economic activity in the future consists simply of the execution of the contracted plans. The Arrow–

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Debreu model with a full set of futures markets appears unsatisfactory in that it denies commonplace observation: futures markets for goods and Arrow securities (contingent contracts payable in money) are not generally available for most dates or a sufficiently varied array of uncertain events; markets do reopen over time. The sequence economy model is an alternative that allows formalization and explanation of these observations.

Several major classes of theoretical model are set in the sequence economy framework: overlapping generations (Balasko and Shell, 1980; 1981; Geanakoplos and Polemarchakis, 1991; Wallace, 1980); temporary equilibrium (Grandmont, 1977; Lucas, 1978), sunspot equilibrium (Chiappori and Guesnerie, 1991), incomplete markets (Geanakoplos, 1990; Magill and Quinzii, 1996). These are general equilibrium models over sequential time emphasizing monetary and financial structure. Each of these areas has a large literature of its own. Typically these models assume a given (incomplete) structure of active financial markets without a detailed foundation for how markets come to be incomplete. This contrasts with the statement of the sequence economy model presented below, which derives market activity and incompleteness endogenously as an equilibrium outcome reflecting transaction costs.

The sequence economy model is particularly suitable to provide a microeconomic foundation for the store-of-value function of money (Hahn, 1971; Starrett, 1973). It is precisely because markets reopen over time that agents may find it desirable to carry abstract purchasing power from one date to succeeding dates. Typically, this will take the form of transactions on spot markets at a succession of dates with money or other financial assets held over time to reflect the (net) excess value of prior sales over purchases. This may occur simply because the model does not provide for futures markets or because futures markets, though available in principle, are in practice inactive. Endogenously determined inactivity of futures markets is the result of transaction costs which tend to make the use of futures markets disproportionately costly compared with spot markets.

There are three principal reasons for the excess cost of futures markets:

1. 1. The necessarily greater complexity of futures contracts may require use of more resources (for example, for record keeping or enforcement) than spot markets.
2. 2. The transaction costs of a futures contract are incurred (partly) at the transaction date, those of an equivalent spot transaction are incurred in the future. The present discounted value of the spot transaction costs incurred in the distant future may be lower than the futures market transaction cost incurred in the present, simply because of time-discounting.
3. 3. Use of a full set of futures markets under uncertainty implies that most contracts transacted become otiose and are left unfulfilled as their effective dates pass and the events on which they were contingent do not occur. There is a corresponding saving in transaction costs associated with reducing the number of transactions required by use of a single spot transaction instead of many contingent commodity contracts, though this reduction may imply a different and inferior allocation of risk-bearing.

We now present a formal pure exchange sequence economy model with transaction costs (Kurz, 1974; Heller and Starr, 1976).

Commodity *i* for delivery at date τ may be bought spot at date τ or futures at any date *t*, $1 \leq t < \tau$. The complete system of spot and futures markets is available at each date (although some markets may be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
inactive). The time horizon is date *K*; each of *H* households is alive at time 1 and cares nothing about consumption after *K*. There are *n* commodities deliverable at each date; in the monetary interpretation of the model spot money is one of the goods. At each date and for each commodity, the household has available the current spot market, and futures markets for deliveries at all future dates. Spot and futures markets will also be available at dates in the future and prices on the markets taking place in the future are currently known. Thus in making his purchase and sale decisions, the household considers without price uncertainty whether to transact on current markets or to postpone transactions to markets available at future dates. There is a sequence of budget constraints, one for the market at each date. That is, for every date, the household faces a budget constraint on the spot and futures transactions taking place at that date, (4) below. The value of its sales to the market at each date (including delivery of money) must balance its purchases at that date.

In addition to a budget constraint, the agent's actions are restricted by a transaction technology. This technology specifies for each complex of purchases and sales at date *t*, what resources will be consumed by the process of transaction. It is because transaction costs may differ between spot and futures markets for the same good that we consider the reopening of markets allowed by the sequence economy model. Specific provision for transaction cost is introduced to allow an endogenous determination of the activity or inactivity of markets. In the special case where all transaction costs are nil, the model is unnecessarily complex; there is no need for the reopening of markets, and the equilibrium allocations are identical to those of the Arrow–Debreu model. Conversely, in the case where some futures markets are prohibitively costly to operate and others are costless, then there is an incomplete array of spot and futures markets and the model is an example of that of Radner (1972).

All of the n-dimensional vectors below are restricted to be non-negative.

- $x_\tau^h(t)$=vector of purchases for any purpose at date *t* by household *h* for delivery at date τ.
- $y_\tau^h(t)$=vector of sales analogously defined.
- $z_\tau^h(t)$=vector of inputs necessary to transactions undertaken at time *t*. The index τ again refers to the date at which these inputs are actually delivered.
- $\omega^h(t)$=vector of endowments at *t* for household *h*.
- $s^h(t)$=vector of goods coming out of storage at date *t*.
- $r^h(t)$=vector of goods put into storage at date *t*.
- $p_\tau(t)$=price vector on market at date *t* for goods deliverable at date τ.

With this notation, $p_{it}(t)$ is the (scalar) spot price of good *i* at date *t*, and $p_{i\tau}(t)$ for $\tau > t$ is the futures price (for delivery at τ) of good *i* at date *t*.

The (non-negative) consumption vector for household *h* is

$$c^h(t) = \omega^h(t) + \sum_{\tau=1}^{t}\left[x_t^h(\tau) - y_t^h(\tau) - z_t^h(\tau)\right] + s^h(t) - r^h(t) \geqq 0, \ (t = 1, \ldots, K). \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

That is, consumption at date t is the sum of endowments plus all purchases past and present with delivery date t minus all sales for delivery at t minus transaction inputs with date t (including those previously committed) plus what comes out of storage at t minus what goes into storage. We suppose that households care only about consumption and not about which market consumption comes from. The household is constrained by its transaction technology, $T^h(t)$, and by its storage technology, $S^h(t)$. $T^h(t)$ specifies the resources, for example, how much leisure time and shoeleather, must be used to carry out a transaction. Let $x^h(t)$ denote the vector of $x_\tau^h(t)$'s [and similarly for $y^h(t)$ and $z^h(t)$]. We insist

$$[x^h(t),\ y^h(t),\ z(t)] \in T^h(t),\ (t = 1, \ldots, K). \tag{2}$$

Naturally, storage input and output vectors must be feasible, so

$$[r^h(t),\ s^h(t+1)] \in S^h(t),\ (t = 1, \ldots, K-1). \tag{3}$$

The budget constraints for household h are then:

$$p(t) \cdot x^h(t) \leqq p(t) \cdot y^h(t),\ (t = 1, \ldots, K). \tag{4}$$

Households may transfer purchasing power forward in time by using futures markets and by storage of goods that will be valuable in the future. Purchasing power may be carried backward by using futures markets. But these may be very costly transactions. In a monetary interpretation of the model, where money and promissory notes are present, the household can either hold money as a store of wealth, or it can buy or sell notes.

Let household h's action at date t be denoted $a^h(t) \equiv [x^h(t)\, y^h(t),\ z^h(t),\ r^h(t),\ s^h(t)]$. Let a^h be a vector of the $a^h(t)$'s, and define x^h, y^h, z^h, r^h and s^h similarly. Define $B^h(p)$ as the set of a^h's which satisfy constraints (1)–(4). The household chooses $a^h(t)$ to maximize $U^h(c^h)$ over $B^h(p)$. Denote the demand correspondence (i.e. the set of maximizing a^h's) by $\gamma^h(p)$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The model can be interpreted as monetary or non-monetary. We think of money as simply a 0th good that does not enter household preferences. Futures contracts in money are discounted promissory notes. $x_{0t}^{h}(t)$ is *h*'s monetary receipts at t, $x_{0\tau}^{h}(t)$ is *h*'s note purchase at *t* due at τ. Money is not treated as numeraire – positivity of its value cannot be assumed – it has a price $p_{0t}(t)$.

The correspondences $Y_t^h(p)$ are always homogeneous of degree zero in $p(t)$, as is seen from the definition of $B^h(p)$. We can therefore restrict the price space to the simplex. Let S^t denote the unit simplex of dimensionality, $n(K–t+1)$. Let $P = X_{t=1}^{K} S^t$, where *X* denotes a Cartesian product.

An equilibrium of the economy is a price vector $p^* \in P$ and an allocation a^{h*}, for each *h*, so that $a^{h*} \in Y^h(p^*)$ for all *h* and

$$\sum_{h=1}^{H} x^{h*} \leqq \sum_{h=1}^{H} y^{h*} \tag{5}$$

(the inequality holds coordinate-wise), where for any good *i*, *t*, τ such that the strict inequality holds in (5) it follows that $p_{i\tau}^*(t) = 0$. The equilibrium of a monetary economy is said to be *non-trivial* (that is, the economy is really monetary) if $p_{0t}^*(t) \neq 0$ for all *t*. Sufficient conditions for existence of equilibrium are continuity and convexity requirements typical of an Arrow–Debreu model appropriately extended. Transaction costs are often thought to be non-convex, leading to approximate equilibrium rather than full equilibrium results (Heller and Starr, 1976).

In the case of fiat (unbacked) money, existence of non-trivial monetary equilibrium requires additional structure designed to maintain positivity of the price of money (boundedness of the price level expressed in monetary terms). This may take a variety of forms: the model may arbitrarily require that fiat money be held or turned in at a finite horizon; households may expect fiat money to be valuable in the future sustaining its value in the present; there may be taxes payable in fiat money. Alternatively, the model may assume an infinite horizon (typical of the overlapping generations model) so that the lack of backing for fiat money need not be experienced (though a nil value of fiat money in equilibrium is still a logical possibility).

In contrast to the Arrow–Debreu economy, a sequence economy equilibrium allocation is not generally Pareto efficient. This is not due simply to the presence of transaction costs; transaction costs technically necessary to a reallocation must be incurred, and they represent no inefficiency. The Arrow–Debreu model, however, uses a lifetime budget constraint. The corresponding constraint here is the sequence of budget constraints in (4). Transfer of purchasing power intertemporally – costless in the Arrow–Debreu model – is here a resource using activity; it requires purchase and sale of assets with resultant transaction cost. But the intertemporal transfer of purchasing power, unlike reallocation of goods among households, is needed not to satisfy technical or consumption requirements but rather to satisfy the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

administrative requirements of sequential budget constraint embodied in (4). Hence technically feasible Pareto-improving reallocations may be prevented in equilibrium by prohibitive transaction costs which would have to be incurred to satisfy the purely administrative requirements of crediting and debiting agents' budgets intertemporally (Hahn, 1971). If trade in monetary instruments is costless, however, then an equilibrium allocation is Pareto efficient (Starrett, 1973). Thus the sequence economy model provides a value-theoretic foundation for the store-of-value role of money.

See Also

- general equilibrium

Bibliography

Balasko, Y. and Shell, K. 1980. The overlapping-generations model, I. The case of pure exchange without money. *Journal of Economic Theory* 23, 281–306.

Balasko, Y. and Shell, K. 1981. The overlapping-generations model. II. The case of pure exchange with money. *Journal of Economic Theory* 24, 112–42.

Chiappori, P.A. and Guesnerie, R. 1991. Sunspot equilibria in sequential markets models. In *Handbook of Mathematical Economics*, vol. 4, ed. W. Hildenbrand and H. Sonnenschein. Amsterdam: North-Holland.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Geanakoplos, J. 1990. An introduction to general equilibrium with incomplete asset markets. *Journal of Mathematical Economics* 19, 1–38.

Geanakoplos, J.D. and Polemarchakis, H.M. 1991. Overlapping generations. In *Handbook of Mathematical Economics*, vol. 4, ed. W. Hildenbrand and H. Sonnenschein. Amsterdam: North-Holland.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45, 535–72.

Hahn, F.H. 1971. Equilibrium with transaction costs. *Econometrica* 39, 417–39.

Heller, W.P. and Starr, R.M. 1976. Equilibrium with non-convex transactions costs: monetary and non-monetary economies. *Review of Economic Studies* 43, 195–215.

Hicks, J.R. 1939. *Value and capital*. Oxford: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kurz, M. 1974. Equilibrium in a finite sequence of markets with transactions cost. *Econometrica* 42, 1–20.

Lucas, Jr., R.E. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–45.

Magill, M. and Quinzii, M. 1996. *Theory of incomplete markets*. Cambridge, MA: MIT Press.

Radner, R. 1972. Existence of equilibrium of plans, prices, and price expectations in a sequence of markets. *Econometrica* 40, 289–303.

Starrett, D.A. 1973. Inefficiency and the demand for ‘money’ in a sequence economy. *Review of Economic Studies* 40, 437–48.

Wallace, N. 1980. The overlapping generations model of money. In *Models of Monetary Economics*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

How to cite this article

Starr, Ross M. "sequence economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000097> doi:10.1057/9780230226203.1512

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sequential analysis

James O. Berger

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Data often arrives sequentially, rather than as a collection. When this occurs – or when the experiment can be designed so that this occurs – there can be a considerable advantage in using statistical methods, called sequential analysis, that are tailored to such situations. Classical frequentist statistical methods require analysis with a pre-specified collection of data, and hence cannot be used in such sequential settings. Interestingly, Bayesian methods can be directly used in sequential settings.

Keywords

Bayesian statistics; dynamic programming; Friedman, M.; sequential analysis; stopping rules; Wald, A.; Wallis, W. A.

Article

Statistical experiments are of either fixed sample or sequential design. A *fixed sample size* experiment is one in which the sample size taken for experimentation is predetermined, while a sequential experiment involves monitoring incoming data to help determine an appropriate time to stop experimentation.

To formalize these notions, suppose the data can be observed one-at-a-time; let X_1, X_2, … denote this possible stream of data. Examples include a series of products coming off an assembly line, a series of missiles being tested for accuracy, and a series of patients participating in a clinical trial.

A key concept is that of a stopping rule, R, which is a description of the manner in which the data stream will be used to determine cessation of the experiment.

Example 1: Consider the stopping rule R_1: stop experimentation after *n* observations have been taken. This stopping rule effectively defines what we earlier called a fixed sample size experiment, since we will take precisely *n* observations.

Example 2: Consider the stopping rules (where $\bar{X}_j = \Sigma_{i=1}^{j} X_i / j$) R_2: stop experimentation if

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
$\bar{X}_{50} > 0.62$ or (failing that) when n=100; R_3: after each new observation, X_j, check whether or not $\bar{X}_j \geq 0.5 + 0.823 / \sqrt{j}$ if so, stop experimentation and otherwise take the next observation. Note that R_2 allows experimentation to stop only after 50 or 100 observations have been taken (this is often called 'group sequential' analysis), while R_3 gives rise to the possibility of stopping after any observation.

To see why stopping rules such as R_2 and R_3 can be desirable, consider a clinical trial investigating a new treatment in which, for the jth participating patient, the observation is a Bernoulli (θ) random variable, X_j, which can assume the values 1 (denoting treatment success) or 0 (denoting treatment failure). Thus θ is the probability of the treatment being successful. Suppose that the standard (old) treatment is known to have a success probability of $\frac{1}{2}$, so it is desired to test the hypothesis (H_0) that $\theta \leq \frac{1}{2}$ (the old treatment is better) versus the hypothesis (H_a) that $\theta > \frac{1}{2}$ (the new treatment is better). A typical fixed sample size test of these hypotheses would proceed by choosing a sample size, say $n = 100$, observing $X_1,\ldots,X_{100}$ from $n = 100$ independent patients, and then rejecting if $\bar{X}_{100} \geq 0.582$. This is an $\alpha = 0.05$ level test. (We make no judgement here concerning the appropriateness of formulating this problem as a statistical hypothesis test.)

Suppose now that the experimenters happen to look at the data after 50 patients have participated in the trial, and observe that, for all 50, the treatment proved successful. This would appear to be overwhelming evidence that the new treatment is better, and would lead reasonable people to stop the clinical trial and recommend adoption of the new treatment. It is a rather surprising fact that this conclusion would be *forbidden* by classical statistics, because the original design called for a sample of size 100. (Classical analyses do not allow deviation from original experimental protocol.) It would have been possible, however, to plan for such a possible eventuality by adopting a sequential design, whereby after every observation (or every few observations) the possibility of stopping is allowed. Indeed, R_2 and R_3 are two such stopping rules, and had either been employed, the above-mentioned clinical trial would certainly have stopped by the time the overwhelming evidence had accumulated.

As indicated in the above example, the advantage of a sequential experimental design is that it allows one to stop the experiment precisely when sufficient evidence has accumulated. The disadvantages of a sequential design are that it can be more expensive (often it is cheaper per observation if the data is collected all at once or in large batches), and that it is harder to analyse from the classical perspective. This last point has to do with the fact that the stopping rule can significantly affect classical statistical measures.

Example 2 (continued): Suppose the stopping rule R_2 had been employed in the clinical trial (that is, an *interim analysis* at the halfway point in the trial had been performed). Also, suppose that, if one did stop after 50 observations (that is, if $\bar{X}_{50} > 0.62$), then H_0 would be rejected, and that, if the trial lasted for all 100 observations (that is, if $\bar{X}_{50} < 0.62$, so that the experiment did not stop at the halfway point), then H_0 would be rejected when $\bar{X}_{100} > 0.582$. It can be shown that, for a *fixed* sample of 50 observations, rejecting H_0 when $\bar{X}_{50} > 0.62$ is an $\alpha = 0.05$ level test, as is rejecting H_0 if $\bar{X}_{100} > 0.582$ for a *fixed* sample size experiment with $n = 100$. For the experiment using R_2, however, it can be shown that the level is $\alpha = 0.095$. (One obtains an error probability larger than each of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

separate $\alpha = 0.05$ because use of R_2 gives 'two chances' to reject H_0.) Thus, if R_1 had been used and $\bar{X}_{100} = 0.582$ had been observed, one could claim significant evidence against H_0 at the $\alpha = 0.05$ level, while if R_2 had been used, one could not claim significance at the $\alpha = 0.05$ level.

It should be mentioned that there is considerable controversy over the issue of whether use of stopping rules should affect statistical conclusions. When classical measures are used, there is no typically a substantial effect. But, interestingly, for certain other statistical measures, such as Bayesian measures, the stopping rule has *no* effect. Thus, employment of the Bayesian approach to statistics allows one to collect data without having to pre-specify a rigid initial stopping rule, greatly increasing the flexibility of experimentation. For discussion of this issue, and support for the Bayesian viewpoint, see Berger and Wolpert (1984) and Berger (1985).

The founder of sequential analysis is generally acknowledged to be Abraham Wald, with Milton Friedman and W. Allen Wallis providing substantial motivational and collaborative support. Early history of sequential analysis is given in Wald (1947), which developed the basic formulation of the problem in terms of stopping rules and analysed a number of basic situations, such as the sequential probability ratio test (for testing between two simple hypotheses). Most of the subsequent work in sequential analysis has focused on either (*a*) evaluating classical measures, such as error probabilities, for special stopping rules (see Siegmund, 1985), or (*b*) determining optimal stopping rules. This last problem is very difficult, and can be rephrased as the problem of deciding if enough information is already available to reach a decision, or if another (or several) observations should be taken. The mathematics of this problem is essentially that of dynamic programming. For general reviews of sequential analysis, see DeGroot (1970), Ghosh (1970), Ghosh, Sen and Mukhopadhyay (1997), Govindarajulu (1981), Berger (1985), Lai (2001), Sen and Ghosh (1991), and Siegmund (1985).

Bibliography

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, J. and Wolpert, R. 1984. *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.

DeGroot, M.H. 1970. *Optimal Statistical Decisions*. New York: McGraw-Hill.

Ghosh, B.K. 1970. *Sequential Tests of Statistical Hypotheses*. Reading, MA: Addison-Wesley.

Ghosh, M., Sen, P.K. and Mukhopadhyay, N. 1997. *Sequential Estimation*. New York: Wiley.

Govindarajulu, Z. 1981. *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation, and Decision Theory*. Columbus: American Science Press.

Lai, T.L. 2001. Sequential analysis: some classical problems and new challenges (with discussion).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Statistica Sinica 11, 303–408.

Sen, P. and Ghosh, B. 1991. *Handbook of Sequential Analysis*. London: Marcel Dekker.

Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer-Verlag.

Wald, A. 1947. *Sequential Analysis*. New York: Wiley.

How to cite this article

Berger, James O. "sequential analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000098> doi:10.1057/9780230226203.1513

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

serial correlation and serial dependence

Yongmiao Hong
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In this article we discuss serial correlation in a linear time series regression context and serial dependence in a nonlinear time series context. We first discuss various tests for serial correlation for both estimated regression residuals and observed raw data. Particular attention is paid to the impact of parameter estimation uncertainty and conditional heteroskedasticity on the asymptotic distribution of test statistics. We discuss the drawback of serial correlation in nonlinear time series models and introduce a number of measures that can capture nonlinear serial dependence and reveal useful information about serial dependence.

Keywords

ARMA models; Durbin–Watson statistic; efficient market hypothesis; entropy; generalized spectral density; homoskedasticity; heteroskedasticity; kernel estimators; Lagrange multipliers; rational expectations; serial correlation; serial dependence; spectral density; statistical inference; time series analysis

Article

1 Introduction

Serial correlation and serial dependence have been central to time series econometrics. The existence of serial correlation complicates statistical inference of econometric models; and in time series analysis, inference of serial correlation, or more generally, serial dependence, is crucial to characterize the dynamics of time series processes. Lack of serial correlation is also an important implication of many economic theories and economic hypotheses. For example, the efficient market hypothesis implies that asset returns are an martingale difference sequence (*m.d.s.*), and so are serially uncorrelated. More generally, rational expectations theory implies that the expectational errors of the economic agent are serially uncorrelated. In this article we first discuss various tests for serial correlation, for both estimated model residuals and observed raw data, and we discuss their relationships. We then discuss serial dependence in a nonlinear time series context, introducing related measures and tests for serial dependence.

2 Testing for serial correlation

Consider a linear regression model

$$Y_t = X_t'\beta^0 + \varepsilon_t,\ t = 1, \cdots, n, \tag{2.1}$$

where Y_t is a dependent variable, X_t is a $k\times1$ vector of explanatory variables, β^0 is an unknown $k\times1$ parameter vector, and ε_t is an unobservable disturbance with $E(\varepsilon_t|X_t)=0$. Suppose X_t is strictly exogenous such that $\mathrm{cov}(X_t, \varepsilon_s)=0$ for all t, s. Then (2.1) is called a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

static regression model. If X_t contains lagged dependent variables, (2.1) is called a dynamic regression model.

For a linear dynamic regression model, serial correlation in $\{\varepsilon_t\}$ will generally render inconsistent the OLS estimator. To see this, we consider an AR(1) model

$$Y_t = \beta_0^0 + \beta_1^0 Y_{t-1} + \varepsilon_t = X_t' \beta^0 + \varepsilon_t,$$

where $X_t=(1, Y_{t-1})'$. If $\{\varepsilon_t\}$ also follows an AR(1) process, we will have $E(X_t \varepsilon_t) \neq 0$, rendering inconsistent the OLS estimator for β^0. It is therefore important to check serial correlation for estimated model residuals, which serves as a misspecification test for a linear dynamic regression model. For a static linear regression model, it is also useful to check serial correlation. In particular, if there exists no serial correlation in $\{\varepsilon_t\}$ in a static regression model, then there is no need to use a long-run variance estimator of the OLS estimator $\hat{\beta}$ (for example, Andrews, 1991; Newey and West, 1987).

2.1 Durbin – Watson test

Testing for serial correlation has been a longstanding problem in time series econometrics. The most well known test for serial correlation in regression disturbances is the Durbin–Watson test, which is the first formal procedure developed for testing first order serial correlation

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \ \{u_t\} \sim i.i.d.(0, \sigma^2)$$

using the OLS residuals $\{e_t\}_{t=1}^n$ in a static linear regression model. Durbin and Watson (1950; 1951) propose a test statistic

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}.$$

Durbin and Watson present tables of bounds at the 0.05, 0.025 and 0.01 significance levels of the d statistic for static regressions with an intercept. Against the one-sided alternative that $\rho > 0$, if d is less than the lower bound d_L, the null hypothesis that $\rho = 0$ is rejected; if ρ is greater than the upper bound d_U, the null hypothesis is accepted. Otherwise, the test is equivocal. Against the one-sided alternative that $\rho < 0$, $4-d$ can be used to replace d in the above procedure.

The Durbin–Watson test has been extended to test for lag 4 autocorrelation by Wallis (1972) and for autocorrelation at any lag by Vinod (1973).

2.2 Durbin's h test

The Durbin–Watson d test is not applicable to dynamic linear regression models, because parameter estimation uncertainty in the OLS estimator $\hat{\beta}$ will have nontrivial impact on the distribution of d. Durbin (1970) developed the so-called h test for first-order autocorrelation in $\{\varepsilon_t\}$ that takes into account parameter estimation uncertainty in $\hat{\beta}$. Consider a simple dynamic linear regression model

$$Y_t = \beta_0^0 + \beta_1^0 Y_{t-1} + \beta_2^0 X_t + \varepsilon_t,$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where X_t is strictly exogenous. Durbin's h statistic is defined as:

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n\hat{\mathrm{var}}(\hat{\beta}_1)}},$$

where $\hat{\mathrm{var}}(\hat{\beta}_1)$ is an estimator for the asymptotic variance of $\hat{\beta}_1$, $\hat{\rho}$ is the OLS estimator from regressing e_t on e_{t-1} (in fact, $\hat{\rho} \approx 1 - d/2$). Durbin (1970) shows that $h \xrightarrow{d} N(0, 1)$ as $n \to \infty$ under null hypothesis that $\rho = 0$.

2.3 Breusch–Godfrey test

A more convenient and generally applicable test for serial correlation is the Lagrange multiplier test developed by Breusch (1978) and Godfrey (1978). Consider an auxiliary autoregression of order p:

$$\varepsilon_t = \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j} + z_t, t = p + 1, \cdots, n. \tag{2.2}$$

The null hypothesis of no serial correlation implies $\alpha_j = 0$ for all $1 \le j \le p$. Under the null hypothesis, we have $nR^2_{uc} \xrightarrow{d} \chi^2_p$, where R^2_{uc} is the uncentred R^2 of (2.2). However, the autoregression (2.2) is infeasible because ε_t is unobservable. One can replace ε_t with the OLS residual e_t:

$$e_t = \sum_{j=1}^{p} \alpha_j e_{t-j} + v_t, t = p + 1, \cdots, n.$$

Such a replacement, however, may contaminate the asymptotic distribution of the test statistic because $e_t = \varepsilon_t - (\hat{\beta} - \beta)' X_t$ contains the estimation error $(\hat{\beta} - \beta)' X_t$ where X_t may have nonzero correlation with the regressors e_{t-j} for $1 \le j \le p$ in dynamic regression models. This correlation affects the asymptotic distribution of nR^2_{uc} so that it will not be χ^2_p. To purge this impact of the asymptotic distribution of the test statistic, one can consider the augmented auxiliary regression

$$e_t = X_t' \gamma + \sum_{j=1}^{p} \alpha_j e_{t-j} + v_t, t = p + 1, \cdots, n. \tag{2.3}$$

The inclusion of X_t will capture the impact of estimation error $(\hat{\beta} - \beta)' X_t$. As a result, the test statistic $nR^2 \xrightarrow{d} \chi^2_p$ under the null

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

hypothesis, where, assuming that X_t contains an intercept, R^2 is the centred squared multi-correlation coefficient in (2.3). For a static linear regression model, it is not necessary to include X_t in the auxiliary regression, because $\{X_t\}$ and $\{\varepsilon_t\}$ are uncorrelated, but it does not harm the size of the test if X_t is included. Therefore, the nR^2 test is applicable to both static and dynamic regression models. We note that Durbin's h test is asymptotically equivalent to the nR^2 test of (2.3) with p=1.

2.4 Box – Pierce – Ljung test

In time series ARMA modelling, Box and Pierce (1970) propose a portmanteau test as a diagnostic check for the adequacy of an ARMA model

$$Y_t = \psi_0 + \sum_{j=1}^{r} \psi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \left\{\varepsilon_t\right\} \sim i.i.d.(0, \sigma^2). \tag{2.4}$$

Suppose e_t is an estimated residual obtained from a maximum likelihood estimator. One can define the residual sample autocorrelation function

$$\hat{\rho}(j) = \frac{\hat{\gamma}(j)}{\hat{\gamma}(0)}, \; j = 0, \pm 1, \cdots, \pm (n-1),$$

where $\hat{\gamma}(j) = n^{-1} \sum_{t=|j|+1}^{n} e_t e_{t-|j|}$ is the residual sample autocovariance function.
Box and Pierce (1970) propose a portmanteau test

$$Q \equiv n \sum_{j=1}^{p} \hat{\rho}^2(j) \xrightarrow{d} \chi^2_{p-(r+q)},$$

where the asymptotic χ^2 distribution follows under the null hypothesis of no serial correlation, and the adjustment of degrees of freedom $r+q$ is due to the impact of parameter estimation uncertainty for the r autoregressive coefficients and q moving average coefficients in (2.4).
To improve small sample performance of the Q test, Ljung and Box (1978) propose a modified Q test statistic:

$$Q^* \equiv n(n+2) \sum_{j=1}^{p} (n-j)^{-1} \hat{\rho}^2(j) \xrightarrow{d} \chi^2_{p-(r+q)}.$$

The modification matches the first two moments of Q^* with those of the χ^2 distribution. This improves the size in small samples, although not the power of the test.

The Q test is applicable to test serial correlation in the OLS residuals $\{e_t\}$ of a linear static regression model, with $Q \xrightarrow{d} \chi^2_p$ under the null hypothesis. Unlike for ARMA models, there is no need to adjust the degrees of freedom for the χ^2 distribution because the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

estimation error $(\hat{\beta}-\beta)'X_t$ has no impact on it, due to the fact that cov(X_t, ε_s)=0 for all *t*, *s*. In fact, it could be shown that the nR^2 and *Q* statistics are asymptotically equivalent under the null hypothesis. However, when applied to the estimated residual of a dynamic regression model which contains both endogenous and exogenous variables, the asymptotic distribution of the *Q* test is generally unknown (Breusch and Pagan, 1980). One solution is to modify the *Q* test statistic as follows:

$$\hat{Q} \equiv n\hat{\rho}'(I-\hat{\Phi})^{-1}\hat{\rho} \xrightarrow{d} \chi^2_p \text{ as } n \to \infty,$$

where $\hat{\rho} = [\hat{\rho}(1), \cdots, \hat{\rho}(p)]'$ and $\hat{\Phi}$ captures the impact caused by nonzero correlation between $\{X_t\}$ and $\{\varepsilon_s\}$. See Hayashi (2000, Section 2.10) for more discussion.

2.5 Spectral density-based test

Much criticism has been levelled at the possible low power of the Box–Pierce–Ljung portmanteau tests, which also applies to the nR^2 test, due to the asymptotic equivalence between the *Q* test and the nR^2 test for a static regression. Moreover, there is no theoretical guidance on the choice of *p* for these tests. A fixed lag order *p* will render inconsistent any test for serial correlation of unknown form. To test serial correlation of unknown form in the estimated residuals of a linear regression model, which can be static or dynamic, Hong (1996) uses a kernel spectral density estimator

$$\hat{h}(\omega) = \frac{1}{2\pi}\sum_{j=1-n}^{n-1} k(j/p)\hat{\gamma}(j)e^{-ij\omega}, \ \omega \in [-\pi, \pi],$$

and compares it with the flat spectrum implied by the null hypothesis of no serial correlation:

$$\hat{h}_0(\omega) = \frac{1}{2\pi}\hat{\gamma}(0), \ \omega \in [-\pi, \pi].$$

Under the null hypothesis, $\hat{h}(\omega)$ and $\hat{h}_0(\omega)$ are close. If $\hat{h}(\omega)$ is significantly different from $\hat{h}_0(\omega)$ there is evidence of serial correlation. A global measure of the divergence between $\hat{h}(\omega)$ and $\hat{h}_0(\omega)$ is the quadratic form

$$L^2(\hat{h}, \hat{h}_0) \equiv \int_{-\pi}^{\pi}\left[\hat{h}(\omega)-\hat{h}_0(\omega)\right]^2 d\omega = \sum_{j=1}^{n-1} k^2(j/p)\hat{\gamma}^2(j).$$

The test statistic is a normalized version of the quadratic form:

$$M_o \equiv \left[n\sum_{j=1}^{n-1} k^2(j/p)\hat{\rho}^2(j) - \hat{C}_o(p)\right] / \sqrt{\hat{D}_o(p)} \xrightarrow{d} N(0, 1)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where the centring and scaling factors

$$\hat{C}_o(p) = \sum_{j=1}^{n-1}(1 - j/n)k^2(j/p),$$

$$\hat{D}_o(p) = 2\sum_{j=1}^{n-2}(1 - j/n)[1 - (j+1)/n]k^4(j/p).$$

This test can be viewed as a generalized version of Box and Pierce's (1970) portmanteau test, the latter being equivalent to using the truncated kernel $k(z)=\mathbf{1}(|z|\leq 1)$, which gives equal weighting to each of the first p lags. In this case, M_o is asymptotically equivalent to

$$M_T \equiv \frac{n\Sigma_{j=1}^{p}\hat{\rho}^2(j) - p}{\sqrt{2p}} \xrightarrow{d} \frac{\chi_p^2 - p}{\sqrt{2p}} \sim N(0,1) \text{ as } p \to \infty.$$

However, uniform weighting to different lags may not be powerful when a large number of lags is employed. For any weakly stationary process, the autocovariance function $\gamma(j)$ typically decays to 0 as lag order j increases. Thus, it is more efficient to discount higher order lags. This can be achieved by using non-uniform kernels. Most commonly used kernels, such as the Bartlett, Pazren and quadratic-spectral kernels, discount higher order lags. Hong (1996) shows that the Daniell kernel $k(z)=\sin(\pi z)/(\pi z)$, $-\infty<z<\infty$, maximizes the power of the M test over a wide class of the kernel functions when $p\to\infty$. The optimal kernel for hypothesis testing differs from the optimal kernel for spectral density estimation.

It is important to note that the spectral density test M applies to both static and dynamic regression models, and no modification is needed when applied to a dynamic regression model. Intuitively, parameter estimation uncertainty causes some adjustment of degrees of freedom, which becomes asymptotically independent when the lag order $p\to\infty$ as $n\to\infty$. This differs from the case where p is fixed.

For similar spectral density-based tests for serial correlation, see Paparoditis (2000), Chen and Deo (2004), and Fan and Zhang (2004).

2.6 Heteroskedasticity – robust tests

All the aforementioned tests assume conditional homoskedasticity or even *i.i.d.* on $\{\varepsilon_t\}$. This rules out high frequency financial time series, which have been documented to have persistent volatility clustering. Some effort has been devoted to robustifying tests for serial correlation. Wooldridge (1990; 1991) proposes a two-stage procedure to robustify the nR^2 test for serial correlation in estimated residuals $\{e_t\}$ of a linear regression model (2.1): (i) regress $(e_{t-1},\ldots,e_{t-p})$ on X_t and save the estimated $p\times 1$ residual vector $\hat{v}_t$; (ii) regress 1 on $\hat{v}_t e_t$ and obtain SSR, the sum of squared residuals; (iii) compare the $n-SSR$ statistic with the asymptotic χ_p^2 distribution. The first auxiliary regression purges the impact of parameter estimation uncertainty in the OLS estimator $\hat{\beta}$ and the second auxiliary regression delivers a test statistic robust to conditional heteroskedasticity of unknown form.

Whang (1998) also proposes a semiparametric test for serial correlation in estimated residuals of a possibly nonlinear regression model. Assuming that $\varepsilon_t=\sigma[Z_t(\alpha)]z_t$, where $\{z_t\}\sim$i.i.d.(0,1), var$(\varepsilon_t|I_{t-1})=\sigma^2[Z_t(\alpha)]$ depends on a random vector with fixed dimension (for example, $Z_t(\alpha) = (\varepsilon_{t-1}^2, \cdots, \varepsilon_{t-K}^2)'$ for a fixed K), but the functional form $\sigma^2(\cdot)$ is unknown. This covers a variety of conditionally heteroskedastic processes, although it rules out non-Markovian processes such as Bollerslev's (1986) GARCH model.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Whang (1998) first estimates $\sigma^2[Z_t(\alpha)]$ using a kernel method, and then constructs a Box–Pierce type test for serial correlation in the estimated regression residuals standardized by the square root of the nonparametric variance estimator.

The assumption imposed on $\mathrm{var}(\varepsilon_t|I_{t-1})$ in Whang (1998) rules out GARCH models, and both Wooldridge (1991) and Whang (1998) test serial correlation up to a fixed lag order only. Hong and Lee (2007) have recently robustified Hong's (1996) spectral density-based consistent test for serial correlation of unknown form:

$$\widehat{M} \equiv \left[n^{-1} \sum_{j=1}^{n-1} k^2(j/p)\hat{\gamma}^2(j) - \hat{C}(p) \right] / \sqrt{\hat{D}(p)},$$

where the centring and scaling factors

$$\hat{C}(p) \equiv \hat{\gamma}^2(0) \sum_{j=1}^{n-1} (1 - j/n) k^2(j/p) + \sum_{j=1}^{n-1} k^2(j/p) \hat{\gamma}_{22}(j),$$

$$\hat{D}(p) \equiv 2\hat{\gamma}^4(0) \sum_{j=1}^{n-2} (1 - j/n)[1 - (j+1)/n] k^4(j/p) + 4\hat{\gamma}^2(0) \sum_{j=1}^{n-2} k^4(j/p) \hat{\gamma}_{22}(j) + 2 \sum_{j=1}^{n-2} \sum_{l=1}^{n-2} k^2(j/p) k^2(l/p) \hat{C}^2(0, j, l),$$

with $\hat{\gamma}_{22}(j) \equiv n^{-1} \Sigma_{t=j+1}^{n-1} [e_t^2 - \hat{\gamma}(0)][e_{t-j}^2 - \hat{\gamma}(0)]$ and $\hat{C}(0, j, l) \equiv n^{-1} \Sigma_{t=\max(j,l)+1}^{n} [e_t^2 - \hat{\gamma}(0)] e_{t-j} e_{t-l}$. Intuitively, the centring and scaling factors have taken into account possible volatility clustering and asymmetric features of volatility dynamics, so the $\widehat{M}$ test is robust to these effects. It allows for various volatility processes, including GARCH models, Nelson's (1991) EGARCH, and Glosten, Jagannathan and Runkle's (1993) Threshold GARCH models.

Martingale tests

Several tests for serial correlation are motivated for testing the *m.d.s.* property of an observed time series $\{Y_t\}$, say asset returns, rather than estimated residuals of a regression model. We now present a unified framework to view some martingale tests for observed data.

Extending an idea of Cochrane (1988), Lo and MacKinlay (1988) first rigorously present an asymptotic theory for a variance ratio test for the *m.d.s.* hypothesis of $\{Y_t\}$. Because the *m.d.s.* hypothesis implies $\gamma(j)=0$ for all $j>0$, one has

$$\frac{\mathrm{var}\left(\Sigma_{j=1}^{p} Y_{t-j}\right)}{p \cdot \mathrm{var}(Y_t)} = \frac{p\gamma(0) + 2p\Sigma_{j=1}^{p}(1 - j/p)\gamma(j)}{p\gamma(0)} = 1.$$

This unity property of the variance ratio can be used to test the *m.d.s.* hypothesis because any departure from unity is evidence against the *m.d.s.* hypothesis.

The variance ratio test is essentially based on the statistic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\mathrm{VR}_0 \equiv \sqrt{n/p}\sum_{j=1}^{p}(1-j/p)\hat{\rho}(j) = \frac{\pi}{2}\sqrt{n/p}\left[\hat{f}(0)-\frac{1}{2\pi}\right],$$

where $\hat{f}(0)$ is a kernel-based normalized spectral density estimator at frequency 0, with the Bartlett kernel $k(z)=(1-|z|)\,\mathbf{1}\,(|z|\leq 1)$ and a lag order p. In other words, VR_0 is based on a spectral density estimator of frequency 0, and because of this, it is particularly powerful against long memory processes, whose spectral density at frequency 0 is infinity (see Robinson, 1994, for an excellent survey).

Under the *m.d.s.* hypothesis with conditional homoskedasticity, Lo and MacKinlay (1988) show that for any fixed p,

$$\mathrm{VR}_0 \xrightarrow{d} N[0, 2(2p-1)(p-1)/3p] \;\; as\; n\to\infty.$$

Lo and MacKinlay (1988) also consider a heteroskedasticity-consistent variance ratio test:

$$\mathrm{VR} \equiv \sqrt{n/p}\sum_{j=1}^{p}(1-j/p)\hat{\gamma}(j)/\sqrt{\hat{\gamma}_2(j)},$$

where $\hat{\gamma}_2(j)$ is a consistent estimator for the asymptotic variance of $\hat{\gamma}(j)$ under conditional heteroskedasticity. Lo and MacKinlay (1988) assume a fourth order cumulant condition that

$$E[(Y_t-\mu)^2(Y_{t-j}-\mu)(Y_{t-l}-\mu)] = 0,\;\; j, l > 0,\;\; j\neq l. \tag{2.5}$$

Intuitively, this condition ensures that the sample autocovariances at different lags are asymptotically uncorrelated; that is, $\mathrm{cov}\left[\sqrt{n}\hat{\gamma}(j), \sqrt{n}\hat{\gamma}(l)\right]\to 0$ for all $j\neq l$. As a result, the heroskedasticity-consistent VR has the same asymptotic distribution as VR_0. However, the condition in (2.5) rules out many important volatility processes, such as EGARCH and Threshold GARCH models. Moreover, the variance ratio test only exploits the implication of the *m.d.s.* hypothesis on the spectral density at frequency 0; it does not check the spectral density at nonzero frequencies. As a result, it is not consistent against serial correlation of unknown form. See Durlauf (1991) for more discussion.

Durlauf (1991) considers testing the *m.d.s.* hypothesis for observed raw data $\{Y_t\}$, using the spectral distribution function

$$H(\lambda) \equiv 2\int_0^{\pi\lambda} h(\omega)\,d\omega = \gamma(0)\lambda + \sqrt{2}\sum_{j=1}^{\infty}\gamma(j)\frac{\sqrt{2}\sin(j\pi\lambda)}{j\pi},\; \lambda\in[0, 1],$$

where $h(\omega)$ is the spectral density of $\{Y_t\}$:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$h(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma(j)\cos(j\omega),\ \omega\in[-\pi,\pi].$$

Under the *m.d.s.* hypothesis, $H(\lambda)$ becomes a straight line:

$$H_0(\lambda) = \gamma(0)\lambda,\ \ \lambda\in[0,1].$$

An *m.d.s.* test can be obtained by comparing a consistent estimator for $H(\lambda)$ and $\hat{H}_0(\lambda) = \hat{\gamma}(0)\lambda$.
Although the periodogram (or sample spectral density function)

$$\hat{I}(\omega) \equiv \frac{1}{2\pi n}\left|\sum_{t=1}^{n}(Y_t-\bar{Y})e^{it\omega}\right|^2 = \frac{1}{2\pi}\sum_{j=1-n}^{n-1}\hat{\gamma}(j)e^{-ij\omega}$$

is not consistent for the spectral density $h(\omega)$, the integrated periodogram

$$\hat{H}(\lambda) \equiv 2\int_0^{\lambda\pi}\hat{I}(\omega)\,d\omega = \hat{\gamma}(0)\lambda + \sqrt{2}\sum_{j=1}^{n-1}\hat{\gamma}(j)\frac{\sqrt{2}\sin(j\pi\lambda)}{j\pi}$$

is consistent for $H(\lambda)$, thanks to the smoothing provided by the integration. Among other things, Durlauf (1991) proposes a Cramer–von Mises type statistic

$$CVM \equiv \frac{1}{2}n\int_0^1\left[\hat{H}(\lambda)/\hat{\gamma}(0)-\lambda\right]^2 d\lambda = n\sum_{j=1}^{n-1}\hat{\rho}^2(j)/(j\pi)^2.$$

Under the *m.d.s.* hypothesis with conditional homoskedasticity, Durlauf (1991) shows

$$CVM \xrightarrow{d} \sum_{j=1}^{\infty}\chi_j^2(1)/(j\pi)^2,$$

where $\{\chi_j^2(1)\}_{j=1}^{\infty}$ is a sequence of *i. i. d.* χ^2 random variables with one degree of freedom. This asymptotic distribution is nonstandard, but it is distribution-free and can be easily tabulated or simulated. An appealing property of Durlauf's (1991) test is its consistency against serial correlation of unknown form, and there is no need to choose a lag order *p*.
Deo (2000) shows that under the *m.d.s.* hypothesis with conditional heteroskedasticity, Durlauf's (1991) test statistic can be robustified as follows:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$CVM = \sum_{j=1}^{n-1} \frac{\left[\hat{\gamma}^2(j) / \hat{\gamma}_2(j)\right]}{(j\pi)^2} \xrightarrow{d} \sum_{j=1}^{\infty} \chi_j^2(1) / (j\pi)^2.$$

where $\hat{\gamma}_2(j)$ is a consistent estimator for the asymptotic variance of $\hat{\gamma}(j)$ and the asymptotic distribution remains unchanged. Like Lo and MacKinlay (1988), Deo (2000) also imposes the crucial fourth order joint cumulant condition in (2.5).

3 Serial dependence in nonlinear models

The autocorrelation function $\gamma(j)$, or equivalently, the power spectrum $h(\omega)$, of a time series $\{Y_t\}$, is a measure for linear association. When $\{Y_t\}$ is a stationary Gaussian process, $\gamma(j)$ or $h(\omega)$ can completely determine the full dynamics of $\{Y_t\}$.
It has been well documented, however, that most economic and financial time series, particularly high-frequency economic and financial time series, are not Gaussian. For non-Gaussian processes, $\gamma(j)$ and $h(\omega)$ may not capture the full dynamics of $\{Y_t\}$. We consider two nonlinear process examples:

- Bilinear (BL) autoregressive process:

$$Y_t = \alpha\varepsilon_{t-1}Y_{t-2} + \varepsilon_t, \ \{\varepsilon_t\} \sim i.i.d.(0, \sigma^2). \tag{3.1}$$

- Nonlinear moving average (NMA) process:

$$Y_t = \alpha\varepsilon_{t-1}\varepsilon_{t-2} + \varepsilon_t, \ \{\varepsilon_t\} \sim i.i.d.(0, \sigma^2). \tag{3.2}$$

For these two processes, there exists nonlinearity in conditional mean: $E(Y_t|I_{t-1}) = \alpha\ \varepsilon_{t-1}Y_{t-2}$ under (3.1) and $E(Y_t|I_{t-1}) = \alpha\ \varepsilon_{t-1}Y_{t-2}$ under (3.2). However, both processes are serially uncorrelated. If $\{Y_t\}$ follows either a BL process in (3.1) or a NMA process in (3.2), $\{Y_t\}$ is not *m.d.s.* but $\gamma(j)$ and $h(\omega)$ will miss it. Hong and Lee (2003a) document that indeed, for foreign currency markets, most foreign exchange changes are serially uncorrelated, but they are all not *m.d.s.* There exist predictable nonlinear components in the conditional mean of foreign exchange markets.
Serial dependence may also exist only in higher order conditional moments. An example is Engle's (1982) first order autoregressive conditional heteroskedastic (ARCH (1)) process:

$$\begin{cases} Y_t = \sigma_t\varepsilon_t, \\ \sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2, \\ \{\varepsilon_t\} \sim i.i.d.(0, 1). \end{cases} \tag{3.3}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

For this process, the conditional mean $E(Y_t|I_{t-1})=0$; which implies $\gamma(j)=0$ for all $j>0$.

However, the conditional variance, $var(Y_t|I_{t-1}) = \alpha_0 + \alpha_1 Y_{t-1}^2$, depends on the previous volatility. Both $\gamma(j)$ and $h(\omega)$ will miss such higher order dependence.

In nonlinear time series modelling, it is important to measure serial dependence, that is, any departure from *i.i.d.*, rather than merely serial correlation. As Priestley (1988) points out, the main purpose of nonlinear time series analysis is to find a filter $h(\cdot)$ such that

$$h(Y_t, Y_{t-1}, \cdots) = \varepsilon_t \sim i.i.d.(0, \sigma^2).$$

In other words, the filter $h(\cdot)$ can capture all serial dependence in $\{Y_t\}$ so that the 'residual' $\{\varepsilon_t\}$ becomes an *i.i.d.* sequence. One example of $h(\cdot)$ in modelling the conditional probability distribution of Y_t given I_{t-1}, is the probability integral transform

$$Z_t(\beta) = \int_{-\infty}^{Y_t} f(y|I_{t-1}, \beta)\,dy,$$

where $f(y|I_{t-1}, \beta)$ is a conditional density model for Y_t given and I_{t-1}, and β is an unknown parameter. When $f(y|I_{t-1}, \beta)$ is correctly specified for the conditional probability density of Y_t given I_{t-1}, that is, when the true conditional density coincides with $f(y|I_{t-1}, \beta^0)$ for some β^0, the probability integral transforms becomes

$$\left\{Z_t(\beta^0)\right\} \sim i.i.d.\,U[0, 1]. \tag{3.4}$$

Thus, one can test whether $f(y|I_{t-1}, \beta)$ is correctly specified by checking the *i.i.d.*U[0,1] for the probability integral transform series.

3.1 Bispectrum and higher-order spectra

Because the autocorrelation function $\gamma(j)$ and the spectral density $h(\omega)$ are rather limited in nonlinear time series analysis, various alternative tools have been proposed to capture nonlinear serial dependence (for example, Granger and Terasvirta, 1993; Tjøstheim, 1996). For example, one often uses the third-order cumulant function

$$C(j, k) \equiv E[(Y_t - \mu)(Y_{t-j} - \mu)(Y_{t-k} - \mu)],\ j, k = 0, \pm 1, \cdots.$$

This is also called the biautocovariance function of $\{Y_t\}$. It can capture certain nonlinear time series, particularly those displaying asymmetric behaviours such as skewness. Hsieh (1989) proposes a test based on $C(j, k)$ for a given pair of (j, k) which can detect some predictable nonlinear components in asset returns.

The Fourier transform of $C(j, k)$,

$$b(\omega_1, \omega_2) \equiv \frac{1}{(2\pi)^2} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C(j, k) e^{-ij\omega_1 - ik\omega_2},\ \omega_1, \omega_2 \in [-\pi, \pi],$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

is called the bispectrum of $\{Y_t\}$. When $\{Y_t\}$ is *i.i.d.*, $b(\omega_1, \omega_2)$ becomes a flat bispectral surface:

$$b_0(\omega_1, \omega_2) \equiv \frac{E(Y_t^3)}{(2\pi)^2}, \omega_1, \omega_2 \in [-\pi, \pi].$$

Any deviation from a flat bispectral surface will indicate the existence of serial dependence in $\{Y_t\}$. Moreover, $b(\omega_1, \omega_2)$ can be used to distinguish some linear time series processes from nonlinear time series processes. When $\{Y_t\}$ is a linear process with *i.i.d.* innovations, that is, when

$$Y_t = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j \varepsilon_{t-j} + \varepsilon_t, \left\{\varepsilon_t\right\} \sim i.i.d.(0, \sigma^2),$$

the normalized bispectrum

$$|\tilde{b}(\omega_1, \omega_2)|^2 \equiv \frac{|b(\omega_1, \omega_2)|^2}{h(\omega_1)h(\omega_2)h(\omega_1 + \omega_2)} = \frac{[E(\varepsilon_t^3)]^2}{2\pi\sigma^6}$$

is a flat surface. Any departure from a flat normalized bispectral surface will indicate that $\{Y_t\}$ is not a linear time series with *i.i.d.* innovations.
The bispectrum $b(\omega_1, \omega_2)$ can capture the BL and NMA processes in (3.1) and (3.2), because the third order cumulant $C(j, k)$ can distinguish them from an *i.i.d* process. However, it may still miss some important alternatives. For example, it will easily miss ARCH (1) with *i.i.d.* N(0,1) innovation $\{\varepsilon_t\}$. In this case, $b(\omega_1, \omega_2)$ becomes a flat bispectrum and cannot distinguish ARCH (1) from an *i.i.d.* sequence. One could use higher order spectra or polyspectra (Brillinger and Rosenblatt, 1967a; 1967b), which are the Fourier transforms of higher order cumulants. However, higher-order spectra have met with some difficulty in practice: Their spectral shapes are difficult to interpret, and their estimation is not stable in finite samples, due to the assumption of the existence of higher order moments. Indeed, it is often a question whether economic and financial data, particularly high-frequency data, have finite higher order moments.

3.2 Nonparametric measures of serial dependence

Nonparametric measures for serial dependence have been proposed in the literature, which avoid assuming the existence of moments. Granger and Lin (1994) propose a nonparametric entropy measure for serial dependence to identify significant lags in nonlinear time series. Define the Kullback–Leibler information criterion

$$I(j) = \int \ln\left[\frac{f_j(x, y)}{g(x)g(y)}\right] f_j(x, y)\,dx\,dy, \; j = 1, 2, \ldots.$$

where $f_j(x, y)$ is the joint probability density of Y_t and Y_{t-j}, and $g(x)$ is the marginal probability density of $\{Y_t\}$. The Granger–Lin

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

normalized entropy measure is defined as follows:

$$e^2(j) = 1 - \exp[-2I(j)],$$

which enjoys some appealing features. For example, $e(j)=0$ if and only if Y_t and Y_{t-j} are independent, and it is invariant to any monotonic continuous transformation. Because $f_j(x, y)$ and $g(x)$ are unknown, Granger and Lin (1994) use nonparametric kernel density estimators. They establish the consistency of their entropy estimator (say $\hat{I}(j)$) but do not derive its asymptotic distribution, which is important for confidence interval estimation and hypothesis testing.

In fact, Robinson (1991) has elegantly explained the difficulty of obtaining the asymptotic distribution of $\hat{I}(j)$ for serial dependence, namely it is a degenerate statistic so that the usual root-n normalization does not deliver a well-defined asymptotic distribution. Robinson (1991) considers a modified entropy estimator

$$\hat{I}_\gamma(j) = n^{-1} \sum_{t=j+1}^{n} C_t(\gamma) \ln\left[\frac{\hat{f}_j(Y_t, Y_{t-j})}{\hat{g}(Y_t)\hat{g}(Y_{t-j})}\right],$$

where $\hat{f}_j(\cdot, \cdot)$ and $\hat{g}(\cdot)$ are nonparametric kernel density estimators, $C_t(\gamma)=1-\gamma$ if t is odd, $C_t(\gamma)=1+\gamma$ if t is even, and γ is a pre-specified parameter. The weighting device $C_t(\gamma)$ does not affect the consistency of $\hat{I}_\gamma(j)$ to $I(j)$ and affords a well-defined asymptotic N(0,1) distribution under the *i.i.d.* hypothesis.

Skaug and Tjøstheim (1993a; 1996) use a different weighting function to avoid the degeneracy of the entropy estimator for serial dependence:

$$\hat{I}_w(j) = n^{-1} \sum_{t=1}^{n} w(Y_t, Y_{t-j}) \ln\left[\frac{\hat{f}_j(Y_t, Y_{t-j})}{\hat{g}(Y_t)\hat{g}(Y_{t-j})}\right],$$

where $w(Y_t, Y_{t-j})$ is a weighting function of observations X_t and X_{t-j}. Unlike using Robinson's (1991) weighting device, $\hat{I}_w(j)$ is not consistent for the population entropy $I(j)$, but it also delivers a well-defined asymptotic N(0,•1) distribution after a root-n normalization.

Intuitively, the use of weighting devices slows down the convergence rate of the entropy estimators, giving a well-defined asymptotic N(0,1) distribution after the usual root-n normalization. However, this is achieved at the cost of an efficiency loss, due to the slower convergence rate. Moreover, this approach breaks down when $\{Y_t\}$ is uniformly distributed, as in the case of the probability integral transforms of the conditional density in (3.4). Instead of using a weighting device, Hong and White (2005) exploit the degeneracy of $\hat{I}(j)$ and use a degenerate U-statistic theory to establish its asymptotic normality. Specifically, Hong and White (2005) show

$$nh\hat{I}(j) + hd_n^0 \xrightarrow{d} N(0, V),$$

where $h=h(n)$ is the bandwidth, and d_n^0 and V are nonstochastic factors. This approach preserves the convergence rate of the unweighted entropy estimator, giving sharper confidence interval estimation and more powerful hypothesis tests. It is applicable

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

when $\{Y_t\}$ is uniformly distributed.

Skaug and Tjøstheim (1993b) also use an Hoeffding measure to test serial dependence (see also Delgado, 1996; Hong, 1998; 2000). The empirical Hoeffding measures are based on the empirical distribution functions, which avoid smoothed nonparametric density estimation.

3.3 Generalized spectrum

Without assuming the existence of higher order moments, Hong (1999) proposes a generalized spectrum as an alternative analytic tool to the power spectrum and higher order spectra. The basic idea is to transform $\{Y_t\}$ via a complex-valued exponential function

$$Y_t \to \exp(iuY_t),\ u \in (-\infty, \infty),$$

and then consider the spectrum of the transformed series. Let $\psi(u) \equiv E(e^{iuY_t})$ be the marginal characteristic function of $\{Y_t\}$ and let $\psi_j(u, v) \equiv E[e^{i(uY_t + vY_{t-|j|})}],\ j = 0, \pm 1, \cdots,$ be the pairwise joint characteristic function of. $(Y_t, Y_{t-|j|})$. Define the covariance function between transformed variables e^{iuY_t} and $e^{ivY_{t-|j|}}$:

$$\sigma_j(u, v) \equiv \mathrm{cov}(e^{iuY_t}, e^{ivY_{t-|j|}}),\ j = 0, \pm 1, \ldots.$$

Straightforward algebra yields $\sigma_j(u, v) = \psi_j(u, v) - \psi(u)\ \psi(v)$, which is zero for all u, v if and only if Y_t and $Y_{t-|j|}$ are independent. Thus $\sigma_j(u, v)$ can capture any type of pairwise serial dependence over various lags, including those with zero autocorrelation. For example, $\sigma_j(u, v)$ can capture the BL, NMA and ARCH (1) processes in (3.1)–(3.3), all of which are serially uncorrelated.

The Fourier transform of the generalized covariance $\sigma_j(u, v)$:

$$f(\omega, u, v) \equiv \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j(u, v) e^{-ij\omega},\ \omega \in [-\pi, \pi],$$

is called the 'generalized spectral density' of $\{Y_t\}$. Like $\sigma_j(u, v)$, $f(\omega, u, v)$ can capture any type of pairwise serial dependencies in $\{Y_t\}$ over various lags. Unlike the power spectrum and higher order spectra, $f(\omega, u, v)$ does not require any moment condition on $\{Y_t\}$. When var(Y_t) exists, the power spectrum of $\{Y_t\}$ can be obtained by differentiating $f(\omega, u, v)$ with respect to (u v) at (0, 0):

$$h(\omega) \equiv \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j) e^{-ij\omega} = -\frac{\partial^2}{\partial u \partial v} f(\omega, u, v)|_{(u,v)=(0,0)},\ \omega \in [-\pi, \pi].$$

This is the reason why $f(\omega, u, v)$ is called the 'generalized spectral density' of $\{Y_t\}$.

When $\{Y_t\}$ is *i.i.d.*, $f(\omega, u, v)$ becomes a flat generalized spectrum as a function of ω:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$f_0(\omega, u, v) = \frac{1}{2\pi}\sigma_0(u, v), \omega \in [-\pi, \pi].$$

Any deviation of $f(\omega, u, v)$ from the flat generalized spectrum $f_0(\omega, u, v)$ is evidence of serial dependence. Thus, $f(\omega, u, v)$ is suitable to capture any departures from *i.i.d.* Hong and Lee (2003b) use the generalized spectrum to develop a test for the adequacy of nonlinear time series models by checking whether the standardized model residuals are *i.i.d.* Tests for *i.i.d.* are more suitable than tests for serial correlation in nonlinear contexts. Indeed, Hong and Lee (2003b) find that some popular EGARCH models are inadequate in capturing the full dynamics of stock returns, although the standardized model residuals are serially uncorrelated.
Insight into the ability of $f(\omega, u, v)$ can be gained by considering a Taylor series expansion

$$f(\omega, u, v) = \sum_{m=0}^{\infty}\sum_{l=0}^{\infty}\frac{(iu)^m (iv)^l}{m!\,l!}\left[\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\operatorname{cov}(X_t^m, X_{t-|j|}^l)\,e^{-ij\omega}\right].$$

Although $f(\omega, u, v)$ has no physical interpretation, it can be used to characterize cyclical movements caused by linear and nonlinear serial dependence. Examples of nonlinear cyclical movements include cyclical volatility clustering, and cyclical distributional tail clustering (for example, Engle and Manganelli's (2004) CAVaR model). Intuitively, the supremum function

$$s(\omega) = \sup_{-\infty < u, v < \infty} |f(\omega, u, v)|, \omega \in [-\pi, \pi],$$

can measure the maximum dependence at frequency ω of $\{Y_t\}$. It can be viewed as an operational frequency domain analogue of Granger and Terasvirta's (1993) maximum correlation measure

$$mmp(j) = \max_{g(\cdot), h(\cdot)} |corr[g(Y_t), h(X_{t-j})]|.$$

Once generic serial dependence is detected using $f(\omega, u, v)$ or any other dependence measure, one may like to explore the nature and pattern of serial dependence. For example, one may be interested in the following questions:

- Is serial dependence operative primarily through the conditional mean or through conditional higher order moments?
- If serial dependence exists in conditional mean, is it linear or nonlinear?
- If serial dependence exists in conditional variance, does there exist linear or nonlinear and asymmetric ARCH?

Different types of serial dependence have different economic implications. For example, the efficient market hypothesis fails if and only if there is no serial dependence in conditional mean.
Just as the characteristic function can be differentiated to generate various moments, generalized spectral derivatives, when they exist, can capture various specific aspects of serial dependence, thus providing information on possible types of serial dependence. Suppose $E[(Y_t)^{2\max(m,l)}] < \infty$ for some nonnegative integers m, l. Then the following generalized spectral derivative exists:

$$f^{(0,m,l)}(\omega, u, v) = \frac{\partial^{m+l}}{\partial u^m \partial v^l} f(\omega, u, v) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\sigma_j^{(m,l)}(u, v)\,e^{-ij\omega},$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $\sigma_j^{(m,l)}(u, v) \equiv \partial^{m+l}\sigma_j(u, v) / \partial u^m \partial v^l$. As an illustrative example, we consider the generalized spectral derivative of order $(m, l)=(1,0)$:

$$f^{(0,1,0)}(\omega, u, v) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\sigma_j^{(1,0)}(u, v)e^{-ij\omega}.$$

Observe $\sigma_j^{(1,0)}(0, v) \equiv \mathrm{cov}(iY_t, e^{ivY_{t-|j|}}) = 0$ for all $v\in(-\infty,\infty)$ if and only if $E(Y_t|Y_{t-|j|})=E(Y_t)$ a.s. The function $E(Y_t|Y_{t-|j|})$ is called the autoregression function of $\{Y_t\}$ at lag j. It can capture a variety of linear and nonlinear dependencies in conditional mean, including the BL and NMA processes in (3.1) and (3.2). (The use of $\sigma_j^{(1,0)}(0, v)$, which can be easily estimated by a sample average, avoids smoothed nonparametric estimation of $E(Y_t|Y_{t-|j|})$.) Thus, the generalized spectral derivative $f^{(0,1,0)}(\omega, u, v)$ can be used to capture a wide range of serial dependence in conditional mean. In particular, the function

$$s(\omega) = \sup_{-\infty<v<+\infty}|f^{(0,1,0)}(\omega, 0, v)|$$

can be viewed as an operational frequency domain analogue of Granger and Terasvirta's (1993) maximum mean correlation measure

$$mm(j) = \max_{h(\cdot)}|\mathrm{corr}(Y_t, h(Y_{t-j})|.$$

See Hong and Lee (2005) for more discussion.

Suppose one has found evidence of serial dependence in conditional mean using $f^{(0,1,0)}(\omega, u, v)$ or any other suitable measure, one can go further to explore whether there exists linear serial dependence in mean. This can be done by using the (1,1)-th order generalized derivative

$$f^{(0,1,1)}(\omega, 0, 0) = -h(\omega),$$

which checks serial correlation. Moreover, one can further use $f^{(0,1,l)}(\omega, u, v)$ for $l\geq 2$ to reveal nonlinear serial dependence in mean. In particular, these higher-order derivatives can suggest that there exist: (i) an ARCH-in-mean effect (for example, Engle, Lilien and Robins, 1987) if $\mathrm{cov}(Y_t, Y_{t-j}^2) \neq 0$, (ii) a skewness-in-mean effect (for example, Harvey and Siddique, 2000) if $\mathrm{cov}(Y_t, Y_{t-j}^3) \neq 0$, and (iii) kurtosis-in-mean effect (for example, Brooks, Burke and Persand, 2005) if $\mathrm{cov}(Y_t, Y_{t-j}^4) \neq 0$. These effects may arise from the existence of a time-varying risk premium, asymmetry of market behaviours, and inadequate account for large losses, respectively.

See Also

- kernel estimators in econometrics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- spectral analysis

I thank Steven Durlauf (editor) for suggesting this topic and comments on an earlier version, and Jing Liu for excellent research assistance and references. This research is supported by the Cheung Kong Scholarship of the Chinese Ministry of Education and Xiamen University. All remaining errors are solely mine.

Bibliography

Andrews, D.W.K. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–58.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedastcity. *Journal of Econometrics* 31, 307–27.

Box, G.E.P. and Pierce, D.A. 1970. Distribution of residual autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association* 65, 1509–26.

Breusch, T.S. 1978. Testing for autocorrelation in dynamic linear models. *Australian Economic Papers* 17, 334–55.

Breusch, T.S. and Pagan, A. 1980. The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies* 47, 239–53.

Brillinger, D.R. and Rosenblatt, M. 1967a. Asymptotic theory of estimates of kth order spectra. In *Spectral Analysis of Time Series*, ed. B. Harris. New York: Wiley.

Brillinger, D.R. and Rosenblatt, M. 1967b. Computation and interpretation of the kth order spectra. In *Spectral Analysis of Time Series*, ed. B. Harris. New York: Wiley.

Brooks, C., Burke, S. and Persand, G. 2005. Autoregressive conditional kurtosis. *Journal of Financial Econometrics* 3, 399–421.

Campbell, J.Y., Lo, A.W. and MacKinlay, A.C. 1997. *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press.

Chen, W. and Deo, R. 2004. A generalized portmanteau goodness-of-fit test for time series models. *Econometric Theory* 20, 382–416.

Cochrane, J.H. 1988. How big is the random walk in GNP? *Journal of Political Economy* 96, 893–920.

Delgado, M.A. 1996. Testing serial independence using the sample distribution function. *Journal of Time Series Analysis* 17, 271–85.

Deo, R.S. 2000. Spectral tests of the martingale hypothesis under conditional heteroscedasticity. *Journal of Econometrics* 99, 291–315.

Durbin, J. 1970. Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. *Econometrica* 38, 422–1.

Durbin, J. and Watson, G.S. 1950. Testing for serial correlation in least squares regression: I. *Biometrika* 37, 409–28.

Durbin, J. and Watson, G.S. 1951. Testing for serial correlation in least squares regression: II. *Biometrika* 38, 159–78.

Durlauf, S.N. 1991. Spectral based testing of the martingale hypothesis. *Journal of Econometrics* 50, 355–76.

Engle, R. 1982. Autoregressive conditional hetersokedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1008.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Engle, R., Lilien, D. and Robins, R.P. 1987. Estimating time varying risk premia in the term structure: the ARCH-M model. *Econometrica* 55, 391–407.

Engle, R. and Manganelli, S. 2004. CARViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics* 22, 367–91.

Fan, J. and Zhang, W. 2004. Generalized likelihood ratio tests for spectral density. *Biometrika* 91, 195–209.

Glosten, R., Jagannathan, R. and Runkle, D. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48, 1779–801.

Godfrey, L.G. 1978. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica* 46, 1293–301.

Granger, C.W.J. and Lin, J.L. 1994. Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis* 15, 371–84.

Granger, C.J.W. and Terasvirta, T. 1993. *Modeling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Harvey, C.R. and Siddique, A. 2000. Conditional skewness in asset pricing tests. *Journal of Finance* 51, 1263–95.

Hayashi, F. 2000. *Econometrics*. Princeton: Princeton University Press.

Hong, Y. 1996. Consistent testing for serial correlation of unknown form. *Econometrica* 64, 837–64.

Hong, Y. 1998. Testing for pairwise serial independence via the empirical distribution function. *Journal of the Royal Statistical Society, Series B* 60, 429–53.

Hong, Y. 1999. Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *Journal of the American Statistical Association* 94, 1201–20.

Hong, Y. 2000. Generalized spectral tests for serial dependence. *Journal of the Royal Statistical Society, Series B* 62, 557–74.

Hong, Y. and Lee, T.H. 2003a. Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* 85, 1048–62.

Hong, Y. and Lee, T.H. 2003b. Diagnostic checking for the adequacy of nonlinear time series models. *Econometric Theory* 19, 1065–121.

Hong, Y. and Lee, Y.J. 2005. Generalized spectral testing for conditional mean models in time series with conditional heteroskedasticity of unknown form. *Review of Economic Studies* 72, 499–51.

Hong, Y. and Lee, Y.J. 2007. Consistent testing for serial correlation of unknown form under general conditional heteroskedasticity. Working paper, Department of Economics, Cornell University, and Department of Economics, Indiana University.

Hong, Y. and White, H. 2005. Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* 73, 837–901.

Hsieh, D.A. 1989. Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business* 62, 339–68.

Ljung, G.M. and Box, G.E.P. 1978. On a measure of lack of fit in time series models. *Biometrika* 65, 297–303.

Lo, A.W. and MacKinlay, A.C. 1988. Stock market prices do not follow random walks: evidence from a simple specification test.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Review of Financial Studies 1, 41–66.

Nelson, D. 1991. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* 59, 347–70.

Newey, W.K. and West, K.D. 1987. A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.

Paparoditis, E. 2000. Spectral density based goodness-of-fit tests for time series models. *Scandinavian Journal of Statistics* 27, 143–76.

Priestley, M.B. 1988. *Non-Linear and Non-Stationary Time Series Analysis*. London: Academic Press.

Robinson, P.M. 1991. Consistent nonparametric entropy-based testing. *Review of Economic Studies* 58, 437–53.

Robinson, P.M. 1994. Time series with strong dependence. In *Advances in Econometrics, Sixth World Congress*, vol. 1, ed. C. Sims. Cambridge: Cambridge University Press.

Skaug, H.J. and Tjøstheim, D. 1993a. Nonparametric tests of serial independence. In *Developments in Time Series Analysis*, ed. S. Rao. London: Chapman and Hall.

Skaug, H.J. and Tjøstheim, D. 1993b. A nonparametric test of serial independence based on the empirical distribution function. *Biometrika* 80, 591–602.

Skaug, H.J. and Tjøstheim, D. 1996. Measures of distance between densities with application to testing for serial independence. In *Time Series Analysis in Memory of E.J. Hannan*, ed. P. Robinson and M. Rosenblatt. New York: Springer.

Tjøstheim, D. 1996. Measures and tests of independence: a survey. *Statistics* 28, 249–84.

Vinod, H.D. 1973. Generalization of the Durbin–Watson statistic for higher order autoregressive processes. *Communications in Statistics* 2, 115–44.

Wallis, K.F. 1972. Testing for fourth order autocorrelation in quarterly regression equations. *Econometrica* 40, 617–36.

Whang, Y.J. 1998. A test of autocorrelation in the presence of heteroskedasticity of unknown form. *Econometric Theory* 14, 87–122.

Wooldridge, J.M. 1990. An encompassing approach to conditional mean tests with applications to testing nonnested hypotheses. *Journal of Econometrics* 45, 331–50.

Wooldridge, J.M. 1991. On the application of robust, regression-based diagnostics to models of conditional means and conditional variances. *Journal of Econometrics* 47, 5–46.

How to cite this article

Hong, Yongmiao. "serial correlation and serial dependence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000485> doi:10.1057/9780230226203.1514

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Shackle, George Lennox Sharman (1903 – 1992)

Peter Earl
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

kaleidics; possibility; Post Keynesian economics; probability; rational expectations; Shackle, G. L. S.

Article

Shackle was born in Cambridge. Financial circumstances compelled him to take an external degree while working first as a bank clerk and then as a schoolmaster; it was only in 1935 that he was able to study under Hayek at the London School of Economics. This was an exciting time to be starting out and later, in one of his best-loved books, *The Years of High Theory* (1967), Shackle was to look back at the problem-solving activities responsible for the interwar theoretical breakthroughs. Within two years, and very much influenced by the latest work of Myrdal and Keynes, he completed his first doctorate (published as Shackle, 1938). By 1940 he was employed in wartime official service, having completed a second thesis that drew on material from his work as assistant to E.H. Phelps Brown at Oxford. Despite the demands of official work, he produced a series of articles on uncertain, crucial choices, whose outcomes may define, for good or bad, the chooser's future possibilities (see especially Shackle, 1942; 1943). These were reworked into his (1949) book and he rose rapidly, after returning to academia as Reader in Economics at Leeds University in 1950, to become Brunner Professor of Economic Science in the University of Liverpool in 1951. His retirement from Liverpool in 1969 saw no easing in his industry or in his desire to see economists deal with knowledge problems as analytical rudiments rather than refinements (see Shackle, 1972).

Although Shackle's (1949) analysis of crucial choices attracted immediate attention, it won few adherents. In this book, as in many of his subsequent works, Shackle argued that probabilistic notions are questionable if choice experiments can destroy any possibility of their own replication. (Post-Keynesians have extended his view in criticizing the rational expectations hypothesis.) Shackle suggested that, in such situations, choosers would come to focus on particularly attention-arresting pairs of *possibilities* (one pair for each scheme of action). A possibility is not something which a chooser would expect to happen, given enough tries, with a particular frequency, but something whose taking place looks *ex ante* surprising (unsurprising) because potentially fatal obstacles to it can (cannot) be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
envisaged. Shackle insisted that possibility is not in general distributive: thoughts about a possible outcome not previously imagined need not affect assignments of potential surprise to its rivals, since potential surprise ratings do not sum to any fixed, bounded value. Despite this, many theorists found his 'potential surprise curves' difficult to distinguish from inverted probability distributions; they also took issue with his view that it is not rational for choosers to weigh together values for possibilities that are mutually exclusive. Behaviouralists were ill-disposed to the large role played by indifference surfaces in his analysis of how ascendant gain/loss pairings were focused upon and then ranked; whereas orthodox theorists (for example, Ford, 1983) argued that it would be irrational for choosers to focus in the way he proposed, and that his selection device – the 'gambler preferences' map – produced the questionable result that an investor will choose a portfolio consisting of no more than two types of financial asset. Much of his noteworthy retirement output (especially his 1974 book) tried to make economists recognize that the incompatibility of speculators' expectations and changes in the 'state of the news' will make the relative demands for durable assets prone to kaleidoscopic instability. To many orthodox model-builders, his kaleidic conception of economic systems had unacceptably nihilistic implications, but it led some Post Keynesians to examine how institutions and policies might be designed to constrain explosive and implosive forces whose precise timings and strengths may be impossible to anticipate.

Selected works

1938. *Expectations, Investment and Income*. Oxford: Oxford University Press.

1942. A theory of investment-decisions. *Oxford Economic Papers*, NS 6(April), 77–94.

1943. The expectational dynamics of the individual. *Economica*, NS 10(May), 99–129.

1949. *Expectation in Economics*. Cambridge: Cambridge University Press.

1967. *The Years of High Theory: Invention and Tradition in Economic Thought 1926–1939*. Cambridge: Cambridge University Press.

1972. *Epistemics and Economics*. Cambridge: Cambridge University Press.

1974. *Keynesian Kaleidics*. Edinburgh: Edinburgh University Press.

Bibliography

Ford, J.L. 1983. *Choice, Expectation and Uncertainty*. Oxford: Martin Robertson.

Ford, J.L. 1994. *G.L.S. Shackle: The Dissenting Economist's Economist*. Aldershot: Edward Elgar.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Earl, Peter. "Shackle, George Lennox Sharman (1903–1992)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000104> doi:10.1057/9780230226203.1515

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

shadow pricing

Ravi Kanbur
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

border prices; decentralization; distribution of income; lump sum taxes; market prices; shadow pricing; social cost

Article

When a businessman evaluates a project, he does it with a view to calculating the prospective profit from it. These calculations can be seen as taking place in two steps. At the first step, all the physical consequences of relevance to the businessman – the inputs to and outputs from the project – are assessed. At the second stage, these inputs and outputs are converted into costs and revenues, using *market prices*. It is natural that a private businessman should use the ruling market prices for costing inputs and for valuing sales, since these are the prices at which transactions take place and hence profit generated.

Consider now the evaluation of a project by a government. Such evaluation will differ at each of the two steps referred to above. At the first step, the government will be interested in *all* of the repercussions of the project, however indirect. This is because it is the government rather than a private businessman concerned with his own narrowly defined activities. At the second step, the government will wish to use not the ruling market prices but prices which reflect social costs and social benefits, in order to calculate what might be termed social profit. These prices are referred to as *shadow prices*, or accounting prices (see Little and Mirrlees, 1974), and the name suggests that they are to be used in lieu of the actual market prices.

Market prices are what they are. But how are shadow prices to be calculated? Clearly they depend on the government's objective function and on the constraints it faces. The shadow prices should be such that the social profit from the project is positive if and only if the project increases the value of the government's objective function. In a general competitive equilibrium, if the government's objective is economic efficiency, then it can be argued that for a small project the shadow prices do in fact coincide with market prices. If the government's objective includes the pursuance of equity, but it has lump sum instruments to carry this out, then shadow prices still coincide with market prices. Basically the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
government should use redistributive lump sum taxation to pursue equity and the project to pursue increases in aggregate economic welfare.

But if the government does not have a sufficient range of instruments to pursue effective redistribution without distortion it may be the case that, even with a full competitive equilibrium, shadow prices may differ from market prices. In addition to this, if the economy is not in a full competitive equilibrium, then the case for using shadow prices different from market prices can be argued strongly.

In programming terms, shadow prices are simply dual to the changes in the government's objective function. One justification for their use is the benefits of decentralization: local project evaluators are better equipped to analyse the physical consequences of a project, and this localized knowledge should be used in conjunction with centrally determined shadow prices to evaluate the social profitability of projects. But the real difficulties arise in specifying the objectives of the government and in specifying its constraints, and this is in turn related to who is thought of as doing the project evaluation.

The standard assumption is one of a unitary government with a given social welfare function – a benevolent dictator. But the reality is one where either the project evaluator is part of a government which is a coalition of interests, or the project evaluation is being done by an international agency which faces a government made up of conflicting and competing objectives. The logical procedure for an international agency should be clear – in evaluating a project it should incorporate a model of the political process to clarify the responses of various government instruments to the project. Sen (1972) gives an illuminating discussion of a project which requires importing an input on which there is already a quota – so that the border price of the input is very different from its domestic scarcity value. The Little and Mirrlees (1974) method of using border prices is predicated on the assumption that it is these prices which represent the transformation possibilities for the economy as a whole. But if the assessment of the political realities is such that this quota will not be removed by the government – because of the overriding influence of interest groups that benefit from the rents generated by the quota – then the domestic scarcity value should be used in costing the input.

Similarly, any project which alters significantly the distribution of income will have repercussions on the political process – and there will be attempts by groups who are adversely affected to restore their standard of living. Project evaluation in general, and shadow pricing in particular, should take these into account. Consider, for example, the shadow cost of labour. If the labour used on the project comes from the agricultural sector, and if this labour is a constraint on output, then agricultural output will fall. If government revenue depends on taxation of this output, this will fall too. If, in turn, government expenditure is a major source of non-agricultural (urban) incomes, then at constant fiscal deficit urban incomes will fall. This change in the distribution of income will be an important element in the shadow cost of labour. But suppose now that the political processes are such as to not allow a decline in urban living standards. Rather, government expenditure remains constant and the fiscal deficit increases. Now it is the increased burden on future generations which has to be taken into account. Either way, it should be clear that a model of the political process is crucial in specifying shadow prices even if the project evaluator (be it an international agency or a project evaluation unit within the government) is clear about what the objectives are.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- cost–benefit analysis

Bibliography

Little, I.M.D. and Mirrlees, J.A. 1974. *Project Appraisal and Planning for Developing Countries*. London: Heinemann.

Sen, A.K. 1972. Control areas and accounting prices: an approach to economic evaluation. *Economic Journal* 82 (Supp.), 486–501.

How to cite this article

Kanbur, Ravi. "shadow pricing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000105> doi:10.1057/9780230226203.1516

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Shapley value

Sergiu Hart
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The *Shapley value* is an a priori evaluation of the prospects of a player in a multi-person game. Introduced by Lloyd S. Shapley in 1953, it has become a central solution concept in cooperative game theory. The Shapley value has been applied to economic, political, and other models.

Keywords

Banzhaf index; coalitions; competitive equilibrium; cooperative game theory; cost allocation; games in coalitional form; large games; market games; perfect competition; political power; Shapley value; Shapley–Shubik index; side payments; transferable utility; value equivalence principle

Article

The *value* of an uncertain outcome (a ‘lottery’) is an a priori measure, in the participant's utility scale, of what he expects to obtain (this is the subject of ‘utility theory’). The question is, how would one evaluate the prospects of a player in a multi-person interaction, that is, in a game?
This question was originally addressed by Lloyd S. Shapley (1953a). The framework was that of n-person games in coalitional form with side-payments, which are given by a set N of ‘players’, say 1, 2, …, n, together with a ‘coalitional function’ v that associates to every subset S of N (‘coalition’) a real number $v(S)$, the maximal total payoff the members of S can obtain (the ‘worth’ of S). An underlying assumption of this model is that there exists a medium of exchange (‘money’) that is freely transferable in unlimited amounts between the players, and moreover every player's utility is additive with respect to it (that is, a transfer of x units from one player to another decreases the first one's utility by x units and increases the second one's utility by x units; the total payoff of a coalition can thus be meaningfully defined as the sum of the payoffs of its members). This requirement is known as existence of ‘side payments’ or ‘transferable utility’. In addition, the game is assumed to be adequately described by its coalitional function (that is, the worth $v(S)$ of each coalition S is well defined, and the abstraction from

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the extensive structure of the game to its coalitional function leads to no essential loss; such a game is called a '*c*-game'). These assumptions may be interpreted in a broader and more abstract sense. For example, in a voting situation, a 'winning coalition' is assigned worth 1, and a 'losing' coalition, worth 0. The essential feature is that the prospects of each coalition may be summarized by one number.

The *Shapley value* associates to each player in each such game a unique payoff – his 'value'. The value is required to satisfy the following four axioms. (EFF) *Efficiency* or *Pareto optimality:* The sum of the values of all players equals $v(N)$, the worth of the grand coalition of all players (in a superadditive game $v(N)$ is the maximal amount that the players can jointly get); this axiom combines feasibility and efficiency. (SYM) *Symmetry* or *equal treatment:* If two players in a game are substitutes (that is, the worth of no coalition changes when replacing one of the two players by the other one), then their values are equal. (NUL) *Null* or *dummy player:* If a player in a game is such that the worth of every coalition remains the same when he joins it, then his value is zero. (ADD) *Additivity:* The value of the sum of two games is the sum of the values of the two games (equivalently, the value of a probabilistic combination of two games is the same as the probabilistic combination of the values of the two games; this is analogous to 'expected utility'). The surprising result of Shapley is that these four axioms *uniquely determine* the values in *all* games.

Remarkably, the Shapley value of a player in a game turns out to be exactly his *expected marginal contribution to a random coalition*. The marginal contribution of a player i to a coalition S (that does not contain i) is the change in the worth when i joins S, that is, $v(S \cup \{i\}) - v(S)$. To obtain a random coalition S not containing i, arrange the n players in a line (for example, 1, 2, …, n) and put in S all those that precede i in that order; all $n!$ orders are assumed to be equally likely. The formula for the Shapley value is striking, first, since it is a consequence of very simple and basic axioms and, second, since the idea of marginal contribution is so fundamental in much of economic analysis.

It should be emphasized that the value of a game is an a priori measure, that is, an evaluation before the game is actually played. Unlike other solution concepts (for example, core, von Neumann–Morgenstern solution, bargaining set), it need not yield a 'stable' outcome (the probable final result when the game is actually played). These final stable outcomes are in general not well determined; the value – which is uniquely specified – may be thought of as their expectation or average. Another interpretation of the value axioms regards them as rules for 'fair' division, guiding an impartial 'referee' or 'arbitrator'. Also, as suggested above, the Shapley value may be understood as the utility of playing the game (Shapley, 1953a; Roth, 1977).

In view of both its strong intuitive appeal and its mathematical tractability, the Shapley value has been the focus of much research and many applications. We can only briefly mention some of these here (together with just a few representative references). The reader is referred to the survey of Aumann (1978) and, for more extensive coverage, to the *Handbook of Game Theory* (Aumann and Hart, vol 1: 1992 [HGT1], vol 2: 1994 [HGT2], vol 3: 2002 [HGT3]), especially Chapters 53–58, as well as parts of Chapters 32–34 and 37.

Variations

Following Shapley's pioneering approach, the concept of *value* has been extended, modified and generalized.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Weighted values

Assume that the players are of unequal ‘size’ (for example, a player may represent a ‘group’, a ‘department’, and so on), and this is expressed by given (relative) weights. This setup leads to ‘weighted Shapley values’ (Shapley, 1953b); in unanimity games, for example, the values of the players are no longer equal but, rather, proportional to their weights [HGT3, Ch. 54].

Semi-values

Abandoning the efficiency axiom (EFF) yields the class of ‘semi-values’ (Dubey, Neyman and Weber, 1981). An interesting semi-value is the *Banzhaf index* (Penrose, 1946; Banzhaf, 1965; Dubey and Shapley, 1979), originally proposed as a measure of power in voting games. Like the Shapley value, it is also an expected marginal contribution, but here all coalitions not containing player *i* are equally likely [HGT3, Ch. 54].

Other axiomatizations

There are alternative axiomatic systems that characterize the Shapley value. For instance, one may replace the additivity axiom (ADD) with a *marginality axiom* that requires the value of a player to depend only on his marginal contributions (Young, 1985). Another approach is based on the existence of a *potential* function together with efficiency (EFF) (Hart and Mas-Colell, 1989) [HGT3, Ch. 53].

Consistency

Given a solution concept which associates payoffs to games, assume that a group of players in a game have already agreed to it, are paid off accordingly, and leave the game; consider the ‘reduced game’ among the remaining players. If the solution of the reduced game is the same as that of the original game, then the solution is said to be *consistent*. It turns out that consistency, together with some elementary requirements for two-player games, characterizes the Shapley value (Hart and Mas-Colell, 1989) [HGT3, Ch. 53], [HGT1, Ch. 18].

Large games

Assume that the number of players increases and individuals become negligible. Such models are important in applications (such as competitive economies and voting), and there is a vast body of work on values of large games that has led to beautiful and important insights (for example, Aumann and Shapley, 1974) [HGT3, Ch. 56].

NTU games

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

These are games ‘without side payments’, or ‘with non-transferable utility’ (that is, the existence of a medium of utility exchange is no longer assumed). The simplest such games, two-person pure bargaining problems, were originally studied by Nash (1950). Values for general NTU games, which coincide with the Shapley value in the side payments case, and with the Nash bargaining solution in the two-person case, have been introduced by Harsanyi (1963), Shapley (1969), Maschler and Owen (1992) [HGT3, Ch. 55].

Non-cooperative foundations

Bargaining procedures whose non-cooperative equilibrium outcome is the Shapley value have been proposed by Gul (1989) (see Hart and Levy, 1999; Gul, 1999) and Winter (1994) for strictly convex games, and by Hart and Mas-Colell (1996) for general games [HGT3, Ch. 53].

Other extensions

This includes games with communication graphs (Myerson, 1977), coalition structures (Aumann and Drèze, 1974; Owen, 1977; Hart and Kurz, 1983), and others [HGT2, Ch. 37], [HGT3, Ch. 53].

Economic applications

Perfect competition

In the classical economic model of perfect competition, the commodity prices are determined by the requirement that total demand equals total supply; this yields a *competitive* (or *Walrasian*) *equilibrium*. A different approach in such setups looks at the cooperative ‘market game’ where the members of each coalition can freely exchange among themselves the commodities they own. A striking phenomenon occurs: various game-theoretic solutions of the market games yield precisely the competitive equilibria. In particular, in perfectly competitive economies every Shapley value allocation is competitive and, if the utilities are smooth, then every competitive allocation is also a value allocation. This result, called the *value equivalence principle*, is remarkable since it joins together two very different approaches: competitive prices arising from supply and demand on the one hand, and marginal contributions to trading coalitions on the other. The value equivalence principle has been studied in a wide range of models (for example, Shapley, 1964; Aumann, 1975). While it is undisputed in the TU case, its extension to the general NTU case seems less clear (it holds for the Shapley NTU value, but not necessarily for other NTU values) [HGT3, Ch. 57].

Cost allocation

Consider the problem of allocating joint costs in a ‘fair’ manner. Think of the various ‘tasks’ (or ‘projects’, ‘departments’, and so on) as players, and let $v(S)$ be the total cost of carrying out the set S of tasks (Shubik, 1962). It turns out that the axioms determining the Shapley value are easily translated into

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
postulates appropriate for solving cost allocation problems (for example, the efficiency axiom becomes ‘total-cost-sharing’). Two notable applications are airport landing fees (a task here is an aircraft landing; Littlechild and Owen, 1973) and telephone billing (each time unit of a phone call is a player; the resulting cost allocation scheme was put into actual use at Cornell University; Billera, Heath and Raanan, 1978) [HGT2, Ch. 34].

Other applications

The value has been applied to various economic models; for example, models of taxation where a political power structure is given in addition to the economic data (Aumann and Kurz, 1977). Further references to economic applications can be found in Aumann (1985) [HGT3, Ch. 58], [HGT2, Ch. 33].

Political applications

What is the ‘power’ of an individual or a group in a voting situation? A trivial observation – though not always remembered in practice – is that the political power need not be proportional to the number of votes (see Shapley, 1981, for some interesting examples). It is therefore important to find an objective method of measuring power in such situations. The Shapley value (known in this setup as the *Shapley–Shubik index*; Shapley and Shubik, 1954) is, by its very nature, a most appropriate candidate. Indeed, consider a simple political game, described by specifying whether each coalition is ‘winning’ or ‘losing’. The Shapley value of a player *i* turns out to be the probability that *i* is the ‘pivot’ or ‘key’ player, namely, that in a random order of all players those preceding *i* are losing, whereas together with *i* they are winning. For example, in a 100-seat parliament with simple majority (that is, 51 votes are needed to win), assume there is one large party having 33 seats and the rest are divided among many small parties; the value of the large party is then close to 50%, considerably more than its voting weight (that is, its 33% share of the seats). In contrast, when there are two large parties each having 33 seats and a large number of small parties, the value of each large party is close to 25% – much less than its voting weight of 33%. To understand this, think of the competition between the two large parties to attract the small parties to form a winning coalition; in contrast, when there is only one large party, the competition is between the small parties (to join the large party).
The Shapley value has also been used in more complex models, where ‘ideologies’ and ‘issues’ are taken into account (thus, not all arrangements of the voters are equally likely; an ‘extremist’ party, for example, is less likely to be the pivot than a ‘middle-of-the-road’ one; Owen, 1971; Shapley, 1977). References to political applications of the Shapley value may be found in Shapley (1981); these include various parliaments (USA, France, Israel), the United Nations Security Council, and others [HGT2, Ch. 32].

See Also

- game theory

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Aumann, R.J. 1975. Values of markets with a continuum of traders. *Econometrica* 43, 611–46.

Aumann, R.J. 1978. Recent developments in the theory of the Shapley Value. *Proceedings of the International Congress of Mathematicians*, Helsinki.

Aumann, R.J. 1985. On the non-transferable utility value: a comment on the Roth–Shafer examples. *Econometrica* 53, 667–78.

Aumann, R.J. and Drèze, J.H. 1974. Cooperative games with coalition structures. *International Journal of Game Theory* 3, 217–37.

Aumann, R.J. and Hart, S., eds. 1992 [HGT1], 1994 [HGT2], 2002 [HGT3]. *Handbook of Game Theory, with Economic Applications*, vols 1–3. Amsterdam: North-Holland.

Aumann, R.J. and Kurz, M. 1977. Power and taxes. *Econometrica* 45, 1137–61.

Aumann, R.J. and Shapley, L.S. 1974. *Values of Non-atomic Games*. Princeton: Princeton University Press.

Banzhaf, J.F. 1965. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review* 19, 317–43.

Billera, L.J., Heath, D.C. and Raanan, J. 1978. Internal telephone billing rates: a novel application of non-atomic game theory. *Operations Research* 26, 956–65.

Dubey, P., Neyman, A. and Weber, R.J. 1981. Value theory without efficiency. *Mathematics of Operations Research* 6, 122–8.

Dubey, P. and Shapley, L.S. 1979. Mathematical properties of the Banzhaf Power Index. *Mathematics of Operations Research* 4, 99–131.

Gul, F. 1989. Bargaining foundations of Shapley value. *Econometrica* 57, 81–95.

Gul, F. 1999. Efficiency and immediate agreement: a reply to Hart and Levy. *Econometrica* 67, 913–18.

Harsanyi, J.C. 1963. A simplified bargaining model for the *n*-person cooperative game. *International Economic Review* 4, 194–220.

Hart, S. and Kurz, M. 1983. Endogenous formation of coalitions. *Econometrica* 51, 1047–64.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hart, S. and Levy, Z. 1999. Efficiency does not imply immediate agreement. *Econometrica* 67, 909–12.

Hart, S. and Mas-Colell, A. 1989. Potential, value and consistency. *Econometrica* 57, 589–614.

Hart, S. and Mas-Colell, A. 1996. Bargaining and value. *Econometrica* 64, 357–80.

Littlechild, S.C. and Owen, G. 1973. A simple expression for the Shapley value in a special case. *Management Science* 20, 370–72.

Maschler, M. and Owen, G. 1992. The consistent Shapley value for games without side payments. In *Rational Interaction: Essays in Honor of John Harsanyi*, ed. R. Selten. New York: Springer.

Myerson, R.B. 1977. Graphs and cooperation in games. *Mathematics of Operations Research* 2, 225–29.

Nash, J.F. 1950. The bargaining problem. *Econometrica* 18, 155–62.

Owen, G. 1971. Political games. *Naval Research Logistics Quarterly* 18, 345–55.

Owen, G. 1977. Values of games with a priori unions. In *Essays in Mathematical Economics and Game Theory*, ed. R. Henn and O. Moeschlin. New York: Springer.

Penrose, L.S. 1946. The elementary statistics of majority voting. *Journal of the Royal Statistical Society* 109, 53–7.

Roth, A.E. 1977. The Shapley value as a von Neumann–Morgenstern utility. *Econometrica* 45, 657–64.

Shapley, L.S. 1953a. A value for *n*-person games. In *Contributions to the Theory of Games, II*, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Shapley, L.S. 1953b. Additive and non-additive set functions. Ph.D. thesis, Princeton University.

Shapley, L.S. 1964. Values of large games VII: a general exchange economy with money. Research Memorandum 4248-PR. Santa Monica, CA: RAND Corp.

Shapley, L.S. 1969. Utility comparison and the theory of games. In *La Décision: agrégation et dynamique des ordres de préférence*. Paris: Editions du CNRS.

Shapley, L.S. 1977. A comparison of power indices and a nonsymmetric generalization. Paper No. P–5872. Santa Monica, CA: RAND Corp.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Shapley, L.S. 1981. Measurement of power in political systems. *Game Theory and its Applications*, Proceedings of Symposia in Applied Mathematics, vol. 24. Providence, RI: American Mathematical Society.

Shapley, L.S. and Shubik, M. 1954. A method for evaluating the distribution of power in a committee system. *American Political Science Review* 48, 787–92.

Shubik, M. 1962. Incentives, decentralized control, the assignment of joint costs and internal pricing. *Management Science* 8, 325–43.

Winter, E. 1994. The demand commitment bargaining and snowballing cooperation. *Economic Theory* 4, 255–73.

Young, H.P. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14, 65–72.

How to cite this article

Hart, Sergiu. "Shapley value." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000106> doi:10.1057/9780230226203.1517

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Shapley – Folkman theorem

Ross M. Starr
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Shapley–Folkman theorem places an upper bound on the size of the non-convexities (loosely speaking, openings or holes) in a sum of non-convex sets in Euclidean N-dimensional space, R^N. The bound is based on the size of non-convexities in the sets summed and the dimension of the space. When the number of sets in the sum is large, the bound is independent of the number of sets summed, depending rather on N, the dimension of the space. Hence the size of the non-convexity in the sum becomes small as a proportion of the number of sets summed; the non-convexity per summand goes to zero as the number of summands becomes large. The Shapley–Folkman theorem can be viewed as a discrete counterpart to the Lyapunov theorem on non-atomic measures (Grodal, 2002).

Keywords

large economies; Lyapunov th; non-convexity; Shapley–Folkman th

Article

The theorem is used to demonstrate the following properties:

- existence of approximate competitive general equilibrium in large finite economies with non-convex preferences (increasing marginal rate of substitution) or non-convex technology (bounded increasing returns; the U-shaped cost curve case);
- convergence of the core to the set of competitive equilibria (Arrow and Hahn, 1972; Anderson, 1978).

It may also be used to characterize the solution of non-convex programming problems (Aubin and Ekeland, 1976).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

For $S \subset R^N$, S compact, define rad(S), the radius of S, as a measure of the size of S. Define $r(S)$, the inner radius of S, and $\rho(S)$ inner distance of S, as measures of the non-convexity (size of holes) of S. Let con$S$ denote the closed convex hull of S (smallest closed convex set containing S as a subset).

$$\mathrm{rad}(S) \equiv \inf_{x \in R^N} \sup_{y \in S} |x - y|;$$

$$r(S) \equiv \sup_{x \in \mathrm{con}S} \inf_{\{T \subset S | T \text{ spans } x\}} \mathrm{rad}(T);$$

$$\rho(S) \equiv \sup_{x \in \mathrm{con}S} \inf_{y \in S} |x - y|.$$

rad(S) is the radius of the smallest closed ball centred in con$S$ containing S. A set of points T is said to span a point x, if x can be expressed as a convex combination (weighted average) of elements of T. $r(S)$ is the smallest radius of a ball centred in the convex hull of S, so that the ball is certain to contain a set of points of S that span the ball's centre. Hence $r(S)$ represents a measure of breadth of non-convexities in S. $\rho(S)$ is the maximum distance from a point in con$S$ to (the nearest point of) S. Hence it represents the smaller of breadth or depth of non-convexities of S.

Let $S_1, S_2, \ldots, S_m$ be a (finite) family of m compact subsets of R^N. The vector sum of $S_1, S_2, \ldots, S_m$, denoted W is a set composed of representative elements of $S_1, S_2, \ldots, S_m$ summed together. W is defined as

$$W \equiv \sum_{i=1}^{m} S_i \equiv \left\{ w | w = \sum_{i=1}^{m} x^i,\ x^i \in S^i \right\}$$

where the sum in the brackets is taken over one element of each S_i.

Theorem: (Shapley–Folkman): Let $S_1, \ldots, S_m$ be a family of m compact subsets of R^N; $W = \Sigma_{i=1}^{m} S_i$. Let $L \geq \mathrm{rad}(S_i)$ for all S_i; let n=min(N, m). Then for any $x \in \mathrm{con}\, W$

1. (i) $x = \Sigma_{i=1}^{m} x^i$, where $x^i \in \mathrm{con} S_i$ and with at most N exceptions, $x^i \in S_i$;
2. (ii) there is $y \in W$ so that $|x - y| \leq L\sqrt{n}$.

Corollary: (Starr): Let $S_1, \ldots, S_m$ be a finite family of compact subsets of R^N. $W = \Sigma_{i=1}^{m} S_i$. Let $L \geq r(S_i)$ for all S_i, n=min(m, N). Then for any $x \in \mathrm{con}\, W$ there is $y \in W$ so that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$|x - y| \leq L\sqrt{n}.$$

Corollary: (Heller): Let $S_1,\ldots, S_m$ be a finite family of compact subsets of R^N; $W = \Sigma_{i=1}^{m} S_i$. Let $L \geq \rho(S_i)$ for all S_i, n=min(m, N). Then for any $x \in \operatorname{con} W$ there is $y \in W$ so that

$$|x - y| \leq Ln.$$

Statements and proofs of the theorem and corollaries along with applications are available in Arrow and Hahn (1972) and Green and Heller (1981). Development of the theorem is due to L.S. Shapley and J.H. Folkman (private correspondence) with publication in Starr (1969). Extensions, alternative proofs, and applications appear in the other references.

See Also

- perfect competition

Bibliography

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46, 1483–7.

Anderson, R.M. 1988. The Second Welfare Theorem with nonconvex preferences. *Econometrica* 56, 361–82.

Arrow, K.J. and Hahn, F.H. 1972. *General Competitive Analysis*. San Francisco: Holden-Day.

Artstein, Z. and Vitale, R.A. 1975. A strong law of large numbers for random compact sets. *Annals of Probability* 3, 879–82.

Artstein, Z. 1980. Discrete and continuous bang-bang and facial spaces or: look for the extreme points. *SIAM Review* 22, 172–85.

Aubin, J.-P. and Ekeland, I. 1976. Estimation of the duality gap in nonconvex optimization. *Mathematics of Operations Research* 1(3), 225–45.

Cassels, J.W.S. 1975. Measure of the non-convexity of sets and the Shapley–Folkman–Starr theorem.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mathematical Proceedings of the Cambridge Philosophical Society 78, 433–6.

Chambers, C.P. 2005. Multi-utilitarianism in two-agent quasilinear social choice. *International Journal of Game Theory* 33, 315–34.

Ekeland, I. and Temam, R. 1976. *Convex Analysis and Variational Problems*. Amsterdam: North-Holland.

Green, J. and Heller, W.P. 1981. Mathematical analysis and convexity with applications to economics. In *Handbook of Mathematical Economics*, vol. 1, ed. K.J. Arrow and M. Intriligator. Amsterdam: North-Holland.

Grodal, B. 2002. The equivalence principle. In *Optimization and Operation Research, Encyclopedia of Life Support Systems (EOLSS)*, ed. U. Derigs. Cologne. Online. Available at http://www.econ.ku.dk/grodal/EOLSS-final.pdf, accessed 5 April 2007.

Hildenbrand, W., Schmeidler, D. and Zamir, S. 1973. Existence of approximate equilibria and cores. *Econometrica* 41, 1159–66.

Howe, R. 1979. On the tendency toward convexity of the vector sum of sets. Discussion Paper No. 538, Cowles Foundation, Yale University.

Manelli, A.M. 1991. Monotonic preferences and core equivalence. *Econometrica* 59, 123–38.

Mas-Colell, A. 1978. A note on the core equivalence theorem: how many blocking coalitions are there? *Journal of Mathematical Economics* 5, 207–16.

Proske, F.N. and Puri, M.L. 2002. Central limit theorem for Banach space valued fuzzy random variables. *Proceedings of the American Mathematical Society* 130, 1493–501.

Starr, R.M. 1969. Quasi-equilibria in markets with non-convex preferences. *Econometrica* 37, 25–38.

Starr, R.M. 1981. Approximation of points of the convex hull of a sum of sets by points of the sum: an elementary approach. *Journal of Economic Theory* 25, 314–17.

Tardella, F. 1990. A new proof of the Lyapunov Convexity Theorem. *Applied Mathematics* 28, 478–81.

Weil, W. 1982. An application of the central limit theorem for Banach space valued random variables to the theory of random sets. *Probability Theory and Related Fields* 60, 203–8.

Zhou, L. 1993. A simple proof of the Shapley–Folkman th. *Economic Theory* 3, 371–2.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Starr, Ross M. "Shapley–Folkman theorem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000107> doi:10.1057/9780230226203.1518

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sharecropping

Pierre Dubois
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Sharecropping is a form of land leasing contract between a tenant and a landlord who share the production. It has a variety of forms and is sometimes linked with credit, lending, or insurance. The apparent inefficiency of sharecropping due to the fact that the tenant receives only a share of the marginal productivity of his labour has attracted economists' attention since Adam Smith. Within the principal–agent paradigm, sharecropping is now thought of as trading off incentives and risk sharing or as reducing transaction costs for a landlord willing to lend out a piece of land.

Keywords

agency costs; arbitrage; collusion; contract repetition; cost sharing; credit; fixed-rent contracts; fixed-wage contracts; incentive contracts; insurance; Laffont, J.-J.; land leasing contracts; lending; limited liability; linear contracts; marketing agreements; monitoring costs; moral hazard; multitask moral hazard models; nonlinear contracts; peasants; principal and agent; risk aversion; risk neutrality; risk sharing; sharecropping; shirking; Stiglitz, J.; tenancy ladder; transaction costs

Article

Sharecropping is a form of land leasing contract in which the tenant shares the final product with the landlord as a partial or total payment of the rent. A landowner leasing his land to a tenant may use several forms of land renting contracts.
'Sharecropping' usually designates all particular forms of land tenancy contracts in which the landlord allows the tenant to cultivate his land in return for a stipulated fraction of the product (the 'share'), possibly combined with other side payments. This institutional contractual agreement prevailed in many parts of the world and many different periods in the history of agriculture, from antiquity (Egypt, Mesopotamia and Greece), the Middle Ages and Renaissance in Europe, through to contemporary economies. Sharecropping is currently most commonly found in less developed countries where

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
agriculture and land rental markets are more active, but it also still exists in many developed countries. The sharecropping relationship assumes a variety of forms and is sometimes linked to agreements involving not only land and labour transactions but also credit, lending, insurance or marketing agreements. Indeed, within sharecropping arrangements landlords may also determine the crops to be grown, may choose to monitor some of the key moments of the agricultural process, and may defray a greater or lesser share of the costs of some inputs (other than labour) with a pre-specified fraction that may not necessarily be equal to the fraction of output retained by the landlord in the payment rule.

Is sharecropping inefficient?

Since Adam Smith, economists have taken an interest in sharecropping because of its apparent inefficiency based on the simple observation that the sharecropper receives only a share of the marginal productivity of his labour but bears its full marginal cost. The persistence of such institutions has thus puzzled many economists. More recently, sharecropping has also constituted the typical example of the principal–agent model, the basic paradigm of contract theory. Similar economic relationships occur in both developed and less developed countries when some party (the principal) delegates the use of some capital to another party in exchange for compensation depending on the returns obtained by the other party (the agent). Examples abound in capital markets (stock markets) where investors may let others use their capital in return for a share of the profits, in vertical relationships between producers and retailers in many industries (food and other consumption goods, or rental services), and in some labour contracts within firms where wages may depend on some measure of performance.
Researchers working on the theory of rural organization have attempted to explain not only the persistence of sharecropping but also the particular features it exhibits.

Sharecropping as an efficient risk-sharing contract

Concerns about the efficiency of sharecropping relationships have gone through several stages in the history of economic thought. While it was first thought to be an inefficient institutional arrangement, since Stiglitz (1974) it has been understood as possibly representing an efficient risk-sharing mechanism in environments where production is risky and other forms of insurance are not available. Sharecropping has the advantage over fixed-rent land leasing contracts of relieving the tenant of some of the risk. By sharing the product, the landlord and tenant also share its fluctuations due to risks related to the weather, diseases and other unpredictable factors affecting agricultural production. Through the payment of a rent contingent on agricultural production, the risk associated with variations in prices of marketed commodities is also shared by both parties. However, if the landlord is less risk averse, he should further protect the risk-averse peasant by simply using wage contracts. Moreover, the same risk-sharing opportunities could be provided without sharecropping simply by having workers combine wage contracts and rental contracts.
However, the landlord's ability to monitor the tenant's labour has also been also called into question. In most places where absentee landlords delegate the use of land to a tenant, it seemed implausible that a contract precisely specifying the labour to be applied could be enforceable. Stiglitz (1974) shows that sharecropping could be an institutional arrangement designed both to share risks and to provide

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
incentives in a situation where monitoring effort (labour supply) is costly. Sharecropping results, then, from a trade-off between incentives and risk sharing. Fixed-rent contracts provide ‘perfect’ incentives by giving the full marginal product to the tenant but at the cost of shifting all the risk on to the tenant, while fixed-wage contracts protect the tenant against production risk but also remove direct incentives to provide effort.

However, the two necessary ingredients of this trade-off have been successively challenged. Cheung (1969) criticizes the need to provide incentives through contracted remuneration, arguing that contracts could simply specify the optimal level of labour that the tenant should provide. Wage contracts would not then imply any inefficiency in the provision of effort and would completely insure tenants against income fluctuations. However, this reasoning implicitly assumes that monitoring is not costly, and empirical tests have shown in some contexts that input provision was actually lower under share contracts than under fixed-rent contracts (see, among others, Shaban, 1987). Conversely, the other side of the trade-off was also challenged by the apparent paucity of evidence as to the effect of risk on contractual forms. In fact, until recently a great deal of empirical work had failed to present evidence as to the effect of risk on the contract incentives that would be consistent with the alleged trade-off. This was mostly due to the significance of other trade-offs determining the choice of contracts (which we examine below) but also to the failure to recognize that the choice of contracts had to be modelled within a more general understanding of how land rental markets function. Dubois (2002) shows that taking into account the endogeneity of the choice to delegate use of land was important in the empirical analysis of contractual choices in order to avoid problems of selection bias in econometric estimates. Ackerberg and Botticini (2002) show that not taking into account the endogenous matching of landlords and tenants could also lead to an apparent absence of correlation between the incentive power of contracts and the crop risk. Referring to direct evidence on risk sharing in village economies using consumption data linked to contract choices, Dubois (2000) demonstrates that the sharecropping institution could actually play a role in consumption risk sharing.

Thus, in order to reduce shirking, the landlord could either expend resources in monitoring the worker or prefer to resort to a sharecropping contract. The persistence of sharecropping can thus be explained by this argument together with a number of other observed features. For example, the landlord has an incentive to encourage the tenant to use inputs (such as fertilizer or manure) which raise the worker's marginal product, therefore resulting in higher worker effort. This explains why, as is often observed in sharecropping contracts, the landlord may be prepared to bear a fraction of the costs of inputs that exceeds the fraction of the product received. Obviously, to implement cost sharing, costs have to be observable and verifiable. Why then does the landlord not simply enforce a specific level of input provision by the tenant? This relates to the information structure as to the appropriate level of input, about which the tenant may have better knowledge given the conditions of production. Sharing costs thus remains a useful incentive.

Another argument against interpreting sharecropping as a risk-sharing contract is that the terms of the contracts should logically vary with the level of risk represented by the specific environment, the crops grown, and both parties’ degree of risk aversion. Empirical observation, however, shows that in practice the terms of sharecropping contracts exhibit little variation, especially when it comes to the share of the product, which is often one-half and sometimes one-third or two-thirds. This could be seen to be the result of an approximation process of optimal contracts, but Allen (1985) provides an interesting

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
explanation of this phenomenon. In a model where landlords initially screen tenants with heterogeneous abilities before entering into fixed-rent contracts with only the more able farmers, a sharecropping contract emerges endogenously in a state of equilibrium. Interestingly, the optimal share of production for the tenant in this case has to be a trade-off between the gains accrued from shirking and leaving the relationship after a given period and the gains from not shirking and being taken on again with a fixed-rent contract following a period of screening. This trade-off clearly depends on the farmers' time preferences. Allen (1985) shows that, given the usual interest rates that apply in less developed countries, the optimal share has to be close to one-half.

Another argument calling into question the aforementioned rationale for sharecropping is that the theory of contracts predicts that optimal incentive contracts should depend on output in general in a nonlinear way. However, nonlinear contracts imply that different tenants would face different marginal prices for their production, giving opportunities for arbitrage and incentives to collude among farmers. Linear contracts may then be a way to avoid this problem in environments where monitoring harvest and trade between tenants may be difficult. Moreover, the gains represented by the use of nonlinear contracts may not be worth the potential additional costs of implementing them.

In addition to such attempts to determine the rationale of sharecropping, the expected effect of this kind of rural organization on agricultural innovation and development has been investigated. The adoption of innovations in agriculture in developing countries in particular has been a major preoccupation. Whether the contractual form affects the adoption of innovations, and if so how, have been the subjects of much investigation.

Sharecropping within the principal – agent paradigm

Most theoretical analysis of sharecropping contracts has been cast within the principal–agent paradigm, in which the landlord is generally considered to be the principal having the bargaining power to make a take-it-or-leave-it offer to the agent (the tenant). This analysis does not explain the decision of the landlord to delegate the use of land and the way landlords and tenants meet in the land rental market. It is only recently that these decisions have been taken into account in the analysis of sharecropping contracts, from both empirical and theoretical points of view. In this light, let us consider a model where the agricultural production function is linearly homogenous in land area (as generally admitted: see Stiglitz, 1974, and Otsuka, Chuma and Hayami, 1992) because agriculture is a spatial activity and induces constant returns to scale in the cultivated area. For a fixed amount of land, denote y the agricultural output of the next crop period, e the tenant's work effort (which can be considered as a measure of efficient labour time), x a state variable as, for example, land fertility at the beginning of the agricultural period, and define an agricultural production function f such that $y = \varepsilon f(x, e)$ where ε is a multiplicative positive random variable with mean one representing weather uncertainty. The effort e can represent labour tasks or other agricultural inputs and can be multidimensional. Its cost is $C(e)$. Assume also that an investment function controls land fertility dynamics such that the next period land fertility is $g(x, e)$ (adding a multiplicative positive random variable with mean one to represent the influence of the weather or other externalities on land fertility or ground quality would not change the following results). According to the contract signed, the principal pays the agent $T(y) = \alpha y + \beta$. The contract parameters (α, β) allow the landlord to propose different kinds of contact, from a fixed-wage

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

contract where $\alpha = 0$, $\beta = w$ with w the wage, to a fixed-rent contract where $\alpha = 1$, $\beta = -R$ with R the rent paid to the landlord, through sharecropping contracts where $0 < \alpha < 1$ and β can be zero or not. Concerning preferences, we define $U(T(y)) - C(e)$ and $y - T(y)$ the agent's and principal's utility functions. We assume that U is increasingly concave because of risk aversion, while we treat the principal as risk neutral.

Moral hazard in sharecropping

When the agent's actions are unobservable to the principal or monitoring costs are prohibitively high, a moral hazard problem arises leading to effort shirking by the agent. The worker chooses his effort level to maximize his expected utility, given the terms of the contract and his outside wage opportunities. Thus, for a crop season, the principal proposes a contract to maximize his welfare given the agent's incentive compatibility (*IC*) constraint and its individual rationality (*IR*) constraint guaranteeing him an exogenous reservation utility denoted U. The maximization programme of the landlord can thus be written as

$$\underset{\alpha,\beta}{Max} E[(1-\alpha)y - \beta]$$

subject to

$$e^* \in \arg\max_{e} EU(\alpha y + \beta) - C(e) \quad (1)$$

$$EU(\alpha y + \beta) - C(e^*) \geq \bar{U} \quad (2)$$

Denoting f_e and f_{ee} the first and second derivatives of the production function with respect to effort, we can show that the solution to this programme is such that the individual rationality constraint (*IR*) is binding and the optimal share of production α^* satisfies the following equation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\alpha^* = 1 + \left(1 - \frac{E\varepsilon U'}{EU'}\right)\frac{f[\alpha^* f_{ee} - C'']}{f_e^2}$$

where e_α is the derivative of effort with respect to the share of output α received by the tenant and that satisfies $e_\alpha = \frac{f_e}{\alpha^* f_{ee} - C''}$. Because of the concavity of *U*, $0 < \frac{E\varepsilon U'}{EU'} < 1$. Moreover, with concavity of production with effort and convexity of cost of effort, the optimal share α^* is strictly lower than 1, thus corresponding to a sharecropping contract.

The exact form of the contract depends on both the properties of *U* and *f*, and the magnitude of uncertainty. First, the greater the (compensated) labour supply elasticity, that is, the more sensitive the worker is to incentives, the greater is the optimal share α^*, that is, the closer is the optimal contract to a rental contract. Second, the trade-off between incentives and risk sharing depends on the riskiness of production through ε and on risk preferences through *U*. With some assumptions on the distribution of ε or on the shape of *U*, it can be shown that the optimal share is lower for more risk-averse agents or more risky environments (Stiglitz, 1974). At the limit, if the tenant is risk neutral, then $\alpha^* = 1$ and a pure rental contract will be used. Conversely, the greater the risk aversion and the greater the risk, the closer the optimal contract is to a pure wage contract ($\alpha^* = 0, \beta^* > 0$). These predictions constituted the focus of many empirical tests that involved examining the determinants of the choice between sharecropping and fixed-rent contracts. A large number of empirical tests (Braido, 2005) have failed to provide evidence that the risk-sharing trade-off could explain the choice between fixed rent and sharecropping.

Also, these optimal sharecropping contracts ($0 < \alpha^* < 1, \beta^*$) involve a fixed payment β^* (either to or from the landlord). In practice, many contractual relations may have an implicit or explicit provision calling for such fixed payments. For example, payments from the landlord to the worker to finance stipulated inputs, like fertilizer, can be interpreted in this manner. However, the empirical observation and measurement of such fixed transfers is generally difficult, which explains why they have not been used for empirical testing of the theory.

In some contexts, given the relative paucity of evidence that the risk-sharing trade-off could explain the determinants of contract choice between fixed rent and sharecropping, other explanations related to transaction costs inherent to landlord–tenant relationships have been offered. One of the most significant transaction costs that seem to plague land rental contracts is linked to the question of land quality maintenance and investment. The risk-sharing argument remains completely silent about investment and land-maintenance problems that were often raised in transaction cost approaches with respect to land-rental contracts. Moreover, as already pointed out by Johnson (1950) and even Adam Smith, the problem of land-fertility maintenance and land overuse can also be cited to explain the choice of contract by landlords. In fact, although observable, land fertility may not be contractible. It may also elude verification due to the complexity of specifying the agricultural tasks related to land-quality maintenance and to difficulties in objectively measuring land quality. A moral hazard problem in land maintenance may thus appear. Delegating farming may lead to land overuse if the landlord and tenant do

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

not have the same opportunity cost of usage of land (Allen and Lueck, 1993). A share contract may then curb the farmer's incentive to exploit land attributes. One way to see this in the previous model is to take into account the value of the land in the landlord's objective. The landlord will then anticipate the consequences of delegating the use of land for the future returns obtained, since the tenant's actions may affect the future land value. The land value $v(z)$, an increasing function of the land fertility index z, can be seen as the result of the expected discounted sum of all future profits obtained by the landlord for a given plot of land of quality z. Then, if the objective of the landlord is now

$$\underset{\alpha,\beta}{Max} E[(1-\alpha)y-\beta+v(z)]$$

subject to (*IC*) and (*IR*), the optimal share of output between the landlord and tenant (Dubois, 2002) is the solution to

$$\alpha^{*}=1+v'(z)\frac{g_e}{f_e}-\left(1-\frac{E\varepsilon U'}{EU'}\right)\frac{f}{e_\alpha f_e}$$

where $z=g(x,e)$. With risk neutrality of the tenant, the optimal share is thus below 1 if the effort of production reduces land fertility ($g_e<0$) because

$$\alpha^{*}=1+v'(z)\frac{g_e}{f_e}.$$

The contract here shows low-powered incentives and generally corresponds to a sharecropping contract even if there is no risk-sharing issue.

Multitasks and contract repetition in sharecropping

But other features of agricultural activity and contractual relationships have been used to explain the observation of sharecropping. Several forms of multitask moral hazard models and dynamic considerations provide interesting insights into this form of contracting.
First, the multitask moral hazard model of Holmström and Milgrom (1987), applied to sharecropping for example by Luporini and Parigi (1996), shows that low-powered incentives can be obtained as a way to mitigate the substitution of effort across tasks that cannot be monitored by the landlord even without risk

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
aversion of the tenant. Luporini and Parigi (1996) consider the two distinct production tasks of subsistence crops and cash crops. In another kind of multitask model, with limited liability of the tenant instead of risk aversion, Ghatak and Pandey (2000) show that a sharecropping contract can be optimal when there is joint moral hazard in effort and in risk factor for output. Other quite different models with multiple labour inputs explain sharecropping differently according to a number of features that can be observed from time to time. In Bardhan and Srinivasan (1971) or Eswaran and Kotwal (1985), both the landowner and the tenant provide labour input. In Roumasset and Uy (1987), a model with an investment task, a production task and two periods provides a study in the reduction of agency costs by monitoring. Bardhan (1989, ch. 7) and Braverman and Stiglitz (1986) have a sharecropping model with a fertilizer input and non-observable labour effort. They determine the efficient incentives on both separable inputs through production sharing and cost sharing.

The other significant dimension of the landlord–tenant relationship that may explain the form of contracts is the fact that these contracts are often repeated and may have variable duration. The repetition of relationships between a landlord and a tenant actually called into question the rationale of supposedly short-term contracts of sharecropping generally observed empirically. Bardhan (1989, ch. 8) uses a two-period model to show the trade-off between production incentives, enhanced in the initial period by the threat of dismissal by the landowner, and land improvement incentives that decrease with a more powered contract. Dutta, Ray and Sengupta (1989) and Bose (1993) study a number of long-term contracts between landowners and landless peasants where infinitely repeated relationships with threats of eviction are examined. Eviction threats can actually serve as an incentive device in repeated sharecropping contracts (Banerjee, Gertler and Ghatak, 2002; Banerjee and Ghatak, 2004). Moreover, in a repeated moral hazard relation, spot-contract sequences may allow the outcomes of long-term contracts, which are Pareto-superior to short-term agreements, to be implemented (Fudenberg, Holmström and Milgrom, 1990; Malcomson and Spinnewyn, 1988).

Sharecropping is also sometimes considered to be part of a ‘tenancy ladder’ in agriculture allowing landless wage workers to become farmers before they become landlords. The farmer's financial constraints and limited liability may then play a significant role in explaining access to land rental markets and contractual forms (sharecropping with variable share or fixed-rent contracts) proposed by the landlord (Shetty, 1988; Ray and Singh, 2001; Laffont and Matoussi, 1995). In Laffont and Matoussi (1995), risk-neutral farmers are offered a sharecropping contract with more or fewer incentives rather than a fixed-rent contract, due to financial constraints that restrict the amount of working capital the tenant can use as affected by the rent and the share of inputs to be paid at the beginning of the crop season.

Finally, many other forms of sharecropping contracts, including some side transfers or those interlinked with credit (Mitra, 1983; Braverman and Stiglitz, 1982) or involving state contingent informal gifts and transfers (Sadoulet, Fukui and de Janvry, 1994), exist and may sometimes completely change the efficiency properties of such contracts. Thus, considerable care and attention should be devoted to describing contractual agreements so as to study such organizations and possibly recommend policy reforms for land rental markets.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- access to land and development
- contract theory
- Laffont, Jean-Jacques
- peasants
- peasant economy
- principal and agent (i)
- principal and agent (ii)
- risk sharing
- Stiglitz, Joseph E.

Bibliography

Ackerberg, D. and Botticini, M. 2002. Endogenous matching and the empirical determinants of contract form. *Journal of Political Economy* 110, 564–91.

Allen, F. 1985. On the fixed nature of sharecropping contracts. *Economic Journal* 95, 30–48.

Allen, D. and Lueck, D. 1993. Transaction costs and the design of cropshare contracts. *RAND Journal of Economics* 24, 78–100.

Bardhan, P. 1989. *The Theory of Agrarian Institutions*. Oxford: Clarendon Press.

Bardhan, P. and Srinivasan, T. 1971. Cropsharing tenancy in agriculture: a theoretical and empirical analysis. *American Economic Review* 61, 48–64.

Banerjee, A., Gertler, P. and Ghatak, M. 2002. Empowerment and efficiency: tenancy reform in West Bengal. *Journal of Political Economy* 110, 239–80.

Banerjee, A. and Ghatak, M. 2004. Eviction threats and investment incentives. *Journal of Development Economics* 74, 469–88.

Bose, G. 1993. Interlinked contracts and moral hazard in investments. *Journal of Development Economics* 41, 247–73.

Braido, L. 2005. Insurance and incentives in sharecropping. In *Insurance: Theoretical Analysis and Policy Implications*, ed. P. Chiappori and C. Gollier. Cambridge, MA: MIT Press.

Braverman, A. and Stiglitz, J. 1982. Sharecropping and the interlinking of agrarian markets. *American Economic Review* 72, 695–715.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Braverman, A. and Stiglitz, J. 1986. Cost-sharing arrangements under sharecropping: moral hazard, incentive flexibility, and risk. *American Journal of Agricultural Economics* 68, 642–52.

Cheung, S. 1969. *The Theory of Share Tenancy*. Chicago: University of Chicago Press.

Dubois, P. 2000. Assurance parfaite, hétérogénéité des préférences et métayage au Pakistan. *Annales d'Economie et de Statistique* 59, 1–36.

Dubois, P. 2002. Moral hazard, land fertility and sharecropping in a rural area of the Philippines. *Journal of Development Economics* 68, 35–64.

Dutta, B., Ray, D. and Sengupta, K. 1989. Contracts with eviction in infinitely repeated principal–agent relationships. In *The Theory of Agrarian Institutions*, ed. P. Bardhan. Oxford: Clarendon Press.

Eswaran, M. and Kotwal, A. 1985. A theory of contractual structure in agriculture. *American Economic Review* 75, 352–67.

Fudenberg, D., Holmström, B. and Milgrom, P. 1990. Short-term contracts and long-term agency relationships. *Journal of Economic Theory* 51, 1–31.

Ghatak, M. and Pandey, P. 2000. Contract choice in agriculture with joint moral hazard in effort and risk. *Journal of Development Economics* 63, 303–26.

Holmström, B. and Milgrom, P. 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55, 303–28.

Johnson, D. 1950. Resource allocation under share contracts. *Journal of Political Economy* 58, 111–23.

Laffont, J.-J. and Matoussi, M. 1995. Moral hazard, financial constraints and sharecropping in El Oulja. *Review of Economic Studies* 62, 381–99.

Luporini, A. and Parigi, B. 1996. Multi-task sharecropping contracts: the Italian Mezzadria. *Economica* 63, 445–57.

Malcomson, J. and Spinnewyn, F. 1988. The multiperiod principal–agent problem. *Review of Economic Studies* 40, 391–408.

Mitra, P. 1983. A theory of interlinked rural transactions. *Journal of Public Economics* 20, 167–91.

Otsuka, K., Chuma, H. and Hayami, Y. 1992. Land and labor contracts in agrarian economies: theories and facts. *Journal of Economic Literature* 30, 1965–2018.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ray, T. and Singh, N. 2001. Limited liability, contractual choice, and the tenancy ladder. *Journal of Development Economics* 66, 289–303.

Roumasset, J. and Uy, M. 1987. Agency costs and the agricultural firm. *Land Economics* 63, 290–302.

Sadoulet, E., Fukui, S. and de Janvry, A. 1994. Efficient share-tenancy contracts under risk: the case of two rice-growing villages in Thailand. *Journal of Development Economics* 45, 225–43.

Shaban, R. 1987. Testing between competing models of sharecropping. *Journal of Political Economy* 95, 893–920.

Shetty, S. 1988. Limited liability, wealth differences and tenancy contracts in agrarian economies. *Journal of Development Economics* 29, 1–22.

Stiglitz, J. 1974. Incentives and risk sharing in sharecropping. *Review of Economic Studies* 41, 219–55.

Williamson, O. 1989. Transaction cost economics. In *Handbook of Industrial Organization*, vol. 1, ed. R. Schmalensee and R. Willig. Amsterdam: North-Holland.

How to cite this article

Dubois, Pierre. "sharecropping." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000111> doi:10.1057/9780230226203.1519

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sharpe, William F. (born 1934)

Steven N. Durlauf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

William Sharpe, 1990 co-winner of the Nobel Prize in economics, is one of the founders of the modern theory of finance. His most famous work involves the development of the capital asset pricing model (CAPM), which is now one of the fundamental tools for understanding equilibrium risk–return relationships for different assets.

Keywords

capital asset pricing model; Euler equations; finance; Markowitz, H. M.; mean variance analysis; Miller, M.; risk and return; Sharpe, W. S.; Sharpe ratio; single factor models; single index models

Article

William F. Sharpe is one of the founders of the modern theory of finance. Born in Boston in 1934, Sharpe received his BA in economics from UCLA in 1955. After graduation and military service, Sharpe joined the Rand Corporation and simultaneously pursued his Ph.D. from UCLA, which he received in 1961. He joined the department of economics at the University of Washington–Seattle in 1961, remaining until 1968. After a two-year stint at UC Irvine, he joined the Stanford Business School, where he remained for the rest of his career. In addition to his scholarly pursuits, he has been an active consultant for financial firms as well as textbook writer.
Sharpe's most famous contribution to economics is his development of the capital asset pricing model (CAPM). This work developed in two stages (Varian, 1993, provides a very clear discussion of the evolution of Sharpe's work on CAPM). First, mentored by Harry Markowitz, who was also at Rand, Sharpe studied the question of the construction of efficient portfolios in the presence of a riskless asset. This led to Sharpe's study of what he dubbed ‘single factor’ models but which are now more often called ‘single index’ models, in which the holding return on a given asset is a linear function of the return on the market portfolio. The single factor approach provides important computational advantages in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
constructing optimal portfolios of the type studied by Markowitz. This doctoral dissertation research was subsequently published in *Management Science* in 1963 as ‘A Simplified Model for Portfolio Analysis’. The analysis of single factor asset return models as a short cut to efficient portfolio construction was followed by Sharpe's investigation of what equilibrium risk–return relationships will emerge in a market of rational agents with mean/variance preferences, leading to his celebrated 1964 *Journal of Finance* paper ‘Capital Asset Prices – A Theory of Market Equilibrium Under Conditions of Risk’. The demonstration that the riskiness of an asset is determined not by the variance of its holding return but rather by the now celebrated ‘beta’ of the asset, defined as the covariance of that holding return with the holding return on the market portfolio as a whole divided by the standard deviation of the market portfolio, is now a canonical idea in economics and is the basis for much of the modern theory of asset pricing. The underlying economic ideas of the CAPM find modern analogs in the use of Euler equations to characterize equilibrium expected asset returns. The main difference in Euler equation approaches from Sharpe's original formulation in this later work is the relaxation of the assumption that market participants have mean variance preferences with respect to asset returns; in its place explicit consumption-based utility functions are used. It is interesting to note that Sharpe's single factor model embodied the CAPM risk–return relationship by construction. Sharpe's Nobel Prize acceptance speech, published as Sharpe (1991b), is interesting for encapsulating his assessment of the model. See capital asset pricing model for discussion of the model in detail as well as its intellectual history; as Sharpe notes in his Nobel autobiography, a number of researchers were working on similar ideas to his.
Sharpe's subsequent research has spanned a wide range of issues in finance. Perhaps most noteworthy, Sharpe's interest in understanding risk and return relationships led to his development, initially in the context of mutual fund evaluation, of what is now called the Sharpe ratio, first discussed in his 1966 *Journal of Business* paper ‘Mutual Fund Performance’. (Sharpe called it the reward to variability ratio.) The ratio is measured by taking the difference between the expected return on an asset (or portfolio) and a benchmark security and dividing by the standard deviation of this difference. When the benchmark security is riskless, the Sharpe ratio provides a simple characterization of the return to risk. Sharpe (1994) gives a nice summary of the statistic and its interpretation.
These achievements led to Sharpe's receipt, with Merton Miller and Harry Markowitz, of the 1990 Nobel Memorial Prize in economics as one of the ‘pioneers in the theory of financial economics and corporate finance’. Varian (1993) provides a lovely discussion of why this joint award was so merited.

See Also

- capital asset pricing model

Selected works

1963. A simplified model for portfolio analysis. *Management Science* 9, 277–93.

1964. Capital asset prices – a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1965. Risk-aversion in the stock market – some empirical evidence. *Journal of Finance* 20, 416–22.

1966. Mutual fund performance. *Journal of Business* 39(1), Part II, 119–38.

1974. Imputing expected returns from portfolio composition. *Journal of Financial and Quantitative Analysis*, June, 463–72.

1976. Corporate pension funding policy. *Journal of Financial Economics* 3, 183–93.

1977. The capital asset pricing model: a ‘multi-beta’ interpretation. In *Financial Decision Making Under Uncertainty*, ed. H. Levy and M. Sarnat. New York: Academic Press.

1978. *Investments*. Englewood Cliffs, NJ: Prentice-Hall.

1978. (With R. Lanstein.) Duration and security risk. *Journal of Financial and Quantitative Analysis*, November, 653–68.

1981. Decentralized investment management. *Journal of Finance* 36, 217–34.

1989. (With G. J. Alexander.) *Fundamentals of Investments*. Englewood Cliffs, NJ: Prentice-Hall.

1991a. Autobiography. In *Les Prix Nobel. The Nobel Prizes 1990*, ed. T. Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1990/sharpe-autobio.html, accessed 11 October 2006.

1991b. Capital asset prices with and without negative holdings. *Journal of Finance* 46, 489–509.

1994. The Sharpe Ratio. *Journal of Portfolio Management* 21(1), 49–58.

2006. *Investors and Markets: Portfolio Choices, Asset Prices, and Investment Advice*. Princeton: Princeton University Press.

Bibliography

Varian, H. 1993. A portfolio of Nobel Laureates: Markowitz, Miller, and Sharpe. *Journal of Economic Perspectives* 7(1), 159–69.

How to cite this article

Durlauf, Steven N. "Sharpe, William F. (born 1934)." The New Palgrave Dictionary of Economics.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000441> doi:10.1057/9780230226203.1520

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Shephard, Ronald William (1912 – 1982)

W.E. Diewert
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cost fiunctions; distance function; duality; homothetic production function; index numbers; production functions; Shephard, R. W.; Shephard's lemma

Article

Shephard was born in Portland, Oregon, or 22 November 1912 and died in Berkeley, California, on 22 July 1982.
He received his BA in Mathematics and Economics in 1935 and his Ph.D. in Mathematics and Statistics in 1940 at the University of California at Berkeley.
During the years 1943–6 he was a statistical consultant at the Bell aircraft corporation. In the years 1949–51 he worked under the direction of Oskar Morgenstern of Princeton University, producing his path-breaking work, *Cost and Production Functions* (1953). During 1950–2, he was a senior economist at the RAND Corporation and during 1952–6 he was the manager of the systems analysis department at the Sandia Corporation. From 1957 to 1980 he was a Professor of Industrial Engineering and Operations Research at the University of California at Berkeley.
Shephard made several fundamental contributions to economics. He was the first to rigorously derive a duality between cost and production functions; that is, given a knowledge of either function, the other may be derived from it. He also introduced the distance function to the economics literature in the course of establishing his duality theorems; the distance function is used to define a theoretical index number concept due to Malmquist. Shephard was also the first to derive the derivative property of the cost function (or Shephard's Lemma) starting from the cost function (the derivations by Hicks and Samuelson started from the production or utility function). Shephard also appreciated the econometric implications of Shephard's Lemma.
Shephard also defined the concept of a homothetic production or utility function: a function is homothetic if it is a monotonic transform of a linearly homogeneous function. He also deduced the implications of a homothetic function for its dual cost function.
Shephard also realized the importance of the assumption of homogeneous weak separability for index

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
number and aggregation theory.
Finally, Shephard postulated an ingenious system of axioms or properties for a production function and then was able to deduce the classical law of diminishing returns to a subset of the factors as a theorem.

See Also

- cost functions

Selected works

1953. *Cost and Production Functions*. Princeton: Princeton University Press.

1970a. *Theory of Cost and Production Functions*. Princeton: Princeton University Press.

1970b. Proof of the law of diminishing returns. *Zeitschrift für Nationalökonomie* 30, 7–34.

How to cite this article

Diewert, W. E. "Shephard, Ronald William (1912–1982)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000116> doi:10.1057/9780230226203.1521

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

shrinkage-biased estimation in econometrics

George G. Judge
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The maximum likelihood estimation principle, unbiasedness and hypothesis testing serve as foundation stones for much that goes on in the lives of theoretical and applied econometricians. In this context, the purpose of these words and other symbols is to review the statistical implications of pursuing these estimation and inference goals and to suggest superior alternatives.

Keywords

estimation; Gauss–Markov theorem; inference; James and Stein estimator; least squares; linear models; maximum likelihood estimator; maximum likelihood; multivariate estimation; pretest estimation; semiparametric estimation; shrinkage-biased estimation in econometrics; Stein rule estimator; testing in econometrics

Article

In economics much empirical research proceeds in the context of incomplete subject matter theories and data based on sampling designs not devised by or known to the researcher. This leads to economic or econometric models in the form of ill-posed inverse problems, and partial or incomplete data, and brings a range of uncertainty to the data-processing and information-recovery process in general and the estimation and inference process in particular. In practice, procedures such as preliminary test statistics, tuning parameters and perhaps a bit of magic are invoked to identify a particular econometric-statistical model on which to base estimation and inference process. Given the uncertainty surrounding the model-discovery and post-data estimation and inference tasks, one basis for coping with or reducing the entropy level is to focus on the statistical implications of shrinkage estimation and the possibilities for combining competing estimation problems. The objective is to demonstrate estimation and inference methods that are free of subjective choices and tuning parameters and that have superior risk performance. In the process we demonstrate simple estimators that are uniformly and non-trivially superior over the unknown parameter space to conventional parametric and pretest estimators used by most applied econometrics researchers.

1 The conventional statistical model and estimator base

In econometrics the most widely used estimation and inference techniques are based on linear statistical models and maximum likelihood (ML) and least squares concepts. These estimation and inference methods are supported by a body of theory that dates back over two centuries to Gauss and Legendre and includes the often cited Gauss–Markov theorem with its best linear unbiased estimator conclusion – a conclusion that appears to be generally accepted by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

applied econometricians. However, for the economic researcher who is interested in parameter estimation this statistical property, which is right on average, may have limited usefulness and also negative statistical implications. It is to these questions that we now turn.

In econometrics, many multivariate estimation problems can be reduced to the canonical form where $\mathbf{b} = (b_1, b_2, \ldots, b_K)'$ is a K variate normal random vector with $\mathbf{b} \sim N_K(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathbf{b}})$. The mean vector β is unknown and the covariance $\boldsymbol{\Sigma}_{\mathbf{b}}$ is usually assumed to be known up to a constant of proportionality. One common problem that gives rise to the above is estimating the location vector β for the linear statistical model when we observe a vector $\mathbf{y}$ such that $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{e} \sim N\left(\mathbf{0}, \sigma^2 I_T\right)$ and $\delta(\mathbf{b}) = \left(X'X\right)^{-1}X'\mathbf{y}$ is the maximum likelihood estimator (MLE) with covariance $\boldsymbol{\Sigma}_{\mathbf{b}} = \sigma^2\left(X'X\right)^{-1}$. The objective is to estimate the unknown vector β by an estimator $\delta(\mathbf{b})$ under the sum of squares of error loss measure

$$L(\boldsymbol{\beta}, \delta(\mathbf{b})) = \|\boldsymbol{\beta} - \delta(\mathbf{b})\|^2 = \sum_{i=1}^{K} (\beta_i - \delta(b_i))^2 \tag{1.1}$$

where the unknown loss is a function of both parameters and data. The usual evaluation of the estimator $\delta(\mathbf{b})$ makes use of the risk function.

$$\rho(\boldsymbol{\beta}, \delta(\mathbf{b})) = E_{\boldsymbol{\beta}}^{\mathbf{b}}\left[L(\boldsymbol{\beta}, \delta(\mathbf{b}))\right] = E_{\boldsymbol{\beta}}^{\mathbf{b}}\left[\|\boldsymbol{\beta} - \delta(\mathbf{b})\|^2\right] \tag{1.1a}$$

where superscripts denote the random quantity over which the expectation is to be taken for fixed β. Under (1.1), the maximum likelihood estimator $\delta(\mathbf{b}) = \mathbf{b}$ is minimax and has constant risk $\rho(\boldsymbol{\beta}, \delta(\mathbf{b})) = \mathrm{tr}\boldsymbol{\Sigma}_{\mathbf{b}} = \sigma^2 \mathrm{tr}\left(X'X\right)^{-1}$.

1.1 The Stein alternative

The first hint of difficulty for the MLE $\delta(\mathbf{b})$ in estimating the multivariate normal mean under quadratic loss (1.1) was when Stein (1955) demonstrated for the orthonormal symmetric case, $\boldsymbol{\Sigma}_{\mathbf{b}} = \sigma^2\left(X'X\right)^{-1} = I_K$, that the conventional estimator $\delta(\mathbf{b})$ is inadmissible when $K \geq 3$. This means there exists another estimator, say $\delta^S(\mathbf{b})$, where $\rho\left(\delta^S(\mathbf{b}), \boldsymbol{\beta}\right) \leq \rho(\delta(\mathbf{b}), \boldsymbol{\beta}) = K$ for all β and $\rho\left(\delta^S(\mathbf{b}), \boldsymbol{\beta}\right) < \rho(\delta(\mathbf{b}), \boldsymbol{\beta}) = K$, for some β. In other words, under the usual measure of statistical performance, there is a superior estimator. Stein's inadmissibility proof did not go on to suggest the components of a risk dominating alternative estimator.

1.1.1 The James and Stein shrinkage estimator

Given the inadmissibility result that suggested there may be under quadratic loss other estimators that risk dominate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the MLE, **b**, James and Stein (1961) demonstrated the estimator

$$\delta^{S}(\mathbf{b}) = \left(1 - \frac{K-2}{\|\mathbf{b}\|^{2}}\right)\mathbf{b} \tag{1.2}$$

that has uniformly smaller risk than **b**. This estimator makes the adjustment in the MLE, **b**, a smooth and known function of the data. The mean of (1.2) is

$$E\left[\delta^{S}(\mathbf{b})\right] = \boldsymbol{\beta} - (K-2)E\left[1/\chi^{2}_{(K+2,\lambda)}\right]\boldsymbol{\beta} \tag{1.2a}$$

with risk

$$\rho\left(\delta^{S}(\mathbf{b}), \boldsymbol{\beta}\right) = K - (K-2)^{2}E\left[1/\chi^{2}_{(K,\lambda)}\right] \tag{1.2b}$$

where $\chi^{2}_{(K+2,\lambda)}$ is a non-central chi square random variable with non centrality parameter $\lambda = \boldsymbol{\beta}'\boldsymbol{\beta}/2$ (see Judge and Bock, 1978). When $\boldsymbol{\beta} = 0$ the risk of (1.2) is 2 and increases to K, the risk of the MLE, **b**, as $\boldsymbol{\beta}'\boldsymbol{\beta} \to \infty$. Consequently, for the values of β close to the origin the risk gain may be considerable.

The James and Stein estimator (1.2) combines the MLE, **b**, and the restricted-fixed vector, $\boldsymbol{\beta} = 0$, and shrinks **b** toward the null mean vector, $\boldsymbol{\beta} = 0$. A more general formulation that introduces explicitly an arbitrary origin, considers a fixed mean vector **r**, and an estimator of the form

$$\delta^{S}_{r}(\mathbf{b}) = \left[1 - \frac{(K-2)}{\|\mathbf{b}-\mathbf{r}\|^{2}}\right](\mathbf{b}-\mathbf{r}) + \mathbf{r} \tag{1.3}$$

which shrinks **b** toward the target vector $\mathbf{r} \in R^{K}$. This estimator has bias and risk characteristics in line with the James and Stein estimator (1.2).

If σ^{2} is unknown, the optimal James and Stein estimator may be written as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\delta^{S}(\mathbf{b}) = \left[1 - ((K-2)/(T-K+2))\left(s/\mathbf{b}'\mathbf{b}\right)\right]\mathbf{b}$$
(1.4)

where $s/\sigma^2 = (T-K)\hat{\sigma}^2/\sigma^2$ has a $\chi^2_{(T-K)}$ random variable distribution that is independent of **b**. Since $\mathbf{b}'\mathbf{b}/\sigma^2$ is distributed as a χ^2_K, the optimal James and Stein estimators (1.4) may be rewritten as

$$\delta^{S}(\mathbf{b}) = [1 - (T-K)(K-2)/(T-K+2)(1/u)]\mathbf{b}$$
(1.5)

where $u = \mathbf{b}'\mathbf{b}/\left(K\hat{\sigma}^2\right)$ is the likelihood ratio statistic which has an F distribution with K and (T–K) degrees of freedom and non-centrality parameter $\lambda = \mathbf{b}'\mathbf{b}/2\sigma^2$. Thus the shrinkage of the MLE is determined by the data and the hypothesis vector, which in this case is, $\boldsymbol{\beta} = \mathbf{0}$. The larger the value of the F test statistic, the smaller is the adjustment made in the MLE. It will be useful to keep in mind this continuous likelihood ratio shrinkage estimator when we discuss pretest estimators in Section 2.

Some of the Stein-like shrinkage estimators have desirable properties from both sampling theory and Bayesian inference points of view. For examples, when σ^2 is unknown the empirical Bayes counterpart is

$$\delta^{B}(\mathbf{b}) = \left[1 - ((K-2)/(T-K))\left(s/\mathbf{b}'\mathbf{b}\right)\right]\mathbf{b}$$
(1.6)

This estimator is dominated by (1.4).

1.1.2 A positive Stein shrinkage rule

Although under quadratic loss Stein rules improve on the minimax MLE and are themselves minimax, they are not admissible and other superior shrinkage rules exist. One such estimator is the positive Stein rule estimator.

$$\delta^{+S}(\mathbf{b}) = \left(1 - \frac{c}{\|\mathbf{b}\|^2}\right)\mathbf{b}I_{[c,\infty)}\left(\|\mathbf{b}\|^2\right) = \mathbf{b} - \left(1 \wedge \frac{c}{\|\mathbf{b}\|^2}\right)\mathbf{b}$$
(1.7)

where $a \wedge \mathbf{b} = \min(a, \mathbf{b})$. Under this formulation, Baranchik (1964) demonstrated that for $(K-2) \le c < 2(K-2)$ the positive rule estimator (1.7) uniformly improves on the James and Stein estimator (1.2) and thus proves its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

inadmissibility. There is no one value of c that is optimal, but Efron and Morris (1973) have demonstrated that rules that restrict c to [(K–2), 2(K–2)], dominate shrinkage rules of c in [(0,(K–2)]. Although the positive rule estimator (1.7) is minimax under quadratic loss, it too is inadmissible.

For the more general non-symmetric case where $\Sigma_{\mathbf{b}}$ is just some positive definite symmetric matrix, the class of pseudo Stein–Bayes rules, $\delta^{S}(\mathbf{b})$, having uniformly smaller risk than the MLE, $\delta(\mathbf{b})$, is very large. For example, if we let $(K-2)/(T-K+2)=a$ and $s=(T-K)\hat{\sigma}^{2}$, we have the following Stein-type estimator proposed by Judge and Bock (1978, p. 240):

$$\delta(\mathbf{b}, s)=\left(1-\frac{as}{\mathbf{b}'X'X\mathbf{b}}\right)\mathbf{b} \tag{1.8}$$

which, under squared loss is minimax if $0 \le a \le 2\left(\mathrm{tr}\left(X'X\right)^{-1}d_L^{-1}-2\right)/(T-K+2)$ and $\mathrm{tr}\left(X'X\right)^{-1} > 2d_L$, where d_L is the smallest characteristic root of $X'X$. To prevent overshrinking, the positive rule estimator $\delta(\mathbf{b}, s)^{+}=c^{+}\mathbf{b}$, which dominates $\delta(\mathbf{b}, s)=c\mathbf{b}$, where $c^{+}=\max\{c, 0\}$, should be used (Judge and Bock, 1978, p. 246).

1.1.3 Implications

In the previous subsections we have considered the problem where the econometrician wishes to estimate the parameters of a K dimensional vector $\boldsymbol{\beta}=(\beta_1, \beta_2, \ldots, \beta_K)'$, where β is the mean of an independent normal random vector $\mathbf{b} \sim N(\boldsymbol{\beta}, I_K)$ and β is unknown. For this problem James and Stein (1961) have demonstrated for $K \ge 3$ that the estimator

$$\delta^{S}(\mathbf{b})=\left(1-(K-2)/\|\mathbf{b}\|^{2}\right)\mathbf{b} \tag{1.9}$$

is uniformly better than the MLE, **b**, under quadratic loss. This result holds for a range of linear statistical models and corresponding ML estimators. Given the high esteem in which the MLE is held, this seems at first blush to be impossible. In this estimator the estimate of each β_i depends not only on b_i, but also on the other b_i whose distributions are apparently independent of b_i. The result is a risk improvement in the MLE regardless of the values of β_i. The reaction of statisticians and econometricians to this seeming magic has been less than overwhelming. However, after a half century, using this shrinkage idea as one way of dealing with model uncertainty seems to be firmly established and, as we see in the next section, has led to the more general idea of combining estimation problems (Efron and Morris, 1973).

1.2 Combining estimation problems

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The variants of the James and Stein estimators discussed in section 1.1 achieve their risk advantages by shrinking to a fixed vector. Another alternative is shrinking to a random vector and in this context Lindley (1962) suggested the estimator

$$\delta^2(\mathbf{b}) = \bar{\mathbf{b}} + \left\{1 - (K-3)/\|\mathbf{b}\|^2\right\}(\mathbf{b} - \bar{\mathbf{b}}) \tag{1.10}$$

that shrinks toward the grand mean $\bar{\mathbf{b}}$. This estimator dominates the MLE competitor when $K \geq 4$ and does especially well risk-wise when the b_i are near each other. Shrinking toward a random vector suggests consideration of two estimation problems and two corresponding estimators that may have different sampling characteristics. Moving in this direction, Green and Strawderman (1991), in the spirit of the James and Stein estimator, proposed an estimator that involves the best weighted linear combination of two estimation problems, where 'best' is defined in terms of quadratic loss. An important point in their formulation is that the quadratic loss criterion is introduced up front and not considered as a result of the estimation process.

Although the Stein-rules of Section 1.1 may be developed from an empirical Bayes base (Judge and Bock, 1978, pp. 173–5), how the shrinkage rule (1.2) came about is a mystery. To make the combining rule transparent we develop the weighted linear combination estimator in some detail. For expository purposes, we continue to consider the orthonormal statistical model

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e};\ \mathbf{e} \sim \left(\mathbf{0}, \sigma^2 I_n\right) \text{ and } X'X = I_K \tag{1.11}$$

and the following ML estimator $\mathbf{b} = \hat{\boldsymbol{\beta}} = \left(X'X\right)^{-1}X'\mathbf{y} \sim \left(\boldsymbol{\beta}, \sigma^2 I_K\right)$, and a biased competitive estimator, $\tilde{\boldsymbol{\beta}} \sim \left(\boldsymbol{\beta} + \delta, \tau^2 I_K\right)$, where δ is a bias vector and $\tau^2 < \sigma^2$. For convenience assume σ^2 and τ^2 are known. The objective is to determine the best linear combination of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ where performance is evaluated in terms of quadratic loss. Under quadratic loss the risk of $\hat{\boldsymbol{\beta}}$ is $\sigma^2 K$ and the risk of $\tilde{\boldsymbol{\beta}}$ is $\tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}$. This suggests that over the parameter space $\boldsymbol{\beta}'\boldsymbol{\beta}$ the risk functions of the two estimators may cross.

Given this situation, the question is whether there exists a weighted linear combination of the two estimators that leads to a combined estimator

$$\gamma\left(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) = \alpha\hat{\boldsymbol{\beta}} + (1-\alpha)\tilde{\boldsymbol{\beta}} \tag{1.12}$$

that risk dominates $\hat{\boldsymbol{\beta}}$. The risk of the linear combination of the two estimators $\gamma(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})$ is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\rho\left(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) = \alpha^2\sigma^2 K + (1-\alpha)^2\left(\tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}\right)$$
(1.13)

and the value of the mixing parameter α that minimizes (1.13) must satisfy

$$d\rho\left(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) / d\alpha = -\alpha\left(\sigma^2 K + \tau K + \boldsymbol{\delta}'\boldsymbol{\delta}\right) + \left(\tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}\right) = 0$$
(1.14)

such that

$$\alpha = \frac{\tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}}{\sigma^2 K + \tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}}$$
(1.15)

Consequently, we may write the minimum risk combination of the two estimators in (1.12) as

$$\gamma\left(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) = \frac{\left(\tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}\right)}{\left(\tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta} + \sigma^2 K\right)}\hat{\boldsymbol{\beta}} + \frac{\sigma^2 K}{\sigma^2 K + \tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}}\tilde{\boldsymbol{\beta}}$$
(1.16)

Since $E\left[\left\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right\|^2\right] = \sigma^2 K + \tau^2 K + \boldsymbol{\delta}'\boldsymbol{\delta}$, if we substitute $\left\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right\|^2$ for its expected value (Judge and Bock, 1978, p. 175), we may rewrite (1.16) as

$$\gamma\left(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) = \tilde{\boldsymbol{\beta}} + \left(1 - \sigma^2 K / \left\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right\|^2\right)\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right)$$
(1.17)

which is in the form of a shrinkage estimator, where $\hat{\boldsymbol{\beta}}$ is shrunk toward the biased estimator $\tilde{\boldsymbol{\beta}}$, with K as an approximation for the best linear combination of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the context of Section 2 it is straightforward to demonstrate that $\gamma(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})$ risk dominates the MLE, $\hat{\boldsymbol{\beta}}$, with K–2 as the uniformly best combination. Building on this work, Judge and Mittelhammer (2004) generalized the orthonormal independent estimator combining problem and developed natural adaptive semiparametric estimation and inference methods for dependent estimators. The resulting semiparametric estimators are free of parametric choices and tuning parameters, and have good asymptotic and finite sample properties and superior risk performance. Extending Stein-like estimation to include a random shrinkage vector greatly extends the applicability of combining type estimators for a range of problems in econometric theory and practice.

2 Traditional pretest estimation

When there is uncertainty about the econometric model and thus the appropriate restrictions or hypothesis to impose, a traditional way to proceed is by statistical hypotheses testing based on the data at hand. The econometric literature abounds with exact and approximate tests for identifying sins of omission or commission relative to a variety of possibly false models. Although the two-stage estimation rule that results is used routinely in applied work, the econometric literature is strangely silent as to the statistical properties of the resulting two-stage estimator, or its sequential counterpart. To see the possible statistical significance of this two-stage process, continue with the orthonormal linear statistical model and notation used in Section 1.2 and follow Judge and Bock (1978; 1983). Under the statistical model and a fixed vector $\mathbf{r}$ and the MLE, $\hat{\boldsymbol{\beta}}$ we may use likelihood ratio procedures to test the null hypothesis $H_0: \boldsymbol{\beta} = \mathbf{r}$ against the hypothesis $\boldsymbol{\beta} \neq \mathbf{r}$, by using the test statistic

$$u = \left(\hat{\boldsymbol{\beta}} - \mathbf{r}\right)'\left(\hat{\boldsymbol{\beta}} - \mathbf{r}\right) / K\hat{\sigma}^2, \tag{2.1}$$

which, if the hypotheses (restrictions) are correct, is distributed as a central F random variable with K and (T–K) degrees of freedom. Of course if the restrictions are incorrect $E\left[\hat{\boldsymbol{\beta}} - \mathbf{r}\right] = (\boldsymbol{\beta} - \mathbf{r}) = \boldsymbol{\delta} \neq \mathbf{0}$, and u is distributed as a non-central F with non-centrality parameter $\lambda = (\boldsymbol{\beta} - \mathbf{r})'(\boldsymbol{\beta} - \mathbf{r}) / 2\sigma^2 = \boldsymbol{\delta}'\boldsymbol{\delta} / 2\sigma^2$. As a test mechanism the null hypothesis is rejected if $u \geq F^{\alpha}_{(K,T-K)} = c$, where c is determined for a given level of the test α by $\int_c^{\infty} dF_{(K,t-K)} = P\left[F_{(K,T-K)} \geq c\right] = \alpha$. This means that by accepting the null hypothesis we use the restricted least squares estimator $\boldsymbol{\beta}^* = \mathbf{r}$ as our estimate of β, and by rejecting the null hypothesis $\boldsymbol{\beta} - \mathbf{r} = \boldsymbol{\delta} = \mathbf{0}$ we use the unrestricted least squares estimator, $\hat{\boldsymbol{\beta}}$. The estimate that results is dependent upon a preliminary test of significance and this means the estimator used by many applied workers is of the form

$$\hat{\boldsymbol{\beta}}^* = \begin{cases} \boldsymbol{\beta}^* & \text{if } u < c, \\ \hat{\boldsymbol{\beta}} & \text{if } u \geq c. \end{cases} \tag{2.2}$$

Alternatively the estimator may be written as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\widehat{\boldsymbol{\beta}}^{*} = I_{(0,c)}(u)\boldsymbol{\beta}^{*} + I_{[c,\infty)}(u)\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} - I_{(0,c)}(u)\left(\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)$$
(2.3)

where $I_A(u)$ is an indicator function that takes the value 1 when $u \in A$ and takes the value 0 otherwise. This specification means that in a repeated sampling context the data, the linear hypotheses, and the selected level of statistical significance determine the combination of the two estimators that is chosen. From (2.3) the mean of the pretest estimator is

$$E\left[\widehat{\boldsymbol{\beta}}^{*}\right] = \boldsymbol{\beta} - E\left[I_{(0,c)}(u)\left(\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)\right]$$
(2.4)

which by theorem 2.1 in Judge and Bock (1978, p. 71) may be expressed as

$$E\left[\widehat{\boldsymbol{\beta}}^{*}\right] = \boldsymbol{\beta} - \delta P\left[\chi^2_{(K+2,\lambda)} / \chi^2_{(T-K)} \le cK / (T-K)\right]$$
(2.5)

Consequently, if $\boldsymbol{\delta} = \mathbf{0}$, the pretest estimator is unbiased. This fortunate outcome aside, the size of the bias is affected by the probability of a random variable with a non-central F distribution being less than a constant, which is determined by the level of the test, the number of hypotheses, and the degree of hypothesis error, δ or λ. Since the probability is always equal to or less than one, the bias of the pretest estimator is equal to or less than the bias of the restricted estimator β^*. Following Judge and Bock (1978, p. 70) and using the discontinuous estimator rule (2.3), we may express the risk function for the pretest estimator, $\widehat{\boldsymbol{\beta}}^{*}$, as

$$\rho\left(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}^{*}\right) = E\left[\left(\widehat{\boldsymbol{\beta}}^{*} - \boldsymbol{\beta}\right)'\left(\widehat{\boldsymbol{\beta}}^{*} - \boldsymbol{\beta}\right)\right] = \sigma^2 K + \left(2\boldsymbol{\delta}'\boldsymbol{\delta} - \sigma^2 K\right) P\left[\frac{\chi^2_{(K+2,\lambda)}}{\chi^2_{(T-K)}} \le \frac{cK}{T-K}\right] - \boldsymbol{\delta}'\boldsymbol{\delta}\, P\left[\frac{\chi^2_{(K+4,\lambda)}}{\chi^2_{(T-K)}} \le \frac{cK}{T-K}\right].$$
(2.6)

Defining the terms in brackets as L(2) and L(4), we may write (2.6) compactly as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\rho\left(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}^{*}\right) = \sigma^2 K + \left(2\boldsymbol{\delta}'\boldsymbol{\delta} - \sigma^2 K\right)L(2) - \boldsymbol{\delta}'\boldsymbol{\delta}L(4),$$

(2.7)

where $1 > L(2) > L(4) > 0$. From the risk function (2.6) the following characteristics of the pretest estimator $\widehat{\boldsymbol{\beta}}^{*}$ emerge: (*i*) if the restrictions are correct and $\boldsymbol{\delta} = 0$, the pretest estimator has a smaller risk than the ML estimator $\widehat{\boldsymbol{\beta}}$ at the origin, $\boldsymbol{\delta} = 0$, and the risk depends on the level of significance α and correspondingly the critical value of the test c; (*ii*) as the hypothesis error δ or λ grows, the risk of the pretest estimator $\widehat{\boldsymbol{\beta}}^{*}$ increases, obtains a maximum after exceeding the risk of the MLE, $\widehat{\boldsymbol{\beta}}$, and then monotonically decreases to approach $\sigma^2 K$, the risk of the MLE; (*iii*) as the hypothesis error $\widehat{\boldsymbol{\beta}} - \mathbf{r} = \delta$ increases and approaches infinity, the risk of the pretest estimator approaches $\sigma^2 K$, the risk of the MLE, from above; and (*iv*) the risk of the pretest estimator, $\widehat{\boldsymbol{\beta}}^{*}$, varies with α the chosen level of significance. Thus, if one is to use the estimator $\widehat{\boldsymbol{\beta}}^{*}$, the question as to the optimal level of significance α remains. Finally, we note that Sclove, Morris and Radhakrishman (1972) demonstrated that the Stein-rule estimator

$$\delta_u^S\left(\widehat{\boldsymbol{\beta}}\right) = I_{[c,\infty)}(u)\left(1 - c^{*}/u\right)\left(\widehat{\boldsymbol{\beta}} - \mathbf{r}\right) + \mathbf{r}$$

(2.8)

where $c^{*} = (T-K)(K-2)K^{-1}(T-K+2)^{-1}$ and u is the likelihood ratio statistic defined in (2.1), is under quadratic loss uniformly superior to the pretest estimator, $\widehat{\boldsymbol{\beta}}^{*}$, thus proving its inadmissibility.

3 Concluding remarks

Post-data evaluation procedures constitute a rejection of the concept of a true econometric model for which econometric theory provides a basis for estimation and inference. In this context, if a researcher is willing to forgo the property of unbiasedness, the Stein family of nonlinear biased estimation rules provides, under quadratic loss and conditions normally found in practice, a uniformly superior alternative. These estimators that shrink maximum likelihood estimates towards zero or some predetermined coordinate enjoy good statistical properties from both sampling theory and Bayesian points of view. Extension of the Stein-rule idea to weighted linear combining estimation problems leads to estimators that can be recommended from the standpoint of simplicity, generalizability and efficiency, and robustness over distribution assumptions and loss functions. If, in an information-theoretic context, one wishes to leave the parametric Stein-rule family, minimum divergence estimators that involve reference and subject distributions and a choice of distance measures offer attractive semiparametric estimation and inference alternatives (Mittelhammer et al., 2005).

Estimators that evolve by a two-stage estimation and then hypothesis-testing process lead to estimation rules that are (*i*) risk inferior, over a large range of the parameter space, to the data-based maximum likelihood estimator, and (*ii*) uniformly risk-inferior to Stein-rule alternatives. Recognizing this unfortunate situation. one of my colleagues once remarked that pretest estimation (single or repeated hypothesis tests) is the biggest unreported scandal in inferential statistics. My only quibble with this statement is that the statistical implications of pretesting have been reported for over a half-century. Hypothesis testing is like an addictive drug and has led econometricians to produce a plethora of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
conditional test statistics, to higher consumption of old and new hypothesis tests, and to model-discovery processes that lead to pretest estimators with negative and unknown sampling performance. Perhaps it is time for econometricians to just say ‘no’!

See Also

- econometrics
- linear models
- maximum likelihood
- semiparametric estimation
- testing

Bibliography

Baranchik, A. 1964. *Multiple Regression and Estimation of the Mean of Multivariate Normal Distribution*. Technical Report 51. Stanford University, Department of Statistics.

Bock, M., Yancey, T. and Judge, G. 1973. The statistical consequences of preliminary test estimators in regression. *Journal of American Statistical Association* 68, 107–10.

Efron, B. and Morris, C. 1973. Combining possibly related estimation problems. *Journal of the Royal Statistical Society*, Series B, 35, 379–421.

Green, E. and Strawderman, W. 1991. James-Stein-type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Associations* 86, 1001–6.

James, W. and Stein, C. 1961. Estimation with quadratic loss. In *Proceedings of the Fourth Berkley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley and Los Angeles: University of California Press.

Judge, G. and Bock, M. 1978. *The Statistical Implications of Pretest and Stein-Rule Estimators in Econometrics*. New York: North-Holland.

Judge, G. and Bock, M. 1983. Biased estimation. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Judge, G., Hill, R., Griffiths, W., Lutkepohl, H. and Lee, T. 1988. *Introduction to the Theory and Practice of Econometrics*, ch. 20. New York: John Wiley.

Judge, G. and Mittelhammer, R. 2004. A semiparametric basis for combining estimation problems under quadratic loss. *Journal of American Statistical Association* 49, 479–87.

Lindley, D. 1962. Discussion of Professor Stein's paper. *Journal of the Royal Statistical Society*, Series B, 24, 285–8.

Mittelhammer, R., Judge, G., Miller, D. and Cardell, S. 2005. Minimum divergence moment based binary response models: estimation and inference. Working Paper No. 998. CUDARE, University of California, Berkeley.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Sclove, S., Morris, C. and Radhakrishman, R. 1972. Non-optimality of preliminary-test estimators for the multinormal mean. *Annals of Mathematical Statistics* 43, 1481–90.

Stein, C. 1955. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley and Los Angeles: University of California Press.

How to cite this article

Judge, George G. "shrinkage-biased estimation in econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000487> doi:10.1057/9780230226203.1522

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sidgwick, Henry (1838 – 1900)

Barton Schultz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Henry Sidgwick is usually regarded as the third greatest classical utilitarian, after Jeremy Bentham and John Stuart Mill, and his masterpiece *The Methods of Ethics* (1874) is a classic of philosophical ethics. But Sidgwick was a many-sided late Victorian intellectual who should also be counted, with Alfred Marshall, as one of the leading figures in the Cambridge School of economics.

Keywords

altruism; British Association; Cambridge School; deductive method; economic biology; economic man; Edgeworth, F.; egoism; ethical socialism; ethics and economics; federalism; Hayek, F. von; hedonism; imperialism; individualism; inductive method; interpersonal utility comparisons; marginal revolution; market socialism; Marshall, A.; Mill, J. S.; Palgrave, R.; *Palgrave's Dictionary of Political Economy* ; Pigou, A.; public goods; Sidgwick, E. (née Balfour); Sidgwick, H.; socialism; utilitarianism; welfare economics

Article

Henry Sidgwick was a Victorian-era philosopher, ethicist, classicist, economist, political and legal theorist, parapsychologist, educational reformer, and literary critic who spent his entire working life at Cambridge University, becoming a central figure in the early Cambridge School of economics. Educated at Rugby and Cambridge, where he studied classics and mathematics and joined the secret discussion society known as the Apostles, he eventually, in 1883, achieved the status of Knightbridge Professor of Moral Philosophy. With his wife, Eleanor Mildred Sidgwick (née Balfour), he helped found both the Society for Psychical Research and Newnham College, Cambridge, one of England's first colleges for women.
Sidgwick published widely, but his major books during his lifetime were *The Methods of Ethics* (first edition 1874), *The Principles of Political Economy* (first edition 1883), and *The Elements of Politics*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(first edition 1891). Best known as an ethical philosopher, he was also very influential in other areas, particularly economics. He worked extensively with the Cambridge and London Charity Organization Societies, and in 1885 he was elected president of the economics and statistics section of the British Association; he also contributed to the first *Dictionary of Political Economy*, edited by R. Palgrave, advised his university on economic policy, and appeared as an expert witness for various government committees on economic policy matters.

The Cambridge School

Sidgwick's importance as an economic thinker has often been underestimated, in part because of his stormy relationship with Alfred Marshall, who is generally regarded as the founder of the Cambridge School (Groenewegen, 1995). Marshall harshly opposed Sidgwick as a ‘University politician’ and criticized his economic work, but Sidgwick clearly played a vital role in shaping both Marshall and their Cambridge context (Backhouse, 2006; Schultz, 2004). He was, as much as Marshall, caught up in the so-called marginalist revolution:

> as Jevons had admirably explained, the variations in the relative market values of different articles express and correspond to variations in the comparative estimates formed by people in general, not of the *total* utilities of the amounts purchased of such articles, but of their *final* utilities; the utilities, that is, of the last portions purchased. (Sidgwick, 1901, p. 82)

But Sidgwick was more involved than Marshall in the methodological debates of the time, seeking to balance both deductive and inductive approaches, and he was also wary of the evolutionary metaphors and talk of ‘economic biology’ to which Marshall was given. Much the better philosopher, Sidgwick's reflective, qualified utilitarianism and analysis of public goods and the role of the state anticipated early 20th-century welfare economics. Indeed, the welfare economics of Marshall's designated successor, A. C. Pigou, was more in line with Sidgwick's views than with Marshall's (O'Donnell, 1979; Backhouse, 2006). As early as 1913, J.S. Nicholson noted that, not only did Pigou ‘apply the same general principle of utility, but the main trend of the argument is the same’ (Nicholson, 1913, p. 420). Baumol (1965) complained that Sidgwick's overall approach and ‘penetrating discussion’ of the ‘Pigouvian problem’ (the possible divergence between private and social benefits and costs) was ‘largely unrecognized’, and quoted from book III of the *Principles*, where Sidgwick argues that ‘there is no general reason’ for supposing that it will always be the case that ‘the individual can always obtain through free exchange adequate remuneration for the services he is capable of rendering to society’ since, among other things, ‘there are some utilities which, from their nature, are practically incapable of being appropriated by those who produce them or who would otherwise be willing to purchase them’ (for example, lighthouses, improvements in climate).
Yet even the accounts of Sidgwick's work that recognize his significance often tend to stress different accomplishments. Blaug (1985, p. 479) suggests that the *Principles* may have been the first work ‘to question the traditional idea that technical change is necessarily capital-using’. Stigler (1982, p. 41), when discussing how Edgeworth, Sidgwick and Marshall gave currency to the work of Cournot and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Dupuit on monopoly and oligopoly, remarked that Sidgwick's *Principles* ‘has two chapters (bk II, ch. IX and X) which are among the best in the history of microeconomics, dealing with the theories of human capital and noncompetitive behavior’. Chapter IX concludes that:

> the possessors of capital, real and personal, as well as persons endowed with rare natural gifts, are likely to have – by reason of their limited numbers – important advantages in the competition that determines relative wages; in consequence of which the remuneration of such persons may – and in England often does – exceed the wages of ordinary labour by an amount considerably larger than is required to compensate them for additional outlay or other sacrifices; such excess tending to increase as the amount of capital owned by any individual increases, but in a ratio not precisely determinable by general considerations. (Sidgwick, 1901, p. 337)

Chapter X investigates how and when self-interested action will lead to combination, and it provocatively argues – beyond Mill and against ‘any economist of repute’ – that in many ordinary cases it is possible for workers to combine and win higher wages without such gain having ‘any manifest tendency to be counterbalanced by future loss’. More generally, and anticipating current analyses of labour markets as involving bargaining processes, Sidgwick argues that, when it comes to wage controversies:

> Economic science cannot profess to determine the normal division of the difference remaining, when from the net produce available for wages and profits in any branch of production we subtract the minimum shares which it is the interest of employers and employed respectively to take rather than abandon the business and seek employment for their labour and capital elsewhere. (1901, p. 355)

Economics and ethics

This fine-grained (but qualitative) analysis, attentive to expert opinion and the line between description and prescription, but yielding conclusions that for the times were quite sceptical or subversive in their implications, is characteristic of Sidgwick's scholarly work in general. Even his work in ethics mostly sought to set aside preaching and polemics, proceeding instead by a painstaking comparative investigation of the rational grounds for the major systematic ethical positions. His celebrated *Methods* sought, with something akin to scientific open-mindedness, ‘to consider simply what conclusions will be rationally reached if we start with certain ethical premises, and with what degree of certainty and precision’ (Sidgwick, 1907, p. viii). Sidgwick concluded that, although commonsense moral rules – do not lie, do not murder, do not break promises – could be largely derived from utilitarianism, there was a ‘dualism of the practical reason’ when it came to egoism (pursue one's own greatest good) and utilitarianism (pursue the greatest good in general), each of which was as ‘rational’ as the other. The dualism of practical reason was in key respects a philosophical representation of collective action problems, and the conception of ultimate good or happiness figuring in both egoism and utilitarianism was a subtle hedonistic account of pleasurable or desirable consciousness that made it clear how difficult

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
yet unavoidable interpersonal comparisons of utility could be. This conception would form the basis for Edgeworth's attempted ‘hedonometry’, or quantification of hedonism, revived in recent years by Kahneman (for example, Edgeworth, 1877; Kahneman, 1999).

Sidgwick himself favoured utilitarianism, though he was haunted by his inability to defend it fully. His utilitarianism nonetheless clearly (and admittedly) informed his economic views, especially when it came to the ‘art’ of political economy, which concerned the normative considerations that came to be called welfare economics. When it came to such normative political theory, Sidgwick largely assumed the utilitarian principle as the normative bottom line (Sidgwick, 1901). Despite their differences, Marshall was willing to give Sidgwick credit for his ‘art’, calling the third book of Sidgwick's *Principles* by ‘far the best thing of its kind in any language’ (in Pigou, 1925, p. 7). He also expressed an admiration for a broader orientation they shared: ‘that we are not at liberty to play chess games, or exercise ourselves upon subtleties that lead nowhere’ (Whitaker, 1996). Neither Sidgwick nor Marshall was enthusiastic about formalization for formalization's sake.

Still, it is not clear precisely what Marshall admired about Sidgwick's welfare economics. Although both stressed the role of education in helping to overcome poverty and economic inequality, Sidgwick arguably went beyond Marshall (and Mill) in setting out the cases of market failure, the limits of laissez-faire, and the limitations of economic analysis in general, both descriptive and normative. Backhouse (2002, p. 271) urges that the truly ‘fundamental part of Sidgwick's argument’ was distinguishing between wealth as ‘as the sum of goods produced, valued at market prices’ and wealth as ‘the sum of individuals’ utilities – what we would now term welfare’. This distinction ‘made it possible, arguably for the first time, to conceive of welfare economics as something distinct from economics in general’, while greatly complicating comparisons of utilities. Analogous arguments are evident in Sidgwick's sceptical account of the possibilities for comparing wealth in cross-cultural or trans-historical contexts. And the notion of unpurchased utilities, not measured by exchange values, allowed that, as Backhouse puts it, ‘if the marginal utility of a particular good were higher for one person than for another, total utility could be raised by redistributing goods to those who valued them most. This would leave wealth at market prices unchanged.’ Such views, conjoined with a belief in declining marginal utility, suggest serious redistributivist possibilities that both Sidgwick and Marshall downplayed:

> Marshall is much more aware of the quantitative side of the problem than is Sidgwick … but no nearer a way to thinking quantitatively about how to achieve the best use of resources. They share both a philosophical viewpoint that inclines them towards egalitarianism and a conservatism that will not risk any interference with incentives, lest output be reduced. (Backhouse, 2006, p. 33)

Economics and politics

Perhaps Sidgwick was overly impressed by the idea that incentives were crucial to production and that complete communism would lead to splendidly equal destitution. Still, he took more gradualist forms of (market) socialism very seriously – so much so that the libertarian Hayek (1960, p. 419) could complain that Sidgwick's *Elements of Politics* scarcely ‘represents what must be regarded as the British liberal tradition and is already strongly tainted with that rationalist utilitarianism which led to socialism’. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Elements and the *Principles* take the laissez-faire principle of individualism – that 'what one sane adult is legally compelled to render to others should be merely the negative service of non-interference, except so far as he has voluntarily undertaken to render positive services' – only as a starting point, from which to run a very long course of qualifications and exceptions: education, child care, poor relief, disease control, public works or goods (the famous lighthouse, pure research, the environment and defence), monopoly, collective bargaining, and others. Sidgwick emphasizes two cases that sharply point up the limitations of economic individualism – the 'humane treatment of lunatics, and the prevention of cruelty to the inferior animals' – because such restrictions hardly aim at securing the freedom of the lunatics or the animals, but are 'a one-sided restraint of the freedom of action of men with a view to the greatest happiness of the aggregate of sentient beings' (Sidgwick, 1919, p. 141). These are only the most conspicuous of the many difficulties with a principle that betrays a naive faith in 'the psychological proposition that every one can best take care of his own interest' and the 'sociological proposition that the common welfare is best attained by each pursuing exclusively his own welfare and that of his family in a thoroughly alert and intelligent manner'.

Cautious as he was, Sidgwick (1903) was ultimately persuaded that the growth of federalism and large-scale state organizations was likely to continue, though he doubted that the social sciences were anywhere near to discovering actual laws of historical development. Moreover, even if he was cautious about economic or political socialism, he was relatively enthusiastic about ethical socialism, that is, about the possibility of humanity growing more altruistic and compassionate, regarding their labour as their contribution to the common good. He was under no illusions whatsoever, not only about the market failing to reflect claims of desert or merit, but also about the limitations of that abstraction, 'economic man', since historical and cultural or national context could dramatically alter the possibilities for moving beyond economic individualism. Perhaps the most disturbing and problematic aspect of his economic and political work was the way in which it lent itself to the racist and imperialistic tendencies of the British Empire (Schultz, 2004). Although he was more sceptical about claims of inherent racial differences than many of his contemporaries, he too often countenanced the possibility of such differences, accepting stereotypes about the varying fitness for physical versus mental labour of different peoples (Schultz and Varouxakis, 2005). And although on the economic side he tended to follow Turgot in doubting the material benefits of colonies, he allowed that there were other forms of remuneration involved:

> there are sentimental satisfactions, derived from justifiable conquests, which must be taken into account, though they are very difficult to weigh against the material sacrifices and risks. Such are the justifiable pride which the cultivated members of a civilised community feel in the beneficent exercise of dominion, and in the performance by their nation of the noble task of spreading the highest kind of civilization. (1919, p. 313)

It was in this highly Eurocentric manner that Sidgwick embraced the global leadership of the 'civilised' nations in (supposedly) spreading international law and morality and limiting the possibility of war (Sidgwick, 1898).

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Edgeworth, Francis Ysidro
- happiness, economics of
- Marshall, Alfred
- Mill, John Stuart
- Pigou, Arthur Cecil
- utilitarianism and economic theory
- welfare economics

Selected works

1898. *Practical Ethics*. London: Swan Sonnenschein.

1901. *The Principles of Political Economy*, 3rd edn. London: Macmillan.

1903. *The Development of European Polity*. London: Macmillan.

1904. *Miscellaneous Essays and Addresses*. London: Macmillan.

1907. *The Methods of Ethics*, 7th edn. London: Macmillan.

1919. *The Elements of Politics*, 4th edn. London: Macmillan.

1997, 1999. *The Complete Works and Select Correspondence of Henry Sidgwick*, gen. ed. B. Schultz. Charlottesville, VA: InteLex Corp.

Bibliography

Backhouse, R. 2002. *The Ordinary Business of Life*. Princeton: Princeton University Press.

Backhouse, R. 2006. Sidgwick, Marshall, and the Cambridge School of economics. *History of Political Economy* 38, 15–44.

Baumol 1965. *Welfare Economics and the Theory of the State*, 2nd edn. Cambridge, MA: Harvard University Press.

Blaug, M. 1985. *Economic Theory in Retrospect*. Cambridge: Cambridge University Press.

Edgeworth, F. 1877. *Old and New Methods of Ethics*. Oxford and London: James Parker and Co.

Groenewegen, P. 1995. *A Soaring Eagle: Alfred Marshall, 1842–1924*. Aldershot, UK and Brookfield,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

US: Edward Elgar.

Hayek, F. 1960. *The Constitution of Liberty*. Chicago: Regnery.

Kahneman, D. 1999. Objective happiness. In *Well-Being: The Foundations of Hedonic Psychology*, ed. D. Kahneman, E. Diener and N. Schwarz. New York: Russell Sage Foundation.

Nicholson, J. 1913. The vagaries of recent political economy. *Quarterly Review* 219.

O'Donnell, M. 1979. Pigou: an extension of Sidgwickian thought. *History of Political Economy* 11, 588–605.

Pigou, A., ed. 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Schultz, B. 2004. *Henry Sidgwick, Eye of the Universe*. New York: Cambridge University Press.

Schultz, B. and Varouxakis, G., eds. 2005. *Utilitarianism and Empire*. Lanham, MD: Lexington Books.

Stigler, G. 1982. *The Economist as Preacher*. Chicago: University of Chicago Press.

Whitaker, J. 1996. *The Correspondence of Alfred Marshall, Economist*, 3 vols. New York: Cambridge University Press.

How to cite this article

Schultz, Barton. "Sidgwick, Henry (1838–1900)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000125> doi:10.1057/9780230226203.1523

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sidrauski, Miguel (1939 – 1968)

Stanley Fischer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adaptive expectations; Sidrauski, M.; steady state capital intensity; superneutrality; Tobin's *q* ; two-sector models

Article

Sidrauski was born and educated in Buenos Aires. He entered the University of Chicago Ph.D. programme in 1963, completed his dissertation in 1966, and accepted an assistant professor appointment at MIT. He died of cancer in September 1968, leaving his wife, Martha, and two-month old daughter, Carmela.

Sidrauski is best known in economics for his article ‘Rational Choice and Patterns of Growth in a Monetary Economy’ (1967a), based on his dissertation written under the supervision of Hirofumi Uzawa and Milton Friedman. The model is one of an economy with a representative intertemporally maximizing household, which derives instantaneous utility from both consumption and the holding of real balances. The thesis contains a careful discussion of the device of putting money in the utility function. The household can hold capital as well as money. Sidrauski derives necessary conditions for a maximum, and then studies the dynamics and steady states of inflation and capital accumulation in the model.

The key result is that steady state capital intensity is invariant to the rate of monetary expansion, and thus that money is superneutral between steady states. Sidrauski indicates in his thesis, which extends the paper in several directions, that the superneutrality result changes if money is given a productive role in the economy. In the dynamic analysis he assumes that expectations of inflation are adaptive.

In a non-maximizing money-and-growth model (1967b) Sidrauski confirms the Tobin result that an increase in the growth rate of money increases steady state capital intensity. This is based on the effects of an increase in the growth rate of money in reducing consumption. Sidrauski shows also in this paper that, with adaptive expectations, an increase in the growth rate of money first causes capital accumulation to fall and only later to increase, taking capital intensity above its initial level.

He published four other articles, including one on exchange rate determination, and a book, *Monetary*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and Fiscal Policy in a Growing Economy. The book, written jointly with Duncan Foley and published posthumously in 1970, develops a three-asset (money, bonds and capital) two-sector Tobinesque growth model. The two sector structure gives a key role in the investment process to the relative price of capital, p_k, which is Tobin's q. The authors succeed in making the model answer questions about the effects on growth and inflation of policy changes such as fiscal expansion, increases in the growth rate of all outside assets, and open market operations. The model has not been widely used despite its usefulness and versatility; this may be because the full employment assumption makes it unattractive for use as a cyclical model, and the absence of explicit maximizing assumptions makes it unattractive to many who study long-term growth.

In his two years at MIT Sidrauski established himself as an excellent teacher and adviser, and as an economist of outstanding promise. Milton Friedman's eulogy (1969) speaks not only of his technical skill and promise, but also of his personal warmth and generosity. It concludes:

> The death of this young man is a grievous loss to our profession and to the world. Here was a man who would have pushed out the frontiers of our subject, would have changed and added to economic analysis, would have enlightened and informed generations of students – struck down at the very beginning of his career, full of promise but as yet almost bereft of fulfillment.

Selected works

1967a. Rational choice and patterns of growth in a monetary economy. *American Economic Review, Papers and Proceedings* 57, 534–44.

1967b. Inflation and economic growth. *Journal of Political Economy* 75, 796–810.

1970. (With D.K. Foley.) *Monetary and Fiscal Policy in a Growing Economy*. London: Macmillan.

Bibliography

Friedman, M. 1969. Miguel Sidrauski. *Journal of Money, Credit and Banking* 1, 129–30.

How to cite this article

Fischer, Stanley. "Sidrauski, Miguel (1939–1968)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000126> doi:10.1057/9780230226203.1524

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sieve extremum estimation

Xiaohong Chen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Semi-nonparametric models are more flexible and robust than parametric models, but are more complex due to the presence of infinite dimensional unknown parameters. This article describes the method of sieve extremum estimation of semi-nonparametric models, which is a general method of optimizing an empirical criterion function over a sequence of approximating parameter spaces (that is, sieves). Widely used sieve spaces and criterion functions are presented as examples, including the sieve M-estimation, series estimation, and sieve minimum distance estimation as special cases. Existing results are cited on asymptotic properties and applications of the method.

Keywords

ARCH models; artificial neural networks; chaos; Engel curve; generalized method of moments; polynomials; maximum likelihood; measurement error models; nonparametric models; semi-nonparametric models; series estimation; sieve extremum estimation; sieve minimum distance; simultaneous equations; splines; spurious regressions; treatment effect; wavelets

Article

1 Introduction

In econometrics literature, a model is called *parametric* if all of its parameters belong to finite-dimensional parameter spaces, and a model is *semiparametric* if its parameters of interests belong to finite-dimensional spaces but its nuisance parameters are in infinite-dimensional spaces. For example, a parameterized joint distribution model (such as logit or probit) is a parametric one and can be estimated by the method of maximum likelihood (ML). A model that only parameterizes some moment restrictions is a semiparametric one and can be estimated by the generalized method of moments (GMM). Since economics problems are complicated, researchers often find parametric likelihood models too restrictive and sensitive to deviations from the parametric specification. Moment-based semiparametric models are less restrictive yet still subject to misspecification of parametric moments.
Due to growing availability of large economic data-sets and computational advances, nonparametric and semi-nonparametric models have become increasingly popular in both theoretical and applied econometrics. A model is *nonparametric* if all of its parameters are in infinite-dimensional parameter spaces, and a model is *semi-nonparametric* if it contains both finite-dimensional and infinite-dimensional unknown parameters of interests. For example, estimation of a conditional mean function E[$Y|X$, W] without specifying its functional form is a nonparametric problem. If $E[Y|X, W]$ is specified as a partially linear form, $\beta' X + h(W)$, with β being unknown finite-dimensional parameter and $h()$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
being unknown function, then it becomes a semi-nonparametric problem. Nonparametric and semi-nonparametric models are more flexible and robust. However, they involve infinite-dimensional parameter spaces; hence it can be computationally difficult to estimate such models using finite data sets. Moreover, even if one could solve the problem of optimizing a sample criterion over an infinite-dimensional parameter space that may not be compact, the resulting estimator may have undesirable large sample properties such as inconsistency and/or a very slow rate of convergence. To tackle the difficulties encountered in semi-nonparametric problems, the *method of sieves* (Grenander, 1981) optimizes an empirical criterion over a sequence of approximating parameter spaces (that is, *sieves*); the sieves are less complex, but their complexity grows with the sample size so as to be dense in the original space. The resulting sieve estimator is consistent under very mild regularity conditions, and can generally reach optimal rate of convergence by balancing the bias part (which diminishes as sieve complexity grows), and the standard derivation part (which grows with sieve complexity). Most commonly used sieves in economics are finite dimensional (compact) approximating parameter spaces; when such sieves are used to estimate a semi-nonparametric model, the computation is as easy as to estimate a parametric model. The sieve method is particularly convenient when unknown functions enter the criterion function (or moment condition) nonlinearly, and/or when semi-nonparametric models contain complicated endogeneity and latent heterogeneity. It can easily incorporate prior information and constraints, often derived from economic theory, such as monotonicity, convexity, additivity, multiplicity, exclusion and non-negativity. It can simultaneously estimate the parametric and nonparametric parts in semi-nonparametric models, typically with optimal convergence rates for both parts.

The method of sieves consists of two key ingredients: a criterion function and sieve parameter spaces. Both the criterion functions and the sieve spaces can be very flexible, and we shall present some examples in the next two sections.

2 Examples of sieve spaces

The infinite-dimensional unknown parameter in a semi-nonparametric model can often be viewed as a member of some function space with certain regularities (for example, having bounded derivatives, monotone, concave). Thus, many deterministic approximation results developed in mathematics can be used to suggest sieves that provide good and computable approximations to an unknown function. Here we present some commonly used sieves in economics applications. Additional ones can be found in Judd (1998), DeVore and Lorentz (1993) and Chen (2006).

2.1 Finite-dimensional linear sieves with bounded support

A sieve is called a ‘(finite-dimensional) linear sieve’ if it is a linear span of finitely many known basis functions. Linear sieves, including power series, Fourier series, splines and wavelets, form a large class of sieves that have been widely used in econometrics and statistics. We now provide some commonly used linear sieves for univariate functions with support χ =[0,1].

Polynomials. Pol(J_n) is the space of polynomials on [0,1] of degree J_n or less:

$$\mathrm{Pol}(J_n) = \left\{ \sum_{k=0}^{J_n} a_k x^k,\ x \in [0,\ 1] : a_k \in \Re \right\}.$$

Trigonometrics. TriPol(J_n) is the space of trigonometric polynomials on [0,1] of degree J_n or less:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\mathrm{TriPol}(J_n) = \left\{ a_0 + \sum_{k=1}^{J_n} [a_k \cos(2k\pi x) + b_k \sin(2k\pi x)],\ x \in [0, 1] : a_k, b_k \in \Re \right\}.$$

CosPol(J_n) is the space of cosine polynomials on [0,1] of degree J_n or less:

$$\mathrm{CosPol}(J_n) = \left\{ a_0 + \sum_{k=1}^{J_n} a_k \cos(k\pi x),\ x \in [0, 1] : a_k \in \Re \right\}.$$

SinPol(J_n) is the space of sine polynomials on [0,1] of degree J_n or less:

$$\mathrm{SinPol}(J_n) = \left\{ \sum_{k=1}^{J_n} a_k \sin(k\pi x),\ x \in [0, 1] : a_k \in \Re \right\}.$$

We note that the classical trigonometric sieve, TriPol(J_n), is well suited for approximating periodic functions on [0,1], while the cosine sieve, CosPol(J_n), is well suited for approximating aperiodic functions on [0,1] and the sine sieve, SinPol(J_n), can approximate functions vanishing at the boundary points (that is, when $h(0)=h(1)=0$).

Splines. Let J_n be a positive integer, and let $t_0, t_1, \ldots, t_{J_n}, t_{J_n+1}$ be real numbers with $0 = t_0 < t_1 < \ldots < t_{J_n} < t_{J_n+1} = 1$. Partition [0,1] into J_n+1 subintervals $I_j=[t_j, t_{j+1})$, $j=0,\ldots,J_n-1$, and $I_{J_n} = [t_{J_n}, t_{J_n+1}]$. We assume that the knots $t_1, \ldots, t_{J_n}$ have bounded mesh ratio:

$$\frac{\max_{0 \le j \le J_n}(t_{j+1} - t_j)}{\min_{0 \le j \le J_n}(t_{j+1} - t_j)} \le c \text{ for some constant } c > 0. \tag{2.1}$$

Let $r \geq 1$ be an integer. A function on [0,1] is a *spline of order r*, equivalently, *of degree* $m \equiv r-1$, *with knots* $t_1, \ldots, t_{J_n}$ if the following hold: (i) it is a polynomial of degree m or less on each interval I_j, $j=0,\ldots,J_n$; and (ii) (for $m \geq 1$) it is ($m-1$)-times continuously differentiable on [0,1]. Such spline functions constitute a linear space of dimension J_n+r. For detailed discussions of univariate splines, see de Boor (1978) and Schumaker (1981). For a fixed integer $r \geq 1$, we let Spl (r, J_n) denote the space of splines of order r (or of degree $m \equiv r-1$) with J_n knots satisfying (2.1). Since

$$\mathrm{Spl}(r, J_n) = \left\{ \sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{J_n} b_j [\max\{x - t_j, 0\}]^{r-1},\ x \in [0, 1] : a_k, b_j \in \Re \right\},$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

we also call Spl(r, J_n) the polynomial spline sieve of degree $m \equiv r - 1$.

2.2 Finite-dimensional linear sieves with unbounded support

In semi-nonparametric econometric applications, sometimes the parameters of interest are functions with unbounded supports. Here we present three finite-dimensional linear sieves that can approximate functions with unbounded supports well. In the following we let $L_2(\chi, \omega)$ denote the space of real-valued functions h such that $\int_\chi |h(x)|^2 \omega(x)\,dx < \infty$ for a smooth weight function $\omega : \chi \mapsto (0, \infty)$.

Orthogonal wavelets. Let $m \geq 0$ be an integer. A real-valued function ϕ is called a 'father wavelet' of degree m if it satisfies the following: (i) $\int_{\mathcal{R}} \varphi(x)\,dx = 1$; (ii) ϕ and all its derivatives up to order m decrease rapidly as $|x| \to \infty$; (iii) $\{\varphi(x-k) : k \in \mathbb{Z}\}$ forms a Riesz basis for a closed subspace of $L_2(\mathcal{R}, leb)$. A real-valued function ψ is called a 'mother wavelet' of degree m if it satisfies the following: (i) $\int_{\mathcal{R}} x^k \psi(x)\,dx = 0$ for $0 \leq k \leq m$; (ii) ψ and all its derivatives up to order m decrease rapidly as $|x| \to \infty$; (iii) $\left\{2^{j/2}\psi(2^j x - k) : j, k \in \mathbb{Z}\right\}$ forms a Riesz basis of $L_2(\mathcal{R}, leb)$.

Given an integer $m \geq 0$, there exist a father wavelet ϕ of degree m and a mother wavelet ψ of degree m, both compactly supported, such that for any integer $j_0 \geq 0$, any function g in $L_2(\mathcal{R}, leb)$ has the following wavelet m–regular multiresolution expansion:

$$g(x) = \sum_{k=-\infty}^{\infty} a_{j_0 k}\varphi_{j_0 k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=-\infty}^{\infty} b_{jk}\psi_{jk}(x),\ x \in \mathcal{R},$$

where

$$a_{jk} = \int_{\mathcal{R}} g(x)\varphi_{jk}(x)\,dx,\ \varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k),\ x \in \mathcal{R}, b_{jk} = \int_{\mathcal{R}} g(x)\psi_{jk}(x)\,dx,\ \psi_{jk}(x) = 2^{j/2}\psi(2^j x - k),\ x \in \mathcal{R},$$

and $\left\{\varphi_{j_0 k},\ k \in \mathbb{Z};\ \psi_{jk},\ j \geq j_0,\ k \in \mathbb{Z}\right\}$ is an orthonormal basis of $L_2(\mathcal{R}, leb)$; see Meyer (1992, theorem 3.3). We consider the finite-dimensional linear space spanned by this wavelet basis. For an integer $J_n > j_0$, set

$$\mathrm{Wav}(m, 2^{J_n}) = \left\{\sum_{k=0}^{2^{j_0}-1}\alpha_{j_0 k}\varphi_{j_0 k}(x) + \sum_{j=j_0}^{J_n-1}\sum_{k=0}^{2^j-1}\beta_{jk}\psi_{jk}(x),\ x \in \mathcal{R} : \alpha_{j_0 k}, \beta_{jk} \in \mathcal{R}\right\}$$

or, equivalently,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\mathrm{Wav}(m, 2^{J_n}) = \left\{ \sum_{k=0}^{2^{J_n}-1} \alpha_k \varphi_{J_n k}(x),\ x \in \mathcal{R}: \alpha_k \in \mathcal{R} \right\}.$$

Hermite polynomials. Hermite polynomial series $\{H_k: k=1,2,\ldots\ldots\}$ is an orthonormal basis of $L_2(\mathcal{R}, \omega)$ with $\omega(x) = \exp\left\{-x^2\right\}$. It can be obtained by applying the Gram–Schmidt procedure to the polynomial series $\{x^{k-1}: k=1,2,\ldots\ldots\}$ under the inner product $\langle f, g\rangle_\omega = \int_{\mathcal{R}} f(x) g(x) \exp\left\{-x^2\right\} dx$. Let HPol($J_n$) denote the space of Hermite polynomials on $\mathcal{R}$ of degree J_n or less:

$$\mathrm{HPol}(J_n) = \left\{ \sum_{k=1}^{J_n+1} a_k H_k(x) \exp\left\{-\frac{x^2}{2}\right\},\ x \in \mathcal{R}: a_k \in \mathcal{R} \right\}.$$

Then any function in $L_2(\mathcal{R}, leb)$ can be approximated by the Pol(J_n) sieve as $J_n \to \infty$.
Laguerre polynomials. The Laguerre polynomial series $\{L_k: k=1,2,\ldots\ldots\}$ is an orthonormal basis of $L_2([0,\infty), \omega)$ with $\omega(x)=\exp\{-x\}$. It can be obtained by applying the Gram-Schmidt procedure to the polynomial series $\{x^{k-1}: k=1,2,\ldots\ldots\}$ under the inner product $\langle f, g\rangle_\omega = \int_0^\infty f(x) g(x) \exp\{-x\} dx$. Let LPol($J_n$) denote the space of Laguerre polynomials on $[0,\infty)$ of degree J_n or less:

$$\mathrm{LPol}(J_n) = \left\{ \sum_{k=1}^{J_n+1} a_k L_k(x) \exp\left\{-\frac{x}{2}\right\},\ x \in [0, \infty): a_k \in \mathcal{R} \right\}.$$

Then any function in $L_2([0,\infty), leb)$ can be approximated by the LPol(J_n) sieve as $J_n \to \infty$.

2.3 Finite-dimensional nonlinear sieves

A popular class of nonlinear sieves in econometrics is single hidden layer feedforward artificial neural networks (ANN) (see for example, Barron, 1993; Hornik et al., 1994; Chen and White, 1999). A typical ANN sieve is the one with a sigmoid activation function:

$$\mathrm{sANN}(k_n) = \left\{ \sum_{j=1}^{k_n} \alpha_j S(\gamma_j' x + \gamma_{0,j}): \gamma_j \in \mathcal{R}^d,\ \alpha_j, \gamma_{0,j} \in \mathcal{R} \right\},$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where S: $R \to R$ is a sigmoid activation function, that is, a bounded non-decreasing function such that $\lim_{u\to-\infty} S(u)=0$ and $\lim_{u\to\infty} S(u)=1$. Some popular sigmoid activation functions include

- heaviside $S(u)=1\{u \geq 0\}$
- logistic $S(u)=1/(1+\exp\{-u\})$
- hyperbolic tangent $S(u)=(\exp\{u\}-\exp\{-u\})/(\exp\{u\}+\exp\{-u\})$
- Gaussian sigmoid $S(u) = (2\pi)^{-1/2}\int_{-\infty}^{u}\exp(-y^2/2)dy$

Additional examples of nonlinear sieves include spline sieves with data-driven choices of knot locations (or free-knot splines), and wavelet sieves with thresholding (Donoho et al., 1995). Nonlinear sieves are more flexible and may enjoy better approximation properties than linear sieves; see for example, Chen and Shen (1998) for the comparison of linear vs. nonlinear sieves.

2.4 Shape-preserving sieves

There are many sieves that can preserve the shape, such as non-negativity, monotonicity and convexity, of the unknown function to be approximated. Here we mention one of such shape-preserving sieves.
Cardinal B-spline wavelets. The cardinal B-spline of order $r \geq 1$ is given by

$$B_r(x) = \frac{1}{(r-1)!}\sum_{j=0}^{r}(-1)^j\binom{r}{j}[\max(0, x-j)]^{r-1}, \tag{2.2}$$

which has support $[0, r]$, is symmetric at $r/2$ and is a piecewise polynomial of highest degree $r-1$. It satisfies $B_r(x) \geq 0$, $\sum_{k=-\infty}^{+\infty} B_r(x-k) = 1$ for all $x \in \Re$, which is crucial to preserve the shape of the unknown function to be approximated. See Chui (1992, ch. 4) for a recursive construction of cardinal B-splines and their properties. Denote

$$\mathrm{SplWav}(r-1, 2^{Jn}) = \left\{\sum_{k=-\infty}^{\infty}\alpha_k 2^{Jn/2}B_r(2^{Jn}x-k),\ x \in \Re: \alpha_k \in \Re\right\}.$$

Any non-decreasing continuous function on $\Re$ can be approximated well by the $\mathrm{SplWav}(r-1, 2^{Jn})$ sieve with non-decreasing sequence $\{\alpha_k\}$ (that is, $\alpha_k \leq \alpha_{k+1}$). See Anastassiou and Yu (1992) on shape-preserving wavelet sieves.

2.5 Infinite-dimensional (nonlinear) sieves

Most commonly used sieve spaces are finite-dimensional truncated series such as those listed above. However, the general theory on sieve extremum estimation can also allow for infinite-dimensional sieve spaces. For example, any function θ that belongs to a Hölder space $\Theta = \Lambda^p([0,1])$ with smoothness $p>1/2$ can be expressed as an infinite Fourier series $\theta(x) = \sum_{k=1}^{\infty}[a_k\cos(kx) + b_k\sin(kx)]$, and its derivative with fractional power $\gamma \in (0, p]$ can also be defined

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in terms of Fourier series:

$$\theta^{(\gamma)}(x) = \sum_{k=1}^{\infty} k^{\gamma}\left[\left(a_k \cos\frac{\pi\gamma}{2} + b_k \sin\frac{\pi\gamma}{2}\right)\cos(kx) + \left(b_k \cos\frac{\pi\gamma}{2} - a_k \sin\frac{\pi\gamma}{2}\right)\sin(kx)\right].$$

More generally, if the parameter space Θ is a typical function space such as a Hölder, Sobolev or Besov space, then any function $\theta \in \Theta$ and its fractional derivatives can be expressed as infinite series of some known Riesz basis $\{B_k(\cdot)\}_{k=1}^{\infty}$ such as splines and wavelets (see for example, Meyer, 1992). An infinite-dimensional sieve space could then take the form:

$$\Theta_n = \{\theta \in \Theta : \theta(\cdot) = \sum_{k=1}^{\infty} a_k B_k(\cdot),\ \mathrm{pen}(\theta) \le b_n\} \text{with } b_n \to \infty \text{ slowly,} \tag{2.3}$$

where pen(θ) is a smoothness (or roughness) penalty term, such as $\mathrm{pen}(\theta) = (\int_x |\theta^{(p)}(x)|^q dx)^{1/q}$, with p>1/2 the smoothness of the function θ, and some $q \geq 1$. For example, Wahba (1990) considered smoothing spline sieve with q=2 to approximate conditional mean function; Koenker, Ng and Portnoy (1994) considered smoothing spline sieve with q=1 to approximate conditional quantile function. See Shen (1997) and van de Geer (2000) for more expressions of pen(θ).

3 Sieve extremum estimation

Let Θ be an infinite dimensional parameter space endowed with a (pseudo-) metric d. A typical semi-nonparametric econometric model specifies that there is a population criterion function $Q: \Theta \to \Re$, which is uniquely maximized at a (pseudo-) true parameter $\theta_o \in \Theta$. The choice of $Q(\cdot)$ and the existence of θ_o are suggested by the identification of an econometric model. The (pseudo-) true parameter $\theta_o \in \Theta$ is unknown but is related to a joint probability measure P_o $(z_1,\ldots,z_n)$, from which a sample of size n observations $\{Z_t\}_{t=1}^{n}$, $Z_t \in \Re^{d_Z}$, $1 \leq d_z < \infty$, is available. Let $\hat{Q}_n : \Theta \to \Re$ be an empirical criterion, which is a measurable function of the data $\{Z_t\}_{t=1}^{n}$ for all $\theta \in \Theta$, and converges to Q in some sense as the sample size $n \to \infty$.

When Θ is infinite dimensional and possibly not compact with respect to the (pseudo-) metric d, maximizing $\hat{Q}_n$ over Θ may not be well-defined; or even if a maximizer $\mathrm{argsup}_{\theta \in \Theta} \hat{Q}_n(\theta)$ exists, it is generally difficult to compute, and may have undesirable large sample properties such as inconsistency and/or a very slow rate of convergence.

The method of sieves provides one general approach to resolve the difficulties associated with maximizing $\hat{Q}_n$ over an infinite dimensional space Θ by maximizing $\hat{Q}_n$ over a sequence of approximating spaces Θ_n, called *sieves* by Grenander (1981), which are less complex but are dense in Θ. Popular sieves are typically compact, non-decreasing $(\Theta_n \subseteq \Theta_{n+1} \subseteq \ldots \subseteq \Theta)$ and are such that for any $\theta \in \Theta$ there exists an element $\pi_n \theta$ in Θ_n satisfying $d(\theta, \pi_n \theta) \to 0$ as $n \to \infty$, where the notation π_n can be regarded as a projection mapping from Θ to Θ_n.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

An *approximate sieve extremum estimate*, denoted by $\hat{\theta}_n$, is defined as an approximate maximizer of $\hat{Q}_n(\theta)$ over the sieve space Θ_n, that is,

$$\hat{Q}_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} \hat{Q}_n(\theta) - O_P(\eta_n), \text{ with } \eta_n \to 0 \text{ as } n \to \infty. \tag{3.1}$$

When η_n=0, we call $\hat{\theta}_n$ in (3.1) the *exact* sieve extremum estimate. The sieve extremum estimation method clearly includes the standard extremum estimation method by setting $\Theta_n = \Theta$ for all *n*. Following White and Wooldridge (1991, theorem 2.2), one can show that $\hat{\theta}_n$ in (3.1) is well defined and measurable under mild sufficient conditions: (i) $\hat{Q}_n(\theta)$ is a measurable function of the data $\{Z_t\}_{t=1}^n$ for all $\theta \in \Theta_n$; (ii) for any data $\{Z_t\}_{t=1}^n$, $\hat{Q}_n(\theta)$ is upper semicontinuous on Θ_n under the metric $d(\cdot,\cdot)$; and (iii) the sieve space Θ_n is compact under the metric $d(\cdot,\cdot)$.

For a semi-nonparametric model, $\theta_o \in \Theta$ can be decomposed into two parts $\theta_o = (\beta_o', h_o')' \in B \times \mathcal{H}$, where *B* denotes a finite dimensional compact parameter space, and $\mathcal{H}$ an infinite dimensional parameter space. In this case, a natural sieve space will be $\Theta_n = B \times \mathcal{H}_n$ with $\mathcal{H}_n$ being a sieve for $\mathcal{H}$, and the resulting estimate $\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n)$ in (3.1) will sometimes be called a simultaneous (or joint) sieve extremum estimate. For a semi-nonparametric model, we can also estimate the parameters of interest (β_o,h_o) by the *approximate profile sieve extremum estimation* that consists of two steps:

- *Step 1*: for an arbitrarily fixed value $\beta \in B$, compute $\hat{Q}_n(\beta, \tilde{h}(\beta)) \geq \sup_{h \in \mathcal{H}_n} \hat{Q}_n(\beta, h) - O_P(\eta_n)$ with η_n=*o* (1);
- *Step 2*: estimate β_o by $\hat{\beta}_n$ solving $\hat{Q}_n(\hat{\beta}, \tilde{h}(\hat{\beta})) \geq \max_{\beta \in B} \hat{Q}_n(\beta, \tilde{h}(\beta)) - O_P(\eta_n)$, and then estimate h_o by $\hat{h}_n = \tilde{h}(\hat{\beta}_n)$.

3.1 Sieve M-estimation

When $\hat{Q}_n(\theta)$ can be expressed as a sample average form $\hat{Q}_n(\theta) = \frac{1}{n}\sum_{t=1}^n l(\theta, Z_t)$, with $l: \Theta \times \mathcal{R}^{d_Z} \to \mathcal{R}$ being the criterion based on a single observation, we also call the $\hat{\theta}_n$ solving (3.1) as an approximate *sieve maximum-likelihood-like (M-)* estimate. This includes sieve maximum likelihood (ML), sieve least squares (LS), sieve generalized least squares (GLS) and sieve quantile regression as special cases.

Example 3.1: (*single spell duration models with unobserved heterogeneity*): Let $G(\tau|\beta, u, x)$ be a parametric structural distribution function of duration *T* conditional on a scalar of unobserved heterogeneity *U*=*u* and a vector of observed heterogeneity *X*=*x*.

The distribution of observed duration given *X*=*x* is

$$F(\tau|\beta, h, x) = \int G(\tau|\beta, u, x)\, dh(u),$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where the unobserved heterogeneity U is modelled as a random factor with unknown distribution function $h(\cdot)$. Let $\{T_i, X_i\}_{i=1}^n$ be an i.i.d. sample of observations. Heckman and Singer (1984) propose sieve ML estimation of this semi-nonparametric single spell duration model with unobserved heterogeneity. Denote $\theta_o = (\beta_o', h_o)' \in B \times \mathcal{H}$ as the unknown true parameters, where B is a compact subset in $\mathcal{R}^d \beta$ and $\mathcal{H}$ is the space of distribution functions. Let $\mathcal{H}_n$ denote a sieve for $\mathcal{H}$ such as the first-order spline sieve basis used in Heckman and Singer (1984). Then the sieve ML estimate is given by

$$\hat{\theta}_n = (\hat{\beta}_n, \hat{h}_n) = \arg \max_{(\beta, h) \in B \times \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \log \left\{ \int g(T_i | \beta, u, X_i) dh(u) \right\}. \tag{3.2}$$

Series estimation is a special case of sieve M-estimation with *concave* criterion functions $\hat{Q}_n(\theta) = \frac{1}{n} \Sigma_{t=1}^n l(\theta, Z_t)$ and *finite-dimensional linear* sieve spaces Θ_n (that is, the sieves Θ_n are linear spans of finitely many known basis functions).

Example 3.2: (*multivariate LS regression*)*:* We consider the series estimation of an unknown multivariate conditional mean function $\theta_o(\cdot) = h_o(\cdot) = E(Y|X = \cdot)$. Here $Z=(Y, X)$, Y is a scalar, X has support X that is a bounded subset of $\mathcal{R}^d$, $d \geq 1$. Suppose $h_o \in \Theta$, where Θ is a linear subspace of the space of functions h with $E[h(X)^2] < \infty$. Let $l(h, Z) = -[Y - h(X)]^2$ and $Q(\theta) = -E\{[Y - h(X)]^2\}$; then both are concave in h and Q is strictly concave in $h \in \Theta$.

Let $\{p_j(X), j=1,2,\ldots\}$ denote a sequence of known basis functions that can approximate any real-valued square integrable functions of X well; see the previous section for specific examples of such basis functions. Then

$$\Theta_n = \mathcal{H}_n = \left\{ h: \mathcal{X} \to \mathcal{R}, h(x) = \Sigma_{j=1}^{k_n} a_j p_j(x) : a_1, \ldots, a_{k_n} \in \mathcal{R} \right\}, \tag{3.3}$$

with $\dim(\Theta_n) = k_n \to \infty$ slowly as $n \to \infty$, is a finite-dimensional linear sieve for Θ, and $\hat{h} = \operatorname{argmax}_{h \in \mathcal{H}_n} \frac{-1}{n} \Sigma_{t=1}^n [Y_t - h(X_t)]^2$ is a series estimator of the conditional mean $h_o(\cdot) = E(Y|X = \cdot)$. Moreover, this series estimator $\hat{h}$ has a simple closed-form expression:

$$\hat{h}(x) = p^{k_n}(x)'(P'P)^- \sum_{i=1}^n p^{k_n}(X_i) Y_i, \; x \in \mathcal{X}, \tag{3.4}$$

with $p^{k_n}(X) = (p_1(X), \ldots, p_{k_n}(X))'$, $P = (p^{k_n}(X_1), \ldots, p^{k_n}(X_n))'$ and $(P'P)^-$ the Moore–Penrose generalized inverse. The estimator $\hat{h}$ given in (3.4) is called a *series LS estimator*.

Many popular semi-nonparametric regression models, such as the partially linear regression of Engle et al. (1986) and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Robinson (1988), the additive regression model of Stone (1985) and Andrews and Whang (1990), can be easily estimated via the series LS estimation.

3.2 Sieve MD estimation of semi-nonparametric conditional moment models

Many economic models imply semi-nonparametric conditional moment restrictions of the form

$$E[\rho(Z_t;\ \theta_o)|X_t] = 0,\ \theta_o \equiv (\beta_o',\ h_o')', \tag{3.5}$$

where $\rho(\cdot;\cdot)$ is a column vector of residual functions whose functional forms are known up to unknown parameters, $\theta \equiv (\beta',\ h')'$, and $\{Z_t' = (Y_t',\ X_t')\}_{t=1}^{n}$ is the data where Y_t is a vector of endogenous variables and X_t is a vector of conditioning variables. Here $E[\rho(Z_t,\ \theta)|X_t]$ denotes the conditional expectation of $\rho(Z_t,\ \theta)$ given X_t, and the true conditional distribution of Y_t given X_t is unspecified (and is treated as a nuisance function). The parameters of interest $\theta_o \equiv (\beta_o',\ h_o')'$ contain a vector of finite dimensional unknown parameters β_o and a vector of infinite dimensional unknown functions $h_o(\cdot) = (h_{o1}(\cdot),\ \ldots,\ h_{oq}(\cdot))'$, where the arguments of $h_{oj}(\cdot)$ could depend on Y, X, known index function $\delta_j(Z,\ \beta_o)$ up to unknown β_o, other unknown function $h_{ok}(\cdot)$ for $k \neq j$, or could also depend on unobserved random variables. This class of models (3.5) includes many semi-nonparametric models with endogeneity and/or latent heterogeneity as special cases. A leading, yet difficult example is the purely nonparametric instrumental variables (IV) regression $E[Y_{1i} - h_o(Y_{2i})|X_i] = 0$ studied by Newey and Powell (2003), Darolles, Florens and Renault (2006), Blundell, Chen and Kristensen (2007), and Hall and Horowitz (2005). A more difficult example is the nonparametric IV quantile regression $E[1\{Y_{1i} \leq h_o(Y_{2i})\} - \gamma|X_i] = 0$ for some known $\gamma \in (0,1)$ considered by Chernozhukov, Imbens and Newey (2007), Horowitz and Lee (2007) and Chen and Pouzo (2007). Both examples belong to the so-called 'ill-posed inverse' problems. See for example, Blundell and Powell (2003), Florens (2003) and Carrasco, Florens and Renault (2006) for additional examples of ill-posed inverse semi-nonparametric models.
Newey and Powell (2003) and Ai and Chen (2003) propose to estimate the model (3.5) by the *sieve minimum distance* (MD) procedure:

$$\sup_{\theta \in \Theta_n} \hat{Q}_n(\theta) = \sup_{\theta \in \Theta_n} -\frac{1}{n}\sum_{t=1}^{n} \widehat{m}(X_t,\ \theta)' \left\{\widehat{\Sigma}(X_t)\right\}^{-1} \widehat{m}(X_t,\ \theta), \tag{3.6}$$

where Θ_n is compact sieves, $\widehat{m}(X_t,\ \theta)$ is any nonparametrically consistent estimate of the conditional mean function $m(X_t,\ \theta) = E[\rho(Z,\ \theta)|X = X_t]$, and $\widehat{\Sigma}(X_t)$ is a possibly nonparametrically consistent estimate of a positive definite weighting matrix $\Sigma(X_t)$.
Example 3.3: (*nonparametric external habit-based consumption asset pricing models*): A consumption-based asset pricing model assumes that at time zero a representative agent maximizes the expected present value of the total utility

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

function $E_0\left\{\Sigma_{t=0}^{\infty}\delta^t u(C_t)\right\}$, where δ is the time discount factor and $u(C_t)$ is period t's utility. The consumption-based asset pricing models imply that for any traded asset indexed by •, with a gross return at time t+1 of $R_{\ell,t+1}$, the following Euler equation holds:

$$E\left(M_{t+1}R_{\ell,t+1}|\mathbf{w}_t\right) = 1,\ \ell = 1, \ldots, N, \tag{3.7}$$

where M_{t+1} is the intertemporal marginal rate of substitution in consumption, and $E(\cdot|\mathbf{w}_t)$ denotes the conditional expectation given the information set at time t (which is the sigma-field generated by $\mathbf{w}_t$). Hansen and Singleton (1982) have assumed that the period t utility takes the power specification $u(C_t) = [(C_t)^{1-\gamma} - 1] / [1-\gamma]$, where γ is the curvature parameter of the utility function at each period, which implies the Euler equation:

$$E\left(\delta_o\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma_o} R_{\ell,t+1} - 1|\mathbf{w}_t\right) = 0,\ \ell = 1, \ldots, N, \tag{3.8}$$

where the unknown scalar parameters δ_o, γ_o can be estimated by Hansen's (1982) GMM. However, this classical power utility-based asset pricing model (3.8) has been rejected empirically.
Chen and Ludvigson (2003) combine the power utility specification with a nonparametric internal habit formation: $E_0\left\{\Sigma_{t=0}^{\infty}\delta^t[(C_t - H_t)^{1-\gamma} - 1] / [1-\gamma]\right\}$, where H_t=$H(C_t,C_{t-1},\ldots,C_{t-L})$ is the period t habit level. Here $H(\cdot)$ is a homogeneous of degree one unknown function of current and past consumption, and can be rewritten as $H(C_t, C_{t-1}, \ldots, C_{t-L}) = C_t h_o\left(\frac{C_{t-1}}{C_t}, \ldots, \frac{C_{t-L}}{C_t}\right)$ with $h_o(\cdot)$ unknown. It is obvious that one needs to impose $0 \le h_o(\cdot) < 1$ so that $0 \le H_t < C_t$. The following external habit specification is a special case of their model:

$$E\left(\delta_o\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma_o} \frac{\left(1 - h_o\left(\frac{C_t}{C_{t+1}}, \ldots, \frac{C_{t+1-L}}{C_{t+1}}\right)\right)^{-\gamma_o}}{\left(1 - h_o\left(\frac{C_{t-1}}{C_t}, \ldots, \frac{C_{t-L}}{C_t}\right)\right)^{-\gamma_o}} R_{\ell,t+1} - 1|\mathbf{w}_t\right) = 0, \tag{3.9}$$

for $\ell = 1, \ldots, N$, where γ_o>0, δ_o>0 are unknown scalar preference parameters, $h_o(\cdot)\in[0,1)$ is an unknown function and $H_{t+1} = C_{t+1}h_o\left(\frac{C_t}{C_{t+1}}, \ldots, \frac{C_{t+1-L}}{C_{t+1}}\right)$ is the habit level at time t+1. Chen and Ludvigson (2003) have applied the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sieve method to estimate this model and its generalization which allows for internal habit formation of unknown form. Their empirical findings, using quarterly data, are in favor of flexible nonlinear internal habit formation.

3.3 Large sample properties

For an infinite-dimensional, compact parameter space Θ, Gallant and Nychka (1987) derived the consistency of sieve M-estimates; Newey and Powell (2003) established the consistency of sieve MD estimates. For an infinite-dimensional, possibly non-compact parameter space Θ, Geman and Hwang (1982) obtained the consistency of sieve MLE with i.i.d. data; White and Wooldridge (1991) obtained the consistency of sieve M-estimates with dependent and heterogeneous data; Chen (2006) and Chen and Pouzo (2007) established the consistency of sieve MD estimates that is applicable to general ill-posed semi-nonparametric problems.

For general theory on convergence rate of nonparametric part, we refer to Wong and Shen (1995) for sieve MLE, Shen and Wong (1994), Birgé and Massart (1998) and van de Geer (2000) for sieve M-estimation with i.i.d. data, Chen and Shen (1998) for sieve M-estimation with time series data, Newey (1997) and Huang (1998) for series LS estimation, and Chen and Pouzo (2007) for sieve MD estimation with i.i.d. data and allowing for ill-posed inverse problems.

For general theory on semiparametric efficiency and root-n normality of sieve estimates of smooth functionals, we refer to Shen (1997) for sieve MLE, Chen and Shen (1998) for sieve M-estimation with time series data, van de Geer (2000) for sieve M-estimation, Ai and Chen (2003) and Chen and Pouzo (2007) for sieve MD estimation allowing for ill-posed inverse problems. See Ai and Chen (2007) for a related result on root-n normality of sieve MD estimates of smooth functionals when the semi-nonparametric conditional moment models could be misspecified.

Unfortunately, so far there is no general theory on limiting pointwise distribution of sieve extremum estimators yet. Nevertheless, such results are established for series estimators (that is, sieve M-estimators with finite-dimensional linear sieves) (see Andrews, 1991b; Newey, 1997; Huang, 2003).

There is also no general theory on data-driven choice of smoothing parameters (‘complexity of sieves’) at the time of writing. There are some results for sieve M-estimation (see, for example, Barron, Birgé and Massart, 1999; Shen and Ye, 2002). There are well developed results for series LS regression and series density estimation (see for example, Li, 1987; Hurvich, Simonoff and Tsai, 1998; Andrews, 1991a; Coppejans and Gallant, 2002; Donald and Newey, 2001). Also see Chen (2006) for a detailed survey of theories on large sample properties of sieve estimation.

4 Economics applications

We conclude this article by listing some applications of the sieve extremum estimation in econometrics; see Chen (2006) for a more detailed review. In microeconometrics, Elbadawi, Gallant and Souza (1983) studied Fourier series LS estimation of demand elasticity. Heckman and Singer (1984) considered sieve ML estimation of a duration model where the unknown error distribution is approximated by a first-order spline. Hausman and Newey (1995) considered power series and spline series LS estimation of consumer surplus. Hahn (1998) used power series and splines in the two-step efficient estimation of the average treatment effect models. Newey, Powell and Vella (1999) considered series estimation of a triangular system of simultaneous equations. Gallant and Nychka (1987) proposed the Hermite polynomial sieve ML estimation of semiparametric sample selection model. Blundell, Chen and Kristensen (2007) considered a profile sieve MD procedure to estimate shape-invariant Engel curves with nonparametric endogenous expenditure. Hirano, Imbens and Ridder (2003) proposed a sieve logistic regression to estimate propensity score for treatment effect models. Chen, Fan and Tsyrennikov (2006) studied sieve MLE of semi-nonparametric multivariate copula models. Chen, Hong and Tamer (2005) made use of spline sieves to estimate nonlinear non-classical measurement error models with an auxiliary sample. Bierens and Carvalho (2007) applied Legendre polynomial sieve MLE to estimate a competing risks model of recidivism.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In time series econometrics, Engle et al. (1986) forecast electricity demand using a partially linear spline regression. Engle and Gonzalez-Rivera (1991) applied sieve MLE to estimate ARCH models where the unknown density of the standardized innovation is approximated by a first order spline sieve. Gallant and Tauchen (1989) employed Hermite polynomial sieve MLE to study asset pricing and foreign exchange rates. Gallant and Tauchen (1996) have proposed the combinations of Hermite polynomial sieve and simulated method of moments to effectively solve many complicated asset pricing models with latent factors, and their methods have been widely applied in empirical finance. White (1990) and Granger and Terasvirta (1993) suggested nonparametric LS forecasting via sigmoid ANN sieve. Chen, Racine and Swanson (2001) used partially linear ANN and ridgelet sieves to forecast US inflation. Shintani and Linton (2004) proposed a nonparametric test of chaos via ANN sieves. Chen and Ludvigson (2003) employed a sigmoid ANN sieve to estimate the unknown habit function in a consumption asset pricing model. Chen and Conley (2001) made use of the shape-preserving wavelet spline sieve to estimate a spatial temporal model with flexible conditional mean and conditional covariance. Phillips (1998) applied orthonormal basis to analyse spurious regressions.

See Also

- generalized method of moments estimation
- spline functions
- wavelets

Bibliography

Ai, C. and Chen, X. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71, 1795–843.

Ai, C. and Chen, X. 2007. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics* 141, 5–43.

Anastassiou, G. and Yu, X. 1992. Monotone and probabilistic wavelet approximation. *Stochastic Analysis and Applications* 10, 251–64.

Andrews, D. 1991a. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* 47, 359–77.

Andrews, D. 1991b. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59, 307–45.

Andrews, D. and Whang, Y. 1990. Additive interactive regression models: circumvention of the curse of dimensionality. *Econometric Theory* 6, 466–79.

Barron, A.R. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* 39, 930–45.

Barron, A., Birgé, L. and Massart, P. 1999. Risk bounds for model selection via penalization. *Probability of Theory Related Fields* 113, 301–413.

Bierens, H. and Carvalho, J. 2007. Semi-nonparametric competing risks analysis of recidivism. *Journal of Applied*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Econometrics 22, 971–93.

Birgé, L. and Massart, P. 1998. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* 4, 329–75.

Blundell, R., Chen, X. and Kristensen, D. 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* 75, 1613–69.

Blundell, R. and Powell, J. 2003. Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications*, ed. M. Dewatripont, L.P. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

Carrasco, M., Florens, J.-P. and Renault, E. 2006. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In *The Handbook of Econometrics*, vol. 6, ed. J.J. Heckman and E.E. Leamer. Amsterdam: North-Holland.

Chen, X. 2006. Large sample sieve estimation of semi-nonparametric models. In *The Handbook of Econometrics*, vol. 6, ed. J.J. Heckman and E.E. Leamer. Amsterdam: North-Holland.

Chen, X. and Conley, T. 2001. A new semiparametric spatial model for panel time series. *Journal of Econometrics* 105, 59–83.

Chen, X., Fan, Y. and Tsyrennikov, V. 2006. Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association* 101, 1228–40.

Chen, X., Hong, H. and Tamer, E. 2005. Measurement error models with auxiliary data. *Review of Economic Studies* 72, 343–66.

Chen, X. and Ludvigson, S. 2003. Land of addicts? An empirical investigation of habit-based asset pricing models. Manuscript, New York University.

Chen, X. and Pouzo, D. 2007. On estimation of semi-nonparametric conditional moment models with possibly nonsmooth moments. Manuscript, New York University.

Chen, X., Racine, J. and Swanson, N. 2001. Semiparametric ARX neural network models with an application to forecasting inflation. *IEEE Transactions on Neural Networks* 12, 674–83.

Chen, X. and Shen, X. 1998. Sieve extremum estimates for weakly dependent data. *Econometrica* 66, 289–314.

Chen, X. and White, H. 1999. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory* 45, 682–91.

Chernozhukov, V., Imbens, G. and Newey, W. 2007. Instrumental variable identification and estimation of nonseparable models via quantile conditions. *Journal of Econometrics* 139, 4–14.

Chui, C. 1992. *An Introduction to Wavelets*. San Diego: Academic Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Coppejans, M. and Gallant, A.R. 2002. Cross-validated SNP density estimates. *Journal of Econometrics* 110, 27–65.

Darolles, S., Florens, J.-P. and Renault, E. 2006. Nonparametric instrumental regression. Mimeo, GREMAQ, University of Toulouse.

de Boor, C. 1978. *A Practical Guide to Splines*. New York: Springer-Verlag.

DeVore, R.A. and Lorentz, G.G. 1993. *Constructive Approximation*. Berlin: Springer.

Donald, S. and Newey, W. 2001. Choosing the number of instruments. *Econometrica* 69, 1161–91.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. 1995. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society, Series B* 57, 301–69.

Elbadawi, I., Gallant, A.R. and Souza, G. 1983. An elasticity can be estimated consistently without a prior knowledge of functional form. *Econometrica* 51, 1731–51.

Engle, R. and Gonzalez-Rivera, G. 1991. Semiparametric ARCH models. *Journal of Business and Economic Statistics* 9, 345–59.

Engle, R., Granger, C., Rice, J. and Weiss, A. 1986. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81, 310–20.

Florens, J.P. 2003. Inverse problems and structural econometrics: the example of instrumental variables. In *Advances in Economics and Econometrics: Theory and Applications*, ed. M. Dewatripont, L.P. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

Gallant, A.R. and Nychka, D. 1987. Semi-non-parametric maximum likelihood estimation. *Econometrica* 55, 363–90.

Gallant, A.R. and Tauchen, G. 1989. Semiparametric estimation of conditional constrained heterogenous processes: asset pricing applications. *Econometrica* 57, 1091–120.

Gallant, A.R. and Tauchen, G. 1996. Which moments to match? *Econometric Theory* 12, 657–81.

Geman, S. and Hwang, C. 1982. Nonparametric maximum likelihood estimation by the method of Sieves. *Annals of Statistics* 10, 401–14.

Granger, C.W.J. and Terasvirta, T. 1993. *Modelling nonlinear economic relationships*. New York: Oxford University Press.

Grenander, U. 1981. *Abstract Inference*. New York: Wiley.

Hahn, J. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66, 315–32.

Hall, P. and Horowitz, J. 2005. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics* 33, 2904–29.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Hansen, L.P. and Singleton, K. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–86.

Hausman, J. and Newey, W. 1995. Nonparametric estimation of exact consumer surplus and deadweight loss. *Econometrica* 63, 1445–67.

Heckman, J. and Singer, B. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 68, 839–74.

Hirano, K., Imbens, G. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–89.

Hornik, K., Stinchcombe, M., White, H. and Auer, P. 1994. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation* 6, 1262–75.

Horowitz, J. and Lee, S. 2007. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* 75, 1191–208.

Huang, J.Z. 1998. Projection estimation in multiple regression with application to functional ANOVA models. *Annals of Statistics* 26, 242–72.

Huang, J.Z. 2003. Local asymptotics for polynomial spline regression. *Annals of Statistics* 31, 1600–35.

Hurvich, C., Simonoff, J. and Tsai, C. 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* 60, 271–93.

Judd, K. 1998. *Numerical Method in Economics*. Cambridge, MA: MIT Press.

Koenker, R., Ng, P. and Portnoy, S. 1994. Quantile smoothing splines. *Biometrika* 81, 673–80.

Li, K. 1987. Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: discrete index set. *Annals of Statistics* 15, 958–75.

Meyer, Y. 1992. *Ondelettes et operateurs I: Ondelettes*. Paris: Hermann.

Newey, W.K. 1997. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79, 147–168.

Newey, W.K. and Powell, J.L 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565–78.

Newey, W.K., Powell, J.L. and Vella, F. 1999. Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67, 565–603.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Phillips, P.C.B. 1998. New tools for understanding spurious regressions. *Econometrica* 66, 1299–325.

Robinson, P. 1988. Root-n-consistent semiparametric regression. *Econometrica* 56, 931–54.

Schumaker, L. 1981. *Spline Functions: Basic Theory*. New York: Wiley.

Shen, X. 1997. On methods of sieves and penalization. *Annals of Statistics* 25, 2555–91.

Shen, X. and Wong, W. 1994. Convergence rate of sieve estimates. *Annals of Statistics* 22, 580–615.

Shen, X. and Ye, J. 2002. Adaptive model selection. *Journal of American Statistical Association* 97, 210–21.

Shintani, M. and Linton, O. 2004. Nonparametric neural network estimation of Lyapunov exponents and a direct test for chaos. *Journal of Econometrics* 129, 1–33.

Stone, C.J. 1985. Additive regression and other nonparametric models. *Annals of Statistics* 13, 689–705.

Van de Geer, S. 2000. *Empirical Processes in M-estimation*. Cambridge: Cambridge University Press.

Wahba, G. 1990. *Spline Models for Observational Data*. Philadelphia: CBMS-NSF Regional Conference Series.

White, H. 1990. Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks* 3, 535–50.

White, H. and Wooldridge, J. 1991. Some results on sieve estimation with dependent observations. In *Non-parametric and Semi-parametric Methods in Econometrics and Statistics*, ed. W.A. Barnett, J. Powell and G. Tauchen. Cambridge: Cambridge University Press.

Wong, W.H. and Shen, X. 1995. Probability inequalities for likelihood ratios and convergence rates for sieve MLE's. *Annals of Statistics* 23, 339–62.

How to cite this article

Chen, Xiaohong. "sieve extremum estimation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_N000163> doi:10.1057/9780230226203.1525

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

signalling and screening

Johannes Hörner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Signalling refers to any activity by a party designed to influence the perception and thereby the actions of other parties. This presupposes that one market participant holds private information that for some reason cannot be verifiably disclosed, and which affects the other participants' incentives. The classic example of market signalling is due to Spence. Consider a labour market in which firms know less than workers about their innate productivity. Under certain conditions, some workers may wish to signal their ability to potential employers, and do so by choosing a level of education that distinguishes them from workers with lower productivity.

Keywords

asymmetric information; introductory pricing; limit pricing; multiple equilibria; political competition; pooling equilibrium; screening; separating equilibrium; signalling; single-crossing conditions; signalling and screening; uniqueness; warranties

Article

Mathematical formulation

We consider here the simplest game-theoretic version of Spence's (1973; 1974) model. A worker's productivity, or *type*, is either θ_H or θ_L, with $\theta_H > \theta_L > 0$. Productivity is private information. Firms share a common prior *p*, with $p = \Pr\{\theta = \theta_H\} \in (0, 1)$. Before entering the job market, the worker chooses a costly education level $e \geq 0$. Workers maximize $U(w, e; \theta) = w - c(e; \theta)$, where w is their wage and $c(e; \theta)$ is the cost of education. Assume that $c(0; \theta) = 0$, $c_e(e; \theta) > 0$, $c_{ee}(e; \theta) > 0$, $c_\theta(e; \theta) < 0$ for all $e > 0$, where $c_e(\cdot; \theta)$ and $c_\theta(e; \cdot)$ denote the derivatives of the cost with respect to education and types, respectively, and $c_{ee}(\cdot; \theta)$ is the second derivative of cost with respect to education. The key assumption made in the literature is that on the cross-derivative: $c_{e\theta}(e; \theta) < 0$. That is, the marginal cost is lower for a high-productivity worker. This *single-crossing* condition ensures that the indifference curves of a high and a low-productivity worker cross at most once, with the indifference curve of the high-productivity worker having a smaller slope where they do. See Figure 1.

Figure 1

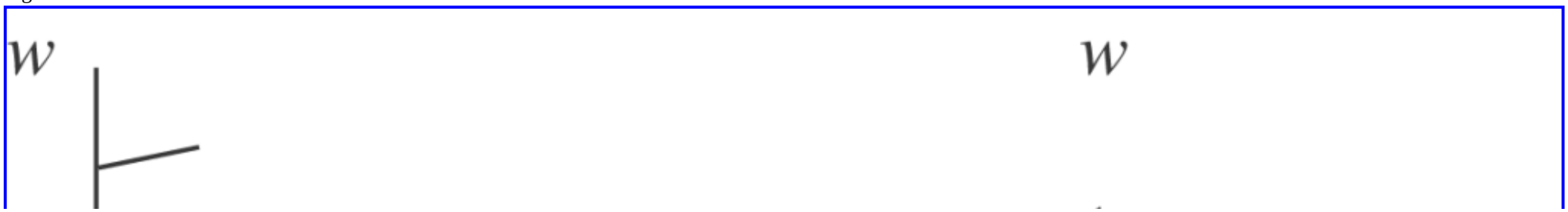

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

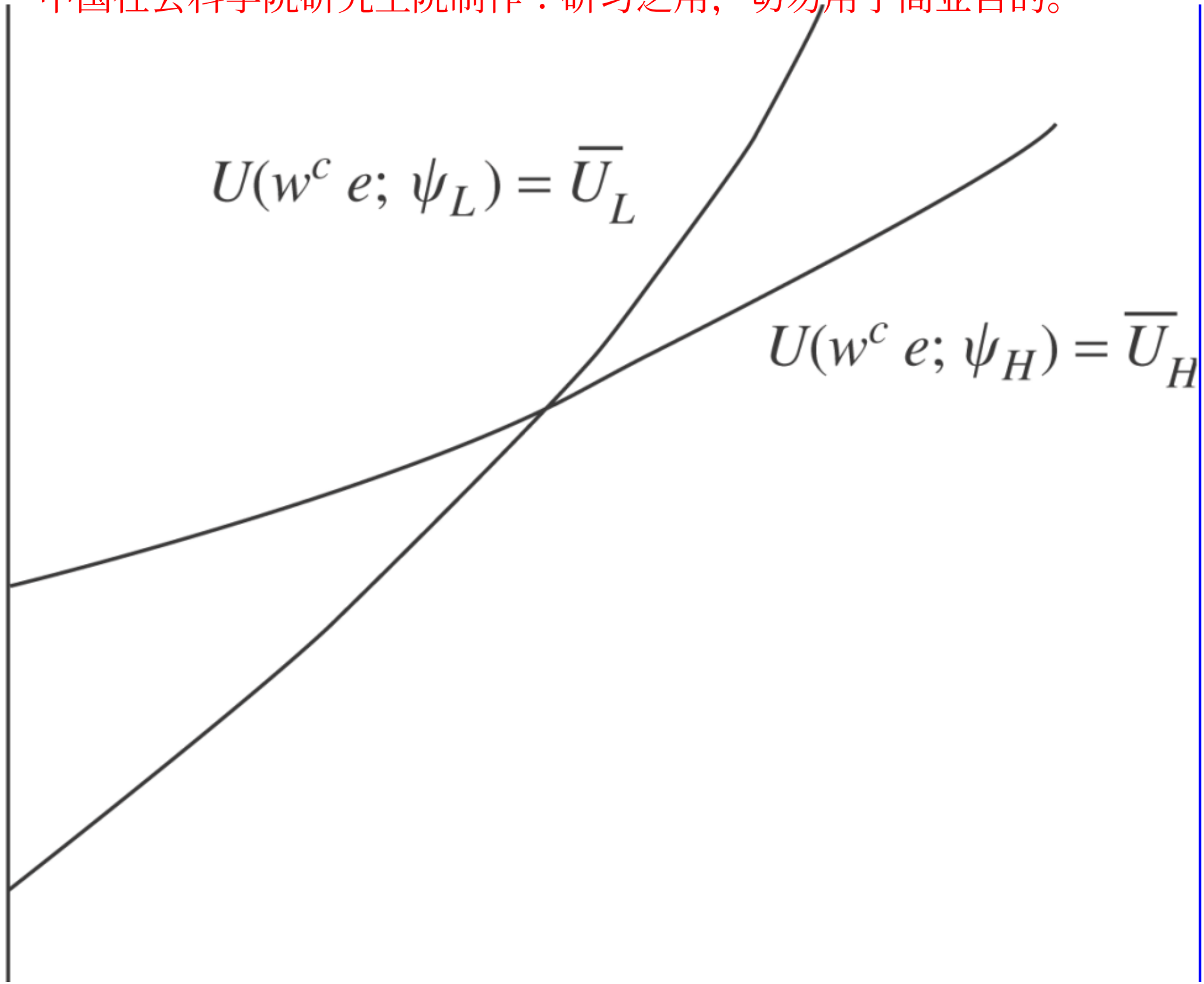

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

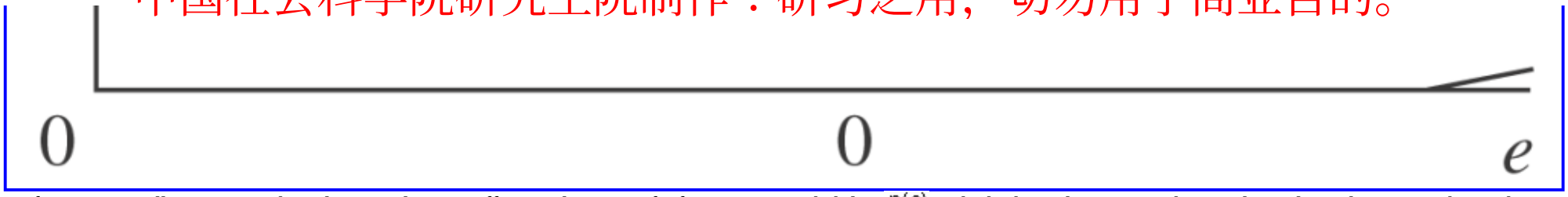

To focus on signalling, assume that education does not affect productivity. If a firm assigns probability $p(e)$ to the high-productivity worker conditional on education *e*, the worker's expected productivity is $(1-p(e))\theta_L + p(e)\theta_H$. If the worker accepts wage *w*, the firm's profit is then $(1-p(e))\theta_L + p(e)\theta_H - w$.

The basic signalling game

There is one worker and two firms. In the first stage, the worker chooses education. Education is observable. In the second stage, firms compete through wages in Bertrand-competition fashion. Following usual arguments, the wage $w(e)$ offered and accepted equals the expected productivity of the worker, given the observed education. A (perfect Bayesian) equilibrium specifies an education function $e^*: \{\theta_L, \theta_H\} \to R_+$, and a belief function $p^*: R_+ \to [0, 1]$, giving respectively the education chosen by each type and the probability assigned to a high-productivity worker conditional on each possible education level, so that the worker's choice is optimal given the wage determined by p^*, and the belief function is derived from this choice using Bayes's rule whenever possible.

Either the worker chooses distinct education levels depending on his productivity or he does not. An equilibrium is *separating* if $e^*(\theta_L) \neq e^*(\theta_H)$, and *pooling* otherwise. In the first case, education perfectly reveals productivity, so that the worker's wage equals his productivity: $w\left(e^*(\theta_i)\right) = \theta_i$, for $i = L, H$. In the second case, education reveals no information, and the wage is equal to his expected productivity given the prior *p*: $w\left(e^*(\theta_i)\right) = (1-p)\theta_L + p\theta_H =: E(\theta)$, for $i = L, H$.

Observe that, in any separating equilibrium, the low-productivity worker gets the lowest possible wage. Since education is costly, this implies that he chooses no education: $e^*(\theta_L) = 0$. The high-productivity worker, on the other hand, gets the highest possible wage. Therefore, the corresponding education level $e^*(\theta_H)$ must be high enough to deter the low-productivity worker from pretending he has high productivity. That is, it must be that $e^*(\theta_H) \geq e'$, where e' solves $\theta_L - c(0; \theta_L) = \theta_H - c\left(e'; \theta_L\right)$. At the same time, the education level $e^*(\theta_H)$ cannot be too high, since the high-productivity worker must choose it. That it, it must be that $e^*(\theta_H) \leq e''$, where e'' solves $\theta_L - c(0; \theta_H) = \theta_H - c\left(e''; \theta_H\right)$. Single-crossing implies that the interval $\left[e', e''\right]$ is non-empty. Indeed, since the indifference curves $\{(e, w): U(w, e; \theta_H) = U(\theta_L, 0; \theta_H)\}$ and $\{(e, w): U(w, e; \theta_L) = U(\theta_L, 0; \theta_L)\}$ cross at $(0, \theta_L)$, the point $\left(e'', \theta_H\right)$ which lies along the first one must be to the right of the point $\left(e', \theta_H\right)$ which lies along the second. See Figure 2. Because a high-productivity worker is more willing to trade off an increase in education to induce an increase in wage, it is possible to find a suitable education level that is worth acquiring if and only if the worker's productivity is high.

Figure 2

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

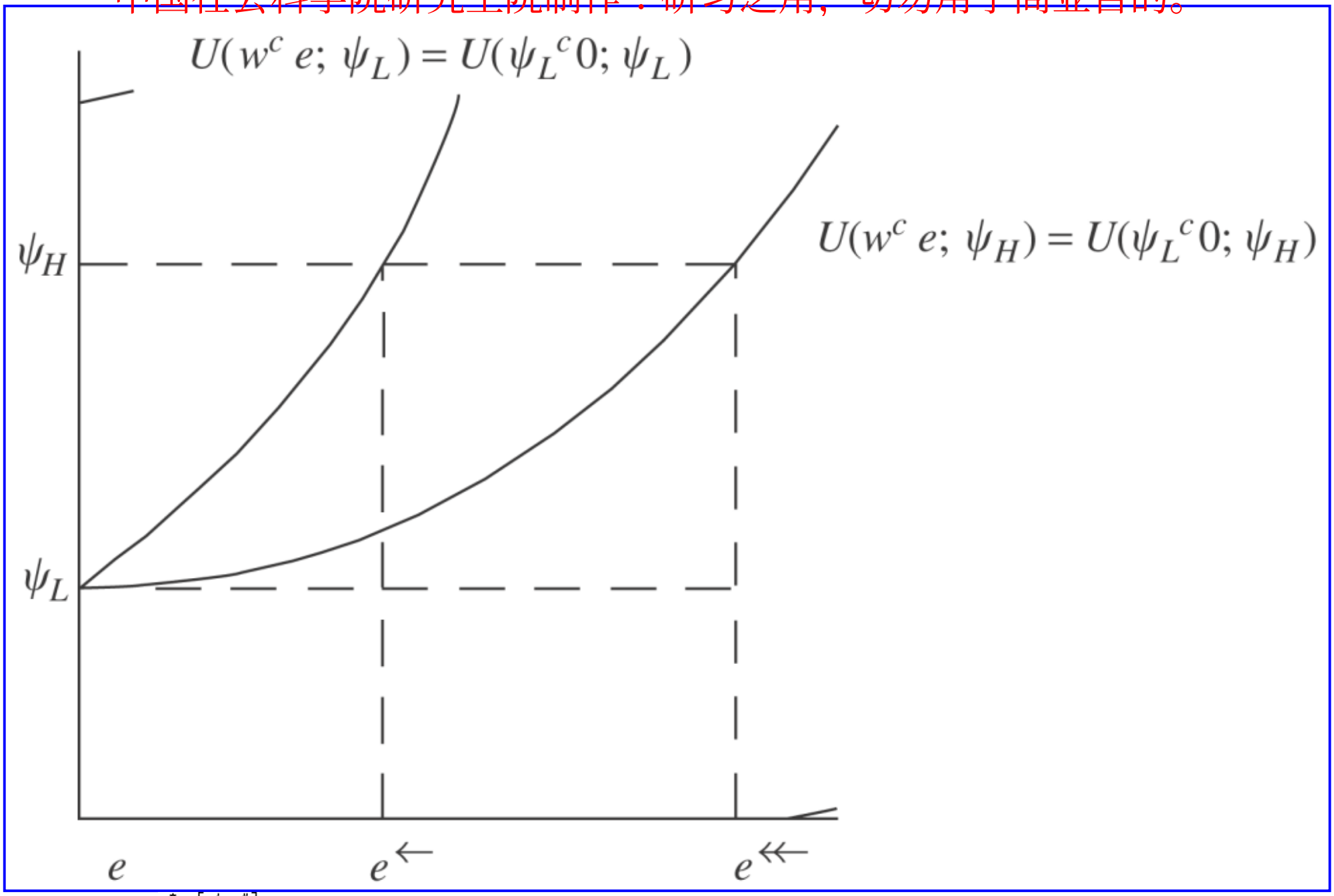

Each education level $e^* \in [e', e'']$ can be supported in equilibrium, for suitable beliefs. For instance, by setting $p^*(e) = 0$ if $e < e^*$, and $p^*(e) = 1$ otherwise, a high-productivity worker optimally chooses $e^*(\theta_H) = e^*$. Observe that these equilibria are Pareto-ranked. The best equilibrium outcome, involving $e^* = e'$, is known as the *Riley outcome* (Riley, 1979).

The low-productivity worker is worse off than in the case in which signalling is not available. Without signalling, the worker would not acquire any education, independently of his type, and he would receive the wage $E(\theta)$. Here, instead, the low-productivity worker earns only θ_L. Surprisingly, the high-productivity worker may also be worse off. As no

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

education is interpreted as evidence of low productivity, the outcome without signalling is no longer available to him in a separating equilibrium, his utility is at most $\theta_H - c(e'; \theta_H)$. Without signalling opportunities, his utility is $E(\theta) - c(0; \theta_H)$. While $E(\theta)$ tends to θ_H as *p* tends to one, e' is independent of *p*. Therefore, if *p* is large enough, the high-productivity is worse off. See Figure 3.

Figure 3

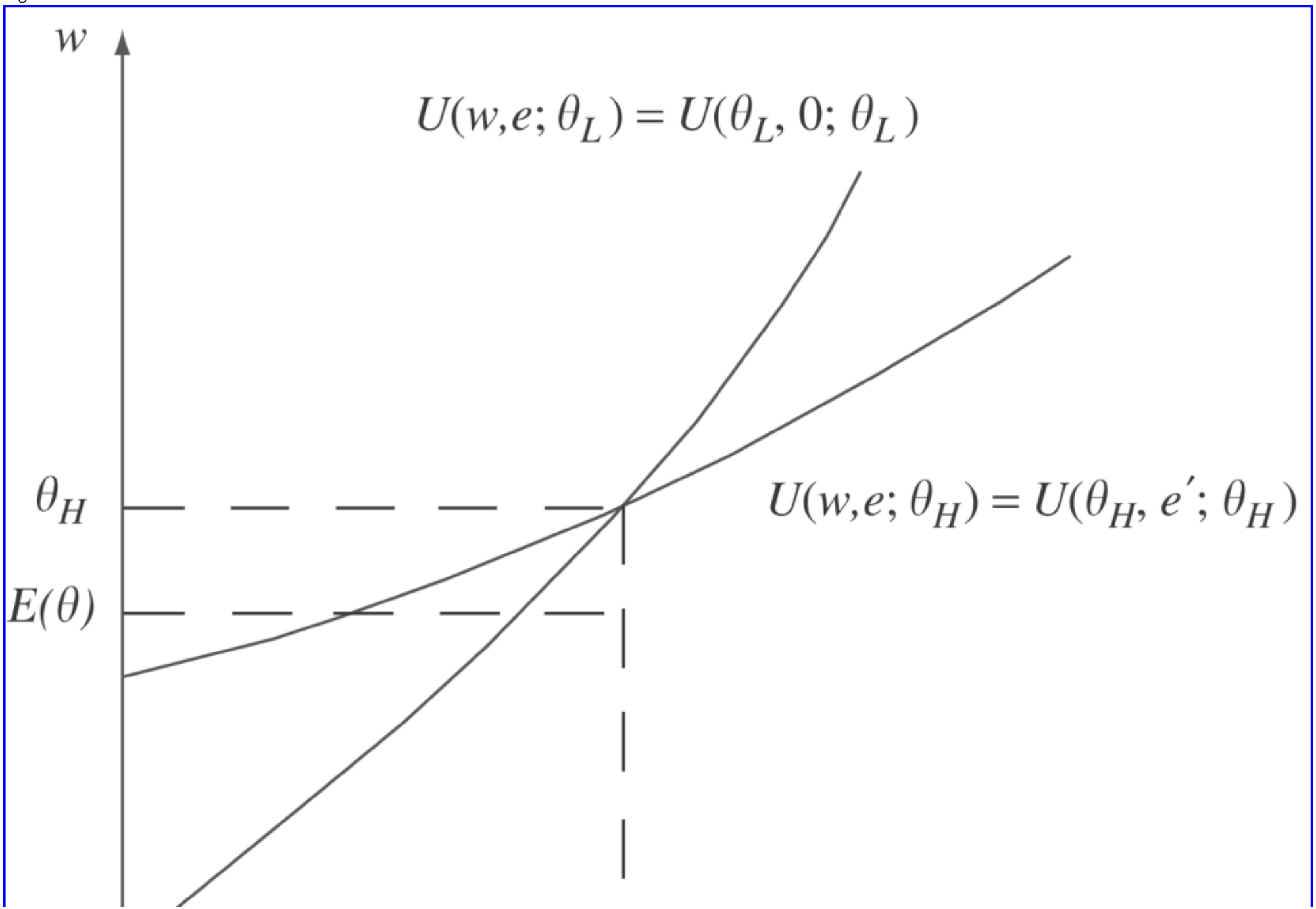

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

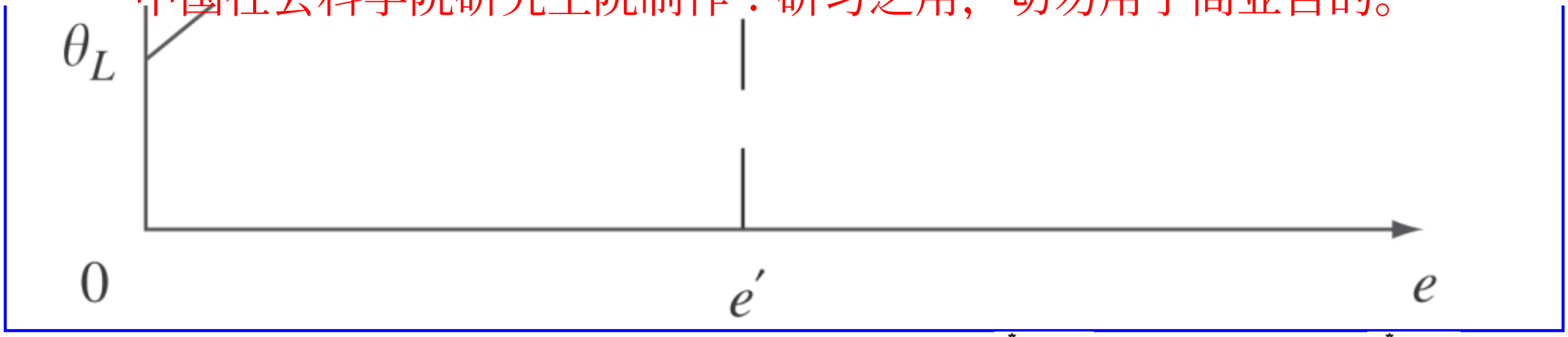

There is also a continuum of pooling outcomes. Let $\hat{e}$ solve $U(\theta_L, 0; \theta_L) = U(E(\theta), \hat{e}; \theta_L)$. Every level of education $e^* \in [0, \hat{e}]$ can be supported by the beliefs $p^*(e) = 0$ if $e < e^*$, $p^*(e) = p$ otherwise. Since education is costly, we need only check that the worker prefers $e(\theta_i) = e^*$ ($i = L, H$) to no education at all. This is true for the low-productivity worker by definition of $\hat{e}$, and follows from single-crossing for the high-productivity worker. Here as well, the outcomes are Pareto-ranked, with the best equilibrium outcome, sometimes referred to as the *Hellwig outcome*, specifying $e^* = 0$ (Hellwig, 1987). In addition to these separating and pooling outcomes, there also exists a continuum of equilibria in mixed strategies.

The basic screening game

While it is standard in the literature to call signalling models those in which the informed party moves first, they are closely related to screening models, in which the uninformed parties take the lead. Classic references include Rothschild and Stiglitz (1976) and Wilson (1977) in the context of insurance markets. In these models, the two firms simultaneously announce a menu of pairs (e, w). Given these contracts, the worker chooses which contract to accept, if any. We sketch here the main results of this model. An equilibrium is *separating* if the worker accepts distinct contracts depending on his type, and *pooling* otherwise.
Observe that, in equilibrium, firms must just break even. Otherwise, if the worker of type $i = L, H$ accepts contract (e_i, w_i), a contract $(e_i, w_i + \varepsilon)$ for small $\varepsilon > 0$ would attract both types of worker, and the firm earning less than half the aggregate profits would gain by offering it.

Also, there can be no pooling equilibrium. Because a pooling contract (e^*, w^*) would have to break even, a firm whose rival offered this contract would gain by offering a contract (e, w) specifying a higher wage and education level, accepted only by the high-productivity worker. See Figure 4.

Figure 4

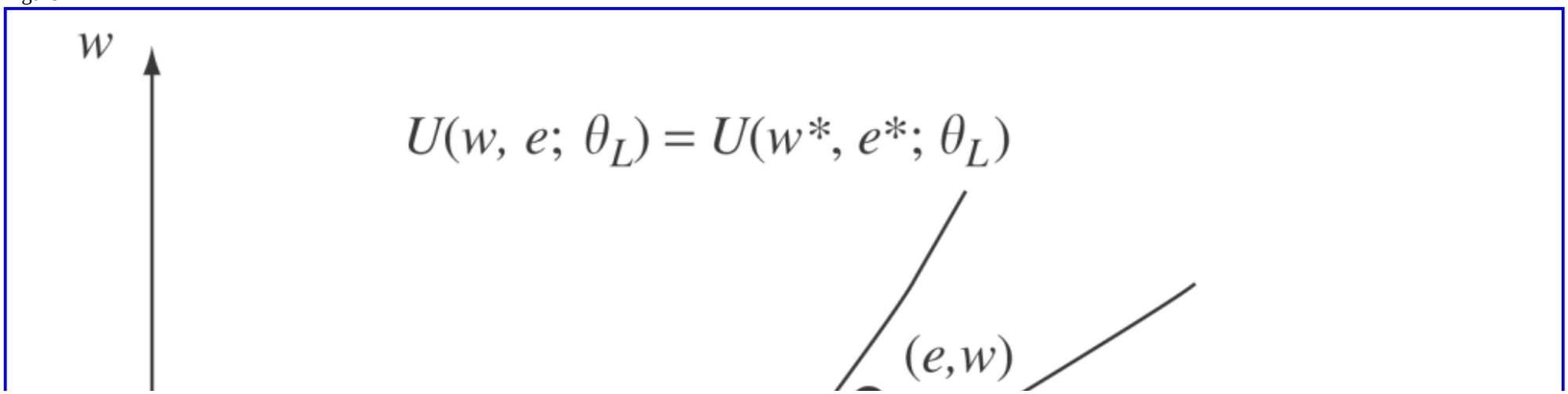

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

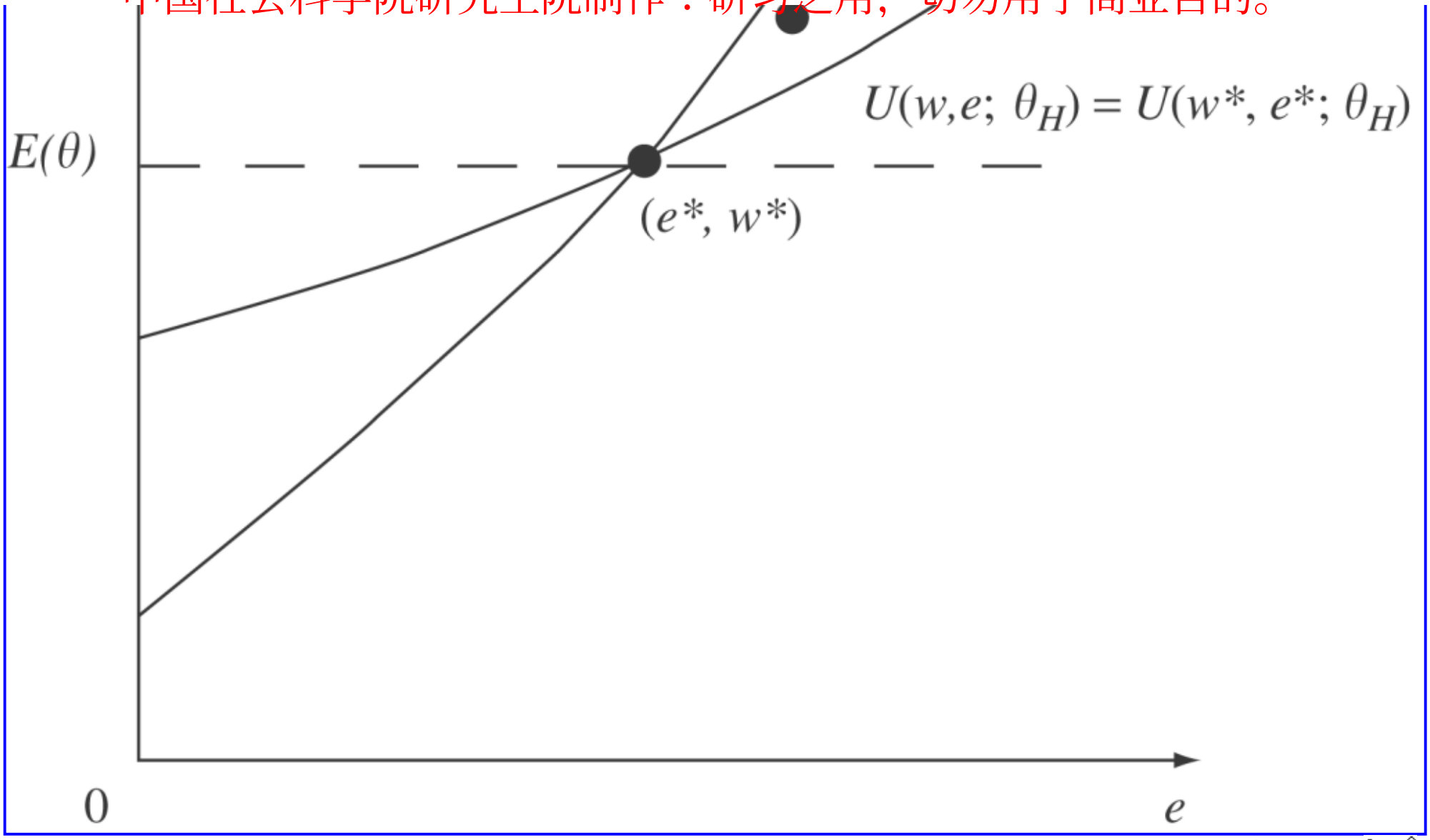

Finally, in any separating equilibrium, wages paid must equal the worker's productivity. In particular, the contract accepted by the low-productivity worker specifies education $e_L = 0$ and wage $w_L = \theta_L$. Indeed, if $(e_L, w_L) \neq (0, \theta_L)$, then the firm whose rival offered (e_L, w_L) would gain by offering a contract with either a slightly lower wage, or a slightly lower education, independently of the worker's type accepting it. Similarly, if the wage accepted by the high-productivity worker fell short of θ_H, a firm whose rival offered the contract accepted by the low-productivity worker would gain by offering a contract specifying a slightly higher wage and education than those specified by the contract accepted by the high-productivity worker. Since $w_i = \theta_i$, $i = L, H$, and $e_L = 0$, it follows that e_H solves $\theta_H - c(e_H;\ \theta_L) = \theta_L - c(0;\ \theta_L)$, that is, $e_H = e'$: if instead the low-productivity worker preferred his contract to (e_H, w_H), at least one firm would gain by offering a contract specifying a wage and education just below w_H and e_H.

However, such an equilibrium need not exist for large *p*. If $E(\theta) - c(0;\ \theta_H) > \theta_H - c\left(e';\ \theta_H\right)$, the contract $(0,\ E(\theta) - \varepsilon)$ for small $\varepsilon > 0$ attracts both types of workers and makes profits. Thus, if the cost of sorting outweighs the gain, no equilibrium exists. As emphasized by Riley (2001), existence requires a strengthening of single-crossing, as marginal cost must be sufficiently lower for a high-productivity worker, given *p*. This is the same condition as earlier, under which the high-productivity worker prefers signalling to be unavailable. While equilibria in mixed strategies exist (Dasgupta and Maskin, 1986), they have not been characterized.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Extensions and refinements

To a large extent, the theoretical literature on signalling has focused on selecting among the equilibria, while the literature on screening has addressed the non-existence issue.
The early literature takes the view that the screening model ignores the dynamic adjustments between firms. To account for these, Wilson (1977) and Riley (1979) define equilibria differently. A set of contracts is a Wilson equilibrium if no firm has a profitable deviation that remains profitable once existing contracts that lose money after the deviation are withdrawn, while it is a Riley, or reactive equilibrium, if no firm has a profitable deviation that remains profitable once new contracts that make money after the deviation are added. Under either definition, equilibria exist. Wilson equilibria involve some pooling, while the unique Riley equilibrium is separating. Hellwig (1987) offers a game-theoretic treatment of Wilson by modelling a second stage in which firms may withdraw any contract offered previously. In the two-type case, the Hellwig outcome can be supported as an equilibrium.
Formal game-theoretic treatments of signalling appear in the 1980s. Many refinements have been applied to and inspired by the basic signalling game, shedding new light on the somewhat ad hoc selection procedures used previously. While sequential equilibrium (Kreps and Wilson, 1982) does not reduce the multiplicity of equilibria, the *intuitive criterion* (Cho and Kreps, 1987) selects the Riley outcome in the basic signalling model. This result, as striking as it is, has several limitations. First, uniqueness does not obtain with more types. Second, the Riley outcome is not necessarily persuasive when the probability of the high-productivity worker is nearly one. As long as this probability is less than 1, the high-productivity worker acquires education $e' > 0$, independently of *p*. But if he is known to be of high productivity, we should expect him not to acquire education, as it serves no signalling purpose. Third, the motivation behind the intuitive criterion also underlies the more stringent *perfect sequential equilibrium* (Grossman and Perry, 1986). Yet such an equilibrium fails to exist in the situation described earlier, in which the basic screening game has no equilibrium. An alternative is offered by the concept of *undefeated equilibrium* (Mailath, Okuno-Fujiwara and Postlewaite, 1993), which selects the Riley outcome when it is also a perfect sequential equilibrium outcome, and the Hellwig outcome otherwise.
In settings with more types, stronger refinements are needed to select the Riley outcome. These include Banks and Sobel's (1987) *divinity* and *universal divinity*, Cho and Kreps's (1987) criterion *D1*, and Kohlberg and Mertens's (1986) *stability* concepts. See also Cho and Sobel (1990). It is worth pointing out that, with a continuum of types and under weak assumptions, the separating outcome is unique in the signalling model (Mailath, 1987), while no equilibrium exists in the screening model (Riley, 2001).
The single-crossing condition has been generalized to multidimensional signals by Engers (1987) for the case of screening and by Cho and Sobel (1990) and Ramey (1996) for the case of signalling. Quinzii and Rochet (1985) consider multidimensional types. Little is known about equilibria when single-crossing fails, as may occur in applications.
Maskin and Tirole (1992) enlarge the set of contracts. In the screening version, firms offer contracts that let the worker choose *ex post* among a set of pairs (e, w). In the signalling version, the worker offers such a contract to the firm. Under weak assumptions, the set of equilibrium outcomes coincides. In particular, only the Riley outcome obtains when the basic screening model has an equilibrium.
Acquiring education takes time, and there is no reason to expect firms to wait until graduation before drawing inferences. In Nöldeke and van Damme (1990) and Swinkels (1999), firms make offers before workers complete their education. Offers are public in Nöldeke and van Damme, and private in Swinkels. As the time between offers diminishes, only the Riley outcome satisfies Kohlberg and Mertens' never weak best response criterion when offers are public, while only the Hellwig outcome is a sequential equilibrium outcome when offers are private.
Following Spence's early suggestion, Nöldeke and Samuelson (1997) extend the basic signalling model by considering a dynamic model in which agents adjust their beliefs and actions to past market outcomes, and introduce perturbations into the process. The dynamic process admits at most two recurrent sets, closely related to the Riley and Hellwig outcomes. Several known refinements reappear in their characterization.

Application

Signalling has found many applications besides education, insurance and labour. Whenever possible, the reader is referred to surveys.

Industrial organization

Signalling helps explain limit (or predatory) pricing. Milgrom and Roberts (1982) show that low price may signal an incumbent's low cost. In Milgrom and Roberts (1986), advertising is a signal that a firm's experience good is of high quality. Bagwell and Riordan (1991) show that introductory pricing can serve the same purpose. In Gal-Or (1989), warranties signal product durability. See Tirole (1988) for a general survey, and Bagwell (2001) for a survey specific to advertising and pricing.

Finance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Myers and Majluf (1984) show that stock issues may signal the firm's value, so that the choice of financing (equity or debt) affects a firm's investment policy when the firm is better informed about returns than investors. In Bhattacharya (1979), dividends signal cash flows, as managers are better informed about those than investors. In Leland and Pyle (1977), the owner's stake signals the firm's underlying quality in an initial public offering, provided the owner is risk averse. These early applications have been extended in several directions, and their predictions empirically tested. See Allen and Morris (2001).

Political science

Signalling has been applied to electoral competition. Banks (1990a) shows how campaign platforms signal the candidates' future actions if elected, while Banks (1990b) argues that agenda-setting signals the bureaucrat's private information about the 'reversion level' if the proposal is turned down. See Banks (1991) for a survey. Lohmann (1993) shows how individuals engage in costly political actions to signal their private information, if politicians are responsive to turnout. Prat (2002) considers a signalling game in which an interest group has private information about candidates, based on which they can offer contributions that are used as campaign advertising.

Social norms

Bernheim (1994) develops a theory of social conformity based on signalling. Agents are motivated both by private tastes and status. When society censures extreme preferences, consumers with centrist preferences may choose to pool, while extremists refuse to conform. Fang (2001) provides a model of social culture rich enough to endogenously generate the single-crossing condition supporting the separating equilibrium. Austen-Smith and Fryer (2005) study signalling when workers have both a social and an economic type, and education affects both the wage and their perception by their peers. Peer pressure may induce educational underinvestment by accepted types.

Biology

Following Zahavi's (1975) 'handicap principle', asserting that animal signals are reliable because they are costly, a large literature on signalling has emerged in biology. Grafen (1990) provides a game-theoretic treatment. See Maynard Smith and Harper (2003) for a survey.

See Also

- cheap talk

Bibliography

Allen, F. and Morris, S. 2001. Finance applications of game theory. In *Game Theory and Business Applications*, ed. K. Chatterjee and W.F. Samuelson. Boston, MA: Kluwer Academic Publishers.

Austen-Smith, D. and Fryer, R.G. 2005. An economic analysis of acting white. *Quarterly Journal of Economics* 119, 551–83.

Bagwell, K. 2001. Introduction. In *The Economics of Advertising*, ed. K. Bagwell. Cheltenham: Edward Elgar.

Bagwell, K. and Riordan, M.H. 1991. High and declining prices signal product quality. *American Economic Review* 81, 224–39.

Banks, J.S. 1990a. A model of electoral competition with incomplete information. *Journal of Economic Theory* 50, 309–25.

Banks, J.S. 1990b. Monopoly agenda control and asymmetric information. *Quarterly Journal of Economics* 104, 445–64.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Banks, J.S. 1991. *Signaling Games in Political Science*. New York: Harwood Academic.

Banks, J.S. and Sobel, J. 1987. Equilibrium selection in signaling games. *Econometrica* 55, 647–61.

Bernheim, B.D. 1994. A theory of conformity. *Journal of Political Economy* 102, 841–77.

Bhattacharya, S. 1979. Imperfect information, dividend policy and 'the bird in the hand' fallacy. *Bell Journal of Economics* 9(1), 259–70.

Cho, I.-K. and Kreps, D.M. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102, 179–221.

Cho, I.-K. and Sobel, J. 1990. Strategic stability and uniqueness in signaling games. *Journal of Economic Theory* 50, 381–413.

Dasgupta, P. and Maskin, E. 1986. The existence of equilibria in discontinuous economic games, I: theory. *Review of Economic Studies* 53, 1–26.

Engers, M. 1987. Signalling with many signals. *Econometrica* 55, 663–74.

Fang, H. 2001. Social culture and economic performance. *American Economic Review* 91, 924–37.

Gal-Or, E. 1989. Warranties as a signal of quality. *Canadian Journal of Economics* 22, 50–61.

Grafen, A. 1990. Biological signals as handicaps. *Journal of Theoretical Biology* 144, 517–46.

Grossman, S. and Perry, M. 1986. Perfect sequential equilibrium. *Journal of Economic Theory* 39, 97–119.

Hellwig, M. 1987. Some recent developments in the theory of competition in markets with adverse selection. *European Economic Review* 31, 319–25.

Kohlberg, E. and Mertens, J.-F. 1986. On the strategic stability of equilibria. *Econometrica* 54, 1003–37.

Kreps, D.M. and Sobel, J. 1994. Signaling. In *Handbook of Game Theory*, vol. 2, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Kreps, D.M. and Wilson, R. 1982. Sequential equilibria. *Econometrica* 50, 863–94.

Leland, H.E. and Pyle, D.H. 1977. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance* 32, 371–87.

Lohmann, S. 1993. A signaling model of informative and manipulative political action. *American Political Science Review* 87, 319–33.

Mailath, G.J. 1987. Incentive compatibility in signaling games with a continuum of types. *Econometrica* 55, 1349–65.

Mailath, G.J., Okuno-Fujiwara, M. and Postlewaite, A. 1993. Belief-based refinements in signalling games. *Journal of Economic Theory* 60, 241–76.

Maskin, E. and Tirole, J. 1992. The principal–agent relationship with an informed principal: the case of common values. *Econometrica* 60, 1–42.

Maynard, S.J. and Harper, D. 2003. *Animal Signals*. New York: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Milgrom, P. and Roberts, J. 1982. Limit pricing and entry under incomplete information: an equilibrium analysis. *Econometrica* 50, 443–59.

Milgrom, P. and Roberts, J. 1986. Price and advertising signals of product quality. *Journal of Political Economy* 94, 796–821.

Myers, S.C. and Majluf, N. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13, 187–221.

Nöldeke, G. and Samuelson, L. 1997. A dynamic model of equilibrium selection in signaling markets. *Journal of Economic Theory* 73, 118–56.

Nöldeke, G. and van Damme, E. 1990. Signalling in a dynamic labour market. *Review of Economic Studies* 57, 1–23.

Prat, A. 2002. Campaign advertising and voter welfare. *Review of Economic Studies* 69, 999–1017.

Quinzii, M. and Rochet, J.-C. 1985. Multidimensional signalling. *Journal of Mathematical Economics* 14, 261–84.

Ramey, G. 1996. D1 signaling equilibria with multiple signals and a continuum of types. *Journal of Economic Theory* 69, 508–31.

Riley, J. 1979. Informational equilibrium. *Econometrica* 47, 331–59.

Riley, J. 2001. Silver signals: twenty-five years of screening and signaling. *Journal of Economic Literature* 39, 432–78.

Rothschild, M. and Stiglitz, J. 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 80, 629–49.

Spence, A.M. 1973. Job market signalling. *Quarterly Journal of Economics* 90, 225–43.

Spence, A.M. 1974. *Market Signaling*. Cambridge, MA: Harvard University Press.

Swinkels, J. 1999. Education signalling with preemptive offers. *Review of Economic Studies* 66, 949–70.

Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

Wilson, C.A. 1977. A model of insurance markets with incomplete information. *Journal of Economic Theory* 16, 167–207.

Zahavi, A. 1975. Mate selection – a selection for a handicap. *Journal of Theoretical Biology* 53, 205–14 .

How to cite this article

Hörner, Johannes. "signalling and screening." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000129> doi:10.1057/9780230226203.1526

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

silver standard

Lawrence H. Officer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Silver was the dominant monetary standard for many centuries. It was supplanted by gold, in England in the early 18th century, in the United States in 1834, and in almost all of western Europe and in most of the British Empire in the 1870s and 1880s. With France no longer bimetallic, the exchange rate between gold-standard and silver-standard currencies fluctuated. The consequence was more countries leaving the silver standard for gold. By the First World War, China was the sole major country still on silver. US silver-purchase policy in the 1930s caused the virtual demonetization of silver worldwide.

Keywords

bimetallism; bullion; central banking; gold standard; gold-exchange standard; monometallism; seigniorage; silver standard

Article

The silver standard, the dominant monetary system for many centuries, lost much importance with the advent of the classical gold standard; and, due to US policy, residual monetary use of silver was virtually eliminated in the 1930s.

Definition of silver standard

A silver standard involves (*a*) a fixed silver content of the monetary unit, (*b*) ‘free coinage’ of silver, that is, privately owned silver in form other than domestic coin convertible into domestic silver coin at, or approximately at, the mint price (the inverse of the silver content of the monetary unit), (*c*) no restrictions on private parties (i) melting domestic coin into bullion, or (ii) importing or exporting silver in any form, and (*d*) full legal-tender status for domestic silver coin.
Other forms of money may exist, but silver is the primary money. Foreign silver coin may be given

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
equal legal-tender status with domestic coin. Gold coin may be in circulation, but its value is in terms of the silver monetary unit and may fluctuate by weight, varying with the market gold–silver price ratio. Paper currency and deposits may exist, but, as liabilities of the issuer or bank, are payable in legal tender, that is, silver coin (or silver-convertible government or central-bank currency).

If silver (whether domestic or foreign coin, or both) constitutes the only money, then, even absent free coinage, the economy is clearly on a silver standard. This conclusion holds with gold coin circulating as well, providing it is circulating by weight or is a minor part of the money supply.

A silver standard might be effective even though the monetary system is legally bimetallic. If the coinage gold–silver price ratio is sufficiently below the market ratio, then gold, undervalued at the mint, will be sold on the world market (even in the form of melted domestic coin), while silver, overvalued, will be imported and coined. Ultimately, an effective silver standard may result.

Depreciation of the silver coinage involves an increased ratio of the legal (face) value of coins relative to silver content, usually by debasement (reducing the silver content, whether weight or fineness, of given-denomination coins) rather than by increasing the denomination of existing (given-weight-and-fineness) coins. In England, the penny (of sterling, 11/12th fineness) was steadily reduced in size from 24 grains in the eighth century to less than 1/3 that weight in 1601.

A silver standard, just as the gold standard, provides a constraint on the money stock. Depreciation of silver coinage was a way of escaping that constraint, even though the authority's objective typically was to increase government revenue (in the form of seigniorage) and/or to change the coinage ratio (under legal bimetallism).

Countries on silver standard to 1870

A silver standard first occurred in ancient Greece. Notwithstanding generally legal bimetallism, silver was everywhere the effective metallic standard – or at least the far-more-important coined metal in the money stock – well into the 18th century. Because of its relative scarcity and high density, gold was always much more valuable than silver on a per-ounce basis: coinage and market ratios were far above unity. So, with most transactions of low value compared with the unit of account, silver was better suited than gold to serve as a medium of exchange. In US history, ‘one dollar’ was both the smallest gold piece and the largest silver piece ever coined.

In England, from the Anglo-Saxon period until the late 13th century, the only coin in existence (with rare exceptions) was the silver penny, with 240 pence coined ideally from one pound of silver and later constituting one pound sterling (where ‘sterling,’ of course, denotes silver). This was a silver standard by default. With coinage of gold, in 1257, there was legal bimetallism; but the practice of denominating gold coins in (silver) shillings and pence was implicit recognition of an effective silver standard. Even the popular, consistently coined, (gold) guinea, first issued in 1663, was left to find its own market value in shillings and pence. However, by the turn of the 18th century, foreign gold–silver price ratios had been falling and, having been increased greatly in 1696, the British coinage ratio was not subsequently reduced enough to compensate. England went briefly on a bimetallic standard, and then on an effective gold standard, legalized in 1774 and 1816.

In the United States, since colonial times a silver standard was in effect, based on the Spanish dollar, the primary circulating silver coin, which varied much in weight and fineness. Yet the dollar was accepted everywhere at face value in terms of local (individual-state) pound–shilling–pence units of account.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Gold coins were rated in dollars according to fine-metal content. The Coinage Act of 1792 placed the United States on a legal bimetallic standard; but the coinage ratio soon fell below the (increasing) world-market ratio. An effective silver standard resulted, until the coinage ratio was corrected in 1834.
In 1870, just before Germany united and established the gold standard (using as financing the French indemnity, emanating from the Franco–Prussian War), Netherlands, Denmark, Norway, Sweden, India, China, Straits Settlements, Hong Kong, Dutch East Indies, Mexico and some German states were on a silver standard. In the 1870s these European countries (and Dutch East Indies) abandoned silver in favour of gold. By 1885 almost all of western Europe – along with the United States, Britain, its dominions and various colonies – was on gold.

Asian abandonment of the silver standard prior to the First World War

Traditionally, Asian countries preferred silver to gold for both monetary and non-monetary use, and the low market ratios in the Far East reflected that fact. The silver standard continued after 1885 in the Asian countries listed above. Further, in the 1880s the Philippines and Japan went on de facto silver.
Until 1873, bimetallic France kept the world market gold–silver price ratio around a narrow band centred on the French coinage ratio of $15\frac{1}{2}$. When France ended bimetallism in 1873, the market ratio lost its anchor and escalated tremendously. The exchange rates between silver-standard and gold-standard currencies also lost their anchor. Following the market gold–silver price ratio, silver currencies depreciated greatly with respect to gold currencies. Exports were enhanced, imports were more expensive, debt and other obligations stated in terms of gold or gold currencies increased greatly in domestic currency, domestic inflation increased, and foreign investment was discouraged due to exchange-rate instability.
The problem of a depreciating currency was especially acute for India, which had the obligation of substantial recurring sterling-denominated ‘home charges’ to Britain (for debt service, pensions, military and other equipment, and so forth). In 1893 India abandoned the silver standard, and in 1898 went on the gold-exchange standard, pegging the (silver) rupee against the pound sterling.
In 1897 Japan switched from a de facto silver standard under legal bimetallism to a monometallic gold-coin standard, using as financing the indemnity received from defeated China in the Sino–Japanese War.
In 1903 the Philippines adopted a gold-exchange standard, with the (silver) peso pegged to the US (gold) dollar. The impetus was transfer of the country from Spain to the United States, thanks to US victory in the Spanish–American War.
Mexico, a large silver producer, with both commodity exporters and silver producers in favour of a continued silver standard, finally adopted a gold-coin standard in 1905. At the beginning of the First World War, the silver standard encompassed only China, Hong Kong and a few minor countries.

Termination of the silver standard

The final blow to the silver standard was delivered by the United States, ironically after it left the gold standard. In December 1933, when the (fluctuating) market price of silver was 44 cents per ounce, President Roosevelt proclaimed that US mints should purchase all new domestically produced silver at a net price (to the depositor, or seller) of 64.65 cents per ounce (half the official, but inoperative, mint

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
price of silver). In 1934 this policy was reinforced by the Silver Purchase Act, which directed the Treasury to purchase silver at home and abroad as long as (*a*) the Treasury stock of gold constituted less than one-quarter its total monetary stock, and (*b*) the market price did not exceed the US official mint price. Subsequently, the president ordered that all silver (with minor exceptions) then situated in the continental United States was to be delivered to US mints, at a net price of 50.01 cents per ounce. In 1935, in response to a higher foreign market price of silver (largely due to the US silver-purchase policy itself!), the president increased the net price for newly produced domestic silver to 71.11 cents.

The reason for the US silver-purchase policy was to provide a subsidy to the (politically powerful) domestic silver producers. Inadvertently, the policy effectively destroyed what remained of the silver standard. The last major country on the silver standard was China. As the gold-standard world suffered monetary and real deflation in 1929–30, the price of silver fell. The Chinese, silver-based, currency (yuan) therefore depreciated against the, gold-based, currencies of important trading partners (Britain, India, Japan). The enhanced competitiveness of export and import-competing industries, and resulting balance-of-payments surplus, prevented deflation. China lost some ‘silver protection’ in 1931, after Britain, India and Japan left the gold standard, as the yuan appreciated against the pound, rupee and yen; but the United States was still on the gold standard, and the yuan continued to fall, slightly, against the dollar. After the United States abandoned the gold standard, in 1933, the yuan appreciated against all four currencies.

While China had lost its ‘silver protection’ from the world depression, it nevertheless retained the silver standard and probably suffered less economically than its main trading partners. Disaster struck with the US silver policy of 1933–4. The huge increase in the US and market price of silver involved a corresponding appreciation of the yuan. Loss of competitiveness, balance-of-payments deficit, export of silver (and gold) to finance the deficit, and deflation followed. China had no choice but to leave the silver standard, effectively in 1934, and legally in 1935.

Other silver-standard, as well as silver-using, countries were also adversely affected by the US policy. Hong Kong followed China, and left the gold standard in 1935. Though not on the silver standard, various Latin American countries had a large silver coinage. These were token coins (face value higher than metallic-content value). Nevertheless, the high US price for silver encouraged the melting and export of these coins. The affected countries resorted to debasement and re-coining in order to retain their silver coinage.

Mexico was a special case. Silver coins constituted a high proportion of its money supply; but, as the world's largest producer of silver, Mexico benefited from a higher price for a major export. However, as other countries left the silver standard, the price of silver began to fall, and this advantage was reduced. Mexico prohibited melting or export of silver coins in 1935, and replaced the coins with paper money. Later, re-coinage occurred, and melting and export were again permitted. Yet the damage had been done, and Mexico was now on a ‘managed paper standard’, having lost the discipline provided by metallic money. In sum, in the 1930s, a US domestic-oriented policy reduced considerably such monetary use of silver as remained.

See Also

- bimetallism

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- commodity money
- gold standard

Bibliography

Bojanic, A.N. 1994. The silver standard in two late adherents: Mexico and India. Ph.D. thesis, Auburn University.

Brandt, L. and Sargent, T.J. 1989. Interpreting new evidence about China and U.S. silver purchases. *Journal of Monetary Economics* 23, 31–51.

der Eng, P.V. 1993. *The Silver Standard and Asia's Integration into the World Economy, 1850–1914*. Canberra: Australian National University.

Eichengreen, B. and Flandreau, M. 1996. The geography of the gold standard. In *Currency Convertibility: The Gold Standard and Beyond*, ed. J.B. de Macedo, B. Eichengreen and J. Reis. London and New York: Routledge.

Einzig, P. 1970. *The History of Foreign Exchange*, 2nd edn. London: Macmillan.

Feavearyear, Sir A. 1963. *The Pound Sterling: A History of English Money*, 2nd edn. Oxford: Clarendon Press.

Friedman, M. 1992. *Money Mischief: Episodes in Monetary History*. New York: Harcourt Brace Jovanovich.

Lai, C. and Gau, J.J. 2003. The Chinese silver standard economy and the 1929 Great Depression. *Australian Economic History Review* 43, 155–68.

Leavens, D.H. *Silver Money*. Bloomington, IN: Principia Press.

Officer, L.H. 1996. *Between the Dollar–Sterling Gold Points*. Cambridge: Cambridge University Press.

Redish, A. 2000. *Bimetallism: An Economic and Historical Analysis*. Cambridge: Cambridge University Press.

Wilson, T. 2000. *Battles for the Standard: Bimetallism and the Spread of the Gold Standard in the Nineteenth Century*. Aldershot: Ashgate.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Officer, Lawrence H. "silver standard." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000205> doi:10.1057/9780230226203.1527

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Simon, Herbert A. (1916 – 2001)

Mie Augier
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses how Simon's vision for behavioural economics (and social science generally) was found in the context of his early work in public administration and political science. It was strengthened as Simon proceeded to make contributions to economics, and found an intellectual home with the establishment of behavioural science in the 1950s.

Keywords

administrative theory; artificial intelligence; behavioural economics; bounded rationality; Carnap, R.; choice; computational capability; computer science; Cowles Commission; decision theory; econometrics; employment contracts; operations research; Organization theory; payoff function; perfect information; satisficing; Simon, H.

Article

Herbert A. Simon was born on 15 June 1916 in Milwaukee, Wisconsin, USA. He received his Ph.D. in political science from the University of Chicago in 1943, and taught at the Illinois Institute of Technology from 1942 to 1949 before going to Carnegie Mellon University in 1949, where he stayed until he died on 9 February 2001. Simon received major awards from many scientific communities, including the A.M. Turing Award (in 1975), the National Medal of Science (in 1986), and the Nobel Prize in Economics (in 1978). During his career, Simon also served on the Committee on Science and Public Policy and as a member of the President's Science Advisory Committee.
Simon made important contributions to economics, psychology, political science, sociology, administrative theory, public administration, organization theory, cognitive science, computer science and philosophy. His best known books include *Administrative Behavior* (1947), *Organizations* (1958, with James G. March), *The Sciences of the Artificial* (1969), *Human Problem Solving* (1972, with Allen Newell), and his autobiography, *Models of My Life* (1991). Although contributing to so many seemingly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
different domains and traditions, Simon's main research interest remained the same: understanding human decision making.

Early life

Simon spent his early years with his parents and his older brother on the West Side of Milwaukee in a middle-class neighbourhood. Attending public schools, Simon at first intended to study biology. However, after he went on a strawberry hunting trip, and discovered that he was colour-blind (he was unable to distinguish the strawberries from the plants), he changed his mind, thinking that colour blindness would be too big a handicap in biology. He then thought briefly about studying physics, but he gave up that idea after discovering that there weren't really any major advances left to be made in physics. ‘They have all these great laws’, he said in conversation. ‘Newton had done it, no use messing around with it.’ As a result, upon finishing high school in 1933 Simon enrolled instead at the University of Chicago, with an interest in making social science more mathematical, and an intention to major in economics. In keeping with his strong wish to be independent, Simon preferred reading alone to taking classes; and as he particularly refused to take the class in accounting, which was required for graduation in economics, he majored instead in political science.

Political science wasn't physics, of course, with all its ‘great laws’. However, as a science it could encompass both theory and practice; and, being an empirical science, it had to take the data seriously. Furthermore, Simon found he was attracted to interdisciplinary thinking (in particularly psychology) in understanding political behaviour. The details of Simon's mature work differ, but the underlying ideas, interdisciplinary thinking and the necessity of bringing together theory and reality remain. Also present from the start was the essential idea of limited rationality, which would stay with Simon as he proceeded to translate his insights in political science and public administration into his work in economics, organization theory, psychology and artificial intelligence.

Early on Simon was invited by Clarence Ridley to participate as a research assistant in a project for the International City Managers’ Association (Simon, 1991, p. 64). Together with Ridley, Simon published the results of this project in several articles as well as a book, *Measuring Municipal Activities* (1938). This brought an invitation to join the University of California's Bureau of Public Administration in order to study local government. While working in Berkeley on directing a study of the administration of state relief programmes, intended to demonstrate how quantitative empirical research could contribute to understanding and improving municipal government problems (1991, p. 82), Simon was also working on an early manuscript of his thesis, which became *Administrative Behavior* (1947), intended to reforming administrative theory. The first working title of *Administrative Behavior* was ‘The Logical Structure of an Administrative Science’. Simon had intended the book to have a heavy philosophical component, in particular reflecting the influence of Rudolph Carnap. Furthermore, Simon introduced the importance of organizations to individual decision making; a theme later elaborated especially in *Organizations* (1958). ‘Human rationality’, he wrote, ‘gets its higher goals and integrations from the institutional settings in which it operates and by which it is molded. … [Therefore] … [t]he rational individual is, and must be, an organized and institutionalized individual’ (1947, pp. 101–2). Simon argued that organizations make it possible to make decisions by constraining the set of alternatives to be taken into account and the considerations that are to be treated as relevant. Organizations can be improved by improving the ways in which those limits are defined and imposed. Finally, *Administrative*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Behavior criticized existing administrative theory for being based on ‘proverbs’ (often contradictory common-sense principles), a perspective he wanted to replace with a more empirically oriented outlook investigating the nature of the decision processes in administrative organizations.

It was in *Administrative Behavior* that Simon first systematically examined the importance of limits to human rationality. ‘The dissertation contains both the foundation and much of the superstructure of the theory of bounded rationality that has been my lodestar for nearly fifty years’ (1991, p. 86). The core chapters of this book were intended to develop a theory of human decision making which was broad and realistic enough to accommodate both ‘those rational aspects of choice that have been the principal concern of the economist, and those properties and limitations of the human decision making mechanisms that have attracted the attention of psychologists and practical decision makers’ (1947, p. xi). Bringing together insights from economics and psychology, Simon laid the foundation for the later establishment of behavioural economics and for organization theory. In Simon's view, the significance of his early work was in replacing ‘economic man’ with ‘administrative man’ by bringing insights from psychology to bear on studying decision-making processes (1947, p. xxv).

While finishing his dissertation Simon moved to Illinois Institute of Technology, in an environment in Chicago in the early 1940s where most of his fellow researchers were believers in rational decision making, Simon remained a strong advocate of the idea of limited rationality. He began to discuss his ideas with prominent economists, in particularly those connected to the Cowles Commission, a group of mathematical economists doing pioneering research in econometrics, linear and dynamic programming, and decision theory, among other things. The economists connected to the Cowles Commission included such well-known names as Kenneth Arrow, Jacob Marshak, Tjalling Koopmans, Roy Radner and Gerard Debreu, and they held regular seminars to discuss their research. During the last years of Simon's stay in Chicago he began attending the Cowles Commission seminars; this became very important to Simon both because, as he noted in his autobiography, his interaction with Cowles almost made him ‘a full time economist’ (1991, p. 140) and because several members of the Cowles Commission would become good friends. Furthermore, Simon seems to have realized that among economists the possibilities for exploring the limits of rationality were themselves limited, for only later did he proceed to construct the broad behavioural programme upon which the foundation of the psychology of decision making in behavioural science could rest. ‘In none of [the] early papers’, wrote Simon, ‘did I challenge the foundations of economic theory strongly’ (1991, p. 270).

In 1949 Simon moved to Pittsburgh to join the newly established School of Industrial Administration at Carnegie Mellon University, an early engineering school trying to become a business school. Business education at that time wasn't oriented mostly towards research, but Simon and colleagues wanted to be different: they wanted to do research. They wanted their research to be relevant for business leaders, while at the same time emphasizing the tools of good science (Cooper, 2002). Early core courses in the programme included ‘quantitative control and business’ (consisting of basically accounting and statistics) taught by Bill Cooper, a sequence of micro and macroeconomics taught by Lee Bach, and organization theory taught by Simon. As a result of their early efforts to build up a research programme at Carnegie Mellon GSIA was picked by the Ford Foundation as one of the foremost places where the new science of behavioural economics could be developed. It became a pioneer in the establishment of business education in the United States, and must be seen as part of the Simon legacy, perhaps as important as his direct intellectual contributions.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Later work and career

Decision making was also the core of Simon's later work, and it became the basis of his other contributions to organization theory, economics, psychology, and computer science. Decision making, as Simon saw it, is purposeful yet not rational, because rational decision making would involve a complete specification of all possible outcomes conditional on possible actions in order to choose the single option that is best. In challenging neoclassical economics, Simon found that such complex calculation is not possible. As a result, Simon wanted to replace the assumption of global rationality in economics with an assumption that was more in correspondence with how humans make decisions, their computational limitations, and how they accessed information in their current environment (1955), thereby introducing the ideas of bounded rationality and satisficing. Satisficing is the idea that decision makers interpret outcomes as either satisfactory or unsatisfactory, with an aspiration level constituting the boundary between the two. Whereas decision makers in neoclassical rational choice theory would list all possible outcomes evaluated in terms of their expected utilities, and then chose the one that is rational and maximizes utility, decision makers in Simon's model face only two possible outcomes, and look for a satisfying solution, continuing to search only until they have found a solution which is good enough. The ideas of bounded rationality and satisficing became important for the subsequent development of economics.

The emphasis on bounded rationality introduced a more psychological and realistic assumption into the analysis. As Simon noted early on:

> [T]he first principle of bounded rationality is that the intended rationality of an actor requires him to construct a simplified model of the real situation in order to deal with it. He behaves rationally with respect to this model, and such behavior is not even approximately optimal with respect to the real world. To predict his behavior, we must understand the way in which this simplified model is constructed, and its construction will certainly be related to his psychological properties as a perceiving, thinking, and learning animal. (1957, p. 199)

Both satisficing and bounded rationality were introduced in 1955, when Herbert Simon published a paper that provided the foundation for a behavioural perspective on human decision making and introduced the ideas of satisficing and bounded rationality. The paper provided a critique of the assumption in economics of perfect information and unlimited computational capability, and replaced the assumption of global rationality with one that was more in correspondence with how humans (and other choosing organisms) made decisions, their computational limitations and how they accessed information in their current environments (1955, p. 99). In Simon's illustration of the problem, the influence of his early ideas outlined in *Administrative Behavior* is clear, echoing the view that decisions are reasoned and intendedly rational yet limited (1947). He first suggested a simple and very general model of behavioural choice which analyses choosing organisms (such as humans) in terms of basic properties to understand what is meant by rational behaviour. He introduced the simplifying assumptions (such as the choice alternatives, the payoff function, possible future states and the subset of choice alternatives which is considered, as well as the information about the probability that a particular outcome will lead to a particular choice) (1955, p. 102). But immediately afterwards he turned to the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
simplifications of this model, stressing that upon careful examination ‘we see immediately what severe demands they make upon the choosing organism’. Whereas in models of rational choice the organism must be able to ‘attach definite payoffs (or at least a definite range of payoffs) to each possible outcome’, Simon suggested that ‘there is a complete lack of evidence that, in actual human choice situations of any complexity, these computations can be, or are in fact, performed’ (1955, p. 103). As a consequence of the lack of computational power, decision makers have to simplify the structure of their decisions (and thus satisfice), one of the most important lessons of bounded rationality.

In a companion paper ‘Rational Choice and the Structure of the Environment’ (1956), Simon introduced the idea that the environment influences decision making as much as do information processing abilities. He examined the influence of the structural environment on the problem of ‘behaving approximately rationally, or adaptively’ in particular environments (1956, p. 130). Simon would later elaborate these ideas in his book, *Sciences of the Artificial*, using the famous ‘ant on the beach’ metaphor to illustrate his idea (1969, pp. 51–3). The ant makes its way from one point to another, using a complex path, the complexity consisting of the patterns of the grains of sand along the way rather than internal constraints. Just so with human beings, Simon argues: ‘Human beings, viewed as behaving systems, are quite simple. The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves’ (1969, p. 53).

Another early important paper (1951) concerned the nature of the employment relation. The paper began by emphasizing the traditional Simon view that models ought to correspond to the empirical realities that are neglected in most economic models of the employment contract (1951, p. 293). He then turned to a concept that was so central to him in *Administrative Behavior*, namely, the concept of authority. Central to the employment relation, Simon said, is the fact that the employer accepts a certain amount of authority of the employee for which he pays a wage and the employee accepts this authority within certain ‘areas of acceptance’ (1951, p. 294). His model applies the idea of satisfaction functions to the employment problem, yet it is still ripe for extension because it still is ‘highly abstract and oversimplified, and leaves out of account numerous important aspects of the real situation’ (1951, p. 302). The model appears to be considerably more realistic in the way it conceptualizes the nature of the employment relationship; yet it is still about ‘hypothetically rational behavior in an area where institutional history and other nonrational elements are notoriously important’ (1951, p. 302). The model suggests a way to reconcile administrative theory and economics through the economic nature of the employment relation; yet it is still limited by its ‘assumptions of rational utility-maximization behavior incorporated in it’ (1951, p. 305). Thus, Simon used the framework of economics (however limited it might be) to discuss an issue he had been interested in since his thesis, and he concluded his analysis by pointing out the limitations of a constrained model and the necessity of accounting also for non-rational elements.

Simon used this behavioural view of decision making to create a propositional inventory of organization theory, together with James March and Harold Guetzkow, which led to the book *Organizations* (1958). The book was intended to provide the inventory of knowledge of the (then almost non-existent) field of organization theory, and also a more proactive role in defining the field. Results and insights from studies of organizations in political science, sociology, economics and social psychology were summarized and codified. The book expanded and elaborated ideas on behavioural decision making, search and aspiration levels, and the significance of organizations as social institutions. ‘The basic features of organization structure and function’, March and Simon wrote, ‘derive from the characteristics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of rational human choice. Because of the limits of human intellective capacities in comparison with the complexities of the problems that individuals and organizations face, rational behavior calls for simplified models that capture the main features of a problem without capturing al its complexities' (1958, p. 151). The book is now considered one of the classics and pioneering works in organization theory.

While Simon opposed some major developments in rational choice economics, he found value in the emerging field of operations research. Although Simon's marriage with operations research was neither entirely happy nor permanent, the fact that operations research was well suited to crossing disciplinary boundaries immediately appealed to him, in addition to its appeal in using computers for heuristic programming. Thus, Simon and Newell wrote (1958, p. 6):

> Even while operations research is solving well-structured problems, fundamental research is dissolving the mystery of how humans solve ill-structured problems. Moreover, we have begun to learn how to use computers to solve these problems, where we do not have systematic and efficient computational algorithms. And we now know, at least in a limited area, not only how to program computers to perform such problem-solving activities successfully; we know also how to program computers to learn to do these things.

Although most of the techniques used in operations research are techniques of constrained maximization, Simon found that they 'formed a natural continuity with my administrative measurement research (1991, p. 108). He found artificial intelligence to be the logical next step in operations research, something which would eventually bring Simon's insights into behavioural economics and organization theory to bear on management science, using empirical studies in decision making in organizations, constructing a mathematical model of the process under study, and then simulating it on a computer (1965).

Simon's interest in operations research is also evident in his work on the design of optimal production schedules, something which ultimately led to the book *Planning, Production, Inventories, and Work Force* (1960). Although initiated at the Cowles Commission, this worked was carried out at Carnegie Mellon University, which provided the context for most of Simon's academic life. It was also at Carnegie that it became clear that Simon was not 'just' another economist. Highly respected among most (if not all) distinguished economists of his time (see, for instance, Samuelson, 2004; Radner, 2004), Simon himself was much more than an economist. For instance, at Carnegie he quickly retooled himself as an organization theorist in order to carry out, with James G. March, a major Ford Foundation study on theories of decision making in organizations. Most important, at Carnegie Simon found both colleagues and an environment which could accommodate and appreciate his broad interests and honour his willingness to cross disciplinary boundaries in pursuing his vision. With the emerging emphasis on behavioural science at Carnegie came many contributions of a cross-disciplinary and interdisciplinary nature. The disciplinary boundary crossing that had previously been, if not difficult, then different from the mainstream became possible and more widespread with the behavioural research focus that Simon helped establish at Carnegie. Having found during his years at Chicago the limits of standard economic theory in dealing with limits to rationality, he turned his attention towards founding a research programme in behavioural economics to accommodate his vision.

Simon thus incorporated his early views on decision making and rationality into his contributions to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
psychology, computer science and artificial intelligence. For example, in his work with Allen Newell he attempted to develop a general theory of human problem solving which conceptualized both humans and computers as symbolic information-processing systems (1972). Their theory was built around the concept of an information processing system, defined by the existence of symbols, elements of which are connected by relations into structures of symbols. The book became as influential in cognitive science and artificial intelligence as Simon's earlier work had been in economics and organization theory. During his amazingly productive intellectual life, Simon worked on many, sometimes different, things; yet he pursued really one vision (Augier and March, 2002). He contributed significantly to many scientific disciplines, yet found scientific boundaries themselves to be less important, even unimportant, vis-à-vis solving the questions he was working on. Even as Simon sought to develop the idea that you could simulate the psychological process of thinking, he tied his interest in economics and decision making closely to computer science and psychology. He used computer science to model human problem solving in a way that was consistent with his approach to rationality. He implemented his early ideas of bounded rationality and means–ends analysis into the heart of his work on artificial intelligence.

See Also

- operations research

Selected works

1938. (With C. Ridley.) *Measuring Municipal Activities*. Chicago: International City Managers' Association.

1947. *Administrative Behavior*. New York: Free Press.

1951. A formal theory of the employment relationship. *Econometrica* 19, 293–305.

1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69, 99–118.

1956. Rational choice and the structure of the environment. *Psychological Review* 63, 129–38.

1957. *Models of Man*. New York: Wiley.

1958. (With A. Newell.) Heuristic problem solving: the next advance in operations research. *Operations Research* 6, 1–10.

1958. (With J. March.) *Organizations*. New York: Wiley.

1960. (With C. Holt, F. Modiliani and J. Muth.) *Planning, Production, Inventories, and Work Force*. Englewood Cliffs: Prentice Hall.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1963. Discussion: problems of methodology. *American Economic Review* 53, 229–31.

1965. *The Shape of Automation (for Men and Management)*. New York: Harper and Row.

1969. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

1972. (With A. Newell.) *Human Problem Solving*. Englewood Cliffs: Prentice-Hall.

1991. *Models of My Life*. Cambridge, MA: MIT Press.

Bibliography

Augier, M. and March, J. 2002. A model scholar. *Journal of Economic Behavior and Organization* 49, 1–17.

Cooper, W. 2002. Auditing and accounting: impacts and aftermaths of R.M. Cyert's research in statistical sampling. In *The Economics of Choice, Change and Organization: Essays in Honor of Richard M. Cyert*, ed. M. Augier and J. March. Cheltenham, UK: Edward Elgar.

Mirowski, P. 2002. *Machine Dreams*. Cambridge: Cambridge University Press.

Radner, R. 2004. The best is the enemy of the good. In *Models of a Man: Essays in Memory of Herbert A. Simon*, ed. M. Augier and J. March. Cambridge, MA: MIT Press.

Samuelson, P. 2004. The Hawkins–Simon theorem revisited. In *Models of a Man: Essays in Memory of Herbert A. Simon*, ed. M. Augier and J. March. Cambridge, MA: MIT Press.

How to cite this article

Augier, Mie. "Simon, Herbert A. (1916–2001)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000455> doi:10.1057/9780230226203.1528

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Simons, Henry Calvert (1899 – 1946)

Herbert Stein
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chicago School; equality; Haig–Simons definition of income; liberty; monetary policy rules; monopoly; price level; progressive taxation; Simons, H. C.; trade unions

Article

Simons was born in Virden, Illinois, and died in Chicago. An economist at the University of Chicago from 1927 to 1946, he was the first professor of economics at the University of Chicago Law School. A leader of the ‘Chicago School’, he had an important influence on American thinking about economic policy.

Simons's central theme was stated in the title of his first writing to attract attention, a 1934 pamphlet: ‘A Positive Program for Laissez Faire: Some Proposals for a Liberal Economic Policy’. The conjunction of the words ‘positive’ and ‘laissez faire’ set him apart from both the conventional conservatives of his time and the conventional liberals (in the American sense of interventionists). Simons visualized a division of labour between the government and the market. The market would determine what gets produced, how and for whom. The government would be responsible for maintaining overall stability, for keeping the market competitive and for avoiding extremes in the distribution of income. This system would preserve liberty by preventing concentration of power, and liberty is the primary virtue, followed closely by equality.

Simons's work was the response of a free society liberal – or, as he preferred, ‘libertarian’ – to the rise of totalitarianism in Europe, to the worldwide depression and to the attempt in the democracies, including the United States, to cope with the depression in ways that Simons regarded as threats to freedom. Simons's close friend, Professor Aaron Director, later said that Simons acted as if the end of the world was at hand. During the period of Simons's work, if not the end of the world at least the end of the free society could realistically be considered a serious possibility. Simons undertook to help to prevent that, by showing that the free society had not failed but that the government had failed to discharge its role in the free society.

The 1934 pamphlet contained the elements of a policy for a free economy that he was to restate and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
refine for the next 12 years, with some changes of emphasis. As he put it in 1934:

> The main elements in a sound liberal program may be defined in terms of five proposals or objectives (in a descending scale of relative importance):
> I. Elimination of private monopoly in all its forms•…•.
> II. Establishment of more definite and adequate ‘rules of the game’ with respect to money … .
> III. Drastic change in our whole tax system, with regard primarily for the effects of taxation upon the distribution of wealth and income•…•.
> IV. Gradual withdrawal of the enormous differential subsidies implicit in our present tariff system•…•.
> V. Limitation upon the squandering of our resources in advertising and selling activities. (1934, p. 57)

In later years the fifth of these items fell from the list.
The first proposal, ‘elimination of private monopoly in all its forms’, was substantially altered later. In 1934 Simons had said, ‘The case for a liberal-conservative policy must stand or fall on the first proposal, abolition of private monopoly; for it is the *sine qua non* of any such policy’ (p. 57). His measures for achieving this included limitation on the absolute size of corporations and on their relative size in their industries. He suggested, for example, that ‘in major industries no ownership unit should produce or control more than 5 per cent of the total output’ (p. 319).
By 1945 he was saying ‘Industrial monopolies are not yet a serious evil’ (1945, p. 35). Simons's concern about private monopoly had always been about its interaction with the state. He feared that government would support private monopolies and then have to become more powerful to control the warring monopolies it had created. The 1934 pamphlet was written at the time of Roosevelt's National Recovery Administration, which was promoting the universal cartelization of business under government aegis. But in 1945 all that was past and the political influence of business seemed too small to be a danger.
In 1945 what he had said about business monopoly he now said about labour unions. In 1934 Simons had expressed concern about labour monopolies, but in a rather subdued way. In the decade after the 1934 pamphlet, labour union membership quadrupled, and this growth showed no sign of diminishing. In his final credo (1945) Simons said ‘the hard monopoly problem is labour organization’. For this problem he could offer no ‘specific’, only a rather uncertain question about ‘the capability of democracy to protect the common interest’ (1945, pp. 35–6).
As the Second World War drew to an end, the preservation of free international trade received more of Simons's attention. Peace was essential for all the goals he cherished. Even the fear of war would require a centralization of power in government that would be incompatible with personal freedom. Simons believed that economic nationalism would be the greatest threat to peace after the Second World War. Therefore, he devoted much of his work in the mid-1940s to arguing for a liberal international economic order.
While the emphasis of some points in Simons's initial policy agenda shifted, two items remained of major importance and constituted Simons's chief contribution aside from the general idea of conjoining ‘positive program’ with ‘laissez faire’. These were the need for monetary certainty and stability and the need to finance government primarily by progressive taxation of ‘income’ defined in a comprehensive

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
way.

His analysis of the monetary problem and proposals for its solution, already outlined in the 1934 pamphlet, were elaborated in a 1936 essay whose title defined the issue for years to come, ‘Rules versus Authorities in Monetary Policy’ . Simons believed that economic instability was due largely to the instability of the financial system. The system rested excessively on private debt, mainly short-term debt. Variations in the quantity and quality of this debt caused destabilizing variations in the quantity of money, in the quantity of ‘near-money’, and thereby of velocity, and in the financial requirements of business. The monetary authority, the central bank, was unreliable in discharging its responsibility to counter these devastating tendencies.

Simons's remedy for this condition was a radical reform of the financial structure and the establishment of a rule to govern the conduct of monetary policy. He regarded as an ‘approximately ideal solution’ one in which all property was held in equity form. Failing that, he would have preferred that all debt be in the form of perpetuities, or at least of very long maturities. He did not, however, expect to achieve even that much. But he was specific in recommending the insulation of the banking system and government finance from the malignancy of short-term debt. Banks would be required to hold reserves in currency and Federal Reserve deposits against 100 per cent of their deposits. The government would have only two kinds of debt: currency and consols.

This arrangement would give the government effective control of the quantity of money, a control that it would exercise by fiscal means – by altering the size of its own debt or the division of the debt between currency and controls. This control would be exercised ‘under simple, definite rules laid down in legislation’, to provide the private sector with the maximum certainty.

Simons wrestled continuously with the question of what the rules should be. His indecision appeared at the beginning, in 1934, when he referred to controlling ‘the quantity, (or through quantity, the value) of effective money’ (1934, p. 57). He debated with himself on this issue in the ‘Rules versus Authorities’ and elsewhere. He recognized that a rule aimed at the price level (or the value of money) would necessarily leave the authority with discretion to decide what quantity of money would achieve the goal. But he also feared that with the existing financial situation the velocity of money would be so variable that a quantity rule would yield great price-level instability. His solution to this dilemma was to opt for the price-level rule until reform of the financial system would reduce the quantity of near moneys and the instability of the debt structure, after which stabilizing the quantity of money would be the preferred rule.

Simons's only two books were on taxation. The first, *Personal Income Taxation* was his doctoral dissertation, written in the early 1930s and published in 1938. The second, *Federal Tax Reform*, was commissioned by the Committee for Economic Development, an organization of businessmen, mainly written in 1943 and published posthumously in 1950. A few main elements ran through all of his work on taxation. The nearly exclusive source of revenue should be taxation of personal income, meaning what has come to be called the Haig–Simons definition of income as the sum of the value of the taxpayers’ consumption plus the addition to his net assets. (The reference is to Robert M. Haig, ‘The Concept of Income’, in *The Federal Income Tax*, ed. R.M. Haig, New York, 1921.) This definition should be applied as comprehensively as possible, for the sake of equity and economic efficiency. Simons fully explored the implication of that for the treatment of capital gains, gifts, income in kind and corporate profits. Finally, he emphasized the use of the progressive income tax as a means of reducing inequality both because reducing inequality was important and because the progressive income tax was a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
way of reducing inequality that was much more compatible with a free economy than other measures commonly proposed for that purpose.

Simons was a leading member of what became known in the 1930s as the ‘Chicago School’. Other members at the time were Frank H. Knight, Lloyd W. Mints and Aaron Director; Jacob Viner shared many of their views but did not consider himself a member. Simons more than the others translated their general attitudes into specific policy proposals, which he advanced forcefully in his own writing and defended in a series of strong reviews of the writings of the opposition.

Simons's great attraction for his colleagues, students and sympathetic readers was a matter of personal and literary style as well as of substance. His writing was polished, ironical, free of technical jargon, statistics or mathematics, rising above ‘mere’ economic analysis to grand pronouncements on eternal subjects. It was not very difficult but difficult enough to leave the reader with a sense of accomplishment at having recognized its merits. He gave his readers and students a feeling of being initiated into a select club that had great insights that politicians, businessmen and most economists were intellectually, morally and ethically incapable of appreciating.

After the Second World War, and after his death in 1946, national discussion and, to some extent, policy turned in Simons's direction. There was no possibility of reverting to the negative conservatism of the pre-war years. But with a greatly enlarged federal budget and debt, and with the experience of inflation, the naive expansionism of Keynes's American disciples was no longer an acceptable policy. In this gap, Simons's ideas filled a need. A ‘modern conservativism’ emerged that accepted government responsibility for overall economic stability, was strongly anti-inflationary, sought a rule to govern stabilization policy, relied on tax changes rather than expenditure changes when positive fiscal measures were needed, opposed protectionism, sought to weaken the power of labour unions and accepted the progressive personal income tax as the main source of federal revenue. Simons's work contributed to this development. By the 1950s some of his principal concepts had become common currency in policy discussion – the combination of positive measures with laissez-faire, the rules-versus-authority issue and the Haig–Simons definition of the tax base. Many of his colleagues and students came into positions from which they could influence public opinion and policy.

By 1960 a new-generation Chicago School had come into prominence. Typified by Milton Friedman and George Stigler, they had been profoundly influenced by Simons as students but were departing substantially from his policy positions. Monetary history convinced them that Simons was wrong in opting for a price-level rule rather than a quantity-of-money rule for monetary policy. They concluded that antitrust activity, on which he had once placed so much emphasis, was on the whole destructive of competition. Whereas Simons never contemplated a peacetime federal budget exceeding 10 per cent of the national income, they were living with one exceeding 20 per cent, and that changed their views of many things. They came to doubt Simons's reliance on rational discussion as a way to improve government policy in a democracy; this led them, in the case of Friedman, to a search for constitutional amendments that would limit the political process or, in the case of Stigler, to concentrating on explaining rather than influencing the process. But still, they all retained the Simons vision of the good free society with a division of responsibility between the government and the market, and through them his voice was still heard 40 years after his death.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Chicago School
- taxation of income

Selected works

1934. A positive program for laissez-faire: some proposals for a liberal economic policy. First published as Public Policy Pamphlet No. 15, ed. H.D. Gideonse. Chicago: University of Chicago Press. Reprinted in Simons (1948).

1936. Rules versus authorities in monetary policy. *Journal of Political Economy* 54, 1–30. Reprinted in Simons (1948).

1938. *Personal Income Taxation*. Chicago: University of Chicago Press.

1945. Introduction: a political credo. First published in Simons (1948).

1948. *Economic Policy for a Free Society*. Chicago: University of Chicago Press. Contains 12 previously published and one previously unpublished article, a prefatory note by A. Director and a complete bibliography of Simons's writings.

1950. *Federal Tax Reform*. Chicago: University of Chicago Press. Contains a prefatory note by A. Director.

How to cite this article

Stein, Herbert. "Simons, Henry Calvert (1899–1946)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000133> doi:10.1057/9780230226203.1529

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Simonsen, Mario Henrique (1935 – 1997)

Mauro Boianovsky
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Brazilian economic theorist and policymaker Mario Henrique Simonsen elaborated in the early 1970s a theory of inertial inflation based on the indexation of economic contracts. He later established that lagged indexation under rational expectations brings about a trade-off between inflation and unemployment. Simonsen further argued that incomes policy may be deployed to speed up the convergence to Nash equilibria. Simonsen also showed that the condition for a country's solvency is that the rate of growth of its exports exceeds the rate of interest, advanced a model of bargaining between banks and indebted developing countries, and formulated a cash-in-advance model.

Keywords

bargaining; cash-in-advance models; Dornbusch, R.; dual economies; Friedman, M.; game theory; incomes policies; inertial inflation; inflation; inflationary expectations; inflation–unemployment trade-off; Kuhn–Tucker theorem; money-in-utility function; nonlinear programming; price indexation; risk aversion; Simonsen curve; Simonsen, M. H.; sovereign debt; staggered wages; structural unemployment; wage indexation; wage rigidities

Article

Simonsen was born on 19 February 1935 in Rio de Janeiro, and died on 9 February 1997 in the same city. A talented mathematical economist, Simonsen played a prominent role in the development of economic research – especially the nature of chronic inflationary processes and their effects on the economic system – and the formulation of economic policy in Brazil from the mid-1960s to the mid-1990s. Simonsen graduated in civil engineering (Universidade do Brasil, 1957) and economics (Faculdade de Economia e Finanças do Rio de Janeiro, 1963), and in 1973 received his doctorate from the School of Graduate Studies in Economics (EPGE) at Getulio Vargas Foundation (FGV, Rio) with a thesis about inflation in Brazil, published a few years earlier (Simonsen, 1970). However, he was

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

essentially a self-taught man as far as economics is concerned. Indeed, he started teaching economics already in 1961, and in 1966 became director of EPGE, the first school of its kind in Brazil, founded a year earlier. He left Rio for Brasilia in 1974, when he assumed office as minister of finance (Fishlow, 1988; Simonsen, 1988d). He resumed his activities as director of EPGE in 1979, a position he held until 1993. From 1979 to 1995 Simonsen was also an outside member of the board of Citicorp. In 1981 he co-organized, with Rudiger Dornbusch, an international conference on indexation held in Rio; the conference volume is today a key reference in the field.

Cash-in-advance model

Simonsen wrote textbooks on microeconomics (1967–69) and macroeconomics (1974a; 1983a). One of the hallmarks of his microeconomics textbook was the application of the Kuhn–Tucker theorem to several problems in consumer and production theories. This is noteworthy in his early elaboration of what we now call cash-in-advance models of demand for money (Clower, 1967). Simonsen (1964; 1967–69) explicitly introduced the cash-in-advance constraint as an inequality in a nonlinear programming problem and provided a diagrammatic illustration of the interior and boundary solutions, which correspond to the money-in-utility function and cash-in-advance cases respectively. His micro book included application of general equilibrium analysis to discuss structural unemployment caused by institutional wage rigidities or by fixed coefficients in underdeveloped dual economies, and the argument that risk aversion provides a solution to the problem of the optimal size of the firm under perfect competition and constant returns.

Inertial inflation

Simonsen's 1974 macro textbook restated his 1970 inflation model. The model is remarkable for introducing into the literature the concept of ‘inertial inflation’. As pointed out by Simonsen (1970), the dependence of the current rate of inflation on its past values means that cold turkey strategies of disinflation are costly. The inertial element – called the ‘feedback component’ – together with the ‘autonomous component’ (supply shocks) and the ‘demand regulation component’ (excess aggregate demand) decide the inflation rate in a given period of time according to a linear formula. The feedback component could be explained at the time either by contract indexation or by adaptive expectations of price changes. Simonsen chose the indexation assumption, because it reflected the institutional wage-setting mechanism of the Brazilian economy in the 1970s. Simonsen's 1970 feedback model differed from the accelerationist Phillips curve by explaining inflation acceleration as a result not of revised expectations but of a reduction in the adjustment interval, captured by changes in the feedback coefficient. Moreover, the model implied that, even if inflation expectations fell to zero, the feedback inertial mechanism would keep working due to wage staggering in the indexation process. On the assumption of zero excess aggregate demand and a less than unity feedback coefficient, the lower limit to the current rate of inflation was given by the autonomous component divided by 1•minus the feedback coefficient. In order to illustrate the argument, Simonsen distinguished between the peak and the average real wage concepts in a sawtooth pattern curve (called the ‘Simonsen curve’ in the Brazilian literature) with the real wage rate as a function of time for a given adjustment interval.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Indexation and incomes policy

Simonsen came back to indexation and wage staggering in the 1980s, when those topics became fashionable after Gray (1976) and Taylor (1979). In both his graduate macro textbook (1983a) and in his conference paper (1983b), he showed that an expanded Gray model under rational expectations and with perfect wage indexation – that is, with individual prices indexed to the current price index – supported Milton Friedman's argument in favour of escalator clauses as a way to ease the side effects of anti-inflationary policies. However, he also showed that, under the more realistic assumption of lagged indexation – with prices indexed to changes in the price index over some past interval – the result would be the opposite: the inflation rate predicted for the current period and later depends upon the inflation rate in the past period (which, by definition, cannot reflect new information about current or future monetary policy). Simonsen's demonstration that lagged indexation brings about a trade-off between inflation and unemployment confirmed his previous conclusion that cold-turkey disinflation is costly. This inflation inertia (or ‘persistence’) result has been obtained in other rational expectation models as well (Fuhrer and Moore, 1995). As shown by Simonsen (1986a), inflation inertia was not a feature of Taylor's 1979 wage staggering model, which could only generate price-level inertia. Simonsen argued for the de-indexation of the Brazilian economy as a crucial step to end chronic inflation, which was eventually achieved through the monetary reform that introduced a new currency in 1994, partly under the influence of his ideas (Andrade and Silva, 1996).

In a series of essays published between 1986 and 1989 and in his 1995 book, Simonsen used game theory to provide a model of inertial inflation and to bring out the role of incomes policies to reduce it. He generalized Townsend's (1978) argument about the correspondence between rational expectations macroeconomics and Nash equilibria, and suggested that the central weakness of the rational expectation hypothesis is the implicit assumption that rational participants in a non-cooperative game promptly move to a Nash equilibrium. From that perspective, inflation inertia is consequence of a coordination failure between wage and price setters in the economy after an observed change in macroeconomic policy. Incomes policy can be used to resolve this coordination failure, in the sense of providing information to speed up the location of Nash equilibria by economic agents.

International debt crisis

Another macroeconomic issue that attracted Simonsen's attention was the debt dynamics of developing countries. Simonsen (1983a, ch. 5; 1985) put forward an analytical framework to derive the conditions for a country's solvency. He advanced a differential equation that splits the rate of change of the country's net foreign debt into the interest payments and the resource gap, and expressed it in the form of the ratio to exports. Simonsen then showed that the condition for solvency is that the rate of growth of exports exceeds the interest rate, which he used to explain the breakdown of competitive recycling in the early 1980s (see also Krueger, 1987). Apart from his new approach to inertial inflation, Simonsen also applied game theory to study the non-cooperative behaviour of banks in providing competitive loans to highly indebted developing countries (1985) and to investigate dynamic bargaining problems between banks and developing countries (1989), partly based on the Rubinstein (1982) bargaining model (cf.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Bulow and Rogoff, 1989). His model attempted to explain why creditors, organized as a cartel, preferred to deal with each country separately, and why debtors did not organize themselves as a cartel.

See Also

- Dornbusch, Rudiger
- sovereign debt
- wage indexation

Selected works

1964. A lei de Say e o efeito liquidez real. *Revista Brasileira de Economia* 18, 41–66.

1965. (With W. Baer.) Profit inflation and policy-making in an inflationary economy. *Oxford Economic Papers* 17, 279–90.

1967–69. *Teoria Microeconomica*, 4 vols. Rio: FGV.

1970. *Inflação: gradualismo vs. tratamento de choque*. Rio: APEC.

1974a. *Macroeconomia*, 2 vols. Rio: APEC.

1974b. The anti-inflationary policy. In *The New Brazilian Economy*, ed. M.H. Simonsen and R. Campos. Rio: Crown.

1983a. *Dinamica Macroeconomica*. São Paulo: McGraw-Hill.

1983b. Indexation: current theory and the Brazilian experience. In *Inflation, Debt, and Indexation*, ed. M. H. Simonsen and R. Dornbusch. Cambridge, MA: MIT Press.

1985. The developing-country debt problem. In *International Debt and the Developing Countries*, ed. J. Cuddington and G. Smith. Washington, DC: World Bank.

1986a. Cinqüenta anos da *Teoria Geral do Emprego*. *Revista Brasileira de Economia* 40, 301–34.

1986b. Rational expectations, income policies and game theory. *Revista de Econometria* 6, 7–46.

1988a. Price stabilization and incomes policies: theory and the Brazilian case study. In *Inflation Stabilization – The Experiences of Israel, Argentina, Brazil and Mexico*, ed. M. Bruno et al. Cambridge, MA: MIT Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1988b. (With R. Dornbusch.) Inflation stabilization: the role of income policies and monetization. In *Exchange Rates and Inflation*, ed. R. Dornbusch. Cambridge, MA: MIT Press.

1988c. Rational expectations, game theory and inflationary inertia. In *The Economy as an Evolving Complex System*, vol. 5, ed. P.W. Anderson, K.J. Arrow and D. Pines. New York: Addison-Wesley.

1988d. Interview. In *An Oral History of Finance: 1967–1988*, compiled by the editors of *Institutional Investor*. New York: William Morrow.

1989. Macroeconomia e teoria dos jogos. *Revista Brasileira de Economia* 43, 315–71.

1995. *30 anos de indexação*. Rio: FGV.

Bibliography

Alberti, V., Sarmento, C. and Rocha, D., eds. 2002. *Mario Henrique Simonsen – um homem e seu tempo*. Rio: FGV.

Andrade, J. and Silva, M. 1996. Brazil's new currency: origin, development and perspectives of the Real. *Revista Brasileira de Economia* 50, 427–67.

Boianovsky, M. 2002. Simonsen and the early history of the cash-in-advance approach. *European Journal of the History of Economic Thought* 9, 57–71.

Bulow, J. and Rogoff, K. 1989. A constant recontracting model of sovereign debt. *Journal of Political Economy* 97, 155–78.

Clower, R. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6, 1–8.

Fishlow, A. 1988. Tale of two presidents. In *Democratizing Brazil: Problems of Transition and Consolidation*, ed. A. Stepan. Oxford: University Press.

Fuhrer, J. and Moore, G. 1995. Inflation persistence. *Quarterly Journal of Economics* 110, 127–59.

Gray, J. 1976. Wage indexation – a macroeconomic approach. *Journal of Monetary Economics* 2, 221–35.

Krueger, A. 1987. Origins of the developing countries' debt. *Journal of Development Economics* 27, 165–87.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

McNelis, P. 1985. Review essay of Simonsen and Dornbusch 1983. *Journal of Money, Credit and Banking* 17, 274–80.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50, 97–110.

Taylor, J. 1979. Staggered wage setting in a macro model. *American Economic Review* 69, 108–13.

Townsend, R. 1978. Market anticipations, rational expectations and Bayesian analysis. *International Economic Review* 19, 481–94.

How to cite this article

Boianovsky, Mauro. "Simonsen, Mario Henrique (1935–1997)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000537> doi:10.1057/9780230226203.1530

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

simplex method for solving linear programs

George B. Dantzig
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The simplex method is used to solve linear programming problems based on pivoting from one iteration to the next. Invented by George Dantzig in 1947, it can be stated in 20 or so instructions for a computer. Commercial codes based on the simplex method, however, usually involve thousands of instructions which are there to take advantage of sparsity (most coefficients of practical problems are zero), to make it easy to start from solutions to variants of the same problem, and to guarantee numerical accuracy of the solution for large-scale systems.

Keywords

basic feasible solutions; Dantzig, G.B.; duality; Lagrange multipliers; linear programming; pivoting; simplex method for solving linear programs

Article

The data for the linear programming problem (LP) is stated below in standard form:
FIND Min $z, x_j \geq 0$:

$$a_{11}x_1 + \cdots + a_{1s}x_s + \cdots + a_{1n}x_n = b_1 \ldots\ldots a_{r1}x_1 + \cdots + a_{rs}x_s + \cdots + a_{rn}x_n = b_r \ldots\ldots a_{m1}x_1 + \cdots + a_{ms}x_s + \cdots + a_{mn}x_n = b_m \text{obj}: c_1x_1 + \cdots + c_sx_s + \cdots + c_nx_n = z - b_0 \quad (1)$$

Obj is the objective or cost equation defining z. In economic applications, the coefficient a_{ij}, depending on sign, is the input or output of item i per unit level of activity j and x_j is the level of activity j to be determined.

Any particular set of values $x^0 = (x_1^0, \ldots, x_n^0)$ that satisfies the first m equations of (1) is called a solution; if in addition $x_j^0 \geq 0$ for all j then x^0 is a feasible solution; if upon substitution into the obj equation of (1), x^0 yields a value of $z = \text{Min } z$, then x^0 is an optimal feasible solution.

The LP, however, could have been given in one of several other ways which, from a mathematical point of view, are all equivalent. Suppose the LP were originally stated as one of minimization of a linear form subject to a system of linear inequalities. It could then be easily converted to (1). For example, the relation $2u + 3v \leq 4$ can be replaced by the equation $2(x_1 - x_2) + 3(x_3 - x_4) + x_5 = 4$ where $x_j \geq 0$ by setting u and v each equal to the difference of two non-negative variables $u = x_1 - x_2$, $v = x_3 - x_4$, and introducing a non-negative *slack* variable x_5.

Commercial software for LP usually allows the user to specify which variables are unrestricted in sign and whether the relation is an equation or an inequality. The software program does not make the above substitutions but uses a modified form of the simplex algorithm designed to handle the mixed case of signed/unsigned variables and equation/inequality relations.

Pivoting defined

The simplex method consists of a sequence of $t = 0, 1, 2 \ldots$ pivot steps (iterations) performed on system (1) which transforms it on each step to a new, mathematically equivalent, system of equations. Any solution of (1) is also a solution for the system of iteration t, and conversely. Thus feasible and optimal feasible solutions remain feasible and optimal after pivoting and so remain for all t. Since the generated systems all have the same solution set as (1), it is not necessary to store in the memory of the computer a record of all the intermediate steps.

We will use the same symbols a_{ij}, c_j, b_i to denote the updated system after pivoting as before. When necessary to distinguish as to which iteration t they pertain, we will use a superscript a_{ij}^t, c_j^t, b_i^t.

To perform a pivot step on system (1) iteration $t = 0$ or a subsequent iteration t, select any term $a_{rs}x_s$ where $a_{rs} \neq 0$, called the *pivot term*. Replace equation r by dividing it through by a_{rs}, then replace each equation $i \neq r$ by subtracting from it the new rth equation multiplied by a_{is}. Do the same thing with the objective equation by subtracting from it the rth equation multiplied by c_s. This eliminates the variable x_s from all equations except the rth. During a pivot step, the current solution x^t is also updated to a new solution x^{t+1} by some rule.

A number of variants of the simplex method based on pivoting from one iteration to the next are used to solve LPs. These include the dual simplex method, the primal-dual method, and the symmetric method. They differ only in the rules used for choosing the pivot term or the way the current solution is updated.

The simplex method to be described was first proposed in 1947; it can be stated in 20 or so instructions for a computer. Commercial codes based on the simplex method, however, usually involve thousands of instructions which are there to take advantage of sparsity (most coefficients of practical problems are zero), to make it easy to start from solutions to variants of the same problem, and to guarantee numerical accuracy of the solution for large-scale systems.

Outline of the procedure

The simplex method consists of phase I which finds a feasible solution if one exists, and phase II which finds an optimal one if one exists. Thus the method can terminate with (a) no feasible solution, (b) an optimal feasible solution, or (c) a class of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

feasible solutions whose values for the objective $z \to -\infty$. Each phase applies a special subroutine called the simplex algorithm to a related but different LP problem. We begin by describing this algorithm.

Simplex algorithm

This algorithm requires the system to be given in canonical form with the right-hand side constants $b_i \geq 0$. The system is said to be in canonical form if we can permute the order of the variables of the first *m* equations so that coefficients of the first *m* variables form an identity matrix, i.e., a square array of all zeros except for a diagonal of all ones. We also require their corresponding terms in the obj equation be zero. We illustrate with an $m = 2$, $n = 5$ example.
FIND Min *z*, $(x_1, \ldots, x_5) \geq 0$:

$$2x_1 + 1x_3 + a_{14}x_4 + 1x_5 = 8 - 3x_1 + 1x_2 - 7x_4 + 1x_5 = 6 \text{obj}: 4x_1 + c_4x_4 + 1x_5 = z - 3. \quad (2)$$

By choosing the constants $c_4 = +5$ or –5 and $a_{14} = -2$ or +1, we have, in fact, four different examples. System (2) is in canonical form because we can reorder the variables so that x_3, x_2 come before the rest. When we do so the matrix of coefficients of x_3, x_2 in the first two equations is the 2×2 identity matrix:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (3)$$

The *ordered* set of *m* indices giving rise to the identity matrix, in the example {3, 2}, is called the basic set; its corresponding variables are called the basic variables; its set of coefficients is called the basis. Each iteration *t* will give rise to varying basic sets of *m* indices.

Termination

The simplex algorithm terminates with an optimal solution when a canonical system is generated on some iteration *t* with $c_j \geq 0$ for all *j*. This is the case in the example if $c_4 = +5$. Note $c_1 = 4$, $c_2 = c_3 = 0$, $c_4 = 5$, $c_5 = 1$. Phase I will always terminate in this way. Phase II can also terminate with a class of feasible solutions in which $z \to -\infty$. This happens when a canonical system is generated on some iteration *t* with some variable x_s whose $c_s < 0$ and all its other coefficients $a_{is} \leq 0$. In the example if $c_4 = -5$ and $a_{14} = -2$, then for variable x_4 this termination condition holds, namely: $c_4 = -5$, $a_{14} = -2$, $a_{24} = -7$.

Basic feasible solutions

The solution obtained by setting the values of all non-basic (independent) variables equal to zero and solving for the basic (dependent) variables is called a basic solution. Since the canonical form for each iteration *t* satisfies $b_i \geq 0$ for all *i*, the basic feasible solution is simply $x_j = 0$ for *j* non-basic and $x_{j_i} = b_i$ where $j_1, j_2, \ldots, j_m$ are the basic set of indices in the order that their coefficients form an identity matrix. In the example, $\{j_1, j_2\} = \{3, 2\}$; the basic feasible solution is $x_3 = 8$, $x_2 = 2$, $x_1 = x_4 = x_5 = 0$. Substituting this solution into the obj equation, we obtain $z = 3$.

Proof of optimality

To prove that the basic feasible solution yields $\text{Min } z = b_0$ when $c_j \geq 0$ for all *j*, we observe for our example with $c_4 = +5$ that the objective equation states that $z = 3 + 4x_1 + 5x_4 + x_5$. Therefore the value of $z \geq 3$ because $4x_1 + 5x_4 + x_5 \geq 0$ for all $x_j \geq 0$ and its lower bound $z = 3$ is attained for the basic feasible solution. Therefore $z = 3$ is minimum.
In general, the value of *z* for the basic feasible solution for iteration *t* is clearly $z = b_0$ and the obj equation can be rewritten $z = b_0 + \Sigma c_j x_j$. Therefore if $c_j \geq 0$ and $x_j \geq 0$ then $z \geq b_0$. Since the lower bound $z = b_0$ is attained for the basic feasible solution, this implies $\text{Min } z = b_0$.
Proof that $z \to -\infty$. We wish to show *z* has no lower bound when for some x_s, $c_s < 0$ and $a_{is} \leq 0$, for all *i*. In the example let $c_4 = -5$ and $a_{14} = -2$, then for x_4, $c_4 = -5$, $a_{14} = -2$, $a_{24} = -7$ which satisfies the termination condition. Setting all non-basic variables=0 *except* x_4 and solving for the basic variables and *z* in terms of x_4, we have:

$$x_3 = 8 + 2x_4,\ x_1 = x_5 = 0, x_2 = 6 + 7x_4, z = 3 - 5x_4. \quad (4)$$

As $x_4 \to +\infty$, a class of feasible solutions is generated in which $z \to -\infty$.
In general, setting all non-basic variables $x_j = 0$ except x_s, and solving for basic x_j and *z* in terms of x_s, we have:

$$x_j = b_i - a_{is}x_s \text{for } j_i \text{ basic} z = b_0 + c_s x_s. \quad (5)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Again we see for $a_{is} \leq 0$, $c_s < 0$, that a class of feasible solutions $x_j \geq 0$ is generated in which $z \to -\infty$.

Improving a basic feasible solution

Let s be such that $c_s = \text{Min } c_j$. If $c_s \geq 0$ the algorithm terminates with an optimal basic-feasible solution. If $c_s \leq 0$, then clearly setting all non-basic $x_j = 0$ except x_s and allowing x_s to increase causes $z = b_0 + c_s x_s$ to decrease towards $-\infty$; hence the more we can decrease x_s the better. However, the values of the basic variables in terms of $x_{j_i} = b_i - a_{is} x_s$ and therefore the maximum increase allowable for x_s in order to keep all x_j non-negative is $x_s = x_s^*$ where

$$x_s^* = \underset{i}{\text{Min}}(b_i / a_{is}) \tag{6}$$

where Min_i is restricted to i such that $a_{is} > 0$. If there are no $a_{is} > 0$, then we have the termination case already discussed of $z \to -\infty$. Otherwise the minimum occurs at some $i = r$ and $x_s^* = b_r / a_{rs}$. When $x_s = x_s^*$, the rth basic variable assumes the value $x_{j_r} = b_r - a_{rs}(b_r / a_{rs}) = 0$. This suggests that the variable x_s replace x_{j_r} as rth basic variable by pivoting on $a_{rs}x_s$.
We illustrate this for our example with $c_4 = -5$ and $a_{14} = 1$. Since $c_4 = \text{Min } c_j$ and $c_4 < 0$, we have $s = 4$. Accordingly we set all non-basic $x_j = 0$ except x_4 and solve for the values of basic variables in terms of x_4. Thus:

$$x_3 = 8 - 1x_4,\ x_1 = x_5 = 0, x_2 = 6 + 7x_4, z = 3 - 5x_4. \tag{7}$$

We are blocked from increasing x_4 indefinitely because x_3 would become negative if $x_4 > 8$ and our class of generated solutions would no longer remain feasible. At $x_4 = 8$ we have two variables positive, $x_4 = 8$ and $x_2 = 62$, and all the rest $x_1 = x_3 = x_5 = 0$. Therefore we drop $j = 3$ from our basic set and replace it by $j = 4$ by pivoting on $a_{14}x_4$. Thus we have:
Iteration $t = 0$

$$\begin{array}{lrrrrrl} & 2x_1 & & +1x_3 & +\boxed{1x_4} & +1x_5 & =8 \\ & -3x_1 & +1x_2 & & -7x_4 & +1x_5 & =6 \\ \text{obj:} & 4x_1 & & & -5x_4 & +1x_5 & =z-3 \end{array}$$

Iteration $t = 1$ (after pivoting using $\boxed{1x_4}$ as pivot term).

$$2x_1 + 1x_3 + 1x_4 + 1x_5 = 8 1x_1 + x_2 + 7x_3 + 8x_4 = 62 \text{obj: } 14x_1 + 5x_3 + 6x_5 = z + 37. \tag{9}$$

We conclude that the basic feasible solution for iteration 1, namely $x_4 = 8$, $x_2 = 62$, $x_1 = x_3 = x_5 = 0$ and $z = -37$ is the optimal feasible solution. If the obj equation for iteration $t = 1$ had some $c_j < 0$, we would have continued the algorithm.

Phase I

To initiate phase I, multiply by –1 all equations of (1) with $b_i < 0$, $i \neq 0$, so that (1) after modification $b_i \geq 0$. Next adjoin auxiliary variables, called artificials, $x_{n+1}, \ldots, x_{n+m}$ as shown below.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Phase I problem, Iteration $t = 0$:
FIND Min w, $x_j \geq 0$:

$$a_{11}x_1 + \cdots + a_{1s}x_s + \cdots + a_{1n}x_n + x_{n+1} = b_1 \ldots\ldots a_{r1}x_1 + \cdots + a_{rs}x_s + \cdots + a_{rn}x_n + x_{n+r} = b_r \ldots\ldots a_{m1}x_1 + \cdots + a_{ms}x_s + \cdots + a_{mn}x_n + x_{n+m} = b_m \text{obj: } d_1x_1 + \cdots + d_sx_s + \cdots + d_nx_n = w - d_0 \quad (10)$$

The obj equation has been replaced by a phase I obj defined by

$$d_j = -\sum_i a_{ij} \text{ and } d_0 = +\sum_i b_i. \quad (11)$$

Note the system is in canonical form with $b_i \geq 0$ so that we are all set to apply the simplex algorithm.

Special rule

Once an artificial variable x_{n+i} is pivoted out of the set of basic variables on some iteration t and becomes non-basic, it is discarded (i.e., all terms involving x_{n+i} are dropped from the canonical form). Hence the pivot term $a_{rs}x_s$ will be one from among the first m rows and n columns of (10).
If we add the first m equations of (10) to the obj equation, we obtain by (11) that:

$$x_{n+1} + x_{n+2} + \cdots + x_{n+m} = w. \quad (12)$$

Thus the phase I objective is equivalent to minimizing the sum of the artificial variables. If a feasible solution to (1) exists, then a feasible solution to (10) exists in which all $x_{n+i} = 0$ and therefore a feasible solution to (10) exists in which $w = 0$. Since $x_{n+i} \geq 0$, it follows for all feasible solutions to the phase I problem, $w \geq 0$ and Min $w \geq 0$. It is therefore impossible in phase I to find a class of solutions in which $w \to -\infty$. If the optimal solution yields Min $w > 0$, the simplex method is terminated with the statement that no feasible solution to (1) exists. If phase I terminates with $w = 0$, then we set up the phase II problem.

Transition to phase II

At the end of phase I if Min $w = 0$, then all artificial variables have value 0 in the basic solution. Usually there are no longer any artificial variables left among the basic ones in the canonical form. When this is the case, we replace the obj equation of phase I by the original one given as input data (1). Next we eliminate from the obj all terms c_jx_j corresponding to $j = j_i$ in the basic set. This is done by subtracting from the obj equation the ith equation of the canonical form multiplied by c_{j_i}. The phase II problem is now in canonical form ready to apply the simplex algorithm to find Min z.
For example, suppose at the end of phase I we have:
FIND Min z. $(x_1, \ldots, x_4) \geq 0$:

$$2x_1 + 1x_3 + a_{14}x_4 + 1x_5 = 8 - 3x_1 + 1x_2 - 7x_4 + 1x_5 = 6 \text{obj: } c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + c_5x_5 = z - 3. \quad (13)$$

The basic set is $\{j_1, j_2\} = \{3, 2\}$. Multiplying the first equation by c_3 and the second by c_2 and subtracting from obj, we eliminate the basic variables x_3, x_2 from the obj equation obtaining an obj equation of the form:

$$\text{obj: } c_1x_1 + c_4x_4 + c_5x_5 = z - b_0. \quad (14)$$

It can happen, however, at the end of Phase I for some iteration t that Min $w = 0$ and some artificial variable, say x_{n+r}, still is basic. Its basic solution value is $x_{n+r} = b_r = 0$. x_{n+r} is gotten rid of by pivoting on any term $a_{rs}x_s$ of the canonical form where $a_{rs} \neq 0$ and $s \leq n$. The new basic solution will have as rth basic variable $x_s = 0$ and x_{n+r}, now non-basic, is then discarded. This process is continued until all artificials are dropped.
There still remains the possibility that a pivot term for some r cannot be found because all $a_{rj} = 0$ for $j = 1, 2, \ldots, n$. In this case it is easy to prove that the rth equation of the original problem is *redundant* and the rth equation of the canonical form of the phase I problem can be discarded or, alternatively, x_{n+r} can be reclassified as belonging among the *true* variables – it will do no harm to include it because in all subsequent iterations its basic solution value will remain zero. Once the artificials are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

made non-basic and removed, we complete the transition as outlined in the paragraph above.
Upon termination of the simplex algorithm applied to the Phase II problem, the software program is directed to print out a statement about the type of termination. In the case of an optimum solution, this is followed by the list of indices of the obj, the basic variables, and alongside them the values of the corresponding b_0 and b_i.
In the case z is unbounded below, the list printed is

$$\begin{vmatrix} \text{obj} \\ j_i \\ s \end{vmatrix} \begin{vmatrix} b_0 \\ b_i \\ 0 \end{vmatrix} \begin{vmatrix} +c_s \\ -a_{is} \\ 1 \end{vmatrix} \text{for } i = 1, \ldots, m. \tag{15}$$

This information permits one to generate z and a feasible solution for any choice of $x_s \geq 0$.

Proof of convergence

It is not difficult to show that if any basic set of indices were to be repeated in some subsequent iteration, the entire canonical form would be repeated including the value of z in the basic solution. In the example (8), the value of z in the basic solution is $z = 3$. After pivoting, see (9), the value of $z = -37$. We see that its value decreased from 3 to –37.
In general, if there is a *positive* decrease in the value of z in the basic solution from one iteration to the next, the canonical form cannot be repeated since the value of z is lower. On the other hand, the iterative process must stop sometime because there is only a finite number of canonical forms. But the only way it could have stopped is via one of the two termination conditions. Hence the iterative process is finite when there is a positive decrease on each iteration. This should not be interpreted, however, as meaning the algorithm is efficient because the number of ways to pick m objects out of n grows exponentially with increasing m and n.

Degeneracy

Should the pivot term occur on a row r whose $b_r = 0$, then the updated value of z in the basic solution is $z = b_0 - b_r(c_s / a_{rs}) = b_0$, i.e., the change in value of z is zero and the proof of convergence given above is no longer applicable. In this case, one or more of the values of the basic variables in a basic solution are zero and the basic solution is said to be *degenerate*. There are examples of canonical forms with degenerate basic solutions, which after a number of pivots return to the original canonical form.
To avoid this possibility of cycling, special rules have been invented that are easy to implement but are not found in commercial codes. Almost all practical problems are degenerate. Failure to provide a rule has never (or almost never) caused the algorithm to cycle in practice. From a theoretical point of view, however, devices that prevent cycling are important because the simplex method is used as a powerful analytic tool for proving theorems like the duality theorem.

Economic interpretations

Feasible. In planning, a feasible solution is a plan or policy that is physically implementable. The plan may be feasible but not necessarily an optimal one.
Prices. Associated with a basic solution is a set of prices $(\pi_1, \pi_2, \ldots, \pi_m)$, also called *Lagrange Multipliers*, which are defined so that if we 'price out' the inputs and outputs of activities associated with basic variables, they *break even*. By this is meant for each j in the basic set:

$$c_j^0 - \sum_i \pi_i a_{ij}^0 = 0, \; j = j_1, \ldots, j_m. \tag{16}$$

where a_{ij}^0, c_j^0 refer to a_{ij}, c_j of iteration $t = 0$.
The value of c_j of iteration t is denoted by c_j^t. It is easy to show that c_j^t can be obtained directly from the data of iteration $t = 0$ by 'pricing out' any activity j in terms of the prices associated with the current basis:

$$c_j^t = c_j^0 - \sum_i \pi_i a_{ij}^0, \; j = 1, \ldots, n. \tag{17}$$

If $c_j^t < 0$ for any activity $j = j^*$, it pays to replace one of the activities of the basic set by activity j^*. The simplex method chooses among the activities j in the non-basic set $j^* = j_s$ that shows the greatest profitability per unit change of activity level x_j, namely s such that $c_s = \text{Min } c_j$ and $c_s < 0$.

Duality

The dual of the LP (1) iteration $t = 0$ is defined by:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

FIND Max **z**, $(\pi_1, \pi_2, \ldots, \pi_m)$, and $(y_1, \ldots, y_m)$:

$$y_j = c_j - \sum_i \pi_i a_{ij},\ j = 1, \ldots, n\mathbf{z} = b_0 + \sum_i \pi_i b_i \tag{18}$$

It is easy to show that $\mathbf{z} \leq z$ for all feasible solutions to the original primal system (1) and feasible solutions to the dual system (18). This implies when feasible solutions to both the primal and dual systems exist that *z* has a finite lower bound and **z** has a finite upper bound. We have shown in this case that an optimal feasible solution exists to the primal system. Moreover for the optimal canonical form of some iteration *t* that π_i defined by (16), satisfies $c_j^t \geq 0$ in (17). Setting $y_j = c_j^t \geq 0$ for all *j*, we see that π_i and $y_i \geq 0$ satisfy (18). It is easy to show that *z* of this basic feasible solution satisfies $\mathrm{Min}\ z = b_0^0 + \Sigma_i \pi_i b_i^0$ so that $\mathbf{z} = z$. It follows therefore that $\mathrm{Max}\ \mathbf{z} = \mathrm{Min}\ z$. This is called the *strong* duality theorem; note that we have proved it using the properties of the simplex algorithm.

Computational experience

Since 1947 the simplex method and its variants have successfully solved each day thousands of large and small scale practical problems. New methods for solving LPs are constantly cropping up. Many LPs have special structures and special algorithms have been developed to solve them. For example there is considerable research on how to efficiently solve large-scale dynamic economic models under uncertainty. One approach makes use of parallel computers, random sampling, methods of decomposing the problem into many subproblems which are solved using the simplex method as a subroutine.

How to cite this article

Dantzig, George B. "simplex method for solving linear programs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000135> doi:10.1057/9780230226203.1531

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

simulation estimators in macroeconometrics

Beth F. Ingram
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The article outlines the method of simulated moments as a technique for estimating the parameters of dynamic, stochastic general equilibrium macroeconomic models. A detailed description is provided for implementation of the method, and its statistical properties are discussed. A brief comparison with other common estimation methods (calibration, generalized method of moments and maximum likelihood) is presented.

Keywords

autocorrelation consistent estimators; Bayesian methods in macroeconometrics; calibration; dynamic, stochastic, general equilibrium (DSGE) models; ergodicity; Euler equations; expected utility; generalized method of moments; heteroskedasticity; maximum likelihood estimation; method of simulated moments; simulation estimators in macroeconometrics

Article

The method of simulated moments (MSM) in macroeconometrics is a method for estimating the parameters of a dynamic, stochastic, general equilibrium (DSGE) model using simulated solution paths for the model's observable variables.

Regardless of the complexity of the DSGE model, implementation of MSM requires only simulated data from that model. Moments of the simulated model are then formally matched to moments of the observed data, and the parameter vector that minimizes the distance between the model and the empirical moments is the MSM estimator. MSM estimates typically have nice statistical properties, including consistency and asymptotic normality.

In general, the solution to a DSGE can be characterized as a vector stochastic process dependent upon k x 1 parameter vector, θ, and m x 1 fundamental shock vector $\{\varepsilon_n, n \geq 0\}$. Consider the simple stochastic growth model in which a representative agent chooses consumption, c_t, labour supply, h_t, and capital, k_t, to maximize expected utility subject to a resource constraint and endowment constraints:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\max_{\{c_n, k_n, h_n > 0\}} E_0 \sum_{n=1}^{\infty} \beta^n \ln(c_n) \; s.t. \; c_n + k_n = A_n k_{n-1}^{\alpha} h_n^{1-\alpha} \; h_n \leq 1 \text{ and } k_0 \text{ given.}$$

Here, $\{A_n, n \geq 0\}$ is an exogenous stochastic process; for example:

$$A_n = \rho A_{n-1} + \varepsilon_n, \; \varepsilon_n \sim N(0, \sigma^2). \tag{1.1}$$

In this example, the parameter vector is comprised of the four parameters in the model, $\theta = [\alpha \beta \rho \sigma^2]$, and the fundamental shock is $\{\varepsilon_n, n \geq 1\}$. The solution to the model is a collection of stochastic processes, $\{c_n, k_n, h_n, n \geq 1\}$ that depend on θ, k_0, and $\{\varepsilon_n, n \geq 1\}$.
It is assumed that a method exists for generating a realization of the stochastic process given a specific parameter vector, $\tilde{\theta}$, and a realization of the shocks, $\{\tilde{\varepsilon}_n\}_{n=1}^{N}$. In the example above, the decision rules for capital and consumption can be derived analytically (labour choice is trivial in this example):

$$k_n = \alpha \beta A_n k_{n-1}^{\alpha}$$
$$c_n = (1 - \alpha\beta) A_n k_{n-1}^{\alpha} \tag{1.2}$$

Given k_0, specific values for the parameters and a realization of the shock process (1.1), a finite realization of consumption and capital can be generated from (1.2). The goal of MSM is to use this simulated data from the model and an observed data set to obtain an estimate for θ. Denote by $\{y_n(\tilde{\theta}, \tilde{\varepsilon}_n)\}_{n=0}^{N}$ the q x 1 vector of data simulated as a solution of a general DSGE model under parameter vector $\tilde{\theta}$ and shock realization $\{\tilde{\varepsilon}_n\}_{n=1}^{N}$; denote the observed data counterpart to these series by $\{x_t\}_{t=1}^{T}$, with N not necessarily equal to T. In the example, the simulated and observed data might consist of such variables as personal consumption expenditures, real GDP, and gross fixed investment. Let $m(\bullet)$ represent an r x 1 vector of functions. For instance,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$m(\bullet) = [y_{1n}(y_{2n} - \bar{y}_2)^2 y_{3n} / y_{4n}]$$

If the model is a true description of the data-generating process for the observed data, then given the true parameter vector, θ_0, the following must hold:

$$E[m(y_n(\theta_0))] = E[m(x_t)] \tag{1.3}$$

That is, the moments implied by the model should be equal to the moments of the observed data when the model is evaluated at the parameter vector that generated the observed data.
In general, the theoretical moments in (1.3) cannot be evaluated analytically, so we employ the empirical counterpart to (1.3), given simulated data from the model and the observed data:

$$\frac{1}{N}\sum_{n=1}^{N} m(y_n(\theta_0)) \approx \frac{1}{T}\sum_{t=1}^{T} m(x_t) \tag{1.4}$$

Since we are using time averages to compute the expectations in (1.3), the equality in (1.4) can be satisfied only asymptotically. Hence, the estimation strategy involves choosing a parameter vector that minimizes the following quadratic form for some *r x r* symmetric weighting matrix W_T:

$$\Delta(\theta) = \left[\frac{1}{N}\sum_{n=1}^{N} m(y_n(\theta)) - \frac{1}{T}\sum_{t=1}^{T} m(x_t)\right] W_T \left[\frac{1}{N}\sum_{n=1}^{N} m(y_n(\theta)) - \frac{1}{T}\sum_{t=1}^{T} m(x_t)\right]' \tag{1.5}$$

Hansen (1982) provides assumptions under which a similar estimator, the generalized method of moments (GMM) estimator, is both consistent and asymptotically normal. Lee and Ingram (1991) show that, under the conditions in Hansen (1982), the MSM estimator will have an asymptotic normal distribution with mean θ_0 and covariance matrix:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\Sigma = (B'WB)^{-1}B'W(1+\tau)\Omega WB(B'WB)^{-1} \tag{1.6}$$

The constant τ is equal to *T*/*N*, and the matrices *W*, *B* and Ω are defined as follows:

$$W = \operatorname*{plim}_{T\to 0} W_T, B \equiv E[\partial m(y_n(\theta)) / \partial \theta], \text{ and } \Omega \equiv Var\left[\frac{1}{T}\sum_{t=1}^{T} m(x_t)\right].$$

Duffie and Singleton (1993) provide an alternative set of assumptions under which the MSM estimator is consistent. Two assumptions deserve particular attention. First, the two stochastic processes $\{y_n(\theta, \varepsilon_n), n \geq 0\}$ and $\{x_t, t \geq 0\}$ must be stationary and ergodic; this ensures that the time average of a moment converges asymptotically to its expectation. Second, eq. (1.3) must have a unique zero at θ_0 in order for the parameter vector to be exactly identified; if more than one value of θ satisfies eq. (1.3), the estimator will not necessarily converge to θ_0.

Clearly, the size of the asymptotic covariance matrix and, thus, the precision of the estimate of θ is determined by the choice of *W*, the length of the simulated series relative to the observed series, the matrix *B*, and the matrix Ω. We discuss each element in turn.

The only restriction on the choice of W_T is that it converge in probability to a constant matrix, *W*. An obvious choice for W_T might be the identity matrix. In that case,

$$\Sigma = (1+\tau)(B'B)^{-1}B'\Omega B(B'B)^{-1}$$

When $W = I_r$, all moments are weighted equally, which may not necessarily be optimal from an efficiency standpoint. Although this is a straightforward choice for *W*, it does not produce the smallest asymptotic covariance matrix. With all else held constant, the smallest asymptotic covariance matrix is attained when W_T is chosen to equal $[(1+\tau)\Omega]^{-1}$. In that case, $\Sigma = (1+\tau)\left[B'\Omega^{-1}B\right]^{-1}$. The asymptotic covariance matrix can be further reduced by choosing *N* to be large relative to *T* (and τ close to zero). As the length of the simulated data series increases relative to the length of the observed series, the term $(1+\tau)$ tends to 1.

The *r x k* matrix *B* is a measure of the sensitivity of the moments to changes in the parameter vector. Note that the number of moments must equal or exceed the number of parameters, $r \geq k$; if not, the *k x k* matrix $B'\Omega^{-1}B$ will not be invertible since the rank of this matrix can be no larger than the rank of its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
constituent matrices. If the partial derivatives of the moments with respect to a particular parameter are close to zero, then B will be close to non-invertible, as will $B'\Omega^{-1}B$, producing large standard errors. One solution is to choose a different set of moments. A smaller asymptotic covariance matrix is achieved when the moments chosen have larger derivatives (in absolute value) with respect to the parameters. Heuristically, the larger the derivative with respect to a parameter, the more informative is the moment about that parameter.

The matrix Ω must be estimated consistently, despite the likely presence of autocorrelation in the moments. Newey and West (1987) provide one such estimator that is consistent in the presence of both autocorrelation and heteroskedasticity. Many statistical packages provide algorithms for implementing heteroskedasticity and autocorrelation consistent (HAC) estimators for covariance matrices; Andrews (1991) provides a comparison of the properties of several such estimators.

To implement the procedure, the researcher must be able to generate realizations drawn from the stationary distribution of the stochastic process $\{V_n(\theta; \varepsilon_n), n \geq 0\}$. The finite realization of the stochastic process, however, depends on a set of starting values (for example, k_0 in the example above). The researcher must draw the starting values from the stationary distribution for the stochastic process; this, however, is problematic in practice. A practical solution is to generate simulations longer than are needed, and to drop a set of observations from the start of the sample.

If the number of moments exceeds the number of parameters to be estimated, $r>k$, this method also provides a test of fit for the model. Under the null hypothesis that the model provides a true description of the observed data, the product of the length of the observed series and the value of eq. (1.5) evaluated at the estimated value of θ is distributed as a chi-square random variable, $T \times \Delta(\hat{\theta}) \sim \chi^2(r-k)$. Of course, failure to reject the model only indicates that the model fits the data along the dimensions implied by the moments used in estimation.

The MSM is one of many approaches for estimating DSGE models. Calibration is similar in the sense that parameters are chosen to equate model and data moments; calibration, as normally implemented, lacks a statistical foundation. Like MSM, generalized method of moments (GMM) is a limited information method in that is uses a subset of the stochastic information implied by the model. GMM uses the conditional moments implied by the stochastic Euler equations produced by the DSGE's optimization problem. The advantage of GMM is that a complete solution to the DSGE need not be generated; however, the Euler equations may involve data series that do not have observable counterparts. In contrast, maximum likelihood estimation (MLE) is a full information method, but requires a complete characterization of the likelihood of the observed data implied by the DSGE. Since the DSGE is apt to be false along some dimension, the likelihood function will be mis-specified. Zhou (2001) and Ruge-Murcia (2003) provide a more detailed analysis and comparison of the statistical properties of MSM, GMM and MLE.

A final issue with MSM (and all classical estimation methods) is that it can yield parameter estimates that violate sensible restrictions on the model's parameter vector. For example, it is quite common to produce estimates of the rate of time discount, β, that exceed 1; Bayesian estimation methods (DeJong, Ingram and Whiteman, 2000) provide one approach to resolving this issue.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Bayesian methods in macroeconometrics
- calibration
- econometrics
- simulation-based estimation

Bibliography

Andrews, D. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–58.

DeJong, D., Ingram, B. and Whiteman, C. 2000. A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* 98, 203–23.

Duffie, D. and Singleton, K. 1993. Simulated moments estimation of Markov models of asset prices. *Econometrica* 61, 929–952.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Lee, B. and Ingram, B. 1991. Simulation estimation of time-series models. *Journal of Econometrics* 47, 197–205.

Newey, W. and West, K. 1987. A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.

Ruge-Murcia, F. 2003. Methods to estimate dynamic stochastic general equilibrium models. Working Paper No. 17-2003, CIREQ, University of Montreal.

Zhou, H. 2001. Finite sample properties of EMM, GMM, QMLE, and MLE for a square-root interest rate diffusion model. *Journal of Computational Finance* 5, 89–122.

How to cite this article

Ingram, Beth F. "simulation estimators in macroeconometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000492> doi:10.1057/9780230226203.1533

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

simulation-based estimation

Eric Renault
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

For a parametric econometric model with possibly latent variables, the simulation tool and Monte Carlo integration provide a versatile minimum distance estimation principle. The general approach is dubbed *simulation-based indirect inference*. It can take advantage of any instrumental piece of information that identifies the structural parameters. Examples include the *simulated method of moments* and its *simulated-score-matching* version. Monte Carlo integration also allows numerical assessment of the criterion to maximize for M-estimation. Asymptotic efficiency is reached by the *simulated maximum likelihood* or a *simulated score* technique. Since the simulator is provided by the structural model, the classical trade-off between efficiency and robustness to misspecification must be revisited.

Keywords

bias correction; bootstrap; efficient method of moments; extremum estimation; GARCH models; generalized method of moments; indirect inference; indirect least squares; maximum likelihood; Monte Carlo methods; parameter-matching estimators; score-matching estimators; simulated expectation maximization; simulated maximum likelihood; simulated method of moments; simulated score matching; simulation-based estimation; simulation-based indirect inference; statistical estimation; stochastic volatility models; white noise

Article

Simulation-based estimation is an application of the general Monte Carlo principle to statistical estimation: any mathematical expectation, when unavailable in closed form, can be approximated to any desired level of accuracy through a generation of (pseudo-) random numbers. Pseudo-random numbers are generated on a computer by means of a deterministic method. (For convenience, we henceforth delete the qualification ‘pseudo’.) Then a well-suited drawing of random numbers (or vectors) $Z_1, Z_2, \ldots, Z_H$ provides the Monte Carlo simulator $(1/H)\sum_{h=1}^{H} g(Z_h)$ of $E[g(Z)]$. Of course, one may also want to resort to many simulators improving upon this naive one in terms of variance reduction, increased smoothness and reduced computational cost. A detailed discussion of simulation techniques is beyond the scope of this article. Nor are we going to study Monte Carlo experiments, which complement a given statistical procedure by the observation of its properties on simulated data. Rather, our focus of interest is to show how Monte Carlo integration may directly help to compute estimators that would be unfeasible without resorting to simulators.
The article is organized as follows. Section 1 is devoted to the most natural use of Monte Carlo integration for estimation, which is finite sample bias correction. It encompasses the parametric bootstrap. More generally, we use throughout the framework of a fully parametric econometric model with possibly latent variables, as defined in Section 1. Section 2 emphasizes that the simulation tool actually provides at least an asymptotic bias correction in much more general settings, such as simultaneity bias, bias due to errors in variables or any kind of misspecification bias. The general approach is dubbed *simulation-based indirect inference* (SII). Instead of using SII only for bias correcting a poor initial estimator, we can actually take advantage of any instrumental piece of information, insofar as it (over)identifies the structural parameters of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

interest. Section 3 is devoted to the *simulated method of moments* (SMM) and its *simulated-score-matching* version. With an asymptotic point of view, it can be seen as a particular asymptotic case of SII when instrumental parameters are some well-chosen moments. Besides computation of moments to match, Monte Carlo integration can also be used for the direct assessment of the criterion to maximize for M-estimation, when it is not available in closed form. The objective of asymptotic efficiency in the context of a parametric model leads us to put forward the *simulated maximum likelihood* (SML) or a *simulated score* technique, both described in Section 4. Some alternative simulated M-estimators, convenient though inefficient, are also reviewed. Concluding remarks in Section 5 are mainly focused on the trade-off between efficiency and robustness to misspecification; the fact that the structural model is also providing a simulator raises new issues for this classical trade-off.

The exposition in this article of simulation-based estimation methods relevant for econometric applications is selective in several respects. We do not present Markov chain Monte Carlo methods and data augmentation techniques. These are especially popular in Bayesian statistics and econometrics, but also relevant for some applications in a classical inference setting. Generally speaking, any kind of random drawing in the parametric space is beyond the scope of this article. Finally, it should be borne in mind throughout that all the simulation-based estimation methods have a non-simulation-based counterpart. While it is well known that SMM and SML are the simulation-based counterparts of GMM (generalized method of moments) and MLE (maximum likelihood estimation) respectively, it may be less known that the approaches of bootstrap and indirect interface make sense even without simulations. The essential characteristic of these techniques is to insert a consistent estimator of the data generating process in a functional of the true data distribution. Simulations are only a tool to evaluate the resulting estimated functional which, more often than not, is not available in closed form. There are, however, interesting exceptions like linear indirect least squares. Moreover, even though we always refer to the general concept of Monte Carlo integration, it does not necessarily involve a large number of simulated paths. Asymptotic theory of simulation-based estimation techniques will be considered when the length of the observed sample path goes to infinity. Depending on the techniques, the number of simulated paths may or may not be constrained to tend to infinity to get consistent estimators.

1 General framework and simulation-based bias correction

Let us denote by θ a vector of *p* unknown parameters. We want to build an accurate estimator $\hat{\theta}_T$ of θ from an observed sample path of length *T*. Let us assume that we have at our disposal some initial estimator, denoted by $\hat{\beta}_T$. Note that we purposely use a letter β different from θ to stress that the estimator $\hat{\beta}_T$ may give a very inaccurate assessment of the true unknown θ^0 we want to estimate. In particular this estimator is potentially severely biased: its expectation $b_T(\theta^0)$ does not coincide with θ^0. The notation $b_T(\theta^0)$ refers to the so-called binding function (Gourieroux and Monfort, 1995) and depends on at least two things: not only on the true unknown value of the parameters of interest but also on the sample size. The bootstrap is a method for estimating the distribution of an estimator, and in particular its expectation, by re-sampling the data. We refer the reader to Hall (1992) and Horowitz (1997) for surveys from which we borrow here. Since the bootstrap estimate is built upon an estimator of the data distribution, it is always recommended to use a parametric estimator of it when available. This is why, since this article is about estimation in parametric models, we focus on the parametric bootstrap, which was first considered in econometrics for the linear regression model with Gaussian errors:

$$y_i = z_i' b + \sigma \varepsilon_i, \quad \varepsilon_i \sim \mathrm{IIN}(0, 1).$$

Of course, bootstrapping is not very useful in such a simple context but it may become relevant if for instance the dependent variable is replaced by its Box–Cox transformation with some unknown parameters (Horowitz, 1997). More generally we allow for any kind of non-linear transformation and also for dynamic models in reduced form, possibly including lagged endogenous variables among the explanatory variables (see Monfort and Van Dijk, 1995 for a thorough exposition of the general framework presented below):

$$y_t = r[z(1, t), y(0, t-1), \varepsilon_t; \theta], \quad t = 1, \ldots, T \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where (ε_t) is a white-noise process whose marginal distribution P_ε is known, (z_t) is a process which is independent of (ε_t), $z(1,t) = (z_\tau)_{1 \le \tau \le t}$, $y(0,t-1) = (y_\tau)_{1 \le \tau \le t-1}$ and $r(\cdot)$ is a known function. Model (0) defines the conditional pdf $f[y_t|x_t;\theta]$ where x_t includes the realization of all the predetermined variables $z(1,t)$ and $y(0,t-1)$. Then, by drawing independently from P_ε, it is possible to simulate values ε_t^h, t=1, …, T and h=1, …, H and to compute:

$$y_t^h(\theta) = r[z(1,t), y(0,t-1), \varepsilon_t^h; \theta], \quad t = 1, \ldots, T, \quad h = 1, \ldots, H.$$

The pdf of $y_t^h(\theta)$ is precisely $f[y_t|x_t;\theta]$. In other words, it is possible to perform *conditional simulations*, that is, to draw, for each t, from the conditional distribution whose pdf is $f[y_t|x_t;\theta]$ for any given value θ of the unknown parameters and for the observed value of x_t. Note that, in all simulation-based estimation methods considered below, the basic drawings ε_t^h will be kept fixed when θ changes.

For the purpose of SMM, it will actually be worthwhile making a distinction between such *conditional simulations* and (unconditional) *path simulations*, which may be the only ones feasible in the more general case of a non-linear state-space model defined as:

$$y_t = r_1[z(1,t), y(0,t-1), y^*(0,t), \varepsilon_{1t}; \theta], \quad t = 1, \ldots, T$$
$$y_t^* = r_2[z(1,t), y(0,t-1), y^*(0,t-1), \varepsilon_{2t}; \theta], \quad t = 1, \ldots, T$$
(2)

where $\varepsilon_t = (\varepsilon_{1t}', \varepsilon_{2t}')'$ is a white-noise process whose marginal distribution P_ε is known, (z_t) is independent of (ε_t), (y_t^*) is a process of latent variables and r_1 and r_2 are two known functions. The big difference between models (1) and (2) is that the latter only recursively defines the observed endogenous variables through a path of latent ones, making conditional simulation impossible. More precisely, from independent random draws ε_t^h, t=1, …, T and h=1, … H in P_ε we can now compute recursively:

$$y_t^{*h}(\theta) = r_2[z(1,t), y^h(0,t-1)(\theta), y^{*h}(0,t-1)(\theta), \varepsilon_{2t}^h; \theta], t = 1, \ldots, T\ ;\ h = 1, \ldots H. \quad y_t^h(\theta) = r_1[z(1,t), y^h(0,t-1)(\theta), y^{*h}(0,t)(\theta), \varepsilon_{1t}^h; \theta], t = 1, \ldots, T\ ;\ h = 1, \ldots H.$$

In other words, while each simulated path $y^h(0,T)(\theta)$, h=1, … H has been correctly drawn from its distribution given the observed path $z(1,T)$ of exogenous variables for each possible value of θ, the draw of $y_t^h(\theta)$ at each given t is conditional to past *simulated* $y^h(0,t-1)(\theta)$ and not to past *observed* $y(0,t-1)$: hence the terminology *path simulations*. Note, however, that the model does not specify the probability distribution of exogenous variables and thus, all simulations are conditional to the observed path $z(1,T)$ of exogenous variables.

In both cases, model (1) or (2), since the spirit of bootstrap is re-sampling from a preliminary estimator, $\hat{\beta}_T$ gives rise to H bootstrap samples $y^h(0,T)(\hat{\beta}_T)$, h=1, … H. For

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

each bootstrap sample, the same estimation procedure can be applied to get *H* estimators denoted as $\beta_T^h(\hat{\beta}_T)$, *h*=1, … *H*. These *H* estimations characterize the bootstrap distribution of $\hat{\beta}_T$ and allow us for instance to approximate the expectation $b_T(\theta^0)$ by $b_T(\hat{\beta}_T)$. Of course, $b_T(\hat{\beta}_T)$ is not known in general but may be approximated at any desired level of accuracy, from the Monte Carlo average $(1/H)\sum_{h=1}^{H}\beta_T^h(\hat{\beta}_T)$, for *H* sufficiently large. As already mentioned, one may imagine non-simulation based versions of bootstrap when the binding function is available in closed form. In any case, the bias-corrected bootstrap estimator is then defined as:

$$\hat{\theta}_T = \hat{\beta}_T - [b_T(\hat{\beta}_T) - \hat{\beta}_T] \quad (3)$$

It is worth mentioning however that this parametric bootstrap procedure requires that we sufficiently trust the initial estimator $\hat{\beta}_T$ to consider that the estimated bias $[b_T(\hat{\beta}_T) - \hat{\beta}_T]$ gives a correct assessment of the true bias $[b_T(\theta^0) - \theta^0]$. This is the reason why Gourieroux, Renault and Touzi (2000) have rather proposed an iterative procedure which, at step *j*, improves upon an estimator $\hat{\theta}_T^j$ by computing $\hat{\theta}_T^{j+1}$ as:

$$\hat{\theta}_T^{j+1} = \hat{\theta}_T^j + \lambda[\hat{\beta}_T - b_T(\hat{\theta}_T^j)] \quad (4)$$

for some given updating parameter λ between 0 and 1. In other words, at each step, a new set of simulated paths $\nu^h(0, T)(\hat{\theta}_T^j)$, *h*=1, … *H* is built and it provides a Monte Carlo assessment $b_T(\hat{\theta}_T^j)$ of the expectation of interest. It is worth noting that this does not involve new random draws of the noise ε. Note that (3) corresponds to the first iteration of (4) in the particular case λ=1 with a starting value $\hat{\theta}_T^1 = \hat{\beta}_T$. While this preliminary estimator is indeed a natural starting value, the rationale for considering λ smaller than 1 is to increase the probability of convergence of the algorithm, possibly at the cost of slower convergence (if faster update would also work). If the algorithm converges, its limit will define an estimator $\hat{\theta}_T$ solution of:

$$b_T(\hat{\theta}_T) = \hat{\beta}_T \quad (5)$$

Gourieroux, Renault and Touzi (2000) study more generally the properties of the estimator (5), which is actually a particular case of SII estimators developed in the next section. The intuition is quite clear. Let us call $\hat{\beta}_T$ the naive estimator. Our preferred estimator $\hat{\theta}_T$ is the value of unknown parameters θ, which, if it had been the true one, would have generated a naive estimator which, in average, would have coincided with our actual naive estimation. In particular, if the bias function $[b_T(\theta) - \theta]$ is linear with respect to θ, we deduce $b_T[E(\hat{\theta}_T)] = E(\hat{\beta}_T) = b_T(\theta^0)$ and thus our estimator is unbiased. Otherwise, unbiasedness is only approximately true to the extent a linear approximation of the bias is reasonable. Since, in the context of stationary first order autoregressive processes, the negative bias of the OLS estimator of the correlation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

coefficient becomes more severely non-linear in the near unit root case, Andrews (1993) has put forward a median-unbiased estimator based on the principle (5) with median replacing expectation. The advantage of median is to be immune to nonlinear monotonic transformations. However, its generalization to multivariate parameters is problematic.

Another well-documented advantage of bootstrap is to provide asymptotic refinements when the initial procedure is not too bad. Gourieroux, Renault and Touzi (2000) have shown that SII does as well as bootstrap in this respect.

2 Simulation-based indirect inference

Let us start from the simple textbook example of a just-identified supply–demand system in equilibrium:

$$Q_t^s = \theta_1 p_t + \theta_2 z_{1t} + u_{1t} Q_t^d = \theta_3 p_t + \theta_4 z_{2t} + u_{2t} Q_t^s = Q_t^d = Q_t.$$

Then the reduced form can obviously be written as a bivariate regresssion of (Q_t, p_t) on (z_{1t}, z_{2t}) and the reduced form regression coefficients β are given as a function $\beta = b(\theta)$ of the structural parameters:

$$\theta = (\theta_1, \theta_2, \theta_3, \theta_4) b(\theta) = (\theta_1 - \theta_3)^{-1}(-\theta_2\theta_3, \theta_1\theta_4, -\theta_2, \theta_4).$$

Under standard assumptions, the vector β of reduced form parameters can be consistently estimated by its OLS counterpart $\hat{\beta}_T$. Moreover, the binding function $\beta = b(\theta)$ relating the vector β of reduced form parameters to the vector θ of structural parameters is clearly one-to-one. Inverting the binding function is a straightforward exercise and suggests computing a consistent estimator $\hat{\theta}_T$ of the structural parameters as $\hat{\theta}_T = b^{-1}(\hat{\beta}_T)$. This estimator has been known since the early days of the simultaneous equations literature as the *indirect least squares* estimator. We conclude from this example that defining an indirect estimator $\hat{\theta}_T$ of the parameters of interest from an initial estimator $\hat{\beta}_T$ and a binding function $b_T(\cdot)$ by solving the equation:

$$b_T(\hat{\theta}_T) = \hat{\beta}_T \tag{6}$$

may be worthwhile in many situations other than the bias-correction setting of Section 1. The vector β of the so-called *instrumental parameters* must identify the *structural parameters* θ but does not need to bear the same interpretation. However, the example of *indirect least squares* is too simple to display all the features of *indirect inference* as more generally devised by Smith (1993) and Gourieroux, Monfort and Renault (1993). Two complications may arise.

First, the binding function is not in general available in closed form and can be characterized only thanks to Monte Carlo integration. Moreover, by contrast with the simple linear example, the binding function does depend in general on the sample size *T*.

Second, most interesting examples allow for over-identification of the structural parameters, for instance through a bunch of instrumental variables in the simultaneous equation case. This is the reason why we refer henceforth to the auxiliary parameters β as instrumental parameters.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The key idea is that, as already explained in Section 1, our preliminary estimation procedure for instrumental parameters not only gives us an estimation $\hat{\beta}_T$ computed from the observed sample path but can also be applied to each simulated path $y^h(0, T)(\theta)$, h=1, … H, always associated to the observed path $z(1,T)$ of exogenous variables. Thus, we end up, for any fixed value of θ, with a set of H 'estimations' $\beta_T^{(h)}(\theta)$. Averaging them, we get a Monte Carlo binding function:

$$\beta_{T,H}(\theta) = (1/H)\sum_{h=1}^{H} \beta_T^{(h)}(\theta).$$

The exact generalization of what we did in Section 1 amounts to defining the binding function $b_T(\theta)$ as the probability limit (w.r.t. the random draw of the process ε) of the sequence $\beta_{T,H}(\theta)$ when H goes to infinity. However, for most non-linear models, the instrumental estimators $\beta_T^{(h)}(\theta)$ are not reliable for finite T but only for a sample size T going to infinity. It is then worth realizing that when T goes to infinity, for any given h=1, …, H, $\beta_T^{(h)}(\theta)$ should tend towards the so-called *asymptotic binding* function $b(\theta)$ which is also the limit of the *finite sample binding* function $b_T(\theta)$.
Therefore, as far as consistency of estimators when T goes to infinity is concerned, a large number H of simulations is not necessary and we will define more generally an *indirect estimator* $\hat{\theta}_T$ as solution of a minimum distance problem:

$$\mathrm{Min}_\theta [\hat{\beta}_T - \beta_{T,H}(\theta)]' \Omega_T [\hat{\beta}_T - \beta_{T,H}(\theta)] \quad (7)$$

where Ω_T is a positive definite matrix converging towards a deterministic positive definite matrix Ω. In case of a completed Monte Carlo integration (H large) we end up with an approximation of the exact binding function-based estimation:

$$\mathrm{Min}_\theta [\hat{\beta}_T - b_T(\theta)]' \Omega_T [\hat{\beta}_T - b_T(\theta)] \quad (8)$$

which generalizes the bias-correction procedure of Section 1. As in Section 1, we may expect good finite sample properties of such an indirect estimator since, intuitively, the finite sample bias is similar in the two quantities, which are matched against each other and thus should cancel out in the matching process.
In terms of asymptotic theory, the main results under standard regularity conditions (see Gourieroux, Monfort and Renault, 1993) are:

1. (i) the indirect inference estimator $\hat{\theta}_T$ converges towards the true unknown value θ^0 insofar as the asymptotic binding function identifies it:

$$b(\theta) = b(\theta^0) \Rightarrow \theta = \theta^0;$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

2. (ii) the indirect inference estimator $\hat{\theta}_T$ is asymptotically normal insofar as the asymptotic binding function first order identifies the true value:

$$\frac{\partial b}{\partial \theta'}(\theta^0) \text{ is of full-column rank;}$$

3. (iii) we get an indirect inference estimator with a minimum asymptotic variance if and only if the limit-weighting matrix is proportional to the inverse of the asymptotic variance Σ_H of $\sqrt{T}[\hat{\beta}_T - \beta_{T,H}(\theta^0)]$;
4. (iv) the asymptotic variance of the efficient indirect inference estimator is the inverse of

$$\frac{\partial b'}{\partial \theta}(\theta^0)(\Sigma_H)^{-1}\frac{\partial b}{\partial \theta'}(\theta^0)$$

with

$$\Sigma_H = \left[1 + \frac{1}{H}\right]\Sigma_\infty.$$

An implication of these results is that, as far as asymptotic variance of the indirect inference estimator is concerned, the only role of a finite number *H* of simulations is to multiply the optimal variance (obtained with *H*=∞) by a factor (1+1/*H*). Actually, when computing the indirect inference estimator (7), one may be reluctant to use a very large *H* since it involves, for each value of θ along a minimization algorithm, computing *H* instrumental estimators $\beta_T^{(h)}(\theta),\ h = 1, \ldots, H$. We will see in Section 3 several ways to replace these *H* computations by only one. However, this will come at a price, which is the probable loss of the nice finite sample properties of (7) and (8).
As a conclusion, let us stress that indirect inference is able, beyond finite sample biases, to correct for any kind of misspecification bias. The philosophy of this method is basically to estimate a simple model, possibly wrong, to get easily an instrumental estimator $\hat{\beta}_T$ while a direct estimation of structural parameters θ would have been a daunting task. Therefore what really matters is to use an instrumental parameter that captures the key features of the parameters of interest, while being much simpler to estimate. For instance, Pastorello, Renault and Touzi (2000) and Engle and Lee (1996) have proposed to first estimate a GARCH model as an instrumental model to indirectly recover an estimator of the structural model of interest, a stochastic volatility model much more difficult to estimate directly. Other natural examples are models with latent variables such that an observed variable provides a convenient proxy. An estimator based on this proxy suffers from a misspecification bias, but we end up with a consistent estimator by applying the indirect inference matching. Examples of this approach are:

- (i) Pastorello, Renault and Touzi (2000), who use Black and Scholes implied volatilities as a proxy of realizations of the latent spot volatility process.
- (ii) Li (2006), who, following a suggestion of Renault (1997), uses observed bids in an auction market as a proxy of latent private values.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

3 Simulated method of moments

SMM, as introduced by Ingram and Lee (1991) and Duffie and Singleton (1993), is the simulation-based counterpart of GMM to take advantage of the informational content of some conditional moment restrictions:

$$E\left\{K[y_t, z(1,t)] \mid z(1,t)\right\} = k[z(1,t); \theta^0]. \tag{9}$$

The role of simulations in this context is to provide a Monte Carlo assessment of the population conditional moment function $k[z(1,t); \theta]$ when it is not easily available in closed form. Typically, with $y_t^h(\theta)$ drawn as above for h=1, …, H (model (1) or (2)), a convenient Monte Carlo counterpart is: $(1/H)\Sigma_{h=1}^{H} K[y_t^h(\theta), z(1,t)]$. Even though we will mainly present SMM in this simple setting, two possible extensions are worth mentioning.

1. First, in dynamic settings, one may want to consider conditional moment restrictions given not only past and current exogenous variables but also past endogenous variables:

$$E\left\{K[y_t, z(1,t), y(0,t-1)] \mid z(1,t), y(0,t-1)\right\} = k[z(1,t), y(0,t-1); \theta^0].$$

SMM can still be applied to this kind of dynamic moment insofar as one is able to draw simulated values $y_t^h(\theta)$ in the conditional probability distribution (corresponding to the value θ of parameters) of y_t given $x_t = [z(1,t), y(0,t-1)]$. In other words, we need conditional simulations and not only path simulations. As explained above, such conditional simulations are not feasible when the structural model is only defined through a recursive form (2). By contrast, either path simulations or conditional simulations work for static moment conditions (9).

2. Second, the introduction of latent variables paves the way for many other Monte Carlo assessments of the population expectation $E\{K[y_t, z(1,t)] \mid z(1,t)\}$. Let us assume to simplify that $y_t = r_1(y_t^*)$ with i.i.d. latent variables y_t^* endowed with a probability distribution with fixed support (independent of the unknown parameters θ). Of course, the density function $f[y_t^* \mid z(1,t); \theta]$ does depend on θ but one may also pick, as a sampling tool called importance function, another distribution with a given density function $\varphi(u_t)$ on the same support. Then, instead of assessing the population expectation with its naive Monte Carlo counterpart

$$(1/H)\sum_{h=1}^{H} K[r_1(y_t^{*h}(\theta)), z(1,t)].$$

one may prefer to resort to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$(1/H)\sum_{h=1}^{H} K[r_1(u_t^h), z(1,t)]\frac{f[u_t^h|z(1,t);\theta]}{\phi(u_t^h)}$$

where the u_t^h, h=1, …, H, are independently drawn from the distribution with density function $\varphi(u_t)$. This kind of *importance sampling* may be helpful, for instance, for removing some nasty lack of smoothness with respect to the unknown parameters.

As far as static conditional moment restrictions like (9) are concerned, the natural way to extend GMM with a Monte Carlo assessment of the population moment is to minimize with respect to the unknown parameters θ a norm of the sample mean of:

$$Z_t\left\{K[y_t, z(1,t) - (1/H)\sum_{h=1}^{H} K\left[y_t^h(\theta), z(1,t)\right]\right\}$$

where Z_t is a matrix of chosen instruments, that is a fixed matrix function of $z(1,t)$. It is then clear that the minimization programme which is considered is a particular case of (7) above with:

$$\hat{\beta}_T = (1/T)\sum_{t=1}^{T} Z_t K[y_t, z(1,t)]$$

and $\beta_{T,H}(\theta)$ defined accordingly. In other words, we reinterpret SMM as a particular case of *indirect inference*, when the instrumental parameters to match are simple moments rather than themselves defined through some structural interpretations. Note, however, that the moment conditions for SMM could be slightly more general since the function $K[y_t, z(1,t)]$ itself could depend on the unknown parameters θ. In any case, the general asymptotic theory sketched above for SII is still valid. It may be a little more involved when using importance sampling, since then the variance of the simulator no longer coincides with the variance of the initial moments, and then computing the asymptotic variance of the SMM estimator is no longer simply akin to multiplying standard formulas by a factor [1+1/H]. However, we still note that the number H of simulated paths does not need to be large for getting consistent and rather accurate estimators.

In contrast with general SII as presented above, an advantage of SMM is that the instrumental parameters to match, as simple moments, are in general easier to compute than estimated auxiliary parameters $\beta_T^{(h)}(\theta)$, h=1, …, H, derived from some computationally demanding *extremum* estimation procedure. Gallant and Tauchen (1996) have taken advantage of this remark to propose a practical computational strategy for implementing indirect inference when the estimator $\hat{\beta}_T$ of the instrumental parameters is obtained as an M-estimator solution of:

$$\mathrm{Max}_{\beta}(1/T)\sum_{t=1}^{T} q_t[y(0,t), z(1,t);\beta].$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The key idea is then to define the moments to match through the (pseudo)-score vector of this M-estimator. Let us denote

$$K[y(0,t), z(1,t); \beta] = \frac{\partial q_t}{\partial \beta}[y(0,t), z(1,t); \beta]$$

and consider an SMM estimator of θ obtained as a minimizer of the norm of a sample mean of:

$$K[y(0,t), z(1,t); \hat{\beta}_T] - (1/H)\sum_{h=1}^{H} K[y^h(0,t)(\theta), z(1,t); \hat{\beta}_T].$$

For a suitable GMM metric, such a minimization defines a so-called *simulated-score matching* estimator $\hat{\theta}_T$ of θ. In the spirit of Gallant and Tauchen (1996), the objective function that defines the initial estimator $\hat{\beta}_T$ is typically the log-likelihood of some auxiliary model. However, this feature is not needed for the validity of the asymptotic theory sketched below. Several remarks are in order:

1. 1. By contrast with a general SMM criterion, the minimization above does not involve the choice of any instrumental variable. Typically, over-identification will be achieved by choosing an auxiliary model with a large number of instrumental parameters β rather than by choosing instruments.
2. 2. By definition of $\hat{\beta}_T$, the sample mean of $K[y(0,t),z(1,t); \beta]$ takes the value zero for $\beta = \hat{\beta}_T$. In other words, the minimization programme above amounts to:

$$\mathrm{Min}_{\theta}\left\| (1/TH)\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\partial q_t}{\partial \beta}[y^h(0,t)(\theta), z(1,t); \hat{\beta}_T]\right\|_{\Omega_T} \quad (10)$$

   for a suitable GMM metric Ω_T.
3. 3. It can be shown (see Gourieroux, Monfort and Renault, 1993) that under the same assumptions as for the asymptotic theory of SII, the *score-matching* estimator is consistent asymptotically normal. We get a score-matching estimator with a minimum asymptotic variance if and only if the limit-weighting matrix Ω is proportional to the inverse of the asymptotic conditional variance of

$$\sqrt{T}\sum_{t=1}^{T}\frac{\partial q_t}{\partial \beta}[y(0,t), z(1,t); b(\theta^0)]$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

given the exogenous variables z. Then the resulting *efficientscore-matching estimator*is asymptotically equivalent to the *efficient indirect inference estimator*.

4. 4. Owing to this asymptotic equivalence, the score-matching estimator can be seen as an alternative to the efficient SII estimator characterized in Section 2. This alternative is often referred to as *efficient method of moments* (EMM) since, when $q_t[y(0,t),z(1,t);\beta]$ is the log-likelihood $f^a[y_t|z(1,t),y(0,t-1);\beta]$ of some auxiliary model, the estimator is as efficient as maximum likelihood under correct specification of the auxiliary model. More generally, the auxiliary model is designed to approximate the true data generating process as closely as possible and Gallant and Tauchen (1996) propose the *semi-nonparametric* (SNP) modelling to this end. These considerations and the terminology EMM should not lead us to believe that *score-matching* is more efficient than *indirect inference*. The two estimators are asymptotically equivalent even though the score-matching approach makes more transparent the required spanning property of the auxiliary model to reach the Cramer Rao efficiency bound of the structural model.
5. 5. Another alleged advantage of the *score-matching* with respect to *parameter-matching* in SII is its low computational cost. The fact is that with a large number of instrumental parameters β, as will typically be the case with a SNP auxiliary model, it may be costly to maximize H times the log-likelihood of the auxiliary model (for each value of θ along an optimization algorithm) with respect to β to compute $\beta_T^{(h)}(\theta)$, h=1, …, H. By contrast, the programme (10) minimizes only once the norm of a vector of derivatives with respect to β. One must realize, however, that not only is this cheaper computation likely to lose the expected nice finite sample properties of SII put forward in the previous section, but also that the point is not really about a choice between matching (instrumental) parameters β or matching the (instrumental) score $\sum_{t=1}^{T}\frac{\partial q_t}{\partial \beta}$. The key issue is rather the way to use H simulated paths of size T each as explained below.
6. 6. It is worth realizing that the sum of TH terms considered in the definition (10) of the score-matching estimator is akin to consider only one simulated path $y^1(0,TH)(\theta)$ of size TH built from random draws as above (conditional simulations or path simulations) from a fictitious path $z^*(1,TH)$ of exogenous variables defined in the following way:

$$z_1^* = z_1, \ldots, z_T^* = z_T, z_{T+1}^* = z_1, \ldots, z_{2T}^* = z_T, z_{2T+1}^* = z_1, \ldots, z_{TH}^* = z_T.$$

If for instance (z_t) is Markov of order 1, such a fictitious path is a correct draw except possibly for Hvalues, which is immaterial when Tgoes to infinity. From such a simulated path, estimation of instrumental parameters would have produced a vector $\beta_{TH}^{(1)}(\theta)$that could have been used for *indirect inference*, that is to define an estimator $\hat{\theta}_T$solution of:

$$\text{Min}_\theta[\hat{\beta}_T - \beta_{TH}^{(1)}(\theta)]'\Omega_T[\hat{\beta}_T - \beta_{TH}^{(1)}(\theta)]. \quad (11)$$

This *parameter-matching*estimator is not more computationally demanding than the corresponding *score-matching*estimator computed from the same simulated path as solution of:

$$\text{Min}_\theta\left\|(1/TH)\sum_{t=1}^{TH}\frac{\partial q_t}{\partial \beta}[y^1(0,t)(\theta), z^*(1,t);\hat{\beta}_T]\right\|\Omega_T.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(12)

Actually, (11) and (12) are numerically identical in the case of just-identification (dim θ =dim β). Then the choice of a GMM metric is immaterial and both estimators are basically the solution $\hat{\theta}_T$ of:

$$(1/TH)\sum_{t=1}^{TH}\frac{\partial q_t}{\partial \beta}[y^1(0,t)(\hat{\theta}_T), z^*(1,t); \hat{\beta}_T] = 0.$$

More generally, the four estimators (7), (10), (11) and (12) are asymptotically equivalent when *T* goes to infinity and the GMM weighting matrix is efficiently chosen accordingly. However, it is quite obvious that only (7) performs the right finite sample bias correction by matching instrumental parameters values estimated on both observed and simulated paths of lengths *T*. The trade-off is thus between giving up finite sample bias correction or paying the price for computing *H* estimated instrumental parameters.

4 Simulated M-estimators

Even though well-chosen moments to match may allow one to get accurate estimators, it is somewhat contradictory to resort to a fully parametric model to perform simulations needed for SMM while Hansen's (1982) GMM was semi-parametric in spirit. *Simulated maximum likelihood* (SML) methods aim at exploiting the whole parametric structure for efficient estimation. The key role of simulations would then be to provide an unbiased simulator of each conditional p.d.f. $f[y_t|z(1,t), y(0,t-1); \theta]$ (also denoted $f[y_t|x_t; \theta]$) because it is not available in closed form. This is typically the case in a model with latent variables, and then conditioning may provide a convenient simulator. The simplest example comes from model (2) when latent variables are exogenous. Let us write it without observed exogenous variables for sake of expositional simplicity:

$$y_t = r_1[y(0,t-1), y^*(0,t), \varepsilon_{1t}; \theta], \quad t = 1, \ldots, T$$
$$y_t^* = r_2[y^*(0,t-1), \varepsilon_{2t}; \theta], \quad t = 1, \ldots, T.$$

(13)

Then $f[y_t|x_t; \theta]$ is nothing but the expectation with respect to the probability distribution of $y^*(0,t)$ of $f[y_t|x_t, y^*(0,t); \theta]$. While the latter is easily deduced from the pdf of ε_{1t} through the measurement equation (the first equation above), the former is in general easy to compute from the evolution equation (the second equation above). Then Monte Carlo integration provides an unbiased estimate of $f[y_t|x_t; \theta]$ with $(1/H)\Sigma_{h=1}^{H} f[y_t|x_t, y^{*h}(0,t)(\theta); \theta]$ where $y^{*h}(0,t)(\theta)$, *h*=1, …, *H*, are independent draws obtained from independent draws in the known distribution of ε_{2t}. Of course, importance sampling must also be relevant in this context. Moreover, for a general model (2), when *y* does cause y^*, this naive approach may be very inefficient (in terms of speed of convergence of the variance towards zero) and one may want to refer to either accelerated versions of importance sampling (Danielsson and Richard, 1993), simulated expectation maximization algorithm (SEM, see for example Shephard, 1993) and other more sophisticated simulators, which are beyond the scope of this article.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Irrespective of the choice of the simulator, the main difficulty with SML is that the logarithm of an unbiased simulator of the likelihood is not an unbiased simulator of the log-likelihood. Therefore, in contrast with SMM, SML is consistent only when both H and T go to infinity. Note, however, that asymptotic bias corrections are possible (see Gourieroux and Monfort, 1996). Under standard regularity conditions, SML is asymptotically efficient insofar as T goes to infinity slower than H^2. Another version of the SML method is the *simulated score* method (see Hajivassiliou, 1993). This is basically a version of SMM where the latent score is used as a simulator for the score of the model on observables, while the later is not tractable analytically. This method should not be confused with *simulated score matching* of the former section, where the score to match was not something to simulate to get it in closed form.

Finally, it is worth mentioning that simulation-based estimation methods are not always to be recommended, even in some fully parametric modelling situations with latent variables making standard maximum likelihood unfeasible. Such situations typically occur with structural econometric models where equilibrium of a market or of a game implies a deterministic relationship between latent variables and observed ones. Then, not only is the above SML theory no longer valid but any other simulation based method will in general be highly inefficient because the dependence between the two blocks of unknown ‘parameters’, namely, structural parameters and latent variables, is sharp. In such contexts, some authors have, however, considered some simulated non-linear least squares methods (Laffont, Ossard and Vuong, 1995 for auction markets) or more generally simulated pseudo-likelihood methods (Laroque and Salanie, 1993). While the focus of interest of Laroque and Salanie (1993) on a disequilibrium model raised more involved issues due to non-differentiability, it is rather clear that some implied state GMM methods (Pan, 2002; Pastorello, Patilea and Renault, 2003) are more efficient than SMM in the contexts of smooth equilibrium relationships like those produced by option pricing theory. The key issue is to take advantage of the deterministic relationship between y_t and y_t^* (for a given value of the parameters θ) to track what would have been an efficient estimation if latent variables had been observed.

5 Concluding remarks

The econometrician's search for a well-specified parametric model (‘quest for the Holy Grail’ as stated by Monfort, 1996) and associated efficient estimators even remain popular when MLE becomes intractable due to highly non-linear dynamic structure including latent variables. The efficiency properties of SML, EMM and more generally of SMM and SII when the set of instrumental parameters to match is sufficiently large to span the likelihood scores are often advocated as if the likelihood score is well specified. However, the likely misspecification of the structural model requires a generalization of the theory of SII as recently proposed by Dridi, Guay and Renault (2007). As for MLE with misspecification (see White, 1982; Gourieroux, Monfort and Trognon, 1984) such a generalization entails two elements.

First, asymptotic variance formulas are complicated by the introduction of sandwich formulas. Ignoring this kind of correction is even more detrimental than for QMLE since two types of sandwich formulas must be taken into account, one for the data generating process (DGP) and one for the simulator (based either on model (1) or model (2)) which turns out to be different from the DGP in case of misspecification.

Secondly, and even more importantly, misspecification may imply that we consistently estimate a pseudo-true value, which is poorly related to the true unknown value of the parameters of interest. Dridi, Guay and Renault (2007) put forward the necessary (partial) encompassing property of the instrumental model (through instrumental parameters β) by the structural model (with parameters θ) needed to ensure consistency towards true values of (part of) the components of θ in spite of misspecification. The key issue is that, since structural parameters are recovered from instrumental ones by inverting a binding function $\beta = b(\theta)$, all components are interdependent. The requirement of encompassing typically means that, if one does not want to proceed under the maintained assumption that the structural model (1) or (2) is true, one must be parsimonious with respect to the number of moments to match or more generally to the scope of empirical evidence that is captured by the instrumental parameters β. In other words, robustness to misspecification requires an instrumental model choice strategy opposite to that commonly used for a structural model: the larger the instrumental model, the larger the risk of contamination of the estimated structural parameters of interest by what is wrong in the structural model. Of course, there is no such thing as a free lunch: robustness to misspecification through a parsimonious and well-focused instrumental model comes at the price of efficiency loss. Efficiency loss means not only lack of accuracy of estimators of structural parameters but also lack of power of specification tests.

See Also

- bootstrap
- Markov chain Monte Carlo methods
- simulation estimators in macroeconometrics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- state space models
- stochastic volatility models

Bibliography

Andrews, D. 1993. Exactly median unbiased estimation of first order autoregressive/unit root models. *Econometrica* 61, 139–65.

Danielsson, J. and Richard, J.F. 1993. Accelerated Gaussian importance sampler with application to dynamic latent variables models. *Journal of Applied Econometrics* 8, S153–73.

Dridi, R., Guay, A. and Renault, E. 2007. Indirect inference and calibration of dynamic stochastic general equilibrium models. *Journal of Econometrics* 136, 397–430.

Duffie, D. and Singleton, K. 1993. Simulated moments estimation of Markov models of asset prices. *Econometrica* 61, 929–52.

Engle, R.F. and Lee, G.G.J. 1996. Estimating diffusion models of stochastic volatility. In *Modeling Stock Market Volatility: Bridging the Gap to Continuous Time*, ed. P.E. Rossi. New York: Academic Press.

Gallant, A.R. and Tauchen, G. 1996. Which moments to match? *Econometric Theory* 12, 657–81.

Gourieroux, C. and Monfort, A. 1995. Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* 11, 195–228.

Gourieroux, C. and Monfort, A. 1996. Simulation-based econometric methods. *Core Lectures*. Oxford: Oxford University Press.

Gourieroux, C., Monfort, A. and Renault, E. 1993. Indirect inference. *Journal of Applied Econometrics* 8, S85–118.

Gourieroux, C., Monfort, A. and Trognon, A. 1984. Pseudo-maximum likelihood methods theory. *Econometrica* 52, 681–700.

Gourieroux, C., Renault, E. and Touzi, N. 2000. Calibration by simulation for small sample bias correction. In *Simulation-Based Inference in Econometrics*, ed. R. Mariano, T. Schuerman and M.J. Weeks. Cambridge: Cambridge University Press.

Hajivassiliou, V.A. 1993. Simulation estimation methods for limited dependent variable models. In *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao and H. Vinod. Amsterdam: North-Holland.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer.

Hansen, L.P. 1982. Large sample properties of generalized method of moments. *Econometrica* 50, 1029–54.

Horowitz, J.L. 1997. Bootstrap methods in econometrics: theory and numerical performance. In *Advances in Econometrics, seventh world congress*, ed. D. Kreps and K. Wallis. Cambridge: Cambridge University Press.

Ingram, B.F. and Lee, B.S. 1991. Estimation by simulation of time series models. *Journal of Econometrics* 47, 197–207.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Laffont, J.J., Ossard, H. and Vuong, Q. 1995. Econometrics of first-price auction. *Econometrica* 63, 953–80.

Laroque, G. and Salanie, B. 1993. Simulation based estimation of models with lagged latent variables. *Journal of Applied Econometrics* 8, S119–33.

Li, T. 2006. Indirect inference in structural econometric models. *Journal of Econometrics*, forthcoming.

Monfort, A. 1996. A reappraisal of misspecified econometric models. *Econometric Theory* 12, 597–619.

Monfort, A. and Van Dijk, H.K. 1995. Simulation-based econometrics. In *Econometric Inference Using Simulation Techniques*, ed. H.K. Van Dijk, A. Monfort and B.W. Brown. New York: Wiley.

Pan, J. 2002. The jump-risk premia implicit in options: evidence from an integrated time-series study. *Journal of Financial Economics* 63, 3–50.

Pastorello, S., Patilea, V. and Renault, E. 2003. Iterative and recursive estimation in structural non-adaptive models. Invited Lecture with discussion. *Journal of Business, Economics and Statistics* 21, 449–509.

Pastorello, S., Renault, E. and Touzi, N. 2000. Statistical inference for random-variance option pricing. *Journal of Business and Economic Statistics* 18, 358–67.

Renault, E. 1997. Econometric models of option pricing errors. In *Advances in Econometrics, Seventh World Congress*, ed. D. Kreps and K. Wallis. Cambridge: Cambridge University Press.

Shephard, N. 1993. Fitting nonlinear time series with applications to stochastic variance models. *Journal of Applied Econometrics* 8, S134–52.

Smith, A. 1993. Estimating nonlinear time series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8, S63–84.

White, H. 1982. Maximum likelihood estimation of mis-specified models. *Econometrica* 50, 1–25.

How to cite this article

Renault, Eric. "simulation-based estimation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000470> doi:10.1057/9780230226203.1532

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

simultaneous equations models

Thomas J. Rothenberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

distributed lags; econometrics; endogeneity and exogeneity; recursive models; reduced-form equations; simultaneous equations models; structural equations

Article

Models that attempt to explain the workings of the economy typically are written as interdependent systems of equations describing some hypothesized technological and behavioural relationships among economic variables. Supply and demand models, Walrasian general equilibrium models, and Keynesian macromodels are common examples. A large part of econometrics is concerned with specifying, testing and estimating the parameters of such systems. Despite their common use, simultaneous equations models still generate controversy. In practice there is often considerable disagreement over their proper use and interpretation.

In building models economists distinguish between *endogenous* variables which are determined by the system being postulated and *exogenous* variables which are determined outside the system. Movements in the exogenous variables are viewed as autonomous, unexplained causes of movements in the endogenous variables. In the simplest systems, each of the endogenous variables is expressed as a function of the exogenous variables. These so-called ‘reduced-form’ equations are often interpreted as causal, stimulus–response relations. A hypothetical experimental is envisaged where conditions are set and an outcome occurs. As the conditions are varied, the outcome also varies. If the outcome is described by the scalar endogenous variable y and the conditions by the vector of exogenous variables x, then the rule describing the causal mechanism can be written as $y=f(x)$. If there are many outcomes of the experiment, y and f are interpreted as vectors; the rule describing how the ith outcome is determined can be written as $y_i=f_i(x)$.

Most equations arising in competitive equilibrium theory are motivated by hypothetical stimulus–response experiments. Demand curves, for example, represent the quantity people will purchase when put in a price-taking market situation. The conditions of the experiment are, in addition to price, all the other determinants of demand. In any given application, most of these determinants are viewed as fixed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
as the experiment is repeated; attention is directed at the handful of exogenous variables whose effects are being analysed. In an *n* good world, there are *n* such equations, each determining the demand for one of the goods as a function of the exogenous variables.

Reduced-form models where each equation contains only one endogenous variable are rather special. Typically, economists propose interdependent systems where at least some of the equations contain two or more endogenous variables. Such models have a more complex causal interpretation since each endogenous variable is determined not by a single equation but *simultaneously* by the entire system. Moreover, in the presence of simultaneity, the usual least-squares techniques for estimating parameters often turn out to have poor statistical properties.

Why simultaneity?

Given the obvious asymmetry between cause and effect, it would at first thought appear unnatural to specify a behavioural economic model as an interdependent, simultaneous system. Although equations with more than one endogenous variable can always be produced artificially by algebraic manipulation of a reduced-form system, such equations have no independent interpretation and are unlikely to be interesting. It turns out, however, that there are many situations where equations containing more than one endogenous variable arise quite naturally in the process of modelling economic behaviour. These so-called ‘structural’ equations have interesting causal interpretations and form the basis for policy analysis. Four general classes of examples can be distinguished.

1. Suppose two experiments are performed, the outcome of the first being one of the conditions of the second. This might be represented by the two equations y1=$f_1(x)$ and $y_2=f_2(x, y_1)$. In this two-step causal chain, both equations have simple stimulus–response interpretations. Implicit, of course, is the assumption that the experiment described by the first equation takes place before the experiment described by the second equation. Sequential models where, for example, people choose levels of schooling and later the market responds by offering a wage are of this type. Such *recursive* models are only trivially simultaneous and raise no conceptual problems although they may lead to estimation difficulties.

2. Nontrivial simultaneous equations systems commonly arise in multi-agent models where each individual equation represents a separate hypothetical stimulus–response relation for some group of agents, but the outcomes are constrained by equilibrium conditions. The simple competitive supply–demand model illustrates this case. Each consumer and producer behaves as though it has no influence over price or over the behaviour of other agents. Market demand is the sum of each consumer's demand and market supply is the sum of each producer's supply, with all agents facing the same price. Although the market supply and demand functions taken separately represent hypothetical stimulus–response situations where quantity is endogenous and price is exogenous, in the combined equilibrium model both price and quantity are endogenous and determined so that supply equals demand.

Most competitive equilibrium models can be viewed as (possibly very complicated) variants of this supply–demand example. The individual equations, when considered in isolation, have straightforward causal interpretations. Groups of agents respond to changes in their environment. Simultaneity results from market clearing equilibrium conditions that make the environments endogenous. Keynesian macromodels have a similar structure. The consumption function, for example, represents consumers’ response to their (seemingly) exogenous wage income – income that in fact is determined by the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
condition that aggregate demand equal aggregate supply.
3. Models describing optimizing behaviour constitute a third class of examples. Suppose an economic agent is faced with the problem of choosing some vector y in order to maximize the function $F(y, x)$, where x is a vector of exogenous variables outside the agent's control. The optimum value, denoted by y^*, will depend on x. If there are G choice variables, the solution can be written as a system of G equations, $y^*=g(x)$. If F is differentiable and globally concave in y, the solution can be obtained from the first-order conditions $f(y^*, x)=0$, where f is the G-dimensional vector of partial derivatives of F with respect to y. The two sets of equations are equivalent representations of the causal mechanism. The first is a reduced-form system with each endogenous variable expressed as a function of exogenous variables alone. The second representation consists of a system of simultaneous equations in the endogenous variables. These latter equations often have simple economic interpretations such as, for example, setting marginal product equal to real input price.
4. Models obtained by simplifying a larger reduced-form system are a fourth source of simultaneous equations. The Marshallian long-run supply curve, for example, is often thought of as the locus of price–quantity pairs that are consistent with the marginal firm having zero excess profit. Both price and quantity are outcomes of a complex dynamic process involving the entry and exit of firms in response to profitable opportunities. If, for the data at hand, entry and exit are in approximate balance, the reduced-form dynamic model may well be replaced by a static interdependent equilibrium model.
This last example suggests a possible reinterpretation of the equilibrium systems given earlier. It can be argued (see, for example, Wold, 1954) that multi-agent models are necessarily recursive rather than simultaneous because it takes time for agents to respond to their environments. From this point of view, the usual supply–demand model is a simplification of a considerably more complex dynamic process. Demand and supply in fact depend on *lagged* prices; hence current price need not actually clear markets. However, the existence of excess supply or demand will result in price movement which in turn results in a change in consumer and producer behaviour next period. When time is explicitly introduced into the model, simultaneity disappears and the equations have simple causal interpretations. But, if response time is short and the available data are averages over a long period, excess demand may be close to zero for the available data. The static model with its simultaneity may be viewed as a limiting case, approximating a considerably more complex dynamic world. This interpretation of simultaneity as a limiting approximation is implicit in much of the applied literature and is developed formally in Strotz (1960).

The need for structural analysis

These examples suggest that systems of simultaneous equations appear quite naturally when constructing economic models. Before discussing further their logic and interpretation, it will be useful to develop some notation. Let y be a vector of G endogenous variables describing the outcome of some economic process; let x be a vector of K ‘predetermined’ variables describing the conditions that determine those outcomes. (In dynamic models, lagged endogenous variables as well as the exogenous variables will be considered as conditions and included in x.) By a simultaneous equations model we mean a system of m equations relating y and x:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$g_i(y, x) = 0 \quad (i = 1, \ldots, m).$$

In the important special case where the functions are linear, the system can be written as the vector equation

$$By + \Gamma x = 0, \tag{1}$$

where B is an $m \times G$ matrix of coefficients, Γ is an $m \times K$ matrix of coefficients, and 0 is an m-dimensional vector of zeros. (The intercepts can be captured in the matrix Γ if we follow the convention that the first component of x is a 'variable' that always takes the value one.) A complete system occurs when $m=G$ and B is non-singular. Then the vector of outcome variables can be expressed as a linear function of the condition variables.

$$y = -B^{-1}\Gamma x = \Pi x. \tag{2}$$

Although the logic of the analysis applies for arbitrary models, the main issues can most easily be illustrated in this case of a complete linear system.
If both sides of the vector equation (1) are premultiplied by any $G \times G$ nonsingular matrix F, a new representation of the model is obtained, say

$$B^* y + \Gamma^* x = 0 \tag{3}$$

where $B^*=FB$ and $\Gamma^* = F\Gamma$. If F is not the identity matrix, the systems (1) and (3) are not identical. Yet if one representation is 'true' (that is, the real world observations satisfy the equation system) then the other is also 'true'. Which of the infinity of possible representations should we select? The obvious answer is that it does not matter. Any linear combination of equations is another valid equation. Any nonsingular transformation is as good as any other. For simplicity, one might as well choose the solved reduced form given by eq. (2).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In practice, however, we are not indifferent between the various equivalent representations. There are a number of reasons for that. First, it may be that some representations are easier to interpret or easier to estimate. The first-order conditions for a profit maximizing firm facing fixed prices may, depending on the production function, be much simpler than the reduced form. Secondly, if we contemplate using the model to help analyse a changed regime, it is useful to have a representation in which the postulated changes are easily described. This latter concern leads to the concept of the degree of *autonomy* of an equation.

In the supply–demand model, it is easy to contemplate changes in the behaviour of consumers that leave the supply curve unchanged. For example, a shift in tastes may modify demand elasticities but have no effect on the cost conditions of firms. In that case, the supply curve is said to be autonomous with respect to this intervention in the causal mechanism. Equations that combine supply and demand factors (like the reduced form relating quantity traded to the exogenous variables) are not autonomous. The analysis of policy change is greatly simplified in models where the equations possess considerable autonomy. If policy changes one equation and leaves the other equations unchanged, its effects on the endogenous variables are easily worked out. Comparative static analysis as elucidated, for example, by Samuelson (1947) is based on this idea. Indeed, the power of the general equilibrium approach to economic analysis lies largely in its separation of the behaviour of numerous economic agents into autonomous equations.

As emphasized by the pioneers in the development of econometrics, it is not enough to construct models that fit a given body of facts. The task of the economist is to find models that successfully predict how the facts will change under specified new conditions. This requires knowing which relationships will remain stable after the intervention and which will not. It requires the model builder to express for every equation postulated the class of situations under which it will remain valid. The concept of autonomy and its importance in econometric model construction is emphasized in the classic paper by Haavelmo (1944) and in the expository paper by Marschak (1953). Sadly, it seems often to be ignored in applied work.

The autonomy of the equations appearing in commonly proposed models is often questionable. Lucas (1976) raises some important issues in his critique of traditional Keynesian macromodels. These simultaneous equations systems typically contain distributed lag relations which are interpreted as proxies for expectations. Suppose, for example, consumption really depends on expected future income. If people forecast the future based on the past, the unobserved expectation variable can be replaced by some function of past incomes. However, since income is endogenous, the actual time path of income depends on all the equations of the model. Under rational expectations, any change in the behaviour of other agents or in technology that affects the time path of income will also change the way people forecast and hence the distributed lag. Thus it can be argued that the traditional consumption function is not an autonomous relation with respect to most interesting policy interventions. Sims (1980) pursues this type of argument further, finding other reasons for doubting the autonomy of the equations in traditional macroeconomic models and concluding that policy analysis based on such models is highly suspect. Although one may perhaps disagree with Sims's conclusion, the methodological questions he raises cannot be ignored.

Even if one accepts the view that structural equations actually proposed in practice often possess limited autonomy and are not likely to be invariant to many interesting interventions, it may still be the case that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

these equations are useful. A typical reduced-form equation contains all the predetermined variables in the system. Given the existence of feedback, it is hard to argue a priori about the numerical values of the various elements of Π. It may be much easier to think about orders of magnitude for the structural coefficients. Our intuition about the behaviour of individual sectors of the economy is likely to be considerably better than our intuition about the general equilibrium solution. As long as there are no substantial structural changes, specification in terms of structural equations may be appropriate even if some of the equations lack autonomy.

Some econometric issues

The discussion up to now has been in terms of exact relationships among economic variables. Of course, actual data do not lie on smooth monotonic curves of the type used in our theories. This is partially due to the fact that the experiments we have in the back of our mind when we postulate an economic model do not correspond exactly to any experiment actually performed in the world. Changing price, but holding everything else constant, is hypothetical and never observed in practice. Other factors, which we choose not to model, do in fact vary across our sample observations. Furthermore, we rarely pretend to know the true equations that relate the variables and instead postulate some approximate parametric family. In practice we work with models of the form

$$g_i(x, y, \theta, u_i) = 0 \quad (i = 1, \ldots, m)$$

where the g's are known functions, θ is a vector of unknown parameters, and the u's are unobserved error terms reflecting the omitted variables and the misspecification of functional form. In the special case where the functions are linear in x and y with an additive error, the system can be written as

$$By + \Gamma x = u, \tag{4}$$

where u is a m-dimensional vector of errors. If B is non-singular, the reduced form is also linear and can be written as

$$y = -B^{-1}\Gamma x + B^{-1}u = \Pi x + v. \tag{5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Equation system (4) as it stands is empty of content since, for any value of x and y, there always exists a value of u producing equality. Some restrictions on the error term are needed to make the system interesting. One common approach is to treat the errors as though they were draws from a probability distribution centred at the origin and unrelated to the predetermined variables. Suppose we have T observations on each of the $G+K$ variables, say, from T time periods or from T firms. We postulate that, for each observation, the data satisfy eq. (4) where the parameters B and Γ are constant but the n error vectors $u_1, \ldots, u_T$ are independent random variables with zero means. Furthermore, we assume that the conditional distribution of u_t given the predetermined variables x_t for observation t is independent of x_t. A least-squares regression of each endogenous variable on the set of predetermined variables should then give good estimates of Π; if the sample size is large and there is sufficient variability in the regressors. However, unless the inverse of B contains blocks of zeros, eq. (5) implies that each of the endogenous variables is a function of all the components of u. In general, every element of y will be correlated with all the endogenous variables, if the correlation is small compared with the sample variation in those variables.

The conclusion that structural parameters in interdependent systems cannot be well estimated using least squares is widely believed by econometric theorists and widely ignored by empirical workers. There are probably two reasons for this discrepancy. First, although the logic of interdependent systems suggests that structural errors are likely to be correlated with all the endogenous variation, if the correlation is small compared with the sample variable in those variables, least squares bias will also be small. Given all the other problems facing the applied econometrician, this bias may be of little concern. Secondly, alternative estimation methods that have been developed often produce terrible estimates. Sometimes the only practical alternative to least squares is no estimate at all – a solution that is rarely chosen.

In some applications, the reduced form parameters Π are of primary concern. The structural parameters B and Γ are of interest only to the extent they help us learn about Π. For example, in a supply–demand model, we may wish to know the effect on price of changes in the weather. Price elasticities of supply and demand are not needed to answer that question. On the other hand, if we wish to know the effect of a sales tax on quantity produced, knowledge of the price elasticities might be essential. Although least squares is generally available for reduced-form estimation (at least if the sample is large), it is not obvious that, without further assumptions, good structural estimates are ever attainable. The key assumption of the model is that the G structural errors are uncorrelated with the K predetermined variables. These GK orthogonality assumptions are just enough to determine (say by equating sample moments to population moments) the GK parameters in Π. But the structural coefficient matrices B and Γ contain G^2+GK elements. Even with G normalization rules that set the units in which the parameters of each equation will be measured, there are more coefficients than orthogonality conditions. It turns out that structural estimation is possible only if additional assumptions (for example, that some elements of B and Γ are known a priori) are made. These considerations lead to three general classes of questions that have been addressed by theoretical econometricians: (1) When, in principle, can good structural estimates be found? (2) What are the best ways of actually estimating the structural parameters, given that it is possible? (3) Are there better ways of estimating the reduced-form parameters than by least squares?

These questions are studied in depth in standard econometrics textbooks and will not be examined here. The answers, however, do have a common thread. If each structural equation has more than K unknown

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
parameters, structural estimation is generally impossible and least squares applied to the reduced form is optimal in large samples. If, on the other hand, each structural equation has fewer than *K* unknown coefficients, structural estimation generally is possible and least squares applied to the reduced form is no longer optimal. In this latter situation, various estimation procedures are available, some requiring little computational effort. However, the sample size may need to be quite large before these procedures are likely to give good estimates.

The assumption that the errors are independent from trial to trial is obviously very strong and quite implausible in time-series analysis. If the nature of the error dependence can be modelled and if the lag structure of the dynamic behavioural equations is correctly specified, most of the estimation results that follow under independence carry over. Unfortunately, with small samples, it is usually necessary to make crude (and rather arbitrary) specifications that may result in very poor estimates. Despite the fact that simultaneous equations analysis in practice is mostly applied to time-series data, it can be argued that the statistical basis is much more convincing for cross-section analysis where samples are large and dependency across observations minimal. Even there, the assumption that the errors are unrelated to the predetermined variables must be justified before simultaneous equations estimation techniques can be applied.

The role of simultaneous equations

Many applied economists seem to view the simultaneous equations model as having limited applicability, appropriate only for a very small subset of the problems actually met in practice. This is probably unwise. Estimated regression coefficients are commonly used to explain how an intervention which changes one explanatory variable will affect the dependent variable. Except for very special cases, this interpretation requires us to believe that the proposed intervention will not affect any of the other explanatory variables and that, in the sample, the errors were unrelated to the variation in the regressors. That is, the mechanism that determines the explanatory variables must be unrelated to the causal mechanism described by the equation under consideration. Unless the explanatory variables were in fact set in a carefully designed controlled experiment, viewing the explanatory variables as endogenous and possibly determined simultaneously with the dependent variable is a natural way to start thinking about the plausibility of the required assumptions.

In a sense, the simultaneous equations model is an attempt by economists to come to grips with the old truism that correlation is not the same as causation. In complex processes involving many decision-makers and many decision variables, we wish to discover stable relations that will persist over time and in response to changes in economic policy. We need to distinguish those equations that are autonomous with respect to the interventions we have in mind and those that are not. The methodology of the simultaneous equations model forces us to think about the experimental conditions that are envisaged when we write down an equation. It will not necessarily lead us to good parameter estimates, but it may help us to avoid errors.

See Also

- econometrics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- maximum likelihood

Bibliography

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12, Supplement, July, 1–115.

Lucas, R. 1976. Econometric policy evaluation: a critique. In *The Phillips Curve and Labor Markets*, ed. K. Brunner and A. Meltzer. Carnegie-Rochester Conference Series on Public Policy No. 1. Amsterdam: North-Holland.

Marschak, J. 1953. Economic measurement for policy and prediction. In *Studies in Econometric Method*, ed. W. Hood and T. Koopmans. New York: Wiley.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Strotz, R. 1960. Interdependence as a specification error. *Econometrica* 28, 428–42.

Wold, H. 1954. Causality and econometrics. *Econometrica* 22, 162–77.

How to cite this article

Rothenberg, Thomas J. "simultaneous equations models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000138> doi:10.1057/9780230226203.1534

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

single tax

Mason Gaffney
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

excess burden of taxation; George, H.; impôt unique; land rent; land tax; land value taxation; marginal-cost pricing; Physiocracy; Quesnay, F.; single tax; tax incidence

Article

Single Tax is a generic label for the programme of Henry George and others to socialize land rent by substituting one heavy tax on land value for most other taxes. It is not an adequate descriptor but a slogan that caught on. ‘Land value taxation’ is more used today, especially for a limited tax by a local authority. In Scotland and England ‘taxation of ground values’ and ‘site-value rating’ are used.
Specifically, the ‘Single Tax’ slogan marked a shift in the movement after 1887 as George swung towards the Centre after purging the Marxists from his United Labour Party, losing Powderley and Gompers, and demurring to the quixotic demands of Fr. Edward McGlynn. He was losing Irish-American support from the hostility of Parnell and the Catholic hierarchy.
Thomas Shearman, a corporate lawyer, coined ‘Single Tax’ to differentiate George's free-market, pro-capitalist programme from those of others who had coalesced around him in the radical and protest awakening of 1879–87. Soon it also served to differentiate Georgism from Bellamy nationalism and Bryan inflationism. In Britain it distinguished Georgists from Wallace's land nationalizers, Hyndman's Marxists, Webb's Fabians, and Parnell's and Chamberlain's movements for peasant proprietorship.
A change of emphasis followed. George had originally striven for rent socialization, redistribution and augmented social spending. Critics on the Right saw too much taxation and levelling. The Single Tax slogan emphasized the counterpart benefits of relief from other taxes. Single Taxers would remove state and local property taxes from buildings and movable capital, and lower most transit and utility rates, often to zero, meeting deficits from the rent fund (anticipating the marginal-cost pricing policies elaborated by Hotelling and Vickrey). Critics on the Left now saw too limited a revenue; so did machine Democrats, militarists, public contractors, and of course landholders seeking developmental public works at the general expense.
The heavy national taxes in America were tariffs. With *Protection or Free Trade?* (1886) George

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
attacked them, invoking Quesnay, Ricardo, Cobden and Bright. Two million copies were printed, equalling his earlier *Progress and Poverty*. He supported Grover Cleveland in 1888 and 1892, hoping that free trade would force Congress to turn to land for revenues. Single Taxers in Congress succeeded in having the Income Tax Act of 1894 include land rent and unearned increments in the base, even though that was likely to be the grounds for its being held unconstitutional, and was. Single Tax began to connote tax limitation.

In *Protection or Free Trade?* George had restated the Physiocratic doctrine of tax incidence, while broadening it to include urban land, which most Physiocrats (except Turgot) had oddly excluded. There is only so much taxable surplus to tap under any system, and most of it lodges in land rent. Single Tax was simply the way to tax this surplus without what is now called the excess burden of indirect taxation. This might refute the charge of inadequacy, but the point has not been widely understood, and some still question revenue adequacy.

Shearman invited more such questions when he went another step from the Left with his ‘Single Tax Limited’, the upper limit being two-thirds of economic rent. To some adherents Single Tax is a tax limitation device. George remained a ‘Single Taxer, Unlimited’. He held that taxation is only a means to justice; justice means every infant has an equal right to the Earth, its use and its rents. He remained a populist who supported Bryan in 1896, even though cold to the free silver panacea.

In Scotland and England George was active and well received by the radical wing of the Liberal Party. But Single Tax continued to mean an extreme position which moderates shunned, even when the Liberal Party put a land tax plank in its platform after Gladstone retired in 1895. Liberals under Asquith and Lloyd George introduced a token land tax in their 1909 budget. The Single Tax bogey frightened the Conservative members of the House of Lords into an intransigent obstructionism that alienated the voters and was used to consolidate the power of the Commons, the Liberals and Lloyd George, who then temporized away the land tax. Labour reintroduced a land tax in 1931 under MacDonald and Snowden, but Neville Chamberlain scuttled it finally in 1934. Post-war Labour governments abandoned Single Tax as being too market-oriented.

Henry George died in 1897. Single Tax remained a power for another generation. Leaders like Shearman and Louis Post sought to professionalize the movement and reconcile it with middle-class values, a timely adaptation to the ethos of progress under scientific management. Somers, Pollock and Zangerle professionalized land assessment. Lawson Purdy helped found the National Tax Association. The Progressive and New Freedom movements absorbed many Single Taxers and reflected some of their ideals.

A century earlier at the court of Louis XV, François Quesnay and his Physiocrats had advanced the ‘*impôt unique*’, an even more limited single tax restricted to farm land, and one-third the rent. Like Shearman, Quesnay argued efficiency and laissez-faire, not redistribution: it was the age of enlightened despotism, not of populism. They (and later their disciple Walras) called it ‘co-proprietorship of land by the state’.

But there is a touch of class-levelling inherent in any proposal to tax land directly, however sugar-coated with the doctrine that landholders gain from removing other taxes, however limited by the safeguard of ‘co-proprietorship’. Physiocracy could beguile a despot dreaming of energizing a decadent gentry, and liberating his people from tax farmers and a jumble of enervating excises that weakened France's economy. It was the fate of Quesnay, and of France, that the privileged gentry proved more sensitive to Physiocracy's threat than others were to its benefits.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Quesnay was closer than he knew to an age of populism; he might better have addressed the new constituency. Indirectly, he did through his influence on Jefferson, transmitted through his disciples Pierre Samuel Du Pont and Destutt de Tracy; and through his influence on Smith, Ricardo and Mill, whose special treatment of land rent set the stage for George's Single Tax.

See Also

- George, Henry

Bibliography

Brown, H.G. 1932. *The Economic Basis of Tax Reform*. Columbia, MO: Lucas Bros.

Douglas, R. 1976. *Land, People and Politics*. London: Allison & Busby.

Geiger, G.R. 1936. *The Theory of the Land Question*. New York: Macmillan.

Groenewegen, P.D., ed. and trans. 1977. *The Economics of A.R.J. Turgot*. The Hague: Martinus Nijhoff.

Holland, D.M., ed. 1970. *The Assessment of Land Value*. Madison: University of Wisconsin Press.

Howe, F.C. 1925. *Confessions of a Reformer*. New York: Charles Scribner's Sons.

Lindholm, R.W. and Lynn, A.D., Jr., eds. 1981. *Land Value Taxation in Thought and Practice*. Madison: University of Wisconsin Press.

Lissner, W., ed. 1941. *American Journal of Economics and Sociology*. New York: Robert Schalkenbach Foundation.

Miller, J.D., ed. 1917. *Single Tax Yearbook*. New York: Single Tax Review Publishing Co.

Scott, A., ed. *Natural Resource Revenues: A Test of Federalism*. Vancouver: University of British Columbia Press.

Shearman, T. 1888. *Natural Taxation*. 2nd edn, New York: Doubleday, 1911.

Skouras, A. 1977. *Land and its Taxation in Recent Economic Theory*. Athens: Papazissis Publishers.

Young, A.N. 1916. *History of the Single Tax Movement in the United States*. Princeton: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Gaffney, Mason. "single tax." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000141> doi:10.1057/9780230226203.1535

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sismondi, Jean Charles Leonard Simonde de (1773 – 1842)

Thomas Sowell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand and supply; aggregate equilibrium income; business cycles; concentration; equilibrium income; gluts; leisure; Malthus, T. R.; Mercier de la Rivière, P.-P.; money; Say's Law; Sismondi, J. C. L. S. de; technological unemployment

Article

A number of concepts and theories that later became important in the history of economics first appeared in the writings of the Swiss economist J.C.L. Simonde de Sismondi. Whether or not these can be considered as his ‘contributions’ to economics is a question not unlike that as to whether a tree that falls in a deserted forest makes a sound. Sismondi developed the first aggregate equilibrium income theory and the first algebraic growth model. Yet both concepts had to be rediscovered and redeveloped by others before they entered the mainstream of economics, long after Sismondi's time. The fact that Sismondi wrote in French may have been part of the reason why his work made so little impact at a time when the development of classical economics was largely the work of British economists. However, the fame achieved by his French contemporary, Jean-Baptiste Say, suggests that language differences alone cannot explain the neglect of Sismondi. His economic writings were neglected in France and Switzerland as well.

When he was born in Geneva in 1773, his name was Jean Charles Leonard Simonde. After an exile in Italy, during which he determined that he was descended from a noble Italian family named Sismondi, he returned to Geneva in 1800 with his new surname, Simonde de Sismondi. However, he was sufficiently tentative about it to use his original name on his first book in economics, *De la richesse commerciale*, in 1803. Sismondi also wrote extensively on history, including a 16-volume history of Italy. All his writings were pervaded by considerations of public policy in general, and the interests of the less fortunate in particular.

Sismondi was born into a prosperous bourgeois family, which was despoiled of much of its wealth during Swiss political upheavals reflecting the contemporary revolution in France. Shifting political fortunes led not only to Sismondi's exile but to two imprisonments as well. After the turmoil subsided,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Sismondi worked at a variety of occupations, including gentleman farmer and professor of philosophy. Sismondi's first venture into economics, the two-volume *De la richesse commerciale*, was intended as a systematic exposition of the ideas of Adam Smith. Yet in it Sismondi also pointed out that he was presenting an ‘absolutely new’ way of looking at aggregate output changes. Crude arithmetic examples depicted output during a given year as a function of investment during a previous year, and showed how a closed economy differed from an economy with international trade, and how the latter differed when there was an export surplus and an import surplus. Algebraic formulas in his footnotes repeated the same arguments presented arithmetically in the text. But the book was little noticed, and so Sismondi's original efforts produced no contribution to the development of economics.

In the wake of the post-Napoleonic War depression, Sismondi turned his attention once more to economics and to issues of aggregate income equilibrium. In 1814, he produced a long article entitled ‘Political Economy’, written in English for the *Edinburgh Encyclopaedia*. In the midst of a summary presentation of classical economics appeared an early version of Sismondi's own theory of aggregate equilibrium income. This theory was elaborated in Sismondi's main economic work, the two-volume *Nouveaux principes d’économie politique* (1819). With this work he entered the controversy over Say's Law and general gluts.

According to Sismondi, the utility of output was balanced against the disutility of work, whether by Robinson Crusoe on an island or by a complex society. But, with different people doing partial balancing in isolation from one another in a complex economy, the aggregate balance was not always continuously assured. Whenever the disutility of labour exceeded the utility of output in a given time period, subsequent time periods would see a decline in aggregate output until the balance was restored. Conversely, when the utility of output exceeded that of labour, output would tend to rise.

The germ of this reasoning had already appeared in *L'Ordre naturel* by the Physiocrat Mercier de la Rivière in 1767. Sismondi elaborated it into a theory of equilibrium income, with which he challenged the reigning view, expressed in Say's Law, that there were no limits to production.

Say's Law, then as later, had many meanings. But one of the contemporary meanings was that there was no such thing as an equilibrium level of aggregate output. Whatever level of output was supplied could always find a demand, and where this did not happen, it was because the assortment of goods did not match consumer preferences, not because the total output was at an unsustainable level. Sismondi rejected this reasoning, arguing that the demand for leisure would at some point outweigh the demand for other goods, and that when production went beyond the point at which this happened, it would be unsaleable at cost-covering prices and so fail to be reproduced in subsequent time periods.

Sismondi understood the full implications of what he was saying and how it contradicted prevailing views. The balance of aggregate supply and demand he considered the most important question in economics, and especially so during the depression following the Napoleonic wars. J.B. Say and the Ricardians maintained that the unsaleability of some goods showed only that insufficient other goods had been produced to exchange with them – that output had the wrong internal proportions, not an excess in the aggregate, and that the proper proportions could be restored at a still higher level of aggregate production. In this view, there could be a partial glut of particular commodities but not a *general* glut of commodities.

Sismondi argued that there could be a general glut of commodities because one of the goods desired was leisure – that is, exemption from the production of commodities. He did not believe that this occurred in the normal course of free market competition but because government policy sometimes artificially

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
fostered production at an unsustainable level. Like the orthodox economists of his time, Sismondi regarded money as an unessential factor, a ‘veil’ obscuring but not fundamentally changing the behaviour of economic aggregates. His disagreements with the classical economists had nothing to do with monetary controversies such as those in 20th-century macroeconomics.

When Sismondi's *Nouveaux principes* appeared in 1819, it was immediately attacked in the *Edinburgh Review* in October of that year, in the midst of a discussion of Robert Owen. Its reasoning was declared a ‘fallacy’ and once more ‘proportions’ – not aggregates – were declared to be the only prerequisites for markets to be cleared and increased output sustained. A glut was there defined as ‘an increase in the supply of a particular class of commodities, unaccompanied by a corresponding increase in the supply of those other commodities which should serve as their equivalents’. In short, there were partial gluts but no general gluts and a still higher level of output was sustainable if properly proportioned internally. The basis of this reasoning was explicitly attributed to ‘the celebrated M. Say’, with a ‘most clear and conclusive’ treatment of the subject added by James Mill.

The appearance of T.R. Malthus's *Principles of Political Economy* the following year added fuel to the debate, for he too challenged Say's Law in the same way. Marx later characterized Malthus's book as simply the ‘English translation’ of Sismondi, but in fact it represented views which Malthus had long expressed in correspondence with Ricardo. Replies to both authors began to appear in both French and English publications during the 1820s, provoking rejoinders in books and articles. Their controversy over general gluts persisted for more than a decade, involving not only the leading economists of the time – Say, Ricardo, Malthus, Sismondi, Torrens, McCulloch and both Mills – but also Samuel Bailey, William Blake, Thomas Chalmers, and others either forgotten or little remembered in the history of economics. These published controversies were supplemented by correspondence between Sismondi and Say, Sismondi and Ricardo, Malthus and Say, and Malthus and Ricardo. Only Say seems to have acknowledged that the theory of aggregate equilibrium income had relevance to one version of Say's Law that was current at the time. In the fifth edition of his *Traité d’économie politique* in 1826 he added three paragraphs to his chapter on the law of markets, discussing ‘the limit to a growing production’ and repeating (without citation) Sismondi's argument that when output's ‘utility is not worth what it cost’, such output is unsustainable. A year later he admitted in a letter to Malthus that his law of markets was ‘subject to some restrictions’ which he had included in the most recent edition of his book. (Unfortunately, the English translation of Say's *Traité* is from the previous edition.) Finally, in a textbook published in 1828–9, Say followed the chapter on his law of markets with a chapter entitled, ‘Limits to Production’ – a phrase from Sismondi.

No such impact or even acknowledgement occurred in British economic writings. There Sismondi and Malthus were answered as if they were arguing for secular stagnation instead of temporary aggregate disequilibrium. John Stuart Mill enshrined this misunderstanding of Sismondi and Malthus in his classic *Principles of Political Economy* in 1848. Thus things stood for nearly a century, until John Maynard Keynes resurrected Malthus, but not Sismondi, as his predecessor in aggregate equilibrium theory.

Sismondi's anticipations of later economic theory were not limited to aggregate income theory. In the course of dealing with that large topic he also proposed a theory of destabilizing responses to overproduction, which would initially take the economy further from equilibrium, though it would ultimately return to equilibrium ‘after a frightful suffering’. He also dealt with the issue of the short-run shut down point of a firm, which he argued would produce even below cost-covering prices if much of its cost was fixed rather than variable. Sismondi also argued against the reigning Malthusian population

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
theory, pointing out fatal ambiguities in the word ‘tendency’ as Malthus used it and using empirical evidence to show that the *historical* tendency was for food supply to grow faster than population.
In many ways Sismondi also anticipated Marx. Sismondi's emphasis on ‘the proletarians’, on an increasing concentration of capital, recurring business cycles, technological unemployment and economic dynamics in general all reappeared (without credit) in Marx's writings.
None of these pioneering efforts by Sismondi received either contemporary acknowledgement or later recognition by the profession. His loose and sometimes inconsistent writings and his emotional assertions made it easy to dismiss him and throw away his insights along with his errors. He left no disciples and his eclecticism provided no dogma around which a school could crystallize.

Selected works

1814. *Political Economy*. New York: Augustus M. Kelley, 1966.

1819. *Nouveaux principes d’économie politique*. Paris: Delaunay, 1827.

Bibliography

Rappard, W.E. 1966. *Economistes Genèvois du XIXe siècle*. Geneva: Libraire Droz, 1966.

Salis, R. de 1932. *Sismondi, 1775–1842*. Paris: Libraire Ancienne Honoré Champion.

Sowell, T. 1968. Sismondi: a neglected pioneer. *History of Political Economy* 1, 62–88.

How to cite this article

Sowell, Thomas. "Sismondi, Jean Charles Leonard Simonde de (1773–1842)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000143>
doi:10.1057/9780230226203.1536

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

size of nations, economics approach to the

Enrico Spolaore
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economics approach to the size of nations proceeds from the trade-off between benefits and costs of larger size. Benefits of scale come from sharing public goods among more taxpayers. Larger countries can also better internalize cross-regional externalities and insure against regional shocks. But a larger size brings about higher heterogeneity of preferences over public policies. The trade-off depends on the domestic political regime and on the international environment. In a world of free trade and low international conflict small countries can prosper, while a large size matters more when trade barriers and conflict are high.

Keywords

national borders; Buchanan, J.; clubs; democracy; economic analysis of conflict; economies and diseconomies of scale; heterogeneity of preferences; globalization; local public goods; median voter; political economy and economics; rent seeking; size of economies; size of nations

Article

The number and size of sovereign states has been at the centre of human history for thousands of years, from the times of Sumerian city states to the post-cold war era. Plato in *The Laws* (360 bc, Book V) calculated the optimal size of a state as 5,040 heads of family. According to Aristotle in *The Politics* (350 bc, Book VII, ch. 4), experience had shown that a very populous state could rarely, if ever, be well governed (a view probably not shared by Aristotle's famous pupil, Alexander the Great). Montesquieu in *The Spirit of Laws* (1748, Book VIII, ch.16) wrote that ‘in a small [republic], the interest of the public is more obvious, better understood, and more within the reach of every citizen’. A theory of optimal size was sketched by Beccaria (1764, p. 91), the Italian philosopher who inspired Bentham's utilitarian approach:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> The more society grows, the smaller part of the whole does each member become, and the republican sentiment diminishes proportionately if the laws neglect to reinforce it. Societies have, like human bodies, their circumscribed limits, increasing beyond which the economy is necessarily disturbed. It would seem that the size of a state ought to vary inversely with the sensitivity [‘*sensibilità*’] of its constituency … A republic grown too vast can escape despotism only by subdividing and then reuniting itself as a number of federated little republics.

These are selective quotations from an enormous philosophical, political and historical literature (see for example Dahl and Tufte, 1973). By contrast, for a long while economists have taken political borders as given. Only in recent years has a small but expanding economic literature started to address questions of country formation and break-up with the tools of economic analysis (for discussions, see Bolton, Roland and Spolaore, 1996; Alesina and Spolaore, 2003; Spolaore, 2006). This research is motivated by the fact that political borders are not a fixed part of the geographical landscape but human-made institutions, affected by the decisions and interactions of individuals and groups who pursue their objectives under constraints. The economics approach to the size of nations can be viewed as a natural extension of the research programme of modern political economics, whose aim is to ‘endogenize’ collective decisions and institutions.

This economic literature has addressed the determination and change of political borders in different political and economic environments and using various solution concepts. Friedman (1977) studies the formation of borders as set by rent-maximizing Leviathans. Findlay (1996) analyses the expansion of empires. Alesina and Spolaore (1997; 2003) derive and compare the number and size of countries under different solution concepts: efficient (that is, welfare-maximizing) borders, voting equilibria, equilibria under unilateral secessions, and equilibria in a world of rent-maximizing Leviathans. Bolton and Roland (1997) study the break-up of nations by direct majority vote, when income distributions differ across regions, and regional median voters have different preferences over redistribution. The relationship between economic integration and the size of countries has been analysed by Alesina and Spolaore (1997) and Alesina, Spolaore and Wacziarg (2000; 2005). Wittman (1991; 2000) focuses on welfare-maximizing solutions. Optimal secession rules have been studied by Bordignon and Brusco (2001). Le Breton and Weber (2003) analyse equilibria when groups of individuals can secede unilaterally, and compensation mechanisms are possible within countries. Contributions to the literature also include Casella and Feinstein (2002), Goyal and Staal (2003), and many others.

When one considers the size of nations from an economic perspective, a natural starting point is the trade-off between benefits and costs from a larger size. Important benefits of scale are associated with the provision of public goods, which are cheaper in per capita terms when more taxpayers pay for them (empirically, smaller countries do have larger governments). Larger countries can also better internalize cross-regional externalities, a point extensively studied in the literature on decentralization and fiscal federalism. Additional benefits from size come from insurance against natural and economic shocks through inter-regional transfers. But size also comes with costs. As countries become larger, congestion may overcome some of the above benefits. More importantly, an expansion of a country's borders is likely to bring about higher heterogeneity of preferences across different individuals. Being part of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
same country implies sharing jointly supplied public goods and policies in ways that cannot satisfy everybody's preferences. Decentralization of some public goods and policies may offer a partial response to heterogeneity. However, many policies that characterize a sovereign state (basic characteristics of the legal system, foreign policy, defence policy) are indivisible and must be shared among the whole population. This induces a trade-off between economies of scale in the provision of public goods and heterogeneity of preferences. This trade-off has played a central role in the economic literature on the size of nations (see Alesina and Spolaore, 1997; 2003). The trade-off depends not only on the degree of heterogeneity of preferences but also on the political regime through which preferences are turned into policies. For example, rent-seeking Leviathans that are less concerned with the preferences of their subjects may pursue expansionary policies that lead to the formation of inefficiently large countries and empires. By contrast, democratization leads to secessions, and, in the absence of effective mechanisms to integrate populations with diverse preferences, self-determination can be associated with excessive fragmentation and costly break-up. Historically, successful societies are those that have managed to minimize the costs of heterogeneity while maximizing the benefits stemming from a diverse pool of preferences, skills and endowments.

Economic analyses of the size of nations have pointed out that the trade-off between benefits and costs of size is also a function of the degree of international economic integration (Alesina and Spolaore, 1997; Alesina, Spolaore and Wacziarg, 2000; 2005; Spolaore and Wacziarg, 2005). The size of the market may or may not coincide with the political size of a country as defined by its borders. Larger nations mean larger domestic markets when political borders imply barriers to international exchange. By contrast, market size and political size would be uncorrelated in a world of perfect free trade in which political borders imposed no costs on international transactions. Therefore, market size depends both on country size and on the trade regime. If market size matters, small countries can prosper in a world of free trade and high economic integration, while a large size is more important for economic success in a world of trade barriers and protectionism. In fact, empirically the effect of size on economic performance (income per capita, growth) tends to be higher for countries that are less open, and the effect of openness is much larger for smaller countries (Alesina, Spolaore and Wacziarg, 2000; 2005; Spolaore and Wacziarg, 2005). As economic integration increases, the benefits of a large political size are reduced, and political disintegration becomes less costly. Conversely, smaller countries tend to benefit from more openness. Hence, economic integration and political disintegration tend to go hand in hand.

The economic literature on the size of nations is connected to other established bodies of research, such as the literature on local public goods and clubs, pioneered by Tiebout (1956) and Buchanan (1965). But while local jurisdictions are not completely autonomous and people are free to move across them, the analysis of nations explicitly focuses on sovereign states that can impose direct barriers to economic exchange and mobility, and can use force in settling disputes with their neighbours. The study of nations is also connected to the literature on customs unions and trade blocs. In so far as modern nations tend to promote free trade within their own borders, countries can be seen as ‘trade blocs’ of regions. In general, free trade areas, customs unions, supranational organizations, confederations and sovereign states can be viewed as points on a continuum of increasing coordination and integration of political functions. A third body of connected research is the economic analysis of conflict, pioneered by Boulding (1962), Tullock (1974), Hirshleifer (1991; 1995), Grossman (1991) and others. Alesina and Spolaore (2005;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
2006) and Spolaore (2004) explicitly link the economic literature on conflict with the economic approach to the size of nations, and analyse how the benefits of size tend to be larger in a world with wars and international conflict. The historical importance of military technology and violent resolution of conflict for the determination of political borders is likely to spur further work linking these variables to the other economic and political factors that affect the number and size of countries over time.
In summary, the economics approach to the size of nations, while in its infancy, builds on an extensive body of concepts and tools in order to explain the endogenous formation and break-up of sovereign states within the framework of modern political economy.

See Also

- globalization
- Tiebout hypothesis
- war and economics

Bibliography

Alesina, A. and Spolaore, E. 1997. On the number and size of nations. *Quarterly Journal of Economics* 112, 1027–56.

Alesina, A. and Spolaore, E. 2003. *The Size of Nations*. Cambridge, MA: MIT Press.

Alesina, A. and Spolaore, E. 2005. War, peace, and the size of countries. *Journal of Public Economics* 89, 1333–54.

Alesina, A. and Spolaore, E. 2006. Conflict, defense spending, and the number of nations. *European Economic Review* 50, 90–120.

Alesina, A., Spolaore, E. and Wacziarg, R. 2000. Economic integration and political disintegration. *American Economic Review* 90, 1276–96.

Alesina, A., Spolaore, E. and Wacziarg, R. 2005. Trade, growth, and the size of countries. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Aristotle. 350 bc. *The Politics*. Online. Available at http://classics.mit.edu/Aristotle/politics.7.seven.html, accessed 22 August 2005.

Beccaria, C. B. di. 1764. *On Crimes and Punishments*. Indianapolis: Bobbs-Merrill, 1963.

Bolton, P. and Roland, G. 1997. The breakups of nations: a political economy analysis. *Quarterly*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Journal of Economics 112, 1057–89.

Bolton, P., Roland, G. and Spolaore, E. 1996. Economic theories of the break-up and integration of nations. *European Economic Review* 40, 697–705.

Bordignon, M. and Brusco, S. 2001. Optimal secession rules. *European Economic Review* 45, 1811–34.

Boulding, K. 1962. *Conflict and Defense: A General Theory*. New York: Harper.

Buchanan, J. 1965. An economic theory of clubs. *Economica* 32, 1–14.

Casella, A. and Feinstein, J. 2002. Public goods in trade: on the formation of markets and political jurisdictions. *International Economic Review* 43, 437–62.

Dahl, R. and Tufte, E. 1973. *Size and Democracy*. Stanford: Stanford University Press.

Findlay, R. 1996. Towards a model of territorial expansion and the limits of empires. In *The Political Economy of Conflict and Appropriation*, ed. M. Garfinkel and S. Skaperdas. Cambridge: Cambridge University Press.

Friedman, D. 1977. A theory of the size and shape of nations. *Journal of Political Economy* 85, 59–77.

Goyal, S. and Staal, K. 2003. The political economy of regionalism. *European Economic Review* 48, 563–93.

Grossman, H. 1991. A general equilibrium model of insurrections. *American Economic Review* 81, 912–21.

Hirshleifer, J. 1991. The technology of conflict as an economic activity. *American Economic Review* 81, 130–34.

Hirshleifer, J. 1995. Anarchy and its breakdown. *Journal of Political Economy* 103, 26–52.

Le Breton, M. and Weber, S. 2003. The art of making everybody happy: how to prevent a secession? *IMF Staff Papers* 50, 403–35.

Montesqueu, Baron de. 1748. *The Spirit of Laws*. Online. Available at http://www.constitution.org/cm/sol_08.htm#016, accessed 22 August 2005.

Plato. 360 bc. *The Laws*. Online. Available at http://classics.mit.edu/Plato/laws.5.v.html, accessed 22 August 2005.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Spolaore, E. 2004. Economic integration, international conflict and political unions. *Rivista di Politica Economica* 94, 3–50.

Spolaore, E. 2006. The political economy of national borders. In *Oxford Handbook of Political Economy*, ed. B. Weingast and D. Wittman. Oxford: Oxford University Press.

Spolaore, E. and Wacziarg, R. 2005. Borders and growth. *Journal of Economic Growth* 10, 331–86.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Tullock, G. 1974. *The Social Dilemma: The Economics of War and Revolution*. Blacksburg, VA: University Publications.

Wittman, D. 1991. Nations and states: mergers and acquisitions; dissolution and divorce. *American Economic Review, Papers and Proceedings* 81, 126–9.

Wittman, D. 2000. The wealth and size of nations. *Journal of Conflict Resolution* 6, 885–95.

How to cite this article

Spolaore, Enrico. "size of nations, economics approach to the." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000213> doi:10.1057/9780230226203.1537

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

skill-biased technical change

Giovanni L. Violante
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Skill-biased technical change is a shift in the production technology that favours skilled over unskilled labour by increasing its relative productivity and, therefore, its relative demand. Traditionally, technical change is viewed as factor-neutral. However, recent technological change has been skill-biased. Theories and data suggest that new information technologies are complementary with skilled labour, at least in their adoption phase. Whether new capital complements skilled or unskilled labour may be determined endogenously by innovators' economic incentives shaped by relative prices, the size of the market, and institutions. The 'factor bias' attribute puts technological change at the center of the income-distribution debate.

Keywords

capital–skill complementarity; elasticity of substitution; factor-neutral technical change; income distribution; information technology; innovation; learning; production functions; research and development; skill premium; skill-biased technical change; Solow, R.; substitutes and complements; technical change; technology–skill complementarity; total factor productivity

Article

Skill-biased technical change (SBTC) is a shift in the production technology that favours skilled (for example, more educated, more able, more experienced) labour over unskilled labour by increasing its relative productivity and, therefore, its relative demand. Ceteris paribus, SBTC induces a rise in the skill premium – the ratio of skilled to unskilled wages.

From factor-neutral to factor-biased technical change

Economic theory views production technology as a function describing how a collection of factor inputs

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

can be transformed into output, and it defines technical change as a shift in the production function, that is, a change in output for given inputs. The traditional measure of economy-wide technological change, introduced by Solow (1957), is aggregate total factor productivity (TFP). Solow defines a TFP advancement as an increase in output that leaves marginal rates of transformations untouched for given inputs; thus, a change in TFP is a form of *factor-neutral* technical change.

For illustrative purposes, suppose that the aggregate production function is constant returns to scale and Cobb–Douglas in aggregate capital (K) and aggregate labor (L) services, that is, $Y = ZK^{\alpha}L^{1-\alpha}$, where Y is aggregate output, α is the elasticity of output to capital, and Z denotes precisely TFP. If output and input markets are competitive, then the share of income going to capital equals α. Solow's (1957) fundamental insight is that, armed with this estimate of α and measures of (Y, K, L) from national accounts, neutral technical change can be quantified 'residually'. This clever and parsimonious approach to growth accounting has dominated the literature for decades, creating an overwhelming consensus that neutral technological improvements are the primary source of growth in income per capita.

However, a key fact that recently emerged from the data highlights the limits of this conceptualization of technical change. Since the mid-1970s, the rental price of skilled labour has soared dramatically relatively to that of unskilled labour despite a major increase in the relative supply of skills: for example, the college wage premium – defined as the ratio between the wage of college graduates and the wage of high-school graduates – jumped from 1.45 in 1965 to 1.7 in 1995, while the relative supply of college skills tripled over the same period. Given the observed movements in the relative quantities, these price changes could not be generated by movements 'along the production function'. Neutral technical change is, by definition, silent on changes in relative prices. Therefore, to make sense of these recent developments one must introduce the concept of *factor-biased* technical change.

For this purpose, I now generalize the aggregate production function above by letting labour input, L, be a constant elasticity of substitution (CES) function of skilled and unskilled labour, L_s and L_u, with factor-specific productivities A_s and A_u:

$$L = [(A_s L_s)^{\sigma} + (A_u L_u)^{\sigma}]^{1/\sigma}, \ \sigma \leq 1. \tag{1}$$

At this point, it is not necessary to specify what makes a worker more skilled than another: it could be education, innate ability or experience. The (log of the) marginal rate of transformation (MRT) between the two labour inputs is

$$\ln(MRT_{s,u}) = \sigma \ln\left(\frac{A_s}{A_u}\right) + (1-\sigma)\ln\left(\frac{L_u}{L_s}\right). \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Note that the TFP term Z does not enter the above equation. A change in the ratio A_s/A_u is a form of factor-biased technical change since it modifies the marginal rates of transformation, at a given input ratio. In particular, under the empirically plausible parametric assumption $\sigma > 0$, technical change is *skill-biased* if A_s/A_u increases. With competitive input markets, the (log of the) skill premium can be read off the right-hand side of (2) as well. Therefore, skill-biased technical change induces an increase in the relative productivity of skilled labour that raises its relative demand and, ceteris paribus, the skill premium.

To take this logic one step further, given an estimate of the elasticity of substitution between types of labour $1/(1-\sigma)$, and time-series on relative wages and relative factor supplies, one can measure skill-biased technical change residually from (2). For example, with an elasticity of substitution of 1.4 (or $\sigma = 0.29$) between college graduates and the rest of the labour force, the dynamics of the US college premium and of the relative supply of college skills imply a growth of skill-biased technical change (that is, of the ratio A_s/A_u) in excess of ten per cent per year from 1963 to 1987 (Katz and Murphy, 1992).

The skill bias of information technologies

Recent shifts in technology have been skill-biased. But SBTC appears all but an unexplained residual very much like Solow TFP, a ‘black box’ that needs to be filled with economic content. What really accounts for this shift in the production process since the mid-1970s? The timing of the rise in the skill premium has coincided with the rapid diffusion of information and communication technologies in the work place. Thus, a natural candidate for this wave of SBTC is the ‘information technology revolution’. Expenditures in information processing equipment and software, as a share of US private non-residential fixed investment, rose from six per cent in 1960 to 40 per cent in 2000. At the heart of these dynamics there is a staggering improvement in the quality and productivity of all those equipment goods relying heavily on semiconductors, like computers, software, and switching equipment underlying much of communication technology.

Ample microeconometric research and several case studies document a statistical correlation between the use of new technologies, like computers, and either the employment share of skilled workers (Bartel and Lichtenberg, 1987) or their wage share (Autor, Katz and Krueger, 1998) across industries. These studies firmly establish that the new technologies are deployed with better-qualified and better-paid labour, but they fail to explain *why*. This deeper question requires a quantitative theory built around an explicit economic mechanism.

Technology-skill complementarity

A large number of economic models in the literature provides a foundation for SBTC (for surveys, see Acemoglu, 2002; Aghion, 2002; Hornstein, Krusell and Violante, 2005). The central tenet of all these theories is technology–skill complementarity and takes three alternative formulations.

The first formulation is built on a defining feature of the post-war US economy: the sharp decline of the constant-quality relative price of equipment investment (Gordon, 1990; Greenwood, Hercowitz and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Krusell, 1997), especially evident for information technologies whose prices fell at ten per cent per year. Krusell et al. (2000) argue that the substantial cheapening of equipment capital is the force behind SBTC. This decline in price led to an increased use of equipment capital in production. At least since Griliches (1969), various empirical papers support the idea that skilled labour is relatively more complementary to equipment capital than is unskilled labour. As a result of capital–skill complementarity in production, the faster growth of the equipment stock pushed up the relative demand for skilled labour and, in turn, the skill premium.

More explicitly, these authors generalize the aggregate production function to:

$$Y = K_s^{\alpha}\left[\gamma[\mu(K_e)^{\rho} + (1-\mu)(L_s)^{\rho}]^{\frac{\sigma}{\rho}} + (1-\lambda)(L_u)^{\sigma}\right]^{\frac{1-\alpha}{\sigma}}, \tag{3}$$

where K_s denotes structures capital and K_e equipment capital. Profit-maximizing behaviour of price-taking firms implies that the skill premium (and the MRT between labor inputs) can be approximated as

$$\ln\left(\frac{w_s}{w_u}\right) \simeq \lambda\frac{\sigma-\rho}{\rho}\ln\left(\frac{K_e}{L_s}\right)^{\rho} + (1-\sigma)\ln\left(\frac{L_u}{L_s}\right). \tag{4}$$

If $\sigma > \rho$, as estimated by Krusell et al., the relative demand for skills increases with the stock of equipment capital. Note the difference between eqs. (2) and (4): capital–skill complementarity gives economic content to the notion of SBTC by replacing an unobserved residual trend (A_s/A_u) with the actual upward trend in equipment–skilled labour ratio. This model replicates well the dynamics of the US skill premium since the mid-1960s. Moreover, historical evidence suggests that complementarity between skilled labour and capital characterized technological developments throughout the entire 20th century (Goldin and Katz, 1998).

The second formulation is inspired by the Nelson–Phelps view of human capital. In the words on Nelson and Phelps (1966, p. 70), ‘educated people make good innovators, so that education speeds the process of technological diffusion’. In particular, they contend that more educated, able or experienced labour deals better with technological change. Skilled workers are less adversely affected by the turmoil created by major technological transformations since it is less costly for them to learn the additional knowledge needed to adopt a new technology. Therefore, rapid technological transitions, such as that witnessed since the mid-1970s, are skill-biased, as more able workers adapt better to change (Greenwood and Yorukoglu, 1997; Caselli, 1999; Galor and Moav, 2000).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Incidentally, this version of the technology–skill complementarity hypothesis, by emphasizing the importance of learning during episodes of drastic technical change, is consistent with the TFP slowdown experienced by most developed economies in the 1980s: upon the arrival of the new information technologies, aggregate productivity can fall temporarily as workers and firms learn how to deploy the new production methods at their best (Hornstein and Krusell, 1996; Aghion, 2002).

The Nelson–Phelps conjecture implies that the rise in the skill premium is transitory: it is only in the early adoption phase of a new technology that those who adapt more quickly can reap some benefits. As time goes by, there will be enough workers learning how to work with the new technology to offset the wage differential. Note the difference from the hypothesis set forth by Krusell et al. (2000), where the effect of capital deepening on the skill premium is permanent.

The third formalization of this hypothesis is based on Milgrom and Roberts (1990). These authors argue that information technologies reduce costs of data storage, communication, monitoring and supervision activities within the firm, which triggers a shift towards a new organizational design. In particular, the layers of the hierarchical structure can be reduced, so that the organization of the firm becomes ‘flatter’. Workers no longer perform routinized, specialized tasks, but they are now responsible for a wide range of tasks within teams. Therefore, adaptable workers who have general skills and who are more versed at multi-tasking activities benefit from this transformation. In other words, the change in technology induces an organizational shift which is skill-biased. An elegant formalization of this hypothesis is contained in Garicano and Rossi-Hansberg (2004).

Microeconomic evidence consistent with all these formulations of the technology–skill complementarity hypothesis is offered by Autor, Levy and Murnane (2003). Based on data on the skill content and tasks of various occupations, they split job requirements into ‘routine’ and ‘non-routine’ tasks and document that, starting from the 1970s, the labour input of non-routine analytic and interactive tasks increased sharply relative to routine cognitive and manual tasks. This shift was concentrated in rapidly computerizing industries and it was pervasive at all educational levels. The authors interpret these findings as evidence that information technologies substituted for unskilled labour employed on simple and more repetitive tasks – more amenable to computerization – and complemented workers endowed with generalized problem-solving, complex communication, and analytical skills.

Endogenous direction of technical change

In the same vein as the endogenous growth literature developed in the 1990s, one could contend that not only the speed – as traditionally argued – but also the *direction* of technical change is endogenous. Profit incentives of innovators determine the amount of R&D activity directed towards different factors of production (Acemoglu, 1998). The main determinants of profit incentives are market size, relative prices and institutions. These forces can shed light on numerous episodes in the history of technology.

Under the assumption that the R&D is fixed, the market size of the innovation determines its revenues. The expansion of educated labour during the postwar period made it profitable to develop machines complementary to skilled workers. The vast rural–urban migration wave towards English cities during the late 18th century opened the way to the development of the factory system and, later, to the Tayloristic assembly line which quickly replaced skilled artisans’ craft shops. Incidentally, this is a notable example of *unskill bias* which proves that, historically, the direction of technical progress has

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
varied.
Profit maximization dictates that, ceteris paribus, innovation be directed towards those factors that are more intensely used in the production of highly priced goods. The recent burst of North–South trade increased the relative price of skill-intensive goods in the North, representing yet another force which pushed towards skill-biased innovations in the post-war period.
Labour institutions that keep wages high despite reductions in productivity induce firms to direct efforts towards labour-saving technologies. Such a fall in labour demand may explain the rise in European unemployment, after the upward wage push secured by the 'labour movement' in the 1970s. The hump-shaped dynamics of the European labour share between 1970 and 1990 validate this conjecture.
The theory of endogenous factor bias in technology is, potentially, far reaching. The main limit, at this early stage of development, is the lack of quantitative analysis of the proposed mechanisms. For example, is the acceleration in the growth of college skills of the 1970s large enough to generate the observed rise in the productivity of skilled labour and in the skill premium, under a plausible model calibration? Such questions remain unanswered to date.
To conclude: traditionally, in the growth literature technological progress is associated with productivity improvements that benefit all workers, and it is viewed as the chief long-run determinant of *average income* levels. The notion of 'skill bias' – and the literature that has recently blossomed around it – has introduced the theoretical possibility that technological progress benefits only a sub-group of workers, placing technical change also at the center of the *income distribution* debate.

See Also

- biased and unbiased technological change
- Griliches, Zvi
- Industrial Revolution
- information technology and the world economy
- Solow, Robert
- substitutes and complements
- technical change
- technology
- total factor productivity

I am grateful to the CV Starr Center for research support.

Bibliography

Acemoglu, D. 1998. Why do new technologies complement skills? Directed technical change and wage inequality. *Quarterly Journal of Economics* 113, 1055–90.

Acemoglu, D. 2002. Technical change, inequality and the labor market. *Journal of Economic Literature* 40, 7–72.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Aghion, P. 2002. Schumpeterian growth theory and the dynamics of income inequality. *Econometrica* 70, 855–82.

Autor, D., Katz, L. and Krueger, A. 1998. Computing inequality: have computers changed the labor market? *Quarterly Journal of Economics* 113, 1169–213.

Autor, D., Levy, F. and Murnane, R. 2003. The skill content of recent technical change: an empirical exploration. *Quarterly Journal of Economics* 118, 1279–334.

Bartel, A. and Lichtenberg, F. 1987. The comparative advantage of educated workers in implementing new technology. *Review of Economics and Statistics* 69, 1–11.

Caselli, F. 1999. Technological revolutions. *American Economic Review* 89, 78–102.

Galor, O. and Moav, O. 2000. Ability biased technological transition, wage inequality within and across groups, and economic growth. *Quarterly Journal of Economics* 115, 469–97.

Garicano, L. and Rossi-Hansberg, E. 2004. Inequality and the organization of knowledge. *American Economic Review* 94, 197–202.

Goldin, C. and Katz, L. 1998. The origins of technology–skill complementarity. *Quarterly Journal of Economics* 113, 693–732.

Gordon, R. 1990. *The Measurement of Durable Goods Prices*. Chicago: University of Chicago Press.

Greenwood, J., Hercowitz, Z. and Krusell, P. 1997. Long-run implications of investment-specific technological change. *American Economic Review* 87, 342–62.

Greenwood, J. and Yorukoglu, M. 1997. 1974. *Carnegie–Rochester Conference Series on Public Policy* 46, 49–96.

Griliches, Z. 1969. Capital–skill complementarity. *Review of Economics and Statistics* 5, 465–68.

Hornstein, A. and Krusell, P. 1996. Can technology improvements cause productivity slowdowns? In *NBER Macroeconomics Annual*, vol. 11, ed. J. Rotemberg and B. Bernanke. Cambridge, MA: MIT Press.

Hornstein, A., Krusell, P. and Violante, G. 2005. The effects of technical change on labor market inequalities. In *Handbook of Economic Growth*, vol. 1B, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Katz, L. and Murphy, K. 1992. Changes in relative wages, 1963–1987: supply and demand factors. *Quarterly Journal of Economics* 107, 35–78.

Krusell, P., Ohanian, L., Ríos-Rull, J. and Violante, G. 2000. Capital skill complementarity and inequality: a macroeconomic analysis. *Econometrica* 68, 1029–54.

Milgrom, P. and Roberts, J. 1990. The economics of modern manufacturing: technology, strategy, and organization. *American Economic Review* 80, 511–28.

Nelson, R. and Phelps, E. 1966. Investment in humans, technological diffusion, and economic growth. *American Economic Review* 56, 69–75.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

How to cite this article

Violante, Giovanni L. "skill-biased technical change." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000493> doi:10.1057/9780230226203.1538

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

slavery

S.L. Egerman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cairnes, J. E.; contract labour; indentured servitude; serfdom; slavery; total factor productivity; Williams, E.

Article

Slavery entails the ownership of one person by another. As a form of labour organization it has existed throughout history, in a large number of different societies. Indeed, while today we regard slavery as ‘the peculiar institution’, by historical standards it is wage-labour markets that are ‘unusual’. Given slavery's long and varied history a precise definition is not always agreed upon, but there are certain general characteristics found in most societies. The status of slave is generally applied to outsiders – individuals not belonging to the dominant nation, religion or race – although the definition of exactly who is an outsider has varied; Orlando Patterson (1982) considers the basic characteristic of slavery to be ‘social death’, with a loss of honour as well as legal rights by the enslaved. So widespread and acceptable was slavery that in Europe and the Americas no movement developed to attack slavery as a system until the late 18th century. In Western thought slavery had long been regarded as a necessity or as a ‘necessary evil’. The pre-19th-century discussions of slavery often pointed to the desirability (on grounds of religion and morality) of ameliorating the conditions of the enslaved or of facilitating the liberation of individual slaves, but it was only in the late 18th century that widespread thought was given to abolishing the institution itself (see Davis, 1966; 1975; 1984).

The economic basis of slavery

In some societies, particularly at low levels of income, slavery was entered into voluntarily as a means of obtaining a minimum level of living. Slavery generally played a quite different role in such cases than in the major slave societies of Greece, Rome, Brazil, the United States South and the West Indies (see Finley, 1980), where slave labour was used on large-scale agricultural units and in mines to produce

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
goods to be sold in foreign or urban markets. These major slave societies were marked by an extensive external trade in slaves – war captives, kidnapped or otherwise acquired.

The economic basis of slavery seems quite straightforward (see, however, Pryor, 1977). While slavery existed in some parts of the world to provide prestige to owners or for purposes of lineage needs, slavery as an economic institution generally persisted because it provided slaveowners with an ability to capture a surplus of the value of production above the costs of the slave's subsistence. In 1900 the Dutch ethnographer H.J. Nieboer presented a detailed comparative study of slavery throughout the world. He argued that at relatively primitive stages of production, among hunters and fishers and in pastoral societies, slavery was generally non-existent. It was only with the development of settled agriculture in areas with productive land still available that slavery became widespread. The role of ‘open resources’ – free land – has been more formally presented by Evsey Domar (1970), who argues that only two of the following three conditions can hold simultaneously: free land, free peasants and non-working landowners. The basic argument is that if it were possible for workers to produce more than their subsistence, with unrestricted mobility they would move to freely available land; it is only with a form of labour coercion (serfdom or slavery) that landowners can obtain an income. The benefit of coerced labour to the slaveowner involves a redistribution of that part of the income above subsistence that would go to a free worker. Thus if everything (for example, crops grown, slave and free productivity) were equal, slavery would provide a means of redistributing the excess above subsistence from the labourer to the landed slaveowner.

While this model of forced redistribution points to the need for slave output to exceed subsistence to make the system desirable to slaveowners, it is incomplete as an explanation for most large-scale slave economies. The critical point has been that free labourers have avoided certain types of labour – producing certain crops, or working in certain locations, as well as limiting their labour force participation. Thus slavery expanded the available labour supply, and was particularly important to the production of certain outputs – from mines and in large-scale agricultural units – for which labour would have been available only at very high prices needed to offset non-pecuniary aspects of the labour process (Barzel, 1977). It is the existence of crops such as sugar, for which free labour cannot easily be attracted, that explains the development of New World slavery and accounts for the fact that where slavery has been economically important the slaves have performed functions quite different from those of free labourers.

The Domar–Nieboer argument, by itself, explains neither the existence of slavery as a form of coerced labour nor even the actual existence of coerced labour. The theoretical point is consistent with any form of coerced labour– slavery as well as serfdom – with the general pattern being that slavery dominated where it was necessary to move more labour into an area. The availability of free land by itself can have a quite opposite impact, as Adam Smith (1776) and others argued was the case for the northern states of the United States. There, free land (and labour mobility) meant a wider distribution of property and a more egalitarian society. The conditions for a successful cartel of slaveowners require methods of restricting the movement of the labour force, measures precluding direct bargaining between labourers and potential owners, and means of identifying and returning those slaves who attempt to leave the system. Why such restrictions would not be enforced against Europeans was no doubt the outcome of various cultural, religious and racial forces, so that the availability of free land provided quite different outcomes in the northern and in the southern states. Importantly, different technologies of crop

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
production – the larger optimum scale for cotton, sugar, rice and tobacco, in comparison with the family farm for grains – meant that some form of coerced labour was needed to attract workers onto the southern plantations, while the prices of slaves became too high to permit large numbers of slave imports into the north.

Between the ending of Roman slavery and the origins of large-scale slavery in the New World settlements of the European powers, slavery persisted throughout Europe, but the relative absence of a large trading sector and the limited need to attract a larger population to new areas of settlement meant that slavery was generally economically unimportant and serfdom was a more frequent form of labour organization.

The transatlantic slave trade

The slave trade westward from Africa began with the movement of African slaves to the offshore islands by Portuguese traders in the middle of the 15th century. From that year until the last slave was landed in Cuba in the late 1860s, more than ten million Africans landed in the New World. Allowing for death in the 'Middle Passage' about 12 million slaves left Africa (Curtin, 1969; Lovejoy, 1983). Higher numbers for the impact of the slave trade on Africa have been presented, allowing for deaths between enslavement and shipment, as well as for estimates of the deaths due to military actions which led to enslavement. While the transatlantic slave trade was most intense in terms of numbers carried per year, it has been estimated that the trans-Saharan slave trade to Arabia and the Middle East may have carried a comparable number over a longer time span. It appears, however, that except for some small areas, the slave trade did not lead to depopulation within Africa. And, although slavery had long existed in Africa, many believe that it was the European contact that transformed African slavery into a harsher institution.

Most countries of western Europe were involved in the Atlantic slave trade, and slaves were sent to all parts of the Americas. Although the high-risk nature meant that some voyages could be very profitable, recent work has cast doubt on the argument that the slave trade provided abnormal profits to European traders, given the African control of the inland traffic and competition among shippers. The first attacks upon slavery were aimed at restricting the transatlantic shipments of Africans. The British, after initial regulatory legislation beginning in 1788, and the United States, as a result of a constitutional compromise permitting its outlawing, ended the slave trade in 1808. Denmark, a minor carrier, had ended the slave trade in 1802. Due to British pressures other countries ended their slave trades, although the 'illegal' slave trade to Cuba and Brazil did not end until after mid-century (Eltis, 1987).

The slave trade was linked to European overseas expansion and played an important part in the settlement of the Caribbean colonies. Because of its early start and late ending Brazil was the largest of the New World recipients of slaves. Large numbers were sent to the British and the French West Indian colonies, whose populations soon became 80 to 90 per cent enslaved blacks. The United States, which was to become the largest slaveholding nation in the 19th century, received only a small part of the African slave trade, its large population being due to the unusually rapid rate of natural increase of the slave population. Cuba rose to dominance as a sugar producer in the 19th century but, unlike the other major sugar-producing Caribbean islands, its population was only about one-third slave, a ratio similar to that in Brazil and in the United States at that time.

The major use of slave labour during the period of the slave trade was in the production of sugar (see Deerr, 1949–50; Klein, 1986). In the case of mainland North America, although slavery existed in every

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
colony, the major uses of slave labour in the 18th century were for tobacco production in the Chesapeake area and rice and indigo production in South Carolina. Cotton was grown in the West Indies and Brazil and, after the invention of the cotton gin in 1793, it became the most important crop produced by slave labour in the United States. These and other slave-grown crops in the Americas, such as coffee and cocoa, were grown on units larger than the family farm.

Slavery in the Americas

In the Americas, after some initial attempts at the enslavement of the native Americans, the condition of slave became limited to Africans and their descendants. The black–white ratio was generally highest in the Caribbean, declining as one moved north and south. There were pronounced differences in the demographic performance of slave populations in the Americas, as reflected in differences between the total number of slaves imported and the number of blacks in various areas. The most dramatic and widely noted differences were seen between the United States and the West Indies. In the West Indies, with few exceptions, the slave population was unable to maintain its numbers, and there was a continued need for new slaves brought from Africa. In the British Caribbean, for example, it is estimated that the approximately two million Africans received before 1808 left a population of only about 780,000 blacks at the time of emancipation (1834). The United States provided a quite different case – unique for a slave population and unusual for any population. There an estimated 600,000 slaves imported resulted in a black population of over 2.3 million in 1830, rising to about 4.4 million in 1860. The relatively high death rates during the period euphemistically known as ‘seasoning’ (the initial period of exposure to the new disease environment) accounts for some of the correlation between imports and mortality. Despite claims at the time there seems little evidence that planters systematically either worked slaves to death or deliberately engaged in the breeding of slaves.

There was both a considerably higher birth rate for slaves in the United States than in the West Indies (a rate about equal to that of United States whites, about 50 per cent above that in Europe and the West Indies), as well as a lower death rate. The higher birth rate was due in part to an earlier onset of menarche in the United States (due to better nutrition), a shorter child-spacing interval (reflecting differences in lactation practices) and a higher frequency of childbearers among adult women (due perhaps to differences in working conditions as well as in nutrition and health care) (see Steckel, 1985).

The lower death rate reflected, in part, the differences between the location and work routines of tropical sugar plantations and the major uses of slave labour in the United States – tobacco and cotton (see Higman, 1984).

Recent studies of the New World slave economies indicate that, despite arguments of contemporaries and subsequent scholars, slavery was expanding throughout its period of existence and that there were no signs that slavery was becoming unprofitable and non-viable on economic grounds. Slave prices in Brazil, Cuba and the United States peaked around 1860 (Moreno Fraginals, Klein and Engerman, 1983).

The United States Civil War ended slavery there, and while prices in Brazil and Cuba declined somewhat from peak levels, they remained higher than they had been before mid-century, until there were clear signals that the system was soon to be ended legislatively. The basic importance of labour coercion in sugar production can also be seen in the drive to import contract labour from India, China and Africa, in the West Indies after the ending of slavery – in some cases (most importantly Cuba) even

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
while slavery still existed.

Whatever alternatives may be argued to have been ultimately in their self-interest, throughout the 19th century (and earlier) slavery was profitable to planters. Rather than facing economic difficulties, planters were benefiting from the increased European demands for sugar, cotton, coffee and other plantation commodities. Despite the theoretical logic of the arguments by Cairnes (1862) and others about the ultimate limits to slavery's profitability, for the major slave powers of the New World emancipation required political or military action to overcome a profitable slave economy in which few planters anticipated an immediate collapse in the productive value of their principal assets.

Emancipation and its economic effects

Several states of the northern United States ended slavery by judicial or legislative measures between the Revolutionary War and the early 19th century, as did several of the formerly Spanish colonies after their independence was achieved in the first part of the 19th century, but these were areas where slavery was relatively unimportant. The major sugar-producing area of Saint-Domingue (now Haiti) ended slavery, as the result of a major slave revolt, by the start of the 19th century, and after violently opposing attempts by its new leaders to reintroduce a plantation economy to produce sugar, it became an area devoted to small-scale peasant production.

The first major area to end slavery after the Haitian Revolution was the British West Indies in 1834. The legislation passed in 1833 in response to a decade of pressure by the antislavery movement, provided for (1) a cash payment to owners and slaves of £20 million based on (but less than one-half of) the 1823–30 market values; and (2) an ‘apprenticeship’ of from four to six years, depending upon the slave's occupation. It is estimated that the value of the monetary compensation, plus the labour dictated by apprenticeship, would be nearly equal to the average 1823–30 value of slaves although, as the slaveowners pointed out, the loss in the value of land due to emancipation was not compensated (Fogel and Engerman, 1974b). The period of apprenticeship was terminated in 1838.

Metropolitan legislations ended slavery, with compensation, in the French West Indies and in the Danish West Indies in 1848, and in the Dutch colonies in 1863. The American Civil War provided an uncompensated end to United States slavery with the passage of the Thirteenth Amendment. The Moret Law of 1870 provided that all those born to slave mothers in Cuba and Puerto Rico (which then ended slavery, with compensation to masters, in 1873) were considered free, subject to a period of compelled labour. The Rio Branco Law of 1871 in Brazil included a similar provision – a ‘law of the free womb’ – with a period of controlled labour. In Cuba slavery was ended in 1880, subject to a proposed eight-year period of *patronato*, which was terminated in 1886. In Brazil slavery was ended without compensation in 1888.

The causes of this century-long process of emancipation have become a major historical controversy, with particular attention given to the movements in England and the United States. For England, the view that emancipation was the outcome of disinterested humanitarianism came under attack with the economic interpretation of Eric Williams (1944), which related the timing of the ending of the slave trade and of slave emancipation to the British Industrial Revolution. While the specific groups and mechanisms remain debated, the linkage now stresses the rise of individualism, the ‘free labour ideology’ and ‘modernization’, all of which meant that slavery was considered an unacceptable

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
arrangement. Similarly for the United States the link between the ‘free labour ideology’ and anti-slavery has become central to the interpretation of various political issues of the 1850s which culminated in the Civil War.

The economic effects of slavery upon production can be seen clearly in the general patterns of output that developed after the emancipation of the slaves in most areas. With few exceptions, emancipation led to initial declines in the level of output, with particularly sharp declines in the production of the staple export commodities (Engerman, 1982). There was a movement of ex-slaves away from the plantation sector into small-scale agriculture. The ending of slavery thus demonstrated anew why most New World slavery had developed – people, given free choice, preferred not to work on plantations producing staple crops for exports, since landowners would (or could) not provide sufficient wages to provide an adequate voluntary plantation force from the local population.

There were, however, several notable exceptions. In Barbados the end of slavery did not end the plantation system nor did it lead to declines in sugar production. Rather the labour-to-land ratio was already so high that land for the ex-slaves to move to was unavailable. Barbados thus maintained its plantation sector, while serving as an area for labour outflow to other parts of the Caribbean through the 19th and 20th centuries. Another important exception was Cuba. By the time of emancipation the rise of the large central mill utilizing cane from smaller farms permitted an alternative there which offset the impact of declines in the plantation labour force, and so the output of sugar did not decline after slavery ended.

The two largest slave economies, the United States and Brazil, were nations where slave labour did not reach the proportionate dominance that it did in most of the Caribbean and they had rather different problems. United States sugar, rice, cotton and tobacco production declined with emancipation, with output recovering pre-Civil War levels subsequently, at speeds that were in inverse relation to the optimum scale of plantation production – tobacco and cotton recovering fastest, and sugar and rice least rapidly. Dramatic regional shifts occurred, with prolonged declines in the older tobacco areas of Virginia and the ultimate transfer of rice production from its antebellum base in South Carolina and Georgia to Louisiana. Cotton production expanded quickly after the post-emancipation decline, recovering antebellum peaks within a decade and the United States regained its dominance in world markets by 1880. Yet plantation production declined, and there emerged a system of small-scale farms, often sharecropped, which were less productive than were antebellum plantations; and while most blacks remained within the cotton sector, increased numbers of whites became involved in the cotton economy. Unlike most other ex-slave societies, with this expansion of cotton production the South was exporting a higher proportion of its agricultural output than before emancipation.

In Brazil, the emancipation of slaves had a sharp initial impact upon the expanding coffee industry. Recovery occurred with a decline in the importance of plantations and a shift in the nature of the labour force, with the attraction of immigrants from southern Europe (mainly Italy) to produce on small units. A move to smaller farms producing sugar for central mills also permitted recovery in the production of sugar, with limited numbers of ex-slaves remaining in sugar production.

This general pattern of ex-slave withdrawal from plantation work was a characteristic of the post-emancipation period throughout the Americas. Important in some parts of the Caribbean after slave emancipation was the attraction of a new labour force from overseas, through indentured labour transported under contract to work on plantations for specified periods of time. The areas of the British Caribbean expanding in the late slave period – Trinidad and British Guiana – regained pre-emancipation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
levels of sugar output within a plantation-based economy. The labour force on the plantations was not primarily ex-slave, but rather indentured labour brought in from Africa, Madeira, China and India, the latter being the predominant group. This system of contract labour had been employed initially in the British Indian Ocean colony of Mauritius. Called ‘a new system of slavery’, contract labour was a widely discussed and regulated form of labour movement by metropolitan powers as well as in areas of outflow and inflow until its abolition in 1917. The importance of the problem of maintaining a labour force on a continuous basis on sugar plantations is seen also in the expansion of contract labour from foreign areas to the newly emerging sugar-producing regions, such as Fiji, Hawaii, Natal and Australia, in the late 19th century. As late as 1880, the production of most cane sugar that entered export markets took place in areas where the predominant plantation labour force was based either on slavery or indentured labour (Engerman, 1983).

Slavery in economic thought

The consideration of slavery in the literature of economics has helped shape subsequent interpretations of the slave economy. Adam Smith (1776) has been the most frequently quoted economist against slavery, his arguments featuring in contemporary debates as well as historical writings. To Smith slavery was an inefficient system: the slave lacked incentives to work as well as to innovate in technological change. Smith explained the existence of slavery in the production of such crops as sugar and tobacco as indicative that these were so profitable that they could afford to utilize slave labour, something not possible with less profitable crops such as corn. Smith drew upon existing arguments, his proposition on relative incentives having a long history going back to the classical world; but Smith's reputation as a political economist served to make his arguments a central component in the emerging anti-slavery argument. Recent views on Smith stress less that he was presenting an empirical proposition than that his remarks on slavery should be regarded as a basic ideological statement.

Several of the classical economists, for example McCulloch (1825) and Mill (1848), agreed with Smith's contention as to the relative effectiveness of slave and free labour when both were undertaking the same type of work, but they stressed that slavery seemed essential for production in areas and in conditions where free labour could not be obtained, particularly in tropical areas for the production of plantation crops. These, however, they regarded either as special cases (where a different economics applied) or else a transient stage (of undefined duration) along the road to free labour.

The most systematic writer on the economics of slavery was John Elliot Cairnes, whose *The Slave Power* (1862) focused on the United States and combined a theoretical analysis of slave labour with a propagandistic attempt to influence British opinion during the Civil War. He argued that slave labour was inefficient – it is given reluctantly; it is unskilful; it is wanting in versatility – and that it precluded southern economic development because of its negative effects upon technology and upon the attitudes of the free white population towards labour. Cairnes did allow that slavery could survive under certain unusual or temporary conditions – the availability of new lands which would offset the retarding effects of the land exhausted by slave labour. This statement, one of theoretical tendencies, was consistent with arguments that expansion was economically necessary for the southern economy. The role of the increasing ratio of labour-to-land in ending slavery's profitability in the long run was also a theme of various American writers of the early 19th century, and the same point re-emerged in the historiography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of United States slavery in the 1920s, with arguments about the natural limits of slavery's expansion, the imminent unprofitability of slavery and the 'needless' Civil War.

Discussions of the economic role of slavery and the causes of its ending were often presented in the 19th and 20th centuries. Yet because of the ideological views of many writers and the emotive implications of slavery it was often difficult to secure entirely accurate descriptions. To the Marxist, slavery is an inefficient economic system, incapable of high levels of productivity and of technical innovation – a system which, however necessary in its time, is incapable of generating sustained economic development. The decline of Roman slavery was thus attributed to its inability to innovate and adapt, and the re-emerged slavery of the modern world was similarly doomed to defeat in economic competition with the capitalistic order. (The relation of slavery and capitalism has itself become a debated subject, even among Marxists.) So, in the examination of the slave economies of the Americas, a perhaps surprising consistency of opinion between those coming from the classical, laissez-faire tradition and those coming from a Marxist perspective meant that a view of slavery as a backward, inefficient economic system had come to dominate the economic and historical literature on slavery.

Economic history and the economics of slavery

In the mid-20th century, work of a detailed empirical nature on slavery in the British Caribbean and in the United States was expanded. This has provided more detailed information as well as having led to new questions and issues being studied. Perhaps the dominant figure in the historiography of the British West Indies has been Eric Williams, the late Prime Minister of Trinidad and Tobago, whose most famous work *Capitalism and Slavery* (1944) dealt with three topics of importance: (1) the link of slavery and racism; (2) the role of slavery and the slave trade in the British Industrial Revolution; and (3) the impact of a declining West Indian economy upon the British abolition of the slave trade and the emancipation of slaves. Williams argued that it was the need for a cheap labour force that led to slavery, and that the justification for enslavement of Africans led to the development of racist beliefs about blacks – a view which remains the subject of a major historical controversy. Using arguments provided by contemporaries justifying the slave trade, Williams traced an important role in financing and in providing markets for British industrial development to the slave economies of the West Indies and the slave trade with Africa. Williams's last two propositions have formed the basis of much of the recent work on slavery in West Indian economic history. Recent writings point to a more limited role for slavery and the slave trade in British industrialization than that advanced by Williams. Even more attention has been devoted to the question of the conditions for the abolition of slavery and its link to a possible decline of the economies of the British West Indies. The thesis of decline has come under strong attack, particularly by Seymour Drescher (1977), who argues that the politically mandated end of the slave trade led to declining West Indian economic fortunes, and not vice versa. The issue of the specifies of the movement to end the slave trade, and the relative contributions of ideological, class and economic forces has become a central historical issue (see Drescher, 1986; Solow and Engerman, 1987).

In the United States there has been a more specific concern by economists and economic historians with issues related to the economics of slavery. A key breakthrough in the new approach to the economics of slavery was an article by Conrad and Meyer (1958) which dealt with the profitability of antebellum slavery. This article was concerned with a major historical issue – was slavery economically

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
unprofitable in the late antebellum period? – and was also intended more generally to demonstrate the value of an economic approach to historical problems. By framing the issue as one of the rate of return on investment, and using available data on interest rates, slave prices, life expectation, costs of consumption by slaves and of their supervision, and crop production and prices, they argued that a planter purchasing a slave at the market price in the late antebellum period would have earned a return equal to that upon alternative assets. The response to this article – both positive and negative, in regard to substance as well as method – has been enormous, and the economics of slavery became one of the most heated and widely discussed topics in American history. It was seen, however, that profitability, as measured by 'normal' profits on an existing asset, did not really adequately answer the question of the possible economic ending of slavery in the absence of the Civil War, since the comparison had been based on market price, and not on the cost of 'producing' a slave.

An analysis by Yasukichi Yasuba (1961) pointed out that, given the illegality of slave imports and the constraints on the demographic expansion of the slave population, the market price of slaves could exceed the costs of rearing slaves, yielding a rent to the slaveowning class (see also Evans, 1962). Yasuba showed that the surplus above rearing costs for slaves was rising in the late antebellum period, peaking just before the onset of the Civil War. Thus, far from being on the verge of economic collapse, slavery was becoming more profitable to the slaveowning class, who did not foresee an economic end to their system in the immediate future. The linking of Easterlin's regional income estimates with Gallman's GNP estimates for 1840 to 1860 indicated that the South was growing about as rapidly in terms of per capita income as was the North, and in 1860 had reached a level of per capita income above that of the agrarian Midwest and most of the rest of the world (Fogel and Engerman, 1971; 1974a). While these estimates cover only a limited time span, they did help to provide a different view of the dynamics of the southern slave economy.

Questions of profitability, viability, and rates of growth of income were not seen, however, as of central historical interest by historians such as Eugene Genovese (1965), who argued that the important questions concerned rather the development and potential for industrialization in the slave economy, reflected particularly in the differences in economic structure in comparison with that of the northern states, as well as the political issues posed by the differing class structures of the two societies. It is argued that the antebellum expansion of the South was due to the demand for one major crop, cotton, leading to a less diversified and industrialized economy than in the North – a growth that could not be sustained. At debate remain the causes and consequences of the southern specialization in agriculture rather than the development of a larger manufacturing sector, and the implications of the limited industrialization, urbanization, and expansion of education (in comparison with the North).

An extensive debate on the efficiency of slavery in United States agriculture was generated by the application of total factor productivity estimates by Fogel and Engerman (1974a). The question was an old one – frequent comparisons of slave versus free labour had been made by contemporaries as part of the anti-slavery argument. In their analysis Fogel and Engerman used a sample of over 5,000 farms in cotton-producing counties, drawn by William Parker and Robert Gallman from the census manuscript schedules. The specific contention that in 1860 southern agriculture was more efficient than northern, in the sense of getting more output per unit of input, became widely discussed and criticized (see David et al., 1976; Wright, 1978). The extensive debate included discussion of alternative measures of factor inputs and adjustments for variations in crop-mix, as well as arguments about the emotive content of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
term ‘efficiency’. Nevertheless, this debate did lead to changes in depictions of the slave economy. More attention was paid to the flexibility of the economy, in terms of shifting patterns of production and location in response to economic stimuli as well as in the use of various mechanisms, in addition to the whip, to elicit work effort from the slaves. More attention was also given to the standard of living provided for the slaves and to their actual work experiences.

Much of the writing on United States slavery in the 1970s, coming from a variety of backgrounds and using different sources, also led to reinterpretations of the nature of slavery and the slave experience. Attention was given to the slave culture, affected, but not destroyed, by the controls of the master. While there remain disagreements about the frequency of the sales of slaves and the extent to which they separated spouses as well as young children, much work has established the strength of the slave family. Descriptions of slave religion, slave culture and the slave family all pointed to the capacity of the slaves to resist being reduced to Sambos – a point with obvious implications for the behaviour of masters as well. Slavery has come to be seen as a system which, with its initial imbalance of power, permitted a range of give-and-take between master and slave (see Genovese, 1974), with the power of the former not as complete, and the impact on the latter not as destructive, as earlier argued.

Bibliography

Barzel, Y. 1977. An economic analysis of slavery. *Journal of Law and Economics* 20, 87–110.

Cairnes, J.E. 1862. *The Slave Power*. London: Parker, Son & Bourn.

Conrad, A.H. and Meyer, J.R. 1958. The economics of slavery in the antebellum south. *Journal of Political Economy* 66, 95–130.

Curtin, P.D, 1969. *The Atlantic Slave Trade*. Madison: University of Wisconsin Press.

David, P.A. et al. 1976. *Reckoning with Slavery*. New York: Oxford University Press.

Davis, D.B. 1966. *The Problem of Slavery in Western Culture*. Ithaca: Cornell University Press.

Davis, D.B. 1975. *The Problem of Slavery in the Age of Revolution 1770–1823*. Ithaca: Cornell University Press.

Davis, D.B. 1984. *Slavery and Human Progress*. New York: Oxford University Press.

Deerr, N. 1949–50. *The History of Sugar*. 2 vols. London: Chapman & Hall.

Domar, E. 1970. The causes of slavery or serfdom: a hypothesis. *Journal of Economic History* 30, 18–32.

Drescher, S. 1977. *Econocide*. Pittsburgh: University of Pittsburgh Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Drescher, S. 1986. *Capitalism and Antislavery*. London: Macmillan.

Eltis, D. 1987. *Economic Growth and the Ending of the Translatlantic Slave Trade*. New York: Oxford University Press.

Engerman, S.L. 1982. Economic adjustments to emancipation in the United States and British West Indies. *Journal of Interdisciplinary History* 13(2), 191–220.

Engerman, S.L. 1983. Contract labour, sugar, and technology in the nineteenth century. *Journal of Economic History* 43, 635–59.

Evans, R., Jr. 1962. The economics of American Negro slavery. In Universities National Bureau Committee for Economic Research, *Aspects of Labor Economics*. Princeton: Princeton University Press.

Finley, M.I. 1980. *Ancient Slavery and Modern Ideology*. New York: Viking Press.

Fogel, R.W. and Engerman, S.L. 1971. The economics of slavery. In *The Reinterpretation of American Economic History*, ed. R.W. Fogel and S.L. Engerman. New York: Harper & Row.

Fogel, R.W. and Engerman, S.L. 1974a. *Time on the Cross*. 2 vols. 1: *The Economics of American Negro Slavery*; 2: *Evidence and Methods*. Boston: Little, Brown.

Fogel, R.W. and Engerman, S.L. 1974b. Philanthropy at bargain prices: notes on the economics of gradual emancipation. *Journal of Legal Studies* 3, 377–401.

Genovese, E.D. 1965. *The Political Economy of Slavery*. New York: Pantheon.

Genovese, E.D. 1974. *Roll, Jordan, Roll*. New York: Pantheon.

Higman, B.W. 1984. *Slave Populations of the British Caribbean, 1807–1834*. Baltimore: Johns Hopkins University Press.

Klein, H.S. 1986. *African Slavery in Latin America and the Caribbean*. New York: Oxford University Press.

Lovejoy, P.E. 1983. *Transformations in Slavery*. Cambridge: Cambridge University Press.

McCulloch, J.R. 1825. *The Principles of Political Economy*. Edinburgh: W. & C. Tait.

Mill, J.S. 1848. *Principles of Political Economy*. 2 vols. London: J.W. Parker.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Moreno Fraginals, M., Klein, H.S. and Engerman, S.L. 1983. The level and structure of slave prices on Cuban plantations in the mid-nineteenth century: some comparative perspectives. *American Historical Review* 88, 201–18.

Nieboer, H.J. 1900. *Slavery as an Industrial System*. The Hague: M. Nijhoff.

Patterson, H.O. 1982. *Slavery and Social Death*. Cambridge, MA: Harvard University Press.

Pryor, F.L. 1977. *The Origins of the Economy*. New York: Academic Press.

Solow, B.L. and Engerman, S.L., eds. 1987. *Caribbean Slavery and British Capitalism*. Cambridge: Cambridge University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. E. Cannan, New York: Modern Library, 1937.

Steckel, R.H. 1985. *The Economics of US Slave and Southern White Fertility*. New York: Garland Publishing.

Williams, E. 1944. *Capitalism and Slavery*. Chapel Hill: University of North Carolina Press.

Wright, G. 1978. *The Political Economy of the Cotton South*. New York: W.W. Norton.

Yasuba, Y. 1961. The profitability and viability of plantation slavery in the United States. *Economic Studies Quarterly* 12, September, 60–7. Repr. in *The Reinterpretation of American Economic History*, ed. R.W. Fogel and S.L. Engerman, New York: Harper & Row, 1971.

How to cite this article

Egerman, S. L. "slavery." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000145> doi:10.1057/9780230226203.1539

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Slichter, Sumner Huber (1892 – 1959)

John T. Dunlop
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Slichter, S. H.; forecasting

Article

Slichter was both a wide-ranging general economist and a scholar in industrial relations, regarding the two disciplines as parts of a seamless whole. He wrote an introductory economics textbook (1931) and was probably the most widely read economist by the general public of his day. He was a highly respected economic forecaster, Paul Samuelson calling him ‘our best economic forecaster for the period 1935–55’, and served as president of the American Economic Association, 1940–1. In industrial relations his two large classics (1941 and 1960) grew out of extended field work.

Slichter took his undergraduate degree in 1913 at the University of Wisconsin (where his father was professor of mathematics), did graduate work there with John R. Commons and completed his doctorate (1918) at the University of Chicago with H.A. Millis. He taught at Cornell for a decade, moved to Harvard Graduate School of Business Administration in 1930 and joined the Department of Economics in 1935. In 1940 he was appointed the first university professor at Harvard.

Among the themes that Slichter stressed were that the American economy was not in danger of stagnation; that the Second World War would not be followed by a depression but rather a boom; that America was becoming a ‘laboristic’ economy in which value judgements of the community reflect those of employees; that the challenges that unions have presented to managements have created superior and better-balanced managements; and that a vigorous and healthy economy is associated with an upward creep in prices that results from strong demand for goods and services and a slow climb in labour costs.

Selected works

1919. *The Turnover of Factory Labor*. New York: D. Appleton.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1931. *Modern Economic Society*. New York: Henry Holt.

1941. *Union Policies and Industrial Management*. Washington, DC: Brookings.

1943. *Present Savings and Postwar Markets*. New York: McGraw-Hill.

1947. *Trade Unions in a Free Society*. Cambridge, MA: Harvard University Press.

1948. *The American Economy: Its Problems and Prospects*. New York: Knopf.

1951. *What's Ahead for American Business*? Boston: Little, Brown.

1960. (With J.J. Healy and E.R. Livernash.) *The Impact of Collective Bargaining on Management*. Washington, DC: Brookings.

1961. *Potentials of the American Economy: Selected Essays of Sumner H. Slichter*, ed. J.T. Dunlop. Cambridge, MA: Harvard University Press. A full bibliography appears on pp. 435–56.

How to cite this article

Dunlop, John T. "Slichter, Sumner Huber (1892–1959)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000148> doi:10.1057/9780230226203.1540

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sliding scales (wages)

Christopher Hanes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

labour market contracts; sliding scales

Article

Theories of long-term employment contracts imply strong private incentives for employers and employees to link wages to product prices, quite apart from the external benefits emphasized by Weitzman (1984), whether or not wages are also indexed to other variables such as consumption goods' prices. Such arrangements have been extremely rare since the Second World War, but many schemes linking wage rates to product prices, referred to generally as 'sliding scales', were observed in Britain and the United States from the 1860s to the 1930s (Howard, 1920; Munro, 1885–6; Palgrave, 1896; Poole, 1938; US Industrial Commission, 1901, pp. 89–98, 135–6). Recent studies of historical sliding scales include Greenfield (1960), Treble (1987), South (1990), and Hanes (2007).

Most sliding scales were a result of negotiation or arbitration between unions and employers' associations in coal or iron-ore mining, or in the metals industries – iron, steel and tinplating. In the United States, sliding scales were also used in zinc, silver and copper mines, in glass manufacture, and in the textile mills of Fall River, Massachusetts, between 1905 and 1910. They were written agreements but not legally enforceable contracts, and were meant to hold for at most one year or for no fixed duration – either side could withdraw after some weeks' notice. Terms specified time- or piece-rate wages, but not employment levels. Nearly all sliding scales set minimum levels below which wages could not fall, no matter what happened to prices; some included maximums as well. Within these limits, wage adjustments took place at predetermined points in time, with intervals ranging from one week to six months, as a continuous or stepped-schedule function of prices observed in the previous interval. A number of scales based wages on the 'margin' between prices of products and material inputs. Agreements were frequently renegotiated in response to changes in external labour-market conditions or costs of non-labour inputs left unaccounted for in the scale's formula. For some sliding scales, prices were taken from press reports of open-market transactions. More frequently, prices were taken from employers' account books, examined by professional accountants approved by both parties or the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
arbitrator. Because firms did not want to reveal sales information to competitors, examiners typically reported only a summary statistic used in the wage formula.

In the United States, unions and their sliding scales were expelled from most iron and steel plants after the 1900s. Where unions survived they continued to negotiate sliding scales up to the 1920s (Daugherty, de Chazeau and Stratton, 1937, pp. 143–4). By the early 1940s, however, sliding scales had nearly disappeared, even in mining (US Bureau of Labor Statistics, 1940, p. 13). In Britain, sliding scales or similar ‘proceeds-sharing’ plans remained widespread within mining and the metals industries throughout the 1930s. These schemes were suspended at the beginning of the Second World War in response to the advent of price controls, and never revived (Burn, 1961, p. 27; Haynes, 1953, p. 26).

Hanes (2007) argues that pre-1940s sliding scales were not examples of indexation as described by theoretical employment-contract models, but rather devices to reduce the frequency of costly strikes in the absence of contracts; that their scope was severely limited by the inability of rank-and-file union members to observe product prices, even in industries where price information was relatively plentiful; and that they disappeared after the 1930s because US unions gained the ability to enter long-term contracts, while British mining and metals industries remained under forms of government control that broke the link between product prices and labour demand.

See Also

- wage indexation

Bibliography

Burn, D. 1961. *The Steel Industry 1939–1959: A Study in Competition and Planning*. Cambridge: Cambridge University Press.

Daugherty, C.R., de Chazeau, M.D. and Stratton, S.S. 1937. *The Economics of the Iron and Steel Industry*. New York: McGraw-Hill.

Greenfield, H.I. 1960. Sliding wage scales. Doctoral dissertation, University of Colombia.

Hanes, C. 2007. The rise and fall of the sliding scale, or why wages are no longer indexed to product prices. Working paper, Department of Economics, State University of New York at Binghamton.

Haynes, W.W. 1953. *Nationalization in Practice: The British Coal Industry*. Boston: Harvard University Graduate School of Business.

Howard, S. 1920. *The Movement of Wages in the Cotton Manufacturing Industry of New England since 1860*. Boston: National Council of American Cotton Manufacturers.

Munro, J.E.C. 1885–6. Sliding scales in the iron industry. *Transactions of the Manchester Statistical*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Society, 1–43.

Munro, J.E.C. 1889–90. Sliding scales in the coal and iron industries from 1885 to 1889. *Transactions of the Manchester Statistical Society*, 119–71.

Palgrave, R.H.I. 1896. Sliding scale (wages). *Dictionary of Political Economy*, vol. 2, London: Macmillan.

Poole, A.G. 1938. *Wage Policy in Relation to Industrial Fluctuations*. London: Macmillan.

South, N.M. 1990. Price contingent wage contracts in British coal and iron and steel 1860–1913: theory and evidence. Ph.D. dissertation, University of Toronto.

Treble, J.G. 1987. Sliding scales and conciliation boards: risk-sharing in the late 19th century British coal industry. *Oxford Economic Papers* 39, 679–98.

US Bureau of Labor Statistics. 1940. Wage-adjustment provisions in union agreements. *Monthly Labor Review* 50, 6–15.

US Industrial Commission. 1901. *Report of the Industrial Commission on the Relations and Conditions of Capital and Labor*. Washington: GPO.

Weitzman, M.L. 1984. *The Share Economy*. Cambridge, MA: Harvard University Press.

How to cite this article

Hanes, Christopher. "sliding scales (wages)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000453> doi:10.1057/9780230226203.1541

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Slutsky, Eugen (1880 – 1948)

Giancarlo Gandolfo

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Born in the Yaroslav province of Russia, Slutsky had troubled years as a student: he enrolled in the department of physics and mathematics at Kiev University, was expelled for taking part in student revolts, went abroad to the Munich Institute of Technology to study engineering and finally graduated in the department of law in 1911 back at Kiev University. He became a member of the faculty at Kiev Institute of Commerce in 1913 and full professor there in 1920. In 1926 he moved to Moscow as a staff member of the Conjuncture Institute; in 1934 he became a staff member of the Mathematical Institute of the University of Moscow and in 1936 a member of the Mathematical Institute of the Academy of Sciences, Moscow, a post which he held until his death.

Slutsky was a mathematician, statistician and economist. His fame as an economist rests mainly on one single contribution (1915), which went unnoticed until the 1930s, when it was discovered independently by Dominedò (1933, p. 790), Schultz (1935, pp. 439ff), and Allen (1936), and subsequently influenced the further development of consumer theory. Hicks – who, together with Allen (Hicks and Allen, 1934), had independently arrived at Slutsky's results – writes: ‘The theory to be set out in this chapter and the two following is essentially Slutsky's•…•The present volume is the first systematic exploration of the territory which Slutsky opened up’ (Hicks, 1939, p. 19). Building on earlier work by Pareto (who had already derived the formulae which express the change in the consumer's demand when any one of its arguments changes, but without seeing their implications), Slutsky showed that the effect of a price change on the quantity demanded can be divided into two effects. One is the effect of a *compensated variation* of price; if a price increases and the consumer is given an income increase so as to make possible the purchase of the same quantities of all the goods previously purchased, the individual – though being in the position to purchase the preceding bundle of goods – will no longer consider it preferable to any other, and there will take place some kind of *residual variation* of demand. This is called the *residual variability* by Slutsky (the substitution effect in Hicks's terminology). It should be noted that the compensated variation of price can also be defined in terms of the income change which leaves the consumer's *real* income unchanged, that is which causes the consumer to remain on the *same* indifference curve (this is the concept used by Hicks, 1939, in the text, while in the mathematical appendix he gives the same definition as Slutsky). Although the two definitions are equivalent for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
infinitesimal changes (as was first shown by Mosak, 1942), Slutsky's is preferable from the operational point of view since it does not require knowledge of the consumer's indifference map. The other effect is the *income effect*, which gives the change in the consumer's purchases when his money income changes at unchanged prices. The two effects turn out to be independent and additive and their algebraic sum gives the *price effect*: this is, in Hicks' terminology, the 'Fundamental Equation of Value Theory', also called the Slutsky Equation.

Slutsky proved the complete properties of the various effects and of the demand curves. The income effect may be either normal (demand increases as income increases: 'relatively indispensable goods' in Slutsky's terminology) or abnormal ('relatively dispensable' goods). The 'own' substitution effect is always negative ('The residual variability of a good in the case of a compensated variation of its price, is always negative', [1915] 1953, p. 42) and the cross substitution effect is symmetric ('The residual variability of the *j*-th good in the case of a compensated variation of the price p_i is equal to the residual variability of the *i*-th good in the case of a compensated variation of the price p_j', [1915] 1953, p. 43). The 'own' price effect, therefore, is necessarily normal in the case of relatively indispensable goods.

Slutsky also proved the relation which implies that the individual demand functions are homogeneous of degree zero. He gave a definition of complementary and competing goods, and made an important methodological point which is usually overlooked in his contribution: he stressed the need for *experiment* in order to obtain all the values of the relevant magnitudes (which cannot be obtained by observation of existing budgets) which enter into the definition. This emphasis on the need for experimental verification of economic laws, which concludes his contribution, is worthy of note and obviously arises from his statistical background.

Slutsky did no other noticeable work in economics but made important contributions to mathematical statistics and probability theory.

In (1914) he suggested the use of a χ^2 variate to test the goodness of fit of a regression line ('line' is taken in the broad sense, i.e. including both straight lines and curves); as a logical consequence, he introduced the concept of *minimum chi-square estimator* ('the most probable values of the coefficients will be those which bring our χ^2 to a minimum', 1914, p. 83) as a general method of fitting regressions. This paper was written several years before R.A. Fisher's work on the subject.

Slutsky was one of the originators of the theory of stochastic processes and time-series analysis. In his renowned (1927) paper he proved several important theorems. One is that the summation of random causes may be the source of cyclic or undulatory processes, and that these waves will show an approximate regularity in the sense that they can be approximated quite well by a relatively small number of terms (sine curves) of the Fourier series. Another is the sinusoidal limit theorem, which states that under certain conditions the summation of random causes will tend to give rise to a specific sine wave. For example, if one takes a moving average (of two terms) of a random series *n* times and then takes the *m*th difference of the result, and lets $n \to \infty$ so that *m/n* tends to a constant *c* between zero and one, it follows that the series will tend to a sine wave with wavelength arc cos $(1-c)/(1+c)$. A corollary of these theorems is the famous Slutsky-Yule Effect (so named because it was also independently discovered by Yule): if a moving average of a random series is taken (for example to determine trend), this may *generate* an oscillatory movement in the series where none existed in the original data.

Slutsky also worked in the theory of probability, where he studied the concept of asymptotic convergence in probability (e.g. 1925, 1928, 1929). He spent the last years of his life in preparing tables

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for the computation of the incomplete gamma-function and the chi-square probability distribution (1950).

Selected works

A bibliography (1912–46) is contained in the memorial article (in Russian) by A.N. Kolmogorov, in *Uspekhi Matematicheskikh Nauk*, Vol. 3, No. 4, 1948. This bibliography is reproduced in Allen (1950). A collection of selected papers was published posthumously in Russian (1960). On Slutsky's life and works see also the memorial article (in Russian) by N. Smirnov, in *Izvestiia Akademiia Nauk SSSR*, Mathematical Series, Vol. 12, 1948, and Allen (1950).

1914. On the criterion of goodness of fit of the regression lines and on the best method of fitting them to the data. *Journal of the Royal Statistical Society* 77, Pt I, December, 78–84.

1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti e Rivista di Statistica* 51 (July), 1–26 Trans. as ‘On the theory of the budget of the consumer’ in *Readings in Price Theory*, ed. K. E. Boulding and G.J. Stigler, London: Allen & Unwin, 1953, 26–56.

1925. Über stochastische Asymptoten und Grenzwerte. *Metron* 5(December), 3–89.

1927. The summation of random causes as the source of cyclic processes. (In Russian.) *Problems of Economic Conditions*, The Conjuncture Institute, Moscow, Vol. 3, No. 1. Revised English version in *Econometrica* 5(April) 1937, 105–46.

1928. Sur un critérium de la convergence stochastique des ensembles de valeurs éventuelles. *Comptes rendu des séances de l'Académie des Sciences* 187, Paris, 17 July to 13 August.

1929. Quelques propositions sur les limites stochastiques éventuelles. *Comptes rendu des séances de l'Académie des Sciences* 189, Paris, 2 September.

1950. *Tables for the computation of the incomplete Gamma-function and the Chi-square probability distribution*. (In Russian.) Ed. A.N. Kolmogorov, Moscow: Akademiia Nauk SSSR (posthumous).

1960. *Selected Works*. (In Russian.) Moscow: Akademiia Nauk SSSR (posthumous).

Bibliography

Allen, R.G.D. 1936. Professor Slutsky's theory of consumers’ choice. *Review of Economic Studies* 3 (February), 120–29.

Allen, R.G.D. 1950. The work on Eugen Slutsky. *Econometrica* 18(July), 209–16.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dominedò, V. 1933. Considerazioni intorno alla teoria della domanda, I-ILe principali premesse e caratteristiche delle curve statiche. *Giornale degli Economisti e Rivista di Statistica* 48(November), 765–807.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press. 2nd edn, 1946.

Hicks, J.R. and Allen, R.G.D. 1934. A reconsideration of the theory of value. *Economica* 1, Pt I, February, 52–76; Pt II, May, 196–219.

Mosak, J. 1942. On the interpretation of the fundamental equation of value theory. In *Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz*, ed. O. Lange, F. McIntyre and T.O. Yntema. Chicago: University of Chicago Press, 69–74.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press Enlarged edn, 1983.

Schultz, H. 1935. Interrelations of demand, price, and income. *Journal of Political Economy* 43 (August), 433–81.

How to cite this article

Gandolfo, Giancarlo. "Slutsky, Eugen (1880–1948)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000150> doi:10.1057/9780230226203.1542

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

small-world networks

Duncan J. Watts
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The ‘small-world hypothesis’ expresses the idea, long an article of popular belief, that every individual in a given population can reach every other via some ‘short’ chain of intermediaries. Exactly what is ‘short’, how individuals ‘reach’ one other, and how the hypothesis, if true, relates to the structure of social networks are all qsts on which recent progress has been made. Here I describe the genesis of the hypothesis, discuss supporting experimental evidence, and explore some recent mathematical models of small-world networks.

Keywords

small-world hypothesis; small-world networks

Article

Small-world networks – so named on account of their resemblance to the ‘small-world hypothesis’ – exhibit short global path lengths in spite of considerable local structure. The small-world hypothesis, long an article of popular belief (Guare, 1990), was first investigated by Pool and Kochen (1978), who posed the following qst: ‘How many pairs of persons in the population can be joined by a single acquaintance, how many by a chain of two persons, how many by a chain of three, etc?’ Pool and Kochen showed that when all individuals choose their acquaintances uniformly at random from the entire population, the number of pairs *i*, *j* connected by a path of length *d* increases exponentially fast in *d*. Thus under quite reasonable assumptions regarding the average number *k* of acquaintances per person, most pairs of individuals in even a very large population should be connected by paths only a few steps long. For example, for the population of the United States at the time – roughly 200 million – and assuming that individuals possessed an average of roughly 1,000 acquaintances, Pool and Kochen estimated that a randomly selected pair could be connected in only three or four steps.
This result, it turns out, is equivalent to a standard result from the theory of random graphs (Bollobas, 1985; Erdos and Renyi, 1959; Solomonoff and Rapoport, 1951), namely, that the average shortest path length separating two randomly chosen individuals should be proportional to log•*N* (for $N \gg k \gg 1$), where *N* is the size of the population. (Strictly speaking, this result holds only when the network is connected, which is not guaranteed for small *k*; hence the second condition $k \gg 1$.) From this result, we can infer our first definition of what it means for the world to be ‘small’:
Definition 1: : A network can be said to exhibit the small-world property when the average shortest path length of the network $L \propto \log N$.
Pool and Kochen noted, however, that social networks, far from being random, exhibit considerable structure; that is, individuals are more likely to know each other when they share certain traits, such as geographical proximity, socio-economic status or common interests. The interesting (but far less tractable) version of the small-world hypothesis is therefore not that path lengths can be short but that path lengths should remain short even when the network in question is far from random. Pool and Kochen also speculated that the increase in *L* resulting from the introduction of social structure to a random network would be a good measure of that structure. As we shall see, however, the validity of the small-world hypothesis actually implies the converse of this claim.

Six degrees of separation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Stimulated by Pool and Kochen's paper, the social psychologist Stanley Milgram, along with his graduate student Jeffrey Travers, decided to test the small-world hypothesis experimentally, using what they called the ‘small-world method’ (Milgram, 1967; Travers and Milgram, 1969):

1. 1. Recruit one or more volunteers to be ‘targets’ of the experiment.
2. 2. For each target, recruit some large number of initial ‘senders’.
3. 3. Give each sender a description of the target sufficient to identify him or her uniquely (for example, name, address, occupation and employer, wife's maiden name, and so on).
4. 4. Instruct each sender to forward a message to the target in one of two ways:
    1. (a) If target is an acquaintance (that is, is known to sender on a first-name basis), forward directly to target.
    2. (b) If not, forward message, target information, and instructions to an acquaintance who is ‘closer’ to target than sender.
5. 5. Repeat steps 3 and 4 with additional senders until chain either *completes* (reaches target) or *terminates* (is not forwarded for any other reason). In Milgram's version, subsequent senders also received a list of previous senders, in order to avoid cycles; however, in practice these cycles are extremely unlikely, and this requirement has subsequently been dropped (Dodds, Muhamad and Watts, 2003).

Travers and Milgram implemented this protocol using a single target (a Boston stockbroker who was an acquaintance of Milgram's) and 296 senders: 100 from Boston, 100 blue-chip stockholders from Omaha, Nebraska, and 96 Omaha residents who were randomly selected from a list of people who had agreed to receive marketing literature. Famously, they found that the chains which reached the target were surprisingly short – approximately six steps on average, where, unsurprisingly, chains initiated by stockholders in Nebraska were shorter than those from randomly selected Nebraskans, and chains that began in Boston were shorter still. However, most chains (roughly 80 per cent) terminated before completion. Travers and Milgram also noted that a disproportionate fraction of completed chains were delivered at the last step by a small number of individuals (one man delivered 16 letters) whom Travers and Milgram dubbed ‘sociometric stars’. Subsequently, small-world experiments have been conducted a number of times for different-sized populations (for reviews, see Garfield, 1979; Kleinfeld, 2002; Kochen, 1989). In most of these studies completion rates are low, but the chains that do complete are short (Dodds, Muhamad and Watts, 2003; Korte and Milgram, 1970).

If one assumes that chains terminate out of insufficient interest or motivation on the part of senders, and not because the underlying network is disconnected or non-navigable (this assumption has been disputed recently by Kleinfeld, 2002, but no internally self-consistent alternative has yet been proposed), then it is possible to compute the hypothetical distribution of chain lengths corresponding to zero attrition (Dodds, Muhamad and Watts, 2003; White, 1970). Because longer chains have more opportunities to terminate, this ‘ideal’ distribution necessarily yields higher estimates of chain length than estimates based solely on completed chains. Nevertheless, the revised estimates – typically between seven and nine steps – are still consistent with the small-world hypothesis, even for reasonably high attrition rates. Small-world experiments therefore suggest two surprising features of large social networks: (*a*) in spite of their considerable structure, any pair of randomly selected individuals is likely to be connected via some short path; and (*b*) these individuals can actually find such a path, given only local information about the network.

Small-world networks

Addressing the first property – that randomly selected individuals in a large network can be connected via a short chain of intermediaries – Watts and Strogatz (1998) analysed a network model that incorporated elements of both social structure and randomness. In the model ‘social structure’ was represented by a uniform one-dimensional lattice, where each node was connected to its *k* nearest neighbours on the lattice, and ‘randomness’ was characterized by a tunable parameter *p*, which specified the probability that a link in the lattice would be *randomly rewired* (see Figure 1). Defining the *clustering coefficient C* of a network as the average probability that two neighbours of a given node would themselves be neighbours, they showed that when $p = 0$ (completely ordered) the network is ‘large’ $(L(0) \sim N / 2k)$ and ‘highly clustered’ $(C(0) \simeq 3 / 4)$, and when $p = 1$ (completely random) it is ‘small’ $(L(1) \sim \log N / \log k)$ and ‘poorly clustered’ $(C(1) \sim k / N)$, thus suggesting that path lengths are short only when clustering is low. In contrast with Pool and Kochen's prediction, however, they showed that the model exhibits a broad region of *p* values in which *C*(*p*) is high relative to its random limit *C*(1), yet *L*(*p*) is roughly speaking as ‘small’ as possible (see Figure 2).

Figure 1

Schematic of the Watts–Strogatz model of partly ordered, partly random networks. Note: *p*=the probability of random edge rewiring.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

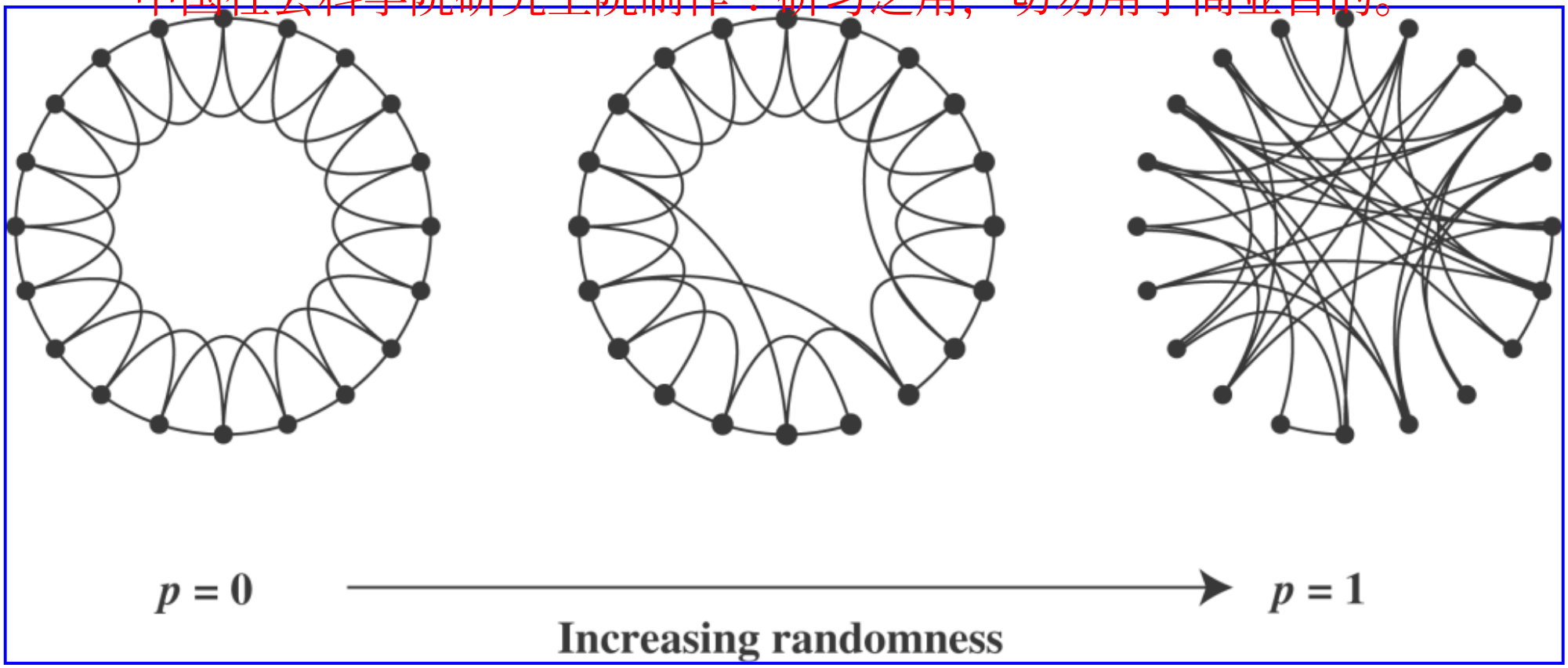

Figure 2
Normalized average path length $L(p)/L(0)$ and clustering coefficient $C(p)/C(0)$ versus p (note log scale)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

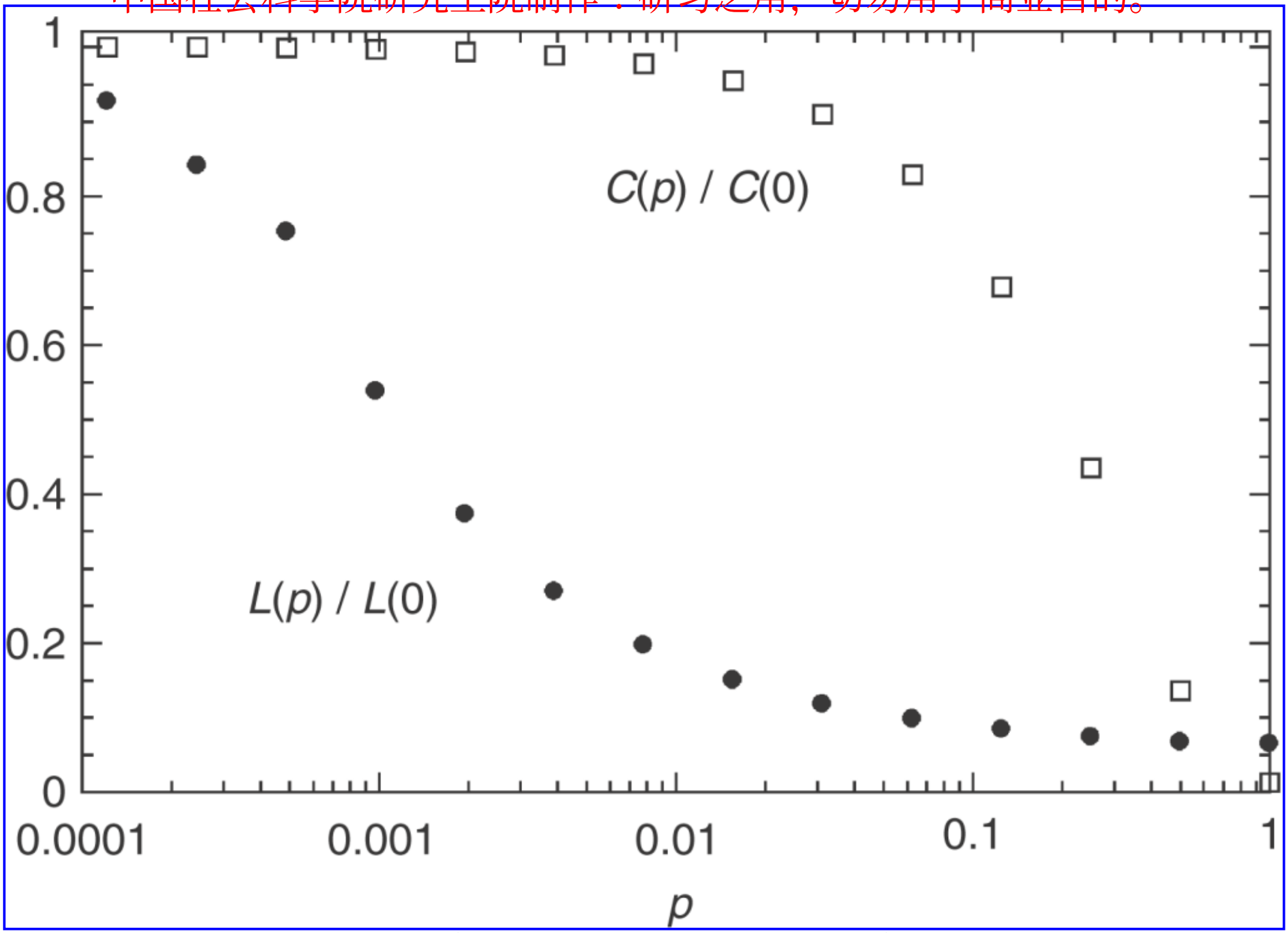

Watts and Strogatz coined the term ‘small-world networks’ to refer to networks in this class. Because the conditions required for any network to exhibit small-world properties (just a small fraction of long-range, random ‘short cuts’) were relatively weak, Watts and Strogatz predicted that many real-world networks – whether social networks or otherwise – would be small-world networks. They then checked this prediction by considering three network data-sets – the affiliation network of movie actors, the power transmission grid of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

western United States, and the neural network of the nematode *C. elegans* – and found that all three examples satisfied the small-world criteria (Watts and Strogatz, 1998). Numerous authors have since investigated the properties of this model, and many large empirical networks have been found to exhibit small-world-like properties (see Newman, 2000; 2003, and Watts, 2004, for reviews of this literature, as well as Newman, Barabasi and Watts, 2006, for a collection of early papers).

Searchable small-world networks

The small-world network models of the kind proposed by Watts and Strogatz, however, do not satisfy the second striking feature of Travers and Milgram's results, namely, that individuals can locate short paths using only local information about the network structure. This point, first raised by Kleinberg (2000a; 2000b), led him to propose a more general class of partly ordered, partly random networks in which the spatial distribution of random links is permitted to vary according to the probability distribution $P_{ij} \propto r_{ij}^{-\gamma}$, where r_{ij} is the distance between nodes *i* and *j* measured on some underlying lattice of dimension *D*, and γ is a tunable parameter. Kleinberg then proved that only when $\gamma = D$ will the network be searchable in the sense that short paths not only exist but are also discoverable: when $\gamma < D$ short paths exist, but cannot be found using only local information; and when $\gamma > D$ the shortest paths, although discoverable, are not short. Subsequently, more realistic models of searchable small-world networks have been proposed, incorporating for example 'social' (Adamic and Adar, 2005; Watts, Dodds and Newman, 2002) as well as geographical (Liben-Nowell et al., 2005) distance, and also heterogeneity of degree (Simsek and Jensen, 2005).

What is a 'short' path anyway?

Although the observed path lengths in small-world experiments do indeed seem to be 'short', it is not clear that they are short in the sense of Definition 1 above; that is, proportional to $\log N$. Indeed, short of conducting controlled experiments in which *N* is varied systematically over several orders of magnitude, strict logarithmic scaling of chain lengths would be difficult to establish empirically. Furthermore, Chung and Lu (2002) have demonstrated that so-called 'scale-free' random networks, whose degree distributions exhibit 'power law' tails (that is, $P[K > k] \propto k^{-\alpha}$ for $k \gg 1$, where $1 < \alpha < 2$), exhibit even shorter average path lengths than in Definition 1 (formally, $L \propto \log\log N$); whereas in contrast Kleinberg (2000b) has suggested that the small-world hypothesis is consistent with longer path lengths (formally, $L \propto (\log N)^{\mu}$, where $\mu \geq 1$). Ultimately, however, it may not matter, as for realistic network sizes all these definitions yield similar answers, and thus are unlikely to be empirically distinguishable. Also, in the presence of chain attrition, the absolute length of chains is more relevant than its scaling behaviour: from a practical perspective, if a chain doesn't complete, it doesn't matter whether or not it is as short as it could be; it is still too long. For these reasons, Watts, Dodds and Newman (2002) have proposed an alternative definition of 'short':

Definition 2: : Given a message failure probability *p* and minimal required chain completion rate *r*, the small-world hypothesis requires (approximately) that average path length $L \leq \log r / \log(1 - p)$.

Sociometric stars

The final result of Travers and Milgram above – that chains tend to be completed by highly connected individuals, called variously 'stars', 'hubs', 'connectors' or 'funnels' – appears to be supported by recent theoretical models (Adamic et al., 2001; Adamic, Lukose and Huberman, 2003), which suggest that when otherwise random networks are characterized by extremely skewed degree distributions, the most connected individuals do indeed serve a critical role in directing decentralized searches. The hubs in these models, however, are analogous to airline network hubs, which tend to be connected directly to a large fraction of total airports in a given region, and, in particular, are connected to most other hubs. In social networks, by contrast, even the most gregarious individuals are estimated to know only on the order of 1,000 others (Bernard et al., 1991; Killworth et al., 1990), and thus are adjacent to a negligible fraction of the entire network. Hubs in social networks, therefore, cannot play the role that they do in so-called 'scale-free' networks. Furthermore, they do not need to: empirical work shows no evidence that 'special individuals' are required to resolve social search problems (Dodds, Muhamad and Watts, 2003), and recent theoretical models suggest that networks can be searchable even when all individuals have exactly the same acquaintance volume (Kleinberg, 2000a; 2000b; Watts, Dodds and Newman, 2002).

See Also

- mathematics of networks

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- network formation
- psychology of social networks
- social networks in labour markets

Bibliography

Adamic, L. and Adar, E. 2005. How to search a social network. *Social Networks* 27, 187–203.

Adamic, L., Lukose, R. and Huberman, B. 2003. Local search in unstructured networks. In *Handbook of Graphs and Networks: From the Genome to the Internet*, ed. S. Bornholdt and H. Schuster. Weinheim: Wiley-VCH.

Adamic, L., Lukose, R., Puniyani, A. and Huberman, B. 2001. Search in power-law networks. *Physical Review E* 64, 46135–43.

Bernard, H., Johnsen, E., Killworth, P. and Robinson, S. 1991. Estimating the size of an average personal network and of an event population: some empirical results. *Social Science Research* 20, 109–21.

Bollobas, B. 1985. *Random Graphs*. New York: Academic Press.

Chung, F. and Lu, L. 2002. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences of the United States of America* 99, 15879–82.

Dodds, P., Muhamad, R. and Watts, D. 2003. An experimental study of search in global social networks. *Science* 301, 827–9.

Erdos, P. and Renyi, A. 1959. On random graphs. *Publicationes Mathematicae* 6, 290–7.

Garfield, E. 1979. It's a small world after all. In *Essays of an Information Scientist*, ed. Garfield. Philadelphia: ISI Press.

Guare, J. 1990. *Six Degrees of Separation: A Play*. New York: Vintage.

Killworth, P., Johnsen, E., Bernard, R., Shelley, G. and McCarty, C. 1990. Estimating the size of personal networks. *Social Networks* 12, 289–312.

Kleinberg, J. 2000a. Navigation in a small world. *Nature* 406–845.

Kleinberg, J. 2000b. The small-world phenomenon: an algorithmic perspective. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*. New York: Association of Computing Machinery.

Kleinfeld, J. 2002. The small world problem. *Society* 39, 61–6.

Kochen, M., ed. 1989. *The Small World*. Norwood, NJ: Ablex.

Korte, C. and Milgram, S. 1970. Acquaintance linking between white and negro populations: application of the small world problem. *Journal of Personality and Social Psychology* 15, 101–18.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. and Tomkins, A. 2005. Geographic routing in social networks. *PNAS* 102, 11623–8.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Milgram, S. 1967. The small world problem. *Psychology Today* 2, 60–7.

Newman, M. 2000. Models of the small world. *Journal of Statistical Physics* 101, 819–41.

Newman, M. 2003. The structure and function of complex networks. *Siam Review* 45, 167–256.

Newman, M., Barabasi, A.-L. and Watts, D., eds. 2006. *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press.

Pool, I. and Kochen, M. 1978. Contacts and influence. *Social Networks* 1, 1–48.

Simsek, O. and Jensen, D. 2005. Decentralized search in networks using homophily and degree disparity. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (IJCAI 2005). Edinburgh.

Solomonoff, R. and Rapoport, A. 1951. Connectivity of random nets. *Bulletin of Mathematical Biophysics* 13, 107–17.

Travers, J. and Milgram, S. 1969. An experimental study of the small world problem. *Sociometry* 32, 425–43.

Watts, D. 2004. The ‘new’ science of networks. *Annual Review of Sociology* 30, 243–70.

Watts, D., Dodds, P. and Newman, M. 2002. Identity and search in social networks. *Science* 296, 1302–5.

Watts, D. and Strogatz, S. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–2.

White, H. 1970. Search parameters for small world problem. *Social Forces* 49, 259–64.

How to cite this article

Watts, Duncan J. "small-world networks." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000481> doi:10.1057/9780230226203.1543

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Smart, William (1853 – 1915)

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; bimetallism; income distribution; single tax; Smart, W.; tariff reform

Article

William Smart was an entrepreneur turned academic economist. Born at Barrhead in Renfrewshire on 10 April 1853, he was educated at the University of Glasgow, where he was later to occupy the (newly created) Adam Smith Chair of Political Economy from 1896 until his death on 19 March 1915. However, his transition from student to professor was interrupted by a successful career in industry which began in the early 1870s and terminated in 1884 when the firm in which he was a partner was sold to the considerable financial advantage of its principals.

Smart's main contribution to economics probably remains his translations into English of the work of Böhm-Bawerk (1890; 1891a), and his edition of von Wieser (1893). In certain circles, Smart is felt to have been primarily responsible for introducing the work of the Austrian School to English readers. As well as making available the originals, Smart published in 1891 his own account of Austrian economics under the title *Introduction to the Theory of Value* – a book which went through three editions during Smart's lifetime. Smart's other work includes a book (1895) dealing principally with wages, consumption and currency. It seems that Smart's advocacy of a bimetallic standard had made his election to the Adam Smith Chair at Glasgow in 1896 more problematic than it might have been. He also wrote on the distribution of income (1899), the single tax (1900), and tariff reform (1904), the last two being essentially contributions to popular debates of the day.

As a young man, and while still a practising businessman, Smart was a Ruskinite. He was a member of the Guild of St George and his first publication was his inaugural address as president of the Ruskin Society of Glasgow (1880). Smart's own account of these intellectual influences on his early development can be found in his *Second Thoughts of an Economist* (1916).

Separate mention should be made of Smart's *Economic Annals of the Nineteenth Century* (1910–17), which he began as a result of the difficulties he had experienced in gathering information in his role as member of the Poor Law Commission in 1905. The simple rationale was to render more accessible

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

official material related to actual economic conditions and debates of the period. Although Smart only saw through to completion two volumes of the *Annals* before he died (which cover less than one third of the 19th century), a glance at the material assembled in the extant volumes is sufficient to confirm their value.

Selected works

1880. *John Ruskin: His Life and Work*. Manchester: A. Heywood & Sons.

1883. *A Disciple of Plato*. Glasgow: Wilson & McCormick.

1890. trans. Böhm-Bawerk E.v., *Capital and Interest*. London and New York: Macmillan & Co.

1891a. trans. Böhm-Bawerk E.v., *Positive Theory of Capital*. London and New York: Macmillan & Co.

1891b. *Introduction to the Theory of Value on the Lines of Menger, Wieser and Böhm-Bawerk*. London: Macmillan. 2nd edn, 1910; 3rd edn, 1914.

1893. ed. Wieser F.v., *Natural Value*. Trans. C.A. Malloch. London and New York: Macmillan & Co.

1895. *Studies in Economics*. London: Macmillan & Co.

1899. *The Distribution of Income*. London: Macmillan & Co.

1900. *The Taxation of Land Values and the Single Tax*. Glasgow: J. MacLehose & Sons.

1904. *The Return to Protection*. London: Macmillan & Co.

1910–17. *Economic Annals of the Nineteenth Century*. 2 vols. London: Macmillan & Co.

1916. *Second Thoughts of an Economist*. London: Macmillan & Co.

How to cite this article

Milgate, Murray. "Smart, William (1853–1915)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000153> doi:10.1057/9780230226203.1544

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Smith, Adam (1723 – 1790)

Andrew Skinner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Smith's conception of ‘economic man’ was primarily a product of his moral philosophy. While defending the motive of self-interest against Hutcheson's claim that it could never be virtuous, he emphasized that self-interested actions take place within a social setting and that humanity is generally motivated by a desire for approbation. Far from an advocate of laissez-faire, Smith envisaged a broad, open-ended agenda of government to rectify market failure, including education to offset the atomizing effects of urbanization. His prowess arguably rests on his sophisticated grasp of the economic process as opposed to any outstanding analytical or conceptual competence.

Keywords

analogy; barriers to entry; Burke, E.; Cannan, E.; capital accumulation; Carmichael, G.; circular flow; circulating capital; city and economic development; civic humanism; civil society; division of labour; economic freedom; economic man; equilibrium; exchange economy; exchange value versus use value; Ferguson, A.; fixed capital; human capital; Hume, D.; Hutcheson, F.; impartial spectator; invisible hand; Justice; Kames, Lord; labour theory of value; Mandeville, B.; market failure; mercantilism; Mill, J.S.; Millar, J.; money; monopoly; natural price; net advantages; Poor Law, old; Physiocracy; price determination; productive and unproductive labour; profit; property; public works; Pufendorf, S. von; Quesnay, F.; quit-rent system; *rate* of exchange; rate of profit; Robertson, W.; rules of behaviour; self-interest; slavery; Smith, A.; special interests; specie-flow mechanism; stages theory of history; stationary state; Steuart, Sir J.; Stewart, D.; structural unemployment; subsistence wage; system of natural liberty; *Tableau économique* ; tax farming; Turgot, A.R.J.; use value versus exchange value; value; wages fund

Article

Biographical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Adam Smith was born in Kirkcaldy, on the east coast of Scotland, and baptized on 5 June 1723. He was the son of Adam Smith, Clerk to the Court Martial and Comptroller of Customs in the town (who died before his son was born) and of Margaret Douglas of Strathendry.

Smith attended the High School of Kirkcaldy, and then proceeded to Glasgow University. He first matriculated in 1737, at the not uncommon age of 14. At this time the university, or more strictly the college, was small. It housed only 12 professors who had in effect replaced the less specialized system of regents by 1727. Of the professoriate, Smith was most influenced by the 'never-to-be-forgotten' Francis Hutcheson (Corr., letter 274, dated 16 November 1787). Hutcheson had succeeded Gerschom Carmichael, the distinguished editor of Pufendorf's *De Officio Hominis et Civis* as Professor of Moral Philosophy.

Smith left Glasgow in 1740 as a Snell Exhibitioner at Balliol College to begin a stay of six years. The atmosphere of the college at this time was Jacobite and 'anti-Scotch'. Smith was also to complain: 'In the university of Oxford, the greater part of the publick professors have, for these many years, given up altogether even the pretence of teaching' (WN,V.i.f.8). But there were benefits, most notably ease of access to excellent libraries, which in turn enabled Smith to acquire an extensive knowledge of English and French literature, which was to prove invaluable, not least in terms of his knowledge of the sciences.

Smith left Oxford in 1746 and returned to Kirkcaldy without a fixed plan. But in 1748 he was invited to give a series of public lectures in Edinburgh, with the support of three men – the Lord Advocate, Henry Home; Lord Kames; and a childhood friend, James Oswald of Dunnikier.

The lectures, which are thought to have be *primarily* (not exclusively) concerned with rhetoric and belles letters, brought Smith £100 a year (Corr. letter 25, dated 8 June 1758). They also seem to have been wide-ranging.

Smith's reputation as a lecturer brought its reward. In 1751 he was elected to the Chair of Logic in Glasgow University, again with the support of Lord Kames. According to John Millar, Smith's most distinguished pupil, he devoted the bulk of his time to the delivery of a system of rhetoric and belles lettres, which was based on the conviction that the best way of:

> explaining and illustrating the various powers of the human mind, the most useful part of metaphysics, arises from an examination of the several ways of communicating our thoughts by speech, and from an attention to the principles of those literary compositions which contribute to persuasion or entertainment. (Stewart, I. 16)

Smith continued to teach the main part of his lecture course on logic after he had been translated to the Chair of Moral Philosophy in 1752. A set of lecture notes, discovered by J.M. Lothian in 1958, relate to the session 1762/3. The notes correspond closely to Millar's description of the course given more than a decade earlier, in that they are concerned with such problems as the development of language, style and the organization of forms of discourse which include the oratorical, narrative and didactical (scientific). Smith was primarily concerned with the study of human nature and with the analysis of the means and forms of communication. He no doubt continued to lecture on these subjects to students of moral philosophy because he rightly believed them to be important (see J.M. Lothian, 1963: W.S. Howell, 1975).

Smith's lectures on language were published in expanded form as *Considerations Concerning the First*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Formation of Language, in the *Philological Miscellany* for 1761. They were reprinted in the third edition of the *Theory of Moral Sentiments* in 1767.

Smith's teaching from the Chair of Moral Philosophy fell into four parts and in effect set the scene for the major published works which were to follow. Again on the authority of John Millar, it is known that Smith lectured on natural theology, ethics, jurisprudence and ‘expediency’ or economics, *in that order*. The lectures on natural theology (a sensitive subject at the time) have not yet been found. But Millar made it clear that the lectures on ethics form the basis for the *Theory of Moral Sentiments* and that the subjects covered in the last part of the course were to be further developed in the *Wealth of Nations* (Stewart, I. 20). As to the third part, on jurisprudence, Millar noted that:

> Upon this subject he followed the plan that seems to be suggested by Montesquieu; endeavouring to trace the gradual progress of jurisprudence, both public and private, from the rudest to the most refined ages, and to point out the effects of those arts which contribute to subsistence, and to the accumulation of property, in producing correspondent improvements or alterations in law and government. (Stewart, I. 19)

Illustration and confirmation of this claim proved impossible until 1896 when Edwin Cannan published an edition of the *Lectures on Jurisprudence*. The notes edited by Cannan are dated 1766, although they were taken in the session 1763/4. This was Smith's last session in Glasgow, so that these lectures, where ‘public’ (broadly constitutional law) precedes ‘private’ jurisprudence (concerning man's rights as a citizen), may reflect a preferred order. A second set of notes, this time relating to the previous session, were also found by J.M. Lothian as recently as 1958 and are here styled LJA.

Academically, the major event for Smith was the publication of the *Theory of Moral Sentiments* in 1759. The book was well received by both the public and Smith's friends. In a delightful letter Hume reminded Smith of the futility of fame and public approbation, and having encouraged him to be a philosopher in practice as well as profession, continued:

> Supposing therefore, that you have duly prepared yourself for the worst by these Reflections; I proceed to tell you the Melancholy News, that your Book has been most unfortunate: For the Public seem disposed to applaud It extremely. (Corr, letter 31, dated 12 April 1759)

The book was to establish Smith's reputation. There was a second revised edition in 1761 and further editions in 1767, 1774, 1781 and 1790.

Charles Townshend was among those to whom Hume had sent a copy of Smith's treatise. Townshend had married the widowed Countess of Dalkeith in 1755 and was sufficiently impressed by Smith's work as to arrange for his appointment as tutor to her son, the young Duke of Buccleuch. The position brought financial security (£300 sterling p.a. for the rest of his life), and Smith duly accepted, formally resigning his chair early in 1764.

Smith and his party left almost immediately for France to begin a sojourn of some two years. At the outset, the visit was unsuccessful, causing Smith to write to Hume, with some humour, that ‘I have begun to write a book in order to pass away the time. You may believe I have very little to do’ (Corr., letter 82, date 5 July 1764, Toulouse).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

But matters improved with Smith's increasing familiarity with the language and the success of a series of short tours. In 1765 Smith, the Duke, and the Duke's younger brother Hew Scott, reached Geneva, giving Smith an opportunity to meet Voltaire, whom he greatly admired as ‘the most universal genius perhaps which France has ever produced’ (Letter, 17). The party arrived in Paris in mid-February 1766, where Smith's fame, together with the efforts of David Hume, secured him a ready entrée to the leading *salons* and, in turn, introductions to *philosophes* such as d'Alembert, Holbach and Helvetius.

During this period Smith met François Quesnay, the founder, with the Marquis de Mirabeau, of the Physiocratic School of economics (Meek, 1962). By the time Smith met Quesnay, the latter's model of the economic system as embodied in the *Tableau économique* (1757, trans. in Meek, 1962) had already been through a number of editions. Quesnay was then working on the *Analyse* (trans. in Meek, 1962), while it is also known that A.R.J. Turgot was currently engaged on his *Reflections on the Formation and Distribution of Riches* (trans. in Meek, 1973).

Smith, who had already developed an interest in political economy, had arrived in Paris at the very point in time that the French School had reached the zenith of its influence and output. The contents of Smith's library amply confirm his interest in this work (Mizuta, 2000).

Smith's stay in Paris had been enjoyable both socially and in academic terms. But it was marred by the developing quarrel between Hume and Rousseau and sadly terminated by the death of Hew Scott. Smith returned to London on 1 November 1766.

Smith spent the winter in London, where he was consulted by Townshend and engaged in corrections for the third edition of the *Theory of Moral Sentiments*. By the spring of 1767 (the year in which Sir James Steuart published his *Principles of Political Oeconomy*) Smith was back in Kirkcaldy to begin a study of some six years. It was during this period that he struggled with the *Wealth of Nations*. Correspondence of the time amply confirms the mental strain involved. But by 1773 Smith was ready to return to London, leaving his friends, notably David Hume, under the impression that completion was imminent. As matters turned out, it took Smith almost three more years to finish his book; a delay which may have been due to part to his increasing concern with the American War of Independence and with the wider issue of the relationship between the colonies and the ‘mother country’ (WN, IV. vii).

An Inquiry into the Nature and Causes of the Wealth of Nations was published by Strahan and Cadell on 9 March 1776, and elicited once more a warm response from Hume:

> Dear Mr. Smith: I am much pleas'd with your Performance, and the Perusal of it has taken me from a State of great Anxiety. It was a Work of so much Expectation, by yourself, by your Friends, and by the Public, that I trembled for its Appearance; but am now much relieved. Not but the Reading of it necessarily requires so much Attention, and the Public is disposed to give so little, that I shall still doubt for some time of its being at first very popular. (Corr., letter 150, dated 1 April 1776)

In fact, the book sold well, with subsequent editions in 1778, 1784, 1786 and 1789.

The year 1776 was marred for Smith by the death of David Hume, after a long illness, and by his concern over the future of the latter's *Dialogues Concerning Natural Religion*. This work, together with Hume's account of ‘My Own Life’ had been left in the care of William Strahan, to whom Smith wrote expressing the hope that the *Dialogue* should remain unpublished, although Hume himself had

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
determined otherwise.
But Smith proposed to ‘add to his life a very well authenticated account’ of Hume's formidable courage during his last illness (Corr., letter 172, dated 5 September 1776). The letter was published in 1777, and as Smith wrote later to Andreas Holt, ‘brought upon me ten times abuse than the very violent attack I had made upon the whole commercial system of Great Britain’ (Corr., letter 208, dated October 1780).
In 1778 Smith was appointed Commissioner of Customs, due in part to the efforts of the Duke of Buccleuch. The office brought an income of £600, in addition to the pension of £300 which the Duke refused to discontinue (Corr., letter 208). Smith settled in Edinburgh, where he was joined by his mother and a cousin, Janet Douglas.
During 1778 Alexander Wedderburn sought Smith's advice on the future conduct of affairs in America. Smith's ‘Thoughts on the State of the Contest with America’ were written in the aftermath of the battle of Saratoga. The Memorandum was first published by G.H. Guttridge in the *American Historical Review* (vol. 38, 1932/3).
In this document, Smith rehearsed a number of arguments which he had already stated in WN (IV.vii.c). He advocated the extension of British taxes to Ireland and to America, provided that representatives from both countries were admitted to Parliament at Westminster in conformity with accepted constitutional practice. Smith noted that ‘Without a union with Great Britain, the inhabitants of Ireland are not likely for many ages to consider themselves as one people’ (WN,V.iii.89). With respect to America, he observed that her progress had been so rapid that ‘in the course of little more than a century, perhaps, the produce of American might exceed that of British taxation. The seat of the empire would then naturally remove itself to that part of the empire which contributed most to the general defence and support of the whole’ (WN, IV.vii.c.79).
But Smith also repeated a point already made in WN; namely, that the opportunity for union had been lost, and proceeded to review the bleak options, now all too familiar, which were actually open to the British government. Military victory was increasingly unlikely (WN, V.i.s.27) and military government, even in the event of victory, unworkable (Corr., letter 383). Voluntary withdrawal from the conflict was a rational but politically impracticable course, given the probable impact on domestic and world opinion (Corr., letter 383). The most likely outcome, in Smith's view, was the loss of the thirteen united colonies and the successful retention of Canada – the worst possible solution since it was also the most expensive in terms of defence (Corr., letter 385).
Smith worked hard as a Commissioner, and to an extent which, as he admitted, affected his literary pursuits (Corr., letter 208). But in this period he completed the third edition of WN (1784), incorporating major developments which were separately published as ‘Additions and Corrections’. The third edition also features an index and a long concluding chapter to Book IV entitled ‘Conclusion of the Mercantile System’.
After 1784 Smith must have devoted most of his attention to the revision of TMS. The sixth edition of 1790 features an entirely new Part VI which includes a further elaboration of the role of conscience, and the most complete statement which Smith offered as to the complex *social* psychology which lies behind man's broadly economic aspirations.
In addition to the essay on the ‘Imitative Arts’, which is mentioned in his letter to Andreas Holt (Corr., letter 208), Smith observed that ‘I have likewise two other great works upon the anvil; the one is a sort of Philosophical History of all the different branches of Literature, of Philosophy, Poetry and Eloquence; the other is a sort of theory and History of Law and Government’ (Corr., letter 248, dated 1 November

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1785, addressed to the duc de la Rochefoucauld).
Smith's literary ambitions also feature in the Advertisement to the 1790 editions of TMS, where he drew attention to the concluding sentences of the first edition of 1759. In these passages Smith makes it clear that TMS and WN are parts of a single plan which he hoped to complete with a published account of ‘the general principles of law and government, and of the different revolutions which they had undergone in the different ages and periods of society’. Smith's ‘present occupations’ and ‘very advanced age’ prevented him from completing this great work, although the approach is illustrated by LJA and LJB, and by those passages in WN which can now be recognized as being derived from them (most notably WN, III and V.i.a.b).
Smith died on 17 July 1790, having first instructed his executors, Joseph Hutton and James Block, to burn his papers, excepting those which were published in *Essays on Philosophical Subjects* (1795).
In what follows, Smith's system will be expounded in terms of the order of argument which he is known to have employed as a lecturer; namely, ethics, jurisprudence and economics. But it will be convenient to begin with his treatment and knowledge of the literature of science.

The literature of science

It should be recalled that each separate component of Smith's system represents scientific work in the style of Newton, contributing to a greater whole which was conceived in the same image. Smith's scientific aspirations were real, as was his consciousness of the methodological tensions which may arise in the course of such work.
Smith's interest in mathematics dates from his time as a student in Glasgow (Stewart, I. 7). He also appears to have maintained a general interest in the natural and biological sciences, facts which are attested by his purchases for the University Library (Scott, 1937, p. 182) and for his own collection (Mizuta, 2000). Smith's ‘Letter to the Authors of the *Edinburgh Review*’ (1756), where he warned against any undue preoccupation with Scottish literature, affords evidence of wide reading in the physical sciences, and also contains references to contemporary work in the French *Encyclopédie* as well as to the productions of Buffon, Daubenton and Reaumur. D.D. Raphael has argued that the Letter owes much to Hume (TMS, pp. 10, 11).
The essay on astronomy, which dates from the same period (it is known to have been written before 1758 and may well date from the Oxford period) indicates that Smith was familiar with classical as well as with more modern sources, such as Galileo, Kepler and Tycho Brahe, a salutary reminder that an 18th-century philosopher could work close to the frontiers of knowledge in a number of fields.
But Smith was also interested in science as a form of communication, arguing in the LRBL that the way in which this type of discourse is organized should reflect its purpose as well as a judgement as to the psychological characteristics of the audience to be addressed.
In a lecture delivered on 24 January 1763 Smith noted that didactic or scientific writing could have one of two aims: either to ‘lay down a proposition and prove this, by the different arguments that lead to that conclusion’ or to deliver a system in any science. In the latter case Smith advocated what he called the Newtonian method, whereby we ‘lay down certain principles known or proved in the beginning, from whence we account for the several phenomena, connecting all together by the same Chain’ (LRBL, ii. 133). Two points are to be noted. First, Smith makes it clear that Descartes rather than Newton was the first to use this method of exposition, even although the former was now perceived to be the author of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
‘one of the most entertaining Romances that have ever been wrote’ (LRBL, ii. 134; see Letter 5). Secondly, his reference to the pleasure to be derived from the ‘Newtonian method’ (LRBL, ii. 134) draws attention to the problem of scientific motivation, a theme which was to be developed in the ‘Astronomy’, where Smith considered those principles ‘which lead and direct philosophical enquiry’. The ‘Astronomy’ takes as given certain results which had already been established in the lectures on language and in the *Considerations*; namely, that men have a capacity for acts of ‘arrangement or classing, or comparison, and of abstraction’ (LRBL, ii. 207; cf. Corr., letter 69, dated 7 February 1763). But the essay on astronomy approaches the matter in hand in a different way by arguing that a mind thus equipped derives a certain pleasure from the contemplation of relation, similarity or order – or as Hume would have put it, from a certain association of ideas. Smith struck a more original note in arguing that when the mind confronts a new phenomenon which does not fit into an already established classification, or where we confront an unexpected association of ideas, we feel the sentiment of surprise, and then that of wonder (Astronomy, II. 9). This is typically followed by an attempt at explanation with a view to returning the ‘imagination’ to a state of tranquillity (Astronomy, II. 6). Looked at in this way, the task of explanation is related to a perceived need, which can only be met if the account offered is coherent and conducted in terms which are capable of account for observed appearances in terms of ‘familiar’ principles. It was Smith's contention that the philosopher or scientist would react in the same way as the casual observer, and that nature as a whole ‘seems to abound with events which appear solitary and incoherent’, thus disturbing ‘the easy movement of the imagination (Astronomy, II. 12). But he also observed that philosophers pursue scientific study ‘for its own sake, as an original pleasure or good in itself’ (Astronomy, III. 3).

The bulk of the essay is concerned to illustrate the extent to which the four great systems of thought which he identified were actually able to ‘soothe the imagination’, these being the systems of Concentric and Eccentric Spheres, together with the theories of Copernicus and Newton. But Smith added a further dimension to the argument by seeking to expose the dynamics of the process, arguing that each thought-system was subject to a process of modification as new observations were made. Smith suggested that each system was subjected to a process of development which eventually resulted in unacceptable degrees of complexity, thus paving the way for the generation of an alternative explanation of the same phenomena, but one which was better suited to meet the needs of the imagination by offering a simpler account (Astronomy, IV. 18, 28). In Smith's eyes, the work of Sir Isaac Newton thus marked the apparent culmination of a long historical process (Astronomy, IV. 76).

The argument as a whole also contains some radical conclusions. There is nothing in the analysis which suggests that the Newtonian (or Smithian) system embodies some final truth. At the same time, Smith seems to have given emphasis to what is now known as the problem of ‘subjectivity’ in science in arguing that scientific thought often represents a reaction to a perceived psychological need. He also likened the pleasure to be derived from great productions of the scientific intellect to that acquired when listening to a ‘well composed concerto of instrumental music’ (Imitative Arts, II. 30). Elsewhere he referred to a propensity, natural to all men, ‘to account for all appearances from as few principles as possible’ (TMS, VII.ii.2.14) and commented further on the ease with which the ‘learned given up the evidence of their senses to preserve the coherence of the ideas of their imagination’ (Astronomy, IV. 35). Smith also emphasized the role of the prejudices of sense and education in discussing the reception of new ideas (Astronomy, IV. 35).

He drew attention to the importance of analogy in suggesting that philosophers often attempt to explain

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the unusual by reference to knowledge gained in unrelated fields, noting that in some cases the analogy chosen could become not just a source of ‘ingenious similitude’ but the great hinge upon which everything turned’ (Astronomy, II. 12).

Smith made extensive use of mechanistic analogies, sometimes derived from Newton, seeing in the universe ‘a great machine’ wherein we may observe ‘means adjusted with the nicest artifice to the ends which they are intended to produce’ (TMS, II.ii.3.5). In the same way he noted that ‘Human society, when we contemplate it in a certain abstract and philosophical light, appears like a great, an immense machine’ (TMS, VII.ii.1.2), a position which leads quite naturally to a distinction between efficient and final causes (TMS, II.ii.3.5), which is not inconsistent with the form of Deism associated with Newton himself. It is also striking that so sympathetic a thinker as Smith should have extended the mechanistic analogy to systems of thought.

> Systems in many respects resemble machines. A machine is a little system created to perform, as well as to connect together, in reality, those different movements and effects which the artist has occasion for. A system is an imaginary machine invented to connect together in the fancy those different movements and effects which are already in reality performed. (Astronomy, IV. 19)

Each part of Smith's contribution is in effect an ‘imaginary’ machine which conforms closely to his own stated rules for the organization of scientific discourse. All disclose Smith's perception of the ‘beauty of a systematical arrangement of different observations connected by a few common principles’ (WN, V.i. f.25). The whole reveals much as to Smith's drives as a thinker, and throws an important light on his own marked (subjective) preference for system, coherence and order.

The Theory of Moral Sentiments

The *Theory of Moral Sentiments* shows clear evidence of a model, and of a form of argument which is in part designed to explain how so self-regarding a creature as man succeeds in erecting barriers against his own passions.

In Part VII of TMS, Smith reviewed different approaches to the questions confronting the philosopher in this field, basically as a means of differentiating his own contribution from them.

In Smith's view there were two main questions to be answered: ‘First, wherein does virtue consist’, and secondly, ‘by what means does it come to pass, that the mind prefers one tenour of conduct to another’? (TMS, VII.i.2). In dealing with the first question, Smith described all classical and modern theories in terms of the emphasis given to the qualities of propriety, prudence and benevolence. In each case, he argued that the identification of a particular quality was appropriate, but rejected what he took to be undue emphasis on any one. He criticized those who found virtue in propriety, on the ground that this approach emphasized the importance of self-command at the expense of ‘softer’ virtues, such as sensibility. He rejected others who found virtue in prudence because of the emphasis given to qualities which are useful, thus echoing his criticism of David Hume in TMS, Part IV. In a similar way, while he admired benevolence, Smith argued that proponents of this approach (notably Francis Hutcheson) had neglected virtues such as prudence.

Smith's criticism of Hutcheson's teaching is remarkable for the emphasis which he gave to self-interest

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and his denial of Hutcheson's proposition that self-love ‘was a principle which could never be virtuous in any degree or in any direction’ (TMS, VII.ii.3.12). Smith also rejected the argument of Mandeville, whose fallacy it was ‘to represent every passion as wholly vicious, which is so in any degree’ (TMS, VII. ii.4.12). Smith contended that ‘The conditions of human nature were peculiarly hard, if these affections, which, by the very nature of our being, ought frequently to influence our conduct, could upon no occasion appear virtuous, or deserve esteem and commendation from anybody’ (TMS, VII.ii.3.18).

A further distinctive element in Smith's approach emerges in his treatment of the second question. He accepted Hutcheson's argument that the perception of right and wrong rests not upon reason but ‘immediate sense and feeling’ (TMS, VII.iii.2.9). But Smith rejected Hutcheson's emphasis on a special sense, the moral sense, which was treated as being analogous to ‘external’ senses, such as sight or touch. But in so doing Smith in effect elaborated on the argument of his teacher, who had already presented moral judgements as being disinterested and as based upon sympathy or fellow-feeling. Smith also enlarged on the role of the *spectator*, which had been a feature of the work done by Hutcheson and Hume.

Smith argued that the spectator may form a judgement with respect to the activities of another person by visualizing how he would have behaved or felt in similar circumstances. It is this capacity for acts of imaginative sympathy which permits the spectator to form a judgement as to the propriety or impropriety of the conduct observed, and as to the ‘suitableness or unsuitableness, the proportion or disproportion which the affection seems to bear to the cause or object which excites it’ (TMS, I.i.3.6). Since we can ‘enter into’ the feelings of another person only to a limited degree, Smith was able to identify the ‘amiable’ virtue of sensibility with the quality of imagination, and that of self-command with a capacity to control expressions or feeling to such an extent as to permit the spectator to comprehend, and thus to ‘sympathize’, with them.

The argument was extended to take account of those actions which have consequences for other people, in suggesting that in such cases the spectator may seek to form a judgement as to the propriety of the *action* taken and of the *reaction* to it. The sense of *merit* is a compounded sentiment, made up of two distinct emotions; a direct sympathy with the sentiments of the agent, and an indirect sympathy with the gratitude of those who receive the benefit of his actions’ (TMS, II.i.5.2). Conversely, a sense of *demerit* is compounded of ‘antipathy to the affections and motives of the agent’ and ‘an indirect sympathy with the resentment of the sufferer’ (TMS, II.i.5.4).

Smith further contended that ‘Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren’ (TMS, III.2.6).

But this general disposition is not of itself sufficient to ensure an adequate degree of control. The first problem which Smith confronted is that of *information*, a problem which arises from the fact that the actual spectator of the conduct of another person is unlikely to be familiar with his *motives*.

Smith solved this problem by arguing that we tend to judge our own conduct by trying to visualize the reaction of an imagined or ‘ideal spectator’ to it, that is, by seeking to visualize the reaction of a spectator, who is necessarily fully informed, with regard to our own motives. Smith gave more and more attention to the role of the ideal spectator in successive editions as an important source of control; that is, to the voice of ‘reason, principle, conscience… the great judge and arbiter of our conduct’ (TMS, III.3.4). Looked at in this way, the argument depends on man's desire not merely for praise, but praiseworthiness (TMS, III.2.32).

The second problem arises from the fact that Smith, following Hume, presents man as an active, self-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
regarding being, whose legitimate pursuit of the objects of ambition, notably wealth, can on some occasions have hurtful consequences for others. The difficulty here is that of *partiality* of view, even where we have the information which is needed to arrive at accurate judgements. When we are about to act, ‘the eagerness of passion will seldom allow us to consider what we are doing with the candour of an indifferent person’, while after we have acted, we often ‘turn away our view from those circumstances which might render … judgement unfavourable’ (TMS, III.4.3–4). The solution to this particular problem is found in man's capacity for generalization on the basis of particular experience:

> It is thus that the general rules of morality are formed. They are ultimately founded upon experience of what, in particular instances, our moral faculties, our natural sense of merit and propriety, approve, or disapprove of. … The general rule … is formed, by finding from experience, that all actions of a certain kind, or circumstanced in a certain manner, are approved or disapproved of. (TMS, III.4.8)

It is these rules that provide the yardstick against which man can judge his actions in all circumstances; rules which command respect by virtue of the desire to be praiseworthy and which are further supported by the fear of God (TMS, III.5.12).

Smith thus offered an explanation of the way in which men were fitted for society, arguing in effect that they typically erect a series of barriers to the exercise of their own (self-regarding) passions, which culminate in the emergence of generally accepted rules of behaviour.

The rules themselves vary in character. Those which relate to justice ‘may be compared to the rules of grammar, the rules of the other virtues, to the rules which critics lay down for the attainment of what is sublime and elegant in composition. The one, are precise, accurate and indispensable. The other, are loose, vague, and indeterminate’ (TMS, III.6.11).

But Smith was in no doubt that the rules of justice were indispensable. Justice ‘is the main pillar that upholds the whole edifice’ (TMS, II.ii.3.4). Smith added that the final precondition of social order was a system of positive law, embodying current conceptions of the rules of justice and administered by some system of magistracy:

> As the violation of justice is what men will never admit to from one another, the public magistrate is under the necessity of employing the power of the commonwealth to enforce the practice of this virtue. Without his precaution, civil society would become a scene of bloodshed and disorder, every man revenging himself at his own hand whenever he fancied he was injured. (TMS, VII.iv.36)

Smith's ethical argument forms an integral part of his treatment of jurisprudence precisely because it is concerned to show how particular rules of behaviour emerge. In LJ the focus is narrower than in TMS, but it is still the spectator that is of critical importance whether Smith is discussing accepted standards of punishment or of law. Attention has also been drawn to the role of the magistrate in this connection (Bagolini, 1975) and of the Legislator (Haakonssen, 1981).

Smith's emphasis in TMS is interesting. He chose to concentrate on the means by which the mind forms judgements as to what is fit and proper to be done or to be avoided as distinct from trying to formulate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
specific rules of behaviour. He had recognized that while the processes of judgement might claim universal validity, specific judgements must be related to experience.

No one living in the age of Montesquieu could fail to be aware of variations in standards of accepted behaviour in different societies at the same point in time, and in the same societies over time. The point at issue seems to have been grasped by Edmund Burke in writing to Smith: ‘A theory like yours founded on the Nature of man, which is always the same, will last, when those that are founded upon his opinions, which are always changing, will and must be forgotten’ (Corr., letter 38, dated 10 September 1759).

But Smith did not deny that common elements could be found on the basis of experience. Although he did not complete his intended account of ‘general principles’ (TMS, VII.iv.37), Smith did provide an argument which related the discussion of private and public jurisprudence to four broad types of socio-economic *environment*, the stages of hunting, pasture, farming and commerce. The importance of the argument in the present context is that it was designed in part to explain the origin of government, thus solving a problem which was only noted in TMS. At the same time the historical dimension throws light on the causes of change in accepted rules of behaviour. As part of the same exercise, Smith supplied a successful account of the emergence of the state of commerce, the stage with which he, as an economist, was primarily concerned.

Emergence of the exchange economy

As we have seen in the last section, Smith's analysis of general rules of behaviour suggests that such rules are the result of man's capacity to form judgements as to what is fit and proper to be done or to be avoided. One implication of this argument is that men, at all times and places, form judgements by using the same mental processes. On the other hand it is clear that judgements formed on particular occasions will be related to experience and to the environment which happens to prevail. This in turn, means that accepted patterns of behaviour may vary between different societies at any point in time, but also that they may vary within a particular society over time. There is a comparative aspect, but also a concern with *change*. The point was caught by Dugald Stewart, Professor of Moral Philosophy in Edinburgh, and an acute commentator on Smith, when he noted that:

> When in such a period of society as that in which we live, we compare our intellectual acquirements, our opinions, manners and institutions with those which prevail among rude tribes, it cannot fail to occur to us as an interesting question, by what gradual steps the transition has been made from the first simple efforts of uncultivated nature, to a state of things so wonderfully artificial and complicated. (Stewart, II. 45)

Stewart appreciated that the problem of change applied to the sciences and the arts but also to the treatment of the ‘astonishing fabric of … political union’ – our main concern at this point.

While Smith's interest in the history of civil society is illustrated very clearly in WN, there can be little doubt that his reputation was enhanced by the discovery of LJ. But at the same time, we should recall that interest in this area of study was widespread notably in Italy and France. The point is neatly caught by Voltaire, whom Smith greatly admired, when he observed that:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> My principal object is to know, as far as I can, the manners of peoples, and to study the human mind. I shall regard the order of succession of kings and chronology as my guides, but not as the objects of my work. (Quoted in Black, 1926, p. 4)

In Scotland, the theme was pursued by David Hume in his *History of England*, but also by Adam Ferguson (*History of Civil Society*, 1767), John Millar, William Robertson and Lord Kames, to name but a few (see Berry, 1997, chs 5 and 6).

At the same time there was a growing interest among both Scottish and French writers in the link between economic organizations (modes of subsistence) and patterns of behaviour in the fields of sociology and politics. The link that was established between the form of economic and the social and political structure was so explicit as to permit William Robertson, Histographer Royal and Principal of Edinburgh University, to state the main propositions with economy and accuracy. First, Robertson noted that:

> In every enquiry concerning the operations of men when united together in society, the first object of attention should be their mode of subsistence. According as that varies, their laws and policy must be different.

Secondly, Robertson drew attention to the relationship between property and power, noting for example that ‘Upon discovering in what state property was at any particular period, we may determine with precision what was the degree of power possessed by the king or the nobility at that juncture’, that is, by the government (Skinner, 1996, p. 99).

Smith managed to isolate four distinct modes of subsistence to which there corresponded different types of social and political structures, together with different patterns of ‘manners’, to use Hume's phrase. The thesis was common among Smith's Scottish contemporaries. The different modes of subsistence are represented by the stages of ‘hunting, pasturage, farming and commerce’ (LJB, p. 149). The most detailed treatment of the first two stages will be found in WN, Book V, where Smith considers the historical provision of defence and justice. The third and fourth stages are examined in WN, Book III. Smith's historical sweep was wide ranging, starting as his discourse did from the record of early classical Greece before proceeding to the *Decline and Fall of the Roman Empire* in the West, and thus to the emergence of the modern state. It should be noted that while the perspective adopted in the third Book is European in its emphasis, the focus gradually narrows to the consideration of British and indeed English experience.

Early history

> When the German and Scythian nations over-ran the western provinces of the Roman empire, the confusions which followed so great a revolution lasted for several centuries. The rapine and violence which the barbarians exercised against the antient inhabitants, interrupted the commerce between the towns and the country. The towns were deserted, and the country was left uncultivated, and the western provinces of Europe, which had enjoyed a considerable degree of opulence under the Roman empire, sunk into the lowest

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> state of poverty and barbarism (WN, III.ii.1).

At the same time, however, Smith argued that the domination of the barbarian nations had generated not only a desert but also an environment from which a particular form of European civilization was ultimately to emerge.

Smith's explanation of this general trend begins with the fact that the primitive tribes which overran the empire had already attained a relatively sophisticated form of the pasturage economy, with some idea of agriculture and of property in land. He argued that they would naturally use existing instructions in their new situation and that in particular their first act would be a division of the conquered territories.

> The chiefs and principal leaders of those nations, acquired or usurped to themselves the greater part of the lands … A great part of them was uncultivated; but no part of them, whether cultivated or uncultivated, was left without a proprietor. All of them were engrossed, and the greater part by a few great proprietors. (WN, III.ii.1)

In this way we move in effect from a developed version of one economic stage to a primitive version of another; from the state of pasture to that of ‘agriculture’. Under the circumstances outlined, property in land became the source of power and distinction, with each estate assuming the form of a separate principality. As a result of this situation, Smith argued, a gradual change took place in the laws governing property, featuring the introduction of primogeniture and entails, designed to protect estates against division and to preserve a ‘certain lineal succession’. The basic point emphasized was the ‘The security of a landed estate … the protection which its owner could afford to those who dwelt on it, depended upon its greatness. To divide it was to ruin it, and to expose every part of it to be oppressed and swallowed up by the incursions of its neighbours’ (WN, III.ii.3).

Such institutions as these quite obviously reflect a change in the mode of subsistence and in the form of property, thus presenting some important contrasts with the previous state of pasture. On the other hand, the great proprietor has still nothing on which to expend his surpluses other than the maintenance of dependants – and at the same time has a positive incentive to do so since they contribute to his military power and security. While Smith carefully distinguished between *retainers* and *cultivators* in this context, he took pains to emphasize that the latter group were in every respect as dependent on the proprietor as the first, and added that ‘Even such of them as were not in a state of villanage, were tenants at will, who paid a rent in no respect equivalent to the subsistence which the land afforded them’ (WN, III.iv.6).

In short, the period was marked by clear relations of power and dependence – but above all by disorder and conflict, and it was from this source that the first important changes in the outlines of the system were to come. As Smith put it by way of summary:

> In those disorderly times, every great landlord was a sort of petty prince. His tenants were his subjects. He was their judge, and in some respects their legislator in peace, and their leader in war. He made war according to his own discretion, frequently against his neighbours, and sometimes against his sovereign. (WN, III.ii.3)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
It was this state of conflict, Smith suggests, which gave the proprietors some incentive to alter the pattern of landholding, in two quite different ways. First, Smith argued that the heavy demands which were inevitably made on their immediate tenants (as distinct from villains) for military service would inevitably change the quit-rent system in terms of which land was normally held. Smith argued in effect that the great lords would naturally begin to grant leases for a term of years, and then in a form which gave security to the tenant's family and ultimately to his posterity. In this way, land came to be held in a *feudal* relationship, which was being designed to give both parties a benefit: the lord, in terms of the supply of military service, and the tenant, security in the use of land. Smith also noted certain consequential developments which reflected the basic purpose of the arrangement, in describing what he called the feudal *casualities*.

Secondly, Smith argued that the need for protection which had altered the relationship between the great lord and his tenants would also lead to patterns of alliance between members of the former group and, therefore, to arrangements which gave some guarantee of mutual service and support. It was for these reasons, Smith argued, that the lesser landowners gradually entered into *feudal* arrangements with those greater lords who could ensure their survival (thus enhancing their ability to do so), just as the great lords would be led to make similar arrangements amongst themselves and with the king. These changes took place about the ninth, tenth and eleventh centuries, and by imposing some shackles on the free enterprise of the proprietors contributed thereby to the emergence of a more orderly form of government.

However, while Smith did describe the *feudal* as a higher form of the agrarian economy than the *allodial*, he also took some pains to emphasize the limited possibilities for economic growth which it presented; limitations which were themselves the reflection of the political institutions now prevailing. He argued that the quit-rent system, so far as it survived, gave no incentive to industry, and that the institution of slavery ensured that it was in the interest of the ordinary individual to ‘eat as much, and to labour as little as possible’ (WN, III.ii.9). In the same way he also cited the disincentive effects of the arbitrary services and feudal taxes which were imposed. But, undoubtedly, Smith placed most emphasis on the continuing problem of political instability:

> The authority of government still continued to be, as before, too weak in the head and too strong in the inferior members, and the excessive strength of the inferior members was the cause of the weakness of the head. After the institution of feudal subordination, the king was as incapable of restraining the violence of the great lords as before. They still continued to make war according to their own discretion, almost continually upon one another, and very frequently upon the king; and the open country still continued to be a scene of violence, rapine, and disorder. (WN, III.iv.9)

Once again, a state of instability was to produce some change in the outlines of the social system, and once again the motive behind this change was *political* rather than *economic* – but now with the kings rather than the great lords as the main actors in the drama.

The exchange economy

The kind of economy which Smith described as appropriate to the agrarian state in its developed form is fundamentally a simple one. It consisted of a division between town and country, that is, between those

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
who produce food and those who make the manufactured goods without which no large country could subsist – the critical point being, however, that such an economy was not wholly based on exchange. The cities which Smith described were small, and composed of those merchants, tradesmen and mechanics who were not bound to a particular place and who might find it in their (economic) interest to congregate together. Smith had in fact relatively little to say about the historical origins of such groupings, but he did emphasize that the inhabitants of the towns were in the same servile condition as the inhabitants of the country, and that the wealth which they did manage under such unfavourable conditions would be subject to the arbitrary exactions of both the king and those lords in whose territories they might happen to reside (WN, III.iii.2).

But evidently some developments must have been possible, for Smith examined the role of cities from the period in time when three distinctive features of royal policy with regard to them were already in evidence. First, Smith noted that cities had often been allowed to farm the taxes to which they were subject, the inhabitants thus becoming ‘jointly and severally answerable’ for the whole sum due (WN, III.iii.3). Second, he noted that in some cases these taxes, instead of being farmed for a given number of years, had been ‘let in fee’, that is ‘forever, reserving a rent certain never afterwards to be augmented’ (WN, III.iii.4). Third, Smith observed that the cities:

> were generally at the same time erected into a commonality or corporation, with the privilege of having magistrates and a town-council of their own, of making bye-laws for their own government, of building walls for their own defence, and of reducing all their inhabitants under a sort of military discipline, by obliging them to watch and ward … (WN, III.iii.6)

It was as a result of following these policies that some kings has achieved the apparently remarkable result of freezing the very revenues which were most likely to increase over time, and at the same time effectively curtailing their own power by erecting ‘a sort of independent republicks in the heart of their own dominions’ (WN, III.iii.7).

Smith advanced two main reasons to explain the apparent paradox. First, he argued that by encouraging the cities the king made it possible for a group of his subjects to defend themselves again the power of the great lords, when he personally was frequently unable to do so, and, secondly, that by imposing a limit on taxation ‘he took away from those whom he wished to have for his friends, and, if one may so do, for his allies, all ground of jealousy and suspicion that he was ever afterwards to oppress them, either by raising the farm rent of their town, or by granting it to some other’ (WN, III.iii.8).

The encouragement given to the cities represented in effect a tactical alliance which was beneficial to both parties. In speaking of the burghers, Smith remarked that ‘Mutual interest … disposed them to support the king, and the king to support them against the lords. They were the enemies of his enemies, and it was his interest to render them as secure and independent of those enemies as he could’ (WN, III. iii.8).

Smith also noted that this development was directly related to the weakness of kings, so that it was likely to be more significant in some countries than in others, and that in general the policy had been successful where employed. He also remarked that the granting of powers of self-government to the inhabitants of the cities had set in motion forces which were ultimately to weaken the authority of the kings through creating an environment within which the forces of economic development could, for the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
first time, be effectively released. In Smith's own words:

> Order and good government, and along with them the liberty and security of individuals, were, in this manner, established in cities at a time when the occupiers of land … were exposed to every sort of violence. But men in this defenceless state naturally content themselves with their necessary subsistence; because to acquire more might only tempt the injustice of their oppressors. On the contrary, when they are secure of enjoying the fruits of their industry, they naturally exert it to better their condition, and to acquire not only the necessaries, but the conveniencies and elegancies of life. (WN, III.iii.12)

The stimulus to economic growth and to further social change was thus seen to emanate from the cities; institutions which had themselves been developed and protected in an attempt to solve a political problem. From this point, Smith's attention shifted to the analysis of the process of economic growth in the manufacturing and trading sectors, before going on to examine its impact on the agrarian sector. Smith clearly recognized that growth was limited by the size of the market and, since the agrarian sector was relatively backward, that the main stimulus to economic growth would have to come from foreign trade. He concluded that cities such as Venice, Genoa and Pisa, all of which enjoyed ready access to the sea, had provided the models for the process. In general Smith laid most emphasis on three sources of encouragement to the development of trade and manufactures. First, he argued that in many cases agrarian surpluses could be acquired by the merchants and used in exchange for foreign manufacturers, and suggested as a matter of fact that the early trade of Europe had largely consisted in the exchange ‘of their own rude, for the manufactured products of more civilized nations’. Secondly, he suggested that over time the merchants would naturally seek to introduce similar manufactures at home (with a view to saving carriage). Such manufactures, it was suggested, would require the use of foreign *materials*, thus inducing an important change in the general pattern of trade. Thirdly, he argued that some manufactures would develop ‘naturally’, that is through the gradual refinement of the ‘coarse and rude’ products which were normally produced at home and which were, therefore, based on domestic materials. Smith suggested that such developments were normally found in those cities which were ‘not indeed at a very great, but at a considerable distance from the sea coast, and sometimes even from all water carriage’ (WN, III.iii.20). He suggested that manufacturers might well develop in areas to which artisans had been attracted by the cheapness of subsistence, thus allowing trade to develop within the locality:

> The manufacturers first supply the neighbourhood, and afterwards, as their work improves and refines, more distant markets. For though neither the rude produce, nor even the coarse manufacture, could, without the greatest difficulty, support the expence of a considerable land carriage, the refined and improved manufacture easily may. In a small bulk it frequently contains the price of a great quantity for rude produce. (WN, III.iii.20).

Smith cited the silk manufacture at Lyons and Spitalfields as examples of the first category of manufactures; those of Leeds, Halifax, Sheffield, Birmingham and Wolverhampton as examples of the second, the natural ‘offspring of agriculture’ (WN, III.iii.19, 20). He also added that manufacturers of the latter kind were generally posterior to those ‘which were the offspring of foreign commerce’ and that the process of development just outlined made it perfectly possible for the city within which economic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
development took place to grow up to a great wealth and splendour, while not only the country in its neighbourhood, but all those to which it traded, were in poverty and wretchedness' (WN, III.iii.13).
In the next stage of analysis, however, it was argued that the situation as outlined was unlikely to continue; that the development of manufacturers and trade within the cities was bound to impinge on the agrarian sector and, ultimately, to destroy the service relationships which still subsisted within it. Essentially, this process may be seen to stem from the fact that the development of trade and manufacturers had given the proprietors a means of expending their wealth, other than in the maintenance of dependants. The development of commerce and manufacturers:

> gradually furnished the great proprietors with something for which they could exchange the whole surplus produce of their lands, and which they could consume themselves without sharing it either with tenants or retainers. All for ourselves, and nothing for other people, seems, in every age of the world, to have been the vile maxim of the masters of mankind. (WN, III.iv.10)

This situation generated two results. First, since the proprietor's objective was now to increase his command over the means of exchange, it would be in his interest to reduce the number of retainers:

> till they were at last dismissed altogether. The same cause gradually led them to dismiss the unnecessary part of their tenants. Farms were enlarged, and the occupiers of land, notwithstanding the complaints of depopulation, reduced to the number necessary for cultivating it, according to the imperfect state of cultivation and improvement in those times. (WN, III.iv.13)

Secondly, since the purpose was now to maximize the disposable surplus, it would be in the proprietor's interest to change the forms of leasehold in order to encourage output and increase his returns. In this way, Smith traced the gradual change from the use of slave labour on the land, to the origin of the 'metayer' system where the tenant had limited property rights, until the whole process finally resulted in the appearance of farmers properly so called 'who cultivated the land with their own stock, paying a rent certain to the landlord' (WN, III.ii.14). Smith added that the same process would, over time, tend to lead to an improvement in the conditions of leases, until the tenants could be 'secured in their possession, for such a term of years as might give them time to recover with profit whatever they should lay out in the further improvement of land. The expensive vanity of the landlord made him willing to accept of this condition …' (WN, III.iv.13).
As a result of these two general trends, the great proprietors gradually lost their powers, both judicial and military, until a situation was reached where 'they became as insignificant as any substantial burgher or tradesman in a city. A regular government was established in the country as well as in the city, nobody having sufficient power to disturb its operations in the one, any more than in the other' (WN, III.iv.15).
Smith thus associated the decline in the feudal powers of the great proprietors with three general trends, all of which followed on the introduction of commerce and manufactures: the dissipation of their fortunes, the dismissal of their retainers, and the substitution of a cash relationship for the service relationships which had previously existed between the owner of land and those who cultivated it. He

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
noted elsewhere that ‘the gradual improvements of arts, manufactures, and commerce, the same causes which had destroyed the power of the great barons, destroyed in the same manner, through the greater part of Europe, the whole temporal power of the clergy’ (WN, V.i.g.25).

As a result, an economic system was generated where the disincentives to ‘industry’ had been removed from the agrarian sector, and where both sectors were, for the first time, fully interdependent at the domestic level.

Smith argued in effect that the *quantitative* development of manufactures based on the cities had eventually produced an important *qualitative* change in creating the institutions of the exchange economy, that is, of the fourth economic stage. It is in this situation that the drive to better our condition, allied to the insatiable wants of man (referred to in the *Theory of Moral Sentiments*), provided the maximum possible stimulus to economic growth, and ensured that the gains accruing to town and country were eventually both mutual and reciprocal. As Smith put it:

> The great commerce of every civilized society, is that carried on between the inhabitants of the town and those of the country. It consists in the exchange of rude for manufactured produce, either immediately, or by the intervention of money, or of some sort of paper which represents money … The gains of both are mutual and reciprocal, and the division of labour is in this, as in all other cases, advantageous to all the different persons employed in the various occupations into which it is subdivided. (WN, III.i.1)

Such an economic and social structure effectively eliminated the direct dependence of the previous period, in that each productive service now commands a price. While the farmer, tradesman or merchant must depend upon his customers, yet ‘Though in some measure obliged to them all … he is not absolutely dependent upon any one of them’ (WN, III.iv.12). It will be noted that the whole process of historical change involved in the transition from the feudal to the commercial state depended on the activities of individuals who were unconscious of the ultimate end towards which such activities contributed. Or, as Smith put it in reviewing the actions of the proprietors and merchants during the latter stage of the historical process which we have outlined:

> A revolution of the greatest importance to the public happiness, was in this manner brought about by two different orders of people, who had not the least intention to serve the public. To gratify the most childish vanity was the sole motive of the great proprietors. The merchants and artificers, much less ridiculous, acted merely from a view to their own interest, and in pursuit of their own pedlar principle of turning a penny wherever a penny was to be got. Neither of them had either knowledge or foresight of that great revolution which the folly of the one, and the industry of the other, was gradually bringing about. (WN, III.iv.17)

Finally it should be noted that while Smith regarded the processes of history as inherently complex, he did nonetheless associate these processes with certain economic, social and constitutional trends. The growth of ‘luxury and commerce’ is represented as the inevitable outcome of normal human drives, and associated with the appearance of new sources of wealth. These new forms of wealth allied to the high

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
degree of personal liberty appropriate to the new patterns of dependence also brought with them a new social and political order – a form of ‘constitution’ which was often cited as an explanation for, and in defence of, the English Revolution Settlement. In this way Whig principles could be put on a sound historical basis; a point which is neatly illustrated by John Ramsay of Ochtertyre's comment on Lord Kames's abandonment of his early Jacobite leanings. Ramsay expressed no surprise that Kames should have finally concluded that the Revolution was ‘absolutely necessary’ after ‘studying history and conversing with first rate people’ – no doubt including Smith. Rather similar sentiments were expressed by John Millar when he remarked that ‘When we examine historically the extent of the tory, and of the whig principle, it seems evident, that from the progress of arts and commerce, the former has been continually diminishing, and the latter gaining ground in the same proportion’ (quoted in Skinner, 1996, p. 91). There is little doubt that Smith shared such opinions, or that he rejoiced in a situation where the personal liberty of the subject had been confirmed at the expense of the absolutist pretensions of kings and the power of the old feudal aristocracy.

The latter theme had been elaborated in Hume's essay ‘Of Refinement in the Arts’. Where ‘luxury nourishes commerce and industry’ Hume wrote, ‘the peasants, by a proper cultivation of the land, become rich and independent; while the tradesmen and merchants acquire a share of the property, and draw authority and consideration to that middling rank of men, who are the best and firmest basis of public liberty’ (1985, p. 277). Hume suggested that this development had brought about major constitutional changes, at least in England. The ‘lower house is the support of our popular government; and all the world acknowledges, that it owed its chief influence and consideration to the increase of commerce, which threw such a balance of property into the hands of the commons. How inconsistent then is it to blame so violently a refinement in the arts, and to represent it as the bane of liberty and public spirit!’ (1985, p. 278).

Hume's perception of the interconnection between economic growth and liberty moved Adam Smith to remark of ‘the most illustrious philosopher and historian of the present age’ (WN, V.i.g.3) that: ‘Mr Hume is the only writer who, so far as I know, has hitherto taken notice of it’ (WN, III.iv.4). This interesting but extraordinary statement was never corrected, perhaps as a tribute to Hume's originality.

The model

Smith's account of the origin of the exchange economy suggests that such an economic structure had to be regarded as a model with history. But he also recognized that this particular institutional structure must be associated with a particular set of ‘customs and manners’, to use Hume's phrase. The link here is with the analyses of the TMS and man's desire for approbation. It is a remarkable fact that the judgements offered with regard to the psychology of the ‘economic man’ are to be found in the TMS rather than in the WN.

For Smith, ‘Power and riches appear … then to be, what they are, enormous and operose machines contrived to produce a few trifling conveniences to the body, consisting of springs the most nice and delicate’ (TMS, IV.i.8). But Smith continued to emphasize that the pursuit of wealth is related not only to the desire to acquire the means of purchasing ‘utilities’ but also to the need for status.

> From whence, then, arises that emulation which runs through all the different ranks of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> men, and what are the advantages which we propose by that great purpose of human life which we call bettering our condition? To be observed, to be attended to, to be taken notice of … are all the advantages which we can propose to derive from it. (TMS, I.iii.2.1)

Smith also suggested that in the modern economy, men tend to admire not only those who have the capacity to enjoy the trappings of wealth, but also the qualities which contribute to that end. Smith recognized that the pursuit of wealth and 'place' was a basic human drive which would involve sacrifices which are likely to be supported by the approval of the spectator. The 'habits of economy, industry, discretion, attention and application of thought, are generally supposed to be cultivated from self-interested motives, and at the same time are apprehended to be very praiseworthy qualities, which deserve the esteem and approbation of everybody' (TMS, IV.2.8). Smith developed this theme in a passage which was added to the TMS in 1790:

> In the steadiness of his industry and frugality, in his steadily sacrificing the ease and enjoyment of the present moment for the probable expectation of the still greater ease and enjoyment of a more distant but more lasting period of time the prudent man is always both supported and rewarded by the entire approbation of the impartial spectator. (TMS, VI.1.11)

The most polished accounts of the emergence of the exchange economy and of the psychology of the 'economic man' are to be found, respectively, in the third book of WN and in Part VI of TMS which was added in 1790. Yet both areas of analysis are old and their substance would have been communicated to Smith's students and understood by them to be what they might be seen to be: a preface to the treatment of political economy.

It is a subtle argument taken as a whole. Nicholas Phillipson had argued that Smith's ethical theory 'is redundant outside the context of a commercial society with a complex division of labour' (1983, pp. 179, 182). John Pocock concluded that:

> A crucial step in the emergence of Scottish social theory, is, of course, that elusive phenomenon, the advent of the four stages scheme of history. The progression from hunter to farmer, to merchant offered not only an account of increasing plenty, but a series of stages of increasing division of labour, bringing about in their turn an increasingly complex organisation of both society and personality. (1983, p. 242)

Early economic analysis

Hutcheson and the Lectures

The early analyses of questions relating to political economy are to be found primarily in two documents: the lectures delivered in 1762–63 and the text discovered by Cannan (1896, LJB). Cannan's discovery is the most significant in respect of both date and content. This version is the most complete and provides an invaluable record of Smith's teaching in this branch of his project in the last year of his

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
professorship (1763–64).

The Cannan version yielded two important results.

First, Cannan was able to confirm Smith's debts to Francis Hutcheson. Hutcheson's economic analysis was not presented by him as a separate discourse, but rather woven into the broader fabric of his lectures on jurisprudence. Perhaps it was for this reason that historians of economic thought had rather neglected him. But the situation was transformed as a result of Cannan's work as he first noted that the *order* of Smith's lectures on ‘expediency’ followed that suggested by Hutcheson, albeit, significantly, in the form of a single discourse. The importance of the connection was noted by Cannan (1896, xxv–xxvi; 1904, xxxvi–xli).

Renewed interest in Hutcheson's *economic* analysis revealed that it had its own history. It is evident that he admired the work of his immediate predecessor in the Chair of Moral Philosophy, Gerschom Carmichael (1672–1729), and especially his translation of, and commentary on, the works of Pufendorf. In Hutcheson's address to the ‘students of Universities’, the *Introduction to Moral Philosophy* (1742) is described thus:

> The learned will at once discern how much of this compend is taken from the writings of others, from Cicero and Aristotle, and to name no other moderns, from Pufendorf's smaller work, *De Officio Hominis et Civis Juxta Legem Naturalem* which that worthy and ingenious man the late Professor Gerschom Carmichael of Glasgow, by far the best commentator on that book, has so supplied and corrected that the notes are of much more value than the text. (Taylor, 1965, p. 25)

It is to W.L. Taylor that we are indebted for the reminder that Carmichael and Pufendorf may have shaped Hutcheson's economic *ideas*, thus indirectly influencing Smith. Taylor concluded that:

> The interesting point for the development of economic thought in all this is the very close parallelism between Pufendorf's *De Officio* and Hutcheson's *Introduction to Moral Philosophy*. Each man covered almost exactly the same field … The inescapable conclusion is that Francis Hutcheson to over almost in whole, from Carmichael, the economic ideas of Pufendorf. (1965, pp. 28–9)

Undoubtedly, both men followed a particular *order* of argument. Starting with the division of labour they sought to explain the manner in which disposable surpluses could be maximized, before going on to emphasize the importance of security of property and freedom of choice. This analysis led naturally to the problem of value and hence to the analysis of the role of money. What is distinctive about the analysis is the attention given to value *in exchange* where both writers emphasized the role of utility and disutility: perceived utility attaching to the commodities to be acquired, and perceived disutility embodied in the labour necessary to create the goods to be exchanged. The distinction between utility anticipated and realized is profoundly striking (Skinner, 1996, ch. 5). This tradition was continued by Smith both in LJ and WN, but with a change of emphasis towards the *measurement* of value – thus explaining Terence Hutchison's point that Smith retained *some* of his heritage (1988, p. 199; ch. 11).

Cannan's account revealed that in his lectures, Smith was concerned with a system featuring the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
activities of agriculture, manufacture and commerce (LJB, p. 210) where these activities are characterized by a division of labour (by sector and process) with the process of exchange facilitated by the use of money. The most polished area of analysis, as in the case of the WN, is that of price theory where Smith deployed his distinctions between ‘market’ and ‘natural’ price in a way which illuminated the processes by virtue of which ‘equilibrium’ positions tended to be attained. Examples such as these refer to particular (‘partial’) cases, but Smith may be said to have added further dimension to the argument by showing an understanding of the fact that the economic system can be seen under a more general aspect (Skinner, 1996, pp. 124–6).

This much is evident in his objection to particular regulations of ‘police’ (policy) on the ground that they distorted the use of resources by breaking what he called the ‘natural balance of industry’ while interfering with the ‘natural connexion of all trades in the stock’ (LJB, pp. 233–4). He concluded: ‘Upon the whole, therefore, it is by far the best police to leave things to their natural course’ (LJB, p. 235).

Smith's understanding of the interdependence of economic phenomena was quite as sophisticated as that of his master. Yet at the same time, it must be noted that his lecture notes do not confirm a clear distinction between factors of production (land, labour, capital) nor between those categories of return which correspond to them (rent, wages, profit). Nor is there any evidence of a macroeconomic model of the system as a whole: a model which Smith first met during his visit to Paris.

Paris, 1700. the Physiocrats

There was a great deal in Physiocratic writing that was to prove unattractive to some, most obviously, perhaps, the doctrine of legal despotism and a political philosophy which envisaged a constitutional monarch modelled upon the Emperor of China.

The attitudes of the disciplines to the teaching of the master, Quesnay, were also a source of aggravation, moving Hume to write to Morellet on the subject of his *Dictionnaire du Commerce*.

> I see that, in your prospectus, you care not to disoblige your economists … But I hope that in your work you will thunder them, and crush them, and pound them, and reduce them to dust and ashes! They are, indeed, the set of men the most chimerical and most arrogant that now exist. I wonder what could engaged our friend, M Turgot, to herd among them. (Hume, Corr, ii.205)

Ironically, Turgot himself was as deeply opposed to authority and received doctrine as Hume had been. Hume's reaction also found an echo in France. Murray Rothbard has reminded us of an amusing passage in the works of Simon Nicolas Linguet (1736–1794), ridiculing the idea that the Physiocrats were not ‘a cult or sect’:

> Not a sect? You have a rallying cry, banners, a march, a trumpeter (Dupont), a uniform for your books, and a sign like freemasons. Not a sect? One cannot touch one of you but all rush to his aid. You laud and glorify each other, and attack and intimidate your opponents in unmeasured terms. (Rothbard, 1995, p. 377)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Smith himself objected to the fact that Quesnay's disciples followed ‘implicitly, and without any sensible variation’ the doctrine of the master such as there was ‘upon this account little variety in the greater part of their works’ (WN, IV.ix.38).
But Smith did recognize that the system:

> with all its imperfections, is, perhaps the nearest approximation to the truth that has yet been published upon the subject of political economy, and is upon that account well worth the consideration of every man who wishes to examine with attention the principles of that very important science. (IV.ix.38)

Quesnay's purpose was both practical and theoretical. As R.L. Meek has indicated, Quesnay announced his purpose in a letter to Mirabeau which accompanied the first edition of the *Tableau*.

> I have tried to construct a fundamental *Tableau* of the economic order for the purpose of displaying expenditure and products in a way which is easy to grasp, and for the purpose of forming a clear opinion about the organisation which the government can bring about. (Meek, 1962, p. 108)

The statement is important in that it confirms the importance of government action in the context of a relatively underdeveloped economy which needed urgent support for the agrarian sector, a reform of the mercantile policies associated with Colbert, and in particular changes in the financial sector and in respect of fiscal policy. But Quesnay's statement also announced a clear understanding of the point that governments can act only on the basis of a knowledge of economic laws. Or, as Meek put it, with pardonable exaggeration:

> With the physiocrats, for the first time in the history of economic thought, we find a firm appreciation of the fact that areas of decision open to policy makers in the economic sphere have certain limits, and that a theoretical model of the economy is necessary to define these limits. (1962, p. 370).

The model in question seeks to explore the interrelationships between output, the generation of income, expenditure and consumption – or in Quesnay's words, a ‘general system of expenditure, work, gain and consumption’ (Meek, 1962, p. 374) which would expose the point that ‘the whole magic of a well ordered society is that each man works for others, while believing that he is working for himself’ (Meek, 1962, p. 70). Again, as Meek put it:

> In this circle of economic activity, production and consumption appeared as mutually interdependent variables, whose action and interaction in any economic *period*, proceeding according to certain socially determined laws, laid, the basis for a reputation of the process in the next economic period. (Meek, 1962, p. 19)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The model has a deliberately *abstract* quality but also a number of deficiencies. There is no clear analysis of the division of labour, as Smith understood the term, and no analysis of the problem of price determination and the allocation of resources. There is no formal allowance made for profit *at this point* and nor is there a division between capitalists and wage labour – to name but a few issues of importance. But despite these criticisms, what Smith would have found in the 'Oeconomical Table' was a *model* of the economic process which represents the working of a macroeconomic process as one which involves a series of withdrawals of commodities (consumption and investment goods) from the market, which is matched in turn by a process of continuous replacement, by virtue of production *in the same time period*, all in the *context of a capital-using system*. Smith could hardly fail to be struck by this model, or by the transformation effected by Turgot, who, in effect, made good the bulk of the analytical deficiencies in Quesnay's account (Meek, 1973).

That Smith benefited from his examination of the French system taken as a whole was quickly noted by Cannan. In referring to the theories of distribution and to the macro-economic dimension, Cannan noted that:

> When we find that there is no trace of these theories in the *Lectures*, and that in the meantime Adam Smith had been to France … it is difficult to understand, why we should be asked, without any evidence, to refrain from believing that he came under physiocratic influence after and not before or during his Glasgow period. (1904, p. xxxi)

Economic analysis

A model of conceptualized reality

The concept of an economy involving a flow of goods and services, and the appreciation of the importance of inter-sectoral dependencies, were familiar in the 18th century. Such themes are dominant features of the work done, for example, by Sir James Steuart and David Hume. But what is distinctive about Smith's work, at least as compared to his *Scottish* contemporaries, is the emphasis given to the importance of *three distinct factors* of production (land, labour, capital) and to the three categories of return (rent, wages, profit) which correspond to them. What is distinctive to the modern eye is the way in which Smith deployed these concepts in providing an account of the flow of goods and services between the sectors involved and between the different socio-economic groups (proprietors of land, capitalists, and wage-labour). The approach is also of interest in that Smith, *following the lead of the French economists*, worked in terms of period analysis – the year was typically chosen, so that the working of the economy is examined within a significant time dimension as well as over a series of time periods. Both versions of the argument emphasize the importance of capital, fixed and circulating.

Smith can be seen to have addressed a series of problems which begin with an analysis of the division of labour, before proceeding to the discussion of value, price and allocation, and thence to the issue of distribution in any one time period and over time.

The analysis offered in the first book enabled Smith to proceed, in WN, Book II, to the discussion of both macro-statics and macro-dynamics, in the context of a model where all magnitudes are dated. What Smith had produced was a model of conceptualized reality which was essentially descriptive, and which was further illustrated by reference into an *analytical* system which, if on occasion subject to ambiguity,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
was none the less so organized as to meet the requirements of the Newtonian ideal. The intellectual system was intended to be *comprehensive*.

Value

Although Smith's model, in its post-Physiocratic form, has several distinct elements, the feature on which he continued to place more emphasis was the *division of labour*. In terms of the content of the model outlined in previous chapters, a division of labour is of course implied in the existence of distinct *sectors* or types of productive activity. But Smith also emphasizes the fact that there was specialization by types of employment, and even within each employment. To illustrate the basic point, Smith chose the celebrated example of the pin; a very ‘trifling manufacture’ which none the less required some 18 processes for its completion.
In Smith's hands, the argument was important for two main reasons. First, he was at some pains to point out that the division of labour (by process) helped to explain the relatively high productivity of labour in modern times; a phenomenon which he ascribed to:

1. 1. The increase in ‘dexterity’ which inevitably results from making a single, relatively simple operation ‘the sole employment of the labourer’.
2. 2. The saving of time which would otherwise be lost in ‘passing from one species of work to another’.
3. 3. The associated use of machines which ‘facilitate and abridge labour, and enable one man to do the work of many’ (WN, I.i.5).

He further observed that the existence of specialization (by employment) necessarily involves a high degree of interdependence, in that each separate *manufacture* tends to rely on the output of other industries for different goods and services. It thus follows that the individual customer who purchases a single commodity must at the same time acquire, in effect, the separate outputs of a ‘great variety of labour’. Smith added:

> If we examine … all of these things, and consider what a variety of labour is employed about each of them, we shall be very sensible that without the assistance and co-operation of many thousands, the very meanest person in a civilised country could not be provided, even according to, what we very falsely imagine, the easy and simple manner in which he is commonly accommodated. (WN, I.ii.11)

However, the aspect of this discussion which is most immediately relevant is the light which it throws on the necessity of exchange. As Smith observed, once the division of labour is established, our own labour can supply us with only a very small part of our wants. He thus noted that even in the barter economy the individual can best satisfy the whole range of his needs by exchanging the surplus part of his own production, receiving in return the products of others. Where the division of labour is thoroughly established, it is to be expected that each individual is in a sense dependent on his fellows, and that ‘Every man thus lives by exchanging, or becomes in some measure a merchant’ (I.iv.1).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
This observation brought Smith directly to the problem of value where he returned to an area which, interestingly, had become more a feature of Hutcheson's lectures than of his own. Here it is noteworthy that he employed the analytical (as distinct from the historical) device of the barter economy. However, despite the attempt to be ‘perspicuous’, these passages remain somewhat difficult largely because Smith uses a single term in handling two distinct but related problems.

> The word VALUE, it is to be observed, has two different meanings, and sometimes expresses the utility of some particular object, and sometimes the power of purchasing other goods which the possession of that object conveys. (WN, I.iv.13).

The first problem concerns the forces which determine the *rate* at which one good, or units of one good, may be exchanged for another; the second is concerned basically with the means by which we can measure the value of the total *stock* of goods created by an individual, and which is used in exchange for others. We may take these issues in turn.

As regards the *rate* of exchange, Smith isolated two relevant factors: the usefulness of the good *to be acquired*, and the ‘cost’ incurred in creating the commodity *to be given up*. The first of the relevant relationships is obviously that which exists between ‘usefulness’ and value. The elements of Smith's argument become apparent in his handling of the famous paradox, namely that:

> The things which have the greatest value in use have frequently little or no value in exchange; and, on the contrary, those which have the greatest value in exchange have frequently little or no value in use. Nothing is more useful than water: but it will scarce purchase anything. A diamond, on the contrary, has scarce any value in use; but a great quantity of other goods may frequently be had in exchange for it. (WN, I.iv.13)

The solution to this paradox can be stated in two stages, where the first involves an explanation as to why two such goods have *some* value, and the second an explanation as to why the two goods have *different* values.

Smith's handling of the first part of the problem is based on his recognition of the fact that both goods are considered to be ‘useful’ although noting that the ‘utilities’ of each are qualitatively different. In the former case (water) we place a value on the good because we can use it in a practical way, while in the latter (diamonds) we place a value on the good because it appeals to our ‘senses’, an appeal which, as Smith observed, constitutes a ground ‘of preference’, or ‘source of pleasure’. He concluded:

> The demand for the precious stones arises altogether from their beauty. They are of no use, but ornaments. (WN, I.xi.c.32)

The utilities of the two goods thus emerge as being qualitatively different, although the significant point is seen to be that *both* have *some* value precisely because they represent sources of satisfaction to the individual.

Smith was then left with the second part of the initial problem, namely the explanation as to why the two goods have *different* values. Here again, the answer provided, while simple, is clear, embodying the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
argument that merit (value) is a function of scarcity. As Smith put it: ‘the merit of an object which is in any degree useful or beautiful, is greatly enhanced by its scarcity’ (WN, I.xi.c.31). Even more specifically he remarked:

> Cheapness is in fact the same thing with plenty. It is only on account of the plenty of water that it is so cheap as to be got for the lifting, and on account of the scarcity of diamonds (for their real use seems not yet to be discovered) that they are so dear. (LJB, pp. 105–6)

Smith introduced the second major element in the problem by observing that the *rate* at which the individual will exchange one good for another must be affected not only by the utility of the good to be acquired, but also by the ‘toil and trouble’ involved in creating the good exchanged. In this connection he recognized that in acquiring the means of exchange (goods in the barter case), the individual must undergo the ‘fatigues’ of labour and thus ‘lay down’ a ‘portion of his ease, his liberty, and his happiness’ (WN, I.v.7).

In dealing with the *rate* of exchange, Smith may be seen to have placed most emphasis on the supply side of the problem, and explicitly argued that in the case of the barter economy ‘the proportion between the quantities of labour necessary for acquiring different objects seems to be the only circumstance which can afford any rule for exchanging them for one another’ (WN, I.vi.1). Thus he suggested that if it takes twice the labour to kill a beaver as it does to kill a deer then ‘one beaver should naturally exchange for … two deer’; an argument which may owe something to Hutcheson's emphasis on labour embodied. Smith left the analysis in this form although it will be apparent that the rate of exchange which he specified could only obtain where the perceived ratios of the utilities and disutilities are acceptable to the respective hunters. These are of course subjective judgements whose presence helps to confirm Cannan's opinion that the ‘germ’ of the WN is to be found in Hutcheson's treatment of value.

This is one way of looking at the problem of exchange value, which clearly shows a parallel with Hutcheson, but Smith seems to have treated it, not as an end in itself, but as a means of elucidating those factors which govern the value of *the whole stock of goods* which the individual creates, and which he proposes to use in exchange. It is of course the presence of this argument in the WN which helps to confirm Taylor's judgement to the effect that the treatment of value was dominated by a concern with the *measurement* of welfare. Looking at the problem *in this way*, Smith went on to argue that:

> The value of any commodity … to the person who possesses it, and who means not to use or consume it himself, but to exchange it for other commodities, is equal to the quantity of labour which it enables him to purchase or command. Labour, therefore, is the real measure of the exchangeable value of all commodities. (WN, I.v.1)

Smith's meaning becomes clear when he remarks that the value of a *stock* of goods must always be in proportion to ‘the quantity … of other men's labour, *or what is the same thing, of the produce* of other men's labour, which it enables him to purchase or command. The exchangeable value of every thing must always be precisely equal to the extent of this power’ (WN, I.v.3). In other words, Smith is here arguing that the real value of the goods which the workman has to dispose of (in effect his income) must be measured by the quantity of goods (expressed in terms of labour units) which he can command, and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
which he receives once the whole volume of (separate) exchanges has taken place.
As Smith observed, a clear difference between the barter and modern economies is to be found in the fact that, while in the former, goods are exchanged for goods, in the latter, goods are exchanged for a sum of money, which may then be expended in purchasing other goods. Under such circumstances the individual, as Smith saw, very naturally estimates the value of his receipts (received in return for undergoing the ‘fatigues’ of labour) in terms of money, rather than in terms of the quantity of goods he can acquire by virtue of his expenditure. However, Smith was at some pains to insist that the real measure of welfare (that is, our ability to satisfy our wants) was to be found in ‘the money's worth’ rather than the money, where the former is determined by the quantity of products (labour ‘commanded’) which either individuals or groups can purchase. On this basis, Smith went on to distinguish between the *nominal* and the *real* value of income, pointing out that if the three original sources of (monetary) revenue in modern times are wages, rent and profit, then the real value of each must ultimately be measured ‘by the quantity of labour which they can, each of them, purchase or command’ (WN, I.vi.9).

Price

Smith regarded rent, wages and profit as the types of return payable to the three ‘great constituent orders’ of society, and as the price paid for the use of the factors of production. The revenues which accrue to individuals and groups in society, and which permit them to purchase commodities, thus appear to be costs incurred by those who create commodities. These points were made quite explicitly by Smith when he remarked:

> As the price or exchangeable value of every particular commodity, taken separately, resolves itself into some one or other or all of those three parts; so that of all the commodities which compose the whole annual produce of the labour of every country, taken complexly, must resolve itself into the same three parts, and be parcelled out among different inhabitants of the country, either as the wages of their labour, the profits of their stock, or the rent of their land. (WN, I.vi.17)

This argument obviously raises the problem of price and its determinants.
To begin with, Smith assumed the existence of given ‘ordinary or average’ rates of wages, profit and rent; rates which may be said to prevail within any given society or neighbourhood, during any given time period. This assumption is of considerable importance, for two main reasons. First, it indicates that in dealing with the problem of price, Smith was working in terms of a given (stable) level of aggregate demand for them. Secondly, the assumption of given rates of return is important in that these rates determine the supply price of commodities.
With these two points forming Smith's major premises, he proceeded to examine the determinants of price, and to produce a discussion which seems to involve two distinct, but related, problems. First, Smith set out to explain the forces which determine the prices of *particular* commodities. Secondly, he would appear to have used the above analysis as a means of explaining the phenomenon of *general* interdependence, and thus those forces which determine the manner in which a given stock of factors of production is allocated between different uses or employments.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In dealing with the first aspect of the problem, Smith implicitly examines the case of a single commodity manufactured by a number of sellers, opening the analysis by establishing a distinction between ‘natural’ and ‘market’ price. *Natural price* is now defined as that amount which is ‘neither more nor less than what is sufficient to pay the rent of the land, the wages of labour, and the profits of the stock … according to their natural rates’ (WN, I.vii.4). In other words, where natural price prevails, the seller is just able to cover his costs of production, including a margin for ‘ordinary or average’ profit. By contrast, *market price* is defined as that price which may prevail at any given point in time, being regulated ‘by the proportion between the quantity which is actually brought to market, and the demand of those who are willing to pay the natural price of the commodity’, the ‘effectual demanders’ (WN I. vii.8). These two ‘prices’ are inter-related, the essential point being that while in the short run the market and natural prices may diverge, in the long run they will tend to coincide. If for example, the quantity offered by the sellers was less than that which the consumers were willing to take at a particular (natural) price, the consequence would be a competition among *consumers* to procure some of a limited stock. The price would then rise above the natural price, and the rewards to factors (notably wages and profit in the short run) would diverge from the natural rates, leading to an influx of resources, and an expansion in the supply of the commodity, thus *tending* towards a return to a position of equilibrium. In making the latter point, Smith took note of the fact that in some cases demand could be postponed to another time period, while in others (for example, perishable necessaries) it could not.

In the second case, where supply exceeds demand, market prices will sink and with them rates of factor payment until factors leave the employment and the supply of commodities is thus reduced. Here again the competition between *suppliers* to rid themselves of excess stock will be affected by the nature of the commodity (durable or perishable). It will be noted that Smith makes allowance for interrelated adjustments in commodity and factor prices, that he makes due allowance for competition *between and among* buyers and sellers, while noting the distinction between durable and perishable goods.

Smith also observed that the result attained, namely that commodities in the long run are sold at their cost of production, can only hold good where there is perfect liberty (as distinct from perfect competition). The cost of production solution is, in short, only to be expected where free competition prevails.

The first stage of the discussion established that in the case of any one commodity, equilibrium will tend to be attained where the good is sold at its natural price, and where each of the relevant factors is paid for at its natural rate. Under these circumstances, equilibrium obtains precisely because there can be no tendency for resources to increase or decrease in this particular type of employment.

Now it is evident that if this process, and this result, holds good for all commodities taken separately, it must also apply to all commodities ‘taken complexly’, at least where a competitive situation prevails. That is, where the conditions which form the assumptions of the competitive case are satisfied *over the whole economy*, a position of equilibrium will tend to be attained where each different type of good is sold at its natural price, and where each factor in each employment is paid at its natural rate. The economy can then be said to be in a position of ‘balance’, since where the above conditions are satisfied there can be no tendency to move resources within or between employments. Where the necessary conditions are not satisfied (for example, as a result of changes in tastes) they will naturally tend to be re-established as a result of simultaneous adjustments in the factor *and* commodity markets. It will be observed that departure from, and re-attainment of, a position of equilibrium depends upon the essentially self-interested actions and reactions of consumers and producers. Smith's treatment of price

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and allocation thus provides one of the best examples of his emphasis on ‘interdependence’ and one of the most dramatic applications of his *analogy* of the Invisible Hand.

While Smith certainly conceived of ‘balance’ in terms of a situation where there was no tendency for resources to move between employments, he also recognized that a position of ‘balance’ need not involve an equality between monetary rates of return. The point follows directly from Smith's recognition of the fact that employments differ qualitatively and that such differences may serve to explain why, even in a position of ‘balance’, different money rates prevail. As Smith put it, ‘certain circumstances in the employments themselves … either really, or at least in the imaginations of men, make up for a small pecuniary gain in some, and counter-balance a great one in others’ (WN, I.x.a.2). Thus, for example, he noted that money wage rates would tend to vary between different types of employment according to the difficulty of learning the trade, the constancy of employment, and the degree of trust involved. In the same way he observed that both wages and profit would vary with the agreeableness of the work, the cost of training, and the probability of success in particular fields. In short, he was suggesting that money rates of return would tend to equality within employments of similar types, so that over the whole economy the relevant ‘balance’ would be one involving net advantages.

Before passing to the next stage of the argument it may be useful to make two points both of which refer to the TMS while at the same time bearing upon the present discussions of the allocative mechanism. The treatment of the doctrine of net advantages is connected to the argument advanced in TMS to the effect that men are motivated by the desire to be approved of. In the present context the arguments suggest that where a profession is widely admired, public approbation may become part of the reward. On the other hand, Smith's analysis suggests that men may only be induced to enter and to remain in particular professions if public disapprobation is compensated by an appropriate monetary reward (the trades of the butcher and the inn-keeper are described as ‘odious’).

Distribution

As we have seen, Smith's analysis of price and its determinants was built upon the assumption of given rates of factor payment of the kind that could prevail in a given time period – say one year. Briefly stated, the argument has three features: Smith attempts to explain *why* a particular form of return is paid to a particular factor of production (labour, capital and land); the nature of those forces which determine the rates of factor payment which prevail at a particular point in time, and finally, those forces which explain trends in factor payment over long periods of time.

Wages: Smith observed that payment for the factor labour is paid for by those classes which require the services involved. The process of wage determination in a *given time period* will then depend on the relative bargaining position of two groups (labour and entrepreneurs) in a situation where the legal advantages typically lay with the ‘masters’. Where labour is scarce to compared with the demand for it, wage rates will tend to be relatively high; relatively low in the opposite case. Smith was thus able to argue that wage rates would be relatively high or low depending on the size of the working population and on the size of the capital stock destined for the employment of the factor (the wages fund). Wage rates may also be relatively high or low depending on the current definition of the subsistence wage, where the latter is defined as a level sufficient to sustain a constant level of population (I.viii.15). The argument leads on to the issue of *long-term adjustment*; Smith's position being that where the wage rate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
is below the subsistence level, population must contract, and where sustained above this level over a period of time, population must expand – the more typical case in the circumstances of, for example, Great Britain and the North American colonies.

In the context of the relatively short run it follows that wage rates may be equal to, above or below the subsistence rate and that the rates paid are related to the size of the working population and the size of the wages fund.

Profit: Smith did not consider that this form of return was payable for the work of ‘inspection and direction’ but rather as the reward accruing to those entrepreneurs who risked their capital in combining this factor with others, such as labour and land. The emphasis upon risk is to be noted.

At least as a broad generalization Smith felt able to argue that at *a given point* in time, the rate of profit prevailing would be determined by the capital stock available, taken in conjunction with the volume of business to be transacted by it. But Smith made an important qualification to this statement in arguing that even where the quantity of stock (capital) remains the same, the rate of profit will be related to the prevailing wage rate.

In the long run, however, Smith suggested that the rate of profit would tend to fall, thus establishing a proposition which was to have enduring vitality. Over time, Smith contended that the rate of profit would tend to decline, partly in consequence of an increase in the capital stock, and partly as a result of the increasing difficulty of finding ‘a profitable method of employing any new capital’. He continued:

> When the stocks of many rich merchants are turned into the same trade, their mutual competition naturally tends to lower its profit, and when there is a like increase of stock in all the different trades carried on in the same society, the same competition must produce the same effect in them all. (I.xi.2)

It then follows that the ‘diminution of profit is the natural effect of prosperity’ (I.x.10).

Rent: is formally defined as the ‘price paid for the use of land’ (I.xi.a.1). Looked at this way, Smith made a point which is reminiscent of the French economists, in arguing that rent constitutes a surplus in the sense that it accrues to the owner of land independently of any effort made by him (I.xi.8) and that rent payments are generally the highest that can ‘be got in the actual circumstances of the land’ (I.xi. a.1). The reference to actual circumstances is important since Smith recognized that rent would vary both with the fertility and the situation of the land.

The analysis serves to suggest that at *any point in time*, or during any annual period, rent payments will be related to the stock of land actually in use where the latter is in turn related to the level of population. The argument also indicates that rent payments will be related not only to the fertility of the land and its situation, but also to the prevailing rates of profit and wages – another reminder of the interdependence of the different rates of return (I.xi.a.8). In the long run, Smith suggested that rent payments in the aggregate would tend to increase owing to the increased use of the available stock of land (I.xi.2). He added that the real value of the landlord's receipts would also increase over time since all ‘those improvements in the productive powers of labour, which tend directly to reduce the real price of manufactures, tend indirectly to raise the real rent of land’ (I.xi.4).

The argument just reviewed thus has a short- and long-run dimension. Smith was concerned with long-run trends in rates of return which, as in the short-run case, are interrelated. Thus Smith suggests that profits will decline as the size of the capital stock increases, that high rates of accumulation of capital

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
will generate high market wage rates, leading to an increase in the level of population and in land use. But so far no explanation has been offered as to the source of the crucial increase in capital; a gap which was to be filled in the analyses of Book II.

Macroeconomics (I)

Smith's analysis of the ‘circular flow’ may be seen as a direct development of certain results already stated in connection with the theory of price. To begin with, it will be recalled that costs of production are incurred by those who create commodities, thus providing individuals with the means of exchange. It therefore follows that if the price of each good in a position of equilibrium comprehends payments made for rent, wages and profit, according to their natural rates, then ‘it must be so with regard to all the commodities which compose the whole annual produce of the land and labour of every country, taken complexly’ (WN, II.ii.2). On this basis, Smith concluded that ‘The whole price or exchangeable value of that annual produce, must resolve itself into the same three parts, and be parcelled out among the different inhabitants’ (WN, II.ii.2). If we ignore the problem of distribution (that is of a given level of income between rent, wages and profit), the result which Smith was endeavouring to establish may be stated to involve a relationship between aggregate output and aggregate income. In his own words, ‘The gross revenue of all the inhabitants of a great country, comprehends the whole annual produce of their land and labour’ (WN, II.ii.5).

It will be evident that a particular level of income, created by a particular level of aggregate output, represents that power to purchase goods which is available to all the members of ‘a great society’. Smith then went on to observe that this level of purchasing power would be divided into two funds, consumption and saving. In fact, Smith offered no *formal* explanation of the forces which would determine the actual distribution of aggregate income or purchasing power between these two uses, at any particular point in time. He did, however, suggest that proprietors and labourers would tend to devote a high proportion of their income to consumption, the latter by virtue of the size of their receipts in relation to their basic needs, and the former by virtue of the habits of ‘expence’ associated with that class. The problem of balancing future against present enjoyments thus appeared to be mainly relevant for the entrepreneurial groups; groups whose functions and objectives dispose them to frugality, at least while actively engaged in the pursuit of fortune.

But Smith did clarify the problems here considered from the standpoint of expenditure. For example, he noted that the proportion of annual income for consumption, taking all groups ‘complexly’, would be used to purchase commodities which were either perishable or durable in character. He also noted that this type of expenditure could involve the purchase of services; services of kind which do not directly contribute to the annual output of commodities in physical terms and which thus cannot be said to contribute to the level of income associated with it. Smith formally described such labour as ‘unproductive’, but did not deny that such services were useful. With regard to *savings* Smith identified two sources and two uses. For example, he identified the agrarian, trading, and manufacturing interests as groups wherein ‘the owners themselves employ their own capitals’ (WN, II.iv.5), as distinct from the monied interest who may lend either for the purpose of consumption or of production.

Smith went on to argue that the undertaker or entrepreneur, engaged in agriculture, manufacture or trade could employ their own or borrowed resources for productive purposes, and divided their capitals into two categories both of which are reminiscent of Physiocratic teaching. *Fixed capital* was defined as that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
portion of savings used to purchase ‘useful machines’ or to improve, for example, the productive powers of land, the characteristic feature being that goods are created, and profits ultimately acquired, by using and retaining possession of the investment goods involved. *Circulating capital* was defined as that portion of savings used to purchase investment goods other than ‘fixed implements’, such as labour power or raw materials, the characteristic feature being that goods are produced through temporarily ‘parting with’ the funds so used. Smith made three points in the context of this discussion:

1. 1. ‘Every fixed capital is both originally derived from, and requires to be continually supported by, a circulating capital’ (WN, II.i.24);
2. 2. ‘No fixed capital can yield any revenue but by means of a circulating capital (WN, II.i.25), while in addition
3. 3. ‘different occupations require very different proportions between the fixed and circulating capitals employed in them’ (WN, II.i.6).

Macroeconomics (II)

While these points are important of themselves, they were to gain further significance when Smith moved to the next stage of his argument: the development of his version of the ‘circular flow’ where, again following the Physiocratic lead, he examined the functioning of the system in a given time period (such as a year). Taking the economic system as a whole, Smith suggested that the total stock of society could be divided into three categories:
There is, first, that part of the total stock which is reserved for immediate consumption, and which is held by *all* consumers (capitalists, labour and proprietors) reflecting purchases made in previous time periods. The characteristic feature of this part of the total stock is that it affords no revenue to its possessors since it consists in ‘the stock of foods, clothes, household furniture, etc. which have been purchased by their proper consumers, but which are not yet entirely consumed’ (WN, II.i.12).
Secondly, there is that part of the total stock which may be described as ‘fixed capital’ and which will be distributed between the various groups in society. This part of the stock, Smith suggested, is composed of the ‘useful machines’ purchased in preceding periods but currently held by the undertakers engaged in manufacture; the quantity of useful buildings and of ‘improved land’ in the possession of the ‘capitalist’ farmers and the proprietors, together with the ‘acquired and useful abilities’ of all the inhabitants (WN, II.i.13–17); that is, human capital.
Thirdly, there is that part of the total stock which may be described as ‘circulating capital’ and which again has several components, these being:

1. 1. The quantity of money necessary to carry on the process of circulation. In this connection Smith observed that ‘The sole use of money is to circulate consumable goods. By means of it, provisions, materials, and finished work, are bought and sold, and distributed to their proper consumers. The quantity of money, therefore, which can be annually employed in any country must be determined by the value of the consumable goods annually circulated within it’ (WN, II. iii.23).
2. 2. The stock of provisions and other agricultural products which are available for sale during the current period, but which are still in the hands of either the farmers or merchants.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

3. 3. The stock of raw materials and work in progress, which is held by merchants, undertakers, or those capitalists engaged in the agricultural sector (including mining, and so on).
4. 4. The stock of manufactured goods (consumption and investment goods) created during the previous period, but which remain in the hands of undertakers and merchants at the beginning of the period examined (WN, II.i.19–22).

The logic of the process can be best represented by artificially separating the activities involved much in the manner of the Physiocratic model with which Smith was familiar. Let us suppose that at the beginning of the time period in question, the major capitalist groups possess the total net receipts earned from the sale of products in the previous period, and that the undertakers engaged in agriculture open by transmitting the total rent due to the proprietors of land, for the current use of that factor. The income thus provided will enable the proprietors to make the necessary purchases of consumption (and investment) goods in the current period, thus contributing to reduce the stocks of such goods with which the undertakers and merchants began the period. Secondly, let us assume that the undertakers engaged in both sectors, together with the merchant groups, transmit to wage labour the content of the wages fund, thus providing this socio-economic class with an income which can be used in the current period. It is worth noting in this connection that the capitalist groups transmit a fund to wage labour which formed a part of their *savings*, providing by this means an income which is available for current *consumption*. Thirdly, the undertakers engaged in agriculture and manufactures will make purchases of consumption and investment goods from each other, through the medium of retail and wholesale merchants, thus generating a series of expenditures linking the two major sectors. Finally, the process of circulation may be seen to be completed by the purchases made by individual undertakers within their own sectors. Once again these purchases will include consumption and investment goods, thus contributing still further to reduce the stocks of commodities which were available for sale when the period under examination began, and which formed part of the circulating capital of the society in question.
Given these points, we can represent the working of the system in terms of a series of flows whereby money income, accruing in the form of rent, wages and profit, is exchanged for commodities in such a way as to involve a series of withdrawals from the ‘circulating’ capital of society. As Smith pointed out, the consumption goods withdrawn from the existing stock may be entirely used up within the current period, used to *increase* the stock ‘reserved for immediate consumption’, or to *replace* the more durable goods, for example, furniture or clothes, which had reached the end of their lives in the course of the same period. Similarly, the undertakers, as a result of their purchases, may *add* to their stocks of raw materials and/or their fixed capital, or *replace* the machines which had finally worn out in the current period, together with the materials used up as a result of current productive activity. Looked at in this way, the ‘circular flow’ could be seen to involve purchases which take goods *from* the circulating capital of *society*, which is in turn matched by a continuous process of *replacement* by virtue of current production of materials and finished goods – where both types of production require the use of the fixed and circulating capitals of *individual* entrepreneurs. It is an essential part of Smith's argument that all available resources will normally be used:

> In all countries where there is tolerable security, every man of common understanding will endeavour to employ whatever stock he can command in procuring either present enjoyment or future profit. If it is employed in procuring present enjoyment, it is a stock

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> reserved for immediate consumption. If it is employed in procuring future profit, it must procure this profit either by staying with him, or by going from him. In the one case it is a fixed, in the other it is a circulating capital. A man must be perfectly crazy who, where there is tolerable security, does not employ all the stock which he commands, whether it be his own or borrowed of other people, in some one or other of those three ways. (WN, II. i.30)

Smith elaborated on this argument in drawing attention to the point that the differing ways in which the *entrepreneurial* classes employ their capitals were interdependent. The point is reminiscent of Turgot:

> A capital may be employed in four different ways: either, first, in procuring the rude produce annually required for the use and consumption of the society; or, secondly, in manufacturing and preparing that rude produce for immediate use and consumption; or, thirdly, in transporting either the rude produce from the places where they abound to those where they are wanted; or, lastly, in dividing particular portions of either into such small parcels as suit the occasional demands of those who want them. (WN, II.v.1).

Macroeconomics (III). the sources of growth

In choosing to examine the working of the economy during a given time period such as a year, Smith gave his model a broadly short-run character although it is obviously one which included a time dimension. At the same time Smith did not seek to formulate *equilibrium* conditions (as Quesnay had done) for the model, at least in the sense that he did not try to develop an argument which used specified assumptions of a quantitative kind as a means of showing the conditions which must be satisfied before the following time period could open under conditions identical to those prevailing in the period actually examined.

Nor in dealing with the ‘flow’ did Smith suggest that the level of output attained during any given period would be exactly sufficient to replace the goods used up during its course. On the contrary, he argued that output levels attained in any year would be likely to exceed previous levels: an important reminder that Smith's predominant concern was with economic growth. In this connection, Smith noted that the ‘annual produce of the land and labour of any nation can be increased in its value by no other means, but by increasing either the number of its productive labourers, or the productive power of those labourers who had before been employed’ (WN, II.iii.32). Smith also observed that both the above sources of increased output required an ‘additional capital’ devoted either to increasing the size of the wages fund or to the purchase of ‘machines and instruments which facilitate and abridge labour’; an additional capital which can only be acquired through net savings.

> By what a frugal man annually saves, he not only affords maintenance to an additional number of productive hands, for that or the ensuing year, but like the founder of a public workhouse, he establishes as it were a perpetual fund for the maintenance of an equal number in all times to come. (WN, II.iii.19)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
It will be observed that net savings attained during the course of a single annual period will lead to higher output and income, where the latter becomes available during the course of the period examined. The argument can be extended from this point, in that higher levels of output and income attained in any one year make it possible to reach still higher levels of savings and investment in subsequent years, thus generating further increases in output and income. Once started, the process of capital accumulation and thus economic growth may be seen as self-generating, indicating that Smith's flow is to be regarded as spiral rather than as a circle of given dimensions. This indeed is the burden of Smith's argument in Book II; a fact which helps to explain some of its recurrent themes.
First, Smith frequently argued that net savings will always be possible during each annual period:

> Whatever a person saves from his revenue he adds to his capital, and either employs it himself in maintaining an additional number of productive hands, or enables some other person to do so, by lending it to him for an interest, that is, for a share of the profits. As the capital of an individual can be increased only by what he saves from his annual revenue or his annual gains, so the capital of a society, which is the same with that of all the individuals who compose it, can be increased only in the same manner. (WN, II.ii.15)

Secondly, Smith emphasized that:

> Parsimony, by increasing the fund which is destined for the maintenance of productive hands, tends to increase the number of those whose labour added to the value of the subject upon which it is bestowed. It tends therefore to increase the exchangeable value of the land and labour of the country. It puts into motion an additional quantity of industry, which given an added value to the annual produce. (WN, II, iii.18)

Smith's basic theme is that economic growth depends upon the accumulation of capital, and he went on from this point to draw attention to those factors which affect its *rate*.
In this connection he noted a number of issues:

1. 1. The incidence of commercial failure since ‘every injudicious project in agriculture, mines, fisheries, trade or manufactures, tends to diminish the funds destined for the maintenance of productive labour’ (II.iii.26).
2. 2. The cost of factors needed to maintain productive assets in a state of normal efficiency (II.ii.7).
3. 3. The area of investment to which a specific injection of capital was applied – it being Smith's contention, for example, that agriculture would support a great quantity of productive labour even than manufactures (see. for example, WN, II.v).
4. 4. The extent to which resources are devoted to the purchase of productive as distinct from unproductive labour. Productive labour for Smith involves the creation of commodities or ‘fixed subjects or vendible commodities’ which may be either investment or consumption goods and which contribute directly to the generation of income. Other forms of labour are described as unproductive although Smith did not deny that such services are useful. For example, he pointed out that the services of artists have a value to those who wish to pay for them. In the same way

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the services provided by governments are essential to the well-being of society. Yet all such services are by definition unproductive:

> The sovereign … with all the officers both of justice and war who serve under him, the whole army and navy, are unproductive labourers. They are the servants of the public, and are maintained by a part of the annual produce of the industry of other people. (II.iii.2)

It therefore follows that the *rate of growth* will be affected especially by the size of the government sector. Smith concluded:

> According … a smaller or great proportion …is in any one year employed in maintaining unproductive hands, the more in the one case and the less in the other will remain for productive, and next years produce will be greater or smaller accordingly. (II.iii.3)

Policy

Smith's analytical apparatus, allied to his judgement with respect to the probable trends of the economy, led him to advance the claims of economic liberty; claims which had already featured in LJ and which date back to his days in Edinburgh (Stewart, IV.25). The argument is repeated in WN, where Smith called upon the sovereign to discharge himself from a duty:

> in the attempting to perform which he must always be exposed to innumerable delusions, and for the proper performance of which no human wisdom or knowledge could ever be sufficient; the duty of superintending the industry of private people, and of directing it towards the employments most suitable to the interests of the society. (WN, IV.ix.51)

The statement is familiar, yet conceals a point of great significance; namely, that while the institutions of the exchange economy are consistent with the emergence of personal freedom (for example, under the law), they are not of themselves sufficient to establish what Smith described as the ‘system of natural liberty’ (WN, IV.ix.51). In fact, one of the most important functions of government is that of *identifying* and *removing* impediments to the effective working of the economy. Smith drew attention, for example, to the adverse effects of the statute of apprenticeship, and of corporate privileges. Regulations of this kind were criticized on the ground that they were both impolitic and unjust: unjust in that controls over qualification for entry to a trade were a violation ‘of this most sacred property which every man has in his own labour’ (WN, I.x.c.12) and impolitic in that such regulations are not of themselves sufficient to guarantee competence. But Smith particularly emphasized that the regulations in question would adversely affect the working of the market mechanism. The ‘statute of apprenticeship obstructs the free circulation of labour from one employment to another, even in the same place. The exclusive privileges of corporations obstruct it from one place to another, even in the same employment’ (WN, I.x.c.42). He also commented on the problems presented by the Poor Laws and the Laws of Settlement (WN, IV. ii.42), which further restricted the free movement of labour from one geographical location to another. Smith objected to positions of privilege, such as monopoly power, which he regarded as creations of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
civil law. The institution was again represented as impolitic and unjust: unjust in that a monopoly position is one of privilege and advantage, and therefore ‘contrary to that justice and equality of treatment which the sovereign owes to all the different orders of his subjects’, impolitic in that the prices at which goods so controlled are sold are ‘upon even occasion the highest that can be got’ (WN, I. vii.27). He added that monopoly is ‘a great enemy to good management’ (WN, I.xi.b.5) and that the institution had the additional defect of restricting the flow of capital to the trades affected as a result of the legal barriers to entry which were involved.

It is useful to distinguish Smith's objection to monopoly from his criticism of one expression of it; namely, the mercantile system of regulation which he described as the ‘modern system’ of policy, best understood ‘in our own country and in our own times’ (WN, IV.2). Smith asserted that mercantile policy aimed to secure a positive balance of trade through the control of exports and imports, a policy whose ‘logic’ was best expressed in terms of the Regulating Acts of Trade and Navigation, which currently determined the pattern of trade between Great Britain and her colonies and which were designed to create in effect a self-sufficient Atlantic Economic Community.

Smith objected to current policies of the type described on the ground that they artificially restricted the market and thus damaged opportunities for economic growth. It was Smith's contention that such policies were liable to that general objection which may be made to all the different expedients of the mercantile system, ‘the objection of forcing some part of the industry of the country into a channel less advantageous than that in which it would run of its own accord’ (WN, IV.v.a.24). In WN Smith placed more emphasis on interference with the allocative mechanism than he had done in LJ, where greater attention had been given to the inconsistency which was involved in seeking a positive balance of trade, an argument which relied heavily on Hume's analysis of the specie flow.

While it is difficult to judge the extent to which the claim for economic liberty explains the contemporary reception of WN, it may have been a major factor, at least in Britain (Schumpeter, 1954, p. 185). There can be no doubt that later generations found Smith's argument (and rhetoric) attractive. The celebrations to mark the 50th anniversary of the book showed a wide and continuing acceptance of the doctrines of free trade. In 1876, at a dinner held by the Political Economy Club to mark the centenary of WN, one speaker identified free trade as the most important consequence of the work done by ‘this simple Glasgow professor’, and predicted that

> there will be what may be called a large negative development of Political Economy tending to produce an important beneficial effect; and that is, such a development of Political Economy as will reduce the functions of government within a smaller and smaller compass. (Black, 1976, p. 51)

This view still commands wide contemporary support.

There can be no argument with Jacob Viner's contention that ‘Smith in general believed that there was, to say the least, a strong presumption against government activity’ (Viner, 1928, p. 14). But as Viner also reminded his auditors during the course of the Chicago conference which celebrated the 150th anniversary of the publication of WN, ‘Adam Smith was not a doctrinaire advocate of laissez-faire. He saw a wide and elastic range of activity for government’ (1928, pp. 153–4). A number of examples, all identified by Viner in a classic article, may briefly be reviewed here.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
First, Smith was prepared to justify specific policies to meet particular needs as these arose; the principle of intervention ad hoc. He defended the use of stamps on plate and linen as the most effectual guarantee of quality (WN, I.x.c.13), the compulsory regulation of mortgages (WN, V.ii.h.17), the legal enforcement of contracts (WN, I.ix.16) and government control of the coinage. In addition, he supported the granting of temporary monopolies to mercantile groups, to the inventors of new machines and, not surprisingly, to the authors of new books (WN, V.i.e.30). He further advised governments that where they were faced with taxes imposed by their competitors, retaliation could be in order, especially if such action had the effect of ensuring the 'repeal of the high duties or prohibitions complained of. The recovery of a great foreign market will generally more than compensate the transitory inconveniency of paying dearer during a short time for some sorts of goods' (WN, IV.ii.39).

Secondly, Smith advocated the use of taxation, not as a means of raising revenue but as a source of social reform, and as a means of compensating for what would now be described as a defective telescopic faculty. In the name of the public interest, Smith supported taxes on the retail sale of liquor in order to discourage the multiplication of alehouses (WN, V.ii.g.4) and differential rates on ale and spirits in order to reduce the sale of the latter (WN, V.ii.k.50). He advocated taxes on those proprietors of land who demanded rents in kind, and on those leases which prescribed a certain form of cultivation. In the same way, Smith argued that the practice of selling a future, for the sake of present, revenue should be discouraged on the ground that it reduced the working capital of the tenant and at the same time transferred a capital sum to those who would use it for the purposes of consumption (WN, V.ii.c. 12) rather than investment which would directly support productive labour.

Smith was well aware, to take a third example, that the modern version of the 'circular flow' depended on paper money and on credit; in effect, a system of 'dual circulation' involving a complex of transactions linking producers and merchants, and dealers and consumers (WN, II.ii.88). It is in this context that he advocated control over the rate of interest, to be set in such a way as to ensure that 'sober people are universally preferred, as borrowers, to prodigals and projectors' (WN, II.iv.15). He was also willing to regulate the small note issue in the interests of a stable banking system. To those who objected to such a proposal Smith replied that the interests of the community required it, and concluded that 'the obligation of building party walls, in order to prevent the communication of fire, is a violation of natural liberty, exactly of the same kind [as] the regulations of the banking trade which are here proposed' (WN, II.ii.94). Although Smith's monetary analysis is not regarded as amongst the strongest of his contributions, it should be remembered that as a witness of the collapse of the Ayr Bank, he was acutely aware of the problems generated by a sophisticated credit structure, and that it was in this context that he articulated a very general principle; namely, that 'those exertions of the natural liberty of a few individuals, which might endanger the security of the whole society, are, and ought to be, restrained by the laws of all governments; of the most free, as well as of the most despotical' (WN, II.ii.94).

Fourthly, emphasis should be given to Smith's contention that a major responsibility of government must be the provision of certain public works and institutions for facilitating the commerce of the society which were 'of such a nature, that the profit could never repay the expense to any individual or small number of individuals, and which it, therefore, cannot be expected that any individual or small number of individuals should erect or maintain' (WN, V.i.c.1). The examples of public works which he provided include roads, bridges, canals and harbours – all thoroughly in keeping with the conditions of the time and with Smith's emphasis on the importance of transport as a contribution to the effective operation of the market and to the process of economic growth. But although the list is short by modern standards,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the discussion is of interest for two main reasons.
First, Smith contended that public works or services should only be provided where market forces have failed to do so; secondly, he insisted that attention should be given to the requirements of efficiency and equity.
As Nathan Rosenberg (1960) has pointed out in an important article, Smith did not argue that governments should *directly* provide relevant services; rather, they should establish institutional arrangements so structured as to engage the motives and interests of those concerned. Smith tirelessly emphasized the point that in every trade and profession 'the exertion of the greater part of those who exercise it, is always in proportion to the necessity they are under of making that exertion' (WN, V.i. f.4); teachers, judges, professors, civil servants and administrators alike.
With regard to equity, Smith argued that public works, such as highways, bridges and canals should be paid for by those who use them and in proportion to the wear and tear occasioned – an expression of the general principle that the beneficiary should pay. He also defended direct payment on the ground of efficiency since only by this means would it be possible to ensure that necessary services would be provided where there was an identifiable need (WN, V.i.d.6).
Yet Smith recognized that it would not always be possible to fund or to maintain public services without recourse to general taxation. In this case he argued that 'local or provincial expenses of which the benefit is local or provincial' ought to be no burden on general taxation since 'It is unjust that the whole society should contribute towards an expense of which the benefit is confined to a part of society' (WN, V.i.i.3). However, he did agree that a general contribution would be appropriate in cases where public works benefit the whole society and cannot be maintained by the contribution 'of such particular members of the society as are most immediately benefited by them' (WN, V.i.i.6).
But here again, the main features of the system of liberty are relevant in that they affect the way in which taxation should be imposed. Smith pointed out on welfare grounds that taxes should be levied in accordance with the canons of equality, certainty, convenience and economy (WN, V.ii.b), and insisted that they should not be raised in ways which infringed the liberty of the subject – for example, through the odious visits and examinations of the tax-gatherer. Similarly, he argued that taxes ought not to interfere with the allocative mechanism (as, for example, taxes on necessities or particular employments) or constitute important disincentives to the individual effort on which the effective operation of the whole system depended (for example, taxes on profits or on the produce of land).

Ethics and history

The policy views which have just been considered are closely related to Smith's economic analysis. Others are only to be fully appreciated when seen against the background of his work on ethics and jurisprudence.
It will be recalled that for Smith moral judgement depends on a capacity for acts of imaginative sympathy, and that such acts can only take place within the context of some social group (TMS, III.i.3). However, Smith also observed that the mechanism of the impartial spectator might well break down in the context of the modern economy, due in part to the size of the manufacturing units and of the cities which housed them.
Smith observed that in the actual circumstances of modern society, the poor man could find himself in a situation where the 'mirror' of society (TMS, III.i.3) was ineffective. The 'man of rank and fortune is by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
his station the distinguished member of a great society, who attend to every part of his conduct, and who thereby oblige him to attend to every part of it himself'. But the 'man of low condition', while 'his conduct may be attended to' so long as he is a member of a country village, 'as soon as he comes into a great city, he is sunk, in obscurity and darkness. His conduct is observed and attended to by nobody, and he is therefore very likely to neglect it himself, and to abandon himself to every sort of low profligacy and vice' (WN, V.i.g.12).

In the modern context, Smith suggests that the individual thus placed would naturally seek some kind of compensation, often finding it not merely in religion but in religious sects; that is, small social groups within which he can acquire 'a degree of consideration which he never had before' (WN, V.i.g.12). Smith noted that the morals of such sects were often disagreeably 'rigorous and unsocial', recommending two policies to offset this.

The first of these is learning, on the ground that science is 'the great antidote to the poison of enthusiasm and superstition'. Smith suggested that government should institute 'some sort of probation, even in the higher and more difficult sciences, to be undergone by every person before he was permitted to exercise any liberal profession, or before he could be received as a candidate for any honourable office of trust or profit' (WN, V.i.g.14). The second remedy was through the encouragement given to those who might expose or dissipate the folly of sectarian bitterness by encouraging an interest in painting, music, dancing, drama – and satire (WN, V.i.g.15).

If the problems of solitude and isolation consequent on the growth of cities explain Smith's first group of points, a related trend in the shape of the division of labour helps to account for the second. In the earlier part of the argument, Smith had emphasized the gain to society at large which arose from improved productivity. But he noted later that this important source of economic benefit could also involve social costs:

> In the process of the division of labour, the employment of the far greater part of those who live by labour, that is, of the great body of the people, comes to be confined to a few very simple operations; frequently to one or two. But the understandings of the greater part of men are necessarily formed by their ordinary employments. The man whose life is spent in performing a few simple operations, of which the effects too are, perhaps, always the same, or very nearly the same, has no occasion to exert his understanding, or to exercise his invention in finding out expedients for removing difficulties which never occur. (WN, V.i.f.50)

Smith went on to point out that despite a dramatic increase in the level of *real income*, the modern worker could be relatively worse off than the poor savage, since in such primitive societies the varied occupations of all men – economic, political and military – preserve their minds from that 'drowsy stupidity, which, in a civilized society, seems to benumb the understanding of almost all the inferior ranks of people' (WN, V.i.f.51). It is the fact the 'labouring poor, that is the great body of the people' will fall into the state outlined that makes it necessary for government to intervene.

Smith's justification for intervention is, as before, market failure, in that the labouring poor, unlike those of rank and fortune, lack the leisure, means or (by virtue of their occupation) the inclination to provide education for their children (WN, V.i.f.53). In view of the nature of the problem, Smith's programme seems rather limited, based as it is on the premise that 'the common people cannot, in any civilized

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
society, be so well instructed as people of some rank and fortune' (WN, V.i.f.54). However, he did argue that they could all be taught 'the most essential parts of education … to read, write, and account' together with the 'elementary parts of geometry and mechanics' (WN, V.i.f.54, 55). Smith added:

> The publick can *impose* upon almost the whole body of the people the necessity of acquiring those most essential parts of education, by obliging every man to undergo an examination or probation in them before he can obtain the freedom in any corporation, or be allowed to set up any trade either in a village or town corporate. (WN, V.i.f.57; italics supplied)

Distinct from the above, although connected with it, is Smith's concern with the decline of martial spirit, which is the consequence of the nature of the fourth, or commercial, stage. He concluded that:

> Even though the martial spirit of the people were of no use towards the defence of society, yet to prevent that sort of mental mutilation, deformity and wretchedness, which cowardice necessarily involves in it, from spreading themselves through the great body of the people would still deserve the most serious attention of government. (WN, V.i.f.60)

Smith went on to liken the control of cowardice to the prevention of 'a leprosy or any other loathsome and offensive disease' – thus moving Jacob Viner to add public health to Smith's already lengthy list of governmental functions (Viner, 1928, p. 150). Such concerns have enabled Winch (1978) to find in Smith evidence of the *language* of an older, classical, concern with the problem of citizenship. Others (for example, see contributions in Hont and Ignatieff, 1983) have located Smith more firmly in the tradition of civic humanism.

The historical dimension of Smith's work also affects the treatment of policy, noting as he did that in every society subject to a process of transition, 'Laws frequently continue in force long after the circumstances, which first gave occasion to them, and which could also render them reasonable, are no more' (WN, III.ii.4). In such cases Smith suggested that arrangements which were once appropriate but are now no longer so should be removed, citing as examples the laws of succession and entail; laws which had been appropriate in the feudal period but which now had the effect of liming the sale and improvement of land. The continuous scrutiny of the relevance of particular laws is an important function of the 'legislator' (Haakonssen, 1981).

In a similar way, the treatment of justice and defence, both central services to be organized by the government, are clearly related to the discussion of the stages of history, an important part of the argument in the latter case being that a gradual change in the economic and social structure had necessitated the formal provision of an army (WN, V.i.a.b).

But perhaps the most striking and interesting features emerge when it is recalled that for Smith the fourth economic stage could be seen to be associated with a particular form of social and political structure which determines the outline of government and the context within which it must function. It may be recalled in this connection that Smith associated the fourth economic stage with the elimination of the relation of direct dependence which had be a characteristic of the feudal agrarian period. Politically, the significant and associated development appeared to been the diffusion of power

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
consequent on the emergence of new forms of wealth which, *at least in the peculiar circumstances of England*, had been reflected in the increased significance of the House of Commons.

Smith recognized that in this context government was a complex instrument, that the pursuit of office was itself a ‘dazzling object of ambition’ – a competitive game with as its object the attainment of ‘the great prizes which sometimes come from the wheel of the great state lottery of British politics’ (WN, IV. vii.c.75).

Yet for Smith the most important point was that the same economic forces which had served to elevate the House of Commons to a superior degree of influence had also served to make it an important focal point for sectional interests – a development which could seriously affect the legislation which was passed and thus affect that extensive view of the common good which ought ideally to direct the activities of Parliament.

It is recognized in the *Wealth of Nations* that the landed, moneyed, manufacturing and mercantile groups all constitute special interests which could impinge on the working of government. Smith referred frequently to their ‘clamourous importunity’, and went so far as to suggest that the power possessed by employers generally could seriously disadvantage other classes in the society’ (WN, I.x.c.61; cf. I. viii.12,13).

Smith insisted that any legislative proposals emanating from this class:

> ought always to be listened to with great precaution, and ought never to be adopted till after having been long and carefully examined, not only with the most scrupulous, but with the most suspicious attention. It comes from an order of men, whose interest is never exactly the same with that of the public, who have generally an interest to deceive and even to oppress the public, and who accordingly have, upon many occasions, both deceived and oppressed it. (WN, I.xi.p.10)

He was also aware of the dangers of manipulation arising from deployment of the civil list (LJA, iv.175–6).

It is equally interesting to note how often Smith referred to the constraints presented by the ‘confirmed habits and prejudices’ of the people, and to the necessity of adjusting legislation to what ‘the interests, prejudices, and temper of the times would admit of’ (WN, IV.v.b.40, 53, and V.i.g; cf. TNS, VI.ii.2.16).

Such passages add further meaning to the discussion of education. An educated people, Smith argued, would be more likely to see through the interested complaints of faction and sedition. He added a warning and a promise in remarking that:

> In free countries, where the safety of government depends very much on the favourable judgment which the people may form of its conduct, it must surely be of the highest importance that they should not be disposed to judge rashly or capriciously concerning it. (WN, V.i.f.61)

Aftermath

J.S. Mill, the archetypal classical economist, of a later period, is known to have remarked that ‘The *Wealth of Nations* is in many parts obsolete and in all, imperfect’. Writing in 1926, Edwin Cannan

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
observed:

> Very little of Adam Smith's scheme of economics has been left standing by subsequent enquirers. No one now holds his theory of value, his account of capital is seen to be hopeless confused, and his theory of distribution is explained as an ill-assorted union between his own theory of prices and the Physiocrat fanciful Economic Table. (1926, p. 123)

In view of authoritative judgements such as these, it is perhaps appropriate to ask what elements in his story should command the attention of the modern historian or economist. A number of points might be suggested.

First, there is the issue of *scope*. As we have seen, Smith's approach to the study of political economy was through the examination of history and ethics. The historical analysis is important in that he set out to explain the origins of the commercial stage. The ethical analysis is important to the economist because it is here that Smith identifies the human values which are appropriate to the modern situation. It is here that we confront the emphasis on the desire for status (which is essentially Veblenesque) and the qualities of mind which are necessary to attain this end: industry, frugality, prudence.

But the TMS also reminds us that the pursuit of economic ends takes place with a social context, and the men maximize their chances of success by respecting the rights of others. In Smith's sense of the term, ‘prudence’ is essentially rational self-love. In a favourite passage from the TMS (II.ii.2.1) Smith noted, with regard to the competitive individual, that:

> In the race for wealth, and honours, and preferments, he may run as hard as he can, and strain every nerve and muscle, in order to outstrip all his competitors. But if he should justle, or throw down any of them, the indulgence of the spectators is entirely at an end. It is violation of fair play, which they cannot admit of.

Smith's emphasis upon the fact that self-interested actions take place within a social setting and that men are motivated (generally) by a desire to be approved of by their fellows, raises some interesting questions of continuing relevance. For example, in an argument which bears upon the analysis of the TMS, Smith noted in effect that the rational individual may be constrained in respect of economic activity or choices by the reaction of the spectator of his conduct – a much more complex case than that which more modern approaches may suggest. Smith made much of the point in his discussion of Mandeville's ‘licentious system’ which supported the view that private vices were public benefits, in suggesting that the gratification of desire should be consistent with observance of the rules of property – as defined by the spectator, that is, by an external agency. In an interesting variant on this theme, Etzioni has noted that we need to recognize ‘at least two irreducible sources of valuation or utility: pleasure and morality’ (1988, 21–4; cf. Oakley, 1994).

Secondly, there is a series of issues which arise from Smith's interest in political economy as a system. The idea of a single all-embracing conceptual system, whose parts should be mutually consistent, is not easily attainable in an age where the division of labour has increased the quantity of science through specialization. Smith was aware of the division of labour in different areas of sciences, and of the fact

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that specialization often led to systems of thought which were inconsistent with each other (Astronomy, IV, 35, 52, 67). But the division of labour within a *branch* of science, for example, economics, has led to a situation where sub-branches of a single subject may be inconsistent with one another.

To take a third point, it may be noted that one of the most significant features of Smith's vision of the economic process lies in the fact that it has a significant time dimension. For example, in dealing with the problem of value in exchange, Smith made due allowance for the fact that the process involves judgements with regard to the utility of the commodities to be acquired, and the disutility involved in creating the goods to be exchanged. In the manner of his predecessors (Hutcheson, Carmichael and Pufendorf), Smith was aware of the distinction between utility (and disutility) anticipated and realized, and, therefore, of the process of adjustment which would inevitable take place through time,

Smith's theory of price, which allows for a wide range of changes in taste, is also distinctive in that it allows for competition among and between buyers and sellers, while presenting the allocative mechanism as one which involves simultaneous and interrelated adjustments in both factor and commodity markets.

As befits a writer who was concerned to address the problems of change, and adjustments to change, Smith's position was also distinctive in that he was not directly concerned with the phenomenon of *equilibrium*. For Smith, the ‘nature’ (supply) price was, as it were:

> The central price, to which the prices of all commodities are continually gravitating … whatever may be the obstacles which hinder them from settling in this centre of response and continuance, they are constantly tending towards it. (WN, I.viii.15)

But perhaps the most intriguing feature of the *macro* model is to be found in the way in which it was linked to the analytics of Book I and in the way in which it was specified. As noted earlier, Smith argued that incomes were generated as a result of productive activity, thus making it possible for commodities to be withdrawn from the ‘circulating’ capital of society. As he pointed out, the consumption goods withdrawn from the existing stock may be used up in the present period, or added to the stock reserved for immediate consumption; or used to replace more durable goods which had reached the end of their lives in the current period. In a similar manner, entrepreneurs and merchants may also add to their stocks of materials, or to their holding of fixed capital, while replacing the plant which had reached the end of its operational life. It is equally obvious that entrepreneurs and merchants may add to, or reduce their *inventories* in ways which will reflect the changed patterns of demand for consumption and investment goods, and their past and current levels of production. Variation in the level of inventories has profound implications for the conventional theory of the allocative mechanism.

Smith's emphasis upon the point that different ‘goods’ have different life-cycles also means that the pattern of purchase and replacement may vary continuously as the economy moves through different time periods, and in ways which reflect the various age profiles of particular products as well as the pattern of demand for them. If Smith's model of the circular flow is to be seen as a spiral, rather than a circle, it soon becomes evident that this spiral is likely to expand (and possibly *contract*) through time at variable rates. This point does not seem to have attracted much attention.

It is perhaps this total vision of the complex working of the economy that led Mark Blaug to comment on Smith's distinctive and sophisticated grasp of the economic *process* and to distinguish this from his contribution to particular *areas* of economic analysis.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Blaug noted that:

> In appraising Adam Smith, or any other economist, we ought always to remember that brilliance in handling purely economic concepts is a very different thing from a firm grasp of the essential logic of economic relationships. Superior technique does not imply superior insight and vice-versa. Judged by standards of analytical competence, Smith is not the greatest of eighteenth century economists. But for an acute insight into the nature of the economic process, it would be difficult to find Smith's equal. (1985, p. 57)

Joseph Schumpeter, not always a warm critic of ‘A. Smith’, yet regarded WN as ‘the peak success of (the) period:

> Though the *Wealth of Nations* contained no really novel ideas, and though it cannot rank with Newton's *Principia* or Darwin's *Origin* as an intellectual achievement, it is a great performance all the same and fully deserved its success. (1954, p. 185)

Writing from a different, but related, point of view, A.L. Macfie noted that ‘the Scottish method was more concerned with giving a broad, well balanced picture seen from different points of view than with logical rigour’ (1967, 22–3).

It has been argued above that Smith's approach to the study of political economy has some distinctive features which deserve the attention of the modern student of the discipline, but which do not seem to loom large in modern teaching. But nor can it be said that the classical system which was to follow Smith did any better.

Richard Teichgraeber's research (1987) revealed that there ‘is no evidence to show that many people exploited his arguments with great care before the first two decades of the nineteenth century’. He concluded:

> It would seem at the time of his death that Smith was widely known and admired as the author of the *Wealth of Nations*. Yet it should be noted that only a handful of his contemporaries had come to see his book as uniquely influential. (1987, p. 363)

But, as we have seen, there were commentators who understood Smith's analytical purpose, notably Thomas Pownall and Smith's biographer, Dugald Stewart. The latter indeed became an important channel of communication between Smith and a later generation of students – many of whom were to contribute to debates published by the *Edinburgh Review* in the early part of the new century (Winch, 1994, p. 91). Amongst the contributors we can number J.R. McCulloch (1789–1864), editor of the *Scotsman*, discipline of Ricardo and a writer who contributed nearly 80 articles to the *Review* between 1818 and 1837.

However, Black made a different point in his ‘Historical Perspective’ in observing that for Smith's early successors the *Wealth* was ‘not so much a classical monument to be inspected, but as a structure to be examined and improved where necessary (1976, p. 44). It was thought that there were ambiguities in respect of Smith's treatment of value, interest, rent, population theory and the theory of economic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
growth. These ambiguities were reduced by the work of T.R. Malthus (1766–1834) whose *Essay on the Principle of Population* was first published in 1798; by the French economist, J.B. Say (1826–1896), *Traite d'economie politique* (1803) and especially by David Ricardo (1772–1823), *Principles* (1817). In this context we can name a number of writers who developed the short-run and dynamic themes associated with Say and Ricardo, such as James Mill (1773–1836), *Elements of Political Economy* (1821) and, of course, his son John Stuart Mill (1806–1873) whose *Principles* brought to close this early version of the classical system.

Among Say's contributions we must number a version of the classical short-run macroeconomic system which accepted the view that the supply of commodities generates income and purchasing power, and also the Smithian assumption that the only use of money was to circulate commodities while also accepting Smith's assumption that an act of savings would normally be matched by a decision to invest. In the relatively short run, the tendency of the economy was to full employment; a situation sustained by self-regulating mechanisms such as Smith's theory of price and allocation; a theory which was not seriously questioned. In Ricardo's case the emphasis was upon a generalized statement of Smith's labour embodied theory of value, a revised theory of rent, and a theory of growth which under the assumptions of constant technique and a closed economy, suggested that the normal progression of an economy was from an advancing state to a stationary state where no further growth was possible. Both models are formal, operating under specified assumptions. They are also essentially mathematical in character even if they clearly do owe much to Smith.

But there are important differences, arising not least from the fact that Smith's own approach was not narrowly ‘mathematical’ (Macfie, 1967, pp. 22–3), as compared, for example, with Ricardo (Baumol, 1962).

There was another difficulty arising from the fact that there was a tendency to assume that the basis of the subject of political economy dated from 1776, the year in which the *Wealth of Nations* first appeared. Donald Winch quotes an important passage from J.-B. Say, Smith's committed, but by no means uncritical, disciple. Say wrote that:

> Whenever the inquiry into the *Wealth of Nations* is perused with the attention it so well merits, it will be perceived that until the epoch of its publication, the science of political economy did not exist. (Quoted in Winch, 1994, p. 103)

Terence Hutchison has argued that the ‘losses and exclusions which ensued after 1776, with the subsequent transformation of the subject and the rise to dominance of the English classical orthodoxy were immense’ (1988, p. 370). One such loss was the Physiocratic concept of the circular flow, to which Smith owed so much. Other losses occurred as a result of ignoring the contributions of Smith's close friend, David Hume, and the work of the latter's friend, Sir James Steuart (1713–1780), *Principles* (1767). The use of the historical method as applied to economic analysis and policy was one such loss and so too was the concern with structural unemployment, and the model of ‘primitive’ (pre-capitalist) accumulation. In addition the classical orthodoxy showed little interest in the problems presented by differential rates of growth in the context of international trade – hardly surprising in view of the fact that Smith largely ignored the problems identified by his two Scottish predecessors.

Ironically, Smith's own work did not always benefit from the work of those who ‘inspected’ the edifice.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Here attention may be drawn to Smith's version of the ‘circular flow’ with its complex focus on period analysis and on the fact that all commodities have different life-cycles. Nor was much attention given to the *Smithian* use of time.

Ironically, the new orthodoxy also made it possible to think of political economy as a discipline which was quite separate from ethics and jurisprudence, thus obscuring Smith's true purpose. In referring to the way in which Smith organized his system of social science (ethics, jurisprudence, economics) Hutchison observed in a telling passage, that Smith was led as it by an Invisible Hand to promote an end which was no part of his original intention, that of ‘establishing political economy as a separate, autonomous disciple’ (1988, p. 355). A.L. Macfie made a related point in observing that ‘it is a paradox of history that the analytics of Book I, in which Smith took his own line, should have eclipsed the philosophical and historical methods in which he so revelled, and which showed his Scots character’ (1967, p. 21).

See Also

- British classical economics

Selected works

Editions and abbreviations. An excellent edition of the *Lectures of Jurisprudence* was brought out by Edwin Cannan in 1896 (Oxford: Clarendon Press). Cannan also prepared a valuable edition of the *Wealth of Nations* in 1904 (London: Methuen). J.M. Lothian edited the *Lectures on Rhetoric* in 1963 (Edinburgh: Nelson).

Subsequent references are to the Glasgow edition of the *Works and Correspondence of Adam Smith* (Oxford, Clarendon Press, 1976–83) and follow the usages of that edition. The edition consists of:

I	*The Theory of Moral Sentiments* (TMS), ed. D.D. Raphael and A.L. Macfie, 1976.
II	*An Inquiry into the Nature and Causes of the Wealth of Nations* (WN), ed. R.H. Campbell, A.S. Skinner and W.B. Todd, 1976.
III	*Essays on Philosophical Subjects* (EPS), ed. D.D. Raphael and A.S. Skinner, 1980.
	This volume includes:
(i)	‘The History of the Ancient Logics and Metaphysics’ (Ancient Logics).
(ii)	‘The History of the Ancient Physics’ (Ancient Physics).
(iii)	‘The History of Astronomy’ (Astronomy).
(iv)	‘Of the affinity between Certain English and Italian Verses’ (English and Italian Verses).
(v)	‘Of the External Senses’ (External Senses).
(vi)	‘Of the Nature of the Imitation which takes place in what are called the Imitative Arts’ (Imitative Arts).
(vii)	‘Of the Affinity between Music, Dancing and Poetry’. Items (i) to (vii), above, were prepared by W.P.D. Wightman.
(viii)	‘Of the Affinity between Certain English and Italian Verses’.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(ix) Contributions to the *Edinburgh Review* (1755–6):

(a) Review of Johnson's Dictionary.

(b) A Letter to the authors of the *Edinburgh Review* (Letter).

(x) Preface to William Hamilton's Poems on general Occasions. Items (viii) to (x), above, were prepared by J.C. Bryce.

(xi) Dugald Stewart, 'Account of the Life and Writings of Adam Smith LL.D' (Stewart), ed. I.S. Ross.

IV *Lectures on Rhetoric and Belles Lettres* (LRBL), ed. J.C. Bryce; general editor, A.S. Skinner, 1983.

This volume includes:

'Considerations concerning the First Formation of Languages' (Considerations).

V *Lectures on Jurisprudence* (LJ), ed. R.L. Meek, P.G. Stein and D.D. Raphael, 1978.

This volume includes:

(i) Student notes for the session 1762–3 (LJA).

(ii) Student notes for the session 1763–4 but dated 1766 (LJB).

(iii) The 'Early Draft' of the *Wealth of Nations* (ED).

(iv) Two Fragments on the Division of Labour, (FA and (FB).

VI *Correspondence of Adam Smith* (Corr.), ed. E.C. Mossner and I.S. Ross, 1977.

This volume includes:

(i) 'A Letter from Governor Pownall to Adam Smith (1776)'.

(ii) 'Smith's thoughts on the State of the Contest with America, February 1778', ed. D. Stevens.

(iii) Jeremy Bentham's 'Letters' to Adam Smith (1787, 1790).

Associated volume

Essays on Adam Smith (EAS), ed. A.S. Skinner and T. Wilson. Oxford: Clarendon Press, 1975.

References to Corr. give letter number and date. References to LJ and LRBL give volume and page number from the MS. All other references provide section, chapter and paragraph number in order to facilitate the use of different editions. For example, Astronomy, II. 4='History of Astronomy, section II, para. 4. Stewart, I.12=Dugald Stewart, 'Account', section I, para. 12. TMS, I.i.5.5=TMS, Part I, section I, chapter 5, para. 5. WN, V.i.f.26=WN, Book V, chapter I, section 6, para. 26.

Bibliography

The literature on Smith is immense. But there are two useful surveys. The first is by Vivienne Brown, 'Mere Inventions of the Imagination: A Survey of Recent Literature on Adam Smith', *Economics and Philosophy* 13 (1997), 281–312. The second has been prepared by Anthony Brewer, 'Let Us Now Praise Famous Men: Assessments of Adam Smith's economics', *Adam Smith Review* 3 (2007).

Bagolini, L. 1975. The topicality of Adam Smith's notion of sympathy. In EAS.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Baumol, W.I. 1962. *Economic Dynamics*. London: Macmillan.

Berry, C.J. 1997. *Social Theory of the Scottish Enlightenment*. Edinburgh: Edinburgh University Press.

Black, J.B. 1926. *The Art of History*. New York: Methuen.

Black, R.D.C. 1976. Smith's contribution in historical perspective. In *Essays in Honour of Adam Smith*, ed. T. Wilson and A.S. Skinner. Oxford: Clarendon Press. Cited as EAS.

Blaug, M. 1985. *Economic Theory in Retrospect*. London: Heinemann.

Brown, M. 1988. *Adam Smith's Economics*. New York: Croom Helm.

Brown, V. 1994. *Adam Smith's Discourse*. London: Routledge.

Campbell, R.H. and Skinner, A.S. 1982. *Adam Smith*. London: Croom Helm.

Cannan, E. 1926. Adam Smith as an economist. *Economica* 6, 123–34.

Etzioni, A. 1988. *The Moral Dimension: Towards a New Economics*. London: Macmillan.

Fitzgibbons, A. 1995. *Adam Smith's System of Liberty, Wealth and Virtue*. Oxford: Clarendon Press.

Groenewegen, P. 1969. Turgot and Adam Smith. *Scottish Journal of Political Economy* 16, 271–87.

Haakonssen, K. 1981. *The Science of a Legislator*. Cambridge: Cambridge University Press.

Hollander, S. 1973. *The Economics of Adam Smith*. Toronto: Toronto University Press.

Hont, I. and Ignatieff, M., eds. 1983. *Wealth and Virtue: The Shaping of Political Economy in the Scottish Enlightenment*. Cambridge: Cambridge University Press.

Howell, W.S. 1975. Adam Smith's lectures on rhetoric: an historical assessment. In EAS.

Hume, D. 1985. *Essays Moral, Political and Literary*, ed. E.F. Miller. Indianapolis: Liberty Fund.

Hutchison, T. 1988. *Before Adam Smith*. Oxford: Blackwell.

Jeck, A. 1994. The macro structure of Adam Smith's theoretical system. *European Journal of the History of Economic Thought* 3, 551–76.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Jensen, H.E. 1984. Sources and contours: conceptualised reality. In *Adam Smith: Critical Assessments*, ed. J.C. Wood. London: Croom Helm.

Koebner, R. 1961. *Empire*. Cambridge: Cambridge University Press.

Lindgren, R. 1973. *The Social Philosophy of Adam Smith*. The Hague: Nijhoff.

Macfie, A.L. 1967. *The Individual in Society: Papers Relating to Adam Smith*. London: Allen and Unwin.

McNally, D. 1988. *Political Economy and the Rise of Capitalism*. Berkeley: University of California Press.

Meek, R.I. 1962. *The Economics of Physiocracy*. London: Allen and Unwin.

Meek, R.L. 1973. *Turgot on Progress, Sociology and Economics*. Cambridge: Cambridge University Press.

Meek, R.L. 1976. *Social Science and the Ignoble Savage*. Cambridge: Cambridge University Press.

Mizuta, H. 2000. *Adam Smith's Library: A Catalogue*. Oxford: Clarendon Press.

Oakley, A. 1994. *Classical Economic Man*. Aldershot: Edward Elgar.

O'Brien, D.P. 1975. *The Classical Economists*. Oxford: Clarendon Press.

Peacock, A.T. 1975. The treatment of the principles of public finance in the *Wealth of Nations*. In EAS.

Peil, J. 1999. *Adam Smith: Economic Science*. Cheltenham: Edward Elgar.

Phillipson, N. 1983. Adam Smith as a civic moralist. In Hont and Ignatieff (1983).

Pocock, J.G.A. 1983. Cambridge paradigms and Scotch philosophers. In Hont and Ignatieff (1983).

Rae, J. 1895. *Life of Adam Smith*. London: Macmillan.

Raphael, D.D. 1985. *Adam Smith*. Oxford: Oxford University Press.

Raphael, D.D. 2007. *The Impartial Spectator: Adam Smith's Moral Philosophy*. Oxford: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Robbins, I. 1953. *The Theory of Economic Policy in English Classical Political Economy*. London: Macmillan.

Rosenberg, N. 1960. Some institutional aspects of the Wealth of Nations. *Journal of Political Economy* 68, 557–70.

Ross, I. 1995. *Life of Adam Smith*. Oxford: Clarendon Press.

Rothbard, M. 1995. *An Austrian Perspective on the History of Economic Thought*. Aldershot: Edward Elgar.

Rothschild, E. 2001. *Economic Sentiments*. Cambridge, MA: Harvard University Press.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen and Unwin.

Scott, W.R. 1900. *Francis Hutcheson: His Life, Teaching and Position in the History of Philosophy*. Cambridge: Cambridge University Press.

Scott, W.R. 1937. *Adam Smith as Student and Professor*. Glasgow: Jacksons.

Sen, A. 1987. *On Ethics and Economics*. Oxford: Blackwell.

Skinner, A.S. 1996. *A System of Social Science: Papers Relating to Adam Smith*, 2nd edn. Oxford: Clarendon Press.

Taylor, W.L. 1965. *Francis Hutcheson and David Hume as Predecessors of Adam Smith*. Durham, NC: University of North Carolina Press.

Teichgraeber, R. 1987. ‘Less abused than I had reason to expect’: the reception of the *Wealth of Nations* in Britain, 1776–1790. *Historical Journal* 30, 337–66.

Tribe, K. 1988. *Governing Economy: The Reformation of German Economic Discourse*. Cambridge: Cambridge University Press.

Viner, J. 1928. Adam Smith and laissez faire. In *Adam Smith, 1776–1926*. Chicago: Chicago University Press.

Vivenza, G. 2001. *Adam Smith and the Classics*. Oxford: Oxford University Press.

Werhane, P.H. 1991. *Adam Smith and his Legacy for Modern Capitalism*. Oxford: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Winch, D. 1978. *Adam Smith's Politics: An Essay in Historiographic Revision*. Cambridge: Cambridge University Press.

Winch, D. 1994. Nationalism and cosmopolitanism. In *Political Economy and National Realities*, ed. M. Albertone and A. Masoero. Turin: Fondazione Luigi Einaudi.

Young, J. 1977. *Economics as a Moral Science: The Political Economy of Adam Smith*. Cheltenham: Edward Elgar.

How to cite this article

Skinner, Andrew. "Smith, Adam (1723–1790)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000154> doi:10.1057/9780230226203.1545

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Smith, Bruce D. (1954 – 2002)

P. Dean Corbae
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

endogenous growth; exchange rate policy; financial intermediation; monetary policy; production externalities; quantity theory of money; Smith, B.

Article

Bruce David Smith was born on 21 September 1954 in St. Paul, Minnesota, and died in Rochester, Minnesota, on 9 July 2002. Smith graduated with a BA in economics from the University of Minnesota in 1977 and obtained his Ph.D. in economics from the Massachusetts Institute of Technology in 1981. Smith's career began as an assistant professor at Boston College (1981–2). From there he moved to the research department at the Federal Reserve Bank of Minneapolis (1982–6), which was where his lifelong research interests in monetary history, monetary theory and financial intermediation blossomed among colleagues like Thomas Sargent, Neil Wallace, Ed Prescott, John Boyd and Warren Weber. After visits to Carnegie-Mellon and the University of California at Santa Barbara, he moved to the University of Western Ontario as an associate professor from 1987 to1990. Smith moved to Cornell University in 1990 and finally to the University of Texas at Austin in 1996, where he was the Fred Hofheinz Regent's Professor of Economics.

Smith's scholarly research consists of nearly one hundred published papers (a complete list of Smith's papers can be found in the *Federal Reserve Bank of Minneapolis Quarterly Review*, 2002). His research includes widely cited papers about monetary history, the causes and consequences of banking panics, the impact of financial market development on per-capita income and growth, the impact of inflation on financial market development, the macroeconomic effects of various types of credit market imperfections, and the optimal conduct of monetary and exchange rate policy.

The importance of financial intermediation for economic growth and stability underlies nearly all of Smith's research. In ‘Taking intermediation seriously’ (2003), his last single-authored paper, Smith argues that the information and spatial frictions which underlie a welfare-improving role for intermediation have important implications for the conduct of monetary policy. In particular, Smith lays out various theoretical models that show not only how excessive monetary expansion can lead to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
decreases in growth, but also how restrictive monetary policy (for instance, the Friedman rule) can be suboptimal.

One of Smith's most widely cited articles, ‘Financial Intermediation and Endogenous Growth’ (1991), co-authored with Valerie Bencivenga, provides a theoretical foundation for a large body of empirical research that finds a correlation between measures of financial market development and real output growth. Bencivenga and Smith provide a model with production externalities where liquidity provided by banks allows private agents to reduce the fraction of their savings held in the form of unproductive liquid assets and to increase their holdings of productive capital assets inducing growth.

Smith viewed economic history as a laboratory for understanding money and the theory of financial intermediation that he argued remains relevant for present-day policy. In ‘Some colonial evidence on two theories of money: Maryland and the Carolinas’, he argues that data from North and South Carolina is inconsistent with the quantity theory of money. For example, while the per capita stock of paper money more than tripled from 1755 to 1760, the price level increased by only seven per cent. On the other hand, data from Maryland, which had a unique method of backing the value of its currency with deliveries of sterling from the Bank of England, appears to be consistent with what Smith calls the Sargent–Wallace approach, where the value of money is determined in the same way that other privately issued assets are priced as the expected present discounted value of future cash flows (in this case, claims to sterling deliveries).

See Also

- financial intermediation
- optimum quantity of money

Selected works

1985. Some colonial evidence on two theories of money: Maryland and the Carolinas. *Journal of Political Economy* 93, 1178–211.

1991. (With V. Bencivenga.) Financial intermediation and endogenous growth. *Review of Economic Studies* 58, 195–209.

2003. Taking intermediation seriously. *Journal of Money, Credit, and Banking* 35, 1319–57.

Bibliography

2002. The published work of Bruce D. Smith. *Federal Reserve Bank of Minneapolis Quarterly Review*, 26(4), 43–52. Online. Available at http://woodrow.mpls.frb.fed.us/research/qr/qr2644.pdf, accessed 10 March 2006.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Corbae, P. Dean. "Smith, Bruce D. (1954–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000497> doi:10.1057/9780230226203.1546

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Smith, Vernon (born 1927)

Catherine C. Eckel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Vernon Lomax Smith shared the 2002 Nobel Prize in Economics for his contributions to the development of experimental economics. Among his insights is the idea that ‘institutions matter’: trading institutions affect the efficiency of a market, and the optimal institution depends on the market context. He has made important contributions to market and institutional design, public goods, bargaining, experimental methodology and the philosophy of science.

Keywords

behavioural finance; bilateral bargaining; Chamberlin, E. H.; competitive equilibrium; constructivist rationality; dictator game; double auction; ecological rationality; experimental economics; experimental methods in economics; posted offer markets; smart markets; Smith, V. L.; social distance

Article

Vernon Lomax Smith is an American economist who shared the 2002 Nobel Prize in Economics (with the psychologist Daniel Kahneman) for his pioneering work on the methodology of laboratory experiments in economics. He is a remarkable scholar and a true pioneer in the quest to understand market institutions (such as auction mechanisms) and nonmarket institutions (such as bargaining rules), as well as the structure and motivation of individual behaviour. His methodological contributions helped overturn the traditional notion of economics as an inherently non-experimental science.
Smith was born on 1 January 1927 in Wichita, Kansas. He studied electrical engineering at the California Institute of Technology, receiving a bachelor's degree in 1949, then earned an MA in economics from the University of Kansas in 1952, and a Ph.D. from Harvard University in 1955. Smith's first teaching position was at the Krannert School of Management at Purdue University (1955–67), where he began his work in experimental economics. After appointments at Stanford, the University of Massachusetts, the Center for Advanced Study in the Behavioral Sciences and Caltech, in 1975 he

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
accepted a position at the University of Arizona, where he was to spend the next 26 years and build a body of research that would result in a Nobel Prize. Since 2001, Smith has been Professor of Economics and Law at George Mason University, where he is a research scholar in the Interdisciplinary Center for Economic Science, and a Fellow of the Mercatus Center.

Author or co-author of over 200 articles and books on capital theory, finance, natural resource economics and experimental economics, Smith has served on editorial boards of journals including (among many others) the *American Economic Review*. He has been influential in professional organizations, serving as president of the Public Choice Society, the Economic Science Association, and the Western Economic Association. His honours are many: he is a Fellow of the Econometric Society, the American Association for the Advancement of Science, and the American Academy of Arts and Sciences, and a Distinguished Fellow of the American Economic Association. He was elected a member of the National Academy of Sciences in 1995. Smith's collected papers were published by Cambridge University Press in 1991, with a second volume in 2000.

The press release issued by the Royal Swedish Academy of Sciences on the occasion of the 2002 Nobel Prize summarizes the contributions that earned Smith the award:

> Vernon Smith has laid the foundation for the field of experimental economics. He has developed an array of experimental methods, setting standards for what constitutes a reliable laboratory experiment in economics. In his own experimental work, he has demonstrated the importance of alternative market institutions, e.g., how the revenue expected by a seller depends on the choice of auction method. Smith has also spearheaded ‘wind-tunnel tests’, where trials of new, alternative market designs – e.g., when deregulating electricity markets – are carried out in the lab before being implemented in practice. His work has been instrumental in establishing experiments as an essential tool in empirical economic analysis. (Press Release: The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, 9 October 2002)

Smith's early contributions focused on markets and how different ways of organizing exchange might lead to different outcomes. The next phase built upon the principles learned in the laboratory to design new institutions, as deregulation and privatization created unprecedented opportunities for the emergence of new markets around the world. Later work turned to the behaviour of agents in bilateral bargaining situations, involving pairs or small groups of participants, and exploring the nature and role of personal social exchange in decision-making in these settings. Recent research has explored the relationship between brain functions and decision-making (2001), the emergence of exchange systems in the absence of extant institutions that are typically imposed exogenously in experimental economies (2006a; 2006b), and the philosophy of science as it relates to the experimental method in economics. In every phase of his broad research agenda, he has produced important insights and critical methodological innovations.

Markets work, but not for the reasons we think

Vernon Smith tells the story of his early experiments (see Smith, 1991, pp. 154–8). As a graduate student at Harvard, he participated in the classroom experiment of Edward Chamberlin. As a teaching

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
exercise, Chamberlin gave students seller costs or buyer values, then instructed them to circulate in the room, and negotiate prices with their counterparts. He then collected the prices, and displayed them, purportedly to illustrate that markets do not work as the perfectly competitive model suggests.

When Smith started teaching at Purdue, he adapted Chamberlin's experiment to his own classes, making a few modifications so that the market more closely resembled a stock market. He distributed costs and values to the participants, as Chamberlin had done, but had traders call out bids and offers. A pit boss recorded the bids and offers on the blackboard, and a trade occurred when a buyer accepted a seller's offer, or a seller accepted a buyer's bid. He also repeated the market, to allow students to learn the mechanics of the trading situation. He says, ‘These two changes seemed to be the appropriate modifications to do a more credible job of rejecting competitive price theory’ (1991, p. 155). To his surprise, the market converged in a couple of rounds to the predicted competitive equilibrium price and quantity. Thinking this might just be a fluke, he repeated it in another class, with the same result. Finally, imagining that the result might depend on the symmetric producer and consumer surpluses in his particular set-up, he ran another market with highly asymmetric surpluses, and once again ‘the darned thing converged to competitive equilibrium’. In all cases, after a few rounds, all trades were taking place within a few cents of the same price. One can almost imagine him saying, ‘Well look at that. Markets work!’ These early ‘double oral auction’ (DOA, also referred to as an oral double auction) market experiments were published as Smith (1962), which also reports tests of the effects of shifts in supply and demand.

This exercise is such a strong illustration of the power of competitive markets that many economists use a classroom version of it to introduce their students to the supply and demand model. Participating in this experiment brings the model to life, and experiencing convergence to competitive equilibrium is an unforgettable lesson for students. By conducting his early experiments in the classroom, Smith inadvertently taught us how to teach economics.

These studies were followed by comparisons of different trading institutions. For example, compared with the posted offer institution found in most retail markets, where sellers post prices and buyers choose whether to buy, the DOA converges more quickly, and responds faster to changes in demand or supply. Smith termed it a ‘disciplining’ institution: the DOA is disciplining in the sense that buyers and sellers quickly discover whether they have left money on the table, and can modify their decisions immediately. In a posted offer market, this is less clear, as buyers who do not win the auction do not find out about other's values.

Another important contribution of Smith's early work is to discover the interactions between market structure and trading institutions in determining the outcome in the market. His work showed that market power is much harder to exercise in the DOA than in a posted offer market, where the monopolist can more easily sustain monopoly prices (Smith, 1981). This result stands outside standard economic theory, as research in industrial organization says little or nothing about the trading institution. A major contribution of experimental economics, and Smith's work in particular, is the insight that ‘institutions matter’. This insight belongs in every Principles of Economics course, though it has not yet made its way into standard texts.

This new-found understanding of how markets and institutions work could then be harnessed to design new markets and institutions, using the laboratory as a ‘wind tunnel’. The development of computer-assisted ‘smart markets’ expanded market design to a whole range of complex policy applications. Smith's work contributed to the design of many such markets, including airport landing slots, natural

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
gas, wholesale electricity, as well as the design of the Arizona Stock Exchange, and the highly profitable auction by the US Federal Communications Commission of the portions of the microwave spectrum used for cellular phones, among other applications.

While Smith is known for his work showing how markets work, he also has explored situations when they might not work so well. This work showed that implementing the common knowledge assumption of perfectly competitive markets and game theory can be difficult, and this failure can lead to unexpected outcomes. For example, Smith and his colleagues have contributed to the growing field of behavioural finance by replicating the propensity of asset markets to bubble and crash. When valuations are based on subjects' beliefs about others' knowledge, beliefs and strategies, prices can deviate substantially from the underlying fundamentals of the asset (1988). In addition, he shows that more information is not necessarily better. When subjects are given information about others' payoffs, fulfilling one of the theoretical assumptions of competitive markets, this can delay or distort convergence to equilibrium (Smith, 1976).

Dominated strategies are for playing

What happens where there is no competitive market to discipline trading behaviour? While people often act in accordance with the rational actor model in competitive situations, the negotiated outcomes of bargaining games allow scope for other motives to emerge. In the 1980s, laboratory tests of game-theoretic models began to produce results that revealed a penchant for fairness on the part of the participants. In a series of papers, Smith and his colleagues explored the importance of social distance between subjects. The greater the social distance, the more likely subjects were to behave selfishly in a dictator game (where one player decides the allocation of an endowment between himself and a counterpart) (1994; 1996a).

Subjects also take advantage of strategies that, according to theory, they should never play. Smith and his colleagues were the first to show that subjects will take advantage of an opportunity to punish a counterpart for unfair behaviour, even when that choice is costly for them (1996b). In this study and others, dominated strategies are chosen by subjects, when they are useful to punish bad behaviour by reducing a counterpart's payoff.

People are more rational than the agents in our models

Smith's Nobel lecture (Smith, 2003) surveys the territory that has been colonized by laboratory experimental economics. He organizes the research by distinguishing two types of rationality: constructivist, by which he means the sort of rationality built into the standard social science model of 'economic man'; and ecological, by which he refers to the use of reason to understand the emergent order and embodied intelligence of cultural rules, norms, and the institutions that result from regular human interactions, but that are not deliberately designed. Experimental economics taught Smith that our understanding of economic phenomena requires both constructivist and ecological rationality. From the discussion it is clear that his own research began with the constructivist model, but from his experience in the laboratory, he grew to appreciate that the richness of human behaviour required an augmented view. Smith's view of rationality grew out of a profound curiosity about and respect for human behaviour, and was shaped by his experience in the laboratory and his careful observation of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
human behaviour in a wide variety of settings.

See Also

- auctions (experiments)
- experimental economics
- experimental methods in economics
- experimental economics, history of

Selected works

1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70, 111–37.

1976. Experimental economics: induced value theory. *American Economic Review* 66, 274–9.

1981. An empirical study of decentralized institutions of monopoly restraint. In *Essays in Contemporary Fields of Economics*, ed. G. Horwich and J.P. Quirk. West Lafayette: Purdue University Press.

1988. (With G. Suchanek and A. Williams.) Bubbles, crashes and endogenous expectations in experimental spot asset markets. *Econometrica* 56, 1119–51.

1991. *Papers in Experimental Economics*. New York: Cambridge University Press.

1994. (With E. Hoffman, K.A. McCabe and K. Shachat.) Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* 7, 346–80.

1996a. (With E. Hoffman and K.A. McCabe.) Social distance and other regarding behavior in dictator games. *American Economic Review* 86, 653–60.

1996b. (With K.A. McCabe and S.J. Rassenti.) Game theory and reciprocity in some extensive form experimental games. *Proceedings of the National Academy of Science* 93, 13421–8.

2000. *Bargaining and Market Behavior*. Cambridge: Cambridge University Press.

2001. (With K.A. McCabe, D. Houser, L. Ryan and T. Trouard.) A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* 98, 11832–5.

2003. Constructivist and ecological rationality in economics. *American Economic Review* 93, 465–508.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

2006a. (With S. Crockett and B.J. Wilson.) Exchange and specialization as a discovery process. Working paper, George Mason University .

2006b. (With E. Kimbrough and B.J. Wilson.) Historical property rights, sociality, and the emergence of impersonal exchange in long-distance trade. Working paper, George Mason University.

How to cite this article

Eckel, Catherine C. "Smith, Vernon (born 1927)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000442> doi:10.1057/9780230226203.1547

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

SNP: nonparametric time series analysis

A. Ronald Gallant
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

SNP is a method of nonparametric multivariate time series analysis. It employs an expansion in Hermite functions to approximate the conditional density of a multivariate process. An appealing feature of the expansion is that it is a nonlinear nonparametric model that directly nests the Gaussian VAR model, the semiparametric ARCH model, the Gaussian GARCH model, and the semiparametric GARCH model. The unrestricted SNP expansion is more general than any of these models. The SNP model is fitted using conventional maximum likelihood together with a model selection strategy that determines the appropriate order of expansion.

Keywords

ARCH models; bootstrap; density; GARCH models; Kalman filter; kernels; maximum likelihood; method of moment estimation; nonparametric time series analysis; semi-nonparametric (SNP) models; splines; statistical inference; vector autoregressions

Article

SNP is a method of multivariate nonparametric time series analysis. SNP is an abbreviation of ‘semi-nonparametric’ which was introduced by Gallant and Nychka (1987) to suggest the notion of a statistical inference methodology that lies halfway between parametric and nonparametric inference. The method employs an expansion in Hermite functions to approximate the conditional density of a multivariate process.

The leading term of this expansion can be chosen through selection of model parameters to be a Gaussian vector autoregression (VAR) model, a semi-parametric VAR model, a Gaussian ARCH model (Engle, 1982), a semiparametric ARCH model, a Gaussian GARCH model (Bollerslev, 1986), or a semiparametric GARCH model, either univariate or multivariate in each case. The unrestricted SNP expansion is more general than that of any of these models. The SNP model is fitted using maximum likelihood together with a model selection strategy that determines the appropriate order of expansion. Because the SNP model possesses a score, it is an ideal candidate for the auxiliary model in connection with efficient method of moment estimation (Gallant and Tauchen, 1996). Due to its leading term, the SNP approach does not suffer from the curse of dimensionality to the same extent as kernels and splines. In regions where data are sparse, the leading term helps to fill in smoothly between data points. Where data are plentiful, the higher-order terms accommodate deviations from the leading term. The method was first proposed by Gallant and Tauchen (1989) in connection with an asset pricing application. A C++ implementation of SNP is at http://econ.duke.edu/webfiles/arg/snp/, together with a User's Guide, which is an excellent tutorial introduction to the method.

Important adjuncts to SNP estimation are a rejection method for simulating from the SNP density developed in Gallant and Tauchen (1992), which can be used, for example, to set bootstrapped confidence intervals as in Gallant, Rossi and Tauchen (1992); nonlinear error shock analysis as described in Gallant, Rossi and Tauchen (1993), which develops the nonlinear analog of conventional error shock analysis for linear VAR models; and re-projection, which is a form of nonlinear Kalman filtering that can be used to forecast the unobservables of nonlinear latent variables models (Gallant and Tauchen, 1998).

As stated above, the SNP method is based on the notion that a Hermite expansion can be used as a general purpose approximation to a density function. Letting z denote an M–

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

vector, we can write the Hermite density as $h(z) \propto [\mathcal{P}(z)]^2 \varphi(z)$ where $\mathcal{P}(z)$ denotes a multivariate polynomial of degree K_z and $\phi(z)$ denotes the density function of the (multivariate) Gaussian distribution with mean zero and variance the identity matrix. Denote the coefficients of $\mathcal{P}(z)$ by a, which is a vector whose length depends on K_z and M. When we wish to call attention to the coefficients, we write $\mathcal{P}(z|a)$.

The constant of proportionality is $1 / \int [\mathcal{P}(s)]^2 \varphi(s) ds$ which makes $h(z)$ integrate to one. As seen from the expression that results, namely

$$h(z) = \frac{[\mathcal{P}(z)]^2 \varphi(z)}{\int [\mathcal{P}(s)]^2 \varphi(s) ds},$$

we are effectively expanding the square root of the density in Hermite functions of the form $\mathcal{P}(z) \sqrt{\varphi(z)}$. Because the square root of a density is always square integrable and because the Hermite functions of the form $\mathcal{P}(z) \sqrt{\varphi(z)}$ are dense for the collection of square integrable functions (Fenton and Gallant, 1996), every density has such an expansion.

Because $[\mathcal{P}(z)]^2 / \int [\mathcal{P}(s)]^2 \varphi(s) ds$ is a homogeneous function of the coefficients of the polynomial $\mathcal{P}(z)$, the coefficients can only be determined to within a scalar multiple. To achieve a unique representation, the constant term of the polynomial part is put to 1. Customarily the Hermite density is written with its terms orthogonalized and the C++ code is written in the orthogonalized form for numerical efficiency. But reflecting that here would lead to cluttered notation and add nothing to the ideas.

A change of variables using the location-scale transformation $y = Rz + \mu$, where R is an upper triangular matrix and μ is an M-vector, gives

$$f(y|\theta) \propto \left\{\mathcal{P}[R^{-1}(y-\mu)]\right\}^2 \left\{\varphi[R^{-1}(y-\mu)] / |\det(R)|\right\}$$

The constant of proportionality is the same as above, $1 / \int [\mathcal{P}(s)]^2 \varphi(s) ds$. Because $\left\{\varphi[R^{-1}(y-\mu)] / |\det(R)|\right\}$ is the density function of the M-dimensional, multivariate, Gaussian distribution with mean μ and variance-covariance matrix $\Sigma = RR'$, and because the leading term of the polynomial part is 1, the leading term of the entire expansion is proportional to the multivariate, Gaussian density function. Denote the Gaussian density of dimension M with mean vector μ and variance matrix Σ by $n_M(y|\mu, \Sigma)$ and write

$$f(y|\theta) \propto [\mathcal{P}(z)]^2 n_M(y|\mu, \Sigma)$$

where $z = R^{-1}(y-\mu)$ for the density above.

When K_z is put to zero, one gets $f(y|\theta) = n_M(y|\mu, \Sigma)$ exactly. When K_z is positive, one gets a Gaussian density whose shape is modified due to multiplication by a polynomial in $z = R^{-1}(y-\mu)$. The shape modifications thus achieved are rich enough to accurately approximate densities from a large class that includes densities with fat, t-like tails, densities with tails that are thinner than Gaussian, and skewed densities (Gallant and Nychka, 1987).

The parameters θ of $f(y|\theta)$ are made up of the coefficients a of the polynomial $\mathcal{P}(z)$ plus μ and R and are estimated by maximum likelihood which is accomplished by minimizing $s_n(\theta) = (-1/n) \Sigma_{t=1}^{n} \log[f(y_t|\theta)]$. As mentioned above, if the number of parameters p_θ grows with the sample size n, the true density and various features of it such as derivatives and moments are estimated consistently (Gallant and Nychka, 1987).

This basic approach can be adapted to the estimation of the conditional density of a multiple time series $\{y_t\}$ that has a Markovian structure. Here, the term ‘Markovian structure’ is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

taken to mean that the conditional density of the M–vector y_t given the entire past $y_{t-1}, y_{t-2}, \cdots$ depends only on L lags from the past. For convenience, we will presume that the data are from a process with a Markovian structure, but one should be aware that, if L is sufficiently large, then non-Markovian data can be well approximated by an SNP density (Gallant and Long, 1997). Collect these lags together as $x_{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_{t-L})$, where L exceeds all lags in the following discussion.

To approximate the conditional density of $\{y_t\}$ using the ideas above, begin with a sequence of innovations $\{z_t\}$. First consider the case of homogeneous innovations; that is, the distribution of z_t does not depend on x_{t-1}. Then, as above, the density of z_t can be approximated by $h(z) \propto [\mathcal{P}(z)]^2 \varphi(z)$ where $\mathcal{P}(z)$ is a polynomial of degree K_z. Follow with the location-scale transformation $y_t = Rz_t + \mu_x$ where μ_x is a linear function that depends on L_u lags

$$\mu_x = b_0 + Bx_{t-1}.$$

(If L_u<L, then some elements of B are zero.) The density that results is

$$f(y|x, \theta) \propto [\mathcal{P}(z)]^2 n_M(y|\mu_x, \Sigma)$$

where $z = R^{-1}(y - \mu_x)$. The constant of proportionality is as above, $1 / \int [\mathcal{P}(s)]^2 \varphi(s) ds$. The leading term of the expansion is $n_M(y|\mu_x, \Sigma)$ which is a Gaussian vector autoregression or Gaussian VAR. When K_z is put to zero, one gets $n_M(y|\mu_x, \Sigma)$ exactly. When Kz is positive, one gets a semiparametric VAR density.

To approximate conditionally heterogeneous processes, proceed as above but let each coefficient of the polynomial $\mathcal{P}(z)$ be a polynomial of degree K_x in x. A polynomial in z of degree K_z whose coefficients are polynomials of degree K_x in x is, of course, a polynomial in (z, x) of degree $K_z + K_x$. Denote this polynomial by $\mathcal{P}(z, x)$. Denote the mapping from x to the coefficients a of $\mathcal{P}(z)$ such that $\mathcal{P}(z|a_x) = \mathcal{P}(z, x)$ by a_x and the number of lags on which it depends by L_p. The form of the density with this modification is

$$f(y|x, \theta) \propto [\mathcal{P}(z, x)]^2 n_M(y|\mu_x, \Sigma)$$

where $z = R^{-1}(y - \mu_x)$. The constant of proportionality is $1 / \int [\mathcal{P}(s, x)]^2 \varphi(s) ds$. When K_x is zero, the density reverts to the density above. When K_x is positive, the shape of the density will depend upon x. Thus, all moments can depend upon x and the density can, in principal, approximate any form of conditional heterogeneity (Gallant and Tauchen, 1989). In practice the second moment can exhibit marked dependence upon x. In an attempt to track the second moment, K_x can get quite large. To keep K_x small when data are markedly conditionally heteroskedastic, the leading term $n_M(y|\mu_x, \Sigma)$ of the expansion can be put to a Gaussian GARCH rather than a Gaussian VAR. SNP uses a modified BEKK expression as described in Engle and Kroner (1995); the modifications are to add leverage and level effects.

$$\Sigma_{x_{t-1}} = R_0 R_0' + \sum_{i=1}^{L_g} Q_i \Sigma_{x_{t-1-i}} Q_i' + \sum_{i=1}^{L_r} P_i (y_{t-i} - \mu_{x_{t-1-i}})(y_{t-i} - \mu_{x_{t-1-i}})' P_i' + \sum_{i=1}^{L_v} \max[0, V_i(y_{t-i} - \mu_{x_{t-1-i}})] \max[0, V_i(y_{t-i} - \mu_{x_{t-1-i}})]' + \sum_{i=1}^{L_w} W_i x_{(1),t-1} x_{(1),t-i}' W_i'.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Above, R_0 is an upper triangular matrix. The matrices P_i, Q_i, V_i, and W_i can be scalar, diagonal, or full M by M matrices. The notation $x_{(1),t-i}$ indicates that only the first column of x_{t-i} enters the computation. The max(0, x) function is applied elementwise. Because $\Sigma_{x_{t-1}}$ must be differentiable with respect to the parameters of $\mu_{x_{t-2-i}}$, the max(0,x) function is approximated by a twice continuously differentiable cubic spline. Defining $R_{x_{t-1}}$ by the factorization $\Sigma_{x_{t-1}} = R_{x_{t-1}} R'_{x_{t-1}}$ and writing x for x_{t-1}, the SNP density becomes

$$f(y|x, \theta) \propto [p(z, x)]^2 n_M(y|\mu_x, \Sigma_x)$$

where $z = R_x^{-1}(y - \mu_x)$. The constant of proportionality is $1 / \int [p(s, x)]^2 \varphi(s) ds$. The leading term $n_M(y|\mu_x, \Sigma_x)$ is Gaussian ARCH if L_g=0 and $L_r > 0$ and Gaussian GARCH if both $L_g > 0$ and $L_r > 0$ (leaving aside the implications of L_v and L_w).

See Also

- ARCH models
- computational methods in econometrics
- impulse response function
- nonlinear time series analysis
- nonparametric structural models

Research for this article was supported by the National Science Foundation.

Bibliography

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–27.

Engle, R.F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

Engle, R.F. and Kroner, K.F. 1995. Multivariate simultaneous generalized ARCH. *Econometric Theory* 11, 122–50.

Fenton, V.M. and Gallant, A.R. 1996. Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics* 74, 77–118.

Gallant, A.R., Hsieh, D. and Tauchen, G. 1997. Estimation of stochastic volatility models with diagnostics. *Journal of Econometrics* 81, 159–92.

Gallant, A.R. and Long, J.R. 1997. Estimating stochastic differential equations efficiently by minimum chi-square. *Biometrika* 84, 125–41.

Gallant, A.R. and Nychka, D.W. 1987. Seminonparametric maximum likelihood estimation. *Econometrica* 55, 363–90.

Gallant, A.R., Rossi, P.E and Tauchen, G. 1992. Stock prices and volume. *Review of Financial Studies* 5, 199–242.

Gallant, A.R., Rossi, P.E. and Tauchen, G. 1993. Nonlinear dynamic structures. *Econometrica* 61, 871–907.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gallant, A.R. and Tauchen, G. 1989. Seminonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications. *Econometrica* 57, 1091–120.

Gallant, A.R. and Tauchen, G. 1992. A nonparametric approach to nonlinear time series analysis: estimation and simulation. In *New Directions in Time Series Analysis, Part II*., ed. D. Brillinger et al. New York: Springer-Verlag.

Gallant, A.R. and Tauchen, G. 1996. Which moments to match? *Econometric Theory* 12, 657–81.

Gallant, A.R. and Tauchen, G. 1998. Reprojecting partially observed systems with application to interest rate diffusions. *Journal of the American Statistical Association* 93, 10–24.

How to cite this article

Gallant, A. Ronald. "SNP: nonparametric time series analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000468> doi:10.1057/9780230226203.1548

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social capital

Partha Dasgupta
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social capital is an aggregate of interpersonal networks. Belonging to a network helps a person to coordinate his strategies with others. Where the state or the market is dysfunctional, communities enable people to survive, even if they do not enable them to live well. But communities often involve hierarchical social structures; and the theory of repeated games cautions us that communitarian relationships can involve allocations where some of the parties are worse off than they would have been if they had not been locked into the relationships. Even if no overt coercion is visible, such relationships could be exploitative.

Keywords

caste system; civil society; common property resources; communitarian institutions; contract enforcement; cooperation; exploitation; human capital; interpersonal networks; Prisoner's Dilemma; public goods; reciprocity; repeated games; reputation; rotating savings and credit associations; social capital; social norms; total factor productivity; trust

Article

Definitions?

The idea of *social capital* sits awkwardly in contemporary economic thinking. Although it has a powerful, intuitive appeal, the object has proven hard to track as an economic good. One can argue (Arrow, 2000) that it is misleading to use the term ‘capital’ to refer to whatever it is that ‘social capital’ happens to be, because capital is usually identified with tangible, durable and alienable objects (for example, buildings and machines), whose accumulation can be estimated and whose worth can be assessed. There is much to agree with this. But in regard to both heterogeneity and intangibility, social capital would seem to resemble knowledge and skills. So one can also argue that, since economists have

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
not shied away from regarding knowledge and skills as forms of capital, we should not shy away in this case either.

In an early definition, social capital was identified with those ‘features of social organization, such as trust, norms, and networks that can improve the efficiency of society by facilitating coordinated actions’ (Putnam, Leonardi and Nanetti, 1993, p. 167). The characterization suffers from a weakness: it encourages us to amalgamate strikingly different objects, namely (and in that order), beliefs, behavioural rules, and such forms of capital assets as interpersonal links (or ‘networks’), without establishing reasons why such an inclusive definition would prove useful in gaining an understanding of our social world. Subsequently, Putnam (2000, p. 19) suggested a redefinition: ‘social capital refers to connections among individuals – social networks and the norms of reciprocity and trustworthiness that arise from them.’ Since then authors have defined social capital even more inclusively, where attitudes towards others make their appearance as well: ‘Social capital generally refers to trust, concern for one's associates, a willingness to live by the norms of one's community and to punish those who do not’ (Bowles and Gintis, 2002, p. F419).

These definitions tell us that ‘social capital’ is an ingredient in the workings of *civil society* (Putnam, 1993; 2000). In a parallel development, the theory and empirics of common-property resources in poor countries (for example, coastal fisheries, village tanks, local forests, pasture lands, and threshing grounds) have revealed the character of those local institutions that enable mutually beneficial courses of action to be undertaken within communities (Dasgupta and Heal, 1979; Jodha, 1986; Ostrom, 1990; Dasgupta and Mäler, 1991; Bromley, 1992; Baland and Platteau, 1996). Development economists have also studied rotating savings and credit associations, irrigation management systems, and credit and insurance arrangements in poor countries (Ostrom, 1990; Udry, 1990; Besley, Coate and Loury, 1992; Grootaert and van Bastelaer, 2002). These studies suggest that social capital is a measure of the worth of *communitarian institutions*.

Where the state is weak or indifferent or rapacious, where markets do not work well or are even non-existent, communities enable people to survive, even if they do not enable them to live well. That may be why scholars writing on social capital have frequently imbued the notion with a warm glow. But there is a dark side to communities, often involving hierarchical social structures (for example, the Hindu caste system), rent-seeking groups, the Mafia, and street gangs. Omniously, the theory of repeated games (Fudenberg and Maskin, 1986) cautions us that communitarian relationships can involve allocations where some of the parties are *worse off* than they would have been if they had not been locked into those relationships; even though no overt coercion is visible, such relationships can be exploitative (see Dasgupta, 2000; 2005).

Why do people keep promises?

In order not to prejudge the character of communities, it is best not to worry about defining social capital, but to ask instead a fundamental question facing *any* group of people who have agreed on a joint course of action: *under what contexts can the members trust one another to try to carry out their terms of the agreement*?

Four points come to mind:

(i) *Mutual affection*. Consider the situation where the people involved care about one another. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
household is the most obvious example of an institution based on affection. To break a promise we have made to someone we care about is to feel bad. So we try not to do so.

(ii) *Pro-social disposition*. Another situation is where people are trustworthy, or where they reciprocate if others have behaved well towards them. Evolutionary psychologists have suggested that we are adapted to have a general disposition to reciprocate. Development psychologists have found that pro-social disposition can be formed by communal living, role modelling, education, and receiving rewards and punishments (be it here or in the afterlife).

We do not have to choose between the two viewpoints; they are not mutually exclusive. Our capacity to have such feelings as shame, guilt, fear, affection, anger, elation, reciprocity, benevolence, jealousy, and our sense of fairness and justice have emerged under selection pressure. Culture helps to shape preferences, expectations, and our notion of what constitutes fairness. Such notions in turn influence behaviour, which is known to differ among societies. But cultural coordinates enable us to identify the situations *in* which shame, guilt, fear, affection, anger, elation, reciprocity, benevolence, and jealousy arise; they do not displace the centrality of those feelings in the human make-up. By internalizing norms of behaviour, people enable the springs of their actions to include them. In short, they have a disposition to obey the norm, be it personal or social. When they do violate a norm, neither guilt nor shame would typically be absent, but frequently they will have rationalized the act. Making a promise is a commitment for such people; and it is essential for them that others recognize it to be so.

People are trustworthy to varying degrees. So, although pro-social disposition is not foreign to human nature, no society could rely on it exclusively; for how is one to tell to what extent someone is trustworthy?

Societies everywhere have therefore tried to establish institutions where people have the incentive to do business with one another. The incentives differ in their details, but they have one thing in common: *those who break agreements without acceptable cause are punished*. Broadly speaking, there are two ways in which the right incentives are created.

(iii) *External enforcement*. It could be that the agreement is translated into an explicit contract and enforced by an established structure of power and authority; that is, an *external enforcer*.

By an external enforcer we imagine here, for simplicity, the state. (There can, of course, be other external enforcement agencies; for example, tribal chieftains, warlords and so forth.) Consider that the rules governing transactions in the formal marketplace involve legal contracts backed by an external enforcer, namely, the state. So it is because you and the supermarket owner are confident that the state has the ability and willingness to enforce contracts that you and the owner of the supermarket are willing to transact when you go there to purchase goods.

What is the basis of that confidence? Simply to invoke an external enforcer for solving the credibility problem will not do; for why should the parties trust the state to carry out its tasks in an honest manner? A possible answer is that the government worries about its reputation. So, for example, a free and inquisitive press in a democracy, aided by a demanding civil society, helps to sober the government into believing that incompetence or malfeasance would mean an end to its rule, come the next election. Knowing that the government worries, the parties trust it to enforce agreements.

The above argument involves a system of interlocking beliefs about one another's abilities and intentions, one that supports an equilibrium in which the agreement is kept. Unfortunately, non-cooperation can also be held together by its own bootstraps. At a non-cooperative equilibrium the parties do not trust one another to keep their promises, because the external enforcer cannot be trusted to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
enforce agreements. To ask whether cooperation or non-cooperation would prevail is to ask which system of beliefs has been adopted by the parties about one another's intentions. Social systems have multiple equilibria.

(iv) *Mutual enforcement in long-term relationships*. Suppose the group of parties in question expect to face similar transaction opportunities in each period over an indefinite future. Imagine, too, that the parties cannot depend on the law of contracts because the nearest courts are far from their residence. There may even be no lawyers in sight. In rural parts of sub-Saharan Africa, for example, much economic life is shaped outside a formal legal system. But even though no external enforcer may be available, people there do transact. Credit involves saying, ‘I lend to you now with your promise that you will repay me’; and so on. But why should the parties be sanguine that the agreements will not turn sour on account of malfeasance?

They would be sanguine if agreements were mutually enforced. The basic idea is this: *a credible threat by members of a community that stiff sanctions would be imposed on anyone who broke an agreement could deter everyone from breaking it*. The problem then is to make the threat credible. As the theory of repeated games has shown, the solution to the credibility problem in this case is achieved by recourse to social norms of behaviour.

By a *social norm* we mean a rule of behaviour (or strategy) that is followed by members of a community. For a rule of behaviour to be a social norm, it must be in the interest of everyone to act in accordance with the rule if all others were to act in accordance with it. Social norms are equilibrium rules of behaviour. The theory of repeated games has shown that, if people discount the future benefits from cooperation at a low enough rate, there are social norms that support cooperation.

As with the case of external enforcement, even when cooperation is a possible equilibrium under mutual enforcement, non-cooperation is an equilibrium too. If each party were to believe that all others would break the agreement from the start, each party would break the agreement at that stage. Failure to cooperate could be due simply to a collection of unfortunate, self-confirming beliefs.

Social capital as interpersonal networks

In common parlance, we reserve the term ‘society’ to denote a collective that has managed to equilibrate at a mutually beneficial outcome. Underlying each of the four contexts I have alluded to in which people trust one another to cooperate is a system of mutual beliefs. Because such a system of beliefs is likely to arise only if the parties know one another (at least indirectly), I believe it is best to regard *social capital* as *interpersonal networks*. The advantage of such a lean notion is that it does not prejudge the asset's quality. Just as a building can remain unused and a wetland can be misused, so a network can remain inactive or be put to use in socially destructive ways. There is nothing good or bad about interpersonal networks: other things being equal, it is the *use* to which a network is put by members that determines its quality.

Interpersonal networks are systems of communication channels linking people to one another. Networks include as tightly woven a unit as a nuclear family or a kinship group, and one as extensive as a voluntary organization, such as Amnesty International. We are born into certain networks and enter new ones. Personal relationships, whether or not they are long term, are emergent features within networks, and involve systems of mutual beliefs. For example, Seabright (1997) has suggested that civic engagements and communal activities heighten the disposition to cooperate (context (ii) above). The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
idea is that trust begets trust and that this gives rise to a positive feedback between civic and communal activities and a disposition to be so engaged. That positive feedback is, however, tempered by the cost of additional engagements (time), which, typically, rises with increasing engagements.

Networks and human capital

How did people who now interact with one another get to connect in the first place? In village economies in poor countries the answer is simple: mostly they have known one another from birth. People engaged in long-term relationships based on social norms (contexts (ii) and (iv) above) – or *communities*, for short – have to know one another, at least indirectly through people they know personally. Communities are *personal* and *exclusive*. Members have names, personalities and attributes. An outsider's word is not as good as an insider's. Markets, in contrast, are impersonal and inclusive.
In his pioneering work, Coleman (1988) saw social capital as an *input* in the production of human capital. In modern, mobile societies, people have to invest resources trying to meet people. Some of the investment is pleasurable, some not. Even so, just as academics are paid for what they mostly like doing anyway (as a return on investment in their education), networking would be expected to pay dividends even when maintaining networks is a pleasurable activity.
Burt (1992) has found among business firms in the United States, controlling for age, education and experience, that employees enjoying strategic positions in networks are more highly compensated than those who are not. His findings confirm that some of the returns from investment in network creation are captured by the investor. However, because of network externalities, not all the returns can be captured by the investor: when *A* and *B* establish a channel linking them, the investment improves both *A*'s and *B*'s earnings, but it also improves the earnings of *C*, who was already connected to *B*. The findings of Burt and his colleagues imply that membership in networks is a component of someone's ‘human capital’. If firms pay employees on the basis of what they contribute to profitability, they would look not only at the conventional human capital employees bring with them (for example, health, education, experience, personality), but also the personal contacts they possess. It would be informative to untangle networks from the rest of human capital. This could reveal the extent to which returns from network investment are captured by the investor. But measurement problems abound. They may be insurmountable because of the pervasive externalities to which they give rise.

Micro-behaviour and macro-performance

How do network activities translate into the macro-performance of economies?
The discussion in the previous section implied that to the extent that the worth of contacts is reflected in wages and salaries, social capital is a component of human capital. It should be noted though that in poor countries, where labour markets can malfunction badly, or can even be non-existent, attributing returns to the various factors of production is especially problematic. But even if we were to leave that problem aside, we know that networks give rise to externalities. This makes the translation from micro-behaviour to macro-performance an especially difficult subject.
To illustrate, consider a simple formulation of economy-wide production possibilities. Let individuals be indexed by j (j=1, 2, …). Let K denote the economy's stock of physical capital and L_j the labour-hours

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
put in by person *j*. I do not specify the prevailing system of property rights to physical capital, nor do I describe labour relations, because to do so would be to beg the questions being discussed here. But it is as well to keep in mind that in a well-developed market economy *K* would be dispersed private property, in others *K* would be in great measure publicly owned, in yet others much would be communally owned, and so forth. It is also worth remembering that in market economies labour is wage based, that in subsistence economies ‘family labour’ best approximates the character of labour relations, and that labour cooperatives are not unknown in certain parts of the world; and so on.

Let h_j be the human capital of person *j* (years of schooling, health). His or her effective labour input is then h_jL_j. h_j is what one may call ‘traditional human capital’ (we leave aside the networks to which *j* belongs). Human capital is embodied in workers. Given the economy's knowledge base and institutions (the latter I take here to be the engagements brought about by the interpersonal networks), human capital in conjunction with physical capital produces an all-purpose output, *Y*, which we may call gross national product (GNP). Each of the aggregate indices requires for its construction prices for the multitude of components that make up the aggregate. In industrial market economies, the required prices are typically market prices. When externalities are pervasive, the construction of such indices poses special problems. Let us therefore assume away problems of aggregation by imagining the economy to possess a single good, *Y*. Problems nevertheless remain in measuring the pathways that link micro-behaviour to macro-performance. Let us study them.

Total factor productivity

Write $H=_j(h_jL_j)$. *H* is aggregate human capital. Now suppose that output possibilities are given by the relationship

$$Y = AF(K, H), \ (A > 0), \tag{1}$$

where *F* is the economy's aggregate production function. *F* is non-negative and is assumed to be an increasing function of both *K* and *H*.

In eq. (1) *A* is *total factor productivity*. It is a combined index of institutional capabilities (including the prevailing system of property rights) and publicly shared knowledge. A macroeconomy characterized by the production function *F* would produce more, other things being equal, if *A* were larger (that is, if publicly shared knowledge was greater or institutional capabilities higher). Of course, the economy would also produce more, other things being equal, if *K* or h_j or L_j were larger. In short, technological possibilities for transforming the services of physical and human capital into output, when embedded in the prevailing institutional structure of the economy, account for eq. (1).

Consider now a scenario where civic cooperation increases in the community: the economy moves from a bad equilibrium system of mutual beliefs to a good one. The increase would make possible a more efficient allocation of resources in production. A question arises: would the increase in cooperation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
appear as a heightened value of A, or would it appear as an increase in H, or as increases in both?
The answer lies in the extent to which network externalities are like public goods. If the externalities are confined to small groups (that is, small groups are capable of undertaking cooperative actions on their own – with little effect on others – and take such actions in the good equilibrium), the improvements in question would be reflected mainly through the h_j's of those in the groups engaged in increased cooperation. On the other hand, if the externalities are economy-wide (as in the case of an increase in quasi-voluntary compliance in the economy as a whole owing to an altered set of beliefs, even about members of society one does not personally know), the improvements would be reflected mainly through A. Either way, the directional changes in macro-performance (though not the magnitude of the changes) would be the same. Other things being equal, an increase in A or in some of the h_j's (brought about by whichever of the mechanisms we have considered) would mean an increase in GNP, an increase in wages, salaries and profits, and possibly an increase in investment in both physical and human capital. The latter would result in faster rate of growth in output and consumption, and, if a constant proportion of income were spent on health, a more rapid improvement in health as well.

Interpreting cross-section findings

It will be useful to connect the macroeconomic account to the findings from less aggregated data. In his analysis of statistics from the 20 administrative regions of Italy, Putnam (1993) found civic tradition to be a strong predictor of contemporary economic indicators. He showed that indices of civic engagement in the early years of the 20th century were highly correlated with employment, income and infant survival in the early 1970s. Putnam also found that regional differences in civic engagement can be traced back several centuries and that, controlling for civic traditions, indices of industrialization and public health have no impact on current civic engagement. As he put it, the causal link appears to be from civics to economics, not the other way round. How do his findings square with the formulation in eq. (1)?
The same sort of question can be asked of even less aggregated data. Narayan and Pritchett (1999) have analysed statistics on household expenditure and social engagements in a sample of some 50 villages in Tanzania, to discover that households in villages where there is greater participation in village-level social organizations on average enjoy greater income per head. The authors have also provided statistical reasons for concluding that greater communitarian engagements result in higher household expenditure rather than the other way round.
To analyse these findings in terms of our macroeconomic formulation, consider two autarkic communities, labelled by i (=1, 2). I simplify by assuming that members of a community are identical. Denote the human capital per person in community i by h_i. By h_i I now mean not only the traditional forms of human capital (health and education), but also network capital. Let L_i denote the number of hours worked by someone in i, by N_i the size of i's population, and by K_i the total stock of physical assets in i. Aggregate output, Y_i, is

$$Y_i = A_i F(K_i, N_i h_i L_i). \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Improvements in civic cooperation are reflected in increases in A, or h, or both. It follows that if civic cooperation were greater among people in community 1 than in community 2, we would have $A_1>A_2$, or $h_1>h_2$, or both. Imagine now that the two communities have the same population size, possess identical amounts of physical capital, and work the same number of hours. GNP in community 1 would be greater than GNP in community 2 (that is, $Y_1>Y_2$). More generally, an observer would discover that, controlling for differences in K and L, there is a positive association between a community's cooperative culture (be it total factor productivity, A_i, or human capital, h_i) and its mean household income (Y_i/N). This is one way to interpret the finding reported in Narayan and Pritchett (1999).

Consider now a different thought-experiment. Imagine that in year 1900 the two communities had been identical in all respects but for their cooperative culture, of which community 1 had more (that is, in 1900, $A_1>A_2$, or $h_1>h_2$, or both). Imagine next that, since 1900, both A_i and h_i have remained constant. Suppose next that people in both places have followed a simple saving rule: a constant fraction s_K (>0) of aggregate output has been invested each year in accumulating physical capital. (For the moment I imagine that net investment in human capital in both communities is nil.) In order to make the comparison between the communities simple, imagine finally that the communities have remained identical in their demographic features. It is then obvious that in year 1970 community 1 would be richer than community 2 in terms of output, wages and salaries, profits, consumption and wealth. This is one way to interpret Putnam's (1993) findings.

Notice that we have not had to invoke possible *increases* in total factor productivity (A_i) or human capital (h_i) to explain why a cooperative culture is beneficial. Total factor productivity and human capital have done all the work in our analysis of the empirical finding: we have not had to invoke secular improvements in them to explain why a more cooperative society would be expected to perform better economically.

As the communities in our thought experiment are both autarkic, there is no flow of physical capital from one to the other. This is an economic distortion for the combined communities: the rates of return on investment in physical capital in the two places remain unequal. The source of the distortion is the enclave nature of the two communities, occasioned in our example by an absence of markets linking them. There would be gains to be enjoyed if physical capital could flow from community 2 to community 1.

Autarky is an extreme assumption, but it is not a misleading assumption. What the model points to is that, to the extent that social capital is exclusive, it inhibits the flow of resources, in this case a movement of physical capital from one place to the other. Put another way, if markets fail to function well, capital does not move from community 2 to community 1 to the extent it ideally should. When social networks within each community block the growth of markets, their presence inhibits economic progress.

Dark matters

Two potential weaknesses of resource allocation mechanisms built on social capital are easy enough to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
identify.
Exclusivity. Networks are exclusive, not inclusive. This means that ‘anonymity’, the hallmark of competitive markets, is absent from the operations of networks. When market enthusiasts proclaim that one person's money is as good as any other person's in the market place, it is anonymity they invoke. In allocation mechanisms governed by networks, however, ‘names’ matter. Transactions are personalized. This implies inefficiencies: resources are unable to move to their most productive uses.
Inequalities. The benefits of cooperation are frequently captured by the more powerful within the network. McKean (1992), among others, has discovered that the local elite (usually wealthier households) capture a disproportionate share of the benefits of common-property resources, such as coastal fisheries and forest products. Her finding is consistent with the possibility that all who cooperate benefit.

Exploitation within networks

The reason why social capital continues to radiate a warm glow in the literature is that the subject has been motivated by examples of the Prisoner's Dilemma. However, one-period games involving the use of common property resources don't give rise to the Prisoner's Dilemma (Dasgupta, 2005). Consider an indefinitely repeated game among N players, in which (*a*) the stage game possesses a unique Nash equilibrium and (*b*) the ‘min-max’ payoffs of the players are lower than their respective payoffs in the Nash equilibrium. As is well known (Fudenberg and Maskin, 1986), if the players discount their future payoffs at a low enough rate, there are social norms that can sustain an outcome where the time-average of the per-period payoff to a player is *less than* the payoff at the unique Nash equilibrium. That player would be worse off in a long-term relationship with the others than if the players were not in a long-term relationship. The social norm sustaining that outcome would be *exploitative* of the player.
Inequality is not the same as exploitation, which is why to demonstrate exploitation in an empirically satisfactory way will prove to be very hard: any such demonstration would involve comparison of an observable state of affairs with a counterfactual. However, some stark examples are suggestive. In Indian villages, access to local common-property resources is often restricted to the privileged (for example, caste Hindus), who are also among the more prosperous landowners. The outcasts (euphemistically called members of ‘schedule castes’) are among the poorest of the poor. Stark inequities exist, too, in patron–client relationships in agrarian societies, which make it very likely that the ‘client’ is worse off in consequence of that relationship than without it. Ogilvie (2003) has unearthed striking differences between the life chances of women in 17th-century Germany (embedded in dense networks) and the life chances of women in 17th-century England (not so embedded in dense networks): English women were better off.

Morals

Social capital is an aggregate of interpersonal networks. From the economic point of view, belonging to a network helps a person to coordinate his strategies with others. We should not prejudge the character of the strategies on which members of a network coordinate. As with any other form of capital asset, social capital can be put to good use or bad.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- common property resources
- cooperation
- repeated games
- social networks, economic relevance of
- social norms

Bibliography

Arrow, K.J. 2000. Observations on social capital. In *Social Capital: A Multifaceted Perspective*, ed. P. Dasgupta and I. Serageldin. Washington, DC: World Bank.

Baland, J.-M. and Platteau, J.-P. 1996. *Halting Degradation of Natural Resources: Is There a Role for Rural Communities?* Oxford: Clarendon Press.

Besley, T., Coate, S. and Loury, G. 1992. The economics of rotating savings and credit associations. *American Economic Review* 82, 782–810.

Bowles, S. and Gintis, H. 2002. Social capital and community governance. *Economic Journal* 112, 419–36.

Bromley, D.W. 1992. *Making the Commons Work: Theory, Practice and Policy*. San Francisco: Institute for Contemporary Studies.

Burt, R.S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Coleman, J.S. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 94, 95–120.

Dasgupta, P. 2000. Economic progress and the idea of social capital. In *Social Capital: A Multifacted Perspective*, ed. P. Dasgupta and I. Serageldin. Washington, DC: World Bank.

Dasgupta, P. 2005. Common property resources: economic analytics. *Economic and Political Weekly* 40, 1610–22.

Dasgupta, P. and Heal, G. 1979. *Economic Theory and Exhaustible Resources*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dasgupta, P. and Mäler, K.-G. 1991. The environment and emerging development issues. *Proceedings of the Annual World Bank Conference on Development Economics, 1990*. Washington, DC: World Bank.

Fudenberg, D. and Maskin, E. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54, 533–56.

Grootaert, C. and van Bastelaer, T. 2002. *The Role of Social Capital in Development: An Empirical Assessment*. Cambridge: Cambridge University Press.

Jodha, N.S. 1986. Common property resources and the rural poor. *Economic and Political Weekly* 21, 1169–81.

McKean, M. 1992. Success on the commons: a comparative examination of institutions for common property resource management. *Journal of Theoretical Politics* 4, 256–68.

Narayan, D. and Pritchett, L. 1999. Cents and sociability: household income and social capital in rural Tanzania. *Economic Development and Cultural Change* 47, 871–89.

Ogilvie, S. 2003. *A Bitter Living: Women, Markets, and Social Capital in Early Modern Germany*. Oxford: Oxford University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Putnam, R.D., Leonardi, R. and Nanetti, R.Y. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, NJ: Princeton University Press.

Putnam, R.D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Seabright, P. 1997. Is cooperation habit-forming? In *The Environment and Emerging Development Issues*, vol. 2, ed. P. Dasgupta and K.-G. Mäler. Oxford: Clarendon Press.

Udry, C. 1990. Credit markets in northern Nigeria: credit as insurance in a rural economy. *World Bank Economic Review* 4, 251–69.

How to cite this article

Dasgupta, Partha. "social capital." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
article?id=pde2008_S000451> doi:10.1057/9780230226203.1549

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social choice (new developments)

Walter Bossert and John A. Weymark
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since the late 1970s there has been a dramatic shift of focus in social choice theory, with structured sets of alternatives and restricted domains of the sort encountered in economic problems coming to the fore. This article provides an overview of some of the recent contributions to four topics in normative social choice theory in which economic modelling has played a prominent role: Arrovian social choice theory on economic domains, variable-population social choice, strategy-proof social choice, and axiomatic models of resource allocation.

Keywords

allotment problem; Arrow's theorem; average utilitarianism; axiomatic models of resource allocation; Bergson–Samuelson social welfare function; classical utilitarianism; critical-level utilitarianism; decomposability; egalitarian equivalence; expected utility; fair allocation; Gibbard–Satterthwaite theorem; impossibility theorem; independence of irrelevant alternatives; interpersonal utility comparisons; majority rule; monotonicity; nondictatorship; option set; Pareto efficiency; population ethics; single-peaked preferences; social choice; social welfare function; strategy-proofness; uniform rule; utilitarianism; Walrasian equilibrium; welfarism

Article

With the exception of the research on single-peaked preferences and their multidimensional generalizations, for the most part the early literature on social choice theory dealt with abstract sets of alternatives and domains of preferences and feasible sets that exhibited little structure. Since the late 1970s there has been a dramatic shift of focus, with structured sets of alternatives and restricted domains coming to the fore. In particular, a great deal of attention has been directed towards the kinds of concrete problems that arise in economics, with alternatives being allocations of goods and preferences and feasible sets satisfying the sorts of restrictions encountered in economic models. In this article, we

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

provide an overview of some of the recent contributions to four topics in normative social choice theory in which economic modelling has played a prominent role: Arrovian social choice theory on economic domains, variable-population social choice, strategy-proof social choice, and axiomatic models of resource allocation. Structured environments have also been considered in positive social choice theory, notably in the political economy literature. See Austen-Smith and Banks (2005) for an introduction to this literature. Other areas of social choice theory have also been active in recent years: see Arrow, Sen and Suzumura (2002; 2008) for recent surveys of these topics.

Arrovian social choice on economic domains

Arrow's theorem (see Arrow, 1963) is concerned with the aggregation of profiles of individual preference orderings into a social ordering of a set of alternatives X. Let $\mathfrak{R}$ denote the set of all orderings of X. In Arrow's theorem, there is a finite set of individuals $N=\{1, \ldots, n\}$ with $n \geq 2$, each of whom has a weak preference ordering R_i on X. An (*Arrovian*) *social welfare* function f assigns a social ordering $R = f(\mathbf{R})$ of X to each profile $\mathbf{R}=(R_1, \ldots, R_n)$ of individual preference orderings in some domain $\mathcal{D}$ of profiles. Arrow's theorem demonstrates that it is impossible for a social welfare function to satisfy independence of irrelevant alternatives, henceforth IIA (the social ranking of a pair of alternatives depends only on the individual rankings of these alternatives), weak Pareto (if everyone strictly prefers one alternative to a second, then so does society), and nondictatorship (nobody's strict preferences are always respected) if the domain is unrestricted ($\mathcal{D} = \mathfrak{R}^n$) and $|X| \geq 3$.

Arrow's theorem is not directly applicable to economic problems. In economic problems, both the social alternatives and the individual preferences exhibit considerable structure and, therefore, a social welfare function only needs to be defined on a restricted domain of preference profiles. For a comprehensive survey of the literature on Arrovian social choice on economic domains, see Le Breton and Weymark (2008).

When X is a subset of the real line $\mathbb{R}$, a preference R_i is *single-peaked* if there is a unique best alternative $\pi(R_i)$ in X, the *peak*, and alternatives on the same side of the peak are worse the further away from the peak they are. Let $\mathcal{S}$ denote the set of all single-peaked preferences on X. If the alternatives in X are different levels of a single public good, it is natural to expect individual preferences to be single-peaked. Black (1948) has shown that ranking pairs of alternatives by majority rule produces a social ordering if the individuals have single-peaked preferences when n is odd. More generally, it follows from results in Moulin (1980) that on $\mathcal{S}^n$, any *generalized median social welfare* function satisfies all the Arrow axioms except his domain assumption with nondictatorship strengthened to anonymity (permuting preferences leaves the social ordering invariant). These functions are defined by first *fixing* n–1 single-peaked preferences (which can be interpreted as being the preferences of phantom voters) and then, for any profile of single-peaked preferences in $\mathcal{S}^n$ of the n real individuals, applying majority rule to the resulting profile of $2n$–1 preferences, both real and phantom. Note that the total number of preferences in one of these profiles is odd, so Black's theorem applies. Each specification of the preferences of the phantom voters defines a distinct generalized median social welfare function. Ehlers and Storcken (2002) have characterized all the social welfare functions on this domain that satisfy IIA and weak Pareto.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

A domain $\mathcal{D}$ of preference profiles is *Arrow inconsistent* if no social welfare function satisfying Arrow's three non-domain axioms exists on $\mathcal{D}$. In a seminal article, Kalai, Muller and Satterthwaite (1979) identified a sufficient condition for $\mathcal{D}$ to be Arrow inconsistent when $\mathcal{D}$ is the Cartesian product of individual preference domains $\mathcal{D}_i$. A set of alternatives is *free* if preference profiles are unrestricted on this set. A domain is *saturating* if (i) there are at least two free pairs, (ii) any two free pairs of alternatives can be connected to each other by means of a series of overlapping free triples, and (iii) any other pair of alternatives is *trivial* in the sense that there is only one way in which any individual ranks these alternatives. When each of the individual preference domains $\mathcal{D}_i$ is the same, saturating preference domains are Arrow inconsistent. Because a free pair is part of a free triple when the domain is saturating, Arrow's theorem implies that there is a dictator on this pair when IIA and weak Pareto are satisfied. The same person must be a dictator on all free pairs because adjacent free triples in the connection procedure have two alternatives in common. On trival pairs, by weak Pareto, everyone is a dictator. This method of showing that a domain is Arrow inconsistent is known as the *local approach*.

Kalai, Muller and Satterthwaite (1979) have also shown that, when $X = \mathbb{R}^m_+$, interpreted as the set of all allocations of *m* divisible public goods, the domain of all profiles of *classical public goods preferences* (that is, continuous, strictly monotonic, and convex preferences) is saturating and, hence, is Arrow inconsistent when $m \geq 2$. Other examples of saturating domains include the set of all expected utility preferences on the set of lotteries on three or more certain outcomes (Le Breton, 1986) and the set of Euclidean spatial preferences on $\mathbb{R}^m_+$ or $\mathbb{R}^m$, that is, preferences for which there is a global best alternative and alternatives are ranked by the negative of their distance from this alternative (Le Breton and Weymark, 2002). The Arrow inconsistency of the spatial preference domain was originally shown by Border (1984) using a different proof strategy.

When alternatives are allocations of private goods and individuals only care about their own consumption, Bordes and Le Breton (1989) have identified a strengthening of the concept of a saturating domain that implies that the domain is Arrow inconsistent. If *X* consists of all the allocations of two or more divisible private goods in which everyone is guaranteed to receive a positive amount of some good, then the domain satisfies this condition if individuals can have any *classical private goods preference*, that is, a preference that is continuous, strictly monotonic, and convex over own consumption (see also Maskin, 1976; Border, 1983).

The examples considered so far all have the feature that the set of alternatives has a Cartesian structure. If *X* incorporates feasiblity constraints, this is not the case. Using a modification of the local approach, Bordes, Campbell and Le Breton (1995) have shown that the domain of classical private goods preferences is Arrow inconsistent if the set of alternatives is the set of feasible allocations with positive consumptions of all goods for an exchange economy with two or more divisible private goods. Bordes and Le Breton (1990) have also adapted the local approach to analyse Arrow consistency in assignment, matching and pairing problems. In an *assignment problem*, one of *n* indivisible objects is assigned to each of the *n* individuals. In a *matching problem*, there are two groups of *n* individuals with each person from one group matched to one person from the other group. In a *pairing problem*, an even number *n* of individuals is grouped in pairs. If the preference domains in these problems are such that individuals only care about which individual or good they are matched, paired or assigned to, but are otherwise unrestricted, then the domain is Arrow inconsistent when $n \geq 4$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The preceding discussion suggests that economic domain restrictions do not provide a satisfactory way of circumventing Arrow's social welfare function impossibility theorem when the set of alternatives is not one-dimensional. This conclusion is reinforced by the results in Redekop (1995) that show that, in order for a subset of a domain of Arrow-inconsistent economic preferences to be Arrow consistent, the sub-domain must be topologically small. Roughly speaking, this requirement severely limits the amount of preference diversity that can be present in the domain.

Arrow's theorem can also be formulated in terms of a social choice correspondence. For each preference profile **R** in its preference domain $\mathcal{D}$, a *social choice correspondence* C specifies the socially optimal alternatives $C(A, \mathbf{R})$ in each agenda A (feasible subset of X) in its agenda domain $\mathcal{A}$. In its choice-theoretic formulation, the Arrow axioms are Arrow's choice axiom (for a fixed preference profile, if agenda A is a subset of agenda B, then the set of alternatives chosen in A consists of the restriction to A of the set of alternatives chosen from B when this restriction is nonempty), independence of infeasible alternatives (the alternatives chosen from an agenda only depend on the preferences for alternatives in this agenda), Pareto optimality (only Pareto optimal alternatives are chosen), and nondictatorship (the chosen alternatives are not always a subset of one individual's best feasible alternatives). Arrow's theorem shows that these conditions are inconsistent if the preference domain is unrestricted and the agenda domain consists of all the finite subsets of X. When the agenda domain is closed under finite unions (as is the case in the choice-theoretic version of Arrow's theorem), Arrow's choice axiom is necessary and sufficient for the chosen alternatives in each admissible agenda to be generated by maximizing a profile-dependent social ordering of X (see Hansson, 1968).

In some economic applications, the ability to restrict the agenda domain, not just the preference domain, has weakened the constraints on the admissible social choice correspondences sufficiently for the Arrovian axioms to be consistent. This observation was first made by Bailey (1979), who noted that the set of feasible allocations in an exchange economy does not contain a finite number of alternatives, and so does not satisfy Arrow's agenda domain assumption. While the example Bailey used to show the consistency of the Arrow axioms is problematic, as Donaldson and Weymark (1988) have shown, if each agenda in the agenda domain is the set of feasible allocations for an exchange economy with divisible private goods (different aggregate endowments yield different agendas) and if each profile in the preference domain is a profile of classical private goods preferences for which no individual is indifferent between a consumption bundle with strictly positive components and one that has zero consumption of some good, then the Arrow axioms are consistent. For example, the *equal division Walrasian social choice correspondence* satisfies these axioms. For each exchange economy, this correspondence selects the set of Walrasian (competitive) equilibrium allocations using an equal division of the aggregate endowment as each individual's endowment vector.

When production is possible, an agenda is the set of feasible allocations given the aggregate resource endowment and the production technologies. Possible restrictions on agendas include compactness, comprehensiveness (that is, they satisfy free disposal), and convexity. When there are only public goods, Le Breton and Weymark (2002) have shown that the Arrow axioms are consistent if the preference domain includes only Euclidean spatial preferences on $\mathbb{R}^m_+$ with $m \geq 2$ and the agenda domain includes only compact sets with nonempty interiors. With these domain assumptions, a social choice correspondence satisfying the Arrow axioms can be constructed by fixing a utility representation for each preference and then using an individualistic Bergson–Samuelson social welfare function to choose

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the best alternatives from each agenda for each preference profile.
In both of these examples, one of the choice-theoretic versions of the Arrow axioms is vacuous. In the exchange economy example, it is Arrow's choice axiom, whereas in the spatial example, it is independence of infeasible alternatives. For a public goods economy with at least two divisible goods, none of the Arrow axioms is vacuous if the agenda domain includes only compact comprehensive sets with nonempty interiors and the preference domain includes only classical public goods preferences. By means of an example, Donaldson and Weymark (1988) have shown that the Arrow axioms are consistent with these domain assumptions. However, their example exhibits dictatorial features and it is not known if the axioms are still consistent if nondictatorship is replaced with anonymity (permuting preferences for a given agenda does not change the set of chosen alternatives). Donaldson and Weymark have also established a private goods version of this possibility theorem.
Arrovian impossibility results have also been obtained with the social choice correspondence framework using a strengthened form of independence of infeasible alternatives, due to Donaldson and Weymark (1988), called independence of Pareto-irrelevant alternatives. This condition requires the chosen alternatives from each agenda to depend only on the preferences over the Pareto optimal alternatives. For example, for public goods economies, Duggan (1996) has shown that this strengthened independence condition and the other Arrow axioms are inconsistent if $X = \mathbb{R}^m_+$ with $m \geq 3$, the agenda domain consists of all the compact, comprehensive and convex subsets of X, and the preference domain is the set of all profiles of classical public goods preferences for which individual preferences are strictly convex in own consumption.
In contrast with the local approach used to analyse social welfare functions, no unifying methodology has been developed to investigate the consistency of Arrow's choice-theoretic axioms, with the consequence that little is yet known about where the boundary between possibility and impossibility for social choice correspondences lies.

Variable-population social choice

The Arrovian framework is based on ordinal preferences that are interpersonally noncomparable and, hence, any social decision rule that makes use of interpersonal utility comparisons, such as classical utilitarianism, is ruled out from the outset. Sen (1974) has argued that this informational poverty plays a fundamental role in precipitating Arrovian impossibilities, and has proposed a generalization of the concept of an Arrovian social welfare function called a social welfare functional to allow for interpersonal utility comparisons. Each individual i is assumed to have a utility function U_i on the set of alternatives X_i in which he is alive and a *social welfare functional* maps each admissible profile of individual utility functions into a social ordering of the set of all alternatives X. In fixed-population social choice, $X_i=X$ for all i. There is an extensive literature that has investigated the implications for the functional form of these functionals of combining different assumptions concerning the measurability and interpersonal comparability of utility with various normative criteria, including analogues of the Arrovian axioms, when there is a fixed population: see Bossert and Weymark (2004) for a survey. In this section, we provide an introduction to the main issues that arise in selecting appropriate social objective functions when the population size is not fixed. A detailed treatment of this topic and further references

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
may be found in Blackorby, Bossert and Donaldson (2005).

Since the 1980s, population ethics has established itself as an important branch of moral philosophy. Parfit (1984) has been particularly influential in bringing this issue to the attention of philosophers and, more generally, to scholars in various disciplines interested in applied ethics. An up-to-date account of variable-population issues in moral philosophy is given by Broome (2004). Although there are many economic applications of variable-population social choice, such as the design of aid packages (that may have population consequences) for developing countries, the choice of budgets devoted to prenatal care, and policies affecting the intergenerational allocation of resources, the economics literature, with few exceptions, did not initially pay much attention to this topic. Much of the recent interest in these issues can be traced to the influential article by Blackorby and Donaldson (1984), who extended the welfarist model of social choice to allow for a variable population.

In this setting, each alternative $x \in X$ is a complete description of the relevant state of affairs including the size and composition of the population. Furthermore, alternatives are interpreted as full histories of the world, from the remote past to the distant future. Thus, the set of those alive in x contains everyone who has ever lived in this alternative and not merely those who are alive in a given period. This assumption is important to avoid counter-intuitive conclusions regarding the termination of lives. As a consequence, ending someone's life does not change population size; it affects the lifetime and, possibly, the lifetime utility of the person in question.

For each $x \in X$, $u_i = U_i(x)$ is the lifetime well-being (utility) of any individual i alive in x and $U(x)$ is the vector of utilities of these individuals. The standard convention is to assign a lifetime utility level of zero to a *neutral* life. A life, taken as a whole, is a neutral life from the viewpoint of the individual leading it if it is as good as a life without any experiences (a state of permanent unconsciousness). Note that it is not necessary to invoke states of non-existence of an individual in order to define the notion of neutrality. In particular, it is not claimed that an individual can gain or lose by being brought into existence. Therefore, an existing person's life is worth living if the individual's lifetime utility is positive.

Welfarism is the principle that the only features of an alternative that are socially relevant are the utilities of the individuals alive in this alternative. Welfarism implies that the social ordering of X for any profile of individual utility functions in the domain of the social welfare functional can be determined by a single *social welfare ordering* of all possible vectors of individual utilities $\mathcal{U} = \cup_{n \in \mathbb{N}} \mathbb{R}^n$, where $\mathbb{N}$ is the set of positive integers; that is, if a social welfare functional is welfarist, there exists an ordering R on $\mathcal{U}$ such that alternative $x \in X$ is at least as good as alternative $y \in X$ for the profile of utility functions U if and only if $U(x) R U(y)$. The set of individuals alive in x and y need not be the same. Thus, given welfarism, the problem of variable-population social evaluation can be reduced to the problem of establishing a social welfare ordering R on the set $\mathcal{U}$ of all utility vectors (of varying dimension). If there are $n \in \mathbb{N}$ individuals alive in an alternative, without loss of generality they can be labelled 1, … , n provided that the standard anonymity property is satisfied. A representation of the restriction of R to fixed-population comparisons is an individualistic Bergson–Samuelson social welfare function.

The most commonly discussed examples of variable-population social welfare orderings are extensions of utilitarianism. According to *average utilitarianism* (AU) (resp. *classical utilitarianism* (CU)), average (resp. total) utilities are used as the criterion to compare any two utility vectors. Formally, for all $n, m \in \mathbb{N}$, all $u \in \mathbb{R}^n$, and all $v \in \mathbb{R}^m$, $u R_{\mathrm{AU}} v$ if and only if $\frac{1}{n}\sum_{i=1}^{n} u_i \geq \frac{1}{m}\sum_{i=1}^{m} v_i$ (resp. $u R_{\mathrm{CU}} v$ if and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

only if $\sum_{i=1}^{n} u_i \geq \sum_{i=1}^{m} v_i$). Clearly, fixed-population comparisons are the same according to R_{AU} and R_{CU}, but this is not necessarily the case if *n* and *m* differ.

Average utilitarianism is rejected by most contributors to this area. Its fundamental problem is that the value of adding a person, *ceteris paribus*, depends on the utilities of those alive. This has rather unfortunate consequences. Suppose, for example, that everyone is extremely well-off in an alternative and we consider the addition of an individual who, if brought into existence, would have a lifetime utility just slightly below the average of the existing population and no one else's utility is affected. According to AU, this person should not be brought into existence. The following example is even more disturbing. Consider a society in which everyone is extremely miserable by all standards (and well below neutrality). AU recommends the *ceteris paribus* addition of anyone with a lifetime utility slightly above the average, even if this utility level is well below neutrality.

Classical utilitarianism suffers from what Parfit (1984) calls the *repugnant conclusion*. A variable-population social welfare ordering *R* implies the repugnant conclusion if, for any population size *n*, for any positive level of utility ξ (no matter how high), and for any level of utility $\varepsilon \in (0, \xi)$ (no matter how close to zero), there exists a population size *m*>*n* such that a population with *n* people in which everyone has a lifetime utility of ξ is considered inferior to a population of *m* individuals each of whom has a lifetime utility of ε; that is, for any situation in which everyone alive has an arbitrarily high level of well-being, there is always a situation of mass poverty (with everyone arbitrarily close to neutrality) that is considered superior.

In order to avoid the repugnant conclusion and, at the same time, the counter-intuitive implications of average utilitarianism, Blackorby and Donaldson (1984) have proposed *critical-level utilitarianism* (CLU) with a positive critical level as an alternative criterion. CLU employs a parameter $\alpha \in \mathbb{R}$ (the critical level) and is defined by letting, for all $n, m \in \mathbb{N}$, all $u \in \mathbb{R}^n$, and all $v \in \mathbb{R}^m$, $uR_{CLU}v$ if and only if $\sum_{i=1}^{n} [u_i - \alpha] \geq \sum_{i=1}^{m} [v_i - \alpha]$. The special case corresponding to $\alpha = 0$ is CU. The parameter α has an intuitive interpretation: it is the level of utility that, if experienced by an additional person, makes the alternative resulting from the *ceteris paribus* addition of such a person to any given society as good as the original. Because the critical level is constant, the problems of AU alluded to above are avoided. If, moreover, α is positive, the repugnant conclusion is avoided because there is a positive difference between the critical level and the level of utility representing neutrality.

In addition to providing a thorough analysis of critical-level utilitarianism and its main alternatives, Blackorby, Bossert and Donaldson (2005) have discussed several extensions of the basic model. For example, the critical-level utilitarian orderings can be generalized by considering transformed utilities rather than the utilities themselves. If the transformation is chosen to be strictly concave, the corresponding social ordering represents inequality aversion in utilities. Furthermore, they have considered orderings that use non-welfare information such as birth dates and lifetimes in addition to lifetime utilities, as well as variants that incorporate uncertainty. Moreover, they have analysed variable-population choice problems and applications.

Strategy-proof social choice

A *social choice* function *g* chooses one alternative from the set of alternatives *X* for each preference

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

profile in the domain $\mathcal{D}$. If it is only known that the true profile is in $\mathcal{D}$, in order to implement the desired choice $g(\mathbf{R})$ when the profile is $\mathbf{R}$, individuals must have an incentive to truthfully report their preferences. *Strategy-proofness* is the requirement that nobody can obtain a preferred outcome by reporting a false preference regardless of what the preferences of the other individuals are. Strategy-proofness places severe constraints on the kinds of social choice functions that can be considered and, on some domains, conflicts with other social desiderata (for introductions to recent developments in strategy-proof social choice theory, see Sprumont, 1995; Barberà, 2001).

The classic result on strategy-proofness is the Gibbard (1973)–Satterthwaite (1975) theorem, which shows that no social choice function g can satisfy both nondictatorship and Pareto optimality when $\mathcal{D} = \mathcal{R}^n$ if $|X| \geq 3$. The same conclusion follows if Pareto optimality is replaced with *unanimity*, the requirement that an alternative is chosen if everybody agrees that it is uniquely best. Either of these conditions implies that the range rg(g) of g is all of X when the domain is unrestricted. A variant of the Gibbard–Satterthwaite theorem states that on an unrestricted domain, if $|\mathrm{rg}(g)| \geq 3$, then strategy-proofness implies that someone must be a dictator on rg(g) (that is, g always chooses one of this person's best alternatives on rg(g)).

More positive results are obtained if it is known that preferences are single-peaked. Moulin (1980) has shown that if $X \subseteq \mathbb{R}$, $\mathcal{D} = \mathcal{S}^n$, and the social choice function g only depends on the peaks of the individual preferences, then g satisfies strategy-proofness if and only if it is a minmax social choice function and it satisfies strategy-proofness, Pareto optimality and anonymity if and only if it is a generalized median social choice function. A *minmax* social choice function g is defined by specifying an alternative x_S in the closure of X for each coalition of individuals with $x_T \leq x_S$ if $S \subseteq T$ and setting

$$g(\mathbf{R}) = \min_{S \subseteq N} \left\{ \max_{i \in S} \left\{ \pi(R_i),\ x_S \right\} \right\},\ \forall \mathbf{R} \in \mathcal{S}^n.$$

For each $\mathbf{R} \in \mathcal{S}^n$, a *generalized median social choice* function chooses the median of the actual individual preference peaks and the fixed peaks of n–1 phantom voters. These functions are minmax rules in which the alternatives x_S are the same for coalitions of the same size. Barberà, Gul and Stacchetti (1993) have provided an alternative characterization of minmax rules in terms of winning coalitions that has proved to be quite useful. If, as is the case with minmax rules, the chosen alternative for each profile depends only on each person's most-preferred alternative(s) on the range, the social choice function satisfies the *tops-only* property. On the domain $\mathcal{S}^n$, Barberà and Jackson (1994) have shown that the tops-only property assumed by Moulin (1980) is implied by strategy-proofness if either rg (g) is an interval or g satisfies Pareto optimality.

The original strategies used to prove the Gibbard–Satterthwaite theorem cannot be adopted to analyse strategy-proofness when preferences are continuous. The problem is that these proofs alter profiles by moving two alternatives to the top two positions in a person's preference, but this is not possible with continuous preferences if X is a connected set, as there can be no second-ranked alternative. This difficulty was overcome by Barberà and Peleg (1990) who established a version of the Gibbard–

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Satterthwaite theorem for continuous preferences on a metric space of alternatives using the option set methodology introduced by Laffond (1980), Satterthwaite and Sonnenschein (1981), and Barberà (1983). An *option set* identifies the set of outcomes that are feasible given the preferences of a subgroup of individuals for some admissible reported preferences of the rest of the population. For example, when there is a dictator *d*, the option set generated by R_d consists of the best alternatives on the range for this preference and the option set generated by any other person's preference is the whole range. The option set methodology proceeds by identifying the structure imposed on option sets by the properties that one wants the social choice function to satisfy.

In order for a social choice function to be strategy-proof, it must ignore most of the information about individual preferences. On many domains in which all admissible preferences have unique best alternatives on the range, strategy-proofness implies the tops-only property provided that the range of the social choice function satisfies some regularity condition. Weymark (2004) has proposed a proof strategy for establishing the tops-only property that avoids the model specificity of earlier proofs.

A social choice function defined on a domain of profiles of separable preferences on a product set of alternatives is *decomposable* if the value chosen for a component depends only on the individual marginal preferences for that component. The first decomposability results were established by Border and Jordan (1983) who, for example, showed that, for the domain of all profiles of separable quadratic preferences on a multidimensional Euclidean space, a social choice function satisfies strategy-proofness and unanimity if and only if it decomposes into strategy-proof, unanimous social choice functions on each component. Furthermore, these one-dimensional mechanisms can be any member of Moulin's class of minmax social choice functions. Since the development of the option set methodology, decomposability results for strategy-proof social choice functions have been established for a number of other domains of separable preferences. For example, Barberà, Gul and Stacchetti (1993) have shown that, if X is a discrete product set in a Euclidean space and individuals can have any separable preference that satisfies a multidimensional analogue of single-peakedness, then the conclusions of Border and Jordan's theorem hold if the range of the social choice function is all of *X*. Whether strategy-proofness and auxiliary conditions such as unanimity imply decomposability depends on how much preference variability is present in the domain. Establishing a decomposability theorem typically involves first showing that the tops-only property is satisfied, as in Barberà, Gul and Stacchetti (1993). Much of the literature on this issue has been synthesized and extended by Le Breton and Sen (1995; 1999).

If *X* is a product set, but only a subset *Z* of *X* is feasible, decomposability results are still possible, but not every combination of the corresponding one-dimensional social choice functions is admissible. For example, using the model in Barberà, Gul and Stacchetti (1993) with the best alternative for each preference required to be in *Z*, Barberà, Massó and Neme (1997) have shown that any social choice function *g* that is strategy-proof and whose range is *Z* is decomposable into one-dimensional minmax rules on each component, but, in order for a combination of such minmax rules to always produce a feasible outcome, *g* must satisfy a rather complicated condition called the *intersection property*.

In the preceding discussion, everyone has the same set of admissible preferences, and so it is possible that they might agree on what is best. When there are private goods and individuals care only about their own consumption, one generally expects there to be distributional conflicts. We illustrate the implications of strategy-proofness with private goods in two problems: the allotment problem and the exchange of divisible goods.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In an *allotment problem*, there is a fixed amount Ω of a divisible good to allocate. If individuals care only about own consumption, each person's preference is defined on $X=[0, \Omega]$. If these preferences are single-peaked, a prominent solution to this problem is the *uniform rule* (see Benassy, 1982) which, for each admissible profile $\mathbf{R} \in \mathcal{S}^n$, chooses the unique allocation $x=(x_1, \ldots, x_n) \in X^n$ (x_i is person i's allocation) for which (i) if $\Omega \leq \sum_{i=1}^{n} \pi(R_i)$, there exists $\lambda \in \mathbb{R}_+$ such that, for all $i \in N$, $x_i=\min\{\pi(R_i), \lambda\}$ and (ii) if $\Omega \geq \sum_{i=1}^{n} \pi(R_i)$, there exists $\lambda \in \mathbb{R}_+$ such that, for all $i \in N$, $x_i=\max\{\pi(R_i), \lambda\}$.

Sprumont (1991) has shown that, if the domain is the set of all profiles of continuous single-peaked preferences on X, then a social choice function satisfies strategy-proofness, Pareto optimality and private-goods anonymity (permuting preferences results in the same permutation of the individual allocations) if and only if it is the uniform rule. Sprumont's article also includes the first explicit theorem about the tops-only property in the strategy-proofness literature.

When X is the set of allocations of an exchange economy with two or more divisible private goods, the general conclusion is that strategy-proofness and Pareto optimality conflict with other desirable properties for a social choice function on a sufficiently rich domain of classical private goods preference profiles. If the aggregate endowment is privately owned and participation in the collective choice procedure is voluntary, the social choice function must satisfy individually rationality; that is, each person is guaranteed a consumption bundle weakly preferred to his endowment. Hurwicz (1972) has shown that strategy-proofness, Pareto optimality and individual rationality are inconsistent for two-person, two-good exchange economies on such a preference domain. This impossibility theorem has only recently been extended to the general n-person, m-good case by Serizawa (2002).

With monotonic preferences, a dictator in an exchange economy always receives the whole endowment. For the domain of classical private goods preference profiles, Zhou (1991) has shown that strategy-proofness, Pareto optimality and nondictatorship are inconsistent when there are at least two goods, but only two individuals. When there are at least three individuals, Satterthwaite and Sonnenschein (1981) have shown by example how to construct Pareto optimal, strategy-proof, nondictatorial social choice functions for this domain. In their example, someone is *bossy* (that is, there is an individual who can change the consumption bundle of someone else by reporting a different preference without affecting his or her own consumption bundle) and, for each profile, one of two individuals receives all of the endowment. It is generally agreed that bossy mechanisms are unsatisfactory. Serizawa and Weymark (2003) have shown that any social choice function that satisfies strategy-proofness and Pareto optimality cannot guarantee everyone a consumption bundle bounded away from the origin on a rich domain of classical private goods preferences.

Given that any strategy-proof and Pareto optimal social choice function g must fail even minimal distributional desiderata on such domains, Barberà and Jackson (1995) have explored the implications of abandoning Pareto optimality. For private ownership exchange economies with classical private goods preferences, they have shown that if g is strategy-proof, nonbossy, and satisfies some other auxiliary conditions, then trade must be restricted to occur in a limited set of fixed proportions with possibly upper limits on the amounts that can be exchanged. In the case of two goods and two individuals, if g satisfies strategy-proofness and individual rationality, there are only two such proportions, and the choice procedure resembles the fixed-price trading rules studied by Benassy (1982) with different buying and selling prices for each good.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Axiomatic models of resource allocation

The literature on axiomatic models of resource allocation has a close affinity to the literature on Arrovian social choice on economic domains. As is the case with Arrovian social choice when there are multiple agendas, the research on axiomatic models of resource allocation investigates the implications of normative criteria (axioms) when both individual preferences and the set of feasible agendas satisfy the kinds of restrictions found in economic models. What distinguishes this literature is the set of axioms considered, many of which rely on the special structure provided by economic models for their definition. In this section, we present a very selective introduction to the models and axioms considered in this literature and describe a few of the theorems that have been obtained. For a comprehensive survey of this literature, see Thomson (2008).

In an *allocation problem*, there is an aggregate social endowment $\Omega \in \mathbb{R}^m_{++}$ of m private goods that are to be allocated among $n \geq 2$ individuals based on their preferences. In the canonical allocation problem, $m \geq 2$ and all goods are divisible. An *economy* is then described by a pair E=(**R**, Ω), where **R** is a profile of classical private goods preferences. Let $\mathcal{E}$ denote the set of all such economies. Given the endowment Ω, the corresponding agenda $A(\Omega)$ is the set of feasible allocations $x=(x_1, \ldots, x_n)$ that exhaust Ω, where $x_i \in \mathbb{R}^m_+$ is person i's consumption bundle.

A *solution* is a mapping that selects a subset of the feasible allocations for each economy in $\mathcal{E}$. Note that a solution φ can be identified with a social choice correspondence C by setting $C(A(\Omega), \mathbf{R})=\varphi(E)$ for all $E \in \mathcal{E}$. A solution satisfies *efficiency* if it always chooses Pareto optimal allocations and it satisfies *no envy* if, at any selected allocation, nobody strictly prefers anyone else's allocated consumption bundle. No envy, which was independently introduced by Tinbergen (1953), Foley (1967), and Kolm (1972), is the fundamental fairness condition considered in this literature. An example of a solution satisfying both efficiency and no envy for this class of economies is the *equal division Walrasian solution* φ^W, which is defined from the equal division Walrasian social choice correspondence in the manner described above.

The literature on fair allocation has expanded the scope of the canonical model in several respects. For example, economies with varying populations or with production have been considered. In addition, variations of this model have been explored, for example, by allowing for indivisibilities or, when there is only one good, preference restrictions such as single-peakedness.

An alternative set-up with public goods has also been examined. The existence of solutions satisfying efficiency and no envy in public goods environments is a more complex matter than in the private goods case, largely because arguments involving pure exchange cannot be invoked when there are public goods. Furthermore, the technology that permits us to transform private goods into public goods is also important. With some additional assumptions, however, efficiency and no envy can be satisfied (see Diamantaras, 1991, for example).

Prominent among the new axioms that have been introduced and used in characterizations of existing and new solutions is *consistency*, which is discussed in detail in Thomson (1990). In order to define consistency, the notion of a solution must be extended to include economies with different numbers of individuals. Let x be an allocation that is selected by such a solution for an (n+1)-person economy. Now suppose that person k leaves the economy with the consumption x_k. Define a *reduced* n-person economy

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

by removing k and subtracting x_k from the total endowment. Consistency demands that the allocation $(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{n+1})$ is selected in the reduced economy.

Other important properties include monotonicity conditions with respect to the quantities of the resources available, with respect to the technology, or with respect to the population. A solution φ satisfies *resource monotonicity* if, whenever the social endowment expands, no one becomes worse off in any chosen allocation. In private goods models with production, an axiom similar in spirit to resource monotonicity is *technology monotonicity*. It requires that if the only difference between two economies is that the technology of one dominates that of the other, then everyone should be at least as well-off in any allocation chosen for the former economy as in any allocation chosen for the latter. *Population monotonicity* is a solidarity axiom. As is the case for consistency, it applies in models with variable population. Suppose that the population is expanded, but the total endowment is unchanged. Population monotonicity demands that the burden imposed on the existing population by the presence of the additional individuals is shared by all its members; no one who is present before the population expansion is better off as a consequence of the population augmentation.

If there is only one divisible good, each economy $E=(\mathbf{R}, \Omega)$ defines an allotment problem, as in the preceding section. When all preference profiles are single-peaked, the *uniform solution* simply applies the uniform rule for the allotment problem to each economy in the domain. In addition to the characterization of this solution presented in the preceding section, there have been axiomatizations of the uniform solution using no envy, consistency, and variants of either resource monotonicity or population monotonicity, among other axioms (see Thomson, 2008).

If some of the goods to be allocated are indivisible, much of the theory developed in the context of perfectly divisible goods still applies. Due to the specific nature of the problem of allocating indivisible objects, some interesting additional results can be obtained. As an illustration, consider an assignment problem in which n indivisible objects are to be allocated to n individuals and there is also a perfectly divisible good ('money') that can be consumed in any amount, positive or negative: see Thomson (2008) for references to contributions that permit the number of goods and individuals to differ. A commodity bundle for person i is now a pair $(t_i, j) \in \mathbb{R} \times N$, where t_i (resp. j) is the amount of money (resp. object) allocated to i. It is assumed that i's preference R_i on $\mathbb{R} \times N$ is strictly monotonic in money and that money can be used to compensate for the receipt of a less desirable good in the sense that, for all $t_i \in \mathbb{R}$ and all $j, k \in N$, there exists $s_i \in \mathbb{R}$ such that i is indifferent between (s_i, k) and (t_i, j). An economy now consists of a preference profile $\mathbf{R}$ with the properties introduced above, an aggregate endowment of money $T \in \mathbb{R}$, and the n indivisible objects. Because the set of objects is fixed, an economy can be characterized by a pair $E=(\mathbf{R},T)$. A feasible allocation for E is a pair (t, ρ), where $t \in \mathbb{R}^n$ is a vector of *balanced* monetary allocations (that is, $\sum_{i=1}^{n} t_i = T$) and $\rho : N \to N$ is a permutation with $\rho(i)$ specifying the object allocated to person i.

Solutions and the efficiency and no-envy axioms are defined in the usual manner. If the monetary allocations are restricted to be non-negative, it is clear that solutions satisfying no envy may not exist. For example, if $T=0$ and everyone regards the same object as being uniquely best regardless of the amount of money received, whoever is allocated this object is envied by everyone else because no monetary compensations are possible. Sufficient conditions for the existence of solutions satisfying no envy with non-negative monetary allocations are discussed in Thomson (2008). As is to be expected,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
these conditions ensure that there is a sufficient amount of money available to carry out the requisite compensation payments.

In the case of perfectly divisible goods, we have noted that the equal division Walrasian solution φ^W satisfies efficiency and no envy. Interestingly, in the indivisible good model considered here, the allocations generated by an adaptation of this Walrasian solution to the present framework are the *only* allocations satisfying no envy (see Svensson, 1983). Moreover, efficiency *follows* as a consequence of no envy. In this model, for an economy E, if everyone is provided with the same endowment $t_0 \in \mathbb{R}_+$ of money, a *Walrasian equilibrium* is a feasible allocation (t, ρ) and a price $p_k \in \mathbb{R}_+$ for each good $k \in N$ such that the bundle $(t_i, \rho(i))$ is weakly preferred by individual i among all bundles that have values no more than t_0. The solution φ^W is then defined by letting $\varphi^W(E)$ be the set of Walrasian equilibrium allocations that can be obtained in this way.

A number of fairness principles besides no envy have been considered in the literature (see Fleurbaey and Maniquet, 2008; Thomson, 2008). Particularly notable among them is *egalitarian equivalence*, which is due to Pazner and Schmeidler (1978). In the canonical allocation problem, egalitarian equivalence requires that, for each economy $E=(\mathbf{R}, \Omega)$, each selected allocation x has the property that there exists a consumption bundle $z_0 \in \mathbb{R}^m_+$ that everyone is indifferent to. Pazner and Schmeidler (1978) have shown that on the domain of economies $\mathcal{E}$ for this problem, solutions exist that satisfy both egalitarian equivalence and Pareto optimality. However, egalitarian equivalence need not satisfy independence of infeasible alternatives, as the egalitarian allocation $(z_0, \ldots, z_0)$ associated with x need not be feasible.

There is now an extensive literature that employs the framework and many of the axioms described in this section to re-examine the foundations of egalitarian theories. If individuals are held responsible in part for the outcomes they receive, conditional versions of egalitarianism demand that individual differences caused by factors beyond the individuals' control should be compensated for, whereas inequities that can be attributed to choices for which an individual is responsible do not attract that kind of equalization. Variants of this theory have been advocated by, for example, Roemer (1993): see Fleurbaey and Maniquet (2008) for a detailed survey of this literature.

Concluding remarks

As noted above, the response of Sen (1974) to Arrovian social welfare function impossibilities was to abandon the ordinal non-comparability of individual utilities built into the Arrow framework. However, he maintained the spirit of IIA by assuming that the social ranking of any two alternatives should depend only on the individual utilities obtained with them. This independence assumption is the cornerstone of the welfarist approach employed in the literatures on social choice with interpersonal utility comparisons and on variable-population social choice.

A different resolution of the Arrovian dilemma has been described and defended in Fleurbaey (2007). Rather than abandoning ordinal non-comparability, Arrow's IIA assumption is relaxed so as (i) to allow the social ranking of a pair of alternatives to depend on how these alternatives are ranked relative to some other alternatives and (ii) to incorporate some principle of fairness. This proposal has been

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

explored in a series of articles by Fleurbaey and various co-authors. For example, Fleurbaey, Suzumura and Tadenuma (2005) have shown that, when the alternatives are the set of all allocations of $m \geq 2$ divisible private goods and the domain of the social welfare function is the set of profiles of classical private goods preferences, then weak Pareto and a private goods version of anonymity are compatible with an independence condition that incorporates fairness considerations of the sort embodied in egalitarian equivalence. Independence conditions such as this or ones based on envy-freeness employ non-local information about preferences, including information about alternatives that may not be feasible if resource constraints are taken into account. This line of research provides a bridge between the literatures on Arrovian social choice on economic domains and axiomatic models of resource allocation discussed above by employing some form of independence condition (as in the former) while at the same time requiring social decisions to be fair (as in many of the latter contributions).

See Also

- Arrow's theorem
- mechanism design
- mechanism design (new developments)
- social choice
- strategic voting
- strategy-proof allocation mechanisms
- voting paradoxes

Bibliography

Arrow, K.J. 1963. *Social Choice and Individual Values*, 2nd edn. New York: Wiley.

Arrow, K.J., Sen, A.K. and Suzumura, K. 2002. *Handbook of Social Choice and Welfare*, vol. 1. Amsterdam: North-Holland.

Arrow, K.J., Sen, A.K. and Suzumura, K. 2008. *Handbook of Social Choice and Welfare*, vol. 2. Amsterdam: North-Holland.

Austen-Smith, D. and Banks, J.S. 2005. *Positive Political Theory II: Strategy and Structure*. Ann Arbor: University of Michigan Press.

Bailey, M.J. 1979. The possibility of rational social choice in an economy. *Journal of Political Economy* 87, 37–56.

Barberà, S. 1983. Strategy-proofness and pivotal voters: a direct proof of the Gibbard–Satterthwaite theorem. *International Economic Review* 24, 413–17.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Barberà, S. 2001. An introduction to strategy-proof social choice functions. *Social Choice and Welfare* 18, 619–53.

Barberà, S., Gul, F. and Stacchetti, E. 1993. Generalized median voting schemes and committees. *Journal of Economic Theory* 61, 262–89.

Barberà, S. and Jackson, M. 1994. A characterization of strategy-proof social choice functions for economies with pure public goods. *Social Choice and Welfare* 11, 241–52.

Barberà, S. and Jackson, M.O. 1995. Strategy-proof exchange. *Econometrica* 63, 51–87.

Barberà, S., Massó, J. and Neme, A. 1997. Voting with constraints. *Journal of Economic Theory* 76, 298–321.

Barberà, S. and Peleg, B. 1990. Strategy-proof voting schemes with continuous preferences. *Social Choice and Welfare* 7, 31–8.

Benassy, J.-P. 1982. *The Economics of Market Disequilibrium*. New York: Academic Press.

Black, D. 1948. On the rationale of group decision-making. *Journal of Political Economy* 56, 23–34.

Blackorby, C., Bossert, W. and Donaldson, D. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.

Blackorby, C. and Donaldson, D. 1984. Social criteria for evaluating population change. *Journal of Public Economics* 25, 13–33.

Border, K.C. 1983. Social welfare functions for economic environments with and without the Pareto principle. *Journal of Economic Theory* 29, 205–16.

Border, K.C. 1984. An impossibility theorem for spatial models. *Public Choice* 43, 293–305.

Border, K.C. and Jordan, J.S. 1983. Straightforward elections, unanimity and phantom voters. *Review of Economic Studies* 50, 153–70.

Bordes, G., Campbell, D.E. and Le Breton, M. 1995. Arrow's theorem for economic domains and Edgeworth hyperboxes. *International Economic Review* 36, 441–54.

Bordes, G. and Le Breton, M. 1989. Arrovian theorems with private alternatives domains and selfish individuals. *Journal of Economic Theory* 47, 257–81.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bordes, G. and Le Breton, M. 1990. Arrovian theorems for economic domains: assignments, matchings and pairings. *Social Choice and Welfare* 7, 193–208.

Bossert, W. and Weymark, J.A. 2004. Utility in social choice. In *Handbook of Utility Theory*, vol. 2, ed. S. Barberà, P.J. Hammond and C. Seidl. Boston: Kluwer.

Broome, J. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Diamantaras, D. 1991. Envy-free and efficient allocations in large public good economies. *Economics Letters* 36, 227–32.

Donaldson, D. and Weymark, J.A. 1988. Social choice in economic environments. *Journal of Economic Theory* 46, 291–308.

Duggan, J. 1996. Arrow's theorem in public goods environments with convex technologies. *Journal of Economic Theory* 68, 303–18.

Ehlers, L. and Storcken, T. 2002. Arrow's theorem in spatial environments. Cahier No. 2002–03, Département de sciences économiques, Université de Montréal.

Fleurbaey, M. 2007. Social choice and just institutions: new perspectives. *Economics and Philosophy* 23, 15–43.

Fleurbaey, M. and Maniquet, F. 2008. Compensation and responsibility. In Arrow, Sen and Suzumura (2008).

Fleurbaey, M., Suzumura, K. and Tadenuma, K. 2005. Arrovian aggregation in economic environments: how much should we know about indifference surfaces? *Journal of Economic Theory* 124, 22–44.

Foley, D.F. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7, 45–98.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Hansson, B. 1968. Choice structures and preference relations. *Synthese* 18, 443–58.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and Organization: A Volume in Honor of Jacob Marschak*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Kalai, E., Muller, E. and Satterthwaite, M.A. 1979. Social welfare functions when preferences are convex, strictly monotonic, and continuous. *Public Choice* 34, 87–97.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Kolm, S.-C. 1972. *Justice et Equité*. Paris: CNRS.

Laffond, G. 1980. Révelation des preferences et utilités unimodales. Doctoral dissertation, Conservatoire National des Arts et Métiers.

Le Breton, M. 1986. Essais sur les fondements de l'analyse économique de l'inégalité. Doctoral dissertation, Université de Rennes 1.

Le Breton, M. and Sen, A. 1995. Strategyproofness and decomposability: weak orderings. Discussion Papers in Economics No. 95-04, Indian Statistical Institute, Delhi Centre.

Le Breton, M. and Sen, A. 1999. Separable preferences, strategyproofness, and decomposability. *Econometrica* 67, 605–28.

Le Breton, M. and Weymark, J.A. 2002. Social choice with analytic preferences. *Social Choice and Welfare* 19, 637–57.

Le Breton, M. and Weymark, J.A. 2008. Arrovian social choice theory on economic domains. In Arrow, Sen and Suzumura (2008).

Maskin, E.S. 1976. Social welfare functions for economics. Unpublished manuscript, Darwin College, Cambridge University.

Moulin, H. 1980. On strategy-proofness and single peakedness. *Public Choice* 35, 437–55.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pazner, E.A. and Schmeidler, D. 1978. Egalitarian equivalent allocations: a new concept of economic equity. *Quarterly Journal of Economics* 92, 671–87.

Redekop, J. 1995. Arrow theorems in economic environments. In *Social Choice, Welfare, and Ethics*, ed. W.A. Barnett, H. Moulin, M. Salles and N.J. Schofield. Cambridge: Cambridge University Press.

Roemer, J.E. 1993. A pragmatic theory of responsibility for the egalitarian planner. *Philosophy and Public Affairs* 22, 146–66.

Satterthwaite, M.A. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Satterthwaite, M.A. and Sonnenschein, H. 1981. Strategy-proof allocation mechanisms at differentiable points. *Review of Economic Studies* 48, 587–97.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sen, A. 1974. Informational bases of alternative welfare approaches: aggregation and income distribution. *Journal of Public Economics* 3, 387–403.

Serizawa, S. 2002. Inefficiency of strategy-proof rules for pure exchange economies. *Journal of Economic Theory* 106, 219–41.

Serizawa, S. and Weymark, J.A. 2003. Efficient strategy-proof exchange and minimum consumption guarantees. *Journal of Economic Theory* 109, 246–63.

Sprumont, Y. 1991. The division problem with single-peaked preferences: a characterization of the uniform rule. *Econometrica* 59, 509–19.

Sprumont, Y. 1995. Strategyproof collective choice in economic and political environments. *Canadian Journal of Economics* 28, 68–107.

Svensson, L.-G. 1983. Large indivisibilities: an analysis with respect to price equilibrium and fairness. *Econometrica* 51, 939–54.

Thomson, W. 1990. The consistency principle. In *Game Theory and Applications*, ed. T. Ichiishi, A. Neyman and Y. Tauman. New York: Academic Press.

Thomson, W. 2008. Fair allocation rules. In Arrow, Sen and Suzumura (2008).

Tinbergen, J. 1953. *Redelijke Inkomensverdeling*, 2nd edn. Haarlem: N.V. DeGulden Pers.

Weymark, J.A. 2004. Strategy-proofness and the tops-only property. Working Paper No. 04-W09, Department of Economics, Vanderbilt University.

Zhou, L. 1991. Inefficiency of strategy-proof allocation mechanisms in pure exchange economies. *Social Choice and Welfare* 8, 247–54.

How to cite this article

Bossert, Walter and John A. Weymark. "social choice (new developments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000503> doi:10.1057/9780230226203.1551

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social choice

Amartya Sen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article is a critical survey of the literature of social choice theory, first formalized by Kenneth Arrow in 1951. Social choice theory deals with the aggregation of some measure of individual welfare into a collective measure. It takes different forms according both to what is being aggregated (interests, judgements, and so on) and to the purpose of the aggregation. The methodology of social choice has greatly clarified a range of hitherto obscure problems.

Keywords

impossibility theorem; aggregation; Arrow, K.; Arrow's theorem; Bentham, J.; Bergson, A.; Black, D.; Borda rule; Borda, J.-C.; Chernoff condition; compensation tests; Condorcet, Marquis de; envy; fairness; independence of irrelevant alternatives; interpersonal utility comparisons; liberty; majority rule; maximin; Mill, J. S.; Pareto principle; positional rules; primary goods; Rawls, J.; revealed preference; rights; Robbins, L.; Samuelson, P.; single-peaked preferences; social choice; social welfare function; transitivity; voting paradoxes; welfare economics

Article

Social choice theory, pioneered in its modern form by Arrow (1951), is concerned with the relation between individuals and the society. In particular, it deals with the aggregation of individual interests, or judgements, or well-beings, into some aggregate notion of social welfare, social judgement or social choice. It should be obvious that the aggregation exercise can take very different forms depending on exactly what is being aggregated (e.g., the personal interests of different people, or their moral or political judgements) and what is to be derived on that basis (e.g., a measure of social welfare, or public decisions regarding what is to be done or what outcomes are to be accepted). The formal similarities between these exercises in the analytical format of aggregation should not make us overlook the diversities in the nature of the exercises performed (see Sen, 1977a, 1986). In fact, the axioms chosen for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
different exercises are often quite divergent, and the general conception of aggregation in social choice theory permits such variation.

1 Welfare economics and social choice

Although the origins of social choice theory – in one form or another – can be traced back at least two hundred years (Borda, 1781; Condorcet, 1785; Bentham, 1789), the formal theory of social choice was initiated by Kenneth Arrow (1951) less than four decades ago. Arrow drew on some existing notions of welfare economics. One concept of a *social welfare function* had been introduced by Bergson (1938). This was defined in a very general form indeed: as a real-value function W(.), determining social welfare, 'the value of which is understood to depend on all the variables that might be considered as affecting welfare' (p. 417). Such a social welfare function – swf for short – might be thought to be a real-valued function defined on X, the set of alternative social states. It is a bit more permissive to see a Bergson social welfare function as an ordering R of X (more permissive because not all orderings can be numerically represented).

Various uses to which a swf can be put in welfare economics were investigated, particularly by Samuelson (1947). His exercises made use of several criteria that a *swf* may be required to satisfy, including the Pareto criterion, demanding that unanimous individual preference over any pair of states should yield the corresponding social preference over that pair.

None of the conditions that Samuelson imposed on a *swf* for his exercises required any general specification of how the social ordering might change if *different* sets (strictly, *n*-tuples) of individual orderings were considered. If any *n*-tuple of individual preference orderings is called a 'profile', then Samuelson's exercises – and those considered by Bergson – were all 'single-profile' problems, without additional requirements of *inter-profile* consistency.

Arrow (1951) defined a social welfare function – henceforth *SWF* (to be distinguished from a Bergson–Samuelson *swf*) – as a functional relation specifying one social ordering R for any given *n*-tuple of individual orderings (Ri), with one ordering Ri for each person i: $R = f(Ri)$).

Note that if a Bergson–Samuelson swf is defined as a social ordering R (rather than as a real-valued function W(.)), then an Arrow SWF is a function the value of which would be a Bergson–Samuelson swf. Arrow's exercise, in this sense, is concerned with the way of arriving at a Bergson–Samuelson swf. Arrow proceeded to impose a few conditions that any reasonable SWF could be expected to satisfy. His 'impossibility theorem' (more formally called 'the General Possibility Theorem') shows that no SWF can satisfy all these conditions together. One of the conditions deals specifically with the multiple-profile characteristics of a *SWF*, viz., the independence of irrelevant alternatives (condition I). This requires that the chosen alternatives from any subset of social states must remain unaltered as long as the individual preferences over this subset remain unaltered, even though the individual preferences may have been revised over other subsets. Another condition is a weak version of the Pareto principle (condition P) which requires that unanimous *strict* preference over a pair must be reflected in the same strict preference for the society. Another requirement is that of unrestricted domain (condition U), which demands that the domain of the *SWF* must include all logically possible *n*-tuples of individual orderings, that is, the *SWF* should be able to specify a social ordering R no matter what the individual orderings happen to be. Finally, there is a condition of non-dictatorship (condition D), which demands that there is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
no individual such that if he or she prefers any *x* to any *y*, then *x* is socially preferred to *y*, no matter what the other individuals prefer.

One version of the ‘impossibility theorem’ of Arrow establishes that, if the set of individuals is finite and the number of distinct social states is at least three, then there is no social welfare function (SWF) satisfying conditions U, I, P, and D.

This result has been the starting point of much of modern social choice theory. Even though the focus has somewhat shifted in recent years from impossibility results to other issues, there is no question at all that Arrow's formulation of the social choice problem in presenting his ‘impossibility theorem’ laid the foundations of social choice theory as it has evolved.

Two *interpretational* issues may be sorted out first before formal social choice theory is considered for a general examination. The first issue concerns the interpretation of ‘social preference’. As has already been remarked, the nature of the social choice exercise can vary in many different ways, and one source of variation is the nature of the end-point that is sought (in particular the interpretation of R). Consider the relation of strict social preference *xPy*. It can be given different interpretations depending on the nature of the exercise. For example, *xPy* can stand for the judgement that society is better off in state *x* than in state *y*. Such a judgement can be the view of a particular individual (in his or her capacity as an aggregating judge), or the mechanical outcome of some institutional process of aggregating judgements (e.g., the result of a voting procedure). Or, alternatively, *xPy* can stand for the statement that, in the choice exactly over the pair (*x*, *y*), *x* alone must be chosen. A further alternative is to interpret *xPy* as the requirement that *y* must not be chosen from any set which contains *x* (whether or not it contains any other alternative). These and other interpretations give different views of ‘social preference’, and careful attention has to be paid to the nature of the exercise depending on the interpretation given. Although Arrow's ‘impossibility theorem’ and similar results apply to *all* the interpretations (and here there is a genuine economy in the general axiomatic method), extensive variations in the relevance of the results to different types of problems must be recognized.

Second, a different source of variation relates to the interpretation of the individual preference orderings. The individual ordering can stand for the ranking of personal well-being, and if so, the exercise is one of *well-being aggregation*. An example may be found in arriving at overall judgements of the well-being of the community based on rankings of individual well-beings. To take a very different type of example, in making a committee decision, the different judgements of the members of the committee may be aggregated together in an overall judgement or an overall decision, and that exercise is one of *judgement aggregation*. This is not to deny that the judgements of members of the committee may, in fact, be influenced by their individual interests, but the nature of the exercise is primarily that of aggregating the possibly divergent judgements of the members of a committee to arrive at an over-all committee view. In some other exercises, for example, in electing a candidate or a member of Parliament or a Mayor, the individual votes may well reflect a clear-cut mixture of individual interests and political beliefs, so that the exercise may have features of interest aggregation as well as judgement aggregation. Once again, it is worth emphasizing that while the formal results such as Arrow's ‘impossibility theorem’ apply to each of the interpretations, the exact substantive content of the result would depend on the particular interpretation chosen.

The specific context of Arrow's exercise was that of supplementing the work of Bergson and Samuelson in deriving social welfare functions for welfare-economic studies. If the individual orderings are interpreted as utility rankings of individuals, and social preferences interpreted as a judgement of social

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
welfare, the Arrow theorem asserted that there is no way of combining individual utility orderings into an overall social welfare judgement satisfying the four specified conditions. The result can be easily translated into a choice-theoretic framework by adopting a choice-based notion of ‘social preference’, e. g., the ‘base’ relation or the ‘revealed preference’ relation of social choice. On this interpretation, it would appear that there is no way of arriving at a social choice procedure specifying what is to be chosen (over pairs, or over larger subsets), satisfying the appropriately interpreted (i.e., in terms of choice) conditions specified by Arrow (see Blair, Bordes, Kelly and Suzumura, 1976; and Sen, 1977a, 1982).

This is, of course, a negative result. A great deal of social choice theory, at least in the early stage, consisted of trying to deal with this result, suggesting different interpretations, different extensions, different ways of ‘resolution’, and other responses to the ‘impossibility’ identified by Arrow.

The main lines of response to Arrow's result will be examined presently. It is, however, worth emphasizing that the ‘impossibility theorem’ must not be seen as primarily a ‘negative’ achievement. The axiomatic method, as used here, can take a set of axioms which look reasonable enough and then derive some joint implications of these axioms. If the implications are unacceptable, the axioms can be re-examined. Interpreted thus, the axiomatic method is a procedure for assessing a set of principles reflected in the axiom structure, and it persistently invites attention to the content and acceptability of the axioms chosen.

Arrow's impossibility result brought out the unviability of the welfare-economic structure that had emerged in the discussion preceding the birth of modern social choice theory. After the rejection of ‘interpersonal comparisons’ of well-being (on this see Robbins, 1932, 1938), it was increasingly accepted that social choices or social judgements would have to be based on individual utility orderings without interpersonal comparisons. The four axioms chosen by Arrow make a good deal of sense in that context, and had indeed been used – formally or informally – in the pre-existing literature. What Arrow's theorem demonstrates is the unviability of that structure. The primary impact of Arrow's initial result was to demand that the entire question of the basis of social welfare judgement be re-examined. While this is, in one sense, a negative result, in another sense it opened up various ways of reformulating the social choice problem as a result of the demonstrated unviability of the pre-existing approach. The later literature in social choice theory bears testimony to the fact that many of these ways have been found to be both feasible *and* useful. Several of these approaches will be examined later on in this note, but the positive contribution of the negative impossibility result presented by Arrow has to be kept in view to see these advances in their appropriate perspective.

2 Variations and extensions of Arrow's impossibility result

The literature of social choice theory contains a large number of theorems that take the form of presenting variations of the type of impossibility identified by Arrow. In fact, Arrow himself has presented several distinct versions. The one presented in 1951 contained a formulational error, which was identified and corrected by Blau (1957). A later version, which was the one cited in the last section, is presented in Arrow (1963). Various other variations can be found in the literature, modifying one condition or another, and presenting impossibility results based on conditions that are more demanding in some respects and less demanding in others (see particularly Blau, 1957; Murakami, 1968; Pattanaik,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1971; Fishburn, 1973, 1974; Hansson, 1976; Brams, 1976; Plott, 1976; Kelly, 1978; Monjardet, 1979; Roberts, 1980b; Chichilnisky, 1982; McManus, 1982; Suzumura, 1983; Hurley, 1985; Nitzan and Paroush, 1985, among many others). Each of Arrow's conditions has been modified in one way or another in these different variants.

One particular variation, which is both illuminating and simple, relates to results presented by Wilson (1972) and Binmore (1976). This shows that given unrestricted domain and independence, all permissible social welfare functions will either have social rankings ‘imposed’ irrespective of individual preferences, or have a dictator, or have a ‘reverse dictator’ (a person such that whenever he prefers *x* to *y*, society prefers *y* to *x*). Arrow's impossibility theorem can be seen as a corollary of this when the Pareto principle is also demanded, since Pareto will eliminate both ‘imposition’ and ‘reverse dictatorship’, leaving dictatorship as the only remaining possibility.

One line of variation that has been very extensively investigated is that of weakening the demand of ‘collective rationality’, i.e., relaxing the requirement that social choice must be based on a social *ordering* (complete, reflexive and transitive). The ‘range’ of the social welfare function is supposed to include only social orderings, and the proposed relaxation weakens that demand. It can be shown that if the transitivity of only *strict* social preference is demanded (without also demanding the transitivity of social indifference), then all of Arrow's conditions can be simultaneously satisfied and there is no impossibility (see Sen, 1969, 1970; see also Schick, 1969).

This condition of transitivity of strict preference, formally called ‘quasi-transitivity’, when imposed on social preferences, for a social welfare function satisfying unrestricted domain, independence and the Pareto principle, has the effect of confining social choice procedures to ‘oligarchies’ (this result was first presented in an unpublished paper by Gibbard, and reported in Sen, 1970). An oligarchic group consists of a set of individuals such that if any one of them prefers any *x* to any *y*, then *x* must be taken to be socially preferred to or indifferent to *y*, and if all the individuals in that group unanimously prefer *x* to *y*, then *x* must be taken to be socially strictly preferred to *y*. One extreme case of oligarchy is that of a one-person oligarchy, which corresponds to Arrow's dictatorship. The other extreme makes the oligarchy group include every individual in the community. In this latter case, the fact that all of them taken together happen to be decisive is not remarkable (it follows in fact immediately from the Pareto principle). But it also gives every member of the community ‘veto’ power in the sense that whenever anyone prefers any *x* to any *y*, this precludes the possibility of *y* being socially strictly preferred to *x*, and this has the effect of producing lots of social indifferences all around (see Sen, 1969).

This ‘veto’ result can be obtained even without demanding quasi-transitivity of social preference, by supplementing the weaker demand of ‘acyclicity’ (i.e. the absence of strict preference cycles) with some other conditions, as has been investigated by Mas-Colell and Sonnenschein, 1972; Schwartz, 1972, 1986; Guha, 1972; Brown, 1974, 1975; Blau, 1976; Blau and Deb, 1977; Monjardet, 1979, and others. Partial ‘veto’ results have been established with still weaker conditions (see Blair and Pollak, 1982; Kelsey, 1984).

On a somewhat different line of investigation, it has been possible to somewhat weaken the condition of full transitivity of social preference and still retain exactly the impossibility identified by Arrow, i.e., dictatorship following from conditions U, I and P. This is easily done by replacing the requirement of ordering by that of having ‘semi-orders’, but it can be relaxed further (see Blair and Pollak, 1979; Blau,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1979).

These investigations of relaxation of collective rationality have not been confined only to the weakening of transitivity of social preference. It is possible to drop the requirement of completeness of social preference, permitting the possibility that many pairs of states *x* and *y* may be not socially rankable vis-à-vis each other, and still the impossibility result may survive if the Arrow conditions are correspondingly redefined to cover this case with sufficient richness of social ranking, in line with Arrow's original motivation (see Barthelemy, 1983; Weymark, 1983).

Yet another line of investigation consists of relaxing the requirement that social choice must be ‘binary’ in nature, in the sense of the choice function being representable by a binary relation (whether or not that binary relation R is called social preference). Some positive possibility results were obtained by Schwartz (1970, 1972), Fishburn (1973), Plott (1973), Bordes (1976), and Campbell (1976), by demanding consistency conditions on choice functions that are weaker than the requirement of binary choice.

One way of doing this is to convert preference cycles into indifference classes. For example, take the case of the so-called ‘paradox of voting’ in which person 1 prefers *x* to *y*, and *y* to *z*, person 2 prefers *y* to *z*, and *z* to *x*, and person 3 prefers *z* to *x*, and *x* to *y*. In this case the majority rule yields *x* being socially preferred to *y*, *y* being socially preferred to *z*, and *z* being socially preferred to *x*, producing a strict preference cycle, with no alternative that is not beaten by another alternative. If, in this case, all the three alternatives are declared socially indifferent, by converting the cycle into an indifference class, then much of Arrow's requirements can be retained. However, one type of consistency will certainly be violated by this formulation of social choice, to wit, relating social choice over the pair (*x*, *y*) to that over the triple (*x*, *y*, *z*). Due to the majority preference for *x* over *y*, and the demand of the ‘independence’ condition (I) that individual preferences only over (*x*, *y*) be considered when choosing over this pair only, *x* must be chosen and *y* rejected in the choice over the pair (*x*, *y*). But in the choice over the triple (*x*, *y*, *z*), even state *y* can be selected as a member of the indifference class, when the majority cycle is converted into indifference. The choosability of *y* from the larger set (*x*, *y*, *z*), and its non-choosability from the smaller set (*x*, *y*) contained in the larger set, does violate a standard condition of consistency of choice, variously called Property α or the ‘Chernoff condition’, or standard ‘contraction consistency’. In the absence of this consistency, the choice function cannot possibly be represented in a binary form, i. e., through a binary relation R such that the choices correspond to the R-maximal elements (with R being derived from the internal properties of choice, e.g., *x*R*y* when *x* is chosen in the presence of *y*). But this condition (Property α) is, in fact, much weaker than the requirement that the choice function be binary.

Since binariness may not in itself be a compelling requirement, the plausibility of this line of resolution of Arrow's impossibility depends on the value of the consistency conditions that these ‘solutions’ may actually satisfy. By imposing some relatively appealing consistency conditions, it can be shown that the dictatorship result of Arrow, and the other related results regarding oligarchy, veto power, etc., derived in the binary framework can reappear easily enough in non-binary choice as well (see Blair, Bordes, Kelly and Suzumura, 1976; Sen, 1977a). It can also be pointed out that even when the social choice procedures do not satisfy binariness in the sense of being *representable* by a binary relation, there would, of course, be binary relations that are *generated* by the choice function. For example, the ‘revealed preference’ relation (*xRy* if *x* is chosen in the presence of *y*) will be defined by any choice

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
function, since the choice of any alternative (say, *x*) from any set containing another alternative (say, *y*), will yield the deduction *x*R*y*. The issue of binariness arises when it is further demanded that is what is chosen from each subset consists exactly of the R-maximal elements of that set, according to that binary relation R. It can be shown that binariness in this form demands much the same thing whether we concentrate on the 'revealed preference' relation, or the 'base relation' (the latter being defined as: *x*R*y* if and only if x is chosen specifically from the pair *x*, *y*). Binariness according to the 'revealed preference' relation is equivalent to that according to the 'base' relation (see Herzberger, 1973).

Although the demands of binariness provide one way of re-establishing Arrow's 'impossibility' results, a different way is not to demand binariness at all, but to translate all of Arrow's demands to one specific binary relation generated by social choice, e.g., the 'revealed preference', or the 'base' relation. The Arrow theorem will hold exactly in the same way for each such interpretation of R, provided the Arrow conditions are correspondingly reinterpreted. In this sense binariness is not a central issue in the inescapability of the 'impossibility' result of Arrow (on this see Sen, 1977a, 1982; on related matters, see Grether and Plott, 1982; Suzumura, 1983; and Matsumoto, 1985).

One general conclusion that seems to emerge from these investigations of relaxation of 'collective rationality' properties is the durability and robustness of Arrow's 'impossibility' result. The tension between different types of principles seems to survive various ways of relaxing these principles, and the particular 'impossibility theorem' of Arrow is a centre piece of a much broader picture. Demands on consistency of social choice can be dramatically changed without the 'impossibility' features disappearing.

3 Domain restrictions

When presenting his impossibility result, Arrow had suggested the possibility that a resolution might be found in terms of restricting the domain of the social welfare function (no longer requiring that it works no matter what the individual preferences happen to be). It is, of course, clear that there are many preference combinations for which such procedures as the method of majority decision will yield perfectly consistent social choice. Arrow (1951) himself had explored a particular type of restriction of individual preferences called 'single-peaked preferences' (earlier discussed by Black, 1948). This corresponds to the case in which the alternatives can be so arranged on a line that everyone's intensity of preference has one peak only, i.e., the preference drops monotonically as we move from left to right, or rises monotonically, or it rises to a peak and then falls. Arrow showed that if individual preferences are single-peaked and the number of voters is odd, then majority decision will yield transitive social preference.

The positive possibility result for single-peaked preferences can be generalized in many different ways. It can be shown that individual preferences being single-peaked in every triple of alternatives is equivalent to the condition that in every triple there is one state such that no one regards it to be 'worst'. It can be shown that a similar agreement on some alternative being regarded as not 'best' would do, and so would an agreement on some alternative being not 'medium'. Altogether, this sufficiency condition is called 'value restriction', and the particular type of agreement (whether 'not best', or 'not worst', or 'not medium') may vary from triple to triple (see Sen, 1966). Also the requirement of oddness of the number of voters can be eliminated if the demand is not for full transitivity of social preference, but only the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
absence of preference cycles and the existence of a majority winner (Sen, 1969). In this general line of investigation, necessary and sufficient conditions for transitivity as well as acyclicity of majority decisions (i.e., for the existence of a majority winner) have been identified by Inada (1969, 1970) and Sen and Pattanaik (1969). The former requirement is called ‘extremal restriction’. The relationships among these and other related conditions are discussed in Inada (1969), Sen (1970), Pattanaik (1971), Fishburn (1973), Salles (1976), Slutsky (1977), Kelly (1978), Monjardet (1979), Blair and Muller (1983), Larsson (1983), Suzumura (1983), Dummett (1984), Arrow and Raynaud (1986), Jain (1986) among many others.

On a different line of analysis, domain conditions can be specified not only in terms of general qualitative correspondence of individual preferences, but also in terms of number-specific requirements on the distribution of voters over the different preferences (see particularly Plott, 1967; Tullock, 1967, 1969; Saposnik, 1975; Slutsky, 1977; Gaertner and Heinecke, 1978; Grandmont, 1978; Dummett, 1984). These domain conditions all deal specifically with the method of majority decision, but the problem can be investigated more generally. Domain conditions for other voting rules have been investigated (see, for example, Pattanaik, 1971). More recently, the necessary and sufficient domain conditions for the existence of any social welfare function satisfying all of Arrow's other conditions (whether or not based on counting majority) have been investigated (see Kalai and Muller, 1977, and Maskin, 1976; see also Dasgupta, Hammond and Maskin, 1979; Kalai and Ritz, 1980, and Chichilnisky and Heal, 1983). These domain restrictions are indeed very demanding, and counterexamples can be found without any loss of plausibility in terms of real-life situations (see particularly Kramer, 1973). But if these restrictions are not fulfilled, then there is no general ‘solution’ to be found in opting for the majority rule, or some other rule like that. Indeed, it can be shown for the majority rule that the cycles that may be generated may well be extremely extensive, yielding ‘total cycles’ involving all social states (see Schofield, 1978; McKelvey, 1979). This line of investigation too, like the one on collective rationality (discussed in the last section), yields rather discouraging results. No general solution of impossibility theorems of the type presented by Arrow can be easily found by opting for a rule like the majority decision, hoping that the domain conditions will be somehow satisfied.

In many economic decisions, it is quite straightforward to see that these conditions will indeed be all violated. However, when the number of alternatives happen to be small, and when there is complex balancing of conflicting considerations, as in many political contexts (elections, committee decisions over rival proposals, etc.), there might possibly be some room for optimism. If cycles or other types of intransitivities turn out to be rather rare in these cases, then the approach of domain restriction may well offer some help. In contrast, in welfare-economic problems, that hope is very limited.

Indeed, if we take such a simple social-welfare problem as the division of a given cake between three or more individuals, with each person voting according to his or her own share of the cake, it can be easily shown that there will indeed be majority cycles. But it is worth noting in this context that the method of majority decision is not particularly appropriate for such economic problems anyway. Any distribution of the given cake can be improved by choosing one of the persons (even the poorest one) and dividing a part of his or her share for the benefit of all others, thereby producing an ‘improvement’ according to the majority rule. Indeed, we can go on ‘improving’ the distribution in this way, following the majority ranking procedure, making the worst-off individual more and more worse off all the time. As a criterion for welfare-economic judgement, majority rule is, in fact, a non-starter. The recognition of this fact

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
makes it less tragic that majority cycles will tend to arise easily enough in many economic problems involving distributional variations. The majority rule would not have offered any ‘real solution’ to the task of making social welfare judgements in this type of economic problems even if it had been fully consistent. It is more in the context of political decisions involving a few diverse alternatives (rather than welfare-economic judgements in general) that majority rule and related decision procedures have some prima facie plausibility. It is, thus, of some interest that it is in the context of these problems that the domain conditions investigated by the social choice literature are of direct relevance and offer some hope.

4 Manipulability and implementation

A different type of problem for voting procedures arises from the possibility of ‘manipulation’ of the decision mechanism by the voters voting ‘dishonestly’. A voting procedure is ‘manipulable’ when it is in the interest of some voter for some set of individual preferences to vote differently from his or her sincere preference.

The ubiquity of the possibility of manipulation had been conjectured for a long time, but it was established only recently in a remarkable theorem first presented by Gibbard (1973), and then by Satterthwaite (1975). A similar result, and a pointer to positive possibility if the conditions are relaxed, was presented by Pattanaik (1973.) The Gibbard–Satterthwaite manipulation theorem establishes that every non-dictatorial voting scheme with at least three distinct outcomes must be manipulable.

Gibbard established this theorem as a corollary of another one dealing with ‘game forms’ in general, of which voting schemes happen to be special cases. A game form does not restrict the strategies to be chosen by the individuals to the orderings of social states (i.e., to ‘ballots’), and each person's strategy set can be any set of signals. Gibbard established that no non-dictatorial game form with at least three possible outcomes can be ‘straightforward’ (a concept first used by Farquharson, 1956), in the sense that each person would have a *dominant* strategy (i.e., a best strategy with respect to his ordering of the outcomes, irrespective of what the strategies of others might be). Thus for every non-dictatorial game form of this type, there is at least one person who does not have a dominant strategy for some preference ordering of outcomes. From this the manipulability theorem follows immediately. If a voting scheme were *non*-manipulable, then everyone would have had a dominant strategy, viz., recording his or her *true* preference irrespective of what others do. Since the existence of dominant strategies is disestablished, so is the existence of *honest* dominant strategies.

Various variations of this discouraging result and some avenues of escape have been investigated in the literature, which is quite vast (but excellent discussions can be found in Barbera, 1977; Pattanaik, 1978; Laffont, 1979; Peleg, 1984; Brams and Fishburn, 1983; Moulin, 1983; and Jain, 1986).

The focus on ‘honest’ revelation of preferences has gradually given way to discussions of equilibrium and of implementation (for an early pointer in this direction, see Dummett and Farquharson, 1961). If the object of the exercise is effectiveness in the sense of getting an appropriate outcome (rather than seeking honesty as such), then the thing to investigate is indeed the existence of an effective mechanism rather than a ‘strategy-proof’ one. If, for example, a non-strategy-proof mechanism yields an equilibrium of dishonest behaviour that produces the same outcome as honest revelation of preferences would, then that mechanism could well be regarded as successful in terms of effectiveness.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The shift in attention towards equilibrium and implementation has opened up new lines of investigation, which are being explored (see particularly Dutta and Pattanaik, 1978; Dasgupta, Hammond and Maskin, 1979; Sengupta and Dutta, 1979; Peleg, 1984; Moulin, 1983). The implementation literature also links up with more standard problems of public economics, in which it has received attention in a somewhat different but related form (see, particularly, Groves and Ledyard, 1977; Green and Laffont, 1979; Laffont, 1979).

5 Information: utility, compensations and fairness

The alleged impossibility of interpersonal comparisons of utility was entirely accepted in the early works on social choice theory. Arrow's (1951) format gave no room to interpersonally comparable utility information, and indeed took utility information in the form of non-comparable ordinal utility rankings. This was entirely in line with the dominant position of welfare economics at that time. Even though there were formats for interpersonal comparisons of utility suggested in some contributions to welfare economics (see particularly Vickrey, 1945, and Harsanyi, 1955), these suggestions were not followed up in the formal social-choice-theoretic literature until much later.

There had been earlier attempts to by-pass the need for utility comparisons by using the notion of compensation tests (e.g., whether the gainers *could* compensate the losers), and this had led to the identification of problems of internal consistency as well as of cogency (see Kaldor, 1939; Hicks, 1939; Scitovsky, 1941; Little, 1950; Samuelson, 1950; Baumol, 1952; Gorman, 1953; Graaff, 1957). The problem of cogency is perhaps deeper, in some ways, than that of consistency. To make sure that gainers have gained so much that they can compensate the loser does, of course, have some immediate plausibility as a requirement. However, the relevance of the compensation tests suffers from the following limitation. If compensations are not paid, then it is not clear in what way the situation can be taken to be an improvement (since those who have lost may well be a great deal poorer, needier or more deserving – whatever our criteria for such judgments might be – than the gainers). And if compensations *are* in fact paid, then *after* the compensation what we observe is a Pareto improvement, so that no compensation tests are in fact needed. Thus, the compensation approach suffers from having to face a choice between being unconvincing or being irrelevant.

Another approach that by-passes the need for interpersonal comparisons proper is that of ‘fairness’, presented first by Foley (1967). Here a person's advantage is judged by comparing his bundle of goods with those enjoyed by others, and a situation is called ‘equitable’ if no individual prefers the bundle of goods enjoyed by another person to his own. If an allocation is both Pareto optimal and equitable then it is called ‘fair’. (There is some non-uniformity of language in the literature, and sometimes the term ‘fair’ has been defined simply as ‘equitable’, e.g., in Feldman and Kirman, 1974 and Pazner and Schmeidler, 1974.) This approach has been pursued by a number of authors (such as Kolm, 1969; Schmeidler and Vind, 1972; Varian, 1974, 1975; Goldman and Sussangkarn, 1978; Archibald and Donaldson, 1979; Crawford, 1979; Crawford and Heller, 1979; Svensson, 1980; Champsaur and Laroque, 1982; Suzumura, 1983, and others). There are interesting problems of the existence of fair allocations and of the consistency of fairness with other principles.

It should be noted that the comparisons involved in the calculus of ‘fairness’ are not interpersonal ones,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
but in fact comparisons of different positions that the same individual might occupy (e.g., having commodity bundles), as it is evaluated by the given person. The criterion of ‘non-envy’ does clearly have some appeal, even though it can be argued, that our deprivations may be related not only to other people's commodity bundles but also to non-commodity features of their advantage. For example, a person with a disability may well prefer to be in somebody else's position without that disability, but that is not the same thing as envying that other person's commodity bundle. As such, it could be argued, that the informational base of the fairness calculus is fundamentally limited.

Another difference between the ‘fairness’ approach and the standard social-choice-theoretic procedures relates to the more limited aim of the former. As Varian (1974) puts it, the fairness criterion in fact limits itself to answering the question as to whether there is a ‘good’ allocation (pp. 64–5). It is certainly true that social choice theory has been abundantly more ambitious, perhaps unrealistically so. On the other hand, it can be argued, that even the features of ‘goodness’ identified by the approach of fairness (e.g., equitability with efficiency) may often fail to be satisfied by any feasible allocation at all, so that the question of ranking the ‘non-good’ allocations is not really avoidable. In addition, it can be argued that insofar as the foundation of the ‘fairness’ approach is based only on comparing the commodity bundles of different persons without going further into the relative advantages enjoyed by the persons (taking everything into account), the criterion of ‘goodness’ used in the ‘fairness’ literature is itself rather a limited one. It is perhaps for these reasons that the use of interpersonal comparisons of well-being in social choice theory (in the literature on social welfare functionals, to be discussed presently) has tended to aim at going a great deal further than the ‘fairness’ literature was programmed to achieve.

6 Social welfare functionals and interpersonal comparisons

The empirical problem of obtaining information on interpersonal comparisons of utility has to be distinguished from the formal problem of accommodating such information within the structure of social choice theory. The format of social welfare functions used by Arrow, and the related formats of collected choice rules (involving such various forms as social decision functions, social choice functions, etc.), make no provision for any utility information finer than that of non-comparable individual orderings. One way of extending that framework is to permit the use of more utility information, through what have been called ‘social welfare functionals’ (SWFL): $R = F(\{U_i\})$. For each set (strictly, *n*-tuple) of utility functions $U_i,\ldots,U_n$ (one function per person), the social welfare functional F determines exactly one social ordering R. However, since utility functions can be nominally varied through alternative presentations without involving any ‘real’ change (e.g., doubling all the utility numbers), any social welfare functional has to be combined with some ‘invariance’ requirement. If two utility *n*-tubles (U_i) and (U_i^*) are judged to be informationally equivalent, differing from each other only in representation, then $F(\{U_i\}) = F(\{U_i^*\})$. The assumed structure of measurability and interpersonal comparability of utilities can be incorporated through specifying these invariance requirements (see Sen, 1970, 1977b; d'Aspremont and Gevers, 1977; Roberts, 1980a; Blackorby, Donaldson and Weymark, 1984).

Arrow's social welfare function is a special case of a social welfare functional with the invariance requirement corresponding to ordinal non-comparability (i.e., if one *n*-tuple of utility functions is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
replaced by another obtained from the first by taking positive, monotonic transformations of each utility function – not necessarily the same for all – then the social ordering R determined by the first *n*-tuple will also be yielded by the second). It is obvious that Arrow's ‘impossibility theorem’ can be translated in the format of social welfare functionals with *ordinal non-comparability*. More interestingly, this result can be generalized to the case of *cardinal non-comparability* also. When individual utilities can be cardinally measured but not in any way interpersonally compared, the same impossibility result continues to hold (see Sen, 1970). On the other hand, introducing interpersonal comparability without cardinality (i.e., using ordinal comparability only) does resolve the Arrow dilemma, and various possible SWFLs exist fulfilling all of Arrow's conditions in this case. An example is provided by Rawls's maximin rule (or the lexicographic version of it), defining these exercises in terms of utility comparison, rather than that of indices of ‘primary goods’, as in Rawls's own framework.

Richer utility information can be systematically used to admit various social choice procedures not admissible in Arrow's framework. The use of various axioms to characterize particular rules utilizing richer utility information can be found in an influential and important contribution by Suppes (1966). In the recent years the more formal frameworks of social choice theory (in particular, that of SWFLs) have been extensively used to derive axiomatically a number of standard aggregation procedures, such as the Rawlsian lexicographic maximin, utilitarianism, and some others (see particularly Hammond, 1976, 1977, 1979; Strasnick, 1976; d'Aspremont and Gevers, 1977; Arrow, 1977; Sen, 1977b; Deschamps and Gevers, 1978, 1979; Maskin, 1978; Gevers, 1979; Roberts, 1980a; Blackorby, Donaldson and Weymark, 1984; d'Aspremont, 1985). While these results are formal and do not address the question of the empirical content of interpersonal comparisons of utility (though this too is discussed by Hammond, 1977), the axiom structures have been related to various empirical insights thrown up by the substantive literature on interpersonal comparisons.

One format that has also been investigated relates to the intermediate possibility of making *partial* interpersonal comparisons of utilities. Various formal structures of partial comparability and partial cardinality have, in fact, been investigated in the social-choice-theoretic literature (see Sen, 1970; Blackorby, 1975; Fine, 1975; Basu, 1979; Bezembinder and van Acker, 1979). This is a less ambitious approach, admitting that not all types of interpersonal comparisons may be possible, and such comparability may be at best partial, with many undecided cases. Nevertheless *some* definite results can be obtained even on the basis of the incomplete structures.

Various other informational frameworks involving richer utility data can be and have been investigated, and some of them lend themselves to fruitful social-choice-theoretic use. One of the structures that need further investigation is the problem of combining *n*-tuples of ‘extended orderings’ reflecting each person's interpersonal comparisons. These are ordinal structures, but instead of there being one interpersonal comparison covering all the individuals in the different possible positions, this starts with the set of interpersonal comparisons made by different individuals (one ‘extended ordering’ per person), and addresses the problem of aggregation in that framework. Some interesting results in this area have been obtained (see Hammond, 1976; Kelly, 1978; Suzumura, 1983; Gaertner, 1983). The task, however, is rather a difficult one, since the information to marshall is extremely extensive, and progress in this area has tended to be rather slow. On the other hand, since social choice theory has to be concerned with the problem of combining different persons’ possibly divergent views, that ‘extended’ problem certainly has a good deal of relevance and potential importance.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

7 Liberty and rights

The informational limitations of the early social-choice-theoretic structures have led to responses in the later literature not only in the form of enriching the utility information (by the use of such structures as social welfare functionals, SWFL), but also that of making more systematic use of *non-utility* information. One of the areas that has been investigated in this context is that of rights in general and of liberty in particular. Liberty can be an important consideration in matters of social choice, but it cannot be adequately captured in terms of utility information, however rich it might be. If it is asserted that a person should be free to do what he or she likes in certain purely personal matters, that assertion is based on the non-utility characteristics of the ‘personal nature’ of these choices, and not primarily on utility considerations. As John Stuart Mill (1859) had argued, even if others might be offended by someone's personal behaviour in such matters as religious practice, it would not be appropriate to count the disutility of the offended in the same way as the utility of the person whose freedom of religious practice is under consideration. Various notions of ‘protected spheres’, ‘personal domains’, etc., have been formalized in the social-choice-theoretic literature in specifying domains of personal liberty.

One of the results obtained in this field that has led to a great deal of controversy concerns the conflict between the Pareto principle and certain minimal conditions of liberty when imposed on a social choice framework with unrestricted (or a fairly wide) domain. The ‘impossibility of the Paretian liberal’, presented in Sen (1970), has led to a variety of responses, including extensions, disputations, and suggestions of different ways of ‘resolving’ the conflict (see Ng, 1971; Batra and Pattanaik, 1972; Gibbard, 1974; Blau, 1975; Seidl, 1975; Campbell, 1976; Kelly, 1976, 1978; Aldrich, 1977; Breyer, 1977; Ferejohn, 1978; Karni, 1978; Suzumura, 1978, 1983; Mueller, 1979; Barnes, 1980; Bernholz, 1980; Breyer and Gardner, 1980; Breyer and Gigliotti, 1980; Fountain, 1980; Gardner, 1980; McLean, 1980; Weale, 1980; Baigent, 1981; Gaertner and Krüger, 1981; Gärdenfors, 1981; Hammond, 1981; Schwartz, 1970, 1972, 1986; Sugden, 1981, 1985; Austen-Smith, 1982; Levi, 1982; Krüger and Gaertner, 1983; Basu, 1984; Kelsey, 1985; Wriglesworth, 1985; Coughlin, 1986; Elster and Hylland, 1986; Gaertner, 1986; Riley, 1986; Webster, 1986, among others). The literature is vast, and covers issues of political compatibility, moral cogency and strategic consistency; it has been critically surveyed and assessed by Suzumura (1983) and Wriglesworth (1985). Various alternative formulations of liberty, in terms of social judgments, social decisions and social institutions can be shown to yield corresponding impossibility results (see Sen, 1983).

It is not really surprising that conditions of liberty or rights which make essential use of non-utility information may clash with exclusively utility-based principles, such as the Pareto principle. Non-utility considerations cannot be immovable objects if utility considerations, even in a rather limited context (as in the Pareto principle), are made into an irresistable force. One role of this type of impossibility result lies in pointing to the possibility that utility data may not be informationally adequate for social judgement or social choice, even when the utility information comes in the most articulate and complete form. Other lessons have also been suggested, and each interpretation has also been substantively disputed.

While impossibility results like this have received good deal of attention, relatively little effort has so far been spent on investigating the positive implications of various theories of rights, liberties and freedom,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in the general area of social choice. The need for caution in formulating the demands of liberty because of problems of internal consistency has in fact been investigated. But the more general question of developing a fruitful and positive theory of rights and liberty within the general structure of social choice theory has not yet been much investigated.

8 Independence and neutrality

The independence of irrelevant alternatives, used by Arrow, plays a major part in the social choice formats in the Arrovian tradition. It is also crucial for Arrow's impossibility theorem. The nature, implications and acceptability of the independence condition have been subjected to a good deal of critical examination in the literature (see particularly Gärdenfors, 1973; Hansson, 1973; Ray, 1973; Fine and Fine, 1974; Fishburn, 1974; Mayston, 1974; Young, 1974a, 1974b; Binmore, 1976; Kelly, 1978; Pattanaik, 1978; Moulin, 1983; Suzumura, 1983; Peleg, 1984; Hurley, 1985; Schwartz, 1986).

One of the objections that was originally raised about the relevance of Arrow's impossibility theorem related to the acceptability of the independence condition. Some authors (in particular Little, 1950 and Samuelson, 1967) argued that seeking inter-profile consistency in any form (including Arrow's ‘independence’ condition) is largely gratuitous. It was also argued that traditional welfare economics had never sought such a condition, and because of the crucial use of condition I, ‘Arrow's work has no relevance to the traditional theory of welfare economics, which culminates in the Bergson–Samuelson formulations’ (Little, 1950, pp. 423–5). ‘For Bergson,’ argued Samuelson (1967), ‘one and only one of the … possible patterns of individuals’ orderings is needed’ (pp. 48–9), and the question of inter-profile consistency does not arise.

In response to this line of objection, several ‘single-profile impossibility theorems’ in the spirit of Arrow's original theorem have been derived and discussed (see particularly Parks, 1976; Kemp and Ng, 1976; Pollak, 1979; Roberts, 1980b; Rubinstein, 1981; Hurley, 1985). These results depend on dropping inter-profile consistency in favour of rather strong intra-profile requirements, typically including some condition of single-profile neutrality, requiring that whatever combination of individual orderings be decisive for establishing xRy should be sufficient for establishing aRb if each individual ranking over (x, y) is the same as that over (a, b) in that given profile. The nature of the alternatives – whether x and y, *or* a and b – is, thus, not to make any difference, in relating individual preferences over particular pairs to social preference over those pairs, for any given profile of individual preferences.

These results are interesting, but it must be noted that the requirements on which they are based (e.g., of single-profile neutrality) are rather strong. Also the dictatorial result that follows from the other conditions is that of single-profile dictatorship, which might not be thought to be as objectionable as the existence of one inter-profile dictator who wins for every possible preference profile (as in Arrow's theorem).

No matter what one thinks of these single-profile impossibility results, it can certainly be argued that the original objection raised by Little and Samuelson about the relevance of inter-profile conditions for social choice theory is hard to sustain. Given the motivation underlying demands for consistency in the relation between individual preferences and social choice, it is not at all clear why such consistency requirements should be thought to be applicable only for a given profile and not between different profiles of individual preferences (no matter how close these profiles are in relevant respects).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
It could, of course, be argued that utility orderings (or preferences) are not an adequate informational base anyway for social choice, and if that position were taken, then the very idea of a social welfare function would have to be rejected in favour of some richer informational formulation, such as a social welfare functional SWFL. If, on the other hand, the motivation underlying the use of a social welfare function is accepted, and it is agreed that for a given preference *n*-tuple (i.e., a given profile), there is only one social ordering, then it is not clear why it would be thought to be perfectly okay that social preferences might change over a given pair when there is a change of individual preferences over some pair of alternatives quite unconnected with this particular one. The need for some interprofile consistency is hard to deny altogether. It could, of course, be argued that Arrow's particular inter-profile condition is not the appropriate one to use for inter-profile consistency, but that would not be an objection to inter-profile conditions as such, only to the particular formulation of Arrow's condition I. It should also be noted that there are other inter-profile conditions that can be used in order to generate impossibility results like Arrow's, without any use of condition I (see in particular Chichilnisky, 1982).

If Arrow's condition I is dropped, a number of alternative possibilities do, in fact, open up for social choice procedures. For one thing, ‘positional’ information can be used to rank alternative social states and to arrive at social choice. In fact, in an early contribution to social choice theory, Borda (1781) had used a decision procedure that violates condition I in arriving at overall rankings based on rank-order weights. This method – often called the Borda rule – is a special case of a general class of ‘positional’ rules. The general properties of ‘positional’ rules have been fruitfully investigated by Gärdenfors (1973) and Fine and Fine (1974), among others. The Borda ruling in particular has also received attention, and various particular rules have been investigated, critically examined and axiomatized (see Young, 1974a; Fishburn and Gehrlein, 1976; Hansson and Sahlquist, 1976; Gardner, 1977; Farkas and Nitzan, 1979; and Nitzan and Rubinstein, 1981, among others).

Positional rules take note of the fact that an alternative *x* preferred to another alternative *y* may be proximate to each other in a person's preference ordering without any other alternative in between, or may be separated by the existence of one or more other alternatives intermediate between the two. The rationale of positional rules relates to attaching importance to the placing of intermediate alternatives in individual preferences, which can be taken as suggesting that the gap between the two must be, other things given, larger. This argument is not entirely convincing. Many intermediate alternatives can be placed in a small interval, while large intervals may happen to be empty because of the contingent fact that there happens to be no other alternative that fits in just there. On the other hand, if information is thought to be extremely hard to get in social choice (a view that was certainly taken by Borda, 1781), then it is not entirely unreasonable to attach some significance to the fact that the placing of intermediate alternatives might be indicative of something. With some implicit assumption of uniformity of distribution of alternatives over the preference line (or some other suitable belief), the positional rules may have some clear rationale, and the Borda rule in particular might be particulary handy and useful. It is possible to use positional information also in the context of richer informational base, e.g., when interpersonal comparisons of utilities are permissible. Indeed ‘interpersonal positional rules’ may have some distinct advantage both (1) over rules that make non-positional use of interpersonally comparable individual orderings, and (2) over non-comparable positional rules. Such interpersonal positional rules may also be demonstrably more reasonable, in some contexts, than voting procedures like the majority rule which use neither interpersonal comparisons nor positional information (on this see Sen, 1977b;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Gaertner, 1983).

9 Concluding remarks

In understanding the literature of social choice theory it is important to bear in mind that while there are considerable analytical similarities between different problems tackled in this vast literature, the interpretations of the results and of their implications must take note of the particular nature of each of the substantively different problems. The axiomatic method, which has been so extensively used in the literature, offers enormous scope for efficient economy, but that economy will be self-defeating if the substantive differences are not carefully taken into account in interpreting exactly the content of the theorems derived. For example, the classic ‘impossibility’ result of Arrow may impose informational constraints that are much more reasonable in aggregating political preferences of different individuals over a small set of alternative proposals (or candidates) than in arriving at aggregative judgements of social justice taking note of conflicting individual interests over possible distributions of commodity vectors.

There is sometimes a temptation to see social choice theory as providing a particular ‘method’ of dealing with problems of aggregation. There is some truth in this diagnosis, in the sense that the discipline of axiomatic procedures has some exacting demands. On the other hand, the axioms can vary a great deal, and the interpretation of the axioms also will vary with the nature of the problems considered. The monolithic view of something called ‘the social-choice-theoretic approach’, which is often referred to both by those who wish to use it and those who wish to criticize it, may be deeply misleading. For some arguments on different sides on this question, see Elster and Hylland (1986).

There are, in fact, two different ways of seeing social choice theory. First, it is a field, and in this field there is scope for having different approaches. There are many problems of interpersonal aggregation, and in the broader sense, social choice theory is a field in which such aggregation – of different types – is studied. Second, social choice theory also provides a method of analysis, in which the insistence on the explicitness of axioms and on the clarity of assumptions imposes exacting formulational demands. Indeed, some of the more notable achievements of social choice theory have come from this insistence on explicitness and clarity (e.g., Arrow's own demonstration of the impossibility of combining a set of assumptions that were being implicitly invoked in the literature of the welfare economics of that period, including eschewing interpersonal comparisons of utility). While the second interpretation is a narrower one than the first, it is nevertheless broad enough to permit different types of axioms to be used, and different political, economic and social beliefs to be incorporated in the axiom structure. Neither interpretation would give any cogency to the search for ‘the social-choice-theoretic approach’.

One reason why social choice theory has received as much attention as it has in the last few decades relates to the importance of the field with which that theory has been concerned (and which characterizes that theory in the *broader* sense). Another reason has been the fruitfulness of making implicit ideas explicit, and of following their implications consistently and clearly. As a methodological discipline, social choice theory has contributed a great deal to clarifying problems that had been obscure earlier. While insistence on clarity at all costs has also some limitations (sometimes the narrowness of the axiom structure used in social choice theory has indeed been seen as a limitation), social choice theory has undoubtedly been a creative tradition among other methodological traditions that can be used to analyse economic, social and political problems involving group aggregation. The vast literature

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
surveyed in this article can be ultimately judged by what has been achieved in terms of clarifying the obscure and illuminating the unclear. Perhaps the successes have been rather mixed, but that fact is not surprising.

See Also

- Arrow's theorem
- constitutions, economic approach to
- public choice
- social choice (new developments)
- social welfare function

Bibliography

Aldrich, J. 1977. The dilemma of a Paretian liberal: Some consequences of Sen's theorem. *Public Choice* 30, 1–21.

Archibald, G.C. and Donaldson, D. 1979. Notes on economic equality. *Journal of Public Economics* 12, 205–14.

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Arrow, K.J. 1963. *Social Choice and Individual Values*. 2nd edn, New York: Wiley.

Arrow, K.J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review* 67, 219–25.

Arrow, K.J. and Raynaud, H. 1986. *Social Choice and Multicriterion Decision-Making*. Cambridge, Mass.: MIT Press.

Austen-Smith, D. 1982. Restricted Pareto and rights. *Journal of Economic Theory* 26, 89–99.

Baigent, N. 1981. Decomposition of minimal libertarianism. *Economic Letters* 7, 29–32.

Barbera, S. 1977. Manipulation of social decision functions. *Journal of Economic Theory* 15, 226–78.

Barnes, J. 1980. Freedom, rationality and paradox. *Canadian Journal of Philosophy* 10.

Barthelemy, J.P. 1983. *Arrow's theorem: Unusual domains and extended codomains*. In Pattanaik and Salles (1983).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Basu, K. 1979. *Revealed Preference of Governments*. Cambridge: Cambridge University Press.

Basu, K. 1984. The right to give up rights. *Economica* 51, 413–22.

Batra, R.M. and Pattanaik, P.K. 1972. On some suggestions for having non-binary social choice functions. *Theory and Decision* 3, 1–11.

Baumol, W.J. 1952. *Welfare Economics and the Theory of the State*. Cambridge, Mass.: Harvard University Press.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*. London: Payne. Reprinted, Oxford: Clarendon Press, 1907.

Bergson, A. 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52, 310–34.

Bernholz, P. 1980. A general social dilemma: Profitable exchange and intransitive group preference. *Zeitschrift für Nationalökonomie* 40, 1–23.

Bezembinder, Th. and van Acker, P. 1979. A note on Sen's partial comparability model. Department of Psychology, Katholieke Universiteit, Nijmegen, The Netherlands.

Binmore, K. 1976. Social choice and parties. *Review of Economic Studies* 43, 459–64.

Black, D. 1948. On the rationale of group decision making. *Journal of Political Economy* 56, 23–34.

Black, D. 1958. *The Theory of Committees and Elections*. London: Cambridge University Press.

Blackorby, C. 1975. Degrees of cardinality and aggregate partial orderings. *Econometrica* 43, 845–52.

Blackorby, C., Donaldson, D. and Weymark, J.A. 1984. Social choice with interpersonal utility comparisons. *International Economic Review* 25, 327–56.

Blair, D.H., Bordes, G., Kelly, J.S. and Suzumura, K. 1976. Impossibility theorems without collective rationality. *Journal of Economic Theory* 13, 361–79.

Blair, D.H. and Muller, E. 1983. Essential aggregation procedures on restricted domains of preferences. *Journal of Economic Theory* 30, 34–53.

Blair, D.H. and Pollak, R.A. 1979. Collective rationality and dictatorship: the scope of the Arrow theorem. *Journal of Economic Theory* 21, 186–94.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Blair, D.H. and Pollak, R.A. 1982. Acyclic collective choice rules. *Econometrica* 50, 931–43.

Blau, J.H. 1957. The existence of a social welfare function. *Econometrica* 25, 302–13.

Blau, J.H. 1975. Liberal values and independence. *Review of Economic Studies* 42, 413–20.

Blau, J.H. 1976. Neutrality, monotonicity and the right of veto: a comment. *Econometrica* 44, 603.

Blau, J.H. 1979. Semiorders and collective choice. *Journal of Economic Theory* 21, 195–206.

Blau, J.H. and Deb, R. 1977. Social decision functions and veto. *Econometrica* 45, 871–9.

Bonner, J. 1986. *Politics, Economics and Welfare*. Brighton: Wheatsheaf.

Borda, J.C. 1781. Memoire sur les élections au scrutin. *Mémoires de l'Académie Royale des Sciences*; English translation by A. de Grazia, *Isis* 44 (1953).

Bordes, G. 1976. Consistency, rationality and collective choice. *Review of Economic Studies* 43, 447–57.

Bose, A. 1975. *Marxian and Post-Marxian Political Economy*. Harmondsworth: Penguin Books.

Brams, S.J. 1975. *Game Theory and Politics*. New York: Free Press.

Brams, S.J. 1976. *Paradoxes in Politics*. New York: Free Press.

Brams, S.J. and Fishburn, P.C. 1983. *Approval Voting*. Boston: Birkhäuser.

Breyer, F. 1977. The liberal paradox, decisiveness over issues, and domain restrictions. *Zeitschrift für Nationalökonomie* 37, 45–60.

Breyer, F. and Gardner, R. 1980. Liberal paradox, game equilibrium and Gibbard optimum. *Public Choice* 35, 469–81.

Breyer, F. and Gigliotti, G.A. 1980. Empathy and respect for the right of others. *Zeitschrift für Nationalökonomie* 40, 59–64.

Brown, D.J. 1974. An approximate solution to Arrow's problem. *Journal of Economic Theory* 9, 375–83.

Brown, D.J. 1975. Aggregation of preferences. *Journal of Economics* 89, 456–69.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Buchanan, J.M. and Tullock, G. 1962. *The Calculus of Consent*. Ann Arbor: University of Michigan Press.

Camacho, A. 1974. Societies and social decision functions. In Leinfellner and Kohler (1974).

Campbell, D.E. 1976. Democratic preference functions. *Journal of Economic Theory* 12, 259–72.

Campbell, D.E. 1980. Algorithms for social choice functions. *Review of Economic Studies* 47, 617–27.

Champsaur, P. and Laroque, G. 1982. Strategic behavior in decentralized planning procedures. *Econometrica* 50, 325–44.

Chichilnisky, G. 1982. Social aggregation rules and continuity. *Quarterly Journal of Economics* 96, 337–52.

Chichilnisky, G. and Heal, G. 1983. Necessary and sufficient conditions for a resolution of the social choice paradox. *Journal of Economic Theory* 31, 68–87.

Condorcet, M.de. 1785. *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à le Pluralité des Voix*. Paris.

Coughlin, P.J. 1986. Rights and the private Pareto principle. *Economica* 53.

Crawford, V.P. 1979. A procedure for generating Pareto-efficient egalitarian-equivalent allocations. *Econometrica* 47, 49–60.

Crawford, V.P. and Heller, W.P. 1979. Fair division with indivisible commodities. *Journal of Economic Theory* 21, 10–27.

Dasgupta, P., Hammond, P. and Maskin, E. 1979. The implementation of social choice rules: some general results on incentive compatibility. *Review of Economic Studies* 46, 185–216.

Dasgupta, P. and Heal, G. 1979. *Economic Theory and Exhaustible Resources*. London: James Nisbet, and Cambridge: Cambridge University Press.

d'Aspremont, C. and Gevers, L. 1977. Equity and informational basis of collective choice. *Review of Economic Studies* 46, 199–210.

d'Aspremont, D. 1985. Axioms for social welfare orderings in Hurwicz, Schmeidler and Sonnenschein (1985).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Davis, O.A., De Groot, M.H. and Hinich, M.J. 1972. Social preference orderings and majority rule. *Econometrica* 40, 147–57.

Deb, R. 1976. On Schwartz's rule. *Journal of Economic Theory* 16, 103–10.

Deschamps and Gevers, L. 1978. Leximin and utilitarian rules: a joint characterisation. *Journal of Economic Theory* 17, 143–163.

Deschamps and Gevers, L. 1979. Separability, risk-taking and social welfare judgements. In Laffont (1979).

Dummett, M. 1984. *Voting Procedures*. Oxford: Clarendon Press.

Dummett, M. and Farquharson, R. 1961. Stability in voting. *Econometrica* 29, 133–43.

Dutta, B. 1980. On the possibility of consistent voting procedures. *Review of Economic Studies* 47, 603–16.

Dutta, B. and Pattanaik, P.K. 1978. On nicely consistent voting systems. *Econometrica* 46, 163–70.

Elster, J. and Hylland, A., eds. 1986. *Foundations of Social Choice Theory*. Cambridge: Cambridge University Press.

Farkas, D. and Nitzan, S. 1979. The Borda rule and Pareto stability: a comment. *Econometrica* 47, 1305–6.

Farquharson, R. 1956. Straightforwardness in voting paradoxes. *Oxford Economic Papers* 8, 80–9.

Farrell, M.J. 1976. Liberalism in the theory of social choice. *Review of Economic Studies* 43, 3–10.

Feldman, A. and Kirman, A. 1974. Fairness and envy. *American Economic Review* 64, 995–1005.

Ferejohn, J.A. 1978. The distribution of rights in society. In Gottinger and Leinfellner (1978).

Fine, B.J. 1975. A note on interpersonal aggregation and partial comparability. *Econometrica* 43, 173–4.

Fine, B.J. and Fine, K. 1974. Social choice and individual ranking. *Review of Economic Studies* 41, 303–22, 459–75.

Fishburn, P.C. 1973. *The Theory of Social Choice*. Princeton: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Fishburn, P.C. 1974. On collective rationality and a generalized impossibility theorem. *Review of Economic Studies* 41, 445–59.

Fishburn, P.C. and Gehrlein, W.V. 1976. Borda's rule, positional voting, and Condorcet's simple majority principle. *Public Choice* 28, 79–88.

Foley, D. 1967. Resource allocation in the public sector. *Yale Economic Essays* 7, 73–6.

Fountain, J. 1980. Bowley's analysis of bilateral monopoly and Sen's liberal paradox in collective choice theory: a note. *Quarterly Journal of Economics* 95.

Gaertner, W. 1983. Equity- and inequity-type Borda rules. *Mathematical Social Sciences* 4, 137–54.

Gaertner, W. 1986. Pareto, independent rights exercising and strategic behaviour. *Journal of Economics: Zeitschrift für Nationalökonomie* 46.

Gaertner, W. and Heinecke, A. 1978. Cyclically mixed preferences: A necessary and sufficient condition for transitivity of the social preference relation. In Gottinger and Leinfellner (1978).

Gaertner, W. and Krüger, L. 1981. Self-supporting preferences and individual rights: the possibility of Paretian libertarianism. *Economica* 48, 17–28.

Gaertner, W. and Krüger, L. 1983. Alternative libertarian claims and Sen's paradox. *Theory and Decision* 15, 211–30.

Gärdenfors, P. 1973. Positional voting functions. *Theory and Decision* 4, 1–24.

Gärdenfors, P. 1981. Rights, games and social choice. *Nous* 15.

Gardner, R. 1977. The Borda game. *Public Choice* 30, 43–50.

Gardner, R. 1980. The strategic inconsistency of Paretian liberalism. *Public Choice* 35, 241–52.

Gevers, L. 1979. On interpersonal comparability and social welfare orderings. *Econometrica* 47, 75–90.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Gibbard, A. 1974. A Pareto-consistent libertarian claim. *Journal of Economic Theory* 7, 338–410.

Goldman, S.M. and Sussangkarn, C. 1978. On the concept of fairness. *Journal of Economic Theory* 19, 210–16.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21, 63–80.

Gottinger, H.W. and Leinfellner, W. 1978. *Decision Theory and Social Ethics: Issues in Social Choice*. Dordrecht: Reidel.

Graaff, J. der. 1957. *Theoretical Welfare Economics*. Cambridge: Cambridge University Press.

Grandmont, J.M. 1978. Intermediate preferences and majority rule. *Econometrica* 46, 317–30.

Green, J. and Laffont, J.-J. 1979. *Incentives in Public Decision Making*. Amsterdam: North-Holland.

Grether, D.M. and Plott, C.R. 1982. Nonbinary social choice: an impossibility theorem. *Review of Economic Studies* 49, 143–9.

Groves, T. and Ledyard, J. 1977. Optimal allocation of public goods: a solution to the ‘free rider’ problem. *Econometrica* 45, 783–810.

Guha, A.S. 1972. Neutrality, monotonicity and the right of veto. *Econometrica* 40, 821–6.

Hammond, P.J. 1976. Equity, Arrow's conditions and Rawls’ difference principle. *Econometrica* 44, 793–804.

Hammond, P.J. 1977. Dual interpersonal comparisons of utility and the welfare economics of income distribution. *Journal of Public Economics* 6, 51–71.

Hammond, P.J. 1979. Equity in two person situation: some consequences. *Econometrica* 47, 1127–36.

Hammond, P.J. 1981. Liberalism, independent rights, and the Pareto principle. In L.J. Cohen et al., *Logic, Methodology and Philosophy of Sciences*, Amsterdam: North-Holland.

Hansson, B. 1973. The independence condition in the theory of social choice. *Theory and Decision* 4.

Hansson, B. 1976. The existence of group preferences. *Public Choice* 28, 89–98.

Hansson, B. and Sahlquist, H. 1976. A proof technique for social choice with variable electorate. *Journal of Economic Theory* 13, 193–200.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Harsanyi, J.C. 1979. Bayesian decision theory, rule utilitarianism and Arrow's impossibility theorem. *Theory and Decision* 11, 289–318.

Herzberger, H.G. 1973. Ordinal preference and rational choice. *Econometrica* 41, 187.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hurley, S. 1985. Supervenience and the possibility of coherence. *Mind* 94, 501–26.

Hurwicz, L., Schmeidler, D. and Sonnenschein, H., eds. 1985. *Social Goals and Social Organization: Essays in Memory of Elisha Pazner*. Cambridge: Cambridge University Press.

Inada, K. 1984. On the economic welfare function. *Econometrica* 32, 316–38.

Inada, K. 1964. A note on the simple majority decision rule. *Econometrica* 37, 490–506.

Inada, K. 1969. On the simple majority decision rule. *Econometrica* 32, 525–31.

Inada, K. 1970. Majority rule and rationality. *Journal of Economic Theory* 2, 27–40.

Jain, S. 1986. Special majority rules: a necessary and sufficient condition for quasi–transitivity with quasi–transitive individual preferences. *Social Choice and Welfare* 3, 99–106.

Kalai, E. and Muller, E. 1977. Characterization of domains admitting non-dictatorial social welfare functions and nonmanipulable voting procedures. *Journal of Economic Theory* 16, 457–69.

Kalai, E. and Ritz, Z. 1980. Characterization of private alternative domains admitting Arrow social welfare functions. *Journal of Economic Theory* 22, 23–36.

Kaldor, N. 1939. Welfare propositions in economics. *Economic Journal* 49, 549–52.

Karni, E. 1978. Collective rationality, unanimity and liberal ethics. *Review of Economic Studies* 45, 571–4.

Kelly, J.S. 1976. The impossibility of a just liberal. *Economica* 43, 67–75.

Kelly, J.S. 1978. *Arrow Impossibility Theorems*. New York: Academic Press.

Kelsey, D. 1984. Acyclic choice without the Pareto principle. *Review of Economic Studies* 51, 693–9.

Kelsey, D. 1985. The liberal paradox: a generalization. *Social Choice and Welfare* 1, 245–50.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kemp, M.C. and Ng, Y.K. 1976. On the existence of social welfare functions, social orderings and social decision functions. *Economica* 43, 59–66.

Kolm, S.Ch. 1969. The optimum production of social justice. In *Public Economics*, ed. J. Margolis and H. Guitton. London: Macmillan.

Kramer, G.H. 1973. On a class of equilibrium conditions for majority rule. *Econometrica* 41, 285–97.

Kramer, G.H. 1978. A dynamic model of political equilibrium. *Journal of Economic Theory* 16, 310–34.

Krüger, L. and Gaertner, W. 1983. Alternative libertarian claims and Sen's paradox. *Theory and Decision* 15.

Laffont, J.J., ed. 1979. *Aggregation and Revelation of Preferences*. Amsterdam: North-Holland.

Larsson, B. 1983. *Basic Properties of the Majority Rule*. Lund: University of Lund.

Leinfellner, W. and Kohler, E., eds. 1974. *Developments in the Methodology of Social Sciences*. Dordrecht: Reidel.

Levi, I. 1982. Liberty and welfare. In Sen and Williams (1982).

Little, I.M.D. 1950. *A Critique of Welfare Economics*. Oxford: Clarendon Press. 2nd edn, 1957.

Machina, M. and Parks, R. 1981. On path independent randomized choices. *Econometrica* 49, 1345–7.

Mas-Colell, A. and Sonnenschein, H.F. 1972. General possibility theorems for group decisions. *Review of Economic Studies* 39, 185–92.

Maskin, E. 1976. Social welfare functions on restricted domain. Mimeo.

Maskin, E. 1978. A theorem on utilitarianism. *Review of Economic Studies* 45, 93–6.

Matsumoto, Y. 1985. Non-binary social choice: revealed preferential interpretation. *Economica* 52, 185–94.

Mayston, D.J. 1974. *The Idea of Social Choice*. London: Macmillan.

McKelvey, R.D. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47, 1085–112.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

McLean, I.S. 1980. Liberty, equality and the Pareto principle: a comment on Weale. *Analysis* 40.

McManus, M. 1982. Some properties of topological social choice functions. *Review of Economic Studies* 49, 447–60.

Mill, J.S. 1859. *On Liberty*. Reprinted, Harmondsworth: Penguin Books, 1974.

Monjardet, B. 1979. Duality in the theory of social choice. In Laffont (1979).

Moulin, H. 1983. *The Strategy of Social Choice*. Amsterdam: North-Holland.

Mueller, D.C. 1979. *Public Choice*. Cambridge: Cambridge University Press.

Murakami, Y. 1968. *Logic and Social Choice*. New York: Dover.

Myerson, R.B. 1983. Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica* 49, 883–97.

Ng, Y.K. 1971. The possibility of a Paretian liberal: impossibility theorems and cardinal utility. *Journal of Political Economy* 79, 1397–402.

Ng, Y.K. 1979. *Welfare Economics*. London: Macmillan.

Nitzan, S. and Paroush, J. 1985. *Collective Decision Making: An Economic Outlook*. Cambridge: Cambridge University Press.

Nitzan, S. and Rubinstein, A. 1981. A further characterization of the Borda Ranking Methods. *Public Choice* 36, 153–8.

Nozick, R. 1974. *Anarchy, State and Utopia*. Oxford: Blackwell.

Parks, R.P. 1976. An impossibility theorem for fixed preferences: A dictatorial Bergson–Samuelson social welfare function. *Review of Economic Studies* 43, 447–450.

Pattanaik, P.K. 1971. *Voting and Collective Choice*. Cambridge: Cambridge University Press.

Pattanaik, P.K. 1973. On the stability of sincere voting situations. *Journal of Economic Theory* 6, 558–74.

Pattanaik, P.K. 1978. *Strategy and Group Choice*. Amsterdam: North-Holland.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Pattanaik, P.K. and Salles, M., eds. 1983. *Social Choice and Welfare*. Amsterdam: North-Holland.

Pazner, E.A. and Schmeidler, D. 1974. A difficulty in the concept of fairness. *Review of Economic Studies* 41, 441–3.

Peleg, B. 1978. Consistent voting systems. *Econometrica* 46, 153–62.

Peleg, B. 1984. *Game Theoretic Analysis of Voting in Committees*. Cambridge: Cambridge University Press.

Plott, C.R. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57, 788–806.

Plott, C.R. 1973. Path independence, rationality and social choice. *Econometrica* 41, 1075–91.

Plott, C.R. 1976. Axiomatic social choice theory: an overview and interpretation. *American Journal of Political Science* 20, 511–96.

Pollak, R.A. 1979. Bergson-Samuelson social welfare functions and the theory of social choice. *Quarterly Journal of Economics* 93, 73–90.

Ray, P. 1973. Independence of irrelevant alternatives. *Econometrica* 41, 987–91.

Riley, J. 1986. *Liberal Utilitarianism: Social Choice Theory and J.S. Mill's Philosophy*. Cambridge: Cambridge University Press.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Robbins, L. 1938. Interpersonal companions of utility. *Economic Journal* 48, 635–41.

Roberts, K.W.S. 1980a. Interpersonal comparability and social choice theory. *Review of Economic Studies* 47, 421–39.

Roberts, K.W.S. 1980b. Social choice theory: the single and multiple-profile approaches. *Review of Economic Studies* 47, 441–50.

Rubinstein, A. 1981. The single profile analogues to multiple profile theorems: mathematical logic's approach. Murray Hill: Bell Laboratories.

Salles, M. 1976. Characterization of transitive individual preferences for quasi-transitive collective preferences under simple games. *International Economic Review* 17, 308–18.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Samuelson, P.A. 1950. Evaluation of real national income. *Oxford Economic Papers* 2, 1–19.

Samuelson, P.A. 1967. Arrow's mathematical politics. In *Human Values and Economic Policy*, ed. S. Hook. New York: New York University Press.

Saposnik, R. 1975. On transitivity of the social preference relation under simple majority rule. *Journal of Economic Theory* 10, 1–7.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Schick, F. 1969. Arrow's proof and the logic of preference. *Journal of Philosophy* 36, 127–44.

Schmeidler, D. and Sonnenschein, H. 1978. Two proofs of the Gibbard–Satterthwaite theorem on the possibility of a strategyproof social choice function. In Gottinger and Leinfellner (1978).

Schmeidler, D. and Vind, K. 1972. Fair net trades. *Econometrica* 40, 637–42.

Schofield, N. 1978. Instability of simple dynamic games. *Review of Economic Studies* 40, 575–94.

Schofield, N. 1983. Generic instability of majority rule. *Review of Economic Studies* 50, 695–705.

Schwartz, T. 1970. On the possibility of rational policy evaluation. *Theory and Decision* 1, 89–106.

Schwartz, T. 1972. Rationality and the myth of the maximum. *Nous* 6, 97–117.

Schwartz, T. 1986. *The Logic of Collective Choice*. New York: Columbia University Press.

Scitovsky, T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* 9, 77–88.

Seidl, C. 1975. On liberal values. *Zeitschrift für Nationalökonomie* 35, 257–92.

Sen, A.K. 1966. A possibility theorem on majority decisions. *Econometrica* 34, 75–9.

Sen, A.K. 1969. Quasi-transivity, rational choice and collective decisions. *Review of Economic Studies* 36, 381–93.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day; Edinburgh: Oliver

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
& Boyd. Republished, Amsterdam: North-Holland.

Sen, A.K. 1977a. Social choice theory: a re-examination. *Econometrica* 45, 58–89.

Sen, A.K. 1977b. On weights and measures: informational constraints in social welfare analysis. *Econometrica* 45, 1539–72.

Sen, A.K. 1982. *Choice, Welfare and Measurement*. Oxford: Blackwell, and Cambridge, Mass.: MIT Press.

Sen, A.K. 1983. Liberty and social choice. *Journal of Philosophy* 80, 5–28.

Sen, A.K. 1986. Social choice theory. In *Handbook of Mathematical Economics* Vol III, ed. K.J. Arrow and M. Intriligator. Amsterdam: North-Holland.

Sen, A.K. and Pattanaik, P.K. 1969. Necessary and sufficient conditions for rational choice under majority decision. *Journal of Economic Theory* 1, 178–202.

Sen, A.K. and Williams, B. 1982. *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Sengupta, M. 1980. The knowledge assumption in the theory of strategic voting. *Econometrica* 49, 1301–4.

Sengupta, M. and Dutta, B. 1979. A condition for Nash stability under binary and democratic group decision functions. *Theory and Decision* 10, 293–310.

Slutsky, S. 1977. A characterization of societies with consistent majority decision. *Review of Economic Studies* 44, 211–26.

Strasnick, S. 1976. Social choice theory and the derivation of Rawls' difference principle. *Journal of Philosophy* 73, 85–99.

Sugden, R. 1981. *The Political Economy of Public Choice*. Oxford: Martin Robertson.

Sugden, R. 1985. Liberty, preference and choice. *Economics and Philosophy* 1, 213–30.

Suppes, P. 1966. Some formal models of grading principles. *Synthese* 6, 284–306.

Suzumura, K. 1978. On the consistency of libertarian claims. *Review of Economic Studies* 45, 329–42. A correction, 46, (1979), 743.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Suzumura, K. 1980. Liberal paradox and the voluntary exchange of rights-exercising. *Journal of Economic Theory* 22, 407–42.

Suzumura, K. 1983. *Rational Choice, Collective Decisions and Social Welfare*. Cambridge: Cambridge University Press.

Svensson, L.G. 1980. Equity among generations. *Econometrica* 48, 1251–6.

Tullock, G. 1967. The general irrelevance of the general possibility theorem. *Quarterly Journal of Economics* 81, 256–70.

Tullock, G. 1969. *Toward a Mathematics of Politics*. Ann Arbor: University of Michigan Press.

Varian, H. 1974. Equity, envy and efficiency. *Journal of Economic Theory* 9, 63–91.

Varian, H. 1975. Distributive justice, welfare economics and the theory of fairness. *Philosophy and Public Affairs* 4, 223–47.

Vickrey, W. 1945. Measuring marginal utility by reactions to risk. *Econometrica* 13, 319–33.

Vickrey, W. 1960. Utility, strategy and social decision rules. *Quarterly Journal of Economics* 75, 507–25.

Ward, B. 1965. Majority voting and alternative forms of public enterprise. In *The Public Economy of Urban Communities*, ed. J. Margolis. Baltimore: Johns Hopkins Press.

Weale, A. 1980. The impossibility of a liberal egalitarianism. *Analysis* 40.

Webster, N. 1986. Liberals and information. *Theory and Decision* 20, 41–52.

Weymark, J.A. 1983. Arrow's theorem with social quasi-orderings. *Public Choice* 42, 235–46.

Wilson, R.B. 1972. Social choice theory without the Pareto principle. *Journal of Economic Theory* 5, 478–86.

Wilson, R.B. 1975. On the theory of aggregation. *Journal of Economic Theory* 10, 89–99.

Wriglesworth, J. 1985. *Libertarian Conflicts in Social Choice*. Cambridge: Cambridge University Press.

Young, H.P. 1974a. An axiomatization of the Borda's rule. *Journal of Economic Theory* 9, 43–52.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Young, H.P. 1974b. A note on preference aggregation. *Econometrica* 42, 1129–31.

How to cite this article

Sen, Amartya. "social choice." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000164> doi:10.1057/9780230226203.1550

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social contract

Peter Vallentyne
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social contract theory is a theory about how the moral assessment of actions, practices, institutions, laws, constitutions, or related items is based – directly or indirectly – on the consent – actual or hypothetical – of the members of society. Hobbes, Locke, Rousseau, and Kant represent the main historical figures. Rawls, Gauthier, and Scanlon are the main contemporary figures.

Keywords

Buchanan, J. M.; consent; contractarianism; Gauthier, D.; Harsanyi, J. C.; Hobbes, T.; Individualism; interests; justice; Kant, I.; Locke, J.; natural rights; Plato; Rawls, J.; reciprocity; Rousseau, J.-J.; social contract; veil of ignorance

Article

Social contract theory, as I shall understand it here, is a theory about how the moral assessment of actions, practices, institutions, laws, constitutions, or related items is based – directly or indirectly – on the consent – actual or hypothetical – of the members of society.

Although social contract theories could be formulated to require merely that a majority of the members society consent to the relevant item, such formulations are not particularly plausible and generally have not been advocated. The standard requirement is that *all* members consent to the item, and I shall assume this in what follows.

Social contract theory is also sometimes understood to be (*a*) an empirical theory about how government actually arose, or (*b*) a metaethical theory about the general nature of morality and moral reasons. Here, however, we shall focus on contractarianism as a substantive theory of justice.

The origins of social contract theory can be traced back to the discussion of justice (in the voice of Glaucon) by Plato (c. 430–347 bce) in *The Republic* (1961), but systematic development of the theory really started in the 17th century, with the beginnings of the modern state and the challenging of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
alleged divine right of kings and aristocrats to rule. The first comprehensive statement of social contract theory came from Thomas Hobbes (1588–1679) in his *Leviathan* (1651), in which he offered a social contract justification for almost unlimited powers of the state. Other important historical figures associated with social contract theory include John Locke (1632–1704) (1690), Jean-Jacques Rousseau (1712–78) (1762), and Immanuel Kant (1724–1804) (1785). Since the publication of John Rawls's *A Theory of Justice* (1971), there has been a significant renewal of interest – among economists, political scientists, and philosophers – in social contract ethical and political theory.

Social contract theory can be used to address several quite different topics. It might be formulated (*a*) as a full theory of *individual morality* (ethics), (*b*) as a theory of *what duties we morally owe each other* (which does not include any impersonal duties), (*c*) as a *theory of political authority* in the sense of the conditions under which we have a duty to obey the dictates of others, (*d*) as a theory of *legitimacy* in the sense of the conditions under which others are not permitted to forcibly interfere with our actions (even if wrong), or (*e*) as a theory of the moral permissibility or *justice of political institutions*. For simplicity, I shall focus on social contract theory as a theory of the justice of political institutions – although most remarks apply to most other versions.

There are two broad kinds of social contract theory: *actual* contract (rights-based) theory and *hypothetical* contract theory (contractarianism or contractualism). I consider each in turn.

Actual social contract theory

Actual social contract theory holds that a set of political institutions (for example) is just if and only if it has been directly consented to by those governed or conforms to the requirements of a constitution to which all have consented. The best-known actual social contract theory is that of Locke (1690) – although it has also been interpreted as a hypothetical contract theory (see below).

In order for consent to have moral force, the agent must be rationally competent and the consent must be free (not coerced) and suitably informed (for example, not based on fraud). Establishing exactly what is required for consent to be valid is a very important topic. For excellent discussion, see Simmons (1993).

Actual consent can be explicit (as in ‘I hereby consent’) or implicit (as when one allows a friend to enter one's house without explicitly granting permission). Identifying the exact conditions under which an action other than explicit consent is nonetheless a case of (implicit) consent is an important topic insightfully analysed by Simmons (1993). Given that explicit consent is extremely rare, most actual social contract theories (plausibly) allow both kind of consent to have moral force.

Actual social contract theories presuppose that individuals have certain (choice-protecting) natural rights of self-governance (a kind of natural freedom). Any restriction of this natural freedom is deemed unjust unless the individual has consented to it. Given the insecurity of life, health, and possessions in the absence of government, it typically makes sense for people to give up some of their natural freedom and submit to the authority of government – on the condition that the political authority thus established does a reasonably good job protecting people's pre-political rights. When they do so, actual social contract theory holds that the resulting political institutions are just.

Actual social contract theory faces three important challenges. One is whether individuals have the natural rights that the theory postulates. A second is that, for almost all existing societies, there never was a relevant universal agreement on political institutions – which entails, given the theory, that no set

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of political institutions is just. A third is that, even if there was a relevant universal agreement in some past generation, it's not clear why this would be relevant for the present generation (the members of which have not consensually given up their natural rights). (For further discussion, see Simmons 1993.)

Hypothetical social contract theory

Given the problem of securing universal actual consent to political institutions (or any set of rules), most contemporary social contract theories appeal to hypothetical consent. They hold that political institutions (for example) are just if and only if (*a*) they would be universally agreed to under specified conditions, or (*b*) they conform to a political constitution that would be agreed to under specified conditions. For brevity, call these theories *contractarian*.

Contractarian theories differ in their specification of the circumstances under which the relevant hypothetical agreement takes place. These conditions include the motivations and beliefs of the parties to the agreement as well as the non-agreement point (the outcome if no agreement is made).

Hobbesian approaches provide realistic, morally neutral, specifications of the circumstances, and attempt thereby to reduce morality to individual or collective rationality. They specify that individuals are mainly interested in promoting their own wellbeing and assume that agents have reasonably full knowledge of their situation. The non-agreement point (state of nature) is a state of war of all against all in which there is neither government nor morally constrained behaviour. Hobbes (1651), Buchanan (1975), and Gauthier (1986) are all in this tradition. (See also Hampton, 1986; Kavka, 1986; and Binmore, 1994; 1998.) A main objection to the Hobbesian approach is that it yields an impoverished conception of morality: individuals are protected by morality only to the extent that cooperation with them is useful to others. Because babies, young children and severely disabled people offer others no benefits from cooperation, knowledgeable predominantly self-interested agents would not enter into agreements with them. Such vulnerable individuals might be protected by the terms of the hypothetical agreement, but that would be only to the extent that others happen to care about them (for example, parents for their children). Absent any such contingent concern, such individuals are deemed merely resources to be exploited.

Lockean forms of contractarianism are similar to Hobbesian ones except that the non-agreement point (state of nature) is not pre-moral. (As indicated above, Locke, 1690, has been read both as an actual social contract theory and as a hypothetical social contract theory.) Although there is no government, individuals have certain natural rights, and generally respect the rights of others. Although the non-agreement point is not the dire Hobbesian state of war of all against all, the absence of a generally accepted adjudication and enforcement agency often leads to feuds when someone believes that her rights have been violated without adequate rectification. In addition, the absence of government makes it difficult to provide various public goods (such as roads and national defence). The result is that it would be rational for all to agree to give up some of their natural rights and to establish a state.

Kantian forms of contractarianism (sometimes called ‘contractualism’) view the social contract device as representing reasonable reciprocity between moral equals who are self-legislating free and equal members of the kingdom of ends. Harsanyi (1953; 1955) and Rawls (1971), for example, allow that individuals are primarily concerned with promoting their own good, but they impose a *veil of ignorance* that blocks parties from having any knowledge of their capacities, positions, or desires. One objection to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
this approach is that it eliminates all individual differences and thus reduces the agreement to the choice of one individual behind the veil of ignorance. Scanlon (1998), by contrast, allows individuals full knowledge of their situations, but stipulates that the parties are motivated by a desire to reach a fair and reasonable agreement (as opposed to simply promoting their own good). (Habermas, 1993, is similar in spirit.) So understood, contractarianism takes ability to justify our conduct to each other as central to morality. Of course, the notion of a fair and reasonable agreement is a moral notion and does an enormous amount of work here.

I shall not here attempt to adjudicate between Hobbesian, Lockean and Kantian contractarianism. Instead, I shall focus on general challenges to contractarianism.

Contractarianism is sometimes charged with ignoring the interests of beings – such as animals, infants and fetuses – that are not able to communicate linguistically, make commitments, and so on. Although some Hobbesian contractarian theories do ignore these interests, this is not an essential part of contractarianism. Some theories (such as Scanlon's) take these interests into account by allowing that trustees representing the interests of such beings are parties to the agreement.

A more fundamental criticism of contractarianism, often raised by Marxists, feminists and communitarians, is that it is *individualistic*. Contractarianism is indeed *normatively individualistic*, which is to say that it claims that the ultimate right-making features are features of individual people (viz. their consent), not irreducible features of collectivities. It does not, however, assume *ontogenetic* (or *developmental*) *individualism*, the view that denies that individual people are shaped and formed by the social context in which they find themselves. Nor does it assume *ontological individualism*, the view that individual persons are ontologically prior to society. Nor is contractarianism committed to the view that people are (inevitably or contingently) *egoistic* or *materialistic* in their desires (for example, caring only about the bundle of material goods that they control). Many contractarian theorists have made such assumptions, but such assumptions are not essential to contractarianism.

The main challenge to contractarianism questions the claim that *hypothetical* consent has any normative force. We generally deem it wrong for someone to take one's car without one's permission, even if one would have consented had one been asked. Moreover, in those cases where hypothetical consent seems to have some moral force, it may be because it is an indicator of what is in that person's *interests*. For example, my hypothetical consent for you to move my car away from the fire is an indication that it is in my *interest* that you do so. It may be that the appeal to my interests is what is doing the moral work here and that hypothetical consent does no real work.

Consent under the right conditions is clearly normatively significant. Whether actual or hypothetical consent can do all the work required of it for contractarian theories is a matter of ongoing debate.

See Also

- Buchanan, James M.
- collective rationality
- contract theory
- Harsanyi, John C.
- Hume, David
- implicit contracts

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- justice
- justice (new perspectives)
- market failure
- Pareto efficiency
- Pareto principle and competing principles
- social norms
- tragedy of the commons

Bibliography

Binmore, K. 1994. *Game Theory and the Social Contract. Volume 1: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, K. 1998. *Game Theory and the Social Contract. Volume 2: Just Playing*. Cambridge, MA: MIT Press.

Buchanan, J. 1975. *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago: University of Chicago Press.

Gauthier, D. 1986. *Morals by Agreement*. London: Oxford University Press.

Habermas, J. 1993. *Justification and Application: Remarks on Discourse Ethics*, trans. C. Cronin. Cambridge, MA: MIT Press.

Hampton, J. 1986. *Hobbes and the Social Contract Tradition*. Cambridge, MA: Cambridge University Press.

Harsanyi, J. 1953. Cardinal utility in welfare economics and the theory of risk taking. *Journal of Political Economy* 61, 434–5. Reprinted in J. Harsanyi, *Essays on Ethics, Social Behavior, and Scientific Explanation*. Dordrecht: Reidel, 1976.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21. Reprinted in J. Harsanyi, *Essays on Ethics, Social Behavior, and Scientific Explanation*. Dordrecht: Reidel, 1976.

Hobbes, T. 1651. *Leviathan*. New York: Penguin Books, 1968.

Kant, I. 1785. *Groundwork of the Metaphysics of Morals*. New York: Harper Torchbooks, 1958.

Kavka, G. 1986. *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Locke, J. 1690. *The Second Treatise of Government*. Indianapolis: Bobbs-Merrill, 1952.

Plato. 1961. *The Republic*. In *The Collected Dialogues of Plato*, ed. E. Hamilton and H. Cairns. Princeton: Princeton University Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Rousseau, J.J. 1762. *Contrat Social*. In *Political Writings*, ed. C.E. Vaughan. Cambridge: Cambridge University Press, 1915.

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

Simmons, A.J. 1993. *On the Edge of Anarchy*. Princeton: Princeton University Press.

How to cite this article

Vallentyne, Peter. "social contract." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000165> doi:10.1057/9780230226203.1552

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social democracy

Ben Jackson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social democracy refers to a political theory, a social movement or a society that aims to achieve the egalitarian objectives of socialism while remaining committed to the values and institutions of liberal democracy. This article examines the historical development of all three forms of social democracy. It shows that social democracy was one of the most creative and durable influences on the politics and economics of the advanced industrialized nations during the 20th century and that, in spite of some setbacks, it retains a distinctive political agenda for the future.

Keywords

aggregate demand; Bernstein, E.; Beveridge, W. H.; budget deficits; business cycle; capital controls; capitalism; corporatism; democracy; earned income tax credits; egalitarianism; equal pay for equal work; equality; full employment; inflation; Keynes, J. M.; labour market regulation; Marshall, T. H.; neoliberalism; pensions; post-industrial economy; poverty; progressive taxation; public expenditure; Rehn–Meidner model (Sweden); revisionism; social citizenship; social democracy; social insurance; socialism; Stockholm School; trade unions; unemployment insurance; utopian socialism; welfare state; women's work and wages

Article

Social democracy refers to a political theory, a social movement or a society that aims to achieve the egalitarian objectives of socialism while remaining committed to the values and institutions of liberal democracy.

Born in an era of sharp ideological polarities and intense social conflicts, social democracy has often been seen as a pragmatic compromise between capitalism and socialism. As Leszek Ko•akowski has put it: ‘The trouble with the social-democratic idea is that it does not stock or sell any of the exciting ideological commodities which totalitarian movements – communist, fascist, or leftist – offer dream-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
hungry youth.' Instead of an 'ultimate solution for all human misfortune' or a 'prescription for the total salvation of mankind', said Ko•akowski, social democracy offers merely 'an obstinate will to erode by inches the conditions which produce avoidable suffering, oppression, hunger, wars, racial and national hatred, insatiable greed and vindictive envy' (Ko•akowski, 1982, p. 11). For much of the 20th century, the comparative modesty of this ambition led many to underestimate social democracy. Both the Marxist Left and the free market Right disparaged the purportedly unimaginative parties and trade unions that were the principal carriers of social democratic ideas, and scorned the vacillating compromises they saw as inherent in a social democratic analysis of politics. But if we review the historical development of social democracy, it becomes clear that such interpretations misjudge its character and significance. In fact, social democracy deserves to be recognized as one of the 20th century's most creative and durable influences on the politics and economics of the advanced industrialized nations.

Formation

What would later be called 'social democracy' first emerged in the late 19th century in the labour movements of north-west Europe. Early non-European outposts were also established in Australia and New Zealand around the same time. In nations such as Britain, France, Germany and Sweden, advocates of the interests of the working class inhabited polities that were characterized by rapid industrialization and the slow, inconsistent emergence of liberal constitutionalism and democratic citizenship. These circumstances created a complex structure of constraints and opportunities for labour movements that differed from those in southern or eastern Europe. In this relatively liberal environment, the politicized elements of the working class could build powerful political parties and trade unions to represent and protect their interests, and ultimately, or so they hoped, use democratic means to abolish the profound poverty and social oppression that working-class leaders saw as the ineluctable consequences of industrialization. The leaders and theorists of these movements, figures such as Keir Hardie, Jean Jaurès, Eduard Bernstein and Hjalmar Branting, were influenced by a variety of ideological traditions, most obviously Marxism, but also progressive liberalism, republicanism and 'utopian' socialism (for one account of such non-Marxist influences see Stedman Jones, 2004). They drew upon all of these intellectual currents as they began to sketch the outlines of a social democratic political theory. While important first approximations of this 'revisionist' socialism were articulated in the late 19th century by the Fabian Society and the Independent Labour Party in Britain, and by the republican socialists led by Jaurès in France (see Tanner, 1997; Berman, 2006, pp. 28–35), the frankest and most influential theoretical case for social democracy in this period was made by Bernstein, who explicitly confronted the forces of Marxist orthodoxy led by Karl Kautsky within the German Sozialdemokratische Partei Deutschlands (SPD). Bernstein's ideas, most fully expressed in his 1899 book *The Preconditions of Socialism*, laid the foundations for subsequent social democratic thinking by directly contesting two doctrines central to the self-understanding of Marxism during this period: historical materialism and the class struggle. Bernstein argued that capitalism was not doomed to collapse of its own accord, as a result of inevitable internal crises and the immiseration of the mass of the population. On the contrary, he thought that capitalism had in fact shown itself to be a flexible and adaptable economic system, capable of sustaining itself for the foreseeable future. While there was therefore no inevitability to the collapse of capitalism, continued Bernstein, it was certainly possible for significant modifications to be made to its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
structure through political action. In a democratic society, progressive social reforms could significantly improve the position of the working class. In this sense, Bernstein provocatively added, ‘what is usually termed “the final goal of socialism” … is nothing to me, the movement is everything’ (Bernstein, 1898, pp. 168–9). Bernstein was equally sceptical of the doctrine of class struggle. In contrast to the widespread Marxist assumption that socialism required the working class to monopolize political power, he stressed that cross-class alliances would be necessary for socialists to enter government, and that socialism was in any case best seen as addressed to the people as a whole rather than as an ideology tethered to only one social group. In these senses, Bernstein presented social democracy as the ‘legitimate heir’ of liberalism, with its aim being ‘the development and protection of the free personality’ (Bernstein, 1899, p. 147).

Bernstein's revisionism was extremely controversial, but his strategic vision appeared increasingly relevant to party leaders after the First World War, as socialist parties began to mobilize greater political support and found themselves on the cusp of power in many nations. The Russian revolution had now established a clear distinction between two different forms of socialist struggle, the reformist and the revolutionary. In response, the socialist parties of north-west Europe (and of Australia and New Zealand) were increasingly drawn towards reformism in practice, if not always in theory. Before the Second World War, however, the experience of such parties in government was for the most part short-lived and ineffective. In Britain, France and Germany, notionally socialist parties endured traumatic periods in government. Faced by economic crisis and ultimately depression, these parties had few intellectual resources to draw upon as they found themselves fighting capitalist crisis armed only with socialist rhetoric. At the 1931 congress of the SPD, the trade union leader Fritz Tarnow aptly summarized the dilemma: ‘Are we standing at the sickbed of capitalism not only as doctors who want to heal the patient, but also as prospective heirs who can't wait for the end and would gladly help the process along with a little poison?’ (quoted in Berman, 2006, p. 110).

However, one party did manage to advance beyond this dilemma, and in effect forged the path that would later be followed by other socialist parties after 1945: the Swedish Socialdemokratiska Arbetarepartiet (SAP). In office more or less continuously from 1932, the SAP's achievement under their leader Per Albin Hansson was twofold. First, in political terms, the Swedish social democrats created a durable and popular political identity that encompassed the industrial working class but reached beyond this social constituency. In particular, the Swedish social democrats forged a cross-class alliance with the agrarian party, creating a coalition of workers and farmers that guaranteed the SAP's hold on office. Second, the social democratic-led governments of this period introduced a range of policies that would later be seen as virtually definitive of social democratic policymaking, although they were not generally regarded as such at this time. Two broad policy agendas were pursued in Sweden: first, the use of counter-cyclical measures to engineer an upturn in the economy. Influenced by the early (pre-*General Theory*) writings of Keynes and the so-called ‘Stockholm School’ of Swedish economists, the social democratic finance minister Ernst Wigforss undertook active state intervention such as public works to fight depression, in effect developing a form of ‘proto-Keynesianism’. Similar remedies had been proposed within the Labour Party in Britain and the German SPD, but only in Sweden were senior social democratic politicians sufficiently open to these ideas and in a position to implement them. Although the Swedish economy's subsequent recovery was impressive, the role of the government's expansionary polices in this recovery is debatable; their importance as a political symbol is less so. In any case, as Erik Lundberg has argued, perhaps the real merit of the SAP-led government's economic programme

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
‘consisted fundamentally in the avoidance of backward or unintelligent policy measures’, something that certainly cannot be said of every government of this period (Lundberg, 1985, p. 9). Second, and complementary to this ‘proto-Keynesianism’, the Swedish social democrats introduced a range of social welfare measures, including unemployment insurance, a housing programme, and enhanced state pensions (for further discussion of this government, see Sassoon, 1996, pp. 42–6; Tilton, 1990, pp. 39–69; Berman, 2006, pp. 152–76).

By the outbreak of the Second World War, the parameters of social democratic ideology had been established. First, social democrats were committed to parliamentary democracy rather than violent insurrection or direct democracy. This not only meant that social democrats saw peaceful, constitutional methods as the best means of reforming capitalism, but also that they saw a system of parliamentary representation as the most plausible form of democratic government and the mass party as the best vehicle for aggregating and advancing their political objectives. These democratic commitments meant that in the early 20th century social democrats often led the struggle to expand the franchise to all men and women. Second, social democrats tailored their electoral appeals to the ‘people’ as a whole and not simply to one social class. From its inception, social democracy has been understood by its advocates as aiming at the construction of cross-class coalitions. A form of ‘social patriotism’ has dominated social democratic political discourse, which presented economic redistribution as synonymous with the national interest. As Per Albin Hansson famously argued in 1928, the Swedish social democrats sought to establish Sweden as a ‘people's home’ (*folkhemmet*) where ‘no one looks down upon anyone else … and the stronger do not suppress and plunder the weaker’ (quoted in Tilton, 1990, p. 127). Third, social democrats believed that it was above all through legislation and government policy that this vision of an egalitarian society would be realized (for further discussion of these three points, see Esping-Andersen, 1985, pp. 4–11; Przeworski, 1985). Animating all three of these basic social democratic assumptions was a distinctive political theory that sought to combine a strong commitment to individual freedom and democracy with the recognition that freedom and democratic participation can only be accessed by all citizens in circumstances of relative material equality (see Jackson, 2007). Nonetheless, there remained significant disagreement before the Second World War about precisely which policies could best advance these political ideals. Many on the Left ultimately believed that some form of collective ownership of capital was the most coherent route to greater equality. The policy instruments characteristically associated with social democracy, and pioneered in Sweden, had yet to be firmly established in the minds and hearts of social democrats themselves.

Golden age

The period from 1945 to the early 1970s is often referred to as the ‘golden age’ of social democracy. One difficulty with this characterization is that social democratic parties were not always in government when many of the reforms identified as ‘social democratic’ were actually implemented. Although the electoral prospects of social democratic parties were considerably brighter than before the war, this was not a period in which social democratic parties established electoral hegemony (except in Scandinavia). A second difficulty with describing this period as a golden age for social democracy is that it could be countered that these years were above all a period of extraordinary success and dynamism for capitalism. The golden age, it might be argued, was in fact a long post-war capitalist boom, which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
brought sustained economic growth and rising living standards for all. Social democracy could therefore be said to be basically irrelevant to this more fundamental economic trend. Both of these caveats need to be taken seriously, but in my view they do not decisively undermine the case for identifying this era as a broadly social democratic one.

Three crucial post-war developments in the political economy of the advanced industrialized nations have rightly been seen as illustrating the powerful influence of social democracy in this period: the establishment of ‘social citizenship’ in Western Europe; the emergence of full employment as a legitimate, and in some cases pre-eminent, objective of government policy; and the increased role given to trade unions in economic policymaking. Post-war social conditions bred much greater popular support for these initiatives than had existed pre-war: the radicalizing impact of a ‘people's war’ against fascism, coupled with deep-seated hostility towards the maladies of unregulated capitalism manifested in the 1930s, meant that both the public and political elites were receptive to new policy frameworks of a broadly social democratic kind. The most influential authors of such frameworks, William Beveridge and John Maynard Keynes, identified themselves as progressive liberals (although it should be noted that in the 1940s, Beveridge was much more sympathetic to socialism than is often recognized: see Harris, 1997, pp. 428–43, 480). However, while the ideas of Beveridge and Keynes significantly modified earlier social democratic thinking, there was no doubt that their proposals meshed with basic social democratic aspirations, enhanced the economic credibility of social democracy as a model for public policy, and offered a practical way forward for parties of the Left previously committed to an imprecise socialist rhetoric. Conservative parties did their best to keep pace with the perceptible march leftwards of public opinion and expert advice, but as they did so they inevitably abandoned their own ideological terrain. The ideological common sense of the age was now based around progressive taxation, public spending and state intervention in the economy.

Social democrats had long emphasized the need to reshape the pattern of resource distribution thrown up by the market so that all citizens could exercise the rights associated with democratic citizenship and be ‘decommodified’, that is have access to an income based on their needs, rather than on their success or otherwise in the labour market (Esping-Andersen, 1990, pp. 21–3). The most advanced welfare states of the immediate post-war period – in Britain and Sweden – were created by social democratic governments. The 1945–51 British Labour government led by Clement Attlee enacted William Beveridge's vision of the welfare state: unemployment insurance, pensions and family allowances all largely followed Beveridge's lead, while Aneurin Bevan's universal and tax-funded National Health Service built on and in some respects deepened Beveridge's ideas. By 1950, the British sociologist T.H. Marshall could write that in Britain social rights had been added to the civil and political rights of each citizen. ‘Social citizenship’ was now a reality (Marshall, 1950). The post-war SAP governments in Sweden were slower off the mark than the Labour Party in undertaking welfare measures, but by the 1950s the celebrated universalism of the Swedish welfare state had begun to emerge. Most importantly, an expanded pension scheme introduced in 1959, the ATP, gave the middle class a stake in the welfare state. As in the 1930s, the SAP were shrewd at building cross-class alliances, this time breaking with the farmers to create a wage-earners’ coalition between blue- and white-collar employees, organized around universal, but also earnings-related, pension rights (Esping-Andersen, 1985, pp. 108–10; Eley, 2002, pp. 318–9).

Of course, the welfare state was not exclusively a social democratic initiative: other ideological traditions and political actors played a role in its creation, notably the Christian Democratic parties of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Austria, Germany and Italy. But social democracy possesses a distinctive vision of welfare, which has dominated the post-war political trajectory of certain nations (notably in Scandinavia) and which has also been influential in other national welfare systems (for example, Britain). The social democratic welfare state is universal in scope and egalitarian in its impact, rather than targeted as a residual measure at those on low incomes. And, even where these social democratic features are absent from a nation's social policy, it has often been the electoral threat from the Left that has prompted the Right to undertake its own social reforms.

The second policy development of this period that can fairly be classified as social democratic is the emergence of full (male) employment as a central political goal. The ‘right to work’, an old leftist slogan, was obviously connected to the ‘social rights’ discussed above. The post-war vision of social citizenship was premised on the assumption that every able-bodied male citizen would be in paid work, both as a matter of social justice, and for the more pragmatic reason that income tax and social insurance payments were needed to fund the new welfare state. The advantages of full employment from a social democratic perspective were clear. As the economist Thomas Balogh wrote, full employment ‘removes the need for servility, and thus alters the way of life, the relationship between the classes. It changes the balance of forces in the economy’ (quoted in Scharpf, 1991, p. 16). In comparison to the labour movement's impotence under the mass unemployment of the 1930s, the more or less full employment of the golden age strengthened the hand of labour in conflicts with capital, significantly increasing the bargaining power of wage-earners.

Prior to the golden age, social democrats had usually maintained that only a substantial socialization of the economy would enable governments to secure full employment. The post-war influence of Keynes was decisive in shifting social democratic thinking on this question, and in enabling social democrats to formulate a radical but economically credible policy agenda. Keynes was thought to have demonstrated that state regulation and intervention could in fact stabilize capitalism, since full employment and steady economic growth could be maintained by an expansionary fiscal policy aimed at sustaining economic demand in periods of economic downturn. This conviction was apparently vindicated by the economic boom of the post-war years, although it is doubtful whether some of the policies conventionally labelled as ‘Keynesian’ played a significant role in sustaining this period of economic stability. Discretionary government intervention and deficit spending did not play a large role in the political economy of the social democratic golden age. Indeed, leading social democratic nations such as Norway and Sweden had the largest budget surpluses in the Organisation for Economic Co-operation and Development (OECD) in the 1960s. More significant was the growth of public spending: the rise of the welfare state was itself a measure that helped to boost aggregate demand and stabilize the business cycle (Glyn, 1995, p. 42).

Cordial relations with the trade union movement were critical to the success of Keynesian policymaking, since under conditions of full employment some form of wage restraint was necessary to avoid inflationary pay settlements. Trade unions now emerged as central players in post-war politics and corporatist wage-bargaining was a key feature of the new economic order. Once again, Sweden was in the vanguard of this development. Two trade union economists, Gösta Rehn and Rudolf Meidner, developed the so-called ‘Rehn–Meidner model’, which incorporated an active labour market policy and solidaristic wage bargaining into a comprehensive economic strategy designed to maintain growth and employment, lower inflation and narrow income inequality. The Rehn–Meidner model was adopted by the SAP in the late 1950s and was an important influence on the policy of successive social democratic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
governments in Sweden. Under these governments, the principle of ‘equal pay for equal work’ was implemented, so that, once the fair rate for a certain job had been agreed by employers and the unions, less efficient firms could no longer enjoy the subsidy of paying their employees lower wages than those undertaking the same job elsewhere in the economy. Poorly performing firms now had an incentive to improve their efficiency or they would go out of business; the active labour market policy was designed to find work for those who found themselves unemployed as a result. Social democratic objectives were therefore harnessed to an attempt to improve economic efficiency and engineer a more dynamic economy (see Turvey, 1952; Tilton, 1990, pp. 189–214; Sassoon, 1996, pp. 203–6). Although no other social democratic party developed an economic strategy of comparable sophistication to the Rehn–Meidner model, the greater involvement of trade unions in economic policy-formation, and the widespread negotiation of agreements over pay and conditions between employers and trade unions, can be seen as a further example of the influence of social democracy after the Second World War.

It is unlikely that, shorn of these three distinctively social democratic elements, post-war capitalism could have secured the same progressive economic outcomes visible in the golden age: steady economic growth, low unemployment, low inflation, rising wages and narrowing income inequality. This radical shift in the character of capitalism inevitably prompted lively debate among social democrats about their political objectives. While many social democrats had earlier assumed that their ultimate objective (however distant) was the creation of a new order called ‘socialism’ (and thus the abolition of capitalism), it now seemed that many of their aspirations could in fact be realized in the context of a reformed capitalism. Reflecting on this point, and on the electoral failures of some social democratic parties in the 1950s and early 1960s, reformist leaders began to articulate a new social democratic revisionism, initiating fresh theoretical controversy with the Left of their parties. The SPD ostentatiously settled its account with Marxism in 1959, with its adoption of the Bad Godesberg Programme. This new statement of SPD priorities dispensed with the socialist end goal that had previously been a constitutive feature of party rhetoric and accepted that what had previously been seen as short-term goals were now exhaustive of social democratic objectives: full employment, a just distribution of wealth, and consistent economic growth. Likewise, in the British Labour Party, revisionists such as the party leader Hugh Gaitskell and his ally Anthony Crosland set out the case for a democratic socialism that discriminated more carefully between ends and means. They defined socialist ends as certain durable ethical principles, in particular the pursuit of greater economic equality, rather than as the attainment of a particular model of economic organization. Equality, argued the British revisionists, could in fact be realized through a variety of means, including some socialization of industry or collective capital ownership, but also through a strong welfare state, the regulation of the labour market, Keynesian economic management, and the progressive taxation of income and wealth. A compelling intellectual case for this was set out in Crosland's 1956 book *The Future of Socialism*, which has remained an essential reference point for debates about social democratic strategy in Britain ever since.

By the late 1960s, revisionist social democracy had therefore acquired a sharper theoretical definition and had some significant policy achievements to its credit. A revival in the electoral popularity of both Willy Brandt's SPD and Harold Wilson's Labour Party saw both parties in government as the 1960s drew to a close. But by then the golden age was also nearing its end.

Crisis and adaptation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
While social democracy was certainly an indispensable ingredient of the social capitalism that permeated western Europe after the Second World War, it would be foolish to neglect the importance of capitalism itself to this distinctive phase of economic history. Strong economic growth and rising productivity had provided the necessary economic conditions for expansive social reforms; by the mid-1970s, these conditions were no longer in place. Instead, social democrats were placed under severe pressure by a threefold crisis.

First, a global economic slowdown was triggered by various factors, including a general decline in the growth of labour productivity and increases in the price of raw materials after the OPEC oil price rises of 1973 and 1979. Growing industrial conflict between unions and employers added to the gloomy economic outlook. The result was a period of rising unemployment and inflation, and low levels of economic growth, across all industrialized nations.

Second, the 1970s saw the beginning of a marked change in the class structure of industrialized economies and a feminization of their workforces. The core constituency of social democratic parties – workers in manufacturing industries – declined as a proportion of the labour force, while there was a substantial increase in the proportion of employees working in service industries. At the same time, there was a steady rise in female employment rates, particularly in the service sector. The male breadwinner model that had been taken for granted in earlier social democratic thinking was now outdated, while the shift to service-sector employment, or so it was claimed, fractured class solidarity and thus the electoral base of social democracy. The growth of service sector employment also created problems for social democratic economic policy: increased employment in private sector services was associated with greater wage inequality, which raised the prospect of a trade-off between the two social democratic goals of full employment and greater income equality. By the 1990s, one social democratic route to resolving this trade-off, namely, increasing employment in public sector services (the policy pursued in Sweden in the 1970s and 1980s) was said to be ruled out by the budgetary constraints imposed on governments by increasingly tax-resistant electorates and financial markets nervous of budget deficits (Iversen and Wren, 1998).

Third, in response to the previous two developments, the intellectual basis of social democracy was subjected to a concerted and ingenious attack. A ‘New Right’ emerged, which was quick to attribute the blame for the economic downturn of the 1970s to ham-fisted Keynesian intervention in the market, over-powerful trade unions enforcing inflationary pay settlements and wasteful public expenditure on welfare programmes, since the latter undermined individual responsibility and required efficiency-inhibiting levels of direct taxation. Although primarily influential in English-speaking countries, the ‘neoliberal’ prescriptions of austere counter-inflationary measures, welfare state retrenchment, privatization and deregulation set the tone for the discussion of economic policy throughout the industrialized nations in the 1980s and 1990s. The policymaking autonomy of national governments was in any case reduced in this period by the (neoliberal-inspired) relaxation of capital controls across the OECD in the 1980s and the associated growth in capital mobility across national boundaries.

These developments led many to claim that the prospects for social democracy within one country were bleak, a point apparently underlined by the electoral dominance of conservative parties in some nations during the 1980s and 1990s, in particular Britain and Germany. There is no doubt that social democrats were defeated on a number of important issues in this period and were forced to adapt their programmatic objectives to take account of the changed economic and political context. But did this necessarily signal the end of the basic social democratic aspirations discussed earlier? There are three

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
reasons for thinking this latter conclusion is too strong. First, although clearly a period of social democratic retreat, social democratic parties remained electorally viable. Second, some established social democratic institutions remained resilient in the face of neoliberal attack. Third, a new revisionism had developed by the end of the 20th century, which sketched out a plausible, if modest, agenda for the pursuit of greater economic equality and full employment in the revivified global capitalism of the new millennium. This survey of social democracy will conclude by elaborating on these three points.

First, then, although frequently remembered as a period of electoral failure for the Left, in fact the 1970s, 1980s and 1990s also saw some significant social democratic victories. France was led by its first socialist head of state, François Mitterrand, from 1981 to 1995; in newly democratized Spain the socialist party under Felipe González remained in power for over a decade after 1982; in Australia Bob Hawke and then Paul Keating presided over the longest ever period of Labor government, from 1983 to 1996. The SAP largely maintained its grip on power in Sweden (albeit with periods out of office between 1976–82 and 1991–4), while the Austrian social democrats were in office throughout the 1970s and 1980s, either as a single-party government or in coalition. By the 1990s, after long periods in opposition, the British Labour Party under Tony Blair and the SPD under Gerhard Schröder had returned to power. This evidence undermines the crude sociological determinism that associates the electoral fortunes of social democratic parties with the proportion of manual workers in the labour force. In fact there has never been an obvious correlation between social democratic party support and the size of the working class. Social democratic parties have had low levels of support when there has been a high proportion of manual workers in the labour force, for example the SPD in the 1950s and 1960s, and have had high levels of support when that proportion has been in decline, for example the SAP after the 1970s (Kitschelt, 1994, p. 41). Of course, electoral success is by itself an insufficient measure of social democratic resilience. All the governments of the Left just mentioned have faced criticism for conceding too much ground to the Right, and some have been accused of implementing the dictates of neoliberal capitalist restructuring at the expense of pursuing a distinctively social democratic course. How valid are these claims?

Here we reach the second point mentioned earlier: although a perceptible swing to the Right occurred in public policy in this period, sometimes under nominally social democratic governments, this backlash was never far-reaching enough to undo the most deeply entrenched of the social democratic institutions established after 1945. Although the welfare state and progressive taxation of income were placed under considerable pressure, they remained at the heart of the public understanding of social justice in most industrialized nations with significant labour movements. Even in Britain, where the Thatcher government carried out the most thorough attack on the achievements of the golden age, Bevan's National Health Service remained intact, and quite substantial redistribution continued to take place through the tax and benefit system. As the 21st century began, 12.3 per cent of the UK population were living in poverty, but the tax and benefit system still reduced by 61 per cent the number of British citizens left in poverty by the market (Glyn, 2006, p. 171). Between 1980 and 2001 social spending actually increased as a share of GDP across the OECD. Although the rate of increase was slower than in previous decades, in this period social spending in northern European countries, those most influenced by social democracy, increased more than in ‘liberal’ economies (Glyn, 2006, pp. 165–6). While the scale of the free-market triumph should therefore be kept in proportion, equally the problems that confronted social democracy as traditionally understood should not be underestimated. In particular,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
after the 1970s the attainment of full employment became a much more elusive and apparently intractable political goal. Some social democratic governments, for example in Austria and Sweden, had initially had some success in responding to the downturn in the 1970s with a classical Keynesian reflation, but in general this period saw an increase in unemployment rates, and greater scepticism among both policymaking elites and public opinion about the capacity of governments to maintain full employment, or at least about whether it was possible to do so while simultaneously narrowing inequality.

Third, this more pessimistic political climate prompted a fresh attempt to revise party programmes in order to adapt social democratic politics to this new ‘post-industrial’ era. The resulting ‘post-industrial social democracy’ can be distilled into three elements (see Vandenbroucke, 2001; White, 1999). First, the acceptance of certain constraints on the policymaking autonomy of national governments and of some strict assumptions about how governments can maintain economic growth and stability. In particular, social democratic politicians ruled out the use of large budget deficits and significant tax increases to fund social democratic goals. Although this was sometimes presented as the result of immutable economic changes (for example ‘globalization’), it is perhaps more accurate to see this commitment as a political one. After the tumultuous events of the 1970s, social democratic parties were obliged to impress on both financial markets and the electorate their competence as economic managers and to signal that there would be no return to the high inflation and sluggish economic growth that marked the end of the golden age. While this commitment undoubtedly set strict limits on the activities of social democratic governments, it nonetheless still left some latitude for egalitarian policy interventions (remember, for example, that budget deficits were never a core feature of social democratic policymaking in the golden age). The second element of this new revisionism therefore stressed the importance of supply-side measures to improve both economic efficiency and social justice. Social democrats became increasingly interested in supplementing *ex post* redistribution through the welfare system with an attempt to equalize the *ex ante* distribution of financial assets and human capital. In particular, investment in education and training, and active labour market programmes, were seen as key instruments that would both boost employment rates and improve productivity. Similarly, social democratic policymakers expressed some interest in ensuring a more equitable distribution of financial assets through universal ‘stakeholder’ grants (Paxton, White and Maxwell, 2006). Third, the reduction of poverty and inequality through the welfare state and labour market regulation remained a social democratic priority, but with a distinctive focus on two issues: first, ensuring that the welfare state contributes to the reduction of unemployment by improving the incentives for low-paid workers to enter, and remain in, the labour market; and second, ensuring that the welfare state adequately supports family life and female participation in the workforce. The increased use of in-work benefits such as earned income tax credits was proposed to address the first of these issues, while the expansion of welfare systems to include universal nursery provision (an objective already accomplished in Scandinavia) was proposed to assist with the second. Although these aspirations were more modest than the equivalent programmes of social democratic parties in the ‘golden age’, they were nonetheless recognizably social democratic, and in crucial respects departed from neoliberalism. In particular, the revisionist emphasis on increasing employment through supply-side measures and generous in-work benefits offered an answer to the claim that in the ‘post-industrial’ economy there must be an inevitable trade-off between employment and equality. However, there was undoubtedly a tension at the heart of this revisionism: these new methods of reducing inequality and unemployment still required significant social spending,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
but the acceptance of tax resistance and budgetary restraint limited the resources available to social democratic governments. Critics were correct to point out that a social democracy cannot be obtained on the cheap.

At the beginning of the 21st century those countries most deeply influenced by social democracy, in essence the Scandinavians, still differed significantly from those in which social democracy had penetrated less thoroughly. In Sweden, 6.4 per cent of the population lived in poverty; in Britain, the equivalent figure was 12.3 per cent, and in the United States 17 per cent (Glyn, 2006, p. 171). As Ko•akowski has suggested, social democracy 'has invented no miraculous devices to bring about the perfect unity of men or universal brotherhood' (Ko•akowski, 1982, p. 11). But when we review the history of social democracy during the 20th century it is clear that it has some non-negligible achievements to its credit. In particular, social democrats have greatly improved the position of the disadvantaged in the industrialized nations, and have ensured that the interests of the poor were represented in political systems formerly monopolized by the rich.

See Also

- Bernstein, Eduard
- Beveridge, William Henry
- Fabian economics
- Keynesianism
- socialism
- welfare state

Bibliography

Berman, S. 2006. *The Primacy of Politics*. Cambridge: Cambridge University Press.

Bernstein, E. 1898. The theory of collapse and colonial policy. In *Marxism and Social Democracy: The Revisionist Debate 1896–1998*, ed. H. Tudor and J.M. Tudor. Cambridge: Cambridge University Press, 1988.

Bernstein, E. 1899. *The Preconditions of Socialism*. Cambridge: Cambridge University Press, 1993.

Crosland, C.A.R. 1956. *The Future of Socialism*. London: Jonathan Cape.

Eley, G. 2002. *Forging Democracy: The History of the Left in Europe*. Oxford: Oxford University Press.

Esping-Andersen, G. 1985. *Politics against Markets: The Social Democratic Road to Power*. Princeton: Princeton University Press.

Esping-Andersen, G. 1990. *The Three Worlds of Welfare Capitalism*. Princeton: Princeton University

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Press.

Glyn, A. 1995. Social democracy and full employment. *New Left Review* 211, 33–55.

Glyn, A. 2006. *Capitalism Unleashed*. Oxford: Oxford University Press.

Harris, J. 1997. *William Beveridge: A Biography*. Oxford: Oxford University Press.

Iversen, T. and Wren, A. 1998. Equality, employment and budgetary restraint: the trilemma of the service economy. *World Politics* 50, 507–46.

Jackson, B. 2007. *Equality and the British Left*. Manchester: Manchester University Press.

Kitschelt, H. 1994. *The Transformation of European Social Democracy*. Cambridge: Cambridge University Press.

Ko•akowski, L. 1982. What is living (and what is dead) in the social-democratic idea? *Encounter* 58, 11–17.

Lundberg, E. 1985. The rise and fall of the Swedish model. *Journal of Economic Literature* 23, 1–36.

Marshall, T.H. 1950. *Citizenship and Social Class*. Cambridge: Cambridge University Press.

Paxton, W., White, S. and Maxwell, D. 2006. *The Citizen's Stake*. Bristol: Policy Press.

Pierson, C. 2001. *Hard Choices? Social Democracy in the 21st Century*. Cambridge: Polity Press.

Przeworski, A. 1985. *Capitalism and Social Democracy*. Cambridge: Cambridge University Press.

Sassoon, D. 1996. *One Hundred Years of Socialism*. London: Fontana.

Scharpf, F. 1991. *Crisis and Choice in European Social Democracy*. Ithaca: Cornell University Press.

Stedman Jones, G. 2004. *An End to Poverty? A Historical Debate*. London: Profile.

Tanner, D. 1997. The development of British socialism 1900–18. In *An Age of Transition: British Politics 1880–1914*, ed. E.H.H. Green. Edinburgh: Edinburgh University Press.

Tilton, T. 1990. *The Political Theory of Swedish Social Democracy*. Oxford: Oxford University Press.

Turvey, R. 1952. *Wages Policy under Full Employment*. London: William Hodge.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Vandenbroucke, F. 2001. European social democracy and the third way: convergence, divisions, and shared questions. In *New Labour: The Progressive Future?* ed. S. White. Basingstoke: Palgrave.

White, S. 1999. Social liberalism, stakeholder socialism and post-industrial social democracy. *Renewal* 7, 29–38.

How to cite this article

Jackson, Ben. "social democracy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000168> doi:10.1057/9780230226203.1553

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social discount rate

Tyler Cowen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists use a social discount rate to evaluate future costs and benefits. To arrive at a suitable magnitude, market interest rates are adjusted for taxes, transactions costs, and risk. Over time the discount rate debate has become more complex. It is agreed that no single discount rate will apply to all choices, and that the proper discount rate will be context-dependent.

Keywords

Arrow, K.; cost–benefit analysis; crowding out; deadweight loss; environmental issues; gamma discounting; infrastructure investment; inter-generational policies; value of life; Posner, R.; productivity of capital; risk; social discount rate; time preference; utility

Article

Discount rates are required to evaluate the future costs and benefits of economic policies. It is widely recognized, for instance, that a dollar today is worth more than a dollar five years from now. But how much more?

All cost–benefit analyses with a temporal element must choose a rate or rates of discount. Discount rates are especially important for evaluating global warming, the loss of biodiversity, hazardous waste disposal, and related environmental issues. Infrastructure investments, such as roads, bridges, and tunnels, often last for generations. In these cases many costs and benefits come in the relatively distant future.

Consider Table 1, which compares present and future benefits using various positive rates of discount:

Estimated number of future benefits equal to one present benefit based on different discount rates

Years in the future	1%	3%	5%	10%

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

30	1.3	2.4	4.3	17.4
50	1.6	4.3	11.4	117.3
100	2.7	19.2	131.5	13,780.6
500	144.7	2,621,877.2	39,323,261,827	4.96×10^{20}

Just as economists use discount rates, so do they require a rate of compounding when measuring compensation for past injustices. For instance, if we are to make restitution for some previous loss, we must decide how to convert past dollar amounts into a present sum.

Economists (for example, Lind et al., 1982) typically suggest that a discount rate for dollars should have two components: the productivity of capital, and the social rate of time preference. The productivity of capital refers to our ability to turn present resources into a larger future value. If, for instance, we have a dollar today we can invest it and reap greater value in the future, at least on average. This is one reason why current resources are worth more than future resources. The social rate of time preference refers to our impatience. Even if we cannot invest capital productively, many of us would rather consume today than one year from now. This provides another reason why current resources might be of greater value than future resources.

Observed market rates of interest will not reflect these values with complete accuracy. First, governments typically tax the return to capital. The social return on capital is therefore greater than the private return. Second, private capital markets face some degree of transactions costs, credit rationing, and bureaucratic regulation. Once again, the social return to capital is likely higher than the private return. Third, government taxation and borrowing (to fund projects) will involve deadweight loss and crowding out. All of these variables should be reflected in any proper measure of the cost of capital, and for these reasons we should not simply take market interest rates as given.

A related debate concerns whether governments should take risk into account when using a social discount rate. Kenneth Arrow (1971) argued that the government should use a riskless rate. A riskless rate, of course, will be much lower. Short-term US Treasury securities often yield no more than a 1 per cent real rate of return; the average rate of return on private capital in the United States can run as high as 10 or 15 per cent.

According to Arrow's reasoning, government faces little or no financial risk from bad policies, given that it can spread costs across a very large number of taxpayers. Nonetheless, this argument has been criticized from at least two directions. First, private corporations can spread their risks across a large number of diversified shareholders. Second, Arrow's argument focuses too much on the purely financial side of risk. Financial losses on a project are, taken alone, mere transfers. The relevant risks include possible variation in the value of the project for its intended beneficiaries. When one measures these risks, the absolute size of the government, or the number of taxpayers, is of little concern.

Changing discount factors over time may lower the effective social discount rate (Weitzman, 2001). In this case the lowest discount rate will have a highest contribution to an expected value calculation made from the vantage point of today. Richard Posner (2004, pp. 153–4) summarizes the argument:

> Suppose there's an equal chance that the applicable interest rate throughout this and future centuries will be either 1 percent or 5 percent. The present value of $1 in 100 years is 36.9 cents if the interest rate used to compute the present value is 1 percent but only .76 cents

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(a shade over three-quarters of a cent) if it is 5 percent. Now consider the 101st year and remember the assumption that the two alternative discount rates are equally probable. If the interest rate used to discount the future to the present value is 1 percent, then the present value of $1 at the end of that year will have shrunk from 36.9 cents to 36.6 cents. If instead the interest rate used is 5 percent, the present value of .76 cents will have shrunk to about .75 cents. This means that the *average* present value of $1 at the end of the 101st year will be 18.68 cents, implying an average discount rate of less than 2 percent, rather than 3 percent. The reason is that the more rapid decline in value under the higher discount rate (5 percent) reduces its influence on present value.

In other words, when there is uncertainty about future discount rates, the lower rates have a greater relative weight the further we look into the future.

In recent times ‘gamma discounting’ has become a popular approach. Weitzman (2001, p. 260) notes, ‘even if every individual believes in a constant discount rate, the wide spread of opinion on what it should be makes the effective social discount rate decline significantly over time.’ This is for the same reason as discussed above. More generally, many individuals will argue that the present is more important than 30 years from now, but that after some point further differences in time should cease to matter. Perhaps what happens 300 years from now is not much less important than what happens 200 years from now. This view has found significant support in polls of both ordinary and distinguished economists (Weitzman, 2001).

Most of the debate has focused on discounting dollars; but the discounting of utility is a separate issue. Most likely, the discount rate on utility should be lower than the discount rate on dollars. Utility is not ‘productive’ over time as is invested capital; we are therefore left with only the rate of time preference as a discount factor. And for inter-generational policies, the rate of social time preference is arguably zero. Before individuals are born, they do not experience a disutility of waiting (Cowen and Parfit, 1992).

Another question is how much discounting should apply to very large changes in individual welfare. Recall that cost–benefit analysis is best suited to analysing small changes at the margin, where market prices (adjusted for risk, taxes, and transactions costs) measure values. So it is reasonable to argue that a dollar today is worth more than a dollar 20 years from now. But what if we are discounting lives across hundreds of years? Is a single life today really worth more than the entire survival of the human race 500 years from now? It is quite easy to generate such a conclusion, for reasonable parameter values, through the uncritical application of discounting. In these cases it appears that a more explicit ethical judgement should override the cost–benefit approach.

In sum, the debate over discount rates has seen continual progress. Most economists now agree that there is no single correct rate of social discount for all problems. Instead, the proper discount rate depends on the problem under consideration, the time horizon, the assumptions about risk, and the magnitude of the associated costs and benefits. A social discount rate is a very useful tool, and it can nicely complement our broader normative judgements. But it does not remove the need to keep the relevant ethical issues in mind.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- risk
- term structure of interest rates
- time consistency of monetary and fiscal policy

Bibliography

Arrow, K. 1971. *Essays in the Theory of Risk-Bearing*. Amsterdam: North-Holland.

Cowen, T. and Parfit, D. 1992. Against the social discount rate. In *Justice Across the Generations: Philosophy, Politics, and Society, Sixth Series*, ed. P. Laslett and J. Fishkin. New Haven: Yale University Press.

Lind, R. et al. 1982. *Discounting for Time and Risk in Energy Policy*. Baltimore, MD: Johns Hopkins University Press.

Posner, R. 2004. *Catastrophe: Risk and Response*. Oxford: Oxford University Press.

Weitzman, M. 2001. Gamma discounting. *American Economic Review* 91, 260–71.

How to cite this article

Cowen, Tyler. "social discount rate." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000464> doi:10.1057/9780230226203.1554

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social insurance and public policy

Jonathan Gruber
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social insurance expenditures are the largest and fastest-growing component of government expenditures in the developed world. The design of social insurance programmes reflects the trade-off between insurance and incentives. This article reviews the impact of social insurance programmes on both insurance against adverse events and incentives for adverse behaviour. It concludes with lessons for optimal social insurance programme design.

Keywords

administrative efficiency; adverse selection; asymmetric information; consumption smoothing; crowding out; defined benefit and defined contribution social insurance; disability insurance; expected utility; experience rating; health insurance; imperfect observability; labour market search; layoffs; leisure; mandatory saving; market failure; moral hazard; paternalism; replacement rates; retirement; risk; self-insurance; social insurance; Social Security in the United States; supply-side risk bearing; unemployment durations; unemployment insurance; workers' compensation

Article

Social insurance expenditures are the largest and fastest-growing component of government expenditures in the developed world. In the United States, for example, only about four per cent of federal government spending in 1953 was devoted to social insurance; by 2003, this had grown eleven-fold to 44 per cent of federal spending (Gruber, 2005). There has been a corresponding growth in research into the behavioural impacts of a wide variety of social insurance programmes. This article reviews that research and its implications for policy, and loosely follows the structure of Gruber (2005), Chapter 12.

At its core, the design of social insurance reflects a trade-off between insurance and incentives. As laid out nicely by Baily (1978), the optimal level of social insurance benefits sets equal the consumption

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
smoothing gains from a social insurance programme to the moral hazard costs of that programme. Programmes that serve largely to displace existing efficient forms of consumption smoothing are less valuable than programmes that greatly increase the ability to smooth consumption around adverse events. Programmes that induce large distortions by insuring adverse behaviour are less valuable than programmes that target those who are truly adversely impacted and not changing their behaviour to qualify. The research on social insurance within economics has largely focused on the latter of these issues, the moral hazard induced by social insurance programmes. More recently there has been more attention paid to the former issue, with an eye towards optimal programme design.

Institutional features of social insurance programmes

Social insurance programmes are distinguished from other types of government spending by four characteristics. Workers participate by ‘buying’ insurance through payroll taxes or mandatory contributions by themselves or their employer. These contributions make them eligible to receive benefits if some measurable event occurs, such as disability or on-the-job injury. These benefits are conditioned only on making contributions and on the occurrence of the adverse event. And they are typically not means-tested: benefits do not depend on the level of one's current income or assets.
The most common social insurance programme the world over is public pensions through programmes such as Social Security in the United States; in the United States, this is the single largest government expenditure (see Gruber and Wise, 1999, for an overview of social security programmes in developed nations). The other major source of social insurance around the world is social health insurance, such as through programmes like Medicare in the United States, which provides universal health insurance coverage to the elderly, or National Health Insurance in Canada, which provides universal health insurance coverage to the entire population; see Cutler (2002) for an overview of health insurance programmes in developed nations.

Why have social insurance?

In the canonical expected utility model, with insurance available at actuarially fair prices, optimizing consumers will choose to fully insure themselves against idiosyncratic risk. Yet, in markets without social insurance, we often observe much less than full insurance. This motivates the desire for government interventions through programmes such as social insurance.
The major theoretical justification for social insurance is *adverse selection* in private insurance markets. As noted in the classic analysis of Akerlof (1970), asymmetric information between buyers and sellers can lead to market failure, whereby trades that would be beneficial to both insured and insurer are not made. Private insurers may be wary that the individuals demanding insurance from them represent the highest risks in the population, leading the insurer to be unwilling to offer insurance at a price that corresponds to the risk of the average person. As shown in the equally classic Rothschild and Stiglitz (1976) model, this can lead to a breakdown of the insurance market, or at best to a situation where high-risk individuals buy full insurance at high prices, while low-risk individuals are only partially insured at low prices. While there is controversy over the extent of adverse selection in some insurance markets, this is clearly an enormous problem in markets such as those for annuities or health insurance. (The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
issue of adverse selection in health insurance markets is discussed in Cutler and Zeckhauser, 2000, who summarize in particular the very compelling Cutler and Reber, 1998, article on this issue. Finkelstein and Poterba, 2004, present evidence for adverse selection in annuities markets.)

While market failure is the most appealing justification to economists for social insurance, other rationales may be much more important to policymakers. One such rationale is administrative efficiency. While the government may be less efficient at producing goods than the private sector, it is clearly more efficient at providing insurance; for example, while health insurance administrative costs are 12 per cent of premiums in the private US health insurance market, they are 1.3 per cent of premiums under Canadian National Health Insurance (Gruber, 2005). Even more important in practice is paternalism. Politicians fear that individuals simply will not insure themselves appropriately against sizable risks, so that social insurance is required to ensure proper protection.

The benefits of social insurance

The arguments presented above suggest a number of reasons why private insurance markets may not make it possible for a risk-averse individual to satisfy his desire for consumption smoothing. Yet they do not suggest that consumption smoothing is completely unavailable, because individuals may have private means to smooth consumption: their own savings, labour supply of family members, borrowing from friends, and so on. The justification for social insurance depends on the extent to which social insurance is more efficient than a consumers' own private consumption smoothing mechanisms. If social insurance is simply displacing, or 'crowding out', equally efficient self-insurance, then there is no gain to the government intervention (but there may be costs, discussed below). The extent to which individuals can self-insure against adverse events will obviously vary with the predictability of the risk (for example, disability may be much less predictable than retirement), the size of the risk (for example, the total income loss from disability or retirement may be much larger than that from unemployment), and the availability of other forms of consumption smoothing (for example, own savings or spousal labour supply).

Tests for the extent of self-insurance against risk have typically proceeded by examining whether individuals can smooth their consumption across adverse events without social insurance, or, relatedly, the extent to which social insurance programmes simply crowd out other forms of self-insurance rather than smoothing consumption. For example, Gruber (1997) and Browning and Crossley (2001) find there is relatively modest consumption smoothing benefits to unemployment insurance during jobless spells. On the other hand, available evidence suggests that retirement income support only partially crowds out private retirement savings, so that there is a real impact on consumption in retirement from programmes such as Social Security. (The evidence on Social Security and savings is described in Gruber, 2005, ch. 13. International evidence on Social Security and consumption is presented in Gruber and Wise, 2008.)

There is also evidence of a crowd-out effect of public health insurance in the United States, both from the general provision of this insurance and the asset limitations that accompany qualification for these programmes (Gruber and Yelowitz, 1999). In developing countries, a very large literature finds that families are typically quite good at insuring themselves against modest risks, but not against large risks like disability (Case, 1995; Gertler and Gruber, 2002).

The limitations of these consumption-based tests, however, are highlighted by Chetty and Looney

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(2006). As they point out, individuals may be using very inefficient means of smoothing consumption, so that even if social insurance doesn't increase consumption smoothing it raises social efficiency. A further difficulty arises because even inefficient consumption smoothing mechanisms (such as savings for idiosyncratic and large risks, which are better insured by pooling across individuals) may have social benefits (if there is too little savings for other reasons such as capital taxation). Clearly, more work is needed to fully evaluate the benefits side of the social insurance equation.

The costs of social insurance

The costs side of social insurance is moral hazard, the adverse behaviour that is encouraged by insuring against an adverse event. As with adverse selection, moral hazard arises due to asymmetric information in insurance markets. By trying to insure against an adverse event (true injury), the insurer may encourage individuals to pretend that an adverse event has happened to them when it actually hasn't. For example, a common type of social insurance in developed nations is insurance against on-the-job injuries (the workers' compensation programme in the United States), which provides partial income replacement for those injured at work. Since work-based injury is often difficult to verify, however (for example for back strains or mental stress), workers may claim to be injured on the job simply to take a partially paid vacation. The rise in leisure induced by the combination of financial incentives and imperfect observability is an example of moral hazard.

Moral hazard can arise along many dimensions. In examining the effects of social insurance, three types of moral hazard play a particularly important role. The first is reduced precaution against entering the adverse state; for example, because individuals have medical insurance that covers illness, they may reduce preventive activities to protect their health. The second is increased expenditures when in the adverse state; for example, because individuals have medical insurance, they may use more medical care than they otherwise would, or because individuals have workers' compensation they don't work hard to rehabilitate their injury. Finally, there may be supplier responses to insurance against the adverse state; for example, physicians may provide too much care to those with health insurance or firms may not be careful enough in protecting workers with injury insurance.

An enormous literature has arisen in public finance to assess the importance of moral hazard in social insurance programmes. The pioneer of this literature was Martin Feldstein, whose work in the early 1970s emphasized the labour market distortions caused by Social Security, unemployment insurance, and other programmes. A large literature has followed up on these issues, as review comprehensively by Krueger and Meyer (2002). In this section, I briefly review the key conclusions from this literature.

Much of the work in this area has been focused on the effects of retirement income security programmes on retirement behaviour: by insuring the elderly against income loss from retirement, these programmes may induce retirement. The literature within countries typically concludes that these programmes are an important, but not the dominant, reason for earlier retirement. Cross-country comparisons such as Gruber and Wise (1999), however, suggest that in the long run these programmes may play the dominant role in determining retirement behaviour.

A second area of much focus has been the impact of unemployment insurance: higher unemployment insurance benefits have been shown to significantly increase unemployment durations (for example, Meyer, 1989), while there is no evidence that the longer durations are resulting in more effective job

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
search. A more limited set of studies has uncovered even larger evidence of responsiveness of injury and illness durations to the financial incentives embedded in workers' compensation and related programmes (for example, Krueger, 1991; Pettersson-Lidbom and Thoursie, 2006). The evidence for programmes that compensate for permanent disability suggests more modest effects on disability rates (for example, Gruber, 2000), which is consistent with the better ability to monitor long-run disability than short-run inability to work.

Unemployment and, to some extent, work-related injury reflect not only the decisions of employees but those of employers also. In most countries, the taxes that finance these insurance programmes are not experience-rated; that is, employers pay tax rates that do not depend on the utilization of the system by their workers. In the United States, there is experience rating, but it is only partial. A small but important literature suggests that this imperfect experience rating is a large source of adverse selection, as workers and firms combine to increase government-subsidized leisure through, for example, temporary layoffs (Topel, 1983; Anderson and Meyer, 2000).

There is also a large literature on moral hazard in medical expenditure programmes. There is clear evidence from the seminal Rand Health Insurance Experiment in the United States that the insured use care excessively, as those randomized into less generous insurance plans used less health care with no adverse impact on health (Newhouse and the Insurance Experiment Group, 1993). At the same time, a number of studies find that having no health insurance is associated with adverse health outcomes (for example, Lurie et al., 1984; Currie and Gruber, 1996a; 1996b). These findings suggest a 'medical effectiveness' curve which relates spending to health improvements: the curve is steep for initial health care use, but flattens out for the excessive care used by those with all health expenditures covered.

As with other social insurance programmes, there are also important supply-side moral hazard issues in health care. A number of studies document the importance of reimbursement structures for the treatment of patients by health providers, finding in particular that prospective reimbursement systems that ensure that providers bear part of the risk for excessive treatment lead to less utilization without adverse health impacts (Newhouse, 1996).

Optimal social insurance

These sets of findings have important implications for the optimal design of social insurance programmes. First, social insurance should be only partial. For example, replacement rates under labour market-based programmes (the extent to which social insurance benefits replace earnings before the adverse event) should be less than full, much less so in several cases. The fact that consumption smoothing is not full, while moral hazard is important, suggests in the framework of Baily (1978) less than full replacement. There is too little work on most of these programmes to state the optimal replacement rates with confidence, but there are clear directions for reform. The enormous moral hazard associated with Workers' Compensation in the United States or with social security in many European nations, for example, suggests that there may be welfare gains from reducing benefit levels.

Another important example is health care, where the optimal insurance policy is one in which individuals bear a large share of medical costs within some affordable range, and are only fully insured when costs become unaffordable. This structure is optimal because coverage for all medical expenditures (what is often called 'first dollar coverage', since all dollars of medical spending, starting

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
with the first dollar, are covered) has little consumption smoothing benefit, but a large moral hazard cost. The consumption smoothing benefit from first dollar coverage is small because there is little utility gain to risk-averse individuals from insuring a small risk. At the same time, this first dollar coverage has substantial moral hazard cost because it encourages individuals to overuse the medical system, demanding care for which the social costs exceed the social benefits. An example of an optimal insurance plan would be Feldstein's (1973) ‘major risk insurance’ plan, in which individuals would face a high co-payment (such as 50 per cent) on all services until they spent a sizable share of their income (such as ten per cent) on medical care, beyond which there would be no more co-payments.

The second lesson is that there should be more supply-side risk bearing in insured markets. For example, more tightly experience rating employers for the costs of unemployed or injured workers could significantly reduce overuse of these social insurance systems. And more risk bearing by providers in the medical system has been shown to reduce utilization without adversely impacting health.

Finally, these findings also open the question of moving from the traditional ‘defined benefit’ approach to social insurance to a ‘defined contribution’ approach where individuals are mandated to provide for their own protection against adverse events. This model has been seen most forcefully in the debate over US Social Security privatization, whereby the traditional programme under which today's workers save collectively for today's retirees would be replaced by one where individuals saved for their own retirement. Partial privatization is in use in a number of countries, but is a source of contentious debate in the United States (see Gruber, 2005, ch. 13, for a brief review of the arguments on both sides of this debate).

This debate has not generally proceeded, however, to other social insurance programmes. For example, Feldstein and Altman (1998) have suggested a programme under which individuals save in their own unemployment insurance accounts for jobless spells. Using such mandatory savings accounts to finance short-run risks can potentially provide the consumption smoothing benefits of unemployment insurance while reducing the moral hazard costs of the programme. This is an approach worth consideration in other social insurance contexts, where the essential question becomes the value of the redistribution that is lost by such ‘self-insurance’ approaches.

See Also

- adverse selection
- moral hazard
- pensions
- social insurance
- Social Security in the United States

Bibliography

Akerlof, G.A. 1970. The market for lemons: quality uncertainty and the market mechanism. *Quarterly Journal of Economic* 84, 488–500.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Anderson, P.M. and Meyer, B.D. 2000. The effects of the unemployment insurance payroll tax on wages, employment, claims and denials. *Journal of Public Economics* 78, 81–106.

Baily, M. 1978. Some aspects of optimal unemployment insurance. *Journal of Public Economics* 10, 379–402.

Browning, M. and Crossley, T. 2001. Unemployment insurance levels and consumption changes. *Journal of Public Economics* 80, 1–23.

Case, A. 1995. Symposium on consumption smoothing in developing countries. *Journal of Economic Perspectives* 9(3), 81–82.

Chetty, R. and Looney, A. 2006. Consumption smoothing and the welfare consequences of social insurance in developing economies. *Journal of Public Economics* 90, 2351–6.

Currie, J. and Gruber, J. 1996a. Saving babies: the efficacy and cost of recent changes in the Medicaid eligibility of pregnant women. *Journal of Political Economy* 104, 1263–96.

Currie, J. and Gruber, J. 1996b. Health insurance eligibility, utilization of medical care, and child health. *Quarterly Journal of Economics* 111, 431–66.

Cutler, D. 2002. Equality, efficiency and market fundamentals: the dynamics of international medical care reform. *Journal of Economic Literature* 40, 881–906.

Cutler, D.M. and Reber, S.J. 1998. Paying for health insurance: the trade-off between competition and adverse selection. *Quarterly Journal of Economics* 113, 433–66.

Cutler, D. and Zeckhauser, R. 2000. The anatomy of health insurance. In *Handbook of Health Economics*, vol. IA, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Feldstein, M.S. 1973. The welfare loss of excess health insurance. *Journal of Political Economy* 81, 251–80.

Feldstein, M.S. and Altman, D. 1998. Unemployment insurance savings accounts. Working Paper No. 6860. Cambridge, MA: NBER.

Finkelstein, A. and Poterba, J. 2004. Adverse selection in insurance markets: policyholder evidence from the U.K. annuity market. *Journal of Political Economy* 112, 183–208.

Gertler, P. and Gruber, J. 2002. Insuring consumption against illness. *American Economic Review* 92, 51–70.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gruber, J. 1997. The consumption smoothing benefits of unemployment insurance. *American Economic Review* 87, 192–205.

Gruber, J. 2000. Disability insurance benefits and labor supply. *Journal of Political Economy* 108, 1162–83.

Gruber, J. 2005. *Public Finance and Public Policy*, 1st edn. New York: Worth Publishers.

Gruber, J. and Wise, D.A. 1999. Introduction and summary. In *Social Security and Retirement around the World*, ed. J. Gruber and D.A. Wise. Chicago: University of Chicago Press.

Gruber, J. and Wise, D.A. 2008. *Social Security and Well-Being Around the World*. Chicago: University of Chicago Press.

Gruber, J. and Yelowitz, A. 1999. Public health insurance and private savings. *Journal of Political Economy* 107, 1249–74.

Krueger, A.B. 1991. Workers' compensation insurance and the duration of workplace injuries. Mimeo, Princeton University.

Krueger, A.B. and Meyer, B.D. 2002. Labor supply effects of social insurance. In *Handbook of Public Economics*, vol. 4, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Lurie, N., Ward, N.B., Shapiro, M.F. and Brook, R.H. 1984. Termination from Medi-Cal – does it affect health? *New England Journal of Medicine* 311, 480–4.

Meyer, B.D. 1989. A quasi-experimental approach to the effects of unemployment insurance. Working Paper No. 3159. Cambridge, MA: NBER.

Newhouse, J. 1996. Reimbursing health plans and health providers: selection vs. efficiency in production. *Journal of Economic Literature* 34, 1236–63.

Newhouse, J.P. and the Insurance Experiment Group. 1993. *Free For All? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

Pettersson-Lidbom, P. and Thoursie, P.S. 2006. Temporary disability insurance and labor supply: evidence from a natural experiment. Mimeo, Stockholm University.

Rothschild, M. and Stiglitz, J. 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90, 629–49.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Topel, R.H. 1983. On layoffs and unemployment insurance. *American Economic Review* 73, 541–59.

How to cite this article

Gruber, Jonathan. "social insurance and public policy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000525> doi:10.1057/9780230226203.1556

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social insurance

Stefania Albanesi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Individuals face uncertainty about their labour income, due to chronic and temporary medical conditions or to unemployment spells. The individual burden of these risks can be eased by social insurance. Income taxation and government transfer programmes, such as disability and unemployment insurance, are the most common forms of social insurance. The challenge in the design of social insurance programmes is the impossibility of fully distinguishing between low income by choice and low income by necessity. This article reviews leading theories of optimal social insurance under private information and points to possible directions of future research.

Keywords

adverse selection; altruism; consumption inequality; consumption smoothing; continuation values; depreciation; disability insurance; Euler equations; expected utility; human capital; immizeration; incentive compatibility; incentives; infinite horizons; lump-sum taxes; Mirrlees, J.; moral hazard; observability; optimal taxation; principal and agent; private information; private insurance; progressive and regressive taxation; promised values; risk sharing; search models of unemployment; self-insurance; social insurance; taxation of capital income; taxation of income; training; unemployment insurance; welfare-to-work programmes

Article

Individuals face uncertainty about their labour income. Chronic medical conditions can arise that reduce or remove their ability to work. Alternatively, unemployment spells can occur, temporarily reducing their earnings. The individual burden of these risks can be eased by private insurance and by social insurance schemes. A social insurance scheme is a government transfer programme whereby individuals who claim a condition or state that reduces their labour income, such as disability or unemployment, obtain a transfer from the government for the duration of this state. Social insurance can also be carried

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
out through income taxation.
The design of optimal social insurance is a classic problem in economics. This problem is interesting because neither private insurance or public arrangements can fully distinguish between low income by choice and low income by necessity. For example, the diagnosis of medical factors is often subjective, and symptoms for certain conditions, such as back pain or recurrent migraines, are hard to verify. If full insurance for chronic medical conditions were available, individuals could exploit this by reporting fake symptoms to claim insurance benefits and stop working. Similarly, an unemployed worker's ability to find new employment depends on the effort she exerts in the job search. If full unemployment insurance is available, an unemployed worker would have no incentive to search for a new job if her search effort cannot be monitored. Private information on earning ability implies that full insurance might defeat itself by removing the incentive to work. Hence, the essential trade-off in the design of optimal social insurance schemes is the one between risk sharing and incentives.

The trade-off between insurance and incentives

Consider the following simple social insurance problem (the set-up for this example is adapted from Diamond and Mirrlees, 1978). Individuals live for two periods. In the first period of their lives, they are endowed with one unit of consumption and they consume. In the second period, they consume and they may work if they are able to. The probability of being able to work is π, a number between zero and 1. Work is publicly observable and the probability distribution of ability is known, but ability is private information. Preferences over consumption in each period are represented by $u(c)$, where $c \geq 0$ is consumption and the function u is strictly increasing and strictly concave. The utility cost of working if able is γ, a number strictly greater than zero. An individual produces one unit of output if she works, zero otherwise.
The government wishes to induce able agents to work in period two and to maximize their lifetime expected utility, given by:

$$u(c_1) + \pi u(c_2^w) + (1 - \pi)\, u(c_2^n) - \pi\gamma,$$

where c_1 is consumption in period one and c_2^w and c_2^n are consumption conditional on work and no work in period two.
There is no private insurance, but the government is able to pool risk by distributing consumption, conditional on work in period two, so that the expected value of lifetime consumption equals the expected value of lifetime work for any individual. Individuals cannot save but the government has access to a storage technology with return $R > 0$. Hence, an individual's lifetime consumption profile must satisfy the constraint:

$$Rc_1 + \pi c_2^w + (1 - \pi)\, c_2^n \leq R + \pi.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(1)

Since the government cannot observe ability, the consumption profile must also satisfy an incentive compatibility constraint:

$$u(c_2^w) - \gamma \geq u(c_2^n),$$

(2)

to ensure that able individuals will work.
The optimal social insurance scheme satisfies the following two inequalities when the incentive compatibility constraint (2) is binding:

$$\frac{u'(c_2^w)}{u'(c_2^n)} < 1,$$

(3)

$$\frac{u'(c_1)}{Eu'(c_2)} < R,$$

(4)

where $Eu'(c_2)$ is the expected marginal utility of consumption in period two.
It immediately follows from (3) that full insurance is not possible and $c_2^w > c_2^n$. Hence, private information implies that the first-best optimum is not attainable. Equation (4) is derived from the government's Euler equation and uncovers another dimension of the incentive problem. It implies that the consumption path associated with the optimal insurance scheme displays a wedge between the marginal intertemporal rate of substitution and the intertemporal rate of transformation, *R*. This intertemporal wedge indicates that the marginal cost of transferring consumption to period two is greater than the value of forgone consumption at time one. The additional cost stems from the need to maintain

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
incentive compatibility. This requires consumption to be conditional on work and therefore stochastic in the second period. Since utility from consumption is strictly concave, a given increase in expected utility requires more resources when consumption is stochastic rather than deterministic.

The intertemporal wedge implies that individuals would prefer to reduce their consumption in the first period. Thus, in general, the optimal intertemporal consumption profile cannot be obtained without preventing access to the capital market or imposing a tax on savings. For example, Golosov and Tsyvinski (2006) study optimal disability insurance in competitive equilibrium and show that the intertemporal wedge requires disability benefits to be asset-tested.

Private information on individuals' ability to work gives rise to adverse selection in the social insurance problem. The resulting partial insurance and the presence of an intertemporal wedge (4) hold generally with adverse selection (Golosov, Kocherlakota and Tsyvinski, 2003) and also characterize social insurance under moral hazard (Rogerson, 1985), as in the unemployment example described in the introduction. Moreover, (4) exemplifies a general feature of social insurance with private information, that is, that the government would want to control trade in related commodities (Atkinson and Stiglitz, 1976; Varian, 1980).

Long-run inequality

Social insurance models with infinite horizons generate normative implications for long-run consumption inequality. The key step in deriving these implications is to formulate the social insurance problem recursively. The pioneering work of Green (1987) provides an early example of a recursive solution to a dynamic social insurance problem with infinite horizon and constant discounting. The main insight is to restate the optimal social insurance scheme as a contract between the government (the principal) and individuals (the agents). Such a contract retains memory of the history of outcomes and assigns current transfers and promises of future transfers based on that history and the current outcome. Despite their large dimensionality, such histories can be summarized by agents' *promised value*. (These results can be found in Spear and Srivastava, 1987; Thomas and Worrall, 1990; Abreu, Pierce and Stacchetti, 1990; Phelan and Townsend, 1991.) This one-dimensional object, which corresponds to an agent's expected lifetime utility at a point in time, encodes an agent's history and permits a recursive formulation of the problem. Incentive compatibility implies that current transfers and promises of future transfers will depend on promised utility and on the current outcome. The government's promises of future transfers can be represented as *continuation values* that determine agents' promised utility in the subsequent period.

For example, in an infinite horizon version of the earlier disability insurance example, the individual history for some period $t > 1$ is the sequence of work outcomes from period 1 to period $t - 1$. This can be summarized by an individual's expected lifetime utility at the beginning of time t, her promised value. The optimal consumption allocation at time t and the continuation value will depend on the promised value and on the current work outcome. The continuation value corresponds to an individual's expected lifetime utility in the subsequent period. Hence, individuals' future consumption depends on current work and on the past history of work.

The history dependence of the consumption path resulting from an optimal insurance scheme implies that the trade-off between insurance and incentives shapes the evolution of the consumption distribution.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Specifically, incentive compatibility implies that not only current transfers but also continuation values will be conditional on current outcomes. An immediate consequence is that the degree of consumption inequality tends to continually increase. An additional implication can be derived from the intertemporal wedge, which implies that individuals will face a downward-sloping path of promised lifetime utility under the optimal social insurance scheme.

Taken together, these results give rise to an extreme conclusion. Consumption inequality should grow without bound, with all individuals in the population converging to their minimum promised lifetime utility, except for a vanishing fraction converging to their bliss point. This *immizeration* property is robust. It obtains in partial (Green, 1987; Thomas and Worrall, 1990) and general (Atkeson and Lucas, 1992) equilibrium, under weak assumptions on preferences (Phelan, 1998), and holds in adverse selection and, in somewhat weaker form (Pavoni, 2004), in moral hazard environments.

The radical implications for consumption inequality generated by optimal social insurance models have prompted research on alternative normative criteria for the government's problem, based on an intergenerational interpretation of the infinite horizon framework. In standard models of social insurance, future generations are considered only indirectly, via the altruism of the earlier ones. Phelan (2006) proposes a government objective with equal weight on all future generations and shows that this implies a finite amount of inequality in the limit. Farhi and Werning (2005) explore a class of social insurance allocations that take into account individuals currently alive, as well as future generations, by assigning the latter a vanishingly small weight in the government's objective. They find that long-run inequality remains bounded and all individuals avoid misery.

The recursive principal–agent approach to social insurance problems underlies most macroeconomic applications. Two recently prominent areas of interest are optimal income taxation with unobserved skills and optimal unemployment insurance with hidden effort.

Optimal income taxation

Income taxes are an important instrument for social insurance. Hence, the basic trade-off between insurance and incentives also underlies the design of optimal tax systems. This intuition drives Mirrlees's (1971) seminal study of optimal income taxes. The main assumption is that labour income is observable but it depends on individual effort and skills, which are private information. Taxes are restricted to depend on labour income only, but, conditional on this, the government is relatively unconstrained. Lump-sum taxes and arbitrarily progressive or regressive tax schemes can all be part of the armoury. Mirrlees studies a static economy and finds that optimal marginal income taxes are low and slightly declining in income. Diamond (1998) and Saez (2001) find that marginal income taxes are high and sharply increasing in income at low income levels. Diamond and Saez's results can be interpreted as a prescription for a rapid phase-out of social benefits for low-income individuals, which is consistent with the US system. The properties of the optimal marginal income tax are very sensitive on the assumed properties of the skill distribution, which explains the difference in findings.

Albanesi and Sleet (2006) and Kocherlakota (2005) apply Mirrlees's approach to dynamic economies and derive implications for optimal capital and labour income taxes. The optimal taxes depend on the agents' history. This is achieved by conditioning the tax payments on the entire history of labour income, or, equivalently, on outstanding wealth when skill shocks are i.i.d. The properties of optimal

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
capital income taxes stem from the intertemporal wedge associated with the optimal consumption path. As noted earlier, this wedge implies that individuals would like to save more for two reasons: their lifetime consumption path is downward-sloping and they face consumption risk, given that the optimal scheme provides only partial insurance. Capital income must then be taxed to prevent this excessive saving. The optimal marginal asset tax, however, has a very specific form: it is decreasing in labour income. Hence, it is stochastic and negatively correlated with consumption. Excessive saving is discouraged by making after-tax returns on assets correlate positively with labour income, and thus reducing the hedging value of holding assets. Albanesi and Sleet (2006) also study the properties of marginal labour income taxes and find that they should be high at low income and decreasing in wealth. Kocherlakota (2006) provides an extensive review of these findings. Albanesi (2006) studies optimal capital income taxes in economies with moral hazard and idiosyncratic capital returns. She shows that the intertemporal wedge can be negative in this class of economies. In this case, the optimal marginal capital income taxes are increasing in income.

Unemployment insurance

Hopenhayn and Nicolini (1997) consider the design of optimal unemployment insurance under moral hazard. The probability of finding a new job depends on the unemployed worker's search effort, which is private information. (Shimer and Werning, 2005, analyse optimal unemployment insurance in a search model with adverse selection.) The unemployed worker is risk averse and cannot borrow or save. Upon finding a new job, the worker will be employed for ever at a fixed wage. The optimal unemployment insurance scheme is self-financing and comprises two elements: a sequence of unemployment benefits and a lump-sum tax levied when the worker finds new employment. The presence of moral hazard implies that the replacement ratio – that is, the fraction of previous salary transferred to the worker in the form of unemployment benefits – must be strictly smaller than one when the worker is searching for a new job. Hence, the optimal scheme provides only partial insurance against unemployment.

The main results are that unemployment benefits should be decreasing over the course of the unemployment spell and the employment tax is increasing with the length of the unemployment spell, under mild conditions on preferences. The decreasing benefits result – first derived in the seminal paper of Shavell and Weiss (1979) – is a manifestation of the intertemporal wedge in this setting. The intertemporal wedge implies that promised utility should be decreasing as long as the worker is unemployed, which requires unemployment benefits to decrease over time. The employment tax result stems from consumption smoothing, since continuation utility rises discretely when a worker becomes employed. By taxing a newly employed worker, the government optimally smooths this jump in consumption. A worker's promised utility declines further for longer unemployment spells, hence, the employment tax is increasing with the length of unemployment.

Pavoni (2004) enriches Hopenhayn and Nicolini's model by introducing a realistic feature: a worker's human capital depreciates over the unemployment spell. The optimal unemployment insurance programme displays two novel features. First, if human capital depreciates rapidly enough during unemployment, benefits are bounded below by a minimal assistance level. Second, the optimal employment tax should decrease with the length of the unemployment spell. These new findings are a consequence of the fact that a worker's wage upon employment depends positively on her human capital.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
As human capital depreciates over the length of the unemployment spell, expected utility upon employment declines. This reduces and eventually eliminates the government's need to decrease benefits over time to induce a decline in promised lifetime utility over the unemployment spell. Similarly, it reduces the value of the employment tax required to smooth consumption in the transition from unemployment to employment.

More recently, Pavoni and Violante (2007) study the optimal design of welfare-to-work programmes. These programmes are a mix of government expenditures on passive policies, such as unemployment insurance and social assistance, and active policies, such as job-search monitoring, training and wage taxes or subsidies, targeted to the unemployed. Most governments in fact use a combination of passive and active policies in dealing with unemployment. There are several novel findings. First, the optimal welfare-to-work programme endogenously generates a permanent policy of last resort, which resembles a social assistance programme. The unemployed worker is given a constant lifetime benefit and is not active in job search or training. Second, optimal unemployment benefits are generally decreasing or constant during unemployment, but they must increase after a successful spell of training. These findings result from the fact that utility conditional on employment decreases with the length of the unemployment spell and increases after job training. The central assumption is once again human capital depreciation over the unemployment spell. Indeed, it is shown that human capital depreciation is necessary for policy transition to be part of an optimal welfare-to-work programme. Pavoni and Violante find that, by providing more insurance to skilled workers and more incentives to unskilled workers, the optimal welfare-to-work programme delivers significant welfare gains with respect to the existing US system.

Concluding remarks

Most studies of optimal social insurance exclude the presence of private insurance contracts. Does the government have a special role in the provision of insurance with private information? The seminal work of Prescott and Townsend (1984) demonstrates that the first welfare theorem holds for a large class of economies with private information. Golosov and Tsyvinski (2007) study an adverse selection economy and allow for private insurance provision. They show that, if all trades are observable and individuals can sign binding contracts with private insurers *ex ante*, the only effect of public provision of social insurance is crowding out of private insurance. On the other hand, if certain trades are not observable, privately provided insurance is not Pareto optimal and government policies can increase welfare. Bisin and Guatioli (2004) examine a moral hazard economy and show that, if agents' contractual relationships with competing insurance providers cannot be monitored, the competitive equilibrium allocation is not Pareto optimal.

The observability of trades allows insurance providers to restrict participation in additional contractual relationships with other agents. Absent this, individuals can undo the incentive effects of an insurance contract by purchasing additional insurance or engaging in trades that provide a limited amount of self-insurance. These results suggest that, if exclusivity of private insurance contracts cannot be enforced, government provision of insurance plays a critical role.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- optimal taxation
- social insurance and public policy
- unemployment insurance

I wish to thank Claudia Olivetti, Nicola Pavoni and Bruce Preston for very helpful comments and discussion.

Bibliography

Abreu, D., Pierce, D. and Stacchetti, E. 1990. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica* 58, 1041–63.

Albanesi, S. 2006. Optimal taxation of risky productive capital with private information. Working Paper No. 12419. Cambridge, MA: NBER.

Albanesi, S. and Sleet, C. 2006. Dynamic optimal taxation with private information. *Review of Economic Studies* 73, 1–30.

Atkeson, A. and Lucas, R. 1992. On efficient distribution with private information. *Review of Economic Studies* 59, 427–53.

Atkinson, A. and Stiglitz, J. 1976. The design of the tax structure: direct versus indirect taxation. *Journal of Public Economics* 6, 55–75.

Bisin, A. and Guatioli, D. 2004. Moral hazard and non-exclusive contracts. *Rand Journal of Economics* 35, 306–28.

Diamond, P. 1998. Optimal taxation: an example with a U–shaped pattern of optimal marginal tax rates. *American Economic Review* 88, 83–95.

Diamond, P. and Mirrlees, J. 1978. A model of social insurance with variable retirement. *Journal of Public Economics* 10, 295–336.

Farhi, E. and Werning, I. 2005. Inequality, social discounting and progressive estate taxation. Working Paper No. 11408. Cambridge, MA: NBER.

Golosov, M., Kocherlakota, N. and Tsyvinski, A. 2003. Optimal indirect and capital taxation. *Review of Economic Studies* 70, 569–87.

Golosov, M. and Tsyvinski, A. 2006. Designing optimal disability insurance: a case for asset testing.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Journal of Political Economy 114, 257–79.

Golosov, M. and Tsyvinski, A. 2007. Optimal taxation with endogenous insurance markets. *Quarterly Journal of Economics* 122, 487–534.

Green, E. 1987. Lending and the smoothing of uninsurable income. In *Contractual Arrangements for Intertemporal Trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

Hopenhayn, H. and Nicolini, J. 1997. Optimal unemployment insurance. *Journal of Political Economy* 105, 412–38.

Kocherlakota, N. 2005. Zero expected wealth taxes: a Mirrlees approach to dynamic optimal taxation. *Econometrica* 73, 1587–621.

Kocherlakota, N. 2006. Advances in dynamic optimal taxation. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, vol. 1, ed. R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Pavoni, N. 2004. Optimal unemployment insurance, with human capital depreciation, and duration dependence. Mimeo, University College London.

Pavoni, N. and Violante, G. 2007. Optimal welfare-to-work programs. *Review of Economic Studies* 74, 283–318.

Phelan, C. 1998. On the long run implications of repeated moral hazard. *Journal of Economic Theory* 79, 174–91.

Phelan, C. 2006. Opportunity and social mobility. *Review of Economic Studies* 73, 487–504.

Phelan, C. and Townsend, R. 1991. Computing multi-period, information constrained optima. *Review of Economic Studies* 58, 853–81.

Prescott, E. and Townsend, R. 1984. Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica* 52, 21–45.

Rogerson, W. 1985. Repeated moral hazard. *Econometrica* 53, 69–76.

Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

205–29.

Shavell, S. and Weiss, L. 1979. The optimal payment of unemployment insurance benefits over time. *Journal of Political Economy* 87, 1347–62.

Shimer, R. and Werning, I. 2005. Liquidity and insurance for the unemployed. Working Paper No. 11689. Cambridge, MA: NBER.

Spear, S. and Srivastava, S. 1987. On repeated moral hazard with discounting. *Review of Economic Studies* 54, 599–617.

Thomas, J. and Worrall, T. 1990. Income fluctuations and asymmetric information: an example of a repeated principal agent problem. *Journal of Economic Theory* 51, 367–90.

Varian, H. 1980. Redistributive taxation as social insurance. *Journal of Public Economics* 14, 49–68.

How to cite this article

Albanesi, Stefania. "social insurance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000494> doi:10.1057/9780230226203.1555

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social interactions (empirics)

Yannis M. Ioannides

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Empirical studies of social interactions address a multitude of definitional, econometric and measurement issues associated with the role of interpersonal and social group influences in economic decisions. Applications range from studies of crime patterns, neighbourhood influences on upbringing and conformist behaviour, mutual influences among classmates and keeping up with roommates in colleges regarding academic and social activities, to herding and to learning about social services. The article reviews several instances of successful identification of effects emanating from others' behaviour as distinct from characteristics of others. Data-sets with increasingly rich contextual information will allow estimation of complex models of economic decisions.

Keywords

congestion; contextual effects; correlated effects; endogenous effects; herding; identification; information sharing; innovation; laboratory experiments; linear models; multiplier effects; natural experiments; neighbourhood effects; neighbours; partial identification; pecuniary externalities; peer effects; rational expectations; reflection problem; self-selection; social capital; social equilibrium; social interactions; social multipliers; social norms; unobservables; well-being

Article

The empirical economics literature on social interactions addresses the significance of the social context in economic decisions. Decisions of individuals who share a social milieu are likely to be interdependent. Recognizing the nature of such interdependence in a variety of conventional and unconventional settings and measuring empirically the role of social interactions poses complex econometric questions. Their resolution may be critical for a multitude of phenomena in economic and social life and of matters of public policy. Questions like why some countries are Catholic and others Protestant, why crime rates vary so much across cities in the same country, why fads exist and survive,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and why there is residential segregation and neighbourhood tipping are all in principle issues that may be examined as social interactions phenomena.

The social context enters in a variety of ways. One is that individuals care not only about their own purely private outcomes – for example, the kinds of cars they drive or the education they acquire – but also about outcomes of others, such as the kinds of cars or the education of their friends. This type of interpersonal effect is known as *endogenous* social effect (or interaction), because it depends on *decisions* of others in the same social milieu. Individuals may also care about personal characteristics of others, that is, whether they are young or old, black or white, rich or poor, trendy or conventional, and so on, and about other attributes of the social milieu that may not be properly characterized as deliberate decisions of others. The latter is known as exogenous social or *contextual* effect. In addition, individuals in the same or similar social settings tend to act similarly because they share common unobservable factors. Such an interaction pattern is known as *correlated* effects. This terminology is due to Manski (1993).

Emergence of social interdependencies is natural if individuals share a common resource or space in a way that is not paid but still generates constraints on individual action. This is also known as pecuniary externalities. Individuals who try to form expectations about future outcomes of current decisions, like occupational choice, may rely on lessons from the actions of others and therefore end up mimicking their behaviour. Endo-genous social interactions are a case of real externalities, a pervasive feature of economic behaviour.

Theorizing in this area must lie in the interface of economics, sociology and psychology, and often is imprecise. Terms like social interactions, neighbourhood effects, social capital and peer effects are often used as synonyms, although they may have different connotations. Empirical distinctions between endogenous, contextual and correlated effects are critical for policy analysis because of the ‘social multiplier,’ as we see further below.

Joint dependence among individuals’ decisions *and* characteristics within a social milieu is complicated further by the fact that in many interesting circumstances individuals in effect choose the social context. For example, individuals choose their friends and their neighbourhoods and thus their neighbourhood effects as well. Such choices involve information that is in part unobservable to the analyst, and therefore require making inferences among the possible factors which contribute to decisions (Brock and Durlauf, 2001; Moffitt, 2001). The present article focuses on highlighting the significance of key empirical findings and owes a lot to Durlauf (2004), the most comprehensive review to date that examines the methodological basis, statistical reliability and conceptual and empirical breadth of the neighbourhood effects literature.

Empirical framework

Let individual i's outcome ω_i, a scalar, be a linear function of a vector of observable individual characteristics, X_i, of a vector of contextual effects, $Y_{n(i)}$, which describe i's neighbourhood $n(i)$, and of the expected value of the ω_j's of the members of neighbourhood $n(i)$, $j \in n(i)$. It is straightforward to incorporate social interactions into economic models in a manner that is fully compatible with economic reasoning, that is, by positing that individuals maximize a utility function subject to constraints and obtain a behavioural equation such as:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\omega_i = k + cX_i + dY_{n(i)} + Jm_{n(i)} + \varepsilon_i, \quad (1)$$

where ε_i is a random error and *k* a constant. Ignore for the moment the fact that individual *i* may have deliberately chosen neighborhood *n*(*i*). The assumption that the expectation of ε_I in (1) is zero, conditional on individual characteristics, on contextual effects and on the event that *i* is a member of neighbourhood *n*(*i*), allows us to focus on the estimation of the model. The critical next step for translating theoretical models into empirical applications is to assume *social equilibrium* and that individuals hold *rational expectations* over $m_{n(i)}$. That is, individuals' expectations are confirmed in that they are exactly equal to what the model predicts on average. So, taking the expectation of ω_i and setting it equal to $m_{n(i)}$ allows us to solve for $m_{n(i)}$. Substituting back into (1) yields a *reduced form equation*, an expression for individual *i*'s outcome in terms of all observables:

$$\omega_i = \frac{k}{1-J} + cX_i + \frac{J}{1-J}cX_{n(i)} + \frac{d}{1-J}Y_{n(i)} + \varepsilon_i. \quad (2)$$

This simple linear model obscures the richness that nonlinear social interactions models make possible, like multiplicity of equilibria (Brock and Durlauf, 2001). Yet it does facilitate studying other aspects. For example, it does confirm that endogenous social effects generate feedbacks which magnify the effects of neighbourhood characteristics. That is, the effect of a unit increase in $Y_{n(i)}$ is $\frac{d}{1-J}$, and not just *d*, as one would expect from (1). It also confirms why it is tempting for empirical researchers to study individual outcomes as functions of all observables. Following the pioneering work of Datcher (1982), a great variety of individual outcomes have been studied in the context of different neighbourhoods and typically significant effects have been found. Deriving causal results requires suitable data.
Manski (1993) emphasized that the practice of including neighbourhood averages of individual effects as contextual effects, $Y_{n(i)} = X_{n(i)}$, may cause failure of identification of endogenous as distinct from exogenous interactions, that is, to estimate *J* separately from *d*. That is, if the neighbourhood attributes *coincide* with the neighbourhood averages of its inhabitants' characteristics, or $Y_{n(i)} = X_{n(i)}$, then regressing individual outcomes on neighbourhood averages of individual characteristics as contextual effects allows us to estimate a function of the parameters of interest, $\frac{Jc+d}{1-J}$, the coefficient of $X_{n(i)}$ in a regression according to (2). A statistically significant estimate of this coefficient implies that at least one type of social interaction is present, either *J* or *d* or both are non-zero. This is known as Manski's

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

reflection problem, which is specific to *linear* models: the equilibrium value of the outcome $m_{n(i)}$ is linearly related to the neighbourhood attributes, and therefore its effect on individual outcomes may not be distinguishable from their 'reflection'.

Complicating the basic model in natural ways, as by assuming correlated effects – like group members sharing group-specific unobservable effects, that is, the performance of students in the same class is affected by the quality of their teacher, which is unobservable, over and above peer effects from classmates – introduces additional difficulties with identification, even if individuals are randomly assigned to groups. Brock and Durlauf (2007) provide an exhaustive analysis of the various possibilities in binary choice situations with unobserved group effects. When more than two choices are available, there may be additional possibilities for identifying choice-specific effects by working with subsets of choices (Brock and Durlauf, 2005). Graham (2006), discussed further below, offers a promising approach for continuous outcomes when individuals are randomly assigned to groups, but is not focusing on the distinction between exogenous and endogenous interactions. Yet more possibilities appear when panel data (that is, repeated observations over time on the same decision-making units) are available. If contextual effects take time to make their impact on the endogenous social effect, the linear dependence is broken and the lack of identification – Manski's reflection problem – is mitigated.

Sometimes, and depending upon the nature of the data as well, it may be impossible, especially in linear models, to identify social interactions in the presence of unobserved group effects. Moving from linear models to binary and other non-linear choice models improves identification even with cross-section data (Brock and Durlauf, 2007).

If it is plausible to exclude some neighbourhood averages of individual covariates, then identification may be possible. Also, if nonlinearities are inherent in the basic model specification, identification again may be possible, even in the case where the contextual effects coincide with neighbourhood averages of individual characteristics. Nonlinearities may eliminate the reflection problem. A noteworthy case in point here is Drewianka (2003), who studies two-sided matching in the marriage market and finds that it allows identification of endogenous *and* exogenous social interactions. The logic of the model requires that the two sides of the market contain an additional source of variation: the greater the number of potential marriage partners, the higher is the probability that a match will occur. There is an inherent multiplier effect at work here. One's prospects of finding a marriage partner depends on the rate at which other people match up, an endogenous social effect. Drewianka's results show that a ten per cent increase in the fraction of the population that is unmarried causes the marriage rate of never-married men to fall by ten per cent and that of never-married women by seven per cent.

An interesting consequence of endogenous social interactions is the amplification of differences in average neighbourhood behaviour across neighbourhoods. In fact, Glaeser, Sacerdote and Scheinkman (2003) use directly such patterns in the data to estimate a *social multiplier*. This is defined for a change in a particular fundamental determinant of an outcome as the ratio of a total effect, which includes a direct effect on an individual outcome plus the sum total of the indirect effects through the feedback from the effects on others in the social group, to the direct effect. It is easy to see it as the ratio of the 'group level' coefficient, the coefficient of $Y_{n(i)}$ in eq. (2), to the 'individual level' coefficient, the coefficient of $Y_{n(i)}$ in eq. (1): $\frac{d}{1-l}\frac{1}{d} = \frac{1}{1-l}$. It follows that a social multiplier greater than 1 implies endogenous social interactions, $0 < l < 1$. This approach must deal, in practice, with dependence across

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
decisions of individuals belonging to the same group, which is implied by non-random sorting in terms of unobservables. It is particularly useful in delivering ranges of estimates for the endogenous social effect and when individual data are hard to obtain.

This is the case with crime data. Glaeser, Sacerdote and Scheinkman (1996) motivate their study of crime and social interactions by the extraordinary variation in incidence of crime across US metropolitan areas over and above differences in fundamentals. If social interactions are present, variations in observed outcomes are larger than would be expected from variations in underlying fundamentals. Glaeser, Sacerdote and Scheinkman (2003) regress actual crime rates against predicted crime rates, which are formed by multiplying percentages of US individuals in each of eight age categories by the crime rate of persons in that category. They perform such regressions at the level of county and state cross-sectionally and for the entire United States over time. Their results imply large social multipliers, which increase with the level of aggregation exactly as their basic theory would predict.

It is possible to modify this basic model in order to study several other areas involving economic decisions akin to social interactions. For example, diffusion of innovations, herding and adoption of norms or other institutions by a population involve ideas that are conceptually related to social interactions. Transmission of job-related information is of particular relevance (see Ioannides and Loury, 2004). Also, *J*, the endogenous social effect, may be negative, as in the case of land development, which is conceivably due to congestion.

Identification of social interactions using observational data on ‘natural experiments’

Several researchers have sought to identify social interactions by exploiting uniquely suitable features of observational data, which are often referred to as ‘natural experiments’. For example, consider outcomes for children from families with several children who share the common influence of unobservable family factors, such as parental values and competence, taste for education and time spent with children, and other unobservables that affect the upbringing of household members living in close proximity. They also share the variation in neighbourhood effects that is produced by families’ residential moves. By using observations on several children from the same family who are separated in age by at least three years, Aaronson (1998) controls for family-specific characteristics. This obviates the need to control for the impact of self-selection in terms of unobservable neighbourhood characteristics. Aaronson uses data from the Panel Study of Income Dynamics and finds large and statistically significant contextual neighbourhood effects, but his models exclude endogenous social effects. His results are robust to changes in estimation techniques and in sample and variable definitions, but are sensitive to the formulation of neighbourhood characteristic proxy. Incomplete specification of family characteristics is an important concern, and its consequences for the robustness of estimated relationships are aptly demonstrated by Ginther, Haveman and Wolfe (2000).

Grinblatt, Keloharju and Ikaheimo (2004) use data for *all* residents of two large Finnish provinces – amounting to millions of observations – and establish that automobile purchase decisions by close residential neighbours influence one another. The measured endogenous neighbourhood effects are strongest among individuals belonging to the same ‘social class’ (especially if they belong to lower-income classes), or when the cars they purchase are of the same make or even the same model. These findings militate in favour of information sharing rather than ‘keeping up with the Joneses’. We note that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
excluding neighbourhood averages of demographics as contextual effects is reasonably plausible in this case: there is no reason why the average *age* of my neighbours should affect directly my taste in *cars*.
Conceptually related is the study of Aizer and Currie (2004), who use data from more than 3.5 million birth certificates from California to examine 'information sharing effects' in the utilization of publicly funded prenatal care. They conclude that it is not information sharing, but is, instead, differences in the behaviour of institutions that explain the established correlations between neighbourhood and ethnic group membership in prenatal care use.
Luttmer (2005) uses data from the US National Survey of Families and Households, augmented with census data from the Public Use Microdata Areas, and examines how self-reported well-being varies with own and neighbours' incomes and with other characteristics. He interprets his findings as direct evidence that people have preferences regarding their neighbours' incomes. That is, after an individual's own income is controlled for, higher earnings of neighbours are associated with lower levels of self-reported happiness on a variety of measures.
Sacerdote (2001) exploits the fact that at Dartmouth College freshman-year room-mates and dorm-mates are randomly assigned, thus producing a natural quasi-experimental setting for studying peer effects. Sacerdote posits that an individual's grade point average is a function of an individual's own academic ability prior to college entrance, of social habits, and of the academic ability and grade point average of his room-mates. Sacerdote finds that peers have an impact on each others' grade point average and on decisions to join social groups such as fraternities. He does not, however, find residential peer effects in other major college decisions, such as choice of college major. He finds peer effects in grade point average at the individual room level – you keep up with your room-mates! – whereas peer effects in fraternity membership occur at both the room level and the entire dorm level – dorms are conformist! These data provide strong evidence for the existence of peer effects in student outcomes, even among highly selected college students who may be otherwise quite homogeneous albeit in close proximity to one another. Peer effects are smaller the more directly a decision is related to labour market activities.

Peer effects in classrooms and schools

Social interactions in classrooms – peer effects – are particularly interesting in understanding schooling as an economic activity and its consequences for inequality of social outcomes. Whether students benefit from classmates with different characteristics and academic performance and whether the effect is different depending upon whether one's classroom peers are more or less able are important for education policy and the actual functioning of schools. In other words, deciding whether or not students should be 'tracked' – that is, administratively segregated in terms of different characteristics – is the sort of policy question which rests on understanding peer effects quantitatively.
Hoxby (2001) posits a relationship between individual academic achievement by a male student in a particular school and grade as the sum of what the mean achievement among males would have been in the absence of peer effects, of a term that is proportional to the percentage of females in the classroom, plus an error. She extends such a relationship to the case of several racial groups, which is particularly appropriate for the Texas Schools Project data that she uses. Her identification strategy involves exploring the panel structure of the data under the plausible assumption that there is natural idiosyncratic variation across successive cohorts in terms of gender, race and other individual attributes. Hoxby finds

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that students are affected by the achievement levels of their peers: an exogenous one-point increase in peers' reading scores raises a student's own score between 0.14 and 0.4 points. Peer effects are stronger intra-race, and there is evidence of contextual effects: both male and female students perform better in classrooms that are more female despite the fact that females' math performance is about the same as that of males.

The role of gender is corroborated by research by Arcidiacono and Nicholson (2005), who use data on the universe of students admitted to US medical schools for a particular year. One positive peer effect in US medical schools that they find pertains to female students, who benefit from attending medical schools that have other female students with relatively high scores on the verbal reasoning section of the Medical College Admission Test.

Of particular interest have been studies of the impact of school racial integration in the US on student performance. Consider Boston's Metropolitan Council for Educational Opportunities (METCO) programme, a voluntary desegregation programme. The programme allows mainly black inner-city kids from Boston public schools to commute to mainly white suburban communities in the Boston area that accommodate them in their public schools. Angrist and Lang (2004) show that, although the receiving districts, which tend to have higher mean academic performance, experience a mean decrease due to the programme, the effects are merely 'compositional', and there is little evidence of statistically significant effects of METCO students on their non-METCO classmates. Analysis with micro data from a particular receiving district (Brookline, Massachusetts) generally confirms this finding, but also produces some evidence of negative effects on *minority* students in the receiving district. METCO is a noteworthy social experiment, which was initiated by civil rights activists seeking to bring about de facto desegregation of schools. Lack of evidence of negative peer effects is particularly useful for informing desegregation policy. Still, there is self-selection in the participants on both sides.

Estimation of social interactions in experimental settings

Experimental data used by social interactions studies come from two types of deliberate experiments: field and laboratory experiments. A well-known field experiment is Project STAR, an experimental programme in the US state of Tennessee that randomly assigned entering kindergarten students into three different class sizes and then randomly assigned teachers to them. A recent study that utilizes Project STAR data is Graham (2006). He seeks to estimate a relationship like (1) by measuring 'excess' variance patterns across groups of exogenously given, but varying, sizes of classrooms that are associated with randomly assigned students and teachers in the presence of correlated effects in the form of unobservable group effects. Graham compares excess variance across small and large classrooms and finds social multipliers between 1.07 and 2.31, and 1.05 to 3.07, for math and reading achievement, respectively. Studies of this type aim at distinguishing excess between-classroom variance that is due to social interactions from that due to group-level heterogeneity.

Duflo and Saez (2003), using experimental data, study how social interactions among employees of a large US university may influence participation in a tax-deferred retirement account plan. The experiment more than tripled the attendance rate of those who received a small monetary reward for participating, doubled that of those not thus 'treated' but who belonged to the same departments as the treated, and significantly increased participation in the target programme by individuals from treated

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
departments, and did so by almost as much as those who did not receive direct encouragement. While clearly the effect of social interactions may coexist with differential treatment and motivational reward effects, social interactions are also relevant for the effects of treatment on attendance and of attendance on participation. The authors conclude that the role of social interactions in amplifying the effect of treatment is unambiguous, in spite of the fact that they cannot distinguish unambiguously between the three different effects.

Moving to Opportunity (MTO) is a set of large randomized field experiments that were conducted by the US Department of Housing and Urban Development in several large US cities. The experiments offered poor households (who were chosen by lottery from among residents of high-poverty public housing projects) housing vouchers and logistical assistance through non-governmental organizations for the purpose of relocating to precisely defined ‘better’ neighborhoods. Several studies based on data from these experiments show that outcomes after relocation improved for children, primarily for females, in terms of education, risky behaviour and physical health, but the effects on male youth were adverse. Regarding outcomes for adults, such as economic self-sufficiency or physical health, the picture is more mixed. Kling, Ludwig and Katz (2005) find that four to seven years after relocation families (primarily female-headed ones with children) lived in safer neighbourhoods that had lower poverty rates than those of a control group that were not offered vouchers. Unfortunately, there is serious controversy over how to interpret these findings in the context of policy design for large-scale policy interventions (Sobel, 2006).

As for laboratory experiments, a notable study is by Ichino and Falk (2006). The experiment involves workers in pairs stuffing envelopes, with control being provided by subjects working alone in a room. These authors find that standard deviations of output are significantly smaller within pairs than between pairs, that is individuals keep up with their neighbours. They also find that social interactions raise productivity: average output per person is greater when subjects work in pairs. They also show that social interactions are asymmetric: low-productivity workers are more sensitive to the behaviour of high-productivity workers. Their setting does reduce some of the noise associated with ‘natural’ experiments but does not allow for contextual effects.

Identification of social interactions with self-selection to groups and sorting

The presence of non-random sorting on unobservables is a major challenge for the econometric identification of social interactions models. The critical role of local public finance in education in the United States has been studied extensively as a link between sorting into residential communities and socio-economic outcomes. Brock and Durlauf (2001) turned adversity into advantage by recognizing that self-selection itself, that is that individuals choose their neighbourhoods making $n(i)$ in eq. (1) endogenous, may provide additional evidence on identification. That is, if it is possible to estimate a neighbourhood selection rule, then correction for selection bias via the mean estimated bias, the so-called Heckman correction term, introduces an additional regressor in the right-hand side of (1) whose neighbourhood average is not a causal effect. Ioannides and Zabel (2002) implement this method successfully using micro data for a sample of households and their ten closest residential neighbours from the American Housing Survey and contextual information for the census tracts in which these individuals reside.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Endogeneity of the average of one's neighbours housing demands, an endogenous social effect, is instrumented by treating housing demands by a group of close neighbours as a simultaneous system of equations. By choosing neighbourhoods, census tracts in this application, individuals choose desirable social interactions. Ioannides and Zabel work with an otherwise standard housing demand model and find a very significant and large endogenous social effect along with very significant contextual effects in the form of unobservable group effects. Several other studies have sought to use instrumental variables to account for self-selection. Still, the identification of valid instruments is often quite hard and requires deep understanding of the actual setting.

Conclusions

Social interactions are ubiquitous. Interest in estimating their effects is expanding rapidly in numerous areas of economics and is motivating important methodological advances. For econometricians, key challenges include social interactions effects on market outcomes coexisting with feedbacks from the characteristics of individual market participants via their impacts on prices, consequences of self-selection and the attendant role of the presence of individual and group unobservables. Fundamentally, and in the light of ever-improving data availability, social interactions empirics will rely increasingly critically on careful theorizing that involves precise definitions of social interactions and justifies stochastic specification, possibly by calling on psychology and sociology to define appropriate boundaries, and must facilitate use of data from different sources. The likely payoff is enormous: better understanding of social forces in the modern economy, with individuals sharing information while self-selecting into social groups and living and working in close proximity to one another, as in firms and cities, the hallmark of modern economic life, and informed design of policy interventions.

See Also

- educational finance
- natural experiments and quasi-natural experiments
- neighbours and neighbourhoods
- psychology of social networks
- social interactions (theory)
- social multipliers
- Tiebout hypothesis

Bibliography

Aaronson, D. 1998. Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *Journal of Human Resources* 23, 915–46.

Aizer, A. and Currie, J. 2004. Networks or neighborhoods? Interpreting correlations in the use of publicly funded maternity care in California. *Journal of Public Economics* 88, 2573–85.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Angrist, J. and Lang, K. 2004. Does school integration generate peer effects? Evidence from Boston's METCO Program. *American Economic Review* 94, 1613–34.

Arcidiacono, P. and Nicholson, S. 2005. Peer effects in medical school. *Journal of Public Economics* 89, 327–50.

Brock, W. and Durlauf, S. 2001. Interaction-based models. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Brock, W. and Durlauf, S. 2005. A multinomial choice model with social interactions. In *The Economy as an Evolving Complex System III*, ed. L. Blume and S. Durlauf. Oxford and New York: Oxford University Press.

Brock, W. and Durlauf, S. 2007. Identification of binary choice models with social interactions. *Journal of Econometrics* 140, 52–75.

Datcher, L. 1982. Effects of community and family background on achievement. *Review of Economics and Statistics* 64, 32–41.

Drewianka, S. 2003. Estimating social effects in matching markets: externalities in spousal search. *Review of Economics and Statistics* 85, 408–23.

Duflo, E. and Saez, E. 2003. The role of information and social interactions in retirement plan decisions: evidence from a randomized experiment. *Quarterly Journal of Economics*, 118, 815–42.

Durlauf, S. 2004. Neighbourhood effects. In *Handbook of Urban and Regional Economics, Volume 4: Cities and Geography*, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Ginther, D., Haveman, R. and Wolfe, B. 2000. Neighborhood attributes as determinants of children's outcomes: how robust are the relationships? *Journal of Human Resources* 35, 603–42.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 1996. Crime and social interactions. *Quarterly Journal of the Economics* 112, 508–48.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 2003. The social multiplier. *Journal of the European Economic Association* 1, 345–53.

Graham, B. 2006. Identifying social interactions through conditional variance restrictions. Working paper, Department of Economics, University of California, Berkeley.

Grinblatt, M., Keloharju, M. and Ikaheimo, S. 2004. Interpersonal effects in consumption: evidence

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
from the automobile purchases of neighbors. Working Paper No. 10226. Cambridge, MA: NBER.

Hoxby, C. 2001. Peer effects in the classroom: learning from gender and race variation. Working Paper No. 7867. Cambridge, MA: NBER.

Ichino, A. and Falk, A. 2006. Clean evidence on peer effects. *Journal of Labor Economics* 24, 39–57.

Ioannides, Y. and Loury, L. 2004. Job information networks, neighborhood effects, and inequality. *Journal of Economic Literature* 42, 1056–93.

Ioannides, Y. and Zabel, J. 2002. Interactions, neighborhood selection, and housing demand. Working paper, Department of Economics, Tufts University.

Kling, J., Ludwig, J. and Katz, L. 2005. Neighborhood effects on crime for female and male youth: evidence from a randomized housing experiment. *Quarterly Journal of the Economics* 120, 87–130.

Luttmer, E. 2005. Neighbors as negatives: relative earnings and well-being. *Quarterly Journal of Economics* 120, 963–1002.

Manski, C. 1993. Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60, 531–42.

Moffitt, R. 2001. Policy interventions, low-level equilibria and social interactions. In *Social Dynamics*, ed. S. Durlauf and H. Peyton Young. Cambridge, MA: MIT Press.

Sacerdote, B. 2001. Peer Effects with random assignment: results for Dartmouth roommates. *Quarterly Journal of Economics* 116, 681–704.

Sobel, Ml. 2006. Spatial concentration and social stratification: does the clustering of disadvantage ‘beget’ bad outcomes?’ In *Poverty Traps*, ed. S. Bowles, S. Durlauf and K. Hoff. Princeton, NJ: Princeton University Press.

How to cite this article

Ioannides, Yannis M. "social interactions (empirics)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000256> doi:10.1057/9780230226203.1557

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social interactions (theory)

José A. Scheinkman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social interactions refer to particular forms of externalities, in which the actions of a reference group affect an individual's preferences. In the presence of strategic complementarities, social interactions help reconcile the observation of large differences in outcomes in the absence of commensurate differences in fundamentals. This article surveys the theoretical literature and discusses different approaches to estimating social interactions.

Keywords

conspicuous consumption; critical mass model; discrete choice models; interactive particle system; multiple equilibria; neighbours and neighbourhoods; network formation; non-market interactions; peer groups; Public Use Microsample Area (PUMA); random field model; residential segregation; Schelling, T.; social interactions; social learning; social multiplier; spatial clustering; statistical mechanics; strategic complementarities; tipping models; urban agglomeration; Veblen, T.

Article

Social interactions refer to particular forms of externalities, in which the actions of a reference group affect an individual's preferences. The reference group depends on the context and is typically an individual's family, neighbours, friends or peers. Social interactions are sometimes called non-market interactions to emphasize the fact that these interactions are not regulated by the price mechanism. Veblen's (1934) analysis of conspicuous consumption – that is, consumption that signals wealth – is perhaps the first contribution to the economic literature on social interactions. Duesenberry (1949) and Leibenstein (1950) are also among the earliest contributors. Although Veblen's *Theory of the Leisure Class* has had a remarkable impact in the social sciences, Schelling's (1971; 1972) pioneering formal analysis of the influence of social groups in behaviour was particularly important for later developments in economics.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Models of social interaction seem particularly adapted to solving a pervasive problem in the social sciences, namely, the observation of large differences in outcomes in the absence of commensurate differences in fundamentals. Many models of social interactions exhibit *strategic complementarities*, which occur when the marginal utility to one person of undertaking an action increases with the average amount of the action undertaken by his peers. Consequently, a change in fundamentals has a direct effect on behaviour and an indirect effect of the same sign. Each person's actions change, not only because of the direct change in fundamentals, but also because of the change in the behaviour of his or her peers. The result of all these indirect effects is the *social multiplier*.

When this social multiplier is large, we expect to see a large variation of aggregate endogenous variables relative to the variability of fundamentals, which seem to characterize phenomena as diverse as stock market crashes, religious differences, and differences in crime rates. In fact, if social interactions are large enough, multiple equilibria can occur – that is, one may observe different outcomes from exactly the same fundamentals. The existence of multiple equilibria also helps us to understand high levels of variance of aggregates.

Social interaction models have implications for the sorting of people and activities across space. As Schelling (1971) demonstrated, when individuals can choose locations, the presence of these interactions may result in segregation across space, even in situations where the typical individual would be content to live in an integrated neighbourhood, provided his group does not form too small a minority. Cities exist because of agglomeration economies which are likely to come from non-market complementarities. In dynamic settings, social interactions can produce *s*-shaped curves which help to explain the observed time series patterns of phenomena as disparate as telephone adoption and women in the workplace. Closely related topics include social learning, where agents learn from observing choices by other agents (for example, Arthur, 1989; Bickhchandani, Hirshleifer and Welch, 1992), and local interaction games (for example, Ellison, 1993; Morris. 2000).

Schelling's critical mass model

In Chapter 3 of his *Micromotives and Macrobehavior* (1978), Schelling discusses a *critical mass model* where he supposes that there is an activity which some individuals will always undertake, others will undertake only if a large enough fraction of the population is engaged in the action, and still others may never undertake. Formally, agents are parameterized by an $x \in [0, 1]$, and can choose between undertaking an action or not. The gain in utility for an agent of undertaking the action is given by $u(x, t)$, where *t* is the fraction of the population engaging in the action. Schelling assumes that $u(x, t)$ decreases with *x*, that is, agents can be inversely ordered by their gains from undertaking the action. He also assumes that $u(x, t)$ increases with *t*, that is, the gain is larger if a larger fraction of the population is engaged in the action. This assumption is exactly what was later named ‘strategic complementarity’. Each agent *x* takes *t* as given and chooses to undertake the action if and only if $u(x, t) \geq 0$. An equilibrium is a fraction t^* such that $u(t^*, t^*) = 0$. Clearly, for such a t^* every agent $x \leq t^*$ will undertake the action while, if $x > t^*$, agent *x* would refrain. Schelling constructs a numerical example where multiple equilibria arise and noticed that even when uniqueness prevails such models display a ‘multiplier effect’. In his example, the presence of a smaller number of individuals that would undertake the activity unconditionally would have a more than proportionate effect on the equilibrium level of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
activity. Granovetter (1978) proposes a very similar model to analyse riots and other collective actions. He noted that as parameter changes some equilibria may disappear, leading to drastic changes in the equilibrium outcomes.

Versions of the critical mass model set out in Schelling (1978) have later been used to study a myriad of economic questions, often with a more detailed micro-economic foundation to justify strategic complementarities. Examples include income inequality (Loury, 1977; Durlauf, 1996), social customs (Akerlof, 1980), the big push in industrialization (Murphy, Shleifer and Vishny, 1989), crime (Sah, 1991), education (Benabou, 1993), savings and consumption norms (Lindbeck, 1997), the transmission of culture (Bisin and Verdier, 2000), and the timing of desertion by soldiers (de Paula, 2005). A continuous action version of the same model, where an agent's utility depends on the average action of the population, is used by Cooper and John (1988) to model macroeconomic coordination failures. Much of this work has ignored market responses to the presence of social interactions. Among the exceptions are Becker and Murphy (2000), who produce a systematic analysis of the effect of prices on market behaviour when social interactions are present, and Pesendorfer (1995), who examines how a monopolist would exploit the presence of non-market interactions.

Models inspired in statistical physics

Schelling's (1971) paper sets out a model where individuals occupy discrete points on the line or plane and interact locally. However, most of the developments that followed use the simpler critical mass model. Follmer (1974) was the first to use explicitly a random field model (also known as an interactive particle system) imported from statistical physics to model social interactions. In these models one typically postulates an individual's interdependence and analyses the equilibrium behaviour that emerges. Typical questions concern the existence and multiplicity of equilibria that are consistent with the postulated individual behaviour, and the sensitivity of these equilibria to parameters. Follmer models an economy in which the preferences of an individual depend on the preference of his peers, and shows that randomness in individual preferences may affect the aggregate, even as the number of agents grows to infinity – a failure of the law of large numbers. Blume (1993) and Brock (1993) recognize the connection between models of discrete choice with interaction effects and some random field models. Glaeser, Sacerdote and Scheinkman (1996) observe that crime rates across large American cities seem to vary too much to be explained by the usual socio-economic variables. They construct a theoretical model connecting the structure of social interactions among individuals with the variation of aggregate behaviour across space, providing a framework for investigating the importance of social interactions. They set up a simple model of local interactions inspired by the voter model in the literature on interacting particle systems (for example, Ligget, 1985) and show that the model is able to generate the large observed variance across aggregates from small amounts of variability in the fundamentals. A simple, one-sided version of their model works as follows. Individuals occupy discrete points on a circle and choose between two actions $\{0, 1\}$. With probability π, the individual chooses action 1 with probability *p* and action 0 with probability $1 - p$. With probability $(1 - \pi)$ he imitates his predecessor's action. The parameter $(1 - \pi)$ can be thought of as a measure of the intensity of social interactions. In a population of *n* individuals, if we write a^i for the action of agent *i* then

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(a^i - p) \to N\left(0, \frac{2-\pi}{\pi}p(1-p)\right).$$

Although the limit average action is always *p*, the (limit) variance of the normalized average action across groups is a function of π. As π converges to zero, this variance becomes arbitrarily large. Social interactions increase the variance of the crime rate across population groups. In a similar vein, Topa (2001) examines the spatial distribution of unemployment with the aid of contact processes (for example, Ligget, 1985), another class of random field models, and shows that social interactions help explain the variation in unemployment rates among Chicago census tracts.

Brock and Durlauf (2001) develop a model that is very much in the spirit of Schelling's critical mass framework. Individuals choose between two actions, and the payoff an individual experiences when taking an action depends on a baseline utility that is common across individuals, on an idiosyncratic preference parameter, and on the distance between his action and the average expected action in the population. By making specific assumptions on the probability distribution of the idiosyncratic preferences, Brock and Durlauf obtain a joint probability measure over choices that is related to that of the mean-field version of the Curie–Weiss model of statistical mechanics. They then show that the model may have one or three equilibria, depending on the values of some parameters. Multiple equilibria are more likely to appear when the baseline utility of the two actions is not very different or when the desire for conformity is strong. Durlauf (1997) and Ioannides (1997; 2006) consider generalizations of the Brock–Durlauf framework with a richer interaction structure that accommodates local interactions. Horst and Scheinkman (2006), who do not use explicitly the language of random fields, also consider infinite systems with arbitrary interaction structures.

Most of this literature is static, but dynamics models, usually involving myopic agents, have been developed, for example, by Blume (1993), Blume and Durlauf (1999), Brock and Durlauf (2001), and Young (1993; 1998), who is especially interested in the evolution of social norms and customs.

The social multiplier

The *social multiplier* measures the ratio of the effect on the average action caused by a change in a parameter to the effect on the average action that would occur if individual agents ignored the change in actions of their peers. This social multiplier can also be thought of as a ratio $\frac{\Delta_P}{\Delta_I}$ where Δ_I is the average response of an individual action to an exogenous parameter (that affects only that person) and Δ_P is the (per capita) response of the peer group to a change in the same parameter that affects the entire peer group. Unless an equilibrium selection mechanism is present, the social multiplier is well defined only if the equilibrium average action is unique, but models that exhibit large social multipliers can explain large differences in outcomes across populations with small differences in exogenous variables. If agents have idiosyncratic random preferences, this same multiplier amplifies the differences in realizations of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
these preferences across samples. In models with continuous actions but otherwise fairly general interaction structures, Glaeser and Scheinkman (2003) and Horst and Scheinkman (2006) show that *moderate social interactions*, a condition that limits the effect of the actions by peers on the optimal choice of an individual, is sufficient for uniqueness of equilibrium. They also show that if, in addition, strategic complementarities are present then the social multiplier exceeds one. Typically, the forces that lead to multiple equilibria also lead to large social multipliers. For instance, in Brock and Durlauf's (2001) model, in the region where uniqueness prevails, the social multiplier is bigger when the desire to conform is stronger and when the fraction of agents that are close to being indifferent between the two possible actions is larger (see Glaeser and Scheinkman, 2003).

Choice of peer group

In several models (for example, Gabszewicz and Thisse, 1996; Benabou, 1993; Glaeser, Sacerdote and Scheinkman, 1996; Mobius, 2000) the peer group that concerns an agent is formed by geographical neighbours. Mobius (1999) shows that, in a context that generalizes Schelling's (1972) tipping model, the persistence of segregation depends on the particular form of the near-neighbour relationship. Kirman (1983), Kirman, Oddou and Weber (1986), and Ioannides (1990) use random graph theory to treat the peer group relationship as random. This approach is particularly useful in deriving properties of the probable peer groups as a function of the original probability of connections. Another literature deals with individual incentives for the formation of networks (for example, Boorman, 1975; Jackson and Wolinsky, 1996; Bala and Goyal, 2000). (A related problem is the formation of coalition in games; for example, Myerson, 1991.) Benabou (1993) and Glaeser and Scheinkman (2001) use Tiebout's equilibrium approach (see for example Bewley, 1981) to model peer group choice.

Empirical issues

Several statistical problems arise in estimating social interactions effects. It is often difficult for a researcher to identify correctly the peer groups. Another problem is that, ideally, one should distinguish between three effects in understanding group behaviour: correlation of individual characteristics, influences of group characteristics on individuals, and the influence of group behaviour on individual behaviour (Manski, 1993). Although the last two effects could both be merged into a social interactions effect, the correlation across individual error terms could remain a problem.

This problem does not arise in randomized experiments that allocate persons into different groups. Katz, Kling and Liebman (2001) and Ludwig, Hirschfeld and Duncan (2001) use data generated by the Moving to Opportunity experiment to provide evidence for the existence of neighbourhood spillovers on juvenile crime. Sacerdote (2001) exploits variation in peer groups generated by the random assignment of freshman room-mates at Dartmouth and finds evidence of peer effects on academic effort, grade point average, and fraternity membership.

In the absence of randomized experiments, Case and Katz (1991) use peer group background characteristics as instruments for peer group outcomes, which in certain cases yield valid estimates of social interactions. They find some evidence that peer behaviour influences self-reported juvenile crime.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

However, as Manski (1993) stresses in the presence of correlations among unobservables, which is particularly likely to arise with sorting, the instrumental variables estimator may overstate social interactions.

Brock and Durlauf (2001) discuss structural identification for their model. In Brock and Durlauf (2000) they provide estimators for the parameters of the model and account for the endogeneity of peer groups. This structural approach leads to a natural behavioural interpretation of the parameters, but it requires individual-level observations and it may suffer from mis-specification.

A less structural approach, proposed by Glaeser, Sacerdote and Scheinkman (1996), is to use the variances of group average outcomes to identify social interactions. Using this methodology they found evidence that social interactions can help explain the large differences in community crime rates. This approach has been formalized further by Graham (2004). Another possibility is to use the logic of the multiplier. The relationship between exogenous variables and outcomes for individuals is compared with the relationship between exogenous variables and outcomes for groups. The ratio is a measure of the size of social interactions. Glaeser, Sacerdote and Scheinkman (2003) apply this method and find evidence of interactions in social group membership in the Dartmouth room-mates data and in crime, and of human capital spillovers at the state and the Public Use Microsample Area (PUMA) level. Yet another alternative is to identify the presence of interactions based on spatial clustering in the behavioural data. (see Topa, 2001; Conley and Topa, 2002). Finally, results in De Paula (2005) suggest that dynamic models might be easier to identify.

See Also

- ergodicity and nonergodicity in economics
- neighbours and neighbourhoods
- social capital
- social interactions (empirics)
- social multipliers
- statistical mechanics

Research supported by the National Science Foundation through grant SES 0350770. I thank Alberto Bisin, Aureo de Paula, Ed Glaeser, and Yannis Ioannides for comments.

Bibliography

Akerlof, G. 1980. A theory of social customs of which unemployment may be one consequence. *Quarterly Journal of Economics* 94, 749–75.

Arthur, W. 1989. Increasing returns, competing technologies and lock-in by historical small events: the dynamics of allocation under increasing returns to scale. *Economic Journal* 99, 116–31.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bala, V. and Goyal, S. 2000. A non-cooperative model of network formation. *Econometrica* 68, 1181–229.

Becker, G. and Murphy, K. 2001. *Social Markets: Market Behavior in a Social Environment*. Cambridge, MA: Belknap-Harvard University Press.

Benabou, R. 1993. Workings of a city: location, education, and production. *Quarterly Journal of Economics* 107, 619–52.

Bewley, T. 1981. A critique of Tiebout's theory of local public expenditures. *Econometrica* 49, 713–40.

Bickhchandani, S., Hirshleifer, D. and Welch, I. 1992. A theory of fads, fashion, custom, and cultural exchange as information cascades. *Journal of Political Economy* 100, 992–1026.

Bisin, A. and Verdier, T. 2000. Beyond the melting pot: cultural transmission, marriage, and the evolution of ethnic and religious traits. *Quarterly Journal of Economics* 115, 955–88.

Blume, L. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5, 387–424.

Blume, L. and Durlauf, S. 1999. Equilibrium concepts for models with social interactions. Mimeo. Cornell University.

Boorman, S. 1975. A combinatorial optimization model for transmission of job information through contact networks. *Bell Journal of Economics* 6, 216–49.

Brock, W. 1993. Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance. *Estudios Economicos* 81, 3–55.

Brock, W. and Durlauf, S. 2000. Interactions-based models. In *Handbook of Econometrics*, ed. J. Heckman and E. Learnerx Amsterdam: North-Holland.

Brock, W. and Durlauf, S. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68, 235–60.

Case, A. and Katz, L. 1991. The company you keep: the effects of family and neighborhood on disadvantaged families. Working Paper No. 3705. Cambridge, MA: NBER.

Conley, T. and Topa, G. 2002. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics* 17, 303–27.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Cooper, R. and John, A. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103, 441–64.

de Paula, A. 2005. Inference in a synchronization game with social interactions. Mimeo. Department of Economics, Princeton University.

Duesenberry, J. 1949. *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.

Durlauf, S. 1993. Nonergodic economic growth. *Review of Economic Studies* 60, 349–66.

Durlauf, S. 1996. A theory of persistent income inequality. *Journal of Economic Growth* 1, 75–93.

Durlauf, S. 1997. Statistical mechanics approaches to socio-economic behavior. In *The Economy as an Evolving, Complex System II*, ed. W. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison-Wesley.

Ellison, G. 1993. Learning, local interaction and coordination. *Econometrica* 61, 1047–72.

Follmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1, 51–62.

Gabszewicz, J. and Thisse, J.-F. 1986. Spatial competition and the location of firms. In *Location Theory*, ed. R. Arnott; *Fundamentals of Pure and Applied Economics 5*, ed. J. Lesourne and H. Sonnenschein. Amsterdam: Harwood Academic.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111, 507–48.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 2003. The social multiplier. *Journal of the European Economic Association* 1, 345–53.

Glaeser, E. and Scheinkman, J. 2001. Measuring social interactions. In *Social Dynamics*, ed. S. Durlauf and P. Young. Cambridge, MA: MIT Press.

Glaeser, E. and Scheinkman, J. 2003. Non-market interactions. In *Advances in Economics and Econometrics: Theory and Applications*, ed. M. Dewatripont, L. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

Graham, B. 2004. Identifying social interactions through excess variance contrasts. Mimeo. Harvard University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Granovetter, M. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83, 1420–43.

Horst, U. and Scheinkman, J. 2006. Equilibria in systems of social interactions. *Journal of Economic Theory* (forthcoming).

Ioannides, Y. 1990. Trading uncertainty and market structure. *International Economic Review* 31, 619–38.

Ioannides, Y. 1997. The evolution of trading structures. In *The Economy as an Evolving, Complex System II*, ed. W. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison-Wesley.

Ioannides, Y. 2006. Topologies of social interactions. *Economic Theory* 28, 559–84.

Jackson, M. and Wolinsky, A. 1996. A strategic model of economic and social networks. *Journal of Economic Theory* 71, 44–74.

Katz, L., Kling, A. and Liebman, J. 2001. Moving to opportunity in Boston: early results of a randomized mobility experiment. *Quarterly Journal of Economics* 116, 607–54.

Kirman, A. 1983. Communication in markets: a suggested approach. *Economic Letters* 12, 1–5.

Kirman, A., Oddou, C. and Weber, S. 1986. Stochastic communication and coalition formation. *Econometrica* 54, 129–38.

Leibenstein, H. 1950. Bandwagon, snob, and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64, 183–207.

Ligget, T. 1985. *Interacting Particle Systems*. New York: Springer Verlag.

Lindbeck, A. 1997. Incentives and social norms in household behavior. *American Economic Review* 87, 370–77.

Loury, G. 1977. A dynamic theory of racial income differences. In *Women, Minorities and Employment Discrimination*, ed. P. Wallace and A. La Mond. Lexington, MA: Lexington Books.

Ludwig, J., Hirschfeld, P. and Duncan, G. 2001. Urban poverty and juvenile crime: evidence from a randomized housing-mobility experiment. *Quarterly Journal of Economics* 116, 665–79.

Mailath, G., Samuelson, L. and Shaked, A. 2001. Evolution and endogenous interactions. In *The Evolution of Economic Diversity*, ed. U. Pagano and A. Nicita. London: Routledge.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Manski, C. 1993. Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60, 531–42.

Morris, S. 2000. Contagion. *Review of Economic Studies* 67, 57–78.

Mobius, M. 2000. The formation of ghettos as a local interaction phenomenon. Working Paper. Massachusetts Institute of Technology.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Industrialization and the Big Push. *Journal of Political Economy* 97, 1003–26.

Myerson, R. 1991. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Pesendorfer, W. 1995. Design innovation and fashion cycles. *American Economic Review* 85, 771–92.

Sacerdote, B. 2001. Peer effects with random assignment: results for Dartmouth roommates. *Quarterly Journal of Economics* 116, 681–704.

Sah, R. 1991. Social osmosis and crime. *Journal of Political Economy* 99, 1272–95.

Schelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1, 143–86.

Schelling, T. 1972. A process of residential segregation: neighborhood tipping. In *Racial Discrimination in Economic Life*, ed. A. Pascal. Lexington, MA: Lexington Books.

Schelling, T. 1978. *Micromotives and Macrobehavior*. New York: Norton.

Topa, G. 2001. Social interactions, local spillovers and unemployment. *Review of Economic Studies* 68, 261–95.

Veblen, T. 1934. *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: Modern Library.

Young, H. 1993. The evolution of conventions. *Econometrica* 61, 57–84.

Young, H. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Scheinkman, José A. "social interactions (theory)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_T000200> doi:10.1057/9780230226203.1558

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social learning

Kaivan Munshi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social learning describes the process whereby individuals learn about a new and uncertain technology from the decisions and experiences of their neighbours. Because information must flow sequentially from one neighbour to the next, social learning provides a natural explanation for the gradual diffusion of new technology that is commonly observed. It can also explain the wide variation in the response to external interventions across communities as a consequence of the randomness in the information signals that they receive. Social learning has been associated with the adoption of new agricultural technology, the fertility transition, and investments in education in developing countries.

Keywords

agriculture; diffusion of technology; family planning; fertility transition; Green Revolution; learning; religion; returns to schooling; social interactions; social learning; social norms; uncertainty

Article

Since the 1950s, people living in developing countries have gained access to technologies, such as high-yielding agricultural seed varieties and modern medicine, that have the potential to dramatically alter the quality of their lives. Although the adoption of these technologies has increased wealth and lowered mortality in many parts of the world, their uptake has been uneven. Entire communities sometimes stubbornly oppose the use of modern medicine or contraceptives. And while high-yielding crop varieties might have spread widely, it took as long as two decades in some cases for this readily available technology to be adopted.

The traditional explanation for the observed differences in the response to new opportunities, across and within countries, is based on heterogeneity in the population (Griliches, 1957; Mansfield, 1968). An alternative explanation, which has grown in popularity in recent years, is based on the idea that individuals are often uncertain about the returns from a new technology. For example, farmers might not

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
know the (expected) yield that will be obtained from a new and uncertain technology and young mothers might be concerned about the side effects from a new contraceptive. In these circumstances, a neighbour's decision to use a new technology indicates that she must have received a favourable signal about its quality and her subsequent experience with it serves as an additional source of information. Because information must flow sequentially from one neighbour to the next, social learning provides a natural explanation for the gradual diffusion of new technology even in a homogeneous population. Social learning can also explain the wide variation in the response to external interventions across otherwise identical communities, simply as a consequence of the randomness in the information signals that they received.

Early contributions by Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992) gave rise to an enormous theoretical literature on social learning (see Bramoulle and Kranton, 2004, for an excellent summary). This article, however, is concerned with a smaller empirical literature on social learning and economic development that has emerged in recent years.

The adoption of new agricultural technology

Following Munshi (2004), consider a simple model of agricultural investment in which there are two technologies: a new high yielding variety (HYV) and a traditional variety. The yield from the uncertain HYV technology for grower *i* in period *t* is specified as

$$y_{it} = y(Z_i) + \eta_{it} \quad (1)$$

where $y(Z_i)$ is the yield under normal growing conditions and Z_i is a vector of soil characteristics and prices. η_{it} is a mean-zero serially independent disturbance term with variance λ_i^2, measuring deviations from normal growing conditions that cannot be observed by the grower.

When $y(Z_i)$ is uncertain, the optimal acreage allocated to HYV by the risk-averse grower is increasing in his best estimate of the expected yield $\hat{y}_{it}$ and decreasing in the variance of that estimate σ_{it}^2, as well as λ_i^2. Because the grower's expected utility is declining in σ_{it}, he will utilize all the information about $y(Z_i)$ that is available to him to arrive at his best estimate $\hat{y}_{it}$.

At the beginning of each period the grower receives an unbiased information signal about the value of his expected yield. He combines this signal with his prior to compute his best estimate of his expected yield, which in turn determines the acreage that he allocates to HYV. Subsequently, he observes his neighbours' acreage decisions, which reveal the signals that they received, as well as all the yields realized in the village. This social information is used to update the grower's prior for the next period. Under the assumption that the expected yield is constant across growers, the acreage function is additively separable in the expected yield, and that the individual information signals are normally

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

distributed, Munshi derives the grower's acreage A_{it} as a linear function of his lagged acreage A_{it-1}, lagged (mean) acreage in the village $\bar{A}_{t-1}$, and lagged yields $\bar{Y}_{t-1}$:

$$A_{it} = \pi_0 + \pi_1 A_{it-1} + \pi_2 \bar{A}_{t-1} + \pi_3 \bar{Y}_{t-1} + \varepsilon_{it}. \quad (2)$$

When information is pooled efficiently within the village, A_{it-1} contains all the information about the expected yield y that was available at the beginning of period t–1, specifically the entire history of information signals and yield realizations up to that time. Conditional on A_{it-1} and fixed individual characteristics subsumed in ε_{it}, $\bar{A}_{t-1}$ represents the new information that was received by the village in period t–1 through the exogenous signals. $\bar{Y}_{t-1}$, in turn, represents the information that was obtained from the yield realizations in that period.

As Manski (1993) notes, neighbours' past decisions will be correlated with the grower's current decision if any determinant of that decision is correlated across neighbours and over time. The prospects for the identification of social learning improve, however, when we focus on the relationship between the current decision and lagged yield realizations. The information signal received by the grower in period t, u_{it}, determines his acreage decision A_{it} (through his expected yield estimate), and so is subsumed in ε_{it}. Growers are never systematically misinformed, $E(u_{it})=y$, and so $\bar{Y}_{t-1}$ and ε_{it} will be correlated as well. One solution to this problem is to difference $\bar{Y}_{t-2}$ from $\bar{Y}_{t-1}$, leaving us with $\bar{\eta}_{t-1} - \bar{\eta}_{t-2}$ from eq. (1). The acreage response to *yield shocks* in the village would then identify the presence of social learning.

By specifying yield to be the sum of a constant term $y(Z_i)$ and an idiosyncratic shock η_{it}, we implicitly assume that input markets function smoothly and that input and output prices do not change over time. In practice, changes in the yield from period t–2 to t–1 could reflect changes in prices or access to scarce resources (such as credit) that are unobserved by the econometrician but directly determine the grower's period-t acreage decision. A spurious yield (shock) effect could in that case be obtained.

To provide additional support for the presence of social learning, Munshi exploits differences in the diffusion of information across crops. Although we have assumed that the expected yield is the same across growers, it will more generally depend on the grower's characteristics. The grower could condition for differences between his own and his neighbours' observed characteristics when learning from them, but the prospects for social learning decline once we allow for the possibility that some of these characteristics will be unobserved, or imperfectly observed. Take the case where all the neighbours' characteristics are unobserved by the grower. He could rely on his own information signals and yield realizations, ignoring information from his neighbours, to obtain a consistent but inefficient estimate of his expected yield. Alternatively, he could continue to learn from his neighbours' decisions and yields, but some bias will then inevitably appear. The testable prediction is that the grower will choose individual learning if the population is heterogeneous and the yield with the new technology is sufficiently sensitive to unobserved characteristics; otherwise he will prefer to learn from his neighbours.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Munshi tests this prediction with farm-level data on wheat and rice cultivation over a three-year period at the onset of the Indian Green Revolution. Rice-growing areas in India are characterized by wide variation in soil characteristics. The early rice varieties were also sensitive to soil characteristics such as salinity, as well as to managerial inputs, which are difficult to observe. As predicted, HYV acreage responds to lagged yield shocks in the village with wheat but not rice. If the view that rice growers were informationally disadvantaged is correct, then these growers should have compensated for their lack of information by experimenting on their own land. Munshi shows that rice adopters allocate more land to HYV than wheat adopters, despite the fact that their farms are smaller and the likelihood of HYV adoption is significantly higher for wheat.

While Munshi assumes that the yield $y(Z_i)$ is exogenous and uncertain, Foster and Rosenzweig (1995) assume that the grower's objective is to learn his optimal (profit-maximizing) input use Z_i. The point of departure for their work is the target-input model of Jovanovic and Nyarko (1996), but we will see that the signal extraction aspect and, hence, the basic structure of the learning process remains the same across these different models of learning. With a slight change of notation, Foster and Rosenzweig assume that the grower attempts to learn the optimal or target input use on his land θ^*,

$$\tilde{\theta}_{ijt} = \theta^* + u_{ijt} \tag{3}$$

where $\tilde{\theta}_{ijt}$ is the optimal input use on plot i for farmer j in period t and u_{ijt} is an i.i.d. random variable with a (known) variance σ_u^2. Notice the similarity with eq. (1), where the grower's objective was to learn the value of the (expected) yield $y(Z_i)$. Previously he collected information on $y(Z_i)$ from various sources to finally arrive at the optimal acreage. In the current set-up, the grower collects information on θ^* from various sources to arrive at his profit-maximizing input level.

Summing the profit over all plots and taking expectations, the grower's expected profit can be expressed as

$$\pi_{jt} = \left[\eta_h - \frac{\eta_{ha}}{A_j}\frac{H_{jt}}{2} - \sigma^2_{\theta jt} - \sigma_u^2\right]H_{jt} + \eta_a A_j, \tag{4}$$

where η_a is the yield from traditional varieties, η_h is the yield from HYV on the plot most suitable for the HYV technology, and η_{ha} represents the loss from using land less suitable for HYV. A_j is the total number of plots and H_{jt} is the number of plots allocated to HYV. The term in square brackets above

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

represents the yield from HYV, which is declining in two sources of error: σ_u^2 and $\sigma_{\theta jt}^2$. σ_u^2 corresponds to λ^2 in Munshi's model and cannot be avoided. However, $\sigma_{\theta jt}^2$, which is associated with suboptimal input use and corresponds to σ_{it}^2 in Munshi's framework, will go to zero as the grower learns the value of θ^* from his own and his neighbours' experiences.

Munshi's test of social learning is based on the relationship between the grower's current HYV acreage and his neighbours' lagged HYV yields. In contrast, Foster and Rosenzweig derive implications for the relationship between the grower's profit (yield) with HYV and cumulative experience with the new technology corresponding to eq. (4):

$$\pi_{jt} = (\eta_h + \beta_{ot} S_{jt} + \beta_{vt} \bar{S}_{-jt}) H_{jt} + \eta_a A_{jt} + \xi_{jt}$$

where the term in parentheses represents the profit (yield) from HYV, which is increasing in the cumulative experience with the new technology on own land S_{jt} and neighbours' land S_{-jt}. The potential sources of spurious correlation that arise in Munshi's analysis evidently apply here as well; time-varying changes in growing conditions or access to scarce resources would affect current HYV yield as well as S_{jt}, $\bar{S}_{-jt}$. Once again it is possible to appeal to the restrictions from the theory to provide additional support for the presence of social learning. While Munshi shows that the effect of neighbours' past decisions and experiences on the grower's current decision will vary across crops, depending on growing conditions and the technology, Foster and Rosenzweig derive predictions for changes in the pattern of learning over time. Foster and Rosenzweig's learning model generates the predictions that (i) β_{ot}, β_{vt} will be declining over time, and (ii) $\frac{\beta_{ot}}{\beta_{vt}} = \frac{\beta_{ot+1}}{\beta_{vt+1}}$. These predictions are successfully tested, consistent with the presence of social learning.

The fertility transition

In Munshi and Myaux's (2006) simple model of the fertility transition no one regulates fertility prior to the inception of a family planning programme. While this remains a potential equilibrium, a new equilibrium in which a sufficient fraction of the community uses convenient modern contraceptives to regulate fertility could also emerge. The object of interest in this case might not be the performance of the new contraceptive (or the side effects associated with its use) but the nature of the future social norm. Individuals gradually learn about the equilibrium that will ultimately prevail in their community as they interact sequentially with each other over time. These changes in beliefs can be mapped into changes in actions: the probability that the individual chooses modern contraceptives in period t is determined by her decision in period t–1 and the probability that she interacted with a user in that period. With random interactions within the community, this last probability is in turn measured by the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
proportion of users in the community in period $t-1$.
As with the identification of social learning in agriculture, estimation of the individual decision-rule described above is complicated by the fact that lagged contraceptive prevalence in the community could proxy for any unobserved determinant of the contraception decision that is correlated across individuals and over time. However, the model of the fertility transition as a process of changing social norms places additional restrictions on the relationship between the individual's contraception decision and lagged contraceptive prevalence in the local area. In particular, social effects should be restricted to the narrow social group within which norms restricting fertility were traditionally enforced. Contraceptive prevalence outside that social group should have no effect on the individual's contraception decision. Munshi and Myaux test these predictions using data from rural Bangladesh. In their research setting, the traditional norm was characterized by early and universal marriage, followed by immediate and continuous child-bearing. Religious authority provided legitimacy and enforced the rules that sustained this equilibrium. Changes in social norms should then have occurred independently within religious groups (Hindus and Muslims) within the village. As predicted, lagged contraceptive prevalence within the individual's religious group within the village has a strong effect on her contraception decision whereas cross-religion effects are entirely absent, both for Hindus and for Muslims.

Education

A recent paper (Yamauchi, 2007) studies social learning and investment in education with the same three-year farm panel at the onset of the Indian Green Revolution that was used by Foster and Rosenzweig (1995) and Munshi (2004). Schooling levels among the growers in the sample were determined long before the unexpected availability of the new HYV technology and so the returns to schooling can be estimated directly at the level of the village using realized incomes. A positive relationship between schooling enrolment among the children and the returns to schooling in the previous generation is then seen to be indicative of social learning.
The usual problem when testing this prediction is that returns to education in the previous generation could be correlated with returns in the current generation, which directly determine school enrolment but are unobserved by the econometrician. Yamauchi consequently derives testable predictions that provide additional support for the presence of social learning at the level of the village. He shows formally that social learning will be faster when the income variance is lower and when there is *greater* heterogeneity in educational attainment in the village. This last prediction is not inconsistent with Munshi's observation that social learning will be slower in heterogeneous populations where neighbours' characteristics are unobserved or imperfectly observed. Schooling is an easily observed characteristic and Yamauchi's insight is that more variance in this characteristic leads to more precise estimates of the returns to schooling. Matching these predictions, schooling enrolment among the children is increasing in the returns to schooling in the village in the previous generation and, more importantly, is increasing in the interaction of the returns and the variance in educational attainment in the village.

Conclusion

How important is social learning in the development process? Foster and Rosenzweig report results from

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
a simulation exercise based on the estimated parameters from their learning model, which compares profits without learning, with learning from own experience, and with learning from own and neighbours' experience. Profitability from the new HYV was lower than profitability from the traditional variety to begin with, but HYV profits exceed traditional profits after four years of experience without learning. With social learning, this point is reached one year earlier. Similarly, Yamauchi's simulations indicate that an increase in schooling inequality within the village could increase enrolment levels by nearly ten per cent. Access to social information appears to be readily available in many practical applications, particularly since there is little cost to the individual from providing information to his neighbours. Thus, the value of interventions that provide the seed for the subsequent spread of such information could be quite high. Understanding how best to design such interventions would seem to be an important area for future research.

See Also

- agricultural economics
- education in developing countries
- fertility in developing countries
- learning and information aggregation in networks
- religion and economic development
- social norms
- technical change
- technology

Bibliography

Banerjee, A.V. 1992. A simple model of herd behavior. *Quarterly Journal of Economics* 117, 797–817.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* 100, 992–1026.

Bramoulle, J. and Kranton, R. 2004. Public goods in social networks: how networks can shape social learning and innovation. Working paper, University of Maryland.

Foster, A.D. and Rosenzweig, M.R. 1995. Learning by doing and learning from others: human capital and technical change in agriculture. *Journal of Political Economy* 103, 1176–209.

Griliches, Z. 1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25, 501–22.

Jovanovic, B. and Nyarko, Y. 1996. Learning by doing and the choice of technology. *Econometrica* 64, 1299–310.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mansfield, E. 1968. *The Economics of Technological Change*. New York: Norton.

Manski, C.F. 1993. Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60, 531–42.

Munshi, K. 2004. Social learning in a heterogeneous population: technology diffusion in the Indian Green Revolution. *Journal of Development Economics* 73, 185–213.

Munshi, K. and Myaux, J. 2006. Social norms and the fertility transition. *Journal of Development Economics* 60, 1–38.

Yamauchi, F. 2007. Social learning, neighborhood effects, and investment in human capital: evidence from Green-Revolution India. *Journal of Development Economics* 83, 37–62.

How to cite this article

Munshi, Kaivan. "social learning." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000527> doi:10.1057/9780230226203.1559

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social multipliers

Mary A. Burke
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social multipliers can be thought of as indicators of the degree of strategic complementarity among interacting agents. A simple choice model is used to derive a formal expression of the social multiplier. The behaviour of the multiplier is described in relation to various factors, such as the strength of complementarity and topology of social interactions. Social multipliers are shown to help explain behaviours, such as criminal activity and labour participation, which vary greatly across social groups or across time out of proportion to variation in fundamentals. Challenges associated with identification of social multipliers are also discussed.

Keywords

comparative statics; continuous and discrete choice models; identification; multiple equilibria; regression-based estimation; selection bias; social interactions; social multipliers; social norms; strategic complementarity

Article

A *social multiplier* arises when choices are subject to social interactions that lead to strategic complementarity, whereby an increase in the level of an action among others in a relevant social group leads to an increase in the same action at the individual level. While multiplier effects may arise in diverse contexts, including that of inter-firm behaviour (see, for example, Cooper and John, 1988), the specific notion of a social multiplier refers to the effects of choice interdependence among individuals in social contexts such as families, neighbourhoods, schools, professional associations, or friendship networks, within which individual choices are thought to be influenced by the choices of others directly, rather than via the effect of others' choices on market variables. Such influences are also referred to as *endogenous* (social) effects. Under such interactions, a change in fundamentals exerts a direct effect on individual action, with the actions of others held fixed, and an indirect effect in the same direction

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
representing the complementary response to the fundamentals-induced change in others' actions. Around a stable equilibrium, the total effect (direct plus indirect) of a change in fundamentals on the average equilibrium action exceeds (in magnitude) the average partial (that is, direct) effect of the change. The social multiplier can be defined as the ratio of the former effect to the latter effect, and therefore exceeds 1 (around a stable equilibrium) under the assumption of strategic complementarity. (The effects of changes in fundamentals around an unstable equilibrium are discussed below.) Social multipliers have been invoked to explain large differences in behaviour across groups that are not readily explained by variation in group characteristics, as well as improbably large swings over time in social practices.

To illustrate formally, consider a choice model in which each agent in a large but finite group (size *N*) chooses the level of an action, $x_i \geq 0$, according to a reaction function, $x_i = f_i(p, \theta_i, \bar{x})$, where *p* is a common exogenous parameter such as the unit price of the action, θ_i is an individual factor such as inherent taste for the action or income (presumed exogenous), and $\bar{x} \equiv \frac{1}{N}\sum_i x_i$ is the (presumed or expected) average action level in the population. (The individual is included in the population average on the assumption that his influence on average action is negligible; this assumption is relaxed below.) The model is said to involve *global* social interactions in the sense that individual action is complementary to the average action in the relevant population, rather than depending only on the actions of some subset of the population. Assume without loss of generality that $\partial x_i / \partial p \leq 0$ and $\partial x_i / \partial \theta_i \geq 0$; strategic complementarity implies $\partial x_i / \partial \bar{x} > 0$. Equilibrium is a set of actions, x_i^e, for $i = 1, 2, 3, \ldots, N$, such that $x_i^e = f_i(p, \theta_i, \bar{x}^e)$ for each *i*, and $\bar{x}^e = \frac{1}{N}\sum_i x_i^e$.

If we evaluate the effects of a price change around a stable equilibrium, the social multiplier is defined by

$$M \equiv \frac{\frac{d\bar{x}^e}{dp}}{\frac{1}{N}\sum_i \frac{\partial x_i}{\partial p}}.$$

(The existence of equilibrium, stable or otherwise, is not guaranteed under the minimal conditions given here, but we assume existence for simplicity. See Glaeser, Sacerdote and Scheinkman, 2003, for discussion of existence conditions.) It can be shown (Becker and Murphy, 2000) that the multiplier can be expressed as

$$\frac{1}{1 - \frac{1}{N}\sum_i \frac{\partial x_i}{\partial \bar{x}}}.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The last expression shows how the strength of the social interactions, as indicated by the term $\frac{1}{N}\sum_i \frac{\partial x_i}{\partial \bar{x}}$, or γ for short, influences the magnitude of the multiplier. For example, if γ =.6, then the multiplier (*M*) equals 2.5. This means that the equilibrium effect on the average action of an exogenous change in price is two and a half times as great (in absolute magnitude, and in the same direction) as the average partial effect of the price change on individual action. When social interactions are 'moderate' in the sense that γ <1 around the given equilibrium (Glaeser and Scheinkman, 2003), the value of the multiplier exceeds 1 (provided also γ >0) and the model exhibits multiplier effects in the intended sense. The consequences of strong social interactions, that is, $\gamma \geq 1$, are discussed below.

In the same model, variation in the vector of individual factors, $\{\theta_1, \theta_2, \ldots, \theta_N\}$, of the population also generates a social multiplier. With some additional assumptions we can define the social multiplier for a change in the average value of θ_i, denoted $\bar{\theta}$. Assuming that the partial effects, $\partial x_i / \partial \theta_i$, are constant and identical across individuals, the effect on the average equilibrium action of a marginal change in $\bar{\theta}$ does not depend on the composition of the changes in the underlying θ_i values, and so is well defined. The social multiplier is now given by the ratio:

$$\frac{\frac{d\bar{x}^e}{d\bar{\theta}}}{\frac{\partial x_i}{\partial \theta_i}}.$$

Again it can be shown that this ratio equals $\frac{1}{1-\gamma}$, where $\gamma = \frac{1}{N}\sum_i \frac{\partial x_i}{\partial \bar{x}}$, so the magnitude of the multiplier around a given equilibrium does not depend on the underlying independent variable (*p* or $\bar{\theta}$). (If the population is small in the sense that a change in any individual x_i exerts a non-negligible effect on $\bar{x}$, then a change in any single θ_i also sets off multiplier effects. See, for example, Cooper and John, 1988; and Glaeser, Sacerdote and Scheinkman, 2003. This analysis may be seen as applying to small group dynamics such as classrooms.)

A critical implication of social multipliers for policy analysis is that failure to consider multipliers may lead to significant underestimation of the aggregate effects of interventions when these are based on estimates of the effect of the intervention at the individual level. Conversely, elasticities measured at the aggregate level will overestimate the average individual response to a change in a variable affecting that individual alone.

Behaviour of the multiplier

Given the formulation that $M = \frac{1}{1-\gamma}$, we see that the multiplier increases in γ over the open interval [0,1), diverging in the limit as γ approaches 1. However, in the neighbourhood of a stable equilibrium,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
it can be shown that $\gamma < 1$, guaranteeing a finite social multiplier. This latter condition says that, for the average individual, a change in average peer action leads to a less than commensurate change in individual action, *ceteris paribus*.

If social interactions are not moderate (that is, $\gamma \geq 1$) in the neighbourhood of equilibrium, the equilibrium is unstable. In addition, when $\gamma \geq 1$ over some or all the domain of mean peer behaviour, multiple equilibria or non-existence of equilibrium may obtain. Comparative statics around unstable equilibria are possible (provided $\gamma \neq 1$), but in such cases the aggregate equilibrium response to a parameter takes the opposite direction from the partial individual response: for example, $\gamma = 1.5$ yields $M = -2$. Consequently, multiplier effects – defined in the sense that the aggregate equilibrium response reinforces the partial individual response – do not apply. Becker and Murphy (2000) define the social multiplier as the value of γ itself and so do not restrict social multipliers to effects around stable equilibria. However, they too discount the relevance of comparative statics around unstable equilibria, instead invoking instability to help explain phenomena such as fads, which are characterized by explosive growth in popularity followed by precipitous declines in same.

Applications

The lesson of social multipliers is that relatively small differences in fundamentals can lead to large differences in outcomes. This insight has been used to help explain large differences in crime rates across cities (Glaeser, Sacerdote and Scheinkman, 1996), in unemployment rates across census tracts (Topa, 2001), in medical treatment rates across hospital market regions (Bell, 2002; and Burke, Fournier and Prasad, 2006), and in stock market participation across social groups (Hong, Kubik and Stein, 2004), among other patterns not readily explained on the basis of the relevant fundamentals.

Conventional wisdom suggests that peer effects may be an important determinant of educational outcomes, and this is the subject of much active research, but conclusive findings on the significance of social multipliers on academic achievement have not yet emerged, given the difficulties of empirical estimation.

Social multiplier effects may accelerate shifts in social norms over time initiated by technological change. Goldin and Katz (2002) link the birth control pill, via direct effects as well as indirect multiplier effects, to the dramatic increases in women's career investment and age of first marriage in the 1970s. Contraception has also been linked to the large increases in out-of-wedlock births since the 1960s (Akerlof, Yellen and Katz, 1996), where the technology's direct impact eventually served, via social interactions, to erode the associated social stigma against out-of-wedlock births. More recently, it has been argued that social multiplier effects have magnified the impact of technological change on obesity rates in the United States in recent decades and led to a larger value of the social norm for body size (Burke and Heiland, 2007).

Extensions

Analysis of social interactions under discrete choice has been described extensively in Brock and Durlauf (2001) and Blume and Durlauf (2003). Utility is given a random component in these models, so

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that individual behaviour varies probabilistically with exogenous factors, and in general mean behaviour does not vary linearly with mean group characteristics. The magnitude of the social multiplier, as well as the existence and number of equilibria, again depends on the intensity of social influence on individual choice, although the analytic representation of the multiplier is less straightforward than in the linear, continuous choice model (see Glaeser and Scheinkman, 2003). This discrete model framework has been applied extensively in the analysis of qualitative, non-market choices such as dropping out of school and engaging in criminal behaviour, choices that are intuitively viewed as particularly susceptible to social influences. The model has the further advantage of being readily amenable to empirical analysis, as the structural choice equation maps directly into a logistic (or, with modification, probit) specification.

Social multipliers also arise in models of *local* social interactions, in which individual action is influenced only by the actions of others in some 'local' subset of the population as defined within a particular spatial model. For example, individuals could be situated at points on a circle and the influential peer group defined as the neighbours to the immediate left and the immediate right. Under this structure, and on the assumption of symmetric and linearly separable reaction functions under continuous choice, the social multiplier (as defined above) equals $\frac{1}{1-\lambda}$, where λ represents the marginal effect of average action among the (local) peer group on individual action. The magnitude of the multiplier therefore reveals the strength of the social interactions but not necessarily their topology. Models on lines, lattices, tori, and other spaces are also possible, as are numerous other specifications of the reference group. Multipliers typically arise, but the specific results derived here do not hold in general. See, for example, Ioannides (2006).

Mean group characteristics, $\bar{\theta}$, may be a direct source of externalities on individual outcomes. In the presence of such an externality, dubbed an *exogenous* or *contextual* effect (Manski, 1993), variation in some mean group characteristic may induce an effect on the mean outcome that exceeds the effect that would be predicted on the basis of the influence of the characteristic at the individual level. For example, if girls tend to have higher test scores than boys, and if an increase in the classroom proportion of girls reduces disruption and so enables higher test scores for all, the total effect on average test scores of an increase in the proportion of girls will exceed the averaged individual gender effects. (However, if girls have lower test scores on average, the two effects of gender composition will oppose each other.) Therefore, even in the absence of *endogenous* effects, for example even if an increase in mean peer achievement has no effect on individual achievement, exogenous interactions may create the appearance of a multiplier effect. While Manski adopts the stance that social multipliers occur only in the presence of endogenous social interactions, on the grounds that contextual effects are merely additive, Glaeser, Sacerdote and Scheinkman (2003) endorse a broader definition of multipliers that includes the contributions of both exogenous and endogenous social interactions. Although the two types of interactions are analytically distinct, it may be impossible to separate their respective effects empirically. Manski (1993) was the first to note this identification problem. Subsequent progress in distinguishing between exogenous and endogenous interactions in the linear model has been made by Brock and Durlauf (2001) and Cohen-Cole (2006).

The canonical derivation of social multipliers relies on a static decision model. However, recent research shows that results may not extend to dynamic decision contexts. In a life-cycle consumption model with social interactions, Binder and Pesaran (2001) show that the emergence of social multipliers depends on

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the distribution of patience parameters in the economy. Bisin, Horst and Ozgur (2006) show that multiplier effects may disappear under forward-looking behaviour and that the magnitude of social multipliers (in both static and dynamic settings) depends on the amount of information people have about other people's types. These findings indicate that the empirical relevance of social multipliers is likely to depend not just on the salience of social influences but also on the nature of intertemporal decision-making, both of these factors being the subject of significant ongoing debate within economics.

Measurement

The researcher who seeks to measure social multipliers encounters a number of formidable identification problems inherent in the empirical analysis of social interactions. One of the most pervasive challenges concerns the non-random selection of individuals into social groups, resulting in a spurious correlation of actions within groups based on similarities in underlying traits, some of which inevitably go unobserved (an example of *correlated effects*, in Manski's terminology). Under such conditions, naive estimation strategies may produce the appearance of a social multiplier where none exists. Even when all relevant sources of correlated effects can be controlled for, it may be possible to identify only the combined effects of endogenous and exogenous social interactions. The attempt to overcome these problems constitutes an area of active research in econometrics. Since identification conditions depend on the nature of the structural model, qualitative research into the empirical context of interest can serve as an important complement to quantitative analysis.

In a model in which individual action depends linearly on mean peer-group action and linear in all other factors, the definition of the social multiplier as a ratio of marginal effects suggests a regression-based estimation strategy. As explicated in Glaeser, Sacerdote and Scheinkman (2003) and Graham and Hahn (2005), the method involves estimating the respective reduced form equations for mean group behaviour and individual behaviour and taking the ratio of the estimated coefficients on mean group characteristics and individual characteristics, using cross-sectional data on groups defined, for example, by geographic boundaries. Under random assignment into groups, this ratio measures the social multiplier at the given level of aggregation. A multiplier value significantly greater than 1 is consistent with the presence of strategic complementarities deriving from endogenous social interactions. However, if exogenous effects of mean group characteristics cannot be ruled out, the measured multiplier captures both endogenous and exogenous effects due to the collinearity of mean characteristics and mean outcomes. In discrete-choice social interactions models, this collinearity does not hold, however, so endogenous and exogenous effects can be identified separately provided social effects in general can be identified (see Brock and Durlauf, 2001; 2007).

Using variants of this ratio method, Glaeser, Sacerdote and Scheinkman (2003) estimate social multipliers on criminal activity in the United States ranging from 1.72 at the county level to 2.8 at the state level and 8.16 at the national level. As the authors acknowledge, the estimates must be viewed sceptically, given that they do not control for sorting on unobservable demographic factors. Although some models predict that multipliers should increase with the level of aggregation, biases due to unobserved factors may also increase with the level of aggregation.

Graham (2006) employs an alternative measurement of social multipliers, also in the linear context, based on ‘excess variance contrasts’ that controls for some types of group-level heterogeneity, in his

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
example the effects of teacher inputs on student achievement. The method relies on variation in average class size across sets of classrooms, where a Tennessee policy ensured random assignment of students and teachers to classrooms on the basis of size. Graham obtains point estimates of 1.76 and 1.97 for the social multipliers on maths and reading achievement in kindergarten, respectively, and shows that his estimation method will detect (true) multiplier effects in many cases in which standard regression methods would not. However, the method requires random assignment into groups on the basis of size, and does not separate exogenous from endogenous effects.

Although nonlinearities in the structural model may alleviate the problem of isolating endogenous social multipliers from exogenous effects, they do not alleviate the problems caused by non-random group selection. Although selection biases cannot be eliminated with complete confidence, they can be mitigated in a number of ways. In addition to including various fixed effects as data permit, researchers may exploit information about the selection process itself, as in Ioannides and Zabel (2003), or by eliminating selection altogether, as in experimental research (see, for example, Ichino and Falk, 2006) and in research that exploits natural experiments (Hoxby and Weingarth, 2006). In each of these last three examples, the evidence points to significant social multipliers (on housing demand, worker output, and school achievement, respectively). But even in cases of random group assignment results may be questioned: recent evidence based on a simulated interaction environment shows that an estimated social multiplier, when measured among randomly assigned groups, may be biased downward relative to its true value (Arcidiacono et al., 2005).

See Also

- consumption externalities
- interdependent preferences
- social interactions (empirics)
- social interactions (theory)
- social learning
- social norms
- social networks, economic relevance of

Bibliography

Akerlof, G., Yellen, J.L. and Katz, M. 1996. An analysis of out-of-wedlock childbearing in the United States. *Quarterly Journal of Economics* 111, 277–317.

Arcidiacono, P., Foster, G., Goodpaster, N. and Kinsler, J. 2005. Estimating spillovers in the classroom with panel data. Mimeo, Duke University.

Becker, G. and Murphy, K.M. 2000. *Social Economics: Market Behavior in a Social Environment*. Cambridge, MA: Belknap-Harvard University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bell, A.M. 2002. Locally interdependent preferences in a general equilibrium environment. *Journal of Economic Behavior and Organization* 47, 309–33.

Binder, M. and Pesaran, M.H. 2001. Life-cycle consumption under social interactions. *Journal of Economic Dynamics and Control* 25, 35–83.

Bisin, A., Horst, U. and Ozgur, O. 2006. Rational expectations equilibria of economies with local interactions. *Journal of Economic Theory* 127, 74–116.

Blume, L.E. and Durlauf, S. 2003. The interactions-based approach to socioeconomic behavior. In *Social Dynamics*, ed. S. Durlauf and H.P. Young. Cambridge, MA: MIT Press.

Brock, W. and Durlauf, S. 2001. Interactions based models. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Brock, W. and Durlauf, S. 2007. Identification of binary choice models with social interactions. *Journal of Econometrics* 140, 52–75.

Burke, M.A., Fournier, G. and Prasad, K. 2006. The emergence of local norms in networks. *Complexity* 11, 65–83.

Burke, M.A. and Heiland, F. 2007. Social dynamics of obesity. *Economic Inquiry* 45, 571–591.

Cohen-Cole, E. 2006. Multiple groups identification in the linear-in-means mode. *Economics Letters* 92, 157–62.

Cooper, R. and John, A. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economic Research* 103, 441–63.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111, 507–48.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 2003. The social multiplier. *Journal of the European Economic Association* 1, 345–53.

Glaeser, E. and Scheinkman, J. 2003. Non-market interactions. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, ed. M. Dewatripont, L.P. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

Goldin, C. and Katz, L. 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. *Journal of Political Economy* 110, 730–70.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Graham, B. 2006. Identifying social interactions through conditional variance restrictions. Working paper, Department of Economics, University of California, Berkeley.

Graham, B. and Hahn, J. 2005. Identification and estimation of the linear-in-means model of social interactions. *Economics Letters* 88, 1–6.

Hong, H., Kubik, J. and Stein, J. 2004. Social interaction and stock-market participation. *Journal of Finance* 59, 137–63.

Hoxby, C. and Weingarth, G. 2006. Taking race out of the equation: school reassignment and the structure of peer effects. Mimeo, Harvard University.

Ichino, A. and Falk, A. 2006. Clean evidence on peer effects. *Journal of Labor Economics* 24, 39–57.

Ioannides, Y. 2006. Topologies of social interactions. *Economic Theory* 28, 559–84.

Ioannides, Y. and Zabel, J.E. 2003. Neighborhood effects and housing demand. *Journal of Applied Econometrics* 18, 563–84.

Manski, C. 1993. Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60, 531–42.

Topa, G. 2001. Social interactions, local spillovers, and unemployment. *Review of Economic Studies* 68, 261–95.

How to cite this article

Burke, Mary A. "social multipliers." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000474> doi:10.1057/9780230226203.1560

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social networks in labour markets

Antoni Calvó-Armengol and Yannis M. Ioannides
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Research in sociology and economics points to an important role for social networks in labour markets. Social contacts mediate propagation of rich and reliable information among individuals and thus help workers to find jobs, and employers to find employees. Recent theoretical advances show that, for agents connected through networks, employment is positively correlated across time and agents, unemployment exhibits duration dependence, and inequality can persist. Recent empirical findings underscore nonlinearities in social interactions and potentially important effects of self-selection. Socioeconomic characteristics can explain substantial spatial dependence in unemployment.

Keywords

information networks; job search; network formation; social networks in labour markets; Stigler, G.; unemployment

Article

The use of social networks is widespread both in employers' recruiting and in workers' job-seeking. Social contacts help workers to find jobs, and employers to find employees. Indeed, social contacts convey rich and reliable information, which they spread widely and fast throughout the labour market. They thus constitute cost-effective search channels that both enrich the information available to both firms and workers, and enhance its quality.

Formal versus informal information sources

The study of social networks in labour markets highlights the nature of labour market transactions as very different from trading in goods, and reflects the importance of idiosyncrasies. The role of job market search and its dealing with frictions goes at least as far back as Stigler (1962). Everyday

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
experience indicates that access to information is heavily influenced by social structure. Individuals use connections with others, such as friends and social and professional acquaintances, to maintain information networks. Rees (1966) first drew attention to differences among workers in their use of the variety of available informational outlets. In this context, *formal* sources of information include state and private employment agencies, newspaper advertisements, union hiring halls, school and college placement services and, more recently, the internet (Kuhn and Skuterud, 2000). *Informal* sources include referrals from employees and other employers, direct inquiries by job seekers and indirect ones through social connections. A recent literature in economics has developed about the details of social interactions that affect the job search process. This literature is complemented by the more extensive sociological analysis of networks. Several sociological works, including notably Granovetter (1974) and Boorman (1975), have been very influential within the economics literature. This article explores the salience within both theoretical and empirical economics research of a social networks approach in the study of labour markets.

Stylized facts

Several stylized facts about labour market networks have been established by empirical work on job information networks (Ioannides and Loury, 2004). The *first* stylized fact is that there is widespread use of friends, relatives, and other acquaintances in job search, and it has increased over time. The *second* stylized fact about job information networks is that the use of friends and relatives in job search often varies by location and by demographic characteristics. Differences in using informal contacts by age, race and ethnicity show conflicting patterns that suggest that important subtleties associated with the operation of social networks are at work. This is confirmed by international comparative evidence. Pellizzari (2004) explores the empirical evidence for the countries of the European Union as of 2003, using the European Community Household Panel, and compares with the United States, using the National Longitudinal Survey of Youth (NLSY). Pellizzari documents large cross-country and cross-industry variation in the wage differentials between jobs found through formal methods and those found through informal ones. Across countries and industries, premiums and penalties are equally frequent. Such differences may be attributed to different recruitment strategies by firms and to different institutional and social practices which may compound the impact of differences in the industrial compositions of economies. The *third* stylized fact about job information networks is that job search through friends and relatives is generally productive. Both employed and unemployed workers who used friends to search for jobs received more offers per contact and accepted more offers per contact than did workers who used other sources of information about job openings. The *fourth* stylized fact about job information networks is that part of the variation in the productivity of job search through networks by demographic group simply reflects differences in usage. In particular, US data suggest that almost one-fifth of the total difference in the probability of gaining employment between black and white youth resulted from racial differences in the use of social contacts.

Recent theoretical treatments

We crudely distinguish two mechanisms through which social contacts impact on the functioning of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
labour market. First, referrals relay information across the two sides of the labour market, firms and workers. Second, workers' connections disseminate job information within the supply side of the labour market through word-of-mouth communication.

Hires mediated by referrals reduce employer uncertainty about prospective workers' productivity, for a number of reasons (Montgomery, 1991). One is that incumbent workers are likely to refer their trusted acquaintances and help them be better informed about their prospective employers. A second reason is that the long-term nature of the relationship between incumbent employees and their employers provides the latter with superior information on the incumbents' productivity-related traits. It is thus not surprising that evidence shows that referral bonuses bring high returns to firms. Yet excessive reliance on referrals deprives firms and individuals who happen to be outside the social networks of firms' workers of mutually beneficial matches.

Recent findings have improved our understanding of the supply-side effects of social networks (Calvó-Armengol and Jackson, 2004; 2006). In their models, workers rely both on own search effort and on information exchange with their social circles to find jobs. Information passing across acquaintances can display a variety of real-life features; for example, when connections differ in terms of intensities, information recipients can be ranked so as to reflect these relational preferences. Calvó-Armengol and Jackson's models are the first to explain several important stylized facts about labour markets, which are hard to explain altogether without an explicit social network model. We turn to those next.

First, information passed from employed individuals to their unemployed acquaintances makes it more likely that these acquaintances will become employed. This generates a positive correlation between employment and wages of networked individuals *within* and *across* periods. Such positive long-run correlation arises despite the short-run rival nature of job information in the following sense: indirect contacts who are two links away in a network are potential competitors for any job held by any common friend. *Second*, duration dependence and persistence in unemployment, both of which are well documented, can be understood as social effects: the longer an individual is unemployed the more likely it is that her social environment is associated with unfavourable future unemployment prospects. This explanation for duration dependence complements more common ones, such as unobserved heterogeneity. This effect resembles an externality and is also responsible for stickiness in aggregate employment dynamics. The closer the economy is to very high employment (or unemployment), the harder it is to leave that state. For similar reasons, parts of the economy can experience a boom while simultaneously other parts of the economy are experiencing a bust.

These are implications of *exogenous* information networks. With an endogenous network that results from agents' participation decisions, the model's predictions are the following. *Third*, the likelihood of dropping out of the labour force is higher for an individual whose social contacts have poor employment experience, or for an individual with few acquaintances. *Fourth*, small differences in initial conditions of different individuals and of network structure can lead to large differences in drop-out rates. Indeed, when an individual drops out, the prospects worsen for all those who remain, and this generates spillover effects in others' decisions to participate or to drop out. Differences in collective employment histories combine with differences in network structure to produce sustained inequality of wages and drop-out rates that feed on each other. So history matters and is responsible for producing persistent income inequality for reasons that are very different from those due to inequalities in human capital investments. Because spillover effects work in reverse, selective and targeted (rather than separate) interventions in the labour market that provide incentives for individuals not to drop out are likely to have amplified

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
effects.

Current empirical treatments

Empirical research has yet to employ fully formal network concepts. It relies typically on concepts of association because of geographic or cultural proximity. There is evidence of persistent correlations in patterns of unemployment in US cities. Socio-economic characteristics, and in particular ethnic and occupational distance, seem to explain a substantial component of the spatial dependence in unemployment. Topa (2001) and Conley and Topa (2002) argue that social interactions can indeed explain the spatial correlation patterns present in the data. Weinberg, Reagan and Yankow (2004) show that one standard deviation improvement in neighbourhood social characteristics and in job proximity raises individuals' hours worked by six per cent and four per cent on average, respectively. Such social interactions have nonlinear effects. The greatest impact is in the worst neighbourhoods. Being in a disadvantaged neighbourhood is more important than the labour activity of one's neighbours per se. Bayer, Ross and Topa (2005) document that people who live close to each other, defined as being in the same census block – a US census block encompasses 3,500 to 5,000 residents of a contiguous geographical area – also tend to work together, that is, in the same census block. Using data from Dartmouth College (where room-mates are assigned randomly), Marmaros and Sacerdote (2002) find large positive correlations between getting help from fraternity or sorority contacts and obtaining prestigious, high-paying jobs. Still, other research points to self-selection as the likely origin of such effects: Oreopoulos (2003) finds that, when neighbourhoods are not selected, neighbourhood quality plays *little* role in determining a youth's eventual earnings, likelihood of unemployment, and welfare participation, while correlations among outcomes for siblings are much higher.
As richer network data become available, further empirical tests of the implications of labour market networks should be developed, which ultimately may call for more elaborate network modelling tools in labour economics. Such research deserves attention.

See Also

- mathematics of networks
- network formation
- social interactions (empirics)
- social interactions (theory)
- psychology of social networks

Bibliography

Bayer, P., Ross, S. and Topa, G. 2005. Place of work and place of residence: informal hiring networks and labor market outcomes. Working Paper No. 11019. Cambridge, MA: NBER.

Boorman, S. 1975. A combinatorial optimization model for transmission of job information through

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
contact networks. *Bell Journal of Economics and Management Science* 6, 216–49.

Calvó-Armengol, A. and Jackson, M. 2004. The effects of social networks on employment and inequality. *American Economic Review* 94, 426–54.

Calvó-Armengol, A. and Jackson, M. 2006. Networks in labor markets: wages and employment dynamics and inequality. *Journal of Economic Theory* (forthcoming).

Conley, T. and Topa, G. 2002. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics* 17, 303–27.

Granovetter, M. 1974. *Getting a Job: A Study of Contacts and Careers*, 1st edn. Cambridge, MA: Harvard University Press.

Ioannides, Y. and Loury, L. 2004. Job information networks, neighborhood effects, and inequality. *Journal of Economic Literature* 42, 1056–93.

Kuhn, P. and Skuterud, M. 2000. Job search methods: internet versus traditional. *Monthly Labor Review* 123(10), 3–11.

Marmaros, D. and Sacerdote, B. 2002. Peer and social networks in job search. *European Economic Review* 46, 870–9.

Montgomery, J. 1991. Social networks and labor market outcomes: towards an economic analysis. *American Economic Review* 81, 1408–18.

Oreopoulos, P. 2003. The long-run consequences of living in a poor neighbourhood. *Quarterly Journal of Economics* 118, 1533–75.

Pellizzari, M. 2004. Do friends and relatives really help in getting a good job? Discussion Paper No. 623. Centre for Economic Performance, London School of Economics.

Rees, A. 1966. Information networks in labor markets. *American Economic Review, Papers and Proceedings* 56, 559–66.

Stigler, G. 1962. Information in the labor market. *Journal of Political Economy* 70(5), 94–105.

Topa, G. 2001. Social interactions, local spillovers and unemployment. *Review of Economic Studies* 68, 261–95.

Weinberg, B., Reagan, P. and Yankow, J. 2004. Do neighborhoods affect work behavior? Evidence from

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the NLSY79. *Journal of Labor Economics* 24, 891–924.

How to cite this article

Calvó-Armengol, Antoni and Yannis M. Ioannides. "social networks in labour markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000471>
doi:10.1057/9780230226203.1562

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social networks, economic relevance of

James Moody and Martina Morris
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social network analysis has a long history in sociology, but interest in this topic has broadened dramatically since 1995. This article reviews the origins and key concepts from the sociological literature on social networks, and outlines the network approach to economic behaviour (see Zuckerman, 2003 or White, 2002 for a broader vision of ‘economic networks’ not covered here due to space limitations). The central themes we trace relate to ‘social embeddedness’ and aggregation – how occupying different positions in complex networks enables and constrains action, as well as the resulting properties of global networks that arise from these actions.

Keywords

division of labour; Durkheim, E.; Simmel, G.; social balance theory; social networks; trust

Article

Classical foundations of social network theory

Sociology's founding theorists often made use of network metaphors, since the discipline is defined by the study of relations; sociology is the study of positions, rather than persons. Durkheim (1984), for example, focused on the primary role played by the division of labour in the idealized transition from traditional to modern society. In traditional society, labour is divided mostly within rather than between households, so outside the household, economic, political, and social relations are based on similarity (‘mechanical solidarity’). Modern society is characterized by a complex division of labour outside the household, resulting in relations based on complementary differences and dependence (‘organic solidarity’). While Durkheim believed that organic solidarity provided much greater potential for growth and development, he also noted that such systems are more vulnerable to breakdowns in connectivity.

Of the classical social theorists, Georg Simmel made the most explicit use of network foundation in his work. What defines *being social* for Simmel is the super-individual quality of a collective (Simmel, 1908, p. 123). A *group* takes on this unique characteristic when its existence cannot be linked to the loss of particular members. When there are only two people involved, there is only one tie, and the group can be dissolved with the loss of either person; so for Simmel, the minimal social group is three persons, and also the first opportunity for structural variation in the ties – a ‘two-path’ or a completed triangle. In his work on *tertius gaudens* (the ‘third who enjoys’), Simmel explores the unique returns to those who occupy an exclusive bridging role between two others. He argues that power and control are a function of how *others* are connected to each other, rather than an individual attribute:

> The power *tertius gaudens* must expend in order to attain his advantageous position does not have to be great in comparison with the power of each of the two parties, since the quantity of his power is determined exclusively by the strength which each of them has relative to each other. (Simmel, 1908)

Simmel's work is the foundation for much current research on power and positions in networks (Blau, 1964; Cook et al., 1983; Burt, 1992).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Key concepts in modern social network analysis

Modern social network analysis focuses on explaining where networks come from and how networks affect outcomes. There are multiple summary measures for network structure that span multiple units of analysis. For nodes, the most basic measure is ‘degree’, the number of ties each node has to others in the network, which aggregates up to the population degree distribution. For dyads, measures can either characterize properties of the dyads themselves (for example, duration and type of relation) or the properties of the two nodes (for example, ‘homophily’, a form of correlation between nodal attributes). Higher-order configurations include triads and other cycles, the size and distribution of connected sets, path lengths and distributions, and many more (see Wasserman and Faust, 1994). Overall network attributes are typically functions of the lower-order properties, and include measures like density, centralization, core-periphery structure and clustering. At each level, researchers ask either ‘what underlying processes produce this network structure?’ or ‘what impact does this structure have on some outcome?’ The endogeneity of some of the implied processes leads to interesting and complicated dynamics, as well as causal ambiguity.

Where do networks come from?

Networks emerge as the result of relationship formation and dissolution. These processes are influenced by several factors: contextual effects (for example, population composition, institutional mediation), individual propensities (for example, expansiveness and attractiveness), dyadic factors (for example, homophily or mutuality), and explicit social rules about relational configurations (such as incest prohibitions or social balance rules for friendship).

Real-world social networks almost always display more clustering and heterogeneity than similar-volume Erdos–Reyni random networks; where ties are distributed uniform random across all possible pairs. Why? Part of the answer is propinquity: individual behaviour is organized by social contexts that filter the pool of available partners (Feld, 1981). Preferences for similarity combine with attribute heterogeneity to cluster ties within groups (Blau, 1977). Heterogeneities in individual propensities to send or receive ties also contribute to departures from a random graph. More active persons and groups will account for a disproportionate share of relations, creating hubs or clusters of activity. Recent ‘scale free’ network research has this simple ‘preferential attachment’ mechanism at its core (Barabasi and Albert, 1999).

Most social network research examines processes that operate above the individual level, focusing on the generative mechanisms explicitly governed by dyadic, triadic, and higher order relational norms. One of the first dyadic models proposed for directed relations was mutuality, which captures the commonly observed propensity for out-ties to be reciprocated. Another common dyadic model is ‘homophily’, the preference for similarity in social relations (Blau, 1977). Friendship pairs tend to be similar with respect to socioeconomic status, gender, race, and delinquent behaviour (Cohen, 1977; McPherson, Smith-Lovin and Brashears, 2006) and romantic ties are strongly assortative by age, education and race (Mare, 1991; Morris, 1995). The general term given to this dyadic process is selective mixing – which may be assortative, disassortative or idiosyncratic.

The most famous model for triads is social balance theory. At its simplest, balance theory predicts that a friend of a friend will also be a friend, as enmity among one's friends leads to strain and is avoided (Davis, 1963; Holland and Leinhardt, 1971). The beauty of balance models is that they provide a clear (if endogenous) link between individual action and network structure. Consider the example below in Figure 1, which maps the potential change space for a single unbalanced triad. The initial triad is unbalanced, since actor a's friends (+) are enemies (–) with each other. This pattern is expected to create dissonance and thus each actor has a motive to change relations, any one of which would result in one of the three balanced outcomes. Each change, however, may alter the balance of other triads containing these nodes, resulting in a chain reaction of tie formation and dissolution over time.

Figure 1

An example of a balance transition

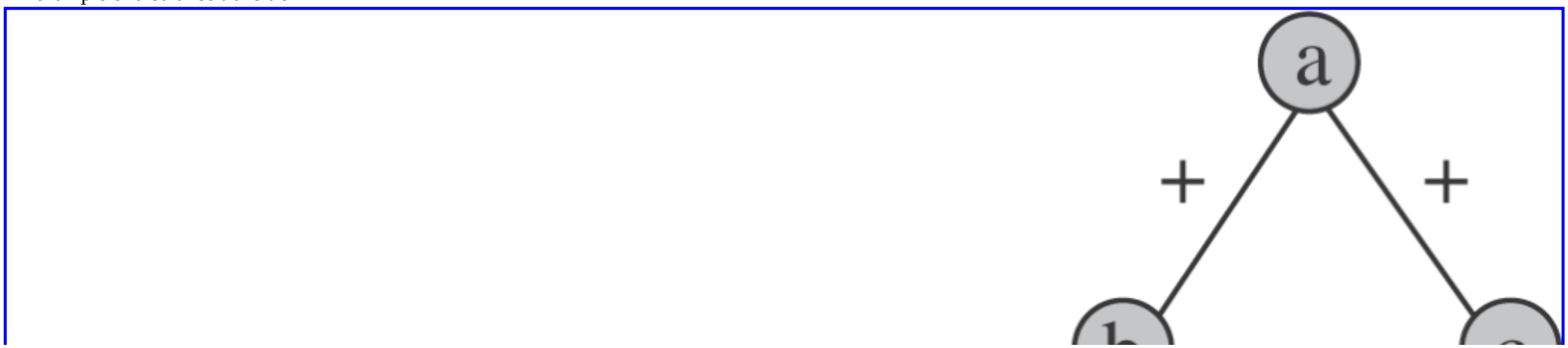

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

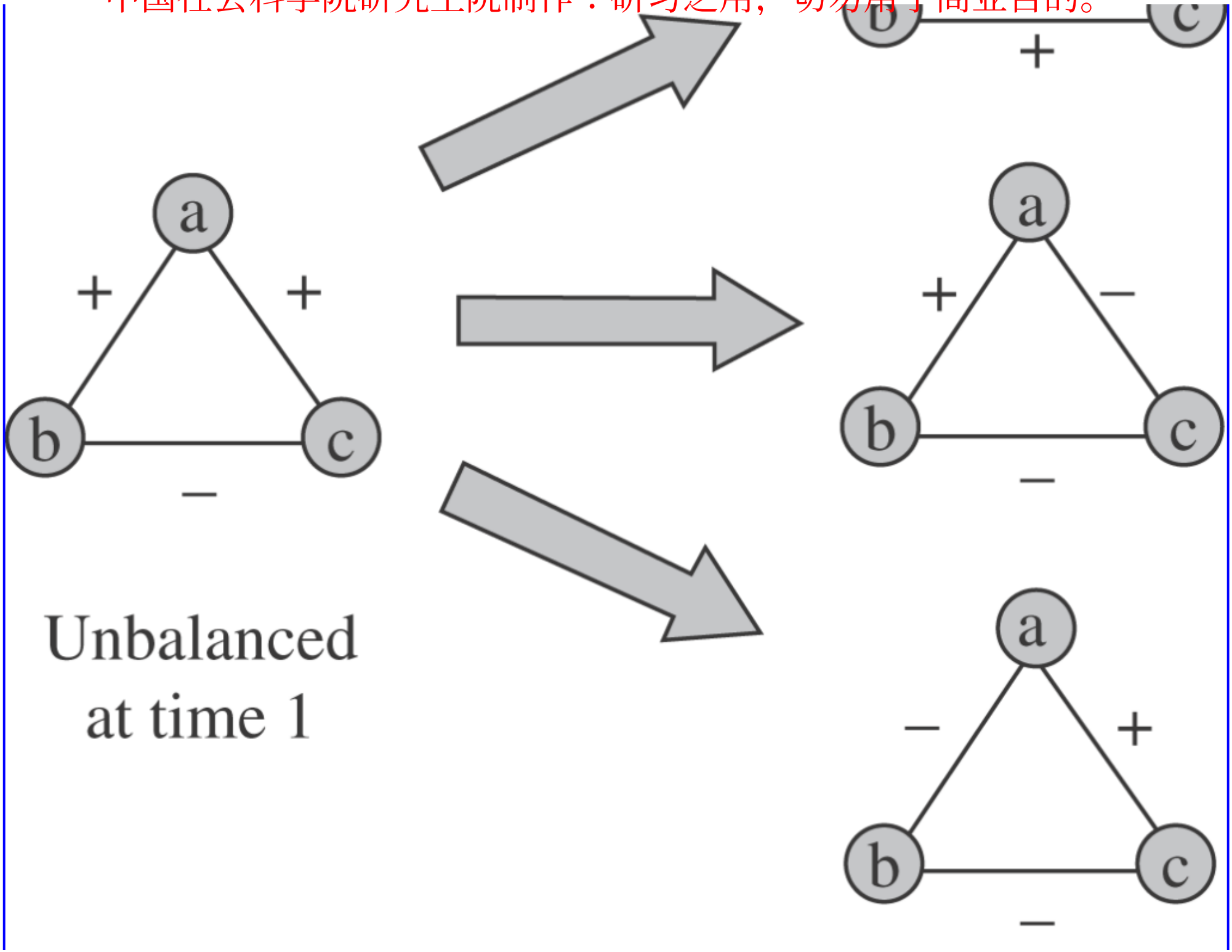

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

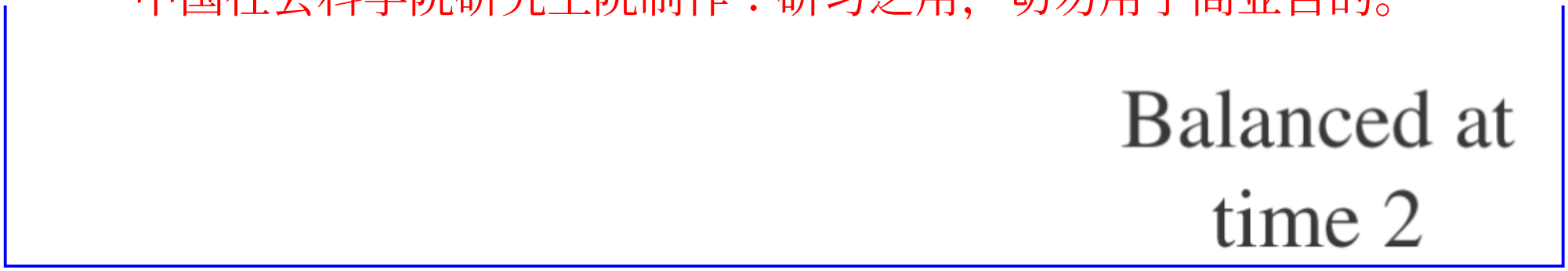

Many triad and higher-order models for generating networks have been explored. Johnsen (1986) generalizes a set of simple balance-like rules to describe complex group-level structures. Moody (1999) demonstrates that a stable macro-structure cannot be assumed from the underlying choice patterns, since triad transitions that create balance from one actor's position will create imbalance for another. Chase (1980) demonstrates that perfect linear hierarchies require transitive triadic dominance relations, and that empirically (at least among chickens) relation formation rules are more important than individual attributes in predicting domination. Bearman, Moody and Stovel (2004) show that a simple rule prohibiting four cycles in heterosexual networks (the smallest possible cycle, equivalent to swapping partners) can generate romantic networks very similar to those observed among adolescents.
Finally, relational rules may also be based on the *timing* of ties. The most well-known example is the norm of serial monogamy in romantic relations. Monogamous networks are completely disconnected at any time point – they are nothing more than collections of isolated dyads. By contrast, if partnerships can be maintained concurrently, larger connected sets can form, and a giant component emerges quickly, even in the absence of nodes with many partners. This example has been well studied because it has implications for the spread of infection (Moody, 2000; Morris and Kretzschmar, 1997; Morris, Goodreau and Moody, 2008).

What do networks do?

The relations that comprise a network define a set of positions. The positions provide individuals with opportunities for access (a person to call for help) and exposure to diffusion (such as a virus through sexual contact or information about a job). The overall structure determines the population dynamics of social exchange and the evolutionary stability of the system.
At the individual level, opportunities for access to resources follow from one's position in the network. The set of direct contacts – called the egocentric or local network – is typically characterized by number, heterogeneity, and strength (Marsden, 1987). Recent work on social capital is in this tradition, building on the strength of ties, questions of trust, and the mix of types of people one is connected to (Lin, 2002). The effect of indirect ties – the partners of one's partners, and beyond – has produced some of the best-known network research, starting with Lee's (1969) *The Search for an Abortionist* and continuing with Granovetter's (1973) classic ‘The Strength of Weak Ties’ on job search. Since the people one is very close to know what one knows, the most profitable information sources come from those connected to other social worlds.
While we may be influenced by our local network partners, their opinions are similarly shaped, so, as with balance theory, the *system* of beliefs (its coherence, polarization and dynamics) co-evolves with network ties. Studies of diffusion over networks have focused on a range of different types of processes, from the spread of information, norms, and innovation, to the spread of infection (Coleman, Katz, and Menzel, 1966; Rogers, 1962; Morris, 2004). Speed, pervasiveness and stability of spread are determined by the variations in the transmission network: density, the length and number of paths, clustering and relationship timing (Pool and Kochen, 1978; Watts and Strogatz, 1998; Moody, 2000; Morris and Kretzschmar, 1997).
The systemic perspective on endogeneous network processes leads naturally to a full game theoretic approach: a set of rules that govern interaction at the micro level which aggregate up to produce complex dynamics at the population level. Topics studied this way include the predominance of matrilateral cross-cousin marriage in classificatory kinship systems (White, 1963), the pervasive spread of unpopular ideas (Centola, Willer and Macy, 2005), and the conditions for the persistence of altruism (Bowles, Choi and Hapfensitz, 2003).

Social networks and economic outcomes

The starting point for most network approaches to economic sociology is Granovetter's (1985) paper on social embeddedness. Granovetter draws a middle ground for economic action between purely autonomous and economically rational actors (‘undersocialized action’) and deeply embedded and largely scripted normative behaviour (‘oversocialized action’). He argues that ‘personal relations and structures’ generate trust, discourage malfeasance and are subject to intentional construction. Here, as always in the network approach, there is both

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

a local and a global dynamic:

> 'Embeddedness' refers to the fact that economic action and outcomes, like all social action and outcomes, are affected by actors dyadic (pairwise) relations *and* by the structure of the overall network of relations. As a shorthand, I will refer to these as the relational and the structural aspects of embeddedness. (Granovetter, 1992, p. 33)

At the dyadic level, the focus has traditionally been on contact strength, with questions turning on issues of trust and exchange in local networks (Uzzi, 1999). At the structural level, the focus is on how larger sets of nodes are mutually reconnected, providing contexts where mutual obligations can be reinforced and behaviour monitored (Moody and White, 2003). The role of networks in shaping economic behaviour is particularly clear in settings where markets are not well developed or where capital resources are low. In China's emerging market economy, for example, firms will often trade with well-trusted associates even if they can find the goods cheaper elsewhere (Keister, 2001). Similar patterns are also found in Western commodity trading markets, however, suggesting there is a benefit to trading with known partners even in well-regulated settings (Baker, 1984; Uzzi, 1996; 1999).
Ron Burt (1992) builds on both Granovetter and Simmel in his research, linking profit opportunities to 'structural holes' – the absence of a tie between two persons who have something to exchange. As in Granovetter, information is more likely to accrue to those who connect such disconnected worlds. As in Simmel, those in a 'middle-man' position can play each of the actors they are connected to off each other, gaining a control advantage in the network. Structural holes are essentially arbitrage opportunities. As such, in a dynamic setting there are incentives for holes to be closed, which can either shorten the window of profitability, or generate action to protect the advantage.

Conclusion

The history of research on social networks in sociology is broad and deep, ranging from the earliest foundation of the discipline to nearly every empirical area under current research. The general theme of this work is to relate individual network positions or network structures either to generative social processes (when explaining network formation) or to substantive outcomes (when using networks to predict behaviour). For economic questions, a network embeddedness approach allows one to identify social bounds to strategic action based on connections with others.
Much of the current research on social networks is computationally intensive. Given the endogeneity inherent in network processes, statistical models cannot assume independence, and dynamic models are not analytically tractable. To analyse network formation, a new class of statistical models is being developed to help distinguish between different mechanisms that lead to similar structural features (Morris, 2003; Handcock et al., 2003; Snijders, 2001; Snijders et al., 2006). The new computational models allow us to examine a wide range of questions without the need for traditional simplifying assumptions. This set of tools promises to put the 'social' back into social science research methods.

See Also

- assortative matching
- economic sociology
- economy as a complex system
- graph theory
- mathematics of networks
- network formation

Bibliography

Baker, W. 1984. The social structure of a national securities market. *American Journal of Sociology* 89, 775–811.

Barabasi, A.-L. and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509–12.

Bearman, P., Moody, J. and Stovel, K. 2004. Chains of affection: the structure of adolescent romantic networks. *American Journal of Sociology* 110, 44–91.

Blau, P.M. 1964. *Exchange and Power in Social Life*. New York: Wiley.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Blau, P.M. 1977. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: The Free Press.

Blau, P.M. and Schwartz, J.E. 1984. *Crosscutting Social Circles: Testing a Macrostructural Theory of Intergroup Relations*. Orlando: Academic Press.

Bowles, S., Choi, J.K. and Hopfensitz, A. 2003. The co-evolution of individual behaviors and social institutions. *Journal of Theoretical Biology* 223, 135–47.

Burt, R.S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Centola, D., Willer, R. and Macy, M. 2005. The emperor's dilemma: a computational model of self-enforcing norms. *American Journal of Sociology* 110, 1009–40.

Chase, I.D. 1980. Social process and hierarchy formation in small groups: a comparative perspective. *American Sociological Review* 45, 905–24.

Cohen, J.M. 1977. Sources of peer group homogeneity. *Sociology of Education* 50, 227–41.

Coleman, J.S., Katz, E. and Menzel, H. 1966. The diffusion of an innovation among physicians. *Sociometry* 20, 253–70.

Cook, K.S., Emerson, R.M., Gillmore, M.R. and Yamagishi, T. 1983. The distribution of power in exchange networks: theory and experimental evidence. *American Journal of Sociology* 89, 275–305.

Davis, J.A. 1963. Structural balance, mechanical solidarity, and interpersonal relations. *American Journal of Sociology* 68, 444–62.

Durkheim, E. 1984. *The Division of Labor in Society*, tr. W.D. Halls. New York: The Free Press.

Feld, S.L. 1981. The focused organization of social ties. *American Journal of Sociology* 86, 1015–35.

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology* 81, 1287–303.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* 91, 481–510.

Granovetter, M. 1992. Problems of explanation in economic sociology. In *Networks and Organizations: Structure, Form and Action*, ed. N. Nohria and R.G. Eccles. Boston, MA: Harvard Business School Press.

Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M. and Morris, M. 2003. *Statnet: An R Package for the Statistical Modeling of Social Networks*. Funding support from NIH grants R01DA012831 and R01HD041877. Available at http://www.csde.washington.edu/statnet, accessed 22 March 2007.

Holland, P. and Leinhardt, J. 1971. Transitivity in structural models of small groups. *Comparative Groups Studies* 2, 107–24.

Johnsen, E.C. 1985. Network macrostructure models for the Davis–Leinhardt set of empirical sociomatrices. *Social Networks* 7, 203–24.

Johnsen, E.C. 1986. Structure and process: agreement models for friendship formation. *Social Networks* 8, 257–306.

Keister, L.A. 2001. Exchange structures in transition: a longitudinal study of lending and trade relations in Chinese business groups. *American Sociological Review* 66, 336–60.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lee, N.H. 1969. *The Search for an Abortionist*. Chicago: University of Chicago Press.

Lin, N. 2002. *Social Capital: A Theory of Social Structure and Action*. Cambridge: Cambridge University Press.

Mare, R.D. 1991. Five decades of educational assortative mating. *American Sociological Review* 56, 15–32.

Marsden, P.V. 1987. Core discussion networks of Americans. *American Sociological Review* 52, 122–31.

McPherson, M., Smith-Lovin, L. and Brashears, M.E. 2006. Social isolation in America: changes in core discussion networks over two decades. *American Sociological Review* 71, 353–75.

Moody, J. 1999. *The Structure of Adolescent Social Relations: Modeling Friendship in Dynamic Social Settings*. Chapel Hill: University of North Carolina.

Moody, J. 2000. The importance of relationship timing for diffusion. *Social Forces* 81, 25–56.

Moody, J. and White, D.R. 2003. Social cohesion and embeddedness: a hierarchical conception of social groups. *American Sociological Review* 68, 103–27.

Morris, M. 1993. Epidemiology and social networks: modeling structured diffusion. *Sociological Methods and Research* 22, 99–126.

Morris, M. 1995. Data driven network models for the spread of infectious disease. In *Epidemic Models: Their Structure and Relation to Data*, ed. D. Mollison. Cambridge: Cambridge University Press.

Morris, M. 2003. Local rules and global properties: modeling the emergence of network structure. In *Dynamic Social Network Modeling and Analysis*, ed. R. Breiger, K. Carley and P. Pattison. Washington, DC: National Academy Press.

Morris, M. 2004. *Network Epidemiology: A Handbook for Survey Design and Data Collection*. International Studies in Demography. Oxford: Oxford University Press.

Morris, M., Goodreau, S. and Moody, J. 2008. Sexual networks, concurrency, and STD/HIV. In *Sexually Transmitted Diseases*, 4th edn. ed. K.K. Holmes, P.F. Sparling, W.E. Stamm, P. Piot, J.N. Wasserheit, L. Corey, M. Cohen and H. Watts. New York: McGraw-Hill.

Morris, M. and Kretzschmar, M. 1997. Concurrent partnerships and the spread of HIV. *AIDS* 11, 641–48.

Myers, D.J. 2000. The diffusion of collective violence: infectiousness, susceptibility, and mass media networks. *American Journal of Sociology* 106, 173–208.

Pool, I.D.S. and Kochen, M. 1978. Contacts and influence. *Social Networks* 1, 5–51.

Rogers, E.M. 1962. *Diffusion of Innovations*. New York: The Free Press.

Simmel, G. 1908. In *The Sociology of Georg Simmel*, ed. K.H. Wolf. New York: The Free Press, 1950.

Snijders, T.A.B. 2001. The statistical evaluation of social network dynamics. *Sociological Methodology* 31, 361–95.

Snijders, T.A.B, Pattison, P.E., Robins, G.L. and Handcock, M.S. 2006. New specifications for exponential random graph models. *Sociological Methodology* 36, 99–153.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Uzzi, B. 1996. The sources and consequences of embeddedness for the economic performance of organizations: the network effect. *American Sociological Review* 61, 674–98.

Uzzi, B. 1999. Embeddedness in the making of financial capital: how social relations and networks benefit firms seeking financing. *American Sociological Review* 64, 481–505.

Valente, T.W. 1995. *Network Models of the Diffusion of Innovations*. Cresskill, NJ: Hampton Press.

Wasserman, S. and Faust, K. 1994. *Social Network Analysis*. Cambridge: Cambridge University Press.

Watts, D.J. and Strogatz, S.H. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–2.

White, H.C. 1963. *Anatomy of Kinship: Mathematical Models for Structures of Cumulated Roles*. Englewood Cliffs, NJ: Prentice-Hall.

White, H.C. 2002. *Markets from Networks: Socioeconomic Models of Production*. Princeton: Princeton University Press.

Zuckerman, E.W. 2003. On networks and markets. *Journal of Economic Literature* 46, 545–65.

How to cite this article

Moody, James and Martina Morris. "social networks, economic relevance of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000511> doi:10.1057/9780230226203.1561

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social norms

H. Peyton Young
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The function of a social norm is to coordinate people's expectations in interactions that possess multiple equilibria. Norms govern a wide range of phenomena, including property rights, contracts, bargains, forms of communication, and concepts of justice. Norms impose uniformity of behaviour within a given social group, but often vary substantially among groups. Over time norm shifts may occur, prompted either by changes in objective circumstances or by subjective changes in perceptions and expectations. The dynamics of this process can be modelled using evolutionary game theory, which predicts that some norms are more stable than others in the long run.

Keywords

asymmetric information; conventions vs norms; coordination games; evolution of norms; expectations; fairness; Hume, D.; local conformity/global diversity effect; multiple equilibria; optimal contracts; punctuated equilibrium effect; reputation; signalling; social capital; social norms; stochastic volatility; stochastically stable norms; transaction costs

Article

Social norms are customary rules of behaviour that coordinate our interactions with others. Once a particular way of doing things becomes established as a rule, it continues in force because we prefer to conform to the rule given the expectation that others are going to conform (Schelling, 1960; Lewis, 1969). This definition covers simple rules that are self-enforcing at a primary level, such as which hand to extend in greeting or which side of the road to drive on, and more complex rules that trigger sanctions against those who deviate from a first-order rule. (We express outrage if someone cuts in front of someone else in a queue.) The former are sometimes called *conventions* and the latter *norms* (Sugden, 1986; Coleman, 1990; Bicchieri, 2006), but in fact there are numerous gradations and levels of response to norm violation that make this dichotomy problematic. Hence I shall use the term ‘norm’ in its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
inclusive sense in what follows.
David Hume (1739) was the first to call attention to the central role that norms play in the construction of social order. Norms define property rights, that is, who is entitled to what. They determine what commodities are accepted as money. They shape our sense of obligation to family and community. They determine the meanings we attach to words. Indeed it is hard to think of a form of interaction that is not governed to some degree by social norms. (For book-length treatments of the subject see Lewis, 1969; Ullman-Margalit, 1977; Sugden, 1986; Young, 1998a; Posner, 2000; Hechter and Opp, 2001; Bicchieri, 2006.)

Norms and equilibria

Norms can be represented as equilibria of suitably defined games; indeed, Hume's analysis of norms can be viewed as one of the earliest examples of game-theoretic reasoning. Nevertheless, not every equilibrium of a game is a norm. First, the term generally applies only to games with *multiple* equilibria. People can queue for service *or* they can push. They can use gold coins *or* glass beads as money. They can drive on the left *or* on the right.
Second, even if a game has multiple equilibria, they do not necessarily qualify as norms. To illustrate the distinction, consider two individuals who get to divide a dollar provided they can agree on how to divide it. Each makes a demand, and if the demands sum to at most one the demands are met; otherwise they get nothing. This is a coordination game and it has many equilibria. For example, if one person demands 43 cents and the other demands 57 cents, the demands are in equilibrium: no one can gain from a unilateral deviation. But this is not a norm; it is an idiosyncratic equilibrium for these two individuals. Fifty–fifty division, by contrast, is a norm because it is usual and customary in games of this kind, and everyone knows it.

Norm enforcement

Broadly speaking there are three different mechanisms by which norms are held in place. Some are sustained by a pure coordination motive. If it is the norm to drive on the left, I adhere to the norm in order to avoid accidents. If gold is the commonly accepted currency, it would be a waste of time to try to conduct my business with glass beads. These are ‘social’ phenomena, because they are held in place by shared expectations about the appropriate solution to a given coordination problem, but there is no need for social enforcement.
Other norms are sustained by the threat of social disapproval or punishment for norm violations (Sugden, 1986; Coleman, 1990). If queuing is the norm, I will be censured if I try to push my way to the front. If duelling is the proper response to an insult, I will lose status in the community if I do not challenge the one who insulted me. If I am expected to avenge the murder of my brother and fail to carry it out, I may be ostracized by other family members. (Exactly why third parties bother to express disapproval or carry out punishments is a matter of debate, but the evidence suggests that they sometimes do so even at considerable personal cost – Fehr, Fischbacher, and Gächter, 2002.)
A third enforcement mechanism arises through the internalization of norms of proper conduct. If it is the norm not to litter, I will avoid littering even in situations where no one can see me. If I eat a meal in a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
foreign city and fail to tip the waiter, I need not fear the consequences because there is no continuing relationship; nevertheless, I may think the worse of myself for having done it. More generally, norms often take on the character of virtuous or right action (Hume, 1739), and departures from a norm can trigger emotions of shame or guilt even when third-party enforcement is absent (Coleman, 1990; Elster, 1989; 1999). This fact is especially useful in large-scale societies, where it may be difficult to monitor others' compliance with equilibrium behaviour.

Norms and efficiency

It remains to be explained why a dictionary on economics, in contrast to sociology or law, should bother with an entry on social norms. What economic purpose do they serve? The answer is that norms coordinate expectations, and thereby reduce transaction costs in interactions that possess multiple equilibria (Wärneryd, 1994).

This point seems clear enough intuitively; it can also be demonstrated experimentally (Roth, 1985). Consider a game in which two players can divide a pile of chips in any way they like, but if they fail to agree on a division within a specified period of time they forfeit all of them. When all the chips can be cashed in for the same amount of money, the norm is to divide the chips equally; moreover, the great majority of players do in fact coordinate in this manner. Notice, however, that *any* division of the chips, not just 50–50, can constitute an equilibrium of the one-shot game if both players expect that it will be played. Now consider a variant in which the players get to cash in the chips for different amounts of money (which is publicly known). In this case there are two potential focal solutions – divide the chips equally and divide the money-value of the chips equally – but there is no norm to steer the players' expectations towards one or the other. As a result, the frequency of disagreement rises substantially.

More generally, a norm has economic value if it creates a uniquely salient or focal solution to a coordination problem, thus reducing the risk of coordination failure. In this sense norms are a form of social capital (Coleman, 1987). This does not mean, however, that norms are invariably welfare-enhancing; indeed some norms would appear not to have *any* direct welfare implications.

Consider norms of etiquette, such as the fine points of table manners. The welfare consequences are so trivial that it is hard to see why anyone bothers with them. No one is harmed, for example, if I wear a hat to dinner or eat peas with my fingers. The fact is, however, that such indiscretions may do serious harm to my reputation. In particular, they signal that I am a person who does not care about social norms, which may lead others to doubt my reliability in more important interactions (Posner, 2000). Complex social rituals allow people to signal their sensitivity to *norms in general*; they also provide a training ground for learning to follow norms, and for disciplining those who fail to do so.

Even when norms do have direct welfare implications, one cannot conclude that societies will opt for efficient norms. It is doubtful, for example, that norms of retribution are efficient, or that pushing is superior to queuing. Yet these are the operative norms in quite a few cases. The problem is that norms are not 'chosen'; they arise from historical accident and the accumulation of precedent. Once expectations converge on an inefficient norm, it can be very difficult to dislodge.

Over the longer run, it is conceivable that societies could somehow extricate themselves from inefficient outcomes. One way that this could happen is that societies with superior norms simply displace societies with inferior norms, through growth, conquest, or migration. Another possibility is that societies with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
inferior norms imitate the practices of more successful ones (Robson and Vega-Redondo, 1996; Boyd and Richerson, 2002). Yet a third possibility is that norm change comes from within, the result of gradual and almost imperceptible changes in expectations that 'tip' the society into a new way of doing things without anyone intending it. I discuss this possibility in more detail below.

Excess uniformity

I have argued that some norms may be inferior from the standpoint of welfare, yet stay in place for long periods of time. Others may have little direct effect on welfare but serve an important signalling function. Another way in which norms can affect economic welfare is by imposing excess uniformity on behaviour. People might be better off if they adapted their actions to their particular circumstances. To illustrate, consider a situation in which a principal and an agent are bargaining over the terms of a contract. To be concrete, think of a landlord bargaining with a prospective tenant over the terms for renting a plot of land. In theory, the optimal contract will depend on a variety of factors that may be idiosyncratic to the contracting parties, including information asymmetries, monitoring costs, and attitudes towards risk (Cheung, 1969). Yet, in practice, contracts often exhibit a high degree of uniformity, and employ 'usual and customary' terms that mask idiosyncratic differences (Bardhan, 1984). In late 20th-century Illinois agriculture, for example, more than 50 per cent of the contracts specified fixed shares between tenant and landlord, and of these more than 90 per cent specified the shares 1/2–1/2, 2/5–3/5, or 1/3–2/3 for landlord and tenant respectively (Young and Burke, 2001). The logic of 'usual and customary' contractual terms is that they create a focal solution in a situation that has many possible solutions, thereby reducing transaction costs. Such norms are not unique to agricultural contracts: for example, building contractors and architects get customary markups over cost, franchisees pay standard percentages to their parent companies, real estate agents receive customary commissions on house sales, and so forth. While such norms may reduce transaction costs, however, the uniformity imposed by the norm may prevent the contracting parties from fully wringing out all of the potential gains in their particular circumstances. Thus, in evaluating the efficiency of a norm, one must consider both the savings in transaction costs and the costs imposed by excess uniformity.

The evolution of norms

If a norm merely represents one equilibrium out of many, how does society settle on a particular one starting from out-of-equilibrium conditions? We may distinguish three ways in which norms become established and change over time: top-down influences, including official edicts and role models; bottom-up influences in which local customs and practices coalesce into norms; and lateral influences in which established norms from one type of interaction are transferred to related types of interactions. The law, for example, operates partly from the top down: statutes and judicial rulings identify norms of acceptable behaviour in people's relations with others. At the same time, the boundary between acceptable and unacceptable behaviour is constantly in flux due to variations in the way that individual cases are resolved by individual courts (a bottom-up effect). And precedents in one domain can be transferred by analogy to other domains (a lateral effect). An example of the latter is the extension of laws regarding persons to those involving corporations. (For more on the interplay between norms and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the law see Ellickson, 1991; Posner, 2000).

As these examples suggest, the evolution of norms is a complex process that involves the interplay of many different forces. One may nevertheless gain insight into the process by examining how small variations in behaviour at the individual level can trigger major norm shifts at the societal level. Consider a symmetric, two-person coordination game G that is played by pairs of agents drawn from a large population. Assume for simplicity that the total number of agents is even, and that in each period everyone is paired with someone else through a random matching process. Each matched pair chooses actions simultaneously and receives the corresponding payoffs in G. Assume that each agent makes a 'trembled best response' to the distribution of choices in the previous period. Specifically, suppose that for some $0 < \varepsilon, \lambda < 1$, each agent chooses a best response with probability $\lambda(1-\varepsilon)$, trembles with probability $\lambda\ \varepsilon$ (in which case he chooses an action uniformly at random), and chooses the same action as before with probability $1-\lambda$ due to inertia (Kandori, Mailath and Rob, 1993; Young, 1993a).

This type of learning process has rather striking implications for the social norms that are most likely to emerge and remain in place for long periods of time. To illustrate, consider a competition between alternative forms of money. Suppose there are m different commodities, indexed $1 \leq k \leq m$. Assume that, in a given pairwise interaction, each player's payoff is a_k if both adopt the *k*th form of money, whereas their payoffs are zero if they adopt different forms. It can be shown that this trembled best-response process selects the efficient equilibrium with high probability, that is, when ε is small the probability is high that, in the long run, almost everyone in the population will be using the form of money with the highest payoff (Kandori and Rob, 1995).

Unfortunately, efficiency may fail when the game does not have the structure of a pure coordination game. Suppose, for example, that a potential form of money generates payoffs in two ways: as a medium of exchange and as a form of adornment. In each pairwise interaction, let $a_k > 0$ be the payoff in each period from using *k* as a medium of exchange (on the assumption that the other player also uses *k*), and let $b_k > 0$ be the payoff from using it instead as jewellery. When there are just two commodities one obtains the symmetric payoff matrix

$$\begin{matrix} a_1 + b_1, a_1 + b_1 & b_1, b_2 \\ b_2, b_1 & a_2 + b_2, a_2 + b_2 \end{matrix}$$

Assume that this is a coordination game: $a_1 + b_1 > b_2$ and $a_2 + b_2 > b_1$. Commodity *k* is *efficient* if it maximizes $a_k + b_k$. However, the evolutionary process defined above selects the *risk-dominant* commodity, that is, the commodity *k* that maximizes $a_k + 2b_k$ (Young, 1998b). The latter criterion gives twice as much weight to the payoffs from adornment (which do not require coordination) as to the payoffs that arise from using the same medium of exchange (which do require coordination). Moreover, this conclusion holds under a wide range of assumptions about the nature of the trembled best-response process (Blume, 2003).

Evolution and fairness norms

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The evolutionary framework outlined above has implications not only for the efficiency of norms but for their distributive properties as well. Consider a situation in which a principal and an agent must agree on the form of contract that will govern their relationship. Different types of contracts will have different distributive implications, some favouring the principal, others favouring the agent. Let us restrict attention to just those contracts that leave both parties better off than they would be under their outside options. To illustrate, suppose that just three contract forms are available: A, B, C. Assume a one-shot, take-it-or-leave-it bargaining process in which a given contract is adopted if and only if both parties simultaneously agree to it; otherwise they fall back on their outside options. Without loss of generality one can assume that the outside options have zero utility for both players. Consider the following example:

		Agents		
		A	*B*	*C*
	A	5, 1	0, 0	0, 0
Principals	*B*	0, 0	3, 3	0, 0
	C	0, 0	0, 0	1, 5

Contract *A* favours the principal, *C* favours the agent, and *B* is a compromise between *A* and *C*. (Of course, in reality there may be many more contract forms, but this does not change the analysis in any fundamental way.) Consider an evolutionary process like the one for money conventions, but with two distinct populations – one of principals, the other of agents – that are randomly matched in each period. Assume, as in the previous model, that they play trembled best responses with inertia, where 'best response' is defined relative to the frequency of play of the opposite population in the previous period. A *contractual norm* is a situation in which the same contract is agreed to by everyone. In this example all three norms are efficient: none of them Pareto dominates another. It can be shown, however, that the evolutionary process favours exactly one of them, namely, the compromise contract *B*. More generally, in evolutionary processes of this sort there tends to be a selection bias toward outcomes that represent a compromise for the two parties, and against extreme outcomes that lie near the boundary of the payoff-possibility set (Young, 1998b). This suggests that *norms of fairness* may result from evolutionary forces, an idea that is explored by Binmore (1994; 2005) and Young (1998a).

General implications

Although evolutionary accounts of norm formation vary in their details, they have several qualitative implications that hold quite generally. One is that different societies often employ different norms for solving the same type of coordination problem. This follows from the fact that norms represent alternative equilibria that can become established through different sequences of chance events. This is known as the *local conformity/global diversity* effect (Young, 1998a). It has been documented in a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
variety of settings, including agricultural contracting (Young and Burke, 2001), and the manner in which subjects divide payoffs in experimental situations (Henrich et al., 2004).

A second implication is that, due to stochastic perturbations, norms occasionally shift, and these shifts tend to be quite rapid compared with the long periods of stasis when a given norm is in place. This is the *tipping* or *punctuated equilibrium* effect (Young, 1998a).

A third implication is that some norms are inherently more stable or durable than others: once established they tend to remain in place for long periods of time even when buffeted by stochastic shocks. These *stochastically stable norms* depend on the payoff structure of the underlying game, and also on the nature of the stochastic perturbations (Foster and Young, 1990; Young, 1993a; Kandori, Mailath and Rob, 1993; Samuelson, 1997). Irrespective of these details, the important point is that some norms are remarkably resilient under changing circumstances. Due to their longevity, such norms may come to be seen as right and necessary, though in fact they are the product of chance and contingency, and are sustained simply because they coordinate people's expectations about how to interact with one another.

See Also

- Hume, David
- social capital

Bibliography

Bardhan, P. 1984. *Land, Labor, and Rural Poverty: Essays in Development Economics*. New York: Columbia University Press.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.

Binmore, K. 1994. *Game Theory and the Social Contract, Vol. 1: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, K. 2005. *Natural Justice*. Oxford: Oxford University Press.

Blume, L. 2003. How noise matters. *Games and Economic Behavior* 44, 251–71.

Boyd, R. and Richerson, P. 2002. Group beneficial norms can spread rapidly in a structured population. *Journal of Theoretical Biology* 215, 287–96.

Cheung, S. 1969. *The Theory of Share Tenancy*. Chicago: University of Chicago Press.

Coleman, J. 1987. Norms as social capital. In *Economic Imperialism: The Economic Approach Applied*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Outside the Field of Economics, ed. G. Radnitzky and P. Bernholz. New York: Paragon House.

Coleman, J. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Ellickson, R. 1991. *Order Without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

Elster, J. 1989. Social norms and economic theory. *Journal of Economic Perspectives* 3(4), 99–117.

Elster, J. 1999. *Alchemies of the Mind*. Cambridge, UK: Cambridge University Press.

Fehr, E., Fischbacher, U. and Gächter, S. 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* 13, 1–25.

Foster, D. and Young, H.P. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38, 219–32.

Harsanyi, J. and Selten, R. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Hechter, M. and Opp, K.-D., eds. 2001. *Social Norms*. New York: Russell Sage Foundation.

Henrich, J., Boyd, R., Bowles, S., Fehr, E. and Gintis, H., eds. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence in 15 Small-Scale Societies*. Oxford: Oxford University Press.

Hume, D. 1739. *A Treatise of Human Nature*. Oxford: Oxford University Press, 1978.

Kandori, M., Mailath, G. and Rob, R. 1993. Learning, mutation, and long-run equilibria in games. *Econometrica* 61, 29–56.

Kandori, M. and Rob, Rl. 1995. Evolution of equilibria in the long run: a general theory and applications. *Journal of Economic Theory* 65, 29–56.

Lewis, D. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Posner, E. 2000. *Law and Social Norms*. Cambridge, MA: Harvard University Press.

Robson, A. and Vega-Redondo, F. 1996. Efficient equilibrium selection in evolutionary games with random matching. *Journal of Economic Theory* 70, 65–92.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Roth, A. 1985. Toward a focal point theory of bargaining. In *Game-Theoretic Models of Bargaining*, ed. A. Roth. Cambridge, UK and New York: Cambridge University Press.

Samuelson, L. 1997. *Evolutionary Games and Equilibrium Selection*. Cambridge, MA: MIT Press.

Schelling, T.C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sugden, R. 1986. *The Economics of Rights, Cooperation and Welfare*. Oxford: Basil Blackwell.

Ullmann-Margalit, E. 1977. *The Emergence of Norms*. Oxford: Oxford University Press.

Wärneryd, K. 1994. Transaction cost, institutions, and evolution. *Journal of Economic Behavior and Organization* 25, 219–39.

Young, H.P. 1993a. The evolution of conventions. *Econometrica* 61, 57–84.

Young, H.P. 1993b. An evolutionary model of bargaining. *Journal of Economic Theory* 59, 145–68.

Young, H.P. 1998a. *Individual Strategy and Social Structure*. Princeton, NJ: Princeton University Press.

Young, H.P. 1998b. Conventional contracts. *Review of Economic Studies* 65, 773–92.

Young, H.P. and Burke, M. 2001. Competition and custom in economic contracts: a case study of Illinois agriculture. *American Economic Review* 91, 559–73.

How to cite this article

Young, H. Peyton. "social norms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000466> doi:10.1057/9780230226203.1563

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social preferences

Jeffrey Carpenter
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Behaviour in a variety of games is inconsistent with the traditional formulation of egoistic decision-makers; however, the observed differences are often systematic and robust. In many cases, people behave as if they value the outcomes accruing to other reference agents. In reaction, behavioural economists have offered and tested a variety of formulations (such as inequality aversion and reciprocity) that capture the social nature of preferences.

Keywords

altruism; collective action; dictator game; fairness; free rider problem; gift exchange; inequality aversion; labour markets; psychological games; public good experiments; reciprocity; self-interest; social dilemma; social norms; social preferences; subgame perfection; trust game; ultimatum game

Article

For the longest time economists reacted allergically to preference formulations that allowed for anything but material self-interest (cf. Binmore, Shaked and Sutton, 1985). The reaction was well founded: by adding elements to the agent's utility function, potentially one allows economic theory to explain everything and, therefore, nothing. Any behaviour can be explained by assuming it is preferred. However, this strong position has sometimes made economics seem out of touch with the world economists try to explain. Even economists care about the outcomes achieved by others, in addition to their own outcomes. Moreover, they also care about how those outcomes are achieved. Only in 1982, however, was the weakness of taking material self-interest for granted demonstrated by Werner Güth and his co-authors, who showed that economic theory failed in the simplest of decision settings (Güth, Schmittberger and Schwarze, 1982), the ultimatum game. In this game a first mover offers a share of a monetary ‘pie’ to a second mover who either accepts the proposal, in which case it is divided as proposed, or rejects the proposal, in which case both players earn nothing. Since then this game has become the workhorse of experimenters intent on exploring carefully the extent to which people behave in ways that are contrary to their material self-interest.

While it is interesting to document the fact that people consider the outcomes of others when they make choices in experimental games, there are at least two other particularly compelling aspects of the research that has developed since the 1980s. First, these deviations from self-interest can be replicated, and have been, both inside and outside the laboratory. Replication suggests that these behaviours are not just errors or flukes, and therefore, although self-interest is a convenient modelling assumption, it should not be used as the basis for policy formulation. Second, this research illustrates that there is a difference between theory failing because of a false assumption and its failing because of flawed logic. Research shows that people do use economic reasoning, but that they, or most of them, are not narrowly self-interested.

The original results of the ultimatum game provided the impetus for a large body of research. Initially, some researchers were convinced that the explanation was not a concern for others but simple error (for example, Binmore, Shaked and Sutton, 1985). However, this explanation was soon swept aside by volumes of evidence from a variety of games that suggested that the payoffs of other players entered into the strategic choices of experimental participants (see the reviews of Bowles, 2004; or Sobel, 2005). Despite all this research, a precise definition of social preference has not been settled upon. In most cases, ‘social preference’ is defined loosely as *a concern for the payoffs allocated to other relevant reference agents in addition to the concern for one's own payoff*. (A largely separate branch of research has focused on altruism and warm glow motives for giving to others, especially in the context of public goods provision. This work is discussed elsewhere in the dictionary.)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Within the standard outcome-oriented definition, research has focused on identifying the more pro-social preferences for altruism and inequality aversion while considerably less attention has been given to their opposites, spite and eminence. The evidence from the hundreds of ultimatum games conducted since 1982 suggests that, on the second-mover side of the game, few people are willing to accept the low offers associated with the subgame perfect equilibrium prediction. In fact, offers of less than 20 per cent of the pie are routinely rejected, and as offers increase they are more likely to be accepted (Camerer, 2003). Turning down positive offers is clearly against one's material self-interest, but it is consistent with aversion to unequal payoffs (inequality aversion). As the stakes increase, the probability of a rejection falls, but even when the pie is as large as three months expenditures the rejection rate is not zero (Cameron, 1999).

Interpreting the motivation of the first mover in the ultimatum game is not as straightforward, though. One hypothesis is that proposers offer half the pie because they are inequality averse. We cannot, however, distinguish this reasoning from that of completely selfish, but astute, proposers who anticipate that low offers will be rejected and offer half because they know it will be accepted. The dictator game evolved to identify the motives of first movers (Forsythe et al., 1994). The dictator game is played just like the ultimatum game except for one very important design change: second movers are passive recipients of whatever they are allocated. In other words, they cannot reject offers. If the enlightened self-interest hypothesis is correct, we would expect to see first movers allocating nothing in the dictator game. This is not the case. Although allocations in the dictator game are susceptible to changes in the presentation of the game (Hoffman et al., 1994; Eckel and Grossman, 1996), it is common for people to allocate positive amounts. In fact, it is common for the behaviour of non-student participants in the two games to be indistinguishable (Carpenter, Burks and Verhoogen, 2005) suggesting that many people prefer equal outcomes.

There is some question as to whether the simple outcome-oriented definition of social preference is sufficient. An example illustrates why. Instead of offers being generated by other participants, imagine second movers in the ultimatum game being assigned offers randomly by a computer programme. If inequality aversion is a sufficient description of the motivations of participants, this change should have no impact on behaviour. However, it does: responders are much less likely to reject computer-generated offers than offers that come from real proposers (Blount, 1995). This indicates that people are also interested in the process and intentions that generate outcomes. The definition of social preference should perhaps be expanded accordingly to *a concern for the payoffs allocated to other relevant reference agents and the intentions that led to this payoff profile in addition to the concern for one's own payoff*.

Expanding the definition of social preference to include a process component allows us to also classify reciprocity – treating only kind acts with kindness – as a social preference. Pure reciprocity, however, is more elusive than inequality aversion because one needs to show that outcomes and intentions matter. Only a few experiments have been conducted to show that intentions matter, but the results are compelling. For example, imagine two binary choice versions of the ultimatum game (Falk, Fehr and Fischbacher, 2003). In game A, the proposer can decide between claiming the lion's share of a ten-dollar pie (8, 2) and sharing the pie equally (5, 5). In game B, the first option is the same (8, 2) but the second is even worse for the second mover because the proposer demands the whole pie (10, 0). Inequality aversion predicts that the (8, 2) offer will be rejected at the same rate in the two games because the other offer is irrelevant – the decision-maker should focus only on the outcome presented. Reciprocity, on the other hand, suggests that one would be much less likely to reject (8, 2) in game B because it is the kinder of the two offers. Indeed, people are almost five times more likely to reject the (8, 2) offer in game A. An alternative approach is to compare the response of participants to different outcome allocations after another participant has made a kind or unkind act to the response when there is no initial move by another participant (Charness and Rabin, 2002). Reciprocity is identified by the subtraction of the first outcomes and intentions experiment from the second baseline inequality-aversion experiment.

In the trust (or investment) game, a first mover decides how much to send to a second mover. Any amount sent is multiplied by $k>1$ before it reaches the second-mover. The second mover then decides how much to send back. Because of the multiplication, sending money is socially efficient yet a first mover should send money only if she trusts the second mover to send back at least enough to cover the investment. The standard interpretation is that the first mover must expect the second mover to be motivated by reciprocity before it makes sense to invest in the partnership (Berg, Dickaut and McCabe, 1995). However, one can just as easily invoke inequality aversion to explain the fact that people tend to send back more when they receive more (Cox, 2004). The same problem exists with the related experiments developed to test for the notion of gift exchange in the labour market context (for example, Fehr and Schmidt, 1999).

Other, more indirect, evidence for reciprocity and the more nuanced definition of social preference comes from the experimental literature on voluntary contributions to public goods. In these settings participants are given an endowment and asked to decide how much to contribute to a ‘group project’. The incentives are of a social dilemma; contributing nothing is a dominant strategy but contributing everything is socially efficient. Playing the public goods game in strategic form asks participants to decide how much they want to contribute conditional on the contributions of others. Half the participants are conditionally cooperative in that they generate contribution schedules that are increasing in the contributions of others (Fischbacher, Gächter and Fehr, 2001). The fact that people condition their contributions according to those of others suggests that intentions and reciprocity matter.

To identify reciprocity separately from inequality aversion one may employ a design in which the two forces pull in different directions. Imagine that one can punish free riders in the public goods game: a participant can impose a penalty p at a cost c. In most cases people punish despite it being dominant to free ride on the punishment done by others (that is, punishment is just a second-order public good), and this tends to stabilize contributions (Fehr and Gächter, 2000). However, in most cases $p>c$, which means cooperators reduce the inequality between themselves and the free rider by punishing. To isolate the role of, in this case negative, reciprocity one can allow $p<c$, which actually increases the inequality. Although they do it less often, people punish when the sanction delivered is lower than the cost, and this is a nice demonstration of reciprocity (Carpenter, 2007).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Several attempts have been made to organize the evidence on social preferences into parsimonious, but flexible, utility functions. One of the most successful outcome-oriented approaches is the Fehr and Schmidt (1999) specification, perhaps because it is relatively easy to work with. Here the utility of player *i* increases in her own payoff, x_i, but decreases in any difference between her payoff and the payoffs of other relevant players. For two-player games this is just:

$$u_i(x_i, x_j) = \begin{cases} x_i - \alpha_i(x_j - x_i) & \text{if } x_i < x_j \\ x_i - \beta_i(x_i - x_j) & \text{if } x_i \geq x_j \end{cases}$$

where α_i is player *i*'s degree of inferiority aversion and β_i is her degree of superiority aversion. It is natural to expect $\alpha_i > \beta_i$.

While this utility function is a good first approximation because it has been shown to be consistent with much of the experimental data (if one is willing to make assumptions about the distribution of α's and β's in the population) it is limited in two ways. First, as one can see in Figure 1, the predictions can be coarse. It is not hard to graph the indifference curves associated with the Fehr–Schmidt specification, but if one superimposes a budget constraint on the indifference mapping there are just two predictions: keep it all or give away half unless the constraint has exactly the same slope as the indifference curve, in which case any amount between nothing and half is possible.

Figure 1

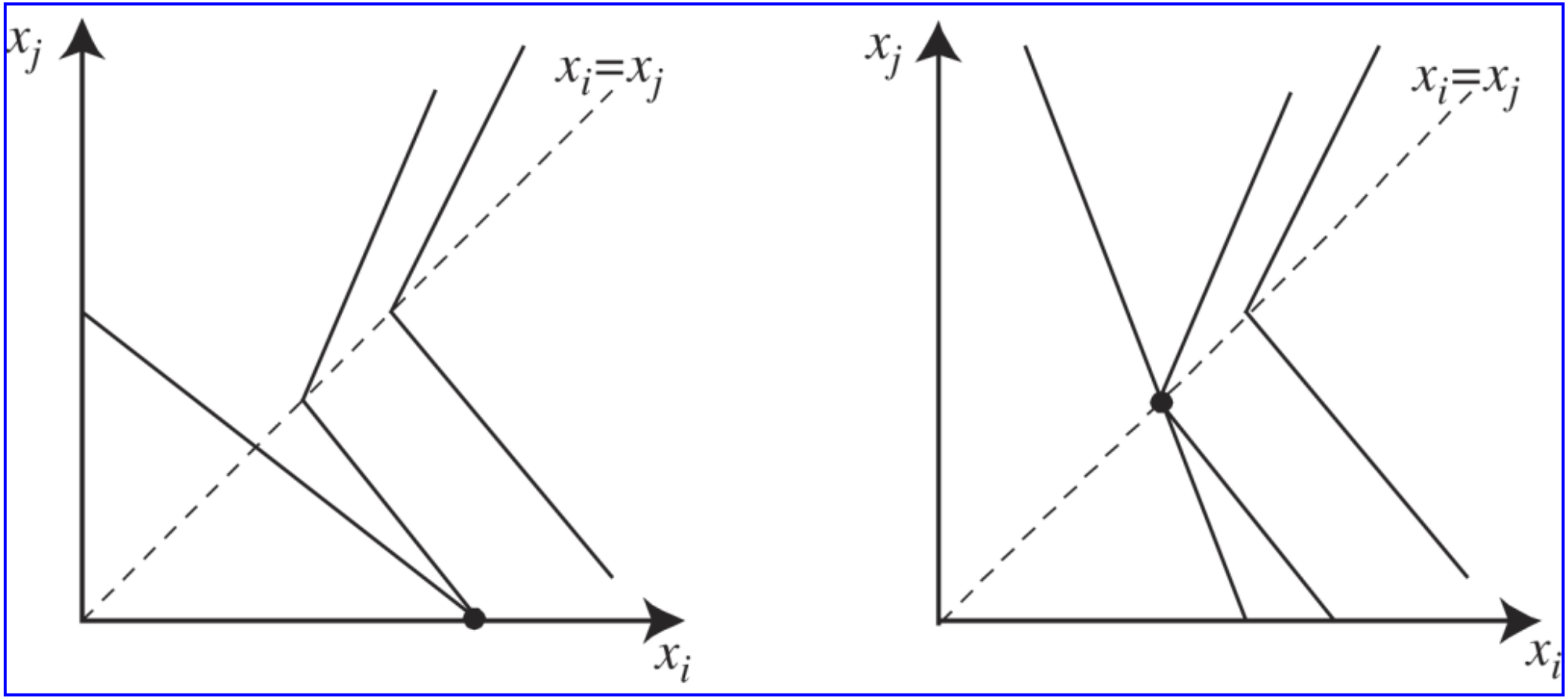

The fact that intentions play no role is a second problem faced by all the outcome-oriented approaches. A trade-off does, however, exist because incorporating intentions makes the specifications considerably harder to work with. The outcome- and process-oriented specifications evolved from the notion of *psychological games*, which posits that utility will

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

depend on both outcomes and beliefs (Geanakoplos, Pearce and Stacchetti, 1989). Beliefs are important because emotional responses are often triggered by expectations about how one should be treated. Perhaps the specification that is easiest to work with is the Charness–Rabin utility function, which incorporates a term θq to capture reciprocal motivations:

$$u_i(x_i, x_j) = (\rho r + \sigma s + \theta q)x_j + (1 - \rho r - \sigma s - \theta q)x_i.$$

The parameters r and s indicate which of the two players has the advantage (r=1 if $x_i > x_j$, s=1 if $x_j > x_i$ and r=s=0 otherwise) and the parameters ρ and σ represent outcome-oriented preferences (Charness and Rabin, 2002). To recover the Fehr–Schmidt specification we simply assume $\sigma < 0 < \rho < 1$ and $\theta = 0$. Reciprocity and intentions are at work if $\theta > 0$ because we set q=−1 if player j has misbehaved and q=0 otherwise.

Why should economists care about social preferences? By ignoring social preferences, economists have incompletely characterized many important interactions (Fehr and Fischbacher, 2002). Because many people are motivated by notions of fairness and reciprocity, social preferences can hinder the dynamics of competition that are assumed to drive equilibria, especially in the context of labour markets. For example, wages may never fall to the competitive equilibrium level because bosses understand that workers are reciprocally motivated. By lowering the wage, the boss also lowers morale and productivity (Bewley, 1999). Likewise, the economic theory of collective action is only narrowly applicable because it fails to realize that most people are predisposed to cooperate, but hate being taken advantage of (Andreoni, 1988). Designing incentives is ultimately more challenging when one accounts for the heterogeneity of social motivations identified in economic experiments.

Future research on social preferences is likely to extend in a number of interesting directions. Experimenters have begun to move from the laboratory to the field to identify the preferences of more representative samples and to investigate the external validity of these preferences (that is, what important behaviours and outcomes do social preferences correlate with?). Within the laboratory it will be interesting to better isolate the role of outcomes versus the role of intentions, to examine the co-evolution of preferences and institutions, and to examine the difference between social preferences and social norms. Is it the case, for example, that norms dictate how one should treat others regardless of whether the prescribed behaviour is consistent with one's underlying preferences?

See Also

- altruism in experiments
- economic man
- public goods experiments
- trust in experiments

Bibliography

Andreoni, J. 1988. Why free ride? Strategies and learning in public good experiments. *Journal of Public Economics* 37, 291–304.

Berg, J., Dickaut, J. and McCabe, K. 1995. Trust, reciprocity and social history. *Games and Economic Behavior* 10, 122–42.

Bewley, T. 1999. *Why Wages Don't Fall During a Recession*. Cambridge, MA: Harvard University Press.

Binmore, K., Shaked, A. and Sutton, J. 1985. Testing noncooperative bargaining theory: a preliminary study. *American Economic Review* 75, 1178–80.

Blount, S. 1995. When social outcomes aren't fair: the effect of causal attribution on preferences. *Organizational Behavior & Human Decision Processes* 62, 131–44.

Bowles, S. 2004. *Microeconomics: Behavior, Institutions and Evolution*. Princeton: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Camerer, C. 2003. *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton: Princeton University Press.

Cameron, L. 1999. Raising the stakes in the ultimatum game: experimental evidence from Indonesia. *Economic Inquiry* 37, 47–59.

Carpenter, J. 2007. The demand for punishment. *Journal of Economic Behavior & Organization* 62, 522–42.

Carpenter, J., Burks, S. and Verhoogen, E. 2005. Comparing students to workers: the effects of social framing on behavior in distribution games. In *Field Experiments in Economics. Research in Experimental Economics*, ed. J. Carpenter, G. Harrison and J. List. Greenwich, CT and London: JAI/Elsevier.

Charness, G. and Rabin, M. 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117, 817–70.

Cox, J.C. 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46, 260–81.

Eckel, C. and Grossman, P. 1996. Altruism in anonymous dictator games. *Games and Economic Behavior* 16, 181–91.

Falk, A., Fehr, E. and Fischbacher, U. 2003. On the nature of fair behavior. *Economic Inquiry* 41, 20–6.

Fehr, E. and Fischbacher, U. 2002. Why social preferences matter – the impact of non-selfish motives on competition, cooperation and incentives. *Economic Journal* 112, 1–33.

Fehr, E. and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90, 980–94.

Fehr, E. and Schmidt, K. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 769–816.

Fischbacher, U., Gächter, S. and Fehr, E. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters* 71, 397–404.

Forsythe, R., Horowitz, J., Savin, N.E. and Sefton, M. 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior* 6, 347–69.

Geanakoplos, J., Pearce, D. and Stacchetti, E. 1989. Psychological games and sequential rationality. *Games and Economic Behavior* 1, 60–79.

Güth, W., Schmittberger, R. and Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3, 367–88.

Hoffman, E., McCabe, K., Shachat, J. and Smith, V. 1994. Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* 7, 346–80.

Sobel, J. 2005. Interdependent preferences and reciprocity. *Journal of Economic Literature* 43, 392–436.

How to cite this article

Carpenter, Jeffrey. "social preferences." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_P000153> doi:10.1057/9780230226203.1564

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Social Security in the United States

Selahattin •mrohoro•lu

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social Security is a pay-as-you-go US federal government programme that provides pensions for retirees and their surviving family members. It is a public annuity that insures against longevity and stock market risks. The payroll tax that finances it discourages private saving and work effort. The projected aging of the US population will strain the programme and necessitate either a reduction in benefits, an increase in payroll taxes, partial privatization, or some combination of these policy reforms, by 2030.

Keywords

adverse selection; commitment; crowding out; dependency ratio; longevity insurance; Medicaid (USA); Medicare (USA); overlapping generations models; payroll tax; pensions; retirement; Social Security in the United States

Article

Social Security is a US federal government programme that provides retirement, survivors, and disability benefits for eligible workers and their dependents. It is largely a pay-as-you-go (PAYG) tax-financed system; current workers are taxed and these revenues are used to finance the old age, survivors (widows, widowers, surviving divorced spouses, surviving children, and parents of deceased qualified workers), and disability insurance payments. Health insurance for the elderly is administered separately in another federal programme called Medicare. Retirement, survivor and disability benefits are financed by Old Age, Survivors, and Disability Insurance (OASDI) taxes, and Medicare is financed by Health Insurance (HI) taxes. In addition to the Social Security Disability Insurance, there is Supplemental Security Income that pays benefits based on financial need. Health insurance for the poor, regardless of age, is called Medicaid, and this is financed jointly from general federal tax revenues and the states.

Eligibility and coverage

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As of 2007, nearly 150 million US workers and their dependents are covered by this compulsory system. (Among the few workers who are exempted from Social Security are federal civilian workers hired prior to 1984, some state and local government workers, and household, agricultural, and self-employed workers whose earnings are too low.)

There are about 40 million retirees and survivors drawing benefits from the system. Eligibility for full retirement benefits typically requires working for ten years, although early retirement at age 62 is possible with reduced retirement benefits. The normal retirement age is currently 66, set to rise to 67 by the time the 1960 cohort retires with full benefits in 2027. Retirement benefits are about 40 per cent of lifetime average earnings of a median wage earner. High-wage earners get less than this average replacement rate and low-wage earners get more than 40 per cent. Up to an additional one-half of the retired worker's full benefits may be paid to a spouse if the spouse has not worked or has low earnings. Under certain conditions, a divorced spouse can also get benefits if the marriage lasted at least ten years. If there are children eligible for Social Security, each receives up to 50 per cent of full benefits, with a maximum of 150–80 per cent of a worker's own benefit payments as a family limit. (For details and the separate Disability Insurance program, see http://www.ssa.gov.)

The tax rate for OASDI in 2007 is 12.4 per cent, paid equally by the employer and the employee (or entirely if the individual is self-employed). There is no tax on earnings over $90,000 in 2005 dollars, with this maximum rising every year by the national average wage index. (The HI tax rate is 2.9 per cent on all earnings.)

A private retirement fund is typically invested in a portfolio of stocks, commercial paper, and government bonds. Social Security is largely a PAYG system with current taxpayers paying for current retirees' benefits. Since the mid-1980s, Social Security has collected more in taxes than it pays in benefits. This surplus is loaned to the US Treasury in return for special-issue Treasury bonds that are used to finance other federal government expenditures. This Trust Fund stood at $1.8 trillion at the end of 2006. Although this sum seems quite large, it is actually quite small compared with the annual OASDI benefits which amounted to $461 billion in 2006. (This sum does not include Disability Insurance payments, which amounted to $91 billion in 2006.)

A short history of social security

In 1889, Germany became the first country to provide old-age insurance on a large scale. Designed by the Chancellor, Otto von Bismarck, the German system required mandatory participation, taxed the employer and the employee, and used government taxes to provide retirement and disability benefits. The retirement age was 70, though it was lowered to 65 in 1916.

In the United States, the German system was viewed as a model. In 1935 The Social Security Act, which covered workers in commerce and industry, was signed by President Roosevelt. The original proposal and draft legislation, the Economic Security Act, was a three-tiered social pension programme: (*a*) old-age welfare payments, (*b*) mandatory, contributory old-age retirement programme, which evolved into what is now called Social Security, and (*c*) voluntary annuity sales through which the federal government would sell certificates to workers who could, upon reaching the retirement age, convert them into monthly annuities to supplement their basic compulsory retirement benefits. However, the Congress rejected this third proposal by President Roosevelt, which was the earliest attempt to allow for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
a private saving programme for retirement with mandatory annuitization at retirement.

Since 1935, many changes were made to the original 1935 act. In 1937, workers were required to pay two per cent of their payroll to support the Social Security system. In 1939, the programme started to cover dependents and survivors. In 1950, coverage was extended beyond commerce and industry, the tax rate was raised to three per cent and benefits were also raised. In 1956, the tax rate was raised to four per cent and disability insurance was introduced. Early retirement for women was permitted in 1956 and for men in 1961, when payroll taxes were raised to six per cent. In 1972, cost-of-living-adjustments (COLAs) were introduced, automatically indexing benefit levels to inflation, and payroll taxes were raised to 9.2 per cent. In 1977, the tax rate jumped to 9.9 per cent.

In 1983 the National Commission on Social Security Reform was formed to suggest ways to deal with the actuarial imbalance of the system. The commission proposed (*a*) a phased-out increase in the retirement age from 65 to 67, (*b*) an increase in the self-employment tax, (*c*) partial taxation of benefits to upper income retirees, and (*d*) expanding the coverage to include federal civilian and non-profit organization employees. The tax rate was raised to 10.8 per cent. The payroll tax rate was raised to 11.4 per cent in 1985 and to 12.4 per cent in 1993. In 1996 the Social Security Trustees (1996) reported that the Social Security system would start to run deficits in 2012, and the trust funds would be exhausted by 2029. They predicted that, in order to keep the benefit levels unchanged, the tax rate would have to rise to 18 per cent.

The simple economics of social security

A simple textbook treatment will be presented here. For a more rigorous treatment of social security and overlapping generations models, the reader should see the seminal papers by Samuelson (1958) and Diamond (1965). For a recent and thorough coverage, see Ljungqvist and Sargent (2005).

Consider a simple overlapping generations model in which individuals live for two periods. Their earnings are y when young, and 0 when old. They face a social security tax $\tau > 0$ when young and receive a retirement benefit b when old. They earn interest on their saving, s at the rate r. There is no uncertainty. The economy is assumed to have perfectly functioning credit markets that allow individuals to borrow any amount at the market rate of interest r. Population grows at the rate $n>0$, and productivity (and hence real income) grows at the rate $g>0$. The lifetime utility function of the individuals is given by

$$\log(c_1) + \beta\log(c_2) \tag{1}$$

where c_1 and c_2 are consumption when young and when old, respectively, and $\beta > 0$ is the subjective discount factor. The budget constraints faced by the individuals are given as

$$c_1 + s = y \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$c_2 = (1 + r)s + b. \tag{3}$$

We can write the lifetime budget constraint facing the agents in present value form

$$c_1 + \frac{c_2}{(1 + r)} = y + \frac{b}{(1 + r)}. \tag{4}$$

The social security system is unfunded, which requires that taxes collected from the young equal benefits given out to the retirees at each period:

$$b = (1 + g)(1 + n)\tau y. \tag{5}$$

The Lagrangian for the individuals' optimization problem is given by

$$L = \log(c_1) + \beta\log(c_2) + \lambda\left[y + \frac{b}{(1 + r)} - c_1 - \frac{c_2}{(1 + r)}\right], \tag{6}$$

where $\lambda > 0$ is the Lagrange multiplier.
The first-order conditions for maximizing the Lagrangian with respect to the choice variables c_1 and c_2 (and λ) yield optimal lifetime consumption quantities as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$c_1 = \frac{(1+r)(1-\tau)y + b}{(1+r)(1+\beta)}, \text{ and}$$
(7)

$$c_2 = \frac{\beta(1+r)(1-\tau)y + \beta b}{(1+\beta)}.$$
(8)

The second-order conditions are satisfied given our choices of functional forms and parameter restrictions.
Using eqs (2), (3) and (8), we obtain the individual's private saving under social security as

$$s^{ss} = \frac{\beta(1-\tau)y - b/(1+r)}{(1+\beta)}.$$
(9)

In the absence of an unfunded social security system, that is, with $\tau = 0$ and $b=0$ saving in this 'laissez-faire' world is given by

$$s^{lf} = \frac{\beta y}{(1+\beta)}.$$
(10)

Equations (9) and (10) show that the introduction of an unfunded social security system reduces private saving. The productive capital stock in an economy is determined to a large extent by private saving, and therefore an unfunded social security system results in a lower capital stock and per capita consumption. (If factor prices were determined by the capital–labour ratio, then the general equilibrium effect would be to mitigate the reduction in private saving as the return to capital would rise in response to the decrease in private saving. However, the direction of the change in saving is still the same, and, in quantitative versions of the life cycle model, the magnitude of this effect is quite large. Also, if labour were endogenous, one could show that a public pension system with no linkage between contributions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and benefits distorts labour supply and causes it to decline, increasing the welfare cost of social security.) What is the impact on the individual's lifetime welfare? To address this issue, we need to compare the lifetime consumption values with and without social security.
Substituting eq. (5) into eqs (7) and (8) yields

$$c_1^{ss} = \frac{(1+r)y - \tau y[(1+r) - (1+g)(1+n)]}{(1+r)(1+\beta)}, \quad (11)$$

$$c_2^{ss} = \frac{\beta(1+r)y - \beta\tau y[(1+r) - (1+g)(1+n)]}{(1+r)(1+\beta)}, \quad (12)$$

where c_1^{ss} and c_2^{ss} denote young- and old-age consumption under social security. Imposing $\tau = 0$ and $b=0$, and using c_1^{lf} and c_2^{lf} to denote young- and old-age consumption under laissez-faire, eqs (11) and (12) become

$$c_1^{lf} = \frac{(1+r)y}{(1+r)(1+\beta)}, \quad (13)$$

$$c_2^{lf} = \frac{\beta(1+r)y}{(1+r)(1+\beta)}. \quad (14)$$

Given our assumptions on the parameters,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$c_1^{lf} > c_1^{ss} \text{ and } c_2^{lf} > c_2^{ss} \text{ if and only if } [(1+r)-(1+g)(1+n)] > 0. \quad (15)$$

In other words, unfunded social security yields lower young- and old-age consumption to the individual if the rate of return on capital, $1+r$ exceeds the rate of growth of the economy, $(1+g)(1+n)$. Historically, the annual growth rate of the US economy has been just under three per cent on average, whereas the return on capital has exceeded this amount significantly, by about two to five percentage points, depending on alternative definitions of the capital stock. (The condition presented in eq. (15) is known as the condition for dynamic efficiency. If the reverse were true, then a reduction in saving would reduce the dynamic inefficiency – or capital overaccumulation – in the economy and improve welfare by raising consumption in both periods.) In fact, reform proposals in the early 21st century to partially privatize social security aim to exploit this return differential in order to build up private retirement funds for old-age consumption.

Pros and cons of unfunded social security

As the previous section demonstrates, the most important welfare cost of an unfunded social security system is that it discourages private saving and the accumulation of capital. There are other social costs associated with a PAYG system, as well as several social benefits. In this section, these costs and benefits will be listed and the empirical evidence about their quantitative importance will be evaluated.

Welfare costs

1. 1. Unfunded social security discourages saving. Young workers with high marginal propensities to save are taxed, and these resources are given to retirees, with low marginal propensities to save.
2. 2. It discourages work effort since the payroll tax paid by the worker has less than perfect linkage, if any, with the retirement benefit that the worker will receive. The labour supply in the economy is adversely affected and so is the demand for labour as firms are reluctant to create new jobs.
3. 3. It distorts the retirement decision and encourages early retirement.
4. 4. It imposes a hardship on workers and companies that are facing borrowing and credit constraints. Removing the high payroll taxes devoted to this mandatory retirement system would bring welcome relief to a large number of households and firms.

Welfare benefits

1. 1. Unfunded social security provides longevity insurance. If an individual lives longer than expected, an annuity that provides for old-age consumption is welfare-enhancing. (Yaari, 1965,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

showed that in a simple two-period overlapping generations model rational individuals without an altruistic motive would find it optimal to annuitize all their wealth. However, the annuity markets in the United States are thin. Indeed, Friedman and Warshawsky, 1990, and Mitchell et al., 1999, document that private annuities are unattractively priced in the United States and that the adverse selection problem likely contributes to the low volume of these contracts. See also Feldstein, 2005, for an excellent survey of issues surrounding social security.)

2. 2. It provides insurance against (privately uninsurable) income shocks over the life cycle.
3. 3. It serves as a partial substitute for a commitment device for individuals with time-inconsistent preferences.
4. 4. It insulates the individuals to a large extent against aggregate shocks such as stock market risks.

Empirical evidence

Auerbach and Kotlikoff (1987) presented one of the earliest attempts at using a large-scale overlapping generations framework to address fiscal policy effects in a calibrated, general equilibrium setting. In their model, social security did not have any potential benefit and it imposed a deadweight loss on the society. Hubbard and Judd (1987) introduced the longevity insurance aspect of social security, but still the negative impact of social security on saving outweighed this potential benefit. •mrohoro•lu, •mrohoro•lu and Joines (1995; 1999) consider variations of the earlier quantitative models but they arrive at the same conclusion.

Diamond (1977) and Feldstein (1985) argue that a fraction of the population might lack the foresight to save for retirement. However, •mrohoro•lu, •mrohoro•lu and Joines (2003) use Strotz's (1956) time-inconsistent preferences to evaluate the welfare benefit of an unfunded social security system in a quantitative model that includes most of the costs and benefits of social security listed in the previous section, and find that social security is a poor commitment device and that individual's welfare is not increased unless for relatively high (and implausible) degrees of short-term discount rates.

Despite the widespread quantitative evidence on the inefficiency of social security, the institution seems resistant to reform. One reason may be due to potentially large transitional costs of privatizing the system. For example, Conesa and Krueger (1999) find large transitional costs, with a majority of the currently alive population suffering welfare losses, and thus blocking any reform proposal. Fuster, •mrohoro•lu and •mrohoro•lu (2007), on the other hand, argue that these transitional costs are easily paid for by a growing economy with strong bequest motives and flexible labour markets. Cooley and Soares (1999) and Boldrin and Rustichini (2000) study the political-equilibrium considerations that allow the introduction and maintenance of an unfunded social security system. Krueger and Kuebler (2006) introduce aggregate uncertainty in the form of ‘investment risk’ but still find that the capital crowding-out effect dominates and social security reduces welfare.

The future of social security

Notwithstanding the economic burden of PAYG systems, as the previous section summarizes, the future

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
holds much worse predictions for OECD countries. Without exception, all social security programmes are facing financing difficulties, due to significant increases in longevity and the decline in fertility. Most European countries and Japan are aging very rapidly. The United States is not far behind. The ratio of the number of retirees to the number of workers, the so-called dependency ratio, was near 20 per cent in 2000, but is expected to rise to almost 50 per cent in 2060, meaning only two workers per retiree. To get an idea of how much of a strain the aging of the US population puts on the current pension system, consider the following two findings from De Nardi, •mrohoro•lu and Sargent (1999). First, responding to a proposal to index retirement age to the increases in longevity so that the dependency ratio is held constant at the current level of 20 per cent, they calculate that the retirement age has to eventually rise to 76. Second, maintaining retirement benefits at their current levels when the population is aging (and the costs of medical services are rising) requires an increase in the Social Security payroll tax rate from its current level of about 10 per cent to an eventual 40 per cent. If a reform is not undertaken, even these high tax rates will not be enough to pay for pensions in the future.

See Also

- overlapping generations model of general equilibrium
- retirement
- social insurance
- social insurance and public policy

Bibliography

Auerbach, A. and Kotlikoff, L. 1987. *Dynamic Fiscal Policy*. Cambridge: Cambridge University Press.

Boldrin, M. and Rustichini, A. 2000. Political equilibria with social security. *Review of Economic Dynamics* 3, 41–78.

Conesa, J. and Krueger, D. 1999. Social Security reform with heterogeneous agents. *Review of Economic Dynamics* 2, 757–95.

Cooley, T. and Soares, J. 1999. A positive theory of social security based on reputation. *Journal of Political Economy* 107, 135–60.

De Nardi, M. •mrohoro•lu, S. and Sargent, T. 1999. Projected U.S. demographics and social security. *Review of Economic Dynamics* 2, 575–615.

Diamond, P. 1965. National debt in a neoclassical debt model. *Journal of Political Economy* 55, 1126–50.

Diamond, P. 1977. A framework for social security analysis. *Journal of Public Economics* 8, 275–98.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Feldstein, M. 1985. The optimal level of Social Security benefits. *Quarterly Journal of Economics* 100, 303–20.

Feldstein, M. 2005. Rethinking social insurance. *American Economic Review* 95, 1–24.

Friedman, B. and Warshawsky, M. 1990. The cost of annuities: implications for saving behavior and bequests. *Quarterly Journal of Economics* 105, 135–54.

Fuster, L., •mrohoro•lu, A. and •mrohoro•lu, S. 2007. Elimination of social security in a dynastic framework. *Review of Economic Studies* 74, 113–45.

Hubbard, G. and Judd, K. 1987. Social security and individual welfare. *American Economic Review* 77, 630–46.

•mrohoro•lu, A., •mrohoro•lu, S. and Joines, D. 1995. A life cycle analysis of social security. *Economic Theory* 6, 83–114.

•mrohoro•lu, A., •mrohoro•lu, S. and Joines, D. 1999. Social security in an overlapping generations economy with land. *Review of Economic Dynamics* 2, 638–65.

•mrohoro•lu, A., •mrohoro•lu, S. and Joines, D. 2003. Time inconsistent preferences and social security. *Quarterly Journal of Economics* 118, 745–84.

Krueger, D. and Kubler, F. 2006. Pareto improving social security reform when financial markets are incomplete. *American Economic Review* 96, 737–55.

Ljungqvist, L. and Sargent, T.J. 2005. *Recursive Macroeconomic Theory*. Cambridge: MIT Press.

Mitchell, O., Poterba, J.M., Warshawsky, M.J. and Brown, J.R. 1999. New evidence on the money's worth of individual annuities. *American Economic Review* 89, 1299–318.

Samuelson, P. 1958. An exact consumption loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66, 467–82.

Social Security Trustees. 1996. *1996 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds*. Washington, DC: U.S. Government Printing Office. Online. Available at http://www.ssa.gov/history/reports/trust/1996, accessed 8 June 2007.

Strotz, R. 1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
23, 165–80.

Yaari, M. 1965. Uncertain lifetime, life insurance, and the theory of the consumer. *Review of Economic Studies* 32, 137–50.

How to cite this article

•mrohoro•lu, Selahattin. "Social Security in the United States." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000524> doi:10.1057/9780230226203.1565

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social status, economics and

Chaim Fershtman

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social status is a social reward that affects the incentive structure facing individuals. If status is provided to educated people, more people will obtain an education. The choice of occupation is affected by the social status associated with different occupations, establishing a link between social status, the equilibrium wage structure and the allocation of workers among occupations. When status is not directly observed, people try to signal it by changing their consumption choices or behaviour. The narrow paradigm of *homo economicus* should be extended to include social status among the basic motivations for economic decisions.

Keywords

conspicuous consumption; economic growth; preference formation; saving; social status, economics and; Veblen effect; Veblen, T.; wage heterogeneity, sources of; wage rigidity; Weber, M.

Article

People are social animals who care about their standing in society and about what other members of society think about them. Stated differently, people care about their ‘prestige’ or the ‘respect’ that they are accorded by individuals with whom they interact. Many people would gladly pay large sums of money for a knighthood. Similarly, the value of the Nobel Prize lies not entirely in the monetary prize itself; hence, we would not be surprised to find that many people might actually be willing to pay to obtain this prize. These observations on human nature are far from new. Hobbes, for example, wrote that: ‘Men are continuously in competition for honour and dignity’ (cited in Hirschman, 1973, p. 634). While the focus of traditional economics has been on the monetary rewards that are exchanged through a market mechanism, sociologists have stressed social status and other social rewards as important motivations for human behaviour. The term ‘social status’ was first introduced by Max Weber as ‘an effective claim to social esteem in terms of negative or positive privileges’ (Weber, 1922, p. 305).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

One of the important factors that distinguishes societies and determines their success is the form of the incentives they provide to their members. Generally speaking, there are three broad types of such incentives: (*a*) private monetary rewards, such as wages and profits; (*b*) social rewards including status and prestige; and (*c*) rules, laws and regulations that enforce certain types of behaviour while penalizing others. Societies may differ in the mix of incentives and rules that they employ. This mixture has a significant effect on economic performance. Social status is thus part of the incentive structure provided to individuals in every society. These incentives affect an individual's choice of actions, occupation, education level and so forth. They thus are of significant economic importance (see also the review in Weiss and Fershtman, 1998).

If prizes, knighthoods and other status symbols were up for sale like any other commodity, their value would be deflated. The value of a medal, award or title will be small if it were obtainable by many people. Thus, the value of status symbols depends on the allocation rule that determines who is eligible to receive a particular symbol, who is excluded from eligibility, and the number (and certainly the identity) of its recipients. This property distinguishes social status from economic rewards. Giving a medal to one individual (may) reduce the medal's value for another individual. Social status may thus be viewed as the ranking of individuals, or groups of individuals, in society. This ranking may be based on personal attributes, actions, occupations or group affiliations. Yet, by definition, if someone climbs up in rank, someone else climbs down.

Ranking matters only if people agree on how ranking is established. For social status to matter, a society must generally agree about the relative position of its members. The crucial feature of social status is that it ‘rests on collective judgement, or rather a consensus of opinion within a group. No one person can by himself confer status on another, and if a man's social position were assessed differently by everybody he met, he would have no social status at all’ (Marshall, 1977, p. 198). An interesting – and relatively unexplored – issue is the role of social status in a multicultural society where every group maintains its own ranking, each of which may be affected by different characteristics. Fershtman, Hvide and Weiss (2006) have shown that the gains made from (social) trade in a culturally diverse society can be translated into higher output and wages.

The categories comprising social status are diverse. We can distinguish between ‘status group’ and ‘individual status’. In the first category, originally conceived by Weber, social status is obtained by affiliation with a group, be it a social class, profession, club, and so on. Members of the status group share a similar status. In the second category, social status is obtained through individual attributes or actions. One should also distinguish between status that is acquired through specific actions or group affiliation and status that has been inherited. People may have social status simply by being born into an aristocratic class. The specific structure for gaining and maintaining status thus plays an important role in determining its effect on the economy.

One of the major problems in incorporating social status into economic models is that it is not directly observed. How do we measure status? How do we identify the ranking that determines who is perceived as important and who is not? And how do we quantify this ranking? Another task necessary for modelling is identification of the variables that affect social status. What determines social status in different societies? Conducting surveys by asking people to rank occupations according to their ‘prestige’ or to state which of the individual's attributes affects his or her social status has been the accepted method for responding to these questions. Treiman (1977), for example, found that the ranking of occupational status is stable over time and similar in different societies. Moreover, status rank has

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
been found to be systematically dependent on occupational attributes; that is, occupations requiring high levels of education and providing high income also confer high social status.

Social status and consumption

Individuals may use consumption choices to signal that they have properties that affect their social status. The most familiar form of such signalling is ‘conspicuous consumption’, a signal relevant whenever relative wealth is a factor in determining social status. Yet the quest for social status may affect other consumption choices as well. For instance, individuals may buy and display books or go to the theatre to signal the level of education they have obtained. They may join clubs, buy a house or hire a maid if such actions signal their desired status.

The concept of ‘conspicuous consumption’ was first introduced by Veblen (1899), who argued that individuals often consume highly attention-getting goods and services in order to signal their wealth and thereby achieve greater social status. ‘In order to hold the esteem of men it is not sufficient merely to possess wealth or power. The wealth and power must be put in evidence, for esteem is awarded only on evidence’ (Veblen, 1899, p. 36). The extreme form of such behaviour is known as the ‘Veblen effect’, witnessed whenever individuals are willing to pay higher prices for functionally equivalent goods (for a discussion, see also Leibenstein, 1950, and Frank, 1985a, 1985b). The Veblen effect may indeed be empirically significant in some luxury good markets (see Creedy and Slottje, 1991; Heffetz, 2004). Veblen distinguished between (*a*) ‘invidious comparison’ – whenever an individual from a higher class consumes conspicuously to distinguish himself from an individual from a lower class, and (*b*) ‘pecuniary emulation’ – whenever an individual from a lower class consumes conspicuously to imitate a member of the upper class. Bagwell and Bernheim (1996) used a signalling model to investigate the conditions under which the Veblen effect may result from the desire to signal wealth (see also Ireland, 1994). Conspicuous consumption may lead to excessive consumption and suboptimal saving but this conclusion depends on the specific details and timing of the ‘conspicuous consumption race’. As Corneo and Jeanne (1996) have shown, if the signalling is typically done late in the life cycle, conspicuous consumption may actually encourage saving.

Letting social status be determined by relative wealth may help to explain some of the puzzles we observe in human behaviour. The empirical evidence indicates that saving continues in old age and hardly declines with wealth. These observations imply that saving behaviour cannot be explained solely by consumption motives. Status concerns derived from relative wealth may provide some explanation of this phenomenon.

It is important to note that the striving for social status is not the only explanation for conspicuous consumption. Such behaviour may also signal professional success or ability. Simply think about a situation in which you need to choose a lawyer without any knowledge about the candidates’ ability. Often a lawyer's dress, the car she drives or how her office is decorated may affect your choice whenever these items may signal ability or success.

Status and the labour market

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Some professions enjoy higher status than others. For example, in most countries, being a physician yields a higher status than being a butcher. In such cases, simply belonging to a profession, and not the individual's characteristics, is rewarded by the relevant professional status. Consequently, individuals choose their occupations not just according to the wages they will be paid but also according to the status associated with that occupation. However, at equilibrium, wages are also affected by the quest for status. Adam Smith, in *The Wealth of Nations* (1776), was the first to state the argument of compensating wage differences by pointing out that: ‘Honour makes a great part of the rewards of all honourable professions. …The most detestable of all employment, that of a public executioner, is, in proportion to the quantity of work done, better paid than any common trade whatever’ (Book I, ch. X). Still, no empirical evidence has been found supporting the phenomenon of high status being associated with low wages, because to do so one would need to control for ability, which is not directly observable. Status concerns may also explain some degree of wage rigidity. Unemployed individuals are often reluctant to accept temporary but low-paid jobs because doing so implies a loss of status (see Blinder, 1988). On the other hand, immigrants are less reluctant to apply for low-status jobs; they tend to stress wages over occupational status partly because their reference group is new immigrants and not the wider society.

The workplace itself is a venue for social interaction. Relative wages may be important in forming ‘local’ status at the firm level. The willingness of workers to exert effort may be affected by social rewards. When workers are concerned with their ‘local’ status, wage inequality across firms will tend to exceed wage inequality within firms. Within each firm, productivity differentials will exceed wage differentials as some reward is derived from higher status (see Frank 1984a, 1984b; for empirical and experimental evidence for the relevance of wage comparisons, see Clark and Oswald, 1996, and Zizzo and Oswald, 2001).

Social status and growth

The great variability in growth across different economies is a major puzzle for economists. While most of the literature offers an economic explanation for this phenomenon, others claim that some of the variation can be attributed to cultural factors. Social status affects growth primarily by affecting individuals’ choices of occupation, investment and education. For example, it has been argued that contempt for entrepreneurs and the high status of the idle gentleman in 19th-century England were the main causes for its economic decline during that period (see Wiener, 1981).

A common feature of recent growth models is the existence of externalities associated with human capital or certain occupations. Each worker, when choosing his level of schooling or occupation, ignores the impact of his choice on overall economic performance. Whenever social status is attached to such activities or occupations, it can be perceived as a corrective mechanism. Baumol (1990) emphasizes the role of social status or social prestige associated with ‘non-productive’ (rent-seeking) activities versus ‘productive’ activities. The implications are simple: a status structure that awards higher status to ‘productive’ activities is conducive to growth. Fershtman and Weiss (1993) used a simple general equilibrium model in which wages and social status are determined endogenously to show that changes in the demand for status, triggered by changes in preferences, may affect growth rates.

But status has a collective nature and may be determined endogenously by the type of people who

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
choose each occupation or profession. The drive for status may thus be counter-productive and induce an inefficient allocation of talent among the different occupations. A large emphasis on status may encourage the 'wrong' individuals – those with low ability and great wealth – to choose a 'productive' occupation or acquire schooling, thereby forcing workers with high ability but little wealth to leave growth-enhancing occupations. This crowding-out effect may, by itself, discourage growth (see Fershtman, Murphy and Weiss, 1996).

Social status as a corrective mechanism

It has long been recognized that activities that generate externalities but cannot be priced are not efficiently regulated by private rewards. It was Arrow (1971) who initially suggested the role of social rewards as a mechanism designed to resolve the inefficiencies arising from externalities (see Elster, 1989, for a critical view for this approach). According to Arrow, an individual who chooses an action or occupation that produces positive externalities is appreciated by other members of the society and obtains social status, whereas an individual who produces negative externalities is treated with contempt (or a negative social status). The use of such a social mechanism is appealing as it implies that the problem of market inefficiencies due to externalities can be resolved or diminished. On the other hand, the use of social rewards is limited in itself and as a corrective mechanism. As mentioned previously, a profusion of medals reduces their value – a property that limits the scope of their use.

Social status and the evolution of preferences

While most economists are sympathetic to the idea that the concern for social status is an important aspect of human decision-making, they remain reluctant to incorporate this variable into mainstream economic modelling. The ruling paradigm of *homo economicus* is that of an individual whose utility depends on his consumption bundle and who makes employment decisions based on the wages to be received for performing a particular job. For sociologists, the dominant paradigm is that of people who, as social animals, wish to maximize their standing in society.

The reluctance to incorporate status concerns into the utility function is rooted in the assumption that models including this variable often allow too broad a range of behaviour and thus ultimately display little predictive power (for more on this view see Postlewaite, 1998). The debate centres on whether status concerns are a 'direct effect', reflecting the fact the people are (also) social animals, or an 'indirect effect', meaning that people care about social status because status affects the goods and services that they and their children will consume (for an illustration of the indirect approach, see Cole, Mailath and Postlewaite, 1992).

Incorporating preferences for status as 'hard-wired' into the utility function raises the question of why people (or other animals) have ingrained preferences for social status. One approach to dealing with the issue, common in economics, is not to deal with the formation of preferences. Social biologists have adopted a different approach. They argue that feelings and social concerns are hard-wired in human actors; moreover, concerns that increase fitness tend to be more common. Fershtman and Weiss (1998a; 1998b) applied an evolutionary approach and showed that caring about social status can be part of stable evolutionary preferences.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- social norms

Bibliography

Arrow, K. 1971. Political and economic evaluation of social effects and externalities. In *Frontier of Quantitative Economics*, ed. M. Intriligator. Amsterdam: North-Holland.

Bagwell-Simon, L. and Bernheim, D. 1996. Veblen effects in a theory of conspicuous consumption. *American Economic Review* 86, 349–73.

Baumol, W. 1990. Entrepreneurship: productive, unproductive and destructive. *Journal of Political Economy* 98, 893–921.

Blinder, A. 1988. The challenge of high unemployment. *American Economic Review* 78, 1–15.

Clark, A. and Oswald, A. 1996. Satisfaction and comparison income. *Journal of Public Economics* 61, 359–81.

Cole, H., Mailath, G. and Postlewaite, A. 1992. Social norms, saving behaviour and growth. *Journal of Political Economy* 100, 1092–125.

Corneo, G. and Jeanne, O. 1996. Social organization, status and saving behaviour. *Journal of Public Economics* 70, 37–52.

Creedy, J. and Slottje, D. 1991. Conspicuous consumption in Australia. Working Paper No. 307. Department of Economics, University of Melbourne.

Elster, J. 1989. Social norms and economic theory. *Journal of Economic Perspectives* 3(4), 99–117.

Fershtman, C., Hvide, H. and Weiss, Y. 2006. Cultural diversity, status concerns and the organization of work. *Research in Labor Economics: The Economics of Immigration and Social Diversity*, ed. S. Polachek, C. Chiswick and H. Rapoport. Amsterdam: Elsevier, JAI Press.

Fershtman, C., Murphy, K. and Weiss, Y. 1996. Social Status, Education and Growth. *Journal of Political Economy* 104, 108–32.

Fershtman, C. and Weiss, Y. 1993. Social status, culture and economic performance. *Economic Journal* 103, 946–59.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fershtman, C. and Weiss, Y. 1998a. Why do we care what others think about us? In *Economics, Values and Organization*, ed. A. Ben Ner and L. Putterman. Cambridge: Cambridge University Press.

Fershtman, C. and Weiss, Y. 1998b. Social rewards, externalities and stable preferences. *Journal of Public Economics* 70, 53–73.

Frank, R. 1984a. Interdependent preferences and the competitive wage structure. *Rand Journal of Economics* 15, 510–20.

Frank, R. 1984b. Are workers paid their marginal products? *American Economic Review* 74, 549–70.

Frank, R. 1985a. *Choosing the Right Pond: Human Behavior and the Quest for Status*. London and New York: Oxford University Press.

Frank, R. 1985b. The demand for unobservable and other nonpositional goods. *American Economic Review* 75, 101–16.

Heffetz, O. 2004. Conspicuous consumption and the visibility of consumers' expenditures. Working paper. Center for Health and Wellbeing, Princeton University.

Hirschman, A. 1973. An alternative explanation of contemporary harriedness. *Quarterly Journal of Economics* 87, 634–7.

Ireland, N. 1994. On limiting the market for status signals. *Journal of Public Economics* 53, 91–110.

Leibenstein, H. 1950. Band wagon, snob, and Veblen effects in the theory of consumers demand. *Quarterly Journal of Economics* 622, 183–207.

Marshall, T. 1977. *Class, Citizenship and Social Development: Essays*. Chicago: University of Chicago Press.

Postlewaite, A. 1998. The Social Basis of Interdependent Preferences. *European Economic Review* 42, 779–800.

Smith, A. 1776. *The Wealth of Nations*, Reprinted, New York: Modern Library, 1934.

Treiman, D. 1977. *Occupational prestige in comparative perspective*. New York: Academic Press.

Veblen, T. 1899. *The theory of the leisure class: An economic study of institutions*. London: Unwin Books, 1994.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Weber, M. 1922. *Economy and Society*. Berkeley: University of California Press, 1978.

Wiener, M. 1981. *English Culture and the Decline of the Industrial Spirit, 1850–1980*. Cambridge: Cambridge University Press.

Weiss, Y. and Fershtman, C. 1998. Social status and economic performance: a survey. *European Economic Review* 42, 801–20.

Zizzo, D. and Oswald, A. 2001. Are people willing to reduce others' income? *Annales d'économie et de statistice*, July/December, 39–65.

How to cite this article

Fershtman, Chaim. "social status, economics and." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000254> doi:10.1057/9780230226203.1566

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

social welfare function

Prasanta K. Pattanaik
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The article deals with the related, though distinct, notions of a social welfare function due to A. Bergson and P. Samuelson on the one hand and K.J. Arrow on the other. After introducing the two formal concepts, it gives a brief outline of Arrow's well-known impossibility theorem, considers some alternative intuitive interpretations of the notion of a social welfare function, and discusses the informational bases of social welfare judgements.

Keywords

aggregation; Arrow's theorem; Bergson–Samuelson social welfare function; cardinal utility; independence of irrelevant alternatives; interpersonal comparison of utilities; majority rule; non-utility information; Pareto principle; preferences; rights; Sen, A.; social choice; social preferences; social welfare function; voting paradoxes

Article

The concept of a social welfare function is central in welfare economics, the branch of economics that explores the implications of various ethical criteria for deciding what promotes social welfare, what public policies the society should choose, and so on. It was first introduced by Bergson (1938), and was subsequently elaborated by Samuelson (1947). A related, though not identical, notion with the same name was introduced by Arrow (1951); as we shall see below, Arrow's interpretation was different from Bergson's earlier concept. The technical literature on the social welfare function is large. The following discussion will, however, focus mainly on conceptual issues relating to the social welfare function.

The basic notation

Consider a given society consisting of individuals 1, 2, …, n, with X a given set of alternative social

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

states. For our purpose, a social state can be thought of as a complete description of all 'relevant' aspects of the state of affairs prevailing in society, but one can also use more limited interpretations of a social state. An element of X can be thought of as a vector, each component of the vector representing some particular feature of the state of society. The elements of X will be denoted by x, y, z, and so forth. Each individual i is assumed to have a preference ordering ('at least as good as') defined over X (an ordering over X is defined to be a binary relation T over X such that T is reflexive (for all x in X, xTx), connected (for all distinct x and y in X, xTy or yTx), and transitive (for all x, y, and z in X, if xTy and yTz, then xTz)). Such a preference ordering will be denoted by R_i, R_i', and so on. An n-tuple of preference orderings over X, with the preference ordering of each individual figuring exactly once in the n-tuple, will be called a preference profile. For all social states x and y, and for every individual i, [xP_iy if and only if xR_iy and not yR_ix] and [xI_iy if and only if xR_iy and yR_ix]. Intuitively, P_i denotes the strict preference relation for individual i and I_i denotes the indifference relation for individual i. R, R', and so forth will denote social weak preference relations ('socially at least as good as') over X. Thus, xRy will denote that x is at least as good as y for the society. Given a social weak preference relation R, one can define P ('socially better than') and I ('socially indifferent to') in terms of R in the same way as P_i and I_i are defined in terms of R_i.

The Bergson – Samuelson social welfare function

The Bergson–Samuelson social welfare function (SWF (BS)) is a function W that specifies exactly one real number, $W(x)$, for each social state x in X. The intended interpretation is this: for all x and y in X, $W(x) \geq W(y)$ denotes that x offers at least as high a level of social welfare as y, and similarly for $W(x) > W(y)$ and $W(x) = W(y)$. $W(x)$, $W(y)$, and so on are thus ordinal indices of social welfare attached to social states. What determines the form of the function? For example, what determines whether x should be assigned a higher number than y or y should be assigned a higher number than x? This, as both Bergson (1938) and Samuelson (1947) pointed out, will depend on the value judgements that we use. It is not surprising that very little can be done with the mathematical notion of an SWF (BS) at this level of generality. Even then, the abstract notion, by itself, makes one thing clear: if one wants to say anything specific about social welfare, one must introduce explicit value judgements.

Specific conclusions emerge from the Bergson–Samuelson social welfare function as one introduces additional value judgements that restrict the form of the function. Thus, assuming that each individual has an (ordinal) utility function U^i defined over X, Samuelson considers an 'individualistic' form for SWF (BS) where the social welfare indices attached to social states depend exclusively on the individual utilities, so that one can write the SWF (BS) as

$$W = F(U^1, \ldots, U^n) \quad \ldots \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Assuming further that social states are simply allocations (a specification of the quantity of each commodity figuring in each consumer's consumption bundle and the quantity of each commodity figuring in each producer's production plan), F is increasing in U^1, U^2, …, and U^n, and each consumer's utility depends only on her consumption bundle, Samuelson (1947, pp. 229–49) derived conditions for the maximization of social welfare subject to the relevant resource constraints and technological constraints of the society.

Arrow's social welfare function

Arrow (1951) introduced a somewhat different concept of a social welfare function (SWF (A)). An SWF (A) is a functional rule G, which, for every possible preference profile, $(R_1,\ldots,R_n)$, belonging to a non-empty class of preference profiles, defines exactly one ordering R over X. Thus, we write

$$R = G(R_1, \ldots, R_n) . \quad \ldots \tag{2}$$

The intuitive interpretation of R figuring in this definition is that it represents the society's weak preference relation over social states, R being constrained to be an ordering. Thus, an SWF (A) gives us a (unique) social ordering of all social states once the individuals' orderings over the social states are given.

What is the relationship between the notion of an SWF (BS) and that of an SWF (A)? Suppose an SWF (BS) has the form given by (1), and suppose the utility functions U^i (i=1,2,…,n) are ordinal so that they have no more significance beyond just the orderings that they, respectively, represent (recall that the Bergson–Samuelson social welfare indices are also ordinal). Then noting that, for every profile of real valued utility functions over X, we have Bergson–Samuelson social welfare indices for social states, such that the ordering implied by the social welfare indices does not change with a change in the profile of utility functions so long as the orderings implied by the utility functions do not change (see Samuelson, 1947, p. 228), we would have a unique social ordering over X for every profile of individual orderings over X and, hence, an Arrow-type social welfare function. Thus, underlying an SWF (BS) as given in (1), there is always an SWF (A). (The converse is not necessarily true since the social ordering R of Arrow may not be representable by a real valued function over X.)

Arrow's impossibility theorem

If Arrow did not impose any restrictions on an SWF (A), then the definition, by itself, would be of no more substantive interest than just the definition of an SWF (BS) without any specific assumptions about the properties of the SWF (BS). Arrow, however, proceeded to introduce specific restrictions regarding the form of an SWF (A). He postulated that an SWF (A), G, must satisfy the following properties.

Universal Domain (UD): the domain of G must include every possible preference profile.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Weak Pareto Criterion (WP): for every profile of individual orderings, $(R_1,\ldots,R_n)$, in the domain of G and for all x and y in X, if xP_iy for every individual i, then xPy.

The weak Pareto criterion, which is a weak version of the familiar Pareto principle, just says that, if every individual strictly prefers social state x to social state y, then the social ordering must rank x higher than y.

Independence of Irrelevant Alternatives (IIA): Consider any two profiles of preference orderings, $(R_1,\ldots,R_n)$ and $(R'_1,\ldots,R'_n)$, in the domain of G and any two social states x and y. If, for every individual i, [xR_iy if and only if xR'_iy] and [yR_ix if and only if yR'_ix], then [xRy if and only if $xR'y$] and [yRx if and only if $yR'x$] where the social orderings R and R' correspond to the preference profiles $(R_1,\ldots,R_n)$ and $(R'_1,\ldots,R'_n)$, respectively.

IIA requires that, if the individual orderings change but everybody's ranking of a pair of social states remains unchanged, then the social ranking of those two social states must remain unchanged though the social ranking over other pairs of social states may change.

Non-dictatorship (ND): there does not exist any individual k such that for all social states x and y and for every profile of individual orderings $(R_1,\ldots,R_n)$ in the domain of G, if xP_ky, then xPy.

ND just says that there should not be any individual such that, whenever she strictly prefers any social state x to any other social state y, x must rank higher than y in the social ordering, irrespective of other people's preferences.

Arrow's (1951) famous impossibility theorem tells us that, if there are at least three social states in X, then there does not exist any SWF (A) that simultaneously satisfies WP, IIA, and ND. The result has the flavour of a paradox since, prima facie, the properties postulated by Arrow for his social welfare function seem plausible. It may be useful to consider two examples to illustrate how Arrow's theorem ‘works’. Consider first the simple majority rule (SMR) which says that, for every preference profile, $(R_1,\ldots,R_n)$, and for all x and y in X, xRy if and only if the number of individuals who consider x to be at least as good as y is greater than or equal to the number of individuals who consider y to be at least as good as x. While the SMR satisfies WP, IIA, and ND, it does not yield a social ordering for every preference profile. Thus, if we have three individuals, 1, 2, and 3, and three alternatives, x, y, and z, then, for the preference profile such that $(xP_1y \;\&\; yP_1z \;\&\; xP_1z)$, $(yP_2z \;\&\; zP_2x \;\&\; yP_2x)$, and $(zP_3x \;\&\; xP_3y \;\&\; zP_3y)$, the SMR gives us $(xPy \;\&\; yPz \;\&\; zPx)$ which is not an ordering (this, in fact is the well known ‘voting paradox’). Let us take a second example, the Borda rule, which for every preference profile specifies the social ordering as follows. On the assumption that X has m elements, if an individual places a social state x in the first position in his preference ordering, then x gets m points from him; if an individual places x in the second position in his preference ordering, then x gets $m-1$ points from him; and so on. (In stating the Borda rule, I have ignored the case where an individual may be indifferent between two social states. For a complete specification of the Borda rule, however, the assignment of points in such cases needs to be specified.) A social state a is considered to be socially at least as good as a social state b under the Borda rule if and only if the sum of all the points received by a from all individuals is greater than or equal to the corresponding sum for b. The Borda rule satisfies all the conditions of Arrow excepting IIA. To see that it violates IIA, let us consider the case where we have two individuals (1 and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

2), three alternatives (x, y, and z), and two preference profiles, (R_1, R_2) and (R_1', R_2') as follows: xP_1y & yP_1z & xP_1z; zP_2y & yP_2x & zP_2x; $xP_1'y$ & $yP_1'z$ & $xP_1'z$; and $zP_2'x$ & $xP_2'y$ & $zP_2'y$. Given the profile (R_1, R_2), each of x and z receives a total of four points, and, hence, we have xIz, but, given the preference profile (R_1', R_2'), x receives a total of five points while z receives a total of four points, and hence, we have $xP'z$. This violates IIA since the social ranking of x and z changes when we go from (R_1, R_2) to (R_1', R_2'), though the ranking of x and z has remained the same for each individual.

An impossibility theorem such as Arrow's (1951) compels us to think what has gone ‘wrong’, which of the requirements are unreasonable, and which of the restrictions need to be discarded or modified to provide a way out of the impasse. In this brief article, I shall not explore these questions, which have been discussed in great detail in the large literature that followed Arrow (1951). Instead, I turn to some basic issues about how one is to interpret the notion of the social welfare function itself.

Alternative intuitive interpretations

Some important questions have been raised by a number of economists (see, for example, Bergson, 1954; Little, 1952; and Sen, 1977a) about the interpretation of a social ranking of social states, such as the social ordering yielded by Arrow's social welfare function and the ordering implied by the welfare indices given to us by a Bergson–Samuelson social welfare function.

It has been claimed by both Little (1952) and Bergson (1954) that the social ordering R figuring in Arrow's definition of a social welfare function is the result of an aggregation procedure or ‘constitution’ that aggregates a given profile of individual preference orderings reflecting the individuals’ judgements or opinions. In contrast, as Bergson pointed out, the welfare indices that come from his social welfare function were intended to reflect a given individual's personal value judgements about what was good for the society (in a somewhat similar fashion, Sen, 1977a, makes a distinction between committee decision and social welfare judgements). Arrow (1963) agreed that he did intend his social welfare function to be a constitution or a rule for aggregating people's opinions, but he claimed that such aggregation was, indeed, the central issue of welfare economics.

An example may be helpful in clarifying the distinction. Suppose someone, say, individual i, says that a complete ban on smoking in all public places is better than a prohibition of smoking only in a few designated public places. Suppose, when asked to give the reason why he thinks so, he gives us as the reason the fact that 99 per cent of the population in the society have the opinion that a complete ban on smoking will be better for society. This may be a good enough reason if i's original statement is a statement about how society should rank the two policies for the purpose of social action, given the existing opinions or judgements of the people. It is, however, possible that individual i made his original statement as his personal judgement about what would promote society's welfare. In that case, we would find his response to the request for justification a little strange: we would feel that he should give ‘independent reasons’ for his statement rather than referring to the judgements of other people.

Typically, in aggregating people's opinions or judgements through a ‘constitution’ or a committee decision procedure, we do not look into the basis of people's opinions; we take them as given and simply try to find out what will be a reasonable way of reconciling different opinions. In contrast, in forming

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
our personal social welfare judgements, typically we do not aggregate individuals' opinions or judgements (in fact, we often question the basis of other people's judgements when they do not conform to our ethical beliefs), though we do take into account other people's well-being. Also, in our personal judgements about social welfare, we often compare the welfare losses or gains of one person with those of another, while, in aggregating opinions or judgements, we rarely take into account the strength with which one individual favours *x* over *y* with the strength with which another person favours *y* over *x* (see Sen, 1977a, p. 159).

Arrow (1963) sees the basic purpose of welfare economics as that of analysing procedures for aggregating individual opinions so as to arrive at social decisions. Therefore, he interprets the social ordering that results from such aggregation as the basis of social action, the aggregation procedure being his social welfare function. Nevertheless, as he pointed out, ordering *R* in the definition of an SWF (A) could also be interpreted as reflecting the social welfare judgement of a given individual, say, *i*. In that case, the SWF (A) would reflect the rule (s) by which *i* derived his personal social welfare judgement regarding social states, given the preferences of the individuals in the society. If, however, we adopt this interpretation of the SWF (A), then it will be appropriate to interpret the preference orderings that constitute the arguments of the SWF (A) as reflecting the individuals' welfares rather than their value judgements or opinions since it is not clear why a person would use other people's judgements to form his own social welfare judgement.

Both the interpretations of a social welfare function would seem to be important for welfare economics conceived in a broad fashion. In some ways, the analysis of personal judgements about social welfare and the aggregation of the opinions of the individuals in society correspond to two distinct phases that can often be discerned in a democratic process. The first stage is the stage of deliberation where people engage in ethical debates about each other's personal social welfare judgements. The second stage is the stage of voting or aggregation of people's judgements, where people's opinions or judgements, as they emerge from the debates and deliberations of the first stage, are taken as given, and attention is focused on arriving at a 'reasonable compromise' on the basis of these judgements (see Little, 1952; and Pattanaik, 2005).

The informational basis of social welfare judgements

Arrow's analytical structure does not permit us to consider certain types of information, which, intuitively, we often regard as important for forming our social welfare judgements. I note two such informational constraints.

Cardinal utility and interpersonal comparisons of utilities: the SWF (A) defines social ordering as a function of the individuals' preference orderings over *X*. Thus, social ordering does not use any cardinal feature of individual welfare, and interpersonal comparisons either of the levels of individual welfare or of individual welfare gains and losses does not play any role in the determination of Arrow's social ordering. The same is also true of the 'individualistic' Bergson–Samuelson social welfare function (see (1)), given Samuelson's assumption that the individual utility functions are all ordinal. Such complete eschewal of cardinal notions of individual welfare and all interpersonal comparisons of individual welfares goes counter to our intuition when one interprets the Arrow's social ordering as reflecting someone's social welfare judgement rather than simply as the result of aggregating opinions through a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
'constitution'. As I have noted earlier, in forming our social welfare judgements, we typically take into account the welfare of individuals. In doing so, we also often resort to interpersonal comparisons of individuals' welfare levels or changes in their welfares. The SWF (A) would not allow us to do this. For example, consider two allocations x=(98, 2) and y=(97, 3) of 100 units of some desired resource between two individuals, 1 and 2. Suppose someone wants to say that a move from x to y, involving a redistribution in favour of 2 will improve social welfare because, at x, 1 has a higher level of utility than 2, and, even after redistribution, 1's welfare continues to be higher than 2's welfare. This justification cannot be given in the Arrow framework since the framework does not permit such interpersonal comparison of welfare levels. Nor can the person justify his social welfare comparison of x and y in the framework by saying that individual 1's utility loss from the switch from x to y is outweighed by 2's gain of utility since Arrow's framework permits neither cardinal individual utilities nor interpersonal comparisons of utility differences. In the literature that followed Arrow (1951), a series of important contributions (see, for example, Harsanyi, 1955; Sen, 1970b, 1977b, 1979; d'Aspremont and Gevers, 1977; and Gevers, 1979) have explored social welfare judgements based on much richer utility information incorporating cardinal and interpersonally comparable individual utilities, and have demonstrated that Arrow-type impossibility results often lose their bite in this expanded analytical structure.

Non-utility information: Sen (1977b) demonstrated that, though the definition of an SWF (A), by itself, does not rule out the possibility of using non-utility information, such as the information contained in the description of social states, in making social welfare judgements, a somewhat stronger version of WP, together with IIA, does rule out that possibility. The stronger version requires that, for all social states x and y and for every preference profile $(R_1,\ldots,R_n)$, [if $x\ I_i\ y$ for all individuals i , then $x\ I\ y$], and, further, [if xR_iy for all individuals i and xP_iy for some individual i, then xPy].

Individual rights based on the notion that an individual should be able to make free choices in affairs relating to his or her private life is an important example of an ethical value based on non-utility information. The concept of an individual's private life, which John Stuart Mill (1859) considered so important, cannot, however, be articulated in terms of individual utilities alone. While i's religion may cause just as much disutility for his neighbours as his playing loud music in early hours of the morning, Mill (1859) would have considered i's religion, but not his playing loud music in early hours of the morning, to be an aspect of i's private life. Sen (1970a) investigated the implications of granting individuals the right to make free choices with respect to their private affairs irrespective of how others feel about their choices. In his celebrated result on the impossibility of the Paretian liberal, Sen (1970a) demonstrated that respect for such individual rights clashes sharply with WP, even if one discards IIA and replaces the Arrow requirement that the social weak preference relation R be an ordering by the much weaker requirement that R be reflexive and connected and P be acyclic (P is said to be acyclic if and only if there do not exist $x_1,x_2,\ldots,x_n$ in X such that x_1Px_2 & x_2Px_3 & … & $x_{n-1}Px_n$ & x_nPx_1). While Sen departed radically from the Arrow format by introducing individual rights that have non-utility information as their basis, he still retained one basic feature of the Arrow format. His analysis was in terms of a social weak preference relation specified by a function of individual orderings over the social states. Sen's formulation of an individual's right to make free choices in his own private life was introduced as a restriction on this function, the restriction being contingent on the individual's preferences over social states that differed only with respect to some features of his private life. An

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
alternative version of Sen's theorem uses the notion of social choice rather than social preference. The point under consideration applies to that version as well. Given any set of feasible social states, social choice from that set is still a function of individual preference orderings over social states, and the rights of an individual are formulated as restrictions on social choices, the restrictions being contingent on the individual's preferences over certain social states.

Several subsequent writers (see, for example, Sugden, 1985; and Gaertner, Pattanaik, and Suzumura, 1992), who argued for a formulation of individual rights in terms of game forms, had to abandon Sen's format altogether, given their conception of an individual's right as the individual's freedom to choose any of the actions or strategies permissible under the right rather than in terms of constraints imposed on social weak preferences (or social choice) by the individual's preferences over certain types of social states. (See, however, Pattanaik and Suzumura, 1996, for an attempt to put the problem of social choice of a rights structure, viewed as a game form, back in the framework of an Arrow-type social welfare function.)

To conclude, it will perhaps be fair to say that, while the concept of a social welfare function has been a powerful analytical tool for investigating implications of value judgements relating to social welfare, the individualistic version (see (1)) of the Bergson–Samuelson formulation, as well as Arrow's formulation of the concept, had certain limitations and that some important developments in welfare economics have their origin in attempts to overcome those limitations.

See Also

- Arrow, Kenneth Joseph
- Arrow's theorem
- Bergson, Abram

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Arrow, K.J. 1963. *Social Choice and Individual Values*, 2nd edn. New Haven: Yale University Press.

Bergson, A. 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52, 310–34.

Bergson, A. 1954. On the concept of social welfare. *Quarterly Journal of Economics* 68, 232–52.

d’ Aspremont, C. and Gevers, L. 1977. Equity and informational basis of collective choice. *Review of Economic Studies* 46, 199–210.

Gaertner, W., Pattanaik, P.K. and Suzumura, K. 1992. Individual rights revisited. *Economica* 59, 161–77.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gevers, L. 1979. On interpersonal comparability and social welfare orderings. *Econometrica* 47, 75–90.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 77–94.

Little, I.M.D. 1952. Social choice and individual values. *Journal of Political Economy* 60, 422–32.

Mill, J.S. 1859. *On Liberty*. New York: Liberal Press, 1956.

Pattanaik, P.K. 2005. Little and Bergson on Arrow's concept of social welfare. *Social Choice and Welfare* 25, 369–79.

Pattanaik, P.K. and Suzumura, K. 1996. Individual rights and social evaluation: a conceptual framework. *Oxford Economic Papers* 48, 194–212.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sen, A.K. 1970a. The impossibility of a Paretian liberal. *Journal of Political Economy* 78, 152–7.

Sen, A.K. 1970b. Interpersonal aggregation and partial comparability. *Econometrica* 38, 393–409.

Sen, A.K. 1977a. Social choice theory: a re-examination. *Econometrica* 45, 53–89.

Sen, A.K. 1977b. On weights and measures: informational constraints in social welfare analysis. *Econometrica* 45, 1539–72.

Sen, A.K. 1979. Interpersonal comparisons of welfare. In *Economic and Human Welfare: Essays in Honour of Tiber Scitovsky*, ed. M. Boskin. New York: Academic Press.

Sugden, R. 1985. Liberty, preference, and choice. *Economics and Philosophy* 1, 213–29.

How to cite this article

Pattanaik, Prasanta K. "social welfare function." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000172> doi:10.1057/9780230226203.1567

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

socialism (new perspectives)

J.E. Roemer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In the Marxian tradition socialism means the end of expropriation of surplus value from labour by capitalists. If markets are needed more to coordinate economic activity than to provide incentives to workers and managers, socialism could be achieved by more equal distribution of share ownership with restrictions on cashing in shares. Social homogeneity may be a necessary condition for the democratic implementation of egalitarian programmes through redistribution, if the welfare state is motivated by either a purely redistributive or an insurance function. Hence the challenge to implementing socialism posed by multiculturalism.

Keywords

asymmetric information; capital; capitalism, people's; central planning; coordination; corporate raider; democracy; economic man; equality; exploitation; family; and fertility; feudalism; Gini coefficient; Hayek, F.; historical materialism; social homogeneity; immigration; incentives; Keynes, J.; labour contracts; labour power; Lange, O.; market socialism; Marx; multiculturalism; prices; principal–agent problem; private ownership; redistribution; Schumpeter, J.; share ownership; social democracy; social insurance; socialism; social wage; solidarity; Soviet Union; state property; surplus labour; syndicalism; unemployment; women's emancipation

Article

Marxian theory

In the Marxian theory of historical materialism, the ruling class in each mode of production has its special method for extracting the economic surplus from the direct producers; that method follows from the characteristic property relations under the mode. Under the slave mode, the surplus produced by slaves is forcibly appropriated by the slave owner; under feudalism, the lord extracts surplus serf labour

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
through the corvée and various forms of taxation. Capitalism, Marx argued, was the first mode in which surplus extraction was not obviously coercive: no capitalist owns his workers or forcibly takes their product. Indeed, under capitalism workers and capitalists form contracts in which labour power is exchanged for a wage. The capitalist keeps the product of the worker's labour.

Indeed, Marx wished to explain capitalist surplus extraction as a process that would emerge under competitive contracting in which workers and capitalists bargain; in the end, competitive markets set the terms of labour exchange. (As Makowski and Ostroy, 1993, have written, prices are what appear after the dust of the competitive brawl has cleared. It is incorrect to think of prices as *directing* trade; rather, bargaining among many pairs of individuals reaches an equilibrium summarized by a price.)

Why is it that capitalists end up getting the better part of the deal – that is, why do they end up with the surplus while the worker ends up with his wage, which in the Marxian view was only enough for him to subsist upon? The answer lies not in the fact that the capitalist is more clever or has the police on his side; it is that capital is scarce relative to the available supply of labour, and workers must bid for the right to use that scarce capital, which provides them with a wage. Were labour scarce, then capital would have to bid for labour, and profits would be bid down to a minimal level at which capitalists were indifferent between continuing to own capital and becoming workers. *Why* capitalism seems to have been characterized, throughout its history, as a situation of capital scarcity is not fully understood. Marx argued that capitalists as a class, perhaps represented by the state, undertook strategies to guarantee a ‘reserve army of the unemployed’ in order to maintain the imbalance. Indeed, the proletarianization of the agricultural periphery is an important process by which labour abundance has been maintained until the present (see Rosa Luxemburg, 1913). Keynes and Schumpeter envisaged a time when capital would cease to be scarce, bringing about the euthanasia of the capitalist class.

Thus, the fundamental source of the accumulation of wealth in the hands of a small class, through profits created in production, is the fact that abundant workers must bid for the ‘privilege’ of using their labour power on privately owned productive assets that increase its productivity immensely. This provides them with a wage greater than they could have earned in the non-capitalist sector (back on the family farm, so to speak, or selling apples from a street cart), and also produces an additional amount that, according to the labour–capitalist bargain, belongs to the capitalist. Capitalists consume a part of this surplus product, and invest the rest in other profit-making activities.

Some writers have argued that capitalism *is* a system that extracts the surplus from workers coercively; they point to the struggles between workers and bosses at the point of production. It is, I believe, important to point out that capitalist accumulation could transpire, in principle, if capitalists were competitive and if coercion at the point of production of the worker by the capitalist and his agents did not occur. That coercion, upon which many have focused as a central evil of capitalism, exists only because labour contracts are incomplete and not costlessly enforceable. Imagine that the worker and capitalist could contract about every eventuality that might occur during production. If, in addition, the contract were costlessly enforceable (imagine an omnipotent arbitrator who is to hand to deal with any disagreement), then there would be no petty coercion at the point of production: capitalists would not try to speed up assembly lines, force workers to work overtime, cheat them of their wages, discipline them in demeaning ways, and so on. I believe that Marx thought that the essence of capitalism was the accumulation of capital *even* under such conditions. That actual capitalism is not perfectly competitive, that contracts are incomplete, and that capitalists and workers will haggle over who is to do what when a situation not described in the contract comes up, is something which makes capitalism more unpleasant

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
than the ideal type would be, but is not of its essence.

Marx (1867) believed that the property relations of each mode of production would last only so long as they succeeded in inducing production in an efficient way. ‘The water mill gives us the feudal lord, the steam engine, the industrial capitalist.’ He believed that eventually productive forces would develop to such an extent that the capitalist mode of extracting economic surplus would no longer be effective. Indeed, the next stage in economic history, Marx conjectured, would be socialism, a period in which the means of production were collectively owned and the economic surplus thereby became the property of the workers.

We must define exploitation in the Marxian sense. Marx said that workers were exploited because the labour required to produce the goods they could purchase with their wages – including the labour needed to reproduce the capital stock used up in that production – was smaller in quantity that the labour expended by workers for which they received those wage goods. The ‘surplus’ labour, the difference between these two quantities, became embodied in goods which, according to the contract, are owned by capitalists and which they sell for profit. Why does the worker put up with this situation? Because he has no access to the means of production; the surplus labour he supplies is, so to speak, the rent he pays the capitalist for access to those means.

Exploitation is defined as the fact that workers labour for more hours than are ‘embodied’ in the goods they receive as the real wage. Note that, although Marx insisted the wage was one of subsistence, this is entirely unnecessary for the argument. All that must be the case, for exploitation to exist, is that the hours of labour embodied in goods which wages purchase are fewer than the hours worked by workers. Marx viewed socialism as the system that would end capitalist exploitation in this sense. There are, however, at least prime facie, several ways of ending exploitation short of collectivizing ownership in the means of production, and hence in collectivizing the product they produce above what workers receive as wages. One is syndicalism, in which groups of workers own their factories collectively; another is people's capitalism, a system in which firms are privately owned by citizens, each of whom owns a small share of all firms. Syndicalism would quickly generate a system with highly unequal ownership of productive assets, so some groups of workers would ‘exploit’ others through trade, not to speak of hiring contract labourers. Designing a people's capitalism in which Marxian exploitation was eliminated is possible in the abstract, but it would be difficult to implement in actuality. One should note that the distribution of shares in firms to citizens that would abolish exploitation could not be an equal one. Consider the situation of a person who does not work out of choice (a ‘surfer’) but collects dividends: he would be exploiting others, in the Marxian sense, because the amount of labour embodied in the goods he can purchase with his income is greater than the labour he expends. Indeed, for exploitation to be abolished (in the sense that the labour accounts balance) those who choose not to work should receive zero shares of the capital stock. It is in fact possible to design a system of share ownership so that, when individuals choose their labour supply to maximize their preferences over labour and income, the income they receive from wages *and their dividends* is precisely enough to purchase goods embodying *exactly the labour they expended* and, in addition, the allocation of labour and goods is Pareto-efficient. This arrangement, called the proportional solution by Roemer and Silvestre (1993), solves an interesting intellectual problem, but it has little importance as a way of solving the problem of capitalist exploitation because of the difficulty of actually computing the shares of firms citizens should receive, when information about preferences is asymmetric. (The proportional solution, however, may be used by a small community – for example, of fishermen – who collectively

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
own a resource, such as a lake, and wish to exploit it efficiently, avoiding congestion and overuse.) Moreover, as we shall see below, it is not necessarily ethically desirable if workers are significantly heterogeneous in skill.

Socialism, then, became identified with collectivization of the means of production. Workers would produce more goods than they consumed (nobody claims that investment should be zero under socialism), but the existence of a surplus product would not constitute exploitation because it would be owned by all. This presumably meant that the state, which represented the working class, would decide upon its use. Whether this would be the case because workers obtain the suffrage and vote in a party to represent them, or because a party proclaiming itself to represent the working class takes power by non-democratic means, is another question.

A terminological point is in order. Some advocates of socialism define it as a system in which everyone reaches his full potential, racism and sexism vanish, and citizens view each other as brothers. This is a mistake. To be true to the theory of historical materialism, socialism should be defined as a nexus of property relations that eliminates capitalist exploitation. Whether such a system possesses other nice characteristics in consequence is a scientific question, one that cannot be settled by definition.

A special word must be said about equality. If workers are highly heterogeneous in skills, eliminating capitalist exploitation does not eliminate inequality in incomes. Nevertheless, there has been a tradition of viewing socialism as a system of quite equal incomes. This is partly due to the level of abstraction of Marx's thinking, in which he often viewed capitalism as characterized by a mass of homogeneous workers struggling against a small elite of homogeneous capitalists. It is, however, also due to the belief that many of the inequalities in workers' skills come from unequal opportunities fostered by capitalism and, were capitalism to be eliminated, workers would therefore *become* more equal in skills. I believe this view of what socialist transformation, in the sense of Marx, would accomplish is too optimistic – on which more below, when we return to the conception of socialism as equality.

Social democracy

In the event, the world saw two major kinds of socialist experiment: one, initiated by the Bolshevik revolution, was brought to power by a communist party which ruled undemocratically, and shunned the use of markets, which, it feared, would bring with them the old capitalist mentality, whereby traders tried to accumulate capital and hence to exploit others. The other was social democracy, in which parties representing workers won state power through democratic means, and attempted to tax profits for the purpose of investment and augmenting workers' consumption (the so-called social wage). The social democratic path did not as a principle abolish private ownership of capital assets, although some firms were nationalized.

In principle, both of these techniques could abolish the kind of exploitation associated with capitalism. If communist parties were perfect agents of their collective principal, the working masses, they could set the rate of accumulation at that level desired by workers (there is a problem here of how to aggregate workers' disparate preferences over that rate), and then invest the surplus in the way that would best meet the interests of workers (another preference aggregation problem). And under social democracy, private capital could be taxed at a rate sufficiently high that, although rates of exploitation would not be zero, they would be small. To keep capital from fleeing to more profitable venues, under that situation, workers would have to be sufficiently skilled so that, even under such a regime, capitalists' profits

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
would be sufficiently great. Thus, guaranteeing highly skilled labour would appear to be a part of the social-democratic formula if capital is freely mobile.

With regard to equalizing the distribution of income, both the Soviet-type economies (the Soviet Union and eastern Europe) and the Nordic social democracies did an excellent job. (Indeed, at least in the Soviet Union, it is arguable that skilled workers contributed more labour, in efficiency units, then they received back in goods.) The major difference was that the Soviet economies equalized incomes at low levels, while the social democracies did so at high levels. To what was due the failure of the Soviet economies? We still do not have a completely satisfying explanation, but it seems as if their abrogation of markets was an important factor.

Although the Soviet-type economies used markets from time to time, beginning with Lenin's introduction of the New Economic Policy in the 1920s, they were never allowed to operate with the kind of freedom that would have fostered technological innovation, and by the 1960s it was the lack of innovation that was largely responsible for the low level of living standards. (Of course, when the state acted to concentrate talent in one sector, such as the space industry, it was able to achieve impressive results, but the Soviet economy never succeeded in fostering innovation across the board.) These problems were foreseen much earlier, however, in the debate around market socialism that began in the 1930s, with Oskar Lange's argument that markets could in large part replace central planning in a socialist economy. Lange (1936) proposed that central planners announce to industrial managers prices for their inputs and outputs, and require the managers to report back with the amounts of inputs they would demand and outputs they would produce at those prices, if they were to equate the prices to marginal costs (a necessary condition for Pareto efficiency). The planners would then sum up, observe the discrepancies in the supply and demand of each commodity, announce a second set of prices, raising those for goods in excess demand and lowering those for goods in excess supply, and go through the exercise again, hoping to eliminate the imbalances. Lange believed this process would converge rapidly to an equilibrium; then the planners would post the equilibrium prices and instruct firms to produce accordingly.

Lange also suggested that each household receive a certain fraction of the firms' profits, perhaps allocated according to family size.

Lange did not deal properly with the demands of consumers. But, even if we assume that these could be incorporated into the scheme, what is the point of his kind of planning? Why not simply let the market run autonomously? Lange had no convincing answer: he did say that the central planning bureau (CPB) would be able to achieve equilibrium much faster than the market, avoiding the disequilibrium phase that he considered to be socially costly. (Today, this seems to be a quaint view, given the millions of commodities that are produced in a complex economy. Indeed, economic theory still has no full explanation of how the market 'finds' the equilibrium, and there are theorems that the kind of 'tâtonnement' Lange proposed would, with high probability, not converge.) Perhaps Lange thought that the CPB would control economic activity through setting interest rates of various kinds, thus directing firms to invest in the directions the planners desired.

Friedrich Hayek (1940), however, offered a critique of another type. He wrote that it was an illusion to believe that managers could respond with their input demands, facing prices announced by planners, because they *did not know* their production functions, and therefore could not compute marginal costs. Capitalist firm managers, he said, learn how much they can produce with given inputs by the discipline of competition. It is the competitive brawl that teaches managers how to cut costs and produce

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
efficiently, and to suppose that managers would know how to do so in the sterilized situation envisaged by Lange was wrong. Indeed, how would the CPB deal with innovation, with new commodities? The secret of real markets, Hayek argued, was that they provide incentives and a mechanism for people (entrepreneurs) with local information about needs and production possibilities to realize their ideas. Thus fixing the set of managers *ex ante* was already dooming the system to conservatism and inefficiency.

It is interesting to note that Hayek never wrote that socialist managers would be opportunistic or self-serving – that they would lie to the CPB in order to influence their allotment of inputs. Hayek postulated that managers were 'loyal and capable'. This is in sharp contrast to the critique of socialist management that emerged after 1970 among Western capitalist economists, when the principal–agent problem was formulated, and 'shirking' and opportunism became central issues.

Markets, incentives and coordination

Indeed, this raises a critical question about the failure of centrally planned socialism: was it due to *lack of incentives* or to *lack of coordination*? Markets perform two functions: they provide incentives for workers and entrepreneurs to improve their skills and discover new commodities so as to increase their income, but they also coordinate economic activity. It may not be simple theoretically to distinguish precisely these two functions, but they are clearly different. Matching of workers to firms, for example, occurs in large part by observing wage offers; firms shop for inputs by observing price offers. Of course, the system does not work perfectly, but there is doubtless a strong element of coordination engendered by a competitive price system. (Price systems do not coordinate some things properly, such as control of externalities and the supplies of public goods, and therein lies a major liberal justification of state intervention.)

The history of the Soviet economy is replete with stories of poor incentives *and* poor coordination: we do not have a complete account of the relative importance of these two failures in the lacklustre performance of centrally planned economies in their late period. One also reads, however, of how hard Soviet workers worked, and how ingenious they were at making do with poor inputs (see, for instance, Burawoy and Lukacs, 1985). I believe it is important to answer the question posed above, for upon it may rest the possibility of a future for socialism.

Suppose markets are needed mainly to generate incentives to work hard, to form skills, to invent, and so on. This implies that it will be difficult to use markets *and* to redistribute income in a relatively equal manner through taxation. After all, if workers form skills in order to increase their incomes, but then their incomes are taxed away, why form skills? But suppose that markets are needed mainly to coordinate economic activity: then in principle wage income (which would adjust competitively to reflect marginal value products) could be taxed to produce an income distribution of equality without harming production. In the second case, workers would form skills and innovate because they enjoyed doing so, or felt valued for their social contributions.

I suspect the coordination problem is relatively more important in the failure of centrally planned economies, and the incentive problem relatively less important, than most currently believe. Many economists, especially, assume that the opportunist kind of behaviour so prevalent in the theory of *homo oeconomicus* is a deep aspect of human nature, and therefore that it must have been rampant in the Soviet Union.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Markets are essential in any complex economy, at least for coordination and perhaps for incentives. But, as we have seen in the Nordic countries, tremendous accomplishments with respect to income distribution can be achieved with taxation and ‘wage solidarity’. One might say that the future of socialism lies in emulating the Nordic social democracies. They may, however, not be easy to emulate, as the solidarity of their citizenries may be due to their homogeneity – linguistic, religious, and ethnic. Perhaps welfare states of that magnitude cannot be achieved in highly heterogeneous societies.

A future for socialism may still, then, require an alternative to conventional private ownership of firms with significant redistribution through taxation, because the solidarity necessary for democratic approval of that degree of redistribution may not evolve in large heterogeneous societies. If firms are not to be privately owned, as they are in the Nordic model, then a central question concerns the way that accountability of firm management is achieved. There is a principal–agent problem between the firm manager (agent) and the shareholder–citizens (principal). How do the latter keep the manager from running off with the profits and even the assets of the firm? The classical solution is that firm ownership must be highly concentrated, so that a small number of shareholders stand to gain huge amounts by carefully monitoring the management. In this view, distributing shares of firms equally to all citizens would destroy management accountability, resulting in unbridled corruption and inefficiency.

Recently a second theory has been proposed: that the guarantor of firm accountability is the corporate raider. When the raider sees the price of a firm's stock fall, because the firm is not performing well (perhaps due to management corruption or lack of imagination), he will buy a majority of shares and reorganize the firm to be efficient, thus increasing its stock price and providing him with a large capital gain. Here, too, if credit markets are imperfect, we need wealthy individuals to keep firms running well.

If these two mechanisms of accountability exhaust the possibilities, then market economies in which firm profits are distributed in a relatively equal manner to citizens are impossible. An apparent alternative, however, exists in Germany and Japan, where firms are monitored by boards of directors consisting largely of officers of banks that have a relationship to the firm. It is beyond my scope to describe this mechanism here: suffice to say it provides an alternative to relying upon hugely wealthy individuals for guaranteeing firm accountability. If market socialism has a future, it may well be with this kind of arrangement: firms will be monitored by bankers from the public sector, whose reputations and careers depend upon doing a good job, or they may be monitored by other stakeholders of the firm.

Another alternative (proposed by Roemer, 1994), with no present real-world examples, is a system in which firm ownership is distributed to citizens in an initially equal way but ownership rights are circumscribed. An owner will collect dividends from the firms in her portfolio, and even trade equity shares on a stock market, but she may not liquidate her equity holdings for cash. This would be accomplished by denominating corporate shares in a special unit of account. The values of shares in that unit would oscillate according to supply and demand, reflecting traders’ views about the future profitability of firms, as in a standard stock market. At death, a citizen's portfolio would escheat to the Treasury, and young adults, at the age of 21, would each receive their endowment of shares. Some inequality in the values of shareholdings would emerge as a consequence of differential luck and skill in the stock market during a lifetime, but that equality would not be passed on to descendants. In other words, this scheme is a method by which the nation's profit income could be distributed in a relatively equal manner to citizens, while the virtues of a stock market, with respect to the valuation of shares, and the disciplining of management are retained.

There are surely possibilities for undermining the intentions of such a system. If there are also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
individuals (such as foreigners) who are allowed to invest in these firms, then possibilities arise for citizens to capitalize their holdings, to cash out their shares. Old citizens will want firms in which they hold shares to sell their assets and pay out the entire value of the firm as dividends. Whether regulation could make the system workable is an open question.

Finally, there is the possibility of state ownership of firms. We do not as yet have a definitive experiment to test whether state ownership can work, for the Soviet-type experiments also involved *lack of democracy*; it is logically possible that democratic accountability could keep state-owned firms running efficiently. There are, however, problems here as well: politicians, to whom state-owned firm managers ultimately report, have their own interests that do not always coincide with those of the public. The electoral mechanism is probably too crude a tool to force politicians to monitor firms in the public interest. (Indeed, state-owned firms often pay their workers too much, to garner their political support.) I conjecture that non-state ownership of firms will be significant in any future socialist experiment.

We return finally to the relationship of socialism to equality. Do socialists believe that an economy which implements ‘from each according to his ability, to each according to his work’, which by definition eliminates Marxian exploitation, is desirable? Most socialists probably desire more equality than this, at least in societies where workers are highly heterogeneous in skills. Thus, socialists have come to be, and perhaps always were, more egalitarian than the Marxian definition would imply. Popular usage suggests that socialism should be defined as a regime of income equality, a departure from the Marxian tradition.

The proposals we have discussed above are all concerned with the allocation of profit income. But is the allocation of profits so important with regard to equalizing the distribution of income? In contemporary advanced economies, profits (including interest and rents) comprise at most one-quarter of national income; even if this part were distributed in an egalitarian manner to all households, and remained of the same size, the distribution of income would still, in most advanced countries, be quite unequal. Should, then, the difference between socialism, popularly conceived, and capitalism lie mainly in the distribution of wage income or of the role of redistributive taxation of labour income?

Rather than trying to define at what Gini coefficient a society becomes socialist, one can be satisfied with ordering regimes in the world with respect to their degree of socialism. The central instruments for socialist implementation then become, as well as the redistribution of profit income, intensive investment in education, with a bias towards rectifying the disadvantages children suffer due to being raised by poorly educated parents, in order to equalize market-determined labour incomes, and redistribution of labour income through taxation. The channel of intensive investment in education of the disadvantaged is important because the provision of skills has value to persons for reasons other than the instrumental one of providing income: education renders life more meaningful and fruitful. But if the education channel alone turns out to be too costly, or too ineffective, to engender the changes in income distribution which are desirable, then other methods must be used as well. The issue of *feasible* socialism, therefore, will hinge upon the package of reforms that are effective and can be realized through democratic means.

Feasible socialism: immigration and unemployment

How politically feasible is socialism – that is, to what degree can we expect democracies to implement the reforms that move societies further along the socialist scale? Here the most hopeful historical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
evidence is provided by the Nordic and north European countries. Two problems seem to be paramount for the continuation of the socialist trajectory in these economies: those of immigration and unemployment.

The welfare states of the north European countries, as mentioned earlier, evolved during the period when their populations were largely homogeneous, along ethnic, linguistic, and religious dimensions. Homogeneity may be a necessary condition for the democratic implementation of significant redistribution if the welfare state is motivated by either a purely redistributive or an insurance function. For, with respect to the insurance function, it is not in the interest of highly educated and high-wage natives in, let us say, Denmark, to pool their risks with poorly educated, low-wage immigrants. And with respect to the purely redistributive function, ethnic, linguistic and religious heterogeneity reduce solidarity, to put it mildly, which must be the motivation of purely redistributive taxation.

Unemployment is a problem not only for the deleterious welfare effects its victims suffer but because it is a severe form of economic inefficiency. If ‘socialist’ countries have high unemployment levels, and ‘capitalist’ countries low levels, eventually the inefficiency of the former may well reduce per capita income significantly below that of the latter, and populations of the socialist countries will begin to find the higher incomes offered, on average, by the capitalist regimes an attractive alternative. If we assume that, in the coming century, the United States (and, let us say, China) continue to offer low-unemployment, low-taxation, high-growth regimes, but with relatively little redistribution, then democratic polities in Europe and the rest of the world may be reluctant to move further along the socialist spectrum. This, of course, assumes that there is some sacrifice in economic growth entailed by redistributive institutions, a point that I have not defended here but have taken for granted, and which may be incorrect. Indeed, a growing literature asserts that equality increases productivity (see Bardhan and Bowles, 2000).

It is perfectly natural for fertility rates to fall when social insurance replaces the family as the source of income in old age: and smaller families, probably more than anything, entail the liberation of women. (They are also, of course, an effect of that liberation.) But European fertility rates now necessitate either a significant flow of immigrants from poorer countries, or a sharp decline in per capita incomes in Europe for retired workers, or an increase in the length of working life (which itself would exacerbate the unemployment problem). So lower fertility renders the progress towards socialism more complex at least, if not infeasible.

Consequently, the issue of multiculturalism becomes a key intellectual problem for socialists. What degree of integration or assimilation of immigrants is necessary for democratic European polities to be willing and interested to continue and perhaps expand their welfare states? (Recall, we refer here not simply to the redistributive motive but to the risk-bearing motive of natives wishing to pool risks with immigrants.) We do not, I think, yet know the answer. And will this degree of assimilation, whatever it turns out to be, be acceptable to poor Southerners or Easterners who are contemplating migration to the North or West?

Socialism, in the sense of equality of incomes, with a democratic implementation requires either a self-interested insurance motive or a selfless solidaristic motive among the majority of voter–citizens. We can hope that, as national populations come to experience more equality, they would come to have a deeper preference for it: socialists, at least, believe that solidaristic preferences can intensify with the experience of equality because equality is a public good, a fact that will be appreciated when it is experienced. (Indeed, we have not, in this article, discussed the *negative* externalities that socialists

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
believe accompany a regime with a highly concentrated ownership of private firms, in which corporate and even state policy is set to further the interests of only the wealthiest sliver of society.) But the initial transitions along this path, taken by relatively self-centred voters, must come from the insurance motive. Here, then, is an important problem for progress towards socialism in our time.

See Also

- labour's share of income
- Marx, Karl Heinrich
- redistribution of income and wealth
- risk sharing
- social democracy
- socialism
- surplus

Bibliography

Bardhan, P. and Bowles, S. 2000. Wealth inequality, wealth constraints and economic performance. In *Handbook of Income Distribution*, ed. A. Atkinson and F. Bourguignon. Amsterdam: Elsevier Science Press.

Burawoy, M. and Lukacs, J. 1985. Mythologies of work: A comparison of firms in state socialism and advanced capitalism. *American Sociological Review* 50, 723–37.

Hayek, F.A. 1940. Socialist calculation: the competitive solution. *Economica* 7, 125–49.

Lange, O. 1936. On the economic theory of socialism. In *On the Economic Theory of Socialism*, ed. B. Lippincott. Minneapolis: University of Minnesota Press, 1956.

Luxemburg, R. 1913. *The Accumulation of Capital*. New York: Monthly Review Press, 1972.

Makowski, L. and Ostroy, J. 1993. General equilibrium and market socialism: clarifying the logic of competitive markets. In *Market Socialism: The Current Debate*, ed. P. Bardhan and J. Roemer. New York: Oxford University Press.

Marx, K. 1867. *Capital*, vol. 1. Basingstoke: Penguin Classics, 1992.

Roemer, J. 1994. *A Future for Socialism*. Cambridge, MA: Harvard University Press.

Roemer, J. and Silvestre, J. 1993. The proportional solution for economies with both private and public

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
ownership. *Journal of Economic Theory* 59, 426–44.

How to cite this article

Roemer, J. E. "socialism (new perspectives)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000449> doi:10.1057/9780230226203.1569

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

socialism

Alec Nove
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A society may be defined as socialist if the major part of the means of production of goods and services is in some sense socially owned and operated, by state, socialized or cooperative enterprises. The practical issues of socialism comprise the relationships between management and workforce within the enterprise, the interrelationships between production units (plan versus markets), and, if the state owns and operates any part of the economy, who controls it and how. ‘Feasible socialism’ would take the form of a mixed economy, with enterprises large and small, many if not most self-managed or cooperative, and some privately owned.

Keywords

alienation; capitalism; Chicago School; class; class struggle; collectivization; communism; division of labour; economic calculation; economic development; Engels, F.; entrepreneurship; exploitation; Fabian Society; inequality; labour theory of value; Lenin, V.; Mao Tse-Tung; Marx, K.; Marxian value analysis; methodological individualism; monopoly capitalism; perfect competition; perfect foresight; proletariat; rate of profit; self-management; social democracy; socialism; unemployment; use value

Article

It is said that the word ‘socialism’ was first used by Pierre Leroux, a supporter of Saint-Simon, in 1832, and was quickly taken up by Robert Owen. The word has meant many different things to different people. It has been used as a synonym for communism, i.e. as a bright vision of a future in which there are neither rich nor poor, neither exploiters nor exploited, in which, to use an expression borrowed from Charles Taylor, ‘generic man is harmoniously united in the face of nature’. It is by definition the solution of most if not all economic problems, the end of ‘alienation’. As such it has religious overtones: Man was at one time in harmony with society, and will become so once again. For others, these utopian-sounding aims are either meaningless or a vague ideal, the higher stage of communism. ‘Socialism’ is,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
so to speak, here on earth, and can be seen (according both to Soviet doctrine and to right-wing critics) in the ‘really existing socialism’ of countries in the Soviet sphere, who claim to be on the way towards a communist future. Still others criticize this ‘really existing socialism’ from the left, declaring it not to be socialism at all; their criteria for what constitutes socialism are not always very clear, some using the marxist vision as their point of departure, others laying stress on the lack of democracy, the hierarchical nature of society, and other departures from what, in their view, ought to be. The term ‘socialism’ is also used, or misused, to describe the aims and programme of the British Labour party, or the state of affairs actually achieved under a series of social-democratic governments in Sweden. The term at one time had an appeal to moderates. Thus the moderate-reforming party of the Third Republic in France chose to call itself Radical-Socialist, though its leaders, such as Edouard Herriot, had no aims which could qualify as socialist. Then, at the extreme right of the political spectrum, Hitler's party was self-described as national-*socialist*.

So one should proceed at an early stage to a definition, or rather to exclusions. Not Hitler, obviously. Nor Herriot either. If one were to adopt a definition which corresponds with Marx's vision of socialism (of which much more below), there is the evident danger of adopting an impossibly rigid criterion by which to judge any real-world society: thus, whatever reasons there may be to criticize or condemn today's USSR, it would be rather pointless to ‘accuse’ it of not having ensured the withering away of the state, or not having ‘surmounted’ (*aufgehoben*) the division of labour. Let us provisionally accept the following as a definition of socialism: a society may be seen to be a socialist one if the major part of the means of production of goods and services are not in private hands, but are in some sense socially owned and operated, by state, socialized or cooperative enterprises. ‘The major part’ is enough. Just as any non-dogmatic socialist would accept that most ‘capitalist’ countries contain sizeable state and cooperative sectors but still deserve the label ‘capitalist’. This leaves three big questions unanswered:

1. (1) What are the relationships between management and workforce *within* the enterprise?
2. (2) How do the production units interrelate? (i.e. by plan, by contractual or market relations, or some combination of both).
3. (3) If the state or other public bodies own and operate any part of the economy, who controls the state, and how. One remembers the remark attributed to Engels, that if state ownership is the criterion of socialism, the first socialist institution was the regimental tailor.

If the world ‘socialist’ was coined in 1832, the idea of socialism long preceded it. Among the first to put forward principles which contain strong socialist elements was Gerard Winstanley, representing the Levellers of Cromwell's time. They believed in equality, wished property to be held in common, opposed concentrations of private wealth. During the French revolution Babeuf denounced inequalities of wealth and advocated the overthrow of the government, which he saw as representing property-owners. Robert Owen could be described as a paternalist, in that he believed in good treatment of his employees (as can be seen even today in the housing he built for his workers in New Lanark), but he also envisaged what would now be called producers’ cooperatives. As essentially a practical man, he can be distinguished from those ‘utopian socialists’ who, before Marx, have painted a series of pictures of imaginary socialist-type societies. Leszek Kolakowski (1981) analyses the ideas of men like Fourier, Saint-Simon, Proudhon, and notes certain elements of similarity with those of Marx, and also some essential differences. They have in common, inter alia, a hate for the ‘bourgeois’ order, a society based

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
upon greed, profit, the mercantile spirit. The French revolution substituted plutocracy for aristocracy. Unlike Marx, they did not consider this to be a progressive stage in the history of mankind, but, like Marx, they stressed the ugly features of capitalist industrialism and wished to do away with it, substituting a new harmony, cooperation, the reassertion of the true rights of Man. They rejected Adam Smith's basic idea that common good is generally attained through the competitive profit-making process. As, for instance, was asserted by Saint-Simon, the basic cause of human misery is free competition and the anarchy of the market. The so-called utopians varied in their approach to the issue of equality: thus in Fourier's ‘phalansteries’ the means of production were held in common, children were to be brought up together, the family would dissolve, there would be provision of subsistence for all, but Fourier would encourage individual enrichment through work (though not the inheritance of riches or unearned incomes). Some advocated violent revolution to achieve their objectives, others hated violence and hoped to persuade their fellow-citizens to adopt freely the ideas of the good and just society of their imagination.

As will be argued later, Marx differed from his predecessors not because he conceived of a realistic alternative to capitalism: there was much that was utopian in his ideas too. However, *firstly* he did not go into detail as to how a future society would function; nothing in Marx is similar to such notions as phalansteries, or radiant cities of 1800 persons with 810 different human characteristics, or the idea that dirty work that needs doing will be done by boys, who, as everyone knows, like dirt; Marx favoured the emancipation of women, but he did not follow Fourier in drawing up a ‘table des termes de l'alternat amoureux’.

Secondly, and more important, he provided a set of powerfully argued historical reasons as to why the desired state of affairs must come to pass. As Engels said at his graveside: ‘Just as Darwin discovered the law of development of organic nature, so Marx discovered the law of development of human history.’ The class struggle, the growth of monopoly capitalism, the proletarianization of the petty bourgeoisie (peasants, shopkeepers, small businessmen of all kinds), the growing misery of the masses, the growth of class-consciousness, the logic and consequences of large-scale industry, the belief that, having spectacularly developed the forms of production, the bourgeois-capitalist relations of production act as fetters on the further development of productive forces, all these things will lead inexorably towards socialism. Ever-deepening crises, the falling rate of profit, the refusal of the poverty-stricken masses to accept their lot, i.e. the accumulation of capitalist contradictions, will bring the system down. The proletariat, having overthrown the bourgeoisie, would inaugurate the classless society. In the marxist tradition there are various interpretations of the relative importance of historic necessity (i.e. inevitableness, a march towards a predestined goal) and voluntariness (deliberate human action designed to achieve the goal). These two principles coexist uneasily, and they can be seen as mutually inconsistent, but they can be reconciled. To take two examples, it *is* meaningful to assert that, should a professional soccer team play a school side, the professionals would ‘inevitably’ win. The same would be (was) true of a conflict between the Germans and say the Luxemburg army. However, the outcome requires human action, on the part of the footballers and the German soldiers respectively.

This calls for two kinds of comments. One relates to the interpretation of history, the other to the utopian elements of so-called scientific socialism.

It hardly needs stressing that capitalism has not evolved in the manner foreseen by Marx. He himself stressed, in a famous passage, that no mode of production passes from the historical scene before its productive potential is exhausted. He believed that capitalism was reaching exhaustion already when he

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
was writing *Das Kapital*. Over a hundred years later it is still not exhausted, and ever-new technological revolutions, while certainly presenting new problems and dilemmas of which we shall speak, continue to enlarge the productive potential of capitalist society. It is also clear that the concept of ‘proletarianization’ was wide of the mark. Yes, great concentrations of ‘monopoly-capital’ do exist, but so do very large numbers of small businesses and a far larger number of ‘professionals’ of all sorts and grades who are, or consider themselves to be, middle class. This fact has given rise to much debate among marxists, typified by the argument between Poulantzas and Erik Wright (see for example, Wright, 1979). We need not go into this argument, which turns on who could or could not be considered to be working class. The political and social fact remains that a large and growing proportion of the citizenry of developed countries do not own the means of production and are emphatically not class-conscious proletarians.

Furthermore, the development of the forces of production has made possible a substantial improvement in the living standards even of those who in any definition are workers. Clearly, they do not have ‘nothing to lose but their chains’. It is neither original nor amusing to say that men who have ‘nothing to lose’ except a three-bedroomed house, a car, a video-tape machine and a holiday in Spain are not very likely to be revolutionaries, or indeed particularly interested in socialism. It is true none the less. Marx himself, and some of his followers, when willing to recognize that living standards could rise, insisted that this does not remove the essential antagonism between labour and capital, the existence of exploitation and alienation. In a sense this is so, though one must avoid an oversimplified zero-sum-game approach; situations arise in which both profits and wages can rise together, as they have done in successful capitalist countries in the twenty-five years that followed the last war. Nor is there any necessary correlation between the depth of human misery and the spirit of revolt. None the less, the lack of support for the socialist alternative in developed countries cannot be treated as merely a temporary aberration. It is also true that revolutions, whatever their merits or necessities, impose grave hardship upon people, notably the masses. The association of the word ‘socialism’ with revolution is therefore an important reason for many ‘proletarians’ *not* to support the socialist idea, at least in developed countries. ‘Underdeveloped socialism’ is a different question, to be tackled later.

Now to the utopian nature of Marx's ‘scientific socialism’. The key points to make are:

1. 1. *Abundance*. Here Marx reflects the optimism of his century, yet natural resources are not inexhaustible. Human needs and wants increase – as indeed Marx himself recognized. Conservationist and ecological socialism can be strongly defended, but this is precisely because resources (even the air we breathe, the water we drink) are finite. It is not the case that the problem of production has been ‘solved’, and that socialists will not require to take seriously the question of the allocation of scarce resources. I define ‘abundance’ as a sufficiency for all reasonable requirements at zero price.
2. 2. *The non-acquisitive ‘new man’*. His (and her) appearance surely presupposes abundance. Marx himself was perfectly clear that a share-poverty ‘socialism’ would reproduce ‘the old rubbish’. Men do not become good by being so persuaded, or by reading good books. If there is enough for everyone, then there is no need to strive to keep things for oneself, one's family, one's locality, one's institution. If there is scarcity, therefore opportunity cost, therefore a situation in which there are mutually exclusive alternatives, then conflict on priorities of resource allocation is inevitable. This does not in fact require any assumption about individual egoism. Even unselfish

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

persons tend to identify the needs they know with the common good. Indeed, in a complex modern society there is no generally accepted and objectively based criterion as to what 'the common good' is. Nor can any individual apprehend the multitude of alternative uses potentially available for the resources he or she desires, either for him/her self or for the given township, library, orchestra, football team, industry or whatever.

3. 3. *The political assumptions*. These are linked with (1) and (2), above. The state withers away, not only because it is assumed that its 'essential' repressive functions are not needed when no ruling class imposes its will on the masses, but also because, to re-cite Charles Taylor, Marx assumed a 'generic man harmoniously united in the face of nature'. Consequently there would be no need for legal institutions, coercive powers, police, indeed any politics as we know them. Civil society and individuals will have merged, the task of the 'administration of things' would not be undertaken by political institutions, would be merely technical. There *is* no marxist *political* theory of socialism.
4. 4. *The economic assumptions*.
    1. (a) *Value theory and economic calculation*. The suppression of the market, of commodity production, of money, seems to involve the 'withering away' of the law of value. What is to replace it? Presumably it will continue to be important to use resources economically to provide the goods and services desired by society. How are calculations to be made? On this Marx is almost totally silent. Engels, in *Anti-Dühring*, speaks of assessing use-values and relating them to the labour-time required to provide for them. This runs at once into several rather evident problems. First is the theoretical one that Marx most emphatically (at the very beginning of the first volume of *Das Kapital*) asserted that different use-values were not comparable, so could not be added up or subtracted. A pen, a cup, a book, a skirt, a light-bulb (to take a few examples at random) satisfy different needs. The one thing they have in common, apart from satisfying various needs, is that they are the products of labour. How, in any case, are Engels's use-values to be computed, by whom, on the basis of what criteria? In a book wholly devoted to marxian use-value (*valeur d'usage*), G. Roland (1985) goes at length into the basically unsatisfactory treatment by Marx of use-value, due apparently to his anxiety to distance himself from subjective value theory. This has created some awkward problems for Soviet pricing theory, or at the very least does nothing to help. The dogmatists insist that Marxian labour-values ought to underlie Soviet prices, or alternatively that these be modified into the equivalence of 'prices of production', but both of these share the characteristic of being based on effort, on cost. This not only fails to give due weight to utility (or user preferences), but also runs into yet another problem, or rather two interlinked problems; measuring labour inputs, and the failure to take into account other scarcities. A few brief remarks are appropriate on each of these points. Can one actually identify the labour content, including the labour embodied in machine and materials, and the 'share' in joint overheads, of hundreds of thousands or even millions of different goods and services? This is a hugely difficult if not impossible task, even if one calculated only in hours of labour. But then what of skilled labour? How is it to be 'reduced' to simple labour? Marx does not handle this 'reduction' satisfactorily in discussing value in capitalist society, and in the end one is left with actual wage ratios as the only usable criterion, which is unhelpfully circular. And then can one

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

treat labour as the only scarce factor? What of land, oil, timber, what of time (not labour-time, but, say, delay in construction)? Novozhilov remarked that the most modern equipment would be scarce even under full communism unless it be assumed that technical progress ceases. Space forbids further remarks about other deficiencies of the labour-theory inherited from Marx. (Thus demand or price must affect labour-content if there are economies or diseconomies of scale, or if relative prices influence choice of techniques.) And if the purist retorts that Marxian value theory is not supposed to apply to socialist economies at all, then he or she must be asked: 'What is your alternative?'. This has (so far) usually taken the form of some surrogate labour-theory (such as hours of human effort), with all the deficiencies of such an approach.

2. (b) '*Simplicity*'. The lack of interest – until comparatively recently – of Marx and marxists in the question of economic calculation under socialism is explicable by a grave misunderstanding, i.e. by the belief that the complexities of modern industrial society are a consequence of commodity production and 'commodity fetishism', which conceal relations which, as Marx said, were inherently 'clear and transparent'. 'Everything will be quite simple without this so-called value', said Engels. Planning under socialism 'will be child's play', said Bebel. 'To organize the entire economy on the lines of the postal service, … under the leadership of the armed proletariat, this is immediate [*sic*] task' (Lenin, in 1917), and so on. But evidently in a modern industrial society with hundreds of millions of people, hundreds of thousands of productive units, millions of products and services (if disaggregated down to specific items, there *are* millions), it is a hugely complex task to discover exactly who needs what, and to identify the most effective means of providing for needs, especially if one bears in mind that any output requires the acquisition (or allocation) of dozens or more of inputs. Barone (in his path-breaking 'Ministry of production in a collectivist state') pointed this out in 1908, but failed to get a hearing from the socialists of his time. It is nonsense to talk of labour under socialism being 'directly social', in the sense of being applied with advance knowledge of needs – contrasting with *ex post* validation through the market under capitalism. This can only be so if perfect knowledge and foresight were assumed, and the need to test *ex post* for possible error assumed to be unnecessary. All socialists (rightly!) reject theories which assume perfect foresight, perfect markets, perfect competition, when put forward by neoclassical model-builders. So, apart from problems of value theory, there is the sheer complexity of marketless, quantitative planning, the formidable obstacles in the way of identifying requirements and providing for their satisfaction.
3. (c) *Political-social implications*. Lest the above be seen as 'merely technical', and so remediable by computers, the objective requirements of marketless planning in a complex industrial economy are centralizing (who but the centre can identify need and ensure the allocation of means of production?), hierarchical, bureaucratic, and concentrate immense power over both people and things in the hands of the state apparatus. The importance of political democracy is undeniable, but the officials (who else?) who plan the output and allocation of sheet steel, sulphuric acid and flour are taking decisions unconnected with democratic voting – save in the sense that such voting should affect broad priorities. There were moments when Marx, Engels, Lenin, showed that they understood the inevitability

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of hierarchy: thus Lenin saw the socialist economy as a sort of ‘single office, a single factory’, with ‘a single will linking all the sub-units together’ to ensure the parts of the economy fitted together ‘like clockwork’ (Lenin, 1962, pp. 157). But whereas clockwork functions automatically (i.e. is not unlike the ‘hidden hand’, or maybe the hidden pendulum), in a marketless economy the parts have to be moved by human beings charged with the purpose. The contrast between ‘the administration of man’ and ‘the administration of things’ (a phrase borrowed by Marx from Saint-Simon) is a false contrast: I am quite unable to ‘administer’ this piece of paper, but I can persuade a secretary to type it, a postman to deliver it to the publisher, and (hopefully!) the publisher decides to tell the printers to print it! All Soviet experience underlines the political and social consequences of the high concentration of hierarchically organized economic power.

5. 5. *Division of labour and ‘alienation’*. There is, and must surely be, a division of labour between productive units (those that produce sulphuric acid, steel or hairdressing services are unlikely also to be making hats, computer software or music). Marx's notion of a universal man, who fishes, looks after sheep and writes literary criticism, without being a professional fisherman, shepherd or critic, makes no sense, other than in the (sensible but weaker) form of aiming at a greater degree of job interchangeability. Thus the author of these lines was once a soldier, then a bureaucrat, then a university teacher, but could not be all of these at once. The vertical division of labour (e.g. between management and those managed) could also be modified by some system of rotation or election, but management is also a skill, and human intelligence is not of itself a guarantee of tolerable administrative ability: we all know of good specialists who could not (and would not wish to) administer anything well. One is then struck by the inherent unreality of such books as by I. Mészáros (1972). Mészáros fully and correctly sets out Marx's view, and he does state that ‘the political road to the supersession of alienation and reification’ is a long one and success is not guaranteed. But he still sees the ‘transcendence of alienation’ as a meaningful goal, as if separation of Man from his product, his subordination to outside forces, the division of labour, can be overcome through the elimination of private ownership. And Kolakowski (1981, p. 172) is surely right when he notes that for Marx ‘the fundamental premise of alienation is already present as soon as goods become commodities’, and that ‘the division of labour leads necessarily to commerce’. So alienation appears to be the inescapable consequence of an inescapable division of labour, so how can it be *aufgehoben*? Private ownership represents a particular manifestation of ‘outside’ control, and it is an important part of any socialist programme to give to labour a greater influence over the work process. But what can one make of Bettelheim (1968), when he criticizes Yugoslav-type self-management enterprises for what is surely the wrong reason: that they are controlled not by the workforce but by the market. It ought to be clear that production is for use, and that *what* is produced ought in the last analysis to conform to user needs, i.e. to be controlled by a force outside the production unit itself. This could be the market, in which bargaining takes place between producer and user. It could be a planning agency, who informs the production unit what is should be doing. *Tertium non datur*.
6. 6. *Labour, wages, ‘the proleteriat’*. Several distinct points need to be made.
    1. (a) *The end of the wages system*. This is not what real workers want. Money wages give freedom of choice, including the choice of hiring the services of each other (to repair the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

roof, baby-mind, drive to work or whatever). Marx's idea of tokens denominated in hours of labour (‘which are not money and do not circulate’) makes very little sense, and not surprisingly has not been applied. If goods are distributed free, this usually limits consumer choice: you take what you are *given*.

2. (b) *Labour direction* is the sole known alternative to material incentives or other forms of inequality. This was understood by Kautsky, Trotsky and Bukharin, when they discussed this question. The term ‘labour market’ has an opprobrious sound, reminiscent perhaps of a slave market. Yet workers are freer, have greater choice, more possibility to bargain, than under direction of labour, necessarily exercised by officials with power over persons.
3. (c) *The proletariat as redemptor humanis* is essentially a religious concept, unrelated to the qualities and desires of the real working class. Eloquent words on this subject have been written by Andre Goroz: ‘No empirical observation or actual experience of struggle can lead to the discovery of the historic mission of the proletariat which, according to Marx, is the constituent of its class being’ (Gorz, 1980, p. 22). Rudolf Bahro wrote that ‘the proletariat, the collective subject of general emancipation, remains a philosophical hypothesis in which is concentrated the utopian element of marxism’, and he added, rightly, that ‘the immediate objectives of subordinate classes and strata are always conservative’ (*sind immer Konservativ*) (1977, p. 174). But if one accepts these and other similar arguments, it follows that, as Lenin said, the working class left to itself will limit itself to ‘trade union’ types of demands, and so it is the task of the revolutionary intelligentsia to provide the revolutionary theory. This in turn leads to what has been called ‘substitutionism’, i.e. a party dominated at the top by non-workers, which in its turn dominates society, an outcome prophesied by Bakunin well over a hundred years ago. It is clearly not the case that, to cite Marx's letter to Weydemeyer in 1852, ‘the class struggle leads necessarily to the dictatorship of the proletariat’, which ‘is but a transition to the withering away of classes’ (letter dated 5 March 1852, Marx, 1962, p. 427).

Marxists may now be impatiently protesting that the above analysis is a vision of full communism, that no one, certainly not Marx himself, expected this to be realized quickly, or even certainly. The much-used words ‘socialisme ou barbarie’ show a recognition that barbarism can be an outcome if the socialist idea fails. Trotsky spoke often of a ‘transitional epoch’ during which money, markets, commodity production, are indeed indispensables. Soviet discussions refer to the indeterminate length of time required to move from ‘socialism’ (i.e. Soviet reality, which they define as socialism) to full communism. For example, a book devoted to the subject and published for the fiftieth anniversary of the revolution duly lists the characteristics of communism (abundance; from each according to his ability, to each according to needs; the elimination of commodity money relations, and so on), but goes on to stress that communism must be preceded by the lower ‘socialist’ phase, and that to try to overleap that phase is ‘a harmful utopia’ (Gatovski, et al (eds), 1967, p. 9, p. 43).

Marx himself used ‘socialism’ and ‘communism’ almost as interchangeable terms. Whether ‘really existing socialism’ should be seen as a transitional society or as socialist is to some extent just a terminological question. In either case it is supposed to be evolving towards fully fledged socialism or communism. But does it? Should it? What are the signs by which such an evolution can be identified? Bettelheim has good evidence for his view that, for Marx and Engels, when the workers acquire the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
means of production, ‘there will be in socialist society, even at the beginning, no commodities, no value, no money, and consequently no prices and no wages’ (Bettelheim, 1968, p. 32). Equally strongly, the French critic Cornelius Castoriadis roundly asserts that ‘Marx knew nothing of transitional societies infinitely contained within each other like Russian dolls or Chinese boxes, which Trotskyists later invented’ (Castoriadis, 1979, p. 299). Marx did specifically say, in the *Grundrisse*, that ‘nothing is more absurd than to imagine that the associated producers’ would choose to interrelate via commodity production, exchange, markets. We have already noted that, in his *Critique of the Gotha Programme*, Marx envisaged an immediate conversion of wages into tokens denominated in hours of labour, not despite but *because* society will still bear the stigmata of pre-socialist attitudes. In the 1920s in the Soviet Union it seemed obvious to the party comrades that reducing the area of market relations was in some sense the equivalent of an advance towards socialism. Indeed those who forced 25 million peasant households to join so-called collective farms thought that this was part of the class struggle, though the effect was to turn independent ‘petty-bourgeois’ households, who did to a considerable degree control their own means of production and their product, into something akin to a new sort of state serfdom. If socialism is to do with the liberation of the ‘direct producers’, then surely this was a march in the wrong direction.

Similarly, can we say that the Hungarian or Soviet reformers of today are wrong in advocating an extension of ‘commodity-money relations’? And if the point is made that such a judgment would be premature, but that communism is still an aim to pursue when circumstances are propitious, it is legitimate to ask: what circumstances can be imagined in which communism/socialism in Marx's sense *could* come about? No wonder the Soviet orthodoxy of today is to speak of ‘mature socialism’ as a long-term stage, with communism seen as a remote objective of no short-term operational significance.

There were also socialist alternatives to Marx, during and since his lifetime. William Morris combined some ideas derived from Marx with ethical socialism and devotion to arts and crafts. Others further developed Christian socialism of various kinds, and indeed much could be made of the contrast between Christian ideals and the mercantile spirit, the ‘dark satanic mills’ (‘and we will build Jerusalem in England's green and pleasant land’). The British Labour party in its origins and for many decades afterwards was heavily influenced by Christian beliefs, especially those based on Methodist and other nonconformist creeds (thereby attracting some contemptuous remarks from Lenin). The Fabian society (Shaw, the Webbs and others) by contrast, preached non-religious (and non-violent) socialism, opposed extremes of inequality, and advocated industrial democracy. However, though they too influenced the Labour party, the Society remained a small intellectual group, with a tendency to believe that an elite (themselves), or even a strong dictator, would show the way. It is perhaps no accident that both Shaw and the Webbs lived to express an admiration for Stalin – even though they themselves would recoil from cruelty and killing. Mention must also be made of G.D.H. Cole and ‘guild socialism’, with decentralized decision-making by producers’ associations.

On the continent, social-democracy nominally retained its allegiance to Marxism. However, already in 1899 Edward Bernstein advocated a non-revolutionary revision of many of Marx's theories. While the leaders of German social-democracy, men like Bebel and Kautsky, rejected Bernstein's ‘revisionism’, it was in fact rooted in the considerable improvement of the workers’ living standards, the weakening of revolutionary spirit. In the end, while retaining marxism as their nominal creed, German and other continental social-democrats (notably the ‘Austro-marxists’, such as Otto Bauer) adopted a non-revolutionary position which differed little from Bernstein's and became a party of moderate reform

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
within capitalist society.

In Russia, side by side with the growth of marxism (initially preached by men such as Plekhanov and Zieber) there arose other and non-marxist socialist currents, sometimes labelled ‘populist’. They believed that a Russian road to some form of socialism could be found, perhaps based on traditional communal institutions, which would enable capitalism to be by-passed. These ideas came from men such as Mikhailovsky and Vorontsov. As we shall see, Marx himself did not reject this possibility. There were also some influential anarchist socialists, owing inspiration to Bakunin, of which Prince Peter Kropotkin was a colourful example.

Since 1945 European social-democrats have tended to abandon their already tenuous allegiance to Marx and Marxism, and it may be hard to discern the extent of commitment to socialism of any sort in the programme and policies of the German and French parties. By contrast, the recent evolution of the Italian communist party has put it close to a social-democratic, evolutionist position. Opinions vary within the British and the Scandinavian Labour parties. Further change may well depend greatly on what happens to contemporary capitalism.

Of course the future may reserve surprises for us all. While material resources may be finite, the scientific-technical revolution may enable us to economise labour on a big scale. The resulting high level of unemployment may be a chronic disease. True, by freeing factory and office labour, we could, in a more rational society, greatly enlarge labour-intensive forms of providing a higher quality of life. But precisely this is opposed, and successfully so, by the New Right, by the ‘Chicago’ ideology, which is vehemently against public expenditures. Yet we may already be reaching a stage in which the *profitable* (privately profitable) use of labour can cover only a portion of those available for work. A possible reading of Marx places emphasis on equating the realm of freedom with freedom *from* work (i. e. from necessity), with a much shorter working week, and Gorz too sees freedom as a situation where one can undertake handicrafts and other hobbies. This would be a paradoxical reversal of the view that the one scarce factor of production is labour, since then it would be the abundant factor, the problem being how to share it out. This would not be the era of abundance. To cite an example, fish could be caught by modern trawlers using fewer fishermen, but dangers of over-fishing would compel a strict limitation on numbers caught. This brings one back to the idea of an environment-preserving, ecologically conscious, employment-sharing socialism as an attractive alternative to capitalism. But this was not Marx's alternative.

A case for socialism can be made, not only along the ‘ecological’ lines mentioned above. In the developed world, massive resources are devoted to persuading people to buy trivia, to keep up with the Joneses. Unemployment is a scourge which is a threat to public order. External diseconomies (and external economies too) frequently cause the pursuit of private micro-profit to conflict with more general interest. The ‘quality of life’ may not be readily quantifiable, but several economists (for instance Kuznets, Tobin) have noted that conventional measures of economic growth by GNP can conceal real losses, or indeed count real *costs* of urban living as a net addition to welfare. The inequalities of income and property-ownership have all too often no visible connection with the contribution to society or to production of the individuals concerned. Schumpeter (cited in Brus, 1980) rightly pointed out that no social system ‘can function which is based exclusively on free contracts … and in which everyone is guided only by personal short-term interest’. Furthermore, fanatics of the New Right are engaged in reducing essential public services, disintegrating where possible the welfare state, cutting back public transport, pursuing dogmatic monetarism, in the naive belief that primitive laissez-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
faire is the best of all possible worlds. It may turn out that that the grave-diggers of capitalism will be those ultra-'liberal' ideologists who fail to understand how modern capitalism really works, that the so-called imperfections (price and wage stickiness, administered prices, oligopoly and so on) are preconditions for the functioning of the system. On the assumption of perfect competition, perfect markets, perfect foresight, there is no role for the entrepreneur, no reason for firms to exist, and logically enough profits tend to zero in equilibrium. The idea that rational investment decisions are possible when we face so many inflationary uncertainties (what will the rate of the dollar, or the rate of interest, be in a year's time?) is somewhat far-fetched, to put it mildly, and inconsistent with meaningful 'rational expectations'. The belief that all markets clear, that unemployment is 'voluntary leisure preference', curable by freeing the labour market, will sound very odd to future generations.

No socialist should deny the need for economic calculations. With no price mechanism it is not possible to calculate or compare cost, or to measure the intensity of wants. Microdemand cannot be derived from voting or from clamour, nor should there be 'dictatorship over needs', to cite the title of a critique of East European socialism (Feher et al., 1983). There really is no alternative to allowing choice, i.e. to 'voting' with money. Choice necessarily involves competition between actual and potential suppliers. Yet the limitations of the price mechanism also require to be clearly seen. As the Hungarian economist Janos Kornai (1971) has pointed out, major decisions are not and cannot be taken on the basis of price information alone. The currently fashionable 'methodological individualism' goes far to deny the very existence of the general interest, distinct from that of individuals composing the society, confining 'public goods' to defence and lighthouses. (Yet it is not even true that the interests of a firm are only the sum total of that of the individuals composing it!)

Socialism as an idea lays stress on the general interest, but has not always avoided overstressing this at the expense of the individuals, for otherwise the dangers of totalitarianism (albeit of a paternalist kind) may loom ahead. The notion that Man is at the mercy of blind forces he cannot control, or of mighty and remote corporations (faceless, *societés anonymes*, or worse still, inhuman computers) sets up a search for a 'socialist' alternative, more human, fairer, and not necessarily less 'efficient' in terms of human welfare. Acquisitiveness and competitiveness may be unavoidable, must indeed be utilized, but do not require to be encouraged. Individualist profit-seeking as the dominant purpose in life, can be regarded by socialists as inhuman and ultimately destructive of society. A greater – not exclusive, but a greater – emphasis on caring for others may be a precondition for survival. More directly destructive would be nuclear war. There was a long-standing attachment of the idea of socialism with that of peace. This can be less confidently argued today, alas (when Chinese and Vietnamese soldiers shot at each other, could they both be 'socialist'?). Experience does show that states aiming to be socialist can commit aggressive acts, and accumulate immense stores of destructive weapons. None the less, the autonomous role of the arms lobby and of hate-propaganda may be particularly associated with militant capitalism.

Socialist ideas in the Third World raise some specific problems. While it is dangerous to generalize about so heterogenous a group of countries, in many of them the logic and spirit of capitalism is rejected. There too, to re-quote Bahro, ordinary people are 'immer konservativ', and it is capitalism which is new, which threatens traditional ties and attitudes. The effect may or may not be to provide mass support for socialist slogans: we have had such phenomena as Khomeini and Moslem fundamentalism by way of reaction. But socialist ideas do attract many, in places as far apart and as different as Chile, India, Egypt, Zimbabwe. Of course many blunders have been committed in the name of pursuing socialist policies, not least in relations with the peasantry. But there are many examples

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
which demonstrate that there are countries where free-market capitalism, far from being associated with free and democratic institutions, requires repressive police-state measures. Pinochet's Chile is but one such example.

The relationship between socialism and economic development is a subject in itself, on which volumes could be written. It has often been pointed out that, paradoxically, marxist-inspired revolutions have occurred in relatively backward countries. Indeed, the Russian Empire in 1917 was in no sense ‘ripe’ for socialism. The preconditions were absent, and the Mensheviks considered themselves to be orthodox marxists when they denounced the Bolsheviks for trying to overleap the predestined historical stages. Lenin, on the contrary, believed that it was possible, indeed essential, to seize power when opportunity offered and *then* to create the preconditions, with (he hoped) the help of revolutions in developed industrial countries. Some of the less agreeable features of the Soviet system can be ascribed to isolation in a hostile world, or to ‘socialism in one country’, though it would be wrong, in the light of later experience (such as the evolution of the relations between the USSR and China) to regard this one factor as decisive. But, true enough, backward countries seeking to introduce ‘socialism’ introduce backward ‘socialism’. It becomes an industrializing ideology, mobilizing the masses and imposing sacrifices for the goal of modernization, of industrialization, with a substantial admixture of nationalism. Whatever may have been their conscious aim, a strong case can be made for the proposition that Lenin and Mao re-established their respective empires, after a period of breakdown and disintegration, which in China's case lasted almost a century.

Marx's attitude to the socialist transformation of backward countries was by no means clear-cut. While his basic model did point to a socialist revolution occurring in highly industrialized capitalist countries, his correspondence with Vera Zasulich showed that he had great difficulty in applying his ideas to Russia. Theodor Shanin has edited a lively and (in the best sense of the word) provocative volume (Shanin, 1984), which does show Marx's perplexity, his partial recognition that there would perhaps be a road which by-passes capitalism. This was far from the view of Russian marxists, and the correspondence with Zasulich remained unpublished until 1924. However, on other occasions Marx took a different view, as when he regarded British rule in India as progressive, in the sense of introducing capitalist relations into a traditionalist society.

Any analysis of ‘really existing socialism’ would have to take account of the major role of nationalism, though this at least would have astonished Marx. It influenced Soviet internal and foreign policies, it surely played a key role in the split between Russia and China, it may be seen in the treatment of the Hungarian minority by the Romanians. The Soviet author Vasili Grossman, in his major novel *Life and Fate* (1985), put into the mouth of one of his characters the thought that the battle of Stalingrad completed the process of transforming Bolshevism into National-Bolshevism (needless to say, the book was not published in the Soviet Union). We are very far from the idea that ‘the workers have no fatherland’, and the proper translation of the Soviet official doctrine of ‘proletarian internationalism’ is ‘acceptance of the leadership of Moscow on all important questions’.

There is one aspect of ‘backward socialism’ which has profound political and social significance. In the USSR, in China, and in many Third World countries, the peasantry formed a large part of the population and there was a sizeable petty bourgeoisie. Far from having exhausted its potentialities, the ‘marketization’ of the economy was still in its early stages. In Marx's model the bulk of the petty bourgeoisie has been eliminated by monopoly capital. But in these countries, in the name of the class struggle, it was destroyed by coercive state policies, i.e. by police measure. Indeed, the police has to be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
ever-watchful in case the banned private activities are reborn. This is one reason, among others, for there being socialist police states, which have only a remote connection with Marx's ‘dictatorship of the proletariat’.

Much could be said about socialist analyses of underdevelopment, and such names as André Gunder Frank, Samir Amin and Arrighi Emmanuel come to mind. How far was underdevelopment due to capitalism and to links with the world capitalist market? Do socialist remedies require a break with that market? Is the poverty of the Third World due to ‘unequal exchange’ and exploitation? Is there any operational meaning in the so-called transfer of values? Thus if (say) Zaire buys a machine from the United States and a precisely similar machine at the same price from India, are ‘values transferred’ in the one case and not in the other?

If Amin is to be believed, such a deal would actually impoverish Zaire if the purchase is from the United States, for presumably the machine would contain much less labour than the similar machine bought from India, or than whatever Zaire exports to America in exchange for it. Yet frankly this is nonsense. Which by no means excludes the possibility, or even the likelihood, of unequal gains from trade.

It is of interest, in the light of some socialist theories of development, to compare the experience of various countries which follow widely different models. In doing so it is evidently important not to select countries which suit a prearranged roman à thèse. Thus Cuba's record on literacy, health, the poor, compares favorably with (say) Guatemala, its economic performance is outshone by South Korea and Singapore, but it would be far-fetched to imagine that Cuba under another Batista would have equalled such countries as these; many factors are involved other than the economic system. More to the point would be to compare South Korea with North Korea: same people, same historical experience until 1945. In this instance South Korea undoubtedly out-performs the North. In Africa the free-market orientated Côte d'Ivoire has done better, even for its poor, than those of its neighbours who have opted for socialist-type solutions, but again, some African countries have achieved an appalling mess for reasons very far removed from socialism: Ghana and Uganda can serve as examples.

Those who assign to capitalism, or the links with the world market, the responsibility for income inequality, unemployment, regional underdevelopment, etc. should be made to study China. China also illustrates the correctness of the idea advanced by Arthur Lewis: the general level of wages in a given country depends not on the relative productivity of specific workers: thus an Indian or Chinese driver of a five-ton truck is probably as ‘productive’ as his American or British equivalent. It is determined by what he called opportunity-cost, notably (in predominantly peasant countries) the very low productivity and rewards available in agriculture. Thus wages in Shanghai, even in the modern industrial sector, are very low indeed. Were China a capitalist country, this would be the effect of the enormous ‘reserve army of labour’ constituted by 800 million peasants, whose income is much lower than that of Shanghai workers. In China it is a matter of public policy that urban wages be not too excessively far above the levels in rural areas. The effect is not dissimilar.

True enough, any comparison between China and India must note the great inequalities of income in India, and also the fact that the lowest strata of the poor in India are very poor indeed, compared with China. However, as was pointed out by Amartya Sen, India since independence has found it politically indispensable to avoid mass famine, while China suffered acutely from the politically imposed effects of the Great Leap Forward: millions died.

Nor should one ignore the big regional disparities in China, or the very considerable inequalities which existed even before Deng's reform policy was adopted. Also Yugoslavia's regional inequalities persist.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Of course in both these instances there are historical and geographic explanations. All that can be said is that these matters resist speedy solutions under all systems.
To return to the developed world, the Soviet model has come to serve as a negative factor, and Western socialists, and indeed Eurocommunists, have tried to distance themselves from it.
The negative influence of the Soviet example is partly due to the revelations about the Stalin terror and Gulag. But, paradoxically, it was the Stalin period which, with all its horrors, did show a high degree of dynamism, high growth rates, evoking some enthusiasm and commitment from many Soviet citizens as well as foreign observers. It was brutal, it was crude, but they were forcing through a huge industrialization programme, preparing for war, fighting it, eventually winning it. There is, unfortunately, some Stalin-nostalgia in the Soviet Union today, analysed vividly by the emigré Viktov Zaslavsky (1982). ‘Really existing socialism’ has become grey, dull, undramatic, inefficient, more than a little corrupt. The ruling stratum under Stalin was young and faced sizeable risks of purge and execution. People could find little to enthuse about under the Brezhnev gerontocracy; the privileged abused their privileges without fear of punishment, shortages and poor quality contrasted with official claims of successes. Of course, under Stalin, things were in fact much worse. There were indeed horrors, but they were little understood outside the Soviet Union. (Thus the brutalities of collectivization and the famine that followed it were fairly successfully concealed from view.) The result was that the Soviet Union and the ‘socialism’ it represented became for a time a pole of attraction for millions. ‘I have seen the future, and it works’, ‘Soviet communism – a new civilization’, to cite two contemporary judgements. Today the Soviet model no longer impresses or convinces. It is not in chaos, it is not about to fall apart, but it is no beacon, can inspire nobody either in or out of the Soviet Union. And this despite the fact that much has gone wrong in the capitalist West. We will see if the new generation of leaders can restore the lost dynamism.
A few left-wing intellectuals transferred their allegiance to Mao. As was the case with some Western admirers of Stalin's Russia of the Thirties, this allegiance or admiration was based on misunderstanding, on ignorance. The ‘Maoists’ simply did not know about the real Great Leap Forward and its millions of victims, or just what the ‘Great Proletarian Cultural Revolution’ was really about. The post-Mao reaction brought them to their senses. The Yugoslav self-management model too has had its admirers, and indeed its principles are attractive, and will be looked at below. However, grave economic problems have hit Yugoslavia. By no means all of them are connected with the self-management model, but the fact remains that the negative aspects now tend to predominate in observers’ minds. Then there was Poland. The ‘Solidarnosc’ story, in the present context, is one which not only highlights governmental economic ineptitude, but more important, makes spectacular nonsense of the communist claims to represent the workers, or to be the advance-guard of the proletariat.
So, to summarize, socialism is not, at present, a politically attractive slogan, and this despite the quite vigorous efforts of the New Right to destroy ‘consensus-capitalism’. Worse, the immediate political programme of (for instance) Labour's left in Great Britain may be a sure recipe for trouble, reminiscent of the tragic errors of the Allende regime in Chile (which I had the sad experience of witnessing: price control, import controls, large wage increases, the disruption of the normal functioning of the market with no coherent idea of how to replace it).
Democratic socialism, however defined, can come only if the majority of the people are convinced that the old order has outlived itself, that major changes in a socialist direction are urgently needed. In a percipient analysis, S.C. Kolm has noted a repeated tendency: a left-wing government is elected, and its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic policies begin to hurt those middle strata (or middle-class, or left-centre parties) whose votes brought this government to power. The result is a rightward shift of opinion, and either the loss of the parliamentary majority (as in France, in 1937–8, for example) or a successful right-wing coup, as in Chile. Some draw far-reaching conclusions about there not being any democratic road to socialism (although, for example, in Chile there was no left-wing majority in Congress, Allende having been elected on a 'reformist' programme and with some support from left-wing Christian Democrats). Whatever may be the actual or anticipated resistance of the powers-that-be, one can only repeat that democratic socialism requires the support over a prolonged period of the democratic majority – and right now this is not available – except for Swedish-style welfare state social-democracy (which has again won an election in Sweden on a welfare-state programme).

Perhaps Sweden is in fact the model we should study, if what we seek is a programme which a moderate, non-revolutionary, democratic-socialist party ought to 'sell' to the electorate. Yes, it is a high-tax solution, but one which the electorate, at least in Sweden, can be persuaded to prefer to any Swedish translation of Thatcherism. In my book on *Feasible Socialism* (Nove, 1983), I rejected the notion that Sweden is a socialist republic ('and not only because it is a monarchy'), and of course there is a large 'capitalist' sector. But there is no serious current of opinion in Sweden which would support a policy of nationalizing the privately owned enterprises, or other drastic changes of existing arrangements. So if this is in fact the practical policy recipe of moderate-socialism or social-democratic parties in Western Europe, then this might be seen as a medium-term objective. Leaving the term 'socialist' as a distant perspective, just as the official Soviet propaganda now views full communism. Just as the Soviet government does not tell people that they actually intend at any particular date to abolish wages and prices, so a Western socialist party should not be committed to 'the introduction of socialism' as a policy for today. But there should be a longer-term objective. What objective?

For reasons already examined at length, it cannot be the socialism/communism foretold by Marx. Then what can it be? Let us examine this subject, bearing in mind the three points made earlier: what relationship between management and workforce; how do productive units interrelate; and what sort of state can be envisaged – bearing in mind that a state there would and must be, with important functions to perform.

So let us look at self-management. Why has its Yugoslav version lost much of its attractiveness? As already suggested, some of the reasons have little to do with the self-management model as such: centrifugal tendencies in a multi-national state with a relatively weak central authority; unwise policies on interest rates (which have been negative in real terms) and on foreign exchange; lack of any effective control over bank credits, to cite some examples. However, certain lessons can none the less be drawn. One is that self-management is not necessarily desired by the workforce, in the sense that many wish to spend long hours sitting in committee-rooms or studying the firm's accounts. However, the formal responsibility of management to the workforce is an important principle, as is the right of participation, which can be exercised when something goes wrong or feelings run high.

A second point relates to the lack of interest of much of the workforce in the longer term. This is a consequence of that fact that the capital assets do not belong to them, and when they leave they have no saleable asset to dispose of. Their only interest is in the income they can earn. This inclines them to a short-term view, to a desire to increase current income rather than invest in the future. One effect is to increase inflationary pressure.

Thirdly, neither the workforce nor the management has any real responsibility for investment decisions,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
past or present. Suppose they prove disastrous, who is to blame? If indeed the initial investment decision (to set up the firm) was mistaken, and it was taken before there could be a workers' council or the election (appointment) of a manager, why should management or labour be penalized? This is one aspect of a wider problem: that of how to cope with failure under socialism (other than be assuming that it will not occur!).

Fourthly, by making the workforce's incomes dependent on the given enterprise's financial results (subject, to be sure, to a legal minimum), one ensures unequal pay for equal work, and thus a chronic source of tension and discontent. Thus suppose citizen A and citizen B both drive five-ton lorries from Zagreb to Split, but A works for a more successful enterprise than B; they may well receive very different pay. The resultant pressure for higher pay in the financially less successful enterprises is yet another source of inflationary pressure.

Fifthly, Yugoslavia suffers from unemployment. Yet material incentives based upon dividing net revenues among the existing labour force builds in a reluctance to employ extra labour, whenever such employment would diminish the sum represented by net (distributable) revenue per head. In choosing between investment variants, there is for the same reason a tendency to choose the more capital-intensive variant, in comparison with the profit-orientated capitalist or the 'plan-fulfilling' Soviet manager.

For what should be obvious reasons, self-management requires a market. The self-managed units decide what to produce by reference to market criteria, and purchase their inputs by freely negotiating contracts with suppliers. Charles Bettelheim was quite right when he wrote that 'commodity production' (i.e. for exchange) must exist so long as units of production are autonomous and not wholly integrated into the plan. Yet he criticizes Yugoslav-type self-management: the workers do not really control their means of production and the product – the market does. This presupposes the existence of some unrealizable alternative, in which what is done and the acquisition of means to do it are controlled by no outside force at all. Yet needs have to be conveyed somehow, if not through negotiating contracts then via instructions from a superior authority.

Another significant moral to draw from Yugoslav experience relates to regional questions. In a country which, for historical and geographical reasons, has a relatively highly developed north and a backward south, measures to correct these disparities have had little success. Experience elsewhere shows that such matters defy solution in very different systems (for instance, compare Italy's *mezzogiorno*, or the megalopolis problem in such countries as Mexico and Brazil). However, the combination of autonomous 'self-managed' units and centrifugal forces, with the centre in a relatively weak position, tends to perpetuate or even reinforce regional inequalities. Indeed – and Soviet experience with *sovnarkhozy* (regional economic councils) points in the same direction – one might conclude that regional power over enterprises is very likely to result in irrationalities. The reason is clear: a local authority has information about the needs of its locality and, unless prevented, will tend to give them priority to the detriment of other localities, with duplication of investments as yet another undesirable consequence. In other words, if one were to imagine a modern industrial society with complex interregional links, there are two possible logical solutions: central control or enterprise autonomy (the 'enterprise' could, in some circumstances, be large or even, in such cases as electricity supply, a centrally controlled monopoly). If power over resources were given to an authority covering one area, it would divert resources for its own purposes, with potentially disruptive effects.

Finally, one must refer to the very considerable literature, of which Ward's fascinating excursion into

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
‘Illyria’ is the original example (Ward, 1958), which appears to prove that self-managed enterprises, in which the workforce's income depends on that enterprise's net revenue, are of their nature inefficient. Some of the conclusions are irrelevant to the real world. Thus Ward's model shows that it would ‘pay’ the firm to reduce output if prices rose, but this would only be so under the assumption of so-called ‘perfect competition’, in which such considerations as real competition do not enter. For example, in real competition one is concerned not to lose customers to one's competitors, who might not be regained if prices fall, as in future they might. Nor are self-managed enterprises likely to dismiss fellow-workers without some extremely strong reasons. None the less, as already noted, they may choose labour-saving, capital-intensive investment variants even when unemployment is a major social problem. It may be necessary (and it surely is possible) to devise fiscal means to counteract this tendency. As for efficiency, this depends (inter alia) on the attitude of the workforce. Would the sense of participation increase commitment and loyalty, and so the quality of the work effort? These considerations seldom figure in economic analysis (with Albert O. Hirschman an honourable exception). Some unimaginative model-builders would doubtless also conclude that the reluctance of Japanese firms to shed labour is ‘inefficient’, yet any loss can be counterbalanced by the sense of ‘belonging’ that goes with security of employment. A recent study of Israeli *kibbutzim* noted that one finds no resistance there to labour-saving innovations, which can be encountered in private firms, because such innovations do not threaten loss of jobs.

There are lessons to be learnt from the experience of the Mondragon cooperatives in northern Spain. Unlike the Yugoslav enterprises, they pay wages, so that there is an identifiable profit. They also ensure that the workforce has shares in the business (if necessary lending them the money to acquire them), and this also gives them a longer-term stake in its prosperity. It is, however, worth recalling that the Mondragon enterprises function in an area of strong local loyalties, just as the *kibbutz* members are committed volunteers. The outcome may be different with different human material.

Socialists must be aware that there are bound to be problems connected with property ownership and long-term responsibility, involving also risk-taking and the consequences of failure. Where uncertainty exists – i.e. in any conceivable situation – there must be the possibility of failure. A capitalist can go bankrupt, but what of ‘socialist bankruptcy’? One cannot ‘solve’ this question simply by assuming either perfect foresight or perfect planning. The existence of genuine autonomy of decision-making is surely an aim desirable in itself, and freedom necessarily involves both uncertainty and freedom to err, to act in ways not necessarily consistent with the general interest or the national plan.

What, then, could a ‘feasible socialism’ be like? Should the word be redefined? Surely a non-utopian definition of socialist values should be counterposed to the crude laissez-faire ideology of the New Right. Some of the traditional slogans associated with socialism have become deservedly unpopular. There are good reasons to associate nationalization with bureaucracy, satisfying neither the workforce nor the customers. It is in a review in *Radical Philosophy* (Spring 1985) that one can read: ‘A regime devoted to equality in its literal sense would have to be authoritarian, ready to crush inequalities whenever they reasserted themselves, as they inevitably and constantly would.’ The New Right's view of ‘liberty’ may be distasteful, but one must recognize that the aims of equality and freedom can conflict with one another. Socialism cannot be happy with a purely acquisitive society. Indeed such a society would fall apart, for why should civil servants, judges, police officers, not be crude income-maximizers, i.e. behave as most doctors seem to do in America? Yet acquisitiveness is not a value to be disparaged, the vast majority of citizens do have material aspirations. Thus a conscientious doctor does his best for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
his patients, even if they cannot pay an economic fee, but he or she is also not averse to acquire a country cottage and go on holiday to Greece. Furthermore, at least since the days of Adam Smith it has been rightly noted that there are worse ambitions than making money: the men who, in the process of competing for power, sent their comrades to be shot in cellars were not seeking to maximize profits. What is to be sought is a balance between (enlightened) self-interest and a sense of social responsibility. Inevitably this differs as between individuals.

Individuals also differ greatly in what might be called ‘producers’ preferences’. Some like to be independent innovators, others prefer routine. Some gladly take responsibility, others prefer to avoid it. Some opt for life in a commune or *kibbutz*, others would be very unhappy there. While Marx's vision of a universal Man is a fantasy, it is not at all a fantasy to provide both for variety and for the opportunity to change one's specialization if the spirit so moves one. A socialism based on one economic model might be a sort of procrustean bed for a sizeable part of the population. (Imagine, for example, *compulsory* communal living!) Hence it seems desirable to redefine ‘socialism’ as a mixed economy: enterprises large and small, many if not most self-managed or cooperative, with some private enterprises too. If the private sector does not play a dominant role, its existence should be consistent with a sensibly defined socialism; otherwise its suppression would be the constant task of a ‘socialist’ police (unless, of course, it proves not to be needed, in which case ‘privateers’ no more require to be banned than to outlaw private water-carriers when everyone has tap water). A major objective would be not only to ensure variety of choice of occupations, but also work for all, when unemployment is in danger of becoming a major social curse. Only in ideological textbooks of economics do labour markets automatically clear. One must anticipate the need to take job-creating action. One must also anticipate that freedom to organize involves freedom to form not only political parties but also interest groups which will press for additional resources. Since money will undoubtedly continue to exist, it would be possible to issue too much of it in the face of pressures, so inflation (and some species of monetarism) will not just go away. Freedom of choice implies both a market and competition, both in consumers’ goods and producers’ goods and services, though there must also be some large-scale natural monopolies (such as electricity, water, public transport), where responsibility of management to the users is as important as its responsibility to its workforce.

Mises, Hayek, and later on also Friedman, have argued that efficiency in resource allocation is impossible under socialism. At a formal level they were answered by Lange, Lerner, Dickinson, but there were and are major practical obstacles in realizing their socialist models, which are anchored (as are so many of the neoclassicals’) in static equilibrium assumptions, and it is unclear why either the central planning board or the managers in Lange's model should act out their parts in the prescribed manner. It should be admitted that the absence of (or severe limits on) a real capital market can cause inefficiencies, that rewards for risk-taking and innovation may well sit uneasily with social or state ownership of capital assets. Nor is this all. Kornai, in his Dublin lecture (Kornai, 1985) pointed to contradictions between the requirements of efficiency and socialist ethics. But the world is full of contradictions, and one usually arrives at some species of compromise; ‘maximization’ in terms of just one objective function can seldom be encountered in really existing societies (a fully fledged and devoted ‘profit maximizer’ would probably suffer a nervous breakdown, if not already dead of cardiac arrest). Mises and company are right to insist that economically meaningful prices are needed, wrong to assert that socialist prices cannot be meaningful (though today's Soviet prices are indeed irrational, reflecting neither use-value nor relative scarcity). But it must be emphasized how far the contemporary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Western system is from the free-market model of the textbooks. Thus in his challenging ‘Profits without production’, Seymour Mellman notes and deplores the narrow concentration on short-term profits, by executives who have no long-term commitment to their corporation (on average they move to another one within five years or so). Current uncertainties about prices, interest rates, inflation, are hardly conducive to ‘rational’ long-term investment decisions. Too often critics of socialist economics (with its imperfections) implicitly compare it with a Chicago utopia, which is in its own way as unreal as a marxist one. Perfect markets and perfect plans are equally utopian.

But in the end much will depend on the ability of contemporary capitalism to surmount its many problems, not least that of mass unemployment and ecological decline (acid rain, deforestation, over-fishing, etc.). The masses will not opt for a different system unless faced with the bankruptcy of the existing one. To repeat, it was Marx who wrote that no mode of production passes from the scene unless and until its productive potential is exhausted. With Soviet-type socialism seen as obsolete, in contradiction with the forces of production, it offers no alternative model. A great deal remains to be done to revive socialism as an aim worthy of effort and sacrifice.

See Also

- command economy
- economic calculation in socialist countries
- Engels, Friedrich
- Marx, Karl Heinrich
- material balances
- planning
- socialism (new perspectives)

Bibliography

Bahro, R. 1977. *Die Alternative: zur Kritik des real existierenden Sozialismus*. Cologne: Europäische Verlaganstalt.

Bettelheim, C. 1968. *La transition vers l’économie socialiste*. Paris: F. Maspero.

Brus, W. 1980. Political system and economic efficiency – the East European context. *Journal of Comparative Economics* 4(1), 40–55.

Castoridis, C. 1979. *Les carrefours du labyrinthe*. Paris: Editions du Seuil. Trans. K. Soper and M.H. Ryle as *Crossroads in the Labyrinth*, Brighton: Harvester, 1984.

Feher, F., Heiler, A. and Markus, G. 1983. *Dictatorship Over Needs*. Oxford: Basil Blackwell.

Gatovsky, L., ed. 1967. *Zakonomernosti i puti sozdaniia materialnotekhnicheskoi bazy kommunizma*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(On legality and the means of creating a material-technical base for communism). Moscow: Akademii Nauk SSSR.

Goroz, A. 1980. *Adieux au prolétariat*. Paris: Editions Galilée.

Grossman, V. 1985. *Life and Fate*. Trans. from the Russian by R. Chandler, London: Collins Harvill.

Kolakowski, L. 1976. *Main Currents of Marxism: its rise, growth and dissolution*. Vol. 1, trans. from the Polish by P.S. Falla, Oxford: Clarendon Pres, 1978.

Kornai, J. 1971. *Anti-Equilibrium. On economic systems theory and tasks of research*. Amsterdam and London: North-Holland.

Kornai, J. 1985. The dilemmas of the socialist economy. Geary Lecture, Dublin, 1979. In *Contradictions and Dilemmas*, ed. J. Kornai, Corvina: Kner Printing House.

Kuznets, S. 1941. *National Income and its Composition, 1919–1938*. New York: National Bureau of Economic Research.

Lenin, V.I. 1962. *Sochineniia* (Works). 5th edn, Vol. 36, Moscow.

Marx, K. 1962. *Marx–Engels Works* (Russian). Vol. XXVIII, Moscow.

Mészáros, I. 1972. *Marx's Theory of Alienation*. 3rd edn, London: Merlin Press.

Nove, A. 1983. *The Economics of Feasible Socialism*. London: Allen & Unwin.

Roland, G. 1985. *Valeur d'usage chez Karl Marx*. Brussels.

Schumpeter, J. 1976. *Capitalism, Socialism and Democracy*. 5th edn, London: Allen & Unwin.

Shanin, T. 1984. *Late Marx and the Russian Road*. London: Routledge & Kegan Paul.

Ward, B. 1958. The firm in Illyria: market syndicalism. *American Economic Review* 48, September, 566–89.

Wright, E.O. 1979. *Class Structure and Income Determination*. New York: Academic Press.

Zaslavsky, V. 1982. *The Neo-Stalinist State: class, ethnicity, and consensus in Soviet society*. New York: Sharpe; Brighton: Harvester.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Nove, Alec. "socialism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000173> doi:10.1057/9780230226203.1568

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

socialist calculation debate

David M. Levy and Sandra J. Peart

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The socialist calculation debate centred on two questions: could a socialist society use planning to replicate the performance of a capitalist society? And could socialism improve on capitalism? Many economists answered both questions in the affirmative until shortly before the demise of the planned economies of the Soviet bloc. The calculation debate was about the interactions among models, motives and incentives. In hindsight, it is evident that the incentives in a planned economy are such that planners are not motivated to promote the public interest.

Keywords

Barone, E.; Barter; capitalism; consumer sovereignty; Dickinson, H.D.; Dobb, M.H.; Eucken, W.; Godwin, W.; Hayek, F.A. von; improvement thesis; incentive compatibility; information aggregation; *keiretsu* ; Lange, O.R.; Lerner, A.P.; Malthus, T.R.; marginal productivity theory; market socialism; Mill, J.S.; Mises, L.E. von; planning; price mechanism; principal and agent; replication thesis; socialism; socialist calculation debate; Soviet Union, economics in; Stakhanovism; Unemployment

Article

The calculation debate centres on two issues. First, is it possible for a socialist society, with common ownership of the means of production, to use planning to replicate the performance of a capitalist society, with private ownership of these means of production? This is the ‘replication’ thesis. Second, is it possible to do ‘better’ than replication? This is the ‘improvement’ thesis. Both theses were widely accepted by economists until shortly before the planned economies passed into history with the demise of the Soviet Union (Lavoie, 1985).

We state the problem in terms of possibility in order to emphasize an ambiguity in the debate. What do we conclude when we observe a centrally planned economy that is unlike any known capitalist economy? Has the replication thesis failed? Not necessarily. Confronting such a situation, Drewnowski

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(1961) combined the replication thesis with a (planner's) revealed preference axiom to argue for the 'improvement' thesis. Some 36 years later Drewnowski (1997) was candid about the profession's failure in foreseeing the collapse of the planned economies:

> an event which will forever dominate the history of the second half of the 20th century, the downfall of communism, or, more correctly, the disintegration of Soviet-type political and economics systems. The striking feature of that great upheaval was that it came as a virtual surprise … the shock of the collapse of seemingly invincible systems was unique in history. (Drewnowski, 1997, p. 919)

Though it was not always recognized as such at the time when the debate was open, in hindsight it is clear that the calculation debate was about the interactions among models, motives and incentives. The question was whether a model that is a 'true' description of a capitalist society might be the basis for a planned economy that satisfies the replication thesis. But to answer in the affirmative, we need to specify the motives of agents and the incentives they confront. If we agree that the model describes what we want to attain, then it can serve as the basis of the public interest. We can stipulate that public interest has some motivational force for everyone and then ask whether the incentives in a planned economy are such that planners will seek the public interest. Again, with hindsight, we can answer that they are not.

As a guide to how the calculation debate played out, we rely on the model-theoretic insights of Walter Eucken. A philosophical economist in the liberal tradition who survived great danger in Germany during the Hitler era (Gerber, 1994), Eucken was brought by F.A. Hayek to the founding meeting of the Mont Pèlerin Society in 1947 as evidence of a surviving liberal tradition and so a hope for a self-governing Germany. In 1948, Eucken's associates Ludwig Erhart (Gerber, 1994, p. 31) who as director of the Office of Economic Administration used the power entrusted in him by the American occupiers to end the centrally administered economy (Mendershausen, 1949).

Eucken (1948) asked whether the same model could be used to describe both a centrally administered and a decentralized market economy. He noted that J.S. Mill argued that two models are required, one for each institutional setting. There is no reason to believe, Mill claimed, that the characters of the people in different institutional settings will be the same over time, Mill (1844, VI, p. x, § 3). Socialism, which may not be feasible now, might become so in the future. Therefore, there is no reason to believe that the same model can encompass both systems. Eucken supported the two-model view, not by appealing to differences in the characters of the agents over time, but by appealing to the economist's understanding at a given moment in time. We will return to this claim below.

Von Mises on Mill

The calculation debate was launched in 1920 by a brief article in which von Mises challenged socialist economists to apply their arguments against market economies to their own systems. On the assumption that the collectivist state would allow private ownership of consumer goods, and thus private exchange, von Mises saw no particular problem with the allocation of consumer goods in such a setting (von Mises, 1920, pp. 90–2). The problem comes with collective ownership without exchange, since there

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
will then be no prices. If prices do not reflect opportunity costs, no ‘rational’ economic system is possible.

Von Mises's 1920 paper cites a small number of socialists, none of whom he respected. That situation changed in *Socialism: An Economic and Sociological Analysis* (German editions 1922 and 1932; English edition 1936), in which he draws the reader's attention to John Stuart Mill's thoughts on socialism:

> The writer who has occupied himself most thoroughly with this problem is John Stuart Mill. All subsequent arguments are derived from his. … They have provided for decades one of the main props of the socialist idea, and have contributed more to its popularity than the hate-inspired and frequently contradictory arguments of socialist agitators. (Von Mises, 1936, pp. 154–5)

It is here that von Mises confronts the second model of economics, the model of socialism that Mill made for people of the future, somewhat like us only a little more sensitive to the opinion of others. Mill explained how a successful communal system must rely upon a different motivational package from that which characterizes market economies. The problem of the commons was central to classical political economy. T.R. Malthus's 1798 *Essay on Population* had addressed William Godwin's proposal of replacing private property with commons. Malthus asked why individuals would renounce or defer marriage and childbearing if someone else were to support them in case of difficulty. For Mill, a workable commons would require a great deal of motivation from the desire for approbation. For a system of equality to hold, the commoners must know that they cannot support each other's children without mass misery. They must understand the self-restraint that characterizes a market economy and, without any incentive other than a desire for praise and an aversion for blame, act on the basis of that understanding (Mill, 1848, II, p. 1, § 3). When embodied in the public understanding, the model must be self-motivating. We will return to the question of self-motivating models below.

Von Mises (1936) surmised that Mill failed to link material rewards with effort in a market economy because he wrote before the advent of marginal productivity theory (1936, p. 155). Von Mises acknowledged the problem of workers who are not paid by the piece – the partial basis of Mill's defence of socialism – and suggests this is the fault of the worker:

> Doubtless the individual working for a time wage has no interest in doing more than will keep his job. But if he can do more, if his knowledge, capability and strength permit, he seeks for a post where more is wanted and where he can thus increase his income. It may be that he fails to do this out of laziness, but that is not the fault of the system. (Von Mises, 1936, p. 155)

To deal with public opinion motivation, von Mises denies Mill's hypothesis that human nature might alter and claims instead that motivation will be the same under both settings:

> It is not impossible that under Socialism the public spirit will be so general that disinterested devotion to the common welfare will take the place of self-seeking. Here

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> Mill lapses into the dreams of the Utopians and conceives it possible that public opinion will be powerful enough to incite the individual to increased zeal for labour, that ambition and self-conceit will be effective motives, and so on.
>
> It need only be said that unfortunately we have no reason to assume that human nature will be any different under Socialism from what it is now. And nothing goes to prove that rewards in the shape of distinctions, material gifts, or even the honourable recognition of fellow citizens, will induce the workers to do more than the formal execution of the tasks allotted to them. Nothing can completely replace the motive to overcome the irksomeness of labour which is given by the opportunity to obtain the full value of that labour. (Von Mises, 1936, p. 157)

Von Mises continues to suggest that 20th-century socialists object to this argument by pointing to heroic acts as counter-examples. He confesses that he does not understand heroes, people who are ready to die for their principles, driven purely by their ‘union of will and deed’ (1936, p. 158). He acknowledges that the fate of civilization itself depends on the heroic (1936, pp. 157–8) – but for the purposes of economic analysis, he is content to deal with the ordinary. Although von Mises was disinclined to pursue the heroic as a motivational force under socialism, there is an important sense in which the Soviet Union seems to have attempted to function on the basis of the heroic: the *Stakhanovite* (someone who joyfully over-fulfils the production quota), a Soviet-era concept of consequence (Kotkin, 1995, pp. 207–15), deserves attention.

Mill, too, was concerned with the character of the majority in any institutional setting; hence, to learn about social transformation one needed to apply the inverse deductive method, Mill (1844, VI, p. x, § 4). The idea that people might develop over time was ruled out of consideration by Hayek (1935a) in his summary of the history of the discussion:

> John Stuart Mill, in his autobiography, numbered himself among the socialists, because he believed that certain ideas would be realized in the distant future. In this connection Cairnes pointed out that true socialism does not consist of a body of ideas which can only be realized if human nature and the conditions of human life are radically transformed. Socialism subsists in the recommendation of certain modes of action and in the utilization of the authority of the state for particular purposes. This appears to me also as the correct view. (Hayek, 1935a, p. 47)

The replication thesis

Fred Taylor's 1928 Presidential Address to the American Economic Association sets the stage for market socialism. He starts with a model of the private market economy:

> First, on the basis of a vast complex of institutions, customs and laws, the citizen adopts a line of conduct which provides him with a money income of greater or less volume. Secondly, that citizen comes on the market with said income demanding from those

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> persons who have voluntarily assumed the rôle of producers, whatever commodities, he, the citizen chooses. Thirdly, the producers promptly submit to the dictation of the citizen in this matter, providing always that said citizen brings along with his demand entire readiness to pay for each commodity a price equal to the cost of producing that commodity. (Taylor, 1929, p. 1)

The ‘correct general procedure’ for market socialism is to maintain consumer sovereignty. Taylor thus argues that the planners must use the market model in the new institutional setting:

> (1) the state would assure to the citizen a given money income and (2) the state would authorize the citizen to spend that income as he chose in buying commodities produced by the state – a procedure which would virtually authorize the citizen to dictate just what commodities the economy authorities of the state should produce. (Taylor, 1929, p. 1)

To implement this under socialism, Taylor proposes ‘trial and error’. The idea is that the planners start somewhere, and then if managers have an algorithm which implements the market model of capitalism, they adjust factor valuations on the basis of shortage or surpluses (Taylor, 1929, p. 8). Equilibrium follows. In H.D. Dickinson's judgement, Taylor's statement of the replication thesis withstood later criticism:

> Taylor has the distinction of being the first writer to point out the way to answer Professor Mises's attack on socialists. Moreover, he answers in anticipation the later criticisms of Professors Hayek and Robbins, that had not been published when his paper first appeared. (Dickinson, 1938, p. 532)

The improvement thesis

Although market socialism literature took off from Taylor, Barone had earlier described an iterative solution to the problem of maximizing consumer surplus, for example, ‘maximum collective welfare’ (Barone, 1908, pp. 270–2), and he discussed the problem of how to determine productive coefficients experimentally (1908, pp. 287–8). Further, with his argument that one can do better than replicate the capitalist solution in the case of multiple prices, he announced the improvement thesis:

> Hence, when the first area is larger than the second it is possible that multiple prices may be consistent with increase welfare for the community. And as such a proceeding is more possible when production is socialized, this is in reality a sound argument in defence of socialized production, in certain cases, when such conditions are proved to exist. (Barone, 1908, p. 283)

As the literature developed, the Taylor solution concept was refined and extended by a host of authorities. Responding to von Mises and developing Taylor's ideas without knowledge of Barone, H.D.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Dickinson (1933, p. 29) added that socialist markets would eliminate or reduce the divergence between private and social costs. For Dickinson, the improvement thesis was made in the welfare metric of orthodox economics characterized by consumer sovereignty (Hutt, 1940).

Maurice Dobb (1933), who had previously supported market socialism entailing consumer sovereignty, abandoned this position in 1933 (1933, p. 591). Instead, Dobb (1933) asked whether the state might actually create tastes (1933, p. 532) and why the rules of equating at the margin would apply to a state which can create technical progress (1933, pp. 596–7). There followed one of the most interesting and important contributions to the debate, Abba Lerner's defence of market socialism. Lerner mentions half a dozen mistakes in Dickinson's algorithm and counters that those in power do not wish to reconcile markets with socialism:

> The cautious guardians of socialism are for retaining the superior strategic fastnesses of simple faith. The heresy must be eradicated. Against the ‘Dickinson’ thesis is the raised the antithetical slogan: ‘The categories of capitalist economy are inapplicable to the socialist society.’ (Dobb, 1933, p. 52)

Hayek entered the debate at this point with his edited collection *Collectivist Economic Planning*, which contained translations of Barone's and von Mises's original papers, as well as Hayek's summary of the discussion and his restatement of the issues in terms of the dynamics of the problem. Here the questions posed were: supposing that a solution were found to a static problem, what incentives does the socialist management have to solve the problem correctly? Why would the model have motivational force? (Hayek, 1935b)

Hayek also noted the ambiguity in the notion of ‘impossible’ in his discussion of Taylor and Dickinson. In his account, their papers

> were directed to show that on the assumption of complete knowledge of all the relevant data, the values and the quantities of the different commodities to be produced might be determined by the application of the apparatus which theoretical economics explains the formation of prices and the direction of production in a competitive system. Now it must be admitted that this is not an impossibility in the sense that it is logically contradictory. But to argue that a determination of prices by such a procedure being logically conceivable in any way invalidates the contention that it is not a possible solution, only proves that the real nature of the problem has not been perceived. (Hayek, 1935b, pp. 207–8)

The great step in algorithm development was then taken by Oskar Lange (1936), with comments and corrections from Lerner (1936). The market algorithm now improved upon the competitive equilibrium by solving for a social optimal in which the same consumer preferences pertained as in the capitalist economy. Market socialism became the Lange–Lerner model.

Hayek's (1940) review of the market socialism advanced by Taylor, Dickinson, and Lange questioned the details of the algorithm, pointing out many gaps in the explanations. Hayek's puzzle as to whether an

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
algorithm based on a model of price-taking competitive equilibrium was appropriate would later blossom into perhaps the most influential single article in the debate, his 1945 'Use of Knowledge in Society' paper. The problem for the capitalist economy was reformulated as a division of knowledge. Prices reflect the decentralized knowledge possessed by all the participants in a market economy. Markets were now viewed in terms of information aggregation. Hayek also notes the importance of the socialist market economists, singling out Lange and Lerner for particular praise, and suggesting that the political aspect of the debate is now at end, leaving only methodological issues to resolve.
The entire focus of the calculation debate was challenged again in 1961 when Drewnowski revived Dobb's position and asked why a socialist economy would wish to mimic or modestly improve upon a market. Why would not the planned economy find a political optimum as distinct from the consumer sovereignty constrained optimum? Indeed, if one looked at existing socialism in the Soviet Union, very little attention was paid to consumer wishes.

Proving the algorithm correct

The line of defence on the replication thesis consisted in taking a model of a private market economy and turning it into an algorithm that might be given to various functionaries of a planned economy. The defences differed on the basis of the model itself, or the algorithm, and in the latter instance the issue of possibility came into play. The question that emerged next was whether the solution of the model and algorithm would be the same. Specifically, would the agents in the model of capitalism make the same choices as those in the planned economy if, in fact, the incentives facing them differed under the two institutional settings?
The debate now moved from theoretical issues of information generation to understanding behaviour in socialist economies in the light of the incentives facing the socialist agents. In 1936, Durbin discussed the responsibility of economists to get things right, arguing that the economist certifies the model's correctness, but not the algorithm's correctness. There is no guarantee that it will be executed as ordered. That is someone else's responsibility:

> "The calculations will not be made." "The mobile resources will be unwilling to move." "The production units that ought to expand will refuse to do so." All these criticisms may or may not be true. They may or may not be the real problems of policy. But they are not problems that the professor of economic theory is competent to discuss. They are problems of social behaviour. They can only be resolved, if they can be resolved at all, by a comprehensive sociological and principally psychological analysis. (Durbin, 1936, p. 678)

Later, reviewing the Taylor and Lange contributions, Dickinson points out the weakness in the algorithm. He focused on whether agents in the socialist algorithm would behave the same way as agents in the capitalist model:

> Mr. Lange … speaks of rules that the Central Planning Board would have to impose on the managers of socialist enterprise. But what guarantee would there be that these rules

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> would, in fact, be observed? Pure equilibrium theory makes it possible to deduce the rules that socialist managers ought to follow: to determine whether, or under what circumstances, socialists managers are likely to follow them requires data that only realistic social studies can provide. (Dickinson, 1938, p. 533)

Is market socialism incentive compatible?

For capitalism and socialism to be described by one model, we require common technology, common preferences and common incentives. At the simplest level we require all three to have a price system. Could a strongly planned economy actually move to the pseudo-planning of market socialism? Eucken (1948) asked whether Lange's solution is incentive compatible in a centrally administered state:

> Would it not perhaps have been possible to graft prices on to the controlling mechanism of the centrally administered economy in the following way? The central administration would have distributed consumption goods by rationing, as well as fixing prices. With regard to consumption goods, demand and supply would have been equated by rationing. But with regard to the factors of production, there would have been no rationing. Entrepreneurs would have applied for these to the state authorities. The factors would have been priced, and then these prices adjusted according to the extent of demand. By this adjustment of prices would not demand and supply have been possible? In this way, the German authorities would have been proceeding in accordance with proposals outlined by, for example, O. Lange. Wouldn't it have been possible to follow this proposal? (Eucken, 1948, p. 93)

Not surprisingly, given what we know today, his answer was an emphatic ‘no’:

> This method of control was out of the question for the central administration, for it would have meant to some extent letting the control of the means of production – in this case leather or iron – out of its hands.

Competition means the end of central authority:

> Competition can be used to improve efficiency, but as a mechanism of direction for an important section of the economy it cannot be applied without abdication of the central authority. (Eucken, 1948, p. 94)

Eucken claimed that the outcomes, as well as the language used to describe outcomes in the two institutional settings, will differ. The critical term is ‘unemployment’. In an institution-free, Post-Keynesian economic model, one could draw a single production possibility curve for any society. Without unemployment, all would attain the frontier (for example, Samuelson, 1948, p. 20). Eucken notes that Keynesian unemployment, which is both privately and socially costly, would not be a problem

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in a centrally administered economy:

> every worker can be taken on regardless of costs. In an exchange economy, workers are dismissed because there exists a measure of scarcity with regard to single units … Workers are dismissed if the return resulting from their employment does not cover the costs. … even it is estimated that the costs of employing several thousand workers on road construction are not covered, the central administration does not have to cut the work short. In these conditions full employment is always attainable. (Eucken, 1948, p. 179)

But without prices to allocate resources, we revert to barter. Here is Eucken and Meyer's (1948) description of the barter economy that will emerge in the centrally administered economy:

> In dirty, overcrowded, and unlighted trains they travel into the farm areas which are a hundred miles away in order to get from the farmers some food in exchange for part of their city rations, part of their wages in kind, if they receive such, or simply for other possessions which they still retain. They collect beechnuts from the forest and leftover ears of grain and potatoes on the harvested fields. In their free time or during their vacations they cut peat, collect wood in the forests, cultivate a small vegetable plot, or search for rabbit food. Housewives spend uncounted hours in lines before stores, in lines before distribution offices for ration coupons, and in lines before various other government offices. (Eucken and Meyer, 1948, pp. 56–7)

The economic consequence of barter is ‘misery’ and risks ‘death by starvation’:

> From the economic point of view, such extra work is senseless waste. From the point of view of the individual German, however, it is exceedingly important because it saves him from ultimate misery and frequently even from death by starvation. (1948, pp. 56–7)

What puts society in this interior of the production possibility set? The ‘unemployment’ problem in capitalism maps to the ‘barter problem’ in planned economies because a move from planning to markets is not in the interests of those who hold state power.

The consequences of the calculation debate

The major principles textbooks of the post-Second World War era (Elzinga, 1992) drew from the planned economies’ lower consumption–income ratio the inference that they were growing faster than market economies. In doing so, they supposed that ‘unemployment’ had the same meaning in both systems (Samuelson, 1970, p. 3; Lipsey and Steiner, 1975, p. 899; McConnell, 1963, p. 751). As a consequence, his insight was temporarily obscured. The realization that planning entailed barter led to the abolition of the German centrally administered economy, but this insight was lost until the Soviet economies were near collapse. Then it became all too obvious that the disequilibrating prices and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
pervasive shortages were the result of planners' pursuit of private self-interest as opposed to the public good (Levy, 1990; Shleifer and Vishny, 1992).

Learning from the Soviet failure

The surprising collapse of the Soviet economies occasioned a useful series of discussions that attempted to understand both the failure of the Soviet economies and the failure of Western economists. It is appropriate that von Mises's and Hayek's works were re-read with considerable care by those on all sides of the debate.

The heroic in the Soviet economy

As noted above, von Mises rejected the possibility of socialism motivated by public opinion. The fact that heroic motivation is a puzzle for him may reflect the fact that economists have lost contact with the classical economists' discussion of sympathetic motivation and the desire for approbation (Peart and Levy, 2005). For Mill, the strength of sympathetic motivation measures a civilization. The willingness of Americans to die because of their obligation towards the enslaved indicated to him the superiority of American civilization (Peart and Levy, 2005). For Mill, the desire for approbation is sufficiently strong that anti-social behaviour is attenuated because an individual wishes to avoid the disapproval that follows when he or she imposes a cost on society. Socialism on a small scale is then unproblematic to the extent that a socialist community has tight consensus concerning costly behaviour. If one disagrees with such proscriptions, one can leave that society. Mill's dissent from large-scale socialism arose from his concern that it would be inconsistent with diversity of opinion: there would be no place to go to.
Scholars have asked why the Soviet political authorities ignored the advice of expert engineers to insist upon the heroic. A belief in the efficacy of heroic motivation might explain the political appeal of the gigantic Soviet engineering failures such as the Great Dnieper Dam, the Steel City of Magnitogorsk and the White Sea Canal (Graham, 1993; Kotkin, 1995). When his engineers asked the authorities to consider what was feasible, Stalin is quoted as having responded: 'There are no fortresses that Bolsheviks cannot storm' (Graham, 1993, p. 42).
In his study of Soviet engineering failures Graham puzzles over why the political leadership gave up the feasible and risked the certainty of the ordinary for a chance at the heroic. Such a choice makes no sense if one is thinking of the incentives of a market economy with a democratic government. But if socialism itself is justified by risking the seemingly impossible, then the political calculus seems inescapable.

> Had all the costs, social and economic, of the Dnieper dam been considered more carefully, and had the benefits of a single enormous hydroelectric power plant been weighted against those of several small ones, including thermal power plants, a different decision probably would have been made. These alternatives, now quite obviously more desirable, were outlined by Russian engineers during the early planning stages. The final decision to go ahead with the giant dam was based not on technical and social analysis but on ideological and political pressure. Stalin and the top leaders of the Community Party wanted the largest power plan ever built in order to impress the world and the Soviet

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> population with their success and that of the Communist social order. (Graham, 1993, p. 52)

If industrial development is viewed as if it required wartime heroics then, on the basis of American experience of the Civil War, one might well expect a mix of volunteers and conscripts. And this, of course, was observed:

> One of the characteristics of industrialization under Stalin was the coexistence of volunteer and forced labour, of heroic self-sacrifice and violent coercion. (Graham, 1993, p. 59)

Heroic motivation also provides some insight into the speed of the Soviet collapse. If people chose socialism over ordinary capitalism with the hope or expectation of the heroic, then, when the 'heroic' was revealed as a sham, there was no reason to believe that the public opinion that supported the system would hold. Events like the explosion at Chernobyl – 'a product of the standard Soviet industrialization policy that emphasized gigantic projects over smaller ones' (Graham, 1993, p. 90) – could no longer be hidden.

The 'impossibility thesis'?

At the beginning of this article we distinguished between the 'replication' thesis and the 'improvement' thesis to identify the source of ambiguity in the question 'Is socialism possible?' Let us return to that theme in the context of a very simple question. Was it possible for the Soviet system to create projects which could have been deemed efficient *ex post* by the standard calculus? Let us be precise and define the heroic as something valued for its own sake. Soviet engineering for instrumental, non-heroic projects, for example military hardware, certainly passes any market test for efficiency. But when the political incentives were such as to view the project as an end it itself then, even when the *ex post* efficient projects were feasible from an engineering point of view, they were not selected by the political authorities. Thus, although efficient projects were possible, in the sense that anything feasible is 'possible', they were not observed and never existed. The debate between Caplan (2004) and Boettke and Leeson (2005) over the 'impossibility' of socialism struggles with this ambiguity, with the question of whether the Soviet Union failed because of the 'impossibility' of calculation or because of the perverse incentives of its rulers.

There is another way in which the 'impossibility' of socialism is ambiguous. Von Mises's argument asserts that a socialist economy cannot replicate the performance of a market economy. From this, can we infer that socialist economies could not exist with the consent of the participants? Surely not. There may well be benefits from socialism that over-compensate the material losses. This would seem an odd reading of von Mises if one supposes, as both sides of the Caplan–Boettke and Leeson debate do, that von Mises is engaged in an argument with Marx and his followers. It is not so odd if he is arguing with Mill's socialism.

Is socialism robust?

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The attractions of socialism to disinterested scholars are evident in Roemer's efforts to re-imagine a socialism that can be defended after the fall of the Soviet Union. In a remarkable display of historical rethinking, Roemer works through the Lange–Hayek argument giving Hayek the better of each question (Roemer, 1994, p. 30). This is not just on technical economic grounds.

> Hayek claims that Lange never justifies disallowing market determination of industrial prices in his proposal along with consume goods' prices and wages. Indeed, Lange actually does but the justification seems weak if not wrong. He says that disequilibrium in industrial prices is very costly in the economy, since these prices determine the prices of all other goods, and that the CPB [Central Planning Board] can find the equilibrium faster than the market can. I am puzzled by this. It seems that perhaps Lange feared that, were he to allow the market to determine all prices in his model (except the interest rate), he would be giving up too much. As Hayek notes … the Lange proposal already makes great concessions to those who opposed pervasive planning, perhaps Lange believed it would not have been politically wise to go further. (Roemer, 1994, pp. 30–1)

Roemer identifies the problem with the Soviet Union as the incentives of the principal–agent sort (Roemer, 1994, pp. 35–9) and he argues that this can be attenuated by allowing democratic competition (Roemer, 1994, pp. 39–40). A model of the political process is proposed in which equilibrium brings with it a weighted average of the utilities of the poor and the rich (Roemer, 1994, p. 65). To deal with investment decisions, Roemer (1994) proposes the creation of institutions akin to the Japanese *keiretsu* (a group of interrelated companies, both in terms of products and owners, which is centred around a bank) sufficiently independent of the state so as not to allow politician–manager interactions (1994, p. 74).

Roemer's model of the political process is not the only one available. A long line of formal voting models worry about the results of changing the sequence of votes. If an agenda-setter can control the voting sequence he is able to force the results (for example, McKelvey, 1979). Faith in the performance of *keiretsu* may have declined somewhat from the early 1990s. Perhaps we should give some thought as to what can go wrong with a model of socialism as compared with a model of capitalism (Levy, 2002). If modellers are putting forward proposals out of sympathy for extra-scientific goals (Peart and Levy, 2005), then the arguments about model uncertainty in the context of least favourable priors take on a new urgency (Brock, Durlauf and West, 2007).

Eucken's argument that moving from a planned economy to market socialism is not incentive compatible does not, of course, preclude moving from a capitalist economy to market socialism. The question is whether that is where society would stay. If market socialism were to be judged a failure, would we return to market capitalism or move to less liberal institutions (Levy, Peart and Farrant, 2005)?

See Also

- Hayek, Friedrich August von

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Lange, Oskar Ryszard
- Lerner, Abba Ptachya
- Mises, Ludwig Edler von
- planning
- socialism

Bibliography

Barone, E. 1908. The ministry of production in the collectivist state. In *Collectivist Economic Planning*, ed. F.A. Hayek. London: Routledge, 1935.

Boettke, P. and Leeson, P. 2005. Still impossible after all these years. *Critical Review* 17, 155–70.

Brock, W.A., Durlauf, S.N. and West, K.D. 2007. Model uncertainty and policy evaluation: some theory and empirics. *Journal of Econometrics* 136, 629–64.

Caplan, B. 2004. Is socialism really ‘impossible’? *Critical Review* 16, 33–52.

Dickinson, H.D. 1933. Price formation in a socialist community. *Economic Journal* 43, 237–50.

Dickinson, H.D. 1938. On the economic theory of socialism. *Economic Journal* 48, 531–33.

Dobb, M.H. 1933. Economic theory and the problem of a socialist economy. *Economic Journal* 43, 588–98.

Drewnowski, J. 1961. The economic theory of socialism: a suggestion for reconsideration. *Journal of Political Economy* 69, 341–54.

Drewnowski, J. 1997. Review of the roundtable talks and the breakdown of communism. *Europe–Asia Studies* 49, 919–20.

Durbin, E.F.M. 1936. Economic calculus in a planned economy. *Economic Journal* 46, 676–90.

Elzinga, K.G. 1992. The eleven principles of economics. *Southern Economic Journal* 58, 861–79.

Eucken, W. 1948. On the theory of the centrally administered economy: an analysis of the German experiment tr. T.W. Hutchison. *Economica* 15, 79–100 and 173–93.

Eucken, W. and Meyer, F.W. 1948. The economic situation in Germany. *Annals of the American Academy of Political and Social Science* 260, 53–63.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gerber, D.J. 1994. Constitutionalizing the economy: German neo-liberalism, competition law and the 'new' Europe. *American Journal of Comparative Law* 42, 25–84.

Graham, L.R. 1993. *The Ghost of the Executed Engineer: Technology and the Fall of the Soviet Union*. Cambridge, MA: Harvard University Press.

Hayek, F.A. 1935a. The nature and the history of the problem. In *Collectivist Economic Planning*, ed. F. A. Hayek. London: Routledge.

Hayek, F.A. 1935b. The present state of the debate. In *Collectivist Economic Planning*, ed. F.A. Hayek. London: Routledge.

Hayek, F.A. 1940. Socialist calculation: the competitive solution. *Economica* 7, 125–49.

Hayek, F.A. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

Hutt, W.H. 1940. Economic institutions and the new socialism. *Economica* 7, 419–34.

Kotkin, S. 1995. *Magnetic Mountain: Stalinism as a Civilization*. Berkeley: University of California Press.

Lange, O. 1936. On the economic theory of socialism: part one. *Review of Economic Studies* 4, 53–71.

Lavoie, D. 1985. *Rivalry and Central Planning: The Socialist Calculation Debate Reconsidered*. Cambridge: Cambridge University Press.

Lerner, A.P. 1936. A note on socialist economies. *Review of Economic Studies* 4, 72–6.

Levy, D.M. 1990. The bias in centrally planned prices. *Public Choice* 67, 213–36.

Levy, D.M. 2002. Robust institutions. *Review of Austrian Economics* 15, 131–42.

Levy, D.M., Peart, S.J. and Farrant, A. 2005. The spatial politics of the road to serfdom. *European Journal of Political Economy* 21, 982–99.

Lipsey, R.G. and Steiner, P.O. 1975. *Economics*, 4th edn. New York: Harper & Row.

Malthus, T.R. 1798. *An Essay on the Principle of Population as It Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*. London: J. Johnson.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

McConnell, C. 1963. *Economics: Principles, Problems, and Policies*, 2nd edn. New York: McGraw-Hill.

McKelvey, R.D. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47, 1085–113.

Mendershausen, H. 1949. Prices, money and the distribution of goods in postwar Germany. *American Economic Review* 39, 646–72.

Mill, J.S. 1844. A System of Logic. In *The Collected Works of John Stuart Mill*, vol. 8, ed. J.M. Robson. Toronto: University of Toronto Press, 1981.

Mill, J.S. 1848. The Principles of Political Economy. In *The Collected Works of John Stuart Mill*, vol. 3, ed. J. Robson. Toronto: University of Toronto Press, 1965.

Mises, L.E. von. 1920. Economic calculation in the socialist commonwealth, tr. S. Adler. In *Collectivist Economic Planning*, ed. F.A. Hayek. London: Routledge, 1935.

Mises, L.E. von. 1936. *Socialism: An Economic and Sociological Analysis*, tr. J.J. Kahane. Indianapolis: Liberty Classics, 1981.

Peart, S.J. and Levy, D.M. 2005. *The "Vanity of the Philosopher": From Equality to Hierarchy in Post-Classical Economics*. Ann Arbor: University of Michigan Press.

Roemer, J.E. 1994. *A Future for Socialism*. Cambridge: Harvard University Press.

Samuelson, P.A. 1948. *Economics: An Introductory Analysis*. New York: McGraw-Hill.

Samuelson, P.A. 1970. *Economics*, 8th edn. New York: McGraw-Hill.

Shleifer, A. and Vishny, R. 1992. Pervasive shortages under socialism. *Rand Journal of Economics* 23, 237–46.

Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19, 1–8.

How to cite this article

Levy, David M. and Sandra J. Peart. "socialist calculation debate." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000535> doi:10.1057/9780230226203.1570

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

soft budget constraint

Gérard Roland
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The concept of ‘soft budget constraint’ was first proposed by Janos Kornai to explain the pervasiveness of shortages under socialism. It has subsequently been understood mostly as a dynamic incentive problem whereby an investor would like to commit not to bail out an agent but ends up deciding on a bailout *ex post*. The concept has had a large number of applications in the transition literature but more broadly in economics as it can shed light on important episodes of bailouts.

Keywords

adverse selection; asymmetric information; bail-outs; business networks; capitalism; central planning; centralized and decentralized banking; Chinese economic reforms; commitment; contract theory; fiscal competition; hard budget constraint; hoarding; incentive; innovation; Kornai, J.; local government; monitoring; paternalism; price liberalization; privatization; ratchet effect; shortage; socialism; soft budget constraint; sunk costs

Article

The concept of ‘soft budget constraint’ was coined by János Kornai in his famous book *Economics of Shortage* (1980). In that book, Kornai developed a comprehensive theory of the socialist economy (completed in 1992 by *The Socialist System*). The starting point was that in the socialist economy firms had soft budget constraints as opposed to the hard budget constraints firms face under capitalism. ‘The classical capitalist firm had a hard budget constraint. If it is insolvent, it will sooner or later become bankrupt.•…•As opposed to this, the budget constraint of the traditional socialist firm is soft. If it works with a loss, that does not yet lead to real bankruptcy, i.e. ceasing operations’ (1980, p. 29). Kornai used the concepts of mainstream economic theory to contrast the situation of the capitalist firm and that of the socialist firm. Interestingly, in standard microeconomic theory only households have a budget constraint, not firms. The latter do not face any financial constraint and are only maximizing profits. Also, a budget

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
constraint is usually supposed to be hard; otherwise, it is not a constraint and would never bind. The concept of a soft budget constraint thus seemed intriguing. It reflected indeed a very important observation about the environment that the managers of socialist firms were facing as compared with the environment faced by the managers of capitalist firms. It immediately implied differences in the incentives managers were facing in the two economic systems, and one could understand that these different incentives had economy-wide consequences.

The basic observation that socialist firms had soft budget constraints and could count on continued financial support in case they were making losses led Kornai to develop a general theory of shortage in the socialist economy. Since the budget constraints of firms under socialism are not binding, one understands that they must necessarily meet a resource constraint, that is, experience shortage. This is, however, only an equilibrium implication. Shortage, or hitting the resource constraint, was thus achieved via the effect of soft budget constraints on the demand behaviour of firms.

The soft budget constraint and the socialist firm

Soft budget constraints had indeed a major effect on the demand behaviour of socialist firms. Demand for labour, input and capital tended to be in general higher than what it would be if firms had hard budget constraints. Moreover, demand was hardly responsive to price variations. Shortage tended to aggravate this phenomenon as it led to a hoarding motive in demand that tended to further aggravate shortages. Shortages were thus ubiquitous. Kornai developed a comprehensive theory of firm behaviour under shortage based on the premise of soft budget constraints. As of 2007 this is the most complete theory of the socialist economy that has been published, and most probably it will never be superseded. It is impossible in this short article even to summarize the exhaustive analysis Kornai made of firm behaviour and its general consequences for the economy. The interested reader is strongly encouraged to read the original analysis.

Like many general theories, Kornai's theory has led to extensions, refinements and also criticisms. Several papers (for example, Kornai and Weibull, 1983; Goldfeld and Quandt, 1988; 1990; 1993; Magee and Quandt, 1994) have examined formally the link between soft budget constraints and the supply and demand behaviour of socialist firms. Gomulka (1985) emphasized that households had hard budget constraints under socialism. In that case, the price system should have been able to eliminate shortages for consumer goods. Firms were, however, not responsive to signals from consumer goods markets under socialism since they were themselves resource-constrained on their supply side.

Why did firms in the socialist economy have soft budget constraints? Kornai related this to paternalism: ‘Paternalism is the direct explanation for the softening of the budget constraint’ (1980, p. 568). While *Economics of Shortage* was written while Kornai was still living under the socialist regime, this explanation remained partial for obvious reasons of self-censorship. In *The Socialist System*, he came back to the question at more length and explicitly defined causal mechanisms from the undivided power of the Communist Party to the dominance of state ownership, the preponderance of bureaucratic coordination to the environment of firms, including softness of budget constraints.

The soft budget constraint as a general incentive problem

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
While debates on soft budget constraints, their causes and consequences, were mostly confined to economists studying the socialist economy, it was recognized that the soft budget constraint syndrome had broader applications than in the context of socialist firms. Bail-outs of large firms and banks occur repeatedly under capitalism. Kornai's concept of the soft budget constraint was widely popularized by a now famous article by Dewatripont and Maskin (1995) in which the soft budget constraint problem was formalized within the context of contract theory. Dewatripont and Maskin formulated the soft budget constraint as a rather general dynamic commitment problem within the framework of contract theory. Specifically, a principal can decide to invest in a project but would like to commit to not refinancing the project if its return is low. However, if he or she cannot commit not to refinance the firm, there are conditions under which they end up bailing out the firm in order to cut its losses. One would thus observe projects that are on the whole unprofitable but that nevertheless get refinanced: in other words they are loss-making firms with soft budget constraints. The reason for this dynamic commitment problem is that the initial funds spent on the firm are sunk costs and only the net return to refinancing is taken into account in the bail-out decision. Dewatripont and Maskin thus showed the soft budget constraint to be a very general incentive problem, not necessarily one to be related only to the context of socialism or government ownership.

The Dewatripont–Maskin model considers the following adverse selection problem. The government faces a population of firms, each needing one unit of funds in initial period 1 in order to start its project. A proportion α of these projects are of the ‘good, quick’ type: after one period, the project is successfully completed, and generates a gross (discounted) financial return $R_g > 1$. Moreover, the manager of the firm (possibly also workers) obtains a positive net (discounted) private benefit E_g. In contrast, there is a proportion $(1-\alpha)$ of bad and slow projects which generate no financial return after one period. If terminated at that stage, managers in the firm obtain a private benefit E_t. Instead, if refinanced, each project generates after two periods a gross (discounted) financial return π_b^* and a net (discounted) private benefit E_b. Initially, α is common knowledge but individual types are private information. A simple result easily follows: if $1 < \pi_b^* < 2$ and $E_b > 0$, refinancing bad projects is sequentially optimal for the government, and bad entrepreneurs who expect to be refinanced apply for initial financing. The government would, however, be better off if it were able to commit not to refinance bad projects, since it would thereby deter managers with bad projects from applying for initial financing, provided $E_t < 0$.

Termination is here, by assumption, a disciplining device that allows the uninformed investor (creditor) to turn away bad types and only finance good ones. The problem is that termination is not sequentially rational if π_b^* is>1: once the first unit has been sunk into a bad project, its net continuation value is positive so that, in the absence of commitment, the soft budget constraint syndrome arises. In this set-up, because irreversibility of investment is such a general economic feature, the challenge for theory is more to explain why hard budget constraints prevail rather than why budget constraints are soft in the first place.

Dewatripont and Maskin considered the soft budget constraint problem within a banking set-up. They found that centralized banking was more prone to the soft budget constraint problem than decentralized banking. In particular, to the extent that decentralized banking makes refinancing more costly because of incentive problems between multiple investors, it tends to favour hard budget constraints (see also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Povel, 2004; Huang and Xu, 1999). Indeed, when refinancing is more costly, investors are less willing to bail out and more willing to terminate projects, thereby changing the incentives of enterprise managers. While the dynamic commitment aspect of the Dewatripont–Maskin model of the soft budget constraint coincides with Kornai's formulation, differences appear when it comes to explain the *cause* of the soft budget constraint problem. Indeed, Kornai emphasized the paternalism of government as the cause of soft budget constraints for the socialist firms. Paternalism plays no role in the Dewatripont–Maskin model. The main driving force is the sunk cost of investment combined with asymmetric information and inability to commit. These apparent differences in the explanation of soft budget constraints are, however, related to the fact that both explanations are located at different levels of abstraction and also related to different concepts of scientific explanation. Indeed, Dewatripont and Maskin develop a logical explanation of soft budget constraints deriving sufficient conditions for soft budget constraints. This logical explanation can then be applied to different empirical and institutional set-ups such as the one they discuss, namely, the comparison between centralized and decentralized banking. From that point of view, one can claim that paternalism is neither a necessary nor a sufficient condition for soft budget constraints. However, it would be misguided to take this as a criticism of Kornai's explanation, which is situated at a different level. His concept of soft budget constraint is an abstraction destined to capture an empirical regularity rooted in the behaviour of socialist firms. The explanation of soft budget constraints by paternalism is an explanation based on empirical plausibility rooted in the institutional context of the socialist economy. Because these are explanations at different levels (pure logic on one hand, a concept grasping an empirical regularity on the other hand), they should not be seen as contradictory and mutually exclusive. Kornai's explanation gives us crucial insights into the mechanisms of the socialist economy and the transition process, while Dewatripont and Maskin give us logical conditions that may be applied to different institutional contexts.

The Dewatripont–Maskin model has made the soft budget constraint concept an integral part of mainstream economic theory. It has also been recognized as a general incentive problem that has played a role in diverse banking crises that have occurred since the 1980s, including the 1997 east Asian crisis.

Models of soft budget constraints under socialism

Soft budget constraint models following the seminal model of Dewatripont and Maskin have been developed in a wide variety of contexts.

First, beyond the models cited above that look mostly at the consequences of soft budget constraints, models were developed to understand various aspects of the socialist economy. For example, Qian (1994) developed a model showing that in the context of the socialist economy, shortages were a good way of alleviating some of the negative consequences of the soft budget constraint. Indeed, he showed that price caps that lead to shortages could be beneficial in terms of somewhat reducing soft budget constraints. Indeed, shortages reduce the likelihood that bad projects will be refinanced. This therefore reduces the incentive to submit poor projects. *A contrario*, with flexible prices and soft budget constraints, enterprises will bid for scarce resources, leading to price inflation and crowding out of consumers who face hard budget constraints. This explains why in many socialist economies, partial price liberalization as in the Soviet Union under President Gorbachev or in Poland, Hungary and China, and even advanced price liberalization as in the former Yugoslavia, led to strong inflationary pressures.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Other models by Wang (1991) and Debande and Friebel (2004) showed similarly that reforms giving more autonomy to socialist enterprises, by relaxing the monitoring of enterprise activities, tended to exacerbate the soft budget constraint phenomenon.

Qian and Xu (1998) developed a model to explain how soft budget constraints had a negative effect on innovation under socialism. This is an important theme since the inferiority of the socialist economy to capitalism in generating innovation is one of the main reasons for its demise. Knowing that firms had soft budget constraints, central planners were very cautious in approving investments with innovations. Indeed, the return to innovations is uncertain and it is important to be able to stop innovations that turn out to be disappointing. The market economy is able to do that quite well because firms have hard budget constraints; but this was not the case in the socialist economy. The caution of central planners in approving investments was relatively greater in areas where innovation was riskier and science was relatively new, precisely in the areas where returns to innovation were greater. This was for example the case for the computer industry where socialist firms were considerably behind capitalist firms. In industries where science was older and where risk was likely to be smaller, the centrally planned economy fared better. This was the case for the aerospace industry.

Dewatripont and Roland (1997) developed a model to analyse the links between the soft budget constraint and another incentive problem that was present in the socialist economy: the ratchet effect. The term ‘ratchet effect’ was coined by Berliner (1952) in his analysis of management behaviour in Soviet-style firms. In such firms, managers were given what appeared to be strong incentives to fulfil their production plans. Indeed, they had inducements to *overfulfil* the plans: each percentage point over the target was rewarded by additional bonuses. Nevertheless, managers tended to pass up the opportunity for these bonuses and instead were conservative in their plan overfulfilment, rarely exceeding two per cent over target. Berliner's explanation for this conservatism was that managers feared that next year's target would be ‘ratcheted up’ (that is, made more demanding) if they exceeded this year's goal. By producing at 110 per cent instead of 102 per cent, their bonus would be higher today, but so would their target tomorrow. Dewatripont and Roland show that the ratchet effect and the soft budget constraint could be interrelated in the sense that the need to bail out weaker firms gave central planners a stronger incentive to ratchet up the plans of the better-performing firms.

Models of soft budget constraints under transition

The issue of soft budget constraints, which was initially fundamentally neglected in the early transition from Communism, especially by the advocates of the Washington consensus and the big-bang approach to transition (apart from some mention here and there), became an increasingly important issue in trying to understand the restructuring process of firms and of banks. First, the transfer of ownership from the state to the private sector changed the incentives to bail out firms but did not automatically eliminate soft budget constraints. Indeed, by using the Dewatripont–Maskin model, it is easy to understand (see for example Kornai, Maskin and Roland, 2003) that privatization, by shifting the financing of firms from the government to the private sector, and in particular to private banks, changes the motive for bailing out firms from *ex post* care for employment and the welfare of those working inside firms to *ex post* profit maximization. However, this only reduces the extent of soft budget constraints since the logic of Dewatripont–Maskin does not require state ownership to generate soft budget constraints. What is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
important is the sunk cost nature of investment that would lead an investor to refinance *ex post* a firm even though it would want to commit not to do so *ex ante*.

Complementary to the changes in ownership, there are other institutional changes during the transition process that would tend to harden budget constraints. Several models were developed along those lines. Segal (1998) examined the role of demonopolization and the development of competition in the hardening of budget constraints. The argument is that demonopolization and competition both increase entry of firms and reduce their profits. The reduction in profitability reduces the incentive to bail out but the increase in entry reduces the likelihood of any single firm of being bailed out. The argument requires a certain amount of excess capacity of firms in the market at any moment in time which is consistent with competition between firms.

Berglöf and Roland (1998) advanced a slightly different argument about the entry of new firms. If newly entering firms have a sufficiently high expected return in a competitive market economy, then the expected return to lending to those firms might be higher than the expected return to bailing out existing firms. This would then yield hard budget constraints as firms with poor projects would know that they were unlikely to be bailed out. However, for the argument to be valid, the expected return of the newly entering firms should be significantly higher than that of older firms. Indeed, with an expected return that is only marginally higher than for old firms, soft budget constraints will still be present. The reason is that the competition for funds is tilted against the newly entering firms because initial investments in older projects are sunk costs. Therefore, the comparison between the return to bailing out existing firms and financing entering firms does not count the initial funding of existing firms, so giving the latter an advantage in the competition for funds. The upshot is that, if the rhythm of innovation in the economy is not sufficiently strong, that is, if newly entering firms do not have an expected profitability that is significantly higher than existing firms, then soft budget constraints may persist.

Yet another argument related to firm restructuring was made by Perotti (1993) and Coricelli and Miles-Ferretti (1993). This is basically an argument about the negative externalities of soft budget constraints. Here, the hardening of budget constraints depends on the existing firms' links with their chains of suppliers and clients. The idea is that banks may be reluctant not to bail out firms if closing the operations of those firms creates financial difficulties for the more healthy firms who do business with the weaker firms. The softness of banks may in turn dull the incentives of firms to restructure. From that point of view, it is harder to restructure an enterprise operating in a business network where budget constraints are very soft. Firms setting up new business networks in an environment with hard budget constraints will then face hard budget constraints themselves.

Another key institutional variable affecting the hardness of budget constraints is the degree of decentralization of banking. This is what the original Dewatripont–Maskin model was about. The general idea is that under decentralized banking renegotiating initial loan contracts is more difficult because it involves more inefficiencies and it thus more costly than under centralized banking. The particular mechanism in the model is the following: if the initial bank is liquidity-constrained and refinancing must involve another bank, the initial bank has less incentive to monitor the firm to ensure a higher return to refinancing than would be the case under centralized banking. Indeed, it must share the returns to monitoring with the other bank, which is assumed to have less experience with the existing project and thus is unable to monitor the use of its money. Because the bank that monitors does not reap the full returns from monitoring, it has less incentive to monitor. Other mechanisms would deliver the same result. Povel (1995) examined a model in which a project is financed from the outset by two banks.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Suppose that an agreement on a restructuring plan is necessary to refinance a poor project and that each bank's assessment of the continuation value of the project is private information. The asymmetric information between banks can give rise to a delay in their negotiating an acceptable restructuring plan. However, if the value of the project declines over time, this delay may render refinancing unprofitable. Huang and Xu (1999) studied a related model in which two banks (investors) agree to lend jointly to a project but have conflicting interests concerning how to organize the project should it be refinanced. This conflict may make it too costly to reach an agreement on a strategy after bail-out. Huang and Xu apply this argument to illuminating the east Asian crisis of the late 1990s. They note that the Korean *jaebeols* were subject to centralized financing and suffered from lack of financial discipline and soft budget constraints. By contrast, Taiwan's economy was characterized by dispersed financial institutions and decentralized banking. In the event, Taiwan suffered much less from the crisis than Korea (even though it, too, was attacked by speculators).

Decentralization of government, under federalism, may also have the effect of hardening budget constraints under certain circumstances, as argued by Qian and Roland (1998) on the basis of the transition experience in China. Decentralization of government was an important feature of Chinese reforms. Among other features, it led local governments to compete with each by investing in infrastructure investment other than to attract capital in order to boost growth in their province or region. This form of fiscal competition leads to overinvestment in infrastructure because the return of infrastructure investment to a province is higher than the return to society as a whole, as local governments do not internalize the effects of attracting capital away from other provinces. A positive side effect of this fiscal competition, however, can be that local governments prefer now to put their money in infrastructure investment rather than in bailing out loss-making enterprises. While decentralization of government can harden the budget constraints of enterprises that are controlled by local government, it can lead to soft budget constraints for local governments. Indeed, local governments can always structure the composition of their expenditures in such a way as to coax additional funds from the central government. Say that local governments are responsible for hospitals: they can strategically underfund hospitals and spend their budget on other items so as to obtain additional funds from the central government to improve hospital services. Zhuravskaya (2000) found evidence in the case of Russia that is consistent with such a story, using a data-set from 35 large cities in 29 regions of the Russian Federation between 1992 and 1997. She found that any increase in own revenues by local governments tends to be offset by a decrease of nearly one to one of shared revenues, which is consistent with soft budget constraints. Moreover, weaker fiscal incentives (a negative correlation between shared revenues and own revenues) tend, everything else equal, to decrease spending on education and health. There is also a negative impact on the quality of health and education, as measured by infant mortality and evening school attendance by children due to crowded schools. This is consistent with the idea that local governments have an incentive to distort their expenditures by neglecting health and education so as to try to attract more grants.

The soft budget constraint phenomenon has been analysed not only in the context of enterprises or local governments. It has been studied in the context of banks. Mitchell (1998), for example, built a model of bank passivity where banks fail to liquidate bad projects because they themselves expect to be bailed out by government. Similarly, Berglöf and Roland (1995) show that banks may strategically exploit the government's concerns for jobs in loss-making firms to extract government subsidies to refinance firms,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in effect forcing the government to pay part of the refinancing of firms. Policies of bank monitoring and recapitalization may help to eliminate soft budget constraints of banks, but policies of recapitalization face incentive problems of their own (see Aghion, Bolton and Fries, 1999).

Other visions of soft budget constraints

Interpretations of the soft budget constraint different from the Dewatripont–Maskin model are also present in the economics profession. Boycko, Shleifer and Vishny (1996), for example, do not identify soft budget constraints as a dynamic commitment problem where the principal would prefer not to commit to bailing out but ends up doing so. They see soft budget constraints as a deliberate choice of a governmental body to subsidize an enterprise in order to prevent management from laying off workers. According to this interpretation, soft budget constraints are not an inefficiency but rather the consequence of government preferences. This interpretation has had some following partly because the Dewatripont–Maskin view of soft budget constraints has often been associated too narrowly with bank–enterprise relationships and because, in reality, soft budget constraints occur more often in government–firm or government–bank relationships.

More recently, Robinson and Torvik (2006) have proposed a political economy theory of the soft budget constraint in a framework where politicians cannot commit to electoral promises. In this context, soft budget constraints, while resulting from an inability to commit to not bailing out a project, can be seen as commitment devices to redistribute transfers to particular constituencies of voters and to secure re-election and create an incumbency advantage. In other words, it emerges as a kind of political patronage.

Empirical work on soft budget constraints

It is very difficult to do empirical work measuring soft budget constraints as *ex post* bail-outs that were not desired *ex ante*. Seen in this light, soft budget constraints must be distinguished from other forms of funding that do not have these features. Moreover, the behaviour of firms is based on expectations of bail-out that are not easy to measure either. It is therefore not surprising that much of the empirical work surrounding soft budget constraints has been about using proxies that can find some justification but are necessarily quite noisy. The empirical work has often been quite detached from the theory. Various studies have looked at tax arrears as a proxy for soft budget constraints (Claessens and Peters, 1997; Schaffer, 1998; and Coricelli and Djankov, 2001). Others have looked at open subsidies (Djankov and Nenova, 2000; Grigorian, 2000) or tax concessions (Alfandari, Fan and Freinkman, 1996; Brown and Earle, 2000; and Shleifer and Treisman, 2000). Other studies have looked at payment arrears of firms and the tendencies of state-owned banks to lend to distressed enterprises (Cull and Xu, 2000; Gao and Schaffer, 1998; Coricelli and Djankov, 2001; Claessens and Djankov, 1998; and Schaffer, 1998).

There have nevertheless been a few serious attempts to study soft budget constraints empirically in line with existing theoretical models. Petterson-Liblom and Dahlberg (2005), for example, have made such an attempt and tried to measure the effect of soft budget constraints on the borrowing behaviour of Swedish municipal governments between 1974 and 1992. They find that, if a municipality expects to be bailed out with certainty, as opposed to a situation where it is not bailed out, it will increase its debt level

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
by 30 per cent. Bail-out expectations obviously cannot be measured directly. However, instead they use observed bail-out as a noisy measure of bail-out expectations and use an instrumental variable approach to get consistent measurement using as instrument observed bail-outs in neighbouring municipalities.
A comprehensive survey of the literature on soft budget constraints can be found in Kornai, Maskin and Roland (2003).

See Also

- adverse selection
- agency problems
- privatization
- socialism
- state capture and corruption in transition economies

Bibliography

Aghion, P., Bolton, P. and Fries, S. 1999. Optimal design of bank bailouts: the case of transition economies. *Journal of Institutional & Theoretical Economics* 155, 51–70.

Alfandari, G., Fan, Q. and Freinkman, L. 1996. Government financial transfers to industrial enterprises and restructuring. In *Enterprise Restructuring and Economic Policy in Russia*, ed. S. Commander, Q. Fan and M. Schaffer. Washington, DC: Economic Development Institute of the World Bank.

Berglöf, E. and Roland, G. 1995. Bank restructuring and soft budget constraints in financial transition. *Journal of the Japanese and International Economies* 9, 354–75.

Berglöf, E. and Roland, G. 1998. Soft budget constraints and banking in transition economies. *Journal of Comparative Economics* 26, 18–40.

Berliner, J. 1952. The informal organization of the Soviet firm. *Quarterly Journal of Economics* 66, 342–65.

Boycko, M., Shleifer, A. and Vishny, R. 1996. A theory of privatization. *Economic Journal* 106, 309–19.

Brown, D. and Earle, J. 2000. Privatization and enterprise restructuring in Russia: new evidence from panel data on industrial enterprises. Working Paper No. 1. Moscow: Russian European Center for Economic Policy.

Claessens, S. and Djankov, S. 1998. Politicians and firms in seven central and eastern European countries. Policy Research Working Paper No. 1954. Washington, DC: World Bank.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Claessens, S. and Kyle Peters, Jr. R. 1997. State enterprise performance and soft budget constraints: the case of Bulgaria. *Economics of Transition* 5, 305–22.

Coricelli, F. and Djankov, S. 2001. Hardened budgets and enterprise restructuring: theory and an application to Romania. *Journal of Comparative Economics* 29, 749–63.

Coricelli, F. and Milesi-Ferretti, M.G. 1993. On the credibility of ‘big bang’ programs: a note on wage claims and soft budget constraints in economies in transition. *European Economic Review* 37, 387–95.

Cull, R. and Xu, L.C. 2000. Bureaucrats, state banks, and the efficiency of credit allocation: the experience of Chinese state-owned enterprises. *Journal of Comparative Economics* 28, 1–31.

Debande, O. and Friebel, G. 2004. A positive theory of ‘give-away’ privatization. *International Journal of Industrial Organization* 22, 1309–25.

Deviatov, A. and Ickes, B.W. 2005. Reputation and the soft budget constraint. Mimeo, New Economic School, Moscow.

Dewatripont, M. and Maskin, E. 1995. Credit and efficiency in centralized and decentralized economies. *Review of Economic Studies* 62, 541–55.

Dewatripont, M. and Roland, G. 1997. Transition as a process of large scale institutional change. In *Advances in Economic Theory*, vol. 2, ed. D. Kreps and K. Wallis. Cambridge: Cambridge University Press.

Dewatripont, M. and Roland, G. 2000. Soft budget constraints, transition and financial systems. *Journal of Institutional and Theoretical Economics* 156, 245–60.

Djankov, S. and Nenova, T. 2000. Why did privatization fail in Kazakhstan? Working paper. Washington, DC: World Bank.

Gao, S. and Schaffer, M.E. 1998. Financial discipline in the enterprise sector in transition countries: how does China compare? Discussion Paper No. 98/01. Edinburgh: Centre for Economic Reform and Transformation, Heriott-Watt University.

Goldfeld, S.M. and Quandt, R.E. 1988. Budget constraints, bailouts and the firm under central planning. *Journal of Comparative Economics* 12, 502–20.

Goldfeld, S.M. and Quandt, R.E. 1990. Output targets, the soft budget constraint and the firm under central planning. *Journal of Economic Behavior and Organization* 14, 205–22.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Goldfeld, S.M. and Quandt, R.E. 1993. Uncertainty, bailouts, and the Kornai effect. *Economics Letters* 41, 113–19.

Gomulka, S. 1985. Kornai's soft budget constraint and the shortage phenemenon: a criticism and restatement. *Economics of Planning* 19(1), 1–11.

Grigorian, D.A. 2000. Ownership and performance of Lithuanian enterprises. Policy Research Working Paper No. 2343. Washington, DC: World Bank.

Huang, H. and Xu, C. 1999. Financial institutions and the financial crisis in east Asia. *European Economic Review* 43, 903–14.

Kornai, J. 1980. *Economics of Shortage*. Amsterdam: North-Holland.

Kornai, J. 1992. *The Socialist System. The Political Economy of Communism*. Princeton: Princeton University Press; Oxford: Oxford University Press.

Kornai, J., Maskin, E. and Roland, G. 2003. Understanding the soft budget constraint. *Journal of Economic Literature* 41, 1095–136.

Kornai, J. and Weibull, J.W. 1983. Paternalism, buyers' and sellers' market. *Mathematical Social Sciences* 6, 153–69.

Magee, K.L. and Quandt, R.E. 1994. The Kornai effect with partial bailouts and taxes. *Economics of Planning* 27, 27–38.

Mitchell, J. 1998. Strategic creditor passivity, regulation and bank bailouts. Discussion Paper No. 1780. London: CEPR.

Perotti, E. 1993. Bank lending in transition economies. *Journal of Banking and Finance* 17, 1021–32.

Petterson-Lidblom, P. and Dalhberg, M. 2005. An empirical approach for estimating the causal effect of soft budget constraints on economic outcomes. Mimeo, Institute for International Economic Studies.

Povel, P. 2004. Multiple banking as a commitment not to rescue. *Research in Finance* 21, 175–99.

Qian, Y. 1994. A theory of shortage in socialist economies based on the soft budget constraint. *American Economic Review* 84, 145–56.

Qian, Y. and Roland, G. 1998. Federalism and soft budget constraint. *American Economic Review* 88, 1143–62.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Qian, Y. and Xu, C. 1998. Innovation and bureaucracy under soft and hard budget constraints. *Review of Economic Studies* 65, 151–64.

Robinson, J.A. and Torvik, R. 2006. A political economy theory of the soft budget constraint. Working Paper No. 12133. Cambridge, MA: NBER.

Schaffer, M.E. 1998. Do firms in transition have soft budget constraints? A reconsideration of concepts and evidence. *Journal of Comparative Economics* 26, 80–103.

Segal, I.R. 1998. Monopoly and soft budget constraint. *Rand Journal of Economics* 29, 596–609.

Shleifer, A. and Treisman, D. 2000. *Without a Map: Political Tactics and Economic Reform in Russia*. Cambridge, MA: MIT Press.

Wang, Y. 1991. Economic reform, fixed capital investment expansion and inflation: a behaviour model based on the Chinese experience. *China Economic Review* 2, 3–27.

Zhuravskaya, E. 2000. Incentives to provide local public goods: fiscal federalism, Russian style. *Journal of Public Economics* 76, 337–68.

How to cite this article

Roland, Gérard. "soft budget constraint." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000475> doi:10.1057/9780230226203.1571

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sohmen, Egon (1930 – 1977)

Charles P. Kindleberger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

exchange rates; Sohmen, Egon

Article

Born in Linz, Austria, on 1 June 1930 and died in Heidelberg, Federal Republic of Germany, on 8 March 1977. He was educated at the Universities of Vienna, Kansas and Tübingen, and at the Massachusetts Institute of Technology, from which he received a doctorate in economics in 1958. He taught economics at Yale University from 1958 to 1961, at the University of Saarlandes in Saarbrücken from 1961 to 1969, and at Heidelberg University from 1969 to his death. Sabbatical leaves were spent at the University of Minnesota in 1963–4, and at the Smithsonian Institution in Washington, DC, in the spring of 1975.
Sohmen played a significant part in the 1960s in making the case for flexible exchange rates respectable. He wrote widely on the subject, attacking the Bretton Woods system and insisting that free floating would produce exchange rates that approached equilibrium continuously, as opposed to fixed-rate systems with their encouragement of speculation, wide departures from equilibrium and ultimate necessities for parity changes. With the adoption of floating rates in 1973, he turned his attention to other problems in economic theory with emphasis on competitive markets.

Selected works

A complete bibliography of Sohmen's work is contained in a volume of essays in his memory, edited by J.S. Chipman and C.P. Kindleberger, *Flexible Exchange Rates and the Balance of Payments* (Amsterdam: North-Holland, 1980). The chief items are:

1961. *Flexible Exchange Rates*. Chicago: University of Chicago Press. Revised edn, 1969.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1964. *International Monetary Problems and the Foreign Exchanges*. Special Papers in International Economics, Princeton: International Finance Section.

1966. *The Theory of Forward Exchange*. Princeton Studies in International Finance, Princeton: International Finance Section.

1976. *Allokationstheorie und Wirtschaftspolitik*. Tübingen: J.C. Mohr (Paul Siebeck).

How to cite this article

Kindleberger, Charles P. "Sohmen, Egon (1930–1977)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000178> doi:10.1057/9780230226203.1868

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Solow, Robert (born 1924)

Alan S. Blinder
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Robert Solow is a leading theorist of economic growth. In one of his two pioneering papers he argues that, with labour fully employed, the long-run economic growth rate is independent of the saving rate. In the other, he argues that only a minority of economic growth can be explained by increases in labour and capital inputs; the residual, which presumably reflects technological innovation, accounts for the majority. He also argues that new capital is more valuable than old capital because it embodies more up-to-date technology. He has made contributions on fiscal policy and wage bargaining as well.

Keywords

American Economic Association; Arrow, K.; budget deficits; CES production function; Cobb–Douglas functions; first turnpike theorem; Great Depression; growth theory; Harrod–Domar growth model; IS-LM model; Keynesianism; Leontief, W.; monetarism; Phillips curve; Samuelson, P.; saving rate; Solow, R.; Solow residual; Swan, T.; Technology

Article

Robert Merton Solow was born in Brooklyn, New York, in August 1924, so he was just five years old when the Great Depression began in the United States and, depending on how it is dated, about 16 years old when it ended. The timing is significant. He has characterized himself as part of ‘the generation of economists that was moved to study economics by the feeling that we desperately needed to understand the Depression’ (1990, p. 183). The search for answers about what makes the macroeconomy tick, and why it occasionally goes badly off track, has been the touchstone of his illustrious career.

Solow was educated in the New York City public schools at a time when that was a great place to be educated – a time and place that produced, among others, Kenneth Arrow, William Baumol and Robert Fogel. (That's just in economics; the number of distinguished scientists is legendary.) Graduating at age 16, he became the first person in his family to attend college when he enrolled at Harvard in 1940. But with the Second World War raging, Harvard did not hold his attention for long, and he volunteered for the US Army in 1942, serving for three years in North Africa and Italy – an experience which, he has written, ‘formed my character’ (1988).

Once back at Harvard, and now married to the love of his life, Barbara, Solow decided to specialize in economics and was fortunate to have Wassily Leontief assigned as his tutor. This happenstance probably also changed his life – and that of the economics profession – forever; for while economics at Harvard may not have been very inspiring at the time, nor particularly modern, nor at all mathematical, Leontief was a shining exception. He took the brilliant young student from Brooklyn under his wings and, six years later, Solow won the first of what were to be many accolades: the coveted David A. Wells Prize for the best Harvard Ph.D. dissertation. The thesis, which Solow never published because ‘I thought I could do it better’ (1988), was a highly original application of the then-novel theory of Markov processes to model the size distribution of wage income in the United States. (I actually read this fine piece of work, in typescript, while writing my own dissertation on the size distribution of income at MIT about 20 years later.)

Solow's first and, in a real sense, only job was at MIT, where he was an assistant professor and then an associate professor of statistics in what was then the Department of Economics and Social Science from 1950 to 1958. It was during that period that he wrote his two classic papers on the theory and empirics of economic growth (described below), which must have made it easy for MIT to promote him to professor of *economics* in 1958. Fifteen years later he was designated an Institute Professor, a post he held until his (*de jure*, not *de facto*) retirement from MIT in 1995. In 1979 he served as president of the American Economic Association. In 1987 he was awarded the Nobel Prize in Economic Science. And in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1999 he was honoured with the National Medal of Science – one of only seven economists ever to have won that distinction. Since 2000 Solow has divided his time between MIT and the Russell Sage Foundation in New York, where he is the foundation's only permanent fellow.

In terms of research *publication* dates, the years 1956–61 in Solow's career were probably as outstanding a quinquennium as any economist has ever had. (The research production dates, one guesses, were probably two years or so earlier.) His two most famous papers, of course, came in 1956 and 1957: ‘A contribution to the theory of economic growth’, *Quarterly Journal of Economics* (1956), and ‘Technical change and the aggregate production function’, *Review of Economics and Statistics* (1957). These two papers, which introduced what came to be called ‘the Solow model’ and ‘the Solow residual’, respectively, are still, almost 50 years later, among the most widely cited papers in all of economics, despite the fact that they are so much a part of the corpus of modern economics that citations are often omitted. At this point the Nobel Prize, had there been one at the time, was in the bag. And indeed, when his Nobel award was announced, these were the two papers principally cited.

The two landmark growth papers were followed by the classic book *Linear Programming and Economic Analysis* (with Paul Samuelson and Robert Dorfman, 1958), which included, among other things, the first turnpike theorem. Next, in a 1960 paper titled ‘Investment and technical progress’, Solow introduced the important – and probably realistic – idea that new technology might have to be ‘embodied’ in new capital. (After all, a vintage 2000 computer does not develop new and more powerful chips just because Intel invents them.) This theoretical innovation opened the door to a new and rich (though more complicated) class of ‘vintage’ models, which then proliferated in the literature. Finally, in collaboration with Arrow, Hollis Chenery, and B. S. Minhas, he invented the constant-elasticity-of-substitution (CES) production function in their paper ‘Capital-labor substitution and economic efficiency’ (1961). This clever functional form has proven to be both a theoretical and an empirical workhorse.

This remarkable burst of intellectual activity and creativity clearly established Robert M. Solow, by then aged 37, as *the* major figure in what was then the sub-field of economic theory that was attracting the most attention (growth theory). And it won him the American Economic Association's John Bates Clark Medal, which is awarded every second year to the most outstanding economist under the age of 40, in 1961. But along the way he also found time to collaborate with Samuelson in bringing the Phillips curve to America in their, ‘Analytical aspects of anti-inflation policy’, *American Economic Review*, 1960. Whew!

The Solow model

Now back to 1956. Why is the Solow model so important in the history of economic thought? Prior to Solow's ‘Contribution’ and the contemporaneous paper by Trevor Swan (1956), growth theory was stuck in an awkward position. Roy Harrod (1939) and Evsey Domar (1946), noting the long-run (relative) fixity of the capital-output ratio, had posited a fixed-proportions technology:

$$Y_t = fK_t \tag{1a}$$

$$Y_t = gL_t, \tag{1b}$$

where f and g, two constants, are the reciprocals of the capital-output and labour-output ratios. Equating investment ($I_t = \frac{dK_t}{dt}$) and saving (S_t), and assuming that saving is proportional to income yields:

$$\frac{dK_t}{dt} = S_t = sY_t, \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and appending an exogenously growing labour force,

$$L_t = L_0 e^{nt}, \quad (3)$$

leads to the simple Harrod–Domar growth model. The model is easily solved for a so-called balanced growth path in which output, capital, and labour are all growing at the same rate, so that ratios like $y = \frac{Y}{L}$ and $k = \frac{K}{L}$ are constant. By (3), that common growth rate must be *n*—or *n*+ μ , where μ is the rate of technical progress, in the presence of exogenous technological progress. (In that case, *L* must be measured in efficiency units.) By (2), the growth rate of capital is $\left(\frac{1}{K}\right)\left(\frac{dK}{dt}\right) = s\left(\frac{Y}{K}\right)$ which, by (1a) is just *sf*. So the fundamental Harrod–Domar equation equates the so-called natural rate of growth, *n*, to the so-called 'warranted' rate of growth, *sf*:

$$sf = n. \quad (4)$$

Output in an economy with a labour force growing at rate *n* can therefore grow at that same rate *indefinitely* as long as *sf* is equal to *n*. But therein lies the problem that bothered Solow. The product *sf* is a number, which is equal to *n* only by coincidence. What if they are unequal? If *sf*>*n*, *K* will grow faster than *L forever*; if *sf*<*n*, it will grow slower – also forever. The economy therefore seems to be poised on a knife-edge between explosive growth of *y* and *k* and implosive contraction. Only if *sf* happens to be equal to *n* is it capable of supporting steady growth.

To Solow, the child of the Depression, this was a most unsatisfactory state of affairs in modelling, because real economies do not behave that way. The Great Depression was such a noteworthy event precisely because it was so *unusual*. History teaches us that capitalist economies do not often either implode or explode. Something approximating steady growth is much more normal. Could all this really be the result of coincidence? Solow thought not.

Furthermore, he had long been interested in production theory and viewed factor proportions – at least at the aggregate level – as more variable than fixed. So he proposed making just one change in the Harrod–Domar model: replacing the fixed-coefficients technology (1) by a more conventional neoclassical production function:

$$Y = F(K, L) \quad (5)$$

which, under constant returns to scale, can be written:

$$\frac{Y}{L} = F\left(\frac{K}{L}, 1\right) \text{ or}$$
$$y = f(k). \quad (6)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Now the solution looks the same as in the Harrod–Domar model because saving (=investment) per head is $sf(k)$, which is constant along a steady-state growth path (defined as a path with constant $k = \frac{K}{L}$). The growth rate of the capital stock is therefore $\frac{I}{K} = \frac{sf(k)}{k}$, and equating this to the growth rate of labour (n) leads to the famous Solow equation:

$$sf(k) = nk. \tag{7}$$

This resembles the Harrod–Domar equation, (4), and in fact, if $f(\cdot)$ is a linear function, the two are identical. But by making $f(\cdot)$ concave, Solow eliminated the bothersome knife-edge. Now no coincidence of constants is necessary to allow steady growth; instead, k will adjust to guarantee it, as illustrated by the famous Solow diagram (see Figure 1).

Figure 1

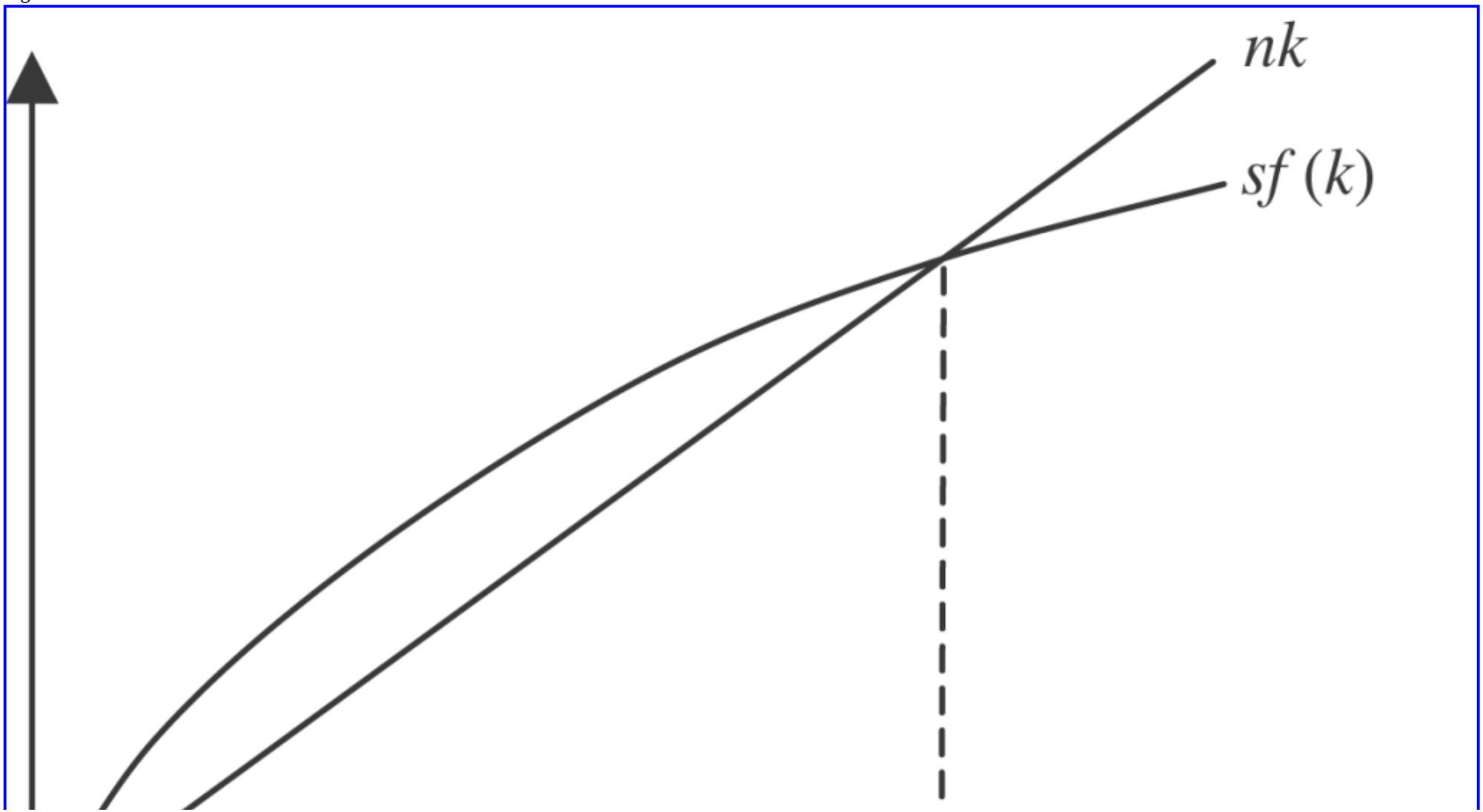

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

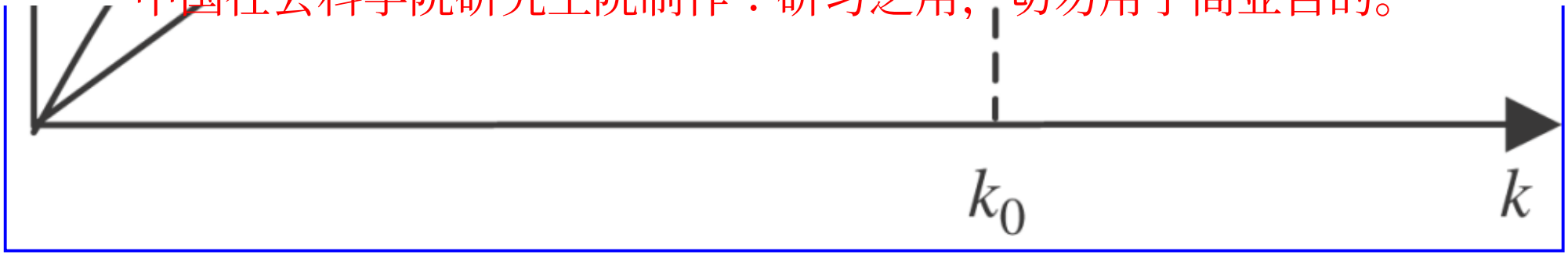

Many *complications* can be and have been added to the basic Solow model – some of them present already in Solow's original 1956 paper. But the two most essential and important lessons are extremely robust. First, the economy's growth path is most likely stable, not perched on a precarious knife edge. Consider the growth rate of $k = \frac{K}{L}$ which, in steady-state equilibrium, is zero. Out of equilibrium, it is:

$$\left(\frac{1}{k}\right)\frac{dk}{dt} = \left(\frac{1}{K}\right)\frac{dK}{dt} - \left(\frac{1}{L}\right)\frac{dL}{dt} = \frac{sf(k)}{k} - n, \text{ so that } \frac{dk}{dt} = sf(k) - nk. \quad (8)$$

Thus k will be *increasing* whenever $sf(k)>nk$ and *decreasing* whenever $sf(k)<nk$. A glance at Figure 1 (or taking the derivative of equation (8) with respect to k) shows that the dynamic model is *stable* – the reverse of the Harrod–Domar conclusion – as long as $sf'(k)<n$ (see again Figure 1.) Thus this new wrinkle achieved Solow's modelling objective: the economy normally stays on track.

But the Solow model accomplished more. Contrary to what had been common intuition until 1956 (and is still common intuition among those not schooled in the Solow model), the model shows that *the economy's growth rate* (n in this case, or $n+\mu$ with exogenous technical progress) *is independent of its saving rate* (s). How can that be true when, in the model, saving is the same as investment? The answer, which was surprising at the time but is now a commonplace, is that a society's propensity to save affects its *level* of output, as shown by equation (7), but not its steady-state *growth rate*.

Of course, this lesson can be learned too well. Figure 2 shows what happens if a Solow economy manages to *increase* its saving rate, from s_0 to s_1. The steady-state level of capital per head will eventually rise from k_0 to k_1, raising output per head from $f(k_0)$ to $f(k_1)$. Thus societies that save more will be richer, just as expected. But Solow pointed out that they *will not* grow faster in steady state; in fact, the growth rates are identical in the two equilibria shown in Figure 2.

Figure 2

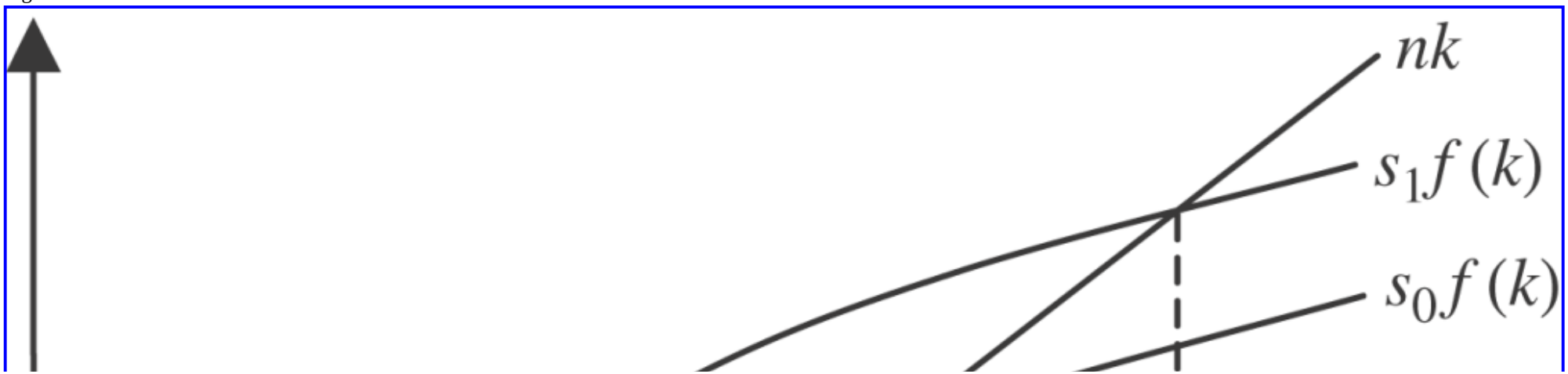

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

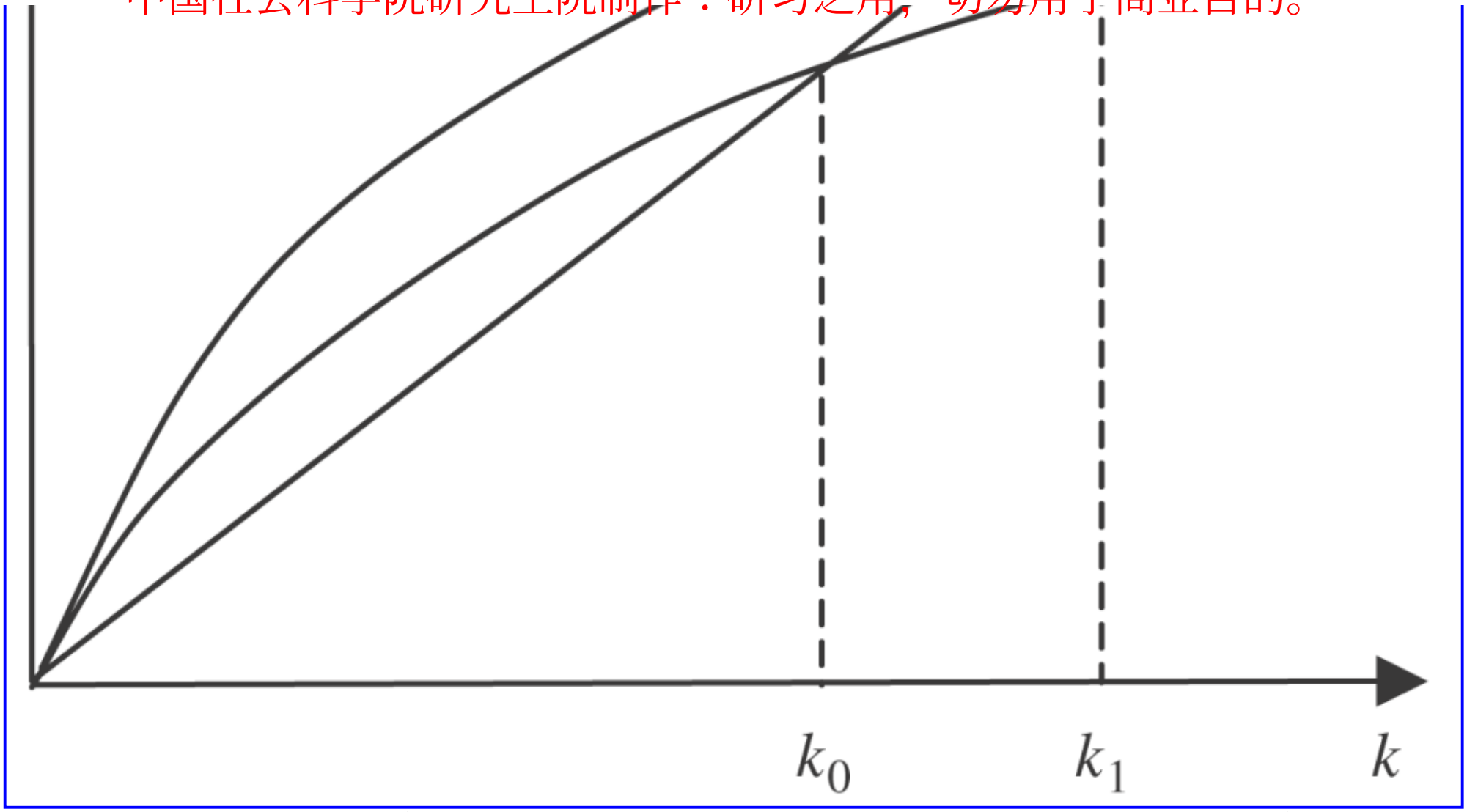

However, we should remember that the transition from one steady state to another can take a very long time (see Atkinson, 1969). During the adjustment period from point A to point B, which may last for decades, the high-saving society will indeed grow faster than the low-saving society. It's just that diminishing returns to capital eventually catch up to it, bringing the growth rate back down to *n*. Furthermore, what later came to be called ‘endogenous growth theory’ argued that capital deepening (which often meant knowledge capital or human capital) can boost the *growth rate* forever, if returns are not diminishing.

The Solow residual

Solow's famous 1957 paper picked up on the diminishing returns point. It had two main purposes: to verify that, as an empirical matter, returns to capital are in fact diminishing; and to estimate the relative contributions of technological progress and capital deepening (rising *k*) to growth.
He began by adding what we now call ‘Hicks-neutral’ technical progress to the production function (5) to get:

$$Y = A_t F(K_t, L_t).$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(9)

Under constant returns and assuming marginal productivity pricing of factor inputs, Euler's theorem implies that:

$$\left(\frac{1}{Y_t}\right)\frac{dY_t}{dt} = \left(\frac{1}{A_t}\right)\frac{dA_t}{dt} + \beta\left(\frac{1}{K_t}\right)\frac{dK_t}{dt} + (1-\beta)\left(\frac{1}{L_t}\right)\frac{dL_t}{dt},$$

where β is capital's factor share. Subtracting the growth rate of labour from both sides leads to the ‘other’ famous Solow equation:

$$\left(\frac{1}{y}\right)\frac{dy}{dt} = \left(\frac{1}{A}\right)\frac{dA}{dt} + \beta\left(\frac{1}{k}\right)\frac{dk}{dt},$$

(10)

where $\left(\frac{1}{A}\right)\frac{dA}{dt}$ is what came to be called *the Solow residual*, that is, the portion of the growth of labour productivity that is not accounted for by capital deepening.
As an empirical matter, Solow (and those who have followed in his tradition) solved (10) for the residual:

$$\left(\frac{1}{A}\right)\frac{dA}{dt} = \left(\frac{1}{y}\right)\frac{dy}{dt} - \beta\left(\frac{1}{k}\right)\frac{dk}{dt}.$$

(11)

Since everything on the right-hand side of (11) is directly measurable, this equation can be used to compute a time series on the growth rate of A. Doing so for the years 1909–1949 led Solow to the conclusion that only about one-eighth of the advance of $y = \frac{Y}{L}$ over those 40 years could be attributed to capital deepening, leaving about seven-eighths to ‘the residual’. This surprising (at the time) conclusion has, of course, changed numerically over the years as Solow's technique was refined (by, for example, Edward Denison, 1962 and others) and replicated on newer data. But the qualitative finding that technological change is far more important than capital deepening has held up extremely well.
What about the shape of the aggregate production function (9)? Does it really display the curvature implied by diminishing returns? By using his synthetic time series on A_t, it was a simple matter of arithmetic to compute a time series on $f(k_t) = \frac{y_t}{A_t}$ (which is *not* constant in the real world) and then to inspect the shape of the *f*(*k*) function. Eyeballing the scatter plot gave Solow ‘an inescapable impression of curvature, of persistent but not violent diminishing returns’ (1957, p. 318). And his regression estimates of a variety of nonlinear functional forms, including the Cobb–Douglas, rejected linearity. (In the Cobb-Douglas case, his original estimate of β was 0.35, not far from current estimates.)
One footnote to the history of economic thought is of interest here. Solow's empirical work included the years of the Great Depression and the Second World War, in contrast to the modern proclivity to omit those years. He noted that the data points for 1943–49 fell noticeably above the smooth estimated *f*(*k*) function, however. Solow commented on that fact, hazarded a few guesses as to why the war years might have been abnormal, but ended up confessing that ‘I leave this a mystery’. Honesty pays. Shortly thereafter Warren Hogan (1958) found a computational error in Solow's original paper and showed that, once the error was corrected, those seven data points were no longer aberrant.

Other work

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

I will be much briefer about some of Solow's other major works.

In a famous paper presented at the December 1959 meetings of the American Economic Association, Samuelson and Solow (1960) not only brought the Phillips curve to America but noted how what we now call a wage-Phillips curve can be transformed into a price-Phillips curve, and offered some reasons why that curve might not represent a stable menu of policy choices. Shortly thereafter, Solow joined the ‘New Frontier’ as a staff member of President Kennedy's original Council of Economic Advisers in 1961–62 and continued as a consultant through the Kennedy-Johnson years. That was quite a staff, including the likes of Kenneth Arrow and Arthur Okun. As Solow put it years later, ‘What I cherish about that experience was the conscious effort to use macroeconomic theory to interpret the world and, in a small way, change it’ (1990, p. 192).

In 1973 Solow and I co-authored ‘Does fiscal policy matter?’ (1973), a widely discussed contribution to the then-raging monetarist-Keynesian debate. The article established several counterintuitive implications that stem from the so-called government budget constraint, one version of which is:

$$\frac{dM}{dt} + P_b\frac{dB}{dt} = G + iB - (Y + iB), \tag{12}$$

where M is the money supply, B is the number of government bonds (paying interest rate i and selling for price P_b), G is government (non-interest) expenditure, and $T(\cdot)$ is the tax function. (While Blinder and Solow, 1973, used a fixed-price model, Tobin and Buiter, 1976, soon established parallel results for a full-employment model with a variable price level). The right-hand side of equation (12) is the budget deficit, and the left-hand side is the value of the money and/or bonds that must be issued to cover it. In steady-state equilibrium, neither M nor B is changing; so the budget must be balanced:

$$G + iB = T(Y + iB). \tag{13}$$

Blinder and Solow used a dynamic IS-LM model with wealth effects to prove the following surprising results:

- The model is always stable under money financing of deficits; but it can be either stable or unstable under bond financing of deficits, depending on parameter values.
- *If* the model is stable under debt financing, then bond-financed deficits are actually *more* expansionary in the long run than money-financed deficits, in stark contrast to standard teaching in macroeconomics. (Remember, monetarists were arguing at the time that bond-financed deficit spending had no effects on aggregate demand.)
- An open-market purchase that creates more money is expansionary in the short run (for the usual reasons), but will be cancelled out in the long run by the government budget constraint.

While the first result is purely technical, the other two have interesting intuitions behind them. Other things equal, more wealth, whether in the form of M or B, leads to higher output. (This assumes that government bonds are net wealth to the private sector.) In the short run, money creation is more expansionary than bond creation for the usual reasons. But for that very reason the budget deficit closes more slowly (and therefore with more cumulative wealth creation) under bond financing, leading ultimately to a larger multiplier (provided the system is dynamically stable), which is the second result.

A similar dynamic adjustment explains the third result, which anticipated parts of the message of both Sargent and Wallace's ‘Unpleasant monetarist arithmetic’ (1981) and the so-called fiscal theory of the price level. Start from an initial steady-state equilibrium with a balanced budget. An expansionary open-market operation, which raises Y, brings in more tax revenue, thereby creating a budget surplus. Under pure money financing $\frac{dB}{dt} = 0$, so (12) implies that the surplus leads to money destruction. And (13) assures us that, with unchanged fiscal policy, the process of money destruction must continue until all of the new money has been withdrawn. Thus, in a real sense, the government budget constraint renders conventional monetary policy (an open-market operation) *impossible* in the long run.

In 1981 Solow teamed with Ian McDonald to produce a widely cited model of wage bargaining. The paper offered a fresh approach to a question that dates back to the earliest days of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the Keynesian period: why, over business cycles, does employment fluctuate so much while real wages fluctuate so little? So-called implicit contract theories tend to predict that both employment (E) and real wages (w) will be stabilized; but empirical observation shows clearly that E is quite cyclical. McDonald and Solow used bargaining theory to explain the phenomenon of highly variable E combined with relatively stable w. Their arguments are lengthy and technical, and defy brief summarization. Suffice it to say that each of several specifications of how efficient bargaining might be conducted points toward this conclusion.

No summary of Robert M. Solow's contributions can be limited to reviewing his research, for Solow has also been a ‘good citizen’ *par excellence*. He has given his name, his wisdom, and his precious time to all manner of worthy causes and organizations, ranging from the Institute for Advanced Study to the Sierra Club. These include the National Academy of Sciences, the Manpower Demonstration Research Corporation, the Center for Advanced Study in the Behavioral Sciences, the Woods Hole Oceanographic Institution, the National Science Board, and the German Marshall Fund. There are many more.

Perhaps even more important are his remarkable achievements as a teacher of economics, particularly Ph.D. students. During the 1950s and early 1960s Samuelson, Solow, and a few of their friends created what came to be considered the world's pre-eminent economics department at MIT, an institution that had no great tradition in the subject. Generations of America's and the world's best graduate students came to study at MIT largely because Samuelson and Solow were there. While there are no objective measures of such things, Solow's clarity, wit, and mastery of a variety of subjects surely made him one of the finest teachers of economics who ever lived.

Perhaps for this reason, he was also the dissertation adviser of choice for scores of MIT's most promising graduate students over a period of time spanning 45 years. The list of Solow dissertation students, particularly in the 1960s, reads like an all-star team. In the two years 1966 and 1967 alone (based on completion dates), he supervised the Ph.D. dissertations of (in alphabetical order) George Akerlof, Robert Gordon, Robert Hall, William Nordhaus, Eytan Sheshinski, Joseph Stiglitz, and Martin Weitzman. Extending the time span back to 1956 and forward to 1971 brings in (now in chronological order) Alain Enthoven, Ronald Jones, Ronald Findlay, Peter Diamond, Ray Fair, Avinash Dixit, Jeremy Siegel, and the present author. More recently, Olivier Blanchard and David Romer were added to the list. It is a remarkable collection, for both its quality and its intellectual diversity. And probably every person on that list came to idolize Solow.

For most leading scholars, a list of students even half that extensive and distinguished would probably be considered their most enduring legacy. But in Robert M. Solow's case the research accomplishments are so fundamental, and have been and will be so enduring, that they eclipse even his amazing accomplishments as a teacher.

See Also

- endogenous growth theory
- fiscal theory of the price level
- Phillips curve (new views)

Selected works

1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

1958. (With R. Dorfman and P. Samuelson.) *Linear Programming and Economic Analysis*. New York: McGraw-Hill.

1960. Investment and technical progress. Proceedings of the First Stanford Symposium. In *Mathematical Methods in the Social Sciences, 1959*, ed. K. Arrow, S. Karlin and P. Suppes. Stanford, CA: Stanford University Press.

1960. (With P. Samuelson.) Analytical aspects of anti-inflation policy. Papers and Proceedings of the Seventy-second Annual Meeting of the American Economic Association. *American Economic Review* 50, 177–94.

1961. (With K. Arrow, H. Chenery, and B. Minhas.) Capital-labor substitution and economic efficiency. *Review of Economics and Statistics* 43, 225–50.

1973. (With A. Blinder.) Does fiscal policy matter? *Journal of Public Economics* 2, 319–37.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1981. (With I. McDonald.) Wage bargaining and employment. *American Economic Review* 71, 896–908.

1988. Autobiography. *Les Prix Nobel: The Nobel Prizes 1987*, ed. W. Odelberg. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/economics/laureates/1987/solow-autobio.html, accessed 17 June 2005.

1990. My evolution as an economist. *Lives of the Laureates: Ten Nobel Economists*, 2nd edn., ed. W. Breit and R. Spencer. Cambridge, MA: MIT Press.

Bibliography

Atkinson, A. 1969. The timescale of economic models: how long is the long run? *Review of Economic Studies* 36, 137–52.

Denison, E. 1962. *The Sources of Economic Growth in the United States and the Alternatives Before Us*. Supplementary Paper 13. Washington, DC: Committee for Economic Development.

Domar, E. 1946. Capital expansion, rate of growth and employment. *Econometrica* 14, 137–47.

Harrod, R. 1939. An essay in dynamic theory. *Economic Journal* 49(193), 14–33.

Hogan, W. 1958. Technical progress and production functions. *Review of Economics and Statistics* 40, 407–11.

Sargent, T. and Wallace, N. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5(3), 1–17.

Swan, T. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Tobin, J. and Buiter, W. 1976. Long-run effects of fiscal and monetary policy on aggregate demand. In *Monetarism: Studies in Monetary Economics*, vol. 1, ed. J. Stein. Amsterdam: North-Holland.

How to cite this article

Blinder, Alan S. "Solow, Robert (born 1924)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000444> doi:10.1057/9780230226203.1572

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sombart, Werner (1863 – 1941)

B. Schefold
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; luxury consumption; Marxism; planning; Sombart, W.

Article

Sombart was born in Ermsleben (Germany), the son of a well-to-do National Liberal member of the Prussian Diet. He studied economics, history, philosophy and law in Berlin, Pisa and Rome. In 1888 he received his Ph.D. from the University of Berlin and became an officer of the Bremen chamber of commerce. Two years later he was appointed extraordinary professor of political economy at the University of Breslau; in 1906 he became full professor at the Handelshochschule in Berlin and in 1917 transferred to the University of Berlin.

Sombart started his career as a left-wing advocate of social reform, influenced by Marxian theory. This was the reason why for a long time he could hold only second-rate positions within the German university system.

His dissertation on the economic and social conditions of the Roman campagna (1888) was brilliant and much less controversial than his later works.

An important work was his description of the German economy in the 19th century (1903). His outstanding study on the historical genesis of modern capitalism from its medieval origins to modern times (1902) may be considered Sombart's magnum opus. The really important edition was the second, published between 1919 and 1927, which differed completely from the first. Its three volumes treated three stages of capitalist development: early capitalism (Frühkapitalismus), high capitalism (Hochkapitalismus) – beginning with the industrial revolution in the 1760s – and late capitalism (Spätkapitalismus), starting with the First World War. The scope of this study was extremely broad. The reader is confronted with an amazing richness of facts. However, the data that Sombart presented were mostly second-hand and contained many speculative notions. Still, this work of ‘unsubstantial brilliance’ was ‘highly stimulating even in its errors’ (Schumpeter).

In the course of his research on *Der moderne Kapitalismus* Sombart published several special studies on the psychology and spirit of capitalism, on the Jews, on war and on luxury. In an outstanding study of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the bourgeois individual in general and the entrepreneur in particular he analysed the emergence of a certain economically oriented mentality and psychology, the ‘Wirtschaftsgeist’ (economic spirit), for the development of capitalism (1913a).

In a voluminous work on the role of the Jews in economic life (1911) he described the Jews much along the lines of Max Weber's analysis of Puritanism, as the most dynamic part of the population, introducing ‘capitalist spirit’ into commerce and industry. At this time, Sombart was *not* anti-Semitic: on the contrary, he perceived the Jews’ contribution to the rise of capitalism positively, regarded them as one of the most valuable ‘species’ of mankind (1912, p. 56) and was in favour of Zionism as the national renaissance of the Jewish people. However, the foundations for his later anti-Semitic turn were laid: his description of the characteristics of the Jews treating them as a ‘species’ was full of prejudices and exaggerations and included a discussion of their ‘race’ and ‘blood’ peculiarities.

Sombart further analysed the development of luxury consumption, which he connected in a very original way with the erotic, and its economic importance as a creator of new markets and industries (1913b, vol. 1). In a similar way Sombart treated war as a creator of new markets due to the services and goods required by the military (1913b, vol. 2). Two years later, during the First World War, Sombart wrote a chauvinistic, strongly anti-English book, which glorified war and militarism (1915).

The best way to observe his political views is to follow his discussion of Marx: Sombart was never a Marxist. But when the third volume of *Capital* appeared, he praised Marx as an outstanding thinker and described his theory in a very positive way (1894), which in turn was warmly welcomed by Engels. Also, in the first edition of his famous work on socialism and social movements (1896), he discussed Marx from a sympathetic point of view. Its tenth edition, now titled *Der proletarische Sozialismus* (*Marxismus*) (1924), was violently anti-socialist and full of hatred and personal insults against Marx. However, this did not hinder Sombart from stating three years later that his *Der moderne Kapitalismus* was written in the Marxian spirit and that he regarded it in a certain way as the conclusion of Marx's work (1927, p. xix).

What had occurred was an evolution in Sombart's assessment of capitalist development. While he still appreciated Marx as a historian of capitalism, he now disliked the latter's optimistic view of the future, his regard for capitalism as the creator of the better world to come (1927, p. xx). Marx had been an admirer of technical progress and of the historical forces which fostered it. Sombart no longer believed that industrial development was automatically beneficial and he realized its destructive potential. He contrasted the uniformity and ugliness of modern civilization with the cultural variety of the pre-industrial past.

This enmity towards industrial development and what he called the ‘economic age’ was the reason why Sombart sought to ally himself with right-wing anti-capitalism and temporarily turned to fascism. When the Nazis came to power he made a contribution towards a programme of German (National) Socialism (1934). Contrary to proletarian socialism, which accepted the industrial society and only intended to redistribute its surplus, Sombart perceived German socialism as rejecting the industrial age (1934, pp. 160–8). Beside the corporative state, the Führerprinzip (leader principle), a state interventionist regulation of the German economy, autarky and a partial re-agrarianization of Germany, he advocated state planning and, as a key idea, control of technological development based on what would now be called technology assessment (1934, pp. 263–7). However, his proposals were not welcomed by the Nazis, who, as Sombart was later to realize, intended to use the most advanced technologies available in order to win political hegemony.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

It is not easy to regard Sombart as member of an economic school. He first rejected and then accepted the deductive method. Although he claimed that *Der moderne Kapitalismus* bridged the gap between the abstract-theoretical and the empirical-historical method (1927, p. xvi), there is no doubt that this book lacked an analytical framework and, though not without theory, was mainly a historical description of a large mass of facts. Sombart even ‘out-Schmollered Schmoller’ (Schumpeter) and had to be regarded as belonging to the younger, or, as Schumpeter called it, the ‘youngest’ historical school. Sombart himself, however, would certainly not have admitted that. He distinguished between three kinds of political economy (1930): the ‘richtende Nationalökonomie’ (judging economics), which was intended to decide what was right and wrong, and whose scientific character Sombart denied; the ‘ordnende Nationalökonomie’ (ordering or systematizing economics) that tried to apply quantitative exact methods, and, finally, the ‘verstehende Nationalökonomie’ (understanding or interpretative economics), which should be a ‘Geistwissenschaft’ (science of social mind) and was both theoretical and historical and tried to grasp the motives of economic life. It goes without saying that Sombart regarded his own work as being in the tradition of the latter.

Sombart extended his claims to being a theorist beyond economics. He also perceived sociology as a ‘Geistwissenschaft’ and developed a new type of sociology, which he called ‘Noo-Soziologie’, which was supposed to be a general theory of culture (1936).

He remains one of the most brilliant and interesting personalities of the German economics profession, being a gifted writer with profound historical insights. While many works of the Historical School are boring collections of facts, the best of Sombart's books are sparkling and still make fascinating reading.

See Also

- Historical School, German

Selected works

1888. Über Pacht- und Lohnverhältnisse in der römischen Campagna. Ph.D. dissertation, Berlin.

1894. Zur Kritik des ökonomischen Systems von Karl Marx. *Archiv für soziale Gesetzgebung und Statistik* 7, 555–94.

1896. *Sozialismus und soziale Bewegungen im 19. Jahrhundert*. Jena: Gustav Fischer. Trans. as *Socialism and the Social Movement*, London: Dent; New York, Dutton, 1909. 10th edn published as *Der Proletarische Sozialismus* (*Marxismus*), 1924.

1902. *Der moderne Kapitalismus: Historisch-systematische Darstellung des gesamteuropäischen Wirtschaftslebens von seinen Anfängen bis zur Gegenwart*. 2 vols, Leipzig. 2nd edn, 3 vols, Munich and Leipzig: Duncker & Humblot, 1916–27.

1903. *Die deutsche Volkswirtschaft im neunzehnten Jahrhundert*. Berlin: Georg Bondi.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1909. *Das Lebenswerk von Karl Marx*. Jena: Gustav Fischer.

1911. *Die Juden und das Wirtschaftsleben*. Leipzig: Duncker & Humblot. Trans. as *The Jews and Modern Capitalism*, London: T.F. Unwin, 1913 .

1912. *Die Zukunft der Juden*. Leipzig: Duncker & Humblot.

1913a. *Der Bourgeois: Zur Geistesgeschichte des modernen Wirtschaftsmenschen*. Munich and Leipzig: Duncker & Humblot. Trans. and M. Epstein as *The Quintessence of Capitalism: A Study of the History and Psychology of the Modern Business Man*, New York: Dutton, 1915.

1913b. *Studien zur Entwicklungsgeschichte des modernen Kapitalismus*. Munich and Leipzig: Duncker & Humblot. Vol. 1: *Luxus und Kapitalismus*. Trans. into English as a report under the auspices of the Works Progress Administration and the Department of Social Science, Columbia University, Project number 465–97–3–81. Vol. 2: *Krieg und Kapitalismus*.

1915. *Händler und Helden. Patriotische Besinnungen*. Munich and Leipzig: Duncker & Humblot.

1930. *Die drei Nationalökonomien*. Munich and Leipzig: Duncker & Humblot.

1934. *Deutscher Sozialismus*. Berlin: Buchholz & Weisswange. Trans and ed. K.F. Geiser as *A New Social Philosophy*, Princeton: Princeton University Press; Oxford: Oxford University Press, 1937.

1936. *Soziologie: Was sie ist und was sie sein soll*. Berlin: Sitzungsberichte der Preussischen Akademie der Wissenschaften.

Bibliography

Crosser, P.K. 1941. Werner Sombart's philosophy of National-Socialism. *Journal of Social Philosophy* 6, 263–70.

Mitchell, W.C. 1929. Sombart's Hochkapitalismus. *Quarterly Journal of Economics* 43 February, 303–23.

How to cite this article

Schefold, B. "Sombart, Werner (1863–1941)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000180> doi:10.1057/9780230226203.1573

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sonnenfels, Joseph von (1733 – 1817)

K. Tribe
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Sonnenfels, J. von; Justi, J. H. G. von; Forbonnais, F. V. G. de

Article

Sonnenfels was born of Jewish parents who shortly afterwards converted to Catholicism; the family moved in 1744 from Moravia to Vienna, where the father taught oriental languages. Joseph first served in the army from 1749 to 1754, when he began to study law and literature at Vienna University. A prominent member of the Enlightenment literati, he was in late 1763 appointed to the newly founded chair in ‘Police and Cameralistic Sciences’ at the University of Vienna. Until his death he was prominent in constitutional reform, also engaging in a campaign for the abolition of torture and of usury. The textbook which he wrote for his own teaching, the *Grundsätze der Polizei Handlungs- und Finanzwissenschaft* (1765, 1769, 1776), remained the official text in the Austrian Empire until 1848, running to eight editions and several abbreviated teaching editions.
The *Grundsätze* devotes a volume each to police, commerce and finance. The leading idea running through all three is the importance of a large population gainfully employed for the general welfare of the state. Coupled with this is a conception of the accumulation of wealth as the ‘multiplication of means of subsistence’, governed however by the necessity of maintaining equilibrium in the society, which is the task of police. The functioning of police is therefore tied less directly to economic welfare than is the case with Justi, to whom Sonnenfels makes reference. The treatment of commerce owes a great deal to Forbonnais’ *Elémens du commerce*, who also laid emphasis on the advantages of a large population and the need for proportion within it. But while Sonnenfels takes much from Forbonnais, he remains more concerned with general political order than the economic structure of an advancing society.

Selected works

1765, 1769, 1776. *Grundsätze der Polizei, Handlungs- und Finanzwissenschaft*. 3 vols. Vienna: Camensina. 8th and final edn, 1818–22.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Tribe, K. "Sonnenfels, Joseph von (1733–1817)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000181> doi:10.1057/9780230226203.1574

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

South Sea bubble

Hans-Joachim Voth
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The South Sea bubble resulted from an equity-for-government debt swap that had gone wrong, and occurred in England in 1720. Prices of South Sea Company stock rose sharply following the announcement of the scheme, and collapsed eight months later. Frequently cited as an example of investors' folly, the factors driving the sharp rise and fall of South Sea Company share prices have remained controversial. The so-called Bubble Act of 1720, passed before the bubble peaked, restricted the development of a vibrant market in publicly traded companies for a century.

Keywords

Bubble Act (1720); Law, John; Mississippi bubble; South Sea bubble; government debt

Article

The South Sea Company was founded in 1711, in the expectation that peace between Spain and England after the end of the War of the Spanish Succession would produce profitable trading opportunities with the 'South Seas' (that is, Spanish America). The company's trading activity remained intermittent and unprofitable throughout the 1710s. In 1719, a new scheme was launched – the conversion of government debt into equity of the South Sea Company. Debt-holders of the 1710 lottery loan were offered the option to convert their holdings into company shares. The government agreed to make interest payments to the company instead of to debt-holders. As old (and illiquid) loans were swapped for liquid company shares, debt-holders gained. The government negotiated a lower rate of interest, and the South Sea Company made a modest profit. The 1719 equity-for-debt swap is generally seen as Pareto-improving.

The 1720 conversion scheme differed in important ways. Key elements included (*a*) the absence of a fixed conversion ratio – higher prices of South Sea stock meant that more debt could be bought with each share, (*b*) issuance of new stock on instalment, with only a small down payment required, (*c*) massive lending against shares, and (*d*) a high degree of corruption in the awarding of the contract. The South Sea conversion also shared important characteristics with John Law's Mississippi scheme in France, which produced a similar run-up (and crash) of prices half a year earlier.

Both the Bank of England and the South Sea Company competed for the contract to convert government bonds into equity. After bribes to MPs, ministers, and members of the court (of about £1.3 million), the South Sea Company won the right to perform the conversion in March 1720. By this time, the price of its shares had increased to 255, from 128 at the beginning of the year. The share prices of other companies moved up and down in parallel with South Sea stock, but less sharply (see Figure 1). The company proceeded to issue fresh shares in four subscriptions, and offered to convert debt into shares on (modestly) generous terms. By late June, prices had risen to 765, and forward prices during the summer rose as high as 950. When regular trading resumed, prices began to weaken, but the fourth subscription was still strongly oversubscribed. In September, prices fell quickly. By the year end they had almost declined to their January level.

Figure 1
Share price of major English companies, 1718–22. *Source*: Neal (1990); data from ICPSR, Study No. 1008.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

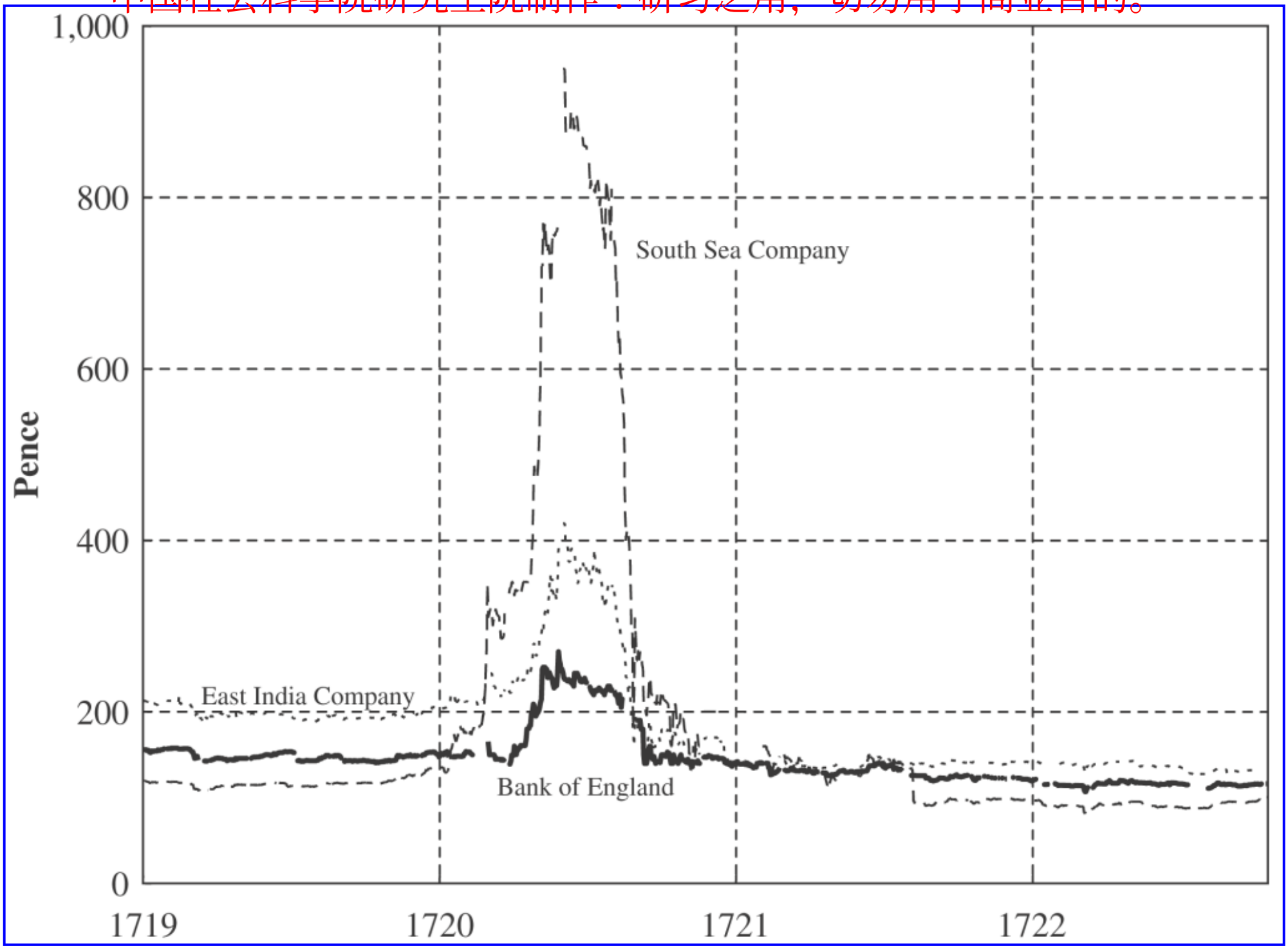

Interpretations

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Since Mackay's classic *Extraordinary Popular Delusions and the Madness of Crowds*, the South Sea bubble has often been cited as a prime example of irrational investor behaviour. In contrast, Peter Garber (2000) argued that the share prices increased in line with ‘changing view(s) of market fundamentals’. If the scheme had succeeded in improving economic conditions in England as a whole (as John Law's logic would predict; Verde, 2004), the firm's large capital base might have allowed it to pursue profitable ventures. Yet most of these remained vague, and the company had no track record of successfully making money from anything other than financial transactions. It is doubtful whether future profits could ever have been high enough to justify the company's market capitalization in the summer of 1720. Even Garber accepts that prices above 400 are hard to square with reasonable expectations of future profits. Easy credit, investor preferences for lottery-like payoffs (as a result of shares being sold with only a small down payment), and restricted free float (caused by company lending against its own shares) may have contributed to the start of the bubble.

Recent work has focused on the reasons why the bubble, once under way, could have expanded greatly. Dale (2004) argues that apparent mispricing of subscription receipts proves investor irrationality, while others have argued that the gap can be explained by the option-like nature of receipts. Temin and Voth (2004) examined the trades of a goldsmith bank, Hoare's, which made large profits buying and selling South Sea stock in 1720. They argue that the bank was aware of the overpricing, but invested in South Sea stock regardless. Predictability of investor sentiment made it rational to ‘ride’ the bubble, and to sell out with a profit as soon as it began to deflate. This strategy is similar to hedge fund behaviour on Nasdaq in the late 1990s (Brunnermeier and Nagel, 2004). If other large investors faced similar incentives, the lack of a coordinated early attack becomes easier to understand. The role of market microstructure imperfections was probably limited, as opportunities to sell short were abundant. However, the nature of the settlement process and the artificial reductions of free float engineered by the company may have contributed to the bubble.

Consequences

The rise and crash of share prices in 1720 had few direct economic consequences. As prices declined, former debt-holders demanded compensation. Parliament investigated the scheme in which it had played such an important role. Directors had most of their assets expropriated. In contrast to the resolution of the Mississippi bubble in France, those who had tendered government bonds for company shares received partial compensation in the form of fresh government debt. The political consequences were possibly more formidable than the immediate economic repercussions. Leading politicians who had taken bribes, such as the Chancellor of the Exchequer, John Aislabie, were forced out of office and incarcerated. Robert Walpole, sometimes referred to as England's first prime minister, distinguished himself both through his opposition to the scheme and competent handling of its fallout. He succeeded Aislabie at the Exchequer and remained in power until 1742.

The collapse of the South Sea bubble is sometimes seen as a factor behind the Bubble Act. This appears to be erroneous, as the Act was passed before the bubble deflated (Carswell, 1993). Its passage and rigorous enforcement after the summer of 1720 probably owed more to the company's efforts to support its own sagging share price. Because of the Act, new equity issues became very rare for almost a century. The Act was repealed only in 1825.

See Also

- bubbles
- Law, John

Bibliography

Brunnermeier, M. and Nagel, S. 2004. Hedge funds and the technology bubble. *Journal of Finance* 59, 2013–40.

Carswell, J. 1993. *The South Sea Bubble*. Stroud: Sutton.

Dale, R. 2004. *The First Crash*. Princeton: Princeton University Press.

Garber, P. 2000. *Famous First Bubbles: The Fundamentals of Early Manias*. Cambridge, MA: MIT Press.

Mackay, C. 1995. *Extraordinary Popular Delusions and the Madness of Crowds*. San Francisco: Essential Library.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Neal, L. 1990. *The Rise of Financial Capitalism: International Capital Markets in the Age of Reason*. New York: Cambridge University Press.

Temin, P. and Voth, H.J. 2004. Riding the South Sea bubble. *American Economic Review* 94, 1654–68.

Verde, F. 2004. Government equity and money: John Law's system in 1720 France. Working paper, Federal Reserve Bank of Chicago.

How to cite this article

Voth, Hans-Joachim. "South Sea bubble." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000186> doi:10.1057/9780230226203.1575

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sovereign debt

Kenneth M. Kletzer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The liabilities of sovereign states are not directly enforceable, for example by court-imposed transfer of collateral. Sovereign borrowing must be sustained by the prospect that indirect sanctions follow default. The credibility of sanctions for default and the roles of third parties and reputations for motivating repayment are discussed. The interaction of sovereignty and externalities between creditors complicates the renegotiation of sovereign debt. The importance of the collective action problem for debt restructuring is reviewed.

Keywords

asymmetric commitment; asymmetric information; bankruptcy; bonds; collective action; commitment; consumption smoothing; debt restructuring; default; extensive-form games; free-rider problem; implicit contracts; international capital flows; international trade; moral hazard; permanent loan autarky; rationing; reputation; risk sharing; sanctions; sovereign debt; sovereign immunity; willingness to pay

Article

A straightforward definition of sovereign debt is simply that it is debt issued by a sovereign state. Sovereignty conveys supreme legal authority within the geographical boundaries of the nation, giving national authorities autonomy over the regulation of economic activity inside the country through legislation, administration, and judicial enforcement. This means that foreign governments cannot interfere with economic activities within sovereign boundaries and cannot enforce contractual relationships without the cooperation of national authorities. Sovereignty denies creditors the right to reach inside national borders to confiscate assets or attach sources of revenue, public or private, to satisfy outstanding debts.

The importance of sovereignty for international capital markets is evidenced by the frequency of sovereign debt crises and defaults in history. Over the last century, debt crises have been associated with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
default on international debt by governments in emerging market economies, but between the 16th and 19th centuries, sovereign default was similarly frequent in Continental Europe (Reinhart, Rogoff and Savastiano, 2003, survey the history of default). Although international debt is at the centre of recent crises, the economics of sovereign debt apply to debt issued in either international or domestic markets held by either foreign or domestic residents.

Repayment incentives

The first concern of the analytical literature on sovereign debt is what motivates borrowers to repay. When the terms of a debt contract cannot be enforced directly by a court, a sovereign debtor can be expected to honour its debts only if it faces adverse consequences for default. The literature focuses on the possibility of sanctions for debt repudiation or default which include the disruption of international trading opportunities by foreign governments or private market participants. For example, a repudiating debtor could face a trade embargo, suspension of trade preferences or loss of access to international capital markets. Each is a case of reduced access to international trade, whether trade takes place contemporaneously or across time.

A consequence of the observation that indirect sanctions provide the incentive to repay debt is that lending to sovereigns is constrained by the willingness of the debtor to repay rather than by its ability to repay. The concept of willingness to pay was first applied to sovereign debt in an analytical model by Eaton and Gersovitz (1981) following its use in earlier writings (for example, by Wallich, 1943). Eaton and Gersovitz emphasize that repayments are made out of the enlightened self-interest of the borrower and that debt limits are determined by the expected present value of the cost to the debtor of sanctions that will be imposed if debt is repudiated. They demonstrate two versions based on punishments given by disruptions of commodity trade and credit embargoes, respectively. Some immediate implications of the Eaton and Gersovitz model are that credit is rationed in equilibrium and that more severe sanctions increase loan supply.

It is helpful to compare lending sustained by the threat of indirect sanctions to conventional domestic lending with collateral. The use of collateral relies on a legal system to enforce debtors' property rights in collateral assets and to transfer the ownership of these assets to creditors in specific events. Under a collateralized loan, the debtor repays to avoid losing collateral, and a solvent debtor should do so only if the value of the collateral exceeds the required repayments. The value of loan collateral determines a non-sovereign debtor's willingness to pay. In a time-consistent equilibrium, lenders should not lend more than the discounted present value of the assets that secure the loan.

Sovereign borrowing is analogous to legally enforced borrowing with collateral if the imposition of sanctions subsequent to default is exogenous. In the first version of the Eaton and Gersovitz (1981) model, static trade sanctions are imposed exogenously. One insight of the paper is that indirect penalties such as the disruption of commodity trade require the cooperation of third parties. Lenders, public or private, can count on repayment only if sovereign nations impose barriers to a recalcitrant debtor's trade. Eaton and Gersovitz consider how repayment incentives might arise within the credit relationship itself using the threat of a credit embargo in response to debt repudiation. Lending and repayment are derived as perfect equilibrium outcomes in an extensive-form game in which a risk-averse borrower with random income seeks to smooth consumption over time. The borrower repays if the net gain from future credit access is at least as large as the utility cost of the current repayment. The gains from credit access

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
depend on the amount that creditors will lend which in turn depends on the expected future repayments by the borrower. Perfect equilibrium is sustained in this model by the threat of permanent loan autarky. The strategies of creditors are restricted so that this punishment is imposed after a borrower defaults in any amount. Therefore, borrowers either repay in full or repudiate their entire debt.

The restriction to a simple choice between full repayment and debt repudiation is one drawback of the basic model. If the gains from international trade or credit market access are stochastic, then allowing state-contingent repayments can raise welfare for borrowers or lenders. Grossman and van Huyck (1988) argue that default and debt renegotiation may result from an implicit state-contingent contract guided by a standard loan contract with fixed repayments. Several authors model debt renegotiation as the outcome of an implicit state-contingent contract with stochastic gains from trade.

Bulow and Rogoff (1989a) analyse a repeated model of bargaining over repayments supported by the threat of trade sanctions. By making an initial loan, a creditor acquires the right to invoke a trade embargo if the debtor fails to repay. Each period the creditor and debtor bargain over the amount paid by the debtor to avoid an embargo. This provides a foundation for the exogenous penalty in the first version of the Eaton and Gersovitz model. The repayment is a price paid by the debtor to purchase the right to impose sanctions from the creditor for one period at a time. The enforcement of property rights is essential to this model. The sovereign initially has the right to free trade which it sells for the price of the loan to the creditor. The creditor then rents access to free trade back to the sovereign debtor. Thus, the Bulow and Rogoff model parallels lending with collateral in that it requires the enforcement of property rights by disinterested parties.

The possibility that lending to sovereigns might be enforced by market participants themselves by denying credit market access to defaulting borrowers is addressed by several authors. Bulow and Rogoff (1989b) question the force of reputations-based punishments for sovereign default and argue that official imposition of sanctions or enforcement of creditor rights is necessary. In particular, they argue that permanent loan autarky as proposed by Eaton and Gersovitz in the consumption-smoothing model of debt does not deter sovereign default when debtors can accumulate foreign assets. They assume that a sovereign borrower can purchase a foreign interest-bearing deposit which it can draw against in default. Under this assumption, a sovereign borrower raises its welfare by defaulting and saving the amount it would have repaid in a foreign deposit. In this model, international capital flows are sustainable in perfect equilibrium, but only if the sovereign lends rather than borrows. An interpretation of Bulow and Rogoff is that reputations alone cannot sustain sovereign borrowing.

This result received a fair amount of attention because it implies that sovereign borrowing cannot be self-enforcing and that earning a reputation for default is insufficient to deter opportunistic debtor behaviour. The special assumption of the paper is that foreign financial obligations to a government are enforceable while its financial obligations are not. The foreign creditors can commit to make future payments to the sovereign even when it is not in their self-interest to do so as events unfold. Contrary to the interpretation of Bulow and Rogoff, the capacity to commit requires exogenous enforcement by another sovereign state.

Kletzer and Wright (2000) model capital flows for smoothing a sovereign borrower's consumption in the absence of any exogenous means of contract enforcement and show that asymmetric commitment is essential for the conclusion of Bulow and Rogoff (1989b). The paper demonstrates self-enforcing equilibrium in an economy that treats both sides of the market as sovereign supported by punishments

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
proof to renegotiation. Permanent credit embargoes are not credible threats because the gains from risk-sharing that motivate lending in the first place give incentives for forgiveness and new lending. Kletzer and Wright prove that sovereign borrowing is sustainable with free entry by new counterparties, as envisioned by Bulow and Rogoff, so that reputations alone are sufficient to enforce repayments. The general result is that credit markets are possible under the anarchy that characterizes international relations between sovereigns.

Kletzer and Wright show that the equilibrium is implemented by one-period loan contracts with state-contingent repayments that are not negative. State-contingent repayments are interpreted as the outcome of an implicit contract guided by non-contingent loan contracts following Grossman and van Huyck (1988) and Bulow and Rogoff (1989a). Non-commitment is exactly the assumption that any net payment (a new loan) made by a lender to the sovereign is voluntary. Introducing one-sided commitment allows conventional insurance, while the absence of commitment is consistent with renegotiation of conventional loans. Kletzer and Wright show that the opportunistic acceptance of defaulting borrower's deposits can only eliminate sovereign borrowing if international insurance contracts are exogenously enforced by other sovereigns.

A different approach to demonstrating the self-enforcement of sovereign borrowing through reputations is taken by Cole and Kehoe (1998). In a model with asymmetric information, actions reveal information about the borrower to counterparties in other financial and trading relationships. The borrower then cares about reputational spillovers to these other relationships (this possibility is noted by Bulow and Rogoff, 1989b). They demonstrate equilibrium repayment in the presence of one-sided enforcement of the sovereign's own foreign loans.

Sovereigns may enforce contractual obligations of domestic borrowers to foreigners because they recognize national benefits from doing so. Sovereign states set the rules inside national borders so that enforcement of private obligations is at the discretion of the sovereign. Until the 1950s, sovereign immunity applied with respect to assets held abroad. A commercial creditor could not pursue legal remedies against a sovereign debtor without its consent in either the United States or UK. By the mid-1970s both countries adopted more restrictive legal theories of sovereign immunity that allow private parties to seek remedy in home courts if the dispute concerns commercial transactions (Brownlie, 2003). Contemporary emerging market bond issues are overwhelmingly issued under the governing law of a few financial centres (primarily, New York and London) and incorporate waivers of sovereign immunity, thus voiding one-sided commitment opportunities.

Insights about the credibility of sanctions can be gained from the historical experience of sovereign default. Although trade sanctions were imposed on some sovereign defaulters before the 1930s, creditor nations appear reluctant to interfere with the international trade of countries in default (see, for example, Eichengreen and Portes, 1989). Ozler (1993), Lindert and Morton (1989) and others find that historically default had limited impact on loan terms for sovereign debtors, suggesting that credit market sanctions may be weak. Short-lived punishment or exclusion from borrowing while existing debts are renegotiated and settled, however, is consistent with credible punishment. Esteves (2005) studies debt renegotiation between 1870 and 1913 and finds that capital market access was disrupted during renegotiation but regained after resolution was reached. Many recent financial crises and debt restructurings conform to this pattern.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Debt restructuring and current issues

Debt renegotiation and restructuring is central to achieving gains from international risk sharing with limited financial instruments. The protracted and costly process of sovereign default resolution and debt restructuring remains the subject of policy concern and debate (for a review, see Eichengreen, 2003). The difficulty of achieving the acquiescence of a large number of creditors to a sovereign debt restructuring agreement has been repeatedly demonstrated for more than a century.

Tirole (2002) identifies the interaction between problems of common agency in lending and debtor sovereignty with inefficiencies in international capital markets that produce crises. He observes that debtor sovereignty impedes solutions adopted in domestic financial markets to the externalities between various creditors due to common agency in lending. With legal enforcement of contracts, bond and loan covenants can establish property rights across heterogeneous creditors in the event of default so that individual creditors internalize the impact of their actions on other creditors. In corporate debt markets, the power of the sovereign can be used to mitigate moral hazard on the part of both the debtor and its creditors, in contrast to the case of sovereign debt.

One of the problems associated with common agency in sovereign lending is the difficulty of achieving collective action among creditors for resolving default. In default or bankruptcy, an important role of court proceedings is to aggregate various debts and resolve conflicting claims between creditors to reach completion of financial restructuring. For a sovereign debtor, creditors who accept reduced repayments in default increase the value of debt held by creditors who do not participate. Free-riding of this type is related to the public goods nature of the capacity to impose sanctions. Any debt rollover, reduction or exchange can be prone to hold out creditors prolonging and contributing to crises. A minority of creditors can also interfere with a debt restructuring by exercising their legal rights under the governing law under which bonds were issued.

The collective action problem for sovereign debt renegotiation (and, hence, lending to sovereigns) has been addressed in various ways over time. Prominent approaches include the formation of bondholder committees, representation of creditors by banks, and official sector intervention. In recent years, two approaches, neither new, have received considerable attention. One of these concerns contractual innovation. An example of the issue is the ability of individual creditors to block debt swaps under unanimous consent clauses required for US corporate borrowing. Collective action clauses allowing less than unanimous consent to restructurings were included in major sovereign debt issues in the United States after active encouragement. An analysis of these clauses is given by Eichengreen, Kletzer and Mody (2004), and the possibility of voiding benefits for hold-outs under unanimous action clauses is discussed by Buchheit and Gulati (2000). The other approach is statutory innovation which may challenge debtor sovereignty. Although such proposals have been advanced over the years, the prospects for an international debt restructuring mechanism are diminished for now in favour the pursuit of market-based innovation.

Bibliography

Brownlie, I. 2003. *Principles of Public International Law*, 6th edn. New York: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Buchheit, L. and Gulati, G. 2000. Exit consents in sovereign bond exchanges. *UCLA Law Review* 48, 59–84.

Bulow, J. and Rogoff, K. 1989a. A constant recontracting model of sovereign debt. *Journal of Political Economy* 97, 155–78.

Bulow, J. and Rogoff, K. 1989b. LDC debt: is to forgive to forget? *American Economic Review* 79, 43–50.

Cole, H. and Kehoe, P. 1998. Models of sovereign debt: partial versus general reputations. *International Economic Review* 39, 55–70.

Eaton, J. and Gersovitz, M. 1981. Debt with potential repudiation: theory and estimation. *Review of Economic Studies* 48, 289–309.

Eichengreen, B. 2003. Restructuring sovereign debt. *Journal of Economic Perspectives* 17(4), 75–98.

Eichengreen, B., Kletzer, K. and Mody, A. 2004. Crisis resolution: next steps. *Brookings Trade Forum 2003*, 279–337.

Eichengreen, B. and Portes, R. 1989. Setting defaults in the era of bond finance. *World Bank Economic Review* 3, 211–39.

Esteves, R. 2005. Quis custodiet quem? Sovereign debt and bondholders' protection before 1914. Mimeo, Department of Economics, University of California, Berkeley.

Grossman, H. and van Huyck, J. 1988. Sovereign debt as a contingent claim: excusable default, repudiation, and reputation. *American Economic Review* 78, 1088–97.

Kletzer, K. and Wright, B. 2000. Sovereign debt as intertemporal barter. *American Economic Review* 90, 621–39.

Lindert, P. and Morton, P. 1989. How sovereign debt has worked. In *Developing Country Debt and the World Economy*, ed. J. Sachs. Chicago: University of Chicago Press.

Ozler, S. 1993. Have commercial banks ignored history? *American Economic Review* 83, 608–20.

Reinhart, C., Rogoff, K. and Savastano, M. 2003. Debt intolerance. *Brookings Papers on Economic Activity* 2003(1), 1–74.

Tirole, J. 2002. *Financial Crises, Liquidity and the International Monetary System*. Princeton, NJ:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Princeton University Press.

Wallich, H. 1943. The future of Latin American dollar bonds. *American Economic Review* 33, 324–35.

How to cite this article

Kletzer, Kenneth M. "sovereign debt." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000498> doi:10.1057/9780230226203.1576

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Soviet and Russian research, development and innovation

Julian Cooper

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

For its level of economic development, the Union of Soviet Socialist Republics (USSR) possessed a research and development (R&D) system that was impressive in terms of personnel and funding, yet the country's achievements in technological innovation were, with the partial exception of military industry, unimpressive. Following the collapse of Communism, the R&D system substantially contracted in scale. After a decade of post-Communist transformation in Russia, funding began to recover but the new market conditions did not prove conducive to successful innovation on a significant scale. The National Innovation System of present-day Russia is still marked profoundly by the Soviet legacy.

Keywords

diffusion of technology; human capital; innovation; intellectual property rights; national innovation system; research and development; Soviet and Russian science policy; soft budget constraint; venture capital

Article

Understanding the functioning and performance of the Soviet R&D system and the innovation-averse character of the planned economy has posed a challenge to economists and science policy specialists since the 1950s. It was the ‘sputnik’ shock of 1957 that triggered serious analysis of the Soviet R&D system. Initial research (De Witt, 1961; Korol, 1965) focused on its scale in terms of personnel and on its basic organizational and behavioural characteristics, a landmark being the 1969 Organisation for Economic Co-operation and Development (OECD) report on science policy in the USSR (Zaleski et al., 1969). This report highlighted the organizational separation of research from production, the dominant role not only in basic research but also in much applied work of the USSR Academy of Sciences, which played a central role in the overall science policy of the country and occupied a position of prestige in society, and the relatively modest role in R&D of the higher educational sector. Later works contributed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to a better understanding of the Soviet research system, its social context and the influence on science and technology of the Communist Party (for example, Kruse-Vaucienne and Logsdon, 1979; Lubrano and Solomon, 1980; and Fortescue, 1986).

By this time there was an appreciation that, notwithstanding the scale of the R&D system, somewhat overstated by Soviet official statistics, overall innovative performance was poor. A pioneering attempt to analyse the economics of innovation was that of Joseph Berliner (1976). For Berliner, Soviet industry was characterized by an all-pervasive lack of responsibility, risk aversion and weak incentives for successful innovation, coupled with mild penalties for failure in the absence of competition. Attempted economic reforms had achieved little to improve the situation, and Berliner was not optimistic that any major structural change would boost innovative performance. The outcome of inadequate innovation was the focus of work undertaken by researchers at the University of Birmingham (Amann, Cooper and Davies, 1977). This demonstrated that from the mid-1950s to the mid-1970s in many sectors of the economy there had been no diminution in the technological gap between the Soviet Union and leading industrial economies. It was found that the rate of diffusion of major new technologies tended to be slower than the rates typical of advanced market economies. Further research into industrial innovation (Amann and Cooper, 1982) built on the findings of Berliner, but with more attention to the specific historical conditions in which particular industrial sectors developed. While the Soviet economic system was by its nature averse to innovation, some sectors showed better performance than others, above all the defence industries, where high-level political intervention and priority resource allocation to some extent overcame the inertia of dysfunctional economic institutions. That these institutions and their implications for the development of science and technology had deep roots was shown by the pioneering work of Bailes (1978) and Lewis (1979). Further insights into poor performance in the adoption of innovations was later provided by the application of a principal–agent framework, notably by Dearden, Ickes and Samuelson (1990).

In the late 1960s and 1970s, recognizing the limits of the domestic civil R&D and innovation systems, the Soviet authorities attempted to boost technological progress by importing the latest technology from the West, with the automobile industry as a pioneer (Holliday, 1979). The results were disappointing. As shown by Hanson (1981), the Soviet system was unable to reap the productivity gains potentially available and was incapable of reproducing the technology itself to diffuse the achievements more widely.

Soviet R&D was heavily militarized. At various times the authorities attempted to harness the resources and skills of the military sector to boost performance in the civil economy. However, ‘spin-off’ from the defence sector was modest; and when a policy of ‘conversion’ was pursued in the final years of the Soviet regime under General Secretary Gorbachev the practical results were modest because of organizational barriers, inadequate incentives and a marked quality gap between the military and civilian sectors.

In terms of personnel and spending, the USSR had a research system of significant size. However, appreciation of its true scale and the productivity of Soviet science was complicated by data problems. According to Soviet official statistics, in 1990 expenditure on science was five per cent of national income. However, this figure was inflated by double counting. Researchers of the Centre for Science Research and Statistics (CSRS), Moscow, concluded that for Russia in 1990 the actual share was two per cent of GDP. Similarly, official data overstated the number of R&D personnel, partly because

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
employment was not expressed in terms of full-time equivalents.

In the USSR the innovation process was always understood, implicitly by government officials and often explicitly by economists, as a linear process; that is, new products and processes were developed on the basis of ideas and inventions originating in basic and applied research, and then development, after which they were ‘introduced’ into the sphere of production and then diffused more widely. Only in the very final years did some analysts become aware of the work of Chris Freeman and other Western science policy specialists who challenged the linear model and argued for a richer understanding involving feedback relationships. However, in present-day Russia the linear model is still widely met and increased spending on science often presented as the principal means of promoting innovation.

Since the collapse of the USSR, there has been no attempt by either Russian or other scholars to revisit the Soviet R&D and innovation system in a comprehensive manner. However, the work of Harrison (2003; 2005) has provided some additional insights into the system during the Stalin era, with a focus on principal–agent problems and incentives. The analysis undertaken before the end of the system drew upon the theoretical advances of Janos Kornai (1980; 1992) only to a limited extent (for example, Hanson and Pavitt, 1987), in particular Kornai's analysis of the role of soft budget constraints. The pioneering work of the late Yurii Yaremenko, perhaps the most innovative of all economists working in the USSR in its final 20 years, was also disregarded (in particular, Yaremenko, 1981). Yaremenko analysed the Soviet economy in terms of a hierarchical, multilevel system, with each level having access to resources, human and material, differentiated by quality. The qualitative heterogeneity of resources became firmly institutionalized. An individual enterprise, or entire branch of production, could rise up the hierarchy and secure access to higher-quality resources only by a policy decision of the political–economic authorities. Lower-level sectors habitually deprived of high-quality resources compensated for their lack by resort to larger quantities of lower-quality inputs. At the upper levels, occupied by defence industries and some priority civil sectors, higher-quality resources permitted the use and development of more advanced technologies, but innovations possible in these privileged conditions were unsuited to diffusion to lower levels of the economy lacking an appropriate resource environment for their successful application.

The collapse of Communism at the end of 1991 had a profound impact on Russian R&D and innovation. Positive factors included the end of ideological controls and censorship, as well as a new freedom for scientists to travel abroad. But for most of the 1990s negative factors predominated: funding collapsed, the number of researchers contracted sharply, many institutes of the Academy of Sciences and industry developed non-research activities in order to survive, and many talented younger scientists found work abroad (Glaziev and Schneider, 1993; Schneider, 1994; Gokhberg, Peck and Gács, 1997). A comparison of the situation in Russia in 2004 with that of 1990, after 15 years of transformation, shows the following principal features: R&D expenditure as a percentage of GDP was 1.35 compared with 2.03, although it reached a low point of 0.74 in 1992; in real terms R&D spending was 42 per cent of the 1990 level; and the number of researchers had fallen to 401,425 compared with 992,571, the decline moderating but still under way. In 2004 almost 50 per cent of all researchers were 50 years old or more, compared with 35 per cent in 1994 (CSRS and Roskomstat data). A landmark was the OECD (1994) review of Russian science, technology and innovation policies, which remains, with its background report, an important source on the system and problems of transition. This argued that post-Communist economic transformation in the prevailing circumstances of Russia would inevitably entail a substantial

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
contraction of the scale of the R&D effort and set out measures designed to convert the Soviet-era legacy into a coherent market-oriented National Innovation System.

Research undertaken by Russian and Western economists and science policy specialists reveals that, notwithstanding reform measures, the Russian R&D system as of 2006 still retains many Soviet characteristics (Dezhina and Saltykov, 2005; Radosevic, 2003). There is still organizational fragmentation, with the majority of R&D organizations being remote from the business sector; a large proportion of research organizations remain in state ownership; budget spending predominates, with only a modest contribution from the private sector; higher education plays a limited role in R&D, the Academy system still absorbs a large share of total expenditure; and the military share of R&D remains substantial, almost as large as in Soviet times. The inertia is such that it cannot be said that Russia possesses a National Innovation System, understood as a coherent set of interrelated institutions promoting innovation as a natural outcome of their day-to-day functioning. This situation has arisen in part because strong vested interests within the Academy of Sciences and industry have pursued their own survival strategies, resisting the government's reform initiatives.

Innovation activity in Russia's market economy remains relatively weak. Explanatory factors include weak competition, inadequate managerial skills, underdeveloped technological capabilities at the company level, weak demand for new products and processes, inadequate finance, with little long-term bank lending and a lack of venture capital, and modest foreign direct investment in manufacturing (OECD, 2006; World Bank, 2006). Russian government initiatives to improve the situation include the creation of special economic zones and the formation of venture capital funds. The domestic intellectual property rights regime has been improved but enforcement remains weak. For Russia, promoting a more effective R&D system and improved innovative performance is becoming a policy priority in the face of competition from other dynamic emerging economies, in particular China and India (Cooper, 2006).

Notwithstanding relative strength in human capital and the possession of a large research system in terms of personnel, the Soviet past still hampers Russia in the field of research, development and innovation, and for economists this represents an interesting case of the costs of institutional inertia.

See Also

- command economy
- emerging markets
- planning
- technical change

Bibliography

Amann, R., Cooper, J.M. and Davies, R.W. 1977. *The Technological Level of Soviet Industry*. New Haven and London: Yale University Press.

Amann, R. and Cooper, J.M. 1982. *Industrial Innovation in the Soviet Union*. New Haven, CT, and London: Yale University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Amann, R. and Cooper, J.M. 1986. *Technical Progress and Soviet Economic Development*. Oxford: Basil Blackwell.

Bailes, K.E. 1978. *Technology and Society under Lenin and Stalin*. Princeton: Princeton University Press.

Berliner, J.S. 1976. *The Innovation Decision in Soviet Industry*. Cambridge, MA: MIT Press.

Cooper, J. 2006. Of BRICs and brains: comparing Russia with China, India, and other populous emerging economies. *Eurasian Geography and Economics* 47, 255–84.

De Witt, N. 1961. *Education and Professional Employment in the USSR*. Washington, DC: National Science Foundation.

Dearden, J., Ickes, B.W. and Samuelson, L. 1990. To innovate or not to innovate: incentives and innovation in hierarchies. *American Economic Review* 80, 1105–24.

Dezhina, I.G. and Saltykov, B.G. 2005. The national innovation system in the making and the development of small business in Russia. *Studies on Russian Economic Development* 16, 184–90.

Fortescue, S. 1986. *The Communist Party and Soviet Science*. Basingstoke: Macmillan.

Glaziev, S. and Schneider, C.M. 1993. *Research and Development Management in the Transition to a Market Economy*. Laxenburg: IIASA.

Gokhberg, L., Peck, M.J. and Gács, J. 1997. *Russian Applied Research and Development: Its Problems and Promise*. Laxenburg: IIASA.

Hanson, P. 1981. *Trade and Technology in Soviet-Western Relations*. London and Basingstoke: Macmillan.

Hanson, P. and Pavitt, K. 1987. *The Comparative Economics of Research, Development and Innovation in East and West: A Survey*. Chur, Switzerland: Harwood Academic Publishers.

Harrison, M. 2003. The political economy of a soviet military R&D failure: steam power for aviation, 1932 to 1939. *Journal of Economic History* 63, 178–212.

Harrison, M. 2005. A Soviet quasi-market for inventions: jet propulsion, 1932–1946. *Research in Economic History* 23, 1–59.

Holliday, G.D. 1979. *Technology Transfer to the USSR 1928–1937 and 1966–1975: The Role of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Western Technology in Soviet Economic Development. Boulder, CO: Westview Press.

Kontorovich, V. 1994. The future of the Soviet science. *Research Policy* 23, 113–21.

Kornai, J. 1980. *The Economics of Shortage*. Amsterdam: North-Holland.

Kornai, J. 1992. *The Socialist System: The Political Economy of Communism*. Oxford: Clarendon Press.

Korol, A. 1965. *Soviet Research and Development: Its Organization, Personnel and Funds*. Cambridge, MA: MIT Press.

Kruse-Vaucienne, U.M. and Logsdon, J.M. 1979. *Science and Technology in the Soviet Union. A Profile*. Washington, DC: Graduate Program in Science, Technology and Public Policy, George Washington University.

Lewis, R. 1979. *Science and Industrialisation in the USSR*. London and Basingstoke: Macmillan.

Lubrano, L.I. and Solomon, S.G. 1980. *The Social Context of Soviet Science*. Boulder, CO: Westview Press.

OECD (Organisation for Economic Co-operation and Development). 1994. *Science, Technology and Innovation Policies. Russian Federation*. Vol. 1: *Evaluation Report*; vol. 2, *Background Report*. Paris: OECD.

OECD. 2006. *OECD Economic Surveys. Russian Federation*. Paris: OECD.

Radosevic, S. 2003. Patterns of preservation, restructuring and survival: science and technology policy in Russia in post-Soviet Era. *Research Policy* 32, 1105–24.

Schneider, C.M. 1994. *Research and Development Management: From the Soviet Union to Russia*. Laxenburg: Physica-Verlag; Heidelberg: IIASA.

World Bank, Russia. 2006. *Russian Economic Report No. 13*. Moscow.

Yaremenko, Y.V. 1981. *Strukturnye izmeneniya v sotsialisticheskoi ekonomike*. Moscow: Mysl'.

Zaleski, E., Kozlowski, J.P., Wienert, H., Davies, R.W., Berry, M.J. and Amann, R. 1969. *Science Policy in the USSR*. Paris: OECD.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Cooper, Julian. "Soviet and Russian research, development and innovation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000500>
doi:10.1057/9780230226203.1577

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Soviet economic reform

Vladimir Kontorovich
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Soviet economic reform’ refers to repeated failed attempts from the 1960s to the 1980s to introduce market-inspired institutions into a command economy. These reforms aimed to increase the discretion of enterprise managers and give them incentives to use local information in executing commands while also attending to customers’ demand. Reform arrangements, such as reduction in the number of commands and stress on profit and sales, clashed with the vital functions of the command mechanism, impaired performance, and were eventually reversed, only to be reintroduced later with similar consequences.

Keywords

central planning; command economy; market socialism; New Economic Policy (NEP) (USSR); profit as an indicator; profit sharing; Soviet economic reform; organizational change

Article

Lively organizational churning took place in the Soviet economy behind the façade of ideologically mandated institutional immobility. A command economy was introduced in 1918, gave way in 1921 to a market socialism-type system known as the New Economic Policy (NEP), and was re-established in the late 1920s and early 1930s. In the remaining 60 years of its existence it was subject to three types of organizational change: the reordering of the command hierarchy; the adjustment of the border between the command core of the economy and the peripheral sectors where the market was allowed to operate; and the introduction of market-inspired institutions into the command core itself, which we call ‘economic reform’ and which is the subject of this article.

The Soviet economy was organized like a pyramid, with party and government leadership at the top, enterprises at the bottom, and ministries for particular sectors in the middle. Its organizational chart was regularly redrawn as ministries and enterprises were merged or split, enterprises were transferred from one ministry to another, and functions were redistributed among staff and line departments. These

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
changes usually came with less fanfare than economic reforms and attracted relatively little interest from researchers, and their effects on economic performance have not been documented.
Administrative allocation of resources in the core of the Soviet economy (industry, transport and communications, construction, distribution of producer goods, foreign trade, part of agriculture) coexisted with market or quasi-market allocation at the interface between these sectors and households, as well as among the households. The boundary between market and command economy was repeatedly adjusted, as when rationing of consumer goods alternated with their sale at fixed prices in state stores. Institutional changes in this category had immediate perceptible effects on consumer well-being, as attested to by contemporaries.
Output commands and input quotas in physical terms never covered all the activities of enterprises. Because of the strains on the planners' information-processing capacity, some commands were formulated in aggregate terms, and some production activities were left out of the central plan. Actual physical details of aggregate commands, as well as decisions on the provision of some goods and services, were left for the producers and users to work out. Also, the command allocation of resources was shadowed by monetary flows. Enterprises paid each other for supplies at centrally fixed prices, kept accounts in monetary terms and received plan targets for cost, total value of output, and other monetary magnitudes.
In the classical Soviet system, enterprise managers' discretion in production decisions was limited to low priority goods. Money circulating in the command sector was meant to passively track planned direction of resources, playing merely the accounting role. Financial targets ranked low in the enterprise plans. The thrust of economic reforms was to make these subordinate features of the economy, which served as props for the command mechanism, at least as important as the latter. Enterprise discretion, the influence of users on suppliers, and reliance on monetary plan targets were all intended to increase.
Economic reforms attracted far greater attention than the other organizational changes because they arguably represented a retreat from the Stalin-era ideological orthodoxy and held out hope for systemic change, fuelling, for example, theorizing on the convergence of socialism and capitalism (Ellman, 1980, p. 200).

The prototypical reform

The first Soviet economic reform was announced in 1965. It replaced the main indicator of performance, namely, gross value of output, with sales, so as to bring the users' evaluation to bear on the producers' performance. To unshackle the initiative of the enterprises, the number of output targets for specific products in physical terms was reduced, and some other targets, including cost reduction, labour productivity, the number of employees, and the average wage were abolished. The volume of profit and its ratio to the enterprise's assets became binding targets. Enterprises were allowed to undertake small-scale investments on their own, and for this purpose to create a development fund financed out of profit, depreciation charges, and proceeds from selling unneeded equipment.
Pecuniary incentives built into the economic mechanism, working automatically, were to play an equal, if not greater, role than specific commands from superiors in guiding enterprises. A substantial bonus fund and a fund for residential construction and social development were created to replace the tiny fund that previously served these purposes. Inspired by the Western experience of profit sharing, reformers

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
made enterprise profit the source of financing for the funds. While previously the evaluation of performance and rewards and punishments were all tied to the degree of plan fulfillment, the size of the funds was tied to the actual growth of profit or sales and to the ratio of profit to assets through a formula. The coefficients of the formula were to be fixed for a number of years, rather than changed by the ministry at its discretion every plan period.

Another element of stable rules of the game for the enterprises was to be provided by firm five-year plan targets. Before, these were changed annually, if not more often, at the discretion of the supervisory body. Enterprises always paid for labour and intermediate inputs, but got their plant and equipment free as a result of centralized investment decisions. The reform introduced a capital charge, a fixed percentage of the value of fixed and working capital to be remitted to the state budget. The intentions to supplant input rationing by wholesale trade in intermediate products and to allow enterprises themselves to determine their wage bills were announced.

Reform was introduced gradually and by 1970 applied to almost all of industry (Schroeder, 1971, p. 38). However, in the process many reform measures were reversed or modified in ways subverting their initial intent. By the early 1970s, commands for production of specific goods had returned to the pre-reform level of detail, and targets for labour productivity and cost reduction were brought back (Iasin, 1989, p. 100). Funds for decentralized investment were first restricted and then used as just another source for financing centrally planned projects (Dyker, 1981, p. 139). Rules for incentive fund formation and for awarding bonuses were changed repeatedly, making the latter dependent on meeting a number of plan targets, rather than on the growth of profit (Ellman, 1984, pp. 87–9). Trade in intermediate goods and operational five-year plans for the enterprises never got off the ground or were realized only on a small scale.

Because of the eventual reversal of so many reform provisions, arguments about their effects are limited to the period of 1966–70, when they were still being implemented. According to the official Soviet data, industrial output continued to grow through this period at the same rate as in the previous five years. The recalculation by the US Central Intelligence Agency (CIA) shows a marginal slowdown in industrial growth (US Congress, 1982, pp. 191–2), and some of the estimates using the CIA data detect a pickup in total factor productivity (TFP) growth or a slowdown in its decline in 1966–70. To be able to attribute it to the effects of reform, one would need a thorough growth accounting exercise, which has not been undertaken for lack of data. Thus, changes in capacity utilization in industry help explain part of the variation in TFP in the 1970s. Capacity utilization in the late 1960s was increasing for reasons unrelated to the reform, though the extent of the increase, and hence its impact, is unknown (Kontorovich, 1990, pp. 44, 46).

There may have been one-shot gains as enterprises strove to qualify for the new incentive rules by assuming more ambitious targets, selling unneeded equipment, and drawing down inventories (Schroeder, 1971, p. 44). Over the longer period, qualitative evidence on the micro level shows the persistence of enterprise behaviour patterns which the reform intended to modify – hiding resources from the planners, meeting plan targets in ways that sacrifice the customers' needs and waste resources, and neglecting the tasks that the planners had not specifically ordered. The reform also had several unintended ill effects on performance. Greater discretion in choosing product mix and enhanced interest in value of output measured in administrative, cost-plus prices led enterprises to shift product mix in the direction of more intermediate input-intensive items, irrespective of customers' needs (Kushnirsky,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1982, pp. 16–17; Iasin, 1989, p. 99). The relationship between the dynamics of output in physical and in value terms shifted, and planners, who did not acknowledge the shift in their calculations, were led to make more wasteful allocation decisions (Khanin, 1991, pp. 41–9).

The treadmill of reform

The expression ‘treadmill of reform’, coined by Schroeder (1979), refers to the post-1965 Soviet attempts at reform followed by retreats. The treadmill appears even more daunting if one considers Soviet history and the collective experience of command economies. A prominent official in charge of heavy industry briefly attempted to do away with rationing of producer goods within his sector as early as 1932, with disastrous results (Davies, 1984, pp. 213–14).

In the late 1940s, as an element of Communist rule, a command economy was imposed on eight east European countries, some of which were the first to try economic reform after Stalin's death in 1953. Most of the elements of the Soviet 1965 reform had been formulated and tried by its east European neighbours in the previous decade, with similar results. In Hungary in 1954, the number of commands issued to the enterprises was reduced, and the intention was announced to promote decentralization through the use of incentives and delivery contracts. The number of commands soon crept back up. In 1956, half a dozen enterprises were experimenting with greater independence and use of profit as the success indicator (Kornai, 1959, pp. 218–21, 234). New incentives based on profit sharing were introduced in the late 1950s, and a capital charge in 1964. These measures proved, in the official estimation, to be ineffective (Bauer, 1987, p. 135; Berend, 1990, pp. 108–9).

Reforms in Poland starting in 1956 included the reduction in the number of commands, provisions for decentralized investment by the enterprises, and creation of a substantial bonus fund, both to be financed out of profit. These changes were rolled back in 1959–60 (Montias, 1962, pp. 294–307, 320–4). In Czechoslovakia in 1958, profit was made the main success indicator for the enterprises, with the latter also given the right to decide on some investment projects and finance them internally. These reforms were reversed in 1960–2 (Myant, 1989, pp. 82–4; Teichova, 1988, p. 148).

By the time the USSR embarked on its first economic reform in 1965, these and other east European countries were already on their second round, to be followed by further retreats, modifications, and new attempts too numerous to list here. Only Hungary escaped from the treadmill, having established a market socialist economy after 1968. The Czechoslovak reform of 1967 and the Polish reform of 1982 had similar aspirations, but the former was aborted after the Soviet invasion of 1968 and the latter was unsuccessful.

Decrees on ‘improvement of the economic mechanism’ adopted in 1979 aimed to correct some of the ill effects and reaffirm some of the neglected planks of the previous attempt. To curb the enterprises’ manipulation of product mix in favour of more material-intensive products, the main target was changed from sales to a variant of net product, and bonus payments were made conditional on fulfilling delivery obligations in terms of volume, assortment, quality and timeliness. The use of fixed rates, rather than ministry discretion, in planning wages, profit distribution, and other financial and physical targets for the enterprises was to be broadened. The intention to make five-year plans operational and to make ministries self-financed was reiterated. An internal investigation of the effects of this reform found that it had no positive impact on performance due to its piecemeal nature (Ellman and Kontorovich, 1998, p.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
109).

Though reforms are discussed as discrete events dated by the year of their announcement, each one was implemented in stages over a number of years, with amendments, modifications, and reversals to follow. This, combined with the redrawing of the organizational chart, such as the merger of enterprises into larger associations initiated in 1973, assured that Soviet industry was in the process of almost continuous administrative upheaval. The tempo of reorganization reached a fever pitch in the system's final years. In 1984 a ‘large scale experiment’, again attempting to cut the number of commands to the enterprises, and extending and strengthening some of the provisions of the 1979 decree, was introduced in five industrial ministries. By 1987 it covered the whole of industry, but in the summer of that year a ‘radical reform’ was announced, to be implemented in 1988.

Its most radical provision was the election of the enterprise managers by their employees. Plan targets for total value of output, profit, and most other aspects of enterprise performance became mere recommendations. Obligatory targets, rechristened as state orders, were kept for output of major products and commissioning of production capacity. They were supposed to cover only a small share of the enterprises’ output. Producers were free to dispose of the output over and above the state orders however they saw fit, though most prices remained fixed. All investment by enterprises was to be financed from their own revenue. Uniform stable long-term rates were to determine the distribution of profit among payments into the state budget and contributions to the investment, bonus, and social development funds, with enterprises retaining a greater share than before. The ministries were forbidden to issue any commands not on the government-approved list, a practice common during the previous reform attempts. Now the enterprises could go to court to overturn such commands and demand compensation for damages they caused.

The effects of the last reform on overall economic performance were drowned out by those of the collapsing political system and of severely misaligned fiscal, monetary, and investment policies. The reform increased financial resources at the enterprises’ disposal, and these were used to boost employee remuneration and decentralized investment, aggravating already significant shortages of consumer and investment goods. Faced with compulsory targets for only part of their output, enterprises reduced production of non-compulsory items and shifted their product mix towards higher-priced items. This disrupted the flow of inputs and made other enterprises unable to meet their obligatory output targets, contributing to an output contraction in many sectors in 1989 and economy-wide in 1990–91 (Krueger, 1993).

The elections of managers were repealed in 1990. The rates governing enterprise profit sharing were manipulated to rein in the growth of employee remuneration. In order to combat supply disruptions, centralized distribution of non-centrally planned products was authorized in mid-1988. While the official objective of curtailing the share of state orders in the output was regularly reiterated, by the end of 1991 only 14 percent of producer goods distribution officially went through non-administrative channels. Amended and revised, the 1987 reform defined the formal economic mechanism of the Soviet economy until its very end. Western experts were both urging Hungarian-type market socialism on the USSR and predicting its imminent adoption (Schroeder, 1979, p. 340). While proposals for such a transformation were floated in the country's final years, they were never realized.

Reasons for the failure of reforms

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The evidence of failure consists of the record of reform decrees modified and overturned; proclaimed intentions never realized in practice; the treadmill, with its repeated attempts to implement the same changes over and over again; and qualitative evidence that reform planks actually implemented failed to effect intended changes in enterprise behaviour, and often had negative unintended consequences.

Politics

Most Western research focused on the Soviet reform of 1965, and the prevailing interpretation attributed its failure to self-interested sabotage by ministry and other middle-level officials (Schroeder, 1979, p. 324). Resistance from the state and party bureaucracy was also seen as a formidable obstacle to the next serious reform. Such an explanation was consistent with the view of Soviet politics that became predominant in Western academia in the 1970s. The totalitarian model, with the top ruler being the sole unchallenged political actor, had been abandoned in favour of a pluralist model with politics propelled by interest-group rivalry.

The ‘bureaucratic resistance’ explanation of reform failure was incomplete at best, for the formal repeal of reform decrees could be done only by the top political authority. It implied the capitulation of the rulers in the face of interest-group pressure. But this would be highly unlikely given the record of Soviet economic history. In 1921, Lenin thought up and introduced his market-socialist NEP, to the dismay of much of the party, which nevertheless did not mount effective resistance. In 1957, Khrushchev abolished ministries, to the unhappiness of officialdom, but without resistance (Berliner, 1983, p. 383). Gorbachev dismantled or emasculated every economic bureaucracy in the space of a few years. Sectoral ministries were merged, had their staffs cut, and lost their power over the enterprises. The planners themselves carried out the abolition of the central planning when ordered to do so. The Central Committee and the regional party committees gave up their economic functions in an equally disciplined fashion. The military-industrial complex, the sector alleged to have the most political power, was ordered to convert to civilian production and complied (Ellman and Kontorovich, 1998).

A more plausible political explanation of the failure of 1965 reform is that the top political echelon was divided, with the General Secretary being less than enthusiastic about it from the beginning (Baibakov, 1998, p. 171). Fear of political destabilization in the wake of the Czechoslovak ferment of 1968 is thought to have further soured the rulers on the reform and brought it to an end.

Systemic incompatibility of reforms

Political explanations imply that the reforms as designed were effective, and their failure was due to faulty implementation. Yet the provisions of the Soviet 1965 reform that were faithfully implemented did not achieve their objectives and produced harmful side effects, some of which are described above. The uniform ineffectiveness and the unintended negative consequences of reforms implemented in different countries over a long period of time suggest flaws in their conceptual foundation, namely, the implanting of market-like institutions into a command economy.

Economic reforms were an exercise in imitation, their intuitive plausibility deriving from the appeal to the workings of competitive markets. Trying to make money, prices, and profit that already existed in the command economy perform some of their market economy functions, the reformers fell into a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
linguistic trap, committing the equivocation fallacy. Enterprise profit from centrally ordered transactions, calculated in terms of administratively set prices in an economy with passive money, had, despite the name, different properties from profit in a competitive market, and could not perform the latter's functions. There is no reason why the pursuit of this profit would lead enterprises to efficient decisions, or why the stress on the volume of sales would make them sensitive to the wishes of their customers.

While market-type institutions in the command economy context lack the efficiency properties of their market economy namesakes, they also interfere with the vital function of central planning. A modern industrial economy, with its narrow specialization, depends for survival on the regular flow of intermediate goods and services from producers to users. In the absence of market, the task of balancing, or assuring a tolerably smooth flow of inputs, falls on the economy's administrators. The complexity of the problem (the large number of products and producers, and continuous shocks to supply) overwhelms the planners' information-processing capabilities. If balancing is done poorly enough, the economy may come to a halt. Because of the difficulty and importance of balancing, it received the lion's share of attention of planners and ministry officials, and economic institutions were moulded so as to make their task easier (Grossman, 1963). Despite all that, the degree of balancing remained a grave problem in the functioning of the Soviet economy.

Most of the economic reform measures, if implemented, endangered the regular flow of supplies (Kontorovich, 1988, pp. 312–13). A reduction in the number of output commands and an increase in the level of their aggregation meant abandoning the balancing of some products, or making it less precise. Tying incentives to growth rather than to plan fulfilment deprived the officials of one of the tools of ensuring that certain quantities of goods were produced and shipped to users at designated intervals (Litwack, 1990). Abiding by stable long-term profit-sharing coefficients and other norms required the planners to give up their discretion to adaptively adjust the plan as its imbalances were being revealed (Litwack, 1991, p. 262). Developing long-range plans with the degree of detail that would make them operational defied the planners' capabilities, already strained by constructing a balanced annual plan. Enterprise investment funds, when effective, diverted inputs from the uses designated by the planners (Dyker, 1981, pp. 136–9; Ellman, 1984, pp. 95–6).

The recognition of systemic incompatibility sheds a different light on the role of bureaucratic resistance. Reforms forced the middle-level economic officials, who were responsible for balancing, to surrender many of the tools for accomplishing the task on which they were being evaluated. Faced with contradictory demands, they procrastinated, violated reform decrees, and exerted informal pressure on the enterprises in order to perform the function that was vital for the economy. To the degree that they succeeded, their actions protected the economy from the harm that systemically incompatible change would have inflicted on it (Kontorovich, 1988, pp. 313–14; Litwack, 1991, p. 264). The formal reversal of reform decrees occurred when the top officials recognized their ill effects.

Reasons for the treadmill

Repeated introduction and abandonment of a policy or an organizational change may result from efficient adaptation to changing circumstances, the shifting outcomes of interest group struggles, or changing ideas or ideology. However, all of these explanations are predicated on the effectiveness of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
recurring change (Siegmund, 1997, pp. 372–4). The treadmill of reform can be explained, in the spirit of North (2005), by considering the rulers' beliefs about the economy. The evidence, centring on the last and most aggressively reformist General Secretary, suggests that they were not aware that the reforms were unworkable. Memoirs of Gorbachev's aides and advisors, as well as his own pronouncements, show that he came to power with the intent of improving the workings of the system and with a belief in its ability to be reformed. This reflected a persistent strand in the Soviet thinking, which viewed the economy as highly malleable, as manifested in the overwhelmingly normative slant of Soviet economics and in much of economic policy.

Gorbachev (1995, p. 117) attributed the failure of the 1965 reform to bureaucratic resistance at the intermediate level of hierarchy, and blamed saboteurs for everything that went wrong with his own reforms. Even after having smashed the system and all its institutes, Gorbachev would insist that 'Sectoral bureaucrats could eat alive anyone, including the Chairman of the Council of Ministers and even the General Secretary' (Abalkin, Medvedev and Kotkovskii, 1995, p. 123). If sabotage was seen as the reason for the previous failure, then trying again, and harder, made sense. The experts on whom the rulers relied for the elaboration of the content of reforms shared this outlook. Some of the best Soviet economists argued that the 1965 reform had a strong positive effect on the economy. They viewed bureaucratic resistance as the main obstacle to a reform's success, as in Zaslavskaia's sensational 'Novosibirsk paper' from 1983 (Hanson, 1992, p. 59). The 1987 reform, incorporating the planks that had been tried and proven unsuccessful over the preceding 30 years, embodied the best the Soviet economists had to offer at the time.

The belief in reformability of the command economy was maintained against long odds. The Soviet rulers had rich experience in supervising large, complex, long-term projects for the creation of weapons systems. With the best domestic experts in charge, every available bit of foreign scientific information was utilized, prototypes were tested experimentally, and failed designs were set aside, while successful ones were developed, resulting in a formidable arsenal. The preparation of reforms outwardly followed the pattern that proved successful in the technical fields: a more or less public discussion among the economists, managers, and officials; blueprints prepared by committees including the representatives of these groups; and experimental trials preceding broad implementation.

Soviet reformers could also draw on foreign expertise, specifically, the east Europeans' evaluation of their own countries' experiences, as well as the work of Western Sovietologists. Thus, Kornai (1959, p. 225) argued that central planning is a coherent system, and piecemeal changes to it, such as less detailed plans or more discretion given to the enterprises, are bound to hurt performance. Granick (1959, p. 123–4) noted that even the minimal use of money and quasi-market institutions in the classical Soviet economic model conflicted with central planning. Grossman (1963, p. 119) demonstrated that the 'command principle' could not coexist with the market mechanism in a stable way, and correctly predicted the fate of the 1965 reform as it was being initiated (Grossman, 1966, p. 54). The Western analysis of reform failures was based on information from Soviet publications.

One would expect the Soviet economists to learn about the futility of reforms from the experiments, or from the experience of their implementation, or from east European and Western publications, and eventually to convey this conclusion to the rulers (though Western analyses from the 1970s and 1980s, stressing bureaucratic resistance, could be read as confirming Soviet domestic misconceptions). This is what happened to the east European economists, who started from the same level of understanding as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
their Soviet colleagues in the 1950s (Kornai, 1959, pp. xiii–xxv). The Hungarian 1968 market reform and the aborted Czechoslovak one in 1967 were preceded by public discussions among economists with a level of sophistication that was not achieved in the USSR until after 1987.

The different pace of learning appears to reflect tighter political constraints in the USSR than in some of its east European satellites. Unlike in the development of new weapons, the rulers had preferences not just over the results of economic reforms, but also over their methods. At various times, they considered particular reform measures to be unacceptable for reasons of political stability and ideological propriety. This not only forestalled public and professional discussions but also created disincentives for individual research. Trying to learn from the results of countries with laxer political constraints simply did not pay in terms of professional advancement. Ideas that were not allowed to be articulated just did not occur to people, at least not to those who might ever be asked for advice. Conversely, the reforms which the rulers did accept could not be undermined by negative experimental results. Economic experiments were ordered by the political authorities and conducted by lower-ranking officials. The latter understood the experiment to embody the official policy and did what they could to make it a success. Choosing the better-performing enterprises to participate in the experiment or affording special treatment to the participants usually assured the result.

In the late 1980s, as political and ideological constraints on acceptable economic advice were being removed, economists' learning greatly accelerated, and proposals for a transition to a market economy were publicly announced. The demise of the command economy, weakened as it was by the final round of economic reforms, came not through the implementation of these radical proposals but as a result of the disintegration of the underlying political system.

See Also

- command economy
- socialism
- Soviet growth record
- Soviet Union, economics in

Bibliography

Abalkin, L.I., Medvedev, V.A. and Kotkovskii, L. Ia., eds. 1995. *Istoricheskie sud'by radikal'noi reformy*. Prague: Laguna.

Baibakov, N.K. 1998. *Ot Stalina do Ieltsina*. Moscow: GazOilPress.

Bauer, T. 1987. Reforming or perfectioning the economic mechanism. *European Economic Review* 31, 132–8.

Berend, I.T. 1990. *The Hungarian Economic Reforms 1953–1988*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Berliner, J.S. 1983. Planning and management. In *The Soviet Economy: Toward the Year 2000*, ed. A. Bergson and H.S. Levine. London: George Allen and Unwin.

Davies, R.W. 1984. The socialist market: a debate in Soviet industry, 1932–33. *Slavic Review* 43, 201–23.

Dyker, D.A. 1981. Decentralization and the command principle – some lessons from Soviet experience. *Journal of Comparative Economics* 5, 121–48.

Ellman, M. 1980. Against convergence. *Cambridge Journal of Economics* 4, 199–210.

Ellman, M. 1984. *Collectivization, Convergence and Capitalism: Political Economy in a Divided World*. London: Academic Press.

Ellman, M. and Kontorovich, V., eds. 1998. *The Destruction of the Soviet Economic System: An Insiders' History*. Armonk, NY: M. E. Sharpe.

Gorbachev, M.S. 1995. *Zhizn' i reformy*. Book 1. Moscow: Novosti.

Granick, D. 1959. An organizational model of Soviet industrial planning. *Journal of Political Economy* 67, 109–30.

Grossman, G. 1963. Notes for a theory of the command economy. *Soviet Studies* 15, 101–23.

Grossman, G. 1966. Economic reforms: a balance sheet. *Problems of Communism* 15(6), 43–55.

Hanson, P. 1992. *From Stagnation to Catastroika. Commentaries on the Soviet Economy, 1983–1991*. The Washington Papers No. 155. Washington, DC: CSIS.

Iasin, Ie.G. 1989. *Khoziaistvennaia sistema i radikal'naia reforma*. Moscow: Ekonomika.

Khanin, G.I. 1991. *Dinamika ekonomicheskogo razvitiia SSSR*. Novosibirsk: Nauka.

Kontorovich, V. 1988. Lessons of the 1965 Soviet Economic Reform. *Soviet Studies* 40, 308–16.

Kontorovich, V. 1990. Utilization of fixed capital and Soviet industrial growth. *Economics of Planning* 23, 37–50.

Kornai, J. 1959. *Overcentralization in Economic Administration*. New York: Oxford University Press, 1994.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Krueger, G. 1993. *Goszakazy* and the Soviet Economic Collapse. *Comparative Economic Studies* 15, 1–18.

Kushnirsky, F.I. 1982. *Soviet Economic Planning, 1965–1980*. Boulder, CO: Westview Press.

Litwack, J.M. 1990. Ratcheting and economic reform in the USSR. *Journal of Comparative Economics* 14, 254–68.

Litwack, J.M. 1991. Discretionary behaviour and soviet economic reform. *Soviet Studies* 43, 255–79.

Montias, J.M. 1962. *Central Planning in Poland*. New Haven, CT: Yale University Press.

Myant, M. 1988. *The Czechoslovak Economy 1948–1988*. Cambridge: Cambridge University Press.

North, D.C. 2005. *Understanding the Process of Economic Change*. Princeton, NJ: Princeton University Press.

Schroeder, G.E. 1971. Soviet economic reform at an impasse. *Problems of Communism* 20(4), 36–46.

Schroeder, G.E. 1979. The Soviet economy on a treadmill of ‘reforms’. In US Congress, Joint Economic Committee, *Soviet Economy in a Time of Change*, vol. 1. Washington, DC: Government Printing Office.

Siegmund, U. 1997. Are there nationalization–privatization cycles? *Economic Systems* 21, 370–4.

Teichova, A. 1988. *The Czechoslovak Economy 1918–1980*. London: Routledge.

US Congress, Joint Economic Committee. 1982. *USSR: Measures of Economic Growth and Development, 1950–1980*. Washington, DC: Government Printing Office, 1982.

How to cite this article

Kontorovich, Vladimir. "Soviet economic reform." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000507> doi:10.1057/9780230226203.1578

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Soviet growth record

Gur Ofer

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Soviet Union initiated and implemented a heroic social experiment that included a rapid and fairly successful, albeit rather distorted, process of modernization and growth. It culminated more than 70 years later in a dead end that required a difficult and costly ‘transition’ in order to join the main road to modern economic growth and a market system. After presenting the record of Soviet economic growth and structural change, this article describes and analyses its particular strategy of extensive growth, its strengths and weaknesses, its relation to the Communist political regime, and the reasons for its non-sustainability and, hence, demise.

Keywords

autarky; autarky: internal and external; balanced growth; capital deepening; capital-led growth; central planning; CES production function; Cobb–Douglas function; command economy; defence; defence expenditure; demographic transition; diffusion of technology; elasticity of substitution; endogenous growth; extensive and intensive growth; fertility; haste; human capital; industrialization; formal and informal institutions; innovation; Kosygin reforms (USSR); marginal productivity of capital; market socialism; military sector; military research and development; patents; price index relativity; Second World War; soft budget constraint; Soviet growth record; switching development strategies; structural change; take-off; technical change; time preference; total factor productivity; transition

Article

The end of 1991 saw the collapse of the Communist regime, which was created 74 years earlier following the October Revolution of 1917, and with it the disintegration of the Soviet Union into 14 newly created independent states. In this way one of the most heroic social experiments of modern times came to a close. That experiment failed to provide a sustainable alternative social order to that offered by the capitalist model of different mixes of market mechanisms and government interventions, and of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
different variants of democratic and free political regimes. The Communist alternative, as presented by the Soviet model, was composed, on the economic side, of central planning and public ownership of most means of production, and of an authoritarian regime on the political and social sides; and both built on the foundation of the political economy teaching of Karl Marx as modified by Vladimir Lenin and his fellow-Bolsheviks. This Communist alternative claimed to be more equitable and just, and at the same time also more efficient, than the capitalist alternative. Specifically, it claimed to achieve faster economic growth, and thereby the potential to catch up with and overtake the capitalist market economies.

The end of the Communist system, indeed its demise from within, provides proof that the system as a whole was not sustainable, and very probably that economic factors had a role to play. This, however, was not the accepted view by everybody during the Soviet period. Throughout that era, more so during its earlier years, the Soviet system presented in the eyes of many an economic challenge (in addition to a military one) to the developed West, the threat of ‘taking over and surpassing’, and an alternative, possibly superior growth strategy to developing countries. This economic challenge, including its military implications, aroused great interest and thereby generated academic and intelligence input into the study of the Communist economic model and development strategy and the resulting outcomes in terms of economic growth, efficiency and equity. This input became all the more difficult, for both Soviet and Western scholars, because of the veil of secrecy that was imposed by the Soviet authorities on all sorts of information and statistics, and the distinct, though in many cases obscure methodology underlying the economic information and data that were published. Often the published data were biased in order to present a rosier picture than the Communist reality.

A survey of Western attempts to estimate the Soviet growth and efficiency record and to evaluate and explain it was prepared by this author during the last years of the Soviet Union (Ofer, 1987). It included, among others, mainstream estimates as well as some of the main debates and disagreements over these estimates and the methodologies used to compile them. Even though some of the figures presented turned out to be somewhat biased in favour of the Soviet record (in the light of data and analysis that became available in later years), the main observations and conclusions regarding the Soviet experience seem to have survived the test of time: among other things, they showed that, following a period of relatively rapid growth and successful industrialization and modernization, the pace of growth slowed down over time up to a virtual stop; that, overall, Soviet growth lagged behind that of many Western and other countries; that Soviet growth was achieved with a greater use of labour and investment resources per unit of output, and a lower contribution to improved productivity, that is, at lower efficiency levels and also with much greater human sacrifice, economic and otherwise; and that the Second World War, the Cold War and imperial ambitions came at a price in terms of lower growth and standards of living. Furthermore, the record showed that Soviet growth was achieved also at high costs of mortgaging future resources and capabilities, thereby making future growth and the transition, at least for a while, even more difficult. There was some degree of greater economic equity among the majority of the people, but it was bought at the cost of lower standard of living and the denial of freedom.

In this article I re-examine the Soviet growth record in the light of new information that became available since the collapse of the Soviet Union, including insights gained from the collapse itself, which was not predicted by most commentators, and from experience gained to date during the transition period. As will be seen, a fair amount of new research and thinking has emerged since the collapse. In most cases the data and analysis in Ofer (1987) and, even more important, the long list of references

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
there will serve as a kind of benchmark for further discussion and, to save space, discussed only briefly here.

The growth record revisited: the basic figures

The Communist revolution took place in 1917, but only in mid-1928 was the Communist economic system and growth strategy put into place. Collectivization of agriculture was combined with a rapid industrialization drive, all organized under a command system of nationalization and central planning. The first decade following the revolution was devoted to taking control of the economy and government, reconstruction following the devastation caused by the First World War, revolution and civil war, and to experimentation and debates over the appropriate economic strategy, integrated with the struggle for political control. It is generally established that by 1928 the GDP in the Soviet Union regained approximately the level of 1913, while GDP per-capita (GDPPC) lagged by at least ten per cent behind. This followed a sharp decline due to war and revolution, reaching a low by 1920 and 1921. Most estimates for the Soviet period start in 1928. The growth record therefore captures performance during the era of the ‘Soviet model’.

Economic growth and productivity can be measured in a number of ways. For growth we will concentrate on the levels and rates of the growth of GDP per capita. Total GDP and various segments thereof will be used when appropriate. The levels and growth rates of productivity will be represented mostly by ‘total factor productivity’ (TFP), that is, output per unit of combined inputs, usually labour and capital. In what follows we use as a base mostly data assembled and estimated by Angus Maddison (1995; 2001) derived from what he considered (and we concur) the best and most reliable sources. For the Soviet Union Maddison relied mostly on the work of Abram Bergson (1961), who put together the most comprehensive body of Soviet national accounting using the same methodology as the CIA (Maddison, 1995, pp. 141–2). In both cases there were more recent adjustments, as explained below. Maddison's monumental work *Monitoring the World* (1995) includes the additional advantage that it provides comparisons of the growth record of many countries, through a uniform methodology and units of accounts. Estimates of variables not included in Maddison's work are presented following a check of their basic consistency with the lead estimates. Remaining disagreements and the underlying alternative estimates will be presented and discussed.

The rate of growth of GDP per capita over the entire 1928–90 period was 2.6 per cent per annum, a fivefold increase. This growth record puts the Soviet Union in the group of fast developers and separates it from a large number of countries, mostly in the Third World, that failed to take off or that started much later. Over that period the Soviet Union transformed itself from a predominantly agricultural economy, but with a considerable industrial base, into a modern industrial and urban economy. Soviet growth also achieved a modest degree of catching up with the US economy and other developed countries. The relative level of GDPPC of the Soviet Union grew from nearly 21 per cent of the US level in 1928 to almost 30 per cent in 1990. Somewhat higher ratios for 1990 are estimated by others. However, similar or better achievements are shown for many advanced countries as well as for a number of developing countries, especially in east Asia (Maddison 1995, Tables 1–4, p. 25). At the same time many countries, mostly of the Third World, remained far behind.

The average growth rate shown above is made up of significantly higher rates up to 1970, of between

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
three and four per cent a year (if the First World War is disregarded) followed by steeply declining growth trend thereafter, down to half of one per cent per year during 1985–90. Some decline in growth rates and sustained growth, following a faster take-off, is considered normal. The Soviet record raises the question of whether growth along the Soviet model was sustainable in the future. Likewise, it is observed that, after reaching the highest relative level of GDPPC of 37.5 per cent of that of the United States in 1970, up from 21 per cent in 1928, the Soviet economy retreated below that level, even below most other developed countries. It is this break in the trend, from catching up to retreat, which started at an early stage that sounds the alarm bell over the merit of the Soviet model and its long-term sustainability.

The trends in the rates of growth of total GDP are similar, with relatively high rates up to around 1970 and sharply declining ones thereafter. The difference between the trends in GDP and GDPPC represents a trend of declining population growth, indeed a very sharp ‘demographic transition’, and declining fertility on top of the heavy losses associated with the collectivization drive and famine, other atrocities by Stalin, and then with the Second World War (see more on this below).

The figures on growth mentioned above are somewhat lower than those calculated by the CIA and accepted by most scholars during the period before the fall of the Soviet regime. The downward adjustments were made, including by the CIA itself (Noren and Kurtzweg, 1993), on the basis of more information that came out from the Soviet Union during the 1980s, of criticism of the then existing estimates by émigré and Soviet scholars, most notably Igor Birman (1989), and on the basis of parallel efforts by others, most notably the International and the European Comparative Projects (Summers and Heston, 1991; Heston, Summers and Aten, 2002; United Nations, 1999 and earlier years) projects of estimating and comparing the national accounts of many countries and their structure, based on the purchasing power parity methodology. The CIA revisions were mostly a result of downgrading the quality of Soviet goods and services and of the productivity of providers of public services below previous estimates, and from higher estimates of hidden inflation rates (especially of military production). Based on these downward revisions, affecting mostly estimates since about 1970 during the last two decades, and the adding of 1985–90 to the series, the 1928–90 annual rates of growth were reduced by about half a percentage point, from 4.2 to 3.7 per cent for total GDP and from 3.0 to 2.6 per cent for GDPPC. The adjusted estimates also revealed an even steeper decline following the break near 1970. The new estimates also reduced the relative standing of the Soviet economy by 1990 and before: the Soviet Union/United States ratio for GDPPC for 1990, estimated previously by the CIA at 0.43 per cent, was adjusted down 0.30–0.34 (the higher figure is from United Nations, 1999, ECP96). While these are quite significant adjustments, they do not change the basic picture of the overall growth record or the declining trend. As we shall see, the factors responsible for these trends also remain pretty much the same.

During the years just before and immediately following the fall, an avalanche of estimates reduced the relative standing of the Soviet economy and its growth record much below the adjustments cited above. These were supported by the economic crisis in the Soviet Union, Russia and the former Soviet states, before and then following the collapse, and included sharp declines in output during the early transition period. Some claimed that the observed fall in measured output after the collapse reflected, among other things, an artificial overestimation of the level of output before the collapse (see, for example, Åslund, 2002, pp. 21–39). Some of the arguments supporting this claim were incorporated into the adjusted

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
estimates mentioned above. However, in some cases such observations were based on a failure to distinguish between the level of output and consumer welfare. First, most of the fall in estimated GDP was due to sharply reduced investment and military expenditure and only a smaller decline in consumption. Second, the price liberalization and the creation of markets reduced shortages and queues, and in this way improved consumer welfare. Finally, the production and consumption preferences of the Russian society shifted from that of the Communist leadership to another, determined by the people, who valued consumption much more and investment and military spending much less. In the eyes of the people, the former GDP was evaluated so much lower. It took a few years until more balanced views on the size of the Soviet GDP returned, and estimates of overall Soviet growth, including the downward adjustments, were generally accepted (Hanson, 2003, p. 249; see also Åslund, 2002, p. 34).

This is, however, not yet the case regarding Soviet growth during 1928–37 (or 1940). Here the gaps among scholars seem to have widened rather than converged since the fall of the Soviet Union. Two Russian economists, Vasili Selyunin and Grigori Khanin, produced an estimate of Soviet growth, with much lower estimates for 1928–40 (Harrison, 1993). These estimates reinforced similar estimates for the same period by Naum Jasny, estimates that were the focus of an older dispute, unearthed in a recent article by Howard Wilhelm (2003). *Per contra*, a recent book by Robert Allen (2003) challenges the entire established view of Soviet growth, based mostly on an upward revision of rates of growth, of GDP and of consumption, also during 1928–37 (Allen, 2003, and Wilhelm, 2003, Appendix A, pp. 212–22). Maddison's figures for that period are taken from Moorstein and Powell (1966, p. 361) which are somewhat higher than the final estimate by Bergson (1961, p. 48).

The estimates by Bergson and by Moorsteen and Powell are calculated with 1937 prices as weights. Given the sharp structural changes, and that of the accompanying relative prices, the choice of price weights, according to the notorious theory on ‘price index relativity’, makes a great difference. In particular, the use of 1928 prices as weights produces significantly faster growth during 1928–37. Allen claims, correctly, that in such cases a geometric average of the two measures or a similar ‘compromise’ is more appropriate. This argument was well known to Bergson and all other scholars working on Soviet growth. Bergson justified his decision to stick to 1937 prices by claiming that the 1928 prices were not free-market prices and extremely distorted. This observation is supported also by Hunter and Szyrmer (1992, ch. 3 and pp. 305–11), who substituted in their calculations for 1928 a set of ‘equilibrium calculated prices’. When these prices were used as weights the rate of growth of GDPPC during 1928–37 turned out to be similar to that based on 1937 prices (see more on Allen's position below). Let us point out in conclusion that any upward adjustment for 1928–40 will have to come at the expense of future growth, of both GDPPC and consumption, thus making the declining trend even steeper. Indeed, despite the above, Allen seems to agree with the overall estimates for GDPPC presented by Maddison (Allen, 2003, pp. 220–2).

A few words have to be added on the changing industrial structure of the Soviet Union. The general patterns of change were similar to those associated with modern economic growth: there was a decline in the share of agriculture, a marked increase in the share of manufacturing and related industries, construction and transportation, combined together as the M sector, and some rise in the share of services – all in terms of shares of the labour force and of GDP. The intensity of the changes was marked during 1928–37 and much less during the period since 1970. The increase in services was slower than normal in market economies, explained by the absence of markets, the view that services were ‘non-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
productive', and the policy of requiring households to supply many services during their non-working hours. The decline in the shares of agriculture was somewhat slower than 'normal', especially of the labour share, reflecting low labour and general productivity, due to the inefficiency of collectivization and low investment. Regarding manufacturing, and the M sector, there was the bias in favour of producer goods at the expense of consumer-good industries. This bias was reflected in the emphasis on investment and defence at the expense of consumption. The Soviet economy, as well as those of other Communist countries, was characterized as 'over-industrialized' in comparison with market economies at similar levels of development.

The growth strategy: extensive and capital-led growth

Growth strategy

What was the growth strategy of the Soviet Union, and how successful was its implementation? What explains the initial take-off and acceleration, and the eventual decline? How do the Soviet growth mechanisms and patterns compare with those in other countries in nature and in effectiveness? Following intensive and prolonged debates during the 1920s, the Soviet leadership opted for a strategy of rapid growth and industrialization based on high rates of investment, mostly in heavy industry and the producer-goods industries – sector 'A', according to the Marxian jargon. It rejected the other option of industrialization along with more 'balanced growth' between the various branches of the economy (Erlich, 1960; Hunter and Szyrmer, 1992). The chosen strategy followed the Marxian doctrine of 'expanded reproduction', as articulated at the time by a growth model developed by G.A. Fel'dman (Allen, 2003, pp. 53–60). It demonstrated that an initial high rate of investment can provide for rapid growth and, after some delay, also an overall higher level of consumption. On the ground this strategy was translated into maximization of capital investment through forced savings and maximum mobilization of the labour force, including of women, while consumption levels were kept at the minimum sustainable level. Both became possible only through the power of the authoritarian regime and the 'command economy', not to mention the merciless rule of and atrocities committed by Stalin. The rates of gross investment climbed over time to over 30 per cent of GDP. Such rates allowed the capital stock to grow much faster than in most other countries, and contributed the lion's share of total growth. The high share of investment left a mere 55 per cent of total output to household consumption, a share that declined further towards the end of the regime to below 50 per cent of GDP, 15–20 per cent below common levels in market economies. The most extreme version of this growth model was manifested during 1928–40, when the industrialization effort drew millions of people from agriculture and rural areas to monumental construction sites in old and newly created cities, and when consumption levels failed to increase, according to the accepted view, or even declined or increased modestly according to dissenters on both sides. There is no doubt, however, that the accompanying human losses and suffering during that period kept personal welfare at extremely low levels. The Second World War added to the losses and suffering, and only during the early 1950s did private consumption levels start to rise to a moderate level. By 1990, consumption per capita in the Soviet Union stood at just a quarter of that in the USA as compared with about a third for GDPPC (investment per capita stood at 55 per cent of that in the USA). While over the Soviet period there was some catching up towards the US level in terms of GDPPC, there was probably very little in terms of private consumption. It can be concluded

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that, while the first part of the Soviet growth strategy, emphasizing investment, was fully implemented, the subsequent rise in consumption was very modest, and failed to arrive at the ‘promised land’. A major reason for this, as is demonstrated below, is the failure of the Soviet system to generate enough productivity growth through technological and other efficiency improvements.

An increasing share of the remaining GDP was devoted to defence and to ‘public consumption’ on education and health care, among other areas. The increase in defence spending, eventually up to about 15 per cent of GDP, is explained by the perceived outside threats, first from Germany and then from the West, and by the ambition of the leadership to build an empire and become a world power (see more below). Investments in human capital through education and health services were extremely important for the industrialization drive. However, they also improved the welfare of the population, typically to a higher degree than in other countries with similar levels of economic development. Especially noteworthy here is the encouragement of women to enrol in professional and higher education.

Extensive growth

Economic growth in general is generated mainly by two sources, inputs, or factors of production, namely, capital and labour, and by increasing productivity in the use of inputs, getting more output per unit of input. Among the factors increasing productivity are technological innovations, improvements in the organization of production, increasing returns to scale, greater effort, and structural shifts from less to more productive activities (such as from agriculture to manufacturing). The common way to account for the contribution of each factor to total growth is through a ‘Cobb–Douglas’ production function, which assumes constant returns to scale. It is estimated as output produced by capital and labour, and at a per capita level. The residual growth left after the contributions of all inputs are accounted for is ‘total factor productivity’ (TFP). Inputs are combined according to their relative weights in the production process.

The Soviet growth model and record described above are commonly characterized as involving ‘extensive’ growth’ or as ‘inputs-led’ or even ‘capital-led growth, driven mostly by increasing contributions of the main inputs and much less by greater productivity. The alternative, ‘modern’ economic growth as defined by among others Simon Kuznets (1966) is characterized in most developed countries as ‘intensive’, whereby productivity growth accounts for a larger share of total growth and the bulk of per capita growth.

But although the mobilization of inputs was fundamental to the Soviet growth strategy, there was no strategic intention to neglect ‘productivity growth’; on the contrary, much was invested in R&D, in quality manpower, material resources and institutional support. Technological innovation and productivity growth were targets of multiple incentives. Yet the outcomes were rather disappointing. While during 1928–40 TFP contributed 1.7 per cent a year to GDP growth, nearly half the per capita growth, this declined to about 0.5 per cent during the 1950s and 1960s and then moved to negative territory for the rest of the period, thereby contributing significantly to the decline in the rates of growth. The higher contribution of TFP during the early period is explained by the dissemination of readily available technologies, and by major structural changes and initial industrialization, advantages that were exhausted later.

The poor outcomes in productivity during the entire post-Second World War period are explained by attributes of the system that limit its capability to generate new technologies and other improvements in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
productivity, by the priority allotted to the military sector, and by systemic barriers to the diffusion of innovations and general TFP measures across the production sector. Other TFP failures resulted from the interaction between the input-driven growth and the systemic nature of central planning. A study by Joseph Berliner (1976) contains a detailed analysis of the technological weaknesses of the Soviet system. There are two major clusters of factors, strongly intrinsic to central planning, that are responsible for the low rate of TFP: first, the rigidity of the system and the high cost of flexibility and change; and second, the strong reliance on quantity and quantitative rather than qualitative achievements and incentives.

The hierarchical, top-down and command character of central planning discourages initiatives from below and suppresses competition: both are believed to be essential generators of innovation. Central planning is also a rigid system that minimizes flexibility in production and supply networks. The high costs of reliable information flows across long hierarchical command ladders provide a great advantage to routine, and to inertia in regard to change, to ‘planning from the achieved level’ as against adopting flexible and innovating plans, and to stable supply networks as against shifting ones – which are needed when new materials or new markets are to be preferred. All are barriers to innovation and dissemination. Control under central planning is so much simpler – indeed only possible – on the basis of quantitative performance measures of output, inputs, supply flows, than on the basis of qualitative performance measures, better or new products, new production materials and processes, and greater labour efficiency. For these reasons, incentives under central planning reward first of all quantitative performance. Production plans are tight, in order to eliminate hidden production reserves and to encourage growth, and incentives to managers are mostly based on plan fulfilment. This also helps to assure smoother supply flows among producers. It comes at the expense of quality, of productivity and of setting aside time and inputs for improvements and innovation. Tight plans create shortages and a seller's market that severely limits the power of buyers to control for quality. There are incentives to innovation, but they are mostly dominated by those for quantitative plan fulfilment. The emphasis on quantity also explains the almost complete absence of exit of inefficient enterprises with obsolete equipment. Their contribution to the fulfilment of branch plans cannot be dispensed with. They are kept alive through the mechanism, termed by Janos Kornai, of the ‘soft budget constraint’, the pumping of additional resources into enterprises that create bottlenecks in the flow of output. On balance, enterprise managers refrain from the introduction of improvements and innovations, whose dissemination across the economy is thereby severely restricted. The extensive nature of the growth strategy as described above also contributed to the bias in favour of quantity over qualities and of inertia over change.

The process of innovation and technological change suffered also from the autarkic nature of the economy, which limited the free flow of advanced technology and production processes from the West. Huge effort and resources were invested in ‘reverse engineering’ of Western technology, and in other cases in ‘reinventing the wheel’, both costly and wasteful alternatives. Civilian innovation also suffered from the high priority accorded to military R&D and to military production in general. Military R&D took the lion's share, by some estimates up to two-thirds or more, of the entire Soviet R&D effort, and on top of this only very limited spillover of military innovations was allowed to benefit civilian production. Furthermore, in order to assure the prompt fulfilment of military production plans and to secure its proper quality, an entire body of priorities was granted to the military at the expense of the civilian sector: in quality of human resources and material inputs, in quality control, in price discrimination and in supply. The superior military and space technological achievements of the Soviet

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Union became possible partly by the imposition of a heavy burden on the civilian sector, and an even heavier burden on civilian R&D and on Soviet TFP.

Part of the relatively high TFP during the 1950s and the decline thereafter can be explained by the process of reconstruction following the devastation caused by the Second World War. However, another part of the TFP decline over time is to be blamed on some of the factors listed above as negatively affecting TFP, such as the increasing difficulty over time in incorporating new and more complex technologies. Another important factor is the increasing complexity of the economy, imposing a heavier burden on the planning process, on the flows of information and on the supply networks. It was expected that at some point the computer revolution may have eased those difficulties, but this did not really happen, partly due to the slow development of information technology (IT) in the Soviet Union, but probably more to the growing complexity of the economy such that even much more advanced IT would not be able to simulate

In addition to external autarky the Soviet economy also suffered from what may be called ‘internal autarky’, the segmentation of the economy along vertical lines corresponding to the planning hierarchy, with very limited horizontal networking. This structure was amenable to the development and dissemination of sector-specific technologies, but was an obstacle to economy-wide ones. This is one explanation of the slow development of IT of all types, including in computing, and other advanced general-purpose technologies. The internal segmentation of the economy also delayed the dissemination of new technologies across branch lines. This is why central planning failed to seize an advantage over the market economy, where the dissemination of innovations is delayed by patent protection. Finally, declining TFP may also have been caused by the trend of deteriorating discipline and work motivation and enforcement during the post-Stalinist era.

After the death of Stalin the Soviet Union became engaged in endless attempts at reform, directed mostly towards partial decentralization, in order to increase flexibility and reduce the cost of change. Most such attempts failed and were abandoned, and gave way to recentralization. The more radical, though still partial, reform measures initiated by President Gorbachev during 1985–91 increased the level of disorganization of the economy and contributed to the disintegration of the centrally planned system.

The above discussion of the low and declining level of TFP fits well the Cobb–Douglas production setting, in the sense that the listed factors affected the entire production process in a neutral way. There were, however, two attempts to link the low productivity performance of the economy to the high rates of capital investment and the nature of this investment. The more notable study, by Martin Weitzman (1970), estimated a constant elasticity of substitution (CES) production function for the Soviet economy and found that, while the level of productivity growth was quite respectable and constant over time, the elasticity of substitution (ES) between capital and labour was very low. Later estimates repeated this finding for the entire post-war period (Easterly and Fischer, 1995). Low and declining ES is consistent with low and declining capital productivity. Low ES focuses on the increasing technological and organizational difficulties in keeping up with the fast trend of capital deepening that resulted from the growth strategy, especially when the rate of growth of the labour force declined over time (see below). This resulted in a fast decline in the marginal productivity of capital. While the substitution of capital for labour is normal along the path of modernization and industrialization, the level and pace of the needed substitution, dictated by the high rate of investment, were too demanding for Soviet planners and innovators. Soviet sources reported thousands of fully equipped work stations with no workers to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
operate them. Low ES is therefore one important argument supporting the non-sustainability in the long run of capital-led growth. But the sources of low ES seem to be very similar to the explanations for low TFP in general listed above.

An interesting interpretation of the same phenomenon is offered by Vladimir Popov (2007). His argument is that it is much easier for the central planners to build and install new enterprises and plants than to re-innovate and replace production lines in existing ones, as was shown above. Accordingly, high rates of growth of output were achieved during the 1950s and early 1960s, when a new generation of enterprises was built. These rates declined later when the production lines aged and the authorities encountered organizational and technical difficulties in replacing them. The outcome was continued use of old equipment for long periods of time, well beyond its normal age in market economies, with very high maintenance and labour costs and long stoppages. The reluctance to retire obsolete enterprises resulted also from the pressures to meet tight production plans, as mentioned above (see Ickes and Ryterman, 1997). In later years, continued investments in new plants, albeit at lower rates, faced extreme labour shortages. This stage arrived during the late 1970s and 1980s, and contributed further to declining marginal utility of capital and to low ES.

The low and declining rates of TFP growth, or for that matter of any ‘residual’ beyond the contribution of capital and labour, made the Soviet growth pattern more and more ‘extensive’, that is, dependent almost exclusively on increasing amounts of labour and capital. At the same time, there are limits to the potential growth of both inputs, and therefore ‘extensive’ growth is bound to decline and eventually stagnate (Bergson, 1973). In order to keep the rate of growth of capital at high levels, under conditions of no or little growth of TFP, and with normally lower growth of labour inputs (see below), the share of investment in GDP must increase constantly, first reducing the share of consumption, and later also its absolute level. In addition to the negative effects on incentives and morale, this eventually becomes also politically unsustainable, even under an authoritarian regime. This danger forced the planners during the 1980s to reduce the rates of growth of material capital, from the previous eight to nine per cent per year to six per cent and below.

The increase in labour inputs was also coming to a virtual halt because of two reinforcing trends: first, the decline in the rate of growth of the population, and second, reaching practically the maximum rate of labour force participation. The first trend is part and parcel of the modernization process, but in the Soviet Union it accelerated beyond the ‘normal’ pace due to heavy material pressures on the population – the depressed standard of living, the meagre provision of housing and of household durables and services, and the pressure on women to participate fully in the labour force. On top of this, most women were forced to spend much time during after-work hours on household chores. As a result the rate of population growth declined from 1.8 per cent a year during the 1950s down to just 0.9 per cent in the 1980s, when much of this remaining growth was concentrated in the Muslim areas. As we shall see, the Soviet Union underwent a too rapid ‘demographic transformation’ that contributed to the decline in growth rates. The proposition suggested by Allen of a danger of population explosion in the Soviet Union under an alternative economic scenario is totally unfounded (Allen, 2003, pp. 111–32).

The rate of growth of labour inputs, however measured, declined during the post-Second World War period from more than 1.5 per cent annually, faster than the population's, down to 0.1 per cent, far below the population's, during 1985–90. During the last decades of the Soviet period the population reached the rates of growth of highly developed countries. The same is true of the rates of labour force participation, especially of women; indeed, on this score the Soviet rates surpassed those in some

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
developed countries. As a result, as time went by the ‘extensive’ model turned more and more into ‘capital-led’. The negative effects on output per capita of the saturation of labour participation rates are clear. The excess decline in the rate of growth of population negatively affects GDPPC through an increased per capita burden of public services such as defence and public administration.
Extensive, or ‘capital-led’ growth, Soviet-style, is therefore first of all a dead end. Second, any given level of GDPPC is achieved with more labour (including during off-work hours) and capital, which leaves the population a smaller share for consumption as compensation, than in market economies at similar levels of development.

Soviet growth strategy: catching up and haste

As mentioned above, the strategy of maximum growth was also motivated by the desire and drive to catch up with the West and surpass it. The rationale was provided by Stalin in 1931: ‘We are fifty or a hundred years behind the advanced countries. We must make good the distance in ten years. Either we do it or they crush us’ (cited by Berliner, 1976, p. 161). This goal was repeated time and again by other Soviet leaders. The growth strategy described above was designed to accomplish just that; but the sense of urgency added one more element to it, namely, ‘virtuous haste’ as termed by Gregory Grossman (1983). Haste is defined here as actions taken in order to bring about higher rates of growth in the near future at the expense of future growth. The economic merits of such an action, or strategy, depend on the rate of time preference of the decision-maker. The more impatient he is, the greater is the room for more worthwhile haste. In what follows we list actions taken by the Soviet leadership that seem to testify to a very high rate of time preference. Alternatively, such actions can result from miscalculations regarding the real costs of their actions in terms of future growth. It seems that both played a role and that miscalculation was also important. Haste can also be looked at as an act of borrowing higher growth rates than available in the present. Since the Soviet Union could hardly borrow abroad, it had to mortgage its own resources and future growth. One can therefore learn about the degree of miscalculation by comparing the rate of interest that the system was ready to pay (its rate of time preference) with the rate it actually paid. This is equivalent to comparing the integral below an intended curve depicting growth rates over the period with the curve of actual growth.
A high rate of time preference seems to contradict the essence of the growth strategy, described above, of high rates of investment and readiness to postpone consumption into the future, which signify low rates of time preference. One way to reconcile this apparent contradiction is to note that patience was imposed on the people rather than on the leadership. The people would probably have preferred a ‘normal’ maximization of utility over time. The objective function of the leadership reflects high time preference, the maximization of present growth even at the expense of future growth. Therefore, compared with an objective function that maximizes the welfare of the people, the policy of haste will be depicted as a steeper down-sloping trend of growth rates, cutting the ‘optimal’ curve from above.
Soviet growth was notorious for over-depleting natural resources and in causing costly damage to the environment. When the rates of oil extraction started to decelerate during the 1970s, pressures to keep oil flowing resulted in over-pumping and the flooding of wells. Overuse of fertilizers and land in order to grow more cotton depleted large areas of arable land and of water sources in Uzbekistan. Lake Baikal and other water and land resources were contaminated by overuse and by industrial and nuclear waste;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
air pollution and other environmental issues were disregarded. Relevant here are also the acceleration of the demographic transition and the early mobilization of labour at the cost of labour shortages later. The high and increasing volume of investment, combined with pressures to complete projects fast, was responsible for lower quality, for a thinner technological content, and for mistaken decisions when taken in a hurry, all with negative effects on future efficiency and growth. Investment is treated by growth theory as a major vehicle for the introduction of new technologies and of processes of ‘endogenous growth’. Haste prevented this from taking place in the Soviet Union despite high investment rates. The haste to construct many new projects simultaneously was also responsible for the permanently large and increasing stock of incomplete projects, further reducing the levels of capital productivity.

Rapid industrialization required vast investments in infrastructure, in urbanization, utilities, and in transport and communication networks. Haste demanded that these should be limited to the minimum level necessary or below, saving investment resources for ‘productive’ projects with shorter payback periods. This was also true of infrastructure services inside enterprises, where the peripheral activities serving production lagged behind. Over time, and especially towards the end of the regime, the level of maintenance of infrastructure services deteriorated. Over the second half of the period the quality of education and health services also deteriorated, a result of conservatism and inertia and little contact with the outside world, as well as mounting resource constraints.

In sum, haste contributed to the trend of declining rates of growth, over and above the factors mentioned earlier. The deficit in growth during the later period can be thought of as payments of interest and payments against the still growing stock of debt. The accumulated debt was endowed to the new regime in the forms of obsolete production capacities, over-depleted natural resources, a large environmental deficit, run-down physical and service infrastructure, a debt that would have to be repaid as a precondition of and barrier to the resumption of growth. It definitely contributed to the initial decline in output and other difficulties during the transition.

More generally, ‘haste’ is part and parcel of the selected growth strategy, and of the economic system and political regime. One justification for such a choice, discussed in the economics literature, but also advanced by the Soviet Communist leaders, is that it provided the (only?) method, or at least an efficient and quick one, of economic take-off, the break away from the vicious circle of the low-development trap. We now know that the selection of the Soviet model of growth had to take into account a future model shift. While it is clear that the choice made by the Soviet leaders was motivated, at least partly, by internal and external power considerations, it is worthwhile examining the merit of such a choice from a purely economic point of view. As mentioned above, the choice of an economic growth strategy was hotly debated in the Soviet Union during the 1920s (Erlich, 1960).

Two-stage development strategies and the cost of switching

In 1991 Russia joined, *ex post* and with no initial intention, a group of countries with two-stage development strategies: a first stage for the purpose of take-off, an initial modernization and industrialization drive, and a second stage, starting with the transition, for joining the ‘normal’ route to ‘modern economic growth’, consisting of a more balanced growth via market mechanisms and private property, moderate government intervention, and a democratic and open society (Rodrik, 2005). The argument justifying a special initial strategy, is that a successful take-off requires more drastic policy steps, more forceful institutions and more determined and intensive government intervention. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic payoff of such a strategy differs across countries. However, usually little attention is paid to the costs of switching to the second stage of normal growth. These costs are higher the wider the differences are between the economic, social and political tools used for the take-off effort and those required for ‘modern economic growth’. Therefore, even if the first stage is considered successful, the overall evaluation and the calculation of net benefits have to take into account these costs too. The costs of switching are made of up three elements: the costs of delaying the shift beyond its appropriate timing, caused by vested interests of the regime in power and fear of the unknown future; the actual costs of switching and transition when they come; and the size of the accumulated debt from ‘haste’, which is an intrinsic part of any take-off. The delay and the transition periods are longer, and all three cost elements higher, the wider the difference is between the economic and institutional set-up of the take-off and the post-switching system. One clear advantage of a uniform strategy of ‘balanced growth’ from the start is that economic modernization and growth go hand in hand with the evolutionary development of the appropriate institutions, so that virtually no radical switching is required.

The case of the Soviet Union is one where the above described gap between the two stages is near the maximum thinkable; so accordingly also are the expected costs. The central political and economic control and the suppression of freedoms of all kinds made it easier to take strategic decisions and to impose a heavier burden on the population. But these came at the costs of destroying and blocking the development of social and political institutions needed eventually for the second stage.

As mentioned above, to this day there are some disagreements as to the merit of the Soviet take-off strategy, standing alone. Most observers see some benefits in, even an economic justification for, the big push forward and the early Soviet industrialization drive. To be sure, nobody condones Stalin's atrocities, and few if any are ready to accept the human costs associated with that period, whatever the achievements. But, assuming that similar, even better results could have been achieved with a milder variant of the same model, one can appreciate some of the economic outcomes – relatively fast growth and radical structural change. Even if one takes into account the destruction caused by the Second World War and the exaggerated concentration on the Cold War and on military build-up, the Soviet economy did succeed by 1970 in accomplishing a degree of catching up and, even more than that, in putting the Soviet Union on the road towards economic modernization. Robert Allen (2003), writing after the collapse of the Soviet regime, goes beyond the above and argues that the take-off strategy saved the Soviet Union from the dismal fate of many developing countries, and that no alternative strategy could have accomplished such a feat. At the same time, the only serious attempt at a counterfactual calculation produced a feasible and better alternative, along a route to more balanced growth, and, what is significant, a strategy that would have reduced considerably the costs of switching (Hunter and Szyrmer, 1992).

The figures presented in the first part of this article suggest a degree of support for the Soviet growth and modernization record during the regime's first three decades or so, but only when no account is taken of the oceans of human suffering inflicted on the population by Stalin and, somewhat less so, by the other leaders. At the same time, the calculation so far, including the cost of shifting, demonstrates without any doubt that, even on the basis of pure economic considerations alone, the Soviet growth project was probably not worthwhile.

The Soviet system was established as a long-term and sustained alternative economic and social system, with no intention whatsoever to switch later. Yet, during the 1950s and early 1960s, the Soviet leaders became worried about the low efficiency of the economy and lagging technology, and the prospect of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
declining growth rates. For a time there were quite open discussions over the proper future course, including various proposals to move to a more decentralized system, in the direction of what was termed ‘market socialism’. The discussions and proposals were limited to the economic mechanisms and did not include any ideas regarding changing the political regime. Yet, in retrospect, this seems to have been the appropriate time for switching to a new system altogether. *Per contra*, the discussions culminated in 1964 in a package of the so-called Kosygin reforms. It was a watered down variant of previous proposals. Only mild attempts were made at implementation, and within a few years the Soviet Union had sunk into the ‘stagnation era’ under Brezhnev and his followers. The actual delay in the shift is fully understandable, given the strong regime and economic system and strategy introduced in 1928. The totalitarian Communist regime, backed by a strong ideological paradigm and an even stronger military backing, could not have been expected to replace itself only on the basis of rational and valid arguments. Would President Gorbachev have acted as he did had he known the final consequences?

The change came 15–20 years later. During 1970–90, the period of delay, the average rate of growth of GDPPC stood at just above one per cent per year or at near zero if 1991 is added in, as compared with 3.4 per cent during 1928–70. It was followed, starting in 1990 or 1991, by the transitional output decline of nearly 40 per cent, which was halted and reversed only from 1999. This decline is partly a manifestation of the difficulties in making the extreme shift of institutions, formal and informal. Over the entire period of delay plus early transition – that is, from 1970 to 1998 – GDPPC *declined* at an average annual rate of 1.3 per cent.

The entire period related to the take-off lasted, therefore, 70 years, from 1928 to 1998, rather than about 30, its segment of rapid growth. The rate of growth associated with this entire period stands at a mere 1.5 per cent per year, less than half the rate during the period of fast growth. This is not a great achievement by comparison with other countries at similar levels of development, or even developed countries. It is even more dismal when the higher sacrifices in terms of labour inputs and lower consumption levels are factored in.

One can think of other dates than 1998 to signal the end of the transition period: one is when Russia resumes its pre-fall GDPPC level of 1990 or 1991, possibly around 2008; yet another date could be when Russia regained the highest level of GDPPC of 37.5 per cent of the US level, reached back in 1970. An optimistic date, based on much guessing, could come between five and ten years after 2008. If true, then the delay plus transition periods will last much longer than the period of rapid growth itself, reducing the hypothetical growth rates still further.

Concluding note

The Soviet Union initiated and implemented a heroic social experiment that included a rapid and fairly successful, albeit rather distorted, process of modernization and growth. It culminated more then 70 years later in a dead end that required a difficult ‘transition’ in order to join the main road to modern economic growth. *Ex ante* it was not intended to be an experiment, nor a temporary phenomenon. There are many lessons to draw: one is that, even on the basis of economic accounting alone, and when the costs of transition are included, it is highly doubtful whether the experiment, judged as a take-off strategy, paid off. It clearly failed as a sustained alternative. Yet the counterfactual analysis, based on the experience of other countries that did better, is always conditioned on the issue of feasibility: could an alternative strategy, as indeed was advocated for the Soviet Union during the 1920s, have pulled the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Soviet Union into take-off and sustained growth?

See Also

- command economy
- Soviet economic reform
- Soviet Union, economics in
- Soviet and Russian research, development and innovation
- transition and institutions

Bibliography

Allen, R.C. 2003. *Farm to Factory: A Reinterpretation of Soviet Industrial Revolution*. Princeton and Oxford: Princeton University Press.

Åslund, A. 2002. *Building Capitalism: The Transformation of the Former Soviet Bloc*. Cambridge: Cambridge University Press.

Bergson, A. 1961. *The Real National Income of Soviet Russia since 1928*. Cambridge, MA: Harvard University Press.

Bergson, A. 1973. Toward a new growth model. *Problems of Communism* 22, 1–9.

Bergson, A. 1997. How big was the Soviet GDP? *Comparative Economic Studies* 39, 1–14.

Bergson, A. and Levine, H.S. 1983. *The Soviet Economy: Toward the Year 2000*. London and Boston: G. Allen & Unwin.

Berliner, J. 1976. *The Innovation Decision in Soviet Industry*. Cambridge, MA: MIT Press.

Birman, I. 1989. *Personal Consumption in the USSR and the USA*. London: Macmillan.

Easterly, W. and Fischer, S. 1995. The Soviet economic decline. *World Bank Economic Review* 9, 341–72.

Erlich, A. 1960. *The Soviet Industrialization Debate, 1924–1928*. Cambridge: Cambridge University Press.

Grossman, G. 1983. The economics of virtuous haste: a view of Soviet industrialization and institutions. In *Marxism, Central Planning, and the Soviet Economy: Economic Essays in Honor of Alexander*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Ehrlich, ed. P. Desai. Cambridge, MA: MIT Press.

Hanson, P. 2003. *The Rise and Fall of the Soviet Economy*. London: Longman.

Harrison, M. 1993. Soviet economic growth since 1928: the alternative statistics of G.I. Khanin. *Europe–Asia Studies* 45, 141–67.

Heston, A.,Summers, R. and Aten, B. 2002. *Penn World Table Version 6.1* (ICPb). Center for International Comparisons at the University of Pennsylvania (CICUP). Online. Available at http://pwt.econ.upenn.edu/php_site/pwt61_form.php, accessed 8 February 2007.

Hunter, H. and Szyrmer, J.M. 1992. *Faulty Foundations: Soviet Economic Policies 1928–40*. Princeton: Princeton University Press.

Ickes, B.W. and Ryterman, R. 1997. Entry without exit: economic selection under socialism. Working paper, Penn State University.

Kuznets, S. 1966. *Modern Economic Growth. Rate, Structure, and Spread*. New Haven and London: Yale University Press.

Maddison, A. 1995. *Monitoring the World Economy*. Paris: OECD.

Maddison, A. 2001. *The World Economy: A Millennial Perspective*. Paris: OECD.

Moorstein, R. and Powell, R.P. 1966. *The Soviet Capital Stock 1928–1962*. Homewood, IL: Richard D. Irwin.

Noren, J. and Kurtzweg, L. 1993. The Soviet economy unravels: 1985–91. In US Congress, Joint Economic Committee. *The Former Soviet Union in Transition*, vol. 1. Washington, DC.

Ofer, G. 1987. Soviet economic growth: 1928–1985. *Journal of Economic Literature* 25, 1767–1933.

Popov, V. 2007. Life cycle of the centrally planned economy: why Soviet growth rates peaked in the 1950s. In *Transition and Beyond: A Tribute to Mario Nuti*, ed. S. Estrin, G. Kolodko and M. Uvalic. Basingstoke: Palgrave Macmillan.

Rodrik, D. 2005. Growth strategies. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Summers, R. and Heston, A. 1991. The Penn World Table (Mark 5): an expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* 106, 327–68.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

United Nations. 1999. *International Comparison of Gross Domestic Product in Europe 1996* (ECP96). Geneva: UN Statistical Commission and Economic Commission for Europe. (Also ECP80, 85, 90, 93).

Weitzman, M. 1970. Soviet postwar economic growth and capital–labor substitution. *American Economic Review* 60, 676–92.

Wilhelm, J.H. 2003. The failure of the American Sovietological economics profession. *Europe–Asia Studies* 55, 59–74.

How to cite this article

Ofer, Gur. "Soviet growth record." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000484> doi:10.1057/9780230226203.1579

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Soviet Union, economics in

Vincent Barnett
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Soviet economics reflected the contradictory attitude of Marxist theory to Western economics, and the different stages of the development of the Soviet economy itself. The most original work occurred in the 1920s, with many contributions by Soviet-based economists to fields like business-cycle analysis and agricultural economics still enduring today. After 1929 the repression of leading intellectuals affected economists in a particularly damaging way, and Soviet economics never recovered its cutting-edge position thereafter. After 1953 a partial thaw led to reform-minded economists gaining confidence, and their efforts provided inspiration for the early stages of market reform after 1985.

Keywords

Abalkin, L.; Aganbegyan, A.; Aslund, A.; Austrian economics; banking crises; Barazov, V.; Bogdanov, A.; Bohm-Bawerk, E. von; bourgeois economics; Brutzkus, B.; Bukharin, N.; capitalism; central planning; Chayanov, A.; Chetverikov, N.; Commons, J.; Conjuncture Institute (USSR); Domar, E.; Econometric Society; Fel'dman, G.; Fisher, I.; Gaidar, Y.; Gerchuk, Y.; *glasnost* ; Gorbachev, M.; Greman, V.; Hayek, F. von; hyperinflation; industrial democracy; Kagarlitsky, B.; Kantorovich, L.; Katsenelinboigen, A.; Keynes, J. M.; Kolmogorov, A.; Kondratieff cycles; Kondratieff, N.; Konyus, A.; Koopmans, T.; Kornai, J.; Kruschev, N.; Kuznets, S.; labour time vouchers; Lange. O.; Lenin, V.; Leontief, W; Liberman, E.; Litoshenko, L.; marginal revolution; market socialism; Marschak, J.; Marx, K.; Marxist economics; Mises, L. von; Mitchell, W.; multi-sector accounting; Nemchinov, V.; New Economic Policy (USSR); Novoshilov, V.; Oparin, D.; Optimal planning; Peasants; *perestroika* ; Pervushin, S.; Petrakov, N.; Popov P.; praxeology; Preobrazhensky, E.; primitive socialist accumulation; privatization; Prokopovich, S.; Rainov, T.; Sachs, J.; secondary economy; shadow pricing; Shaposhnikov, N.; Shatalin, S.; shock therapy; Slutsky, E.; Socialism; Soviet economic reform; Soviet Union, economics in the; Stalin, J.; Strumilin, S.; system of optimally functioning economy (SOFE); Tugan-Baranovsky, M.; Vainshtein, A.; Veblen, T.; Yavlinsky, G.; Yurovsky, L.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Article

Economics in the Soviet Union progressed through a number of distinct stages of development that were linked to the evolution of the Soviet system itself. It was also replete with multi-level contradictions. Actively building upon the foundations laid down by the British classical economists, Karl Marx's prognosis of capitalist collapse had envisaged a new role for economic ideas as the blueprint for the creation of a socialist utopia; even so, official Soviet doctrine came to view Western ‘bourgeois’ economics with unreserved hostility. After 1929 much of Soviet economics declined dramatically to become a subservient puppet of its Communist Party masters, yet before 1929 the Russian contribution to mainstream economic theory had bloomed to a new level of international respectability. Even under Joseph Stalin's gaze the works of some Western economists were translated into Russian – J.M. Keynes's *General Theory* was published in the USSR in 1948 – but other Western authors who were critical of the Soviet system were banned. And just as the USSR collapsed at the end of the 1980s and central planning was finally laid to rest, a number of Western economists were actively planning exactly how to transform Soviet-type economies into their market-based opposites. To summarize the results of the Soviet experiment for economics widely interpreted, it was an essential failure coupled with an episodic success.

1917 – 1929

The first problem encountered after 1917 was whether ‘economics’ as a discipline would be required at all, given that planning was supposed to be a science of collectively organized production. Did economic laws as natural regulators apply only to commodity production, or would new economic laws be developed in socialism? N.I. Bukharin's *Economics of the Transition Period* (1920) exemplified the liquidationist call for an end to all monetary accounting, anticipating that the naturalization of economic thinking would occur. This approach was however soon discarded when the New Economic Policy (NEP) was introduced in 1921. Bolshevik leaders then enthusiastically embraced the principles of ‘sound money’ and market exchange. For example V.I. Lenin declared that what was needed was ‘less politics and more economics’ and advised that communists should ‘learn to trade’. But against Bukharin's proposal for an industrial democracy, Lenin still declared that politics should take precedence over economics and that the proletarian dictatorship must prevail. This contradictory attitude remained an undercurrent in all official Soviet proclamations on economics until 1991.

Despite such inconsistencies there is no doubt that the 1920s were the decade in which Russian and Ukrainian economists contributed the most to international developments in economic theory. Immediately recognizable names like N.D. Kondratiev, E.E. Slutsky and A.V. Chayanov were only the tip of an iceberg in terms of the contributions made by Soviet-based economists to fields like business-cycle analysis, agricultural economics, monetary theory and the economics of planning. Kondratiev and Slutsky were actively involved in cross-country collaborations through personal connections with Western economists like Wesley Mitchell and through their membership of international bodies such as the Econometric Society. Moreover, if the influence of Russian-born émigré economists is also considered, then the international impact of economics originating in the Soviet Union before Stalin's rise to dominance is difficult to deny.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
To take the relevant fields of economics in turn: with respect to business-cycle analysis, Kondratiev and his colleagues in the Moscow Conjuncture Institute built upon the pre-revolutionary contributions of M. I. Tugan-Baranovsky, and added various extra dimensions of analysis including Mitchell-style empiricism, greater statistical sophistication (such as Fourier analysis) and a more direct interest in contemporary policy concerns (the effects of peasant taxation). Kondratiev (1922) began the decade with a detailed study of the Russian grain market, using differing levels of farm marketability to explain its collapse during the war and revolution, before embarking upon a more general analysis of the world economy (1925) using a three-cycle schema of long, medium and short cycles that was later employed by Joseph Schumpeter. The notion of 50-year-long cycles generated by the periodic creation of basic capital goods quickly achieved some international notoriety. Other members of the Conjuncture Institute focused on topics such as scientific discovery as causation (T.I. Rainov), the methodology of cycle analysis (N.S. Chetverikov), seasonal fluctuations (Ya.P. Gerchuk) and the theory of economic prognosis (A.L. Vainshtein). Outside the Conjuncture Institute, S.A. Pervushin (1928) analysed cyclical movements as being composed of various shifts in price relations, and suggested that in Russia from 1890 to 1913 domestic factors such as the harvest were increasing in importance as causative influences. Even before 1917, agriculture had been a key focus for Russian economists; after 1917, this interest blossomed into a multitude of different approaches. As director of the Institute for Agricultural Economics, Chayanov (1925) focused on the structure and optimal size of farms, and the motivating drive of peasants. He argued that peasants should not be modelled as *homo economicus*, but rather a labour-consumption balance was more appropriate to them, in which both the monetary and the non-monetary needs of the family were evaluated against the drudgery of labour performed. L.N. Litoshenko, by contrast, viewed peasant proprietors as driven solely by acquisitive motives, and supported rural class differentiation as a means of improving agricultural techniques. E.A. Preobrazhensky (1926) identified the agrarian sector of the economy as a source of funds for state-induced industrialization that should be ‘pumped over’ into the industrial sector by price policy, as part of a process called ‘primitive socialist accumulation’. Against this idea Bukharin encouraged peasants to ‘enrich themselves’ as a means of fostering growth in both agriculture and industry. And Kondratiev provided detailed forecasts of international grain markets using the latest insights of cycle theory in order to facilitate Russian agricultural exports as a means of financing industrial development (Barnett, 1998).

Monetary theory experienced an unexpected boost in the early Soviet context through the need to establish a stable currency in civil war conditions. Debates over the role of money in a socialist economy occurred from 1917 onwards, with contributors from the left and the right clashing over fundamentals. Tugan-Baranovsky advocated a system of paper money in which metallic reserves would not circulate, while Marxists like S.G. Strumilin proposed various non-monetary accounting schemes such as labour time vouchers and energy units. Preobrazhensky celebrated the profligate issue of paper currency as a method of political confrontation, but L.N. Yurovsky's idea (1925) for a parallel gold-backed currency to counter the resultant hyperinflation eventually won out after 1921. Keynes's design for a currency board for north Russia in 1918 was quickly overtaken by events. These issues were part of a larger debate over the significance of war communism (1918–1920), a system that A.A. Bogdanov characterized as being driven by acute shortages, rather than being the result of the rational application of socialist economics. Most obviously of all, the economics of planning attracted a huge amount of effort from many Soviet

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economists, especially in relation to industrialization strategy. In 1917 Tugan-Baranovsky outlined a method of planning using marginalist techniques, but this soon became doctrinal heresy. The first detailed Soviet effort was the GOELRO electrification plan of 1920 that employed an engineering approach, but as NEP progressed a more sophisticated planning methodology developed. Debate centred on questions such as the weight of extrapolation from current trends (genetic planning) against desired ultimate goals (teleological planning), the nature of the interrelation between state industry and private agriculture, and the role of Quesnay-type balances in the planning process (Barnett, 2005).

For example, building on the agricultural census developed by pre-revolutionary *zemstvo* (local government) statisticians, P.I. Popov pioneered the preparation of an economy-wide balance for 1923/24 from within the Central Statistical Administration. As part of the genetic approach to planning developed inside the State Planning Agency, V.G. Groman and V.A. Bazarov (1925) uncovered various empirical regularities that operated in the NEP economy, such as the law of market saturation, using models adapted from the natural sciences. From within the Conjuncture Institute, N.N. Shaposhnikov applied the net present value principle to the planning of capital investment projects, and Kondratiev prepared a detailed plan for agriculture and forestry for 1924–8 that was based on an indicative ‘perspectives’ approach. By the end of the 1920s, both yearly ‘control figures’ and five-year imperative plans had been developed, the latter containing hundreds of pages of figures that plotted the progress of all branches of Soviet industry as centralized directives.

Slutsky's economics-related work in the 1920s was in a theoretical league of its own. His first important contribution was to the mathematical modelling of currency emission, where he compared various complex formulae with the reality of Soviet monetary policy after 1917, in order to calculate the income that the state received from emission. He then turned his attention to the praxeological foundations of economics and provided a set of axioms for describing the parameters of an economic system. And in the same year as he published the groundbreaking paper suggesting that cyclical processes in the economy might be modelled as the summation of independent chance causes (1927), he also wrote a critique of Bohm-Bawerk's conception of value. An important feature of Slutsky's random cycles was periodic disarrangement and consequently regime change, which served to distinguish them from approximately regular business cycles. The 1927 article was the result of many years of work on the theory of stochastic processes that eventually produced a new conception of the stochastic limit. After the closure of the Conjuncture Institute in 1930 Slutsky was not arrested, and he continued to pursue related topics such as the use of the extrapolation method in relation to random processes. Slutsky's mathematical work was also employed in the set theoretic approach to probability theory developed by A.N. Kolmogorov.

In various unconnected fields, in 1924 A.A. Konyus drew upon the contributions of Irving Fisher to develop a sophisticated cost of living index that lay between the Paasche and Laspeyres indices, and in 1926 he suggested the idea that consumer preferences could be represented in terms of both prices and income. In 1925 Strumilin assessed the economic benefits of education by evaluating the length of study undertaken against the increased skills obtained, suggesting that education should be taken to the point where marginal cost equalled marginal revenue. And in 1928 G.A. Fel'dman developed a two-sector model of economic growth in which the ratio of the capital stock of the producer and consumer goods sectors was related to projected growth rates, predicting that (for a stable growth path) investment must be divided between the two sectors in identical proportion to the stock of capital. This model had direct implications for Soviet planners.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Finally, the emigration of economists like Jacob Marschak, Simon Kuznets, Evgeny Domar and W.W. Leontief after 1917 transferred some of the existing themes of Russian economics to the USA. Leontief's input-output approach (1941), Marschak's Walrasian market socialism, Domar's growth theory and Kuznets' work on secular trend all owed an important debt to their Russian origins, although such work was transformed by the American context. But the 1920s were unquestionably successful in terms of producing many influential developments of relevance outside of the immediate Soviet context.

1929 – 1953

After 1929 the storm clouds of Stalinism poured down upon Soviet economists with uncompromising force. Key figures like Bukharin and Kondratiev were sentenced to long periods in jail, and important centres like the Conjuncture Institute were closed: in response Kondratiev wrote an account of the methodology of economic statics and dynamics. Lesser members of the various economics groupings like Vainshtein and Konyus were dispersed. What took the place of the pioneering Soviet economics of the 1920s was a polarization to ideological and technical extremes.

Vacuous general statements about the current direction of Soviet policy, such as Stalin's 'law of the harmonious development of the national economy', sat alongside the minute detail of the planning of every branch of Soviet industry. This polarization facilitated the Orwellian discrepancy between officially declared aims (economic equality for all) and actual government policies (a system of slave labour camps). The central question of deciding overall plan targets was resolved politically in the Politburo (with assistance from Gosplan and the Commissariats), and the advisory function of economists evaporated. Soviet economic discourse consequently became an instrument of its political masters. However, there remained limited scope for dissent in tangential fields such as mathematics. For example, L.V. Kantorovich initiated his conception of optimal planning in 1939, in response to the problem of distributing the manufacture of parts to available machine tools so as to produce the maximum number of sets of components, or a method of machine loading to obtain the highest productivity. To generalize this idea, an optimal plan was one in which the proposed product assortment was optimally distributed amongst firms at the lowest possible cost of production. Shadow prices or 'objectively determined valuations' were to be used in this process. Kantorovich won the Nobel Prize (jointly with Tjalling Koopmans) in 1975 for this work on linear programming.

Further afield, émigré economists like Boris Brutzkus and S.N. Prokopovich contributed to the debates over the nature of the Soviet system in the 1930s. Brutzkus (1935) took an Austrian-type approach that focused on the centralizing and information-gathering issues, while Prokopovich (1924) criticized the reliability of official Soviet statistics and the state control of property. The socialist calculation debate that viewed the problem either as a computational issue relating to solving sets of equations (Oscar Lange) or as erroneously assuming perfect knowledge (F.A. Hayek) failed to fully engage with the reality of Soviet planning, where bureaucratic and interest group factors were dominant. The Institutionalist response was quite different, with Thorstein Veblen (1921) analysing the Soviet success in Russia in terms of long-prevalent national traditions of economic organization, and John Commons viewing the USSR as the outcome of collective action. Keynes (1925) characterized Bolshevism as business in subordination to religion.

In terms of contextual influences, it is difficult to exaggerate the effect that the Second World War had

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
on Soviet society. The Soviet economy had been placed on a war footing since the mid-1930s, and centralized accounting of material production was seen as crucial to military success. Hence the Soviet system was the direct result of the needs of a war economy, including features such as consumer rationing and strict hierarchical control of industrial production. When the war was finally over, planners found it difficult to adapt to a consumer-led approach. Consequently a semi-underground second economy developed in the USSR that was built upon informal connections, yet was tolerated because of its role in facilitating plan fulfilment. This was theorized by Aron Katsenelinboigen (1977) as a network of coloured markets with differing degrees of non-legality, black being the most extreme. This aspect of the Soviet economy encouraged criminality and had serious consequences for the market reforms of the late 1980s.

1953 – 1985

After Stalin's death in 1953, a thaw began in Soviet intellectual life that had positive consequences for Soviet economics. Previously repressed economists like Vainshtein resurfaced, alongside the coming of age of a new generation of Soviet economists who demonstrated a greater mathematical sophistication than their predecessors. Key members of the mathematical school included N.Ya. Petrakov, V.S. Nemchinov and V.V. Novoshilov, and their organizational base was the Central Economic Mathematical Institute of the Academy of Sciences. Alongside the concern for mathematics went an interest in cybernetics, in Kantorovich-type optimality problems and in economic reforms in general.

The outcome of this new infusion was the development of the concept of a system of optimally functioning economy (SOFE), in which the idea of an optimal plan was taken a number of steps further by adding the notions that plans should be calculated using optimal prices and with concern for resource scarcity and incentive rewards. It also included the idea of an economy as a hierarchical structure in which decision-making should occur at various levels appropriate to the specific task being considered, such as the national economy, a given industry or an individual enterprise, rather than always at the apex of the pyramid. Another important centre of this period was the Institute of Economics and Organization of Industrial Production in Siberia, where in 1965 Abel Aganbegyan stirred controversy by presenting a very negative evaluation of the Soviet economy as antiquated and undemocratic. Foreign commentators also highlighted problems endemic to the Soviet system such as gigantomania (large units were easier to plan), investment cycles linked to production bottlenecks, and soft budget constraints (Nove, 1990).

The Khrushchev era witnessed the first major attempts at economic reform since Stalin, initiated by E.G. Liberman's campaign (1962) to allow enterprises the capacity to plan their own production programmes. While some reforms were implemented in 1965 that attempted to improve industrial management, the impetus for change diminished after 1968, when Soviet tanks were sent into Czechoslovakia. In the 1970s Soviet economics continued to develop along various mathematical, ideological and empiricist paths, including the suggestion of oxymorons such as ‘planned markets’, while the Soviet economy entered a period of relative decline. Burdened by the need to maintain nuclear parity with the West and by imperial overstretch (Afghanistan), its superpower status was increasingly difficult to sustain. Answers about how to solve these problems were not always forthcoming from within official Soviet economics, and, if they were, they were rarely immediately heeded. For example, it took Aganbegyan's reformist economics nearly 20 years to be taken seriously by Soviet leaders.

However, in less contentious areas worthwhile contributions came in the post-Stalin period from

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Vainshtein on measuring national wealth and D. I. Oparin on multi-sector accounting. The standard of empirically focused economic investigation without obvious ideological significance was usually high, and it was often easy to separate out the pseudo-Marxian framework from the valuable detailed content. However, in areas with direct policy significance the quality and reliability of the economics declined dramatically. For example, official statistics on growth rates in the USSR after 1928, perhaps the most contentious subject of all for Soviet economists, certainly suffered from ideological interference, but even Western economists who tackled this topic objectively (such as Kuznets) came to somewhat ambiguous conclusions. In general Soviet economic theory between 1929 and 1985 was thoroughly dominated by political concerns, but this did not necessarily invalidate every aspect of the all of the work done in this period.

1985 – 1991

With Mikhail Gorbachev's accession to power, the twin tracks of cultural openness (*glasnost*) and economic restructuring (*perestroika*) were opened. This was presented as a return to a NEP-style mixed economy, with ‘socialist markets’ and entrepreneurial cooperation being hailed as solutions. This opened the floodgates to the reprinting of material by repressed economists like Bukharin and Litoshenko, and to the open discussion in Soviet economics journals of topics like the neoclassical theory of production and distribution. Soon after this, Western authors like Hayek and von Mises were being translated. In 1989, a government programme was issued acknowledging that ‘the market’ must take precedence over the plan. Some of the key economists of the Gorbachev period were Aganbegyan, Stanislav Shatalin and Leonid Abalkin, who were part of an old guard that had always advocated reform. They were, however, quickly overtaken in the audacity of their transition programmes by a younger generation of economists such as Grigory Yavlinsky and Yegor Gaidar. Various concepts were applied to explain the dilemmas facing the Soviet economy in this period, such as monetary overhang (savings without prospect of being spent), market Stalinism (state decrees pronouncing market activity) and then spontaneous privatization (impromptu transfer of state property).

A key point to recognize is that, in the USSR at the end of the 1980s and after being suppressed for so long, pro-market economics was a ‘revolutionary’ set of ideas that gave Russian advocates a sense of liberation from decades of stale dogma, although some had made the intellectual change much earlier. The main focus of the early proponents of market reform was mass privatization of state-owned property and the liberalization of state-controlled prices, as contained in Yavlinsky's 500-days programme of 1990. Macroeconomic stabilization policies (such as balancing the state budget) took third place. This shock-therapy approach was influenced by Western economists like Jeffrey Sachs, and also by eastern European theorists such as Janos Kornai. Advice on upholding the appropriate sequence of economic reforms, as articulated by Ronald McKinnon (1991) based on previous experience in Latin America and Asia, was not fully heeded in the ‘transition fever’ of the time. Many Western economists were actively involved in propagating market-friendly ideas and policies even in the early Gorbachev period, and they often overwhelmed any lingering loyalty to Marxist economics through the greater sophistication and more rigorous technical framework of their work.

It should also be acknowledged that the particular conception of ‘the market’ espoused by many Russian economists at this time was somewhat one-sided, an amalgam of favourable assessments such as Frank Knight's conception of entrepreneurial capacity, early Austrian capital theory and Milton Friedman-style

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
monetarism, complemented by Kornai's account (1980) of the Soviet shortage economy. The Keynesian tradition was difficult to discern. In discarding the socialist heritage, the baby (market critiques) had been thrown out with the bath water (central planning). In hindsight, many market advocates appeared rather naive about the consequences of releasing the genie of unfettered self-interest from the Soviet bottle. The question of whether ‘the market’ as a universal general mechanism (Adam Smith's ‘invisible hand’) actually existed, as against the idea of many different types of markets as institutions that were influenced by local and national conventions, grew in significance as the reforms progressed. Perhaps one reason for the sidestepping of Keynes's work was that it had been openly discussed even under Stalin, and hence was seen by some as tainted.

In terms of results, the consequences of the post-Soviet transition are fundamentally contested, with some economists celebrating the creation of a successful market economy in Russia (Anders Aslund) while others protested against oligarch-controlled mafia capitalism (Boris Kagarlitsky). What is unquestionable is that the initial rush to privatization and liberalization has been tempered by more recent concern for the development of stable legal institutions and a mature business culture. If banking panics and financial crashes are a sign of a functioning market system, then in the summer of 1998 Russia experienced a classic example, sparked by the government defaulting on its debts.

Conclusion

With the collapse of the USSR at the end of the 1980s, and China's conversion to ‘capitalism with a socialist mask’, the West has undoubtedly triumphed in the battle of comparative economic systems. However, it did so at significant ideological cost. Firstly, state intervention in the economy had to be embraced by many Western governments after 1936. Secondly, the goals of full employment, a welfare state and (more recently) fairness in international exchange have become accepted policies in many developed countries, even if they are not often achieved. And thirdly, Western economics was transformed after 1945, with the mainstream integration of ideas such as public goods, monopolistic competition, cooperative games, social cost and even status goods, that would make it unrecognizable to someone like Bukharin versed in the Austrian approach of 1914, and even more so to Marx, who had never even recognized the ‘marginal revolution’ of the 1870s. The refusal of Soviet economists to officially engage with new developments in Western economics was thus understandable, since had they had done so they would have realized the simplistic caricature that had been foisted onto them.

Furthermore, the existence of the USSR altered to some extent the purpose of mainstream economics as it was conducted in the West. Before 1917, economists served the state mainly as academic theorists and as advisers on specific topics, for example on monetary or fiscal policy. After 1945, with the onset of the Cold War, the role of some economists widened to providing more general advice on economic development issues. Moreover, the responsibility of a few Western economists increased even further than this. For example W. W. Rostow, famous for his stages theory of economic growth, was a development economist with a detailed knowledge of British trade cycles. Summoned to serve in the Kennedy administration he became a key advocate of increased US involvement in Vietnam, arguing for more American troops on the ground and heightened bombing of the North, in order to contain the expansion of communism in South-East Asia. That it was a professionally trained economic historian giving advice to US presidents on geopolitics was indicative.

To conclude, the USSR collapsed spectacularly, and its official economic ideology was derisory even

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
when first issued, yet many Russian/Soviet economists produced work of lasting value that was influential at the time of its first issue, and is still being referred to today. All things considered, the Soviet contribution was (in spite of Stalin's efforts) perhaps the most influential experimental failure in the development of economic ideas thus far.

See Also

- economic calculation in socialist countries
- socialism
- Soviet economic reform
- Soviet growth record
- Stalinism, political economy of

Bibliography

Aganbegyan, A. 1988. *The Challenge: Economics of Perestroika*. London: Hutchinson.

Aslund, A. 1995. *How Russia Became a Market Economy*. Washington, DC: Brookings Institution.

Barnett, V. 1998. *Kondratiev and the Dynamics of Economic Development*. London: Macmillan.

Barnett, V. 2005. *A History of Russian Economic Thought*. London: Routledge.

Bazarov, V. 1925. On ‘recovery processes’ in general. In *Foundations of Soviet Strategy for Economic Growth*, ed. N. Spulber. Indiana: Indiana University Press, 1964.

Bergson, A. and Kuznets, S. 1963. *Economic Trends in the Soviet Union*. Cambridge, MA: Harvard University Press.

Brus, W. and Laski, K. 1989. *From Marx to the Market*. Oxford: Clarendon.

Brutzkus, B. 1935. *Economic Planning in Soviet Russia*. London: Routledge.

Bukharin, N. 1920. *The Politics and Economics of the Transition Period*. London: Routledge and Kegan Paul, 1979.

Chayanov, A. 1925. *The Theory of Peasant Economy*. Homewood: Irwin, 1966.

Ellman, M. 1973. *Planning Problems in the USSR*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Fel'dman, G. 1928. On the theory of growth rates of national income. In *Foundations of Soviet Strategy for Economic Growth*, ed. N. Spulber. Indiana: Indiana University Press, 1964.

Groman, V. 1925. On certain regularities empirically observable in our economy. In *Foundations of Soviet Strategy for Economic Growth*, ed. N. Spulber. Indiana: Indiana University Press, 1964.

Kagarlitsky, B. 1995. *Restoration in Russia*. London: Verso.

Kantorovich, L. 1965. *The Best Use of Economic Resources*. Oxford: Pergamon.

Katsenelinboigen, A. 1977. Coloured markets in the Soviet Union. *Soviet Studies* 29, 62–85.

Keynes, J. M. 1925. *A Short View of Russia*. London: Hogarth Press.

Kondratiev, N. 1922. Regulation of the grain market and the supply of grain. In *The Works of Nikolai Kondratiev*, ed. N. Makasheva, W. Samuels, and V. Barnett. London: Pickering and Chatto, 1998.

Kondratiev, N. 1925. The long waves in economic life. *Review of Economic Statistics* 17 (1935), 105–15.

Konyus, A. 1924. The problem of the true index of the cost of living. *Econometrica* 7 (1939), 10–29.

Kornai, J. 1980. *Economics of Shortage*. Amsterdam: North Holland.

Leontief, W. 1941. *The Structure of American Economy*. New York: Oxford University Press.

Liberman, E. 1962. The plan, profits and bonuses. In *Socialist Economics*, ed. A. Nove and M. Nuti. Harmondsworth: Penguin, 1972.

McKinnon, R. 1991. *The Order of Economic Liberalization*. Baltimore, MD: Johns Hopkins University Press.

Nove, A. 1990. *Studies in Economics and Russia*. London: Macmillan.

Pervushin, S. 1928. Cyclical fluctuations in agriculture and industry in Russia. *Quarterly Journal of Economics* 42, 564–92.

Preobrazhensky, E. 1926. *The New Economics*. Oxford: Clarendon, 1965.

Prokopovich, S. 1924. *The Economic Condition of Soviet Russia*. London: King and Son.

Slutsky, E. 1927. The summation of random causes as the source of cyclic processes. *Econometrica* 5

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(1937), 105–45.

Sutela, P. 1991. *Economic Thought and Economic Reform in the Soviet Union*. Cambridge: Cambridge University Press.

Veblen, T. 1921. *The Engineers and the Price System*. New York: Viking.

Yurovsky, L. 1925. *Currency Problems and Policy of the Soviet Union*. London: Parsons.

How to cite this article

Barnett, Vincent. "Soviet Union, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000483> doi:10.1057/9780230226203.1580

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Spain, economics in

Alfonso Sánchez Hormigo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Due to the delay in the development of its economy and scientific thought, Spain became a recipient of economic ideas from more advanced centres. Spain's decline from empire to decadence led to a search for strategies which would remedy the situation. Strong institutional obstacles delayed the reception of foreign theories, but did not prevent them from becoming known. This process of reception and redevelopment of foreign ideas was not always linear, nor was it uniform in time.

Keywords

agriculture and economic development; Álvarez, V. A.; Canga Argüelles, J.; Arriquíbar, N. de; autarky; Azpilcueta, M. de; Fernández Baños, O.; Bastiat, F.; Bernácer, G.; Bielfeld, Baron von; Borrego, A.; Bullionists; Campomanes, conde de; Canard, N.-F.; Cao-Pinna, V.; Castañeda, J.; Cobden, R.; Colmeiro, M.; Condillac, E. B. de; corporatism; Danvila, B.; Valle, E. M. del; Dormer, D. J.; Enlightenment; Flórez Estrada, A.; exchange theory; Garnier, G.; Genovesi, A.; George, H.; Gournay, Marquis de; Herrenschwand, J.; Ibn Jaldún; Jones, R.; Jovellanos, G. M. de; Justi, J. H. G. von; Keynesianism; Flores de Lemus, A.; Melon, J. F.; Mercado, T. de; mercantilism; Mora, J. J. de; neoclassical synthesis; Olariaga, L.; Olavide, P. de; Alonso Ortíz, J.; Ortíz, L.; Pebrer, P.; physiocracy; political arithmetic; purchasing power parity; quantity theory of money; Fuentes Quintana, E.; Ramos, E.; Ricardo, D; Roma i Rossell, F.; School of Salamanca; Sardá, J.; Say, J.-B.; Schmoller, G. von; scholastic economics; single tax; Sismondi, J. C. L. Simonde de; Smith, A.; Spain, economics in; Stackelberg, H. F. von; Struzzi, A.; subjective theory of value; Torres, M. de; Ulloa, B. de; Uztáriz, J. de; Varela, M.; Villeneuve-Bargemont, J. P. A. de; Vitoria, F. de

Article

The development of economic thought in Spain should be understood as occurring in a country which,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
due to the delay in the development of its economy and scientific thought, became a recipient of economic ideas from more advanced centres. Spain went from empire to decadence, and awareness of this led people to search for strategies which would remedy the situation. The delay in the reception of foreign theories was determined by strong institutional obstacles, though these did not prevent theories from becoming known. This process of reception and redevelopment of foreign ideas was not always linear, nor was it uniform in time. The Spanish case is of interest when explaining and analysing and the effectiveness of different theories once they cross the borders of the country in which they have been developed.

The conquest of America and the world of scholasticism

The first traces of economic thought in the Iberian Peninsula are found in the age of the Arab civilization. A writer and politician of Tunisian origin, between the 14th and 15th centuries, Ibn-Jaldún, wrote *The Muqaddima*, a treatise on the science of civilization, in which he added an explanation of the birth and decline of dynasties, and which contains the first global vision of economics. *The Muqaddima* also contains an analysis of the cyclical nature of economics, as well as an embryonic model of development.

Later, the conquest of America set a new stage for the prosperous Hispanic monarchy. The transport of precious metals from the Indies and the increase in commercial exchange sparked off a revolution in prices which accompanied the commercial revolution. Population increase and the cultivation of new lands brought about a decline in productivity and, together with a lack of technical innovation, the problems of pauperism and the monetary and financial problems associated with unprecedented inflation, which originated in the discovery of the New World, led a group of theologians and members of the colonial administration to set forth their opinions on the situation of a country which had begun a long process of decline.

The first diagnoses came from a group of doctors, naturalists, politicians, philosophers and theologians, scholastic latecomers, members of the so-called School of Salamanca, whose founder was Francisco de Vitoria (1483–1546). These writers were to be found at the prestigious University of Salamanca, though they had been educated at the most important European universities. They were witnesses to the problems which originated during the conquest of America, and though the trade practices of wealthy Spaniards returning to Spain from Latin America were always ahead of theological doctrine, the scholars attempted to reconcile economics and morality by using natural law based on probabilism and casuistry. This led them to apply Thomistic philosophy to resolve practical affairs and to explain economic problems which had arisen in the world of the Counter-Reformation before the scenario which had opened with the discovery of America. They used the manuals of confessors and wrote a set of works in which they tackled economic problems with the utmost insight, amongst which stood out a subjective theory of value, the quantity theory of money, the theory of exchange, the general doctrine of interest and the workings of the market, as well as questions of fiscal policy in relation to distributive justice.

The cleric Martín de Azpilcueta early on formulated the quantity theory of money in his *Comentario resolutorio de cambios* of 1556, and, through considering of the variations in exchange rates, came close to the purchasing power parity theory. His successor, Tomás de Mercado, author of *Suma de Tratos y Contratos* of 1571, managed to integrate quantity theory into a theory of prices. The Spanish scholastics built a bridge between medieval scholasticism and modern philosophy. They were contemporary with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the writers of the first mercantile system but, unlike them, they were privileged by having had a philosophic-cum-systematic training of Aristotelian–Thomistic origin, which enabled them to focus on economic questions not only from a moral viewpoint but also from a more analytical one.
The ideas of the School of Salamanca were disseminated throughout the rest of Europe and the American colonies from the universities of Coimbra, Alcalá, Mexico, Lima, Rome and Paris. Together with Azpilcueta and Mercado, their principal members included clerics such as Domingo de Soto, Luís de Molina and Juan de Lugo. Scholastic doctrines at the end of the 16th century coexisted with those of the first mercantilist writers, in some cases resulting in an authentic symbiosis.

The age of mercantilism

A group of writers, royal advisers, social reformers and Spanish political economists swelled the ranks of the bullionists, among whom the first of the so-called mercantilists were often found. However, Schumpeter discovered among them a few economics writers whose ideas were distinct from primitive bullionism, and whom he considered to be authors of a quasi-systems, such as Luís Ortíz, who wrote a Memoir in 1558, the year following the first bankruptcy of Phillip II of the House of Austria. Aware of Spain's economic decline, Ortíz investigated a set of remedies which constituted a programme of industrial development.
Ortíz attributed the problems of the Spanish economy to internal price rises in Castile, to the inability of national industry to meet demand from the American continent, and to social disdain towards the skilled classes. The remedies consisted of prohibiting exports of raw materials from Spain and preventing imports of foreign goods. In addition, he investigated Spain's political and social institutions in relation to the workings of the economic system.
With an optimistic vision based on the potentialities of the Spanish economy, some Spanish mercantilists saw the fundamental problem not in the limitations of nature itself, but in social and institutional factors. This led to their proposing the elimination of idleness, the abolition of laws which considered manual labour contemptible, limitations on of luxuries for the rich, and changes in a taxation system whose burden fell unjustly on farm workers, aggravating depopulation. Some, such as Ortíz, saw market unification as necessary, the nobility and the Church being the main obstacle.
The expulsion of the ‘Moriscos’ at the beginning of the 17th century undermined the demographic base, already weakened by epidemics in the previous century; this was combined with a reduction in shipments of precious metals from the Indies. The disastrous policies of the King's advisors only added to the problems caused by the territorial and political dispersal of the Spanish monarchy. Some writers, such as Sancho de Moncada, in his work *Restauración política en España* of 1619, summed up the efforts made by the best writers of the time. He was concerned with quantifying the crisis and identifying the variables which brought it about and the links between them.
As a whole, the mercantilists at the beginning of the 17th century concluded that the root of all evil was international trade, since it enabled foreigners to extract from Spain undeveloped raw materials, as well as silver and gold. The remedy consisted in prohibiting manufactured foreign imports, as well as prohibiting the export of raw materials from Spain. This would stimulate the domestic market and internal spending, benefiting the people by increasing the monarchy's revenue. Some mercantilists, such as Martínez de Mata, insisted on increasing production when establishing a relation between cost and the production function.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Not all voices clamoured for prohibition; they explained the crisis in terms of the enrichment of foreigners and the unfair commerce which Spain maintained with them; some, exceptionally, such as Alberto Struzzi, who published a *Diálogo sobre el comercio destos Reinos de Castilla* in 1624, and, later, Diego José Dormer, author of several politico–historical discourses in 1684, defended economic plans which would reactivate commercial relations with foreign countries, with a different strategy based on improving the domestic competitiveness of Spain's economic sectors.

A second mercantilist phase began in Spain with the economic recovery at the end of the 17th century, during the reigns of the last monarch of the House of Austria, Charles II, and the first Bourbon, Philip V. The political economists of the period, amongst whom Jerónimo de Uztáriz and Bernardo de Ulloa were outstanding, were already more seasoned writers. The former, the most important of the Spanish mercantilists, wrote the *Theórica y práctica de Comercio y de Marina* in 1724, a work which aroused the interest of Adam Smith, and which was translated into several European languages. Uztáriz obtained quantitative data on the Spanish economy, on the basis of which he proposed the lowering of production costs in order to lower domestic prices as a solution. He also proposed a free trade zone with America.

In general, the ‘mature’ Spanish mercantilists concerned themselves with questions such as political unification, the elimination of internal customs and excessive indirect taxes, which accumulated in a chain effect and raised final sale prices. They equally supported privileged commercial companies and the existence of a strong navy, which could be used not only for military purposes but also for commercial ones. Their fundamental aim was to encourage industry, in terms of which they continued to defend the protection of domestic industry from foreign competition, in this way relegating agriculture, though only in order of preference. Foreign countries were no longer considered enemies, but were now an example to be followed, especially the Netherlands, France and the United Kingdom.

A characteristic of the Spanish mercantilists in this last period was their belief in the need to establish a more complex analysis of underdevelopment, with reference to the disadvantages of international trade (with consequent problems for the balance of trade) as well as the problems stemming from low labour productivity, which had its origin in the socio–political and cultural structure of the period. To break out of this vicious cycle of underdevelopment, they proposed that industry should be the key to development and consumption the driving force of industry.

The Enlightenment and economic reform

The long expansive cycle that began at the end of the 17th century had brought about a marked population increase, crop expansion, growth in industrial production and greater development of internal and external trade. Spain, however, was still a feudal society in which the land ownership system had not changed and, despite crop development, agrarian production techniques had not substantially changed.

Although Spanish mercantilism had been characterized as a ‘hardy perennial’, the age of Enlightenment in Spain started under the reign of the House of Bourbon of Charles III (1759–88), in which mercantilist stances in their most extreme version were abandoned, bringing about internal market liberalization and a set of reforms which began to liberate the economy.

For the economic writers of the period, the problem was still the same as in the past: to remove obstacles to economic growth, but with a clear awareness that this should be combined with certain institutional and legal changes, provided that they did not affect the absolute power of the monarch in any way.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Despite the survival of mercantilist ideas, they now had new analytical tools at their disposal, thanks especially to economic ideas received from abroad, such as those of the new British political arithmeticians and of the French Physiocrats and agrarian reformers, together with those of the so-called Gournay group, and later, those of Adam Smith himself, which would circulate uninterruptedly until shortly before the end of the century, when the multiple wars with France and the United Kingdom largely hindered the circulation of ideas.

Some politicians who enjoyed great power and who were close to the monarch, such as the Count of Campomanes, author of *The Discurso sobre el fomento de la industria popular* (1774), amongst other works, and a genuine inspirer of this first transitional phase between mercantilism and enlightenment, drew up a renewal programme whose central element was economic reform. His modernizing ideas were founded on a gradual transformation consisting of (*a*) liberalizing colonial commerce, (*b*) abolishing the price-regulated ‘tax’ on cereals, (*c*) stimulating their commerce, and (*d*) putting an end to the increase in the considerable unproductive properties of the Church, and to the excessive interventionism of guilds. All in all, he designed a coherent system of development based on domestic economic freedom, together with protection against foreign trade, which granted a key role to the promotion of agriculture based on the figure of the independent farm worker, the expansion of ‘popular industry’ and occupational increase.

The Enlightenment saw the creation of the Sociedades Económicas de Amigos del País, organizations which encouraged the study of regional economy, of agricultural techniques and of economic science in general. In these societies, after the Italian model, the first professorships in civil economy and commerce were created, giving rise to the beginning of the institutionalization of political economy in Spain at the end of the 18th century. These societies disseminated economic ideas developed in other European countries, such as those of the British political arithmeticians of the 17th century and those of the French thinkers, mercantilists, advocates of agrarian reform, and physiocrats, namely, Cantillon, Melon, Forbonnais, Mirabeau, Montesquieu, Turgot, and, later on, Condillac, Necker and Adam Smith. The Spanish economics writers, of whom the most prominent were Pablo de Olavide, Enrique Ramos, Nicolás de Arriquíbar, Bernardo Danvila and Francesco Roma i Rossell, were acquainted with the ideas of physiocracy and were especially attracted to their advocacy of agrarian reform, even if the theoretical and analytical core on which they were based, as well as ideas of exclusive agricultural productivity and the single tax, were not accepted as a whole. In this sense, many of the first enlightened Spaniards were more in tune with the ideas developed by the so-called Gournay group, which better reconciled the importance of agriculture, industry and commerce, and in which indigenous mercantilism was in its turn very pronounced.

It was also in this period that knowledge of the *Wealth of Nations* began to spread. However, it was not translated until the 1790s. Not all viewpoints were in favour of liberal agrarianism; others opted for industrial development, the influence of German sources being obvious, especially in the territories of the Crown of Aragon, by which the ideas of the Baron von Bielfeld and von Justi led to more industrialist formulations and justified greater participation by the state and by the nobility in the development of economic activities. Enlightened Spaniards of this first period found themselves at a crossroads between tradition and innovation, and although the influence of mercantilism remained, they created a favourable environment so that in the following years the enlightenment movement would manifest itself with greater strength, giving impetus to the modernization of Spain.

A second stage of enlightenment coincided with the French Revolution. At that time Gaspar Melchor de Jovellanos was commissioned by the *Real Sociedad Económica Matritense* to write the most important

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
work on economics in 18th century Spain, the *Informe de Ley agraria* of 1795, based on the concept of the economy as the main 'government science'. He designed a programme of pragmatic, gradual and moderate liberal reforms, which included the redistribution of land, freedom from leasing, and imposing limits on property inheritance. He also proposed abolishing the privileges of the powerful stock-breeding organization, the 'Mesta', and the liberalization of domestic commerce, which, however, did not include freedom to export or with revision of the taxation system. All of this was contemplated with a programme of public investment and special attention to education, since it conferred decisive importance on human capital. Jovellanos, very well-acquainted with the works of European economists such as Cantillon, Galiani, Mirabeau, Turgot and Necker, among many others, was one of the Spanish economists who read Smith's work with discernment, as a result of which the idea of self-interest as the driving force of economic activity would be consistently present in his work, as it would be for many Spanish economists of the time.

In short, it may be said that in Spain the Enlightenment occurred later than in other countries and had different overtones arising not only from Spain's political and institutional situation but also from other factors such as its comparative economic backwardness. This did not present an obstacle to the enlightened Spaniards when disseminating the main economic ideas synthesized in the United Kingdom, France, Italy and in the Germanic countries; in any case, the canonical plan for the dissemination of economic ideas through the mercantilism–physiocracy–liberal economy chain did not appear to be suitable for Spain, where, along with a more persistent neo-mercantilism, fundamentally of an Italian and French nature, the weak presence of physiocracy in its theoretical aspects could be detected, and also a strong influence from the British and French agrarian reformers, who were widely dispersed throughout the peninsula.

The enlightened Spaniards applied the ideas they received to the solution of the problems of their age. In this sense it may be considered to have been an active reception, stimulated both by the economic societies in which the Economy was studied and which promoted the translation of the principal texts of the economists mentioned above, and by economic publications, which multiplied during the second half of the century. These disseminated ideas, making the last decade of the century an especially fertile one, in which, despite heightened censorship because of the French Revolution, the principal works of the European economists saw the light of day.

Classical economics in the liberal age

The Court of the Spanish Inquisition had already denounced Lorenzo Normante, the first lecturer in political economy, who had begun his classes in the Aragonese Economics Society under the influence of Melon, Genovesi and Condillac, for considering economic ideas which were not in accordance with dictates of the Catholic religion. The same trial dealt with the *Wealth of Nations*, which was censored by the Inquisition in 1792, being accused of tolerance and naturalism.

Despite this, Smith's influence had already been felt by some Spanish economists, who generally made an adapted interpretation of his work. Government collaboration and the suppression of his identity allowed a compendium of his work to be translated by the Marquis of Condorcet in the same year, and only two years later the first complete edition saw the light of day, with a few adjustments here and there to avoid the rigours of the censor. *The Wealth of Nations* was translated by one of the Spanish economists on whom the influence of Smith was clear, the diplomat José Alonso Ortíz. Smith's work

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
quickly became a reference point for the teaching of economics, even though it must be emphasized that its late translation and the prevailing doctrinal influences in Spain simultaneously brought several interpretations of Smith's work into being, especially those of an agrarian nature, which coincided with an avalanche of influences which converged at the beginning of the 19th century.

The introduction of Smith's work was accompanied by the first versions of the mathematical economics of N.-F. Canard, of the neo-physiocratic doctrines of Germain Garnier, or the partially physiocratic ones of Jean Herrenschwand, and those of the other advocates of agrarian reform of a non-physiocratic origin, both British and French. Together with these, the doctrines of the French economist J.B. Say had special relevance. His *Traité d'Économie Politique* was translated early in Spain (1804–7) and used as an official text to teach economics, when Spanish economics took its first step towards institutionalization in universities at the beginning of the 19th century.

Nevertheless, the influence of Smith lasted throughout the first decades of the 19th century; however, by that time the influence of the French economist was decisive. This strong influence (eight editions in Spanish of the *Traité* alone) can be explained by its support for keeping the same groups in power and its more pragmatic nature, which favoured its reception, together with greater accessibility of the French language in Spain, and because it was not subject to the vicissitudes of censorship. The fact is that it inspired all university texts for the first three decades of the 19th century. Smith's ideas were used by his followers to defend non-Malthusian interpretations with respect to population and an industrial development model, modified with a defence of customs prohibition. Some authors, such as Gonzalo de Luna, made an interpretation of the ideas of Say and Smith in a neo-mercantilist vein, which defended industrial development models without altering the political bases of the old regime.

Only those who were in exile for political reasons when the absolute monarch, Fernando VII, returned escaped such influences. They lived in the United Kingdom and became acquainted with the ideas of the classical British authors, especially Ricardo, McCulloch and James Mill. Amongst them stood out José Canga Argüelles, the Chancellor of the Exchequer in the brief liberal period which began in 1820 and author of *Elementos de la Ciencia de la Hacienda* (1825), and the most important Spanish economist of the 19th century, Álvaro Flórez Estrada, the main introducer in Spain of the ideas of Ricardo and James Mill, and in whose works the influence of Sismondi and Richard Jones could also be appreciated. In Flórez Estrada's *Curso de Economía*, published for the first time during his exile in London in 1828, a limited influence of Ricardo's ideas (the *Principles* were not translated until the 20th century) could be seen. Also disseminated with the ideas of Say and Ricardo, though with less intensity, were those of Malthus (almost exclusively his ideas on population), Sismondi (his ideas on the agrarian development model) and Bentham, Condillac and Gaetano Filangieri (though their influence was not solely in the sphere of economics).

Say's influence dwindled from the 1840s, allowing greater doctrinal plurality. The influence of Say's disciples left a clear mark on the main Spanish economists of the time, such as Eusebio María del Valle, Andrés Borrego and Manuel Colmeiro. It was a period characterized by eclecticism, in which distancing from the ideas of the classical school is associated on one hand with criticism of deductive methodology, and on the other with recognizing the negative effects of the development of capitalism on the lowest social classes. This favoured reception of Sismondi's ideas and those of the social Christianity of Alban de Villeneuve-Bargemont, so influential in France. All this included a distancing from free trade and favouring industrial protection.

Richard Cobden's journey to Spain in 1844, in his crusade for the free trade league, united with reception

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of the ideas from the French liberal economists, directed most Spanish economists towards free trade, a trend supported by such texts as *Cinco proposiciones sobre los males que causa la Ley de Aranceles a la nación*, published in 1837 by the liberal economist Pablo Pebrer, and the writings of another exile, José Joaquín de Mora, which sparked off the lengthy so-called protectionism–free trade debate, which continued until the beginning of the protectionist era with the Restoration in 1874. The decisive influence of the ideas of Frederic Bastiat and his *Harmonies Économiques* established an age of economic optimism and defence of liberal ideas in their French version, an influence which separated Spanish economists from knowledge of the most relevant developments in the classical British economy, minimizing the influence of economists like John Stuart Mill, an economist of more advanced and complex interpretation, especially after the events of 1848. Perhaps it was the sparseness of the theoretical analysis in Bastiat's work which favoured its dissemination, together with his outspoken opposition to the situation of social confrontation, which by that time was beginning to spread from France, fuelled by criticism of the republican groups and of incipient socialism in its utopian aspects. The debate was situated in the territory of common-sense economic discourse, which brought about pragmatic empiricism and led to the pronounced stagnation of economic science. In the last three decades of the century even marginalist ideas were unknown, perhaps due to a lack of mathematical preparation of the economists, most of them university professors in law schools, or perhaps due to their lack of dedication to investigation. Neither did the historicism of the German Historical School of Schmoller make its mark, and the influence of the Historical School was reduced to methodological relativism which legitimized public intervention, both commercially and socially. Neither did incipient theoretical Marxism make its mark; Bakunin's anarchism was introduced first, and when this happened it was initially separated from the workers' movement, and with a level of popularization in which the ideas of Marx arrived clearly transformed.

Undoubtedly, Spanish economists lost the thread of economic science in the second half of the 19th century; the main criticisms of classical thought came from their own ranks, where the protectionist change of tack that took place in the 1870s later combined with concern for social problems. This led to a defence of liberalism tempered by historicist influence which legitimized protectionism and social reform when not influenced by fundamentalist pseudo-religious postures which would later be taken up in corporatist thought. Likewise, the German Philosopher Krause's ideas, received through the influence of the philosophy of H. Ahrens, contributed to elaborating the first labour legislation of a paternalist nature in defence of the working class, which in these authors, amongst whom stand out G. de Azcárate, A. Álvarez Buylla and J.M. Piernas Hurtado, was compatible with free exchange and economic liberalism. In this breeding ground, tinged with a spirit of regeneration and extra-scientific approaches, the populist ideas of Henry George had a great impact, his *Progress and Poverty* having a tardy but enthusiastic reception in Spain.

The economic modernization of Spain

In the last few years of the 19th century and in the first few of the 20th century, the educational reform project carried out by the Institución Libre de Enseñanza allowed Spanish undergraduates to travel to foreign universities and become acquainted with doctrines developed beyond Spanish borders, especially in Germany. This was how the renewal of economic studies began, after its long stagnation in the previous century.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Antonio Flores de Lemus, the principle Spanish economist of the first half of the 20th century, was educated at the turn of the century in the lecture halls of German universities under the tuition of Schmoller and Wagner, and, on returning to Spain, became the great disseminator of a neo-historicist and realist stream in economic science which dominated Spanish science until the outbreak of the civil war in 1936. A university department chair in the service of the Inland Revenue, he did research on official economic statistics. The reformation of the Inland Revenue and monetary problems were some of the questions in which his intervention was decisive, such as the period between 1927 and 1929, when the possible return to the gold standard was being considered. Some of his successors, such as Luís Olariaga, an expert in monetary matters, Francisco Bernis, an Inland Revenue scholar, economist and statistician, Olegario Fernández Baños, or Germán Bernácer, who developed a macroeconomic model with Keynesian connotations, distanced themselves from the neo-historicist stance in order to embrace more recent developments in economic science.

At the same time the influence of corporatist doctrines was equally making itself felt, based on *fin de siècle* conservative Christian thought, and an outward admiration for the Fascist movement and the corporatist doctrines of the Italians which quickly took root in the most conservative sectors of the Spanish intelligentsia. The Spanish corporatist model inspired economic policy during the dictatorship of Primo de Rivera (1923–30) and Franco's fledgling dictatorship, where its influence united with economic planning, on which the economic policy in the first years of Franco's autarky was based.

In any case, corporatism in Spain had neither the significance nor the duration which it enjoyed in Portugal, and had no level of analysis, being reduced to plans which were ostensibly superior to capitalism and socialism, but in which the economy was completely subordinated to political ends.

On the other hand, the marginalist revolution was almost unknown in Spain during the 19th century and for a good part of the 20th, with the exception of some teachers in the engineering schools. Later, thanks to the creation of the Faculty of Political and Economic Sciences at the University of Madrid (1943), neoclassical microeconomics began to spread in university lecture halls. Equally influential was the arrival in Spain in 1941 of Heinrich Von Stackelberg, an economist of German origin with a mathematical education. He was mainly responsible for training future university lecturers in economic theory, participating in courses at the Faculty of Economic Sciences, created only three years before his death. The delay in the reception of neoclassical thought is attributable to the economists' poor mathematical preparation and perhaps to the neo-historicist focus which prevailed in Spain in the first decades of the 20th century.

The reception of Keynesianism in Spain

Knowledge of the work of Keynes was hindered by the fact that the *General Theory* appeared only months before the outbreak of the Spanish civil war, though previously the worldwide dissemination of *The Economic Consequences of the Peace* had allowed Keynesian ideas to be known through the press, and Keynes himself had visited Spain in 1930. In the intense debate held on monetary questions in Spain in the period 1927–30, no influence can be detected of the ideas contained in the *Tract on Monetary Reform*. With the benefit of hindsight, during Keynes's visit, agreement can be detected between Keynes's ideas and those of the Spanish economists who had rejected the revaluation of the peseta because of its repercussions on production and employment, especially with the *Dictamen sobre la implantación del patron oro* drawn up in 1929. In the years leading up to the Spanish civil war,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Keynesian ideas were interpreted in the context of attempts to escape from the depression through economic policy which did not conform with the orthodox line: the promotion of public works at the cost of a budget deficit, the abandonment of the gold standard and the granting of credit facilities, in the framework of moderate protectionism.

The later reception of the *General Theory* was complicated, not only because of the civil war but also because the tradition of a more neo-historicist slant had prevailed in the academic world until then. There was first ‘underground’ dissemination among a depleted group of Spanish economists during the civil war and the first few years of the Second World War. Later, however, in the 1940s, there was a wide dissemination of Keynesian literature within the context of wider discussion of post-war reconstruction policies and the framework necessary for stability. There was a varied reception to Keynesian thought by some economists such as Manuel de Torres, Emilio de Figueroa, or Joan Sardá, who in some cases considered the primary formulations of the neoclassical synthesis, while others combined the ideas of the *General Theory* with some contributions from Austrian and Scandinavian economists.

Another group of experienced economists, very knowledgeable of Keynesian literature, such as Germán Bernácer and Luís Olariaga, harboured reservations about Keynesian ideas, especially on an analytical level.

On the economic and political level, the interpretation is more complex due to the political situation during General Franco's long dictatorship. In this context, Keynesian proposals were used by the regime itself and by some of its principal advisors, such as Higinio Paris Eguilaz, to encourage employment policies and economic stability. This meant misuse on the part of an official authority which used a ‘bastardized’ Keynesianism identified with systematic interventionism, resulting in a confusing clash between Keynesianism and corporatist dirigisme, in which Franco's ‘autarky’ was propped up by an interventionist recipe of price and wage controls, rationing, credit manipulation, and an import substitution policy, rounded off by a system of licences and official authorizations. This model distanced itself completely from Keynesianism since, in reality, its economic strategy relied upon a revision of the assisted capitalist model which had begun in the time of the dictatorship of Primo de Rivera (1923–30), of clearly corporatist influence, and was now being updated in the light of pre-war national socialist experiences.

Together with these, there was a group of expert economists which used Keynesianism for establishing the basis of a more rational economy, and it was precisely from this group that the main criticism came of the autarkic regime, of its disastrous and suffocating interventionism and of its self-destructive effects. Within this new generation of economists educated at the new Facultad de Ciencias Económicas there were also subtle differences, among those who embraced the postures of the Ordo group and the school of Freiburg (liberals linked to research institutes, tolerated, but critical of Franco's regime), as well as some university lecturers who defended mixed viewpoints in which well-received Hayekian liberalism was tempered from viewpoints which appealed to a ‘liberating’ intervention along the lines of Röpke or Eucken.

The 1950s saw a considerable expansion of Keynesianism through debate by several disciples of Professor Manuel de Torres, such as Enrique Fuentes Quintana, Manuel Varela and Emilio de Figueroa on the applicability in Spain of the ideas contained in the *General Theory*; the debate was useful for advancing Keynesian ideas on unemployment towards more relevant problems which ran deep in Spanish economic backwardness, such as the limitations of Spanish industry and the consequences of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
very restricted international trade. It also tested awareness of the need to maintain applicable economic policy measures, with better knowledge of economic macro-magnitude principles, which led to the reactivation of the first studies in national accounting and the first estimations on the national revenue and its provincial distribution. In 1954, studies for the drafting of the first input–output table for the Spanish economy began under the direction of Professors Manuel de Torres and Valentín Andrés Álvarez who, with the assistance of the Italian economist Vera Cao-Pinna and the Instituto per la Cogiuntura di Roma, allowed the first table to be drawn up in 1958, with data referring to the Spanish economy of 1954.

Resistance to this expansion of Keynesianism was less than in other countries; abandonment of the neo-historicist line helped, of a realistic nature which had dominated the teaching of economics up to the Spanish civil war, along with the fact that, unlike in Portugal, economic corporatism did not attain an influence beyond the first moments of Franco's regime.

The clearest influence on Keynesian economic policy was the Stabilization Plan (1959), which helped to bring about a thawing in Franco's autarky, which by then had become untenable. A paradox existed in that, in reality, this late application of the Keynesian programme consisted of a ‘cooling off’ and restriction on the economy, with a view to an opening up of the Spanish economy to foreign markets.

Later on, a more mature assimilation of Keynesian ideas came about, the main focus of which was the Bank of Spain's Research Department, where an econometric model was developed for the Spanish economy on a Keynesian base. Later, recognition of the effectiveness of monetary policy to combat inflation and macroeconomic imbalance contributed in an indirect way to the favourable penetration of the ideas of Milton Friedman, which were almost unknown in Spain until the 1980s.

Concluding remarks

Unlike in other countries, the academic institutionalization of economic studies, begun in the 19th century, did not gel into the creation of specific centres for the study of economic science, which until then had been taught in law schools, and later in schools of commerce at non-graduate level.

From the beginning of the 20th century there were several attempts to create a specific centre, which were not successful until 1943, with the creation of the School of Political, Economic and Commercial Sciences at the University of Madrid. This was the main factor which led to the consolidation of the academic study of economics in Spain, and to the fact that economic theory, first microeconomics and later macroeconomics, became known in the Spanish academic world. From these same lecture halls later in the 1960s Keynesian ideas spread, in their diverse formulations, initially the neoclassical synthesis, and later on in the 1960s and 1970s Keynesian macroeconomics in its truest interpretation and its later developments, as with the monetarist theories and those of the new macroeconomics.

Assimilation of economic theory and the training of economists constituted an essential step in that the policies of stabilization and the opening up to foreign markets, first within Franco's regime in a phase in which Spain, once the autarky had been exhausted and abandoned, began to be incorporated into the main international economic organizations and later on during the democratic transition. These policies allowed change from a closed economic model to an open economy, from which integration into the international economic panorama became possible.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Bernácer, Germán
- British classical economics
- cameralism
- Catholic economic thought
- France, economics in (before 1870)
- Gournay, Jacques Claude Marie Vincent, Marquis de
- heterodox economics
- Historical School, German
- Italy, economics in
- mercantilism
- physiocracy
- Portugal, economics in
- religion, economics of
- scholastic economics
- Stackelberg, Heinrich von

Bibliography

Almenar, S. 2005. Chair, tribune and seat: Spanish economists in parliament (1844–1923). An exploration. In *Economists in Parliament in the Liberal Age (1848–1920)*, ed. M. Augello and M. Guidi. Aldershot: Ashgate.

Almenar, S. and Llombart, V. 2001. Spanish societies, academies and economic debating societies. In *The Spread of Political Economy and the Professionalisation of Economists*, ed. M. Augello and M. Guidi. London and New York: Routledge.

Almenar, S. and Llombart, V. 2007. *History of Spanish Economic Thought*. New York and London: Routledge.

Almenar, S. and Sánchez Hormigo, A. 2004. Bibliografía de historia del pensamiento económico en España. In *Economía y economistas españoles*, vol. 9, ed. E. Fuentes Quintana. Barcelona: Galaxia Gutenberg-Funcas.

Azpilcueta, Martín de. 1556. *Comentario resolutorio de cambios*, ed. J.M. Pérez Prendes and L. Pereña. Madrid: CSIC, 1965.

Campomanes, Conde de. 1774–1775. *Discurso sobre el fomento de la industria popular y Discurso sobre la educación popular de los artesanos y su fomento*, ed. J. Reeder. Madrid: Instituto de Estudios Fiscales, 1975.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Cardoso, J.L. and Lluch, E. 1999. Las teorías económicas contempladas a través de una óptica nacional. In *Economía y economistas españoles*, vol. 1, ed. E. Fuentes Quintana. Barcelona: Galaxia Gutenberg-Funcas.

Colmeiro, M. 1861. *Biblioteca de economistas españoles de los siglos XVI, XVII y XVIII*, ed. L. Perdices, J. Reeder. Madrid: IEF, Fundación ICO y Real Academia de Ciencias Morales y Políticas, 2005.

Estapé, F. 1990. *Introducción al pensamiento económico. Una perspectiva española*. Madrid: Espasa Calpe.

Flórez Estrada, A. 1828. *Curso de Economía Política*, ed. S. Almenar, 2 vols. Madrid: Instituto de Estudios Fiscales, 1980.

Fuentes Quintana, E., ed. 1999–2004. *Economía y economistas españoles*, 9 vols. Barcelona: Galaxia Gutenberg-Funcas.

Grice-Hutchinson, M. 1993. *Economic thought in Spain: Selected Essays*. Aldershot: Edward Elgar.

Jovellanos, G.M. de. 1795. Informe en el Expediente de Ley Agraria. In *G.M. de Jovellanos, Escritos económicos*, ed. V. Llombart. Madrid: I.E.F., Fundación ICO, Real Academia de Ciencias Morales y Políticas, 2000.

Llombart, V. 2004. Traducciones españolas de economía política: 1700–1812: catálogo bibliográfico y una nueva perspectiva. *Cyber Review of Modern Historiography*, No. 9. Firenze: Firenze University Press.

Lluch, E. 1973. *El pensament econòmic a Catalunya (1760–1840). Els orígens ideològics del proteccionisme i la pressa de consciència de la burgesia catalana*. Barcelona: Edicions 62.

Lluch, E. 1980. Sobre la historia nacional del pensamiento económico. In *A. Flórez Estrada, Curso de Economía*, ed. S. Almenar. Madrid: I.E.F.

Lluch, E. 1999. Las historias nacionales del pensamiento económico y España. In *Economía y economistas españoles*, vol. 1, ed. E. Fuentes Quintana. Barcelona: Galaxia Gutenberg-Funcas.

Martín Rodríguez, M. 1989. La institucionalización de los estudios de Economía Política en la Universidad Española (1784–1857). In *Marqués de Valle Santoro, Elementos de Economía Política con aplicación particular a España*. Madrid: I.E.F.

Ortiz, L.E. 1558. *Memorial del contador Luis de Ortiz a Felipe II*, ed. J. Larraz. Madrid: Instituto de

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
España, 1970.

Perdices, L. and Reeder, J. 2003. *Diccionario de Pensamiento Económico en España (1500–2000)*. Madrid: Síntesis, Fundación ICO.

Perrotta, C. 1993. Early Spanish mercantilism: the first analysis of underdevelopment. In *Mercantilism: The Shaping of an Economic Language*, ed. L. Magnusson. London: Routledge.

Schwartz, P. 2000. The Wealth of Nations censured. Early translations in Spain. In *Contributions to the History of Economic Thought in. Essays in Honour of R.D. Collison Black*, ed. E. Murphy R. Prendergast. London: Routledge.

Smith, R.S. 1968. English economic thought in Spain, 1776–1848. In *The Transfer of Ideas: Historical Essays*, ed. C.D. Godwin et al. Durham, NC: South Atlantic Quarterly.

Smith, R.S. 1971. Spanish mercantilism: a hardy perennial. *Southern Economic Journal* 38, 1–11.

Uztáriz, Jerónimo de. 1724. *Theórica y práctica de comercio y de marina*. Madrid: Aguilar, 1968.

Velarde, J. 1974. *Introducción a la historia del pensamiento económico español en el siglo XX*. Madrid: Editora Nacional.

How to cite this article

Hormigo, Alfonso Sánchez. "Spain, economics in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000526> doi:10.1057/9780230226203.1581

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spatial econometrics

Timothy G. Conley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Spatial econometrics is concerned with modelling dependent observations indexed by points in a space. Complicated patterns of interdependence can be parsimoniously described in terms of these points' locations. Covariances between observations, for example, can be modelled as functions of their distances. Index spaces are *not* limited to the physical space or times inhabited by economic agents and can be as abstract as required by the economics of the application. This entry discusses the use of generalized method of moments and other common estimators with spatial data, as well as simultaneous equation methods specialized to certain types of spatial data.

Keywords

data generation processes; generalized method of moments; heteroskedasticity and autocovariance; index space; inference; interactions-based models; kernels; locations/distances; maximum likelihood estimators; nonparametric estimators; simultaneous equations models; simultaneous spatial autoregression; spatial correlation; spatial econometrics; spatial weights matrix; spectral density; spectral methods; time series models; unobservable variables

Article

Spatial econometrics is concerned with models for dependent observations indexed by points in a metric space or nodes in a graph. The key idea is that a set of locations can characterize the joint dependence between their corresponding observations. Locations provide a structure analogous to that provided by the time index in time series models. For example, near observations may be highly correlated but, as distance between observations grows, they approach independence. However, while time series are ordered in a single dimension, spatial processes are almost always indexed in more than one dimension and not ordered. Even small increases in the dimension of the indexing space permit large increases in the allowable patterns of interdependence between observations. The primary benefit of this modelling

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
strategy is that complicated patterns of interdependence across sets of observations can be parsimoniously described in terms of relatively simple and estimable functions of objects like the distances between them.

The fundamental ingredients in any spatial model are the index space and locations for the observations. In contrast to the typical time series situation where calendar observation times are natural indices and immediately available, the researcher will often need to decide upon an index space and acquire measurements of locations/distances. The role of measured locations/distances is to characterize the interdependence between economic agents' variables, particularly those that are unobservable – for example, regression error terms. The appropriate index space depends on the economic application, and its choice is inherently a judgement call by the researcher. Fortunately, the economics of the application often provide considerable guidance and the index space/metric(s) can be tailored to promote a good fit between the economic model and the empirical work. For example, when local spillovers or competition are the central economic features, obvious candidate metrics are measures of transaction/travel costs limiting the range of the spillovers or competition. If productivity measurement were the focus, distances between observed firms or sectors could be based upon economic mechanisms that might generate co-movement in productivity – for example, measures of similarity between production technologies. Index spaces are not limited to the physical space or times inhabited by the agents and can be as abstract as required by the economics of the application.

Locations/distances are almost never perfectly measured, and this puts a premium on empirical methods that are robust to their mismeasurement. Even if the ideal metric were physical distance, usually agents' physical locations are imprecise, known only within an area – for example, census tract or county. At best this will result in imprecise distance information between agents, and if inter-agent distances are approximated with measurements based on these areas, such as distance between centroids, errors result. Moreover, in the great majority of applications the ideal metric is *not* physical distance and must be either estimated or approximated, inevitably resulting in some amount of measurement error.

There are two main approaches to modelling a spatial data generation process (DGP). The first is to model explicitly a population residing in an underlying metric space and the process of drawing an observed sample from this population. The second is to model the data-set of observed agents' outcomes as being determined by a system of simultaneous equations. In the remainder of this article, I discuss each of these approaches in turn for the simplest case of cross-sectional data. It is important to note, however, that the methods in the following section – covariance and generalized method of moments (GMM) estimation, spatial correlation robust inference – can be directly applied to panel or repeated cross-section data by simply including time as one of the components in the spatial index (*s* defined below). Most if not all cluster/group effect models can be considered a special case of spatial models with a binary metric indicating common group/cluster membership. See Wooldridge (2003) for an excellent review of these models. I do not discuss them here because their associated empirical techniques and sampling schemes do not translate well to more general spatial models. I conclude with a brief discussion of areas of econometrics where links to spatial econometrics are perhaps underappreciated.

1 Models for samples from a population

This section discusses spatial econometric models that view the data as being a sample from some

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
arbitrarily large population (see, for example, Conley, 1999, for a more formal treatment). The population of individuals is assumed to reside in some metric space, typically $\Re^k$ or an integer lattice, with each individual *i* located at a point s_i.

The basic model of dependence characterizes dependence between agents' random variables via their locations. The data are assumed to be weakly dependent (perhaps after de-trending). (Andrews, 2005, is an important exception that explicitly considers strong cross sectional dependence arising from common shocks.) If two agents' locations s_i and s_j are close, then their random variables φ_{s_i} and φ_{s_j} may be highly dependent. As the distance between s_i and s_j grows large, φ_{s_i} and φ_{s_j} become essentially independent. Notions of weak dependence can be formalized in essentially the same manner as for time series, for example, with mixing coefficients. Under regularity conditions limiting the strength of dependence, laws of large numbers and central limit results can be obtained for properly normalized averages of ϕ_s. See, for example, Takahata (1983) or Bolthausen (1982). These approximations almost always use what is called an increasing domain approach to limits, with the corresponding thought experiment being that, as the sample size grows, an envelope containing the locations would be growing without bound.

When one works within this framework, it is often useful to approach an empirical problem in two steps. First, decide upon a (small) set of metrics based on the economics of the application, and then consider statistical modelling of dependence as a function of the metrics. It is much easier to conduct statistical modelling given a metric than to try to simultaneously vary both the model specification and the metric itself.

Statistics that describe spatial correlation patterns are simple to construct. Any statistic relating co-variation of φ_{s_i} and φ_{s_j} to some measure of their proximity could be used to characterize patterns in dependence. Classic references are Moran (1950) and Geary (1954), and the text by Cliff and Ord (1981) contains a good treatment. One useful approach is based on nonparametric estimation of a covariance function (see for example Conley and Topa, 2002, or Conley and Ligon, 2002). The ϕ_s process is covariance stationary if its expectation is the same at all locations and $cov(\varphi_s, \varphi_{s+h})$ depends only on the relative displacement *h*. For high-dimensional *h*, it is useful to consider a special case called isotropy where covariances depend only on the length of *h*; covariance depends upon distance but not direction. Take an isotropic covariance stationary ϕ_s with expectation zero for simplicity. Its covariance function f can be expressed in a regression equation involving distances $d_{i,j}$:

$$E(\varphi_{s_i}\varphi_{s_j}|s_i, s_j) = f(d_{i,j}). \tag{1}$$

The function *f* in eq. (1) can be estimated parametrically or, as is particularly useful in preliminary data analysis, via a nonparametric regression of $\varphi_{s_i}\varphi_{s_j}$ on $d_{i,j}$. Investigation of correlation patterns when

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

there is more than one candidate metric can by done by simply letting *f* be a function of more than one distance measure.

In cases where ϕ_s is not isotropic or non-stationary, *f* can still be interpretable as a measure of average co-movement. If the process is covariance stationary but not isotropic, an estimate at a given distance d_0, call it $\hat{f}(d_0)$, will converge to a weighted average of $cov(\varphi_s, \varphi_{s+h})$ for displacements *h* that have length d_0. The relative weights of different directions *h* will depend on their frequency of sampling. An analogous interpretation of *f* is available when ϕ_s is non-stationary, $cov(\varphi_s, \varphi_{s+h})$ depends on *s*, but still weakly dependent with averages of $cov(\varphi_s, \varphi_{s+h})$ across *s* remaining convergent. In this case, $\hat{f}(d_0)$ will converge to a weighted average of $cov(\varphi_s, \varphi_{s+h})$ across those *h* with length d_0 and across all *s*. Typically, this is still a valuable measure of co-movement. If non-stationarity is suspected, it is also very useful to construct localized versions of measures of spatial correlation. Localized *f* estimates for subregions of the locations can easily be constructed by just confining the observations used to estimate (1); see Anselin (1995) for extensive treatment of localized versions of Moran (1950) and Geary (1954) measures of spatial correlation.

Estimates of *f* can also be viewed directly as test statistics for the null hypothesis of independence. Under the null hypothesis of independence, the sampling distribution of an *f* estimator can be approximated and compared to the realized value of *f* estimates to test the hypothesis of independence. Such tests for independence remain valid even with measurement errors in distances (see Conley and Ligon, 2002).

Parameter estimation via moment conditions

In most econometric applications, the parameters of interest can be estimated using GMM. GMM estimation with weakly spatially dependent data is straightforward, and the spatial dependence is relevant for inference and efficiency (see Conley, 1999). Consider instrumental variables (IV) estimation in the linear model with outcome y_{s_i}, regressors x_{s_i}, and instruments z_{s_i}

$$y_{s_i} = x'_{s_i}\beta + u_{s_i}$$
and
$$Ez_{s_i}u_{s_i} = 0$$
(2)

The IV estimator is identified by the moment condition (2): that the instruments are not correlated with the error term. Since this is a moment condition with respect to the marginal distribution of the data across agents, it is valid with or without spatial dependence. The familiar solution remains: $\beta = \left(Ez_{s_i}x'_{s_i}\right)^{-1} Ez_{s_i}y_{s_i}$. Consistent estimates of β can be obtained using sample averages to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

approximate these expectations since a law of large numbers applies to weakly dependent spatial data. Thus, the usual IV estimator, $\hat{\beta}_N = (\frac{1}{N}\Sigma_{i=1}^{N} z_{s_i} x'_{s_i})^{-1} \frac{1}{N}\Sigma_{i=1}^{N} z_{s_i} y_{s_i}$, remains consistent with weak spatial dependence. It is of course feasible to construct $\hat{\beta}_N$ without any knowledge of locations/ distances, so it is trivially robust to measurement error in them. The impact of such spatial dependence is only upon inference, getting correct standard errors or testing.

This logic carries over to any GMM estimator of a parameter θ_0 that is identified from a moment condition involving a (potentially nonlinear) function g:

$$Eg(\varphi_{s_i}, \theta_0) = 0.$$

The majority of econometric models with nonlinearity or limited dependent variables can be estimated via some choice for *g*. Under mild regularity conditions, θ_0 can be consistently estimated by minimum distance methods using $\frac{1}{N}\Sigma_{i=1}^{N} g(\varphi_{s_i}, \cdot)$ to approximate $Eg(\varphi_{s_i}, \cdot)$. A GMM estimator is the argument minimizing the criterion function, $J_N(\theta)$, which takes the same form as with time series or independent data:

$$J_N(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N} g(\varphi_{s_i}, \theta)\right]' \Omega \left[\frac{1}{N}\sum_{i=1}^{N} g(\varphi_{s_i}, \theta)\right],$$

where Ω is some positive definite matrix. Just as for the time series case (Hansen, 1982), an efficient GMM estimator can be obtained by taking Ω to be a consistent estimator of the limiting variance-covariance matrix of $\frac{1}{\sqrt{N}}\Sigma_{i=1}^{N} g(\varphi_{s_i}, \theta_0)$, whose form depends on the spatial covariance structure of the data. One such covariance matrix estimator is described in the following subsection.

Inference

The usual approach to inference using large sample approximations can be employed with weakly spatially dependent data. Returning to the IV model, the typical approximation for the distribution for $\hat{\beta}_N$ is based on the expression:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\sqrt{N}(\hat{\beta}_N - \beta) = \left(\frac{1}{N}\sum_{i=1}^{N} z_{s_i} x'_{s_i}\right)^{-1}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} z_{s_i} u_{s_i}\right].$$

(3)

Under regularity conditions, the first term in the product converges to the matrix $Ez_{s_i}x'_{s_i}$. The second term in brackets has a limiting normal distribution:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} z_{s_i} u_{s_i} \Rightarrow N(0, V)$$

(4)

where V is the limiting variance-covariance matrix of

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} z_{s_i} u_{s_i}.$$

V contains terms of the form $Ez_{s_i}u_{s_i}z'_{s_i}u_{s_i}$ and cross-covariance terms, $Ez_{s_i}u_{s_i}z'_{s_j}u_{s_j}$, that will be non-zero for at least some i,j pairs. With weak dependence, the covariance between variables indexed i and j will eventually vanish as the distance between s_i and s_j grows.

In some cases, V has a nice form. For example, suppose locations were on an integer lattice, Z^k; samples consist of all integer coordinates in a region (assumed to grow as $N \to \infty$); and variables are covariance stationary. In this case, V can be expressed as an infinite sum of a covariance function:

$$V = \sum_{h \in Z^k} Cov(z_s u_s, z_{s+h} u_{s+h}).$$

(5)

With integer locations on the line, this expression coincides with its analog for covariance stationary time series.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

With a consistent estimate of *V*, call it $\hat{V}_N$, the approximate distribution implied by (3) and (4) can be used for inference:

$$\sqrt{N}(\hat{\beta}_N - \beta) \overset{Approx}{\sim} N\left(0, \left(\frac{1}{N}\sum_{i=1}^{N} z_{s_i} x'_{s_i}\right)^{-1} \hat{V}_N \left(\frac{1}{N}\sum_{i=1}^{N} z_{s_i} x'_{s_i}\right)^{-1}\right).$$

There are of course many ways *V* could be estimated. If it were assumed to have a parametric form, for example, by parameterizing the covariance function in (5), then consistent estimates could be obtained by GMM. Perhaps the most popular approach has been nonparametric estimation of *V* following Conley (1996; 1999). This approach is analogous to time series heteroskedasticity and autocovariance (HAC) consistent covariance matrix estimation, and can be viewed as a smoothed periodogram spectral density estimator. (See Priestley, 1981, for a discussion of the vast literature on spectral methods in time series and some extensions to spatial processes. Spectral methods for spatial processes date back to at least the 1950s; for example, Whittle, 1954; Bartlett, 1955; Grenander and Rosenblatt, 1957; Priestley, 1964). With the use of residuals $\hat{u}_{s_i}$ to approximate u_{s_i}, *V* can be estimated as a weighted sum of cross products $z_{s_i}\hat{u}_{s_i} z'_{s_j}\hat{u}_{s_j}$:

$$\hat{V}_N = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} K_N(s_i, s_j) \cdot z_{s_i}\hat{u}_{s_i} z'_{s_j}\hat{u}_{s_j}.$$

$K_N(\cdot, \cdot)$ is a kernel used to weight pairs of observations, with close observations receiving a weight near 1 and those far apart receiving weights near zero. $K_N(s_i,s_j)$ is commonly specified to be uniform kernel that is 1 if s_i and s_j are within a cut-off distance and zero otherwise. (This indicator function K_N is not guaranteed to provide positive definite (PD) covariance matrix estimates; however, this is very rarely a problem in practice. PD estimates can be insured by an alternate choice of kernel; see Conley, 1999.) $\hat{V}_N$ will be consistent if as $N \to \infty$, $K_N(s, s+h) \to 1$ for any given displacement *h*, but slowly enough so that the variance of $\hat{V}_N$ collapses to zero.

In practice, this estimator will require a decision about the exact form of $K_N(\cdot, \cdot)$. With a uniform kernel, this is just an operational definition of which observations are near and which are far. A conservative distinction between near and far observations can be made even with multiple candidate metrics by assigning a far classification only when all metrics agree. There is no need for the data to be covariance stationary, nor is the specific sampling framework here necessary. Analogous HAC methods can be applied to weakly dependent but non-stationary data, including that generated by simultaneous

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

equations DGPs like those discussed in the following section 2 (see Pinkse, Slade and Brett, 2002; Kelejian and Prucha, 2007).

The main reason nonparametric estimators like $\hat{V}_N$ are often preferred to parametric models for V is their robustness to measurement errors in locations/distances. Parametric V estimators are generally inconsistent with such errors, while $\hat{V}_N$ remains consistent under mild conditions. $\hat{V}_N$ can be consistent with spatially correlated and even endogenous errors; a sufficient condition is simply that they be bounded (Conley, 1999). With location/distance errors, the weight assigned to pair *i*,*j* can be altered relative to the weight $K_N(\cdot,\cdot)$ would assign with exact locations. But $\hat{V}_N$ remains consistent, because the altered weights will still satisfy the necessary conditions for consistency of $\hat{V}_N$: the weight on observations at any true displacement will still converge to 1, slowly enough. Even when working with parametric models of *V*, $\hat{V}_N$ remains of interest since the discrepancy between it and a parametric *V* estimator can provide a useful joint test for the absence of location/distance errors and proper parametric specification (Conley and Molinari, 2007).

More important than $\hat{V}_N$ remaining consistent is its robustness in practice to moderate amounts of location error. Consider the impact of introducing location error for $\hat{V}_N$ defined with a kernel $K_N(s_i,s_j)$ equal to 1 if s_i and s_j are within L_N units, and zero otherwise. If the magnitude of measurement error is moderate relative to L_N, then the weights on most pairs of points would be unchanged if erroneously measured locations were used in place of true locations. Changes in weights occur only for those points whose true distance is near enough to the cut-off L_N that location errors result in measured and true distances being on opposite sides of L_N. With moderate amounts of location error, these pairs of observations with true distance near L_N will usually not be a large portion of the sample, so $\hat{V}_N$ will tend to be close to its value with true locations. Similar results obtain for other kernels as weights arising from moderately mismeasured locations remain close to those received with true locations (see Conley and Molinari, 2007).

2 Population simultaneous equation models

The second approach to modelling spatial data is with a simultaneous equations model, most directly interpretable as a model for a *population* of *N* agents. This approach explicitly specifies a joint model for the population, in contrast to typical models in Section 1, where the joint determination of outcomes in the population is not explicitly treated. These simultaneous equation models are directly applicable to situations where the entire population of agents is observed, like all US states or counties or even all firms in an industry. Typical applications include studies of games being played among these agents or of spillovers across agents; see, for example, Case, Hines and Rosen (1993) and Pinkse, Slade and Brett (2002).

The most common type of model is a simultaneous spatial autoregression (SAR). Its simplest formulation for an $N \times 1$ outcome vector Y_N is:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$Y_N = \rho W_N Y_N + \varepsilon_N, \tag{6}$$

with scalar parameter ρ and IID shocks ε_N (typically Gaussian). The $N \times N$ matrix W_N is commonly referred to as a 'spatial weights' matrix and assumed known. W_N has zero main-diagonal elements, and its off-diagonal elements reflect some notion of interaction. Typical W_N contain (i,j) elements that are non-zero only if locations i and j are adjacent on a graph or elements inversely related to distances between locations. W_N is usually row-standardized so that its rows sum to 1. The parameter space is restricted so that $(I - \rho W_N)^{-1}$ exists and the model has reduced form:

$$Y_N = (I - \rho W_N)^{-1} \varepsilon_N.$$

Thus Y_N is a linear combination of the ε_N IID shocks. Though SAR models are finite (usually) irregular lattice models, their origins date to at least the infinite regular lattice models of Whittle (1954). Textbook treatment of SARs can be found in Anselin (1988).

Typical specifications for W_N imply a great deal of heterogeneity across observations. Variances will typically differ across the elements of Y_N by construction unless $\rho = 0$. Unconditional heteroskedasticity is thus coupled with spatial dependence. Covariances between pairs of agents will differ in patterns that are of course determined by W_N but will depend on the entire structure of this matrix and will not generally follow a simple pattern in terms of some metric. For example, with W_N defined based upon a graph, covariance between agents i and j will not be a function of their graph distance, though it can be characterized in terms of properties of the graph (Martellosio, 2004). A given graph will 'hard-wire' patterns in correlations across agents. For example Wall (2004) notes, with model (6) for US states with W_N based on adjacency, that Missouri and Tennessee are constrained to be the least spatially correlated states, while relative correlations between other pairs of states change depending on ρ. Even with a more flexible parameterization – for example, specifying the elements of W_N to be flexible functions of distance, as in Pinkse, Slade and Brett (2002) – there is still a tendency for heterogeneity in the implied joint distribution to be difficult to anticipate. While this complicates their use as statistical models, as discussed below, it is in my view likely to be a desirable property in a structural model. For example, if the model's joint distribution is to be taken seriously as capturing equilibrium outcomes for N asymmetric agents playing a game, then one would expect 'hard-wired' heterogeneity depending on the exact structure of the game.

Though the population of agents is observed, large sample approximations taking limits as $N \to \infty$ are still potentially useful. However, the requisite limit theorems technically differ from those referenced in Section 1. Since the DGP is changing as N grows, triangular array limit results are required. Consistency

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and distribution results for Gaussian maximum likelihood estimators (MLEs) with spatial dependence have existed at least since Mardia and Marshall (1984). An extensive set of SAR limiting distribution results is obtained by Lee (2004a) for likelihood-based estimators under a variety of conditions upon 'spatial weights' matrices like W_N. Quite useful limit theorem results can also be found in Kelejian and Prucha (2001). Correct specification of W_N is essential for these results, as SAR estimators will generally be inconsistent when there is measurement error in locations/distances used to specify this matrix (the same holds true for other parametric models of dependence structure).

A great deal of the literature has focused on computational issues involving MLEs. Non-trivial W_N matrices make computation of normalizing constants challenging. Substantial progress has been made in techniques for computing MLEs by exploiting sparseness or specific structure of 'spatial weights' matrices and re-parameterization to facilitate computation (see Pace and Barry, 1997; Barry and Pace, 1999; LeSage and Pace, 2007). These numerical techniques allow likelihood-based inference for even very large data-sets in certain applications or specifications. It is also feasible, of course, to estimate SAR parameters without computing MLEs, by using only a subset of the implications of the model to obtain method of moments estimates (see Kelejian and Prucha, 1999, and Lee, 2007, and subsequent work by these authors). This literature has been successful in addressing most computational issues with SAR models.

The key remaining difficulties in using SAR models are in terms of model specification and interpretation. Even for the simplest SAR model (6), it is hard to characterize implications of different ρ without explicitly calculating their implied joint distributions. The parameter ρ is not a simple correlation coefficient; in general it is not comparable across different specifications for W_N. In my experience, explicit calculations of descriptive measures of the implied joint distributions for many different ρ are required to understand whether varying this parameter will trace out a useful path through the space of joint distributions.

Unless one has access to virtually complete data on a population, SAR models are very difficult to properly specify as structural models. To take an optimistic case, suppose model (6) with Gaussian ε applied to a population of N agents, but a subset of agents were sampled. The likelihood of such a sample is well-defined, and in principle its form could be found by integrating out all the unobserved variables. But this calculation requires the exact form of W_N, which depends on all the unobserved agents, a full structure which will rarely be observed if only a small fraction of the agents are sampled. Proper specification of W_N is perhaps feasible only if the vast majority of the population is sampled – for example, if only a few states or counties are missing.

Even with complete data on a population, SARs are difficult to specify because they are inherently fragile. Changing a single element of W_N will in general influence the entire joint distribution of Y and it is difficult to intuitively understand the impact of a given change in W_N. Increasing flexibility by parameterizing W_N by taking its elements to be a series expansion in distance(s), as in Pinkse, Slade and Brett (2002), is of limited help. There remains only an indirect link between the series expansion and the implied joint distribution. It is hard to see how much additional flexibility in, for example, allowed covariance structure is gained by adding another term in the expansion.

I think these difficulties should be considered a consequence of modelling a large-dimensional system of structural simultaneous equations rather than SAR-specific problems. It seems likely to be difficult to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
anticipate changes in equilibrium outcomes resulting from changes in individual agents' decision rules or best-response functions in any modelling framework. In my view, SARs remain a useful first step towards the goal of constructing good large-dimensional structural simultaneous equation models.
Of course SAR models need not be intended as structural models; they can be viewed, for instance, as tools to incorporate spatial dependence into forecasting models. A mis-specified but parsimonious model might still forecast well. However, the cumbersome relation between specification of 'spatial weights' and the implied joint distribution makes it hard to fashion parsimonious SAR models. This seems ample reason to avoid their use in forecasting. Directly specifying measures of dependence like covariances as a parsimonious function of distance appears far easier, even if the true DGP were an SAR.

3 Links between spatial econometrics and other areas

Work on interactions-based models has much in common with simultaneous equations-style spatial models (see Brock and Durlauf, 2001, for an extensive review). In these models, the behaviour of individuals is influenced by the characteristics and/or behaviour of others. Insofar as the relevant set of 'others' can be described in a spatial framework, they can be thought of as spatial econometric models. Much of this work is theory, taking the approach of specifying conditional probability measures to capture individuals' behaviours and then deriving the implied properties of the compatible joint distribution(s). Empirical work with these models has just begun and will share many of the same challenges described above; some can even be cast directly as SARs (see Lee, 2004b).
Spatial models are potentially very useful in modelling high-dimensional vector time series. Limited degrees of freedom with typical length samples require substantial restrictions upon the DGP to make progress. The potential of spatial models to capture complicated interdependence with a small number of parameters (given auxiliary location/distance information) makes them well suited for use in characterizing a variety of restrictions upon high-dimensional vector DGPs. Good examples of the benefits of spatial approaches to this type of time series modelling are Chen and Conley (2001), Giacomini and Granger (2004), and Bester (2005a; 2005b).

See Also

- generalized method of moments estimation
- heteroskedasticity and autocorrelation corrections
- social interactions (empirics)
- spectral analysis
- stratified and cluster sampling
- statistical mechanics

Bibliography

Andrews, D. 2005. Cross-section regression with common shocks. *Econometrica* 73, 1551–85.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Boston: Kluwer Academic Publishers.

Anselin, L. 1995. Local indicators of spatial association. *Geographical Analysis* 27, 93–115.

Barry, R. and Pace, R. 1999. A Monte Carlo estimator of the log determinant of large sparse matrices. *Linear Algebra and its Applications* 289, 41–54.

Bartlett, M. 1955. *An Introduction to Stochastic Processes*. Cambridge: Cambridge University Press.

Bester, C. 2005a. Random field and affine models for interest rates: an empirical comparison. Working paper, University of Chicago.

Bester, C. 2005b. Bond and option pricing in random field models. Working paper, University of Chicago.

Bolthausen, E. 1982. On the central limit theorem for stationary mixing random fields. *Annals of Probability* 10, 1047–50.

Brock, W. and Durlauf, S. 2001. Interactions-based models. In *Handbook of Econometrics* 5, ed. J. Heckman and Leamer. Amsterdam: North-Holland.

Case, A., Hines, J. and Rosen, H. 1993. Budget spillovers and fiscal policy interdependence: evidence from the states. *Journal of Public Economics* 52, 285–307.

Chen, X. and Conley, T. 2001. A new semiparametric spatial model for panel time series. *Journal of Econometrics* 105, 59–83.

Cliff, A. and Ord, J. 1981. *Spatial Processes*. London: Pion Limited.

Conley, T. 1996. Econometric modeling of cross-sectional dependence. Ph.D. thesis, University of Chicago.

Conley, T. 1999. GMM estimation with cross sectional dependence. *Journal of Econometrics* 92, 1–45.

Conley, T. and Ligon, E. 2002. Economic distance, spillovers, and cross country comparisons. *Journal of Economic Growth* 7, 157–87.

Conley, T. and Molinari, F. 2007. Spatial correlation robust inference with errors in location or distance. *Journal of Econometrics* 140(1), 76–96.

Conley, T. and Topa, G. 2002. Socio-economic distance and spatial patterns in unemployment. *Journal*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Applied Econometrics 17, 303–27.

Geary, R. 1954. The contiguity ratio and statistical mapping. *Incorporated Statistician* 5, 115–45.

Giacomini, F. and Granger, C. 2004. Aggregation of space–time processes. *Journal of Econometrics* 118, 7–26.

Grenander, U. and Rosenblatt, M. 1957. Some problems in estimating the spectrum of a time series. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 7, 77–93.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Kelejian, H. and Prucha, I. 1999. A Generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40, 509–33.

Kelejian, H. and Prucha, I. 2001. On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* 104, 219–57.

Kelejian, H. and Prucha, I. 2007. HAC estimation in a spatial framework. *Journal of Econometrics* 140 (1), 131–54.

Lee, L. 2004a. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72, 1899–925.

Lee, L. 2004b. Identification and estimation of spatial econometric models with group interactions, contextual factors and fixed effects. Working paper, Ohio State University.

Lee, L. 2007. GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics* 140(1), 155–89.

LeSage, J. and Pace, R. 2007. A matrix exponential spatial specification. *Journal of Econometrics* 140 (1), 190–214.

Mardia, K. and Marshall, R. 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* 71, 135–46.

Martellosio, F. 2004. The correlation structure of spatial autoregressions. Working paper, University of Southampton.

Moran, P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, 17–23.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Pace, R. and Barry, R. 1997. Quick computation of regressions with a spatially autoregressive dependent variable. *Geographical Analysis* 29, 232–47.

Pinkse, J., Slade, M. and Brett, C. 2002. Spatial price competition: a semiparametric approach. *Econometrica* 70, 1111–53.

Priestley, M. 1964. Analysis of two-dimensional processes with discontinous spectra. *Biometrika* 51, 195–217.

Priestley, M. 1981. *Spectral Analysis and Time Series*, 2 vols. New York: Academic Press.

Takahata, H. 1983. On the rates in the central limit theorem for weakly dependent random fields. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* 64, 445–56.

Wall, M. 2004. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference* 121, 311–24.

Whittle, P. 1954. On stationary processes on the plane. *Biometrika* 2, 434–49.

Wooldridge, J. 2003. Cluster-sample methods in applied econometrics. *American Economic Review* 93, 133–8.

How to cite this article

Conley, Timothy G. "spatial econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000450> doi:10.1057/9780230226203.1582

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spatial economics

Gilles Duranton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides a general overview of spatial economics, which covers location theory, spatial competition, and regional and urban economics. After a brief review of the main theoretical traditions, the fundamental role of non-convexities and imperfect competition is highlighted. The main challenges faced by theoretical and empirical research are also discussed, followed by a broader discussion of the relationship between this field of research and other sub-fields of economics and other disciplines.

Keywords

Alonso, W.; Debreu, G.; economic geography; Hecksher–Ohlin trade theory; Hotelling, H.; Krugman, P.; new economic geography; Non-convexity; Ricardo, D.; spatial economics; spatial impossibility theorem; systems of cities; Thünen, J. von; urban agglomeration; Weber, A.

Article

What is spatial economics? In a nutshell, spatial economics is concerned with the allocation of (scarce) resources over space and the location of economic activity. Depending on how this definition is read, the realm of spatial economics may be either extremely broad or rather narrow. On the one hand, economic activity has to take place somewhere so that spatial economics may be concerned with anything that economics is concerned about. On the other hand, location analysis focuses mostly on one economic question, namely, location choice. This is only one decision among a large number of economic decisions.

Which boundaries for spatial economics?

In practice, we can distinguish three sets of questions for which the importance of the spatial dimension is very different. Consider first the core questions of spatial economics. For example, why are there

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
cities? Why do some regions prosper while others do not? Why do we observe residential segregation? Why do firms from the same industry cluster? These are intrinsically 'spatial' questions, that is, questions in which the spatial dimension plays a dominant role. For instance, it would be difficult to speak meaningfully about the existence and growth of cities without some explicit consideration of space. Then there is a second group of issues concerned with, *inter alia*, the analysis of technological spillovers, the determinants of trade flows, or even the functioning of social networks. These issues all have a spatial dimension but its importance remains to be determined. Put differently, these are 'contested' issues between spatial economists and other economists. Finally, there is an extremely broad range of economic questions for which the spatial dimension is likely to be less important. For example, what are the drivers of investment? How important are firing costs to explain unemployment? What are the returns to education? To answer these questions, the main role of space is to provide possibly a major source of variation for empirical research. If, for instance, different regions of a country have different education systems with, say, different age limits for compulsory schooling, this variation can be used to produce some meaningful estimates of the returns to education. Even if we take this third set of questions to lie outside spatial economics, it nonetheless remains the case that spatial economics is concerned with a broad and heterogeneous set of questions, which involve very different spatial scales (from the very small to the very big) with imprecise boundaries. It is quite possible that this breadth and heterogeneity has hindered the development of the field. This is also what makes it interesting.

The centrality of spatial frictions

To see what makes spatial economics specific, it is useful to reformulate the question about its definition in the following way. Is spatial economics only about adding a spatial dimension? In the *Theory of Value*, Debreu (1959) answers affirmatively. A commodity is defined by all its characteristics including its location: the same good traded in different locations must be treated as different commodities. This 'answer' runs into serious problems, as pointed most clearly by Starrett (1974). Consider the extreme case of homogenous space where firms face the same convex production set, and consumer preferences are the same (and locally not satiated). Transporting commodities between locations is costly. Then the *spatial impossibility theorem* states that, with a finite number of locations, consumers, and firms, no equilibrium involves transportation. The intuition behind this result is straightforward: since economic activities are perfectly divisible and agents have no objective reason to distinguish between locations, each location operates in autarchy to save on transport costs. To avoid this very counterfactual result (no trade), one of the assumptions behind the spatial impossibility theorem needs to be relaxed. If one takes transport costs as an unavoidable fact of life, one must assume either some non-homogeneity of space or some non-convexity of production sets.

As shown by a first branch of trade theory, it is possible to develop a framework for spatial economics that builds only on local productivity differences. This approach was pioneered by Ricardo (1821), who developed a theory of land use based on relative fertility. This approach was later generalized to consider exogenous technological differences for all types of goods. A second branch of trade theory builds instead on differences in factor endowments over space. This is the so-called Hecksher–Ohlin theory of trade. The Ricardian and Hecksher–Ohlin approaches have led to sophisticated theories of location and trade that rely on the existence of (exogenous) 'comparative advantages' across locations.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
As shown by a large body of theoretical work in international trade, these approaches can be readily incorporated in the Arrow–Debreu framework. Although these approaches are central to the sister discipline of international trade, they played a much less important role in the development of spatial economic theory.

The pioneers

Instead, spatial economics has focused on the existence of non-convexities in the presence of transport costs. A key reason for this focus is that, although comparative advantage constitutes an appealing explanation for understanding trade flows at the world level, it provides at best a partial explanation for the location patterns of industries within countries, and it is at pains to explain major concentrations of population in large metropolitan areas. Instead, non-convexities in production or consumption seem to hold more promise for providing convincing answers to the core questions of spatial economics. The easiest way to model these non-convexities is to assume some indivisibility in a partial equilibrium framework. This type of work was pioneered by von Thünen (1826). In his model, a competitive farming sector bids for some homogenous land. The key non-convexity is that the output must be sold at a central market. With costly transport costs, farmers are willing to bid for land up to the point where the rent at a given distance from the market is equal to the gross revenue from the output minus the cost of non-land factors and minus transport costs. With a competitive land market, land goes to the highest bidder and the equilibrium typically involves concentric rings of specific land use around the central market.

To understand land use patterns in cities, Alonso (1964) developed an approach that was based on similar principles. His model again assumes a homogenous space, but replaces von Thünen's market by a central business district to which residents must commute at a cost to find work. This very parsimonious microeconomic model manages to replicate key stylized facts about land use and land prices within cities using rigorous microeconomic modelling. This is a showcase for the power of microeconomic approaches. It has spawned a large literature, which first extended the basic model in a number of directions and then went on to model multi-centric cities (for further details, see Fujita, 1989).

Independently of the ‘Thünen tradition’ that relies on an exogenous focal point for trade or production, another tradition was developed following Weber's (1909) work. Weber deals with the location problem of an indivisible and competitive plant that faces transport costs in order to ship its inputs from their sources and its outputs to their markets. With the use of essentially linear-programming techniques, the optimal location (which minimizes total transport costs) can be derived. Like Alonso's monocentric model, Weber's approach has been extended in many directions to consider, among others, more flexible production functions and the optimal location of public facilities.

Hotelling (1929) also explored the location problem faced by producers but went in a very different direction. His fundamental insight is that, because of indivisibilities, there will not be infinitely many producers at each point so that Weber's price-taking assumption is not tenable. With a small number of producers the location decision will involve more than minimizing transport costs since location also affects the competitive process. To make his point, he assumes evenly distributed consumers over a finite segment, each consuming one unit of a homogenous good. The market is served by two firms that need to choose a location and each customer patronizes the firm that minimizes the sum of the ‘mill’

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
price and shipping costs. At a first stage, firms choose a location and then compete in price. This deceptively simple game has received a lot of attention. Firms face a fundamental trade-off between central locations, which allow them to capture a larger share of the market, and more peripheral locations, which allow them to mitigate the intensity of competition. The resolution of this trade-off depends on the fine details of the assumptions being made (and particularly how an increase in the distance affects the price setting power of producers). Difficulties with existence of equilibrium have also turned out to be an important issue in this literature.

Modern approaches to spatial economic theory

Non-convexities in production lie at the heart of spatial economics. The literature discussed so far treats them as exogenous. It was not long before the literature started to worry about what these non-convexities were about. Nowhere was this worry stronger than in the ‘new urban economics’ literature, where Alonso's assumption of an exogenously given central business district quickly started to look very ad hoc. To understand central business districts or, more generally, why economic activity agglomerates, spatial economics had to provide microeconomic foundations for (local) increasing returns. Being able to generate increasing returns from plausible assumptions without leading to a degenerate market structure (for example, a monopoly firm for the entire economy) was a fundamental challenge for spatial economics. This was also true for many other fields such as industrial organization and international trade, where increasing returns were also needed to explain key stylized facts. Spatial economists were fortunate because they could draw on the insights provided by Adam Smith (1776) and Alfred Marshall (1890). Although Smith's argument about the division of labour being limited by the extent of the market pre-dates Marshall's *Principles* by more than a century, Marshall's ‘magic trilogy’ proved much more influential. Following Marshall, local increasing returns could arise because of knowledge spillovers, linkages between input suppliers and final producers, and thick local labour market interactions. What the modern literature on the microfoundations of increasing returns has achieved is a formalization of these insights (see Duranton and Puga, 2004, for an extensive review of this literature). Three main mechanisms can be used to generate local increasing returns: sharing, matching, and learning. Sharing mechanisms show how small non-convexities like small fixed costs paid by heterogeneous producers can be spread across larger quantities as market size increases and thus yield aggregate increasing returns. Matching mechanisms explore how larger markets might improve the probability and quality of matching. Finally, learning mechanisms explore the benefits of local size for the creation and diffusion of knowledge.

The second major problem faced by spatial economics is that many fundamental issues having to do with regional and urban development call for general equilibrium modelling. For instance, some cities can afford to specialize because they can trade with other cities. Hence, looking at one isolated city in the tradition of Thünen and Alonso may not be enough for some purposes. Similarly, the agglomeration of economic activity in core regions may occur because firms find larger markets there and because consumers find cheaper and more diverse supplies. These two forces are mutually reinforcing. This is the famous circular and cumulative causation mechanisms first emphasized by Myrdal (1957).

To model spatial economies, two main approaches came to dominate the intellectual landscape. The first follows the work of Henderson (1974) and is know as the ‘urban systems’ approach. In this type of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
framework, cities arise endogenously as the result of a trade-off between agglomeration economies and urban crowding. Both types of forces are modelled with the use of various microeconomic foundations. Cities can also trade with one another and the workers decide where to work. This strand of literature has been successful in replicating many stylized facts about urban systems, from the tendency of many cities to specialize while others diversify to their role in the innovation process.

Following the work of Krugman (1991), ‘the new economic geography’ is the second main general equilibrium approach in spatial economics. This approach puts trade costs (rather than commuting costs in urban systems) at the heart of the agglomeration–dispersion trade-off. Agglomeration in the larger market is beneficial for firms because it gives them better access to consumers. Following this, workers also want to be in the larger market in order to be able to buy goods without having to pay inter-regional trade costs. Krugman's model is based on Dixit and Stiglitz's (1977) model of product differentiation and offers a formalization of Myrdal's circular and cumulative causation. It goes beyond that because agglomeration is not always an equilibrium outcome. This is because, under agglomeration, most goods sold in the periphery need to be shipped from the core and thus prices there may be quite high. In turn, this can make it profitable for firms to locate in the periphery. When trade costs are high, the even dispersion of manufacturing is indeed the unique equilibrium in Krugman's model. On the other hand, when trade costs are low, serving the residual demand in the periphery can be achieved at a low cost and agglomeration occurs. This strand of literature has grown exponentially since 1990, culminating with Fujita, Krugman, and Venables's (1999) book.

The difficulties of spatial empirical work

What about the evidence, then? Ultimately, it is observation that should allow us to judge of the relevance of our theories. To discuss very briefly what the issues for empirical work are, it is useful to retain the partial versus general equilibrium distinction made above. A typical ‘partial equilibrium’ question that has attracted much attention over the years is that of location choices of new firms. To look at the determinants of location choice, one would like to somehow explain location choices in terms of a bunch of possible determinants. This is a difficult exercise, for several reasons. The first one has to do with the nature of the problem. Location choices are not continuous. Because they are discrete, firms decide to locate ‘somewhere’ rather than be spread continuously. Put differently, new firms choose between discrete alternatives so that one has to use discrete choice methods, which are more complex to implement than standard regression approaches. Then there is a whole range of possible determinants for location choices. This makes this type of exercise very data-intensive and particularly prone to missing variables biases. It is also likely that location decisions are made not only on the basis of the characteristics of the spatial units where firms locate, but are also influenced by what happens in neighbouring units. More generally, it is likely to be the case that different determinants of location matter at different spatial scales. To take these concerns into account, spatial econometrics has developed a set of tools. Spatial econometrics resembles standard time-series analysis in that it takes the values of the explanatory variables of the neighbouring spatial units (as well as their error term) into account. The fundamental complication is that spatial dependence can ‘go both ways’, unlike time dependence in time series. An alternative would be to ignore spatial units altogether and work directly on continuous space. Although there have been some developments in that direction (see Cressie, 1993,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for a review), data limitations confine this type of approach to a small range of problems.
When one looks at more general equilibrium issues, many of the difficulties of partial equilibrium analysis are still there while other concerns also become prominent. Take the analysis of regional disparities as an example. A first issue is that such general equilibrium problems have several endogenous variables. In our example regional population and income are likely to be simultaneously determined. To analyse this type of question, two polar approaches (and everything in between) may be adopted. A more descriptive approach consists in focusing on one particular variable, say, local income, and trying to explain its spatial variation using a range of potential factors as indicated by theoretical models. Many of these factors such as the local population are then likely to be endogenous. This requires finding appropriate instruments for such endogenous variables (since unfortunately natural experiments are even scarcer in this field than elsewhere in economics). In some cases, good instruments may not be available. In contrast to descriptive analysis, structural approaches require writing down a particular model and deriving a set of equations that can then be estimated. The main problem faced by this type of approach is that many possible models are likely to have some explanatory power. To return to our example, regional disparities in a country are likely to be caused by the factors highlighted by the urban systems approach (local external effects and so forth) and those highlighted by the new economic geography (trade costs and pecuniary externalities), as well as factor endowments, institutions, and so on. The list of plausible determinants for regional disparities is long and it is very problematic to impose strong priors regarding the validity of one specific model. For many issues in spatial economics, model selection is in fact a huge concern (see Sutton, 2000, for more). Despite these difficulties, it is fair to say that much has been learnt about cities and regions since the mid-1970s (see Rosenthal and Strange, 2004, and Head and Mayer, 2004, for recent reviews).

The road ahead (?)

What current challenges does spatial economics face? On the theoretical front, three main problems remain open. The first is to provide a unified general equilibrium approach to spatial economics and end the often uneasy coexistence between urban systems and the new economic geography. Despite some attempts, as of 2005 there is no such unified framework, and providing one will be difficult. The main obstacles are about modelling. General equilibrium models of spatial economics entail making detailed assumptions about the spatial structure, the production structure, and the mobility of people, goods and ideas, all this under increasing returns. In such cases, nonlinearities occur everywhere and analytical solutions are the exception rather than the rule. Despite this, a general but tractable model of cities and regions is probably worth fighting for. A second key challenge regards the microfoundations of trade costs. Trade costs play a fundamental role in many models but their microeconomic foundations have received only scant attention. This will probably involve looking beyond transport costs and open the black box of the multiplicity of transactions costs associated with trade between different parties. A third major challenge regards the development of a ‘theory of proximity’ (for lack of a better name). Such theory would provide some answers as to why direct interactions between economic agents matter and how. Non-market interactions will no doubt loom large in any theory of proximity.
On the empirical front, a first key challenge is to develop new tools for spatial analysis. With very detailed data becoming available, new tools are needed. Ideally, all the data work should be done in continuous space to avoid border biases and arbitrary spatial units. We are still a long way from being

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
able to do so. The second main challenge for empirical work is of a very different nature. Applied work has over the years managed to produce a reasonable set of estimates regarding a range of issues such as the intensity of local externalities or the determinants of urban growth, among others. No doubt, further progress is necessary and will occur but the main challenge is now to understand the mechanisms behind the elasticities or the decompositions that have been produced. For instance, the elasticity of local productivity to the density of economic activity is now well circumscribed between two and five per cent. We ignore nearly everything about the relative importance of the possible mechanisms behind such numbers. Finally, being able to distinguish between theories – for instance, between factor endowments, urban systems and new geography to explain regional patterns of economic activity – is also a fundamental task where research has barely begun to make progress.

To conclude, one may want to raise the issue of the position of this field within economics and its relationship with other areas of investigation. It is fair to say that, with the advent of Alonso's modern urban economics and that of strategic models of location, spatial economics traded its breadth of knowledge against some depth on much smaller subset of questions. Since the mid-1970s, spatial economics has managed to broaden again its focus by remaining open to outside influences. For instance, the new economic geography finds its roots in international trade theory, while much modern empirical work is heavily influenced by modern applied labour economics and industrial organization. For spatial economics, there is scope for further expansion. Over the years, housing and real estate economics have become fairly detached from the rest of spatial economics and the time may be ripe for new encounters and new cross-fertilizations. A similar statement also holds for local public economics. Finally, outside economics, the part of geography that deals with economic issues, ‘economic geography’, has a focus that considerably overlaps with spatial economics. The relationship between the two disciplines has been fraught with difficulties. On the one hand many geographers react very negatively to the renewed interest by economists in spatial issues. On the other hand, economists tend to ignore the work done by economic geographers. Despite these difficulties, geographers may learn something from the economists’ more rigorous approach while the greater breadth of geographers may offer a great source of inspiration for economists.

See Also

- central place theory
- GIS data in economics
- housing supply
- international trade theory
- location theory
- monocentric versus polycentric models in urban economics
- neighbours and neighbourhoods
- new economic geography
- spatial econometrics
- systems of cities
- urban agglomeration

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- urban economics
- urban growth
- urban housing demand
- urban production externalities
- urban transportation economics

Bibliography

Alonso, W. 1964. *Location and Land Use*. Cambridge, MA: Cambridge University Press.

Cressie, N. 1993. *Statisticsfor Spatial Data*. New York: John Wiley.

Debreu, G. 1959. *Theory of Value*. New Haven: Yale University Press.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Duranton, G. and Puga, D. 2004. Micro-foundations of urban agglomeration economies. In *Handbook of Regional and Urban Economics*, vol. 4, ed. V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Fujita, M. 1989. *Urban Economic Theory: Land Use and City Size*. Cambridge: Cambridge University Press.

Fujita, M., Krugman, P. and Venables, A. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Head, K. and Mayer, T. 2004. The empirics of agglomeration and trade. In *Handbook of Regional and Urban Economics*, vol. 4, ed. V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Henderson, J. 1974. The sizes and types of cities. *American Economic Review* 64, 640–56.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39, 41–57.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99, 484–99.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Myrdal, G. 1957. *Economic Theory and Under-developed Regions*. London: Duckworth.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ricardo, D. 1821. *The Principles of Political Economy*, 3rd edn. Homewood, IL: Irwin, 1963.

Rosenthal, S. and Strange, W. 2004. Evidence on the nature and sources of agglomeration economies. In *Handbook of Regional and Urban Economics*, vol. 4, ed. V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Printed for W. Strahan and T. Cadell.

Starrett, D. 1974. Principles of optimal location in a large homogeneous area. *Journal of Economic Theory* 9, 418–48.

Sutton, J. 2000. *Marshall's Tendencies: What Can Economists Know?* Cambridge. MA: MIT Press.

von Thünen, J. 1826. *Isolated State: An English edition of Der isolierte Staat*, trans. C. Wartenberg and P. Hall. Oxford: Pergamon Press, 1966.

Weber, A. 1909. *Alfred Weber's Theory of the Location of Industries*, trans. C. Friedrich. Chicago: University of Chicago Press, 1929.

How to cite this article

Duranton, Gilles. "spatial economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000195> doi:10.1057/9780230226203.1583

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spatial market integration

Christopher B. Barrett
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

arbitrage; competitive equilibrium; law of one price; Pareto efficiency; parity bounds model; price variability; purchasing power parity; quotas; spatial market integration; tradability

Article

Markets aggregate demand and supply across actors distributed in space. At the international level, monetary policy, exchange rate adjustment and the distribution of the gains from trade depend fundamentally on how well prices equilibrate across countries, as vast literatures on the law of one price and purchasing power parity emphasize (Froot and Rogoff, 1995; Anderson and van Wincoop, 2004). At the national level, well-functioning markets ensure that macro-level economic policies (for example, with respect to exchange rates, trade, and fiscal or monetary policy) change the incentives and constraints faced by micro-level decision-makers. Macroeconomic policy commonly becomes ineffective without strong market transmission across space of the signals sent by central governments. Similarly, well-functioning markets underpin growth stimuli originating in micro-level phenomena. For example, without good access to distant markets that can absorb excess local supply, firms' adoption of improved production technologies will tend to cause producer prices to drop, erasing the gains from technological change and thereby dampening incentives for firms to adopt new technologies that can stimulate economic growth. Poorly integrated markets thereby choke off the prospective gains from technological change. Markets also play a fundamental role in managing risk associated with demand and supply shocks in that well-integrated markets facilitate adjustment in net export flows across space, thereby reducing price variability faced by consumers and producers. Finally, the spatial extent of markets has profound implications for antitrust policy (Stigler and Sherwin, 1985).

The micro-level realities of markets in much of the world, however, involve poor communications and transport infrastructure, limited rule of law, and restricted access to commercial finance, all of which can sharply limit the degree to which markets function as effectively as textbook models typically assume. A long-standing empirical literature documents considerable commodity price variability across space, especially in developing countries, with various empirical tests of market integration suggesting

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
significant and puzzling forgone arbitrage opportunities (Fackler and Goodwin, 2001). The international trade literature similarly finds substantial and sometimes persistent deviations from the law of one price and from purchasing power parity, even among advanced market economies (Froot and Rogoff, 1995; Anderson and van Wincoop, 2004). These results raise important questions about the nature of spatial market integration in actual economies.

The concept of spatial market integration

Although contemporary economics rests fundamentally upon the concept of markets, the discipline struggles with the important and practical challenges of clearly defining a market empirically and of establishing whether markets are efficient in allocating scarce goods and services (Barrett, 2001). Much of the problem revolves around the concept of 'market integration' one employs and the empirical evidence thereby needed to demonstrate that condition. In macroeconomics and international economics, a common conceptualization of market integration focuses on 'tradability', the notion that a good is traded between two economies or that market intermediaries are indifferent between exporting from one nation to another and not doing so. Tradability signals the transfer of excess demand from one market to another, as captured in actual or potential physical flows. Positive trade flows are sufficient to demonstrate spatial market integration under the tradability standard. But prices need not be equilibrated across markets. Spatial market integration conceptualized as tradability is therefore consistent with Pareto-inefficient distributions.

For this reason, the primary approach one finds in the spatial market integration literature focuses instead on the notion of competitive equilibrium and Pareto efficiency manifest in zero marginal profits to arbitrage. At the heart of most analyses of market integration lies the Enke–Samuleson–Takayama–Judge (ESTJ) spatial equilibrium model (Enke, 1951; Samuelson, 1952; Takayama and Judge, 1971), in which the dispersion of prices in two locations for an otherwise identical good is bounded from above by the cost of arbitrage between the markets when trade volumes are unrestricted and bounded from below when trade volumes reach some ceiling value (for example, associated with a trade quota). More precisely, in ESTJ spatial equilibrium

$$p^0 = p^1 + \tau^{10} \text{ if } q^{10} \in (0, q^{10*}) \leq p^1 + \tau^{10} \text{ if } q^{10} = 0 \geq p^1 + \tau^{10} \text{ if } q^{10} = q^{10*}$$

where p^0 and p^1 are the prices in two spatially distinct markets, 0 and 1, respectively, τ^{10} is the cost of moving the good from market 1 to market 0, q^{10} is the physical volume of trade between the two markets and q^{10*} is a maximal permitted trade volume between the two markets (for example, due to a trade quota). These equilibrium conditions imply both firm-level profit maximization and long-run competitive equilibrium at market level. The strict equality reflects the form of competitive equilibrium assumed under the law of one price. If trade occurs and is unrestricted, the marginal trader earns zero profits and prices in the two markets co-move perfectly. The theory, however, implies multiple competitive equilibria. The first weak inequality reflects a segmented equilibrium in which no trade

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
occurs. Prices can be uncorrelated within the price band created by the costs of inter-market arbitrage. The latter weak inequality reflects binding trade quotas that may yield positive marginal quasi-rents to arbitrage.

Note that trade is neither necessary nor sufficient for the attainment of ESTJ competitive equilibria. Hence the difference between tradability-based and efficiency-based conceptualizations of market integration. In the prevailing view, spatial market integration occurs when the ESTJ equilibrium condition holds, irrespective of whether trade occurs.

Empirical estimation methods

The empirical challenge of measuring spatial market integration arises because the ESTJ equilibrium condition involves four variables – prices, transactions costs, trade volumes and trade volume quotas – yet few studies employ more than price data. Spatial price analysis studies typically test for co-movement in time series of prices measured simultaneously at different places. But even with proper controls for autocorrelation or non-stationarity, such studies inevitably impose great structure on the nature of market relationships: for example, linear pricing, continuous unidirectional tradability, and stationary transactions costs series. Tests of the hypothesis of market efficiency thereby become indistinguishable from tests of the veracity of the assumptions that underpin model specification. Simple linear time series tests of market integration-cum-equilibrium using co-integration, error correction or Granger causality models have therefore drawn considerable criticism (Barrett, 1996; Baulch, 1997; Fackler and Goodwin, 2001).

More recent innovations use mixture distribution estimation methods in an attempt to integrate price data with transactions costs, trade volume data, or both, while relaxing some of the strong assumptions that underpin conventional time series methods of testing for market integration. Baulch's (1997) parity bounds model (PBM) that integrates price and transactions costs series is perhaps the best known of these methods. Barrett and Li (2002) extended the PBM to incorporate trade data. These methods have their shortcomings too, however. They rely on inherently arbitrary distributional assumptions in estimation and typically ignore the time series properties of the data, not permitting analysis of the dynamics of inter-temporal adjustment to short-run deviations from long-run equilibrium and potentially important distinctions between short-run and long-run integration (Ravallion, 1986).

A fragile empirical foundation for guiding policy

Even satisfaction of the ESTJ spatial equilibrium condition does not imply welfare maximization unless the costs of commerce and the quasi-rents associated with binding trade quotas are minimized. In order for markets to fulfil the promise they offer for risk management, efficient distribution of production according to comparative advantage, clear transmission of policy signals, and maintenance of micro-level incentives to innovate, there should be neither segmented competitive equilibria nor effective trade quotas. When the costs of commerce are high or trade restrictions bind, it can be difficult to draw out clear implications for policy even from empirical analyses that take seriously the implications of ESTJ spatial equilibrium. Given limited data, in particular a paucity of data on transactions costs and trade volumes, and the intrinsic limitations of existing empirical methods, economists still have only a fragile

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
empirical foundation for reaching clear, strong judgements about spatial market integration as a guide for corporate or government policy.

See Also

- agricultural markets in developing countries
- purchasing power parity
- spatial economics

Bibliography

Anderson, J. and van Wincoop, E. 2004. Trade costs. *Journal of Economic Literature* 42, 691–751.

Barrett, C. 1996. Market analysis methods: are our enriched toolkits well suited to enlivened markets? *American Journal of Agricultural Economics* 78, 825–9.

Barrett, C. 2001. Measuring integration and efficiency in international agricultural markets. *Review of Agricultural Economics* 23, 19–32.

Barrett, C. and Li, J. 2002. Distinguishing between equilibrium and integration in spatial price analysis. *American Journal of Agricultural Economics* 84, 292–307.

Baulch, B. 1997. Transfer costs, spatial arbitrage, and testing for food market integration. *American Journal of Agricultural Economics* 79, 477–87.

Enke, S. 1951. Equilibrium among spatially separated markets: solution by electrical analogue. *Econometrica* 19, 40–7.

Fackler, P. and Goodwin, B. 2001. Spatial price analysis. In *Handbook of Agricultural Economics*, vol. 1B, ed. B. Gardner and G. Rausser. Amsterdam: Elsevier.

Froot, K. and Rogoff, K. 1995. Perspectives on PPP and long-run real exchange rates. In *Handbook of International Economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: Elsevier.

Ravallion, M. 1986. Testing market integration. *American Journal of Agricultural Economics* 68, 102–9.

Samuelson, P. 1952. Spatial price equilibrium and linear programming. *American Economic Review* 42, 283–303.

Stigler, G. and Sherwin, R. 1985. The extent of the market. *Journal of Law and Economics* 28, 555–85.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Takayama, T. and Judge, G. 1971. *Spatial and Temporal Price Allocation Models*. Amsterdam: North-Holland.

How to cite this article

Barrett, Christopher B. "spatial market integration." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000448> doi:10.1057/9780230226203.1584

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spatial mismatch hypothesis

Yves Zenou
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

efficiency wage model; housing market; human capital; labour market; labour productivity; minimum wage model; racial segregation; spatial mismatch hypothesis; unemployment; urban segregation

Article

First formulated by Kain (1968), the spatial mismatch hypothesis states that, residing in urban segregated areas distant from and poorly connected to major centres of employment growth, black workers face strong geographic barriers to finding and keeping well-paid jobs. In the US context, where jobs have been decentralized and blacks have stayed in the central parts of cities, the main conclusion of the spatial mismatch hypothesis is that distance to jobs is the main cause of high unemployment rates and low earnings among blacks. The spatial mismatch literature has focused on race under the presumption that (inner-city) blacks are not residing close to (suburban) jobs, either because they are discriminated against in the (suburban) housing market or because they want to live near members of their own race. Most of this literature has focused on black workers, and it is only recently that the analysis has been extended to other minority workers, especially Hispanics.

Since Kain's study, hundreds of others have been conducted trying to test the spatial mismatch hypothesis (see, in particular, the literature survey by Ihlanfeldt and Sjoquist, 1998). The usual approach is to relate a measure of labour-market outcomes, typically employment or earnings, to another measure of job access, typically some index that captures the distance between residences and centres of employment. Some control variables (typically human capital variables) are also included.

The main econometric problem with this test is that residential location is endogenous, since families are not randomly assigned residential locations but instead choose them (Ihlanfeldt, 2005). Thus, self-selection rather than distance to jobs may explain why black workers have adverse labour market outcomes. This problem has been dealt with mainly by exploiting inter-city variations in black residential centralization to estimate the effect of job access on black employment (Weinberg, 2000). Another way is to focus the analysis on youth who still reside with their parents, since residential

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
location is decided by parents for their children (Raphael, 1998). Given the limits of these approaches, the general conclusions for youth workers are: (*a*) poor job access indeed worsens labour-market outcomes, (*b*) black and Hispanic workers have worse access to jobs than white workers, and (*c*) racial differences in job access can explain between one-third and one-half of racial differences in employment. The theoretical foundations of these empirical results, however, remain unclear. If researchers do agree on the causes (housing discrimination and/or social interactions) and on the consequences of the spatial mismatch hypothesis (higher unemployment rates and lower wages for black workers), the economic mechanisms and thus the policy implications are difficult to identify.

A first theoretical view (Brueckner and Zenou, 2003) is to argue that suburban housing discrimination skews black workers towards the central business district (CBD) and thus keeps black residences remote from the suburbs. Since black workers who work in the suburban business district (SBD) have more costly commutes, few of them will accept SBD jobs, which makes the black CBD labour pool larger than the SBD pool. Under either a minimum wage or an efficiency wage model, this enlargement of the CBD pool leads to a high unemployment rate among CBD workers.

Another theory (Zenou, 2002) has proposed that distance has a negative impact on workers' productivity. Indeed, because of the lack of good public transportation in large US metropolitan areas, especially from the central city to the suburbs, blacks have relatively low productivity at suburban jobs because they arrive late to work due to the unreliability of the mass transit system, which causes them to frequently miss transfers. If this is true, then firms may draw a red line beyond which they will not hire workers. So, if housing discrimination against blacks forces them to live far away from jobs, then firms will be reluctant to hire black workers because they have relatively lower productivity than whites.

See Also

- housing policy in the United States
- poverty traps
- residential segregation
- unemployment
- urban economics
- wage inequality, changes in

Bibliography

Brueckner, J. and Zenou, Y. 2003. Space and unemployment: the labour-market effects of spatial mismatch. *Journal of Labor Economics* 21, 242–66.

Ihlanfeldt, K. 2005. A primer on spatial mismatch within urban labor markets. In *A Companion to Urban Economics*, ed. R. Arnott and D. McMillen. Boston: Blackwell Publishing.

Ihlanfeldt, K. and Sjoquist, D. 1998. The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. *Housing Policy Debate* 9, 849–92.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kain, J. 1968. Housing segregation, Negro employment, and metropolitan decentralization. *Quarterly Journal of Economics* 82, 32–59.

Raphael, S. 1998. The spatial mismatch hypothesis and black youth joblessness: evidence from the San Francisco Bay area. *Journal of Urban Economics* 43, 79–111.

Weinberg, B. 2000. Black residential centralization and the spatial mismatch hypothesis. *Journal of Urban Economics* 48, 110–34.

Zenou, Y. 2002. How do firms redline workers? *Journal of Urban Economics* 52, 391–408.

How to cite this article

Zenou, Yves. "spatial mismatch hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000457> doi:10.1057/9780230226203.1585

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

specie-flow mechanism

William R. Allen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balance of payments; Cantillon, R.; comparative advantage; Gervaise, I.; Hume, D.; law of one price; Malynes, G. de; Marshall–Lerner condition; mercantilism; monetary approach to the balance of payments; price level; price-specie-flow mechanism; quantity theory of money; specie-flow mechanism

Article

The ‘specie-flow mechanism’ is an analytic version of automatic, or market, adjustment of the balance of international payments. In competitive markets with specie-standard institutions, behaviour will lead to national price levels and income flows consistent with equilibrium in the international accounts, commonly interpreted in this context to mean zero trade balances.

The classic exposition of the mechanism, for the better part of two centuries all but universally accepted, at least as a first approximation, was provided by David Hume in a 1752 essay, ‘Of the Balance of Trade’. While it is appropriate to associate the essence of the model with Hume, all the ingredients of Hume's argument had long been available. There were even notable prior attempts to fit the analytic pieces into a self-contained model. Further, even if we give to Hume all the considerable credit due to his systematic, compact statement, his version is not the whole of the specie-flow mechanism; and the specie-flow mechanism is not the whole analysis of balance of payments adjustment.

Hume's presentation is a simple application of the quantity theory of money in a setting of international trade and its financing. With a pure 100 per cent reserve gold standard, and beginning with balance in the international accounts, a decrease in the money stock of country A results in a directly proportionate fall in its price level, which is also a decrease relative to the initially unaffected price levels of other countries; as country A's price level falls, consumer response, in Hume's account, will reduce A's imports and increase its exports; when the exchange rate is bid to the gold point, the export trade balance will be financed by gold inflow, which will raise prices in A and lower prices abroad until the international price differentials and net trade flows are eliminated. The line of causation runs from changes in money to changes in prices to changes in net trade flows to international movements of gold that eliminate the earlier price differentials and thereby correct the trade imbalance and stop the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
shipment of gold. In equilibrium, the distribution of gold among countries (and regions within countries) yields national (and regional) price levels consistent with zero trade balances.

This theory of trade equilibration links with the Ricardian theory of production specialization. In a comparative advantage model of two countries, two commodities, and labour input, country A has absolute advantages of different degrees in both goods. To have two-way trade, the wage rate of country A must be greater than that abroad, within the wage-ratio range specified by the proportions of A's productive superiority in the two goods. Gold will flow until the international wage ratio yields domestic prices that equate total import and export values.

The conclusion that trade imbalances, and thus gold flows, cannot long obtain was in fundamental contrast to the mercantilistic emphasis on persistent promotion of an export balance and indefinite accumulation of gold. Still, the mercantilists decidedly associated gold inflows with export surpluses of goods and services; a good many writers had posited a direct relation between the money stock and the price level; similarly, it had been indicated that relative national price level changes would affect trade flows. However, while we should bow to such predecessors of Hume as Isaac Gervaise (1720) and Richard Cantillon (1734) and perhaps nod to Gerard de Malynes (1601) for attempts to construct adjustment models, Hume put the elements together with unmatched elegance and awareness of implication – and influence.

Hume's version was specifically a *price*-specie-flow mechanism, with the prices being national price levels (and exchange rates). Even as a price mechanism, the model has problems.

While it is reasonable to presume that price levels will move in the same directions (even if not in the same proportions) as the huge changes in the money stock envisioned by Hume, there remain questions of the impact on import and export expenditures. Vertical demand schedules in country A for imports and in other countries for A's exports would leave the physical amounts of imports and exports unresponsive to price changes. If, following Hume, we upset the initial equilibrium by a large decrease in money and thus in prices in country A, foreign expenditure on A's goods will fall proportionately with the fall in A's prices. The import balance of A will be financed with gold outflow, resulting in a further fall in A's prices and export value and an increase in prices abroad and in A's import expenditure. The gold flow, rather than correcting the trade flow, will increase the import trade balance of A when demand elasticities are zero (or sufficiently small). The import and export demand (and supply) elasticity conditions required for price (including exchange rate) changes to be equilibrating – conditions which are empirically realistic – came much later to be summarized in the ‘Marshall–Lerner condition’. Under the most unfavourable circumstances of infinite supply elasticities and initially balanced trade, all that is required for stability is that the arithmetic sum of the elasticities of foreign demand for A's exports and of A's demand for imports be greater than unity.

Aside from the nicety of specifying elasticity conditions for stability, is it appropriate to couch the model in terms of diverging national price levels or of changes in a country's import prices compared to its export prices? Suppose country A has a commodity export balance, resulting perhaps from a shift in international demands reflecting changed preferences in favour of A's goods or imposition of a tariff by A or a foreign crop failure. As gold flows in, A's expenditures expand and prices are expected to rise. Prices of A's *domestic* goods (which do not enter foreign trade) do rise; but prices of *internationally* traded goods are affected little, if at all, for the increase in A's demand for such goods is countered by decrease in demand for them in gold-losing countries. Consumers in A, facing the domestic–international price divergence, shift to now relatively cheapened international goods (imports and A-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
exportables) from more expensive domestic goods, thus increasing import volume and value and also absorption of exportables. Producers in A shift out of international goods into domestic, thus reducing exports and expanding imports. Corresponding, but opposite, substitutions and shifts are diffused among the other countries. These respective domestic adjustments in consumption and production would continue until the gold flow ceases and the trade imbalance is corrected.

Substantial modern empirical research, however, is more supportive of Hume's changes in the terms of trade or of transitory divergences in relative prices of traded and non-traded goods than of the assumed invariant applicability of the equilibrium ‘law of one price’ commonly adopted in the modern ‘monetary approach’ to the balance of payments.

When gold flows into country A, portfolio equilibria of individuals and firms are upset, with cash balances now in excess. People try to spend away redundant balances. Expenditure rises and money income becomes larger. With greater income, demands for goods – including foreign goods – increase: at any given commodity price, quantity demanded has become larger. Import quantities and values rise. Changes in money give rise abroad to opposite portfolio adjustments and changes of income, thereby decreasing A's exports. In all this, there are some changes (upward in A and downward abroad) in prices of domestic goods and production factors, but the adjustment process entails income changes as well as price changes.

Some such role of changes in money income and demand schedules was noted – in different contexts and with different degrees of clarity and emphasis – by many writers in the 19th and early 20th centuries. But single-minded emphasis on income, with little or no explicit role for the money stock and prices, came only with application to balance of payments adjustment of the national income theory of J. M. Keynes. However, such application – with its regalia of marginal propensities and secondary, supplemental repercussions of multipliers – is not contingent on, or uniquely associated with, an international gold standard. Further, neglect of money in the foreign-trade multiplier analysis is a grievous omission. Equilibrium in the income model is characterized by equating of the flows of income leakages (saving, tax payments, imports) and income injections (investment, government expenditure, exports). But such equality of total leakages and injections permits a continuing trade imbalance. And a trade imbalance financed by a gold flow – or accompanied by money change generally – leads to further change in income; that is, income had not reached a genuine equilibrium.

The actual world, even with the classical gold standard in the generation prior to the Second World War, has not conformed well in institutions and processes with the construct of Hume. A world generally of irredeemable paper money and universally of demand deposits along with fractional-reserve banking and discretionary money policy – a world including the International Monetary Fund arrangement of indefinitely pegged exchange rates – has relied on selected adjustment procedures more than on automatic adjustment mechanisms. So Hume's model in its own terms is inadequate and in important empirical respects is even inappropriate. But it provided analytical coherency and expositional emphasis in an early stage of a discussion which continues to evolve.

Bibliography

Blaug, M. 1985. *Economic Theory in Retrospect*. Cambridge: Cambridge University Press.

Darby, M. and Lothian, J. 1983. *The International Transmission of Inflation*. Chicago: University of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Chicago Press.

Fausten, D. 1979. The Humean origin of the contemporary monetary approach to the balance of payments. *Quarterly Journal of Economics* 93, 655–73.

Rotwein, E., ed. 1970. *David Hume: Writings on Economics*. Madison: University of Wisconsin Press.

Yeager, L. 1976. *International Monetary Relations: Theory, History, and Policy*. 2nd edn. New York: Harper & Row.

How to cite this article

Allen, William R. "specie-flow mechanism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000198> doi:10.1057/9780230226203.1586

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

specification problems in econometrics

Edward E. Leamer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The theory of econometrics presumes a ‘specification’ that selects a sharp borderline between (*a*) assumptions that are maintained and (*b*) questions that the data are allowed to address. For example, one might treat the list of variables in a regression model as known with certainty but the ‘coefficients’ of these variables as uncertain. In practice, a wide and fuzzy border between the maintained assumptions and the uncertain assumptions causes great ambiguity in the inferences economists draw from their non-experimental data.

Keywords

Bayesian inference; criticism; data-instigated models; Durbin–Watson statistic; Econometrics; frequentist econometrics; goodness of fit; linear regression; robustness; sensitivity analysis; simplification; specification; statistical estimation; statistical inference; subjective probability

Article

Specification problems in econometrics arise because economic theory often identifies a generally agreed upon framework (such as market determination of price and volume) but leaves up to the individual analyst the translation of the framework into a fully defined empirical model. With virtually no guidance from theory, a data analyst is expected to choose a list of relevant variables, the functional form, the separation of variables into endogenous and exogenous, the dynamics, and the error distributions. Substantial doubt about these assumptions is a characteristic of the analysis of non-experimental data and much experimental data as well. If this uncertainty is left unattended, it can cause serious doubt about the corresponding inferences.

To emphasize the distinction between the general framework and an instance of the framework on which the data analysis rests, the specific set of assumptions used to draw inferences from a data set is called a ‘specification’. The treatment of doubt about the specification is called ‘specification analysis’. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
research strategy of trying many different specifications is called a ‘specification search’.

Estimation, sensitivity analysis and simplification searches: three treatments of specification ambiguity

When an inference is suspected to depend crucially on a doubtful assumption, two kinds of actions can be taken to alleviate the consequent doubt about the inferences. Both require a list of alternative assumptions. The first approach is *statistical estimation*, which uses the data to select from the list of alternative assumptions and then makes suitable adjustments to the inferences to allow for doubt about the assumptions. (I am including under the heading ‘estimation’ the kind of two-valued estimation problem that economists usually call hypothesis testing.) The second approach is a *sensitivity analysis* that uses the alternative assumptions one at a time, thereby demonstrating either that all the alternatives lead to essentially the same inferences or that minor changes in the assumptions make major changes in the inferences. For example, a doubtful variable can simply be included in the equation (estimation), or two different equations can be estimated, one with and one without the doubtful variable (sensitivity analysis).

The borderline between the techniques of estimation and sensitivity analysis is not always clear since a specification search can be either a method of estimation of a general model or a method of studying the sensitivity of an inference to choice of model. Stepwise regression, for example, which involves the sequential deletion of ‘insignificant’ variables and insertion of ‘significant’ variables is best thought to be a method of estimation of a general model rather than a study of the sensitivity of estimates to choice of variables, since no attempt is generally made to communicate how the results change as different subsets of variables are included.

A fundamental distinction between estimation and sensitivity analysis is whether the logic is two-valued or three-valued. The logic of traditional econometric estimation is two-valued: either one takes the action or one does not. A sensitivity analysis offers a third possible conclusion: the data cannot be relied upon to make the decision.

If the data are very informative, estimation is the preferred approach. But parameter spaces can always be enlarged beyond the point where data can be helpful in distinguishing alternatives. When abbreviated parameter spaces appear to be used, there usually lurks behind the scene a much larger space of assumptions that ought to be explored. If this larger space has been explored through a pre-testing procedure and if the data are sufficiently informative in indicating that estimation is the preferred approach, then adjustments to the inferences are in order to account for the pre-testing bias and model uncertainty. If the data are not adequately informative about the parameters of the larger space, we need to have answers to the sensitivity question whether ambiguity about the best method of estimation implies consequential ambiguity about the inferences. A data analysis should therefore combine estimation with sensitivity analysis, and only those inferences that are clearly favoured by the data or are sturdy enough to withstand minor changes in the assumptions should be retained.

Estimation and sensitivity analyses are two phases of a data analysis. Simplification is a third. The intent of simplification is to find a simple model that works well for a class of decisions. A specification search can be used for simplification, as well as for estimation and sensitivity analysis. Confusion among these three kinds of searches ought to be eliminated since the rules for a search and measures of success properly depend on the intent.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Criticism and revision: a deeper problem

These first three kinds of specification searches fall within the reach of traditional econometric theory, both frequentist and Bayesian. But neither theory of inference can deal with another reason for exploring more than one model: a data anomaly that forces the analyst to alter the model and thus to carry out an action that was wholly unanticipated and unplanned. This is not allowed within traditional theories of inference, which presume that responses to data are fully planned and completely committed before the data are observed. From a frequentist standpoint, a fully committed plan is needed to determine sampling properties. From a Bayesian perspective, subjective probabilities fully determine the responses to the data, and, if elicitation of probabilities is allowed to be determined after the data are explored, then we double count the data evidence – once to form the 'priors' and again to update the priors.

In settings in which the theory and the method of measurement are clear, responses can be conveniently planned in advance. In practice, however, most data analysts have very low levels of commitment to whatever plans they may have formulated before reviewing the data. Even when planning is extensive, most analysts reserve the right to alter the plans if the data are judged 'unusual'. A review of the planned responses to the data after the data are actually observed can be called *criticism*, the function of which is either to detect deficiencies in the original family of models that ought to be remedied by enhancements of the parameter space or to detect inaccuracies in the original approximation of prior information. When either the model or the prior information is revised, the planned responses are discarded in favour of what at the time seem to be better responses.

The form that criticism should take is not clear cut. Much of what appears to be criticism is in fact a step in a process of estimation, since the enhancement of the model is completely predictable. An example of an estimation method masquerading as criticism is a t-test to determine if a specific variable should be added to the regression. In this case the response to the data is planned in advance and undergoes no revision once the data are observed.

Criticism and the prospect of the revision of planned responses create a crippling dilemma for both classical and Bayesian inference. According to classical inference, the choice of procedure for analysing the data should be based entirely on sampling properties, but these are impossible to compute unless the response to every conceivable data set is planned and fully committed. A Bayesian has problems whether or not a criticism is successful. When a criticism is successful, that is to say when the family of models is enhanced or the prior distribution is altered in response to anomalies in the data, there is a severe double counting problem if estimation then proceeds as if the model and prior distribution were not data-instigated. Even if criticism is not successful, the prospect of successful criticism makes the inferences from the data weaker than conventionally reported because the commitment to the model and the prior is weaker than is admitted.

Estimation: choice of variables for linear regression

Specification problems are not limited to, but are often discussed within, the context of the linear regression model, probably because the most common problem facing analysers of economic data is doubt about the exact list of explanatory variables. The first results in this literature addressed the effect of excluding variables that belong in the equation, and including variables that do not. Let y represent the dependent variable, x an included explanatory variable, and z a doubtful explanatory variable. If we

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

assume that:

$$E(y|x, z) = \alpha + x\beta + z\theta.$$
$$\mathrm{Var}(y|x) = \sigma^2,$$
$$E(z|x) = c + xr$$
$$\mathrm{Var}(z|x) = s^2$$

where α, β, θ, c, r, σ^2 and s^2 are unknown parameters, and y, x, and z are observable vectors, then β can be estimated with z included in the equation:

$$b_{.z} = \left(x' M_z x\right)^{-1}\left(x' M_z y\right),$$

or with z excluded:

$$b = \left(x' x\right)^{-1} x' y,$$

where

$$M_z = I - z\left(z' z\right)^{-1} z'.$$

The first two moments of these estimators are straightforwardly computed:

$$E(b_{.z}) = \beta,\ E(b) = \beta + r\theta \mathrm{Bias}(b_{.z}) = 0\ \mathrm{Bias}(b) = r\theta \mathrm{Var}(b_{.z}) = \sigma^2\left(x' M_z x\right)^{-1} \mathrm{Var}(b) = \sigma^2\left(x' x\right)^{-1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where

$$\text{Bias}(b) = E(b) - \beta.$$

A quick algebraic calculation reveals that $\text{Var}(b_{.z}) \geq \text{Var}(b)$. These moments form two basic results in 'specification analysis' made popular by Theil (1957): (*a*) if a relevant variable is excluded $(\theta \neq 0)$, the estimator is biased by an amount that is the product of the coefficient of the excluded variable times the regression coefficient from the regression of the included on the excluded variable; (*b*) if an irrelevant variable is included $(\theta = 0)$, the estimator remains unbiased, but has an inflated variance.

This bias result can be useful when the variable *z* is unobservable and information is available on the probable values of θ and *r*, since then the bias in *b* can be corrected. But if both *x* and *z* are observable, these results are not useful by themselves because they do not unambiguously select between the estimators, one doing well in terms of bias but the other doing well in terms of variance. The choice will obviously depend on information about the value of $r\theta$ since a small value of $r\theta$ implies a small value for the bias of *b*. The choice will also depend on the loss function that determines the trade-off between bias and variance.

For mathematical convenience, the loss function is usually assumed to be quadratic:

$$L\left(\beta^*, \beta\right) = \left(\beta^* - \beta\right)^2,$$

where β^* is an estimator of β. The expected value of this loss function is known as the mean squared error of the estimator, which can be written as the variance plus the square of the bias:

$$\text{MSE}(b_{.z}, \beta) - \text{MSE}(b, \beta) = r\left(\text{Var}\left(\theta^*_{.x}\right) - \theta\theta'\right)r'$$

where $\theta^*_{.x}$ is the least squares estimator of θ controlling for *x*, and where the notation allows *x* and *z* to represent collections of variables as well as singlets. Through inspection of this formula we can derive the fundamental result in this literature. The estimator based on the restricted model is better in the mean squared error sense than the unrestricted estimator if and only if θ is small enough that $\text{Var}\left(\theta^*_{.x}\right) - \theta\theta'$ is positive definite. If θ is a scalar, this condition can be described as 'a true *t* less than one', $\theta^2 / \text{Var}\left(\theta^*_{.x}\right) < 1$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This result is also of limited use since its answer to the question 'Which estimator is better?' is another question, 'How big is θ?' A clever suggestion is to let the data provide the answer, and to omit *z* if its estimated *t* value is less than one. Unfortunately, because the estimated *t* is not exactly equal to the true *t*, this two-step procedure does not yield an estimator that guarantees a lower mean squared error than unconstrained least squares. Thus the question remains: how big is θ? For more discussion consult Judge and Bock (1983).

Since the choice of estimator depends on what we already know about the parameter, we need to find a way to include in the analysis some explicit dependence on the prior state of knowledge. A Bayesian analysis allows the construction of estimators that make explicit use of prior information about θ. It is convenient to assume that the information about θ takes the form of a preliminary data set in which the estimate of θ is zero (or some other number, if you prefer). Then the Bayes estimate of β is a weighted average of the constrained and unconstrained estimators:

$$b_B = \left(v^{-1} + v'^{-1}\right)^{-1}\left(v^{-1}b_{.z} + v'^{-1}b\right)$$

where v' is the prior variance for *h*, and *v* is the sampling variance for $\theta^*_{.x}$. Instinct might suggest that this compromise between the two estimators would depend on the sampling variances of $b_{.z}$ and *b*, but the correct weights are inversely proportional to prior variance and the sample variance for θ.

A card-carrying Bayesian regards this to be the solution to the problem. Others will have a different reaction. What the Bayesian has done is only to enlarge the family of estimators. The two extremes are still possible since we may have $v'=0$ or $v' \to \infty$ but in addition there are the intermediate values $v'>0$. Thus the Bayesian answer to the question is another question: 'What is the value of v'?'

Sensitivity analysis

At this point we have to switch from the estimation mode to the sensitivity mode, since precise values of v' will be hard to come by on a purely a priori basis and since the data usually will be of little help in selecting v' with great accuracy. A sensitivity analysis can be done from a classical point of view simply by contrasting the two extreme estimates, $b_{.z}$ and *b*, corresponding to the extreme values of v'.

A Bayesian approach allows a much richer set of sensitivity studies. A mathematically convenient analysis begins with a hypothetical value for v', say v'_0, which is selected to represent as accurately as possible the prior information that may be available. A neighbourhood around this point is selected to reflect the accuracy with which v'_0 can be chosen. For example, v' might be restricted to lie in the interval

$$v'_0 / (1 + c) < v' < v'_0 (1 + c),$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where c measures the accuracy of v_0'. The corresponding interval of Bayes's estimates b_b is

$$1 / \left(v_0'(1+c)+v\right) < (b_B - b_{.z}) / v(b - b_{.z}) < (1+c) / \left(v_0' + v(1+c)\right),$$

where it is assumed that $(b - b_{.z})>0$. If this interval is large for small values of c, then the estimate is very sensitive to the definition of the prior information. For example, suppose that interest focuses on the sign of β. Issues of standard errors aside, if $b_{.z}$ and b are the same sign, then the inference can be said to be sturdy since no value of c can change the sign of the estimate b_b. But if $b>0>b_{.z}$, then the values of c in excess of the following will cause the interval of estimates to overlap the origin:

$$c^* = \max\left[u - 1, u^{-1} - 1\right],$$

where $u = -\left(v / v_0'\right)(b / b_{.z})$ Thus if u is close to one, the inference is fragile. This occurs if differences in the absolute size of the two estimates are offset by differences in the variances applicable to the coefficient of the doubtful variable. Measures like these can be found in Leamer (1978; 1982; 1983b).

Robustness

When a set of acceptable assumptions does not map into a specific decision, the inference is said to be fragile. A decision can then sensibly be based on a minimax criterion that selects a ‘robust’ procedure that works well regardless of the assumption. The literature on ‘robustness’ such as that reviewed by Krasker, Kuh and Welsch (1983) has concentrated on issues relating to the choice of sampling distribution, but could be extended to choice of prior distribution.

Simplification, proxy selection and data selection

A specification search involving the estimation of many different models can be a method of estimation or a method of sensitivity analysis. Simplification searches are also common, the goal of which is to find a simple quantitative facsimile that can be used as a decision-making tool. For example, a model with a high R^2 can be expected to provide accurate forecasts in a stable environment, whether or not the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
coefficients can be given a causal interpretation. In particular, if two explanatory variables are highly correlated, then one can be excluded from the equation without greatly reducing the overall fit since the included variable will take over the role of the excluded variable. No causal significance necessarily attaches to the coefficient of the retained variable.

A specification search can also be used to select the best from a set of alternative proxy variables, or to select a data subset. These problems can be dealt with by enlarging the parameter space to allow for multiple proxy variables or unusual data points. Once the space is properly enlarged, the problems that remain are exactly the same as the problems encountered when the parameters are coefficients in a linear regression, namely estimation and sensitivity analysis.

Data-instigated models

The subjects of estimation, sensitivity analysis and simplification deal with concerns that arise during the planning phase of a statistical analysis when alternative responses to hypothetical data-sets are under consideration. A distinctly different kind of specification search occurs when anomalies in the actual data suggest a revision in a planned response, for example, the inclusion of a variable that was not originally identified. This is implicitly disallowed by formal statistical theories that presuppose the existence of a response to the data that is planned and fully committed. I like to refer to a search for anomalies as ‘Sherlock Holmes inference’, since, when asked who might have committed the crime, Holmes replied, ‘No data yet … It is a capital mistake to theorize before you have all the evidence. It biases the judgements.’ This contrasts with the typical advice of theoretical econometricans: ‘No theory yet. It is a capital mistake to look at the data before you have identified all the theories.’

Holmes is properly concerned that an excessive degree of theorizing will make it psychologically difficult to see anomalies in the data that might, if recognized, point sharply to a theory that was not originally identified. On the other hand, the econometrician is worried that data evidence may be double counted, once in the Holmesian mode to instigate models that seem favoured by the data and again in the estimation mode to select the instigated models over original models. Holmes is properly unconcerned about the double counting problem, since he has the ultimate extra bit of data: the confession. We do not have the luxury of running additional experiments and the closest that we can come to the Holmesian procedure is to set aside a part of the data set in hopes of squeezing a confession after we have finished identifying a set of models with a Holmesian analysis of the first part of the data. Unfortunately, our data-sets never do confess, and the ambiguity of the inferences that is clearly present after the Holmesian phase lingers on with very little attenuation after the estimation phase. Thus we are forced to find a solution to the Holmesian conundrum of how properly to characterize the data evidence when models are instigated by the data, that is to say, how to avoid the double counting problem. Clearly, what is required is some kind of penalty that discourages but does not preclude Holmesian discoveries. Leamer (1978) proposes one penalty that rests on the assumption that Holmesian analysis mimics the solution to a formal pre-simplification problem in which models are explicitly simplified before the data are observed in order to avoid observation and processing costs that are associated with the larger model. Anomalies in the data-set can then suggest a revision of this decision. Of course, real Holmesian analysis cannot actually solve this sequential decision problem since in order to solve it one has to identify the complete structure that is simplified before observation. But we can nonetheless act as if we were solving this problem, since by doing so we can compute a very sensible kind of penalty for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Holmesian discoveries (Leamer, 1978).

Criticism

Criticism refers to a search for data anomalies that might force a revision of the model. Criticism and data-instigated models are not as frequent as they may appear. As remarked before, much of what is said to be criticism is only a step in a method of estimation, and many models that seem to be data-instigated are in fact explicitly identified in advance of the data analysis. For example, forward stepwise regression, which adds statistically significant variables to a regression equation, cannot be said to be producing data-instigated models because the set of alternative models is explicitly identified before the data analysis commences and the response to the data is fully planned in advance. Stepwise regression is thus only a method of estimation of a general model. Likewise, various diagnostic tests that lead necessarily to a particular enhancement of the model, such as a Durbin–Watson test for first-order autocorrelation and Ramsey's (1969) so-called specification error test for a special kind of nonlinearity, select but do not instigate a model.

‘Goodness of fit’ tests that do not have explicit alternatives are sometimes used to criticize a model. However, the Holmesian question is not whether the data appear to be anomalous with respect to a given model but rather whether there is a plausible alternative that makes the data appear less anomalous and, most importantly, more understandable. Goodness of fit tests have nothing to do with understanding. They measure statistical properties that may or may not be meaningful in the context in which the data are being studied. In large samples, all models have large goodness of fit statistics, and the size of the statistic is no guarantee, or even a strong suggestion, that there exists a plausible alternative model that is substantially better than the one being used.

Unexpected parameter estimates are probably the most effective criticisms of a model. A Durbin–Watson statistic that indicates a substantial amount of autocorrelation can be used legitimately to signal the existence of left-out variables in settings in which there is strong prior information that the residuals are white noise. Apart from unexpected estimates, graphical displays and the study of influential data points may stimulate thinking about the inadequacies in a model.

See Also

- Bayesian econometrics
- econometrics
- statistical inference

Bibliography

Box, G.E.P. 1980. Sampling and Bayes’ inference in scientific modelling and robustness. *Journal of the Royal Statistical Society* 143, 383–430.

Judge, G.G. and Bock, M.E. 1983. Biased estimation. In *Handbook of Econometrics*, vol. 1, ed. Z.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Griliches and M. Intriligator. Amsterdam: North-Holland.

Krasker, W.S., Kuh, E. and Welsch, R.E. 1983. Estimation for dirty data and flawed models. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Leamer, E.E. 1978. *Specification Searches*. New York: Wiley.

Leamer, E.E. 1982. Sets of posterior means with bounded variance priors. *Econometrica* 50, 725–36.

Leamer, E.E. 1983a. Let's take the con out of econometrics. *American Economic Review* 73, 31–43.

Leamer, E.E. 1983b. Model choice and specification analysis. In *Handbook of Econometrics*, vol. 1, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Ramsey, J.B. 1969. Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society* 32, 350–71.

Theil, H. 1957. Specification errors and the estimation of economic relationships. *Review of the International Statistical Institute* 25, 41–51.

How to cite this article

Leamer, Edward E. "specification problems in econometrics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000200> doi:10.1057/9780230226203.1587

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spectral analysis

Timothy J. Vogelsang
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Spectral analysis is a statistical approach for analysing stationary time series data in which the series is decomposed into cyclical or periodic components indexed by the frequency of repetition. Spectral analysis falls within the frequency domain approach to time series analysis. The spectral density function plays the central role and it summarizes the contributions of cyclical components to the variation of a stationary time series. The spectral density at frequency zero is particularly important because of its direct link to the variance of a time series sample average, that is, the long-run variance.

Keywords

alias effect; ARMA models; autocovariances; autoregressive spectral density estimator; bandwidth; Bartlett kernel; bias; cross-spectral densities; cycles; econometric methodology; fixed-*b* asymptotics; frequency; frequency domain; generalized method of moments; Granger causality; heteroskedasticity and autocorrelation covariance matrix estimation; kernel estimators; long-run variance; mean square error; nonparametric estimators; probability; quadratic spectral kernel; spectral analysis; spectral density; spectral density matrix; spectral representation theorem; time domain; time series analysis; truncation lag; variance; white noise

Article

Spectral analysis is a statistical approach for analysing stationary time series data in which the series is decomposed into cyclical or periodic components indexed by the frequency of repetition. Spectral analysis falls within the frequency domain approach to time series analysis. This is in contrast with the time domain approach in which a time series is characterized by its correlation structure over time. While spectral analysis provides a different interpretation of a time series from time domain approaches, the two approaches are directly linked to each other.
Statistical spectral analysis tools were first developed in the middle of the 20th century in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

mathematical statistics and engineering literatures, and many of the important early contributions are discussed in the classic textbook treatment by Priestley (1981). The label 'spectral' has been adopted because of the close link to the physics of light. While the analogy with the physics of light is fairly useless in economics, economists recognized by the 1960s that spectral analysis is a useful empirical tool for understanding the cyclical nature of many time series, and it provides a powerful theoretical framework for developing econometric methodology: for example, the theoretical underpinnings of Granger causality (Granger, 1969) are based in spectral analysis. Since the 1960s, spectral analysis tools have become standard parts of the time series econometrics toolkit, and have influenced a broad range of areas within econometrics. A comprehensive list of references would be long but some notable examples are: band spectral regression (Engle, 1974), generalized method of moments (GMM) (Hansen, 1982), heteroskedasticity autocorrelation (HAC) covariance matrix estimation and inference (Newey and West, 1987; Andrews, 1991; Kiefer and Vogelsang, 2005), unit root testing (Phillips and Perron, 1988), cointegration (Stock and Watson, 1988; Phillips and Hansen, 1990), semiparametric methods (Robinson, 1991), structural identification of empirical macroeconomics models (Blanchard and Quah, 1989; King et al., 1991), testing for serial correlation (Hong, 1996), measures of persistence (Cochrane, 1988), measures of fit for calibrated macro models (Watson, 1993), estimation of long memory models (Geweke and Porter-Hudak, 1983).

Let y_t, t=1,2, … denote a second-order stationary time series with mean $\mu = E(y_t)$ and autocovariance function $\gamma_j = cov(y_t, y_{t-j})$. Most empirical economists find it natural to characterize relationships between random variables in terms of correlation structure, and γ_j conveniently summarizes the statistical structure of y_t. Autocovariances are fundamental population moments of a time series not directly connected to any specific modelling choice. In contrast, the idea of decomposing y_t into cyclical components may appear to impose restrictions on y_t; but an important result, the spectral representation theorem, indicates that nearly any stationary time series can be represented in terms of cyclical components. By using notation from Hamilton (1994), nearly any stationary (discrete-time) time series can be represented as

$$y_t = \mu + \int_0^{\pi} [\alpha(\omega)\cos(\omega t) + \delta(\omega)\sin(\omega t)]\, d\omega, \tag{1}$$

where ω denotes frequency and $\alpha(\omega)$ and $\delta(\omega)$ are mean zero random processes such that for any frequencies $0 < \omega_1 < \omega_2 < \omega_3 < \omega_4 < \pi$,

$$cov\left(\int_{\omega_1}^{\omega_2} \alpha(\omega)\, d\omega, \int_{\omega_3}^{\omega_4} \alpha(\omega)\, d\omega\right) = 0, \quad cov\left(\int_{\omega_1}^{\omega_2} \delta(\omega)\, d\omega, \int_{\omega_3}^{\omega_4} \delta(\omega)\, d\omega\right) = 0,$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and for any frequencies $0 < \omega_1 < \omega_2 < \pi$ and $0 < \omega_3 < \omega_4 < \pi$,

$$cov\left(\int_{\omega_1}^{\omega_2} \alpha(\omega)\, d\omega,\ \int_{\omega_3}^{\omega_4} \delta(\omega)\, d\omega\right) = 0.$$

It is fundamental that a stationary time series can be decomposed into (random) cyclical (cosine and sine) components.

A useful way of interpreting (1) is to measure how the cyclical components contribute to the variation in y_t. Similar to the way in which the area under a density of a random variable determines the probability of a range of values, the area under the spectral density of y_t measures the contribution to the variance of y_t from the cyclical components for a range of frequencies.

Let $f(\omega)$ denote the spectral density of y_t where $\omega \in [-\pi, \pi]$. It can be shown that $f(\omega) \geq 0$ and that $f(-\omega) = f(\omega)$. A fundamental property of $f(\omega)$ is that $var(y_t) \equiv \gamma_0 = \int_{-\pi}^{\pi} f(\omega)\, d\omega$ and the contribution to the variance of y_t from components with frequencies $\omega \in (\omega_1, \omega_2)$ where $-\pi < \omega_1 < \omega_2 < \pi$ is given by $\int_{\omega_1}^{\omega_2} f(\omega)\, d\omega$. Therefore, loosely speaking, frequencies for which $f(\omega)$ takes on large values correspond to cyclical components that make relatively large contributions to the variation in y_t. For those more comfortable thinking in terms of the time domain cycle length, it is easy to convert frequency to cycle length. Suppose $f(\omega_1)$ corresponds to a peak (global or local) of $f(\omega)$; then components with frequencies close to ω_1 make important contributions to the variation in y_t. Consider $\cos(\omega_1 t)$ and $\sin(\omega_1 t)$ and rewrite them as $\cos(2\pi \frac{\omega_1 t}{2\pi})$ and $\sin(2\pi \frac{\omega_1 t}{2\pi})$. Recall that the cosine and sine functions are periodic with period 2π in their argument. Therefore, $\cos(2\pi \frac{\omega_1 t}{2\pi})$ and $\sin(2\pi \frac{\omega_1 t}{2\pi})$ repeat whenever $\frac{\omega_1 t}{2\pi}$ is an integer. Setting $\frac{\omega_1 t}{2\pi} = 1$ indicates that the functions repeat every $t = \frac{2\pi}{\omega_1}$ time periods. The quantity $\frac{2\pi}{\omega_1}$ is called the period corresponding to frequency ω_1.

For a concrete example consider a monthly time series where $f(\omega)$ has a peak at $\omega = \frac{\pi}{6}$. Thus cycles with period $2\pi/(\pi/6)=12$ months (annual cycles) are important for variation of y_t. Suppose quarterly cycles (period is three months) are also important, then $f(\omega)$ will have a peak at $\omega = \frac{2\pi}{3}$. The highest frequency for which we can learn about y_t is $\omega = \pi$ or two-period cycles because cycles that last fewer than two periods do not have data observed within the cycle. This practical limitation on the frequency is called the 'alias effect'. The length of the sample size, T, also limits what can be learned about long cycles. For cycles that last T time periods, that is, frequency $\omega = \frac{2\pi}{T}$, we observe data for exactly one

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

cycle. For cycles longer than T time periods the data does contain information about those cycles, but the information is incomplete because only part of the cycle is observed. It is difficult to learn about very low-frequency components from the data, and in particular it is difficult to learn about $f(0)$.

What does $f(\omega)$ look like? The spectral representation theorem implicitly defines the integral of $f(\omega)$ but not $f(\omega)$ itself. Because the variance of y_t is the area under $f(\omega)$, there is a direct link between $f(\omega)$ and γ_j. Suppose $\sum_{j=-\infty}^{\infty}|\gamma_j| < \infty$, then the spectral density can be expressed as

$$f(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\cos(\omega j)\gamma_j = \frac{1}{2\pi}\left[\gamma_0 + 2\sum_{j=1}^{\infty}\cos(\omega j)\gamma_j\right] \quad (2)$$

where the last expression uses cos(0)=1, cos(−ωj)=cos(ωj) and $\gamma_{-j} = \gamma_j$. It straightforward to show that a converse relationship holds

$$\gamma_j = \int_{-\pi}^{\pi} f(\omega)\cos(\omega j)\, d\omega.$$

This dual relationship between $f(\omega)$ and γ_j makes spectral analysis a powerful analytical tool beyond the direct interpretation of the spectral density in assessing importance of cyclical components.

For the class of stationary autoregressive moving average (ARMA) models $f(\omega)$ takes on a simple form. Let L denote the lag operator, $Ly_t=y_{t-1}$, and define lag polynomials $\varphi(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \ldots - \varphi_p L^p$ and $\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q$. Suppose y_t is a stationary *ARMA*(p, q) process given by $\varphi(L)(y_t - \mu) = \theta(L)\varepsilon_t$ where ε_t is a mean zero uncorrelated time series (white noise process) with $var(\varepsilon_t) = \sigma_\varepsilon^2$. Then

$$f(\omega) = \frac{\sigma_\varepsilon^2\,\theta(e^{i\omega})\,\theta(e^{-i\omega})}{2\pi\varphi(e^{i\omega})\,\varphi(e^{-i\omega})} \quad (3)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $i = \sqrt{-1}$. If the lag polynomials can be factored as $\varphi(L) = (1 - \lambda_1 L)(1 - \lambda_2 L)\ldots(1 - \lambda_p L)$ and $\theta(L) = (1 - \delta_1 L)(1 - \delta_2 L)\cdots(1 - \delta_q L)$, then $f(\omega)$ can be written as

$$f(\omega) = \frac{\sigma_\varepsilon^2 \prod_{j=1}^{q}(1 + \delta_j^2 - 2\delta_j \cos(\omega))}{2\pi \prod_{j=1}^{p}(1 + \lambda_j^2 - 2\lambda_j \cos(\omega))}.$$

In the *AR*(1) case $f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi}\left(1 + \varphi_1^2 - 2\varphi_1 \cos(\omega)\right)^{-1}$. If $\varphi_1 > 0$ (typical for many macroeconomic and finance time series), $f(\omega)$ has a single peak at $\omega=0$. As ω increases, $f(\omega)$ steadily declines. As φ_1 approaches one, the peak at $\omega=0$ increases and sharpens. Thus, variation of autoregressive time series with strong persistence is driven primarily by short frequency/long cycle components. At the other extreme, when the time series is uncorrelated ($\varphi_1 = 0$), the spectral density is constant/flat for all ω, so cyclical components contribute equally at all frequencies to the variation in y_t. An uncorrelated series is called a white noise process because of the analogue to white light which is comprised equally of all visible frequencies of light (all colours).

The special case of $f(0)$ is important for inference in time series models because the asymptotic variance of many time series estimators depends on $f(0)$. For example, consider $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t$, the natural estimator of μ. A simple calculation gives

$$var(\bar{y}) = T^{-1}\left(\gamma_0 + 2\sum_{j=1}^{T-1}\left(1 - \frac{j}{T}\right)\gamma_j\right).$$

If $\sum_{j=-\infty}^{\infty}|\gamma_j| < \infty$, then $\lim_{T\to\infty} T var(\bar{y}) = \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j = 2\pi f(0)$. Therefore, the asymptotic variance of a sample average, often called the long-run variance, is proportional to the spectral density at frequency zero. Inference about the population mean would require a standard error, that is, an estimate of $f(0)$. The link between asymptotic variances and zero frequency spectral densities extends to estimation of linear regression parameters and nonlinear estimation obtained using GMM. The estimation of asymptotic variance matrices that are proportional to a zero frequency spectral density is commonly known as HAC covariance matrix estimation.

Estimates of the spectral density can be obtained either parametrically or nonparametrically. For the case of ARMA models, parametric estimators are straightforward in principle and involve replacing the lag polynomial coefficients in (3) with estimators. Although estimation methods for ARMA models are well established, there are numerical and identification issues that can complicate matters when an MA

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
component is included, especially in the case of vector time series. In contrast, pure AR models are easy to estimate (including the vector case) and, in principle, AR models can well approximate a stationary time series with suitable choice of the lag order. For these reasons, autoregressive spectral density estimators are the most widely used parametric estimators (see, for example, Berk, 1974; Perron and Ng, 1998; den Haan and Levin, 1997). One important practical challenge of implementing autoregressive spectral density estimators is the choice of autoregressive lag order. Advice in the literature on choice of lag order often depends on the intended use of the spectral density estimator.

Nonparametric estimators of the spectral density are appealing at the conceptual level because they do not depend on specific parameterization of the model. In principle nonparametric estimators are flexible enough to provide good estimators for a very wide range of stationary time series. In practice, though, implementation of nonparametric estimators can be a delicate matter, and large sample sizes are required for accuracy. Notwithstanding the practical challenges, nonparametric spectral density estimators are widely used in econometrics primarily because of the central role they play in HAC covariance matrix literature due to the influential contributions by Newey and West (1987) and Andrews (1991). These so-called Newey–West or Newey–West–Andrews standard errors are routinely used in practice; yet many empirical researchers are unaware of the direct link to nonparametric spectral density estimation.

Nonparametric estimators are obtained by estimation of (2) using sample autocovariances

$$\hat{\gamma}_j = T^{-1} \sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y}).$$

The challenge is that $f(\omega)$ depends on an infinite number of autocovariances of which only a finite number can be estimated. The highest-order autocovariance that be estimated is γ_{T-1}, but it is estimated badly because there is only one observation, $(y_T - \bar{y})(y_1 - \bar{y})$. Plugging the $\hat{\gamma}_j$ into (2) gives the estimator

$$I_T(\omega) = \frac{1}{2\pi}\left[\hat{\gamma}_0 + 2\sum_{j=1}^{T-1} \cos(\omega j)\hat{\gamma}_j\right],$$

which is the periodogram. Like $f(\omega)$ the periodogram is non-negative. For $\omega \neq 0$ the periodogram is asymptotically unbiased but its variance does not shrink as the sample size grows, so it is not a consistent estimator. At frequency zero, the situation is even more problematic because simple algebra can used to show that $I_T(0) = 0$. This result holds as long as $\hat{\gamma}_j$ is computed using a quantity that sums to zero (like $y_t - \bar{y}$). Therefore, $I_T(0)$ is useless for estimating $f(0)$. Fortunately, the periodogram can be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

modified to obtain better estimates of $f(\omega)$. Consider

$$\hat{f}(\omega) = \frac{1}{2\pi}\left[\hat{\gamma}_0 + 2\sum_{j=1}^{T-1} k\left(\frac{j}{M}\right)\cos(\omega j)\hat{\gamma}_j\right]$$

where $k(x)$ is a weighting function or kernel such that $k(0)=1$, $k(-x)=k(x)$, $\int_{-\infty}^{\infty} k(x)^2 dx < \infty$ and $|k(x)| \leq 1$. The number M is called the bandwidth or, for some $k(x)$ functions, the truncation lag. The kernel downweights the higher order $\hat{\gamma}_j$, and the bandwidth controls the speed at which downweighting occurs. A recent paper by Phillips, Sun and Jin (2006) achieves downweighting by exponentiating the kernel, for example by using $k(j/T)^{\rho}$, where ρ controls the degree of downweighting.

While a large number of kernel functions have been proposed and analysed since the 1940s, two have become widely used in econometrics: the Bartlett kernel and the quadratic spectral (QS) or Bartlett–Priestley kernel. These kernels are in the class of kernels that guarantee $\hat{f}(\omega) \geq 0$. The Bartlett kernel is

$$k(x) = \begin{cases} 1 - |x| & \text{for} |x| \leq 1 \\ 0 & \text{for} |x| > 1 \end{cases},$$

and it puts linearly declining weights on $\hat{\gamma}_j$ up to lag $M-1$ and weight zero on higher lags so that M plays the role of a truncation lag. Consistency of zero frequency Bartlett kernel estimators was established by Newey and West (1987) in a very general setting. The QS kernel is

$$k(x) = \frac{25}{12\pi^2 x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5)\right),$$

and it does not truncate; weight is placed on all $\hat{\gamma}_j$. The weights decline in magnitude as j increases but some weights can be negative. Andrews (1991) showed in a general setting that the QS kernel minimizes the approximate mean square error of $\hat{f}(0)$ for a particular class of kernels.

The idea of downweighting $\hat{\gamma}_j$ is natural and is not merely a technical trick. For any stationary time series, $\lim_{j\to\infty}|\gamma_j| = 0$, therefore downweighting $\hat{\gamma}_j$, or replacing it with zero when j is large, is similar

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

to replacing an unbiased, high variance estimator with a biased, small variance estimator. If γ_j shrinks quickly as *j* increases, the bias induced by downweighting is small. For $\hat{f}(\omega)$ to be a consistent estimator $M \to \infty$ and $M/T \to 0$ as $T \to \infty$. These conditions suggest that while downweighting is required it cannot be too severe. Unfortunately, these conditions do not restrict the value of *M* that can be used for a given sample because, given *T*, any value of *M* can be embedded in a rule that satisfies these conditions. For example, suppose *T*=100. Then the bandwidth rules $M = 10\sqrt{T}$ and $M = 0.2\sqrt{T}$ satisfy the conditions for consistency but yield very different bandwidths of *M*=100 and *M*=2.

The finite sample properties of $\hat{f}(\omega)$ are complicated and depend on *M* and *k*(*x*). Formulas for the exact bias and variance of $\hat{f}(\omega)$ have been worked out by Neave (1971) when $\omega \neq 0$, and by Ng and Perron (1996) for $\omega = 0$. Because the exact formulas are complicated, approximations are often used. The variance can be approximated by

$$\frac{T}{M} var(\hat{f}(\omega)) \approx \begin{cases} V & \text{for } 0 < \omega < \pi \\ 2V & \text{for } \omega = 0, \pi \end{cases} \tag{4}$$

where $V = f(\omega)^2 \int_{-\infty}^{\infty} k(x)^2 dx$. An approximation for the bias was derived by Parzen (1957) and it depends on the behaviour of *k*(*x*) around *x*=0. For the Bartlett and QS kernels the approximate bias formulas are $-\frac{1}{M}\sum_{j=-\infty}^{\infty} |j| \gamma_j$ and $-\frac{18}{125}\pi^2 \frac{1}{M^2} \sum_{j=-\infty}^{\infty} j^2 \gamma_j$ respectively. Under suitable regularity conditions, an asymptotic normality result holds:

$$\sqrt{\frac{T}{M}}\left(\hat{f}(\omega) - E(\hat{f}(\omega))\right) \to^d N(0, V).$$

According to these approximations, the variance is proportional to $f(\omega)^2$ and increases as *M* increases whereas the bias depends on additional nuisance parameters but decreases as *M* increases. These well known results are discussed at length in Priestley (1981) and are the source of commonly held intuition that says that, as *M* increases, bias decreases but variance increases. This intuition is usually valid but only holds for $\hat{f}(0)$ when *M* is small. As *M* increases the relationship between bias/variance and *M* is more complicated, as discussed by Ng and Perron (1996). Recall that if no downweighting of $\hat{\gamma}_j$ is used, then $\hat{f}(0)$ becomes $I_T(0) = 0$. Obviously, this estimator has a large bias and zero variance. As *M* increases, less downweighting is used and, once *M* is large enough, the bias/variance relationship flips

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

with bias increasing and variance decreasing in M.

An asymptotic approximation that can capture this more complex bias/variance bandwidth relationship for $\hat{f}(0)$ can be obtained using fixed-b asymptotics. Suppose $\hat{f}(0)$ is embedded into a sequence of random variables under the assumption that $b=M/T$ is a fixed constant with $b\in(0,1)$. Neave (1970) first used this approach to derive an alternate asymptotic variance formula for $\hat{f}(\omega)$. Let $B(r)$ denote a standard Brownian bridge, that is, $B(r)=W(r)-rW(1)$ where $W(r)$ is a standard Wiener process, and let $\Rightarrow$ denote weak convergence. Under suitable regularity conditions Kiefer and Vogelsang (2005) show that $\hat{f}(0) \Rightarrow f(0)Q(b)$, where

$$Q(b) = \frac{2}{b}\left(\int_0^1 B(r)^2 dr - \int_0^{1-b} B(r+b)B(r)\,dr\right)$$

for the Bartlett kernel and

$$Q(b) = -\frac{1}{b^2}\int_0^1\int_0^1 k''\left(\frac{r-s}{b}\right)B(r)B(s)\,drds$$

for the QS kernel with analogous results for $\omega \neq 0$ obtained by Hashimzade and Vogelsang (2007). Phillips, Sun and Jin (2006) obtain similar results for exponentiated kernels. The fixed-b asymptotic result approximates $\hat{f}(0)$ by the random variable $Q(b)$ which is similar to a chi-square random variable. When $\hat{f}(0)$ is used to construct standard errors of an estimator like $\bar{y}$, fixed-b asymptotics provides an approximation for $t=(\bar{y}-\mu)/\sqrt{2\pi\hat{f}(0)/T}$ of the form $t \Rightarrow W(1)/\sqrt{Q(b)}$. This limiting random variable is invariant to $f(0)$ but depends on the random variable $Q(b)$. Because $Q(b)$ depends on M (through $b=M/T$) and $k(x)$, the fixed-b approximation captures much (but not all) of the randomness in $\hat{f}(0)$. In contrast, the standard approach appeals to a consistency result for $\hat{f}(0)$ to justify approximating $\hat{f}(0)$ by $f(0)$ and t is approximated by a $N(0,1)$ random variable that does not depend on M or $k(x)$. Theoretical work by Jansson (2004) and Phillips, Sun and Jin (2005) has established that the fixed-b approximation for t in the case of Gaussian data is more accurate than the standard normal approximation. Some results for the non-Gaussian case have been obtained by Goncalves and Vogelsang (2006).

The fixed-b approximation provides approximations of the bias and variance of $\hat{f}(0)$ that are polynomials in b. For the Bartlett kernel Hashimzade, Kiefer and Vogelsang (2005) show that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$E(\hat{f}(0) - f(0)) \approx f(0)(-b + \frac{1}{3}b^2)$ and

$$var(\hat{f}(0)) \approx var(f(0)Q(b)) = f(0)^2\left(\frac{4}{3}b - \frac{7}{3}b^2 + \frac{14}{15}b^3 + \frac{2}{9}b^4 - \frac{1}{15b^2}(2b-1)^5 1(b > \frac{1}{2})\right), \quad (5)$$

where $1(b > \frac{1}{2}) = 1$ if $b > \frac{1}{2}$ and 0 otherwise. Because $b=M/T$, the bias and variance of $\hat{f}(0)$ are approximated by high order polynomials in M/T. The leading term in the variance exactly matches the standard variance formula (4). Because of the higher order terms, the fixed-b variance is more closely related to the exact variance. A plot of the variance polynomial would show that as M increases, variance is initially increasing but once M becomes large enough, variance decreases in M. The fixed-b bias can be combined with the Parzen bias to give

$$E(\hat{f}(0) - f(0)) \approx -\frac{1}{M}\sum_{j=-\infty}^{\infty}|j|\gamma_j + f(0)\left(-\frac{M}{T} + \frac{1}{3}\left(\frac{M}{T}\right)^2\right). \quad (6)$$

This combined formula better approximates the behaviour of the exact bias. As M increases, the first term shrinks, but the second and third terms increase in magnitude. Depending on the relative magnitudes of $\sum_{j=-\infty}^{\infty}|j|\gamma_j$ and $f(0)$, bias will be decreasing in M when M is small, but as M increases further bias becomes increasing in M. It is interesting to note that $1/M$ and M/T terms in (6) match the terms in the type of bias approximations used by Velasco and Robinson (2001) in third order Edgeworth calculations.

Bandwidth choices that minimize approximate mean square error (MSE) of $\hat{f}(0)$ were used by Andrews (1991) and Newey and West (1994) where the bias and variance were approximated using only the leading terms in (5) and (6). A simple, closed form, solution is obtained for M that depends on $\sum_{j=-\infty}^{\infty}|j|\gamma_j$ and $f(0)$. Andrews (1991) recommends plugging in parametric estimators of these unknown quantities, whereas Newey and West (1994) recommend using nonparametric estimators. Including the higher order terms provided by (5) and (6) would allow a higher order approximation to the MSE. Given the polynomial structure of (5) and (6) with respect to M, the first order condition to this optimization problem is a high order polynomial in M with coefficients that depend $\sum_{j=-\infty}^{\infty}|j|\gamma_j$ and $f(0)$. Given plug-in estimates, obtaining the value of M that minimizes the approximate MSE would

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
amount to numerically finding the root of a polynomial, which is not difficult. Such an analysis does not appear to exist in the econometrics literature.

While the focus of this article has been on the spectral analysis of a univariate time series, extending the concepts, notation, and estimation methods to the case of a vector of time series is straightforward. A vector of time series can be characterized by what is called the spectral density matrix. The diagonal elements of this matrix are the individual spectral densities. The off-diagonal elements are called the cross-spectral densities. The cross-spectral densities in general can be complex valued functions even when the data is real valued. The cross-spectral densities capture correlation between series and co-movements of series can be characterized in terms of cross-amplitude, phase and coherency, which are real valued functions. Many of the ideas and concepts in the original Granger (1969) causality paper were expressed in terms of cross-spectral densities.

See Also

- generalized method of moments estimation
- heteroskedasticity and autocorrelation corrections
- serial correlation and serial dependence

Bibliography

Andrews, D.W.K. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–54.

Berk, K. 1974. Consistent autoregressive spectral estimates. *Annals of Statistics* 2, 489–502.

Blanchard, O. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.

Cochrane, J.H. 1988. How big is the random walk in GNP? *Journal of Political Economy* 96, 893–920.

den Haan, W.J. and Levin, A. 1997. A practictioner's guide to robust covariance matrix estimation. In *Handbook of Statistics: Robust Inference*, vol. 15, ed. G. Maddala and C. Rao. New York: Elsevier.

Engle, R. 1974. Band spectrum regression. *International Economic Review* 15, 1–11.

Geweke, J. and Porter-Hudak, S. 1983. The estimation and application of long memory time series. *Journal of Time Series Analysis* 4, 221–38.

Goncalves, S. and Vogelsang, T.J. 2006. Block bootstrap puzzles in HAC robust testing: the sophistication of the naive bootstrap. Working paper, Department of Economics, Michigan State University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Granger, C. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.

Hamilton, J.D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.

Hashimzade, N., Kiefer, N.M. and Vogelsang, T.J. 2005. Moments of HAC robust covariance matrix estimators under fixed-*b* asymptotics. Working paper, Department of Economics, Cornell University.

Hashimzade, N. and Vogelsang, T.J. 2007. Fixed-*b* asymptotic approximation of the sampling behavior of nonparametric spectral density estimators. *Journal of Time Series Analysis*.

Hong, Y. 1996. Consistent testing for serial correlation of unknown form. *Econometrica* 64, 837–64.

Jansson, M. 2004. The error rejection probability of simple autocorrelation robust tests. *Econometrica* 72, 937–46.

Kiefer, N.M. and Vogelsang, T.J. 2005. A new asymptotic theory for heteroskedasticity autocorrelation robust tests. *Econometric Theory* 21, 1130–64.

King, R., Plosser, C., Stock, J. and Watson, M. 1991. Stochastic trends and economic fluctuations. *American Economic Review* 81, 819–40.

Neave, H.R. 1970. An improved formula for the asymptotic variance of spectrum estimates. *Annals of Mathematical Statistics* 41, 70–7.

Neave, H.R. 1971. The exact error in spectrum estimates. *Annals of Mathematical Statistics* 42, 901–75.

Newey, W.K. and West, K.D. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.

Newey, W.K. and West, K.D. 1994. Automatic lag selection in covariance estimation. *Review of Economic Studies* 61, 631–54.

Ng, S. and Perron, P. 1996. The exact error in estimating the spectral density at the origin. *Journal of Time Series Analysis* 17, 379–408.

Parzen, E. 1957. On consistent estimates of the spectrum of a stationary time series. *Annals of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mathematical Statistics 28, 329–48.

Perron, P. and Ng, S. 1998. An autoregressive spectral density estimator at frequency zero for nonstationarity tests. *Econometric Theory* 14, 560–603.

Phillips, P.C.B. and Hansen, B.E. 1990. Statistical inference in instrumental variables regression with i (1) processes. *Review of Economic Studies* 57, 99–125.

Phillips, P.C.B. and Perron, P. 1988. Testing for a unit root in time series regression. *Biometrika* 75, 335–346.

Phillips, P.C.B., Sun, Y. and Jin, S. 2005. Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. Working Paper No. 2005-12, Department of Economics, UCSD.

Phillips, P.C.B., Sun, Y. and Jin, S. 2006. Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation. *International Economic Review* 47, 837–94.

Priestley, M.B. 1981. *Spectral Analysis and Time Series*, vol. 1. New York: Academic Press.

Robinson, P. 1991. Automatic frequency domain inference on semiparametric and nonparametric models. *Econometrica* 59, 1329–63.

Stock, J. and Watson, M. 1988. Testing for common trends. *Journal of the American Statistical Association* 83, 1097–107.

Velasco, C. and Robinson, P.M. 2001. Edgeworth expansions for spectral density estimates and studentized sample mean. *Econometric Theory* 17, 497–539.

Watson, M.W. 1993. Measures of fit for calibrated models. *Journal of Political Economy* 101, 1011–41.

How to cite this article

Vogelsang, Timothy J. "spectral analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000201> doi:10.1057/9780230226203.1588

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

speculative bubbles

Miguel A. Iraola and Manuel S. Santos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We review some work on speculative bubbles on asset prices. Some empirical evidence suggests that asset prices fluctuate more than their fundamental values, and so asset pricing bubbles do occur. Empirical tests on bubbles are inconclusive. Also, most commonly used models in macroeconomics and finance can generate bubbles only under rather specialized assumptions.

Keywords

agency problems; arbitrage; asymmetric information; backward induction; behavioural finance; bounded rationality; fiat money; fiscal theory of the price level; fundamental theorem of asset pricing; infinite horizons; liquidity constraints; martingales; Mississippi bubble; noise traders; overlapping generations; present value; public debt; quantity theory of money; South Sea bubble; speculative bubbles; stock price volatility; strategic behaviour; transversality condition; tulipmania

Article

We maintain that a speculative bubble exists if the market price of an asset differs from its fundamental value – the expected present value of the stream of future dividends attached to the asset. In an economy with a finite sequence of trading dates, the fundamental theorem of asset pricing (see Dybvig and Ross, 1987) guarantees that the equilibrium market price of any asset equals its fundamental value. But in some economies with an infinite sequence of trading dates, this result does not hold, and speculative bubbles may arise. An investor might buy an asset at a price higher than its fundamental value if she expects to sell it later on at a higher price – Harrison and Kreps (1978) call this process ‘speculative behaviour’. In general equilibrium models, however, agents take prices as given and trade assets to transfer income across time and states. These models do not contemplate ‘speculative behaviour’ as it is usually understood. Therefore, the term ‘speculative bubble’ may seem inappropriate in some theoretical frameworks. Santos and Woodford (1997) talk broadly about ‘asset pricing bubbles’.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
There have been famous historical examples of sudden asset price increases followed by an abrupt fall as the Dutch ‘tulipmania’ (1634–7), the ‘Mississippi bubble’ (1719–20) and the ‘South Sea bubble’ (1720). Kindleberger (1978) argues that these are examples of bubbles, whereas Garber (2000) provides market-fundamental explanations for these episodes. More recently, we have seen sharp changes in stock and housing markets. The Japanese stock and land prices experienced a sharp rise in the late 1980s and a dramatic fall in the early 1990s. During the ‘technology bubble’, the Nasdaq Composite Index rose by more than 300 per cent between August 1996 and March 2000, and then fell sharply, reaching the August 1996 level in October 2002. This pattern has been especially intense for the Internet-related sector (Ofek and Richardson, 2003).

There is a vast literature following the variance-bound tests proposed by LeRoy and Porter (1981) and Shiller (1981) that finds significant excess volatility of stock prices (see Gilles and LeRoy, 1991, for a survey). The violation of these variance bounds suggests that asset prices are not determined by fundamental values (see Flood and Hodrick, 1990, and Cochrane, 1992 for a discussion). Various tests have been proposed to detect the presence of rational bubbles in asset prices (see Camerer, 1989, and Cuthbertson, 1996, for a survey). But these tests have important shortcomings. Estimating the fundamental values of an asset is usually a complex task. Hence, rejections of the null hypothesis could be due to an incorrect specification of the fundamental value and not necessarily to the existence of a bubble (Flood and Hodrick, 1990). Even in the most famous apparent bubble episodes, some authors have provided a fundamentalist explanation (see, for example, Donaldson and Kamstra, 1996; Pástor and Veronesi, 2006). To avoid the uncertainty associated with the specification of the fundamental value, Diba and Grossman (1988a) develop a test to detect bubbles based on the investigation of the stationary properties of asset prices and dividends. The main drawback of this test, as Evans (1991) shows, is its limited power to detect periodically collapsing bubbles. Given the severe problems in establishing empirically the existence of bubbles, it is of great importance to understand the theoretical conditions under which bubbles may exist.

If all traders are rational, a backward induction argument precludes the existence of bubbles for assets traded at a finite sequence of dates. More specifically, assume that the economy ends at time T, and there is an asset that provides a dividend of d_T at time T. Then the price of the asset at $T–1$ must be equal to the present value of d_T. By backward induction a bubble cannot exist at any point in time t less that T. Hence, a rational bubble begins on the first date of trading. Moreover, in present value terms the size of the bubble must be constant. (This is usually called the martingale property of bubbles.) Diba and Grossman (1988b) argue that negative rational bubbles cannot exist because it would imply that investors expect that the price of the asset will become negative at a finite future date. Tirole (1982) concludes that, in an economy with a finite number of infinitely lived traders, any asset must be valued according to its market fundamental. However, Tirole (1985) shows that under certain circumstances a deterministic overlapping generations economy allows for the existence of bubbles. In infinite-horizon optimization economies, bubbles are not compatible with the transversality condition: the present value of optimal asset holdings must converge to zero. But by definition the discounted price of the asset will converge to the size of the bubble. Hence, either the asset is in zero net supply or the size of the bubble is equal to zero.

Santos and Woodford (1997) explore the existence of asset pricing bubbles in an infinite-horizon

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
competitive framework, allowing for potentially incomplete markets, arbitrary borrowing limits and incomplete participation of agents (this framework considers jointly economies with a finite number of infinitely lived households and overlapping generations economies). They show that the price of any asset in positive net supply must be equal to its fundamental value, provided that the present value of aggregate wealth is finite. This latter condition is satisfied empirically (see Abel et al., 1989) since in industrialized economies the aggregate share of income that goes to capital is greater than the investment rate. Loewenstein and Willard (2000) extend these results to a finite horizon economy where assets are negotiated continuously. Some key conditions underlying the negative results of Santos and Woodford (1997) are rational expectations, symmetric information and competitive behaviour.

This analysis has important implications for monetary theory because it precludes the existence of valued fiat money as a store of wealth in a broad class of economies. Santos (2006) extends these results to an economy with liquidity constraints and proves that these constraints must be binding infinitely often for all agents in the economy. Hence, in his simple model the aggregate value of the money supply must be equal to the value of aggregate output infinitely often. This is in the spirit of the quantity theory of money. On a related matter, the absence of rational bubbles guarantees that the initial real value of public debt is equal to the present value of future net public revenues. This is a necessary condition to establish the validity of the ‘fiscal theory of the price level’ (Sims, 1994; Woodford, 1995).

The presence of bubbles has also been explored in theoretical frameworks with asymmetric information or boundedly rational agents. Allen, Morris and Postlewaite (1993) find necessary conditions for the existence of bubbles in a model with asymmetric information and a finite sequence of trading dates, and provide examples satisfying these conditions. The existence of a bubble is possible because there is private information which is not common knowledge (all agents know that all agents know, and so on, ad infinitum) that the stock price will fall. Everybody realizes that the stock is overpriced but each agent expects to sell at a higher price before the true value becomes publicly known.

Bubbles may appear in the presence of agency problems associated with short-run optimization behaviour. Allen and Gorton (1993) show that for some compensation schemes a manager may purchase a stock with some prospect of capital gains although with certainty the price will fall below its current level at some point in the future. Allen and Gale (2000) develop a model in which intermediation by the banking sector leads to an agency problem that results in asset bubbles. Investors borrow from banks to buy a risky asset, and they can default in the case of low payoffs. Hence risky assets are more attractive, and therefore investors bid up asset prices.

The behavioural finance literature (see Barberis and Thaler, 2003; Shleifer, 2000 for a survey) often assumes that some agents – called noise traders – are not fully rational. In models in which noise traders and rational agents coexist, the price of an asset can deviate from its fundamental value if rational agents are limited in their capacity to eliminate the mispricing. Shleifer (2000) describes bubbles as the interaction between a significant number of positive feedback investors (who buy securities when prices rise and sell when prices fall), and rational arbitrageurs who anticipate the bursting of bubbles. In this framework, rational arbitragers buy initially after a good-news event to increase the price of the asset and to stimulate the demand of the positive feedback traders; later, they undo their position before the bubble explodes. Abreu and Brunnermeier (2003) develop a model in which noise traders coexist with rational arbitrageurs who become aware of the existence of a bubble sequentially. These rational arbitrageurs would like to exit the market just before the bubble bursts, because before bursting the asset

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
displays high capital gains. The bubble can explode for exogenous reasons, or endogenously when a sufficient number of arbitrageurs decide to abandon the market. In this setting some news could facilitate synchronization and, as a consequence, the bursting of the bubble. Scheinkman and Xiong (2003) develop a model in which overconfidence generates disagreements among agents regarding asset fundamentals. They show that the price of an asset can be above its fundamental value.

In summary, asset prices seem rather volatile – more than their fundamental values. By definition this implies the existence of speculative bubbles. Most empirical exercises to detect the presence of bubbles seem inconclusive. The conditions under which general equilibrium models generate bubbles seem rather fragile, since optimizing agents are unwilling to accumulate arbitrary amounts of wealth. Most recent work has explored the existence of bubbles in economies with limited rationality, asymmetric information and strategic behaviour. The main challenge for these approaches is to explain the mechanisms that lead agents to hold overpriced assets. Specifically, if agents accumulate those assets for arbitrary reasons, then these exercises will not be very enlightening.

See Also

- arbitrage pricing theory
- excess volatility tests
- noise traders
- present value

Bibliography

Abel, A.B., Mankiw, N.G., Summers, L.H. and Zeckhauser, R.J. 1989. Assessing dynamic efficiency: theory and evidence. *Review of economic studies* 56, 1–20.

Abreu, D. and Brunnermeier, M.K. 2003. Bubbles and crashes. *Econometrica* 71, 173–204.

Allen, F. and Gale, D. 2000. Bubbles and crises. *Economic Journal* 110, 236–55.

Allen, F. and Gorton, G. 1993. Churning bubbles. *Review of Economic Studies* 60, 813–36.

Allen, F. Morris, S. and Postlewaite, A. 1993. Finite bubbles with short sale constraints and asymmetric information. *Journal of Economic Theory* 61, 206–29.

Barberis, N. and Thaler, R. 2003. A survey of behavioral finance. In *Handbook of the Economics of Finance*, vol. 1, ed. G.M. Constantinides, M. Harris and R.M. Stulz. Amsterdam: North-Holland.

Camerer, C. 1989. Bubbles and fads in asset prices. *Journal of Economic Surveys* 3, 3–41.

Cuthbertson, K. 1996. *Quantitative Financial Economics: Stocks, Bonds and Foreign Exchange*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Chichester: Wiley & Sons Ltd.

Cochrane, J. 1992. Explaining the variance of price dividend ratios. *Review of Financial Studies* 5, 243–80.

Diba, B.T. and Grossman, H.I. 1988a. Explosive rational bubbles in stock prices. *American Economic Review* 78, 520–30.

Diba, B.T. and Grossman, H.I. 1988b. The theory of rational bubbles in stock prices. *Economic Journal* 98, 746–54.

Donaldson, R.G. and Kamstra, M. 1996. A new dividend forecasting procedure that rejects bubbles in asset prices: the case of 1929's stock crash. *Review of Financial Studies* 9, 333–83.

Dybvig, P.H. and Ross, S.A. 1987. Arbitrage. In *The New Palgrave: A Dictionary of Economics*, vol. 1, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Evans, G.W. 1991. Pitfalls in testing for explosive bubbles in asset prices. *American Economic Review* 81, 922–30.

Flood, R.P. and Hodrick, R.J. 1990. On testing for speculative bubbles. *Journal of Economic Perspectives* 4(2), 85–101.

Garber, P.M. 2000. *Famous First Bubbles*. Cambridge, MA: MIT Press.

Gilles, C. and LeRoy, S. 1991. Econometric aspects of the variance-bounds tests: a survey. *Review of Financial Studies* 4, 753–91.

Harrison, J.M. and Kreps, D.M. 1978. Speculative investor behavior in a stock market with heterogeneous expectations. *Quarterly Journal of Economics* 92, 323–36.

Kindleberger, C.P. 1978. *Manias, Panics and Crashes: A History of Financial Crises*. New York: Basic Books.

LeRoy, S.F. and Porter, R.D. 1981. The present value relation: test based on implied variance bounds. *Econometrica* 49, 555–74.

Loewenstein, M. and Willard, G.A. 2000. Rational equilibrium asset-pricing bubbles in continuous trading markets. *Journal of Economic Theory* 91, 17–58.

Ofek, E. and Richardson, M. 2003. DotCom mania: the rise and fall of internet stock prices. *Journal of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Finance 58, 1113–37.

Pástor, L. and Veronesi, P. 2006. Was there a Nasdaq bubble in the late 1990s? *Journal of Financial Economics* 81, 61–100.

Santos, M.S. 2006. The value of money in a dynamic equilibrium model. *Economic Theory* 27, 39–58.

Santos, M.S. and Woodford, M. 1997. Rational asset pricing bubbles. *Econometrica* 65, 16–57.

Scheinkman, J.A. and Xiong, W. 2003. Overconfidence and speculative bubbles. *Journal of Political Economy* 11, 1183–219.

Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71, 421–36.

Shleifer, A. 2000. *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford: Oxford University Press.

Sims, C.A. 1994. A simple model of the determination of the price level and the interaction of monetary and fiscal policy. *Economic Theory* 4, 381–99.

Tirole, J. 1982. On the possibility of speculation under rational expectations. *Econometrica* 50, 1163–81.

Tirole, J. 1985. Asset bubbles and overlapping generations. *Econometrica* 53, 1499–528.

Woodford, M. 1995. Price level determinacy without control of a monetary aggregate. *Carnegie-Rochester Conference Series on Public Policy* 43, 1–46.

How to cite this article

Iraola, Miguel A. and Manuel S. Santos. "speculative bubbles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000205> doi:10.1057/9780230226203.1589

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Spence, A. Michael (born 1943)

Edward Lazear
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Michael Spence is a pioneer in the economics of information. His most famous work is on signalling. Spence's fundamental insight is that individuals can take actions that provide information to others, even though the actions themselves have no effect on productivity or on that which is desired by the buyer. This work has proven fundamental to understanding phenomena ranging from education to advertising.

Keywords

adverse selection; advertising; Akerlof, G.; asymmetric information; education; information economics; productivity; R&D subsidies; signaling; Spence, A. M.; statistical discrimination

Article

Andrew Michael Spence was awarded the Nobel Prize in 2001 for his work on information economics. His pioneering doctoral dissertation on signalling formed the basis of an enormous subsequent literature in the economics of information. Spence's insight is that individuals can take actions that provide information to others, even though the actions themselves have no effect on productivity or on that which is desired by the buyer.

In Akerlof's classic 1970 paper on lemons ‘The Market for “Lemons”: Quality Uncertainty and the Market Mechanism’, a market characterized by adverse selection is described where goods that are put on the used market are not a random selection of all goods. Spence asked whether there were actions that could be taken that might identify one type of seller from another in a market for lemons. For example, sellers who knew that their goods were particularly high-quality might offer warranties, whereas those who were concerned about the quality of their products would refrain from doing so.

This simple insight stimulated Spence to create an entire new way of thinking about information in markets. His dissertation, which, published in 1974, became one of the most influential of the second half of the 20th century, is a work that is well ahead of its time, incorporating complexity of equilibrium concepts, game-theoretic strategies and notions of multiplicity that were not to be fleshed out until a number of years later.

Education as signalling

Spence's classic paper ‘Job Market Signaling’ (1973a) considers whether education might be used merely as a signal of worker quality rather than as a tool to enhance productivity. There are two assumptions behind it. First, information is asymmetric. Workers know their own productivity, but firms do not. Second, there is a negative correlation between the cost of acquiring education and worker productivity. That is, the individuals who are most productive in the labour market are also those who find it cheapest to acquire education. The easiest way to think of this is that the most able people can pass an exam with fewer hours of study than the least able people.

Spence shows that an equilibrium exists where high-ability types acquire more schooling than low-ability types, even though schooling has no inherent effect on productivity. The simplest analysis is shown in Figure 1.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Figure 1

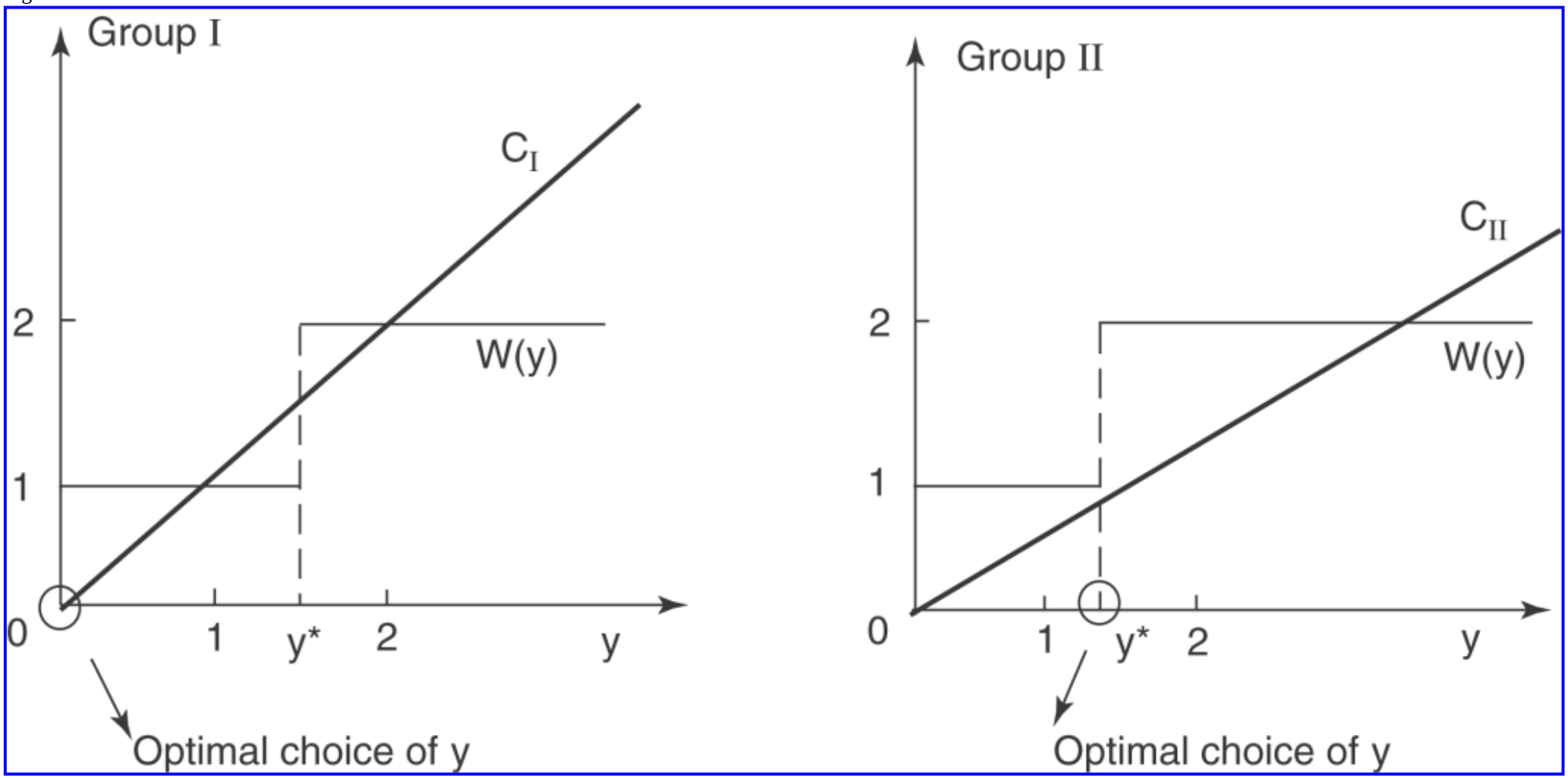

Suppose that there are two groups, the high-ability group with costs shown by C_{II} and a low ability group shown C_I. Suppose further that the employers make the assumption that individuals who have y^* or more education are high-ability, whereas those with $<y^*$ education are of low ability. The low-ability group has a cost of schooling equal to y, whereas the high-ability group has a cost of schooling equal to $y/2$. The productivity of high-quality workers is 2, whereas the productivity of low-quality workers is 1. Spence shows that the employers' expectations will be validated in equilibrium.

First, consider a high-ability group. If workers are deemed to be high-ability, then their output and wage is 2. If workers are deemed to be low-ability, then their output and wage is 1. Low-productivity individuals, whose cost is given by C_I, find it best to accept the low wage and to invest in zero units of schooling, as shown by the point at the origin in the diagram labelled group I. But because high-ability individuals have sufficiently low costs, it is profitable for them to acquire y^* of schooling because $2(-y^*/2) > 1$.

The assumptions behind this model lead immediately to predictions on which signals may survive in the market. In order for a signal to be effective, it must be correlated with productivity, and also have different costs across high-ability and low-ability groups. So, for example, skill in origami is rarely suggested as a signal in labour markets because origami prowess is unlikely to be highly correlated with productivity in most jobs.

The basic insight that comes out of the signalling structure can be extended. In the job market signalling paper itself, Spence lays out an early version of the statistical discrimination argument, where male and female workers may end up with different levels of education, but the signal that a given level of education conveys in the male population may be quite different from that conveyed by the same level of education in the female population. This results from multiple equilibria in the basic signalling model, a point that Spence makes clear even in his very early analysis.

When Spence presented his Nobel lecture in 2001 (Spence, 2002), he revisited the signalling analysis and discovered that differential costs were unnecessary. Indeed, it is possible to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

have costs that are the same across two groups but benefits that differ, such as the case in the labour market setting where some workers choose to be measured and are paid piece rates and other workers choose not to be measured and are paid straight salaries. Those workers who choose to be measured are of higher ability (Lazear, 1986).

Additionally, a large literature on advertising traces its roots to the Spence signalling analysis. Certain firms choose to bear the cost of advertising because the value to them is higher than the value to other firms. For example, a firm with a very good product benefits more by having information about its product available to the public than a firm with a bad product. While both types of products may benefit in the short run from signalling through advertising, the firm with the better product can enjoy longer-run gains, and therefore higher return to the information conveyed by advertising.

In an extension ‘Time and Communication in Economic and Social Interaction’ (1973b) Spence considers how willingness to spend time acts as a signal. When a political leader attends an event at which he has no role, he signals that he values the organization that has sponsored the event. Indeed, the fact that he has no role makes the signal stronger. Another example involves pricing sporting events at less than the market-clearing level, while using time spent in a queue as an allocative mechanism. Although the wasted time creates deadweight loss, it may be that time is a better signal of interest in the sporting event than is the money that an individual is willing to spend on the tickets. To the extent that the quality of the event depends on the active participation of the fans, it may be optimal to use time in addition to money to allocate tickets to potential buyers.

Time operates differently from other costs in the traditional signalling model that applied in the job market. The high cost of time actually may preclude using education as a signal. If the most able individuals have the most valuable use of time in other (unobservable) activities, then the correlation between cost of schooling and productivity may be reversed. The highly productive individuals in the labour market would also be the ones with the high cost of schooling because the time that they allocate to schooling is so valuable. In this case schooling would not be an effective signal. Spence works out situations where the schooling may be an adverse signal of productivity in his 2001 Nobel address.

Spence notes in his early signalling paper that signalling creates a divergence between social and private value. While individuals may have an incentive to signal and firms may have an incentive to pay for that signal, the signal in the purest case reflects wasted resources. In the simplest version, where information has no sorting or allocative role, individuals who acquire education receive higher earnings, and those who do not receive lower earnings. The market equilibrium has higher inequality of income than one without signalling, but no higher output in the labour market. Because signalling is in itself unproductive, resources devoted to schooling are wasteful. A law that said ‘firms are precluded from paying on the basis of education’ would be welfare enhancing.

Spence after signalling

Spence's interest in the economics of information has motivated him to consider issues that arise in product markets. Specifically, he questions whether the information that is present in a market provides appropriate signals to producers and consumers, or whether those signals would lead to inefficient outcomes under certain circumstances. Information, he reasons, is an important component of markets that provides incentives and alters behaviour. One of the first questions that Spence considers is whether the information that is transmitted by the market induces firms to produce the correct quality of a good. Spence (1975) shows that, as a general matter, the market produces the wrong quality, although not necessarily a quality that is too high or too low. Price is determined by the marginal consumer, but the rents associated with the particular quality of a good depend on the area under the demand curve up to the quantity produced, and not just the value that the marginal consumer places on it. If the marginal consumer's valuation of higher quality is lower than the average consumer's, then the market underproduces quality. If the marginal consumer's valuation is higher than the average consumer's, then the market overproduces quality. (A related idea is explored in Spence and Owen, 1977.)

Spence's forays into quality and product choice have led him to consider more general questions relating to the interaction between information and industry structure. In ‘Cost Reduction, Competition, and Industry Performance’ (1981a), Spence examines the trade-off between a duplication of R&D expenditures and having the competitive industry that makes the market statically efficient. If there is only one firm, then that firm will charge a monopoly price. However, if there is a multiplicity of firms, then each firm must do its own R&D when R&D is appropriable. The output is closer to the competitive equilibrium with more firms than with fewer. The optimal degree of competition trades off replication of R&D expenditures against competitive outcomes. R&D in this world creates natural monopoly-like characteristics for the industry. Just as it is inefficient to duplicate electrical lines that go from power plants to individual houses, it is also inefficient to force every firm to do its own R&D. One solution discussed is to have an R&D consortium with appropriate subsidies. The firms in the consortium can share the information generated, and the subsidies may provide enough incentives to undertake efficient R&D activity.

A related idea is explored in the ‘The Learning Curve and Competition’ (1981b). The learning curve makes costs a function of past output. The more firms there are in an industry, the less output that any single firm produces. Each firm learns more slowly and costs are higher when there is a large number of firms than when there is a small number of firms. But the advantage of a large number of firms is that output is closer to the competitive level. Unfortunately, the competitive level is at a higher cost than if output were produced by one firm alone. This trade-off is analysed and a market structure is suggested.

Michael Spence as a person

An American by birth, Spence was born in New Jersey, when his father, a Canadian, was spending time on a project in Washington during the Second World War. His mother, also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

an American by birth, was raised in Winnipeg. The Spence family moved to Toronto from rural Canada when Mike was nine years old. Mike lived in Toronto until he went to Princeton in 1962. Having grown up in Canada, Mike acquired important skills, in particular the ability to play hockey. An excellent athlete, Spence played hockey for four years while at Princeton. He graduated summa cum laude and went on to be a Rhodes Scholar, attending Oxford University, where he studied mathematics. After getting his training at Oxford, he enrolled for a Ph.D. at Harvard. His most important mentors were Kenneth Arrow, Thomas Schelling and Richard Zeckhauser. In 1972 he completed his landmark thesis on signalling and was awarded a Ph.D. Mike taught at the Harvard's J. F. Kennedy School of Government for a couple of years. In 1973 he moved to Stanford University's economics department, but was lured back to Harvard in 1976. After a few more years doing research, Mike became the Dean of the Faculty of Arts and Sciences in 1984, and stayed in that position until he moved back to Stanford in 1990, when he took over as Dean of the Graduate School of Business. He retired from that position in 1999.
Mike's output in family life has been as high-quality as his output in research. He has three children. Graham, born in 1979, graduated from Princeton; Catherine, born in 1982, graduated from Columbia; and Marya, born in 1985, enrolled at Harvard in 2003. Michael Spence, a great friend and exceptional individual, excels at everything that he attempts. Mike's hobbies include windsurfing and motorcycle riding.

See Also

- Akerlof, George Arthur
- signalling and screening
- Stiglitz, Joseph E.

Selected works

1973a. Job market signaling. *Quarterly Journal of Economics* 87, 355–74.

1974. *Market Signaling: Informational Transfer in Hiring and Related Processes*. Cambridge, MA: Harvard University Press.

1973b. Time and communication in economic and social interaction. *Quarterly Journal of Economics* 87, 651–60.

1975. Monopoly quality and regulation. *Bell Journal of Economics* 6, 417–29.

1977. (With B. Owen.) Television programming, monopolistic competition, and welfare. *Quarterly Journal of Economics* 91, 103–26.

1981a. Cost reduction, competition, and industry performance. *Econometrica* 52, 101–22.

1981b. The learning curve and competition. *Bell Journal of Economics* 12, 49–70.

2002. Signaling in retrospect and the informational structure of markets. *Les Prix Nobel 2001*. Stockholm: The Nobel Foundation. Online. Available at http://nobelprize.org/economics/laureates/2001/spence-lecture.pdf, accessed 10 June 2006.

Bibliography

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Lazear, E. 1986. Salaries and piece rates. *Journal of Business* 59, 405–31.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Lazear, Edward. "Spence, A. Michael (born 1943)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000452> doi:10.1057/9780230226203.1590

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Spencer, Herbert (1820 – 1903)

M. Donnelly
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

evolutionism; social Darwinism; Spencer, H.

Article

Though largely ignored today, Herbert Spencer was one of the most influential scientists and philosophers of the late 19th century. He was born at Derby to a family of Dissenters, and educated at home. As a young man he worked on the London & Birmingham Railway, acquiring considerable practical knowledge of civil engineering and, through his observation of railway cuts, expertise in geology. He had no university training, but read extremely widely in an array of fields.

He was an early and enthusiastic partisan of Darwin and of evolutionist ideas. In 1860 he published a prospectus for a ‘system of synthetic philosophy’, a general compendium of knowledge, which was to occupy him for much of the rest of his life. He set out to survey, from the ‘evolutionary point of view’, the fields of biology, psychology, sociology and ethics, publishing in turn *First Principles* (1862); *Principles of Biology* (2 vols, 1864–7); *Principles of Sociology* (3 vols, 1876–96) and *Principles of Ethics* (2 vols, 1879–93).

His most important work bearing on social policy was the polemical *The Man Versus the State* (1884). Spencer was a highly vocal champion of social Darwinism, applying the principle he termed ‘survival of the fittest’ to a broad variety of struggles, including economic competition. He conceived society by analogy to an organism, arguing that it developed according to immanent processes of growth, and hence that the positive actions and interventions of politicians were likely to be harmful or superfluous.

Selected works

1862. *First Principles*. London.

1864–7. *The Principles of Biology*, 2 vols. London: Williams & Norgate.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1870–72. *The Principles of Psychology*, 2 vols. 2nd edn. London: Williams & Norgate. (First published 1855.)

1876–96. *Principles of Sociology*, 3 vols. London: Williams & Norgate.

1879–93. *The Principles of Ethics*, 2 vols. London: Williams & Norgate.

1884. *The Man Versus the State*. Indianapolis: Liberty Classics, 1982.

Bibliographic addendum

Offer, J., ed. 2000. *Herbert Spencer: Critical Assessments*. London: Routledge, 2000.

Weinstein, D. 1998. *Equal Freedom and Utility: Herbert Spencer's Liberal Utilitarianism*. Cambridge: Cambridge University Press.

How to cite this article

Donnelly, M. "Spencer, Herbert (1820–1903)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000209> doi:10.1057/9780230226203.1591

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Spiethoff, Arthur August Kaspar (1873 – 1957)

Ingo Barens
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; German Historical School; investment; Juglar, C.; Kondratieff cycles; methodology of economics; Spiethoff, A. A. K.

Article

Spiethoff was born on 13 May 1873 in Düsseldorf and died on 4 April 1957 in Tübingen. He was a student of Adolph Wagner and research assistant to Gustav Schmoller. In 1908 he was appointed professor at Prague University and from 1918 held a chair of political economy at Bonn University until his retirement in 1939. He was the long-time editor of *Schmollers Jahrbuch* and (with Edgar Salin) of *Hand- und Lehrbücher aus dem Gebiet der Sozialwissenschaften*.
Spiethoff is best known for his path-breaking research into business cycles, as well as for his studies on methodology, culminating in his concept of ‘economic style’ (*Wirtschaftsstil*). In his methodological studies Spiethoff, strongly influenced by the German Historical School, sought for a solution to the antinomy of history and theory: the quest for generalizing statements about an ever-changing reality. He stressed the distinction between two methods of inquiry: pure economic theory (brought to perfection by Quesnay, Ricardo, Thünen, Menger, Jevons and Pareto) and ‘observational’ (*anschauliche*) or ‘economic Gestalt theory’ (in the tradition of the mercantilists, List, Sombart and Schmoller). Pure economic theory, whether or not it exclusively deals with timeless phenomena, such as those common to all forms of economic life, abstracts and isolates arbitrarily, depending on the particular purpose in view. ‘Observational theory’, on the other hand, takes its time-conditioned data from the real world and abstracts only from their historical uniqueness to isolate the regular and essential features. It thus yields an ‘explanatory description’ – that is, an effigy, or replica, of reality – ‘purged of historical accidents’ (Spiethoff, 1953b, p. 76). With its findings derived from time-conditioned data, ‘economic Gestalt theory’ is a ‘historical’ theory, the validity and applicability of its generalizations dependent on the existence and dominance of a certain ‘economic style’, representing uniformities of economic life in a certain historical epoch (for example, the ‘economic styles’ of medieval town economy, of free market capitalism or of interventionism). Spiethoff's ultimate aim was an all-embracing general economic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
theory that would include as many different ‘historical’ theories as there are ‘economic styles’, together with the pure theory of timeless phenomena.

The foremost field of application of Spiethoff's methodological approach has always been his research on business cycles. In his writings (see Schweitzer, 1941, for a comprehensive bibliography), starting from the work of Clément Juglar, he emphasized three points: first, the necessity not to focus exclusively on crisis or overproduction, but instead to visualize the phenomenon of cyclical fluctuations as an entity; second, the strategic role to be ascribed to capital investment in the explanation of business cycles; and third, the fact that booms and depressions should not be considered as merely an accidental and insignificant concomitant of economic activity but must be understood as the essential form of capitalist life itself. This basic perception made him one of the founding fathers of modern business cycle research.

In keeping with this notion of ‘time-conditioned’ theory, Spiethoff considered his findings to be valid only for a certain ‘economic style’, representing an age marked by the prevalence of a highly developed capitalist economy and a free-market system. This era lasted from 1820 to 1913, with the capitalist economy not yet fully developed in earlier periods, while increasingly becoming subject to manipulation, planning and management in later times. Spiethoff's striving for ‘historical’ generalization took the form of distilling a ‘typical cycle’ to give account of the recurrent and essential features of all historically known business cycles. This ‘typical cycle’, now generally accepted, consists of three ‘cyclical stages’ (upswing, crisis and downswing), two of which may be subdivided into five ‘cyclical phases’: the downswing comprises the recession phase with investment declining and a ‘first revival’ during which the decline in investment is halted, while the upswing includes the ‘second revival’ with rapidly increasing investment, the boom phase characterized by rising interest rates and, finally, ‘capital scarcity’ with declining investment paving the way for the next downturn.

With his observation of ‘cyclical periods’, during which years of either boom or depression preponderate, Spiethoff in fact anticipated what later would become known as the Kondratieff cycle or ‘long wave’.

Selected works

1933. Overproduction. In *Encyclopaedia of the Social Sciences*, vol. 11. New York: Macmillan.

1952. The ‘historical’ character of economic theories. *Journal of Economic History* 12, 131–9.

1953a. Pure theory and economic Gestalt theory: ideal types and real types. In *Enterprise and Secular Change: Readings in Economic History*, ed. F.C. Lane and J.C. Riemersma. London: George Allen & Unwin.

1953b. Business cycles. *International Economic Papers* 3, 75–171. Originally published in German as ‘Krisen’, in *Handwörterbuch der Staatswissenschaften*, vol. 6, Jena: G. Fischer, 1923.

1955. *Die wirtschaftlichen Wechsellagen: Aufschwung, Stockung, Krise*. 2 vols. Zurich: Polygraph.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Clausing, G., ed. 1933. *Der Stand und die nächste Zukunft der Konjunkturforschung. Festschrift für Arthur Spiethoff*. Berlin: Duncker & Humblot.

Clausing, G. 1958. Arthur Spiethoffs wissenschaftliches Lebenswerk. *Schmollers Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft* 78, 257–90.

Lane, F.C. 1956. Some heirs of Gustav von Schmoller. In *Architects and Craftsmen in History: Festschrift für Abbott Payson Usher*, ed. J.T. Lombie. Tübingen: Mohr.

Lane, F.C. and Riemersma, J.C. 1958. Introduction to Arthur Spiethoff. In *Enterprise and Secular Change: Readings in Economic History*, ed. F.C. Lane and J.C. Riemersma. London: George Allen & Unwin.

Schweitzer, A. 1941. *Spiethoff's Theory of the Business Cycle*, vol. 8. Laromie: University of Wyoming Publications (contains an extensive bibliography).

How to cite this article

Barens, Ingo. "Spiethoff, Arthur August Kaspar (1873–1957)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000212> doi:10.1057/9780230226203.1592

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spline functions

Dale J. Poirier
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Spline functions are smooth piecewise functions that are popular tools in approximation theory and which arise naturally in economics.

Keywords

least squares; linear regression models; maximum likelihood estimation; New Jersey Income-Maintenance Experiment; nonparametric regression; spline functions; structural change

Article

In the everyday use of the word, a ‘spline’ is a flexible strip of material used by draftsmen in the same manner as French curves to draw a smooth curve between specified points. The mathematical *spline function* is similar to the draftsman's spline. It has roots in the aircraft, automobile and shipbuilding industries. Formally, a spline function is a piecewise continuous function with a specified degree of continuity imposed on its derivatives. Usually the pieces are polynomials. The abscissa values, which define the segments, are referred to as ‘knots’, and the set of knots is referred to as the ‘mesh’.

The terminology and impetus for most contemporary work on spline functions can be traced to the seminal work of I.J. Schoenberg (1946), although the basic idea can be found in the writings of E.T. Whittaker (1923) and, in Schoenberg's (1946, p. 68) own modest opinion, in the earlier work of Laplace. Today the literature on spline functions comprises an integral part of modern approximation theory. Useful monographs covering splines are De Boor (2001), Eubank (1988), Green and Silverman (1994), Poirier (1976), Schumaker (1981), and Wahba (1990). The many important contributions of Grace Wahba in the 1970s and 1980s (for example, Kimeldorf and Wahba, 1970; Wahba, 1978; 1983) united the approximation theory and the emerging statistics literatures involving spline functions.

Given a degree *d* and a knot vector $t = [t_1, t_2, \ldots, t_K]'$, where $t_1 < t_2 < \ldots < t_K$, the collection of polynomial splines having *s* continuous derivatives forms a linear space. For example, the collection of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

linear splines with knot sequence *t* is spanned by the functions

$$1, x, (x - t_1)_+, \ldots, (x - t_K)_+.$$

where $(\cdot)_+ = \max(\cdot, 0)$. This set is called the *truncated power basis* of the space. In general, the basis for a spline space of degree *d* and smoothness *s* is made up of monomials up to degree *d* together with terms of the form $(x - t_k)_+^{s+j}$, where $1 \leq j \leq d - s$. For example, cubic splines have $d = 3$ and $s = 2$ so that the basis has elements

$$1, x, x^2, x^3, (x - t_1)_+^3, \ldots, (x - t_K)_+^3.$$

Unfortunately, these truncated power functions have poor numerical properties. For example, in linear regression problems the condition of the design matrix deteriorates rapidly as the number of knots increases. A popular alternative representation is the so-called *B-spline basis* (see De Boor. 2001). These functions are constructed to have support only on a few neighbouring intervals defined by the knots. The importance of spline functions in approximation theory is explained by the following *best approximation property*. Consider the data points $(x_i, y_i)\ (i = 1, 2, \ldots, n)$ and suppose without loss of generality that $0 < x_1 < x_2 < x_3 < \ldots < x_n < 1$. Given $\lambda > 0$, consider the optimization problem

$$\min_{f(\cdot)} \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_0^1 [D^m f(x)]^2 dx, \tag{1}$$

where D^m denotes the differentiation operator of degree *m*, $f(\cdot)$ is a function defined on [0, 1] such that $D^j f,\ j \leq m - 1$, is absolutely continuous, and $D^m f$ is in the set of measurable square integrable functions on [0, 1]. The first term in (1) comprises the familiar least squares measure of fit and the second term comprises a measure of the smoothness in $f(\cdot)$. The parameter λ measures the trade-off between fit and smoothness. The solution to (1) is a polynomial *smoothing spline* of degree $2m - 1$ with knots at all the abscissa data points. As $\lambda \to 0$, the solution is referred to as an *interpolating spline* and it fits the data exactly. The choice of λ is crucial and the method of cross-validation is a popular method for choosing λ (see for example Craven and Wahba, 1979, or Green and Silverman, 1994, pp. 30–8). The most popular choice for m is $m = 2$ yielding a *natural cubic spline* as the solution to (1). The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
adjective ‘natural’ implies that the second derivative equals zero at the endpoints.
To interpret the first term in (1) as the log-likelihood of normal linear regression model, smoothing splines can be viewed as the outcome of penalized (reflected in the second term in (1)) maximum likelihood estimation. A Bayesian interpretation of smoothing splines, provided by Kimeldorf and Wahba (1970) and expanded by Silverman (1985) and Wahba (1978; 1983), views (1) as a log-posterior density with $\exp[-\frac{1}{2}\lambda\int[D^m f(x)]^2 dx]$ serving as a prior density over the space of all smooth functions.
In econometrics spline functions are most often employed to parametrize a regression function. For example, splines were the functional form chosen to paremetrize the treatment in the first major social experiment in economics: the New Jersey Income-Maintenance Experiment. Such *regression splines* usually include only a few knots and not necessarily at the design points. This usage may simply reflect the flexibility and good approximation properties of splines, or the attempt to capture *structural change*. For example, a researcher may believe the relationship between two variables *y* and *x* is locally a polynomial, but that at precise points in terms of *x* the relationship ‘changes’, not in a discontinuous fashion in level but rather continuously in derivative of order $2m-1$. Common choices for such *x* variables are time, age, education or income, to name a few, with a nearly unlimited number of choices of candidates for *y* variables.
In statistics spline functions are used in isotonic regression, histogram smoothing, density estimation, interpolation of distribution functions for which there is no closed-form analytic representation, and nonparametric regression. In the latter case spline smoothing corresponds approximately to smoothing by a kernel method with bandwidth depending on the local density of design points.
While spline functions have proved to be valuable approximation tools, they also arise naturally in their own right in economics. Income tax functions with increasing marginal tax rates constitute a *linear spline*, as do familiar ‘kinked’ demand curves and ‘kinked’ budget sets. *Quadratic splines* serve as useful ways of generating asymmetric loss functions for use in decision theory. In distributed lag analysis, spline functions have been used as natural generalizations of Almon polynomial lags. *Periodic cubic splines* have proved useful in seasonal adjustment and in analysis of electricity load curves. Spline functions in these applications are attractive partly because, given the knots, the spline can be expressed as linear functions of unknown parameters, hence facilitating statistical estimation.
Knots play different roles in the approximation theory and structural change literatures. In the former they are largely nuisance parameters, and, apart from parsimony considerations, the number and location of the knots are of no particular importance other than that they serve to define a smooth best-fitting curve. When viewed as *change points*, however, the knots become parameters of interest. In applications involving structural change, the number of potential knots is small, and their location reflects subject-matter considerations. For example, when fitting a time trend with a regression spline, the knots may reflect the effect on the dependent variable of a specific event of interest – for example, a war. A prior distribution can then be specified over the interval bounded by the start and end of the war.
Estimation of the number and location of the knots is hindered by numerical and statistic complications. The knots enter spline functions nonlinearly, and there are typically numerous local minima in the residual sum-of-squares surface. Many of these local minima correspond to knot vectors with *replicate knots*, that is, knots which pile up on top of each other, signalling that further discontinuities in the derivatives of the function are required. When knot locations are set free, knots move to areas where the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
function is less smooth. If, in addition, the *number* of knots is unknown, the difficulties multiply. For example, under the null hypothesis that adjacent intervals are identical, the location of the unnecessary knot is unidentified.

Different solutions have emerged to the problem of unknown location and number of knots. Some introduce a large number of potential knots from which a subset is to be selected (for example, Halpern, 1973; Friedman and Silverman, 1989; Smith and Kohn, 1996). The problem then becomes one of variable selection where each knot corresponds to a column of a design matrix from which a 'significant' subset is to be determined. In some Bayesian nonparametric regression studies (for example, DiMatteo, Genovese and Kass, 2001; Smith, Wong and Kohn, 1998; Denison, Mallick and Smith, 1998), knot locations are treated as parameters and given prior distributions. Additional constraints are usually imposed to keep knots some minimum distance apart. The definitive treatment of the problem of unknown location and number of knots has not yet emerged.

Early applications of splines to *multivariate* problems (see Green and Silverman, 1994, ch. 7) involved tensor product spaces that of necessity depended on the choice of coordinate system. An example is the two-dimensional *thin plate spline* of Wahba (1990) which simulates how a thin metal plate would behave if forced through some control points. This is similar to the one-dimensional draftsman's spline. The tensor product structure of these spaces implicitly defines the domain of an unknown function to be a hyperrectangle, and this can restrict the ability to capture important features in the data that are not oriented along one of the major axes. There is a considerable literature on constructing and representing smooth, piecewise polynomial surfaces over meshes in many variables. In particular, much has been written about the case in which the underlying partition consists of triangles or high-dimensional simplicies. Because of their invariance to affine transformations, *barycentric coordinates* (that is, coordinates expressed as weighted combinations of the vertices of the triangle) are used to construct spline spaces over such meshes. The *triogram* methodology of Hansen, Kooperberg and Sardy (1998) employs continuous, piecewise linear (planar) bivariate splines defined over adaptively selected triangulations in the plane. Analogous to stepwise knot addition and deletion in a univariate spline space, the underlying triangulation is constructed adaptively by adding and deleting vertices.

See Also

- statistical inference
- statistics and economics
- structural change
- two-stage least squares and the *k*-class estimator

Bibliography

Craven, P. and Wahba, G. 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of cross-validation. *Numerische Mathematik* 31, 377–403.

De Boor, C. 2001. *A Practical Guide to Splines*. New York: Springer-Verlag.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Denison, D., Mallick, B. and Smith, A. 1998. Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B* 60, 333–50.

DiMatteo, I., Genovese, C. and Kass, R. 2001. Bayesian curve-fitting with free knot splines. *Biometrika* 88, 1055–71.

Eubank, R. 1988. *Spline Smoothing and Nonparametric Regression*. New York: Marcel-Dekker.

Friedman, J. and Silverman, B. 1989. Flexible parsimonious smoothing and additive modeling. *Technometrics* 31, 3–21.

Green, P. and Silverman, B. 1994. *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Halpern, E. 1973. Bayesian spline regression when the number of knots is unknown. *Journal of the Royal Statistical Society: Series B* 35, 347–60.

Hansen, M., Kooperberg, C. and Sardy, S. 1998. Triogram models. *Journal of the American Statistical Association* 93, 101–19.

Kimeldorf, G. and Wahba, G. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics* 41, 495–502.

Poirier, D. 1976. *The Econometrics of Structural Change with Special Emphasis on Spline Functions*. Amsterdam: North-Holland.

Schoenberg, I. 1946. Contributions to the problem of approximation of equidistant data by analytic functions: Parts I and II. *Quarterly Journal of Applied Mathematics* 4, 45–99; 112–41.

Schumaker, L. 1981. *Spline Functions: Basic Theory*. New York: Wiley.

Silverman, B. 1985. Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society: Series B* 47, 1–52.

Smith, M. and Kohn, R. 1996. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* 75, 317–43.

Smith, M., Wong, C. and Kohn, R. 1998. Additive nonparametric regression with autocorrelated errors. *Journal of the Royal Statistical Society: Series B* 60, 311–31.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Wahba, G. 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B* 40, 364–72.

Wahba, G. 1983. Bayesian ‘confidence intervals’ for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B* 45, 133–50.

Wahba, G. 1990. *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Whittaker, E. 1923. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society* 41, 63–75.

How to cite this article

Poirier, Dale J. "spline functions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000213> doi:10.1057/9780230226203.1593

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spontaneous order

Peter Boettke and Jennifer Dirmeyer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A spontaneous order is a pattern that is recognizable with human reason but is not created by design. The market process is a spontaneous order that is created by the process of entrepreneurship and competition. A spontaneous order is self-correcting. Prices are formed through the actions of self-interested individuals in the market, buying and abstaining from buying. They make possible an extended order so complex that it is able to communicate information worldwide to producers and consumers alike so that production plans may be coordinated to consumer demands despite any change in resources, tastes or income.

Keywords

entrepreneurship; invisible hand; language; markets; money; price system; property rights; rules of behaviour; social norms; spontaneous order

Article

> I am convinced that if it were the result of deliberate human design… this mechanism would have been acclaimed as one of the greatest triumphs of the human mind.
> —F.A. Hayek (1945, p. 87)

A spontaneous order is a recognizable pattern that is produced by a process that is neither directed by deliberate design nor created for a specific purpose, though it may produce useful results. The economy is a spontaneous order of dizzying magnitude, with billions of individuals coordinating their plans in order to satisfy their individual desires, with no overarching direction. One of most important characteristics of a spontaneous order is that it consists of patterns that are recognizable with human reason despite the fact that they are not the result of human design. In the market, the distribution and allocation of resources according to some principle or predictable relationship, such as the fact that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
prices correspond to opportunity costs or that the price of a good equals the firm's marginal cost, is an example of such a pattern. These systematic and predictable patterns, while obviously recognizable with human reason, are not the result of any person or group actively trying to achieve this result.

The origin of money or a medium of exchange can be explained purely by self-interested actions on the part of individuals who had no intention of creating it. ‘As each economizing individual becomes increasingly more aware of his economic interest, he is led, by this interest, without any agreement, without legislative compulsion, and even without regard to the public interest, to give his commodities in exchange for other, more saleable, commodities even if he does not need them for an immediate consumption purpose’ (Menger, 1871, p. 260). As more people begin realizing the gain to be had from collecting goods that others want, a medium of exchange spontaneously arises even though no person was intending this result, only their own gain. In this example, it is easy to see that emergence of a medium of exchange was ‘of human action but not of human design’ (Hayek, 1967).

Hayek (1948, p. 78) defines the economic problem as ‘how to secure the best use of resources known to any member of society, for ends whose relative importance only these individuals know’. In the market, the spontaneous order that serves this function is the systematic elimination of ‘mutual ignorance on the part of potential market participants’ (Kirzner, 1992, p. 44). The mechanism that produces this result is the price system. The price system permits the spread and use of dispersed knowledge of time and place so the actions of individuals are coordinated. The self-interested actions of entrepreneurs searching for profit opportunities and competing for potential trading partners is what insures that prices embody the relevant information that people need in order to adjust their plans to the constantly changing circumstances of the market. Wealth creation throughout the market system is driven by the continuous discovery of new opportunities to trade, and the ceaseless desire to innovate in a manner that will either lower the costs associated with previous trades or open the opportunity for new trades to be made. This result, while spectacular, is not an end to be actively sought. That the spontaneous order of the market process is means-driven implies that the information that is generated through this process is not available by any other means. We cannot know the utility function of even one other individual (perhaps not even our own), much less that of every individual in society. The only way to generate the allocation of resources that best provides for the varying ends of individuals is to allow the market process to function. That is why we stress that the process that generates a spontaneous order is ends-independent. Although it performs an extremely useful function in the coordination of knowledge and action, this coordination is not the result of action aimed at coordination, but the result of action aimed at gaining profit.

The fact that the generation of a spontaneous order relies not on purpose but on the elements following particular rules of behaviour is significant in that such behavioural rules need to be self-enforcing. In the market process, the behaviour that produces the order is profit-seeking, and as such is self-enforcing as long as people prefer more to less. However, there is an additional criterion that must be met for the order to emerge: ‘…people must also observe some conventional rules, that is, rules which do not simply follow from their desires and their insight into relations of cause and effect, but which are normative and tell them what they ought or ought not to do’ (Hayek 1973, p. 45). There must be an accepted and well-defined sphere over which the individual has control. The importance of clearly defined and enforced property rights cannot be overstated in terms of its impact on the emergence and maintenance of a spontaneous market order.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Features of a spontaneous order

The unique characteristics of a spontaneous order imply certain features that are in contrast to planned or designed orders. Since the order is a result of a process that depends only on the individual elements following certain rules, the complexity and scale of a spontaneous order can be much greater than that of a planned order. The limitations of planned orders are readily apparent when we first consider the necessity of channelling everything through one mind. What is amazing is the possible complexity of the spontaneous order. Leonard Reed's ‘I, Pencil’ (1958) is among the best expositions of the degree of complexity that the market process achieves, and, since he wrote of the amazingly complicated task of producing pencils, it is safe to say that the process has become even more complex when one considers such products as automobiles, airplanes and computers. (See Adam Smith, 1776, pp. 15–16, for a similar examination of the common woollen coat worn by a day labourer and the multitude of exchanges that must be coordinated to produce even this homely product. See also Milton and Rose Friedman, 1980, pp. 1–29, for a classic discussion of the implications of the ‘I, Pencil’ story for our understanding of the market order.) This ability to maintain orders of such complexity is the source of the continued productive and technological growth that has characterized modern history. The importance of this feature of the spontaneous order lies in the astounding ability not just to develop such a complex order but, more importantly, to maintain it though circumstances are changing at almost every instant.

A related feature that spontaneous orders have in common is their abstract nature. Because the order is not the result of deliberate action, the individuals within it do not need to understand or even be aware of the process at work. This feature allows for a level of abstraction that is beyond the understanding of the human mind. The evidence for this in nature is abundant. Although the frontier of scientific knowledge is constantly being pushed out, it seems that there is no limit to what we still have to learn about the workings of the universe. This is true also in the realm of social sciences, but to an even greater extent. Since there is no one mind that must understand the implications of every action, there is almost no limit to experimentation in the market. The profit incentive provides for the generation of endless variation in attempts to better satisfy consumer desires. The abstract quality of the market process is that the goal, satisfying consumers, is never clearly defined. It is the market test, the ‘systematic plan changes generated by the flow of market information released by market participation’ that directs activity toward this end, not a comprehensive plan (Kirzner, 1973, p. 10).

Finally, spontaneous orders are dependent on the structure within which the elements are acting. (An excellent discussion of this point can be found in Vernon Smith's 2002 Nobel Prize address, where he contrasts ‘ecological rationality’ with ‘constructivist rationality’; see V. Smith, 2003.) A purposeful design is dependent on the conception of the designer and his control over the elements. The specific state of a spontaneous order is determined not only by the characteristics of the elements but also by the structure that determines motivation. In the market the entrepreneurial profit incentive is a characteristic of the process. However, this characteristic produces coordination only if the easiest way to get a profit is to provide buyers and sellers with opportunities to better their trading positions. If, instead, the best way to profit is to manipulate the political process in order to confiscate others’ wealth, then the order that emerges will not be recognizable as the coordination of knowledge and action.

Conclusion

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Social norms, language, money and the price system are all examples of spontaneously emerging institutions in the social realm which, while certainly products of purposive human action, are nevertheless not of direct human design. The role of the political process is to define the structure within which the filter processes and the equilibrium processes of the 'invisible hand' operate. Understanding this concept means admitting a limit to the social states that are achievable. It means realizing that 'most of these social states that are romantically imagined to be possible are inconsistent with the motivational postulate of economics, with human nature as it exhibits its uniformities' (Buchanan, 1997, p. 37). Instead the political process must be confined to creating 'conditions in which an orderly arrangement can establish and ever renew itself' (Hayek, 1960, p. 161). That the market process is a spontaneous order means that our span of control is limited, yes, but there is no need to despair! For recognizing the spontaneous order of the market economy as an ongoing entrepreneurial process of discoveries for mutual benefit and wealth creation also means that we can achieve states of civilization that could not be imagined were it necessary that we designed them.

See Also

- invisible hand

Bibliography

Buchanan, J. 1997. There is science of economics. *The Collected Works of James M. Buchanan*, vol. 12. Indianapolis: Liberty Press.

Friedman, M. and Friedman, R. 1980. *Free to Choose*. New York: Harcourt Brace Jovanovich.

Hayek, F.A. 1960. *The Constitution of Liberty*. Chicago: University of Chicago Press.

Hayek, F.A. 1948. The use of knowledge in society. In *Individualism and Economic Order*. Chicago: University of Chicago Press.

Hayek, F.A. 1967. The results of human action but not of human design. In *Studies in Philosophy, Politics and Economics*. London: Routledge & Kegan Paul.

Hayek, F.A. 1973. *Law Legislation and Liberty, Volume 1: Rules and Order*. Chicago: University of Chicago Press.

Kirzner, I. 1973. *Competition and Entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, I. 1992. *The Meaning of the Market Process*. New York: Routledge.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Menger, C. 1871. *Principles of Economics*. New York: New York University Press, 1981.

Read, L. 1958. I, pencil: my family tree as told to Leonard E. Read. *The Freeman*: *Monthly Magazine of the Foundation for Economic Education*, December. Reprinted in *The Freeman* 46(5) (1996).

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Chicago: University of Chicago Press, 1976.

Smith, V. 2003. Constructivist and ecological rationality in economics. *American Economic Review* 93, 465–508.

How to cite this article

Boettke, Peter and Jennifer Dirmeyer. "spontaneous order." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000480> doi:10.1057/9780230226203.1594

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sports, economics of

Stefan Szymanski and Andrew Zimbalist
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economic analysis of professional sport proceeds from the uncertainty of outcome and competitive balance. Members of sports leagues seek to restrain economic competition and claim that they need to do so in order to make contests more attractive. In 1956 Rottenberg analysed this ‘competitive balance defence’, arguing that the reserve clause would not have an effect on the distribution of playing talent in a league, since players would migrate to teams where their economic return was highest, regardless of whether the club or the player owned the right to the revenue stream.

Keywords

antitrust policy; baseball; Coase theorem; competitive balance; marginal revenue product; rents; sport, economics of; tournaments

Article

Why study the economics of sport? In purely financial terms it is not a very important part of the economy. In 1998, out of the $8 trillion US economy, the commercial sports industry generated only $17.7 billion of revenue – a sizeable sum to be sure, but only 0.2 per cent of US GDP and one-quarter the size of the automotive repair industry ($69.6 billion) in that year (U.S. Bureau of the Census, 1999, Table 6.1). We have no field of study on the economics of car repair. We advance five reasons for special interest in sport.
First, the cultural significance of sport is huge. Either by choice or because of economic constraints, organizers have extracted only a portion of the consumer surplus generated by sporting events;
Second, professional sports as business activities are unique in that the service is (sporting) competition itself, and one supplier cannot create the product without the cooperation of his or her (sporting) rival.
Third, investing public money in building facilities to host sports teams and sporting events such as the Olympics is widely advocated by politicians and boosters, and using cost–benefit methods to evaluate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic impacts is an important contribution of economists;
Fourth, the sports industry is also a marvellous laboratory for testing and applying economic theory: it is Frederick Winslow Taylor's fantasy, where practically every aspect of worker (player) performance is observed and most are measured.
Fifth, when properly handled, sport can be one of the most effective ways of introducing students to fundamental economic concepts.

Sport and culture

Games and sports are a universal cultural phenomenon. Many animal species engage in what humans call play, and perhaps it is the lengthy adolescence characteristic of humans which has facilitated such an astonishing range of game playing. All ancient civilizations had their sports, and athletic contests are characteristic of even the least developed societies. In most civilizations sport has been connected with religion, as was true, for example, of the ancient Olympic Games. Modern sports, however, are characterized by secular rules. They also tend to be more formalized than traditional games, encouraging the gathering of statistical records for the purposes of evaluation and comparison.
Almost all of the popular modern sports were formalized somewhere between 1840 and 1900: for example, baseball (1846), soccer (1848), Australian football (1859), boxing (1865), cycling (1867), rugby union (1871), tennis (1874), American football (1874), ice hockey (1875), basketball (1891), rugby league (1895), motor sport (1895) and the modern Olympic Games (1896). Many historians have noted the coincidence between the creation of the rigid, scientific sports, and the advent of rigid industrialized societies in the United Kingdom and United States during the 19th century. The only major modern sports whose organization pre-dates this era are golf, cricket and horse racing, all of which had established rules and clubs from the mid-18th century.
The fact that almost all the sports which today dominate the world originated either in the United Kingdom or the United States also reflects Anglo-Saxon economic dominance. Soccer was spread at the end of the 19th century, when British expatriates or foreign students studying in Britain taught local elites the virtues of British culture and British games. American economic hegemony has given further impetus either to sports that were invented in Britain and taken up by Americans (such as golf and tennis) or to indigenous inventions such as basketball.
The cultural significance of sport has made most countries reluctant to see it exploited for business purposes. Even in the United States, the National Collegiate Athletic Association (NCAA) insists in retaining the myth of amateurism within the confines of the billion-dollar college sports industry. The myth of amateurism prevailed until recent years in a wide variety of international sports. Increasing commercialization is still opposed today in world soccer, and even in baseball the fans object to the idea of corporate sponsorship emblazoned on the sacred team shirt or on the bases.
This reluctance to mix business with sport explains in part how sport can attain such cultural centrality without achieving an equivalent economic status. Another reason is the difficulty involved in capturing rents. Each match in a championship is a mini-soap opera, involving high drama and the unexpected, and much of the pleasure derives from discussion among fans both before and after the event, but the protagonists have few opportunities to secure property rights over these external effects. Even the news media that pore ceaselessly over the latest sporting gossip pays little for the privilege. Another explanation may be the elastic supply of sporting contests (even if demand is frequently inelastic).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Sporting competition is generally cheap to supply. Thus, even if the fans complain bitterly about the prices they pay, only a small fraction of personal disposable income is devoted to sport. Nonetheless, sports consumption is also a luxury good, and total expenditure rises rapidly as national income increases.

The business of professional sport

The starting point for the economic analysis of professional sport is the uncertainty of outcome and competitive balance. Almost from their inception, professional sports leagues have sought to restrain economic competition in the name of creating a more attractive contest. Thus, baseball's National League, the world's first professional sports league, was created in the United States in 1876 and introduced the reserve clause, a mechanism for tying each player in perpetuity to his club in 1879. From the 1880s, the clubs in the league argued that this restraint was necessary to ensure that all the best players did not end up playing for the biggest teams. Were this to happen, they reasoned, the championship would become predictable, fans would lose interest and the competition would collapse. In economic terms, they were arguing that a non-cooperative equilibrium would lead to a less balanced distribution of results than would be selected by a planner interested in maximizing interest in the sport. Given the frequent use of this argument by sports leagues to defend antitrust challenges levelled against economic restraints in the sporting labour and product markets, it might reasonably be termed the ‘competitive balance defence’.

Economic analysis of the competitive balance defence began with Simon Rottenberg's 1956 article in the *Journal of Political Economy*. Rottenberg argued that the reserve clause would have no effect on the distribution of playing talent in a league, since players would always migrate to teams where their economic return (or marginal revenue product) was highest, regardless of whether the club or the player owned the right to the revenue stream. This is clearly an articulation of the Coase theorem, despite being published four years before Coase's ‘On the problem of social cost’. Moreover, references to empirical tests of the Coase theorem commonly present sports leagues as an outstanding example. Much of the sports literature is an elaboration of these issues. El-Hodiri and Quirk (1971) developed the first theoretical model of revenue sharing, and used this to advance the proposition that revenue sharing would not affect the distribution of talent in a league, also for Coasean reasons. Much of the early literature involved examining a sport league's objective function, which in North America typically has been considered to be characterized by profit maximization (see, for example, Jones, 1969), while Sloane (1971) drew attention to the not-for-profit structure and culture that obtained in British sports, notably football, and in the rest of the world (for a detailed comparison of the American and European sports business models see Szymanski and Zimbalist, 2005). Since the 1990s there has been renewed interest in the theoretical modelling of league structures, focusing on the North American model (see, for example, Fort and Quirk, 1995; Vrooman 1995) and the European model (Késenne, 2000; Hoehn and Szymanski, 1999). Szymanski (2003) advances a synthesis based on the contest or tournament framework, and there is a good deal of current research in this vein.

Much of this analysis has been applied in the context of antitrust either in the United States or in the European Union. The courts, in line with most economists, consider leagues as cartels that impose restraints on economic competition in labour and product markets. In North America, the competitive

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
balance defence has been applied in cases such as *NCAA v. Board of Regents of the University of Oklahoma*, 468 U.S. 85 (1984), which argued that collective selling of broadcast rights was necessary to maintain competitive balance. In this case, as in several others, the defence failed, not because the courts did not accept that leagues were special cases where ancillary restraints could be justified, but because the restraints were excessive given their stated aims (see, for example, Flynn and Gilbert, 2001). As a result, most labour market restraints in sports have been built into collective bargaining agreements that are exempt from antitrust, while broadcasting rights generally enjoy the exemption granted by Congress in the 1962 Sports Broadcasting Act. In Europe, where equivalent exemptions do not exist, labour market restraints akin to the reserve clause have been ruled illegal (by the European Court of Justice in the 1995 *Bosman* case), while collective selling of broadcast rights has been treated differently in EU member states, some outlawing collective selling (Italy and Spain) and others permitting it (Germany and the UK).

Economic impact of sports teams and facilities

Top-level professional sports teams have an immense cultural impact on their communities but very little, if any, positive economic impact. All independent empirical research concurs on this point (see, for example, Siegfried and Zimbalist, 2000). There are three principal reasons why no positive economic development effect should be expected from a new team or facility. First, most of the money spent at a facility is replacement for spending at other entertainment activities elsewhere in the metropolitan area. Second, leakages of spending at a ballpark or arena out of the local economy tend to be much greater than at other locally owned entertainment venues, thereby depressing sports expenditure multipliers. Third, public subsidies for facility construction and maintenance create a negative budgetary impact, engendering either higher taxes or lower services, thereby dragging down economic expansion. Nonetheless, the existence of externalities, public good benefits and consumer surplus (consumer demand for sporting contests is generally assumed to be inelastic) may justify some level of public support for facility construction.

Sport as laboratory

Sports involve forms of competitive behaviour that are readily observed and measured. Often we know the precise output and productivity of agents, and their compensation packages are frequently a matter of public record. Hence, it is not surprising that labour economists have looked to the sports field for natural experiments of theories such as incentive design, teamwork, and discrimination.

Gerald Scully (1974) pioneered attempts to empirically estimate a player's marginal revenue product (MRP) in baseball, and much of the subsequent research on labour market performance has followed his innovation. He also applied his method, which involved estimating individual contributions to winning through measures of batting performance, and then the value of winning for team revenues, to compare the relationship between wages and MRPs for players from different races, thus providing a test of discrimination. Disputes over Scully's MRP methodology has kept a lively debate going in the literature (see, for example, Zimbalist, 1992).

Another example of the use of sports data to test labour market theories is Ehrenberg and Bognanno

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(1990), who used scores of professional golfers in different tournaments to test the incentive effects of different prize structures. Other important studies include the use of trading in baseball cards to test for discrimination on the part of fans (Nardinelli and Simon, 1990) and the effect of strikes on demand in baseball (Schmidt and Berri, 2004).

Sports in the classroom

Sports economics has become extremely popular in the classroom, often because many students find it easiest to appreciate economic reasoning in the context of an activity with which they have a certain degree of familiarity. As the discussion of the competitive balance defence suggests, using sport to teach economics can be risky given that sports are often a special case. However, using sport to study concepts such as demand or labour market incentives can be illuminating.

For example, the case of spectator demand can be used to develop a number of key economic concepts. Since the marginal cost to an owner of an additional fan attending any given game is approximately zero, profit and revenue maximization at the ballpark are congruent. This implies that a profit-maximizing strategy for an owner is to set ticket prices so that the price elasticity of demand equals 1. Much of the early empirical literature, however, yielded estimates of inelastic demand at existing prices. This literature suffered from various deficiencies: the use of weighted average, rather than individual, ticket prices; the absence of important control variables; and the failure to specify the model properly. Possibly significant control variables include the probability of home team victory, the uncertainty of outcome, the quality of the visiting team and importance of the contest in league competition, the weather, the age of the stadium, the number of star players on the teams, the distribution and level of local income, and whether the game is televised. Modelling should be affected by the existence of a stadium capacity constraint, the positive dynamic (demonstration) effect of having a full or nearly full stadium on future attendance, and the simultaneity effect whereby higher attendance increases the home field advantage, which improves home team performance and increases attendance. Finally, the profit-maximizing owner will seek to maximize not just gate revenues but all stadium revenues (ticket sales, net concessions, catering, signage, memorabilia, and parking). Many of these considerations will cause ticket prices to appear to be set below the point of unitary elasticity when, in fact, they are set at revenue-maximizing levels.

Conclusion

The economics of sport is a relatively new and rapidly growing field. While the literature is gaining sophistication and relevance, there is still much ground to be covered. Nonetheless, there are some important areas of research that have not been covered properly here, such as research in college athletics in the United States or studies of team production in sports.

In our view the most pressing area for economic research remains the competitive balance defence, both in terms of theoretical foundations and in terms of empirical evidence.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Ehrenberg, R. and Bognanno, M. 1990. Do tournaments have incentive effects? *Journal of Political Economy* 98, 1307–24.

El-Hodiri, M. and Quirk, J. 1971. An economic model of a professional sports league. *Journal of Political Economy* 79, 1302–19.

Flynn, M. and Gilbert, R. 2001. An analysis of professional sports leagues as joint ventures. *Economic Journal* 111, F27–46.

Fort, R. and Quirk, J. 1995. Cross subsidization, incentives and outcomes in professional team sports leagues. *Journal of Economic Literature* 33, 1265–99.

Hoehn, T. and Szymanski, S. 1999. The Americanization of European football. *Economic Policy* 28, 205–40.

Jones, J. 1969. The economics of the National Hockey League. *Canadian Journal of Economics* 2, 1–20.

Késenne, S. 2000. Revenue sharing and competitive balance in professional team sports. *Journal of Sports Economics* 1, 56–65.

Nardinelli, C. and Simon, C. 1990. Customer racial discrimination in the market for memorabilia: the case of baseball. *Quarterly Journal of Economics* 105, 575–96.

Neale, W. 1964. The peculiar economics of professional sports. *Quarterly Journal of Economics* 78, 1–14.

Rottenberg, S. 1956. The baseball players' market. *Journal of Political Economy* 64, 242–58.

Scully, G. 1974. Pay and performance in major league baseball. *American Economic Review* 64, 915–30.

Schmidt, M. and Berri, D. 2004. The impact of labor strikes on consumer demand: an application to professional sports. *American Economic Review* 94, 344–57.

Siegfried, J. and Zimbalist, A. 2000. The economics of sports facilities and their communities. *Journal of Economic Perspectives* 14(3), 95–114.

Sloane, P. 1971. The economics of professional football: the football club as a utility maximiser. *Scottish Journal of Political Economy* 17(2), 121–46.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Szymanski, S. 2003. The economic design of sporting contests. *Journal of Economic Literature* 41, 1137–87.

Szymanski, S. and Zimbalist, A. 2005. *National Pastime: How Americans Play Baseball and the Rest of the World Plays Soccer*. Washington, DC: Brookings Institution.

U.S. Bureau of the Census. 1999. *Service Annual Survey: Current Business Reports* Washington, DC: U. S. Department of Commerce. Online. Available at http://www.census.gov/prod/2000pubs/bs98.pdf, accessed 12 August 2005.

Vrooman, J. 1995. A general theory of professional sports leagues. *Southern Economic Journal* 61, 971–90.

Zimbalist, A. 1992. Salaries and performance in major league baseball. In *Diamonds Are Forever: The Business of Baseball*, ed. P. Sommers. Washington, DC: Brookings Institution.

How to cite this article

Szymanski, Stefan and Andrew Zimbalist. "sports, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_E000235> doi:10.1057/9780230226203.1595

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

spurious regressions

Clive W.J. Granger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Simulations have shown that if two independent time series, each being highly autocorrelated, are put into a standard regression framework, then the usual measures of goodness of fit, such as *t* and R-squared statistics, will be badly biased and the series will appear to be ‘related’. This possibility of a ‘spurious relationship’ between variables in economics, particularly in macroeconomics and finance, restrains the form of model that can be used. An error-correction model will provide a solution in some cases.

Keywords

autocorrelation; Durbin–Watson statistic; econometrics; ordinary least squares (OLS); regression; serial correlation; spurious regression; Weiner process

Article

For the first three-quarters of the 20th century the main workhorse of applied econometrics was the basic regression

$$Y_t = a + bX_t + e_t. \quad (1)$$

Here the variables are indicated as being measured over time, but could be over a cross-section; and the equation was estimated by ordinary least squares (OLS). In practice more than one explanatory variable *x* would be likely to be used, but the form (1) is sufficient for this discussion. Various statistics can be used to describe the quality of the regression, including R^2, *t*-statistics for β, and Durbin–Watson

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

statistic d which relates to any autocorrelation in the residuals. A good fitting model should have $|t|$ near 2, R^2 quite near one, and d near 2.

In standard situations the regression using OLS works well, and researchers used it with confidence. But there were several indications that in special cases the method could produce misleading results. In particular, when the individual series have strong autocorrelations, it had been realized by the early 1970s by time series analysis that the situation may not be so simple; that apparent relationships may often be observed by using standard interpretations of such regressions. Because a relationship appears to be found between independent series, they have been called 'spurious'. Note that, if $b = 0$, then e_t must have the same time series properties as Y_t, that is, it will be strongly autocorrelated, and so the assumptions of the classical OLS regression will not be obeyed. The possibility of getting incorrect results from regressions was originally pointed out by Yule (1926) in a much cited paper that discussed 'nonsense correlations'. Kendall (1954) also pointed out that a pair of independent autoregressive series of order one could have a high apparent correlation between them; and so if they were put into a regression a spurious relationship could be obtained.

The magnitude of the problem was found from a number of simulations. The first simulation on the topic was by Granger and Newbold (1974), who generated pairs of independent random walks, from (1) with $a = b = 1$. Each series had 50 terms and 100 repetitions were used. If the regression is run, using series that are temporarily uncorrelated, one would expect that roughly 95 per cent of values of $|t|$ on b would be less than 2. This original simulation using random walks found $|t| \leq 2$ on only 23 occasions; out of the 100, $|t|$ was between 2 and 4 on 24 occasions, between 4 and 7 on 34 occasions, and over 7 on the other 19 occasions.

The reaction to these results was to reassess many of the previously obtained empirical results in applied time series econometrics, which undoubtedly involved highly autocorrelated series but had not previously been concerned by this fact. Just having a high R^2 value and an apparently significant value of b was no longer sufficient for a regression to be satisfactory or its interpretations relevant. The immediate questions were how one could easily detect a spurious regression and then correct for it. Granger and Newbold (1974) concentrated on the value of the Durbin–Watson statistic: if the value is too low, it suggests that the regressions results cannot be trusted. Remedial methods such as using a Cochrane–Orcutt technique to correct autocorrelations in the residuals, or differencing the series used in a regression, were inclined to introduce further difficulties and could not be recommended. The problem arises because the equation is mis-specified; the proper reaction to having a possible spurious relationship is to add lagged dependent and independent variables until the errors appear to be white noise, according to the Durbin–Watson statistic. A random walk is an example of a I(1) process, that is, a process that needs to be differenced to become stationary. Such processes seem to be common in parts of econometrics, especially in macroeconomics and finance. One approach that is widely recommended is to test whether X_t, Y_t are I(1) and, if so, to difference before one performs the regression. There are many tests available; a popular one is due to Dickey and Fuller (1979).

A theoretical investigation of the basic unit root, ordinary least squares, spurious regression case was undertaken by Phillips (1986). He considered the asymptotic properties of the coefficients and statistics of eq. (1), $\hat{a}$, $\hat{b}$, the t-statistic for b, R^2 and the Durbin–Watson statistics $\hat{\rho}$. To do this he introduced the link between normed sums of functions of unit root processes and integrals of Weiner processes. For

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

example, if a sample X_t of size T is generated from a driftless random walk, then

$$T^{-2}\sum_{1}^{T} X_t^2 \to \sigma_\varepsilon^2 \int_0^1 W^2(t)\,dt,$$

where σ_ε^2 is the variance of the shock, and $W(t)$ is a Weiner process. As a Weiner process is a continuous time random process on the real line [0,1], the various sums involved are converging and can thus be replaced by integrals of a stochastic process. This transformation makes the mathematics of the investigation much easier, once one becomes familiar with the new tools. Phillips is able to show that

- the distributions of the t-statistics for $\hat{a}$ and $\hat{b}$ from (1) diverge as t becomes large, so there is no asymptotically correct critical values for these conventional tests;
- $\hat{b}$ converges to some random variable whose value changes from sample to sample;
- Durbin–Watson statistics tend to zero; and
- R^2 does not tend to zero but to some random variable.

What is particularly interesting is not only that do these theoretical results completely explain the simulations but also that the theory deals with asymptotics, $T \to \infty$, whereas the original simulations had only $T = 50$. It seems that spurious regression occurs at all sample sizes.
Haldrup (1994) has extended Phillips's result to the case for two independent I(2) variables and obtained similar results. (An I(2) variable is one that needs differencing twice to get to stationarity, or, here, difference once to get to random walks.) Marmol (1998) has further extended these results to fractionally integrated I(d) processes. Durlauf and Phillips (1988) regress I(1) process on deterministic polynomials in time, thus polynomial trends, and found spurious relationships.
Although spurious regressions in econometrics are usually associated with I(1) processes, which were explored in Phillips's well-known theory and in the best known simulations, what is less appreciated is that the problem can also occur, although less clearly, with stationary processes.
Table 1 shows simulation results from independent series generated by two first order autoregressive models with coefficients a_1 and a_2 where $0 \square a_1 = a_2 \square 1$ and with inputs e_{xt}, e_{yt} both Gaussian white noise series, using regression 1 estimated using OLS with sample sizes varying between 100 and 10,000.

Regression between independent AR(1) series

Sample series	$a = 0$	$a = 0.25$	$a = 0.5$	$a = 0.75$	$a = 0.9$	$a = 1.0$
100	4.9	6.8	13.0	29.9	51.9	89.1
500	5.5	7.5	16.1	31.6	51.1	93.7
2,000	5.6	7.1	13.6	29.1	52.9	96.2
10,000	4.1	6.4	12.3	30.5	52.0	98.3

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$a_1 = a_2 = a$ percentage of $|t| > 2$

Source: Granger, Hyung and Jeon (2001).

It is seen that sample size has little impact on the percentage of spurious regressions found (apparent significance of the *b* coefficient in (1)). Fluctuations down columns do not change significantly with the number of iterations used. Thus, the spurious regression problem is not a small sample property. It is also seen to be a serious problem with pairs of autoregressive series which are not unit root processes. If $a = 0.75$, for example, then 30 per cent of regressions will give spurious implications. Further results are available in the original paper but will not be reported in detail. The Gaussian error assumption can be replaced by other distributions with little or no change in the simulation results, except for an exceptional distribution such as the Cauchy. Spurious regressions also occur if $a_1 \neq a_2$, although less frequently, and particularly if the smaller of the two *a* values is at least 0.5 in magnitude.

The obvious implications of these results is that applied econometricians should not worry about spurious regressions *only* when dealing with I(1), unit root, processes. Thus, a strategy of first testing whether a series contains a unit root before entering into a regression is not relevant. The results suggest that many more simple regressions need to be interpreted with care when the series involved are strongly serially correlated. Again, the correct response is to move to a better specification, using lags of all variables.

Concerns about spurious regressions produced interest in tests for unit roots, of which there are now many; and empirical works with time series will usually test between I(1) or I(0), or may sometimes consider more complicated alternatives. If series are found to be I(1), simple regressions have been replaced with considerations of cointegration and construction of error-correction models.

A recent survey of studies of spurious relationships is Pilatowska (2004).

See Also

- Durbin-Watson statistic

Bibliography

Baltagi, B. 2001. *A Companion to Theoretical Econometrics*. Oxford: Blackwell.

Dickey, D. and Fuller, W. 1979. Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.

Durlauf, S. and Phillips, P. 1988. Trends versus random walks in time series analysis. *Econometrica* 56, 1333–54.

Granger, C. 2001. Spurious regression in econometrics. In *A Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Granger, C., Hyung, N. and Jeon, Y. 2001. Spurious regression with stationary series. *Applied*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Economics 33, 899–904.

Granger, C. and Newbold, P. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.

Haldrup, N. 1994. The asymptotics of single-equation cointegration regressions with I1 and I2 variables. *Journal of Econometrics* 63, 153–81.

Kendall, M. 1954. *Exercises in Theoretical Statistics*. London: Griffin.

Marmol, F. 1998. Spurious regression theory with non-stationary fractionally integrated processes. *Journal of Econometrics* 84, 233–50.

Phillips, P. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–40.

Phillips, P. 1988. New tools for understanding spurious regressions. *Econometrica* 66, 1299–325.

Pilatowska, M. 2004. Realization of the congruence postulate as a method of avoiding the effects of a spurious relationship. In *Dynamic Economic Models*, vol. 6, ed. Z. Zieli•ski. Torun: Nicolaus Copernicus University.

Yule, G. 1926. Why do we sometimes get nonsense correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society* 89, 1–64.

How to cite this article

Granger, Clive. "spurious regressions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000218> doi:10.1057/9780230226203.1596

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sraffa, Piero (1898 – 1983)

John Eatwell and Carlo Panico
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capital theory; classes; classical theory of growth; classical theory of value and distribution; diminishing returns; Dmitriev, V. K.; Dobb, M. H.; equilibrium; fixed capital; forced saving; gold standard; Hayek, F. A. von; imperfect competition; increasing returns; inflation; invariable standard of value; Keynes, J. M.; labour theory of value; laws of returns; managed currency; marginal productivity of capital; money; multi-product industries; natural rate of interest; neoclassical theory of cost; neoclassical theory of the rate of profit; neoclassical theory of value and distribution; neo-Ricardian economics; non-negative prices; non-substitution theorem; own rate of return; partial equilibrium; perfect competition; Perron–Frobenius theorem; price formation; quantity theory of money; rate of profit; real wages; reswitching of technique; returns to scale; Ricardo, D.; single-product industries; Sraffa, P.; standard commodity; surplus

Article

In the history of economics, Piero Sraffa is an enigma. His reputation as a major economic theorist rests on but three works: the *Economic Journal* article of 1926, the Introduction to his edition of Ricardo's *Principles* (the first of the 11 volumes of the complete *Works and Correspondence* of *David Ricardo*, which established Sraffa as the finest scholar to have edited a major work in the literature of economics), and the 99 pages of *Production of Commodities by Means of Commodities*, a sparse, terse collection of logical propositions, the significance of which is a matter of often heated debate.

A reclusive figure, of great personal warmth and puckish humour, Sraffa spent most of his life in Cambridge. Yet his influence extended far beyond academic economics.

Throughout the 1930s he spent every Thursday afternoon and evening, during term, with Ludwig Wittgenstein. It was Sraffa who forced Wittgenstein to accept that the theory of language advanced in the *Tractatus Logico-Philosophicus* was logically inadequate, paving the way for the recognition of the social content of signs and language presented in *Philosophical Investigations*. In the preface of the latter, Wittgenstein acknowledged the importance of Sraffa's criticism of his arguments over many years,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and added ‘I am indebted to *this* stimulus for the most consequential ideas of this book’. He later commented that after discussion with Sraffa he felt ‘like a tree stripped of all its branches’, but that a consequence of this drastic pruning was healthier growth.

Of perhaps wider significance was Sraffa's close friendship with Antonio Gramsci. After Gramsci's imprisonment, Sraffa led the effort to ameliorate the harsh conditions in which he was held, and to secure his release. The essential contents of Gramsci's letters from prison, written to his sister-in-law Tatiana, were channelled through Sraffa in Cambridge to the exiled Italian Communist Party. And it was Sraffa who ensured, by establishing an account with unlimited credit at a Milanese bookshop, that Gramsci was supplied with the materials he needed to work on his *Prison Notebooks*.

The philosophical debates with Wittgenstein, and his political and intellectual commitment to socialism, point to important elements in Sraffa's intellectual make-up. His economics, always rigorous, became in the 1930s increasingly formal, to the extent that the search for logically precise and unambiguous expression inhibited his writing. His socialism demanded an economics that was concrete; that, however abstract, was appropriate to the interpretation of real economic institutions and phenomena.

The compelling empiricism of Marxian socialist thought, and the rejection of the use of subjective concepts, are themes running throughout Sraffa's economics. Economics should be constructed from variables and relationships which are, at least in principle, observable and measurable. The classical analysis of value and distribution is constructed on just such ‘empirical’ foundations, whilst neoclassical theory, based as it is on unverifiable hypotheses concerning individual choice, evidently is not.

But although Sraffa's contribution to economic theory may have been motivated by these methodological concerns, its substance involves the logic of theoretical argument, in particular the demonstration of the logical consistency of the classical analysis of value and distribution, and, as a corollary, the logical deficiencies of neoclassical theory.

Life and works

Piero Sraffa was born in Turin on 5 August 1898. His father was Angelo Sraffa, a professor of commercial law who later became Chancellor of the Bocconi University in Milan. The Piazza Sraffa in Milan is named for Angelo, not Piero. Piero's mother was Irma Tivoli.

Sraffa was educated at the Liceo d'Azeglio in Turin, where he was greatly influenced by Umberto Cosmo, who introduced him to socialist ideas and, in 1919, to Antonio Gramsci. Sraffa's studies at the University of Turin, from 1916 to 1920, were interrupted by military service, which he spent as both a ski instructor and an engineer, blowing up bridges to stem the Austrian advance. He attended relatively few lectures. Nonetheless his honours thesis, ‘Monetary Inflation in Italy During and After the War’, was considered by his supervisor, Luigi Einaudi, to be quite brilliant. It was published in 1920.

After graduation, Sraffa worked for a few weeks in a bank to learn some banking ‘from the inside’. He then went to the London School of Economics (1921–2) where he attended lectures by Cannan, Foxwell and Gregory.

During his stay in London, Sraffa visited Cambridge, bearing a letter of introduction to Keynes from Mary Berenson, a friend of Sraffa's family who, ten years earlier, had entertained Keynes and other young Cambridge graduates in her villa near Florence. Keynes was at the time engaged in the debate on the reconstruction of the international monetary system, and had agreed to be editor of a weekly supplement to the *Manchester Guardian*, dealing with the monetary and financial problems of Europe.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Keynes asked Sraffa to contribute an article on the Italian banking system, which was at the time experiencing a severe crisis. The article proved to be too long for the newspaper, and was published in the *Economic Journal* instead, a shorter version appearing in the *Manchester Guardian*.
These two articles were also published in Italian, and caused the Fascist regime considerable irritation. Now back in Italy, Sraffa was accused by Mussolini of ‘banking defeatism’ and ‘sabotage of Italian finance’. Keynes invited Sraffa to return to England until things had calmed down. However, on 23 January 1923 Sraffa was detained at Dover, and after being questioned for three hours was informed that he had been refused permission to land by order of the Home Secretary. The reason for the denial of entry is still not clear. Keynes secured the removal of Sraffa's name from the list of ‘undesirables’ in 1924.
Fortunately the situation in Italy was less severe than had been expected, though Sraffa resigned the job he had obtained as Director of the Bureau of Labour Statistics of the Province of Milan. The Bureau had been established by the socialist provincial administration in 1922, and was experiencing difficulties with the Fascist government. A few months later he was appointed to a lectureship in Political Economy and Public Finance at the University of Perugia.
The preparation of his lectures at Perugia stimulated him to write ‘Sulle relazioni fra costo e quantita prodotta’ (1925). As a result of this article Sraffa was appointed to a professorship in Political Economy at the University of Cagliari, a post he held *in absentia* to the end of his life, donating his salary to the support of the library. Edgeworth's high opinion of the article led to an invitation to submit a version to the *Economic Journal* (1926). This, in turn, led to Sraffa's being offered the lectureship in Cambridge, which he took up in October 1927.
Before leaving Italy, Sraffa had pursued his interest in monetary problems by translating Keynes's *Tract on Monetary Reform* into Italian, and writing several short reviews of books on money and banking for the *Giornale degli Economisti* (1925b; 1926b; 1926c; 1927b).
In 1919, Sraffa had joined the Socialist Students’ Group at the University of Turin and had participated actively in the political life developing around *Ordine Nuovo*, the magazine founded in 1919 by Gramsci, Tasca, Terracini and Togliatti, the group who were to play the crucial role in the split from the Socialist Party at the Congress of Livorno in 1921 and the foundation of the Italian Communist Party. Whilst in London Sraffa wrote three articles for *Ordine Nuovo* on the condition of the working class in England and the role of trade unions. In 1924 he published an open letter to Gramsci in *Ordine Nuovo* criticizing the Communist Party for its dogmatic refusal to contemplate an alliance with other democratic groups against Fascism. A few years later Gramsci, now imprisoned, accepted Sraffa's argument.
Sraffa also opposed the orthodox Party line in two letters to *Stato Operaio* in 1927. In a discussion of the devaluation of the Italian lira, he criticized the prevalent view that policy decisions are always mechanically and ‘*directly* dictated by the *immediate* interests of the banks and the big industrialists’ (1927a, p. 180). He advanced instead the view that political bodies such as the Fascist Party have their own interests which can enter into the dynamics of the decision process.
In October 1927 Sraffa began his lectures in Cambridge, presenting courses on the theory of value and on the relationship between banks and industry in continental Europe. He was to lecture for only three years, finding the very process increasingly difficult. Joan Robinson, who attended the lectures on her return from India, recalled them vividly, not least because Sraffa liked to develop a dialogue with his class – a procedure unknown in Cambridge. In 1930 Sraffa was appointed Marshall Librarian, and also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
placed in charge of graduate studies. He gave up lecturing for good.
Shortly after arriving in Cambridge, Sraffa had shown Keynes the set of propositions (derived from Marx's reproduction schemes) which were to grow into *Production of Commodities*. But this work was somewhat overwhelmed both by the intense debate in Cambridge surrounding Keynes's *Treatise on Money* and, later, *The General Theory* (it was Sraffa who organized the famous ‘circus’ which discussed the *Treatise* in 1931), and by Sraffa assuming in 1930 the editorship of the Royal Economic Society edition of *The Works and Correspondence of David Ricardo*. Sraffa's work on the theory of value, and his interest in monetary theory, coalesced in a critical review of Hayek's *Prices and Production* (Sraffa, 1932a; 1932b).
Sraffa's participation in political debate was necessarily limited after his arrival in Cambridge. He maintained contact with the leadership of the Italian Communist Party in Paris, and in 1927 wrote to the *Manchester Guardian* denouncing the ill treatment of the imprisoned Gramsci. He visited Gramsci and attempted, to no avail, to use the influence of his uncle, an eminent judge, to secure Gramsci's release. Following Gramsci's death in 1937, it was Sraffa who conveyed to Togliatti Gramsci's wishes concerning the editing of the *Quaderni dal Carcere*.
In another service to a friend, Sraffa travelled to Austria following the *Anschluss* to inform Wittgenstein's family that Ludwig had renounced his Austrian citizenship.
In 1939 Sraffa was elected to a Fellowship of Trinity College (he had previously held dining rights at King's), a post he took up shortly after the outbreak of war. When Italy entered the war in June 1940, Sraffa was interned as an enemy alien on the Isle of Man. Keynes managed to extricate him by the end of the summer. Sraffa never gave up his Italian citizenship.
By the late 1940s, the publication of the edition of Ricardo had been long delayed. This was partly due to the reorganization required when in 1943, after six volumes were already in the press, Ricardo's letters to Mill and, amongst other writings, the papers on Absolute and Exchangeable Value were discovered. But delay was also caused by the difficulty Sraffa was having in writing the introductions to the volumes, particularly the introduction to the *Principles*. The second problem was solved after 1948 with the assistance of Maurice Dobb. Dobb and Sraffa would discuss each paragraph in detail. Dobb would write it up. Sraffa would revise what Dobb had written. Dobb would rewrite. And so on, until the job was done. The first four volumes of the *Works and Correspondence of David Ricardo* were published in 1951, the next five volumes in 1952, a bibliographical miscellany formed the tenth volume published in 1955, and, after a number of false starts by others, a general index, compiled by Sraffa himself, was published as the eleventh volume of the set in 1971.
The edition is widely acknowledged to be a scholarly masterpiece. George Stigler (1953) commented,

> Ricardo was a fortunate man. He lived in a period – then drawing to a close – when an untutored genius could still remake economic science•…•. And now, 130 years after his death, he is as fortunate as ever: he has been befriended by Sraffa – who has been befriended by Dobb.
> Keynes told us, in 1933, that Sraffa, ‘from whom nothing is hid’, would give us the full works of Ricardo within the year. The truth of the first part of the statement had as its cost the falsification of the second, and it has been a splendid bargain. For Sraffa's *Ricardo* is a work of rare scholarship. The meticulous care, the constant good sense, and the erudition make this a permanent model for such work; and the host of new materials seem to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

suggest that Providence meets half-way the deserving scholar.

The *Ricardo* completed, Sraffa could return to the work he had done on the theory of value and distribution in the 1930s and early 1940s. Old notes were reassembled, including a proof of the Perron–Frobenius theorem on non-negative square matrices, which the Trinity mathematician Besicovitch (an analyst with no prior knowledge of the theorem) had provided on a postcard delivered on Christmas Day 1944. The result of assembling these old notes was *Production of Commodities by Means of Commodities* (1960).

This book was greeted with almost universal puzzlement. It seemed to present, in odd, formal terms, propositions which had become familiar with the development of linear models in the 1940s and 1950s. Had Sraffa's brilliant insights of 30 years before been overtaken by events in mainstream theory? Of the earlier reviewers only Dobb (1961), Meek (1961) and Newman (1962) grasped the fact that this was a work of profound significance, with implications for the logical foundations of both classical and neoclassical theories of value and distribution.

Sraffa spent the rest of his life in Cambridge though up to 1973 he visited Italy in every vacation, staying in his apartment in Rapallo, and going to Rome to attend meetings of the economic section of the Academia dei Lincei. Other than economics and politics, Sraffa's great interest was the collection of books. He assembled a magnificent collection of economics books and pamphlets, including a first edition of *Kapital* inscribed by Marx himself (which he later presented to the Istituto Gramsci) and a copy of the *Wealth of Nations* containing Adam Smith's bookplate. He shared this enthusiasm with Keynes. Together they discovered, identified, and wrote an introduction to an edition of David Hume's *An Abstract of a Treatise on Human Nature* (1938). When he died, on 3 September 1983, Sraffa left his collection to Trinity College.

Contributions to economic theory

The early years

Sraffa's dissertation (1920) dealt with central *practical* issues occupying writers on monetary matters at the end of the First World War – the causes and consequences of inflation, the stabilization of internal prices and exchange rates within an unstable international monetary system, the argument for restoring the gold standard and revaluing the currency to the pre-war gold parity.

Sraffa argued that since the abandonment of the gold standard at the beginning of the war had been followed by halving of the purchasing power of gold, then a return to the gold standard would require a rise in the value of the metal, forcing countries which fixed parity either to devalue or to bear the consequences of deflation. Since, in these circumstances, the monetary authorities could not achieve stability of both prices and the rate of exchange, it was better to opt for the former. There is no law which forced the authorities to stabilize the currency at the pre-war level. The *normal* value of the currency is completely ‘conventional’, that is, it can be at *any* level that common opinion expects it to be (Sraffa, 1920, p. 42). Sraffa, in opposition to Einaudi, and in common with the position which Cassel, Hawtrey and Keynes were to urge on the Genoa Economic Conference in April 1922, favoured a ‘managed currency’.

Sraffa's ‘practical’ case embodied an implicit theoretical argument. He stressed the role of the state,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
moulded by the pressure of the major economic classes, in determining the distribution of income. Monetary policy was thus considered in terms of its impact on the real wage. Sraffa, like Keynes in *A Tract on Monetary Reform*, accepted a version of the quantity theory. But whereas Keynes's views on the determination of the distribution of income were essentially neoclassical, Sraffa's position was more akin to that of the classical economists and Marx. Keynes argued that social forces and monetary factors have only temporary effects on the distribution of income, influencing only the disequilibrium real wage rate. Unless, that is, they were able to effect the real factors which determined the equilibrium distribution (Keynes, 1923, p. 27). Sraffa, however, argued that monetary policies, and hence inflation and deflation, are aspects of the social conflicts that directly regulate the equilibrium or normal real wage rate (Sraffa, 1920, pp. 25, 40–2).

A similar view of the role of economic institutions in social conflict was spelt out in the *Economic Journal* and *Manchester Guardian* articles. The articles focused on the financial needs of newly developing Italian industry, and on the evolution of the links between industry and the banks. His study of the formation of large groups or ‘concentrations’ of financial and industrial power is similar to that in Hilferding's *Finance Capital*. He stressed the enormous economic and political power which such groups can acquire and outlined the way in which conflicts within the groups might affect economic policy. He also revealed the accounting tricks which had been used by two major banks to disguise their financial difficulties, and showed how the authorities had evaded legal restrictions in order to favour some major financial groups (Sraffa, 1922b, p. 676). No wonder Mussolini was so upset.

Laws of returns

Sraffa's early writings, although imbued with a certain critical radicalism, provided no hint of the theoretical tour de force that was to come. It is true that he had emphasized the role of social classes and institutions in the normal (as opposed to disequilibrium) operation of the economy, but in Sraffa's examination of the neoclassical (predominantly Marshallian) theory of cost (1925; 1926) these concerns with the ‘objective’ characteristics of economic activity were transformed into a penetrating critique of the logical foundations of the theory of the equilibrium of the competitive firm and of the supply curve. Sraffa's starting point in the *Annali di Economia* was a distinction between an analysis in which the relationship between cost and quantity produced was determined by ‘objective’ factors, such as the ordering of different qualities of land, and a relationship which was based on ‘subjective’ factors, namely the marginal disutility which accompanies the offer of increased quantities of factor services. The former relation is, as Wicksteed had argued (1914), essentially descriptive; it is the latter which is, in neoclassical theory, analytic. The supply curve is simply the demand curve ‘reversed’ (Wicksteed, 1914).

In the *Economic Journal* Sraffa made the same point in a rather different way. In classical economics, he argued, the ‘laws of returns’ did not derive from a unified analysis of cost. Quite the contrary. The discussion of increasing returns was associated with the analysis of accumulation, most notably Adam Smith's examination of the relationship between the extent of the market and the division of labour. Diminishing returns, on the other hand, were the distinctive component of the theory of rent. The suggested symmetry of increasing and diminishing returns is a quite different construction, characteristic of the neoclassical supply curve.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
But if cost is subjective disutility, how does a phenomenon which is defined by personal psyche manifest its influence in the determination of the equilibrium of the competitive firm, and in the determination of the supply curve of the industry? The apparent similarity between the determination of equilibrium in individual choices, and the equilibrium of the firm, is quite spurious. For whereas all the determinants of individual choice – preferences and endowments – are peculiar to the individual, the determinants of the equilibrium in production in a competitive economy – the technical conditions of production and the supply of factor services – are external to the firm.

Sraffa then demonstrated that neither increasing *nor* diminishing returns are compatible with the assumption of perfect competition in the determination of the supply curve of an industry, except in the peculiar case in which economies or diseconomies of scale are external to the firm but internal to the industry.

Diminishing returns are incompatible with perfect competition, since the presumption of price taking precludes any impact of the output of individual firms, or, in Marshall's *ceteris paribus* world, the output of individual industries, on prices, unless it is assumed that endowments are fixed in individual firms, or are peculiar to individual industries.

Increasing returns are also incompatible with assumption of perfect competition – other than those which are external to the firm, but internal to the industry.

Only the assumption of constant returns to scale is compatible with the assumption of perfect competition:

> In normal cases the cost of production of commodities produced competitively … must be regarded as constant in respect of small variations in the quantity produced. And so, as a simple way of approaching the problem of competitive value, the old and now obsolete theory which makes it dependent on the cost of production alone appears to hold its ground as the best available. (Sraffa, 1926a, pp. 540–1)

There were two ways out of the conundrum, either to adopt the general equilibrium reasoning which Sraffa had deployed so effectively against the notion of the supply curve, or to abandon the assumption of perfect competition. Marshall's theory must be abandoned (Sraffa, 1930a; 1930b).

The first course was ruled out on the grounds that examination of ‘the conditions of simultaneous equilibrium in numerous industries’, though a well-known approach, is far too complex; ‘the present state of our knowledge … does not permit of even much simpler schemata being applied to the study of real conditions’ (1926a).

The second course recognizes both the ‘everyday experience … that a very large number of undertakings – and the majority of those which produce manufactured consumers’ goods – work under conditions of individual diminishing costs’, and that

> the chief obstacle against which they have to contend when they want gradually to increase their production does not lie in the cost of production … but in the difficulty of selling the larger quantity of goods without reducing the price, or without having to face increased marketing expenses. This … is only an aspect of the usual descending demand curve, with the difference that instead of concerning the whole of a commodity, whatever its origin, it relates only to the goods produced by a particular firm … (1926a).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sraffa's second option launched the Cambridge analysis of imperfect competition, first in Richard Kahn's fellowship dissertation (1931) at King's, then in Joan Robinson's *Economics of Imperfect Competition*.

Apart from his contribution in the *Economic Journal* symposium on increasing returns, Sraffa did not participate further in the debate on the Marshallian theory of cost. The reasons are not hard to seek. First, imperfect competition theory, instead of providing a new, more concrete approach to the analysis of value and distribution, was simply absorbed into neoclassical theory. The fact that imperfectly competitive models do not provide a foundation for a theory of value, seemed to enhance the status of partial equilibrium analysis, rather than hasten its rejection; with the competitive theory of value still holding sway at the level of general equilibrium (a neat rationale is provided by Hicks, 1946, pp. 83–4). The survival of the ‘U’ shaped cost curve as an analytical tool, constructed from the presumption of increasing, then diminishing returns, is in no small part attributable to the longevity provided by models of the imperfectly competitive firm. Nonetheless, the appearance of the ‘U’ shaped curve in models of the competitive firm, more than 60 years after Sraffa clearly demonstrated the illegitimacy of the construction, is an indication of just how intellectually disreputable theoretical economics can be.

Second, Sraffa's implicit identification of classical and Marxian theory with the notion that competitive value is ‘dependent on the cost of production’ is clearly wrong, as examination of neoclassical models which take account of ‘simultaneous equilibrium in numerous industries’ readily demonstrates. Sraffa had deployed general equilibrium reasoning to demolish the theory of the competitive firm and the industry supply curve. Further criticism of neoclassical theory would require consideration of general equilibrium models of value and distribution. And a constructive rehabilitation of classical theory would require a general analysis too. It would require, that is, an analysis of ‘the process of diffusion of profits throughout the various stages of production and of the process of forming a normal level of profits throughout all the industries of a country’ – the problem Sraffa acknowledged was ‘beyond the scope of this article’ (1926a, p. 550).

Monetary theory

There was no sign of Sraffa's emerging critique of neoclassical theory in his review of Hayek's *Prices and Production* (1932a; 1932b). Instead, the review displays some similarities between Sraffa's position and that held by Keynes soon after the publication of the *Treatise*.

Sraffa argued that Hayek had failed to identify the essential properties of money by neglecting the fact that

> money is not only a medium of exchange, but also a store of value and the standard in terms of which debts, and other legal obligations, habits, opinions, conventions, in short all kinds of relations between men, are more or less rigidly fixed. (Sraffa, 1932a, p. 43)

The absence of any conception of wage agreements and debts fixed in money terms prevented Hayek from analysing correctly the effects on the distribution of income of a general fall or rise in prices. Since money had been thoroughly ‘neutralized’ it could not effect the distribution of income or the rate of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
accumulation. Hence Hayek could characterize ‘forced saving’ as a disequilibrium phenomenon, with no permanent effects.

Sraffa argued that this conclusion was contrary to a ‘common sense’ view of the economy. During a period of inflation

> one class has, for a time, robbed another class of a part of their incomes; and has saved the plunder. When the robbery comes to an end, it is clear that their victims cannot possibly consume the capital which is now well out of their reach. (Sraffa, 1932a, p. 48; see also 1932b, p. 249)

Sraffa's view that class conflict determines the *normal* real wage, and that monetary policy may be part of that conflict, is an echo of Sraffa's earlier position, yet nothing is said on the theory of distribution as such.

The most enduring construction in the article is Sraffa's invention of the concept of the own rate of return. Sraffa utilized the idea to elucidate the concept of equilibrium underpinning much of Hayek's discussion. In particular he demonstrates that whilst in disequilibrium there may be as many ‘natural’ rates of interest (that is, own rates of return) as there are commodities, competition will tend to equalize these natural rates just as competition eliminates any divergence between market prices and normal prices – indeed these are two aspects of the same process.

Yet Sraffa does not consider how the equilibrium rate of interest is determined. Nor does he criticize Hayek's association of the rate of interest with the length of the production process.

Ricardo

Sraffa's edition of *The Works and Correspondence of David Ricardo* proved to be more than a great scholarly achievement. For in his introduction to *The Principles of Political Economy and Taxation* Sraffa presented an entirely new interpretation of Ricardo's theory of value and distribution. Sraffa's interpretation established a new, theoretically consistent version of the surplus approach to the analysis of distribution in the *Essay on Profits*. Further, he demonstrated that this approach was sustained in the *Principles* by Ricardo's use of the labour theory of value, and that, contrary to the accepted view of Ricardo's analysis presented by Jacob Hollander (1904), Ricardo did not retreat from his use of the labour theory of value in successive versions of the *Principles*.

In the *Essay on Profits* Ricardo stated that it is the rate of profit in agriculture which determines the rate of profit in the economy as a whole. Sraffa argued that

> The rational foundation of the principle of the determining role of the profits of agriculture, which is never explicitly stated by Ricardo, is that in agriculture the same commodity, namely corn, forms both the capital (conceived as composed of the subsistence necessary for workers) and the product; so that the determination of profit by the difference between total product and capital advanced, and also the determination of the ratio of this profit to the capital, is done directly between quantities of corn without any question of valuation. (Sraffa, 1951, p. xxxi)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The beautiful simplicity of this interpretation – the rate of profit being determined in the agricultural sector as a ratio of quantities of corn, and in the other sectors as a ratio of values, with the price ratio between corn and other commodities adjusting so as to equalize the rate of profit (corn, being the wage good, is part of the capital in all sectors) – suggested itself in the preparation of *Production of Commodities by Means of Commodities*. If there is but one 'basic' commodity in the economy (that is, but one commodity which enters directly or indirectly the production of all others), then not only must all the inputs to that commodity consist of itself, but the general rate of profit must be determined by the ratio of surplus of the commodity produced to its means of production.

This powerful result, in which the rate of profit is clearly determined as the ratio of surplus to means of production was sustained in the *Principles*, where it was generalized to incorporate the fact that surplus and means of production will consist of heterogeneous 'bundles' of commodities. The homogeneity necessary to find the ratio of surplus to means of production was achieved by evaluating the two bundles in terms of the labour embodied directly and indirectly in their production.

As is well known, this generalization foundered on the fact that commodities do not exchange at their labour values, and hence that the ratio, evaluated in terms of labour values, does not measure the rate of profit. As Sraffa pointed out, Ricardo's persistent struggle with this difficulty was expressed as the fact that prices might change due to a change in distribution when labour values (dependent upon conditions of production) were unchanged. Hence he sought an 'invariable standard of value' which would tie movements in price to movements in labour values alone:

> Ricardo was not interested for its own sake in the problem of why two commodities produced by the same quantities of labour are not of the same exchangeable value. He was concerned with it only in so far as thereby relative values are affected by changes in wages. The two points of view of difference and of change are closely linked together; yet the search for an invariable measure of value, which is so much at the centre of Ricardo's system, arises exclusively from the second and would have no counterpart in an investigation of the first. (Sraffa, 1951, p. xlix)

Sraffa was able to demonstrate that Ricardo had continued the search for an invariable standard to the end of his life, sustaining thereby the exposition of his theory of distribution in terms of the labour theory of value. The conclusive proof of Sraffa's argument was found in Ricardo's papers on *Absolute and Exchangeable Value* discovered together with other papers and letters in 1943, their existence having been previously unknown.

Sraffa's interpretation of Ricardo had a considerable impact at the time of its publication, not least because there was great interest in the analysis of growth at the time. The analysis of distribution plays a central role in the classical theory of growth (accumulation by the capitalists is determined by their share of the product). The problems in the theory of the rate of profit posed by the neoclassical analysis of growth, and in some versions of Keynesian growth theory, excited interest in Ricardo's approach. However, the real importance of Sraffa's new interpretation was for the understanding of Marx's analysis of value and distribution, which is based on Ricardo's theory, and for the general rehabilitation of the surplus approach to value and distribution, which had, for so long, been regarded as logically deficient.

Production of Commodities by Means of Commodities

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The subtitle of Sraffa's book (1960) is *Prelude to a Critique of Economic Theory*, and in the Preface he suggested that ‘If the foundation holds, the critique may be attempted later, either by the writer or by someone younger and better equipped for the task’ (Sraffa, 1960, p. vi), echoing Ricardo's comment in the Preface to his *Principles* (1817, p. 6).

Production of Commodities is a peculiarly sparse book. The argument has been pared to the absolute minimum to sustain the propositions which Sraffa wishes to advance. Yet the precision and logical elegance of the argument are ‘the work of an artist working in the medium of economic theory’ (Newman, 1962).

The theoretical essence of the book may be distilled from the argument of Part I in which Sraffa deals with single-product industries and circulating capital. There Sraffa demonstrates that the approach to the analysis of value and distribution adopted by Ricardo and by Marx is logically consistent. Taking as data the size and composition of output, the conditions of reproduction, and the real wage it may be shown that (1) in an economy which is capable only of reproducing itself, relative prices are determined by the conditions of production; and (2) that, in an economy which is capable of producing a physical surplus over and above the needs of reproduction, relative prices are determined by the conditions of production of basic commodities, and the manner in which the surplus is distributed. If, in the latter case, the surplus is distributed as a rate of profit, then the data determine relative prices and that rate of profit. The economically meaningful solution – that with non-negative prices – is unique. (The prices of non-basics depend upon their own conditions of production and the prices of basics, but the prices of basics are not affected by the prices of non-basics.) These propositions had already been advanced by Dmitriev (1898), though they were not well known.

Sraffa then drops the assumption that the real wage is given. The degree of freedom thus introduced into the analysis is expressed in the locus of the rates of profit associated with any particular values of the wage (in terms of the *numéraire*). For any given value of the wage there is a unique rate of profit (and associated prices), and vice versa. There is a maximum wage, when the rate of profit is equal to zero; and a maximum rate of profit when the wage is equal to zero. Closure of the model requires either that the real wage be given (that is, determined outside the determination of the rate of profit and normal prices) as in classical theory; or that the rate of profit be given.

Sraffa's suggestion that the rate of profit is ‘susceptible of being determined from outside the system of production, in particular by the level of the money rates of interest’ (1960, p. 33), is essentially symmetrical with the classical approach in which the real wage is ‘given’. For the classical economist the real wage is determined by social and historical forces, circumstances which may be analysed quite separately from the determination of relative prices and the rate of profit. Likewise, it may be argued that the money rate of interest, to which, in a competitive economy, the rate of profit must conform, is determined by the normal operations of monetary institutions, especially the state. This position is reminiscent of Sraffa's earlier work in monetary theory, and of Keynes's remark that ‘the rate of interest is a highly conventional, rather than a highly psychological, phenomenon’ (Keynes, 1936, p. 203).

While the ‘data’ of Sraffa's analysis of value and distribution are identical with the data of the analyses advanced by Ricardo and by Marx (other than in his not taking the real wage as given), and hence his results are a validation of their arguments, his method of solution is different. Whereas Ricardo and Marx sought to determine the rate of profit as a ratio of aggregates, Sraffa solves for the rate of profit

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and prices simultaneously. Indeed, his argument demonstrates the necessity of doing so. Yet in his construction of the standard commodity, Sraffa seeks to recreate the clarity of the classical derivation of the rate of profit from the ratio between surplus and means of production.

Sraffa first constructs from the given conditions of production a ‘standard system’, an hypothetical economy in which the composition of means of production and net product (wages and profits) are the same. If the wage is expressed as a proportion of standard net product, w, then the proportion of net product accruing to profits is $(1-w)$. If the ratio of total net product to surplus is R – a ratio which may be evaluated because the composition of inputs and net output is the same – then the rate of profit will be equal to $R(1-w)$.

The standard system is therefore a direct descendant of the agricultural sector in the *Essay on Profits*, the rate of profit is expressed as a ratio between two physical quantities.

Sraffa then demonstrates that if the standard net product is adopted as *numéraire*, and hence as the measure in terms of which the wage is expressed, then the rate of profit will be equal to $R(1-w)$, exactly as in the standard system in which the relationship between the wage and the rate of profit is expressed in purely physical terms.

The purpose of this construction is, Sraffa tells us, to ‘give transparency to a system and render visible what was hidden’ (1960, p. 23). The rate of profit is seen to be determined by the magnitude of surplus. Yet the use of the standard commodity must be distinguished from Ricardo's use of the ‘corn sector’, or the use of the labour theory of value by Ricardo and Marx. In Sraffa's case, the rate of profit is *determined* by the solution of the simultaneous equations, the standard commodity is a purely auxiliary construction. In the case of Ricardo and Marx, the rate of profit is determined (albeit imperfectly) by calculating ratio of surplus to means of production by means of the labour theory of value.

It cannot be said that the standard commodity is entirely successful as a means of rendering visible what might otherwise be hid. It is, perhaps, too complex, lacking the simple force of the labour theory. It has the virtue, however, of being analytically correct.

Considerable puzzlement was engendered by Sraffa's statement in the Preface of his book that ‘The investigation is concerned exclusively with such properties of an economic system as do not depend on changes in the scale of production …‘ (1960, p. v). The absence of any reference to demand led unsuspecting readers to equate his results with the non-substitution theorem, and hence with the assumption of constant returns to scale. However, a careful reading of Sraffa's analysis reveals that no knowledge of any relationship between *changes* in outputs and *changes* in inputs, or between price and quantity is *necessary* for the solution of the equations, and hence for the determination of the rate of profit and prices (given the wage). This contrasts with neoclassical theory, in which the determination of prices is dependent upon knowledge of functional relationships between supply and demand. If, in Sraffa's analysis, quantities should change, then any consequential change in conditions of production will result in changes in prices.

In Part II of his book, Sraffa extends his analysis to multi-product industries and fixed capital, and to the analysis of economies with more than one non-reproducible input. As might be expected, the analysis is considerably more complex, and in some cases the results less clear-cut (the solution of the system may not, for example, be unique, and the definition of basics and non-basics is more abstract than is the case with single-product industries). Yet the basic structure of classical analysis is preserved – the prices, the rate of profit, and other distributive variables (say, land rents), are determined by the conditions of production, given the wage.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Sraffa's analysis is a triumphant restatement of the classical analysis of value and distribution. It is therefore somewhat misleading to refer to a ‘Sraffa-based critique of Marx’, a phrase which implies, perhaps unintentionally, that Sraffa has developed a method of analysis which is conceptually different from that advanced by his predecessors in the theory of surplus value. This is not the case.

The label ‘neo-Ricardian’ which is often attached to Sraffa's work (a term which he himself vehemently rejected) is also unfortunate, implying as it does that the argument of *Production of Commodities* is in some way a solution to problems posed by Ricardo, but not encountered by Marx or by other surplus theorists. The confusion may have derived from simplistically identifying Sraffa with Ricardo, given his edition of Ricardo's works, or from a confusion of the standard commodity with Ricardo's invariable standard of value (even though the latter cannot exist). It may also derive from a fear that any weakening of ‘commitment’ to the labour theory of value implies a rejection of surplus theory. This is to confuse a tool which is used to solve an analytical problem, with the data of the problem. The labour theory of value is not a datum in surplus theory (if it were, Quesnay would not be a surplus theorist), it is a means of demonstrating that the rate of profit is determined by the magnitude of surplus (less rent).

Almost as a by-product of his examination of the fundamentals of classical theory, Sraffa produced a decisive critique of the neoclassical theory of the rate of profit (and hence of the neoclassical theory of long-run normal prices). An examination of the relationship between the changes in the distribution of income and the consequent changes in relative prices leads to the conclusion that such changes ‘cannot be reconciled with *any* notion of capital as a measurable quantity independent of distribution and prices’ (1960, p. 38). In Part III of *Production of Commodities* Sraffa extended his examination of changes in distribution and prices to the case in which changes in distribution lead to changes in the technique of production. He demonstrates that as distribution is varied switches between methods of production, according to which is cheapest, do not follow any particular pattern. Indeed, a technique which is cheapest when the rate of profit is low, may be superseded by another technique at a higher rate of profit, and at a yet higher rate of profit the first technique may again prove to be cheapest and so supersede the second technique. In other words, competitive choice of technique does not result in any particular ordering of techniques. Most notably, the capital intensity of production is *not* an inverse function of the rate of profit, as is implied by the concept of the marginal productivity of capital.

The discussion of ‘reswitching’ (see Symposium, 1966; Garegnani, 1970) following Levhari's failed attempt (1965) to demonstrate that Sraffa's result was confined to decomposable systems, blossomed into a general critique of the logical foundations of the neoclassical theory of the rate of profit. The conclusion of the debate may be stated as:

> it is not possible, using the data of neoclassical theory – the preferences of individuals, the technology, and the size and distribution of the endowment – to determine the normal long-run rate of profit and the associated prices.

Neoclassical models of competitive value which are consistent in their own terms (say, the model presented in Debreu's *Theory of Value*, 1959) do not determine a long-run equilibrium, in which stocks of produced means of production are adjusted to the demand for them and, in consequence, there is a uniform rate of profit. The definition of equilibrium used in such models is different from the traditional long-run equilibrium (Garegnani, 1976; Milgate, 1979). If the model in Debreu (1959) were constrained

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

to yield a uniform rate of profit, it would be over-determined to degree k–1, where k is the number of reproducible means of production. Paradoxically, this latter result is derivable from Hahn's (1982) attempt to refute the above conclusion.

Conclusion

Piero Sraffa's consistently critical approach to the neoclassical theory of value and distribution was motivated by a distaste for ‘subjective’ models, but was conducted in purely logical terms, at least in his later works.

His admiration for the ‘objective’ structure of classical theory, and for the ‘openness’ of that structure which permits the incorporation of concrete institutional factors into the formal analysis, led him to attempt to establish that theory on logically more rigorous grounds than had hitherto been available.

The critical debate set off by *Production of Commodities* has been somewhat blunted by the change in the notion of equilibrium used in general equilibrium theory (the implications of this change for the operational content of economic theory, that is, for the relationship between the theory and the competitive market economy it purports to analyse, have not as yet been satisfactorily analysed).

The well-known propensity of economists to ignore uncomfortable results has also led to the critique being viewed as an esoteric debate in capital theory, with little general significance. This view is clearly wrong. Any critique of the neoclassical theory of value and distribution is a critique of the entire corpus of neoclassical analysis, for the theory of price formation is central to all neoclassical results. (See, for example, Eatwell and Milgate, 1984, on the relevance of these results for the theory of output and employment.)

But it was the revival of the classical (and Marxian) approach to value and distribution, with all the consequences that has for the study of employment, accumulation, technical change, and so on, which was Sraffa's central concern. *Production of Commodities* was designed to lay the groundwork for that revival.

See Also

- capital theory (paradoxes)
- reswitching of technique
- surplus

Selected works

1920. *L’ Inflazione Monetaria in Italia*. Milan: Primiata Scuola Tipografica Salesiana.

1921a. Open shop drive. L’ Ordine Nuovo, 5 July, 3.

1921b. Industriali e governo inglese contr i lavoratori. *L’ Ordine Nuovo*, 24 July, 3.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1921c. Labour leaders. *L' Ordine Nuovo*, 4 August, 1–2.

1922a. The bank crisis in Italy. *Economic Journal* 32, 179–97.

1922b. Italian banking to-day. The Manchester Guardian Commercial: Reconstruction in Europe, Supplement, 7 December, 675–6.

1924a. Problemi di oggi e di domani. *L' Ordine Nuovo*, 1–15 April, 4.

1924b. Obituary – Maffeo Panteloni, *Economic Journal* 34, 648–53.

1925a. Sulle relazioni tra costi e quantità prodotta. *Annali di Economia* II, 277–328.

1925b. A short review of Hastings, H.B., *Cost and Profit Their: Relation to Business Cycles*. Boston: Houghton Mifflin, 1923. Giornale degli Economisti 66, 389–90.

1926a. The laws of return under competitive conditions. *Economic Journal* 36, 535–50.

1926b. A short review of Lehfeldt, R.A., *Money*. London: Oxford University Press, 1926. Giornale degli Economisti 67, 230.

1926c. A short review of Segre, M., Le banche nell'ultimo decennio, con particolare riguardo al loro sviluppo patologico nel dopo guerra, Milano: La Stampa Commercial, 1926. Giornale degli Economisti 67, 230.

1927a. Il vero significato della 'quoto 90', two letters to A Tasca with a reply. *Stato Operaio* 1 (November–December 1927), 1089–95. Reprinted in *Capitalismo Italiano del novecento*, Bari: Laterza, 1972.

1927b. A short review of Phillips, H.W., *Modern Foreign Exchange and Foreign Banking*. London: Macdonald and Evans, 1926. Giornale degli Economisti 68, 610.

1927c. The methods of Fascism. The case of Antonio Gramsci. *Manchester Guardian*, 24 October.

1930a. Symposium on 'Increasing Returns and the Representative Firm'. A criticism. *Economic Journal* 50, 89–92.

1930b. Symposium on 'Increasing Returns and the Representative Firm'. Rejoinder. *Economic Journal*, 50, 93.

1930c. An alleged correction of Ricardo. *Quarterly Journal of Economics* 44, 539–44.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1932a. Dr. Hayek on money and capital. *Economic Journal* 42, 42–53.

1932b. Money and capital: a rejoinder. *Economic Journal* 42, 249–51.

1938. *D. Hume: An Abstract of a Treatise on Human Nature (1740)*. Cambridge: Cambridge University Press. (Introduction with J.M. Keynes.)

1951–73. *The Works and Correspondence of David Ricardo*. 11 vols. Cambridge: Cambridge University Press.

1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

1962. Production of Commodities. A comment. *Economic Journal* 72, 477–9.

Bibliography

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New York: Wiley.

Dmitriev, V.K. 1898. The theory of value of David Ricardo: an attempt at a rigorous analysis. In *Economic Essays on Value, Competition, and Utility*, trans D. Fry, ed. D.M. Nuti. Cambridge: Cambridge University Press, 1974.

Dobb, M. 1961. Review of Production of Commodities by Means of Commodities. De Economist. Reprinted In *A Critique of Economic Theory*. ed. E.K. Hunt and J.G. Schwartz. Harmondsworth: Penguin, 1972.

Eatwell, J. 1975. Mr Sraffa's standard commodity and the rate of exploitation. *Quarterly Journal of Economics* 89, 543–55.

Eatwell, J. and Milgate, M., eds. 1984. *Keynes's Economics and the Theory of Value and Distribution*. London: Duckworth.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37, 407–36.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in Modern Capital Theory*, ed. M. Brown, K. Sato and P. Zarembka. Amsterdam: North Holland.

Garegnani, P. 1978. Sraffa's revival of Marxist economic theory. *New Left Review* No. 112, 71–5.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hahn, F. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6, 353–74.

Hicks, J. 1946. *Value and Capital*. 2nd edn. Oxford: Clarendon Press.

Hollander, J. 1904. The development of Ricardo's theory of value. *Quarterly Journal of Economics* 18, 455–91.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Levhari, D. 1965. A non-substitution theorem and switching of techniques. *Quarterly Journal of Economics* 79, 98–105.

Meek, R.L. 1961. Mr Sraffa's rehabilitation of classical economics. *Scottish Journal of Political Economy* 8, 119–36.

Milgate, M. 1979. On the origin of the notion of ‘intertemporal equilibrium’. *Economica* 46, 1–10.

Newman, P. 1962. Production of Commodities by Means of Commodities. *Schweizerische Zeitschrift fur Volkwirtschaft und Statistik* 98, 58–75.

Pasinetti, L.L. 1985. In memoria di Piero Sraffa: economista italiano a Cambridge. *Economica politica* 2, 315–32.

Ricardo, D. 1817. *On the Principles of Political Economy and Taxation*. London: Murray.

Roncaglia, A. 1978. *Sraffa and the Theory of Prices*. London: Wiley.

Roncaglia, A. 1983. Piero Sraffa: una bibliografia ragionata. *Studi Economici* 21, 137–61.

Roncaglia, A. 1984. Sraffa e le banche. *Rivista Milanese di Economia* 10, April–June, 104–12.

Stigler, G.J. 1953. Sraffa's Ricardo. *American Economic Review* 43, 586–99.

Symposium. 1966. Symposium on paradoxes in capital theory. *Quarterly Journal of Economics* 80, November.

Wicksteed, P. 1914. The scope and method of political economy in the light of the ‘marginal’ theory of value and distribution. *Economic Journal* 24, 1–23.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Eatwell, John and Carlo Panico. "Sraffa, Piero (1898–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000219> doi:10.1057/9780230226203.1597

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sraffian economics (new developments)

Michael Mandler
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Sraffian claim that the forces of competition can fail to determine the distribution of income stands opposed to the neoclassical conclusion that competitive equilibria are determinate. We show that when Sraffa is embedded in a general equilibrium model the indeterminacy claim can sometimes be valid. Sraffians go on to argue that economies use just the right number of production activities to generate exactly one dimension of indeterminacy. If, however, the devices Sraffians use to block extra dimensions of indeterminacy are applied systematically, then all indeterminacy can disappear. Sraffians also argue that capital-theory paradoxes lead to instability of equilibrium, but this position is hard to reconcile with the fact that if consumer demand satisfies the weak axiom then production economies are tâtonnement stable, even in the presence of linear activities and capital-theory paradoxes.

Keywords

aggregate capital; aggregation; capital accumulation; capital theory paradoxes; determinacy of equilibrium; differentiability; differentiable production function; distribution of income; existence of equilibrium; expectations; factor prices; factor tâtonnement; general equilibrium; growth accounting; indeterminacy; intensive rent; intertemporal rates of substitution; labour supply; Leontiev utilities; linear activities; market clearing; neoclassical growth theory; overlapping generations; power; price effect; rent; reverse capital deepening; savings and investment; Sraffian economics; stability of equilibrium; steady state; supply and demand; tâtonnement; total factor productivity; value of capital; Walras' law

Article

1 Introduction

In an earlier era, Sraffians took aim at the neoclassical assertion that the demand for and supply of labour and other resources determine factor incomes. In stalking the big game, a smaller question served as bait: is there a homogenous substance, aggregate capital, whose marginal product determines the return to capital? After a brief episode of disagreement in the 1960s, the neoclassical side conceded that in models with even a minimal disaggregation of capital goods the answer to the smaller question was 'no'. Despite this bloodletting, the chase petered out. When the hunter and hunted ran out of formal modeling disagreements, a settlement was drawn up.

The agreement stipulated that there are two neoclassical theories. The first is an aggregative model that tells the familiar parables of Solow growth theory: increases in savings raise the ratio of the value of capital to labour employed, the rate of interest falls as production becomes more capital intensive, and so on. But once a multiplicity of capital goods is introduced these parables no longer hold true. General equilibrium theory, on the other hand, places no limit on the number of capital or consumption goods and still gives a coherent, determinate account of markets and price determination and hence of the distribution of factor incomes.

While both parties could agree to this decree, they took different views as to who walked away with the more valuable share of the community property. By the time of the split in the early 1970s, general equilibrium had already been singled out as the jewel of microeconomic theory. The results that had to be jettisoned were confined to the steady-state effects of capital accumulation, leaving the prize results of general equilibrium theory – the existence of equilibrium and its welfare properties – untouched. Moreover, a multiplicity of consumption goods will by itself (without multiple capital goods) imply that the distribution of income cannot be determined by *the* marginal products of capital and labour. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
disavowal by high theory of aggregate capital therefore seemed a small loss. Hahn (1982), which marked the end of engagement with Sraffian dissent, argued not only that general equilibrium theory left the line of neoclassical succession intact but also that all legitimate Sraffian results could be obtained from suitably specialized general equilibrium models. The resilience of general equilibrium theory to the Sraffian assault had the curious effect that aggregative empirical work could proceed unaffected by the Sraffa episode. Even a literature such as growth accounting for which the Sraffa critique was pertinent showed no influence – it took off just as the Sraffa attack reached its height. Since general equilibrium theory viewed all forms of aggregation as suspect, overlooking Sraffian concerns about capital aggregation seemed to be one of the ordinary compromises that applied work demands.

The Sraffian reaction was more complicated. Some held that modern general equilibrium theory, although internally consistent, fails because it does not explain how relative prices converge through time to long-run values, in this view the proper goal of economic science (Garegnani, 1976; Eatwell, 1982; Kurz and Salvadori, 1995). Another strand simply promoted Sraffian and classical economics more generally as a distinct type of economic theory (for example, Harcourt, 1974; Marglin, 1984). Sraffians take as their starting place a wage or wage-share of output that is determined by non-economic forces, for example by political power or by social consensus. The general equilibrium model in contrast explains factor prices as the outcome of the endogenous play of supply and demand. Different assumptions, different theories: let the evidence decide.

An agreement to disagree should leave all parties dissatisfied. If Sraffian economics and general equilibrium theory are merely two contenders, each with its own starting place about the causal forces that move factor prices, to be adjudicated by empirical test, then all the wrangling in the 1960s was for nought. For if wages are determined by political power, say, rather than supply and demand, then political power could remain the prime determinant even if capital always aggregated perfectly or in models where prices do not converge to long-run values. What makes the Sraffa–neoclassical debate significant is its critical dimension, the Sraffian arguments that the forces of competition cannot pin down the distribution of income, that any supply-and-demand theory is riddled with internal flaws. The neoclassical side of the debate has contributed its share of confusion: the mere existence of competitive equilibria does not speak to the adequacy of a supply-and-demand account of factor incomes. When translated into the language of general equilibrium, the Sraffian complaints presumably concern the determinacy and stability of equilibrium, not existence or optimality. Yet Hahn (1982), for example, treats determinacy only casually and leaves stability unaddressed. Fortunately, some decades of delay after the noisy 1960s and 1970s, the literature on Sraffa has turned to these points.

The Sraffian complaints about supply-and-demand theories of price determination can be spelled out in two ways. The first appears in Sraffa's *Production of Commodities by Means of Commodities* (1960): the laws of competitive markets do not fully determine factor prices or the distribution of income. Competition requires that the same rate of return is earned in every sector; when laid out as a system of equations, this requirement leaves one more variable than equation, thus revealing a single dimension of indeterminacy, or, as Hahn (1982) put it, a ‘missing equation’. Hahn and other neoclassical economists responded that the missing equation would be found as soon as supply-equals-demand equalities are incorporated into Sraffa's model. Sraffians vacillate on market clearing for factors of production: the land market has to clear, but the labour market does not. This asymmetry drives the single dimension of indeterminacy. As we will see, if the same market-clearing conditions that Sraffians impose on land markets to quash extra dimensions of indeterminacy are applied to the labour market then even the standard single dimension of indeterminacy can disappear. This conclusion would seem to undercut Sraffa's book: if indeterminacy stems solely from failing to require the labour market to clear then Sraffians hardly need an elaborate model to press their point. In essentially any setting, the deletion of labour-market clearing will turn the wage into a free variable and hence leave the distribution of income indeterminate. On this score at least, there would be no need to object to the aggregate neoclassical production function.

But the story is not so simple: the neoclassical presumption that the full gamut of market-clearing conditions necessarily brings determinacy is not correct. Although the ingredients have to be recombined, the Sraffian tradition takes just the right modeling steps that lead factor prices in general equilibrium models to be indeterminate. Sraffians have long insisted that linear activities provide a more faithful representation of technology than the differentiable production functions that dominate practical work in neoclassical economics. Although linear activities by themselves do not generate indeterminacy, they can when endowments of capital goods are governed by rational savings decisions rather than by chance. The Sraffian view of the economy as an ongoing cycle of reproduction thus paves the way for factor-price indeterminacy. On the other hand, the particular way Sraffa and his followers have spelled out their long-run view of the economy, by requiring that relative prices be constant through time, undermines their ‘missing equation’ criticism: linear activities models with constant relative prices have determinate factor prices. And the aggregation of capital has no bearing on the matter: the determinacy of a supply-and-demand theory of factor incomes depends on how many activities operate compared with the number of scarce factors, not on the number of capital goods.

The second completion of the Sraffa critique focuses on the potential for the value of capital per worker to behave badly, for example to increase in response to a rise in the interest rate. Although it might seem that this possibility could by itself lead to instability, this turns out not to be the case. Instability can arise in general equilibrium but it stems from the demand side of the model, not the failure of capital goods to aggregate.

Little of the Sraffian–neoclassical settlement therefore withstands scrutiny. While a couple of assertions in Solow growth theory about steady states hinge on whether the economy has a single sector and whether capital aggregates, the operation of competitive markets does not. The neoclassical confidence that the general equilibrium model answers all Sraffian challenges is equally misplaced: the Sraffian indeterminacy thesis can be reclaimed. As for neoclassical growth theory, its main message that the return to saving diminishes as savings increase can be re-expressed to avoid the limitations of single-sector models. But here too the Sraffian tradition points the way to important corrections. The characteristic neoclassical equality between an economy's interest rate and its marginal rate of transformation is an artifact of differentiable technologies. With linear activities this equality need not obtain, although for the failure of the neoclassical maxim to be robust, we must follow the neoclassical program, rejected by some Sraffians, of letting utility functions determine consumption.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This article's focus on determinacy and stability aims to complement Paul Samuelson's Sraffian economics, which draws the lessons to be learned from Sraffa regarding aggregative parables.

2 The single dimension of Sraffian indeterminacy

We set a benchmark model of linear activities. Let there be *N* goods, one type of labour, *L* types of land, and finitely many activities. Each activity *i* when operated at the unit level requires an investment one period in advance of $a_i = (a_{1i}, \ldots, a_{Ni}) \geq 0$ units of the *N* material inputs and then a contemporaneous application of $\ell_i \geq 0$ units of labour and $\Lambda_i = (\Lambda_{1i}, \ldots, \Lambda_{Li}) \geq 0$ units of the *L* land types to produce the outputs $b_i = (b_{1i}, \ldots, b_{Ni}) \geq 0$. The level at which activity *i* is operated is given by y_i. We assume to begin that the prices $p = (p_1, \ldots, p_N) \geq 0$ of the material goods purchased as inputs equal the prices of the same material goods when sold as outputs one period later. Profit maximization requires that no activity makes positive economic profits and that any activity *i* in use $(y_i > 0)$ makes zero economic profits. Let *r* be the intertemporal interest rate, $w \geq 0$ be the wage, and $\rho = (\rho_1, \ldots, \rho_L) \geq 0$ be the rental rates on land. So profit maximization dictates, for each activity *i*,

$$p_1 b_{1i} + \ldots + p_N b_{Ni} \leq (1 + r)(p_1 a_{1i} + \ldots + p_N a_{Ni}) + w\ell_i + \rho_1 \Lambda_{1i} + \ldots + \rho_L \Lambda_{Li} \tag{2.1}$$

and that equality holds for any *i* such that $y_i > 0$.

The Sraffa literature equivocates on market clearing for resources. While land types are required to have a 0 rental rate when in excess supply, the situation for labour is often left unspecified. Let e_{Λ_k} denote the supply of type *k* land and e_ℓ denote the supply of labour. We then have the market-clearing conditions

$$\sum_i \Lambda_{ki} y_i \leq e_{\Lambda_k}$$
$$\sum_i \Lambda_{ki} y_i < e_{\Lambda_k} \Rightarrow \rho_k = 0, \tag{2.2}$$

for land type $k = 1, \ldots, L$. The analogous conditions for labour are

$$\sum_i \ell_i y_i \leq e_\ell, \quad \sum_i \ell_i y_i < e_\ell \Rightarrow w = 0.$$

Letting *y* denote a vector activity levels, an *equilibrium* is a $(p, w, 1 + r, y)$ that satisfies (2.1) and (2.2). When we impose labour-market clearing, we say so explicitly. In the background lurk additional market-clearing conditions for produced goods, which we introduce in Section 3.

Since we are interested only in relative prices, we can normalize prices by choosing one of the goods or a bundle of goods as numéraire. We set

$$p_1 + \ldots + p_N = 1. \tag{2.3}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the basic Sraffa model, the focus of the first generation of literature, each activity produces only one good and uses no land, and only one activity is available to produce each good *j* and is therefore given the index *j*. So in this case we can rescale each (b_i, a_i, $\bullet_i$) so that b_{ii} (the sole non-zero coordinate of b_i) equals 1. Assuming that each activity is in use, then (2.1) gives us the classical Sraffa equalities:

$$p_i = (1 + r)(p_1 a_{1i} + \ldots + p_N a_{Ni}) + w\ell_i, \; i = 1, \ldots, N. \tag{2.4}$$

The simplest version of Sraffa's missing equation or single dimension of indeterminacy thesis amounts to the observation that (2.4) and the normalization (2.3) comprise $N + 1$ equations but contain $N + 2$ price variables $(p, w, 1 + r)$. A complete argument that any $(p, w, 1 + r) \gg 0$ that satisfies (2.3) and (2.4) is locally contained in a one-dimensional set of points that solve the same equalities might seem to require a rank condition, but the linearity and homogeneity of the model make this unnecessary (see the *Note on the dimension of indeterminacy of the basic Sraffa model* at the end of the text). Since typically we can parameterize the solutions to (2.3) and (2.4) by *w* or *r*, the distribution of income is indeed indeterminate. Competition, Sraffians suggest, does not pin down a division of social wealth between capital and labour.

The indeterminacy of the above model has little significance from the neoclassical point of view: the only alarming possibility would be if market-clearing equalities for some reason could not close the model. The Sraffian literature does not engage this argument, however, and instead takes either *w* or *r* to be exogenous, set by political factors or by a macroeconomic determination of the interest rate. The rationale for this practice is perpetually unclear: is it that market-clearing equalities cannot fill the indeterminacy gap or does some principle trump the laws of supply and demand?

We now document the efforts of Sraffa and his followers to maintain precisely a single dimension of indeterminacy. ‘Single dimension of indeterminacy’ is not standard terminology; a more conventional but equivalent description would be that Sraffians aim to show that (2.1) and (2.3) locally determine prices once *w* or *r* has been set, and hence that, given *w* or *r* and given an equilibrium, a small exogenous change in demand will leave prices unaffected. Indeed, the view that prices in the long-run are affected only by technology and the distribution of income, not the composition of demand, has long been a top item of the Sraffian theoretical agenda.

Single-dimensional indeterminacy faces three threats – rent, joint production, and the choice of technique – the same topics that dominate the second half of Sraffa's book and the second generation of the Sraffian literature. Although we will see that the arguments available against extra dimensions of indeterminacy can sometimes be turned against the presence of any indeterminacy at all, our position is that zero, one, and more than one are all plausible equilibrium possibilities for the dimension of indeterminacy. It may seem in some of our exhibits that demand and market-clearing should eradicate any indeterminacy, but Section 3 will show that they are compatible.

Exhibit A. land and rent

If we add an additional non-produced factor with a positive price to a Sraffa model the number of endogenous price variables will increase. If no further activities are drawn into production, the dimension of indeterminacy will normally go up.

For an elementary example, let there be one produced good $(N = 1)$, one type of land $(L = 1)$ with endowment $e_\Lambda > 0$, and multiple activities. Let the rental rate of the single land type be ρ and again normalize activities so that when activity *i* is operated at the unit level one unit of output is produced. Suppose activity *i* with $(a_{1i}, \ell_i, \Lambda_i) \gg 0$ is in use and $y_m = 0$ for $m \neq i$. Then (2.1) reduces to

$$1 = (1 + r)a_{1i} + w\ell_i + \rho\Lambda_i, \tag{2.5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$1 \leq (1+r)a_{1m} + w\ell_m + \rho\Lambda_m,\ m \neq i, \tag{2.6}$$

where we have substituted in our normalization $p_1 = 1$. To ensure that (2.2) does not constrain ρ to equal 0, the land must be fully employed: $\Lambda_i y_i = e_\Lambda$. With ρ as an additional free variable, if $(\bar{w}, 1+\bar{r}, \bar{\rho}) \gg 0$ satisfies (2.5) and (2.6) and each inequality in (2.6) is strict, an additional dimension of indeterminacy obtains: if we independently vary *w* and $1+r$ a small amount then ρ can be adjusted so that (2.5) and (2.6) remain satisfied.

Another type of equilibrium occurs when two or more activities are in use. If two activities *i* and *j* are in use, then (2.5) is replaced by two equalities,

$$1 = (1+r)a_{1i} + w\ell_i + \rho\Lambda_i, 1 = (1+r)a_{1j} + w\ell_j + \rho\Lambda_j. \tag{2.7}$$

Condition (2.6) now holds for $m \neq i, j$, and full employment for land is given by $\Lambda_i y_i + \Lambda_j y_j = e_\Lambda$. Evidently the argument for an additional dimension of indeterminacy now fails; we cannot independently vary *w* and $1+r$ and expect to satisfy both equalities in (2.7) using the single free variable ρ.

The Sraffa literature largely focuses on the second type of equilibrium with a single dimension of indeterminacy. Is this the more likely type? To make the best case, notice that in the first type of equilibrium the produced input must be accumulated in an amount that leads the stock of land to be fully utilized using only activity *i* – otherwise (2.2) would require $\rho = 0$. If e_c is the stock of the produced input accumulated each period, then to fully employ both e_c and the entire land supply e_Λ using only activity *i* there must be an activity level $y_i \geq 0$ such that $a_{1i}y_i = e_c$ and $\Lambda_i y_i = e_\Lambda$. Since e_c must therefore equal $e_\Lambda a_{1i} / \Lambda_i$, perhaps one could conclude that the accumulation of this exact amount is unlikely to occur. But consider how the shape of the production possibilities set changes as e_c changes. If we fix *w* arbitrarily (and suppose implicitly there is no constraint on labour supply), then, outside of exceptional values for *w* and barring flukes in the input usage coefficients, at most two activities can have the least cost per unit of output and be employed by profit-maximizing firms. If exactly two activities are in use, then the economy can raise its usage of the produced input and increase output by switching the mix of the two activities towards whichever activity economizes on the use of land and uses the produced input intensively – the activity *j* with the higher a_{1j} / Λ_j ratio. This remixing delivers a linear increase in current output as e_c rises. Since increases in e_c must come from the previous period, the production possibilities frontier (PPF) for the current and previous period's consumption is also linear at points where two activities are in use. Once remixing has been exhausted (a 'switch point'), a new activity with a higher a_{1j} / Λ_j must be adopted if more capital is to be used to produce more current output. At the switch point, the first type of equilibrium occurs where one activity is in use and the PPF exhibits a kink (non-differentiability). But optimizing agents may well choose to save a quantity of the produced input so as to end up at a kink in the PPF. See Figure 1, which pictures the tangency between a kink in a PPF and a smooth indifference curve, where consumption at other periods is fixed at optimal levels. (In a multi-agent model, one may interpret the indifference curve as the boundary of the set of consumptions that can Pareto improve on the optimum.) Such one-activity-in-use optima are robust to perturbations of the model. Hence, contrary to Sraffian practice, the first type of equilibrium with the extra dimension of indeterminacy should not be excluded.

Figure 1

Production possibility frontier and indifference curve

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

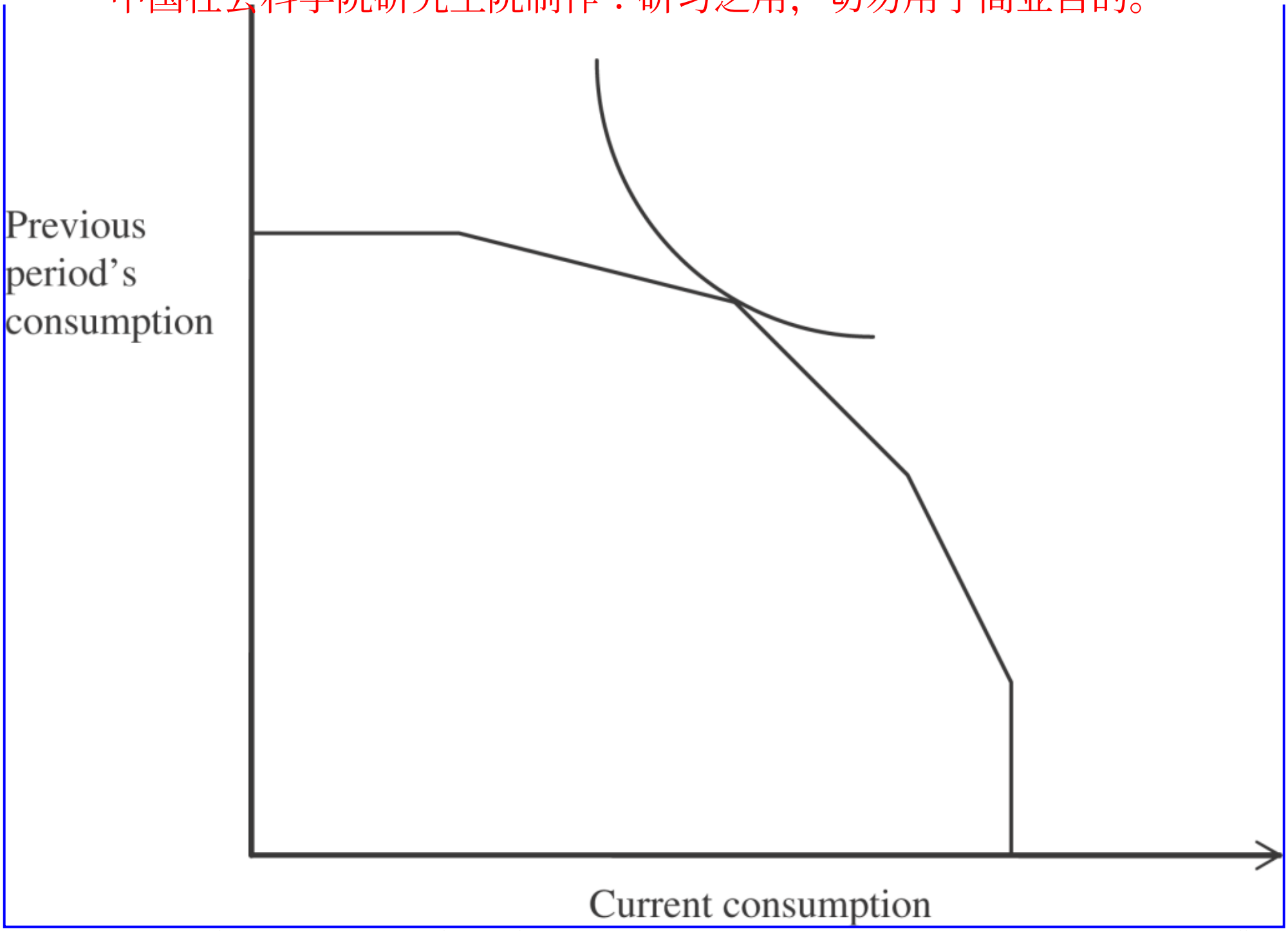

We will see in Exhibit C that if nevertheless we dismissed the first type of equilibrium as unlikely then we would be compelled also to dismiss equilibria with the traditional Sraffian single dimension of indeterminacy. The view that a single dimension of indeterminacy obtains across modeling environments therefore cannot be upheld.

The Sraffian theory of rent often considers cases where a given type of land is used in only one sector out of many. A different rationale is then available for concluding that a single

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dimension of indeterminacy will be the norm. Let $N > 1$ and suppose some good j is produced by single-output activities that use various types of land in addition to labour and material inputs. We consider the simplest scenario, extensive rent, where each type of land is used by only one activity (Sraffa, 1960, ch. 11; Quadrio-Curzio, 1980; for more general models, see Salvadori, 1986, and Bidard, 2004, ch. 17). We assign each land type the same index as the activity that uses it. Renormalizing again so that activities produce one unit of output when run at the unit level, the profit-maximization conditions for the production of j appear as

$$p_j \leq (1+r)(p_1 a_{1i} + \ldots + p_N a_{Ni}) + w\ell_i + p_i \Lambda_{ii} \tag{2.8}$$

for each activity i that produces j and where equality holds if $y_i > 0$. The market-clearing requirements (2.2) for land types are assumed to hold. Let the remainder of the economy's sectors satisfy the assumptions of the basic Sraffa model: for each good $k \neq j$, one single-output activity that uses no land is available and in operation. To avoid the distraction of feedback from changes in p_j into j's cost of production, we assume that j does not directly or indirectly enter into the production of any other good. We pick some good besides j as numéraire and fix w. Finally, letting c_i denote the non-land cost of production $(1+r)(p_1 a_{1i} + \ldots + p_N a_{Ni}) + w\ell_i$ for an activity i that produces j, we assume that w and technology coefficients are such that no ties occur among the c_i: if i and m are distinct activities that produce j then $c_i \neq c_m$.

As in the previous $N = 1$ example, the presence of an extra dimension of indeterminacy will depend on the extent of production. But since $N > 1$ we may view the extent of the production of j as a consequence of the demand for j rather than of different levels of accumulation. As increases in demand progressively raise p_j the economy will first use the type i land for which c_i is lowest. When this land type is exhausted p_j must rise further, until the type i for which c_i is second lowest earns the rate of return r, at which point production can expand further. And so on. The supply ‘function’ thus consists of steps where the ‘horizontals’ indicate that some type i is partly but not fully utilized and the ‘verticals’ that a set of land types is fully utilized but that p_j is not yet high enough for the lowest cost of the remaining types to be drawn into production.

On the horizontals the standard one dimension of indeterminacy obtains. If l types of land are used to produce j, the single zero-profit equality in (2.4) for p_j is replaced by l equalities from (2.8). But the additional $l - 1$ equalities are matched by $l - 1$ additional endogenous rental rates – the one land type that is partly utilized is constrained to have a 0 rental rate. Hence, since (2.4) and (2.3) generate a single dimension of indeterminacy, so do the horizontal equilibria. On the verticals an additional dimension of indeterminacy appears: with l types of land in use, $l - 1$ additional zero-profit equalities are again present, but now, since the last type of land to be brought into production is no longer constrained to have a 0 rental rate, there are l additional endogenous factor prices.

The Sraffa literature concentrates on the horizontals rather than the verticals, in line with the Sraffian tradition of taking demand to be exogenous. If the demand for j were completely inelastic – unresponsive to price – then it would be an unlikely accident if this inelastic demand happened to coincide with one of the verticals of the supply function. The horizontals have the added advantage for Sraffians that any small shift in demand will leave the equilibrium at the same step and hence have no effect on any price. From the neoclassical point of view, completely inelastic demand seems far-fetched and does not obtain even when agents have Leontiev utilities. An inelastic demand argument for the horizontal equilibria also sometimes has no bite. When $N = 1$ there is no division of demand into separate outputs; only an inelastic accumulation of the produced input can then allow escape from an extra dimension of indeterminacy.

But in the $N > 1$ case and if we grant an elastic demand function for j, the additional indeterminacy of the vertical equilibria is hardly a reason for worry. More land is brought into cultivation because of demand-led increases in p_j; at a vertical equilibrium the demand for j therefore can pin down p_j and determine each rental rate. While the vertical equilibria dash the Sraffian hope to show that demand is locally irrelevant for price determination, they have no broader significance. The ease with which demand disposes of additional dimensions of indeterminacy underscores the pressing need for demand functions in the Sraffa model; without explicit demands, we will never be able to check whether any apparent case of indeterminacy is the genuine article.

Outside of our attention to extra dimensions of indeterminacy, the above account of rent stays close to Sraffa (1960). Sraffa pays heed to the supply-and-demand restrictions on rental rates: he complies with the rule that factors in excess supply must have a zero price and argues that when the scale of production expands the price of a scarce factor used in production should increase. While Sraffa applies these principles only to land, they pertain to labour as well, as we will see in Exhibit C.

Exhibit B. joint production

The simplest case of joint production occurs when $N = 2$ and there is no land. Profit maximization then requires

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$p_1 b_{1i} + p_2 b_{2i} \leq (1+r)(p_1 a_{1i} + p_2 a_{2i}) + w\ell_i, \tag{2.9}$$

for each activity *i* and with equality if $y_i > 0$. We again consider two types of equilibria. In the first type, one activity *i* is in use and each of the unused activities satisfies (2.9) with strict inequality. Then just one equality constrains the four prices $(p_1, p_2, w, 1+r)$, and so, given the normalization (2.3), there are two dimensions of indeterminacy. In the second type of equilibrium, two activities are in use, in which case just the standard single dimension of Sraffian indeterminacy obtains.
Dual to the dimensions of indeterminacy in the two types of equilibria are the dimensions of possible net productions of goods. In first type, the net production in a steady state must lie in the one-dimensional cone

$$\{((b_{1i} - a_{1i})y_i, (b_{2i} - a_{2i})y_i) : y_i \geq 0\}$$

while in the second type, with activities *i* and *j* active, net production lies in the two-dimensional cone

$$\{((b_{1i} - a_{1i})y_i + (b_{1j} - a_{1j})y_j, (b_{2i} - a_{2i})y_i + (b_{2j} - a_{2j})y_j) : (y_i, y_j) \geq 0\}.$$

The first type of equilibrium might therefore seem implausible: if labour is inelastically supplied, then this supply determines y_i and hence pins down exactly one vector of net outputs in the one-dimensional cone. But this restriction does not undermine the one-activity equilibria; there is ample room for the relative price p_1/p_2 to equilibrate demand to the fixed supply. While the one-activity equilibria therefore cannot be dismissed as pathological, it would seem, as in the extensive rent example in Exhibit A, that the additional indeterminacy will vanish once we introduce explicit market-clearing conditions. See the ‘sheep’ example in Sraffian economics for an illustration of how an economy can move from one type of equilibrium to the other as demand shifts. Many Sraffians contend that the two-activity equilibria – called ‘square’ since the number of activities in use equals the number of goods – are more likely (Steedman, 1976; Schefold, 1978a; 1978b; 1988; Lippi, 1979). One interesting rationale for this view, Schefold (1990), argues that, if agents always consume goods in fixed proportions, as with Leontiev utilities, then price adjustment will not be able to bring demand in line with supply in the one-dimensional cone. But if the fixed proportions of consumption vary from person to person, then a change in p_1/p_2 will have a differential effect on the scale of consumption of dissimilar agents. The ratio of demands for the two outputs will then vary with p_1/p_2 and equilibration can occur (Salvadori, 1982; 1990; Bidard, 1997; and see also the Samuelson–Schefold interchange in Bharadwaj and Schefold, 1990, and for an overview of the extensive literature Salvadori and Steedman, 1988).

Exhibit C. choice of activities

So far, we have considered how an extra dimension of indeterminacy can arise. When a choice of activities is available to produce one or more goods even the standard single dimension of Sraffian indeterminacy can disappear. Let each activity produce just one good. We suppose there is no land and consider equilibria – $(p, w, 1+r, y)$ that satisfy (2.1) and (2.3) – where each good is produced. If we ignore whether the labour market clears, the standard single dimension of Sraffian indeterminacy will obtain and we may index equilibrium prices $(p, w, 1+r)$ by *w*. For most values of *w*, and if we bar flukes of the production coefficients, the resulting equilibrium $(p, w, 1+r)$ will permit exactly *N* activities to satisfy their 0-economic profit conditions, that is, earn exactly the rate of return *r*. If an additional $N+1$st activity were required to satisfy its 0-profit condition, then the prices $(p, w, 1+r)$ that satisfy (2.1) and (2.3) will be pinned down. Hence, with a menu of only finitely many activities available, there are only finitely many *w* at which $N+1$ zero-profit equalities could be satisfied at an equilibrium obeying (2.1) and (2.3) – these are the ‘switch points’ at which the economy moves from one set of *N* cost-minimizing activities to another. Since there are only finitely many switch points, the required values for *w* might seem like flukes. But once we impose market clearing for labour, equilibrium can demand a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
switch point and Sraffian indeterminacy then disappears.

For an example, let $N = 1$ and thus by normalization $p_1 = 1$. Suppose two activities are available with input coefficients (a_{11}, $\bullet_1$) and (a_{12}, $\bullet_2$). If both activities are in use, then the profit-maximization condition (2.1) and the normalization (2.3) reduce to the two equalities

$$1 = (1 + r)a_{11} + w\ell_1,\ 1 = (1 + r)a_{12} + w\ell_2. \tag{2.10}$$

Flukes in coefficients aside, (2.10) will determine a unique $(w, 1 + r)$. In contrast, if only one activity is in use and the idle activity makes strictly negative profits (its inequality in (2.1) is strict), then the standard single dimension of Sraffian indeterminacy obtains. But now consider market clearing, which we impose on both labour and the material input. If labour is inelastically supplied in the quantity $e_\bullet$ and if e_c is the amount of the material input accumulated each period, then full employment of the input and labour requires

$$a_{11}y_1 + a_{12}y_2 = e_c,\ \ell_1 y_1 + \ell_2 y_2 = e_\ell, \tag{2.11}$$

where if $y_i > 0$ then activity i satisfies its zero-profit condition with equality. Evidently equilibria with both activities in use can be robust (they are not accidents of the parameters). For typical values of the model's parameters – e_c, $e_\bullet$, and the a_{ij} coefficients – (2.10) and (2.11) will have a unique solution $(\bar{w}, 1 + \bar{r}, \bar{y})$ and any small variation in the parameters will through a small adjustment of $(w, 1 + r, y)$ lead to a new unique solution. So if we begin with a model that has a two-activity equilibrium $(\bar{w}, 1 + \bar{r}, \bar{y})$ that is strictly positive in each coordinate then as the model is perturbed a two-activity equilibrium will continue to be present. Marginal products for e_c and $e_\bullet$ are also well-defined at these equilibria and equal $(w, 1 + r)$. We have taken the savings/accumulation level e_c to be exogenous, but we could let e_c be a function of the prices $(w, 1 + r)$ without affecting the robustness of the two-activity equilibria. Indeed, a two-activity equilibrium could well be the unique equilibrium – as when, for example, the accumulation level e_c is increasing in r. (That is, if $(1 + r, w)$ and $(1 + r', w')$ both satisfy (2.10) and $r > r'$ then e_c is strictly larger with $(1 + r, w)$ than with $(1 + r', w')$.) The robustness argument in no way hinges on there being a single material good. With two goods and three activities, the analogues to (2.10) to (2.11) would each consist of three variables and equations, and again indeterminacy would disappear.

The Sraffian dismissal of cases where $N + 1$ activities are in use rarely receives explicit defence. The rationale presumably is that the labour market is not required to clear, in which case there is no reason to suppose that w should equal one of the unusual values where $N + 1$ activities all earn the same rate of return.

The indeterminacy-reducing effect of factor-market clearing has already appeared. In the $N = 1$ example in Exhibit A we saw the dimension of indeterminacy drop from 2 to 1 when two activities are in use rather than one. Indeed, that example and the present example are essentially the same: r and the rental rate on land were the endogenous price variables in Exhibit A whereas r and the wage are the endogenous price variables here (w also appears in Exhibit A but we treated w as a parameter and ignored labour-market clearing). And just as in Exhibit A, the one-activity-in-use equilibrium requires a special configuration for (e_c, $e_\bullet$): when one activity i is in use and $(w, 1 + r) \gg 0$ we must have $e_c = e_\ell a_{1i} / \ell_i$. Similar conclusions hold when $N > 1$ (see Section 3).

Sraffians cannot have it both ways: if the case against an extra dimension of indeterminacy in the presence of land – that the special resource configurations are unlikely – is compelling, then consistency would seem to demand rejection of the single dimension of indeterminacy in the classical Sraffa setting. Of course, one may argue instead that labour unlike land is traded in a market that does not clear. But then the indeterminacy of the wage becomes an assumption rather than a conclusion: in any model where the labour market does not clear, the wage can be treated as a free parameter. We will expand on this point in the next section.

Gathering our exhibits together, we can summarize concisely what determines the extent of indeterminacy. Counting labour as an input, the dimension of indeterminacy equals the difference between the number of positively priced (hence fully utilized) inputs and the number of activities in use (see factor prices in general equilibrium). In basic Sraffian indeterminacy, $N + 1$ inputs are used by N activities: hence 1 dimension of indeterminacy. In the $N = 1$ extensive-rent example with 1 activity in use, 3 inputs are used by 1 activity: 2 dimensions of indeterminacy. In the extensive-rent example where l types of land are in use and all l are fully utilized, $N + 1 + l$ inputs have a positive price and are used by $N + l - 1$ activities: again 2 dimensions of indeterminacy. In the joint production example with 2 goods, 3 inputs have a positive price: 2 dimensions of indeterminacy occur when 1

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
activity is in use and 1 dimension of indeterminacy occurs when 2 activities are in use. Finally, in our choice-of-activities example indeterminacy disappears when 2 positively priced inputs are used by 2 activities.
As we will now see, a general equilibrium account of factor-price indeterminacy also reports a dimension of indeterminacy equal to the difference between the number of scarce inputs and the numbers of activities in use.

3 Sraffian indeterminacy with explicit market clearing

The forces of supply and demand have been nipping at the heels of Sraffian indeterminacy: the labour market clearing requirement in Exhibit C sometimes eliminates indeterminacy, and our informal appeals to output demand functions appeared to eclipse the extra indeterminacy that arose in Exhibits A and B. So perhaps a careful inclusion of market clearing for all goods will snuff all the indeterminacy out. This turns out not to be the case.
We include labour among the markets that are required to clear. The only formal feature of labour in Sraffian models that distinguishes it from land is that homogenous labour is used in every sector whereas a specific type of land need not be. (In reality, different varieties of labour are used in different industries and some type of land is used in every industry.) So we treat labour (and stocks of other inputs) as we have previously modelled land: for a positive price to rule, demand must equal supply. Labour markets are distinctive of course – labour can require an efficiency wage to induce effort, wage contracts can serve as decades-long insurance contracts, workers can be in unions, and so forth – and perhaps these special traits lead labour markets not to clear. But if we simply exempt labour from market clearing by fiat, one purpose of the Sraffa model is undermined. Wage indeterminacy will obtain whenever the labour market does not have to clear – whether or not capital aggregates, relative prices are constant through time, or linear activities describe technology. Models that allow the labour market not to clear in effect *assume* that markets do not pin down the distribution of income; they do not demonstrate that principle.
We now distinguish explicitly between material goods when they are inputs at an earlier point in time and outputs at a later point. Two periods will be enough; material inputs and labour will be supplied inelastically at time 1, and labour supplied and output sold at period 2. Relative prices will no longer be restricted to remain proportional through time. Relative prices that vary from period to period run counter to Sraffian tradition but are indispensable: if indeterminacy is to survive in the presence of market clearing, additional free price variables are imperative.
The prices of the *N* material inputs supplied at time 1 will be denoted $p^1 = (p_1^1, \ldots, p_N^1)$ while the prices of the goods sold at time 2 will be $p^2 = (p_1^2, \ldots, p_N^2)$. As in the basic Sraffa model, suppose just *N* activities are available, one for each produced good, and let *y* denote the activity levels. Output demand is given by the demand function

$$x(p^1, p^2, w, 1+r) = (x_1(p^1, p^2, w, 1+r), \ldots, x_N(p^1, p^2, w, 1+r)),$$

and the exogenous supply of labour and material inputs is given by e_ℓ and the *N*-vector *e*. Together these ingredients must obey Walras' law:

$$\frac{1}{1+r} p^2 \cdot x(p^1, p^2, w, 1+r) = \frac{1}{1+r} w e_\ell + p^1 \cdot e.$$

An equilibrium at which $(p^1, p^2, w, 1+r, y) \gg 0$ satisfies

$$p_i^2 = (1+r)(p_1^1 a_{1i} + \ldots + p_N^1 a_{Ni}) + w\ell_i, \; i = 1, \ldots, N, \tag{3.1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$x_i(p^1, p^2, w, 1+r) = y_i, \; i = 1, \ldots, N, \tag{3.2}$$

$$a_{j1}y_1 + \ldots + a_{jN}y_N = e_j, \; j = 1, \ldots, N, \tag{3.3}$$

$$\ell_1 y_1 + \ldots + \ell_N y_N = e_\ell, \tag{3.4}$$

$$p_1^1 + \ldots + p_N^1 = 1, \tag{3.5}$$

$$p_1^2 + \ldots + p_N^2 = 1. \tag{3.6}$$

There are two normalizations, (3.5) and (3.6), since the model uses an interest rate *r* rather than present value prices. The explicit inclusion of demand ensures that the model takes into account the indeterminacy-reducing effect of demand that we saw in Exhibits A and B.

The equilibria described by the above equalities typically will exhibit indeterminacy. If we fix *y* at an equilibrium value, then the market-clearing equalities for the material inputs and labour, (3.3) and (3.4), will remain satisfied as $(p^1, p^2, w, 1+r)$ varies. Of the remaining $2N+2$ equalities, one of the remaining market-clearing or zero-profit equalities is redundant due to Walras' law. But since the remaining $2N+1$ equalities have the $2N+2$ endogenous variables $(p^1, p^2, w, 1+r)$, one dimension of indeterminacy will typically obtain. The qualification 'typically' is necessary because a rank condition must hold in order to prove indeterminacy via the implicit function theorem (Mandler, 1999).

Several points give the above reasoning a Sraffian flavour. First, just as in Sraffa's book, aggregate quantities remain fixed and hence the indeterminacy operates on prices alone (though as the prices consistent with the fixed aggregate quantities change, individual incomes and individual consumption vary). Of course, unlike Sraffa, we know that markets

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

clear at the fixed aggregate quantities. Second, linear activities are essential. With a differentiable technology a given vector of aggregate quantities would be incompatible with multiple equilibrium price vectors (Mandler, 1997). Third, it is no accident that the present supply-and-demand model and the basic Sraffian model both display a single dimension of indeterminacy. The degree of indeterminacy is the same because first, with inelastic input supply the market-clearing conditions for inputs do not restrict prices, and second, the $N - 1$ independent market-clearing equalities for output are exactly counterbalanced by the $N - 1$ new second-period prices (we lose one price due to the added normalization (3.6)). This match between added equilibrium conditions and added prices variables means that if we open the door to the variations considered in the previous section – inelastically supplied land, joint production, choice of activities – the dimensions of indeterminacy of the two approaches will still coincide. In both models, the dimension will equal the difference between the number of inelastically supplied first-period factors that have a positive price and the number of activities in use.

Indeterminacy therefore need not always obtain when the Sraffa price equations are embedded in a supply-and-demand model. If more than one activity per produced good is available, then equilibria may have $N + 1$ rather than *N* activities in use. This possibility should come as no surprise since the $N = 1$ no-indeterminacy example in Exhibit C had two activities in use. Indeed if $N = 1$ and we introduced choice among activities, the present supply-and-demand model would be exactly the model in Exhibit C (the market-clearing equality omitted in Exhibit C is superfluous by Walras' law and (3.5)–(3.6) imply $p^1 = p^2$). Despite their neglect in the Sraffian literature, equilibria where the number of activities in use exceeds the number of produced goods are perfectly plausible if all factor markets are required to clear.

To complete the case for the compatibility of Sraffian indeterminacy and market clearing, we must deal with a famous counterargument that with overwhelming probability any equilibrium will have at least as many activities in use as positively priced factors (Mas-Colell, 1975; Kehoe, 1980) – we faced a similar argument in Exhibit A. Conditions (3.3) and (3.4) consist of $N + 1$ equalities in the *N* unknowns *y*; hence for almost every endowment (*e*, $e_{\bullet}$) there will exist no *y* that obeys these equalities. Consequently for these generic endowments there will be no equilibria satisfying (3.1)–(3.6). If only these *N* activities are available, then at the generic endowments one of the material inputs or labour will be in excess supply and have a 0 price. Hence for generic endowments the number of positively priced factors will not exceed the number of activities in use. But the seemingly unusual endowments (e, e_ℓ) at which (3.3) and (3.4) do have a solution can arise systematically (see factor prices in general equilibrium and Mandler, 1995). The material endowments *e* are the outcome of past savings–investment decisions; agents will not knowingly accumulate so much of a material input that it ends up in excess supply. Even when resources can be used productively no matter how great their supply, the endowments where $N + 1$ resources are used by *N* activities can still arise – those endowments appear at kinks on the production possibilities frontier (for example, see Figure 1 where the material input is accumulated just to the point where the available land is fully utilized using only one activity). This view of capital as a set of accumulated goods rather than a random endowment of nature fits well with Sraffa's view of production as circular.

We have had to do some damage to the Sraffian tradition to embed its indeterminacy claims in a market-clearing model. Inelastic factor supply finds no echo in the Sraffa literature. Not all factors have to be supplied inelastically for indeterminacy to obtain – it would be enough if some subset of *k* factors with positive prices were supplied inelastically and were used by fewer than *k* activities – but some must be.

More heretical from the Sraffian point of view, we have had to let relative prices vary across time periods. Time-varying prices allow output prices to clear the output markets without constraining input prices. The $N = 1$ case obscures this feature of the indeterminacy since then there are no relative output prices. But when $N > 1$ relative prices will be constant through time in a market-clearing model only if the economy is in a steady state, and steady states typically are determinate. For example, an overlapping generations model, even with a linear activity analysis technology, has locally unique steady states (Mandler, 1999). Indeed steady states will be determinate in virtually any model where markets clear and saving responds to the rate of return (including Marxian models where investment is increasing in the profit rate). There are two reasons. First, in any given period of a steady state, the prices of that period's given stocks of producible factors are constrained to equal the prices of the same factors being produced for the next period; the indeterminacy arguments we gave earlier therefore cannot be applied. Second, for factors that are not produced, endowment levels then *should* be seen as random parameters, and it would therefore be a fluke if some set of *k* such factors were used by fewer than *k* activities without one of them being in excess supply.

4 Sraffian instability

One may also read the Sraffian critique of neoclassical economics as arguing that the failure of capital to aggregate can lead the savings–investment market to be unstable. The case for instability relies on 'reverse capital deepening', where the ratio of the value of an economy's capital goods relative to the number of workers employed increases as a function of the interest rate. Consider a constant-relative-price Sraffa model with one or more activities available to produce each good, no joint production and hence no fixed capital, and where the economy is in a steady state. Set some consumption good to be the numéraire. Then, if we fix the economy's vector of outputs per worker, the ratio of the value of capital to labour is well-defined at any given *r*. An increase in *r* can affect the value-of-capital to labour ratio in two ways. First, for any produced good *j*, a 'real effect' can change which activity produces *j* at lowest cost, which in turn will alter the vector of capital goods per worker used in the production of *j*. Second, even if the activities in use remain the same (and with the composition of output still fixed), a change in *r* will affect the relative prices of capital goods. This 'price effect' will typically change the value of the capital goods each worker uses. So although the real effect of an increased interest rate might lead to a decrease in the quantity of each capital good used per worker, the price effect can cause the value of capital per worker to rise; hence reverse capital deepening can occur. This possibility can appear in an economy with one consumption good and just one capital good. An increase in the interest

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

rate will then lower the physical capital to labour ratio, but the price of the capital good can increase by enough that the ratio of the value of capital to labour rises (Bloise and Reichlin, 2009). Since this chain of events can happen with a single capital good, it is, strictly speaking, misleading to identify reverse capital deepening with the failure of capital goods to aggregate.

Here is the instability scenario. To ensure that Sraffian 'long-run' prices rule, we must assume that before and after a shock the economy is in a steady state, where relative prices are constant. If for simplicity there is no fixed capital, the steady state assumption implies that the value of capital per worker equals investment per worker. Consequently, with reverse capital deepening, investment per worker can exceed savings per worker when the interest rate is above its equilibrium value, thus pushing the interest rate higher, further away from equilibrium. Similarly, investment per worker can fall short of savings per worker when the interest rate is below equilibrium, pushing the interest rate lower.

The difficulty with this reasoning is that there is no market where long-run or steady-state savings and investment meet. When a shock to savings or investment occurs, the only instability that could undermine the economy must appear in markets at some specific set of dates – presumably the markets concurrent with the shock. Those markets, however, can equilibrate only at the out-of-long-run-equilibrium prices that obtain when the economy with the pre-shock endowments of resources begins its transition to a new post-shock long run. Long-run prices are therefore of dubious pertinence to the stability issue.

Garegnani (2000) and Schefold (2005) have argued that Sraffian instability surfaces in market-clearing general equilibrium models as a failure of tâtonnements to converge to equilibrium prices. This innovation in the Sraffian agenda has cleared away the cobwebs from the well-rehearsed interchange where Sraffians grouse that capital goods do not aggregate and Walrasians reply that equilibria in the Arrow–Debreu model exist.

The traditional model of a tâtonnement does not apply to an economy with linear activities: whenever positive economic profits can be earned, firms will want to expand without bound, and hence excess demands and tâtonnement price adjustments will be ill-defined. The most detailed and specific proposal to embed Sraffian instability in a general equilibrium model, Schefold (2005), steps around this problem by assuming that output prices are always set so that no activity makes positive profits, and letting the tâtonnement operate only on factor prices, which are the primary object of interest. Given a vector of factor prices, one may calculate the prices for outputs that minimize costs and then the consumer demand for outputs that result from these factor and output prices. The profit-maximizing decisions of firms, assuming they produce these output levels, then determine factor demands, and the difference between factor demand and factor supply leads to a tâtonnement price adjustment. Even in this setting, factor demand can be multi-valued since there can be many cost-minimizing factor combinations that produce any given output vector. To tackle this problem, one may define the tâtonnement with a differential inclusion rather than a differential equation (Mandler, 2005, and for differential inclusions generally see Aubin and Cellina, 1984).

Goods are distinguished by the date at which they appear and are of two types, *factors* which are not produced and *outputs* which can be produced. Factors can include the initial period's stock of capital goods as well as various types and dates of labour, land, and raw materials. Technology is given by a matrix of linear activities A where each activity (a column of A) produces only one output but may use any of the non-produced factors and produced goods as inputs. We follow the standard sign convention where positive entries in A denote outputs and negative entries indicate inputs, index goods so that outputs come first and factors second, and assume that positive quantities of all outputs can be produced simultaneously. To permit intermediate goods, an output can have negative as well as positive entries in A. Let A_o denote the output rows of A and let A_f denote the factor rows of A. For an arbitrary vector of factor prices p_f, we may find the competitive output prices $p_o(p_f)$ by solving the cost minimization problem $\min y - p_f A_f y$ subject to $A_o y \geq (1, \ldots, 1)$, $y \geq 0$, and setting $p_o(p_f)$ equal to the Lagrange multipliers at a solution to this problem. With output prices set in this way, consumers' output and factor excess demands become functions of p_f alone. Let $x_o(p_f)$ and $x_f(p_f)$ denote these functions, which we assume obey Walras' law. The demand for factors by firms is a x_f in the set

$$X_f(p_f) = \{x_f : x_f \text{ maximizes } (p_o(p_f), p_f) \cdot (x_o(p_f), x_f) \quad \text{subject to } (x_o(p_f), x_f) = Ay,\ y \geq 0\}.$$

As we mentioned, $X_f(p_f)$ may have multiple elements.

An equilibrium is a p_f such that $x_f(p_f) \in X_f(p_f)$. A *factor tâtonnement* is then a function $p_f(t)$, differentiable almost everywhere, such that when differentiable there is a $x_f \in X_f(p_f(t))$ with

$$\dot{p}_f(t) = x_f(p_f(t)) - x_f.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Due to the sign convention governing A and since $x_f(p_f)$ is an excess demand, $x_f(p_f)$ and x_f will both usually be negative.

Since the factors can appear at any date, the model can cover a classical Sraffa economy with just labour and produced goods as inputs, so long as the economy is finite-lived: the factors subject to the tâtonnement would consist of the initial period's stocks of capital goods and labour at all dates, while all capital goods that appear after the initial period would be classified as outputs.

While our initial story of reverse-capital-deepening instability was driven by responses of the value of capital to the interest rate, comparable stories are possible that refer just to the non-produced factors that appear in the above factor tâtonnement. Suppose in the classical Sraffa setting that some bundle of activities is cost-minimizing at both low and high r's and some other set of activities is cost-minimizing at intermediate r's (this is called 'reswitching'). And suppose further that a steady state with a small labour supply will use one of these sets of activities and a steady state with a large labour supply will use the other set. Then, if the economy initially is in a steady state at either a high or intermediate r and has the small labour supply, an exogenous shift to the large labour supply would lower r in a new steady state and hence raise w, hardly the intuitive price response to a supply increase. This tale compares steady states and tracks the movement of the wage through time, whereas all prices in a factor tâtonnement respond simultaneously to disequilibrium. Schefold (2005) nevertheless suggests that reswitching can lead a factor tâtonnement to be unstable.

Evaluation of this claim faces an immediate difficulty: no matter how well-behaved firms' factor demands are, consumer behaviour, which here appears as the factor supply function $x_f(p_f)$, can by itself lead to instability. To block this path, let us assume that demand obeys the weak axiom, the traditional tool used in general equilibrium theory to tame an exchange economy's demand function and ensure tâtonnement stability. In the present setting the weak axiom states that, for any p_f and $p_f', p_o(p_f') \cdot x_o(p_f) + p_f' \cdot x_f(p_f) \leq 0$ and $(x_o(p_f'), x_f(p_f')) \neq (x_o(p_f), x_f(p_f))$ imply $p_o(p_f) \cdot x_o(p_f') + p_f \cdot x_f(p_f') > 0$. Just as in an exchange economy, the weak axiom implies that a factor tâtonnement is stable (Mandler, 2005). Thus, no matter how many potential capital theory paradoxes are packed into the technology, if price adjustments are guided by excess demand and demand obeys the weak axiom, stability obtains. In fact, if the weak axiom is satisfied then in a factor tâtonnement the distance between the out-of-equilibrium prices that the auctioneer calls out and any equilibrium price vector will decline monotonically.

A tâtonnement is a highly artificial model of how an economy responds to disequilibrium: price adjustments are governed by 'notional' consumer demands, which cannot be satisfied at nonequilibrium prices, rather than by rationed or constrained demands that could be. On top of this problem, an intertemporal tâtonnement requires the prices of goods that appear at different time periods to adjust simultaneously. Perhaps in a more realistic setting the paradoxes of capital theory will turn out to be a distinct source of instability – but the case remains to be made.

5 Back to growth theory

Sraffa saw the economy as embedded in time, with endowments of produced inputs determined by the accumulation of capital, and he modelled technology using the plausible primitive of linear activities, not production functions packaged with suspicious differentiability assumptions designed to make factor returns determinate. These points add up to an effective criticism of a supply-and-demand theory of factor pricing, but the details of the argument need to be rearranged. The impossibility of capital aggregation plays no role.

But does the Sraffian stress on capital aggregation at least serve as an effective criticism of growth theory and the parables of the Solow model? On the surface, the fact that in a comparison of steady states the value of capital per worker or consumption per worker can increase with the interest rate may appear to undermine boilerplate neoclassical maxims on how to allocate resources through time. For example, it might seem that increases in savings could raise interest rates and lower future consumption. Unfortunately, as in the analysis of stability, the Sraffian focus on steady-state comparisons and on the value of capital per worker misleads. The move from one steady state to another involves the adjustment of myriad individual consumption and savings decisions at multiple points in time: consequently the impact of a change in savings today on steady-state consumption can diverge from the impact on consumption at a specific future date with all other consumption levels held constant. And when capital goods and consumption goods are separate commodities, changes in the relative prices of capital goods can break the linkage between increases in savings – sacrifices of present consumption – and increases in the value of capital; thus an increase in savings that lowers r *and* the value of capital per worker is not remarkable.

Following Solow (1963), define an economy's rate of return between the present and some future date t as the return in consumption at t as present-day consumption is sacrificed. If there is one consumption good per period, the gross rate of return between the present and t is the ratio of the gain in consumption at t relative to the quantity of today's consumption forgone, holding consumption at all other dates fixed. No reference to the value of capital is involved. With this definition, the familiar neoclassical maxims reappear: with linear activities and free disposal production possibilities set, a sacrifice of consumption today must lead to a weak increase in consumption at t (holding all other consumption levels fixed) and the rate of return between today and t must weakly diminish in the quantity of consumption sacrificed. Comparisons of steady states can also be cleansed of references to the value of capital, but here a less than impressive set of claims is available. If again there is one consumption good per period and we avoid settings with infinitely many agents, such as the overlapping generations model, then any increase in steady-state consumption per worker entails an increase of the amount of at least one of the capital goods used per worker,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and the move to a steady state with higher consumption per worker requires a sacrifice of consumption at some set of dates prior to arrival at the new steady state. The no-free-lunch moral of neoclassical growth theory rears its head.

But beneath these broad conclusions lies a vein of caveats, so far unmined by the Sraffian movement. When technology is described by linear activities, the rate of return can differ depending on whether it is defined by decreases or increases in present consumption. The reason is that the production possibilities sets are convex may well exhibit a kink at precisely the consumption levels chosen either by private agents in a market economy or by a benevolent planner who maximizes a sum of utilities (see Figure 1, but read the axes as present consumption and date *t* consumption). In the market-economy case, an exact match between the market interest rate *r* and the rate of return can then fail to obtain, though *r* must lie between the lower bound given by the rate of return for small increases in present consumption and the upper bound given by the rate of return for small decreases. In the planning case, the mismatch is between agents' intertemporal rates of substitution and the technological rate of return. Neoclassical growth theory has largely ignored these discrepancies, perhaps because of the blinders of its long reliance on differentiable production functions. But as in static factor pricing, linear activities in a growth setting open up a conceptual gap between prices and material rates of return: *r* no longer has to align with the physical return to sacrifices in consumption.

The presence of a kink in a production possibilities set will hinge on the number of activities in use, just as with factor-price indeterminacy. An irony crops up here: it is only in an optimal growth exercise that maximizes agents' utilities that an allocation at which the production possibilities set is kinked normally would be selected. Consider a planner with access to linear activities and stocks of resources at dates from the present (period 1) to the distance future (period *T*) that can be used to produce a sequence of consumption levels (again one consumption good per period). Resources are inelastically supplied and an arbitrary number of intermediate capital goods is permitted. Any consumption sequence $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_T)$ that is on the economy's PPF must satisfy the property that, for any $t = 1, \ldots, T$, $\bar{x}_t$ solves the problem of maximizing x_t subject to $(x_t, (\bar{x}_i)_{i \neq t})$ being in the production possibilities set. Pick one such problem with $t > 1$ where, if we view $\bar{x}_1$ as a parameter, the solution $x_t(\bar{x}_1)$ is non-constant. Consider those activities in use at some initial solution whose usage levels change as $\bar{x}_1$ changes. If no subset of *k* of these activities utilizes or produces more than *k* goods at the initial optimum, then (barring flukes of the production coefficients) the function $x_t(\bar{x}_1)$ will be differentiable at the initial $\bar{x}_1$: the PPF is smooth. (The good x_t should not be included in the count of the number of goods utilized or produced.) In the remaining cases, the initial $\bar{x}_1$ is a point where $x_t(\bar{x}_1)$ is not differentiable and the PPF is kinked; here inputs are accumulated just to the point where some set of *k* activities uses or produces more than *k* goods. Since almost every $\bar{x}$ on the PPF does not sit at a kink, a planner who selects a consumption stream arbitrarily could safely ignore the nondifferentiable points and declare that given the selection $\bar{x}$ the forces of technology alone determine the marginal rate of intertemporal transformation between any two time periods. If, furthermore this planner decentralized the economy's investment decisions to private entrepreneurs the planner would have to choose this rate as the market rate of interest. Curiously, the Sraffian hostility to utility maximization and substitution in consumption (see Exhibits A and B) also leads to the conclusion that a $\bar{x}$ at a PPF kink is an unusual event; hence the Sraffian view comes to the aid of the neoclassical identification of interest rates with rates of technological transformation. On the other hand a planner who maximizes a sum of agent utilities could well choose a consumption stream at a kink (see Exhibit A and Figure 1). While the Sraffian emphasis on linear activities serves as a welcome corrective to the neoclassical habit of assuming that any function or surface is differentiable, in the end it is utility functions that prevent a linear activities growth model from providing a purely technological determination of the interest rate.

As we have seen, this lesson goes beyond growth theory. Market economies also gravitate to kinks on PPFs. Consequently, even when an economy has a single consumption good per period, which allows consumption output to be modelled with an aggregate production function, that production function may well not be differentiable when evaluated at the factor endowments that arise in equilibrium. Empirical work that relies on a differentiability assumption – for example, the classical growth-accounting estimates of total factor productivity (Solow, 1957; Kendrick, 1961; Denison, 1962) – is therefore subject to coherent Sraffian criticism.

6 Conclusion

The Sraffian insistence on linear activities casts critical light on the instinctive neoclassical habit of assuming that interest rates and marginal rates of transformation must be equal and that production functions must be differentiable. Another more abstract Sraffian principle proves just as illuminating: economic activity is ongoing, not a one-time exchange among agents with disparate endowments and preferences. As we have seen, it is when the endowments of capital goods are determined by rational accumulation rather than by chance that factor-price indeterminacy can appear.

The Sraffian view of equilibrium, which revives earlier classical ideas, fits well with subsequent mainstream developments. Modern macroeconomics, both new Keynesian and new classical, has rejected models where the government's actions, such as an aggregate demand stimulus, systematically surprise agents; instead government actions are governed by a distribution that agents know. The new understanding of expectations is not driven by a belief that agents are never surprised or never hold an incorrect model of the economy but in order to pinpoint results that are immune to invalidation as agents learn and adapt to their environment, a precept close to the Sraffian view of equilibria as ongoing. The Sraffian perspective has wide application. For example, while the production of capital endowments by past equilibrium activity can lead to factor-price indeterminacy, a similar dependence of the present on past equilibrium decisions can eliminate some disturbing features of other brands of indeterminacy (Mandler, 2002). In overlapping-generations indeterminacy, market clearing is compatible with agents at the beginning of economic time unanimously anticipating any future price path that lies in a multidimensional set. But if one sees an equilibrium as ongoing, rather than commencing anew each period, the indeterminacy problem disappears after an equilibrium gets under way. The agents in an economy that has

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

already followed an anticipated price path for a number of periods and that experiences no shock will continue to hold their previously formed expectations, and given those expectations equilibrium prices at each period will be locally unique.

See Also

- capital theory (paradoxes)
- determinacy and indeterminacy of equilibria
- factor prices in general equilibrium
- general equilibrium
- neo-Ricardian economics
- Sraffian economics.

Note on the dimension of indeterminacy of the basic Sraffa model
We may rewrite (2.4) as $p(I-(1+r)A)=w\ell$, where A is the matrix whose ith column is a_i and $\ell=(\ell_1,\ldots,\ell_N)$. Due to the homogeneity of (2.4) in (p, w), we can replace (2.3) with w=1 without changing the relative prices $\frac{1}{\|p\|}p$ in any solution $(p, w, 1+r)$ to (2.3) and (2.4) or changing the dimension of the set of solutions. If at a solution $(\bar{p}, \bar{w}, 1+\bar{r})\gg 0$ to $p(I-(1+r)A)=\ell$, $I-(1+\bar{r})A$ has rank N, then, for r near $\bar{r}$, $p=\ell(I-(1+r)A)^{-1}$ solves $p(I-(1+r)A)=\ell$. Hence locally there is a one-dimensional set of solutions. If $I-(1+\bar{r})A$ has rank <N, then $\{p: p(I-(1+\bar{r})A)=\ell\}$ has dimension $\geq$1, and so locally the solution set contains a set of dimension $\geq$1.

Bibliography

Aubin, J. and Cellina, A. 1984. *Differential Inclusions*. Berlin: Springer-Verlag.

Bharadwaj, K. and Schefold, B., eds. 1990. *Essays on Piero Sraffa*. London: Unwin Hyman.

Bidard, C. 1997. Pure joint production. *Cambridge Journal of Economics* 21, 685–701.

Bidard, C. 2004. *Prices, Reproduction, and Scarcity*. Cambridge: Cambridge University Press.

Bloise, G. and Reichlin, P. 2009. An obtrusive remark on capital and comparative statics. *Metroeconomica*.

Denison, E. 1962. *The Sources of Economic Growth in the United States*. Washington, DC: Committee for Economic Development.

Eatwell, J. 1982. Competition. In *Essays in Classical and Marxian Political Economy*, ed. I. Bradley and M. Howard. London: Macmillan.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays on Modern Capital Theory*, ed. M. Brown, K. Sato and P. Zarembka. Amsterdam: North-Holland.

Garegnani, P. 2000. Savings, investment, and capital in a system of general intertemporal equilibrium. In *Critical Essays on Piero Sraffa's Legacy in Economics*, ed. H. Kurz. Cambridge: Cambridge University Press.

Hahn, F. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6, 353–74.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Harcourt, G. 1974. The Cambridge controversies: the afterglow. In *Contemporary Issues in Economics*, ed. M. Parkin and A. Nobay. Manchester: Manchester University Press.

Kehoe, T. 1980. An index theorem for general equilibrium models with production. *Econometrica* 48, 1211–32.

Kendrick, J. 1961. *Productivity Trends in the United States*. Princeton: Princeton University Press.

Kurz, H. and Salvadori, N. 1995. *Theory of Production: A Long-period Analysis*. Cambridge: Cambridge University Press.

Lippi, M. 1979. *I prezzi di produzione*. Bologna: Il Mulino.

Mandler, M. 1995. Sequential indeterminacy in production economies. *Journal of Economic Theory* 66, 406–36.

Mandler, M. 1997. Sequential regularity in smooth production economies. *Journal of Mathematical Economics* 27, 487–504.

Mandler, M. 1999. Sraffian indeterminacy in general equilibrium. *Review of Economic Studies* 66, 693–711.

Mandler, M. 2002. Classical and neoclassical indeterminacy in one-shot versus ongoing equilibria. *Metroeconomica* 53, 203–22.

Mandler, M. 2005. Well-behaved production economies. *Metroeconomica* 56, 477–94.

Marglin, S. 1984. *Growth, Distribution, and Prices*. Cambridge, MA: Harvard University Press.

Mas-Colell, A. 1975. On the continuity of equilibrium prices in constant-returns production economies. *Journal of Mathematical Economics* 2, 21–33.

Quadrio-Curzio, A. 1980. Rent, income distribution, and orders of efficiency and rentability. In *Essays on the Theory of Joint Production*, ed. L. Pasinetti. London: Macmillan.

Salvadori, N. 1982. Existence of cost-minimizing systems within the Sraffa framework. *Zeitschrift für Nationalökonomie* 42, 281–98.

Salvadori, N. 1986. Land and choice of techniques within the Sraffa framework. *Australian Economic Papers* 25, 94–105.

Salvadori, N. 1990. Comment. In *Essays on Piero Sraffa*, ed. K. Bharadwaj and B. Schefold. London: Unwin Hyman.

Salvadori, N. and Steedman, I. 1988. Joint production analysis in a Sraffian framework. *Bulletin of Economic Research* 40, 165–95.

Schefold, B. 1978a. Multiple product techniques with properties of single product systems. *Zeitschrift für Nationalökonomie* 38, 29–53.

Schefold, B. 1978b. On counting equations. *Zeitschrift für Nationalökonomie* 38, 253–85.

Schefold, B. 1988. The dominant technique in joint production systems. *Cambridge Journal of Economics* 12, 97–123.

Schefold, B. 1990. On changes in the composition in output. In *Essays on Piero Sraffa*, ed. K. Bharadwaj and B. Schefold. London: Unwin Hyman.

Schefold, B. 1997. *Normal Prices, Technical Change, and Accumulation*. London: Macmillan.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Schefold, B. 2005. Reswitching as a cause of instability of intertemporal equilibrium. *Metroeconomica* 56, 438–76.

Solow, R. 1957. Technical changes and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

Solow, R. 1963. *Capital Theory and the Rate of Return*. Amsterdam: North-Holland.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1976. Positive profits with negative surplus value: a reply to Wolfstetter. *Economic Journal* 86, 873–6.

How to cite this article

Mandler, Michael. "Sraffian economics (new developments)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000536> doi:10.1057/9780230226203.1599

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sraffian economics

Paul A. Samuelson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Sraffa's 1960 book *The Production of Commodities by Means of Commodities* has spawned an extensive literature but still needs to relate itself to the vast post-1945 finite mathematics literature of von Neumann, Dantzig, Kuhn-Tucker-Gail and Bellman. The Sraffa-Leontief circulating capital model provides by itself a prism with which to diffract the paradigms of Marx, Ricardo, and various brands of neoclassicism. It can serve to help judge Ricardo's editor and to illuminate the unity in Sraffa's scientific vision, from before 1926 until death in 1983.

Keywords

Arrow–Debreu model of general equilibrium; capital theory; circulating capital; Clark, J.; competitive equilibrium; Darwinian arbitrage; Engels, F.; factor-price trade-off frontier; Hawkins–Simon conditions; Heckscher–Ohlin trade theory; increasing returns; joint production; Keynes, J.; labour theory of value; Lange. O.; Lerner, A.; marginalism; Marx, K.; Marxist economics; non-substitution th; Pareto efficiency; Pareto optimality; rate of profit; Ricardo, D.; Robinson, J.; Samuelson, P.; Smith, A.; Sraffa, P.; Sraffa–Leontief circulating capital model; Sraffian economics; standard commodity; stationary state; Stigler, G.; Stolper–Samuelson th; surplus value; von Mises, L.; von Neumann model; Walrasian analysis

Article

Piero Sraffa, born in 1898 at Turin as the son of a well-off professor of law, lived from the twenties to his death in 1983 the quiet bachelor life of a don at King's and Trinity Colleges, Cambridge. Though his published and unpublished works are few, Sraffa has four claims to fame in the science of economics and the history of ideas.
(i) His 1926 article, ‘The Laws of Returns Under Competitive Conditions’, was a seminal progenitor of the monopolistic competition revolution. It alone could have justified a lifetime appointment.
(ii) An intimate of Keynes and Wittgenstein, Sraffa is said to have speeded Wittgenstein on his second philosophical road to Damascus by a rail station query, ‘What then is the meaning of this [Sicilian] gesture?’ The young Sraffa provided books and pin money to the marxist Antonio Gramsci jailed by Mussolini, and he

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
remained quietly interested in leftist matters. Sraffa was an organizer of the famous 1931–5 Cambridge ‘Circus’ which included Joan and Austin Robinson, Roy Harrod, James Meade and many others. Using Richard Kahn as messenger Gabriel, Maynard Keynes derived much benefit for his nascent *General Theory* from their brilliant group. Except for the chapter 17 discussion of *own* rates of interest, where Keynes must have benefited from Sraffa's 1932 polemic with Friedrich Hayek, there are few signs of a Sraffian interest in the macro-economics of effective demand. Sharing with Keynes an antiquarian's preoccupation with rare books, Sraffa and Keynes jointly discovered, identified, and edited the valuable *Abstract of A Treatise on Human Nature* that David Hume had published anonymously as a puff for his great initial work on philosophy.

(iii) Sraffa's editing of *The Works and Correspondence of David Ricardo*, a lone-wolf effort over a quarter of a century (aided much toward the end by Maurice Dobb) is one of the great scholarly achievements of all time, ranking in its perfections with the team efforts of the editors of Horace Walpole and James Boswell.

(iv) Finally, in the seventh decade of his life, Piero Sraffa published a classic in capital theory, *The Production of Commodities by Means of Commodities* (1960). As with Mozart if not Mendelssohn, Sraffa's death leaves posterity wistful that his full potential never came into print: what would we not give the good fairies, if somewhere in the attic of a country house there should be discovered a manuscript presenting Sraffa's planned critique of marginalism?

A fresh survey of Sraffa requires the reverse of a chronological order. First comes his 1960 book, which has spawned an extensive literature but still needs – if the technology is to be adequately handled – to have Sraffa's special equalities embedded in the general inequalities–equalities of the 1937 von Neumann model. The essentially completed Sraffa–Leontief circulating capital model provides by itself a prism with which to diffract the paradigms of Marx, Ricardo, and various brands of neoclassicism; and, self-reflexively, it can serve to help judge Ricardo's editor. The unity in Sraffa's scientific vision, from before 1926 until death at age 85, then becomes visible.

Truly general time phasing

The polemics of Böhm-Bawerk, Knight, and other capital theorists are illuminated, and seen through, by the 1937 von Neumann general model once that is made explicitly *time-phased* and open-ended, to allow (a) for *primary* factors (labour, land, …) not necessarily producible within the system, and (b) for net consumptions of outputs not necessarily ploughed back into the system for self-propelled growth.

We can summarize the *n* outputs produced by *J* activities, where each *j*th activity uses as inputs a vector of *M* primary factors and a vector of commodity-inputs, while producing a vector of joint products:

$$q_i^{t+1} = \sum_{j=1}^{J} b_{ij} \min\left[Q_{1j}^t / a_{1j}, \ldots, Q_{nj}^t / a_{nj}, L_{1j}^t / l_{1j}, \ldots, L_{Mj}^t / l_{Mj} \right] \quad (1)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$= Q_{i1}^{t+1} + \cdots + Q_{iJ}^{t+1} + C_i^{t+1},$$

$$i = 1, \ldots, n \gtreqless J, \quad Q_{ij}^t \geqq 0, \quad C_i^{t+1} \geq 0, \quad L_{mj}^t \geq 0$$

(2)

$$a = [a_{ij}] = \begin{bmatrix} a_{11} & \cdots & a_{1J} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nJ} \end{bmatrix} \geqq 0; b = [b_{ij}] = \begin{bmatrix} b_{11} & \cdots & b_{1J} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nJ} \end{bmatrix} \geq 0 l = [l_{mj}] = \begin{bmatrix} l_{11} & \cdots & l_{1J} \\ \vdots & & \vdots \\ l_{M1} & \cdots & l_{MJ} \end{bmatrix} \geq 0; \quad J \gtreqless M \gtreqless n$$

(3)

The *a*'s are input/output coefficients; *l*'s are Walrasian factor coefficients; *b*'s are joint-product proportions (as with 1 bu. wool and 2 pounds mutton per sheep). We use standard notation: for a matrix or vector z, $z > 0$ means all its elements are positive; $z = 0$ means all elements zero; $z \geqq 0$ means no element negative, with all possibly zero; $z \geq$ means no element negative, but at least one being strictly positive; z' is the transpose of *z*: thus, $[1 \cdots 1]'$ is a column vector of ones. *I* is the identity matrix, with 1's in the main diagonal and zero elsewhere:

$$I = [\delta_{ij}], \text{ as in } I = [1], \; I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ldots, \text{etc.}$$

For systems in a *stationary* state, regardless of the variable we have $z^{t+1} = z^t = z$. An important case is where supplies of primary factors are specified to be positive constants:

$$\sum_1^J L_{mj} = L_m, \quad m = 1, \ldots, M.$$

Remark: the circulating capital models occupying most pages in Sraffa (1960) are a special case of (3) where each column of (b_{ij}) can be written to contain a single ‘one’ with the remaining elements zero. In general, any or all a_{ij}'s could be assumed by a mathematician to be zero; but, to be interesting, every *b* must have at least one positive element in each of its rows and its columns, so that every good is produced somewhere and every activity produces at least one good. An activity might not require any *direct* primary input; but, if we

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

are not to be in the Land of Cockaigne, where lollipops grow freely on trees and don't even require picking, indirectly or directly every good must require some primary factor(s).

General Hawkins – Simon conditions

For a circulating-capital system to be ‘productive’, in the sense of providing positive net consumptions, *a* must satisfy simple Hawkins–Simon conditions, such as that powers of *a*, a^k go to zero as $k \to \infty$. When *b* involves joint products, matters are more complex and the literature needs the equivalent of the following two constructive tests:

Axiom 1: (No Land of Cockaigne). The following standard linear programming problem must have a solution of zero:

Subject to

$$bx \leqq 0, \quad x \geqq 0 [b - a] x \geqq 0, \max_x [1 \cdots 1] [b - a] x = Z^* = 0$$

Axiom 2: (Generalized Hawkins–Simon). The following standard LP problem will have a positive solution if, and only if, the system is ‘productive’.

Subject to

$$bx \leqq [L_1 \cdots L_M] > 0, \quad x \geqq 0 [b - a] x \geqq C \geqq 0 C \geqq (c \cdots c)' \geqq C \max_{x, C} c = c^* > 0$$

Example: Suppose, as in the joint-production passages of Sraffa (1960), that $(b - a)$ is $n \times n$, with $l = n$. It will then be *sufficient* that $b - a$ have row sums positive for the productiveness axiom to be satisfied. (Contrary to what seems to be suggested by the mathematician C. F. Manara (1979), these row-sum conditions are *not* necessary—as $b - a$ with diagonal elements near to 1's and off-diagonals of –10 and –0.01 demonstrates.)

Competitive pricing relations

In the absence of any uncertainty, or restrictions on entry and knowledge, perfect competition will be led in (1) by Darwinian arbitrage to equality of the profit rate in all processes positively operated. In matrix terms, with *P* a non-negative row vector of *n* goods prices, *W* a non-negative row vector of *M* primary-factor prices, and *r* the common profit or interest rate, we have the general dynamic equalities–inequalities:

$$p^{t+1} b \leqq \left(p^t a + W^t l\right)\left(1 + r^t\right), \quad \left(p^t, p^{t+1}, W^t, r^t\right) \geq 0$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(4)

$$P^{t+1}q^{t+1} = \left[P^t a x^{t+1} + W^t l x^{t+1}\right]\left(1 + r^t\right) = \left[P^t a q^t + W^t L^t\right]\left(1 + r^t\right)$$

(4′)

where x^{t+1} is the column-vector of the J 'intensities' at which the respective activities are carried on:

$$0 \leqq x_j^{t+1} \leqq L_{mj}^j / l_{mj}, \quad m = 1, \ldots, M; \quad j = 1, \ldots, J \leqq Q_{ij} / a_{ij}, \quad i = 1, \ldots, n \geqq q_i^{t+1} / b_{ij}$$

(4″)

The convention is understood in (4) and (1) that when a denominator vanishes, the term in which it appears is ignorable. In (4) P's and W's are measured in any numeraire and equality of r's in all processes used does not imply equality of different goods' *own* rates of interest when P^{t+1} is not proportional to P^t.
Wherever a strong inequality holds in (4), that activity must cease to operate, in accordance with the perfect-competition *duality* conditions:

$$0 = x_j^{t+1}\left[\sum_{i=1}^{n} P_i^{t+1} b_{ij} - \sum_{i=1}^{n} P_i^t a_{ij}(1 + r^t) - \sum_{m=1}^{M} W_m^t l_{mj}\left(1 + r^t\right)\right] j = 1, \ldots, J$$

(4•)

In stationary equilibrium, $(P^t, W^t, r^t)=(P, W, r)$ and all time scripts can be omitted. All *own* rates of interest are then the same.
Then (4) becomes

$$Pb \leqq (Pa + Wl)(1 + r)$$

(5a)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$Pbx = (Pax + Wlx)(1 + r)$$
(5b)

These equality-of-profit-rate conditions are not merely competitive arbitrage conditions. Taken together with the steady-state version of (1)–(2), namely with

$$(b - a)x \geqq C \geq 0, \qquad lx \leqq L > 0$$
(5c)

(5a)–(5b) are the necessary and sufficient conditions for *inter-temporal* production efficiency (or inter-temporal Pareto efficiency on the allocation side). Unless there exists, at the observed steady state, an existent (P, W, uniform r), the society must be planning wastefully. So aside from Marx's 1867 innovation being a backward step in positivistic realism, it was a bad (avoidable) blunder from the standpoint of social planning and efficiency – a point never glimpsed by Marx or his admiring editor Engels. Because Sraffa's brief text never grapples with the *inter-temporal* relations implied by *his* steady states, his readers are left unaware of this technocratic property of dualistic competitive pricing.
Sraffa (1960) considers for the most part very special cases of (5), and of (1)–(2), cases for which n happens to equal J with $b - a$ square and of rank n and with all equalities holding in (4). Save for the brief chapter on land, he mostly works with labour as the only primary factor and seems to presuppose that $[l_{11} \cdots l_{1n}][b - a(1 + r)]^{-1}$ happens to be positive in an open neighbourhood above r equal to zero. (*Remark*: in many technologies, even when J greatly exceeds n, the number of activities operated at a *positive* level will be equal to n; this endogenous fact should not obscure the more general truth, that changes in demand will generically alter the choice of n viable activities – so that, as will be seen, von Neumann's rectangular matrices cannot be sidestepped.)

Properties of competitive equilibrium

The following twenty properties of Sraffian steady-state systems are straightforwardly verifiable:
1. When goods require more than one primary input, as for example labour and land, a shift in demand away from a land-intensive good and toward a labour-intensive good, will at each ruling profit rate raise the price ratio of the latter relative to the former; and it will tend to raise labour's distributive share at the expense of land's. *The labour theory of value*, even in the absence of the complication of time-phasing and interest, *thus generally fails and the theory of distribution cannot be separated from the complications of value theory* (of supply–demand pricing theory).
2. Even when labour is the only primary factor, time-phasing means that Smith's bipartite formula of wage-*plus*-interest is indeed a necessary statement of price and of national income. Save in singular cases where all goods happen to have exactly the same percentage of direct-wage cost to total cost – what Marx called the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
case of ‘equal organic compositions of capital’ – a change in the interest rate must alter relative commodity prices – again vitiating the simple labour theory of value.

3. Suppose there are no joint products, all raw materials being used up in a single employment. It follows from the last paragraph that competitive prices in time-phased systems must generally at positive interest rates differ from the ‘marked-up values’ of Marx's *Capital* (1967, Volume I, ch. IX) which replace a uniform industry-by-industry rate of profit by a uniform rate of surplus value: the 1867 marked up values mark up the direct wage costs only, with allegedly no mark up earned on raw-materials outlays for produced goods (for ‘constant’ capitals). Ratios of these peculiar ‘marked-up values’ agree with ratios of zero-profit-rate prices, and do so no matter how great the capitalists’ surplus! So, in consequence of paragraph numbered 2 above, the 1867 Marxian constructs do systematically depart from realistic competitive prices and need to be ‘transformed’ into correct Sraffian competitive prices *by abandoning them* – as L. von Bortkiewicz had demonstrated in 1907.

4. Staying with the assumption of no joint products, we can deduce the existence of a *factor-price trade-off frontier*: at any specified Sraffian interest rate, there is defined a convex trade-off frontier between the maximal real wage rate of labour and real rent rate of land (where *any* good is the numeraire for measuring such real factor prices); a rise in the interest rate must shift inward the convex contour relating real factor prices (but equal upward increments of r can induce inward shifts in the frontier that both accelerate and decelerate). Though Sraffa does not use this name for the frontier, for the case where labour is the sole primary factor, he recognizes the properties of this basic frontier.

5. The amount of net consumptions produced in the stationary state will be maximal when the competitive system chooses those golden-rule techniques that are supported by a zero rate of interest. (If all primary factors grow at a Harrod natural rate of $[1+g]t$, the maximal per capita net consumptions will be realized only when competition mandates use of the golden-rule techniques supportable by an interest rate of g: $1 + r = 1 + g$.) This paragraph states something different from the earlier remarks around equations (5) concerning inter-temporal Pareto efficiency.

6. At any interest rate r, prevailing in Sraffa's time-phased system, the competitively viable techniques observed can be verified from (5) to achieve inter-temporal Pareto optimality. By contrast, consider the 1867 Marxian techniques that maximize the rate of surplus value for a specified vector of relative primary-factor prices. Not only are these techniques unrealistic describers of the positive facts about the laws of motion of capitalism. In addition they would achieve a Planner's nightmare, in general producing permanently in the steady state less of goods than the system is capable of and involving more of society's scarce primary factors than is technologically needed. When Sraffa's readers begin to worry about consumption preferences over different time periods, they will additionally have to revert back from 1867 Marxian values to (5)'s dualistic competitive pricing to restore *inter-temporal* Pareto efficiency of consumption.

How demand-tastes affect pricing and distribution

The above half-dozen rules of a Sraffian system are valid either for *his* von Neumann technologies or for *all* versions of neoclassical technologies (involving convexity and first-degree homogeneity). Although Sraffa reserved judgement for half a century on whether he wanted to assume constant returns to scale, experiments with returns laws that depart from that property will be found to rob his algebra of *any* interesting economic applications, as the paucity of results on this point in the literature of the last quarter of a century attests. (Thus, specify for the coal and iron example of the opening page of Sraffa (1960) that in the iron industry doubling inputs quadruples output, while in the wheat industry tripling inputs doubles output. Then none of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the pricing relations have other than empty defal content!) The Perron input/output matrices don't define *existent* prices of production.
Further properties of steady-state competitive systems are the following.
7. Prior to Sraffa (1960), the Leontief literature had established and generalized the 1949 Non-Substitution Theorem. When labour is the only primary factor and there are no joint products, at any observed profit rate the competitive prices at which all goods are positively produced must be independent of the composition of steady-state demand: even when alternative techniques could be reasonably substituted, changes in demand can *never mandate their new use*. Also, with labour the only primary factor, a shift in demand toward goods that involve a high relative fraction of direct-wage cost must be at the observed profit rate raise the wage-profit share. As already indicated, when labour must cooperate with land, shifts in demand toward or away from goods that are relatively labour-rather-than-land-intensive must at the observed competitive profit rate raise or lower their relative prices and presumptively alter the distribution of income between workers and landlords. The Ricardian dream of ridding distribution theory from the complications of (consumer-demand) value theory is seen to be, in general, a pipe dream – as Ricardo realized in his occasional lapses into good sense (as for example, when recognizing that Napoleonic-war shift of demand toward labour-intensive soldiering would raise the wage share prior to a repopulating of the countryside). At a given interest rate, in the absence of technical innovation and non-labour primary factors, a rise in the money wage rate cannot force a permanent substitution of machines for labour since machines' steady-state costs then rise proportionally with the wage rate.

Some joint-product phenomena

The following properties of joint-product systems are also common to Sraffian technologies of the von Neumann type and to *all* versions of neoclassical technologies.
8. When joint products are admitted – surely the realistic case – the classical economists' hope to deduce steady-state price ratios from technology and supply alone is *generally* frustrated. When one species of sheep produces wool and mutton in joint proportions – or when one round trip of a ship supplies east and west transportation jointly – each alteration in tastes and demands for the joint products alters their steady-state price ratio and does so even under perfect Arrow–Debreu certainty.
Sraffa's favourite case of joint products would involve as many independent activities used as there are goods: $l = n$. For a simple example, consider 2 sheep species, each producing 2 products: say, species 1 produces 2 of wool and 1 of mutton, while species 2 produces 1 of wool and 2 of mutton; for simplicity let both sheep require the same inputs and cost, which might as well be labour only.
So long as consumer's demand involves relative expenditures on the two goods not too unequal – no good ever attracting more than two-thirds of consumers' dollars – Sraffa is correct in expecting cost-technology alone to determine competitive pricing of $P_{\text{mutton}} / P_{\text{wool}} = 1.0$. But, as soon as people want to spend more than two-thirds of their incomes on mutton, Sraffa loses his equality of number of goods and number of positively used activities: only the meat-intensive species is competitively viable; consumer demand functions are then price determining.
Let us add a third sheep species, producible like the others but yielding 1.75 of wool and 1.75 of mutton. Now $l = 3 > 2 = n$. When people singularly spend exactly half their incomes on wool and mutton, only the species 3 is viable and $P_m / P_w = 1.0$ as set by demand. When people spend a bit more on mutton than wool, less than $l = 3$ activities are positively viable, namely, $2 = n < l = 3$, as species 3 and species 2 alone survive. For a range of demands, $P_m / P_w = 1 / 3$, a numerical value that can be calculated from Sraffa's 2 cost-prices of production relations. However, as the demand for mutton relative to that of wool runs the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
gamut of possible ratios, we reach the limit of each Sraffian horizontal step and must traverse the staircase's vertical risers with market price being consumer-demand determined. In realistic cases, *J* is large relative to *n*; and there are many different *n*-by-*n* square sub-matrices that consumer-demand functions will endogenously select out of the rectangular *n*-by-*J* matrix of technology.

In sum, globally demand generally helps determine relative prices of Sraffa's joint products between zero and infinity, doing so along a staircase's vertical risers and (locally horizontal) step segments.

This general rule of global dependence on demand of joint-products' price ratios does admit of an exceptional case where an aspect of a Non-Substitution theorem does obtain. Let us, so to speak, introduce so weak a degree of jointness of production that we are still in a close neighbourhood of indecomposable non-joint-production. This will occur when a positive *net* amount of each good is available only from a single process, and where each process does produce one such positive amount net. Under this stipulation the square matrix $(b_{ij} - a_{ij})$, after feasible renumbering of the goods and of the processes, will be specified to have positive elements in its diagonal and negative off-diagonal elements. Its general Hawkins–Simon conditions will also require that the inverse $[b - a]^{-1}$ exists and is positive. Under these strong stipulations, as is well known, the same Non-Substitution theorem that holds for circulating-capital and exponential-depreciation models will hold for joint production.

Otherwise, however, demand conditions can in general have an essential effect on relative prices; and can do so even when we grant Sraffa special indulgences (such as equality of *J* and *n*, non-singularity of $b - a$, and labour the only primary factor). Thus, often, even when $(l_{1j})(b - a)^{-1}$ is positive, some elements of $(b - a)^{-1}$ will be negative – with the result that for some ratios of consumptions, $(b - a)^{-1}C$ will not define positive feasible gross outputs, and price ratios will have to be influenceable by demand-tastes.

Sraffian artifacts: standard commodity baskets

The following properties are special and hold only for singular subsets of von Neumann technologies or for singular subsets of neoclassical technologies.

9. Postulate no joint products, no primary factors other than labour, no alternative techniques for producing goods, and that all goods considered are basics in the sense that a is indecomposable so that every good requires for its production something directly or indirectly of every good as input. Then, Sraffa (1960) deduces the existence of a market basket of goods or *standard commodity*, with unique positive weights $\left(Q_1^*, \ldots, Q_n^*\right)'$, such that the *real wage expressed in terms of the standard commodity* and paid to the workers postfactum (at the *end* of the period when they work at the beginning of the period) is a declining *linear* relation:

$$w = 1 - \left(r / r^*\right) \tag{6a}$$

where *r** is the maximum positive profit rate the system can pay, with the column vector (Q_i^*) uniquely

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
definable by the eigenvector relation:

$$a(1+r^*)\left[Q_i^*\right] = \left[Q_i^*\right] > 0\,[l_{11}\cdots l_{1n}]\,[I-a]^{-1}Q^* = 1 \tag{6b}$$

Some real prices rise faster with the profit rate than the standard commodity's price does; some rise less fast. When the real *standard*-basket wage rate is half its zero-profit-rate level, the post-factum wage share of the basket's cost is also one half; and similarly for any other fraction.

Sraffa, for reasons not easy to understand, thought that (6)'s truth somehow provided Ricardo with a defence for his labour theory of value. Even in the restricted case where (6) does validly obtain, one perceives no successful resurrection of Ricardo's desired labour theory of value or absolute standard of value that is provided by Sraffa's demonstration of the existence of this standard commodity: price ratios are still not equal to zero-profit-rate price ratios as the crude labour theory of value wants them to be.

10. It is by now understood that a Sraffian Standard Commodity often fails to exist, for a variety of reasons:

(i) In circulating-capital and exponentially depreciating models, as soon as competition mandates a switch in technology when interest rates or relative primary factor prices change, there will then generally not be any market-basket weights that entail a linear tradeoff frontier between the real wage and the profit rate.

(ii) In these same models, realistic decomposabilities can negate the existence of positive weights that yield linearity.

(iii) In still other cases, as shown by Takahiro Miyao in 1977, for various vectors of direct-labour coefficients, there can be an infinity of positive weights that produce the same normalized linearity as the Sraffian standard.

(iv) In joint-product cases, as C. F. Manara (1979) instanced. $(b - a[1+r])$ may have only complex eigenvalues and eigenvectors. Again, no Sraffian standard validity obtains. (For such Manara cases, at some feasible interest rates certain processes cease to be competitively viable; so Ricardo can never find his middling composite, which is neither too time-consuming nor too time-economizing, to provide him with the chimera of an absolute standard of value for making comparisons across time and space.)

(v) There exist many joint-product cases in which Sraffa's eigenvector of $(b - a[1+r])$ is real, but involves some negative elements. Playing the game of defining market baskets with negative weights cannot, by some analogy with foreign-trade exports and imports or debtor-creditor ownings and owings, make economic sense useful to a Ricardian critical of Adam Smith or of J. B. Clark.

(vi) There are many joint-product cases in which Sraffa's eigenvector for $(b - a[1+r])$ is all positive. But still there may be no standard market basket that yields a *linear* factor-price frontier validity applicable over the *whole* interval of feasible profit rates. (On one side of some critical r, all of Sraffa's $J(=n)$ processes are competitively viable. On the other side of that r, less than n processes can earn the competitive profit rate. The *true* market cost of Sraffa's nominated market basket there ceases to obey the linear law that allegedly Ricardo's value theory could benefit from.)

(vii) Instead of abandoning the hunt for a chimera, some Sraffians sought comfort in the belief that the important joint-product cases are not those of the wool-mutton type but rather are those of the new-machine-old-machine type or are of the permanently-durable-land type; and nursed a hope that for these important cases, the ‘pathologies’ of non-existent standard commodities might be absent.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

If anyone ever believed this, it was an illusion. Under (i) above, we saw land–labour models in which induced changes in a_{ij} coefficients take place, which induce violation of the Sraffian desired linearity. A locus linear piecewise is not linear.

Also, a model where the only jointness is of the durable-good type may well not admit of a standard commodity defined over the whole interval of feasible profit rates. The following counterex settles the issue of possible non-existence:

$$b = \begin{bmatrix} 1 & 1 \\ 7/8 & 0 \end{bmatrix}, \quad a = \begin{bmatrix} 7/8 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad [l_{11} l_{12}] = [1 \;\; 9] \tag{7}$$

Here is its story. Good 1 can be consumed directly or be used as a new machine to produce, in cooperation with labour, Good 1 itself along with the joint product of an old machine. The used machine, called Good 2, can be used with labour to produce Good 1 and worthless scrap. (Later, we'll recognize that Good 2 might be an object of final utility for its own sake, distinct from Good 1's marginal and total utility.)

This ex's numbers involve the need in Process 2 for (relatively) much labour and little of the old machine to produce 1 of Good 1. Process 1, by contrast, produces 1 of Good 1 with relatively little labour and a fair amount of new machines: for Process 1, labour with 7/8 of a new machine, produces 1 of a new machine and 7/8 of an old machine; for Process 2, 9 of direct labour, with 1/2 of an old machine, produces 1 of a new machine and scrap.

Calculation shows that competition must then work out as follows:

At very low profit rates, the used machine is so consuming of high-wage direct labour that it (and Process 2) cannot be used in production under viable competition. The price ratio P_2/P_1 will then be set completely by utility tastes for Goods 1 and 2: with enough yen for Good 2, P_2/P_1 can be anything from zero to infinity at low profit rates.

Between the critical profit rates of 1.59%, and 109.67% both processes could be used competitively. If both are useable throughout that interval, then over that interval (and only it) Sraffa will get his desired linear standard commodity; but it must fail him at the lower interval of profit rates. Worse still, even in the higher interval where algebra does yield him a linear function, economics can veto the relevance of Sraffa's standard function: as soon as people have strong enough marginal utility for Good 2, *all* of it is bid away from production uses! Process 2 becomes competitively unviable and Sraffian cost-prices lose relevance; demand becomes decisive, frustrating the primitive classicist's yearning to determine prices from supply considerations alone. The 1960 purported defence of Ricardo's absolute standard has collapsed.

Moral

The Walrasian paradigms are in general unavoidable in the most unrestricted von Neumann paradigm. What began as a classicist-inspired critique of ‘marginalism’ ends up as a demonstration that the classical-economics paradigm does need to be broadened into the post-1870 mainstream economics – even when smooth Clarkian marginal products are scrupulously avoided! Where a critique succeeds, and is needed, is in exposing how special are the one-sector, homogeneous-scalar-capital paradigms of Clark.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The genuine Fisher, von Neumann, Arrow–Debreu structure of time-phased general equilibrium stands confirmed by the Sraffa–Leontief probings.

1960 light on Clarkian oversimplifications

Heroic works clarifying faults of neoclassical parables have been done by Joan Robinson, Luigi Pasinetti, Pierangelo Garegnani, Bertram Schefold, J. S. Metcalfe and I. Steedman, C. F. Manara, V. K. Dimitriev, and many others.

11. Before she knew of Sraffa's 1960 model, Joan Robinson had usefully debunked the notion that some aggregate Platonic Kapital, $K = \Sigma P_j k_j$, enters into real-world production functions and defines (aggregate) *marginal net product* of Kapital, $\partial [K(t+1) - K(t)] / \partial K(t)$, to give the real-world interest or profit rate. Sraffa's model showed once and for all the falsity of the following neoclassical apologetics:

> Roundabout [‘mechanized’] methods are productive and the interest rate measures the incremental social product obtainable by extending the degree of roundaboutness through the effective action of *saving* (by the exercise of painful ‘waiting’ and ‘abstinence’).

Sraffa showed this: as soon as there are more capital goods than 1, it is impossible to say of every pair of techniques which one is the more roundabout, time-intensive, or mechanized. Increases in the interest rate above zero can first raise various P_2/P_1 ratios and then later lower them. So above a critical high interest rate, competition may revert back to a technique that had been viable only at very low interest rates.

> Sraffian *reswitching* thus implies: lowering the interest rate may result in lower steady-state consumption levels, prior to its ultimately raising them to the maximal golden-rule level.

Oskar Lange once wrote, sardonically but seriously, that Ludwig von Mises, the enemy of socialism, deserved a statue in the socialist Hall of Fame for compelling Abba Lerner and Lange (to say nothing of earlier Pareto, Barone and Fred Taylor) to work out how socialists might devise efficient decentralizing-pricing planning algorithms. One can insist, seriously, that a neoclassical Hall of Fame deserves Sraffa's statue.

1. 12. Actually all of Sraffa's findings about the impossibility of defining ‘more roundaboutness’ and ‘less roundaboutness’ apply precisely to smoothly differentiable *vector*-capital Clarkian models. Neoclassicists should reproach themselves, and thank Robinson and Sraffa, for belated recognition that the locus of (stationary-state consumption, interest rate) need not be a one-way trade-off. This recognition is achievable quite apart from the dramatic case of *double-switching*.

Example. Let $[L, K_1, K_2, C_1, C_2, \dot{K}_1 = dK_1/dt, \dot{K}_2]$ depict, for a 2-sector neoclassical economy in which marginal products are *not* illusory, its labour, its 2 heterogeneous capital good's stocks, its 2 consumptions, and its net investments. Its production function and steady-state profit-rate are given by

$$L = F[K_1, K_2; C_1 + \dot{K}_1, C_2 + \dot{K}_2]$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(8a)

$$r = -(\partial F/\partial K_i)/(\partial F/\partial \dot{K}_i), \quad i = 1, 2$$

(8b)

where F is a first-degree-homogeneous, smooth, convex function. Set L by convention at a plateau of unity, and for simplicity specify C_2/C_1 always to be unity; and, as a condition of stationarity, make all $\dot{K}_i$ vanish. Then the above pair of relations permit us to determine the level of consumption as a function of the profit rate: $C_1 = f(r)$. What neoclassicals insufficiently realized before 1960, away from $r = 0$ there is no necessity for C_1 to fall as r rises. This valid Sraffa–Robinson point does not score against 'marginalism'. Their razor cuts as deeply against the discrete-activity technology!

'Marginalism' per se is not what deserves the razor of Sraffa's critique. As already said, whatever Clarkian marginalism can display for good or ill is already capable of being displayed by a von Neumann discrete-activities technology.

Space permits only a few further mentions of paradoxical neoclassical phenomena that can result from positive profit rates in time-phased models. A critique to reveal them, as shown by several valuable works of Metcalfe and Steedman, cuts as much within von Neumann-activities models as within smooth marginalist models of the marginal-product type.

13. The Non-Substitution Theorem, which makes relative costs and prices, at each interest rate, independent of the composition of final demands when only one primary factor is present, is common to discrete and marginalist models. When more than one primary factor is present, the Non-Substitution Theorem becomes a Substitution Theorem – again, generally both for discrete and marginalist models.

14. Both for discrete and marginalist no-joint-product models, for each interest rate there is a convex trade-off between various real factor returns (expressed in terms of any good as numeraire). For both models, it is not necessarily true that the trade-off relation between the interest rate itself and real factor returns is convex. (Point 14 has a considerable overlap with Point 4.)

15. For both models, the substitutions of technique that win out under competition are Pareto efficient. For both models, any technique observed to be viable under competition is a golden-rule technique for that Harrod-growth rate which is equal to the observed interest rate!

16. For both models, when the real wage rises relative to the real land-rent (independently of the good used as numeraire), any mandated technical substitution must (if anything) involve a shift to lower embodied-dated-labour contents of each good and to higher embodied-dated-land contents. This is as true for a positive interest rate as for a zero one, since we deal exclusively with golden-rule technologies. (The difference, when r is zero, is that there is then no need for *dating* of the direct-and-indirect labour and land contents.)

17. As shown by Metcalfe and Steedman, Pasinetti, Samuelson, and others, in both models one must guard against a tempting fallacy.

In both models, when r is zero, the change in technique induced by a rise in the Wage/Rent ratio must economize on the actual Labour/Land ratios observed to be used in the various industries in the stationary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
state. Something like this remains true for very low interest rates. (For economy of space, explanatory qualifications are omitted.) However, for large enough *r*, what are called embodied-dated-labour contents and embodied-dated-land contents can be substantially different from synchronized-stationary-state labour and land actually observed to be used. So beware of believing that observed $(\Delta W_m)(\Delta L_m)$ must be negative when *r* is positive. This has implications for the correct Heckscher–Ohlin and Stolper–Samuelson trade theorems under time-phasing.

18. Related to the above is the following observation. At *r*=0, the competitive system is out on its true *stationary-state production-possibility-frontier*. At positive *r*, although the system moves to a new *steady*-state (not stationary-state!) golden-rule set of techniques, what it can produce and does produce in the stationary state is generally not out on the stationary-state production-possibility frontier – but rather is inside that frontier. As demand changes, what we observe inside that frontier need not trace out well-behaved concave loci.

Some Marxians, bemused by notions of ‘unequal exchange’, think that the above phenomena are Pareto inefficient: the system and the world get less product than is producible in the zero-interest-rate golden-rule state. Yes, of course. But that is unavoidable in any scenario where the economic system could only obtain the produced inputs needed for Schumpeter's golden-rule zero-interest rate utopia by doing current ‘waiting’ and sacrificing of current consumptions. The curse of the poor societies is their poverty – even when *intertemporal* Pareto efficiency is always obtaining. Again, all this is as true of *discrete* as of *marginalist* technologies.

Sraffian refutation of Marxism

On the basis of this elaborate description of findings about Sraffian time-phased systems, one can apply the results to appraise and correct (1) neoclassical economics, (2) Marxian economics, (3) Ricardian and classical economics. Its thrust on neoclassical economics was sampled in the previous section.

20. Sraffian economics, as earlier passages make clear, devastatingly repudiates that central part of Marx's economics, Capital, Volume I (1867) which proposed a new paradigm involving an equal ‘rate of surplus value’ by industries or departments. Sraffa and Darwinian arbitrage require an *equal* ‘rate of profit’ by industries or departments. Under exploitative capitalism, Marx misidentifies what is out there to be observed in the competitive market, for the reason that Marx has the capitalists garnering too little in the industries using much capital goods and garnering too much in those using relatively little capital goods. It is a gratuitous error, and a sterile one, since Marx's paradigm does not help predict the laws of motion of competitive capitalism, or help understand the average magnitude of the profit level around which Marx's errors spread. Equal profit rates are not a capitalist shibboleth; dual to the primal variables of the von Neumann technology is an equalized profit rate.

Ian Steedman, a scholar sympathetic to the *Weltanschauung* of Marx and of economic reform, has documented the Sraffian rejection of 1867 Marxian rate of surplus value in his book, *Marx After Sraffa* (1981). So here it need only be said: in the end, from the posthumous publication of *Capital*, volume III (1894), one can recognize that the ‘transformation’ from Marx's 1867 [marked-up] ‘values’ to bourgeois ‘prices’ involves abandoning the 1867 relations and returning to the pre-Marx and post-Marx cost-profit relations of the Sraffian model.

One cannot quite leave it at that. Karl Marx does deserve a statue in the Sraffian Hall of Fame. Not, of course, for his sterile detour into the rate of surplus value. As documented by Samuelson in the 1974 Lloyd Metzler *Festschrift*, Marx was the first scholar after Quesnay to grapple explicitly with input/output capital models. Implicitly, Marx was the first to use (a_{ij}) coefficients. Moreover, in his Tableaus of Steady Reproduction and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Expanded Reproduction, *Capital*, volume II (1885), Karl Marx was the first to present coherent row-and-column arrays of steady-state equilibrium. This is Marx's imperishable contribution to analytical economics, and it is impervious to the deflating of hyperbole concerning Marx as allegedly a great *Mathematical* economist.

1926 reconsidered from 1960

There were two main parts to Sraffa's celebrated 1926 paper. The first part, which today we realize was by far the more important one, dealt with the phenomena of *increasing returns to scale*. So long as demand is insufficiently large to take a firm beyond such an initial phase of increasing returns or decreasing costs, the firm cannot find a maximum-profit equilibrium while still remaining a perfect competitor. Cournot knew this in 1838, and it was not new doctrine even then. The turn of the century literature on trusts and industrial organization, in America and Germany, was never in doubt on this. By 1926 the whole issue would have been old hat save for Alfred Marshall's tergiversations on the compatibility of *increasing* (internal, statical and reversible) *returns* and perfection of competition. Marshall, at the time of his death in 1924, was at the height of his prestige, with the greatest capacity for good and for potential confusion. Therefore, it was important for Piero Sraffa to restate elegantly that real-life firms, with localized and segmented markets, were in imperfect-competition equilibrium with falling marginal production costs that were offset by selling and transport costs and by price declines inducible by expansion of their own outputs and sales. Independently, E.H. Chamberlin and J.M. Clark were saying much the same thing at the time in America. But the ruling establishment of Cambridge, understandably, could learn something best from one of its own publishing in the *Economic Journal*. The 1960 book has no relation to this part of Sraffa's early work.

Within the mid-1920s, Sraffa's other thesis generated a disproportionate amount of interest. Not only was the familiar downward-sloping Marshallian supply curve to be ruled out as incompatible with perfection of competition; the young Sraffa was newly arguing that upward-sloping supply curves were also of vacuous importance for Marshallian partial equilibrium. All that Sraffa left his reader, then, was a horizontal, *constant-cost* competitive supply curve.

This is plain wrong. Sraffa's 1960 book demonstrates that, when primary factors other than a single homogeneous labour exist, rightward shifting Marshallian and Walrasian demand curves will generally trace *rising* price intersections on the relevant supply curves. Joan Robinson's famous 1941 *Economic Journal* article on rising supply price was the first East Anglian recognition of the formal comparative statics of general equilibrium. I doubt that she or Piero ever noticed the incompatibility with 1926 Sraffa; or the incompatibility of Heckscher–Ohlin and Stolper–Samuelson in the foreign-trade literature with Sraffa's thesis of *constant* costs and implied *linear* production-possibility frontiers. The pre- and post-Ricardian classical literature is full of examples of wine grown on special vineyard lands: for generations students have followed Viner's 1931 example of calling this ‘the pure Ricardian case’ – while the modern teachers of the young know this as the Jones–Samuelson–Haberler specific-factors model.

Students of rhetoric should be interested to analyse the elements of style that enable the erudite author of a faulty thesis to persuade himself and several generations of thinkers of its truth and importance. Because Sraffa wrote so little, and wrote so rarely on the mainstream topics of contemporary scholarship, his skills as a writer have perhaps been insufficiently noticed.

Knowing what we now know – that it is a mistake to believe that constant-cost cases exhaust the categories of admissible competitive price – we are in a position to study the young Sraffa's extra-scientific motivations. Why does a sophisticated intelligence make *this* mistake? Just as inside a fat man is a thin man trying to get out, so outside the Sraffa of the post-World War I heyday of Walras–Marshall neoclassicism, there was

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
already in 1925 an atavistic classical economist trying to get back in. This is why one can say that, from before 1925 to after 1960, there is a discernible consistency in Piero Sraffa's thought and ideology. An objective reader will want to be alert to this tendency.

1960 verdicts on 1817 Ricardianism

Sraffa's models, we have by now seen, tellingly reject the following Ricardian stereotypes:

1. (*a*) Prices are determinable by the labour theory of value.
2. (*b*) Land and rent can be ignored by concentrating on external-marginal land with zero rent (or on internal margins in every positive-rent acre).
3. (*c*) The complication for pricing of time and the interest rate can somehow be avoided or ameliorated by defining an intermediate standard commodity or workbasket, which is less time consuming and interest-rate-inflated than the most time consuming goods (old wine and tall trees) and is more time consuming and interest-rate inflated than the most directly produced goods (shrimp picked up by labour on the seashore).
4. (*d*) One can correctly understand the distribution of income among workers, landowners, and capitalists independently of the complications of demand theory (consumers' demand functions, marginal utility, revealed preferences, etc.).
5. (*e*) It is superficial to base goods' and factors' pricings on mere supply and demand.
6. (*f*) Adam Smith committed some grave errors in decomposing price and national income eclectically into wage-plus-rent-plus-profit components, and in enunciating his notion of 'labour command'.

Enough has been given of current misunderstandings related to Ricardo for the present exegesis. What does a close rereading of Smith and Ricardo reveal under the light of the post-1960 analysis?
Most of both scholars' actual inferences about the real world are compatible with the post-1960 findings. Smith's scorecard is certainly not inferior to Ricardo's, even after we make allowances for the latter's tendency to proclaim as being universal what is only likely.
Ricardo began to write on microeconomics because he thought he discerned basic logical flaws in Smith's system. Just as Steedman showed, in the Ronald Meek *Festschrift*, that Marx's criticisms of Ricardo could not stand up to modern examination, Robert L. Bishop (1985) has shown that Ricardo's criticisms of Smith similarly cannot stand up.
Both Smith and Ricardo tried to compare, by one scalar parameter, the diverse price vector of two times and places: $[P_i/P_1, P_i/W_1, W_m/W_1; r]$ for China and Scotland, or for the Englands of 1780 and 1688. *We* know that just cannot be done: then, now, or ever. Smith's 'labour command' notions merely proceeded from the prosaic observation that people always have about two-thirds of a 24-hour day available to them: $\Sigma P_j C_j / W$ is the hours people have to work for what they consume; because per capita C's tend to rise only with longterm productivity, the above ratio tends to decline only slowly as people enjoy more leisure.
Ricardo never supplanted this imperfect measure by a better one, for all his palaver about absolute standards of value. Worse: gratuitously he attributed to Adam Smith *unwarranted* deviations from the labour theory of value by virtue of Smith's labour command passages. Actually, there is no valid connection. Smith only introduces deviations between what $[P_i/P_j, P_i/W]$ are and what they would be under a labour theory of value when these deviations are warranted by (1) scarcities of needed lands and natural resources, and (2) timephasing of production that involves produced goods as inputs and which takes place when the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
competitive market displays positive interest and profit rates.
For all of Ricardo's scolding of Smith as a recanter from the labour theory of value, Ricardo up to his dying month admitted the cogency of the point (2) – that goods of the same content but involving manifestly different time involvements and relative profits would have prices that systematically fail to be proportional to their embodied labour contents. And on point (1), as we have already seen, Smith was right; and Ricardo was wrong in his belief that he could get rid of the complication of rent by use of external (or internal) margins. George Stigler's Pickwickian 1958 defence of 'Ricardo's 93% labour theory of value' on the positivistic grounds that his labour theory of value allegedly averages out in practice to errors of only about 7 per cent makes one wonder what scientists would think of a defence of Lamarck as being 49.99% right; or of the stone-fire-air-and water paradigm as offering a theory of matter that is at least 0.001% as accurate as Mendelev's 93-element periodic table. For some comparisons a 7% error becomes a 70% or a 700% error.
Ricardo does seem to make two major advances on Smith. Paradoxically for Sraffa's hero, *Ricardo made a giant step beyond* Smith toward marginalism. His ubiquitous numerical examples presuppose almost a continuum of alternative doses of labour-and-produced-goods applied to the same acre(s) of land. Inside Ricardo there is a von Thünen and a J.B. Clark striving to be born!
A second Ricardian advance is not so important or clear-cut. You must read Smith closely to perceive his understanding that the rent of inelastically supplied land is price determined rather than price-determining. In Ricardo the point is made crystal clear; in the chapter on land in Sraffa (1960), you must read with sophistication to perceive the point.
Also, Ricardo stresses, indeed overstresses, the point that the profit rate would not have to decline – as saving brings into existence new capital stocks and enlarged populations – if new lands were ever available in unlimited and redundant supply. A good point, even if readers of his expositions might be forgiven for misunderstanding him to imply that the only reason for a drop in interest and profit rates is a rise in rent: actually, as Smith knew, a persistent excess in the growth rate of capitals relative to population must in many realistic technologies raise the real wage and lower profit rates *even in the presence of superabundant lands*.
The post-1960 Sraffian analysts who grade Ricardo's blue books must often mark down his submissions. It is time therefore to study how Ricardo's 1951 editor dealt with these issues.

1960 light on the 1951 editor of Ricardo

The history of humane letters involves only history. Samuel Johnson's mistakes may be more interesting than his correct observations. To the antiquarian, antiquarianism is all there is to the history of the humanities.
The history of scientific thought is a two-fold matter. We are interested in Newton's alchemy and biblical prophecy because we are interested in Newton the man and scientist. At the same time his stepsister's theology is likely to elicit a yawn from even the most besotted antiquarian. How Newton discerned that a homogeneous sphere of non-zero radius attracts as if all its mass were at its center point, that is part of the history of cumulative science. Say that this attitude involves an element of Whig history if you will, but remember that working scientists have some contempt for those historians and philosophers of science who regard efforts in the past that failed as being on a par with those that succeeded, success being measurable by latest-day scientific juries who want to utilize hindsight and example post knowledge.
Economics is in between belles lettres and cold science. Serious economists below the age of 60 will judge Sraffa's edition of Ricardo both for its antiquarian and its scientific interests and insights. How then will they judge it?
From an antiquarian view the work is a jewel of perfection. Reviewers' enthusiasm has been unbounded. By luck and Sraffa's energetic skills, virtually every scrap written by David Ricardo has been made available to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the interested reader. This is a boon to scholars who lack the slightest interest in the history of thought for its own sake: Baconian scientific observation of Ricardo's economics has now been made possible by Sraffa's labours.

Editorial emendations have also been done in the new edition with skill and brevity. You might almost say that the editor has for the most part stayed chastely out of the act, letting David and his friends speak their pieces without an accompaniment of Greek Chorus expressions of approval or disapproval.

From the scientific viewpoint, and now a minority viewpoint is being expressed here, there is something anticlimactic about the great Sraffa edition of Ricardo. It is not just that we see, as if imprisoned in amber, the backward and forward gropings of a scholar who from his 1814 entrance into microeconomics until his death in 1823 makes almost no progress in resolving his self-created ambiguities and problematics. Somehow one had hoped that the whole picture would be a prettier scientific picture, so that the editor's Herculean framings would be for a more worthwhile object.

There is, however, no point in lamenting that Ricardo was only what he was. It is the ‘road not taken’ by the editor that occasions a twinge of regret. From the scientist's rather than the antiquarians’ viewpoint, we appreciate from an editor and commentator what Jacob Viner gave economists in his magnificent 1937 *Studies in the Theory of International Trade* and what Eli Heckscher supplied in his *Mercantilism*. It is what Clifford Truesdell's lengthy introductions to the collected works of Euler provide, and what Abraham Pais succeeds in bringing off in his 1984 survey of the scientific physics of Albert Einstein. Admittedly old Edwin Cannan carried to excess his patronizing reviews of past economic giants, not only faulting them for their sins in failing to believe what Cannan believed in 1928 but also managing to convict them of the crime of not being so smart as himself. Surely, there is a golden mean somewhere between Cannan's dominating the act and Sraffa's avoiding getting into it?

Fortunately, in his Introduction to Ricardo's *Principles* (written late in the day, with the help of Maurice Dobb), Piero Sraffa does let himself go a little bit. Thus, he conjectures that Ricardo, in a lost 1814 manuscript or letter or conversation, may have worked out a model in which the profit rate is determined within agriculture, as a ratio of (so to speak) corn to corn; and, Sraffa all but says, in such a model distribution theory is successfully emancipated from value theory. Unlike Viner and Cannan, who can be very hard indeed on the guinea pigs they are judging, one reads Sraffa in his Introduction as being quite indulgent of Ricardo. When he quotes Ricardo as purporting to get rid of the complication of rent by concentrating on the external margin, Sraffa never seems tempted to add that this is a *non sequitur*. When Ricardo tries to overdifferentiate his product from Smith's, Sraffa never writes: ‘Of course, when Smith made the emergence of positive interest cause a divergence of price from labour contents, he was doing what Ricardo often admits must be done – namely, formulating a two-factor rather than a one-factor model of pricing.’ The critique of mainstream twentieth century that Sraffa never lived to articulate was evidently festering inside the editor of Ricardo during the 1930–1951 period and serving to soften his critical judgments.

Salutations

Did any scholar have so great an impact on economic science as Piero Sraffa did in so few writings? One doubts it. And there cannot be many scholars in any field whose greatest works were published exclusively in their second half century of life.

Piero Sraffa was much respected and much loved. With each passing year, economists perceive new grounds for admiring his genius.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Sraffian economics (new developments)

Bibliography

Bishop, R. 1985. Competitive value when only labor is scarce. *Quarterly Journal of Economics* 100, November, 1257–92.

Bortkiewicz, L. von. 1907. On the correction of Marx's fundamental theoretical construction in the third volume of *Capital*. Trans. in *Karl Marx and the Close of his System*, ed. P.M. Sweezy, New York: August M. Kelley, 1949.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37, 407–36.

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 36 (2), 291–325.

Manara, C.F. 1979. Sraffa's model for the joint production of commodities by means of commodities. In *Essays on the Theory of Joint Production*, ed. L.L. Pasinetti, New York: Columbia University Press.

Marx, K. 1867, 1885, 1894. *Capital*, Vols. I, II, III. English translations in C.H. Kerr (ed.), Chicago: Chicago University Press, or other editions.

Metcalfe, J.S. and Steedman, I. 1972. Reswitching and primary input use. *Economic Journal* 82, March, 140–57.

Miyao, T. 1977. A generalization of Sraffa's standard commodity and its complete characterization. *International Economic Review* 18, February, 151–62.

Neumann, J. von. 1945. A model of general economic equilibrium. In *Review of Economic Studies* 13, 1–9. Trans. of an article in *Ergebnisse eines mathematischen Kolloquiums*, Vol. 8, ed. K. Menger, Leipzig, 1937.

Pasinetti, L.L. 1977. *Lectures on the Theory of Production*. New York: Columbia University Press.

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Samuelson, P.A. 1975. Marx as mathematical economist: steady-state and exponential growth equilibrium. In *Trade, Stability, and Macroeconomic, Essays in Honor of Lloyd A. Metzler*, ed. G. Horwich and P. Samuelson, New York: Academic Press, 269–307. Also ch. 225 in *The Collected Scientific Papers of Paul A. Samuelson*, Vol. III, Cambridge, Mass.: MIT Press, 1972.

Samuelson, P.A. 1975. Trade pattern reversals in time-phased Ricardian systems and intertemporal

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

efficiency. *Journal of International Economics* 5, November, 309–63. Also ch. 251 in *The Collected Scientific Papers of Paul A. Samuelson*, Vol. IV, Cambridge, Mass.: MIT Press, 1977.

Samuelson, P.A. 1983. Thünen at two hundred. *Journal of Economic Literature* 21, December, 1468–88.

Schefold, B. 1971. *Theorie der Kuppelproduktion (Mr. Sraffa on Joint Production)*. Privately printed, Basel.

Seton, F. 1957. The transformation problem. *Review of Economic Studies* 25, 149–60.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36, December, 535–50.

Sraffa, P., ed. 1951–73. *The Works and Correspondence of David Ricardo*. Cambridge: Cambridge University Press, for the Royal Economic Society; with editorial collaboration by M.H. Dobb.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx After Sraffa*. London: New Left Books. Reprinted, London: Verso, 1981.

Steedman, I. 1982. Marx on Ricardo. In *Classical and Marxian Political Economy: Essays in Honour of Ronald L. Meek*, ed. I. Bradley and M. Howard, London: Macmillan.

Stigler, G.J. 1958. Ricardo and the 93% labor theory of value. *American Economic Review* 48, June, 357–67.

How to cite this article

Samuelson, Paul A. "Sraffian economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000220> doi:10.1057/9780230226203.1598

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

s-S models

John Leahy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The s-S model is the canonical model of infrequent and discrete action. It has been used to explain inertia in durable goods, investment, money demand, and cash management, and to provide microfoundations for price stickiness and the real effects of money. In the model, fixed costs make small adjustment impractical. Agents allow their state to drift in response to shocks until it reaches an adjustment trigger before setting it to a target value. This article reviews the microeconomic comparative statics of the optimal s-S policy, as well as the implications of discrete individual adjustment aggregate dynamics.

Keywords

adjustment costs; adverse selection; Arrow, K. J.; curse of dimensionality; frictions; inventory theory; Markov processes; Marschak, J.; microfoundations; neutrality of money; one-sided rules; s-S models; state dependence; sticky prices; Taylor's theorem; two-sided rules

Article

The s-S model is the canonical model of inaction arising from costs of adjustment. Individuals do not always react to changes in their environment. Consumers rarely buy a new house or car after every fluctuation in their permanent income. Firms often leave prices fixed for months even though information is arriving at a much greater frequency. Fixed costs of adjustment provide a natural explanation of this inertial behaviour. If an agent faces a fixed cost to taking some action and if the loss to non-adjustment is small in the neighbourhood of the optimal choice, then it will pay to leave things be until the benefits of adjustment exceed the costs. Bar-Ilan and Blinder (1996) call this the ‘optimality of usually doing nothing’.
The term s-S derives from inventory theory. In Arrow, Harris and Marschak's (1951) seminal paper, a firm allows its inventory holdings to decline below a level s before placing an order that replenishes inventories to a level S. Subsequently, the term s-S has come to denote an entire class of models of discrete and infrequent adjustment in which the optimal strategy is characterized by a set of triggers and targets. s-S models have been applied to explain inertia in a variety of microeconomic settings, including money demand, cash management, pricing, durable goods, and investment. In macroeconomics, the principal application has been to provide microfoundations for price stickiness and thereby the real effects of money. This is the menu cost model of price stickiness.

Microeconomics: the basic idea

The hallmark of the s-S policy is the combination of inaction and discrete adjustment. The basic idea can be simply illustrated in a static setting. Consider an agent who must choose x to minimize some twice differentiable, concave payoff function $\pi(x)$. The agent is endowed with a value x_0. The wrinkle is that there is a fixed cost k to changing x from its initial value. The optimal policy which balances the costs and benefits of adjustment is illustrated in Figure 1. If x_0 is less than S_L or greater than S_H, the benefit of adjusting to S^* outweighs the fixed cost of adjustment k. If x_0 is between S_L and S_H , inaction is optimal, and consequently $[S_L, S_H]$ is referred to as the range of inaction. The points S_L and S_H are, respectively,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the upper and lower adjustment triggers. S^* is the adjustment target.

Figure 1

An s-S policy

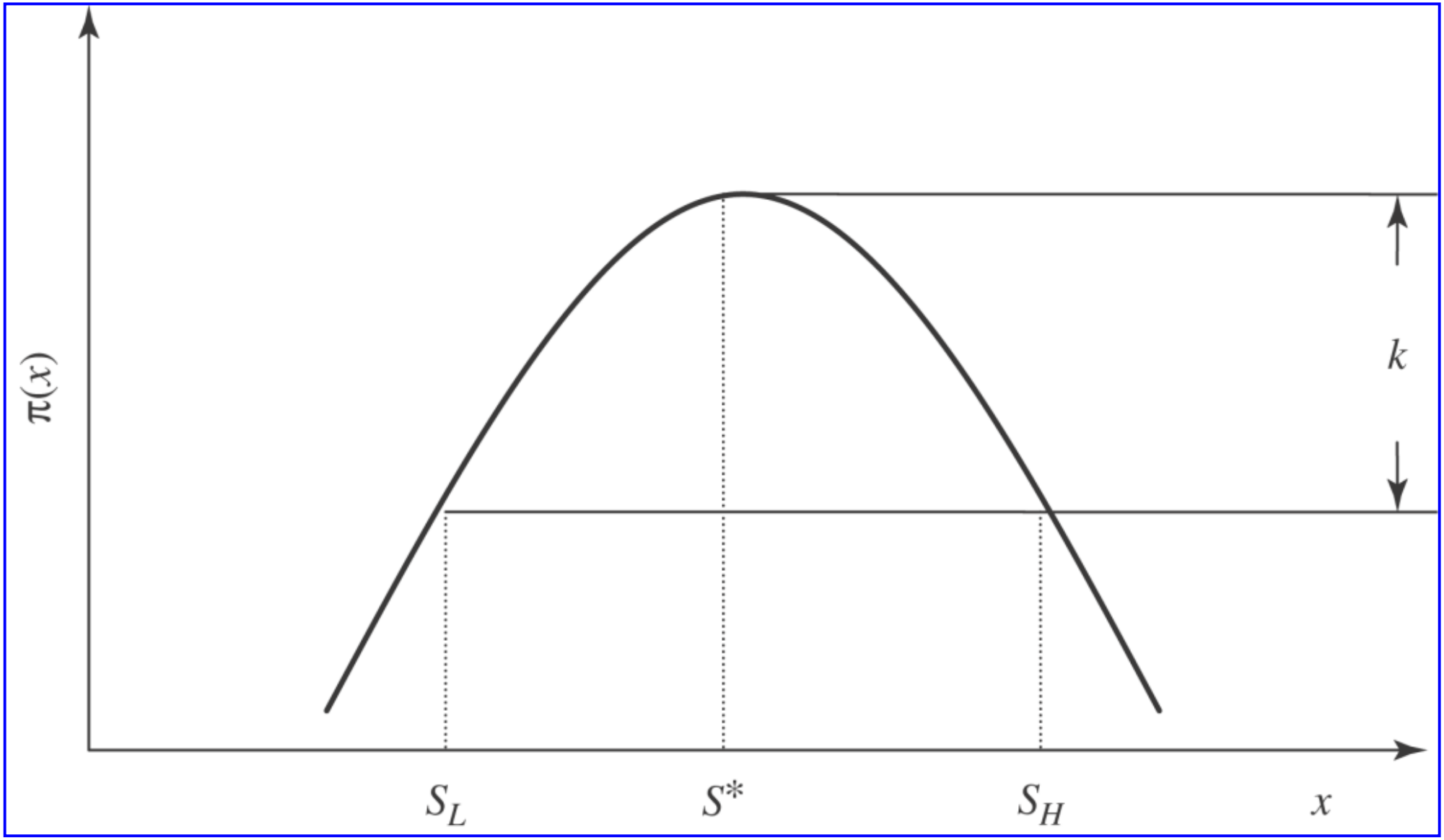

The combination of inaction and discrete adjustment is a direct result of the fixed cost of adjustment k. If instead the cost of adjustment were some twice differentiable convex function $c(x - x_0)$ with $c(0) = c'(0) = 0$, then there would be no inaction. Since the marginal cost of adjustment is zero at x_0, it would always be optimal to move closer to S^*. One of the immediate lessons of Figure 1 is that the range of inaction $[S_L, S_H]$ may be large even if the fixed cost k is small. This is because the payoff function is normally very flat in the neighbourhood of S^*. Formally, if k is small we may apply Taylor's theorem and approximate $\pi(x)$ in the neighbourhood of S^* by a quadratic $\pi(S^*) + \frac{1}{2}\pi''(S^*)(x - S^*)^2$. Recall S^* is the optimal choice so the first-order term is zero. S_L and S_H are the points at which this payoff function is equal to $\pi(S^*) - k$. Solving for S_L and S_H yields

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\left|S_L - S^*\right| = \left|S_H - S^*\right| = \sqrt{\frac{k}{-\frac{1}{2}\pi''(S^*)}}.$$

The size of the range of inaction is increasing in the adjustment cost and decreasing in the concavity of the loss function. Here second order adjustment costs are sufficient to generate a first-order range of inaction (Mankiw, 1985).

One-sided rules

Adding exogenous forces that act on *x* makes the problem dynamic. If the increments to *x* in the absence of adjustment are Markov, then the optimal trigger-target strategy will be stationary. The literature considers two cases. The first is monotonic drift in *x*. This leads to 'one-sided' s-S rules. Agents let *x* drift until it passes a trigger *s* and then reset it to some value *S*. One-sided rules arise naturally in inventory theory (Arrow, Harris and Marschak, 1951), money demand (Baumol, 1952) and pricing under inflation (Sheshinski and Weiss, 1977). Scarf (1959) provides a general proof of the optimality of one-sided s-S policies, requiring only that π is concave.

The canonical model is due to Sheshinski and Weiss. In their model, time is continuous and indexed by *t*. They interpret $x = \ln p_i - \ln p$ as the log difference between a firm's price p_i and a price index *p*. Payoffs therefore depend on the firm's relative price p_i/p. They assume that the price index grows at rate *g* and that the firm discounts future payoffs at a rate ρ. Let *V* denote the profits of a firm that has just paid the fixed cost of adjustment and must choose a new price. The firm's optimization problem becomes:

$$V = \max_{\ln S, T} \frac{\int_0^T \pi(\ln S - gt)\, e^{-\rho t} dt - e^{-\rho T} k}{1 - e^{-\rho T}}.$$

Here ln•*S* is natural logarithm of the target price *S*, and *T* is the time between successive price adjustments. The adjustment trigger $s = Se^{-gT}$. The first order conditions are:

$$\int_0^T \pi'(\ln S - gt)\, e^{-\rho t} dt = 0, \text{ and } \pi(\ln s) = \rho(V - k).$$

The first equation says that the present value of marginal losses over the cycle should be set to zero. If the firm did not discount profits and the loss function were quadratic, this would mean that the trigger *S* and target *s* would be symmetrically placed about the static optimal price. With discounting the firm backloads losses; *S* is closer to the optimum than *s*. The second equation says that, at the trigger, the instantaneous profit is equal to the cost of delaying the next cycle.

Most of the comparative statics of this model are similar to those of the static model outlined above. An increase in the cost of adjustment *k*, an increase in the rate of inflation *g*, or a reduction in the concavity of π causes the target *S* to rise and the trigger *s* to fall. One somewhat surprising result is that an increase in *g* has an ambiguous effect on the frequency of adjustment. For given s-S bands an increase in *g* increases the frequency of adjustment. The widening of the bands, however, works in the opposite direction.

Two-sided rules

If we relax the assumption that the shocks to *x* are monotonic, then the optimal policy will involve two triggers, one above and one below the target. Such two-sided rules have been used to study cash management (Miller and Orr, 1966), pricing (Barro, 1972), durable consumption (Grossman and Laroque, 1990) and investment (Abel and Eberly, 1994).

As an example of a two-sided model, take the static example from the beginning of this section and assume that *x* follows a Brownian motion in the absence of adjustment. In this dynamic model, the agent would not necessarily adjust to the static optimum; the shape of the profit function and the drift of the shock may cause the agent to pick a target on one

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
side or the other. Even so, many of the comparative statics of this example are similar to the static model. The range of inaction tends to widen if the profit function becomes less concave or the adjustment cost rises.

One entirely new effect is that of the variance of the shock to *x* on the adjustment policy. If *x* follows a driftless Brownian motion, then an increase in the variance of *x* leads to wider bands and to more frequent adjustment. The intuition is simple. If the agent keeps the bands fixed, then the increase in the variance of *x* makes adjustment more frequent. The costs of adjustment rise relative the benefits. Wider bands are optimal. If the agent widens the bands in proportion to the standard deviation of the shock, the frequency of adjustment is unchanged but the benefits to adjustment rise. More frequent adjustment is optimal.

The widening of the bands in this case reflects an option value to waiting. s-S adjustment is like exercising an option: the greater is uncertainty, the more this option must be in the money. Dixit (1991) shows that, given this option value, fourth-order adjustment costs lead to a first-order range of inaction.

Macroeconomic models

The work of Blinder (1981) and Caplin (1985) inspired interest in the aggregate implications of s-S policies. Blinder provided several examples that illustrate the difficulty of aggregating s-S policies and the potential complexity of the resulting dynamics. Caplin demonstrated that the s-S model of inventories can explain the commonly observed finding in the inventory literature that production is more volatile than sales, an observation that is at odds with the production smoothing model of inventories (the idea is that the discreteness of orders in the s-S model adds to the variability in demand). In recent years, a large body of research has examined the relationship between microeconomic frictions and aggregate dynamics. Two themes run through this literature. The first is a need to surmount difficult modelling issues brought on by the heterogeneity inherent in discrete adjustment. The second is that aggregate dynamics may look very different from the microeconomic dynamics.

The curse of dimensionality

The main difficulty encountered in constructing aggregate models of with s-S behaviour is that the cross-sectional distribution of agents' deviations from their optimum becomes a state variable. This distribution determines how many agents are near their triggers and hence the amount of adjustment that will take place in the near future. To handle the high dimension of this distribution, current models follow one of three approaches. Some, like Caplin and Spulber (1987), Caplin and Leahy (1991, 1997), Danziger (1999) and Gertler and Leahy (2006) make distributional assumptions that reduce the number of state variables. Others, like Blinder (1981), Bertola and Caballero (1990), and Caballero and Engel (1991; 1999), ignore equilibrium considerations and construct aggregates by integrating across actions of isolated individuals facing possibly correlated shocks. A third group, Dotsey, King and Wolman (1999), Willis (2002), and Caplin and Leahy (2006), solve stochastic general equilibrium models, employing approximations that reduce dimensionality.

The Caplin – Spulber model

The Caplin–Spulber model illustrates a few basic lessons in s-S aggregation. Three log-linear relationships form the backbone of the model: the aggregate price index *p* is the average of individual prices p_i; output *y* equals real balances $m - p$; and a firm's optimal price p^* rises with the price level and with real balances, $p^* - p = \alpha(m - p)$. Caplin and Spulber close the model with three assumptions: *m* is continuous and monotonically increasing; firms follow one-sided s-S pricing policies (when $p_i - p^*$ falls to *s*, it is raised to *S*) and the initial distribution of relative prices is uniform between *s* and *S*.

The surprising result is that money is neutral in this setting in spite of the fact that all prices adjusted infrequently. The reason is simple. Output can only change if the distribution of prices shifts relative to the money supply. As the money supply increases, firms' prices tend to fall relative to p^*. The few firms that hit *s* adjust to *S*. In this way, firms rearrange themselves within the interval (*s*,*S*). Given the uniformity assumption, the distribution of $p_i - p^*$ is unchanged. The price index moves with the money supply because a small number of firms change their prices by a large amount.

The Caplin–Spulber model makes two important points. First, discrete adjustment at the microeconomic level does not necessarily imply discrete adjustment at the macroeconomic level. Heterogeneity in adjustment times tends to smooth macroeconomic dynamics. Second, the aggregate implications of s-S dynamics depend on the evolution of the distribution of idiosyncratic deviations from the optimum. In this case, the distribution does not change, and the s-S model looks exactly like its frictionless counterpart.

The Caplin – Leahy model

Caplin and Leahy (1991) make three simple amendments to the model of Caplin and Spulber: they assume that money is non-monotonic and follows a Brownian motion, that firms

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

follow symmetric, two-sided pricing policies, and that the initial distribution of prices is uniform with a range equal to the distance between the trigger and the target. In this setting, they show that the distribution retains this shape. It rises and falls within the range of inaction like an elevator.

The Caplin–Leahy model provides an example of an economy in which the distribution of idiosyncratic deviations changes over time. The model also illustrates a new phenomenon: the state-dependence of the effects of aggregate shocks. When the distribution of prices is interior to the range of inaction, shocks to the money supply shift the distribution and output changes. When the distribution of prices reaches the edge of the range of inaction, further movements in the money supply cause price adjustment. The distribution of deviations rearranges itself as in the Caplin–Spulber model and money is neutral.

The Bertola – Caballero – Engel model

In a series of papers, Bertola, Caballero and Engel develop a very flexible framework for modelling aggregate and distributional dynamics with idiosyncratic shocks. Caballero and Engel (1991) consider prices and one-sided adjustment, and Bertola and Caballero (1990) consider durables and the two-sided adjustment. In this discussion, I will focus on durables and two-sided adjustment.

The model takes the neoclassical model without adjustment costs to be its benchmark. This neoclassical model would predict that an individual would like to hold a quantity $x_i^*(z_i)$ where z_i is a vector of individual characteristics. They then postulate that individuals follow s-S policies: if $x_i - x_i^* \notin (S_L, S_H)$ then the individual adjusts his holdings so that $x_i - x_i^* = S^*$. They assume that innovations in x_i^* are i.i.d. across agents i and over time, and are characterized by a stochastic matrix P_t, where P_t may depend on the aggregate state of the economy at date t. Since P_t has a finite number of states, the $x_i - x_i^*$ take on a finite number of values; denote these deviations by $\hat{x}_i$. The stochastic matrix P_t and the s-S adjustment policy, induce transitions on the $\hat{x}_i$. Denote these transitions by the stochastic matrix $\hat{P}_t$.

This model provides a simple accounting of aggregate dynamics. Let X_t denote the aggregate holdings of durables and let the column vector f_t denote the cross-sectional density of the $\hat{x}_i$. Then $X_i = \int x_i^* di + f_t' \hat{x}$ and $f_{t+1} = f_t \hat{P}_{t+1}$. In this view, s-S dynamics provide a theory of the error term in the neoclassical model. Deviations from the neoclassical model are associated with fluctuations in the density f.

The model allows one to consider the relative roles of idiosyncratic and common shocks. Shocks that are common across firms tend to shift f as they do in the model of Caplin and Leahy. Idiosyncratic shocks, however, tend to mute the effects of aggregate shocks. If only idiosyncratic shocks are present the error $f_t' \hat{x}$ settles down to an ergodic density of $\hat{P}$. If idiosyncratic shocks dominate, the microeconomic frictions are lost in the aggregate like in the model of Caplin and Spulber.

Caballero (1993) applies this framework to show that an aggregate s-S model may look like a partial adjustment model. A shock causes some agents to adjust and leaves others near the adjustment trigger. The result is that subsequent idiosyncratic shocks lead to further adjustment.

Like Caplin and Spulber and Bertola and Caballero, Golosov and Lucas (2004) emphasize the fact that the firms that adjust in an s-S model are those with the greatest desire to adjust. This differentiates s-S models from models in which the time between price adjustments is fixed such as the Taylor or Calvo models of price adjustment.

Equilibrium interactions

The simplicity of the Bertola–Caballero–Engel model comes from the absence of equilibrium interactions. The question arises whether the introduction of equilibrium interactions would enhance or limit the differences between the s-S model and its frictionless counterpart. As usual, the answer is ‘it depends’.

On the one hand, Thomas (2002) shows that endogenous prices may blunt s-S dynamics. If an unusually large number of people are about to purchase a car, the price of cars should rise, thereby dissuading some agents from making purchases. Endogenous prices movements therefore tend to smooth any deviations of the s-S model from the frictionless neoclassical benchmark. Caplin and Leahy (2004) argue that in a one-sided model the aggregate dynamics become observationally equivalent to a model without adjustment costs.

On the other hand, Ball and Romer (1990) argue that in the presence of strategic complementarities, non-adjustment by some agents can encourage non-adjustment by others. For example, if due to increasing returns the profitability of investment by any single firm is increasing in aggregate investment, then each firm may forgo investment because others have decided to forgo investment. Strategic complementarities may therefore cause the range of inaction to widen, allowing for greater differences from the frictionless benchmark.

s-S models remain an active area of research. At the microeconomic level, recent work has shown how s-S frictions may trap information leading to interesting boom–bust cycles (Caplin and Leahy, 2006) and how adverse selection may amplify Ss frictions (House and Leahy, 2004). At the macroeconomic level, work progresses on the importance of fixed costs in explaining price inertia (Golosov and Lucas, 2003; Gertler and Leahy, 2006; Midrigan, 2006).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- adjustment costs
- Arrow, Kenneth Joseph
- microfoundations
- monetary business cycle models (sticky prices and wages)
- options
- Phillips curve
- stochastic optimal control

Bibliography

Abel, A. and Eberly, E. 1994. A unified model of investment under uncertainty. *American Economic Review* 84, 1369–84.

Arrow, K., Harris, T. and Marschak, J. 1951. Optimal inventory policy. *Econometrica* 19, 205–72.

Ball, L. and Romer, D. 1990. Real rigidities and the non-neutrality of money. *Review of Economic Studies* 57, 183–203.

Baumol, W. 1952. The transactions demand for cash: an inventory theoretic approach. *Quarterly Journal of Economics* 66, 545–56.

Bar-Ilan, A. and Blinder, A. 1996. Consumer durables: evidence on the optimality of usually doing nothing. *Journal of Money, Credit and Banking* 24, 258–72.

Barro, R. 1972. A theory of monopolistic price adjustment. *Review of Economic Studies* 39, 17–26.

Bertola, G. and Caballero, R. 1990. Kinked adjustment costs and aggregate dynamics. In *NBER Macroeconomics Annual 1990*, ed. O. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Blinder, A. 1981. Retail inventory behavior and business fluctuations. *Brookings Papers on Economic Activity* 1981(2), 443–505.

Caballero, R. 1993. Durable goods: an explanation for their slow adjustment. *Journal of Political Economy* 101, 351–84.

Caballero, R. and Engel, E. 1991. Dynamic (S,•s) economies. *Econometrica* 59, 1659–86.

Caballero, R. and Engel, E. 1999. Explaining investment dynamics in US manufacturing: a generalized (S,•s) approach. *Econometrica* 67, 741–82.

Caplin, A. 1985. The variability of aggregate demand with (S,•s) inventory policies. *Econometrica* 53, 1395–410.

Caplin, A. and Spulber, D. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102, 703–25.

Caplin, A. and Leahy, J. 1991. State-dependent pricing and the dynamics of money and output. *Quarterly Journal of Economics* 106, 683–708.

Caplin, A. and Leahy, J. 1994. Business as usual, market crashes and wisdom after the fact. *American Economic Review* 84, 548–65.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Caplin, A. and Leahy, J. 1997. Aggregation and optimization with state-dependent pricing. *Econometrica* 65, 601–23.

Caplin, A. and Leahy, J. 2004. On the relationship between representative agent models and (S,•s) models. In *Productivity, East Asia Seminar on Economics 13*, ed. T. Ito and A. Rose. Chicago: University of Chicago Press.

Caplin, A. and Leahy, J. 2006. Equilibrium in a durable goods market with lumpy adjustment. *Journal of Economic Theory* 128, 187–213.

Danziger, L. 1999. A dynamic economy with costly price adjustments. *American Economic Review* 89, 878–901.

Dixit, A. 1991. Analytic approximations in models of hysteresis. *Review of Economic Studies* 58, 141–52.

Dotsey, M., King, R. and Wolman, A. 1999. State dependent pricing and the general equilibrium dynamics of money and output. *Quarterly Journal of Economics* 104, 655–90.

Gertler, M. and Leahy, J. 2006. A Phillips curve with an Ss foundation. Working Paper No. 11971. Cambridge, MA: NBER.

Grossman, S. and Laroque, G. 1990. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. *Econometrica* 58, 25–51.

Golosov, M. and Lucas, R. 2003. Menu costs and Phillips curves. Working Paper No. 10187. Cambridge, MA: NBER.

House, C. and Leahy, J. 2004. An Ss model with adverse selection. *Journal of Political Economy* 112, 581–614.

Midrigan, V. 2006. Menu costs, multiproduct firms and aggregate fluctuations. Working paper, Ohio State University.

Mankiw, N. 1985. Small menu costs and large business cycles: a macroeconomic model of monopoly. *Quarterly Journal of Economics* 100, 529–39.

Miller, M. and Orr, D. 1966. A model of demand for money by firms. *Quarterly Journal of Economics* 80, 413–35.

Scarf, H. 1959. The optimality of (S,•s) policies in the dynamic inventory problem. In *Mathematical Methods in Social Sciences*, ed. K. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

Sheshinski, E. and Weiss, Y. 1977. Inflation and the cost of price adjustment. *Review of Economic Studies* 44, 287–303.

Thomas, J. 2002. Is lumpy investment relevant for the business cycle? *Journal of Political Economy* 110, 508–34.

Willis, J. 2002. General equilibrium of a monetary model with state-dependent pricing. Working paper, Federal Reserve Bank of Kansas City.

How to cite this article

Leahy, John. "s-S models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.dictionaryofeconomics.com/article?id=pde2008_S000473> doi:10.1057/9780230226203.1600

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stability and growth pact

Thierry Warin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Stability and Growth Pact was designed in 1997 and implemented with the inception of the euro in 1999. An innovative tool in essence, it provides, first, a practical definition of the concept of fiscal sustainability by imposing a ceiling of three per cent and 60 per cent respectively on the budget deficit and public debt. Second, it offers guidelines for governments' public finances. Third, it offers a way to coordinate national public finances to achieve an optimal fiscal–monetary policy mix within the Eurozone. Even before some countries breached the Pact, the economic literature argued about its rationales.

Keywords

bonds; budget deficits; business cycles; default risk premia; Economic and Monetary Union (EMU); European Central Bank; Eurozone; fiscal rules vs. discretion; fiscal–monetary policy mix; free-rider problem; Medium Term budgetary Objective (EU); monetization; moral hazard; optimal fiscal policy; political budget cycles; public debt; public finance; Stability and Growth Pact; stabilization; Treaty of Amsterdam (EU); Treaty of Maastricht (EU); uncovered interest parity

Article

Proposed by Germany in 1995, backed by France, and created by the Treaty of Amsterdam in 1997, the Stability and Growth Pact (SGP) is the European Union's (EU) answer to concerns about fiscal unsustainability. It consisted initially of three regulations: (*a*) on the strengthening of the surveillance and coordination of economic policies (Council of the European Union, 1997b), (*b*) on speeding up and clarifying the implementation of the excessive deficit procedure (Council of the European Union, 1997c, and (*c*) on the SGP (Council of the European Union, 1997a). The SGP extended fiscal discipline into the Economic and Monetary Union (EMU) after 1 January 1999 in the manner foreseen by the Treaty of Maastricht (1992) for the convergence period 1993–8: budget deficit and public debt must be,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
respectively, no more than three per cent of GDP and 60 per cent of GDP.
When, in May 1998 the European Union was deciding which of the EU–15 countries would enter into the EMU in light of the Treaty of Maastricht criteria, some countries, including Germany, would have failed to qualify under the debt criterion. Since then, the debt criterion has been interpreted in terms of trend rather than level in the Treaty of Maastricht, and as a consequence in the Treaty of Amsterdam. The nature of the SGP thus changed within a year of its ratification and a year before its implementation. *De facto* the rule was then to abide by the deficit criterion, leaving the debt criterion to be interpreted in trend. The change in the interpretation of the Treaty of Maastricht resulted in a change in the interpretation of the Treaty of Amsterdam, weakening its original double objective.
What happens when a country does not abide by the three per cent rule? If it is a first violation, the country will have to make a non-interest-bearing deposit with the European Commission. The amount of this deposit comprises a fixed component equal to 0.2 per cent of GDP, and a variable component linked to the size of the deficit. Each following year the Council may decide to intensify the sanctions by requiring an additional deposit, though the annual amount of deposits may not exceed the upper limit of 0.5 per cent of GDP. A deposit is converted into a fine if, in the view of the Council, the excessive deficit has not been corrected after two years. After three consecutive years of violation, the country will see its three deposits become a fine. While the three per cent limit might seem tight, the probability that a country will fail to abide by the Pact for three years in a row, and thus be fined, was originally perceived as low. However, revised numbers from Eurostat indicate that Greece has always been above the three per cent deficit ceiling since its entry into the EMU in 2001. Additionally, Portugal's deficit was greater than three per cent in 2001, 2004 and 2005, Germany's deficits exceeded three per cent from 2002 to 2005, France's deficits exceeded three per cent from 2002 to 2004, and Italy, the UK and the Netherlands breached the three per cent rule in 2004. Further, the deposit rules were not enforced. Not only did Portugal not have to make a deposit, but France and Germany were not fined. The credibility of the SGP was dramatically weakened for the second time.
On 20 March 2005 the European Union decided to try to improve the credibility of the Pact: the Council adopted a report entitled ‘Improving the Implementation of the Stability and Growth Pact’ (Council of the European Union, 2005). The report was endorsed by the European Council in its conclusions of 22 March, and is now an integral part of the SGP. On 27 June 2005 two additional regulations amended Regulations 1466/97 and 1467/97 (Council of the European Union, 1997b; 1997c). The European Council unanimously agreed to introduce some flexibility into the SGP, creating a *de facto* SGP II. This flexibility was particularly introduced via the concept of ‘relevant factors’, which are country-specific. Examples of relevant factors are: (*a*) budgetary efforts towards increasing or maintaining at a high level financial contributions to foster international solidarity and to achieve European policy goals (notably the unification of Europe if this has a detrimental effect on the growth and fiscal burden of a Member State), (*b*) structural reforms (for example, pensions, social security), (*c*) policies supporting R&D, and (*d*) medium-term budgetary efforts (consolidating during good economic times, a reduction in debt levels, and an increase in public investment). The European Council is the final judge of the relevance of a given factor.
In the meantime, the European Union introduced a ‘code of conduct’ to counterbalance the new flexibility that had been introduced. This code of conduct established a country-specific Medium Term budgetary Objective (MTO) that serves as the actual deficit target around which some flexibility is allowed so long as the country abides by the three per cent reference value. To better understand the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
effects of these amendments, the background economics literature is reviewed, as well as the institutional design of the original SGP and its amended version.

The rationales for a supra-national fiscal rule

One of the major goals of the SGP was to make fiscal discipline a permanent feature of the EMU. Safeguarding sound government finances was regarded as a means to strengthen the conditions for price stability in the Eurozone. However, it was also recognized that the loss by individual countries of the exchange rate instrument in the EMU must be offset by automatic fiscal stabilizers at the national level to help economies adjust to asymmetric shocks. The three per cent deficit limit was calculated under the assumption that the long run average nominal GDP growth rate is five per cent, whereas the long-run inflation rate is two per cent.

Beetsma (2001) provided a summary of the different arguments in favour of a fiscal rule. Arguments in support of a Europe-wide fiscal rule are of three types: benefits to domestic governments; benefits to other governments; and collective benefits.

Benefits to domestic governments

The main benefit to domestic governments, namely, public finance sustainability, has been studied by several researchers, including Amador (2000), Ballabriga and Martinez-Mongay (2005), Bohn (1995), Mongelli (1999), Nielsen (1992), and Perotti, Strauch and von Hagen (1998). The SGP aims at ensuring the sustainability of EU public finances, and hence is supposed to prevent governments hampering growth through unsustainable fiscal policies. For illustration purposes, it should be noticed that the primary balance as a percentage of GDP is close to zero or even positive (a surplus) for most of the euro area members. Thus, what pushes countries like France, and Germany above the three per cent deficit ceiling seems to be, primarily, interest payments on debt.

Does Europe really need a fiscal rule to prevent unsustainable domestic public finances? If the answer is ‘yes’, it is because financial markets do not work properly. If government bond yields include risk premia, increasing indebtedness may cause bond yields to rise, thus raising the cost of borrowing and imposing discipline on governments. Market discipline of this kind may be especially relevant and important in the EMU, in which governments of the Member States can issue debt but do not have the possibility of monetizing and inflating away excessive debt. Spreads between European bonds have narrowed considerably since 1991 for the Eurozone members. Bernoth, von Hagen and Schuknecht (2004) explain that, for Deutschmark/euro denominated bonds, EMU membership reduces the linear effect of debt on default risk premia. Accordingly, EMU members enjoy a lower risk premium than before, but this benefit declines with the size of public debt compared with Germany's. This is consistent with the view that markets anticipate fiscal support for EMU countries in financial distress unless these countries had previously been highly undisciplined. Thus, the disciplinary function of credit markets still exists. If the domestic benefit provided by the SGP to governments is not striking, does the SGP discipline governments in their relationships with others, or ease the coordination of fiscal policies between governments?

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Benefits to other governments

A first issue is the likelihood of free-riding behaviours: the fear that governments would run higher deficits in a monetary union. Indeed, under the uncovered interest parity assumption, the understanding was that countries would run deficits that would be financed partly by the Eurozone through an overall rise in the Eurozone bonds interest rate. Undeniably, a country not belonging to a monetary union and running a high deficit will have to face a rise in its domestic interest rate. In a monetary union and integrated capital market, a country running a high deficit will face a much lower rise in its interest rate since it is now equalized at the European level. This may create the incentive for some countries to free-ride on others. If every country behaves this way, on the one hand the overall interest rate rises, and on the other hand the Eurozone faces a greater risk of public finance unsustainability. In the extreme scenario, Eichengreen and Wyplosz (1998) argue that the financial turmoil triggered by a default on the debt of any member country would have significant cross-border effects. In this context, focusing his discussion of free-riding and the SGP on the effects of centralized monetary policy combined with decentralized fiscal policy, Uhlig (2002) regards the SGP as necessary in preventing free-riding in the form of excessively high deficits.

The second issue is moral hazard, which differs from free-riding to the extent that it is ‘post-contractual opportunism’. In other words, once countries belong to the EMU, countries’ loss functions change: without the SGP, governments could weigh more the use of fiscal policies to increase the likelihood of a re-election rather than keeping their public finances in line with the Maastricht guidelines. Dixit (2001) and Dixit and Lambertini (2001) demonstrate that fiscal discretion leads to equilibrium levels of output and inflation very different from Pareto-optimal choices. The SGP should thus prevent countries from changing their attitudes once within the Eurozone.

Collective benefits

Collective benefits of the SGP are at least threefold: the coordination of domestic fiscal policies; the policy-mix argument; and the reinforcement of the ECB's credibility. First, there is the question of fiscal coordination among member countries. A lack of coordination could lead to asymmetric economic shocks on both the aggregate demand and aggregate supply in every country due to large differences in fiscal policies. The coordination of fiscal policies is intended to eliminate those large differences in fiscal policies across countries, and thus implicitly create a Europe-like fiscal policy. The coordination argument is different from the policy-mix argument in the sense that it addresses only the coordination of fiscal policies, whereas the policy-mix argument addresses the question of the coordination of the European monetary policy to a European-like fiscal policy.

Second, Beetsma (2001) and Issing (2002) analyse the policy-mix argument and claim that the advent of a central monetary authority was important in establishing the correct mix of fiscal and monetary policy within the Eurozone. Article 99(3) of the Treaty establishing the European Community stipulates: ‘In order to ensure closer coordination of economic policies and sustained convergence of the economic performances of the Member States, the Council shall, on the basis of reports submitted by the Commission, monitor economic developments in each of the Member States’ (Treaty establishing the European Community, 2002). Beetsma and Uhlig (1999) build a model of centralized monetary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
policymaking and decentralized fiscal policymaking and find that a monetary union combined with an appropriately designed fiscal rule will be strictly preferred to fiscal autonomy, as there are benefits to coordination of a Europe-wide monetary policy with a Europe-wide fiscal policy.
The third reason is the maintenance of the credibility of the European Central Bank (ECB) through insuring its primacy as a monetary authority. As noted by Buti and van den Noord (2004, p. 6), the EMU is, '[commonly] seen as a regime of monetary leadership where fiscal policy is to support the central bank in its task to keep inflation in check'. This power is drawn from the following European Council resolution which accompanies the Pact: '[it] is also necessary to ensure that national budgetary policies support stability oriented monetary policies' (Buti and van den Noord, 2004, p. 6). Around the time that the Maastricht Treaty was drafted, Beetsma and Bovenberg (1999) showed that the European budgetary situation could undermine the credibility of the future European Central Bank. If a country's fiscal situation becomes unsustainable, other countries might be forced to bail out the insolvent national government. Alternatively, the European Central Bank could be forced to monetize national debts, and thereby create additional inflation in the EU although this would be forbidden in theory by the statutes of the ECB. In this regard, the SGP is a secondary safety device.

Institutional design

Formally, the SGP consists of three elements: a political commitment, a preventive element, and a dissuasive element.

The political commitment

Peer support and peer pressure are an integral part of the Stability and Growth Pact: the Council and the Commission are expected to motivate countries to adhere to the pact, and make public their positions and decisions at all appropriate stages of SGP procedure. The idea is to make the SGP more transparent. Member States may also establish a committee of experts to advise them on the main macroeconomic projections, a notion that has roots in the economic literature (Wyplosz, 2005). With this aim, Council Regulation 1466/97 reinforces the multilateral examination of budget positions and the coordination of economic policies.
The SGP foresees the submission of all Member States to stability and convergence programmes. Stability and convergence programmes must present information on the adjustment path and the expected path of the general government debt ratio, as well as the main assumptions made about expected economic development. New to SGP II and in line with the recommendations of the literature, structural reforms are encouraged by the possibility of taking them into account on the path towards adjustment.

The preventive arm of the SGP

The preventive arm of the Pact was, for the first time, given real substance with the implementation of the Medium Term budgetary Objective (MTO). In the 1997 version of the SGP, the MTO was the same for every country: a close-to-balance or surplus budget. Since 2005 the MTO has been given a new

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
definition and is part of a broader new addition to the SGP: the code of conduct. Member States have to define a specific MTO in cyclically adjusted terms. Thus, cycles are now taken into consideration. As recommended by the literature, country specificities must be taken into account. This new device means that surpluses from periods of economic growth are required to be used for debt and deficit reduction. The goals of the MTO are threefold. The first is to provide a margin with respect to the three per cent of GDP deficit ceiling. This margin is calculated by taking into account the past output volatility and budgetary sensitivity to output fluctuations of each Member State. The second goal is fiscal sustainability, for instance, taking into account the economic and budgetary impact of aging populations. Influenced by the economic literature (Blanchard and Giavazzi, 2004; Buti, Eijffinger and Franco, 2003; Fatas, 2005), the third goal is to take into account the need for public investment and represents the structural side of the SGP. The MTOs are revised every four years or whenever a major reform is implemented.

The Council also has the leeway to issue an ‘early warning’ to Member States before an excessive deficit has occurred. Articles 6(2) and 10(2) of Council Regulation 1466/97 state that

> In the event that the Council identifies significant divergence of the budgetary position from the medium-term budgetary objective, or the adjustment path towards it, it shall, with a view to giving early warning in order to prevent the occurrence of an excessive deficit, address, in accordance with Article 103 (4) a recommendation to the Member State concerned to take the necessary adjustment measures.

The dissuasive element

If a country breaches the three per cent value for three consecutive years, it is considered to be in violation of the SGP. In order to dissuade countries from excessive deficits, Council Regulation 1467/97 establishes the Excessive Deficit Procedure (EDP). When the council decides that an excessive deficit exists, it makes recommendations to the Member State and establishes a deadline of six months (raised from four) for corrective policies to be implemented. If a Member State fails to implement the policies based on the Council's decisions, the Council imposes sanctions (deposits, and then fines), which are levied within ten months of the first report of an excessive deficit. A country cannot avoid the deposits, and ultimately the fine, unless the Council decides to abrogate some or all of the sanctions. Abrogation depends on the significance of the progress made by the participating Member State concerned in correcting the excessive deficit, if the breach has resulted from an unusual event or a major economic decline (that is, an annual decline in real GDP of at least two per cent), or if the country's deficit is due to ‘relevant factors’. Any fines already imposed are not reimbursable. Interest on the deposits lodged with the Commission, and the yield from fines, are distributed among Member States without an excessive deficit, in proportion to their share in the total GNP of eligible Member States.

Assessing the Pact

The economic literature and the Pact

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Before the creation of the SGP, the economic literature reflected on the need of a fiscal rule for Europe. In 1997 this fiscal rule materialized into the SGP. Since then, the literature has addressed criticisms to the scientific justifications of the specific design of the SGP, as well as the proposed alternatives. Most of the justifications covered by the literature are inscribed in the following notions: sustainability of public finances; free-riding and moral hazard; coordination and policy mix; and finally the credibility of the ECB.

As for sustainability, the literature is divided with respect to the actual effects of the SGP. Prior to the SGP, financial markets played an active role in disciplining governments, and this still holds true (Bernoth, von Hagen and Schuknecht, 2004). Moreover, some authors challenge the arbitrariness of the definition of sustainability implicit in the SGP (the 3 per cent–60 per cent rule). On the one hand, since budget composition is different across countries, De Grauwe (2003) argues that countries should be able to choose their own debt target instead of the 60 per cent ceiling, and as a consequence have different deficit targets. On the other hand, Coeure and Pisani-Ferry (2005) call for a better concept of sustainability, including, for example, pension regimes. This is something that has been addressed in the amended version of the SGP through the notion of ‘relevant factors’, one of them being the change in pension expenditures.

As for free-riding, the SGP may prevent it (Warin and Wolff, 2005), but it seems not to prevent moral hazard. Indeed, although included in the definition of the Pact, the dissuasive arm seems to malfunction due to moral hazard behaviours: De Haan, Berger and Jansen (2003) explain that this is, likely, one of the reasons why some countries – Germany and France, for instance – decide to put more emphasis on solving their internal troubles by relaxing their fiscal policies instead of strictly abiding by the letter of the SGP. The SGP should prevent countries from changing their attitudes once within the Eurozone, but a recent literature on the political budget cycle (PBC) and the SGP explains that incumbent governments within the Eurozone display an inclination to raise public expenditure, and thus deficits, before an election, although they have to abide by the SGP (Mink and de Haan, 2005). Buti and van den Noord (2003) analyse the fiscal policies over the 1999–2002 period and find some evidence of expansionary fiscal policies motivated by near-term elections. This result is confirmed by von Hagen (2003) who concludes that there is evidence that fiscal policies were used during the period 1998–2002 before elections. Those results mean that the SGP does not seem to prevent moral hazard behaviours.

As for coordination, the SGP may not represent the optimal means of dealing with this problem. Eichengreen (1990), Cohen (1990), and Branson (1990) study the necessity for a federal budget to augment national budgets. MacDougall (1977), De Grauwe (1990), Italianer and Vanheukelen (1993), and Bryson (1994) analyse the need for a centralized budget as a way of establishing automatic stabilizers with income transfers from better-off to worse-off countries.

As for the ECB's credibility, every EMU member enjoys a lower risk premium than before the creation of the euro, which can be explained by many reasons, such as the liquidity of the market and the improved credibility for the central bank in charge of the European monetary policy. Did the SGP play a role? Since Germany and France did not abide by the SGP for some years, it is difficult to grant this benefit exclusively to the SGP.

In this context, the SGP has had mixed results. The amended version of 2005 embodies some of the changes called for by the economic literature. The lack of flexibility was one of the main criticisms. For instance, Cooper and Kempf (2000) call for some flexibility at the fiscal level, as the ECB lacks the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
tools necessary for stabilization in the presence of country-specific shocks. The new definition of the MTO based on country specificities seems to go into this direction, although this is misleading. Indeed, the MTO in SGP I was ignored by the countries which considered only the three per cent of GDP reference value. Its new definition, based on country specificities, is now at the core of the assessment of countries' fiscal policies. In other words, it seems that for national policies what matters is no longer three per cent but their specific medium-term objectives, by definition lower than three per cent. In this regard, the new preventive arm seems to be tighter than the original SGP, and hence less flexible.
Flexibility refers to the idea of an optimal fiscal policy even though it can be above the three per cent deficit limit. In order to target this optimal fiscal policy and prevent at the same time the existence of a political budget cycle, authors such as Wyplosz (2005), Beetsma and Debrun (2005), Annet, Decressin and Deppler (2005), and Marinheiro (2005) argue for the strong version of institutional reform – the creation of an independent Fiscal Policy Committee (FPC), and a reconfiguring of the debt targets so that they are established, country by country, on a basis of the starting position. This would not automatically mean the end of a fiscal rule in Europe, but it would mean the end of the SGP. However, SGP II seems to go in this direction. Indeed, it allows countries to decide whether they need an independent committee to scrutinize their domestic fiscal policy. In fact, this independent committee is not the same as the FPC: the FPC could decide an optimal fiscal policy above the three per cent deficit limit, which is not the case with a national independent committee.
Another solution to introduce some flexibility without renegotiating the Treaty of Amsterdam, as well as giving some weight back to the debt criterion, is proposed by Pisani-Ferry (2002): allowing countries to opt out of the Excessive Deficit Procedure based on the deficit, and abide by the 60 per cent of GDP debt criterion. In this spirit, a country can have a deficit greater than three per cent as long as its debt is below 60 per cent. But before we consider other amendments, what is the future of the Pact from an institutional perspective?

The future of the Pact

The changes produced by SGP II to SGP I are twofold and concern both the preventive and dissuasive arms of the Pact. First, there is new definition of the medium-term objective (preventive arm). However, it is acceptable if countries do not abide by the medium-term objective as long as they do not go over the reference value of three per cent: regulation 1466/97 explains,

> [in] order to enhance the growth-oriented nature of the Pact, major structural reforms which have direct long-term cost-saving effects, including by raising potential growth, and therefore a verifiable impact on the long-term sustainability of public finances, should be taken into account when defining the adjustment path to the medium-term budgetary objective for countries that have not yet reached this objective and in allowing a temporary deviation from this objective for countries that have already reached it.

This amendment does not relax the reference value of three per cent, but relaxes the constraint imposed by the definition of the medium-term budgetary objective. The amendment adds: 'in order not to hamper structural reforms that unequivocally improve the long-term sustainability of public finances, special

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
attention should be paid to pension reforms introducing a multipillar system that includes a mandatory, fully funded pillar, because these reforms entail a short-term deterioration of public finances during the implementation period.'

The second change was to the dissuasive arm and deals with exceptional circumstances. The dissuasive arm is looser than it was. The European Commission is asked to prepare a report in case of a breach of the deficit reference value by a Member State. If the breach is not justified by an economic downturn (a recession of at least two per cent of GDP) or an exceptional external event, countries have to make deposits to the European Commission that will be transformed into fines in the third consecutive year if a country could not abide by the reference value for three years in a row. The amended regulation loosens the constraint by introducing the notion of relevant factors. Moreover, before asking for deposits when a country breaches the deficit reference value for the first or second time, the Commission should look at the medium-term economic position of a country, at relevant factors, and at the overall quality of public finances.

In the long run, the most important question in assessing the effects of SGP II versus SGP I is to know whether the preventive arm – tighter than under SGP I – will outweigh the loosening of the dissuasive arm. The answer is in the hands of national governments. In retrospect, the SGP does not seem to provide an ideal answer to the branches of the literature studying the potential need for a fiscal rule. This is not surprising, since the SGP is, by its nature, as much a politically designed rule extending the Treaty of Maastricht as an economically designed one.

See Also

- European Central Bank
- fiscal federalism

Bibliography

Amador, J. 2000. Fiscal policy and budget deficit stability in a continuous time stochastic economy. Working Paper No. 384. Universidade Nova de Lisboa.

Annet, A., Decressin, J. and Deppler, M. 2005. Reforming the stability and growth pact. Policy Discussion Paper 05/02, IMF.

Ballabriga, F.C. and Martinez-Mongay, C. 2005. Sustainability of EU public finances. Economic Papers No. 225, European Commission.

Beetsma, R. 2001. Does EMU need a stability pact? In *The Stability and Growth Pact – The Architecture of Fiscal Policy in EMU*, ed. A. Brunila, M. Buti and D. Franco. Basingstoke: Palgrave.

Beetsma, R. and Bovenberg, A. 1999. The interaction of fiscal and monetary policy in a monetary union: balancing credibility and flexibility. In *The Economics of Globalization: Policy Perspectives from Public Economics*, ed. A. Razin and E. Sadka. Cambridge: Cambridge University Press, 1999.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Beetsma, R. and Debrun, X. 2005. Implementing the stability and growth pact: enforcement and procedural flexibility. Working Paper No. 433, European Central Bank.

Beetsma, R. and Uhlig, H. 1999. An analysis of the stability and growth pact. *Economic Journal* 109, 546–71.

Bernoth, K., von Hagen, J. and Schuknecht, L. 2004. Sovereign risk premia in the European government bond market. Working Paper No. 369, European Central Bank.

Blanchard, O. and Giavazzi, F. 2004. Reforms that can be done: improving the SGP through a proper accounting of public investment. Discussion Paper No. 4220, CEPR.

Bohn, H. 1995. The sustainability of budget deficits in a stochastic economy. *Journal of Money, Credit and Banking* 27, 257–71.

Branson, W. 1990. Intégration des marchés financiers et croissance durant la transition vers l'UEM. In *Vers l'Union économique et monétaire européenne*, ed. Ministère de l'Economie. Paris: La Documentation Française.

Bryson, J. 1994. Fiscal policy coordination and flexibility under European monetary union: implications for macroeconomic stabilization. *Journal of Policy Modeling* 16, 541–57.

Buti, M., Eijffinger, S. and Franco, D. 2003. Revisiting the stability and growth pact: grand design or internal adjustment? Economic Papers No. 180, European Commission.

Buti, M. and van den Noord, P. 2003. Discretionary fiscal policy and elections: the experience of the early years of EMU. Economics Department Working Paper No. 351, OECD.

Buti, M. and van den Noord, P. 2004. Fiscal policy in EMU: rules, discretion and political incentives. Economic Papers No. 206, European Commission.

Coeure, B. and Pisani-Ferry, J. 2005. Fiscal policy in EMU: towards a sustainability and growth pact. *Oxford Review of Economic Policy* 21, 598–617.

Cohen, D. 1990. Union monétaire, solvabilité des Etats et politiques budgétaires. In *Vers l'Union économique et monétaire européenne*, ed. Ministère de l'Economie. Paris: La Documentation Française.

Council of the European Union. 1997a. Resolution of the European council on the stability and growth pact. Amsterdam, 17 June. *Official Journal* C 236, 02/08/1997, 1–2.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Council of the European Union. 1997b. Council regulation (EC) No 1466/97 of 7 July 1997 on the strengthening of the surveillance of budgetary positions and the surveillance and coordination of economic policies. *Official Journal* L 209, 02/08/1997, 1–5.

Council of the European Union. 1997c. Council regulation (EC) No 1467/97 of 7 July 1997 on speeding up and clarifying the implementation of the excessive deficit procedure. *Official Journal* L 209 , 02/08/1997, 6–11.

Council of the European Union. 2005. Council report to the European council on improving the implementation of the stability and growth pact. 21 March. Online. Available at http://www.eu2005.lu/en/actualites/documents_travail/2005/03/21stab/stab.pdf, accessed 30 August 2006.

Cooper, R. and Kempf, H. 2000. Designing stabilisation policy in a monetary union. Working Paper No. 7607. Cambridge, MA: NBER.

De Grauwe, P. 1990. La discipline budgétaire dans les unions monétaires. In *Vers l'Union économique et monétaire européenne*, ed. Ministère de l'Economie. Paris: La Documentation Française.

De Grauwe, P. 2003. *Economics of Monetary Union*, 5th edn. Oxford: Oxford University Press.

De Haan, J., Berger, H. and Jansen, D.-J. 2003. The end of the stability and growth pact? Working Paper No. 1093, CESifo.

Dixit, A. 2001. Games of monetary and fiscal interactions in the EMU. *European Economic Review* 45, 589–613.

Dixit, A. and Lambertini, L. 2001. Monetary–fiscal policy interactions and commitment versus discretion in a monetary union. *European Economic Review* 45, 977–87.

Eichengreen, B. 1990. Is Europe an optimal currency area? Discussion Paper No. 478, CEPR.

Eichengreen, B. and Wyplosz, C. 1998. The stability pact: more than a minor nuisance. *Economic Policy* 26, 67–113.

Fatas, A. 2005. Is there a case for sophisticated balanced-budget rules? Economics Department Working Papers No. 466, OECD.

Issing, O. 2002. On macroeconomic policy co-ordination in EMU. *Journal of Common Market Studies* 40, 345–58.

Italianer, A. and Vanheukelen, M. 1993. Proposals for community stabilization mechanisms: some

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

historical applications. *European Economy* 5, 493–510.

MacDougall, D. 1977. *Report of the Study Group on the Role of Public Finance in European Integration*. Brussels: Commission of the European Communities.

Marinheiro, C. 2005. Has the stability and growth pact stabilized? Evidence from a panel of 12 European countries and some implications for the reform of the Pact. Working Paper No. 1411, CESifo.

Mink, M. and de Haan, J. 2005. Has the stability and growth pact impeded political budget cycles in the European union? Working Paper No. 1532, CESifo.

Mongelli, F. 1999. The effects of the European economic and monetary union (EMU) on fiscal sustainability. *Open Economies Review* 10, 31–61.

Nielsen, S. 1992. A note on the sustainability of primary budget deficits. *Journal of Macroeconomics* 14, 745–54.

Perotti, R., Strauch, R. and von Hagen, J. 1998. Sustainability of public finances. CEPR and ZEI.

Pisani-Ferry, J. 2002. Fiscal discipline and policy coordination in the Eurozone: assessment and proposals. In *Budgetary Policy in E(M)U: Design and Challenges*, Ministerie van Financiën. The Hague.

Treaty establishing the European community. 2002. *Official Journal* C 325, 24 December.

Uhlig, H. 2002. One money, but many fiscal policies in Europe: what are the consequences? Discussion Paper No. 3296. London: CEPR.

von Hagen, J. 2003. Fiscal discipline and growth in Euroland: experiences with the stability and growth pact. *Wirtschaftspolitische Blätter* 50, 163–83.

Warin, T. and Wolff, L. 2005. Europe's deficit free riders: a panel data analysis. *European Political Economy Review* 3, 5–17.

Wyplosz, C. 2005. Fiscal policy: institutions versus rules. *National Institute Economic Review* 191, 64–78.

How to cite this article

Warin, Thierry. "stability and growth pact." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
dictionaryofeconomics.com/article?id=pde2008_S000488> doi:10.1057/9780230226203.1601

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stable population theory

Shripad Tuljapurkar
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A population's age structure and growth are determined by rates of fertility, mortality and migration. Stable population theory provides a widely useful mathematical framework, described here, that connects a fixed set of rates to the population dynamics they generate. This theory makes it possible to trace causes and consequences of population change, to establish methods for estimating rates, and to make projections of future population. Much of the power of stable theory rests on the fact that the key features of population dynamics with fixed rates can be generalized, as discussed here, to rates that vary over time.

Keywords

continuous and discrete time models; dependency ratio; economic demography; fertility; forecasting; generation times; international migration; Markov processes; mortality; population growth; population momentum; population projections; stable population theory; stochastic stable theory

Article

Demographers are centrally concerned with the relationship between a population's age structure and its rates of mortality, fertility and immigration. Alfred Lotka (Sharpe and Lotka, 1911; Lotka, 1939) devised the fundamental mathematical form of this relationship for a population with no migration, in the event that age-specific mortality and fertility do not change over time.

Renewal equation

Lotka dealt primarily with females; we shall say more about males shortly. Mortality is described by an instantaneous mortality rate $\mu(a)$ at each age a, and determines the probability $l(a)$ of surviving from birth to age a, as $l(a) = \exp\left(-\int_0^a ds\mu(s)\right)$. Demographers refer to $l(a)$ as survivorship. Fertility is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
described by the rate $m(a)$ of female births per female of age a. The goal of the theory is to track changes in population number and age composition over time. Suppose that we know the numbers of females at all ages at some initial time that we denote as $t = 0$. Writing $B(t)$ for the rate at which females are born to the total population at any later time $t > 0$, Lotka derived what is called the 'renewal equation',

$$B(t) = \int_0^t ds\, l(s)\, m(s)\, B(t-s) + h(t).$$

To understand this equation, note first that the number of females $n(s,t)$ at age s at time t must be simply the number of survivors, $l(s)\, B(t-s)$, of the $B(t-s)$ females born at time $t-s$. Thus, the first term on the right side of the equation sums births to females whose ages range from zero to t, and who were born at any time between zero and t. The second term $h(t)$ represents births to all females who were alive at the initial time $t = 0$, and whose ages are therefore greater than t. As time passes, the females whose children are counted in $h(t)$ get older and eventually die. Thus after a long time passes (that is, for large values of t) only the first term on the right of the renewal equation remains. Lotka found that the solutions to this simpler equation are of a particular form that describes what is called a stable population.

Classical stable theory

A *stable population* has an unchanging age structure and a constant exponential growth rate r; both structure and growth rate are determined by vital rates (mortality, fertility). In a stable population, births at time $t-a$ must differ from births at time t by an exponential factor in the growth rate, $B(t-a) = e^{-ra} B(t)$. Also, a stable population obeys the renewal equation at long times t when the term $h(t)$ is zero. Inserting the above exponential relationship into the renewal equation shows that the growth rate r satisfies what is called Lotka's *characteristic equation*,

$$1 = \int_0^\infty da\, l(a)\, m(a)\, e^{-ra}.$$

In a stable population, the number of females at age s at time t is $n(s,t) = l(s)\, B(t-s) = l(s)\, e^{-rs} B(t)$. Hence the proportion of a stable population that is between ages a and $a + da$ in a stable population is $u(a) = C\, l(a)\, e^{-ra}$, with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$C = \frac{1}{\int_0^\infty da\, l(a)\, e^{-ra}}.$$

The per-capita birth rate (also called the crude birth rate) in a stable population is $b = \int_0^\infty da\, m(a)\, u(a)$, and the per-capita death rate (also called the crude death rate) in a stable population is $d = \int_0^\infty da\, \mu(a)\, u(a)$. The stable growth rate is $r = (b - d)$ (as may be shown by using the relationship $\mu(a) = -(1/l(a))(dl(a)/da)$ in d and integrating by parts).

Given unchanging vital rates and any initial population in the full renewal equation, Lotka showed that the population obtained by solving the renewal equation must eventually approach a stable population whose growth rate is determined by the characteristic equation. This result explains the adjective stable in the term, stable age distribution. Lotka's original proof was put on a secure mathematical footing by Feller (1941; 1971); such stability is characteristic of many populations that undergo mortality and renewal (including, for example, light bulbs or laptop computers that die at a rate that depends on their age and must therefore be replaced at some rate). The property of stability of the age distribution when mortality and fertility are constant in time is known as ‘demographic strong ergodicity’.

In Lotka's classical theory of demography, male births are accounted for by noting that the human sex ratio at birth is remarkably constant over time and place, close to 1.05 male births for every female birth, except in cases where deliberate preference for one or the other sex leads to an excess mortality of the less favoured sex. Thus the numbers of males born are computed simply as a constant multiple of the number of females born. Male age structure will obviously become stable along with the female age structure, but male mortality and thus survivorship are usually different from female.

In practice, the renewal equation, in which time is a continuous variable, is often replaced by a discrete-time version in which time advances in discrete units. In that case the age composition of a population is represented by a vector of population numbers in successive discrete age classes, and the renewal equation is replaced by a matrix recursion. Leslie (1945) formulated and analysed the properties of the discrete equation, which has much the same properties as the continuous version. Coale (1972), working with time as a continuous variable, and Keyfitz (1977), working with time as a discrete or a continuous variable, provide authoritative discussions of the mathematics and application of stable population theory.

Applying classical stable theory

Stable theory and the characteristic equation yield powerful, fundamental insights into the relationship between mortality, fertility, population growth rate and age structure. We mention a few important examples here. The fact that a stable population's age structure is proportional to $l(a)e^{-ra}$ shows that a population's age pyramid is steeper for faster growing populations and shallower for low-mortality populations. Economists are often interested in the proportions of a population at young (usually under 20 yrs), working (20 to 65 yrs) and old (over 65 yrs) ages, and in the dependency ratio (the sum of young and old divided by the number working). With mortality fixed, the dependency ratio for a stable

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
population is high when the growth rate r is either large and negative (fewer young, many old) or large and positive (many young, fewer old), and there is an optimal growth rate at which dependency ratio is minimized. With growth rate fixed, the dependency ratio increases when mortality declines because more people survive to reach old age. These properties of stable populations carry over qualitatively to real populations, illuminating the effects of changes in mortality and fertility on population composition. In particular many industrialized countries in the 21st century have small positive or even negative growth rates as well as low mortality; today's populations thus have a larger fraction of older individuals and a smaller fraction of young than the higher-fertility and lower-mortality populations of the 19th and early 20th centuries. Changes in population age structure play an important role in the theory of economic demography, especially in understanding the role of transfers between different age segments of a population (Lee, 1994).

In stable theory, the contribution of fertility to population growth rate is described by the net replacement rate,

$$\mathrm{NRR} = \int_0^\infty da\, l(a)\, m(a),$$

which is the expected lifetime reproduction of a female. A stable population grows ($r > 0$), declines ($r < 0$) or is stationary depending on whether the NRR is greater than, less than, or equal to one. For any given pattern of mortality the condition $\mathrm{NRR} = 1$ defines replacement fertility, the fertility rate at which $r = 0$. The generation time of a stable population is the average number of years over which a cohort of mothers produces its daughters, and is defined as

$$T = \int_0^\infty da\, a\, l(a)\, m(a)\, e^{-ra}.$$

Generation times in contemporary populations are set primarily by the age pattern of reproduction. Historically generation times have ranged from 20 to 25 years for human populations, being on the lower end in high-fertility populations that have a high growth rate.

Stable population theory has found wide application in ecology and evolutionary biology, as discussed extensively by Caswell (2001). In the biological context, the stable population growth rate r is often used to measure the fitness (in a Darwinian sense) of a particular biological combination of fertility and mortality. Another useful characterization of fertility and mortality patterns is given by the reproductive value $v(a)$ of an individual of age a in a stable population (Fisher, 1930),

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$v(a) = \int_a^{\infty} db \frac{l(b)}{l(a)} m(b)\, e^{-r(b-a)}.$$

Clearly $v(a)$ is the discounted present value of an individual's future reproduction after age a, using the discount rate r. The characteristic equation tells us that $v(0) = 1$, so the reproductive value measures the future contribution of an individual of age a relative to a newborn. In evolutionary theory, and in some questions in economic demography, we are interested in the effect on population growth rate of a change in either mortality or fertility at a particular age. An increase in fertility at age a changes r by an amount proportional to the reproductive value $v(a)$, whereas an increase in mortality at age a changes r by an amount proportional to the product $l(a)e^{-ra}\, v(a)$.

An elegant and important concept derived from stable theory is that of population momentum (Keyfitz, 1971). Suppose that a stable population is growing at some rate $r > 0$ and that the population makes an instantaneous transition to replacement fertility. Stable theory implies that with this reduced fertility the population will eventually become stable and stationary with growth rate $r = 0$. Keyfitz asked: what is the ratio of the size of the final stationary population to the size of the population just before the fertility transition occurs? The answer is that the ratio equals *population momentum*. Write $v_0(a)$ for the reproductive value in the final stationary population, T_0 for the generation time in the final stationary population, and $u(a)$ for the stable structure of the initially growing population. Then Keyfitz's population momentum equals $(1/T_0)\int_0^{\infty} da\, v_0(a)\, u(a)$. In the real world, fertility transitions take time and so actual momentum is generally larger than Keyfitz's momentum, as shown by Li and Tuljapurkar (1999).

Stable theory has also been extended to include the effects of migration, with much attention focused on the effects of immigration into low-fertility populations. Arthur and Espenshade (1988) and Feichtinger and Steinmann (1992) analyse the case where a stream of immigrants of known age distribution and total number is added annually to a population with below-replacement fertility. This case is relevant to a number of industrialized countries. Over time, the population's age structure will again converge to a stable age structure that is determined jointly by the age structure of the immigrant flow, the vital rates of the resident population, and the rate at which immigrants' vital rates converge to those of the residents. Feichtinger and Steinmann point out that the general theory of stable populations with immigration is closely connected with the theory of manpower systems and other social processes; for a review of the latter see Vassiliou (1997).

The concepts and mathematics of stable theory are relevant to the dynamics of populations that are structured by variables other than age. In demography, we may be interested in the joint distribution of age and parity (a female's parity is the number of offspring that she has had), or of age and health status. In biological applications, we may be concerned with a stage variable (for example, size or developmental state) rather than with age: this is especially the case for plants, insects and other organisms in which age is not directly observable in nature, or in which vital rates depend directly on size or state rather than on age. In such cases, Lotka's renewal arguments can nonetheless be applied, except that we must keep track of the distribution of individuals according to both their age and their

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

stage. Vital rates now include mortality and fertility as function of age and stage, as well as rates at which individuals move between stages during the course of life. It should be obvious that such stage-based demography is relevant to economic analyses of the human life cycle in which we are interested in the transitions made by individuals between stages such as marriage, divorce and employment. The stable theory of stage-based demography parallels the theory we describe here, and an account of the biological theory is given by Caswell (2001).

Stable theory has been an essential component of successful efforts by demographers to create widely applicable models of vital rates, as nicely discussed by Preston, Heuveline and Guillot (2000). The authors also describe an extension of stable theory, in which a population's different age segments change at different rates over time. Thus, if the growth rate of individuals aged *a* at time *t* is *r*(*a*, *t*), the population density of individuals at age $a + b$ at a later time $t + b$ is given by

$$n(a + b, t + b) = n(a, t)\,\frac{l(a + b)}{l(a)} \times \exp\left\{\int_0^b ds\, r(a + s, t + s)\right\}.$$

This expression allows us to use observations at different times on population structures (such as censuses or surveys) to estimate the growth rates of particular cohorts. Since populations in the real world are rarely stable, this approach is often useful.

Demographic weak ergodicity

Ansley Coale (1957) pointed out that mortality and fertility rates in practice will vary with time, violating the assumptions of strong demographic ergodicity and calling into question the use of classical stable theory. He argued, however, that human populations should forget their more remote history, in the sense that today's population composition should be most strongly influenced by the recent rather than the distant past. Lopez (1961; 1967) provided a mathematical framework for this argument by defining *demographic weak ergodicity*: if mortality and fertility rates change with time in some arbitrary (but demographically sensible) way, and two populations with distinct initial age structures are subject to the same sequence of changing rates, then as time goes by the age structures of these two populations will become proportional. Lopez's original proof has been strengthened and extended in more recent work, as discussed for example by Tuljapurkar (1982). An important consequence of demographic weak ergodicity is that we are justified in focusing attention on relatively recent changes in mortality and fertility as being the key to current and future age structures.

Stochastic stable theory

The work of Coale and Lopez provided the impetus for a powerful generalization of stable theory to cases when mortality and fertility rates change with time in a stochastic fashion. Analysis of historical mortality and fertility rates for any real population reveals that these rates change with time. Some

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
changes are slow and secular, as has been true for the 20th century decline in mortality and for the transition to low fertility in some parts of the world. But many changes in rates occur over time intervals of a generation or less and are often quickly reversed: 20th century examples include fertility swings in industrialized countries, and short-term (decadal or faster) variability in age-specific mortality rates. Cohen (1976) formulated a mathematical description of such variability for a discrete-time population model in which fertility and mortality rates vary over time in response to some underlying stochastic process. (Cohen's original model assumed a finite-state Markov process but his results have been extended to many other stochastic processes.) The vital rates (mortality, fertility) vary over time but have a stationary probability distribution. A central assumption on the variability in rates is that any initial population structure (for example, with only children present, or only adults present) will eventually lead to a population in which every age class is represented. Given this condition, Cohen showed that, if the same random sequence of vital rates is applied to two populations with different initial age structures, the two populations will change subsequently so that their age structures become proportional over time. Thus the stochastic sequence of mortality and fertility rates generates a stationary stochastic sequence of population age structures that is maintained over time. This property is called *demographic stochastic weak ergodicity*. Let $X(t)$ denote the time sequence of vital rates and $Y(t)$ the time sequence of population structures. Then there is a joint stationary probability distribution of these two quantities. The stationary stochastic sequence of age structures is a *time-varying stable population*.

The property of stochastic weak ergodicity allows demographers to focus attention on the time-varying stable population. We note here some useful properties of this kind of stable population; see Tuljapurkar (1990) and Caswell (2001) for further discussion. The growth rate r of a classical stable population is replaced here by a long-run stochastic growth rate that satisfies a stochastic analog of the Lotka characteristic equation. The age structure itself can be described in terms of its moments, means, variances, covariances and autocorrelations. Variability over time in population structure reflects both the time-averages of mortality and fertility and the variances and covariances of these vital rates over time. Finally, stochastic theory allows us to characterize population trajectories in terms of probabilities of future events; for example, we can compute the probability that a dependency ratio will become unusually high or low over some specified time interval. Stochastic stable theory has been useful in demographic applications, especially to forecasting and fiscal problems (see below), and in a variety of ecological applications.

Population forecasts

The age structures of human populations can be effectively described, analysed, and even projected with reasonable accuracy using classical one-sex renewal theory. In many situations demographers are called upon to make extremely long-term forecasts of population number and composition; for example, the United States Social Security Administration's trustees require a 75-year forecast. In most forecasts made by institutions such as census bureaus or the United Nations, experts first make a set of alternative projections of mortality, fertility and immigration, and then generate projections along these alternative futures. These alternative futures are referred to as scenarios. In most cases, the vital rates in these projected scenarios settle to constant values after some initial period of time, often over a generation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
length or so. Correspondingly, the long-run forecast populations always approach classical stable populations that correspond to the alternative long-run rates.

Stochastic stable population theory generalizes this forecasting approach considerably by projecting vital rates as non-stationary stochastic processes. In most cases, the stochastic variability in the rates is fixed or changes very slowly, and there is a long-run slow secular change in the time-average vital rates. With vital rates projected in this way it is possible to generate stochastic population forecasts. At long times, these forecasted population structures usually approach stable time-varying populations. A major advantage of using stochastic stable theory is that probabilistic projections can be made of population structures, dependency ratios, and associated quantities of policy interest (Lee and Tuljapurkar, 2000).

See Also

- economic demography
- population aging
- population dynamics

Bibliography

Arthur, W.B. and Espenshade, T.J. 1988. Immigration policy and immigrants' ages. *Population and Development Review* 14, 315–26.

Caswell, H. 2001. *Matrix Population Models: Construction, Analysis and Interpretation*, 2nd edn. Sunderland, MA: Sinauer.

Coale, A.J. 1972. *The Growth and Structure of Human Populations: A Mathematical Investigation*. Princeton: Princeton University Press.

Cohen, J.E. 1976. Ergodicity of age structure in populations with Markovian vital rates, I: countable states. *Journal of the American Statistical Association* 71, 335–9.

Feichtinger, G. and Steinmann, G. 1992. Immigration into a population with fertility below replacement level – the case of Germany. *Population Studies* 46, 275–84.

Feller, W. 1941. On the integral equation of renewal theory. *Annals of Mathematical Statistics* 12, 243–67.

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*, vol. 2, 2nd edn. New York: Wiley.

Fisher, R.A. 1930. *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Keyfitz, N.C. 1971. On the momentum of population growth. *Demography* 8, 71–80.

Keyfitz, N.C. 1977. *Introduction to the Mathematics of Population – With Revisions*. New York: Addison-Wesley.

Lee, R.D. 1994. The formal demography of population aging, transfers, and the economic life cycle. In *The Demography of Aging*, ed. L. Martin and S. Preston. Washington, DC: National Academy Press.

Lee, R. and Tuljapurkar, S. 2000. Population forecasting for fiscal planning: issues and innovations. In *Demography and Fiscal Policy*, ed. A. Auerbach and R. Lee. Cambridge: Cambridge University Press.

Leslie, P.H. 1945. On the use of matrices in certain population mathematics. *Biometrika* 33, 183–212.

Li, N. and Tuljapurkar, S. 1999. Population momentum. *Population Studies* 53, 255–62.

Lopez, A. 1961. *Problems in Stable Population Theory*. Princeton: Office of Population Research.

Lopez, A. 1967. Asymptotic properties of a human age distribution under a continuous net maternity function. *Demography* 4, 680–7.

Lotka, A.J. 1939. A contribution to the theory of self-renewing aggregates, with special reference to industrial replacement. *Annals of Mathematical Statistics* 10, 1–25.

Preston, S.H., Heuveline, P. and Guillot, M. 2000. *Demography: Measuring and Modeling Population Processes*. Malden, MA: Blackwell.

Sharpe, F.R. and Lotka, A.J. 1911. A problem in age-distribution. *Philosophical Magazine* 21, 435–8.

Tuljapurkar, S. 1982. Population dynamics in variable environments. IV: Weak ergodicity in the Lotka equation. *Journal of Mathematical Biology* 14, 221–30.

Tuljapurkar, S. 1990. *Population Dynamics in Variable Environments*. Berlin: Springer Verlag.

Vassiliou, P.-C.G. 1997. The evolution of the theory of non-homogeneous Markov systems. *Applied Stochastic Models and Data Analysis* 13, 159–76.

How to cite this article

Tuljapurkar, Shripad. "stable population theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 02 January 2009 <http://www.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dictionaryofeconomics.com/article?id=pde2008_S000226> doi:10.1057/9780230226203.1602

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Stackelberg, Heinrich von (1905 – 1946)

Wilhelm Krelle

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Heinrich von Stackelberg was born on 31 October 1905, in Kudinowo, near Moscow, where his father was the director of a factory. The homeland of the family was the Baltic state of Estonia, although his mother was born in Argentina of Spanish descent. The family escaped the Russian revolution, retiring first to Yalta in the Crimea, and afterwards to Germany. They initially settled in Ratibor, Silesia but moved to Cologne in 1923. He completed his high school education in Cologne, studied economics at the University of Cologne, obtaining his ‘Diplomvolkswirt’ (master of economics) in 1927, ‘Dr. rer.pol.’ in 1930, and his habilitation in 1935.

He began his scientific career in 1928 as an assistant professor at the University of Cologne (1928–35). From 1935 until 1941 he was ‘Dozent’ and ‘ausserordentlicher Professor’ (associate professor) at the University of Berlin, and from 1941 until 1944 full professor at the University of Bonn. During World War II he was for some time drafted to military service. In 1944 and 1945 he held a guest professorship at the University of Madrid. He died at the early age of 41 in Madrid on 12 October 1946.

Stackelberg was the most gifted theoretical economist in Germany during his time. His habilitation thesis *Marktform und Gleichgewicht* (1934) has had a lasting influence on price theory. ‘Stackelberg asymmetric duopoly’ is known all over the world. His contributions to Austrian capital theory are the basis for all modern extensions of this theory. His textbook *Grundzüge der theoretischen Volkswirtschaftslehre* (1943) was the first ‘modern’ introduction to economics in the sense that it is based on a coherent theory of household and firm behaviour. Moreover, Stackelberg contributed to several other fields: cost theory, exchange rate theory, saving theory and others. In Germany he was one of the few leading economists who introduced mathematics into economics and took up the Anglo-Saxon approach in price and cost theory (Edgeworth, Marshall, Hicks, Harrod, Chamberlin and others).

The difficulty of oligopoly theory consists in the fact that the oligopolists are in a game theoretic situation which, in general, cannot be put into the form of a pure maximum problem. Stackelberg's seminal idea was that this can nevertheless be done if – in the case of a duopoly – one firm takes a ‘dependent’ position (i.e. takes the actual price or production of the other firm as given) and the other an ‘independent one (i.e. knows this behaviour and fixes its price or production accordingly so that it maximizes its profits or other utility indices). If both firms wish to be in the ‘dependent’ position, a Cournot-type equilibrium results; on the other hand if both firms wish to be in the ‘independent’

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
position, a contradiction arises since each firm assumes a behaviour of the other which is incompatible with its actual behaviour. If they nevertheless fix their prices (or production) at that level, a ‘Bowley’-type oligopoly solution, as Stackelberg calls it, would emerge. Since it is unclear which position the firms will take, Stackelberg considered the oligopoly as a market form without equilibrium. *Marktform und Gleichgewicht* (1934) is comparable with Chamberlin's *The Theory of Monopolistic Competition* (1933) and Joan Robinson's *The Economics of Imperfect Competition* (1933), but goes further in the analysis and in mathematical rigour.

Stackelberg accepted Austrian capital theory, which emphasizes the time structure of production (‘zeitlicher Aufbau der Produktion’). The main drawback of this theory is that one of its basic concepts, namely the average gestation period, could not be well defined and measured for a modern interdependent economy. In ‘Kapital und Zins in der stationären Verkehrswirtschaft’ (1941) Stackelberg suggests the following solution. In a simple economy where the original factor input takes place in period O and the product ripens by nature (such as in the production of wood), the subsistence fund S, the yearly income (=harvest) Y, the interest factor $q=1+r$, where r=rate of interest, and the gestation period T satisfy the relation $S=Y/q^T$. Stackelberg defines an economy as equivalent to this simple economy, if they correspond with respect to S, Y and r, where S is identified with labour income L. Thus the average gestation period may be calculated by $T=(\log Y - \mathrm{Log}\, L)/\log q$.

In the article ‘Beitrag zur Theorie des individuellen Sparens’ (1939), Stackelberg deals with the problem: why does a household save? What are the effects of interest rate expectations on household saving? He took up the conceptual framework of Hicks and Allen (1934) and applied it to the allocation of expenditures in the time space. He derived Böhm-Bawerk's law of under-evaluation of future commodities from the law of declining marginal rates of substitution and showed how the optimal allocation of expenditure in time depends on it.

Stackelberg was a neoclassical economist. In his opinion, Keynes really added nothing new to available economic knowledge. He also considered Keynes's interest rate theory as a special case of Böhm-Bawerk's theory of exchange of present against future commodities (see ‘Zins und Liquidität’, 1947). Stackelberg kept intimate relations with that group of German economists (Walter Eucken, Erwin v. Beckerath and others) who during the war prepared the transition of the German economy to a free enterprise system. In spite of his untimely death, his influence especially on economic theory in Germany was most important in the sense that he initiated the reorientation of German economic thinking to the Anglo-Saxon approach. His very original contributions to economic theory have had a lasting effect.

Selected works

1932. *Grundlagen einer reinen Kostentheorie*. Vienna: Julius Springer.

1934. *Marktform und Gleichgewicht*. Vienna and Berlin: Julius Springer.

1938a. Probleme der unvollkommenen Konkurrenz. *Weltwirtschaftliches Archiv*, 95–141.

1938b. Arbeitszeit and Volkswirtschaft. *Jahrbuch des Arbeitswissenschaftlichen Instituts*, Berlin, 61–86.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1938c. Das Brechungsgesetz des Verkehrs. *Jahrbüchr für Nationalökonomie und Statistik*, 680–96.

1939a. Beitrag zur Theorie des individuellen Sparens. *Zeitschrift für Nationalökonomie*, 167–200.

1939b. Theorie der Vertriebspolitik und Qualitätsvariation. *Schmollers Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft im Deutschen Reich*, 43–85.

1940. Die Grundlagen der Nationalökonomie, Bemerkungen zu dem gleichnamigen Buch von Walter Eucken. *Weltwirtschaftliches Archiv* 41, 245–85.

1941. Kapital und Zins in der stationären Verkehrswirtschaft. *Zeitschrift für Nationalökonomie*, 25–61.

1941a. Elemente einer dynamischen Theorie des Kapitals. *Archiv für mathematische Wirtschafts-und Sozialforschung*, 8–29, 70–93.

1943. *Grundzüge der theoretischen Volkswirtschaftslehre*. Stuttgart and Berlin: Kohlhammer.

1944. Theorie des Wechselkurses bei vollständiger Konkurrenz. *Jahrbücher für Nationalökonomie und Statistik*, 1–65.

1947. Zins und Liquidität. Eine Auseinandersetzung mit Keynes. *Schweizerische Zeitschrift fü Volkswirschaft und Statistik*, 311–28.

1951. *Grundlagen der theoretischen Volkswirtschaftslehre*. Tübingen and Zurich: J.C.B. Mohr and Polygraphischer Verlag.

Bibliography

Chamberlin, E. 1933. *The Theory of Monopolistic Competition*. Cambridge, Mass: Harvard University Press.

Hicks, J.R. and Allen, R.G.D. 1934. A reconsideration of the theory of value. *Economica*, NS 1, 196–219.

Robinson, J. 1933. *The Economics of Imperfect Competition*. London: Macmillan.

How to cite this article

Krelle, Wilhelm. "Stackelberg, Heinrich von (1905–1946)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009
<http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000227>
doi:10.1057/9780230226203.1603

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Stakhanovism

Lewis H. Siegelbaum
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Stakhanovism

Article

Stakhanovism was a movement begun in the Soviet Union in 1935 to increase labour productivity by the popularization of work techniques reputedly initiated by workers themselves.

On 31 August 1935, Aleksei Grigorevich Stakhanov, a 30-year-old miner in the Donets Basin, cut 102 tons of coal during his six-hour shift. This amount represented 14 times his quota, and within a few days was hailed by *Pravda* as a world record. Anxious to celebrate and reward individuals' achievements in production that could serve as stimuli to other workers, the Soviet Union's Communist Party launched the Stakhanovite movement, or Stakhanovism. The title 'Stakhanovite', conferred on workers and peasants who set production records or otherwise demonstrated mastery of their assigned tasks, quickly superseded that of 'shock worker' (*udarnik*). Day by day throughout the autumn of 1935, the campaign intensified, culminating in an All-Union Conference of Stakhanovites in industry and transportation which met in the Kremlin in late November. Outstanding Stakhanovites mounted the podium to recount how, defying their quotas and often the scepticism of fellow workers and bosses, they applied new techniques of production to achieve stupendous results for which they were rewarded with wages that reached dizzying heights. Stalin captured the upbeat mood of the conference when, by way of explaining how such records were possible only in the 'land of socialism', he uttered the phrase, 'Life has become better, and happier too.' Widely disseminated, and even set to song, Stalin's words served as the motto of the movement.

The year 1936 was declared a Stakhanovite year. Competitions among workers during designated Stakhanovite months, ten-day periods (*dekady*), and shifts spread the movement everywhere in the Soviet Union. Not a single place of work was without its Stakhanovites. Even the Gulag got into the act. The highest proportion of Stakhanovites could be found in the extractive industries, the energy sector, and railway transportation, where upwards of 40 per cent of all workers were so designated by August 1936. Young male workers who had passed technical training courses, were classified as at least semi-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
skilled and had an average of three to five years experience were most likely to be represented among Stakhanovites. However, what was characteristic in industry was not the case in agriculture, where the most prominent Stakhanovites were women. They included Pasha Angelina – brigade leader of the first all-female tractor brigade – and Maria Demchenko, the sugar-beet cultivator. The enormous publicity surrounding these and other female collective- and state-farm workers suggests a continuation of the party's efforts, begun during collectivization, to forge an alliance with rural women against the previously dominant patriarchy of peasant families. Lending support to this interpretation is the rather frequent mention in rural female Stakhanovites' testimonials of having been orphans and of overcoming the resistance of unenlightened husbands.

Stakhanovism encompassed lessons about not only how to work but how to live. In addition to providing a model for success on the shop floor, in the mine, or in the field, it conjured up images of the good life. Many of the qualities Stakhanovites were supposed to exhibit at work – cleanliness, neatness, punctuality, preparedness, and a keenness for learning – were applicable at home, too. These qualities were associated with *kulturnost* ('culturedness'), the acquisition of which marked an individual as a New Soviet Man or Woman. Advertisements for perfume in journals intended for Stakhanovites, articles about Stakhanovites on shopping sprees, photographs of Stakhanovites sharing their happiness with their families, newsreels showing them moving into comfortable apartments and driving new automobiles presented to them as gifts, all symbolized *kulturnost*. Wives of male Stakhanovites had an important part to play in the movement as helpmates preparing nutritious meals, keeping their apartments clean and comfortable, and otherwise creating a cultured environment in the home so that their husbands were well rested and eager to work with great energy. It was also important to demonstrate that Stakhanovites were admired by their workmates and considered worthy of holding public office.

Stakhanovites, however, were not necessarily popular. Even before the raising of output norms in early 1936, workers who had not been favoured with the best conditions and consequently struggled to fulfil their norms expressed resentment of Stakhanovites by verbally and even physically abusing them. Foremen and engineers, only too well aware that 'recordmania' and the provision of special conditions for Stakhanovites created disruptions in production and bottlenecks in supplies, also on occasion 'sabotaged' the movement. At least that was the accusation made against many who often served as scapegoats for the failure of Stakhanovism to fulfil its promise of unleashing the productive forces of the country. Thus, at least in an indirect way, Stakhanovism fed the Terror of 1936–8.

In the course of those years, quite a few Stakhanovites received special educational training followed by promotion into the ranks of management; others were sent on tours of worksites where they demonstrated their skills; many became deputies in the Supreme Soviets of the USSR and its constituent republics. Eventually, Stakhanovism was routinized, becoming merely another task for party and trade union committees to carry out. It continued into the war and even enjoyed something of a revival in the post-war years when it was exported to eastern Europe. Stakhanov himself served in a number of honorific administrative positions in the coal-mining industry before descending into alcoholism and disability. When he died in 1977 he was all but forgotten, although the eastern Ukrainian town of Kadievka (Lugansk oblast) where he had worked and set his record had its name changed to Stakhanov. The 50th anniversary of Stakhanovism in 1985 was observed in the Soviet Union with an outpouring of popular and scholarly literature, museum displays, lectures, and special exhibitions, all of which were intended to inspire the 'Stakhanovites of the 1980s'. It is hard to imagine Stakhanovism inspiring anything in the post-Soviet era except perhaps ridicule.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- labour discipline
- Soviet Union, economics in
- Soviet economic reform
- Stalinism, political economy of

Bibliography

Buckley, M. 1999. Was rural Stakhanovism a movement? *Europe-Asia Studies* 51, 299–314.

Co-operative Publishing Society of Foreign Workers in the U.S.S.R. 1936. *Labour in the Land of Socialism: Stakhanovites in Conference*. Moscow.

Davies, R. and Khlevnyuk, O. 2002. Stakhanovism and the Soviet economy. *Europe-Asia Studies* 54, 867–904.

Siegelbaum, L. 1988. *Stakhanovism and the Politics of Productivity in the USSR, 1935–1941*. New York: Cambridge University Press.

How to cite this article

Siegelbaum, Lewis H. "Stakhanovism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000499> doi:10.1057/9780230226203.1604

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Stalinism, political economy of

Paul R. Gregory
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes the economic working arrangements put in place in the Soviet Union in the early 1930s by Stalin from the vantage point of Mises' and Hayek's scepticism about the feasibility of socialism. It describes the high-level management of the economy, the dictator's curse, the ad hoc nature of planning, the handling of principal–agent problems, the management of worker morale, the manner in which investment was maximized, the dictator's aversion to rules, and the use of coercion to achieve economic goals. Stalin's use of forced labour in his Gulag system is described as coercion failure.

Article

'Stalinism' refers to a political–economic system of state ownership and administrative-resource allocation directed by a monopoly political party (in Stalin's case, the Politburo of the Central Committee of the Communist Party, or Stalin himself), which combines economic incentives and 'repression' to achieve its economic and political goals (Gregory, 2004). Stalinism was first created in the Soviet Union as Stalin gained totalitarian authority and has subsequently been practised in modified forms by dictators in China (Mao), Cuba (Castro), North Korea (Kim Jong Il), Iraq (Saddam Hussein) and Cambodia (Pol Pot). 'Early' Stalinism dates from the expulsion from ruling circles of Stalin's last significant opposition in 1929 and 1930, when Stalin still had to build majorities within the Politburo (Khlevnyuk, 1996). 'High Stalinism' dates from late 1934 until his death in March 1953, when Stalin's decisions could no longer be challenged and his top officials carried out orders rather than participated in collective decisions. By the mid-1930s, the Politburo had ceased to meet regularly, and the party and state were run by informal groups appointed directly by Stalin. Stalinism is distinguished from the 'administrative-command' system of the post-Stalin era by the latter's less extreme use of repression and by its collective leadership.

The economic policy of Stalinism was marked by extremely high rates of capital accumulation in heavy industry and defence, with lesser priority for consumer goods and services. Nevertheless, there were investment cycles both in Stalin's Russia and throughout the Soviet empire, even though investment was dictated politically rather than by market forces. Although some interpret these investment cycles as temporary bouts of moderation, Stalin deliberately reduced investment when he feared that low real wages would harm work effort or, worse, lead to uncontainable civil unrest. Stalin used his secret police to gauge the mood of workers and peasants for this purpose.

Although Stalin's predecessor, V.I. Lenin, established the institutions of terror, such as a special secret police for political enemies and an arbitrary system of 'socialist legality', it was Stalin who initiated 'mass operations' against his enemies. In January 1930, he ordered the arrest, deportation, and execution of kulaks (wealthier peasants and virtually any perceived regime opponent). In this 'dekulakization' campaign, more than two million peasants were deported to 'special settlements' or to the 'corrective labour camps' of the Gulag. The term 'Gulag' denotes the Main Administration of Camps under the jurisdiction of the interior ministry. Stalin used the assassination of the Leningrad party boss in December of 1934 to purge the party and state leadership of remaining political enemies. Their executions were pronounced at the Moscow show trials, the first being the 1936 trial of G. Zinoviev and L. Kamenev, expelled Politburo members and opponents of Stalin. The purge of the party elite broadened into 'mass operations', or the Great Terror, in July 1937 initiated by telegrams from Stalin and operational plans by NKVD head, N. Ezhov, which set execution and imprisonment quotas for 65 regions. The NKVD was the Peoples' Commissariat for Internal Affairs, which was charged with carrying out terror operations. The Great Terror was tightly controlled by Stalin from start to finish. Although the original quotas called for 70,000 executions, almost 700,000 executions took place between July 1937 and November 1938. Although various theories exist as to why Stalin ordered these mass executions, it appears that he wished to create a new generation of party leaders to replace the 'Old Bolsheviks' who had lost their revolutionary fervour, and also wished to rid the Soviet Union of 'socially harmful' classes, who could have formed fifth columns during the Second World War.

After the Great Terror, Stalin used 'lesser terror', whereby he imposed criminal penalties on huge numbers of ordinary people for workplace violations, such as theft of state or

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

collective property, no matter how petty. He continued, however, 'greater terror' with massive deportations of national groups in the 'mobilization years' of 1940–6. Although many expected a relaxation of terror at the end of the war, in fact the Gulag's population more than doubled. Stalin's lesser terror decrees remained in effect until his death in March of 1953, although many of them had fallen into disuse.

Stalin as dictator

From approximately 1932 until his death Stalin was a true dictator: he had his way on every matter and was not afraid to abuse and humiliate his associates (Rees, 2004; Gorlizki and Khlevnyuk, 2004). As Khlevnyuk (2001a, p. 325, emphasis added), concluded, 'Stalin himself was not merely a symbol of the regime but the leading figure who made the principal decisions and initiated *all* state actions of *any* significance.' After 1930, Stalin increasingly bypassed formal procedures as reflected in the declining frequency of Politburo meetings (Rees and Watson, 1997, p. 12) and the use of ad hoc subcommittees that he personally appointed (Wheatcroft 2004, p. 91). Stalin continued to bind his associates into complicity by requiring each Politburo member to approve his decisions once he had made them (Gorlizki and Khlevnyuk, 2004).

Despite his dominance, Stalin faced massive principal–agent problems with his associates. His correspondence is full of concern about 'paper fulfilment' and of angry calls for monitoring fulfilment and increased responsibility for designated officials (Gregory, 2004, pp. 165, 266).

The erosion of collective rule is consistent with Hayek's (1944) insight that the rise of a sole dictator is inevitable in such an environment. Stalin's ascendancy is explained by the need for a tie-breaker as the Politburo members quarrelled, but, more importantly, by the fact that Stalin was more ambitious, brutal and controlled than his rivals. There were no 'moderates' or 'extremists' within the Politburo after 1930. Stalin did not have to confront major ideological or policy differences. The divisions that did exist were on lines of narrow self-interest based on departmental position.

Stalin made top-level appointments personally, was deeply suspicious of professional administrators and technocrats, and trusted only a few old Bolsheviks. Stalin was particularly concerned about rent seeking by those within his narrow circle who represented branch or territorial interests, 'who cause us to deceive each other' (Khlevnyuk et al., 2001, p. 80), and 'who turn our Bolshevik party into a conglomerate of branch groups' (Rees and Watson, 1997, p. 16).

Planning under the Stalinist system was quite different from its textbook description. The state planning agency, Gosplan, prepared only highly aggregated plans, stating: 'Gosplan is not a supply organization and cannot take responsibility either for centralized specification of orders by product type or by customer or the regional distribution of products' (cited in Gregory, 2004, p. 139). Gosplan refused to plan actual transactions, labelling them 'syndicate work' (Belova and Gregory, 2002, p. 271). Gosplan only reluctantly represented the state in inter-ministerial conflicts, claiming that 'we are simply not equipped to deal with such matters' (Belova and Gregory, 2002, p. 271). In short, after its 1929 purge and subsequent politicization, Gosplan limited its exposure by doing as little as possible. The ultimate power to direct resources belonged to the dictator. Stalin did not want a planning board with immense powers or numerous staff. He did not trust information from those accountable for results, a phenomenon Wintrobe (1998) labelled the 'dictator's curse'. Truth-telling was the specialized task of Gosplan and other agencies such as the NKVD, which became Stalin's solution to the wider principal–agent problem (Belova and Gregory, 2002, pp. 269–73).

Hayek's (1944, p. 82) assertion that a totalitarian system 'cannot tie itself down in advance to general and formal rules that prevent arbitrariness… It must constantly decide questions which cannot be answered by formal principles only' was true of Stalinist practice. There were few formal rules; the rules that existed were subject to override. Fresh guidelines were issued to plan each new year or quarter, rather than general planning rules being carried over. Ministries operated without corporate-governance charters (Gregory and Markevich, 2002, pp. 793–4). 'Administrative' enforcement was encouraged through appeals to vertical superiors (Belova, 2001). Planning procedures were complicated, contradictory, and confusing (Markevich, 2003). Enterprise usually received a few output and assortment assignments midway through the plan period, while secondary targets for costs and productivity were worked out retrospectively for reporting purposes. All plans, labelled 'draft' or 'preliminary', were no more than informal agreements which could be changed subsequently by virtually any superior. The 'correcting' and 'finalizing' of plans was a never-ending process; the 'final' plan remained always on the horizon (Markevich, 2003).

Resources were allocated in the course of the 'battle for the plan' during which superiors were barraged by requests to intervene. Intervention was generally arbitrary but was based on some implicit rules of thumb, such as the priority of heavy industry and the military: 'All orders for the Ministry of Defense must be fulfilled exactly according to the schedule not allowing any delays' (Gregory, 2004, pp. 160–1). Plan interventions created havoc for producers. The most important industrial leader of the 1930s expressed his frustration as follows: 'They give us every day decree upon decree; each one is stronger and without foundation' (Khlevnyuk, 1993, p. 32).

Stalin himself favoured such ad hoc administrative allocation: 'Only bureaucrats can think that planning work ends with the creation of the plan. *The creation of the plan is only the beginning*. The real direction of the plan develops only after putting the plan together' (Stalin, 1937, p. 413; emphasis added). Ad hoc allocation could not have been better designed for the exercise of political influence. Everything was tentative and subject to arbitrary change by someone higher up in the chain of command. Savvy politicians like Stalin would have been able to weigh the political benefits of satisfying an influential regional or industrial leader.

Stalin's unwillingness to bind himself in advance to rules cascaded down through the political system, preventing the emergence of a 'law-governed' economy. The commitment

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

failure outlined by Kornai, Maskin and Roland (2003) – the predatory principle forcing agents to break formal rules so as to exploit them – seems to fit the Stalin model. In the Soviet case producers could break rules citing the threat to production from the rule, while superiors reserved the right to punish hapless scapegoats for breaking the same rules.

It was the power to live outside formal rules that sentenced Stalin and his Politburo to lives of toil, drudgery and tedium (Gregory, 2004, pp. 68–72). Threats of resignation and pleas for lengthy vacations were commonplace. A representative Politburo meeting, held on 5 March 1932, had 69 participants and 171 points on its agenda (Khlevnyuk et al., 1995, p. 232). The greatest burden fell on Stalin who in 1934, a typical year, spent 1,700 hours in official meetings, the equivalent of more than 200 eight-hour days (Khlevnyuk, 1996, pp. 190–1). Virtually every communication requested a decision from him.

Accumulation and consumption

Dictators could aim for economic growth (Olson, 1993; Glaeser et al., 2004) or for self-enrichment, or share rents to build loyalty and political power (Wintrobe, 1998). Stalin was clearly obsessed with accumulation (hardly a surprise given Marx's emphasis on accumulation), which was captured in the growth models of Preobrazhenskii and Fel'dman in the early Soviet period (Erlich, 1960; Spulber, 1964). At the core of Stalin's strategy to ‘build socialism’ were massive programmes for the hydroelectric dams, machinery complexes, vehicle works, blast furnaces, railways, and canals that were included on its itemized ‘title lists’ of approved projects.

Although Politburo meetings in the 1930s left few formal minutes, records reveal that the Politburo consistently set the *nominal* investment budget, grain collections, and foreign exchange, three variables related to investment (Gregory, 2004, chs 4 and 5). The investment budget allotted ‘investment rubles’ to industrial and regional agencies for construction and machinery, although no one appeared to know the real investment that resulted. Grain collections were designed to contribute to a budget surplus through the excess of state sale prices over purchase prices. Stalin personally directed foreign exchange to foreign capital goods rather than the luxury goods sometimes demanded by the Bolshevik elite.

If Stalin's goal was indeed to maximize investment, two facts are, at first glance, confusing. First, Stalin was extremely concerned about consumption. In Stalin's words, the ‘provisioning of workers’ was one of ‘the most contested issues’ before the Politburo, and trade was ‘the most complicated ministry’ (Gregory, 2004, pp. 93–4). Stalin personally ordered consumer goods to cities where labour productivity was declining (Gregory, 2004, ch. 4). The Politburo decided retail trade plans, prices, assortment, and even the opening of new stores. The second confusing fact is that Stalin reduced the nominal investment budget on two occasions, in 1933 and 1937 (Davies, 2001). Such evidence can be interpreted either as unstable dictatorial preference or as a consistent rule of thumb to decide the volume of investment. Stalin's capacity for calculation, patience, and self-control suggest the stable preferences approach is correct.

Figure 1 illustrates the model that Stalin and his Politburo used to set investment and consumption. The figure captures the Marxian concept of the surplus product, the gap between output and consumption, as the outcome of a distributive struggle. The model has theoretical precursors in Schrettl (1982; 1984), and is set out more fully by Gregory (2004, ch. 4). It belongs to a general class of models in which a ruler's freedom of action is circumscribed by social ‘tolerance limits’ (Kornai, 1980, pp. 211–14) or a revolution or disorder constraint (Acemoglu and Robinson, 2000). If the Politburo increased investment too much, it risked provoking the workers to provide less effort or even rebel.

Figure 1
The investment maximization model

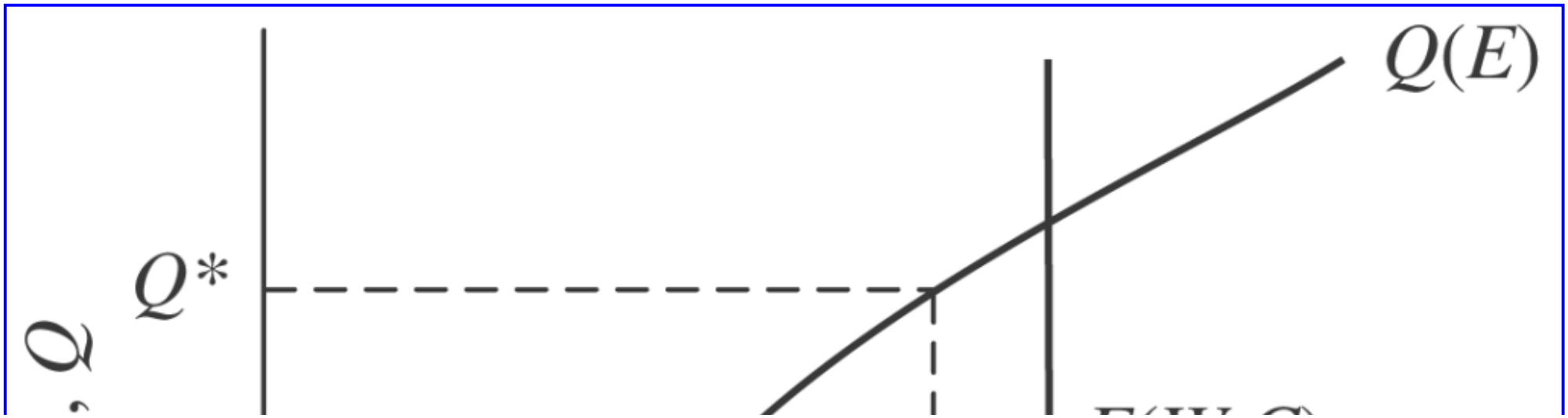

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

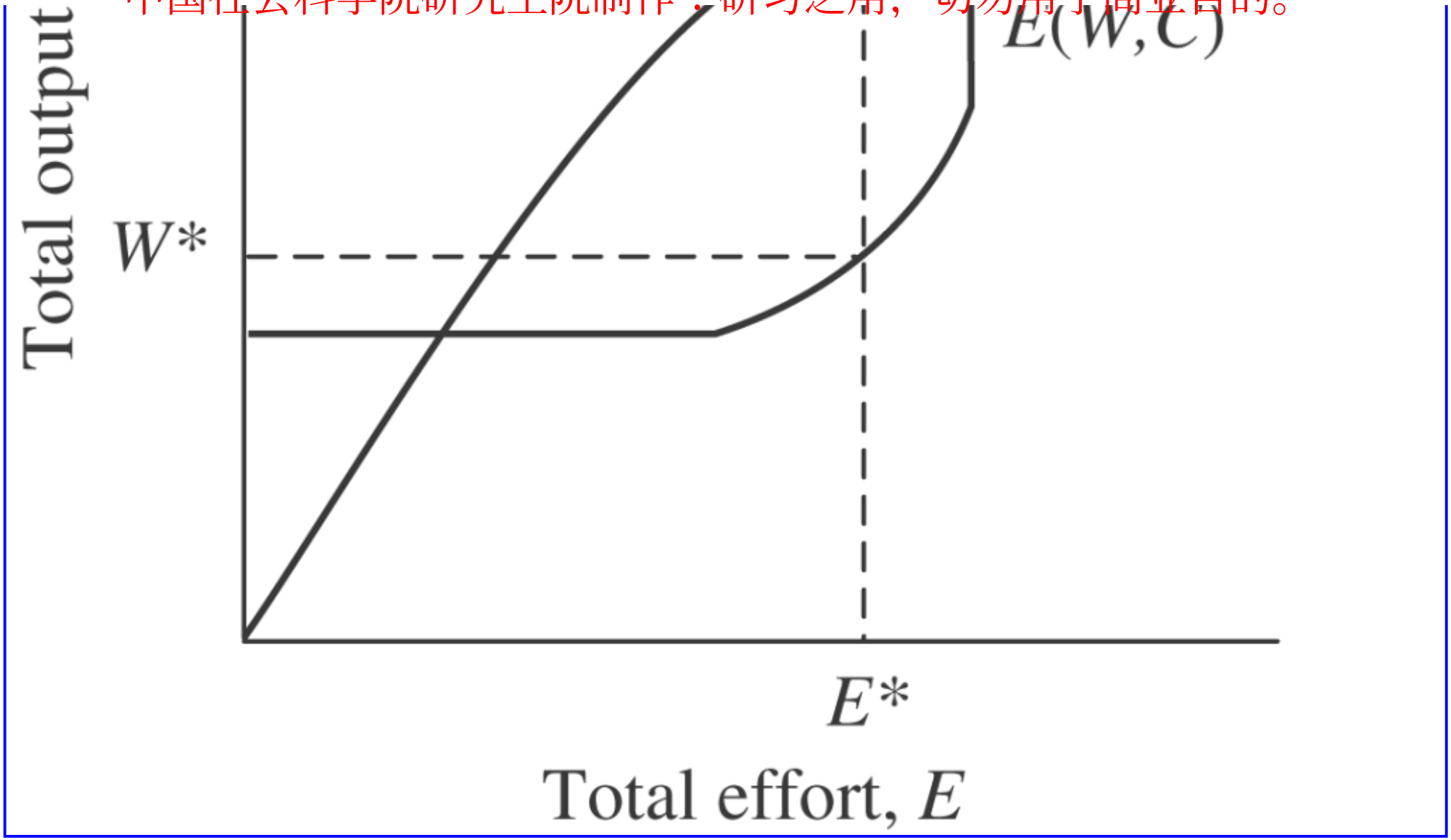

In the Stalinist system, the demand for labour was always enough for full employment and all able-bodied persons were required by law to work (Granick, 1987); thus, employment, *N*, was fixed exogenously; individual effort, *e*, was variable, so total effort, $E = e \cdot \bar{N}$, was variable although employment was not. Total output, *Q*, depended on total effort, *E*. Total effort varied with the real wage, *w*, as follows. The aggregate wage bill *W*, the consumer goods received by workers, is measured along the vertical axis in the same units as output, and is proportional to the real wage given that employment is fixed, that is, $W = w \cdot \bar{N}$. There is a reservation wage, analogous to a tolerance limit or disorder constraint, below which effort is zero; there is also a 'fair' wage at which effort is maximized. This effort curve bears some resemblance to Akerlof (1984) as applied at the micro-level. As the economy moves from the fair wage to the reservation wage, effort declines; at the lower limit unrest threatens to boil over into strikes and rebellion. Thus the effort curve intersects the horizontal axis at the reservation wage and becomes vertical at the fair wage. Effort also depends on the level of repression or coercion, *C*, discussed below. To maximize effort, the dictator would pay the fair wage and get the maximum output, but this would not maximize the surplus. To maximize the surplus, $Q - W$, he would choose the intermediate wage, effort, and output levels denoted W^*, E^*, and Q^*.

An effort curve of this shape makes the consequences of plan mistakes asymmetric: paying the workers too little could be much worse than paying them too much. Paying workers too little not only cuts effort but also risks outright confrontation with the state. An investment-maximizing dictator must tread a fine line between the pursuit of investment and the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
triggering of serious disorder.
Stalin managed worker morale and effort in two ways. When investment and consumption were about right in the aggregate, plan mistakes were taken care of by reallocating consumer goods to those left short. In the case of aggregate mistakes, when too much investment threatened to provoke the workers, investment was cut back, such as in 1933 and 1937. In this sense, Stalin's behaviour was stable and consistent, given the constraints that he perceived.
The fair wage was set by a mass psychology that was unpredictable and hard to manipulate. If workers concluded from the propaganda of economic successes that they were being cheated, the fair wage would rise, forcing the Politburo to cut investment back. Stalin used the vast informant network of the NKVD to monitor protests, strikes, anti-Soviet statements, and factory-wall graffiti, and eavesdrop to gauge mass opinion (Berelovich, 2000). Stalin had obvious political motives to do this, but within our framework wages and fairness lay at the cross-hairs of politics and economics.
Figure 1 suggests other options. Stalin and his closest subordinates could seek to manipulate the effort curve by offering ideological rewards in place of material payoffs. The attempt to transform *homo economicus* into *homo sovieticus* led, however, to a vicious circle of wage equalization and declining productivity (Kuromiya, 1988; Davies, 1996). The Stakhanovite movement, the most publicized mobilization campaign of the 1930s, was driven by progressive piece rates that permitted participating workers to drive up their incomes by overfulfilling norms (Davies and Khlevnyuk, 2002). Stalin abandoned it because it tended to raise fair-wage aspirations among non-participating workers, and also threatened inflation (Gregory, 2004, pp. 104–6). Stalin also saw targeted rationing as a way to force accumulation without a loss of effort of high-priority workers: ‘He who does not work *on industrialization* shall not eat’ (cited in Gregory, 2004, p. 98; emphasis added). Finely targeted rationing, however, required a massive bureaucracy and proved to be a blunt instrument (Davies and Khlevnyuk, 1999).

Coercion and accumulation

The dictator could also use coercion to cause workers to lower their reservation wage without reducing effort. As long as coercion displaces the effort curve in Figure 1 downward while leaving the production curve undisturbed, the surplus is increased and coercion is ‘successful’. Stalin conducted three notable experiments with coercion to foster accumulation: the forced collectivization of the peasantry, the criminalization of workplace behaviour, and the use of forced labour.
Politically, collectivization aimed to impose Soviet power in the countryside and eliminate the stratum of richer peasants. Economically, collectivization was to give the state agricultural products at low state-dictated delivery prices, which could be sold domestically and abroad for a profit. In a word, collectivization's aim was to lower peasant living standards while controlling their effort administratively. Collectivization was triggered by the peasants’ perceived unwillingness to contribute sufficiently to investment-led industrialization (Wheatcroft and Davies, 2002; Davies and Wheatcroft, 2004). The collective farms enabled Moscow to replace local decision making with central plans on sown acreage and obligatory deliveries. Acreage expanded but yields collapsed, while the share delivered to the state increased. Excessive procurements, bad weather, and plan errors combined to strip the countryside of grain; first the livestock were slaughtered, then the farmers themselves starved, threatened by severe punishments including death for theft of agricultural products. Davies and Wheatcroft (2004) dispel Conquest's (1987) notion that Stalin manufactured the famine of 1932/33 to kill class enemies; rather they show an inept leadership subsequently trying to ameliorate the effects of its own bungling.
In 1932/33, Stalin intentionally directed food to those able to work in the fields and denied it to those already hospitalized by hunger (Davies and Wheatcroft, 2004, ch. 13). Ellman (2000) has also applied the entitlement theory of Sen (1983) to the 1946/7 famine and shows that the role of the state was essentially negative: it again selected those who died by denying them entitlements. Thus, concentrating grain stocks in the hands of the Soviet state actually increased the number of deaths. The 5.5 million–6.5 million famine deaths in 1932/33 far exceeded deaths in pre-revolutionary famines (Davies and Wheatcroft, 2004, pp. 402–3).
A variety of studies, including Millar (1974), conclude that collectivization did not increase the ‘agricultural surplus’ defined as the gap between the value of output produced in agriculture and output consumed in agriculture, due to the need to shift investment resources into agriculture to make up for the loss of animal draft power slaughtered during collectivization.
As the 1940s began, Stalin redirected coercion from specific class enemies to the entire public-sector work force. A battery of intimidating laws criminalized workplace violations which had previously been punished by administrative sanctions within the enterprise. The law of 26 June 1940 (Kozlov, 2004, vol. 1) made absenteeism, defined as any 20•minutes’ unauthorized absence or even idling on the job, a criminal offence, punishable by up to six months’ corrective labour with a 25 per cent reduction in pay. Repeat offences counted as unauthorized quitting, punishable by two to four months’ imprisonment. Enterprise managers were made criminally liable for failure to report worker violations. In August 1940 the minimum sentence for petty theft at work and ‘hooliganism’ was set at one year's imprisonment. The notorious decree of June 1947 raised the minimum sentence for any theft of state or socialized property to five and seven years imprisonment. These punitive laws remained on the books until Stalin died.
A report prepared as background for Khrushchev's secret de-Stalinization speech of February 1956 (Kozlov, 2004, vol. 1, statistical appendix) shows that from 1940 through June 1955 a total of 35.8 million persons were sentenced for criminal offences. With repeat offenders not allowed for, this would represent about one-third of the adult population of roughly 100 million. Of the 35.8 million, 15.1 million were imprisoned and a quarter of a million were executed. These laws placed cumulative totals of millions of people in prisons

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and camps outside of the 'civilian' workforce. The goal of lesser terror was to increase the effort for the same real wage by criminally punishing work indiscipline. Unlike collectivization, there have been no decisive evaluations of the success or failure of these draconian policies. What is known is that they fell into misuse and were formally removed after Stalin's death. If they had been successful, they would have been retained.

The Gulag

Collectivization, the Great Terror, the repression of state employees in the 1940s, and the arrests of 'national contingents' created huge flows into the Gulag, the interior ministry's chief administration of labour camps, created in 1930 to manage camps which at their peak housed more than 2.5 million inmates. Similar numbers of deportees were confined to labour settlements in the remote interior. The forced labourers were engaged in forestry, mining and construction, where they made up substantial shares of employment, but never more than about three per cent of the total workforce including farm workers (Khlevnyuk, 2001; 2003). The cumulative total of persons sentenced to the Gulag in the course of its existence, probably in excess of 20 million, remains the subject of debate. The Gulag's own central catalogs are inconsistent (Kozlov, 2004, vol. 2); it appears that even the Gulag did not know the correct number.

Political repression strategy (collectivization, terror, war) rather than economics dictated the Gulag's development, but, once created, the Gulag represented a tempting source of cheap labour. The Gulag's consistent economic raison d'être was to explore and colonize regions that were resource-rich but inhospitable, since forced labour could be ordered around the country at will (Khlevnyuk, 2003). Subsistence wages combined with the enforcement of effort through close supervision were supposed to promote the low-cost accumulation depicted in Figure 2.

Figure 2
'Failed' coercion?

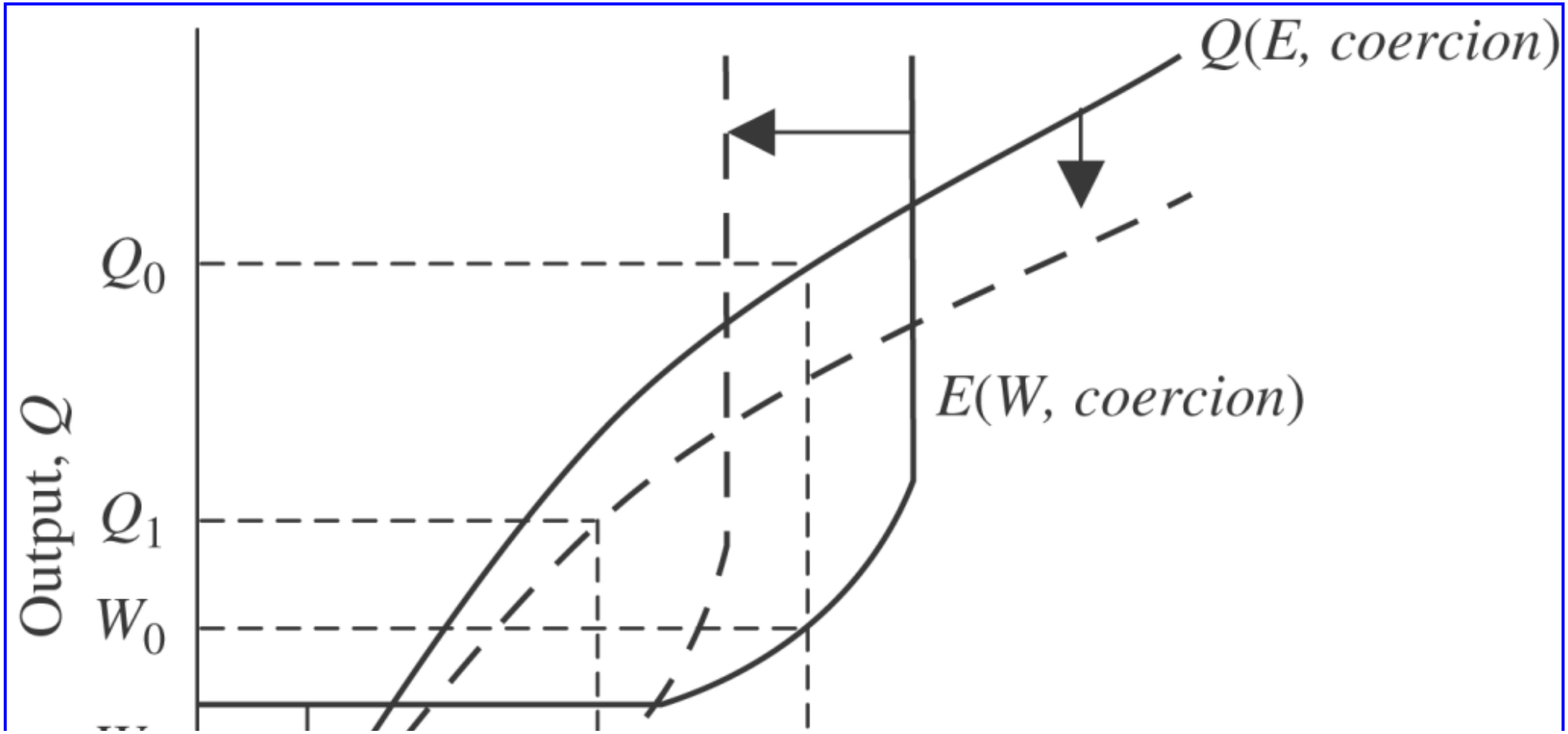

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

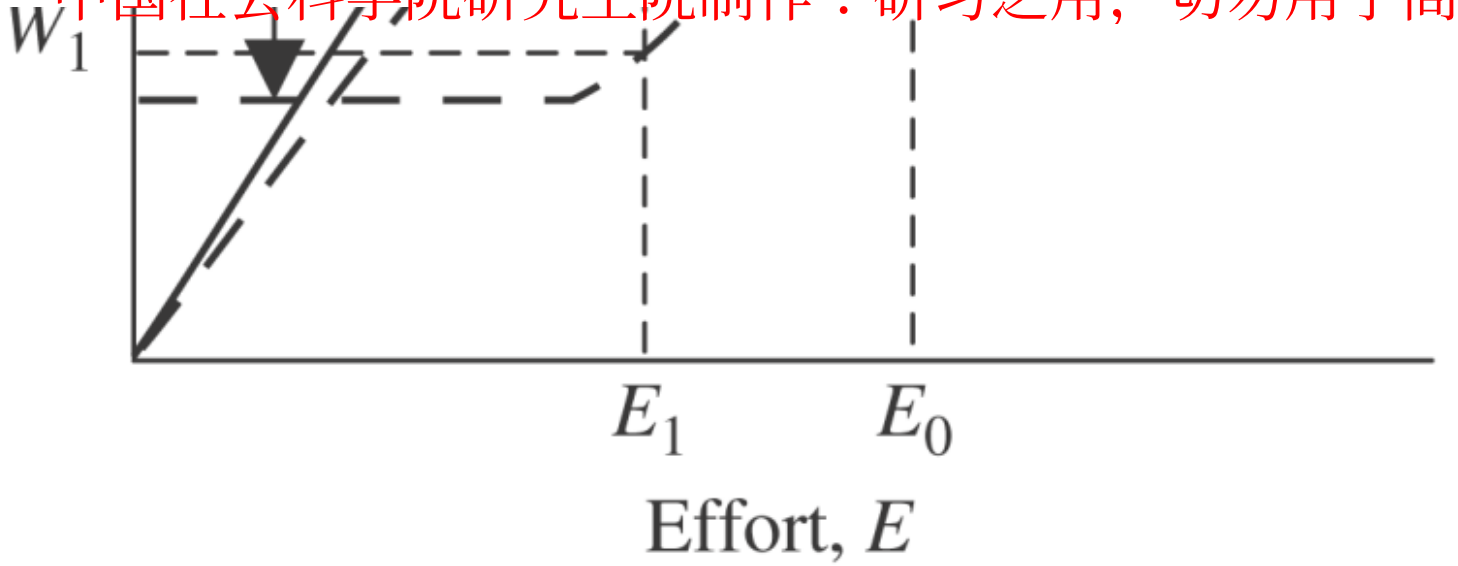

The primacy of politics over economics is shown by the fact that the NKVD and its successor the MVD (the Ministry for Internal Affairs) did not lobby for expansion. The NKVD projected a shrinking number of inmates for the third five-year plan (1938–42), just as the first victims of Stalin's Great Terror began to flood in (Gregory, 2003, p. 4). In the late 1940s Gulag officials proposed to release all but the most dangerous prisoners from camps (Tikhonov, 2003), but this was unacceptable to Stalin. In 1953, within three months of Stalin's death, MVD chief Lavrenty Beriia released one and a half million prisoners, 60 per cent of the Gulag's inmates according to a plan prepared five years earlier. The MVD was increasingly alarmed by the Gulag's economic and social costs. Its economic costs were reflected in growing financial deficits; the social costs were measured by high rates of recidivism. Although the camps were supposed to segregate hardened criminals from youth offenders, the camps were mixing bowls and the high turnover spread the culture and mores of camp life throughout society.

Hopes for huge profits from the Gulag were quickly dashed. In the Far Eastern camps (Nordlander, 2003) early optimism about huge surpluses in gold mining was replaced by pessimism as output per inmate fell precipitiously. The fact the White Sea–Baltic Canal (Morukov, 2003) was finished on time and on budget stimulated high expectations, until its major construction flaws became apparent. The Gulag leaders underestimated the risks of building Noril'sk (Borodkin and Ertz, 2003) and exposed the illusions that inmates could be coerced into supplying effort without economic rewards (Ertz, 2003).

By the post-war years, Gulag officials had concluded that the camps were operating at a loss. Labour productivity was much lower than that of free workers, while guarding detainees was very expensive; in 1950 there was one guard to ten inmates, leading to the widespread practice of ‘unguarded’ prison contingents. Prisoners formed protective networks and actually operated a number of camps (Heinzen, 2004). The arsenal of punishments was not sufficient to motivate prisoners, and trade-offs were complicated: prisoners placed on reduced rations for failing to meet work quotas could not work effectively. The most effective incentive systems, such as early release for exemplary work, deprived the Gulag of its best workers. Material incentives played an ever larger role in motivating penal labour (Borodkin and Ertz, 2003; 2005; Ertz, 2005). In the last years of the Gulag, prisoners were paid civilian wages (albeit at lower scales) and the distinctions between penal and free labour became blurred.

Evaluating repression

Effective coercion requires that penalties be accurately assessed and targeted, and that the agents of repression be well informed about offenders and the costs of their crimes. The relationship between true effort and punishment was ‘noisy’, and oppressive law could do little more than ensure that workers or inmates were physically at work and did not steal too much. Agricultural controllers could order the collective farms to sow more land but could not assess whether the land was being farmed efficiently (Davies and Wheatcroft, 2004). In industry, attempts to pin ‘normal’ effort down to objective technological criteria proved fruitless (Davies and Khlevnyuk, 2002, p. 877; Filtzer, 2002, pp. 232–41).

The investigation of low effort could yield an error of Type I that punished the innocent, or a Type II error that acquitted the guilty. Type I errors are reflected in the high rates of penalization that condemned hard workers along with neer-do-wells, drunks and thieves. Virtually every worker became liable to prosecution for some offence, including one-time and accidental violations. Rational managers might wish to prosecute only problem workers and repeat offenders, but the laws penalized even petty offences, and managers who failed to report offences were threatened with the same. As a result, the innocent were bundled along with the guilty in extraordinarily large numbers.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In March 1937, Nikolai Ezhov, Stalin's chief instrument of the Great Terror, told his officials to expect 'some innocent victims … Better that ten innocent people should suffer than one spy get away. When you chop wood, chips fly' (Montefiore, 2003, p. 194). The attitude of Vyacheslav Molotov, Stalin's prime minister (interviewed by Chuev, 1991, p. 416) was, 'never mind if extra heads fall', even when one of those 'heads' was that of his own wife. Type II errors were evidenced by the fact that, although penalization rates were very high, offending rates were even higher (Filtzer, 2002, pp. 167–8). A judicial system that was supposed to 'make the chips fly' somehow failed to chop the wood. The combination of severe penalization and low conviction probability for the guilty is consistent with high-cost policing and justice administration (Becker, 1968, p. 184); the high rate of conviction of innocents, however, is a cost of the dictator's efforts to achieve a lower rate of offending than society was willing to tolerate (Djankov et al., 2003). Although the Gulag did not generate an *internal* surplus, it could have displaced the effort curve of civilian workers if they expected the Gulag wage as their punishment for low effort. But if workers expect Type I errors, they will be punished regardless of effort; if they expect to benefit from Type II errors, they can shirk without fear.

Error rates were not exogenous. They were fashioned by the counteractions of those threatened, who could take steps to reduce their risks. Workers and managers diverted effort from production into mutual insurance: since the threat was shared among them, they could agree to cover up each other's shortcomings. Post-war managers tolerated lateness and absence to maintain goodwill, and underreported such violations, while pursuing quitters who undermined morale and the factory's capacity to fulfil the plan (Filtzer, 2002). The rural police and courts pooled risks with the rural community in sheltering the young offenders who had deserted factories or technical schools (Kozlov, 2004, vol. I). Mutual insurance tended to cut the individual risk of punishment. Regional party officials defied even the most powerful central organizations to protect their own (Harris, 1999, pp. 156–63). High-level patrons could protect the most egregious embezzlers (Belova, 2001).

Figure 2 explains the phenomenon of 'failed coercion'. The threat of punishment raised the 'noise' of the system as managers and regional officials reported false results and formed mutual protection networks, shifting the output–effort curve down. The leftward movement of the effort curve captures the diversion of activity from production to the avoidance of repression. Failed coercion reduces the surplus and lowers output.

Faced with widespread enforcement failures at lower levels, Stalin forced the legal system, local party offices, and the militia to increase arrest and conviction rates or suffer penalties themselves. The most common method of forcing repression was to distribute quotas by region and profession to officials at lower levels (Kozlov, 2004, vol. 1). In the Great Terror of 1937/38 local officials had to work feverishly to achieve a set number of confessions per day (Vatlin, 2004). To fulfil such plans the police officials imputed individual guilt from increasingly trivial differences in behaviour. Whether or not these measures reduced the Type II errors, they seem likely to have encouraged false denunciation and confession and so added to the errors of Type I.

After Stalin

Stalin's successors reverted to collective leadership and toned down repression. They also inherited an economy in secular decline. Thus, the story of the post-Stalin leadership is that of unsuccessful attempts to reform the Stalinist system without altering its basic characteristics, other than mass repression.

See Also

- Akerlof, George Arthur
- inflation
- Preobrazhensky, Evgenii Alexeyevich
- rationing
- rent seeking
- Soviet Union, economics in
- Stakhanovism

Bibliography

Acemoglu, D. and Robinson, J. 2000. Political losers as barriers to economic development. *American Economic Review* 90(2), Papers and Proceedings, 126–30.

Akerlof, G. 1984. Gift exchange and efficiency wages: four views. *American Economic Review* 74(2), Papers and Proceedings, 79–83.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Becker, G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76, 169–217.

Belova, E. 2001. Economic crime and punishment. In *Behind the Façade of Stalin's Command Economy: Evidence from the State and Party Archives*, ed. P. Gregory. Stanford, CA: Hoover Institution Press.

Belova, E. and Gregory, P. 2002. Dictator, loyal and opportunistic agents: the Soviet archives on creating the Soviet economic system. *Public Choice* 113, 265–86.

Berelovich [Berelowitch], A. 2000. *Sovetskaia Derevnia Glazami VChK–OGPU–NKVD, 2: 1923–1929*. Moscow: Rosspen.

Borodkin, L. and Ertz, S. 2003. Coercion versus motivation: forced labor in Norilsk. In *The Economics of Forced Labor: The Soviet Gulag*, ed. P. Gregory and V. Lazarev. Stanford, CA: Hoover Institution Press.

Borodkin, L. and Ertz, S. 2005. Forced labor and the need for motivation: wages and bonuses in the Stalinist camp system. *Comparative Economic Studies* 47, 418–35.

Chuev, F. 1991. *Sto sorok besed s Molotovym*. Moscow: Terra.

Conquest, R. 1987. *The Harvest of Sorrow: Soviet Collectivization and the Terror-Famine*. Oxford: Oxford University Press.

Davies, R. 1996. *The Industrialisation of Soviet Russia, 4: Crisis and Progress in the Soviet Economy, 1931–1933*. Basingstoke: Macmillan.

Davies, R. 2001. Why was there a Soviet investment cycle in 1933–7? PERSA Working Paper No. 16. Department of Economics, University of Warwick. Online. Available at http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/harrison/archive/persa/016fulltext.pdf, accessed 12 September 2006.

Davies, R. and Khlevnyuk, O. 1999. The end of rationing in the Soviet Union, 1934–1935. *Europe–Asia Studies* 51, 557–610.

Davies, R. and Khlevnyuk, O. 2002. Stakhanovism and the Soviet economy. *Europe–Asia Studies* 54, 897–903.

Davies, R. and Wheatcroft, S. 2004. *The Industrialisation of Soviet Russia, 5: The Years of Hunger: Soviet Agriculture, 1931–1933*. Basingstoke: Palgrave.

Djankov, S., Glaeser, E., La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. 2003. The new comparative economics. *Journal of Comparative Economics* 31, 595–619.

Erlich, A. 1960. *The Soviet Industrialization Debate, 1924–1928*. Cambridge, MA: Harvard University Press.

Ellman, M. 2000. The 1947 Soviet famine and the entitlement approach to famines. *Cambridge Journal of Economics* 24, 603–30.

Ertz, S. 2003. Building Norilsk. In *The Economics of Forced Labor: The Soviet Gulag*, ed. P. Gregory and V. Lazarev. Stanford, CA: Hoover Institution Press.

Ertz, S. 2005. Trading effort for freedom: workday credits in the Soviet camp system. *Comparative Economic Studies* 47, 476–90.

Filtzer, D. 2002. *Soviet Workers and Late Stalinism: Labour and the Restoration of the Stalinist System after World War II*. Cambridge: Cambridge University Press.

Glaeser, E., La Porta, R., Lopez-de-Silane, F. and Shleifer, A. 2004. Do institutions cause growth? *Journal of Economic Growth* 9, 271–303.

Gorlizki, Y. and Khlevnyuk, O. 2004. *Cold Peace: Stalin and the Soviet Ruling Circle, 1945–1953*. New York: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Granick, D. 1987. *Job Rights in the Soviet Union: Their Consequences*. New York: Cambridge University Press.

Gregory, P. 2003. An introduction to the economics of the Gulag. In *The Economics of Forced Labor: The Soviet Gulag*, ed. P. Gregory and V. Lazarev. Stanford, CA: Hoover Institution Press.

Gregory, P. 2004. *The Political Economy of Stalinism: Evidence From the Soviet Secret Archives*. New York: Cambridge.

Gregory, P. and Markevich, A. 2002. Creating Soviet industry: the house that Stalin built. *Slavic Review* 61, 4787–814.

Harris, J. 1999. *The Great Urals: Regionalism and the Evolution of the Soviet System*. Ithaca, NY: Cornell University Press.

Hayek, F. 1944. *The Road to Serfdom*. Chicago: University of Chicago Press.

Heinzen, J. 2004. Corruption in the Gulag: dilemmas of officials and prisoners. *Comparative Economic Studies* 47, 2456–75.

Khlevnyuk, O. 1993. *Stalin i Ordzhonikidze: konflikty v Politburo v 30–e gody*. Moscow: Rossiia molodaia.

Khlevnyuk, O. 1996. *Politburo. Mekhanizmy politchecheskoi vlasti v 1930-e gody*. Moscow: ROSSPEN.

Khlevnyuk, O. 2001a. Stalinism and the Stalin period after the ‘Archival Revolution’. *Kritika* 2, 319–28.

Khlevnyuk, O. 2001b. The economy of the Gulag. In *Behind the Façade of Stalin's Command Economy: Evidence from the State and Party Archives*, ed. P. Gregory. Stanford, CA: Hoover Institution Press.

Khlevnyuk, O. 2003. The economy of the OGPU, NKVD, and MVD of the USSR, 1930–1953: the scale, structure, and trends of development. In *The Economics of Forced Labor: The Soviet Gulag*, ed. P. Gregory and V. Lazarev. Stanford, CA: Hoover Institution Press.

Khlevnyuk, O. and Devis, R.U. [Davies, R.W.], Kosheleva, L., Ris, E. [Rees, E.] and Rogovaia, L., eds. 2001. *Stalin i Kaganovich. Perepiska. 1931–1936 gg*. Moscow: ROSSPEN.

Khlevnyuk, O., Kvashonkin, A., Kosheleva, L. and Rogovaia, L. eds. 1995. *Stalinskoe Politbiuro v 30-e gody. Sbornik dokumentov*. Moscow: AIRO-XX.

Kornai, J. 1980. *The Economics of Shortage*, 2 vols. Amsterdam: North-Holland.

Kornai, J., Maskin, E. and Roland, G. 2003. Understanding the soft budget constraint. *Journal of Economic Literature* 41, 1095–136.

Kozlov, V., ed. 2004. *Istoriia Stalinskogo Gulaga*, 6 vols. Moscow: ROSSPEN.

Kuromiya, H. 1988. *Stalin's Industrial Revolution: Politics and Workers, 1928–1932*. Cambridge: Cambridge University Press.

Markevich, A. 2003. Was the Soviet Economy planned? Planning in the People's Commissariats in the 1930s. PERSA Working Paper No. 9. Department of Economics, University of Warwick.

Millar, J. 1974. Mass collectivization and the contribution of Soviet agriculture to the first Five-Year Plan: a review article. *Slavic Review* 33, 4750–66.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Montefiore, S. 2003. *Stalin: The Court of The Red Tsar*. London: Weidenfeld and Nicolson.

Morukov, M. 2003. The White Sea–Baltic Canal. In *The Economics of Forced Labor: The Soviet Gulag*, ed. P. Gregory and V. Lazarev. Stanford, CA: Hoover Institution Press.

Nordlander, D. 2003. Magadan and the economic history of Dalstroi in the 1930s. In *The Economics of Forced Labor: The Soviet Gulag*, ed. P. Gregory and V. Lazarev. Stanford, CA: Hoover Institution Press.

Olson, M. 1993. Dictatorship, democracy, and development. *American Political Science Review* 87, 3567–76.

Rees, E., ed. 2004. *The Nature of Stalin's Dictatorship: The Politburo, 1924–1953*. Basingstoke: Palgrave.

Rees, E. and Watson, D. 1997. Politburo and Sovnarkom. In *Decision-Making in the Stalinist Command Economy, 1932–37*, ed. E. Rees. Basingstoke and London: Macmillan.

Schrettl, W. 1982. *Consumption, Effort, and Growth in Soviet-Type Economies: A Theoretical Analysis*. Ann Arbor, MI: University Microfilms International.

Schrettl, W. 1984. Anspruchsdenken, Leistungsbereitschaft und Wirtschaftszyklen. In *Wachstumsverlangsamung und Konjunkturzyklen in unterschiedlichen Wirtschaftssystemen*, ed. A. Schuller. Berlin: Duncker and Humblot.

Sen, A. 1983. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Oxford University Press.

Spulber, N. 1964. *Soviet Strategy for Economic Growth*. Bloomington, IN: Indiana University Press.

Stalin, I. 1937. *Voprosy Leninizma*, 10th edn. Moscow: Gospolitizdat.

Tikhonov, A. 2003. The end of the Gulag. In *The Economics of Forced Labor: The Soviet Gulag*, ed. P. Gregory and V. Lazarev. Stanford, CA: Hoover Institution Press.

Vatlin, A.Iu. 2004. *Terror raionnogo masshtaba: 'massovye operatsii' NKVD v Kuntsevskom raione Moskovskoi oblasti. 1937–1938 gg*. Moscow: ROSSPEN.

Wheatcroft, S. 2004. From Team Stalin to degenerate tyranny. In *The Nature of Stalin's Dictatorship: The Politburo, 1924–1953*, ed. E. Rees. Basingstoke: Palgrave.

Wheatcroft, S. and Davies, R. 2002. The Soviet famine of 1932–3 and the crisis in agriculture. In *Challenging Traditional Views of Russian History*, ed. S. Wheatcroft. Basingstoke: Palgrave.

Wintrobe, R. 1998. *The Political Economy of Dictatorship*. Cambridge: Cambridge University Press.

How to cite this article

Gregory, Paul R. "Stalinism, political economy of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000501> doi:10.1057/9780230226203.1605

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

standards of living (historical trends)

Richard H. Steckel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social scientists generally agree on the considerable value of measuring and investigating long-term trends and socio-economic differences in the standard of living or quality of life. Over the past two centuries, researchers have refined the methodology for making such assessments but economic historians continue to be challenged by lack of raw data before the early 20th century. While it is impossible to assemble a comprehensive picture, useful information is available about changes or differences in GDP per capita and life expectancy from the early 19th century onward. It was during this era that large differences across countries and regions appeared.

Keywords

child mortality; Great Depression; index number problems; industrialization; international migration; life expectancy; longevity; new anthropometric history; nutrition; standards of living; urbanization

Article

Methodology

Research specialists agree that the standard of living has many elements, such as material goods and services, health, socio-economic fluidity, education, inequality, and political and religious freedom. Opinions differ on the precise measures to be used within each category and on the weights that should be attached to each. Health, for example, is measurable by length of life, morbidity (illness or disability) and physical fitness. Conceivably, one might attempt comparisons using all feasible measures, but this is expensive and time-consuming, and in any event good measures within categories are often highly correlated.
Weighting is a contentious issue in any attempt to summarize the standard of living, or otherwise compress diverse measures into a single number. Economists and other social scientists know that tastes are individualistic and diverse but they recognize general tendencies. Here I consider the available historical information on material standards and health.

Material aspects

Measure

The most widely used measure of the material standard of living is Gross Domestic Product (GDP) per capita, adjusted for changes in the price level (inflation or deflation). This measure reflects only economic activities that flow through markets, omitting productive endeavours unrecorded in market exchanges, such a preparing meals at home or maintenance done by the homeowner. It ignores work effort required to produce income and does not consider conditions surrounding the work environment, which might affect health and safety. Crime, pollution and congestion, which many people consider important issues affecting their quality of life, are also excluded from GDP. Moreover, technological change, relative prices and tastes affect the course of GDP and the products and services that it includes, which creates what economists call an ‘index number’ problem that is not readily solvable.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Nevertheless most economists believe that real GDP per capita does summarize or otherwise quantify important aspects of the average availability of goods and services.

Time trends

Table 1 shows the course of the material standard of living in the United States from 1820 to 1998. Over this period of 178 years real GDP per capita increased by 21.7 times, or an average of 1.73 per cent per year. Although the evidence available to estimate GDP directly is meagre, this rate of increase was probably many times higher than experienced during the colonial period. This conclusion is justified by considering the implications of extrapolating the level observed in 1820 ($1,257) backward in time at the growth rate measured since 1820 (1.73 per cent). Under this supposition, real per capita GDP would have doubled every 40 years (halved every 40 years going backward in time) and so by the mid-1700s there would have been insufficient income to support life. Because the cheapest diet able to sustain good health would have cost nearly $500 per year, the tentative assumption of modern economic growth contradicts what actually happened. Moreover, historical evidence suggests that important ingredients of modern economic growth, such as technological change and human and physical capital, accumulated relatively slowly during the colonial period.

GDP per capita in the United States, 1820–1998

Year	GDP per capita (1990 international dollars)	Annual growth rate (%) from previous period
1820	1,257	
1870	2,445	1.34
1913	5,301	1.82
1950	9,561	1.61
1973	16,689	2.45
1990	23,214	1.94
1998	27,331	2.04

Source: Maddison (2001, tables A-1c and A-1d).

Cycles

Although real GDP per capita is given for only seven dates in Table 1, it is apparent that economic progress has been uneven over time. If annual or quarterly data were given, it would show that business cycles have been a major feature of the economic landscape since industrialization began in the 1820s. By far the worst downturn in US history occurred during the Great Depression of the 1930s, when real per capita GDP declined by approximately one-third and the unemployment rate reached 25 per cent.

Regions

The aggregate numbers also disguise regional differences in the standard of living. In 1840 personal income per capita was twice as high in the Northeast as in the North Central States. Regional divergence increased after the Civil War when the South Atlantic became the nation's poorest region, attaining only one-third of the living standards in the Northeast. Regional convergence occurred in the 20th century, and industrialization in the South significantly improved the region's economic standing after the Second World War.

Health

Life expectancy

Two measures of health are widely used in economic history: life expectancy at birth (or average length of life) and average height, which measures nutritional conditions during the growing years. Table 2 shows that life expectancy has approximately doubled since the mid-19th century, reaching 76.7 years in 1998. If depressions and recessions have adversely affected the material standard of living, epidemics have been a major cause of sudden declines in health in the past. Fluctuations during the 19th century are evident from the table, but as a rule growth rates in health have been considerably less volatile than those for GDP, particularly during the 20th century.

Life expectancy at birth

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in the United States, 1850–1998

Year	Life expectancy
1850	38.3
1860	41.8
1870	44.0
1880	39.4
1890	45.2
1900	47.8
1910	53.1
1920	54.1
1930	59.7
1940	62.9
1950	68.2
1960	69.7
1970	70.8
1980	73.7
1990	75.4
1998	76.7

Source: Haines (2006).

Childhood mortality greatly affects life expectancy, which was low in the mid-1800s substantially because mortality rates were very high for this age group. For example, roughly one child in five born alive in 1850 did not survive to age one, but today the infant mortality rate is less than one per cent. The period since 1850 has witnessed a significant shift in deaths from early childhood to old age. At the same time, the major causes of death have shifted from infectious diseases originating with germs or micro-organisms to degenerative processes that are affected by lifestyle choices such as diet, smoking and exercise.

Time trends

The largest gains were concentrated in the first half of the 20th century, when life expectancy increased from 47.8 years in 1900 to 68.2 years in 1950. Factors behind the growing longevity include the ascent of the germ theory of disease, programmes of public health and personal hygiene, better medical technology, higher incomes, better diets, more education, and the emergence of health insurance.

Explanations

Numerous important medical developments contributed to improving health. The research of Pasteur and Koch was particularly influential in leading to acceptance of the germ theory in the late 1800s. Prior to their work, many diseases were thought to have arisen from miasmas or vapours created by rotting vegetation. Thus, swamps were accurately viewed as unhealthy, but not because they were home to mosquitoes and malaria. The germ theory gave public health measures a sound scientific basis, and shortly thereafter cities began cost-effective measures to remove garbage, purify water supplies, and process sewage. The notion that ‘cleanliness was next to Godliness’ also emerged in the home, where bathing and the washing of clothes, dishes and floors became routine.

The discovery of Salvarsan in 1910 was the first use of an antibiotic (for syphilis), which meant that the drug was effective in altering the course of a disease. This was an important medical event, but broad-spectrum antibiotics were not available until the middle of the century. The most famous of these early drugs was penicillin, which was not manufactured in large quantities until the 1940s. Much of the gain in life expectancy was attained before chemotherapy and a host of other medical technologies were widely available. A cornerstone of improving health from the late 1800s to the middle of the 20th century was therefore prevention of disease by reducing exposure to pathogens. Also important were improvements

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in immune systems created by better diets and by vaccination against diseases such as smallpox and diphtheria.

Heights

Since the early 1980s, historians have increasingly used average heights to assess health aspects of the standard of living. Average height is a good proxy for the nutritional status of a population because height at a particular age reflects an individual's history of *net* nutrition, or diet minus claims on the diet made by work (or physical activity) and disease. The growth of poorly nourished children may cease, and repeated bouts of biological stress – whether from food deprivation, hard work, or disease – often leads to stunting or a reduction in adult height. The average heights of children and of adults in countries around the world are highly correlated with their life expectancy at birth and with the log of the per capita GDP in the country where they live.

Applications

This interpretation for average heights has led to their use in the study of the health of slaves, health inequality, living standards during industrialization, and trends in mortality. The first important results in the ‘new anthropometric history’ dealt with the nutrition and health of Americans slaves as determined from stature recorded for identification purposes on slave manifests required in the coastwise slave trade. The subject of slave health has been a contentious issue among historians, in part because vital statistics and nutrition information were never systematically collected for slaves (or for the vast majority of the American population in the mid-19th century, for that matter). Yet the height data showed that children were astonishingly small and malnourished while working slaves were remarkably well fed. Adolescent slaves grew rapidly as teenagers and were reasonably well-off in nutritional aspects of health.

Time trends

Table 3 shows the time pattern in height of native-born American men obtained in historical periods from military muster rolls, and for men and women in recent decades from the National Health and Nutrition Examination Surveys. This historical trend is notable for the tall stature during the colonial period, the mid-19th century decline, and the surge in heights of the 20th century. Comparisons of average heights from military organizations in Europe show that Americans were taller by two to three inches. Behind this achievement were a relatively good diet, little exposure to epidemic disease, and relative equality in the distribution of wealth. Americans could choose their foods from the best of European and Western Hemisphere plants and animals, and this dietary diversity combined with favourable weather meant that Americans never had to contend with harvest failures. Thus, even the poor were reasonably well fed in colonial America.

Average height of native-born US men and women by year of birth, 1710–1970

Year	Centimeters		Inches	
	Men	Women	Men	Women
1710	171.5		67.5	
1720	171.8		67.6	
1730	172.1		67.8	
1740	172.1		67.8	
1750	172.2		67.8	
1760	172.3		67.8	
1770	172.8		68.0	
1780	173.2		68.2	
1790	172.9		68.1	
1800	172.9		68.1	
1810	173.0		68.1	

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1820	172.9		68.1	
1830	173.5		68.3	
1840	172.2		67.8	
1850	171.1		67.4	
1860	170.6		67.2	
1870	171.2		67.4	
1880	169.5		66.7	
1890	169.1		66.6	
1900	170.0		66.9	
1910	172.1		67.8	
1920	173.1		68.1	
1930	175.8	162.6	69.2	64.0
1940	176.7	163.1	69.6	64.2
1950	177.3	163.1	69.8	64.2
1960	177.9	164.2	70.0	64.6
1970	177.4	163.6	69.8	64.4

Source: Steckel (2006) and sources therein.

Cycles and explanations

Loss of stature began in the second quarter of the 19th century when the transportation revolution of canals, steamboats and railways brought people into greater contact with diseases. The rise of public schools meant that children were newly exposed to major diseases such as whooping cough, diphtheria, and scarlet fever. Food prices also rose during the 1830s and growing inequality in the distribution of income or wealth accompanied industrialization. Business depressions, which were most hazardous for the health of those who were already poor, also emerged with industrialization. The Civil War of the 1860s and its troop movements further spread disease and disrupted food production and distribution. A large volume of immigration also brought new varieties of disease to the United States at a time when urbanization brought a growing proportion of the population into closer contact with contagious diseases. Estimates of life expectancy among adults at ages 20, 30 and 50, which was assembled from family histories, also declined in the middle of the 19th century. In the 20th century, heights grew most rapidly for those born between 1910 and 1950, an era when public health and personal hygiene took vigorous hold, incomes rose rapidly and there was reduced congestion in housing. The latter part of the era also witnessed a larger share of income or wealth going to the lower portion of the distribution, implying that the incomes of the less well-off were rising relatively rapidly. Note that most of the rise in heights occurred before modern antibiotics were available, which means that disease prevention was a more significant cause of improving health than the ability to alter its course after onset. The growing control that humans have exercised over their environment, particularly increased food supply and reduced exposure to disease, may be leading to biological (but not genetic) evolution of humans with more durable vital organ systems, larger body size, and later onset of chronic diseases.

Recent stagnation

Between the middle of the 20th century and the present, however, the average heights of American men have stagnated, increasing by only a small fraction of an inch over the since the 1950s. Table 3 refers to the native born, so recent increases in immigration cannot account for the stagnation. In the absence of other information, one might be tempted to suppose that environmental conditions for growth are so good that most Americans have simply reached their genetic potential for growth. Unlike in the United States, heights and life expectancy have continued to grow in Europe, which has the same genetic stock as that from which most Americans descend. By the 1970s several American health indicators had fallen behind those in Norway, Sweden, the Netherlands and Denmark. While American heights were essentially flat after the 1970s, heights continued to grow significantly in Europe. The Dutch men are now the tallest, averaging six feet, about two inches more than American men. Lagging heights leads to questions about the adequacy of health care and lifestyle choices in America. As discussed below, it is doubtful that lack of resource commitment to health care is the problem because America invests a greater share of GDP than

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the Netherlands. Greater inequality and less access to health care could be important factors in the difference. But access to health care alone, whether due to low income or lack of insurance coverage, may not be the only issues – health insurance coverage must be used regularly and wisely. In this regard, Dutch mothers are known for regular pre- and post-natal checkups, which are important for early childhood health.

Note that significant differences in health and the quality of life follow from these height patterns. The comparisons are not part of an odd contest that emphasizes height, nor is 'big' per se assumed to be beautiful. Instead, we know that, on average, stunted growth has functional implications for longevity, cognitive development and work capacity. Children who fail to grow adequately are often sick, suffer learning impairments and have a lower quality of life. Growth failure in childhood has a long reach into adulthood because individuals whose growth has been stunted are at greater risk of death from heart disease, diabetes and some types of cancer. Therefore it is important to know why Americans are falling behind.

International perspective

Per capita GDP comparisons

Table 4 places American economic performance in perspective relative to other countries. In 1820 the United States was fifth in world ranking, roughly 30 per cent below the leaders (United Kingdom and the Netherlands), but still two to three times better off than the poorest sections of the globe. It is notable that in 1820 the richest country (the Netherlands at $1,821) was approximately 4.4 times better off than the poorest (Africa at $418), but by 1950 the ratio of richest to poorest had widened to 21.8 ($9,561 in the United States versus $439 in China), which is roughly the level it is today (in 1998, it was $27,331 in the United States versus $1,368 in Africa). These calculations understate the growing disparity in the material standard of living because several African countries today fall significantly below the average, whereas it is unlikely that they did so in 1820 because GDP for the continent as a whole was close to the level of subsistence.

GDP per capita by country and year, 1820–1998

1990 international dollars

Country	1820	1870	1913	1950	1973	1998	Ratio 1998 to 1820
Austria	1,218	1,863	3,465	3,706	11,235	18,905	15.5
Belgium	1,319	2,697	4,220	5,462	12,170	19,442	14.7
Denmark	1,274	2,003	3,912	6,946	13,945	22,123	17.4
Finland	781	1,140	2,111	4,253	11,085	18,324	23.5
France	1,230	1,876	3,485	5,270	13,123	19,558	15.9
Germany	1,058	1,821	3,648	3,881	11,966	17,799	16.8
Italy	1,117	1,499	2,564	3,502	10,643	17,759	15.9
Netherlands	1,821	2,753	4,049	5,996	13,082	20,224	11.1
Norway	1,104	1,432	2,501	5,463	11,246	23,660	21.4
Sweden	1,198	1,664	3,096	6,738	13,493	18,685	15.6
Switzerland	1,280	2,202	4,266	9,064	18,204	21,367	16.7
United Kingdom	1,707	3,191	4,921	6,907	12,022	18,714	11.0
Portugal	963	997	1,244	2,069	7,343	12,929	13.4
Spain	1,063	1,376	2,255	2,397	8,739	14,227	13.4
United States	1,257	2,445	5,301	9,561	16,689	27,331	21.7
Mexico	759	674	1,732	2,365	4,845	6,655	8.8
Japan	669	737	1,387	1,926	11,439	20,413	30.5
China	600	530	552	439	839	3,117	5.2
India	533	533	673	619	853	1,746	3.3
Africa	418	444	585	852	1,365	1,368	3.3

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

World	667	867	1,510	2,114	4,104	5,709	8.6
Ratio of richest to poorest	4.4	7.2	8.9	20.6	21.7	20.0	

Source: Maddison (2001, table B-21).

Per capita GDP growth

It is clear that the poorer countries are better off today than they were in 1820 (by 3.3 times in both Africa and India). At the simplest level, the explanation is that the countries that are now rich grew much faster after 1820. The last column of Table 4 shows that Japan realized the most spectacular gain, climbing from approximately the world average in 1820 to the fifth richest today, with more than a thirtyfold increase in real per capita GDP. All countries that are rich today had rapid increases in their material standard of living, realizing more than tenfold increases since 1820. The underlying reasons for this diversity of economic success is a central question in the field of economic history.

Life expectancy

Table 5 shows that disparities in life expectancy have been much less than those in per capita GDP. In 1820 all countries were bunched in the range of 21 to 41 years, with Germany at the top and India at the bottom, giving a ratio of less than two to one. It is doubtful that any country or region has had a life expectancy below 20 years for long periods of time because death rates would have exceeded any plausible upper limit for birth rates, leading to population implosion. The 20th century witnessed a compression in life expectancies across countries, with the ratio of levels in 1999 being 1.56 (81 in Japan versus 52 in Africa). Japan has also been a spectacular performer in health, increasing life expectancy from 34 years in 1820 to 81 years in 1999. Among poor unhealthy countries, health aspects of the standard of living have improved more rapidly than the material standard of living relative to the world average. Because many public health measures are cheap and effective, it has been easier to extend life than it has been to promote material prosperity, which has numerous complicated causes.

Life expectancy at birth by country and year, 1920–1999

Country	1820	1900	1950	1999
France	37	47	65	78
Germany	41	47	67	77
Italy	30	43	66	78
Netherlands	32	52	72	78
Spain	28	35	62	78
Sweden	39	56	70	79
United Kingdom	40	50	69	77
United States	39	47	68	77
Japan	34	44	61	81
Russia	28	32	65	67
Brazil	27	36	45	67
Mexico	n.a.	33	50	72
China	n.a.	24	41	71
India	21	24	32	60
Africa	23	24	38	52
World	26	31	49	66

Source: Maddison (2001, table 1-5a).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Height comparisons

Figure 1 compares stature in the United States and the United Kingdom. Americans were very tall by global standards in the early 19th century as a result of their rich and varied diets, low population density, and relative equality of wealth. Unlike other countries that have been studied (France, the Netherlands, Sweden, Germany, Japan and Australia), both the United States and the UK suffered significant height declines during industrialization (as defined primarily by the achievement of modern economic growth) in the 19th century. Note, however, that the amount and timing of the height decline in the UK has been the subject of a lively debate. See for example the February 1993 issue of the *Economic History Review* for papers by Roderick Floud, Kenneth Wachter and John Komlos; only the Floud–Wachter figures are given here.

Figure 1

Average height of soldiers in Britain and of native-born American soldiers, 1710–1970. *Sources*: Steckel (2006, Fig. 12) and Floud, Wachter and Gregory (1990, Table 4.8).

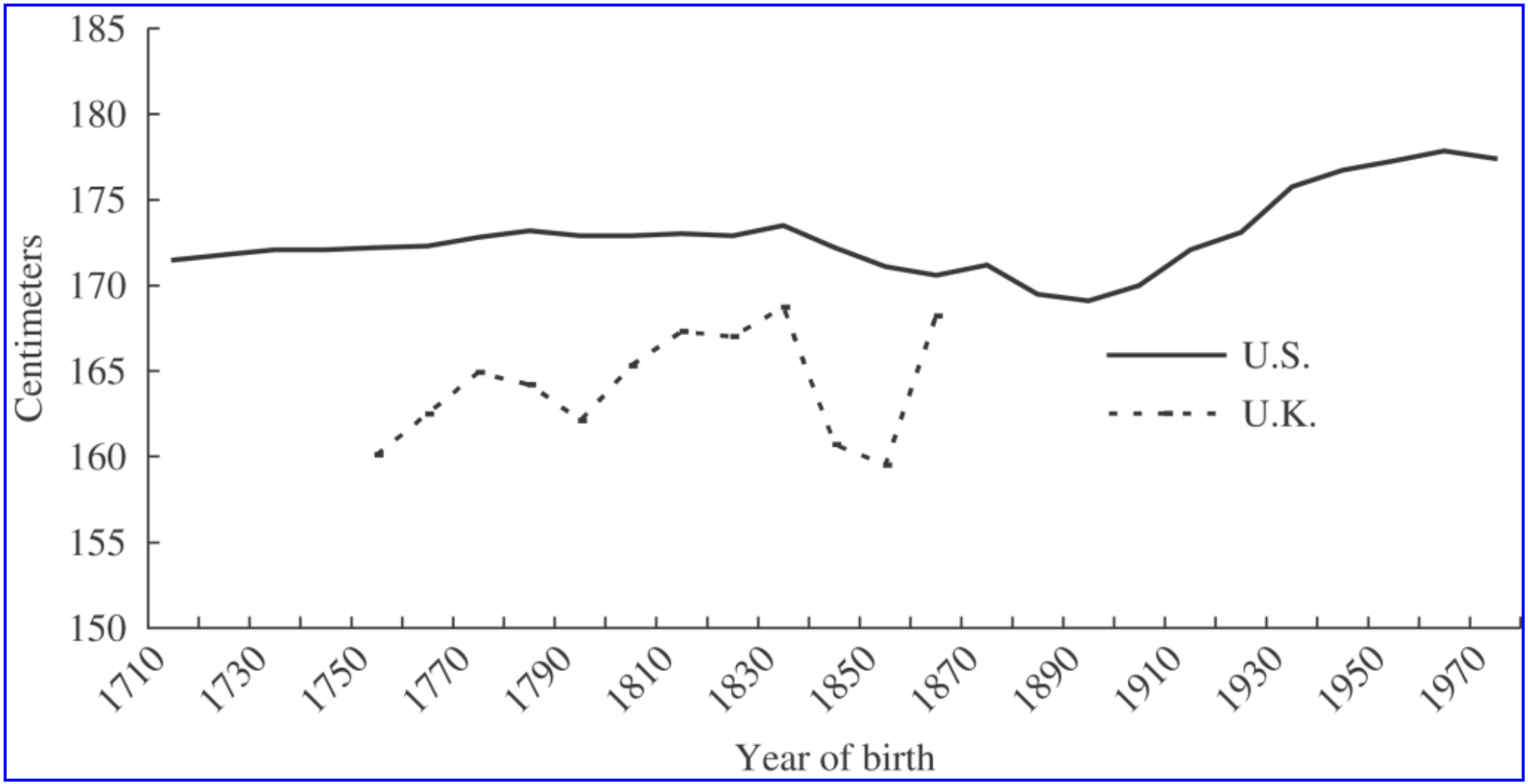

One may speculate that the timing of the declines shown in the Figure 1 is probably more coincidental than emblematic of linkage among similar causal factors across the two countries. While it is possible that growing trade and commerce spread disease, as in the United States, it is more likely that a major culprit in the UK was rapid urbanization and associated increased in exposure to diseases. This conclusion is reached by noting that urban-born men were substantially shorter than the rural-born, and between the periods of 1800–30 and 1830–70 the share of the British population living in urban areas leaped from 38.7 per cent to 54.1 per cent.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- anthropometric history
- cliometrics

Bibliography

Easterlin, R.A. 1971. Regional income trends, 1840–1950. In *The Reinterpretation of American Economic History*, ed. R.W. Fogel and S.L. Engerman. New York: Harper and Row.

Engerman, S.L. 1997. The standard of living debate in international perspective: measures and indicators. In *Health and Welfare during Industrialization*, ed. R.H. Steckel and R. Floud. Chicago: University of Chicago Press.

Floud, R., Wachter, K.W. and Gregory, A.S. 1990. *Height, Health, and History: Nutritional Status in the United Kingdom, 1750–1980*. Cambridge: Cambridge University Press.

Haines, M. 2006. Vital statistics. In *Historical Statistics of the United States: Millennial Edition*, ed. S. Carter et al. New York: Cambridge University Press.

Komlos, J. 1998. Shrinking in a growing economy? The mystery of physical stature during the industrial revolution. *Journal of Economic History* 58, 779–802.

Maddison, A. 2001. *The World Economy: A Millennial Perspective*. Paris: OECD.

Meeker, E. 1980. Medicine and public health. In *Encyclopedia of American Economic History*, ed. G. Porter. New York: Scribner.

Pope, C.L. 1992. Adult mortality in America before 1900: a view from family histories. In *Strategic Factors in Nineteenth Century American Economic History: A Volume to Honor Robert W. Fogel*, ed. C. Goldin and H. Rockoff. Chicago: University of Chicago Press.

Porter, R., ed. 1996. *The Cambridge Illustrated History of Medicine*. Cambridge: Cambridge University Press.

Steckel, R.H. 1986. A peculiar population: the nutrition, health, and mortality of American slaves from childhood to maturity. *Journal of Economic History* 46, 721–41.

Steckel, R.H. 1995. Stature and the standard of living. *Journal of Economic Literature* 33, 1903–40.

Steckel, R.H. 1998. Strategic ideas in the rise of the new anthropometric history and their implications for interdisciplinary research. *Journal of Economic History* 58, 803–21.

Steckel, R.H. 2000. Industrialization and health in historical perspective. In *Poverty, Inequality and Health*, ed. D. Leon and G. Walt. Oxford: Oxford University Press.

Steckel, R.H. 2006. Health, nutrition and physical wellbeing. In *Historical Statistics of the United States: Millennial Edition*, ed. S. Carter, S. Gartner, M. Haines, A. Olmstead, R. Sutch and G. Wright. New York: Cambridge University Press.

Steckel, R.H. and Floud, R., eds. 1997. *Health and Welfare during Industrialization*. Chicago: University of Chicago Press.

How to cite this article

Steckel, Richard H. "standards of living (historical trends)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
id=pde2008_S000533> doi:10.1057/9780230226203.1606

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

state capture and corruption in transition economies

Konstantin Sonin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

State capture by industrial lobbyists is a significant obstacle to normal economic development of formerly command (socialist) economies, at both the local and the national levels. It is prevalent in transition economics because of an excessively concentrated industrial structure and low labour mobility, both horizontal and vertical, a high level of discretion of public officials in economic affairs, and generally weak political institutions. Most of these features might be traced back to the pre-transition legacy.

Keywords

administrative corruption; agency problems; barriers to entry; civil society; Coase theorem; command economy; concentration; control rights; corporate governance; corruption; decentralization; enforcement of contracts; fiscal federalism; influence; innovation; interjurisdictional mobility; local government; oligarchs; property rights protection; quotas and tariffs; rents; rule of law; social networks; special interests; state capture; Tiebout hypothesis; transition and institutions; transition economies; wealth effects

Article

In an ideal democratic society, citizens determine economic policy by a direct voting procedure or by selecting appropriate candidates at polls. In the real world, economic policy is affected either directly by self-interested elected officials or bureaucrats, or indirectly by special interests such as industrial lobbyists or even large individual enterprises. The actual channels of the (primarily negative) influence of special interests on economic policy are called ‘state capture’. In most contexts, state capture necessarily involves ‘corruption’, that is, abuse of public office for private gain.
In any single instance of state capture, there are winners and losers. Typically, winners are politically important or simply large firms or whole industries and bureaucrats that receive favours from those

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
firms. On the losing side, there are small, politically unimportant businesses, and, ultimately, the public interest and consumer welfare. For transition economies, the overall growth rate of the enterprise sector is estimated to be ten per cent lower in a capture state than in a state where the effect is less pronounced (Hellman, Jones and Kaufmann, 2003).

The principal characteristics of transition economies, as related to the issue of state capture, are an excessively concentrated industrial structure inherited from the planned economy, a high level of discretion of public officials with respect to economic policy, weak institutions of political control such as party system or independent media, suboptimal allocation of authority between government layers, and low labour mobility, both horizontal and vertical. In short, all of these factors might be traced to the command economy legacy, one-party political system, and legal system subordinated to political authorities, which were characteristic for most now-in-transition countries in the mid-1980s.

In transition economies, political channels that transmit information and, if necessary, money from firms and other special interests to politicians are not only less efficient but also less institutionalized. Furthermore, conceptualization of a specific exchange, both by parties to the transaction and by a student of transition, might vary widely. For example, what is considered as a fully legal lobbying activity or campaign contribution in an OECD country might be thought of as a bribe or even outright extortion in some other economy.

Theory and transition specifics

A bureaucrat's ability to extract bribes, the manifest form of corruption, is primarily determined by her ability to create rents by shaping the playing field for business (Rose-Ackerman, 1999; Shleifer and Vishny, 1993). In a perfect analogy with a monopolist's behaviour, if a certain kind of business activity requires obtaining a licence from a regulating body, the bureaucrat who is able to collect kickbacks for granting a licence has incentives to keep the number of licences issued less than socially optimal, thus increasing the ‘price’ of those licences that are actually issued. Furthermore, this provides incentives for bureaucrats to create new rent-seeking opportunities by introducing as much licensing and regulation as possible. A monopolistic bureaucrat sets a lower bribe level than a chain of successive monopolies, but the total volume of bribes is obviously higher, and so the impact on social welfare of centralized versus decentralized corruption is ambiguous.

Typically, special interests, through the channels of state capture, seek protection from competition such as barriers to entry to the local market. At the international level, protection usually takes the form of either import quotas or tariff protection. At the local level, it takes the form of licensing, small-scale regulation, and preferential treatment such as tax exemptions granted to individual firms. State capture, as a stable, long-standing relationship between existing businesses and incumbent politicians, has adverse effects on business and politics. In business, state capture prevents innovation by new or potential entrants, and, as a result, by the incumbent firms as well, and thus constrains market development. In politics, it decreases the chances of challengers to mount an aggressive campaign against an incumbent, which in turn reduces the electoral accountability of incumbents.

It is possible to further refine the concept of state capture. Hellman, Jones, and Kaufmann (2003), based on their econometric analysis of a large sample of firms in transition economies, argue that is helpful to distinguish state capture from influence by special interests and administrative corruption, with each of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the three forms of business involvement in politics having distinct causes and consequences. In this refined definition, *state capture* is the process by which firms shape rules of the game through semi-institutionalized bribes paid to public officials and politicians; *influence* is the same process without direct transfers (this category corresponds to costly 'informational lobbying' à la Grossman and Helpman, 2001); and *administrative corruption* encompasses all 'petty' forms of bribery related to law enforcement and regulation.

Political power and economic power

At the beginning of the transition, most of the political institutions that help to mitigate agency problems in the developed world such as independent courts and media, grass-roots political parties, and an institutionalized civil society were virtually non-existent. Inadequate provision of basic public goods such as property rights protection and enforcement of contracts, which are crucial for economic development, forced economic agents to seek alternative tools, for example by supplementing their productive investment with investment in private protection. For large businesses, state capture was a potentially powerful tool. In this view, bribes to bureaucrats or other forms of privately financed purchase of a public good becomes a strategy of an economic agent to increase efficiency and predictability in his business relations.

A specific feature of many transition economies, as compared with developing countries with a similar level of GDP per capita, is that they have been left with the remnants of the command economy with its highly centralized industry. The effect was more pronounced in industrially developed countries such as Russia or Slovakia, and less in countries such as Vietnam or Albania.

The existence of enterprises with very large employment levels produced a specific form of state capture. The managers of large enterprises, regardless of whether they were profitable, have a large menu of political instruments in their hands from which to choose. In particular, they could either rely on ties inherited from the times of the plan, or operate in a newly formed web of quasi-market exchanges including various forms of barter (Gaddy and Ickes, 2002). In Russian practice, quasi-market exchanges have often relied on government power to set individual tax rates and energy tariffs for enterprises. In the Hellman, Jones, and Kaufmann (2003) classification this often amounts to influence, not state capture, as government intervention is caused not by direct bribery, but by a complex chain of exchanges made possible by agents' participation in the same social network. The downside of this phenomenon is that inefficient enterprises are not driven out of the market, and at the same time provide political pressure to deter entry of new enterprises.

Proponents of 'big bang' reforms argued that establishing the rule of law, including institutions of property rights protection, requires the creation of a 'grass-roots' demand for the rule of law. In practice, it meant that the former state property should first go into the hands of private owners, and only then would those owners become natural proponents of a system of property rights protection. With the high degree of concentration present in transition economies, and with undeveloped factor markets, this forecast (from private ownership to demand for the rule of law to property rights) has not been borne out. The rich – the beneficiaries of early privatization – have obvious incentives to protecting their property privately, for example via state capture. Consequently, they do not have incentives to lobby for the establishment of well-working state institutions such as independent courts or efficient bureaucrats.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Instead, they seek to increase their political influence and modify the existing state institutions so that resources and wealth continue to be redistributed in their favour (Polishchuk and Savvateev, 2004; Sonin, 2003). (This is an example of how the Coase theorem is not valid when there are wealth effects.) In Russia, the country that entered transition after the longest period of Communist rule, and was the most industrialized of the transition economies, a decade after economic reforms were launched most of the productive assets were concentrated in the hands of a few individuals, the so-called oligarchs (see Oligarchs). In part, it is an efficiency requirement that, when property rights are poorly protected, control rights over assets become concentrated: the larger a single owner's share is, the greater her incentives are to pursue improvements, such as in corporate governance. (Recent cross-country studies show that the worse the general protection and enforcement of property rights are, the greater is the concentration of control rights.) On the other hand, wealth inequality very often imposes heavy costs on the economy, primarily because it produces widespread inequality of opportunity. Still, the main problem is that the oligarchs who rely on state capture for protection of their property rights and enforcement of contracts do not form a natural constituency for the rule of law. This effect is especially strong when economic inequality is accompanied by underdeveloped democratic institutions, which is typically the case in formerly Communist countries.

Decentralization and interjurisdictional mobility

A critical check on the extent of state capture comes from institutions of federalism. The traditional approach to fiscal federalism focused on externalities in the provision of public goods that arise from preference heterogeneity in different jurisdictions. In the 1990s, a new approach emerged emphasizing the accountability of government officials (agency problems) at both the central and the local levels (for example, Qian and Weingast, 1997; see also Bardhan and Mookherjee, 2006). All former command economies started transition with an overly centralized government structure, and faced the problem of optimal reallocation of economic and political authority.

The effect of decentralization on state capture is twofold. First, more authority allocated to the local level makes the agency problem at this level less prevalent, thus increasing accountability and reducing corruption. On the other hand, special interests might have much more influence over local government bodies; therefore, shifting authority downward might make state capture both more desirable and easier to achieve. In Bardhan and Mookherjee (2006) a reallocation of authority towards local government has a number of consequences. First, the amount of bribes collected by the central government decreases; second, local governments become captured by local special interests. Since local capture might be more easily supported by a social network, the total amount of bribes in the economy declines. However, since corruption at the central level is more money-based, the special interests are less entrenched than at the local level, and thus local capture brings more economic inefficiency. Ultimately, decentralization, while reducing bribe-based corruption measures, reduces economic efficiency.

Starting with Tiebout (1956), interjurisdictional mobility has been considered a major constraint on the local governments’ power to abuse their prerogatives. When mobility is high and subjects can relocate from a jurisdiction with a predatory or hostile government, the local government's monopoly power over laws, regulation, and their execution is compromised. (See Slinko, Yakovlev and Zhuravskaya, 2005, for unique evidence on political capture at the local level.) Since the capacity to extract bribes is increasing

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in the bureaucrats' power to manipulate regulation, high mobility – of both firms and individuals – would force local governments to compete in providing a business-friendly environment. For a local government, the incentives to devote resources to, for example, fighting corruption, increase with the resources devoted by neighbouring governments.

The endless transition

In most transition countries, the transition from the command economy started in the late 1980s and early 1990s. It might be argued that most of the economic problems they face are no longer those of transition, but those of economic development.
Since the transition began, fear of the *Leviathan state* has been swiftly replaced by the fear of the *capture state*, where large and powerful businesses tilt the playing field through bribery, media ownership, huge campaign contributions, and direct participation in politics. Lurid stories told about some extreme forms of capture, such as the use of corrupt secret service and police officers against business competitors, have crowded out images of millions perishing in forced-labour camps or dying in numerous famines caused by Communist economic management. There is a false sense of symmetry between the state manipulating its subjects and subjects manipulating the state: while the former has proved to be dangerous on a large scale, the second is a mere obstacle to economic development. There are several ways to deal with this obstacle. A backward-looking way is through restoration of the repressive capacity of the state. Another is through the development of political and civic institutions that may put a check on elected officials and bureaucrats, and through the institutionalization of business influence on politics in an efficient way.

See Also

- oligarchs
- political institutions, economic approaches to
- transition and institutions

Bibliography

Bardhan, P. and Mookherjee, D. 2000. Capture and governance at local and national levels. *American Economic Review* 90, 135–9.

Bardhan, P. and Mookherjee, D. 2006. Decentralization, corruption and government accountability: an overview. In *Handbook of Economic Corruption*, ed. S. Rose-Ackerman. Northampton, MA: Edward Elgar.

Gaddy, C. and Ickes, B. 2002. *Russia's Virtual Economy*. Washington, DC: Brookings Institution Press.

Grossman, G. and Helpman, E. 2001. *Special Interest Politics*. Cambridge, MA: MIT Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hellman, J., Jones, G. and Kaufmann, D. 2003. Seize the state, seize the day: state capture and influence in transition economies. *Journal of Comparative Economics* 31, 751–73.

Persson, T. and Tabellini, G. 2000. *Political Economics*. Cambridge, MA: MIT Press.

Polishchuk, L. and Savvateev, A. 2004. Spontaneous (non)emergence of property rights. *Economics of Transition* 12, 103–27.

Qian, Y. and Weingast, B. 1997. Federalism as a commitment to preserving market incentives. *Journal of Economic Perspectives* 11(4), 83–92.

Roland, G. 2000. *Transition and Economics: Politics, Markets, and Firms*. Cambridge, MA: MIT Press.

Rose-Ackerman, S. 1999. *Corruption and Government: Causes, Consequences, and Reform*. Cambridge: Cambridge University Press.

Shleifer, A. and Vishny, R. 1993. Corruption. *Quarterly Journal of Economics* 108, 601–17.

Slinko, I., Yakovlev, E. and Zhuravskaya, E. 2005. Laws for sale: evidence from Russia. *American Law and Economics Review* 7, 284–318.

Sonin, K. 2003. Why the rich may favor poor protection of property right. *Journal of Comparative Economics* 31, 715–31.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

How to cite this article

Sonin, Konstantin. "state capture and corruption in transition economies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000478> doi:10.1057/9780230226203.1607

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

state space models

Andrew Harvey
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The state space form opens the way to the statistical treatment of a wide range of dynamic models in a unified framework. For models formulated in unobserved components it offers algorithms for filtering, signal extraction and prediction. Data irregularities can be handled and recent work on computational methods has extended the range of nonlinear and non-Gaussian models that can be adopted for practical use.

Keywords

ARIMA models; dynamic stochastic general equilibrium (DSGE) models; finite sample computation; Hodrick–Prescott filter; International Labor Organization (ILO); Kalman filter; linear rational expectations model; Markov chain Monte Carlo; maximum likelihood; mean square errors; missing observations; nowcasting; output gap; particle filtering; Phillips curve; prediction; smoothing; state space form; state space models; state vector; stochastic volatility models; structural time series models; Wiener–Kolomogorov (WK) filter

Article

State space models is a rather loose term given to time series models, usually formulated in terms of unobserved components, that make use of the state space form for their statistical treatment.
At the simplest level, structural time series models (STMs) are set up in terms of components such as trends and cycles that have a direct interpretation. Signal extraction, or smoothing, provides a description of these features that is model-based and hence avoids the ad hoc nature of procedures such as moving averages and the Hodrick–Prescott filter. While smoothing uses all the observations in the sample, filtering yields estimates at a given point in time that are constructed only from observations available at that time. For key time series, filtered estimates form the basis for ‘nowcasting’ in that they give an indication of the current state of the economy and the direction in which it is moving. They also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
provide the starting point for forecasts of future observations and components.

Local level model

A simple model with permanent and transitory components illustrates the basic ideas of filtering and smoothing. Suppose that the observations consist of a random walk component plus a random irregular term, that is

$$y_t = \mu_t + \varepsilon_t,\ \varepsilon_t \sim NID(0, \sigma_\varepsilon^2),\ t = 1, \ldots, T \quad (1)$$

$$\mu_t = \mu_{t-1} + \eta_t,\ \eta_t \sim NID(0, \sigma_\eta^2), \quad (2)$$

where the irregular and level disturbances, ε_t and η_t respectively, are mutually independent and the notation $NID(0, \sigma^2)$ denotes normally and independently distributed with mean zero and variance σ^2. The signal–noise ratio, $q = \sigma_\eta^2 / \sigma_\varepsilon^2$, plays the key role in determining how observations should be weighted for prediction and signal extraction. In a large sample, filtering is equivalent to a simple exponentially weighted moving average; the higher q is, the more past observations are discounted. When q is zero, the level is constant and all observations have the same weight. The reduced form of the model has the first differences following a first-order moving average process, that is

$$\Delta y_t = \xi_t + \theta \xi_{t-1},\ \xi_t \sim NID(0, \sigma_\xi^2)$$

where $\theta = [(q^2 + 4q)^{1/2} - 2 - q] / 2$. This produces the same forecasts. However, the structural form in (1) also yields nowcasts and smoothed estimates of the level, μ_t, throughout the series. In the middle of a large sample, the smoothed estimates are approximately equal to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\frac{1+\theta}{1-\theta}\sum_{j}(-\theta)^{|j|}y_{t+j},\ q>0,$$

(3)

while the filtered estimates are

$$(1+\theta)\sum_{j\geq 0}(-\theta)^{j}y_{t-j},\ q>0.$$

(4)

At the end of the sample this estimate also yields the forecast of future levels and future observations. Note that, although the above expressions are useful for displaying the weighting of the observations, finite sample computation is best done by a simple forward recursion for filtering and a subsequent backward one for smoothing.

State space form

The state space form (SSF) is a simple device whereby a dynamic model is written in terms of just two equations. The model in (1) and (2) is a special case. The general linear SSF applies to a multivariate time series, $\mathbf{y}_t$, containing N elements. These observable variables are related to an $m\times 1$ vector, α_t, known as the *state vector*, through a *measurement equation*

$$\mathbf{y}_t=\mathbf{Z}_t\boldsymbol{\alpha}_t+\mathbf{d}_t+\boldsymbol{\varepsilon}_t,\ t=1,\ldots,T$$

(5)

where $\mathbf{Z}_t$ is an $N\times m$ matrix, $\mathbf{d}_t$ is an $N\times 1$ vector and ε_t is an $N\times 1$ vector of serially uncorrelated disturbances with mean zero and covariance matrix $\mathbf{H}_t$. In general the elements of α_t are not observable. However, they are assumed to be generated by a first-order Markov process,

$$\boldsymbol{\alpha}_t=\mathbf{T}_t\boldsymbol{\alpha}_{t-1}+\mathbf{c}_t+\mathbf{R}_t\boldsymbol{\eta}_t,\ t=1,\ldots,T$$

(6)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $\mathbf{T}_t$ is an $m \times m$ matrix, $\mathbf{c}_t$ is an $m \times 1$ vector, $\mathbf{R}_t$ is an $m \times g$ matrix and η_t is a $g \times 1$ vector of serially uncorrelated disturbances with mean zero and covariance matrix, $\mathbf{Q}_t$. Equation (6) is the *transition equation*. The specification of the system is completed by assuming that the initial state vector, α_0, has a mean of $\mathbf{a}_0$ and a covariance matrix $\mathbf{P}_0$, and that the disturbances ε_t and η_t are uncorrelated with the initial state. The disturbances are often assumed to be uncorrelated with each other in all time periods, though this assumption may be relaxed to allow contemporaneous correlation, the consequence being a slight complication in some of the filtering formulae.

The definition of α_t for any particular statistical model is determined by construction. Its elements may or may not be identifiable with components that have a substantive interpretation, for example as a trend or a seasonal. From the technical point of view, the aim of the state space formulation is to set up α_t in such a way that it contains all the relevant information on the system at time *t* and that it does so by having as small a number of elements as possible. The SSF is not, in general, unique.

The Kalman filter (KF) is a recursive procedure for computing the optimal estimator of the state vector at time *t*, based on the observations up to and including $\mathbf{y}_t$. In a Gaussian model, the disturbances ε_t and η_t, and the initial state, are all normally distributed. Because a normal distribution is characterized by its first two moments, the Kalman filter can be interpreted as updating the mean and covariance matrix of the conditional distribution of the state vector as new observations become available. The conditional mean minimizes the mean square error and when viewed as a rule for all realizations it is the minimum mean square error estimator (MMSE). Since the conditional covariance matrix does not depend on the observations, it is the unconditional MSE matrix of the MMSE. When the normality assumption is dropped, the KF is still optimal in the sense that it minimizes the mean square error within the class of all linear estimators. Given initial conditions, $\mathbf{a}_0$ and $\mathbf{P}_0$, the Kalman filter delivers the optimal estimator of the state vector as each new observation becomes available. When all *T* observations have been processed, it yields the optimal estimator of the current state vector based on the full information set. When the initial conditions cannot be specified a diffuse prior is often placed on the initial state. This amounts to setting $\mathbf{P}_0 = \kappa\mathbf{I}$, and letting the scalar κ go to infinity. Stable algorithms for handling diffuse priors are set out in Durbin and Koopman (2001).

Prediction is carried out straightforwardly by running the KF without updating. Mean square errors of the forecasts are produced at the same time. Smoothing is carried out by a backward filter initialized with the estimates delivered by the KF at time *T*. The aim is to compute the optimal estimator of the state vector at time *t* using information made available after time *t* as well as before. Efficient smoothing algorithms are described in Durbin and Koopman (2001, pp. 70–3). The weights are implicit, but Koopman and Harvey (2003) give an algorithm for computing and displaying them at any point in time.

The state space smoother is far more general than the classic Wiener–Kolomogorov (WK) filter. The WK filter computes weights explicitly and for simple models it is possible to obtain expressions for the estimator in the middle of a doubly infinite sample without too much difficulty. Formula (3) is a case in point. However, the WK filter is limited to time-invariant models and even here it has no computational advantages over the state space fixed-interval smoothing algorithm. In the second edition of his celebrated text describing the WK filter, Whittle (1984, p. xi) writes ‘In its preoccupation with the stationary case and generating function methods, the 1963 text essentially missed the fruitful concept of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
state structure. This … has now come to dominate the subject.'

The system matrices $\mathbf{Z}_t$,$\mathbf{H}_t$,$\mathbf{T}_t$,$\mathbf{R}_t$ and $\mathbf{Q}_t$ may depend on a set of unknown parameters, and one of the main statistical tasks will often be the estimation of these parameters. Thus in the random walk plus noise model, (1), the parameters σ_n^2 and σ_ε^2 will usually be unknown. As a by-product, the KF produces a vector of prediction errors or *innovations* and in a Gaussian model these can be used to construct a likelihood function that can be maximized numerically with respect to the unknown parameters.

Since the state vector is a vector of random variables, a Bayesian interpretation of the Kalman filter as a way of updating a Gaussian prior distribution on the state to give a posterior is quite natural. The mechanics of filtering, smoothing and prediction are the same irrespective of whether the overall framework is Bayesian or classical. Smoothing gives the mean and variance of the state, conditional on all the observations. For the classical statistician, the conditional mean is the MMSE, while for the Bayesian it minimizes the expected loss for a symmetric loss function. With a quadratic loss function, the expected loss is given by the conditional variance. The real differences between classical and Bayesian treatments arise when the parameters are unknown. In a Bayesian framework, the hyperparameters, as they are often called, are random variables. The development of simulation techniques based on Markov chain Monte Carlo (MCMC) has now made a full Bayesian treatment a feasible proposition. This means that it is possible to simulate a distribution for the state that takes account of hyperparameter uncertainty.

Applications

The use of unobserved components opens up a new range of possibilities for economic modelling. Furthermore, it provides insights and a unified approach to many other problems. The examples below give a flavour.

The local linear trend model generalizes (1) by the introduction of a stochastic slope, β_t, which itself follows a random walk. Thus

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t,\ \eta_t \sim NID(0, \sigma_\eta^2), \beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim NID(0, \sigma_\zeta^2), \tag{7}$$

where the irregular, level and slope disturbances, ε_t, η_t and ζ_t, respectively, are mutually independent. If both variances σ_η^2 and σ_ζ^2 are zero, the trend is deterministic. When only σ_ζ^2 is zero, the slope is fixed and the trend reduces to a random walk with drift. Allowing σ_ζ^2 to be positive but setting σ_η^2 to zero gives an integrated random walk trend, which when estimated tends to be relatively smooth. Signal extraction of the trend by setting the signal–noise ratio, $q = \sigma_\zeta^2 / \sigma_\varepsilon^2$, to 1/1600 gives the Hodrick–Prescott filter for quarterly data. Adding a cyclical component to (1) provides a vehicle for detrending based on a model the parameters of which can be estimated from the data.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Orphanides and van Norden (2002) have recently stressed the importance of tracking the output gap in *real time*. Given the parameter estimates, real time estimation of components such as the output gap is just an exercise in filtering. However, as new observations become available the estimate of the gap at a particular point in time can be improved by smoothing. Harvey, Trimbur and van Dijk (2007) adopt a Bayesian approach which has the advantage of giving the full distribution of the output gap. Statistics such as the probability that the output gap is increasing are readily calculated. Following Kuttner (1994), Harvey, Trimbur and van Dijk (2007) also construct an unobserved components model relating the output gap to inflation in what is effectively a Phillips curve relationship.

A number of authors, beginning with Sargent (1989), have estimated the structural parameters of dynamic stochastic general equilibrium (DSGE) models using state space methods. The linear rational expectations model is first solved for the reduced-form state equation in its predetermined variables. Once this has been done, the model is put in state space form and the parameters are estimated by maximum likelihood. Alternatively a Bayesian approach can be adopted; see Smets and Wouter (2003, p. 1138).

Data irregularities

Some of the most striking benefits of the structural approach to time series modelling become apparent only when we start to consider more complex problems. In particular, the SSF offers considerable flexibility with regard to dealing with data irregularities, such as missing observations and observations at mixed frequencies. Missing observations are easily handled in the SSF simply by omitting the updating equations while retaining the prediction equations. Filtering and smoothing then go through automatically and the likelihood function is constructed using prediction errors corresponding to actual observations. With flow variables, such as income, the issue is one of temporal aggregation. This may be dealt with by the introduction of a cumulator variable into the state. The study by Harvey and Chung (2000) on the measurement of British unemployment provides an illustration of how mixed frequencies are handled and how using an auxiliary series can improve the efficiency of nowcasting and forecasting a target series. The challenge was how to obtain timely estimates of the underlying change in unemployment. Estimates of the numbers of unemployed according to the International Labor Organization (ILO) definition are given by the Labour Force Survey (LFS), which consists of a rotating sample of approximately 60,000 households. These estimates have been published on a quarterly basis since the spring of 1992, but from 1984 to 1991 estimates were available for the spring quarter only. Another measure of unemployment, based on administrative sources, is the number of people claiming unemployment benefit. This measure, known as the claimant count, is available monthly, with very little delay and is an exact figure. It does not provide a figure corresponding to the ILO definition, but it moves roughly in the same way as the LFS figure. The first problem is how to extract the best estimate of the underlying monthly change in a series which is subject to sampling error and which may not have been recorded every month. The second is how to use a related series to improve this estimate. These two issues are of general importance, for example in the measurement of the underlying rate of inflation or the way in which monthly figures on industrial production might be used to produce more timely estimates of national income. State space methods deal with the mixed frequencies in the target series, with the rather complicated error structure coming from the rotating sample (see Pfeffermann, 1991) and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
with the different frequency of the auxiliary series.

Continuous time

Continuous time STMs observed at discrete intervals can easily be put in SSF (see Harvey, 1989, ch. 9). An important case is the continuous time version of (7) where the smoothed trend is a cubic spline. Setting up such a model for a cubic spline enables the smoothness parameter to be estimated by maximum likelihood and the fact that irregularly spaced data may be handled means that it can be used to fit a nonlinear function to cross-sectional data. The model can easily be extended, for example to include other components, and it can be compared with alternative models using standard statistical criteria (see Kohn, Ansley and Wong, 1992).

Nonlinear and non-Gaussian models

Some of the most exciting recent developments in time series have been in nonlinear and non-Gaussian models. For example, it is possible to fit STMs with heavy-tailed distributions on the disturbances, thereby making them robust with respect to outliers and structural breaks. Similarly, non-Gaussian models, designed to deal with count data and qualitative observations, can be set up with stochastic components. In the general formulation of a state space model, the distribution of the observations is specified conditional on the current state and past observations, that is

$$p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathbf{Y}_{t-1}),\ t = 1, \ldots, T \tag{8}$$

where $\mathbf{Y}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots\}$. Similarly the distribution of the current state is specified conditional on the previous state and observations so that

$$p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \mathbf{Y}_{t-1}). \tag{9}$$

The initial distribution of the state is given as $p(\alpha_0)$. In a linear Gaussian model the conditional distribution in (8) and (9) are characterized by their first two moments and so they are specified by the measurement and transition equations. The Kalman filter updates the mean and covariance matrix of the state. In more general models, computer-intensive methods, using techniques such as importance sampling, have to be applied (see Durbin and Koopman, 2001). Within a Bayesian framework, methods are normally based on MCMC. Particle filtering is often used for signal extraction; see the review in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Harvey and de Rossi (2006).

The use of state space methods highlights a fundamental distinction in time series models between those motivated by description and those set up to deal directly with forecasting. This is epitomized by the contrast between STMs on the one hand and autoregressions and autoregressive-integrated-moving average (ARIMA) models on the other. In a linear Gaussian world, the reduced form of an STM is an ARIMA model and questions regarding the merits of STMs for forecasting revolve round the gains, or losses, from the implied restrictions on the reduced form and the guidance, or lack of it, given to the selection of a suitable model (see the discussion in Harvey, 2006). Once nonlinearity and non-Gaussianity enter the picture, the two approaches can be very different. Models motivated solely by forecasting tend to be set up in terms of a distribution for the current observations conditional on past observations rather than in terms of components. For example, changing variance can be captured by a model from the generalized autoregressive conditional heteroscedasticity (GARCH) class, where conditional variance is a function of past observations, as opposed to a stochastic volatility (SV) model in which the variance is a dynamic unobserved component. The readings in Shephard (2005) describe SV models and discuss the use of computationally intensive methods for estimating them.

The realization that the statistical treatment of a wide range of dynamic models can be dealt with directly in a unified framework is important. For engineers, using state space methods is a natural way to proceed. For many economists, brought up with regression and autoregression, state space is an alien concept. This is changing. State space methods are now becoming an important part of the toolkit of econometricians and economists.

See Also

- data filters
- Kalman and particle filtering
- prediction formulas

Bibliography

Durbin, J. and Koopman, S. 2001. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Harvey, A. 1989. *Forecasting, Structural Time Series Models and Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. 2006. Forecasting with unobserved components time series models. In *Handbook of Economic Forecasting*, vol. 1, ed. G. Elliot, C. Granger and A. Timmermann. Amsterdam: North-Holland.

Harvey, A. and Chung, C.-H. 2000. Estimating the underlying change in unemployment in the UK (with discussion). *Journal of the Royal Statistical Society, Series A* 163, 303–39.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Harvey, A. and de Rossi, G. 2006. Signal extraction. In *Palgrave Handbook of Econometrics*, vol. 1, ed. K. Patterson and T. Mills. Basingstoke: Palgrave Macmillan.

Harvey, A., Trimbur, T. and van Dijk, H. 2007. Trends and cycles in economic time series: a Bayesian approach. *Journal of Econometrics* 140(2), 618–49.

Kohn, R., Ansley, C. and Wong, C.-H. 1992. Nonparametric spline regression with autoregressive moving average errors. *Biometrika* 79, 335–46.

Koopman, S. and Harvey, A. 2003. Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics and Control* 27, 1317–33.

Kuttner, K. 1994. Estimating potential output as a latent variable. *Journal of Business and Economic Statistics* 12, 361–8.

Orphanides, A. and van Norden, S. 2002. The unreliability of output gap estimates in real-time. *Review of Economics and Statistics* 84, 569–83.

Pfeffermann, D. 1991. Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics* 9, 163–75.

Sargent, T. 1989. Two models of measurements and the investment accelerator. *Journal of Political Economy* 97, 251–87.

Smets, F. and Wouter, R. 2003. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association* 1, 1123–75.

Shephard, N. 2005. *Stochastic Volatility*. Oxford: Oxford University Press.

Whittle, P. 1984. *Prediction and Regulation*, 2nd edn. Blackwell: Oxford.

How to cite this article

Harvey, Andrew. "state space models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000509>
doi:10.1057/9780230226203.1609

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

state-dependent preferences

Edi Karni

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

State-dependent preferences pertains to situations involving decision making in the face of uncertainty, in which the states resolving the uncertainty are of direct concern to the decision maker, and affect his or her evaluation of the consequences. The presence of state-dependent preferences raises fundamental issues concerning the representation of the decision maker's preference relations and, in particular, the definition and interpretation of subjective probabilities. These difficulties are explained and an approach to resolving them is discussed.

Keywords

decisions under uncertainty; expected utility hypothesis; moral hazard; risk aversion; Savage, L.; Savage's subjective expected utility model; state-dependent preferences; uncertainty

Article

Theories of individual decision making under uncertainty pertain to situations in which a choice of a course of action, by itself, does not determine the outcome. To formulate these theories Savage (1954) introduced what has become the standard analytical framework. It consists of three sets: a set *S*, of *states of the world* (or *states*, for short); an arbitrary set *C*, of *consequences*; and the set *F*, of all the functions from the set of states to the set of consequences. Elements of *F*, referred to as ‘acts’, represent courses of action, consequences describe anything that may happen to a person, and states are the resolutions of uncertainty, that is, ‘a description of the world so complete that, if true and known, the consequences of every action would be known’ (Arrow, 1971, p. 45). Decision makers are characterized by preference relations, $\succsim$, on *F*. With few exceptions, preference relations are taken to be complete (that is, for all *f* and *g* in *F*, either $f \succsim g$ or $g \succsim f$) and transitive binary relations on *F*. The symbols $f \succsim g$ have the interpretation ‘the course of action *f* is preferred or indifferent to the course of action *g*’. The strict preference relation, $\succ$, and the indifference relation, $\succ$, are the asymmetric and symmetric parts of $\succsim$,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
respectively.

To speak loosely, a preference relation is state-dependent when the prevailing state of nature is itself of direct concern to the decision maker. For example, taking out a health insurance policy is choosing an action whose consequences – the indemnities – depend on the realization of the decision maker's state of health. In this example, the state is the decision maker's state of health. It affects the decision maker's well-being directly, and indirectly, through the payoff prescribed by the health insurance policy. The preference relation may display ordinal state dependence, in which case the underlying state may affect the decision maker's preferences by altering his ordinal ranking of the consequences; or cardinal state dependence, by altering his risk attitudes; or both.

To define state dependence formally, it is convenient to adopt the model of Anscombe and Aumann (1963). In this model the state space is finite, and the consequences are lotteries, that is, probability distributions that assign strictly positive probability to a finite number of outcomes. Denote by $L(X)$ the set of lotteries on an arbitrary set, X, of possible outcomes. Given a preference relation $\succsim$ on F; a state s; and f, f', g, g' in F, define a preference relation on F conditional on s, $\succsim_s$, by $f \succsim_s f'$ if $g \succsim g'$ whenever $f(s) = g(s)$, $f'(s) = g'(s)$ and $g(s') = g'(s')$ for all $s' \in S - \{s\}$. Because acts are functions, $f(s)$ is defined uniquely. Thus $\succsim_s$ defines a preference relation on $L(X)$ conditional on s. This induced preference relation is also denoted by $\succsim_s$.

A state $s \in S$ is said to be *null* if $f \succsim_s f'$ for all $f, f' \in F$, otherwise it is *non-null*.

Definition. A preference relation $\succsim$ on F is *state dependent* if $\succsim_s \neq \succsim_{s'}$ for some non-null s and s' in S.

Because consequences are lotteries, if a preference relation $\succsim$ on F displays state dependence, then $\succsim_s$ and $\succsim_{s'}$ must differ on the ranking of some lotteries in $L(X)$. This may be due to distinct attitudes toward risk and/or distinct ordering of outcomes, that is, degenerate lotteries that assign the given outcomes probability one. Circumstances in which the dependence of the decision maker's preferences on the state constitutes an indispensable feature of the decision problem include the choice of health insurance coverage (see Arrow, 1974; Karni, 1985); the choice of flight insurance coverage (see Eisner and Strotz, 1961); the choice of optimal consumption and life insurance plans in the face of uncertain life span (see Yaari, 1965; Karni and Zilcha 1985); and the provision of collective protection (see Cook and Graham, 1977).

Subjective expected utility representations

Preferences among acts are a matter of personal judgement, presumably combining the decision maker's valuation of the consequences and his or her beliefs regarding the likely realization of alternative events (that is, subsets of the state space). Subjective expected utility theory pertains to preference relations whose structures allow the decision maker's valuations of the consequences to be expressed numerically, by a utility function; his or her beliefs to be quantified by a (subjective) probability measure on the set of states; and the acts to be evaluated by the expectations of the utility of the corresponding consequences with respect to the subjective probability. In other words, the theory depicts the decision makers' choice among alternative acts as expected utility maximizing behaviour.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

By the classic von Neumann–Morgenstern expected utility theorem, $\succsim$ on F satisfies the axioms of expected utility theory (that is, $\succsim$ is a complete and transitive binary relation satisfying the Archimedean and independence axioms) if and only if there exist real-valued functions $w(\cdot, s)$ on X, $s \in S$, such that for all $f, g \in F$,

$$f \succsim g \Leftrightarrow \sum_{s \in S} \sum_{x \in X} w(x, s) f(x, s) \geq \sum_{s \in S} \sum_{x \in X} w(x, s) g(x, s). \quad (1)$$

Furthermore, the functions $\{w(\cdot, s) | s \in S\}$ are unique up to cardinal unit comparable transformation. (That is, if some other utility functions $\{w'(\cdot, s) | s \in S\}$ represent $\succsim$ on F, in the sense of equation (1), there exit $b > 0$ and real numbers $a(s)$, one for each state, such that $w'(s) = bw(s) + a(s)$ for all $s \in S$.) The function $w(\cdot, \cdot)$ captures the decision maker's valuation of the outcomes and his or her beliefs about the likely realization of the states. The axioms of expected utility theory do not imply a unique decomposition of w into subjective probability distribution on S and utility on outcome-state pairs. Indeed, let $p(s)$, $s \in S$, be any list of positive numbers that sum up to 1, where $p(s) = 0$ if and only if s is null. Define $u(x, s) = w(x, s) / p(s)$ for all non-null $s \in S$ and all $x \in X$, and $u(x, s) = \bar{u}$ if s is null. Then, by equation (1), $\succsim$ is represented by $\sum_{s \in S} p(s) \sum_{x \in X} u(x, s) f(x, s)$. The question is as follows: are there additional conditions that would imply a *unique* decomposition of $w(x,s)$ into a product of utility representing the (possibly state-dependent) valuation of the outcomes and probabilities representing beliefs that govern the decision maker's choice among acts?

Anscombe and Aumann (1963) show that a preference relation is non-trivial (that is, $f \succ g$ for some $f, g \in F$) satisfying the axioms of expected utility theory and state independence, that is, $\succsim_s = \succsim_{s'}$, for all non-null $s, s' \in S$ if and only if there exist a real-valued function, u, on X, and a subjective probability distribution, π, on S such that, for all $f, g \in F$,

$$f \succsim g \Leftrightarrow \sum_{s \in S} \pi(s) \sum_{x \in X} u(x) f(x, s) \geq \sum_{s \in S} \pi(s) \sum_{x \in X} u(x) g(x, s). \quad (2)$$

Moreover, u is unique up to positive linear transformation, and π is unique satisfying $\pi(s) = 0$ if and only if s is null.

The subjective expected utility representation (2) separates risk attitudes, represented by the utility function, from beliefs, represented by the subjective probabilities. However, the uniqueness of the probabilities depends crucially on the premise that constant acts are constant utility acts. This premise is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

not implied by the axioms. In particular, *state-independent preferences do not imply state-independent utility function*. To see why, let γ be a strictly positive real-valued function on S and $\Gamma = \sum_{s \in S} \gamma(s)\pi(s)$. Define $\hat{u}(x, s) = u(x) / \gamma(s)$, for all $x \in X$ and $s \in S$, and let $\hat{\pi}(s) = \gamma(s)\pi(s) / \Gamma$, for all $s \in S$. Then, by equation (2) and the uniqueness of u, for all $f, g \in F$,

$$f \succsim g \Leftrightarrow \sum_{s \in S} \hat{\pi}(s) \sum_{x \in X} \hat{u}(x, s) f(x, s) \geq \sum_{s \in S} \hat{\pi}(s) \sum_{x \in X} \hat{u}(x, s) g(x, s). \tag{3}$$

Thus, the utility–probability pair $(\hat{u}, \hat{\pi})$ induces a subjective expected utility representation of $\succsim$ that is equivalent to the one induced by the pair (u, π). There are infinitely many distinct utility-probability pairs that represent the same preference relation in the sense of equation (2). Moreover, because π and $\hat{\pi}$ are distinct, even if beliefs exist a priori and are coherent enough to allow their representation by probabilities, it is not evident which of the infinitely many probability distributions consistent with $\succsim$ actually represents the decision makers beliefs. But if the probabilities that figure in the representation are meaningless, there seems to be no compelling reason to prefer the expected utility representation (2) over the more general additive representation (1). On the contrary, because the additive representation does not require that the preferences be state independent, it is applicable to the analysis of problems, such as the demand for health and life insurance, in which the assumption of state-independent preferences is clearly inadequate.

Hypothetical preferences and subjective expected utility representations of state-dependent preferences

An alternative analytical framework, introduced by Karni and Schmeidler (1981), postulates the existence of a preference relation on hypothetical lotteries, whose prizes are outcome-state pairs. This preference relation is assumed to satisfy the axioms of expected utility and to be consistent with the actual preference relation on acts. Because the hypothetical lotteries imply distinct, hence incompatible, marginal distributions on the state space, preferences among such lotteries are *introspective* and may be expressed verbally only as hypothetical choices. Decision makers are supposed to be able to conceive of such hypothetical lotteries and to invoke, for the purpose of their evaluation, the same mental processes that govern their actual decisions.

To express these ideas formally, denote by $L(X \times S)$ the set of all probability distributions on $X \times S$ that assign strictly positive probabilities to a finite number of outcome-state pairs. A lottery $\ell \in L(X \times S)$ is said to be *non-degenerate* if $\sum_{y \in X} \ell(y, s) > 0$ for all $s \in S$. Denote by $\widehat{\succsim}$ an introspective preference relation on $L(X \times S)$. For each $s \in S$, define the conditional introspective preferences, $\widehat{\succsim}_s$, on $L(X \times S)$ analogously to the definition of $\succsim_s$ (that is, $\ell \widehat{\succsim}_s \ell'$ if and only if $\ell \widehat{\succsim} \ell'$, for all $\ell, \ell' \in L(X \times S)$ such that $\ell(\cdot, s') = \ell'(\cdot, s')$ for all $s' \in S - \{s\}$).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To speak loosely, the introspective preference relation $\widehat{\succsim}$ and the actual preference relation $\succsim$ are consistent if they are induced by the same utilities. Formally, define a mapping, *H*, from $L(X \times S)$ to *F* as follows: For each non-degenerate $\ell \in L(X \times S)$, let $H(\ell)(x, s) = \ell(x, s) / \sum_{y \in X} \ell(y, s)$, for all $(x, s) \in X \times S$. A state *s* is said to be *obviously null* if $f \succsim_s f'$ for all *f* and f' in *F* and there are $\ell, \ell' \in L(X \times S)$ such that $\ell \widehat{\succ}_s \ell'$. A state *s* is *obviously non-null* if $f \succ_s g$ for some *f* and *g*. A state *s* is *essential* if there are ℓ and ℓ' in $L(X \times S)$ such that $\ell \widehat{\succ}_s \ell'$.

Strong consistency: For all $s \in S$ and non-degenerate ℓ and ℓ' in $L(X \times S)$, $H(\ell) \succ_s H(\ell')$ implies $\ell \widehat{\succ}_s \ell'$, and if *s* is obviously non-null, then $\ell \widehat{\succ}_s \ell'$ implies $H(\ell) \succ_s H(\ell')$.

Theorem (Karni and Schmeidler, 1981): Let $\succsim$ be a non-trivial binary relation on *F* and $\widehat{\succsim}$ a binary relation on $L(X \times S)$. Then each of the two relations satisfies the axioms of expected utility and jointly they satisfy strong consistency if and only if there exist a real-valued function, *u*, on $X \times S$ and a probability distribution, π, on *S* such that, for all *f* and *g* in *F*,

$$f \succsim g \Leftrightarrow \sum_{s \in S} \pi(s) \sum_{x \in X} u(x, s) f(x, s) \geq \sum_{s \in S} \pi(s) \sum_{x \in X} u(x, s) g(x, s) \tag{4}$$

and, for all ℓ and ℓ' in $L(X \times S)$,

$$\ell \widehat{\succsim} \ell' \Leftrightarrow \sum_{s \in S} \sum_{x \in X} u(x, s) \ell(x, s) \geq \sum_{s \in S} \sum_{x \in X} u(x, s) \ell'(x, s). \tag{5}$$

Moreover, the function u is unique up to cardinal unit comparable transformation, the probability π restricted to the event of all essential states is unique, and for *s* obviously null $\pi(s) = 0$ and for *s* obviously non-null $\pi(s) > 0$.

The subjective expected utility representation in (4) applies whether the preference relation, $\succsim$, is state-dependent or state-independent. Furthermore, as Karni and Mongin (2000) observed, because the utility function is identified using hypothetical lotteries, the probability measure π in the representation theorem above quantifies the decision maker's beliefs. A similar result in a somewhat different framework is proved in Karni (2003); a probabilistically sophisticated version of this approach appears in Grant and Karni (2004).

A weaker version of this result, based on restricting the consistency condition to a subset of hypothetical lotteries that have the same marginal distribution on *S*, due to Karni, Schmeidler and Vind (1983), yields

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
a subjective expected utility representation with state-dependent preferences. Wakker (1987) has extended the theory of Karni, Schmeidler and Vind (1983) to include the case in which the set of consequences is a connected topological space. However, the arbitrary choice of the subset of hypothetical lotteries renders the probabilities in these works arbitrary.

Other theories that yield subjective expected utility representations invoke preferences on conditional acts (that is, preference relations over the set of acts conditional on events). Fishburn (1973) and Karni (2007) advanced such theories assuming consequence sets that have distinct structures. Skiadas (1997) proposed a non-expected utility model, based on hypothetical preferences, that yield a representation with state-dependent preferences. In this model, acts and states are primitive concepts, and preferences are defined on act-event pairs. For any such pair the consequences (utilities) represent the decision maker's expression of his holistic valuation of the act. The decision maker is not supposed to be aware whether or not the given event occurred; hence, his evaluation of the act reflects, in part, his anticipated feelings, such as disappointment aversion.

Moral hazard and state-dependent preferences

rèze (D1961; 1987) and Karni (2006) present distinct theories of individual decision making under uncertainty with moral hazard and state-dependent preferences. Both assume that decision makers can exercise some control over the likely realization of events.

Drèze does not specify the means by which this control is exercised, relying instead on their manifestation in the decision maker's choice behaviour. In particular, departing from Anscombe and Aumann's (1963) ‘reversal of order’ assumption, Drèze assumes that decision makers strictly prefer that the uncertainty of the lottery payoff be resolved before that of the acts, presumably to allow them to exploit this information by taking action to affect the likely realization of the underlying states. Drèze obtains a unique separation of state-dependent utilities from a set of probability distributions over the set of states of nature. Choice is represented as expected utility maximizing behaviour where the expected utility associated with any given act is itself the maximal expected utility with respect to the probabilities in the set.

Karni (2005b) replaces the state space with a set of effects – phenomena on which decision makers can place bets and whose realization they can influence by their actions. In Karni's theory the choice set consists of action-bet pairs. Actions affect the decision maker's well-being directly (for example, actions may correspond to levels of effort) and indirectly (through their impact on the decision maker's beliefs); bets are functions from effects to monetary payoffs. Karni gives necessary and sufficient conditions for the existence of subjective expected utility representations with unique, action-dependent, subjective probabilities; effect-dependent utility functions representing the evaluation of wealth; and a distinct function that captures the direct impact of the choice of action on the decision maker's well-being.

Attitudes toward risk

As with state-independent preferences, the economic analysis of many decision problems involving state-dependent preferences requires measures of risk aversion. Such measures are developed in Karni (1985).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- expected utility hypothesis
- Savage's subjective expected utility model
- uncertainty

Bibliography

Anscombe, F. and Aumann, R. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 43, 199–205.

Arrow, K. 1971. *Essays in the Theory of Risk Bearing*. Chicago: Markham Publishing Co.

Arrow, K. 1974. Optimal insurance and generalized deductibles. *Scandinavian Actuarial Journal*, 1–42.

Cook, P. and Graham, D. 1977. The demand for insurance and protection: the case of irreplaceable commodities. *Quarterly Journal of Economics* 91, 143–56.

Drèze, J. 1961. Les fondements logiques de l'utilité cardinale et de la probabilité subjective. *La Décision*. Colloques Internationaux de CNRS.

Drèze, J. 1987. Decision theory with moral hazard and state-dependent preferences. In *Essays on Economic Decisions under Uncertainty*. Cambridge: Cambridge University Press.

Eisner, R. and Strotz, R. 1961. Flight insurance and the theory of choice. *Journal of Political Economy* 69, 355–68.

Fishburn, P. 1973. A mixture-set axiomatization of conditional subjective expected utility. *Econometrica* 41, 1–25.

Grant, S. and Karni, E. 2004. A theory of quantifiable beliefs. *Journal of Mathematical Economics* 40, 515–46.

Karni, E. 1985. *Decision Making under Uncertainty: The Case of State-Dependent Preferences*. Cambridge, MA: Harvard University Press.

Karni, E. 2003. On the representation of beliefs by probabilities. *Journal of Risk and Uncertainty* 26, 17–38.

Karni, E. 2006. Subjective expected utility theory without states of the world. *Journal of Mathematical*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Economics 42(3), 168–88.

Karni, E. 2007. Foundations of Bayesian theory. *Journal of Economic Theory* 132(1), 168–88.

Karni, E. and Mongin, P. 2000. On the determination of subjective probability by choice. *Management Science* 46, 233–48.

Karni, E. and Schmeidler, D. 1981. An expected utility theory for state-dependent preferences. Working Paper No. 48–80. Foerder Institute for Economic Research, Tel Aviv University.

Karni, E., Schmeidler, D. and Vind, K. 1983. On state-dependent preferences and subjective probabilities. *Econometrica* 51, 1021–31.

Karni, E. and Zilcha, I. 1985. Uncertain lifetime, risk aversion and life insurance. *Scandinavian Actuarial Journal*, 109–23.

Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley.

Skiadas, C. 1997. Subjective probability under additive aggregation of conditional preferences. *Journal of Economic Theory* 76, 242–71.

Wakker, P. 1987. Subjective probabilities for state-dependent continuous utility. *Mathematical Social Sciences* 14, 289–98.

Yaari, M. 1965. Uncertain lifetime, life insurance and the theory of the consumer. *Review of Economic Studies* 32, 137–50.

How to cite this article

Karni, Edi. "state-dependent preferences." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000246>
doi:10.1057/9780230226203.1608

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

statistical decision theory

James O. Berger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Decision-making can be done scientifically. One can assemble the available information in a form directly usable in decision-making, mathematically assess the consequences of decisions, and combine both to reach optimal decisions. This article discusses the basis of such scientific decision-making, explaining the key concepts of utility, prior information, and maximization of expected utility. Statistical decision theory enlarges the framework of decision-making to include ‘choice among statistical procedures’. We introduce and contrast the competing Bayesian and frequentist approaches to statistical decision theory.

Keywords

Bayesian decision theory; decision theory; frequentist decision theory; game theory; invariance principle; minimax optimality; minimax principle; risk; optimality; statistical decision theory; uncertainty; utility; Wald, A.

Article

Decision theory is the science of making optimal decisions in the face of uncertainty. Statistical decision theory is concerned with the making of decisions when in the presence of statistical knowledge (data) which sheds light on some of the uncertainties involved in the decision problem. The generality of these definitions is such that decision theory (we drop the qualifier ‘statistical’ for convenience) formally encompasses an enormous range of problems and disciplines. Any attempt at a general review of decision theory is thus doomed; all that can be done is to present a description of some of the underlying ideas.
Decision theory operates by breaking a problem down into specific components, which can be mathematically or probabilistically modelled and combined with a suitable optimality principle to determine the best decision. Section 1 describes the most useful breakdown of a decision problem – that into actions, a utility function, prior information and data. Section 2 considers the most important optimality principle for reaching a decision – the Bayes principle. The frequentist approach to decision theory is discussed in Section 3, with the minimax principle mentioned as a special case. Section 4 compares the various approaches.
The history of decision theory is difficult to pin down, because virtually any historical mathematically

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

formulated decision problem could be called an example of decision theory. Also, it can be difficult to distinguish between true decision theory and formally related mathematical devices such as least squares estimation. Decision theory became clearly formulated as a science through work of John von Neumann and Oscar Morgenstern, culminating in their book *Theory of Games and Economic Behavior* (1944), and Abraham Wald, culminating in his book *Statistical Decision Functions* (1950). (The books do discuss some of the earlier history of decision theory.) General introductions to decision theory can be found, at an advanced level, in Blackwell and Girshick (1954) and Savage (1954); at an intermediate level in Raiffa and Schlaifer (1961), Ferguson (1967), De Groot (1970), Berger (1985), and French and Rios Insua (2000); and at a basic level in Raiffa (1968), Lindley (1985), and Winkler (1972).

1 Elements of a decision problem

In a decision problem, the most basic concept is that of an action *a*. The set of all possible actions that can be taken will be denoted by *A*. Any decision problem will typically involve an unknown quantity or quantities; this unknown element will be denoted by θ.

Example 1: A company receives a shipment of parts from a supplier, and must decide whether to accept the shipment or to reject the shipment (and return it to the supplier as unsatisfactory). The two possible actions being contemplated are:

- a_1: accept the shipment, a_2: reject the shipment.
- Thus $A = \{a_1, a_2\}$. The uncertain quantity which is crucial to a correct decision is:
- θ =the proportion of defective parts in the shipment.

Clearly action a_1 is desirable when θ is small enough, while a_2 is desirable otherwise.

The key idea in decision theory is to attempt a quantification of the gain or loss in taking possible actions. Since the gain or loss will usually depend upon θ as well as the action *a* taken, it is typically represented as a function of both. In economics this function is generally called the *utility function*, following the work of Frank Ramsey in the 1920s, and is denoted by $U(\theta, a)$. It is to be understood as the gain achieved if action *a* is taken and θ obtains. (The scale for measuring 'gain' will be discussed later.) In the statistical literature it is customary to talk in terms of loss instead of gain, with typical notation $L(\theta, a)$ for the loss function. Loss is just negative gain, so defining $L(\theta, a) = -U(\theta, a)$ results in effective equivalence between the two formulations (whatever maximizes utility will minimize loss).

Example 1 (continued): The company determines its utility function to be given by:

$$U(\theta, a_1) = 1 - 10\theta, \; U(\theta, a_2) = -0.1.$$

To understand how these might be developed, note that if a_2 is chosen the shipment will be returned to the supplier and a new shipment sent out. This new shipment must then be processed, all of which takes time and money. The overall cost of this eventuality is determined to be 0.1 (on the scale being used). The associated utility is –0.1 (a loss is a negative gain). Note that this cost is fixed: that is, it does not depend on θ.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

When a_1 is chosen, quite different considerations arise. The parts will be utilized with, say, gain of 1 if none is defective. Each defective part will cause a reduction in income by a certain amount, however, so that the true overall gain will be 1 reduced by a linear function of the proportion of defectives. $U(\theta, a_1)$ is precisely of this form. The various constants in $U(\theta, a_1)$ and $U(\theta, a_2)$ are chosen to reflect the various importance of the associated costs.

The scale chosen for a utility function turns out to be essentially unimportant, so that any convenient choice can be made. If the gain or loss is monetary, a suitable monetary unit often can provide a natural scale. Note, however, that utility functions can be defined for any type of gain or loss, not just monetary. Thus, in ex 1, the use of defective parts could lead to faulty final products from the company, and affect the overall quality image or prestige of the company. Such considerations are not easily stated in monetary terms, yet can be important to include in the overall construction of the utility function. (For more general discussion of the construction of utility functions, see Berger, 1985.)

The other important component of a decision problem is the information available about θ. This information will often arise from several sources, substantially complicating the job of mathematical modelling. We content ourselves here with consideration of the standard statistical scenario where there are available (*a*) *data*, *X*, from a statistical experiment relating to θ; and (*b*) background or *prior* information about θ, to be denoted by $\pi(\theta)$. Note that either of these components could be absent.

The data, *X*, is typically modelled as arising from some *probability density* $p_\theta(X)$. This, of course, is to be interpreted as the probability (or probability density) of the particular data value when θ obtains.

Example 1 (continued): It is typically too expensive (or impossible) to test all parts in a shipment for defects, so that a statistical sampling plan is employed instead. This generally consists of selecting, say, *n* random parts from the shipment, and testing only these for defects. If *X* is used to denote the number of defective parts found in the tested sample, and if *n* is fairly small compared with the total shipment size, then it is well known that $p_\theta(X)$ is approximately the binomial density:

$$p_\theta(X) = \frac{n!}{X!(n-X)!}\theta^X(1-\theta)^{n-X}.$$

The *prior* information about θ is typically also described by a probability density $\pi(\theta)$. This density is the probability (or mass) given to each possible value of θ in the light of beliefs as to which values of θ are most likely.

Example 1 (continued): The company has been receiving a steady stream of shipments from this supplier and has recorded estimates of the proportion of defectives for each shipment. The records show that 30 per cent of the shipments had θ between 0.0 and 0.025, 22 per cent of the shipments had θ between 0.025 and 0.05, 15 per cent had θ between 0.05 and 0.075, 11 per cent had θ between 0.075 and 0.10, 13 per cent had θ between 0.10 and 0.15, and the remaining 9 per cent had θ bigger than 0.15. Treating the varying θ as random, a probability density which provides a good fit to these percentages is the beta (1,14) density given (for $0 \le \theta \le 1$) by:

$$\pi(\theta) = 14(1-\theta)^{13}.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(For example, the probability that a random θ from this density is between 0.0 and 0.025 can be calculated to be 0.30, agreeing exactly with the observed 30 per cent.) It is very reasonable to treat θ for the current shipment as a random variable from this density, which we will thus take as the prior density.

2 Bayesian decision theory

When θ is known, it is a trivial matter to find the optimal action: simply maximize the gain by maximizing $U(\theta, a)$ over a. When θ is unknown, the natural generalization is to first 'average' $U(\theta, a)$ over θ, and then maximize over a. The correct method of 'averaging over θ' is to determine the overall probability density of θ, to be denoted $\pi^*(\theta)$ (and to be described shortly), and then consider the *Bayesian expected* utility:

$$U^*(a) = E^{\pi^*}[U(\theta, a)] = \int U(\theta, a)\pi^*(\theta)\mathrm{d}\theta.$$

(This last expression assumes that θ is a continuous variable taking values in an interval of numbers. If it can assume only one of a discrete set of values, then this integral should be replaced by a sum over the possible values.) Maximizing $U^*(a)$ over a will yield the optimal *Bayes action*, to be denoted by a^*.
Example 1 (continued): Initially, assume that no data, X, are available from a sampling inspection of the current shipment. Then the only information about θ is that contained in the prior $\pi(\theta)$; $\pi^*(\theta)$ will thus be identified with $\pi(\theta) = 14(1-\theta)^{13}$. Calculation yields:

$$U^*(a_1) = \int_0^1 (1-10\theta)\,14(1-\theta)^{13}\mathrm{d}\theta = 0.33,\ U^*(a_2) = \int_0^1 (-0.1)\,14(1-\theta)^{13}\mathrm{d}\theta = -0.1.$$

Since $U^*(a_1) > U^*(a_2)$, the Bayes action is a_1, to accept the shipment.
When data, X, are available, in addition to the prior information, the overall probability density π^* for θ must combine the two sources of information. This is done by *Baye's theorem* (from Bayes, 1763), which gives the overall density, usually called the *posterior density*, as:

$$\pi^*(\theta) = p_\theta(X)\cdot\pi(\theta)\,/\,m(X),$$

where:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$m(X) = \int p_\theta(X)\pi(\theta)\mathrm{d}\theta$$

(or a summation over θ if θ assumes only a discrete set of values), and $p_\theta(X)$ is the probability density for the experiment with the observed values of the data X inserted.
Example 1 (continued): Suppose a sample of $n = 20$ items is tested, out of which $X = 3$ defectives are observed. Calculation gives that the posterior density of θ is:

$$\pi^*(\theta) = p_\theta(3)\cdot\pi(\theta)\,/\,m(3) = \left[\frac{20!}{3!17!}\theta^3(1-\theta)^{17}\right]\cdot\left[14(1-\theta)^{13}\right]/\,m(3) = (185,504)\theta^3(1-\theta)^{30},$$

which can be recognized as the beta (4,31) density. This density describes the location of θ in the light of all available information. The Bayesian expected utilities of a_1 and a_2 are thus:

$$U^*(a_1) = \int_0^1(1-10\theta)\pi^*(\theta)\mathrm{d}\theta = \int_0^1(1-10\theta)(185,504)\theta^3(1-\theta)^{30}\mathrm{d}\theta = -0.14,$$

and

$$U^*(a_2) = \int_0^1(-0.1)\pi^*(\theta)\mathrm{d}\theta = -0.1\,.$$

Clearly a_2 now has the largest expected utility and should be the action chosen; in other words, the lot of parts should be rejected.

3 Frequentist decision theory

An alternative approach to statistical decision theory arises from taking a ‘long run’ perspective. The idea is to imagine repeating the decision problem a large number of times, and to develop a decision strategy which will be optimal in terms of some long-run criterion. This is called the *frequentist* approach, and is essentially due to Neyman, Pearson and Wald (see Neyman and Pearson, 1933; Neyman, 1977; Wald, 1950).
To formalize the above idea, let $d(X)$ denote a decision strategy or decision rule. The notation reflects the fact that we are imaging repetitions of the decision problem which will yield possibly different data X, and must

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
therefore specify the action to be taken for any possible X. The utility of using $d(X)$ when θ obtains is thus $U(\theta, d(X))$. The statistical literature almost exclusively works with loss functions instead of utility functions; for consistency with this literature we will thus use the loss function $L(\theta, d) = -U(\theta, d)$. (Of course, we want to minimize loss.)

The first step in a frequentist evaluation is to compute the *risk function* (expected loss over X) of d, given by:

$$R(\theta, d) = E_\theta\left\{L(\theta, d(X))\right\} = \int L(\theta, d(X))\, p_\theta(X)\, dX.$$

(Again, this integral should be a summation if X is discrete valued.) For a fixed θ this risk indicates how well $d(X)$ would perform if utilized repeatedly for data arising from the probability density $p_\theta(X)$. For various common choices of L this yields familiar statistical quantities. For instance, when L is 0 or 1, according to whether or not a correct decision is made in a two action hypothesis testing problem, the risk becomes the 'probabilities of type I or type II errors'. When L is 0 or 1, according to whether or not an interval $d(X)$ is contains θ, the risk is 1•minus the 'coverage probability function' for the confidence procedure $d(X)$. When $d(X)$ is an estimate of θ and $L(\theta, d) = (\theta - d)^2$, the risk is the 'mean squared error' commonly considered in many econometric studies. (If the estimator $d(X)$ is unbiased, then this mean squared error is also the variance function for d.)

Example 2: Example 1, involving acceptance or rejection of the shipment, is somewhat too complicated to handle here from the frequentist perspective; we thus consider the simpler problem of merely estimating θ (the proportion of defective parts in the shipment). Assume that loss in estimation is measured by *squared error*; that is:

$$L(\theta, d(X)) = [\theta - d(X)]^2.$$

A natural estimate of θ, based on X (the number of defectives from a sample of size n), is the sample proportion of defectives $d_1(X) = X / n$. For this decision rule (or *estimator*), the risk function when X has the binomial distribution discussed earlier (so that X takes only the discrete values 0,1,2, …n) is given by:

$$R(\theta, d) = \sum_{X=0}^{n} \left(\theta - \frac{X}{n}\right)^2 p_\theta(X) = \theta(1 - \theta) / n.$$

The second step of a frequentist analysis is to select some criterion for defining optimal risk functions (and hence optimal decision rules). One of the most common criteria is the *minimax principle*, which is based on consideration of the maximum possible risk:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$R^*(d) = \max_{\theta} R(\theta, d).$$

This indicates the worst possible performance of $d(X)$ in repeated use, and hence has some appeal as a criterion based on a cautious attitude. Using this criterion, an optimal decision rule is, of course, defined as one which minimizes $R^*(d)$, and is called a *minimax decision rule*.
Example 2 (continued): It is easy to see that:

$$R^*(d_1) = \max_{\theta} R(\theta, d_1) = \max_{\theta} \frac{\theta(1-\theta)}{n} = \frac{1}{4n}.$$

However, d_1 is not the minimax decision rule. Indeed, the minimax decision rule turns out to be:

$$d_2(X) = \left(X + \sqrt{n}/2\right) / \left(n + \sqrt{n}\right),$$

which has $R^*(d_2) = 1/[4(1+\sqrt{n})^2]$ (compare with Berger, 1985, p. 354). The minimax criterion here is essentially the same as the minimax criterion in game theory. Indeed, the frequentist decision problem can be considered to be a zero-sum two-person game with the statistician as player II (choosing $d(X)$), an inimical 'nature' as player I (choosing θ), and payoff (to player I) of $R(\theta, d)$. (Of course, it is rather unnatural to assume that nature is inimical in its choice of θ.) (For further discussion of this relationship, see Berger, 1985, ch. 5.)
Minimax optimality is but one of several criteria that are used in frequentist decision theory. Another common criterion is the invariance principle, which calls for finding the best decision rule in the class of rules which are 'invariant' under certain mathematical transformations of the decision problem. (See Berger, 1985, ch. 6, for discussion.)
There also exist very general and elegant theorems which characterize the class of acceptable decision rules. The formal term used is 'admissible': a decision rule, d, is *admissible* if there is no decision rule, d^*, with $R(\theta, d^*) \le R(\theta, d)$, the inequality being strict for some θ. If such a d^* exists, then d is said to be *inadmissible*, and one has obvious cause to question its use. Very common decision rules, such as the least squares estimator in three or more dimensional normal estimation problems (with sum of squares error loss), can turn out rather astonishingly to be inadmissible, so this avenue of investigation has had a substantial impact on decision theory. A general discussion, with references, can be found in Berger (1985).

4 Comparison of approaches

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

For solving a real decision problem, there is little doubt that the Bayesian approach is best. It incorporates all the available information (including the prior information, $\pi(\theta)$, which the frequentist approach ignores), and it tends to be easier than the frequentist approach by an order of magnitude. Maximizing $U^*(a)$ over all actions is generally much easier than minimizing something like $R^*(d)$ over all decision rules; the point is that, in some sense, the frequentist approach needlessly complicates the issue by forcing consideration of the right thing to do for each possible X, while the Bayesian worries only about what to do for the actual data X that are observed. There are also fundamental axiomatic developments (see, Ramsey, 1931; Savage, 1954; and Fishburn, 1981, for a general review) which show that only the Bayesian approach is consistent with plausible axioms of rational behaviour. Basically, the arguments are that situations can be constructed in which the follower of any non-Bayesian approach, say the minimax analyst, will be assured of inferior results.
Sometimes, however, decision theory is used as a formal framework for investigating the performance of statistical procedures, and then the situation is less clear. In Example 2, for instance, we used decision theory mainly as a method to formulate rigorously the problem of estimating a binomial proportion θ. If one is developing a statistical rule, $d(X)$, to be used for binomial estimation problems in general, then its repeated performance for varying X is certainly of interest. Furthermore, so the argument goes, prior information may be unavailable or inaccessible in problems where routine statistical analyses (such as estimating a binomial proportion θ) are to be performed, precluding use of the Bayesian approach.
The Bayesian reply to these arguments is that (*a*) optimal performance for each X alone will guarantee good performance in repeated use, negating the need to consider frequentist measures explicitly; and (*b*) even when prior information is unavailable or cannot be used, a Bayesian analysis can still be performed with so called objective prior densities: see Bernardo and Smith (1994), and Berger (2006).
Example 2 (continued): If no prior information about θ is available, one might well say that choosing $\pi(\theta) = 1$ reflects this lack of knowledge about θ. A Bayesian analysis (calculating the posterior density and choosing the action with smallest Bayesian expected squared error loss) yields, as the optimal estimate for θ when X is observed:

$$d_3(X) = (X + 1) / (n + 2).$$

This estimate is considerably more attractive than, say, the minimax rule $d_2(X)$ (see Berger, 1985, p. 375).
In practical applications of decision theory, it is the Bayesian approach which is dominant, yet the frequentist approach retains considerable appeal among theoreticians. A general consensus on the controversy appears quite remote at this time. This author sides with the Bayesian approach in the above debate, while recognizing that there are some situations in which the frequentist approach might be useful. For an extensive discussion of these issues, see Berger (1985); Chernoff and Moses (1959) provide an insightful introduction.

See Also

- Bayesian statistics
- decision theory in econometrics
- game theory

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- utility

Bibliography

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53, 370–418.

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, J. 2006. The case for objective Bayesian analysis. *Bayesian Analysis* 1, 385–402.

Bernardo, J. and Smith, A. 1994. *Bayesian Theory*. New York: John Wiley.

Blackwell, D. and Girshick, M.A. 1954. *Theory of Games and Statistical Decisions*. New York: Wiley.

Chernoff, H. and Moses, L. 1959. *Elementary Decision Theory*. New York: John Wiley.

De Groot, M.H. 1970. *Optimal Statistical Decisions*. New York: McGraw-Hill.

Ferguson, T.S. 1967. *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Fishburn, P.C. 1981. Subjective expected utility: a review of normative theories. *Theory and Decision* 13, 139–99.

French, S. and Rios Insua, D. 2000. *Statistical Decision Theory*. London: Arnold.

Lindley, D.V. 1985. *Making Decisions*. New York: Wiley.

Neyman, J. 1977. Frequentist probability and frequentist statistics. *Synthese* 36, 97–131.

Neyman, J. and Pearson, E.S. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society* 231, 289–337.

Raiffa, H. 1968. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Reading, MA: Addison-Wesley.

Raiffa, H. and Schlaifer, R. 1961. *Applied Statistical Decision Theory*. Boston: Division of Research, Graduate School of Business Administration, Harvard University.

Ramsey, F.P. 1931. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*. London: Kegan, Paul, Trench and Trubner. Reprinted in *Studies in Subjective Probability*, ed. H. Kyburg and H. Smokler, New York: Wiley, 1964.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wald, A. 1950. *Statistical Decision Functions*. New York: Wiley.

Winkler, R.L. 1972. *An Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart & Winston.

How to cite this article

Berger, James O. "statistical decision theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000251> doi:10.1057/9780230226203.1610

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Eco

statistical discrimination

Andrea Moro
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Statistical discrimination is a theory of inequality between demographic groups based on stereotypes that do not arise from prejudice or racial and gender bias. When rational, information-seeking decision makers use aggregate group characteristics, such as group averages, to evaluate individual personal characteristics, individuals belonging to different groups may be treated differently even if they share identical observable characteristics in every other aspect. Discrimination can be the agents' efficient response to asymmetric beliefs, or discriminatory outcomes may display an element of inefficiency: the disadvantaged group could perform better if beliefs were not asymmetric across groups (but beliefs are asymmetric because the disadvantaged are not performing as well as the dominant group).

Keywords

discrimination; inequality; statistical discrimination; wage determination

Article

Statistical discrimination is a theory of inequality between demographic groups based on stereotypes that do not arise from prejudice or racial and gender bias. It occurs when rational, information-seeking decision makers use aggregate group characteristics to evaluate relevant personal characteristics of the individuals with whom they interact. Because group-level statistics, such as group averages, are used as a proxy for the individual variables, individuals belonging to different groups may be treated differently even if they share

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

identical observable characteristics in every other aspect.
Some examples may help clarify the concept.

- When an insurance company chooses life insurance premia, the customer's likelihood of dying is one of the most relevant variables affecting the company profitability. A seemingly irrelevant proxy, gender, is highly correlated with death frequencies at every age. It is therefore optimal for the company to adopt a policy setting different premia for men and women who share similar characteristics.
- Employers usually place value on job attachment, for reason such as the costs of specific human capital investment. Historically, women have had lower labour market attachment than men, perhaps because of a higher propensity to be involved directly in child-rearing. In evaluating workers with otherwise identical characteristics, employers may prefer to hire male over identical female candidates. This is because employers assess probabilistically higher profitability from hiring a man.
- Highway police are often accused of searching cars driven by minorities more frequently than other cars. While such a policy may be viewed as unfair, it may not be the outcome of prejudice if police officers hold (perhaps biased) beliefs that minorities are more likely to engage in criminal activities. They use such beliefs to maximize the probability of arrest over a given time frame.

All of the above examples share the following features: the decision maker a) is a rational utility maximizing agent engaged in perfecting the available information; b) has incomplete information about some outcome-relevant individual characteristics; and c) holds asymmetric beliefs regarding the average value of relevant variables across groups. These beliefs can be interpreted as stereotypes.
The examples are also different in one important aspect. In the first example, differences in the company's beliefs are rooted in a ‘technological’, exogenous difference between groups, the death frequencies. In the other examples, asymmetric beliefs may feed back into differences in the type of behaviour that generates them. To clarify: it is possible that, because employers are less likely to hire women in jobs that require labour market attachment, then women are more likely to be involved in child-rearing than men, and are less prone to acquire the skills that are necessary to seek and perform well in those jobs, confirming the asymmetric belief that employers hold regarding labour

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

market attachment. In this case, beliefs are endogenous, or self-confirming. The distinction between these two sources of inequality is important. In the first case, discrimination is the agents' efficient response to asymmetric beliefs. In the second case, discriminatory outcomes may display an element of inefficiency: the disadvantaged group could perform better if beliefs were not asymmetric across groups (but beliefs are asymmetric because the disadvantaged are not performing as well as the dominant group).
We proceed by presenting in greater detail two additional labour-market-related examples of these two flavours of statistical discrimination.

Discrimination when the quality of information differs exogenously across groups

Consider the example of an employer that does not observe with certainty the skill level of her prospective employees. The population of workers of a given group has a skill distribution Φ. Workers draw from Φ their skill p, which is assumed to be equal to the value of their product when employed. Employers know the distribution, but only observe a noisy signal of productivity, $s=p+e$, where e is a zero-mean error distributed according to Φ_e. The employer infers p from s using the available information. In equilibrium, if the labour market is competitive and all employers share the same type of information, workers are paid according to their expected productivity conditional on the value of the signal. It can be shown that if Φ is a normal distribution, with mean μ and standard deviation σ, and if the error is also normal, with mean 0 and variance σ_e, then, using conditional expectations, the employers' best estimate of p is a weighted average of the workers' signal and the population average, with weights that depend on the relative size of the variance of the skill and the error distribution (see Phelps, 1972). Formally, the expected worker's productivity conditional on the signal is:

$$E(p|s) = \frac{\sigma_e^2}{\sigma^2 + \sigma_e^2} p + \frac{\sigma^2}{\sigma^2 + \sigma_e^2} s \qquad (1)$$

Intuitively, if the signal is very noisy (that is, if the variance of e is very high),

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the expected conditional value of workers productivity is close to the population average regardless of the signal's value. On the other extreme, if the signal is very precise (σ_e is close to zero), then the signal provides a precise estimate of the worker's ability.

If workers belong to two identifiable demographic groups, then it is rational for the employer to condition her inference also on the group identity of the worker. Suppose for example that the signal emitted by minorities is 'noisier' than the signal of non-minorities (perhaps because tests are race-biased). Then it follows that minorities with high signals will receive lower wages than same-signal workers from the dominant group, and the opposite happens to workers with low signals.

While this model is capable of explaining differential treatment for same-signal workers from different groups, on average workers of the two groups receive the same wage, which is equal to the average productivity μ. Group differences between groups' average wages can be obtained by extending the model to include employers' risk aversion (Aigner and Cain, 1977), or workers' pre-market investment in human capital (Lundberg and Startz, 1983). All of these approaches still require the assumption of some form of exogenous group difference, for instance in the signal's quality, an assumption which has been questioned.

Equilibrium discrimination with *ex-ante* identical groups

Recognizing these limitations, Arrow (1973) proposed an alternative model where inequality occurs even with identical groups' fundamentals. In his model, employers' asymmetric beliefs about members of different groups are self-confirming.

A formalization of this approach is presented in Coate and Loury (1993). There are two job-tasks, a simple task that anybody can perform, and a complex task requiring skills that can only be acquired through prior costly investment in human capital. Workers are heterogeneous in the cost of investment, but the cost distribution is the same across groups. There is a linear technology: each worker has the same productivity in the low-skill job, but workers have zero productivity in the high-skill job if they made no investment in human capital. If they do invest, their productivity is higher in the high-skill job than in the low-skill job. Wages are set exogenously and the high-skill job is paid with higher wages than the simple job.

Employers do not observe skill level, and assign workers to tasks according to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

an imperfect signal that is correlated with investment (this may be viewed as the outcome of a test where the likelihood of receiving a higher grade is higher for workers that have invested in human capital). No differences in the quality or informativeness of the test across groups need to be assumed. The task assignment depends on the expected skill, which depends both on the individual signal and on the group average, for reasons identical to those outlined in Phelps's model above, and formalized in equation (1). The optimal task assignment rule is to set a threshold and assign all workers with signals above such a threshold to the high-skill job, and workers below that threshold to the low-skill job. The marginal worker, with a signal equal to the threshold, has an expected productivity in the high-skill job identical to her productivity in the low-skill job.

A crucial feature of the model is that human capital investment provides a positive informational externality: investing not only helps a worker's own chances of being assigned to the high-skill job, but also increases the probability of being qualified to perform the high-skill job for every member of her group. This externality may generate multiple equilibria with different fractions of aggregate human capital investment.

For an intuition on how equilibria may be characterized, consider a model with only one group of workers. Suppose that in equilibrium only a few workers acquire human capital. In this case, employers' assessment of the probability that a worker has invested in human capital, given her signal, will be low even if the signal is relatively high. This is because Bayes's rule implies that the posterior probability of one worker having invested in human capital is increasing with the prior – the group's aggregate investment, (in equation (1) the workers' expected productivity depends on the group's average). Therefore, the optimal threshold for task assignment is set relatively high. Because it is difficult to obtain a high-skill job, the expected wage gain from investing is small, so only the few workers with relatively low cost of investment will acquire human capital, which confirms the original assumption that few workers invest.

There may also be equilibria with a high fraction of workers acquiring human capital. In such equilibria, employers will set a lower threshold in order to assign workers to a high-skill job relative to the low-investment equilibrium. Because high-skill jobs are more accessible, this task assignment rule provides higher returns to investment in human capital, which is a necessary condition to support this as an equilibrium. Note that the threshold cannot be too low, because if the high-skill job becomes too easily accessible, there are no

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

incentives to invest in human capital.

Turning to the two-group model, all outcomes where both groups replicate one of the equilibria of the one-group model are equilibria of the two-group model. These symmetric equilibria display no inequality. However, if the base model displays multiple equilibria, and if groups coordinate on equilibria with different fractions of workers investing in human capital, the group with lower investment will exhibit lower average wages, a higher fraction of workers employed in the low-skill job, and a higher threshold required for assignment to the high-skilled job.

While such a model illustrates the possibility for inequality to arise even when groups are *ex ante* identical, it cannot predict which group will be discriminated against, or why a symmetric equilibrium was not selected. The linearity of the technology implies that groups are treated separately, as if they were living in different islands: expected marginal productivities of workers depend only on their own signal and on the aggregate investment of their own group. Therefore, in this environment, statistical discrimination exists because of a *coordination failure*: the disadvantaged group fails to coordinate on the 'good' equilibrium, but the dominant group has nothing to lose if the disadvantaged group could solve the coordination failure.

A version of the model with a more general technology provides an alternative source of discrimination in which groups have conflicting interests (Moro and Norman, 2004). Consider a production function exhibiting a complementarity between tasks. Then, the marginal product of a worker in each task is affected by aggregate investment in human capital in both groups. Specifically, the expected marginal product in the high-skill job of a given worker depends negatively on aggregate investment in human capital from members of the other group. This is because when more members of the other group acquire human capital, the higher aggregate availability of skills decreases the marginal product of a skilled worker. Hence, incentives to acquire skills decrease when more members of the other group acquire skills. The complementarity generates incentives for groups to *specialize*, and asymmetric equilibria may exist even if there is a unique symmetric equilibrium. While there is an element of self-fulfilling prophecy, asymmetric equilibria here are the result of specialization rather than coordination failure. In such equilibria, the discriminated group cannot coordinate on a better outcome without a simultaneous coordination on a worse outcome by the other group: the dominant group always gains from discrimination. While in this model there always exist symmetric equilibria, group size is a relevant

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

factor, and the roles of the two groups can be reversed only if group sizes are identical.

Empirical evidence of statistical discrimination

There is a vast body of literature documenting racial and gender wage inequality. (See the bibliographies for the articles in this *Dictionary* on black-white labour market inequality in the United States, and women's work and wages.) In such literature, group differences that cannot be explained by differences in observable characteristics are attributed to prejudicial preferences, and little attention has been devoted to testing whether statistical discrimination plays a role in determining such differences. The main problem is to find ways to identify, using available data, to what extent group differences are caused by prejudicial attitudes, or by asymmetric beliefs (self-confirming or otherwise) and incentives.

Altonji and Pierret (2001) observe that if firms statistically discriminate, then as firms learn over time about workers' productivity, differences on the observed variables should fall over time. The data supports this proposition. Another example of an ‘outcome-based’ approach in the identification of the source of discrimination is the study by Knowles, Persico and Todd (2001), testing discrimination against minorities in motor vehicle searches by police officers. They find no evidence of racial animus in the data.

Attempts to identify different sources of discrimination include experimental or quasi-experimental data. (See Anderson, Fryer and Holt, 2006 for a survey of the literature, which also includes sources from the psychology literature.) A different approach is to estimate statistical discrimination models directly. Moro (2003) finds that adverse equilibrium selection did not play a role in exacerbating wage inequality during the last part of the 20th century. Fang (2006) estimates a statistical discrimination model to assess the prevalence of a signalling component to the college wage premium. While the estimates match wage distributions reasonably well, they are not designed to answer questions about model validation.

See Also

- affirmative action
- black–white labour market inequality in the United States
- racial profiling
- taste-based discrimination

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- women's work and wages

Bibliography

Aigner, D.J. and Cain, G. 1977. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review* 30, 749–76.

Altonji, J.G. and Pierret, C.R. 2001. Employer learning and statistical discrimination. *Quarterly Journal of Economics* 116(1), 313–50.

Arrow, K.J. 1973. The theory of discrimination. In *Discrimination in Labor Markets*, ed. O. Ashenfelter and A. Rees. Princeton, N.J.: Princeton University Press.

Anderson, L., Fryer, R. and Holt, C. 2006. Discrimination: experimental evidence from psychology and economics. In *Handbook on Economics of Discrimination*, ed. W. Rogers. Cheltenham, UK/Northampton, Mass.: Edward Elgar.

Coate, S. and Loury, G.C. 1993. Will affirmative action policies eliminate negative stereotypes? *American Economic Review* 83, 1220–40.

Fang, H. 2006. Disentangling the college wage premium: estimating a model with endogenous education choices. *International Economic Review* 47(4), 1151–85.

Knowles, J., Persico, N. and Todd, P. 2001. Racial bias in motor vehicle searches: theory and evidence. *Journal of Political Economy* 109(1), 203–29.

Lundberg, S.J. and Startz, R. 1983. Private discrimination and social intervention in competitive labor markets. *American Economic. Review* 73, 340–7.

Moro, A. 2003. The effect of statistical discrimination on black-white wage inequality: estimating a model with multiple equilibria. *International Economic Review* 44(2), 457–500.

Moro, A. and Norman, P.A. 2004. A general equilibrium model of statistical discrimination. *Journal of Economic Theory* 114(1), 1–30.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Phelps, E.S. 1972. The statistical theory of racism and sexism. *American Economic Review* 62(4), 659–61.

How to cite this article

Moro, Andrea. "statistical discrimination." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_S000544> doi:10.1057/9780230226203.1905

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

statistical inference

D.V. Lindley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic facts are typically established by means of statistical inference, which is concerned with how data supporting general claims about the world deduced from economic models should be calculated. Normally statistical inference proceeds by either estimation (point or interval) or testing. There is no general agreement on how statistical inference should be performed, but in some common situations there is good agreement about the numerical results. Of the three main schools – named after Fisher; Neyman, Pearson and Wald (NPW); and Bayes – the Bayesian alone is logically coherent.

Keywords

analysis of variance; Bayes' theorem; causality in economics; de Finetti, B.; estimation; exchangeability; Fisher, R.A.; frequentist probability; interval estimates; least-squares estimates; Lehmann, E.; Likelihood; Lindley, D. V.; log-likelihood functions; maximum likelihood; minimax principle; Neyman, J.; nuisance parameters; Pearson, E.S.; point estimates; probability distribution; Savage, L. J.; significance tests; statistical inference; subjective probability; testing; Wald, A.

Article

Deduction is the process whereby we pass from a general statement to a particular case: the reverse procedure, from the particular to the general, is variously called induction, or inference. Statistical inference is ordinarily understood to involve repetition or averaging, as when an inference is made about a population on the basis of a sample drawn from it. Economic facts are typically established by means of statistical inference. Economists construct a model of the world and deduce from it implications for the real world. These are checked against the available data, leading to some degree of support for the model. Statistical inference is concerned with how this support should be calculated.

Statistical inference incorporates a parameter θ which describes the model. In the simplest cases θ is a real number but in many models it is a set of numbers or even a function. The other basic element is the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
data *x* being observations made on the actual economic system. So θ corresponds to the general element and *x* to the particular. The model describes how the data follow from the parameter value. This is usually in the form of a probability distribution $p(x|\theta)$: the probability of *x*, given the value of θ. The problem of statistical inference is to make some statement about θ given the value of *x*. A simple example is provided by a model that says one variable *y* has linear regression on another *z*, the regression line having equation $y = \alpha + \beta z$ and the parameter being the pair $(\alpha, \beta) = \theta$. Data may then be collected for several pairs (y_i, z_i), $i = 1, 2, \ldots, n$ and an inference made about θ. The probability specification will ordinarily be that, for any *z*, *y* is normally distributed about $\alpha + \beta z$ with constant variance σ^2. If σ^2 is unknown then it will need to be included with α and β in θ.

Two types of inference statement are ordinarily made about θ: estimation and testing. The main distinction being that in testing some values of θ are singled out for special consideration, whereas in estimation all values of θ are treated equally. In the regression example, the hypothesis may be made that *z* does not affect *y* in the sense that $\beta = 0$. It would then be usual to test the hypothesis $\beta = 0$. In estimation, on the other hand, $\beta = 0$ plays no special role and the reasonable values of β on the basis of *x* are required. Estimation takes two forms, point and interval. In the former θ is estimated by a single number, the point estimate; or in the multidimensional case by a set of numbers. In the latter an interval, or region, of values of θ which are reasonably supported by the data is given. In the regression example

$$b = \sum (y_i - y_.)(z_i - z_.) / \sum (z_i - z_.)^2$$

is the least-squares point estimate of β, *y*. and *z*. being the means of the *y*- and *z*- values respectively. An interval estimate would be of the form $b \pm ts$, where *s* is the standard deviation evaluated from the data and *t* is the value obtained from Student's *t*-distribution. Point estimates are usually inadequate because they do not include any expression of the uncertainty that exists about the parameter: interval estimates are much to be preferred and usually, as in the regression case, start with the point estimate *b* and construct the interval about it. Interval estimates and tests are often related by the fact that the interval contains those parameter values which would not be judged significant were a test of that value to be carried out.

There is no general agreement on how statistical inference should be performed though, in some common situations, there is good agreement about the numerical results. It is possible to recognize three main schools named after Fisher; Neyman, Pearson and Wald (NPW); and Bayes.

The Fisherian school is the least formalized and is the one most favoured by scientists, especially those on the biological side, in medicine and agriculture. Because of its lack of a strict mathematical structure it is the hardest to describe succinctly, yet, because of this it is often the easiest to use. The name is entirely apposite since it is essentially the creation of one man, R.A. Fisher (1925, 1935). Estimation is based on the log-likelihood function $L(\theta) = \log p(x|\theta)$. Here $p(x|\theta)$, the probability of data *x* given parameter θ, is considered as a function of θ for the observed values of the data, now considered as fixed. A point estimate of θ is provided by the maximum likelihood value $\hat{\theta}$, that maximizes, over θ, $L(\theta)$. The precision of $\hat{\theta}$ can be found using minus the second derivative of $L(\theta)$ at $\hat{\theta}$. An interval

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

estimate is then of the form $\hat{\theta} \pm s$, where *s* depends on the measure of precision. Extensions to the multi-dimensional dimensional case are readily available and, although cases are known where the method is unsatisfactory, it often works extremely well and is deservedly popular. In the case of normal means, maximum likelihood and least-squares estimates agree. A Fisherian test of the hypothesis that θ is equal to a specified value θ_0 is found by constructing a statistic *t*(*x*) from the data *x* and calculating the probability, were $\theta = \theta_0$, of getting the value of *t*(*x*) observed, or more extreme. This probability is called the significance level: the smaller it is, the more doubt is cast on θ having the value θ_0. The best-known example is the *F*-test for the equality of means in an analysis of variance. It is typical of the Fisherian approach that few rules are available for the choice of the statistic *t*(*x*). His genius was enough to produce reasonable answers in important cases. Often *t*(*x*) is based on a point estimate of θ.

In some ways NPW is a formalized version of Fisher's approach. It has been much developed in the United States, though even there much applied work is Fisherian and it is the theoreticians who espouse NPW. There are many good expositions: for example, Lehmann (1959, 1983). Statistical inferences are thought of as decisions about θ and the merit of a decision is expressed in terms of a loss function measuring how bad the decision is when the true value is θ. If *t*(*x*) is a point estimate of the real parameter θ, squared error $\{t(x) - \theta\}^2$ is the loss function ordinarily used, the loss diminishing the nearer the estimate is to the true value. In testing, the decisions are to reject or to accept the null value θ_0 being tested. The simplest loss function is zero for a correct decision and some constant, positive value for each incorrect one. The probability of rejection of $\theta = \theta_0$ when in fact it is true is typically the significance level in Fisher's approach. Having the concept of a decision and a loss function, it becomes possible to ask the question, what is the best decision (estimate or test)? The criterion used to answer this is the expected loss, the expectation being over the data values according to the probability specification $p(x|\theta)$. Thus, for point estimate *t*(*x*), the expected loss is $\int\{t(x) - \theta\}^2 p(x|\theta)\mathrm{d}x$. The problem then is to choose *t*(*x*) to minimize this function. There is a substantial difficulty in that this expected loss depends on θ, which is unknown. Consequently additional criteria have to be used in order to select the optimum decision. For example, the decisions may be restrained in some way, as when a point estimate is restricted to be unbiased. A basic result is that the only sensible decisions are those which arise from the following procedure. Select a probability distribution *p*(θ) for θ and minimize the expected loss obtained by averaging over both *x* and θ – in the point estimation case, $\int\int\{t(x) - \theta\}^2 p(x|\theta)\, p(\theta)\mathrm{d}x\mathrm{d}\theta$. This expectation being a number, the minimization is usually possible without ambiguity. However, the choice of *p*(θ) remains to be made. It is important to notice that in NPW theory the distribution of θ is merely introduced as a device for producing a reasonable decision (the technical term is ‘admissible’) and is not necessarily held to express opinions about θ.

The third system of inference is named, quite inappropriately, after the discoverer of Bayes's,' theorem. Laplace was the first significant user. Inference is a passage from the special *x* to the general θ on the basis of a model $p(x|\theta)$ going in the opposite direction, from θ to *x*. In the Bayesian view, inference is similarly accomplished by a probability distribution $p(\theta|x)$ of θ, given *x*. The two distributions are related by Bayes's,' theorem, $p(\theta|x) \propto p(x|\theta)\, p(\theta)$, where *p*($\theta$) is a distribution for θ. NPW and Bayes are similar in their introduction of probabilities for θ. A basic difference is that the Bayesian approach recognizes *p*(θ) as a statement of belief about θ, and not, as does NPW, just as a technical device. With this strong statement about *p*(θ) both *x* and θ have probabilities attached and the full

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
force of the probability calculus can be employed: in particular, to make the inference $p(\theta|x)$. Now the inference is couched, not in terms of estimates or tests, but by means of a probability distribution. If $p(\theta|x)$ is centred around $t(x)$, say as its mean, then $t(x)$ may be conveniently thought of as a point estimate of θ. If θ_0 is of special interest $p(\theta_0|x)$ may be used as a test of the hypothesis that $\theta = \theta_0$. But the full inference is the distribution $p(\theta|x)$. Consequently, once the big step of introducing $p(\theta)$ has been made, the inference problem is solved by use of the probability calculus: no other considerations are needed. For example, typically θ is multi-dimensional $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ and only a few parameters are of interest, the remainder are called nuisance parameters. If only θ_1 matters, inferences about it are easily made by the marginal distribution $p(\theta_1|x)$ found by integrating out the nuisance parameters from $p(\theta|x)$. The regression example above for slope β (α and σ^2 being nuisance) provides an illustration.

Until World War I, Bayesian and non-Bayesian views had alternated in popularity, but the work of R.A. Fisher was so influential that it led to an almost complete suppression of the Bayesian view, which was reinforced by the work of Neyman, Pearson and Wald. Savage (1954) renewed interest in the Bayesian approach by providing it with its axiomatic structure, following Ramsey (1931) whose original ideas had lain unappreciated. Savage was much influenced by the work of de Finetti (his most accessible work is 1974/5) who provided a new view of probability that has had considerable impact upon subsequent thinking. Today the three disciplines lie uneasily together.

The Bayesian approach is the most formalized of the three inferential methods because everything is expressed within the single framework of the probability calculus, which is itself very well formalized. It has been relatively little used largely because of the perceived difficulty of assigning a distribution to θ. An important property of this method is that it is easily extended to include decision-making. As with NPW theory, a class of decisions d is introduced together with a loss function $l(d, \theta)$ expressing the loss in selecting d when θ obtains, and choosing that decision d that minimizes the expected (over θ) loss $\int l(d, \theta)\, p(\theta|x)\, d\theta$, using the inference $p(\theta|x)$. (This is in contrast to the NPW approach, using the expectation over x.)

There are two basic differences between the Bayesian paradigm and the other two. These concern the logical structure, and the likelihood principle. Both the Fisherian and NPW paradigms tackle an inference problem by thinking of several, apparently sensible procedures, investigating their properties and choosing that procedure which overall has the best properties. Fisher's work on maximum likelihood and its demonstrated superiority to the method of moments provides an example. In neither of these approaches are there general procedures: for example, there is no way known of constructing an interval estimate. Within NPW, Wald did introduce the minimax principle but it is generally unsatisfactory in the inference context and has not been used in practice. Against this, the techniques that are available, like maximum likelihood and analysis of variance, are easy to use and interpret (though the interpretation is often wrong: see equations (1) and (2) below). The lack of a formal structure has enabled statisticians to extemporize and come up with valuable concepts and techniques that are of substantial practical value though sometimes with weak justifications. The Bayesian paradigm proceeds differently. It begins by laying down reasonable, elementary properties to be demanded of an inference and then, by deduction, discovers which procedures have these properties. In that sense it is the complete opposite of the Fisherian and NPW views that start with the procedures. It is the method used in other branches of mathematics where the basic properties provide the axioms for the subsequent, logical development.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Though there are important variants, all the axiom systems proposed lead to the result that the only inference procedures satisfying them are those that use probability: that the only sensible inference for θ, given *x*, is a probability statement about θ, given *x*. The Bayesian position is therefore a deduction from simple requirements about our inferences. NPW comes near to recognizing this in its technical introduction of $p(\theta)$. The Fisherian view never addresses the problem.

The second difference between the Bayesian and other views involves the likelihood principle. The model provides $p(x|\theta)$ which, for fixed θ, is a probability for *x*. Considered as a function of θ for fixed *x*, it is called the likelihood for θ (given *x*). It was an important contribution of Fisher's to emphasize the distinction between the probability and likelihood aspects, and to show us, for example, in the maximum likelihood estimate, the importance of the likelihood function. However, Fisher did not consider the likelihood to be the only tool for inference. In a significance test, based on a statistic $t(x)$, he used the significance level, which is an integral over values of *x* giving more extreme values to *t* than that observed, for the tested value θ_0. Clearly this cannot be calculated from the likelihood function which holds *x* fixed and varies θ. NPW uses the expected (over *x*) loss and therefore does not use the likelihood function. On the other hand, the only feature of the data used in a Bayesian procedure is the likelihood, supplementing it with the distribution for θ. The likelihood principle says that if two data sets, *x* and *y*, have the same likelihood, then the inferences from *x* and *y* should be the same. Most statistical procedures in common use today violate the principle, but Bayesian procedures do not. The latter part of that statement is clearly true from Bayes's,' theorem which, in order to calculate the inference, uses only the likelihood. Here is an example of its violation when an unbiased estimate is used. Given θ, *x* is a random sample from a population in which each value is either 1 or 0 with probabilities θ and $1-\theta$. In one case the sample is selected to be of size *n* and *r* of the values are found to be 1. In the second case, *r* is chosen and the population sampled until *r* 1's have been observed, the total sample being of size *n*. In each case the likelihood is $\theta^r(1-\theta)^{n-r}$ and so, by the likelihood principle, the inferences should be the same. However, in the first case the unbiased estimate of θ based on (*r*, *n*) is the familiar *r/n*: in the second case it is $(r-1)/(n-1)$. Significance tests of $\theta = \frac{1}{2}$, say, are different in the two cases because ‘more extreme’ in one case means more extreme values of *r* for fixed *n*, and in the other more extreme values of *n* for fixed *r*. There are many impressive arguments in favour of the likelihood principle, even outside the Bayesian paradigm, yet it is not accepted by most statisticians and almost all inferential procedures used today violate it: maximum likelihood estimation is the obvious exception.

There is another interesting consequence of the axiomatic, Bayesian approach leading to the probabilistic form of inference, and that is that any non-probabilistic inference will somewhere violate the basic properties set out in the axioms. Indeed, it is true that every non-Bayesian procedure has a counter-example where it behaves in an absurd fashion. In illustration let $(l(x),u(x))$ be a confidence interval for θ at level α based on data *x*. The precise meaning of this is that

$$p(l(x) < \theta < u(x)|\theta) = \alpha, \text{ for all } \theta. \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Notice that this is a probability statement about x, given θ, based on $p(x|\theta)$. In words, the probability that the random interval $(l(x),u(x))$ contains θ is α, for all given θ. It is easy to produce examples for x in which the interval is the whole line; $l(x) = -\infty$, $u(x) = \infty$, and $\alpha = 0.95$. Here we are 95 per cent confident that θ is real. This is absurd in the case of the observed x, although it is true that for 95 per cent of x's the statement will be true. Contrast this with the Bayesian statement that

$$p(l(x) < \theta < u(x)|x) = \alpha, \text{ for all } x, \tag{2}$$

based on $p(\theta|x)$. This is about θ, given x: in words, the probability is α that θ lies between $l(x)$ and $u(x)$. Clearly, with $\alpha < 1$, it could never happen that the interval is the whole real line.

A key ingredient in any form of inference is clearly probability, whose laws are well understood. But there is considerable dispute over the interpretation of probability: disputes which have practical consequences. There are two broad groups: subjective and frequentist views. In the subjective view, a probability is an expression of the subject's belief. Thus (2) expresses a belief that θ lies between the numbers $l(x)$ and $u(x)$. In the frequentist view, probabilities are related to observed frequencies. Thus (1) says that the frequency with which the interval contains θ is α. The latter are objective, in the sense that the frequencies can be objectively observed by all subjects. The great majority of statisticians today adopt the frequency view, claiming an objectivity for their methods. Most Bayesians hold to the subjective approach, claiming that economists have to express beliefs about the system they are discussing. It is undoubtably true that many users of statistics think of the frequency statements, like (1), as belief statements, like (2). It had been thought that the two views were opposites but de Finetti showed that the frequentist view of probability is a special case of the subjective view, namely when the data are believed to be exchangeable. The values $x_1, x_2, \ldots, x_n$ are exchangeable if their probability distribution is invariant under permutation of the x's. A random sample from a population would ordinarily be judged to possess this invariance. The case mentioned earlier where each x_i is either 1 or 0, with probabilities θ and $1 - \theta$ respectively, is the standard example. Here θ is a frequency probability, or chance, about which there are beliefs $p(\theta)$ changed by the data x to new beliefs $p(\theta|x)$. Resistance to the Bayesian approach and subjective probability has centred around the genuine difficulty of assessing beliefs, especially when there is little knowledge of the parameter. Rather than face the formidable, and perhaps impossible, task of measuring belief, statisticians have concentrated on frequentist methods, sometimes ignoring their defects. A related difficulty with the subjective approach is the lack of objectivity in the sense that two subjects may, on the basis of the same data, have different beliefs. The Bayesian response is that this reflects reality and if each economist were to express all his beliefs probabilistically, we would have a clearer appreciation of the situation; and, in any case, different beliefs come together with increasing amounts of data. This is why observational studies are so important. Economics is predominantly frequentist but does have a substantial school, particularly in econometrics, of the Bayesian persuasion. The close connection between that view and decision-making

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
makes it more attractive to the economist than to the laboratory scientist who sees himself as acquiring knowledge, not making decisions.

Inference that is statistical, involving repetition, is naturally allied to the frequency view: whereas inference, in general, has no frequency basis. But de Finetti's observation connecting exchangeability (which is essentially a finite, frequentist property) with subjectivity shows that the Bayesian view embraces both statistical and non-statistical inference. Consequently, the subjective, Bayesian paradigm has enormous potential, encompassing almost all problems of passing from the special to the general. The guilt (corresponding to θ) of a defendant in a law court on the basis of evidence *x* is an example. The likelihood principle says that the only relevant features are the probabilities of the evidence on the assumptions of innocence, and of guilt. Whether this potential will be realized depends in large part on overcoming the practical difficulties of assessment of beliefs.

Statistical inference depends on a probability specification $p(x|\theta)$ for data *x*, given parameter θ. If NPW it uses, in addition, a loss function: if Bayesian it introduces an additional probability specification for θ, $p(\theta)$. An important topic studies how the inference is affected by changes in any of the specifications. The inference is said to be robust if the change has little effect on it. For example, it is usual to choose $p(x|\theta)$ to be normal, largely because this assumption is relatively easy to handle and leads to many, simple and powerful answers. We might ask what happens if the normal is replaced by the very similar Student's *t*-distribution with its rather longer tails. For the mean μ of the normal, the sample mean is, by any standard, an excellent point estimate of μ: a trimmed mean, in which a few extreme observations are discarded, is not quite as good but is still reasonable. With the *t*-distribution however, the situation is reversed and the trimmed mean behaves better than the sample mean. The former is more robust. Of recent years a lot of work has been put into the study of robust inference procedures to replace less robust ones like least squares.

The scientist who is able to plan his experiment, either in the laboratory or in the field, has a much simpler inference problem than the economist who, almost entirely, has to rely on data that have arisen naturally instead of being planned. The planned experiment can take cognizance of factors additional to those the scientist is directly interested in. This can be done either by explicitly including them in the experiment, or by a suitable randomization procedure that has a high chance of eliminating any unwanted effects. The economist is usually denied both opportunities, though sometimes extra factors can be included. The inference procedure should therefore recognize uncertainties that the laboratory experiment has eliminated. This is not always done and inference in economics remains less satisfactory than in other sciences. The concept of causation is harder to understand in economics. In the regression of *y* on *z* above, it is easy to think of *z* causing changes in *y*: but it may be that changes in *y* and *z* are both caused by related fluctuations in a third variable *w*. The attempt by econometricians to avoid this difficulty by including many variables has led to complexities of interpretation due to the high dimensionality of the problem.

Statistical inference is ordinarily thought of as a passage from data *x* to parameter θ but there is another form in which the inference is from past data *x* to future observations *y*, with no explicit reference to a parameter. This is often called prediction, and the obvious application is to time series with $x = (x_1, x_2, \ldots, x_n)$, x_i being the value of some quantity at time t_i and $y = x_{n+1}$; so that the quantity has been observed up to time t_n and it is required to predict its value at t_{n+1}. The usual way to proceed is to model the time series in some parametric form involving θ and to infer the value of θ on the basis of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

x. The model will specify $p(y|x, \theta)$ and one possibility is to predict y using $p(y|x, \hat{\theta})$, where $\hat{\theta}$ is a point estimate of θ. In the Bayesian view $p(y|x)$ is directly available for prediction, where $p(y|x) = \int p(y|x, \theta)\, p(\theta|x)\, d\theta$, using standard probability calculations. It is arguable that all practical inference problems are of this type and that the model, and θ, are only introduced as a means of solving them.

Many statistical procedures are complicated and require extensive computations: in some cases, one does not even know how to find a procedure. One possibility is to use approximation techniques and find a procedure which loses only little information in comparison with the optimum method and yet is simple. Asymptotic methods often provide such approximations. Data often consist of a random sample, or of a time series $(x_1, x_2, \ldots, x_n)$ involving n observations. It is often possible to study the limiting behaviour as n increases without limit. For example, with random samples, the maximum likelihood estimate θ is asymptotically normally distributed with mean equal to the true value and variance σ^2/n, with σ^2 calculable in terms of the second derivative of the loglikelihood. Although this is only true as n goes to infinity, it can be used to produce a 95 per cent confidence interval for θ of the form $\theta \pm 1.96\sigma / n^{1/2}$. This is then an approximation, for large n, to the exact interval. Asymptotic methods have been very successful though it is often difficult to know how fast the limit is approached and whether a particular n is large or not. Stirling's asymptotic formula is remarkably accurate for n as low as 3. Some asymptotic results are not realized until n is well into the thousands.

The present position in statistical inference is historically interesting. The bulk of practitioners use well-established methods like least squares, analysis of variance, maximum likelihood and significance tests: all broadly within the Fisherian school and chosen for their proven usefulness rather than their logical coherence. If asked about their rigorous justification most of these people would refer to ideas of the NPW type: least-squares estimates are best, linear unbiased; F-tests have high power and maximum likelihood values are asymptotically optimal. Yet these justifications are far from satisfactory: the only logically coherent system is the Bayesian one which disagrees with the NPW notions, largely because of their violation of the likelihood principle. The practitioner is most reluctant to adopt this logical approach because of its apparent impracticality. The impracticality is largely an illusion and current work is energetically overcoming it. So the next few decades should be interesting as the various theories get amended and one emerges triumphant, or some new ideas avoid the contradictions. Whatever happens, inference will surely remain one of the most important of subjects, simply because of the ubiquity of inference problems in all aspects of human endeavour.

See Also

- maximum likelihood

Bibliography

De Finetti, B. 1974–5. *Theory of Probability: a Critical Introductory Treatment*. 2 vols, London: Wiley. Translation from the 1970 Italian original by A. Machi and A. Smith.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fisher, R.A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R.A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Lehmann, E.L. 1959. *Testing Statistical Hypotheses*. New York: Wiley.

Lehmann, E.L. 1983. *The Theory of Point Estimation*. New York: Wiley.

Ramsey, F.P. 1931. *The Foundations of Mathematics and Other Logical Essays*. London: Kegan, Paul, Trench, Trubner.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

How to cite this article

Lindley, D. V. "statistical inference." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000252> doi:10.1057/9780230226203.1611

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

statistical mechanics

Steven N. Durlauf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Statistical mechanics models constitute a mathematical framework that is useful in describing the aggregate behaviour of interacting populations. While the methods originate in physics, they have proven useful in modelling socio-economic phenomena. This article describes the basic properties of statistical mechanics models and discusses their use in theoretical and empirical economics.

Keywords

Arrow–Debreu model of general equilibrium; conditional probability; continuous choice spaces; coordination games; diffusion of technology; discrete time models; emergence; evolutionary game theory; financial market fluctuations; game theory; information transmission; Kolmogorov extension theorem; law of large numbers; Markov processes; mathematics and economics; microfoundations; multinomial choice models; multiple equilibria; multiple steady states; nonergodicity; physics and economics; reflection problem; statistical mechanics; technical change; trade networks; unemployment; universality

Article

Statistical mechanics is a branch of physics which studies the aggregate behaviour of large populations of objects, typically atoms. A canonical question in statistical mechanics is how magnets can appear in nature. A magnet is a piece of iron with the property that atoms tend on average to be spinning up or down; the greater the lopsidedness the stronger the magnet. (Spin is binary.) While one explanation would be that there is simply a tendency for individual atoms to spin one way rather than another, the remarkable finding in the physics literature is that interdependences in spin probabilities between the atoms can, when strong enough, themselves be a source of magnetization. Classic structures of this type include the Ising and Currie–Weiss models (cf. Ellis, 1985).

Economists, of course, have no interest in the physics of such systems. On the other hand, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
mathematics of statistical mechanics has proven to be useful for a number of modelling contexts. As illustrated by the magnetism example, statistical mechanics models provide a language for modelling interacting populations. The mathematical models of statistical mechanics are sometimes called ‘interacting particle systems’ or ‘random fields’, where the latter term refers to interdependent populations with arbitrary index sets, as opposed to a variables indexed by time. It is the mathematics of statistical mechanics models that economists have found valuable in studying the evolution of populations.

Statistical mechanics models are useful to economists as these methods provide a framework for linking microeconomic specifications to macroeconomic outcomes. A key feature of a statistical mechanical system is that, even though the individual elements may be unpredictable, order appears at an aggregate level. At one level this is an unsurprising property; laws of large numbers provide a similar linkage. However, in statistical mechanics models properties can emerge at an aggregate level that are not describable at the individual level. Magnetism is one example of this as it is a feature of a system, not an individual element; the existence of aggregate properties without individual analogues is sometimes known as ‘emergence’. Emergent properties are in fact why statistical mechanics models appear to be such an intriguing set of tools for economists since they suggest that there may be aspects of macroeconomic outcomes that are not reducible to the microeconomic specification from which they derive. As such, emergence is a way to make progress on understanding aggregate behaviour in the presence of heterogeneous agents. This is especially important in light of results by Hugo Sonnenschein and others that show that the Arrow–Debreu type general equilibrium framework does not, by itself, impose many restrictions on which data can be observed. (See aggregate demand theory and aggregation (theory) for the basic results and different efforts to overcome this lack of empirical content to general equilibrium theory.) In order to produce empirical implications, statistical mechanics models impose stronger (and in many ways different) restrictions on individual interrelationships than are found in Arrow–Debreu models, so in this sense are clearly less general. What is interesting is that the aggregate properties of statistical mechanics models often do not depend on details of the interaction structure, a property known as ‘universality’.

The general structure of statistical mechanics models may be understood as follows. Consider a population of elements ω_i, where i is an element of some arbitrary index set I. Let $\underset{\sim}{\omega}$ denote vector all elements in the population and $\underset{\sim}{\omega}_{I-i}$ denote all the elements of the population other than i. Concretely, each ω_i may be thought of as an individual choice. A statistical mechanics model is specified by the set of probability measures

$$\mu(\omega_i|\underset{\sim}{\omega}_{I-i}) \tag{1}$$

for all i. These probability measures describe how each element of a system behaves given the behaviour of other elements. Following our example, (1) can be interpreted as describing the probabilities of a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

given person's choice given the choices of others. The objective of the analysis of the system is to understand the joint probability measures for the entire system,

$$\mu(\underset{\sim}{\omega}), \tag{2}$$

that are compatible with the conditional probability measures (1). Thus, the goal of the exercise is to understand the probability measure for the population of choices given the conditional decision structure for each choice. Stated this way, one can see how statistical mechanics models are conceptually similar to various game-theory models, an idea found in Blume (1993), who uses statistical mechanics methods to study the convergence properties of populations in which individual agents interact with their neighbours via a sequence of coordination games.

This formulation of statistical mechanics models, with conditional probability measures representing the micro-level description of the system, and associated joint probability measures the macro-level or equilibrium description of the system, also illustrates an important difference between physics and economics reasoning. For the physicist, treating conditional probability measures as primitive objects in modelling is natural. One does not ask ‘why’ one atom's behaviour reacts to other atoms. In contrast, conditional probabilities are not natural modelling primitives to an economist. The microeconomic foundations of a model (that is, the specification of preferences, technology, beliefs and possibly institutional framework) produce conditional probabilities as descriptions of how individuals behave in the environment. Hence, one does not start by taking as a given a conditional probability description that imposes the requirement that the likelihood that an individual student drops out of high school is an increasing function of the drop-out decisions of other students. Rather, one specifies a decision problem for the student in which peer influence matters, possibly through a direct desire to conform or via information communicated by the decisions of others. This decision problem will have a probabilistic structure in which the individual outcome ω_i depends on others in the population, just as in the standard statistical mechanics case, but the form of this dependence is derivative from the specification of the decision problem. A defect of a number of economic models using statistical mechanics is the tendency to follow the physics literature and treat (1) as an appropriate way to formulate microfoundations.

Dynamic versions of statistical mechanics models are usually modelled in continuous time. One considers the process $\omega_i(t)$ and, unlike the atemporal case, probabilities are assigned to at each point in time to the probability of a change in the current value. Operationally, this means that for sufficiently small δ

$$\mu(\omega_i(t+\delta)|\omega_i(t+\delta) \neq \omega_i(t)) = f(\underset{\sim}{\omega}_{i-I}(t), \omega_i(t))\delta + o(\delta). \tag{3}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

What this means is that at each t there is a small probability that $\omega_i(t)$ will change value; such a change is known as a flip when the support of $\omega_i(t)$ is binary. This probability is modelled as depending on the current value of element i as well as on the current (time t) configuration of the rest of the population. Since time is continuous whereas the index set is countable, the probability that two elements change at the same time is 0 when the change probabilities are independent. Systems of this type lead to the question of the existence and nature of invariant or limiting probability measures for the population, that is, the study of

$$\lim_{t \Rightarrow \infty} \mu(\underset{\sim}{\omega}(t) | \underset{\sim}{\omega}(0)). \tag{4}$$

Discrete time systems can of course be defined analogously; for such systems a typical element is $\omega_{i,t}$. In such cases, it perhaps most natural to assume that changes in the individual elements of the system are simultaneous.

The conditional probability structure described by (1) can lead to very complicated calculations for the joint probabilities (2). In the interests of analytical tractability, physicists have developed a set of methods referred to as mean field analyses. These methods typically involve replacing the conditioning elements in (1) with their expected values, that is

$$\mu(\omega_i | E(\underset{\sim}{\omega}_{l-i})). \tag{5}$$

A range of results exist on how mean field approximation relate to the original probabilities models they approximate. From the perspective of economic reasoning, mean field approximations have a substantive economic interpretation as they implicitly mean that agents make decisions based on their beliefs about the behaviours of others rather than the behaviours themselves. Brock and Durlauf (2001a) develop an environment in which the equilibrium set of choices, modelled as an expectational form of a noncooperative (that is, Nash) equilibrium, turns out to the mean field approximation of a model that is interpretable as a social planning equilibrium in which the planner determines all choices, accounting for complementarities in payoffs across individuals.

Properties

Existence

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The first question one naturally asks for environments of the type described concerns the existence of a joint or invariant probability measure over the population of elements in which conditional probabilities for the behaviours of the elements have been specified. Existence results of this type differ from classic results such as the Kolmogorov extension theorem in that they concern the relationship between conditional probabilities and joint probabilities, rather than the relationship (as occurs in the Kolmogorov case) between joint probabilities measured on finite sets of elements versus an infinite collection that represents the union of the various elements. Liggett (1985; 1991) provides a comprehensive survey of results. These results are quite technical but do not, in my judgement, require conditions that appear to be reasonable from the perspective of socio-economic systems, at least in the sense that they do not seem to have any interesting behavioural content.

Uniqueness or multiplicity

The existence of a joint or invariant measure says nothing about how many such measures exist. When there are multiple measures compatible with the conditional probabilities, the system is said to be nonergodic. Notice that for the dynamical models the uniqueness question involves the dependence of the invariant measure on the initial configuration on $\underset{\sim}{\omega}(0)$ or $\underset{\sim}{\omega}_0$. Heuristically, for atemporal models, nonergodicity is thus the probabilistic analog to multiple equilibria, whereas for temporal models nonergodicity is the probabilistic analog to multiple steady states.

One of the fascinating features of statistical mechanics models is their capacity to exhibit nonergodicity in nontrivial cases. Specifically, nonergodicity can occur when the various direct and indirect connections between individuals in a population create sufficient aggregate interdependence across agents. As such statistical mechanics models use richer interactions structure than appear, for example in conventional time series models. To see this, consider a Markov chain $\Pr(\omega_t = 1|\omega_{t-1} = 1) \neq 1$ and $\Pr(\omega_t = 0|\omega_{t-1} = 0) \neq 1$, then $\lim_{j \Rightarrow \infty} \Pr(\omega_{t+j}|\omega_0)$ will not depend on ω_0. However, suppose that $I = Q$, that is, the index set is the set of integers so that we are considering the evolution of a countable collection of elements. Suppose further that the system has a local Markov property of the form $\Pr(\omega_{i,t}|\underset{\sim}{\omega}_{t-1}) = \Pr(\omega_{i,t}|\omega_{i-1,t-1}\omega_{i,t-1}\omega_{i+1,t-1})$; in words, the behaviour of a particular $\omega_{i,t}$ depends on its value at $t-1$ as well as its 'nearest neighbours'. In this case, it is possible that $\lim_{j \Rightarrow \infty} \Pr(\omega_{i,t+j}|\underset{\sim}{\omega}_0)$ does depend on $\underset{\sim}{\omega}_0$ even though no conditional probability $\Pr(\omega_{i,t}|\omega_{i-1,t-1}\omega_{i,t-1}\omega_{i+1,t-1})$ equals 1. The reason for this is that, in the case of an evolving set of interacting Markov processes, there are many indirect connections. For example, the realization of $\omega_{i-2,t-2}$ will affect $\omega_{i,t}$ because of its effect on $\omega_{i-1,t-1}$; no analogous property exists when there is a single element at each point in time. In fact, the number of elements at time $t-k$ that affect $\omega_{i,t}$ is, in this example, growing in *k*. This does not mean that such a system necessarily has multiple invariant measures, merely that it can when there is sufficient sensitivity of $\Pr(\omega_{i,t}|\underset{\sim}{\omega}_{t-1}) = \Pr(\omega_{i,t}|\omega_{i-1,t}, \omega_{i,t-1}, \omega_{i+1,t-1})$ to the realizations of $\omega_{i-1,t}\omega_{i,t-1}$ and $\omega_{i+1,t-1}$. For many statistical mechanics models, this dependence can be reduced to a single parameter. For example, a dynamic version of the Ising model may be written

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$\Pr(\omega_{i,t}|\omega_{\sim t-1}) \propto e^{J(\omega_{i-1,t-1}+\omega_{i,t-1}+\omega_{i+1,t-1})}$; so J fully characterizes the degree of dependence in the system. In statistical mechanics models, one often finds threshold effects, that is, when J is below some $J < \bar{J}$, the system exhibits a unique invariant measure whereas, if $J > \bar{J}$, multiple measures exist.

Applications

The earliest uses of statistical mechanics models in economics appear to be Follmer (1974) and Allen (1982). Follmer analyses the question of when idiosyncratic preference shocks affect aggregate prices. He models these shocks as binary and shows that, if the shocks obey nearest neighbour-type interdependence, then it is possible for the shocks to affect the aggregate price level. This occurs specifically if the interdependences are strong enough to produce multiple invariant measures, that is, the law of large numbers breaks down for the shocks. Allen (1982) applies statistical mechanics ideas to the diffusion of technical change.

Statistical mechanics ideas were largely dormant until the early 1990s, when a number of researchers independently began using the tools, often in very different contexts. One area where this work has proven valuable is game theory. Blume (1993; 1995) employed statistical mechanics methods to understand the role of different interactions structures in evolutionary game theory. Brock (1993) provides a wide-ranging exploration of the relationship between various types of statistical mechanics models and particular socio-economic environments, with particular attention to the difference between environments with and without a social planner.

Other authors have applied statistical mechanics to particular substantive contexts. Durlauf (1993) employs a discrete time model to study economic growth. For this work, the motivation was twofold: first, to formalize the idea that local spillover effects can create a development trap and, second, to identify how leading sectors can expand and thereby lead to a take-off to sustained industrialization and growth. Bak et al. (1993) analyse a model in which industrial demand linkages can cause idiosyncratic shocks to produce aggregate fluctuations. Kelly (1994) shows how these models can explain how shocks of different sizes lead to very different macroeconomic consequences. Other applications include financial market fluctuations (Horst, 2005), information transmission (Kosfeld, 2005), technical change (Auerswald et al., 2000) and trade networks and unemployment (Oomes, 2003).

Current theoretical research using statistical mechanics models has attempted to extend their use to more general specifications than have appeared in the physics and mathematics literatures. Brock and Durlauf (2006) extend various properties of statistical mechanics models to contexts where agents face more than two choices; the choices are not ordered so this approach creates links between statistical mechanics models and multinomial choice models in the various social sciences. Other authors have focused on continuous choice spaces (for example, Bisin, Horst and Özgür, 2006; Horst and Scheinkman, 2006). Ioannides (2006) considers how a range of alternative interaction structures affect aggregate outcomes. Bisin, Horst and Özgür, (2006) consider issues of self-consistent beliefs for different local interaction structures. Horst and Scheinkman (2006) consider the relationship between conditional probabilities of the form (1) and the underlying decision problems of agents, thus facilitating better microfoundations when these methods are used. These various directions seem promising in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
allowing statistical mechanics methods to describe richer socio-economic environments.
Researchers have also begun to bring statistical mechanics models to empirical work. The econometric analyses of statistical mechanics models of Brock and Durlauf (2001a; 2001b; 2006; 2007), Conley and Topa (2003) and Topa (2001) have begun to be studied. As initially discussed in Manski (1993), complicated identification problems exist in uncovering behavioural interdependences even when individual-level data are available. One important message from Brock and Durlauf (2001a; 2001b) is that the nonlinearities that are embedded in the probability structure of statistical mechanics models are important in overcoming what Manski has called the reflection problem. Further, Brock and Durlauf (2007) shows how the presence of multiple equilibria can be used to uncover behavioural interdependences in the presence of group level unobservables.
Once one moves to econometric applications, it is essential to allow for richer forms of individual heterogeneity than are found in the various theoretical models. Indeed, most theoretical models assume that individual agents are described by the same conditional probability measure; one exception is Glaeser, Sacerdote and Scheinkman (1996). At this point, essentially nothing is known about the properties of statistical mechanics models in which empirically salient forms of heterogeneity have been introduced. For this reason, I believe that advances in the use of statistical mechanics methods in economic theory and econometrics will prove to be strong complementary.

Additional reading

Thompson (1988) is a standard physics textbook on statistical mechanics. Badii and Politi (1997) provide a useful discussion of statistical mechanics that segues from physical to statistical and computational models. Liggett (1985; 1991) are magisterial mathematical treatments of the probability structures that underlie statistical mechanics models as I have described them. Kinderman and Snell (1980) is an informal but enjoyable treatment and useful for building intuition. In the statistical mechanics literature, models where there is heterogeneity in the interaction weights linking individual elements are known as ‘spin glasses’; see Fischer and Hertz (1991) for a readable treatment. Durlauf (1997) develops a statistical mechanics framework that nests a number of models that have appeared in economics, particularly those associated with complex systems; related perspectives are found in Ioannides (1997).

See Also

- economy as a complex system
- emergence
- ergodicity and nonergodicity in economics
- Markov processes
- social interactions (empirics)
- social interactions (theory)
- spatial econometrics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Allen, B. 1982. Some stochastic processes of interdependent demand and technological diffusion exhibiting externalities among adopters. *International Economic Review* 23, 595–608.

Auerswald, P., Kauffman, S., Lobo, J. and Shell, K. 2000. The production recipes approach to technological innovation: an application to learning by doing. *Journal of Economic Dynamics and Control* 24, 389–450.

Badii, R. and Politi, A. 1997. *Complexity: Hierarchical Structures and Scaling in Physics*. New York: Cambridge University Press.

Bak, P., Chen, K., Scheinkman, J. and Woodford, M. 1993. Aggregate fluctuations from independent sectoral shocks: self-organized criticality in a model of production and inventory dynamics. *Ricerche Economiche* 47, 3–30.

Bisin, A., Horst, U. and Özgür, O. 2006. Rational expectations equilibria of economies with social interactions. *Journal of Economic Theory* 127, 74–116.

Blume, L. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5, 387–424.

Blume, L. 1995. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior* 11, 111–45.

Brock, W. 1993. Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance. *Estudios Economicos* 8, 3–55.

Brock, W. and Durlauf, S. 2001a. Discrete choice with social interactions. *Review of Economic Studies* 68, 235–60.

Brock, W. and Durlauf, S. 2001b. Interactions-based models. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Brock, W. and Durlauf, S. 2006. A multinomial choice model with social interactions. In *The Economy as an Evolving Complex System III*, ed. L. Blume and S. Durlauf. New York: Oxford University Press.

Brock, W. and Durlauf, S. 2007. Identification of binary choice models with social interactions. *Journal of Econometrics* 140, 52–75.

Conley, T. and Topa, G. 2003. Identification of local-interactions models with imperfect location data.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Journal of Applied Econometrics 18, 605–18.

Durlauf, S. 1993. Nonergodic economic growth. *Review of Economic Studies* 60, 349–66.

Durlauf, S. 1997. Statistical mechanics approaches to socioeconomic behavior. In *The Economy as an Evolving Complex System II*, ed. W. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison-Wesley.

Ellis, R. 1985. *Entropy, Large Deviations, and Statistical Mechanics*. New York: Springer-Verlag.

Fischer, K. and Hertz, J. 1991. *Spin Glasses*. New York: Cambridge University Press.

Follmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1, 51–62.

Glaeser, E., Sacerdote, B. and Scheinkman, J. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111, 507–48.

Horst, U. 2005. Financial price fluctuations in a stock market model with many interacting agents. *Economic Theory* 25, 917–832.

Horst, U. and Scheinkman, J. 2006. Equilibria in systems of social interactions. *Journal of Economic Theory* 130, 44–77.

Ioannides, Y. 1997. The evolution of trading structures. In *The Economy as an Evolving Complex System II*, ed. W. Arthur, S. Durlauf and D. Lane. Reading, MA: Addison-Wesley.

Ioannides, Y. 2006. Topologies of social interactions. *Economic Theory* 28, 559–84.

Kelly, M. 1994. Big shocks versus small shocks in a dynamic stochastic economy with many interacting agents. *Journal of Economic Dynamics and Control* 18, 397–410.

Kindermann, R. and Snell, J. 1980. *Markov Random Fields and Their Applications*. Providence, RI: American Mathematical Society. Online. Available at http://0-www.ams.org.library.lemoyne.edu/online_bks/conm1/conm1-whole.pdf, accessed 13 September 2006.

Kosfeld, M. 2005. Rumours and markets. *Journal of Mathematical Economics* 41, 646–64.

Liggett, T. 1985. *Interacting Particle Systems*. New York: Springer-Verlag.

Liggett, R. 1991. *Stochastic Interacting Systems: Contact, Voter, and Exclusion Models*. New York: Springer-Verlag.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Manski, C. 1993. Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* 60, 531–42.

Oomes, N. 2003. Local trade networks and spatially persistent unemployment. *Journal of Economic Dynamics and Control* 27, 2115–49.

Thompson, C. 1988. *Classical Equilibrium Statistical Mechanics*. New York: Oxford University Press.

Topa, G. 2001. Social interactions, local spillovers, and unemployment. *Review of Economic Studies* 68, 261–95.

How to cite this article

Durlauf, Steven N. "statistical mechanics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000463>
doi:10.1057/9780230226203.1612

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

statistics and economics

Aris Spanos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Some statisticians and economists might find it surprising to learn that statistics and economics share common roots going back to ‘Political Arithmetic’ in the mid-17th century. The primary objective of this article is to revisit the common roots and trace the parallel development of both disciplines up to and including the 20th century, and to attempt to signpost certain methodological lessons that were missed along the way to the detriment of both disciplines. The emphasis is primarily on methodological developments, with less attention paid to institutional developments.

Keywords

ARIMA models; Bayes, T.; Bernoulli, J.; Bowley, A. L.; central limit theorems; cointegration; convergence in distribution; Cournot, A. A.; Cowles Commission; Davenant, C.; Econometric Society; Edgeworth, F.Y.; error-correction models; Farr, W.; Fisher, I.; Fisher, R. A.; frequentist approach to inference; Galton, F.; Gauss, C. F.; Gauss–Markov theorem; generalized method of moments; graphical techniques; Graunt, J.; Haavelmo, T.; Heckman, J. J.; Hume, D.; identification; index numbers; induction; inverse probability; Jevons, W. S.; King, G.; Koopmans, T. C.; Laplace, P.-S.; law of large numbers; least squares; Legendre, A.-M.; life tables; marginal revolution; mathematics and economics; Mills, F. C.; Mortality; Neyman, J.; nonparametric methods; Pearson, K.; Petty, W.; Playfair, W.; political arithmetic; political economy; Probability; Quetelet, A.; reliability of inference; Royal Statistical Society; semiparametric methods; simultaneous equations models; specification; spurious regressions; statistical adequacy; statistical description; statistical inference; statistical models; Statistical Society of London; statistics and economics; stochastic processes; structural models; unit roots; Walras, L.; Yule, G. U.

Article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The close interrelationship between economics and statistics, going back to their common roots in 'Political Arithmetic', played a crucial role in availing the development of both disciplines during their practical knowledge (pre-academic) period. Political economy was first separated from political arithmetic and became an academic discipline – the first social science – at the end of the 18th century, partly as a result of political arithmetic losing credibility. Statistics emerged as a 'cleansed' version of political arithmetic, focusing on the collection and tabulation of data, and continued to develop within different disciplines including political economy, astronomy, geodesy, demography, medicine and biology; however, it did not become a separate academic discipline until the early 1900s.

During the 19th century the development of statistics was institutionally nurtured and actively supported by the more empirically oriented political economists such as Thomas Malthus who helped to create section F of the Royal Society, called 'Economic Science and Statistics', and subsequently to found the Statistical Society of London. The teaching of statistics was introduced into the university curriculum in the 1890s, primarily in economics departments (see Walker, 1929).

The close relationship between economics and statistics was strained in the first half of the 20th century, as the descriptive statistics tradition, associated with Karl Pearson, was being transformed into modern (frequentist) statistical inference in the hands of Fisher (1922, 1925, 1935a, 1956) and Neyman and Pearson (1933), and Neyman (1935, 1950, 1952). During the second half of the 20th century this relationship eventually settled into a form of uneasy coexistence. At the dawn of the 21st century there is a need to bring the two disciplines closer together by implementing certain methodological lessons overlooked during the development of modern statistics.

1 The 17th century: political arithmetic, the promising beginnings

If one defines statistics broadly as 'the subject matter of collecting, displaying and analysing data', the roots of the subject are traditionally traced back to John Graunt's (1620–74) *Natural and Political Observations upon the Bills of Mortality*, published in 1662 (see Hald, 1990; Stigler, 1986), the first systematic study of demographic data on birth and death records in English cities. Graunt detected surprising regularities stretching back over several decades in a number of numerical aggregates, such as the male/female ratio, fertility rates, death rates by age and location, infant mortality rates, incidence of new diseases and epidemics, and so on. On the basis of these apparent regularities, Graunt proceeded to draw certain tentative inferences and discuss their implications for important public policy issues. Hald summarized the impact of this path-breaking book as follows:

> Graunt's book had immense influence. Bills of mortality similar to the London bills were introduced in other cities, for example, Paris in 1667. Graunt's methods of statistical analysis were adopted by Petty, King and Davenant in England; Vauban in France; by Struyck in the Netherlands; and somewhat later by Sussmilch in Germany. Ultimately, these endeavours led to the establishment of governmental statistical offices. Graunt's investigation on the stability of the sex ratio was continued by Arthuthnott and Nicolas Bernoulli. (Hald, 1990, p. 103)

Graunt's book had close affinities in both content and objectives to several works by his close friend

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
William Petty (1623–87) on ‘Political Arithmetick’ published during the 1670s and 1680s; Graunt and Petty are considered joint founders of the ‘political arithmetic’ tradition (Redman, 1997). The fact that Graunt had no academic credentials and published only the single book led to some speculation in the 1690s, which has persisted to this day, that Petty was the real author of *The Bills of Mortality*. The current prevailing view (see Greenwood, 1948; Kreager, 1988) is that Petty's potential influence on Graunt's book is marginal at best. Stone aptly summarizes this view as follows:

> Graunt was the author of the book associated with his name. More than likely, he discussed it with his friend; Petty may have encouraged him to write it, contributed certain passages, helped obtaining the Bills for the county parish … at Romsey, the church in which Petty's baptism is recorded and in which he is buried; he may even have suggested the means of interpolating the numbers of survivors between childhood and old age. But all this does not amount to joint let alone sole authorship. (Stone, 1997, p. 224)

Hull (1899), one of Petty's earliest biographers and publisher of his works, made a strong case against Petty being the author of the ‘Bills of Mortality’ by comparing his methodological approach to that of Graunt:

> Graunt exhibits a patience in investigation, a care in checking his results in every possible way, a reserve in making inferences, and a caution about mistaking calculation for enumeration, which do not characterize Petty's work to a like degree.
> The spirit of their work is often different when no question of calculation enters. Petty sometimes appears to be seeking figures that will support a conclusion which he has already reached; Graunt uses his numerical data as a basis for conclusions, declining to go beyond them. He is thus a more careful statistician than Petty, but he is not an economist at all. (Hull, 1899, pp. xlix and lxxv)

Both Graunt and Petty used limited data to draw conclusions and make predictions about the broader populations, exposing themselves to severe criticisms as to the appropriateness and reliability of such inferences. For instance, using data on christenings and burials in a single county parish in London, they would conjure up estimates of the population of London (which included more than 130 parishes), and then on the basis of those estimates, and certain contestable assumptions concerning mortality and fertility rates, proceed to project estimates of the population of the whole of England. The essential difference between their approaches is that Graunt put enough emphasis on discussing the possible *sources of error* in the collection and compilation of his data, as well as in his assumptions, enabling the reader to assess the reliability (at least qualitatively) of his inferences. Petty, in contrast, was more prone to err on the side of political expediency by drawing inferences that would appeal to the political powers of his time (see Stone, 1997).

Graunt and Petty considered statistical analysis a way to draw *inductive inferences* from observational data, analogous to performing experiments in the physical sciences (see Hull, 1899, p. lxv). Political arithmetic stressed the importance of a new method of quantitative measurement – ‘the art of reasoning by figures upon things relating to the government’ – and was instrumental in the development of both

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
statistics and economics (see Redman, 1997, p. 143). The timing of this emphasis on quantitative measurement and the collecting of data was not coincidental. The *empiricist* turn pioneered by Francis Bacon (1561–1626) had a crucial impact on intellectual circles such as the London Philosophical Society and the British Association, with which Graunt and Petty were associated – these circles included Robert Boyle, John Wallis, John Wilkins, Samuel Hartlib, Christopher Wren and Isaac Newton. As summarized by Letwin:

> The scientific method erected by Bacon rested on two main pillars: natural history, that is, the collection of all possible facts about nature, and induction, a careful logical movement from those facts of nature to the laws of nature. (Letwin, 1965, p. 131)

Graunt and Petty were also influenced by philosopher John Locke (1632–1704), through personal contact. Locke was the founder the British empiricist tradition, which continued with George Berkeley (1685–1753) and David Hume (1711–76). Indeed, all three philosophers wrote extensively on political economy as it relates to empirical economic phenomena, and Locke is credited with the first use of the most important example of analytical thinking in economics, the demand-supply reasoning in determining price (see Routh, 1975).

Graunt's and Petty's successors in the political arithmetic tradition, Gregory King (1648–1712) and Charles Davenant (1656–1714) continued to emphasize the importance of collecting data as the only objective way to frame and assess sound economic policies. Their efforts extended the pioneering results of Grant and Petty and provided an improved basis for some of the original predictions (such as the population of England), but they did not provide any new methodological insights into the analysis of the statistical regularities originally enunciated by Graunt. The enhanced data collection led to discussions of how certain economic variables should be measured over time, and a new literature on index numbers was pioneered by William Fleetwood (1656–1723). The roots of national income accounting, which eventually led to the current standardized macro-data time series, can be traced back to the efforts of these early pioneers in political arithmetic (see Stone, 1997).

According to Hald:

> His [Graunt's] life table was given a probabilistic interpretation by the brothers Huygens; improved life tables were constructed by de Witt in the Netherlands and by Halley in England and used for the computation of life annuities. The life table became a basic tool in medical statistics, demography, and actuarial science. (Hald, 1990, p. 1034)

The improved life tables, with proper probabilistic underpinnings, were to break away from the main political arithmetic and become part of a statistical/probabilistic tradition that would develop independently in Europe in the next two centuries, giving rise to a new literature on life tables and insurance mathematics (see Hald, 1990).

A methodological digression. This was a crucial methodological development for data analysis because it was the first attempt to provide probabilistic underpinnings to Graunt's statistical regularities. Unfortunately, the introduction of probability in the life tables was of limited scope and had no impact on the broader development of political arithmetic, which was growing during the 18th century without

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
any concerns for any probabilistic underpinnings. Without such underpinnings, however, one cannot distinguish between real regularities and artifacts.

2 The 18th century: the demise of political arithmetic

At the dawn of the 18th century political arithmetic promised a way to provide an objective basis for more reliable framing and assessment of economic and social policies. As described by Petty, the method of political arithmetic replaces the use of ‘comparative superlative words, and intellectual arguments’ with ‘number, weight, or measure; to use only arguments of sense; and to consider only such causes as have visible foundations in nature, leaving those that depend on the mutable minds, opinions, appetites, and passions of particular men, to the consideration of others’ (Hull, 1899, p. 244).
English political institutions, including the House of Commons, the House of Lords and the monarchy, took full advantage of the newly established methods of political arithmetic and encouraged, as well as financed, the collection of new data as needed to consider specific questions of policy (see Hoppit, 1996). Putting these methods to the (almost exclusive) service of policy framing by politicians carried with it a crucial danger for major abuse. An inherent problem for social scientists in general has always been to distinguish between inferences relying on sound scientific considerations and those motivated by political or social preferences and leanings.
The combination of (*a*) the absence of sound probabilistic foundations that would enable one to distinguish between real regularities and artefacts, and (*b*) the inbuilt motivation to abuse data in an attempt to make a case for one's favourite policies, led inevitably to extravagant and unwarranted speculations, predictions and claims. These indulgences eventually resulted in the methods of political arithmetic losing credibility. The extent of the damage was such that Greenwood, in reviewing ‘Medical Statistics from Graunt to Farr’, argued:

> One may fairly say on the evidence here summarized that the eighteenth-century political arithmeticians of England made no advance whatever upon the position reached by Graunt, Petty and King. They were second-rate imitators of men of genius. (Greenwood, 1948, p. 49)

An important component of the evidence provided by Greenwood was the ‘population controversy’, which often involved idle speculation in predicting the population of England. This speculation began with Graunt with a lot of cautionary notes attached, but it continued into the 18th century with much less concern about the possible errors that could vitiate such inferences. The discussions were from two opposing schools of thought: the *pessimists*, who claimed that the population was decreasing, and the *optimists*, who argued the opposite; their conflicting arguments were based on the same bills of mortality popularized by Graunt. Neither side had reliable evidence for its predictions because the data provided no sound basis for reliable inference. All predictions involved highly conjectural assumptions of fertility and mortality rates, the average number of people living in each house, and so on. The acrimonious arguments between the two sides revealed the purely speculative foundations of all such claims and contributed significantly to the eventual demise of political economy (see Glass, 1973, for a detailed review).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The above quotation from Greenwood might be considered today as an exaggeration, but it describes accurately the prevailing perception at the end of the 18th century. An unfortunate consequence of disparaging the methods of political arithmetic was the widely held interpretation that it provided decisive evidence for the ineffectiveness of Bacon's *inductive method*. Indeed, one can argue that this cause was instrumental in the timing of the emergence of *political economy* at the end of the 18th century, as the first social science to break away from political arithmetic. Adam Smith (1723–90) declared: ‘I have no great faith in political arithmetick’ (1776, p. 534). James Steuart (1712–80) was even more critical:

> Instead of appealing to political arithmetic as a check on the conclusions of political economy, it would often be more reasonable to have recourse to political economy as a check on the extravagances of political arithmetic. (quoted by Redman, 1997)

During the late 18th century, political economy defined itself by contrasting its methods with those of political arithmetic, arguing that it did not rely only on tables and figures in conjunction with idle speculation, but was concerned with the theoretical issues, causes and explanations underlying the process that generated such data. Political economists contrasted their primarily deductive methods to the discredited inductive methods utilized by political arithmeticians. As argued by Hilts:

> Of importance to the history of statistics in England was the fact that the political economists were fully conscious of their deductive proclivities and saw political economy as methodologically distinct from the inductive science of statistics. (Hilts, 1978, p. 23)

At this point it should be emphasized that the terms induction and deduction had different connotations during the 18th century, and care should be taken when interpreting some of the claims of that period (see Redman, 1997). Despite the criticisms by leading political economists of the inductive method, broadly understood as using the data as a basis of inference, the tradition of collecting, compiling and charting data as well as drawing inferences concerning broad tendencies on such a basis, continued to grow throughout the 18th and 19th centuries, and was influential in the development of political economy. Some political economists such as Thomas Malthus (1766–1834) and John McCulloch (1789–1864) continued to rely on the British empiricist tradition of using data as a basis of inference, but were at great pains to separate themselves from the 18th century's discredited political arithmetic tradition. Indeed, the leading political economists of that period, including Adam Smith and David Ricardo (1772–1823), used historical data extensively in support of their theories, conclusions and policy recommendations developed by deductive arguments (see Backhouse, 2002a).

At the close of the 18th century, the only bright methodological advance in the withering tradition of political arithmetic was provided by William Playfair's (1759–1823) *The Commercial and Political Atlas*, published in 1786. This book elevated the analysis of tabulated data to a more sophisticated level by introducing the power of graphical techniques in displaying and analysing data. Playfair introduced several innovating techniques such as hachure, shading, colour coding, and grids with major and minor divisions of both axes to render the statistical regularities in the data even more transparent. In a certain sense, the graphical techniques introduced by Playfair made certain empirical regularities more

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

transparent and rendered certain conclusions easier to draw. The graphs in this book represent economic time series, measuring primarily English trade (imports/exports) with other countries during the 18th century. Indeed, Playfair's writings were mainly on political economy; his first book, *Regulation of the Interest of Money*, was published in 1785 (see Harrison, 2004).

In what follows the developments in probability theory will be discussed only when they pertain to the probabilistic underpinnings of statistical analysis; for a more detailed and balanced discussion see Hald (1990; 1998; 2007). The probabilistic underpinnings literature on probability developed independently from political arithmetic in England, and there was no interaction between the two until the mid-19th century.

Viewed from today's vantage point, the primary problem with Grant's inferences based on data pertaining to a single parish in London, was how ‘representative’ the data were for the population of London as a whole, which included more than 130 other parishes. This problem was formalized much later in terms of whether the data can be realistically viewed as a ‘random sample’ from the population of London. Defining what a random sample is, however, requires probability theory, which was not adequately understood until the late 19th century (see Peirce, 1878).

Jacob Bernoulli. The first important result relating to the probabilistic underpinnings of statistical regularities was Jacob Bernoulli's (1654–1705) Law of Large Numbers (LLN), published posthumously in 1713 by his nephew Nicolas Bernoulli (1687–1759). Bernoulli's theorem showed that *under certain circumstances*, the relative frequency of the occurrence of a certain event *A*, say $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{m}{n}$ (*m* occurrences of $\{X_i=1\}$ and $n-m$ occurrences of $\{X_i=0\}$ in *n* trials) provides an estimate of the probability $\mathbb{P}(A) = p$ whose accuracy increases as *n* goes to infinity. In modern terminology $\bar{X}$ constitutes a consistent estimator of *p*. Bernoulli went on to use this result in an attempt to provide an interval estimator of the form: *p* is in $(\bar{X} \pm \varepsilon)$ for some $\varepsilon > 0$, but his estimator was rather crude (see Hald, 1990).

A methodological digression. The *circumstances* assumed by Bernoulli were specified in terms of the trials being *independent and identically distributed* (IID). It turned out that the same probabilistic assumption defines the notion of a *random sample* mentioned in relation to the probabilistic underpinnings concerning Graunt's statistical regularities, though the two literatures were developing independently. The role of these probabilistic underpinnings was not made explicit, however, until the early 1920s (see Section 4.2). Indeed, the role of the IID assumptions is often misunderstood to this day. For instance, Hilts argues:

> Mathematically the theorem stated [LLN], in very simplified language, that an event which occurs with a certain probability, appears with a frequency approaching that probability as the number of observations is increased. (Hilts, 1973, p. 209)

Strictly speaking, the LLN says nothing of a sort, because, unless the trials are IID, the result does *not* follow. This insight was clearly articulated by Uspensky:

> It should, however, be borne in mind that little, if any, value can be attached to the practical applications of Bernoulli's theorem, unless the conditions presupposed in this theorem are at least approximately fulfilled: independence of trials and constant

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

probability of an event for every trial. (Uspensky, 1937, p. 104)

Laplace. The first successful attempt to integrate data analysis with the probabilistic underpinnings should be credited to Pierre-Simon Laplace (1749–1827), a famous French mathematician and astronomer, and Thomas Bayes (1702–61), a British mathematician and Presbyterian minister. In papers published in 1764 and 1765 (see Hald, 2007) respectively, they proposed the first inverse probability (posterior-based) interval for *p* for the form *p* of the form '*p* is in $(\bar{x} \pm [\varepsilon|\bar{x}])$' for some $\varepsilon > 0$, by assuming a prior distribution $p \sim U(0,1)$ that is *p* is a uniformly distributed random variable (see Hacking, 1975). This gave rise to the *inverse probability* approach (known today as the Bayesian approach) to statistical inference, which was to dominate statistical induction until the 1920s, before the Fisherian revolution. In 1812 Laplace (see Hald, 2007) also provided the first frequentist interval estimator of *p* of the form *p* is in $(\bar{X} \pm \varepsilon)$ for some $\varepsilon > 0$. The difference between this result and a similar result by Bernoulli is that Laplace used a more accurate approximation based on *convergence in distribution* as the basis of his result; the first *central limit theorem* supplying an asymptotic approximation of the binomial by the Normal distribution (see Hald, 1990).

3 The 19th century: political economy and statistics

The demise of political arithmetic by the early 19th century was instrumental in contributing to the creation of two separate fields: political economy and statistics. Political economy was created to provide more reasoned explanations for the causes and contributing factors giving rise to economic phenomena. Statistics was demarcated by the narrowing down of the scope of political arithmetic in an attempt to cleanse it from the unwarranted speculation that undermined its credibility during the 18th century.

3.1 The Statistical Society of London

Given their common roots, the first institution created to foster the development of the field of statistics, the Statistical Society of London, was created in 1834 with the active participation of several political economists, including Thomas Malthus and Richard Jones (1790–1855), who, together with John Drinkwater (1801–51), Henry Hallam (1777–1859) and Charles Babbage (1791–1871), were to found the Society after some prompting from Quetelet, who visited England in 1833. Other political economists who played very active roles in the early stages of the Society included Thomas Tooke (1774–1858), John R. McCulloch (1789–1864) and Nassau Senior (1790–1864). The first council included notable personalities such as Earl FitzWilliam (1748–1833), William Whewell (1794–1866), G. R. Porter (1792–1852) and Samuel Jones-Loyd (1796–1883).
In an attempt to protect themselves from the disrepute on speculation based on data brought about by political arithmeticians, the new society was founded upon the explicit promise to put the emphasis, not on inference, but upon the collection and tabulation of data of relevance to the state. The founding document stated:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> The Statistical Society of London has been established for the purposes of procuring, arranging, and publishing Facts calculated to illustrate the condition and prospects of the Society. (*Journal of the Statistical Society of London*, 1834, p. 1)

The seal on the cover of the *Journal of the Statistical Society of London* (*JSSL*) was a wheatsheaf around which was written '*aliis exterendum*' ('to be threshed by others'). That is, the aim of the society is to painstakingly gather the facts and let others draw whatever conclusions might be warranted:

> The Statistical Society will consider it to be the first and most essential rule of its conduct to exclude carefully all Opinions from its transactions and publications – to confine its attention rigorously to facts – and, as far as it may be found possible, to facts which can be stated numerically and arranged in tables. (*JSSL*, 1834, pp. 1–2)

Of particular interest is the way the statement of the aims of the society separated statistics from political economy:

> The Science of Statistics differs from Political Economy because although it has the same end in view, it does not discuss causes, nor reason upon probable effects; it seeks only to collect, arrange, and compare, that class of facts which alone can form the basis of correct conclusions with respect to social and political government. (*JSSL*, 1834, p. 2)

The overwhelming majority of the published papers in the *JSSL* were in the political arithmetic tradition of Graunt, relating primarily to economic, medical and demographic data, with two major improvements: ameliorated methods for the collection and tabulation of data giving rise to more accurate and reliable data, and more careful reasoning being used to yield less questionable inferences. This is particularly true for data relating to life tables and mortality rates associated with epidemics. The best examples of such an output are given by William Farr (1807–83), who is considered to be the founder of medical statistics because his analysis of such data contributed to medical advances and crucial changes in policies concerning public health (see Greenwood, 1948). For a more extensive discussion of the methodological and institutional developments associated with data collection and tabulation in England and France see Schweber (2006) and Desrosières (1998).

By the 1850s it became apparent that the early founding declaration of the society to publish papers that stay away from 'Opinions' – drawing conclusions on the basis of data – was unrealistic, unattainable and unjustifiable in the minds of the members of the society. Despite this initial promise, slowly but surely *JSSL* publications began to go beyond the mere reporting and tabulation of data relating to economic, political, demographic, medical, moral and intellectual issues, including poverty figures and education statistics. The motto '*aliis exterendum*' was removed in 1857 from their seal to reflect the new vision of the society (see RSS, 1934).

3.2 The probabilistic underpinnings in the 19th century

During the early 19th century, a completely separate tradition in statistical analysis of data was being

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

developed in Europe (mainly in France and Germany) in the fields of *astronomy* and *geodesy*. This literature was developing completely independently of political arithmetic, but by the 1840s the two traditions had merged in the hands of Adolphe Quetelet (1796–1874): see Porter (1986).

Legendre and Gauss. In the early 19th century the analysis of astronomical and geodesic data by Adrien-Marie Legendre (1752–1833), Carl Friedrich Gauss (1777–1855) and Laplace introduced curve-fitting as a method to summarize the information in data (see Farebrother, 1999). In modern notation the simplest form of curve-fitting can be expressed in the form of a linear model $\mathbf{y}=\mathbf{X}\beta+\varepsilon$, where $\mathbf{y}:=(y_1, y_2,\ldots,y_n)$ and $\mathbf{X}:=(\mathbf{x}_1,\mathbf{x}_2,\ldots,x_n)$ denote a vector and a matrix of observations, respectively, $\boldsymbol{\beta}:=(\beta_1, \beta_2, \ldots, \beta_m)$ a vector of unknown parameters and $\boldsymbol{\varepsilon}:=(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ a vector of errors. Legendre (1805) is credited with inventing least squares as a mathematical approximation method, by proposing the minimization of $\mathcal{E}(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ as a way to estimate β. Gauss (1809) should be credited with providing the probabilistic underpinnings for this estimation problem by transforming the mathematical approximation error into a generic statistical error:

$$\varepsilon_k(x_k, m) = \varepsilon_k \sim \mathrm{NIID}(0, \sigma^2),\ k = 1, 2, \ldots, n, \ldots, \tag{1}$$

where NIID(0, σ^2) stands for 'Normal, Independent and Identically Distributed with mean 0 and variance σ^2'. Laplace provided the first justification of the Normality assumption based on the central limit theorem in 1812 (see Hald, 2007). What makes Gauss's contribution all-important from today's vantage point is that the probabilistic assumptions in (1) provide the framework that enables one to assess the reliability of inference. Ironically, Gauss's embedding of the mathematical approximation problem into a statistical model is rarely appreciated as the major contribution that it is (see Spanos, 2008). Instead, what Gauss is widely credited with is the celebrated Gauss-Markov theorem (see Section 4.6).

Quetelet. The 'law of error' was elevated to a most important method in analysing social phenomena by Adolphe Quetelet (1796–1874), a Belgian astronomer and polymath, in the 1840s. His statistical analysis of data differed in that his methods were integrated with the probabilistic underpinnings that were lacking in the analysis of political arithmeticians; his probabilistic perspective was primarily influenced by the work of Joseph Fourier (1768–1830), a French mathematician and physicist. Quetelet's most important contribution was to explicate Graunt's regularities in terms of the notion of probabilistic (chance) regularity which combined the unpredictability at the individual level with the abiding regularity at the aggregate level. By fitting the Normal curve over the histogram of a great variety of social data, his objective was to eliminate 'accidental' influences and determine the average physical and intellectual features of a human population, including normal and abnormal behaviour. His *modus operandi* was the notion of the 'average man' (see Desrosières, 1998). The 'average man' began as a simple way of summarizing the systematic characteristic of a population, but in some of Quetelet's later work, 'average man' is presented as an ideal type, and any deviations from this ideal were interpreted as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
errors of nature.

A methodological digression. In addition to the substantive issues raised by his approach to ‘social physics’ (see Cournot, 1843), the methodological underpinnings of Quetelet's statistical analysis were rather weak. When the Normal curve is fitted over a histogram of data $\mathbf{x}:=(x_1,x_2,\ldots,x_n)$ in an attempt to summarize the statistical regularities, one implicitly assumes that data $\mathbf{x}$ constitutes a realization of an IID process $\{X_k, k=1,2,\ldots,n,\ldots\}$ (see Spanos, 1999); these are highly questionable assumptions for most of the data used in Quetelet (1942). The concern to evaluate the precision (reliability) of an inference, introduced earlier by Laplace and Gauss, was absent from Quetelet's work. Hence, his analysis of statistical regularities did not give rise to any more reliable inferences than those of the political arithmetic tradition a century earlier; the necessity to assess the validity of the premises (NIID assumptions) for inductive inference was not clearly understood at the time.

3.3 The ‘ mathematical ’ turn

In the last quarter of the 19th century there was a concerted effort to render both statistics and economics more rigorous by introducing the language of mathematics into both disciplines. In statistics this effort was spearheaded by Edgeworth, Galton and Pearson and in economics by Edgeworth, Jevons, Walras and Irving Fisher. The mathematical turn of this period was motivated by the strong desire to emulate the physical sciences and introduce quantification into these fields, which involved both calculus and probability theory (see Backhouse, 2002a; 2002b).

Galton. Quetelet's use of the Normal curve to analyse social data had a powerful influence on Francis Galton (1822–1911), who provided a different interpretation to the ‘law of error’. Galton (1869) interpreted the variation around the mean, not as errors from the ideal type, but as the very essence of nature's variability. Using this variability he introduced the notions of regression and correlation in the 1890s as a way to determine relationships between different data series $\{(x_k, y_k), k=1,2,\ldots,n\}$. Regression and correlation opened the door for providing statistical explanations which revolutionized statistical modelling in the biological and social sciences (see Porter, 1986). Retrospectively, Galton was the founder of the biometrics tradition, which had a great influence on the development of statistics in the 20th century in the hands of Karl Pearson (1857–1936) and Udny Yule (1871–1951).

Pearson significantly extended the summarization of data in the form of smoothing histograms, by introducing a whole family of new frequency curves – known today as the Pearson family – to supplement the Normal curve, and applied these techniques extensively to biological data, with notable success. He also provided clear probabilistic underpinnings for Galton's regression and correlation methods. *Yule* (1897) established a crucial link between the Legendre–Gauss least-squares and the linear regression model, by showing that least-squares can be used to estimate the parameters of linear regression, bringing together two seemingly unrelated literatures (see Spanos, 1999). This was an important breakthrough that, unfortunately, also introduced a confusion between two different perspectives on empirical modelling: curve-fitting as a mathematical approximation method, and the probabilistic perspective where regression is viewed as a purely probabilistic concept defined in terms of the first moment of a conditional distribution (see Stigler, 1986; Spanos, 2008). Yule published a highly influential textbook in statistics in 1911 in which he successfully blended the biometric with the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
'economic statistics' tradition.

Edgeworth. Of particular interest for the fundamental interaction between statistics and economics is the case of Francis Edgeworth (1845–1926), primarily an economist. His mathematical self-training enabled him to provide a bridge between the theory of errors tradition going back to Gauss and Laplace, the biometric tradition of Galton and Pearson, and the more traditional economic statistics of the 19th century focusing on economic time series data and index numbers. His direct influence on statistics, however, was rather limited because the style and mathematical level of his writings were too demanding for the statisticians of the late 19th century. Bowley (1928), 'at the request of the Council prepared a summary of his mathematical work which may have served to make his achievement known to a wider circle' (see RSS, 1934, p. 238). Edgeworth contributed crucially to the mathematization of economics and the theory of index numbers (see Backhouse, 2002b).

William Stanley Jevons (1835–82), English economist and logician. In his book *The Theory of Political Economy* (1871), he used calculus to formulate the marginal utility theory of value, and the notion of partial equilibrium, which provided the foundation for the marginalist revolution in economics (see Backhouse, 2002a).

Léon Walras (1834–1910) was a French mathematical economist, one of the protagonists in the marginalist revolution and the innovator of general equilibrium theory. His perspective on the use of mathematics in economics was greatly influenced by *Augustin Cournot* (1801–77), a French philosopher, mathematician and economist. Cournot is credited with the notion of functional relationships among economic variables, which led him to the supply and demand curves (see Backhouse, 2002a).

In the United States the process of mathematization began somewhat later with *Irving Fisher* (1867–1947), who followed in the footsteps of Walras, Jevons and Edgeworth in introducing mathematics into economics and making significant contributions to the theory of index numbers (see Backhouse, 2002b).

These early pioneers in the mathematization of economics shared a vision of using statistics to provide pertinent empirical foundations to economics (see Moore, 1908). Fisher described this goal as a life-long ambition:

> I have valued statistics as an instrument to help fulfill one of the great ambitions of my life, namely, to do what I could toward making economics into a genuine science. (Fisher, 1947, p. 74)

The same vision was clearly articulated much earlier by Jevons:

> The deductive science of Economics must be verified and rendered useful by the purely empirical science of statistics. (Jevons, 1871, p. 12)

Indeed, Neville Keynes attributed to *statistics* a much greater role in the quantification of economics than hitherto:

> The functions of statistics in economic enquiries are: … descriptive, … to suggest

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> empirical laws, which may or may not be capable of subsequent deductive explanation, … to supplement deductive reasoning by checking its results, and submitting them to the test of experience, … the elucidation and interpretation of particular concrete phenomena, … enabling the deductive economist to test and, where necessary, modify his premisses, … measure the force exerted by disturbing agencies. (Keynes, 1890, pp. 342–46)

At the dawn of 20th century, pioneers such as Moore (1908; 1911), who aspired to help in securing empirical foundations for economics, had several advantages – for example, the institutionalization of the collection and compilation of economic data via the establishment of government statistical offices, the systematic development of index numbers, and so on. The mathematization of economics provided them with economic models amenable to empirical enquiry (see Backhouse, 2002a; 2002b). In addition, at the end of the 19th century there were several developments in statistical methods, including least-square curve-fitting, regression, correlation, periodogram analysis and trend modelling that seemed tailor-made for analysing economic data (see Mills, 1924; Stigler, 1954; Heckman, 1992; Hendry and Morgan, 1995).

4 The 20th century: a strained relationship

To enliven the discussion of the tension created in the 1920s between economic statistics and statistical inference, the account below refers to the confrontation between the two protagonists who represented the different perspectives, Bowley and Fisher.

4.1 Economic statistics as against statistical inference

The early 20th century statistics scene was dominated by Karl Pearson (1857–1936) and his research in biology at the Galton Laboratory established in 1904. Pearson's research at this laboratory consolidated the biometrics tradition, whose primary outlet was the in-house journal *Biometrika*. Pearson established the department of ‘Applied Statistics’ at University College in 1911, which, at the time, was the only place one could study for a degree in statistics (see Walker, 1958).

Arthur Bowley (1869–1957) was a typical successful ‘economic statistician’ of the early 20th century who authored one of the earliest textbooks in statistics, *Elements of Statistics* (1901), while a part-time lecturer at the London School of Economics. Bowley understood statistics as comprising two different but interrelated components, the *arithmetic* and the *mathematical*. The former was concerned with statistical techniques as they relate to measurement, compilation, interpolation, tabulation and plotting of data, as well as the construction of index numbers; this constitutes *Part I – General Elementary Methods*, and comprises the first 258 pages of Bowley (1902). The mathematical dimension (*Part II – The Application of the Theory of Probability to Statistics*, the last 74 pages of Bowley, 1902) was concerned with the use of probability theory in minimizing and evaluating the errors associated with particular inferences. The last 12 pages of Bowley (1902) are devoted to a discussion of ‘regression and correlation’ as expounded by Pearson (1896) and Yule (1897).

Bowley (1906) illustrated what he meant by ‘errors’ using the ‘probable error’ for the arithmetic average

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
$\bar{x}_n := \frac{1}{n}\sum_{k=1}^{n} x_k$ of the data $(x_1,x_2,\ldots,x_n)$ as:

$$\bar{x} \pm SD(\bar{x}), \tag{2}$$

with $SD(\bar{x})$ denoting the standard deviation of $\bar{x}$. Taking the Normal distribution as an example he argued that the claim in (2) can be interpreted as saying that ‘the chance that a given observation should be within this distance of the true average is 2:1’ (1906, p. 549). This interpretation can be best understood as based on a Bayesian credible interval evaluation, instead of a frequentist confidence interval developed in the 1930s.

From this perspective, Bowley interpreted the work of Pearson and Edgeworth as concerned with providing different ways to evaluate these ‘probable errors’ (for example, $SD(\bar{x})$) using either a fitted frequency curve or an asymptotic approximation, respectively (see Bowley, 1906, p. 550). It is interesting to note that in the 5th edition of Bowley's statistics book, published in 1926, *Part II* increased threefold to 210 pages, but contains no reference to Fisher's work, which, at the time, was well on its way to transform Karl Pearson's descriptive statistics into modern statistical inference.

Ronald Fisher (1893–1962) pioneered a recasting of statistics (1915; 1921; 1922), moving away from the Edgeworth–Pearson reliance on large sample approximations based on inverse probability (Bayesian) methods, and focusing on finite sample frequentist inference relying on sampling distributions. This recasting was initially inspired by Gossett's (1908) derivation of the student's t distribution for a given sample size *n*. Fisher made this recasting explicit in his 1921 paper by severely criticizing the inverse probability (Bayesian) approach and articulating a more complete picture of his approach to statistical inference in his 1922 classic paper.

In the early 1920s Bowley was a professor of statistics (second in fame only to Karl Pearson) at the London School of Economics (LSE), known primarily for his contributions in the area of survey sampling, and Fisher was a young statistician at Rothamstead Agricultural Station trying to make sense of a 200-year accumulation of experimental data. Bowley was aware of Fisher's early work: we know that in 1924 Bowley requested and promptly received Fisher's offprints for the LSE library (see Box, 1978, p. 171). Moreover, by some accident of faith, the two were neighbours at Harpenden, interacting socially as bridge companions (see Box, 1978, p. 85). Indeed, Bowley encouraged Fisher to publish his correction of Pearson's (1900) evaluation of degrees of freedom associated with his goodness-of-fit test (see Fisher, 1922a).

The next academic encounter between the professor and the young aspiring statistician was in 1929 when Fisher applied for an academic position in Social Biology at the LSE, but was turned down in favour of Lancelot Hogben (see Box, 1978, p. 202). Fisher's first academic position was at University College as Professor of Eugenics, in 1933. The tension between their different perspectives on statistics became public in their first showdown at Fisher's presentation to the Royal Statistical Society in 18 December 1934 entitled ‘The Logic of Inductive Inference’, where he attempted to explain his published work on recasting the problem of statistical induction since his 1922 paper. Bowley was appointed to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
move the traditional vote of thanks and open the discussion, and after some begrudging thanks for Fisher's ‘contributions to statistics in general’ – by then Fisher's 1925 book had made him famous – he went on to disparage his new approach to statistical inference based on the likelihood function by describing it as abstruse, arbitrary and misleading. His comments were predominantly sarcastic and discourteous, and went as far as to accuse Fisher of giving insufficient credit to Edgeworth (see Fisher, 1935, pp. 55–7). The litany of churlish comments and currish remarks continued with the rest of the old guard: Isserlis, Irwin and the philosopher Wolf (1935, pp. 57–64), who was brought in by Bowley to undermine Fisher's philosophical discussion on induction. Jeffreys complained about Fisher's criticisms of the Bayesian approach (1935, pp. 70–2). To Fisher's support came Egon Pearson, Neyman and, to a lesser extent, Bartlett. Pearson (1935, pp. 64–5) argued that:

> When these ideas [on statistical induction] were fully understood … it would be realized that statistical science owed a very great deal to the stimulus Professor Fisher had provided in many directions. (Pearson, 1935, pp. 64–5)

Neyman was equally supportive, praising Fisher's path-breaking contributions, and explaining Bowley's reaction to Fisher's critical review of the traditional view of statistics as understandable attachment to old ideas (1935, p. 73).

Fisher, in his reply to Bowley and the old guard, was equally contemptuous:

> The acerbity, to use no stronger term, with which the customary vote of thanks has been moved and seconded … does not, I confess, surprise me. From the fact that thirteen years have elapsed between the publication, by the Royal Society, of my first rough outline of the developments, which are the subject of to-day's discussion, and the occurrence of that discussion itself, it is a fair inference that some at least of the Society's authorities on matters theoretical viewed these developments with disfavour, and admitted with reluctance. … However true it may be that Professor Bowley is left very much where he was, the quotations show at least that Dr. Neyman and myself have not been left in his company. … For the rest, I find that Professor Bowley is offended with me for ‘introducing misleading ideas’. He does not, however, find it necessary to demonstrate that any such idea is, in fact, misleading. It must be inferred that my real crime, in the eyes of his academic eminence, must be that of ‘introducing ideas’. (Fisher, 1935, pp. 76–82)

Fisher's reference to ‘his academic eminence’, although containing a dose of sarcasm, it was not totally out of place. Bowley became a member of the Council of the Royal Statistical Society as early as 1898, served as its Vice-President in 1907–8 and again in 1912–14, and President in 1938–40. He was awarded the society's highest honour, the Guy Medal in gold, in 1935; he received the Guy in silver as early as 1895. In contrast, Fisher had no academic position until 1933, and even that came with the humiliating stipulation that he would *not* teach statistics from his new position as Professor of Eugenics at University College (see Box, 1978, p. 258).

Fisher made it clear that he associated the ‘old guard’ in statistics with Bowley-type economic statistics:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences. This particular dependence of social studies upon statistical methods has led to the unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth methods adequate to the treatment of economic data, in so far as these exist, have mostly been developed in the study of biology and the other sciences. (Fisher, 1925, p. 2)

The unbridgeable gap between Bowley and the ‘old guard’ on one side, and Fisher, Neyman and Pearson on the other, was apparent six months earlier when Bowley was assigned the same role for Neyman's first presentation. Despite the fact that Neyman began his presentation by praising Bowley for his earlier contributions to survey sampling methods, he grouped him with Fisher and accused him of the same abstruseness:

> I am not certain whether to ask for an explanation or to cast a doubt. It is suggested in the paper that the work is difficult to follow and I may be one of those who have been misled by it. I can only say I have read it at the time it appeared and since, and I have read Dr Neyman's elucidation of it yesterday with great care. I am referring to Dr Neyman's confidence limits. I am not at all sure that the ‘confidence’ is not a ‘confidence trick’. (Neyman, 1934, pp. 608–9)

His ‘confidence trick’ remark is not very surprising in view of Bowley's own interpretation of (2) in inverse probabilistic (Bayesian) terms. Predictably, Egon Pearson and Fisher came to Neyman's rescue from the rebukes of old guard.
Retrospectively, Bowley's charge of abstruseness, levelled at both Fisher and Neyman, might best be explained in terms of David Hume's (1711–76) ‘tongue in cheek’ comment two centuries earlier:

> The greater part of *mankind* may be divided into two classes; that of *shallow* thinkers, who fall short of the truth; and that of *abstruse* thinkers, who go beyond it. The latter class are by far the most rare; and I may add, by far the most useful and valuable. They suggest hints, at least, and start difficulties, which they want, perhaps, skill to pursue; but which may produce fine discoveries, when handled by men who have a more just way of thinking.•…•All people of *shallow* thought are apt to decry even those of *solid* understanding, as *abstruse* thinkers, and methaphysicians, and refiners; and never will allow any thing to be just which is beyond their own weak conceptions. (Hume, 1987, pp. 253–4)

In summary, the pioneering work of Fisher, Egon Pearson and Neyman, was largely ignored by the Royal Statistical Society (RSS) establishment until the early 1930s. By 1933 it was difficult to ignore their contributions, published primarily in other journals, and the ‘establishment’ of the RSS decided to display its tolerance to their work by creating ‘the Industrial and Agricultural Research Section’, under the auspices of which both papers by Neyman and Fisher were presented in 1934 and 1935 respectively. In their centennial volume published in 1934, the RSS acknowledged the development of ‘mathematical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
statistics', referring to Galton, Edgeworth, Karl Pearson, Yule and Bowley as the main pioneers, and listed the most important contributions in this sub-field which appeared in its *Journal* during the period 1909–33, but the three important papers by Fisher (1922a; 1922b; 1924) are conspicuously absent from that list. The list itself is dominated by contributions in vital, commercial, financial and labour statistics (see RSS, 1934, pp. 208–23). There is only one reference to Egon Pearson, for his 1933 paper 'Control and Standardization of Quality of Manufactured Products' – the very paper used as self-justification by the RSS in creating the new section. It is interesting to note that by the late 1920s the revolutionary nature of Fisher's new approach to statistics was clearly recognized by many. Tippet (1931) was one of the earliest textbook attempts to blend the earlier results on regression and correlation within Fisher's new approach. In the United States, Hotelling (1930) articulated a most elucidating perspective on Fisher's approach.

4.2 The Fisher – Neyman – Pearson approach

The main methods of the Fisher–Neyman–Pearson (F–N–P) approach to statistical inference, point estimation, hypothesis testing and interval estimation, were in place by the late 1930s. The first complete textbook discussion of this approach, properly integrated with its probabilistic underpinnings, was given by Cramer (1946). The methodological discussions concerning the form of inductive reasoning underlying the new frequentist approach, however, were to linger on until the 1960s and beyond; see the exchange between Fisher (1955), Pearson (1955) and Neyman (1956).
One of the most crucial insights of the F–N–P approach to statistical inference, which set it apart from previous approaches to statistics, was the explicit specification of the premises of statistical induction in terms of the notion of a *statistical model*:

> The postulate of randomness thus resolves itself into the question, 'Of what population is this a random sample?' which must frequently be asked by every practical statistician. (Fisher, 1922, p. 313)

He defined the initial choice of the statistical model in the context of which the data will be interpreted as a 'representative sample' as the problem of *specification*, emphasizing the fact that: 'the adequacy of our choice may be tested *posteriori*' (1922, p. 314). Indeed, the first three tests discussed in Fisher (1925, pp. 78–94) are misspecification (M-S) tests for the Normality, Independence and Identically Distributed assumptions. Fisher (1922; 1925; 1935), and later Neyman (1938/1952; 1950), emphasized the importance of both model specification and validation vis-à-vis the data:

> Guessing and then verifying the 'chance mechanism', the repeated operations of which produces the observed frequencies. (Neyman 1977, p. 99)

Pearson (1931a, 1931b) was among the first to discuss the implications of non-Normality as well as develop M-S tests for it; see Lehmann (1999) for the early concern about the consequences of misspecification in the 1920s.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The F–N–P discernments concerning statistical model *specification*, *M-S testing*, and *respecification* can be summarized in the form of what might be called the F–N–P *perspective* (articulated in Spanos, 2006a) which can be summarized as follows:

1. 1. Every statistical (inductive) inference is based on certain *premises*, in the form of (*a*) a *statistical model* $\mathcal{M}$ parameterizing the probabilistic structure of an observable stochastic process $\{\mathbf{Z}_t, t \in \mathbb{N}\}$, and (*b*) a set of data $\mathbf{Z}:=(\mathbf{z}_1,\ldots,\mathbf{z}_n)$, viewed as a 'typical realization' of this process.
2. 2. A statistical model is specified in terms of a complete and internally consistent set of probabilistic assumptions concerning the underlying stochastic process $\{\mathbf{Z}_t, t \in \mathbb{N}\}$. For example, the Normal/linear regression model is specified in terms of assumptions [1]–[5] (Table 1) concerning the observable process $\{(y_t|\mathbf{X}_t = \mathbf{x}_t), t \in \mathbb{N}\}$, and not the errors.
3. 3. *Statistical adequacy*. Securing the validity of assumptions [1]–[5] vis-à-vis the data in question is necessary for establishing 'statistical regularities' and ensuring the reliability of inference (see Spanos, 2006a; 2006b; 2006c).

The Normal/linear regression model

Statistical GM:	$y_t = \beta_0 + \boldsymbol{\beta}_1^{\top} \mathbf{x}_t + u_t, \quad t \in \mathbb{N},$
[1] Normality:	$(y_t\|\mathbf{X}_t = \mathbf{x}_t) \sim \mathrm{N}(\cdot, \cdot),$
[2] Linearity:	$E(y_t\|\mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}_1^{\top} \mathbf{x}_t,$ linear in $\mathbf{x}_t,$
[3] Homoskedasticity:	$Var(y_t\|\mathbf{X}_t = \mathbf{x}_t) = \sigma^2,$ free of $\mathbf{x}_t,$
[4] Independence:	$\{(y_t\|\mathbf{X}_t = \mathbf{x}_t), \quad t \in \mathbb{N}$ an independent process,
[5] t-invariance:	$\boldsymbol{\theta} := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ do not change with t.

The importance of the F–N–P perspective stems from the fact that the *statistical model* enables one:

1. (i) to assess the validity (statistical adequacy) of the premises for inductive inference – by testing the assumptions using misspecification tests; and
2. (ii) to provide relevant error probabilities for appraising the reliability of the associated inference (see Spanos, 2006a).

It is well known that the *reliability* of any inference procedure depends crucially on the validity of the pre-specified *statistical model* vis-à-vis the data in question. The optimality of these procedures is defined by their capacity to give rise to valid inferences (*trustworthiness*), which is calibrated in terms of the associated error probabilities – how often these procedures lead to erroneous inferences (see Mayo, 1996). In the case of confidence interval estimation the calibration is usually gauged in terms of minimizing *the coverage error probability*: the probability that the interval does *not* contain the true value of the unknown parameter(s). In the case of hypothesis testing the calibration is ascertained in terms of minimizing *the type II error probability* – the probability of accepting the null hypothesis when false, for a given *type I error probability* (see Cox and Hinkley, 1974). It is also known, but often insufficiently appreciated, that when any of the model assumptions are invalid, the *reliability of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
inference is called into question (see Pearson, 1931a; Bartlett, 1935, for early discussions). Departures from the model assumptions will give rise to a discrepancy between the *nominal* error probabilities (valid premises), and the *actual* error probabilities (misspecified premises), giving rise to unreliable inferences (see Spanos and McGuirk, 2001, Spanos, 2005).

Although the nature of the F–N–P statistical induction became clear by the late 1930s, the form of the underlying *inductive reasoning* was clouded by a disagreement between the two protagonists (see Mayo, 2005). Fisher argued for 'inductive inference' spearheaded by his significance testing (see Fisher, 1955; 1956), and Neyman argued for 'inductive behaviour' based on Neyman–Pearson testing (see Neyman, 1956; Lehmann, 1993; Cox, 2006). Neither account, however, gave satisfactory answers to the question 'when do data **Z** provide *evidence* for (or against) a hypothesis or a claim *H*?' The *pre-data* error–probabilistic account of inference seemed inadequate for a *post-data* evaluation of the inference reached to provide a clear evidential interpretation of the results (see Hacking, 1965).

The F–N–P paradigm, in addition to (*a*) the pre-data as against post-data error probabilities, still grapples with some additional philosophical/methodological issues including (*b*) the fallacies of acceptance and rejection (for example statistical as against substantive significance), (*c*) double use of data, (*d*) statistical model selection (specification) as against model validation, (*e*) structural as against statistical models. These and other methodological issues have been extensively debated in other social sciences such as psychology and sociology (see Morrison and Henkel, 1970; Lieberman, 1971; Godambe and Sprott, 1971), but largely ignored in economics until recently.

Mayo (1996) argued convincingly that some of these chronic methodological issues and problems can be addressed by supplementing the Neyman–Pearson approach to testing (see Pearson, 1966) with a post-data assessment of inference based on *severe testing reasoning*. This extended frequentist approach to inference, called *the error-statistical approach*, has been used by Mayo (1991) to address (*c*), by Mayo and Spanos (2006) to address the fallacies of acceptance and rejection, and by Spanos (2006b; 2007) to deal with the issues (*d*) and (*e*), respectively.

4.3 Economic statistics in the early 20th century

In the 1930s applied economists were more keyed to Bowley's traditional view of economic statistics than to F–N–P statistical inference perspective. Indeed, Bowley was elected president (the first from Britain) of the Econometric Society for 1938–9. The more economics-oriented 'statistics textbooks' written in the 1920s and 1930s, including Bowley (1920/1926/1937), Mills (1924/1938), Ezekiel (1930), Davis and Nelson (1935) and Secrist (1930), largely ignored the new statistical inference paradigm. Their perspective was primarily one of 'descriptive statistics', supplemented with the Pearson–Yule curve-fitting perspective on correlation and regression, and certain additional focus on the analysis of time series data, including index numbers (see Persons, 1925).

Economic statistics, as exemplified in Mills (1924), provided the framework for the work at the National Bureau of Economic Research (NBER), of which Mills was a staff member. The empirical work on business cycles by Burns and Mitchell (1946) represents an excellent use of descriptive statistics in conjunction with graphical methods, as understood at the time. Their detailed, carefully crafted and painstaking statistical analysis of business cycles, however, suffers from the same crucial weakness as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
all descriptive statistics: the premises for inductive inference (the underlying statistical model) is not explicitly specified, and as a result one cannot assess the reliability of inferences based on such statistics. For instance, without clearly specified probabilistic premises one can easily misidentify temporal dependence type cycles with regular business cycles (see Spanos, 1999).

The conventional wisdom at the time is summarized by Mills (1924) in the form of a distinction between ‘statistical description vs. statistical induction’. In statistical description measures such as the sample mean $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$, the sample variance $s_x^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x}_n)^2$, the correlation $r = \frac{\sum_{k=1}^{n}(x_k - \bar{x}_n)(y_k - \bar{y}_n)}{\sqrt{\left[\sum_{k=1}^{n}(x_k - \bar{x}_n)^2\right]\left[\sum_{k=1}^{n}(y_k - \bar{y}_n)^2\right]}}$, and so on, ‘provide just a summary for the data in hand’ and ‘may be used to perfect confidence, as accurate descriptions of the given characteristics’ (1924, p. 549). However, when the results are to be extended *beyond* the data in hand – statistical induction – their validity depends on certain inherent a priori assumptions such as (*a*) the ‘uniformity’ for the *population* and (*b*) the ‘representativeness’ of the *sample* (1924, pp. 550–2).

A methodological digression. Unfortunately, Mills's misleading argument concerning descriptive statistics lingers on even today. The reality is that there are *appropriate* and *inappropriate* summaries of the data, which depend on the inherent probabilistic structure of the data. For instance, if data $\{(x_k, y_k), k=1,\ldots,n\}$ are *trending*, like most economic time series, the summary statistics $(\bar{x}, s_x^2, \bar{y}, s_y^2, r)$ represent artefacts – highly misleading descriptions of the features of the data in hand. When viewed in the context of a probabilistic framework, $(\bar{x}, s_x^2, \bar{y}, s_y^2, r)$ are unreliable estimators of $E(X_k)$, $E(Y_k)$, $Var(X_k)$, $Var(Y_k)$, $Corr(X_k, Y_k)$; they provide reliable and precise estimates only when certain probabilistic assumptions concerning the underlying the vector process $\{(X_k, Y_k)\ k \in \mathbb{N}\}$, such as independent and identically distributed (IID), are valid for the data in hand. Any departures from these premises require one to qualify the reliability and precision of these estimates. In an important sense one of Fisher's lasting contribution to statistics was to (*a*) make the IID assumptions explicit as part of the problem of specification, by formalizing Mills's a priori ‘uniformity’ and ‘representativeness’ assumptions, and (*b*) render them empirically testable. It is important to note that ignoring statistical adequacy is a very different criticism of Burns and Mitchell than that of Koopmans (1947); see below.

The paper by Yule (1926), entitled ‘Why Do We Sometimes get Nonsense Correlations between time series’?, provided a widely discussed wakeup call in economics, because it raised serious doubts about the appropriateness of the linear regression model when the data $\{(x_k, y_k), k=1,\ldots,n\}$ constitute time series, by pointing out the risk of getting *spurious* results. As commented in Spanos (1989b), the source of the spurious (nonsense) correlation problem is *statistical inadequacy* (see Section 4.8 below). Yule's (1927) autoregressive (AR(p)) and Slutsky's (1927) moving average (MA(q)) models can be viewed as attempts to specify statistical models to capture the temporal dependence in time series data.

Stochastic processes. The AR(p) and MA(q) models were given proper probabilistic underpinnings by Wold (1938) using the newly developed theory of stochastic processes by Kolmogorov and Khitchin in the early 1930s (see Doob, 1953). This was a crucial and timely development in probability theory which extended significantly the intended scope of the F–N–P approach beyond the original IID frame-up, by introducing several dependence and heterogeneity concepts, such as Markov dependence,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
stationarity and ergodicity (see Spanos, 1999, ch. 8).

4.4 The Econometric Society and the Cowles Commission

The vision statement of the Econometric Society founded in 1930 read:

> Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems. (Frisch, 1933, p. 106)

The impression among quantitatively oriented economists in the early 1930s was that the F–N–P sampling theory methods were inextricably bound up with agricultural experimentation. It was generally believed that these methods are relevant only for analysing ‘random samples’ of experimental data, as Frisch argued:

> In problems of the kind encountered when the data are the result of *experiments* which the investigator can control, the sampling theory may render very valuable services. Witness the eminent works of R.A. Fisher and Wishart on problems of agricultural experimentation. (Frisch, 1934, p. 6)

In place of the statisticians’ linear regression Frisch proposed his errors-in-variables scheme, which treated all observable variables symmetrically by decomposing them into a latent systematic (deterministic) component and a white-noise error with economic theory providing relationships among the systematic components. Fisher's reaction to Frisch's scheme was that economists were perpetuating a major *confusion* between ‘statistical’ regression coefficients and ‘coefficients in abstract economic laws’ (see Bennett, 1990, p. 305).

Tinbergen's (1939) empirical modelling efforts were in the spirit of the Pearson–Yule curve-fitting tradition, which paid little attention to the validity of the premises of inference. In reviewing this work Keynes (1939) destructively criticized the use of regression in econometrics and raised numerous substantive and statistical problems, but not the reliability of inference problem (see Spanos, 2006a).

The first attempt to bring together Frisch's errors-in-variables scheme with Fisher's linear regression model was made by Koopmans (1939), which had no success. Koopmans’ primary influence on econometrics was as a leading figure in the Cowles Commission in Chicago in the 1940s (see Heckman, 1992).

The first successful attempt to bring the F–N–P methods into econometrics modelling was made by Haavelmo (1944), who argued convincingly against the prevailing view that sampling methods are only applicable to random samples of experimental data (see Spanos, 1989a). Contrary to this view, the F–N–P perspective provides the proper framework for modelling time series data which exhibit both dependence and heterogeneity:

> For no tool developed in the theory of statistics has any meaning … without being referred

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> to some stochastic scheme. (Haavelmo, 1944, p. iii)
>
> … economists might get more useful and reliable information (and also fewer spurious results) out of their data by adopting more clearly formulated probabilistic models. (1944, p. 114)

The part of Haavelmo's monograph that had the biggest impact on the development of econometrics was, however, the technical ‘solution’ to the simultaneity problem that was formalized and extended by the Cowles Commission in the form of the simultaneous equations model (SEM): see Koopmans, 1950. Despite the introduction of frequentist methods of inference by the Cowles Commission, the theory-driven specification of the *structural model*:

$$\Gamma^{\top}\mathbf{y}_k = \Delta^{\top}\mathbf{x}_k + \boldsymbol{\varepsilon}_k,\ \boldsymbol{\varepsilon}_k \sim \mathrm{N}(0, \boldsymbol{\Omega}),\quad k \in \mathbb{N}, \tag{3}$$

(using the traditional notation, see Spanos, 1986), leaves any inferences concerning the structural parameters (Γ, Δ, Ω) highly susceptible to the unreliability of inference problem. *Methodological digression*. The unreliability of inference arises primarily because it is often insufficiently appreciated that the statistical reliability of such inference depends crucially on the *statistical adequacy* of the (implicit) *reduced form model*:

$$\mathbf{y}_k = \mathbf{B}^{\top}\mathbf{x}_k + \mathbf{u}_k,\ \mathbf{u}_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}),\quad k \in \mathbb{N}. \tag{4}$$

That is, unless (4), viewed as multivariate linear regression model (assumptions [1]–[5] in Table 1 in vector form), is statistically adequate ([1]–[5] are valid for the data in question), any inference based on (3) is likely to be unreliable. Note that *identification* refers to being able to define the structural parameters (Γ, Δ, Ω) uniquely in terms of the statistical parameters $(\mathbf{B}, \boldsymbol{\Sigma})$. In practice (4) is not even estimated explicitly, let alone have its assumptions [1]–[5] tested thoroughly before drawing any inferences concerning (Γ, Δ, Ω) (see Spanos, 1986; 1990). A more expedient way one that highlights the reliability issue, is to view (3) as a structural model which is embedded into the statistical model (4), giving rise to a special type of substantive information restrictions. Hence, the theory-dominated perspective of the Cowles Commission, despite the importance of the technical innovations introduced in dealing with simultaneity, has (inadvertently) undermined the problem of statistical adequacy in empirical modelling (see Spanos, 2006a). As argued by Heckman:

> The Haavelmo–Cowles way of doing business – to postulate a class of models *in advance*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> *of looking at the data* and to consider identification problems within the prescribed class – denies one commonly used process of inductive inference that leads to empirical discovery.•…•The Haavelmo program as interpreted by the Cowles Commission scholars refocused econometrics away from the act of empirical discovery and toward a sterile program of hypothesis testing and rigid imposition of a priori theory onto the data. (Heckman, 1992, pp. 883–4)

Koopmans (1947), in reviewing Burns and Mitchell (1946), criticized their focusing on the purely empirical nature of their results without any guidance from economic theory. He pronounced their empirical findings as representing the 'Kepler stage' of data analysis, in contrast to the 'Newton stage', where the original empirical regularities were given a structural (theoretical) interpretation using the law of universal gravitation (LUG). What Koopmans (1947) neglected to point out is that it was not the theory that guided Kepler to the regularities, but the statistical regularities exhibited by the data. Indeed, Kepler established these regularities 60 years before Newton was inspired by them to come up with his LUG. The Cowles Commission approach, which Koopmans misleadingly associates with the Newton stage, was equally (if not more) vulnerable to the reliability of inference problem. There is no reason to believe that the reduced form (4) implied by the structural form (3), which was specified in complete ignorance of the probabilistic structure of the data, will constitute a statistically adequate model. The specification of statistical models relying exclusively on substantive information is not conducive to reliable/precise inferences. The crucial difference between Kepler's empirical results and those in Burns and Mitchell (1946) and Klein (1950) – based largely on Koopmans's preferred approach – is that Kepler's constitute real statistical regularities in the sense that his estimated model of elliptical motion, viewed retrospectively in the context of the linear regression model (Table 1), is statistically adequate; assumptions [1]–[5] are valid for his original data (see Spanos, 2008).

4.5 Textbook econometrics. the Gauss – Markov perspective

The textbook approach to econometrics was largely shaped in the early 1960s by two very successful textbooks by Johnston (1963) and Goldberger (1964) by viewing the SEM as an extension/modification of the classical linear model. These textbooks demarcated the intended scope of econometrics to be the 'quantification of theoretical relationships', and reverted back to the 'curve-fitting' perspective of the Legendre–Gauss 19th century tradition, instead of adopting the F–N–P perspective (see Spanos, 1995; 2007).

The cornerstone of textbook econometrics is the so-called *Gauss–Markov theorem*, which is based on the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \mathbf{0},\ E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\top}) = \sigma^2\mathbf{I}_n, \tag{5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $\mathbf{I}_n$ is the identity matrix. In the context of (5), Gauss in 1823 (see Hald, 2007) proved that the least squares estimator $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ has minimum variance within the class of *linear* and *unbiased* estimators of β. For the sake of historical accuracy it is important to point out that Markov had nothing to do with this theorem (see Neyman, 1952, p. 228). This theorem, and the perspective it exemplifies, provide the central axis around which textbook econometrics revolves (see Greene, 2003).

A methodological digression. Spanos (1986) challenged the traditional interpretation that the Gauss–Markov theorem provides a formal justification for least squares via the optimality of the estimators it gives rise to, arguing that the results of this theorem provide a poor basis for *reliable* and *precise* inference. This is primarily because the Gauss–Markov theorem yields the mean and variance of $\hat{\boldsymbol{\beta}}_{LS}$ but *not* its sampling distribution, that is $\hat{\boldsymbol{\beta}}_{LS} \overset{?}{\sim} D(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$. Hence, even the simplest forms of inference, like testing $H_0: \boldsymbol{\beta} = \mathbf{0}$ would require one to use either inequalities like Chebyshev's to approximate the relevant error probabilities (Spanos, 1999, pp. 550–2), or invoke asymptotic approximations; neither method would, in general, give rise to reliable and precise inferences (Spanos, 2006a, pp. 46–7).

The Gauss–Markov ‘curve-fitting’ perspective promotes ‘saving the theory’ by attributing the stochastic structure to the error term and favouring broad premises (weak assumptions) in an attempt to protect the inference from the perils of misspecification. This move, however, relegates the essentialness of ensuring the reliability and precision of inference. Weak assumptions, such as the Gauss–Markov assumptions in (5), do not guarantee reliable inferences, but they usually give rise to much less precise inferences than specific premises comprising assumptions such as [1]–[5] (Table 1): Spanos, 2006a. As perceptively noted by Heckman:

> In many influential circles, ambiguity disguised as simplicity or ‘robustness’ is a virtue. The less said about what is implicitly assumed about a statistical model generating data, the less many economists seem to think is being assumed. The new credo is to let sleeping dogs lie. (Heckman, 1992, p. 882)

In addition, the ‘error-fixing’ strategies of the textbook approach, designed to deal with departures from the linearity, homoskedasticity, no-autocorrelation assumptions, do not usually address the reliability of inference problem (Spanos and McGuirk, 2001).

Some of the important technical developments in both econometrics and statistics since the 1980s, such as the *generalized method of moments* (see Hansen, 1982), as well as certain *nonparametric* (see Pagan and Ullah, 1999) and *semiparametric* methods (see Horowitz, 1998), are motivated by this Gauss–Markov perspective. These methods, although very useful for a number of different aspects of empirical modelling, do not provide the answer to statistical misspecification, and often compromise the reliability/precision of substantive inferences (see Spanos, 1999, pp. 553–5).

4.6 Demarcating the boundaries of modern statistics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As argued above, the F–N–P perspective has been largely ignored in empirical modelling in economics, despite the wholesale adoption of Fisher's estimation and the Neyman–Pearson testing methods. One of the primary obstacles has been the problem of blending the substantive subject matter and statistical information and their roles in empirical modelling. Many aspects of empirical modelling, in both the physical and social sciences, implicate both sources of information in a variety of functions, and others involve one or the other, more or less separately. For instance, the development of *structural* (theoretical) *models* is primarily based on *substantive information*; that activity, by its very nature, cannot be separated from the disciplines in question, but where does this leaves statistics? It renders the problem of demarcating its boundaries as a separate discipline extremely difficult (see Lehmann, 1990; Cox, 1990).

A methodological digression. Spanos (2006c) argued that the lessons learned in blending the substantive and statistical information in econometric modelling can help delineate the boundaries of statistics as a separate discipline. Certain aspects of empirical modelling, which focus on *statistical information* and are concerned with the nature and use of statistical models, can form a body of knowledge that is shared by all applied fields. *Statistical model specification, the use of graphical techniques* (going back to Playfair), *misspecification* (*M-S*) *testing and respecification*, together with the relevant inference procedures, constitute aspects of statistical modelling that can be developed *generically* without requiring any information concerning ‘what substantive variables the data **Z** quantify or represent’. All these aspects of empirical modelling belong to the *realm of statistics* and can be developed generically without any reference to substantive subject matter information. This, in a sense, will broaden the scope of modern statistics because the current literature and textbooks pay little attention to some of these aspects of modelling (see Cox and Hinkley, 1974).

The statistical and substantive information can be amalgamated, without compromising their integrity, by embedding structural models into adequate statistical models, which would provide the premises for statistical inference. That is, the substantive restrictions need to be thoroughly tested and accepted in the context of the statistical model in order for the resulting empirical model to enjoy both statistical and substantive meaning (see Spanos, 2006b; 2007).

4.7 The Box – Jenkins turn in statistics

An important development in statistics that had a lasting effect on econometrics and created a tension with textbook econometrics, was the publication of Box and Jenkins (1970). Building on the work of Wold (1938), they proposed a new statistical perspective on time series modelling which placed it within the F–N–P modelling framework where the premises of inference is specified by a *statistical model*. In addition to transforming descriptive time series analysis into statistical inference proper, the Box–Jenkins approach introduced several noteworthy innovations into empirical modelling that influenced empirical modelling in economics.

1. (i) Modelling begins with a family of *statistical models* in the form of the ARIMA(p,d,q):

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$y_t^* = \alpha_0 + \sum_{k=1}^{p} \alpha_k y_{t-k}^* + \sum_{\ell=1}^{q} \beta_\ell \epsilon_{t-\ell} + \epsilon_t,\ \epsilon_t \sim \mathrm{NIID}(0, \sigma^2),\ \ t \in \mathbb{N}, \tag{6}$$

where $y_t^* := \Delta^d y_t$, that was thought to capture adequately the temporal dependence and heterogeneity (including seasonality) in time series data.

2. (ii) Statistical modelling was viewed as an *iterative process* that involves several stages, *identification*, *estimation*, *diagnostic checking*, and *prediction*.
3. (iii) *Diagnostic checks*, based on the residuals from the fitted model, offered a way to detect model inadequacies with a view to improve the original model.
4. (iv) *Exploratory data analysis* (EDA) was legitimized as providing an effective way to select (identify) a model within the ARIMA(p,d,q) family.
5. (v) The deliberate choice of a more *general specification* in order to put the model 'in jeopardy' (see Box and Jenkins, 1970, p. 286) is exploited in assessing the adequacy of a selected model.

The Box–Jenkins approach constituted a major departure from the rigid textbook approach, where the model is assumed to be specified by economic theory in advance of any data. Indeed, the predictive success of the ARIMA(p,d,q) models in the 1970s exposed the statistical inadequacy of traditional econometric models, sending the message that econometric models could ignore the temporal dependence and heterogeneity of times series data at their peril (see Granger and Newbold, 1986). The weaknesses of traditional econometric modelling techniques brought out by the Box–Jenkins modelling motivated several criticisms from within econometrics, including those by Hendry (1977) and Sims (1980), that led to the autoregressive distributed lag (ADL(p,q)) and the vector autoregressive (VAR(p)) family of models, respectively. The LSE tradition (see Hendry, 1993), embraced and extended the Box–Jenkins innovations (i)–(v), rendering the general-to-specific approach the backbone of its empirical modelling methodology (see Hendry, 1995).

4.8 Unit roots and cointegration

The Box–Jenkins ARIMA(p,d,q) modelling approach raised the question 'how does one decide on the value of $d \geq 0$ in $\Delta^d y_t$, that is appropriate to induce stationarity?' It turned out that the value of d is related to the number of unit roots in the $AR(m)$ representation, $y_t = \gamma_0 + \sum_{k=1}^{m} \gamma_k y_{t-k} + u_t$, of the underlying stochastic process $\{y_t, t \in \mathbb{N}\}$. Efforts to answer this question led to the unit root 'revolution', initiated by Dickey and Fuller (1979) in the statistics literature. This had an immediate impact on the econometrics literature, which generalized and extended the initial results in a number of different directions (see Phillips and Durlauf, 1986; Phillips, 1987). This literature eventually led to further important developments, which brought out a special relationship (cointegration) among unit root

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

processes and error-correction models (see Engle and Granger, 1987; Johansen, 1991; Hendry, 1995).

A methodological digression. The (non-standard) sampling distribution results associated with unit roots were used by Phillips (1986) to shed light on the chronic problem of spurious regression raised by Yule (1926). This problem was revisited by Granger and Newbold (1974) using simulations of time series data $\{(x_t, y_t), t=1,\ldots,n\}$ generated by two *uncorrelated* Normal unit root processes:

$$y_t = y_{t-1} + \varepsilon_{1t},\ x_t = x_{t-1} + \varepsilon_{2t},$$

$$E(\varepsilon_{1t}) = 0,\ \ E(\varepsilon_{2t}) = 0,\ \ E(\varepsilon_{1t}^2) = \sigma_{11},\ \ E(\varepsilon_{2t}^2) = \sigma_{22},\ \ E(\varepsilon_{1t}\varepsilon_{2t}) = 0.$$

Their results demonstrated that when these data were used to estimate the linear regression model, $y_t = \beta_0 + \beta_1 x_t + u_t$, the inferences based on the estimated model were completely unreliable. In particular, they noted a huge discrepancy between the *nominal* (α =.05) and *actual* ($\hat{\alpha} = .76$) error probabilities when testing the hypothesis β_1=0.

In a very influential paper, Phillips (1986) explained this by deriving analytically the (non-standard) sampling distributions of the least-squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ under the above unit root scheme, showing how different they were from the assumed distributions. What was not sufficiently appreciated was that the discrepancy between the nominal and actual error probabilities is a classic symptom of unreliable inferences emanating from a statistically misspecified model, that is *misspecification*, due to ignoring the temporal dependence/heterogeneity in the data, is the real source of spurious regression. One would encounter similar unreliabilities when the data exhibit deterministic trends or/and Markov dependence, or/and non-Normalities (see Spanos and McGuirk, 2001, Spanos, 2005). Deriving the sampling distributions under all scenarios of possible misspecifications is impractical (there is an infinity of such scenarios), and does *not* address the unreliability of inference issue. What is needed is to respecify the original model to account for the disregarded information that gave rise to the detected departures. For instance, for the above Granger and Newbold data, if one were to estimate the dynamic linear regression model:

$$y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 x_{t-1} + \alpha_3 y_{t-1} + \varepsilon_t,\ t \in \mathbb{N},$$

the above noted unreliabilities would disappear (see Spanos, 2001).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

4.9 Recent developments in microeconometrics

Arguably, some of the most important developments in econometrics since 1980 have taken place in an area broadly described as microeconometrics (see Manski and MacFadden, 1981; Heckman and Singer, 1984; Cameron and Trivedi, 2005) for a recent textbook survey. This area includes *discrete and limited dependent* and *duration models* for *cross-section data*, as well as *panel data* models. The roots of these statistical models go back to the statistical literature on the probit/logit and analysis of variance models (see Agresti, 2002, ch. 16), but they have been generalized, extended and adapted for economic data.
A welcome facet of microeconometrics is the specification of statistical models that often takes into consideration the probabilistic structure of the data (see Heckman, 2001). Unfortunately, this move does not often go far enough in securing statistical adequacy. This becomes apparent when one asks, ‘what are the probabilistic assumptions providing a complete specification for the probit/logit, duration and the fixed and random effect models?’ Without such complete specifications, one would not even know what potential errors to probe for to secure statistical adequacy.
While these developments in microeconometrics are of great importance, their potential value has been offset by the insufficient attention paid to the task of ensuring reliability and precision of inference. Their statistical results are still largely dominated by the Gauss–Markov perspective, in the sense that:

1. (i) the probabilistic structure of the models in question is specified, almost exclusively, in terms of unobservable *error terms*,
2. (ii) the error probabilistic assumptions are often vague and incomplete, and invariably involve non-testable orthogonality conditions,
3. (iii) the statistical analysis focuses primarily on constructing consistent and asymptotically Normal estimators, and
4. (iv) respecification is often confined to ‘error-fixing’.

In view of (i)–(iv), even questions of ensuring statistical adequacy cannot be posed unequivocally for these statistical models.
Spanos (2006a; 2006d) proposed complete specifications for these statistical models in terms of probabilistic assumptions relating to the observable stochastic processes involved, but there is a long way to go to develop adequate misspecification testing and the respecification results needed to ensure the reliability and precision of inference when applying these statistical models to actual data.

5 Conclusion

The demise of political arithmetic by the end of the 18th century, due to the unreliability of the inferences its methods gave rise to, contains important lessons for both economics and statistics. Petty's attitude of ‘seeking figures that will support a conclusion already reached by other means’ lingers on in applied econometrics more than three centuries later. The problem then was that, in addition to the quality and the accuracy of data, the probabilistic underpinnings of establishing statistical regularities were completely lacking. Fisher's recasting of statistical induction has changed that, and it is now known

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that the explicit specification of the *statistical model* enables one to (*a*) assess the validity of the premises for inductive inference, and (*b*) provide relevant error probabilities for assessing the reliability of ensuing inferences. It has taken several decades to understand how one can assess the model assumptions vis-à-vis the observed data using *misspecification tests* (see Spanos, 1999), but one hopes it will take less time before modellers understand the necessity to implement such tests with the required care and thoroughness to ensure the *reliability* of the resulting statistical inferences (see Spanos, 2006a). The Box–Jenkins modelling approach exposed the inattention to statistical adequacy in traditional econometric modelling and strengthened the call for adopting the F–N–P perspective. This will bring modern statistical inference closer to econometrics to the benefit of both disciplines. Careful implementation of this perspective will certainly improve the reliability of empirical evidence in economics and other applied disciplines. Moreover, the *ab initio* separation of the statistical and substantive information can help demarcate and extend the intended scope of statistics. The error-statistical extension/modification of frequentist statistics (Mayo, 1996) can address some of the inveterate problems concerning inductive reasoning and broaden the intended scope of statistical inference in these disciplines by enabling one to consider questions of substantive adequacy, shedding light on causality issues, omitted variables and confounding effects (see Spanos, 2006b).

See Also

- Bowley, Arthur Lyon
- Davenant, Charles
- Edgeworth, Francis Ysidro
- Fisher, Ronald Aylmer
- King, Gregory
- Petty, William

Bibliography

Agresti, A. 2002. *Categorical Data Analysis*, 2nd edn. New York: Wiley.

Backhouse, R.E. 2002a. *The Ordinary Business of Life: A History of Economics from the Ancient to the Twenty-First Century*. Princeton, NJ: Princeton University Press.

Backhouse, R.E. 2002b. *The Penguin History of Economics*. London: Penguin Books.

Bartlett, M.S. 1935. Some aspects of the time-correlation problem in regard to tests of significance. *Journal of the Royal Statistical Society* 98, 536–43.

Bennett, J.H. 1990. *Statistical Inference and Analysis: Selected correspondence of R.A. Fisher*. Oxford: Clarendon Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bernoulli, J. 1713. *Ars Conjectandi*. Basilea: Thurnisius, tr. E. D. Sylla, Baltimore: Johns Hopkins University Press, 2006.

Bowley, A.L. 1902, 1920, 1926, 1937. *Elements of Statistics*, 2nd, 4th, 5th and 6th edns. London: Staples Press.

Bowley, A.L. 1906. Address to the economic science and statistics. *Journal of the Royal Statistical Society* 69, 540–58.

Bowley, A.L. 1928. *F.Y. Edgeworth's Contributions to Mathematical Statistics*. Clifton: Augustus M. Kelley.

Box, G.E.P. and Jenkins, G.M. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Box, J.F. 1978. *R.A. Fisher: The Life of a Scientist*. New York: Wiley.

Burns, A.F. and Mitchell, W.C. 1946. *Measuring Business Cycles*. New York: NBER.

Cameron, A.C. and Trivedi, P.K. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Cournot, A. 1843. *Exposition de la théorie des chances et des probabilités*. Paris: Hachette.

Cox, D.R. 1990. Role of models in statistical analysis. *Statistical Science* 5, 169–74.

Cox, D.R. 2006. *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

Cox, D.R. and Hinkley, D.V. 1974. *Theoretical Statistics*. London: Chapman & Hall.

Cramer, H. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Davis, H.T. and Nelson, W.F.C. 1935. *Elements of Statistics*. Indiana: Principia Press.

Desrosières, A. 1998. *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Dickey, D.A. and Fuller, W.A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.

Doob, J.L. 1953. *Stochastic Processes*. New York: Wiley.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Engle, R.F. and Granger, C.W.J. 1987. Co-integration and error-correction: representation, estimation and testing. *Econometrica* 55, 251–76.

Ezekiel, M. 1930. *Methods of Correlation Analysis*, 2nd edn. New York: Wiley.

Farebrother, R.W. 1999. *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. New York: Springer.

Fisher, I. 1947. Response of Irving Fisher. *Journal of the American Statistical Association* 42, 4–5.

Fisher, R.A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 507–21.

Fisher, R.A. 1921. On the ‘probable error’ of a coefficient deduced from a small sample. *Metron* 1, 2–32.

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A* 222, 309–68.

Fisher, R.A. 1922a. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society* 85, 87–94.

Fisher, R.A. 1922b. The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society* 85, 597–612.

Fisher, R.A. 1924. The conditions under which χ^2 measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society* 87, 442–50.

Fisher, R.A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R.A. 1935. The logic of inductive inference. *Journal of the Royal Statistical Society* 98, 39–54, 55–82.

Fisher, R.A. 1935a. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fisher, R.A. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society, B* 17, 69–78.

Fisher, R.A. 1956. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd.

Frisch, R. 1933. Editorial. *Econometrica* 1, 1–4.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Frisch, R. 1934. *Statistical Confluence Analysis by Means of Complete Regression Schemes*. Oslo: Universitetets Okonomiske Institutt.

Galton, F. 1869. *Hereditary Genius: An Inquiry into its Laws and Consequences*. London: Macmillan.

Gauss, C.F. 1809. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Hamburg: F. Perthes & I.H. Besser.

Glass, D.V. 1973. *Numbering the People*. Farnborough: Saxton House.

Godambe, V.P. and Sprott, D.A. 1971. *Foundations of Statistical Inference: A Symposium*. Toronto: Holt, Rinehart and Winston.

Goldberger, A.S. 1964. *Econometric Theory*. New York: Wiley.

Gossett, W.S. 1908. The probable error of the mean. *Biometrika* 6, 1–25.

Granger, C.W.J. and Newbold, P. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.

Granger, C.W.J. and Newbold, P. 1986. *Forecasting Economic Time Series*, 2nd edn. London: Academic Press.

Graunt, J. 1662. *Natural and Political Observations upon the Bills of Mortality*. London: John Martyn. Repr. in Hull (1899).

Greene, W.H. 2003. *Econometric Analysis*, 5th edn. Englewood Cliffs, NJ: Prentice-Hall.

Greenwood, M. 1948. *Medical Statistics from Graunt to Farr*. Cambridge: Cambridge University Press.

Haavelmo, T. 1944. The probability approach to econometrics. *Econometrica* 12(Suppl), 1–115.

Hacking, I. 1965. *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hacking, I. 1975. *The Emergence of Probability*. Cambridge: Cambridge University Press.

Hald, A. 1990. *History of Probability and Statistics and Their Applications before 1750*. New York: Wiley.

Hald, A. 1998. *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hald, A. 2007. *A History of Parametric Statistical Inference from Bernoulli to Fisher*. New York: Springer.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 97, 93–115.

Harrison, B. 2004. *Oxford Dictionary of National Biography*. Oxford: Oxford University Press.

Heckman, J.J. 1992. Haavelmo and the birth of modern econometrics: a review of the history of econometric ideas by Mary Morgan. *Journal of Economic Literature* 30, 876–86.

Heckman, J.J. 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* 109, 673–748.

Heckman, J.J. and Singer, B. 1984. Econometric duration analysis. *Journal of Econometrics* 24, 63–132.

Hendry, D.F. 1977. On the time series approach to econometric model building. In *New Methods in Business Cycle Research*, ed. C.A. Sims. Minnesota: Federal Reserve Bank of Minneapolis.

Hendry, D.F. 1993. *Econometrics: Alchemy or Science?* Oxford: Blackwell.

Hendry, D.F. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D.F. and Morgan, M.S. 1995. *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press.

Hilts, V.L. 1973. Statistics and social science. In *Foundations of Scientific Method: the Nineteenth Century*, ed. R.N. Giere and R.S. Westfall. Bloomington, IN: Indiana University Press.

Hilts, V.L. 1978. Aliss exterendum, or, the origins of the Statistical Society of London. *Isis* 69, 21–43.

Hotelling, H. 1930. British statistics and statisticians today. *Journal of the American Statistical Association* 25, 186–90.

Hoppit, J. 1996. Political arithmetic in eighteenth-century England. *Economic History Review* 49, 516–40.

Horowitz, J.L. 1998. *Semiparametric Methods in Econometrics*. New York: Springer.

Hull, H.C. 1899. *The Economic Writings of Sir William Petty*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hume, D. 1987. *Essays, Moral, Political and Literary*, ed. E.F. Miller Indianapolis, ID: Liberty Fund.

JSSL (Journal of the Statistical Society of London). 1834. Prospects of the objects and plan of operation of the statistical society of London. Reprinted in *Journal of the Statistical Society of London*, 1869, 385–7.

Jevons, W.S. 1871. *The Theory of Political Economy*. New York, NY: Augustus M. Kelley, 1911.

Johansen, S. 1991. Estimation and hypothesis testing of cointegrating vector of Gaussian vector autoregressive models. *Econometrica* 59, 1551–80.

Johnston, J. 1963. *Econometric Methods*. New York: McGraw-Hill.

Keynes, J.M. and Tinbergen, J. 1939–40. Professor's Tinbergen's method. *Economic Journal* 49, 558–68; A Reply, by J. Tinbergen, and Comment by Keynes, 50, 141–56.

Keynes, J.N. 1890/1917. *The Scope and Method of Political Economy*. Fairfield, NJ: Augustus M. Kelley.

Klein, L.R. 1950. *Economic Fluctuations in the United States 1921–1941*. Cowles Commission for Research in Economics, Monograph No. 11. New York: Wiley.

Kolmogorov, A.N. 1933. *Foundations of the theory of Probability*, 2nd edn. New York: Chelsea.

Koopmans, T.C. 1939. *Linear Regression Analysis of Economic Time Series*. Netherlands Economic Institute, Publication No. 20. Haarlem: F. Bohn.

Koopmans, T.C. 1947. Measurement without theory. *Review of Economics and Statistics* 29, 161–72.

Koopmans, T.C. 1950. *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph, No. 10. New York: Wiley.

Kreager, P. 1988. New light on Graunt. *Population Studies* 42, 129–40.

Legendre, A.M. 1805. *Nouvelles Méthodes pour la détermination des orbites des comètes*. Paris: Courcier.

Lehmann, E.L. 1990. Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science* 5, 160–8.

Lehmann, E.L. 1993. The Fisher and Neyman–Pearson theories of testing hypotheses: one theory or

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
two? *Journal of the American Statistical Association* 88, 1242–9.

Lehmann, E.L. 1999. ‘Student’ and small-sample theory. *Statistical Science* 14, 418–26.

Letwin, W. 1965. *The Origins of Scientific Economics*. New York: Anchor Books.

Lieberman, B. 1971. *Contemporary Problems in Statistics: A Book of Readings for the Behavioral Sciences*. Oxford: Oxford University Press.

Manski, C.F. and MacFadden, D. 1981. *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.

Mayo, D.G. 1991. Novel evidence and severe tests. *Philosophy of Science* 58, 523–52.

Mayo, D.G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, D.G. 2005. Philosophy of statistics. In *Philosophy of Science: An Encyclopedia*, ed. S. Sarkar and J. Pfeifer. London: Routledge.

Mayo, D.G. and Spanos, A. 2004. Methodology in practice: statistical misspecification testing. *Philosophy of Science* 71, 1007–25.

Mayo, D.G. and Spanos, A. 2006. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science* 57, 323–57.

Mills, F.C. 1924. *Statistical Methods*. New York: Henry Holt, 1938.

Moore, H.L. 1908. The statistical complement of pure economics. *Quarterly Journal of Economics* 23, 1–33.

Moore, H.L. 1911. *The Law of Wages*. New York: Macmillan.

Morgan, M.S. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Morrison, D.E. and Henkel, R.E. 1970. *The Significance Test Controversy: A Reader*. Chicago: Aldine.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97, 558–625.

Neyman, J. 1935. On the problem of confidence intervals. *Annals of Mathematical Statistics* 6, 111–6.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Neyman, J. 1950. *First Course in Probability and Statistics*. New York: Henry Holt.

Neyman, J. 1952. *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd edn. Washington, DC: U.S. Department of Agriculture.

Neyman, J. 1956. Note on an article by Sir Ronald Fisher. *Journal of the Royal Statistical Society B* 18, 288–94.

Neyman, J. 1977. Frequentist probability and frequentist statistics. *Synthèse* 36, 97–131.

Neyman, J. and Pearson, E.S. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*. Series A 231, 289–337.

Pagan, A.R. and Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Pearson, E.S. 1931a. The analysis of variance in cases of non-normal variation. *Biometrika* 23, 114–33.

Pearson, E.S. 1931b. Note on tests for normality. *Biometrika* 22, 423–24.

Pearson, E.S. 1955. Statistical concepts in the relation to reality. *Journal of the Royal Statistical Society*. Series B 17, 204–7.

Pearson, E.S. 1966. The Neyman–Pearson story: 1926–34. In *Research Papers in Statistics: Festschrift for J. Neyman*, ed. F.N. David. New York: Wiley.

Pearson, K. 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London*. Series A 187, 253–318.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* 50, 157–175.

Peirce, C.S. 1878. The probability of induction. *Popular Science Monthly* 12, 705–18.

Persons, W.M. 1925. Statistics and economic theory. *Review of Economics and Statistics* 7, 179–97.

Phillips, P.C.B. 1986. Understanding spurious regression in econometrics. *Journal of Econometrics* 33, 311–40.

Phillips, P.C.B. 1987. Time series regressions with a unit root. *Econometrica* 55, 227–301.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Phillips, P.C.B. and Durlauf, S.N. 1986. Multiple time series regression with integrated processes. *Review of Economic Studies* 53, 473–95.

Playfair, W. 1786. *The Commercial and Political Atlas*. London: T. Burton, 1801.

Porter, T.M. 1986. *The Rise of Statistical Thinking 1820–1900*. Princeton, NJ: Princeton University Press.

Quetelet, A. 1942. *A Treatise on Man and the Development of his Faculties*. Edinburgh: Chambers.

Redman, D.A. 1997. *The Rise of Political Economy as a Science*. Cambridge, MA: MIT Press.

Routh, G. 1975. *The Origins of Economic Ideas*. London: Macmillan.

RSS (Royal Statistical Society). 1934. *Annals of the Royal Statistical Society 1834–1934*. London: The Royal Statistical Society.

Schweber, L. 2006. *Disciplining Statistics: Demographic and Vital Statistics in France and England, 1830–1885*. Durham, NC: Duke University Press.

Secrist, H. 1930. *An Introduction to Statistical Methods*, 2nd edn. New York: Macmillan.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Slutsky, E. 1927. The summation of random causes as the source of cyclic processes. In Russian; English tr. in *Econometrica* 5, 1937.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R.H. Campell, A.S. Skinner and W.B. Todd. Oxford: Clarendon Press.

Spanos, A. 1986. *Statistical Foundations of Econometric Modelling*. Cambridge: Cambridge University Press.

Spanos, A. 1989a. On re-reading Haavelmo: a retrospective view of econometric modeling. *Econometric Theory* 5, 405–29.

Spanos, A. 1989b. Early empirical findings on the consumption function, stylized facts or fiction: a retrospective view. *Oxford Economic Papers* 41, 150–69.

Spanos, A. 1990. The simultaneous equations model revisited: statistical adequacy and identification. *Journal of Econometrics* 44, 87–108.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Spanos, A. 1995. On theory testing in econometrics: modeling with nonexperimental data. *Journal of Econometrics* 67, 189–226.

Spanos, A. 1999. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press.

Spanos, A. 2001. Time series and dynamic models. In *A Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Spanos, A. 2005. Misspecification, Robustness and the Reliability of Inference: the simple t-test in the presence of Markov dependence. Working Paper, Virginia Polytechnic Institute and State University.

Spanos, A. 2006a. Econometrics in retrospect and prospect. In *New Palgrave Handbook of Econometrics*, vol. 1, ed. T.C. Mills and K. Patterson. London: Macmillan.

Spanos, A. 2006b. Revisiting the omitted variables argument: substantive vs. statistical adequacy. *Journal of Economic Methodology* 13, 179–218.

Spanos, A. 2006c. Where do statistical models come from? Revisiting the problem of specification. In *Optimality: The Second Erich L. Lehmann Symposium*, ed. J. Rojo. Lecture Notes-Monograph Series, vol. 49. Beachwood, OH: Institute of Mathematical Statistics.

Spanos, A. 2006d. Revisiting the statistical foundations of panel data models. Working paper, Virginia Polytechnic Institute and State University.

Spanos, A. 2008. Curve-fitting, the reliability of inductive inference and the error-statistical approach. *Philosophy of Science*.

Spanos, A. and McGuirk, A. 2001. The model specification problem from a probabilistic reduction perspective. *Journal of the American Agricultural Association* 83, 1168–76.

Stigler, G.J. 1954. The early history of the empirical studies of consumer behavior. *Journal of Political Economy* 62, 95–113.

Stigler, S.M. 1986. *The History of Statistics: the Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stone, J.R.N. 1954. *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920–1938*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Stone, R. 1997. *Some British Empiricists in the Social Sciences 1650–1900*. Cambridge: Cambridge University Press.

Tinbergen, J. 1939. *Statistical Testing of Business Cycle Research*, 2 vols. Geneva: League of Nations.

Tippet, L.H.C. 1931. *The Methods of Statistics*. London: Williams & Norgate.

Uspensky, J.V. 1937. *Introduction to Mathematical Probability*. New York: McGraw-Hill.

Walker, H.M. 1929. *Studies in the History of Statistical Method*. Baltimore: Williams & Wilkins.

Walker, H.M. 1958. The contributions of Karl Pearson. *Journal of the American Statistical Association* 53, 11–22.

Wold, H.O. 1938. *A Study in the Analysis of Stationary Time Series*. Uppsala: Almquist & Wicksell.

Yule, G.U. 1897. On the theory of correlation. *Journal of the Royal Statistical Society* 60, 812–54.

Yule, G.U. 1911. *An Introduction to the Theory of Statistics*. London: Griffin & Co.

Yule, G.U. 1926. Why do we sometimes get nonsense correlations between time series? – a study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89, 1–64.

Yule, G.U. 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London*. Series A 226, 267–98.

How to cite this article

Spanos, Aris. "statistics and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000502>
doi:10.1057/9780230226203.1613

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

status and economics

Sheryl Ball
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Social status both affects and is affected by earnings and wealth. Research on status examines status-seeking behaviour, and the impact of acquired and endowed status. Status characteristics such as wealth and education can be acquired; others such as beauty, gender, or race are endowed. Status hierarchies appear to waste resources, as agents expend effort and income acquiring position, but such social structures may also benefit societies.

Keywords

animus; black–white labour market inequality in the United States; conspicuous consumption; incomplete information; positional externalities; reciprocity; social norms; statistical discrimination; status and economics; status seeking; trust; Veblen, T.; wage discrimination; women's work and wages

Article

The study of status by economists is not new, although some find it surprising since status initially seems to be a purely social phenomenon. A person's status is a ranking in a hierarchy that is socially recognized. People may participate in a number of status hierarchies defined by the different social groups of which they are members. Status may be defined and recognized narrowly, as when it is based on skills or accomplishments, or widely, when it is valued and recognized by an entire society. People within a society may differ in their assessment of the importance of a given status characteristic, so consistent rankings may be impossible. Research on status has focused on several questions. Why do individuals value status? What kinds of status-seeking behaviours do we observe? Is status-seeking a good thing?

Status-seeking behaviour

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Adam Smith (1759) noted that individuals are willing to expend effort to attain status, a practice that Veblen (1926), who coined the term ‘conspicuous consumption’, viewed as wasteful. Social psychologists study the subject, and have developed *status characteristics theory*. Status-seeking acts vary by culture, and include purchasing status goods or houses in the ‘right’ neighbourhood and seeking education that is not ‘useful’ (Veblen, 1926), as well as contributing money to charity (Getzner, 2000) and planning weddings that strain the family's financial resources (Bloch, Vijayendra and Desai, 2004). In these examples status-seeking is clearly ‘costly’ to the individual. Assessing the full impact of status also requires evaluating several possible indirect consequences of individuals’ status-seeking decisions. Bernheim (1994) models behaviour that is a costly signal about an individual's social desirability. Agents possess unobservable characteristics and tastes for associating with others with whom they are similar. Individuals seek the esteem of others, where esteem is based on public perceptions of their types. The result is the development of social norms whereby individuals choose similar actions in order to increase their popularity. In this case, the underlying characteristics may have cultural value, for example, whether an individual is ‘selfish’ or ‘generous’. Sub-populations within a society might value these characteristics differently, however, resulting in different sets of social norms. This means that this type of status characteristic can only be ranked within its cultural context.

In other cases, the status characteristic may have value for its own sake in addition to its status value. Here returns to status are not based solely on an absolute level of status, but also on one's relative status ranking. Thus status-seeking becomes a contest where individuals are willing to incur costs in order to win. Positional externalities are inherent in the status contest since actions that increase one individual's relative status decrease another's. The resulting positional arms races generally reduce social welfare and probably account for much of Veblen's distaste for status-seeking behaviour.

Earnings or wealth are straightforward examples of this type of positional externality. As every department chair knows, people care not only about their absolute earnings but also about relative earnings. Bolton (1991) models bargaining with agents who care about their absolute and relative earnings and finds that this model organizes some experimental data better than a pure self-interest model. In the laboratory, subjects behave as if they care about relative earnings. Concern about relative earnings may explain why risk-averse individuals may make decisions that, on the surface, seem more consistent with risk-seeking behaviour (Robson, 1992.)

Fershtman, Murphy and Weiss (1996) argue that status-seeking can affect economic growth. They model social status as a reward for education-seeking in economic growth-enhancing occupations. If individuals differ only by ability or income, then awarding status to growth-enhancing occupations will increase growth. However, if individuals differ by both wealth and learning ability, the ‘wrong’ agents may acquire education; inefficiency arises because low-talent/high-income workers invest in education to obtain higher status occupations. This reduces wages in the high-status occupations and crowds out high-talent/low-income workers. In addition to the inefficient distribution of talent in the economy, status-seeking then reduces economic growth.

Endowed status

While the status characteristics discussed above result from decisions an individual makes and may be productivity related, this is not necessarily the case. Other status characteristics, such as race, are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
inherited. In other cases, individuals may have some, but not complete, control, for example, for physical attractiveness. Nevertheless, both experimental and empirical work documents the effect of these types of status on economic decision-making.

Artificial status. Even artificial status may act like a type of power in economic interactions. Ball et al. (2001) report results from experiments designed to control for productivity-related attributes and to isolate the effect of status. Status is awarded by the experimenters in order to ensure that all subjects recognize the relevant status characteristics. Half the participants are publicly awarded gold stars based on the sum of their answers to a trivia quiz, and then trade in experimental markets where the traders on one side of the market have stars, and those on the other side do not. They find that earnings average 15 per cent higher for the high-status traders, and that the pattern of higher earnings persists even when status is randomly awarded. Thye (2000) finds a similar result in a bargaining game. Kumru and Vesterlund (2005) use the same procedure to induce status differences, in a sequential voluntary contributions game. They find that low-status agents tend to mimic the decisions of high-status agents, which creates an incentive for high-status agents to raise their own contributions. This suggests that status may play a role in increasing welfare in games where the socially optimal outcome is not the equilibrium in a pure self-interest model of behaviour.

Beauty. Economists' theory of incomplete information explains why costly status credentials that signal productivity may be linked to higher income. This occurs, for example, when job candidates with college degrees are preferred over those with no college education for jobs requiring no higher education-related skills. A systematic link between earnings and status characteristics that seem entirely unrelated to productivity, such as beauty, is more puzzling. For example, Hamermesh and Biddle's (1994) find a 15 per cent wage premium to more beautiful workers; this is consistent with status characteristics theory, but probably not with the marginal productivity theory of wages. Equally puzzling is Hamermesh and Parker's (2005) finding that more attractive faculty members earn higher teaching evaluations. In a trust-game experiment, Wilson and Eckel (2006) find that people are more likely to trust attractive people, although they are actually no more trustworthy than less attractive people. Other examples come from Solnick and Schweitzer (1999) on ultimatum games and Mulford et al. (1998) and Kahn, Hottes and Davis (1971) in Prisoner's Dilemma games.

In an experimental study of employment, Mobius and Rosenblatt (2006) are able to decompose the positive effect of beauty into three mechanisms. More attractive people are more confident, and confidence increases productivity. If one holds confidence fixed, however, more attractive people are wrongly considered more able by employers, a stereotype that may result in hiring mistakes. Beautiful workers also may have better communication and social skills, which may be beneficial when tasks require working with others. The result concerning confidence may explain the higher evaluations of attractive teachers.

Gender. In almost all societies men hold higher status than women. In the United States, for example, women earn less money than men, even after attempting to control for productivity related attributes such as education (Altonji and Blank, 1999; Darity and Mason, 1998.) Experimental tasks can allow researchers to look at environments where productivity does not affect outcomes in any way. Earnings differentials persist even in these environments. Women also may face discrimination in hiring. Goldin and Rouse (2000) collect hiring data for symphony orchestra musicians where in some cases musicians

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
were auditioned from behind a screen. They find that women were more likely to be hired or advanced to the next round of auditions in the 'blind' condition and less likely than men to be advanced in the 'non-blind' condition.

Differential treatment by gender also exists in experimental bargaining games. Solnick (2001) and Eckel and Grossman (2001) conducted ultimatum bargaining games where subjects knew their opponents' gender. Both studies find that offers made by men and women are about the same, while offers made to women are significantly lower than those made to men. Field experiments also find differences in bargaining by gender; women are initially offered higher prices in the market for cars (Ayers and Siegelman, 1995) and trading cards (List, 2004).

Race. Race is a form of status, and also follows the pattern of low status being related to low earnings. Even after controlling for productivity related attributes, African-Americans still earn less than whites (Altonji and Blank, 1999; Darity and Mason, 1998). Economic research on the motivations for race-related wage discrimination dates back to Becker (1957). Arrow (1972) argues that status in working environments is related to discrimination in yet another way. Conferring high status on some individuals can compensate for 'nearness' to individuals with whom an agent prefers not to associate. This hypothesis explains the result that highly educated African-Americans face a relatively larger wage gap. In this case education, another status characteristic, works against their earnings potential.

Any status characteristic needs to be observable in order to affect others' behaviour. Since African-Americans have less education than whites on average, Arrow argues that using education as a criterion in employment screening can be an effective means of discrimination. Visual cues may form another basis for discrimination. Historically, skin colour was used as a signal of social status – since people who worked outside tended to tan, so lighter skin was a signal of wealth and high status. Arrow (1972) describes skin colour as a 'cheap' signal about race. Darity and Mason (1998) survey a number of studies on race and skin colour in the United States and find that even within races, darker skin colour is related to lower earnings. This suggests that race-related status hierarchies are not bilateral. Eckel and Wilson (2004) show subjects pictures of their opponents prior to playing a trust game. A different sample of subjects from the same subject pool evaluated the pictures for a number of characteristics. They find that trust and reciprocity are significantly related to skin colour as well as perceived 'friendliness' and 'reliability' of the recipients. Field experiments also find differences in bargaining by race; minorities are initially offered higher prices in the market for cars (Ayers and Siegelman, 1995) and trading cards (List, 2004).

Names. Names are a signal of status in societies where there is segmentation based on race or ethnicity. Bertrand and Mullianathan (2003) find that employers in the United States are less likely to call the sender of a résumé identified with an African-American sounding name than they are the sender of an identical resume with a white-sounding name. Fryer and Levitt (2004) calculate an index of 'distinctively black'-sounding names, but once they control for other socio-economic variables, names are not associated with economic disadvantage. Fershtman and Gneezy (2001) conduct laboratory experiments in Israel that pair subjects with partners who have names that denote a distinctive Ashkenazic or Eastern Jewish heritage. They find that subjects with an Ashkenazic partner are three times more likely to make an efficient transfer in a trust game, although similar amounts are transferred in dictator games, regardless of the ethnic heritage of the recipient. This suggests that while people of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Ashkenazic heritage are of higher status, more highly educated, and, in the experiment, more trusted, they are not more worthy when it comes to income distributions in an experiment.

Discussion

Amidst all the evidence that status affects economic decision-making, the puzzle is: why? Discrimination research often focuses on two possible explanations, namely, animus and statistical discrimination. While animus seems a consistent explanation for preferring individuals endowed with high status, it is less compelling when considering status that can be obtained. Statistical discrimination is consistent with Bernheim's (1994) model of social norms; however, Goldin and Rouse's (2000) result that women are hired more frequently than men in a setting where gender is unknown, and less frequently when it is known, suggests that status may be an inaccurate proxy for a person's hidden attribute of interest.

Status may serve a role in helping to prevent coordination failures in games with multiple plausible outcomes. If all individuals use the strategy ‘defer to high-status individuals’ in a game such as the ‘battle of the sexes’, status can serve to largely eliminate dominated outcomes. Status may also provide a means of avoiding low payoff equilibrium outcomes in Prisoner's Dilemma type games. Kumru and Vesterlund's (2005) result that low-status individuals follow the behaviour of high-status leaders in public goods games suggests a positive role for status in guiding behaviour.

Evolutionary arguments provide the most plausible explanation for the existence of status-related behaviour. Tastes that increase the likelihood that individuals survive and produce offspring are most likely to persist in a population. Among social animals the highest-ranking individual is generally the largest with size being linked to sexual maturity, strength and health. These high-status individuals generally enjoy preferential access to food and mates and produce a disproportionate number of society's offspring. Status hierarchies can also enhance the transmission of useful information if the prestige of high-status, successful individuals attracts close observation by others (Henrich and Gil-White, 2001). To the extent that status characteristics are passed along to offspring, this strategy produces a stronger ‘next generation’ than one without status hierarchies. This argument suggests that current tastes may result from what was a successful survival strategy for our distant ancestors. On the other hand, in most parts of the world humans have access to food and medical care that reduce the importance of status preference as a survival strategy. Given the inefficiencies that status-seeking produces, therefore, these tastes may no longer be optimal.

See Also

- behavioural game theory
- experimental economics
- psychological games
- women's work and wages

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Altonji, J.G. and Blank, R.M. 1999. Race and gender in the labor market. In *Handbook of Labor Economics*, vol. 3C, ed. O. Ashenfelter and D. Card. Oxford: Elsevier.

Arrow, K. 1972. *Models of Job Discrimination. Racial Discrimination in Economic Life*. Lexington, MA: Lexington Books.

Ayers, I. and Siegelman, P. 1995. Race and gender discrimination in bargaining for a new car. *American Economic Review* 85, 304–21.

Ball, S., Eckel, C., Grossman, P. and Zame, W. 2001. Status in markets. *Quarterly Journal of Economics* 116, 161–88.

Becker, G. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bernheim, B.D. 1994. A theory of conformity. *Journal of Political Economy* 102, 841.

Bertrand, M. and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94, 991–1013.

Bloch, F., Rao, V. and Desai, S. 2004. Wedding celebrations as conspicuous consumption: signaling social status in rural India. *Journal of Human Resources* 39, 675–95.

Bolton, G.E. 1991. A comparative model of bargaining: theory and evidence. *American Economic Review* 81, 1096–136.

Darity, W.A., Jr. and Mason, P.L. 1998. Evidence on discrimination in employment: codes of color, codes of gender. *Journal of Economic Perspectives* 12(2), 63–90.

Eckel, C. and Grossman, P. 2001. Chivalry and solidarity in ultimatum games. *Economic Inquiry* 39, 171–88.

Eckel, C. and Wilson, R.K. 2004. Initiating trust: the conditional effects of skin color on trust among strangers. Working paper, Rice University.

Fershtman, C. and Gneezy, U. 2001. Trust and discrimination in a segmented society: an experimental approach. *Quarterly Journal of Economics* 116, 351–77.

Fershtman, C., Murphy, K.M. and Weiss, Y. 1996. Social status, education, and growth. *Journal of Political Economy* 104, 108–32.

Fershtman, C. and Weiss, Y. 1993. Social status, culture and economic performance. *Economic Journal*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
103, 946–59.

Fryer, R.G. and Levitt, S.D. 2004. The causes and consequences of distinctively black names. *Quarterly Journal of Economics* 119, 767–805.

Getzner, M. 2000. Hypothetical and real economic commitments, and social status, in valuing a species protection programme. *Journal of Environmental Planning and Management* 43, 541–59.

Goldin, C. and Rouse, C. 2000. Orchestrating impartiality: the impact of ‘blind’ auditions on female musicians. *American Economic Review* 90, 715–41.

Hamermesh, D.S and Biddle, J.E. 1994. Beauty and the labor market. *American Economic Review* 84, 1174–94.

Hamermesh, D.S. and Parker, A. 2005. Beauty in the classroom: instructors’ pulchritude and putative pedagogical productivity. *Economics of Education Review* 24, 369–76.

Henrich, J. and Gil-White, F. 2001. The evolution of prestige: freely conferred deference as a mechanism for enhancing cultural transmission. *Evolution and Human Behavior* 22, 165–96.

Kahn, A., Hottes, J. and Davis, W.L. 1971. Cooperation and optimal responding in the prisoner's dilemma game: effects of sex and physical attractiveness. *Journal of Personality and Social Psychology* 17, 267–79.

Kumru, C. and Vesterlund, L. 2005. The effects of status on voluntary contribution. Working paper, University of Pittsburgh.

List, J.A. 2004. The nature and extent of discrimination in the marketplace: evidence from the field. *Quarterly Journal of Economics* 119, 49–89.

Mobius, M.M. and Rosenblat, T.S. 2006. Why beauty matters. *American Economic Review* 96, 222–35.

Mulford, M., Orbell, J., Shatto, C. and Stockard, J. 1998. Physical attractiveness, opportunity and success in everyday exchange. *American Journal of Sociology* 103, 1565–93.

Robson, A.J. 1992. Status, the distribution of wealth, private and social attitudes to risk. *Econometrica* 60, 837–57.

Solnick, S. 2001. Gender differences in the ultimatum game. *Economic Inquiry* 39, 189–200.

Solnick, S.J. and Schweitzer, M.E. 1999. The influence of physical attractiveness and gender on

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

ultimatum game decisions. *Organizational Behavior and Human Decision Processes* 79, 199–215.

Smith, A. 1759. *The Theory of Moral Sentiments*, ed. D.D. Raphael and A.L. MacFie. Indianapolis, IN: Liberty Press/Liberty Classics, 1982.

Thye, S.R. 2000. A status value theory of power in exchange relations. *American Sociological Review* 65, 407–32.

Van Huyck, J. and Battalio, R. 2002. Prudence, justice, benevolence, and sex: evidence from similar bargaining games. *Journal of Economic Theory* 104, 227–46.

Veblen, T. 1926. *A Theory of the Leisure Class: An Economic Study of Institutions*. New York, NY: Vanguard Press.

Wilson, R.K. and Eckel, C.C. 2006. Judging a book by its cover: beauty and expectations in the trust game. *Political Research Quarterly* 59, 189–202.

How to cite this article

Ball, Sheryl. "status and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000522> doi:10.1057/9780230226203.1614

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Steindl, Josef (1912 – 1993)

K. Laski

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

firm size; Kalecki, M.; oligopoly; stagnation; Steindl, J.

Article

Steindl was born in Vienna on 14 April 1912. He studied economics in Vienna and received his Ph.D. working under Richard Strigl. He worked in the Austrian Institute for Economic Research (AIER) from 1935 to 1938, the year of his emigration to England. He was a lecturer at Balliol College, Oxford, from 1938 to 1941, and then a research worker at the Oxford Institute of Statistics. He worked there with Michal Kalecki, who left a lasting mark on his theoretical work. He returned to Austria in 1950. He was barred from teaching at the University of Vienna for ideological reasons and resumed his job at AIER, where he worked until his retirement in 1978. In 1970, however, the University of Vienna bestowed upon him a honorary professorship. He was visiting professor at Stanford University in 1974/5.
Steindl dealt with the economic problems of the size of firms (1945) and of the distribution of firms according to size (1965a). He explained the pattern of size distribution of firms by means of random processes (birth and death processes). Other fields of interest were education (1967) and technology (for example, 1980).
The research which may prove longest lasting is his work on the development and the present phase of capitalist economies. His main work (1952) deals with the tendency to stagnation of the mature capitalist economy. His point of departure was that oligopoly leads to increased profit margins and consequently to a fall in effective demand. The ensuing decline in the degree of capacity utilization causes *ceteris paribus* a lower level of investment and a decline in the rate of growth in mature capitalist economies. The slowing down of capital growth reduces further the utilization of capacity and leads to a cumulative process of declining growth. Steindl thus treats the utilization parameter differently from Kalecki, for whom it is a purely passive variable. Another difference consists in the explanation of the growth trend of the capitalist economy without having recourse to exogenous factors like innovations.
Maturity and Stagnation in American Capitalism was largely ignored during a period of high employment and intensive growth. Only when the old weaknesses of unemployment and stagnation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

reappeared did the book arouse wider interest and prove its lasting significance. The evolution of his ideas is shown in the introduction to (1976) and in the penetrating analysis both of present economic trends (for example, 1979; 1985a; 1985b) and of the present state of economics (1984).

Selected works

1945. *Small and Big Business: Economic Problems of the Size of Firms*. Oxford: Basil Blackwell.

1952. *Maturity and Stagnation in American Capitalism*. Oxford: Basil Blackwell. Republished, New York: Monthly Review Press, 1976.

1965a. *Random Processes and the Growth of Firms: A Study of the Pareto Law*. London: Griffin.

1965b. The role of manpower requirement in the educational planning experience of the Austrian EIP. In *Manpower Forecasting in Educational Planning*. Paris: OECD.

1979. Stagnation theory and stagnation policy. *Cambridge Journal of Economics* 3, 1–14.

1980. Technical progress and evolution. In *Research, Development and Technological Innovation*, ed. D. Sahal. Lexington, MA: Lexington Books.

1981. Ideas and concepts of long run growth. *Banca Nazionale del Lavoro Quarterly Review* 136 (March), 35–48.

1982. The role of household savings in the modern economy. *Banca Nazionale del Lavoro Quarterly Review* 140(March), 69–88.

1984. Reflections on the present state of economics. *Banca Nazionale del Lavoro Quarterly Review* 148 (March), 3–14.

1985a. Distribution and accumulation. *Political Economy: Studies in the Surplus Approach* 1(1).

1985b. Structural problems in the present crisis. *Banca Nazionale del Lavoro Quarterly Review* 154 (September), 223–32.

How to cite this article

Laski, K. "Steindl, Josef (1912–1993)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000254>

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

doi:10.1057/9780230226203.1615

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Steuart, Sir James (1713 – 1780)

Original article by Walter Eltis, revised and expanded by Andrew Skinner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An 18th-century Scottish political economist, Steuart's interest in economic management, contrasting with the economic liberalism of Adam Smith, reflected his long exile on the European continent and his awareness of backward conditions generally. Believing that markets do not always clear, Steuart advocated import protection, subsidies for exports and agriculture, and public works to combat unemployment. He thought the resultant high taxation would encourage people to work hard to obtain the necessities of life. Whereas Smith ignored him in *Wealth of Nations*, Marx gave him his due in *Capital*, and in the 20th century his monetary theory was recognized by some as anticipating that of Keynes.

Keywords

agricultural subsidies; balance of payments; deduction; export subsidies; Hamilton, A.; Hume, D.; induction; industrial policy; infant trade; Law, J.; Malthus's theory of population; Marx, K.H.; Mercier de la Rivière, P.-P.; money supply; primitive accumulation; protection; public debt; public works; Scottish Historical School; Smith, A.; specie flows; stages theory of history; Steuart, Sir J.; taxation

Article

Biographical

The Steuart family owned two estates, Goodtrees, which is near Edinburgh, and Coltness on the outskirts of Glasgow. Goodtrees was the seat of Sir James Steuart, the second baronet, Solicitor-General and a member of the Union Parliament. Sir James married Anne Dalrymple, the eldest daughter of the Lord President of the Court of Session, by whom he had five children of whom James was the only son. James was born on 10 October 1713, presumably at Goodtrees.
James attended the Parish School at North Berwick, proceeding in due course to Edinburgh University

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
where he studied, inter alia, constitutional and Scots law. Thereafter James made the expected progression and passed the Bar examinations in 1735 at the age of 22.

Steuart became the third baronet in 1717 on the death of his father but did not spend time either enjoying his new status or his standing as an advocate. Rather he embarked upon a Foreign Tour (1735–40). It was during this period that he lost his remarkable mother, an event which may have affected his future fate.

Steuart travelled with a fellow advocate, Carnegie of Boysack, and the pair initially went to Holland where they pursued further study. But in due course they travelled through France, settling for a period in Avignon. Avignon was at this time a Papal Territory and a haven for those Scots who had been ‘out’ in the Jacobite Rebellion of 1715. It was here that Steuart met the Duke of Ormond, a fervent supporter of the Cause who in turn directed Steuart's steps to Madrid where he met the Earl Marischall, another of the architects of the ill-fated ’15. It may have been the influence of these two men that directed Steuart's steps to Rome in the late 1730s. Steuart seems to have been captivated by the Old Pretender and his staff (there is very little mention of Prince Charles) and in a way which was to have a profound influence upon his future.

Steuart met Lord Elcho in Lyons, en route home, and it may be that he persuaded the future commander of the Prince's Life Guards to join the movement. In any event Steuart was active on behalf of the Jacobites after his return to Scotland in 1740 and it was because of this that he was sent to France as ambassador in 1745, following the success at Prestonpans. But after the battle of Culloden in April 1746 Steuart entered a long period of exile and to begin with maintained an active link with the Party. But the early 1750s saw a withdrawal from the Jacobite interest and Steuart, together with Lord Elcho, eventually settled in Angoulême where they lived in some style, with the support of Elcho's mother (Wemyss, 2003).

Steuart was bored, however, and it was probably significant that the exiled Parlement of Paris came to the locality in 1753. It was here that Steuart met Mercier de la Rivière, the latter-day Physiocrat so much admired by Adam Smith, with whom Steuart formed a long and lasting friendship. When the Parlement returned to Paris in 1754 Steuart followed where he was entertained by Mercier de Rivière and probably introduced to Montesquieu and Mirabeau.

The scientific opportunities were considerable, but in fact Steuart left Paris and France in 1755 to avoid compromising his position further in the event of hostilities with Britain. He left Paris in short before the dissemination of the *Tableau économique*. The first two books of the *Principles* were completed in the isolation of Tübingen (Germany) by August 1759. His work on the *Policy of Grain* and the *Dissertation on the German Coin* belong to this period.

The Steuart family left Tübingen in 1761 following Lord Barrington's successful attempt to have Steuart's son, also James, appointed as a coronet in the British Dragoons. They travelled west to Rotterdam and Antwerp before settling temporarily in the Spa. It was here that Steuart was arrested by the French authorities and subsequently imprisoned. The arrest was thought to be due to Steuart's close knowledge of the weakness of the French economy, although another gloss has been put upon the event by Paul Chamley. Chamley indicated that Steuart had been caught in the possession of plans for the invasion of Santo Domingo (Haiti): plans which had been prepared by Mercier de la Rivière, ‘who had a personal pecuniary interest in an English invasion of the island and may also have realised that it would do his friend Steuart no harm in the eyes of London if he were arrested by the French’ (Chamley, 1965, pp. 44–6; Skinner, in Steuart, 1998, vol. 1, pp. xlv–xlvi).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Steuart returned to England in 1763, the year of peace with France, under the mistaken belief that the British government had acted upon his behalf. Steuart did not in fact receive a pardon for past misdemeanours until 1771. But in the meantime he enjoyed the protection of Lord Barrington, sometime Secretary at War, whom he had first met on the Foreign Tour.

After the frenetic period when he brought the *Principles* (1767) to completion, he pursued, or rather continued to pursue, work of an academic nature as the *Works* amply confirm. He also found time to write a series of letters on the American War (Raynor and Skinner, 1994) which are chiefly interesting for his suspicion of military victory and his advocacy of free trade with the Colonies whatever the outcome. These letters were written between 1775 and 1778.

Steuart was apparently a good neighbour, actively interested in the economic affairs of the locality (Lanarkshire) and in the politics of the region. Steuart died on 26 November 1780. He was interested in the family vault at Cambusnethan (Lanarkshire) which is now sadly in ruins. Coltness has been demolished apart from some remnants of the original stable block.

Steuart married Lady Frances Wemyss (Lord Elcho's sister) on 25 October 1743 and their son, also James, was born the following year. Sir James Steuart-Denham had a distinguished military career. He served mainly in Ireland and on his death in 1839 was the Senior General in the British army, notable for his reform of cavalry tactics. He married Alicia Blacker of Carrick but there were no children. (The name ‘Denham’ was added in 1773 following the transfer of the estate of Westshields to the third baronet on the death of Archibald Denham) (Skinner, 2006, p. 73).

The Principles: methodology

It should be noted that one of the most important features of Sir James Steuart's career was his extensive knowledge of the Continent. The Foreign Tour (1735–40) and exile as a result of his association with the Jacobites meant that by the end of the Seven Years’ War Sir James had spent almost half of his life in Europe. In this time he mastered four languages (French, German, Spanish and Italian), a fact which may help to explain Joseph Schumpeter's judgement that ‘there is something un-English (which is not merely Scottish) about his views and his mode of presentation’ (1954, p. 176n).

In the course of his travels Steuart visited a remarkable number of places which included Antwerp, Avignon, Brussels, Cadiz, Frankfurt, Leyden, Liège, Madrid, Paris, Rome, Rotterdam, Tübingen, Utrecht, Venice and Verona. He seems, moreover, consistently to have pursued experiences which were out of the common way. For example, when he settled at Angoulême he took advantage of his situation to visit Lyons and the surrounding country. During his residence in Tubingen, he undertook a tour of the schools in the Duchy of Württemburg. Earlier he had spent no less than 15 months in Spain where he was much struck by the irrigation schemes in Valencia, Mercia and Granada, the mosque in Cordoba and the painful consequences of the famine in Andalusia in the spring of 1737. In fact very little seems to have been lost on him and it is remarkable how often specific impressions found their way into the main body of the *Principles*. In his major book Steuart noted the economic consequences of the Seven Years’ War in Germany, the state of agriculture in Picardy, the arrangement of the kitchen gardens round Padua, and the problem of depopulation in the cities of the Austrian Netherlands.

Steuart drew attention to the difficulties under which he laboured in the preface to the *Principles* precisely because he thought they would be of interest to the reader. He pointed out that the ‘composition’ was the ‘successive labour of many years spent in travelling’ (1966, p. 304) during which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
he had examined different countries ‘constantly, with an eye to my own subject’:

> I have attempted to draw information from every one with whom I have been acquainted: this however I found to be very difficult until I had attained to some previous knowledge of my subject. Such difficulties confirmed to me the justness of Lord Bacon's remark, that he who can draw information by forming proper questions, is already possessed of half the science. (1966, pp. 5–6)

Steuart wrote very much in the style of a man finding his way through a new field. This, added to the fact that nearly eight years separate the first and last books, presented obvious problems; problems of which Steuart was always conscious but which he viewed with very mixed feelings:

> Had I been master of my subject on setting out, the arrangement of the whole would have been rendered more concise; but had this been the case, I should never had been able to go through the painful deduction which forms the whole train of my reasoning and upon which … the conviction it carries along with it in a great measure depends. (1966, p. 7)

Steuart sought to establish a system of thought whose content met the requirements of Newtonian methodology. The leading feature of Steuart's method is objective empiricism. He was thus entirely in accord with his friend Hume (Skinner, 2005) but like Hume, he recognized that the mere collection of facts was not of itself sufficient. The first step on the route to knowledge is the collection and description of facts; the second, the statement of certain ‘principles’ reached through a process of induction. Steuart also recognized that the scientist can only advance by concerning ‘himself’ with *cause* and *consequence*, that is, by thinking deductively. He solved the problem of how to combine the two techniques by using induction to establish his basic hypotheses, or ‘principles’, and deduction for what Hasbach described as the ‘clarification of phenomena’ (1891).

Steuart was quite clear as to the techniques of reasoning to be employed. The rules were simple, if difficult to obey: observation, induction, deduction, verification. There remained the question of the technique to be followed in building up a body of knowledge and here Steuart's answer was equally clear. The scientist should begin with the simple (and thus apparently abstract) case and gradually take account of more and more complex (and thus ‘realistic’) cases. The first objective must be clarity and Steuart thus recognized that the attainment of the second, relevance, can only come about through the use of the abstraction in the early stages of study. He argued that in building up a body of knowledge:

> Every branch of it must, in setting out, be treated with simplicity and all combinations not absolutely necessary must be banished from the theory. (1966, p. 227)

But, since the object is relevance and since the ‘more extensive any theory be made, the more it will be useful’, it follows that as we proceed ‘combinations will crowd in and every one of these must be attended to’ (1966, p. 227). Steuart always employed this technique in dealing with a body of knowledge; that is, he gradually builds up his argument in a series of steps which progressively increase in complexity.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
At the same time, Steuart recognized that theoretical edifices constructed in this manner present the economist with particular difficulties which arise from the nature of the subject matter itself. In Steuart's view, the economist or social scientist can only show ‘how consequences *may* follow from one another; to foretell what *must* follow is exceedingly difficult if not impossible’ (1966, p. 365). While we can and must establish general principles these do not provide rules of behaviour which must always hold good. Steuart thus concluded (somewhat ironically, in the context of a critique of Hume's quantity theory) that:

> I think I have discovered that in this, as in every other part of political economy, there is hardly such a thing as a general rule to be laid down’ (1966, p. 339)

Given the need for a systematic statement of particular principles, established in accordance with the discipline of an appropriate methodology, there remained the problem of establishing a useful ‘method’ in respect of the *organization* of the discourse as a whole.

> The thing to be done is to fall upon a distinct method … by contriving a chain of ideas, which may be directed towards every part of the plan, and which at the same time, may be made to arise methodically from one another. (1966, p. 28)

Here again, Steuart followed Hume's lead.
The ‘plan’ is contained in the first two books and is based upon a theory of economic development. Steuart's dominant theme was to be change and growth, and it is this which gives his work cohesion.

The historical perspective

Steuart opened his account with ‘society in the cradle’ before going on to trace the origins of, and the process of transition between, the various stages of the progress of man.
In this context, Steuart made use of a theory of stages, now recognized as a piece of apparatus which was central to the work of the Scottish Historical School. He cites, for example, the Tartars and Indians as relatively primitive socio-economic types of organization (1966, p. 56) while concentrating primarily on the third and fourth stages – the stages of agriculture and commerce. In the former case, Steuart observed that those who lacked the means of subsistence could acquire it only through becoming dependent on those who owned it; in the latter, he noted that the situation was radically different in that all goods and services command a price. He concluded, in passages of quite striking clarity:

> I deduce the origin of the great subordination under the feudal government, from the necessary dependence of the lower classes for their subsistence. They consumed the produce of the land, as the price of their subordination, not as the reward of their industry in making it produce.

He continued:

> I deduce modern liberty from the independence of the same classes, by the introduction of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

industry, and circulation of an adequate equivalent for every service. (1966, pp. 208–9).

Steuart also observed that ‘an opulent, bold and spirited people, having the fund of the Prince's wealth in their own hands, have it also in their own power, when it becomes strongly their inclination, to shake off his authority’ (1966, p. 216).

The alteration in the distribution of power which was reflected in the changing balance between proprietor and merchant led Steuart to the conclusion that ‘industry must give wealth and wealth *will* give power’ (1966, p. 213). As an earnest of this position, he drew attention (significantly in his Notes on Hume's *History*) to the reduced position of the Crown at the end of the reign of Elizabeth: a revolution which appears ‘quite natural when we set before us the causes which occasioned it. Wealth must give power; and industry, in a country of luxury, will throw it into the hands of the commons’ (1966, p. 213n).

It was perhaps for this reason that Steuart's French translator, Senovert (1789), advised his readers that of the advantages to be gained from a reading of the *Principles*, ‘Le premier sera de convaincre, sans doute, que la révolution qui s'opère sous nos yeux était dans l'ordre des choses nécessaires’ (1966, p. 24n). Senovert, in short, was convinced of the inevitability of the Revolution and believed that the *Principles* confirmed the point.

Economic analysis

To economists he has always been Sir James Steuart, because that is how he appears on the title page of his 1767 book. This is subtitled, *An Essay on the Science of Domestic Policy in Free Nations, in which are particularly considered, Population, Agriculture, Trade Industry, Money, Coin, Interest, Circulation, Banks, Exchange, Public Credit and Taxes*. It offers a detailed, comprehensive and often original account of the application of economic argument to this enormous range of questions.

The population theory with which the book opens anticipates much that Malthus went on to say, and Marx even suggested in the first volume of *Capital* that, ‘admirers of Malthus do not even know that the first edition of the latter's work on population contains, except in the purely declamatory part, very little but excerpts from Steuart’ (Marx, 1867, p. 333).

His analysis of the balance of payments has also been much admired. He went considerably further than Hume by incorporating a detailed analysis of the capital account, and this led him to the conclusion (among several where he differs from Hume) that a country with a persistent capital account deficit will be unable to find an equilibrium price level at which specie flows cease.

Steuart's travels on the Continent during his 18 years of exile from 1745 to 1763 acquainted him with monetary developments in Paris and Amsterdam, and this enriched his theoretical and empirical chapters on money and banking. But it is his analysis of economic policy which has attracted much modern attention. The contrast between his analysis and Smith's in *The Wealth of Nations* published just nine years later is especially marked. Steuart's years of exile had given him a detailed knowledge of economic and financial policy on the Continent, and in particular in France, Germany and Holland, and he advocated a degree of state intervention into every aspect of economic life, which contrasted sharply with the principles that Smith enunciated. Skinner (1981) has suggested that it was precisely Steuart's long years of residence on the Continent that led him to evolve a ‘system’ which was so much more

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
dirigiste than that of his great Scottish contemporary.

In his book, Steuart offers extensive and detailed advice to an idealized statesman, who is assumed to possess unlimited knowledge and whose ‘inclinations are always to be virtuous and benevolent’ (1966, p. 333). Steuart believed that markets do not always clear, and this was especially the case with the labour market, where there was always liable to be an imbalance between ‘demand’ and the supply of ‘work’. Manufacturers, merchants and workers sought to ‘consolidate’ any high living standards they temporarily achieved into permanently higher incomes, and they often achieved this by restricting competition. Once prices and wages were consolidated at high levels, employment necessarily suffered as soon as foreign manufacturers began to produce more cheaply. With these assumptions about the behaviour of workers and entrepreneurs, and the impotence or non-existence of corrective market forces, there was an extensive range of policies through which state intervention could be expected to increase wealth, welfare and employment.

As soon as domestic production became overpriced, imports would undermine domestic employment and the creation of wealth, and Steuart therefore proposed that ‘a branch of trade should be cut off’ where the Statesman shall find,

> upon examining the whole chain of consequences, … the nation's wealth not at all increased, nor her trade encourages, in proportion to the damage at first incurred by the importation. (1966, p. 293)

In addition to protecting industry against imports, Steuart advocated export subsidies, because he saw the alternative to, for instance, subsidizing exports of fish by £250,000 so that what cost £1,000,000 could be sold overseas for £750,000, as the total loss of £750,000 of potential domestic output. Without the subsidy,

> those employed in the fishery will starve; … the fish taken will either remain upon hand, or be sold by the proprietors at a great loss; they will be undone, and the nation for the future will lose the acquisition of £750,000 a year. (1966, pp. 256–7)

Steuart was also concerned that as industry and population grew, the price of subsistence would rise as the population forced farming onto inferior land, where ‘the progress of agriculture demands an additional expence’. In order to ‘preserve the intrinsic value of goods at the same standard as formerly; [the Statesman] must assist agriculture with his purse, in order that exportation may not be discouraged’ (1966, p. 200).

As well as seeking to avert the influence of agricultural diminishing returns by subsidizing agriculture in order to keep export costs down, Steuart actually proposed the setting up of a ‘policy of grain’ in ‘the Common Markets of England’, where the government would buy up all the grain that farmers were prepared to produce at ‘the minimum price expedient for the farmers’, and sell all that could be marketed at ‘the maximum price expedient for the wage-earners’, and store any excess in state granaries. Steuart actually drafted this anticipation of the European Economic Community's agricultural policies of the 1970s and the 1980s in 1759 while he was in exile in Tübingen.

Steuart also anticipated post-Second World War industrial policies, for he argued that a Statesman should not hesitate to intervene directly in the finance and management of any new undertaking where

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
he saw economic potential, and should

> inquire into the capacity of those at the head of it; order their projects to be laid before him; and when he finds them reasonable and well planned, he ought to take unforeseen losses upon himself … the more care and expence he is at in setting the undertaking on foot, the more he has a right to direct the prosecution of it towards the general good. (1966, p. 391)

Steuart was a powerful advocate of public works to create employment whenever there was an excess supply of labour. The government should always finance the employment of ‘the deserving and the poor’, and they should be employed to extend a nation's social and economic infrastructure rather than for unproductive purposes:

> If a thousand pounds are bestowed upon making a firework, a number of people are thereby employed, and gain a temporary livelihood. If the same sum is bestowed for making a canal for watering the fields of a province, a like number of people may reap the same benefit, and hitherto accounts stand even; but the firework played off, what remains, but the smoke and stink of the powder? Whereas the consequence of the canal is a perpetual fertility to a formerly barren soil. (1767, vol. 1, p. 519)

All these interventionist policies needed to be financed, and Steuart actually welcomed the high taxation this would entail. He argued that taxes redistribute income and wealth and create employment, for they

> advance the public good, by drawing from the rich, a fund sufficient to employ both the *deserving* and the *poor* in the service of the state. (1767, vol. 1, pp. 512–13)

They also increase the power and prestige of the Statesman, for

> By taxes the Statesman is enriched, and by means of this wealth, he is enabled to keep his subjects in awe, and to preserve his dignity and consideration. (1966, p. 304)

Economists who believe in the efficacy of market forces have often been concerned that high taxation may have adverse supply side effects, but Steuart actually believed that taxation would often have *favourable* supply side effects. High taxes

> may discourage idleness; and idleness will not be totally rooted out, until people be forced, in one way or other, to give up superfluity and days of recreation … When the hands employed are not diligent, the best expedient is to raise the price of their subsistence by taxing it. (1966, pp. 691–5)

Steuart was aware that this analysis of the social and economical benefits from high taxation would not be popular with his contemporaries, and that ‘the politics of my closet is very different from those of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
century in which I live', but he comforted himself with the thought that 'reason is reason', and that in another century these startling opinions would be acknowledge as correct (1767, vol. 1, p. 514).
Steuart's industrial policies amount (as Eltis, 1986, has suggested) to the setting up of a corporate state with a social contract between producers who are protected against foreign competition and whose employment is guaranteed, and the state to whom they pay high taxes. Some of these are then returned to inefficient producers, while the rest furthers the state's social and political objectives.
In addition to welcoming high taxation as a tool for the finance of industrial policies, Steuart was an advocate of state banks which would issue paper money. By making money less scarce, he believed that they would reduce interest rates and so benefit industry and commerce. He argued that John Law's Mississippi Scheme could have been successful in France with only a few minor modifications in the manner it was set up and administered, and that this could have established the long-term rate of interest at two per cent in France.
The many kinds of government expenditure Steuart so strongly advocated could also be financed through borrowing, and here again Steuart was ahead of his time. He believed that in the limit, whatever a government could raise from taxation could be devoted to the payment of interest on public debt so that at a five per cent rate of interest, governments could borrow 20 times their tax revenues:

> If no check be put on the augmentation of public debts, if they be allowed constantly to accumulate, and if the spirit of a nation can patiently submit to the natural consequences of such a plan, it must end in this, that all property, that is income, will be swallowed up in taxes; and these will be transferred to the creditors.

But even in that state of affairs where all property income is paid as interest to those who have lent to the government does not represent the limit of the state's power to borrow. It can go on to tax the recipients of debt interest and so provide the wherewithal to finance still further borrowing, for these taxes 'may be mortgaged again to a new set of men, who will retain the denomination of creditors' (1767, vol. 2, pp. 633–4). Some may doubt that governments can at the same time continue to borrow, and defraud those from whom they borrowed in the past by taxing away their interest so that this provides the finance for still further borrowing. Won't there be a refusal to go on lending to such governments? No, opines Steuart, because

> The prospect of a second revolution of the same kind with the first would be very distant; and in matters of credit, which are constantly exposed to risk, such events being beyond the reach of calculation, are never taken into any man's account who has money to lend. (1966, p. 647)

Hence Steuart was perceptive enough to appreciate that sovereigns (and sovereign governments) can continually defraud their creditors, while new lenders will still queue up to be defrauded because the prospect of this will be so distant and problematical that it has a negligible influence on the immediate willingness to lend.
Steuart's book was well received at first, but Smith, who believed that economies would make full use of their labour and capital in the complete absence of government-inspired employment policies, and at the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
same time wholly distrusted the omniscience and benevolence of governments, greatly weakened Steuart's reputation as a serious economist by totally ignoring the existence of his book in *The Wealth of Nations*. Four years before its publication, he wrote that ‘Without once mentioning [Steuart's book], I flatter myself, that every false principle in it, will meet with a clear and distinct confutation in mine’ (1977, p. 164).

In the 19th century Marx gave Steuart his due, and there are 13 references to him in the first volume of *Capital*. Several 19th-century German economists have compared Steuart's historical and institutional approach to political economy favourably with Smith's deductive methodology, but most accolades to the richness and originality of Steuart's contribution only emerged after the Keynesian revolution.

His monetary and employment theory have been much praised, most comprehensively by Vickers (1959), though Hutchison (1978) and Schumpeter (1954) have also recognized his Keynesian anticipations. Steuart's monetary theory has much more in common with Keynes than the mere proposition that sufficient monetary expansion will reduce interest rates to two per cent. In Steuart's argument, money expenditure is not closely linked to the money supply, for idle balances will often be freely held, and the price level depends upon

> demand and competition … Let the specie of a country … be augmented or diminished, in ever so great a proportion, commodities will still rise and fall according to the principles of demand and competition … Let the quantity of coin be ever so much increased, it is the desire of spending it alone which will raise prices. (1966, pp. 344–5)

But Steuart's monetary and employment theory describe only one element of his thought which has anticipated modern developments. S.R. Sen, the distinguished Indian economic planner who published an important book on Steuart in 1957, commended him as ‘the first Economic Adviser to the Government of India’, praised his case for detailed intervention into every aspect of economic life and suggested that ‘it would not be any great exaggeration to say that A.P. Lerner's chapter on functional finance seems almost a paraphrase of Steuart’ (Sen, 1957, p. 122). Twenty years later, Akhtar (1979), of the New York Federal Reserve Bank restated Steuart in 30 equations, and compared his growth theory favourably with Smith's.

The classical counter-revolution of the 1980s has, of course, challenged the case for detailed state intervention which became so fashionable after the Keynesian revolution, and Steuart's dirigisme has been criticized by Anderson and Tollison (1984). It will be evident that there has been a more extensive response to the interventionist political economy of Sir James Steuart in the 20th century than there was in his own time.

Reception

Contemporary reaction was mixed. Hume is said to have been critical of the ‘form and style’ of the book while James Boswell considered the work to be ‘irregular and fanciful’ (Skinner, 1966, p. xlvi). Hugh Blair wrote to Hume that ‘Sir James’ Book is the most ponderous piece of lumber that I have ever looked into’ (NLSms. 23153). One contemporary review was cautious, noting that:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> We have no idea of a statesman having any connection with the affair, and we believe that the superiority which England has at present over all the world, in point of commerce, is owing to her excluding statesman from the executive part of all commercial concerns. (*Critical Review* 23 (1767), p. 412)

The *Monthly Review* (36, p. 464) went so far as to accuse Steuart of imbibing prejudices abroad ‘by no means consistent with the present state of England and the genius of Englishmen’. Steuart replied:

> Can it be supposed, that during an absence of near twenty years, I should in my studies have all the while been modelling my speculations of English notions? If, from this work I have any merit at all, it is by divesting myself of English notions, so far as to be able to expose in a fair light the sentiments and policy of foreign nations, relatively to their own situation. (1966, pp. 4–5. This passage occurs in the second edition of the *Principles*, published in the *Works*)

But if Steuart did not fare well among at least some of the key figures of the Enlightenment (and later!) the situation was rather different upon the Continent and elsewhere. During the 1780s the text was twice translated into German while there was a French version in 1789. Kobayashi (in Steuart, 1998) has suggested that Steuart's model of ‘primitive accumulation’ may help to explain the popularity of his work in contemporary Ireland and Germany. Keith Tribe (1988, p. 133) on the other hand, noted that ‘until the final decade of the eighteenth century Sir James Steuart's *Inquiry* was better known and more frequently cited than Smith's *Wealth of Nations*’.
But perhaps the most intriguing link is with North America. The pirated Dublin edition of 1770 was circulated widely in the Colonies and attracted the attention of Alexander Hamilton who was naturally concerned about the economic prospects of the infant republic. Hamilton rejected Smith's ‘fuzzy philosophy’ in favour of a policy of protection as a means of counterbalancing the competitive advantages of the British economy in the years following the Peace of Paris (1783). This perspective seems to have been widely shared, and is essentially a variant of Steuart's stage of ‘infant trade’.

Selected works

1767. *An Inquiry into the Principles of Political Oeconomy; being an Essay on the Science of Domestic Policy in Free Nations*, 2 vols. London.

1805. *Works, Political, Metaphysical and Chronological*, 6 vols. London.

The *Works* includes a revised edition of the 1767 edition of the Principles.

1966. *Principles*, 2 vols, ed. A.S. Skinner. Edinburgh: Oliver & Boyd for the Scottish Economic Society.

1998. *Principles*, 4 vols (variorum), ed. A.S. Skinner with N. Kobayashi and H. Mizuta. London: Pickering and Chatto.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Biography

Chamley (1963; 1965); Skinner (in Steuart, 1966; 1998). See also the article on Steuart in the *Oxford Dictionary of National Biography* (2004).

Bibliography

Akhtar, M.A. 1978. Steuart on growth. *Scottish Journal of Political Economy* 26, 57–74.

Akhtar, A. 1979. An analytical outline of Sir James Steuart's macroeconomic model. *Oxford Economic Papers* 31, 283–302.

Anderson, G.M. and Tollison, R.B. 1984. Sir James Steuart as the apotheosis of mercantilism and his relation to Adam Smith. *Southern Economic Journal* 51, 456–68.

Chamley, P. 1963. *Economie Politique et Philosophie chez Steuart et Hegel*. Paris: Librairie Dalloz.

Chamley, P. 1965. *Documents Relatifs à Sir James Steuart*. Paris: Librairie Dalloz.

Coltness Collections, The. 1842. Edinburgh: Maitland Club.

Davie, E.G. 1967. Anglophone and Anglophile. *Scottish Journal of Political Economy* 14, 291–304.

Eagly, R.V. 1961. Sir James Steuart and the aspiration effect. *Economica* 28, 53–61.

Eltis, W. 1986. Sir James Steuart's corporate state. In *Ideas in Economics*, ed. R.D.C. Black. London: Macmillan.

Eltis, W. 1987. Steuart, Sir James. In *The New Palgrave: A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Grossman, H. 1943. The evolutionist revolt against classical political economy. *Journal of Political Economy* 51, 506–22.

Hasbach, W. 1891. *Untersuchungen über Adam Smith*. Leipzig: Duncker & Humblot.

Hirschman, A.O. 1977. *The Passions and the Interests: Political Arguments for Capitalism before its Triumph*. Princeton: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hont, I. 1983. The rich country–poor country debate in Scottish political economy. In *Wealth and Virtue: The Shaping of Political Economy in the Scottish Enlightenment*, ed. I. Hont and M. Ignatieff. Cambridge: Cambridge University Press.

Hutchison, T. 1978. *On Revolutions and Progress in Human Economic Knowledge*. Cambridge: Cambridge University Press.

Hutchison, T. 1988. *Before Adam Smith*. Oxford: Basil Blackwell.

Johnston, E.A.G. 1937. *Predecessors of Adam Smith*. New York: Kelley, 1960.

Jones, P., ed. 1988. *Philosophy and Science in the Age of Enlightenment*. Edinburgh: John Donald.

King, J.E. 1988. *Economic Exiles*. London: Macmillan.

Low, J.M. 1952. An eighteenth century controversy in the theory of economic progress. *Manchester School of Economic and Social Studies* 20, 311–20.

Marx, K. 1867. *Capital*. Moscow: Progress Publishers for Lawrence & Wishart, 1974.

Meek, R.L. 1967. The rehabilitation of Sir James Steuart. In *Economics and Ideology and other Essays*. London: Chapman and Hall.

Perelman, M. 1983. Classical political economy and primitive accumulation. *History of Political Economy* 15, 451–94.

Raynor, D. and Skinner, A.S. 1994. Sir James Steuart: nine letters on the American Conflict, 1775–1778. *William and Mary Quarterly* 51, 775–6.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Sen, S.R. 1957. *The Economics of Sir James Steuart*. London: Bell.

Skinner, A.S. 1981. Sir James Steuart: author of a system. *Scottish Journal of Political Economy* 38, 20–42.

Skinner, A.S. 2005. David Hume and James Steuart. In *The Reception of David Hume in Europe*, ed. P. Jones. London: Thoemmes.

Skinner, A.S. 2006. Sir James Steuart, *Principles of Political Economy*. In *A History of Scottish Economic Thought*, ed. A and S. Dow. London: Routledge.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Tortajada, R. 1999. *The Economics of James Steuart*. London: Routledge.

Vickers, D. 1959. *Studies in the Theory of Money 1690–1776*. Philadelphia: Chilton.

Vickers, D. 1979. Sir James Steuart. *Journal of Economic Literature* 8, 1190–5.

Tribe, K.P. 1988. *Governing Economy: The Reformation of German Economic Discourse*. Cambridge: Cambridge University Press.

Yang, H.S. 1994. *The Political Economy of Trade and Growth: An Analytical Interpretation of Sir James Steuart's Inquiry*. Cheltenham: Edward Elgar.

Wemyss, A. 2003. *Elcho of the '45*, ed. J.S. Gibson. Edinburgh: Saltire Society.

How to cite this article

Eltis, Walter and Andrew Skinner. "Steuart, Sir James (1713–1780)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000259> doi:10.1057/9780230226203.1616

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Stewart, Dugald (1753 – 1828)

Nicholas Phillipson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

political economy; Smith, A.; Stewart, D.

Article

Stewart was the most important early commentator on Adam Smith's work. He was born in Edinburgh in 1753 and died there in 1828. He was the brilliant and well-connected son of an Edinburgh professor and was destined for an academic career from the earliest age. Educated at Edinburgh and Glasgow Universities, Stewart was taught by Adam Ferguson and Thomas Reid and became a close acquaintance of Adam Smith. He was appointed to the Edinburgh Chair of Moral Philosophy on Ferguson's retirement in 1785 and held it until 1810, when ill-health forced his retirement. A charismatic and influential teacher, his vast erudition and synthetic skill was shaped by an acute sensitivity to the ideological responsibilities of the pedagogue. He was a prolific writer whose contemporary reputation was built on the first volume of his *Elements of the Philosophy of the Human Mind* (vol. 1, 1792; vol. 2, 1815; vol. 3, 1826) and its companion text book *Outlines of Moral Philosophy* (1793). These works circulated widely in the universities of Britain, America and the continent in the early 19th century and did much to establish Scottish Common Sense Philosophy as the most influential vehicle of elite education in the age of the American, French and Industrial Revolutions. Stewart's collected works were published posthumously in 1854–60.

Stewart's *Account of the Life and Writings of Adam Smith LL.D* (1793) was frequently republished, often as an introduction to Smith's works. He discussed the *Wealth of Nations* in relation to the *Theory of Moral Sentiments* and both in relation to Smith's abortive plan for publishing a theory of jurisprudence. At Edinburgh he lectured on the principles of government and political economy in 1800–8 to an influential group of students who were to do much to form Whig and Tory opinion in the early 19th century. These lectures were intended for publication but the manuscript was accidentally destroyed and never rewritten. Their substance can, however, by inferred from a posthumous text which was compiled from his notes and published with his collected works.

Stewart was the first academic to detach the study of political economy from that of the theory of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

government and to treat each as a distinct branch of political science and it is in this methodological innovation rather than for any particular economic theory that his importance for political economy lies. His lectures were addressed to those ‘who study Political Economy with a view to the improvement of the theory of legislation’ (Stewart, vol. 9, p. 255). He defined political economy as the sum of ‘all those speculations which have for their object the happiness and improvement of Political Society’. This, not ‘the mistaken notions concerning Political Liberty which have been so widely disseminated in Europe by the writing of Mr Locke’, was the only proper foundation on which a true science of government could be raised (Stewart, vol. 8, pp. 10, 23). In general, Stewart's lectures offered an intelligent, critical presentation of the arguments of the *Wealth of Nations* illuminated by occasional information about Smith's last thoughts, by a sustained and sympathetic reappraisal of the Physiocrats and by a persistent preoccupation with perfectibility, progress, and the gradual improvement of the British Constitution.

Selected works

1854–60. *The Collected Works of Dugald Stewart*, 10 vols, ed. Sir W. Hamilton. Edinburgh: Thomas Constable & Co.

Bibliography

Collini, S., Winch, D. and Burrow, J. 1893. *That Noble Science of Politics: A Study in Nineteenth-century Intellectual History*. Cambridge: Cambridge University Press.

Phillipson, N.T. 1983. Dugald Stewart and the pursuit of virtue in Scottish university education; Dugald Stewart and Scottish moral philosophy in the Enlightenment. In *Universities, Society and the Future*, ed. N. Phillipson. Edinburgh: Edinburgh University Press.

How to cite this article

Phillipson, Nicholas. "Stewart, Dugald (1753–1828)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000260>
doi:10.1057/9780230226203.1617

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sticky wages and staggered wage setting

Christopher J. Erceg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Staggered wage contracts have become a widely utilized framework for modelling nominal wage stickiness and for assessing its macroeconomic consequences, including for the transmission of monetary shocks. This article provides a heuristic description of the staggered contracts model, motivates its salient features, and discusses its evolution.

Keywords

contract multiplier; Great Depression; inflation inertia; Keynes, J. M.; monetary transmission mechanism; output gap; staggered wage contracts; sticky wages and staggered wage setting; Thornton, H.; time-dependent' pricing; wage indexation; wage rigidity

Article

Nominal wages are regarded as 'sticky' if they fail to adjust to the level that would prevail in an equilibrium with costless wage adjustment and full information. Modern analysis of sticky wages, including their allocative effects and policy implications, owes an enormous debt to Keynes (1936). Keynes believed that nominal wages were likely to adjust less rapidly than prices in response to nominal shocks. He inferred that real wages would move counter-cyclically, so that a monetary contraction would push up the real wage and reduce employment. Keynes applied his theory to explain the collapse of employment during the Great Depression, and also argued that wage stickiness had major implications for the choice of a monetary regime. But while Keynes's work is justly celebrated, it is important to recognize that the belief that nominal wage rigidities play a central role in the monetary transmission channel was held by some pre-eminent classical economists of the early 19th century, such as Henry Thornton.

The staggered wage contracts model: background and development

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the wake of the Keynesian revolution, considerable interest developed in constructing quantitative models that incorporated the hallmarks of Keynes's framework, including sticky wages. A shortcoming of these early models was in their characterization of expectations as adaptive. By the mid-1970s, the first models incorporating sticky wages into a general equilibrium framework with rational expectations appeared in seminal work by Fisher (1977) and Gray (1976). These authors assumed that wages were set a fixed number of periods in advance, with the predetermined wage set so that the labour market was expected to clear at the ‘maturity’ of the contract. While a considerable innovation, the ‘Fischer–Gray’ contract formulation effectively constrained the real effects of monetary shocks to be no longer than the duration of the longest contract.

Taylor (1980) introduced staggered wage contracts as a mechanism for allowing monetary shocks to exert real effects lasting beyond the length of the longest contract, a feature he termed the ‘contract multiplier’. Taylor's wage-contracting model was meant to be consistent with several empirical observations about wage-setting: (*a*) wages are typically set in nominal terms and remain unchanged for sustained periods; (*b*) wage-setting tends to be asynchronous across different groups of workers; and (*c*) workers appear to take the wages set by other workers into account when adjusting their own wage, as well as aggregate demand. (See Taylor, 1999, for a comprehensive survey of staggered contracting models.)

Specifically, Taylor divided workers into *N* cohorts, and assumed that each cohort was constrained to adjust its contract wage at fixed intervals once every *N* periods. To illustrate the key features accounting for a contract multiplier, it is helpful to consider the special case in which wage contracts last two periods. In this case, Taylor's model can be expressed as three equations:

$$x_T = (1/2)^*(w_T + E_T w_{T+1}) + (1/2)^* g(y_T + E_T y_{T+1}) \quad (1)$$

$$w_T = (1/2)^*(x_T + x_{T-1}) \quad (2)$$

$$m_T = y_T + w_T \quad (3)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where variables are in logs, and E_T denotes the conditional expectation operator. The first equation determines the contract wage x_T, which is the fixed wage each member of the cohort currently readjusting its wage receives over the life of the contract. The contract wage is specified as depending on average economy-wide wages expected over the contract life, and on current and future output (with the parameter g determining the sensitivity). The second equation expresses the average economy-wide wage w_T as a simple average of the contract wages still in effect. Finally, the model includes a simple quantity theory relation between the exogenous money stock m_T and nominal demand (the price level is a constant markup over the average wage). If we substitute (2) into (1), it is evident that contract wages have both a forward- and a backward-looking component, with the latter playing a crucial role in allowing the model to generate persistence.
Solving Taylor's model for the contract wage yields:

$$x_T = h * x_{T-1} + (1-h) * m_T \quad h = (1-g^{\wedge}(1/2))^{\wedge}2/(1-g) \tag{4}$$

where the money supply is assumed to follow a random walk. If g is sufficiently small, the composite parameter h is close to unity, implying that a monetary shock has a small initial effect on the contract wage (and on average wages or prices), and thus exerts a large and persistent effect on output; by contrast, h approaches zero as g approaches unity, consistent with monetary shocks exerting large immediate effects on wages and prices, but little effect on output. Importantly, since g is assumed to be a free parameter, Taylor's model can rationalize an arbitrarily high degree of output persistence even if contracts last only two periods.
Taylor's staggered contracts model represented a major innovation in so far as it seemed to provide an empirically realistic model of monetary transmission within a rational expectations framework. However, while the staggered contracts framework became widely utilized for generating persistent responses to monetary shocks, the assumption that nominal wage stickiness was the primary source of monetary non-neutrality proved less durable. Thus, when Calvo (1983) developed an alternative staggered contracts framework that departed from Taylor's by specifying contract durations to be random, he assumed that prices rather than wages were sticky. This shift towards specifying sticky prices as the source of monetary non-neutrality was motivated by empirical evidence that appeared inconsistent with counter-cyclical real wages, and persisted until the late 1990s.

Some critiques, and the model's evolution

The staggered contracts model of Taylor and Calvo has evolved in response to several critiques. First, Fuhrer and Moore (1995) criticized its inability to account for inflation persistence or for the output costs of disinflation. In particular, while one-time changes in money could have persistent real effects,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
these authors showed that permanent changes in money growth had fleeting effects on inflation and output. Two alternative approaches have emerged in response to this critique. One route has effectively embedded additional persistence into the contracting structure by assuming that some agents have adaptive expectations (as in Roberts, 1997) or that workers not receiving a signal to change their wage (or price) follow a mechanical indexation scheme (Christiano, Eichenbaum and Evans, 2005). An alternative approach adopted by Ball (1995) and Erceg and Levin (2003) retains rational expectations, but assumes that agents learn gradually about shocks. Either approach may account for inflation inertia and prolonged output losses due to disinflation; however, in the latter inflation persistence is not intrinsic as in the indexation schemes, but instead depends on features that determine the speed of learning (including policymaker credibility and transparency).

Chari, Kehoe and McGratten (2000) challenged the ability of Taylor-style contracts to generate a contract multiplier in a version with explicit micro-foundations. These authors argued that the key parameter '*g*' in eq. (1) should not be regarded as a free parameter, since in their model it was determined by structural parameters characterizing tastes and technology; moreover, no reasonable calibration could account for a low enough value of '*g*' to deliver a sizeable contract multiplier. This challenge spawned an expansive literature showing how it was possible to account for a sizeable contract multiplier in a more richly specified micro-founded model. While Chari, Kehoe and McGratten included only price rigidities, Erceg (1997) and Huang and Liu (2002) demonstrated that a *combination* of wage and price rigidities could help generate a substantial contract multiplier in a micro-founded setting in which workers acted as monopolistic competitors in the labour market and set wages in a staggered fashion; an extensive literature (including these authors) has shown that persistence may also be enhanced through various real rigidities.

Finally, Caplin and Spulber (1987) introduced state-dependent pricing into the staggered contracts framework, allowing agents to choose when to adjust their contracts rather than constraining them to adjust at exogenous intervals ('time-dependent' pricing). Their paper suggested that this innovation had pivotal implications for the monetary transmission mechanism: money could be neutral under some conditions. However, in subsequent analysis in a state-dependent framework with maximizing agents, Dotsey, King and Wolman (1999) found that the dynamic responses to a monetary policy shock were qualitatively similar to those of the standard Taylor model with time-dependent contracts.

In light of these critiques, the recent literature has tended to incorporate both wage and price rigidities within a micro-founded staggered contracts framework that allows for some form of inflation inertia. Christiano, Eichenbaum and Evans (2005) showed that such a model provides a good quantitative characterization of the responses of macro variables to a monetary policy shock. Importantly, the nearly acyclical empirical response of the real wage to a monetary policy shock has also helped renew support for incorporating sticky wages (as well as sticky prices) into macro models; with sticky prices alone, the real wage would be strongly pro-cyclical. Finally, interest in sticky wages has also been buttressed by the welfare analysis of Erceg, Henderson and Levin (2000). These authors showed that a combination of wage and price rigidities may account for the policymaker's apparent trade-off between stabilizing inflation and the output gap, and has important normative implications for the design of policy rules.

A significant shortcoming of recent models is that their empirical support comes primarily from aggregate data. Moreover, current micro-founded sticky wage models appear deficient in their characterization of worker–firm attachments, and because they fail to take account of the sizeable costs

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of renegotiating labour contracts. Models that are developed to address such limitations may well have substantially different normative implications from current models, even if their dynamic properties remain similar.

See Also

- monetary transmission mechanism

Bibliography

Ball, L. 1995. Disinflation and imperfect credibility. *Journal of Monetary Economics* 35, 5–24.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–98.

Caplin, A. and Spulber, D. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102, 703–25.

Chari, V., Kehoe, P. and McGratten, E. 2000. Sticky price models of the business cycle: can the contract multiplier solve the persistence problem? *Econometrica* 69, 1151–79.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal wage rigidities and the dynamic effects of shocks to monetary policy. *Journal of Political Economy* 113, 1–45.

Dotsey, M., King, R. and Wolman, A. 1999. State-dependent pricing and the general equilibrium dynamics of money and output. *Quarterly Journal of Economics* 104, 655–90.

Erceg, C. 1997. Nominal rigidities and the propagation of monetary disturbances. International Finance Discussion Paper 590. Washington, DC: Board of Governors of the Federal Reserve System.

Erceg, C., Henderson, D. and Levin, A. 2000. Optimal monetary policy with staggered wage and price contracts. *Journal of Monetary Economics* 46, 281–313.

Erceg, C. and Levin, A. 2003. Imperfect credibility and inflation persistence. *Journal of Monetary Economics* 50, 915–44.

Fisher, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85, 191–205.

Fuhrer, J. and Moore, G. 1995. Inflation persistence. *Quarterly Journal of Economics* 110, 127–59.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gray, Joanna 1976. Wage indexation: a macroeconomic approach. *Journal of Monetary Economics* 2, 221–35.

Huang, Kevin and Liu, Z. 2002. Staggered price-setting, staggered wage-setting, and business cycle persistence. *Journal of Monetary Economics* 46, 405–33.

Humphrey, T. 2004. Classical deflation theory. *Federal Bank of Richmond Economic Quarterly* 90, 11–32.

Keynes, J. 1936. *The General Theory of Interest, Employment, and Money*. London: Macmillan.

Roberts, J. 1997. Is inflation sticky? *Journal of Monetary Economics* 39, 173–96.

Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88, 1–24.

Taylor, J. 1999. Staggered wage and price setting in macroeconomics. In *Handbook of Macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

How to cite this article

Erceg, Christopher J. "sticky wages and staggered wage setting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000467> doi:10.1057/9780230226203.1618

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Stigler, George Joseph (1911 – 1991)

David M. Levy and Sandra J. Peart
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

George Stigler combined technical competence, erudition, and wit. Stigler asked Frank Knight's question of how we know the goals of those we study. Early on, he offered a Knightian challenge to welfare economics, defending classical economic policy against the new orthodoxy. Later, he became an exponent of the orthodoxy. If goals are fixed, information is useful as a means to achieving them. Characterizing information as a commodity, Stigler described competitive equilibrium as a distribution of prices in a market. Policy advice is endogenous and we return to Adam Smith's analysis in which the philosopher is no better than those he studies.

Keywords

addiction; American Economic Association; Arrow, K.; auctions; Becker, G.; Chicago School; classical economics; Coase theorem; Coase, R.; collusion; cooperation; Dantzig, G.; dismal science; equality; Friedman, M.; History of Economics Society; industrial organization; Knight, F. H.; linear programming; Malthus, T. R.; Mandeville, B.; Mill, J. S.; Mont Pelerin Society; negative income tax; new welfare economics; oligopoly; prisoner's dilemma; regulation; regulatory capture; revealed preference; Rosen, S.; Samuelson, P.; science as consensus; science, economics of; shadow prices; Simons, H.; Smith, A.; Stigler, G.; time preference; Viner, J.; Wallis, W.; welfare economics

Article

George Stigler begins his autobiographical statement for the 1982 Nobel Prize in Economic Sciences (Stigler, 1983a) with these words:

> I was born in Renton, a suburb of Seattle, Washington, in 1911. I was the only child of Joseph and Elizabeth Stigler, who had separately migrated to the United States at the end of the 19th century, my father from Bavaria and my mother from what was then Austria-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> Hungary (and her mother was in fact Hungarian). I attended schools in Seattle through the University of Washington, from which I was graduated in 1931. I spent the next year at Northwestern University.
> My main graduate training was received at the University of Chicago from which I received the Ph.D. in 1938. The University of Chicago then had three economists – each remarkable in his own way – under whose influence I came. Frank H. Knight was a powerful, sceptical philosopher, at that time vigorously debating Austrian capital theory but gradually losing interest in the details of economic theory. Jacob Viner was the logical disciplinarian, and equally the omniscient student of the history of economics. Henry Simons was the passionate spokesman for a rational, decentralized organization of the economy. I was equally influenced by two fellow students, W. Allen Wallis and Milton Friedman.

His statement ends this way:

> I met my wife, Margaret L. Mack, at the University of Chicago. We were married in 1936. She died in 1970. I have three sons, Stephen (a statistician), David (a lawyer), and Joseph (a social worker). We are a close-knit family, and each summer we gather at a cottage on the Muskoka Lakes in Canada.

The Nobel Committee adds, ‘George J. Stigler died on December 1, 1991’.
What is not mentioned in the texts between these paragraphs are the honours. The presidency of the American Economic Association in 1964, of the History of Economics Society in 1977, and of the Mont Pelerin Society in 1977 along with the Nobel Prize in 1982 testify to a unique career. Specialist accounts focus on his contributions to the study of regulation (Peltzman, 1993), industrial organization (Demsetz, 1993) and the economics of science (Diamond, 2005). Biographical sketches by those who knew him well (Becker, 1993; Friedman, 1993; Wallis, 1993), as well as his 1988 *Memoirs*, attempt to link his life and work. Following his death, McCann and Perlman (1993) assessed Stigler's career; Longawa (1993) provided the bibliography. More than a decade later, we attempt to capture the permanent challenge of Stigler's work here.
Economists today, trained in a straightforward mathematical discipline, find it difficult to appreciate Stigler's combination of technical competence, scholarly erudition (Becker, 1993; Rosen, 1993; Rosenberg, 1993), and mordant wit (Friedland, 1993). The erudition is a serious problem. A student who has read neither Aristotle nor Bernard Mandeville may fail to appreciate the change of attitude between an article by Stigler that opened with a passage from Aristotle's *Ethics* (1943, p. 355) and his identification of revealed preference with the philosophy advanced in Mandeville's *Fable of the Bees* (1966, p. 68). Moreover, there is what Friedland calls ‘shyness’ and Friedman refers to as ‘sensitivity’. Stigler's 1988 autobiographical account spells out his contributions to the ongoing scientific discussion, but it hides virtuosity behind a veil of modesty.

Virtuosity and modesty

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
We begin with a contribution that is not discussed in Stigler's autobiography but which might have made the career of an ordinary scholar, what is known in the operations research literature as ‘Stigler's diet problem’. Given a scientific consensus on nutrients, what is the least-cost diet? We start with 9 nutrients and 77 foods. Recent accounts of Stigler's treatment remain appreciative:

> Stigler used trial and error, and mathematical insight and agility to solve his (9×77) set of inequalities. Based on cost and nutrient content, he was able to ‘weed’ the original 77 foods down to 15 as the eliminated foods were dominated by those in the list of 15. … Stigler's diet for 1939 data cost $39.93 per year. … Stigler's 1939 diet problem was the first ‘large-scale’ problem that was solved using the simplex method (Dantzig 1963 [pp. 551–67]). In 1947, nine clerks, using hand-operated desk calculators, pivoted away for 120 clerk-days and found the linear programming minimum cost of $39.69. Stigler knew what he was doing! (Garille and Gass, 2001, p. 2)

Authorities preface their judgement of the technical virtuosity of the performance with this judgement of the oddity of it all:

> Stigler's diet problem is a prime example of an OR model that faithfully describes the real-world situation but whose solution validity is close to zero. As Stigler (1945, p. 312) cautioned: ‘No one recommends these diets for anyone, let alone everyone.’ (Garille and Gass, 2001, p. 2)

What was the effort about, then? Stigler explains when he compares his solution to those of ‘competent dieticians’ which cost two or three times as much as his:

> The dieticians take account of the palatability of goods, variety of diet, prestige of various foods, and other cultural facets of consumption. … the particular judgments of the dieticians as to minimum palatability, variety, and prestige are at present highly personal and non-scientific, and should not be presented in the guise of being parts of a scientifically-determined budget. The second reason is that these cultural judgments, while they appear modest enough to government employees and even to college professors, can never be valid in such a general form. No one can now say with any certainty what the cultural requirements of a particular person may be … If the dieticians persist in presenting minimum diets, they should at least report separately the physical and cultural components of these diets. (1945, p. 314)

So, for Stigler, claims about the goals of individuals that had neither scientific nor philosophical weight were embedded in the dieticians’ solutions. At a minimum, he insisted, these claims should be made transparent.
Later editions of Stigler's textbook discuss his contribution to linear programming:

> This method of isolating products is intimately related to a method known as linear

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> programming, and the 'shadow prices' of that method are the implicit alternative cost of inputs. (1966, p. 119)

To which he adds the footnote, 'See almost any other book on economics' and a reference to an article by Paul Samuelson on Frank Knight's theorem on linear programming.

We do not learn that Stigler's data and solution were the test case for George Dantzig to prove the worth of his simplex. Instead, we see played out Stigler's distaste for the economist who overemphasizes his own originality, a view reiterated in the essays on J.S. Mill's originality. There, too, Stigler endorsed Mill's stance of impartiality between ideas he created and those he adopted from others (1955; 1982, pp. 96–7).

We turn next to the problem that closes Stigler's diet paper – how we, as scientists, know the goals of individuals outside some very narrow physical sense.

Knight's discipline

Stigler's writings with an autobiographical component invariably stress the importance of Frank H. Knight as exemplar. In a self-assessment that is viewed sceptically by McCann and Perlman (1993), Stigler wrote 'A more improbable Moses, if Knight would ever forgive the analogy, could not be designed' (1988, p. 17). He stressed the impact of Knight's personal integrity, his life as a scholar who renounced money and fame as impediments to truth seeking (1988, p. 18). Is it then something of a surprise that the opening chapter in *Memories of an Unregulated Economist* asks 'Are economists good people?' Here, what worries Stigler most seems to have been the possibility that scholarship might be compromised by monetary rewards, specifically in the case of consulting. He thinks this is not so in his own case, but he recognizes that scholarship might also be influenced by sympathy for the client's case (1988, p. 133). Indeed, the idea that a statistician might be tempted by sympathy for the client is a concern in statistical consulting (Vardeman and Morris, 2003). Such ethical issues are rarely considered in economics.

Stigler was Knight's discipline for a long time. In his 1943 *Economica* review of *Theory of Competitive Price* (the partially completed 1946 *Theory of Price*), Lachmann pronounced the book to be a coherent version of Knight's distribution theory in which classical productive inputs are replaced by anonymous factors and cost is opportunity forgone. This view has triumphed so completely, outside perhaps of neo-Ricardian theorizing, that the radicalism of the Knight–Stigler position at the time is now largely forgotten.

In 1943 Stigler put forward a Knightian challenge to new welfare economics, defending classical economic policy against the new orthodoxy. First, he offered the objection:

> Consider theft; our present policy toward this means of livelihood probably has adverse effects on the national income. Prevention of theft and punishment of thieves involves substantial expenditures for policemen, courts, jails, locks, insurance salesman, and the like. By compensating successful thieves for the amounts they would otherwise steal, we save these resources and hence secure a net gain. (If this policy leads to an undue increase in declarations of intent to steal, the retired successful thieves – who, after all, have

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> special talents in this direction – may be persuaded to assume the police functions.) (1943, p. 356)

Since it would ‘outrage our moral sensibilities to pay voluntary tribute to thieves’, something must be wrong. Stigler sketched the alternative:

> The familiar admonition not to argue over differences in tastes [*de gustibus non est disputandum*] leads not only to dull conversations but to bad sociology. It is one thing to recognize that we cannot *prove*, by the usual tests of adequacy of proof, the superiority of honesty over deceit or the desirability of a more equal income distribution. But it is quite another thing to conclude that therefore ends of good policy are beyond the realm of scientific discussion.
> For surely the primary requisite of a working social system is a consensus on ends. The individual members of the society must agree upon the major ends which that society is to seek. (1943, p. 357)

The 1943 paper is cited in the first edition of the full *Theory of Price* as defending a Knightian consensus view of the law (Knight 1947, p. 62):

> it is the fundamental tenet of those who believe in free discussion that matters of fact and logic can (eventually) be agreed upon by competent men of good will, that matters of taste cannot be … (1946, pp. 15–16)

Consensus in deep goals was as critical to policy as consensus about theory and fact was to science. The defence of new welfare economics was brief and devastating (Samuelson, 1943). Here is K.J. Arrow's judgment:

> Professor Stigler has made it a burden of reproach to the new welfare economics that it does not take into account the consensus on ends. It is not clear from his discussion whether he regards the agreed-on ends as being obvious from introspection or casual observation … or as requiring special inquiry; his comments seem rather to incline in the former direction, in which case he lays himself open to Professor Samuelson's request for immediate enlightenment on various economic issues. (1951, pp. 83–4)

We find much of the Knightian spirit also in the 1946•minimum wage law article. Stigler did not reprint this article in the *Citizen and the State*, where he described it mischievously as an example of how the misguided economist, lacking a market failure to correct, ‘lamented the intervention of the state’ (1975, p. x). The reader can predict the judgement about the topic but perhaps not the remedy. Given that minimum wage laws fail to ameliorate poverty, what should we do? The shared goal of equal treatment drives Stigler's results in a surprising direction:

> One principle is fundamental in the amelioration of poverty; those who are equally in need

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> should be helped equally. If this principle is to be achieved, there must be an objective criterion of need; equality can never be achieved when many cases are judged (by many people) 'on their merits' … It is the corollary of this position that assistance should not be based upon occupation. The poor farmer, the poor shopkeeper, and the poor miner are on an equal footing. (1946, pp. 364–5).

Stigler endorsed a negative income tax in this context:

> There is a great attractiveness in the proposal that we extend the personal income tax to the lowest income brackets with negative rates in these brackets. Such a scheme could achieve equality of treatment with what appears to be a (large) minimum of administrative machinery. (1946, p. 365)

Stigler's belief at the time that greater equality is a shared goal might explain his sharp reaction to 'suggestions' that he and Friedman tone down their egalitarianism in their joint study *Roofs and Ceilings* (Hammond and Hammond, 2006). The full history of 'natural experiments' in economics has yet to be written. When it is, the study of how the San Francisco housing market responded to the great earthquake (Friedman and Stigler, 1946) will take pride of place (Rockoff, 1991).

The work on regulation cited in the Nobel award falls between his Knightian and his Mandevillean periods. Neither electricity (Stigler and Friedland, 1962) nor securities regulation (Stigler, 1964) attained the articulated goals of regulation policy.

Mandeville's disciple

Stigler renounced the procedure of imputing goals from articulated speech when his papers on regulation were collected:

> It seems unfruitful, I am now persuaded, to conclude from the studies of the effects of various policies that those policies which did not achieve their announced goals, or had perverse effects (as with a minimum wage law), are simply mistakes of the society. A policy adopted and followed for long time, or followed by many difference states, could not usefully be described as a mistake: eventually its real effects would become known to interested groups. To say that such policies are mistaken is to say that one cannot explain them. (1976, p. x)

Ten years before, in the 1966 edition of *The Theory of Price*, Stigler tells the student about the 'penetrating' Bernard Mandeville, the philosopher of revealed preference. This is the passage he quotes:

> I don't call things Pleasures which Men say are best, but such as they seem to be most pleased with; … John never cuts any Pudding, but just enough that you can't say he took none; this little Bit, after much chomping and chewing you see goes down with him like chopp'd Hay; after that he falls upon the Beef with a voracious Appetite, and crams

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> himself up to his Throat. Is it not provoking to hear John cry every Day that Pudding is all his Delight, and that he don't value the Beef of a Farthing? (1966, p. 68)

Instead of supposing policy goals are revealed in discussion, he now follows Mandeville in imputing goals from consequences (Stigler, 1971). So, regulatory capture is not a ‘failure’ of regulation, but is instead the very *point* of regulation.

An intriguing variation on Mandeville is presented in Stigler and Becker (1977). The title, ‘De Gustibus Non Est Disputandum’, teases Knight and the early Stigler. Stigler and Becker posit that individuals are homogeneous with respect to unobservable ends of life revealed in the material world by means of physical goods and other inputs. Attracting much attention as the first of a sequence of papers on rational addiction, the paper perhaps ought to be seen as an attempt to recover common goals in choice. Following the supposition that goals are better revealed in mute choices than articulated speech, in the Tanner Lectures (Stigler, 1982), Stigler now defends the sort of productivity ethics once denounced by Knight.

Information as commodity

In Stigler's judgement (and that of Becker), the 1961 article ‘Economics of Information’ was the most important of the contributions listed in the citation for the Nobel award. The problem is deceptively straightforward: how long will the rational consumer search for a lower price? From the claim that it pays to search more for higher-priced commodities, Stigler obtains the result that, the higher the price of a commodity, the lower is its percentage deviation from the central tendency. The importance of characterizing a competitive equilibrium in terms of a statistical distribution of prices instead of a point mass from ‘one price in the market’ is hard to overstate. This approach explains Stigler's scepticism regarding models of oligopoly that involve price movements in one direction but not another (1947; 1978). Stigler's insistence on the veil of language also helps explain why he maintains that we recover prices at which transactions were made even if they differ from reported prices (Stigler and Kindahl, 1970).

The implications that flow from Stigler's notion that information is a commodity extend to his theory of oligopoly (Stigler, 1964). Here, Stigler takes joint profit maximization as the default and explains oligopoly on the basis of the probability of detecting collusion. Now that the economics literature has rediscovered Adam Smith's principle of sympathetic behaviour as an explanation of group formation and cooperative behaviour (Sally, 2001), Stigler's argument that cooperation is the default in small groups reveals great prescience. Perhaps most unsettling to economists is the additional implication regarding the endogeneity of economic advice. For, if information is a commodity and economists provide information to clients, then we, too, are inside the economic process. This endogeneity of economic advice raises causality questions that are central to ‘Do Economists Matter?’ (Stigler, 1976). The issue here is whether our opinions are the cause or the effect? Now this is an identification problem!

Science as consensus

As we read the record, the Knightian emphasis on consensus underscores Stigler's view of scientific

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

practice. For Stigler, as for his teachers, the past of economics was part of what has been called the extended present. Indeed, Rosen (1993) and Rosenberg (1993) explain, perhaps more clearly than did Stigler himself, how Stigler drew inspiration from the past. What distinguishes him from Knight is the view that we might justifiably exclude views that are sufficiently far from the centre:

> There is merit in excluding the lunatic from discourse. If a man tells me he is Napoleon, or that matter, Josephine, discussion would serve no purpose. … Occasionally the lone dissenter with the absurd view will prove to be right – a Galileo with a better scheme of the universe, a Babbage with a workable computer – but if we give each lunatic a full, meticulous hearing, we should be wasting vast time and effort. So long as we do not suppress the peaceful lunatic, we leave open the possibility that he may convince others that he is right.
> ••… The larger the group [with a common outlook], the more certain we can be that it is not *insane* in the sense of being divorced from apparent fact and plausible reasoning. (Stigler, 1975, pp. 3–4)

This view explains Stigler's reading of the Sraffian account of classical economics and the challenge offered by Samuel Hollander (Hollander, 1979; Stigler, 1990; Hollander, 1990). In the case of the Chicago reaction to Ronald Coase's ‘mistake’ in the discussion of externalities, the outcome differed. Stigler recounts how the response was to invite Coase to dinner to talk about the ‘mistake’ which, by the end of the evening, became the ‘Coase theorem’ (Stigler, 1988, pp. 75–8).

Stigler's confidence in the relationship between the size of the community asserting a factual claim and the probability that the claim is correct depends upon the independence of inquiry. Yet once we believe that common acceptance warrants belief, we violate the independence of acceptance upon which the probability claim rests. If we accept a result *because* it is widely accepted, can we really think the result is even more firmly established for the next ‘researcher’? This is, perhaps, the most significant weakness of Stigler's generation, the neglect of what in retrospect turn out to be game-theoretic issues. And here Stigler is particularly stubborn (Demsetz, 1993). One technical point at which game theory is missing occurs in Stigler's assertion (1966, pp. 94–5) that the institutional framework fails to affect the equilibrium in the case of auctions. Stigler's (1952) work on T.R. Malthus fails to appreciate Malthus's argument that William Godwin's proposal for communism would create a large-number prisoner's dilemma. Godwin himself came to understand Malthus's prisoner's dilemma argument (Godwin, 1801, p. 74).

If there is systematic error in one aspect of our understanding of the past, then our demand for coherence may well impose, as a general equilibrium condition, errors in all those aspects that are connected. So one error about the past leads to others. As evidence of how the Malthus error has cascaded, consider Stigler on the most famous characterization of economics: ‘[Malthus's] pessimism was the source for the characterization of economics as “the dismal science”’ (1988, p. 5). A great deal of scholarship has now demonstrated that this is precisely wrong (Persky, 1990; Levy, 2001; Peart and Levy, 2005).

Chicago School

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Stigler's Nobel lecture (1983b) is an important defence of the progressive nature of economics. To make the progressive case, one needs to know how the discipline changes over time. In an important sense, the Chicago School doctrine of motivational homogeneity (Stigler and Becker, 1977) is a return to the classical economic doctrine of analytical egalitarianism against contending schools of thought that had focused instead on racial, ethnic or class differences. These latter accounts prevailed in early 20th-century economic analysis, and presupposed unequal economic competence (Peart and Levy, 2005). In early neoclassical economics, inferiority was adduced from claims about positive time preference (Peart, 2000). Stigler never succumbed to this temptation (Stigler, 1941, pp. 212–19; Stigler and Becker, 1977). The Chicago defence of the egalitarian roots of classical economics in the face of racist claims by eugenicists and other ‘progressive’ ‘experts’ (Peart and Levy, 2005), suggests that, fundamentally, economics has hardly progressed in its foundational elements. We find the same defense of homogenous capacity in Adam Smith arguing against Plato, in J.S. Mill against Carlyle, and in Chicago. But perhaps this ought not to surprise economists. Whatever is important for policy will be contested. The answer to Samuelson's request, which Arrow echoed, then perhaps lies in considering whether ‘common ends’, policy goals, are recommended by those who believe themselves able to rule others or by those who believe themselves to be essentially the same as others. In his defence of equal capacity of economic agents, George Stigler will be remembered as second to none in the 20th century.

See Also

- Arrow, Kenneth Joseph
- Becker, Gary S.
- Chicago School
- egalitarianism
- Knight, Frank Hyneman
- Malthus, Thomas Robert
- Mandeville, Bernard
- Mill, John Stuart
- revealed preference theory
- Ricardo, David
- Smith, Adam
- Sraffa, Piero
- welfare economics

Selected works

1941. *Production and Distribution Theories: The Formative Period*. New York: Macmillan.

1943. The new welfare economics. *American Economic Review* 33, 355–9.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1945. The cost of subsistence. *Journal of Farm Economics* 27, 303–14.

1946. The economics of minimum wage legislation. *American Economic Review* 36, 358–65.

1946. *The Theory of Price*. New York: Macmillan.

1946. (With M. Friedman.) *Roofs or Ceilings? The Current Housing Problem*. Irvington-on-Hudson, NY: Foundation for Economic Education.

1947. The kinked oligopoly demand curve and rigid prices. *Journal of Political Economy* 55, 432–49.

1952. The Ricardian theory of value and distribution. *Journal of Political Economy* 60, 187–207.

1955. The nature and role of originality in scientific progress. *Economica* n.s. 22, 293–302.

1961. The economics of information. *Journal of Political Economy* 69, 213–25.

1962. (With C. Friedland.) What can regulators regulate? The case of electricity. *Journal of Law and Economics* 5, 1–16.

1964. A theory of oligopoly. *Journal of Political Economy* 72, 44–61.

1966. *The Theory of Price*. New York: Macmillan.

1970. (With J. Kindahl.) *The Behavior of Industrial Prices*. New York: NBER.

1971. Smith's travels on the ship of state. *History of Political Economy* 3, 265–77.

1975. *The Citizen and the State: Essays on Regulation*. Chicago: University of Chicago Press.

1976. Do economists matter?*Southern Economic Journal* 42, 347–54.

1977. (With G. Becker.) De gustibus non est disputandum. *American Economic Review* 67, 76–90.

1978. The literature of economics: the case of the kinked oligopoly demand curve. *Economic Inquiry* 16, 185–204.

1982. *The Economists as Preacher, and Other Essays*. Chicago: University of Chicago Press.

1983a. Autobiography. In *Les Prix Nobel. The Nobel Prizes 1982*, ed. W. Odelberg. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1982/stigler-bio.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
html, accessed 13 September 2006.

1983b. Nobel Lecture: The process and progress of economics. *Journal of Political Economy* 91, 529–45.

1988. *Memoirs of an Unregulated Economist*. New York: Basic Books.

1990. Ricardo or Hollander?*Oxford Economic Papers* n.s. 42, 765–8.

Bibliography

Arrow, K. 1951. *Social Choice and Individual Values*, 2nd edn. New York: John Wiley & Sons.

Becker, G. 1993. George Joseph Stigler: January 17, 1911–December 1, 1991. *Journal of Political Economy* 101, 761–7.

Dantzig, G. 1963. *Linear Programming and Extensions*. Princeton: Princeton University Press.

Demsetz, H., 1993. George J. Stigler: midcentury neoclassicalist with a passion to quantify. *Journal of Political Economy* 101, 793–808.

Diamond, A., Jr. 2005. Measurement, incentives and constraints in Stigler's economics of science. *European Journal of the History of Economic Thought* 12, 635–61.

Friedland, C. 1993. On Stigler and Stiglerisms. *Journal of Political Economy* 101, 780–3.

Friedman, M. 1993. George Stigler: a personal reminiscence. *Journal of Political Economy* 101, 768–73.

Garille, S. and Gass, S. 2001. Stigler's diet problem revisited. *Operations Research* 49, 1–13.

Godwin, W. 1801. *Thoughts Occasioned by the Perusal of Dr. Parr's Spital Sermon*. London: G. G. & J. Robinson.

Hammond, J. and Hammond, C. 2006. *Making Chicago Price Theory: Friedman–Stigler Correspondence, 1945–1958*. London: Routledge.

Hollander, S. 1979. *The Economics of David Ricardo*. Toronto: University of Toronto Press.

Hollander, S. 1990. A reply to Professor Stigler and Dr. Peach. *Oxford Economic Papers* n.s. 42, 769–71.

Knight, Frank H. 1947. *Freedom and Reform*. New York: Harper & Bros.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lachmann, L. 1943. *The Theory of Competitive Price*. *Economica* n.s. 10, 264–5.

Levy, D. 2001. *How the Dismal Science Got Its Name: Classical Economics and the Ur-Text of Racial Politics*. Ann Arbor: University of Michigan Press.

Longawa, V. 1993. George J. Stigler: A bibliography. *Journal of Political Economy* 101, 849–62.

McCann, C. Jr. and Perlman, M. 1993. On thinking about George Stigler. *Economic Journal* 103, 994–1014.

Peart, S. 2000. Irrationality and intertemporal choice in early neoclassical thought. *Canadian Journal of Economics* 33, 175–88.

Peart, S. and Levy, D. 2005. *The 'Vanity of the Philosopher': From Equality to Hierarchy in Post-Classical Economics*. Ann Arbor: University of Michigan Press.

Peltzman, S. 1993. George Stigler's contribution to the economic analysis of regulation. *Journal of Political Economy* 101, 818–32.

Persky, J. 1990. Retrospectives: a dismal romantic. *Journal of Economic Perspectives* 4(4), 165–72.

Rockoff, H. 1991. History and economics. *Social Science History* 15, 239–64.

Rosen, S., 1993. George J. Stigler and the industrial organization of economic thought. *Journal of Political Economy* 101, 809–17.

Rosenberg, N. 1993. George Stigler: Adam Smith's best friend. *Journal of Political Economy* 101, 833–48.

Sally, D. 2001. On sympathy and games. *Journal of Economic Behavior & Organization* 44, 1–30.

Samuelson, P. 1943. Further commentary on welfare economics. *American Economic Review* 33, 604–7.

Vardeman, S. and Morris, M. 2003. Statistics and ethics: some advice for young statisticians. *American Statistician* 57, 21–6.

Wallis, W. 1993. George J. Stigler: In Memoriam. *Journal of Political Economy* 101, 774–9.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Levy, David M. and Sandra J. Peart. "Stigler, George Joseph (1911–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000262> doi:10.1057/9780230226203.1619

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Stiglitz, Joseph E. (born 1943)

Karla Hoff
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Joseph E. Stiglitz, 2001 Nobel Laureate in Economics, helped to create the theory of markets with asymmetric information and was one of the founders of modern development economics. He played a leading role in an intellectual revolution that changed the characterization of a market economy. In the new paradigm, the price system only imperfectly solves the information problem of scarcity because of the many other information problems that arise in the economy: the selection over hidden characteristics, the provision of incentives for hidden behaviours and for innovation, and the coordination of choices over institutions.

Keywords

adverse selection; agency problems; Arrow, K. J.; Arrow–Debreu model; asymmetric information; bankruptcy; capital controls; convexity; coordination problems; corruption; development economics; East Asian financial crisis (1997); economic geography; economic rents; efficient markets hypothesis; endogenous growth; Freeman, H.; hidden characteristics; imperfect information; incentive contracts; incomplete contracts; incomplete markets; information economics; institutional economics; insurance; International Monetary Fund; law of the single price; market failure; market socialism; Mirrlees, J.; monitoring costs; moral hazard; multiple equilibria; neoclassical economics; noise; non-market institution; optimal taxation; partial equilibrium; perfect markets; pooling contracts; pooling equilibrium; principal and agent; privatization; quality; reputation; risk; Ross, S.; rule of law; search costs; selection problems; separating contracts; separating equilibrium; sharecropping; signalling; Stiglitz, J. E.; subjective probability; transition and institutions; transparency; uncertainty; underdevelopment traps; variance; Washington Consensus

Article

Joseph E. Stiglitz helped to create the economics of information, which analyses equilibrium in markets in which there are asymmetries of information among the market participants. For that work, he received the Nobel Prize in Economics in 2001, jointly with George A. Akerlof and A. Michael Spence. Stiglitz's work demonstrated the many and sometimes subtle ways in which markets can fail to lead to efficient outcomes. His work elucidated a broad set of phenomena that had largely been ignored before 1970 because they were outside the limits of the standard paradigm: incentive contracts, bankruptcy, quantity rationing, financial structure, equilibrium price distributions, innovation, and dysfunctional institutions. This work contributed to a paradigm shift in economics. In the new paradigm, the price system only imperfectly solves the information problem of scarcity because of the many other information problems that arise in the economy. Stiglitz has also proved central theorems in many fields: development economics, finance, trade theory, public economics, and industrial organization.
The broad plan from which much of Stiglitz's work originates had two central goals. The first was to show that many of the implications of the standard neoclassical model do not remain valid once the assumption of perfect information is dropped. His famous paper on adverse selection (Rothschild and Stiglitz, 1976) opens with these words:

> Economic theorists traditionally banish discussions of information to footnotes. Serious consideration of costs of communication, imperfect knowledge, and the like would, it is believed, complicate without informing.•…•[T]his comforting myth is false. Some of the most important conclusions of economic theory are not robust to considerations of imperfect information. (1976, p. 629)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The second goal was to provide a better theoretical understanding of the workings of the economic system as a whole. As Stiglitz (2007) explains,

> By the time I had finished my graduate studies, I had realized that the model of the economic system that was being taught – and that was at the center of policy analyses – was *not* a model of a modern capitalist economy. It was little more than a fancy version of a primitive agriculture exchange/production economy, slightly updated to include manufacturing – so long as there were diminishing returns. There was but a short distance between Ricardo and Walras, and between Walras and Samuelson.•…
> …Capital was nothing more than seed that was harvested but not consumed…
> …technology was stagnant (or at most exogenous).

A critical assumption of the standard neoclassical model is that there is a price for each quality of good in the market and for each action one would wish to contract for. Buyers have no problem ascertaining quality, and firms produce the quality that they have agreed to produce. Firms do not need to motivate their workers. Lenders do not worry about borrowers repaying. Owners do not worry about managers taking the right actions. Stiglitz in his lectures to students in the 1980s gave the example of a stylist cutting hair: in the standard model, there would be a price for each hair that he cut.

A real-life experiment that helped economists evaluate the standard neoclassical theory were the experiences with market socialism in Eastern Europe, in which government owned the firms but there was a manager of each firm whose job it was to maximize profits of that firm, facing market prices. Stiglitz (1994) argues that if the neoclassical model were an accurate characterization of a market economy, then market socialism would have been successful. Because of the importance of incentive problems and of non-price institutions (such as banks) within the economy, the inefficiencies that arose under market socialism were not accidents, but rather the inevitable consequence of (*a*) the limitations of the information contained in prices, and (*b*) the gap between the set of goods for which markets can practically exist and the much broader set of present and future goods and actions on which welfare depends.

Two ideas inform much of Stiglitz's work.

1. The ‘control/information’ system of market economies embraces far more than the price system of the neoclassical model. The exchange problem is intertwined with the process of selection over hidden characteristics, the provision of incentives for hidden behaviours and for innovation, and the coordination of choices over institutions.
2. Competitive equilibrium in economies with imperfect information and missing markets is not, in general, Pareto efficient. Market outcomes can be improved on by government intervention, e.g., taxes and subsidies. A simple illustration is that if the care that the insured take to avoid an accident is not observable to the insurance company, then commodities like fire extinguishers that decrease insured agents’ losses should be subsidized, while commodities like alcohol that increase their losses should be taxed (Greenwald and Stiglitz, 1986, p. 247).

Until Arrow's work on medical care (Arrow, 1963), the only reasons for a missing market that had been well explored were environmental externalities and the inability, or undesirability, to exclude from use (the problem of public goods). Stiglitz's contributions would help to radically broaden the understanding of the sources of externalities to include information externalities, group reputation effects, agglomeration effects, knowledge spillovers, and pecuniary externalities (see, for example, Greenwald and Stiglitz, 1986 on pecuniary externalities and Hoff, 2001 on coordination failures). In the process, Stiglitz's work would help to change the profession's understanding of capitalism, although the policy recommendations of economists did not change as much as Stiglitz had hoped.

Citations are an objective, if imperfect, measure of influence. Kim, Morse and Zingales (2006) compiled a list of the 146 articles published in economics journals from 1970 to 2002 that had received by June 2006 more than 500 citations from the ISI Web of Science/Social Science Citation Index. Six of Stiglitz's papers appear on this list (no other author has more). In descending order of number of citations, the papers are Stiglitz and Weiss (1981), Rothschild and Stiglitz (1970), Dixit and Stiglitz (1977), Shapiro and Stiglitz (1984), Rothschild and Stiglitz (1976), and Grossman and Stiglitz (1980). This article will place each of these papers in the context of Stiglitz's research programme.

Biographical data

Stiglitz's early experiences shaped his lifelong professional interests in understanding how an economy handles risk, and in bringing economic theory to bear on real-world problems. Stiglitz was born in Gary, Indiana, a city marred, in his words, by ‘huge inequality, poverty, and discrimination’ (Stiglitz, 2007). He was the middle of three children. After the failure of an earlier business, his father became an independent insurance agent. One part of his job was to find new insurers for firms whose businesses had burned down and whose insurance policies had been cancelled. Stiglitz's mother worked in the family insurance business when Stiglitz was young and later taught elementary school in a low-income inner city neighbourhood of Gary. After retiring from elementary education, she worked in adult remedial education, where she encountered some of the same students whom she had taught as children in inner-city schools.

Stiglitz's genius was recognized early. In high school, he was assigned independent study in lieu of some of the regular classes, which he had outstripped. (His father apparently took

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
this the wrong way: he was concerned that something might be wrong with the other children in the school.) Stiglitz followed his older brother to Amherst College and graduated in 1964. There he studied economics, physics, history and mathematics, and was president of his senior class. In that position, he was a maverick: he tried unsuccessfully to stop the college from funding the training of Amherst sports teams in Bermuda during college vacations, and organized an exchange programme between Amherst and a segregated college in the US South.

He obtained a Ph.D. from the Massachusetts Institute of Technology (MIT) in 1967. Stiglitz (2007) writes that particularly important influences at MIT on his later work were his statistics teacher, Harold Freeman, who taught the recently developed theory of subjective probability, and Ken Arrow, with whom he took a class as a second-year graduate student. At that time, Arrow was writing the final formalization of the Arrow–Debreu paradigm. Realizing the model's limitations, Arrow was also beginning a research agenda into the consequences of imperfect information.

Stiglitz joined the faculty of MIT in 1966, spent time between 1969 and 1971 at the Institute for Development Studies at the University of Nairobi under a Rockefeller Foundation grant, and then moved from university to university: Yale (1967–74), St Catherine's College, Oxford (Visiting Fellow, 1973–74), Stanford (1974–76), All Souls' College, Oxford (1976–79), the Institute of Advanced Studies at Princeton (1978–79), Princeton University (1979–88), Stanford (1988–2000), and Columbia University (2000 to date).

Stiglitz received the John Bates Clark Medal in 1979, awarded biennially by the American Economic Association for the most distinguished work by an economist under the age of 40. In 1987, Stiglitz became founding editor of the *Journal of Economic Perspectives*.

In 1993 Stiglitz joined President Clinton's Council of Economic Advisors, which he chaired in 1995–97. In 1997, Stiglitz was appointed to the position of Chief Economist of the World Bank. The East Asian financial crisis occurred in 1997–99, and Stiglitz argued publicly against the policies of the International Monetary Fund (IMF) towards the crisis. Disagreements arose both about the consequences of the policies (given the uncertainties about both the structure of the economy and future events), and about welfare judgments of the acceptable trade-offs between competing goals. Stiglitz's positions led to conflict with other officials in Washington. In November 1999, he resigned from the World Bank and returned to academia. At Columbia, he co-founded the Initiative for Policy Dialogue, which studies policy issues and provides training to policymakers from developing countries.

The economics of uncertainty

The economics of uncertainty is concerned with the principles that an individual uses in evaluating a random distribution of returns. Applications extend from how individuals allocate their portfolios between safe assets and risky assets, to how farmers allocate their land among different crops. The work in the 1960s by James Tobin and others equated an increase in risk with an increase in variance. Rothschild and Stiglitz (1970) set forth an alternative definition of an increase in risk. Comparing income distributions with the same mean, they proposed a definition that corresponded to a preference ordering among every expected-utility-maximizer with a concave utility function. This ordering was not the same as a ranking based on increases in variance. In a companion paper, Rothschild and Stiglitz (1971) demonstrated the usefulness of their definition in deriving comparative static results. They showed that such results depended on a simple criterion: the concavity or convexity of a 'first-order condition' characterizing the individual's optimal decision with respect to the random variable that was the source of risk. That work unified work in an area that until then had been in great confusion.

The economics of information

In the 1960s, James Mirrlees (1971) began working on the problem of how a government could design an optimal tax schedule, taking into account that government can observe individuals' incomes, but not their ability and effort. Given this asymmetry of information, the analytical problem is to distribute a given tax burden according to differences in ability to pay. Stiglitz recognized the similarities between this problem and the problems that arise in *markets* with asymmetric information. For example, insurance companies and banks want to design a menu of offers that will maximize their profits, taking into account that they do not know each individual's risk of accident or bankruptcy and the care that an individual expends to avoid the insured for event. Employers want to design labour contracts to maximize productivity, taking into account that they can observe only imperfectly workers' ability and effort. Together with a small group of pioneers in the 1970s, and influenced in particular by Akerlof (1970) and Spence (1974), Stiglitz devised models in which these kinds of problems could be analysed. The work on hidden characteristics (adverse selection) and incentive problems (moral hazard) came to be the core of the economics of information.

Hidden characteristics

A selection problem arises whenever there is imperfect information about the characteristics of the items being transacted, and different sides of the transaction know different things (so information is *asymmetric*). This problem is pervasive. Rothschild and Stiglitz (1976) constructed a celebrated model of the insurance market in which individuals differ only in terms of their privately known risk type, and the insurance firms know the overall distribution of risks in the population. The model uses the canonical textbook apparatus of consumer choice – budget lines and indifference curves – but produces very surprising (and counter-Walrasian) results.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To illustrate Stiglitz's *modus operandi* – to begin with highly simplified models of particular markets that allow him to identify a general principle – I describe the Rothschild and Stiglitz (1976) model in detail.

Consider an individual whose situation is described by his income if he is lucky enough to avoid an accident (W_{NA}) and his income if an accident occurs (W_A). His initial endowment point E is (W, W–d), where d represents the damages incurred in the accident. An individual purchases insurance in order to alter his pattern of income across these two states of nature.

Begin with the benchmark case in which insurance companies know an individual's probability of accident. Given this probability, let $\bar{W}$ denote his expected income. Then competitive equilibrium would be at a point A, illustrated in Figure 1, where the insurance company breaks even and the individual's indifference curve, denoted $\bar{U}^L$, is tangent to the budget line. Since risk-averse individuals who are offered a break-even price for insurance will choose full insurance, the equilibrium is along the 45 degree line. The line that goes through the endowment point E and the point A is the locus of contracts at which an insurance company breaks even (‘the fair-odds line’).

Figure 1

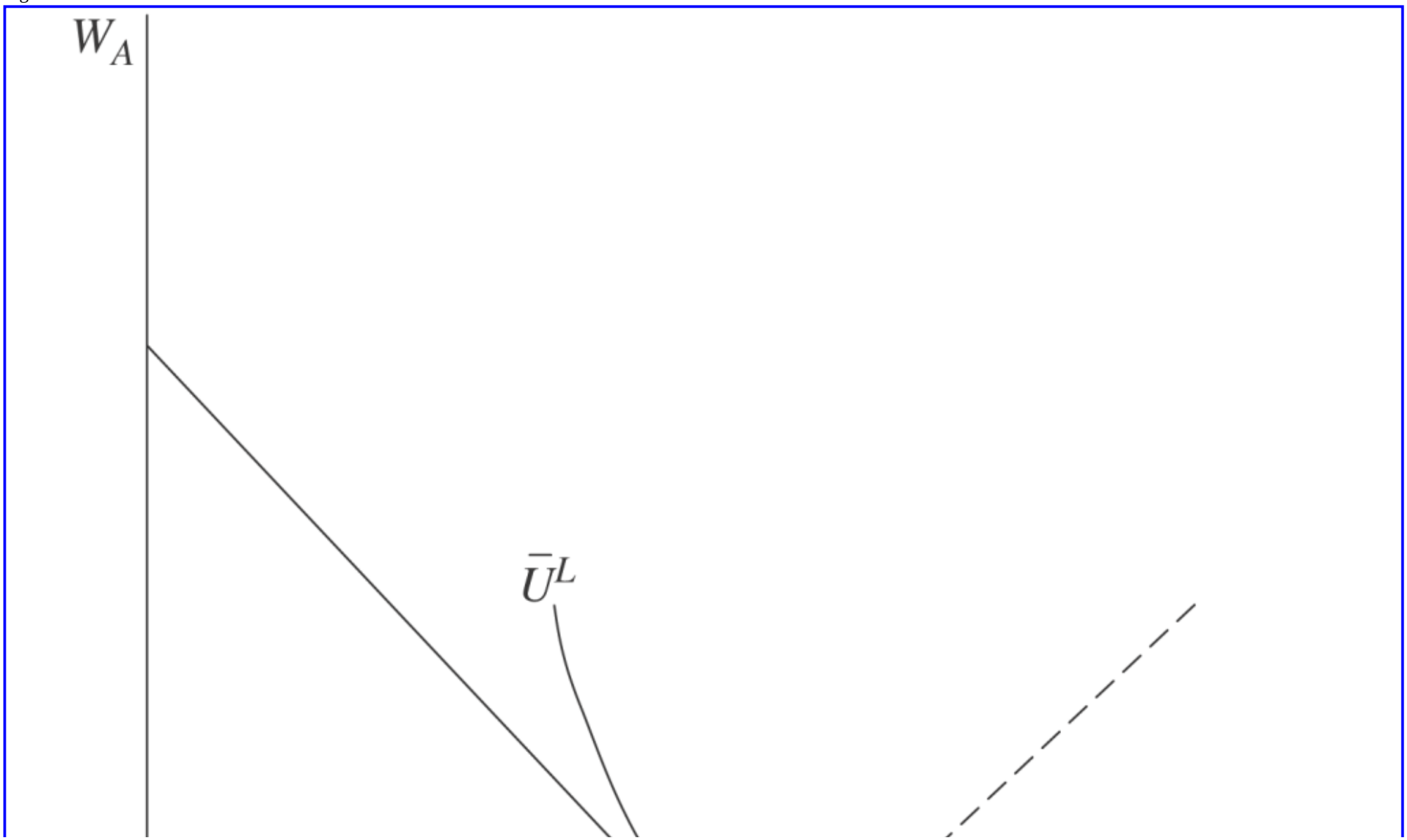

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

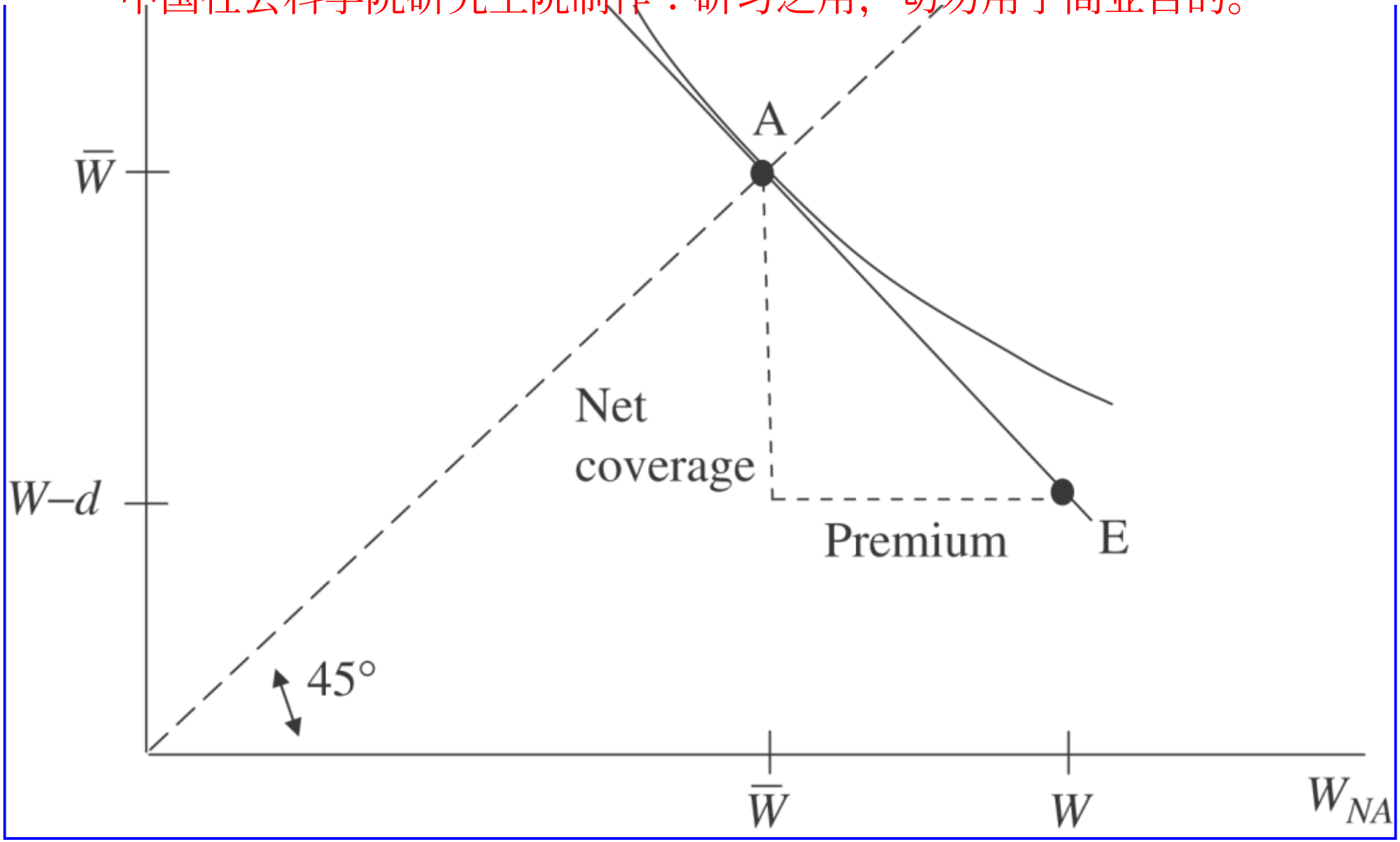

Next, bring in asymmetric information. Suppose that individuals are of two types that differ in their accident probability, and each individual knows his own type. Given the differences in risks, the high-risk individual has lower expected wealth (denoted $\underline{W}$ in Figure 2) than the low-risk individual (denoted $\bar{W}$). If the insurance company could observe who was low-risk and who was high-risk, then, on the same reasoning as above, the equilibrium contracts would be at A and B in Figure 2. A higher accident probability gives rise to a flatter indifference curve (the two types' indifference curves satisfy the 'single crossing property'), and also to more costly insurance (a flatter fair-odds line). If, however, the insurance firms do not know the characteristics of individuals, then clearly they cannot offer contracts A and B. For in that case all individuals would claim they were the low-risk type and choose the contract A, and the insurance companies would not break even. Offers that survive the competitive process cannot specify a price at which customers choose to buy all the insurance they want, because the high-risk individuals would always purchase more insurance at that price than the low-risk individuals, and the insurance firms would not break even. Competitive offers of contracts instead consist of both a price and a quantity.

Figure 2

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

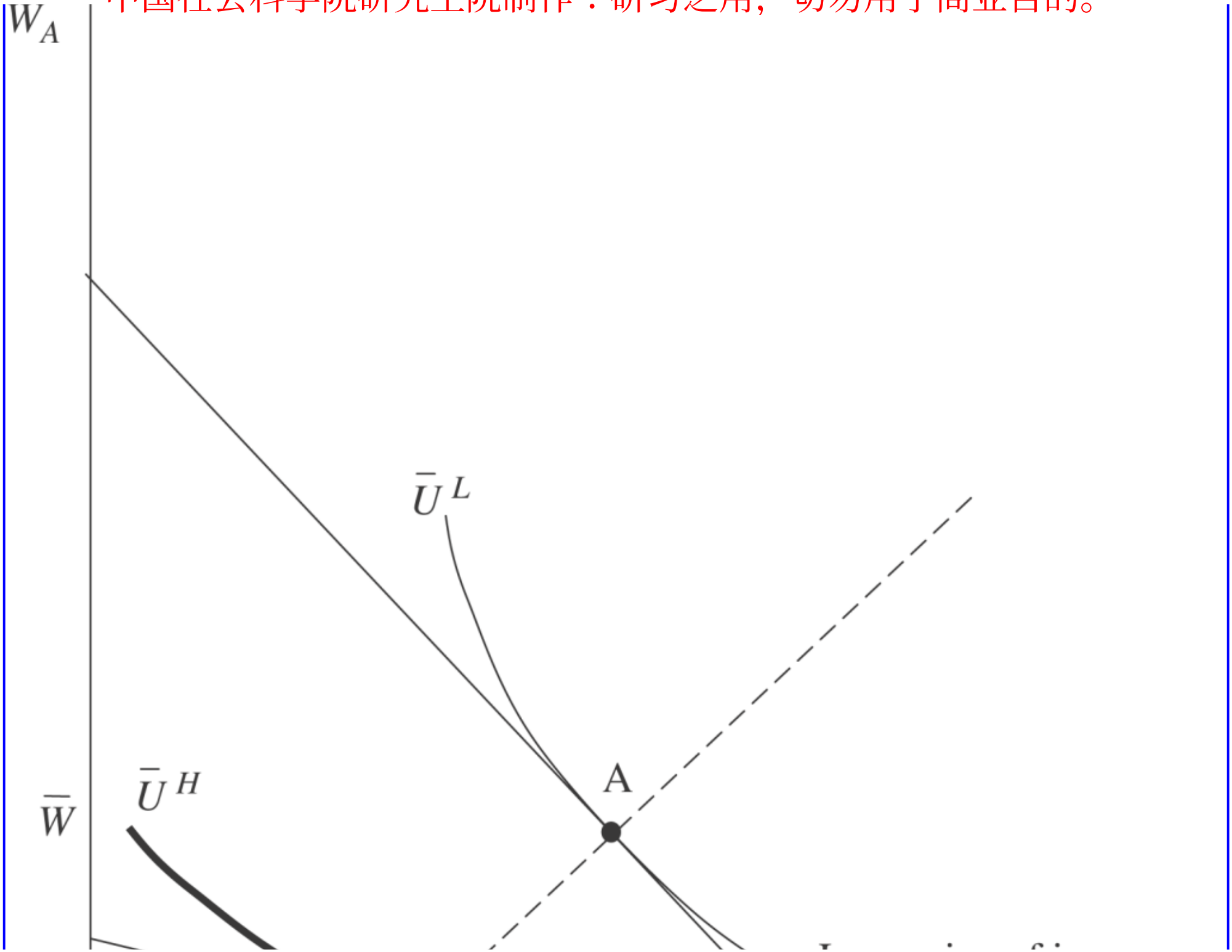

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

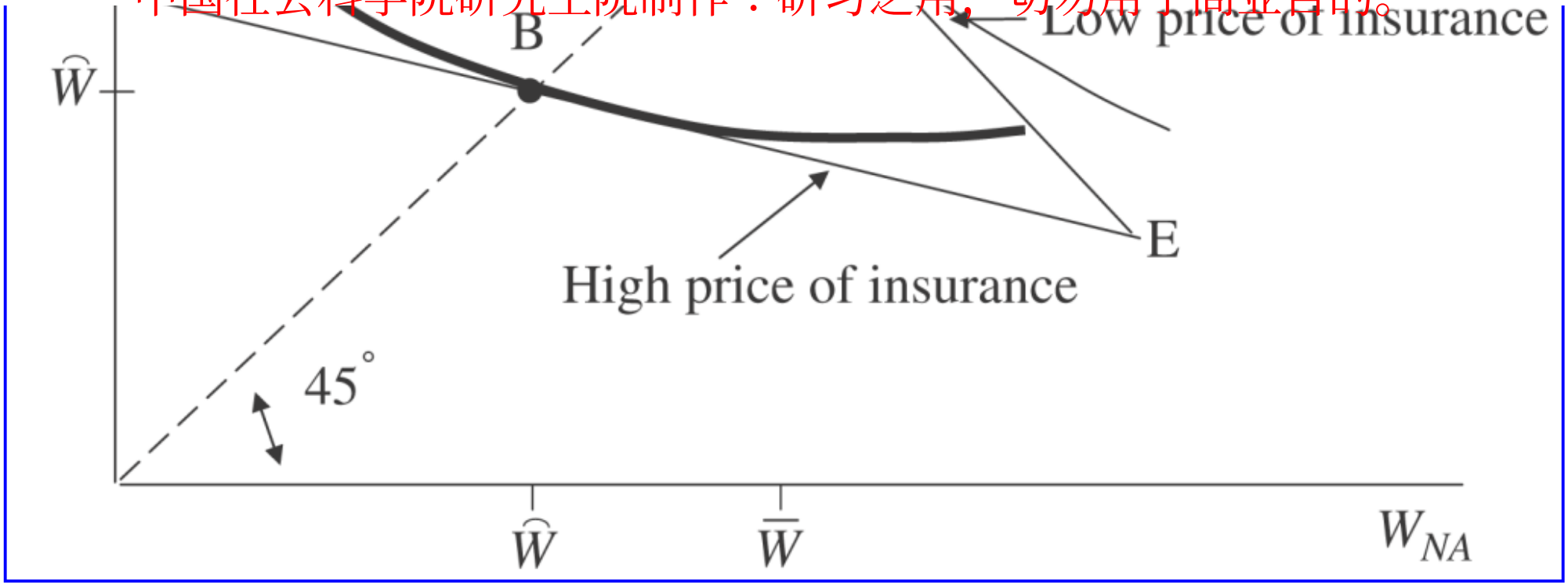

Consider then the price and quantity offer at point C in Figure 3, which would break even if all individuals purchased it (a ‘pooling’ contract). Rothschild and Stiglitz demonstrate that this *cannot* be equilibrium. Any contract such as C′ in the shaded area in Figure 3 – with slightly *less* insurance coverage than C but at a *lower* price per dollar of coverage – would attract only the low-risk individuals. Given that, the contract C generates losses and would be withdrawn.

Figure 3

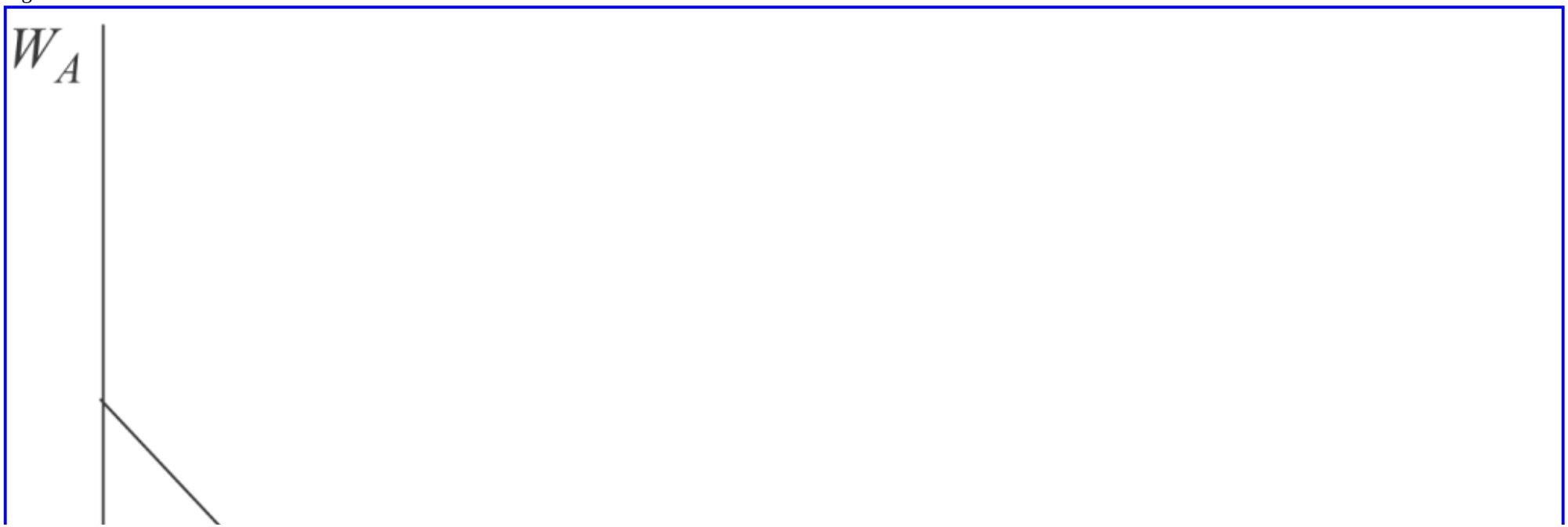

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

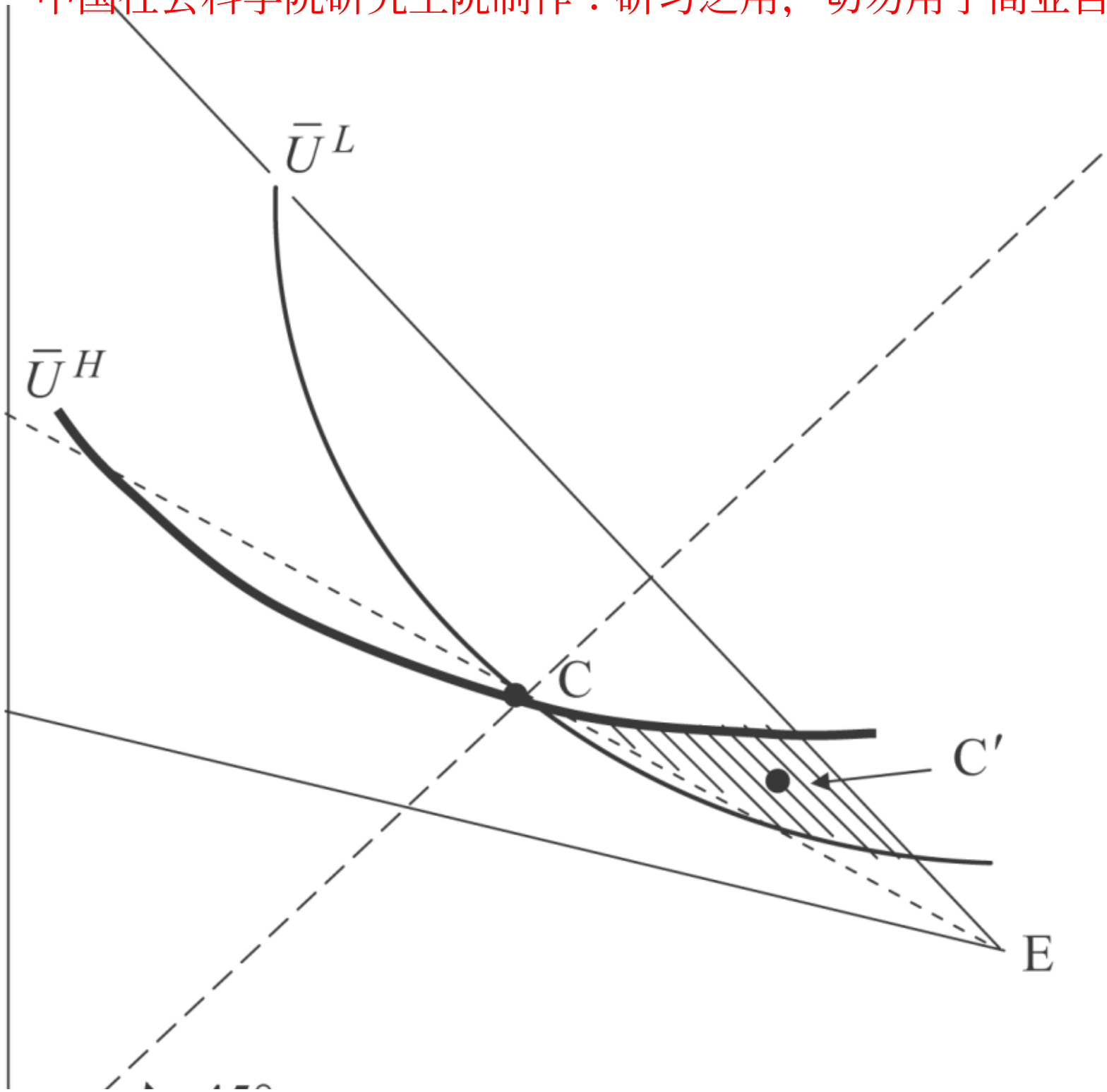

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

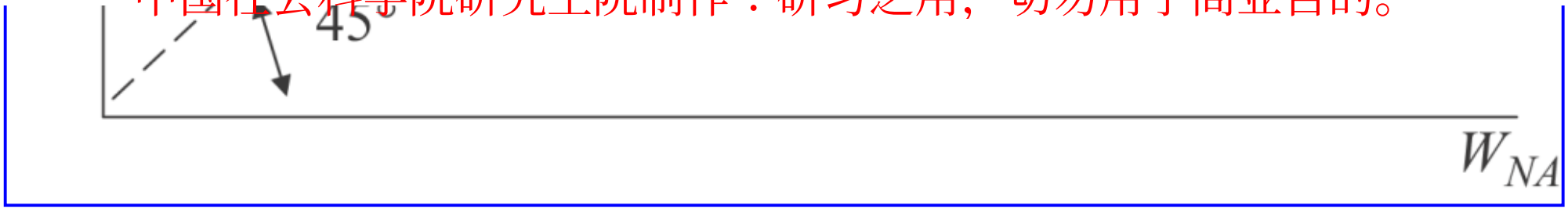

The only possible equilibrium, which is illustrated in Figure 4, is one in which the market distinguishes types by offering a contract B with complete coverage (which will be chosen by the high-risk individuals), and a contract D with partial insurance coverage (which the low-risk individuals prefer to full insurance at B and which the high-risk individuals do not prefer to full insurance at B). In this case, the market ‘solves’ the screening problem, but at the cost of foreclosing otherwise feasible and desirable exchanges.

Figure 4

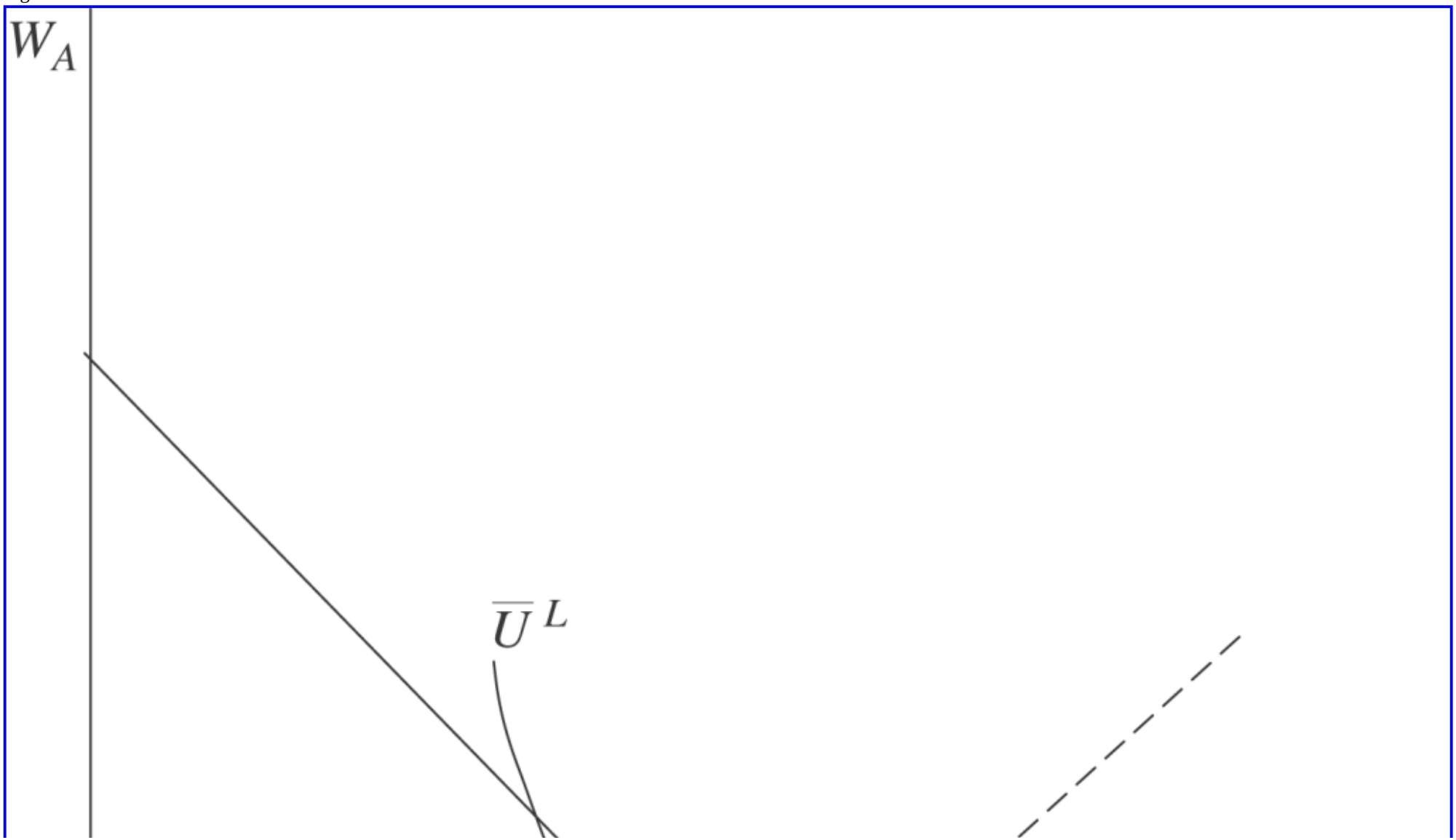

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

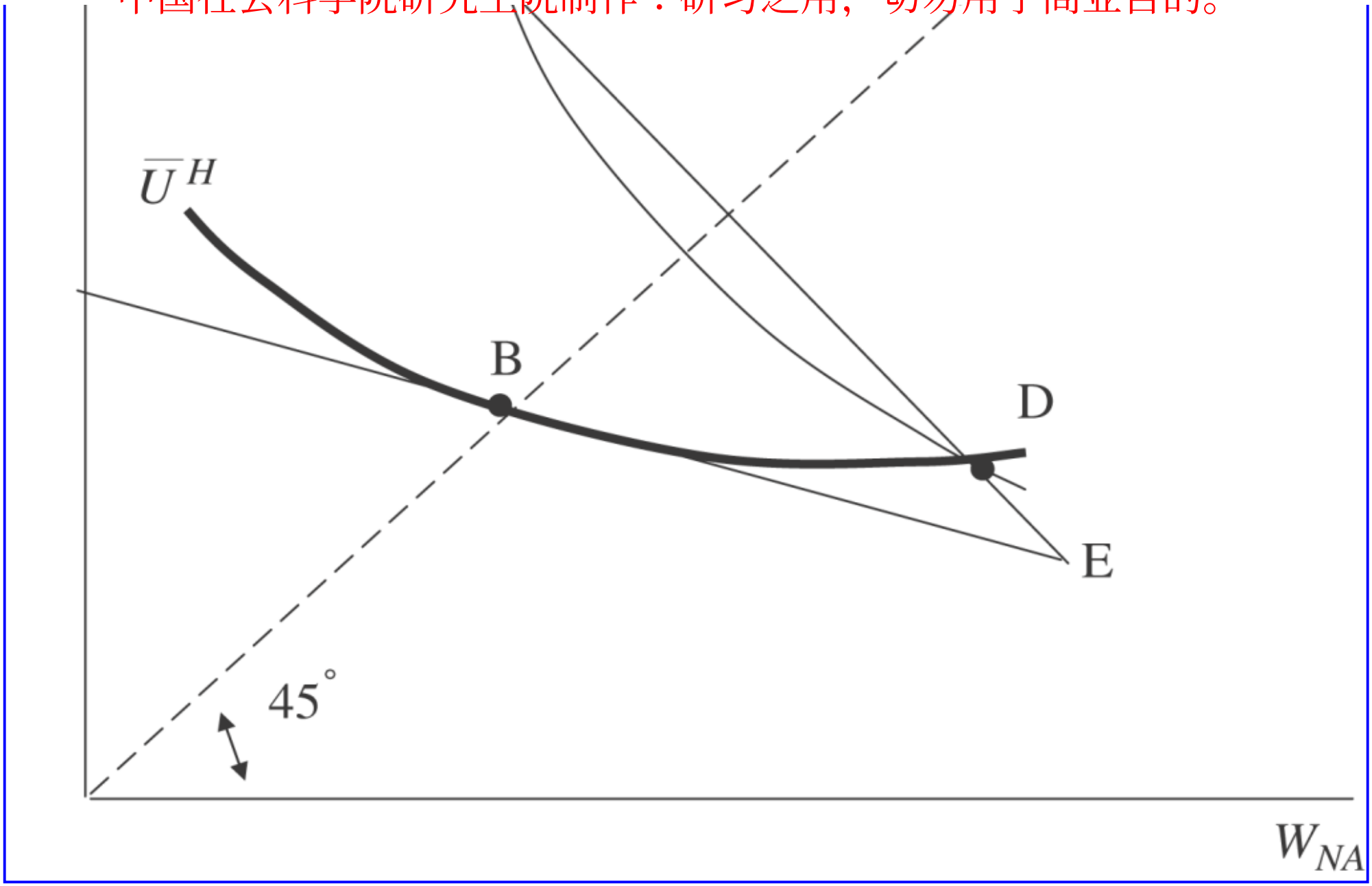

If, however, there are so few high-risk types that low-risk individuals are strictly better off at a pooling contract, then there can be no competitive equilibrium at all! The pooling contract with full insurance breaks the candidate separating equilibrium, but a contract at a slightly lower price and lower quantity of insurance cream skims the low-risk individuals. Thus the contract with full insurance cannot break even and is withdrawn.

With these intuitive graphs, Rothschild and Stiglitz demonstrate the non-robustness to considerations of imperfect information of two central results of the neoclassical model – that equilibrium is characterized by supply equals demand, and that equilibrium always exists. In papers published over the next two decades, Stiglitz demonstrated how market responses to the screening problem could explain puzzles in equity, credit, and labour markets. For instance, in equity markets, when insiders in a firm have more information than outsiders, the controlling insiders' willingness to issue equity conveys a signal that says that on average the shares are overpriced. The market responds by lowering the price. This discourages

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
firms from issuing new shares and provides an explanation for the fact that firms have limited access to equity (Greenwald, Stiglitz and Weiss, 1984).
In credit markets, if prospective borrowers have more information than lenders about the riskiness of their investments, there are situations in which a lender will set his interest rate below the market-clearing rate. He will not wish to raise his interest rate to what the market will bear if an increase in the rate would lead the lowest-risk borrowers to drop out of the market and reduce the lender's expected return. In this case, credit rationing will occur, as demonstrated in Stiglitz and Weiss (1981).
In many markets, individuals can, at a cost, provide credible information about their characteristics. Then ‘hidden characteristics’ become public information. This led Stiglitz to examine the incentives to acquire and transmit information. A central point is that the *private* returns to the provision of information differ from the *social* returns; thus the level of information that is public in a signalling equilibrium has, in general, no efficiency properties. To see this, consider the following simple example from Stiglitz (1975).
Suppose that there are two ability types whose productivity is A^H and A^L (the more able can do in one hour what the less able take A^H/A^L hours to do). Suppose a fraction of the population p is high ability and a fraction $1-p$ is low ability. Ability is private information but, at a cost C, an individual can reveal it (as would occur, for instance, if there is a credential that only a high-ability type is able to obtain but that certifies a skill unrelated to productivity). Then there exist two equilibria – a *separating* equilibrium and a *pooling* equilibrium – if

$$A^L < A^H - C < pA^H + [1-p]A^L.$$

If all other high-ability types screen, then the first inequality implies that the remaining high-ability individual has an incentive to screen. In doing do, he earns more than his alternative wage in the screening equilibrium, A^L. This establishes that a separating equilibrium exists.
However, if no other high-ability types screen, then both types are paid their average productivity, and the second inequality implies that the remaining high-ability individual has no incentive to screen. This establishes that a pooling equilibrium exists, as well, and that it Pareto dominates the separating equilibrium. In the separating equilibrium, each worker fails to take into account the effect of his decision to screen on the wage of unscreened individuals. Individual decisions create a diffuse externality.
The fact that the investor in information must obtain a positive expected private return from his information-gathering activities led Stiglitz to a fundamental result in finance. There had long been a theory, the *efficient markets theory*, which states that the observation of prices in capital markets suffices to reveal all relevant private information. Grossman and Stiglitz (1980) showed that the theory was incorrect: if information is costly and markets are competitive, then there must be an ‘equilibrium degree of disequilibrium’ – persistent discrepancies between prices and ‘fundamental values’ that provide incentives for individuals to obtain information. In capital markets, prices serve two functions: besides being used in the conventional way to clear markets, they also convey information. When individuals invest in information and thereby learn that the return to a security is going to be high (or low), they bid its price up (or down), and thus the price system makes that information publicly available. But if all information were publicly conveyed, there would be no incentives for individuals to invest in information.
Differential information can be a source of pure economic rents. Stiglitz argued that firms exploit that fact by creating ‘noise’. Knowing that it is costly for customers to search, Salop and Stiglitz (1977) showed that stores can exploit that by varying their prices to extract rents from customers with high search costs. The market equilibrium prices serve to discriminate (imperfectly) among individuals with different search costs. These results overturned a standard theory, *the law of the single price* (a given commodity is sold at the same price in all stores). In a similar vein, Edlin and Stiglitz (1995) argued that managers will have an incentive to enhance asymmetries of information between them and rival managers and boards of directors, and thereby limit the scope for takeovers.

Hidden actions and agency theory

In the standard neoclassical model, there are no conflicts of interest between economic actors. *Principal–agent* theory introduces conflicts of interest by specifying actions that cannot be observed. In this theory, a principal, who delegates a task to an agent, designs a contract that makes payment depend on observable circumstances (for example, revenues) that are correlated with the desired, but unobservable, actions of the agent. Stephen Ross, James Mirrlees, and Joe Stiglitz contemporaneously developed principal–agent theory.
Stiglitz's initial contribution was stimulated by his observations in Kenya during parts of 1969–71. He analysed a puzzle that had been recognized at least since Alfred Marshall – the apparent inefficiency of the institution of sharecropping, which assigns the tenant only a share of the marginal return to his effort. Stiglitz (1974) showed that sharecropping could be advantageous to tenants and landlords because of the *savings in monitoring costs* compared to a wage system with costly monitoring, the *increases in output* compared with a wage system with imperfect monitoring, and the *reduction in risk borne by tenants* compared with a system where workers pay fixed land rents but do not have access to risk markets.
Four insights in this paper and Stiglitz's other work in principal–agent theory have been important for further developments in economics.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1. 1. It is sometimes useful to take the *transaction*, rather than the *market*, as the unit of analysis.
2. 2. There is a trade-off between incentives and insurance if the principal has greater ability to bear risk than the agent. Since the first-order effect of distorting incentives is zero, the provision of a small level of insurance is in this case always welfare-increasing.
3. 3. The distribution of wealth influences the extent of agency problems, in both rich and poor countries.
4. 4. Agency problems are pervasive in a complex modern economy. Agency theory has contributed to new theories of public finance, of corporate governance, and of positive political economy.

However, in many situations, incentive contracts cannot be written because an individual's individual contribution to output is not well observed. In that case, pure economic rents can play an important role in providing incentives, as in Shapiro and Stiglitz (1983). But integrating the idea of rents into a model of a competitive economy initially posed a puzzle. If price exceeds marginal cost, why doesn't competition lead to price-cutting?
An antecedent to Shapiro and Stiglitz (1984) was the model in Shapiro (1983) of rents as an incentive to quality that is unobservable at the time of purchase. In that model, firms develop a reputation for quality by the goods that they produce. The prospect of the loss of rents to a firm that ‘milks’ its reputation by selling at a high price less than the promised quality induces the firm to live up to consumers’ expectations. Competition does not lead to price-cutting because consumers come to learn that, if the price is too low, firms do not have an incentive to maintain their reputation, and therefore the offer of high-quality goods at a low price is not credible.
Shapiro and Stiglitz (1984) extends to a labour market the idea of rents as an incentive device for difficult-to-monitor effort. Workers in this model are identical (so there is no selection issue). Firms observe at random intervals whether a worker is working or shirking. To elicit effort, each firm would want to offer a higher wage than other firms so that, if it finds a worker shirking, he suffers a cost when he is fired. But if it benefited one firm to raise its wage, it would benefit all firms. This might seem like the dilemma where, if every spectator in the stadium stands up to get a better view, no one sees any better. But, when all employers raise their wages, those actions have a real effect on the economy: unemployment emerges since the higher wage rations firms’ demand for workers. Now a worker who is fired cannot immediately find another job. This makes job loss costly. Unemployment creates an incentive to work on the job rather than shirk, and so competitive equilibrium will be characterized by unemployment and pure labour rents.

Macroeconomics

Informational imperfections limit the scope of equity and credit markets, as well as insurance and labour markets. In a series of papers with Bruce Greenwald and Andrew Weiss (for example, Greenwald, Stiglitz and Weiss, 1984), Stiglitz drew out the implications for the fluctuations in output and employment that have characterized capitalism throughout its history. The central argument was this. Limitations in the scope of equity markets in the presence of significant bankruptcy costs lead firms to behave in a risk-averse manner. They pay attention to own risk, while traditional theory suggests that the only risk that firms should care about is the correlation with the stock market. Higher levels of investment or production entail increased debt, and, as debt is increased, the bankruptcy probability is increased. Firms will therefore produce and invest only up to the point where expected marginal returns equal expected marginal bankruptcy costs. This has four implications that contrast with what would occur with perfect markets:

1. 1. *Amplification of small shocks*. Changes in the net worth of the firm or in the riskiness of the environment affect the production and investment decisions of the firm (in contrast to the standard theory). For a highly leveraged firm, small changes in demand can result in large changes in output and employment. Thus, disturbances to the economy tend to be amplified.
2. 2. *Persistence*. If for some reason net worth is reduced at a given time, production falls in subsequent periods. Only gradually will production be restored to normal, as net worth builds up again.
3. 3. *Risk-averse banks*. Banks are a specialized kind of firm whose production activity is making loans. A reduction in the net worth of banks and an increase in the riskiness of their environment will lead them to contract their output – that is, to make fewer loans, which has multiplier effects throughout the economy.
4. 4. *Worsening of the applicant pool for loans during a recession*. For any given bankruptcy cost, there is a critical net worth such that, below that net worth, firms act in a risk-loving manner, and, above that net worth, in a risk-averse manner. If the economy moves into a recession and firms find their net worth decreases, good (that is, risk-averse) firms reduce their loan applications, bad (that is, risk-loving) firms increase their loan applications, so that there is an increasing proportion of bad (that is, low net worth) applicants. These effects may be so strong as to lead to a situation where banks make no loans at all!

During Stiglitz's tenure as Chief Economist of the World Bank, the contrast between his perspective on macroeconomics and the perspective based on well-functioning markets came to a head. These are two starkly different ways of looking at the world. If there are well-functioning markets, then opening up capital markets will lead to efficient outcomes. This view was identified with the US Treasury Department and the IMF in the 1990s. During the 1997–99 East Asian financial crisis, a condition of IMF financial support was that the East Asian economies adopt contractionary fiscal and monetary policies. The contractionary monetary policy would raise interest rates and, at some point, reverse private capital

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
outflows and restore the ability of the East Asian countries to repay their foreign debts. It was argued that the basic reason for the East Asian financial crisis was lack of transparency, or corruption, in the business practices of these economies, which frightened away foreign investors.
In contrast, Furman and Stiglitz (1998) argued that a lack of transparency did not cause the crisis (although it aggravated the effects of the downturn once it began). They argued that small developing countries are financially fragile. There are pervasive externalities in banks' and firms' decisions to obtain short-term loans from abroad. Each bank and each firm takes the risk environment as given, and yet the aggregate set of decisions determines the risk of a financial crisis. This meant that some limits on free capital markets were appropriate in developing countries. Moreover, Furman and Stiglitz argued that policies that increased interest rates in the East Asian economies would greatly erode the net worth of debtors, and the erosion of their net worth would lead to a recession that could not easily be reversed.
Stiglitz's views ultimately were influential. However, the openness of his conflict with the IMF and US Treasury frayed his relationships with many people in Washington and hastened his departure from the World Bank.

Development economics

Whereas macroeconomics remains split between different schools with contrasting views on the importance of market imperfections, the centrality of market imperfections in the field of development economics is not questioned. Before the development of the economics of information (and also the development of game theoretic models of political economy), economists lacked a broad framework for understanding of the sources of the imperfections in markets. Economists who tried to design policies to fit developing country markets generally assumed rigidities in markets, but did not explain them by reference to a choice-based perspective. Abhijit Banerjee (2001, p. 465) has characterized development economics in this era as the 'ugly duckling' of economics: 'It was full of strange assumptions and contrary logic, and all the other [fields of] economics made fun of it.'
Stiglitz's work in development economics played a major role in transforming the field. His models were important in establishing (*a*) that positive feedback mechanisms can give rise to multiple equilibria and underdevelopment traps; (*b*) that, because the causes of market failures and constraints on growth vary greatly from setting to setting, analysis has to be done case by case; and (*c*) that non-market institutions need have no efficiency properties. Important applications of these ideas are below.
Trade-off between diversity of goods and scale economies. In a path-breaking model, Dixit and Stiglitz (1977) posed a question seemingly unrelated to development. This paper addressed the question: will a market solution yield the socially optimum kinds and quantities of commodities if there are multiple possible varieties of goods, each produced by a single firm with increasing returns to scale in production? The desire by consumers for diversity meant that there would be many firms, but not necessarily the optimal number. Dixit and Stiglitz used a modelling assumption that turned out to be very useful analytically. By assuming a continuum of goods, their set-up lets the modeller respect the discrete nature of many location decisions and yet analyse the model in terms of the behaviour of continuous variables like the share of manufacturing in a particular region.
This model became a building block in models in the new fields of endogenous growth theory and economic geography. To understand the flavour of this work, consider an economy with three sectors: a low-technology sector, an advanced sector, and an intermediate sector that produces an array of non-traded, i.e., domestic, goods, modelled as Dixit–Stiglitz commodities, which are inputs into the advanced sector. An expansion of the advanced sector increases the demand for non-traded inputs, and so lowers their average costs and increases the available variety. With a greater variety of intermediate inputs, production in the advanced sector is more efficient. It can thus be the case that, when many other firms enter the advanced sector, it pays the remaining firms in the traditional sector to do so; but, when all other firms remain in the traditional, low-technology sector, it pays the remaining firm to do so, too. A low-level equilibrium can therefore be sustained even when the economy is fully open to international trade.
Breakdown of the Washington consensus. The standard neoclassical model predicts that growth is inevitable in capital-poor market economies: over time, all economies will converge in per capita income. This model led a generation of economists to a simple set of policy prescriptions that would set the preconditions for growth: maintain macroeconomic stability (since high inflation interferes with the workings of the price system), limit government ownership of enterprise, and deregulate ('stabilize, privatize, and liberalize'). This so-called Washington Consensus has broken down, in part because it has become clear that there are no sure-fire formulas for success, and in part because of the evolution of economic theory away from the perfect markets paradigm. The three central developments in economic theory have been the economics of information, game theory, and institutional economics. In the new economic theory, development is no longer seen primarily as a process of capital accumulation, but instead as a process of organizational change. Evidence of the breakdown of the Washington consensus is that a recent World Bank volume that reviews economic growth in developing countries in the 1990s states that 'The central message of this volume is that there is no unique universal set of rules [to promote growth]…[W]e need to get away from formulae and the search for elusive "best practices".' (World Bank, 2005, p. xiii).
Dysfunctional institutions. In the past, many scholars have made the argument that institutions that emerge out of individual actions are necessarily optimal: they are there because their benefits outweigh the costs. Stiglitz's work on sharecropping (Stiglitz 1974) exemplifies that approach. However, as Stiglitz has often remarked, that analysis is partial equilibrium. That analysis studies the optimal contract while holding fixed everything else in the economy.
In many cases, however, contracts that individuals enter into impose externalities on other agents. There may be no forces that ensure the Pareto efficiency of the set of contracts that individuals adopt. For instance, when insurance is provided through family and friends as well as through the market, the informal insurance will *raise* welfare if it provides (sufficient) peer monitoring (Arnott and Stiglitz, 1990), but, otherwise, such insurance will *lower* welfare because the additional insurance exacerbates moral hazard and so raises the cost of market insurance – an effect that no individual internalizes. The analysis of the inefficiency of contracting choices generalizes widely, for example, to technological change

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(Acemoglu, 1997), to neighbourhood formation (Hoff and Sen, 2005), and to institutional change, as I discuss next.

One of the most important economic transformations in modern history began with the collapse of Communism in Eastern Europe and the former Soviet Union…. The transition process in the 1990s entailed the creation of a new set of economic and political institutions and, in most countries, produced an unexpectedly deep and prolonged depression. In Russia and many other transition countries, the rapid transfer of state enterprises to private hands (‘Big Bang privatization’) did not lead to a political demand for institutional reforms needed to govern private property, as many economists had expected. Hoff and Stiglitz (2004; 2007) investigate the influence of economic policies on the demand for a rule of law. They show that Big Bang privatization can create powerful incentives to strip assets and to delay the establishment of a rule of law. This may result in a long period of economic decline. The cause is that no individual takes account of the effect of his economic choices on long-run institutional change.

Evaluation

The high level of idealization in much of Stiglitz's formal work, and the surprising (or at least counter-Walrasian) results often obtained, have led his harshest critics to see in his work a predilection for the intriguing exception rather than the general rule: granted that market failures occur, how much do they really matter? From a staunch admirer, Avinash Dixit, one hears the statement that a paper by Stiglitz begins with the phrase, ‘Assume there are two types’. However, the statement that ‘there are two types’ (or ‘two actions’) in Stiglitz's papers of the 1970s and 1980s marked a radical departure from the standard model, which implicitly assumes that in each market there is only one type (or one action); that is, that information in the market is symmetric. This modest relaxation of the perfect information assumption reveals that symmetric information is essential to the results of the standard neoclassical paradigm.

Stiglitz's work demonstrates that the standard theory misconstrues many of the virtues of the market. The standard theory exaggerates the role of prices in conveying information about scarcity, and fails to take account of the difficulties of making the price system work. At the same time, the standard theory fails to recognize some central virtues of the market – its ability to address problems of selection, incentives, information gathering, and innovation – because the standard paradigm is silent about all these problems.

Compared to the state of economics in 1970, mainstream theory can now accommodate a far broader ranger of phenomena. But there is no unified single framework, as there was in the Walrasian paradigm. Instead, there is a fragmented collection of disparate models. Distinct models are capable of explaining the same phenomena but are difficult to distinguish empirically.

A further problem that remains for future work is that multiple forms of private information exist within any sector. With multiple incentive problems, it is necessary to consider the distortion among incentives that results when incentives for more easily observed actions are created at the expense of less easily observed actions.

In these ways, Stiglitz's theoretical work has contributed to the resurgence of empirical work in economics. Kim, Morse and Zingales (2006) document a reversal in the previous 30 years in the importance of theoretical work, which dominated the profession in the 1970s and 1980s, and gave way to the primacy of empirical work in the early 1990s. Much of that empirical work is a response to a body of theory that established that neither markets nor governments work perfectly. Stiglitz's demonstration that imperfect information undermines the results of the standard neoclassical model has shifted not only models of thought in economics, but also the relative importance of different sources of knowledge about the economy.

See Also

- adverse selection
- credit rationing
- moral hazard
- principal and agent (i)
- principal and agent (ii)
- selection bias and self-selection
- Washington Consensus

Selected works

1970. (With M. Rothschild.) Increasing risk: I. A definition. *Journal of Economic Theory* 2, 225–43.

1971. (With M. Rothschild.) Increasing risk: II. Its economic consequences. *Journal of Economic Theory* 3, 66–84.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1974. Incentives and risk sharing in sharecropping. *Review of Economic Studies* 41, 219–55.

1975. The theory of ‘screening’, education, and the distribution of income. *American Economic Review* 65, 283–300.

1976. (With M. Rothschild.) Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90, 630–49.

1977. (With S. Salop.) Bargains and ripoffs: a model of monopolistically competitive price dispersion. *Review of Economic Studies* 44, 493–510.

1977. (With A. Dixit.) Monopolistic competition and optimal product diversity. *American Economic Review* 67, 297–308.

1980. (With S. Grossman.) On the impossibility of informationally efficient markets. *American Economic Review* 70, 393–408.

1981. (With A. Weiss.) Credit rationing in markets with imperfect information. *American Economic Review* 71, 393–410.

1984. (With C. Shapiro.) Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

1984. (With B. Greenwald and A. Weiss.) Informational imperfections in capital markets and macroeconomic fluctuations. *American Economic Review* 74, 194–99.

1986. (With B. Greenwald.) Externalities in economies with imperfect information and incomplete markets. *Quarterly Journal of Economics* 101, 229–64.

1991. (With R. Arnott.) Moral hazard and non-market institutions: dysfunctional crowding out or peer monitoring. *American Economic Review* 81, 179–90.

1993. (With B. Greenwald.) Financial market imperfections and business cycles. *Quarterly Journal of Economics* 108, 77–114.

1994. *Whither Socialism*. Cambridge, MA: MIT Press.

1995. (With A. Edlin.) Discouraging rivals: managerial rent-seeking and economic inefficiencies. *American Economic Review* 85, 1301–12.

1998. (With J. Furman.) Economic crises: evidence and insights from East Asia. *Brookings Papers on Economic Activity* 1998(2), 1–114.

2004. (With K. Hoff.) After the Big Bang: obstacles to the emergence of the rule of law in post-Communist societies. *American Economic Review* 94, 753–63.

2007. (with K. Hoff.) Exiting a Lawless State. Manuscript.

2007. *Selected Scientific Papers of Joseph E. Stiglitz*, vol. 1. Oxford: Oxford University Press.

Bibliography

Acemoglu, D. 1997. Training and innovation in an imperfect labour market. *Review of Economic Studies* 64, 445–64.

Akerlof, G. 1970. The market for ‘lemons’: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53, 941–69.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Banerjee, A. 2001. Comment on K. Hoff and J.E. Stiglitz, ‘Modern Economic Theory and Development’. In *The Future of Development Economics in Perspective*, ed. G. Meier and J. E. Stiglitz. Oxford: Oxford University Press.

Hoff, K. 2001. Beyond Rosenstein-Rodan: the modern theory of coordination problems in development. *Annual World Bank Conference on Development Economics 2000*, 145–176.

Hoff, K. and Sen, A. 2005. Homeownership, community interactions, and segregation. *American Economic Review* 95, 1167–89.

Kim, E.H., Morse, A. and Zingales, L. 2006. What has mattered to economics since 1970? *Journal of Economic Perspective* 20(4), 189–202.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Shapiro, C. 1983. Premiums for high quality products as returns to reputations. *Quarterly Journal of Economics* 98, 659–80.

Spence, A.M. 1974. *Market Signaling: The Information Structure of Hiring and Related Processes*. Boston, MA: Harvard University Press.

World Bank. 2005. *Economic Growth in the 1990s: Learning from a Decade of Reform*. Washington, DC: World Bank.

How to cite this article

Hoff, Karla. "Stiglitz, Joseph E. (born 1943)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000447> doi:10.1057/9780230226203.1620

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stigma

Lawrence E. Blume and Steven N. Durlauf
From The New Palgrave Dictionary of Economics Online, ,
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The many contemporary characterizations of stigma share the assumption that individuals have characteristics that devalue them in the eyes of others. The observer's response leads to status loss and discrimination. Stigmatized attributes, their linkage to a particular set of stereotypes and the general response to those stereotypes are shared within a particular ingroup, and thus are fundamental to social identity. Stigmatization affects the allocation of resources through direct discrimination, self-fulfilling beliefs, and identity threat.

Keywords

Becker, Gary; benefits; contact hypothesis; crime; demographics; discrimination; Goffman, Erving; identity threat; Myrdal, Gunnar; outgroup; racial discrimination; resource allocation; self-fulfilling beliefs; slavery; stereotypes; stigma; vicious circle; welfare programmes

Article

Contemporary research on stigma begins with Erving Goffman's (1963) essay, *Stigma: Notes on the Management of a Spoiled Identity*. He writes:

> While the stranger is present before us, evidence can arise of his possessing an attribute that makes him different from others in the category of persons he is available to be, and of a less desirable kind…He is thus reduced in our minds from a whole and usual person into a tainted, discounted one. Such an attribute is a stigma, especially when its discrediting effect is very extensive; sometimes it is also called a failing, a shortcoming, a handicap. It constitutes a special discrepancy between virtual and real social identity. (Goffman, 1963, pp. 2–3)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Contemporary characterizations of stigma (and there are many) share the assumption that individuals have (or are assumed by others to have) characteristics that devalue them in the eyes of others. These characteristics trigger a stereotype response on the part of the observer which leads to status loss and discrimination against the status holder. Goffman observes that the essence of stigma is relational, the linking of attributes to stereotypes. Stigma is distinct from a simple stereotype response because of its social component. Stigmatized attributes, their linkage to a particular set of stereotypes and the general response to those stereotypes are shared within the ingroup, and thus are fundamental to the determination of social identity. The attributes tagged by the social process of stigmatization may have little or nothing to do with individual function, and little or nothing to do with the individual's self-perceived identity as opposed to their social identity (to the extent that these can be meaningfully separated). Triggers can be exogenous characteristics (types), such as gender or ethnicity, or behaviours, such as being a felon, or uneducated. They can be observed (race) or merely inferred (religion). They certainly vary across cultures, space and time (Crocker, Major and Steele, 1998).

Stigma and resource allocation

Stigma is a profoundly important social process, with significant implications for individuals' ability to achieve desired social outcomes. There are at least three channels through which stigmatization affects the allocation of resources: (1) direct discrimination, (2) self-fulfilling beliefs, and (3) identity threat.

Discrimination

Direct discrimination in the United States was dramatically described in Gunnar Myrdal's seminal *An American Dilemma* (1944), and brought to the centre of economic research in Gary Becker's classic *The Economics of Discrimination* (1957). Becker's analysis emphasizes what has come to be known as taste-based discrimination, which can be understood as the assignment of negative attributes to an individual on the basis of their membership in an outgroup. Becker's notion of taste-based discrimination has been extended to that of stigma in work such as Price, Darity and Headen (2008), who explore how former slave status increased the probability that an individual would be lynched in the post-bellum South. These authors remark that ‘The consequences of racial stigma follow from a decision-making criterion by whites, who view blacks as stigmatized by slavery, which warrants standards of treatment in black–white social interactions that are lower than the standards warranted in typical human interactions (Price et al., p. 168). The empirical salience of this approach in understanding contemporaneously inequality is explored by Charles and Guryan (2007).

Goffman's notion of stigma is of course deeper that simply identifying a fixed effect associated with race or some other criterion for group membership, which is often how stigma is instantiated in empirical work. Besley and Coate (1992) undertake a theoretical analysis of welfare stigma which considers two sources of stigma. The first type corresponds to Goffman's notion of stigma as ascription. Individuals who are on welfare are stigmatized when participation in the programme is used to evaluate whether the individual is a shirker as opposed to disabled. A second notion of stigma developed by Besley and Coate focuses on taxpayer resentment, and is a function of the gap between what is regarded as an appropriate level of benefits and actual benefits. The two models differ in their conclusions concerning how stigma

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
varies with respect to the size and composition of the poor population. It is worthwhile observing that while economists have largely been concerned with the group implications of stigma, resource redirection resulting from stigma cuts across many group boundaries. A striking example is Crandall's (1995) observation that the parents of heavy daughters are less likely to pay for their college education than are the parents of average-weight women.

Self-fulfilling beliefs

The problem of self-fulfilling beliefs was first and beautifully described not by an economist but by a sociologist. Robert Merton's ‘The self-fulfilling prophecy’ (1947) not only coined a phrase, but carefully expressed the idea that beliefs give rise to decisions which confirm the beliefs, which would not otherwise be true. One of his leading examples addresses the stigmatization of black workers by pre-war union leaders as strikebreakers and scabs:

> History creates its own test of the theory of self fulfilling prophecies. That Negroes were strike-breakers because they were excluded from unions (and from a large range of jobs) rather than excluded because they were strikebreakers can be seen from the virtual disappearance of Negroes as scabs in industries where they have gained admission to unions in the last decade.’ (Merton, 1947, p. 197)

Curiously, self-fulfilling prophecies are not unrelated to Myrdal's (1944) principle of cumulation, or cumulative causation, or the *vicious circle*, wherein beliefs and institutional structures are self-reinforcing. Myrdal's principle of cumulation, which appears earlier in his (1939) *Monetary Equilibrium*, is a multiplier effect. In *An American Dilemma* Myrdal identified a social multiplier effect. He argued that the degradation of daily life for the black community further reinforced the beliefs of the whites, whose collective political and social conditions instantiated the conditions at the outset.

The modern expression of Merton's self-fulfilling prophecy is an expectations equilibrium, as in rational expectations models and in Nash equilibrium. This was expressed initially in Arrow's (1972, 1973) models of statistical discrimination. In these labour market equilibrium models, employers believe the unobservable human capital investments of the outgroup to be low, and therefore do not hire them. Consequently, the return to human capital investment by outgroup members is low. They choose not to invest, and so the ingroup employers' beliefs are correct. Social psychologists too have recognized self-fulfilling prophecies as a general phenomena. Their formulation has more to do with social process than the equilibrium formulations economists employ (Darley and Fazio, 1980).

The role of beliefs in the persistence of statistical discrimination has been a major source of theoretical developments. The contact hypothesis, due to Allport (1954), claims that stigmatic associations will be reduced if in- and outgroups come into closer contact. One justification for affirmative action programmes is that they do exactly this; they force otherwise separate groups to come into closer contact. Coate and Loury (1993) have described a theory of affirmative action in labour markets where statistical discrimination is prevalent. They find that affirmative action policies have ambiguous effects on employers' beliefs. Affirmative action may fail because to implement an affirmative action policy, employers will have to lower standards for hiring outgroup employees. But if standards are lowered, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
incentives for human capital investment are reduced.

Chaudhuri and Sethi (2008) offer an interesting variant of the contact hypothesis. The argument of the contact hypothesis is that groups in contact with each other will abandon stereotypes in favour of more refined and nuanced views of the meaning of group membership. That is, contact changes beliefs. Chaudhuri and Sethi suppose instead the existence of positive externalities in human capital acquisition. Integration lowers the cost of skill acquisition to the relatively unskilled group. Their model exhibits a discontinuity, a threshold of integration above which negative stereotypes cannot persist in equilibrium.

Coate and Loury's analysis indicates how statistical discrimination models are of particular interest because of their ability to produce multiple equilibria such that in one equilibrium the outgroup is stigmatized, whereas in another employers discriminate against members of their own group, and in a third equilibrium no discrimination occurs at all. Whether some equilibria are more salient, more likely to arise, is a question similar to that addressed in the extensive literature on Nash equilibrium refinements. Blume (2005, 2006) develops learning models of statistical discrimination, in which payoff parameters of the model determine whether discrimination will persist in the long run and show how this persistence can in fact occur.

Identity threat

Identity threat, or stereotype threat, is a channel distinct from expectational equilibrium through which stigma affects individual outcomes. In situations where one's attributes or behaviours risk confirming the plausibility of negative stereotypes, for others or even for oneself, members of stigmatized groups may respond behaviourally.

> Consider the stereotypes elicited by the terms *yuppie*, *feminist*, *liberal*, or *White male*. Their prevalence in society raises the possibility for potential targets that the stereotype is true of them, and also that other people will see them that way. When the allegations of the stereotype are importantly negative, this predicament may be self-threatening enough to have disruptive effects of its own. (Steele and Aronson, 1995, p. 797)

Steele and Aronson (1995) investigated the performance of black students in difficult verbal exams. Through different ways of varying the task description and of activating stereotypes about the academic performance of black students, they were able to conclude that the stereotypes led black students to underperform relative to whites when the test was presented as diagnostic of academic skills, but not when it was presented differently. Steele (1992) has argued that the pressure of negative stereotypes causes students to disidentify with school and school tasks, reduce motivation to succeed, and even creates black peer pressure to disidentify, all of which leads to bad schooling outcomes and all that follows. While economists' concern with stigma has mostly been focused on discrimination and statistical discrimination, most recent stigma research by social psychologists is concerned with identity threat (Major and O'Brien, 2005).

Relatively little work has been undertaken by economists on identity threat, although identity has become a topic of interest. Not surprisingly, a few experimentalists have replicated the findings of experimental social psychology, only using the economist's war horses: ultimatum games, and risky and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
intertemporal choice problems. Benabou and Tirole (2006) build a cognitive model of belief management. The idea of their model is that individuals are uncertain and even forgetful of their deep motivations. They make inferences from their past choices, and so their past choices come to define their self-perception. Two of their key findings are that identity investments are largest in situations where self-information is scarce, and when one is most uncertain of one's identity. They argue that the non-monotonicity of identity investment can explain a confirmatory response to identity threat, for instance by opting out of threatening situations such as academic achievement exams.

Implications of stigma

Stigma has been used to explain empirical phenomena that seem puzzling under narrow conceptualizations of rational choice theory. One well-studied example concerns stigma and participation in public assistance programmes. In a widely cited paper, Moffitt (1983) estimates welfare participation regressions which incorporate both fixed disutility to participation and a monetary cost that is proportional to the benefit. He argues that these explain why many eligible poor persons do not use available welfare programmes. In this context, the failure to pursue public assistance represents the puzzle; stigma associated with the receipt of welfare is argued to explain why some eligible individuals forgo assistance. The empirical analysis in this paper finds evidence of a fixed stigma cost. A limitation of this type of analysis is that it does not allow for other explanations of nonparticipation; for example, Daponte, Sanders and Taylor (1999) find that ignorance about public assistance programmes is a major source of nonparticipation; survey evidence is reported that suggests that stigma-type reasons are primary for about 6 per cent of those eligible but not participating in the US Food Stamps programme.
Stigma has also been used to understand the dynamics of nonmarital fertility; here the fact to be understood is the rise of nonmarital fertility in light of the effects on the socioeconomic prospects of mother and child. Akerlof, Yellen and Katz (1996) argue that a large component of the observed increases in out-of-wedlock birth rates in the United States between the late 1960s and the late 1980s is a result of the decline in the rate with which an out-of-wedlock conception would end in a within-wedlock birth: that is, the rate of ‘shotgun marriage’. Their formal argument is that changes in the technology of contraception and the onset of legally available abortions are a technology shock which decreases the bargaining power of those unable to take advantage of them. Stigma, however, plays two roles in their analysis. One they explicitly mention is that an increase in the out-of-wedlock birth rate, by increasing the number of unwed mothers, destigmatizes the category, which lowers the cost of not terminating a pregnancy. The stigma reduction changes the parameters of the relevant model in a way that reinforces the decline in shotgun weddings and the increase of out-of-wedlock births.
A second role for stigma, not explicitly mentioned, is in male decision making. They quote a male in a pre-shock ethnographic study of shotgun marriage who says, ‘If a girl gets pregnant you married her. There wasn't no choice. So I married her.’ Why was there ‘no choice’? Men who refused to do the ‘right thing’ were stigmatized. Akerlof, Yellen and Katz (1996) quote an Internet (and therefore post-shock) contributor to a newsgroup who wrote, ‘Since the decision to have the child is solely up to the mother … I don't see how both parents have responsibility to that child.’ This is to say, turning a pre-nuptial pregnancy from a birth by necessity into a birth by woman's choice led to the destigmatization of the male decision not to marry. Nechyba (2001) develops a dynamic model of nonmarital fertility based on

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
stigma, in which stigma is determined by a weighted average of the nonmarital fertility of past generations.

A third area where stigma has been invoked concerns crime. An early example of this type of argument is Rasmusen (1996). Rasmusen considers models in which conviction for a crime provides information to an employer either because of on-the-job misbehaviour (crime thus representing a moral hazard problem) or due to its informativeness concerning productivity-related unobservables (which represents an adverse selection problem). Rasmusen demonstrates that both models can have multiple, Pareto-ranked equilibria, and then nonetheless extols stigma manipulation as a deterrence strategy.

O'Flaherty and Sethi (2008) use racial stereotypes in a two-sided statistical discrimination model to explain black–white crime patterns in the United States. Blacks are more likely than either whites or Hispanics to be involved in crime, both as victims and as offenders. Furthermore, while crimes by whites against whites, by blacks against blacks and by blacks against whites are common, crimes by whites against blacks are rare. We might have expected white crimes against blacks to be abundant if it were thought that police were racist or that they devalued black safety, or that black witnesses were regarded by juries as less credible than whites. O'Flaherty and Sethi explain this pattern by constructing an incomplete information model of criminal/victim interaction. A potential criminal must decide whether or not to commit a crime, and a victim must decide whether or not to resist. If all victims believe that blacks are more likely to engage in violence during a crime, and if all criminals believe that whites are less likely to resist, then these beliefs can be sustained in equilibrium and the equilibrium has the black/white crime patterns described above.

Blume (2002) has built a dynamic game model of criminal behaviour wherein apprehended criminals pay both a direct penalty (fine or incarceration) and a stigma cost. For some time after conviction individuals are publicly labelled as criminals. The cost of being stigmatized is decreasing in the number of individuals similarly stigmatized – if everyone is a criminal, crime bears no stigma. The incentive effects of direct penalties are those of the standard Beckerian model. Similarly, shifting upward the stigma cost function decreases crime. Increasing the duration of stigmatization increases the cost of being a criminal, but it can nonetheless have a perverse incentive effect because it can lead to more people being stigmatized at any given time. The stream of stigma costs comes from farther down the cost curve. Consequently, labelling individuals as felons for their entire post-incarceration life may have a negative impact on crime deterrence.

Kendall and Tamura (2008) make a subtle argument about the impact of stigma on the relationship between out-of-wedlock births and crime. They observe that over the period from 1923 to 2002 in the United States, a rate of ten non-marital births per 1,000 live births is associated with increases of between 2.5 per cent and 5 per cent in murder and property crime rates. This relationship, however, is complicated by the match quality of the parents. In the 1940s and 1950s, when out-of-wedlock birth was stigmatized and termination was costly, only the lowest-quality matches, matches that would not have beneficial to the child, did not take place. From the 1970s, however, when the stigma pressure to marry declined, matches did not take place that would in all likelihood have benefited the child. Thus the positive relationship between out-of-wedlock fertility and future crime would have increased.

Finally, recent work by Glenn Loury (2002) has invoked racial stigma as an alternative to discrimination in understanding persistent inequality between blacks and whites. Loury defines racial stigma as ‘dishonorable meanings socially inscribed on arbitrary bodily marks, of “spoiled collective identities”‘ (2002, p. 59), thus employing Goffman's language. While Loury's discussion of racial stigma

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
involves ideas that closely relate to statistical discrimination, he departs from this approach in his emphasis on the imprecision of the use of group identity to draw individual inferences. This is seen in Loury's discussion of racial profiling in policing. He argues that:

> [U]ncovering specification error in this example – learning that a group's average behavior is a poor predictor of what to expect from almost everyone in the group – requires a law enforcement agent to avoid constructing his subjects' social identities ‘from the outside’ But, and this is crucial, given his initial beliefs, he will expect it to be a waste of time and resources to retain idiosyncratic data. He perceives little gain from distinguishing individuals because he begins with the conviction that they are all alike. And yet unless he tracks individuals he cannot learn that he is dead wrong in this presumption. If he is not forced to track individuals then, short of a coincidental sequence of fortuitous occurrences, he is likely to persist in his erroneous belief and act accordingly. (Loury, 2002, pp. 63–4)

We believe that Loury's argument represents an important new way to understand stigma. Whereas statistical discrimination refers to self-confirmed beliefs, racial stigma in the Loury sense refers to unfalsified beliefs. Members of society face the problem that their experiences do not provide rich enough data to overrule prejudices. In our view, however, we need not equate prejudices with initial beliefs. Rather, we see prejudices as a way in which agents respond to information limitations. Given initial beliefs and experiences, racial stigma occurs, in our view, when an actor ascribes the least favourable set of characteristics to members of an outgroup that are consistent with the actor's experiences. From this vantage point, stigma is a form of ambiguity aversion, in which the history of racial oppression in the United States means that, in dealing with blacks, whites assume the worst. We recognize that Loury might not accept this interpretation of his views, but we believe that this interpretation is a useful way to understand and extend his thinking.

Concluding comments

While stigma is often invoked as an explanation, its use is often imprecise. In theoretical models, stigma is often conflated with a generic form of social interactions. For example, the stigma variable and associated assumed influence of the variable on preferences in Nechyba's (2001) analysis could just as easily have been labelled a conformity effect. Similar problems exist in empirical work. Pager's widely publicized (2003) field experiment to study the effects on black and white job applicants of a criminal background invokes stigma as a mechanism, but it is unclear whether her findings are due to direct discrimination, statistical discrimination, racial stigma as argued by Loury, or something else. Bertrand and Mullainathan's (2004) analysis, finding that resumés with black-sounding names receive fewer interview requests than those with white-sounding names, is admirable in its refusal to mechanically invoke stigma as an explanation.

In our view, the imprecision of much work on stigma can be addressed in two ways. First, it seems important to return to Goffman's ideas about ascription and spoiled identity in order to avoid conflating stigma with the presence of some ill-defined social interaction. Second, in empirical work, it is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
important to avoid treating the presence of stigma as a primitive, and rather to model explicitly the mechanism by which stigma emerges, if we are serious about stigma as a structural determinant.

See Also

- ambiguity and ambiguity aversion
- black–white labour market inequality in the United States
- identity
- social interactions (empirics)
- social interactions (theory)
- social multipliers

Bibliography

Akerlof, G.A., Yellen, J. and Katz, M. 1996. An analysis of out-of-wedlock childbearing in the United States. *Quarterly Journal of Economics* 111, 277–317.

Allport, G.W. 1954. *The Nature of Prejudice*. Boston, MA: Addison-Wesley.

Arrow, K. 1972. Some mathematical models of race in the labor market. In *Racial Discrimination in Economic Life*, ed. A. Pascal. Lexington, MA: Lexington.

Arrow, K. 1973. The theory of discrimination. In *Discrimination in Labor Markets*, ed. O. Ashenfelter and A. Rees. Princeton, N.J.: Princeton University Press.

Becker, G. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Benabou, R. and Tirole, J. 2006. Identity, dignity and taboos: beliefs as assets. Unpublished, Princeton University. Available at http://www.princeton.edu/~rbenabou/identity%20december%20f.pdf, accessed 16 July 2008.

Bertrand, M. and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94, 991–1013.

Besley, T. and Coate, S. 1992. Understanding welfare stigma: taxpayer resentment and statistical discrimination. *Journal of Public Economics* 48, 165–83.

Blume, L. 2002. *Stigma and Social Control*. Unpublished, Institut für Höhere Studien. Available at: http://www.ihs.ac.at/index.php3?id=330&sr=298, accessed 24 July 2008.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Blume, L. 2005. Learning and statistical discrimination. *American Economic Review* 95, 118–21.

Blume, L. 2006. The dynamics of statistical discrimination. *Economic Journal* 116, F480–F498.

Chaudhuri, S. and Sethi, R. 2008. Statistical discrimination with peer effects. Can integration eliminate negative stereotypes? *Review of Economic Studies* 75, 579–96.

Charles, K. and Guryan, J. 2007. Prejudice and the economics of discrimination, NBER Working Paper No. 13661.

Coate, S. and Loury, G. 1993. Will affirmative-action policies eliminate negative stereotypes? *American Economic Review* 83, 1220–40.

Crandall, C.S. 1995. Do parents discriminate against their heavyweight daughters? *Personality and Social Psychology Bulletin* 21, 724–35.

Crocker, J., Major, B. and Steele, C. 1998. Social stigma. In *Handbook of Social Psychology*, vol. 2, ed. S. Fiske, D. Gilbert and G. Lindzey. Boston, MA: McGraw-Hill.

Daponte, B., Sanders, S. and Taylor, L. 1999. Why do low-income households not use food stamps? *Journal of Human Resources* 34, 612–28.

Darley, J.M. and Fazio, R.H. 1980. Expectancy confirmation processes in the social interaction sequence. *American Psychologist* 35, 867–81.

Goffman, E. 1963. *Stigma: Notes on the Management of Spoiled Identity*. New York: Simon and Schuster.

Kendall, T. and Tamura, R. 2008. Unmarried fertility, crime, and social stigma. Unpublished, Clemson University. Available at http://mpra.ub.uni-muenchen.de/8031/1/MPRA_paper_8031.pdf, accessed 16 July 2008.

Loury, G.C. 2002. *The Anatomy of Racial Inequality*. Cambridge, MA: Harvard University Press.

Major, B. and O'Brien, L.T. 2005. The social psychology of stigma. *Annual Review of Psychology* 56, 393–421.

Merton, R.K. 1947. The self-fulfilling prophecy. *Antioch Review* 8, 193–210.

Moffitt, R. 1983. An economic model of welfare stigma. *American Economic Review* 73, 5, 23–1035.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Myrdal, G. 1939. *Monetary Equilibrium*. London: William Hodge.

Myrdal, G. 1944. *An American Dilemma: The Negro Problem and Modern Democracy*. New York: Harper and Row.

Nechyba, T. 2001. Social approval values and AFDC: a reexamination of the illegitimacy rate. *Journal of Political Economy* 109, 637–72.

O'Flaherty, B. and Sethi, R. 2008. Racial stereotypes and robbery. *Journal of Economic Behavior and Organization*, forthcoming.

Price, G., Darity, W. and Headen, A. 2008. Does the stigma of slavery explain the maltreatment of blacks by whites? The case of lynchings. *Journal of Socio-Economics* 37, 1, 67–93.

Pager, D. 2003. The mark of a criminal record. *American Journal of Sociology* 108, 937–75.

Rasmusen, E. 1996. Stigma and self-fulfilling expectations of criminality. *Journal of Law and Economics* 39, 519–34.

Steele, C.M. 1992. Race and the schooling of black Americans. *Atlantic Monthly* April, 68–78.

Steele, C.M. and Aronson, J. 1995. Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology* 69, 797–811.

How to cite this article

Blume, Lawrence E. and Steven N. Durlauf. "stigma." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000539> doi:10.1057/9780230226203.1874

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stochastic adaptive dynamics

H. Peyton Young
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economic systems often involve large numbers of agents whose behaviour, and patterns of interaction, have stochastic components. The dynamic properties of such systems can be analysed using stochastic dynamical systems theory. A key feature of these processes is that their long-run behaviour often differs substantially from the behaviour of the deterministic process obtained by taking expectations of the random variables. Furthermore, unlike the deterministic dynamics, the theory yields sharp predictions about the probability of being in different equilibria independently of the initial conditions.

Article

Stochastic adaptive dynamics require analytical methods and solution concepts that differ in important ways from those used to study deterministic processes. Consider, for example, the notion of asymptotic stability: in a deterministic dynamical system, a state is locally asymptotically stable if any sufficiently small deviation from the original state is self-correcting. We can think of this as a first step toward analysing the effect of stochastic shocks; that is, a state is locally asymptotically stable if, after the impact of a small, one-time shock, the process evolves back to its original state.
This idea is not entirely satisfactory, however, because it treats shocks as if they were isolated events. Economic systems are usually composed of large numbers of interacting agents whose behaviour is *constantly* being buffeted by perturbations from various sources. These *persistent* shocks have substantially different effects from *one-time* shocks; in particular, persistent shocks can accumulate and tip the process out of the basin of attraction of an asymptotically stable state. Thus, in a stochastic setting, conventional notions of dynamic stability – including *evolutionarily stable strategies* – are inadequate to characterize the long-run behaviour of the process. Here we shall outline an alternative approach that is based on the theory of large deviations in Markov processes (Freidlin and Wentzell, 1984; Foster and Young, 1990; Young, 1993a).

Types of stochastic perturbations

Before introducing formal definitions, let us consider the various kinds of stochastic shocks to which a system of interacting agents may be exposed. First, there is the interaction process itself whereby agents randomly encounter other agents in the population. Second, the agents' behaviour will be *intentionally* stochastic if they are employing mixed strategies. Third, their behaviour may be *unintentionally* stochastic if their payoffs are subject to unobserved utility shocks. Fourth, mutation processes may cause one type of agent to change spontaneously into another type. Fifth, in and out-migration can introduce new behaviours into the population or extinguish existing ones. Sixth, the system may be hit by aggregate shocks that change the *distribution* of behaviours. This list is by no means exhaustive, but it does convey some sense of the range of stochastic influences that arise quite naturally in economic (and biological) contexts.

Stochastic stability

The early literature on evolutionary game dynamics tended to sidestep stochastic issues by appealing to the law of large numbers. The reasoning is that, when a population is large, random influences at the individual level will tend to average out, and the aggregate state variables will evolve according to the *expected* (hence deterministic) direction of motion.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

While this approximation may be reasonable in the short and medium run, however, it can be quite misleading when extrapolated over longer periods of time. The difficulty is that, even when the stochastic shocks have very small probability, their accumulation can have dramatic long-run effects that push the process far away from its deterministic trajectory. The key to analysing such processes is to observe that, when the aggregate stochastic effects are 'small' and the resulting process is ergodic, the long-run distribution will often be concentrated on a very small subset of states – possibly, in fact, on a *single* state. This leads to the idea of *stochastic stability*, a solution concept first proposed for general stochastic dynamical systems by Foster and Young (1990, p. 221): 'the stochastically stable set (SSS) is the set of states S such that, in the long run, it is nearly certain that the system lies within every open set containing S as the noise tends slowly to zero.' The analytical technique for computing these states relies on the theory of large deviations first developed for continuous-time processes by Freidlin and Wentzell (1984), and subsequently extended to general finite-state Markov chains by Young (1993a). It is in the latter form that the theory is usually applied in economic contexts.

An illustrative example

The following simple model illustrates the basic ideas. Consider a population of *n* agents who are playing the 'Stag Hunt' game:

	A	*B*
A	10, 10	0, 7
B	7, 0	7, 7

The *state* of the process at time *t* is the current number of agents playing *A*, which we shall denote by $z_t \in Z = \{0, 1, 2, \ldots, n\}$. Time is discrete. At the start of period *t*+1, one agent is chosen at random. Strategy A is a best response if $z_t \leq .7n$ and *B* is a best response if $z_t \geq .7n$. (We assume that the player includes herself in assessing the current distribution, which simplifies the computations.) With high probability, say $1-\varepsilon$, the agent chooses a best response to the current distribution of strategies; while with probability ε she chooses *A* or *B* at random (each with probability $\varepsilon/2$).

We can interpret such a departure from best response behaviour in various ways: it might be a form of experimentation, it might be a behavioural 'mutation', or it might simply be a form of ignorance – the agent may not know the current state. Whatever the explanation, the result is a *perturbed best response process* in which individuals choose (myopic) best responses to the current state with high probability and depart from best response behaviour with low probability.

This process is particularly easy to visualize because it is one-dimensional: the states can be viewed as points on a line, and in each period the process moves to the left by one step, to the right by one step, or stays put. Figure 1 illustrates the situation when the population consists of ten players.

Figure 1

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Transitions for the perturbed best response process in the Stag Hunt game and a population of ten agents. *Note*: Each vertex represents the number of agents playing action A at a given time. Solid arrows are transitions with high-probability, dotted arrows are transitions with low (order- ε) probability.

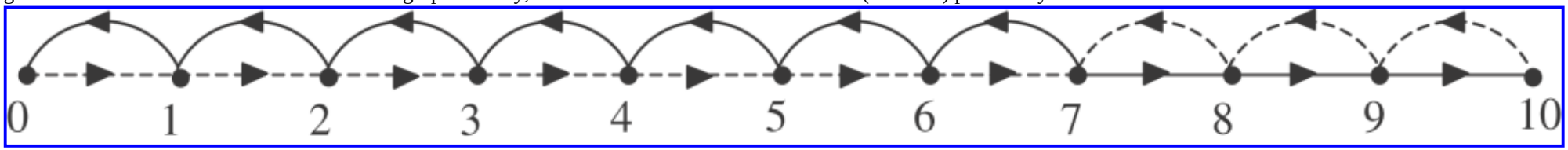

The transitions indicated by solid arrows have high probability and represent the direction of best response, that is, the main flow of the process. The dashed arrows go against the flow and have low probability, which is the same order of magnitude as ε . (The process can also loop by staying in a given state with positive probability; these loops are omitted from the figure to avoid clutter.)

In this example the transition probabilities are easy to compute. Consider any state *z* to the left of the critical value $z^*=7$. The process moves right if and only if one more agent plays *A*. This occurs if and only if an agent currently playing *B* is drawn (an event with probability $1 - z / 10$) and this agent mistakenly chooses *A* (an event with probability ε /2). In other words, if $z<7$ the probability of moving right is $R_z = (1 - z / 10)(\varepsilon / 2)$. Similarly, the probability of moving left is $L_z = (z / 10)(1 - \varepsilon / 2)$. The key point is that the right transitions have much smaller probability than the left transitions when ε is small. Exactly the reverse is true for those states $z > 7$. In this case the probability of moving right is $R_z = (1 - z / 10)(1 - \varepsilon / 2)$, whereas the probability of moving left is $L_z = (z / 10)(\varepsilon / 2)$. (At $z=7$ the process moves left with probability .15, moves right with probability .35, and stays put with probability .50.)

Computing the long-run distribution

Since this finite-state Markov chain is irreducible (each state is reachable from every other via a finite number of transitions), the process has a unique long-run distribution. That is, with probability 1, the relative frequency of being in any given state *z* equals some number μ_z *independently of the initial state*. Since the process is one-dimensional, the equations defining μ are particularly transparent, namely, it can be shown that for every $z<n$, $\mu_z R_z = \mu_{z+1} L_{z+1}$. This is known as the *detailed balance condition*. It has a simple interpretation: in the long run, the process transits from $z+1$ to *z* as often as it transits from *z* to $z+1$.

The solution in this case is very simple. Given any state *z*, consider the directed tree T_z consisting of all right transitions from states to the left of *z* and all left transitions from states to the right of *z*. This is called a *z-tree* (see Figure 2).

Figure 2

The unique 3-tree

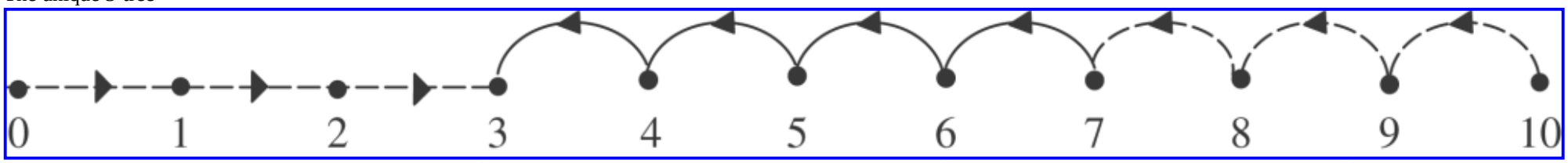

An elementary result in Markov chain theory says that, for one-dimensional chains, the long-run probability of being in state *z* is proportional to the product of the probabilities on the edges of T_z:

$$\mu_z \propto \prod_{y<z} R_y \prod_{y>z} L_y. \quad (1)$$

This is a special case of the *Markov chain tree theorem*, which expresses the stationary distribution of any finite chain in terms of the probabilities of its *z*-trees. (Versions of this result go back at least to Kirchhoff's work in the 1840s; see Haken, 1978, s. 4.8. Freidlin and Wentzell, 1984, use it to study large deviations in continuous-time Wiener processes.)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Formula (1) allows us to compute the order-of-magnitude probability of each state without worrying about its exact magnitude. Figure 2 shows, for example, that μ_3, the long-run probability of state z=3, must be proportional to ε^6, because the 3-tree has six dotted arrows, each of which has probability of order ε. Using this method we can easily compute the relative probabilities of each state.

Stochastic stability and equilibrium selection

This example illustrates a general property of adaptive processes with small persistent shocks. That is, the persistent shocks act as a *selection mechanism*, and the *selection strength increases the less likely the shocks are*. The reason is that the long-run distribution depends on the probability of escaping from various states, and the critical escape probabilities are *exponential* in ε. Figure 1 shows, for example, that the probability of all-B (the left endpoint) is larger by a factor of $1/\varepsilon$ than the probability of any other state, and it is larger by a factor of $1/\varepsilon^4$ than the probability of all-*A* (the right endpoint). It follows that, as ε approaches zero, the long-run distribution of the process is concentrated entirely on the all-*B* state. It is the unique stochastically stable state.

While stochastic stability is defined in terms of the limit as the perturbation probabilities go to zero, sharp selection can in fact occur when the probabilities are quite large. To illustrate, suppose that we take ε =.20 in the above example. This defines a very noisy adjustment process, but in fact the long-run distribution is still strongly biased in favour of the all-*B* state. It can be shown, in fact, that the all-*B* state is nearly 50 times as probable as the all-*A* state. (See Young, 1998b, ch. 4, for a general analysis of stochastic selection bias in one-dimensional evolutionary models.)

A noteworthy feature of this example is that the stochastically stable state (all-B) does not correspond to the Pareto optimal equilibrium of the game, but rather to the risk dominant equilibrium (Harsanyi and Selten, 1988). The connection between stochastic stability and risk dominance was first pointed out by Kandori, Mailath and Rob (1993). Essentially their result says that, in any symmetric 2×2 game with a uniform mutation process, the risk dominant equilibrium is stochastically stable provided the population is sufficiently large. The logic of this connection can be seen in the above example. In the pure best response process (ε =0) there are two absorbing states: all-*B* and all-*A*. The basin of attraction of all-*B* is the set of states to the left of the critical point, while the basin of attraction of the all-*A* is the set of states to the right of the critical point. The left basin is bigger than the right basin. To go from the left endpoint into the opposite basin therefore requires more 'uphill' motion than to go the other way around. In any symmetric 2×2 coordination game the risk dominant equilibrium is the one with the widest basin, hence it is stochastically stable under uniform stochastic shocks of the above type.

How general is this result? It depends in part on the nature of the shocks. On the one hand, if we change the probabilities of left and right transitions in an arbitrary way, then we can force any given state – including non-equilibrium states – to have the highest long-run probability; indeed this follows readily from formula (1). (See Bergin and Lipman, 1996.) On the other hand, there are many natural perturbations that do lead to the risk dominant equilibrium in 2×2 games. Consider the following class of perturbed best response dynamics. In state z, let $\Delta(z)$ be the expected payoff from playing A against the population minus the payoff from playing B against the population. Assume that in state z the probability of choosing *A* divided by the probability of choosing *B* is well-approximated by a function of form $e^{h(\Delta(z))/\beta}$ where $h(\Delta)$ is non-decreasing in Δ, strictly increasing at Δ=0, and skew-symmetric ($h(\Delta)=-h(-\Delta)$). The positive scalar β is a measure of the noise level. In this set-up, a state is *stochastically stable* if its long-run probability is bounded away from zero as $\beta \to 0$. Subject to some minor additional regularity assumptions, it can be shown that, in any symmetric 2×2 coordination game, if the population is large enough, the unique stochastically stable state is the one in which everyone plays the risk-dominant equilibrium (Blume, 2003).

Unfortunately, the connection between risk dominance and stochastic stability breaks down – even for uniform mutation rates – in games with more than two strategies per player (Young, 1993a). The difficulty stems from the fact that comparing 'basin sizes' works only in special situations. To determine the stochastically stable states in more general settings requires finding the path of least resistance – the path of greatest probability – from every absorbing set to every other absorbing set, and then constructing a rooted tree from these critical paths (Young, 1993a). (An absorbing set is a minimal set of states from which the *unperturbed* process cannot escape.) What makes the one-dimensional situation so special is that there are only two absorbing sets – the left endpoint and the right endpoint – and there is a unique directed path going from left to right and another unique path going from right to left. (For other situations in which the analysis can be simplified, see Ellison, 2000; Kandori and Rob, 1995.)

There are many games of economic importance in which this theory has powerful implications for equilibrium selection. In the non-cooperative Nash bargaining model, for example, the Nash bargaining solution is essentially the unique stochastically stable outcome (Young, 1993b). Different assumptions about the one-shot bargaining process lead instead to the selection of the Kalai–Smorodinsky solution (Young, 1998a; for further variations see Binmore, Samuelson and Young, 2003). In a standard oligopoly framework, marginal cost pricing turns out to be the stochastically stable solution (Vega-Redondo, 1997).

Speed of adjustment

One criticism that has been levelled at this approach is that it may take an exceedingly long time for the evolutionary process to reach the stochastically stable states when it starts

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

from somewhere else. The difficulty is that, when the shocks have very small probability, it takes a long time (in expectation) before enough of them accumulate to tip the process into the stochastically stable state(s). While this is correct in principle, the waiting time can be very sensitive to various modelling details. First, it depends on the size and probability of the shocks themselves. As we have already noted, the shocks need not be small for sharp selection to occur, in which case the waiting time need not be long either. (In the above example we found that an error rate of 20 per cent still selects the all-B state with high probability.) Second, the expected waiting time depends crucially on the *topology* of interaction. In the above example we assumed that each agent reacts to the distribution of actions in the whole population. If instead we suppose that people respond only to actions of those in their immediate geographic (or social) neighbourhood, the time to reach the stochastically stable state is greatly reduced (Ellison, 1993; Young, 1998b, ch. 6). Third, the waiting time is reduced if the stochastic perturbations are not independent, either because the agents act in a coordinated fashion, or because the utility shocks among agents are statistically correlated (Young, 1998b, ch. 9; Bowles, 2004).

Path dependence

The results discussed above rely on the assumption that the adaptive process is ergodic, that is, its long-run behaviour is almost surely independent of the initial state. Ergodicity holds if, for example, the number of states is finite, the transition probabilities are time-homogeneous, and there is a positive probability of transiting from any state to any other state within a finite number of periods. One way in which these conditions may fail is that the weight of history grows indefinitely. Consider, for example, a two-person game G together with a population of potential row players and another population of potential column players. Assume that an initial history of plays is given. In each period, one row player and one column player are drawn at random, and each of them chooses an ε-trembled best reply to the opposite population's previous actions (alternatively, to a random sample of fixed size drawn from the opponent's previous actions). This is a stochastic form of fictitious play (Fudenberg and Kreps, 1993; Kaniovski and Young, 1995). The *proportion* of agents playing each action evolves according to a stochastic difference equation in which the magnitude of the stochastic term decreases over time; in particular it decreases at the rate $1/t$. This type of process is not ergodic. It can be shown, in fact, that the long-run proportions converge almost surely either to a neighbourhood of all-*A* or to a neighbourhood of all-*B*, where the relative probabilities of these two events depend on the initial state (Kaniovski and Young, 1995). Processes of this type require substantially different techniques of analysis from the ergodic processes discussed earlier; see in particular Arthur, Ermoliev and Kaniovski (1987), Benaim and Hirsch (1999) and Hofbauer and Sandholm (2002).

Summary

The introduction of persistent random shocks into models with large numbers of interacting agents can be handled using methods from stochastic dynamical systems theory; moreover, there is virtually no limit on the dimensionality of the systems that can be analysed using these techniques. Such processes can exhibit path dependence if the weight of history is allowed to grow indefinitely. If instead past actions fade away or are forgotten, the presence of persistent random shocks makes the process ergodic and its long-run behaviour is often easier to analyse. An important feature of such ergodic models is that some equilibrium states are much more likely to occur in the long run than others, and this holds independently of the initial state. The length of time that it takes to reach such states from out-of-equilibrium conditions depends on key structural properties of the model, including the size and frequency of the stochastic shocks, the extent to which they are correlated among agents, and the network topology governing agents' interactions with one another.

See Also

- agent-based models
- evolutionary economics

Bibliography

Arthur, W.B., Ermoliev, Y. and Kaniovski, Y. 1987. Adaptive growth processes modelled by urn schemes. *Kybernetika* 6, 49–57; English trans in *Cybernetics* 23, 779–89.

Benaim, M. and Hirsch, M.W. 1999. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior* 29, 36–72.

Bergin, J. and Lipman, B.L. 1996. Evolution with state-dependent mutations. *Econometrica* 64, 943–56.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Binmore, K., Samuelson, L. and Young, H.P. 2003. Equilibrium selection in bargaining models. *Games and Economic Behavior* 45, 296–328.

Blume, L.E. 1995. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior* 11, 111–45.

Blume, L.E. 2003. How noise matters. *Games and Economic Behavior* 44, 251–71.

Bowles, S. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton, NJ: Princeton University Press.

Ellison, G. 1993. Learning, local interaction, and coordination. *Econometrica* 61, 1047–71.

Ellison, G. 2000. Basins of attraction, long run equilibria, and the speed of step-by-step evolution. *Review of Economic Studies* 67, 17–45.

Foster, D. and Young, H.P. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38, 219–32.

Freidlin, M. and Wentzell, A. 1984. *Random Perturbations of Dynamical Systems*. Berlin: Springer-Verlag.

Fudenberg, D. and Kreps, D. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5, 320–67.

Haken, H. 1978. *Synergetics*. Berlin: Springer-Verlag.

Harsanyi, J. and Selten, R. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge MA: MIT Press.

Hofbauer, J. and Sandholm, W. 2002. On the global convergence of stochastic fictitious play. *Econometrica* 70, 2265–94.

Kandori, M., Mailath, G. and Rob, R. 1993. Learning, mutation, and long-run equilibria in games. *Econometrica* 61, 29–56.

Kandori, M. and Rob, R. 1995. Evolution of equilibria in the long run: a general theory and applications. *Journal of Economic Theory* 65, 383–414.

Kaniovski, Y. and Young, H.P. 1995. Learning dynamics in games with stochastic perturbations. *Games and Economic Behavior* 11, 330–63.

Karlin, S. and Taylor, H.M. 1975. *A First Course in Stochastic Processes*. New York: Academic Press.

Vega-Redondo, F. 1997. The evolution of Walrasian behaviour. *Econometrica* 65, 375–84.

Young, H.P. 1993a. The evolution of conventions. *Econometrica* 61, 57–84.

Young, H.P. 1993b. An evolutionary model of bargaining. *Journal of Economic Theory* 59, 145–68.

Young, H.P. 1998a. Conventional contracts. *Review of Economic Studies* 65, 776–92.

Young, H.P. 1998b. *Individual Strategy and Social Structure*. Princeton. NJ: Princeton University Press.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Young, H. Peyton. "stochastic adaptive dynamics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_A000221> doi:10.1057/9780230226203.1621

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stochastic dominance

Russell Davidson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Stochastic dominance is a term which refers to a set of relations that may hold between a pair of distributions. A very common application of stochastic dominance is to the analysis of income distributions and income inequality, the main focus in this article. The concept can, however, be applied in many other domains, in particular financial economics, where the distributions considered are usually those of the random returns to various financial assets. In what follows, there are often clear analogies between things expressed in terms of income distributions and financial counterparts.

Keywords

cumulative distribution functions; headcount ratio; inequality; Kolmogorov–Smirnov test; Lorenz curve; Pareto principle; Pigou–Dalton principle of transfers; poverty gap; poverty indices; poverty lines; restricted stochastic dominance; separability; social welfare function; statistical inference; stochastic dominance

Article

In order to determine whether a relation of stochastic dominance holds between two distributions, the distributions are first characterized by their *cumulative distribution functions* (CDFs). For a given set of incomes, the value of the CDF at income y is the proportion of incomes in the set that are no greater than y. In the context of a random variable Y, the value of the CDF of the distribution of Y at y is the probability that Y should be no greater than y.

Suppose that we consider two distributions A and B, characterized respectively by CDFs F_A and F_B. Then distribution B dominates distribution A stochastically at first order if, for any argument y, $F_A(y) \geq F_B(y)$. This definition often looks as though it is the wrong way round, but a moment's reflection shows that it is correct as stated. If y denotes an income level, then the inequality in the definition means that the proportion of individuals in distribution A with incomes no greater than y is no smaller than the proportion of such individuals in B. In other words, there is at least as high a proportion of poor people in A as in B, if poverty means an income smaller than y. If B dominates A at first order, then, whatever poverty line we may choose, there is always more poverty in A than in B, which is why we say that A is the dominated distribution.

Higher orders of stochastic dominance can also be defined. To this end, we define repeated integrals of the CDF of each distribution. Formally, we define a sequence of functions by the recursive definition

$$D^1(y) = F(y),\ D^{s+1}(y) = \int_0^y D^s(z)\,dz, \text{ for } s = 1, 2, 3, \ldots .$$

Thus the function D^1 is the CDF of the distribution under study, $D^2(y)$ is the integral of D^1 from 0 to y, $D^3(y)$ is the integral of D^2 from 0 to y, and so on. By definition, distribution B

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dominates A at order s if $D_A^s(y) \geq D_B^s(y)$ for all arguments y. The lower limit of 0 is used for clarity of exposition; in general it is the lowest income in the pooled distributions. The definition makes it clear that first-order dominance implies dominance at all higher orders, and more generally that dominance at order s implies dominance at all orders higher than s. Since the implications go in only one direction, it follows that higher-order dominance is a weaker condition than lower-order dominance. I will give a more detailed interpretation of the functions D^s shortly in the context of poverty indices.

In theoretical arguments, it is sometimes desirable to distinguish weak from strong stochastic dominance. The above definitions are of weak dominance. For strong dominance, it is required that the inequality should be strict for at least one value of the argument y. In empirical investigations, the distinction is of no interest, since no statistical test can detect a significant difference between weak and strong inequalities.

Some applications make use of the concept of *restricted stochastic dominance*. This means that the relevant inequality is required to hold over some restricted range of the argument y rather than for all possible values. In empirical work, it is often only restricted dominance that can usefully be studied, since with continuous distributions there are usually too few data in the tails of the distributions for statistically significant conclusions to be drawn. Again, for measures of poverty, it is only dominance over the range of incomes up to the poverty line that is of interest.

Relation between stochastic dominance and welfare

When studying either income inequality or poverty, one is automatically in a normative context. Most modern studies make explicit or implicit use of a *social welfare function* (SWF). In a paper by Blackorby and Donaldson (1980) (henceforth BD), various ethically desirable criteria are developed and the sorts of SWF that respect these criteria are characterized.

One of these criteria is the *anonymity* of individuals. If we take all the worldly goods of a rich man and give them to a poor man, and then give the few worldly goods of the poor man to the rich man, then social welfare should be unchanged. Formally, a SWF that respects this requirement is symmetric with respect to its arguments, which are the incomes of the members of society.

Another requirement is the *Pareto principle*. According to it, we should rank situation B better than situation A if at least one individual is better off in B than in A, and no one is worse off. In order for a SWF to respect the Pareto principle, it must be increasing in all its arguments.

Another requirement, for measures of poverty only, is that a poverty index should not depend at all on the incomes of the non-poor. BD show that this implies a separability condition on the SWF. If in addition we require that the poverty index should be defined for arbitrary poverty lines, then the separability condition becomes a requirement of additive separability. The SWF can therefore be written as

$$W(y_1, \ldots, y_N) = \sum_{i=1}^{N} u(y_i), \tag{1}$$

where the ‘utility’ function u is increasing in its argument. Alternatively, the SWF can be any increasing transform of the function W. In all cases, the SWF is symmetric and increasing in its arguments, and so satisfies BD's ethical criteria.

It can be seen that first-order stochastic dominance of A by B means that B has higher social welfare than A for all SWFs of the form (1). This can be shown by a simple integration by parts, under the assumption that the function u is differentiable. In fact, this dominance is also a necessary condition for B to have higher welfare than A for all SWFs of the form (1).

It follows from the above argument that, if we use first-order stochastic dominance as a criterion for ranking distributions, then we need not restrict attention to a specific SWF, since any SWF of the form (1) gives the same ranking if one distribution dominates the other at first order.

A more restricted class of SWFs is given by functions of the form (1) where we impose the additional restriction that the second derivative of u is negative. It turns out that all the SWFs of this more restricted class give a unanimous ranking of two distributions if one dominates the other at second order. This sort of result can be extended to higher orders of dominance. As the dominance condition becomes progressively weaker, the class of SWFs that give unanimous rankings becomes progressively smaller, subject to more and more restrictive conditions on the function u and its derivatives.

Relation between stochastic dominance and poverty

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The so-called *headcount ratio* is sometimes used as a measure of the amount of poverty in a given income distribution. This ratio is the proportion of individuals in the distribution with incomes below, or equal to, the poverty line. If this line is denoted by *z*, then the headcount ratio is the value of the CDF at *z*. If we have two populations, *A* and *B*, characterized by two CDFs, F_A and F_B, then, for poverty line *z*, the headcount ratio is higher in *A* than in *B* if and only if $F_A(z) > F_B(z)$. If the inequality $F_A(y) > F_B(y)$ holds for all values of *y* up to *z*, then we have restricted first-order stochastic dominance up to *z*.

Corresponding to any income *y* less than *z*, we define the *poverty gap* as $z - y$. When we restrict attention to the welfare of people with incomes less than *z*, it is convenient to use a function π that measures the disutility of the poverty gap, rather than a function *u* of the sort used in a SWF. Thus we have a class of *poverty indices*, defined as follows:

$$\Pi(z) = \int_0^z \pi(z - y)\, dF(y),$$

If $\pi' > 0$, which means that the disutility of poverty increases with the poverty gap, it can be shown that, for all poverty indices of the above form, there is more poverty in *A* than in *B* if *B* dominates *A* at first order over the range of incomes less than or equal to *z*.

Earlier, a sequence of functions D^s was introduced, these functions being repeated integrals of the CDF. A useful explicit representation of these functions is given by the formula

$$D^s(z) = \frac{1}{(s-1)!}\int_0^z (z - y)^{s-1} dF(y). \quad (2)$$

The formula clearly holds for $s = 1$, if we remember that $0! = 1$. It is not hard to show by induction that it also holds for integers greater than 1.

For $s = 2$, the formula becomes

$$D^2(z) = \int_0^z (z - y)\, dF(y),$$

from which we see that, for given *z*, $D^2(z)$ is the average poverty gap for poverty line *z*. If, for all $z \in [z_-, z_+]$, $D_A^2(z) > D_B^2(z)$, then it follows that the average poverty gap is greater in *A* than in *B* for all poverty lines in the interval $[z_-, z_+]$. But this condition is just restricted second-order stochastic dominance of *A* by *B* over that interval.

As with welfare functions, this result can be extended. By progressively restricting the admissible class of poverty indices, in particular by imposing signs on the derivatives of π, it can be seen that all poverty indices in these more restricted classes unanimously see more poverty in *A* than in *B* if there is a progressively higher order of stochastic dominance; see Davidson and Duclos (2000) for more details. An essential reference on poverty measurement is Atkinson (1987), in which the axiomatic approach of BD is extended to poverty measurement. See also three papers by Foster and Shorrocks (1988a; 1988b; 1988c).

Relation between stochastic dominance and inequality

If a richer person in distribution *A* transfers some income to a poorer person in such a way that the richer person stays richer after the transfer, the post-transfer distribution *B* stochastically dominates *A* at second order. The Pigou–Dalton principle of transfers says that ‘Robin-Hood’ transfers of the sort described should improve welfare. But it is easy to see that distribution *B* does not dominate *A* at first order, and indeed this is right and proper according to the Pareto principle, since the richer person is worse off after the transfer.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This example shows that, when we discuss inequality, we are not talking about the same thing as welfare. Any reasonable measure of inequality must declare that there is no inequality if everyone has the same income, even if everyone is in abject poverty.

The classical tool for studying inequality is the *Lorenz curve*. For any proportion *p* between zero and one, the ordinate of the corresponding point on the Lorenz curve for a given income distribution is the proportion of total income that accrues to the first 100*p* per cent of people when they are sorted in order of increasing income. By construction, the Lorenz curve fits into the unit square, lies below the 45-degree line that is the diagonal of that square, and is (weakly) convex. Figure 1 displays a typical Lorenz curve.

Figure 1

A typical Lorenz curve

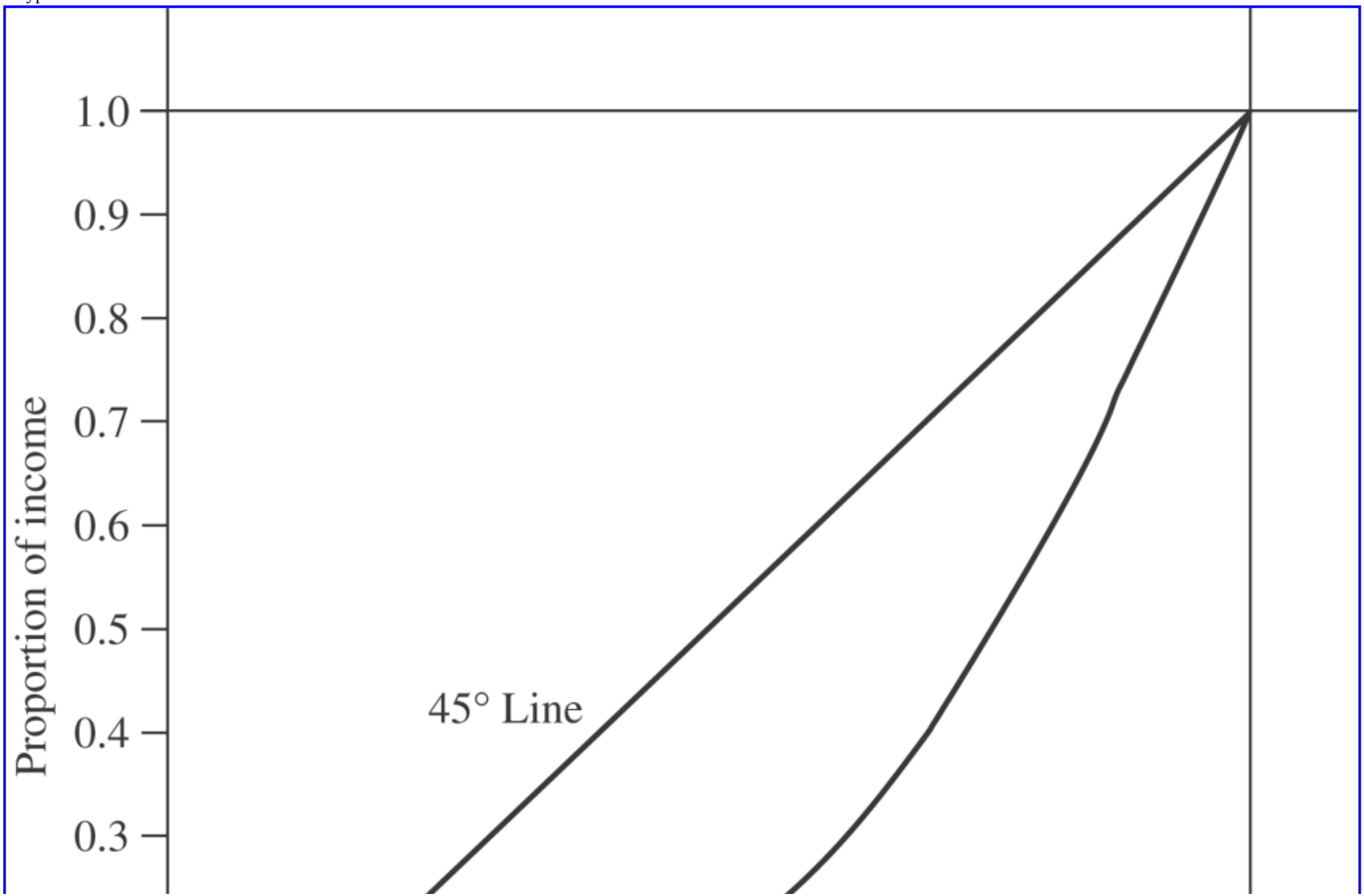

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

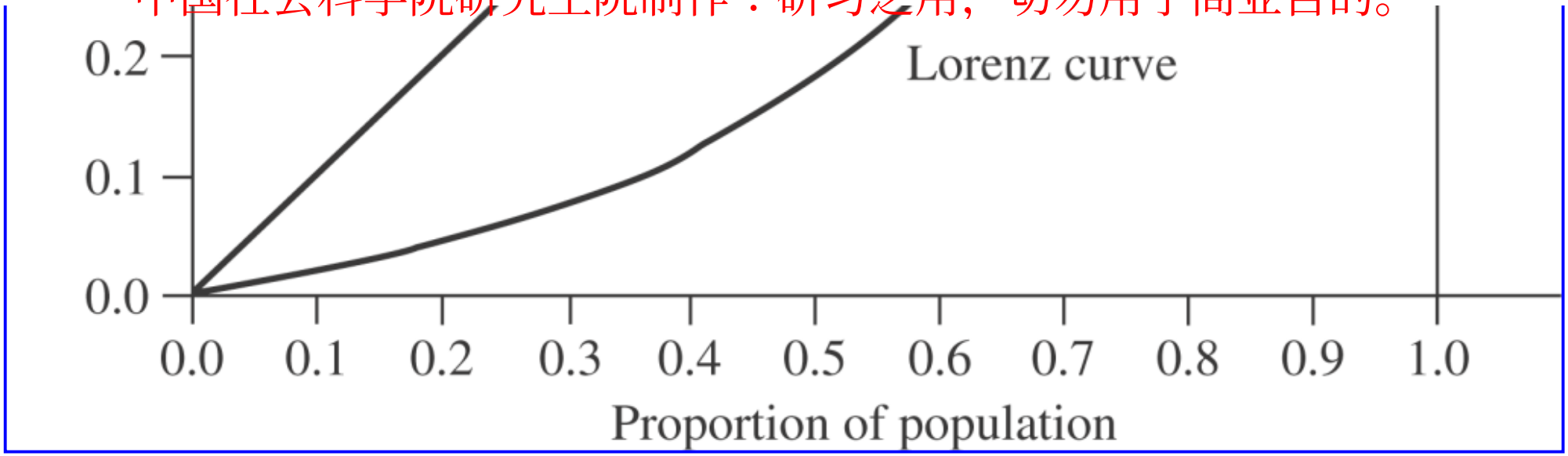

A distribution *B* is said to *Lorenz dominate* another distribution *A* if the Lorenz curve of *B* lies everywhere above that of *A*. We then say that there is less inequality in *B* than in *A*. But this comparison of *A* and *B* is not a welfare comparison, and, in particular, does not allow a comparison of poverty. This defect is remedied by the concept of *generalized Lorenz dominance*, based on the generalized Lorenz curve introduced by Shorrocks (1983). The ordinates of this curve are the Lorenz ordinates multiplied by the average income of the distribution. It turns out that generalized Lorenz dominance is the same thing as second-order stochastic dominance. Either one of these concepts implicitly mixes notions of welfare and inequality, as shown by the fact that the function *u* in a SWF of form (1) that respects second-order dominance has a negative second derivative, which implies diminishing marginal (social) utility of income. The discussion of the previous section shows that higher-order dominance criteria put more and more weight on the welfare of the poorest members of society.

Graphical representation and quantiles

Consider the setup in Figure 2, where the CDFs of two distributions *A* and *B* are plotted. The functions D^2 used for second-order dominance comparisons can be evaluated for a given argument, like z_1 in the figure, as the areas beneath the CDFs, by the usual geometric interpretation of the Riemann integral. We see that distribution *B* dominates *A* at second order because, although the CDFs cross, the areas between them are such that the condition for second-order dominance is always satisfied. Thus the vertical line *MN* marks off a large positive area between the graphs of the two CDFs up to the point at which they cross, and thereafter a small negative area bounded on the right by *MN*.

Figure 2
Generalized Lorenz and second-order dominance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

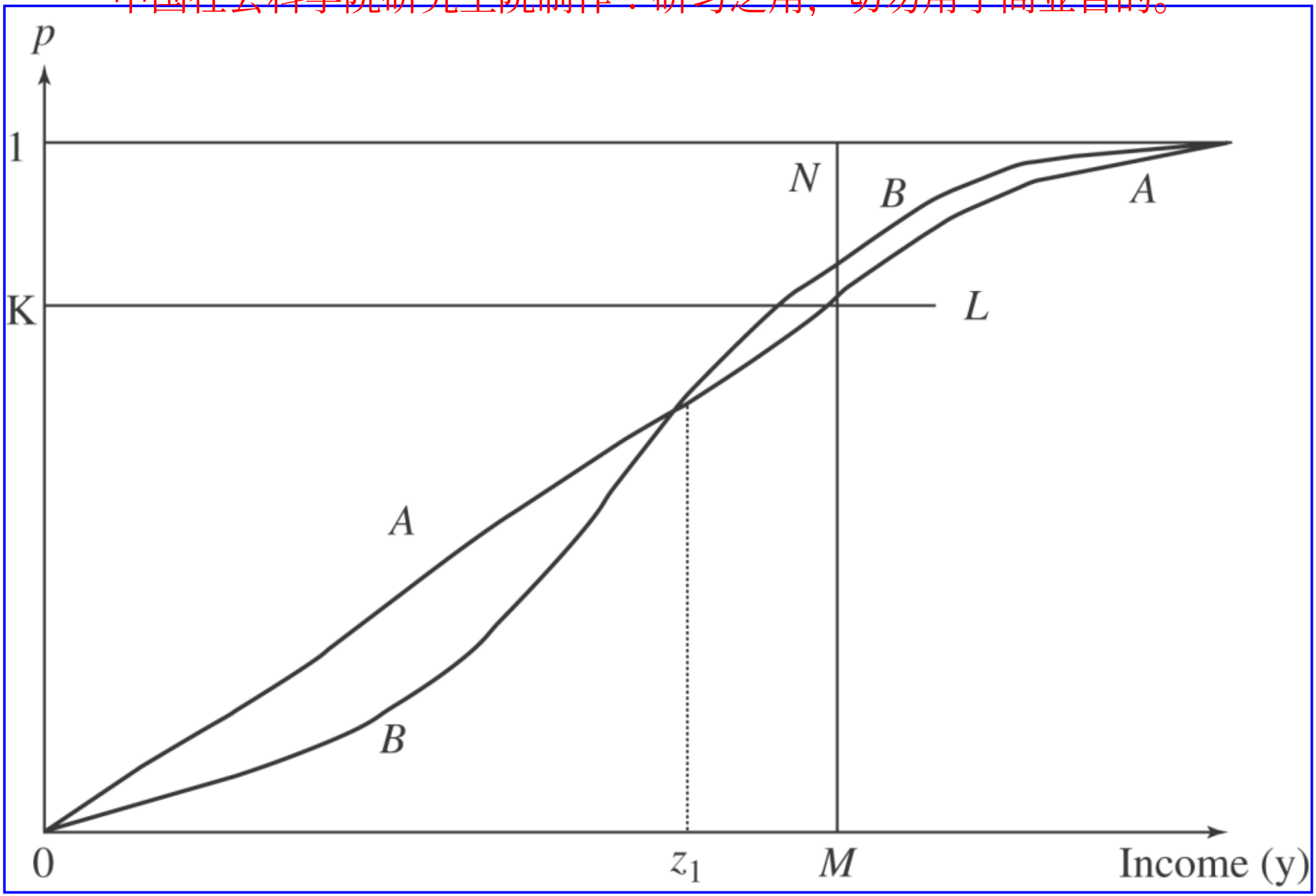

For generalized Lorenz dominance, it can be shown that what must be non-negative everywhere is the area between the two curves, bounded not by a vertical line like *MN* but rather by a horizontal line like *KL*. This area is the difference between the areas under two *quantile functions*, a quantile function being by definition the inverse of the CDF. Although it is tedious to demonstrate it algebraically, it is intuitively clear that, if the areas bounded on the right by vertical lines like *MN* are always positive, then so are the areas bounded above by horizontal lines like *KL*. This is why generalized Lorenz dominance and second-order stochastic dominance are equivalent conditions. The whole theory of stochastic dominance can be developed using quantiles rather than incomes; this is called a *p-approach*. Such approaches are used to advantage in Jenkins and Lambert (1997; 1998), Shorrocks (1998), and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

also Spencer and Fisher (1992).

Another thing that emerges clearly from Figure 2 is that the threshold income z_1 up to which first-order stochastic dominance holds is always smaller than the threshold z_2 up to which we have second-order dominance. In the figure, we have second-order dominance everywhere, and so we can set z_2 equal to the highest income in either distribution. More generally, we can define a threshold z_s as the greatest income up to which we have dominance at order *s*. The z_s constitute an increasing sequence.

A result shown in Davidson and Duclos (2000) is that, if the distribution *B* dominates *A* at first-order over a range [0,*z*], with $z > 0$, then, no matter what happens for incomes above *z*, there is always *some* order *s* such that *B* dominates *A* at order *s* over the full range of the two distributions, provided only that that range is finite.

Estimation and inference

Suppose that we have a random sample of *N* independent observations y_i, $i = 1, \ldots, N$, from a population. Then it follows from (2) that a natural estimator of $D^s(z)$ (for a non-stochastic *z*) is

$$\hat{D}^s(z) = \frac{1}{(s-1)!}\int_0^z (z-y)^{s-1}\, d\hat{F}(y) = \frac{1}{n(s-1)!}\sum_{i=1}^{N}(z-y_i)^{s-1}\mathrm{I}(y_i \le z), \tag{3}$$

where $\hat{F}$ denotes the empirical distribution function of the sample and I(.)is an indicator function equal to 1 when its argument is true and 0 otherwise. For *s*=1, the formula (3) estimates the population CDF by the empirical distribution function. For arbitrary *s*, it has the convenient property of being a sum of independent and identically distributed (IID) variables, which makes it easy to show that (3) is consistent and asymptotically normal. The asymptotic variance is also easy to estimate in a distribution-free manner, by which is meant that no parametric assumptions need be made about the distributions under study.

When two distributions are compared for stochastic dominance, two kinds of situations typically arise. The first is when there are two independent populations, with random samples from each. The other arises when we have independent paired drawings from the same population. For instance, one variable could be before-tax income, and the other after-tax income for the same individual. Explicit expressions for the asymptotic variance of the difference between the estimates of $D^s(z)$ for the case of independent samples were given as early as 1989 in an unpublished thesis (Chow, 1989). The sampling distribution of a related estimator for poverty indices and independent samples is found in Kakwani (1993), Bishop, Chow and Zheng (1995) and Rongve (1997). For a different approach to inference on stochastic dominance, see Anderson (1996). A comprehensive approach to inference on stochastic dominance is found in Davidson and Duclos (2000).

The approach proposed by McFadden (1989) is based on the supremum of the difference between the estimates (3) for two independent populations. For $s = 1$, this turns out to be a variant of the Kolmogorov–Smirnov test, with known properties. For higher values of *s*, although it is easy to compute the statistic, its asymptotic properties under the null are not analytically tractable. However, simulation-based methods can surmount this difficulty; see Barrett and Donald (2003).

A somewhat vexed question in testing for dominance is whether to test the null hypothesis of dominance or that of non-dominance. The latter has the advantage that, if the null is rejected, all that remains is dominance. More generally, the former approach rejects the null of dominance only when there is clear evidence against it, and the latter accepts the alternative of dominance only when there is clear evidence in its favour. The former approach is more common in the literature; see for instance Richmond (1982), Beach and Richmond (1985), Wolak (1989), and Bishop, Formby and Thistle (1992). The latter is discussed in an unpublished paper (Howes, 1993) and in Kaur, Prakasa Rao and Singh (1994).

See Also

- Dalton, Edward Hugh John Neale
- ethics and economics
- Gini ratio
- income mobility

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- inequality (measurement)
- Pigou, Arthur Cecil
- poverty
- social welfare function
- tax incidence

Bibliography

Anderson, G. 1996. Nonparametric tests of stochastic dominance in income distributions. *Econometrica* 64, 1183–93.

Atkinson, A. 1987. On the measurement of poverty, *Econometrica* 55, 749–64.

Barrett, G. and Donald, S. 2003. Consistent tests for stochastic dominance. *Econometrica* 71, 71–103.

Beach, C. and Richmond, J. 1985. Joint confidence intervals for income shares and Lorenz curves. *International Economic Review* 26, 439–50.

Bishop, J., Chow, K. and Zheng, B. 1995. Statistical inference and decomposable poverty measures. *Bulletin of Economic Research* 47, 329–40.

Bishop, J. Formby, J. and Thistle, P. 1992. Convergence of the South and non-South income distributions, 1969–1979. *American Economic Review* 82, 262–72.

Blackorby, C. and Donaldson, D. 1980. Ethical indices for the measurement of poverty. *Econometrica* 48, 1053–62.

Chow, K. 1989. Statistical inference for stochastic dominance: a distribution free approach. Ph.D. thesis, University of Alabama.

Davidson, R. and Duclos, J.-Y. 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* 68, 1435–64.

Foster, J. and Shorrocks, A. 1988a. Poverty orderings. *Econometrica* 56, 173–7.

Foster, J. and Shorrocks, A. 1988b. Poverty orderings and welfare dominance. *Social Choice and Welfare* 5, 179–98.

Foster, J. and Shorrocks, A. 1988c. Inequality and poverty orderings. *European Economic Review* 32, 654–62.

Howes, S. 1993. Asymptotic properties of four fundamental curves of distributional analysis. Unpublished paper, STICERD, London School of Economics.

Jenkins, S. and Lambert, P. 1997. Three ‘I's of poverty curves, with an analysis of UK poverty trends. *Oxford Economic Papers* 49, 317–27.

Jenkins, S. and Lambert, P. 1998. Three ‘I's of poverty curves and poverty dominance: TIPs for poverty analysis. *Research on Economic Inequality* 8, 39–56.

Kakwani, N. 1993. Statistical inference in the measurement of poverty. *Review of Economics and Statistics* 75, 632–39.

Kaur, A., Prakasa Rao, B. and Singh, H. 1994. Testing for second-order stochastic dominance of two distributions. *Econometric Theory* 10, 849–66.

McFadden, D. 1989. Testing for stochastic dominance. In *Studies in the Economics of Uncertainty*, ed. T. Fomby and T. Seo. New York: Springer-Verlag.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Richmond, J. 1982. A general method for constructing simultaneous confidence intervals. *Journal of the American Statistical Association* 77, 455–60.

Rongve, I. 1997. Statistical inference for poverty indices with fixed poverty lines. *Applied Economics* 29, 387–92.

Shorrocks, A. 1983. Ranking income distributions. *Economica* 50, 3–17.

Shorrocks, A. 1998. Deprivation profiles and deprivation indices. In *The Distribution of Household Welfare and Household Production*, ed. S. Jenkins et al. Cambridge: Cambridge University Press.

Spencer, B. and Fisher, S. 1992. On comparing distributions of poverty gaps. *Sankhya: The Indian Journal of Statistics Series B* 54, 114–26.

Wolak, F. 1989. Testing Inequality constraints in linear econometric models, *Journal of Econometrics* 41, 205–35.

How to cite this article

Davidson, Russell. "stochastic dominance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000266> doi:10.1057/9780230226203.1622

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stochastic frontier models

Hung-Jen Wang
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The stochastic frontier model was first proposed in the context of production function estimation to account for the effect of technical inefficiency. The inefficiency causes actual output to fall below the potential level (that is, the production frontier) and also raises production cost above the minimum level (that is, the cost frontier). Recent applications of the model are found in many fields of study including labour, finance, and economic growth. In these applications, the observed outcome (of wages, investment, and so on) is modelled as being deviating from a frontier level in one direction owing to factors such as information asymmetry.

Keywords

aftermarkets; allocative inefficiency; convergence; copulas; cost functions; duality; financing constraints; fixed-effect panel estimators; labour market search models; likelihood functions; nonparametric estimation; production function estimation; production functions; semiparametric estimation; stochastic cost frontiers; stochastic frontier models; technical inefficiency; technological catch-up; truncated distributions

Article

The stochastic frontier model was first proposed by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) in the context of production function estimation. The model extends the classical production function estimation by allowing for the presence of technical inefficiency. The idea is that, although the production technology is common knowledge to a group of producers, efficiency in using that technology in the production process may vary by producers, with the degree of efficiency depending possibly on factors such as experience, management skills, and so on. Given the technology, fully efficient producers may realize the full potential of the technology and obtain the maximum possible output for given inputs, while less efficient producers see their output fall short of the maximum

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

possible level. Therefore, the underlying technology defines a *frontier* of production, and actual outputs observed in the data may fall below the frontier because of the presence of technical inefficiency.

A stochastic production frontier model can be specified as

$$\ln y_i = \ln y_i^* - u_i,\ u_i \geq 0, \tag{1}$$

$$\ln y_i^* = f(\mathbf{x}_i; \boldsymbol{\beta}) + v_i, \tag{2}$$

where y_i is the observed output of producer i, y_i^* is the potential output which is subject to a zero-mean random error v_i, $\mathbf{x}_i$ and β are vectors of inputs and the corresponding coefficients, respectively, and $u_i \geq 0$ is the effect of technical inefficiency. Equation (2) defines the *stochastic frontier* of the production function; it is stochastic because of v_i. Given that $u_i \geq 0$, observed log of output ($\ln y_i$) is bounded below the frontier. The value of $100*u_i$ is the percentage by which output can be increased using the same inputs if production is fully efficient. The model without u_i reduces to the classical specification of a production function.

A popular empirical strategy in estimating the above model is to impose distributional assumptions on u_i and v_i, from which a likelihood function can be derived and estimated. For instance, one may assume that

$$v_i \sim N(0, \sigma^2), \tag{3}$$

$$u_i \sim N^+(\mu, \sigma_u^2), \tag{4}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $N^{+}(\cdot)$ indicates the positive truncation of a normal distribution. The positive truncation gives non-negative values of u_i and hence ensures that firms are constrained by the technology frontier. By making μ and/or σ_u^2 functions of observables (such as ages and years of schooling), one can model the determinants of inefficiency.

The distribution assumption of (4) encompasses many of the models in the literature as special cases. For instance, the half-normal distribution of u_i proposed by Aigner, Lovell and Schmidt (1977) is obtained by restricting $\mu = 0$ and σ_u^2 to be a constant. The half-normal density has a mode at 0 which implies that the majority of the producers are clustered near full efficiency level. The assumption may be unnecessarily restrictive, in particular for industries in which certain degree of inefficiency is expected for the producers. The assumption is relaxed by having $\mu \neq 0$ to allow the mode to depart from 0. Since limited theory is available in guiding the choice of u_i's distribution, various distribution assumptions are explored in the literature for their flexibility in shaping the distribution (for example, the Gamma distribution of Greene, 1980) and/or for checking the robustness of estimation results.

It is often of great empirical interest to estimate the degree of inefficiency (u_i) for each producer (observation). The observation-level estimates are obtained using the estimator $E(u_i|v_i - u_i)$ proposed by Jondrow et al. (1982). The value of $100 \times E(u_i|v_i - u_i)$ gives the percentage by which output is increased if production is fully efficient. Similarly, an efficiency index is estimated using $E(\exp(-u_i)|v_i - u_i)$ (Battese and Coelli, 1988). The estimated value gives the actual output as a share of potential output, and the value is bounded between 0 and 1. A likelihood ratio test of the null hypothesis that u_i equals 0 can be performed to test for the presence of inefficiency. It amounts to testing the model against its OLS counterpart (the model without u_i). The distribution of the test statistic, however, is non-standard, because the value of 0 is on the boundary of u_i's support. Alternatively, given that an obvious difference between v_i and $v_i - u_i$ is the skewness of the latter, Schmidt and Lin (1984) suggest a simple test based on the sample skewness of the OLS residuals. If $v_i - u_i$ is the correct specification, the residuals would skew to the left and the null hypothesis of a normal error would be rejected.

If panel data is available, the model may be written as (for the ease of illustration, assume that the deterministic part of the frontier function is linear):

$$\ln y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} - u_i,\ u_i \geq 0, \tag{5}$$

where α is a constant. One may impose distributional assumptions on v_{it} and u_i to derive the likelihood function of the model. Alternatively, a distribution-free approach suggested by Schmidt and Sickles (1984) is available. In this approach, one defines $\alpha_i = \alpha - u_i$ and assumes that the u_i is an individual-specific parameter. With the definition of α_i substituted into (5), the model is estimated by standard

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

fixed-effect panel estimators which yield consistent estimates of α_i for a large T. Since $\alpha_i = \alpha - u_i$ and $u_i \geq 0$, one then recovers the estimated values of α and u_i using the normalization equations of $\hat{\alpha} = max\{\hat{\alpha}_i\}$ and $\hat{u}_i = \hat{\alpha} - \hat{\alpha}_i$. This normalization procedure amounts to counting the most efficient firm in the sample as 100 per cent efficient.

By duality, technical inefficiency in the production also leads to a higher cost of production. The estimated cost of technical inefficiency often has important policy implications, and estimation can be done using a stochastic cost frontier model in a cost minimization framework. The model specification is:

$$\ln C_i = \ln C_i^* + u_i', \; u_i' \geq 0, \tag{6}$$

$$\ln C_i^* = g(\mathbf{w}_i, y_i; \gamma) + v_i', \tag{7}$$

where C_i is the observed cost of producer i, C_i^* is the efficient level of cost which is subject to a zero-mean random error v_i', $\mathbf{w}_i$ is the vector of input prices, γ is the vector of coefficients, and $u_i' \geq 0$ is the effect of inefficiency on the cost of production. Equation (7) defines the stochastic cost frontier, and the observed cost lies above the frontier. The value of $100 * u_i'$ measures the extra cost as a percentage of the minimum cost. Econometric analysis of (6) and (7) is similar to that of the production function model. A notable difference is that the cost model's OLS residuals skew to the right if inefficiency presents in the data.

An advantage of a cost function approach over a production function approach is that the issue of *allocative inefficiency* can be addressed in addition to the technical inefficiency. Allocative inefficiency refers to the use of improper input combinations, that is, the marginal rate of technical substitution between inputs departs from the input price ratio. The improper input mix increases the cost of production, and the effect is not the same as technical inefficiency. Because the analysis of allocative inefficiency requires information of input prices, it is usually carried out in a cost minimization framework. To jointly estimate both technical and allocative inefficiency, Schmidt and Lovell (1979) provide the solution technique for a cost system in which the production technology is Cobb-Douglas. Kumbhakar (1997) presents a theoretical solution for a model with a translog cost function, and the difficulty in the empirical implementation of this model is discussed and resolved in Kumbhakar and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Wang (2006) and Kumbhakar and Tsionas (2005).

Although the stochastic frontier model is most often applied to the estimation of production and cost functions, an increasing body of research has adopted the methodology to other fields of study. Hofler and Murphy (1992) apply this estimation approach to labour market search models. Due to the costs of search, observed wages tend to fall below the maximum offers that are available in the market; this shortfall is analogous to a technical inefficiency. Another application is found in the study of financing constraints on investment, where Wang (2003) models the frictionless level of investment as the frontier, and actual investment falls below the frontier because of financing constraints. This approach allows Wang to quantify the effect of financing constraints on investment (represent by u_i), which is infeasible with the conventional linear-regression approach. In an application to economic growth, Kumbhakar and Wang (2005) employ the stochastic frontier approach and model growth convergence as countries' movements towards the world production frontier. A country may fall short of producing the maximum possible output because of technical inefficiency, and the phenomenon of technological catch-up is observed if the country moves towards the world production frontier over time. By making u_i a function of time and other macro variables, Kumbhakar and Wang test and confirm the convergence hypothesis. The stochastic frontier model also finds applications in finance. For example, a long-standing issue in the finance literature is whether the initial public offering (IPO) underpricing – the phenomenon whereby the initial offer price of an IPO is below the closing day's bid price – is deliberate on the firm's part or not. Hunt-McCool, Koh and Francis (1996) adopt the stochastic frontier model to investigate the issue, in which u_i measures the difference between the maximum predicted offer price and the actual offer price. The advantage of the stochastic frontier model in this application is that it can be used to measure the level of deliberate underpricing in the pre-market without using aftermarket information. Kumbhakar and Lovell (2000) offer an excellent review of the existing models in the stochastic frontier literature. The more recent developments in the literature aim to make the model more flexible. For instance, correlations between v_i and u_i are made possible through copula functions. If time series or panel data are available, then it is possible to make u_t or u_{it} serially correlated. Semiparametric and nonparametric estimation methods are also adopted to estimate the frontier of the production function (for example, $f(x_i; \beta)$) and the frontier of the cost function (for example, $g(w_i, y_i; \gamma)$) so that they are not restricted to specific functional forms.

See Also

- X-efficiency

Bibliography

Aigner, D., Lovell, C.A.K. and Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6, 21–37.

Battese, G.E. and Coelli, T.J. 1988. Prediction of firm-level technical efficiencies with a generalized

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
frontier production function and panel data. *Journal of Econometrics* 38, 387–99.

Greene, W.H. 1980. On the estimation of a flexible frontier production model. *Journal of Econometrics* 13, 101–15.

Hofler, R.A. and Murphy, K.J. 1992. Underpaid and overworked: measuring the effect of imperfect information on wages. *Economic Inquiry* 30, 511–29.

Hunt-McCool, J., Koh, S.C. and Francis, B.B. 1996. Testing for deliberate underpricing in the IPO premarket: a stochastic frontier approach. *Review of Financial Studies* 9, 1251–69.

Jondrow, J., Lovell, C.A.K., Materov, I.S. and Schmidt, P. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19, 233–8.

Kumbhakar, S.C. 1997. Modeling allocative inefficiency in a translog cost function and cost share equations: an exact relationship. *Journal of Econometrics* 76, 351–6.

Kumbhakar, S.C. and Lovell, C.A.K. 2000. *Stochastic Frontier Analysis*. Cambridge, UK and New York: Cambridge University Press.

Kumbhakar, S.C. and Tsionas, E.G. 2005. The joint measurement of technical and allocative inefficiencies: an application of Bayesian inference in nonlinear random-effects models. *Journal of American Statistical Association* 100, 736–47.

Kumbhakar, S.C. and Wang, H.-J. 2005. Estimation of growth convergence using a stochastic production frontier approach. *Economics Letters* 3, 300–5.

Kumbhakar, S.C. and Wang, H.-J. 2006. Estimation of technical and allocative inefficiency: a primal system approach. *Journal of Econometrics* 134, 419–40.

Meeusen, W. and van den Broeck, J. 1977. Technical efficiency and dimension of the firm: some results on the use of frontier production functions. *Empirical Economics* 2, 109–22.

Schmidt, P. and Lin, T.F. 1984. Simple tests of alternative specifications in stochastic frontier models. *Journal of Econometrics* 24, 349–61.

Schmidt, P. and Lovell, C.A.K. 1979. Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers. *Journal of Econometrics* 9, 343–66.

Schmidt, P. and Sickles, R.C. 1984. Production frontiers and panel data. *Journal of Business and Economic Statistics* 2, 367–74.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Wang, H.-J. 2003. A stochastic frontier analysis of financing constraints on investment: the case of financial liberalization in Taiwan. *Journal of Business and Economic Statistics* 3, 406–19.

How to cite this article

Wang, Hung-Jen. "stochastic frontier models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000443>
doi:10.1057/9780230226203.1623

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Onlir

stochastic optimal control

A.G. Malliaris
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The purpose of this article is, first, to state the stochastic optimal control problem and, second, to explain how it differs from deterministic optimal control and why that difference is crucial in economic problems. The article presents intuitively the methodology of optimal stochastic control and provides an illustration from optimal stochastic economic growth as an application of this mathematical technique in economics.

Keywords

applied control; Bellman, R.; continuous time models; deterministic optimal control; discrete time models; dynamic economic models; dynamic programming; Hamilton–Jacobi–Bellman equation; principle of optimality; pure randomness; stochastic optimal control; Taylor's th; uncertainty; white noise; Wiener process

Article

In the long history of mathematics, stochastic optimal control is a rather recent development. Using Bellman's principle of optimality along with measure-theoretic and functional-analytic methods, several mathematicians such as H. Kushner, W. Fleming, R. Rishel, W.M. Wonham and J.M. Bismut, among many others, made important contributions to this new area of mathematical research during the 1960s and early 1970s. For a complete mathematical exposition of the continuous time case, see Fleming and Rishel (1975), and for the discrete time case, see Bertsekas and Shreve (1978).
The assimilation of the mathematical methods of stochastic optimal control by economists was very rapid. Several economic papers started to appear in the early 1970s, among which we mention Merton (1971) on consumption and portfolio rules using continuous time methodology and Brock and Mirman (1972) on optimal economic growth under uncertainty using discrete time techniques. Since then, stochastic optimal control methods have been applied in most major areas of economics such as price theory, macroeconomics, monetary economics and financial economics. Chang (2004) offers a rigorous mathematical exposition of stochastic control methods and numerous examples from economics.
In this article we (a) state the stochastic optimal control problem, (b) explain how it differs from deterministic optimal control and why that difference is crucial in economic problems, (c) present intuitively the methodology of optimal stochastic control and, finally, (d) give an illustration from optimal stochastic economic growth.
Consider the problem:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$J[k(t), t, \infty] = \max E_t \int_t^{\infty} e^{-\rho s} u[k(s), v(s)] ds$$
(1)

subject to the conditions

$$dk(t) = T[k(t), v(t)]dt + \sigma[k(t), v(t)]dZ(t), k(t) \text{given}.$$
(2)

Here $v = v(t) = v(t, \omega)$ is the control random variable, $k=k(t)=k(t, \omega)$ is the state random variable, $\rho \geq 0$ is the discount on future utility, u denotes a utility function, T is the drift component of technology, σ is the diffusion component, dZ is a Wiener process and E_t denotes expectation conditioned on $k(t)$ and $v(t)$.

We note immediately that (1) and (2) generalize the deterministic optimal control by incorporating *uncertainty*. The modelling of economic uncertainty is achieved by allowing both the control and state variables to be random and more importantly by postulating that condition (2) is described by a stochastic differential equation of the Itō type.

In the problem described by (1) and (2), if $\sigma(k, v)=0$ and if k and v are assumed to be real variables instead of random, then (1) and (2) reduce to the special case of deterministic optimal control. Thus, the stochastic optimal control problem differs from the deterministic optimal control in the sense that the former generalizes the latter, or equivalently, in the sense that the latter is a special case of the former. This is crucial mathematical difference.

For the economist, the generalization achieved from stochastic optimal control means that the analysis of dynamic economic models becomes more realistic. The economic theorist who uses stochastic optimal control in positive or in welfare economics, in free market or centrally planned economies allows for randomness. Measurement errors, omission of important variables, non-exact relationships, incomplete theories and other methodological complexities are modelled in stochastic optimal control by allowing the control and state variables to be random, and also, by incorporating *pure randomness* through the white noise factor d$Z(t)$. The random variable d$Z(t)$ describes increments in the Wiener process $\{Z(t), t \geq 0\}$ that are independent and normally distributed with mean, E[dZ(t)]=0 and variance Var[dZ(t)]=dt.

In particular, eq. (2) is a significant economic generalization of the analogous equation in deterministic control. The reader may recall that in deterministic control the constraint is given by $\dot{k} \equiv dk(t) / dt = T[k(t), v(t)]$. Because d$k(t)$ in (2) is a random variable we can compute its mean and variance. They are given by

$$E[dk(t)] = T[k(t), v(t)]; \mathrm{Var}[dk(t)] = \sigma^2[k(t), v(t)]dt.$$

Thus (2) is a meaningful generalization of its counterpart in deterministic control because it involves means, standard deviations and pure randomness in capturing the complexities of economic reality. A comprehensive analysis of Itō equations, such as (2), is given in Malliaris and Brock (1982).

The problem in (1) and (2) is a stochastic analogue of the deterministic one studied in Arrow and Kurz (1970, pp. 27–51). A standard technique for our problem, as in the case of Arrow and Kurz, is Bellman's (1957, p. 83) ‘Principle of Optimality’ according to which ‘an optimal policy has the property that, whatever the initial state and control are, the remaining decisions must constitute an optimal policy with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

regard to the state resulting from the first decision'. The problem in eqs (1) and (2) is studied here for the undiscounted, finite horizon case, that is, for $\rho = 0$ and $N < \infty$.
Using Bellman's technique for dynamic programming, eqs (1) and (2) can be analysed as follows:

$$J[k(t), t, N] = \max E_t \int_t^N u(k, v)\mathrm{d}s = \max E_t \int_t^{t+\Delta t} u(k, v)\mathrm{d}s + \max E_{t+\Delta t} \int_{t+\Delta t}^N u(k, v)\mathrm{d}s = \max E_t \int_t^{t+\Delta t} u(k, v)\mathrm{d}s + \tag{3}$$

Observe that Taylor's theorem is used to obtain (3) and therefore it is assumed that J has continuous partial derivatives of all orders less than three in some open set containing the line segment connecting the two points $[k(t), t]$ and $[k(t + \Delta t), t + \Delta t]$. Let (2) be approximated and write

$$\Delta k = T(k, v)\Delta t + \sigma(k, v)\Delta Z + o(\Delta t)\,. \tag{4}$$

Insert (4) into (3) and use the multiplication rules

$$(\Delta Z) \times (\Delta t) = 0,\ (\Delta t) \times (\Delta t) = 0 \text{ and } (\Delta Z) \times (\Delta Z) = \Delta t$$

to get

$$0 = \max E_t\left[u(k, v)\Delta t + \left(J_k T + J_t + \frac{1}{2}J_{kk}\sigma^2\right)\Delta t + J_k\sigma\Delta Z + o(\Delta t)\right]. \tag{5}$$

For notational convenience let

$$\Delta J = \left[J_t + J_k T + \frac{1}{2}J_{kk}\sigma^2\right]\Delta t + J_k\sigma\Delta Z. \tag{6}$$

Using (6), eq. (5) becomes

$$0 = \max E_t[u(k, v)\Delta t + \Delta J + o(\Delta t)]. \tag{7}$$

This is a partial differential equation with boundary condition $[(\partial J) / \partial K][k(N), N, N] = 0$. Pass E_t

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

through the parentheses of (7) and, after dividing both sides by Δt, let $\Delta t \to 0$ to conclude

$$0 = \max\left[u(k, v) + J_t + J_k T(k, v) + \frac{1}{2} J_{kk} \sigma^2(k, v)\right]. \tag{8}$$

This last equation is usually written as

$$-J_t = \max\left[u(k, v) + J_k T(k, v) + \frac{1}{2} J_{kk} \sigma^2(k, v)\right] \tag{9}$$

and is known as the Hamilton–Jacobi–Bellman equation of stochastic control theory.
Next, we define the costate variable $p(t)$ as

$$p(t) = J_k[k(t), t, N]$$

from which it follows that its partial derivative with respect to k is

$$p_k = \partial p / \partial k = J_{kk}. \tag{10}$$

Therefore, we may rewrite (9) as

$$-J_t = \max H(k, v, p, \partial p / \partial k). \tag{11}$$

where H is the functional notation of the expression inside the brackets of (9). Assume next that a function v exists that solves the maximization problem of (11) and denote such a function by

$$v^0 = v^0(k, p, \partial p / \partial k). \tag{12}$$

Note that v^0 is a function of $k(t)$ and t alone, along the optimum path, because J_k is a function of $k(t)$ and t alone. In the applied control literature, and more specifically in economic applications, v^0 is called a *policy function*. Assuming then that a policy function v^0 exists, (11) may be rewritten as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$-J_1 = \max H(k, v, p, \partial p / \partial k) = H\left[k, v^0(k, p, \partial p / \partial k), p, \partial p / \partial k\right] = H^0(k, p, \partial p / \partial k). \quad (13)$$

This last equation is again a functional notation of the right-hand side expression of (9) under the assumption of the existence of an optimum control v^0, that is,

$$H^0(k, p, \partial p / \partial k) = u(k, v^0) + pT(k, v^0) + \frac{1}{2}\frac{\partial p}{\partial k}\sigma^2(k, v^0). \quad (14)$$

Equipped with the above analysis we can now state

Proposition 1: (Pontryagin stochastic maximum principle). Suppose that $k(t)$ and $v^0(t)$ solve for $t \in [0, N]$ the problem:

$$\max E_0 \int_0^N u(k, v) \mathrm{d}t$$

subject to the conditions

$$\mathrm{d}k = T(k, v)\mathrm{d}t + \sigma(k, v)\mathrm{d}Z, \; k(t) \text{given}.$$

Then, there exists a costate variable $p(t)$ such that for each t, $t \in [0, N]$:

1. (1) v^0 maximizes $H(k, v, p, \partial p / \partial k)$ where

$$H(k, v, p, \partial p / \partial k) = u(k, v) + pT(k, v) + \frac{1}{2}\frac{\partial p}{\partial k}\sigma^2;$$

2. (2) the costate function p(t) satisfies the stochastic differential equation

$$\mathrm{d}p = -H_k^0 \mathrm{d}t + \sigma(k, v^0) J_{kk} \mathrm{d}Z; \text{ and}$$

3. (3) the transversality condition holds

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$p[k(N), N] = \frac{\partial J}{\partial k}[k(N), N, N] = 0, p(N)k(N) = 0.$$

Finally, we briefly illustrate the stochastic optimal control technique to the stochastic Ramsey problem studied in Merton (1975). The problem is to find an optimal saving policy s^0 to

$$\text{maximize } E_0 \int_0^T u(c)\mathrm{d}t \tag{15}$$

subject to

$$\mathrm{d}k = \left[sf(k) - (n - \sigma^2)k\right]\mathrm{d}t - \sigma k \mathrm{d}Z \tag{16}$$

and $k(t) \geq 0$ for each t. Here, u is a strictly concave, von Neumann–Morgenstern utility function of per capita consumption c for the representative consumer and $f(k)$ is a well-behaved production function. Note that $c=(1 - s)f(k)$ and that eq. (16) generalizes Solow's equation of neoclassical economic growth. Uncertainty enters (16) via randomness in the rate of growth of the labour force. Let

$$J[k(t), t, T] = \max_s E_t \int_t^T u[(1 - s)f(k)]\mathrm{d}t.$$

The Hamilton–Jacobi–Bellman equation is given by

$$0 = \max\left\{u[(1 - s)f(k)] + J_t + J_k\left[sf(k) - (n - \sigma^2)k\right] + \frac{1}{2}J_{kk}\sigma^2 k^2\right\} \tag{17}$$

which yields

$$\frac{\mathrm{d}u}{\mathrm{d}c}\left[(1 - s^0)f(k)\right] = J_k. \tag{18}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To solve for s^0, in principle, one solves (18) for s^0 as a function of k, $T-t$ and J_k, and then substitutes this solution into (17) which becomes a *partial differential equation for J*. Once (17) is solved, then its solution is substituted back into (18) to determine s^0 as a function of k and $T-t$. The nonlinearity of the Hamilton–Jacobi–Bellman equation causes difficulties in finding a closed form solution for the optimal saving function. However, if we let $\sigma=0$ in (17), one obtains the classical Ramsey rule of the certainty case.

From the fact that numerous economic qsts involve uncertainty and can be formulated as stochastic optimal control problems, one may conclude that economic interest is likely to be lively in this area for some time to come.

See Also

- dynamic programming
- Markov processes
- Wiener process

Bibliography

Arrow, K.J. and Kurz, M. 1970. *Public Investment, the Rate of Return, and Optimal Fiscal Policy*. Baltimore, MD: Johns Hopkins Press.

Bellman, R. 1957. *Dynamic Programming*. Princeton: Princeton University Press.

Bertsekas, D.P. and Shreve, S.E. 1978. *Stochastic Optimal Control: The Discrete Time Case*. New York: Academic Press.

Brock, W.A. and Mirman, L. 1972. Optimal economic growth and uncertainty: the discounted case. *Journal of Economic Theory* 4, 479–513.

Chang, F.R. 2004. *Stochastic Optimization in Continuous Time*. New York: Cambridge University Press.

Fleming, W.H. and Rishel, R.W. 1975. *Deterministic and Stochastic Optimal Control*. New York: Springer.

Malliaris, A.G. and Brock, W.A. 1982. *Stochastic Methods in Economics and Finance*. Amsterdam: North-Holland.

Merton, R.C. 1971. Optimal consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* 3, 373–413.

Merton, R.C. 1975. An asymptotic theory of growth under uncertainty. *Review of Economic Studies* 42, 375–93.

How to cite this article

Malliaris, A. G. "stochastic optimal control." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000269>
doi:10.1057/9780230226203.1624

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Onlir

stochastic volatility models

Neil Shephard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Stochastic volatility (SV) is the main concept used in the fields of financial economics and mathematical finance to deal with the endemic time-varying volatility and codependence found in financial markets. Here I trace the origins of SV and provide links with the basic models used today in the literature. I briefly discuss some of the innovations in the second generation of SV models and discuss the literature on conducting inference for SV models. I talk about the use of SV to price options, and consider the connection of SV with realized volatility.

Keywords

asset pricing; Black–Scholes–Merton prices; Brownian motion; Dambis–Dubins–Schwartz theorem; financial econometrics; generalized method of moments (GMM) estimators; Kalman filter; Markov chain Monto Carlo (MCMC) methods; Markov processes; martingales; multivariate models; option pricing theory; options; probability; quadratic variation (QV) process; realized volatility; stochastic discount factor (SDF) approach; stochastic volatility

Article

Stochastic volatility (SV) is the main concept used in the fields of financial economics and mathematical finance to deal with the endemic time-varying volatility and codependence found in financial markets. Such dependence has been known for a long time; early commentators include Mandelbrot (1963) and Officer (1973). It was also clear to the founding fathers of modern continuous time finance that homogeneity was an unrealistic if convenient simplification; for example, Black and Scholes (1972, p. 416) wrote, '… there is evidence of non-stationarity in the variance. More work must be done to predict variances using the information available.' Heterogeneity has deep implications for the theory and practice of financial economics and econometrics. In particular, asset pricing theory is dominated by the idea that higher rewards may be expected when we face higher risks, but these risks change through time in complicated ways. Some of the changes in the level of risk can be modelled stochastically, where the level of volatility and degree of codependence between assets is allowed to change over time. Such models allow us to explain, for example, empirically observed departures from Black–Scholes–Merton prices for options and understand why we should expect to see occasional dramatic moves in financial markets.
The outline of this article is as follows. In the first section I trace the origins of SV and provide links with the basic models used today in the literature. In the second section I briefly discuss some of the innovations in the second generation of SV models. In the third section I briefly discuss the literature on conducting inference for SV models. In the fourth section I talk about the use of SV to price options. In the fifth section I consider the connection of SV with realized volatility. An extensive review of this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

literature is given in Shephard (2005).

The origin of SV models

The origins of SV are messy. I will give five accounts, which attribute the subject to different sets of people. Clark (1973) introduced Bochner's (1949) time-changed Brownian motion (BM) into financial economics. He wrote down a model for the log-price M as

$$M_t = W_{\tau_t}, \; t \geq 0, \tag{1}$$

where W is Brownian motion (BM), t is continuous time, τ is a time change and $W \quad \tau$, where denotes independence. The definition of a time-change is a non-negative process with non-decreasing sample paths, although Clark also assumed τ has independent increments. Then $M_t | \tau_t \sim N(0, \tau_t)$. Further, so long (for each t) as $E\sqrt{\tau_t} < \infty$, then M is a martingale (written $M \in \mathcal{M}$) for this is necessary and sufficient to ensure that $E|M_t| < \infty$. More generally, if (for each t) $\tau_t < \infty$, then M is a local martingale (written $M \in \mathcal{M}_{loc}$). Hence Clark was solely modelling the instantly risky component of the log of an asset price, written Y, which in modern semimartingale (written $Y \in \mathcal{SM}$) notation we would write as

$$Y = A + M.$$

The increments of A can be thought of as the instantly available reward component of the asset price, which compensates the investor for being exposed to the risky increments of M. The A process is assumed to be of finite variation (written $A \in \mathcal{FV}$).
To the best of my understanding, the first published direct volatility clustering SV paper is that by Taylor (1982). His discrete time model of daily returns, computed as the difference of log-prices

$$y_i = Y_i - Y_{i-1}, \; i = 1, 2, \ldots,$$

where I have assumed that $t = 1$, represents one day to simplify the exposition. He modelled the risky part of returns, $m_i = M_i - M_{i-1}$, as a product process

$$m_i = \sigma_i \varepsilon_i. \tag{2}$$

Taylor assumed ε has a mean of zero and unit variance, while σ is some non-negative process, finishing the model by assuming $\varepsilon \quad \sigma$. Taylor modelled ε as an autoregression and

$$\sigma_i = \exp(h_i / 2),$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where h is a non-zero mean Gaussian linear process. The leading example of this is the first order autoregression

$$h_{i+1} = \mu + \varphi(h_i - \mu) + \eta_i, \; \eta_i \sim NID(0, \sigma_\eta^2). \tag{3}$$

In the modern SV literature the model for ε is typically simplified to an i.i.d. process, for we deal with the predictability of asset prices through the A process rather than via M. This is now often called the log-normal SV model in the case where ε is also assumed to be Gaussian. In general, M is always a local martingale.

A key feature of SV, which is not discussed by Taylor, is that it can deal with leverage effects. Leverage effects are associated with the work of Black (1976) and Nelson (1991), and can be implemented in discrete time SV models by negatively correlating the Gaussian ε_i and η_i. This still implies that $M \in \mathcal{M}_{loc}$, but allows the direction of returns to influence future movements in the volatility process, with falls in prices associated with rises in subsequent volatility.

Taylor's discussion of the product process was pre-dated by a decade in the (until recently) unpublished Rosenberg (1972). Rosenberg introduces product processes, empirically demonstrating that time-varying volatility is partially forecastable and so breaks with the earlier work by Clark. He suggests an understanding of aggregational Gaussianity of returns over increasing time intervals and pre-dates a variety of econometric methods for analysing heteroskedasticity.

In continuous time the product process is the standard SV model

$$M_t = \int_0^t \sigma_s dW_s, \tag{4}$$

where the non-negative spot volatility σ is assumed to have càdlàg sample paths (which means it can possess jumps). The squared volatility process is often called the spot variance.

The first use of continuous-time SV models in financial economics was, to my knowledge, by Johnson (1979), who studied the pricing of options using time-changing volatility models in continuous time (see also Johnson and Shanno, 1987; Wiggins, 1987). The best-known paper in this area is Hull and White (1987). Each of these authors desired to generalize the Black and Scholes (1973) approach to option pricing models to deal with volatility clustering. In the Hull and White approach, σ^2 follows the solution to the univariate SDE

$$d\sigma^2 = \alpha(\sigma^2)dt + \omega(\sigma^2)dB,$$

where B is a second Brownian motion and $\omega(\cdot)$ is a non-negative deterministic function.

The probability literature has demonstrated that SV models and their time-changed BM relatives are fundamental. This theoretical development will be the fifth strand of literature that I think of as representing the origins of modern stochastic volatility research. Suppose we simply assume that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$M \in \mathcal{M}_{loc}^{c}$, a process with continuous local martingale sample paths. Then the celebrated Dambis–Dubins–Schwartz theorem shows that *M* can be written as a time-changed Brownian motion. Further, the time-change is the quadratic variation (QV) process

$$[M]_t = \mathrm{p}-\lim_{n\to\infty}\sum_{j=1}^{n}\left(M_{t_j}-M_{t_{j-1}}\right)\left(M_{t_j}-M_{t_{j-1}}\right)', \tag{5}$$

for any sequence of partitions $t_0 = 0 < t_1 < \ldots < t_n = t$ with $\sup_j\{t_j - t_{j-1}\} \to 0$ for $n \to \infty$. What is more, as *M* has continuous sample paths, so must [*M*]. Under the stronger condition that [*M*] is absolutely continuous, then *M* can be written as a stochastic volatility process. This latter result, which is called the martingale representation theorem, is due to Doob (1953). Taken together, this implies that time-changed BMs are canonical in continuous sample path price processes, and SV models are special cases of this class. A consequence of the fact that, for continuous sample path time-change BM, $[M] = \tau$, is that in the SV case

$$[M]_t = \int_0^t \sigma_s^2\, ds.$$

The SV framework has an elegant multivariate generalization. In particular, write a *p*-dimensional price process *M* as (4) but where σ is a matrix process whose elements are all càdlàg, *W* is a multivariate BM process. Further $[M]_t = \int_0^t \sigma_s \sigma_s' ds$.

Second-generation model building

Univariate models

General observations

In initial diffusion-based models the volatility was Markovian with continuous sample paths. Research in the late 1990s and early 2000s has shown that more complicated volatility dynamics are needed to model either options data or high frequency return data. Leading extensions to the model are to allow jumps into the volatility SDE (for example, Barndorff-Nielsen and Shephard, 2001; Eraker, Johannes and Polson, 2003) or to model the volatility process as a function of a number of separate stochastic processes or factors (for example, Chernov et al., 2003; Barndorff-Nielsen and Shephard, 2001).

Long memory

In the SV literature considerable progress has been made on working with both discrete and continuous time long-memory SV. This involves specifying a long-memory model for σ in discrete or continuous time.
Breidt, Crato and de Lima (1998) and Harvey (1998) looked at discrete time models where the log of the volatility was modelled as a fractionally integrated process. In continuous time there is work on modelling the log of volatility as fractionally integrated Brownian motion by Comte and Renault (1998). More recent work, which is econometrically easier to deal with, is the square-root model driven by fractionally integrated BM introduced in an influential paper by Comte, Coutin and Renault (2003) and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the infinite superposition of non-negative OU processes introduced by Barndorff-Nielsen (2001).

Jumps

In detailed empirical work a number of researchers have supplemented standard SV models by adding jumps to the price process or to the volatility dynamics. Bates (1996) was particularly important as it showed the need to include jumps in addition to SV, at least when volatility is Markovian. Eraker, Johannes and Polson (2003) deals with the efficient inference of these types of models. A radical departure in SV models was put forward by Barndorff-Nielsen and Shephard (2001), who suggested building volatility models out of pure jump processes called non-Gaussian OU processes. Closed form option pricing based on this structure is studied briefly in Barndorff-Nielsen and Shephard (2001) and in detail by Nicolato and Venardos (2003). All these non-Gaussian OU processes are special cases of the affine class advocated by Duffie, Pan and Singleton (2000) and Duffie, Filipovic and Schachermayer (2003).

Multivariate models

Diebold and Nerlove (1989) introduced volatility clustering into traditional factor models, which are used in many areas of asset pricing. In continuous time their type of model has the interpretation

$$M_t = \sum_{j=1}^{J} \int \beta_{(j)s} dF_{(j)s} + G_t,$$

where the factors $F_{(1)}, F_{(2)}, \ldots, F_{(J)}$ are independent univariate SV models and G is correlated multivariate BM. Some of the related papers on the econometrics of this topic include King, Sentana and Wadhwani (1994) and Fiorentini, Sentana and Shephard (2004), who all fit this kind of model. These papers assume that the factor loading vectors are constant through time.
A more limited multivariate discrete time model was put forward by Harvey, Ruiz and Shephard (1994) who allowed $M_t = C\int_0^t \sigma_s dW_s$, where σ is a diagonal matrix process and C is a fixed matrix of constants with a unit leading diagonal. This means that the risky part of prices is simply a rotation of a p-dimensional vector of independent univariate SV processes.

Inference based on return data

Moment-based inference

The task is to carry out inference on $\theta = (\theta_1, \ldots, \theta_K)'$, the parameters of the SV model based on a sequence of returns $Y = (Y_1, \ldots, Y_T)'$. Taylor (1982) and Melino and Turnbull (1990) calibrated their models using the method of moments. Systematic studies, using a GMM approach, of which moments to heavily weight in SV models was given in Andersen and Sørensen (1996), Genon-Catalot, Jeantheau and Larédo (2000), Sørensen (2000) and Hoffmann (2002).
A difficulty with using moment-based estimators for continuous time SV models is that it is not straightforward to compute the moments y. In the case of no leverage, general results for the second order properties of y and their squares were given in Barndorff-Nielsen and Shephard (2001). Some quite general results under leverage are also given in Meddahi (2001).
In the discrete time log-normal SV models the approach advocated by Harvey, Ruiz and Shephard (1994) has been influential. Their approach was to remove the predictable part of the returns, so we

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

think of $Y = M$ again, and work with $\log y_i^2 = h_i + \log \varepsilon_i^2$. If the volatility has short memory, then this form of the model can be handled using the Kalman filter, while long-memory models are often dealt with in the frequency domain. Either way, this delivers a Gaussian quasi-likelihood which can be used to estimate the parameters of the model. The linearized model is non-Gaussian due to the long left-hand tail of $\log \varepsilon_i^2$ which generates outliers when ε_i is small.

Simulation-based inference

In the 1990s a number of econometricians started to use simulation-based inference to tackle SV models. To discuss these methods it will be convenient to focus on the simplest discrete time log-normal SV model given by (2) and (3).

MCMC allows us to simulate from $\theta, h | y$, where $h = (h_1, \ldots, h_T)'$. Discarding the h draws yields samples from $\theta | y$. Summarizing yields fully efficient parametric inference. In an influential paper, Jacquier, Polson and Rossi (1994) implemented an MCMC algorithm for this problem. A subsequent paper by Kim, Shephard and Chib (1998) gave quite an extensive discussion of various MCMC algorithms. This is a subtle issue and makes a very large difference to the computational efficiency of the methods (see, for example, Jacquier, Polson and Rossi, 2004; Yu, 2005).

Kim, Shephard and Chib (1998) introduced the first filter using a so-called particle filter. As well as being of substantial scientific interest for decision making, the advantage of a filtering method is that it allows us to compute marginal likelihoods for model comparison and one-step-ahead predictions for model testing.

Although MCMC-based papers are mostly couched in discrete time, a key advantage of the general approach is that it can be adapted to deal with continuous time models by the idea of augmentation. This was fully worked out in Elerian, Chib and Shephard (2001), Eraker (2001) and Roberts and Stramer (2001).

A more novel non-likelihood approach was introduced by Smith (1993) and later developed by Gourieroux, Monfort and Renault (1993) and Gallant and Tauchen (1996) into what is now called indirect inference or the efficient method of moments. Here I briefly give a stylized version of this approach.

Suppose there is an auxiliary model for the returns (for example, GARCH) whose density, $g(y, \psi)$, is easy to compute and, for simplicity of exposition, has $\dim(\psi) = \dim(\theta)$. Then compute its MLE, which we write as $\hat{\psi}$. We assume this is a regular problem so that $\partial \log g(y, \hat{\psi}) / \partial \psi = 0$ recalling that y is the observed return vector. Simulate a very long process from the SV model using parameters θ, which we denote by y^+, and evaluate the score using not the data but this simulation. This produces

$$\left. \frac{\partial \log g(y^+; \psi)}{\partial \psi} \right|_{\psi = \hat{\psi}}, \quad y^+ \sim f(y, \theta).$$

Then move θ around until the score is again zero, but now under the simulation. Write the point where this happens as $\tilde{\theta}$. It is called the indirect inference estimator.

Options

Models

SV models provide a basis for realistic modelling of option prices. We recall the central role played by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Johnson and Shanno (1987) and Wiggins (1987). The best-known paper in this area is by Hull and White (1987), who looked at a diffusion volatility model with leverage effects. They assumed that volatility risk was unrewarded and priced their options either by approximation or by simulation. Hull and White (1987) indicated that SV models could produce smiles and skews in option prices, which are frequency observed in market data. The skew is particularly important in practice, and Renault and Touzi (1996) proved that can be achieved in SV models via leverage effects.
The first analytic option pricing formulae were developed by Stein and Stein (1991) and Heston (1993). The only other closed form solution I know of is the one based on the Barndorff-Nielsen and Shephard (2001) class of non-Gaussian OU SV models. Nicolato and Venardos (2003) provided a detailed study of such option pricing solutions; see also the textbook exposition in Cont and Tankov (2004, ch. 15). Slightly harder computationally to deal with is the more general affine class of models highlighted by Duffie, Filipovic and Schachermayer (2003).

Econometrics of SV option pricing

In theory, option prices themselves should provide rich information for estimating and testing volatility models. I discuss the econometrics of options in the context of the stochastic discount factor (SDF) approach, which has a long history in financial economics and is emphasized in, for example, Cochrane (2001) and Garcia, Ghysels and Renault (2006). For simplicity I assume interest rates are constant. We start with the standard Black–Scholes (BS) problem, which will take a little time to recall, before being able to rapidly deal with the SV extension. We model

$$\mathrm{d}\log Y = (r + p - \sigma^2/2)\mathrm{d}t + \sigma \mathrm{d}W, \mathrm{d}\log \widetilde{M} = h\mathrm{d}t + b\mathrm{d}W,$$

where $\widetilde{M}$ is the SDF process and r the riskless short rate, and σ, h, b and p, the risk premium, are assumed constant for the moment.
We price all contingent payoffs $g(Y_T)$ as $C_t = E(\frac{\widetilde{M}_T}{\widetilde{M}_t} g(Y_T) | \mathcal{F}_t)$, the expected discounted value of the claim where $T > t$. For this model to make financial sense we require that $\widetilde{M}_t Y_t$ and $\widetilde{M}_t \exp(tr)$ are local martingales, which is enough to mean that adding other independent BMs to the $\log \widetilde{M}$ process makes no difference to C or Y, the observables. These two constraints imply, respectively, $p + b\sigma = 0$ and $h = -r - b^2/2$. This means that (C^{BS}, Y) is driven by a single W.
When we move to the standard SV model we can remove this degeneracy. The functional form for the SV Y process is unchanged, but we now allow

$$\mathrm{d}\log \widetilde{M} = h\mathrm{d}t + a\mathrm{d}B + b\mathrm{d}W, \mathrm{d}\sigma^2 = \alpha \mathrm{d}t + \omega \mathrm{d}B,$$

where we assume that $B \quad W$ to simplify the exposition. The SV structure means that p will have to change through time in response to the moving σ^2. B is again redundant in the SDF (but not in the volatility) so the usual SDF conditions again imply $h = -r - \frac{1}{2}a^2$ and $p + b\sigma = 0$. This implies that the move to the SV case has little impact, except that the sample path of $\sigma^2 \quad W$. So the generalized BS (GBS) price is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$C_t^{GBS}(\sigma_t^2) = E\left(\frac{\widetilde{M}_T}{\widetilde{M}_t} g(Y_T) | \mathcal{F}_t\right) = E\left\{C_t^{BS}\left(\frac{1}{T-t}\int_t^T \sigma_u^2 du\right) | \sigma_t^2, Y_t\right\}.$$

Now C^{GBS} is a function of both Y_t and σ_t^2, which means that (C^{GBS}, Y) is not degenerate. From an econometric viewpoint this is an important step, meaning inference on options is just the problem of making inference on a complicated bivariate diffusion process. When we allow leverage back into the model, the analysis becomes slightly more complicated algebraically.
In some recent work econometricians have been trying to use data from underlying assets and option markets to jointly model the dynamics of (C^{GBS}, Y). The advantage of this joint estimation is that we can pool information across data types and estimate all relevant effects which influence Y, σ^2 and $\widetilde{M}$. Relevant papers include Chernov and Ghysels (2000), Pastorello, Patilea and Renault (2003), Das and Sundaram (1999) and Bates (2000).

Realized volatility

The advent of very informative high-frequency data has prompted econometricians to study estimators of the increments of the quadratic variation (QV) process and then to use this estimate to project QV into the future in order to predict future levels of volatility. The literature on this starts with independent, concurrent papers by Andersen and Bollerslev (1998), Barndorff-Nielsen and Shephard (2001) and Comte and Renault (1998). Some of this work echoes earlier important contributions from, for example, Rosenberg (1972) and Merton (1980).
A simple estimator of $[Y]$ is the *realized QV process*

$$[Y_\delta]_t = \sum_{j=1}^{\lfloor t/\delta \rfloor} (Y_{\delta j} - Y_{\delta(j-1)})(Y_{\delta j} - Y_{\delta(j-1)})',$$

thus as $\delta \downarrow 0$ so $[Y_\delta]_t \xrightarrow{p} [Y]_t$. If $A \in \mathcal{FV}^c$, then $[Y] = [M]$, while if we additionally assume that M is SV then $[Y_\delta]_t \xrightarrow{p} \int_0^t \sigma_s \sigma_s' ds$.
In practice it makes sense to look at the increments of the QV process. Suppose we are interested in analysing daily return data, but in addition have higher-frequency data measured at the time interval δ. The i-th daily realized QV is defined as

$$V(Y_\delta)_i = \sum_{j=1}^{\lfloor 1/\delta \rfloor} (Y_{i+\delta j} - Y_{i+\delta(j-1)}) \times (Y_{i+\delta j} - Y_{i+\delta(j-1)})' \xrightarrow{p} V(Y)_i = [Y]_i - [Y]_{i-1},$$

the i-th daily QV. The diagonal elements of $V(Y_\delta)_i$ are called realized variances and their square roots are called realized volatilities.
Andersen et al. (2001) have shown that to forecast the volatility of future asset returns a key input should be predictions of future daily QV. Recall, from Ito's formula, that, if $Y \in \mathcal{SM}^c$ and $M \in \mathcal{M}$, then

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

writing F_t as the filtration generated by the continuous history of Y up to time t then

$$E(y_i y_i' | \mathcal{F}_{i-1}) \simeq E(V(Y)_i | \mathcal{F}_{i-1}).$$

A review of some of this material is given by Barndorff-Nielsen and Shephard (2006a).

A difficulty with this line of argument is that the QV theory tells us only that $V(Y_\delta)_i \xrightarrow{p} V(Y)_i$; it gives no impression of the size of $V(Y_\delta)_i - V(Y)_i$. Jacod (1994) and Barndorff-Nielsen and Shephard (2002) have strengthened the consistency result to provide a univariate central limit theory

$$\frac{\delta^{-1/2}([Y_\delta]_t - [Y]_t)}{\sqrt{2\int_0^t \sigma_s^4 \mathrm{d}s}} \xrightarrow{d} N(0, 1),$$

while giving a method for consistently estimating the integrated quarticity $\int_0^t \sigma_s^4 \mathrm{d}s$ by using high-frequency data. This analysis was generalized to the multivariate case by Barndorff-Nielsen and (Shephard 2004a).This type of analysis greatly simplifies parametric estimation of SV models, for we can now have estimates of the volatility quantities SV models directly parameterize. Barndorff-Nielsen and Shephard (2002), Bollerslev and Zhou (2002) and Phillips and Yu (2005) study this topic from different perspectives.

Recently there has been interest in studying the impact of market microstructure effects on the estimates of realized covariation. This causes the estimator of the QV to become biased. Leading papers on this topic are Zhou (1996), Fang (1996), Bandi and Russell (2003), Hansen and Lunde (2006) and Zhang, Mykland and Aït-Sahalia (2005). Further, one can estimate the QV of the continuous component of prices in the presence of jumps using the so-called realized bipower variation process. This was introduced by Barndorff-Nielsen and Shephard (2004b; 2006b).

See Also

- capital asset pricing model
- options
- options (new perspectives)
- realized volatility

My research is supported by the Economic and Social Science Research Council (UK) through the grant 'High frequency financial econometrics based upon power variation'.

Bibliography

Andersen, T. and Bollerslev, T. 1998. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39, 885–905.

Andersen, T., Bollerslev, T., Diebold, F. and Labys, P. 2001. The distribution of exchange rate volatility. *Journal of the American Statistical Association* 96, 42–55. Correction published in vol. 98 (2003), p. 501.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Andersen, T. G. and Sørensen, B. 1996. GMM estimation of a stochastic volatility model: a Monte Carlo study. *Journal of Business and Economic Statistics* 14, 328–52.

Bandi, F. and Russell, J. 2003. Microstructure noise, realized volatility, and optimal sampling. Mimeo. Graduate School of Business, University of Chicago.

Barndorff-Nielsen, O. 2001. Superposition of Ornstein–Uhlenbeck type processes. *Theory of Probability and its Applications* 45, 175–94.

Barndorff-Nielsen, O. and Shephard, N. 2001. Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society, Series B* 63, 167–241.

Barndorff-Nielsen, O. and Shephard, N. 2002. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* 64, 253–80.

Barndorff-Nielsen, O. and Shephard, N. 2004a. Econometric analysis of realised covariation: high frequency covariance, regression and correlation in financial economics. *Econometrica* 72, 885–925.

Barndorff-Nielsen, O. and Shephard, N. 2004b. Power and bipower variation with stochastic volatility and jumps (with discussion). *Journal of Financial Econometrics* 2, 1–48.

Barndorff-Nielsen, O. and Shephard, N. 2006a. Variation, jumps, and high frequency data in financial econometrics. In *Advances in Economics and Econometrics: Theory and Applications*, vol. 1, ed. R. Blundell, P. Torsten and W. Newey. Cambridge: Cambridge University Press.

Barndorff-Nielsen, O. and Shephard, N. 2006b. Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* 4, 1–30.

Bates, D. 1996. Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. *Review of Financial Studies* 9, 69–107.

Bates, D. 2000. Post-’97 crash fears in the S-&P 500 futures option market. *Journal of Econometrics* 94, 181–238.

Black, F. 1976. Studies of stock price volatility changes. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 177–81.

Black, F. and Scholes, M. 1972. The valuation of option contracts and a test of market efficiency. *Journal of Finance* 27, 399–418.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–54.

Bochner, S. 1949. Diffusion equation and stochastic processes. *Proceedings of the National Academy of Science of the United States of America* 85, 369–70.

Bollerslev, T. and Zhou, H. 2002. Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Journal of Econometrics* 109, 33–65.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Breidt, F., Crato, N. and de Lima, P. 1998. On the detection and estimation of long memory in stochastic volatility. *Journal of Econometrics* 83, 325–48.

Chernov, M. and Ghysels, E. 2000. A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *Journal of Financial Economics* 56, 407–58.

Chernov, M., Gallant, A., Ghysels, E. and Tauchen, G. 2003. Alternative models of stock price dynamics. *Journal of Econometrics* 116, 225–57.

Clark, P. 1973. A subordinated stochastic process model with fixed variance for speculative prices. *Econometrica* 41, 135–56.

Cochrane, J. 2001. *Asset Pricing*. Princeton: Princeton University Press.

Comte, F., Coutin, L. and Renault, E. 2003. Affine fractional stochastic volatility models. Mimeo. University of Montreal.

Comte, F. and Renault, E. 1998. Long memory in continuous-time stochastic volatility models. *Mathematical Finance* 8, 291–323.

Cont, R. and Tankov, P. 2004. *Financial Modelling with Jump Processes*. London: Chapman and Hall.

Das, S. and Sundaram, R. 1999. Of smiles and smirks: a term structure perspective. *Journal of Financial and Quantitative Analysis* 34, 211–40.

Diebold, F. and Nerlove, M. 1989. The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *Journal of Applied Econometrics* 4, 1–21.

Doob, J. 1953. *Stochastic Processes*. New York: John Wiley and Sons.

Duffie, D., Filipovic, D. and Schachermayer, W. 2003. Affine processes and applications in finance. *Annals of Applied Probability* 13, 984–1053.

Duffie, D., Pan, J. and Singleton, K. 2000. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* 68, 1343–76.

Elerian, O., Chib, S. and Shephard, N. 2001. Likelihood inference for discretely observed non-linear diffusions. *Econometrica* 69, 959–93.

Eraker, B. 2001. Markov chain Monte Carlo analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics* 19, 177–91.

Eraker, B., Johannes, M. and Polson, N. 2003. The impact of jumps in returns and volatility. *Journal of Finance* 53, 1269–300.

Fang, Y. 1996. Volatility modeling and estimation of high-frequency data with Gaussian noise. Ph.D. thesis. Sloan School of Management, MIT.

Fiorentini, G., Sentana, E. and Shephard, N. 2004. Likelihood-based estimation of latent generalised

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

ARCH structures. *Econometrica* 12, 1481–517.

Gallant, A. and Tauchen, G. 1996. Which moments to match. *Econometric Theory* 12, 657–81.

Garcia, R., Ghysels, E. and Renault, E. 2006. The econometrics of option pricing. In *Handbook of Financial Econometrics*, ed. Y. Aït-Sahalia and L. Hansen. Amsterdam: North-Holland.

Genon-Catalot, V., Jeantheau, T. and Larédo, C. 2000. Stochastic volatility as hidden Markov models and statistical applications. *Bernoulli* 6, 1051–79.

Gourieroux, C., Monfort, A. and Renault, E. 1993. Indirect inference. *Journal of Applied Econometrics* 6, S85–S118.

Hansen, P. and Lunde, A. 2006. Realized variance and market microstructure noise (with discussion). *Journal of Business and Economic Statistics* 24, 127–61.

Harvey, A. 1998. Long memory in stochastic volatility. In *Forecasting Volatility in Financial Markets*, ed. J. Knight and S. Satchell. Oxford: Butterworth-Heinemann.

Harvey, A., Ruiz, E. and Shephard, N. 1994. Multivariate stochastic variance models. *Review of Economic Studies* 61, 247–64.

Heston, S. 1993. A closed-form solution for options with stochastic volatility, with applications to bond and currency options. *Review of Financial Studies* 6, 327–43.

Hoffmann, M. 2002. Rate of convergence for parametric estimation in stochastic volatility models. *Stochastic Processes and their Application* 97, 147–70.

Hull, J. and White, A. 1987. The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42, 281–300.

Jacod, J. 1994. Limit of random measures associated with the increments of a Brownian semimartingale. Preprint No. 120. Laboratoire de Probabilitiés, Université Pierre et Marie Curie, Paris.

Jacquier, E., Polson, N. and Rossi, P. 1994. Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business and Economic Statistics* 12, 371–417.

Jacquier, E., Polson, N. and Rossi, P. 2004. Stochastic volatility models: univariate and multivariate extensions. *Journal of Econometrics* 122, 185–212.

Johnson, H. 1979. Option pricing when the variance rate is changing. Working paper. University of California, Los Angeles.

Johnson, H. and Shanno, D. 1987. Option pricing when the variance is changing. *Journal of Financial and Quantitative Analysis* 22, 143–51.

Kim, S., Shephard, N. and Chib, S. 1998. Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies* 65, 361–93.

King, M., Sentana, E. and Wadhwani, S. 1994. Volatility and links between national stock markets.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Econometrica 62, 901–33.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Meddahi, N. 2001. An eigenfunction approach for volatility modeling. Cahiers de recherche No. 2001–29. Department of Economics, University of Montreal.

Melino, A. and Turnbull, S. 1990. Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* 45, 239–65.

Merton, R. 1980. On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics* 8, 323–61.

Nelson, D. 1991. Conditional heteroskedasticity in asset pricing: a new approach. *Econometrica* 59, 347–70.

Nicolato, E. and Venardos, E. 2003. Option pricing in stochastic volatility models of the Ornstein–Uhlenbeck type. *Mathematical Finance* 13, 445–66.

Officer, R. 1973. The variability of the market factor of the New York stock exchange. *Journal of Business* 46, 434–53.

Pastorello, S., Patilea, V. and Renault, E. 2003. Iterative and recursive estimation in structural non-adaptive models. *Journal of Business and Economic Statistics* 21, 449–509.

Phillips, P. and Yu, J. 2005. A two-stage realized volatility approach to the estimation for diffusion processes from discrete observations. Discussion Paper No. 1523. Cowles Foundation, Yale University.

Renault, E. and Touzi, N. 1996. Option hedging and implied volatilities in a stochastic volatility model. *Mathematical Finance* 6, 279–302.

Roberts, G. and Stramer, O. 2001. On inference for nonlinear diffusion models using the Hastings–Metropolis algorithms. *Biometrika* 88, 603–21.

Rosenberg, B. 1972. The behaviour of random variables with nonstationary variance and the distribution of security prices. Working paper 11, Graduate School of Business Administration, University of California, Berkeley. Reprinted in N. Shephard (2005).

Shephard, N. 2005. *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press.

Smith, A. 1993. Estimating nonlinear time series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8, S63–S84.

Sørensen, M. 2000. Prediction based estimating equations. *Econometrics Journal* 3, 123–47.

Stein, E. and Stein, J. 1991. Stock price distributions with stochastic volatility: an analytic approach. *Review of Financial Studies* 4, 727–52.

Taylor, S. 1982. Financial returns modelled by the product of two stochastic processes – a study of daily sugar prices 1961–79. In *Time Series Analysis: Theory and Practice*, vol. 1, ed. O. Anderson.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Amsterdam: North-Holland.

Wiggins, J. 1987. Option values under stochastic volatilities. *Journal of Financial economics* 19, 351–72.

Yu, J. 2005. On leverage in a stochastic volatility model. *Journal of Econometrics* 127, 165–78.

Zhang, L., Mykland, P. and Aït-Sahalia, Y. 2005. A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100, 1394–411.

Zhou, B. 1996. High-frequency data and volatility in foreign-exchange rates. *Journal of Business and Economic Statistics* 14, 45–52.

How to cite this article

Shephard, Neil. "stochastic volatility models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000506> doi:10.1057/9780230226203.1625

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Onl

stock price predictability

Wayne E. Ferson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The question of predictability in stock returns has important and broad economic implications. Predictability relates directly to the efficiency of the capital markets in allocating resources to their highest valued uses. But the interpretation of predictability, and the evidence for its very existence, remain controversial. This article provides a review of the arguments and evidence for stock return predictability. The evidence for weak-form predictability, based on the information in past stock prices, is more fragile and less compelling than the evidence for semi-strong form predictability, based on publicly available information more generally.

Keywords

arbitrage; ARMA processes; asset pricing models; behavioral finance; capital markets; data mining; dividend–price ratio; efficient markets hypothesis; financial econometrics; finite sample bias; frictions; market inefficiency; martingales; momentum; multiple comparisons; predictability; regime shifts; selection bias; semi-strong form predictability; serial dependence; spurious regressions; standard error estimation; statistical econometrics; stock price predictability; stock market returns; strong-form predictability; structural breaks; trading strategies; variance ratio; weak-form predictability

Article

The interest in predicting stock prices or returns is probably as old as the markets themselves. Fama (1970) reviews early work and provides some organizing principles. This article concentrates selectively on developments following Fama's review. In that review, Fama describes increasingly fine information sets in a way that is useful in organizing the discussion. Weak-form predictability uses the information in past stock prices. Semi-strong form predictability uses variables that are obviously publicly available, and strong form uses anything else. While there is a literature characterizing strong-form predictability (for example, analysing the profitability of corporate insiders' trades), this article concentrates on the first two categories of information.

Early studies, reviewed by Fama (1970), concluded that a martingale or random walk was a good model for stock prices, values or their logarithms. Thus, the best forecast of the future price was the current price. However, predicting price or value changes, and thus rates of return, is more challenging and controversial. The current financial economics literature reflects two, often-competing, views about predictability in stock returns. The first argues that predictability represents exploitable inefficiencies in the way capital markets function. The second view argues that predictability is a natural outcome of an efficient capital market.

The exploitable inefficiencies view of return predictability argues that, in an efficient market, traders would bid up the prices of stocks with predictably high returns, thus lowering the future return and removing any predictability at the new price (for example, Friedman, 1953; Samuelson, 1965). However, market frictions or human imperfections are assumed to impede such price-correcting, or 'arbitrage' trading. Predictable patterns can thus emerge when there are important market

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

imperfections, like trading costs, taxes, or information costs, or important human imperfections in processing or responding to information, as studied in behavioural finance. These predictable patterns are thought to be exploitable in the sense that an investor who could avoid the friction or cognitive imperfection could profit from the predictability at the expense of other traders.

The 'efficient markets' view of predictability is described by Fama (1970) and has an entry in this dictionary. According to this view returns may be predictable if required expected returns vary over time in association with changing interest rates, risk or investors' risk aversion. Predictability may be expected in an efficient capital market. If the required expected returns vary over time, there may be no abnormal trading profits and thus no incentive to exploit the predictability. Write the return as $R = E(R|\Omega) + u$, where Ω is the information at the beginning of the period and u is the unexpected return. Since $E(u|\Omega) = 0$, the unexpected return cannot be predicted ahead of time. Thus, predictability, in the efficient markets view, rests on systematic variation through time in the expected return. Modelling and testing for this variation is the focus of the conditional asset pricing literature (reviewed by Ferson, 1995).

While this article focuses on stock return predictability, not all of the predictability associated with stock prices involves predicting the levels of returns. A large literature models predictable second moments of returns (for example, using ARCH and GARCH-type models; see Engle, 2004, or stochastic volatility models). Predictability studies have also examined the third moments (for example, Harvey and Siddique, 2001).

Weak-form predictability

Much of the literature on weak-form predictability can be characterized through an autoregression. Let R_t be the continuously compounded rate of return over the shortest measurement interval ending at time t. Let $r(t, t+H) = \Sigma_{j=1,\ldots,H} R_{t+j}$. Then,

$$r(t, t+H) = a_H + \rho_H r(t-H, t) + \varepsilon(t, t+H) \quad (1)$$

is the autoregression, ρ_H is the autocorrelation, and H is the return horizon. Studies can be grouped according to the return horizon. An alternative to the autoregression is the variance ratio statistic: $\mathrm{Var}\{r(t, t+H)\} / H\mathrm{Var}(R_t)$, proposed by Working, 1949, and studied for stock returns by Lo and MacKinlay, 1988, and others. Cochrane, 1988, shows that the variance ratio is a function of the autocorrelation in returns. Kaul, 1996, provides an analysis of various statistics that have been used to evaluate weak-form predictability, showing how they can be viewed as combinations of autocorrelations at different lags, with different weights assigned to the lags.

Many studies measure small but statistically significant serial dependence in daily or intra-daily stock return data. Serial dependence in daily returns can arise from end-of-day price quotes that fluctuate between bid and ask (Roll, 1984) or from non-synchronous trading of the stocks in an index (for example, Fisher, 1966; Scholes and Williams, 1977). These effects do not represent predictability that can be exploited with any feasible trading strategy. Spurious predictability due to such data problems should clearly not be attributed to time variation in the expected discount rate for stocks. On the other hand, much of the literature on predictability allows that high-frequency serial dependence may reflect changing conditional means. For example, Lo and MacKinlay (1988) and Conrad and Kaul (1988) model expected returns within the month as following an autoregressive processes.

Conrad and Kaul (1988; 1989) study serial dependence in weekly stock returns. They point out that, if the expected returns, $E(R|\Omega)$, follow an autoregressive process, the actual returns would be described by the sum of an autoregressive process and a white noise, and thus follow an ARMA process. The autoregressive and moving average coefficients would be expected to have the opposite signs: if current expected returns increase, it may signal that future expected returns are higher, but

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

stock prices fall in the short run because the future cash flows are discounted at a new, higher rate. The two effects offset and returns could have small autocorrelations. Estimating ARMA models, Conrad and Kaul find that the autoregressive coefficients for weekly returns on stock portfolios are positive, near 0.5, and can explain up to 25 per cent of the variation in the returns on a portfolio of small-firm stocks.

Even with weekly returns, however, some of the measured predictability can reflect non-synchronous trading effects. Lo and MacKinlay (1990) use statistical models that attempt to separate out the various effects in measured portfolio returns. Boudoukh, Richardson and Whitelaw (1994) use stock index futures contracts, which are not subject to non-synchronous trading, and find little evidence for weak-form predictability at a weekly frequency.

Much of the literature on weak-form predictability studies broad stock-market indexes or portfolios of stocks, grouped according to the market capitalization (size) or other characteristics of the firms. But another significant stream of the literature studies *relative* predictability. Stocks have relative predictability if the future returns of one group of stocks is predictably higher than the returns of another group. Thus, if a trader could buy the stocks in the high-return group and sell short the stocks in the low-return group, the trader can profit even if both groups go up (or down). In a weak-form version of relative predictability, past stock prices or returns are used to form the groups. If past winner (loser) stocks have predictably higher (lower) returns, we have continuation or 'momentum'. If past winner stocks can be predicted to have lower future returns, we have reversals. Relative predictability can be evaluated by viewing eq. (1) as a cross-sectional regression, an approach taken by Jegadeesh (1990), who finds continuation in monthly returns. Lehman (1990) finds some evidence for reversals in the weekly returns of US stocks.

Jegadeesh and Titman (1993) find that relatively high-return stocks over the previous year tend to repeat their performance over future three- to twelve-month horizons. They study US data for 1927–89 but focus on the 1965–89 period. The magnitude of the effect is striking. The top 20 percent winner stocks over the last six months can outperform the loser stocks by about one percent per month for the next six months. This momentum effect has spawned a huge subsequent literature which is largely supportive of the momentum effect but which has not reached a consensus about its causes. The efficient markets view of predictability suggests that momentum trading strategies should be subject to greater risk exposures which justify their high returns. Most efforts at explaining the effect by risk adjustments have failed. (There are some partial successes. For example, Ang, Chen and Xing, 2006, associate some of the momentum strategy profits with high exposure to 'downside risk', that is, likely negative returns to the strategy when the market return is negative.)

The momentum effect has inspired a number of behavioural models, suggesting that momentum may occur because markets under-react to news in the pricing of stocks. For example, one argument (Daniel, Hirshleifer and Subrahmanyam, 1998) is that traders have 'biased self attribution', meaning that they think their private information is better than it really is. As a result they do not react fully to public news about the value of stocks, so the news takes time to get impounded in market prices, resulting in momentum. In another argument, traders suffer a 'disposition effect', implying that they tend to hold on to their losing stocks longer than they should, which can lead to momentum. These arguments suggest that traders who can avoid these cognitive biases may profit from momentum trading strategies. However, Lesmond, Schill and Zhou (2004) and Korajczyk and Sadka (2004) measure the trading costs of momentum strategies and conclude that the apparent excess returns to the strategies are consumed by trading costs.

Perhaps the most controversial evidence of weak-form predictability involves long-horizon returns. Fama and French (1988) use autoregressions like (1) to study predictability in portfolio returns, measured over one-month to multi-year horizons. They find U-shaped patterns in the autocorrelations as a function of the horizon, with negative serial dependence, or mean reversion, at four- to five-year horizons. Mean reversion can be consistent with either view of predictability. If expected returns are stationary (reverting to a constant unconditional mean) but time-varying, mean reversion can occur in an efficient market. Mean reversion would also be expected if stock values depart temporarily from the fundamental, or correct prices, but are drawn back to that level. DeBondt and Thaler (1985) find that past high-return stocks perform poorly over the next five years and vice versa – a form of relative predictability. They interpret reversals in long-horizon relative returns as indicating that the market overreacts to news about stock values, and then eventually

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

corrects the mistake. The reversal effect was shown to occur mainly in the month of January by Zarowin (1990) and Grinblatt and Moskowitz (2003), which is interpreted as related to 'tax loss selling'. In this story investors sell loser stocks at the end of the year for tax reasons, thus depressing their prices, and buy them back in the new year, subsequently raising their prices. McLean (2006) finds that reversals are concentrated in stocks with high idiosyncratic risks, which is thought to present a deterrent to arbitrage traders who might otherwise correct temporary errors in the market prices.

Like momentum, behavioural models attempt to explain reversals as the result of cognitive biases. Models of Barberis, Shleifer and Vishny (1998), Daniel, Hirshleifer and Subrahmanyam (1998) and Hong and Stein (1999) argue that both short-run momentum and long-term reversals can reflect under- and overreaction to news about stock values. Research in this area continues, and it is fair to say that the jury is still out on the issue of weak-form predictability.

Semi-strong form predictability

Studies of semi-strong form predictability can be described with the regression:

$$r(t, t+H) = \alpha_H + \beta_H' Z_t + v(t, t+H), \quad (2)$$

where Z_t is a vector of variables that are publicly available by time t. Many predictor variables have been analysed in published studies, and it is useful to group them into categories. The first category of predictor variables comprises 'valuation ratios', which are measures of cash flows divided by the stock price. Keim and Stambaugh (1986) use a constant numerator in the ratio and 'detrend' the price. Rozeff (1984), Campbell and Shiller (1988) and Fama and French (1989) use dividend–price ratios, Pontiff and Schall (1998) and Kothari and Shanken (1997) use the book value of equity divided by price. Boudoukh et al. (2007) add share repurchases and other non-cash payouts, respectively, to the dividend measure. Lettau and Ludvigson (2001) propose a macroeconomic variation on the valuation ratio: Aggregate consumption divided by a measure of aggregate wealth. All of these studies find the regression coefficients β_H to be significant.

Rozeff (1984) and Berk (1995) argue that valuation ratios should generally predict stock returns. Consider the simplest model of a stock price, P, as the discounted value of a fixed flow of expected future cash flows or dividends: $P = c / R$, where c is the expected cash flow and R is the expected rate of return. Then, $R = c / P$, and the dividend price ratio *is* the expected return of the stock. If predictability is attributed to the expected return, as in the efficient markets view, then a valuation ratio should be a good predictor variable.

Predictability of stock returns with valuation ratios is also related to the expected growth rates of future dividends or cash flows. Consider the Gordon (1962) constant-growth model for a stock price: $P = c / (R - g)$, where g is the future growth rate. Then, $c / P = R - g$. This suggests that if dividend–price ratios vary, either across stocks or over time, then expected returns should vary, and/or expected cash flow growth rates should vary, and the dividend–price ratio should be able to predict one or the other. Campbell and Shiller (1988) show that the intuition from this example holds to a good approximation in models where the growth rates and expected returns are not held fixed over time. They find that market dividend–price ratios do not significantly predict future cash flow growth rates. Cochrane (2006) uses this result to re-evaluate the empirical evidence for stock return predictability using dividend–price ratios. He essentially argues that, if you know that the dividend–price ratio does not forecast future cash flow growth, then it must forecast future stock returns.

Studies of semi-strong form predictability in stock index returns typically report regressions with small R-squares, as the fraction of the variance in returns that can be predicted with the lagged variables is small – say, 10–15 percent or less for monthly to annual return horizons. The R-squares are larger for longer-horizon returns – up to 40 percent or more for four- to five-year horizons. This

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

is interpreted as the result of expected returns that are more persistent than returns themselves, as would be expected if returns are expected returns plus noise. Thus, the variance of the sum of the expected returns accumulates with longer horizons faster than the variance of the sum of the returns, and the R-squares increase with the horizon (for example, Fama and French, 1989). However, small R-squares can mask economically important variation in the expected returns.

Stocks are long 'duration' assets, so a small change in the expected return can lead to a large change in the asset value. To illustrate, consider another example using the Gordon model, where the dividend is $c = kE$, E is the earnings and k is the dividend–payout ratio. The price–earnings ratio, $P/E = 15$, the payout ratio, $k = 0.6$, and the expected growth rate, $g = 3\%$. The expected return is $R = 7\%$. Suppose there is a shock to the expected return, *ceteris paribus*. A change of one per cent in R leads to approximately a 20 per cent change in the asset value. Of course, this overstates the effect to the extent that a shock that changes the required return also changes the expected future cash flows. But the example suggests that small changes in expected returns can produce economically significant changes in asset values. Consistent with this argument, studies such as Kandel and Stambaugh (1996), Campbell and Viceira (2002) and Fleming, Kirby and Ostdiek (2001) show that optimal portfolio decisions can be affected to an economically significant degree by return predictability, even when the amount of predictability, as measured by R-squared, is small.

The second category of semi-strong form predictor variables for stock returns includes calendar and seasonal effects. The list of effects that have been related to stock returns and the list of studies are too long to cite here. (See Haugen and Lakonishok, 1988, and Schwert, 2003, for reviews.) Some examples include the season (winter versus summer), the month of the year (especially, high returns in January), the time of the month (first versus second half), holidays, the day of the week (low returns on Mondays), the time of the day, the amount of sunlight (as in seasonal affective disorder), and even the frequency of geomagnetic storms.

The third category of predictor variables in eq. (2) is a catch-all, 'other' variables. Prominent among these are bond yields and yield spreads. Fama and Schwert (1977) were among the first to observe that the level of short-term Treasury yields predicts returns in eq. (2) with a negative coefficient. They interpreted the short-term yield as a measure of expected inflation. Ferson (1989) argues that the regressions imply that the systematic risk of stocks that determines the expected returns must vary over time with changes in interest rates. Keim and Stambaugh (1986) study the yield spreads of low-quality over high-quality bonds and find predictive ability for stock returns, and Campbell (1987) studies a number of yield spreads in shorter-term Treasury securities. Another interesting set of predictor variables includes measures of the conditional variance or volatility of stock returns (for example, Merton, 1980). Fama and French (1989) assemble a list of variables from studies in the 1980s and describe their relations with US business cycles.

Of course, many other semi-strong form predictor variables have been proposed, and more will doubtless be proposed in the future. Some recent variables include the fraction of equity issues in new issues of corporate securities (Baker and Wurgler, 2000), firms' investment plans (Lamont, 2000), the average 'idiosyncratic' or firm-specific component of past return volatility (Goyal and Santa-Clara, 2003), the level of corporate cash holdings, the aggregate rate of dividend initiation (Baker and Wurgler, 2000), share issuance, and the political party currently in office (Santa-Clara and Valkanov, 2003).

Methodological issues

Even though the regressions (1) and (2) seem pretty straightforward, interpreting the predictability evidence for stock returns based on these regressions is not. It can be argued that one of the greatest contributions of the literature on stock return predictability is the methodological lessons it has taught researchers in the field. Perhaps the most difficult issues involve selection bias and data mining. Additional issues that have been addressed in the literature include small sample biases in the coefficients, standard error estimation, multiple comparisons, efficient estimation, regime shifts, spurious regression and the interactions among these effects.

Selection bias and data mining are serious concerns. Data mining refers to sifting through the data in search of predictive or associative patterns. There are two kinds of data mining. Sophisticated data mining accounts for the number of searches undertaken when evaluating the statistical significance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of the finding (for example, White, 2000). This is important because, if 100 independent variables are examined, we expect to find five that are 'significant' at the five per cent level, even if there is no predictive relation. Naive data mining does not account for the number of searches. The big problem, given the strong interest in predicting stock returns among academics and practitioners, and the many studies using the same data, is that it is difficult to account for the number of searches. There probably have been at least as many regressions run using the Center for Research in Security Prices (CRSP) database as there are numbers in the database. Compounding this problem are various selection biases. Perhaps most difficult is the fact that only 'significant' results get circulated and published in academic papers. No one knows how many insignificant regressions were run before those results were found.

A reasonable response to these concerns is to see whether the predictive relations hold out of sample. This kind of evidence is mixed. Some studies find support for predictability in step-ahead or out-of-sample exercises (for example, Fama and French, 1989; Pesaran and Timmerman, 1995). Semi-strong form variables show some ability to predict returns outside of the US data where they were originally studied (for example, Harvey, 1991; Solnik, 1993; Ferson and Harvey, 1993; 1999). Other studies conclude that predictability using many of the semi-strong form variables does not hold outside of the original samples (for example, Goyal and Welch, 2003; Simin, 2008). Even this evidence is difficult to interpret, because a variable could have real predictive power yet still fail to outperform a naive benchmark when predicting out of sample (Campbell and Thompson, 2005).

A large literature has addressed statistical issues in predictive regressions. Boudoukh and Richardson (1994) provide an insightful review. Ferson, Sarkissian and Simin (2003) study the interaction between data mining and spurious regression effects. The interaction among the various statistical issues with predictive regressions has received relatively little research to date, and this is likely to be an active field in the future.

Perspective

Does the current body of evidence lead to the conclusion that there actually is predictability in stock returns? I think there are good reasons to be sceptical of predictability and good reasons to believe in predictability.

Why should we be sceptical? First, the logic of the older efficient markets literature is compelling to many. In that view, traders would bid up the prices of stocks with predictably high returns, thus lowering their returns and removing any predictability at the new price. Furthermore, data mining and selection bias conspire to make us see predictable patterns where none may exist.

For one example, studies of relative predictability, such as momentum, have sliced and diced common stocks into portfolios based on many characteristics of the data, at which point the effect often retreats into subsets of stocks, sub-periods of time, phases of the business cycle or other parts of the data. Studies find momentum to be concentrated according to industries, the size of the firm (more momentum in large stocks), the price of the share (more when the price is above five dollars per share), and so on. The effect does not appear prior to 1940, appears stronger after 1968, and appears stronger during economic expansions than contractions.

The evidence for long-term return reversals has a similar problem. Reversals have been found to be concentrated in small stocks, low-priced stocks, the month of January, and high idiosyncratic risk stocks, and to be more pronounced in earlier samples than in more recent data. For each approach to slicing and dicing the data there is a clever story. That is not all bad. By digging into subsamples it should be possible, in principle, to isolate what is driving an effect. But many of the patterns, especially those documented in the weak-form predictability literature, appear to be sample-specific. Finding results that vary with the sample period stimulates research featuring structural breaks and regime shifts. This reader is left more with concerns about naive data mining, improper multiple comparisons and statistical issues than with an understanding of why and where these weak-form effects occur systematically.

Predictive regressions are also subject to a host of statistical issues. There are finite sample biases and problems associated with structural breaks and regime shifts. It is hard to get reliable standard errors for the regressions. There are potential spurious regression problems. And, as we are now beginning to understand, these effects can interact with each other. If semi-strong form predictability

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

is spurious, as a result of statistical bias and naive data mining, we would expect predictor variables to appear in the empirical literature, and then to fail to work with fresh data. To some extent, the literature has evolved in this way.

There are also many good reasons to believe that stock returns are predictable. First of all, theory suggests that some amount of predictability is likely. If expected returns vary over time with some degree of persistence, predictability is expected. Most people find it easy to believe that expected stock returns and risks might be different coming out of a recession, for example, than going into one, and the predictability evidence tends to confirm such common-sense patterns. Studies find that predictability using lagged variables is largely explained by asset pricing models if they allow the risk premiums to vary over time (for example, Ferson and Harvey, 1991; Ferson and Korajczyk, 1995; Avramov and Chordia, 2006). The behavioural models of predictability are compelling to many, as we see ourselves making the same cognitive errors as those made by the agents in those models.

The momentum effect has been found to hold out of sample relative to the original study of Jegadeesh and Titman (1993). Jegadeesh and Titman (2001) find momentum in data for 1990–98. Momentum is also found in stock markets outside the United States.

The evidence of semi-strong form predictability has also survived a number of out-of-sample tests, working in other countries and over different time periods. Many of the variables identified as predictors for stock returns also seem to have some predictive power for other types of securities such as bonds and futures, and also for the growth rates in ‘fundamental’ macroeconomic data. The evidence for semi-strong form predictability has survived corrections for a host of statistical and data problems.

Some studies that find semi-strong form stock market predictability, measured directly using lagged variables, has weakened in recent samples. It may be that the predictability was never really there, or that it was ‘real’ when first publicized but diminished as traders attempted to exploit it. Ferson, Heuson and Su (2005) examine semi-strong form predictability by regressing individual stocks on firm-specific predictors, then measuring the average covariances of the fitted values. It can be shown that the variance of the expected return on a large portfolio is approximately the average covariance. They find no evidence that predictability, measured in this way, is weaker in recent sub-periods. As the firm-specific predictors have not been examined extensively in the literature, they may be less subject to naive data mining biases. Campbell and Thompson (2005), using step-ahead tests, also find that semi-strong form predictability holds up in recent data.

Conclusion

The issue of predictability in stock returns has important and broad economic implications. For example, it relates to the efficiency of capital markets in allocating resources to their highest-valued uses. But the interpretation of predictability, and the evidence for its very existence, remain controversial. This reader finds the evidence for weak-form predictability (using the information in past stock prices) to be more fragile and less compelling than the evidence for semi-strong form predictability (using publicly available information more generally).

For the field of financial economics and asset pricing in particular, allowing for the possibility of predictability in stock returns through time-variation in expected returns, risk measures and volatility, has been one of the most significant developments since the mid-1980s. Such conditional asset pricing models provide a rich setting for the study of the dynamic behaviour of asset markets and problems such as the evaluation of portfolio manager performance. The issue of predictability has stimulated numerous advances in the statistical and econometric methods of financial economics. Research on predictability in asset markets is likely to continue, and remain both useful and controversial, for a long time.

See Also

- arbitrage
- data mining

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- efficient markets hypothesis
- excess volatility tests
- rational expectations
- spurious regressions
- stochastic volatility models
- time series analysis

The author acknowledges financial support from the Collins Chair in Finance at Boston College and the helpful comments of Steven Durlauf and Timothy Simin.

Bibliography

Ang, A., Chen, J. and Xing, Y. 2006. Downside risk. *Review of Financial Studies* 19, 1191–239.

Ariel, R.A. 1987. A monthly effect in stock returns. *Journal of Financial Economics* 18, 161–74.

Avramov, D. and Chordia, T. 2006. Asset pricing models and financial market anomalies. *Review of Financial Studies* 19, 1001–40.

Baker, M. and Wurgler, J. 2000. The equity share in new issues and aggregate stock returns. *Journal of Finance* 55, 2219–57.

Barberis, N., Shleifer, A. and Vishny, R. 1998. A model of investor sentiment. *Journal of Financial Economics* 49, 307–43.

Berk, J. 1995. A critique of size–related anomalies. *Review of Financial Studies* 8, 275–86.

Boudoukh, J. and Richardson, M. 1994. The statistics of long–horizon regressions revisited. *Mathematical Finance* 4, 103–20.

Boudoukh, J., Richardson, M. and Whitelaw, R. 1994. A tale of three schools: insights on the autocorrelations of short-horizon returns. *Review of Financial Studies* 7, 539–73.

Boudoukh, J., Michaely, R., Richardson, M. and Roberts, M. 2007. On the importance of measuring payout yield: implications for empirical asset pricing. *Journal of Finance* 62, 877–915.

Campbell, J.Y. 1987. Stock returns and the term structure. *Journal of Financial Economics* 18, 373–99.

Campbell, J.Y., Lo, A. and MacKinlay, A.C. 1997. *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

Campbell, J.Y. and Shiller, R. 1988. The dividend price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195–228.

Campbell, J.Y and Thompson, S. 2005. Predicting the equity premium out of sample: can anything beat the historical average? Working Paper No. 11468. Cambridge, MA: NBER.

Campbell, J.Y. and Viceira, L. 2002. *Strategic Asset allocation: Portfolio Choice for Long-Term Investors*. New York: Oxford University Press.

Cochrane, J.H. 2005. *Asset Pricing*. Princeton, NJ: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Cochrane, J.H. 2006. The dog that did not bark: a defense of return predictability. Working Paper No. 12026. Cambridge, MA: NBER.

Cochrane, J.Y. 1988. How big is the random walk in GNP? *Journal of Political Economy* 96, 893–920.

Conrad, J. and Kaul, G. 1988. Time-variation in expected returns. *Journal of Business* 61, 409–25.

Conrad, J. and Kaul, G. 1989. Mean reversion in short-horizon expected returns. *Review of Financial Studies* 2, 225–40.

Daniel, K., Hirshleifer, D. and Subrahmanyam, A. 1998. Investor psychology and security market under and over-reaction. *Journal of Finance* 53, 1839–85.

DeBondt, W. and Thaler, R. 1985. Does the stock market overreact? *Journal of Finance* 40, 793–805.

Engle, R. 2004. Risk and volatility: econometric modelling and financial market practice. *American Economic Review* 94, 405–20.

Fama, E.F. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25, 383–417.

Fama, E. and French, K. 1988. Permanent and temporary components of stock prices. *Journal of Political Economy* 96, 246–73.

Fama, E. and French, K. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25, 23–49.

Fama, E.F. and Schwert, G.W. 1977. Asset returns and inflation. *Journal of Financial Economics* 5, 115–46.

Ferson, W.E. 1989. Changes in expected security returns, risk and the level of interest rates. *Journal of Finance* 44, 1191–217.

Ferson, W.E. 1995. Theory and empirical testing of asset pricing models. In *Handbooks in Operations Research and Management Science, Volume 9: Finance*, ed. R. A. Jarrow, V. Maksimovic and W.T. Ziemba. Amsterdam: North-Holland.

Ferson, W.E. and Harvey, C.R. 1991. The variation of economic risk premiums. *Journal of Political Economy* 99, 385–415.

Ferson, W.E. and Harvey, C.R. 1993. The risk and predictability of international equity returns. *Review of Financial Studies* 6, 527–66.

Ferson, W.E. and Harvey, C.R. 1999. Conditioning variables and cross-section of stock returns. *Journal of Finance* 54, 1325–60.

Ferson, W.E. and Korajczyk, R.A. 1995. Do arbitrage pricing models explain the predictability of stock returns? *Journal of Business* 68, 309–49.

Ferson, W., Heuson, A. and Su, T. 2005. Weak and semi-strong form stock return predictability revisited. *Management Science* 51, 1582–92.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ferson, W.E., Sarkissian, S. and Simin, T. 2003. Spurious regressions in financial economics? *Journal of Finance* 58, 1393–414.

Fisher, L. 1966. Some new stock market indexes. *Journal of Business* 39, 191–225.

Fleming, J., Kirby, C. and Ostdiek, B. 2001. The economic value of volatility timing. *Journal of Finance* 56, 329–52.

Friedman, M. 1953. The case for flexible exchange rates. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Gordon, M.J. 1962. *The Investment, Financing and Valuation of the Corporation*. Homewood, IL: Irwin.

Goyal, A. and Santa-Clara, P. 2003. Idiosyncratic risk matters. *Journal of Finance* 63, 975–1007.

Goyal, A. and Welch, I. 2003. Predicting the equity premium with dividend ratios. *Management Science* 49, 639–54.

Grinblatt, M. and Moskowitz, T. 2003. Predicting price movements from past returns: the role of consistency in tax-loss selling. *Journal of Financial Economics* 71, 541–79.

Harvey, C.R. 1991. The world price of covariance risk. *Journal of Finance* 46, 111–57.

Harvey, C.R. and Siddique, A. 2001. Autoregressive conditional skewness. *Journal of Financial and Quantitative Analysis* 34, 465–87.

Haugen, R. and Lakonishok, J. 1988. *The Incredible January Effect*. Homewood, IL: Dow Jones-Irwin.

Hong, H. and Stein, J. 1999. A unified theory of underreaction, momentum trading and overreaction in asset markets. *Journal of Finance* 45, 265–95.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *Journal of Finance* 45, 881–98.

Jegadeesh, N. and Titman, S. 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance* 48, 65–91.

Jegadeesh, N. and Titman, S. 2001. Profitability of momentum portfolios: an evaluation of alternative explanations. *Journal of Finance* 56, 699–720.

Kandel, S. and Stambaugh, R. 1996. On the predictability of stock returns: an asset allocation perspective. *Journal of Finance* 51, 385–424.

Kaul, G. 1996. Predictable components in stock returns. In *Handbook of Statistics 14*, ed. G.S. Maddala and C.R. Rao. Amsterdam: North-Holland.

Keim, D. and Stambaugh, R. 1986. Predicting returns in the stock and bond markets. *Journal of Financial Economics* 17, 357–90.

Kim, M., Nelson, C.R. and Startz, R. 1991. Mean reversion in stock returns? A reappraisal of the statistical evidence. *Review of Economic Studies* 58, 515–28.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Korajczyk, R. and Sadka, R. 2004. Are momentum profits robust to trading costs? *Journal of Finance* 59, 1039–82.

Kothari, S.P. and Shanken, J. 1997. Book-to-market, dividend yield and expected market returns: a time-series analysis. *Journal of Financial Economics* 44, 169–203.

Lamont, O. 2000. Investment plans and stock returns. *Journal of Finance* 55, 2719–46.

Lehmann, B.N. 1990. Fads, martingales and market efficiency. *Quarterly Journal of Economics* 105, 1–28.

Lesmond, D., Schill, M. and Zhou, C. 2004. The illusory nature of momentum profits. *Journal of Financial Economics* 71, 349–80.

Lettau, M. and Ludvigson, S. 2001. Consumption, aggregate wealth and expected stock returns. *Journal of Finance* 56, 815–850.

Lo, A. and MacKinlay, A.C. 1988. Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies* 1, 41–66.

Lo, A. and MacKinlay, A.C. 1990. An econometric analysis of infrequent trading. *Journal of Econometrics* 45, 181–211.

McLean, D. 2006. Why do momentum and long-term reversals persist? An evaluation of mispricing explanations. Working paper, University of Alberta.

Merton, R.C. 1980. On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics* 8, 323–61.

Pesaran, M.H. and Timmermann, A. 1995. Predictability of stock returns: robustness and economic significance. *Journal of Finance* 50, 1201–28.

Pontiff, J. and Schall, L. 1998. Book-to-market as a predictor of market returns. *Journal of Financial Economics* 49, 141–60.

Roll, R.R. 1984. A simple implicit measure of the effective bid–ask spread in an efficient market. *Journal of Finance* 39, 1127–40.

Rouwehorst, G. 1998. International momentum portfolios. *Journal of Finance* 53, 267–84.

Rozeff, P. 1984. Dividend yields are equity premiums. *Journal of Portfolio Management* 1(1), 68–75.

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 41–63.

Santa-Clara, P. and Valkanov, R. 2003. The presidential puzzle: political cycles and the stock market. *Journal of Finance* 58, 1841–72.

Scholes, M. and Williams, J. 1977. Estimating beta from nonsynchronous data. *Journal of Financial Economics* 5, 309–27.

Schwert, G.W. 2003. Anomalies and market efficiency. In *Handbook of the Economics of Finance*,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

ed. G.M. Constantinides, M. Harris and R.M. Stulz. Amsterdam: North-Holland.

Sharpe, W.F. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–42.

Simin, T. 2008. The (poor) predictive performance of asset pricing models. *Journal of Financial and Quantitative Analysis*.

Solnik, B. 1993. The unconditional performance of international asset allocation strategies using conditioning information. *Journal of Empirical Finance* 1, 33–55.

White, H. 2000. A reality check for data snooping. *Econometrica* 68, 1097–126.

Working, H. 1949. The investigation of economic expectations. *American Economic Review* 39, 150–66.

Zarowin, P. 1990. Size, seasonality and stock market overreaction. *Journal of Financial and Quantitative Analysis* 25, 113–25.

How to cite this article

Ferson, Wayne E. "stock price predictability." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000529> doi:10.1057/9780230226203.1626

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Onl

stock price volatility

Stephen J. Taylor
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The volatility of a stock or stock index can be calculated either from historical prices or from the prices of option contracts. Several methods and their relative forecasting accuracy are reviewed. The most accurate methods require either very frequent price measurements or option prices for several strikes.

Keywords

ARCH models; asset price variability; Black–Scholes formula; conditional volatility; forecasting; implied volatility; long memory; options; realized volatility; risk premium; smile effects; stochastic volatility models; stock price volatility; value at risk; volatility clustering

Article

Volatility is a measure of asset price variability over some period of time, which typically describes the standard deviation of an investment's return in a particular context. The measurement and prediction of volatility is important when designing an equity portfolio. It is also important for risk managers, particularly when they calculate ‘value at risk’.

Volatility has been defined in at least five different ways. The simplest of these regards volatility as a constant numerical parameter σ, associated with the assumption that investment returns during identical durations of time have both identical standard deviations and independent distributions. This assumption is contradicted for stock prices by substantial changes in price variability as time progresses. These changes are often referred to as the phenomenon of volatility clustering; as first noted by Mandelbrot (1963, p. 418), this is the property of prices that ‘large changes tend to be followed by large changes – of either sign – and small changes tend to be followed by small changes’.

A second definition of volatility views it as a stochastic process, whose dynamic properties and parameters can be estimated from historical time series (see stochastic volatility models). This entry covers realized volatility, conditional volatility and implied volatility, and it concludes with observations about forecasting methods. Chapters about all of these topics and stochastic volatility are included in Taylor (2005).

An answer to the fundamental question ‘Why does stock price volatility change?’ has proved elusive. Occasional crises explain some of the higher observed levels of volatility. These crises may be economic, political or financial, examples being the Great Depression, the Watergate tapes episode and the crash of 19 October 1987. Macroeconomic variables, such as inflation, unemployment and GNP, have an impact upon equity volatility. Scheduled announcements about these variables usually coincide with a very short period of higher volatility, but at other times macro variables explain only a small proportion of the variability of volatility. Stock volatility depends partially on the level of the market. When prices fall, the value of firm equity relative to debt increases, financial leverage increases and it is observed that volatility increases on average. Volatility is positively correlated with trading volume, but this does not imply a causal relationship

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

between these two variables. Instead, information flows are likely to have a common impact upon both volatility and volume – as the amount and the importance of information increase both volatility and volume should theoretically increase.

Realized volatility

Investment returns can be defined by changes in the logarithms of stock prices, with appropriate adjustments for dividend payments. The standard deviation of a set of n returns calculated from the latest price and n previous prices provides a simple measure of realized or historical volatility. These standard deviations are often calculated from daily closing prices and scaled to provide an annualized quantity. The conversion from daily to annual volatility requires multiplication by $\sqrt{N}$, with N the number of daily returns available in one year. A typical level in the United States of the annualized volatility is 15 per cent for a diversified stock index, 30 per cent for a large firm and 45 per cent for a randomly selected firm.
More accurate realized volatilities can be calculated from more frequent price observations. Andersen et al. (2001) essentially evaluate standard deviations for five-minute returns, for each stock in the Dow Jones Industrial Average index. They show that realized volatility has an approximate lognormal distribution and a long-memory property. The distribution of daily returns, conditional upon their realized volatility, is approximately normal. This contrasts with the high peaks, fat tails and excess kurtosis of the unconditional distribution of returns, all of which are created by the stochastic behaviour of volatility.

Conditional volatility

A conditional variance for a future return can be calculated from historical returns and a time series model, whose structure and parameters can respectively be identified and estimated from previous returns. A massive literature on autoregressive conditional heteroscedasticity (ARCH) commences with Engle (1982). The most appropriate simple choice for the conditional variance h_t of the daily stock return r_t during period t is given by the following recursive equation of Glosten, Jagannathan and Runkle (1993):

$$h_t = \omega + (\alpha + \gamma d_{t-1})(r_{t-1} - \mu)^2 + \beta h_{t-1} \quad (1)$$

with

$$d_{t-1} = 1 \text{ if } r_{t-1} < \mu \text{ and } d_{t-1} = 0 \text{ otherwise.} \quad (2)$$

The expected return μ and the volatility parameters ω, α, β and γ can all be estimated easily by maximizing the log-likelihood function for some historical set of returns.
The dummy variable d_{t-1} in (2) is important when modelling stock indices. It is required to accommodate the asymmetric tendency for volatility to rise more sharply when stock market prices fall significantly than when they rise by the same percentage amount. Most US estimates of the asymmetry ratio, $(\alpha + \gamma) / \alpha$, exceed three, and values as high as nine have been estimated from recent data. Possibly the best explanation for the asymmetric effect is that the correlations between the returns from different firms tend to increase when the market falls.
The duration of volatility shocks is characterized by the persistence parameter, $\varphi = \alpha + \beta + 0.5\gamma$ for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(1), that is often estimated above 0.97 for series of daily returns. Volatility does not appear to be a unit-root process. Typical estimates of the half-life parameter, $H = \log(0.5) / \log(\varphi)$ trading periods, vary from a few months to one year.
Typical annualized standard deviations for the above model can be calculated from daily S&P 100 index returns between 1991 and 2000: they vary from 7 per cent to 47 per cent, with interquartile range from 10 per cent to 18 per cent, and median equal to 13 per cent. The average level rises from 11 per cent for 1991–95 to 18 per cent for 1996–2000, reflecting the dotcom bubble, and it has fallen since 2000.
We note that many other ARCH models have been proposed for stock volatility. Including a long-memory component often appears to improve the statistical fit, although the calculations and the mathematical structure are then far more complicated.

Implied volatility

Option markets provide a rich source of information about market beliefs concerning future volatility. The Black–Scholes formula for a European call option assumes that stock prices follow a special continuous-time process that has constant volatility σ. The formula provides a theoretical price $c(S, T, K, r, q, \sigma)$ that depends on the unknown quantity σ and five observable variables: the current stock price S, the time until expiry of the option T, the strike price K, the risk-free interest rate r and the dividend yield q (here assumed constant for expositional convenience). As the theoretical price is an increasing function of σ, for any market price c_M that excludes arbitrage profits there is a unique value of σ that solves the equation $c(S, T, K, r, q, \sigma) = c_M$. This unique solution is called the implied volatility. Similar definitions apply to put and to American options.
At any moment in time a matrix of implied volatilities can be calculated, with the rows and columns respectively defined by the two contractual parameters, T and K. The entries in an empirical matrix are not identical as predicted by Black–Scholes theory, primarily because their assumed continuous-time process oversimplifies reality. The mean-reverting, stochastic behaviour of volatility creates term effects – for a fixed K, the implied volatilities tend to increase (decrease) as T increases when their level is below (above) the long-run average. The deviation of future price distributions from the lognormal shape and the level of the correlation ρ between price and volatility shocks creates so-called smile effects – for a fixed T, implied volatilities decrease as K increases for stock indices (as $\rho < 0$) but can define a U-shaped function for individual securities (as ρ can be near to zero) (see also stochastic volatility models)
Implied volatilities are on average slightly higher than realized or conditional volatilities. One explanation is that option prices incorporate risk premia that compensate their owners for bearing the risk that volatility changes and/or prices jump.

Volatility forecasts

Historical information about prices can be converted into forecasts of future volatility, from the current time until any chosen future time, by using time-series methods to extrapolate from current and past values of either realized or conditional volatilities. Implied volatilities can also be converted into forecasts. Conceptually they have the advantage that they can incorporate beliefs about volatility changes arising from future events. Set against this are two issues: (*a*) the necessity of extracting a single number from a matrix of candidate values, which refer to option lifetimes that rarely match the forecast horizon, and (*b*) the tendency of unadjusted implied volatilities (that embed risk-neutral probabilities) to be biased predictors of future realized volatilities.
A high proportion of the empirical evidence shows that option-based forecasts of realized index volatility are more accurate than historical forecasts, particularly when the forecast horizon is at least one month. Blair, Poon and Taylor (2001) show that an option-based forecast calculated from a weighted average of several implied volatilities outperforms historical forecasts that utilize daily and more frequent returns. Jiang and Tian (2005) obtain the same conclusion when they use a model-free measure of implied volatility, which provides a theoretically optimal method for extracting the market's risk-neutral variance expectation from a complete set of option prices.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- ARCH models
- options
- stochastic volatility models

Bibliography

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. 2001. The distribution of realized stock return volatility. *Journal of Financial Economics* 61, 43–76.

Blair, B.J., Poon, S.-H. and Taylor, S.J. 2001. Forecasting S & P 100 volatility: the incremental information content of implied volatilities and high frequency index returns. *Journal of Econometrics* 105, 5–26.

Engle, R.F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

Glosten, L.R., Jagannathan, R. and Runkle, D. 1993. Relationship between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48, 1779–801.

Jiang, G.J. and Tian, Y.S. 2005. The model-free implied volatility and its information content. *Review of Financial Studies* 18, 1305–42.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Taylor, S.J. 2005. *Asset Price Dynamics, Volatility, and Prediction*. Princeton: Princeton University Press.

How to cite this article

Taylor, Stephen J. "stock price volatility." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000532> doi:10.1057/9780230226203.1627

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Onl

Stockholm School

Björn Hansson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cassel, G.; comparative statics; disequilibrium approaches; dynamic methods; equilibrium approaches; ex ante and example post; expectations; Hammarskjöld, D.; intertemporal equilibrium; Lindahl, E. R.; Lundberg, E. F.; Myrdal, G.; normal rate of interest; Ohlin, B. G.; perfect foresight; price level; sequence analysis; Stockholm School; temporary equilibrium; Wicksell, J. G. K.

Article

The Stockholm School existed between the years 1927 and 1937 as a separate and distinctive school of economics. The School was given its name by Ohlin in his review of the *General Theory* (see Ohlin, 1937). A ‘School’ is here defined as an interrelated development of a common theme among its members, who included Myrdal, Lindahl, Hammarskjöld, Ohlin and Lundberg. The common theme is the development of *dynamic methods*, which refer to notions such as temporary equilibrium and intertemporal equilibrium.
The development of dynamic methods is an original contribution. There is no evidence that the members were influenced in any significant way by other contemporary economists. After 1937 there are clear indications that the Swedish economists took up ideas from Hicks and Samuelson (cf. Hansen, 1951). The Stockholm School generally applied their concepts to macroeconomic problems and there are similarities to Keynes's *General Theory*, but they did not construct the principle of effective demand (Hansson, 1982; Landgren, 1960; Patinkin, 1982; Steiger, 1971).

Equilibrium approaches

The first contribution to the development of dynamic method is Myrdal's dissertation from 1927, *Prisbildningsproblemet och föränderligheten* (The problem of price formation and change). His method includes anticipations, that is, among the data or the immediate determinants of relative prices. The direct stimulus to this idea seemed to have come from what Myrdal considered to be Cassel's incomplete handling of the dynamic problem.
The inclusion of anticipations implies that future anticipated changes have effects on the economic process long before they actually take place. The theoretical determination of an equilibrium has therefore to include the anticipated consequences of probable changes. Myrdal's method of putting the anticipated effects alongside the other data has been called ‘the method of expectation’, which means the inclusion of expectations as explicit variables in a formal equilibrium theory (cf. Hicks, 1973, p. 143 n. 11).
Lindahl's construction of intertemporal and temporary equilibrium are both examples of an *equilibrium approach*, which means that for each individual and commodity the anticipated price achieves a balance of demand and supply and all expectations are therefore fulfilled. Lindahl also considered the assumption of perfect foresight a necessary condition for the determination of a price situation as a state of equilibrium.
Lindahl's aim in the article ‘The Place of Capital in the Theory of Price’ (1929) was to analyse the effects of including capital goods in the determination of a static equilibrium. Intertemporal

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

equilibrium emerges from the shortcomings of comparative statics in handling some of the problems related to this analysis and it is not aimed at solving macro-theoretical problems. Hayek had already constructed intertemporal equilibrium in 1928 but there is no evidence that Lindahl at this time was aware of Hayek's contribution.

Intertemporal equilibrium can analyse dynamic conditions, since there is a movement in the system. However, the mathematical formulation shows that there is no 'movement' in any meaningful sense, since it is a simultaneous determination of prices, quantities and interest rates for all periods under the assumption of equilibrium within each period. Lindahl's development of temporary equilibrium grew out of this inadequacy of intertemporal equilibrium. He looked for a method which should include the analysis both of relative prices in each period and of the price relations between different periods, and temporary equilibrium was meant to solve this problem.

The notion of a period plays an important role both in intertemporal and temporal equilibrium as methods to analyse dynamic conditions. The principal difference between dynamic and stationary conditions is that in the former case the factors determining the prices – the data – are constantly changing. The dynamic case implies practically continuous changes in data but Lindahl assumed for analytical reasons that:

> In order to analyse such a dynamic process, we imagine it to be subdivided into periods of time to short that the factors *directly* affecting prices, and therefore also the prices themselves, can be regarded as *unchanged in each period*. All such changes are therefore assumed to take place at the transition points between periods. The development of prices can then be expressed as a series of successive price situations. (Lindahl, 1930, p. 158)

To divide the dynamic process into periods was in the beginning just introduced as an heuristic device, but later on Hammarskjöld gave it an analytical base.

The application of temporary equilibrium implies that for each period the prices are in equilibrium states in the sense that there will be equality between supply and demand during the period; the determination of prices is expressed as a system of equations for each period. The dynamic process is then analysed as a series of temporary equilibria. It shows that temporary equilibrium was developed before the publication of Hicks's *Value and Capital* (1939).

Lindahl's analysis was almost immediately criticized by both Lundberg and Myrdal. Lundberg's criticism relates only to intertemporal equilibrium but his critique is also valid for temporary equilibrium. The main weakness is that the accommodations to the disturbances are unexplained and outside the model, and the successive sequence of equilibria is therefore not explained (cf. Lundberg, 1930, p. 157; Myrdal, 1939, p. 122). The missing element is an analysis of the accommodation process, that is to say to explain the 'link' between consecutive periods. In fact, Lindahl himself later described his method as introducing 'dynamical problems within the static framework' (Lindahl, 1939a, p. 10), since this method could employ the entire static apparatus for the analysis of a dynamic sequence.

Disequilibrium approaches

In *Monetary Equilibrium* (1931) Myrdal attempted a critical reconstruction of Wicksell's normal rate of interest; the starting-point was Lindahl's analysis in *The Rate of Interest and the Price Level*. However, the most important contribution to the School is the construction of the famous *ex ante/ex post* calculus. The *ex ante* anticipations are the driving force in the dynamic process, but the *ex post* results do still play a role since they are a basis for the forthcoming *ex ante* calculations.

The method is constructed in such a way that there is always an *ex post* balance. But the interesting problem is to analyse the changes during the period which are required to bring about the *ex post* balance. These changes must be the result of inconsistent anticipations or due to exogenous changes during the period. It was precisely Lindahl's insufficient analysis of the balancing changes that was criticized by Myrdal and by Lundberg: the intervening changes could not be explained because time is divided into a number of short equilibrium periods during which no changes occur. However, with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the application of *ex ante* and *ex post*, which makes a proper analysis of the intervening changes, it is possible to be released from the straitjacket of the equilibrium approach implied by temporary equilibrium.

Hammarskjöld was the first within the Stockholm School to give an explicit explanation of the means by which two periods may be connected in a disequilibrium approach. Both Lindahl and Myrdal held, with greater or lesser clarity, that unexpected changes in the price level and the concomitant income changes would keep the process going, but their equations have no formal connections with plans in subsequent periods. Before Hammarskjöld nobody within the Stockholm School had shown that there was a relation between fixed plans and the length of the period, but after his contribution it was generally assumed that the duration of unchanged plans determines the length of the unit period.

Like the other members of the Stockholm School, Ohlin was interested in explaining changes in the price level, but he also wanted to give the most general analysis of the driving force behind price movements. His explanation is a restatement of an old Wicksellian idea. An explanation of price changes must look at the factors influencing the demand and supply of both consumption goods and investment goods. It is therefore not surprising that Ohlin found the idea of the disequilibrium in the capital market as representing the driving force – the hallmark of the Wicksellian approach – as a too narrow explanation. In particular, since he could construct examples where a disequilibrium in the consumer goods market was the main factor behind a price movement.

Ohlin's specific extension of Wicksellian macro-theory was his explicit treatment of *autonomous* changes in consumption, which seemed to have been Ohlin's own evaluation of his contribution to the Stockholm School (cp. Steiger, 1976, p. 356 n.22). But he made no significant contribution to the dynamic method.

Sequence analysis

Lindahl's ‘Note on the dynamic pricing problem’, a short (four pages) and privately circulated paper from 1934, lays the foundation of sequence analysis. His unpublished manuscript ‘Introduction to the theory of price movements in a closed community’ (1935) gives his idea on a general dynamic approach, which should be the basis for all theories, and most of these ideas were later reproduced in ‘The Dynamic Approach to Economic Theory’ (1939b).

To define a determinate sequence it is necessary to prove that certain human actions will necessarily follow from a definite situation at a given point of time. However, human actions or the results of human actions are not deterministic and can not be explained in the same way as events in the physical world, but this is only a problem for the empirical relevance of the theory and it does not show that the theory is inconsistent. The construction of a dynamic theory ‘solves’ the problem. To develop his general dynamic approach Lindahl had to postulate that individual actions represent the fulfilment of certain plans, which at the beginning of the period are determined according to explicit principles. It obviously takes a certain amount of time to realize these actions and a period of time is resolved. As far as changes in plans are concerned, Lindahl assumed that these take place at the transition point between two consecutive periods. It is implied that a period is defined by unchanged plans, which was already hinted at by Hammarskjöld. The notion of plan had played a central role within the Stockholm School almost from its beginning, but now it is explicitly stated as the pivot for the dynamic method.

To develop sequence analysis Lindahl assumed: at an arbitrary point in time (t) the plans for production and consumption are given for a certain period of time (t to t+1), which means that if the prices are known then the individual actions are determined for the period; the supply prices are given during the period and all changes in prices and plans take place at the transition point between two consecutive periods, that is, a *fixprice* method. The problem to be solved is then the following:

> the analysis of what happens during the said period, that is, the determination of the situation at the point t+1 as it *results* from the situation at the point t. When this problem is solved, the situation at the point t+2 can in the same manner be explained as it results from the situation at the point t+1, and so on. The solution of this problem implies,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

therefore, […], the solution of the whole dynamic problem. (Lindahl, 1934, p. 204)

This solution contributes to the 'whole dynamic problem' in the sense that the same method could be employed for the period (t+1 to t+2) at point t+1. This is an example of a *single-period analysis*, which shows that is going to happen during one single period; how certain *ex ante* plans at the beginning of the period lead to determinate *ex post* results at the end of the period. But it does not determine the whole dynamic process (from t to t+1 and onwards) at point t. It is also necessary to have a *continuation analysis* to show the effects of the *ex post* results in the current period on the plans for the subsequent period. It is obvious that both parts are necessary components in a sequence analysis which is supposed to determine a process spanning several consecutive periods.
To solve the single-period analysis Lindahl assumed that the spending plans are assumed to be realized as far as quantities are concerned; it is implied that there are enough unemployed factors of production and sufficient stocks of goods and therefore all adjustments during a period can take place via changes in stocks. The role of this assumption is to show that the actions during the period can be directly deduced from the plans once the prices are given at the beginning of the period. Hence, the result of the period is a necessary outcome of the *ex ante* plans for the current period.
To determine the *ex post* results from the given *ex ante* plans at time t does not imply that the ongoing process after time t+1 can be determined at t without further assumptions. The crucial assumption of the continuation analysis concerns the relation between *ex ante* plans for the forthcoming period and the *ex post* results for the current period. Lindahl now assumed: if the plans are fulfilled, and there are no changes in the exterior events, then it is possible to postulate a simple functional relation between the *ex ante* plans for the next period and the *ex post* results of the current period. The application of this assumption for several periods gives as a result that the whole dynamic process can be deduced from the data given at the beginning of the first period. Thus, in Lindahl's analysis each single period is in equilibrium and this type of sequence analysis can be interpreted as a sort of moving equilibrium. Lundberg left out the postulate that the plans have to be fulfilled within each period, which is the difference between equilibrium and disequilibrium sequence analysis.
Lundberg's development of sequence analysis, in *Studies in the Theory of Economic Expansion* (1937), used a disequilibrium approach but at the same time it is an equilibrium process. It is this two-sided character of disequilibrium and equilibrium in Lundberg's sequence analysis which is important to understand.
Lundberg criticized static equilibrium but such equilibrium notions might still be useful to explain the development of aggregate relations. In fact, if a sequence analysis should be possible then some equilibrium relations must hold even out of equilibrium, and equilibrium constructions play therefore a fundamental role in sequence analysis. The question is to what extent equilibrium constructions may be used in a sequence analysis which starts with aggregate categories. Lundberg's analysis is a *disequilibrium sequence analysis*, while Lindahl in 1934 pursued an *equilibrium sequence analysis*. In Lundberg's analysis the equilibrium notion is represented by the fixed response functions – an expectation function of a constant form – which is presupposed for the existence of equilibrium through time (cf. Hahn, 1952, p. 804). Sequence analysis belongs therefore to the class of equilibrium processes, since constant expectation functions imply that behaviour is invariant over a certain period of time which is a crucial aspect of a general notion of equilibrium. Hence, disequilibrium sequence analysis is an equilibrium process, but, at the same time, it is a disequilibrium approach since expectations are not fulfilled within each period.

Assessment

After the late 1930s there was almost no contribution to the development of dynamic method from the original members of the School. Most of them had also left the world of pure academics: Myrdal worked on the problem of black people in the United States in the early 1940s and was later minister for trade in the social democratic government; Hammarskjöld was already in the mid-1930s a prominent civil servant; Ohlin became leader of the Liberal Party in the early 1940s; Lundberg was in 1937 appointed as the first director of the Swedish Business Cycle Research Institute; Lindahl was

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the only one who pursued a pure academic career and now and then he would dwell on his original ideas, but there was no further development.

See Also

- Lindahl, Erik Robert
- Myrdal, Gunnar
- Ohlin, Bertil Gotthard
- Wicksell, Johan Gustav Knut

Bibliography

Hammarskjöld, D. 1933. *Konjunkturspridningen*. Stockholm: P.A. Norstedt.

Hansen, B. 1951. *A Study in the Theory of Inflation*. London: George Allen & Unwin.

Hansson, B.A. 1982. *The Stockholm School and the Development of Dynamic Method*. London: Croom Helm.

Hicks, J.R. 1973. Recollections and documents. *Economica* 40, 2–11. As reprinted in J.R. Hicks, *Economic Perspectives*, Oxford: Clarendon Press, 1977.

Landgren, K.-G. 1960. *Den 'nya ekonomien' i Sverige*. Stockholm Almqvist & Wiksell.

Lindahl, E. 1929. Prisbildningsproblemets uppläggning fra˙n kapitalteoretisk synpunkt. *Ekonomisk Tidskrift* 31, 31–81. Trans. as 'The Place of Capital in the Theory of Price', in Lindahl (1939a).

Lindahl, E. 1930. *Penningpolitikens medel*. Malmö: Förlagsaktiebolaget. Trans. as 'The Rate of Interest and the Price Level', in Lindahl (1939a).

Lindahl, E. 1934. A note on the dynamic pricing problem. Stencil, Gothenburg, 23 October 1934. As reprinted in Steiger (1971).

Lindahl, E. 1935. Introduction to the theory of price movements in a closed community. Chapter 1 (the only surviving) of the planned book *Monetary Policy and its Theoretical Basis*. Unpublished manuscript from the beginning of 1935.

Lindahl, E. 1939a. *Studies in the Theory of Money and Capital*. London: George Allen & Unwin.

Lindahl, E. 1939b. The dynamic approach to economic theory. Part I of Lindahl (1939a).

Lundberg, E. 1930. Om begreppet ekonomisk jämvikt. *Ekonomisk Tidskrift* 32(4), 133–60.

Lundberg, E. 1937. *Studies in the Theory of Economic Expansion*. Reprinted, New York: Kelley & Millman, 1955.

Myrdal, G. 1927. *Prisbildningsproblemet och föränderligheten*. Uppsala and Stockholm: Almqvist & Wiksell.

Myrdal, G. 1939. *Monetary Equilibrium*. A translation and extended version of the German edition of 1933, which is a translation and an extension of an article in *Ekonomisk Tidskrift* (1931). Reprinted, New York: Kelley, 1965.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ohlin, B. 1933. Till frågan om penningteoriens uppläggning. *Ekonomisk Tidskrift* 35, 45–81. Trans. as ‘On the Formulation of Monetary Theory’, *History of Political Economy* 10 (1978), 358–88.

Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment. *Economic Journal* 47, Part I, March, 53–69; Part II, June, 221–40.

Patinkin, D. 1982. *Anticipations of the General Theory*? Oxford: Basil Blackwell.

Steiger, O. 1971. *Studien zur Enstehung der neuen Wirtschaftslehre in Schweden*. Berlin: Duncker & Humblot.

Steiger, O. 1976. Bertil Ohlin and the origins of the Keynesian revolution. *History of Political Economy* 8, 341–66.

How to cite this article

Hansson, Björn. "Stockholm School." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000288> doi:10.1057/9780230226203.1628

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Onl

stocks and flows

Glenn W. Harrison
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

auctioneer; flows and stocks; Hicks, J. R.; Marshall. A.; price determination; rational expectations; rational expectations equilibrium; stocks and flows; temporary equilibrium

Article

The essence of stock–flow analyses of individual or market behaviour is an explicit recognition of the interdependence of current production, consumption and asset-holding plans. The term ‘stock–flow analysis’ is used here generically to refer to any theories dealing simultaneously with the economic activities of production, consumption and asset-holding. At a market level this implies a distinction between plans to purchase a good in the current period in order to consume it during the current period (flow demand) and for the purpose of holding it at the end of the current period as an asset (stock demand), and an analogous distinction between plans to supply the good from current production (flow supply) and past production (inventories, or stock supply). In this article we review the nature of the distinction between stocks and flows and then illustrate the importance of the distinction for alternative theories of market price determination. Detailed references to the folklore of the stock–flow literature may be found in Burstein (1982), Bushaw and Clower (1957), Clower (1968) and Harrison (1980). We focus here on markets: see Archibald and Lipsey (1958), Clower (1963), Clower and Burstein (1960), Hadar (1965; 1971, ch. 11) and Johnson (1971) for important contributions to theories of individual behaviour in a stock–flow environment.

Whether acknowledged or not, there are three respects in which stocks and flows are commonly distinguished. We shall refer to these, for convenience, as the dimensional, behavioural and heuristic distinctions. It is crucial to keep these conceptually separate. For purposes of exposition we shall discuss the analytical characterization of the trading and exchange processes pertaining to a market period of given length. For convenience and familiarity presume that market-price determination occurs instantaneously at discrete points of time spaced equally apart in calendar time. Hicks (1946; ch. 9) chose to visualize this point at the beginning of each ‘week’. Providing we do explicitly identify some such reference date, there is no reason to visualize it at any particular point during the week. For the sake of tradition we follow Hicks's vision of traders coming together at the beginning of the week. Their trading plans, which we shall characterize precisely in a moment, are presumed to be coordinated in some fashion by the activities of an auctioneer in setting a particular price vector that is to prevail until the next meeting at the Bourse (the beginning of the week). For present purposes we shall assume that prices are set so as to satisfy the Hicksian temporary equilibrium condition that planned purchases equal planned sales.

By viewing what happens at the Bourse as happening at a point in time we are led to characterize trading plans, for whatever purpose, *dimensionally* as stocks. At a given price a trader plans to buy or sell a particular quantity of a good; when the auctioneer cries out a particular price for a particular good, the traders each respond with no more than a positive or negative number (positive for purchase offers and negative for sales offers, let us presume). The auctioneer is not in the least interested in what the trader plans to do with the bottle of whisky he bid for – for all he cares, the trader may sip it continuously for the ensuing week, gulp it down completely at some instant during

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the week, or just add it to his stockpile. It is only his *trading* plans that are relevant to the price determination that occurs at the Bourse, and they are unambiguously measured at the instant the Bourse is open. Thus we have the first stock–flow distinction – trading plans are dimensionally stocks (measured at some reference date) and not flows (measured per some market period).

The illustration of the opening of the Bourse is just one of many possible characterizations. Such a conception permits us to focus on the important implications of a behavioural stock–flow distinction implicit in many writings. Referring to his own temporary equilibrium condition, Hicks argues (1965; p. 85) that:

> As long as we hold to the principle of price determination by 'equilibrium of demand and supply', on which that theory is based, we have no call to attend to anything but transactions. We do not need to distinguish between stocks and flows; for stocks and flows enter into the determination of equilibrium in exactly the same way.
>
> There can, in competitive conditions, be no more than one price for the same commodity at the same time; and even in conditions that are only partially competitive, it does not have one price at stock and another as flow. The supply and demand that are equated, in the single period of Temporary Equilibrium theory, may (and probably will) contain stock elements as well as flow elements. Supply comes partly from stock carried over, partly from new production; demand is partly a demand for carry-forward. Expectations of futures prices affect both elements; interest affects both elements. The analysis does not require that stock and flow should be separated into compartments. It is not the case that there is one stock equilibrium and one flow equilibrium. There is one 'stock–flow' equilibrium of the single period; and that is all.

The stock–flow distinction implicit in Hicks's discussion is the behavioural one that calls the activities of current period production and consumption 'flow activities' and the activity of asset-holding a 'stock activity'. It cannot be emphasized too strongly that the terms stock and flow are in this respect merely euphemisms for asset-holding and production/consumption behaviour, respectively. Thus we shall distinguish those trading plans related to asset-holding plans from those trading plans related to current production and/or consumption plans. We may refer to the former as excess stock demand and the latter as excess flow demand.

There are two important reasons for wanting to make such a behavioural distinction. The first concerns the alternative ways in which we may then choose to characterize trading plans as being 'coordinated'. In terms of the Hicksian temporary equilibrium condition, market price is set such that at that price excess market demand (the sum of excess stock and excess flow demand) is zero. In terms of the Marshallian temporary equilibrium condition, market price is set such that excess stock demand is zero. Indeed, a failure to recognize the behavioural stock–flow distinction has led some authors to fail to see that these (and other) price determination conditions are alternatives; these implications of the behavioural stock–flow distinction are taken up below.

The second reason for making the behavioural distinction is that it raises the important difference between *temporary* (current period) market equilibrium and *full* (stock–flow, stationary, long-run static) market equilibrium. Clower (1968) argues that this may be the only reason for adopting a stock–flow analysis of market behaviour. This point may be put more accurately by saying that a recognition of stock and flow trading plans leads to a non-trivial explicit dynamic model of market behaviour from period to period. The overwhelming majority of received theory is written in terms of stock *or* flow trading plans; in such cases, a Hicksian temporary equilibrium in any period is, *ceteris paribus*, a full equilibrium. To account for observed non-stationary time series of market prices and/or inventories, recourse in such cases must be had either to persistent exogenous shocks to the system and/or 'ad hoc' adjustment lags in production or expectations formation.

We now highlight the point of separating the dimensional and behavioural stock–flow distinctions. When describing the Hicksian temporary equilibrium condition, we shall talk about flow trading plans rather than production and consumption plans. The reason for doing so is that we may wish to visualize the latter plans as being dimensionally flows. Similarly, we may wish to visualize asset-holding plans as being dimensionally stocks – we may accordingly choose to reference-date stock demand as planned holdings as at the end of the period and stock supply as existing inventory at the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

beginning of the period. The point is that '… we have no call to attend to anything but transactions', and all trading plans are dimensionally comparable, obviously leaving the theorist great scope in terms of the 'real' economic processes he may wish to discuss in terms of this scenario of market price determination. This scope is, however, often restricted by a third heuristic stock–flow distinction in common use.

The heuristic distinction asserts that current production and consumption are economic activities occurring in 'real time', while portfolio formation plans are not. It is often an implicit presumption, based intuitively but not necessarily on the behavioural stock–flow distinction. The rate at which whisky may be currently consumed at some price is, let us say, fixed in terms of calendar time units by taste considerations. Hence, it is argued, the number representing flow demand trading plans at that price varies directly with the market period length adopted. At all prices I, for example, like to sip one bottle of whisky a day; if the market opens once a week my (perfectly inelastic) flow demand is seven; if it opens daily, it is one. In each case my planned purchases remain dimensionally a stock. With respect to asset–holding behaviour it is argued conversely that the number representing (end of market period) stock demand is invariant to the market period length adopted. One thinks perhaps of speculators who are only concerned with inter-period price changes. It makes no difference to them, so it is argued, if the price changes daily or weekly (storage costs aside). The argument is far less plausible in the case of whisky retailers, however. The heuristic stock–flow distinction involves, in Patinkin's words,

> implicitly defining a 'flow' not as a quantity whose dimensions are $1/T$, but as a quantity whose magnitude is directly proportionate to h [the length of the market/planning period]; similarly, the implicit definition of a 'stock' is that a quantity whose magnitude is independent of h. Clearly such 'stocks' and 'flows' can be added together. (Patinkin, 1965, p. 521)

We might just add at the end of this quote: '… for a period of given, fixed length with respect to the conceptual experiments underlying these relations'. An obvious extension to this received presumption is to consider the factors influencing the cost of adjustment of portfolios in real time (cf. Clower and Howitt, 1978). With respect to production, consumption and portfolio formation, we may think of transaction costs (as distinct from learning by doing and indigestion) as common to all three activities. This extension implies, for the time being, an elaboration rather than refinement of the basic price theory under discussion here. It does clarify, however, the basis (such as it is) of a stock–flow dichotomy in the literature, focusing on price determination *at an instant* (for example, the approach of Tobin, 1969, may be viewed as an application of the Marshallian temporary equilibrium condition along with the explicit assumption of a perfectly elastic flow supply schedule). Knowing this, one would hardly wish to make this received presumption an article of modelling faith.

We now consider the importance of the behavioural stock–flow distinction for alternative theories of market price determination. For expository purposes consider an isolated market model for some good which is, with respect to the market period length adopted, capable of being produced, consumed and held as an asset. Such a good is referred to as a stock–flow good. A perishable good which may only be currently produced and currently consumed is called a pure flow good and a non-augmentable durable a pure stock good. Define the following notation with respect to any market period t: D_t: planned stock demand to hold the good at the end of the current period; S_t: existing stock supply of inventory of the good at the beginning of the current period; d_t: planned flow demand to consume the good over the current period; s_t: planned flow supply of current period production of the good; p_t: money-price of one unit of the good in the current market period; x_t: excess flow demand $\equiv d_t - s_t$; and X_t: excess stock demand $\equiv D_t - S_t$. Stock supply in period t is defined as the accumulated sum of past excess flow supply: $S_t \equiv s_0 - x_{t-1} - x_{t-2} - \ldots - x_0$.

The alternative temporary equilibrium conditions introduced earlier are defined as follows:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\text{Hicksian:}\quad x_t + X_t = 0;$$

and

$$\text{Marshallian:}\quad X_t = 0.$$

It is clear from these specifications of the two temporary equilibrium conditions that they do not in general predict that the current market price will be the same value. Moreover, the periodic rate of accumulation of inventory also depends on the theory employed. This accumulation process feeds back on the market price in period t+1, due to an induced shift of the excess stock demand schedule. The properties of the first-order difference equations implied by the alternative theories are well known: Bushaw and Clower (1957, chs 3 and 4) discuss the Hicksian condition, and Clower (1954a) the Marshallian condition.

A full (or stationary, or stock flow) equilibrium sequence in this market is characterized not only by a stationary market price from period to period, but also by zero accumulation of inventory holdings from period to period. The three conditions defining such a static equilibrium sequence (only two of which are independent) are

$$x_t + X_t = 0 \quad X_t = 0 \quad x_t = 0.$$

These conditions define the static equilibrium for both Hicksian and Marshallian theories of market behaviour.

Presuming linear functional forms for the basic trading relations, the alternative theories are illustrated in Figures 1 and 2 and, their common static equilibrium in Figure 3. Note that S_{t+1} in Figures 1 and 2 will be located to the right of S_t by the amount of net new production shown. The market supply schedule $S_{t+1}+s_{t+1}$ correspondingly shifts to the right implying, *ceteris paribus*, a new temporary equilibrium price p_{t+1} and further net new production. In Figures 1 and 2 current period excess flow supply is shown by the amount $q_1 - q_2$. In Figure 3, S^* and p^* denote static equilibrium values of these endogenous variables.

Figure 1

Hicksian temporary equilibrium

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

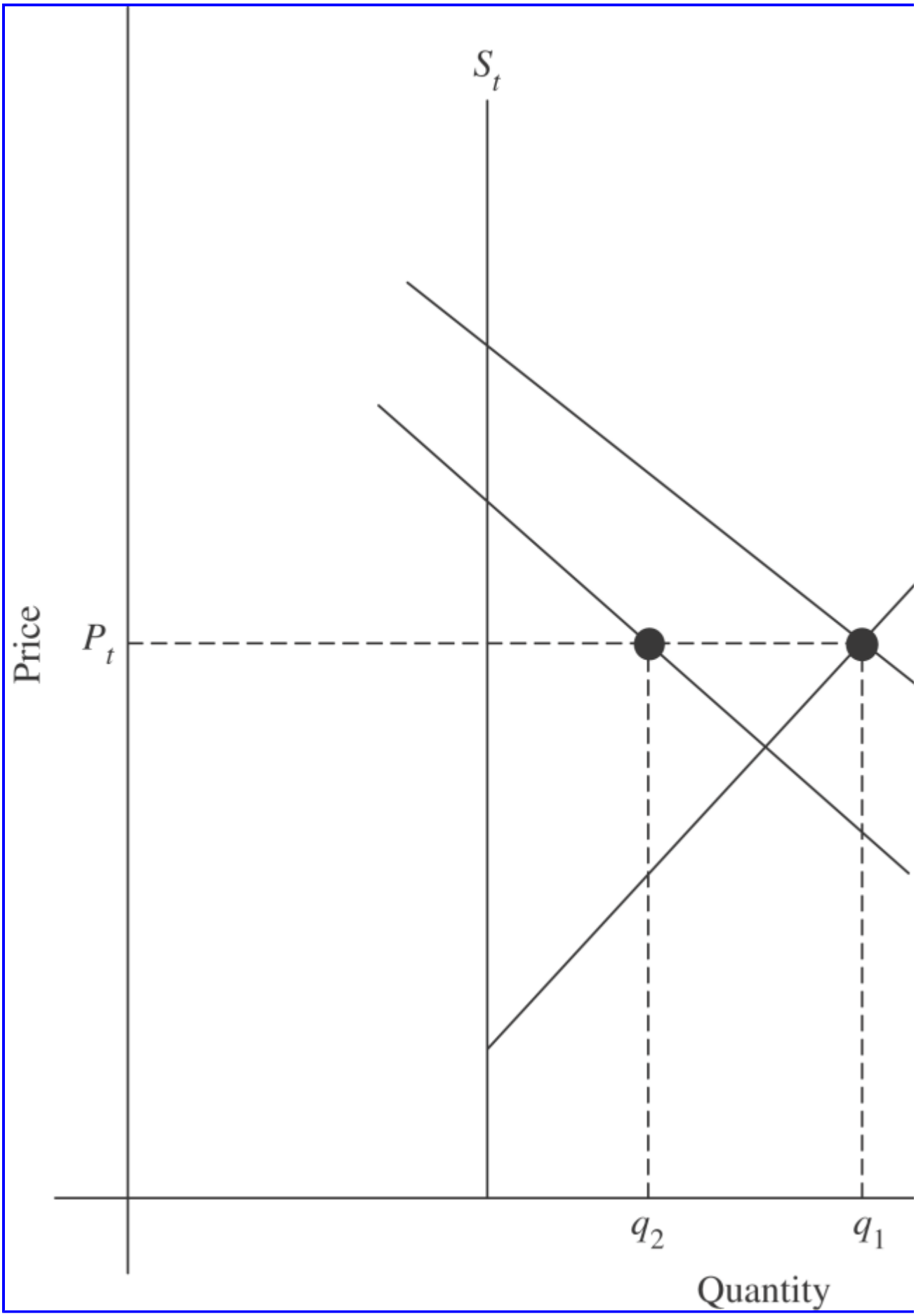

Figure 2
Marshallian temporary equilibrium

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

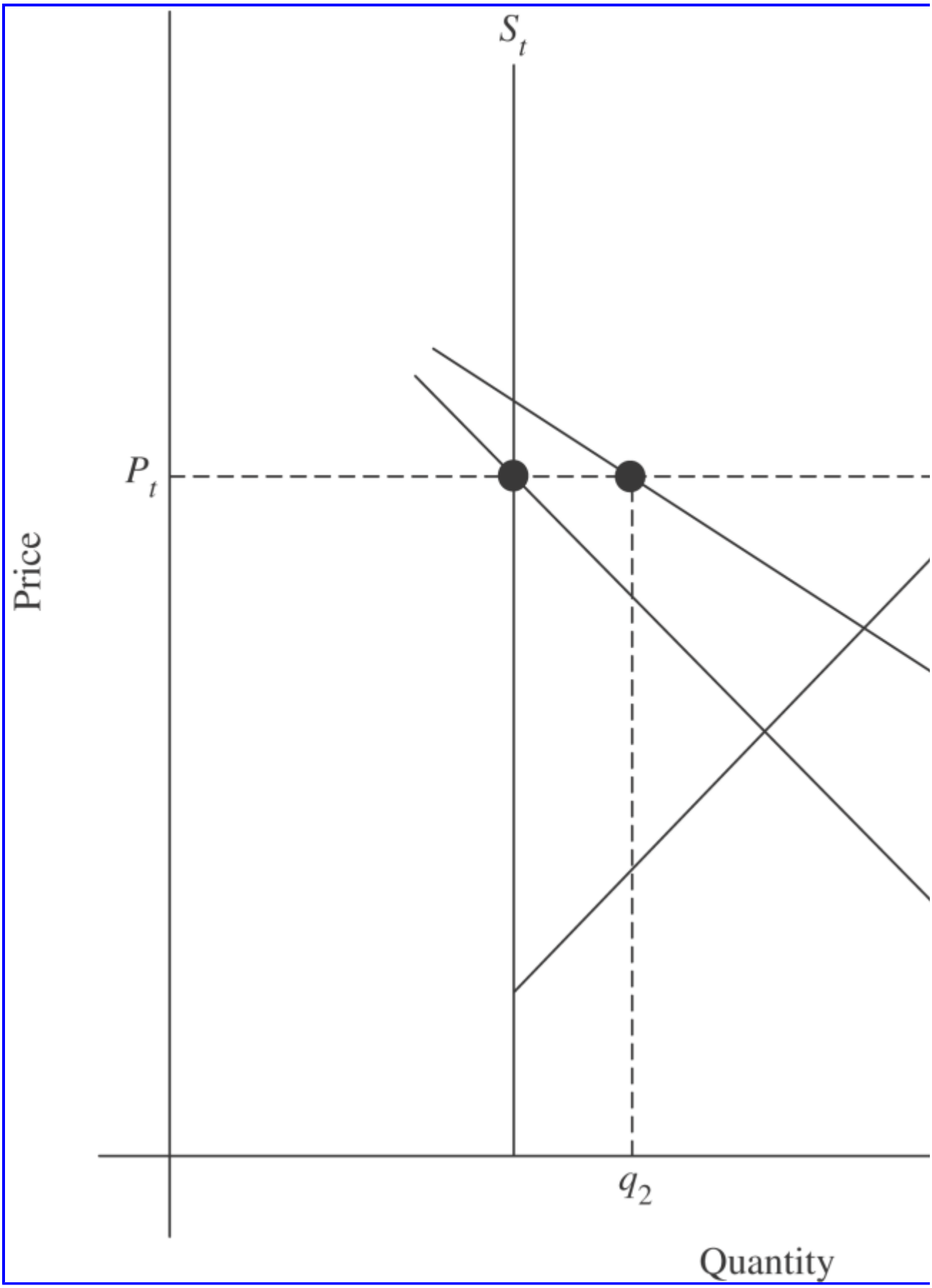

Figure 3
Full equilibrium (Hicksian and Marshallian)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

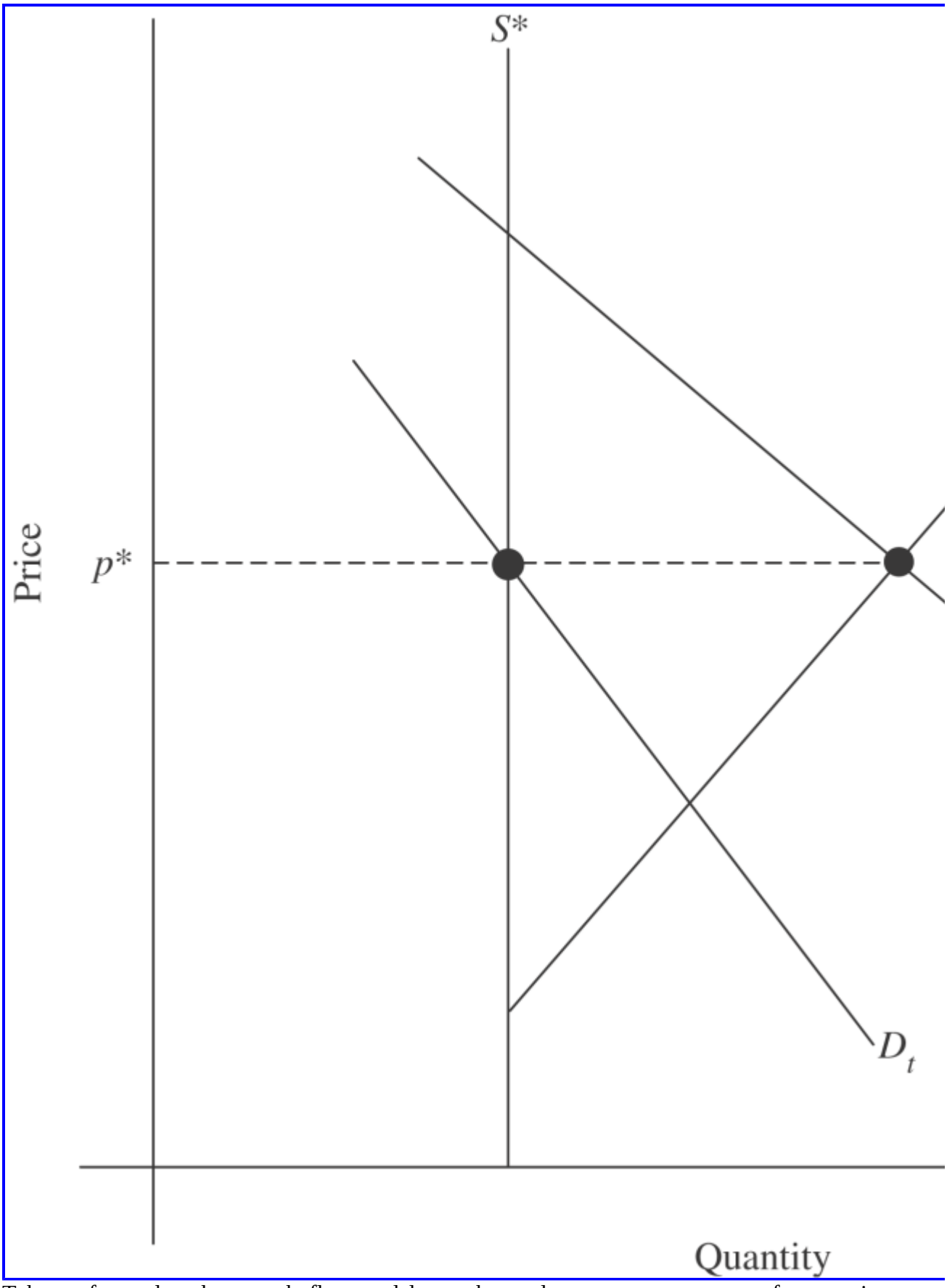

Taken at face value, these stock–flow models may be used to generate sequences of non-stationary temporary equilibria that eventually converge on some full equilibrium (presuming the system to be stable). Indeed Clower (1954a; 1954b) and Bushaw and Clower (1957, pp. 63–75) illustrate these deterministic paths at length. What of the dynamic behaviour of the basic stock–flow model, however, when economic agents have some form of rational expectations (RE) about the future

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

course of the system?

Muth (1961, pp. 321–9) presented a model of an isolated market with certain stock–flow characteristics. If we amend that model to include flow supply as a function of the current period price, it can be shown that Muth's derivation of the RE equilibrium remains essentially valid. Using our notation, the system is:

$$d_t = -\beta p_1 D_t = \alpha[E(p_{t+1}) - p_t] s_t = \gamma p_t + \varepsilon_t D_t + d_t = S_t + s_t,$$

where $E(p_{t+1})$ denotes the price expected in period t to prevail in period t+1, this expectation based on all information about the history of this system up to and including period t. The first three equations refer to the deviations of endogenous variables from their *full equilibrium* values in the absence of any stochastic influences (following Muth). The form of this stock demand equation may be rationalized along several lines (for example, the 'supply of storage' literature uses a similar relation to explain spot/futures differentials); for current purposes we need not pursue these. It can be shown that, if we accept the RE hypothesis that $E(p_{t+1}) = E(p_{t+1}|\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$ and we assume that the stochastic disturbances ε_t are independent, the expected price is given by $E(p_{t+1}) = \lambda p_t$. The parameter λ in this case depends on the coefficients of flow demand, flow supply, and stock demand in a very specific way (see McCallum, 1972). It can also be shown that the use of RE in the simple stock–flow model leads to more robust dynamic market behaviour when compared to alternative expectations schemes.

These results, however, are not intuitively appealing. Stock–flow analysis emphasizes the distinction between temporary and full equilibrium, and the study of the stability (or otherwise) of sequences of temporary equilibria. The notion of rational expectations postulates that agents know the nature of the market they operate in, and pattern their current behaviour on expectations based on that knowledge. Why, then, do agents in Muth (1961, section 4) only consider the current *temporary* equilibrium of their market? It would seem more logical for *speculators* to be concerned with expectations of the *full* equilibrium of the market (note that the first model introduced by Muth, 1961, p. 317, is a pure flow market with a one-period supply lag; in this case it may seem more appropriate to focus on the expected price next period, given that producers are facing a point-input point-output decision problem). Burstein (1982) recognizes this point, arguing correctly that

> under rational expectations, transactors, knowing the structure of the economy and the structure of the sub-economies in which they operate, will optimize relative to 'best estimates' of future data; under characteristic formulations of stock–flow disequilibrium, transactors do not optimize even relative to data-sequences ground out by deterministic models; in the non-stochastic world of standard stock–flow theory, 'rational' transactors would achieve an equilibrium at the onset of the process – one spanning the *phase space* of the system. The upshot would support the basic insights of earlier stock–flow analysts but would require a quite different paradigm of market behavior.

In the amended stock–flow model with RE, a similar derivation with the added *full equilibrium* condition leads to the result $E(p_{t+1})$=0. Recall that we are specifying the basic structural model in deviation form (thus a zero here implies that expected price is equal to the expected full equilibrium price). There are many ways to relax this result formally, as evidenced in recent literature on exchange rate dynamics. One major determinant of dynamics in stock–flow models is therefore the assumed 'term-structure' of expectations (rational or otherwise).

Bibliography

Archibald, G.C. and Lipsey, R.G. 1958. Monetary and value theory: a critique of Patinkin. *Review of Economic Studies* 26, 1–22.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Burstein, M.L. 1982. Stock-flow analysis. In *McGraw-Hill Encyclopedia of Economics*. New York: McGraw-Hill.

Bushaw, D.W. and Clower, R.W. 1957. *Introduction to Mathematical Economics*. Homewood, IL: Irwin.

Clower, R.W. 1954a. An investigation into the dynamics of investment. *American Economic Review* 44, 64–81.

Clower, R.W. 1954b. Productivity, thrift, and the rate of interest. *Economic Journal* 64, 107–15.

Clower, R.W. 1963. Permanent income and transitory balances: Hahn's paradox. *Oxford Economic Papers* 15, 177–90.

Clower, R.W. 1968. Stock–flow analysis. In *International Encyclopedia of the Social Sciences*, vol. 12. New York: Macmillan and Free Press.

Clower, R.W. and Burstein, M.L. 1960. On the invariance of demand for cash and other assets. *Review of Economic Studies* 28, 32–6.

Clower, R.W. and Howitt, P.W. 1978. The transactions theory of the demand for money: a reconsideration. *Journal of Political Economy* 86, 445–66.

Hadar, J. 1965. Comparative statics of stock–flow equilibrium. *Journal of Political Economy* 73, 304–9.

Hadar, J. 1971. *Mathematical Theory of Economic Behavior*. Reading, MA: Addison-Wesley. 1980.

Harrison, G.W. 1980. The stock–flow distinction: a suggested interpretation. *Journal of Macroeconomics* 2, 111–28.

Hicks, J.R. 1946. *Value and Capital*, 2nd edn. Oxford: Clarendon Press.

Hicks, J.R. 1965. *Capital and Growth*. Oxford: Clarendon Press.

Johnson, M.B. 1971. *Household Behaviour: Consumption, Income and Wealth*. Harmondsworth: Penguin.

McCallum, B.T. 1972. Inventory holdings, rational expectations, and the law of supply and demand. *Journal of Political Economy* 80, 386–93.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–35.

Patinkin, D. 1965. *Money, Interest and Prices*, 2nd edn. New York: Harper & Row.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1, 15–29.

How to cite this article

Harrison, Glenn W. "stocks and flows." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000290>
doi:10.1057/9780230226203.1629

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Stone, John Richard Nicholas (1913 – 1991)

Angus Deaton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Richard Stone was a great British empiricist. He and James Meade developed a systematic double-entry approach to national income accounting that later became the basis of the UN's System of National Accounts and for which he received the Nobel Memorial Prize in 1984. His work on the analysis of consumer behaviour is an outstanding example of how to use theory in the service of measurement, and contains perhaps the first case of the econometric estimation of the parameters of preferences. He was founding director of Cambridge's Department of Applied Economics and, through it, the father of a generation of applied econometricians.

Keywords

Afriat, S. N.; Aitchison, J.; Anderson, O. N.; Brown, J. A. C.; Brumberg, R.; Cambridge Growth Model; Clark, C. G.; Cobb–Douglas functions; Cochrane, D.; consumers' expenditure; Cowles Commission; Davenant, C.; demand theory; demographic accounts; double-entry bookkeeping; Duesenberry, J. S.; Durbin, J.; Durbin–Watson test; Econometric Society; econometrics; Engel curve; Farrell, M. J.; Frisch, R. A.; Geary, R. C.; Hoffman, A. J.; household income; Houthakker, H. S.; Keynes, J. M.; King, G.; Klein, L. R.; Koopmans, T. C.; Leontief input–output matrix; Leontief, W.; marginal propensity to consume; Meade, J. E.; national accounting, history of; national income accounting; Orcutt, G. H.; Petty, W.; Pigou, A. C.; policy evaluation; Prais, S. J.; price indexes; principal components analysis; revealed preference; Royal Economic Society; Ruggles, R.; Samuelson, P. A.; savings behaviour; seasonal adjustment; social accounting matrix; stock-adjustment model; Stone index; Stone, J. R. N.; structural estimation; Tinbergen, J.; Tintner, G.; Tobin, J.; Watson, G. S.; Wold, H. O.

Article

Sir Richard Stone, knighted in 1978 and Nobel Laureate in economics in 1984, was the outstanding

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
figure in post-war British applied econometrics. His work in social accounting has had a profound influence on the way that measurement is carried out in economics, and his econometric model building changed the way that economists analyse those measurements. In contrast to many of his British contemporaries, he was a scientist and scholar whose command of methodology and theory was always at the service of the interpretation and measurement of the evidence. He was the inheritor of the British empiricist tradition in economics that saw its first flowering among the ‘political arithmeticians’ of the English Restoration, men such as William Petty, Gregory King and Charles Davenant. To a large extent, he abstained from providing short-term policy advice, preferring to concentrate on the advancement of his science. But his contributions have had an incalculable effect on economic policy and his career provides eloquent testimony to the long-run social value of scientific scholarship in economics and a contrast to the sometimes unenviable record of his contemporaries who involved themselves in the day-to-day conduct of British economic policy.

Richard Stone was born in 1913, attended Westminster School, and set out to follow his father's profession by reading law at Gonville and Caius College, Cambridge. He moved to economics midway through his undergraduate career, and came under the influence of Colin Clark, who was then lecturing in statistics to the economists and who was himself deeply involved in the measurement of national income (see particularly Clark, 1937). Stone's interest in modelling, in measurement and in estimation was immediate. During the summer prior to his graduation from Cambridge, he set out to estimate a two-factor Cobb–Douglas production function, a pioneering effort, the results of which excited little interest or understanding from ‘the Prof’, as Pigou was known, perhaps the first evidence of a Cambridge attitude to econometrics that was later to be reinforced by Maynard Keynes's reactions to Tinbergen's work (Keynes, 1939) and was to be maintained long after similar perceptions had died out elsewhere.

After a brief spell in the City of London, during which he devoted his spare time to producing a monthly bulletin of current economic trends, Stone moved at the outset of the Second World War to Whitehall, where eventually he came to work, with James Meade and initially under his direction, on the construction of wartime national accounts. At Keynes's instigation, their results were published in the 1941 government White Paper, *An Analysis of the Sources of War Finance and an Estimate of the National Income and Expenditure in 1938 and 1940*. In 1945, and again under Keynes's stimulus, the Cambridge Department of Applied Economics was founded and Richard Stone was appointed its first director with an indefinite tenure in the position. Stone brought enormous distinction and worldwide recognition to the department until he was manoeuvred out of the directorship by the Cambridge ‘Keynesians’ in the mid-1950s; he remained in Cambridge as the P.D. Leake Chair of Finance and Accounting until his retirement in 1980. The 1984 Nobel Prize in economics is perhaps the greatest of many professional honors bestowed on Sir Richard. He was a Fellow of King's College, Cambridge from 1945 and of the Econometric Society from 1946. He was president of the Econometric Society in 1955 and President of the Royal Economic Society from 1978 to 1980.

The work for which Stone received the 1984 Nobel Prize in economics was his ‘fundamental contributions to the development of national accounts’ that ‘greatly improved the basis for empirical economic analysis’. The full history of the development of modern national income accounting remains to be written, and any attempt is beyond the scope of an article such as this. It is of course not true that Stone was responsible for the basic concepts of national product, consumption, investment and so on, nor that he provided the first estimates of these magnitudes for the United Kingdom or anywhere else (see, for example, Stone's brief history of the subject in his Nobel Memorial Lecture: Stone, 1984a).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
What Stone (along with Meade, whose original vision Stone developed and made his own) should be credited with is the construction of an interlocking *system* of balanced national accounts, and the implementation of that system on a worldwide basis. Stone's system of national accounts, the SNA, published by the United Nations Statistical Office in 1953 with several subsequent revisions, is not simply a set of tables containing the national income magnitudes, but a set of interlocking accounts in which the principles of double-entry bookkeeping are scrupulously maintained. Each outlay for each agent must be matched somewhere else by an inflow for some other agent, so that each entry in each account must appear somewhere else in some other account. Of course, this is of value only because each account, whether for production, accumulation, consumption or international trade, is independently filled in so that in the end the whole system provides its own complete set of internal consistency checks. Of course, there are always errors and omissions, and some magnitudes cannot be independently measured from both sides of the account, but the credibility and usefulness of each of the numbers hinge on the systematic framework in which they are set. It was Richard Stone, first with James Meade in the Cabinet Office in London, and later on the world stage at the United Nations and the Organization for European Co-operation and Development, who was largely responsible for the way in which national accounts are today collected and presented throughout the world (Stone, 1947; OEEC, 1952).

Stone always favored the presentation of his national accounts in a matrix format, so that each account appears as the row (incomings) and column (outgoings) of a single matrix. In this social accounting matrix (SAM), the standard magnitudes such as national product, consumption or the balance of trade all have their place, but the detailed entries provide a rich picture of the structure and functioning of the economy. For example, the Leontief input–output matrix of inter-industry transactions is the sub-matrix corresponding to the detail of the production accounts. Demand patterns of households appear in the sub-matrix with industries in the rows and households in the columns, while the incomes generated in production flow into households through the value added sub-matrix. Such social accounting matrices can be disaggregated to show any amount of data, and they can be supplemented by balance sheet data (the opening and closing stocks corresponding to the national income flows); and they can be related to socio-demographic variables in a set of demographic accounts. For a typically elegant and lucid account of this with simple examples, see again Stone (1985). One of the most important features of such ‘tableaux économiques’ is that it is almost impossible to look at them for long without being led into attempts to model the behaviour that they reveal. For some cases, the SAM is close to *being* a model; the input–output matrix can be thought of both as a record of transactions, and as a succinct description of the technology of production. Similarly, the links between production, accumulation and consumption lead naturally to models of the allocation of household income between saving and the purchases of goods and services. Together with his first wife, Stone had published one of the very first empirical papers on the marginal propensity to consume (Stone and Stone, 1938), and his work on modelling, particularly of consumer behaviour, continued along with his work on national accounts through the late 1940s and 1950s.

Perhaps Stone's greatest work lies in his empirical analysis of consumer behaviour and the contributions to econometric methodology that came with it. In a series of papers (Stone, 1945; 1948; 1951; Stone and Prais, 1953) that culminated in 1954 in a book, *The Measurement of Consumers’ Expenditure and Behaviour in the United Kingdom, 1920–38*, which to this day remains one of the classics of applied econometrics, Stone presented models that analysed the determination of consumers’ expenditures. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
book contains a dazzling display of all of the elements of the econometrician's art as of the mid-1950s, and there is very much that can be learned from it even today. There is a great deal of very careful and painstaking description of the data, not tucked away where the details cannot be seen, but proudly and prominently displayed for readers to see and quarrel with should they choose. There is a masterly exposition of the theory of demand and of revealed preference, and there is a chapter on econometric methodology that reads like a text until one realizes that this is where the texts originated. The standard matrix algebra formulation of the general linear model $Y = X\beta + U$ appears in its modern form, together with such now standard diagnostics as the Durbin–Watson test, then just invented in the Department of Applied Economics by two young statisticians.

For each of the commodities that he analyses, Stone begins with a loglinear formulation in which the logarithm of the quantity of the good is related to the logarithm of income and the logarithms of other prices, together with a number of other factors that vary from commodity to commodity. For example, the demand for beer is influenced by the average strength of beer as measured by its specific gravity. Stone's major practical problem is lack of degrees of freedom; with only 19 annual observations, disentangling the separate effects of prices, income, and other influences requires generous application of theory and/or of prior information. Stone uses both. In the first place, he uses the Slutsky decomposition to absorb the income effects of prices into the income term through what is now known as a Stone index, thus converting the latter into real rather than money income. Second, he uses zero degree homogeneity to convert prices to relative prices, saving one degree of freedom. Third, he uses elasticities estimated from Engel curve analysis on cross-sectional household budget data to estimate the income elasticities so that, with these imposed, the time-series data are liberated to estimate as many price effects as precisely as possible. Fourth, Stone recognizes the difficulties presented by strong positive autocorrelation in the residuals and to counteract them takes first differences of model and data prior to estimation. Stone's recognition of the non-stationarity of his data, and his first-differencing procedure, though less than perfect, is much superior to and less misleading than the ignoring of the problem that characterized most applied work for the quarter-century after Stone's book. His general procedure set up, Stone then goes on to analyse commodities one by one, reporting results and testing alternative specifications with a care and conviction that has been a model for generations of those of us who have tried to follow him.

The other work of Stone's that is of lasting importance is his paper on the linear expenditure system that appeared in the *Economic Journal* in 1954, the same year that the book appeared. The transition from the models of the book to the model in the paper is in some respects one of the most important transitions in modern applied econometrics, and the methodological issues that are involved are still far from settled. In Stone's book, the influence of the theory of demand is pervasive throughout the discussion of specification and interpretation, but the functional form of the demand equations is essentially ad hoc, the double logarithmic form having been widely adopted because of its convenient parameterizations of the elasticities which are routinely used to describe demand behaviour. The consequences of using such an equation, and of treating demand equations one by one, is that certain aspects of the theory cannot be used nor easily tested. In particular, the symmetry of the compensated substitution effects could not be imposed within the analysis of the book, much as it would have been desirable to do so to gain degrees of freedom and precision of estimation. In the *EJ* paper, Stone comes up with a solution. Starting from a *system* of expenditure equations that are linear in prices and total expenditure, the theoretical requirements of adding-up, homogeneity, and symmetry are imposed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
algebraically to yield a set of estimating equations, the linear expenditure system, that is fully consistent with demand theory. Although the model cannot be estimated by linear methods, Stone invents an iterative Gauss–Seidel procedure that allows him to obtain estimates for a small system using the interwar data. There are many things to admire in this paper, and many things that can be criticized, especially with the benefit of hindsight. The linear expenditure system is a rather primitive model, and Stone's estimation technique was a poor one; similar things could no doubt be said about many great innovations. It is also true that Stone did not solve out for the linear expenditure system utility function, even though the theory of the model had been fully analysed some years before in papers by Klein and Rubin (1947–8), Samuelson (1947–8), and Geary (1950–1). The real originality and importance of the paper lie elsewhere. Nowhere is the previous literature had anyone ever had the extraordinary idea that it might be possible to use economic theory to confront the data so directly; demand equations had been estimated before, but no one had ever attempted to estimate the parameters of a *utility* function. Economic theory might be used as a general guide as to what to look for, but not to yield estimating equations directly. The two main currents in applied econometrics today, structural estimation of ‘deep’ parameters versus more eclectic, atheoretical, or implicitly theoretical estimation, can be seen in Stone's paper and book of 1954. Today, when structural estimation is so familiar, it is easy to forget that ‘taking theory to the data’ is a relatively young methodology. I believe that Stone's linear expenditure system is a major landmark along the route that leads to where we are now.

In an article of this length it is impossible to give any detail on more than a tiny fraction of Stone's contributions to economics, although see my own more detailed (and somewhat more personal) memoir (Deaton, 1993). In addition to his work on the detail of commodity expenditures, there is a set of important papers on savings behaviour (Stone and Rowe, 1962; Stone, 1964a; 1966; 1973) and on the development of the stock-adjustment model for explaining the dynamic demands for durable goods (Stone and Rowe, 1957; 1958; 1960). A fuller appreciation of this work and other papers on demand analysis can be found in Johansen (1985) and in Houthakker (1985). Stone published important work on the theory of price indexes (1956), on seasonal adjustment (1970), and on methods of handling errors of measurement in national accounts (Champernowne, Stone and Meade, 1942; Stone, 1984b). He was one of the first to use principal components analysis as a practical data reduction procedure in economics (Stone, 1947). Over many years, he supervised the construction of the Cambridge Growth Model, in which social accounting matrices and behavioural equations for demand and production were integrated so as to provide a tool for planning and policy evaluation (see in particular Stone and Brown, 1962a); Stone 1964b). He also extended his work on economic accounting to incorporate demographic accounts (Stone, 1971; 1975; Stone and Weale, 1986). At the very end of his life, he acknowledged his debts to his predecessors in a set of quantitative biographies of 12 British empiricists in the social sciences, William Petty, Charles Davenant, Gregory King, William Fleetwood, Arthur Young, Patrick Colquhoun, John Graunt, Edmond Halley, William Farr, Frederick Morton Eden, Florence Nightingale, and Charles Booth (Stone, 1997).

There is another very great contribution that Stone has made to economics and econometrics that is not reflected in his own published work, but in that of those who have been associated with him over the years. Stone was never really a teacher in the conventional way. He was a reluctant lecturer, especially to students, and he participated very little in the routine of Cambridge instruction over more than 30 years of formal attachment to the faculty. However, his personal influence has been extraordinarily strong, partly because of the compelling lucidity of his writings, but also by the example he set to the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
stream of economists and statisticians who spent time in the Department of Applied Economics with him. That stream flowed for many years, but there is no doubt that the best years were at the beginning, in the late 1940s and early 1950s, when Stone himself was working on demand and on the econometric techniques of estimating demands. I have no complete list of those who passed through, but a partial list of those who were there for extended periods includes Brumberg working on life-cycle models, Houthakker working on revealed preference and applied demand analysis, Prais working on family budgets, and Tobin working on demand analysis and on rationing. On the more statistical side, Durbin, Watson, Cochrane, Orcutt and Anderson spent time in the Department working on autocorrelation in economic time series, early visitors included Tintner and Duesenberry, Geary, Klein, Leontief, Samuelson, Koopmans, Wold, Frisch, Ruggles and Hoffman. Farrell began his academic life in Stone's department and did fine empirical work on dynamic demands and on aggregation theory. Prest worked on demand analysis and on time-series problems. Alan Brown worked on Engel curves and wrote a distinguished book with Aitchison on the uses of lognormal distribution (Aitchison and Brown, 1957). Afriat began his work on price indexes in the department. Not only did all of this work owe much to Stone's presence and to the existence of the Department of Applied Economics, but the joint output of all of these people represents an explosion of econometric and economic knowledge that has never been exceeded in the history of the subject and has perhaps been equalled only by the work of the Cowles Commission.

See Also

- consumer expenditure
- demand theory
- economic demography
- Meade, James Edward
- national accounting, history of

Selected works

1938. (With W.M. Stone.) The marginal propensity to consume and the multiplier. *Review of Economic Studies* 6, 1–24.

1942. (With D.G. Champernowne and J.E. Meade.) The precision of national accounts estimates. *Review of Economic Studies* 9, 111–25.

1945. The analysis of market demand. *Journal of the Royal Statistical Society* 108, 1–98.

1947. On the interdependence of blocks of transactions. *Journal of the Royal Statistical Society* 9 (Supplement), 1–45.

1948. The analysis of market demand: an outline of methods and results. *Review of the International*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Statistical Institute 16, 23–35.

1951. The demand for food in the United Kingdom before the war. *Metroeconomica* 3, 8–27.

1953. (With S.J. Prais.) Forecasting from econometric equations: a further note on derationing. *Economic Journal* 62, 565–83.

1954a. Linear expenditure systems and demand analysis: an application to the pattern of British demand. *Economic Journal* 64, 511–27.

1954b. (With D.A. Rowe et al.) *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920–1938*, vol. 1, Cambridge: Cambridge University Press.

1956. *Quantity and Price Indexes in National Accounts*. Paris: OECD.

1957. (With D.A. Rowe.) The market demand for durable goods. *Econometrica* 25, 423–43.

1958. (With D.A. Rowe.) Dynamic demand functions: some econometric results. *Economic Journal* 68, 256–70.

1960. (With D.A. Rowe.) The durability of consumers' durable goods. *Econometrica* 28, 407–16.

1962a. (With J.A.C. Brown.) *A Computable Model of Economic Growth. No. 1 of A Programme for Growth*. London: Chapman and Hall.

1962b. (With D.A. Rowe.) A post-war expenditure function. *Manchester School of Economic and Social Studies* 30, 187–201.

1964a. Private saving in Britain, past, present and future. *Manchester School of Economic and Social Studies* 32, 79–112.

1964b. *The Model in its Environment. No. 5 in A Programme for Growth*. London: Chapman & Hall.

1966. Spending and saving in relation to income and wealth. *L'industria* No. 3, 350–61.

1970. *Mathematical Models of the Economy and Other Essays*. London: Chapman & Hall.

1971. *Demographic Accounting and Model Building*. Paris: OECD.

1973. Personal spending and saving in post-war Britain. In *Economic Structure and Development: Essays in Honour of Jan Tinbergen*, ed. H.C. Bos, H. Linnemann and P. de Wolff. Amsterdam: North-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Holland.

1975. *Towards a System of Social and Demographic Statistics*. Geneva: United Nations.

1984a. *The accounts of society. Nobel Memorial Lecture in Les Prix Nobel 1984*. Stockholm: Almqvist and Wicksell. Reprinted in *Journal of Applied Economics* 1 (1986), 5–28.

1984b. Balancing the national accounts: the adjustment of initial estimates. In *Demand, Equilibrium and Trade*, ed. A. Ingham and A.M. Ulph. London: Macmillan.

1986. (With M.R. Weale.) Two populations and their economies. In *Integrated Analysis of Regional Systems*, ed. P.W.J. Batey and M. Madden. London Papers in Regional Science No. 15. London: Pion.

1998. *Some British Empiricists in the Social Sciences 1650–1900*. Cambridge: Cambridge University Press for the Raffaele Mattioli Foundation.

Bibliography

Aitchison, J. and Brown, J.A.C. 1957. *The Lognormal Distribution, with Special Reference to its Uses in Economics*. Cambridge: Cambridge University Press.

Clark, C. 1937. *National Income and Outlay*. London: Macmillan.

Deaton, A.S. 1993. John Richard Nicholas Stone 1913–1991. *Proceedings of the British Academy* 82, 475–92.

Geary, R.C. 1950–51. A note on ‘A constant utility index of the cost of living’. *Review of Economic Studies* 18, 65–6.

Houthakker, H.S. 1985. *Richard Stone and the analysis of consumer demand*. Cambridge, MA: Harvard Institute of Economic Research, Processed.

Johansen, L. 1985. Richard Stone's contribution to economics. *Scandinavian Journal of Economics* 87, 4–32.

Keynes, J.M. 1939. Professor Tinbergen's method. *Economic Journal* 49, 306–18.

Keynes, J.M. 1940. On a method of statistical business-cycle research: comment. *Economic Journal* 50, 154.

Klein, L.R. and Rubin, H. 1947–8. A constant utility index of the cost of living. *Review of Economic*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Studies 15, 84–7.

OECC (Organization for European Economic Cooperation). 1952. *A Standardised System of National Accounting*. Paris: OECC.

Samuelson, P.A. 1947–8. Some implications of linearity. *Review of Economic Studies* 15, 88–90.

How to cite this article

Deaton, Angus. "Stone, John Richard Nicholas (1913–1991)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000292> doi:10.1057/9780230226203.1630

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

strategic and extensive form games

Martin J. Osborne
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The basic theory of strategic and extensive games is described. Strategic games, Bayesian games, extensive games with perfect information, and extensive games with imperfect information are defined and explained. Among the solution concepts discussed are Nash equilibrium, correlated equilibrium, rationalizability, subgame perfect equilibrium, and weak sequential equilibrium.

Keywords

auctions; bargaining game; Bayesian games; behavioural strategy; Bertrand oligopoly; correlated equilibrium; Cournot oligopoly; Cournot, A.; extensive games with imperfect information; extensive games with perfect information; Hotelling, H.; mixed strategy Nash equilibrium; Nash equilibrium; Prisoner's Dilemma; rationalizability; strategic games; subgame perfect equilibrium

Article

1 Introduction

Game theory is a collection of models designed to understand situations in which decision-makers interact. This article discusses models that focus on the behaviour of individual decision-makers. These models are sometimes called ‘non-cooperative’.

2 Strategic games

2.1 Definition

The basic model of decision-making by a single agent consists of a set of possible actions and a preference relation over this set. The simplest theory of the agent's behaviour is that she chooses a member of the set that is best according to the preference relation.
The model of a *strategic game* extends this model to many agents, who are referred to as *players*. Each player has a set of possible actions and a preference relation over action *profiles* (lists of actions, one for each player).
Definition 1: A ***strategic game with deterministic preferences*** *consists of*

- *a set N* (*the set of* ***players***)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and for each player $i \in N$

- *a set* A_i *(the set of player i's possible* ***actions****)*
- *a preference relation* $\succsim_i$ *over the set* $\times_{j \in N} A_j$ of action profiles.

A strategic game $(N, (A_i), (\succsim_i))$ *is* ***finite*** *if the set N of players and the set* A_i *of actions of each player i are finite.*

The fact that each player's preferences are defined over the set of action profiles allows for the possibility that each player cares not only about her own action but also about the other players' actions, distinguishing the model from a collection of independent single-agent decision problems.

Notice that the model does not have a temporal dimension. An assumption implicit in the solution notions applied to a game is that each player independently commits to an action before knowing the action chosen by any other player. Notice also that no structure is imposed on the players' sets of actions. In the simplest cases, a player's set of actions may consist of two elements; in more complex cases, it may consist, for example, of an interval of real numbers, a set of points in a higher dimensional space, a set of functions from one set to another, or a combination of such sets. In particular, an action may be a contingent plan, specifying a player's behaviour in a variety of possible circumstances, so that the model is not limited to 'static' problems (see Section 3.1.1). Thus, although the model has no temporal dimension, it may be used to study 'dynamic' situations under the assumption that each player chooses her plan of action once and for all.

A few examples give an idea of the range of situations that the model encompasses. The most well-known strategic game is the *Prisoner's Dilemma*. In this game, there are two players ($N = \{1, 2\}$, say), each player has two actions, *Quiet* and *Fink*, and each player's preference relation ranks the action pair in which she chooses *Fink* and the other player chooses *Quiet* highest, then (*Quiet*, *Quiet*), then (*Fink*, *Fink*), and finally the action profile in which she chooses *Quiet* and the other player chooses *Fink*. In this example, as in most examples, working with payoff representations of the players' preference relations is simpler than working with the preference relations themselves. Taking a payoff function for each player that assigns the payoffs 3, 2, 1, and 0 to the four outcomes, we may conveniently represent the game in the table in Figure 1. (Any two-player strategic game in which each player has finitely many actions may be represented in a similar table.)

Figure 1

The Prisoner's Dilemma

		Player 2	
		Quiet	*Fink*
Player 1	*Quiet*	2, 2	0, 3
	Fink	3, 0	1, 1

This game takes its name from the following scenario. The two players are suspected of joint involvement in a major crime. Sufficient evidence exists to convict each one of a minor offence, but conviction of the major crime requires at least one of them to confess, thereby implicating the other (that is, one player 'finks'). Each suspect may stay quiet or may fink. If a single player finks, she is rewarded by being set free, whereas the other player is convicted of the major offence. If both players fink, then each is convicted but serves only a moderate sentence.

The game derives its interest not from this specific interpretation but because the structure of the players' preferences fits many other social and economic situations. The combination of the desirability of the players' coordinating on an outcome and the incentive on the part of each player individually to deviate from this outcome is present in situations as diverse

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

as duopolists setting prices and countries involved in an arms race.

Another example of a strategic game models oligopoly as suggested by Cournot (1838). The players are the n firms, each player's set of actions is the set of possible outputs (the set of non-negative real numbers), and the preference relation of player i is represented by its profit, given by the payoff function u_i defined by

$$u_i(q_1, \ldots, q_n) = q_i P\left(\sum_{j=1}^{n} q_j\right) - C_i(q_i),$$

where q_i is player i's output, C_i is its cost function, and P is the inverse demand function, giving the market price for any total output. Another strategic game that models oligopoly, associated with the name of Bertrand, differs from Cournot's model in taking the set of actions of each player to be the set of possible prices (which requires profit to be defined as a function of prices).

A strategic game that models competition between candidates for political office was suggested by Hotelling (1929). The set of players is a finite set of candidates; each player's set of actions is the same subset X of the line, representing the set of possible policies. Each member of a continuum of citizens (who are not players in the game) has single-peaked preferences over X. Each citizen votes for the candidate whose position is closest to her favourite position. A density function on X represents the distribution of the citizens' favourite policies. The total number of votes obtained by any player is the integral with respect to this density over the subset of X consisting of points closer to the player's action (chosen policy) than to the action of any other player. A player's preferences are represented by the payoff function that assigns 1 to any action profile in which she obtains more votes than every other player, $1/k$ to any action profile in which she obtains at least as many votes as any other player and $k \geq 2$ players tie for the highest number of votes, and 0 to any action profile in which she obtains fewer votes than some other player.

2.2 Nash equilibrium

Which action profile will result when a strategic game is played? Game theory provides two main approaches to answering this qst. One isolates action profiles that correspond to stable 'steady states'. This approach leads to the notion of Nash equilibrium, discussed in this section. The other approach, discussed in Section 2.5, isolates action profiles that are consistent with each player's reasoning regarding the likely actions of the other players, taking into account the other players' reasoning about each other and the player in qst.

Fix an n-player strategic game and suppose that for each player in the game there exists a population of K individuals, where K is large. Imagine that, in each of a long sequence of periods, K sets of n individuals are randomly selected, each set consisting of one individual from each population. In each period, each set of n individuals plays the game, the individual from population i playing the role of player i, for each value of i. The selected sets change from period to period; because K is large, the chance that an individual will play the game with the same opponent twice is low enough not to enter her strategic calculations. If play settles down to a steady state in which each individual in each population i chooses the same action, say a_i^*, whenever she plays the game, what property must the profile a^* satisfy?

In such a (deterministic) steady state, each individual in population i knows from her experience that every individual in every other population j chooses a_j^*. Thus we can think of each such individual as being involved in a single-person decision problem in which the set of actions is A_i and the preferences are induced by player i's preference relation in the game when the action of every other player j is fixed at a_j^*. That is, a_i^* maximizes i's payoff in the game given the actions of all other players. Or, looked at differently, a^* has the property that no player i can increase her payoff by changing her action a_i^* *given the other players' actions*. An action profile with this property is a *Nash equilibrium*. (The notion is due to Nash, 1950; the underlying idea goes back at least to Cournot, 1838.) For any action profile b, denote by (a_i, b_{-i}) the action profile in which player i's action is a_i and the action of every other player j is b_j.

Definition 2: A ***Nash equilibrium*** *of the strategic game* $\langle N, (A_i), (\succsim_i)\rangle$ *is an action profile* a^* *for which*

$$a^* \succsim_i (a_i, a^*_{-i}) \quad \text{for all } a_i \in A_i$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

for every player $i \in N$.

By inspection of the four action pairs in the *Prisoner's Dilemma* (Figure 1) we see that the action pair (*Fink*, *Fink*) is the only Nash equilibrium. For each of the three other action pairs, a player choosing *Quiet* can increase her payoff by switching to *Fink*, given the other player's action.

The games in Figure 2 immediately answer three qsts. Does every strategic game necessarily have a Nash equilibrium? Can a strategic game have more than one Nash equilibrium? Is it possible that every player is better off in one Nash equilibrium than she is in another Nash equilibrium? The left-hand game, which models the game 'Matching pennies', has no Nash equilibrium. The right-hand game has two Nash equilibria, (*B*,*B*) and (*C*, *C*), and both players are better off in (*C*, *C*) than they are in (*B*, *B*).

Figure 2

Two strategic games

	B	*C*
B	1, −1	−1, 1
C	−1, 1	1, −1

	B	*C*
B	1, 1	0, 0
C	0, 0	2, 2

In some games, especially ones in which each player has a continuum of actions, Nash equilibria may most easily be found by first computing each player's best action for every configuration of the other players' actions. For each player *i*, let u_i be a payoff function that represents player *i*'s preferences. Fix a player *i* and define, for each list a_{-i} of the other players' actions, the set of actions that maximize *i*'s payoff:

$$B_i(a_{-i}) = \{a_i \in A_i : a_i \text{ maximizes } u_i(a_i, a_{-i}) \text{ over } a_i \in A_i\}.$$

Each member of $B_i(a_{-i})$ is a *best response* of player *i* to a_{-i}; the function B_i is called player *i*'s *best response function*. (Note that it is set-valued.) An action profile a^* is a Nash equilibrium if and only if

$$a_i^* \in B_i(a_{-i}^*) \text{ for every player } i.$$

In some games, the set $B_i(a_{-i})$ is a singleton for every player *i* and every list a_{-i}. For such a game, denote the single element by $b_i(a_{-i})$. Then the condition for the action profile a^* to be a Nash equilibrium may be written as

$$a_i^* = b_i(a_{-i}^*) \text{ for every player } i,$$

a collection of *n* equations in *n* unknowns.

Consider, for example, a two-player game in which each player's set of actions is the set of non-negative real numbers and the preference relation of each player *i* is represented by the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

payoff function u_i defined by

$$a_i(c + a_j - a_i)$$

where $c > 0$ is a constant. In this game each player *i* has a unique best response to every action a_j of the other player (*j*), given by $b_i(a_j) = \frac{1}{2}(c + a_j)$. The two equations $a_1 = \frac{1}{2}(c + a_2)$ and $a_2 = \frac{1}{2}(c + a_1)$ immediately yield the unique solution $(a_1, a_2) = (c, c)$, which is thus the only Nash equilibrium of the game.

2.3 Mixed strategy Nash equilibrium

In a steady state modelled by the notion of Nash equilibrium, all individuals who play the role of a given player choose the same action whenever they play the game. We may generalize this notion. In a *stochastic steady state*, the rule used to select an action by individuals in the role of a given player is probabilistic rather than deterministic. In a *polymorphic steady state*, each individual chooses the same action whenever she plays the game, but different individuals in the role of a given player choose different deterministic actions.

In both of these generalized steady states an individual faces uncertainty: in a stochastic steady state because the individuals with whom she plays the game choose their actions probabilistically, and in a polymorphic steady state because her potential opponents, who are chosen probabilistically from their respective populations, choose different actions. Thus, to analyse the players' behaviour in such steady states, we need to specify their preferences regarding lotteries over the set of action profiles. The following extension of Definition 1 assumes that these preferences are represented by the expected value of a payoff function. (The term 'vNM preferences' refers to von Neumann and Morgenstern, 1944, pp. 15–31; 1947, pp. 204–221, who give conditions on preferences under which such a representation exists.)

*Definition 3: A **strategic game** (with vNM preferences) consists of*

- *a set N (the set of **players**)*

and for each player $i \in N$

- *a set A_i (the set of player i's possible **actions**)*
- *a function $u_i : \times_{j \in N} A_j \to \mathbb{R}$ (player i's **payoff function**, the expected value of which represents i's preferences over the set of lotteries over action profiles).*

A probability distribution over A_i, the set of actions of player *i*, is called a *mixed strategy* of player *i*. The notion of a *mixed strategy Nash equilibrium* corresponds to a stochastic steady state in which each player chooses her mixed strategy to maximize her expected payoff, given the other players' mixed strategies.

*Definition 4: A **mixed strategy Nash equilibrium** of the strategic game $(N, (A_i), (u_i))$ is a profile α^* in which each component α_i^* is a probability distribution over A_i that satisfies*

$$U_i(\alpha^*) \geq U_i(\alpha_i, \alpha^*_{-i}) \quad \text{for every probability distribution } \alpha_i \text{ on } A_i$$

for every player $i \in N$, where $U_i(\alpha)$ is the expected value of $u_i(a)$ under α.

Suppose that each player's set of actions is finite and fix the mixed strategy of every player $j \neq i$ to be α_j. Then player *i*'s expected payoff when she uses the mixed strategy α_i is a weighted average of her expected payoffs to each of the actions to which α_i assigns positive probability. Thus, if α_i maximizes player *i*'s expected payoff given α_{-i}, then so too do

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

all the actions to which α_i assigns positive probability. This observation has two significant consequences. First, a mixed strategy Nash equilibrium corresponds not only to a stochastic steady state but also to a polymorphic steady state. (The equilibrium probability $\alpha_i^*(a_i)$ is the fraction of individuals in population *i* that choose a_i.) Second, the fact that in a mixed strategy Nash equilibrium each player is indifferent between all the actions to which her mixed strategy assigns positive probability is sometimes useful when computing mixed strategy Nash equilibria.

To illustrate the notion of a mixed strategy Nash equilibrium, consider the games in Figure 2. In the game on the left, a player's expected payoff is the same (equal to 0) for her two actions when the other player chooses each action with probability $\frac{1}{2}$, so that the game has a mixed strategy Nash equilibrium in which each player chooses each action with probability $\frac{1}{2}$. The game has no other mixed strategy Nash equilibrium because each player's best response to any mixed strategy other than the one that assigns probability $\frac{1}{2}$ to each action is either the action *B* or the action *C*, and we know that the game has no equilibrium in which neither player randomizes.

The game on the right of Figure 2 has three mixed strategy Nash equilibria. Two correspond to the Nash equilibria of the game in which randomization is not allowed: each player assigns probability 1 to *B*, and each player assigns probability 1 to *C*. In the third equilibrium, each player assigns probability $\frac{2}{3}$ to *B* and probability $\frac{1}{3}$ to *C*. This strategy pair is a mixed strategy Nash equilibrium because each player's expected payoff to each of her actions is the same (equal to $\frac{2}{3}$ for both players).

The notion of mixed strategy Nash equilibrium generalizes the notion of Nash equilibrium in the following sense.

- If a^* is a Nash equilibrium of the strategic game $(N, (A_i), (\succsim_i))$, then the mixed strategy profile in which each player *i* assigns probability 1 to a_i^* is a mixed strategy Nash equilibrium of any strategic game with vNM preferences $(N, (A_i), (u_i))$ in which, for each player *i*, u_i represents $\succsim_i$.
- If a^* is a mixed strategy Nash equilibrium of the strategic game with vNM preferences $(N, (A_i), (u_i))$ in which for each player *i* there is an action a_i^* such that $\alpha_i^*(a_i^*) = 1$, then a^* is a Nash equilibrium of the strategic game $(N, (A_i), (\succsim_i))$ in which, for each player *i*, $\succsim_i$ is the preference relation represented by u_i.

The following result gives a sufficient condition for a strategic game to have a mixed strategy Nash equilibrium.

Definition 5: A strategic game with vNM preferences $(N, (A_i), (u_i))$ in which the set *N* of players is finite has a mixed strategy Nash equilibrium if either (a) the set A_i of actions of each player *i* is finite or (b) the set A_i of actions of each player *i* is a compact convex subset of a Euclidean space and the payoff function u_i of each player *i* is continuous.

Part (*a*) of this result is due to Nash (1950, 1951) and part (*b*) is due to Glicksberg (1952).

In many games of economic interest the players' payoff functions are not continuous. Several results giving conditions for the existence of a mixed strategy Nash equilibrium in such games are available; see, for example, Section 5 of Reny (1999).

As I have noted, in any mixed strategy Nash equilibrium in which some player chooses an action with positive probability less than 1, that player is indifferent between all the actions to which her strategy assigns positive probability. Thus, she has no positive reason to choose her equilibrium strategy: any other strategy that assigns positive probability to the same actions is equally good. This fact shows that the notion of a mixed strategy equilibrium lacks robustness. A result of Harsanyi (1973) addresses this issue. For any strategic game *G*, Harsanyi considers a game in which the players' payoffs are randomly perturbed by small amounts from their values in *G*. In any play of the perturbed game, each player knows her own payoffs, but not (exactly) those of the other players. (Formally the perturbed game is a Bayesian game, a model described in Section 2.6.) Typically, a player has a unique optimal action in the perturbed game, and this game has an equilibrium in which no player randomizes. (Each player's equilibrium action depends on the value of her own payoffs.) Harsanyi shows that the limit of these equilibria as the perturbations go to zero defines a mixed strategy Nash equilibrium of *G*, and almost any mixed strategy Nash equilibrium of *G* is associated with the limit of such a sequence. Thus we can think of the players' strategies in a mixed strategy Nash equilibrium as approximations to collections of *strictly* optimal actions.

2.4 Correlated equilibrium

One interpretation of a mixed strategy Nash equilibrium is that each player conditions her action on the realization of a random variable, where the random variable observed by each player is independent of the random variable observed by every other player. This interpretation leads naturally to the question of how the theory changes if the players may observe random variables that are not independent.

To take a simple example, consider the game at the right of Figure 2. Suppose that the players observe random variables that are perfectly correlated, each variable taking one value,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

say x, with some probability p, and another value, say y, with probability $1-p$. Consider the strategy that chooses the action B if the realization of the player's random variable is x and the action C if the realization is y. If one player uses this strategy, the other player optimally does so too: if the realization is x, for example, she knows the other player will choose B, so that her best action is B. Thus the strategy pair is an equilibrium.

More generally, the players may observe random variables that are partially correlated. Equilibria in which they do so exist for the game at the right of Figure 2, but the game in Figure 3 is more interesting.

Figure 3

A strategic game

	B	C
B	6, 6	2, 7
C	7, 2	0, 0

Consider the random variable that takes the values x, y and z, each with probability $\frac{1}{3}$. Player 1 observes only whether the realization is in $\{x,y\}$ or is z (but not, in the first case, whether it is x or y), and player 2 observes only whether it is in $\{x,z\}$ or is y. Suppose that player 1 chooses B if she observes $\{x,y\}$ and C if she observes z, and player 2 chooses B if she observes $\{x,z\}$ and C if she observes y. Then neither player has an incentive to change her action, whatever she observes. If, for example, player 1 observes $\{x,y\}$, then she infers that x and y have each occurred with probability $\frac{1}{2}$, so that player 2 will choose each of her actions with probability $\frac{1}{2}$. Thus her expected payoff is 4 if she chooses B and $\frac{7}{2}$ if she chooses C, so that B is optimal. Similarly, if player 1 observes z, she infers that player 2 will choose B, so that C is optimal for her. The outcome is (B, B) with probability $\frac{1}{3}$, (B, C) with probability $\frac{1}{3}$, and (C, B) with probability $\frac{1}{3}$, so that each player's expected payoff is 5.

An interesting feature of this equilibrium is that both players' payoffs exceed their payoffs in the unique mixed strategy Nash equilibrium (in which each player chooses B with probability $\frac{2}{3}$ and obtains the expected payoff $\frac{14}{3}$).

In general, a *correlated equilibrium* of a strategic game with vNM preferences consists of a probability space and, for each player, a partition of the set of states and a function associating an action with each set in the partition (the player's *strategy*) such that, for each player and each set in the player's partition, the action assigned by her strategy to that set maximizes her expected payoff given the probability distribution over the other players' actions implied by her information. (The notion of correlated equilibrium is due to Aumann, 1974.)

The appeal of a correlated equilibrium differs little from the appeal of a mixed strategy equilibrium. In one respect, in fact, most correlated equilibria are more appealing: the action specified by each player's strategy for each member of her partition of the set of states is strictly optimal (she is not indifferent between that action and any others). Nevertheless, the notion of correlated equilibria has found few applications.

2.5 Rationalizability

The outcome (*Fink*, *Fink*) of the *Prisoner's Dilemma* is attractive not only because it is a Nash equilibrium (and hence consistent with a steady state). In addition, for each player, *Fink* is optimal and *Quiet* is suboptimal *regardless* of the other player's action. That is, we may argue solely on the basis of a player's rationality that she will select *Fink*; no reference to her belief about the other player's action is necessary.

We say that the mixed strategy α_i of player i is *rational* if there exists a probability distribution over the other players' actions to which it is a best response. (The probability

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

distribution may entail correlation between the other players' actions; we do not require it to be derived from independent mixed strategies.) With this terminology, the only rational action for each player in the *Prisoner's Dilemma* is *Fink*.

This definition of rationality puts no restriction on the probability distribution over the other players' actions that justifies a player's mixed strategy. In particular, an action is rational even if it is a best response only to a belief that assigns positive probability to the other players' not being rational. For example, in the game on the left of Figure 4, *Q* is rational for player 1, but all the mixed strategies of player 2 to which *Q* is a best response for player 1 assign probability of at least $\frac{1}{2}$ to *Q*, which is not rational for player 2. Such beliefs are ruled out if we assume that each player is not only rational, but also believes that the other players are rational. In the game on the left of Figure 4 this assumption means that player 1's beliefs must assign positive probability only to player 2's action *F*, so that player 1's only optimal action is *F*. That is, in this game the assumptions that each player is rational and that each player believes the other player is rational isolate the action pair (*F*, *F*).

Figure 4

Two variants of the *Prisoner's Dilemma*

	Q	*F*
Q	3, 2	0, 3
F	2, 0	1, 1

	Q	*F*
Q	4, 2	0, 3
X	1, 1	1, 0
F	3, 0	2, 2

We may take this argument further. Consider the game on the right of Figure 4. Player 1's action *Q* is consistent with player 1's rationality and also with a belief that player 2 is rational (because both actions of player 2 are rational). It is not, however, consistent with player 1's believing that player 2 believes that player 1 is rational. If player 2 believes that player is rational, her belief must assign probability 0 to player 1's action *X* (which is not a best response to any strategy of player 2), so that her only optimal action is *F*. But if player 2 assigns positive probability only to *F*, then player 1's action *Q* is not optimal.

In all of these games — the *Prisoner's Dilemma* and the two in Figure 4 — player 1's action *F* survives any number of iterations of the argument: it is consistent with player 1's rationality, player 1's belief that player 2 is rational, player 1's belief that player 2 believes that player 1 is rational, and so on. An action with this property is called *rationalizable*, a notion developed independently by Bernheim (1984) and Pearce (1984). (Both Bernheim and Pearce discuss a slightly different notion, in which players are restricted to beliefs that are derived from *independent* probability distributions over each of the other player's actions. Their notion does not have the same properties as the notion described here.)

The set of action profiles in which every player's action is rationalizable may be given a simple characterization. First define a strictly dominated action.

Definition 6: Player i's action a_i in the strategic game with vNM preferences $\langle N, (A_i), (u_i)\rangle$ is ***strictly dominated*** *if for some mixed strategy α_i of player i we have*

$$U_i(\alpha_i, a_{-i}) > u_i(a_i, a_{-i}) \quad \text{for every } a_{-i} \in \times_{j \in N \setminus \{i\}} A_j,$$

where $U_i(\alpha_i, a_{-i})$ is player i's expected payoff when she uses the mixed strategy α_i and the other players' actions are given by a_{-i}.

Note that the fact that α_i in this definition is a mixed strategy is essential: some strictly dominated actions are not strictly dominated by any action. For example, in the variant of the game at the left of Figure 4 in which player 1 has an additional action, say *Z*, with $u_1(Z, Q) = 0$ and $u_1(Z, F) = 5$, the action *F* is not strictly dominated by any action, but is strictly dominated by the mixed strategy that assigns probability $\frac{3}{4}$ to *Q* and probability $\frac{1}{4}$ to *Z*.

We may show that an action in a finite strategic game is not rational if and only if it is strictly dominated. Given this result, it is not surprising that actions are rationalizable if they

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

survive the iterated elimination of strictly dominated actions, defined precisely as follows.

Definition 7: Let $G = (N, (A_i), (u_i))$ *be a strategic game. For each* $j \in N$, *let* $X_j^1 = A_j$, *and for each* $j \in N$ *and each* $t \geq 1$, *let* X_j^{t+1} *be a subset of* X_j^t *with the property that every member of* $X_j^t \setminus X_j^{t+1}$ *is strictly dominated in the game* $(N, (X_i^t), (u_i^t))$, *where* u_i^t *denotes the restriction of the function* u_i *to* $\times_{j \in N} X_j^t$. *If no member of* X_j^T *for any* $j \in N$ *is strictly dominated, then the set* $\times_{j \in N} X_j^T$ ***survives iterated elimination of strictly dominated actions***.

The procedure specified in this definition does not pin down exactly which actions are eliminated at each step. Only strictly dominated actions are eliminated, but not *all* such actions are necessarily eliminated. Thus the definition leaves open the question of the uniqueness of the set of surviving action profiles. In fact, however, this set is unique; it coincides with the set of profiles of rationalizable actions.

Proposition 8: In a finite strategic game the set of action profiles that survives iterated elimination of strictly dominated actions is unique and is equal to the set of profiles of rationalizable actions.

Every action of any player used with positive probability in a correlated equilibrium is rationalizable. Thus the set of profiles of rationalizable actions is the largest ‘solution’ for a strategic game that we have considered. In many games, in fact, it is very large. (If no player has a strictly dominated action, all actions of every player are rationalizable, for example.) However, in several of the games mentioned in the previous sections, each player has a single rationalizable action, equal to her unique Nash equilibrium action. This property holds, with some additional assumptions, for Cournot's and Bertrand's oligopoly games with two firms and Hotelling's model of electoral competition with two candidates. The fact that in other games the set of rationalizable actions is large has limited applications of the notion, but it remains an important theoretical construct, delineating exactly the conclusion we may reach by assuming that the players take into account each others’ rationality.

2.6 Bayesian games

In the models discussed in the previous sections, every player is fully informed about all the players’ characteristics: their actions, payoffs, and information. In the model of a Bayesian game, players are allowed to be uncertain about these characteristics. We call each configuration of characteristics a *state*. The fact that each player's information about the state may be imperfect is modelled by assuming that each player does not observe the state but rather receives a *signal* that may depend on the state. At one extreme, a player may receive a different signal in every state; such a player has perfect information. At another extreme, a player may receive the same signal in every state; such a player has no information about the state. In between these extremes are situations in which a player is partially informed; she may receive the same signal in states ω_1 and ω_2, for example, and a different signal in state ω_3.

To make a decision, given her information, a player needs to form a belief about the probabilities of the states between which she cannot distinguish. We assume that she starts with a *prior belief* over the set of states, and acts upon the posterior belief derived from this prior, given her signal, using Bayes's Law. If, for example, there are three states, ω_1, ω_2, and ω_3, to which her prior belief assigns probabilities $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{4}$, and she receives the same signal, say *X*, in states ω_1 and ω_2, and a different signal, say *Y*, in state ω_3, then her posterior belief assigns probability $\frac{2}{3}$ to ω_1 and probability $\frac{1}{3}$ to ω_2 when she receives the signal *X* and probability 1 to ω_3 when she receives the signal *Y*.

In summary, a Bayesian game is defined as follows. (The notion is due to Harsanyi, 1967/68.)

Definition 9: A ***Bayesian game*** *consists of*

- *a set N* (*the set of* ***players***)
- *a set* Ω (*the set of* ***states***)

and for each player $i \in N$

- *a set* A_i (*the set of player i's possible* ***actions***)
- *a set* T_i *(the set of* ***signals*** *that player i may receive) and a function* $\tau_i : \Omega \to T_i$, *associating a signal with each state (player i's* ***signal function****)*
- *a probability distribution* p_i *over* Ω (*player i's* ***prior belief***), *with* $p_i(\tau_i^{-1}(t_i)) > 0$ *for all* $t_i \in T_i$
- *a function* $u_i : (\times_{j \in N} A_j) \times \Omega \to \mathbb{R}$ (*player i's* ***payoff function***, *the expected value of which represents i's preferences over the set of lotteries on the set* $(\times_{j \in N} A_j) \times \Omega)$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This definition allows the players to hold different prior beliefs. In many applications every player is assumed to hold the same prior belief.

A widely studied class of Bayesian games models auctions. An example is a single-object auction in which each player knows her own valuation of the object but not that of any other player and believes that every player's valuation is independently drawn from the same distribution. In a Bayesian game that models such a situation, the set of states is the set of profiles of valuations and the signal received by each player depends only on her own valuation, not on the valuation of any other player. Each player holds the same prior belief, which is derived from the assumption that each player's valuation is drawn independently from the same distribution.

The desirability for a player of each of her actions depends in general on the signal she receives. Thus a candidate for an equilibrium in a Bayesian game is a profile of functions, one for each player; the function for player i associates an action (member of A_i) with each signal she may receive (member of T_i). We refer to player i after receiving the signal t_i as *type* t_i of player i. A Nash equilibrium of a Bayesian game embodies the same principle as does a Nash equilibrium of a strategic game: each player's action is optimal given the other players' actions. Thus, in an equilibrium, the action of each type of each player maximizes the payoff of that type given the action of every other type of every other player. That is, a Nash equilibrium of a Bayesian game is a Nash equilibrium of the strategic game in which the set of players is the set of pairs (i,t_i), where i is a player in the Bayesian game and t_i is a signal that she may receive.

Definition 10: A ***Nash equilibrium of a Bayesian game*** $\langle N, \Omega, (a_i), (T_i), (\tau_i), (p_i), (u_i)\rangle$ *is a Nash equilibrium of the following strategic game.*

- *The set of players is the set of all pairs* (i,t_i) *such that* $i \in N$ *and* $t_i \in T_i$.
- *The set of actions of player* (i,t_i) *is* A_i.
- *The payoff of player* (i,t_i) *when each player* (j,t_j) *chooses the action* $a(j,t_j)$ *is*

$$\sum_{\omega \in \Omega} Pr(\omega | t_i)\, u_i((a_i, \hat{a}_{-i}(\omega)), \omega),$$

where $\hat{a}_j(\omega) = a(j, \tau_j(\omega))$ *for each* $j \in N$.

To illustrate this notion, consider the two-player Bayesian game in which there are two states, each player has two actions (B and C), player 1 receives the same signal in both states, player 2 receives a different signal in each state, each player's prior belief assigns probability $\frac{1}{3}$ to state 1 and probability $\frac{2}{3}$ to state 2, and the payoffs are those shown in Figure 5. A Nash equilibrium of this Bayesian game is a Nash equilibrium of the three-player game in which the players are player 1 and the two types of player 2 (one for each state). I claim that the strategy profile in which player 1 chooses B, type 1 of player 2 (that is, player 2 after receiving the signal that the state is 1) chooses C, and type 2 of player 2 chooses B is a Nash equilibrium. The actions of the two types of player 2 are best responses to the action B of player 1. Given these actions, player 1's expected payoff to B is $\frac{2}{3}$ (because with probability $\frac{1}{3}$ the state is 1 and player 2 chooses C and with probability $\frac{2}{3}$ the state is 2 and player 2 chooses B) and her expected payoff to C is $\frac{1}{3}$. Thus player 1's action B is a best response to the actions of the two types of player 2.

Figure 5

A Bayesian game

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

State 1 (probability $\frac{1}{3}$)

	B	C
B	1, 0	0, 1
C	1, 1	1, 0

State 2 (probability $\frac{2}{3}$)

	B	C
B	1, 1	0, 0
C	1, 0	0, 1

3 Extensive games

Although situations in which players choose their actions sequentially may be modelled as strategic games, they are more naturally modelled as extensive games. In Section 3.1 I discuss a model in which each player, when choosing an action, knows the actions taken previously. In Section 3.2 I discuss a more complex model that allows players to be imperfectly informed. (The notion of an extensive game is due to von Neumann and Morgenstern, 1944, and Kuhn, 1950; 1953.) The formulation in terms of histories is due to Ariel Rubinstein.)

3.1 Extensive games with perfect information

An *extensive game with perfect information* describes the sequential structure of the players' actions. It does so by specifying the set of sequences of actions that may occur and the player who chooses an action after each subsequence. A sequence that starts with an action of the player who makes the first move and ends when no move remains is called a *terminal history*.
Definition 11: An ***extensive game with perfect information*** *consists of*

- *a set N (the set of* ***players****)*
- *a set H of sequences (the set of* ***terminal histories****) with the property that no sequence is a proper sub-history of any other sequence*
- *a function P (the* ***player function****) that assigns a player to every proper subsequence of every terminal history*

and for each player $i \in N$

- *a preference relation* $\succsim_i$ *over the set H of terminal histories.*

The restriction on the set H is necessary for its members to be interpreted as terminal histories: if (x,y,z) is a terminal history then (x,y) is not a terminal history, because z may be chosen after (x,y). We refer to subsequences of terminal histories as *histories*.
The sets of actions available to the players when making their moves, while not explicit in the definition, may be deduced from the set of terminal histories. For any history h, the set of actions available to $P(h)$, the player who moves after h, is the set of actions a for which (h,a) is a history. We denote this set $A(h)$.
Two simple examples of extensive games with perfect information are shown in Figure 6. In the game on the left, the set of terminal histories is $\{(X, w), (X, x), (Y, y), (Y, z)\}$ and the player function assigns player 1 to the empty history (a subsequence of every terminal history) and player 2 to the histories X and Y. The game begins with player 1's choosing either X or Y. If she chooses X, then player 2 chooses either w or x; if she chooses Y, then player 2 chooses either y or z. In the game on the right, the set of terminal histories is $\{(W, x, Y), (W, x, Z), (W, y), X\}$ and the player function assigns player 1 to the empty history and the history (W,x), and player 2 to the history W.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Figure 6
Two extensive games with perfect information. *Note*: Player 1's payoff is the first number in each pair.

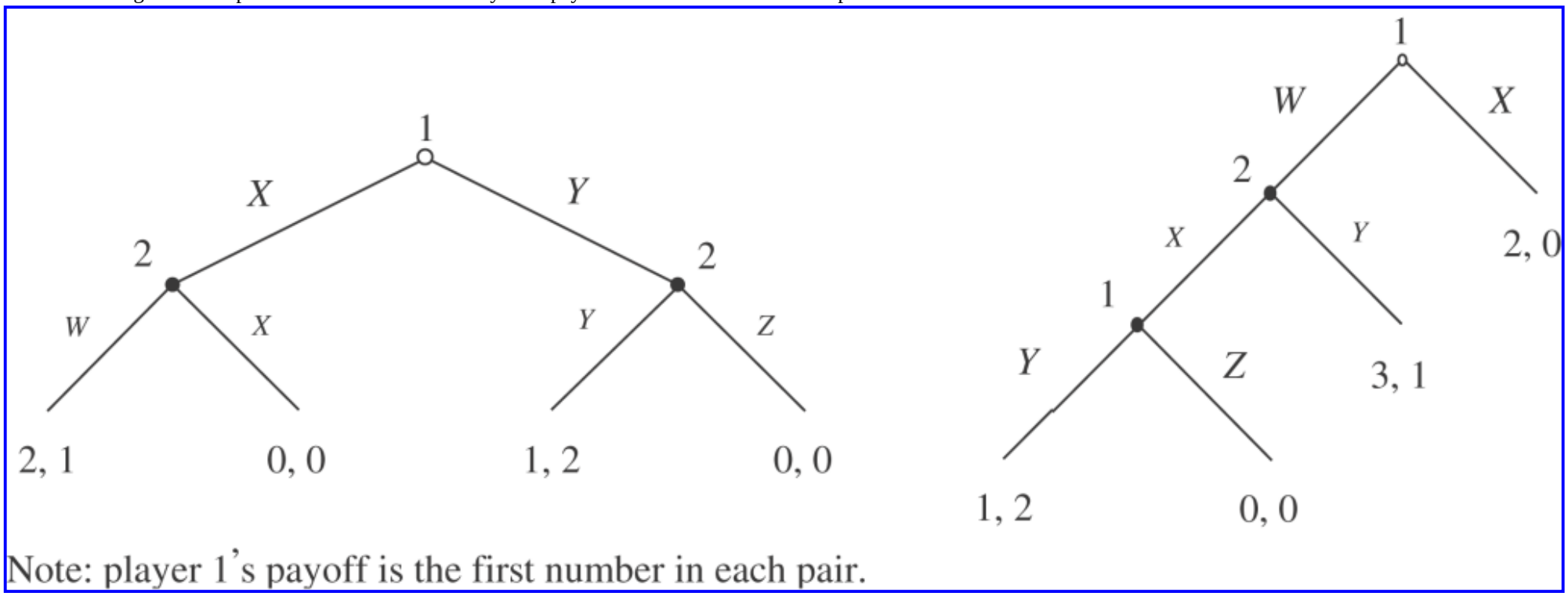

Another example of an extensive game with perfect information is a sequential variant of Cournot's model of oligopoly in which firm 1 chooses an output, then firm 2 chooses an output, and so on. In this game, the set of terminal histories is the set of all sequences $(q_1, \ldots, q_n)$ of outputs for the firms; the player function assigns player 1 to the empty history and, for $k = 1, \ldots, n-1$, player $k+1$ to every sequence $(q_1, \ldots, q_k)$. (Because a continuum of actions is available after each non-terminal history, this game cannot easily be represented by a diagram like those in Figure 6.)

A further example is the bargaining game of alternating offers. This game has terminal histories of infinite length (those in which every offer is rejected).

3.1.1 Strategies

A key concept in the analysis of an extensive game is that of a *strategy*. The definition is very simple: a strategy of any player *j* is a function that associates with *every* history *h* after which player *j* moves a member of *A*(*h*), the set of actions available after *h*.

Definition 12: A ***strategy*** *of player j in an extensive game with perfect information* $(N, H, P, (\succsim_i))$ *is a function that assigns to every history h* (*subsequence of H*) *for which* $P(h) = j$ *an action in A*(*h*).

In the game at the left of Figure 6, player 1 has two strategies, *X* and *Y*. Player 2 has *four* strategies, which we may represent by *wy*, *wz*, *wy*, and *xz*, where the first component in each pair is the action taken after the history *X* and the second component is the action taken after the history *Y*. This example illustrates that a strategy is a complete plan of action, specifying the player's action in every eventuality. Before the game begins, player 2 does not know whether player 1 will choose *X* or *Y*; her strategy prepares her for both eventualities.

The game at the right of Figure 6 illustrates another aspect of the definition. Player 1 in this game has *four* strategies, *WY*, *WZ*, *XY*, and *XZ*. In particular, *XY* and *XZ* are distinct strategies. (Remember that a player's strategy assigns an action to *every* history after which she moves.) I discuss the interpretation of strategies like these in Section 3.1.3.

3.1.2 Nash equilibrium

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

A Nash equilibrium of an extensive game with perfect information is defined in the same way as a Nash equilibrium of a strategic game: it is a strategy profile with the property that no player can increase her payoff by changing her strategy, given the other players' strategies. Precisely, first define the *outcome* $O(S)$ of a strategy profile s to be the terminal history that results when the players use s. (The outcome $O(X,wy)$ of the strategy pair (X,wy) in the game on the left of Figure 6, for example, is the terminal history (X,w).)

Definition 13: A ***Nash equilibrium*** *of the extensive game with perfect information* $\langle N, H, P, (\succsim_i)\rangle$ *is a strategy profile* s^* *for which*

$$O(s^*) \succsim_i O(s_i, s^*_{-i}) \quad \text{for all } s_i \in S_i$$

for every player $i \in N$*, where* S_i *is player i's set of strategies.*

As an example, the game on the left of Figure 6 has three Nash equilibria, (X,wy), (X,wz) and (Y,xy). (One way to find these equilibria is to construct a table like the one in Figure 1 in which each row is a strategy of player 1 and each column is a strategy of player 2.)

For each of the last two equilibria, there exists a history h such that the action specified by player 2's strategy after h is not optimal for her in the rest of the game. For example, in the last equilibrium, player 2's strategy specifies that she will choose x after the history X, whereas only w is optimal for her after this history. Why is such a strategy optimal? Because player 1's strategy calls for her to choose Y, so that the action player 2 plans to take after the history X has no effect on the outcome: the terminal history is (Y,y) regardless of player 2's action after the history X.

I argue that this feature of the strategy pair (Y,xy) detracts from its status as an equilibrium. Its equilibrium status depends on player 1's believing that if she deviates to X then player 2 will choose x. Given that only w is optimal for player 2 after the history X, such a belief seems unreasonable.

Suppose that player 1 forms her belief on the basis of her experience. If she *always* chooses Y, then no amount of experience will enlighten her regarding player 2's choice after the history X. However, in a slightly perturbed steady state in which she very occasionally erroneously chooses X at the start of the game and player 2 chooses her optimal action whenever called upon to move, player 1 knows that player 2 chooses w, not x, after the history X.

If player 1 bases her belief on her reasoning about player 2's rational behaviour (in the spirit of rationalizability), she reaches the same conclusion. (Note, however, that this reasoning process is straightforward in this game only because the game has a finite horizon and one player is indifferent between two terminal histories if and only if the other player is also indifferent.)

In either case, we conclude that player 1 should believe that player 2 will choose w, not x, after the history X. Similarly, the Nash equilibrium (X,wz) entails player 1's unreasonable belief that player 2 will choose z, rather than y, after the history Y. We now extend this idea to all extensive games with perfect information.

3.1.3 Subgame perfect equilibrium

A *subgame perfect equilibrium* is a strategy profile in which each player's strategy is optimal not only at the start of the game, but also after every history. (The notion is due to Selten, 1965.)

Definition 14: A ***subgame perfect equilibrium*** *of the extensive game with perfect information* $\langle N, H, P, (\succsim_i)\rangle$ *is a strategy profile s* for which*

$$O_h(s^*) \succsim_i O_h(s_i, s^*_{-i}) \quad \text{for all } s_i \in S_i$$

for every player $i \in N$ *and every history h after which it is player i's turn to move (that is,* $P(h) = i$*), where* S_i *is player i's set of strategies and* $O_h(s)$ *is the terminal history consisting of h followed by the sequence of actions generated by s after h.*

For any non-terminal history h, define the *subgame* following h to be the part of the game that remains after h has occurred. With this terminology, we have a simple result: a strategy profile is a subgame perfect equilibrium if and only if it induces a Nash equilibrium in every subgame. Note, in particular, that a subgame perfect equilibrium is a Nash equilibrium of the whole game. (The function O in Definition 2 is the same as the function $O_\varnothing$ in Definition 3, where $\varnothing$ denotes the empty history.) The converse is not true, as we have seen: in the game at the left of Figure 6, player 2's only optimal action after the history X is w and her only optimal action after the history Y is y, so that the game has a single subgame perfect

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
equilibrium, (X,wy), whereas it has three Nash equilibria.

Now consider the game at the right of Figure 6. Player 1's only optimal action after the history (W,x) is Y; given that player 1 chooses Y after (W,x), player 2's only optimal action after the history W is x; and given that player 2 chooses x after the history W, player 1's only optimal action at the start of the game is X. Thus the game has a unique subgame perfect equilibrium, (XY,x).

Note, in particular, that player 1's strategy XZ, which generates the same outcome as does her strategy XY regardless of player 2's strategy, is *not* part of a subgame perfect equilibrium. That is, the notion of subgame perfect equilibrium differentiates between these two strategies even though they correspond to the same 'plan of action'. This observation brings us back to a question raised in Section 3.1.1: how should the strategies XZ and XY be interpreted?

If we view a subgame perfect equilibrium as a model of a perturbed steady state in which every player occasionally makes mistakes, the interpretation of player 1's strategy XY is that she chooses X at the start of the game, but, if she erroneously chooses W and player 2 subsequently chooses x, she chooses Y. More generally, a component of a player's strategy that specifies an action after a history h precluded by the other components of the strategy is interpreted to be the action the player takes if, after a series of mistakes, the history h occurs. Note that this interpretation is strained in a game in which some histories occur only after a long series of mistakes, and thus are extremely unlikely.

In some finite horizon games, we may alternatively interpret a subgame perfect equilibrium to be the outcome of the players' calculations about each other's optimal actions. If no player is indifferent between any two terminal histories, then every player can deduce the actions chosen in every subgame of length 1 (at the end of the game); she can use this information to deduce the actions chosen in every subgame of length 2; and she can similarly work back to the start of every subgame at which she has to choose an action. Under this inpt, the component Y of the strategy XY in the game at the right of Figure 6 is player 2's belief about player 1's action after the history (W,x) and also player 1's belief about the action player 2 believes player 1 will choose after the history (W,x). (This interpretation makes sense also under the weaker condition that, whenever one player is indifferent between the outcomes of two actions, every other player is also indifferent – a sufficient condition for each player to be able to deduce her payoff when the other players act optimally, even if she cannot deduce the other players' strategies.)

This inpt, like the previous one, is strained in some games. Consider the game that differs from the one at the right of Figure 6 only in that player 1's payoff of 3 after the history (W,y) is replaced by 1. The unique subgame perfect equilibrium of this game is (XY, x) (as for the original game). The equilibrium entails player 2's belief that player 1 will choose Y if player 2 chooses x after player 1 chooses W. But choosing W is inconsistent with player 1's acting rationally: she guarantees herself a payoff of 2 if she chooses X, but can get at most 1 if she chooses W. Thus it seems that player 2 should either take player 1's action W as an indication that player 1 believes the game to differ from the game that player 2 perceives, or view the action as a mistake. In the first case the way in which player 2 should form a belief about player 1's action after the history (W,x) is unclear. The second case faces difficulties in games with histories that occur only after a long series of mistakes, as for the interpretation of a subgame perfect equilibrium as a perturbed steady state.

The subgame perfect equilibria of the games in Figure 6 may be found by working back from the end of the game, isolating the optimal action after any history given the optimal actions in the following subgame. This procedure, known as *backward induction*, may be used in any finite horizon game in which no player is indifferent between any two terminal histories. A modified version that deals appropriately with indifference may be used in any finite horizon game.

3.2 Extensive games with imperfect information

In an extensive game with perfect information, each player, when taking an action, knows all actions chosen previously. To capture situations in which some or all players are not perfectly informed of past actions we need to extend the model. A general extensive game allows arbitrary gaps in players' knowledge of past actions by specifying, for each player, a partition of the set of histories after which the player moves. The interpretation of this partition is that the player, when choosing an action, knows only the member of the partition in which the history lies, not the history itself. Members of the partition are called *information sets*. When choosing an action, a player has to know the choices available to her; if the choices available after different histories in a given information set were different, the player would know the history that had occurred. Thus for an information partition to be consistent with a player's not knowing which history in a given information set has occurred, for every history h in any given information set, the set $A(h)$ of available actions must be the same. We denote the set of actions available after the information set I_i by $A(I_i)$.

*Definition 15: An **extensive game** consists of*

- *a set N (the set of **players**)*
- *a set H of sequences (the set of **terminal histories**) with the property that no sequence is a proper subhistory of any other sequence*
- *a function P (the **player function**) that assigns a player to every proper subsequence of every terminal history*

and for each player $i \in N$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- *a partition $\mathcal{I}_i$ of the set of histories assigned to i by the player function (player i's **information partition**) such that for every history h in any given member of the partition (**information set**), the set A(h) of actions available is the same*
- *a preference relation $\succsim_i$ over the set H of terminal histories.*

(A further generalization of the notion of an extensive game allows for events to occur randomly during the course of play. This generalization involves no significant conceptual issue, and I do not discuss it.)

An example is shown in Figure 7. The dotted line indicates that the histories *X* and *Y* are in the same information set: player 2, when choosing between *x* and *y*, does not know whether the history is *X* or *Y*. (Formally, player 2's information partition is $\{\{X, Y\}, \{Z\}\}$. Notice that $A(X) = A(Y) (= \{x, y\})$, as required by the definition.)

Figure 7

An extensive game with imperfect information. *Note*: The dotted line indicates that the histories *X* and *Y* are in the same information set.

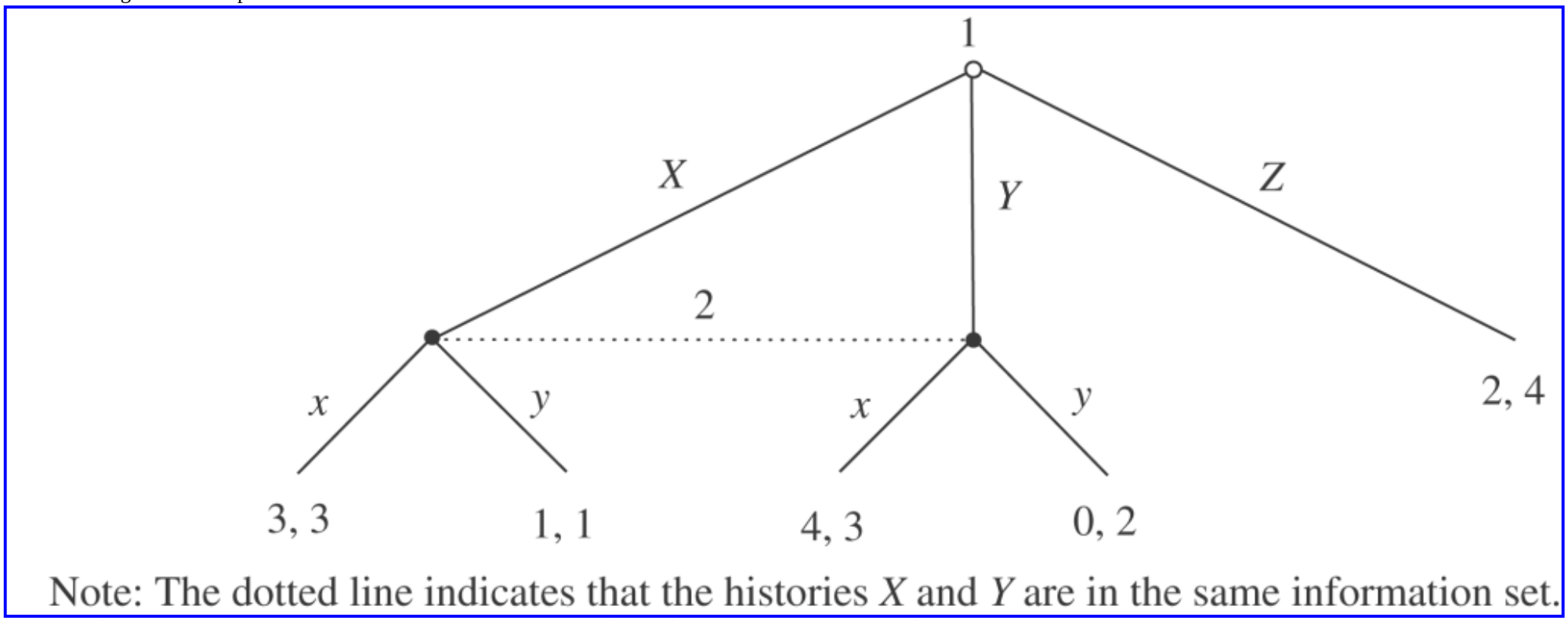

A strategy for any player *j* in an extensive game associates with each of her information sets I_j a member of $A(I_j)$.

*Definition 16: A **strategy** of player j in an extensive game $(N, H, P, (\mathcal{I}_i), (\succsim_i))$ is a function that assigns to every information set $I_j \in \mathcal{I}_j$ of player j an action in $A(I_j)$.*

Given this definition, a Nash equilibrium is defined exactly as for an extensive game with perfect information (Definition 2) — and, as before, is not a satisfactory solution. Before discussing alternatives, we need to consider the possibility of players' randomizing.

In an extensive game with perfect information, allowing players to randomize does not significantly change the set of equilibrium outcomes. In an extensive game with imperfect information, the same is not true. A straightforward way of incorporating the possibility of randomization is to follow the theory of strategic games and allow each player to choose her strategy randomly. That is, we may define a mixed strategy to be a probability distribution over (pure) strategies. An approach more directly suited to the analysis of an extensive game is to allow each player to randomize independently at each information set. This second approach involves the notion of a behavioural strategy, defined as follows.

*Definition 17: A **behavioural strategy** of player j in an extensive game $(N, H, P, (\mathcal{I}_i), (\succsim_i))$ is a function that assigns to each information set $I_j \in \mathcal{I}_j$ a probability distribution over the actions in $A(I_j)$, with the property that each probability distribution is independent of every other distribution.*

For a large class of games, mixed strategies and behavioural strategies are equivalent: for every mixed strategy there exists a behavioural strategy that yields the same outcome

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
regardless of the other players' strategies, and vice versa. (This result is due to Kuhn, 1950; 1953.) In the discussion that follows, I work with behavioural strategies.

The notion of subgame perfect equilibrium for an extensive game with perfect information embodies two conditions: whenever a player takes an action, (*a*) this action is optimal given her belief about the other players' strategies and (*b*) her belief about the other players' strategies is correct. In such a game, each player needs to form a belief only about the other players' future actions. In an extensive game with imperfect information, players need also to form beliefs about the other player's *past* actions. Thus, in order to impose condition *b* on a strategy profile in an extensive game with imperfect information, we need to consider how a player choosing an action at an information set containing more than one history forms a belief about which history has occurred and what it means for such a belief to be correct.

Consider the game in Figure 7. If player 1's strategy is *X* or *Y*, then the requirement that player 2's belief about the history be correct is easy to implement: if player 1's strategy specifies *X* then she believes *X* has occurred, whereas if player 1's strategy specifies *Y* then she believes *Y* has occurred. If player 1's strategy is *Z*, however, this strategy gives player 2 no basis on which to form a belief — we cannot derive from player 1's strategy a belief of player 2 about player 1's action. The main approach to defining equilibrium avoids this difficulty by specifying player 1's belief as a component of an equilibrium. Precisely, we define a belief system and an assessment as follows.

Definition 18: A ***belief system*** *is a function that assigns to every information set a probability distribution over the set of histories in the set. An* ***assessment*** *is a pair consisting of a profile of behavioral strategies and a belief system.*

We may now define an equilibrium to be an assessment satisfying conditions *a* and *b*. To do so, we need to decide exactly how to implement *b*. One option is to require consistency of beliefs with strategies only at information sets reached if the players follow their strategies, and to impose no conditions on beliefs at information sets not reached if the players follow their strategies. The resulting notion of equilibrium is called a weak sequential equilibrium. (The name 'perfect Bayesian equilibrium' is sometimes used, although the notion with this name defined by Fudenberg and Tirole, 1991, covers a smaller class of games and imposes an additional condition on assessments.)

Definition 19: An assessment (β, μ), *where* β *is a behavioural strategy profile and* μ *is a belief system, is a* ***weak sequential equilibrium*** *if it satisfies the following two conditions.*

Sequential rationality.: Each player's strategy is optimal in the part of the game that follows each of her information sets, given the other players' strategies and her belief about the history in the information set that has occurred. Precisely, for each player i and each information set I_i *of player i, player i's expected payoff to the probability distribution over terminal histories generated by her belief* μ_i *at* I_i *and the behaviour prescribed subsequently by the strategy profile* β *is at least as large as her expected payoff to the probability distribution over terminal histories generated by her belief* μ_i *at* I_i *and the behaviour prescribed subsequently by the strategy profile* (γ_i, β_{-i}), *for each of her behavioural strategies* γ_i.

Weak consistency of beliefs with strategies.: For every information set I_i *reached with positive probability given the strategy profile* β, *the probability assigned by the belief system to each history h in* I_i *is the probability of h occurring conditional on* I_i *being reached, as given by Bayes' law.*

Consider the game in Figure 7. Notice that player 2's action *x* yields her a higher payoff than does *y* regardless of her belief. Thus in any weak sequential equilibrium she chooses *x* with probability 1. Given this strategy, player 1's only optimal strategy assigns probability 1 to *Y*. Thus the game has a unique weak sequential equilibrium, in which player 1's strategy is *Y*, player 2's strategy is *x*, and player 2's belief assigns probability 1 to the history *Y*.

Now consider the game in Figure 8. I claim that the assessment in which player 1's strategy is $(\frac{1}{2}, \frac{1}{2}, 0)$, player 2's strategy is $(\frac{1}{2}, \frac{1}{2})$, and player 2's belief assigns probability $\frac{1}{2}$ to *X* and probability $\frac{1}{2}$ to *Y* is a weak sequential equilibrium. Given her beliefs, player 2's expected payoffs to *x* and *y* are both 2, and given player 2's strategy, player 1's expected payoffs to *X* and *Y* are both $\frac{5}{2}$ and her payoff to *Z* is 2. Thus each player's strategy is sequentially rational. Further, player 2's belief is consistent with player 1's strategy. This game has an additional weak sequential equilibrium in which player 1's strategy is *Z*, player 2's strategy is *y*, and player 2's belief assigns probability 1 to the history *Y*. Note that the consistency condition does not restrict player 2's belief in this equilibrium, because player 1 chooses neither *X* nor *Y* with positive probability.

Figure 8

An extensive game with imperfect information

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

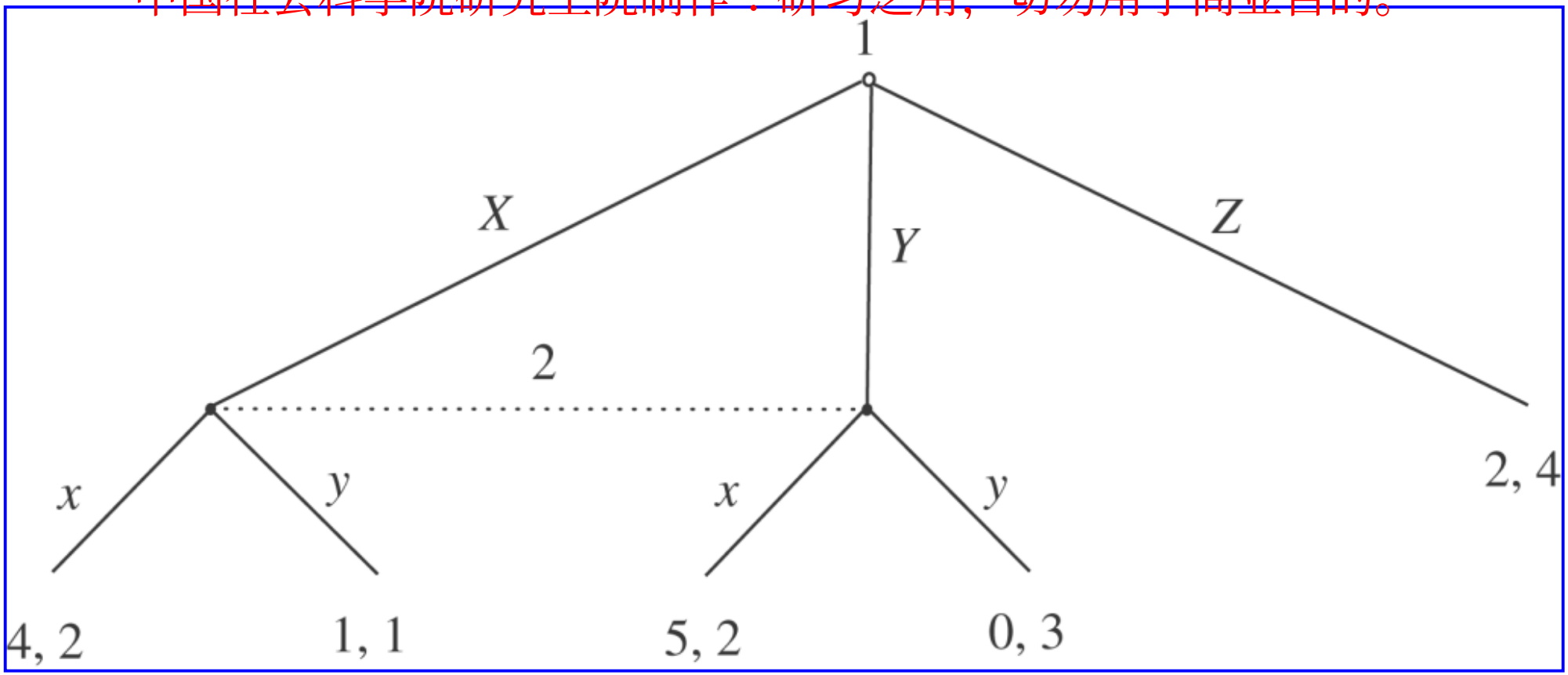

In some games the notion of weak sequential equilibrium yields sharp predictions, but in others it is insufficiently restrictive. Some games, for example, have weak sequential equilibria that do not satisfy a natural generalization of the notion of subgame perfect equilibrium. In response to these problems, several ‘refinements’ of the notion of a weak sequential equilibrium have been studied, including sequential equilibrium (due to Kreps and Wilson, 1982) and perfect Bayesian equilibrium (due to Fudenberg and Tirole, 1991).

See Also

- bargaining
- epistemic game theory: an overview
- epistemic game theory: complete information
- epistemic game theory: incomplete information
- Nash equilibrium, refinements of
- non-cooperative games (equilibrium existence)

I am grateful to Jean Guillaume Forand for comments and to the Social Sciences and Humanities Research Council of Canada for financial support.

Bibliography

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bernheim, B. 1984. Rationalizable strategic behavior. *Econometrica* 52, 1007–28.

Cournot, A. 1838. Recherches sur les principes mathématiques de la théorie des richesses. Paris: Hachette. Trans. N. Bacon as Researches into the Mathematical Principles of the Theory of Wealth. New York: Macmillan, 1897.

Fudenberg, D. and Tirole, J. 1991. Perfect Bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory* 53, 236–60.

Glicksberg, I. 1952. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proceedings of the American Mathematical Society* 3, 170–4.

Harsanyi, J. 1967/68. Games with incomplete information played by 'Bayesian' players, Parts I, II, and III. *Management Science* 14, 159–82, 320–34, and 486–502.

Harsanyi, J. 1973. Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory* 2, 1–23.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39, 41–57.

Kreps, D. and Wilson, R. 1982. Sequential equilibria. *Econometrica* 50, 863–94.

Kuhn, H. 1950. Extensive games. *Proceedings of the National Academy of Sciences of the United States of America* 36, 570–6.

Kuhn, H. 1953. Extensive games and the problem of information. In *Contributions to the Theory of Games, Volume II (Annals of mathematics studies, 28)*, ed. H. Kuhn and A. Tucker. Princeton, NJ: Princeton University Press.

Nash, Jr., J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America* 36, 48–9.

Nash, Jr., J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

Pearce, D. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52, 1029–50.

Reny, P. 1999. On the existence of pure and mixed strategy Nash equilibria in discontinuous games. *Econometrica* 67, 1029–56.

Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetrgheit. *Zeitschrift für die gesamte Staatswissenschaft* 121, 301–24, 667–89.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd edn. Princeton, NJ: Princeton University Press.

How to cite this article

Osborne, Martin J. "strategic and extensive form games." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000523> doi:10.1057/9780230226203.1631

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

strategic trade policy

Barbara J. Spencer and James A. Brander
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Strategic trade policy refers to trade policy that affects the outcome of strategic interactions between firms in an actual or potential international oligopoly. A main idea is that trade policies can raise domestic welfare by shifting profits from foreign to domestic firms. A well-known application is the strategic use of export subsidies, but import tariffs as well as subsidies to R&D or investment for firms facing global competition can also have strategic effects. Since intervention by more than one government can lead to a Prisoner's Dilemma, the theory emphasizes the importance of trade agreements that restrict such interventions.

Keywords

Cournot competition; Cournot oligopoly; Cournot–Nash equilibrium; Cramer's rule; economies of scale; export subsidy; game theory; international oligopoly; international trade (theory); international trade policy; intra-firm trade; intra-industry trade; Krugman, P.; monopolistic competition; multinational corporations; oligopoly theory; optimum tariff; Prisoner's Dilemma; profit shifting; research and development; Ricardo, D.; strategic trade policy; technology transfer

Article

International trade policy is one of the oldest subject areas in economics, having generated serious academic debate at least as far back as the classical period of ancient Greece, well over two thousand years ago. A very informative description of classical Greek and Roman thought on international trade and trade policy is provided by Irwin (1996, ch. 1). Interestingly, for example, both Plato and Aristotle were at best ambivalent about the virtues of open trade. Our modern understanding of international trade policy is based largely on the principle of comparative advantage as developed by David Ricardo (1807) and has been the focus of much political as well as academic debate in the two centuries since Ricardo. Consideration of strategic trade policy is a relatively recent addition to the trade policy debate, having started in the early 1980s. Although definitions of the term differ slightly, we believe the following definition captures the important concepts:

Definition: : *Strategic trade policy refers to trade policy that affects the outcome of strategic interactions between firms in an actual or potential international oligopoly*.

As the definition suggests, the term ‘strategic’ in this context arises from consideration of the strategic interaction between firms. It does not refer to military objectives or the importance of an industry. Strategic interaction requires that firms recognize that their payoffs in terms of profit or other objectives are directly affected by the decisions of rivals or potential rivals. As a result, firms recognize that their own choices concerning such variables as output, price and investment depend on the decisions of other firms. The existence of strategic interaction is the defining characteristic of oligopoly.

The term ‘trade policy’ is interpreted broadly here as any policy directed primarily at the level or pattern of trade. In particular, policies that change the incentives for investment or research and development (R&D) in the context of international oligopoly represent an important application in the literature.

The requirement that the oligopoly be ‘international’ implies that production is actually or potentially carried out in two or more countries. Trade policy instruments set by one country then tend to affect the strategic choices of firms located in that country differently from firms located abroad. Strategic trade policy typically exploits these differential effects so as to achieve a domestic objective at the expense of welfare in other countries. In much of the literature the domestic objective is to maximize aggregate domestic welfare, but there is nothing that rules out political economy objectives, such as the use of trade policy to reward special interest groups that provide large donations to the government.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Most applications of strategic trade policy assume that firms differ by ownership as well as country of location. This assumption focuses attention on the importance of the policy in shifting profits from foreign to domestic firms. Indeed, strategic profit-shifting is often viewed as the hallmark of strategic trade policy. According to our definition, however, strategic trade policy can apply even if all firms in the industry are owned by residents of just one country. For example, a country might be interested in fostering exports by foreign multinationals that compete with firms located abroad. Potential sources of domestic gain would include rents such as above-normal wages, captured by domestic employees of the multinational, and taxes on the multinational's profits.

A brief history of the origins of strategic trade policy

Strategic trade policy was one of the early applications of oligopoly theory in international economics. Formal treatment of oligopoly (and monopolistic competition) in international trade theory did not become well-established until the 1980s. Perhaps the first formal application was by Brander (1981), who explained intra-industry trade in identical commodities. Prior to the 1980s, most trade theory relied on the assumption of perfect competition, although monopoly also received some attention. There was an early ‘distortions literature’ that concerned second-best policies in imperfectly competitive markets, but strategic interaction between firms was not modelled (see, for example, Bhagwati, Ramaswami and Srinivasan, 1969). In the light of the empirical importance of competition between large firms in world markets, the introduction of oligopoly into international trade was a significant step forward in improving the relevance of international trade theory. Oligopoly turned out to be central for understanding and explaining a number of important phenomena that could not be understood in a perfectly competitive framework. In addition to strategic trade policy, these included intra-industry trade, intra-firm trade, multinational corporations, and the role of economies of scale, R&D and technology transfer in international trade.

In applying strategic trade policy, one key difference between oligopoly and other market structures is the existence of profits (or ‘rents’) that can be shifted from one firm to another in a given industry by altering the strategic interactions between firms. Under monopoly, profit-shifting between firms does not arise as there is only one firm. In standard models of perfect competition and monopolistic competition, long-run profits are zero so there are no profits to shift. There are variations of the basic models of perfect competition or monopolistic competition in which firms are heterogeneous, only marginal firms earn zero profits, and infra-marginal firms can earn positive profits. However, such firms are not explicitly engaged in strategic interactions or ‘games’ with one another, so altering the outcome of strategic interactions is not an issue.

As the previous paragraph suggests, application of basic game theory is a feature of strategic trade policy that distinguishes it from much of the previous work in international economics. In addition to considering games between firms, strategic trade policy places particular emphasis on the sequential structure of decision-making, making it one of the first areas of application of game theory where the implications of sequential rationality were clearly understood.

Two papers often cited as pioneering contributions to strategic trade policy are Spencer and Brander (1983) and Brander and Spencer (1985). Both papers assume an international duopoly in which a domestic and a foreign firm compete based on Cournot oligopoly in a third-country market. Spencer and Brander (1983) develops a three-stage game in which a subsidy to R&D (or the combination of an R&D tax and an export subsidy) can increase domestic welfare by shifting profits from the foreign to the domestic firm. The R&D subsidy makes it credible for the domestic firm to commit to a higher level of R&D, causing the foreign firm to reduce its R&D and exports. Brander and Spencer (1985) uses a simpler two-stage game so as to emphasize the profit-shifting role of export subsidies in a more standard international trade setting in which a second good is used to achieve trade balance.

However, an earlier paper, Brander and Spencer (1981), may in fact be the first application of strategic trade policy. In Brander and Spencer (1981) a foreign firm chooses between entry deterrence based on the model of Dixit (1979) and Stackelberg leader–follower competition with a domestic entrant. The paper sets out cost conditions under which the domestic country can gain by increasing its import tariff above the entry-inducing level. The optimum tariff shifts sufficient profits from the foreign to the domestic firm so as to more than offset the loss in consumer surplus and tariff revenue. A drawback of the paper is that the game theory structure of the entry-deterrence model does not satisfy sequential rationality: it is not subgame perfect. The foreign firm prevents entry by setting its exports to the domestic country at a level that reduces domestic profits to zero. This is not subgame perfect since, if the domestic firm were to enter, the foreign firm would maximize profits by reducing its exports so as to accommodate entry. In a sequentially rational structure the domestic firm would be aware of this reaction and entry would not be deterred. In a subsequent paper, Brander and Spencer (1984) examine the strategic use of a tariff to shift profits to a domestic firm in a sequentially rational structure in which a domestic and foreign firm engage in Cournot competition. Other early contributions to strategic trade policy include Krugman (1984), Dixit (1984) and Eaton and Grossman (1986).

Numerical examples

We first illustrate the idea that governments can use trade policy instruments to shift profits from foreign to domestically owned firms, thereby raising national economic welfare at the expense of other countries. The example draws on Brander (1986) and Krugman (1987).

Suppose that only two firms, Boeing, an American firm, and Airbus, a European firm, are capable of producing a certain type of passenger aircraft. To focus on profit-shifting, we abstract from effects on consumer welfare in Europe and America by assuming that the aircraft are all exported to a third country. The profit earned by each country's firm minus the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

cost of any subsidy is then the appropriate measure of each country's national benefit. Consider an initial situation where the third-country market is profitable if there is only one producer, but both firms would make losses if they both enter and must share the market. The European government is considering whether to subsidize the entry of Airbus.

Figure 1 shows the profits of each firm depending on whether or not each firm enters. The game tree on the left illustrates profits if there is no intervention or 'free trade', while the game tree on the right illustrates profits if Europe commits to pay a subsidy of 6 to Airbus in the event that it enters. The box diagrams show Boeing's profit as the first number in each cell, while the second number is the profit of Airbus.

Figure 1

Intervention by Europe

		EUROPE			
		Non-intervention		Entry subsidized by 6	
		Airbus		Airbus	
		Enter	*Not enter*	*Enter*	*Not enter*
Boeing	*Enter*	– 5, – 5	50, 0	– 5, 1	50, 0
	Not enter	0, 50	0, 0	0, 56	0, 0

The outcome of the game in Figure 1 is indeterminate under non-intervention. If either firm enters, the other firm loses from entry. Thus, if Boeing enters while Airbus does not, then Boeing earns 50, but both Boeing and Airbus lose 5 if Airbus enters. By contrast, if Airbus is given a subsidy of 6 when it enters, the outcome is a Nash equilibrium in which Airbus enters and Boeing does not. The subsidy makes entering a dominant strategy for Airbus. If Boeing enters, Airbus earns 1 from entry, which is better than the zero it gets if it does not enter. If Boeing does not enter, then Airbus gains 56 by entering. Given that Airbus enters, Boeing will not enter, for it will lose 5.

To move back one stage to Europe's subsidy decision, it is clear that Europe is made better off by the subsidy. If Boeing would have entered under no intervention, Europe gains 50 by preventing the entry of Boeing: Airbus earns 56, but Europe's payoff is reduced by 6 due to the cost of the subsidy to taxpayers. If there is a 50 per cent chance that Airbus would have captured the market in the absence of intervention, the expected gain to Europe is 25 from intervention. It is notable that a small subsidy can give rise to a large payoff as a result of the effect of the subsidy in changing the outcome of the strategic interaction between firms.

As this example illustrates, strategic trade policy requires that governments have the ability to commit to policy: that is, government policy must be 'credible'. This requirement is captured in the game-theoretic structure by the order in which parties make decisions. Credibility of policy requires that the government move first by committing to its policy, prior

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to the decisions by firms. Commitment means that the government cannot subsequently change its policy. If Boeing did not believe that Europe would follow through on its subsidy, then the subsidy would not have any effect.

Having established the advantage to Europe from a subsidy, the obvious next question is whether the US government would also have an incentive to subsidize the entry of Boeing. In this example, the outcome of the policy game is indeterminate. Both countries lose if both countries subsidize leading to the entry of both firms, but each country has an incentive to subsidize if the other does not. Consequently, to gain a richer insight, we slightly change the example so that each government is considering an export subsidy in a situation where, without intervention, both firms earn profits from exports to the third-country market.

Figure 2 illustrates the payoffs to each country in a symmetric game in which both Airbus and Boeing earn 25 in the absence of intervention (bottom-right cell). If Europe subsidizes exports, say by 6, and the US does not, the profits of Airbus are assumed to increase by 16, so that, net of the subsidy, Europe earns 35 (bottom-left cell). The subsidy makes it credible that Airbus expands its sales at the expense of Boeing, which then earns 5. Due to the overall expansion in sales, the buyers of aircraft enjoy lower prices and the net industry profit (after the European subsidy is subtracted) falls from 50 to 40. Consequently, for the subsidy to benefit Europe, the shift in sales from Boeing to Airbus must be sufficient to offset the fall in price. The same situation applies to the United States if it subsidizes exports but Europe does not. If both countries subsidize exports, the expansion of sales by both firms reduces net industry profit to 20, with each country gaining 10 (top-left cell).

Figure 2

Intervention by both Europe and the United States

		EUROPE	
		Subsidy	No subsidy
United States	Subsidy	10, 10*	35, 5
	No subsidy	5, 35	25, 25

This policy game involves a Prisoner's Dilemma. In a non-cooperative one-shot game in which countries move simultaneously, the dominant strategy is for each country to subsidize its exports. Consequently, at the Nash equilibrium, both countries use strategic trade policy with a payoff of 10 each (upper-left cell). However, both countries would be better off if they could cooperate so as to achieve the higher payoff of 25 (bottom-right cell). As pointed out by Spencer and Brander (1983), one means of cooperation would be to negotiate a trade agreement that binds the countries to free trade. Also, if the game is repeated through time, a government might hope that current cooperation (that is, choosing not to intervene) might induce future cooperation from the other government.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The analysis of strategic trade policy provides a helpful framework for understanding the incentives facing governments in trade policy negotiations. In particular, strategic trade policy provides a rationale for the Prisoner's Dilemma mentality that pervades real-world trade policy negotiations. Under perfect competition, countries would not use export subsidies since they would simply benefit foreign consumers. In a world of strategic trade policy it makes sense that each government might reasonably view reducing or eliminating its own subsidies or tariffs as ‘giving up’ something, but might be willing to do so if other countries do the same. Each country faces a unilateral incentive to use activist trade policy but all can benefit if they can collectively agree to abandon such policies.

A formal model of the role of export subsidies

In the examples given so far, payoffs to firms and countries have been specified as convenient numbers. For a convincing analysis, it is important to model the underlying structure that gives rise to these payoffs. In this section we provide an algebraic demonstration of the argument for the profit-shifting effects of an export subsidy.

As in Brander and Spencer (1985), a domestic and a foreign firm are assumed to act as Cournot competitors in exporting to a third-country market. Entry barriers, such as high fixed costs, prevent entry. Let *x* and *y* represent the exports of the domestic and foreign firm, *c* and *c** their respective marginal costs, and $p = p(x + y)$ the price of the homogeneous product. If an export subsidy, *s*, is applied per unit of domestic exports, the profit functions of the domestic and foreign firm are:

$$\pi(x, y, s) = xp(x + y) - cx + sx \text{ and } \pi^*(x, y, s) = yp(x + y) - c^* y. \quad (1)$$

At the Cournot–Nash equilibrium, each firm sets output to maximize its profit given the output level of its rival. The first order conditions are $\pi_x = 0$ and $\pi^*_y = 0$, where subscripts denote partial derivatives. From total differentiation of the first order conditions with respect to *x*, *y* and *s*, and use of Cramer's rule, it follows that a domestic subsidy always raises domestic exports: that is, $dx / ds = -\pi^*_{yy} / D > 0$ where $\pi^*_{yy} < 0$ and $D / \pi_{xx}\pi^*_{yy} - \pi_{xy}\pi^*_{yx} > 0$ from the second order and stability conditions. To ensure that a domestic export subsidy reduces the output of the foreign firm (that is, $dy / ds = \pi^*_{yx} / D < 0$), an important assumption identified by Brander and Spencer (1985) is

$$\pi^*_{yx} = p' + yp'' < 0 \text{ and } \pi_{xy} = p' + xp'' < 0 \quad (2)$$

where prime and double prime represent first and second derivatives. Condition (2) is now known as the requirement that *x* and *y* be ‘strategic substitutes’: an increase in *x* reduces the rival firm's marginal profit from an increase in *y* and vice versa. Given Cournot competition, this holds for linear demand (since $p'' = 0$) and, more generally, if the inverse demand curve is not too convex.

The effect of the export subsidy is illustrated in Figure 3 showing the best response or reaction functions of each firm. Since domestic marginal cost falls, the subsidy increases domestic exports for any given level of the rival's exports, as shown by the outward shift in the best-response function of the domestic firm. As a result, the subsidy moves the Cournot–Nash equilibrium from point N to point S, reducing the output of the foreign firm.

Figure 3

A domestic export subsidy

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

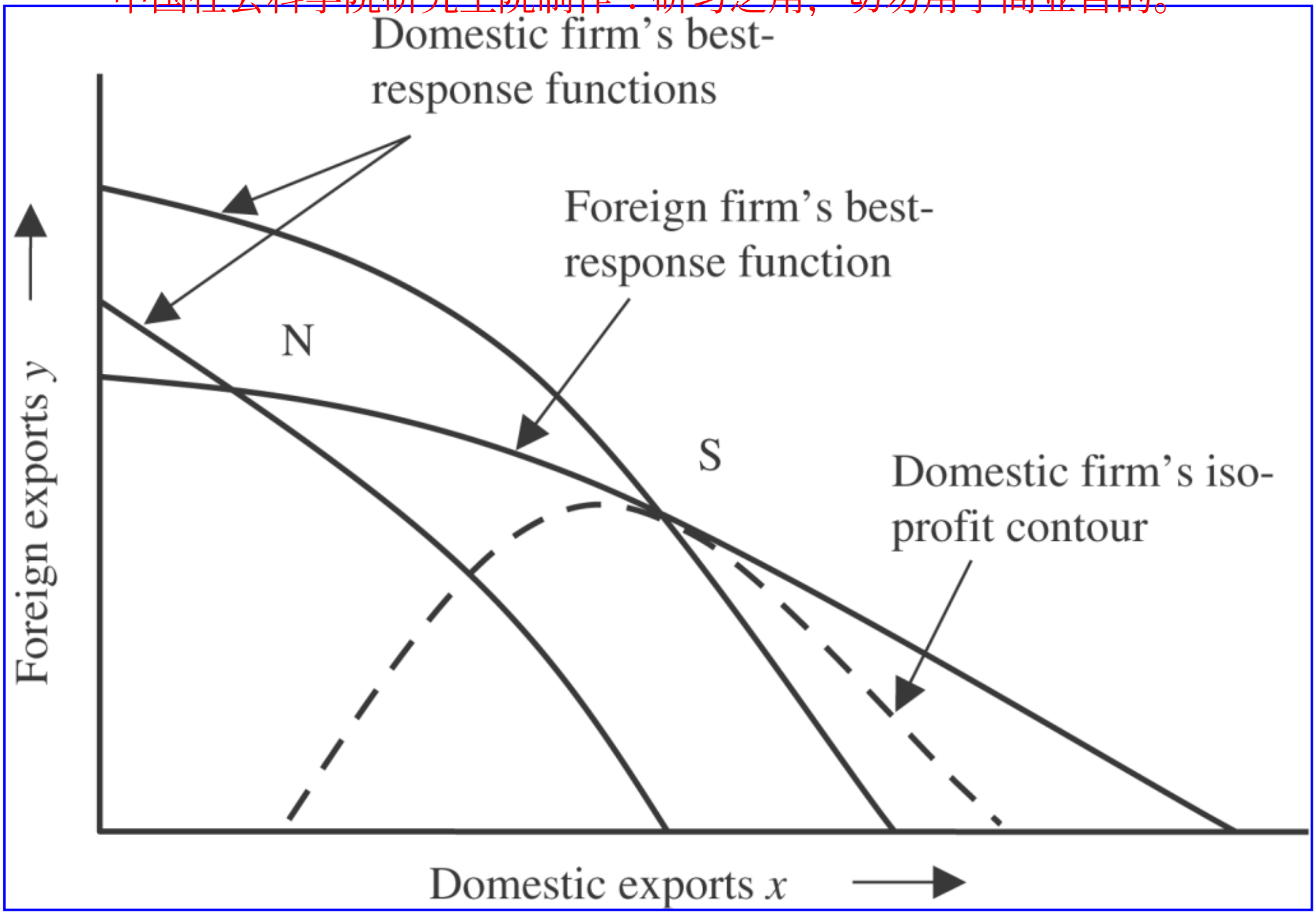

The optimal subsidy is determined by maximizing domestic welfare with respect to *s*. As there is no domestic consumption, welfare, denoted *W*, consists only of the profit of the domestic firm minus the cost of the subsidy:

$$W(s) = \pi(x(s), y(s); s) - sx(s) = xp(x + y) - cx.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(3)

Setting $dW/ds = \pi_y dy/ds - s dx/ds = 0$ yields $s = -\pi_y \pi^*_{yx} / \pi^*_{yy} > 0$. Consequently, an export subsidy increases domestic profits by more than the subsidy payment leading to a rise in domestic welfare. But why is it that we need government intervention to do this?

At the initial Cournot–Nash equilibrium, domestic profits are lower than they would be if the domestic firm were able to somehow act first as a Stackelberg leader so as to take into account the reaction of the foreign firm. Government commitment to an export subsidy makes it credible for the domestic firm to expand within the confines of a Cournot–Nash equilibrium in which no one firm has the ability to act first. Indeed, if only one government intervenes, the optimal export subsidy increases domestic exports to the (unsubsidized) Stackelberg-leader level. As illustrated in Figure 3, the profits of a domestic leader-firm (and domestic welfare) are maximized at S, which is the point of tangency between the domestic firm's iso-profit curve and the foreign firm's reaction function.

Further analysis of this case shows that both governments have a unilateral incentive to subsidize exports. There is a domestic profit-shifting gain even if the foreign government also subsidizes exports. The outcome is a Prisoner's Dilemma in which both are made worse off than if they could agree not to use export subsidies. Thus, as previously discussed, an understanding of strategic trade policy helps us make sense of international trade agreements that disallow export subsidies.

Limitations and extensions

There are significant difficulties in implementing strategic trade policy. A main problem is that strategic trade policy incentives depend very much on the nature of the underlying oligopolistic interaction. In particular, the strategic argument for export subsidies requires that outputs be strategic substitutes, which typically holds for a Cournot duopoly. However, as shown by Eaton and Grossman (1986), outputs are typically strategic complements under Bertrand competition ($\pi^*_{yx} > 0$ and $\pi_{xy} > 0$), giving rise to an incentive to tax rather than to subsidize exports. Other conditions, such as a greater number of domestic relative to foreign firms, can also change optimal policy from a subsidy to a tax.

These findings imply that governments need to know a lot about a particular industry in order to correctly identify whether to target exports with a subsidy or a tax. As a practical matter, there is also the political economy argument that governments might choose to target unprofitable ‘sunset industries’ rather than profitable industries (see Spencer, 1986, for consideration of what should be targeted). In addition, the argument for subsidies is weakened once we recognize that the marginal cost of raising revenue to pay for a subsidy is increased by the distortionary effects of taxation. However, strategic trade policies that use taxes or tariffs are made more attractive by recognizing the full value of government revenue.

Notwithstanding the various limitations of strategic trade policy, the basic insight that governments may have a unilateral incentive to influence the outcome of strategic interactions in international oligopoly remains. Also, strategic trade policy has been analysed in a wide range of contexts and is robust to a range of generalizations. These extensions include consideration of the effects of unionization of the industry, dynamic effects on investment and R&D, vertical integration and trade in intermediate and final goods, and extension to general equilibrium.

See Also

- international trade theory
- strategic and extensive form games

Bibliography

Bhagwati, J., Ramaswami, V. and Srinivasan, T. 1969. Domestic distortions, tariffs and the theory of optimum subsidy: some further results. *Journal of Political Economy* 77, 1005–10.

Brander, J. 1981. Intra-industry trade in identical commodities. *Journal of International Economics* 11, 1–14. Reprinted in *Intra-Industry Trade*, ed. P. Lloyd and H. Grubel. Camberley, UK: Edward Elgar, 2002.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Brander, J. 1986. Rationales for strategic trade and industrial policy. In *Strategic Trade Policy and the New International Economics*, ed. P. Krugman. Cambridge, MA: MIT Press.

Brander, J. and Spencer, B. 1981. Tariffs and the extraction of foreign monopoly rent under potential entry. *Canadian Journal of Economics* 14, 371–89. Reprinted in *Readings in International Trade*, ed. J. Bhagwati. Cambridge, MA: MIT Press, 1987.

Brander, J. and Spencer, B. 1984. Tariff protection and imperfect competition. In *Monopolistic Competition and International Trade*, ed. H. Kierzkowski. Oxford: Oxford University Press. Reprinted in *Imperfect Competition and International Trade*, ed. G. Grossman, Cambridge, MA: MIT Press, 1992.

Brander, J. and Spencer, B. 1985. Export subsidies and international market share rivalry. *Journal of International Economics* 18, 83–100. Reprinted in *International Trade, Volume 1. Welfare and Trade Policy*, ed. P. Neary, Aldershot, UK. and Brookfield, VT: Edward Elgar, 1995.

Dixit, A. 1979. A model of duopoly suggesting a theory of entry-barriers. *Bell Journal of Economics* 10, 20–32.

Dixit, A. 1984. International trade policy for oligopolistic industries. *Economic Journal* 94(supplement), 1–16.

Eaton, J. and Grossman, G. 1986. Optimal trade and industrial policy under oligopoly. *Quarterly Journal of Economics* 101, 383–406. Reprinted in *Readings in International Trade*, ed. J. Bhagwati. Cambridge, MA: MIT Press, 1987.

Irwin, D. 1996. *Against the Tide: An Intellectual History of Free Trade*. Princeton NJ: Princeton University Press.

Krugman, P. 1984. Import protection as export promotion: international competition in the presence of oligopoly and economies of scale. In *Monopolistic Competition and Product Differentiation and International Trade*, ed. H. Kierzkowski. Oxford: Oxford University Press.

Krugman, P. 1987. Is Free Trade Passé? *Journal of Economic Perspectives* 1(2), 131–44.

Ricardo, D. 1817. On the Principles of Political Economy and Taxation. In *The Works and Correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Spencer, B. 1986. What should trade policy target? In *Strategic Trade Policy and the New International Economics*, ed. P. Krugman. Cambridge, MA: MIT Press. Reprinted in *Structural Change, Industrial Location and Competitiveness*, ed. J. Oxley and B. Yeung, Cheltenham, UK and Northampton, MA: Edward Elgar, 1998.

Spencer, B. and Brander, J. 1983. International R&D rivalry and industrial strategy. *Review of Economic Studies* 50, 707–22.

How to cite this article

Spencer, Barbara J. and James A. Brander. "strategic trade policy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000300> doi:10.1057/9780230226203.1632

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

strategic voting

Timothy Feddersen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Strategic voting in elections occurs when a voter submits a ballot in an election with the intention of maximizing the likelihood of a good election outcome given his expectation of how others are voting. Strategic voting is typically contrasted with sincere voting. When election rules permit ballots that amount to a rank ordering of alternatives, a voter is said to vote sincerely if his ballot ranks more preferred alternatives above less preferred ones. There is evidence of strategic voting in real elections, and an extensive theoretical literature demonstrates incentives for strategic voting under almost all election rules.

Keywords

approval voting; Arrow's theorem; binary agenda voting; collective choice; mechanism design; plurality voting; sincere voting; social choice; strategic voting

Article

A voter in an election is said to vote *strategically* (or *tactically*) if he casts a ballot that maximizes his expected payoff from voting in the election. Strategic voting is typically contrasted with *sincere* voting. When election rules permit ballots that amount to a rank ordering of all the alternatives, a voter is said to vote sincerely if his ballot ranks his more preferred alternatives above less preferred ones. Readers interested in an early definition and discussion of strategic voting should see Farqhuarson (1969). The term ‘tactical voting’ is also used to describe voting behaviour in real world mass elections (see Niemi, Whitten and Franklin, 1992.)

For many voting methods sincere voting can be difficult to define. Approval voting (see Weber 1977; 1995, and also Brams and Fishburn, 1978), for example, specifies that voters must cast a ballot in which they may vote for as many candidates or alternatives as they wish. Thus, if three or more candidates were on the ballot a voter might cast one vote for 0, 1, 2 or even all of the candidates. There is no unique

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
map from a voter's strict preferences over the candidates into a ballot that necessarily gives some candidates the same number of votes. Moreover, even when sincere voting is clearly defined strategic and sincere voting models often predict equivalent behaviour.

Interest in strategic voting increased as a consequence of central results in social choice theory. Social choice theory explores the relationship between individual preferences and collective choice rules including common methods of voting. Interested readers may consult Austen-Smith and Banks (1999) for a comprehensive survey of the social choice theory literature. Arrow's (1963) General Possibility Theorem demonstrated that collective choices among three or more alternatives using choice rules that are minimally responsive to citizen's preferences cannot be guaranteed to satisfy minimal rationality conditions unless the method of choice is dictatorial (Arrow, 1963). Gibbard (1973) and Satterthwaite (1975) independently exploit Arrow's theorem to show that all non-dictatorial voting methods are *manipulable*, that is, preference profiles can arise in which, given all other citizens truthfully report their preferences, at least one citizen will have an incentive to misrepresent his preferences.

The Gibbard and Satterthwaite theorems show that almost all voting methods over three or more alternatives create incentives for strategic voting. Riker (1982) illustrates the broad range of voting methods in which three or more citizens may have incentives to vote strategically. In a legislative assembly in which three legislators vote over a set of alternatives $\{x, y, z\}$ using a *binary agenda* method, let us suppose that the legislature first votes between x and y with the alternative receiving the most votes in the first round facing z in a second round of voting. If legislators have cyclic preferences, that is, a majority prefers x to y, y to z and z to x, then at least one legislator will have an incentive to vote strategically between x and y in the first round. The reason is that if x wins in the initial round then in the final round z beats x, while if y wins in the initial round then it also wins in the final round against z. Strategic voters will anticipate the result in the final round and realize that a vote for x in the first round is really a vote for z to be the final outcome. If voters behave strategically a majority will vote for y over x in the first round even though a majority actually prefers x to y. Banks (1985) provides results for a model of strategic voting over binary agendas.

There are also incentives for strategic voting in simultaneous elections with three or more alternatives. Consider an election in which voters must choose between three alternatives (A, B and C) using the plurality rule – the alternative receiving the most votes wins the election. Under plurality rule a voter may cast a vote for at most one of the alternatives. Suppose a voter prefers alternative A to B and B to C. Under the standard interpretation of sincere voting this voter should vote for A. However, if the voter is strategic then he will determine how to vote by comparing the relative likelihood of his vote being *pivotal* and able to change the outcome of the election: from B to A, from C to A; and from C to B (since the voter likes candidate C the least, he or she will never vote for C). Suppose this voter expects that his vote is relatively unlikely to be pivotal between A and either of the other alternatives, compared with the probability his vote may be pivotal between B and C. In that case the voter has an incentive to strategically vote for B. Myerson and Weber (1993) highlight the importance of comparing the relative likelihood of pivot events in determining incentives for strategic voting. Myerson (2002) provides formulae for calculating the relative likelihood of pivot events under various voting rules.

Incentives for strategic voting exist even in two alternative elections when voters have identical preferences and private information (see Austen-Smith and Banks, 1996, and Feddersen and Wolfgang, 1996; 1998, for a discussion of strategic voting incentives in two alternative elections with private

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
information and common values). Consider an election in which a jury of three or more people must decide whether to acquit or convict a defendant. In order to convict the defendant the jury must vote unanimously for conviction, otherwise the defendant is acquitted. Assume that each juror has a private signal known only to himself (guilty or innocent) and that all jurors prefer to convict the defendant if and only if a majority of the private signals are guilty. In this case a sincere voter who observes the guilty signal would vote to convict while one who observes an innocent signal would vote to acquit. A strategic juror who expects that others are voting sincerely properly reasons that the only event in which his vote could be pivotal would be when all the other voters are voting guilty. In this case, he knows that a majority of the private signals are guilty, and therefore he prefers to vote guilty even though he has observed the innocent signal.

The demonstration of the ubiquity of settings in which citizens have incentives to vote strategically has led formal political theorists to increasingly focus on analysing collective choice situations as games and to consider the design of political institutions as a mechanism design problem. Readers may consult Austen-Smith and Banks, 2005, for survey of the literature on strategic models of politics. Chapter 3, in particular, has a careful discussion of the implications of the Gibbard and Satterthwaite theorems for voting models.

Whether or not voters or legislators vote strategically or sincerely in real elections is a difficult question to resolve empirically because preferences are not directly observable. Riker (1982; 1986) provides cases of legislative voting that are consistent with the strategic voting hypothesis. Cox (1997) examines cross-national data on voting patterns as a function of different electoral systems and finds evidence of strategic voting behaviour in mass elections. Degan and Merlo (2006) discuss the empirical difficulties associated with testing the strategic voting hypothesis in mass elections and find evidence that suggests ‘virtually all’ voter behaviour in US elections can be explained by the sincere voting hypothesis. Guarnaschelli, McKelvey and Palfrey (2000) provide evidence from laboratory experiments that is consistent with the predictions of strategic voting models in two-candidate elections with private information.

See Also

- social choice
- voting paradoxes

Bibliography

Arrow, K.J. 1963. *Social Choice and Individual Values*, 2nd edn. New Haven: Yale University Press.

Austen-Smith, D. and Banks, J.S. 1996. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review* 82, 405–22.

Austen-Smith, D. and Banks, J.S. 1999. *Positive Political Theory I: Collective Preference*. Ann Arbor: University of Michigan Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Austen-Smith, D. and Banks, J.S. 2005. *Positive Political Theory II: Strategy and Structure*. Ann Arbor: University of Michigan Press.

Banks, J. 1985. Sophisticated voting outcomes and agenda control. *Social Choice and Welfare* 1, 295–306.

Brams, S.J. and Fishburn, P.C. 1978. Approval voting. *American Political Science Review* 72, 831–47.

Cox, G.W. 1997. *Making Votes Count*. Cambridge: Cambridge University Press.

Degan, A. and Merlo, A. 2006. Do voters vote sincerely? Unpublished manuscript, University of Pennsylvania.

Farqhuarson, R. 1969. *Theory of Voting*. New Haven: Yale University Press.

Feddersen, T.J. and Wolfgang, P. 1996. The swing voter's curse. *American Economic Review* 86, 408–24.

Feddersen, T.J. and Wolfgang, P. 1998. Convicting the innocent: the inferiority of unanimous jury verdicts. *American Political Science Review* 92, 23–35.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Guarnaschelli, S., McKelvey, R.D. and Palfrey, T.R. 2000. An experimental study of jury decision rules. *American Political Science Review* 94, 407–23.

Myerson, R.B. and Weber, R.J. 1993. A theory of voting equilibria. *American Political Science Review* 87, 102–14.

Myerson, R.B. 2002. Comparison of scoring rules in Poisson voting games. *Journal of Economic Theory* 103, 219–51.

Niemi, R.G., Whitten, G. and Franklin, M.N. 1992. Constituency characteristics, individual characteristics and tactical voting in the 1987 British general election. *British Journal of Political Science* 22, 229–40.

Satterthwaite, M.A. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Riker, W.H. 1982. *Liberalism Versus Populism*. Prospect Heights: Waveland Press, Inc.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Riker, W.H. 1986. *The Art of Political Manipulation*. New Haven: Yale University Press.

Weber, R.J. 1977. Comparison of voting systems. Cowles Foundation Discussion Paper No. 498, Yale University.

Weber, R.J. 1995. Approval voting. *Journal of Economic Perspectives* 9(1), 39–49.

How to cite this article

Feddersen, Timothy. "strategic voting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000531>
doi:10.1057/9780230226203.1633

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

strategy-proof allocation mechanisms

Mark A. Satterthwaite
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An allocation mechanism is a function mapping agents' preferences into allocations. Each agent's preferences, however, are private to himself, so in reporting them he can misrepresent them if that achieves a more preferable allocation. A strategy-proof allocation mechanism eliminates incentives to misrepresent: each agent, no matter what his preferences are and no matter what preferences other agents report, maximizes by reporting his preferences truthfully. While the Gibbard–Satterthwaite theorem establishes that no useful strategy-proof allocation mechanisms exist in general, useful ones do exist for a number of specific, economically important classes of allocation problems.

Keywords

allocation mechanisms; Arrow's th; combinatorial auctions; dominant strategy; Gibbard–Satterthwaite th; impossibility theorems; majority rule; median voter rule; misrepresentation of preferences; non-dictatorial allocation mechanisms; Pareto optimal allocation mechanisms; single-peaked preferences; strategy-proof allocation mechanisms; voting by quota; dictatorial allocation mechanisms; Vickrey–Clarke–Groves auction

Article

An allocation mechanism is a function mapping agents' preferences into final allocations. For example, the competitive allocation mechanism calculates market-clearing prices to select a feasible, Pareto optimal, final allocation that varies with agents' preferences. This simple view, however, of the mechanism as a map from preferences to final allocations is inadequate because it ignores agents' propensity to maximize. The problem is that each agent can misrepresent his preferences in reporting them to the mechanism because they are private to him and not verifiable. Therefore, he will tend to misrepresent whenever he realizes that misreporting his true preferences will result in a more preferable allocation than a truthful report. If, as is the case with the competitive mechanism, agents do have

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

incentives to misrepresent their preferences, then the resulting allocation is optimal only with respect to their reported preferences, not to their true preferences.

Strategy-proof allocation mechanisms neutralize the complications that strategic misrepresentation creates. Informally, an allocation mechanism is strategy-proof if each agent's maximizing choice of what preference ordering to report depends only on his own preferences and not on the preferences that other agents report. If the mechanism is strategy-proof, then each agent's incentives are to disregard his expectations concerning other agents' reports and truthfully report his own preference ordering because, no matter what, it secures the allocation that is maximal among those that his choice of reported preferences could secure. That is, truth telling is every agent's dominant strategy if the mechanism is strategy-proof. Strategy-proofness makes understanding the strategic choices of agents trivial: they always play their dominant strategy of reporting their true preferences. Whenever a mechanism fails strategy-proofness, game theoretic methods with all their complications become essential to understand its true optimality properties.

Gibbard – Satterthwaite th

Strategy-proof mechanisms are desirable, but do they exist? That most voting procedures and the competitive mechanism fail strategy-proofness suggests the conjecture that many, if not all, attractive allocation mechanisms also fail strategy-proofness. This conjecture, which Dummett and Farquharson (1961, p. 34) first made, turns out to be true. Gibbard (1973) and Satterthwaite (1975) independently proved that for general environments no strategy-proof allocation mechanisms exist that satisfy minimal requirements for responsiveness to agents' preferences.

Precise statement of this fundamental result requires some notation. Let $I = \{1, 2, \ldots, n\}$, $n \geq 2$, be a fixed set of agents who must select a single alternative from a set $X = \{x, y, z, \ldots\}$ of $|X|$ distinct, final allocations. Each agent $i \in I$ has transitive preferences P_i over the allocations X. Let P_i represent strict preference and I_i represent indifference. Thus for each $x, y \in X$ and each $i \in I$, only three possibilities exist: xP_iy (agent i strictly prefers x over y), yP_ix (agent i strictly prefers y over x), or xI_iy (agent i is indifferent between x and y). Not every transitive ordering X is necessarily admissible as a preference ordering P_i. For example, for a particular $x, y \in X$, xP_iy might be the only admissible ordering because allocation x dominates allocation y in terms of the usual non-satiation axiom of consumer demand theory. Therefore, let Σ represent all possible transitive preference orderings over X and let $\Omega_i \subset \Sigma$ represent the set of all transitive preference orderings over X that are admissible. Thus P_i is an admissible ordering for agent i only if $P_i \in \Omega_i$. Let $\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n$ be the product of all agents' sets of admissible preference orderings. If every transitive ordering is admissible, then preferences are said to be unrestricted and $\Omega = \Sigma^n$. Call the triple (I, X, Ω) the environment.

An n-tuple $P = (P_1, \ldots, P_n) \in \Omega$ is called a preference profile. An allocation mechanism is a function $f : \Omega \to X$ that maps each admissible preference profile into a final allocation. Agent i can manipulate allocation function f at profile $P \in \Omega$ if an admissible ordering $P_i' \in \Omega_i$ exists such that:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$f\left(P_1, \ldots, P_{i-1}, P_i', P_{i+1}, \ldots, P_n\right) P_i f(P). \tag{1}$$

The interpretation of (1) is this. Preference ordering P_i is agent *i*'s true preferences. The other agents report preference orderings $P_{-i} = (P_1, \ldots, P_{i-1}, P_{i+1}, \ldots, P_n)$. If agent *i* reports his preferences truthfully, then the outcome is $f(P_i, P_{-i}) \equiv f(P)$. If he misrepresents his preferences to be P_i', then the outcome is $f(P_i', P_{-i}) \equiv f(P_1, \ldots, P_{i-1}, P_i', P_{i+1}, \ldots, P_n)$. Relation (1) states that agent *i* prefers $f(P_i', P_{-i})$ to $f(P_i, P_{-i})$. Therefore agent *i* has an incentive to manipulate *f* at profile *P* by misrepresenting his preferences to be P_i' rather than P_i.

An allocation mechanism *f* is strategy-proof if no admissible profile $P \in \Omega$ exists at which *f* is manipulable. This means that even if, for instance, agent *i* has perfect foresight about the preferences the other *n*–1 agents will report, agent *i* can never do better than to report his true preferences P_i. Truth is always every agent's dominant strategy. Presumably this is sufficient to induce every agent always to report his preferences truthfully. Gibbard (1973) and Satterthwaite (1975) show that strategy-proof allocation mechanisms generally do not exist.

Theorem: . If admissible preferences are unrestricted $(\Omega = \Sigma^n)$ and at least three possible allocations exist $(|X| \geq 3)$, then no strategy-proof allocation mechanism *f* exists that is non-dictatorial and Pareto optimal.

An allocation mechanism is dictatorial if, for and all profiles $P \in \Omega^n$, an agent *i* exists such that $f(P) \in \max_x P_i$ where

$$\max_x P_i = \{x : x \in X \text{ and, for all } y \in X, \text{ not } y P_i x\}.$$

That is, a dictatorial mechanism always gives the dictator one of the allocations that he most prefers. A non-dictatorial mechanism is a mechanism that is not dictatorial. A mechanism satisfies Pareto optimality if and only if, for any profile $P \in \Omega^n$ and any $x, y \in X$, $x P_i y$ for all $i \in I$ implies $f(P) \neq y$. Pareto optimality implies that, if unanimity exists among the agents that an allocation *x* is the most preferred feasible allocation, then the mechanism picks *x*.

The theorem can be proved either through appeal to Arrow's impossibility theorem (1951) for social-welfare functions or through a direct argument. Reny (2001) constructs parallel direct proofs for both the Arrow and the Gibbard–Satterthwaite theorems that make transparent their identical foundations.

The theorem is an impossibility theorem because mechanisms that violate either non-dictatorship or Pareto optimality are unattractive. If, however, the environment $\langle J, X, \Sigma \rangle$ contains only two allocations $(|X| = 2)$, then impossibility no longer obtains: majority rule is an attractive, non-dictatorial, Pareto

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
optimal, and strategy-proof mechanism. Normally, however, for economic environments $|X| > 2$. Therefore the theorem implies that existence can be obtained (if at all) only for environments $\langle J, X, \Omega \rangle$ where admissible preferences Ω are restricted to a strict subset of Σ^n.

Environments with restricted admissible preferences

In classical economic environments, whether for private goods or for public goods, admissible preferences are naturally constrained so that for all agents admissible preference orderings Ω_i are a strict subset of Σ. Consequently the Gibbard–Satterthwaite theorem does not rule out strategy-proof, non-dictatorial, and Pareto optimal mechanisms. Nevertheless, with one significant exception, classical economic environments do not admit useful strategy-proof mechanisms.

Consider an exchange economy with private goods first. It involves an enormous restriction in Ω relative to Σ^n: each agent cares only about his private allocation from the economy's endowment and, over his possible private allocations, has smooth preferences that generate classical indifference surfaces. Despite this, building on a long series of papers that includes Hurwicz and Walker (1990), Serizawa and Weymark (2003) show that no strategy-proof allocation mechanism exists that always prescribes allocations that are Pareto optimal and, additionally, provide every agent with a consumption bundle that exceeds some positive, minimum level. Thus strategy-proofness and optimality is incompatible with an acceptable distribution of real income.

Turn next to the canonical spatial model of public goods. Let there be $m \geq 2$ public goods so that allocations lie in the non-negative orthant of R^m_+. Each agent's admissible preferences Ω_i includes preference orderings characterized by a most preferred allocation $x \in R^m_+$ surrounded by convex indifference surfaces of less preferred allocations. Zhou (1991) shows that every strategy-proof mechanism in this environment is dictatorial provided that its range within R^m_+ has dimension at least two. This, as Zhou points out, is precisely analogous to the Gibbard–Satterthwaite theorem since Zhou's requirement that the range have dimension two parallels Gibbard and Satterthwaite's twin requirements that *f is* Pareto optimal and *X* includes at least three allocations.

If, however, in this set-up there is only one public good (*m*=1) so that all feasible allocations lie on the non-negative half line R^1_+, then the set of admissible preference profiles reduces to profiles of ‘single-peaked preferences’ and impossibility switches to possibility. A preference ordering $P_i \in \Omega_i$ is single-peaked if it is characterized by a most preferred allocation $x \in R^1_+$ with the desirability of other allocations *y* strictly decreasing as the distance $|x - y|$ increases. Given that for all agents Ω_i contains only single-peaked preferences, strategy-proof and Pareto optimal mechanisms exist. Ching (1997) describes the full set of these mechanisms. The simplest of these rules is the generalization of majority rule that picks the median of the agents’ most preferred allocations. Other, more complicated rules in this set are augmented median voter rules that make use of ‘phantom voters’.

Even though useful strategy-proof mechanisms do not exist in general for classical economic environments, they do exist in some cases for the particular environment of a small-scale allocation problem. Two important examples are briefly discussed here. The first concerns a committee *J* whose *n*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

members are considering a set $K = \{a_1, \ldots, a_m\}$ of $m \geq 2$ proposals and must decide which one of its subsets $B \subset K$ should be approved and implemented. Each member has preferences P_i over the 2^m subsets of K; thus if $A, B \subset K$ and $AP_i \bullet B$, this means that member i prefers that the set A of proposals be approved and implemented rather than the set B of proposals. For each $P_i \in \Omega_i$, let $G(P_i) = \{a \in K | \{a\} P_i \varnothing\}$. Denote member i's 'good' proposals. These are just the proposals a that i would vote to approve if the committee were restricted to choosing between approving the single proposal {a} and no proposals at all. A member i's preferences are separable if, for each ordering $P_i \in \Omega_i$ and each subset $A \subset K$,

$$A \cup \{a\} P_i A \Leftrightarrow a \in G(P_i).$$

In words, adding proposal a to the set A of approved proposals improves the outcome in i's eyes if and only if a is one of the proposals he deems good.

Barbera, Sonnenschein and Zhou (1991) fully characterize all useful strategy-proof voting rules when preferences are restricted to be separable. The simplest member of this class of strategy-proof rules is 'voting by quota'. Under that rule, which is defined by a positive integer Q, each committee member casts a ballot listing the proposals that he judges to be good. Any proposal that is listed on at least Q of the ballots is declared approved. For instance, for election to a club that currently has 100 members, each proposed member might need to be named on the ballots of at least 60 current members in order to be allowed to join.

The second example concerns the Vickrey–Clarke–Groves combinatorial auction in which a seller has a set $K = \{a_1, \ldots, a_m\}$ of $m \geq 2$ of unique objects that he is selling to a set J of $n \geq 2$ buyers. An allocation is a vector $x = \{x_1, x_2, \ldots, x_m\}$ where the value of $x_j \in \{0, 1, 2, \ldots, n\}$ indicates to which buyer i object j is assigned (with the convention that the seller is labelled as buyer 0). Thus if $x_3 = 2$, then the third object is assigned to the second buyer.

Buyer i has quasi-linear utility if his utility (in dollars) for an allocation x and monetary transfer t_i to the seller takes the form $v_i(x) = u_i(x) - t_i$ where t_i is the payment buyer i makes to the seller and $u_i(x)$ is the value that he places on allocation x. Let Ω_i consist of all possible quasi-linear utility functions. The efficient allocation x^* assigns the objects to those buyers who most highly them:

$$x^* \in \arg\max_{x \in X} \sum_{k \in I} u_k(x)$$

where X is the set of all possible allocations. If buyer i declines to participate and only the other the other $n - 1$ buyers bid, then $x^*_{-i} \in \arg\max_{x \in X} \sum_{k \in J \setminus i} u_i(x)$ is the efficient allocation for the remaining buyers

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$k \in I \setminus i$. The value of the *m* objects if they are optimally allocated to all *n* buyers is $V^* = \Sigma_{k \in I} u_k(x^*)$ and their value if they are optimally allocated to the $n - 1$ buyers excluding *i* is $V^*_{-i} = \Sigma_{k \in I \setminus i} u_k(x^*_{-i})$. With this notation in place, the Vickrey–Clarke–Groves auction consists of three steps. First, each buyer $i \in I$ sends the seller a list of bids for each possible allocation: $u_i(x)$ for all $x \in X$. Second, the seller computes and implements the optimal allocation y^*. Third, the seller collects the Vickrey–Clarke–Groves payments from each buyer for that bundle of goods the buyer is allocated under y^*. Buyer *i*'s payment is $V^*_{-i} - (V^* - u_i(x^*))$: the total value the other buyers accrue if *i* does not participate less their total value if *i* does participate. This is the opportunity cost of the bundle agent *i* receives in the optimal allocation. Given that Ω_i contains only quasi-linear utility functions, this auction is strategy-proof: each buyer's dominant strategy is to report truthfully his $u_i(\cdot)$ to the seller. The reason it is strategy-proof is that (*a*) truthful reporting guarantees that he receives a bundle of goods if and only if the value he places on it exceeds its opportunity cost and (*b*) his bundle's opportunity cost is independent of his report. Groves and Loeb (1975) derive the efficiency and incentive properties of this mechanism, Green and Laffont (1979) develop its full theory, and de Vries and Vohra (2003) review its application to combinatorial auctions.

See Also

- Arrow's theorem
- game theory
- social choice

Bibliography

Arrow, K. 1951. *Social Choice and Individual Values*. New York: Wiley.

Barbera, S., Sonnenschein, H. and Zhou, L. 1991. Voting by committees. *Econometrica* 59, 595–609.

Ching, S. 1997. Strategy-proofness and ‘median voters’. *International Journal of Game Theory* 26, 473–90.

de Vries, S. and Vohra, R.V. 2003. Combinatorial auctions: a survey. *INFORMS Journal of Computing* 15, 284–309.

Dummett, M. and Farquharson, R. 1961. Stability in voting. *Econometrica* 29, 33–44.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–602.

Green, J. and Laffont, J.-J. 1979. *Incentives in Public Decision-making*. Studies in Public Economics,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

vol. 1. Amsterdam: North-Holland.

Groves, T. and Loeb, M. 1975. Incentives and public inputs. *Journal of Public Economics* 4, 211–26.

Hurwicz, L. and Walker, M. 1990. On the generic nonoptimality of dominant-strategy allocation mechanisms: a general theorem that includes pure exchange economies. *Econometrica* 58, 683–704.

Reny, P. 2001. Arrow's theorem and the Gibbard–Satterthwaite theorem: a unified approach. *Economics Letters* 70, 99–105.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Serizawa, S. and Weymark, J. 2003. Efficient strategy-proof exchange and minimum consumption guarantees. *Journal of Economic Theory* 109, 246–63.

Zhou, L. 1991. Impossibility of strategy-proof mechanisms in economies with pure public goods. *Review of Economic Studies* 58, 107–19.

How to cite this article

Satterthwaite, Mark A. "strategy-proof allocation mechanisms." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000301> doi:10.1057/9780230226203.1634

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stratification

Stephen L. Morgan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Stratification’ refers to a structure of inequality where individuals occupy differentiated positions that are ranked hierarchically according to broadly recognized standards. Prominent in 20th-century sociology, the term was used by Parsons and his students to explain why individuals in the most functionally important positions in society receive the greatest rewards for their services. In sociology, the most important legacy of stratification research is the cross-national study of intergenerational mobility between occupational categories. Recently, economists have joined sociologists in studying the relationship between increasing inequalities within the labour markets of industrialized countries and rates of intergenerational mobility.

Keywords

class; inequality; intergenerational income mobility; Parsons, T.; mobility; social status; stratification

Article

‘Stratification’ is a term used to characterize a structure of inequality where (*a*) individuals occupy differentiated structural positions and (*b*) the positions are situated in layers (or strata) that are ranked hierarchically according to broadly recognized standards. The implied reference to sedimentary layers from geology reflects the relative permanence of the posited structure and the long history that is assumed to have generated it. Stratification researchers focus primarily on the empirical study of (*a*) the sources of the rankings that generate the hierarchy of strata, (*b*) the mobility of individuals between strata, and (*c*) the mechanisms of integration that allow societies to cope with the existence of persistent inequalities between strata.

The structural orientation of stratification scholarship can be contrasted with distributional approaches to the study of inequality that have dominated economics. Modelling the distribution of valued resources across individuals makes possible explanations of change in response to short-run interventions and shocks from unforeseen exogenous events. For stratification researchers, short-run variation in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
inequality is considered to be noise that dissipates as social inequalities are reproduced.

The development of the concept of stratification

Although the sociologist Pitirim Sorokin is often credited with first developing and then using the concept of stratification in empirical work, the clearest lineage emerges in the work of Talcott Parsons and that of his students. In his essay ‘An analytical approach to the theory of social stratification’, Parsons (1940, p. 841) wrote: ‘Social stratification is regarded here as the differential ranking of the human individuals who compose a given social system and their treatment as superior and inferior relative to one another in certain socially important respects.’ Parsons (1940, p. 849) then wrote that the ‘status of any given individual in the system of stratification in a society may be regarded as a resultant of the common valuations underlying the attribution of status to him’ in dimensions such as achievements, possessions, authority, and power.

In 1945 Parsons's students, Kingsley Davis and Wilbert Moore, wrote ‘Some principles of stratification’ in which they specified a clear (but ultimately controversial) conception of the sources and inevitably of stratification. Adopting the functionalist framework championed by Parsons, Davis and Moore maintained that society is a functioning social system, directly analogous to a living organism, which survives because it determines necessary social positions, recruits appropriate individuals to fill each position, and induces individuals to perform their assigned duties. To foster efficiency, the social system attaches differential rewards to alternative positions, where the sizes of the rewards are based on (*a*) the functional importance of the position to the society as a whole and (*b*) the counterfactual scarcity of individuals willing to take the position in the absence of appropriate rewards. Davis and Moore (1945, p. 243) claimed that ‘Social inequality is thus an unconsciously evolved device by which societies insure that the most important positions are conscientiously filled by the most qualified persons.’ (As discussed later, this view of the sources, functional necessity, and inevitably of stratification was challenged almost immediately by scholars wishing to focus on the mutability of the structure of inequality as well as the power dynamics that pervade it.)

Thirteen years after his initial essay on the topic, and eights years after the seminal Davis and Moore piece, Parsons (1953, p. 92) began ‘A revised analytical approach to the theory of social stratification’ with the bold assertion: ‘It has come to be rather widely recognized in the sociological field that social stratification is a generalized aspect of the structure of all social systems, and that the system of stratification is intimately linked to the level and type of integration of the system as a system.’ Parsons then discussed how societies cope with the functional necessity of stratification by developing norms and value standards that, by and large, attribute differences in attainment to differences in achievement. Parsons and his colleagues assumed that moderately high levels of intergenerational mobility are essential for the efficiency and integration of society. Functionally important positions must be staffed by the most qualified individuals and hence based on past achievements rather than social origins. And, to ensure integration and social order, reward for achievement rather than reward for social origins must be reasonably expected and then observed. In later work, Parsons (1959) specified the social processes that develop and then transmit these norms of achievement in his essay ‘The school class as a social system’. He argued that schools serve two primary functions in society – socialization and allocation – which they fulfil in a simultaneous four-part process: (*a*) emancipation of children from exclusive

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
attachment to their parents, (*b*) inculcation of values and norms that cannot be taught by parents, (*c*) differentiation of the school class on actual achievement and on differential valuation of achievement, and finally (*d*) an allocation of individuals to positions in the adult role system. Parsons (1959, p. 145) wrote: ‘Differentiation of the class along the achievement axis is inevitably a source of strain, because it confers higher rewards and privileges on one contingent than on another within the same system. … [The] common valuation [of achievement] helps make possible the acceptance of the crucial differentiation, especially by the loser in the competition.’

As a result of this scholarship, the term ‘stratification’ gained popularity in sociology, becoming the name for the entire sub-field of inquiry concerned with the causes and consequences of inequality. Thereafter, the term diffused throughout the social sciences and was drawn upon by historians and anthropologists to frame comparative studies of inequality. Within economics, the term has been used less frequently and with no consistent definition. For early studies of racial–ethnic inequality, one can find the term used in the dissimilar work of Closson (1896) and Myrdal (1944). More recently, the term has been used to refer to the determinants of labour market earnings that arise from family background rather than one's own skills (see Heckman and Hotz, 1986). For Durlauf (1994) and Benabou (1996), the term is used to refer to persistent neighbourhood differences in average levels of family income and well-being.

Complexity and contention

The foregoing presentation of the Parsonian perspective on stratification is sanitized in two important respects. First, it ignores alternative perspectives which rejected the Parsonian vision at the time it was being proposed and which later grew into the neo-Marxist scholarship of the 1960s onward. This scholarship led Pierre Bourdieu (1984, p. 245) to look back at the stratification literature of the 1950s and 1960s and declare that ‘the opposition between theories which describe the social world in the language of stratification and those which speak the language of class struggle corresponds to two ways of seeing the social world.’

Second, the presentation overly formalizes what for some was an informal term, often used as shorthand for the simple notion of a systematic pattern of inequality. In fact, many of the scholars who worked within the stratification tradition used a mixture of class-based and stratification-based terminology in the course of the empirical analysis that was their main interest. For example, Sorokin (1927, p. 11) wrote: ‘Social stratification means the differentiation of a given population into hierarchically superposed classes.’ Edward Shils (1962, p. 249) wrote: ‘The class system, or the system of stratification of a society, is the system of classes in their internal and external relationships.’ And Melvin Kohn (1969, p. 11) wrote: ‘we use a multidimensional index of class, based on the two dimensions of stratification that appear to be the most important in contemporary American society – occupational position and education.’

Empirical mobility research in sociology and economics

The most important legacy of stratification research is the empirical study of mobility between strata, however defined. Mobility researchers have comprehensively modelled rates and patterns of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
intergenerational and intra-generational movement between strata (see Morgan, Grusky and Fields, 2006, for a review and examples from both sociology and economics). Two approaches to mobility have dominated sociology. In the first, mobility is modelled by accounting for movement between aggregated occupations, sometimes labelled social classes. Accordingly, intergenerational mobility is analysed via inspection of cross-classifications of parents' and children's occupations. In the early literature, levels of mobility were summarized by alternative indices, often derived in the course of analysis of cross-classifications drawn from different societies. The later literature dispensed with mobility indices, focusing instead on the fine structure of patterns of mobility using new log-linear modelling techniques. With the publication of Blau and Duncan's *American Occupational Structure* in 1967, a second approach to the study of mobility reached maturity in sociology, later labelled status attainment research. In this tradition, sociologists focus on the causes and consequences of differences in socio-economic status (often defined in terms of scores attached to occupational titles, based on the average educational attainment and earnings of incumbents). Levels of social mobility are measured by intergenerational correlations of socio-economic status, and these correlations are then decomposed with the use of intervening variables in structural equations models.

An important concern of this literature has been the impact of structural change over time on mobility outcomes and hence the extent to which such change has altered the stratification order. In particular, the degree to which shifts in occupational distributions generate upward mobility has been studied extensively. Such outcomes were welcomed in the middle of the 20th century, and elaborated in scholarship where it was argued that the growth of higher-status occupations is an inevitable outcome of the process of industrialization (and also, by implication, that Marxist claims of the inevitability of class polarization under capitalism had been exaggerated). Perhaps reflecting the growing pessimism and radicalism of sociology in the 1960s and 1970s, such structurally induced upward mobility was deemed less theoretically meaningful than levels of mobility purged of these effects. The study of what came to be known as pure exchange mobility then became possible with the development of log-linear modelling techniques that could be used to ascertain margin-free measures of mobility. This work is best represented by the cross-national research of Erikson and Goldthorpe (1992), which supported the claim that industrialized societies can be characterized by broadly similar patterns of intergenerational occupational mobility (in spite of national mythologies that claim that some societies are more open and meritocratic than others).

In economics, mobility has been a topic of theoretical and empirical work as well, even though it is not connected directly with any tradition of stratification research in sociology. Rather, the early work arose out of labour economics, based on the 'unified approach to intergenerational mobility and inequality' (Becker and Tomes, 1979, p. 1154), which brought together human capital theory with dynastic investment models for family behaviour. As with the status attainment tradition in sociology, economists working in this area often seek single-number expressions for levels of mobility, generally intergenerational correlations of income (although estimated as elasticities from regressions of log earnings across generations; see Solon, 1999). In contrast to the sociological literature, economists have argued recently that there are substantial cross-country differences in mobility, with greater intergenerational mobility of earnings in mainland Europe than in either the United States or the United Kingdom (see Solon, 2002). In addition to the large economics literature on earnings mobility, a well-developed literature on the intergenerational dynamics of wealth inequality now exists (see Mulligan,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1997; Charles and Hurst, 2003).

Economists have begun to focus, like sociologists, on categorical representations of the structure of inequality, examining placement within the distribution of earnings and wealth (using either fixed categories across generations or relative ranks within distributions). When analysed as cross-classifications of quantiles, these methods are similar in spirit and method to the between-social-class mobility studies of sociology. In fact, Björklund and Jäntti (1997) refer to income groups as income classes, and reference the log-linear tradition of mobility research in sociology.

Economists have also become interested in the extent to which increasing inequalities within the labour markets of industrialized countries between the 1970s and the 1990s can be seen as less consequential to the extent that they have been accompanied by increasing chances of intergenerational mobility (see Welch, 1999). Relatedly, some economists have sought to determine the extent to which increasing chances of upward mobility sustained support for the market reforms in eastern Europe and the former Soviet Union that increased inequality (see Birdsall and Graham, 2000). This work is reminiscent of the concern with societal integration that is most closely associated with Parsons in sociology, and it may represent a shared territory which both sociologists and economists will further cultivate.

See Also

- class
- income mobility
- inequality (global)
- inequality (international evidence)
- inequality (measurement)
- intergenerational income mobility
- poverty

Bibliography

Becker, G. and Tomes, N. 1979. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87, 1153–89.

Benabou, R. 1996. Heterogeneity, stratification, and growth: macroeconomic implications of community structure and school finance. *American Economic Review* 86, 584–609.

Birdsall, N. and Graham, C., eds. 2000. *New Markets, New Opportunities? Economic and Social Mobility in a Changing World*. Washington, DC: Brookings Institution.

Björklund, A. and Jäntti, M. 1997. Intergenerational income mobility in Sweden compared to the United States. *American Economic Review* 87, 1009–18.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Blau, P. and Duncan, O. 1967. *The American Occupational Structure*. New York: Wiley.

Bourdieu, P. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press.

Charles, K. and Hurst, E. 2003. The correlation of wealth across generations. *Journal of Political Economy* 111, 1155–82.

Closson, C. 1896. Ethnic stratification and displacement. *Quarterly Journal of Economics* 11, 92–107.

Davis, K. and Moore, W. 1945. Some principles of stratification. *American Sociological Review* 10, 242–9.

Durlauf, S. 1994. Spillovers, stratification, and inequality. *European Economic Review* 38, 836–45.

Erikson, R. and Goldthorpe, J. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford: Oxford University Press.

Heckman, J. and Hotz, V. 1986. An investigation of the labor market earnings of Panamanian males evaluating the sources of inequality. *Journal of Human Resources* 21, 507–42.

Kohn, M. 1969. *Class and Conformity: A Study in Values*. Chicago: University of Chicago Press, 1977.

Morgan, S., Grusky, D. and Fields, G., eds. 2006. *Mobility and Inequality: Frontiers of Research from Sociology and Economics*. Stanford: Stanford University Press.

Mulligan, C. 1997. *Parental Priorities and Economic Inequality*. Chicago: University of Chicago Press.

Myrdal, G. 1944. *An American Dilemma: The Negro Problem and Modern Democracy*. New York: Harper.

Parsons, T. 1940. An analytical approach to the theory of social stratification. *American Journal of Sociology* 45, 841–62.

Parsons, T. 1953. A revised analytical approach to the theory of social stratification. Pp. In *Class, Status, and Power: A Reader in Social Stratification*, ed. R. Bendix and S. Lipset. Glencoe: Free Press.

Parsons, T. 1959. The school class as a social system. In *Social Structure and Personality*, ed. T. Parsons. New York: Free Press, 1964.

Shils, E. 1962. *Center and Periphery: Essays in Macrosociology*. Chicago: University of Chicago Press,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1975.

Solon, G. 1999. Intergenerational mobility in the labor market. In *Handbook of Labor Economics*, vol. 3, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Solon, G. 2002. Cross-country differences in intergenerational earnings mobility. *Journal of Economic Perspectives* 16(3), 59–66.

Sorokin, P. 1927. *Social and Cultural Mobility*. New York: Free Press, 1959.

Welch, F. 1999. In defense of inequality. *American Economic Review* 89, 1–17.

How to cite this article

Morgan, Stephen L. "stratification." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000472> doi:10.1057/9780230226203.1635

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stratified and cluster sampling

Jeffrey M. Wooldridge
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The random sampling paradigm, typically introduced in basic statistics courses, ensures that a sample of data is, loosely speaking, ‘representative’ of the underlying population. When the population parameters are identified, many common estimation techniques, including least squares, maximum likelihood, and instrumental variables, have desirable statistical properties under random sampling. Unfortunately, while random sampling is convenient, it can be, and often intentionally is, violated when cross-sectional data and panel data are collected. Two important deviations from random sampling are stratified sampling and cluster sampling, or perhaps a combination.

Keywords

cluster correlation; cluster sampling; exogenous sampling; heteroskedasticity; multinomial sampling; probability sampling; sampling; stratified sampling; survey sampling; two-stage sampling; unbiased estimators; variable probability sampling; variance; weighted least squares

Article

With stratified sampling, some segments of the population are over- or under-represented by the sampling scheme. For example, a survey of income and demographic characteristics may oversample those with below-median incomes. Intuitively, it is clear that using descriptive statistics from such a sample will not necessarily produce satisfying estimates of population moments. Fortunately, if we know enough information about the stratification scheme, we can often modify standard econometric methods and consistently estimate population parameters.

There are two common types of stratified sampling, namely, standard stratified (SS) sampling and variable probability (VP) sampling. A third type of sampling, typically called multinomial sampling, is practically indistinguishable from SS sampling, but it generates a random sample from a modified population (thereby simplifying certain theoretical analyses). See Cosslett (1993), Imbens and Lancaster (1996), and Wooldridge (1999) for further discussion. We focus on SS and VP sampling here.

SS sampling begins by partitioning the sample space (set of possible outcomes), say **Z**, into *G* non-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

overlapping, exhaustive groups, $\{\mathbf{Z}_g : g = 1, \ldots, G\}$. Then a random sample is taken from each group g, say $\{Z_{gi} : i = 1, \ldots, N_g\}$, where N_g is the number of observations drawn from stratum g and $N = N_1 + N_2 + \cdots + N_G$ is the total number of observations. If Z is a random vector representing the population, and taking values in $\mathbf{Z}$, each random draw from stratum g has the same distribution as Z conditional on Z belonging to $\mathbf{Z}_g$. Therefore, the resulting sample consists of independent but not identically distributed observations. But, unless we are told, we have no way of knowing that our data came from SS sampling.

What if we want to estimate the mean or expected value of Z from an SS sample? It turns out we cannot get an unbiased or consistent estimator of $E(Z)$ unless we have some additional information. Typically, the information comes in the form of population frequencies for each of the strata. Specifically, let $Q_g = P(Z \in \mathbf{Z}_g)$ be the probability that Z falls into stratum g; the Q_g are often called the 'aggregate shares'. For example, if Z represents the distribution of wealth in a country, Q_g is simply the fraction of people in the population whose wealth falls into stratum g (probably defined as intervals in this case). Sometimes very precise estimates of the population frequencies Q_g can be obtained from very large surveys or censuses, in which case they can be treated as known. In some cases, small random samples (sometimes called a 'supplementary sample') are collected and these can be used to estimate the Q_g.

If we know the Q_g (or can consistently estimate them), then $E(Z)$ is identified by a weighted average of the expected values for the strata:

$$\mu_Z \equiv E(Z) = Q_1 E(Z|Z \in \mathbf{Z}_1) + \ldots + Q_G E(Z|Z \in \mathbf{Z}_G) \quad (1)$$

Because we can estimate each of the conditional means using the random sample from the appropriate stratum, an unbiased estimator of μ_Z is simply

$$\hat{\mu}_Z = Q_1 \bar{Z}_1 + Q_2 \bar{Z}_2 \ldots + Q_G \bar{Z}_G \quad (2)$$

where $\bar{Z}_g$ is the sample average for stratum g. Using an asymptotic analysis where the number of observations in each stratum gets large, $\hat{\mu}_Z$ is also a consistent estimator of μ_Z. The variance of $\hat{\mu}_Z$ is easily estimated, too, because

$$Var(\hat{\mu}_Z) = Q_1^2 Var(\bar{Z}_1) + \cdots + Q_G^2 Var(\bar{Z}_G) \quad (3)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and the variance of each sample average can be estimated by the usual unbiased variance estimator within each stratum.

SS sampling is convenient when the members from each stratum are easily identified before sampling. Variable probability sampling is more convenient for telephone or email surveys, where little, if anything, is known ahead of time about those being contacted. With VP sampling, each stratum g is assigned a nonzero sampling probability, p_g. A random draw Z_i is kept with probability p_g if Z_i falls into stratum g. With VP sampling, the population is sampled N_o times. Typically, N_o is not reported along with VP samples. Instead, we know how many data points were kept, and we call this N. Because of the randomness in whether an observation is kept, N is properly viewed as a random variable. As discussed in Wooldridge (1999), it is handy for deriving properties of estimates to define a selection indicator, S_i, which is equal to unity if observation i is kept, and zero otherwise. When S_i=1, we observe the stratum for the observation (or at least its associated sampling probability).

A key formula underlying estimation of the mean from a population under VP sample is

$$E(Z) = N_o^{-1} \sum_{i=1}^{N_o} E[(S_i / p_{g_i}) Z_i]. \quad (4)$$

Because observations within each stratum g are kept with the same probability p_g, it can be shown that $E[(S_i / p_{g_i}) Z_i] = E(Z_i) = \mu_Z$; see, for example, Wooldridge (1999). In other words, weighting an observation by the inverse of its probability of being kept in the sample restores its unbiasedness for the population mean. Equation (4) cannot be used directly in estimation because N_o is typically unknown. A consistent estimator is

$$\hat{\mu}_Z = \frac{1}{\left(\sum_{i=1}^{N} w_i\right)} \sum_{i=1}^{N} w_i Z_i \quad (5)$$

where $w_i = 1 / p_{g_i}$ is the inverse probability weight for the kept observations (N of them): $\hat{\mu}_Z$ is simply a weighted average using the observed data points. Observations with $p_{g_i} = 1$ are always kept and are therefore fully represented in the sample. If, say, $p_{g_i} = 1/3$, the observation needs to be given three times the weight so that its frequency in the sample reflects its population frequency. Wooldridge (1999) uses the selection indicator setup to show that this estimator (and many others that use inverse probability weighting, or IPW) is consistent under VP sampling.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the context of SS sampling, (5) is also consistent if the weights are chosen appropriately. Letting $H_g=N_g/N$ be the fraction of observations falling into stratum g, the appropriate weight is $w_i = Q_{g_i}/H_{g_i}$. This shows explicitly that observations underrepresented in the sample relative to the population ($Q_{g_i} > H_{g_i}$) receive a weight greater than one.

Virtually all standard estimation methods, including least squares, instrumental variables, quantile estimation, maximum likelihood and quasi-ML, can be appropriately modified for SS and VP samples. For example, if we partition Z_i as $Z_i=(X_i,Y_i)$ where Y_i is a scalar and X_i is a row vector, then a linear least squares problem is

$$\min_{\beta} \sum_{i=1}^{N} w_i \cdot (Y_i - X_i\beta)^2. \tag{6}$$

As has been demonstrated by many authors, including Jewell (1985) and Wooldridge (2001), the solution to (6), $\hat{\beta}$, sometimes called a 'weighted least squares' estimator, is consistent for $\beta_o = [E(X'X)]^{-1}E(X'Y)$, the vector in the linear least squares projection of Y on X. Nonlinear regression functions can replace linear regression functions, as demonstrated by Wooldridge (2001). The squared residual in (6) can be replaced with a log-likelihood function, as in Manski and Lerman (1977).

It is important to understand that the weighting in (6) is not used to correct for heteroskedasticity in the underlying population; therefore, one must use care when estimating the asymptotic variance of $\hat{\beta}$. The usual 'sandwich' form of the asymptotic variance estimator, of the type appearing in White (1982), $\hat{A}^{-1}\hat{B}\hat{A}^{-1}$, turns out to be consistent in many cases. (Here, $\hat{A}$ is an estimate of the expected Hessian of the objective function and $\hat{B}$ is outer product of the weighted score.) In particular, the sandwich estimator is consistent under VP sampling with known sampling probabilities; see Wooldridge (1999) for a general treatment that applies to a wide variety of estimation methods. (The sandwich form is also consistent under exogenous sampling of any type, a case we consider in the next section.)

Interestingly, in cases where the usual sandwich estimator is inconsistent, it is always conservative. These include VP sampling with known population frequencies, Q_g, and SS sampling (see, for example, Cosslett, 1993; Imbens and Lancaster, 1996; Wooldridge, 1999; 2001). If the sandwich standard errors are suitably small, one might be satisfied with conservative standard errors. But the adjustment to the usual sandwich estimator is straightforward, provided stratum membership is known for each observation. In constructing $\hat{B}$, one should subtract off the mean of the score within each stratum, which reduces the total variation in the outer product of the score (see Wooldridge, 1999; 2001, for further discussion). Fortunately, statistical software that supports survey sampling commands computes the correct variance matrix estimator provided the strata identifiers are indicated for each observation.

Using modern software, it is easy to analyse data sets that involve both SS and VP sampling. Often, strata are first assigned and then within strata a VP sampling scheme is adopted. The sampling weights reflect both the initial stratification and the variable probability sampling.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Exogenous sampling

When we partition the random vector Z as $Z=(X,Y)$, and we are interested in modelling some feature of the conditional distribution $D(Y|X)$, such as a conditional mean or even the full conditional density, then it makes sense to define the notion of 'exogenous' sampling. For most purposes, it is sufficient to equate exogenous sampling with strata defined in terms of the conditioning variables X only.

Exogenous sampling has three important implications. First, one need not apply weighting to consistently estimate the population parameters. Second, if one does employ weighting, the usual sandwich-type asymptotic variance matrix estimators are consistent. The adjustments used under endogenous sampling described above simply introduce estimation error. Third, there is typically a set of assumptions under which weighting is less efficient than not weighting. For correctly specified maximum likelihood, it is always inefficient to use sampling weights under exogenous sampling. The same is true for nonlinear regression with a correctly specified conditional mean and homoskedasticity. See Wooldridge (2002) for references and further discussion.

Is there a reason to use sampling weights if we know that the sampling is a function of exogenous variables? The answer is 'yes', for two reasons. First, if, say, there is heteroskedasticity in a regression model, it may be more efficient to use sample weights than not, although the most efficient method is to weight based on the form of heteroskedasticity, ignoring the sampling weights. Second, weighting always consistently estimates the parameters of a population optimization problem even in the presence of misspecification. So, if we use linear regression but the conditional mean is not linear, using sample weights nevertheless consistently estimates the parameters in the population linear projection; the unweighted estimator does not.

Cluster sampling

Cluster samples are obtained from one of two basic sampling schemes. One type arises when disaggregated units present themselves naturally as relatively small clusters in the population, and then those clusters are sampled. For example, to study the effects of school inputs on a national fourth-grade mathematics test, we might randomly sample public schools and obtain test results for all students in each school, or for a random sample of students within each school. Under this sampling scheme it makes sense to think of each fourth grader as belonging to his or her cluster (school). While there is randomness in outcomes within a school, typically we think that the outcomes for students within a school will be correlated, due to both observed and unobserved school characteristics. When we apply econometric methods to cluster samples we generally need to account for within-cluster correlation, as was recognized by Scott and Holt (1982) for ordinary least squares. These days, so-called 'cluster-robust' standard errors are routinely computed by many statistical packages.

A cluster sample presents itself in much the same way as a stratified sample: a cluster or group identifier is included for each observation. But, because clusters are sampled, valid inference requires accounting for within-cluster correlation (except in the rare case where it is not present). Typically, this correlation is dealt with through 'cluster-robust' estimation of the matrix B in the middle of the sandwich formula (although, in some cases, generalized least squares-type methods are used). Theoretically, the cluster-robust variance matrix estimators are valid when the number of clusters is 'large' and the cluster sizes are relatively small, although these estimators are employed sometimes for small group sizes. Wooldridge (2003) contains a survey.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sometimes clustering arises in the context of survey sampling, which also involves some kind of probability sampling. Two-stage sampling is the simplest example. In the first stage, one chooses 'primary sampling units' (PSUs), which might be local labour markets, cities, or census block groups, for example. The PSUs are typically not exhaustive because they are selected from a large group of potential PSUs. In the second stage, observations within each PSU are obtained by VP sampling. (If the PSUs are based on a stratification of the initial population, the sampling probabilities reflect that, too.) An observation comes with its PSU identifier and a sampling weight. When sampling is endogenous, weighting is needed to obtain consistent estimators of the population parameters. Additionally, one must account for the within-cluster correlation. For example, we can apply OLS pooled across all observations but with inverse probability weighting. For each g, let M_g denote the number of observations for PSU g. Then, the $M_g \times 1$ vector y_g is the data on the response variable for PSU g and the $M_g \times K$ matrix X_g is the collection of data on the covariates. If $\hat{\beta}$ is the IPW least squares estimator, then its asymptotic variance is estimated by the sandwich form

$$\widehat{Avar(\hat{\beta})} = \left(\sum_{g=1}^{G}\sum_{i=1}^{M_g} X_{gi}' X_{gi} / p_{gi}\right)^{-1} \left(\sum_{g=1}^{G}\sum_{i=1}^{M_g}\sum_{h=1}^{M_g} \hat{u}_{gi}\hat{u}_{gh} X_{gi}' X_{gh} / (p_{gi} p_{gh})\right) \times \left(\sum_{g=1}^{G}\sum_{i=1}^{M_g} X_{gi}' X_{gi} / p_{gi}\right)^{-1} \quad (7)$$

where we need to explicitly index observations by PSU (g) and unit within PSU (i), and the sampling probabilities can differ within PSU. This rather daunting formula is simply a general sandwich form, where the inner term properly accounts for the weighting and within-cluster correlation. Without the probability weights, (7) properly accounts for clustering. If in the middle term we drop the summands with $i \neq h$, and assume constant sampling probabilities within each PSU, then (7) is the proper formula to account for VP sampling without cluster correlation.

See Also

- instrumental variables
- matching estimators
- nonlinear panel data models
- quantile regression
- selection bias and self-selection

Bibliography

Cosslett, S.R. 1993. Estimation from endogenously stratified samples. In *Handbook of Statistics*, Vol. 11, ed. G.S. Maddala, C.R. Rao and H.D. Vinod. Amsterdam: Elsevier.

Imbens, G.W. and Lancaster, T. 1996. Efficient estimation and stratified sampling. *Journal of Econometrics*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

74, 289–318.

Jewell, N.P. 1985. Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika* 72, 11–21.

Manski, C.F. and Lerman, S. 1977. The estimation of choice probabilities from choice-based samples. *Econometrica* 45, 1977–88.

Scott, A.J. and Holt, D. 1982. The effect of two-state sampling on ordinary least squares methods. *Journal of the American Statistical Association* 77, 848–54.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.

Wooldridge, J.M. 1999. Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica* 67, 1385–406.

Wooldridge, J.M. 2001. Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory* 17, 451–70.

Wooldridge, J.M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J.M. 2003. Cluster sample methods in applied econometrics. *American Economic Review* 93, 133–8.

How to cite this article

Wooldridge, Jeffrey M. "stratified and cluster sampling." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_C000557>
doi:10.1057/9780230226203.1636

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

strikes

John Kennan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An economic strike is a suspension of production while workers and their employer argue about how to divide the surplus from their relationship. Modern economic theories of strikes assume that at least one side has private information about the surplus, viewing the lost production as a cost of extracting information. Empirically, strikes are quite rare. There is evidence that strike incidence is high at the peak of the business cycle, but strike duration seems to fall when the economy is strong. Strike activity is evidently influenced by the legislative environment, and particularly by legislation restricting the use of replacement workers.

Keywords

Attrition; bargaining; collective bargaining; employment surplus; Hicks paradox; opportunity cost; private information; replacement workers; screening; signalling; strikes

Article

The value of an employee's labour is generally greater than the wage paid by the employer: that is after all the point of the employment relationship. This gives rise to a surplus to be divided between the worker and the employer. A strike is a suspension of production while the two sides argue about how this surplus is to be divided.

Under ideal competitive conditions, the employment surplus is negligible: each employer competes with many other employers, who bid up the wage until it matches the value of the employee's labour, and each worker competes with many other workers who bid down the wage until it matches the value of the worker's alternative use of time. An employee who strikes for a higher wage is replaced by an equivalent worker who is willing to accept the competitive wage, and an employer who attempts to cut the wage is replaced by another employer who pays the competitive wage.

Thus strikes occur only in non-competitive labour markets, where there is a surplus worth fighting about. Even then, it is not easy to explain why strikes happen. Indeed, if one could explain both the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
occurrence of strikes and the terms of settlement, then strikes would be quite pointless, since the settlement could be reached without the waste associated with strikes. This is sometimes called the Hicks paradox, since the theoretical difficulty of a complete theory of strikes was first articulated by Hicks (1932). As Hicks observed (1932, pp. 146–7), ‘Any means which enables either side to appreciate better the position of the other will make settlement easier; adequate knowledge will always make a settlement possible.’

Building on Hicks's observation, modern economic theories of strikes assume that at least one side has private information about the size of the surplus to be divided. The apparent waste associated with a strike is then seen as a cost of obtaining information.

The main idea can be illustrated by using a simple example. Consider a union negotiating a one-year labour contract with an employer who has private information about the market value of the product being produced. The union is not strong enough to maintain a strike indefinitely, but it can strike for a period of length s (measured in years). Moreover, the union has the power to make offers that the employer must accept or reject. In the most favourable case, v_H is the employer's demand price for labour (that is, the highest wage that the employer would pay), and in the worst case the demand price is v_L, where both are measured relative to the workers’ supply price (which is thus normalized to zero). If these are in fact the only two possibilities, then it is easy to see that the union should demand either a low wage or a high wage, leaving either the low or high employer type indifferent between acceptance and rejection. A strike occurs when the low employer type rejects the high wage demand.

Thus strikes arise when the union is relatively confident that the employer can afford to pay a high wage, but this confidence is in fact misplaced. If p is the probability that the employer is in the high state, the union chooses between v_L for sure, or a higher wage W with probability p. This higher wage leaves the high-state employer indifferent between acceptance, with profit $v_H - W$, or rejection, with profit $(1-s)(v_H - v_L)$, since rejection entails a strike of length s, followed by agreement at the low wage. Thus $W = v_L + s(v_H - v_L)$, and the union threatens a strike if $pW + (1-p)(1-s)v_L > v_L$, that is if

$$p > \frac{v_L}{v_H}.$$

Thus the union's decision as to whether to use the strike threat is influenced by two factors: (a) the probability p that the more favourable state is realized, and (b) the importance of private information, represented by the ratio v_L/v_H, or equivalently by the spread $v_H - v_L$ as a proportion of the opportunity cost v_L. Strengthening either of these factors can tip the balance in favour of the strike threat. Whether a strike actually happens depends on the realized state of demand. Thus an increase in p at some point triggers the use of the strike threat, but further increases in p reduce the probability that a strike will actually occur. The strength of the union, represented by its ability to commit to a strike of length s, has no influence on whether a strike occurs, although it obviously affects the duration of a strike if it does occur, and the terms of settlement if there is no strike.

This simple model is analysed in more detail in Kennan (1986). A much more extensive analysis,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
allowing for private information on both sides, is presented in Kennan and Wilson (1993), with applications to legal conflicts as well as labour negotiations. A general treatment of games with private information, with many bargaining examples, can be found in Myerson (1991).

Three main categories of private information bargaining models can be used to interpret data on labour market negotiations. First, the simple model discussed above is an example of a screening model, in which an uninformed bargainer makes offers that are acceptable only if the informed bargainer knows that the realized surplus is relatively large. More general versions of the screening model assume that once an offer has been rejected, another offer will be made after some specified length of time, since take-it-or-leave-it offers are generally not credible. This leads naturally to a theory of strike durations in which the union makes a declining sequence of wage demands such that in more favourable demand states the employer finds it more profitable to accept an early offer rather than suffer a long strike, and conversely for an employer in a less favourable demand state. In signalling models (the second category), offers are made by bargainers who have private information; this leads to complications arising from a desire to remain ‘inscrutable’, rather than making an offer that reveals valuable information. In attrition models (the third category), the parties fight ‘to the death’ until one side concedes everything: no compromise is allowed.

In a series of papers, Cramton and Tracy (1992; 2003) have presented detailed analyses of collective bargaining negotiations in North America, using a model that includes both screening and signalling components. They emphasize that unions can (and very often do) apply pressure by refusing to sign a new contract after the old contract expires, while continuing to work under the terms of the old contract rather than launching a strike. It is assumed that the employer has private information about the size of the surplus, and that the union makes the first offer. If this offer is refused, the union either continues to work under the old contract, or calls a strike, depending on how optimistic the union is about the size of the surplus relative to the opportunity cost of a strike, as represented by the wage under the old contract. An employer who refuses the initial offer waits some time before making a counter-offer, and the wage settlement then gives each side half or the actual surplus (where the employer's delay is just enough to signal that the surplus is no bigger than it actually is).

Empirical analysis of strike activity

There are many well-known examples of long and hard-fought strikes involving large numbers of workers. But strikes are in fact quite rare, by any measure. Many workers are not covered by collective bargaining agreements; relatively few wage negotiations involve strikes, and most strikes are fairly short. In Britain in 1926 (the year of the General Strike) about nine workdays per worker were lost due to strikes. In 1979, the loss due to strikes was a little more than one day per worker. These are the exceptional cases. In the 79 years following 1926, the number of workdays lost in Britain was fewer than two hours per year per worker. In the United States, idleness due to strikes never exceeded 0.5 per cent of total working days in any year during the period 1948–2005; the average loss was 0.1 per cent per year. Similarly, in Canada over the period 1980–2005, the annual number of work days lost due to strikes never exceeded one day per worker; on average over this period work time lost due to strikes was about one-third of a day per worker. Although the data are not readily available for a broad sample of developed countries, the pattern described above seems quite general: days lost due to strikes amount to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
only a fraction of a day per worker per annum, on average, exceeding one day only in a few exceptional years.
In recent years, the number of workdays lost due to strikes has fallen far below even these low levels. For example, in the United States since 1990, the average loss was about 0.016 per cent, which is about 20 minutes per worker per annum. According to International Labour Organization (ILO) data, similar declines have occurred quite generally in developed countries; even in Spain, which historically has had high rates of strike activity, the average loss since 1990 was about one-fifth of a day per worker per annum. If strikes are caused by private information about rents, as has been argued in the recent theoretical work described above, then a fall in the costs of acquiring information must lead to a decrease in strike activity. It is undeniable that information costs have fallen sharply as computers have improved, and it is tempting to conclude that this is the reason for the decline in strike activity.

Cyclical fluctuations

The relationship between strike activity and business cycle fluctuations is analysed extensively in the literature. The main conclusion from early work in this area is that strikes are more frequent when general economic conditions are good. Although this conclusion is supported by a considerable body of evidence, it is of limited interest because it lumps together strikes of all sorts, including many minor disputes that occur during the term of ongoing labour contracts. More recent work has attempted to determine whether economic conditions affect the incidence of contract strikes. This work largely relies on North American data, because unions in the United States and Canada generally negotiate contracts covering clearly defined periods of a few years, so that one can count the number of negotiations that might lead to a strike, and use this to measure strike incidence.
The empirical results on strike incidence are well summarized by Card (1990); some more recent findings are reviewed by Cramton and Tracy (2003). Surprisingly, the evidence indicates that strike incidence and duration move in opposite directions over the business cycle. Strike incidence is generally found to be pro-cyclical, although the relationship between strikes and general economic conditions is not strong enough to dominate other sources of variation, so that a long time series is needed to establish the result. Although less work has been done on cyclical movements in strike duration, there is solid evidence that duration moves counter-cyclically. Some attempts have also been made to distinguish between the effects of cyclical fluctuations in product markets and in labour markets. There is no clear pattern in these results, and the theoretical significance of the distinction is also unclear. Indeed, if the probability distribution governing the private information about the size of the surplus changes, changes in the incidence and duration of strikes are to be expected, but it should not matter whether the source of this change in the distribution is the product market or the labour market.

Effects of collective bargaining legislation

Strike activity is clearly affected to a large extent by laws governing the tactics available to workers and employers as they negotiate over the division of the employment surplus. This is a large subject in itself, which cannot be dealt with here; Cramton and Tracy (2003) give an overview of some of the main issues with respect to North American legislation. One topic worth considering briefly is the use of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
replacement workers.

The availability of replacement workers directly affects the employment surplus. For example, in the simple screening model described above, suppose that the existing workforce can be permanently replaced, at a cost C (including legal costs, and costs of providing security for the replacement workers). If $\rho < 1$ is the productivity of the replacements, relative to the incumbents, the condition governing the union's strike decision becomes

$$\rho > 1 - \frac{v_H - v_L}{v_H + \frac{C}{1-\delta}}.$$

If the incumbent workers can be replaced without cost, the effect of ρ is merely to scale down the surplus, with no effect on strike incidence. But when *C* is positive, strike incidence falls. The reason for this is that as the cost of replacements increases, the surplus increases by the same amount in both states of demand, so the opportunity cost of a strike rises while the potential gain is unchanged. Thus if the use of permanent replacement workers is banned (or made more difficult), strike incidence falls.

In the case of temporary replacement workers (who are employed only while a strike is going on) the effects on strike incidence may be quite different: a ban on temporary replacements means that the union can obtain a larger share of the surplus in the favourable demand state, so the union makes a more aggressive demand, and strike incidence rises. Cramton and Tracy (2003) review the theoretical implications of banning temporary replacements, and also review the empirical relationship between differences in strike incidence and differences in labour laws. Much remains to be done in this area.

See Also

- bargaining
- collective bargaining
- industrial relations
- litigation, economics of

Bibliography

Card, D. 1990. Strikes and bargaining: a survey of the recent empirical literature. *American Economic Review* 80, 410–15.

Cramton, P.C. and Tracy, J.S. 1992. Strikes and holdouts in wage bargaining: theory and data. *American Economic Review* 82, 100–21.

Cramton, P.C. and Tracy, J.S. 2003. Unions, bargaining and strikes. In *International Handbook of Trade*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Unions, ed. J.T. Addison and C. Schnabel. Cheltenham: Edward Elgar.

Hicks, J. 1932. *The Theory of Wages*, 2nd edn. 1963. New York: Macmillan.

Kennan, J. 1986. The economics of strikes. In *Handbook of Labor Economics*, ed. O. Ashenfelter and R. Layard. New York: North-Holland.

Kennan, J. and Wilson, R. 1993. Bargaining with private information. *Journal of Economic Literature* 31, 45–104.

Myerson, R.B. 1991. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

How to cite this article

Kennan, John. "strikes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000304> doi:10.1057/9780230226203.1637

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Strotz, Robert H. (1922 – 1994)

Robert M. Coen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

adjustment costs; behavioural economics; Cowles Commission; investment; Strotz, R.

Article

Robert Strotz's highly original work in economic theory foreshadowed the contemporary field of behavioural economics and introduced the formal analysis of adjustment costs.

He trained at the University of Chicago, receiving his Ph.D. in 1951. Appointment in 1947 as an instructor at nearby Northwestern University, where he would spend his entire academic career, kept him physically as well as intellectually close to researchers at the Cowles Commission (then in Chicago), from whom he drew inspiration. In 1955 he succeeded Ragnar Frisch as managing editor of *Econometrica*, a post he held until 1968, nurturing the journal in both size and quality to its current pre-eminence.

Strotz's most widely recognized contribution appeared in a commissioned review of investment literature. To analyse effects of interest rates on investment timing, he formulated a model in which interest costs are an increasing function of investment, and applied the calculus of variations to solve for the optimal time path of investment. The paper popularized a new approach to dynamic optimization and clarified links between the short and long runs.

Strotz was the first to analyse the apparent irrationality of consumers who make optimal future plans, then violate them when the future arrives, even though their expectations of future conditions are realized. To stick to plans, they must pre-commit to carrying them out, a common form of behaviour that economists had heretofore been unable to explain. This novel paper was slow to gain recognition but is now viewed as a seminal work in behavioural economics.

Strotz also made contributions to consumer decision theory, distributional ethics, and concepts of causality. His writing was always replete with memorable examples to elucidate difficult points; and a playful wit was often in evidence, for example, when he introduced a paradoxical theorem in distributional ethics with the warning, ‘I wish to indicate … that the views expressed are not necessarily my own’. Had he not largely withdrawn from economics to become dean at Northwestern in 1966 and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

president in 1970, his influence and fame as a scholar would no doubt be even greater.

See Also

- adjustment costs
- consumer expenditure
- demand theory
- intertemporal choice
- simultaneous equations models

Selected works

1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23, 165–80.

1958. How income ought to be distributed: a paradox in distributive ethics. *Journal of Political Economy* 66, 189–205.

1959. The empirical implications of a utility tree. *Econometrica* 25, 269–80.

1960. (With H. Wold.) Recursive vs. nonrecursive systems: an attempt at synthesis. *Econometrica* 28, 417–27.

1963. (With R. Eisner.) Determinants of business investment. In *Commission on Money and Credit: Impacts of Monetary Policy*. Englewood Cliffs, NJ: Prentice Hall.

How to cite this article

Coen, Robert M. "Strotz, Robert H. (1922–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000504>
doi:10.1057/9780230226203.1638

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

structural change, econometrics of

Pierre Perron

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article is concerned with methodological issues related to estimation, testing and computation in the context of structural changes in linear models. The topics covered are: methods related to estimation and inference about break dates for single equations with or without restrictions, with extensions to multi-equations systems where allowance is also made for changes in the variability of the shocks; tests for structural changes including tests for single or multiple changes and tests valid with unit root or trending regressors, and tests for changes in the trend function of a series that can be integrated or trend-stationary.

Keywords

break dates; cointegration; computation; convergence; estimation; heteroskedasticity; hypothesis testing; Lagrange multiplier tests; likelihood ratio; linear models; long-run variance; quasi-maximum likelihood; serial correlation; spurious regression; structural change, econometrics of; temporal dependence; testing; trending variables; unit roots; variance; vector autoregressions; Wald test; Wiener process

Article

This article covers methodological issues related to estimation, testing and computation for models involving structural changes. The amount of work on this subject since the 1950s is truly voluminous in both the statistics and econometrics literature. Accordingly, any survey is bound to focus on specific aspects. Our aim is to review developments as they relate to econometric applications based on linear models. Recently, substantial advances have been made to cover models at a level of generality that allows a host of interesting practical applications. These include models with general stationary regressors and errors that can exhibit temporal dependence and heteroskedasticity, models with trending variables and possible unit roots and cointegrated models, among others. Advances have been made pertaining to computational aspects of constructing estimates, their limit distributions, tests for structural

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

changes, and methods to determine the number of changes present. For a more extensive review the reader is referred to Perron (2006).

We consider the following multiple linear regression with m breaks (or m+1 regimes):

$$y_t = x_t'\beta + z_t'\delta_j + u_t,\ t = T_{j-1} + 1, \ldots, T_j, \quad (1)$$

for j=1,…, m+1. In this model, y_t is the observed dependent variable; both x_t (p×1) and z_t (q×1) are vectors of covariates and β and δ_j (j=1,…, m+1) are the corresponding vectors of coefficients; u_t is the disturbance. The break dates (T_1,…,T_m) are explicitly treated as unknown (the convention that T_0=0 and T_{m+1}=T is used). The purpose is to estimate the unknown regression coefficients together with the break points when T observations on (y_t, x_t, z_t) are available. This is a partial structural change model since the parameter vector β is not subject to shifts. When p=0, we obtain a pure structural change model where all coefficients are subject to change. Note that using a partial structural change model can be beneficial in terms of obtaining more precise estimates and having more powerful tests.

The estimates are obtained by minimizing the overall sum of squared residuals

$$\sum_{i=1}^{m+1} \sum_{t=T_{i-1}+1}^{T_i} [y_t - x_t'\beta - z_t'\delta_i]^2.$$

Let $\hat{\beta}(\{T_j\})$ and $\hat{\delta}(\{T_j\})$ denote the estimates based on the given m-partition (T_1,…,T_m) denoted $\{T_j\}$. Substituting these in the objective function and denoting the resulting sum of squared residuals as S_T(T_1, …,T_m), the estimated break points $(\hat{T}_1, \ldots, \hat{T}_m)$ are such that

$$(\hat{T}_1, \ldots, \hat{T}_m) = \operatorname{argmin}_{(T_1,\ldots,T_m)} S_T(T_1, \ldots, T_m), \quad (2)$$

with the minimization taken over a set of admissible partitions (see below). The parameter estimates are those associated with the partition $\{\hat{T}_j\}$, that is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\hat{\beta} = \hat{\beta}(\{\hat{T}_j\}), \quad \hat{\delta} = \hat{\delta}(\{\hat{T}_j\}).$$

This framework includes many contributions as special cases depending on the assumptions imposed; for example, single change, changes in the mean of a stationary process, and so on. However, since estimation is based on the least-squares principle, even if changes in the variance of u_t are allowed at the same dates as the breaks in the parameters of the regression they are not exploited to increase the precision of the break date estimators unless a quasi-likelihood framework is adopted (see below).

The assumptions and their relevance

To obtain theoretical results about the consistency and limit distribution of estimates of the break dates, some conditions need to be imposed on the regressors, the errors, the set of admissible partitions and the break dates. To our knowledge, the most general set of assumptions in the case of weakly stationary regressors are those in Perron and Qu (2006). Some are simply technical (for example, invertibility requirements), while others restrict the potential applicability of the results.

The assumptions on the regressors specifies that for $w_t = (x_t', z_t')'$, $(1/l_i)\sum_{t=T_i^0+1}^{T_i^0+[l_i v]} w_t w_t' \to pQ_i(v)$ a non-random positive definite matrix uniformly in $v \in [0,1]$. It allows their distribution to vary across regimes. It, however, requires the data to be weakly stationary stochastic processes. This can be relaxed on a case-by-case basis though the proofs then depend on the nature of the relaxation. For instance the scaling used forbids trending regressors, unless they are of the form $\{1,(t/T),\ldots,(t/T)^p\}$, say, for a polynomial trend of order p. Casting trend functions in this form can deliver useful results in many cases. However, there are instances where specifying trends in unscaled form, that is, $\{1,t,\ldots,t^p\}$, can deliver much better results, especially if level and trend slope changes occur jointly. Results using unscaled trends with p=1 are presented in Perron and Zhu (2005). A comparison of their results with other trend specifications is presented in Deng and Perron (2006).

Another important restriction is implied by the requirement that the limit be a fixed, as opposed to stochastic, matrix. This, along with the scaling, precludes integrated processes as regressors (that is, unit roots). In the single break case, this has been relaxed by Bai, Lumsdaine and Stock (1998) who considered structural changes in cointegrated relationships in system of equations. Kejriwal and Perron (2006a) provide general results for multiple structural changes in a single cointegrating vector. Consistency still applies but the rate of convergence and limit distributions of the estimates are different. The assumptions on u_t and $\{w_t u_t\}$ impose mild restrictions on the vector $w_t u_t$ and permits a wide class of potential correlation and heterogeneity (including conditional heteroskedasticity) and lagged dependent variables. It rules out errors that have unit roots. However, unit root errors can be of interest, for example when testing for a change in the deterministic component of the trend function for an integrated series, in which case the estimates are consistent (see Perron and Zhu, 2005). The set of conditions is not the weakest possible. For example, Lavielle and Moulines (2000) allow the errors to be strongly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dependent (long memory processes) but consider only the case of multiple changes in the mean. It is also assumed that the minimization problem defined by (2) is taken over all partitions such that $T_i - T_{i-1} \geq \varepsilon T$ for some $\varepsilon > 0$. This is not restrictive in practice since ε can be small. Another assumption specifies that the break dates are asymptotically distinct, that is, we have $T_i^0 = [T\lambda_i^0]$, where $0 < \lambda_1^0 < \ldots < \lambda_m^0 < 1$. It dictates the asymptotic framework adopted, whereby all segments increase in length in the same proportions as *T* increases.

Under these conditions, the break fractions λ_i^0 are consistently estimated, that is, $\hat{\lambda}_i \equiv (\hat{T}_i / T) \to_p \lambda_i^0$ and that the rate of convergence is *T*. Note that the estimates of the break dates are not consistent themselves, but the differences between the estimates and the true values are bounded by some constant, in probability. Also, this implies that the estimates of the other parameters have the same distribution as would prevail if the break dates were known. Kejriwal and Perron (2006a) obtain similar results with *I* (1) regressors for a cointegrated model subject to multiple changes, using the static regression or a dynamic regression augmented with leads and lags of the first differences of the *I*(1) regressors.

Allowing for restrictions on the parameters

Perron and Qu (2006) consider the issues in a broader framework whereby arbitrary linear restrictions on the parameters of the conditional mean can be imposed in the estimation. The class of models considered is

$$y_t = z_t' \delta_j + u_t, \; t = T_{j-1} + 1, \ldots, T_j,$$

where $R\delta = r$, with *R* a *k* by $(m+1)q$ matrix with rank *k* and *r*, a *k* dimensional vector of constants. The assumptions are the same as discussed above. There is no need for a distinction between variables whose coefficients are allowed to change and those whose coefficients are not allowed to change. A partial structural change model is obtained specifying restrictions that impose some coefficients to be identical across all regimes. This is a useful generalization since it permits a wider class of models of practical interests – for example, a model with a specific number of states less than the number of regimes, or one where a subset of coefficients may be allowed to change over only a limited number of regimes. Perron and Qu (2006) show that the same consistency and rate of convergence results hold. Moreover, the limit distribution of the estimates of the break dates are unaffected by the imposition of valid restrictions, but improvements can be obtained in finite samples. The main advantages of imposing restrictions are that more powerful tests and more precise estimates are obtained.

Method to compute global minimizers

To estimate the model, we need global minimizers of the objective function (2). A standard grid search

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
requires least squares operations of order $O(T^m)$ and becomes prohibitive when the number of breaks is greater than two. Bai and Perron (2003a) discuss a method based on a dynamic programming algorithm that is very efficient (see also Hawkins, 1976). Indeed, the additional computing time needed to estimate more than two break dates is marginal compared with the time needed to estimate a two break model. Consider the case of a pure structural change model. The basic idea is that the total number of possible segments is at most $T(T+1)/2$ and is therefore of order $O(T^2)$. One then needs a method to select which combination of segments yields a minimal value of the objective function. This is achieved efficiently using a dynamic programming algorithm. For models with restrictions (including the partial structural change model), an iterative procedure is available, which in most cases requires very few iterations (see also Perron and Qu, 2006). Hence, even with large samples, the computing cost is small.

The limit distribution of the estimates of the break dates

With the assumptions on the regressors, the errors and given the asymptotic framework adopted, the limit distributions of the estimates of the break dates are independent of each other. Hence, for each break date, the analysis is the same as that of a single break. This holds because the distance between each break increases at rate T, and the mixing conditions on the regressors and errors impose a short memory property so that events that occur a long time apart are independent. This independence property does not hold with integrated data (see below).

The limit distribution of the estimates of the break dates depends on: (*a*) the magnitude of the change in coefficients (with larger changes leading to higher precision); (*b*) the (limit) sample moment matrices of the regressors for the segments pre and post break (allowed to be different); (*c*) the so-called 'long-run' variance of $\{w_t u_t\}$, which accounts for serial correlation in the errors (also allowed to be different pre and post break); (*d*) whether the regressors are trending or not. In all cases, the nuisance parameters can be consistently estimated and appropriate confidence intervals constructed, which need not be symmetric given that the data and errors can have different properties before and after the break.

A feature of the limit distribution is that, for given fixed magnitude of change, it depends on the finite sample distribution of the errors. To get rid of this dependence, the asymptotic framework is modified with the change in parameters getting smaller as T increases, but slowly enough for the estimated break fraction to remain consistent. The limit distribution obtained in Bai (1997a) and Bai and Perron (1998) applies to the case with no trending regressors. With trending regressors, a similar result is still possible (on the assumption of trends of the form (t/T)) and the reader is referred to Bai (1997a) for the case where z_t is a polynomial time trend. For an unscaled linear trend, see Perron and Zhu (2005).

The simulations in Bai and Perron (2006) show that the shrinking shifts asymptotic framework provides useful approximations to the finite sample distributions. The coverage rates are adequate, in general, unless the shifts are quite small, in which case the confidence interval is too narrow. But in such cases, the breaks are unlikely to be detected by test procedures. On the other hand, Deng and Perron (2006) show that the shrinking shift asymptotic framework leads to a poor approximation in the context of a change in a linear trend and that the limit distribution based on a fixed magnitude of shift is preferable.

In a cointegrating regression with I(1) variables, Kejriwal and Perron (2006a) show that, if the coefficients of the integrated regressors are allowed to change, the estimated break fractions are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
asymptotically dependent so that confidence intervals need to be constructed jointly. Methods to construct such confidence intervals are discussed. If, however, only the intercept and/or the coefficients of the stationary regressors are allowed to change, the estimates of the break dates are asymptotically independent.

Estimating breaks one at a time

Bai (1997b) and Bai and Perron (1998) showed that it is possible to consistently estimate all break fractions sequentially, that is, one at a time. When estimating a single break model in the presence of multiple breaks, the estimate of the break fraction will converge to one of the true break fractions, the one that is dominant in the sense that taking it into account allows the greatest reduction in the sum of squared residuals. Then, allowing for a break at the estimated value, a one-break model can be applied to each segment which will consistently estimate the second dominating break, and so on.
Bai (1997b) considers the limit distribution of the estimates and shows that they are not the same as those obtained when estimating all break dates simultaneously. Except for the last estimated break date, the limit distributions depend on the parameters in all segments of the sample. To remedy this problem, he suggested a repartition procedure, which re-estimates each break date conditional on the adjacent break dates. The limit distribution is then the same as when the break dates are estimated simultaneously.

Estimation in a system of regressions

Estimating structural changes in a system of regressions is relatively recent. Bai, Lumsdaine and Stock (1998) consider estimating a single break date in multivariate time series allowing stationary or integrated regressors as well as trends. They show that the width of the confidence interval decreases when series having a common break are treated as a group and estimation is carried using quasi-maximum likelihood (QML). Bai (2000) considers a segmented stationary vector autoregression (VAR) model when the breaks can occur in the parameters of the conditional mean, the covariance matrix of the error term, or both. The most general framework is that of Qu and Perron (2007), who consider models of the form

$$y_t = (I \otimes z_t')S\beta_j + u_t$$

for $T_{j-1} + 1 \le t \le T_j \; (j = 1, \ldots, m+1)$, where y_t is an n-vector of dependent variables and z_t is a q-vector that includes the regressors from all equations, and $u_t \sim (0, \Sigma_j)$. The matrix S is of dimension nq by p with full column rank (usually a selection matrix that specifies which regressors appear in each equation). The set of basic parameters in regime j consists of the p vector β_j and of Σ_j. Qu and Perron (2007) also allow for the imposition of a set of r restrictions of the form $g(\beta, vec(\Sigma))=0$, where

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$\beta = (\beta_1', \ldots, \beta_{m+1}')'$, $\Sigma = (\Sigma_1, \ldots, \Sigma_{m+1})$ and $g(\cdot)$ is an r dimensional vector. Both within- and cross-equation restrictions are allowed, and in each case within or across regimes. The assumptions on the regressors z_t and the errors u_t are similar to those discussed above. Hence, the framework permits a wide class of models including VAR, SUR, linear panel data, change in means of a vector of stationary processes, and so on. Models with integrated regressors (that is, models with cointegration) are not permitted.

Allowing for general restrictions on the parameters β_j and Σ_j permits a very wide range of special cases that are of practical interest: (*a*) partial structural change models, (*b*) block partial structural change models where only a subset of the equations is subject to change; (*c*) changes in only some elements of the covariance matrix Σ_j; (*d*) changes in only the covariance matrix Σ_j, while β_j is the same for all segments; (*e*) models where the breaks occur in a particular order across subsets of equations; and so on.

The method of estimation is again QML (based on normal errors) subject to the restrictions. Qu and Perron (2007) derive the consistency, rate of convergence and limit distribution of the estimated break dates. They obtain a general result stating that, in large samples, the restricted likelihood function can be separated in two parts: one involving only the break dates and the true values of the coefficients, so that the estimates of the break dates are not affected by the restrictions; the other involving the parameters of the model, the true values of the break dates and the restrictions, showing that the limiting distributions of these estimates are influenced by the restrictions but not by the estimation of the break dates. The limit distributions for the estimates of the break dates are qualitatively similar to those discussed above. Though only root-T consistent estimates of (β, Σ) are needed to construct asymptotically valid confidence intervals, it is likely that more precise estimates of these parameters will lead to better finite sample coverage rates. Hence, it is recommended to use the estimates obtained imposing the restrictions even though imposing restrictions does not have a first-order effect on the limiting distributions of the estimates of the break dates. To make estimation possible in practice, Qu and Perron (2007) present an algorithm which extends the one discussed in Bai and Perron (2003a) using, in particular, an iterative generalized least squares (GLS) procedure to construct the likelihood function for all possible segments. The theoretical analysis shows that substantial efficiency gains can be obtained by casting the analysis in a system of regressions.

Qu and Perron (2007) also consider a novel aspect to the problem of multiple structural changes labelled ‘locally ordered breaks’. This applies when the breaks across two equations are ‘ordered’ in the sense that we have the prior knowledge that the break in one equation occurs after the break in the other. The breaks are ‘local’ in the sense that the time span between their occurrence is expected to be short. Hence, the breaks cannot be viewed as occurring simultaneously, nor can the break fractions be viewed as asymptotically distinct. An algorithm to estimate such models is presented. Also, a framework to analyse the limit distribution of the estimates is introduced. Unlike the case with asymptotically distinct breaks, the distributions of the estimates of the break dates need to be considered jointly.

Tests that allow for a single break

To test for a structural change at an unknown date, Quandt (1960) suggested the likelihood ratio test

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

evaluated at the break date that maximizes it. This is a non-standard problem since one parameter is only identified under the alternative hypothesis. This problem was treated under various degrees of specificity that culminated in the general treatment by Andrews (1993). The basic method is to use the maximum of the likelihood ratio test over all possible values of the parameter in some pre-specified set. In the case of a single change, this translates into the statistic $\sup_{\lambda_1 \in \Lambda_\varepsilon} LR_T(\lambda_1)$, where $LR_T(\lambda_1)$ denotes the likelihood ratio evaluated at some $T_1=[T\lambda_1]$ and the maximization is restricted over break fractions that are in the set $\Lambda_\varepsilon=[\varepsilon_1, 1-\varepsilon_2]$. The limit distribution is given by

$$\sup_{\lambda_1 \in \Lambda_\varepsilon} LR_T(\lambda_1) \Rightarrow \sup_{\lambda_1 \in \Lambda_\varepsilon} \frac{[\lambda_1 W_q(1) - W_q(\lambda_1)]'[\lambda_1 W_q(1) - W_q(\lambda_1)]}{\lambda_1(1-\lambda_1)} \tag{3}$$

with $W_q(\lambda)$ a vector of independent Wiener processes of dimension q, the number of coefficients that are allowed to change (this result holds with non-trending data). The limit distribution depends on Λ_ε. If $\varepsilon_1=\varepsilon_2=0$, the test diverges under the null hypothesis, and critical values grow and the power of the test decreases as ε_1 and ε_2 get smaller. Hence, the range over which we search for a maximum must be small enough for the critical values not to be too large and for the test to retain decent power, yet large enough to include break dates that are potential candidates. In the single break case, a popular choice is $\varepsilon_1=\varepsilon_2=.15$.

Andrews (1993) also considered tests based on the maximal value of the Wald and Lagrange multiplier (LM) tests and shows that they are asymptotically equivalent, that is, they have the same limit distribution under the null hypothesis and under a sequence of local alternatives. All tests are also consistent and have non-trivial local asymptotic power against a wide range of alternatives, namely, for which the parameters of interest are not constant over the interval specified by Λ_ε. This does not mean, however, that they all have the same behaviour in finite samples. Indeed, the simulations of Vogelsang (1999) for the special case of a change in mean with serially correlated errors, showed the sup LM_T test to be seriously affected by the problem of non-monotonic power, in the sense that, for a fixed sample size, the power of the test can rapidly decrease to zero as the change in mean increases.

For Model (1) with *i.i.d.* errors, the LR and Wald tests have similar properties, so we shall discuss the Wald test. For a single change, it is defined by (up to a scaling by q):

$$\sup_{\lambda_1 \in \Lambda_\varepsilon} W_T(\lambda_1; q) = \sup_{\lambda_1 \in \Lambda_\varepsilon} \frac{\hat{\delta}' H'(H(\bar{Z}' M_X \bar{Z})^{-1} H')^{-1} H\hat{\delta}}{SSR_k/(T-2q-p)} \tag{4}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $\bar{Z} = diag(Z_1, \ldots, Z_{m+1})$ with $Z_i = (z_{T_{i-1}+1}, \ldots, z_{T_i})'$, H is the conventional matrix such that $(H\delta)' = (\delta_1' - \delta_2')$ and $M_X = I - X(X'X)^{-1}X'$. Here SSR_k is the sum of squared residuals under the alternative hypothesis. Note that break point that maximizes the Wald test is the same as the estimate obtained by minimizing the sum of squared residuals provided the minimization problem (2) is restricted to the set Λ_ε, that is, $\sup_{\lambda_1 \in \Lambda_\varepsilon} W_T(\lambda_1; q) = W_T(\hat{\lambda}_1; q)$. When serial correlation and/or heteroskedasticity in the errors is permitted, the Wald test must be adjusted to account for this. In this case, it is defined by

$$W_T^*(\lambda_1; q) = (T - 2q - p)\hat{\delta}' H'(H\hat{V}(\hat{\delta})H')^{-1}H\hat{\delta}, \tag{5}$$

where $\hat{V}(\hat{\delta})$ is an estimate of the variance covariance matrix of $\hat{\delta}$ that is robust to serial correlation and heteroskedasticity; that is, a consistent estimate of

$$V(\hat{\delta}) = \text{plim}_{T \to \infty} T(\bar{Z}' M_X \bar{Z})^{-1} \bar{Z}' M_X \Omega M_X \bar{Z}(\bar{Z}' M_X \bar{Z})^{-1} \tag{6}$$

where Ω is the covariance matrix of the errors. Note that it can be constructed allowing identical or different distributions for the regressors and the errors across segments. This is important because if a variance shift occurs at the same time and is not taken into account, inference can be distorted (Pitarakis, 2004).

The computation of the robust version of the Wald test (5) can be involved. Since the estimate of λ_1 is T-consistent even with correlated errors, an asymptotically equivalent version is to first take the supremum of the original Wald test, as in (4), to obtain the break point, that is imposing $\Omega = \sigma^2 I$. The robust version is obtained by evaluating (5) and (6) at this estimated break date, that is, using $W_T^*(\hat{\lambda}_1; q)$ instead of $\sup_{\lambda_1 \in \Lambda_\varepsilon} W_T^*(\lambda_1; q)$, where $\hat{\lambda}_1$ is obtained by minimizing the sum of squared residuals over the set Λ_ε. This is especially convenient when testing for multiple structural changes.

Optimal tests

Andrews and Ploberger (1994) consider a class of tests that are optimal, in the sense that they maximize

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
a weighted average of the local asymptotic power function. They are weighted functions of the standard Wald, LM or LR statistics for all permissible fixed break dates. Using either of the three basic statistics leads to tests that are asymptotically equivalent. Here, we shall proceed with the version based on the Wald test. On the assumption that equal weights are given to all break fractions in some trimmed interval [$\varepsilon_1, 1-\varepsilon_2$], the optimal test for distant alternatives is the following so-called Exp-type test

$$Exp-W_T = \log\left(T^{-1}\sum_{T_1=[T\varepsilon_1]+1}^{T-[T\varepsilon_2]} \exp\left(\frac{1}{2}W_T(T_1/T)\right)\right).$$

For alternatives close to the null value of no change the optimal test is the *Mean*-W_T test

$$Mean-W_T = T^{-1}\sum_{T_1=[T\varepsilon_1]+1}^{T-[T\varepsilon_2]} W_T(T_1/T)$$

Andrews and Ploberger (1994) provide critical values for both tests for a range of values for symmetric trimmings $\varepsilon_1 = \varepsilon_2$ (which can be used for some non symmetric trimmings as well). The *Mean*-W_T has highest power for small shifts and the *Exp*-W_T performs better for moderate to large shifts. None of them uniformly dominates the *Sup*-W_T test and Andrews and Ploberger (1994) recommend the *Exp*-W_T test. The *Sup*-W_T test is not a member of the class of tests that maximize some weighted version of the local asymptotic power function, though it is admissible.
Kim and Perron (2006) approach the optimality issue from a different perspective using the approximate Bahadur measure of efficiency. They show that tests based on the Mean functional are inferior to those based on the Sup and Exp (which are as efficient) when using the same base statistic. When considering tests that incorporate a correction for potential serial correlation in the errors: (*a*) for a given functional, using the LM statistic leads to tests with zero asymptotic relative efficiency compared with using the Wald statistic; (*b*) for a given statistic the Mean-type tests have zero relative efficiency compared to using the Sup and Exp versions, which are as efficient. These results are in contrast to those of Andrews and Ploberger (1994) and the practical implication is that the preferred tests should be the Sup or Exp–Wald tests. Any test based on the LM statistic should be avoided.

Non-monotonicity in power

The Sup–Wald and Exp–Wald tests have monotonic power when only one break occurs under the alternative. As shown in Vogelsang (1999), the Mean–Wald test can exhibit a non-monotonic power

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

function, though the problem has not been shown to be severe. All of these, however, suffer from some important power problems when the alternative is one that involves two breaks (for example, Vogelsang, 1997). This suggests that a test will exhibit a non-monotonic power function if the number of breaks present under the alternative hypothesis is greater than the number of breaks explicitly accounted for in the construction of the tests. Hence, though a single break test is consistent against multiple breaks, substantial power gains can result from using tests for multiple structural changes.

Tests for multiple structural changes

The literature on tests for multiple structural changes is relatively scarce. Here, the problem with the *Mean*-W_T and *Exp*-W_T tests is practical implementation as they require the computation of the Wald test over all permissible partitions of the sample, a number of order $O(T^m)$, which is prohibitively large when $m>2$. Consider instead the Sup-Wald test. With i.i.d. errors, maximizing the Wald statistic is equivalent to minimizing the sum of squared residuals when the search is restricted to the same possible partitions of the sample. As discussed above, this problem can be solved with a very efficient algorithm. This is the approach taken by Bai and Perron (1998). In the context of model (1) with i.i.d. errors, the Wald test for testing the null hypothesis of no change versus the alternative hypothesis of k changes is given by

$$W_T(\lambda_1, \ldots, \lambda_k; q) = \left(\frac{T-(k+1)q-p}{k}\right)\frac{\hat{\delta}'H'(H(\bar{Z}'M_X\bar{Z})^{-1}H')^{-1}H\hat{\delta}}{SSR_k}$$

where H now is the matrix such that $(H\delta)' = (\delta_1' - \delta_2', \ldots, \delta_k' - \delta_{k+1}')$. The Sup–Wald test is defined by

$$\sup_{(\lambda_1,\ldots,\lambda_k)\in\Lambda_{k,\varepsilon}} W_T(\lambda_1, \ldots, \lambda_k; q) = W_T(\hat{\lambda}_1, \ldots, \hat{\lambda}_k; q),$$

where

$$\Lambda_{k,\varepsilon} = \{(\lambda_1, \ldots, \lambda_k); |\lambda_{i+1} - \lambda_i| \geq \varepsilon, \lambda_1 \geq \varepsilon, \lambda_k \leq 1-\varepsilon\}$$

and $(\hat{\lambda}_1, \ldots, \hat{\lambda}_k) = (\hat{T}_1/T, \ldots, \hat{T}_k/T)$, with $(\hat{T}_1, \ldots, \hat{T}_k)$ the estimates of the break dates obtained by minimizing the sum of squared residuals by searching over partitions defined by the set $\Lambda_{k,\varepsilon}$. When

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

serial correlation and/or heteroskedasticity in the residuals is allowed, the test is

$$W_T^*(\lambda_1, \ldots, \lambda_k; q) = \left(\frac{T-(k+1)q-p}{k}\right)\hat{\delta}'H'(H\hat{V}(\hat{\delta})H')^{-1}H\hat{\delta},$$

with $\hat{V}(\hat{\delta})$ as defined by (6). Again, the asymptotically equivalent version with the Wald test evaluated at the estimates $(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ is used to make the problem tractable. The limit distribution of the tests under the null hypothesis is the same in both cases, again on the assumption of non-trending data. Critical values are presented in Bai and Perron (1998; 2003b). The importance of the choice of ε for the size and power of the test is discussed in Bai and Perron (2003a; 2006). They also discuss variations in the construction of the test that allow imposing various restrictions on the nature of the errors and regressors, which can help improve power.

Double maximum tests

Often, one may not wish to pre-specify a particular number of breaks. Then a test of the null hypothesis of no structural break against an unknown number of breaks given some upper bound *M* can be used. These are called ‘double maximum tests’. The first is an equal-weight version defined by $UD\max W_T(M, q) = \max_{1 \le m \le M} W_T(\hat{\lambda}_1, \ldots, \hat{\lambda}_m; q)$. The second test applies weights to the individual tests such that the marginal p-values are equal across values of *m* and is denoted *WD* max $F_T(M, q)$ (see Bai and Perron, 1998, for details). The choice *M*=5 should be sufficient for most applications. In any event, the critical values vary little as *M* is increased beyond 5.
The double maximum tests are arguably the most useful to apply when trying to determine if structural changes are present. First, there are types of multiple structural changes that are difficult to detect with a single break test (for example, two breaks with the first and third regimes the same). Second, as discussed above, is the potential non-monotonic power problem when the number of changes is greater than specified. Third, the power of the double maximum tests is almost as high as the best power that can be achieved using the test that accounts for the correct number of breaks (for example, Bai and Perron, 2006).

Sequential tests

Bai and Perron (1998) also discuss a test of • versus •+1 breaks, which can be used to estimate the number of breaks using a sequential testing procedure. For the model with • breaks, the estimated break points denoted by $(\hat{T}_1, \ldots, \hat{T}_\ell)$ are obtained by a global minimization of the sum of squared residuals. The strategy proceeds by testing for the presence of an additional break in each of the (•+1) segments obtained using the partition $\hat{T}_1, \ldots, \hat{T}_\ell$. We reject in favour of a model with (•+1) breaks if the minimal

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
value of the sum of squared residuals over all segments where an additional break is included is sufficiently smaller than that from the • breaks model. The break date selected is the one associated with this overall minimum. The limit distribution of the test is related to that of a test for a single change. Bai (1999) considers the same problem allowing the breaks to be global minimizers of the sum of squared residuals under both the null and alternative hypotheses. The limit distribution of the test is different. A method to compute the asymptotic critical values is discussed and the results extended to the case of trending regressors.

These tests can form the basis of a sequential testing procedure by applying them successively starting from •=0, until a non-rejection occurs. The estimate of the number of breaks thus selected will be consistent provided the significance level used decreases at an appropriate rate. The simulation results of Bai and Perron (2006) show that such estimate of the number of breaks is better than those obtained using information criteria as suggested by, for example, Liu, Wu and Zidek (1997) (see also Perron, 1997). But this sequential procedure should not be applied mechanically. In several cases, it stops too early. The recommendation is to first use a double maximum test to ascertain if any break is at all present. The sequential tests can then be used starting at some value greater than 0 to determine the number of breaks.

Tests for restricted structural changes

Consider testing the null hypothesis of 0 break versus an alternative with *k* breaks in a model which imposes the restrictions $R\delta = r$. In this case, the limit distribution of the Sup–Wald test depends on the nature of the restrictions so that it is not possible to tabulate critical values valid in general. Perron and Qu (2006) discuss a simulation algorithm to compute the relevant critical values given some restrictions. Imposing valid restrictions results in tests with much improved power.

Tests for structural changes in multivariate systems

Bai, Lumsdaine and Stock (1998) considered a Sup–Wald test for a single change in a multivariate system. Qu and Perron (2007) extend the analysis to the context of multiple structural changes. They consider the case where only a subset of the coefficients is allowed to change, whether it be the parameters of the conditional mean, the covariance matrix of the errors, or both. The tests are based on the maximized value of the likelihood ratio over permissible partitions assuming i.i.d. errors. The tests can be corrected for serial correlation and heteroskedasticity when testing for changes in the parameters of the conditional mean assuming no change in the covariance matrix of the errors. However, when the tests involve potential changes in the covariance matrix of the errors, the limit distributions are only valid assuming a Normal distribution for these errors.

An important advantage of the general framework analysed by Qu and Perron (2007) is that it allows studying changes in the variance of the errors in the presence of simultaneous changes in the parameters of the conditional mean, thereby avoiding inference problems when changes in variance are studied in isolation. Also, it allows for the two types of changes to occur at different dates, thereby avoiding problems related to tests for changes in the parameters when a change in variance occurs at some other

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
date.
These tests are especially important in light of Hansen's (2000) analysis. First note that the limit distribution of the tests in a single equation system has the stated limit distribution under the assumption that the regressors and the variance of the errors have distributions that are stable across the sample. Hansen shows that when the regressors are not stationary the limit distribution changes and the tests can be distorted, especially when a change in variance occurs. He proposes a fixed regressor bootstrap method to construct valid tests. But both problems of changes in the distribution of the regressors and the variance of the errors can be handled using the framework of Qu and Perron (2007). If a change in the variance of the residuals is a concern, one can perform a test for no change in some parameters of the conditional model allowing for a change in variance since the tests are based on a likelihood ratio approach. If changes in the marginal distribution of some regressors are a concern, one can use a multi-equations system with equations for these regressors.

Tests valid with I(1) regressors

With *I*(1) regressors, a case of interest is a system of cointegrated variables. For testing, Hansen (1992) considered the null hypothesis of no change in both coefficients. The tests considered are the Sup and Mean–LM tests directed against an alternative of a one time change in the coefficients. Hansen also considers a version of the LM test directed against the alternative that the coefficients are random walk processes. Kejriwal and Perron (2006b) provide a comprehensive treatment of issues related to testing for multiple structural changes at unknown dates in cointegrated regression models using the Sup–Wald test. They allow both *I*(0) and *I*(1) variables and derive the limiting distribution of the Sup–Wald test under the null hypothesis of no structural change against the alternative hypothesis of a given number of cointegrating regimes. They also consider the double maximum tests and provide critical values for a wide variety of models that are expected to be relevant in practice. The asymptotic results have important implications for inference. It is shown that, in models involving both *I*(1) and *I*(0) variables, inference is possible as long as the intercept is allowed to change across regimes. Otherwise, the limiting distributions of the tests depend on nuisance parameters. Simulation experiments show that with serially correlated errors the commonly used Sup, Mean and Exp–LM tests suffer from the problem of non-monotonic power in finite samples both with a single and multiple breaks. Kejriwal and Perron (2006b) propose a modified Sup–Wald test that has good size and power properties.
Note, however, that the Sup and Mean–Wald test will also reject when no structural change is present and the system is not cointegrated. Hence, the application of such tests should be interpreted with caution. No test are available for the null hypothesis of no change in the coefficients allowing the errors to be *I*(0) or *I*(1). This is because when the errors are *I*(1), we have a spurious regression and the parameters are not identified. To be able to properly interpret the tests, they should be used in conjunction with tests for the presence or absence of cointegration allowing shifts in the coefficients (see the discussion and references in Perron, 2006). A partial solution to this problem is the following. If a spurious regression is present, the number of breaks selected will always (in large samples) be the maximum number of breaks allowed. Thus, selecting the maximum allowable number of breaks can be indicative of the presence of *I*(1) errors (using a Sup–Wald test uncorrected for serial correlation in the errors). The same is true when information criteria are used to select the number of breaks.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Tests valid whether the errors are I(1) or I(0)

The issue of testing for structural changes in a linear model with errors that are either *I*(0) or *I*(1) is interest when the regression is a polynomial time trend (for example, testing for a change in the slope of a linear trend). The problem here is to devise a procedure that has the same limit distribution in both the *I*(0) and *I*(1) cases. The first to provide such a solution is Vogelsang (2001). He also accounts for correlation with an autoregressive approximation so that the Wald test has a non-degenerate limit distribution in both the *I*(0) and *I*(1) cases. The novelty is that he weights the statistic by a unit root test scaled by some parameter. For any given significance level, a value of this scaling parameter can be chosen so that the asymptotic critical values will be the same. Vogelsang's simulations show, however, the test to have little power in the *I*(1) case so that he resorts to advocating the joint use of that test and a normalized Wald test that has good properties in the *I*(1) case but has otherwise very little power in the *I*(0) case.
Perron and Yabu (2007b) builds on the work of Perron and Yabu (2007a) who analysed the problem of hypothesis testing on the slope coefficient of a linear trend model. The method is based on a feasible quasi generalized least squares approach that uses a super-efficient estimate of the sum of the autoregressive parameters α when $\alpha=1$. The estimate of α is the OLS estimate from an autoregression applied to detrended data and is truncated to take a value 1 whenever it is in a $T^{-\delta}$ neighbourhood of 1. This makes the estimate 'super-efficient' when $\alpha=1$ and implies that inference can be performed using the standard Normal or Chi-square distribution for all $|\alpha| \leq 1$. Theoretical arguments and simulation evidence show that $\delta=1/2$ is the appropriate choice. Perron and Yabu (2007b) analyse the case of testing for changes in level or slope of the trend function of a univariate time series. When the break dates are known, things are similar. When the break dates are unknown, the limit distributions of the Exp, Mean and Sup functionals of the Wald test across all permissible breaks dates is no longer the same in the *I*(0) and *I*(1) cases. However, the limit distribution is nearly the same using the Exp functional. Hence, it is possible to have tests with nearly the same size in both cases. To improve the finite sample properties of the test, use is made of a bias-corrected version of the OLS estimate. This makes possible a testing procedure that has good size and power properties in finite samples.

Summary

There has been tremendous progress since the early 1990s in developing methods to analyse structural changes for a variety of cases that are of practical interest. Still, much remains to be done, in particular in providing tools to analyse changes in the variance of the errors without the need to assume a Normal distribution.

See Also

- cointegration
- dummy variables
- linear models

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- regime switching models
- regression-discontinuity analysis
- structural change
- time series analysis
- unit roots

Bibliography

Andrews, D.W.K. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821–56 (Corrigendum, 71, 395–7).

Andrews, D.W.K. and Ploberger, W. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383–414.

Bai, J. 1997a. Estimation of a change point in multiple regression models. *Review of Economic and Statistics* 79, 551–63.

Bai, J. 1997b. Estimating multiple breaks one at a time. *Econometric Theory* 13, 315–52.

Bai, J. 1999. Likelihood ratio tests for multiple structural changes. *Journal of Econometrics* 91, 299–323.

Bai, J. 2000. Vector autoregressive models with structural changes in regression coefficients and in variance-covariance matrices. *Annals of Economics and Finance* 1, 303–39.

Bai, J., Lumsdaine, R.L. and Stock, J.H. 1998. Testing for and dating breaks in multivariate time series. *Review of Economic Studies* 65, 395–432.

Bai, J. and Perron, P. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66, 47–78.

Bai, J. and Perron, P. 2003a. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18, 1–22.

Bai, J. and Perron, P. 2003b. Critical values for multiple structural change tests. *Econometrics Journal* 6, 72–8.

Bai, J. and Perron, P. 2006. Multiple structural change models: a simulation analysis. In *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, ed. D. Corbea, S. Durlauf and B.E. Hansen. New York: Cambridge University Press.

Deng, A. and Perron, P. 2006. A comparison of alternative asymptotic frameworks to analyze structural

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
change in a linear time trend. *Econometrics Journal* 9, 423–47.

Hansen, B.E. 1992. Tests for parameter instability in regressions with I(1) processes. *Journal of Business and Economic Statistics* 10, 321–35.

Hansen, B.E. 2000. Testing for structural change in conditional models. *Journal of Econometrics* 97, 93–115.

Hawkins, D.M. 1976. Point estimation of the parameters of piecewise regression models. *Applied Statistics* 25, 51–7.

Kim, D. and Perron, P. 2006. Assessing the relative power of structural break tests using a framework based on the approximate Bahadur slope. Working paper 2006-63, Boston University.

Kejriwal, M. and Perron, P. 2006a. The limit distribution of the estimates in cointegrated regression models with multiple structural changes. Working paper 2006-64, Boston University.

Kejriwal, M. and Perron, P. 2006b. Testing for multiple structural changes in cointegrated regression models. Working paper 2006-51, Boston University.

Lavielle, M. and Moulines, E. 2000. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis* 21, 33–59.

Liu, J., Wu, S. and Zidek, J.V. 1997. On segmented multivariate regressions. *Statistica Sinica* 7, 497–525.

Perron, P. 1997. L'estimation de modèles avec changements structurels multiples. *Actualité Économique* 73, 457–505.

Perron, P. 2006. Dealing with structural breaks. In *Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory*, ed. K. Patterson and T.C. Mills. Basingstoke: Palgrave Macmillan.

Perron, P. and Qu, Z. 2006. Estimating restricted structural change models. *Journal of Econometrics* 134, 373–99.

Perron, P. and Yabu, T. 2007a. Estimating deterministic trends with an integrated or stationary noise component. Working paper 2007-21, Boston University.

Perron, P. and Yabu, T. 2007b. Testing for shifts in trend with an integrated or stationary noise component. Working paper 2007-25, Boston University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Perron, P. and Zhu, X. 2005. Structural breaks with stochastic and deterministic trends. *Journal of Econometrics* 129, 65–119.

Pitarakis, J.-Y. 2004. Least squares estimation and tests of breaks in mean and variance under misspecification. *Econometrics Journal* 7, 32–54.

Qu, Z. and Perron, P. 2007. Estimating and testing multiple structural changes in multivariate regressions. *Econometrica* 75, 459–502.

Quandt, R.E. 1960. Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association* 55, 324–30.

Vogelsang, T.J. 1997. Wald-type tests for detecting breaks in the trend function of a dynamic time series. *Econometric Theory* 13, 818–49.

Vogelsang, T.J. 1999. Sources of nonmonotonic power when testing for a shift in mean of a dynamic time series. *Journal of Econometrics* 88, 283–99.

Vogelsang, T.J. 2001. Testing for a shift in trend when serial correlation is of unknown form. Working paper, Cornell University.

How to cite this article

Perron, Pierre. "structural change, econometrics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000512>
doi:10.1057/9780230226203.1640

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

structural change

Kiminori Matsuyama
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Structural change is a complex, intertwined phenomenon, not only because economic growth brings about complementary changes in various aspects of the economy, such as the sector compositions of output and employment and the organization of industry, but also because these changes in turn affect the growth process. Using a simple two-sector model, we highlight some driving forces behind structural change, attempt to convey the complexity of the phenomenon and identify some key issues discussed in the literature.

Keywords

agricultural productivity growth; development failure; economic growth; Engel's Law; exogenous productivity growth; product diversity; resource curse; Rostow, W.; rural–urban migration; specialization; stages theory of growth; staple trap; structural change; total factor productivity

Article

Structural change can occur as a consequence of significant shocks, such as plagues, wars, revolutions, the discovery of a continent, and major technological advances. Here, however, we confine ourselves to the structural change experienced by an economy over the course of its development. It is a complex, intertwined phenomenon, not only because economic growth brings about complementary changes in various aspects of the economy, such as the sector compositions of output and employment, the organization of industry, the financial system, income and wealth distribution, demography, political institutions, and even the society's value system, but also because these changes can in turn affect the growth processes.

Earlier work on the subject attempted to establish some stylized facts, that is, the patterns of development followed by most countries. Among the best known are Fisher (1939), Clark (1940), Kuznets (1966) and Chenery and Syrquin (1975), who postulated that, as the economy grows, the production shifts from the primary (agriculture, fishing, forestry, mining) to the secondary (manufacturing and construction) to the tertiary sector (services). Also notable is Rostow (1960), who argued that the economy passes through various stages of development, from the traditional stage to the take-off stage to the mass consumption stage. This literature is mostly descriptive, trying to provide a sweeping overview of the development process, with the emphasis on the multifaceted nature of structural change.

In contrast, recent work tends to be more analytical, using formal models designed to focus on a few specific aspects of structural change. There is also an increasing awareness that the two-way causality between economic growth and structural change can provide possible explanations for development failures.

From the rural agricultural society to the urban industrial society

Structural change caused by exogenous productivity growth

To illustrate some of the driving forces behind structural change, consider a simple two-sector model, adopted from Matsuyama (1992a). The *j*-th sector (*j*=1,2) produces its output

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

with $Y_j = A_j F_j(n_j)$, where A_j is its total factor productivity (TFP), and $F_j(\cdot)$ is an increasing, concave production function, and n_j is the employment share, with $n_1 + n_2 = 1$. Consumers have Stone–Geary preferences, $U(C_1, C_2) = \beta \log(C_1 - \gamma) + \log(C_2)$ with $\gamma > 0$. In competitive equilibrium, the marginal values of labour in the two sectors are equalized;

$$A_1 F_1'(n) = p A_2 F_2'(1 - n), \tag{1}$$

where p is the relative price of good 2 and $n = n_1 = 1 - n_2$ is the first sector's employment share. Since the consumer demand satisfies $C_1 = \gamma + (\beta p) C_2$, the goods market equilibrium (in the closed economy) is given by $A_1 F_1(n) = \gamma + A_2 F_2(1 - n)(\beta p)$. Combining the two conditions yields

$$F_1(n) - \beta F_2(1 - n) F_1'(n) / F_2'(1 - n) = \gamma / A_1, \tag{2}$$

which implicitly defines n as a decreasing function of A_1, $n=N(A_1)$. By interpreting the first sector as agriculture and the second as industry, this offers one explanation for the transformation from the decline of agriculture and the rise in industry: Engel's Law. Because the demand for agricultural goods has lower income elasticity than the demand for manufacturing goods, agricultural productivity growth helps to release labour for industry. This mechanism plays an important role in, for example, Murphy, Shleifer and Vishny (1989), Matsuyama (1992a, s. 2), Laitner (2000), Caselli and Coleman (2001), and Gollin, Parente and Rogerson (2002).

The above argument suggests that productivity gains in agriculture push the workers out of agriculture. There is an entirely opposite view that productivity gains in industry pull the workers out of agriculture. To capture this view, let us change the consumer's preferences to $U(C_1, C_2) = C_1 + pC_2$, with p now given exogenously. One interpretation is that the economy has the two different techniques to produce a single consumption good. The first is traditional, land-based, craft production, and the second is modern, capitalistic, manufacturing. In this case, eq. (1) alone determines the equilibrium allocation, which means that n increases with A_1/A_2. Thus, faster productivity growth in the modern sector (or relative stagnation in the traditional sector) induces more workers to abandon the traditional sector. This captures in essence the structural change mechanism envisioned by Lewis (1954) and many others (see Hansen and Prescott, 2002, for a neoclassical treatment). Alternatively, this case can be interpreted as the case of a small open economy, where the two sectors, agriculture and industry, produce different goods, but the relative price p is determined exogenously in the world market. Then, a higher A_1/A_2 increases n, by shifting the country's comparative advantage towards agriculture, contrary to what Engel's Law suggests.

Productivity growth caused by structural change

Let us introduce some dynamics into the above model by making the total factor productivity (TFP) of the second sector endogenous. More specifically, let $A_{2t} = A(Q_t)$, where $A(\cdot)$ is an increasing function, and Q_t is the stock of the experience accumulated in the second sector through learning-by-doing, and follows the law of motion, $dQ_t / dt = H(1 - n_t)$, where $H(\cdot)$ is an increasing function with $H(1 - n_c) = 0$. This captures the idea that, with higher employment, the firms in the second sector can learn faster; but without a critical mass of employment, $1–n_c$, productivity declines. Let us also assume that learning-by-doing is external, so that the firms take A_{2t} as given when choosing their level of employment. For simplicity, A_1 is assumed to be exogenous and constant.

Then, the equilibrium condition in the closed economy is given by eq. (2) at any t, so that $n_t = N(A_1)$ and $dQ_t / dt = H(1 - N(A_1))$, which is increasing in A_1. In other words, a higher A_1, by releasing labour from the first sector, leads to faster productivity growth in the second sector. This captures a version of ‘the staple theory of growth’ (Watkins, 1963), which argues that a productive primary sector triggers the growth of the industry. In a small open economy, however, the equilibrium condition is $A_1 F_1'(n_t) = pA(Q_t) F_2'(1 - n_t)$,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which suggests that a higher A_1 implies a higher n_t for any level of Q_t. For $A_1 < pA(Q_t)F_2'(1-n_c)/F_1'(n_c)$, $n_t < n_c$ and hence $H(1-n_t) > 0$, which leads to productivity growth in industry and a steady decline in n_t. For $A_1 > pA(Q_t)F_2'(1-n_c)/F_1'(n_c)$, $n_t > n_c$ and hence $H(1-n_t) < 0$, which leads to a productivity decline in industry and a steady increase in n_t. This suggests the so-called *staple trap* or *resource curse*, the situation where the abundance of natural resources prevents the country from growing. Indeed, even a temporary boom in the resource sector could lead to a permanent decline in industry, the so-called *Dutch Disease* (see Matsuyama, 1992a, s. 3, for more detailed analysis).

Impediments to structural change

The above analysis assumes perfect labour mobility across sectors, equating the marginal value of labour instantly. Many studies have modelled various impediments that slow down the reallocation of labour. In the Lewis (1954) model, the workers earn the average (not marginal) value of labour in the traditional sector, which causes its overemployment. In the Harris and Todaro (1970) model, moving to the urban area is necessary but not sufficient to find a high-wage job in the modern sector, which leads to the urban–rural wage gap, offset by the risk of unemployment in the urban area. In Matsuyama (1991, 1992b), only the young can migrate to the urban sector. In Banerjee and Newman (1998), credit constraints prevent some workers from moving. In Caselli and Coleman (2001) and Lucas (2004), the frictions come from the need to acquire skill or accumulate human capital. These models have implications that a reduction in such frictions accelerates structural change.

The circular causality between productivity growth and structural change

Even without any frictions, structural change and development may fail to materialize due to the circular causality. As pointed out above, productivity growth can cause structural change, which in turn leads to further growth in productivity. The circular causality, however, is a double-edged sword, as lack of productivity growth and lack of structural change can reinforce each other, creating a vicious circle of poverty. To illustrate this point, let us go back to a version of the above model, where the economy has traditional and modern sectors, both producing perfectly substitutable goods, so that the equilibrium is given solely by eq. (1). Now, modify it by assuming that the modern sector is subject to economies of scale, so that its TFP increases with its employment share, as $A_2 = A(1-n)$. For simplicity, let us assume that higher productivity is entirely due to external economies so that the firms take A_2 as given when choosing their level of employment. Then, the equilibrium is given by $A_1F_1'(n) = pA(1-n)F_2'(1-n)$. This could generate multiple equilibria, each corresponding to a different level of development, as shown in Figure 1. One of them, E, may be viewed as the state of underdevelopment, where the modern sector cannot attract much workers due to low productivity, and hence cannot take advantage of the scale economies, implying low productivity. Yet there is another equilibrium, E', characterized by high productivity and a high employment share of the modern sector. A move from E to E' implies both productivity growth and a change in sectoral composition, generating the same observations as models of structural change with exogenous productivity growth; but, in this model, the causality goes in both directions.

Figure 1

Development failures due to two-way causality between productivity growth and structual change

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

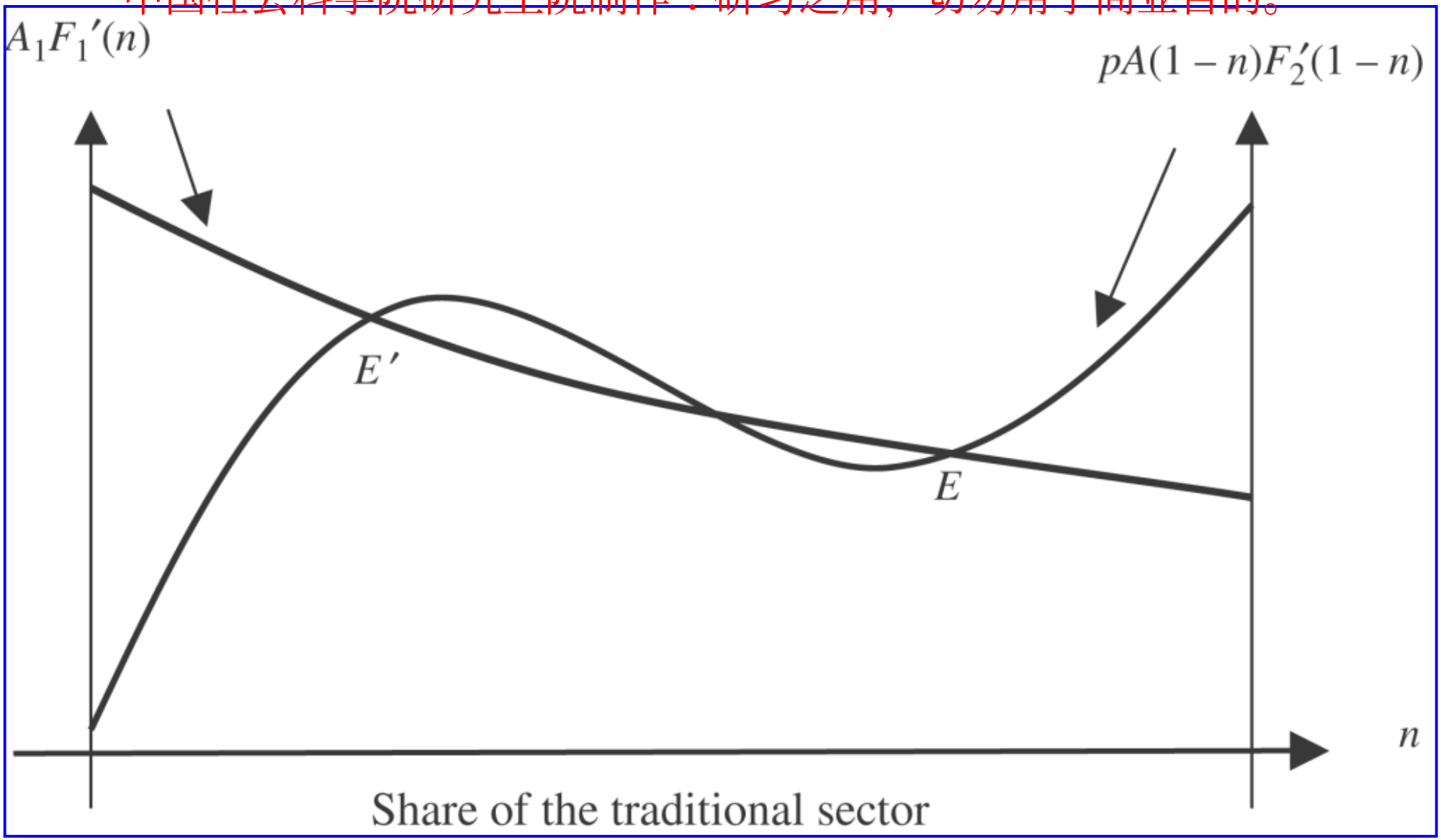

The notion of underdevelopment as a low-level equilibrium generates many conceptual and methodological issues which are poorly understood. These issues are discussed at length in Matsuyama (1991; 1995; 1997). There are also significant misunderstandings regarding its policy implications; see Matsuyama (1996a).

Other aspects of structural change

The transformation from rural agricultural society to urban industrial society is just one of many aspects of structural change discussed in the literature. Due to space constraints, we mention just two.

From old to new industries

The compositions of output and employment change also within the manufacturing sector. Economic growth requires a continuous shift from one industry to another, as existing

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

industries become older and new industries are born. Several authors have studied how the economy may succeed or fail in such structural change. The key issue is whether productivity gains in one industry help or hinder growth in the next industry. See, for example, Stokey (1988), Lucas (1993), Matsuyama (2002), and Foellmi and Zweimüller (2002).

Increasing product diversity and specialization

Productivity growth is often associated with a greater indirectness of production, as many advanced technologies require a wide variety of highly specialized inputs and services. In poor countries, the lack of local support industries forces the use of relatively simple production methods in downstream industries, which in turn implies a small market size for specialized inputs, which prevents a network of support industries from springing up in the economy. In contrast, rich countries are characterized by a network of highly specialized firms producing a wide range of products. This aspect of structural change, discussed by Young (1928), has been formalized by Romer (1987), Ciccone and Matsuyama (1996), Fafchamps and Helms (1996), and Rodriguez-Clare (1996), among others. For surveys, see Matsuyama (1995; 1997). See also Saint-Paul (1992) and Acemoglu and Zilibotti (1997), who use similar arguments to model the interaction between the development of financial markets and economic growth, thereby capturing some of the issues discussed by Gurley and Shaw (1955).

Concluding comments

Most existing studies of structural change, whether descriptive or theoretical, examine the experience of a country in isolation, and fail to take into account the interactions between countries. This can be misleading. (Recall that, in the above model, productivity gains in agriculture can have opposite implications on sector compositions in the closed and small open-economy cases.) Of course, some notable exceptions exist; for example, Brezis, Krugman, and Tsiddon (1993), Krugman and Venables (1995), Matsuyama (1992a; 1996b; 1998), Puga and Venables (1996), and Fafchamps (1997). However, more research will be needed in this area. The central question is whether structural change in one country will slow down or speed up structural change in other countries.

See Also

- poverty traps

Bibliography

Acemoglu, D. and Zilibotti, F. 1997. Was Prometheus unbounded by chance?: Risk, diversification and growth. *Journal of Political Economy* 105, 709–51.

Banerjee, A. and Newman, A. 1998. Information, the dual economy, and development. *Review of Economic Studies* 65, 631–53.

Brezis, E., Krugman, P. and Tsiddon, D. 1993. Leapfrogging in international competition: a theory of cycles in national technology leadership. *American Economic Review* 83, 1211–19.

Caselli, F. and Coleman II, W. 2001. The US structural transformation and regional convergence. *Journal of Political Economy* 109, 584–616.

Chenery, H. and Syrquin, M. 1975. *Patterns of Development, 1957–1970*. London: Oxford University Press.

Ciccone, A. and Matsuyama, K. 1996. Start-up costs and pecuniary externalities in economic development. *Journal of Development Economics* 49, 33–57.

Clark, C. 1940. *The Conditions of Economic Progress*. London: Macmillan.

Fafchamps, M. 1997. Mobile capital, local externalities, and industrialization. *Journal of Comparative Economics* 25, 345–65.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fafchamps, M. and Helms, B. 1996. Local demand, investment multipliers, and industrialization. *Journal of Development Economics* 49, 61–92.

Fisher, A. 1939. Primary, secondary and tertiary production. *Economic Record* 15, 24–38.

Foellmi, R. and Zweimüller, J. 2002. Structural change and Kaldor facts of economic growth. Working Paper No. 111. Institute for Empirical Research in Economics, University of Zurich.

Gollin, D., Parente, S. and Rogerson, R. 2002. The role of agriculture in development. *American Economic Review* 92, 160–4.

Gurley, J. and Shaw, E. 1995. Financial aspects of economic development. *American Economic Review* 45, 515–38.

Hansen, G. and Prescott, E. 2002. Malthus to Solow. *American Economic Review* 92, 1205–17.

Harris, J. and Todaro, M. 1970. Migration, unemployment, and development: a two-sector analysis. *American Economic Review* 60, 126–42.

Krugman, P. and Venables, A. 1995. Globalization and inequality of nations. *Quarterly Journal of Economics* 110, 857–80.

Kuznets, S. 1966. *Modern Economic Growth*. New Haven, CT: Yale University Press.

Laitner, J. 2000. Structural change and economic growth. *Review of Economic Studies* 67, 541–61.

Lewis, A. 1954. Economic development with unlimited supplies of labour. *Manchester School* 22, 139–91.

Lucas, R. Jr. 1993. Making a miracle. *Econometrica* 61, 251–72.

Lucas, R. Jr. 2004. Life earnings and rural-urban migrations. *Journal of Political Economy* 112, S29–S59.

Matsuyama, K. 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. *Quarterly Journal of Economics* 106, 617–50.

Matsuyama, K. 1992a. Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory* 58, 317–34.

Matsuyama, K. 1992b. A simple model of sectoral adjustments. *Review of Economic Studies* 59, 375–88.

Matsuyama, K. 1995. Complementarities and cumulative processes in models of monopolistic competition. *Journal of Economic Literature* 33, 701–29.

Matsuyama, K. 1996a. Economic development as coordination problems. In *The Role of Government in East Asian Development, Comparative Institutional Analysis*, ed. M. Aoki, H. Kim and M. Okuno-Fujiwara. Oxford: Oxford University Press.

Matsuyama, K. 1996b. Why are there rich and poor countries? Symmetry-breaking in the world economy. *Journal of the Japanese and International Economies* 10, 419–39.

Matsuyama, K. 1997. The 1996 Nakahara Lecture: complementarity, instability and multiplicity. *Japanese Economic Review* 48, 240–66.

Matsuyama, K. 1998. Productivity-based theory of manufacturing decline: an international perspective. Mimeo. Northwestern University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Matsuyama, K. 2002. The rise of mass consumption societies. *Journal of Political Economy* 110, 1035–70.

Murphy, K., Shleifer, A. and Vishny, R. 1989. Income distribution, market size, and industrialization. *Quarterly Journal of Economics* 104, 537–64.

Puga, D. and Venables, A. 1996. The spread of industry; spatial agglomeration and economic development. *Journal of the Japanese and International Economies* 10, 440–64.

Rodriguez-Clare, A. 1996. The division of labor and economic development. *Journal of Development Economics* 49, 3–32.

Romer, P. 1987. Growth based on increasing returns due to specialization. *American Economic Review* 77, 56–62.

Rostow, W. 1960. *The Stages of Economic Growth: A Non-Communist Manifesto*. London: Cambridge University Press.

Saint-Paul, G. 1992. Technology choice, financial markets and economic development. *European Economic Review* 36, 763–81.

Stokey, N. 1988. Learning-by-doing and the introduction of new goods. *Journal of Political Economy* 96, 701–17.

Watkins, M. 1963. A staple theory of economic growth. *Canadian Journal of Economics and Political Science* 29, 141–58.

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

How to cite this article

Matsuyama, Kiminori. "structural change." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000440> doi:10.1057/9780230226203.1639

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

structural unemployment

C. Freeman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

active labour market policies; aggregate demand; cyclical unemployment; frictional unemployment; Kondratieff cycles; Okun curve; principle of substitution; segmented labour markets; structural unemployment; technical change; technological unemployment; unemployment; uneven development; wage flexibility; wage rigidity

Article

Economists have generally distinguished between various types of unemployment, in terms of both their characteristic features and their underlying causal mechanisms. Some of these distinctions are non-controversial and relatively straightforward both in their definition and in their measurement. Thus, for example, there is not too much disagreement about the scale and nature of ‘seasonal unemployment’ or about the need for ‘seasonal adjustments’ to the regular monthly series of unemployment statistics published in all industrial countries.

However, when it comes to such categories as ‘structural’ unemployment and ‘cyclical’ unemployment, and even more in the case of ‘voluntary’ unemployment, ‘natural’ unemployment or ‘technological’ unemployment, there is intense disagreement not only with respect to their importance but even with respect to their very existence. In the case of structural unemployment, there is agreement that such a phenomenon does indeed exist, but there is a wide area of disagreement about its extent and even more about its causes.

The distinction between ‘frictional’ and ‘structural’ unemployment has never been a precise and unambiguous one. Nevertheless, there is general agreement that whereas *frictional* unemployment is a transitory and very short-lived form of unemployment based on minor imperfections in the labour market, *structural* unemployment is a more intractable and persistent phenomenon. Some frictional unemployment is generally regarded as an inevitable accompaniment of a dynamic economy, since there will probably always be imperfections in the available information about job opportunities, in the speed of response and in mobility. However, such frictional unemployment can be reduced to a very low level (that is, less than half of one per cent) of the aggregate labour force, as the experience of many countries

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
showed in the high boom period after the Second World War.

Structural unemployment could in principle be caused by a variety of forms of ‘mismatch’ in the labour market with more persistent characteristics. The ‘mismatch’ might, for example, be caused by *regional* disparities in the availability of new employment opportunities, and in the decline of older industries. The persistence of high unemployment levels in southern Italy over prolonged periods, when there were labour shortages in northern and in central Italy, is one instance of this type of structural disequilibrium in regional development, which exists to a greater or lesser degree in most industrialized countries.

Another form of ‘mismatch’ relates to the *skill* profile of the labour force. Whereas the concept of ‘frictional’ unemployment presupposes a rapid process of retraining or learning-by-doing (or none at all), in practice the skill requirements for new job opportunities may differ very substantially from the skills of those in search of work. The degree of mismatch may be so great and so deeply rooted in educational and cultural aspects of society, that it is necessary to analyse ‘segmented’ labour markets, each with its own specific characteristics. If mobility between ‘segments’ is very low, then it is possible to explain the persistence of, for example, very high levels of unemployment among young, black, unskilled workers in the big cities of the United States, side by side with labour shortages in other segments of the market associated with high professional and skill requirements, and lower levels of general unemployment (Doeringer and Piore, 1971; Edwards, Reich and Gordon, 1975).

Whereas almost all economists would agree that such problems as change in skill requirements and the uneven development of different regions of a country may give rise to problems of structural adjustments, they have differed greatly in their views about the severity of such problems, about their underlying causes and about the appropriate policy prescriptions. They have differed in particular about the feasibility of substitution between labour and capital and the time lags involved.

Ricardo provoked an intense and continuing debate amongst economists with his famous remark that ‘The opinion entertained by the labouring class that the employment of machinery is frequently detrimental to their interests, is not founded on prejudice or error, but is conformable to the correct principles of political economy’ (Ricardo, 1821, p. 387).

Although controversy still continues about the interpretation of his remarks and he himself revised some of his earlier formulations because they were misunderstood, it is nevertheless clear that Ricardo was drawing attention to the fact that rapid technical change in the form of mechanization of existing processes of production could in principle give rise to serious unemployment problems for the labour force, at least in specific industries and regions. Although he certainly acknowledged that what would now be called ‘compensation mechanisms’ could ultimately lead to the growth of new employment, both in the machine-building industries and elsewhere in the system, he was pointing out that there could be substantial time lags in the adjustment of the capital stock and the mobility of the labour force. In today's terminology he was emphasizing that rapid process innovation could lead to structural unemployment.

This conclusion has never been a comfortable one for neoclassical economics, which has generally sought to minimize the severity and complexity of these problems, and to put the emphasis almost exclusively on wage flexibility. One of the reasons for this was the neoclassical ‘principle of substitution’ between capital and labour which, provided wages and interest rates are flexible, should in theory assure that there cannot be more than short-term disequilibrium in the labour market. Since entrepreneurs are assumed to be free to make a rational choice between a wide spectrum of alternative combinations of labour and capital, they will substitute labour for capital in the event of a surplus of labour and an appropriate fall in wage rates and vice versa in the case of a shortage of labour. This

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
means that a self-regulating mechanism will always tend to clear the labour market.

Indeed, in the period before the First World War, when neoclassical theories were first established, it was often assumed that structural unemployment could not and would not be a serious problem, and that the problem of ‘technological unemployment’ simply did not exist (Gourvich, 1940). The remaining problems of imperfect information, imperfect mobility and retraining could be handled through such institutional innovations as ‘labour exchanges’(Beveridge, 1909).

The relative neglect of the issue of unemployment by most mainstream neoclassical theory led Keynes (1936) to complain when he came to write his *General Theory of Employment, Interest and Money* that he could not find any adequate statement of a theory of employment in the classical tradition. The somewhat myopic way in which Say's Law was invoked to rule out the possibility of persistent high levels of unemployment was one reason for the severity of Keynes's onslaught on the classical tradition. The rather complacent mainstream view was shattered by the deep depression of the 1930s, and even before that by high levels of unemployment in many countries already in the 1920s. Although traditional theorists continued to emphasize the issue of wage flexibility, most professional economists in the 1930s and 1940s ultimately followed Keynes in recognizing that there were other fundamental problems in maintaining a high level of employment. The Keynesian school can probably best be distinguished from the neoclassical school by its rejection of the notion that equilibrium necessarily implies full employment. Keynes (1936) denied both to wages and to interest rates the self-regulating equilibrating functions which neoclassical theory had assumed. Thus, within a Keynesian framework of analysis there is far more scope for positive structural adjustments policies designed to deal with the problems of regional disequilibrium, skill mismatch and technical change. The problem of structural unemployment is by no means assumed out of existence, but is tackled within an overall framework, emphasizing the cardinal importance of aggregate demand. Beveridge (1944) suggested that, given a strong commitment to ‘full employment’ policies on the part of central government, structural unemployment could be reduced to a relatively low figure, perhaps one per cent of the labour force.

An alternative critique of the neoclassical general equilibrium theory of employment came from Schumpeter (1939) and from other economists, who, for want of a better description, might be designated as ‘structuralists’. In his theory of economic development, Schumpeter stressed in particular the role of major technical innovations as a disequilibriating phenomenon. Such revolutionary new technologies could give rise to ‘creative gales of destruction’ in which old industries, technologies, crafts and employment were decimated by the rise of investment in new products and processes and the opening up of new market.

Schumpeter (1939) regarded the process of technical change as inherently uneven and disequilibrating, giving rise to cyclical behaviour in the system:

> Economists have a habit of distinguishing between, and contrasting cyclical and technological unemployment. But it follows from our model that, basically, cyclical unemployment is technological unemployment … We have seen, in fact, in our historical survey, that periods of prolonged supernormal unemployment, coincide with the periods in which the results of inventions are spreading over the system … (vol. 2, p. 515)

Schumpeter (1952) criticized the Keynesian model for its neglect of technical change and for its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
concentration on the short-term business cycle:

> … it limits applicability of the analysis to a few years at most – perhaps the duration of the ‘40 months’ cycle’ – and in terms of phenomena to the factors that would govern the greater or the smaller utilisation of an industrial apparatus *if* the latter remains unchanged. All the phenomena incident to the creation and change in the apparatus, that is to say the phenomena that dominate the capitalist process are thus excluded from consideration. (p. 480)

In Schumpeter's view the most important problems were associated with long-term processes or Kondratiev cycles. Thus, in his analysis, periods of ‘supernormal unemployment’ occurred approximately every half century or so, that is, in the 1820s, in the 1880s and in the 1930s, when problems of structural adjustment to technical change were particularly severe. Lederer (1931) had already pointed out that severe structural unemployment could arise from the problems of adjustment in the capital stock. The shift of capital investment from ‘static’ branches to the innovative sectors would be hampered by the inertia and rigidity of the existing pattern of investment; ‘capital shortage’ unemployment could be a serious problem side by side with surplus capacity in declining branches of the economy. Thus capital mismatch may give rise to structural unemployment as well as skill mismatch, since the ‘principle of substitution’ does not in fact operate in the simple and instantaneous manner postulated in neoclassical models.

In the prolonged boom after the Second World War Keynesian ideas predominated both in academic economics and in the policy advice offered to governments. Unemployment fell to historically low levels in most industrial OECD countries; female participation rates rose to much higher levels, and in many countries there was substantial net immigration. In these circumstances there was again some tendency to assume that the problems of structural unemployment had been largely resolved, this time through a combination of aggregate demand management policies and active labour market and regional policies. In such countries as Sweden, the German Federal Republic and Austria in particular, active labour market policies were followed which laid great stress on the training and retraining of the labour force to cope with changing skill requirements, and these did indeed help to minimize structural unemployment problems. In the 1970s and 1980s, however, there came renewed recognition that the problem had still not completely disappeared and with general unemployment rates in the OECD area often two or three times as high as in the 1960s, and a rising proportion of long-term unemployed within the total, structural problems of adjustment once more moved to the centre of the stage.

The increasing seriousness of structural mismatch unemployment within the OECD area was shown by ‘Okun curve’ analysis in papers presented to the OECD conference on employment and structural change (Soete and Freeman, 1985) and in the OECD's own publications (OECD, 1985). This evidence showed that the problem was no longer just a cyclical one since the level of unemployment associated with any particular degree of capacity utilization had tended upwards in the 1970s and 1980s, most notably in Europe, but also in the USA and even in Japan. The Secretary-General of the OECD, M. Paye, used the results of this analysis to emphasize the magnitude of the problems of structural adjustment confronting all the OECD countries, and to point to major problems of ‘mismatch’ both in relation to the skill composition of the work force and in relation to the capital stock.

This marked some degree of consensus within the economics profession that the wave of new

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
technology associated with computerization and microelectronics did indeed raise important problems of structural adjustment. The mismatch of skill requirements was very widely recognized and almost all countries initiated special training and retraining programmes to cope with this problem. It was also increasingly recognized that mismatch in the capital stock could give rise to problems of capital shortage unemployment, particularly in countries which had experienced structural rigidity in adapting to new technology. The severe international disequilibria arising from differential rates of technical change once again aroused anxiety over those problems of international structural adjustment which had so much concerned both Ricardo and Keynes.

However, opinions continue to diverge about the relative significance of wage flexibility, incomes policies, interest rates, national and international monetary and demand policies, industry and technology policies in overcoming the structural unemployment problems. Both neoclassical economists and Keynesians continued to insist on the crucial importance on the one hand of wage rigidity and on the other hand of aggregate demand in explaining the persistence of high unemployment and indicating the appropriate remedies (Layard and Nickell, 1985).

See Also

- unemployment

Bibliography

Beveridge, W.H. 1909. *Unemployment: a Problem of Industry*. London: Longmans, Green.

Beveridge, W.H. 1944. *Full Employment in a Free Society*. London: George Allen & Unwin.

Doeringer, P.B. and Piore, M. 1971. *Internal Labor Markets and Manpower Analysis*. Lexington, MA: D.C. Heath.

Edwards, R.C., Reich, M. and Gordon, D.M., eds. 1975. *Labor Market Segmentation*. Lexington, MA: D. C. Heath.

Freeman, C. and Soete, L.L.G. 1987. *Technical Change and Full Employment*. Oxford: Blackwell.

Gourvich, A. 1940. *Survey of Economic Theory on Technological Change and Employment*. Reprinted, New York: A.M. Kelley, 1966.

Keynes, J.M. 1936. *General Theory of Employment, Interest and Money*. London: Macmillan; New York: Harcourt, Brace.

Layard, R. and Nickell, S. 1985. The causes of British unemployment. *National Institute Economic Review* 111, 62–85.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lederer, E. 1931. *Technischer Fortschritt und Arbeitslosigkeit*. Tubingen. Trans. as *Technical Progress and Unemployment*, Geneva: ILO, 1938.

OECD. 1985. *Economic Outlook* 37, June. Paris: OECD.

Ricardo, D. 1821. *Principles of Political Economy and Taxation*. 3rd edn, In *The Works and Correspondence of David Ricardo*, ed. P. Sraffa, Vol. 1, Cambridge: Cambridge University Press, 1951.

Schumpeter, J.A. 1939. *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process*, 2 vols. New York: McGraw-Hill.

Schumpeter, J.A. 1952. *Ten Great Economists*. London: Allen & Unwin.

Soete, L.L.G. and Freeman, C. 1985. New technologies, investment and employment growth. In OECD, *Employment Growth and Structural Change*. Paris: OECD.

How to cite this article

Freeman, C. "structural unemployment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000311>
doi:10.1057/9780230226203.1641

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

structural vector autoregressions

Jesús Fernández-Villaverde and Juan F. Rubio-Ramírez
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Structural vector autoregressions (SVARs) are a multivariate, linear representation of a vector of observables on its own lags. SVARs are used by economists to recover economic shocks from observables by imposing a minimum of assumptions compatible with a large class of models. This article reviews the relation of SVARs to dynamic stochastic general equilibrium models, discusses the normalization, identification, and estimation of SVARs, and concludes with an assessment of the advantages and drawbacks of SVARs.

Keywords

bootstrap; dynamic stochastic general equilibrium models; estimation; identification; Markov chain Monte Carlo methods; neoclassical growth theory; normalization; reduced-form representation; Sims, C. A.; structural vector autoregressions; vector autoregressions

Article

Structural vector autoregressions (SVARs) are a multivariate, linear representation of a vector of observables on its own lags and (possibly) other variables as a trend or a constant. SVARs make explicit identifying assumptions to isolate estimates of policy and/or private agents' behaviour and its effects on the economy while keeping the model free of the many additional restrictive assumptions needed to give every parameter a behavioural interpretation. Introduced by Sims (1980), SVARs have been used to document the effects of money on output (Sims and Zha, 2006a), the relative importance of supply and demand shocks on business cycles (Blanchard and Quah, 1989), the effects of fiscal policy (Blanchard and Perotti, 2002), or the relation between technology shocks and worked hours (Galí, 1999), among many other applications.

Economic theory and the SVAR representation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dynamic economic models can be viewed as restrictions on stochastic processes. From this perspective, an economic theory is a mapping between a vector of k economic shocks w_t and a vector of n observables y_t of the form $y_t=D(w^t)$, where w^t represents the whole history of shocks w_t up to period t. The economic shocks are those shocks to the fundamental elements of the theory: preferences, technology, informational sets, government policy, measurement errors, and so on. The observables are all variables that the researcher has access to. Often, y_t includes a constant to capture the mean of the process. The mapping $D(\cdot)$ is the product of the equilibrium behaviour of the agents in the model, implied by their optimal decision rules and consistency conditions like resource constraints and market clearing. The construction of the mapping $D(\cdot)$ is the sense in which economic theory tightly relates shocks and observables. Also, the mapping $D(\cdot)$ can be interpreted as the impulse response of the model to an economic shock.

Often, we restrict our attention to linear mappings of the form $y_t=D(L)w_t$, where L is the lag operator. For simplicity of exposition, w_t will be i.i.d. random variables and normally distributed, $w_t \sim N(0, \Sigma)$. More involved structures – for example, allowing for autocorrelation between the shocks – can be accommodated with additional notation.

We pick the neoclassical growth model, the workhorse of dynamic macroeconomics, to illustrate the previous paragraphs. In its basic version, the model maps productivity shocks, the w_t of the theory, into observables, y_t, like output or investment. The mapping comes from the optimal investment and labour supply decisions of the households, the resource constraint of the economy, and the law of motion for productivity. If the productivity shocks are normally distributed and we solve the model by linearizing its equilibrium conditions, we obtain a mapping of the form $y_t=D(L)w_t$ described above.

If $k=n$, that is, we have as many economic shocks as observables, and $|D(L)|$ has all its roots outside the unit circle, we can invert the mapping $D(L)$ (see Fernández-Villaverde, Rubio-Ramírez and Sargent, 2005) and obtain

$$A(L)\, y_t = w_t$$

where $A(L) = A_0 - \Sigma_{k=1}^{\infty} A_k L^k$ is a one-sided matrix lag polynomial that embodies all the (usually nonlinear) cross-equation restrictions derived by the equilibrium solution of the model. In general, $A(L)$ is of infinite order. This representation is known as the SVAR representation. The name comes from realizing that $A(L)y_t=w_t$ is a vector autoregression (VAR) generated by an economic model (a ‘structure’).

Reduced form representation, normalization, and identification

Consider now the case where a researcher does not have access to the SVAR representation. Instead, she

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

has access to the VAR representation of y_t:

$$y_t = B_1 y_{t-1} + B_2 y_{t-2} + \cdots + a_t,$$

where $Ey_{t-j}a_t=0$ for all j and $Ea_t a_t' = \Omega$. This representation is known as the reduced-form representation. Can the researcher recover the SVAR representation using the reduced-form representation? Fernández-Villaverde, Rubio-Ramírez, and Sargent (2005) show that, given a strictly invertible economic model, that is, $|D(L)|$ has all its roots strictly outside the unit circle, there is one and only one identification scheme to recover the SVAR from the reduced form. In addition, they show that the mapping between a_t and w_t is $a_t = A_0^{-1} w_t$. Hence, if the researcher knew A_0^{-1}, she could recover the SVAR representation from the reduced form, noting that $A_j=A_0B_j$ for all j and $w_t=A_0a_t$.

Hence, the recovery of w_t from y^T requires the knowledge of the dynamic economic model. Can we avoid this step? Unfortunately, the answer is, in general, 'no', because knowledge of the reduced-form matrices B_i and Ω does not imply, by itself, knowledge of the A_i and Σ, for two reasons.

The first is normalization. Reversing the signs of two rows or columns of the A_i does not matter for the B_i. Thus, without the correct normalization restrictions, statistical inference about the A_i's is essentially meaningless. Waggoner and Zha (2003) provide a general normalization rule that maintains coherent economic interpretations.

The second is identification. If we knew A_0, each equation $B_i = A_0^{-1} A_i$ would determine A_i given some B_i. But the only restrictions that the reduced-form representation imposes on the matrix A_0 comes from $\Omega = A_0 \Sigma A_0'$. In this relationship, we have $n(3n+1)/2$ unknowns (the n^2 distinct elements of A_0 and the $n(n+1)/2$ distinct elements of Σ) for n^2 knowns (the $n(n+1)/2$ distinct elements of Ω). Thus, we require n^2 identification restrictions. Since we can set the diagonal elements of A_0 equal to 1 by scaling, we are left with the need of $n(n-1)$ additional identification restrictions (alternatively, we could scale the shocks such that the diagonal of Σ is composed of ones and leave the diagonal of A_0 unrestricted). These identification restrictions are dictated by the economic theory being studied.

The literature, however, has often preferred to impose identification restrictions that are motivated by the desire to be compatible with a large class of models, instead of just one concrete model and its whole set of cross-equations restrictions. The hope is that, thanks to this generality, the inferences drawn from SVAR can be more robust and can compensate for the lack of efficiency derived from not implementing a full information method.

The most common identification restriction has been to assume that Σ is diagonal. This assumption relies on the view that economic shocks are inherently different sources of uncertainty that interact only through their effect on the decisions of the model's agents. Since this assumption imposes $n(n-1)/2$ restrictions, we still require $n(n-1)/2$ additional restrictions.

To find these additional restrictions, economists have followed two main approaches: short-run

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
restrictions and long-run restrictions. Sims (1980) pioneered the first approach when he proposed to impose zeros on A_0. The motivation for such a scheme comes from the idea that there is a natural timing in the effect of economic shocks. For example, to place a zero on A_0, we can use the intuition that monetary policy cannot respond contemporaneously to a shock in the price level because of informational delays. Similarly, institutional constraints, like the timing of tax collections, can be exploited for identification (Blanchard and Perotti, 2002). Sims (1980) ordered variables in such a way that A_0 is lower triangular. Sims and Zha (2006a) present a non-triangular identification scheme on an eight-variable SVAR.

The long-run restrictions were popularized by Blanchard and Quah (1989). These restrictions are imposed on $A(1) = A_0 - \Sigma_{k=1}^{\infty} A_k$. Since $A^{-1}(1)=D(1)$, long-run restrictions are justified as restrictions on the long-run effects of economic shocks, usually on the first difference of an observable. For example, Blanchard and Quah (1989) assume that there are two shocks ('demand' and 'supply') affecting unemployment and output. The demand shock has no long-run effect on unemployment or output. The supply shock has no long-run effect on unemployment but may have a long-run effect on output. These differences in their long-run impacts allow Blanchard and Quah to identify the shocks and trace their impulse response function.

New identification schemes have been proposed to overcome the difficulties of the existing approaches. See, for instance, Uhlig (2005) for an identification scheme of monetary policy shocks based on sign restrictions that hold across a large class of models.

Estimation

Why is the previous discussion of the relation between the reduced and structural form of a VAR relevant? Because the reduced form can be easily estimated. An empirically implementable version of the reduced-form representation truncates the number of lags at the *p*th order:

$$y_t = \bar{B}_1 y_{t-1} + \cdots + \bar{B}_p y_{t-p} + \bar{a}_t$$

where $E\bar{a}_t\bar{a}_t' = \bar{\Omega}$. We use hats in the matrices and the error $\bar{a}_t$ to indicate that they do not correspond exactly to the reduced form of the model but to the truncated version. The effects of the truncation on the accuracy of inference delivered by SVARs are unclear (see Chari, Kehoe and McGrattan, 2005, and Christiano, Eichenbaum and Vigfusson, 2007, for two opposite assessments). The resulting truncated VAR can be taken to the data using standard methods: GMM, maximum likelihood, or Bayesian.

The Bayesian approach is especially attractive. SVARs are proliferatively parametrized. The number of parameters in $B(L)$ grows with the square of the number of variables and the number of lags. Consequently, given the short period of data typically available to macroeconomists, classical methods become unreliable. A careful use of prior information alleviates the problem of overparametrization and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

improves the quality of the inference. The advent of modern simulation techniques, especially Markov chain Monte Carlo methods, has made the implementation of the Bayesian paradigm straightforward, even for sophisticated priors.

The point estimates $\hat{\bar{B}}_i$ for I=1,…,p and $\hat{\bar{\Omega}}$ can be used to find estimates of $\hat{A}_i$ and $\hat{\Sigma}$ by solving $\hat{\bar{B}}_i = \hat{A}_0^{-1}\hat{A}_i$ for I=1,…,p, and $\hat{\bar{\Omega}} = \hat{A}_0\hat{\Sigma}\hat{A}_0'$. With an estimate of $\hat{A}_0$ and the $\bar{a}_t$, we can get $\hat{\bar{w}}_t = \hat{A}_0\hat{\bar{a}}_t$. Thus, the reduced form plus the identifying restrictions deliver both an estimate of the economic shocks and the impulse response of the variables in the economy to those shocks. Confidence intervals for point estimates and error bands for impulse response functions can be estimated by resorting to Markov chain Monte Carlo techniques or the bootstrap.

In an important contribution, Sims and Zha (2006b) have extended the estimation of SVARs to allow for changes in equation coefficients and variances. This article opens the door for the analysis of richer dynamic models with parameter instability, arguably a more realistic description of observed aggregate variables.

Assessment of SVARs

SVARs offer an attractive approach to estimation. They promise to coax interesting patterns from the data that will prevail across a set of incompletely specified dynamic economic models with a minimum of identifying assumptions. Moreover, SVARs can be easily estimated, even with commercial software and freely available routines from the Internet. In the hands of skilful researchers, SVARs have contributed to the understanding of aggregate fluctuations, have clarified the importance of different economic shocks, and have generated fruitful debates among macroeconomists.

However, SVARs have also been criticized. We mention only three criticisms. First, it has been argued that the economic shocks recovered from an SVAR do not resemble the shocks measured by other mechanisms, such as market expectations embodied in future prices. Second, the shocks recovered from an SVAR may reflect variables omitted from the model. If these omitted variables correlate with the included variables, the estimated economic shocks will be biased. Third, the results of many SVAR exercises, even simple ones, are sensitive to the identification restrictions. Related to this criticism is the view that many of the identification schemes are the product of a specification search in which researchers look for ‘reasonable’ answers. If an identification scheme matches the conventional wisdom, it is called successful; if it does not, it is called a puzzle or, even worse, a failure (Uhlig, 2005). Consequently, there is a danger that economists will get stuck in an a priori view of the data under the cloak of formal statistical inference.

See Also

- real business cycles
- time series analysis
- vector autoregressions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Blanchard, O.J. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.

Blanchard, O.J. and Perotti, R. 2002. An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *Quarterly Journal of Economics* 117, 1329–68.

Chari, V.V., Kehoe, P.J. and McGrattan, E.R. 2005. A critique of structural VARs using real business cycle theory. Working Paper No. 631, Federal Reserve Bank of Minneapolis.

Christiano, L.J., Eichenbaum, M. and Vigfusson, R. 2007. Assessing structural VARs. In *NBER Macroeconomics Annual 2006*, vol. 21, ed. D. Acemoglu, K. Rogoff and M. Woodford. Cambridge, MA: MIT.

Fernández-Villaverde, J., Rubio-Ramírez, J.F. and Sargent, T.J. 2005. A,B,C's (and D's) for understanding VARs. Technical Working Paper No. 308. Cambridge, MA: NBER.

Galí, J. 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations? *American Economic Review* 89, 249–71.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Sims, C.A. and Zha, T. 2006a. Does monetary policy generate recessions? *Macroeconomic Dynamics* 10, 231–72.

Sims, C.A. and Zha, T. 2006b. Were there regime switches in U.S. monetary policy? *American Economic Review* 96, 54–81.

Uhlig, H. 2005. What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics* 52, 381–419.

Waggoner, D.F. and Zha, T. 2003. Likelihood preserving normalization in multiple equation models. *Journal of Econometrics* 114, 329–47.

How to cite this article

Fernández-Villaverde, Jesús and Juan F. Rubio-Ramírez. "structural vector autoregressions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
id=pde2008_S000491> doi:10.1057/9780230226203.1642

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

structuralism

Stephanie Blankenburg, José Gabriel Palma and Fiona Tregenna
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Structuralist analysis advocates a focus on a system in its totality and on the interrelations between its elements rather than on individual elements in isolation: for instance, understanding the world economy as a system within which the centre and periphery are intrinsically linked, with many economic problems of the periphery deriving from that interaction. Structuralist development economics, as associated with, for example, Raúl Prebisch and Celso Furtado, essentially turns on two phenomena considered inherent to development in countries of the periphery that export almost exclusively commodities: foreign-exchange constrained growth and the tendency towards deterioration in the terms of trade.

Keywords

agency; business cycles; centre and periphery; class; comparative advantage; complementary capital; cost-push inflation; Economic Commission for Latin America and the Caribbean (ECLAC); economic development; foreign exchange-constrained growth; Furtado, C.; gains from trade; Haberler, G.; Heckscher–Ohlin–Samuelson trade theory; import-substituting industrialization; institutions; international division of labour; periphery; Perroux, F.; Prebisch, R.; sectoral growth; Sombart, W.; specialization; stages of growth; structural change; structuralism; structures; terms of trade

Article

Structuralism is essentially a theoretical approach that challenges the methods of empiricism and positivism. Structuralism features in several disciplines across the humanities and the social sciences, but not as a cohesive school of thought. In so far as there are common elements, the conception of *an integrated system of distinguishable yet mutually constitutive elements* could be said to be the most important feature. These elements derive their meaning in relation to one another – such as, for instance in economics, the understanding of development and underdevelopment as related, mutually constitutive

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
processes within an integrated world economic system. This is as distinct from the analysis of the elements as relatively independent units.

The emergence of structuralism is typically traced to the linguistic work of De Saussure, who analysed language as a system whose elements can be defined only through their relations of equivalence or opposition to one another, with these relations forming the structure. The term ‘structuralism’ was apparently coined by Jakobson, a member of the Prague School of linguistics, in 1929. Structuralism is also strongly associated with the work of Lévi-Strauss in anthropology, notably in his analysis of the structures within cultures through which meanings are produced and reproduced. Other writers often associated with structuralism include Althusser, Barthes, Derrida, Godelier, and Lacan (although not all identified themselves as structuralist nor would be universally regarded as such).

Structuralism can be considered as having *methodological, epistemological, and ontological* dimensions. However, in different disciplines and for different writers variously identified as ‘structuralist’, not all these dimensions necessarily feature.

As a *methodology*, structuralist analysis advocates a focus on a system in its totality and on the interrelations between its elements, rather than on individual elements in isolation. For instance, in terms of economic analysis, this might point towards an emphasis on understanding the world economy as a unified system, with the economic dynamics of its constituent parts – centre and periphery – being defined in terms of their interrelationship. Such a methodological approach relates closely to the seminal focus in structuralist linguistics on synchronic rather than diachronic analysis (that is, an emphasis on a comprehensive analysis of a linguistic structure at a given conjuncture, even at the expense of historical or comparative approaches). Structuralism can be distinguished in this respect from historicist analysis. It would also tend to offer non-narrative explanations (in terms of an emphasis on analysis of the underlying dynamics rather than on descriptive explanations).

In terms of *epistemology*, structuralism requires the penetration of appearance in order to grasp the deep underlying structure. In this sense, structuralism is anti-phenomenological and may also be considered anti-empiricist. From this perspective, structuralist approaches in economics hold that there is a set of economic and social structures that are themselves unobservable, yet generate observable social and economic phenomena. The latter could not be properly understood unless the analysis focuses on the unobservable underlying structures.

In terms of *ontology*, structuralism tends to favour explanations as to how structures cause or at least condition or asymmetrically constitute aspects such as agency. This ontological approach is particularly relevant to structuralist political theory and structuralist Marxism (notably in Althusser), and stands in opposition to humanist and historicist interpretations of Marxism. There are two dimensions here: structure and agency. Notwithstanding important differences between structuralist approaches, a shared characteristic is the ontological primacy accorded to structure over an event or a phenomenon. Agents are regarded as often unaware of the economic structure, or the totality of the social relations of production, so that there is a gap between their discourse and the collective social practice that constitutes the objective economic structure. This is expressed in Althusser's theory of social practice as a process of transformation without a subject: by transforming the social and natural environment through work, people determine the economic structure, but not as subjects, through their agency, but through internalized social organization and practice. Hence, structuralism seeks to explain social phenomena by reference to the underlying structure of the mode of production and the social organization – or practice – that determines it. This approach can be distinguished from humanist

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
approaches that privilege agency, stressing the role of human consciousness and action in social change.
In modern economics, structuralism is mostly associated with the United Nations Economic Commission for Latin America and the Caribbean (ECLAC), whose work merged into a coherent school of thought in the late 1950s. However, at least some of the core elements of ECLAC thought can be traced to earlier Continental European contributions, in particular to the ‘structuralist’ economic school in France. The earliest use of the modern concept of economic structure – after Marx – can be found in the work of Wagemann, who introduced the notion of structure in his business cycle studies. For Wagemann, certain types of business cycles correspond to specific underlying economic structures. A system is determined by all the particularities of a country and its people, or by the totality of the data. In a similar vain, Åckerman developed a ‘structuralist’ view of the economy in relation to the study of business cycles. From a dynamic point of view, Åckerman defines economic structures as those structures that are invariable in the short term.
Wagemann's and Åckerman's work was well known to Perroux, who is probably the best-known representative of French economic structuralism (and who was the main intellectual influence in Furtado's early work, including his doctoral dissertation at the Sorbonne; see Furtado, 1995). Perroux (1939) defined structural economics as the science of the relations characteristic of an economic system (ensemble) situated in time and space. Central to Perroux's approach was the view that, over and above the ‘givens’ of neoclassical theory (preferences, resources and technology), the analysis of institutions and structures over time had to be at the heart of economic analysis. An important innovative contribution by Perroux concerns his theory of domination that is central to ECLAC's conception of economic systems: rather than being constituted by relationships between equal agents, the economic world is conceptualized in terms of hidden or explicit relationships of ‘force’, ‘power’ and ‘constraints’ between dominant and dominated entities. Perroux applies the theory of domination to different levels of analysis, first, to markets (such as the effects of the level of unemployment on relations of dominance in the labour market), next to the theory of firms, especially regarding imperfect competition, and finally to the international economy, in terms of the relationship between ‘dominant’ and ‘dominated’ economies, particularly concerning trade and finance.
French economic ‘structuralism’ clearly anticipates important aspects of ECLAC thought. This is not surprising, given the close personal links between the main protagonists on both sides. In addition to the Perroux–Furtado relationship, Wagemann immigrated to Chile where he worked with leading ECLAC economists and promoted the ideas of Sombart who, apart from Perroux, was also an important influence on Prebisch and Furtado.
ECLAC first explicitly termed its own analysis ‘structuralist’ in the mid-1950s in its controversies with the monetarist view of inflation (see Noyola, 1957; Pinto, 1963; and Sunkel et al., 1963; see also Kaldor, 1957). In brief, the ECLAC argument was that money was endogenous in Latin America's persistent inflation, which was seen as a cost-push phenomenon that mainly originated in the supply rigidities of the agricultural sector (an idea initially advanced by Austin Robinson in his early works on India).
‘Anglo-American’ development economics of the 1950s and early 1960s – associated with the work of Lewis, Rosenstein-Rodan, Nurkse, Chenery and Syrquin, among others – also centred around the analysis of structural change.
From the very beginning ECLAC's analysis was structuralist in the sense that it was associated with both a view of the world economy as a system within which the centre and the periphery are intrinsically related to one another, and that most economic problems of the periphery (such as inflation, stop/go

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
macroeconomics, inequality and unemployment) derive from the specific economic structure that emerged from that interaction (see ECLAC, 1964; Love, 1977).

ECLAC's analysis was also structuralist in the sense that it tried to focus on *underlying* structures and relationships, as opposed to epiphenomena. More specifically, the hub on which the whole of ECLAC's analysis of underdevelopment turned was the idea that the *structure of production* in the centre and in the periphery differed substantially. That of the centre was seen as homogeneous and diversified, that of the periphery as heterogeneous and specialized – heterogeneous because economic activities with remarkably different productivity–growth dynamics existed side by side, with two extremes of a modern export sector and subsistence agriculture; specialized because the production of commodities for exports had very limited backward and forward linkages with the rest of the economy (see ECLAC, 1969).

It was this structural difference between the two types of economy that underpinned the different function of each pole in the international division of labour. These structural differences could not be defined or understood in static terms, as the transformation of either pole would be conditioned by the interaction between them. Centre and periphery formed a single system, dynamic by its very nature. This approach differs from the traditional Marxist view of the time, which still saw the periphery simply as a ‘backward’ region that had not yet succeeded in revolutionizing relations of production and the class structure in ways similar to the transformation achieved by more advanced capitalist countries (see dependency).

The nucleus of ECLAC analysis was the critique of the (‘non-product specific’) Solow-type growth models, and of the (‘static resource endowment specific’) comparative advantages approach found in the Heckscher–Ohlin–Samuelson-type models of international trade (see Palma, 2005). It aimed to show that growth was both a ‘product specific’ phenomenon, and that comparative advantages can also be *acquired*. The key issue here is that, according to ECLAC, in terms of both growth and welfare the international division of labour which conventional trade theory claimed to be ‘naturally’ produced by (static) comparative advantages was of much greater benefit to the centre (where manufacturing production is concentrated) than to the commodity-exporting periphery. The greatest asymmetries could be found in short-term welfare gains from trade, and in the effectiveness of exports as a long-term engine of growth.

From this perspective, ECLAC's structuralism did not question the basic postulate of mainstream economics – that free interaction of rational and selfish agents in the market leads to ‘equilibrium’. What it argued was both that this equilibrium was not optimal for long-term growth and welfare in the periphery, and – following the post-war Keynesian tradition – that there was something that the periphery could do to improve upon this suboptimal equilibrium.

The ECLAC analysis essentially turns on two basic phenomena which are considered inherent to the development of countries in the periphery that export almost exclusively commodities: foreign exchange-constrained growth, and the tendency toward deterioration of the terms of trade. In terms of foreign exchange-constrained growth, as long as domestic production of tradables continues to be concentrated in primary commodities, the periphery is bound to experience external disequilibrium leading to foreign exchange-constrained growth. Successful commodity exporters will tend to have a weak manufacturing sector not just because they can afford deficits in their trade balance in manufacturing, but also due to ‘Dutch Disease’ effects (see Palma, 2005, and de-industrialization, ‘premature’ de-industrialization and the Dutch Disease). Given that income elasticity for imported manufactures in the periphery is much

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
greater than that for commodities in the centre, growth in commodity-exporting countries is likely to be foreign exchange-constrained (see McCombie and Thirlwall, 1994). The need for a long-term trade balance between centre and periphery would imply that, for a given rate of growth of real income in the centre, the disparity between the income elasticities of imports at each pole will impose a limit upon the rate of growth of real income in the periphery. This growth will tend to be less than that of the centre in proportion to the degree of disparity between the respective income elasticities of demand for imports. If the periphery attempts to surpass this limit, subsequent trade deficits would force it to decelerate. Although in the short term the periphery can borrow its way out of trouble, the only long-term alternative to this trend towards growth divergence is for the periphery to try to reduce this disparity in income elasticity for imports. That is, to make a significant effort to satisfy the highly income-elastic demand for manufactures through import substitution, and/or to diversify its export trade towards more income-elastic products. While Latin America tried to overcome its foreign-exchange constrained growth almost exclusively through a process of import-substituting industrialization, East Asia more ambitiously attempted both routes (while also curbing domestic demand for luxury consumer goods). Thus, given ECLAC's assumptions, only a process of industrialization can enable the periphery to enjoy a rate of growth of real income higher than that determined by the growth rate in the centre and by the disparity between income elasticities of demand for imports in both poles. Therefore, for ECLAC growth ends up being ‘product-specific’ in a double sense. From a demand point of view, only a process of industrialization can lift the foreign-exchange constraint on growth. From a supply point of view, only manufacturing is a good engine of growth (in a Kaldorian sense) due to its potential for sustained high productivity growth. This is because ‘learning by doing’, dynamic economies of scale, increasing returns, externalities and spillover effects are more prevalent in manufacturing than elsewhere in the economy.

In terms of the second major problem that ECLAC considered inherent in the development of the commodity-exporting periphery – the tendency toward deterioration of its terms of trade (and the asymmetries which this brings with it in terms of the welfare gains from trade between centre and periphery) – in ECLAC thought this is also a logical analytical deduction from the phenomena of specialization and heterogeneity. That is, contrary to common perception, the terms of trade issue is not the starting point of ECLAC thought, but, given its assumptions and hypotheses, a logical analytical deduction.

There are demand and supply forces behind this tendency to deterioration in the terms of trade of the periphery. The fundamental problem is the specific effect that economic growth has on the terms of trade (see Prebisch, Raúl for a detailed analysis of this point; see also Singer, 1949; and Kindleberger, 1956).

In sum, according to ECLAC, it is possible for the commodity-rich periphery to escape from the negative growth and welfare effects of passive integration into the world economy, through the transformation of its economic structure. The central element in this structural transformation is to accelerate industrialization in the periphery. Thus Prebisch often summarized ECLAC's task as having been that of showing that rapid industrialization was an unavoidable prerequisite for development; in fact, he appears at times to use the concepts ‘industrialization’ and ‘development’ as synonyms (see Prebisch, Raúl).

If industrialization were perceived as a necessary (and sometimes in the work of some structuralist writers apparently even sufficient) condition for a rapid and sustainable rate of economic growth, this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
process could not be expected to occur spontaneously. It would be inhibited by the international division of labour which the centre would attempt to impose, and by various structural obstacles internal to the peripheral economies. Consequently, a series of measures was proposed to promote deliberate, ‘forced’, or ‘state-led’ (import-substituting) industrialization. These included state intervention in the economy, both through trade and industrial policies and as a direct productive agent. Among the economic policies suggested in order to create incentives (rents) to direct resources towards domestic manufacturing and associated activities were those of ‘healthy’ protectionism, exchange-rate management, the attraction of foreign capital, and the stimulation and orientation of domestic investment. State intervention in directly productive activities was recommended in areas where large amounts of slow-maturing investment were needed, and particularly where this need coincided with the production of essential inputs and services for manufacturing, and of ‘complementary capital’ (such as infrastructure and utilities).

The various elements of ECLAC thought are brought together by its internal unity and its structuralist nature. The two most important problems of the development of the economy in the periphery (that is, foreign exchange-constrained growth and the tendency to deterioration in the terms of trade) derive directly from the characteristics of its structure of production and of its interaction with the centre. The possibility of tackling them is thus seen in terms of an ideal pattern of transformation, indicating the conditions of structural proportionality which must hold if these problems are to be avoided. This leads to the formulation, tacitly or explicitly, of the law of proportionality in the transformation. This would avoid heterogeneity and specialization and thus allow the escape from external disequilibria (and stop/go growth), eventually also counteracting the tendency towards deterioration of the terms of trade. It would also assist the periphery in achieving full productive utilization of its labour force, a more equal distribution of income, and lower inflation.

Nevertheless, it is also in this very structuralist nature that the limitations of ECLAC thought lie. At this level of analysis no consideration is given to the social relations of production which developed with import-substituting industrialization, nor to the problems arising from the broader social transformation that inevitably followed. Furthermore (and mainly as a result of its own intellectual ‘structural rigidities’, arising from the fact that ECLAC is a UN organization), structuralist thinkers never properly addressed the issue that it is one thing to use trade and industrial policies to create rents to divert resources towards more ‘dynamic’ activities, but quite another for the state to have the institutional capabilities necessary to ensure that the capitalist elite uses those rents effectively.

ECLAC proposes an ideal model of sectoral growth – and hence of overall growth – designed to avoid the reproduction of the tendencies peculiar to economic development in the commodity-rich periphery. From this model are derived the necessary conditions of domestic accumulation that ensure a balanced (proportional) transformation of the different production sectors. Nevertheless, even when pushed to the limits of its potential internal coherence, the structural approach is inadequate for the analysis of the long-term evolution of the economic system as a whole, as this clearly involves more than the transformation of the structure of production alone (see Hirschman, 1971 and Rodríguez, 2006). ECLAC theories describe and examine certain aspects of the development of the forces of production (to the extent that the theories deal with the productivity of labour and the degree of diversification and homogeneity of the structures of production). However, they do not touch on relations of production, nor, as a result, on the manner in which the two interact. In fact, initial attempts to introduce into the traditional ECLAC analysis various ‘social’ and ‘political’ aspects (for example, Prebisch, 1963), far from strengthening the analysis, revealed its fragility (see Palma, 1978).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Furthermore, the analysis of the asymmetries of development cannot be undertaken solely in terms of the patterns of accumulation necessary to avoid the creation of certain disproportionalities between the different sectors of production, as their feasibility depends more upon the general conditions in which accumulation occurs globally than upon the possibility of accumulation and structural change in the periphery. If the intention is to analyse the asymmetries of the centre–periphery system, it is inadequate just to consider the inequality of the development of the forces of production. It is just as necessary to bear in mind that those forces of production develop in the broader context of the generation, appropriation and utilization of the economic surplus. This process, and the relations of exploitation upon which it is based, are not reproduced purely *within* each pole, but also *between* the two poles of the world economy (see Rodríguez, 2006).

It is not surprising that ECLAC immediately attracted considerable criticism, particularly as it went beyond theoretical pronouncements to offer packages of policy recommendations. It was criticized from sectors of the Left for failing to sufficiently denounce the mechanisms of exploitation within the capitalist system, and for criticizing the conventional theory of international trade only from ‘within’ (see for example Frank, 1967). On the other hand, from the other end of the political spectrum the reaction was immediate: ECLAC's policy recommendations were totally heretical from the perspective of conventional economic theory, and threatened the political interests of significant sectors of both the traditional capitalist elite, and of foreign capital and finance. A leading critic in academic circles was Haberler (1961), who accused ECLAC of failing to take due account of economic cycles, and argued that single factorial terms of trade would be a better indicator than the simple relationship between the prices of exports and imports.

On the political front, the right accused ECLAC of being the ‘Trojan horse of Marxism’, on the grounds of the degree of coincidence between parts of both analyses. In both cases the principal economic obstacle was located externally (international division of labour imposed by the centre), and both shared the conviction that without a strenuous effort to remove the obstacles to development (the traditional sectors that benefited from the status quo, and their external allies) the process of industrialization would be significantly impeded. Furthermore, the coincidence between crucial elements in the analysis of the two respective lines of thought is made more evident by the fact that both schools underwent simultaneous exercises in the revision of established lines of thought (see dependency). Moreover, both reformulations had one extremely important element in common: a growing *pessimism* regarding the possibility of capitalist development in the periphery.

Aspects of the ECLAC analysis re-emerged in the 1980s in some North American academic circles (the most imaginative contribution is that of Taylor, 1983), but rather than revitalizing structuralism as a new method of enquiry into economic analysis these only succeeded in integrating some of the key assumptions and hypotheses of the traditional ECLAC analysis into mainstream economic thinking. In the South there was also a (relatively short-lived) attempt to develop a ‘neo-structuralist’ school by reformulating classical ECLAC thought using modern economic analysis and techniques (see Fajnzylber, 1983; and Sunkel, 1991; see also Rodríguez, 2006). The main aim was to build a coherent set of economic ideas as an alternative to the then rapidly emerging neoliberalism. Very few ‘neo-structuralists’ stayed the course for long (except mainly for José Antonio Ocampo and Lance Taylor).

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- de-industrialization, ‘premature’ de-industrialization and the Dutch Disease
- dependency
- Díaz-Alejandro, Carlos
- Furtado, Celso
- Prebisch, Raúl
- terms of trade

Bibliography

ECLAC (United Nations Economic Commission for Latin America and the Caribbean). 1964. *El desarrollo económico de América Latina en la post-guerra*. Santiago: UN (CEPAL).

ECLAC. 1969. *El pensamiento de la CEPAL*. Santiago: Editorial Universitaria.

Fajnzylber, F. 1983. *La industrialización trunca de América Latina*. México, DF: Nueva Imagen.

Frank, A.G. 1967. *Capitalism and Underdevelopment in Latin America*. New York: Monthly Review Press.

Furtado, C. 1995. *Retour à la vision globale de Perroux et Prebisch*. Grenoble: Presse Universitaires Grenoble.

Haberler, G. 1961. Terms of trade and economic development. In *Economic Development of Latin America*, ed. H.S. Ellis. New York: St. Martin's Press.

Hirschman, A. 1971. *A Bias for Hope*. New Haven: Yale University Press.

Kaldor, N. 1957. La inflación chilena y la estructura de la producción. *Panorama Económico* 180, 738–52.

Kindleberger, C.P. 1956. *The Terms of Trade: A European Case Study*. New York: MIT Press.

Love, J.L. 1977. *Center–Periphery and Unequal Exchange: Origins and Growth of an Economic Doctrine*. Houston: LASA.

McCombie, J.S. and Thirlwall, P. 1994. *Economic Growth and the Balance of Payments Constraints*. London: St. Martin Press.

Noyola, J. 1957. Inflación y desarrollo económico en Chile y México. *Panorama Económico* 170, 391–99.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Palma, J.G. 1978. Dependency: a formal theory of underdevelopment, or a methodology for the analysis of concrete situations of underdevelopment? *World Development* 6, 881–924.

Palma, J.G. 2005. Four sources of de-industrialisation and a new concept of the 'Dutch Disease'. In *Beyond Reforms: Structural Dynamic and Macroeconomic Vulnerability*, ed. J.A. Ocampo. Stanford: Stanford University Press.

Perroux, F. 1939. Pour un approfondissement de la notion de structure. In *Mélanges économiques et sociaux offerts à Emile Witmeur*. Paris: Librairie du Recueuil Sirey.

Pinto, A. 1963. El análisis de la inflación: 'estructuralistas' y 'monetaristas'; un recuento. *Economía* 79, 15–30.

Prebisch, R. 1963. *Hacia una dinámica del desarrollo económico*. Mexico: Fondo de Cultura Económica.

Rodríguez, O. 2006. *El estructuralismo latinoamericano*. Mexico: Siglo XXI.

Singer, H. 1949. *Postwar Price Relations Between Underdeveloped and Industrialized Countries*. New York: United Nations.

Sunkel, O. 1963. La inflación chilena: un enfoque heterodoxo. In Sunkel et al. (1963).

Sunkel, O. 1991. *El desarrollo desde dentro. Un enfoque neoestructuralista para la América Latina*. Mexico, DF: Fondo de Cultura Económica.

Sunkel, O., Maynard, G., Seers, D. and Olivera, J. 1963. *Inflación y estructura económica*. Buenos Aires: Paidós.

Taylor, L. 1983. *Structuralist Macroeconomics*. New York: Basic Books.

How to cite this article

Blankenburg, Stephanie, José Gabriel Palma and Fiona Tregenna. "structuralism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000312> doi:10.1057/9780230226203.1643

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

stylized facts

Lawrence A. Boland
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

assumptions; capital accumulation; explanation; income distribution; Kaldor, N.; Keynesian–classical model; models; neoclassical theory; stylized facts

Article

The concept of ‘stylized facts’ is usually attributed to Nicholas Kaldor, who discussed this concept in a well-known 1958 Corfu conference paper (1961) on capital accumulation and economic growth. While the term ‘stylized facts’ is widely used today in many varied contexts, Kaldor had a specific use in mind. Kaldor (1961, p. 177) noted that everyone agrees ‘that the basic requirement of any model is that it should be capable of explaining the characteristic features of the economic process as we find them in reality’. But how are we to explain results of a theoretical model which are contrary to what we observe in reality? Too often, Kaldor complains, the contrary results are explained away by simply noting that the assumptions of the model did not account for changes in such things as knowledge or merely assumed away uncertainty and technological progress. For Kaldor, such a method of explaining away discrepancies between the results of the theoretical model and the facts of the world we see outside of our window is of very little ‘interpretive value’.

The problem that Kaldor is concerned with is not the simplistic view that recognizes that assumptions of models and theories ‘must necessarily be based on abstractions’. Rather, it is the more difficult one of being careful to choose a type of abstraction that is ‘appropriate to the characteristic features of the economic process as recorded by experience’. In Kaldor's view, when choosing between competing theoretical approaches the proponents of the two competing approaches ‘ought to start off with a summary of the facts’ which both regard as relevant to the task at hand.

Since Kaldor wished to focus our attention on the difficult problem of choosing an appropriate type of abstraction for the economic world we hope to explain, he wanted to avoid unproductive debate over details of historical accuracy. He said that we should be free to start off with a ‘stylized’ view of the facts to be explained. Specifically, we should be free to concentrate on broad tendencies, ignoring individual detail, and proceed with what he calls the ‘as if’ method. His use of an ‘as if’ method is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
different from the common neoclassical use of an ‘as if’ method. Unlike neoclassical economists who employ simplistic *assumptions* ‘as if’ they were true, Kaldor would have us explain ‘stylized’ facts as if they truly represented the reality we want to explain. As long as proponents of competing explanations can come to an agreement regarding the ‘stylized’ facts that both wish to explain, the comparative appropriateness of competing explanatory abstractions can be brought into clear and decisive focus.

In Kaldor's day, the issue was the construction of a theoretical model of economic growth and capital accumulation. The competitor to his Keynesian–classical model was anyone promoting a neoclassical model. To focus the debate, Kaldor suggested six ‘stylized’ facts as a starting point. These included such things as a steady rate of growth of production and productivity of labour, a growing capital–labour ratio, a steady rate of profit on capital, steady capital–output ratios over long periods, high correlation between the share of profits in income and the share of investment in output, with allowance for the fact that there are differences between societies in the rates of growth of labour productivity and of total output. With this in mind, Kaldor went on to claim that none of his suggested ‘stylized facts’ can be plausibly ‘explained’ by the assumptions of neoclassical models, while the alternative model of income distribution and capital accumulation which he presents is capable of explaining some if not all of his ‘stylized facts’.

Economists today who claim to explain ‘stylized facts’ usually are not fully aware that Kaldor used the term only in the context of a theoretical comparison between two competing approaches. For a while in the 1970s and 1980s most neoclassical model-builders claimed to be explaining stylized facts merely as a convenient simplification of the model-building process. There was rarely a mention of a competing approach. If these model-builders thought they were following Kaldor's lead, they were clearly mistaken. Today, stylized facts are used mainly in macroeconomics (for example, D. Romer, 1996, p. 15) and is consistent with the primary methodological reason Kaldor gave for using stylized facts: ‘facts, as recorded by statisticians, are always subject to snags and qualifications, and for that reason are incapable of being accurately summarized.’ The difference is that today, when one claims to be using stylized facts, one is claiming that there exists a ‘scientific consensus’ that those facts are an acceptable characterization of the data one is trying to explain. So, the issue is never about a debate between competing models or approaches but merely a laying out of a commonly accepted task for the model-builder. A successful model is expected to at least explain those ‘stylized facts’.

Interestingly, when the theoretical domain at issue is an economy's economic growth, Kaldor's stylized facts are still part of the picture that any growth model would have to explain (see for example P. Romer, 1989, pp. 53–4). Even if we were to leave aside Kaldor's methodological purpose for using ‘stylized facts’ and instead see them as an indication of the commonly accepted characterization of what we wish to explain, the model is still open to dispute as to whether the claimed stylized facts do in fact represent an acceptable characterization. To this extent things have not changed in the last four or five decades. Discussing Kaldor's stylized facts, Robert Solow (1970, p. 2) said that there ‘was no doubt that they are stylized, though it is possible to question whether they are facts’. Nevertheless, Kaldor is to be admired for insisting that we identify our facts to be explained before engaging in a critical comparison of competing explanatory models.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- models

Bibliography

Kaldor, N. 1961. Capital accumulation and economic growth. In *The Theory of Capital*, ed. F. Lutz. London: Macmillan.

Romer, D. 1996. *Advanced Macroeconomics*. New York: McGraw-Hill.

Romer, P. 1989. Capital accumulation in the theory of long-run growth. In *Modern Business Cycle Theory*, ed. R. Barro. Cambridge, MA: Harvard University Press.

Solow, R. 1970. *Growth Theory: An Exposition*. Oxford: Clarendon Press.

How to cite this article

Boland, Lawrence A. "stylized facts." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000315> doi:10.1057/9780230226203.1644

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

subsistence and subsistence wages

Krishna Bharadwaj
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Böhm-Bawerk, E. von; Boisguillbert, P.; Cantillon, R.; Child, J.; circular flow; internal migration; international migration; Locke, J.; Mandeville, B.; market wage; mercantilism; natural wage; Necker, J.; net product; Petty, W.; primitive accumulation; Quesnay, F.; Steuart, J.; subsistence; subsistence wages; Turgot, A. R. J.; wage differentials; wages fund

Article

In ancient economies where slave labour was the predominant form of labour, and for a long time after, when the material conditions of the labourer were not much better, the earnings of labour were viewed as not very different from the feed of horses. There were no ‘theories’ of wages but it was automatically supposed that they receive subsistence, certain fixed quantities of the necessaries of life. As working horses need to be maintained in adequate supply, the worker or the slave would be so provisioned as to be enabled to work and reproduce.

During the second half of the 17th and early 18th century with the rapid rise of commercial capitalism, intensification of competition among trading nations seeking new markets and sources of supply in Europe, discussions arose on management of labour and of wages – in particular, the advantages of maintaining cheap labour. In pursuance of mercantilist ideas, cheap labour was considered a favourable factor in competition. ‘National Wealth’, wrote Mandeville ‘consists not in money’ but in a ‘multitude of laborious poor’. It was also believed that hard work could be compelled out of the poor only by extreme need and want. ‘Men have nothing to stir them up to be serviceable but their wants which it is Prudence to relieve but Folly to cure’ (Mandeville, 1714, p. 194). It was believed that low wages not only spur productivity but also yield commercial advantage since cheap labour meant cheap produce. Thus the mercantilist policy sought to encourage population and selective migration to keep prices of necessaries high through taxes, if necessary, and set wages at a low subsistence level. ‘An increase of people in the country to such a degree as may make things necessary to life dear and thereby force general industry from each member of the family’ seemed a statesman's maxim (William Temple, 1673, p. 116). Discussions on wages were, at this stage, an issue of labour policy and the idea of fixed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
subsistence rested on the extant conditions, as observed. No attempt either to explain the level or the mechanisms stabilizing or restoring that level were discussed. However, it was already recognized, as a practical premise, that taxes on necessaries of subsistence were likely to be accompanied by a rise in money wages (cf. among others, Thomas Mun, 1664, and John Locke, 1692).

It is in William Petty, rightly considered by Marx as among the founders of political economy, that a theoretical role was assigned to the norm of subsistence and the difference between the labourer's subsistence and his produce brought to the fore. Petty was also concerned with quantitative relations within production and with evolving a measure of value.

> The most important consideration in Political Oeconomics [is], viz how to make a Par and Equation between lands and labour, so as to express the value of anything by either alone•…•wherefore the days food of an adult man•…•and not the days labour, is the common measure of value•…•(Petty, 1691, p. 181)

Thus subsistence became a constant measure of value. Defining the value of a commodity in terms of the quantity of food necessary for the day's payment for an adult, he extended the same measure for valuation of different types of labourers. The unit of food was thus a standard of commodity wage. Already by Petty's time the legislative fixation of wages and actual compliance to it were being violated. Attention was gradually drawn to the observed stable level of subsistence, not as an axiom or as a statutory fixation but as an economic condition. The mechanisms that tend to keep wages down were thus explored. John Locke (1692) attempted to explain its existence in terms of psychological inertia among the poor induced by the lowness of wage itself. While active struggle occurred between the land owners and merchants,

> the labourer's share, being seldom more than a bare subsistence, never allows that body of men time or opportunity to raise their thoughts above that unless when some common and great distress, uniting them in one universal ferment, makes them forget respect and emboldens them to carve to their wants with armed force; and then sometimes they break it upon the rich, and sweep all like a deluge. (Locke, 1692, p. 57)

The phenomenon of wages gained importance with the gradual breakdown of guilds, spread of commercial capital and ‘labour’ becoming a commodity. The role played by migration of labour to level wages down to a subsistence level appeared prominently in the writings of Josiah Child who highlighted the effects of national and international migration and brought to the fore also the possibility that higher wage levels may go along with national riches. In contrast to the English emphasis on migration, in France, Boisguilbert and Cantillon took up the question of variations in wages and their effect on accumulation. They, too, while not explaining the *level* of the wage, took it as a datum to draw out the implications of an increase in wage on agricultural surplus and accumulation. The hypothesis of wage as a necessary part of productive consumption led them to consider the net product and the circular process of reproduction. Boisguilbert argued that a higher wage, cutting into the revenues of the landlord, would inhibit further expansion of agriculture and, in turn, diminish the demand for labour, thus putting a downward pressure on wages to revert to their subsistence level. Agriculture was, for him, the branch of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the economy which conditioned the national economy – a position that continued with the Physiocrats – and wage rises in agriculture and its adverse effects on accumulation would spread to other spheres of the national economy. However Boisguilbert did not discuss the possibility of wages getting depressed below subsistence or analyse its implications. Cantillon (1755) who also developed the important role of the agricultural surpluses in forming and shaping the effective demand for products and, in turn, demand for various categories of labour, emphasized the mechanism by which the numbers of farmers, artisans and labourers adjust their supplies of labour to match the level and pattern of demands. He emphasized the role of territorial migration, intra-occupational mobility and also the population mechanism through which supplies of labour adjusted themselves to needs.

This however did not explain the *level* of the wage norm but only the tendencies towards uniformity. While accepting as given the limitation of wages of the unskilled labourer to the amounts necessary for subsistence, he discussed the wage differentials among various occupations arising from ‘varying conditions and circumstances’ – such as

> the number of tradesmen in a given branch of industry, the period of time necessary for learning the trade, the skill and the quality of the labour, the risks and dangers connected with it and finally, the degree of responsibility required of the person entrusted with the performance of a given task. (1755, pp. 25–7)

(An echo of these, we find in Adam Smith's discussion of the natural differences in wages of different labourers.) Unlike his predecessors, Cantillon did give some thought to the content of the subsistence wage and, on certain assumptions (that half the children born die before 17, one-third under one year, and that the labour of the wife, on account of her necessary attendance on the children, is no more than sufficient to provide for herself), he worked out that a worker, to maintain himself and the family, must have double what he requires for his own subsistence, which may be ‘somewhat exceeding that of a slave’.

With the Physiocrats, the assumption of the *given* necessary wage, constituting part of ‘productive consumption’, being ‘advanced’, played an important role in the theory of production and accumulation. Quesnay argued that a tax on necessaries would inevitably be shifted to entrepreneurs. Further, a ‘high price of bread’ was advantageous, not only to the agriculturists but also to the workers (whose money wages would rise sympathetically). ‘It would’, he argued, ‘encourage agriculture, increase the revenue of the nation, increase the wages of the worker and insure a life of comfort, plenty and convenience which would attract people to the land and keep them where they partake all the advantages’. Contrary effects would follow with constraints on exports or a fall in the price of bread. Quesnay, unlike Boisguilbert, did not stress forces preventing the rise of wages above subsistence. He was concerned, it would seem, more with establishing the harmony of interests among the workers and the agriculturists.

It was, however, Turgot who offered a more complete theoretical treatment of subsistence wages:

> The exchange value of food products, profits, the level of wages, and the population are phenomena which are mutually inter-connected and interdependent. The balance among them is established in accordance with a peculiar natural proportion and the proportion is constantly maintained if trade and competition are completely free. (1844, pp. 437–8)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The interaction was visualized as follows. A higher wage (above the natural) increases costs of production, reduces net product, decreases profits. On the other hand, a reduction below the proper level reduces efficiency of labour and also reduces consumption demand, leading eventually to a fall in prices of the produce. A high wage may encourage growth of population or lead to immigration and then, competition among the labourers would tend to lower wages again. Thus for the continued and harmonious economic reproduction, wages and profits must bear a ‘natural proportion’. Thus Turgot attempted to combine, in a complementary fashion, the analysis of Boisguilbert (effect of wage changes on the price of the product, prosperity of entrepreneurs and their demand for labour) and that of Sir Josiah Child (effects of wage changes on migration of labour and its effect on competition).

Somewhat as a reaction to the Physiocratic notion of the ‘natural order’ the writing of Jacques Necker (1775) reflected the political unease of the impending revolution in France. He attributed the tendency of wages to fall to the subsistence minimum to ‘social forces’, particularly to the highly inequitable relationship of power and need that bound the owners ‘who force others to serve them and the propertyless who serve the owners’. The price of labour could adjust to the price of bread only after a time, worsening the state of the worker in the meanwhile. He thus foresaw a ‘dark struggle’ between the owners and the workers. The numerousness of the propertyless, their consequent immediate need and acute competition rendered their bargaining power hopelessly weak, while the growing concentration of property, accentuated further by technical innovations, enhancing productivity of labour but not wages, widened the gap relentlessly. He also pointed out that the accumulation of luxuries at the cost of means of production was leading to a decline in demand for labour. Many of Necker's ideas were to find an echo in Adam Smith.

Steuart (1767) whose approach to wages was more from the point of view of establishing a norm of subsistence for a suitable wage policy beneficial to the development of industry, offered many interesting observations on the process of ‘primitive accumulation’. An interesting contribution on the discussion of subsistence was his distinction between the needs, ‘physically necessary’ and ‘politically necessary’; the former, being ‘ample subsistence where no superfluity is implied’ and the latter, ‘proceeding from the affections of his mind, are formed by habit and education, and when once regularly established create another kind of necessity’ (Steuart, 1767, p. 312). Already the subsistence as strict physiological necessity was being modified with ‘custom’, ‘habit’, ‘rank-based conventions’ playing an important part.

With Adam Smith, Ricardo and Marx, a more integrated view of distribution emerged, along with a distinction between natural wage and market wage. While the forces that tend to put down wages were recognized, the assumption of a fixed wage was replaced by a ‘given’ wage, determined by a complex of historical and economic factors.

The Physiocratic notion of capital as ‘wages advanced’ and the idea that the demand for labour was limited by the provision of subsistence, developed in later classical theory into the concept of wage fund and to an explanation of wage determined by the ‘proportion of capital to labour’. In an entirely different theoretical framework, Böhm-Bawerk used also the notion of subsistence fund to represent given endowments of ‘capital’. In both, the explanation of wage was different from that of the ‘natural wage’.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Boisguilbert, P. 1707. *Traité de la nature, culture, commerce et intérêt des grains*. In *Economistes et financiers du xviiie siècle*, ed. E.D. Daire. Paris: Guillaumin, 1851.

Cantillon, R. 1755. *Essai sur la nature du commerce en général. Traduit de l'anglais*. London: F. Gyles. Trans. H. Higgs, London: Macmillan, 1931.

Child, Sir J., 1690. *A New Discourse of Trade*. London.

Locke, J. 1692. *Some Considerations of the Consequences of the Lowering of Interest and Raising the Value of Money*. London.

Mandeville, B. de. 1714. *Fable of the Bees or Knaves Turned Honest*. Oxford: Clarendon Press, 1924.

Mun, T. 1664. *England's Treasure by Foreign Trade*. London. Ed. W.J. Ashley, London: Macmillan, 1894.

Necker, J. 1775. *De la législation et du commerce des grains*. Paris.

Petty, W. 1691. *Political Anatomy of Ireland*. Repr. In *Economic Writings of Sir William Petty*, ed. C.H. Hull, Cambridge: Cambridge University Press, 1899.

Ricardo, D. 1817. *Principles of Political Economy and Taxation*. In *The Works and Correspondence of David Ricardo*, vol. 1, ed. P. Sraffa and M.H. Dobb. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. E. Cannan, New York: Modern Library, 1937.

Steuart, Sir J. 1767. *An Inquiry into the Principles of Political Economy*. London: Oliver & Boyd.

Temple, Sir W. 1673. *An Essay Upon the Advancement of Trade in Ireland*. London.

Turgot, A.R.J. 1844. *Observations sur le mémoire de M. Graslin*. In *Oeuvres de Turgot*, ed. E. Daire and H. Dussard. Paris: Guillaumin.

How to cite this article

Bharadwaj, Krishna. "subsistence and subsistence wages." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000322> doi:10.1057/9780230226203.1645

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

substitutes and complements

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Economists have found it surprisingly hard to nail down the obvious intuition that in some rough unspecified way tea and coffee are substitutes, and bacon and eggs complements. Not that they can't do it. Quite the opposite, they have all too many ways of doing it, so much so that even if their less attractive inventions are discarded still an abundance is left, each with its own usefulness and charm. Yet because complementarity and income effects are the two chief impediments to sharp results in microeconomic theory, this question of appropriate definition cannot be shrugged off. Neither does it help that what ‘appropriate’ means tends to vary from problem to problem.

Two important examples exhibit briefly and in turn the fair face of substitutability and the ugly mug of complementarity. In the theory of Walrasian adjustment processes for multiple markets, a famous result traceable in its origins to Metzler (1945) is that if all commodities are (gross) market substitutes then equilibrium is reached from any initial position, i.e. the system is *globally stable* (see Negishi, 1962). Secondly, in capital theory Hatta (1976) has shown that in order for the output/labour ratio to fall as the interest rate falls, i.e. in order for *capital perversities* to appear, it is *necessary* that at least one input pair be (net) complements, which in turn implies that there must be at least three inputs in total.

Despite these examples and purely for reasons of space, this essay will concentrate entirely on substitution and complementarity in the theory of consumer's demand, to the neglect of production. In doing this it differs only in degree rather than kind from the literature of the subject, of which surveys may be found in Schultz (1938, pp. 22–4, and chs 18 and 19, written with the help of Milton Friedman), Stigler (1950, section VI; reprinted in 1965), Georgescu-Roegen (1952) and Samuelson (1974).

Two further disclaimers are in order. Neither the important role in the Austrian theory of value played by what Menger called complementary goods (e.g. 1950, ch. I.3), nor the question of the relations between complementarity and changes in tastes or endowments (begun by Lange, 1940, and analysed by Hicks, 1956, pp. 161–8) is discussed here.

I The Auspitz – Lieben Definitions

Passing recognition that certain pairs of goods can be called substitutes and other pairs complements

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

must have occurred ever since economics became a serious subject. However, that is far from developing formal definitions and results. A clear though crude attempt to derive explicit comparative static propositions for substitutes and ‘co-elements’ may be found in chapter IX of Donisthorpe (1876) but a more adequate theory had to await formulation of utility functions in general form, i.e. $u(x, y,\ldots)$, rather than in the additively separable form $u^1(x) + u^2(y) + \cdots$ employed for example by Jevons.

So it is not unexpected that the first hint of a definition is to be found in *Mathematical Psychics* (1881, p. 34). Edgeworth was concerned to prove that his newly-minted indifference curves between ‘sacrifice objectively measured’ x, and ‘objectively measured remuneration’ y, were convex. Writing u_x, u_{xx}, … for $\partial u(.,.)/\partial x$, $\partial^2 u(.,.)/\partial x^2$, etc., Edgeworth's five assumptions may be expressed: $u_x < 0$, $u_y > 0$, and u_{xx}, u_{yy}, and u_{xy} all ‘continually *negative*. (Attention is solicited to the interpretation of the third condition.)’ Maddeningly, he stopped right there. Here, the invited attention will be delayed for two paragraphs.

At various times Fisher (1892, Part II), Edgeworth (1897, p. 21; 1925, Vol. 1, p. 117 n.1), and Pareto (e. g. 1927, pp. 268–9, 575–6) have each been said to have introduced the first formal definitions of substitutes and complements. However, Stigler (1950; 1965, p. 131) pointed out that in fact the credit belongs to Auspitz and Lieben (1889, p. 482; 1914, Texte, pp. 318–19), for whom two commodities x and y were *complements*, *independent*, or *competitive* according as

$$u_{xy} > 0;\ u_{xy} = 0;\ u_{xy} < 0;$$

These A–L definitions (as they will be called here) have three important virtues and one fatal flaw. First, they are intuitively appealing. It seems commonsensical to say that y is a complement of x if an increase in the latter raises the marginal utility of the former, and so on. Secondly, for sufficiently smooth utility functions Young's Theorem implies that always $u_{xy}=u_{yx}$, so that the A–L definitions are symmetrical; if (and only if) x is a substitute for y then y is a substitute for x. Finally, the relations between any pair of commodities involve only that pair alone and the individual agent. In particular, they do not ‘depend upon the incidents of a comparatively advanced regime, such as the distribution of money among different purchases’ (Edgeworth, 1925, Vol. II, p. 465), i.e. they do not depend upon market phenomena. For positivistically inclined economists like Slutsky ([1915] 1951, pp. 52–56) this last property appeared more vice than virtue.

However, the flaw truly is (as Bertrand would say) ‘une objection peremptoire’. Suppose that u is transformed to $v=F(u)$, where $F' > 0$. Then u_x is transformed to $v_x = F' u_x$ and u_{xx} to $v_{xx} = F' u_{xx} + F'' u_x^2$ and similarly for u_y and u_{yy}, respectively; in particular, u_{xy} is changed to $v_{xy} = F' u_{xy} + F'' u_x u_y$. So by cunning choice of F the sign of u_{xy} can be changed. For example, F can be taken to be very convex by making F'' positive and large, in this way changing a negative u_{xy} into a positive v_{xy}. Substitutes can be made complements, or complements substitutes, all without any change

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
occurring in the agent's tastes. The A–L definitions are therefore meaningless in any theory of demand that abandons cardinal utility.

Now to interpret Edgeworth's assumption $u_{xy}<0$. According to the A–L definitions, $u_{xy}<0$ means that sacrifice x is a substitute for remuneration y; increasing the amount of labour supplied, for example, lowers the marginal utility of the commodity received, so that leisure (non-labour) is then a complement for y (c.f. Stigler, 1965, p. 99, fn 84). While the argument of the last paragraph would seem to imply that this A–L interpretation is obviously meaningless, as so often with Edgeworth matters are actually rather more subtle. His sufficient condition C for strict convexity of indifference curves, given in (1881, pp. 35–6), is indeed the same as the modern strong quasiconcavity condition on the bordered Hessian of u, as given by Hicks (1939, p. 306). Hence C is invariant under strictly monotonic transformations F, and so one cannot pick F merely to change from $u_{xy}<0$ to $v_{xy}>0$. The F chosen to do this trick must also simultaneously transform u_x, u_y, u_{xx}, and u_{yy} in such a way as to preserve C as well.

II Johnson and after

Pareto himself, so vehement against cardinal utility, seems never to have seen the inconsistency between that epistemological position and his continued use of the A–L definitions, and in this he was followed by Zawadzki (1914, pp. 171–4). But not by Slutsky, who was quite explicit that the A–L definitions were meaningless in his radically new theory of demand ([1915] 1951, pp. 54–5). However, unlike Hicks and Allen (1934) he did not see the possibilities, opened up by the new theory, of defining substitutes and complements in a way free of dependence on cardinally measurable utility.

One reason for this omission may have been Slutsky's apparent unawareness of the work of the logician W.E. Johnson (1913; 1968) who, from the partial-equilibrium isolation of Cambridge and without specific reference to anybody else, had already offered definitions not dependent on a particular utility index. Let $V = u_x / u_y$ and $W=V^{-1}$, so that in Hicks–Allen language V is the *marginal rate of substitution of y for x*. Then Johnson (1968, 108) defined x and y to be complementary if both $\partial V/\partial x$ and $\partial W/\partial y$ are negative, while they are competitive if either $\partial V/\partial x$ or (disjunctively) $\partial W/\partial y$ is positive. He shows later (p. 114) that, at least in the case of two goods, x and y are complementary if and only if (iff) each of them is normal (i.e. has a positive income elasticity), whereas $\partial V/\partial x > 0 (\partial W/\partial y > 0)$ holds iff x is normal (inferior) and y inferior (normal).

These definitions are clearly invariant to permissible transforms of u, and equally obviously symmetrical; moreover, they are at least formally independent of market phenomena. The main difficulties in accepting them are their lack of intuitive appeal (which Johnson did nothing to mitigate) and their excessively stringent implications. An *ex post facto* rationalization of the definition of complements can however be given, using arguments couched in the language of cardinally measurable utility (a similar account can be given of his substitutes). Thus an increase in x has two effects: (a) If y is a complement of x in the rough everyday sense, then as in the A–L case we would expect u_y to rise, and hence the rate of compensation for the loss of x by this now more highly valued y to fall; and (b) By the law of diminishing marginal utility u_x is lowered, and so again this reduces the necessary rate of compensation by y for the loss of x.

The implication that if both goods are normal they must be Johnsonian complements seems

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
unacceptably restrictive, though perhaps little more so than the conclusion that if there are only two goods they must be Hicks–Allen substitutes. The latter result has of course been familiar and broadly accepted for over half a century, though not without serious criticism (see e.g. Pearce, 1958, pp. 136ff).
The textbook definitions of substitutes and complements used today first appeared in Hicks and Allen (1934), although recent versions are much simpler and more transparent than the original. Given that their work was in large part a response to the lack of invariance of the A–L definitions (Hicks, 1981, p. 27, fn42), it is not surprising that substitution and complementarity played so prominent (to modern eyes, too prominent) a role in both parts of their paper. Moreover, Hicks' 'literary' expositions of this topic, not only in (1934) but also in (1939, ch. 3) and (1956, ch. 16), seem overly complex and unrelated to his mathematical expositions, to such a degree that Samuelson (1947, pp. 183–9) was led to allege actual inconsistencies between them, a criticism that he later formally withdrew (1950, p. 379, fnl; 1974, p. 1286).
One reason why it is hard to understand Hicks's analysis is that he works always with a numéraire commodity ('money'), which itself enters into the agent's preferences. Thus in considering Hicks–Allen substitutability between a pair of goods *x* and *y*, for example, he has to consider the effect of a change in the 'money' price of *x* on the compensated demand for *y*, which necessarily involves three goods, *x*, *y* and 'money', not two. It is simpler and clearer to treat all goods symmetrically, by normalizing their prices not by use of a numéraire but by income or total expenditure, assumed always to be positive. Such a normalization will be used throughout the next section, even though that prevents any direct comparison of its analysis with much of the earlier literature on the subject, especially that influenced by Hicks.

III Some modern definitions

Assume that the representative agent (trader, consumer) has preferences $\succsim$ defined over the commodity space R^n, and focus attention on his (or her) "target" bundle $z^t=(x^t, y^t, \ldots)$, assumed for present purposes to be given exogenously. Then his "better set" B^t is $\{z: z \in R^n, z \succsim z^t\}$, which will be assumed always to be convex, closed and such that if *z* is in it then so is λz for all $\lambda \geq 1$. If preferences are incomplete a utility function will not exist, but in any case those preferences can be represented by an *s-gauge* function for B^t, defined as follows (see the entry on GAUGE FUNCTIONS):

$$j(z|z^t) = \sup\{\mu > 0: z \in \mu B^t\} \text{ if } z \neq O = 0 \text{ if } z = O \tag{1}$$

where *O* denotes the origin. This function $j(\cdot|z^t)$ is a generalization of the so-called distance function, discussed for example in Deaton (1979) and Deaton and Muellbauer (1980, pp. 53–7). It is easy to show that it is concave and positively homogeneous of degree 1 (phd1) in *z*. Next, assume that the agent faces

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

a vector of prices $p=(p_x, p_y, \ldots)$ determined on competitive markets, and give him the following problem of cost-minimization for the target z^t:

Find $z \in R^n$ that achieves $\inf\langle z, p\rangle$ subject to $z \in B^t$ where $\langle .,. \rangle$ denotes the inner product of z and p. The value of any solution to this problem is functionally dependent on its parameters p and z^t, a fact summarized by writing down the expenditure function $e(p+z^t)$. Assume now and always that $e(p+z^t)$ is positive and divide each market price by this amount, thus arriving at a vector of normalized (indeed personalized) prices $w=(w_x, w_y, \ldots)$. Following this normalization, the function $e(\cdot|z^t)$ becomes the *cost function* $c(\cdot|z^t)$, where $c(w|Z^t)=1$, a pure number. This function is concave and phd1 in w, and is actually the (concave) support function of B^t.

So defined, the *s*-gauge and cost functions are dual to each other, in the sense that under general conditions $c(\cdot|z^t)$ is the *s*-gauge function of the set B^{t*} in the (normalized) price space that is polar to B^t; and dually, under similar conditions $j(\cdot|z^t)$ is the concave support function of B^{t*}.

We can now expeditiously set out various modern definitions of substitutes and complements. For this purpose I will assume where necessary and without blushing that each function concerned is as differentiable as desired.

A Hicks – Allen

By the well-known Shephard's Lemma, the compensated or Hicksian demand function h_i for the *i*th commodity is given by

$$\forall i h_i\left(w|z^t\right) = \partial c\left(w + z^t\right) / \partial w_i \tag{2}$$

Then good x is a *substitute* (*net* or *Hicks-Allen* substitute) for good y if

$$\partial h_x\left(w|z^t\right) / \partial w_y > 0 \tag{3}$$

i.e. increasing the price of y leads the agent to buy more x, all the time keeping to the better set B^t and minimizing the cost of doing so (as Marshall would say, obeying the Law of Substitution). Similarly, x is a *complement* (*net* or *Hicks–Allen* complement) if the inequality in (3) is reversed; otherwise, x is *independent* of y.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

It follows from (2) and (3) that it is equivalent to say that *x* is a substitute for *y* if

$$\partial^2 c\left(w|z^t\right) / \partial w_y \partial w_x > 0$$
(4)

For sufficiently smooth $c(\cdot|z^t)$, the left-hand-side (LHS) of (4) is the same as $\partial^2 c(w|z^t) / \partial w_x \partial w_y$, from which it follows that if *x* is a substitute (complement, independent) of *y*, then *y* is a substitute (complement, independent) of *x*. The Hicks-Allen definitions are symmetrical.

Since $c(\cdot|z^t)$ is phd1 it follows from (2) that $h_i(\cdot|z^t)$ is phd0, and thence from Euler's Theorem that

$$\forall i \left\{ w, \nabla h_i\left(w|z^t\right)\right\} = 0$$
(5)

Because $c(\cdot|z^t)$ is concave, $\partial^2 c(w|z^t) \partial w_i^2 \geq 0$ for every *i*. So it follows from this and (5) that, taking the summation over all *j*=1, 2, …, *n*–1 for $j \neq i$,

$$\forall i \sum w_j \partial h_i(w|z^t) / \partial w_j \geq 0$$
(6)

Hence, while all goods can be substitutes for each other, it cannot happen that all goods are complements, for that would imply the negation of the inequality in (6). It is in this sense that, in a Hicks-Allen world, substitution predominates. Moreover, in the two good (*x*, *y*) case (6) implies that $\partial h_x(w + z^t) / \partial w_y \geq 0$, so that *x* and *y* cannot ever be complements. An extension of this reasoning shows that in the three good case, at most one pair of goods can be complements; and so on.

B Hicks – Deaton

By design, the *s*-gauge function $j(\cdot|z^t)$ has just the same mathematical properties as its polar transform $c(\cdot|z^t)$. So the following result is polar to Shephard's Lemma and is obtained in precisely the same way

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(cf Deaton, 1979, 394):

$$\forall i w_i = \partial j\left(z|z^t\right) / \partial z_i \tag{7}$$

The interpretation of (7) is that it yields the compensated or Hicksian *inverse* demand functions. Roughly speaking, these show how the “marginal valuation” of (or the “marginal-willingness-to-pay” for) good *i* varies, as one moves around the lower boundary of B^t (see Hicks, 1956, ch. 16, which uses a numéraire, and Deaton, 1979, which does not). Note the similarities to and contrasts with the definitions of Johnson, which did not involve movements around this lower boundary but, instead, movements away from it.

Denote these inverse Hicksian functions by H_i, so that from (7)

$$\forall i H_i\left(z|z^t\right) = \partial j\left(z|z^t\right) / \partial z_i \tag{8}$$

If *x* and *y* are complements in the rough everyday sense, one would expect that as one has more *y* one would be willing to pay more for a marginal unit of *x*, while if they are substitutes one would be willing to pay less. This is the intuitive basis for the following Hicks–Deaton definitions. A good *x* is a *q*-substitute for *y* if $\partial H_x(z|z^t) / \partial y < 0$, and a *q-complement* if this inequality is reversed; otherwise, *x* and *y* are *q-independent* (the language, though not the precise definitions, is due to Hicks, 1956, ch. 16). From (7) and (8) these definitions are symmetrical.

Since $j(\cdot|z^t)$ is phd1 each $H_i(\cdot|z^t)$ is phd0. From this and the fact that $j(\cdot|z^t)$ is concave, analogues to (5) and (6) are readily derived and imply, for example, that in the Hicks–Deaton world *q*-complements predominate, and that in a two-good world there cannot be *q*-substitutes. It obviously follows that it is *not* true in general that if *x* and *y* are Hicks–Allen substitutes then they are *q*-substitutes, and similarly for the two definitions of complements.

Note that the gradient mappings $\nabla j(\cdot|z^t)$ and $\nabla c(\cdot|z^t)$ are inverse to each other, a property which persists even when compensated demand functions (direct and inverse) are generalized to compensated demand correspondences (see Theorem 9 in the entry on GAUGE FUNCTIONS).

C Samuelson's money-metric definitions

Samuelson (1974, pp. 1272–3) proposed the following ‘local measure of money-metric

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

complementarity'. The good *x* is a substitute for, independent of, or a complement of *y*, according as $\partial^2 c(w|z^t) / \partial x \partial y$ is negative, zero, or positive, i.e. the test criterion is the behaviour of (minimized) cost as the target bundle z^t is varied. The original paper should be consulted for Samuelson's rationale of this measure, which does not appear to have become popular.

IV Gross substitutes and complements

Stigler (1950; 1965, pp. 134–5) argued that once the intuitively appealing A–L definitions have succumbed to the ordinalist critique, there is no point in stopping at such halfway non-intuitive houses as the Johnson or the Hicks–Allen definitions. Instead, one should go the whole way to 'simple criteria such as the cross-elasticity of demand', moving completely from psychological to market data.

For individual ordinary (Marshallian) demand functions $z_i = f_i(p, \omega)$, *i*=1, 2, …, *n*, where *w* is the individual's wealth and prices are normalized by a numeraire, these criteria in essence reduce to the following definitions, given originally by Mosak (1944, p. 45): A good *x* is a *gross substitute* for *y* if $\partial f_x(p, \omega) / \partial p_y$ is positive, a *gross complement* if it is negative, and *independent* of *y* otherwise. Actually, in Mosak's original definitions these three situations corresponded not to *x* being a gross substitute for *y*, etc., but to *y* being a gross substitute for *x*, etc. Subsequent literature has continued this ambiguity, which is by no means trivial, for the chief problem with these 'simple criteria' is that they are not symmetrical. Hence *x* can be a gross complement for *y*, and *y* simultaneously a gross substitute for *x*, a highly non-intuitive state of affairs.

The proof of such possible non-symmetry is standard textbook fare and so is only quickly sketched here. The standard Hicks decomposition of the effect on the (ordinary) demand for a good *x* of a simple price change in a good *y*, utility level τ^* and chosen bundle $z^* = (x^*, y^*, \ldots)$ is,

$$\partial f_x(p, \omega) / \partial p_y = \partial h_x\left(p, \tau^*\right) / \partial p_y - y^* \partial f_x(p, \omega) / \partial \omega \tag{9}$$

Suppose *x* is a gross substitute for *y*, so that the LHS of (9) is positive. Suppose also that *x* is a Hicks–Allen substitute for *y*, so that the first term on the RHS is also positive (h_x being the compensated demand function for *x*). Consider now the similar Hicks decomposition for good *y*. By the symmetry of Hicks–Allen substitutes the first term on its RHS will be the same as the corresponding term of (9), and so positive. But if *y* is a normal good and x^* sufficiently large, then the whole RHS might be negative, and so *y* a gross complement for *x*. If on the other hand *x* is a Hicks–Allen complement for *y* the first term on the RHS of the decomposition for *y* will be negative, and so if it is a normal good, again it will be a gross complement for *x*.

Given a fixed distribution of endowments the individual ordinary demand functions for each good may be aggregated to a market demand function (or an excess demand function) for that good, and gross substitutes and complements then defined in terms of these market functions, as done for example by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Metzler (1945). Mosak (1944, pp. 46–7) inverted such market functions and defined ‘gross complementarity [and substitutability] in the inverse sense’, which differs from ordinary gross complementarity in very roughly the same way that q-complements differ from Hicks–Allen complements.

Bibliography

Auspitz, R. and Lieben, R. 1889. *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot. French translation published in two volumes (Texte et Album), Paris: Giard, 1914.

Baumol, W.J. and Goldfeld, S.M. 1968. *Precursors in Mathematical Economics. An Anthology*. Reprints of Scarce Works in Political Economy No. 19, London: London School of Economics.

Boulding, K.E. and Stigler, G.J. 1951. *Readings in Price Theory*. London: George Allen & Unwin Ltd.

Deaton, A. 1979. The distance function in consumer behaviour with applications to index numbers and optimal taxation. *Review of Economic Studies* 46, 391–406.

Deaton, A. and Muellbauer, J. 1980. *Economics and Consumer Behaviour*. Cambridge: Cambridge University Press.

Donisthorpe, W. 1876. *Principles of Plutology*. London: Williams and Norgate.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Edgeworth, F.Y. 1897. Teoria pura del monopolio. *Giornale degli Economisti* 33. Translated in Edgeworth (1925), Vol. I, 111–42.

Edgeworth, F.Y. 1925. *Papers Relating to Political Economy*, 3 vols. London: Macmillan.

Fisher, I. 1892. Mathematical investigations in the theory of value and prices. *Transactions of the Connecticut Academy*, IX. Republished, New Haven: Yale University Press, 1925.

Georgescu-Roegen, N. 1952. A diagrammatic analysis of complementarity. *Southern Economic Journal* 19, 1–20.

Hatta, T. 1976. The paradox in capital theory and complementarity of inputs. *Review of Economic Studies* 43, 127–42.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hicks, J.R. 1956. *A Revision of Demand Theory*. Oxford: Clarendon Press.

Hicks, J.R. 1981. *Wealth and Welfare*. Cambridge, Mass.: Harvard University Press.

Hicks, J.R. and Allen, R.G.D. 1934. A reconsideration of the theory of value. *Economica*, N.S. 1, 52–76, 196–219. Reprinted in Hicks (1981), 3–55.

Johnson, W.E. 1913. The pure theory of utility curves. *Economic Journal* 23, 483–513. Reprinted in Baumol and Goldfeld (1968), 97–124.

Lange, O. 1940. Complementarity and interrelations of shifts in demand. *Review of Economic Studies* 8, 58–63.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Trans. by J. Dingwall and B. Hoselitz as *Principles of Economics*, Glencoe, Ill: Free Press, 1950.

Metzler, L.A. 1945. Stability of multiple markets: the Hicks conditions. *Econometrica* 13, 277–92.

Mosak, J.L. 1944. *General Equilibrium Theory in International Trade*. Cowles Commission Monograph No. 7, Bloomington: Principia Press.

Negishi, T. 1962. The stability of a competitive economy: a survey article. *Econometrica* 30, 635–69.

Pareto, V. 1909. *Manuel déconomie politique*. Paris: Giard. 2nd edn, 1927.

Pearce, I.F. 1964. *A Contribution to Demand Analysis*. Oxford: Clarendon Press.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Samuelson, P.A. 1950. The problem of integrability in utility theory. *Economica*, N.S. 17, 355–85.

Samuelson, P.A. 1974. Complementarity – an essay on the 40th anniversary of the Hicks–Allen revolution in demand theory. *Journal of Economic Literature* 12, 1255–89.

Schultz, H. 1938. *The Theory and Measurement of Demand*. Chicago: University of Chicago Press.

Slutsky, E. 1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti* 51. Trans. in Boulding and Stigler (1951), 27–56.

Stigler, G.J. 1950. The development of utility theory. *Journal of Political Economy* 58, 307–27, 373–96. Reprinted in Stigler (1965), 66–155.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Stigler, G.J. 1965. *Essays in the History of Economics*. Chicago: University of Chicago Press.

Zawadzki, W.L. 1914. *Les mathematiques appliquées a l'économie politique*. Paris: Rivière. Reprinted New York: Burt Franklin Research & Source Works Series No. 93, n.d.

How to cite this article

Newman, Peter. "substitutes and complements." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000323>
doi:10.1057/9780230226203.1646

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sugar industry

Wallace P. Mullin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

collusion; entry; firm, theory of; information sharing among firms; market power; mergers; non-price agreements; price fixing; sugar industry; Sugar Institute (USA); sugar tariff (USA); transaction cost economics

Article

The sugar industry is the set of firms involved in processing sugar from sugar cane or sugar beets and distributing it to consumers. The industry is both ancient and modern. Sugar production has occurred for thousands of years. But the developments since 1880 have received the most attention from economists. Over subsequent decades, the industry experienced several structural changes and legal regimes. The industry's evolution provides a window into central questions such as the sources of market power, and the theory of the firm. Much of this research has focused on the United States, although it was a global industry.

A firm or a group of firms is able to exercise market power when they are able to charge prices above marginal cost and earn supra-competitive profits over a prolonged period of time.

One possible source of market power is through mergers with competitors. In the United States, the trust movement and the merger wave at the turn of the 20th century were early and dramatic attempts. In 1887, the Sugar Trust was formed, later reorganized as the American Sugar Refining Company (ASRC). Eichner's exhaustive study (1969) details the rise of ASRC, its attempts to maintain its market power as a dominant firm, its decline in market share in part due to an antitrust action, and its subsequent position as an oligopoly leader.

For a firm to retain market power, it must prevent entry from eroding its profits. After 1887, two major entry episodes sparked price wars. Zerbe (1969) interprets this as a return to competition. In contrast, Genesove and Mullin (2006) agree with Eichner (1969) that ASRC's actions were predatory. Genesove and Mullin (2006) show that price fell below direct measures of marginal cost, and below counter-factual competitive prices. Predation forced the prey to sell out on favourable terms, and may have

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
deterred future entry. Predation worked by strengthening ASRC's reputation.

An important empirical issue is how to measure or infer market power when a researcher is unable to observe marginal cost, and therefore the price–cost margin. A large literature uses static oligopoly models to estimate market conduct and unobserved cost components. Genesove and Mullin (1998) evaluate this approach by comparing the estimated measures to direct measures. The approach is largely validated, although partial cost information can still be useful.

Besides attempts by a single firm to achieve market dominance, market power may also be created by a group of firms acting collusively, or attempting to soften competition between themselves. That was the primary aim of the Sugar Institute, a trade association of cane sugar refiners that operated from 1928 to 1936. During that time, price fixing was illegal, but non-price agreements and non-price discussions were permitted.

Genesove and Mullin (2001) explore the theory of collusion by a detailed examination of the Institute's workings as revealed by the internal business memoranda prepared by one of its participants. The analysis reveals how firms may alter their environment to enhance the probability of detecting price-cutting through both specific rules and institutional structure. In particular, the Institute served as a forum for communication. In practice, these communications provided advance notification of changes in business practices, a method of coordinating behaviour, and determination of guilt when a firm was accused of violating the industry's Code of Ethics.

The Institute also facilitated the exchange of firm-level information. The Institute collected production and delivery data from the individual firms and returned them in aggregated form. But attempts to exchange sales data were stymied by the larger firms, and so firm heterogeneity was an important impediment to more successful information sharing (Genesove and Mullin, 1999).

The sugar industry had important international dimensions.

The early 20th century saw intense and visible debate in the United States on trade policy. Ellison and Mullin (1995) examine the factors, such as location of corporate headquarters, constituent employment in the sugar industry, and constituent shareholding in sugar firms, that affected legislators’ votes on sugar tariff legislation. They find that political efficacy lay in large, unconcentrated interest groups, such as sugar cane and beet farmers. Ironically, the latter groups would not have existed absent prior protective tariff legislation (Krueger, 1990).

Sugar cane was best grown and processed into raw sugar in the tropics. Dye (1998) shows that much of the development of the Cuban sugar industry can be explained by modern theories of the firm. For example, the adoption of new technologies by mills was strongly influenced by asset-specificity concerns as predicted by transaction cost economics.

See Also

- antitrust enforcement
- cartels
- information sharing among firms

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dye, A. 1998. *Cuban Sugar in the Age of Mass Production*. Stanford: Stanford University Press.

Eichner, A. 1969. *The Emergence of Oligopoly: Sugar Refining as a Case Study*. Baltimore, MD: Johns Hopkins University Press.

Ellison, S. and Mullin, W. 1995. Economics and politics: the case of sugar tariff reform. *Journal of Law and Economics* 38, 335–66.

Genesove, D. and Mullin, W. 1998. Testing static oligopoly models: conduct and cost in the sugar industry, 1890–1914. *Rand Journal of Economics* 29, 355–77.

Genesove, D. and Mullin, W. 1999. The Sugar Institute learns to organize information exchange. In *Learning by Doing in Markets, Firms, and Countries*, ed. N. Lamoreaux, D. Raff, and P. Temin. Chicago: University of Chicago Press.

Genesove, D. and Mullin, W. 2001. Rules, communication and collusion: narrative evidence from the Sugar Institute case. *American Economic Review* 91, 379–98.

Genesove, D. and Mullin, W. 2006. Predation and its rate of return: the sugar industry, 1887–1914. *Rand Journal of Economics* 37, 47–69.

Krueger, A. 1990. The political economy of controls: American sugar. In *Public Policy and Economic Development*, ed. M. Scott and D. Lal. New York: Oxford University Press.

Zerbe, R. 1969. The American Sugar Refinery Company 1887–1914: the story of a monopoly. *Journal of Law and Economics* 12, 339–75.

How to cite this article

Mullin, Wallace P. "sugar industry." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000459> doi:10.1057/9780230226203.1647

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sunspot equilibrium

Karl Shell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Sunspots’ is short for an *extrinsic* random variable, that is, one that does not affect economic fundamentals but can affect economic outcomes. Sunspots are said to matter when the allocation of resources depends in a non-trivial way on the realization of the sunspot random variable. Sunspot equilibria are instances of ‘excess volatility’. They arise even when expectations are fully rational. Separate sources of sunspot equilibria include unbounded time horizons, incomplete markets, restricted market participation, imperfect competition, non-convexities, externalities, asymmetric information and financial indeterminacy. Sunspot equilibria are typically not mere randomizations over certainty equilibria.

Keywords

asymmetric information; banking crises; bank runs; bootstrap method; bubbles; business cycle; calibration; contingent commodities; correlated equilibrium; endogenous uncertainty; excess volatility; expectations; extrinsic uncertainty; financial fragility; general equilibrium with incomplete markets; imperfect competition; indivisible goods; infinite horizons; intrinsic uncertainty; Jevons, W. S.; law of large numbers; learning; lottery equilibrium; market games; market uncertainty; mixed stragey equilibrium; monetary search models; monopolistic competition; multiple equilibria; non-convexities; options and derivatives; overlapping generations models; perfect foresight; rational expectations; real business cycles; restricted market participation; self-fulfilling prophecy; sunspot equilibrium; sunspot immunity theorem; symmetry breaking; volatility

Article

The volatility of market outcomes such as the price level, stock market prices, unemployment rates, interest rates, and exchange rates and what to do about this are important subjects in macroeconomics. Some of the observed randomness in market outcomes is the result of shocks to the fundamentals (preferences, technologies, and endowments) that are transmitted through the economy. Uncertainty about the economic fundamentals is intrinsic uncertainty. The general-equilibrium model extended by Arrow (1953; 1964) to include uncertainty provides an explanation of how volatility in the fundamentals is transmitted through the economy, resulting in volatile prices and quantities. This is not the only possible source of the volatility in economic outcomes. The market economy is a social system. In attempting to optimize her own actions, each agent must attempt to predict the actions of the other agents. A, in forecasting the market strategy of B, must forecast B's forecasts of the forecasts of others including those of A herself. An entrepreneur is uncertain about the moves of his customers and his rivals, and they of his moves. It is not surprising that this process may generate uncertainty in outcomes even in the extreme case in which the fundamentals are non-stochastic. The uncertainty generated by the economy is market uncertainty. It is either created by the economy or adopted from outside the economy as a means of coordinating the plans of individual agents. Market uncertainty is not transmitted through the fundamentals. It can be driven by extrinsic uncertainty.

Extrinsic uncertainty

‘Sunspots’ is shorthand for ‘the extrinsic random variable’ (or ‘extrinsic randomizing device’) upon which agents coordinate their decisions. In a proper sunspot equilibrium, the allocation of resources depends in a non-trivial way on sunspots. In this case, we say that *sunspots matter*; otherwise, sunspots do not matter. Sunspot equilibrium was introduced by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Cass and Shell (see Shell, 1977; Cass and Shell, 1982; 1983). Sunspot models are complete general-equilibrium models that offer an explanation of excess volatility. It was by no means a new idea that economies can and do generate excess volatility, but the sunspots model is the first general-equilibrium model to exhibit excess volatility even when agents are fully rational. The sunspots model also allows for non-rational agents, but the excess volatility from this source – while possibly empirically substantial – is less novel.

'Sunspots' is an unfair spoof on Jevons (1884), who in serious empirical work attempted to explain the business cycle by relating it to the observed (through telescopes) cycle of actual sunspot activity. To the extent that actual sunspot activity does affect economic fundamentals (such as crop yields and cancer risk), this is an instance of intrinsic uncertainty, but the effects of actual sunspots on fundamentals are probably very small. If actual sunspots have only a minor effect on the fundamentals, but they do have a substantial effect on the economy, they must serve a role in the economy beyond their effects on the fundamentals. Manuelli and Peck (1992) show that a sunspot equilibrium can be interpreted as the limit of traditional rational-expectations equilibria as the uncertainty in the fundamentals vanishes (see Spear, Srivastava and Woodford, 1990). Roughly speaking, 'Jevons equilibrium' becomes 'Cass–Shell equilibrium' as the effects of actual solar activity on the fundamentals disappear. Cass–Shell sunspot equilibria are easy to interpret because in the basic sunspots models the only uncertainty is extrinsic uncertainty; hence any volatility in outcomes is excess volatility. Engineers compute 'gain' in noise as the volatility of the output signal divided by the volatility of the input signal. In a sunspot equilibrium, the gain is $+\infty$.

Overlapping generations sunspots

The first sunspots model, Shell (1977), is of an overlapping-generations exchange economy with taxes and transfers denominated in fiat money. This OG model is based on the very simple (degenerate) example used in Shell (1971) to show that restrictions on market participation are inessential in the Samuelson perfect-foresight (non-stochastic) OG model. The only stochastic element in the 1977 paper is sunspot-driven extrinsic uncertainty about the price level. Shell used the fact that there is a continuum of equilibria (parameterized by the initial price level) in the non-sunspots version to construct the sunspot equilibrium allocation by a bootstrap method. This particular sunspot equilibrium, while bootstrapped from multiple certainty equilibria, is *not* a mere randomization over certainty equilibria, contrary to what some popularizers of sunspot equilibrium have claimed. Sunspot equilibria can be randomizations over certainty equilibria, but typically they are not. In unpublished work in about 1975, Cass and Shell generalized the OG sunspots analysis from the degenerate linear model to the concave-utility-function OG model of Gale (1973). Peck (1988) showed that, for the 'Samuelson' and related cases in the OG economy, sunspots can be active in every period for economies with even non-stationary environments.

Azariadis (1981) translated the pure-exchange OG model into a macro-oriented Lucas-style OG model with capital investment and endogenous labour supply. Azariadis showed instances – based on a backward-bending offer curve – of economies that exhibit long-run stationary sunspot cycles. Azariadis and Guesnerie (1986) employed the backward-bending offer curve to exhibit economies with long-run deterministic cycles. They thus established a link between sunspot cycles and deterministic cycles. Roughly speaking, if there is room to condition expectations on sunspots, there is also room to condition them on calendar time and vice versa (see Cass and Shell, 1980).

Sunspot immunity

To better understand how sunspot equilibria arise, consider two simple, related examples. In the first, the economy is immune from sunspots. In the second, all competitive equilibria are sunspot equilibria. Consider the two-consumer, one-good, two-states-of-nature, competitive exchange economy. Draw the Edgeworth box. Measure consumption x in state α ('sunspots') on the horizontal and consumption in state β ('no sunspots') on the vertical. Endowments ω lie on the minor diagonal, because endowments are by definition independent of the state of nature. For the same reason, the Edgeworth box is a square. Assume that consumers possess smooth, strictly concave von Neumann–Morgensten utility functions. Competitive equilibrium exists. There are two cases: (1) Consumers share the same probability beliefs π about the occurrence of sunspots. Indifference curve tangency and hence competitive equilibrium occurs only on the minor diagonal with contingent claims prices p proportional to the probabilities π. Sunspots do not matter. This is an instance of the *Cass–Shell sunspot immunity theorem* (1983, Proposition 3). It holds when the box is square, that is, whenever there is no aggregate uncertainty (Figure 1).

Figure 1
Sunspots do not matter

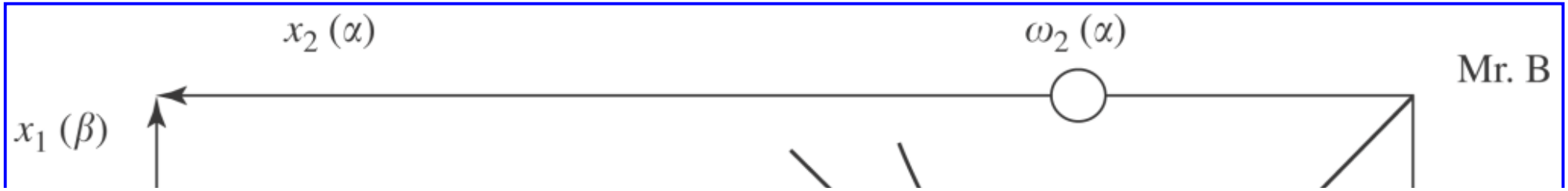

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

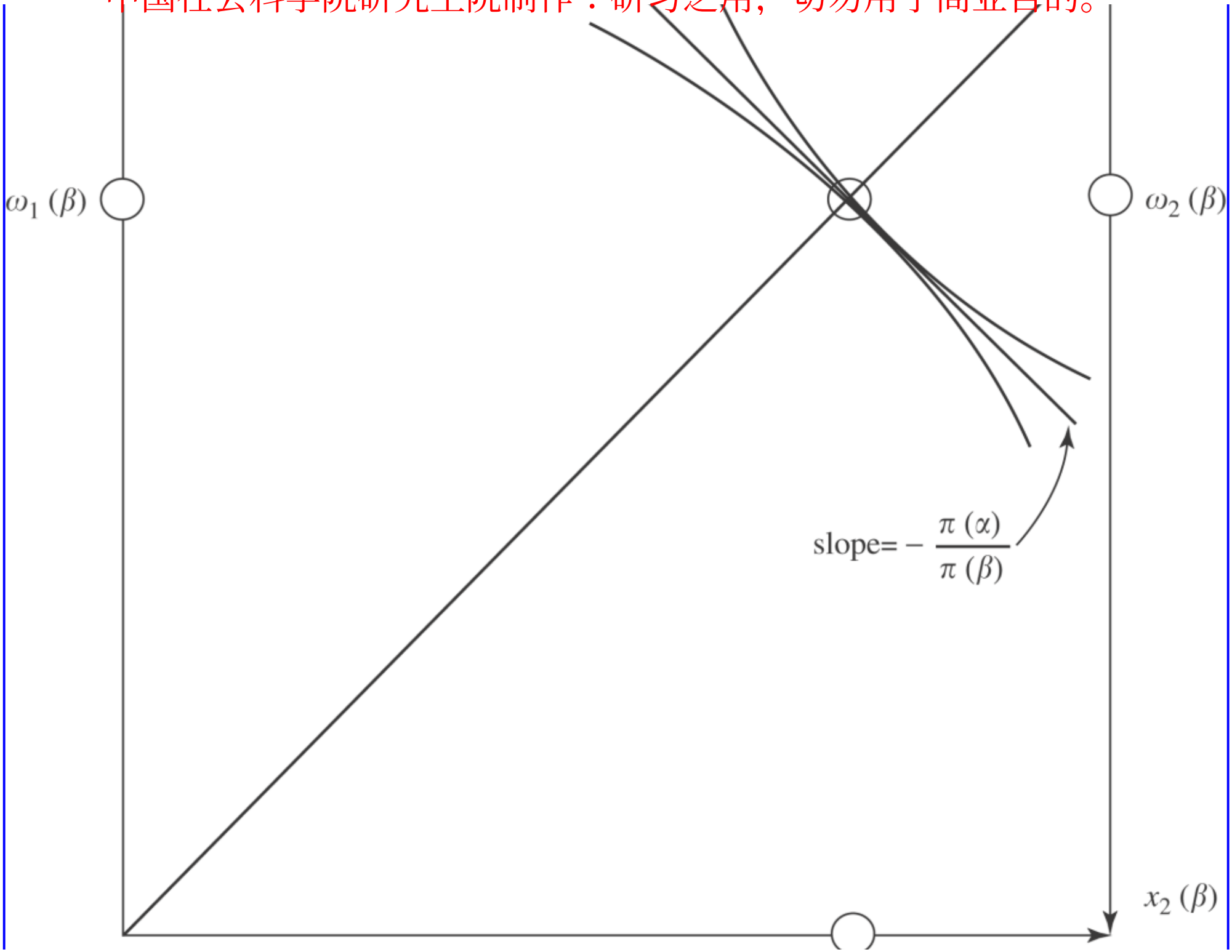

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

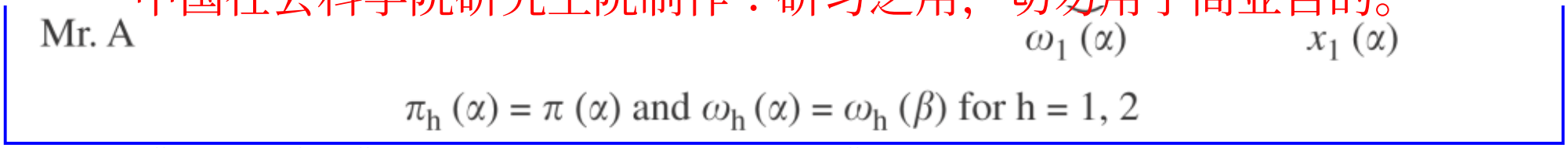

(2) Consumers differ in their beliefs. Indifference curves will be tangent to each other but always off the minor diagonal. Sunspots matter (Figure 2).

Figure 2

Sunspots matter

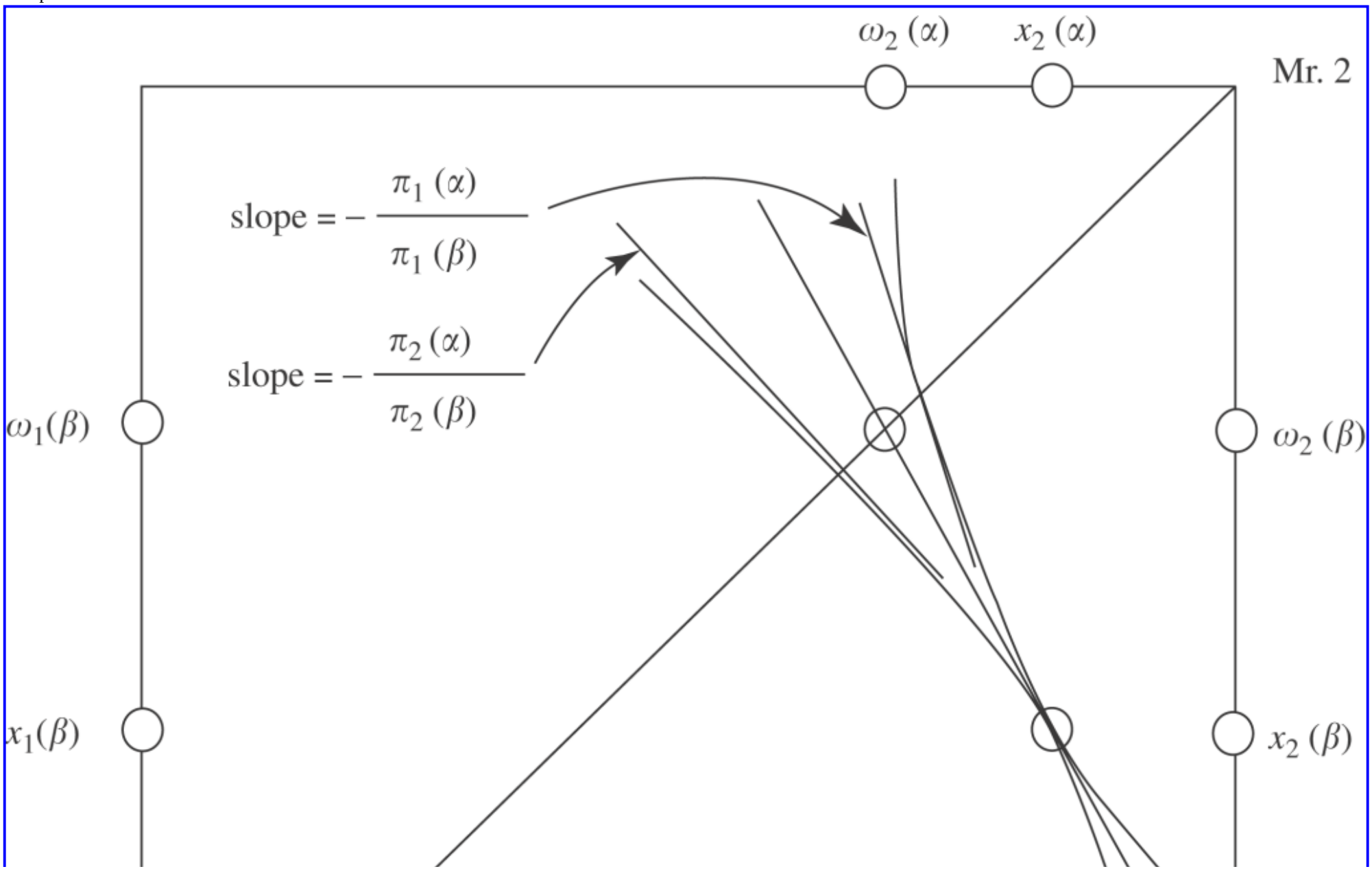

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

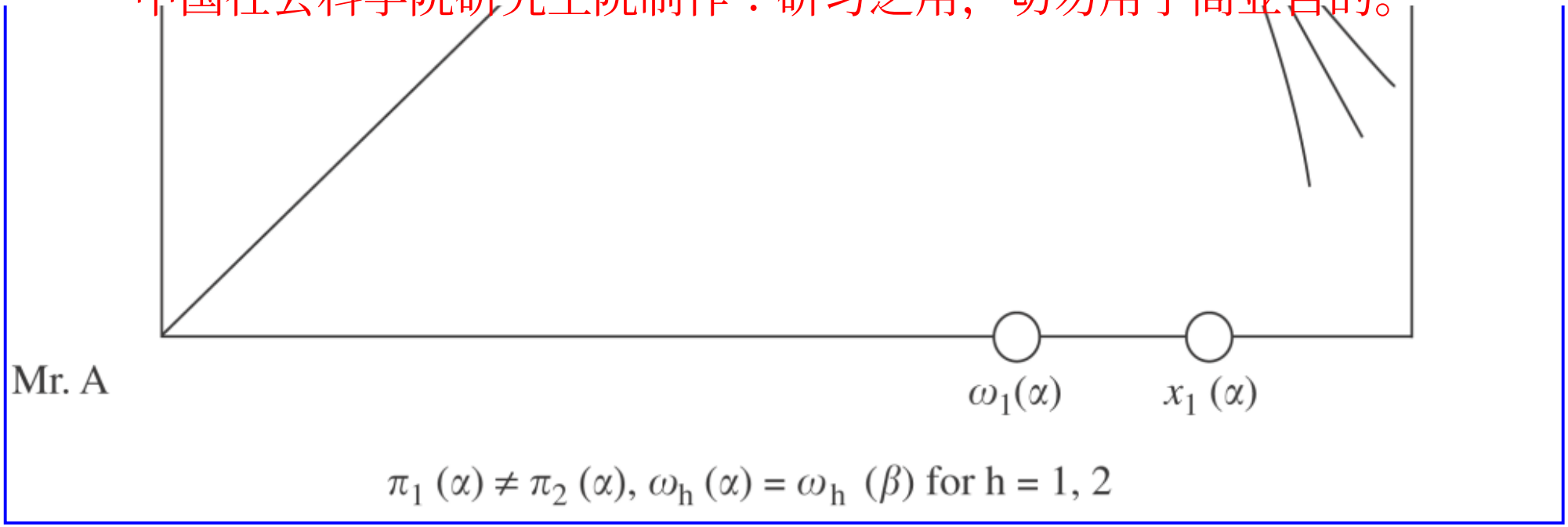

Sources of sunspot equilibria in OG economy

Heterogeneity of (probability) beliefs is a source of sunspot equilibria, but it is hardly the only source. The sunspots immunity theorem is based on a finite model with strictly convex preferences and convex production and a full range of perfectly competitive markets. Any departure from these assumptions could be a possible source of sunspots that matter. (See Shell, 1987, p. 550, for the so-called Philadelphia Pholk 'Theorem' on how to find sunspots that matter.) For example, the usual overlapping-generations models (including the one in Shell, 1977) fail to fit the assumptions of the immunity theorem in three ways.

1. 1. *There are restrictions on participation in the securities markets*. If a random variable is realized before your birth, you cannot buy securities dependent on its realization; see Cass and Shell (1983) for analysis of sunspot equilibria caused solely by restricted market participation. Balasko, Cass and Shell (1995) also focus on restrictions on market participation. If there are no (or sufficiently few) restricted agents, then sunspots do not matter in convex, finite economies. If all individuals are restricted, then sunspot equilibria are randomizations over non-sunspot equilibria. Otherwise, the typical sunspot equilibrium is not a mere lottery over non-sunspot equilibria.
2. 2. *The securities market is incomplete*. There is only one money. Completeness of the market would require instead state-contingent Arrow securities for each state of nature at each date. General equilibrium with incomplete markets, sometimes studied under the acronym GEI, is an important area in financial economics that was spawned by the sunspot-equilibrium model. Cass (1989; 1992), Balasko and Cass (1989), and others have played central roles in developing the GEI model and placing it in the sunspots-equilibrium literature. It is worth noting, however, that incomplete markets do not necessarily lead to sunspot equilibria; see, for example, Antinolfi and Keister (1998), who show that with only a few options (puts and calls) with the right strike prices the economy can be immune from sunspots even when there are many sunspot states of nature.
3. 3. *The OG model is not a finite model*. There are a countable number of individuals and a countable number of dated commodities. There can be sunspot equilibria in the OG economy even if we assume that markets are completed with Arrow securities for every state and every date, and that, contrary to actual biology and demography, agents are not restricted in the trades of these securities. In this thought experiment, they can even buy securities to hedge against events that occur before their natural lifetimes (see Cass and Shell, 1989). The unbounded horizon permits bubbles in the form of public debt that need not be retired. If a bubble is possible in an infinite-horizon economy, then it is likely that there can also be a stochastic (or sunspot) bubble. The infinite horizon is in itself a source of sunspot equilibria. Sunspots can be an imperfect substitute for fiat money in the 'Samuelson' case of the OG model (see Cass and Shell, 1989).

Non-convexities as a source of sunspot equibibria

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Each of these three departures from the finite, perfect-market competitive equilibrium economy is *in itself* a separate source of sunspot equilibrium allocations. So the OG model is a natural and – it turns out – a relatively easy place to find sunspots that matter. It is also natural to expect that non-convexities create a role for sunspot equilibria. Random allocations would seem to offer the possibility of at least partially 'convexifying' the certainty economy. This turns out to be the case. Shell and Wright (1993) analyse sunspot equilibrium in competitive, exchange economies with an indivisible good. They show that the Rogerson (1988) indivisible-labour lottery equilibrium can be decentralized as a sunspot equilibrium even in finite economies (as well as in continuum-of-agents economies). Unlike the situation in the finite, convex economy, (1) sunspot equilibrium allocations in these non-convex economies are Pareto optimal among stochastic allocations and often strictly dominate the best allocations available in the related economy that does not have access to randomization; (2) the certainty allocations do not necessarily reappear in the sunspots model as non-sunspot equilibrium allocations. Goenka and Shell (1997b) extend the Shell–Wright analysis to non-convex production (see also Goenka and Shell, 1997a). An earlier paper on sunspots in non-convex economies is Guesnerie and Laffont (1991). Indivisible labour and sunspots are central to a recent contribution to the theory of money and search (see Rocheteau et al., 2007). Previous monetary search models have required for tractability the apparently restrictive assumption that agents possess quasi-linear utility functions. If one assumes that labour is indivisible and is allocated to work or leisure by a sunspot process, then von Neumann–Morgenstern agents act as if their utilities are quasi-linear.

Lotteries and sunspots

What is the relationship between the sunspot-equilibrium concept and the lottery-equilibrium concept introduced by Prescott and Townsend (1984a; 1984b)? The original motivations for the two concepts were very different. The first sunspots papers focused on stochastic allocations that are Pareto non-optimal, cases where sunspots lead to inefficient allocations because of restrictions on market participation, incomplete markets, the infinite horizon, or imperfect competition. The first Prescott–Townsend lottery equilibrium papers focused on random allocations that partially remedy the effects of 'non-convexities' in the certainty economy due to moral hazard constraints. Because sunspots form the basis for coordination of individual plans, the sunspot equilibrium notion is directly applicable in economies with few agents, many agents, or even a continuum of agents. The original lottery equilibrium notion was applicable only to economies with a continuum of agents, in which detailed coordination is not necessary because of the law of large numbers. An important formal difference between these two stochastic equilibrium concepts is based on how commodities are defined. In the sunspots model, the commodity might be chocolate delivered in state α. In the lottery model, the commodity might be chocolate delivered with probability π. If the sunspot random variable used in each case is continuous (that is, has a non-atomic density function), then lottery equilibrium allocations are always sunspot equilibrium allocations (see Garratt et al., 2002). For lottery equilibria to make sense in finite economies (without the law of large numbers), the lottery equilibrium notion must be suitably adjusted as in Garratt (1995) to ensure that in equilibrium materials balance for every realization of the randomizing device. After making the Garratt adjustment, it is shown by Garratt et al. (2002) that for economies with a finite number of agents *or* a continuum of agents, and a continuous randomizing device, the set of sunspot equilibrium allocations is identical to the set of lottery equilibrium allocations. Garratt, Keister and Shell (2004) show that this equivalence does not always hold when the randomizing device is finite. Kehoe, Levine and Prescott (2002) establish the equivalence of sunspot equilibrium allocations and lottery equilibrium allocations in economies with a continuum of agents facing 'non-convexities' caused by moral hazard constraints. Prescott and Shell (2002) provide a review of the sunspot and lottery literatures and attempt to highlight the relatively strong similarities and the non-trivial differences between the two concepts.

Imperfect competition, correlated equilibria and sunspot equilibria

While the notion of sunspot equilibrium was not immediately accepted by macroeconomists, game theorists were not at all shocked by the idea of stochastic outcomes in non-stochastic environments. Think mixed strategy and – more generally – correlated equilibrium. Peck and Shell (1991) analyse in market games the relationship of sunspot equilibria to correlated equilibria defined by Aumann (1974; 1987). Peck and Shell show that every correlated equilibrium allocation can be decentralized as a sunspot equilibrium allocation, but the converse is not true. Correlated equilibria are self-enforcing while sunspot equilibria allow for transfer of incomes across states of nature. The market game is the leading general-equilibrium model of imperfect competition. It is shown by Peck and Shell that unless endowments are Pareto optimal, there is always a proper sunspot equilibrium due to imperfect competition. Imperfect and monopolistic competition are highly prone to sunspot effects. Imperfect competition is one of the very useful building blocks for calibrating applied sunspot models to business cycle data. Peck and Shell (1991) also incorporate asymmetric information into the sunspots model. Earlier examples of sunspot equilibria in which the randomizing device provides asymmetric (but correlated) information to agents are Azariadis and Guesnerie (1982) and Maskin and Tirole (1987).

Detailed market structure matters in imperfectly competitive economies. Peck and Shell (1989) construct two different securities games from the same certainty market game. In one, there is a full spectrum of Arrow financial securities. In the other, there is a full spectrum of real contingent commodities. A sunspot Nash equilibrium allocation in the Arrow securities game which is not a mere lottery over certainty Nash equilibrium allocations is *never* a Nash equilibrium allocation to the contingent-commodities game. The two games differ because the market power of individual agents depends on the way markets are organized.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Symmetry breaking

Yves Balasko (1983) provides a general definition of extrinsic uncertainty. Modelling how extrinsic uncertainty affects technologies and endowments is straightforward: endowments and input–output pairs are independent of the realization of the randomizing device. Modelling *ex ante* preferences is more subtle. Sunspots affect preferences only through allocations. Otherwise uncertainty would be intrinsic; the randomness would not be sunspot randomness. In particular, if the states of nature are renamed, say: α becomes β, the allocation in α becomes the allocation in β, and the probability $\pi(\alpha)$ becomes $\pi(\beta)$, then *ex ante* utility is unaffected. This generalizes von Neumann–Morgenstern utility. Balasko (1990) shows how sunspot equilibrium is an instance of symmetry-breaking in economics. The non-sunspot equilibrium is the symmetric solution to symmetric equations. The sunspot equilibrium breaks the symmetry of endowments, technologies and preferences. It is an asymmetric solution to the symmetric equations.

Sunspots and empirical business cycle research

Benhabib and Farmer (1994), Farmer and Guo (1994), and Galí (1994) launched the field of applied sunspot business cycle analysis (see also Benhabib and Farmer, 1996; Benhabib and Nishimura, 1998; Benhabib and Wen, 2004). They made only simple adjustments to the standard real business cycle (RBC) model of Kydland and Prescott (1982) in their set-ups. For example, Benhabib and Farmer (1994) – following the lead of Spear (1991) – introduced an externality in production, leading to aggregate increasing returns to scale. Without this adjustment, sunspots would not matter because the standard RBC model (based on a single infinite-lived individual) is equivalent to a planning model, so when preferences and technology are convex, sunspots cannot matter. With this adjustment, there can be a multiplicity of certainty equilibrium paths, which leads to the existence of equilibrium fluctuations driven by sunspots. Farmer and Guo (1994) calibrated a discrete-time version of the Benhabib–Farmer model to match business cycle facts while employing only sunspot uncertainty. Yi Wen (1998) replaced the Benhabib–Farmer externality with capacity utilization, reducing the size of the externality to a more reasonable level, while still matching business cycle facts without positing any intrinsic uncertainty. Calibration of sunspot-driven business cycles is a major and growing area. There is barely room in this review to scratch the surface. The applied sunspots business cycle calibrators were wise in deviating only in relatively small steps from the well-established RBC model for their experiments; otherwise they would have been less likely to get the attention of the calibration community. On the other hand, one guesses that sunspot allocations will be easier to find and easier to match to data in the overlapping-generations economy.

Sunspots, bank runs and economic fragility

Some sharp economic downturns have been attributed to ‘panics’ or ‘bursting bubbles’ in financial markets. People ‘run’ on a bank or other financial institution when they expect others to run. In their classic bank-run model, Diamond and Dybvig (1983) highlight the fragile nature of financial intermediaries. Banks attempt to smooth consumption between depositors who turn out to be patient (and can afford to wait) and those who turn out to be impatient (and need to withdraw early). The problem is that the ‘patient’ people might panic, attempting to withdraw early and causing a run on the bank. In the standard bank contract, there are two equilibria to the post-deposit game: (1) the (good) no-run equilibrium and (2) the (bad) run equilibrium. However, the run equilibrium is not an equilibrium for the pre-deposit game: if individuals know in advance that there will be a run on the bank, they will not make a deposit. Hence, in the formal model bank runs are not possible. Diamond and Dybvig suggest that sunspots will play a role in panic-based runs. Peck and Shell (2003; 2005) validate their intuition. It is shown that panic-based bank runs can be part of a sunspot equilibrium for the pre-deposit game even when partial suspension of convertibility is allowed. Sunspot-driven bank runs are possible, but they are typically not mere randomizations over certainty equilibria. If the probability of the run is small, the optimal banking contract tolerates runs. If the probability of the run is large, then the optimal banking contract is run-proof. Ennis and Keister (2003) exploit these ideas to investigate the implications of the possibility of bank runs on economic growth. Gu (2006) considers an asymmetric-information, extrinsic randomizing device in the banking set-up. If the sunspots signals are highly correlated, there exists a proper correlated equilibrium for some banking contracts. In the equilibrium, depending upon the realization of the signals, either a full bank run, or a partial bank run, or no bank run will occur.

What policies should be taken to stabilize or even immunize the economy from sunspot fluctuations? Complete immunization might not be feasible and when feasible it might not be desirable. For example, we know that while avoiding bank runs can be feasible, the optimal banking contract tolerates runs at small probabilities. See Ennis and Keister (2005) for some other examples. Grandmont (1985; 1986) designs government policies that immunize the economy from sunspot effects. Grandmont's policies set taxes according to feedback rules that render the current price level as predetermined and thus immunizing it from sunspots. Smith (1994) considers the policy of inflation-rate targeting in an overlapping-generations economy. If the government maintains a target price-level path by standing ready to exchange money for interest-bearing assets, this immunizes the economy from sunspots but at the cost of substantial inefficiency. Other policies that target a given price level lead to smaller inefficiencies even though they do not immunize the economy from

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

sunspot effects. Woodford (1994) considers a cash-in-advance economy in which the government either fixes the constant money growth rate or fixes a constant nominal interest rate. Woodford finds that there is a unique equilibrium under the interest rate rule, but the corresponding constant money growth economy is susceptible to sunspot shocks. Woodford (1994) differs from Smith (1994), because Woodford uses a transversality condition not appropriate in Smith's OG environment. For Woodford, under the interest rate rule all price histories but one result in too rapid accumulation of government debt. Keister (1998) studies a model with segmented asset markets in which the amount of sunspot-driven volatility in consumption depends on the government's tax-transfer policy. A policymaker concerned about inequality may choose to accept sunspot volatility in order to achieve some redistribution. It has been proposed that narrow banks, banks that are restricted to holding only liquid assets, are more stable than wide banks, banks that are unrestricted in asset holding. Peck and Shell (2005) show that narrow banks are subject to sunspot-based panic runs while wide banks are immune from these. On the other hand, wide banks are subject to running out of funds in the face of intrinsic shocks, while the narrow banks are immune from these shocks because of their over-investment in the liquid asset. Goenka (1994a) shows that restrictions on government institutions intended to increase bureaucratic accountability can also increase the fragility of the economy in the face of sunspot shocks. For example, forcing government agencies to finance their separate budgets through agency-specific taxes can introduce sunspot instability and inefficiency.

Stability of sunspot equilibria

Can sunspot equilibria be dismissed as less ‘stable’ or more ‘fragile’ than non-sunspot equilibria? The current answer to the question seems to be ‘no’, meaning sunspot equilibria cannot in general be dismissed. For example: (1) Woodford (1990) shows that under some plausible assumptions, the economy will learn to believe in sunspots. (2) Balasko, Cass and Shell (1995) show that if the parameters of the sunspot economy are slightly perturbed then sunspot equilibrium allocations will typically move to nearby sunspot equilibrium allocations. Of course, these stability results are only for specific models.

It has been possible to sketch only a few of the many excellent contributions to the very rich and extensive literature on sunspot equilibrium. Sunspots now play important roles in both descriptive and normative economics. They naturally arise in dynamic economies in which expectations play a central role. They matter when markets are incomplete or participation in them is restricted. They matter when the horizon is infinite. They matter when preferences and/or technologies are non-convex. … Sunspots matter.

See Also

- animal spirits
- calibration
- multiple equilibria in macroeconomics

Bibliography

Antinolfi, G. and Keister, T. 1998. Options and sunspots in a simple monetary economy. *Economic Theory* 11, 295–315.

Araújo, A.P. and Maldonado, W.L. 2000. Ergodic chaos, learning, and sunspot equilibria. *Economic Theory* 15, 163–84.

Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie* 11, 41–7, Colloques Internationaux du CNRS.

Arrow, K.J. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31, 91–6.

Aumann, R.J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96.

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55, 1–18.

Aumann, R.J., Peck, J. and Shell, K. 1988. Asymmetric information and sunspot equilibria: a family of simple examples. Working Paper No. 88-34, Center for Analytic Economics, Cornell University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25, 380–96.

Azariadis, C. and Guesnerie, R. 1982. Prophéties créatrices et persistence des théories. *Revue Economique* 33, 787–806.

Azariadis, C. and Guesnerie, R. 1986. Sunspots and cycles. *Review of Economic Studies* 58, 725–37.

Balasko, Y. 1983. Extrinsic uncertainty revisited. *Journal of Economic Theory* 31, 203–10.

Balasko, Y. 1990. Equivariant general equilibrium theory. *Journal of Economic Theory* 52, 18–44.

Balasko, Y. and Cass, D. 1989. The structure of financial equilibrium with exogenous yields: the case of incomplete markets. *Econometrica* 57, 135–62.

Balasko, Y., Cass, D. and Shell, K. 1995. Market participation and sunspot equilibria. *Review of Economic Studies* 62, 491–512. Repr. in *Equilibrium*, ed. D. Walker. London: Edward Elgar, 2000.

Barnett, W., Geweke, J. and Shell, K., eds. 1989. *Economic Complexity: Chaos, Sunspots, Bubbles, and Nonlinearity*. New York: Cambridge University Press.

Basak, S., Cass, D., Licari, J.M. and Pavlova, A. 2006. Multiplicity and sunspots in general financial equilibrium with portfolio constraints. *Mimeo, Penn Institute for Economic Research*.

Benhabib, J. and Farmer, R.E. 1994. Indeterminacy and increasing returns. *Journal of Economic Theory* 63, 19–41.

Benhabib, J. and Farmer, R.E. 1996. Indeterminacy and sector-specific externalities. *Journal of Monetary Economics* 37, 397–419.

Benhabib, J. and Nishimura, K. 1998. Indeterminacy and sunspots with constant returns. *Journal of Economic Theory* 81, 58–96.

Benhabib, J. and Wen, Y. 2004. Indeterminacy, aggregate demand and the real business cycle. *Journal of Monetary Economics* 51, 503–30.

Bhattacharya, J., Guzman, M. and Shell, K. 1998. Price level volatility: a simple model of money taxes and sunspots. *Journal of Economic Theory* 81, 401–30.

Billot, A., Chateauneuf, A., Gilboa, I. and Tallon, J.-M. 2000. Sharing beliefs: between agreeing and disagreeing. *Econometrica* 68, 685–94.

Cass, D. 1989. Sunspots and incomplete financial markets: the leading example. In *The Economics of Imperfect Competition and Employment: Joan Robinson and Beyond*, ed. G. Feiwel. London: Macmillan.

Cass, D. 1992. Sunspots and incomplete financial markets: the general case. *Economic Theory* 2, 341–58.

Cass, D. and Shell, K. 1980. In defense of a basic approach. In *Models of Monetary Economies*, ed. J. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Cass, D. and Shell, K. 1982. Les tâches solaires ont-elles de l'importance? *Cahiers du séminaire d'économétrie* 24, 93–127.

Cass, D. and Shell, K. 1983. Do sunspots matter? *Journal of Political Economy* 91, 193–227.

Cass, D. and Shell, K. 1989. Sunspot equilibrium in an overlapping-generations economy with an idealized contingent-commodities market. In Barnett, Geweke and Shell (1989).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Chatterji, S. and Chattopadhyay, S. 2006. Functional sunspot equilibria. *Journal of Mathematical Economics* 42, 22–35.

Chattopadhyay, S. 1996. Pareto optimal improvements for sunspots: the golden rule as a target for stabilization. *Economic Theory* 8, 123–35.

Chattopadhyay, S. and Gottardi, P. 1999. Stochastic OLG models, market structure, and optimality. *Journal of Economic Theory* 89, 21–67.

Chattopadhyay, S. and Muench, T. 1999. Sunspots and cycles reconsidered. *Economics Letters* 63, 67–75.

Chiappori, P.A. and Guesnerie, R. 1989. On stationary sunspot equilibria of order k. In Barnett, Geweke and Shell (1989).

Cozzi, G. and Giordani, P.E. 2006. Do sunspots matter under complete ignorance? *Research in Economics* 60, 148–54.

Dávila, J., Gottrdi, P. and Kajii, A. 2007. Local sunspot equilibria reconsidered. *Economic Theory* 31, 401–25.

Diamond, D. and Dybvig, P. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91, 401–19.

Dos Santas Ferreira, R. and Dufourt, F. 2006. Free entry and business cycles under the influence of animal spirits. *Journal of Monetary Economics* 53, 311–28.

Ennis, H.M. and Keister, T. 2003. Economic growth, liquidity, and bank runs. *Journal of Economic Theory* 109, 220–45.

Ennis, H.M. and Keister, T. 2005. Government policy and the probability of coordination failures. *European Economic Review* 49, 939–73.

Farmer, R.A. and Guo, J.-T. 1994. Real business cycle and the animal spirit hypothesis. *Journal of Economic Theory* 63, 42–73.

Farmer, R.A. and Woodford, M. 1997. Self-fulfilling prophecies and the business cycle. *Macroeconomic Dynamics* 1, 740–69.

Forges, F. and Peck, J. 1995. Correlated equilibrium and sunspot equilibrium. *Economic Theory* 5, 33–50.

Gale, D. 1973. Pure exchange equilibrium of dynamic economic models. *Journal of Economic Theory* 6, 12–36.

Galí, J. 1994. Monopolistic competition, business cycles, and the composition of aggregate demand. *Journal of Economic Theory* 63, 73–96.

Garratt, R. 1995. Decentralizing lottery allocations in markets with indivisible commodities. *Economic Theory* 5, 295–313.

Garratt, R. and Keister, T. 2002. A characterization of robust sunspot equilibria. *Journal of Economic Theory* 107, 136–44.

Garratt, R., Keister, T. and Shell, K. 2004. Comparing sunspot equilibrium and lottery equilibrium allocations: the finite case. *International Economic Review* 45, 351–86.

Garratt, R., Keister, T., Qin, C. and Shell, K. 2002. Equilibrium prices when the sunspot variable is continuous. *Journal of Economic Theory* 107, 11–38.

Goenka, A. 1994a. Fiscal rules and extrinsic uncertainty. *Economic Theory* 4, 401–16.

Goenka, A. 1994b. Rationing and sunspot equilibria. *Journal of Economic Theory* 62, 424–42.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Goenka, A. and Matta, S. 2007. Manipulation via endowments and sunspot equilibria. *Economic Theory* (online).

Goenka, A. and Prechac, C. 2006. Stabilizing sunspots. *Journal of Mathematical Economics* 42, 544–55.

Goenka, A. and Shell, K. 1997a. When sunspots don't matter. *Economic Theory* 9, 169–78.

Goenka, A. and Shell, K. 1997b. Robustness of sunspot equilibria. *Economic Theory* 10, 79–98.

Gottardi, P. and Kajii, A. 1999. The structure of sunspot equilibria: the role of multiplicity. *Review of Economic Studies* 66, 713–32.

Grandmont, J.-M. 1985. On endogeneous, competitive business cycles. *Econometrica* 53, 995–1046.

Grandmont, J.-M. 1986. Stabilizing competitive business cycles. *Journal of Economic Theory* 40, 57–76.

Gu, C. 2006. Asymmetric information and bank runs. Mimeo, Cornell University.

Guesnerie, R. 1986. Stationary sunspot equilibria in an *n*-commodity world. *Journal of Economic Theory* 40, 103–27.

Guesnerie, R. and Laffont, J.-J. 1991. Notes on sunspot equilibria in finite economies. In *Essays in Honor of E. Malinvaud*, ed. P. Champsaur. Cambridge, MA: MIT Press.

Guesnerie, R. and Woodford, M. 1992. Endogenous fluctuations. In *Advances in Economic Theory: Proceedings of the Sixth World Congress of the Econometric Society*, ed. J.-J. Laffont. Cambridge: Cambridge University Press.

Hansen, G. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16, 309–28.

Harrison, S. 2001. Indeterminacy in a model with sector-specific externalities. *Journal of Economic Dynamics and Control* 25, 747–64.

Hens, T. 2000. Do sunspots matter when spot market equilibria are unique? *Econometrica* 68, 435–41.

Hens, T. and Pilgrim, B. 2004. Sunspot equilibria and the transfer paradox. *Economic Theory* 24, 583–602.

Jackson, M. and Peck, J. 1991. Speculations and price fluctuations with private, extrinsic signals. *Journal of Economic Theory* 55, 274–95.

Jaimovich, N. 2007. Firm dynamics and markup variations: implications for sunspot equilibria and endogenous economic fluctuations. *Journal of Economic Theory* 137, 300–25.

Jevons, W.S. 1884. *Investigations in Currency and Finance*. London: Macmillan.

Kajii, A. 1998. Sunspots and the sequential regularity of competitive equilibria. *Journal of Economic Theory* 78, 187–94.

Kajii, A. 1997. On the role of options in sunspot equilibria. *Econometrica* 65, 977–86.

Kajii, A. and Ui, T. 2006. Agreeable bets with multiple priors. *Journal of Economic Theory* 128, 299–305.

Kajii, A. 2007. Welfare gains and losses in sunspot equilibria. *Japanese Economic Review* 58, 329–44.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kamihigashi, T. 1996. Real business cycles and sunspot fluctuations are observationally equivalent. *Journal of Monetary Economics* 37, 105–17.

Kehoe, T.J., Levine, D.K. and Prescott, E.C. 2002. Lotteries, sunspots, and incentive constraints. *Journal of Economic Theory* 107, 39–69.

Keister, T. 1998. Money taxes and efficiency when sunspots matter. *Journal of Economic Theory* 83, 43–68.

Kydland, F.E. and Prescott, E.C. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–1370.

Lucas, R.E., Jr. 1972. Expectations and the Neutrality of Money. *Journal of Economic Theory* 4, 103–24.

Manuelli, R. and Peck, J. 1990. Exchange rate volatility in an equilibrium asset pricing model. *International Economic Review* 31, 559–74.

Manuelli, R. and Peck, J. 1992. Sunspot-like effects of random endowments. *Journal of Economic Dynamics and Control* 16, 193–206.

Marimon, R., Spear, S.E. and Sunder, S. 1993. Expectationally driven market volatility: an experimental study. *Journal of Economic Theory* 61, 74–103.

Mas-Colell, A. 1992. Three observations on sunspots and asset redundancy. In *Economic Analysis of Markets and Games*, ed. P. Dasgupta, D. Gale, O. Hart and E. Maskin. Cambridge, MA: MIT Press.

Maskin, E. and Tirole, J. 1987. Correlated equilibria and sunspots. *Journal of Economic Theory* 43, 364–73.

Peck, J. 1988. On the existence of sunspot equilibria in an overlapping generations model. *Journal of Economic Theory* 44, 19–42.

Peck, J. and Shell, K. 1989. On the nonequivalence of the Arrow-securities game and the contingent-commodities game. In Barnett, Geweke and Shell (1989).

Peck, J. and Shell, K. 1991. Market uncertainty: correlated and sunspot equilibria in imperfectly competitive economies. *Review of Economic Studies* 58, 1011–29.

Peck, J. and Shell, K. 2003. Equilibrium bank runs. *Journal of Political Economy* 111, 103–23.

Peck, J. and Shell, K. 2005. Bank portfolio restrictions and equilibrium bank runs. Working Paper 99-07, Center for Analytic Economics, Cornell University, Ithaca, July 1999. Revised 25 February 2005.

Pietra, T. 1992. The structure of the set of sunspot equilibria in economies with incomplete financial markets. *Economic Theory* 2, 321–40.

Prescott, E.C. and Shell, K. 2002. Introduction to sunspots and lotteries. *Journal of Economic Theory* 107, 1–10.

Prescott, E.C. and Townsend, R.M. 1984a. General competitive analysis in an economy with private information. *International Economic Review* 25, 1–20.

Prescott, E.C. and Townsend, R.M. 1984b. Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica* 52, 21–46.

Rigotti, L. and Shannon, C. 2005. Uncertainty and risk in financial markets. *Econometrica* 73, 203–243.

Rocheteau, G., Rupert, P., Shell, K. and Wright, R. 2007. General equilibrium with nonconvexities money. *Journal of Economic Theory* (online).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Rogerson, R. 1988. Indivisible labor, lotteries and equilibrium. *Journal of Monetary Economics* 21, 3–16.

Schmitt-Grohé, S. 1997. Comparing four models of aggregate fluctuations due to self-fulfilling expectations. *Journal of Economic Theory* 72, 96–147.

Shell, K. 1971. Notes on the economics of infinity. *Journal of Political Economy* 79, 1002–11.

Shell, K. 1977. Monnaie et allocation intertemporelle. Communication to the Roy-Malinvaud seminar. Mimeo, Paris, November. (Title and abstract in French, text in English.)

Shell, K. 1987. Sunspot equilibrium. In *The New Palgrave: A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. New York: Macmillan. Repr. in *The New Palgrave: General Equilibrium*, ed. J. Eatwell, M. Milgate and P. Newman. New York: Macmillan, 1989.

Shell, K. and Wright, R. 1993. Indivisibilities, lotteries, and sunspot equilibria. *Economic Theory* 3, 1–17.

Shell, K. and Smith, B.D. 1992. Sunspot equilibrium. In *New Palgrave Dictionary of Money and Finance*, vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Smith, B.D. 1988. Legal restrictions, ‘sunspots’, and Peel's Bank Act: the real bills doctrine versus the quantity theory reconsidered. *Journal of Political Economy* 96, 3–19.

Smith, B.D. 1989. Legal restrictions, ‘sunspots’, and cycles. *Journal of Economic Theory* 47, 369–92.

Smith, B.D. 1991. Interest on reserves and sunspot equilibria: Friedman's proposal reconsidered. *Review of Economic Studies* 58, 93–105.

Smith, B.D. 1994. Efficiency and determinacy of equilibrium under inflation targeting. *Economic Theory* 4, 327–44.

Spear, S.E. 1984. Sufficient conditions for the existence of sunspot equilibria. *Journal of Economic Theory* 34, 360–70.

Spear, S.E. 1985. Rational expectations in the overlapping generations model. *Journal of Economic Theory* 35, 251–75.

Spear, S.E. 1989. Are sunspots necessary? *Journal of Political Economy* 97, 965–73.

Spear, S.E. 1991. Growth, externalities, and sunspots. *Journal of Economic Theory* 54, 215–23.

Spear, S.E., Srivastava, S. and Woodford, M. 1990. Indeterminacy of stationary equilibrium in stochastic overlapping generations models. *Journal of Economic Theory* 50, 265–84.

Strzalecki, T. 2006. Subjective beliefs and ex-ante agreeable trade. Working paper, Northwestern University.

Suda, S., Tallon, J.M. and Villanacci, A. 1992. Real indeterminacy of equilibria in a sunspot economy with inside money. *Economic Theory* 2, 309–19.

Tallon, J.-M. 1998. Do sunspots matter when agents are choquet-expected-utility maximisers? *Journal of Economic Dynamics and Control* 22, 357–368.

Tirole. 1985. Asset bubbles and overlapping generations. *Econometrica* 53, 1071–100.

Townsend, R.M. 1987. Arrow-Debreu programs as microfundations for macroeconomics. In *Advances in Economic Theory: Fifth World Congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Wang, P.F. and Wen, Y. 2006. Imperfect competition and sunspots. Working Paper No. 2006-015A, Federal Reserve Bank of St Louis.

Wang, P.F. and Wen, Y. 2007. Imperfect competition and indeterminacy of aggregate output. *Working Paper, Federal Reserve Bank of St. Louis*.

Wen, Y. 1998. Capacity utilization under increasing returns to scale. *Journal of Economic Theory* 81, 7–36.

Woodford, M. 1986. Stationary sunspot equilibria in a finance constrained economy. *Journal of Economic Theory* 40, 128–37.

Woodford, M. 1990. Learning to believe in sunspots. *Econometrica* 58, 227–307.

Woodford, M. 1994. Money policy and price level indeterminacy in a cash-in-advance economy. *Economic Theory* 4, 345–80.

Zilcha, I. 1991. Characterizing efficiency in stochastic overlapping generations models. *Journal of Economic Theory* 55, 1–16.

How to cite this article

Shell, Karl. "sunspot equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000325> doi:10.1057/9780230226203.1648

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

supermodularity and supermodular games

Xavier Vives
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The mathematical concept of supermodularity formalizes the idea of complementarity and opens the way for a rigorous treatment of monotone comparative statics and games with strategic complementarities. The approach is based on lattice methods and provides conditions under which optimal solutions to optimization problems change in a monotone way with a parameter. The theory of supermodular games exploits order properties to ensure that the best response of a player to the actions of rivals is increasing in their level. It yields strong results that apply to a wide range of games including dynamic games and games of incomplete information.

Keywords

assortative matching; Bertrand oligopoly; comparative statics; complementarity; correspondence principle; Cournot competition; existence of equilibrium; global games; incomplete information games; indivisibilities; Kakutani's fixed point th; Leontieff utility function; mixed strategy outcomes; monotone comparative statics; multiple equilibria; patent races; stochastic dominance; strategic complementarities; strategic substitutability; supermodular games; supermodularity; Tarski's fixed point th

Article

The concept of complementarity is well established in economics at least since Edgeworth (1881). The basic idea of complementarity is that the marginal value of an action increases with the level of other actions available. The mathematical concept of supermodularity formalizes the idea of complementarity. The theory of monotone comparative statics and supermodular games provides the toolbox to deal with complementarities. This theory, developed by Topkis (1978; 1979), Vives (1985; 1990) and Milgrom and Roberts (1990a), in contrast to classical convex analysis, is based on order and monotonicity properties on lattices (see Topkis, 1998; Vives, 1999; and Vives, 2005, for detailed accounts of the theory and applications). Monotone comparative statics analysis provides conditions under which optimal solutions to optimization problems change monotonically with a parameter. The theory of supermodular games exploits order properties to ensure that the best response of a player to the actions of rivals increases with their level. Indeed, this is the characteristic of games of strategic complementarities (the term was coined in Bulow, Geanakoplos and Klemperer, 1983). The power of the approach is that it clarifies the drivers of comparative statics results and the need of regularity conditions; it allows very general strategy spaces, including indivisibilities and functional spaces such as those arising in dynamic or Bayesian games; it establishes the existence of equilibrium in pure strategies (without requiring quasi-concavity of payoffs, smoothness assumptions, or interior solutions); it allows a global analysis of the equilibrium set when there are multiple equilibria, which has an order structure with largest and smallest elements; and, finally, it finds that those extremal equilibria have strong stability properties and there is an algorithm to compute them. We will provide an introduction to the approach and some definitions, move on to the basic monotone comparative statics results, and provide the basic results for supermodular games.

Preliminaries and definitions

A binary relation $\geq$ on a nonempty set X is a *partial order* if $\geq$ is reflexive, transitive, and antisymmetric (a binary relation is antisymmetric if $x \geq y$ and $y \geq x$ implies that $x=y$). A partially ordered set $(S, \geq)$ is *completely ordered* if for x and y in S either $x \geq y$ or $y \geq x$. An upper bound on a subset $A \subset X$ is $z \in X$ such that $z \geq x$ for all $x \in A$. A greatest element of A is an element of A that is also an upper bound on A. Lower bounds and least elements are defined analogously. The greatest and least elements of A, when they exist, are denoted max A and min A, respectively. A *supremum* (resp., *infimum*) of A is a least upper bound (resp., greatest lower bound); it is denoted sup A (resp., inf A). A *lattice* is a partially ordered set $(X, \geq)$ in which any

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

two elements have a *supremum* and an *infimum*. Any interval of the real line with the usual order is a lattice since any two points have a *supremum* and an *infimum* in the interval. However, the set in $\mathbb{R}^2$ {(1,0), (0,1)} is not a lattice with the vector ordering (the usual component-wise ordering), since (1,0) and (0,1) have no joint upper bound in the set. However, if we add the points (0,0) and (1,1) the set becomes a lattice with the vector ordering (see Figure 1 and let x=(0,1) and y=(1,0)). A lattice $(X,\geq)$ is *complete* if every non-empty subset has a *supremum* and an *infimum*. Any compact interval of the real line with the usual order, or product of compact intervals with the vector order, is a complete lattice. Open intervals are lattices but they are not complete (for example, the *supremum* of the interval (a, b) does not belong to (a, b)). A subset L of the lattice X is a *sublattice* of X if the *supremum* and *infimum* of any two elements of L belong also to L. A lattice is always a sublattice of itself, but a lattice need not be a sublattice of a larger lattice. Let $(X,\geq)$ and $(T,\geq)$ be partially ordered sets. A function $f:X\rightarrow T$ is *increasing* if, for x, y in X, $x\geq y$ implies that $f(x)\geq f(y)$.

Figure 1

Illustration of supermodular payoff on lattice. *Source*: Vives (1999, p. 25).

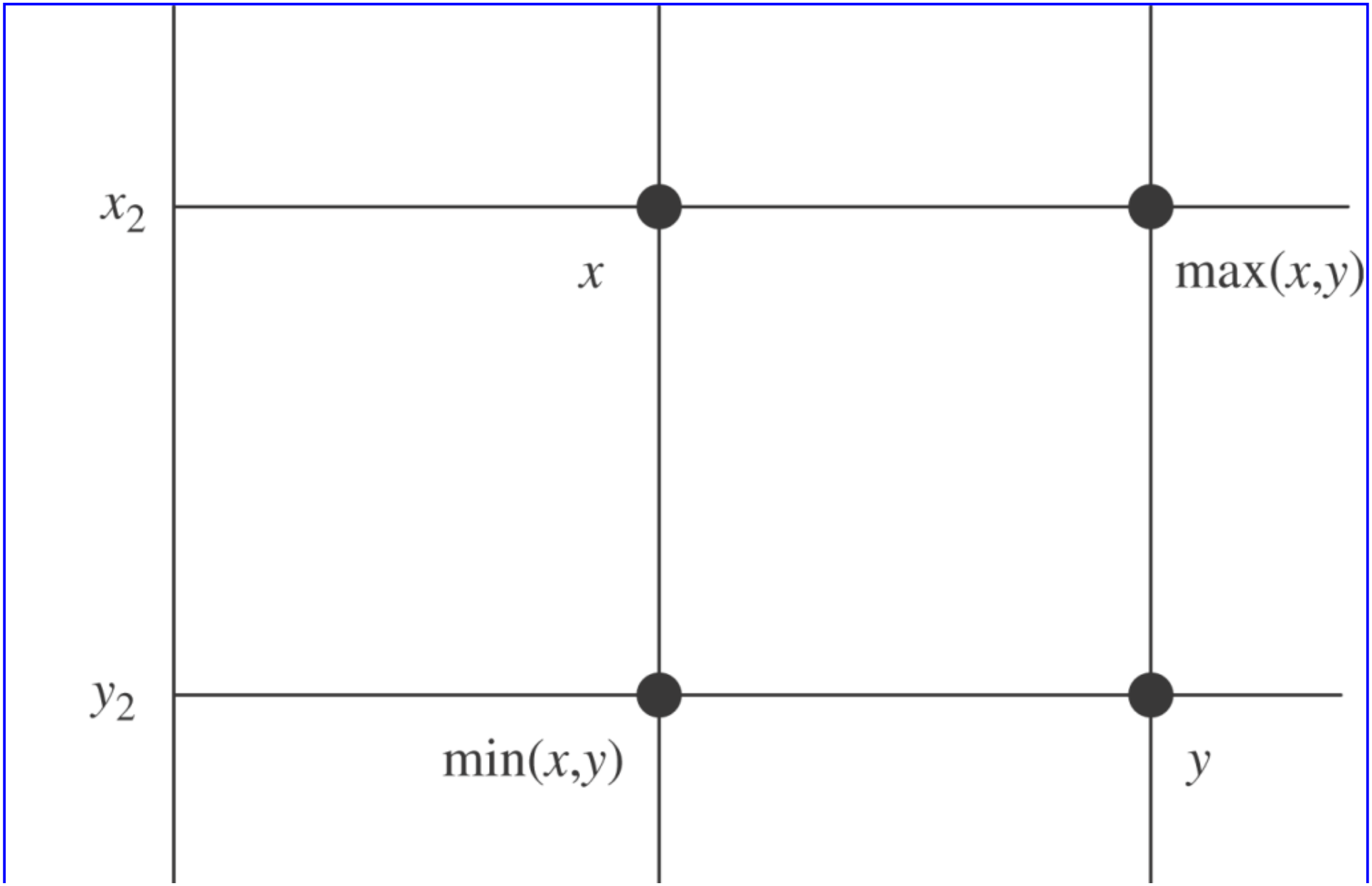

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

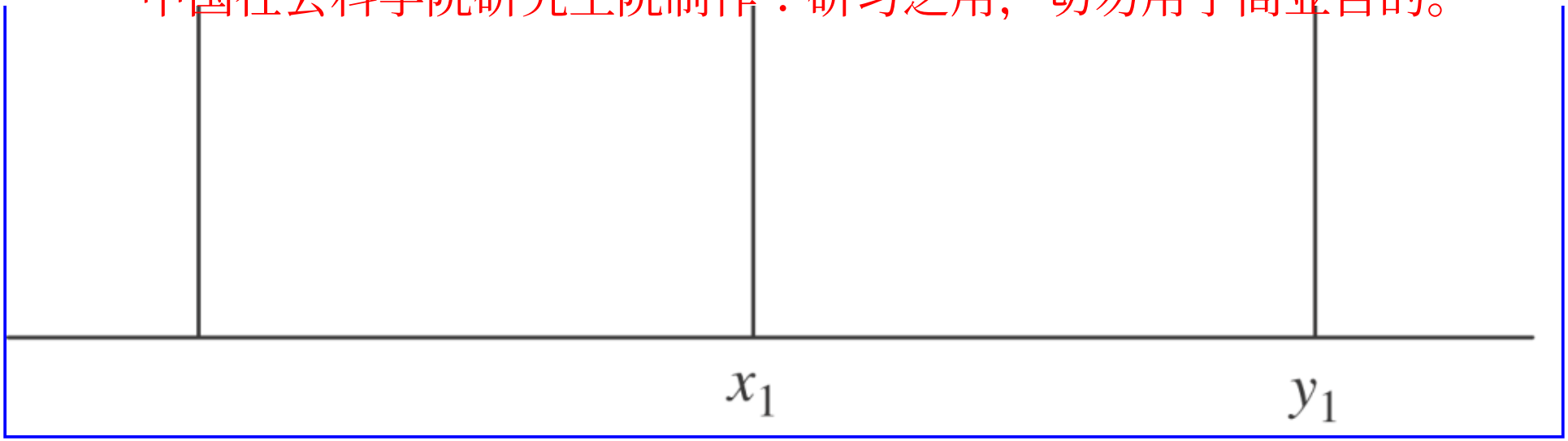

Supermodular functions

A function $g: X \to \mathbb{R}$ on a lattice X is *supermodular* if, for all x,y in X, $g(\inf(x, y)) + g(\sup(x, y)) \geq g(x) + g(y)$. It is *strictly supermodular* if the inequality is strict for all pairs x, y in X that neither $x \geq y$ nor $y \geq x$ holds. A function f is (*strictly*) *submodular* if $-f$ is (strictly) supermodular; a function f is (*strictly*) *log-supermodular* if log f is (strictly) supermodular. Let X be a lattice and T a partially ordered set. The function $g: X \times T \to R$ has (*strictly*) *increasing differences* in (x, t) if $g(x', t)-g(x, t)$ is (strictly) increasing in t for $x' > x$ or, equivalently, if $g(x, t') - g(x, t)$ is (strictly) increasing in x for $t' > t$. Decreasing differences are defined analogously.
Supermodularity is a stronger property than increasing differences: if T is also a lattice and if g is (strictly) supermodular on $X \times T$, then g has (strictly) increasing differences in (x, t). However, the two concepts coincide on the product of completely ordered sets: in such case a function is supermodular if and only if it has increasing differences in any pair of variables.
Both concepts formalize the idea of complementarity: increasing one variable raises the return to increase another variable. For example, the Leontieff utility function $U(x) = \min\{a_1x_1, \ldots, a_nx_n\}$ with $a_i \geq 0$ for all i is supermodular on $\mathbb{R}^n$. The complementarity idea can be made transparent by thinking of the rectangle in $\mathbb{R}^n$ with vertices $\{\min(x, y), y, \max(x, y), x\}$ and rewriting the definition of supermodularity as $g(\max(x, y))-g(x) \geq g(y) - g(\min(x, y))$. Consider, for example, points in $\mathbb{R}^2$ $x=(x_1, x_2)$ and $y=(y_1, y_2)$ with the usual order. Then going from $\min(x, y)=(x_1, y_2)$ to y, for given y_2, increases the payoff less than going from x to $\max(x, y)=(y_1, x_2)$, for given $x_2 \geq y_2$ (see Figure 1).
If X is a convex subset of $\mathbb{R}^n$ and if $g:X \to R$ is twice-continuously differentiable, then g has increasing differences in (x_i, x_j) if and only if $\partial^2 g(x) / \partial x_i \partial x_j \geq 0$ for all x and $i \neq j$. For decreasing differences (or submodularity) we would have $\partial^2 g(x) / \partial x_i \partial x_j \leq 0$. This characterization has a direct counterpart with the concept of (weak) cost complementarities if g is a cost function and $x \geq 0$ the production vector. If $\partial^2 g(x) / \partial x_i \partial x_j > 0$ for all x and $i \neq j$, then g is strictly supermodular. The differential characterization of supermodularity can be motivated by the figure as before. As an example consider assortative matching when types x and y in [0,1] produce $f(x, y)$ when matched and nothing otherwise. If $\partial^2 f / \partial x \partial y > 0$ then in a core allocation matching is positively assortative, that is, matched partners are identical (Becker, 1973; see Shimer and Smith, 2000, for a dynamic model with search where it is required also that $\log \partial f / \partial x$ and $\log \partial^2 f / \partial x \partial y$ are supermodular).
Positive linear combinations and pointwise limits preserve the complementarity properties (supermodularity/increasing differences) of a family of functions $g_n: X \times T \to \mathbb{R}$. Supermodularity is also preserved under integration. This has important consequences for comparative statics under uncertainty and games of incomplete information (see Vives, 1990; and Athey, 2001).
Supermodularity is also preserved under the maximization operation. Supermodularity is unrelated to convexity, concavity or returns to scale. Indeed, any real-valued function on a completely ordered set (say the reals) is both supermodular and submodular. This fact also makes clear that supermodularity in Euclidean spaces, in contrast to concavity or convexity, has no connection with continuity or differentiability properties. Note also that if g is a twice-continuously differentiable function, supermodularity only puts restrictions on the cross partials of g while the other concepts impose restrictions also on the diagonal of the matrix of second derivatives.

Monotone comparative statics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Let X be a compact rectangle in Euclidean space and let T be a partially ordered set. Let $g: X \times T \to \mathbb{R}$ be a function that (*a*) is supermodular and continuous on X for each $t \in T$ and (*b*) has increasing differences in (x, t). Let $\Phi(t) = \operatorname{argmax}_{x \in X} g(x, t)$. Then (Topkis, 1978):

1. 1. $\Phi(t)$ is a non-empty compact sublattice for all t;
2. 2. Φ is increasing in the sense that, for $t' > t$ and for $x' \in \Phi(t')$ and $x \in \Phi(t)$, we have $\sup(x', x) \in \Phi(t')$ and $\inf(x', x) \in \Phi(t)$; and
3. 3. $t \mapsto \sup \Phi(t)$ and $t \mapsto \inf \Phi(t)$ are increasing selections of Φ.

Several remarks are in order: (i) The continuity requirement of g can be relaxed. In more general spaces the requirement is for X to be a complete lattice and for g to fulfil an appropriate order continuity property. (ii) If g has strictly increasing differences in (x, t), then all selections of Φ are increasing. (iii) If solutions are interior, and $\partial g / \partial x_i$ is strictly increasing in t for some i, then all selections of Φ are strictly increasing (Edlin and Shannon, 1998). (iv) Milgrom and Shannon (1994) relax the complementarity conditions to ordinal complementarity conditions (quasi-supermodularity and a single crossing property), and develop necessary and sufficient conditions for monotone comparative statics.

Let us illustrate the result when $T \subset \mathbb{R}$ and g is twice-continuously differentiable on $X \times T$. Suppose first that $X \subset \mathbb{R}$, g is strictly quasi-concave in x (with $\partial g / \partial x = 0$ implying that $\partial^2 g / (\partial x)^2 < 0$), and that the solution to the maximization problem $\Phi(t)$ is interior. Then, using the implicit function theorem on the interior solution, for which $\partial g(\Phi(t), t) / \partial x = 0$, we obtain that Φ is continuously differentiable and $\Phi' = -\frac{\partial^2 g}{\partial x \partial t} / \frac{\partial^2 g}{(\partial x)^2}$. Obviously, sign Φ' =sign $\frac{\partial^2 g}{\partial x \partial t}$. The solution is increasing (decreasing) in t if there are increasing, $\frac{\partial^2 g}{\partial x \partial t} \geq 0$, (decreasing, $\frac{\partial^2 g}{\partial x \partial t} \leq 0$) differences. The monotone comparative statics result asserts that the solution $\Phi(t)$ will be monotone increasing in t even if g is not strictly quasi-concave in x, in which case $\Phi(t)$ need not be a singleton or convex-valued, provided that $\frac{\partial^2 g}{\partial x \partial t}$ does not change sign. For example, if we consider a single-product monopolist with revenue function $R(x)$ and cost function $C(x, t)$, where x is the output of the firm and t a cost efficiency parameter, we have $g(x, t) = R(x) - C(x, t)$. If $C(\cdot)$ is smooth and $\frac{\partial^2 C}{\partial x \partial t} \leq 0$, an increase in t reduces marginal costs. Then if $R(\cdot)$ is continuous the comparative static result applies, and the largest $\overline{\Phi}(t)$ and the smallest $\underline{\Phi}(t)$ monopoly outputs are increasing in t. If $\frac{\partial^2 C}{\partial x \partial t} < 0$, then all selections of the set of monopoly outputs are increasing in t. It is worth noting that the comparative statics result is obtained with no concavity assumption on the profit of the firm.

Suppose now that $X \subset \mathbb{R}^k$. If g is strictly concave in x (with the Jacobian of $\left(\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_k}\right)$ with respect to x, H_x, negative definite) and the solution to the optimization problem $\Phi(t) = (\Phi_1(t), \ldots, \Phi_k(t))$ is interior, then $\Phi(t)$ is continuously differentiable, and $\left(\frac{\partial \Phi_1}{\partial t}, \ldots, \frac{\partial \Phi_k}{\partial t}\right) = -H_x^{-1}\left(\frac{\partial^2 g}{\partial x_1 \partial t}, \ldots, \frac{\partial^2 g}{\partial x_k \partial t}\right)$. If the off-diagonal elements of H_x are nonnegative $\frac{\partial^2 g}{\partial x_i \partial x_j} \geq 0$, $j \neq i$, then all the elements of $-H_x^{-1}$ are nonnegative and the diagonal elements are positive (McKenzie, 1959). A sufficient condition for $\frac{\partial \Phi_i}{\partial t} \geq 0$ for all i is that $\frac{\partial^2 g}{\partial x_i \partial t} \geq 0$ for all i (the statement also holds with strict inequalities). As before, even if H_x is not negative definite, the assumptions that $\frac{\partial^2 g}{\partial x_i \partial x_j} \geq 0$, $j \neq i$, and that $\frac{\partial^2 g}{\partial x_i \partial t} \geq 0$ imply that the solution set $\Phi(t)$ has the monotonicity properties stated in the monotone comparative statics result. Note that when X is multidimensional, the restriction that g be supermodular on X, ensuring that for any components i and j an increase in the variable x_j raises the marginal return of variable x_i, when coupled with increasing differences on $X \times T$ is needed to guarantee the monotonicity of the solution. For example, let us consider a multiproduct monopolist. If the revenue function $R(\cdot)$ is continuous and supermodular on X, the cost function $C(\cdot)$ continuous and submodular on X, and $C(\cdot)$ displays decreasing differences in (x, t), the comparative static result follows. That is, the largest $\overline{\Phi}(t)$ and the smallest $\underline{\Phi}(t)$ monopoly output vectors are increasing in t. In the differentiable case, $\frac{\partial^2 R}{\partial x_i \partial x_j} \geq 0$ and $\frac{\partial^2 C}{\partial x_i \partial x_j} \leq 0$, for all $i \neq j$, and $\frac{\partial^2 C}{\partial x_i \partial t} \leq 0$ for all i. The result hinges on revenue and cost complementarities among outputs, and the impact of the efficiency parameter on marginal costs, and not concavity of profits.

Now let us consider a team problem. Suppose that n persons share a common objective $g(x_1, \ldots, x_n, t)$ where the action of player i x_i is in the rectangle $X_i \subset \mathbb{R}^{k_i}$ for each i and t is a payoff relevant parameter. If g is supermodular on $X = \prod_{i=1}^{n} X_i$ and has strictly increasing differences in (x, t), then any optimal solution is increasing in the level of the parameter. For example, the optimal production $g(x, t)$ of the firm (seen as a team problem) is increasing in the level of information technology t (which raises the marginal productivity of any worker of the firm).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Supermodular games

Consider the game (A_i, π_i; $i \in N$) where for each i=1, …, n in the set of players N, A_i is the strategy set, a subset of Euclidean space, and π_i the payoff of the player (defined on the cross product of the strategy spaces of the players A). Let $a_i \in A_i$ and $a_{-i} \in \prod_{j \neq i} A_j$ (that is, denote by a_{-i} the strategy profile (a_1, …, a_n) except the ith element). The strategy profiles are endowed with the usual component-wise order. We will say that the game (A_i, π_i; $i \in N$) is (*strictly*) *supermodular* if for each i, A_i is a compact rectangle of Euclidean space, π_i is continuous and (i) supermodular in a_i for fixed a_{-i} and (ii) displays (strictly) increasing differences in (a_i, a_{-i}). We will say that the game (A_i, π_i; $i \in N$) is *smooth* (*strictly*) *supermodular* if furthermore π_i (a_i, a_{-i}) is twice continuously differentiable with (i) $\partial^2 \pi_i / \partial a_{ih} \partial a_{ik} \geq 0$ for all $k \neq h$, and (ii) $\partial^2 \pi_i / \partial a_{ih} \partial a_{jk} \geq (>) 0$ for all $j \neq i$ and for all h and k, where a_{ih} denotes the hth component of the strategy a_i of player i. Condition (i) is the strategic complementarity property in own strategies a_i. Condition (ii) is the strategic complementarity property in rivals' strategies a_{-i} .

In a more general formulation strategy spaces need only be complete lattices and this includes functional spaces such as those arising in dynamic or incomplete information games. The complementarity conditions can be weakened to define an 'ordinal supermodular' game (see Milgrom and Shannon, 1994). Furthermore, the application of the theory can be extended by considering increasing transformations of the payoff (which do not change the equilibrium set of the game). For example, we will say that the game is log-supermodular if π_i is nonnegative and log π_i fulfils conditions (i) and (ii). This is the case of a Bertrand oligopoly with differentiated substitutable products, where each firm produces a different variety and marginal costs are constant, whenever the own-price elasticity of demand for firm i is decreasing in the prices of rivals, as with constant elasticity, logit, or constant expenditure demand systems.

In the duopoly case (n=2) the case of strategic substitutability can also be covered. Indeed, suppose that there is strategic complementarity or supermodularity, in own strategies ($\partial^2 \pi_i / \partial a_{ih} \partial a_{ik} \geq 0$ for all $k \neq h$, in the smooth version) and strategic substitutability in rivals' strategies or decreasing differences in (a_i, a_{-i}) ($\partial^2 \pi_i / \partial a_{ih} \partial a_{jk} \leq 0$ for all $j \neq i$ and for all h and k, in the smooth version). Then the game obtained by reversing the order in the strategy space of one of the players, say player 2, is supermodular (Vives, 1990). Cournot competition with substitutable products displays typically strategic substitutability between the (output) strategies of the firms.

In a supermodular game best responses are monotone increasing even when π_i is not quasi-concave in a_i. Indeed, in a supermodular game each player has a largest, $\overline{\Psi}_i(a_{-i})=\sup \Psi_i(a_{-i})$, and a smallest, $\underline{\Psi}_i(a_{-i})=\inf \Psi_i(a_{-i})$, best reply, and they are increasing in the strategies of the other players. Let $\overline{\Psi} = (\overline{\Psi}_1, \ldots, \overline{\Psi}_n)$ and $\underline{\Psi} = (\underline{\Psi}_1, \ldots, \underline{\Psi}_n)$ denote the extremal best reply maps.

Result 1: In a supermodular game there always *exist* a largest $\overline{a}=\sup \{a \in A: \overline{\Psi}(a) \geq a\}$ and a smallest $\underline{a}=\inf \{a \in A: \underline{\Psi}(a) \leq a\}$ equilibrium (Topkis, 1979).

The result is shown applying Tarski's fixed point theorem to the extremal selections of the best-reply map, $\overline{\Psi}$ and $\underline{\Psi}$, which are monotone because of the strategic complementarity assumptions. Tarski's theorem (1955) states that if A is a complete lattice (for example, a compact rectangle in Euclidean space) and $f: A \to A$ an increasing function then f has a largest $\sup\{a \in A: f(a) \geq a\}$ and a smallest $\inf\{a \in A: a \geq f(a)\}$ fixed point. There is no reliance on quasi-concave payoffs and convex strategy sets to deliver convex-valued best replies as required when showing existence using Kakutani's fixed point theorem. The equilibrium set can be shown also to be a complete lattice (Vives, 1990; Zhou, 1994). The result proves useful in a variety of circumstances to get around the existence problem highlighted by Roberts and Sonnenschein (1976). This is the case, for example, of the (log-)supermodular Bertrand oligopoly with differentiated substitutable products.

Result 2: In a *symmetric* supermodular game (that is, a game with payoffs and strategy sets exchangeable against permutations of the players) the extremal equilibria $\overline{a}$ and $\underline{a}$ are symmetric and, if strategy spaces are completely ordered and the game is strictly supermodular, then *only* symmetric equilibria exist (see Vives, 1999).

The result is useful to show uniqueness since if there is a unique symmetric equilibrium then the equilibrium is unique. For example, in a symmetric version of the Bertrand oligopoly model with constant elasticity of demand and constant marginal costs, it is easy to check that there exists a unique symmetric equilibrium. Since the game is (strictly) log-supermodular, we can conclude that the equilibrium is unique. The existence result of symmetric equilibria is related to the classical results of McManus (1962; 1964) and Roberts and Sonnenschein (1976).

Result 3: In a supermodular game if there are positive spillovers (that is, the payoff to a player is increasing in the strategies of the other players) then the largest (smallest) equilibrium point is the Pareto best (worst) equilibrium (Milgrom and Roberts, 1990a; Vives, 1990).

Indeed, in many games with strategic complementarities equilibria can be Pareto ranked. In the Bertrand oligopoly example the equilibrium with higher prices is Pareto dominant for the firms. This has proved particularly useful in applications in macroeconomics (for example, Cooper and John, 1988) and finance (for example, Diamond and Dybvig, 1983).

Result 4: In a supermodular game:

1. (*a*) Best-reply dynamics approach the interval [$\underline{a}$, $\overline{a}$] defined by the smallest and the largest equilibrium points of the game. Therefore, if the equilibrium is unique it is globally stable. Starting at any point A^+ (A^-) in the intersection of the upper (lower) contour sets of the largest (smallest) best replies of the players, best-reply dynamics lead monotonically downwards (upwards) to an equilibrium (Vives, 1990; 1999). This provides an iterative procedure to find the largest (smallest) equilibrium (Topkis, 1979) starting at sup A (inf A) (see Figure 2).
   Figure 2

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Best-reply dynamics in a supermodular game (with best-reply functions $r_1(\cdot)$, $r_2(\cdot)$). *Source*: Vives (1999, p. 54).

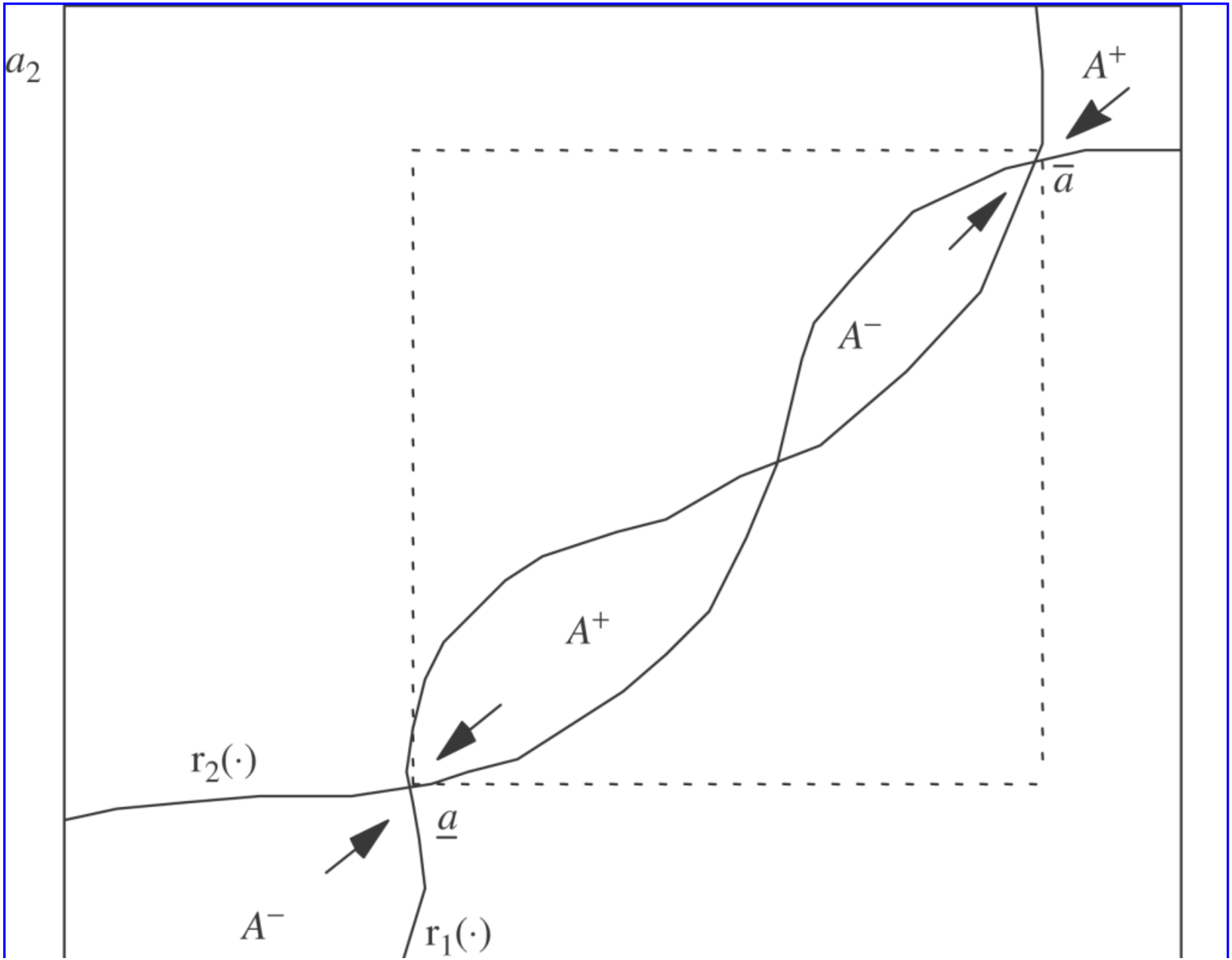

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

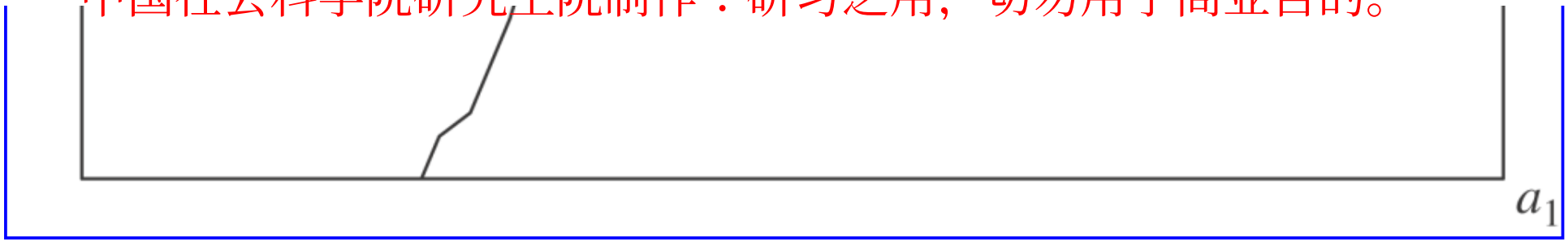

2. (*b*) The extremal equilibria correspond to the largest and smallest serially undominated strategies (Milgrom and Roberts, 1990a). Therefore, if the equilibrium is unique, the game is dominance solvable. Rationalizable (Bernheim, 1984; Pearce, 1984) or mixed strategy outcomes must lie in the interval $[\underline{a}, \overline{a}]$.

In the Bertrand oligopoly example with linear, constant elasticity, or logit demands, the equilibrium is unique and therefore it is globally stable, and the game is dominance solvable. In the team example it is clear that an optimal solution will be a Nash equilibrium of the game among team members. If the equilibrium is unique, then best-reply dynamics among team members will converge to the optimal solution. This need not be the case if there are multiple equilibria (see Milgrom and Roberts, 1990b, for an application to the theory of the firm).

Result 5: Let us consider a supermodular game with parameterized payoffs $\pi_i(a_i, a_{-i}; t)$ with t in a partially ordered set T. If $\pi_i(a_i, a_{-i}; t)$ has increasing differences in (a_i, t) (in the smooth version $\partial^2 \pi_i / \partial a_{ih} \partial t \geq 0$ for all h and i), then the largest and smallest equilibrium points increase with an increase in t, and starting from any equilibrium, best reply dynamics lead to a (weakly) larger equilibrium following the parameter change. The latter result can be extended to adaptive dynamics, which include fictitious play and gradient dynamics (see Lippman, Mamer and McCardle, 1987; and Sobel, 1988, for early results; and Milgrom and Roberts, 1990a; Milgrom and Shannon, 1994; and Vives, 1999, for extensions). It is worth noting that continuous equilibrium selections that do not increase monotonically with t predict unstable equilibria (Echenique, 2002). The result yields immediately that an increase in an excise tax in a (log-)supermodular Bertrand oligopoly raises prices at an extremal equilibrium.

The basic intuition for the comparative statics result is that an increase in the parameter increases the actions for one player, for given actions or rivals, and this reinforces the desire of all other players to increase their actions because of strategic complementarity. This initiates a mutually reinforcing process that leads to larger equilibrium actions. This is a typical positive feedback in games of strategic complementarities. In this class of games, unambiguous monotone comparative statics obtain if we concentrate on stable equilibria. We can understand this as a multidimensional version of Samuelson's (1947) correspondence principle, which was obtained with standard calculus methods applied to interior and stable one-dimensional models.

A patent race

Let us consider n firms engaged in a memory-less patent race that have access to the same R&D technology. The winner of the patent obtains the prize V and losers obtain nothing. The (instantaneous) probability of innovating is given by $h(x)$ if a firm spends x continuously, where h is a smooth function with $h(0)=0$, $h' > 0$, $\lim_{x \to \infty} h'(x) = 0$, and $h'(0) = \infty$. It is assumed also that h is concave but a region of increasing returns for small x may be allowed. If no patent is obtained the (normalized) profit of a firm is zero. The expected discounted profits (at rate r) of firm i investing x_i if rival $j \neq i$ invests x_j is given by

$$\pi_i = \frac{h(x_i)\, V - x_i}{h(x_i) + \Sigma_{j \neq i} h(x_j) + r}.$$

Lee and Wilde (1980) restrict attention to symmetric Nash equilibria of the game and show that, under a uniqueness and stability condition at a symmetric equilibrium x^* expenditure intensity increases with n. The classical approach requires assumptions to ensure a unique and stable symmetric equilibrium and cannot rule out the existence of asymmetric equilibria. Suppose that there are potentially multiple symmetric equilibria and that going from n to n+1 new equilibria appear. What comparative static result can we infer then? Using the lattice approach we obtain a more general comparative statics result that allows for the presence of multiple symmetric equilibria (Vives, 1999, Exercise 2.20; and 2005, Section 5.2). Let $h(0)=0$ with h strictly increasing in $[0, \overline{x}]$, with $h(x)\, V - x < 0$ for $x \geq \overline{x} > 0$. Under the assumptions the game is strictly log-supermodular and from Result 2 only symmetric equilibria exist. Let $x_i = x$ and $x_j = y$ for $j \neq i$. Then $\log \pi_i$ has (strictly) increasing differences in (x, n) for all y (y>0) and, according to Result 5, the expenditure intensity x^* at extremal equilibria is increasing in n.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Furthermore, starting at any equilibrium, an increase in n will raise research expenditure with out-of-equilibrium adjustment according to best-reply dynamics. This will be so even if new equilibria appear, or some disappear, as a result of increasing n. Finally, if h is smooth with $h' > 0$ and $h'(0) = \infty$, then $\partial \log \pi_i / \partial x_i$ is strictly increasing in n and (at extremal equilibria) x^* is *strictly* increasing in n. This follows because, under our assumptions, equilibria are interior and must fulfil the first-order conditions.

The results can be applied to dynamic and incomplete information games, which have complex strategy spaces. For example, in an incomplete information game, if for given types of players the *ex* post game is supermodular, then the Bayesian game is also supermodular and therefore there exist Bayesian equilibria in pure strategies (Vives, 1990). If, furthermore, payoffs to any player have increasing differences between the actions of the player and types, and higher types believe that other players are also of a higher type (according to first-order stochastic dominance), then extremal equilibria of the Bayesian game are monotone increasing in types (Van Zandt and Vives, 2007). This defines a class of monotone supermodular games. An example is provided by global games, introduced by Carlsson and Van Damme (1993) and developed by Morris and Shin (2002) and others with the aim of equilibrium selection. Global games are games of incomplete information with type space determined by each player observing a noisy private signal of the underlying state. The result is obtained applying iterated elimination of strictly dominated strategies. From the perspective of monotone supermodular games we know that extremal equilibria are the outcome of iterated elimination of strictly dominated strategies, that they are monotone in type (and therefore in binary action games there is no loss of generality in restricting attention to threshold strategies), and the conditions put to pin down a unique equilibrium in the global game amount to a lessening of the strength of strategic complementarities (see Vives, 2005, Section 7.2).

See Also

- epistemic game theory: incomplete information
- game theory
- global games.

Bibliography

Athey, S. 2001. Single crossing properties and the existence of pure strategy equilibria in games of incomplete information. *Econometrica* 69(4), 861–89.

Becker, G. 1973. A theory of marriage: part I. *Journal of Political Economy* 81, 813–46.

Bernheim, B.D. 1984. Rationalizable strategic behavior. *Econometrica* 52, 1007–28.

Bulow, J., Geanakoplos, J. and Klemperer, P. 1985. Multimarket oligopoly: strategic substitutes and complements. *Journal of Political Economy* 93, 488–511.

Carlsson, H. and van Damme, E. 1993. Global games and equilibrium selection. *Econometrica* 61, 989–1018.

Cooper, R. and John, A. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103, 441–63.

Diamond, D.W. and Dybvig, P.H. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91, 401–19.

Echenique, F. 2002. Comparative statics by adaptive dynamics and the correspondence principle. *Econometrica* 70, 833–44.

Edlin, A.S. and Shannon, C. 1998. Strict monotonicity in comparative statics. *Journal of Economic Theory* 81, 201–19.

Lee, T. and Wilde, L. 1980. Market structure and innovation: a reformulation. *Quarterly Journal of Economics* 94, 429–36.

Lippman, S., Mamer, J. and McCardle, K. 1987. Comparative statics in non-cooperative games via transfinitely iterated play. *Journal of Economic Theory* 41, 140–66.

McKenzie, L. 1959. Matrices with dominant diagonals and economic theory. In *Mathematical Methods in the Social Sciences*, ed. K. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

McManus, M. 1962. Numbers and size in Cournot oligopoly. *Yorkshire Bulletin of Economic and Social Research* 14, 14–22.

McManus, M. 1964. Equilibrium, numbers and size in Cournot oligopoly. *Yorkshire Bulletin of Economic and Social Research* 16, 68–75.

Milgrom, P. and Roberts, J. 1990a. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica* 58, 1255–77.

Milgrom, P. and Roberts, J. 1990b. The economics of modern manufacturing: technology, strategy, and organization. *American Economic Review* 80, 511–28.

Milgrom, P. and Shannon, C. 1994. Monotone comparative statics. *Econometrica* 62, 157–80.

Morris, S. and Shin, H.S. 2002. Global games: theory and applications: advances in economics and econometrics. In *Proceedings of the Eighth World Congress of the Econometric Society*, ed. M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge: Cambridge University Press.

Pearce, D. 1984. Rationalizable strategy behavior and the problem of perfection. *Econometrica* 52, 1029–50.

Roberts, J. and Sonnenschein, H. 1976. On the existence of Cournot equilibrium without concave profit functions. *Journal of Economic Theory* 13, 112–17.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Shimer, R. and Smith, L. 2000. Assortative matching and search. *Econometrica* 68, 343–70.

Sobel, M. 1988. Isotone comparative statics in supermodular games. Mimeo, State University of New York.

Tarski, A. 1955. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics* 5, 285–309.

Topkis, D. 1978. Minimizing a submodular function on a lattice. *Operations Research* 26, 305–21.

Topkis, D. 1979. Equilibrium points in non-zero sum n-person submodular games. *SIAM Journal of Control and Optimization* 17, 773–87.

Topkis, D. 1998. *Supermodularity and Complementarity*. Princeton: Princeton University Press.

Van Zandt, T. and Vives, X. 2007. Monotone equilibria in Bayesian games of strategic complementarities. *Journal of Economic Theory* 134, 339–50.

Vives, X. 1985. Nash equilibrium in oligopoly games with monotone best responses. Working Paper No. 85-10, CARESS, Unibversity of Pennsylvania.

Vives, X. 1990. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics* 19, 305–21.

Vives, X. 1999. *Oligopoly Pricing: Old Ideas and New Tools*. Boston: MIT Press.

Vives, X. 2005. Complementarities and games: new developments. *Journal of Economic Literature* 43, 437–79.

Zhou, L. 1994. The set of Nash equilibria of a supermodular game is a complete lattice. *Games and Economic Behavior* 7, 295–300.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Vives, Xavier. "supermodularity and supermodular games." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000486> doi:10.1057/9780230226203.1649

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

superneutrality

Jean-Pierre Danthine
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Money is said to be superneutral – or long-run neutral – if changes in the steady-state rate of growth of the money supply do not affect the value of real economic variables. Superneutrality depends on the hypothesis that the marginal productivity of capital is not affected by the level or the growth rate of money balances. It may hold true in steady state equilibria where the marginal utility of consumption is constant over time. Qualitatively, superneutrality is a fragile, knife-edged result that fails in a variety of contexts. Empirically, however, it is not clear that deviations from superneutrality are quantitatively significant.

Keywords

capital accumulation; imperfect competition; inflation; labour-leisure choices; leisure; menu costs; money supply; natural rate of unemployment; neoclassical growth theory; neutrality of money; new Keynesian macroeconomics; optimal growth model; Phillips curve; real business cycles; structural vector autoregressions; superneutrality; uncertainty; vector autoregressions

Article

Money is said to be neutral when a change in the quantity of money affects only the level of prices, and not the level of output or of other real variables. By extension, money is said to be superneutral – or long-run neutral – if changes in the steady-state rate of growth of the money supply do not affect the value of real economic variables (with the exception of money balances, whose steady-state real value usually decreases as the rate of inflation, and thus the cost of holding money balances, increases).
Tobin (1965) may be viewed as having sparked the modern discussion of this issue. In a growth model with a constant savings ratio, he showed that an increased inflation rate, by lowering the return to money, encourages portfolio holders to increase their demand for real assets at the expense of money holdings, and thereby displaces the equilibrium time path of the economy to one of higher per capita

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
output and capital.

Casting the same question in the context of an optimal growth model, Sidrauski (1967) showed that when the savings rate becomes endogenous the superneutrality result obtains: the steady-state values of per capita capital stock and output are unaffected by the growth rate of the money supply.

Sidrausky's model is one in which the representative, infinitely lived agent maximizes the discounted sum of future utilities arising from consumption and real money balances. She may hold two assets, money balances depreciating at the inflation rate π and real capital *k* depreciating at a rate δ. At any date *t*, the first-order condition determining the level of investment is of the type

$$MU_t = \beta MU_{t+1}[MP_{t+1} + (1 - \delta)], \tag{1}$$

where MU_t refers to the marginal utility of consumption at time *t*, β is the discount factor and MP_{t+1} is the marginal productivity of capital. The LHS of eq. (1) is the utility value of sacrificing one unit of consumption in any period *t*, while the RHS is the gain in utility from transforming that unit of commodity into capital and producing MP_{t+1} units of the good with a depreciation rate δ. In an economy with identical agents, this first-order condition for the representative individual's problem also defines the equilibrium values of the corresponding aggregates, consumption, real balances (entering the *MU*), capital stock and possibly other variables (entering the *MP*). Since, by definition, *MU* is constant in the steady state, it drops out from both sides of the equation and the steady-state capital stock level k^* must satisfy

$$\beta[MP(k^*, \cdot) + (1 - \delta)] = 1. \tag{2}$$

If, as in Sidrauski's model, the marginal product of capital depends only on the capital stock, eq. (2) uniquely determines the steady-state level of capital *independently of monetary considerations*: hence the superneutrality results. It is then easy to show that real money balances are inversely related with the rate of money creation in the steady state.

It is clear from eq. (2) that Sidrauski's superneutrality result is relatively fragile to changes in the hypotheses or structure of the model. We give a few illustrations below for models which are close in spirit to Sidrauski's original model. Further departures from the neoclassical growth framework, for instance in the directions illustrated by the New Keynesian literature, are bound to yield the same conclusion.

First of all, superneutrality depends on the hypothesis that the MP of capital is not affected by the level or the growth rate of money balances. Suppose instead that money enters a production function with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
standard properties (money balances acting as firms' working capital), then higher inflation, by penalizing money holding and thus decreasing steady-state money balances, diminishes the productivity (or the return) to the complementary input, capital. Money is not superneutral but the effect of a higher rate of money creation is opposite to the one suggested in the portfolio view of Tobin, in which money and capital are substitutable assets.

In the same vein, the real business cycle literature has introduced labour-leisure choices as an important additional element of the neoclassical growth model. In a model where money holding is motivated by a cash-in-advance constraint, Cooley and Hansen (1989) show that an increase in steady-state inflation leads individuals to shift away from activities that require cash, such as consumption, for activities that do not require cash, such as leisure. Labour supply thus shifts leftward at higher inflation rates and output falls. Thus, superneutrality fails to hold in this context as well with the opposite of the Tobin effect obtaining.

The derivation of eq. (2), and hence superneutrality, also hinges on the fact that the MU has to be constant in the steady state of a model with infinitely lived agents. The 'Tobin effect' reappears if one breaks away from this property, as shown by Fisher (1979), who looks at the effect of money growth on capital accumulation in the transition to a new state, by Weiss (1980) who casts the problem in an overlapping generations framework, and by Danthine, Donaldson and Smith (1987) who introduce uncertainty, and thus a richer definition of the steady state, into Sidrauski's model. Using numerical techniques, however, the last named authors are able to quantify such an effect of money on growth and reveal that this effect appears to be quantitatively small. In particular, it is dwarfed by the impact of a change in money growth on real balances and prices. They conclude that, while in a strict sense superneutrality is violated in their setting, it holds as a reasonable approximation.

The New Keynesian literature – for example, Goodfriend and King (1997) – is the latest development relevant for the superneutrality issue. New Keynesian models differ from the optimal growth model discussed above on two grounds: they move away from perfect competition, introducing market power by monopolistically competitive producers; they additionally feature nominal rigidities such as 'menu costs' preventing prices from adjusting instantaneously to nominal changes. Such rigidities imply the possibility of a trade-off between inflation, and real variables and quite naturally superneutrality fails to hold in this context. The basic result is that the distortions arising from menu costs are minimized if prices are stable, that is, at a zero rate of inflation. Positive inflation rates entail output (and welfare) losses as relative prices are distorted by the inability of a fraction of producers to adjust their prices at the time of the inflationary shock. There is a countervailing effect, however: the mark-up set by monopolistic producers is affected by the presence of inflation. Early Keynesian analyses had recognized that inflation would erode the market power of firms, and indeed imperfect competition includes a bias in favour of inflation. Goodfriend and King, however, show this bias to be quantitatively very small so that the output maximizing rate of inflation is very close to zero.

Further away from a Walrasian perspective on labour markets, the superneutrality issue is related to the question of the slope of the long-run Phillips curve. Thus Phelps's natural rate of unemployment hypothesis (1967) can be developed into an argument in favour of superneutrality. And, at the empirical level, recent attempts at rediscovering a long run trade-off between inflation and unemployment can be reinterpreted in the same vein. Notable is King and Watson (1994), who use VAR methodology with various identification hypotheses on United States data to conclude that there is evidence of a long-run

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
negative trade-off between unemployment and inflation, although the slope of the trade-off has declined since 1972. The authors perform causality tests suggesting that changes in unemployment are causing changes in inflation, a result that hints at a reality where lower unemployment places workers in a position to extract higher wages. Modelling such a reality requires a richer perspective on the labour market than the one adopted in the literature reviewed so far.

The very large empirical literature concentrating on the inflation and growth nexus is also relevant here. One of the long-run monetary facts of McCandless and Weber (1995) is that the rate of inflation and the growth rate of output are essentially uncorrelated. This is somewhat different from much of what is reported elsewhere in the literature, where typically a small negative correlation between the inflation rate and the growth rate of output is uncovered. Barro (1996), however, stresses that 'the clear evidence for adverse effects of inflation comes from the experience of high inflation'. McCandless and Weber argue that one high-inflation outlier is responsible for the standard result.

More directly focusing on superneutrality, Bullard and Keating (1995) use a structural VAR on long-run data from 16 countries. Their general conclusion is that 'superneutrality describes well most of the postwar economies we study'.

See Also

- neutrality of money
- optimum quantity of money

Bibliography

Akerlof, G.A., Dickens, W.T. and Perry, G.L. 1996. The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1996(1), 1–59.

Barro, R.J. 1996. Inflation and growth. *Federal Reserve Bank of St. Louis Review* 78(3), 153–69.

Bullard, J. and Keating, J. 1995. Superneutrality in postwar economies. *Journal of Monetary Economics* 36, 477–96.

Cooley, T.F. and Hansen, G.D. 1989. The inflation tax in a real business cycle model. *American Economic Review* 79, 733–47.

Danthine, J-P., Donaldson, J. and Smith, L. 1987. On the superneutrality of money in a stochastic dynamic macroeconomic model. *Journal of Monetary Economics* 20, 475–99.

Fisher, S. 1979. Capital accumulation on the transition path in a monetary optimizing model. *Econometrica* 47, 1433–9.

Goodfriend, M. and King, R.G. 1997. The new neoclassical synthesis and the role of monetary policy. In

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

NBER Macroeconomics Annual, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

King, R.G. and Watson, M.W. 1994. The post-war U.S. Phillips curve: a revisionist econometric history. *Carnegie-Rochester Conference Series on Public Policy* 41, 157–219.

McCandless, G.T., Jr. and Weber, W.E. 1995. Some monetary facts. *Federal Reserve Bank of Minneapolis Quarterly Review* 19(3), 2–11.

Phelps, E. 1967. Phillips curves, expectations of inflation and optimal unemployment over time. *Economica* 34, 254–81.

Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review, Proceedings* 57, 534–44.

Tobin, J. 1965. Money and economic growth. *Econometrica* 33, 671–84.

Weiss, L. 1980. The effect of money supply on economic welfare in the steady state. *Econometrica* 48, 565–76.

How to cite this article

Danthine, Jean-Pierre. "superneutrality." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000330>
doi:10.1057/9780230226203.1650

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

superstars, economics of

Walter Y. Oi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Gigantic incomes and rare talents attract attention and elicit a search for an explanation. Sherwin Rosen has provided us with an elegant neoclassical model whose market equilibrium is characterized by a superstar. Given a fixed cost of consumption, customers flock to talented sellers. If there are no costs of production, the most talented seller becomes a superstar. Discrete gaps in the talent distribution allow less talented sellers to survive but they must charge lower prices. A competitive market equilibrium thus exhibits a skewed distribution of earnings and outputs.

Keywords

Marshall, A.; motion pictures, economics of; Rosen, Sherwin; Simon, H.; superstars, economics of; Yule distribution

Article

The settings in which superstars are found share two common elements. ‘First, a close connection between personal reward and the size of one's own market, and second a strong tendency for market size and reward to be skewed toward the most talented people in the activity’ (Rosen, 1981, p. 845). The modern world seems to be characterized by an increasing number of activities dominated by superstars. What is the defining characteristic of a superstar? Why do we have more of them today? In this article, I first review the original Rosen model. I next turn to the Yule distribution which yields the skewness in returns. The final section relaxes the assumption of a fixed distribution of talents among potential sellers that is costlessly known to all agents.

Skewness in equilibrium

An individual derives utility from a composite good *x* and a flow of services *y* which is a function of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
quantity N and the quality z of each unit.

$$U = U(x, y), \quad y = g(z, N). \tag{1}$$

Gone with the Wind was an A movie, while *Getting Gertie's Garter* was a B movie. I can get a larger service flow y by seeing better movies or going to more movies. Consumption of y takes time. There is a fixed cost of consumption per unit irrespective of quality. The cost of the service flow depends on quantity N times the full cost per unit.

$$C = [p + s]N. \tag{2}$$

Suppose that quantity and quality multiplicatively determine the service flow, $y=Nz$. If x is the numeraire, the budget constraint is given by,

$$x + vy = I, \quad v = [p(z) + s] / z. \tag{3}$$

A consumer will substitute quality for quantity and demand a higher-quality good if $p(z)$ rises at a modest rate as z is increased.
On the supply side, the quality of a unit of output is a function of the talent of a seller q and her market output m.

$$z = h(q, m), \; h_q > 0, \; h_m \leq 0. \tag{4}$$

A seller's output is a function of her talent, $m=m(q)$ which is chosen to maximize profits.

$$R(q) = p(q, m)^{*} m - c(m), \tag{5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $c(m)$ is the total cost of producing m units of output, and $p(q,m)$ is the price that consumers will pay for a unit of output supplied by a type q seller. A seller's net revenue is a convex function of her talent. She will supply a service flow y if $R(q) \geq k$ where k is the return in some alternative activity. Turn first to the market equilibrium when production is characterized by what Rosen called a public goods technology. There are no costs of production, $c(m) = 0$. The price of a unit of output must satisfy the equilibrium condition,

$$p = vq - s. \tag{6}$$

where v is the implicit price for a unit of the service flow. Since one seller can costlessly supply the entire market, the implicit price for y and the price p for a unit of output yielding $q = z$ units of the service flow are bid down to the zero profit level. The superstar located at the right extreme of the distribution of talents $\Phi(q)$ is the only supplier, but she cannot raise the price above $p_0 = (vq_0 - s)$ because the second most talented seller is nearly a perfect substitute. If, however, there is a discrete gap in the talent distribution so that $q_0 = q_1 + \varepsilon$. The superstar can charge a price, $p_0 = (vq_0 - s)$, while the second best has to settle for a lower price, $p_1 = (vq_1 - s)$. The superstar can realize a total rent of $v\varepsilon N$ where N is the market demand.

If there are diseconomies of scale, several sellers producing differentiated services can survive. Each firm equates marginal cost to marginal revenue. More talented sellers charge higher prices and supply larger outputs. The quantity–talent gradient is steeper than the price–talent gradient. A highly talented seller is less contaminated by the presence of competitors. A high-q seller can and does handle larger crowds who are drawn to her in order to spread the fixed cost of consumption. One unit of a high-quality product yields a larger service flow. Technological advances in communication, transportation and packaging enabled the most talented sellers to increase their share of the total market. Alfred Marshall (1920, p. 686) observed, ‘But so long as the number of persons who can be reached by a human voice is strictly limited, it is not very likely that any singer will make an advance on the ten thousand pounds said to have been earned in a season by Mrs. Billington at the beginning of the last century’. The skew distribution of earnings in the Rosen model is the outcome of a competitive market equilibrium. ‘(a) All sellers maximize profits and cannot earn larger amounts in other activities, and (b) all buyers maximize utility and cannot improve themselves by purchasing from another seller’ (Rosen, 1981, p. 846).

A stochastic model

A skew distribution is not limited to certain economic activities. Herbert A. Simon (1955, p. 426) noted five other examples: distribution of (*a*) words in a book, (*b*) scientists by number of papers published, (*c*)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
cities by population, (*d*) incomes by size, and (*e*) biological genera by number of species. Simon argued that ‘one is led to the conjecture that if these phenomena have any property in common, it can only be a similarity in the structure of the underlying probability mechanism’. A class of such distributions was described by G. Udny Yule (1924).

$$f(i) = \psi B(i, \rho + 1). \tag{7}$$

where ψ and ρ are constants, and $B(i, \rho + 1)$ is the Beta function. Chung and Cox (1994) describe a process that can generate the Yule distribution. Each consumer sequentially buys one record. After the last consumer has made her purchase, the process repeats itself and so on, where ($\text{record}_s \neq \text{srecord}_t$) with $(s, t = 1, 2, \ldots n)$. It is assumed that (*a*) the probability that consumer $k + 1$ chooses a record that was already chosen by *i* of the previous *k* consumers is proportional to *i*, and (*b*) there is a constant probability δ that $k + 1$ chooses a record that was not yet chosen by the previous *k* consumers. Chung and Cox report that in the 1958–89 period 1,377 recording artists had at least one gold record. The list was headed by the Beatles with 46 gold records, followed by Elvis Presley with 45. Setting $\rho = 1$, the Yule distribution predicts that half of the sample should report having exactly $i = 1$ gold record; the actual fraction was $f(1) = .485$. A chi-square test confirms the goodness of fit of the data to the Yule distribution.

M. Adler (1985) motivates this process by arguing that consumption requires knowledge. One has to talk to others to learn about an artist's work. Consumers minimize the cost of searching for knowledgeable discussants by choosing the most popular artist. It can be supposed that consumers initially believe that all artists are equally likely to become stars. Each consumer lives for *n* periods and revise their prior distributions after each period. If a slight majority of consumers choose an artist as their first choice, that artist will snowball into a star because her majority will steadily increase. Notice that there need be no dispersion in the talents of potential superstars. One performer attains a slight majority of consumers who spread the word to other consumers. The model is vague about how long it takes to establish a stable skewed Yule distribution. Elvis Presley skyrocketed into stardom quickly. Nat King Cole was struggling on Avalon Boulevard in Los Angeles in 1938. It took him nearly two decades to become a star. The speed with which information spreads ought to influence the time to reach stardom. Notice that nothing had to be assumed about the fixed cost of consumption or about the technology of production to obtain the J-shaped Yule distribution.

Talent and superstars

Superstars are found in certain kinds of economic activities where there is a concentration of output among a few individuals, marked skewness in the distribution of income. Do we identify these activities by measures of concentration and inequality of incomes? The individuals at the top not only supply a higher-quality product but they also enjoy massive scale economies. Borghans and Groot (1998) contend

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that these activities are mainly media events in show business, arts and letters, and sports. These industries are characterized by competition for the number one position. Those at or near the top acquire endogenous property rights which yield additional income. ‘The extra reward is not due to the way in which the superstar performs but the fact that people are not satisfied to see other players once the superstar has been identified’ (Borghans and Groot, 1998, p. 555.). Their model also shows that income increases with the extent of the market, and market size is more important than the superiority of the superstar. Small differences in talent can be associated with large differences in income when output can be almost costlessly replicated.

Rosen assumed that there was a fixed distribution of talent among potential sellers. Borghans and Groot argue that a superstar acquires an endogenous property right which entitles her to a super-normal return. Talent-like ability is an attribute that defies easy quantification. Alan Krueger (2005) approximates talent q by the length of a performer's biography in the *Rolling Stones Encyclopedia of Rock & Roll*. Campbell McConnell is the author of the economics principles text which is allegedly the best-seller and hence is at the top of the talent distribution (McConnell and Brue, 2005). Defining and measuring talent in sports ought to be simpler. There are lots of statistics to identify the best shortstop. What is amazing is the sharp rise in salaries. A major league ballplayer received an average salary of $29,303 in 1970, which was 4.23 times the annual earnings in manufacturing. The ratio rose to 6.93 in 1977. The average salary of all major league ballplayers exceeded one million dollars in 1992 (The Baseball Archive). (The average salary in 1997 was 1.38 million dollars, which was 48.1 times the annual earnings in manufacturing.) The averages conceal the wide dispersion among individual players. Free agency and cable television apparently account for rising incomes for these superstars. The extent of the market again seems to be more important than the superiority of the talent.

In the Chung and Cox model, all sellers are equally talented. Consumer choices are proportional to the previous frequency of purchases which generated the J-shaped Yule distribution of record sales. Performers differ in Adler's world. Individuals search for information about performers through discussions with friends and experts. Performers could lobby the experts, bribe the disc jockeys, or invest in training resulting in a different distribution of ‘talents’. Will the search for information and actions by potential sellers generate a stable distribution of talents?

Gigantic incomes and rare talents attract attention, elicit conjectures, and encourage the search for an explanation. The phenomenon of superstars calling for an explanation is the shape of the income distribution with its long right tail. It is not the talent of the superstar or the determinants of the star's income. Sherwin Rosen has provided us with an elegant neoclassical model. Consumers patronize the most talented sellers to spread the fixed cost of consumption. Talented sellers have access to a low marginal cost technology enabling them to dominate their nearby competitors. Market size and rewards are skewed towards these talented superstars. Rosen's is a neat competitive market equilibrium explanation for all kinds of superstars.

See Also

- motion pictures, economics of
- Rosen, Sherwin

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Adler, M. 1985. Stardom and talent. *American Economic Review* 75, 208–12.

Beckmann, M. 1987. Rank. In *The New Palgrave: A Dictionary of Economics*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. Basingstoke: Macmillan.

Borghans, L. and Groot, L. 1998. Superstardom and monopolistic power: why media stars earn their marginal contribution to welfare. *Journal of Institutional and Theoretical Economics* 154, 546–71.

Chung, K. and Cox, R. 1994. A stochastic model of superstardom: an application of the Yule distribution. *Review of Economics and Statistics* 76, 771–5.

Krueger, A. 2005. The economics of real superstars: the market for rock concerts in the material world. *Journal of Labor Economics* 23, 1–30.

Marshall, A. 1920. *Principles of Economics*, 8th edn. New York: Macmillan, 1952.

McConnell, C. and Brue, S. 2005. *Economics, Principles, Problems, and Policies*, 16th edn. New York: McGraw Hill.

Rosen, S. 1981. The economics of superstars. *American Economic Review* 71, 845–58.

Rosen, S. 1983. The economics of superstars: reply. *American Economic Review* 73, 460–2.

Simon, H. 1955. On a class of skew distribution functions. *Biometrika* 42, 425–40.

The Baseball Archive. Online. Available at http://www.baseball1.com/bb-data/bbd-mas.html, accessed 29 December 2005.

Yule, G. 1924. A mathematical theory of evolution, based on the conclusion of Dr. J.C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London*. Series B 213, 21–87.

How to cite this article

Oi, Walter Y. "superstars, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_E000239>
doi:10.1057/9780230226203.1651

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

‘ supply and demand ’

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cournot, A. A.; Hutcheson, F.; Law, J.; Longfield, M.; Malthus; Marcet, J.; Mill, J. S.; Ricardo, D.; Steuart, J.; supply and demand; Torrens, R.; West, E.; Whewell, W.

Article

Although the notion of supply and demand in the context of market price determination potentially goes back to the economic writings of the Greek philosophers, the terminology itself is of more recent origin. It was not given any prominence in a chapter title or table of contents until well into the second decade of the 19th century (see Groenewegen, 1973), though during the previous two decades the phrase was used in the literature with increasing frequency. Its first use in English writings appears to have occurred in 1767 (see Thweatt, 1983). The discussion here is confined to English usage: French, Italian and German developments are omitted. This unfortunately means ignoring the interesting distinction made in German by Marx between ‘Zufuhr und Nachfrage’ (supply and demand) and ‘Angebot und Nachfrage’ (offer and demand) with its analytical connotations for more modern developments in economics (see Schefold, 1981).

It is certain that by the end of the 17th century the expression ‘supply and demand’ was not in use by English economic writers. For example, Locke (1691, pp. 45–6, 59, 61), despite the importance of the concept for his general argument, followed the contemporary practice and expressed the notion either in terms of the ‘proportion of the number of Buyers and Sellers’ or, more novel, the ‘quantity in proportion to [the] vent’. This terminology was explicitly criticized by John Law (1705, p. 5) who stated, ‘The Prices of Goods are not according to the quantity in proportion to the Vent, but in proportion to the Demand.’ By the end of the 1730s the demand part of the phrase was being used by prominent authorities like Erasmus Philips, Jacob Vanderlint and Bishop Berkeley and during the 1730s and 1740s was partially transformed by Francis Hutcheson (1755, II, pp. 53–4) into the argument that ‘prices of goods depend on these two jointly, the *Demand* … and the *Difficulty* of acquiring …’.

Another Scottish writer, Sir James Steuart, has the distinction of combining supply and demand together

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
on a number of occasions in the context of price determination and competitive analysis and thereby originating the first major use of the phrase. For example, in his chapter 'Of Demand', Steuart (1767, p. 153) argues, 'The nature of demand is to encourage industry; and when it is regularly made, the effect of it is, that the supply for the most part is found to be in proportion to it, and then the demand is simple.' Steuart (for example, 1767, p. 184) used the phrase on a number of other occasions and presumably it is from this source that it spread to authors like Adam Smith, Malthus, Thornton, James Mill, Horner, Brougham, Lauderdale and others, whose practice in this respect has been documented by Thweatt (1983). Some other aspects of the adoption in economics of the phrase 'supply and demand' may be mentioned. By the middle of the second decade of the 19th century usage of the phrase was still relatively rare. Only a few examples of more than 20 uses in a single work have been identified, no more than a dozen works use it more than ten times and most of the works Thweatt (1983) examined used the phrase only sparingly. Secondly, Scottish use of the new terminology was decisive with Steuart as the pioneer, Smith as a considerable influence and the *Edinburgh Review* circle as most important disseminator of the new terminology. This is not to say that English sources were unimportant. Malthus, for example, used it no less than 20 times in his influential second edition of the *Essay on Population* (1803); Torrens used it 29 times in his *The Economists Refuted* (1808) and over 70 times in his *Essay on Money and Paper Currency* (1812). However, the phrase 'supply and demand' did not really gain prominence until 1817 when Ricardo (1817, p. 382) used it in a chapter heading, while in the previous year, Mrs Marcet (1816, p. 296) had used it in the summary table of contents at the start of her Conversation XV on value and price.

Three further observations can be made on the etymology of the phrase 'supply and demand'. First, the very slow adoption of the word 'supply' in conjunction with the much earlier use of 'demand' may be explained by the distinct usages associated with that word in expressions like 'granting supply' used in the language of public finance and its associated military connotations of supply troops. Secondly, the predominance of Scottish usage of the new terminology combined with their undisputed role in originating its usage in English economic discourse suggests the adaptation of the word from the French *supplier*, a verb with a less restricted meaning than English usage of the verb 'to supply'. Thirdly, and most importantly, up to the 1830s either term was rarely used in its modern sense, that is, as a function of price. English pioneers in this modern usage appear to have included West (1826), Whewell (1930) and Longfield (1833) but systematic exposition of the practice had to await the work of Cournot (1838) and to a lesser extent that of John Stuart Mill (1848), and was of course especially developed by Marshall.

Bibliography

Cournot, A.A. 1838. *The Mathematical Principles of the Theory of Wealth*. Trans. N.T Bacon with notes of I. Fisher, London: Macmillan, 1897; Homewood, IL: Irwin Economic Classics, 1963.

Groenewegen, P.D. 1973. A note on the origin of the phrase, supply and demand. *Economic Journal* 83, 505–9.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hutcheson, F. 1755. *A System of Moral Philosophy*. Glasgow: Robert and Andrew Foulis.

Law, J. 1705. *Money and Trade Consider'd, with a Proposal for Supplying the Nation with Money*. Edinburgh: Anderson.

Locke, J. 1691. *Some Considerations of the Consequences of the Lowering of Interest and Raising the Value of Money*. 2nd edn, London, 1696.

Longfield, M. 1833. *Lectures on Political Economy, Delivered in Trinity and Michaelmas Terms, 1833*. Dublin: Milliken, 1834.

Marcet, J. 1816. *Conversations on Political Economy*. 5th edn, London, 1824.

Malthus, T.R. 1803. *Essay on the Principle of Population*. 2nd edn. London: J. Johnson.

Mill, J.S. 1848. *Principles of Political Economy*. London: Parker.

Ricardo, D. 1817. *Principles of Political Economy and Taxation*. In *Works and Correspondence of David Ricardo*, vol. 1, ed. P. Sraffa and M.H. Dobb. Cambridge: Cambridge University Press, 1951.

Schefold, B. 1981. Nachfrage und Zufuhr in der klassischen Ökonomie. In *Studien zur Entwickelung der ökonomischen Theorie*, vol. 1, ed. F. Neumark. Berlin.

Steuart, Sir, J. 1767. *An Inquiry into the Principles of Political Oeconomy*. Ed. A.S. Skinner, Edinburgh and London: Oliver & Boyd for the Scottish Economic Society, 1966.

Thweatt, W.O. 1983. Origins of the terminology, supply and demand. *Scottish Journal of Political Economy* 30, 287–94.

Torrens, R. 1808. *The Economists Refuted*. London and Dublin.

Torrens, R. 1812. *Essay on Money and Paper Currency*. London.

West, E. 1826. *Price of Corn and Wages of Labour*. London.

Whewell, W. 1830. Mathematical exposition of some doctrines of political economy. In *Transactions of the Cambridge Philosophical Society*. New York: A.M. Kelley, 1971.

How to cite this article

Groenewegen, Peter. "‘supply and demand’." The New Palgrave Dictionary of Economics. Second

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000333>
doi:10.1057/9780230226203.1652

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

supply chains

Martin A. Lariviere
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A supply chain encompasses all the resources and processes required to fulfill the demand for a product. A properly functioning supply chain is critical for a firm to be able to equate supply and demand at a reasonable cost. We first review common measures of supply chain performance and inventory related costs. We next discuss how facility choices, transportation modes, inventory policies and information structures interact to determine the profitability and responsiveness of a supply chain. We conclude with common criteria for categorizing supply chains.

Keywords

bullwhip effect; excess demand; flow time; information technology; inventory and holding costs; inventory investment; inventory policy; inventory theory; supply chains; supply-chain profitability

Article

Supply chains are all the resource and processes required to fulfill the demand for products.
A supply chain encompasses physical assets such as factories, warehouses, and trucks as well as non-physical assets such as product design organizations, forecasting processes, and inventory tracking systems. As such, the supply chain affects nearly everything a firm does. A smoothly operating supply chain is essential for satisfying customers and consequently necessary for supporting marketing campaigns. Similarly, funding facilities and inventories is among the biggest capital needs of the firm and hence central to finance decisions.

Evaluating supply chains

Two important metrics of supply-chain performance are flow time and profitability. Flow time is the average amount of time a unit of material spends in the supply chain. Flow time directly affects the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
amount of inventory in a supply chain. For a fixed rate of material flow, inventory increases in proportion with the flow time. Flow time also affects the firm's ability to react to shifts in market demand and its exposure to product obsolescence.

Supply-chain profitability is a function of revenues and costs. Revenues depend on the supply chain's ability to manage the flow of material, avoiding market downs while still satisfying demand. On the cost side, production and transportation costs will be driven by the product's design and the firm's choice of facilities. Additionally, there are costs associated with managing inventory, notably fixed costs of acquiring inventory and holding costs. The latter include out-of-pocket costs such as insurance and the opportunity cost of tying up capital in inventory. There may also be goodwill costs for missing sales (on the assumption that excess demand is lost) or forcing customers to wait (on the assumption that excess demand is back-ordered).

Additional measures reflect service quality. The fill rate is the fraction of product demand filled from inventory. The service level is the probability that all demand in a selling season is satisfied. It is straightforward to define these for a single product, but customers often order multiple items, and a high line-item fill rate may not result in a high order fill rate.

Determinants of supply-chain performance

Assuming supply equals demand is a staple of economic analysis, but equating the two in practice is challenging. Managing a supply chain requires a variety of decisions that vary in the costs and timing involved. Following Chopra and Meindl (2004), we group these decisions around four drivers of supply-chain performance: facilities, transportation, inventory, and information.

The facilities of a supply chain can broadly be classified as either processing material or storing material. Processing should be understood as adding value to the product; this may involve complex manufacturing or simply packaging. Storing material may involve more than keeping products safe. Distribution centres often ‘break bulk’, taking in large quantities (for example, truckloads) but sending out small quantities (for example, cases).

Investing in facilities is both time-consuming and costly. Consequently, the design of a facility network is a long-term, strategic choice that must balance current and future needs. Primary decisions are the number and locations of facilities. Centralized processing facilities usually result in lower average cost per unit produced. Centralizing storage facilities decreases the amount of inventory required to provide a given service level (Eppen, 1979). For either type of facility, centralization savings must be balanced with increased travel times and costs. Location also affects the available workforce, the supply of raw materials, tariffs and many other factors.

Supply-chain design must also consider the capacity of facilities and their operating characteristics. Capacity affects both cost and responsiveness. More capacity implies a greater investment, but this does not mean that a firm should build the minimum capacity necessary to cover expected demand. Markets are dynamic and uncertain, and a supply chain must be able to react to surges in demand. Having capacity above expected demand allows the firm to respond without imposing excessive delays on customers. The importance of additional capacity typically increases with the level of uncertainty in demand. An example of a relevant operating characteristic of a facility is its flexibility in handling different product mixes. Mix flexibility depends on both capital choices and workforce training. Much like carrying extra capacity, it adds cost but enables greater responsiveness.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Once facilities are specified, there is the question of how to move material through the network. Transportation modes represent trade-offs in speed and cost. Airfreight is faster than rail but also more expensive. Transportation decisions are intimately tied to inventory and facility choices. Because shorter transportation times allow for lower inventory, paying for premium service may be warranted. Facility decisions dictate the viable transportation decisions; a factory, for example, must be near a railhead in order to utilize trains. While firms often have one primary mode of transportation, it is not unusual for multiple transportation methods to be used. Transportation decisions thus include the rules governing which shipments are sent by which methods. In addition, a firm must decide the extent to which it will aggregate shipments over time. When a shipment method has a fixed capacity (for example, a truckload), combining today's shipments with tomorrow's may result in a higher utilization of capacity and a lower average cost at the expense of delay.

Inventory is what most people associate with supply chains. The primary reasons for carrying inventory are economies of scale and the time to execute supply processes. Economies of scale arise when there is a fixed cost for initiating production or transporting goods that is independent of the quantity involved. Producing or shipping in a batch then results in lower average cost but creates inventory in excess of the firm's immediate needs. When demand is relatively stable, such ‘cycle stocks’ are simply a cost of doing business. When demand shifts rapidly, fixed costs can greatly hinder a firm's ability to adapt to the market. Supply-chain improvement initiatives consequently concentrate on reducing fixed costs. Beyond simply lowering costs, reduced fixed costs allow for smaller inventories and shorter flow times.

When fulfilling demand is time-consuming, ‘pipeline’ inventory of material in production or in transit is inevitable even with perfectly stable demand. When demand is uncertain, the firm must hold additional inventory. ‘Safety stocks’ buffer customers from production and distribution processes so demand can be filled without delay. Safety stocks increase both cost and flow time. Reducing safety stocks while maintaining service levels requires more accurate forecasts and shorter lead times.

Inventory decisions take place at different levels. At a low level, there are the basic questions of when and how much to order. Inventory policies depend on the costs of the system and desired service performance. For example, an order size may be found from the classical economic order-quantity model which trades off holding costs with fixed ordering costs. A reorder point, which dictates when to place an order, can be found from the service level the firm needs to offer the market.

At a higher level, there is also the question of what form of inventory to hold. Should the firm hold finished goods, sub-assemblies or merely components? In a multi-product environment, delaying product differentiation (much like centralizing inventory) allows the firm to provide a given level of service with less inventory. In some settings (for example, laundry detergent), customers demand quick service, and holding finished goods is inevitable. In others (for example, paint), the final processing steps are sufficiently quick and customer tastes sufficiently heterogeneous that a postponement strategy is viable. Because customers will tolerate a delay to get exactly what they want, the firm may postpone differentiation and save on inventory costs. The appropriate form of inventory depends on both customer needs and available technology. Supply-chain management thus reaches back to product and process design.

The final supply-chain performance driver is information. Managing inventory requires an estimate of the distribution of demand and how it varies over time, not just a forecast of mean demand. One must also know the firm's pricing and promotional plans and estimate how these will affect demand. The better the estimate of demand, the less safety stock is required. The difficulty of forecasting demand

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
interacts with other decisions the firm has made. If the firm exploits postponement and holds only sub-assemblies, it may forecast demand at the sub-assembly level rather than at the end-product level. Also, not all levels of a supply chain will necessarily have the same information available to them. Demand information often becomes distorted and more variable as one moves up a supply chain away from the market, a phenomenon termed the ‘bullwhip effect’ (Lee, Padmanabhan and Whang, 1997).

Beyond knowing demand, a supply chain also needs to know supply. Tracking the amount of inventory on hand seems a simple matter of comparing what has come in with what has gone out. In practice, maintaining accurate inventory records is a complicated process. A supermarket carries tens of thousands of distinct products, and there is no way to maintain correct records without information technology that integrates receiving databases with point-of-sales scanners. Even these systems are not perfectly accurate because they do not account for misplaced or stolen items. Further, this deals with only one location. Optimal decisions may depend on the available inventory at multiple locations as well the status of goods in production or transit.

Modern information systems have greatly affected supply chains, both at individual locations and between locations. They have had an additional benefit in reducing the effort and cost of placing orders. Electronic data interchange systems reduce the fixed costs of placing orders and hence result in lower cycle stocks.

Classifying supply chains

Supply chains can be classified by their physical structure, their strategic focus and their economic organization. The physical structure reflects how material moves through a supply chain. The simplest supply chains have a serial structure. Each echelon (or level) consists of a single node and all material moves from the highest to the lowest echelon. Alternatively, one can have an assembly system in which multiple nodes at a higher level all supply a single node at a lower level. Thus, not all material follows the same path through the supply chain. Distribution systems are essentially the reverse of assembly systems. A single point at an upper echelon supplies multiple points at a lower echelon. These structures can be combined. Looking upstream from an auto assembly plant, one sees an assembly structure with parts arriving from many facilities. Looking downstream, one sees a distribution system as completed cars are sent to multiple dealers.

The physical structure dictates the decisions that must be made. The analysis of serial systems focuses on what the inventory policy should be at each echelon. Assembly systems with one end product are similar, but if there are multiple products one must also determine how orders are to be filled as a function of available inventory. In a distribution system, one needs to consider how to ration stock among the multiple end points when the central higher node has limited inventory.

Following Fisher (1997), the strategic focus of supply chains can be either efficient or responsive. Efficient supply chains are designed to produce and deliver goods as cheaply as possible even though such a design may greatly increase flow times. Responsive supply chains sacrifice low cost for speed. This classification is useful because there is no one best supply-chain design. Firms must make trade-offs to align their supply chains with the markets they serve. Efficient supply chains are appropriate for ‘functional’ products such as consumer staples with long life cycles and low variety. ‘Innovative’ products, such as fashion clothing and consumer electronics with short life cycles and high variety, require greater responsiveness.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Finally, one may classify supply chains based on their economic organization, the degree to which they are centralized or decentralized. We have assumed a completely centralized supply chain and used the terms ‘firm’ and ‘supply chain’ interchangeably. However, essentially no supply chain today is completely controlled by one party. Rather, most are decentralized with multiple decision makers acting in their own interests. A consumer electronics company may hire one firm to produce an application-specific integrated circuit, contract with a second to assemble the final product, and rely on a third to transport the goods, all while selling through independent retailers.

Designing a supply chain consequently goes beyond determining the supply chain drivers discussed above. One must specify not just where a facility will be located but who will own and operate it. Decentralized supply chains thus raise issues that go beyond traditional supply-chain analysis. For example, who is ultimately responsible for unsold merchandise can make a dramatic difference in the performance of a decentralized supply chain, but it is not an issue in a centralized one (Pasternack, 1985).

See Also

- operations research

Bibliography

Cachon, G. and Terwiesch, C. 2006. *Matching Supply and Demand: An Introduction to Operations Management*. New York: McGraw-Hill/Irwin.

Chopra, S. and Meindl, P. 2004. *Supply Chain Management*, 2nd edn. Upper Saddle River, NJ: Pearson Prentice Hall.

de Kok, A. and Graves, S., eds. 2003. *Supply Chain Management: Design, Coordination and Operation*. Handbooks in Operations Research and Management Science. Amsterdam: Elsevier.

Eppen, G. 1979. Effects of centralization on expected costs in a multi-location newsboy problem. *Management Science* 25, 498–501.

Fisher, M. 1997. What is the right supply chain for your product? *Harvard Business Review* 75(2), 105–16.

Lee, H., Padmanabhan, V. and Whang, S. 1997. Information distortion in a supply chain: the bullwhip effect. *Management Science* 43, 546–58.

Pasternack, B. 1985. Optimal pricing and return policies for perishable commodities. *Marketing Science* 4, 166–76.

Tayur, S., Ganeshan, R. and Magazine, M. 1998. *Quantitative Models for Supply Chain Management*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Boston: Kluwer.

How to cite this article

Lariviere, Martin A. "supply chains." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000510> doi:10.1057/9780230226203.1653

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

surplus

Anthony Brewer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The surplus produced by an economy, in the sense discussed here, is its output over and above the necessary subsistence of the labour force as well as other costs. This concept has a long history in economics, with the focus sometimes on the uses of the surplus for luxury consumption, state spending, or new investment, and sometimes on income distribution – if workers are paid a subsistence wage then the surplus accrues to non-wage earners.

Keywords

aggregate demand; Baran, P.A.; Cantillon, R.; capital accumulation; classical economics; excess saving; Hobson, J. A.; Hume, D.; imperialism; labour theory of value; luxuries; Marx, K. H.; natural price; necessities; neo-Ricardian economics; net product; Petty, W.; Physiocracy; price of production; profit and profit theory; Quesnay, F.; rent; Ricardian distribution theory; Ricardo, D.; Sraffa, P.; stationary state; subsistence; surplus; surplus value; Tableau économique; Turgot, A. R. J.; unemployment; value; von Neumann, J.

Article

The key question in defining the surplus is: what should be counted as the cost of maintaining the labour force? If we deduct from output only what is physically necessary to maintain workers at a minimum standard of existence, the resulting surplus measure would represent the amount available for everything else. It could be relevant, for example, to a very poor country or to a situation of total war. In a very crude version of classical economics, wages might be held down to the subsistence minimum by Malthusian population pressures with the surplus over bare subsistence going to others. In modern developed economies wages are well above physical subsistence, so this definition of surplus is of little interest. At the other end of the scale, we could count the actual wage, whatever it might be, as the cost of subsistence of the labour force, so surplus is simply defined as total income less wages, equal by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
definition to total non-wage incomes.
Modern proponents of the surplus approach typically claim that (real) wages are determined prior to other variables, at some socially acceptable level (perhaps representing workers’ bargaining power), so wages can be treated as fixed, with profits (plus other non-wage incomes) determined by the remainder, or surplus. With very strong assumptions, the static equilibrium profit rate and prices can then be determined, given the real wage.

Necessities and luxuries: from Petty to Smith

Early writers on economic issues often used a concept of surplus, focusing on the production of necessities (frequently identified with food) and drew conclusions about the relation between sectors, not sources of income. The key idea is that people need a certain amount of food (and other necessities), but no more, since ‘[t]he desire of food is limited in every man by the narrow capacity of the human stomach’ (Smith, 1776, p. 181). The agricultural sector has to feed itself and everyone else. The size of the non-agricultural sector is limited by the agricultural surplus, that is, the output of food less the amount consumed by the agricultural sector itself. Using a somewhat broader definition of necessities, William Petty gave a hypothetical example (with an implausibly high level of productivity for the period): ‘if there be 1000 men in a territory, and if 100 of them can raise necessary food and raiment for the whole 1000’ (Petty, 1899, vol. 1, p. 30). His worry was about employment – only 100 are needed to provide necessities, so what will the rest do? In his example, he suggested a variety of employments (200 produce for export, 400 produce luxuries, and so on) but some might remain unemployed.
David Hume used the idea of an agricultural surplus quite differently. ‘The land may easily maintain a much greater number of men, than those who are immediately employed in its culture, or who furnish the more necessary manufactures to such as are so employed’ (1955, p. 6). The question he asked was: why should farmers work to produce more than they themselves want to eat? They might be forced to do so, in a feudal system, but that was unlikely to generate much of a surplus. If, however, they could buy attractive manufactures, they would have an incentive to produce and market a surplus. ‘When a nation abounds in manufactures and mechanic arts, the proprietors of land, as well as the farmers, study agriculture as a science, and redouble their industry and attention’ (1955, p. 11). For Hume, this was not simply a theoretical point – in his *History of England* (1778) he described how the introduction of luxuries, initially from abroad, changed the behaviour of successive generations down to his own time.
Adam Smith told a similar story. ‘When by the improvement and cultivation of the land the labour of one family can provide food for two, the labour of half the society becomes sufficient to provide food for the whole’ (1776, p. 180). Although the need for food is limited, ‘the desire of the conveniences and ornaments of building, dress, equipage, and household furniture, seems to have no limit’ (p. 181). Smith's account of development in Europe drew on Hume, and stressed that as landlords became more concerned with luxury spending than with maintaining their political power, they switched to forms of tenure which gave better incentives to farmers.
The concept of surplus sketched in the last three paragraphs is a surplus generated in the necessity-producing sector which supports everyone else – soldiers, the upper class and their servants, luxury producers, and so on. There is, however, no presumption that farmers, or workers, are reduced to bare subsistence. Part of the surplus may be extracted as rent, but part accrues to the producers of necessities

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
themselves and allows them to purchase other things.

Cantillon and Quesnay

In Richard Cantillon's theory, written before Hume but published later, the land is fundamental and everyone works, directly or indirectly, for the landlords. Whether the owner manages the land himself or lets it out to farmers, the farmers and labourers must 'have a living', but the 'overplus [in the French original, '*surplus*'] of the land' goes to the owner (1755, p. 6). Wages are set by social convention, at levels which vary between different places. Here, then, the distribution of income between landlords and the rest is determined by the surplus over a socially determined subsistence level. François Quesnay developed Cantillon's model further, using the famous *Tableau économique* to show the relations between the different sectors of the economy (Meek, 1963). Only agriculture generates a net product (*produit net*) or surplus of output over costs, since the 'sterile' manufacturing sector covers its costs, but no more.

Cantillon treated agricultural productivity as essentially given. Quesnay, by contrast, thought that French agriculture had declined, bringing the whole economy down with it. Investment, he argued, is needed to maintain agricultural productivity. Excessive taxes and restrictions on trade had so reduced agricultural returns that farmers were unable to invest. The net product plays two roles in Quesnay. Reforms which restore returns to a healthy level accrue first to farmers, allowing them to invest, but when leases come up for renewal the owners will be able to capture the increased net product as rent. Whether either Cantillon or Quesnay allowed for any net profit on investment to accrue to tenant farmers in the long run is a matter of debate.

In sum, Quesnay and his followers, the Physiocrats, took a critical step towards classical economics by recognizing the need for investment (financed out of the net product), but did not explicitly recognize the return to investment as a continuing source of income (and a share of the surplus) or make any provision for continued investment beyond the periodic renegotiation of rents. This next step should be credited to Anne Robert Jacques Turgot, who anticipated key elements of Adam Smith's system ten years before the *Wealth of Nations*.

Classical economics

David Ricardo plays a central role in the surplus interpretation of classical economics (hence the surplus approach is alternatively known as neo-Ricardian economics). He based himself on Smith but concentrated particularly on the distribution of income between wages, profits, and rent, which he thought Smith had not treated adequately. The familiar textbook summary of Ricardian distribution theory makes a good starting point. Land is of varied qualities, and there are diminishing returns to investment on any given piece of land. As cultivation expands (to feed a growing population), marginal returns fall. Marginal land yields no rent so, with a given real wage, profit is the difference between costs (including wages) and returns on marginal land. Better land, of course, yields higher returns, but landowners capture intra-marginal returns as rent. Capital mobility keeps non-agricultural profits in line with returns on marginal land. As the economy grows, the margin moves out and the surplus at the margin falls, bringing the profit rate down. In this reading of Ricardo, then, profit is the surplus over

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
given wages on marginal land, while rent and profits absorb the surplus for the whole economy.
There are two main problems with this reading. First, costs and returns have to be calculated in terms of prices to give profit as a percentage return on investment. Ricardo sought to solve this problem with a labour theory of value – prices proportional to labour inputs. This way, costs and returns could be calculated without reference to the profit rate, which is the unknown in the calculation. Unfortunately, labour values are not compatible with equalization of profit rates across industries. Ricardo had no answer to this problem.
Second, and more relevant here, the surplus reading of Ricardo requires real wages to be treated as independent of profits and prices. It is not clear that this assumption adequately represents Ricardo's wage theory. He defined the natural price of labour, like the natural price of anything else, as the cost of producing it, hence ‘that price which is necessary to enable the labourers … to subsist and perpetuate their race, without either increase or diminution’ (1817, p. 93). But he then admitted that the wage must rise above this level in a growing economy ‘for an indefinite period’ (p. 95). On average, therefore, the actual profit rate must be below the rate calculated using the natural (subsistence) wage throughout the process of growth, until the stationary state is reached. Since growth influences wages and hence profits, and the growth rate of the economy depends (inter alia) on the profit rate and the total size of the surplus, the surplus cannot play the independent causal role that the simple version suggests.
Adam Smith had rejected a subsistence wage theory much more explicitly than Ricardo. He had a notion of agricultural surplus, as noted above, but his (rather confused) theory of profit does not fit the Ricardian surplus pattern at all. If we look at the classical school as a whole, Ricardo's emphasis on factor returns and on surplus as an explanation of non-wage incomes is only a relatively small element in a large body of work devoted to economic growth, to policy issues and incentives, and so on.

Marx

If there is one writer who fits what has been called the surplus approach to income distribution, it is Karl Marx. This is not surprising, because the surplus approach was largely inspired by Marx and by his view of his predecessors.
The substance of Marx's analysis of profit (or surplus value) is very similar to Ricardo's, though Marx's terminology and conclusions are very different. Like Ricardo, Marx adopted a labour theory of value, but where Ricardo saw the labour theory as a theory of price (and one which he knew could not be sustained, as noted above), Marx distinguished between value and price, simply defining value as labour embodied and thus making the labour theory of value true by definition. In the first two volumes of *Capital* (Marx, 1867–94), he assumed that prices were proportional to values (as defined), without giving any justification.
The value of labour power (the labour value of the real wage) is determined by subsistence needs, including a (rather ill-defined) ‘historical and moral element’. The value produced by each worker depends on the hours worked. Given the assumption that everything (including labour power) sells at a price reflecting its (labour) value, the difference between the value produced and the value of labour power accrues to the capitalist employer as profit.
In the unfinished third volume of *Capital*, Marx (1867–94) tried to link the analysis of the first two volumes to the determination of what he called ‘prices of production’ (long-run equilibrium prices or Smith's ‘natural prices’). The key idea is that total profit is determined by total surplus value, as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
analysed above, but that relative prices adjust to equalize profit rates across industries. His procedure for doing this is not now regarded as sound. In the third volume he also allowed for some part of surplus value to be absorbed by incomes other than profit (such as rent).

In sum, Marx explained profit by the surplus of hours worked over the hours needed to produce subsistence, and hence by output over a socially determined subsistence wage. The main problem of substance with this theory is that his subsistence wage is a matter of definition, without a satisfactory account of how wages are determined in the market. When he did provide such an account (in the splendid account of accumulation in *Capital*, volume I, chapter 25), wages in a growing economy turn out to be determined simultaneously with profit and accumulation (as in Smith and Ricardo), so the wage is not determined prior to other variables after all.

The 20th century

By the time Marx's *Capital* was published, the mainstream in economics was moving away from a subsistence wage theory, and hence away from surplus theories of profit. More generally, late 19th- and 20th-century mainstream (neoclassical) economics had little use for the notion of surplus because economic variables (including incomes, consumption demand, investment and savings decisions, and so on) were all to be explained in a common framework, with no distinction between necessities and other goods or between different types of income.

John Hobson, an avowed heretic, was one of the few to keep the notion of surplus alive. In his account, one part of output corresponds to the costs of production, including both a subsistence wage (with a ‘moral’ element as well as the bare physical minimum) and a return on investment sufficient to induce investors to keep their investment unchanged. The remainder is the surplus. In a growing economy, factor returns have to be higher to induce growth: the part of the surplus absorbed in this way is the ‘productive surplus’ or ‘costs of growth’. The rest is the ‘unproductive surplus’ or ‘unearned increment’ (1910, ch. 4). In modern society the surplus increases with increased productivity, but much of it accrues to a small minority who earn so much that they cannot spend it all. The resulting excess saving causes unemployment, a search for investment opportunities abroad, and hence imperialist expansion.

Paul Baran's influential *Political Economy of Growth* (1957) has much in common with Hobson. Baran defined the ‘actual surplus’ as actual output minus actual consumption (of all sorts), making it effectively equal to new net investment. Up to a certain level, at least, investment determines surplus, not the other way round, because higher/lower investment will induce higher/lower aggregate demand, bringing the actual surplus (in Keynesian terms, saving) into line. By contrast, the ‘potential surplus’ is the difference between maximum potential output and the minimum acceptable level of consumption, showing the maximum possible investment level. Surplus concepts like these provided a rather loose framework for radical discussions of growth, unemployment, the role of military spending, and the like in the mid-late 20th century.

John von Neumann's classic paper on general equilibrium deserves mention, since it embodies a surplus account of growth, assuming that ‘all income in excess of necessities of life will be reinvested’ (1945, p. 2). It was important in the development of mathematical techniques for handling general equilibrium, but the surplus aspect was not widely developed.

Piero Sraffa's *Production of Commodities by Means of Commodities* (1960) can be seen as a restatement

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of the Ricardo–Marx theory of profit and prices, but avoiding the problems caused by the labour theory of value. Put simply, Sraffa wrote down equations for a many-commodity system showing the conditions for an equal profit rate across all industries, and showed that there remained one degree of freedom, representing the distribution of income between wages and profits (assumed to be the only forms of income). Sraffa did not commit himself on how to close the system, but the majority of his followers have chosen to take the real wage as fixed. When this is done, the profit rate and the set of relative prices are simultaneously determined. Sraffa's system does indeed provide a logically coherent version of the surplus theory of profit free of the labour theory of value. Beyond that, it is not clear why one would want to assume that wages (and the whole pattern of outputs and inputs) are determined independently of and prior to the determination of prices and profits.

Conclusion

Classical economists were acutely aware that wages were not very far above subsistence. An important theme in their work was indeed the generation of a surplus over subsistence and the use of part of the surplus for investment and growth, but an interpretation which insists on treating wages as determined prior to other variables is at best a first approximation, used by some classical writers but only with heavy qualifications. In a modern context, it is hard to see what purpose a concept of surplus serves when wages are far above subsistence.

See Also

- Baran, Paul Alexander
- Cantillon, Richard
- Hume, David
- labour theory of value
- Marx, Karl Heinrich
- neo-Ricardian economics
- Petty, William
- Quesnay, François
- Smith, Adam
- Sraffa, Piero
- Turgot, Anne Robert Jacques, Baron de L'Aulne
- von Neumann, John

Bibliography

Baran, P. 1957. *Political Economy of Growth*. New York: Monthly Review Press.

Cantillon, R. 1755. *Essay on the Nature of Commerce in General*. New Brunswick, NJ: Transaction

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Publishers, 2001.

Hobson, J. 1910. The Industrial System: An Inquiry into Earned and Unearned Income. London: Routledge/Thoemess Press, 1992.

Hume, D. 1778. The History of England from the Invasion of Julius Caesar to the Revolution in 1688. Indianapolis: Liberty Fund, 1983.

Hume, D. 1955. *Writings on Economics*, ed. E. Rotwein. Edinburgh: Nelson.

Marx, K. 1867–94. *Capital*, vols 1–3. Moscow: Foreign Languages Publishing House, 1957–62.

Meek, R. 1963. *The Economics of Physiocracy: Essays and Translations*. Cambridge, MA: Harvard University Press.

Petty, W. 1899. *The Economic Writings of Sir William Petty*, 2 vols, ed. C. Hull. Cambridge: Cambridge University Press.

Ricardo, D. 1817. On the principles of political economy and taxation. In *The Works and Correspondence of David Ricardo*, vol. 1, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. A. Campbell, A. Skinner and W. Todd. Oxford: Clarendon Press, 1976.

Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

von Neumann, J. 1945. A model of general economic equilibrium. *Review of Economic Studies* 13, 1–9.

How to cite this article

Brewer, Anthony. "surplus." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000516> doi:10.1057/9780230226203.1654

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

survey data, analysis of

Jeff Dominitz and Arthur van Soest

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

An important advantage of survey data over, for example, administrative data is the opportunity to directly measure subjective phenomena, such as respondents' expectations on some future outcomes or their preferences over consumption bundles. The existing literature shows that data on expectations and preferences reported in carefully designed household surveys can greatly enhance the empirical content of models of economic decisions and choices. A particularly promising route is to combine subjective data with data on actual behaviour.

Keywords

consumer confidence; contingent valuation; expectations; expected utility; health insurance; household surveys; independence of irrelevant alternatives; intertemporal choice models; job search; Katona, G.; mortality; multinomial models; precautionary savings; prediction; preferences; rational expectations; reservation wage; revealed preference theory; risk aversion; risk preference; stated preferences; subjective data; subjective probability; survey data, analysis of; symmetric loss functions; time preference; willingness to pay

Article

When economists analyse survey data, they must confront characteristics of the data-generating process that may distinguish these data from other types, such as administrative records.
Discussions of survey data analysis, such as the text by Chambers and Skinner (2003), typically consider problems caused by response errors, non-response, partially censored responses and complex sampling schemes. These problems are not unique to survey data: administrative records may not perfectly capture the phenomena of interest, may be incomplete, may be 'masked' to preserve confidentiality, and may be generated by something very different from simple random sampling or stratified sampling. Detailed discussions of methods for addressing such problems are included elsewhere in this dictionary and in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
standard econometrics texts.
In this article, we focus on survey data on subjective phenomena. Such data cannot be found among administrative records but are commonly reported in surveys. We discuss methods for interpreting subjective data and utilizing them in econometric analysis of individual behaviour, highlighting recent innovations in the measurement of expectations and preferences. The collection and analysis of subjective data was an important component of mainstream economic research until the mid-20th century, when it fell out of favour. Since then, the standard empirical method for the study of choice behaviour has been one of revealed preference analysis exclusively utilizing data on observed choices and attributes in combination with strong assumptions on expectations and preferences.
Yet the empirical basis for rejecting the use of subjective data was very limited (see, for example, Dominitz and Manski, 1997a; 2004). In fact, Tobin (1959, p. 11) concluded his generally negative analysis of expectations and attitudes data with a call to ‘investigate the questions [of] which attitudes are the most important ones to investigate in periodic surveys and what is the best way to use these data in combination with other economic information’. But this call went unheeded by economists, who ‘are taught early in their careers to believe only what people do, not what they say’ (Manski, 1990).
The prevailing scepticism has recently been challenged by researchers who seek to use data on subjective phenomena to weaken assumptions made on individual behaviour and to assess the credibility of maintained assumptions. To be viable, this approach requires careful design and administration of the survey instrument and proper interpretation of survey responses. Of particular concern are loosely worded survey questions that are subject to multiple interpretations by respondents and researchers alike. These concerns apply to the collection and analysis of all forms of survey data but are particularly salient in the context of subjective data. Importantly, even if subjective data suffer more severe response problems than do other forms of data, researchers must confront the limitations of the main alternative to directly measuring expectations and preferences, that is, making strong assumptions on the form of expectations and preferences in order to infer them from realizations data alone.
In this article, we demonstrate that the existing literature provides good reasons to believe that data on expectations and preferences reported in carefully designed household surveys can greatly enhance the empirical content of models of economic decisions and choices. In so doing, we endorse Manski's (2004) argument in favour of combining subjective data with other data to estimate models of choice behaviour, and we reiterate Tobin's (1959) call to determine the circumstances under which this combination is most fruitful.

Expectations

Published reports on consumer confidence indices are almost certainly the best-known output from analyses of household survey data on expectations. We begin by summarizing the history of consumer confidence measurement, an important component of the broader history of expectations data collection (see also Dominitz and Manski, 2004). We then discuss recent developments in the analysis of such qualitative expectations data and in methods for eliciting quantitative expectations in the form called for by modern economic theory – that is, subjective probability distributions.
Measures of consumer confidence were developed during the mid-20th century by George Katona and his colleagues at the University of Michigan's Survey Research Center, where much pioneering work in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic surveys has been conducted. The Index of Consumer Sentiment aggregates responses to expectations and attitudes questions asked in the monthly Survey of Consumers, with ordered categories that are coded as positive, neutral, or negative. Responses to the following question, for instance, were and still are included:

> Now looking ahead – do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?

The Federal Reserve Board formed a committee to assess the value of expectations and attitudes data soon after initiation of the Michigan surveys, which the Board funded. The committee produced negative findings on the ability of these and other consumer sentiment data to predict individual savings and consumption reported in follow-up interviews (Tobin, 1959). Katona (1957) and others argued that the indices predict aggregate economic outcomes. Over the next half-century, consumer confidence measures would become widely discussed indicators of the state of the economy and would find some use in macroeconomic studies of aggregate economic behaviour (Ludvigson, 2004). The qualitative data on expectations and attitudes that form the basis for these measures, however, were generally not thought to be of value for use in microeconomic studies of individual behaviour. Recent research has highlighted limitations that are generic to traditional, qualitative questions eliciting expectations.

The form of these expectations questions necessarily limits the predictive value of individual responses. Manski (1990) formally modelled respondents who report best predictions (that is, minimize expected loss) when asked yes/no expectations questions concerning future binary outcomes. These respondents say ‘yes’ if the subjective probability that the outcome will occur exceeds some threshold. Tobin (1959) and Juster (1966) had previously posited similar models of survey response. Manski derives sharp bounds on the correspondence between reported expectations and outcomes in a best-case scenario, where all respondents form rational expectations and minimize symmetric loss functions. The latter condition yields a threshold probability of 0.5. Even then, in the absence of an aggregate shock, all we would expect to find in a follow-up interview is that the outcome will occur for (*a*) more than half of all ‘yes’ respondents and (*b*) less than half of all ‘no’ respondents. Das, Dominitz, and van Soest (1999) extended Manski's model to ordered-category expectations of the form used in the consumer confidence questions and conducted a test of rational expectations using income expectations and realizations in a panel survey.

Dominitz and Manski (1997b) emphasize that the vague wording of many expectations questions further limits the interpersonal comparability and hence the predictive value of responses, as respondents must determine which possible outcomes would, for instance, constitute being ‘better off financially’, ‘worse off’, or ‘about the same’. Still, many researchers have formally modelled such qualitative responses, under strong identifying assumptions, and used the data to learn about expectations formation and the relationship between expectations and realizations. Pesaran and Weale (2006) review work that analyses the qualitative and quantitative reports of expectations.

The limitations generic to qualitative survey data on expectations need not apply to quantitative expectations data. Some surveys use questions of the ‘What do you expect…?’ form to elicit point expectations of continuous variables. This question format is typical in surveys of professional forecasters (see Keane and Runkle, 1990), but has also been used in household surveys. Bernheim

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(1989; 1990) studied expectations of retirement age and Social Security benefits reported in the Retirement History Survey, a survey that followed about 11,000 Americans aged 58–63 in 1969 through the 1970s. Lancaster and Chesher (1983) used point expectations of wage offers, in addition to individual reports of the subjective reservation wage, to identify a structural model of job search. These point expectations are typically interpreted as the mean of the subjective distribution, but other models of survey response are certainly plausible. Bernheim (1989), for instance, presents evidence that respondents report the mode of the subjective distribution of retirement age rather than the mean. To clarify the expectations of interest and to obtain information on uncertainty about prospective outcomes, economists designing surveys in the early 1990s began eliciting expectations in the form of subjective probabilities, as previously proposed by Juster (1966). Early examples are found in the Health and Retirement Study (HRS) survey, of which Tom Juster was the Principal Investigator, as well as the Survey of Economics Expectations (SEE), the Bank of Italy's Survey of Household Income and Wealth (SHIW), and the Dutch Center panel.

When the outcome of interest is binary, the probability of its occurrence summarizes the subjective probability distribution. Dominitz and Manski (1997b), for example, use SEE data to study economic insecurity arising from the prospective loss of a job or of health insurance coverage. When the outcome of interest takes on many values, a sequence of probabilities may be elicited to describe the subjective probability distribution. For instance, Hurd and McGarry (2002) use HRS data to study survival probabilities, which they find to be predictive of mortality. Dominitz (2001) demonstrates how SEE income expectations data can be fruitfully combined with income realizations data to estimate income expectations models of the type commonly adopted in research on consumption and savings, but allowing for greater heterogeneity. Guiso, Jappelli and Terlizzese (1992) use SHIW income expectations data to study the relationship between subjective uncertainty and precautionary savings. In each case, the authors impose parametric assumptions on the subjective probability distribution to identify the entire distribution from a handful of subjective probabilities.

Preferences

Standard econometric analysis of behaviour reported in a household survey combines data on household attributes and observed choices to make inferences on preference parameters and to generate predictions of choices not observed in the data. This application of *revealed preference* (RP) analysis typically requires strong assumptions on the expectations, choice sets, and preferences of households. In some cases, the data on choices and household attributes are sufficiently rich so as to yield credible inferences and predictions under weak assumptions. Many other situations, however, will require strong, untestable, and perhaps untenable assumptions. This may be the case, for instance, if the set of available alternatives varies considerably across households but the survey collects data only on the alternative selected by each household. The restrictiveness of the maintained assumptions may become particularly problematic when choices are made in a life cycle context with uncertainty, where subjective distributions of the future consequences of current choices are important.

The addition of survey questions on subjective phenomena may allow researchers to test and relax the maintained assumptions. First, questions on future expectations can be used, as discussed above. Second, respondents can be asked to make choices in hypothetical situations or to evaluate hypothetical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
opportunities, thereby providing *stated preference* (SP) data. Analysis of SP data has been commonplace for quite a long time in marketing research and transportation studies, but considerable scepticism remains in economics. Louviere, Hensher and Swait (2000) review the advantages and drawbacks of using SP data in place of or in combination with data on actual behaviour.

Consider hypothetical choice questions in which respondents are offered a number of alternatives and asked to choose one. Strong assumptions on choice sets required for analysing RP data can be avoided because the alternatives are explicitly given and thus observed by the researcher. This even applies to choices in a life cycle context where the question may specify the distribution of future outcomes associated with each alternative (for example, Barsky et al., 1997); therefore, the researcher may not need to elicit expectations or infer them from realizations data. Moreover, the range of offered alternatives can be manipulated to extract maximal information on preferences and to estimate preference parameters more efficiently than would be possible with RP data.

As with revealed preference analysis of discrete choice behaviour, hypothetical choices may be modelled using a random utility framework in which the utility of each alternative depends upon its specified attributes and an idiosyncratic disturbance term (see McFadden, 1986). For empirical implementation, the multinomial logit model has been the standard (McFadden, 1973), but more general models are available, such as multinomial probit to avoid the assumption of independence of irrelevant alternatives or mixed logit to account for unobserved heterogeneity across respondents (for example, Revelt and Train, 1998).

Still, even with hypothetical choice data some assumptions on choice sets or preference structures will be required, because it is not generally possible to fully specify the characteristics of each alternative. If so, then one may assume that unobserved attributes enter utility additively and are the same for all choice alternatives, so that they do not matter for the choice. Another possibility is to assume that the respondent takes an expectation over the distribution of unobserved attributes, given the information that is provided (Manski, 1999).

Much of the scepticism about SP data seems to arise from the extensive literature on problems with contingent valuation (CV) studies that seek to measure willingness to pay (WTP) for or willingness to accept in exchange for a non-market good, a frequent subject of analysis in environmental economics. In one type of CV survey, respondents get information about the current state (for example, of the environment) and a proposed change (for example, a quality improvement). They are then asked whether they would vote in favour of the change, given that this change would require a certain monetary cost (such as a tax increase). Survey responses are used to estimate the population distribution of the subjective valuation of the good at issue (see Mitchell and Carson, 1989).

Research has shown that CV studies can yield systematically misleading assessments of the value of a good (for example, Diamond and Hausman, 1994). Hanemann (1994) argues that many of the problems with CV arise from the way in which the survey is conducted. He concludes that the existing criticism is often justified but not valid in all circumstances: a well-designed and rigorously conducted CV study can yield reliable estimates of WTP. Note that some criticisms of the collection and analysis of these data actually question the basic assumptions of revealed preference analysis rather than simply the elicitation of preferences (for example, Ariely, Prelec and Loewenstein, 2003).

The problems with CV studies can also apply to SP experiments. For example, they can be sensitive to framing effects because respondents tend to choose the middle category, to anchoring or status quo bias

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
if one of the alternatives is clearly specified as the benchmark, or to yea-saying if questions have a yes or no format (Schwarz et al., 1985). Recent evidence suggests, however, that serious problems can largely be avoided by a careful and precise wording of the questions (Louviere, Hensher and Swait, 2000).

An obvious way to test and validate SP data is to compare them with RP data (see Louviere, Hensher and Swait, 2000). Among economic studies, Euwals, Melenberg and van Soest (1998) show that stated preferences for changes in working hours are predictive of changes in actual hours, and Kapteyn (1994) compares data on actual expenditures on certain commodities with data on subjective income evaluations, and finds that the two types of data are consistent with a common underlying utility function. The conclusion from studies such as these seems to be that RP and SP data are often consistent with each other, once an appropriate model is used to allow for different sources of idiosyncratic noise. In such a case, SP data can lead to more efficient estimates than RP data, and opportunities for combining the two data sources can be exploited to more accurately estimate parameters of interest.

Louviere, Hensher and Swait (2000) focus on static models of consumer choice, but the potential advantages of SP data also apply to dynamic models of consumer behaviour under uncertainty. Models of intertemporal choice have become increasingly important in many areas of empirical economics as researchers have gained access to rich panel data on households in the Panel Study of Income Dynamics and many subsequent national surveys in the United States and Europe.

Two central parameters of interest describe the agent's time preference and risk aversion. Standard RP analysis would yield inferences on these preference parameters based solely on, for example, observed savings and investment decisions. The rate of time preference indicates how the utility of current consumption is traded off for future consumption (for example, paid for with current savings). Risk aversion parameters indicate how risky alternatives (such as risky assets) are valued relative to less risky ones (such as risk-free assets). A standard identifying assumption in econometric models of intertemporal choice requires that time preferences and risk preferences be invariant across households.

In recent years, economists have used SP data to estimate the distribution of time and risk preference parameters in heterogeneous populations. Frederick, Loewenstein and O’Donoghue (2002) summarize empirical measures of rates of time preference (by economists and others) via RP and SP methods and use this to argue that the standard model of intertemporal choice (a single discount rate) is misspecified. Barsky et al. (1997) use SP data from an experimental module of the HRS. For risk preferences, the survey asks respondents to choose between two jobs, either the current job with the current income or a new job with higher expected income but also with some income risk. For time preferences, the survey asks respondents to make a sequence of choices over alternative consumption profiles across time. Their results show that SP data can be usefully applied to economic choices in a life cycle context, but they also pay attention to the potential pitfalls of using SP data with incompletely specified alternatives, such as the tendency to prefer the current job for reasons other than income.

Similar survey methods have been introduced in the Netherlands. Kapteyn and Teppa (2003) use SP data from the Dutch Center panel to estimate a structural model with habit formation in which preference parameters vary with individual attributes. Donkers, Melenberg and van Soest (2001) use SP data on choices between hypothetical lotteries to identify a structural model that generalizes a standard expected utility model by allowing for reference-dependent utility, loss aversion, and probability weighting.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Conclusion

We have described recent developments in the collection of subjective data on expectations and preferences and the estimation of behavioural models using these data. Combining data on expectations and preferences with each other and with data on actual choice behaviour is an important goal of this endeavour. An ambitious example is Erdem, Keane and Strebel (2005), who estimate a dynamic structural model of information acquisition and purchase decisions by consumers who choose among brands of personal computer. To identify the model, they utilize reports of price change expectations and stated assessments of brand quality in combination with data on actual purchase behaviour.
It is far too early to conclude that the general scepticism among economists about data on subjective phenomena has been overcome. However, we strongly believe that recent innovations in the measurement and analysis of expectations and preferences are enriching the empirical content of economic models and hold great promise for the future.

See Also

- bias correction
- categorical data
- household surveys
- longitudinal data analysis
- nonlinear panel data models
- partial identification in econometrics
- propensity score
- selection bias and self-selection
- stratified and cluster sampling

Bibliography

Ariely, D., Prelec, D. and Loewenstein, G. 2003. Coherent arbitrariness: stable demand curves without stable preferences. *Quarterly Journal of Economics* 118, 73–105.

Barsky, R., Juster, F., Kimball, M. and Shapiro, M. 1997. Preference parameters and behavioral heterogeneity: an experimental approach in the health and retirement survey. *Quarterly Journal of Economics* 112, 537–80.

Bernheim, B. 1989. The timing of retirement: a comparison of expectations and realizations. In *The Economics of Aging*, ed. D. Wise. Chicago: NBER and University of Chicago Press.

Bernheim, B. 1990. How do the elderly form expectations? An analysis of responses to new information. In *Issues in the Economics of Aging*, ed. D. Wise. Chicago: NBER and University of Chicago Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Chambers, R. and Skinner, C. 2003. *Analysis of Survey Data*. Hoboken, NJ: Wiley.

Das, M., Dominitz, J. and van Soest, A. 1999. Comparing predictions and outcomes: theory and application to income changes. *Journal of the American Statistical Association* 94, 75–85.

Diamond, P. and Hausman, J. 1994. Contingent valuation: is some number better than no number? *Journal of Economic Perspectives* 8(4), 45–64.

Dominitz, J. 2001. Estimation of income expectations models using expectations and realizations data. *Journal of Econometrics* 102, 165–95.

Dominitz, J. and Manski, C. 1997a. Using expectations data to study subjective income expectations. *Journal of the American Statistical Association* 92, 855–67.

Dominitz, J. and Manski, C. 1997b. Perceptions of economic insecurity: evidence from the Survey of Economic Expectations. *Public Opinion Quarterly* 61, 261–87.

Dominitz, J. and Manski, C. 2004. How should we measure consumer confidence? *Journal of Economic Perspectives* 18(2), 51–66.

Donkers, B., Melenberg, B. and van Soest, A. 2001. Estimating risk attitudes using lotteries – a large sample approach. *Journal of Risk and Uncertainty* 22, 165–95.

Erdem, T., Keane, M. and Strebel, J. 2005. Learning about computers: an analysis of information search and technology choice. *Quantitative Marketing and Economics* 3, 207–46.

Euwals, R., Melenberg, B. and van Soest, A. 1998. Testing the predictive value of subjective labour supply data. *Journal of Applied Econometrics* 13, 567–86.

Frederick, S., Loewenstein, G. and O'Donoghue, T. 2002. Time discounting and time preference: a critical review. *Journal of Economic Literature* 40, 351–401.

Guiso, L., Jappelli, T. and Terlizzese, D. 1992. Earnings uncertainty and precautionary saving. *Journal of Monetary Economics* 30, 307–37.

Hanemann, W. 1994. Valuing the environment through contingent valuation. *Journal of Economic Perspectives* 8(4), 19–43.

Hurd, M. and McGarry, K. 1995. Evaluation of the subjective probabilities of survival in the HRS. *Journal of Human Resources* 30, S268–S292.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Hurd, M. and McGarry, K. 2002. The predictive validity of subjective probabilities of survival. *Economic Journal* 112, 966–85.

Juster, F. 1966. Consumer buying intentions and purchase probability: an experiment in survey design. *Journal of the American Statistical Association* 61, 658–96.

Kapteyn, A. 1994. The measurement of household cost functions: revealed preference versus subjective measures. *Journal of Population Economics* 7, 333–50.

Kapteyn, A. and Teppa, F. 2003. Hypothetical intertemporal consumption choices. *Economic Journal* 113, C140–C152.

Katona, G. 1957. Federal Reserve Board Committee reports on consumer expectations and savings statistics. *Review of Economics and Statistics* 39, 40–6.

Keane, M. and Runkle, D. 1990. Testing the rationality of price forecasters: new evidence from panel data. *American Economic Review* 80, 714–34.

Lancaster, T. and Chesher, A. 1983. An econometric analysis of reservation wages. *Econometrica* 51, 1661–76.

Louviere, J., Hensher, D. and Swait, J. 2000. *Stated Choice Methods*. Cambridge: Cambridge University Press.

Ludvigson, S. 2004. Consumer confidence and consumer spending. *Journal of Economic Perspectives* 18 (2), 29–50.

Manski, C. 1990. The use of intentions data to predict behavior: a best-case analysis. *Journal of the American Statistical Association* 85, 934–40.

Manski, C. 1999. Analysis of choice expectations in incomplete scenarios. *Journal of Risk and Uncertainty* 19, 49–65.

Manski, C. 2004. Measuring expectations. *Econometrica* 72, 1329–76.

McFadden, D. 1973. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1986. The choice theory approach to market research. *Marketing Science* 5, 275–97.

Mitchell, R. and Carson, R. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Method. Baltimore: John Hopkins University Press.

Pesaran, M. and Weale, M. 2006. Survey expectations. In *Handbook of Economic Forecasting*, vol. 1, ed. G. Elliott, C. Granger and A. Timmermann. Amsterdam: North-Holland.

Revelt, D. and Train, K. 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics* 80, 647–57.

Schwarz, N., Hippler, H.-J., Deutsch, B. and Strack, F. 1985. Response scales: effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly* 49, 388–95.

Tobin, J. 1959. On the predictive value of consumer intentions and attitudes. *Review of Economics and Statistics* 21, 1–11.

How to cite this article

Dominitz, Jeff and Arthur van Soest. "survey data, analysis of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_A000229> doi:10.1057/9780230226203.1655

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

sustainability

Daniel W. Bromley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Sustainability concerns the specification of a set of actions to be taken by present persons that will not diminish the prospects of future persons to enjoy levels of consumption, wealth, utility, or welfare comparable to those enjoyed by present persons. Sustainability grows out of a need for intertemporal ethical rules when one generation can determine the endowment of natural and constructed capital that will be passed on to all subsequent generations. Economic models of sustainability seek axiomatic guidance for the selection of rules regarding natural resource use. Ecologists approach sustainability from a related – though not identical – ethical stance.

Keywords

exhaustible resources; Hartwick's rule; Hotelling, H.; intertemporal efficiency; justified commitment; liability rule; natural assets; precautionary principle; pure theory of savings; Ramsey, F.; safe minimum standard; sustainability

Article

The World Commission on Environment and Development (1987) popularized an idea that can be traced to Frank Ramsey's seminal work on the pure theory of savings (1928). Sustainable development meets '…the needs of the present without compromising the ability of future generations to meet their own needs'.
Some natural scientists discuss sustainability as it relates to human well-being. For instance, environmental sustainability

> seeks to improve human welfare by protecting the sources of raw materials used for human needs and ensuring that the sinks for human wastes are not exceeded, in order to prevent harm to humans. (Goodland, 1995, p. 3)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In a related article, Goodland and Daly (1996, p. 1003) assert that environmental sustainability means ‘maintaining natural capital’. The strict maintenance of natural capital has come to be called ‘strong sustainability’. There are also efforts to connect sustainability with the idea of the earth's carrying capacity:

> *Sustainability* characterizes any process or condition that can be maintained indefinitely without interruption, weakening, or loss of valued qualities. Sustainability is a necessary and sufficient condition for a population to be at or below carrying capacity … Carrying capacity … always embodies the concept of sustainability. (Daily and Ehrlich, 1996, p. 992).

Notice two key premises here. First, reflecting what Pigou called a ‘faulty telescopic faculty’, the absence of some form of collective action implies that the future is not secure if left to the whims of atomistic agents seeking their best advantage. Second, natural resources are an indispensable asset upon which the future depends. In addition to these two premises, Lele and Norgaard (1996) remind us of the central value judgement inherent in discussions about sustainability, namely, that we *ought* to care about the future.

Modelling sustainability

Following Ramsey's early work, Harold Hotelling (1931) developed the general theory of production from non-renewable resources. But economists started to devote consistent attention to the problem of sustainable production from nature only in the mid-1970s, when the *Review of Economic Studies* published a symposium issue on the subject (Dasgupta and Heal, 1974; Solow, 1974; Stiglitz 1974a; 1974b). An economy produces goods and services in the current period under constant returns to scale, by using previously constructed capital and a pool of labour, and by drawing down some increment of a non-renewable natural resource (such as oil, titanium, copper, magnesium) (Solow, 1986). If population is held constant, and if there is no technical change, Hotelling's condition asks that the shadow value of the natural resource left in the ground should show instantaneous rates of increase that exactly match the current value of the marginal product of the reproducible capital. If labour and constructed capital are fully employed, if the conditions of intertemporal efficiency are met, and if the economy invests in reproducible capital at each instant the exact amount (value) by which the non-renewable capital stock is being diminished, then society will just be able to maintain a constant stream of consumption into the infinite future. This latter investment policy concerning the augmentation of reproducible capital is known as Hartwick's rule (Hartwick, 1977; 1978a; 1978b). Martin Weitzman (1976) showed that the maximum welfare along a competitive path from any point into the future is formally identical to what might be obtained from the path of constant consumption.

Unavoidable inconveniences

The formal models of Ramsey, Hotelling, Dasgupta and Heal, Stiglitz, Solow, Hartwick and Weitzman

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
set out the essential logic of how economists think about sustainability. The models' austerity is their strength. But, William of Ockham notwithstanding, spare and austere models do not always offer coherent policy guidance. This recognition gave rise to two dominant concepts of sustainability in economics – weak and strong.

The early models deriving optimal trajectories of constant consumption (and utility) provided the conceptual basis for *weak sustainability* – replace used-up or degraded natural capital with constructed capital in order to assure continued consumption (utility). To the environmental community, weak sustainability is not sustainability at all. Indeed, a better term for this view of sustainability is derived from the realm of medicine, where we encounter organ transplants and replacement therapy. (Technically, replacement therapy is the regular introduction of necessary chemicals into the body that are no longer produced naturally.) To the environmentalist, *sustainability as replacement* resembles a doctor who says to you, 'go ahead and drink all you want, and when your liver or kidneys eventually fail we will give you mechanical ones, or we will put you on dialysis'.

Recall that for *some* aspects of nature it may not matter if we run out of them. That is, for many objects, it is the attributes they bring to us, not the unified bundle that we call 'coal' or 'titanium', that are the object of our interest. If one worries about the exhaustion of coal, that is simply an expression of concern about where energy will come from when the coal is gone. But the Grand Canyon and Victoria Falls are not coal or titanium. And this brings us to *strong sustainability*. Strong sustainability concerns specific bundles of attributes that are regarded as valuable in their own right; that is, they are valuable, not because of what they will *produce* for us, but because of what they *mean* to us. Their meaning is the cultural imprint we seek to project on to future persons as our legacy to them. The issue of strong sustainability concerns differing perceptions across a population – whether living now or in the future – of the imagined purposes of nature. What, exactly, is nature for? Its meaning *to* us is what it does *for* us. When we figure out what nature is for and craft policies accordingly, we are also prefiguring how future persons will come to see the natural assets we have bequeathed to them. We transmit, indeed we impose, our values on to future persons.

For those who worry about nature in terms of its meaning to contemporary and future persons, the idea of replacement sustainability is ethically defective. However, the term 'strong sustainability' is also unhelpful. Strong sustainability seems to suggest that all (or most) things must be locked up and preserved indefinitely. Few take it this far, but the terminology suggests as much. The central issue here is a justifiable concern for particular settings and circumstances that ought not to be compromised under any plausibly foreseeable circumstances. In fact, the matter can be put as *justified commitments* to the protection of specific settings and circumstances that must not to be compromised under any plausibly foreseeable circumstances. The concept of 'justified commitment' suggests that the advocates of lasting protection of particular settings and circumstances are under an obligation to offer a plausible and coherent justification to the wider political community. Why, exactly, does this particular setting or object deserve the sort of irrevocable protection you seek for it? Give me good reasons for doing so (Bromley, 2006). Notice that the concept of justified commitments requires the general agreement of a large number of others to whom the reasons advanced are found compelling. Only then can we say that the reasons advanced were sufficient. This is what it means to formulate public policy in a democracy.

Several important clarifications and elaborations have entered the literature since the early formal models were advanced (Asheim, 1994; Pezzey and Toman, 2002). A paper by Asheim, Buchholz and Tungodden (2001) derives the properties of a set of economic pathways that are non-decreasing and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
efficient. It suggests that, since any such path is sustainable, efficiency and equity can be cited to indict an unsustainable path as ethically flawed. Richard Howarth (1997; 1998) has offered a synthesis of much of the early work, and he writes that 'optimal' policies depend on ethical propositions that are impossible to resolve from within economics. We must look outside our models and our analytical apparatus for clear guidance. That is, we must turn to discussions of what is thought to be 'right' rather than what is thought to be 'good'. An intergenerational (intertemporal) golden rule offers some help, but it can be a weak thing indeed. Moral fibre is strongest when not threatened by the sharp blade of self-interest. Like Ulysses, we need to bind ourselves to *some* mast since we cannot be trusted to do the right thing for the future. The Sirens are too alluring. But which mast will provide sufficient restraint?

Some possible alternatives

One alternative concerns the *precautionary principle*: (*a*) safeguard options for future generations by protecting thresholds of resilience in desirable states of nature; and (*b*) contain the fundamental uncertainties associated with economic activity – either by restricting the level of that activity to preserve a degree of system predictability, or by containing the quantifiable risks associated with innovative activities that test the resilience of the system (Perrings, 2001, p. 196).

A related approach, advanced by S. V. Ciriacy-Wantrup, is the *safe minimum standard* (Ciriacy-Wantrup, 1968; Bishop, 1978; Woodward and Bishop, 1997). Unique natural assets – particular species, habitats, and situations crucial to ecosystem functions – must be protected unless the costs of doing so are too high. Unfortunately, there are no clear rules to determine whether a level of costs is 'too high'. But a focus on the safe minimum standard (SMS) of conservation serves to reframe the decision problem along the following lines: 'are you *really* sure you want to do that?' The SMS connects to the earlier idea of justified commitments.

In this vein, Bromley (1989) offers an approach that features the 'projection of rights' onto future persons. Specifically, he suggests two possible approaches: (*a*) an inalienability rule; and (*b*) a liability rule. Under inalienability, revered natural assets must be justified to a sufficiently large share of the polity as being inviolable. Think of this as *civic sanctification*. Once recognized, future persons hold inalienable rights, and members of all prior generations, including the present generation, are the reciprocal bearers of a duty towards those future right holders. This approach requires that irrevocable legal steps be taken by those of us now living to make sure that natural assets will be protected in the future. National parks, wilderness areas, and the US Endangered Species Act are of this sort – moral commitments towards future persons that are assured through organized and official legal obligation.

Under a liability rule there are two possible options. First, if those of us now living violate our commitment to the future – if we are unable to resist the Sirens – then we owe future persons some form of offsetting compensation. Perhaps we agree to set aside yet larger areas of wilderness, or make special efforts to increase the protection of an expanded list of endangered habitats or valuable species. Second, we atone for our violation of the interests of future persons by establishing an annuity that would pay dividends to future persons in some amount thought commensurate with the harm we have done them.

While neither approach adequately accounts for future uncertainty, they both acknowledge our obligation to the future, and both approaches force us to indemnify that obligation. If we are going to become richer by squandering some part of nature that morally belongs in their (future persons) bequest

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
package, then we need to set aside something to make future persons better off than they would otherwise be.

In a recent paper that builds on Bromley (1989; 1990) and Stern (1997), Gerlagh and Keyser (2001) advocate a 'trust fund policy'. The trust fund of Gerlagh and Keyser is a long-lived organization – comparable to a central bank – that oversees the entitlement of all members of present and future generations to an equal *claim* over natural resources. The arrangement would have precise rules of conduct to guarantee that members of future generations receive their claims independently of the preferences of the members of intervening generations.

See Also

- ethics and economics
- exhaustible resources
- Hotelling, Harold
- intertemporal equilibrium and efficiency
- precautionary principle
- Ramsey, Frank Plumpton

I am grateful for the comments of Richard C. Bishop and Richard T. Woodward.

Bibliography

Asheim, G. 1994. Net national product as an indicator of sustainability. *Scandinavian Journal of Economics* 96, 257–65.

Asheim, G., Buchholz, W. and Tungodden, B. 2001. Justifying sustainability. *Journal of Environmental Economics and Management* 41, 252–68.

Bishop, R. 1978. Endangered species and uncertainty: the economics of a safe minimum standard. *American Journal of Agricultural Economics* 57, 10–18.

Bromley, D. 1989. Entitlements, missing markets and environmental uncertainty. *Journal of Environmental Economics and Management* 17, 181–94.

Bromley, D. 1990. The ideology of efficiency: searching for a theory of policy analysis. *Journal of Environmental Economics and Management* 19, 86–107.

Bromley, D. 1998. Searching for sustainability: the poverty of spontaneous order. *Ecological Economics* 24, 231–40.

Bromley, D. 2006. *Sufficient Reason: Volitional Pragmatism and the Meaning of Economic Institutions*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Princeton: Princeton University Press.

Ciriacy-Wantrup, S. 1968. *Resource Conservation: Economics and Policy*. Berkeley: University of California Press.

Daily, G. and Ehrlich, P. 1996. Socioeconomic equity, sustainability, and earth's carrying capacity. *Ecological Applications* 6, 991–1001.

Dasgupta, P. and Heal, G. 1974. The optimal depletion of exhaustible resources. *Review of Economic Studies* 41, 3–28.

Gerlagh, R. and Keyzer, M. 2001. Sustainability and the intergenerational distribution of natural resource entitlements. *Journal of Public Economics* 79, 315–41.

Goodland, R. 1995. The concept of environmental sustainability. *Annual Review of Ecology and Systematics* 26, 1–24.

Goodland, R. and Daly, H. 1996. Environmental sustainability: universal and non-negotiable. *Ecological Applications* 6, 1002–17.

Hartwick, J. 1977. Intergenerational equity and the investing of rents from exhaustible resources. *American Economic Review* 67, 972–4.

Hartwick, J. 1978a. Substitution among exhaustible resources and intergenerational equity. *Review of Economic Studies* 45, 347–54.

Hartwick, J. 1978b. Investing returns from depleting renewable resource stocks and intergenerational equity. *Economic Letters* 1, 85–8.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39, 137–75.

Howarth, R. 1997. Sustainability as opportunity. *Land Economics* 73, 569–79.

Howarth, R. 1998. An overlapping generations model of climate-economy iterations. *Scandinavian Journal of Economics* 100, 575–91.

Lele, S. and Norgaard, R. 1996. Sustainability and the scientist's burden. *Conservation Biology* 10, 354–65.

Perrings, C. 2001. Modelling sustainable ecological–economic development. In *The International Yearbook of Environmental and Resource Economics, 2001/2002*, ed. H. Folmer and T. Tietenberg.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Cheltenham: Edward Elgar.

Pezzey, J. and Toman, M. 2002. Progress and problems in the economics of sustainability. In *The International Yearbook of Environmental and Resource Economics, 2001/2002*, ed. H. Folmer and T. Tietenberg. Cheltenham: Edward Elgar.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38, 543–49.

Solow, R. 1974. Intergenerational equity and exhaustible resources. *Review of Economic Studies* 41, 29–45.

Solow, R. 1986. On the intergenerational allocation of resources. *Scandinavian Journal of Economics* 88, 141–9.

Stiglitz, J. 1974a. Growth with exhaustible natural resources: efficient and optimal growth paths. *Review of Economic Studies* 41, 123–37.

Stiglitz, J. 1974b. Growth with exhaustible natural resources: the competitive economy. *Review of Economic Studies* 41, 139–52.

Stern, D. 1997. The capital theory approach to sustainability: a critical appraisal. *Journal of Economic Issues* 31, 145–73.

World Commission on Environment and Development [Bruntland Commission]. 1987. *Our Common Future*. Oxford: Oxford University Press.

Weitzman, M. 1976. On the welfare significance of a national product in a dynamic economy. *Quarterly Journal of Economics* 90, 156–62.

Woodward, R. and Bishop, R. 1997. How to decide when experts disagree: uncertainty-based choice rules in environmental policy. *Land Economics* 73, 492–507.

How to cite this article

Bromley, Daniel W. "sustainability." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000482> doi:10.1057/9780230226203.1656

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Svennilson, Ingvar (1908 – 1972)

Bjorn Hansson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

investment; planning; Stockholm School; structural transformation; Svennilson, Ingvar; technical change

Article

Svennilson was born and died in Stockholm. With a PhD in economics (1938) at the University of Stockholm, he was co-editor of *Ekonomisk Tidskrift* (1939–56); professor of economics at the University of Stockholm (1947–72) and vice-president of the University of Stockholm (1958–66). His dissertation (1938) tackled the problem of ‘planning in an unplanned world’ and is a theoretical analysis of the investment decisions taken by individual firms under risk and uncertainty. Plans are based on uncertain future events, which implies a choice of strategy in a multiperiod model. The strategy embodies the idea of alternative planning, which means that the firms have plans how to adapt their actions periodically in accordance with the realized events. The ideas and the concepts used in the dissertation are developed from earlier Swedish works on sequence analysis, such as Lindahl's ‘The rate of interest and the price level’ (1930).

Throughout his career Svennilson published many empirical and theoretical works on the problem of structural transformation and technical change. In Svennilson (1954) structural maladjustment plays the major role in the lagging and unstable growth of the European economy during the interwar period. He always tried to incorporate technical change in the analysis of capital formation, as in Svennilson (1956). In Svennilson (1964) he made an attempt to give a systematic understanding of ‘the residual factor’ in the production function, and the equilibrium is shown to be determined by the growth rate as well as the functional distribution of income.

Svennilson was for 20 years (1947–67) chairman and expert to the Long-Term Planning Commission and he obviously had a significant influence on these documents, which are published almost every fifth year as a Swedish Government Official Report; the reports play an important role in the debates on economic policy. From 1942 to 1952 he was director of the Industrial Institute for Economic and Social Research, which is funded by private industry. This organizational and intellectual background explains why Svennilson always encouraged closer cooperation between government and private research and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

planning.

Selected works

1933. (With G. Bagge and E. Bagge Lundberg.) *Wages, Cost of Living and National Income in Sweden 1860–1930*. London: P.S. King & Son.

1938. *Ekonomisk Planering. Teoretiska Studier*. Uppsala: Almqvist & Wiksell.

1954. *Growth and Stagnation in the European Economy*. Geneva: UN Economic Commission for Europe.

1956. Capital accumulation and national wealth in an expanding economy, In *25 Essays in Honour of Erik Lindahl*, Stockholm: Svenska Tryckeribolaget.

1964. Economic growth and technical progress: an essay in sequence analysis. In *The Residual Factor and Economic Growth*, Paris: OECD.

Bibliography

Lindahl, E. 1930. The rate of interest and the price level. In *Studies in the Theory of Money and Capital*. London: George Allen and Unwin.

Lundberg, E. 1972. Ingvar Svennilson: a note on his scientific achievements and a bibliography of his contributions to economics. *Scandinavian Journal of Economics* 74, 313–28.

How to cite this article

Hansson, Bjorn. "Svennilson, Ingvar (1908–1972)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000345>
doi:10.1057/9780230226203.1867

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Swadeshi movement

Susan Wolcott
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economics in India; Gandhi, M.; protection; swadeshi movement

Article

The word ‘swadeshi’ is derived from the Bengali, *svadesi*, or from the Sanskrit, *svadesin*. Literally, it means from ‘one's own country’ (Leadbetter, 1993, p. 95). It came to mean the use in India of Indian manufactures or services in preference to imported goods or European provided services. Whatever emotional support it may have enjoyed in India, as an economic movement it was effective only in its final phase.

Swadeshi became a rallying cry of Indian nationalists as early as the mid-19th century. In part, it was a response of the decline of the indigenous hand spinning and weaving industries as India became a major importer of British factory produced cotton cloth. Chandra (1966) notes that as early as the 1870s Indian politicians would appear at public meetings dressed in self-spun cloth. He writes, ‘the tide of swadeshi went on rising all over the country during the 15 years from 1880 to 1895’ (Chandra, 1966, p. 126). It was ‘inflamed’ by the removal of the revenue tariff on imported cotton cloth in 1878–82, which was believed to have been done in deference to Lancashire interests. It was further aroused by the imposition of an excise tax on Indian factory production in 1896, which was done to keep Lancashire and India on an even footing when a revenue tariff was reinstated.

But though Chandra may be correct about the degree of hostile feeling these moves engendered, in practice it is difficult to see any real economic effect. Imports of British cloth rose throughout this period. Also, though many indigenous factories started and advertised themselves as swadeshi enterprises, none were successful because they were swadeshi. Consider two examples. The grandson of Dwarkanath Tagore, the very successful Bengali entrepreneur, started a shipping firm to break the exclusive European monopoly of river navigation. The firm was a spectacular failure (Ray, 1984, p. 151). Tata, a Bombay entrepreneur, started the Swadeshi Mill in 1896, which was a successful textile mill. But many Indian cotton textile mills started in this period succeeded, including those owned and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
operated by Europeans. Chandra attributes the ineffectualness of the movement in this period to lack of support by prominent Indian cotton textile mill owners, who themselves feared exciting popular opposition to mill-manufactured cloth, and to their fear of arousing the animosity of their British machinery suppliers. He points out that the call for swadeshi did not even get the official endorsement of the Indian National Congress in this early phase (Ray, 1984, p. 141).

The next phase of the swadeshi movement occurred in Bengal between 1905 and 1908. In 1905, Lord Curzon, Viceroy of India, decided that the Bengal presidency was too large to administer, and separated it into two parts. Bengali intellectuals believed that this was at least in part a strategy to divide and conquer. The protest which followed evolved from agitation for the revocation of the partition to agitation ‘to enable the educated Bengalis to break out of the narrow confines of service and professions into the wider fields of commercial and industrial enterprise’ (Ray, 1984, p. 150). Ray argues that this movement was the culmination of several years of rising prices which had put economic pressure on middle class lawyers, administrators and teachers, most of whom were on fixed incomes. The deep resentment of educated Bengalis towards partition provided the issue which finally goaded them to action. A boycott of British goods was officially called on 7 August 1905 at a mass meeting in Calcutta. And, as before, swadeshi enterprises were begun. But this incarnation of swadeshi also remained the purview of the middle classes. Not only were the lower classes largely uninvolved, but the Marwaris, the wealthy caste of merchants, ‘prudently stayed away from these schemes’ (Ray, 1984, p. 164). The movement gradually lost coherence.

Mahatma Gandhi ushered in the final phases of the colonial swadeshi movement. The 1920 Congress at Nagpur for the first time passed a swadeshi resolution. It called on Indians to give up their titles, to shun official ceremonies, to withdraw from government colleges and schools, and to boycott courts, offices and legislative councils. Also, foreign goods should be boycotted, and support given to native enterprises. Finally, there was a call for ‘reviving handspinning in every house and handweaving on the part of millions of weavers who have abandoned their ancient and honourable calling for want of encouragement’ (Leadbetter, 1993, p. 110). This phase is also sometimes called the non-cooperation movement, and lasted from 1920 to 1922. Bose and Jalal (1998, p. 141) write that this boycott of British goods and institutions was much more effective than had been that of 1905. As evidence they point out that the Prince of Wales was greeted in Indian villages by closed shutters on his 1921 visit. Leadbetter (1993) also writes that the boycott of British textile goods was effective, in part because this time it received the support of at least some major mill owners. But none of these authors presents evidence of an economic impact. The one empirical study of the textile industry found no effect. Wolcott (1991) estimated an Indian import demand function for British cotton textile goods covering the years 1890–1914 and 1920–38. She finds no structural break after 1920. Gandhi called another boycott in 1930–32; this is referred to as the ‘civil disobedience movement’. Leadbetter writes that Indian business leaders were even more supportive during this phase, and Wolcott does find a significant, exogenous decline of 28 per cent in sales of British goods after 1930.

See Also

- India, economics in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Bose, S. and Jalal, A. 1998. *Modern South Asia: History, Culture, Political Economy*. London: Routledge.

Chandra, B. 1966. *The Rise and Growth of Economic Nationalism in India*. New Delhi: People's Publishing House.

Leadbetter, S. 1993. *The Politics of Textiles, the Indian Cotton-Mill Industry and the Legacy of Swadeshi 1900–1985*. New Delhi: Sage Publications.

Ray, R. 1984. *Social Conflict and Political Unrest in Bengal 1875–1927*. Delhi: Oxford University Press.

Wolcott, S. 1991. British myopia and the collapse of Indian textile demand. *Journal of Economic History* 51, 367–84.

How to cite this article

Wolcott, Susan. "Swadeshi movement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000458> doi:10.1057/9780230226203.1657

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Swan, Trevor W. (1914 – 1989)

Paul A. Johnson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The first professor of economics at the Australian National University, Trevor W. Swan was an Australian economist known for his foundational contributions to economic growth theory and the theory of economic policy in small open economies. Independently, he and Robert Solow simultaneously formulated what has become known as the Solow–Swan model of economic growth. He devised ‘the Swan diagram’ familiar to generations of students of international economics. A graduate of the University of Sydney, Swan became an Officer of the Order of Australia in 1988 ‘for service to education and to government particularly in the field of economics’.

Keywords

Cobb–Douglas functions; factor substitution; golden rule; Harrod–Domar growth model; international reserves; Klein, L. R.; neoclassical growth theory; open economy macroeconomics; returns to scale; Solow–Swan model; Swan, T. W.; Tinbergen, J.

Article

Trevor W. Swan is perhaps best known for Swan (1956), his foundational contribution to the theory of economic growth. This paper, and that of Solow (1956), relaxes the extant Harrod–Domar assumption of a fixed capital–labour ratio by replacing the fixed-proportions technology with one permitting factor substitution. While the exposition in Swan (1956) differs from that in Solow (1956), the basic economic model in the two papers is the same, and the Solow–Swan model, as it often called, marks the beginning of neoclassical growth theory. The model in Swan (1956) employs a Cobb–Douglas production function. However, the notes on the model, circulated prior to the paper's publication but after Swan's first public presentation of the model, and later published as Swan (2002), employ a more general constant returns to scale technology. Pitchford (2002) discusses Swan's first public presentation of the model and the relationship between the version in the notes and that in Swan (1956). Dixon (2003) compares the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

expositions of Solow (1956) and Swan (1956), and Solow (1997) discusses Swan's contribution to the Solow–Swan model. As is now well known, the model offers two central predictions, namely: (*a*) a change in the (exogenous) saving rate has a long-run effect on the level of output per worker but not on its growth rate; and, (*b*) in the absence of continued technological progress, the long-run growth rate of output per worker is zero. After demonstrating this ‘anti-accumulation, pro-technology’ result Swan (1956, p. 338) speculates that ‘… accumulation may give rise to external economies, so that the true social yield of capital is greater than any “plausible” figure based on common private experience’. Such externalities can now be found in models such as Romer (1986) and Lucas (1988). Swan (1964), which was written in 1960, develops the golden rule of economic growth. In this paper Swan also expresses his doubts about the utility of contemporaneous economic growth theory when dealing with applied economic development problems.

Swan also made early contributions to other areas including both macroeconomic model building and the macroeconomics of small open economies. Much of the motivation for his work seems to have been provided by his interest in macroeconomic policy issues. In Swan (1989), which was written in 1945, Swan specifies, calibrates and simulates a Keynesian model of the Australian economy. This was just the second such empirical macroeconomic model ever built – the first being that of Tinbergen and the third that of Klein. The model, sophisticated for its time, specifies consumption, investment and imports as functions of lagged income. Investment is constrained to be non-negative. Imports also include an autonomous component that depends on their domestic currency price. Prices move with marginal cost derived from the aggregate production function which, with a fixed level of technology and a fixed capital stock, also determines the level of employment. The exogenous variables are exports, public investment, and the autonomous part of imports. Simulation of the model shows that it replicates closely the Australian macro-experience of the 1930s. In what may be the first use of such a model for the purpose, Swan uses counterfactual simulations of his model to study policies that would have ‘preserved Australia from the Great Depression’.

In Swan (1960), which was written in 1953, and Swan (1963), which was written in 1955, Swan discusses the problem of simultaneously achieving both internal balance (full employment) and external balance (a current account balance of zero) in a small open economy with a fixed exchange rate. Swan (1963) introduces ‘the Swan diagram’ familiar to generations of students of international economics. This diagram, which still appears in international economics texts (Caves, Frankel and Jones, 1999, p. 323 f, for example), shows the internal and external balance schedules with the real exchange rate on the vertical axis and domestic demand on the horizontal one. The internal balance schedule is negatively sloped because real depreciation (a rise in the real exchange rate), and the consequent increase in net exports, is necessary to maintain internal balance as domestic demand falls. The external balance is positively sloped as real appreciation, and the consequent decrease in net exports is necessary to maintain external balance as domestic demand, and hence imports, fall. The curves divide the plane ‘into four zones of economic unhappiness’. An economy in the zone below both curves, for example, suffers from both unemployment and a current account deficit. The intersection of the two curves gives the combination of real exchange rate and domestic demand consistent with both internal and external balance. The correct policy response when away from this intersection depends on the zone in which the economy lies and its location in the zone. For example, in the zone below both curves, if the economy is left of the intersection real depreciation and an increase in domestic demand is appropriate but, right of the intersection, real depreciation and a reduction in domestic demand is called for. After noting that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
both targets are likely to be more responsive to changes in demand than to changes in the real exchange rate in the short-run, Swan suggests tying the real exchange rate to a long-run value and using international reserves as a buffer against temporary external imbalances.

Swan was born in Sydney, Australia, and graduated from the University of Sydney in 1940 with a first-class honours degree in economics and a university medal. He held various government positions during and after the Second World War before being appointed as the first professor of economics at the Australian National University in 1950, a position he held until he retired in 1983. In addition to this, and other academic appointments, he was active in Australian economic policymaking until the mid-1980s, serving as an advisor to many post-war Australian governments and on the Board of the Reserve Bank of Australia. In 1988 Swan became an Officer of the Order of Australia ‘for service to education and to government particularly in the field of economics’. Butlin and Gregory (1989) and P. Swan (2006) provide more extensive biographies of Swan, including bibliographies of his published and unpublished works. This article owes much to both.

Selected work

1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

1960. Economic control in a dependent economy. *Economic Record* 36, 51–66.

1963. Longer-run problems of the balance of payments. In *The Australian Economy*, ed. H. Arndt and W. Corden. Melbourne: Cheshire.

1964. Of golden ages and production functions. In *Economic Development with Special Reference to East Asia*, ed. K. Berrill. London: Macmillan.

1989. The principle of effective demand – a ‘real life’ model. *Economic Record* 65, 378–98.

2002. Economic growth. *Economic Record* 78, 375–80.

Bibliography

Butlin, N. and Gregory, R. 1989. Trevor Winchester Swan 1918–1989. *Economic Record* 65, 369–77.

Caves, R., Frankel, J. and Jones, R. 1999. *World Trade and Payments*, 8th edn. New York: Addison Wesley.

Dixon, R. 2003. Trevor Swan on equilibrium growth with technological progress. *Economic Record* 79, 487–90.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Pitchford, J. 2002. Trevor Swan's 1956 economic growth ‘seminar’ and notes on growth. *Economic Record* 78, 381–7.

Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94, 1002–37.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Solow, R. 1997. Trevor Winchester Swan. In *An Encyclopedia of Keynesian Economics*, ed. T. Cate, G. Harcourt and D. Colander. Cheltenham: Edward Elgar.

Swan, P. 2006. Trevor Winchester Swan AO. ANU Inaugural Trevor Swan Distinguished Lecture, Australian National University, 23 May. Online. Available at http://www.crawford.anu.edu.au/pdf/seminars/trevor_swan.pdf, accessed 2 March 2007.

How to cite this article

Johnson, Paul A. "Swan, Trevor W. (1914–1989)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000530>
doi:10.1057/9780230226203.1658

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Sweezy, Paul Marlor (1910 – 2004)

John Bellamy Foster
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Harvard-trained economist and co-editor of *Monthly Review*, Paul Sweezy was among the most influential economists and Marxist intellectuals of the 20th century. His contributions extended over six decades from the early 1930s to the early 1990s. He played a role in the development of imperfect-competition analysis and in debates surrounding the Great Depression. His *Theory of Capitalist Development* (1942) provided the premier exposition of Marxian economics, after Marx. *Monopoly Capital* (1966, with Paul Baran) was the most influential economic analysis emanating from the US New Left. With Harry Magdoff he extended this analysis into the 1970s, ’80s and early ’90s.

Keywords

American Economic Association; Association for Evolutionary Economics; Baran, P.; Bortkiewitz, L. von; concentration and centralization of capital; Dobb, M.; falling rate of profit; Foster, J. Bellamy; Great Depression; Huberman, L.; imperfect competition; Kalecki, M.; Keynesian Revolution; kinked demand curve; Leontief, W.; Magdoff, H.; Marx, K.; Marxism; monopoly capitalism; multinational corporations; New Deal; O'Conner, H.; oligopolistic pricing; realization crisis; Schumpeter, J.; Socialism; Soviet Union, economics in; stagnation controversy; Steindl, J.; Stone, I.; surplus; Sweezy, P.; technological innovation; transformation problem; underconsumptionism; Underdevelopment; Union for Radical Political Economics; Wood, E. Meikins; world system theory

Article

One of the leading figures in Western Marxism in the 20th century and co-editor from 1949 to 2004 of *Monthly Review*, Sweezy is known both for his leading contributions to economics and his influence on the development of socialist thought.
Born on 10 April 1910 in New York, the son of a vice-president of the First National Bank of New York (part of J.P. Morgan's financial empire), he obtained his early education at Exeter and Harvard

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

University. At Harvard he was president of the *Harvard Crimson* and studied economics, receiving his BA (*magna cum laude*) in 1932. In 1932 he left Cambridge, Massachusetts, for a year of graduate study at the London School of Economics. Awakened by the Great Depression, and responding to the intellectual ferment in Britain during what was to be a turning point in world history, Sweezy gained sympathy for the Marxist perspective to which he was introduced for the first time. Returning to the United States in 1933 to do graduate studies at Harvard, he found the academic climate much changed, with Marxism becoming a topic of intense interest in some of the larger universities. As he recalled many years later,

> It was under these circumstances that I acquired a mission in life, not all at once and self-consciously but gradually and through a practice that had a logic of its own. That mission was to do what I could to make Marxism an integral and respected part of the intellectual life of the country, or, put in other terms, to take part in establishing a serious and authentic North American brand of Marxism. (1981a, p. 13).

In pursuing this goal at Harvard, Sweezy received direct help and indirect inspiration from the great conservative economist Joseph Schumpeter, whose analysis of the origins, development and imminent decline of capitalism revealed a complex, critical appreciation of the Marxian schema. Sweezy's 1943 essay on ‘Professor Schumpeter's Theory of Innovation’, which compared Schumpeter's analysis of entrepreneurial development to Marx's theory of accumulation, was to be one of the path-breaking studies in this area. In the 1930s, while in his twenties, Sweezy published more than 25 major publications – books, articles, government reports, and reviews – on economic topics. He was a member of the founding editorial board of the *Review of Economic Studies*, originally established in 1933 to promote the work of younger economists.

Imperfect competition, economic stagnation, and war

Receiving his Ph.D. in 1937, Sweezy assumed a position as instructor at Harvard until 1939, when he rose to the rank of assistant professor. During these years he played a central role in two of the major areas of debate in economics: (*a*) the theory of imperfect competition, and (*b*) the issue of secular stagnation. Sweezy's interest in the monopoly question began early in his career, as shown by his first book (winner of the David A. Wells prize), *Monopoly and Competition in the English Coal Trade, 1550–1850* (1938a). His article ‘Demand Under Conditions of Oligopoly’ (1939a), in which he presented the kinked demand curve analysis of oligopolistic pricing, remains one of the classic essays in modern price theory. Along with a small group of Harvard and Tufts economists, Sweezy was one of the authors and signatories of the influential Keynesian tract, *An Economic Program for American Democracy* (1938b), which provided a rationale for a sustained increase in public spending during the final years of the New Deal. While continuing to carry out his teaching responsibilities at Harvard, Sweezy worked for various New Deal agencies (including the National Resources Committee and the Temporary National Economic Committee) investigating the concentration of economic power. His study, ‘Interest Groups in the American Economy’, published as an appendix to the NRC's well-known report, *The Structure of the American Economy* (1939b), was to be an important guide to later research. A famous Harvard debate between Sweezy and Schumpeter (over which Wassily Leontief presided) on the stagnation controversy

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of the late 1930s was recounted decades later by Paul Samuelson, who had been in the audience, as one of the great events of his own lifetime (Samuelson, 1969).

From the lecture notes to his Harvard course on the economics of socialism, Sweezy produced his seminal work, *The Theory of Capitalist Development* (1942). Containing a comprehensive review of Marxian economics up until the time of the Second World War, this study also did much to determine the character of later Marxian theory through its advocacy of Laudislau von Bortkiewitz's solution to the ‘transformation problem’, its presentation of a logically acceptable ‘underconsumptionist’ model of accumulation and crisis, and its elaboration of Marxian views on monopoly capitalism. Rapidly translated into several languages, *The Theory of Capitalist Development* soon established Sweezy's reputation as the foremost Marxian economist of his generation.

During the Second World War Sweezy served in the Office of Strategic Services (OSS) and was assigned to the monitoring of British plans for post-war economic development. With a number of years still remaining in his Harvard contract when the war ended, he opted to resign his position rather than resume teaching, recognizing that his political and intellectual stance would hinder his receiving tenure. In this period, Sweezy authored numerous articles on the history of political economy and socialism, some of which were reprinted in his book, *The Present as History* (1953), and edited a volume (1949) containing three classic works on the ‘transformation problem’: *Karl Marx and the Close of His System* by Eugene Böhm-Bawerk, *Böhm-Bawerk's Criticism of Marx* by Rudolf Hilferding, and ‘On the Correction of Marx's Fundamental Theoretical Construction in the Third Volume of *Capital*’ by Ladislaus von Bortkiewicz (which Sweezy translated into English). His 1950 critique (in Sweezy et al., 1976) of Maurice Dobb's *Studies in the Development of Capitalism*, in which Sweezy, following his interpretation of Marx, emphasized the role of the world market in the decline of feudalism, launched the famous debate over the transition from feudalism to capitalism which has played a key role in Marxian historiography ever since. Sweezy's work in this area established him as one of the pioneers of what came to be known as ‘world system theory’.

McCarthyism and the early Monthly Review years

With the financial backing of literary critic F.O. Matthiessen, Sweezy and the Marxist historian Leo Huberman founded *Monthly Review* (subtitled ‘An Independent Socialist Magazine’) in 1949 as an intellectual resource for a US left threatened by anti-Communist hysteria. Albert Einstein wrote his article ‘Why Socialism?’ for the first issue. Two years later they began publishing books under the imprint of Monthly Review Press, when it came to their attention that in the repressive climate of the times even such celebrated authors as I.F. Stone and Harvey O'Conner were unable to find publishers for their book manuscripts.

In 1953, at the height of the McCarthyite period in the United States, the state of New Hampshire conferred wide-ranging powers on its Attorney General to investigate ‘subversive activities’. On this basis, Sweezy was summoned to appear before the state Attorney General on two occasions in 1954. Adopting a principled opposition to the proceedings, he refused to answer questions regarding (*a*) the membership and activities of the Progressive Party, (*b*) the contents of a guest lecture delivered at the University of New Hampshire, and (*c*) whether or not he believed in Communism. As a result, he was declared in contempt of court and consigned to the county jail until purged of contempt by the Superior Court of Merimack County, New Hampshire. His case was appealed and worked its way through the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
state and federal courts to the US Supreme Court, which ruled in his favour in *Sweezy v. New Hampshire* on 17 June 1957 (US Supreme Court, 1956). *Sweezy v. New Hampshire* is remembered historically as having dealt the death knell to the witch hunt on US campuses.

Despite the adverse ideological climate, Sweezy continued to author articles on all aspects of Marxian theory, adding up to hundreds of essays by the 1990s. Publication of Paul Baran's book, *The Political Economy of Growth* (1957), marked the beginning of Marxian dependency theory and helped establish *Monthly Review*'s primary identity as a backer of Third World liberation struggles. Visiting Cuba shortly after the revolution, Huberman and Sweezy co-authored two influential works on the transformation of Cuban economic society: *Cuba: Anatomy of a Revolution* (1960) and *Socialism in Cuba* (1969).

Monopoly Capital: an essay on the American economic and social order

The appearance in 1966 of *Monopoly Capital* by Baran and Sweezy (published two years after Baran's death) represented a turning point in Marxian economics. Although described by the authors as a mere ‘easy-sketch’, it rapidly gained widespread recognition as the most important attempt thus far to bring Marx's *Capital* up to date, as well as providing a formidable critique of prevailing Keynesian orthodoxy. Where Sweezy himself was concerned, *Monopoly Capital* reflected dissatisfaction with the analysis of accumulation and crisis advanced in *The Theory of Capitalist Development*. His earlier study had been written when mainstream economics was undergoing rapid change due to the Keynesian ‘revolution’ and the rise of imperfect competition theory. Thus, he had provided a detailed elaboration of both Marx's theory of realization crisis (or demand-side constraints in the accumulation process), and of work by Marx and later Marxian theorists on the concentration and centralization of capital. As with mainstream theory, however, these two aspects of Sweezy's analysis remained separate. Hence he failed to develop an adequate explanation of the concrete factors conditioning investment demand in an economic regime dominated by the modern large enterprise. It was essentially this critique of Sweezy's early efforts that was provided by Josef Steindl in *Maturity and Stagnation in American Capitalism* (1952, pp. 243–46), who went on to show how a more unified theory could ‘be organically developed out of the underconsumptionist approach of Marx’ based on Michal Kalecki's model of capitalist dynamics, which had connected the phenomenon of realization crisis to the increasing ‘degree of monopoly’ in the economy as a whole.

In fact, it was out of this argument, as outlined by Steindl, that the underlying framework for Baran and Sweezy's own contribution in *Monopoly Capital* was derived. What emerged in their analysis was ‘a Kaleckian version of Marxism, in which monopolist corporations control markets but have problems recovering their surplus because of their underinvestment’ (Toporowski, 2004).

In this analysis Marx's fundamental ‘law of the tendency of the rate of profit to fall’ associated with accumulation in the era of free competition had been replaced, in the more restrictive competitive environment of monopoly capitalism, by a law of the tendency of the surplus to rise (with ‘surplus’ defined as the gap, at any given level of production, between output and socially necessary costs of production). Under these circumstances, the critical economic problem was one of surplus absorption. Capitalist consumption tended to account for a decreasing share of capitalist demand as income grew, while investment was hindered by the fact that it took the form of new productive capacity, which could not be expanded for long periods of time independently of final, wage-based demand. Although there

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
was always the possibility of new ‘epoch-making innovations’ emerging that would help absorb the potential economic surplus, all such innovations – resembling the steam engine, the railroad and the automobile in their overall effect – were few and far between. Hence, Baran and Sweezy concluded that the monopoly capitalist system had a powerful tendency toward stagnation, rooted in long-term underinvestment, largely compensated for thus far through the promotion of economic waste by means of ‘the sales effort’ (including its penetration into the production process) and military expenditures, and through the expansion of the financial sector. All such ‘countervailing influences’ were, however, of a self-limiting character and would lead to a doubling-over of contradictions in the not-too-distant future. *Monopoly Capital*'s international influence was enormous, making Baran and Sweezy arguably the best known US economists outside the United States. The publication of *Monopoly Capital* coincided with the rise of the New Left, largely in response to the Vietnam War. The work of Baran and Sweezy thus constituted the initial theoretical common ground for a younger generation of radical economists in the United States who formed the Union for Radical Political Economics in 1968. In 1971, Sweezy delivered the Marshall Lecture at Cambridge University. From 1974 to 1976 he served on the executive of the American Economic Association.

The re-emergence of conditions of relative economic stagnation in the 1970s, not long after *Monopoly Capital* was published, confirmed much of Baran and Sweezy's analysis, pointing to the importance of many of the factors that they had emphasized, including: increasing aggregate concentration; the ascendance of multinational corporations; global surplus capacity; the clustering of technological innovations; continuing high-level military spending; intensification of the sales effort; and the soaring of finance and speculation. Yet, despite the fact that *Monopoly Capital* had both foreseen the advent of stagnation and highlighted many of the key historical factors that increasingly attracted attention as the 1970s and 1980s unfolded, Baran and Sweezy's magnum opus lost much of its lustre among left economists within 20 years of its publication. It seemed to many to be a product of the period of the unquestioned domination of US corporations and undisputed US economic hegemony, rather than the more globalized competitive environment that followed. In the long run, however, the theory continued to gather adherents and was seen as representing a crucial stage in the analysis of the global political economy (see Foster, 2002).

With Harry Magdoff (who replaced Huberman as co-editor of Monthly Review after the latter's death in 1968), Sweezy continued to strengthen the analysis begun in *Monopoly Capital* in the decades following its publication, utilizing the original framework to explain the re-emergence of stagnation and the rise of financial instability, as well as conditions of Third World underdevelopment, in *The Dynamics of US Capitalism* (1970), *The End of Prosperity* (1977), *The Deepening Crisis of US Capitalism* (1979), *Four Lectures on Marxism* (1981a), *Stagnation and the Financial Explosion* (1987), and *The Irreversible Crisis* (1988). In all of these works, the role of financial expansion as a means of utilizing the surplus and countering stagnation, while simultaneously creating growing problems of financial instability, was emphasized – well before this became a central concern in economics as a whole. If there was one criticism to be levelled at *Monopoly Capital*, as Sweezy pointed out again and again, it was that it had not placed enough emphasis on the ballooning of the financial superstructure of the capitalist economy as a response to stagnation – though *Monopoly Capital* had briefly addressed this problem at the end of the chapter on the sales effort. In ‘The Triumph of Financial Capital’ (1994, p. 2) he declared: ‘In earlier times no one ever dreamed that speculative capital, a phenomenon as old as capitalism itself, could grow to dominate a national economy, let alone the whole world. But it has.’

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Socialism and ecology

Sweezy's influential analysis of post-revolutionary societies, developed over half a century, maintained a critical distance from these societies, emphasizing their class-exploitative nature. Although the Soviet Union had been brought into being by a revolution under genuine socialist leadership, he argued, it had not continued down the socialist track and had resurrected class relations – giving it an inner logic that was neither capitalist nor socialist. Hence, Sweezy designated it as a political-economic system *sui generis*, which he labelled *Post-Revolutionary Society*—the title of a compendium (1981b) of his writings in this area. As early as 1971 in *On the Transition to Socialism*, which grew out of a lengthy debate with Charles Bettelheim, he wrote:

> When the bureaucratically administered economy runs into difficulties (as it certainly must), there are two politically opposite ways in which a solution can be sought. One is to weaken the bureaucracy, politicize the masses, and entrust increasing initiative and responsibility to the workers themselves. This is the road forward to socialist relations of production. The other way is to put increasing reliance on the market (as was the case with the New Economic Policy under Lenin) but as an ostensible step toward a more efficient 'socialist' economy. This is in fact to elevate profit-making to the guiding role in the economic process and to tell the workers to mind their own business, which is to work hard so that they can consume more. It is to recreate the conditions in which commodity fetishism flourishes along with its associated false and alienated consciousness. It is, I submit, the road back to class domination and ultimately the restoration of capitalism. (1971, pp. 53–4)

In the concluding paragraph of his *Post-Revolutionary Society*, written five years before the rise of Gorbachev, Sweezy declared that the Soviet system had 'entered a period of stagnation, different from the stagnation of the advanced capitalist world but showing no more visible signs of a way out'.
Indeed, Sweezy argued beginning in the early 1960s – most notably in *Modern Capitalism and Other Essays* (1972) – that those interested in the future of socialism should look for leadership not at present to the working class in the advanced capitalist states, nor to the new class societies of the Soviet Union and eastern Europe, but rather to the insurgent populations of the periphery of world capitalism. It was here, in Third World liberation struggles, if anywhere, that the revolutionary proletariat in the Marxist sense ('the focal point of all inhuman conditions') continued to struggle in the name of humanity itself.
From the 1970s on Sweezy was to turn frequently to the environmental issue, first with his 1973 article on 'Cars and Cities' and later notably in articles on 'Capitalism and the Environment' (1989a) and 'Socialism and Ecology' (1989b). If socialism did not 'guarantee' the salvation of the environment, he argued, it was nonetheless a necessary condition for its salvation (1990).
During the 1990s Monthly Review went through transformations in its editorial leadership, with Ellen Meiksins Wood joining Sweezy and Magdoff as co-editor of the magazine for three years (1997–2000), followed by John Bellamy Foster and Robert W. McChesney. Yet *Monthly Review* retained its essential identity, with Sweezy remaining co-editor despite declining direct involvement in his final years. He died of congestive heart failure at his home in Larchmont, New York, on 27 February 2004.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In 1983 Sweezy was granted an honorary doctorate of literature from Jawaharlal Nehru University in India. In 1999 he received the Veblen–Commons Award from the Association for Evolutionary Economics. The results of a survey of its readers by the *Post-Autistic Economics Review* in Issue no. 36 (2006) listed him as one of the 15 ‘greatest twentieth-century economists’. Indeed, as Janek Toporowski had already observed in the *Royal Economic Society Newsletter* in April 2004, ‘If he is to be measured by the impact of his ideas, then he stands among the greatest economists of the twentieth century.’

See Also

- Baran, Paul Alexander
- Braverman, Harry
- Marx, Karl Heinrich
- monopoly capitalism
- socialism (new perspectives)

Selected works

1938a. *Monopoly and Competition in the English Coal Trade, 1550–1850*. Cambridge, MA: Harvard University Press.

1938b. (With R. Gilbert et al.) *An Economic Program for American Democracy*. New York: Vanguard.

1939a. Demand under conditions of oligopoly. *Journal of Political Economy* 47, 568–73.

1939b. Interest groups in the American economy. In US National Resources Committee, *The Structure of the American Economy*, Pt I. Washington, DC: US Government Printing Office.

1942. *The Theory of Capitalist Development: Principles of Marxian Political Economy*. New York: Monthly Review Press.

1943. Professor Schumpeter's theory of innovation. *Review of Economics and Statistics* 25, 93–96.

1949. ed. *Karl Marx and the Close of His System by Eugen Böhm-Bawerk and Böhm-Bawerk Criticism of Marx by Rudolf Hilferding*. London: Merlin Press, 1976.

1953. *The Present as History*. New York: Monthly Review Press.

1960. (With L. Huberman.) *Cuba: Anatomy of a Revolution*. New York: Monthly Review Press.

1966. (With P. Baran.) *Monopoly Capital: An Essay on the American Economic and Social Order*. New

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

York: Monthly Review Press.

1969. (With L. Huberman.) *Socialism in Cuba*. New York: Monthly Review Press.

1970. (With H. Magdoff.) *The Dynamics of US Capitalism*. New York: Monthly Review Press.

1971. (With C. Bettelheim.) *On the Transition to Socialism*. New York: Monthly Review Press.

1972. *Modern Capitalism and Other Essays*. New York: Monthly Review Press.

1973. Cars and Cities. *Monthly Review* 24(11), 1–18.

1976. (With others.) *The Transition from Feudalism to Capitalism*. London: New Left Books.

1977. (With H. Magdoff.) *The End of Prosperity*. New York: Monthly Review Press.

1979. (With H. Magdoff.) *The Deepening Crisis of U.S. Capitalism*. New York: Monthly Review Press.

1981a. *Four Lectures on Marxism*. New York: Monthly Review Press.

1981b. *Post-Revolutionary Society*. New York: Monthly Review Press.

1987. (With H. Magdoff.) *Stagnation and the Financial Explosion*. New York: Monthly Review Press.

1988. (With H. Magdoff.) *The Irreversible Crisis*. New York: Monthly Review Press.

1989a. Capitalism and the environment. *Monthly Review* 41(2), 1–15.

1989b. Socialism and ecology. *Monthly Review* 41(4), 1–8.

1990. Nineteen Eighty Nine. *Monthly Review* 41(11), 18–21.

1994. The triumph of financial capital. *Monthly Review* 46(2), 1–11.

Bibliography

Baran, P. 1957. *The Political Economy of Growth*. New York: Monthly Review Press.

Foster, J. 1986. *The Theory of Monopoly Capitalism*. New York: Monthly Review Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Foster, J. 2002. Monopoly capital and the new globalization. *Monthly Review* 53(8), 1–16.

Foster, J. 2004. The commitment of an intellectual: Paul M. Sweezy (1910–2004). *Monthly Review* 56 (5), 5–39.

Samuelson, P. 1969. Memories. *Newsweek*, 2 June.

Steindl, J. 1952. *Maturity and Stagnation in American Capitalism*. New York: Monthly Review Press, 1976.

Toporowski, J. 2004. Paul M. Sweezy. *Royal Economic Society Newsletter*, No. 125, 17–18.

US Supreme Court. 1956. Sweezy *v*. New Hampshire. *US Reports* 354, October Term.

How to cite this article

Foster, John Bellamy. "Sweezy, Paul Marlor (1910–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000350> doi:10.1057/9780230226203.1659

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Swift, Jonathan (1667 – 1745)

S.C. Stimson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bastiat, C. F.; legal tender; mercantilism; Swift. J.; Young, A.

Article

Dean of St Patrick's Dublin, the austere Rabelais, the party pamphleteer from whom Rousseau learnt to detest politics and society, the high churchman from whom Voltaire and Lessing learnt their religion, the author of *Gulliver's Travels*, Swift is a writer to whose economic views critics are often unjust. *The Humble Petition of the Colliers, Cooks, Cook-maids*, etc., against the use of focused rays by a supposed company instead of fires, represents that this ‘will utterly ruin … your petitioners … and trades on them depending, there being nothing left to them after the said invention but warming of cellars and dressing of suppers in the wintertime’. And ‘whereas the said’ company ‘talk of making use of the moon by night as of the sun by day, they will utterly ruin the numerous body of tallow chandlers’, and so the tallow tax will fail. The fame of Bastiat is chiefly based on his expansion of this parable in the seventh of his *Sophismes Economiques* (1846), of which his admirers still say ‘nothing is more brilliant, nothing more French’.

Swift's *Maxims controuled in Ireland*, suggested perhaps by Sir William Temple (*Works*, 1814 edn, vol. i. p. 177), exposes, after the manner of Bastiat, popular economic fallacies which deceived Temple, Locke, and Child, whom he had studied; e.g. that a large population, high prices for land, dear provisions, and big towns (cf. Barbon) *must* imply wealth, and that low interest *must* be due to much money. For ‘must’, he says, you should write ‘may’, thus, in trading countries like Holland and England, low interest and high capital values for land were effects of the causes alleged, but in Ireland of the absence of trade, and therefore of a demand for loans. He perceived ‘that in the arithmetic of the customs two and two, instead of making four, make sometimes only one’ (Smith, *Wealth of Nations*, book v, ch. ii).

Otherwise Swift belongs to his age. Thus the king of Brobdingnag's belief that ‘whoever could make two ears of corn or two blades of grass to grow … where only one grew before would deserve better of mankind … than the whole race of politicians’, and the echo of this belief at the end of the last of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Drapier's Letters resembles Molesworth's ideas (1723), and afterwards became a favourite motto with Arthur Young. He thinks with Locke that traces fall mainly upon land, which, like Harrington, he overrates (*Works*, 1824 edn, vol. iii, p. 518); and his anger against ploughlands being turned into sheep-runs makes him akin to Latimer, Boulter (*Letters*, 24 February 1727), and, as he himself said, to Ajax. He pillories the trading spirit in his abuse of the Dutch; and wishes a weavers' corporation to regulate prices and qualities, and to punish offenders by 'warnings' (*Works*, vol. vii, pp. 49, 50, 137). He denounces 'the restriction' and urges Irishmen to raise by way of reply what Berkeley called 'a wall of brass a thousand cubits high' round Ireland; and thinks that this could be done by a resolution to consume home-made goods instead of 'unwholesome drugs and unnecessary finery' imported from India and elsewhere (*Proposal for the Universal Use of Irish Manufacture*). His sumptuary mercantilism is the same as that of Sir William Temple, Polexfen and Berkeley.

For the rest he advocated national education and beggars' badges, and adopted Temple's fallacy that high rents caused high prices, Prior's facts and fallacies on absentees, and Molesworth's views on rack-rents; he deplored with Temple the destruction of timber, and opposed Boulter's lowering of the gold coin, and Berkeley's proposal for a bank; and like Prior, Berkeley, James King, Simon, and others, he advocated an Irish mint in his Drapier's Letters (1723–4). In these Swift's economic objection to Wood's copper – if stripped of its figures, which Swift meant to be figures of speech – was that the coin being hammered and not milled was easily forged, was base, excessive, and not convertible by the patentee; further, the patent did not make it legal tender, so that when this was known it would at once lose its mint value unless it should, by an abuse of the royal prerogative, be made full legal tender, in which case it would drive out gold and then depreciate. Swift did not, nor could any writer at that time, analyse the latter process, and he omitted a third possibility, that it might be made limited legal tender.

If this omission was uncandid what shall we say of his critics Leslie Stephen (*Swift*, pp. 153ff.) and Moriarty (*Swift*, 1893, p. 211), who assume that the coin was legal tender up to $5\frac{1}{2}$d.? Further, token coins, if redundant and difficult to convert, are open to these objections, so that the omission weakens but does not vitiate the argument. Lastly, facts and dates indicate that there was a likelihood that these coins would be made legal tender either to an unlimited or to a dangerously high extent. Wood boasted, 9 February 1722, that his coins were or would be made legal tender (Coxe, *Walpole*, ii. 371); and from Lady-day 1722, when his coining rights began, to 16 September 1723, the terms were unknown even in Dublin Castle; had the whole amount, £108,000 been floated in the dark, the hands of ministers must have been forced and most of Swift's fears realized. Again, the crown rent was £100, and Walpole's report valued it at £800; under the Armstrong–Knox patent of 1680 the copper need not be quite so good, and was only limited legal tender, but the rent was only £16. Further, there was virtually no silver in Ireland (Sir John Browne, *Scheme*, 1729; British Museum, Add. MSS 34358, pp. 74, 79); and every one was either bimetallic or silver-monometallic; and only thirty-three years before James II had substituted full legal tender brass for silver. Further, the customs officers were practically ordered (Coxe, p. 393) to receive these coins without limit, and in the efforts referred to in Coxe (pp. 346–438) and Monck Mason (Appendix, note *c*) to dissuade ministers from making them legal tender, no limit is mentioned. Lastly, it was clear ever since the first letter that the patent would always be onerous; yet when it was revoked, 14 August 1725, the treasury paid Wood instead of Wood paying the treasury; a compact with Wood to make the coins legal tender would explain this. Ruding cites against these arguments Walpole's *Report of the Privy Council*, 24 July 1724, which disclaimed any intention to make the coins legal tender; and argues that because danger was averted it was not real. Yes, but Swift's first

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Letter was published November (?) 1723, and doubtless caused the report, just as the second letter doubtless caused its publication, and the third letter criticized it. (Faulkner's reprint (1725) misprints ‘four’ for ‘three’ in the third paragraph of the first letter, and so makes the date autumn 1724. A similar sentence occurs in the seventh letter (vol. vii, p. 52) where ‘four’ is correct; the seventh is therefore a year later than the first letter, and its date is the end of October 1724. Lord Midleton probably refers to the first letter as written but not yet published, 1 November 1723 (Coxe, vol. ii, p. 372).) In the next two letters the storm centre shifts from economics to politics; the next is retrospective, the last prospective. Or it will be said ‘how absurd to think that Walpole would do what James II did and in the same way!’ Of course the patent was a mere blunder; if the coin were private it ought to have been, like promissory notes, convertible into legal tender coin by the issuer; if public, it ought to have been legal tender; and it was neither. But blunders often have the same effect as crimes. To conclude, Swift described, with popular but not misleading rhetoric, a grave economic peril which he more than anyone averted.

J.D. Rogers

Reprinted from *Palgrave's Dictionary of Political Economy*.

With regard to Swift's contributions to economic thought, it might be added that his unmatched gifts for satire tower in general import over his undoubtedly sincere proposals for actual reform. This disparity can be seen, for example, by comparing his ‘Modest Proposal’ with the seventh of the *Drapier's Letters*, which Swift entitled ‘An Humble Address to Both Houses of Parliament’ (1724). Somewhat less than half of the latter is given over to the larger project of slinging arrows at those who tolerated the Englishman Wood's monopoly in the issue of ‘Hogsheads’ (copper halfpennies). The remainder is taken up with Swift's concrete proposals for the economic restoration of Ireland – proposals which indicate the extent of its economic subjection to England. These proposals include: creation of a national mint, reform of the structure of rent paid to absentee landlords and salaries paid to absentee governmental placeholders, cessation of the forced importation of English goods, and cessation of the devastating practice of deforesting and depopulating the Irish countryside for the creation of pastureland. The last of these reforms was particularly urgent. For, if the ‘Hogshead’ symbolized Ireland's political subjugation and economic dependence on England, Swift was quite aware that the vista of its countryside populated by sheep rather than men signalled something yet more sinister. However, in style, these ‘humbly’ formulated and half-plaintive proposals for reform clearly lack the bite and moral outrage of Swift's direct attacks on Wood. By comparison, the satirical ‘A Modest Proposal for Preventing the Children of Ireland from Being a Burden to their Parents or Country’ (1729) far more effectively ridicules the extent of England's exploitation of Ireland through the image of cannibalism. If England's economic policy towards Ireland is to remain the same, then the ‘modest proposal’ of creating a market for the raising and consumption of Irish children, reveals more starkly both its aims and absurd destructiveness than do the more sober proposals of the Drapier's letters. One may wonder whether in writing this ‘modest proposal’ Swift presumed correctly that the English, like the ministers of Yahoodom, feared far more the sting of having their policies proven foolish than of having them shown to be unjust.

How to cite this article

Stimson, S. C. "Swift, Jonathan (1667–1745)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000351>
doi:10.1057/9780230226203.1660

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

switching costs

Paul Klemperer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Switching costs arise when transactions, learning, or pecuniary costs are incurred by a user who changes suppliers (including for ‘follow-on’ or ‘aftermarket’ products such as refills and repairs). The *ex post* market power that switching costs give suppliers *need not* create inefficiencies, and early ‘bargain’ prices can compensate consumers for later ‘rip-off’ pricing. More often, however, switching costs make new entry hard, distort firms’ product ranges, raise firms’ profits and lower consumer and social welfare. Similar issues arise in ‘shopping-cost’ markets. Policymakers should scrutinize markets where firms deliberately choose incompatibility.

Keywords

aftermarkets; antitrust cases; compatibility; competition policy; consumer protection; credit card industry; demand inertia; entry; experience goods; follow-on products; IBM; incompatibility; learning costs; lock-in; loss leaders; market power; market share; Microsoft; network effects; oligopoly; penetration pricing; price wars; product variety; search costs; shopping costs; state dependence; switching costs; transaction costs

Article

A product exhibits classic switching costs if a buyer will purchase it repeatedly and find it costly to switch from one seller to another. For example, there are high transaction costs in closing an account with a bank and opening another with a competitor; there may be substantial learning costs involved in switching between computer-software packages; and switching costs can also be created by nonlinear pricing as, for example, when an airline enrols passengers in a ‘frequent flyer’ programme that gives them free trips after flying a certain number of miles with that airline.
Switching costs also arise if a buyer will purchase ‘follow-on’, or ‘aftermarket’, products such as service, refills or repairs, and finds it difficult to switch from the supplier of the original product. In

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
short, switching costs are created whenever the consumer makes an investment specific to his current seller that must be duplicated for any new seller.

Large switching costs lock in a buyer once he makes an initial purchase, so unless sellers specify all the future prices and qualities of their products a long-term relationship is governed by short-term contracts. This creates *ex post* market power, for which firms compete *ex ante*; they use strategies such as penetration pricing, price wars, and introductory offers to fight for market share that will generate future profits (see, for example, Klemperer, 1987a; 1989).

The central question in the literature is the extent to which this fierce *ex ante* competition for buyers is an adequate substitute for more standard period-by-period competition without switching costs.

In the simplest models, firms' low 'bargain' prices to new customers exactly compensate buyers for the high 'rip-off' prices they will pay after lock-in, so buyers' total 'life-cycle' payments are unaffected by their switching costs, and the absence of any price commitments leads to no inefficiencies. But things do not usually work so well.

Most theoretical models confirm the popular intuition that switching costs raise firms' profits and lower social welfare. Switching costs can segment even an otherwise undifferentiated market as firms concentrate on exploiting their established customers and do not compete aggressively for their rivals' buyers. Unless new firms will enter the market in the future, an existing firm would usually expect to earn future profits even if it made no current sales (because it could then usually capture a few of its rivals' lower-switching cost customers, by setting a lower price than its rivals). So rent dissipation is low, allowing oligopolists to extract positive profits overall (as illustrated by Farrell and Shapiro, 1988; Padilla, 1995; Chen, 1997; Taylor, 2003; and others). Furthermore, consumers facing switching costs care about expected future prices as well as about current prices, and are therefore generally less sensitive to current prices than they would be in the absence of switching costs. So firms generally have less incentive to cut prices, and prices and profits are therefore higher, than they would be in the absence of switching costs (see Klemperer, 1987b; Beggs and Klemperer, 1992). Although it is possible to construct models in which switching costs lower profits by making special assumptions about consumers' expectations and tastes (see von Weizsäcker, 1984), and there is little convincing empirical evidence, the limited laboratory evidence suggests switching costs raise profits (see Cason and Friedman, 2002).

Even when *ex ante* competition fully dissipates *ex post* rents, it may do so in unproductive ways such as through socially inefficient marketing. The 'lo-hi' pricing patterns distort buyers' choices – for example, if a printer is priced below cost but the locked-in ink is expensive, buyers may buy an over-specified printer but then use it too little from the social viewpoint. Such pricing also gives customers wrong signals about whether to switch. If consumers do switch, direct costs are incurred (Klemperer, 1988) and, if consumers avoid those costs by not switching, that obstructs otherwise efficient matching between buyers and sellers. Product variety is less sustainable in the market than if niche products are compatible with mainstream products and so don't require new users to incur switching costs.

Switching costs also hamper forms of entry that must persuade customers to pay those costs, in particular, large-scale entry that seeks to attract other firms' customers (for instance to achieve minimum viable scale, if the market is not growing quickly) (Klemperer, 1987c). The difficulty of new entry may be broadly efficient *given* switching costs, but is nevertheless a social cost of switching costs.

So switching costs often damage competition, and firms may therefore also dissipate further resources

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
creating and defending incompatibilities – as, for example, when Gillette famously, and repeatedly, changed the design of its razors to prevent competing manufacturers from selling compatible blades. Likewise, it is alleged that both IBM and Microsoft have deliberately obstructed compatibility between their products and those of their competitors, and Kodak tried to prevent independent repair firms from servicing its photocopiers.

The bargain-then-rip-off pricing structure is clearest when new and locked-in customers are clearly distinguished and can be charged separate prices, as, for example, when prices are individually negotiated, or when locked-in customers buy separate ‘follow-on’ products, such as parts and service, rather than repeatedly buying the same good.

If, instead, each firm has to set a single price to old (locked-in) and new customers (see Beggs and Klemperer, 1992), that price must compromise between a high price to exploit current locked-in buyers and a lower price to attract more buyers to lock in and exploit later. The implications for competition and welfare are similar to those for the bargain-then-rip-off case above, except that here switching costs also create a ‘fat-cat’ effect: firms with large customer bases set higher prices, because they have more to gain from harvesting current customers than winning new ones.

On the one hand, the ‘fat-cat’ effect is a further force for high prices – firms price less aggressively both because they recognize that, if they win fewer customers today, their rivals will be ‘fatter’ and therefore less aggressive tomorrow, and because consumers recognize this too and so are less impressed by lower current prices. On the other hand, it actually facilitates entry that focuses purely on new customers – since an incumbent's incentive to set a high price against its locked-in buyers creates a price umbrella under which a new entrant can come in.

The fat-cat effect means large shares tend to shrink and small shares to grow; when firms’ shares are similar, they return to stable steady state after any shock. More generally, the trade-off between harvesting old customers and investing in new ones depends on interest rates, the state of the business cycle, expectations about exchange-rates, and so on, with implications for macroeconomics and international trade (Chevalier and Scharfstein, 1996; Froot and Klemperer, 1989; Klemperer, 1995).

Some of the same issues as with switching costs arise when shops advertise only some of their prices: customers become ‘locked in’ as they bear the costs of going to a shop, and only afterwards learn its other prices. Just as with dynamic switching costs, this tends to produce bargains on advertised (‘loss leader’) prices and corresponding rip-offs on unadvertised prices (see Lal and Matutes, 1994).

Also closely related are ‘shopping-cost’ markets where consumers face costs of using different suppliers for different goods in a single period but all prices are advertised (though neither time nor commitment problems are central in these markets). Shopping costs encourage firms to offer a full product line – for example, a supermarket stocks a broad range of products to encourage consumers to shop only there – and so help explain multi-product firms. Indeed firms’ product ranges may be too broad from the social viewpoint (Klemperer and Padilla, 1997), but may also be too similar to each other (Klemperer, 1992) so that there is too little variety in the market as a whole. Shopping costs also make single-product entry hard. However, the ‘mix-and-match’ literature, beginning with Matutes and Regibeau (1988), suggests that firms typically prefer to be compatible (no shopping costs) rather than incompatible (infinite shopping costs), at least in symmetric single-period duopolies.

Other literature related to switching costs includes that on network effects (see, for example, Farrell and Klemperer, 2007, and the companion-piece to this article, network goods (theory)), search costs (see, for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
example, Stiglitz, 1989), and ‘experience goods’ (see, for example, Schmalensee, 1982).
The theoretical literature on switching costs described above arguably began with Selten's (1965) model of ‘demand inertia’ (which assumed a firm's current sales depended in part on history, even though it did not explicitly model consumers’ behaviour in the presence of switching costs), and then took off in the 1980s with contributions from von Weizsäcker (1984), Klemperer (1983; 1987a; 1987b; 1987c), Farrell and Shapiro (1988), and others. But, although there is an extensive empirical marketing literature on brand loyalty (or ‘state dependence’) which often reflects, or has equivalent effects to, switching costs (summarized in Seetharaman, Ainslie and Chintagunta, 1999), the empirical economics literature on switching costs is smaller and more recent than the theoretical literature.
Only a few studies attempt to directly measure switching costs. Where microdata on individual consumers’ purchases are available, a discrete choice approach can be used to explore the determinants of a consumer's probability of purchasing from a particular firm (for example, Greenstein, 1993, on computer systems procurement; Shum, 2004, on breakfast cereal purchases), but, because switching costs are usually both consumer-specific and not directly observable, and microdata on individual consumers’ purchase histories are seldom available, less direct methods of assessing the level of switching costs are often needed (for example, Kim, Kliger and Vale, 2003) estimate a first-order condition and demand and supply equations for Norwegian bank loans, and Shy (2002) uses data on prices and market shares for the Israeli cellular phone market.
One defect of most empirical studies is that few of them model the dynamic effects of switching costs that are the main focus of the theoretical literature; most of them assume consumers myopically maximize current utility without considering the future effects of their choices.
Nevertheless, the empirical literature does suggest that switching costs play an important role in many industries, including credit cards, cigarettes, supermarkets, air travel, phone services, electricity suppliers, bookstores, and automobile insurance (see Farrell and Klemperer, 2007, for references, and Klemperer, 1995, for more examples of markets with switching costs); as technology continues to develop, products become more complex, and services become more important, the importance of switching costs seems likely to increase further.
Because switching costs very often make competition, and especially entry, less effective, I (and many others) favour cautiously pro-compatibility public policy. Policymakers should look particularly carefully at markets where incompatibility is strategically chosen rather than inevitable. Because buyers’ early choices are often crucial, and depend on their expectations about firms’ future behaviour, provision of information and consumer protection against deception, and so on, are more important than usual. And competition policy must recognize that the analyses of mergers, monopolization, intellectual property, and predation are all affected by switching costs; it is unsurprising that switching costs have featured importantly in many of the world's best-known and most significant antitrust cases, including the IBM, the Kodak, and the European Microsoft cases.
Farrell and Klemperer (2007) contains a recent and comprehensive survey of switching costs.

See Also

- network goods (empirical studies)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- network goods (theory)

The views expressed here are personal and should not be attributed to the UK Competition Commission or to any individual member other than myself. Furthermore, although some observers thought some of the behaviour discussed above warranted regulatory investigation, I do not intend to suggest that any of it violates any applicable rules or laws.

Bibliography

Beggs, A. and Klemperer, P.D. 1992. Multiperiod competition with switching costs. *Econometrica* 60, 651–66.

Cason, T.N. and Friedman, D. 2002. A laboratory study of customer markets. *Advances in Economic Analysis and Policy* 2(1), Article 1.

Chen, Y. 1997. Paying customers to switch. *Journal of Economics and Management Strategy* 6, 877–97.

Chevalier, J. and Scharfstein, D. 1996. Capital-market imperfections and countercyclical markups: theory and evidence. *American Economic Review* 86, 703–25.

Farrell, J. and Klemperer, P.D. 2007. Coordination and lock-in: competition with switching costs and network effects. In *Handbook of Industrial Organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: North-Holland.

Farrell, J. and Shapiro, C. 1988. Dynamic competition with switching costs. *RAND Journal of Economics* 29, 123–37.

Froot, K.A. and Klemperer, P.D. 1989. Exchange rate pass-through when market share matters. *American Economic Review* 79, 637–54.

Greenstein, S.M. 1993. Did installed base give an incumbent any (measurable) advantage in federal computer procurement? *RAND Journal of Economics* 24, 19–39.

Kim, M., Kliger, D. and Vale, B. 2003. Estimating switching costs: the case of banking. *Journal of Financial Intermediation* 12, 25–56.

Klemperer, P.D. 1983. Consumer switching costs and price wars. Working paper, Graduate School of Business, Stanford University.

Klemperer, P.D. 1987a. Markets with consumer switching costs. *Quarterly Journal of Economics* 102, 375–94.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Klemperer, P.D. 1987b. The competitiveness of markets with switching costs. *RAND Journal of Economics* 18, 138–50.

Klemperer, P.D. 1987c. Entry deterrence in markets with consumer switching costs. *Economic Journal* 97(Supplement), 99–117.

Klemperer, P.D. 1988. Welfare effects of entry into markets with switching costs. *Journal of Industrial Economics* 37, 159–65.

Klemperer, P.D. 1989. Price wars caused by switching costs. *Review of Economic Studies* 56, 405–20.

Klemperer, P.D. 1992. Equilibrium product lines: competing head-to-head may be less competitive. *American Economic Review* 82, 740–55.

Klemperer, P.D. 1995. Competition when consumers have switching costs. *Review of Economic Studies* 62, 515–39.

Klemperer, P.D. and Padilla, A.J. 1997. Do firms' product lines include too many varieties? *RAND Journal of Economics* 28, 472–88.

Lal, R. and Matutes, C. 1994. Retail pricing and advertising strategies. *Journal of Business* 67, 345–70.

Matutes, C. and Regibeau, P. 1988. Mix and match: product compatibility without network externalities. *RAND Journal of Economics* 19, 221–34.

Padilla, A.J. 1995. Revisiting dynamic duopoly with consumer switching costs. *Journal of Economic Theory* 67, 520–30.

Schmalensee, R. 1982. Product differentiation advantages of pioneering brands. *American Economic Review* 72, 349–65.

Seetharaman, P.B., Ainslie, A. and Chintagunta, P.K. 1999. Investigating household state dependence effects across categories. *Journal of Marketing Research* 36, 488–500.

Selten, R. 1965. Spieltheoretische behandlung eines oligopolmodells mit nachfrägetragheit. *Zeitschrift für die Gesamte Staatswissenschaft* 121, 301–24 and 667–89.

Shum, M. 2004. Does advertising overcome brand loyalty? Evidence from the breakfast cereals market. *Journal of Economics and Management Strategy* 13, 241–72.

Shy, O. 2002. A quick-and-easy method for estimating switching costs. *International Journal of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Industrial Organization 20, 71–87.

Stiglitz, J.E. 1989. Imperfect information in the product market. In *Handbook of Industrial Organization*, vol. 1, ed. R. Schmalensee and R.D. Willig. Amsterdam: North-Holland.

Taylor, C. 2003. Supplier surfing: competition and consumer behavior in subscription markets. *RAND Journal of Economics* 34, 223–46.

von Weizsäcker, C.C. 1984. The cost of substitution. *Econometrica* 52, 1085–116.

How to cite this article

Klemperer, Paul. "switching costs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000460> doi:10.1057/9780230226203.1661

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

symmetry breaking

Kiminori Matsuyama
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Symmetry breaking creates asymmetric outcomes in the symmetric environment. It is the key concept for understanding self-organized pattern formations in natural sciences as well as in economics. We explain the logic of symmetry breaking and some methodological issues, and discuss applications to urban and regional economics, international economics, growth and development, economic fluctuations, and occupational choice.

Keywords

agglomeration forces; comparative advantage; fashion cycles; inequality between nations; investment distortions; racial segregation; regional economics; resource constraint forces; social inequality; spatial symmetry; symmetry breaking; temporal symmetry; total factor productivity; urban economics; urban segregation

Article

Symmetry breaking creates asymmetric outcomes in the symmetric environment. It is the key concept for understanding self-organized (that is, endogenous) pattern formations. For example, cosmologists wonder why matter in the universe is distributed in clusters, leaving much of the universe empty. Earth scientists study the formation of wave patterns, such as jet streams, ocean currents, and continental drifts. Material scientists study phase transitions, how molecules align themselves when they reach the critical temperature. Molecular biologists ask how life began in the primordial soup of amino acids, and developmental biologists attempt to explain how living organisms acquire forms through cell division and morphogenesis (Weyl, 1969; Prigogine, 1980). Similar questions about pattern formations also exist in economics. Why are there rich and poor countries? Why are industries clustered? Why are there booms and recessions? Why are some ethnic groups under-represented in certain jobs or neighbourhoods?

A simple model can illustrate the role of symmetry breaking in self-organized pattern formations. Consider an economy with two inherently identical regions. Each region has an equal amount of an immobile factor, say, land or labour. In addition, there is a single mobile factor, capital, whose allocation is measured along the horizontal axis in Figure 1. Two curves show how the rates of return to capital in the two regions change with the allocation of capital. To the extent that capital has to compete for the use of the immobile factor, the rate of return to capital in one region declines as more capital is allocated to that region. Because the two regions are inherently identical, the two curves are symmetric around the midpoint. When capital is allocated evenly, the rates of return in the two regions are equalized; hence, this is an equilibrium allocation. Furthermore, this allocation is stable: if capital is allocated unevenly, the region with less capital offers a higher rate of return and so attracts capital from the other region, and the resulting capital flow restores the equilibrium allocation. Thus, the ‘centrifugal forces’ of the resource constraint prevent one region from attracting more capital than the other.

Figure 1

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

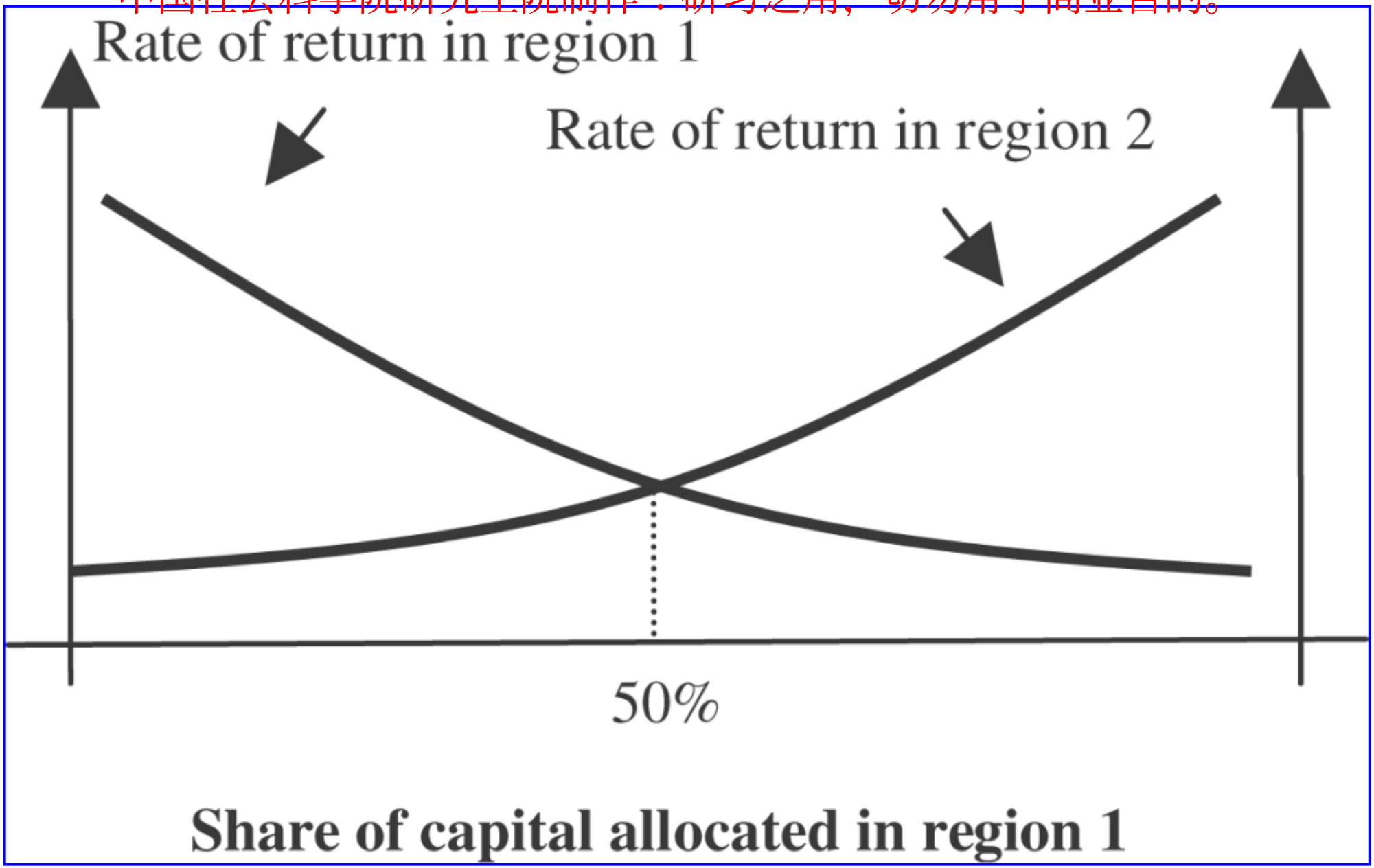

Now let us add some agglomeration economies so that the productivity of capital can be enhanced by concentration. If the ‘centripetal forces’ of agglomeration economies dominate the ‘centrifugal forces’ of the resource constraint, the rate of return in one region may increase as more capital is allocated to that region, at least over a certain range (Figure 2). The model now has two stable equilibria and both imply uneven spatial allocations of capital. The even allocation of capital is still an equilibrium, but unstable. If slightly more than half of capital is allocated in one region due to some small historical accidents, the rate of return becomes higher there and so more capital flows from the other region. Once this process gets started, it would feed on itself, and the allocation of capital would be further away from the equal division. One region emerges as the developed core and the other is left behind as the underdeveloped periphery. The model does not say which of the two asymmetric outcomes would emerge, but it does say that the observed outcome has to be asymmetric. Thus, the loss of stability of the symmetric outcome, ‘symmetry breaking’, generates the formation of the core-periphery patterns.

Figure 2

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

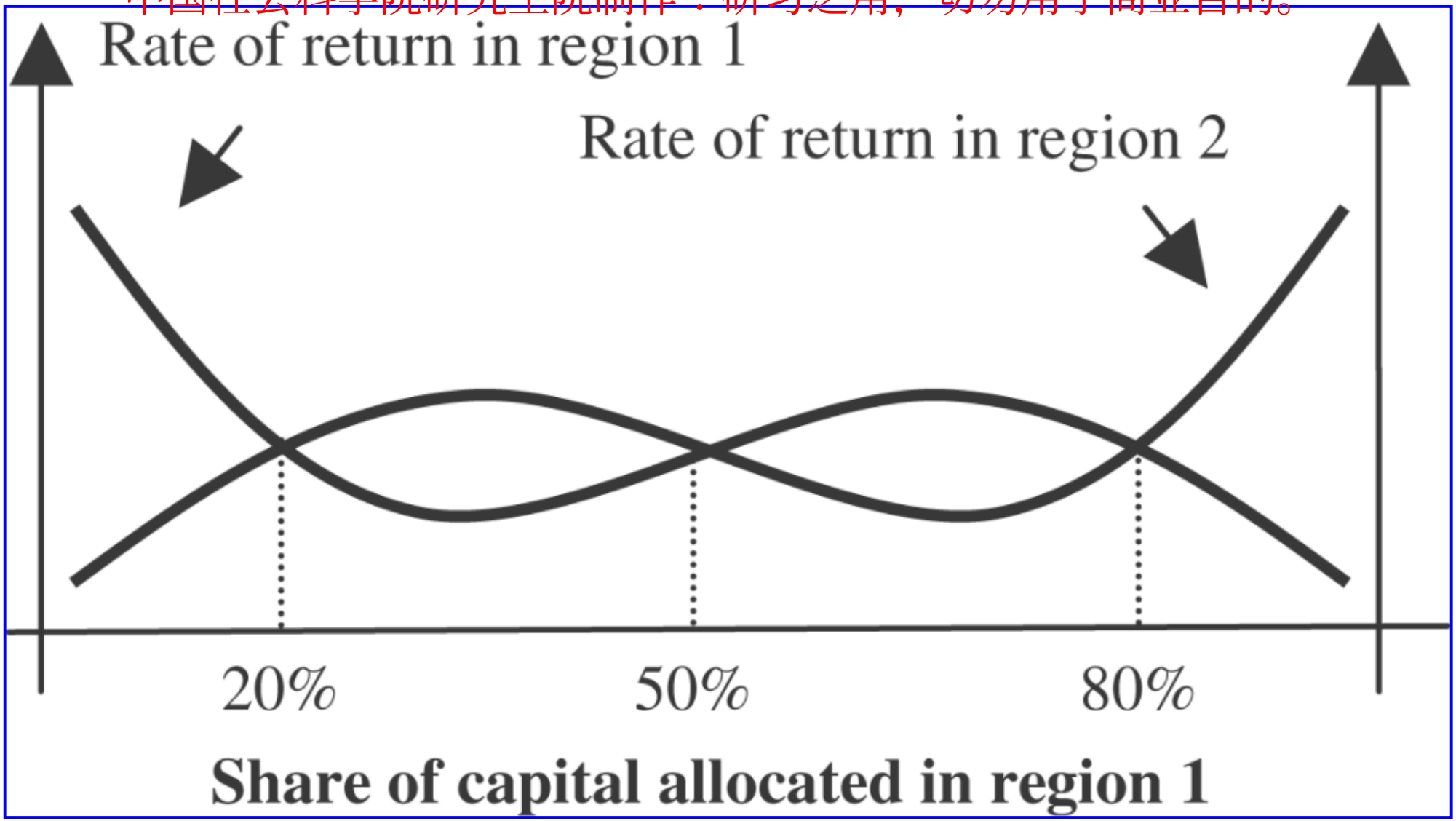

What separates the two situations depicted in Figures 1 and 2 is the relative magnitude of the centripetal to centrifugal forces, which may in turn be determined by a certain parameter of the model, say, λ. Figure 3 schematically traces out how stable outcomes might change as λ increases. For $\lambda < \lambda_c$, there is a unique stable outcome, which is symmetric; no spatial asymmetry appears. As it reaches to a critical value of the parameter, λ_c, the balance between the two forces is tipped, at which point there will be an abrupt change. For $\lambda > \lambda_c$, there are two stable asymmetric outcomes (shown by the solid branches of the curve) and the symmetric outcome becomes unstable (as indicated by the dashed line). So, as λ exceeds its critical value, patterns are formed. Figure 3 shows that the rate at which a stable asymmetric outcome responds to a change in λ is arbitrarily large above the critical value. This is a fairly generic feature of models in which two competing forces are at work. At the onset of pattern formation, even a small change in the environment can be amplified to create a large effect.

Figure 3

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

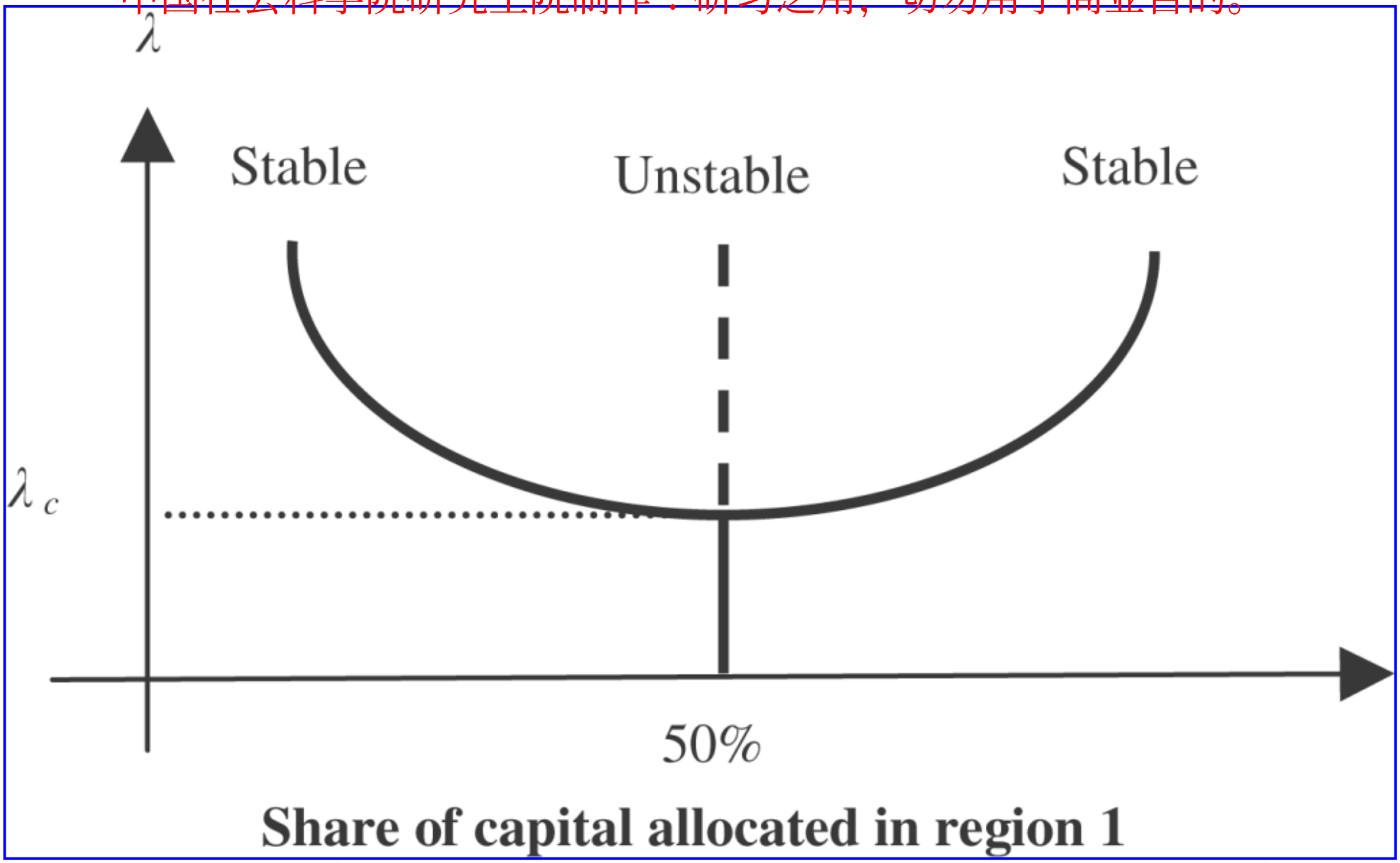

Some physical analogies may be apt here. If water is heated at atmospheric pressure, nothing drastic happens until it reaches 100°C, when it suddenly starts boiling. If the temperature is lowered, the water suddenly turns to ice at 0°C. Such abrupt changes in the phase of matter are the results of a competition between the attractive intermolecular forces, which tend to order the system, and the individual kinetic energies of the molecules, which have the opposite effect. In a liquid phase, the random motion of water molecules is too violent for intermolecular interaction to hold them in place. Increasing the pressure, which favours the intermolecular forces, or reducing the temperature, which reduces the kinetic energies of the molecules, eventually tips the balance between the two. Once the critical point is crossed, the attractive forces will be strong enough to keep molecules firmly in place and the crystal structure of ice will be developed. Many similar examples exist in nature. At room temperature iron can be magnetized, but when it is heated above the Curie point it loses its ferromagnetic property. Another example would be superconductivity, in which electronic resistance disappears discontinuously at the critical temperature.

Symmetry breaking aims to explain the formation of the patterns or the structure. In contrast, most economic analysis treats the structure of the economy as given, which can be a useful short cut for many purposes. For example, if we are interested only in constructing a model consistent with the observed patterns, we could simply assert that the regions are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
not similar and that some regions have more capital and produce more outputs simply because they have better technologies. Although most economists might accept this as an 'explanation', it merely raises another question in the mind of many. Why are some regions more productive than others? One might respond to this question by pointing out that the market works more efficiently in some regions than in others, or that people in some regions are more diligent, or that people in other regions lack entrepreneurial spirits, and so on – which would raise yet another question. As long as we try to explain spatial variations in one variable by introducing spatial variations in another variable, there may be no end to this process. The logic of symmetry breaking does not require any exogenous source of variations, because of the circular causation. In the case of Figure 2, not only does better technology attract more capital but more capital leads to better technology, due to agglomeration economies. Such two-way causality can bring about the instability of the symmetric outcome, which leads to the formation of the core-periphery patterns. In other words, symmetry breaking explains such variations purely as an outcome of the internal mechanisms of the system, that is, in a self-organized manner.

This is not to say that some exogenous sources of variations, such as the climate and geography, are unimportant factors in explaining pattern formations. Rather, the central message of symmetry breaking is that such exogenous variations do not have to be large; even small exogenous variations can be amplified to generate large variations. In short, a small cause can create a big effect. Once pointed out, this may seem obvious. Yet some economists may look for a big change in the environment when trying to explain a big movement of a certain economic variable. But the logic of symmetry breaking suggests that their efforts might very well turn out to be futile. When somebody proposes a hypothesis that the difference in X is due to the difference in Y, one often hears the criticism that the difference in Y is too small to have caused the difference in X. But the logic of symmetry breaking suggests that such a criticism may be unwarranted.

The logic of symmetry breaking has always been central in urban and regional economics (see Fujita, Krugman and Venables, 1999). Symmetry breaking also gives us a fresh perspective on many familiar questions in other areas of economics. In international trade, the patterns of comparative advantage are linked to cross-country differences in productivity, but the sources of the productivity differences are often left unexplained. If productivity improves through accumulation of experience, there will be a two-way causality between trade and technology (Krugman, 1987; Lucas, 1988; Matsuyama, 1992a; see Grossman and Helpman, 1995, for a survey). In the presence of such two-way causality, many of the observed patterns of trade can arise endogenously due to the cumulative effects of small historical accidents.

In the growth and development literature, cross-country differences in per capita income are often attributed to difference in total factor productivity (TFP) or in investment distortions, but these differences themselves are often left unexplained. But lower productivity or greater distortions may not only be a cause of low income, but the low income may also be a cause of lower productivity or greater distortions. With such two-way causality, cross-border movement of goods, as in Krugman and Venables (1995) and Matsuyama (1996; 1999a), or capital, as in Boyd and Smith (1997) and Matsuyama (2004a), amplifies small inherent differences across the countries, which makes balanced development (spatial symmetry) unstable, and the world economy may inevitably evolve into a system of rich and poor, where the countries are divided between a developed core, characterized by high income, high TFP, small investment distortions, and an underdeveloped periphery, characterized by low income, low TFP, and large investment distortions. If so, the problem of an underdeveloped country cannot be looked at in isolation; instead, it has to be examined as a part of the interrelated whole because the rich may be rich in part due to the presence of the poor and the poor may be poor in part due to the presence of the rich.

Likewise, social inequality and a class structure might emerge endogenously, even if people were identical in every conceivable dimension, which suggests the unattainability of a classless society (Freeman, 1996); Matsuyama, 2000; 2006). Francois (1998) and Burdett and Wright (1998) suggest that an outcome in which the two sexes are treated equally may be unstable.

In macroeconomics, booms and recessions are often attributed to fluctuating TFP; yet why TFP fluctuates is often left unexplained. Endogenizing TFP, either through innovation as in Aghion and Howitt (1992) and Matsuyama (1999b), or through distorted allocation of the credit as in Azariadis and Smith (1998) and Matsuyama (2004b; 2005), could cause instability in the stationary path (the temporal symmetry), which generates temporal patterns of booms and recessions, in which it appears that fluctuations are driven by the movements of TFP.

In residential choices, symmetry breaking forces us to ask why an integrated or mixed neighbourhood is difficult to sustain (see Schelling, 1978; Miyao, 1978; Benabou, 1993). And why are certain ethnic groups over-represented or under-represented in certain jobs? The standard model of occupational choice in labour economics would suggest that there are large innate differences in skills across ethnic groups. The logic of symmetry breaking suggests that, if the informational externalities in the process of skill acquisition or job search are greater within the same ethnic group, even small initial differences in skill distributions or some historical accidents may end up sorting different ethnic groups into different occupations. Or think of the problem of comparative economic systems. One might be tempted to attribute differences in labour market practices or in financial systems across countries to the differences in regulations or in national cultures. However, symmetry breaking suggests an alternative view. Due to some institutional complementarities across different practices and across different firms, either most firms in a country would adopt these practices, or else very few firms would adopt them. They may be more prevalent in some countries than in others only because of certain historical accidents. Furthermore, the diversity of national economic systems may be an inevitable feature of an integrated world (Aoki, 2001). Last but not least, spatial and temporal variations in social behaviour, such as fashion cycles (Matsuyama 1992b), may also be better understood through the lens of symmetry breaking.

For further discussion on the methodological issues raised by symmetry breaking, circular causation, and multiple equilibria, see Matsuyama (1995a; 1995b; 1997; 2002).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- growth and cycles
- poverty traps
- structural change

Bibliography

Aghion, P. and Howitt, P. 1992. A model of growth through creative destruction. *Econometrica* 60, 323–51.

Aoki, M. 2001. *Towards a Comparative Institutional Analysis*. Cambridge, MA: MIT Press.

Azariadis, C. and Smith, B. 1998. Financial intermediation and regime switching in business cycles. *American Economic Review* 88, 516–36.

Benabou, R. 1993. Workings of a city: location, education, and production. *Quarterly Journal of Economics* 108, 619–52.

Boyd, J. and Smith, B. 1997. Capital market imperfections, international credit market and nonconvergence. *Journal of Economic Theory* 73, 335–64.

Burdett, K. and Wright, R. 1998. Two-sided search with non-transferable utility. *Review of Economic Dynamics* 1, 220–45.

Francois, P. 1998. Gender discrimination without gender differences. *Journal of Public Economics* 68, 1–32.

Freeman, S. 1996. Equilibrium income inequality among identical agents. *Journal of Political Economy* 104, 1047–64.

Fujita, M., Krugman, P. and Venables, A. 1999. *The Spatial Economy*. Cambridge, MA: MIT Press.

Grossman, G. and Helpman, E. 1995. Technology and trade. In *Handbook of International Economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: North-Holland.

Krugman, P. 1987. Narrow moving bands, the Dutch Disease, and the economic consequences of Ms. Thatcher. *Journal of Development Economics* 27, 41–55.

Krugman, P. and Venables, A. 1995. Globalization and inequality of nations. *Quarterly Journal of Economics* 110, 857–80.

Lucas, R., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Matsuyama, K. 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. *Quarterly Journal of Economics* 106, 617–50.

Matsuyama, K. 1992a. Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory* 58, 317–34.

Matsuyama, K. 1992b. Custom versus fashion. CMS-EMS DP No. 1030. Evanston, IL: Northwestern University.

Matsuyama, K. 1995a. Complementarities and cumulative processes in models of monopolistic competition. *Journal of Economic Literature* 33, 701–29.

Matsuyama, K. 1995b. Comment on Paul Krugman's ‘Complexity and Emergent Structure in the International Economy’. In *New Directions in Trade Theory*, ed. A. Deardorff, J.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Levinsohn and R. Stern. Ann Arbor: University of Michigan Press.

Matsuyama, K. 1996. Why are there rich and poor countries? Symmetry-breaking in the world economy. *Journal of the Japanese and International Economies* 10, 419–39.

Matsuyama, K. 1997. The 1996 Nakahara Lecture: complementarity, instability, and multiplicity. *Japanese Economic Review* 48, 240–66.

Matsuyama, K. 1999a. The 1999 Fukuzawa Lecture: geography of the world economy. CMS EMS DP No. 1239. Evanston, IL: Northwestern University.

Matsuyama, K. 1999b. Growing through cycles. *Econometrica* 67, 335–47.

Matsuyama, K. 2000. Endogenous inequality. *Review of Economic Studies* 67, 743–59.

Matsuyama, K. 2002. Explaining diversity: symmetry-breaking in complementarity games. *American Economic Review* 92, 241–6.

Matsuyama, K. 2004a. Financial market globalization, symmetry-breaking, and endogenous inequality of nations. *Econometrica* 72, 853–84.

Matsuyama, K. 2004b. The good, the bad, and the ugly: an inquiry into the causes and nature of credit cycles. CMS-EMS DP No. 1391. Evanston, IL: Northwestern University.

Matsuyama, K. 2005. Credit traps and credit cycles. Working paper. Evanston, IL: Northwestern University.

Matsuyama, K. 2006. The 2005 Lawrence R. Klein Lecture: Emergent Class Structure. *International Economic Review* 47.

Miyao, T. 1978. Dynamic instability of a mixed city in the presence of neighborhood externalities. *American Economic Review* 68, 454–63.

Prigogine, I. 1980. *From Being to Becoming: Time and Complexity in the Physical Sciences*. New York: W. H. Freeman.

Schelling, T. 1978. *Micromotives and Macrobehaviors*. New York: Norton.

Weyl, H. 1969. *Symmetry*. Princeton, NJ: Princeton University Press.

How to cite this article

Matsuyama, Kiminori. "symmetry breaking." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000465> doi:10.1057/9780230226203.1662

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

systems of cities

J. Vernon Henderson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Research on systems of cities examines five issues. How do cities form and what are the agglomeration benefits driving city formation and sizes? Second, in a hierarchy of cities, what is the role of big versus small cities? Third, how does a system of cities evolve under economic and population growth; and how does urban growth based on localized knowledge spillovers potentially define national economic growth? Fourth, how do governance, institutions, and public policy affect city formation and sizes? Finally where do cities locate and what the effects of history, climate and natural resource location?

Keywords

agglomeration; cluster analysis; congestion; endogenous growth models; geography; Gibrat's Law; Henry George Theorem; human capital; industry concentration; information spillovers; innovation; knowledge externality; monopolistic competition; Nash equilibrium; new economic geography; protection; quality-ladder model; systems of cities; transport costs; urban agglomeration; urban growth; urban specialization; Zipf's Law; zoning regulation

Article

Modelling and empirical work on systems of cities focuses on five related sets of questions. First, why do cities form, with so much of economic activity in countries geographically concentrated in cities? The answer has to do with agglomeration benefits, as described below. Second, given there is a hierarchy of cities, with cities of different sizes and types in a country, how do cities interact with each other? What are the potential trade patterns across cities and how do they correspond to the roles of big versus small cities? Third, how does a system of cities evolve under economic and population growth; and how does urban growth intersect with, or even define, national economic growth? In growth theory, endogenous growth is based on knowledge spillovers and sharing, and evidence suggests that much of that interaction must occur at the level of individual cities.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The fourth set of questions asks how governance, institutions, and public policy affect city formation and sizes, which then in turn affect economic efficiency. Cities require enormous public infrastructure investments which affect urban quality of life, in particular health and safety and commuting and congestion costs. Institutions governing land markets, property rights, local government autonomy, and local financing affect the city formation process and city sizes. And national government policies concerning trade, labour policies and national investment in communications and transport infrastructure affect the shape of the urban system. The final set of questions has to do with where cities locate. What are the effects of history, climate and natural resource locations, including rivers and natural harbours, on the location of current urban agglomerations?

To see how economics has tackled, and in some cases avoided, aspects of these questions, we start by describing the economics of a representative city. Then we turn to how to model the urban hierarchy and the role of trade, institutions, geography, and economic growth.

The representative city

In models of a representative city dating back to Mills (1967), as applied to systems of cities in Henderson (1974) and subsequent work, any city is viewed as having an inverted-U relationship between real income or utility per worker and city employment or population. As city sizes increase from low levels, workers benefit from localized agglomeration benefits of increasing scale. This leads to increases in per worker incomes, giving the rising part to the inverted U. These benefits can involve information spillovers in input and output markets from economic agents in close proximity, where information decay over space is very rapid. Thus close spatial clustering of firms is required for efficient information transfer. The benefits can involve search and matching economies in local labour markets. In addition, to draw from the new economic geography literature (Krugman, 1991), they can involve economies of diversity in local intermediate inputs where bigger local markets can support more varieties of local intermediate inputs, which enhances the efficiency of final export good producers in the city. Models of the micro-foundations of these ideas and econometric evidence all yield a monotonic increasing relationship between city employment size and nominal wages per worker. These static agglomeration economies can be augmented with knowledge spillovers, particularly in growth contexts. Local knowledge accumulation as represented by educational attainment enhances productivity, whereby knowledge accumulation either enriches (multiplies) local information spillovers or simply represents improved levels of local technology.

If agglomeration benefits were the end of the story, in each country there would be only one massive city containing the entire urban population. In systems of cities models, the limit on city sizes is modelled as being internal to the city. Cities must house workers, where there are scale diseconomies in city living, including per capita infrastructure costs, pollution, accidents, crime, and commuting costs. Most urban models consider an explicit internal spatial structure of cities, whereby all production occurs at a point in a city, the central business district (CBD). Surrounding the CBD is a residential area from which people must commute to the CBD. People residing nearer the city centre pay higher rents, while those residing further away, who must commute further, pay lower rents as compensation. As city sizes increase, average commuting distances, times and expenditures all rise, as do average housing and land rents paid by residents. The former are resource costs limiting the efficiency of increasing city sizes. The latter – rents – are an income whose distribution is considered below. The rising costs of increasing city

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sizes can be modelled also in cities with multiple business districts.
Once cities become large enough, at some point the marginal production benefits of increasing city size are surpassed by the marginal costs of rising commuting and other diseconomy costs. At the point of overtake, the inverted-U of real income per worker against city size peaks and then declines with further increases in city size. Although they are not well captured in theoretical models, two other key considerations limit the benefits of increasing city sizes. One is the ‘extent’ of the market for a city's product, represented in Mills (1967) and the new economic geography literature. As cities grow they have to sell their products further and further afield, encountering rising transport costs. The other is spatially dispersed natural resource deposits. Most processing of deposits is weight-reducing. Transport cost considerations result in materials being initially processed in cities at dispersed locations, where natural resources are found, before being shipped to other population and production centers.

City formation

An inverted U tells us that efficient city sizes will be limited. If cities all operate with the same level of technology and market demand for products, national welfare is maximized if all cities operate at the peak to their inverted U's. The critical issue is what institutions and market mechanisms determine how new cities form, and whether cities operate at their peak points. We start with a specific mechanism and discuss how it generalizes, as well as what happens in the absence of any mechanism. Suppose there is an unexhausted supply of identical city sites in the economy, each owned by a different land developer in a nationally perfectly competitive urban land development market (Henderson, 1974, Helsley and Strange, 1990). A developer acts both as a developer and as a ‘private’ local government, where, for an occupied city, she collects local land rents, specifies city population, provides any public services, and offers subsidies, or inducements to firms or people to locate in that city, in competition with other cities. Population is freely mobile. In choosing city attributes, she attempts to maximize profits (land rents less subsidies), but competition in national land markets with other developers for residents drives her profits to zero in equilibrium.
This solution has a variety of properties heralded in the urban literature. First, in simple models, it is a unique equilibrium that is Pareto efficient; it is the only coalition-proof equilibrium; and it is a free mobility equilibrium where the developer-specified populations are self-enforcing, that is, no worker wants to move to another city at the equilibrium configuration. Second, the solution reflects the Henry George Theorem (Flatters, Henderson and Mieszkowski, 1974), under which land rents collected in a city when it is at its efficient size exactly equal the revenues needed to fund all public services and costs of internalizing externalities. In the simple case where there are no public services, the subsidy per worker exactly equals the gap between the social and private marginal product of labour within the city. Thus the subsidy prices scale externalities. A final property of the equilibrium is that, in a large economy with many cities, there are effective constant returns to scale nationally. Increases in national population are accommodated by increased city numbers, where each city is at efficient size (Kanemoto, 1980; Henderson and Ioannides, 1981). If there are not many cities, there is a ‘lumpiness’ or integer problem in solving the model more generally. If there is enough national urban population to support only 3.5 cities of efficient size, an equilibrium is going to have no more than three cities, all a greater size than that where the inverted U peaks.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
As shown in Henderson and Becker (2001), the land developer solution arises also under an alternative mechanism where the private government/developer is replaced by a city government that can tax land rents, provide public services, subsidize entrants to the city and potentially exclude residents through zoning regulations ('no-growth' restrictions). These autonomous democratic local governments act in a political competition model to maximize the welfare of the representative local voter. In doing so, they also achieve efficient results and the Henry George Theorem applies. Note that critical ingredients are required for either the private government/developer or city government solution to work. Local governments must be able to tax away urban land rents to use as revenues to subsidize entry; otherwise cities will be too small since they can't offer the efficient level of subsidies to induce enough entry and internalize scale externalities. More critically, there must be freely functioning national land markets with autonomous local developers or local governments that are free to set up cities and potentially limit their sizes. That is, either developers must be able to assemble sufficient land to house an entire city or local governments have must have jurisdiction over the same land area. These institutions give a 'complete' national land market for city formation.

Absent such institutions and complete markets, cities form only through 'self-organization'. In simple models with perfect mobility of all resources, the result is potentially enormously oversized cities (Henderson, 1974; Henderson and Becker, 2001). In the inverted-U context, Nash equilibrium city size in terms of atomistic worker migration decisions lies between the efficient size where real income per worker is maximized and a limit size to the far right which is a 'bifurcation' point (Krugman, 1991). At that point, city size is so large with such enormous diseconomies that the population is indifferent between being in this oversized city and in a rural settlement of size 1 (the size of a community formed by a defecting migrant). The problem is the familiar one of coordination failure. For example, in a large economy growing in population, we would like population increases to be accommodated by increases in city numbers all of efficient size. But timely formation of the next city to accommodate population growth requires en masse movement of population (drawn in small amounts from each of the many existing cities) into a new city of efficient size. Without coordination in the form of developers or city governments, no such en masse movement is possible, so people wait to migrate from existing cities to a new city until existing cities are all grossly overgrown at their 'bifurcation point'.

Types of cities: foundations for an urban hierarchy

Henderson (1974) conceives of economies composed of cities specializing in different products. His reasoning is that, in particular for standardized manufactured products, economies of scale are ones of 'localization' or are internal to the own industry. So textile producers learn from each other about technology and market conditions by clustering together, but they learn little of interest from any, say, steel producers nearby. But there are costs of collocating steel and textiles in the form of urban diseconomies such as commuting times and congestion. Thus it is efficient (and a unique equilibrium) to have separate steel and textile-type cities. Given differences between the steel and textile industries in the degree of local external scale economies, these city types will have different sizes. With free migration, real incomes per worker in each type of city will be equalized, with both types of cities at the peaks of their particular inverted U's of real income against city size. The numbers of each type of city, along with steel and textile prices, adjust to ensure equalized real incomes across city types; in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
equilibrium the numbers of each type of city yield supplies of steel and textiles equal to national demand for these products.

This idea of urban specialization has two key implications. First, cities trade extensively with each other. Each city has a non-traded goods sector, with housing and retail and personal services, and an export sector where, with urban specialization, widespread inter-city trade is a key aspect of national economic efficiency. Second, while free migration equalizes real incomes across city types, in the bigger types of cities costs of living will be higher. With equalized real incomes, nominal wages must also be higher in bigger cities, reflecting their greater scale economies. Empirical evidence from different countries suggests that, as cities move from a small population (say, 50,000) to very large metro areas, the cost of living and nominal wages both may typically double.

The original system of cities models ignore intercity transport costs; transport costs of goods across cities are zero or infinite, as for housing. The new economic geography literature (Krugman, 1991) appropriately brings transport costs back to the front burner. In simple urban models it is possible to introduce generalized intercity transport, or trade costs without a specific geography, as in Abdel-Rahman (1996). Then, whether cities are all completely specialized depends on the extent of these trade costs. To put cities in ‘real’ geography so as to have endogenous distances and levels of transport costs across cities is very difficult. Fujita, Krugman and Mori (1999) simulate a situation with monopolistic competition in product markets, where cities sell to each other and also to an agriculture population spread along a line. Cities spread out to serve farmers, but the centripetal force limiting the spread is intercity trade costs. In empirical application of urban models, it is helpful to adapt monopolistic competition models to systems of cities models, so each city has a market potential which characterizes the extent of the market for its products. Au and Henderson (2006) do this in the context of estimating the shape to the inverted U curve of real income against city size, for Chinese cities, allowing different types of cities to exist, each with its own inverted U. This is actually the first time the inverted U has been estimated. China has unusual data, such as urban area value-added, and unusual circumstances in the form of migration restrictions that leave some cities undersized and others oversized that allow inverted U's to be traced out.

There are two empirical literatures on the extent to which production patterns differ across space. The first describes the extent of urban specialization. Starting with Bergsman, Greenston and Healy (1972), researchers use cluster analysis to group cities into categories based on similarity of production patterns – correlations (or minimum distances) in the shares of different industries in local employment. Using 1990 data for the USA, Black and Henderson (2003) group 317 metro areas into 55 clusters, ‘defining’ 55 city types based on patterns of specialization for 80 two-digit industries. There are textile, primary metals, machinery, electronics, oil and gas, transport equipment, health services, insurance, entertainment, diversified market centre, and so on, types of cities, where anywhere from five to 33 per cent of local employment is typically found in just one industry. Specialization in manufactured products, especially among smaller cities, tends to be absolute; many cities have absolutely zero employment in industries such as computers, electronic components, aircraft, instruments, and metal-working machinery.

A related but different perspective is the literature on spatial concentration of industry – the extent to which a particular industry is found in a few as opposed to many locations. The key paper is by Ellison and Glaeser (1999), who model the counterfactual to concentration: the spatial patterns that would arise

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
if location patterns of different industries were determined by random draws across national employment for all industries, where the total number of draws per location is determined by each location's share of national total employment. Ellison and Glaeser develop a specific index measuring the degree of industry concentration relative to this counterfactual, and argue that almost all industries display some degree of spatial concentration. In a related paper, Duranton and Overman (2005) develop a framework to test statistically whether observed industry spatial distributions are statistically significantly different from 'counterfactual' distributions if locations were chosen randomly. They examine the distribution of all pairwise distances between plants in an industry, where distributions which have a greater mass of short pairwise distances are more spatially concentrated.

This literature on specialization and concentration of manufacturing activities tends to overlook two key facts about the process. First, simple indices of urban diversity indicate that smaller cities are very specialized and larger cities highly diversified. Second, in modern economies like the USA the ratio of business services to manufacturing tends to rise sharply as we move up the urban hierarchy. Given these facts, what is the role of large metro areas in an economy and their relationship to smaller cities? A recent literature has focused on the role of large metro areas as centres of innovation, headquarters, and business services. In examining one key aspect of this role, Duranton and Puga (2005) model functional specialization. Manufacturing activities require business service activities, summarized as 'headquarter functions', for organization of production and final sale of goods. Firms may locate headquarters in large cities, away from production facilities in smaller specialized cities, so as to purchase local intermediate business service inputs such as R&D, marketing, financing, exporting, and so on. Urban agglomeration economies in these metro areas arise essentially from shopping agglomeration benefits – the wide diversity of intermediate business service firms in large metro areas from which headquarters can shop in making their outsourcing decisions. In a related earlier paper, Duranton and Puga (2000) develop another aspect of the role of large metro areas. They are modelled as diversified cities, where new firms may experiment with different technologies until they learn the one that best satisfies their needs; once firms have chosen a standard technology they decentralize to smaller specialized types of cities. This model captures a critical phenomenon. Large diversified metro areas may serve as high-tech, innovative, and R&D incubators where new products are born; and then, following the product cycle model, once production of a product is standardized it moves to small cities or offshore, where land and wage costs are cheaper.

The effect of trade and other policies on urban systems

As noted earlier, at the national level in a large economy with many cities, at the limit there are constant returns to scale, or replicability. If national population doubles, the numbers of cities of each type and national output of each good simply double, with individual city sizes, relative prices and real incomes unchanged. As such, in neoclassical trade models, with two goods and two factors, basic international trade theorems (Rybczynski, factor price equalization, and Stolper–Samuelson) hold (Hochman, 1977). This gives an urban flavour to national policies (Renaud, 1981). For example, trade protection policies favouring the steel industry in relatively large cities over, say, the textile industry in smaller cities will alter national output composition towards steel production and increase the number of large relative to small cities. National urban concentration, or the fraction of the population living in bigger versus

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
smaller types of cities, will correspondingly rise. Similarly, subsidizing an input such as capital for a high-tech product again, say, produced in a larger type of city will cause the numbers of that type of city to increase, again raising urban concentration. A minimum wage policy which sets a minimum nominal wage for an economy may bind in smaller types of cities which have lower equilibrium nominal wages, but not bind in larger types of cities. This distortion will increase the sizes and costs of products produced in smaller types of cities, reducing their numbers and national output, as production moves to products in cities which face no effective constraint (Henderson, 1988).

Growth in a system of cities

In a large system of cities, given replicability, in Henderson and Ioannides (1981) in an exogenous growth model, increases in national population growth are accommodated by increases in numbers of cities, as noted earlier. With exogenous technological change which enhances agglomeration economies or reduces urban commuting costs, efficient city sizes will increase also. Black and Henderson (1999) develop these notions further in an endogenous growth model, where there are two types of cities. Growing dynastic families allocate their members across the two types of cities, which have different technologies and intensities of human capital usage. The only capital in the model is human capital, and as such there is no formal market for it. Families operate as an intra-family capital ‘market’ with members in low capital-intensity cities lending to those in high capital-intensity cities. As such, not only do people in high-intensity human capital cities earn a cost-of-living (positive or negative) premium relative to other types of cities, but their nominal wages must be high enough to pay back the human capital they have borrowed – to earn the returns on their higher required levels of human capital. Under conditions allowing for steady state growth, workers accumulate human capital continuously. Within cities the local stock of human capital generates a knowledge externality, which improves production efficiency – so efficient city sizes continuously increase. These urban knowledge externalities are the source of national economic growth. As economies grow, on the urban side there is parallel growth. Both types of cities grow at the same rate in terms of size; and, with sufficient national population growth, both types of cities also grow in number at the same rate.
For evidence on urban-based growth through knowledge accumulation, Glaeser, Scheinkman, and Shleifer (1995) in a cross-section city growth framework show that cities which had higher median years of schooling in 1960 grew faster from 1960 to 1990 than other cities. Similarly, Black and Henderson (1999) find in a panel context for 1940–90 for the USA that cities with higher fractions of college-educated grow faster over each decade. In terms of parallel growth, papers by Eaton and Eckstein (1997) on France and Japan and by Dobkins and Ioannides (2001) and Black and Henderson (2003) on the USA establish basic facts about the evolution of urban systems. First, there is a wide relative size distribution of cities in large economies that is stable over time. Big and small cities coexist in equal proportions over long periods of time. This fact can be established by direct comparisons of current versus historical relative size distributions. More formally, writers make a division of cities by relative sizes usually into discrete cells, or states. Then, on the assumption of a first-order Markov process, transition matrices and steady-state size distributions can be calculated, where in many cases stationarity of the transition matrix over time can't be rejected. Steady-state distributions tend to be close to beginning- and end-period distributions.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Second, within relative size distributions for a country, individual cities are generally growing in population size over time; and what is considered a big city in absolute terms changes over time. In addition, city numbers are increasing over time with entry of new cities. City sizes in the USA, Japan, and France grew during the 20th century at average annual rates of 1.2–1.5 per cent, depending on the country and exact time interval: rates which involve city sizes increasing 3.3–4.5-fold every century. Worldwide, with an absolute cut-off point of 100,000 to be an urban area, numbers of cities worldwide doubled from 1960 to 2000. And the wide size distribution of cities means that much of the urban action is in smaller cities rather than mega-cities. In 2000, between 70 and 75 per cent of the world's urban population lived in cities of under 2 million.

Third, within countries as they urbanize and grow in population there is entry of new cities and both rapid growth and decline of cities nearer the bottom of the urban hierarchy. However, at the top of the hierarchy city size rankings are remarkably stable over time. In the transition matrix the probability of a city moving out of the highest state, once there, is remarkably low. In Japan and France, the 39–40 largest cities in 1925 and 1876, respectively, were all in the top 50 in 1985 and 1990 respectively; and, at the top, absolute rankings were unchanged (Eaton and Eckstein, 1997). In the USA, for 1900–90 cities in the top decile of rankings stay in that decile indefinitely, with newer cities joining the decile as the total number of cities expands. Why is this? A common answer is physical infrastructure. Large cities have huge historical capital stocks of streets, buildings, sewers, water mains and parks that are cheaply maintained and almost infinitely lived, which give them a persistent comparative advantage over cities without that built-up stock. A second answer is based on concepts in Arthur (1990): large cities offer persistent accumulations of knowledge, allowing them to outcompete smaller cities for new industries, which also makes large cities more adaptable.

Zipf's Law

A heralded empirical ‘fact’ about size distributions of cities within countries is that, at least at the upper tail, they are well approximated by a Pareto distribution, with Zipf's Law applying in many cases (Ioannides and Overman, 2003). If one ranks all cities from smallest to highest (rank 1), Zipf's Law is interpreted as stating that rank times population size across cities in the hierarchy is a constant. Gabaix (1999) argues that if there is a stochastic process whereby individual city growth rates follow Gibrat's Law – the growth rate in any period is unrelated to initial size – then the size distribution that emerges will follow Zipf's Law. One focus of theoretical work has been to try to model why something like Zipf's Law would emerge as cities evolve in an economy. Essential to the modelling is to have shocks to productivity or preferences that follow a random walk; but, to get the result in a model where there is an endogenous number of cities, as opposed to just fixing the number of cities, is difficult and requires very specific assumptions. Rossi-Hansberg and Wright (2004) adapt the Black and Henderson (1999) model, with many types of specialized cities. They group industries and specialized city types into sets. Within each set, industries and city types have the *same* technology, but each individual industry draws its own permanent shock each instant. To have Gibrat's Law lead to Zipf's Law, along with Gabaix (1999) they must impose an arbitrary lower bound on the sizes to which cities can fall. These assumptions lead to Zipf's Law holding for each set of industries, and they show that one can aggregate across sets of industries to get Zipf's Law.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Duranton (2005) tries to model micro-foundations for the stochastic process affecting city sizes and as a result ends up modelling an important, overlooked aspect of city evolution. Duranton has many footloose, mobile products in a quality-ladder model and a limited fixed number of cities. The latest innovation in each product is produced by the monopolist holding the patent, and only this top quality is marketed for any product. Investment raises the probability of a firm making a successful innovation and being the patented next step up in the quality ladder in an industry. But innovation can also lead to the next step up in a different industry – that is, there can be cross-industry innovation. To partake of a winning innovation in a different industry, a footloose industry must locate in the city where the innovator is, so industries move around with cross-industry innovation. Presumably co-location of the inventor and production makes the information needed for the transition to mass production cheaper to exchange. Innovation probabilities depend on R&D expenditures. Industry jumps from city to city according to where the latest innovation is, and city growth also follows a stochastic process. The resulting stochastic process of city growth and decline results in steady-state size distributions that are similar to Zipf's Law. Duranton's formulation has the nice feature that cities have patterns of production specialization which change over time, consistent with the data. Moreover, in this context cities that emerge as big are likely to stay big indefinitely, since they hold so many industries in which an innovation might be realized.

Geography

A variety of studies have examined the role of geography, primarily natural features, in the spatial configuration of production and growth of cities. As noted earlier, in terms of formal modelling of geography we have made little formal progress, given the complexity of the problem (Fujita, Krugman and Mori, 1999). Empirically, Beeson, DeJong and Troeskan (2001) look at US counties from 1840 to 1990. They show that iron deposits, other mineral deposits, river location, ocean location, river confluence, heating degree days, cooling degree days, mountain location, and precipitation all affect the base 1840 county population significantly. However, for 1840–1990 *growth* in county population, only ocean location, mountain location, precipitation, and river confluence matter, with the 1840 population controlled for. That is, first-nature items strongly affected 1840 and hence indirectly 1990 populations; but growth from 1840 to 1990 is independent of many first-nature influences. Other studies focus on the geography of markets and the role of neighbours in influencing city evolution. Dobkins and Ioannides (2001) show that growth of neighbouring cities influences own-city growth, and cities with neighbours are generally larger than isolated cities. Black and Henderson (2003) show that normalized market potential variables encourage growth, if geography is controlled for. High market potential helps explain why north-east US cities maintain reasonable growth, given that for historical reasons they are in the most densely populated area, despite the hypothesized natural advantages of the West.

See Also

- location theory
- spatial economics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- urban agglomeration
- urbanization
- urban production externalities

I thank Yannis Ioannides and Will Strange for helpful comments on the first draft.

Bibliography

Abdel-Rahman, H. 1996. When do cities specialize in production? *Regional Science and Urban Economics* 26, 1–2.

Abdel-Rahman, H. and Fujita, M. 1990. Product variety, Marshallian externalities, and city sizes. *Journal of Regional Science* 30, 165–85.

Arthur, B. 1990. Silicon Valley locational clusters: when do increasing returns to scale imply monopoly. *Mathematical Social Sciences* 19, 235–51.

Au, C. and Henderson, J. 2006. Are Chinese cities too small? *Review of Economic Studies*, in press.

Beeson, P., DeJong, D. and Troeskan, W. 2001. Population growth in US counties, 1840–1990. *Regional Science and Urban Economics* 31, 669–700.

Bergsman, J., Greenston, P. and Healy, R. 1972. The agglomeration process in urban growth. *Urban Studies* 9, 263–88.

Black, D. and Henderson, J. 1999. A theory of urban growth. *Journal of Political Economy* 107, 252–84.

Black, D. and Henderson, J. 2003. Urban evolution in the USA. *Journal of Economic Geography* 3, 343–73.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Dobkins, L. and Ioannides, Y. 2001. Spatial interactions among U.S. cities: 1900–1990. *Regional Science and Urban Economics* 31, 701–32.

Duranton, G. 2005. Urban evolutions: the fast, the slow, and the still. Mimeo, University of Toronto.

Duranton, G. and Overman, H. 2005. Testing for localization using micro-geographic data. *Review of Economic Studies* 72, 1077–106.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Duranton, G. and Puga, D. 2000. Nursery cities: urban diversity process innovation, and the life cycle of products. *American Economic Review* 91, 1454–77.

Duranton, G. and Puga, D. 2005. From sectoral to functional urban specialization, *Journal of Urban Economics* 57, 343–70.

Eaton, J. and Eckstein, Z. 1997. Cities and growth: evidence from France and Japan. *Regional Science and Urban Economics* 27, 443–74.

Ellison, G. and Glaeser, E. 1999. The geographic concentration of US manufacturing: a dartboard approach. *Journal of Political Economy* 105, 889–927.

Flatters, F., Henderson, J.V. and Mieszkowski, P. 1974. Public goods, efficiency, and regional fiscal equalization. *Journal of Public Economics* 3, 99–112.

Fujita, M., Krugman, P. and Mori, T. 1999. On the evolution of hierarchical urban systems', *European Economic Review* 43, 209–51.

Gabaix, X. 1999. Zipf's Law for cities: an explanation. *Quarterly Journal of Economics* 114, 739–67.

Glaeser, E., Scheinkman, J. and Schleifer, A. 1995. Economic growth in a cross-section of cities. *Journal of Monetary Economics* 36, 117–34.

Helsley, R. and Strange, W. 1990. Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* 20, 189–212.

Henderson, J. 1974. The sizes and types of cities. *American Economic Review* 61, 640–56.

Henderson, J. 1988. *Urban Development: Theory, Fact and Illusion*. New York: Oxford University Press.

Henderson, J. and Ioannides, Y. 1981. Aspects of growth in a system of cities. *Journal of Urban Economics* 10, 117–39.

Henderson, J. and Becker, R. 2001. Political economy of city sizes and formation. *Journal of Urban Economics* 48, 453–84.

Hochman, O. 1977. A two factor three sector model of an economy with cities. Mimeo.

Ioannides, Y. and Overman, H. 2003. Zipf's Law for cities: an empirical examination. *Regional Science and Urban Economics* 33, 127–37.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kanemoto, Y. 1980. *Theories of Urban Externalities*. Amsterdam: North-Holland.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99, 483–99.

Mills, E. 1967. An aggregative model of resource allocation in a metropolitan area. *American Economic Review* 57, 197–210.

Renaud, B. 1981. *National Urbanization Policy in Developing Countries*. New York: Oxford University Press.

Rossi-Hansberg, E. and Wright, E. 2004. Urban structure and growth. Mimeo. Stanford, CA: Stanford University.

How to cite this article

Henderson, J. Vernon. "systems of cities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_S000446>
doi:10.1057/9780230226203.1663

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tableau économique

Loïc Charles
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The *Tableau économique* is an important landmark in the history of economics and the earliest attempt to provide a calculable model that can help government in policy making. François Quesnay, who invented the *Tableau* in 1758, provided several versions of it, both in equilibrium and in disequilibrium, to account for the many cases and policies he discussed in his economic writings.

Keywords

hoarding; input–output analysis; Leontief, W.; luxury consumption; Marx, K. H.; mathematics and economics; Mirabeau, V. R.; physiocracy; productive and unproductive consumption; Quesnay, F.; *Tableau économique*

Article

The *Tableau économique* is generally considered the first economic model in its own right ever conceived. François Quesnay wrote the first version (or ‘edition’) in December 1758 and sent it to his most prominent disciple, the marquis de Mirabeau, a few days later. Fortunately, the letter and a manuscript copy of this first version have survived (INED, 2005, pp. 391–403). Over the next ten years, Quesnay produced numerous versions of his *Tableau économique*. Some of these (the second and third versions) he printed privately in Versailles. The others were published in books he co-wrote with Mirabeau, in articles he published in physiocratic periodicals and in *Physiocratie*, a collection of his own essays edited and published in 1767 by Pierre Samuel Du Pont de Nemours, one of his young disciples. There are also several manuscript versions in Quesnay's own hand in Mirabeau's papers at the French National Archives (INED, 2005). However, Quesnay's deep interest in this visual device did not appeal to his contemporaries (Van den Berg, 2002). Even his disciples, who like Mirabeau (1775, pp. 203–4) praised him, calling him a new ‘Socrates’ and the ‘Confucius of Europe’, felt uneasy with the *Tableau*, and, after Quesnay died in 1774, the *Tableau* was ignored in the history of economics until Karl Marx

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
rediscovered it in 1862 (Gherke and Kurz, 1995). It was only with the transformation of economics into a mathematical and diagrammatic science that the *Tableau économique* gained a more important place in economics. At the same time, the *Tableau* has become an object of controversy for both economists and historians. Of the several problems that have been raised, here we have selected the most significant ones. What is the *Tableau économique*? What are its origins? What was its purpose for Quesnay? And how to interpret it? We will conclude this article by assessing the place of Quesnay's *Tableau économique* in the history of economics.

Our understanding of the nature of the *Tableau économique* and its place in Quesnay's economics has deepened since the rediscovery of the 'third version (or edition)' of the *Tableau* by Marguerite Kuczynski in 1965 and the publication in 2005 of several manuscript versions of the *Tableau* from Quesnay's own hand (Kuczynski and Meek, 1972; INED, 2005). It is now clear that the *Tableau économique* per se is only the figure and as such 'just a small part of Quesnay's model' (Pressman, 1994, p. 5). Hence, the 'first edition' consists of three parts: the *Tableau* itself, Quesnay's marginal commentary explaining its working, and the 22 'remarks on the variations in the distribution of the annual revenue of a nation', an expansion of the 14 maxims from the article 'Grains' published in 1757 and which can be best described as a list mixing hypotheses and policy recommendations (see INED, 2005, pp. 198–212). The second and third versions of the *Tableau*, compiled in 1759, are based on the same structure, despite formal changes. The 'Remarks' (23 in the second and 24 in the third version) are presented as an 'extract of the royal economic memoirs of M. de Sully'. The commentary, still in the margins in the second version, is transformed into an 'Explanation of the *Tableau économique*' in the third version. In the version of the *Tableau* found in the 'Analyse de la formule arithmétique du Tableau économique' published in *Physiocratie* (1767), only the *Tableau* and its explanation (now called 'analysis') remained in the same text; the maxims are now expanded to 30, with more numerous and longer notes, and presented in an independent text, 'General maxims for the economic government of an agricultural kingdom'.

The origins of the *Tableau* have been seen as a problem since the end of the 19th century, when it was first remarked that there was a likeness between the zigzag diagram of the first three versions and the process of blood circulation. This hypothesis was developed by Foley (1974) and further refined by Christensen (1994). However, other scholars have criticized this theory and pointed out there was no direct evidence in Quesnay's writings to support it (Rieter, 1983; Eltis, 1984, pp. 20–1). Since 1990, several studies have established that the *Tableau* can be seen as a sort of technological device designed to calculate economic quantities as well as display economic laws. Its features derived from different areas of knowledge, covering, for example, art works, machines and early modern arithmetic, which Quesnay brought together to produce a unique concept (Charles, 2000; 2003; 2004; Rieter, 1990; Wise 1990). Quesnay invented the *Tableau économique* as a visual calculating machine or, as he and Mirabeau put it in the introduction to *Philosophie rurale*, an 'arithmetic rule' and a 'formula of calculation' (Mirabeau and Quesnay, 1763, pp. xix, xxiii).

This leads us to the more general problem of the economic interpretation of the *Tableau*, which has captivated economists since Karl Marx. There have been several articles and books providing interpretations, but no general consensus has emerged and important theoretical aspects of the *Tableau* have been controversial ever since Marx's rediscovery of the *Tableau*. One issue that had been discussed at length by commentators is whether the different versions, in particular the zigzag and the late version

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
from ‘Analyse de la formule arithmétique’, are consistent. The most prominent scholar of physiocracy, R.L. Meek (1962), thought they were, as did several commentators writing under his influence (Barna, 1975; Eagly, 1969; Pressman, 1994). Indeed, the various versions of the *Tableau économique* are based on a single microeconomic unit: a farm made of one plough, four horses and 120 acres of land which is considered as the best technique of production available to agricultural entrepreneurs (Eltis, 1984, pp. 3–13; Herlitz, 1996, p. 17; Cartelier, 1998, p. 249). There are also features common to all versions of the *Tableau*. First, all the *Tableaux* are divided into three columns representing the three classes of citizen relevant to Quesnay's economic representation of society: the productive class (the *classe productive* composed of farmers and professions linked to the primary sector and the marketing of its products), the landowning class (the *classe des propriétaires* which includes landowners, the state and the clergy) and the unproductive class (the *classe stérile*, corresponding roughly to those working in commerce and industry). Second, the *Tableaux* feature a general interdependence between the three classes in the form of rows linking them: these rows signal the mutual expenditures between the classes that take place in the economic system.

However, the two main versions of the *Tableaux* presented significant differences. First, the *Tableau* in zigzag is an open system that functions as a ‘table of expenditure and reproduced produce’, and includes a propagation effect that resembles the Keynesian multiplier and accelerator effect (Herlitz, 1996, p. 13; Hishiyama, 1960, pp. 124–6; Meek, 1962, p. 293). Conversely, the mature version of the *Tableau économique* was designed to provide ‘a consistent account of social reproduction’ as a whole and leave out many details of the process of expenditure and reproduction to concentrate on the general results (Barna, 1975; Herlitz, 1996). These two main versions of the *Tableau* are also based on different economic models (Cartelier, 1982; Herlitz, 1961; 1996). In the zigzag version, emphasis is put on the process of expenditure: it is the spending of the landowning class that comes first in the graphical representation, and is seen to initiate the circulation of wealth in the economy (see the two diagrams in Quesnay, François). In Quesnay's economic model, the prominence of the landowning class is made clear by the fact that the equilibrium of the whole system depended on their expenditure. When landowners spend one-half of their revenue on agricultural products and the other on industrial goods, the *Tableau* (and the model) is in equilibrium and simply reproduces itself without changes. When landowners spend more on industrial goods – indulging in ‘*luxe de décoration*’ (excess consumption of luxury goods) – equilibrium is disrupted and the *Tableau* as well as the economy are in decline. Conversely, when they spend more on agricultural products – indulging in ‘*faste de subsistence*’ (increased consumption of subsistence goods) – the economy grows and produces a larger economic surplus.

In the latest version of the *Tableau*, emphasis has switched to the expenditure of the productive class in annual advances (invested in production) and in rent payments to the landowners. In this version, the prominent role goes to the farmers and their advances (which do not figure in the zigzag version), which initiate the process of reproduction of wealth. Conversely, the landowners’ expenditure now appears contingent and the landowning class unnecessary for the functioning of the system: the fact that it is the landowners who seize the disposable surplus in the form of rent is arbitrary and linked to the historical context of the society depicted by Quesnay (*Ancien Régime* France), but has no economic justification (see Cartelier, 1982). The coexistence of two alternative versions of the *Tableau* corresponded to two policy issues underlined by Quesnay in his economic maxims: (*a*) the necessity of productive advances

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and (*b*) both the moral and the economic imperative of virtuous spending on the part of the landowning class (no excess luxury consumption). At the same time, it signals problems that Quesnay was unable to resolve satisfactorily. As several commentators have noted, in the latest version of the *Tableau* the landowners' consumption has no bearing on the economic equilibrium, hence there is no need for the hypothesis that landowners spend half their money on agricultural products and the other half on industrial goods; it is enough that they spend all their revenue and do not hoard (Barna, 1975; Bilingsoy, 1994; Cartelier, 1982; Negishi, 1989).

The place of the *Tableau économique* in the history of economics has increased in importance dramatically from the 1940s onwards. It has been interpreted either as a forerunner of neoclassical general equilibrium (Samuelson, 1982; Schumpeter, 1954), as a Leontief input–output system and more generally a linear system (see Phillips, 1955; Maitel 1972; Barna, 1975; Bilingsoy 1994), while Marxist authors have interpreted the *Tableau* as a rationalization of *Ancien Régime* society (Fox-Genovese, 1976; Gleicher, 1982). The first interpretation, notwithstanding the authority of Schumpeter (and Samuelson!) holds only at the more general level since there is no trace of marginal analysis in Quesnay's economic model. There is much more ground to link the *Tableau économique* to input–output analysis (for a different opinion, see Pressman, 1994, ch. 5). Indeed, Leontief himself has suggested, if rather cryptically, that Quesnay's work had been an important landmark in the development of his own ideas (Leontief, 1936, p. 105). Moreover, like Leontief, Quesnay was interested in providing both a theoretical and an empirical model of the economy at the same time (Barna, 1975; Meek, 1962, p. 296).

According to the Marxist interpretation, the *Tableau économique* exemplified the contradiction of Quesnay's social thought, and more generally of *Ancien Régime* France, caught between the feudal order, characterized by the prominence of landlords, and the burgeoning capitalist economy, characterized by the role played by the capital investments of farmers. Since the fall of the Berlin Wall in 1989 this interpretation has been on the wane among economists and historians alike.

More recently, the attention of economists interested in Quesnay's economics has turned towards several disequilibrium and underemployment-of-resources equilibrium *Tableaux* (Barna, 1976; Charles, 2000; Eltis 1984, ch. 2; 1996; Pressman, 1994, chs 4, 6 and 7). These are particularly noteworthy since they are concerned with the possibility of growth or decline and are linked to specific policy issues. There is no place here to detail these different figures (more than 50 in total!), but it may be useful to list the different economic cases investigated which give rise to these *Tableaux*. First, Quesnay used the *Tableau* to study the effects of hoarding and of excessive consumption of luxury goods by the landowning class. Second, Quesnay investigated the consequences of an unjust tax system with the case of a tax on the productive sector (in Quesnay's theory tax should be levied on the landowning class) and the case of the *Ancien Régime*'s (costly) taxation system (the *fermes générales*). Third, Quesnay discussed the joint cases of low agricultural prices due to impediments in external trade and of the economic effects of the establishment of free export in agricultural products (and the rise in prices it causes). Finally, other *Tableaux* are used to show the consequences of policies encouraging the unproductive industrial sector at the expense of the productive agricultural sector.

All in all, Quesnay's *Tableau économique* is now considered by most economists as an important landmark in the history of economics and the earliest attempt to provide a calculable model that can be used by governments for policymaking.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- classical growth model
- Du Pont de Nemours, Pierre Samuel
- Ephémérides du citoyen ou chronique de l'esprit national
- physiocracy
- Quesnay, François

Bibliography

Barna, T. 1975. Quesnay's *Tableau* in modern guise. *Economic Journal* 85, 485–96.

Barna, T. 1976. Quesnay's model of economic development. *European Economic Review* 8, 315–38.

Bilingsoy, C. 1994. Quesnay's *Tableau économique*: analytics and policy implications. *Oxford Economic Papers* 46, 519–33.

Cartelier, J. 1982. De l'ambiguïté du Tableau économique. *Cahiers d'économie politique* 9, 39–63.

Cartelier, J. 1998. Quesnay, François, and the Tableau économique. In *The Elgar Companion to Classical Economics*, ed. H. Kurz and N. Salvadori. Cheltenham and Northampton: Edward Elgar.

Charles, L. 2000. From the *Encyclopédie* to the *Tableau économique*: Quesnay on freedom of the grain trade and economic growth. *European Journal of the History of Economic Thought* 7, 1–22.

Charles, L. 2003. The visual history of the *Tableau Économique*. *European Journal of the History of Economic Thought* 10, 527–50.

Charles, L. 2004. The Tableau Économique as rational recreation. *History of Political Economy* 36, 445–74.

Christensen, P.P. 1994. Fire, motion, and productivity: the proto-energetics of nature and economy in François Quesnay. In *Natural Images in Economic Thought*, ed. P. Mirowski. Cambridge: Cambridge University Press.

Eagly, R. 1969. A physiocratic model of dynamic equilibrium. *Journal of Political Economy* 77, 66–84.

Eltis, W. 1984. *The Classical Theory of Economic Growth*. London: Macmillan.

Eltis, W. 1996. The *Grand Tableau* of François Quesnay's economics. *European Journal of the History*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Economic Thought 3, 21–43.

Foley, V. 1974. An origin of the Tableau économique. *History of Political Economy* 5, 121–50.

Fox-Genovese, E. 1976. *The Origins of Physiocracy: Economic Revolution and Social Order in Eighteenth-Century France*. Ithaca: Cornell University Press.

Gherke, C. and Kurz, H. 1995. Karl Marx on physiocracy. *European Journal of the History of Economic Thought* 2, 53–90.

Gleicher, D. 1982. The historical bases of physiocracy. *American Journal of Economics and Sociology* 46, 328–60.

Herlitz, L. 1961. The *Tableau économique* and the doctrine of sterility. *Scandinavian Economic History Review* 9, 11–51.

Herlitz, L. 1996. From spending reproduction to circuit flow and equilibrium: the two conceptions of *Tableau économique*. *European Journal of the History of Economic Thought* 3, 1–20.

Hishiyama, I. 1960. The *Tableau économique* of François Quesnay. *Kyoto University Economic Review* 30, 1–46.

INED (Institut National d'Études Démographiques). 2005. François Quesnay. *Œuvres économiques complètes et autres textes*, 2 vols, ed. C. Théré, L. Charles and J.C. Perrot. Paris: INED.

Kuczynski, M. and Meek, R.L. 1972. *Quesnay's Tableau économique*. London and New York: Macmillan and Augustus M. Kelley.

Leontief, W. 1936. Quantitative input and output relations in the economic system of the United States. *Review of Economic Statistics* 18, 105–25.

Maitel, S. 1972. The *Tableau Économique* as a Leontief model: an amendment. *Quarterly Journal of Economics* 86, 504–7.

Meek, R. 1962. *The Economics of Physiocracy*. London: George Allen & Unwin.

Mirabeau, V.-M. 1775. Eloge funèbre de François Quesnay. *Nouvelles Ephémérides Economiques* 1, 197–216.

Mirabeau, V.-M. and Quesnay, F. 1763. *Philosophie rurale, ou économie générale et politique de l'agriculture réduite à l'ordre immuable des loix physiques et rurales, qui assurent la prospérité des*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

empires. Amsterdam: Libraires associés.

Negishi, T. 1989. Expenditure patterns and international trade in Quesnay's *Tableau économique*. In *Developments in Japanese Economics*, ed. R. Sato and T. Negishi. Tokyo: Academic Press.

Phillips, A. 1955. The *Tableau économique* as a simple Leontif model. *Quarterly Journal of Economics* 69, 134–44.

Pressman, S. 1994. *Quesnay's Tableau économique. A Critique and Reassessment*. New York: Augustus M. Kelley.

Quesnay, F. 1767–68. *Physiocratie ou constitution naturelle du gouvernement le plus avantageux au genre humain*, ed. P.-S. Du Pont [de Nemours]. Merlin: Leyde and Paris.

Rieter, H. 1983. Zur Reception der Kreislaufanalogie in der Witschaftswissenchaft. *Studien zur Entwicklung der ökonomischen theorie III*. Berlin: Duncker-Humblot.

Rieter, H. 1990. Quesnays *Tableau Économique* als Uhren-Analogie. *Studien zur Entwicklung der ökonomischen theorie IX*. Berlin: Duncker-Humblot.

Samuelson, P.A. 1982. Quesnay's *Tableau économique* as a theorist would formulate it today. In *Classical and Marxian Political Economy*, ed. I. Bradley and M. Howard. London: Macmillan.

Schumpeter, J.A. 1954. *History of Economic Analysis*. London: Allen & Unwin.

Van den Berg, R. 2002. Contemporary responses to the *Tableau économique*. In *Is There Progress in Economics? Knowledge, Truth and the History of Economic Thought*, ed. S. Boehm et al. Cheltenham: Edward Elgar.

Wise, N.M. 1990. Mediations. Enlightenment balancing acts, or the technologies of rationalism. In *World Changes: Thomas Kuhn and the Nature of Science*, ed. P. Horwich. Cambridge, MA: Harvard University Press.

How to cite this article

Charles, Loïc. "Tableau économique." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000245> doi:10.1057/9780230226203.1664

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

taking (eminent domain)

Perry Shapiro
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Compensation for a public taking of private property (typically land) can affect both government's and landowners' decisions, and compensation rules affect both real and perceived fairness. Arguably, no compensation is efficient; landowners will account for probable loss when making their investment decisions, since such decisions are determined by events outside the discretion of either the landowners or the government (though if the taking decision is made to benefit government or landowners, zero compensation may be inefficient). But zero compensation is unconstitutional (in the United States and Australia) and inequitable. A trade-off between efficiency and equity is therefore normally unavoidable.

Keywords

compensation calculus; efficiency–equity trade-off; Epstein, R.; Fairness; just compensation; land use; public property; rent seeking; taking; Fifth Amendment (US Constitution)

Article

The Fifth Amendment of the US Constitution ends with 'nor shall private property be taken for public use without just compensation'. Similarly, the Australian Constitution, in section 51(xxxi), allows '[t]he acquisition of property on just terms …'. Both constitutions give government the power to condemn property for a public purpose and both specify the requirement for compensation. However, they leave unanswered the formula for computing 'just compensation'. Some dispute that government should be permitted to force the sale of private property through condemnation, but it is widely held that many functions of government, particularly those that require the assemblage of contiguous plots of land, would be impractically complicated without the power of eminent domain. For most, the question is not whether the government should have the right to compulsory acquisition but, rather, what compensation should be paid to the private landowners.

There are two distinct issues when considering compensation for a public taking of private property. The first is that compensation rules can affect the decisions of both government and landowners. The rate at

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
which landowners must be paid for their condemned property may affect the public's choice of how much land to condemn. In addition, the knowledge of potential compensation can influence landowners' choice of property improvements. If compensation is based on the market value of land, landowners will over-improve their property because they do not consider the possibility of a taking (moral hazard) (Blume, Rubinfeld and Shapiro, 1984); and if improvements increase compensation landowners are inclined to over-invest in their property to favorably affect their settlement with the government (rent seeking) (Fischel, 1995 p. 296).

The second issue is that compensation rules affect both real and perceived fairness, and it is likely that equity is what just compensation is about. In the US Supreme Court 1960 decision on *Armstrong* v. *United States*, the majority opinion was that '[t]he Fifth Amendment guarantee … [is] designed to bar Government from forcing some people alone to bear public burdens which, all fairness and justice, should be borne by the public as a whole.' A large group of people, those whose property is not condemned, benefit at the expense of a much smaller minority who must surrender their property. Even when compensation is equal to the pre-taking market value of property, as is most commonly the case, the owners of condemned property lose relative to those escaping condemnation.

Blume, Rubinfeld and Shapiro (1984) use an example of land in a river valley to illustrate potential moral hazard loss. With known probability, p, the price of oil will rise to a price sufficiently high that it is in the public interest to dam the river and flood the valley. The structures invested on the land are all lost under the reservoir waters. With known loss probability, p, the efficient level of investment on the land equates the expected marginal product (the product of p and the marginal product of capital) with r, the market return on capital. If the landowners are fully compensated for both their lost land and immovable capital, the probability of dam-caused flooding will not affect the level of investment on the land. Their investment choice will equate the marginal product of invested capital with the market rate of return (MP=r). The result is that an inefficient amount of capital will be invested in the river valley. The conclusion is that no compensation is efficient because landowners will correctly account for probable loss when making their investment decision.

The recommendation for no compensation is based on the presumption that, while uncertain, the decisions whether or not to condemn land and, if so, how much land are determined by events outside the discretion of either the landowners or the government. However, if the taking decision is made to benefit either the government or the landowners, zero compensation induces inefficient decisions (Fischel and Shapiro, 1988; 1989). For instance, if the government represents the interest of a subset of its citizens (for example, the majority) and does not act to maximize social welfare, the need to make compensatory payments to the owners of condemned land will put a beneficial constraint on the government's propensity to condemn an inefficiently large amount of land. However, even if the government is venal and self-serving, it is never efficient to compensate landowners for 100 per cent of lost value. For the sake of social efficiency the landowners, even in the face of bureaucratic venality, must account for the probability of a taking, even if the amount is inefficiently chosen.

Efficiency is only part of the public policy story. While it might be efficient for government to take land without compensation, it nonetheless offends our notion of what is fair. It is unlikely that any policy as draconian as the one suggested would be adopted for the physical acquisition of private property. (Public regulations that restrict the use of private property commonly are not thought to require compensation.) Uncompensated condemnation is not consistent with a constitution that spells out both the powers and the limits of government, as does the US Constitution. Uncompensated condemnation appears as a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
bullying big government strong-arming a small minority of landowners.

Frank Michelman (1967) proposes a compensation calculus incorporating fairness. Michelman includes an explicit and quantifiable measure he labels ‘demoralization cost’. Demoralization is the personal (psychological) reaction to a government leviathan that runs roughshod over a landowning minority. It is manifest in two different ways: the outpouring of sympathy for the downtrodden, and a concern that the same can happen to you. Citizens empathize with the taken and, simultaneously, worry about the sanctity of their own property rights.

Whether or not Michelman's calculus is used, it is important for real public policy to balance the potential inefficiencies resulting from compensation with the inequities without it. While in most cases it is impossible to achieve efficiency without sacrificing some degree of fairness, or to achieve a fair outcome without sacrificing efficiency, there are special cases for which this is not true. If somehow the interests of the private landowners and the government can be aligned with social welfare, the investment and taking choice can be equitable as well as efficient.

In his discussion of disproportionate impact, Richard Epstein (1985) argues that, if prospective compensation does not affect investment choices, the interests of the landowners and government are the same if takings require full compensation. The case against full compensation is that it induces inefficient investment choices. Without the resource-use concerns, land prices serve to direct government to make only welfare-increasing decisions about condemnation.

For certain types of public projects – those for which changes in land values reflect all the benefits – compensation equal to the post-taking enhanced land value has favorable efficiency and equity consequences. It is equitable because those who lose their property are rewarded equally to those who are lucky enough to escape condemnation. It is efficient because the possibility of condemnation is independent of individual property improvement.

If the benefits of the public taking are specific with measurable market values, it is possible to devise a compensation scheme that achieves both equity and efficiency. However, the dual goals are unattainable when project benefits are diffuse and immeasurable. With these more common cases, it is necessary to consider the trade-off between equity and efficiency. The Michelman calculus is useful in expressing this trade-off if the underlying quantities (discouragement costs) are truly measurable.

See Also

- justice
- property law, economics and

Bibliography

Blume, L., Rubinfeld, D. and Shapiro, P. 1984. The taking of land: when should compensation be paid? *Quarterly Journal of Economics* 99, 71–92.

Epstein, R. 1985. *Taking: Private Property and the Power of Eminent Domain*. Cambridge, MA: Harvard University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fischel, W. 1995. *Regulatory Takings, Law Economics and the Power of Eminent Domain*. Cambridge, MA: Harvard University Press.

Fischel, W. and Shapiro, P. 1988. Taking, insurance and Michelman: comments on economic interpretation of just compensation law. *Journal of Legal Studies* 17, 269–93.

Fischel, W. and Shapiro, P. 1989. A constitutional choice model of compensation for takings. *International Journal of Law and Economics* 9, 115–28.

Michelman, F. 1967. Property, utility and fairness: comments on the ethical foundations of just compensation law. *Harvard Law Review* 80, 1165–258.

How to cite this article

Shapiro, Perry. "taking (eminent domain)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_P000325>
doi:10.1057/9780230226203.1665

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tarbell, Ida Minerva (1857 – 1944)

Alice H. Amsden
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Fordism; industrial relations; marginalism; scientific management; Tarbell, I. M.; Taylorism

Article

In a profession where women have been denied the liberties of expression otherwise permitted to men of lesser quality, Ida M. Tarbell made her mark as an economic journalist. She had a keen sense of ethical issues, but regrettably, her blend of reformism and conservatism was sometimes bewildering.

Born in 1857 in Erie County, Pennsylvania, Tarbell is best known for her *History of the Standard Oil Company* (1904), a two-volume attack on the ruthlessness of the oil monopolies (her father was ruined by them). As a muckraker, Tarbell could be expected to favour state intervention in wage setting, then a hotly debated issue. But ironically, whereas the progenitor of marginal productivity theory, J.B. Clark, edged towards arbitration to reduce labour strife, Tarbell embraced Taylorism, whose logic of work atomization builds on the marginalist principle. From 1912 to 1915 Tarbell toured factories she handpicked to study industrial conditions. Favourably struck with Fordism, she wrote a contemporary equivalent of the ‘excellently managed corporation’ entitled *New Ideals in Business* (1916). The ideals were scientific management, humanistic labour relations and a belief in the fundamental goodness of entrepreneurs.

In her feminist outpourings, Tarbell is better remembered for the way she lived than by what she wrote. Unusual for a woman at the time, Tarbell moved to Paris after college to study women in the French Revolution, as praised by Woodrow Wilson for her ‘common sense’ views on the tariff (which she opposed), attended the Paris Peace Conference, corresponded with notables, including Richard T. Ely, interviewed Mussolini, and shunned marriage for a career. The same Tarbell, however, fought against women's suffrage and in *The Business of Being a Woman* (1912) advised members of her sex to stay at home.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1904. *The History of the Standard Oil Company*. New York: McClure, Phillips.

1911. *The Tariff in Our Times*. New York: Macmillan.

1912. *The Business of Being a Woman*. New York: Macmillan.

1915. *The Ways of Woman*. New York: Macmillan.

1916. *New Ideals in Business: An Account of Their Practice and Their Effects Upon Men and Profits*. New York: Macmillan.

1926. *The Life of Elbert H. Gray: the Story of Steel*. New York: D. Appleton.

1936. *The Nationalizing of Business, 1878–1898*. New York: Macmillan.

How to cite this article

Amsden, Alice H. "Tarbell, Ida Minerva (1857–1944)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000009>
doi:10.1057/9780230226203.1666

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tariff versus quota

Arvind Panagariya
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Bhagwati (1965) first demonstrated that if perfect competition prevails in all markets, a tariff and import quota are equivalent in the sense that an explicit tariff reproduces an import level that, if set alternatively as a quota, produces an implicit tariff equal to the explicit tariff, and vice versa. This equivalence breaks down, for example, if the domestic production is monopolized. In this case, replacing an explicit tariff by an import quota set at the level equal to the imports under the explicit leads to a higher implicit tariff. Many other cases of the breakdown of the equivalence also arise.

Keywords

directly unproductive profit-seeking; tariff versus quota; tariffs; uncertainty; voluntary export restraints

Article

The ‘tariffs versus quota’ literature was stimulated by the seminal contribution by Bhagwati (1965). Bhagwati defined the two instruments as equivalent if an explicit tariff reproduces an import level that, if set *alternatively* as a quota, produces an implicit tariff equal to the explicit tariff and vice versa.

Equivalence and its breakdown

Bhagwati (1965) demonstrated that the tariff–quota equivalence necessarily obtains when perfect competition prevails in all markets. This is shown most simply in the small-country context. By definition, the small country faces a perfectly elastic supply at a given price in the world market. In Figure 1, DD and SS represent the demand and supply curves in this (small) country and P^* fixed world prices. Under free trade, that country produces Q_0, consumes C_0 and imports Q_0C_0. The imposition of an explicit tariff *t* per unit raises the internal price in the country to $P=P^*+t$ and the output and consumption move to Q_1 and C_1, respectively. Imports decline to Q_1C_1. The consumer surplus declines by the trapezium formed by the sum of the areas marked *b*, *e*, R and *f*. Of this, area *b* is recovered by producers as extra surplus and area R by the government as tariff revenue. Areas *e* and *f* are lost entirely and called deadweight losses. Area *e* is lost because the marginal cost of production of Q_0Q_1 exceeds the world price. Area *f* is lost because the tariff forces the consumers to stop before the marginal benefit at P exceeds the marginal social cost of obtaining the goods at P^*.

Figure 1
Equivalence under perfect competition and non-equivalence under monopoly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

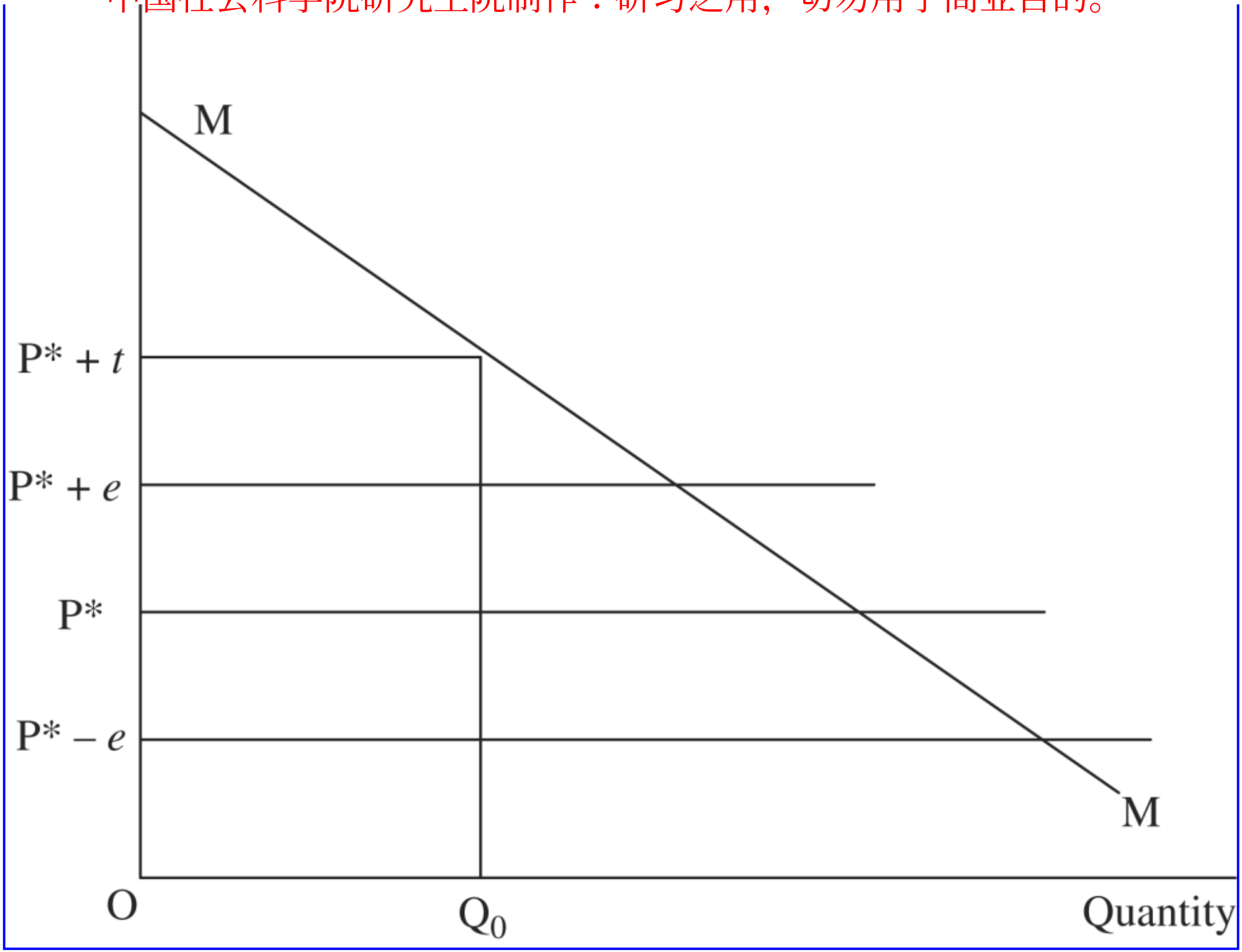

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

If we now replace the tariff by the import Q_1C_1 and auction the quota licences competitively, imports would equal the quota. Subtracting these imports from DD at each price, we obtain D′D′ as the demand facing domestic producers. The internal price now obtains at the intersection of SS and D′D′. But by construction, this is $P=P^*+t$, with t now representing the implicit tariff. Explicitly, t is now the price of the licence per unit of imports. The total revenue from the auction of the licences equals R. The outcome is identical to that under a tariff in every way.

Suppose now that a monopoly producer supplies the domestic output with SS representing his marginal cost curve. Under the tariff t, the monopolist cannot raise the price above P^*+t so that the outcome is no different from under perfect competition. But if we replace the tariff by import quota Q_1C_1, with the quota licences auctioned competitively, the monopolist faces the demand curve D′D′. Associated with D′D′ is a marginal revenue curve (which, for the sake of simplicity, is not shown in Figure 1) whose intersection with SS gives the monopoly output Q_M. The price the monopolist changes at this quantity is P_M, which is higher than P^*+t. The equivalence breaks down. Non-equivalence also obtains if we replace domestic competitive suppliers by an oligopoly rather than a monopoly (Helpman and Krugman, 1992). All these results can be generalized to the large-country case.

Alternatively, non-equivalence obtains if we assume perfect competition in demand and supply but not the allocation of quota. For example, if the holder of the quota licence is a monopolist, he would maximize the quota rents. The solution in this case may involve leaving some licences, thereby raising the domestic price above P^*+t.

Retaining perfect competition in all markets, non-equivalence also arises if the quota takes the form of a voluntary export restraint (VER). Under the VER, enforcement of the quota is the responsibility of the exporting country. In this case, the exporting country captures the quota rent, and the welfare loss from the quota to the importing country is larger than under the tariff or direct import quota.

A further case of non-equivalence arises in the presence of uncertainty. This is shown in a simple manner using a construction from Pelcovits (1976). In Figure 2, suppose the import demand by home country is MM and the world price can be either P^*+e or P^*-e, each with a probability of one half. Suppose further that we want to restrict expected imports to OQ_0. Regardless of which world price is realized, an import quota will hit the target exactly with the domestic price given by the height of MM at Q_0 and denoted P^*+t. If a tariff is to be used to achieve the same objective, assuming risk-neutral behaviour, we would set the tariff at t. The domestic price in this case could be either P^*+t+e or P^*+t-e, each with a probability of one half. The reader can verify that the expected welfare losses under the quota and tariff would be different, which implies non-equivalence.

Figure 2

Non-equivalence under uncertainty

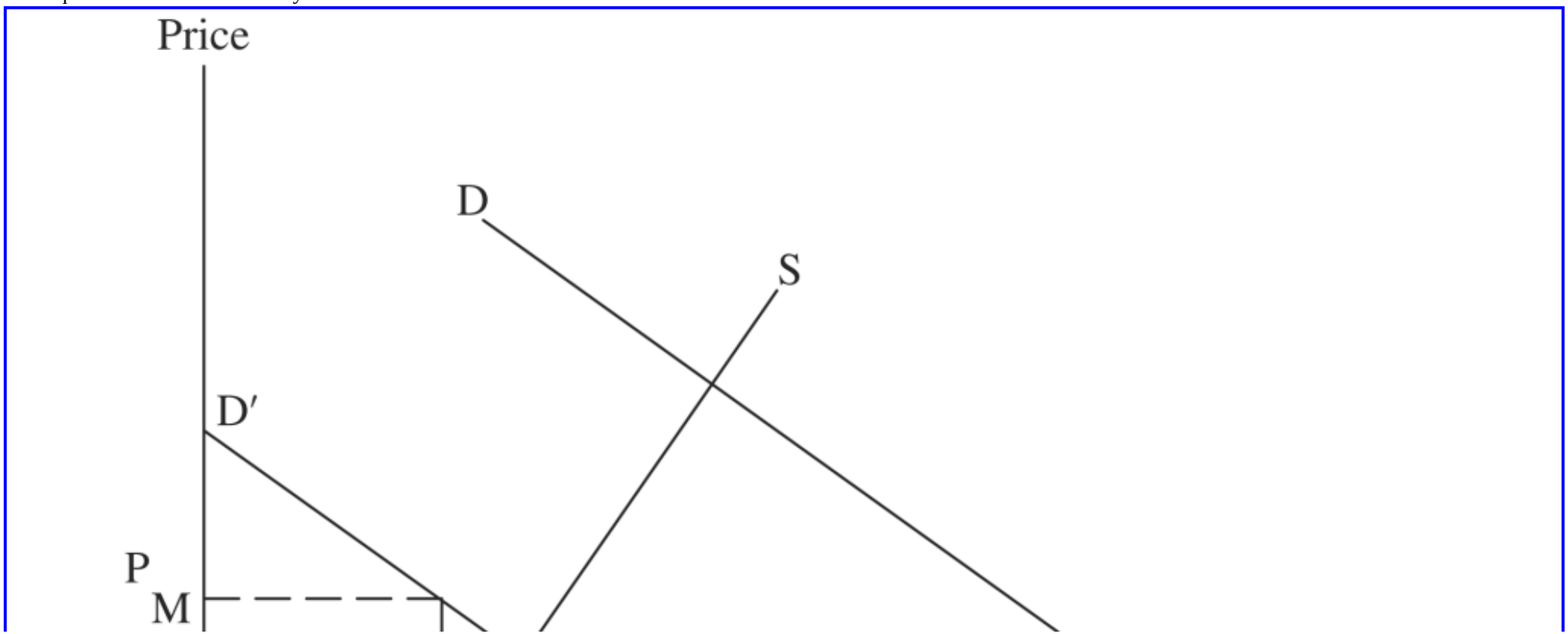

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

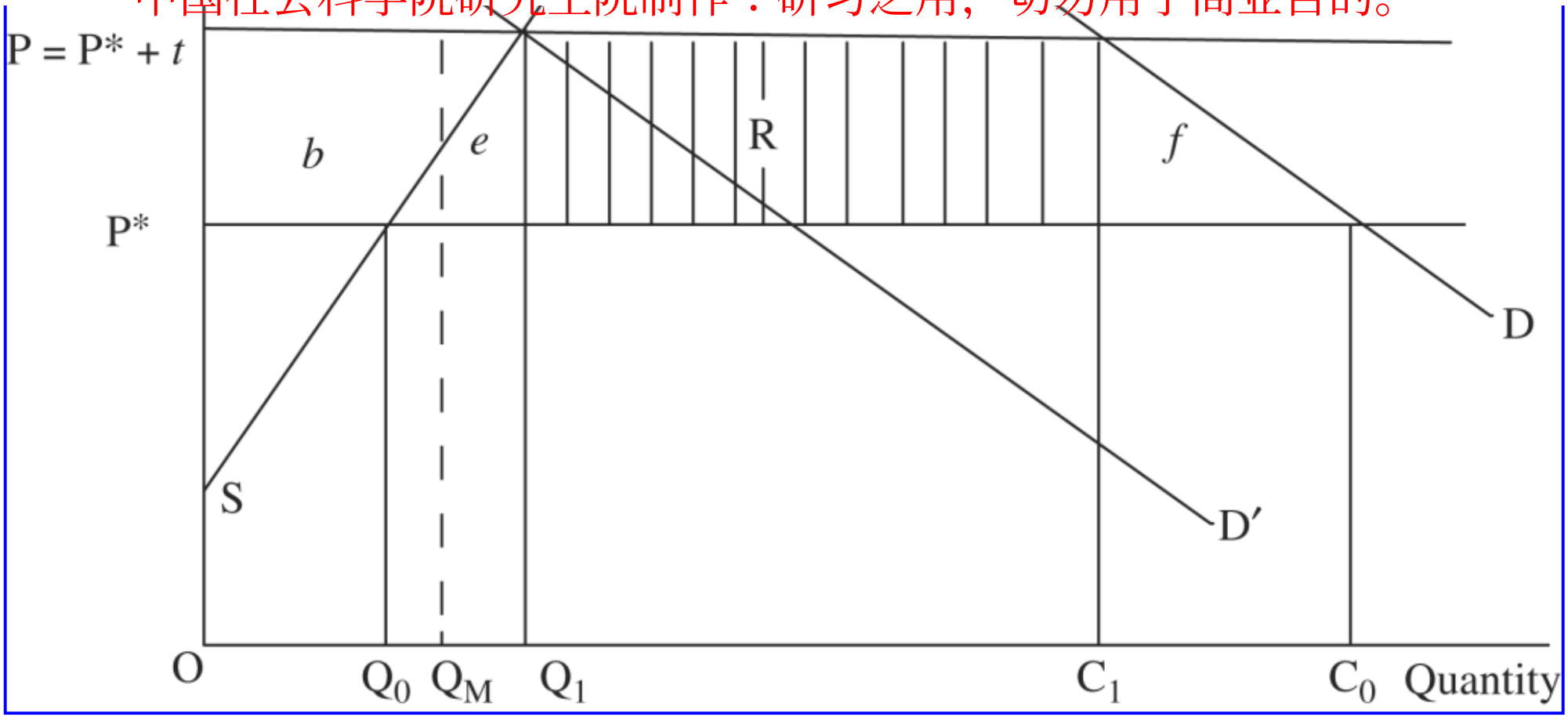

Panagariya (1980) considers the equivalence of optimal tariff and quota *structures* in a small-country model with two or more imports. He considers a government wishing to restrict the value (at the world prices) of two or more imports to a pre-specified level. If perfect competition prevails in all markets, the optimal policy is either an explicit uniform tariff on the restricted set of goods or import quotas on them at levels that imply a uniform implicit tariff at the same rate. If domestic production of these goods is monopolized, however, optimal tariffs are still uniform, but optimal import quotas are characterized by implicit tariffs that are generally non-uniform.

In two companion papers, Panagariya (1981; 1982) brings out yet another aspect of tariff–quota non-equivalence in the presence of domestic monopoly. He considers a large-country, general-equilibrium model in which domestic industry is monopolized. He shows that in such a model, exogenous changes in quotas and tariffs lead to qualitatively different outcomes. For example, if quota is the instrument, tightening it always improves the terms of trade. But if tariff is the instrument, raising it may lead to deterioration in the terms of trade.

Finally, Rodriguez (1974) and Tower (1975) have independently considered the outcomes when two countries optimally choose trade interventions in a Nash non-cooperative game within a two-good general-equilibrium model. They show that if the countries choose tariff as the instrument, an equilibrium characterized by finite positive trade between them generally exists. But if the countries employ quotas, such equilibrium does not exist.

Welfare ranking of tariffs and quotas

When tariffs and quotas are not equivalent and we use them to target some variable in the economy, a natural question concerns the welfare ranking of the two instruments. For example, if we assume that the domestic production is monopolized but perfect competition prevails everywhere else and the objective is to restrict imports to a specified level at the lowest cost, tariff is superior to quota. This is readily seen in the small-country case in which the tariff forces the monopolist to behave like a competitor whereas the quota allows him to earn positive monopoly rents. Similarly, on the assumption that there is perfect competition on all sides in the product market, if the quota holder is a monopolist, tariff would once

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
again be the superior instrument.
Two additional ranking results are due to Panagariya (1980) and Pelcovits (1976). The former considers the ranking of tariffs and quotas when a small country aims to restrict the value of a subset of imports (at world prices) to a fixed level and these goods are subject to the monopoly distortion. He finds tariffs to be an instrument superior to quotas. Pelcovits considers the welfare ranking in the presence of uncertainty. In the small-country context, on the assumption that the world price is stochastic and the country wishes to constrain the expected imports at a pre-specified level, he asks whether quotas yield higher expected welfare or tariffs. Using a construction similar to that in Figure 2, he shows that the answer is ambiguous.

Welfare outcomes with pre-existing tariffs and quotas

A final question of interest is how the welfare outcomes differ when a parameter is altered in the presence of tariffs versus quotas. The first set of contributions in this category comes from the so-called piecemeal trade reforms literature that asks how welfare changes as we relax one trade barrier at a time. Corden and Falvey (1985) demonstrate that in a small country with many imports, if the country restricts trade by quotas only, the relaxation of any quota necessarily improves welfare. Intuitively, the relaxation of the quota reduces the distortion in that that good has no effect on the distortion in the other goods since their imports face the same quota as before. Therefore, the net effect of the change on welfare is positive. This is not true if imports are restricted by tariffs. A reduction in any one tariff directly improves welfare by expanding the imports of that good. But it may indirectly lower welfare by reducing the imports of substitute goods subject to tariff distortions. If the latter effect dominates, the net effect is a reduction in welfare.
Building on the work of Johnson (1967) and Kemp and Negishi (1970), Eaton and Panagariya (1979; 1982) derive conditions under the presence of tariffs on a subset of imports that can lead to an improvement in the terms of trade or growth in a small open economy to result in a loss of welfare. It is readily shown, however, that if import quotas restrict imports instead, improvement in the terms of trade or growth cannot lead to a decline in welfare. Just as in Corden and Falvey, when quotas are in place, their distortionary effect remains unchanged when the terms of trade improve or growth takes place. Therefore, the direct benefits from improved terms of trade or growth determine the final outcome. In the presence of tariffs, tariff distortion worsens if the improvement in the terms of trade or growth is accompanied by a contraction of imports of one or more tariff-ridden goods.
Finally, Bhagwati and Srinivasan (1982) alternatively consider the effect of directly unproductive profit-seeking (DUP) activities in the presence of tariffs and quotas. They show that in the former case the DUP activity may paradoxically raise welfare if it draws resources out of the import-competing good and therefore leads to an expansion of imports of the tariff-ridden good. This cannot happen in the latter case, however, since the imports of the quota-ridden good cannot rise beyond the fixed import quota.

See Also

- tariffs
- trade policy, political economy of

Bibliography

Bhagwati, J. 1965. On the equivalence of tariffs and quotas. In *Trade, Growth and the Balance of Payments: Essays in Honor of G. Haberler*, ed. R.E. Baldwin, J. Bhagwati, R.E. Caves and H.G. Johnson. Chicago: Rand McNally.

Bhagwati, J. and Srinivasan, T.N. 1982. The welfare consequences of directly unproductive profit seeking (DUP) lobbying activities: price vs. quantity distortions. *Journal of International Economics* 13, 15–33.

Corden, W.M. and Falvey, R.E. 1985. Quotas and the second best. *Economics Letters* 18, 67–70.

Eaton, J. and Panagariya, A. 1979. Gains from trade under variable returns to scale, commodity taxation, tariffs and factor market distortions. *Journal of International Economics* 9, 481–501.

Eaton, J. and Panagariya, A. 1982. Growth and welfare in a small, open economy. *Economica* 49, 409–19.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Helpman, E. and Krugman, P. 1992. *Trade Policy and Market Structure*. Cambridge, MA: MIT Press.

Johnson, H.G. 1967. The possibility of income losses from increased efficiency or factor accumulation in the presence of tariffs. *Economic Journal* 77, 151–4.

Kemp, M.C. and Negishi, T. 1970. Variable returns to scale, commodity taxes, factor market distortions and their implications for trade gains. *Swedish Journal of Economics* 72, 1–11.

Panagariya, A. 1980. Import targets and the equivalence of optimal tariff and quota structures. *Canadian Journal of Economics* 13, 711–15.

Panagariya, A. 1981. Quantitative restrictions in international trade under monopoly. *Journal of International Economics* 11, 15–31.

Panagariya, A. 1982. Tariff policy under monopoly in general equilibrium. *International Economic Review* 23, 143–56.

Pelcovits, M. 1976. Quotas versus tariffs. *Journal of International Economics* 6, 363–70.

Rodriguez, C.A. 1974. The non-equivalence of tariffs and quotas under retaliation. *Journal of International Economics* 4, 295–8.

Tower, E. 1975. The optimum quota and retaliation. *Review of Economic Studies* 42, 623–30.

How to cite this article

Panagariya, Arvind. "tariff versus quota." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_Q000013> doi:10.1057/9780230226203.1667

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tariffs

Tibor Scitovsky
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tariffs are taxes levied on goods imported or (less often) exported as they cross a geographical border. They raise revenue but are normally evaluated by reference to their impact on the economy, which usually means the protection they provide to domestic producers and their effect on the terms of trade. Tariffs can exploit a country's monopoly or monopsony position in world markets, but only if that is not already exploited by private firms within the country. An import duty can be used as countervailing power to prevent a country being exploited by a foreign exporter's use of *his* monopoly power.

Keywords

ad valorem tariffs; cartels; devaluation; effective tariffs; export tariffs; extortion tax; free trade; Great Depression; Hamilton, A.; import substitution; infant-industry protection; Lerner, A.; List, F.; monopoly; monopsony; protection; Scitovsky, T.; specific tariffs; tariffs; terms of trade; value added

Article

Tariffs are taxes levied on foreign trade: on the importation and, less often, the exportation of goods as they cross the border of a country or other geographical area. Since they are easy to enforce and collect and seem to be (and partly are) paid by foreigners, tariffs have been an important and popular source of government revenue from the earliest times. In early days, the ostensible purpose of tariffs was to pay the government levying them for the protection it afforded to foreign traders on its territory. In modern times, arguments for and against tariffs as well as the determination of their level focus on their impact or supposed impact on the economy.

Tariffs nowadays are paid in money and specified either as so much money per unit of merchandise (specific tariffs) or as a given percentage of its value (ad valorem tariffs). With demand a diminishing function of price, tariffs reduce the quantity of dutiable goods imported or exported; and, with the price elasticity of demand also a diminishing function of price, Government's revenue from the tariff (i.e. the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
product of its level and the quantity on which it is levied) first increases then diminishes as the tariff rate is raised. Accordingly, there is a rate of tariff that maximizes tariff revenue; but the proper criterion for judging the desirability of tariffs and determining their level is not the amount of revenue they yield but their impact on the whole economy, which is usually discussed under two headings: the protection they provide to domestic producers and their effect on the terms of trade.

Protective tariffs

Tariffs on imports raise their domestic prices, thereby shifting demand from imports to their domestic substitutes and increasing the profitability of the latter's production. Import duties also lower the purchasing power of income over imports and import substitutes (collectively known as importables) but add to the money incomes of producers of imports substitutes, their employees and suppliers. Accordingly, the tariff-imposing country's real national income may be raised or lowered, depending on whether the sum of the Government's tariff revenue and the additional incomes generated exceeds or falls short of the loss of purchasing power over importables. That, however, is still only a small part of a full cost-benefit calculation, which must also take into account other costs and benefits of the tariff.

By far the most important among the costs is the danger of retaliation by the foreign countries whose export industries are hurt, or believed to be hurt by the first country's import duties. That cost is especially great when the trade restrictions other countries impose in retaliation to the first country's tariff lead, in their turn, to further retaliations, and so to a general overall reduction in the volume of trade and its gains.

For the impact of import duties is to discourage the imports on which they are levied. It is true that they also stimulate domestic activity and domestic income generation, which, in the long run, may well counteract their restrictive effect on imports, at least to the extent of more or less offsetting the reduction in *overall* imports. But the combined influence of the restrictive short- and expansionary long-run effects of tariffs would have to keep unchanged not only the overall value of total imports but also their structure by country of origin in order to eliminate the economic justification and pressure for retaliation; whereas even the commodity composition of imports would have to remain unchanged in order to eliminate the political pressure for retaliation as well. Needless to say, those conditions necessary to obviate retaliation are not likely to be fulfilled.

The benefits of import duties include increased employment, an improved balance of trade, the enhanced stability of a more diversified economy, the political and economic advantages of greater self-sufficiency, and the increased efficiency of protected industries when their comparative disadvantages are remediable and can be remedied through learning by doing. Some of those advantages, however, are mutually exclusive. Tariffs, for example, that greatly stimulate the domestic economy are unlikely to improve the balance of trade – a fact that was strikingly brought home to many of the developing countries that engaged in import-substitution policies.

Of the benefits listed, by far the most important is the last-mentioned, which is a permanent benefit secured by temporary tariff protection. It has also received the most attention in the professional literature under the name of the infant industry argument. Trade restriction to nurture budding industries was well known and much practised already during the mercantilist period; but after the advent of economic liberalism, the argument in its favour needed to be reasserted. Its best known and most influential statements in modern times are those of Alexander Hamilton and Friedrich List. Hamilton's

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
celebrated 'Report on Manufactures' to the US Congress (1791) had a great influence on American tariff policy, and its prediction of the hoped-for consequences of protection turned out to be a remarkably accurate forecast of the country's subsequent economic development. List's similar argument half a century later (List, 1841) had even more influence on both US and German foreign-trade policy.

The US and German protective tariffs of the 19th century, however, which seem to have been so successful in promoting those countries' economic development, were very much more moderate than the mid-20th-century import barriers behind which India, Pakistan and the Latin-American and other developing countries pursued their not very successful import-substitution policies (Little et al., 1970).

That raises the question of what level and structure of protective tariffs are the most conducive to a country's economic development. We cannot answer here that much-debated and highly controversial question; but something must be said about effective tariffs, a statistical tool designed to help the search for an answer.

Effective tariffs

The height of a tariff levied on imports of a good (also known as the nominal tariff) is not a good measure of the degree of encouragement of its domestic manufacture. For one thing, a manufacturer almost never creates a whole good, only a greater or lesser contribution to it, which is called his value added or effective price; and a given percentage tariff on imports, which enables domestic manufacturers of its substitutes to raise their prices by a like amount, makes a greater, often very much greater *percentage* addition to their value added, in a proportion that is the inverse of the ratio in which their value added stands to price. For example, if the value added in cloth manufacture is 40 per cent of price, then a 20 per cent nominal tariff on imported cloth enables domestic cloth manufacturers to increase their value added by 50 per cent.

For another thing, tariffs are often levied on final, intermediate and primary goods alike; and an import duty on a primary or intermediate good, while encouraging its domestic production, also *dis*courages the domestic manufacture of all those other goods that use it as an input. An import duty on yarn, for example, discourages domestic cloth manufacture by reducing the value added cloth manufacturers can earn.

The concept of effective tariff (ET) is designed to measure the degree of encouragement provided to given productive activities by the combined effect of the nominal tariffs imposed on their outputs *and* inputs. A simple formula for the effective tariff protection on the manufacture of good *j* is:

$$ET_j = \frac{t_j}{1 - \Sigma a_{ij}} - \frac{\Sigma a_{ij} t_i}{1 - \Sigma a_{ij}},$$

where t_j is the nominal ad valorem tariff on good *j*, t_i are the nominal tariffs on its several inputs, and the a_{ij} show the share of the cost of input *i* in the price of good *j* at free trade prices. Note that the two terms show the contributions of the two factors discussed in the text, note also that the denominator represents

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
value added as a proportion of price.

The terms-of-trade effect of tariffs

In contrast to all the attention economists, politicians and the general public have paid to the protection that tariffs provide to domestic industry, the tendency of tariffs to improve the terms on which a country trades its exports for imports, and thereby to increase its share in the gains from international specialization, have been very much neglected. The subject attracted some attention at the end of the last and the beginning of this century, but mainly as a theoretician's intellectual exercise and an economic curiosum.

While protection results from import duties' raising the *domestic* price that *domestic* buyers have to pay for importables, they improve the terms on which the duty-imposing country trades its exports for imports, provided that they lower the *foreign* price that the *foreign* producers of it receive for them. Similarly, an export duty will also improve a country's terms of trade if it raises the *foreign* price that *foreign* buyers of its exports have to pay for them. Accordingly, tariffs improve a country's terms of trade if the foreign supply of its imports or the foreign demand for its exports is less than perfectly elastic; and a given tariff has the greater impact on the terms of trade, the lower are those elasticities (Bickerdike, 1906; Kaldor, 1940).

The advantage of a tariff that improves the terms of trade can be given two interpretations. First, when a tariff changes the foreign price of imports and/or exports to the foreigners' disadvantage, it causes *them* to pay part of the tariff – a clear and obvious gain for the tariff-imposing country. Secondly, the same gain can also be looked upon as a monopoly or monopsony profit extracted from foreigners by the tariff, which in turn closely resembles the profit margin a monopolist adds on to marginal cost, or a monopsonist subtracts from marginal worth, when he sets his profit-maximizing price. Indeed, when perfect competition among a country's export producers causes them to equate prices to marginal costs and causes its importers to equate the marginal value product of imports to their prices, then export and import duties coincide exactly with a monopolist's and monopsonist's profit margins.

Such a situation resembles a cartel agreement among domestic competitors with respect to their foreign transactions, except that the monopoly or monopsony profits generated accrue to the State in the form of tariff revenue and that the private producers and traders are made worse off than they would be under free trade, because the tariff reduces the volume of their business. From the point of view of the country's national welfare, however, tariffs can be beneficial, in the sense of increasing the sum of the country's private and public real income, just as monopolistic or monopsonistic pricing can increase the monopolist's or monopsonist's profit. Indeed, there are optimum tariffs, which maximize a country's gain from trade, and whose level depends on the price elasticities of the foreign supply of imports and the foreigners' demand for exports, just as the monopolist's profit maximizing profit margin depends on the price elasticity of demand he faces (Scitovsky, 1942).

Tariffs, like monopoly pricing, redistribute income in favour of those imposing them in a way that inflicts a greater loss on those hurt than the gain they secure for those favoured. For that reason, it is important to prevent competitive tariff impositions and increases, whereby each country retaliates in self-defence to the tariffs imposed by others, and so contributes to a general impoverishment of all or almost all, due to the all-round reduction of international specialization and of the gain it generates. Yet, that happened during the 1930s depression; and it can easily happen when each country believes itself to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
have a small enough share in world trade to erect or raise tariffs unpunished and retaliation is effective or believed to be effective in recapturing some of the lost gain of the retaliating country. Free trade therefore is not a stable situation, unless imposed by a dominant country, such as Great Britain in the 19th century or the United States during the period following World War II (Scitovsky, 1942; Kahn, 1947). What happens when free trade is not enforced, which countries gain, which lose from tariffs and retaliatory tariffs, and what is the nature of the path and final outcome of competitive trade restrictions has received considerable attention in the theoretical literature (Kaldor, 1940; Scitovsky, 1942; Graaff, 1949–50; Johnson, 1953–4; Gorman, 1957–8), but is too complex to summarize here.

Also, the subject has remained a theoretical exercise and faded into the background. Yet, it has two aspects that, though largely overlooked in the literature, deserve mention here. One is that tariffs can exploit a country's monopoly or monopsony position in world markets *only* if that is not already exploited by private firms within the country. The other is that an import duty can be used as countervailing power to prevent a country's being exploited by a foreign exporter's use of *his* monopoly power.

A country's only large producer of an exportable product or its single importer of a foreign product enjoys, of course, the same monopoly or monopsony position in world markets as does the country as a whole. Accordingly, he can, and usually does, exploit that position to his own – as well as to his country's – advantage by setting the profit-maximizing monopoly or monopsony price. The same is approximately true also if, instead of a single monopolist, a few large firms act in open or tacit oligopolistic collusion in setting monopoly prices. When they do that, tariffs for the purpose of exploiting the country's bargaining position in world markets are not only redundant but harmful, because, added to a producer's monopoly (or subtracted from an importer's monopsony) price, they are liable to push the foreign price beyond its profit-maximizing level, thereby inflicting a loss on domestic exporters or importers that exceeds the government's tariff revenue. In short, tariffs and monopolistic profit margins can substitute one for another, complement each other, but cannot be used to exploit the same monopoly or monopsony position twice over.

That explains, for example, why export tariffs and other export restrictions have been imposed almost exclusively on primary products and only in countries where those are grown by many small growers under competitive conditions. Export duties on coffee and the Ghanaian State monopoly for the export of cocoa are the obvious examples. The industrial countries, which export manufactures, have no need for export duties to exploit their monopoly position in world markets, because the large manufacturers of their exportables are usually able to charge monopoly prices on their own, thus making export duties redundant.

The same argument also explains why Britain practised and preached free trade up to the end of the 19th century. Her heavy manufactures were produced and exported by large, monopolistic firms, her light manufactures (textiles), though produced competitively, were exported by large wholesale merchants, and some of her primary-product imports were also handled by large British firms, most of them able to set prices that exploited their foreign and domestic monopoly positions alike and rendered tariffs superfluous.

We come now to the use of an import duty to offset a foreign exporter's monopoly and diminish or eliminate his monopoly profits. Ross Shepherd has shown (Shepherd, 1978) that a variable import duty which varies, and is *expected to vary*, directly with the foreign price of an imported good, raises the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
country's apparent price elasticity of demand for that import and correspondingly reduces its manufacturer's monopoly power and with it his profit maximizing price. Indeed, under constant cost conditions, a suitable duty will leave unchanged both the volume imported and the domestic price paid for it by domestic consumers, while expropriating the foreign exporter's monopoly profit. Abba Lerner, who seems to have arrived at the same conclusion independently, advocated imposing such a variable duty (which he called ‘extortion tax’) on oil imports, thereby creating an incentive for OPEC's members to break ranks by reducing price (Lerner, 1980).

In closing, it is worth noting some similarities and differences between the imposition of tariffs and devaluation. A uniform ad valorem duty on all imports combined with a uniform ad valorem subsidy (negative duty) of the same magnitude on all exports is identical to a devaluation of that magnitude in its effects on the balance of trade but leaves unchanged all other international transactions and financial obligations. For that reason, countries anxious not to increase the burden on domestic debtors of foreign debt denominated in foreign currencies have used such and similar policies as means of improving their balance of trade in preference to devaluation. Also, since devaluation worsens a country's terms of trade when the foreign demand for some of its exports is very inelastic, it may be combined with a duty or other restraint on those of its exports (usually primary products), thereby to prevent the deterioration of its terms of trade, or import restriction may be substituted for devaluation.

See Also

- optimal tariffs

Bibliography

Bickerdike, C.F. 1906. The theory of incipient taxes. *Economic Journal* 16, 529–35.

Gorman, W.M. 1958. Tariffs, retaliation, and the elasticity of demand for imports. *Review of Economic Studies* 25, 133–62.

Graaff, J. de V. 1949. On optimum tariff structures. *Review of Economic Studies* 17, 47–59.

Johnson, H.G. 1954. Optimum tariffs and retaliation. *Review of Economic Studies* 21, 142–53.

Kahn, R.F. 1947. Tariffs and the terms of trade. *Review of Economic Studies* 15, February, 14–19.

Kaldor, N. 1940. A note on tariffs and the terms of trade. *Economica* N.S. 7, 377–80.

Lerner, A.P. 1980. OPEC – a plan – if you can't beat them, join them. *Atlantic Economic Journal*, September, 1–3.

List, F. 1841. *The National System of Political Economy*. Trans. S.S. Lloyd, London: Longmans, Green

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

& Co., 1885.

Little, I.M.D., Scitovsky, T. and Scott, M.K. 1970. *Industry and Trade in Some Developing Countries: A Comparative Study*. London: Oxford University Press.

Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9, Summer, 89–110.

Shepherd, A.R. 1978. *International Economics: A Micro-Macro Approach*. Columbus: Charles E. Merrill.

How to cite this article

Scitovsky, Tibor. "tariffs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000012> doi:10.1057/9780230226203.1668

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tarshis, Lorie (1911 – 1993)

D.E. Moggridge
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate supply function; Keynesianism; Tarshis, L.; underinvestment

Article

Tarshis was born in Toronto, Canada, on 22 March 1911. After a commerce degree at the University of Toronto, he went to Trinity College, Cambridge, where he took a BA in 1934 and a Ph.D. in 1939. His years in Cambridge, 1932–6, which coincided with the emergence of Keynes's *General Theory*, shaped much of his subsequent professional life. His notes for Keynes's annual series of eight lectures on his work in progress for the years 1932–5 have become an important source for those interested in tracing the evolution of Keynes's views. The two Cambridge revolutions of the 1930s, Keynes's and imperfect competition, focused the analysis of his Ph.D. dissertation, ‘The Distribution of Labour Income’. From this came two classic articles in 1938 and 1939 which, along with a contemporaneous piece by John Dunlop (1938), forced Keynes to reconsider his generalization that real and money wages moved inversely over the trade cycle and its implications for the assumption of perfect competition that underlay the analysis of the book (Keynes, 1939).

By then Tarshis had moved to the United States, first to Tufts University (1936–9, 1942–6) and subsequently to Stanford (1946–71). While at Tufts, along with his Cambridge classmate R.B. Bryce, he played a significant role in spreading Keynes's ideas among the Harvard community of economists. Then in 1938 he participated with several other economists in the manifesto *An Economic Program for American Democracy*. Only seven of them eventually signed it – R.V. Gilbert, G.H. Hildebrand Jr., A. W. Stuart, M.Y. and P.M. Sweezy, Tarshis and J.D. Wilson – the government or other connections of the rest preventing them from doing so. The *Program* was ‘Keynesian in analysis, stagnationist in diagnosis and all-out in prescription’, and was ‘instrumental’ in driving home to New Deal Washington the need for more spending to overcome the fatal flaw of contemporary capitalism, underinvestment (Stein, 1969, pp. 165–7). His move to Stanford coincided with another effort at Keynesian persuasion, *The Elements of Economics*, the first unashamedly Keynesian introductory textbook. Dogged by controversy over its supposed ‘left wing’ views, it was much less successful than the slightly later

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

competing text of Paul Samuelson.

During the subsequent 40 years, despite his heavy teaching commitments where he probably left his greatest mark, Tarshis continued to publish regularly. His contributions related to international finance, the microeconomics of Keynes (most notably the aggregate supply function) and contemporary policy issues.

Selected works

1938a. Real wages in the United States and Great Britain. *Canadian Journal of Economics and Political Science* 4, 362–76.

1938b. (With R.V. Gilbert et al.) *An Economic Program for American Democracy*. New York: Vanguard Press.

1939. Changes in real and money wages. *Economic Journal* 49, 150–4.

1947. *The Elements of Economics*. Boston: Houghton Mifflin.

1979. The aggregate supply function in Keynes's *General Theory*. In *Economics and Human Welfare: Essays in Honour of Tibor Scitovsky*, ed. M.J. Boskin. New York: Academic Press.

1984. *World Economy in Crisis*. Toronto: Lorimer.

Bibliography

Dunlop, J.T. 1938. The movement of real and money wage rates. *Economic Journal* 48, 413–34.

Keynes, J.M. 1939. Relative movements of real wages and output. *Economic Journal* 49, 34–51.

Stein, H. 1969. *The Fiscal Revolution in America*. Chicago: University of Chicago Press.

How to cite this article

Moggridge, D. E. "Tarshis, Lorie (1911–1993)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000013> doi:10.1057/9780230226203.1669

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Eco

taste-based discrimination

Kerwin Kofi Charles and Jonathan Guryan
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists typically account for differences in economic outcomes between ethnic groups with explanations having to do with differences in skill, explanations emphasizing informational problems associated with accurately assessing skill, as in statistical discrimination models, or explanations that rely on the presence of prejudice, the key element of taste-based discrimination models. This article defines taste-based discrimination and briefly outlines the economics of associated models. It discusses empirical implications of these models, and reviews empirical tests from the literature. It speculates about possible avenues for future research likely to enrich the insights forthcoming from the standard taste-based model. Although this article focuses on discrimination arising from racial prejudice, taste-based discrimination subsumes negative preferences towards groups of individuals of many alternative types, including different age, gender or religious groups.

Keywords

discrimination; prejudice; race; segregation; taste-based model; wage levels

Article

1 Becker's taste-based discrimination models

Individual preferences

In his seminal work, Becker (1971; 1st edn 1957) formally demonstrated how negative racial feeling, or prejudice, on the part of individual members of a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

majority group (here described as whites) could be related in a market environment to negative outcomes for members of a discriminated-against group (here described as blacks). He chose a representation of racial prejudice which, in addition to being precise and intuitively appealing, lent itself readily to tractable representation in an economic model. Prejudice in Becker's framework is represented as an aversion to cross-racial (or more generally cross-group) contact. Since this aversion renders cross-racial interactions psychically costly, the strength of an agent's aversion can also be thought of as the price the person would be willing to pay to avoid the interaction. Becker studies the effect of prejudice among three distinct types of white agents – employers, employees and customers.

Given the representation of racial prejudice, it follows straightforwardly that, in their market interactions that involve blacks, prejudiced agents act *as if* the relevant price mediating that interaction is the actual market price plus an amount determined by the agent's level of prejudice. Thus prejudiced employers with a taste for discrimination, d_i, deciding about hiring black workers view themselves as paying not the market wage w they pay to white workers but rather the price $w+d_i$. (This functional form assumes the disutility of interaction is linear in the number of black employees. Other functional forms are of course possible.) When the prejudiced person is an employee, holding a taste for discrimination Δ_i and contemplating a wage offer of $\hat{w}$ to work at a firm alongside black workers, he views himself to be working for a wage of $\hat{w} - \Delta_i$, rather than the wage $\hat{w}$ he would consider himself to be receiving were all his co-workers white. Finally, in the third example studied by Becker, a prejudiced customer with a taste for discrimination κ_i views himself as paying $p+\kappa_i$ per unit for goods sold by black sellers, rather than the market price p he would consider himself to be paying were the sellers white. It should be clear that the parameters d_i, Δ_i and κ_i each reflect the disutility that a prejudiced agent receives from interacting with blacks.

Market implications

What does individual prejudice as represented above imply about equilibrium wages and prices? Keeping closely to Becker's original presentation, we briefly describe how tastes interact in a market setting, where there is optimizing behaviour and competition, to determine the level of market discrimination for each of the three models.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Employer discrimination

If black (b) and white (a) workers are equally productive and perfect substitutes in production, an employer i will have utility given by

$$U_i = f(K, L_a + L_b) - w_a L_a + w_b L_b - d_i L_b$$

where $f(.)$ is a constant returns to scale production function, w_j is the market wage paid to workers from group $j \in (a, b)$, K is capital, and L_j is the number of workers hired from group j. Taste for discrimination, $d_i \geq 0$, among all employers varies according to some arbitrary distribution Ω.

Since in this model black and white workers are perfect substitutes in production, each employer simply hires the type of worker who is less costly, at the margin, *to him*. Thus an employer hires only black workers if $w_b + d_i < w_a$, and he hires only white workers if the strict inequality is reversed. Notice that an employer's workforce is strictly segregated by race, unless his racial prejudice d_i is such that $w_b + d_i = w_a$.

What is the equilibrium in the short run, when the number and size of firms are fixed? Imagine a central planner choosing black and white wages and allocating black and white workers to employers so that the markets for both types of workers clear. The planner allocates black workers to the least prejudiced employers: that is, to those with $d_i = 0$ first and then, if necessary, to those with the lowest values of d_i. If the distribution of d_i is smooth, the last employer to be allocated a black worker must be indifferent between hiring black and white workers. In equilibrium, less prejudiced employers hire blacks, more prejudiced employers hire whites, and the equilibrium black–white wage gap $w_a^* - w_b^*$ is equal to the prejudice of the employer who is indifferent between hiring blacks and whites at the equilibrium wage, or d_i^*.

The model sharply distinguishes individual prejudice from market discrimination. In particular, the equilibrium black–white wage gap is not determined by the average level of prejudice among employers, but by the prejudice of a marginal discriminator. Even in a market in which some employers are prejudiced, there need be no racial wage gap in equilibrium provided there are non-prejudiced employers to hire all the black workers.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Notice also that since prejudiced employers and black workers have an incentive to avoid interacting, there is market pressure towards segregation, so that segregation and observed discrimination are effectively substitutes. The more segregation there is in equilibrium, the smaller is the wage gap.

In general, the equilibrium black–white wage gap increases as the prejudice of the marginal discriminator increases – either because the fraction of the workforce that is black increases or because of changes in the distribution of prejudice among all employers. In the first case, the presence of more blacks in the market means that market clearance requires that blacks be allocated to ever more prejudiced employers at the margin. In the second case, higher levels of prejudice in *the part of the distribution from which the marginal employer is likely to be drawn* will increase market discrimination. Since blacks represent a small minority in most markets, the sorting that characterizes the equilibrium guarantees that only higher prejudice among those employers in the left tail of the distribution (below the median) should lower black wages; higher prejudice among the most prejudiced employers in the market should have no effect on the equilibrium wage gap since the marginal employer is very unlikely to be drawn from among these persons.

Employee discrimination

As discussed above, a prejudiced employee behaves as if the wage he is offered by a firm with black workers were the actual offered wage $\hat{w}$ minus the disutility he gets from interacting with blacks at work, Δ_i. As in the employer discrimination model, market forces generate a tendency toward segregation. In the employee discrimination case, because of the preferences of their workers, employers have an incentive to segregate their workforces. An employer who hires both black and prejudiced white workers is forced to pay a premium to the whites to induce them to work for him. He does not pay that premium if his workforce is all the same race. Each firm therefore prefers to employ either only white workers or some combination of black and unprejudiced white workers.

If, in equilibrium, firms are able to segregate perfectly by race, there will be no equilibrium wage gap. Only if there are impediments to perfect segregation, large enough to ensure that blacks work with prejudiced co-workers, can employee prejudice lead to wage discrimination. In the likely event these frictions are such that more prejudiced employees are especially unlikely to work with black co-workers, reductions in segregation lead to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

increases in the racial wage gap.

Customer discrimination

In the third type of taste-based model discussed by Becker, some customers care about the race of workers producing the goods they purchase. They consequently regard the price they pay for goods made by a firm with l_b black workers to be the charged price, P_b, plus their disutility which increases with the number of black inputs into production, or $\kappa_i l_b$. Since all employers are unprejudiced, and since the profits of firms hiring black and white workers must be the same in equilibrium, any per unit price difference in equilibrium must be reflected in a difference in wages paid to black and white workers. (P_a is the price of a good produced exclusively by white workers.) If there is a price difference in the good, the most prejudiced customers will buy goods produced by whites, and the least prejudiced customers will buy goods produced by blacks. The marginal discriminator is that consumer who is indifferent between buying goods made by blacks and those made by whites, given his level of prejudice and the equilibrium prices charged. If there are enough unprejudiced customers relative to the number of blacks in the market, there will be no difference in prices between the two types of goods, and no difference in equilibrium wages by race. If this condition does not hold, then there will be a racial wage gap.

2 Long-run implications

Traditional view

Much early discussion and criticism about taste-based models centred on the nature of the long run, when firms can freely enter and exit from the market. The employer version of the taste-discrimination model has historically been the focus of this criticism, as it is in this particular prejudice model that the long-run implications appear, at first blush, to be most troubling for the standard model.

To see the essence of the criticism of the employer prejudice model, note that if there is an equilibrium wage gap in the short run, employers that hire only white workers have higher labour costs than do firms that hire only black workers. Since workers are equally productive by assumption, non-discriminating firms earn greater profits. The return to capital is thus different

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

across firms, and in the long run, capital should flow to those firms with the highest return – those with the least prejudiced employers. With a constant returns to scale production function, this mechanism continues until all employers with $d_i>0$ shut down and leave the market. In the long-run equilibrium, no prejudiced employer survives and any wage gap caused by prejudice is eliminated. Becker himself originally outlined this argument in his original work, and it was famously and forcefully repeated later by Arrow (1972). The influence of Arrow's criticism was such that it probably discouraged work on taste-based prejudice models, relatively few of which have appeared subsequently in the theoretical literature. At the same time it probably encouraged work on statistical discrimination, which is today the dominant paradigm for the theoretical study of discrimination. Indeed, the earliest versions of statistical discrimination model were presented by Arrow (1972) in the same paper in which his criticism was lodged, followed quickly by Phelps's (1972) seminal work.

Recent theoretical work

In recent years a number of authors have presented employer prejudice models of discrimination that modify Becker's original assumptions about competition or the wage-setting process, or introduce job search frictions to resurrect the prediction of long-run racial wage gaps resulting directly from taste-based discrimination.

In an important paper, Black (1995) relaxes the perfect competition assumption, introducing costly job search into a model in which some employers refuse to hire black workers and others are non-discriminatory. Without perfect information, black workers know that some fraction of the employers they randomly encounter while searching would refuse to hire them. Non-discriminatory employers consequently enjoy local monopsony power over their black workers, in the sense that reductions in their wage will not lead all such workers to leave. In equilibrium, the presence of employers that refuse to hire blacks causes blacks to receive, from those employers that *do* hire them, lower wages than otherwise identical whites.

More recently Lang, Manove and Dickens (2005) described a model in which employers post binding wage offers. Because firms cannot base wage offers on the race of the worker, discriminatory employers exercise their preferences by choosing to hire based on race. Since job search is costly, black workers choose to apply to firms at which they believe they have a good chance of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

being hired. These firms are those that post low wage offers. In equilibrium, there is both segregation and a wage gap.

Our own recent work (Charles and Guryan, 2007, 2008) calls into question the basic logic of Arrow's critique of the Becker model. The conclusion that discriminatory employers are driven to shut down as a result of competition with non-discriminatory employers is based on an assumption about employers' alternative labor market opportunities, which may be as a worker at a firm. By its reliance on zero profit as the condition for shutting down, this conclusion is implicitly based on the assumption that an employer's outside option is valued only according to the monetary wage paid at that firm, and not according to the race of his potential co-workers. If instead it is assumed that a prejudiced employer takes his distaste for racial contact along with him to his role as a co-worker, then the conclusion is different. Whether a prejudiced employer shuts down then depends on his likelihood of finding a job that does not involve contact with black co-workers. It follows that the ability of the market to segregate is the key mediator between individual discriminatory tastes and market discrimination. In a model with racial preferences that are portable across roles (employer or worker) and in which there is an impediment to segregation, racial wage gaps can persist in the long run even in the face of perfect competition and free entry and exit.

Interestingly, we might think of the model described in Charles and Guryan (2007, 2008) as one in which the choice of whether to be an employer or an employee is endogenized. In this case, Becker's worker discrimination model is essentially a generalization of his employer discrimination model.

The predictions of this model are essentially the same as those from Becker's short-run employer discrimination equilibrium.

3 Empirical assessment of taste-based prejudice models

There is a vast empirical literature devoted to studying racial gaps in economic outcomes, but very little of that work can be said to test directly the implications of taste-based models. Part of the reason is that it is difficult to establish that observed racial wage differences are truly the result of any form of discrimination and not of unmeasured skill or ability differences across the groups. The suspicion that unmeasured skill differences account in part for observed wage and other differences by race is strengthened by results, such as those from Neal and Johnson (1996), showing that adding rarely used controls like test scores to wage regressions results in a significant reduction

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in the amount of the wage gap that is unexplained. Even when it can be reasonably argued that the unexplained gaps are unlikely to be the result of some measure of skill for which the researcher has failed to control, there remains the problem that such results are consistent with forms of discrimination that have nothing at all to do with taste-based prejudice. This is true even of the important and carefully done audit studies in various markets, in which blacks and whites, or their resumes, are sent at random to different employers or firms in analysts, attempts to identify differential treatment. (See for example Heckman, 1998; Bertrand and Mullainathan, 2004.)
Charles and Guryan (2008) test the predictions of Becker's model empirically, collecting data on explicit measures of prejudice from the General Social Survey (GSS) to construct indices of the distribution of individual discriminatory tastes. They then compare various measures of prejudice at the labour market (in this US example, state) level, and compare these with the measured black–white wage gap in the market. The results are remarkably supportive of Becker's model. Black–white wage gaps are larger in states with a higher fraction of a black workforce. Racial wage gaps are larger in more prejudiced states, but the relationship holds in a particular way. Black relative wages are negatively related to the degree of prejudice in the left tail of the prejudice distribution, but not to variation in the median or right tail. This finding is consistent with the sorting mechanism that Becker describes, and that tends to make the marginal discriminator someone less prejudiced than the average person. Further supporting the view that the market's ability to segregate is a key mediator between individual tastes and market outcomes, the study found that black–white wage gaps are larger in states where there is more integration of – more contact between – blacks and whites in the workplace.

4 Areas for future work and conclusion

Despite the historical importance of taste-based models in the economics literature on discrimination, we believe there are many areas for future theoretical and empirical work. We briefly discuss only a few of these, listing them in the form of questions.

What precisely is taste-based prejudice?

Becker represents racial prejudice as distaste for interaction, and virtually all subsequent prejudice models have followed Becker's lead. However, racial

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

prejudice might manifest itself in various other ways, with possible important implications for market outcomes. Prejudice might, for example, affect information processing. In particular, racial prejudice might cause people infected by it not to update negative stereotypes about members of a group, even in the face of contradictory evidence. This representation of prejudice is closely related to work by authors like Loury (2002) and Coate and Loury (1993) about the formation of stereotypes, and to Myrdal's (1944) work on ‘vicious cycles’. Much work remains to be done with this representation of prejudice. Nor has much work been done with prejudice represented as a preference for one's own group (nepotism) rather than an aversion to interactions with another group. (See Goldberg, 1982 for an important exception.)

A massive literature in social psychology examines the formation of prejudice and stereotypes, including the question of whether prejudice is a preference towards members of an in-group or dislike of members of an out-group. (It would be impossible to summarize that entire literature in this article, but for a good overview, the interested reader should see Fiske, 1998. An important early study is Allport, 1954.)

The interesting experimental work of Tajfel et al. (1971) suggests that agents care about the utility of their ‘own’ group, but may care more about actions that maximize the *difference* between their group's outcomes and those of the out-group. Exploring the implications of these and other insights in formal economic models would vastly enrich our understanding of the effect of prejudice.

From where does racial prejudice come?

Are group-based preferences instinctual or learned? What are the root causes of prejudice? Is prejudice the result of ignorance or unfamiliarity? The answers to questions like these are important for designing policies aimed at reducing discrimination, or even the prevalence of the tastes themselves. Research in psychology and sociology, such as that by Tajfel et al. (1971) and the famous ‘Robbers’ Cave' study by Sherif et al. (1961/1988), suggests that the entire notion of in- and out-groups, with associated negative and positive feelings, might even arise among people randomly assigned to groups. At the same time, some of this research suggests that in- and out-group sentiments are especially strong when there is competition over scarce resources. Future work by economists might incorporate these insights to how prejudice could arise endogenously from, say, residential segregation. Or, future work might

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

assess whether, following sectoral reallocation or local demand shocks, there is a change in the extent to which prejudice varies with competition over scarce jobs.

Is prejudice conscious?

Most analyses of prejudice in economics assume that the prejudiced agent is aware of his prejudicial sentiments. But research suggests that one way that people might conserve their limited cognitive resources is to group objects into categories and then to generate summary beliefs for those groups rather than a separate one for each individual (see e.g. Allport, 1954; Tajfel, 1981; Taylor, 1981). Humans may therefore develop prejudicial beliefs of which they are not overtly conscious. An example of a test of such subconscious prejudice is the Implicit Attitudes Test (Banaji and Greenwald,1995; Greenwald, McGhee and Schwartz, 1998). Results in the lab (Wittenbrink, Judd and Park, 1997) and in the field (Price and Wolfers, 2007) support the possibility that automatic psychological responses contribute to discriminatory actions. Further investigation of whether prejudice is subconscious and the associated empirical consequences of that being the case would be an obvious useful area for future work.

In addition to answering questions like these, there is clearly a need to subject the various prejudice models to additional empirical tests. Charles and Guryan (2008) is the only paper of which we are aware that tests the basic predictions of the employer discrimination model about the relationship between equilibrium wage gaps and the distribution of prejudice. Clearly there should be more of this type of work, across different areas and over different time periods. To that end, there might be great benefit to economists from collecting data on prejudice themselves rather than relying on data collected by others. For example, no prejudice measure of which we are aware elicits from respondents an answer to the question how much they would be willing to pay to avoid interacting with members of a particular group. Economists collecting such data would naturally cast their prejudice questions in this form, allowing for an almost exact translation between the theoretical construct of interest to economists and the measure used to study it in the data. Similarly, given the central theoretical importance of the nature and frequency of cross-group interaction within firms to observed racial wage gaps in taste-based models, there is likely to be a great deal to be learned from empirical analyses that jointly study wage discrimination and segregation. Here too there would seem to be substantial returns from the collection of new data, as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the available data on actual interactions within firms is coarse and available in very few data sources.

Taste-based models were the first models of discrimination written down by economists. While these models almost certainly cannot account for all of the large and durable differences observed in the market across racial, gender and other dimensions, they nonetheless yield valuable insights about why putatively unproductive traits like race or gender might be correlated systematically with worse market outcomes. Although most work on discrimination has focused on explanations like statistical discrimination or systematic differences in unmeasured skill, there is evidence of a resurgence in interest in taste-based models, given the appearance of a number of papers in the recent literature. In our view, renewed interest in, and investigation of, taste-based models will inevitably lead to a richer understanding of the nature and reasons for differences in economic outcomes in the population.

See Also

- anti-discrimination law
- labour market discrimination

Bibliography

Allport, G.W. 1954. *The Nature of Prejudice*. Reading, Mass.: Addison-Wesley.

Arrow, K. 1972. Some mathematical models of race in the labor market. In *Racial Discrimination in Economic Life*, ed. A.H. Pascal. Lexington, Mass.: Lexington Books, pp. 187–204.

Banaji, M.R. and Greenwald, A.G. 1995. Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology* 68, 181–98.

Becker, G.S. 1971. *The Economics of Discrimination*, 2nd edn. Chicago, Ill.: University of Chicago Press.

Bertrand, M. and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94(4), 991–1013.

Black, D.A. 1995. Discrimination in an equilibrium search model. *Journal of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Labor Economics 13(2), 309–34.

Charles, K.K. and Guryan, J. 2007. Prejudice and the economics of discrimination. NBER Working Paper no. 13661.

Charles, K.K. and Guryan, J. 2008. Prejudice and wages: an empirical assessment of Becker's. *The Economics of Discrimination. Journal of Political Economy* 116(5), 773–809.

Coate, S. and Loury, G.C. 1993. Will affirmative-action policies eliminate negative stereotypes? *American Economic Review* 83(5), 1220–40.

Fiske, S.T. 1998. Stereotyping, prejudice and discrimination. In *The Handbook of Social Psychology*, ed. D.T. Gilbert, S.T. Fiske and G. Lindzey. New York: Oxford University Press.

Goldberg, M.S. 1982. Discrimination, nepotism, and the long-run wage differential. *Quarterly Journal of Economics* 97(2), 307–19.

Greenwald, A.G., McGhee, D.E. and Schwartz, J.K.L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* 74, 1464–80.

Heckman, J.J. 1998. Detecting Discrimination. *Journal of Economic Perspectives* 12(2), 101–16.

Lang, K., Manove, M. and Dickens, W. 2005. Racial discrimination in labor markets with posted wage offers. *American Economic Review* 95(4), 1327–40.

Loury, G.C. 2002. *The Anatomy of Racial Inequality*. Cambridge, Mass.: Harvard University Press.

Myrdal, G. 1944. *An American Dilemma: The Negro Problem and American Democracy*. New York: Harper & Brothers.

Neal, D.A. and Johnson, W.R. 1996. The role of premarket factors in black–white wage differences. *Journal of Political Economy* 104(5), 869–95.

Phelps, E.S. 1972. The statistical theory of racism and sexism. *American Economic Review* 62, 659–61.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Price, J. and Wolfers, J. 2007. Racial discrimination among NBA referees. NBER Working Paper no. 13206.

Sherif, M., Harvey, O.J., White, B.J., Hood, W.R. and Sherif, C.W. 1961/1988. *The Robbers Cave Experiment: Intergroup Conflict and Cooperation*. Middletown, Conn.: Wesleyan University Press.

Tajfel, H. 1981. *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge, Mass.: Cambridge University Press.

Tajfel, H., Flament, C., Billig, M. and Bundy, R. 1971. Social categorization and intergroup behavior. *European Journal of Social Psychology* 1(2), 149–78.

Taylor, S.E. 1981. A categorization approach to stereotyping. In *Cognitive Processes in Stereotyping and Intergroup Behavior*, ed. D.L. Hamilton. Hillsdale, NJ: Elrbaum, pp. 83–114.

Wittenbrink, W., Judd, C.M. and Park, B. 1997. Evidence for racial prejudice and the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology* 72, 262–74.

How to cite this article

Charles, Kerwin Kofi and Jonathan Guryan. "taste-based discrimination." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_T000251> doi:10.1057/9780230226203.1906

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tâtonnement and recontracting

Takashi Negishi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

auctioneer; barter; Edgeworth, F. Y.; exchange; general equilibrium; gross substitutability; Hicks, J. R.; kinked demand curve; law of indifference; limit theorem; Marshall, A.; money; neutrality of money; non-recontracting models; non-tâtonnement models; numéraire; recontracting; sticky prices; tâtonnement; tâtonnement and recontracting; Walras, L.; Walras's Law

Article

In the current theory of general economic equilibrium, recontracting and tâtonnement (a French word meaning ‘groping’) are used interchangeably to denote a simplifying assumption that no actual transactions, and therefore no production and consumption activities, take place at disequilibria when prices are changed according to the law of supply and demand (Kaldor, 1934; Arrow and Hahn, 1971, pp. 264, 282). Historically speaking, however, this usage is somewhat confusing, since recontracting is originally due to Edgeworth, who developed it in a direction different from that in which Walras developed his tâtonnement (Walker, 1973).

Though different interpretations are given as to whether Walras explicitly excluded disequilibrium transactions from the beginning (Patinkin, 1956, p. 533; Newman, 1965, p. 102; Jaffé, 1967; 1981), it is clear that Walras developed his theory of tâtonnement so as to exclude such transactions. To do this there are at least three methods of tâtonnement. First, we may assume that price-taking traders facing market prices cried by the *auctioneer* reveal their plans of demand and supply to the auctioneer but do not make any trade contract among themselves until the auctioneer declares that equilibrium is established. Alternatively, traders may be assumed to make trade contracts (Walras, 1926, p. 242, suggested the use of tickets when production is involved) but recontract is assumed always to be possible, in the sense that contract can be cancelled without consent of the other party if market prices are changed. Finally, the effect of past contracts can be nullified by offering new demands (supplies) to offset past supplies (demands), even if it is assumed that past contracts are effective and would be carried out at the current prices when the equilibrium is established (Morishima, 1977, pp. 28–30). Since

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
any changes in prices make the contract unfavourable to one of the parties which then wishes to cancel the trade contract, there is no difference between the three methods of tâtonnement in the behaviour of demand, supply and prices. Recontracting in this sense of tâtonnement is, however, quite different from that developed in Edgeworth's theory of recontract.

We shall start by the consideration of why this assumption of tâtonnement is necessary for the Walrasian theory of general equilibrium, which is the foundation of neoclassical economic theory. The reason lies in the structure of Walrasian economics, dichotomized between real and monetary theories. Then we analyse formal models of tâtonnement including the original one due to Walras and the modified version developed in modern theories of general equilibrium. It is followed by our assessment of the theoretical achievements and empirical relevance of Walrasian tâtonnement economics. Edgeworth's theory of recontract is reviewed in its relation to the Walrasian theory of tâtonnement. Finally, an evaluation is made on the recent studies of tâtonnement and recontracting, to show in which direction further progress should be made.

1.•Walras (1874–7) insisted that complicated phenomena can be studied only if the rule of proceedings from the simple to the complex is always observed. To understand the fundamental nature of Walrasian economics, it is convenient to make (as did Hicks, 1934) a comparison of Walrasian and Marshallian ways of applying this rule to the study of complicated economic phenomena. Both Walras and Marshall (1890) start with a very simple model of an economy and then proceed to more complex models. There is an important difference, however, between Walrasian general equilibrium analysis and Marshallian partial equilibrium analysis.

Walras first decomposes a complicated economy of the real world into several fundamental components like consumer-traders, entrepreneurs, consumers’ goods, factors of production, newly produced capital goods, and money. He then composes a simple model of a pure exchange economy by picking up a very limited number of such components, that is, individual consumer-traders and consumer's goods, disregarding the existence of all other components. Travel from this simple model to the complex proceeds by adding one by one those components so far excluded, that is, entrepreneurs and factors of production first, then newly produced capital goods, and finally money. In this journey each intermediate model, enlarged from a simpler one and to be enlarged into a more complex one, is still a closed and self-compact logical system. However, each of them is as unrealistic as the starting model, with the exception of the last, into which all the components of a real world economy have been introduced.

Marshall on the other hand studies a whole complex of a real world economy as such. Of course, he also simplifies his study at first by confining his interest to a certain limited number of aspects of the economy. But he does it not by disregarding the existence of other aspects but by assuming that other things remain equal. In this sense most of Marshall's models of an economy, though realistic, are open and not self-sufficient, since some endogenous variables (that is, the ‘other things’) remain unexplained and have to be given exogenously.

The simplest model of Walrasian economics is that studied in the theory of exchange, where goods to be exchanged among individual consumer-traders are assumed simply to be endowed to them and not considered as produced at cost. There exist no production activities in this hypothetical world. The corresponding simplest model of Marshall is that of the market day, in which goods to be sold are produced goods, although the amount available for sale is, for the time being, assumed to be constant. Production does exist in this temporary model, though the level of output is assumed to be unchanged. In

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that Walrasian model which includes production capital goods are introduced as a kind of factor of production but investment (that is, the production of new capital goods) simply does not exist. On the other hand, in Marshallian short-run theory, which is also a theory of production, investment is actually undertaken though the amount of currently available capital is given. In all of the Walrasian models of exchange, of production and of credit and capital formation there exists no money at all, until it is finally introduced in the theory of circulation and money. In Marshallian models on the other hand money exists from the beginning, though its purchasing power is sometimes assumed to be constant.

In other words, Marshallian theories correspond respectively to special states of the real world economy. The market day (temporary) and short-run models are just as realistic as the long-run model, where capitals are fully adjusted. Thus Marshallian models are practically useful to apply to what Hicks (1934) called particular problems of history or experience. On the other hand, Walrasian models are in general not useful for such practical purposes. They are designed to show the fundamental significance of such components of the real world economy as entrepreneurs and production, investment and the rate of interest, inventories and money, and so on, by successively introducing them into simple models which are then developed into more complex ones. Walras' theoretical interest was not in the solution of particular problems but in what Hicks called the pursuit of the general principles which underlie the working of a market economy.

From our standpoint we must emphasize that all exchanges have to be non-monetary (that is, direct exchanges of goods for goods) in all the Walrasian theories of exchange, production and capital formation and credit, since money has not yet been introduced. Relative prices (including the rate of interest) and hence consumption and production activities are determined in non-monetary real models without using money, while the role of the model of circulation and money lies only in the determination of the level of absolute prices by the use of the money (Morishima, 1977, ch. 11; Negishi, 1979, ch. 2). Thus Walrasian economics is completely dichotomized between non-monetary real theories and monetary theory, in the sense that all non-monetary real variables are determined in the former and money is neutral, that is it does not matter for the determination of such variables. 'That being the case, the equation of monetary circulation, when money is not a commodity, comes very close, in reality, to falling outside the system of equations of (general) economic equilibrium' (Walras, 1926, pp. 326–7).

2.•In each of his non-monetary theories Walras tried to show the existence of a general equilibrium in its corresponding self-compact closed model. General equilibrium is of course a state in which not only each individual consumer-trader (entrepreneur) achieves the maximum obtainable satisfaction (profit) under given conditions but also demand and supply are equalized in all markets. In a large economy, how can we make such a situation possible without introducing money? What kind of process of exchange should we consider in order to establish a general equilibrium without using money? Even in the most simple case of an exchange economy, it seems in general almost impossible to satisfy all individual traders by barter exchanges, unless mutual coincidence of wants accidentally prevails everywhere. Walras ingeniously solved this difficulty by his famous tâtonnement, a preliminary process of price (and quantity) adjustment which precedes exchange transactions and/or effective contracts. Suppose that all the individual consumer-traders and entrepreneurs meet in a big hall. Since all of them are assumed to be competitive price takers it is convenient to assume (though Walras himself did not do so explicitly) the existence of an *auctioneer* whose only role is to determine prices. At the start the auctioneer calls all prices (including the price of a bond) at random. Individual consumers and entrepreneurs make decisions on the supply and demand of all goods, factors of production and of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
bond, assuming that the prices cried by the auctioneer are fixed and that whatever amount they wish can actually be supplied and demanded at these prices. If total demand equals total supply for every good (including the factors of production and the bond) exchange takes place (or contracts are made) at these prices, and the problem is solved.

Generally, however, this will not be the case, in which event no exchange transaction should take place at all, even for a good for which total demand is equal to total supply, and every mutually agreed contract should be cancelled. The auctioneer cancels the earlier prices, which failed to establish a general equilibrium, and calls new prices by following the law of supply and demand, that is, raising (lowering) the price of each good for which the demand is larger (smaller) than the supply. The same procedure is repeated until general equilibrium is established. Actual exchange transactions take place and enforceable contracts are made only when every party can actually realize its plan of demand and supply.

Prices change in the process of tâtonnement and it is generally impossible for a single trader to purchase or sell whatever amount he wishes at going prices. Nevertheless, each trader behaves on the assumption that prices are unchanged and that unlimited quantities of demand and supply can be realized at the current prices. This conjecture is justified by the very fact that no exchange transactions are made and no trade contracts are in effect during the tâtonnement, until general equilibrium is established where prices are no longer changed, and every trader can purchase and sell exactly the amount he wishes at going prices.

In a monetary economy of the real world, where of course the tâtonnement assumption cannot be made and some exchange transactions actually take place before general equilibrium is established, even a competitive trader without power to control prices has to expect price changes and to try to sell when the price is high and to buy when the price is low, though he may not always succeed in doing so. This leads to the separation of sales and purchases, a separation which is made possible only by the use of money as the medium of exchange and the store of value. In Walrasian non-monetary real models where the tâtonnement assumption is made, on the other hand, sales and purchases are synchronized when general equilibrium is established so that there is no need for money, and indeed there is no reason why the role of medium of exchange should be exclusively assigned to a single item called money. Since equilibrium prices are already fixed and unchanged almost any non-perishable good can be used if necessary as a medium of exchange.

Walras considered tâtonnement even in his final model, that is, that of circulation and money. Since disequilibrium transactions are thus excluded and there is no uncertainty, there is no room here for money as a store of value. We have to assume therefore that people demand money only for the sake of convenience in transactions. Since all actual transactions are carried out at general equilibrium after the preliminary tâtonnement is over, however, this rationale for the demand for money is not at all convincing. The only role left for money is to determine its own price, that is, the general level of prices.

3.•Walras gave two solutions for general equilibrium of each of his non-monetary real models, as well as his monetary model. The first solution is the demonstration that the number of unknowns is equal to the number of independent equations, which Walras called the scientific or mathematical solution. But how can we find a solution of such equations, particularly when the number of equations is very large? The second solution of general equilibrium given by Walras (1926, pp. 162–3, 170–72) is tâtonnement itself, which is suggested by the mechanism of free competition in markets and is called the practical or empirical solution. Taking the example of the simple model of exchange these two solutions may be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

reformulated in modern notation as follows.

Consider an exchange economy of m goods and denote the price of and the excess demand for the jth good by p_j and E_j respectively. One condition for general equilibrium is that demand is equal to supply in all markets, that is

$$E_j(p_1, \ldots, p_m) = 0, \; j = 1, \ldots, m. \tag{1}$$

In view of Walras's Law that

$$\sum_j p_j E_j \equiv 0, \tag{2}$$

only (m–1) equations of (1) are independent, while we can assign the role of numéraire to the mth good so that p_m=1 since only relative prices are relevant in a non-monetary economy. Therefore (1) is replaced by

$$E_j(p_1, \ldots, p_{m-1}) = 0, \; j = 1, \ldots, m-1. \tag{3}$$

Equations (1) or (3) are derived from the competitive behaviour of individual consumer-traders. The ith consumer-trader is assumed to maximize his utility $U_i(x_{i1}, \ldots, x_{im})$, subject to the budget constraint

$$\sum_j p_j x_{ij} = \sum_j p_j y_{ij} \tag{4}$$

where x_{ij} and y_{ij} denote respectively the gross demand for the jth good by the ith consumer-trader and the given initial holding of the jth good of the ith consumer-trader. The excess demand for the jth good is then defined as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$E_j = \sum_i x_{ij} - \sum_i y_{ij}.$$
(5)

It is to be noted that excess demand is not defined in (1) and (3) explicitly as a function of the y_{ij}'s. The reason is that the y_{ij}'s are given constants and are assumed not to change through the process of exchange until the demand plans of all consumer-traders are simultaneously realized when general equilibrium is established. In other words the assumption of tâtonnement is already implicitly made in the mathematical or theoretical solution of general equilibrium.
The original form of Walrasian tâtonnement is the process of successive adjustment in each single market. Suppose the initial set of prices cried by the auctioneer $(p_1, \ldots, p_{m-1})$ does not satisfy the condition (3) of general equilibrium, and we are for example in a situation described by

$$\begin{gathered} E_1(p_1, \ldots, p_{m-1}) > 0 \\ E_2(p_1, \ldots, p_{m-1}) < 0 \\ \ldots \\ E_{m-1}(p_1, \ldots, p_{m-1}) > 0. \end{gathered}$$
(6)

The price of the first good p_1 is now adjusted by reference to its excess demand E, and increased in the situation (6) until an equilibrium in the first market is established, that is

$$E_1\left(p_1', p_2, \ldots, p_{m-1}\right) = 0.$$
(7)

Here E_1 is assumed to be decreasing with respect to p_1, an assumption which, writing the partial derivative of the excess demand function for the jth good with respect to the kth price by E_{jk}, may be symbolized by $E_{11}<0$.
Under the new price system $\left(p_1', p_2, \ldots, p_{m-1}\right)$ the remaining $m-1$ markets may or may not be in equilibrium. If the second market is out of equilibrium, again under the assumption that $E_{22}<0$, the price

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of the second good is changed from p_2 to p_2' so as to satisfy

$$E_2(p_1', p_2', p_3, \ldots, p_{m-1}) = 0. \quad (8)$$

(Generally, this will upset the equilibrium in the first market (7).) Under the price system $(p_1', p_2', p_3, \ldots, p_{m-1})$, then, the price of the third good p_3 is adjusted if the third market (where $E_{33}<0$) is out of equilibrium, upsetting the equilibrium in the second market (8) just established. In this way the last, $m-1$th market, where $E_{m-1,m-1} < 0$, is eventually cleared by changing the price system from $(p_1', \ldots, p_{m-2}', p_{m-1})$ into $(p_1', \ldots, p_{m-2}', p_{m-1}')$ so as to satisfy

$$E_{m-1}(p_1', \ldots, p_{m-2}', p_{m-1}') = 0. \quad (9)$$

By this time all the markets except the last, which were once cleared successively, have generally been thrown out of their respective equilibria again. Neither the price system we have just arrived at, $(p_1', \ldots, p_{m-1}')$, nor the initial system, $(p_1, \ldots, p_{m-1})$, is part of a general equilibrium. The question then is which of the systems is closer to a true general equilibrium that satisfies (3). Walras argued that the former price system is closer to equilibrium than the latter since for example $E_1(p_1', \ldots, p_{m-1}') \neq 0$ is closer to 0 than $E_1(p_1, \ldots, p_{m-1}) \neq 0$. The reason for this, according to Walras, is that the change from p_1 to p_1' which established (7) exerted a direct influence that was invariably in the direction of zero excess demand so far as the first good is concerned. But the subsequent changes from p_2 to $p_2', \ldots, p_{m-1}$ to p_{m-1}', which jointly moved the first excess demand away from zero, exerted only indirect influences, some in the direction of equilibrium and some in the opposite direction, at least so far as the excess demand for the first good is concerned. So up to a certain point they cancelled each other out. Hence, Walras concluded, by repeating the successive adjustment of $m-1$ markets along the same lines, that is, changing prices according to the law of supply and demand, we can move closer and closer to general equilibrium.

4.•Walras's argument for the convergence of the tâtonnement process to general equilibrium was intended to be, if successful, the first demonstration of the existence of competitive general equilibrium (Wald, 1936). As we said above, it was merely an argument for the plausibility of such convergence of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the process of tâtonnement, and cannot be considered as a rigorous demonstration of existence of equilibrium. Whether indirect influences of the prices of other goods on the excess demand of a given good cancel each other out will certainly depend on substitutability and complementarity between goods. For example, indirect influences are *not* cancelled out and the excess demand of a good *is* increased if the prices of all gross substitutes are raised and the prices of all gross complements are lowered. In addition to the Walrasian stability condition for a single market, that is, $E_{jj} < 0$ for all *j*, therefore, some conditions on the cross-effects of prices on excess demands, that is on E_{jk}, $j \neq k$, have to be imposed so as to demonstrate convergence.

It was Allais (1943, vol. 2, pp. 486–9) who first demonstrated the convergence of Walrasian tâtonnement by assuming gross substitutability, that is, $E_{jk} > 0$ for all $j \neq k$. To see whether the price system moves closer and closer to the general equilibrium, which he assumes to be at least locally unique, Allais defines the distance *D* of a price system from the equilibrium price system as the sum of the absolute values of the value of excess demand for all goods, including the numéraire. The convergence of tâtonnement is then demonstrated by showing that this distance *D* is always decreased by changes in prices that are made in accordance with the law of supply and demand. His demonstration may be reformulated in our notation as follows.

The distance to the general equilibrium is defined as

$$D = \sum_j \left| p_j E_j \right| \tag{10}$$

where the summation runs from *j*=1 to *j*=*m*, and E_j is defined as a function of $p_1, \ldots, p_{m-1}$ as in (3). In view of Walras' Law (2), *D* can be replaced either by the summation of positive excess demands

$$D_1 = \sum_j p_j \max(0, E_j) \tag{11}$$

or by the summation of negative excess demands

$$D_2 = -\sum_j p_j \min(0, E_j) \tag{12}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where max(0, E_j) denotes E_j if it is positive and 0 if E_j is negative, and min(0, E_j) denotes E_j if it is negative and 0 if E_j is positive. From (2), that is $D_1 - D_2 = 0$, it is clear that

$$D = 2D_1 = 2D_2 \tag{13}$$

so that whether D is increasing or decreasing can be seen by checking whether D_1 or D_2 (whichever is more convenient) is increasing or decreasing.
Suppose E_1 to be positive as in (6) and that p_1 is raised following the law of supply and demand. From (12), we have

$$\partial D_2 / \partial p_1 < 0 \tag{14}$$

since $E_{j1} > 0$ for any j such that $E_j < 0$, from gross substitutability. In other words, a change in the price of the first good from p_1 to p_1' so as to satisfy (7) decreases the sum of negative excess demands D_2 and therefore the distance D to the general equilibrium. Suppose next that $E_2(p_1', p_2, \ldots, p_{m-1})$ is negative and p_2 is lowered to p_2' so as to satisfy (8). From (11) this time, we have

$$\partial D_1 / \partial p_2 > 0 \tag{15}$$

since $E_{j2} > 0$ for any j such that $E_j > 0$ from gross substitutability. In other words, a decrease in the price of the second good from p_2 to p_2' decreases the sum of positive excess demands D_1 and therefore the distance D to the general equilibrium.
Generally, if E_j is positive and p_j is raised D is decreased, which can be seen from the fact that D_2 is decreased. Similarly, if E_j is negative and p_j is lowered again D is decreased, which can be seen from the consideration of the behaviour of D_1. Out of the general equilibrium D remains positive and there exists at least one non-numéraire good with non-zero excess demand, so that its price is changing. The distance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to the general equilibrium always decreases out of equilibrium, and therefore we can move closer and closer to that equilibrium by changing prices according to the law of supply and demand, provided that gross substitutability is assumed.

Though Walras discussed the behaviour of the process of successive adjustment, he was not against the consideration of *simultaneous adjustment processes* in all markets (Uzawa, 1960; Jaffé, 1981). If we assume that adjustments take place not only simultaneously but also continuously, the tâtonnement process that each rate of change of price is governed by excess demand can be described by a set of differential equations,

$$dp_j/dt = a_j E_j(p_1, \ldots, p_{m-1}),\ j = 1, \ldots, m-1, \tag{16}$$

where t denotes time and the a_j's are positive constants signifying the speed of adjustment in the jth market. The study of the behaviour of the solutions of (16), that is prices as functions of t, which was initiated by Samuelson (1941) is called the study of the *stability of competitive equilibrium* and has been extensively carried out by many mathematical economists (Arrow and Hahn, 1971, pp. 263–323; Negishi, 1972, pp. 191–206). It is well known that gross substitutability is also a sufficient condition for the *convergence of adjustment processes* like (16).

5.•The idea of tâtonnement was clearly suggested to Walras from the observation of how business is done in some well organized markets in the real world, like the stock exchanges, commercial markets, grain markets, fish markets. As a matter of fact, Walras was well informed of the actual operation of the Paris Stock Exchange where disequilibrium transactions actually did not occur (Jaffé, 1981).

Tâtonnement is therefore not entirely unrealistic as a model of adjustment in such special markets. However, it is certainly very unrealistic to apply such a model of special markets to the whole economy, since preliminary adjustments are usually not made before exchange transactions and effective contracts take place, even in markets where competition, though not so well organized, functions fairly satisfactorily. Of course, Walras would have admitted this, since tâtonnement was for him not so much a description of the process of adjustment in the markets of the real world as it was the demonstration of the existence of general equilibrium, that is a limit to which tâtonnement converges. It should be so interpreted not only in the case of successive tâtonnement, which reminds us of the Gauss–Seidel method of solving a set of simultaneous equations, but also in the case of simultaneous tâtonnement (16), where time t is not real calendar time, but hypothetical process time. This is no wonder, since Walrasian non-monetary models are not intended to be faithful descriptions of the real world. They are designed rather to make clear the significance of each component of the market economy and to uncover the general principles that underlie its working.

One may feel that such an interpretation of Walrasian tâtonnement is too strict and that the behaviour of not so well organized markets can be described approximately by the tâtonnement model. Walrasian tâtonnement may be interpreted as something like the laws of motion, that work strictly speaking only in the ideal frictionless world but which can be applied approximately to the real world. The law of supply

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and demand can certainly be applied even in markets where there is no auctioneer, traders are dispersed, and exchange transactions take place and effective contracts are made before equilibrium of demand and supply is established.

Prices are formed differently in each exchange transaction by negotiation between relevant parties of traders. The law of indifference tends to prevail, however, if the transmission of information is nearly perfect, since atomistic traders know the difficulty of purchasing (selling) at prices lower (higher) than the prices offered by competitors and there are, furthermore, arbitrage activities. If demand falls short of supply, it is suicidal for atomistic sellers to offer a price higher than the average market price, while an atomistic purchaser is unable to consider a price lower than the average market price when demand exceeds supply. With disequilibrium of supply and demand, exchange transactions can take place only if demanders (suppliers) can find suppliers (demanders). If demand is deficient therefore sellers consider cutting prices or increasing marketing costs in order to attract more purchasers, since a drastic increase in sales is expected from slight falls in price or slight increases in marketing costs when information is nearly perfect. By observing such behaviour by the sellers, the purchasers also insist on price cuts. Thus price falls in the face of excess supply. Similarly, market prices rise as the result of a similar process of disequilibrium exchange transactions in the face of excess demand, in which the roles of sellers and purchasers are interchanged from the case of excess supply.

Therefore, we can extend (16) to

$$\mathrm{d}p_j/\mathrm{d}t = a_j E_j(p_1, \ldots, p_{m-1}, y_{11}, \ldots, y_{nm}),\ j = 1, \ldots, m-1, \tag{17}$$

where the E_j's are again derived from (5) but have now to be considered explicitly as functions of the y_{ij}'s that is the stock of the jth good held by the ith consumer-trader, $i = 1, \ldots, n$, since the y_{ij}'s are no longer constants but instead are changed by disequilibrium transactions among the n consumer-traders. Here we cannot discuss in detail how the y_{ij}'s are changed as a result of transactions at disequilibria, and have to be content with the general assumption that their rates of change depend on everything, that is, we have

$$\mathrm{d}y_{ij}/\mathrm{d}t = F_{ij}(p_1, \ldots, p_{m-1}, y_{11}, \ldots, y_{nm}),\ i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{18}$$

where the F_{ij}'s are unknown functions incorporating rules for exchange transactions out of equilibria. Models of an economy with (17) and (18) are called non-tâtonnement models or *non-recontracting models*.

Generally, if a non-tâtonnement or non-recontracting process (17) and (18) converges, it does so to an

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
equilibrium that is different from that arrived at by the tâtonnement process (16), since changes in the y_{ij}'s due to disequilibrium exchange transactions have effects on (17) which do not exist in the case of (16). As Newman (1965, p. 102) correctly pointed out, however, the difference can be safely neglected, if the speed of price adjustment in every market is very high (that is, the a_j's in (17) are very large), since then markets arrive at equilibrium prices so rapidly that the effects of disequilibrium transactions are prevented from becoming serious. Although the possibility of disequilibrium transactions is not institutionally excluded and there may well be some, most transactions are actually carried out at equilibrium so that it looks as if the assumption of tâtonnement is satisfied. In this sense, tâtonnement models can be used to describe the behaviour of non-tâtonnement or non-recontracting markets in the real world.

6.•Although the tâtonnement model can be applied to markets that are not so well organized if the transmission of information is nearly perfect and the speed of price adjustment is rapid, the general equilibrium tâtonnement model (16) is still not a realistic description of the real world economy. The reason is that the role of money as the medium of exchange and a store of value is very important in the real world, while as we saw it is highly limited in a model where most exchange transactions are simultaneously carried out at equilibrium. To make our model more realistic so that sales and purchases take place at disequilibria and are separated by the use of money, therefore, we have to get rid of tâtonnement by arguing that the speed of price adjustment is not rapid in (17), so that disequilibrium transactions cannot be ignored.

If the transmission of information is perfect, the law of supply and demand can be applied even in not so well organized markets where no auctioneer exists and disequilibrium transactions take place. This is because every seller (purchaser) perceives an infinitely elastic demand (supply) curve and expects that a drastic increase in sale (purchase) is made possible by a slight reduction (increase) in price. If total demand falls short of total supply in a market, then every trader willingly reduces price or accepts a reduction in it. Similarly, if total demand exceeds total supply in a market every trader willingly raises price or accepts a rise in it.

The transmission of information may not be so perfect, however, in markets where traders are dispersed and so cannot meet in a big hall as they do in the case of Walrasian tâtonnement. Suppose that a market is segmented and transmission of information is perfect among closely related traders, but that it is not so between different segments. Individual traders are assumed to keep contact with current trade partners and not to leave the segment of the market in which they are currently located in search of more favourable trade conditions, unless either they are well informed of such conditions in other segments or trade conditions change unfavourably in the original segment. Possibly because of consideration of cost, traders are constrained by inertia and do not move unless shocked by information on other segments or by changes in the original segment.

Then even an atomistic seller (purchaser) does not perceive an infinitely elastic demand (supply) curve. A seller expects that sales cannot be increased very much by reduction of price since only those purchasers who are currently buying from him are well informed of the price reductions, and this information is not perfectly transmitted to those purchasers who belong to other segments of the market and who are not buying from him. When total demand falls short of total supply and other sellers do not raise the price, it cannot be expected that ‘their’ purchasers leave them in search of cheaper sellers. The same seller has to expect, however, that sales will be drastically reduced if the price is raised, since those customers who are currently buying will be well informed of this and will leave to search for cheaper

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sellers, which they can find easily when total demand falls short of total supply and there are many other sellers willing to sell more at the unchanged price.

Atomistic sellers perceive kinked demand curves, with a downward sloping segment for levels of sale higher than the current one, and an almost infinitely elastic segment for levels of sale lower than the current one, when the market is in excess supply. It is very likely then that price does not fall and remains sticky in the face of excess supply (Reid, 1981, pp. 65–6, 96–9; Negishi, 1979, p. 36). It may not pay to reduce price if demand cannot be increased very much. Similarly, an atomistic purchaser perceives a kinked supply curve with an upward sloping segment for levels of purchase higher than the current one, and an almost infinitely elastic segment for levels of purchase lower than the current one, when the market is in excess demand. Since the transmission of information is imperfect and the purchaser cannot attract many sellers by raising price, it may not pay to raise price even if a larger purchase is wanted at the current price. It is very likely, therefore, that price does not rise and remains sticky in the face of excess demand.

Thus prices may be sticky and may not be adjusted quickly by demand and supply in not-well-organized markets in the real world. The speed of adjustment in (17) need not be rapid enough to allow one to ignore the effects of disequilibrium transactions, so that the tâtonnement process (16) cannot then be regarded as a realistic description of adjustment in real-world markets. Walrasian tâtonnement models are, of course, not designed to describe such markets empirically. They are constructed to show how the market mechanism works beautifully under ideal conditions. No one can deny that Walrasian economics succeeded in accomplishing this purpose. The market mechanism, however, does not work so beautifully in the real world. It certainly manages to work somehow but quite often at the cost of prolonged disequilibria in markets, such as involuntary unemployment in the labour market and excess capacity in goods markets. This is why we have to supplement Walrasian economics by launching out into the study of non-Walrasian economics.

7.•*Recontracting*. Since the idea of recontracting is due originally to Edgeworth, who developed it in a way different from Walras's tâtonnement, the implication of Edgeworth's theory of recontract has to be carefully considered in its relation to the theory of tâtonnement in Walrasian economics. These two theories are different from each other in at least two ways, namely with respect to the law of indifference (the uniformity of prices) and to the provisional nature of revocability of trade contracts. The first problem is discussed below, while the second will be considered in the next section.

There have been different interpretations as to whether Edgeworth's *Mathematical Psychics* (1881) excluded disequilibrium transactions or assumed the irrevocability of contracts (Walker, 1973; Creedy, 1980). Even if we assume that disequilibrium transactions are excluded, however, the theory of recontract in Edgeworth is different from the theory of tâtonnement in Walrasian economics. The law of indifference (that is, the existence of uniform market prices even in disequilibria) is imposed as an axiom in the original Walrasian as well as in modern Walrasian economic theories. This axiom may be justified either through arbitrage activities or by the existence of the auctioneer, and enables individual traders to act as price takers who have only to adjust their plans of supply and demand to the given prices. Such an axiom is not imposed in Edgeworth's recontracting model.

To demonstrate his famous limit theorem (Bewley, 1973), Edgeworth starts with a simple two-good two-individual model of exchange, where a trader X offers a good x to a trader Y in exchange for a good y. If we consider the so-called Edgeworth box diagram, any point on the contract curve, where each of two individual traders is not worse off than before exchange, can be a final settlement of trade contract

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
which cannot be varied by recontract. To narrow down the range of possible final settlements Edgeworth introduces a second *X* and a second *Y*, each respectively identical to the first, both in tastes and initial endowments. Since identical traders have to be treated equally in any final settlement, we can still use the same box diagram. Now it can be shown that no final settlement of contract can contain points on the contract curve which give ‘small’ gains from trade to the *X* traders. Otherwise, it is ‘possible for *one* of the *Y*s (without the consent of the other) to *recontract* with the two *X*s, so that for all those three parties the recontract is more advantageous than the previously existing contract’ (Edgeworth, 1881, p. 35). Similarly, it is possible to exclude as final settlements those points which give ‘small’ trade gains to *Y* traders.

In this way Edgeworth shows that the range of possible final settlements shrinks as the number of identical traders grows. If there are infinitely many traders the only remaining final settlements turn out to be precisely the points of Walrasian equilibrium, each with a uniform price line, that is the common tangent to indifference curves of *X* and *Y* passing through the point of initial endowments. In the terminology of the modern theory of cooperative games, the core of the exchange game (that is, those allocations not blocked by any coalitions of players) consists only of the Walrasian equilibria when the numbers of the *X*s and the *Y*s are each infinitely large. Thus Edgeworth tries to show that the recontracting process in the large economy, where traders obtain a free flow of information through the making and breaking of provisional contracts, leads to the same uniform prices that are given by the auctioneer to price-taking traders in Walrasian equilibria. Though there are no uniform market prices and individual traders are not assumed to be price-takers in Edgeworth's recontracting process, the resulting equilibrium exchanges are the same as those obtained through Walrasian tâtonnement in a large economy. In such an economy, therefore, where information is perfect, we can safely argue as if there were uniform market prices and as if traders were price-takers. In a sense, Edgeworth justified the Walrasian axiom, since axioms of theories should be assessed not by themselves but by the results derived from them. Even if the Walrasian axiom is not itself realistic, the results derived from it can be as realistic as those derived from more realistic but more complicated axioms.

In later writings Edgeworth confirmed his early position on Walras and the uniformity of prices. Walras

> describes *a* way rather than *the* way by which economic equilibrium is reached. … Walras's laboured description of prices set up or ‘cried’ in the market is calculated to divert attention from a sort of higgling which may be regarded as more fundamental than his conception, the process of *recontract*…. The proposition that there is only one price in a perfect market may be regarded as *deducible* from the more axiomatic principle of recontract. (*Mathematical Psychics*, p. 40 and context: Edgeworth, 1925, vol. 2, pp. 311–23)

> We may add that even the existence of a uniform rate-of-exchange between any two commodities is perhaps not so much axiomatic as deducible from the process of competition in a perfect market. (Edgeworth, 1925, vol. 2, p. 453)

8.•It is possible to interpret Edgeworth's theory of bilateral exchange (Edgeworth, 1925, vol. 2, pp. 316–19) as a theory of a process where not only the rate of exchange is variable but also contracts are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
irrevocable. Starting from a situation with initial holdings, two goods are actually exchanged so as always to increase the utility of each of the two traders. Since exchanges are irrevocable, however, where on the contract curve this process of exchange will terminate depends on the path of exchanges as well as on the initial holdings. Hence it contrasts strongly with Walrasian tâtonnement, the equilibrium of which depends only on the initial holdings. There is no confusion, however, between this theory of Edgeworth and Edgeworth's theory of recontract interpreted in the sense of tâtonnement, since the modern extension of the former theory to the case of multiple traders is rightly called the theory of Edgeworth's *barter* process (Uzawa, 1962; Fisher, 1983, pp. 29–31).

Incidentally, Edgeworth's idea that exchanges necessarily take place only in the direction of increasing utilities can be relevant only in a barter economy. In a monetary economy an exchange of one good against another is decomposed into an exchange of the first good against money and an exchange of money against the second good. Even though the completed exchange of the two goods increases utility, its first half need not do so since in the course of the exchange process one may temporarily receive more money than one plans to keep eventually. In other words, one may impose a rule for non-monetary goods of no overfulfilment of demand and supply plans in the process of exchange, but this cannot be done for money, which has to act both as the medium of exchange and as a store of value beyond the current period.

In view of the current usage of the concept of recontracting in the sense of tâtonnement, what is confusing is the fact that Edgeworth sometimes, and particularly in his later writings, applied his recontracting model to situations where exchange transactions actually take place at disequilibria. To show that his model is of more than academic interest Edgeworth considered the case of a labour market, which each day ends in disequilibrium after exchange transactions have taken place at disequilibrium rates of exchange. From day to day, as the traders’ knowledge of the state of the disequilibrium in the market changes they progressively modify their behaviour, changing the rate of exchange in such a way that the market converges to equilibrium.

Since labour service is perishable within a day and the number and dispositions of the traders are assumed to be unchanged, this process over a sequence of days is formally equivalent to the recontracting process within a day, even though in the former process contracts made on the previous days are irrevocable while in the latter disequilibrium contracts are revocable. Edgeworth insisted that in this example of a process over a sequence of days (Edgeworth, 1925, vol. 1, p. 40) traders do recontract, in the full sense of *Mathematical Psychics*. Since contracts made in earlier days are irrevocable, however, in this case to recontract implies that a new contract is made which is different from that carried out on the previous day. It does not imply the cancellation of contracts already made.

Only a formal similarity exists between these two processes of recontracting, which is due to the assumption that disequilibrium exchange transactions do not really involve a permanent redistribution of wealth. Although labour service is perishable within a day, however, the money paid against labour service certainly is not and it is likely that a redistribution of wealth does take place over a sequence of days. Even from a formal point of view, then, Edgeworth's model of the labour market is rather a pioneering instance of *non-recontracting models*.

9.•No one can deny that the rigorous demonstration of the dynamic stability of tâtonnement under certain sufficient conditions has substantially improved on the original argument for the plausibility of its convergence that was made by Walras. More importantly, however, the recent studies on stability have helped us to understand the underlying economic assumptions of the Walrasian tâtonnement

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
process itself, and made us realize its considerable differences from most price adjustment processes in actual economies. The similar studies of Edgeworth's recontracting process have been helpful in the same way.

As we have shown, Walrasian tâtonnement is a realistic approximation to some actual adjustment processes, provided that the transmission of information is perfect and the speed of adjustment is rapid, as is roughly the case in well organized markets. The problem that remains to be studied, therefore, is the nature of adjustment processes when these conditions are not satisfied, that is, when markets are not so well organized. This is the problem of non-recontracting models in non-Walrasian or disequilibrium economies.

See Also

- auctioneer
- general equilibrium
- Walras, Léon

Bibliography

Allais, M. 1943. *Traité d'économie pure*. Paris: Imprimerie Nationale. 2nd edn, 1952.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Bewley, T.F. 1973. Edgeworth's conjecture. *Econometrica* 41, 425–54.

Creedy, J. 1980. Some recent interpretations of *Mathematical Psychics*. *History of Political Economy* 12, 267–76.

Edgeworth, F.Y. 1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: C. Kegan Paul & Co.

Edgeworth, F.Y. 1925. *Papers Relating to Political Economy*, 3 vols. London: Macmillan.

Fisher, F.M. 1983. *Disequilibrium Foundations of Equilibrium Economics*. Econometric Society Monographs in Pure Theory. Cambridge: Cambridge University Press.

Hicks, J.R. 1934. Léon Walras. *Econometrica* 2, 338–48.

Jaffé, W. 1967. Walras's theory of *tâtonnement*: a critique of recent interpretations. *Journal of Political Economy* 75, 1–19.

Jaffé, W. 1981. Another look at Léon Walras's theory of *tâtonnement*. *History of Political Economy* 13,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
313–36.

Kaldor, N. 1934. A classificatory note on the determinateness of equilibrium. *Review of Economic Studies* 1(February), 122–36.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Morishima, M. 1977. *Walras' Economics: A Pure Theory of Capital and Money*. Cambridge: Cambridge University Press.

Negishi, T. 1972. *General Equilibrium Theory and International Trade*. Amsterdam: North-Holland.

Negishi, T. 1979. *Microeconomic Foundations of Keynesian Macroeconomics*. Amsterdam: North-Holland.

Newman, P. 1965. *The Theory of Exchange*. Englewood Cliffs, NJ: Prentice-Hall.

Patinkin, D. 1956. *Money, Interest, and Prices: An Integration of Monetary Theory*. Evanston: Row, Peterson.

Reid, G.C. 1981. *The Kinked Demand Curve Analysis of Oligopoly: Theory of Evidence*. Edinburgh: Edinburgh University Press.

Samuelson, P.A. 1941. The stability of equilibrium: comparative statics and dynamics. *Econometrica* 9, 97–120.

Uzawa, H. 1960. Walras' tâtonnement in the theory of exchange. *Review of Economic Studies* 27, 182–94.

Uzawa, H. 1962. On the stability of Edgeworth's barter process. *International Economic Review* 3, 218–32.

Wald, A. 1936. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie* 7, 637–70.

Walker, D.A. 1973. Edgeworth's theory of recontract. *Economic Journal* 83, 138–49.

Walras, L. 1874–7. *Éléments d'économie politique pure ou théorie de la richesse sociale*. Definitive edn, Lausanne, 1926. Trans. W. Jaffé as *Elements of Pure Economics*, London: George Allen and Unwin; Homewood, IL: Richard D. Irwin, 1954.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Negishi, Takashi. "tâtonnement and recontracting." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000015>
doi:10.1057/9780230226203.1670

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Taussig, Frank William

Warren J. Samuels
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Clark, J. B.; comparative advantage; economic sociology; marginal productivity theory; marginalism; mathematics and economics; minimum wages; monopoly; protection; Schumpeter, J. A.; specie-flow mechanism; Taussig, F. W.; trade unions; unemployment insurance; wages fund

Article

Taussig was born on 28 December 1859 in St Louis, Missouri, and died on 11 November 1940 in Cambridge, Massachusetts. After starting college at Washington University, St Louis, he transferred to Harvard University, where he received the BA (1879), Ph.D. (1883) and LLB (1886). He also studied at the University of Berlin.

Taussig was one of the foremost US economists for half a century. He was on the Harvard faculty from 1885 to 1935, where he was a magisterial teacher and edited the *Quarterly Journal of Economics* from 1896 to 1936. A member of several government commissions, he was the first chairman of the US Tariff Commission, 1917–19, and an adviser to President Woodrow Wilson. He was president of the American Economic Association in 1904 and 1905. His *Principles of Economics* (1911) was the foremost US textbook for generations of economists and non-economists.

Taussig was accurately called ‘the American Marshall’ by Joseph Schumpeter because of his professional stature. He shared with Alfred Marshall an identification with the Ricardo–Mill tradition coupled with a willingness to integrate the ideas of marginalism; a scepticism of the mathematicization of economics; a desire to moderate conflict within the discipline; an understanding that economics was or should be more an organon of analysis, a collection of tools, than a body of doctrine; a sympathy for the working class; and a view that economics was to remain political economy, to include what Schumpeter called ‘economic sociology’. He preferred J.S. Mill's *Principles of Political Economy* to any modern text, including Marshall's *Principles*, because in his view it prevented delusions as to economic questions being easy of solution. Like Marshall, too, he considered the Austrian system needlessly complex.

His principal work as an economic theorist lay in wage theory and in international trade theory and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
policy. In the former, he attempted to resuscitate a modified version of the wages fund theory, centring on the relative inelasticity of the short-run supply of consumer goods, which he combined with marginal productivity theory. He stressed, however the role of non-competing groups (as part of his great realism in matters of stratification), criticized John Bates Clark's moralistic version of marginal productivity theory and argued that the frequent superior advantage in bargaining of employers meant that marginal productivity was only an upper limit, in the absence of effective competition.

In the field of international trade, in which he was the principal US figure for decades, his major concerns were the complexities of comparative advantage, the role of the specie flow mechanism and the international trade mechanism under non-specie monetary systems, and the history and analysis of protection. His position on protectionism was complex: he affirmed free trade, accepted the infant-industry argument with considerable scepticism of its application in practice, favoured gradual lessening of tariffs and (or but) affirmed a stable tariff system rather than policies of disruptive shifts and shocks. In other policy controversies he strenuously opposed the free coinage of silver as inflationary; criticized unemployment insurance and minimum wages as violating traditional individual initiative; thought progressive taxation less important than the elimination of monopoly and the use of education to diffuse opportunity; and, understanding of worker reliance on unions, blamed labour–management conflict on the failure of management to exercise the responsibilities of wealth and power and to be more understanding of worker interests and ambitions.

Taussig's political economy, or economic sociology, set him off from most other leading orthodox economists. He saw society as both a structure and a struggle for power and privilege, in which an instinct of domination and an impulse of emulation were prevalent if not dominant. These ideas pervaded his *Principles*, in which he presented, for example, a functional analysis of the leisure class and, in his *Inventors and Money Makers* (1915b) and (with C.S. Joslyn) *American Business Leaders* (1932), a further analysis of both the complex psychological bases of economic behaviour and the role of leadership in the successful operation of the market mechanism. Schumpeter wrote that Taussig was ‘among those few economists who realize that the method by which a society chooses its leaders, in what for its particular structure, is the fundamental social function … is one of the most important things about a society, most important for its performance as well as for its fate’ (Schumpeter, 1951, p. 217). In these respects, Taussig's work was compatible with institutionalism, but his establishment position apparently kept those in the later tradition from fully appreciating his contributions. Taussig's ideas here, supplementing those of Friedrich von Wieser, had some influence on Schumpeter himself.

See Also

- wages fund

Selected works

1888. *The Tariff History of the United States*, 8th edn, revised, New York: Putnam's, 1931.

1896. *Wages and Capital*. New York: D. Appleton.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1911. *Principles of Economics*, 4th edn. New York: Macmillan, 1939.

1915a. *Some Aspects of the Tariff Question*, 3rd edn, enlarged. Cambridge, MA: Harvard University Press, 1931.

1915b. *Inventors and Money Makers*. New York: Macmillan.

1927. *International Trade*. New York: Macmillan.

1932. (With C.S. Joslyn.) *American Business Leaders: A Study in Social Origins and Social Stratification*. New York: Macmillan.

Bibliography

Opie, R. 1941. Frank William Taussig (1959–1940). *Economic Journal* 51, 347–68.

Schumpeter, J.A. 1951. *Ten Great Economists*. New York: Oxford University Press.

How to cite this article

Samuels, Warren J. "Taussig, Frank William." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000016> doi:10.1057/9780230226203.1671

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tawney, Richard Henry (1880 – 1962)

J.M. Winter
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capitalism; individualism; Protestant ethic; religion and economic development; Tawney, R. H.; Weber, M.

Article

R.H. Tawney was an economic historian and socialist philosopher whose Anglican beliefs lay at the heart of his influential studies of the enduring problem of the ethics of wealth distribution. As Professor of Economic History at the London School of Economics from 1921 to 1958, he became the doyen of a school of thought which defined the subject as the exploration of the resistance of groups and individuals in the past to the imposition on them of capitalist modes of thought and behaviour.

In his first book, *The Agrarian Problem in the Sixteenth Century* (1912) – written to provide an appropriate text for his pioneering tutorial classes for the Workers' Educational Association – he examined patterns of rural development, protest and litigation surrounding the enclosure of land in Tudor England. After service in the British Army during the First World War – he was severely wounded on the first day of the Battle of the Somme – Tawney returned to his scholarship and developed the arguments which appeared in perhaps his best-known work, *Religion and the Rise of Capitalism* (1926). Here he showed how alien to the teachings of the Reformation was the assumption that religious thought had no bearing on economic behaviour. Tawney captured in classical prose the clash within religious opinion that preceded that abnegation of the social responsibility of the churches and suggested that 'religious indifferentism' was but a phase in the history of Christian thought.

In *Religion and the Rise of Capitalism* Tawney crystallized a number of ideas he had begun to consider in the pre-1914 period. In a commonplace book he kept from 1912 to 1914, Tawney jotted down notes on many of his religious and historical preoccupations. Among them is the simple query, 'I wonder if Puritanism produced any special attitude toward economic matters'. Over the following decade, he gathered evidence on this subject, and presented preliminary statements in the Scott Holland Memorial Lectures at King's College, London, in 1922, and in the lengthy introduction he wrote for a 1925 edition of Thomas Wilson's *Discourse on Usury* of 1569.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In the notes Tawney left concerning this facet of his historical research, there is no evidence whatsoever that he drew on the celebrated essay of Max Weber, originally published in 1905, on *The Protestant Ethic and the Spirit of Capitalism*. Indeed, a full appreciation of Tawney's Anglican concerns requires a divorce between the two partners of the so-called ‘Tawney–Weber’ thesis.

It is true that both men believed that (in Tawney's words) ‘The fundamental question to be asked, after all, is not what kind of rules a faith enjoins, but what kind of character it values and cultivates.’ They agreed as well that there was in Calvinism a corrosive force which undermined traditional doctrines of social morality in ways which would have shocked the early reformers. And they shared the view that in Protestant teaching there was an important emphasis in religious terms on the ‘inner isolation of the individual’ which reinforced a more general individualism of social and economic behaviour.

But what differentiates their work is the uses to which they put their interpretations of Protestantism. Weber's essay was but one part of a comprehensive study of the sociology of religion. It reflects his overriding concern with the development of what he termed the rational bureaucratic character of modern society. In both these facets of his work he charted the progressive, relentless, and irreversible demystification of the world.

Weber's essay helped foster a belief in the bleak permanence of the spirit of capitalism which Tawney laboured to refute throughout his work. *Religion and the Rise of Capitalism* was written precisely to counter the view that social indifferentism in religious thought and individualism in economic thought were unchangeable features of modern life. If Weber's purpose was to describe the demystification of the world, Tawney's was to help in the demystification of capitalism, by stripping it of some of its most powerful ideological supports, derived from one reading of the Protestant tradition.

Anglicanism is, of course, a house of many mansions, in which there is room for reactionaries and socialists alike. The view that capitalism is unchristian because it stultifies the common fellowship of men of different means and occupations has never been more than a minority view. But, at precisely the same time as he was writing *Religion and the Rise of Capitalism*, Tawney joined a number of other influential Anglicans who spoke out against capitalism as a way of life which violated the moral precepts of their faith.

This position was as evident in his essays in political philosophy as it was in his scholarship in economic history. In *The Acquisitive Society* (1921) and in *Equality* (1931), Tawney argued that capitalism was an irreligious system of individual and collective behaviour, since it was based on the institutionalization of distinctions between men based on inherited or acquired wealth. For a Christian, such divisions manifested a denial of the truth that all men are equally children of sin and equally insignificant in the eyes of the Lord. What Matthew Arnold had called the ‘religion of inequality’ was really the obverse of a Christian way of looking at the world.

Tawney's legacy has been particularly pervasive, because his voice had a resonance which appealed to many who did not share his religious outlook. This was in part because he wrote with the moral outrage of Marx and with the grace and eloquence of Milton. His strength lay too in the fact that his was a distinctively English voice. This did not prevent his advocacy of the comparative method in the study of economic history, best evidenced in his book *Land and Labour in China* (1932), written after an eight-month mission to China as an educational adviser to the League of Nations, and in a history of the American labour movement he wrote while adviser to Lord Halifax, British Ambassador to Washington during the Second World War.

But Tawney's influence lies more centrally in his writings on the moral issues posed by capitalist

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic development in Britain. His call for an alternative to the cash nexus – firmly within the tradition of Owen, Ruskin and Morris – has continued to strike a chord among many people not of religious temperament who have sought indigenous answers to the problems of a society crippled by the injuries of class.

See Also

- Fabian economics
- Weber, Max

Selected works

1912. *The Agrarian Problem in the Sixteenth Century*. London: Longmans.

1914. (ed., with A.E. Bland and P.A. Brown.) *English Economic History: Selected Documents*. London: Bell.

1921. *The Acquisitive Society*. London: Bell.

1924. (ed., with E. Power.) *Tudor Economic Documents*. London: Longmans.

1926. *Religion and the Rise of Capitalism*. London: Murray.

1927. ed. *Economic History: The Collected Papers of George Unwin*. London: Macmillan.

1931. *Equality*. London: Allen & Unwin.

1932. *Land and Labour in China*. London: Allen & Unwin.

1953. *The Attack and other Essays*. London: Allen & Unwin.

1958. *Business and Politics under James I: Lionel Cranfield as Merchant and Minister*. Cambridge: Cambridge University Press.

1964. *The Radical Tradition*. London: Allen & Unwin.

1972. *R.H. Tawney's Commonplace Book*, ed. D.M. Joslin and J.M. Winter. Cambridge: Cambridge University Press.

1978. *History and Society. Essays by R.H. Tawney*, ed. J.M. Winter. London: Routledge & Kegan Paul.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1979. *The American Labour Movement and Other Essays*, ed. J.M. Winter. Brighton: Harvester Press.

Bibliography

Terrill, R. 1974. *R.H. Tawney and His Times*. Cambridge, MA: Harvard University Press.

Winter, J.M. 1974. *Socialism and the Challenge of War*. London: Routledge & Kegan Paul.

How to cite this article

Winter, J. M. "Tawney, Richard Henry (1880–1962)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000017>
doi:10.1057/9780230226203.1672

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tax competition

Michael Keen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tax competition refers to strategic tax-setting in a non-cooperative game between jurisdictions – whether countries or states or provinces within a federation – with each setting some parameters of its tax system in relation to the taxes set by others. This creates a potential case for international tax coordination, though the revenue impact has as yet been modest (at least in OECD countries). Conflicting national interests make it difficult to design effective coordination schemes.

Keywords

average effective marginal tax; beggar-thy-neighbour; capital controls; cooperation; cross-border shopping; excise taxes; flat tax; marginal effective tax rate; tax competition; tax havens; taxation of corporate profits; transfer pricing

Article

Tax competition refers to strategic tax-setting in a non-cooperative game between jurisdictions – which can be countries, or states or provinces within a federation – with each setting some parameters of its tax system in relation to the taxes set by others.

Broadly read, this definition of tax competition encompasses essentially all tax policy, since every decision potentially requires some view as to the tax strategies formed in other jurisdictions. Typically though, the term is intended to focus on explicit interactions in tax-setting. And the central policy concern to which such interactions point is the potential for efficiency losses (or, as will be seen, gains) – and hence potential gains (or losses) from cooperation – to the extent that each policymaker ignores (or at least attaches less importance to) the impact of its own tax decisions on other jurisdictions, creating fiscal externalities between them.

These cross-border effects might, in principle, lead to taxes ending up too *high* from a collective perspective, rather than too low, as a result of ‘tax exporting’. Most obviously, countries with power in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the world market for some commodity may find it in their interest to explicitly tax those exports. Similar effects may also be at work in relation to corporate taxation: for example, to the extent that profits arising domestically accrue to foreigners, there is an incentive to tax them heavily (since the well-being of foreigners is presumably less valued than the welfare of domestic citizens).

But the main concern in this area has been with the potential for a ‘race to the bottom’ arising from the possibility that a cut in one country's tax rate will make other countries worse off – losing tax base and/or real economic activity. If each policymaker ignores these harmful cross-border effects, there is a risk of a ‘beggar thy neighbour’ situation in which tax rates end up generally too low in terms of the collective interest.

There are, though, two reasons why this may not pose as great a policy problem as it sounds. First, if policymakers are to some degree self-serving rather than wholly benevolent, then the constraints that tax competition imposes on them may on balance be beneficial for the citizenry. Second, even though governments do not cooperate explicitly, since the tax-setting game between them is played many times, they may find ways to tacitly sustain the cooperative outcome and so avoid inefficiencies.

Tax competition can affect many aspects of tax design, but the policy concerns it raises are naturally greatest where tax bases are most mobile. At an international level, this has meant a focus on the taxation of capital income and excises on such readily transportable and conventionally heavily taxed commodities as cigarettes and alcohol. For brevity, this article follows most of the literature (Wilson, 1999) in focusing on these two, and moreover focuses, within the former, on corporate taxation.

Excise tax competition

The international norm is that commodity taxes are levied on the ‘destination principle’, meaning that tax is charged according to the jurisdiction in which consumption occurs: French wine consumed in the United Kingdom, for example, is taxed at the United Kingdom rate. This leaves some room for strategic tax setting (a point developed by Friedlander and Vandendorpe, 1968): prohibited from levying an import tariff on a fellow member of the European Union, the United Kingdom, for example, might be tempted to set a relatively high excise on wine, domestic and imported, so as to dampen import demand and generate a terms of trade benefit or as a means of rent shifting. The scope for tax competition is greatly increased, however, where – as a consequence of legal shopping across borders by consumers and smuggling (for brevity, we refer to both as ‘cross-border shopping’) – the destination principle proves difficult to enforce, so that commodities are in effect taxed according to where they are produced: the ‘origin principle’.

With taxation on an origin basis – at least de facto, and to a potentially significant extent – countries have an incentive to set a relatively low tax rate in order to protect their tax base and perhaps gain base at the expense of others. And there are clearly instances in which cross-border shopping has been a significant influence in tax-setting. The classic example is the reduction in Canadian cigarette taxes in the early 1990s in response to widespread smuggling across the US border. Also, in the United Kingdom cross-border shopping with France has for many years been cited explicitly as constraining the rate applied to alcoholic drink. There is, it should be noted, a further complication with the excises. This is the risk that cigarettes will not even bear the tax charged by a low-tax country, but will simply be diverted to consumption without payment of any tax (Keen, 2002a).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Corporate tax competition

Competition in capital income taxation is potentially most powerful when the tax is levied on a ‘source’ or ‘territorial’ basis (that is, on the income derived in particular jurisdictions) rather than on a ‘residence’ basis (on the income that residents in country derive from their income around the world). This is because under a pure residence system the only way that individuals or companies can take advantage of lower tax rates offered in other countries is by changing their residence: simply by relocating their investments to low-tax jurisdictions, or transfer pricing profits into subsidiaries located there, they would not ultimately avoid the taxes in their country of residence.

In practice, capital income taxes often have a significant element of source taxation even when formally levied on a residence basis. Individuals may simply locate their savings in low-tax jurisdictions and fail to report the proceeds, being especially secure in this when the source country provides a measure of banking secrecy. At corporate level, exemption may be explicit, with outright exemption of corporation's earnings from abroad: this is the case, for example, in the Netherlands and (for countries with which it has a double tax treaty) Canada. Or it may be implicit: while many countries, including the United States and United Kingdom, in principle apply the residence principle, their taxes typically apply only when a multinational's subsidiary abroad pays dividends to the parent, so that those taxes can be deferred (and hence reduced in present value) by delaying repatriation. This brings the system closer to one of de facto source taxation.

Countries that apply the residence principle find it is coming under increasing pressure, particularly for individuals but also for corporations. One sign of this has been the emergence of corporate inversions in the United States and elsewhere, with companies shifting their residence to low-tax jurisdictions. Another is the increased emphasis on controlled foreign corporation (CFC) rules, under which profits of subsidiaries earned abroad, typically in low-tax jurisdictions, may be brought into tax even if not repatriated to the parent.

Importantly, the impact of the corporate tax on business decisions depends on more than just the rate of the tax: it depends on what allowances are available for depreciation, financial costs and other expenses. The question then arises as to precisely which aspects of the corporate tax countries might compete over. There are three candidate tax rates.

First, attention has naturally focused on the headline statutory rate of corporation tax. This is especially relevant to firms’ decisions regarding income shifting, meaning essentially the use of devices other than real investment – transfer pricing, financial arrangements and so on – to shift taxable receipts from countries in which the statutory rate is high to those in which it is low, and deductible expenses in the opposite direction. Statutory rates have fallen substantially since the 1980s. In the Organisation for Economic Co-operation and Development (OECD) countries, the top rate of corporation tax fell from 41 per cent in 1986 to around 27 per cent in 2007. And this reduction in statutory rates has not been confined to developed countries: in sub-Saharan Africa, for example, it fell, on average, by about ten points over the 1990s (Keen and Simone, 2004).

Second, decisions as to the level of investment in a given country depend on the marginal effective tax rate (METR), which is a summary measure of the combined impact of the statutory rate and the allowances available on the return that a firm must earn in order to provide investors with the after-tax return they require. This is an indicator of the effect of the tax system on the incentive to invest at the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
margin. If the METR is zero, for example, then the tax system has no impact on the marginal decision to invest even though it may collect revenue by taxing the return infra-marginal investments. Strikingly, for the OECD countries the METR has remained broadly stable over the 1990s (Devereux, Griffith and Klemm, 2002) – which indicates that the reduction in statutory rates has been offset to some degree by a broadening of the corporate tax base.

Third, the choice as to the country in which to locate a given discrete investment (a factory, for example) depends on a comparison of the average effective marginal tax rate (AETR) in each, this quantity reflecting the present value of taxes to be paid – including on infra-marginal profits – over the life of the project. In practice, the AETR often tracks the statutory tax rate quite closely (Devereux and Griffith, 2003), so that it too has fallen substantially over the 1980s.

The overall picture thus suggests that the developed countries have been competing aggressively over the statutory rate of corporation tax, consistent with a desire to benefit from (or prevent) income shifting, but have broadened the tax base sufficiently to leave overall incentives to real investment broadly unchanged.

In what sense is this tax competition?

It could be that these trends simply reflect common developments in a range of countries rather than direct interaction between them – perhaps a common desire to improve incentives by reducing top rates of personal income tax, with the reduction in the corporate tax an adjunct to this, driven by the need to prevent disincorporation. Or – consistent with the basic notion of tax competition above – it could be a form of ‘yardstick competition’, with governments perceiving that their electorates assess them in part by comparing them with neighbouring countries, and so use the tax system to signal a supportive attitude to business or wider competence in economic management. Distinguishing empirically between these two hypotheses – correlation due to competition for tax base or real investment, versus correlation due to common shocks or yardstick competition – is not easy. Brueckner (2003) provides a review of the empirical literature in this area. In the international context, the emerging empirical evidence does seem to suggest that tax competition in pursuit of tax base or investment is important: Devereux, Lockwood and Redondo (2003), for example, find that the correlation between corporate tax rates became greater as capital controls became weaker, which would not be the case if the correlation were due to common domestic shocks. More anecdotally, one recent sign of tax competition closer to yardstick form than to competition for mobile base may be the spread, in the last few years, of ‘flat tax’ systems characterized by low tax rates on personal income: while one can see an argument for a low corporate tax rate to attract mobile capital, it is hard to see a similar base gain from setting the tax rate on labour income at the same, low rate.

This account of developments in the various corporate tax rates suggests that the pattern of any corporate tax competition has been quite complex. Why is it that countries have apparently competed over the tax rate but not over the base? One possibility is that this enables them to attract the more mobile and profitable investments without overly jeopardizing revenue from the less mobile corporate tax base. But there is no direct evidence for this, so the pattern of competition remains something of a puzzle.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Revenue effects

There has been much talk of tax competition leading to the erosion – perhaps the elimination – of revenue from the corporate tax, and from capital income taxes more generally.

For the developed countries as a group, however, this has not happened. Indeed, for them corporate tax revenues since the 1990s have if anything increased. In the EU15, for example, corporate tax revenues rose from 2.5 per cent of GDP in 1990 to 3.2 per cent in 2004. Country experiences vary, however. In Japan, corporate tax revenue has decreased over the same period from 6.5 to 3.8 per cent of GDP, and in Germany from 2.3 to 0.6 per cent. While these revenue figures need to be interpreted with great caution, there is no sign (as of 2007) of a collapse.

Quite why revenue has held up in the developed countries is not fully understood. Base-broadening measures have in many cases been adopted alongside the reduction in statutory rates, as seen above, but it is not clear that these have been sufficient to provide a full explanation. For the United Kingdom, there is evidence (Devereux and Klemm, 2005) that while base-broadening played some role, the buoyancy of corporate tax revenues has reflected also growth of the corporate sector, especially the financial sector, that did not seem to be largely attributable to the tax itself. It may well be that in other countries too – though not, it seems, all – an increase in the share of profits in GDP is part of the reason. For the United States, Auerbach (2006) argues that the resilience of corporate tax revenue reflects increased volatility of profits: the asymmetric treatment of positive and negative profits (the former taxed, the latter attracting no rebate but only carry forward) implies that this would lead to an increase in expected tax revenues. It is by no means clear, of course, that any such trends – and, hence, the resilience of corporate tax revenues – can be expected to continue.

There is tentative but emerging evidence of quite a different story in developing countries. These, too, have seen a marked reduction in statutory rates of tax. But this has not been offset by an expansion of the base – revenues in relation to GDP appear to have been falling, by about one-fifth in the poorest of them since the 1990s. Given their pressing needs for revenue and greater reliance on corporate taxes – around 13 per cent of total revenues at the start of the 1990s, compared to about nine per cent for the high-income countries – it may be that corporate tax competition is a more pressing concern for developing countries than for developed.

Basic welfare concerns

The policy concern raised by tax competition is that the failure to coordinate – each jurisdiction attaching relatively little importance to the impact of its decisions on others – will lead to a worse outcome than could be obtained from some form of cooperation in tax-setting. The fear is that the revenue pressures which emerge will lead either to reduced government expenditure or reliance on other tax instruments, notably taxes and social contributions on labour income, that are more distortionary and/or less coherent with equity objectives. In the limit (and leaving aside location-specific rents, most classically from natural resources), the corporate tax, for instance, would be reduced to a benefit tax – that is, one which simply charges companies for the value of the services they receive. Competition for mobile capital may also distort the composition of public spending, as well as its level, with too much focus on public infrastructure complementary with capital and too little on public expenditure on items

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of benefit to consumers: too many airports, not enough libraries. The argument is developed by Keen and Marchand (1997), and there is emerging evidence to suggest it is of some real importance.

Winners and losers

This lack of coordination does not mean, however, that *all* countries will be worse off as a result of tax competition. The literature strongly suggests, in particular, that small countries are likely to set particularly low tax rates, and in doing so to be better off than they would be under schemes of cooperation that do not explicitly involve the large countries paying them transfers (Kanbur and Keen, 1993). This is because in setting low tax rates they have little to lose in terms of revenue from their domestic tax base but a lot to gain (both in revenue terms and, perhaps, for the scale and profitability of their financial sector) from attracting tax base initially located abroad. And of course many tax havens are indeed small jurisdictions.

Thus coordination is not necessarily Pareto-improving: inducing all countries to participate may require transfers to those who win from tax competition and/or the exercise or threat of some form of punishment for non-cooperation.

Tax competition may be good: ‘ taming the beast ’

Some see tax competition as a good thing, providing a market discipline that serves to limit the size of government, supplementing inadequate constitutional and electoral constraints. In this view, developed first by Buchanan and others taking the view of government as ‘Leviathan’, the inefficiencies from non-coordination reduce the welfare of the policymakers themselves but increase that of the citizenry. There is another more subtle argument to the effect that tax competition may be beneficial even when policymakers are wholly benevolent. This is because the possibility of capital flight to low-tax countries provides a way in which others can commit not to impose heavy *ex post* taxes on savings once they have been made (so overcoming a basic time consistency problem in taxing capital income): see Kehoe (1989).

Much of the policy debate on the desirability or otherwise of tax competition quickly becomes a sterile statement of ideology, with it being seen as either wholly bad or wholly good. But there is a strand of the literature that has sought to better understand the trade-off at issue, between improved efficiency of the tax system as a result of coordination and the possibility that some or all of the additional revenue this can be used to generate will be wasted. One model leads to a simple test: a coordinated increase in tax rates from the non-cooperative equilibrium increases the citizens’ welfare if and only if $\lambda < MDL/(1+MDL)$, where λ is the proportion of public expenditure that is wasted and *MDL* is the marginal deadweight loss from raising an additional euro of revenue (Edwards and Keen, 1996). This at least enables one to narrow down the scope of policy disagreement. Suppose, for example, it is agreed that the marginal deadweight loss from taxation is at least 15 cents per euro. Then all those who believe that government wastes no more than 13 cents per euro of its spending, at the margin, should agree that a small coordinated increase in the corporate tax rate would be beneficial. The key difficulty remains, however, of determining what proportion of spending is ‘wasted’ – or indeed what ‘waste’ means in this context, since different people may clearly take different views, for example, of the social value of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
spending that is essentially redistributional.
Besley and Smart (2007) provide a more subtle analysis of these political economy issues in a model of electoral competition, in which candidates may be either Leviathans or benevolent. The results that emerge are more nuanced, but include, importantly, the observation that intensified tax competition is likely to be beneficial to the citizenry only if pure Leviathans are sufficiently rare: otherwise the electoral process offers them little respite from exploitation.

Quantification

There have been a few attempts to estimate the welfare costs of tax competition, using computable general equilibrium models. On the corporate tax side, Parry (2003) finds the loss to be relatively small: about three per cent of capital tax revenues, and even less when a modest degree of 'Leviathanism' is present. Sørensen (2004a) also finds relatively small gains: less than one per cent of GDP. On the excises, Keen (2002a) stresses the difficulty of inferring the extent of any problem from the extent of observed cross-border shopping: tax competition could be so fierce, for example, that this is zero in equilibrium and yet a welfare loss is suffered from the inefficiently low equilibrium tax rate. But illustrative calculations in this case also suggest a relatively modest welfare loss: rarely more than two per cent of tax revenue.

Ensuring that all participating countries gain

The first problem in coordinating taxes is to ensure that all participants gain. For the EU, this is explicit in the unanimity rules; in other contexts, it is implicit in the national sovereignty of potential participants. In principle, the inefficiency being addressed implies that all could gain from coordination *if* accompanied by compensating transfers to some participants. This means, more specifically, making payments to any countries that gain from tax competition. This appears unlikely in practice, because of the appearance of rewarding those who are gaining at the expense of others. It may be that the best way to deal with this is by combining coordination with other measures, perhaps not in the tax area, that tend to benefit the winners: this is the 'Hicksian optimism' that by adopting a series of reforms that are potentially Pareto-improving one will arrive at an actual Pareto improvement. The alternative is the exercise of threat.

The 'third country' problem

Coordination among a subset of countries may be undermined if other countries do not also participate, a point that has been explicit, for example, in the negotiations leading to the EU Savings Directive. This raises the same issues of compensation and cajoling just discussed. The third country problem does not mean that coordination among that subset alone would actually leave them worse off (Konrad and Schjelderup, 1999). Simulation exercises do though suggest that, when capital mobility is high, the welfare gains may be far smaller when only a subset of countries participate (Sørensen, 2004a).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Full harmonization

There are circumstances in which harmonization – a term that has come to mean complete uniformity not only of rates but also of bases (especially problematic for the corporate tax, but not a trivial issue for the excises either) – is collectively beneficial. Keen (1989), for example, shows that starting from a Nash equilibrium in the setting of destination-based commodity taxes, convergence to an appropriately weighted average of the initial tax rates is Pareto-improving. This result supposes, however, that taxes are used only for strategic reasons: by adding a revenue motive the conditions for Pareto gains become much stronger, as shown in Lockwood (1997). In practice, moreover, such full convergence is not only politically unlikely but is overly restrictive as a means of dealing with coordination failure. The search has been for looser measures of coordination.

A minimum tax rate

Minimum rates are in principle an attractive way of limiting any ‘race to the bottom’, leaving potentially considerable leeway for national discretion in tax-setting. Even those countries initially below the minimum, and so required to raise their tax rates, may benefit from the adoption of a minimum tax: this is because when they raise their rates, countries above the minimum will be less threatened by their low rates and so will tend to set higher rates than they otherwise would. This increase in the rates set by those above the minimum reduces the damage to those forced to raise their rates, and may even cause the latter to gain (Kanbur and Keen, 1993). In this way, imposing a minimum rate may be Pareto-improving.

All this assumes, however, that countries compete only over the statutory rate of tax. That may be a reasonable assumption in relation to excise taxation (for which the EU has indeed adopted minimum rates), since there is relatively little scope for game-playing on the base itself. For the corporate tax, however, the danger is that unless there is also agreement on a common base for the corporate tax, countries may instead compete by narrowing their tax bases, offering more generous allowances for investment, and so on. Tax competition would thus simply manifest itself in a different form.

Formula apportionment

This refers to a corporate tax system – of the kind operated by the states in the United States and the provinces of Canada – under which the profits earned by multi-jurisdictional enterprises are allocated across those jurisdictions by means of some formula intended to capture the extent of its activities in each, and each jurisdiction then taxes the profits allocated to it at whatever rate it chooses.

The advantage of such a scheme is that it eliminates the incentive for multinationals to move profits between jurisdictions by transfer-pricing or financial arrangements, since these have no effect on aggregate profits and hence also no effect on the taxes charged in each jurisdiction. This in turn means that the jurisdictions have no incentive to set low tax rates in order to encourage such income-shifting. Companies will have an incentive to distort their activities across jurisdictions, however, in so far as this affects the weights by which their profits are allocated across jurisdictions. This makes it important, for example, that (in contrast to common practice in the United States) the capital stock should *not* enter the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
formula. If it does, companies will have an incentive to invest in jurisdictions with a low tax rate; and there will be an incentive to offer low tax rates in order to attract such investment. Indeed, the net effect could be that tax competition is actually worse under formula apportionment than under separate accounting (Sørensen, 2004b): encouraging a firm to invest a little more in a country produces revenue proportional to the *marginal* return on that investment under separate accounting, but proportional to the firm's *average* profit under formula apportionment. Since average returns tend to exceed marginal, this makes it more tempting to attract capital by offering a low tax rate. The general difficulty here is that under formula apportionment the corporate tax becomes to some degree a tax on whatever is used to define the weights. Thus, similar but perhaps less problematic effects arise with using sales and some measure of employment in the weights, these being the other main candidates.

Ring-fenced corporate tax regimes: good or bad?

A recurrent theme in attempts to identify especially ‘harmful’ aspects of corporate tax competition is the idea that special schemes which are ‘ring-fenced’ in the sense of being restricted to particular investors or sectors are especially damaging.
This may not be correct (Keen, 2002b; Janeba and Smart, 2003). Allowing countries to compete very aggressively over particularly mobile aspects of the corporate tax base while maintaining higher rates on the rest may lead to an outcome that is better for all concerned than one in which they are required to set the same tax rate on all parts of the base. The reason is simple: it may ultimately be less damaging for countries to compete very aggressively over a narrow base than to compete less aggressively over a wider one.

Codes of conduct

The attempt to prevent the spread of, and to roll back, particular practices by means of non-binding rules of the game has been a remarkable development in international taxation over recent years, with such a code of conduct being adopted in the EU and a similar approach being adopted as part of the OECD's harmful tax practice project. The question arises as to how much further such an approach can be pushed. For by focusing on special schemes and excluding general levels of corporate taxation – and even putting aside the possible reservation on ring-fencing just mentioned – the codes arguably miss the central issue: too low a general level of corporate taxation.

Information sharing: reinforcing destination and residence principles

A quite different strategy is to limit the scope for tax competition by seeking to strengthen the application of the destination and residence principles (for commodity and income taxes respectively). The former is difficult to do given the general trend towards seeking fewer border formalities rather than more, an effect amplified by the increased importance of international services that cannot be taxed as they cross the border. For the income tax, strengthening resident taxation would involve, in particular, limiting the scope for deferral of taxes by leaving earnings aboard, in turn entailing an extension and firmer enforcement of CFC rules.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Another key measure that has been the focus of much attention in recent years is the strengthening of international information exchange on tax matters. In relation to capital income, this is a key element of the EU savings directive (which requires member states to either provide information to others on the interest income earned by their residents or levy a withholding tax, most of the proceeds being transferred to the country of residence) and the OECD's harmful-tax project.

It may seem obvious that low-tax countries can only lose from sharing information, through a reduction in their tax base and associated activities. Bacchetta and Espinosa (1995), however, show that they may benefit from a strategic effect of such sharing: for the higher-tax countries will then be inclined to set higher rates than they otherwise would (prospective outflows being diminished by the prospect of discovery abroad), which also enables the low-tax country to raise its tax rate too. Building on this insight, Huizinga and Nielsen (2003) explore the choice between information sharing and the use of withholding taxes in a repeated game, while Keen and Ligthart (2006) show that, if the difference in country size (and hence non-cooperative tax rates) is large enough, then sharing some of the revenue raised as a result of shared information with the source country (contrary to standard practice) has an adverse impact on total revenue raised but provides a device for securing a Pareto gain. How powerful the strategic effects of information exchange are likely to be, and indeed how effective such measures are likely to be in a technical sense (given, not least, the absence of common taxpayer identification numbers across countries) remains to be seen.

Tax competition in federations

The discussion so far has been concerned with ‘horizontal’ tax competition, in the sense that the interaction has been between jurisdictions with their own distinct tax bases. In federal systems, however, different levels of government commonly share tax bases: this may be explicit – in the United States, for example, corporate income is taxed at both federal and state level – or implicit (the base of a federal income tax, for instance, would overlap substantially with that of a state sales tax). The latter gives rise to ‘vertical’ tax competition, which in itself might be expected to lead to taxes that are too high from the collective perspective: a lower-level government that increases its own tax rate is liable to take less than full account of the impact on federal revenues of the consequent contraction of the shared tax base.

Two aspects of vertical tax externalities have received particular attention. First, how does tax-setting at the two levels interact? Theory leaves this indeterminate: intuitively, the optimal tax rate set by a lower-level jurisdiction is likely to depend (inversely) on the elasticity of the base, the effect of a change in the federal tax rate is in principle typically unclear (see, for example, Keen, 1998). Empirically, Hayashi and Boadway (2001), for example, find that higher federal rates of corporate taxation in Canada are associated with lower provincial tax rates. Second, how does the interplay between horizontal and vertical externalities play out (with the former pointing in most models to tax rates being too low, and the latter to their being too high)? In a model of capital income tax competition, Keen and Kotsogiannis (2002) show that this turns on the relative magnitudes of the elasticities of the demand for capital (which shapes the aggressiveness of horizontal tax competition) and of the supply of savings (shaping the responsiveness of the shared tax base). In a model of excise taxation, capturing both directions of externality, Devereux, Lockwood and Redoano (2007) show that the balance between the two depends on the ease of cross-border arbitrage and the price elasticity of demand. Using data for the US states,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
they also find evidence of significant vertical interactions in the setting of gasoline taxes, with the federal tax tending to be positively associated with state taxes.

Conclusion

The marked reduction in statutory corporate tax rates over the 1980s, which seems to be largely if perhaps not entirely attributable to international tax competition, has not been matched by a similar reduction in corporate tax revenues – at least in developed countries. Quite why remains something of a puzzle, and the possibility of more marked reductions in the future cannot be ruled out. The welfare significance of these developments also remains a matter of dispute, most fundamentally because the view that tax competition may provide a useful discipline on government has not been developed to a point at which it has firm empirical substance. The case for coordination thus remains uneasy, and is perhaps strongest in developing countries given both the more apparent impact on revenues and the clearer need there for stronger revenue mobilization. If coordination is sought, the key difficulty is to ensure that it takes a form from which all participants gain – in the absence of explicit transfers, this is not easy to assure, and may require packaging measures within some broader agreement. Recent policy initiatives have focused more on administrative measures than on substantive policy restrictions: it remains unclear how effectively these will deal with the underlying problems that motivate their use.

See Also

- excise taxes
- taxation of corporate profits
- tax havens
- taxation of foreign income

Bibliography

Auerbach, A. 2006. The future of capital income taxation. *Fiscal Studies* 27, 399–420.

Bacchetta, P. and Espinosa, M.P. 1995. Information sharing and tax competition among governments. *Journal of International Economics* 39, 103–21.

Besley, T. and Smart, M. 2007. Fiscal restraints and voter welfare. *Journal of Public Economics* 91, 755–73.

Brueckner, J. 2003. Strategic interaction among governments: an overview of empirical studies. *International Regional Science Review* 26, 175–88.

Devereux, M. and Klemm, A. 2005. Why has the U.K. corporation tax raised so much revenue? *Fiscal Studies* 25, 367–88.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Devereux, M. and Griffith, R. 2003. Evaluating tax policy for location decisions. *International Tax and Public Finance* 10, 107–26.

Devereux, M., Lockwood, B. and Redoano, M. 2007. Horizontal and vertical indirect tax competition: theory and evidence from the USA. *Journal of Public Economics* 91, 451–79.

Devereux, M., Griffith, R. and Klemm, A. 2002. Corporate income tax reforms and tax competition. *Economic Policy* 35, 451–95.

Devereux, M., Lockwood, B. and Redoano, M. 2003. Capital account liberalization and corporate taxes. IMF Working Paper 03/180. Online. Available at http://0-www.imf.org.library.lemoyne.edu/external/pubs/ft/wp/2003/wp03180.pdf, accessed 3 July 2007.

Edwards, J.S.S. and Keen, M.J. 1996. Tax competition and Leviathan. *European Economic Review* 40, 113–43.

Friedlander, A.F. and Vandendorpe, A.L. 1968. Excise taxes and the gains from trade. *Journal of Political Economy* 76, 1058–68.

Hayashi, M. and Boadway, R. 2001. An empirical analysis of intergovernmental tax interaction. *Canadian Journal of Economics* 34, 481–503.

Huizinga, H. and Nielsen, S.B. 2003. Withholding taxes or information exchange: the taxation of international interest flows. *Journal of Public Economics* 87, 39–72.

Janeba, E. and Smart, M. 2003. Is targeted tax competition less harmful than its remedies? *International Tax and Public Finance* 10, 259–80.

Kanbur, R. and Keen, M. 1993. Jeux Sans Frontières: tax competition and tax coordination when countries differ in size. *American Economic Review* 83, 877–92.

Keen, M. 1989. Pareto-improving indirect tax harmonization. *European Economic Review* 33, 1–12.

Keen, M. 1998. Vertical tax externalities in the theory of fiscal federalism. *IMF Staff Papers* 45, 454–85. Online. Available at http://0-www.imf.org.library.lemoyne.edu/external/Pubs/FT/staffp/1998/09-98/pdf/keen.pdf, accessed 3 July 2007.

Keen, M. 2002a. Some international issues in commodity taxation. *Swedish Economic Policy Review* 9, 11–45.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Keen, M. 2002b. Preferential regimes can make tax competition less harmful. *National Tax Journal* 54, 757–62.

Keen, M. and Kotsogiannis, C. 2002. Does federalism lead to excessively high taxes? *American Economic Review* 92, 363–70.

Keen, M. and Ligthart, J. 2006. Incentives and information exchange in international taxation. *International Tax and Public Finance* 13, 163–80.

Keen, M. and Marchand, M. 1997. Fiscal competition and the pattern of public spending. *Journal of Public Economics* 66, 33–53.

Keen, M. and Simone, A. 2004. Tax policy in developing countries: some lessons from the 1990s, and some challenges ahead. In *Helping Countries Develop: The Role of the Fiscal Policy*, ed. S. Gupta, B. Clements and G. Inchauste. Washington DC: IMF.

Kehoe, P. 1989. Policy cooperation among benevolent governments may be undesirable. *Review of Economic Studies* 56, 289–96.

Konrad, K. and Schjelderup, G. 1999. Fortress building in global tax competition. *Journal of Urban Economics* 46, 156–67.

Lockwood, B. 1997. Can international commodity tax harmonization be Pareto-improving when governments supply public goods? *Journal of International Economics* 43, 387–408.

Mintz, J. and Tulkens, H. 1986. Commodity tax competition between member states of a federation. *Journal of Public Economics* 29, 133–172.

Parry, I.W.H. 2003. How large are the welfare costs of tax competition? *Journal of Urban Economics* 54, 39–60.

Sørensen, P.B. 2004a. International tax coordination: regionalism versus globalism. *Journal of Public Economics* 88, 1187–214.

Sørensen, P.B. 2004b. Company tax reform in the European Union. *International Tax and Public Finance* 11, 91–115.

Wilson, J.D. 1986. A theory of interregional tax competition. *Journal of Urban Economics* 19, 296–315.

Wilson, J.D. 1999. Theories of tax competition. *National Tax Journal* 52, 269–304.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Zodrow, G. and Mieszkowski, P. 1986. Pigou, property taxation and the underprovision of public goods. *Journal of Urban Economics* 19, 356–370.

How to cite this article

Keen, Michael. "tax competition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000018> doi:10.1057/9780230226203.1673

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tax compliance and tax evasion

Joel Slemrod

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tax evasion is widespread, always has been, and probably always will be. Variations in duty and honesty can explain some of the across-individual and, perhaps, across-country heterogeneity of evasion. But the stark differences in compliance rates across taxable items that line up closely with detection rates suggest strongly that deterrence is a power factor in evasion decisions. Although the normative theory of taxation has been extended to tax system instruments such as the intensity of enforcement, the empirical knowledge for operationalizing these rules is sparse.

Keywords

civic virtue; deterrence; reciprocity; tax compliance; tax evasion; tax incidence

Article

No government can announce a tax system and then rely on taxpayers' sense of duty to remit what is owed. Some dutiful people will undoubtedly pay what they owe, but many others will not. Over time the ranks of the dutiful will shrink, as they see how they are being taken advantage of by the others. Thus, paying taxes must be made a legal responsibility of citizens, with penalties attendant on noncompliance. But even in the face of those penalties, substantial tax evasion exists – and always has.
Determining the extent of evasion is not straightforward, for obvious reasons. Because tax evasion is both personally sensitive and potentially incriminating, self-reports are vulnerable to substantial underreporting. Moreover, the dividing line between illegal tax evasion and legal tax avoidance is blurry. Under US law, tax evasion refers to a case in which a person, through commission of fraud, unlawfully pays less tax than the law mandates. Tax evasion is a criminal offence under federal and state statutes, subjecting a person convicted to a prison sentence, a fine, or both. An overt act is necessary to give rise to the crime of income tax evasion; therefore, the government must show wilfulness and an affirmative act intended to mislead. Some tax understatement is, however, inadvertent error, due to ignorance of or confusion about the tax law (as is some overpayment of taxes). Although the theoretical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
models of this issue generally refer to wilful understatement of tax liability, the empirical analyses cannot precisely identify the taxpayers' intent and therefore cannot precisely separate the wilful from the inadvertent. Nor can they, in complicated areas of the tax law, precisely distinguish the illegal from the legal.

The most careful and comprehensive estimates of the extent and nature of tax noncompliance anywhere in the world have been made for the federal taxes that the US Internal Revenue Service (IRS). The IRS comes up with its estimates by combining information from random intensive audits with information obtained from ongoing enforcement activities and special studies about sources of income, such as tips and cash earnings of informal suppliers like nannies and housepainters, that can be difficult to uncover even in an intensive audit.

The latest tax gap estimate, released in February 2006 (IRS, 2006) but pertaining to the 2001 tax year, estimated the overall gross tax gap estimate to be 345 billion dollars, which amounts to 16.3 per cent of estimated actual (paid plus unpaid) tax liability. Of the 345 billion dollar estimate, the IRS expects to recover 55 billion dollars, resulting in a 'net tax gap' – that is, the tax not collected – for tax year 2001 of 290 billion dollars, which is 13.7 per cent of the tax that should have been reported.

About two-thirds of all underreporting happens on the individual income tax; the corporation income tax makes up slightly more than ten per cent and the payroll tax gap makes up about one-fifth of total underreporting. For the individual income tax, understated income, as opposed to overstating of exemptions, deductions, adjustments, and credits, accounts for over 80 per cent of underreporting of tax. Business income, rather than wages or investment income, accounts for about two-thirds of the understated individual income. Taxpayers who were required to file an individual tax return, but did not, accounted for slightly less than ten per cent of the gap.

There are wide variations in the rate of misreporting as a percentage of actual income by type of income (or offset). Only one per cent of wages and salaries and four per cent of taxable interest and dividends are underreported. In large part this is because wages and salaries, as well as interest and dividends, must be reported to the IRS by those who pay them; in addition, wages and salaries are subject to employer withholding. Self-employment business income is not subject to information reports or withholding, and its estimated noncompliance rate is sharply higher. An estimated 57 per cent of non-farm proprietor income is not reported – 68 billion dollars – which by itself accounts for more than a third of the total estimated underreporting for the individual income tax. All in all, over half of underreporting is attributable to the underreporting of business income, of which non-farm proprietor income is the largest component.

All in all, there is substantial evidence that the extent of evasion for sole proprietor income is high compared to such income sources as wages, salaries, interest and dividends, and may be more than half of true income. Other components of taxable income for which information reports are nonexistent or of limited value, such as other non-wage income and tax credits, also have relatively high estimated misreporting rates. The IRS reports (IRS, 2006) that the net misreporting rate is 53.9, 8.5, and 4.5 per cent for income types subject to 'little or no,' 'some,' and 'substantial' information reporting, respectively, and is just 1.2 per cent for those amounts subject to both withholding and substantial information reporting.

Little is known about how the level of noncompliance, and its proportion to actual income, varies by income class. One study based on IRS audit data for 1988 suggested that higher-income people evade *less*, in relation to the size of their true income, than those with lower incomes, but for a number of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
reasons this study is not conclusive. Other studies suggest that married filers, taxpayers younger than 65, and men have significantly higher average levels of noncompliance than others. Within any group defined by income, age, or other demographic category, there are some who evade, some who do not, and even some who overstate tax liability. It is not known to what extent this heterogeneity is explained by different ‘tastes’ for evasion or different opportunities to evade.

Noncompliance is also a factor with businesses, both in their role as withholding agents for taxes that are not statutorily levied on businesses, and also for taxes that are levied on businesses, such as the corporation income tax. Based largely on operational data, the IRS estimates that noncompliance with the corporation income tax in 2001 was 30 billion dollars, which corresponds to a noncompliance rate of 17 percent. Of this 30 billion dollars, noncompliance by corporations with over 10 million dollars in assets make up 25 billion. But the estimated noncompliance rate of the larger companies is lower, 14 per cent compared to 29 per cent for corporations with less than 10 million dollars of assets. Because these estimates are largely based on deficiencies proposed by the examination teams of operational audits, and because most big corporations are routinely audited, these tax gap estimates are subject to several caveats. Because of the complexity of the tax law, exactly what is actual tax liability – and therefore what is actual tax noncompliance – is often not clear. In any given audit, some noncompliance may be missed, and there will also be mistakes in characterizing as noncompliance what is legitimate tax planning. Knowing that the resolution of the ultimate tax liability is often a long process of negotiation, the tax liability according to the originally filed return, as well as the initial deficiency assessed by the examination team, may be partly a tactical ‘opening bid’ that is neither party's best estimate of the ‘true’ tax liability.

It is difficult to compare the magnitude and nature of tax evasion in the United States with other countries, in part because no other country has undertaken a broad-based analysis of tax evasion like that undertaken in the United States. Based on less extensive analysis, the Swedish Tax Agency has estimated the total gap as a percentage of taxes at eight per cent in 2000. Although no official estimate for the United Kingdom has been released, a government document has speculated that it is likely that the United Kingdom has a tax gap of a similar magnitude to that of Sweden and the United States. Many studies suggest that noncompliance rates in developing countries are considerably higher.

Economics models have tried to put these facts into a coherent model. The standard economics framework for considering an individual's choice of whether and how much to evade taxes is a deterrence model in which taxpayers make these decisions in the same way they would approach any risky decision or gamble – by maximizing expected utility – and are influenced by possible penalties no differently than any other contingent cost. Optimal tax evasion depends on the chance of getting caught and penalized, the size of the penalty for evasion, and the individual's degree of risk aversion.

Attempts to empirically verify the predictions of the deterrence model of tax evasion have focused on the effect on evasion of enforcement intensity and the level of tax rates, but have been plagued by the same measurement issues that arise in assessing the magnitude of tax noncompliance. Perhaps the most compelling empirical support for the deterrence model is the cross-sectional variation in noncompliance rates across types of income and deductions. Line item by line item, there is a clear negative correlation between the noncompliance rate and the presence of enforcement mechanisms such as information reporting and employer withholding. A striking example of the link from a lack of deterrence to tax compliance involves state use taxes, which are due on sales purchased from out-of-state vendors but consumed in the state of residence. These taxes are largely unenforceable (except perhaps for some

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
expensive items like cars), and noncompliance rates are in the range of 90 per cent. The effect on noncompliance of the penalty for detected evasion, as distinct from the probability that a given act of noncompliance will be subject to punishment, has not been compellingly established empirically.
Although the deterrence model has dominated the economics literature, some have argued that it predicts a compliance rate much lower than what we actually observe, and that factors such as duty and reciprocal altruism can explain this. Some have argued that many taxpayers comply with tax liabilities because of ‘civic virtue’, and that more punitive enforcement policies may crowd out such intrinsic motivation by making people feel that they pay taxes because they have to, rather than because they want to. Others argue that tax evasion decisions depend on perceptions of the fairness of the tax system or what the government uses tax revenues for. But such individual judgements can be complex; for example, expenditures on warfare might be tolerated in a patriotic period, but rejected during another period characterized by anti-militarism. These patterns suggest that a form of reciprocal altruism may be at work where taxpayer behaviour depends on the behaviour, motivations, and intentions not of any subset of particular individuals, but of the government itself. In support of this view, surveys show a positive relationship across countries between attitudes towards tax evasion and professed trust in government.
There is, however, no clear evidence that tax compliance behaviour can be easily manipulated by the government to lower the cost of raising resources. Appeals to patriotism to induce citizens to pay their taxes (and, often, buy war bonds) are common; the US Secretary of Treasury during the First World War, William Gibbs McAdoo, referred to these campaigns as ‘capitalizing patriotism’. That such campaigns are successful during ordinary (non-war) times in convincing taxpayers to forego the cost-benefit calculus and comply has not been compellingly demonstrated. Recent randomized field experiments in the state of Minnesota and in Switzerland have found no evidence that appeals to taxpayers’ consciences, stressing either the beneficial effects of tax-funded projects or conveying the message that most taxpayers were compliant, had a significant effect on compliance.
The difficulties of separating out whether people pay their taxes because they feel they ‘ought to’ or whether they fear the penalties attendant to not doing so is well illustrated by some evidence from a recent survey sponsored by the Internal Revenue Service (IRS Oversight Board, 2006). While 96 per cent of those surveyed in 2005 mostly or completely agreed that ‘It is every American's civic duty to pay their fair share of taxes’, 62 per cent also said that ‘fear of an audit’ had a great deal or somewhat of an influence on whether they report and pay their taxes ‘honestly’.
Tax evasion has policy implications because it affects the distribution of the tax burden as well as the resource cost of raising taxes. Variations in compliance rates by income class can to some extent be offset by adjustments in the rate schedule, but it is practically impossible to offset variations within an income class, so that evasion creates horizontal inequity because equally well-off people end up with different tax burdens.
Tax evasion also imposes efficiency costs. The most obvious are the resources taxpayers expend to implement and camouflage noncompliance, and the resources the tax authority expends to address this. In addition, when the tax system is otherwise close to optimal it provides a socially inefficient incentive to engage in those activities for which it is relatively easy to evade taxes. For example, because the income from house painting can be done on a cash basis and is therefore harder to detect, this occupation is more attractive than otherwise. Although a supply of eager and cheap house painters undoubtedly is greeted warmly by prospective buyers of that service, the work of the extra people drawn to house

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
painting, or any activity that facilitates tax evasion, would have higher value in some alternative occupation.

The same argument applies to self-employment generally, as the enhanced opportunity for noncompliance inefficiently attracts people who would otherwise be employees. The opportunity for noncompliance can distort resource allocation in a variety of other ways, such as causing companies that otherwise would not find it attractive to set up a financial subsidiary, or set up operations in a tax haven, to facilitate or camouflage abusive avoidance or evasion.

The mere presence of tax evasion does not imply a failure of policy. Just as it is not optimal to station a police officer at each street corner to eliminate robbery and jaywalking, it is not optimal to completely eliminate tax evasion. Recognizing tax evasion introduces a new set of policy instruments whose optimal setting is at issue, such as the extent of audit coverage, the strategy for choosing audit targets, and the penalty imposed on detected evasion. It also invites a rethinking of standard taxation problems. One important issue is how many resources to devote to enforcing the tax laws. One superficially intuitive rule – increase the probability of detection until the marginal increase of revenue thus generated equals the marginal resource cost of so doing – is incorrect. Although the cost of hiring more auditors, buying better computers and the like is a true resource cost, the revenue brought in does not represent a net gain to the economy, but rather a transfer from private (noncompliant) citizens to the government. The correct rule equates the marginal social benefit of reduced evasion, which is not well measured by the increased revenue, to the marginal resource cost. The distinction suggests that unregulated privatization of tax enforcement, in which profit-maximizing firms would maximize revenue collection net of costs, would lead to socially inefficient overspending on enforcement. The social benefit is related to the reduced risk bearing that comes with reduced tax evasion and a reduction in the resource misallocations generated by evasion. Some have suggested that the basic framework of social welfare maximization is inappropriate, and have argued that there should be a specific social welfare discount applied to the utility of those who are found to be guilty of tax evasion and thus are known to be ‘antisocial’; the standard normative model applies no such discount, so that noncompliant taxpayers do not per se receive a lower social welfare weight than compliant taxpayers.

No one has yet compellingly translated this theoretical characterization of optimal enforcement into a statement about how much evasion should be tolerated. But its implication for interpretation of the tax gap is clear and was stated by former IRS Commissioner Lawrence Gibbs, who said that the tax gap estimates are not intended to be measures of the potential for additional enforcement yields because some would not be ‘cost-effective’ to collect. An economist would substitute the term ‘socially optimal’ for ‘cost-effective,’ but the spirit of Gibbs's remark is essentially correct. Just as there is an important difference between oil reserves and ‘economically recoverable’ oil reserves, there is a difference between tax evasion and economically (read optimally) recoverable tax evasion.

The normative theory has not yet made much progress in guiding policy regarding the key tools of tax administration, especially the role of information reporting by arms-length parties. The ability of the IRS to rely on reports by firms about wages and salaries paid to employees explains why the (optimal) noncompliance rate of labour income is so much lower than for self-employment income, for which no such information reports exist. The ability to match firm-to-firm sales is touted by advocates as a major administrative advantage of value-added taxes, and the difficulty of monitoring firm-to-consumer sales and to distinguish them from firm-to-firm sales has been noted as the Achilles heel of administering a retail sales tax. Overall, when relatively disinterested third parties can be required to provide

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
information, as they are with wages and salaries, high compliance rates can be achieved at fairly low cost. But when there are only interested parties involved, an alternative mechanism must be found – such as the requirement in an invoice-credit value added tax that taxes on input purchases can be deducted only if the seller produces an invoice for taxes remitted – or else compliance will be low in the absence of costly auditing.

The ubiquity and importance of evasion call into question one of the canons of undergraduate public finance textbooks – that the incidence and efficiency of taxes does not in the long run depend on which side of the market the tax is levied. Once the reality of tax evasion is recognized, the incidence and efficiency of a tax system may depend critically on which side of the market *remits* the tax to the government and which side must report its transactions to the government. A uniform value-added tax and a uniform national retail sales tax may look identical in a world of no evasion or administrative costs, but have very different effects in the real world.

Tax evasion is widespread, always has been, and probably always will be. Variations in duty and honesty can explain some of the across-individual and, perhaps, across-country heterogeneity of evasion. But the stark differences in compliance rates across taxable items that line up closely with detection rates suggest strongly that deterrence is a powerful factor in evasion decisions. Given the current state of theory and evidence on tax evasion, it is not clear in what way or how much enforcement might be most efficiently increased. Although the normative theory of taxation has been extended to tax system instruments such as the intensity of enforcement, the empirical knowledge for operationalizing these rules is sparse.

Bibliography

IRS (Internal Revenue Service). 2006. Updated estimates of the TY 2001 individual income tax underreporting gap. Overview. 22 February. Washington, DC: Office of Research, Analysis, and Statistics, U.S. Department of the Treasury.

IRS Oversight Board. 2006. 2005 Taxpayer attitude survey. U.S. Department of the Treasury. Online. Available at http://www.ustreas.gov/irsob/releases/2006/02212006.pdf, accessed 28 June 2007.

Slemrod, J. 2007. Cheating ourselves. *Journal of Economic Perspectives* 21(1), 25–48.

Slemrod, J. and Bakija, J. 2004. *Taxing Ourselves: A Citizen's Guide to the Debate over Taxes*, 3rd edn. Cambridge, MA: MIT Press.

Slemrod, J. and Yitzhaki, S. 2002. Tax avoidance, evasion and administration. In *Handbook of Public Economics*, vol. 3, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

How to cite this article

Slemrod, Joel. "tax compliance and tax evasion." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000235>
doi:10.1057/9780230226203.1674

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tax expenditures

Daniel N. Shaviro

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Labelling certain provisions in the tax law as tax expenditures has been criticized for lacking an ‘agreed conceptual model’ for distinguishing between integral tax rules and interpolations reflecting spending rather than tax policy. However, the tax expenditure concept can be reformulated as relying on a distinction between (*a*) the distributional goals that might underlie the use of a tax base such as income or consumption, and (*b*) allocative goals such as encouraging particular activities or investments. Tax expenditure estimates could be prepared using both measures, including negative tax expenditures (that is, tax penalties) as well as positive ones.

Keywords

consumption tax base; consumption taxation; Haig–Simons income taxation; income tax base; tax expenditures; tax penalties; taxation of corporate profits; taxation of income

Article

The practice of labelling certain provisions in the tax law as ‘tax expenditures’ is widely attributed to Stanley Surrey, the longtime Harvard law professor and, from 1961 to 1969, Assistant Secretary of the Treasury for Tax Policy in the United States. Surrey introduced the term in a 1967 speech, in which he urged official measurement of the revenue cost of all tax expenditures, which he defined as ‘special’ benefits in the income tax law. Surrey argued that publication of a tax expenditure budget would encourage and facilitate treating special tax rules on a par with similarly motivated direct spending rules. This proposal had earlier antecedents, having been part of the federal budgetary process in Germany since at least 1959. In the United States, however, it proved more controversial, reflecting Surrey's use of it as a tool, not just of budgetary policy, but also in tax policy debate, where he was well-known as an advocate of progressive, comprehensive income taxation. In keeping with his tax policy views, Surrey, after leaving the Treasury, pressed the argument that tax expenditures should generally be eliminated from the US income tax, with any that served meritorious social goals being replaced by direct

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
appropriations. Surrey's advocacy may have encouraged some with different tax policy views to regard the tax expenditure budget as special pleading for his views, merely masquerading as anodyne budgetary reporting.

Concern that the tax expenditure budget unduly served Surrey's particular views became more widespread with the rise in tax policy circles, beginning in the 1970s, of support for replacing the US federal income tax with a consumption tax. Various tax expenditures from an income tax standpoint, such as the exclusion of interest from bonds issued by state or local governments, are correct from a consumption tax standpoint. If ‘tax expenditures’ should be eliminated presumptively, then using a normatively controversial income tax standpoint to define them could be viewed as unduly aiding those who favour moving the current ‘hybrid’ US system closer to the income tax pole rather than to the consumption tax pole.

In the face of these criticisms, Surrey arguably won the battle concerning tax expenditure analysis, but lost the war. The Congressional Budget and Impoundment Control Act of 1974 made tax expenditure estimates mandatory both in the President's annual budget and in certain reports by Congressional committees. These estimates generally are static, measuring the level of utilization of a given provision, rather than how much revenue would be raised by repealing it. The tax expenditure concept has remained intellectually controversial, however. Moreover, it has not noticeably discouraged the use of ‘special’ tax benefits, other than perhaps temporarily if it helped to inspire the landmark Tax Reform Act of 1986.

What is a tax expenditure?

In official US estimates, tax expenditures are defined as pro-taxpayer departures from a ‘normal’ income tax base. This is not uncommon, although practices vary around the world. The normal income tax base has a number of features that depart from theoretically pure Haig–Simons income taxation. For example, it features a realization requirement under which unrealized gains and losses have no immediate tax consequences, includes double taxation of corporate income, and makes no adjustments for inflation. It also treats as tax expenditures some items whose preferentiality is controversial – for example, the itemized deductions for medical expenses and for state and local income taxes paid.

Criticism of tax expenditure analysis has focused both on what some view as the arbitrariness of the normal income tax base in any of its currently used variants, and on the lack of any ‘agreed conceptual model’ (Bittker, 1969, p. 258) for identifying special tax benefits. Such a model is needed to support the view that a particular provision, although located in the tax code, is actually a spending rule.

A deeper problem is that ‘taxes’ and ‘spending’ cannot meaningfully be distinguished, even though the former involves cash flow to the government while the latter involves cash flow from the government. By way of illustration, consider US federal income taxation of Social Security benefits, which was introduced during the Reagan administration and increased during the Clinton administration. While both administrations classified the changes as benefit cuts, Republican critics of the Clinton proposal argued that it was a tax increase. Arguably, these critics were formally correct, in that the reduction in net benefits was accomplished via income tax payments. However, if exactly the same reduction in net benefits had been accomplished by reducing gross benefits (that is, the amounts paid out by the Social Security administration), it evidently would have been a ‘spending cut’. Tax expenditure analysis requires discerning a substance to distinction between taxes and spending that does not depend in this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
way on form, or else by definition everything in the tax code would be a tax rule.
Still, the intuition underlying tax expenditure analysis is hard to dismiss. Suppose, for example, that the US Congress decided to replace $1 billion of military spending with a $1 billion tax credit, offered to the same taxpayers who would have received the direct appropriation in exchange for the same goods or services. Nominally, this switch would lower both 'taxes' and 'spending' by $1 billion. In substance, however, little would have changed. Tax expenditure analysis would treat the credit as 'spending' through the tax code, thus preventing the change in form from being misperceived as a change in substance.

Distribution and allocation

The distinction that tax expenditure analysis draws between tax and spending rules can be restated in terms of Richard Musgrave's (1959, p. 5) conceptual division between the public sector's distribution and allocation functions. Apportioning the burden of paying for government through a measure of ability to pay, such as income or consumption, is conceptually a distributional enterprise. Thus, a rule within the income tax law, such as the hypothetical military suppliers' credit, that appears to serve primarily allocative purposes (furnishing goods and services for military use) can logically be viewed as extraneous to the distribution function, even if its placement in the tax code is desirable (for example, on administrative grounds). The same reasoning applies in reverse if a set of tax rules serving primarily allocative purposes includes provisions that appear to serve primarily distributional aims. Thus, suppose a Pigouvian pollution tax offered rebates to low-income polluters. One could extend this reasoning to cover clearly distinguishable allocative or distributional functions as well – for example, the inclusion of education subsidies, such as lower tax rates for pollution by schools, in a Pigouvian pollution tax.
In each of these cases, the extraneous provision could be termed a tax expenditure, albeit without any necessary implication that it should be eliminated or moved. The reason for this linguistic exercise might be to increase public understanding of the provision at issue, and in particular to prevent 'tax cuts' from being distinguished from 'spending increases' on purely formal grounds where their substance is identical.
While this restatement of the distinction that Surrey attempted to draw between taxes and spending can go a long way to rationalize tax expenditure analysis, it does not support all aspects of current practice. For example, in the USA child tax credits are classified as tax expenditures, but personal exemptions (deductions for dependents, including children) are not. Yet the two provisions have similar effects, and both could be viewed as relating to a distributional goal of having tax burdens depend on family size. Thus, neither the distinction between them nor the treatment of either as a tax expenditure is highly robust.

Income tax base versus consumption tax base

The distinction between distribution and allocation functions does not address the issue of how to handle distinctions between the income and consumption tax bases, as illustrated by the question of whether the US exclusion for state and local government bond interest is a tax expenditure. Here the answer would depend on whether the distribution branch was assumed to follow income or consumption tax norms.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Under the latter norm, the anomalous result would be taxing other interest income, rather than exempting municipal bond interest. From either distributional perspective, however, the distinction in the tax treatment for interest depending on who paid it is likely to seem anomalous, even if desirable for allocative reasons. One possible solution discussed in recent US government budgets is to prepare alternative tax expenditure listings, one from an income tax baseline and the other from a consumption tax baseline.

Administrative departures from a pure income tax or consumption tax

One reason for the practice of computing tax expenditures relative to a ‘normal’ income tax base, rather than Haig–Simons income, is the notion that the provisions being analysed are substitutes for direct spending. Thus, if the main reason for not taxing unrealized asset appreciation is administrative, the legislature is unlikely to be choosing between alternative implementations of the resulting allocative policy. However, if administrative considerations limit the departures from a given theoretical base that are treated as tax expenditures, a system with those departures may easily be confused with one that actually implemented the theoretical ideal. One possible response to this dilemma is to create a separate category in tax expenditure analysis for departures from a given theoretical base that appear to be primarily administratively motivated (Shaviro, 2004, p. 218).

Negative tax expenditures (tax penalties)

Under current practice, tax expenditure analysis is limited to measuring the static revenue effect of departures from a given baseline that favour the taxpayer. Departures that disfavour a taxpayer are ignored, rather than being treated as negative tax expenditures or tax penalties. However, the rationale for measuring departures in one direction arguably should apply symmetrically.
An example of an unmeasured tax penalty in the current US income tax is the disallowance of business expense deductions for bribes. However socially desirable the disallowance rule may be, it reflects a departure from simply measuring net income in cases where the bribe was economically motivated. The practical importance of measuring tax penalties would increase if income tax rules were being analysed from a consumption tax as well as an income tax baseline, since this would cause a variety of common income tax features, such as taxing particular kinds of interest income, to constitute tax penalties.

Bibliography

Bittker, B. 1969. Accounting for federal ‘tax subsidies’ in the national budget. *National Tax Journal* 22, 244–61.

Musgrave, R. 1959. *The Theory of Public Finance*. New York: McGraw Hill.

Shaviro, D. 2004. Rethinking tax expenditures and fiscal language. *Tax Law Review* 57, 187–232.

Surrey, S. 1973. *Pathways to Tax Reform*. Cambridge, MA: Harvard University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Surrey, S. and McDaniel, P. 1985. *Tax Expenditures*. Cambridge, MA: Harvard University Press.

United States Government. 2003. Tax expenditures. Ch. 6 in Fiscal Year 2003. *Budget of the United States Government*. Analytical Perspectives. Online. Available at http://www.gpoaccess.gov/usbudget/fy03/pdf/spec.pdf, accessed 9 February 2007.

How to cite this article

Shaviro, Daniel. "tax expenditures." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000240> doi:10.1057/9780230226203.1675

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tax havens

James R. Hines Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tax havens are low-tax jurisdictions that offer businesses and individuals opportunities for tax avoidance. The 45 major tax haven countries in the world today are small, affluent, and generally well governed. They attract disproportionate shares of world foreign direct investment, and, largely as a consequence, their economies have grown much more rapidly than the world as a whole since the 1980s. The effect of tax havens on economic welfare in high-tax countries is unclear, though the availability of tax havens appears to stimulate economic activity in nearby high-tax countries.

Keywords

foreign investment; tax avoidance; tax havens; transfer pricing

Article

Tax havens are low-tax jurisdictions that offer businesses and individuals opportunities for tax avoidance.
There are roughly 45 major tax havens in the world today. Examples include Andorra, Ireland, Luxembourg and Monaco in Europe, Hong Kong and Singapore in Asia, and the Cayman Islands, the Netherlands Antilles, and Panama in the Americas. These tax havens are generally small and affluent, in total comprising just 0.8 per cent of world population, though accounting for 2.3 per cent of world income (Hines, 2005). Low-tax jurisdictions are also common within countries, at various times taking the form of special economic zones in China, offshore possessions and local enterprise zones in the United States, and tax-favoured regions including eastern Germany, southern Italy, eastern Canada, and others. Tax havens are widely used by international investors; in 1999, 59 per cent of US multinational firms with significant foreign operations had affiliates in one or more tax havens (Desai, Foley and Hines, 2006b).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Tax haven experiences

Countries offer low tax rates in the belief that, by doing so, they attract greater investment and economic activity than would otherwise have been forthcoming. Countries with low tax rates permit investors to retain most of their locally earned pre-tax income; other considerations equal, therefore, countries with lower tax rates should be expected to offer a broader range of attractive opportunities, and therefore draw larger volumes of foreign investment, than countries with higher tax rates.

The possibility of using tax havens to facilitate avoidance of taxes that would otherwise be owed to governments of other countries adds to the attractiveness of tax haven investments. For individuals, who are taxed by their home governments on income earned in tax havens, tax avoidance typically entails wilful income misreporting. For businesses, tax avoidance can be accomplished by the use of financial arrangements, such as intrafirm lending, that locate taxable income in low-tax jurisdictions and tax deductions in high-tax jurisdictions. In addition, firms are often able to adjust the prices at which affiliates located in different countries sell goods and services to each other. Most governments require that firms use arm's-length prices, those that would be used by unrelated parties transacting at arm's length, for transactions between related parties, in principle thereby limiting the scope of tax-motivated transfer price adjustments. In practice, however, the indeterminacy of appropriate arm's length prices for many goods and services, particularly those that are intangible, or for which comparable unrelated transactions are difficult to find, leaves room for considerable discretion. As a result, transactions with tax haven affiliates can be used to reallocate income from high-tax locations to the tax haven affiliates themselves or else to other low-tax foreign locations. This, in turn, increases the appeal of locating investment in foreign tax havens.

As a result of these incentives, American firms exhibit unusual activity levels and income production in foreign tax havens (Hines, 2005). Of the property, plant and equipment held abroad by American firms in 1999, 8.4 per cent was located in tax havens, considerably more than would be predicted strictly on the basis of the sizes of tax haven economies. Employment abroad by American firms was likewise unusually concentrated in foreign tax havens, with 6.1 of total foreign employee compensation, and 5.7 per cent of total foreign employment, located in tax haven affiliates. American firms located 15.7 per cent of their gross foreign assets in the major tax havens in 1999; the major foreign tax havens accounted for 13.4 per cent of their total foreign sales, and a staggering 30 per cent of total foreign income in 1999. Much reported tax haven income consists of financial flows from other foreign affiliates that parent companies owned indirectly through their tax haven affiliates.

Tax haven countries have enjoyed very rapid economic growth rates that coincide with dramatic inflows of foreign investment. Tax havens averaged 3.3 per cent annual per capita real GDP growth from 1982 to 1999, whereas the world averaged just 1.4 per cent annual real per capita GDP growth over the same period. Controlling for country size, initial wealth, and other observable variables, does not change the conclusion that the period of globalization has been favourable for the economies of countries with very low tax rates (Hines, 2005).

The policy of offering foreigners very low tax rates is potentially costly to tax haven governments, if doing so reduces tax collections that might otherwise have been used to fund worthwhile government expenditures. It is far from clear, however, that tax haven countries face significant trade-offs of this nature. Governments have at their disposal many tax instruments, including personal income taxes, property taxes, consumption or sales taxes, excise taxes, and others, that can be used to finance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
expenditures. Furthermore, even very low rates of direct taxation of business investment may yield significant tax revenues if economic activity expands in response. In fact, the public sectors of tax haven countries are of comparable sizes to those of other countries, though there is evidence that they may be somewhat smaller than would have been predicted on the basis of their populations and affluence alone (Hines, 2005).

Characteristics of tax havens

Tax havens are small countries, commonly below one million in population, and are generally more affluent than other countries. In addition, tax havens score very well on cross-country measures of governance quality that include measures of voice and accountability, political stability, government effectiveness, rule of law, and control of corruption. Indeed, there are almost no poorly governed tax havens. Poorly governed countries, of which the world has many, almost never become tax havens (Dharmapala and Hines, 2006).

An important reason why better-governed countries are more likely than others to become tax havens is that the potential returns are greater: higher foreign investment flows, and the economic benefits that accompany them, are more likely to accompany tax reductions in well-governed countries than they are tax reductions in poorly governed countries. Evidence from the behaviour of American firms is consistent with this explanation, in that tax rate differences among well-governed countries are associated with much larger effects on US investment levels than are tax rate differences among poorly governed countries (Dharmapala and Hines, 2006).

Impact on other countries

Tax havens are viewed with alarm in parts of the high-tax world, where there are concerns that the availability of foreign tax haven locations may divert economic activity from countries with higher tax rates, and erode their tax bases. Alternatively, tax havens could encourage investment in other countries, if the ability to relocate taxable income into tax havens improves the desirability of investing in high-tax locations, or if low tax rates reduce the cost of goods and services that are inputs to production or sales in high-tax countries. In fact, the evidence indicates that foreign tax haven activity appears to stimulate activity in nearby high-tax countries, a one per cent greater likelihood of establishing a tax haven affiliate being associated with two-thirds of a per cent greater investment and sales in nearby non-haven countries (Desai, Foley and Hines, 2006a).

The empirical regularity that tax havens stimulate economic activity in high-tax countries does not resolve the impact of tax havens on the welfare of high-tax countries. Tax avoidance carries mixed implications for governments of nearby countries, since it may erode tax bases and therefore tax collections, implying that the greater economic activity associated with nearby tax havens might come at a high cost in terms of forgone government revenues. One possibility is that countries would prefer to subject mobile international companies to lower tax rates than they do other firms, but are prevented from doing so by political considerations or the practical difficulty of distinguishing multinational from domestic firms. In such a setting, countries might benefit from permitting multinational firms to obtain tax reductions by using affiliates in tax havens, thereby implicitly subjecting these mobile firms to lower

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
tax burdens than other taxpayers.

In 1998, the Organization for Economic Co-operation and Development (OECD) introduced its Harmful Tax Practices initiative, the purpose of which was to discourage OECD member countries and certain tax havens from pursuing policies that were thought to harm other countries by unfairly eroding tax bases. As part of this initiative, the OECD produced a List of Un-Cooperative Tax Havens, identifying countries that have not committed to sufficient exchange of information with tax authorities in other countries. The concern was that the absence of information exchange might impede the ability of OECD and other countries to tax their resident individuals and corporations on income or assets hidden in foreign tax havens. As a result of the OECD initiative, along with diplomatic and other actions of individual nations, many countries and jurisdictions outside the OECD have committed to improve the transparency of their tax systems and to facilitate information exchange. While there remain a few tax havens that have not made such commitments, the vast majority of the world's tax havens rely on low tax rates and other favourable tax provisions to attract investment, rather than using the prospect of local transactions that will not be reported.

See Also

- tax competition
- tax treaties
- taxation of foreign income
- transfer pricing

Bibliography

Desai, M.A., Foley, C.F. and Hines, J.R., Jr. 2006a. Do tax havens divert economic activity? *Economics Letters* 90, 219–24.

Desai, M.A., Foley, C.F. and Hines, J.R., Jr. 2006b. The demand for tax haven operations. *Journal of Public Economics* 90, 513–31.

Dharmapala, D. and Hines, J.R., Jr. 2006. Which countries become tax havens? Working Paper No. 12802. Cambridge, MA: NBER.

Hines, J.R., Jr. 2005. Do tax havens flourish? In *Tax Policy and the Economy*, vol. 19, ed. J.M. Poterba. Cambridge, MA: MIT Press.

Hines, J.R., Jr. and Rice, E.M. 1994. Fiscal paradise: Foreign tax havens and American business. *Quarterly Journal of Economics* 109, 149–82.

Rose, A.K. and Spiegel, M.M. 2007. Offshore financial centers: parasites or symbionts? *Economic Journal* 117, 1310–35.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Slemrod, J. and Wilson, J.D. 2006. Tax competition with parasitic tax havens. Working Paper No. 12225. Cambridge, MA: NBER.

How to cite this article

Hines, James R., Jr. "tax havens." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000248> doi:10.1057/9780230226203.1676

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tax incidence

Gilbert E. Metcalf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tax incidence is the study of who bears the burden of a tax. It distinguishes between statutory incidence (the legal requirement to remit a tax) and economic incidence (the change in real income or wealth resulting from a tax). Considerable advances have been made since the mid-1980s in our understanding of the burden of taxes in imperfectly competitive models as well as in intertemporal models. In particular, analysing lifetime tax burdens can give markedly different results for many taxes. Increases in computing power and the availability of large-scale data-sets have also enriched our understanding of tax incidence.

Keywords

ad valorem taxes; commodity taxes; consumption tax; excise taxes; factor taxes; imperfect competition; lifetime income; lump-sum taxes; progressive and regressive taxation; property taxation; statutory and economic tax incidence; tax incidence; taxation of corporate profits

Article

Tax incidence is the study of who bears the economic burden of a tax. Broadly put, it is the positive analysis of the impact of taxes on the distribution of welfare within a society. It begins with the very basic insight that the person who has the legal obligation to pay a tax may not be the person whose welfare is reduced by the tax. The statutory incidence of a tax refers to the distribution of tax payments based on the legal obligation to remit taxes to the government. Thus, for example, the statutory burden of the payroll tax in the United States is shared equally between employers and employees. Economists, quite rightly, focus on the economic incidence, which measures the changes in economic welfare in society arising from a tax. The standard view of the economic burden of the payroll tax in the United States is that it is borne entirely by employees.
Economic incidence differs from statutory incidence because of changes in behaviour and consequent

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
changes in equilibrium prices. Consumers buy less of a taxed product, so firms produce less and buy fewer inputs – which changes the net price of each input. Thus the job of the incidence analyst is to determine how those other prices change, and how those changes affect different kinds of individuals.

The distributional impact of a tax (or system of taxes) depends in part on how the question is framed. An absolute incidence analysis considers the burden of a change in taxes without regard to the use of proceeds. A differential incidence analysis carries out a revenue-neutral change in tax by raising one tax while lowering another. Typically, a lump-sum tax is changed to effect revenue neutrality. A balanced budget incidence analysis considers the burden of a change in taxes along with an equivalent change in spending. In his classic analysis of the US tax system, Pechman (1985) carried out a differential incidence analysis and concluded that the total system of taxes in the United States was broadly proportional. Taking into account government transfers financed by taxes, on the other hand, Browning and Johnson (1979) argued that the US tax system was progressive.

In addition to framing the incidence question precisely, incidence results can depend on the time frame for analysis. Pechman's analysis ranks households by their annual income. It is well known that annual income can be a poor proxy for measuring the well-being and consumption potential of a household, because of measurement error and lifetime income considerations. Lifetime income considerations are particularly important for assessing the distributional implications of a consumption tax, since consumption to annual income ratios are very high in the lowest annual-income deciles. Fullerton and Rogers (1993) replicate the Pechman analysis using a lifetime income framework, and conclude that the overall incidence of the US tax system is similar to that obtained in Pechman's annual income framework, though the forces driving incidence results differ somewhat.

In a perfectly competitive partial equilibrium framework, the economic incidence of a tax is unaffected by which side of the market the tax is levied on. Thus the statutory requirement to share the payroll tax in the United States equally between employer and employee has no bearing on the ultimate incidence of the tax. Second, the economic burden of a tax is borne more heavily by the side of the market that is less elastic (in absolute value). Thus, the share of the payroll tax borne by the employee is, to a first-order approximation, equal to $\varepsilon_D / (\varepsilon_S + \varepsilon_D)$ where ε_S (ε_D) is the labour supply (demand) elasticity.

This burden share formula suggests that no more than 100 per cent of the tax can be shifted to a party. In an imperfectly competitive market, commodity tax overshifting can occur (in the sense that the consumer price rises by more than the tax rate). Moreover, *ad valorem* and excise taxes, which have equivalent burden impacts in a competitive market when set to collect the same revenue, now can have different burden impacts. Delipalla and Keen (1992) show that in markets with oligopoly supply *ad valorem* taxes are less likely to lead to overshifting than excise taxes. Once one allows for imperfect competition, many counter-intuitive results can obtain, including a commodity tax *reducing* the consumer price (for example, Cremer and Thisse, 1994). Fullerton and Metcalf (2002) develop the analysis of tax incidence under imperfect competition further and provide some hypothetical results.

Harberger (1962) is the progenitor of the modern field of general equilibrium incidence analysis. In addition to providing a framework for analysing the corporate income tax, Harberger's approach can be used to analyse a wide array of taxes. He models the corporate income tax as a partial factor tax, that is, a tax on the use of one factor in one sector. The tax thus affects relative factor prices and relative output prices. Harberger concludes that capital is likely to bear approximately the full burden of the corporate income tax. Capital mobility means that the burden is on all capital, not just corporate capital.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Harberger's analysis assumed a closed economy. In a small open economy with international capital mobility, corporate tax drives capital abroad so that domestic savers earn the same net return as before the tax is imposed. This drives down the domestic capital stock, and thus the domestic wage rate, and the burden of the tax falls on labour (as an immobile factor). While the immobile local factor bears a burden from the tax, Bradford (1978) shows that worldwide capital in the aggregate suffers a loss exactly offset by gains to immobile factors in the rest of the world, resulting from the outflow of capital from the country imposing the tax. In contrast, Gravelle and Smetters (2001) argue that imperfect substitutability of domestic and foreign products can limit or even eliminate the incidence borne by labour, even in an open economy model. They find that the tax is borne by domestic capital, as in the original Harberger model.

While Harberger's analysis (and subsequent work) showed the importance of general equilibrium effects, it lacked a fully dynamic characterization of savings and investment, channels through which important burden shifting could occur. Feldstein (1974), for example, argues that much (if not all) of the burden of a tax on capital income is shifted to workers in the form of lower wages as a result of decreased investment reducing the capital–labour ratio.

Once investment is considered, the incidence of a tax in a dynamic model can also be affected by the distinction between old and new capital. Old capital is capital in place prior to a tax change. For example, Auerbach and Kotlikoff (1987) show that a consumption tax and a wage tax – two approaches to exempting capital income from taxation – differ only in their tax treatment of old capital. In the absence of transition rules, the former subjects old capital to a lump-sum tax, while the latter does not. In addition to distributional implications, the presence of old capital complicates the attribution of economic incidence. Consider a new property tax that has been in place for many years in a community. Carrying out an incidence analysis today, we might allocate the burden of the tax to current owners based on their property values. This approach would be consistent with the ‘old’ view of property taxation (see Fullerton and Metcalf, 2002, for more on the incidence of the property tax). But with capitalization effects the tax burden should properly be allocated to the property tax owners at the time of the enactment of the tax: more precisely, it would be allocated to the owners at the time that potential buyers and sellers of property in the community become aware that the tax would be enacted. Without offsetting benefits from the property tax revenues, potential homeowners will be willing to pay less for housing. In equilibrium, housing values would fall by the present discounted value of the stream of future tax payments at the time of enactment, and the owners at that time would bear the entire burden of the tax.

To return to the corporate income tax, an increase in the tax rate generates lump-sum taxes on previously installed capital through capitalization effects. As Auerbach (2005) emphasizes, the tax treatment of corporate capital is sufficiently complicated to ensure that assigning its burden is a hazardous exercise, but both in the short and the long run it is probably the case that some portion of the tax falls specifically on shareholders due to the tax on old capital, among other factors.

Careful tax incidence analysis is essential to understanding the distributional implications of a country's tax system. The field of incidence analysis has progressed dramatically since the mid-1980s, as new research has yielded fresh insights into the burden of taxes in imperfectly competitive models and in intertemporal models. The increase in computing power and the availability of large-scale data-sets have also enriched our understanding of tax incidence. Despite all the advances that have occurred, the topic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of tax incidence will probably continue to be an area of productive research, yielding further insights in the years to come.

See Also

- income taxation and optimal policies
- progressive and regressive taxation
- public finance
- redistribution of income and wealth
- taxation of corporate profits

Bibliography

Auerbach, A. 2005. Who bears the corporate tax? A review of what we know. Working Paper No. 11686. Cambridge, MA: NBER.

Auerbach, A.J. and Kotlikoff, L.J. 1987. *Dynamic Fiscal Policy*. New York: Cambridge University Press.

Bradford, D.F. 1978. Factor prices may be constant but factor returns are not. *Economics Letters* 1, 199–203.

Browning, E.K. and Johnson, W.R. 1979. *The Distribution of the Tax Burden*. Washington, DC: American Enterprise Institute.

Cremer, H. and Thisse, J.F. 1994. Commodity taxation in a differentiated oligopoly. *International Economic Review* 35, 613–33.

Delipalla, S. and Keen, M. 1992. The comparison between ad valorem and specific taxation under imperfect competition. *Journal of Public Economics* 49, 351–67.

Feldstein, M. 1974. Incidence of a capital income tax in a growing economy with variable savings rates. *Review of Economic Studies* 41, 505–13.

Fullerton, D. and Metcalf, G.E. 2002. Tax Incidence. In *Handbook of Public Economics*, vol. 4, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Fullerton, D. and Rogers, D.L. 1993. *Who Bears the Lifetime Tax Burden*? Washington, DC: Brookings Institution.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gravelle, J.G. and Smetters, K. 2001. Who bears the burden of the corporate tax in the open economy? Working Paper No. 8280. Cambridge, MA: NBER.

Harberger, A.C. 1962. The incidence of the corporation income tax. *Journal of Political Economy* 70, 215–40.

Pechman, J. 1985. *Who Paid the Taxes: 1966–85*? Washington, DC: Brookings Institution Press.

How to cite this article

Metcalf, Gilbert E. "tax incidence." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000019> doi:10.1057/9780230226203.1677

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tax shelters

David A. Weisbach
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tax shelters take advantage of unintentional gaps in the tax base often caused by subtle mismatches in complex tax rules. The optimal line between allowable tax planning and illegitimate tax shelters depends on the cost of closing these gaps compared with the revenue raised, relative to the efficiency costs of other sources of funds. The definition of illegitimate tax shelters, therefore, depends on parameters such as the tax base and rate structure as well as the expected taxpayer response to different possible definitions.

Keywords

elasticity of taxable income; tax avoidance; tax base; tax compliance; tax evasion; tax rules; tax shelters

Article

The term ‘tax shelters’ generally refers to any tax reducing activity other than outright evasion or traditionally modelled responses to taxation, such as changes in labour supply or savings. The term sometimes includes investing in explicitly tax-favoured assets such as homes, life insurance or tax exempt bonds. At other times, however, the term is used to mean only tax-reducing activities inconsistent with the intent of the tax law, in which case it may not include these activities. The concept is closely related to, but usually thought to be narrower than, the notion of tax avoidance.

To define the term more precisely has proven to be impossible. The US Treasury Department (1999) observed that tax shelters come in the ‘guises of Proteus’ and argued that no single definition was appropriate. The tax law itself defines shelters for purposes of certain penalties as a plan or arrangement, a significant purpose of which is the avoidance or evasion of federal income tax, a definition sufficiently broad as to be almost meaningless. Rather than attempt to define shelters, Treasury (1999) has listed factors common to many shelter transactions, including (1) lack of economic substance; (2) inconsistent financial and accounting treatment; (3) presence of tax-indifferent parties; (4) complexity; (5) unnecessary steps or novel investments; (6) promotion or marketing; (7) confidentiality; (8) high

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
transaction costs; and (9) risk reduction arrangements. The Joint Committee on Taxation (1999) took a similar approach, recommending the use of tax shelter indicators or factors for purposes of triggering enhanced disclosure requirements or penalties. Other investigations have defined tax shelters as complex transactions, marketed by sophisticated professionals, used by corporations or wealthy individuals to obtain significant tax benefits in a manner never intended by the tax law (Levin, 2003; GAO, 2003). Perhaps the most pithy definition, by Michael Graetz, is that a tax shelter is ‘a deal done by very smart people that absent tax considerations, would be very stupid’ (Herman, 1999).

The definition of tax shelters matters for two related reasons. First, it points to behavioural responses to taxation that are left out of the usual analysis of labour/leisure or investment distortions and that are also not included in many analyses of tax evasion. Tax evasion, for example, is usually modelled as a report/non-report decision that imposes risk on the individual but otherwise has no direct affect on behaviour. Tax shelters, in contrast to evasion, normally involve actual although subtle changes to behaviour, such as leasing rather than owning, using a different organizational form, or using hybrid financing instruments. A less than optimal use of legal forms can produce social losses beyond merely the risk of audit. Analyses of investment distortions normally compute marginal effective tax rates on different activities. Tax shelters can significantly reduce effective tax rates on certain activities, which means that the distortions might be substantially larger than otherwise thought. Second, the definition has a set of legal consequences, such as disallowance of tax benefits, penalties, disclosure rules, and additional audits.

The appropriate legal consequences of the definition of tax shelters and an analysis of the economic effects of sheltering need to be tied together. In particular, the scope of activities that should be subject to various policies must be determined based on an analysis of the consequences of such policies, not a definition. There is no clear line between various tax-reducing activities, ranging from working less, being paid in tax-free fringe benefits, investing in tax-favoured assets, entering into traditional shelters, and false reporting. The treatment of a class of these activities as tax shelters, others as criminal evasion, and others as allowable should depend on which activities are optimally subject to a given set of policy instruments.

This approach means that, to define shelters, we need a general theory of tax instruments, including the tax base, the penalty structure, the drafting of legal rules, reporting regimes, and the audit rate. As emphasized by Feldstein (1995; 1999), given some tax base and set of audit, penalty, and similar policies, the private marginal cost of all tax reducing activities will be equal. Therefore, the elasticity of taxable income can be used as a summary measure of the efficiency of the tax system. Slemrod and Kopczuk (2002) observe that the elasticity of taxable income is, in part, a policy variable rather than a preference because policymakers can control the size of the tax base, auditing mechanisms, penalties, and other variables that affect opportunities to reduce taxable income. First order conditions for an arbitrary tax instrument (assuming an optimal linear income tax) set the marginal administrative cost of the instrument equal to sum of marginal (indirect) utility from the change in the instrument and the marginal revenue. Marginal revenue is made up of two components: revenue directly from the change in taxation of the item at issue with the elasticity held constant, and revenue from the change in the elasticity of taxable income.

Viewed in this way, tax shelters are similar to any other gap in the tax base, and appropriate responses to shelters (and definition of shelters) depend on relative cost of obtaining that source of funds. Although

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
this general formulation does not tell us anything about the particular definition of tax shelters or which particular mix of instruments is optimal, it does focus attention on the relevant factors. For example, it seems clear that there should be substantial sanctions for fraud because any other rule would produce a very high elasticity of taxable income – without such sanctions, any increase in the tax rate, starting from zero, would produce substantial reductions in reported income. A similar conclusion holds for many shelters. If they became well known and inexpensive, the elasticity of taxable income becomes unduly high. Similarly, an important effect of imposing a tax on a shelter is reducing the elasticity of taxable income rather than raising revenue directly from the tax on the shelter activity. Because few would engage in the shelter activity without the tax benefits, any revenue from a direct tax on that activity is likely to be small. Most important, the elasticity measure emphasizes that the optimal definition of tax shelters depends on what other instruments are in use, such as the scope of the tax base, the rates, and the audit and evasion rules. For example, *ceteris paribus*, the higher the tax rates, the broader the optimal definition of shelters is likely to be.

Another approach, more grounded in law than in economics, is to focus on the drafting of tax rules. The primary cause of shelters, in this view, is the imperfect interactions of statutory rules. Treasury (1999) referred to these interactions as discontinuities. Given limited resources, drafters of tax rules can cover only general cases. Taxpayers have a private incentive to find unusual interactions of otherwise reasonable rules and structure transactions to take advantage of them. Weisbach (1999) argued that the solution to this problem, long followed by US law, involves general rules for common behaviours and ambiguous standards, so-called anti-abuse rules, that prevent intentional use of unintended, tax-reducing interactions in the general rules. This approach balances the uncertainty of the ambiguous standards (and the potential principal–agent problems of a revenue-maximizing tax agency overusing the standards) with the benefit of tax rules that need to cover only general cases and, therefore, that can be simpler.

A related approach is to focus on the industrial organization of the tax shelter industry. The designing and implementation of tax shelters takes significant resources. While privately beneficial, most of this expenditure of resources has no social benefit. On the other hand, advice about compliance (as opposed to shelters) does have a social benefit. The regulatory problem is to distinguish these activities. Various approaches have been considered, such as limiting the use of contingency fee arrangements based on tax savings, requiring disclosure by advisors of client lists, and direct sanctions, including criminal penalties, for giving inappropriate tax advice.

Reliable estimates of the number or size of tax shelters are difficult to obtain because of their secrecy and because of the ambiguity in the definition of shelters. There is no tax shelter equivalent to the measurements of the tax gap (which is a measure of non-compliance). Shelters have long been considered a problem in the tax law, and anecdotal evidence of sheltering activity has frequently driven tax policy. Important Supreme Court decisions on shelters date back to the 1930s or earlier. In 1934, the Treasury Department attempted to prosecute the former Secretary of the Treasury, Andrew Mellon, for tax evasion. Although the grand jury refused to indict, Mellon was eventually ordered to pay $400,000 in back taxes for what might be considered shelters. As Secretary, Mellon had solicited from Internal Revenue Service ‘memorandum setting forth the various ways by which an individual may legally avoid tax’ (Brownlee, 1996). In 1969, an outgoing Secretary of the Treasury revealed that in 1967 no income taxes were paid on 155 tax returns with gross incomes of 200,000 dollars or more, including 21 of returns of millionaires (Zelizer, 1998). This revelation led immediately to a variety of tax law changes to prevent the use of shelters.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The late 1970s and early 1980s saw a significant rise in tax shelter activity. In 1980, the Commissioner of the Internal Revenue Service stated that ‘about 200,000 individual tax returns representing about 18,000 shelter schemes are now at various stages of the examinations’ process’ (Kurtz, 1980). In 1985, there were over 20,000 tax shelter cases pending in the tax court, and, as of 1982, these cases made up approximately one-third of the court's docket (Collinson, 1985). Limitations on tax shelter losses that were part of the Tax Reform Act of 1986 were estimated to raise almost $53 billion over the five-year revenue window. Birnbaum and Murray (1987) report that these provisions were central to the compromise that allowed the 1986 Tax Reform Act to pass.

Notwithstanding the tax shelter limitations enacted in the 1986 Act, tax shelter activity was thought to rebound significantly in the 1990s. A Senate Subcommittee investigation reported that a single major accounting firm had more than 500 tax shelter products in its inventory and that it sold these products aggressively to individuals and corporations (Levin, 2003). Some 19 lawyers or accountants associated with these shelters were indicted in the largest criminal tax case in US history. Graham and Tucker (2006) study 44 instances of tax shelters by examining reported cases (which should represent only a small fraction of actual shelters). They find that typical tax shelter deduction was more than one billion dollars per firm and that the median shelter produced a deduction sufficient to shield income of approximately nine per cent of asset value. Settlements from a single type of shelter, known as ‘Son of Boss’ sold largely to individuals produced more than $3.7 billion in additional taxes in 2005.

See Also

- excess burden of taxation
- tax compliance and tax evasion
- tax havens

Bibliography

Birnbaum, J.H. and Murray, A.S. 1987. *Showdown at Gucci Gulch*. New York: Random House.

Brownlee, W.E. 1996. *Federal Taxation in America: A Short History*. Cambridge: Cambridge University Press.

Collinson, D. 1985. New York State Bar Association Tax Section Committee on Practice and Procedure, ‘Managing the Tax Court Docket’, *Tax Notes Today* 85, 146–93 (24 July).

Desai, M. 2003. The divergence between book and tax income. *Tax Policy and the Economy* 17, 169–206.

Feldstein, M.S. 1995. The effect of marginal tax rates on taxable income: a panel study of the 1986 tax reform act. *Journal of Political Economy* 103, 551–72.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Feldstein, M.S. 1999. Tax avoidance and the deadweight loss of the income tax. *Review of Economics and Statistics* 81, 674–80.

GAO (General Accounting Office). 2003. Internal revenue service: challenges remain in combating abusive tax shelters. Testimony before the Committee on Finance. GAO-04-104T. Washington, DC: US Senate.

Graham, J.R. and Tucker, A. 2006. Tax shelters and corporate debt policy. *Journal of Financial Economics* 81, 563–94.

Herman, T. 1999. Tax report. *Wall Street Journal*, 10 February, p. A-1.

Joint Committee on Taxation. 1999. Study of present-law penalty and interest provisions as required by Section 3801 of the Internal Revenue Service Restructuring and Reform Act of 1998 (including provisions relating to corporate tax shelters). JCS 3-99. Washington, DC: US Government Printing Office.

Kurtz, J. 1980. Kurtz on ‘abusive tax shelters’. *Tax Notes* 10, 213.

Levin, C. 2003. *U.S. tax shelter industry: the role of accountants, lawyers, and financial professionals. Report prepared by the Minority Staff of the Permanent Subcommittee on Investigations*. Washington, DC: US Senate.

Manzon, G., Jr. and Plesko, G. 2002. The relation between financial and tax reporting measures of income. *Tax Law Review* 55, 175–214.

Slemrod, J. and Kopczuk, W. 2002. The optimal elasticity of taxable income. *Journal of Public Economics* 84, 91–112.

U.S. Treasury Department. 1999. *The Problem of Corporate Tax Shelters: Discussion, Analysis, and Legislative Proposals*. Washington, DC: U.S. Treasury Department.

Weisbach, D. 1999. Formalism in the tax law. *University of Chicago Law Review* 66, 860–86.

Zelizer, J.E. 1998. *Taxing America: Wilbur D. Mills, Congress, and the State, 1945–1975*. Cambridge: Cambridge University Press.

How to cite this article

Weisbach, David A. "tax shelters." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000237> doi:10.1057/9780230226203.1678

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tax treaties

Charles E. McLure, Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tax treaties coordinate how signatories tax transnational income flows to avoid double taxation and prevent fiscal evasion. Treaties affect the division of tax revenues and signal commitment to international ‘rules of the game’. Most of the 2,000 plus extant bilateral tax treaties are based on the OECD Model Tax Convention. Residence countries avoid double taxation of foreign-source business income by exempting it or providing credits for source-country taxes. Source-countries typically reduce withholding taxes on interest, dividends and royalties. Despite apparent differences, the economic effects of taxing worldwide income, with credits for foreign taxes, may resemble those of exempting it.

Keywords

arm's length prices; business income; capital export neutrality; capital import neutrality; capital taxation; characterization of income; division of tax revenues; double non-taxation; double taxation; electronic commerce; exchange of information; fiscal evasion; foreign-source income, taxation of; formulary apportionment; General Agreement on Tariffs and Trade; intangible assets; Mutual Agreement Procedures; non-discrimination; OECD Model Tax Convention; permanent establishment; portfolio investment; residence country; signalling; source country; tax competition; tax evasion; tax exemption; tax harmonization; tax havens; tax sparing; tax treaties; taxation of foreign income; transfer prices; treaty shopping; UN Model Tax Convention; withholding taxes

Article

The term ‘tax treaties’ is commonly used – as here – to describe bilateral treaties that coordinate how signatories apply their taxes on income and capital to transnational economic activity. (A few multilateral treaties addressing broader objectives, for example, that establishing the European Union, deal to a limited extent with these or other taxes, as do some treaties – notably the multilateral General Agreement on Tariffs and Trade – that deal primarily with other forms of taxation.) The primary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
objectives of tax treaties are avoidance of double taxation and prevention of fiscal evasion. Secondary objectives include the division of tax revenues between treaty partners and signalling that signatories will abide by international ‘rules of the game’. Avoidance of double non-taxation has recently received increased attention. While treaties generally regulate taxation of both individuals and legal entities, the latter are by far the more important and the focus of this article.

Over 2,000 bilateral tax treaties are currently in force. Because tax treaties require several years of negotiation, they are expected to remain in force for several decades, and their wording is rather general, to allow reinterpretation. The vast majority of tax treaties are based on the OECD Model Tax Convention (OECD, 2005), whose origins can be traced to the work of the League of Nations during the 1920s (see Graetz and O'Hear, 1997). The ‘Commentary’ that accompanies and interprets this Model Convention, often also called the OECD Model Tax Treaty, is frequently revised to deal with unforeseen issues (for instance, those involving electronic commerce discussed below). Developing countries generally prefer the United Nations (UN) Model Tax Convention, which is more favourable to their interests as source countries. Although they have historically had difficulty getting more powerful developed countries to accept its terms, this has changed recently (see Kosters, 2004). Some countries, including the United States, publish their own model treaties. Although based on the OECD Model, these models deal with special concerns or features of the country's tax system, such as the US concern with ‘treaty shopping’ considered below. The OECD website discusses many issues covered here.

Avoiding double taxation

Nations have the legal right to tax both income arising within their borders and the income of their residents, whatever its geographic source. In the absence of treaties, there is a risk of ‘double international taxation’ by the ‘source country’ where income arises and the ‘residence country’ of the taxpayer, even though domestic legislation may unilaterally provide relief from double taxation. Moreover, because of differences in definitions of residence or source, more than one country may impose either a residence- or a source-based tax. Double taxation impedes transnational transactions and investment flows.

Tax treaties commonly assign to source countries the primary right to tax business income resulting from direct investment, but to residence countries the primary right to tax other forms of income, including that from portfolio investment. Treaties generally provide one of two methods that residence countries can use to avoid double taxation of foreign-source business income, including dividends from subsidiaries: exemption of foreign-source income and credits for taxes paid to source countries. By comparison, they typically provide that source-country withholding taxes on interest, dividends, and royalties will be reduced reciprocally, sometimes to zero. The latter provisions have given rise to ‘treaty shopping’, the practice of establishing subsidiaries in a country solely for the purpose of benefiting from its treaties. The United States now insists on a ‘limitations of benefits’ provision that eliminates most treaty shopping.

Treaties specify rules for the characterization of income (for example, as business profits, dividends, interest, royalties, capital gains and service income) and the geographical ‘sourcing’ of each type of income. The former task is particularly challenging in cases such as income from the provision of computer software, which might reasonably be characterized as business profits from the sale of goods, royalties or service income.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Treaties limit source-country taxation of business profits to income earned by a permanent establishment (PE), indicated by the presence of a fixed place of business or a dependent agent in the country. This provision has important implications for the division of revenues between source and residence countries. Having its origin in a world of physical products, the definition of a PE, and especially its application to modern business models such as electronic commerce that may not require a physical presence in the source jurisdiction, has recently been the subject of considerable controversy.
Treaties generally provide that arm's length prices – prices that would prevail in transactions between unrelated parties – are to be used in valuing transactions (including financial transactions) between related parties. While determining transfer prices is relatively straightforward for some homogeneous commodities that are widely traded (such as oil and wheat), it is difficult – or even conceptually impossible – for many unique intangible assets that have no market outside a given multinational corporation. The OECD has issued guidelines for the determination of transfer prices, some of which rely on formulas, but has steadfastly refused to sanction formulary apportionment (OECD, 1995).
Countries may disagree over the proper transfer prices, despite treaty provisions for mutual agreement procedures intended to resolve these and other conflicts in interpreting and applying treaties (for example, the residence of a taxpayer). In such cases double taxation may occur.

Preventing fiscal evasion

To prevent fiscal evasion, tax treaties provide for the exchange of information between tax authorities. For example, if a person deposits funds in a foreign bank, but does not report the resulting interest income, the fiscal authorities of that country could report the interest to their counterparts in the taxpayer's country of residence. In fact, exchange of information has been less useful than this description suggests. First, the fiscal authorities of the country of residence must identify the suspected tax cheat (‘no fishing’) and cannot require provision of information not collected in the normal course of operations or in violation of domestic law. Tax evaders can utilize legal entities whose identities are not known to the tax authorities of their country of residence to make investments.
Tax havens – low-tax jurisdictions that have bank secrecy and related laws that allow ownership of assets to be concealed – pose a particularly important threat to tax compliance. Since tax havens generally do not participate in treaty networks and resist exchange of information, it has been relatively simple to evade taxes by channelling investments to or through them. During the 1990s the OECD undertook a project on ‘harmful tax competition’ that included pressure on tax havens to exchange information.

Signalling

Some developing countries and countries in transition from socialism deviate from widely recognized standards for taxing business profits, for example by not allowing deductions for all business expenses or enacting tax laws that favour domestic taxpayers over foreigners. The existence of a tax treaty provides a signal to potential investors that the signatories will play by the internationally accepted ‘rules of the game’, including taxation of *net* business income and non-discrimination, and assurance that their country of residence will defend them in the event of deviations from those rules. (A taxpayer

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
may ordinarily appeal to the ‘competent authority’ of its country of residence to insure that both countries are abiding by the terms of the treaty.) For example, by concluding a tax treaty with Canada (its first), Mexico demonstrated readiness to join the OECD, the North American Free Trade Association and the World Trade Organization. Also, non-discrimination rules prevent source countries from levying higher taxes on non-resident investors than on local investors, in order to benefit from the availability of foreign tax credits offered by the residence country.

Multilateral treaties

Some advocate a World Tax Organization analogous to the World Trade Organization. Such a body might foster harmonization of income tax bases, far-reaching cooperation among tax administrations, or even equalization of tax rates. Harmonization of tax bases – and especially of tax rates – is unlikely to occur soon, if ever, because, inter alia, defining the tax base and setting tax rates are important aspects of sovereignty, there is no benchmark for the ideal income tax base, and the laws of nations exhibit numerous complex differences.

Economic issues

At first glance the two methods of preventing double taxation of business profits have quite different economic consequences. If all residence countries employed the exemption method, all income from foreign investment in a particular source country would bear only the tax of the source country and *capital import neutrality* would prevail. On the other hand, if all residence countries taxed all foreign-source income currently and allowed credits for all source-country taxes, all investment made by residents of a particular country would bear the same tax and *capital export neutrality* would prevail. There are, however, at least three reasons why the economic effects of taxing worldwide income, with credits for foreign taxes, may resemble those of an exemption system.

First, the parent's tax on most income of foreign subsidiaries is deferred until the income is distributed. (The primary exception occurs when the residence country taxes the undistributed income of certain controlled foreign corporations currently, commonly as an anti-abuse technique.) Thus systems that ostensibly tax the worldwide income of residents and systems that exempt foreign-source income produce similar results.

Second, foreign tax credits are commonly limited to the average tax rate paid on both foreign and domestic-source income. When ‘excess foreign tax credits’ exist, not all taxes levied by high-tax nations can be credited, again producing results similar to those of exemption systems.

Third, the tax treaties of some nations (but not the United States) allow credits for taxes that developing countries choose not to collect, because of tax incentives such as holidays and investment incentives. Such ‘tax sparing’ implies that tax incentives benefit investors, as under an exemption system, rather than being neutralized by higher taxes in the residence country.

See Also

- tax havens

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- taxation of foreign income
- transfer pricing
- World Trade Organization

Bibliography

Graetz, M.J. and O'Hear, M.M. 1997. The ‘original intent’ of U.S. international taxation. *Duke Law Review* 46, 1021–109.

Kosters, B. 2004. The United Nations Model Tax Convention and its recent developments. *Asia-Pacific Tax Bulletin* 10, 4–11.

OECD (Organisation for Economic Co-operation and Development). 1995. *Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations*. Paris: OECD.

OECD. 2005. *Model Tax Convention on Income and on Capital*. Paris: OECD. Online. Available at http://www.oecd.org/dataoecd/52/34/1914467.pdf, accessed 17 February 2007.

OECD. Online. Available at http://www.oecd.org, accessed 17 February 2007.

Vann, R.J. 1998. International aspects of income tax. In *Tax Law and Drafting*, vol. 2, ed. V. Thuronyi. Washington, DC: International Monetary Fund.

How to cite this article

McLure, Charles E., Jr. "tax treaties." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000225> doi:10.1057/9780230226203.1679

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

taxation and poverty

John Karl Scholz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Low-income US households typically pay Social Security payroll taxes, state and local sales taxes, and possibly, state income taxes. Federal income taxes in the United States and the United Kingdom, among other countries, provide tax subsidies to low-income working families, particularly those with children. These ‘in-work benefits’ raise the incomes of poor families, modestly increase employment, and have negligible effects on hours of work. The design and effectiveness of these provisions depend on details of the tax system, such as the unit of taxation, the degree to which people file tax returns, and the ability of the tax authority to enforce tax rules.

Keywords

Earned Income Tax Credit (USA); hours of work; implicit tax rates; negative income tax; payroll taxes; Social Security in the United States; tax compliance; taxation and poverty; value-added tax; Working Tax Credit (UK)

Article

Tax systems around the world can have substantial effects on the income available to families with low-skill workers. Key factors affecting the tax burdens of poor families include the set of taxes used in the economy, the specific exemptions and deductions contained in the system, and the special provisions targeting low-income households. To discuss these issues, this article focuses primarily on the experiences of the United States, but much of the discussion applies to tax systems in other developed and developing countries. For a broader treatment of taxation in developing economies, see, for example, Burgess and Stern (1993) and Gordon and Li (2005).

The primary taxes borne by low-income US households are the Social Security payroll tax, state and local sales taxes, and in some states, state income taxes. Roughly 41 per cent of US families pay more in payroll taxes than individual income taxes. If we (appropriately) assume the employer's share of payroll

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
taxes are borne by workers, payroll taxes exceed income taxes for 71 per cent of US families. For most low-earning individuals, the net present value of Social Security benefits still exceeds the present discounted value of taxes paid (Liebman, 2002), but these families are much more likely than others to be intertemporally credit constrained, so, if the payroll tax is fully borne by workers, the 14.2 per cent combined employer–employee tax results in a substantial reduction in after-tax resources available for consumption. (14.2 per cent is the sum of the employer and employee shares of payroll taxes, which equals 15.3 per cent, divided by market earnings increased by the employer's tax share – 1.0765 – with the idea that, without the payroll tax, employers would increase wages by their share of the tax.) The perceived regressivity of the Social Security payroll tax was one factor leading to the adoption of the Earned Income Tax Credit in the mid-1970s.

Sales taxes and their international cousins, value-added taxes (VAT), also raise concerns among policymakers that they impose inappropriate burdens on low-income households. Consequently, these taxes frequently exempt items such as food, clothing, and medicine that are thought to typically compose larger shares of poor families' budgets than is the case for other families. Zero-rating (excluding) items raises a fundamental issue in taxation. Should tax systems be designed to raise the revenues necessary for the operation of government in the most efficient way possible, leaving expenditure policy to address distributional concerns, or should taxes be designed to address equity issues directly? Exempting (or zero-rating) items in a VAT or consumption tax reduces efficiency (for example, see Ballard, Scholz and Shoven, 1987). Whether policymakers deem the exemptions as being necessary depends on political considerations and the strength of other available institutions to redistribute resources to poor families.

The federal individual income tax is conspicuously absent from my list of taxes reducing the incomes of poor families. Until around 1974, the federal income tax imposed positive average and marginal tax rates on families with incomes at the US poverty line, so, along with payroll and sales taxes, income taxes (at both the federal and, in some circumstances, the state level) reduced the incomes of low-income working families. In the absence of other tax provisions targeting low-income families or individuals, the threshold at which families began to pay income taxes was determined largely by the size of the standard deduction and exemptions, and whether these provisions were indexed for inflation. In 1974 the difference in average tax rates, combining income and payroll taxes, between a one-adult, two-child family with income at the poverty line and a two-adult, two-child family with income three times the poverty line was 9.2 percentage points, or the difference between 13.2 per cent and 22.4 per cent. By 2005 the difference was 36.9 percentage points, or the difference between –15.3 per cent and 21.6 per cent. (These calculations are made with the NBER's TAXSIM model: see Feenberg and Coutts, 1993 for a discussion of TAXSIM.)

By far the most important factor affecting the tax treatment of low-income families in the United States since 1977 has been the development and expansion of tax provisions targeted to low-income taxpayers that are 'refundable' – the Treasury pays out the value of the credit regardless of whether the taxpayer otherwise has positive tax liability. The most important of these is the Earned Income Tax Credit, though in recent years a portion of the child credit has also been made refundable. Refundable credits can result in negative average tax rates for working poor families with children.

The antecedent for current tax provisions targeting low-income families is negative income tax (see Moffitt, 2004, for a nice discussion). The negative income tax (NIT) was to provide a basic income guarantee that would be clawed back as earnings increase. In the mid-1970s US policymakers came

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
close to enacting a NIT, and its labour market and family formation effects were studied extensively in a series of closely watched, widely publicized social experiments (see, Robins, 1985; Cain and Wissoker, 1990 for further details).

The United States implemented an Earned Income Tax Credit (EITC) in 1975. The EITC provides a subsidy to earnings up to a specific income threshold. For example, in 2004 the EITC gave a 40 per cent earnings subsidy up to 10,750 dollars to a family with two or more children. Taxpayers with earnings between 10,750 and 14,040 dollars received the maximum credit of 4,300 dollars. The maximum credit for families with one child is 2,604 dollars; for childless workers it is 390 dollars. The credit was reduced by 21.06 per cent of earnings between 14,040 and 34,458 dollars. Hence, there are three distinct ranges of the EITC: the subsidy, flat and phase-out ranges of the credit.

The political appeal of the EITC, and similar programmes in other countries such as the Working Tax Credit in the United Kingdom and an EITC-like earnings subsidy to be implemented in South Korea, rests on at least two factors. First, earnings subsidies like the EITC are thought to encourage work and they are sometimes justified as part of a set of policies to ‘make work pay’. There is considerable evidence that this perception is accurate: the EITC has positive employment effects so, in contrast to many alternative ways of redistributing income from higher- to lower-income families, the EITC does not substantially harm labour market incentives. Second, by adding the EITC to an existing individual income tax, implementation costs are relatively low, particularly compared with programmes that require their own bureaucracy.

The static labour supply model implies the EITC will have an unambiguous, positive incentive effect on employment. The empirical evidence is consistent with these incentive effects: the EITC has a statistically significant and large effect on labour force participation of single women with children. Grogger (2003), for example, concludes that the EITC ‘may be the single most important policy measure for explaining the decrease in welfare and the rise in work and earnings among female-headed families in recent years’ (2003, p. 408). For more on the EITC, see Dickert, Houser and Scholz (1995); Eissa and Liebman (1996); Keane and Moffitt (1998); Ellwood (2000); Meyer and Rosenbaum (2000, 2001); and Hotz, Mullin and Scholz (2005). Eissa and Hoynes (2004) focus on the employment and hours decisions of secondary workers in married families and find small, negative effects of the credit on work. Hotz and Scholz (2003) survey EITC research.

The static labour supply model implies an ambiguous incentive effect of the EITC on hours in the phase-in range of the credit and unambiguous negative incentive effects on hours in the flat and phase-out ranges. Studies estimating the effects of the EITC on hours of work for working households find no bunching of taxpayers at the beginning and end of the phase-out range, as might be expected if the EITC significantly affects hours and taxpayers are cognizant of the discontinuities in implied marginal tax rates generated by the credit (Liebman, 1997). It is not surprising that negative effects on hours for people already in the labour market are small, since the precise relationship between the EITC and hours worked is likely to be poorly understood by most taxpayers. Most EITC recipients pay a third party to prepare their tax returns, and it is difficult to infer the implicit tax rates embodied in the credit from the look-up table that accompanies the EITC instructions. This confusion is less likely to mitigate positive participation effects, since, for these to be operative, taxpayers need only to understand that there is some tax-related bonus to work. Abundant anecdotal evidence indicates that taxpayers have this understanding (see, for example, DeParle, 1999).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The UK Working Tax Credit has an interesting design feature when compared with the EITC. Instead of phasing in, it imposes an hours threshold that triggers eligibility, thereby increasing the number of households receiving positive employment and hours incentives in relation to a credit on the first dollar of earnings. All households working fewer than 16 hours will see an increase in the after-tax return to work (and, since they do not receive any credit if they have fewer than 16 hours of work, there is no incentive to 'buy' more leisure). Hours limits impose a potentially significant additional administrative burden – because hours information is typically not required to implement an income tax – so their desirable labour market incentive effects must be balanced against the additional costs that arise from administering the hours requirement. Blundell and Hoynes (2004) find the EITC seems to have a larger effect on employment than the WTC predecessor, even though average EITC benefits are somewhat smaller. This may in part be because the incentive effects of in-work benefits in the United Kingdom are dulled by integrations with the rest of the tax and benefit system.

The unit of taxation in most countries around the world is the individual, not the family as is the case in the United States. Most policymakers (including those in the United Kingdom), however, believe that it is essential to target tax benefits on the basis of family income, since it is widely believed that families pool resources when making economic decisions. UK tax authorities meet this goal by having taxpayers claim eligibility by submitting a form to the tax authorities during the year, while the claim is recalculated at the end of the year based on *family* income. The UK experience shows that it is possible to have a credit with family-based eligibility in a tax system where individuals are the unit of taxation.

Less is known about the effects of the EITC on other aspects of behaviour. Dickert-Conlin and Houser (2002) and Eissa and Hoynes (2004) provide some evidence that the EITC encourages the existence of female-headed families. Heckman, Lochner and Cossa (2002) examine the effects of the EITC on skill formation. While they emphasize that much more needs to be done, they reach a tentative conclusion that the EITC has little impact on average skill levels in the economy. The EITC appears to reach those who are eligible – participation rates among eligible taxpayers is high (Scholz, 1994). Lastly, the EITC also suffers from high rates of non-compliance (Internal Revenue Service, 2002): many taxpayers who are not eligible end up claiming and receiving the credit. There is probably a trade-off between a policy with low administrative costs, like the EITC, and high rates of non-compliance.

See Also

- anti-poverty programmes in the United States
- low-income housing policy
- nutrition and public policy in advanced economies
- poverty
- poverty alleviation programmes
- welfare state

Bibliography

Ballard, C.L., Scholz, J.K. and Shoven, J.B. 1987. The value-added tax: a general equilibrium look at its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
efficiency and incidence. In *The Effects of Taxation on Capital Accumulation*, ed. M. Feldstein. Chicago: University of Chicago Press.

Blundell, R. and Hoynes, H. 2004. Has ‘in-work’ benefit reform helped the labour market? In *Seeking a Premier Economy: The Economic Effects of the British Economic Reforms, 1980–2000*, ed. R. Blundell, D. Card and R. Freeman. Chicago: NBER and University of Chicago Press.

Burgess, R. and Stern, N. 1993. Taxation and development. *Journal of Economic Literature* 31, 762–830.

Cain, G.G. and Wissoker, D.A. 1990. A reanalysis of marital stability in the Seattle-Denver income-maintenance experiment. *American Journal of Sociology* 95, 1235–69.

DeParle, J. 1999. Once a forlorn avenue, tax preparers now flourish. *New York Times*, 21 March.

Dickert, S., Houser, S. and Scholz, J.K. 1995. The earned income tax credit and transfer programs: a study of labor market and program participation. In *Tax Policy and the Economy*, vol. 9, ed. J.M. Poterba. Cambridge, MA: NBER and the MIT Press.

Dickert-Conlin, S. and Houser, S. 2002. EITC and marriage. *National Tax Journal* 55, 25–39.

Eissa, N. and Liebman, J.B. 1996. Labor supply response to the earned income tax credit. *Quarterly Journal of Economics* 111, 605–37.

Eissa, N. and Hoynes, H.W. 2004. Taxes and the labor market participation of married couples: the earned income tax credit. *Journal of Public Economics* 88, 1931–58.

Ellwood, D.T. 2000. The impact of the earned income tax credit and social policy reforms on work, marriage, and living arrangements. *National Tax Journal* 53, 1063–105.

Feenberg, D.R. and Coutts, E. 1993. An introduction to the TAXSIM model. *Journal of Policy Analysis and Management* 12, 189–94 Online. Available at http://0-www.nber.org.library.lemoyne.edu/taxsim/, accessed 14 June 2007.

Gordon, R. and Li, W. 2005. Tax structure in developing countries: many puzzles and a possible explanation. Working Paper No. 11267, Cambridge, MA: NBER.

Grogger, J. 2003. The effects of time limits, the EITC, and other policy changes on welfare use, work, and income among female-headed families. *Review of Economics and Statistics* 85, 394–408.

Heckman, J.J., Lochner, L. and Cossa, R. 2002. Learning-by-doing vs. on-the-job training: using variation induced by the EITC to distinguish between models of skill formation. Working Paper No.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
9083. Cambridge, MA: NBER.

Hotz, V.J. and Scholz, J.K. 2003. The earned income tax credit. In *Means-Tested Transfer Programs in the United States*, ed. R. Moffitt. Chicago: University of Chicago Press and NBER.

Hotz, V.J., Mullin, C. and Scholz, J.K. 2005. Examining the effect of the earned income tax credit on the labor market participation of families on welfare. Mimeo, UCLA and Wisconsin. Online. Available at http://www.ssc.wisc.edu/~scholz/Research/EITC_Draft.pdf, accessed 14 June 2007.

Internal Revenue Service. 2002. *Compliance estimates for earned income tax credit claimed on 1999 returns*. Washington, DC: Inland Revenue Service.

Keane, M. and Moffitt, R. 1998. A structural model of multiple welfare program participation and labor supply. *International Economic Review* 39, 553–89.

Liebman, J. 1997. The impact of the earned income tax credit on incentives and income distribution. *Tax Policy and the* 11, 83–119.

Liebman, J.B. 2002. Redistribution in the current U.S. social security system. In *The Distributional Aspects of Social Security and Social Security Reform*, ed. M. Feldstein and J.B. Liebman. Chicago and London: University of Chicago Press.

Meyer, B.D. and Rosenbaum, D.T. 2000. Making single mothers work: recent tax and welfare policy and its effects. *National Tax Journal* 53, 1027–61.

Meyer, B.D. and Rosenbaum, D.T. 2001. Welfare, the earned income tax credit, and the labor supply of single mothers. *Quarterly Journal of Economics* 116, 1063–114.

Moffitt, R.A. 2004. The idea of a negative income tax: past, present, and future. In *Focus* 23, 1–4. Online. Available at http://www.irp.wisc.edu/publications/focus/pdfs/foc232a.pdf, accessed 14 June 2007.

Robins, P.K. 1985. A comparison of the labor supply findings from the four negative income tax experiments. *Journal of Human Resources* 20, 567–82.

Scholz, J.K. 1994. The earned income tax credit: participation, compliance, and anti-poverty effectiveness. *National Tax Journal* 47, 59–81.

How to cite this article

Scholz, John Karl. "taxation and poverty." The New Palgrave Dictionary of Economics. Second Edition.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000236>
doi:10.1057/9780230226203.1680

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

taxation of corporate profits

Alan J. Auerbach
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Corporate profits taxes account for a relatively small share of revenues in leading industrial countries but represent a potentially important source of economic distortion. The incidence of corporate taxes has traditionally been assigned to owners of capital, but more recent theories have suggested that many other groups, from shareholders to owners of other domestic factors of production, may share the burden, and that the burden itself may be overstated. Although commonly described as taxes on income, corporate profits taxes may have quite different bases, making the economic effects potentially quite different from those of a tax on corporate source income.

Keywords

capital accumulation; capital gains and losses; capital gains taxation; capital mobility; computable general equilibrium model; deadweight loss; debt finance; depreciation; dividend taxation; double taxation; efficiency effect of taxation; equity finance; excise taxes; Harberger, A. C; imperfect competition; insurance; investment risk; portfolio investment; progressive and regressive taxation; proportional taxation; tax avoidance; tax competition; tax distortions; tax incidence; taxation of capital income; taxation of corporate profits; vintage capital

Article

Industrial countries commonly levy a tax on the earnings of corporations.
Corporate income taxes account for a relatively small share of revenues in these countries. As of 2000, the share of total government revenues accounted for by the corporate tax (OECD, 2002, Table 13) reached a high among the G-7 countries in Japan, at 14 per cent, and a low in Germany, at five per cent. In a number of these countries, this share had dropped substantially from the 1960s. For example, the US share stood at 16 per cent of revenues in 1965, almost double its 2002 value of nine per cent. The decline in importance of corporate tax revenues has been attributed to a variety of overlapping factors,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
including tax policy, shifts in business activities out of corporate form, more aggressive tax avoidance behaviour, financial innovation, and increasing tax competition and capital mobility among jurisdictions. In sorting through these potential explanations, economists continue to develop models of corporate tax incidence and efficiency effects and to consider the rationale for a separate tax on corporate income.
The US corporate tax dates at least to a corporate profits excise tax in 1909, four years before a constitutional amendment cleared the way for the 1913 introduction of a general income tax that included a corporate profits tax. In the years since, two key measures have been the relative taxation of distributed and retained earnings and the relative taxation of corporate and non-corporate capital income. The original US income tax imposed a ‘normal’ one per cent tax rate on both retained and distributed earnings but also imposed a graduated individual income surtax of up to six per cent on a tax base that excluded retained corporate earnings, leaving retained earnings subject to a much lower overall tax rate than distributed earnings and leaving corporate income as a whole facing a lower tax rate than income from non-corporate sources, which could not escape the surtax. Through the years, however, although the tax rate on retained earnings generally remained below that on distributed earnings, increases in the corporate level tax caused both rates to rise relative to tax rates on non-corporate income (Bank, 2006).
By the post-Second World War era, the corporate income tax was viewed as imposing an extra tax burden on activities within the corporate sector, even with the favourable treatment of retained earnings. This set the stage for the extremely influential analysis by Harberger (1962 and 1966, respectively) of the incidence and efficiency effects of the corporate income tax.

Harberger's contributions

Dividing the US economy into two sectors according to whether production was predominantly carried out by corporate or non-corporate businesses, Harberger (1962) characterized the corporate tax as an additional tax levied on capital income originating in the corporate sector, layered on top of the individual income tax collected on capital income from both sectors. He then estimated incidence through the changes in factor prices and product prices that would result from a small increase in the corporate tax. Harberger's main conclusion was that, under reasonable assumptions regarding the two sectors’ production elasticities of substitution and consumers’ elasticity of substitution between the two sectors’ products, the corporate income tax was borne fully by owners of capital, economy-wide. This finding has two important elements. First, capital bears the entire tax; it is not shifted to labour or consumers. Second, it is all capital, not just corporate capital, that bears the tax.
Harberger's second contribution (1966) is his estimates of the inefficiency resulting from the higher tax imposed on corporate capital. As a result of the corporate tax, he saw the social (before-tax) rate of return as higher in the corporate sector. Real national income would increase if the tax distortion were eliminated and capital were reallocated to equalize rates of return across sectors. Harberger estimated the deadweight loss of the tax differential between corporate and non-corporate investment to be about seven per cent of the taxes collected on corporate earnings.
Harberger's analyses spawned a vast literature that extended and challenged his initial results. The simplicity of Harberger's technique – comparative static analysis of small changes in a two-sector model – proved not to be a major source of concern given that similar findings resulted from analysis using a multi-sector computable general equilibrium model (Shoven, 1976). But Harberger also relied on several

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
other simplifying assumptions including: (*a*) free mobility of factors across sectors; (*b*) that the corporate tax can be viewed as an add-on tax on capital income originating in the corporate sector; (*c*) no risk; (*d*) competitive markets and constant returns to scale; (*e*) a closed economy; and (*f*) fixed economy-wide factor supplies. These and other assumptions have been examined in the literature.

Dynamics

It is reasonable to think of the shifts predicted by the Harberger model as occurring over time with after-tax returns to capital slowly equalizing in the two sectors and with corporate assets initially dropping in value to reflect the gap in after-tax returns, consistent with the *q*-theory of investment envisioned by Tobin (1969) and developed by Hayashi (1982), Summers (1981) and others. Thus, a corporate income tax with gradual adjustment to Harberger's long-run outcome would be borne partially by current owners of corporate capital, through an initial drop in asset values, and partially by future investors in corporate and non-corporate capital, through lower rates of return. The inefficiency of the corporate tax would be changed by gradual adjustment, with weaker responsiveness to tax changes translating into smaller present-value deadweight losses.

Investment provisions

As modelled by Harberger, the base of the corporate income tax equals income from all corporate capital. In particular, income from capital goods of different vintages is taxed at the same rate. In reality, capital goods of different ages receive different treatment, even though they are subject to the same statutory corporate tax rate, because of differences in depreciation provisions and other incentives provided to new investment. With accelerated depreciation and investment incentives, new assets are more attractive than old ones of the same productivity. Put another way, the effective tax rate on new investment may be lower than the effective tax rate on existing capital. The differential should be capitalized into the value of existing assets.
Calculations in Auerbach (1983) found that the value ratio of old to new corporate fixed capital in the United States should have been around 0.8 in the early 1980s, with a reduction in both investment incentives and the statutory corporate rate reducing capitalization substantially by the next decade (Auerbach, 1996). Thus, there is a second component of the corporate tax borne by corporate asset holders rather than by all capital, and potentially lower deadweight loss as well to the extent of the tax wedge being shifted from new capital to existing assets. Auerbach (1983) found the marginal effective corporate tax rate in the United States to be well below the average effective corporate rate.

Corporate financial policy and taxation

With corporations having the option to issue debt, the interest payments which are deductible at the corporate level, and to retain earnings, thereby trading off current dividends for capital gains on which taxes may be lower and can be deferred, how much ‘double taxation’ does corporate capital actually face? In the extreme, if corporations finance all their investment by borrowing, there is no corporate tax imposed on investment; indeed, corporate tax liability is reduced, because nominal interest payments – a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
portion of which simply compensates lenders for a loss in purchasing power – are tax deductible. Based on these attributes, Stiglitz (1973) concluded that firms should pay no dividends, retain all their earnings, and meet any additional financing needs by issuing debt. While this theory explains why some equity would exist, it fails to explain why so much equity exists. Here, the theory of Miller (1977) comes in. Miller focused on the heterogeneity of individual investors, arguing that, under a progressive tax system, there may be some investors in high enough tax brackets that the extra taxation at the corporate level is more than offset by the preferential individual tax treatment of equity income. Under Miller's theory, investors with a tax preference for equity would hold equity, those with a tax preference for debt would hold debt, and corporations would be indifferent between the two, issuing enough of the two securities to satisfy the demands of investors.

Even with investor heterogeneity, is it plausible that a significant share of investors will have a tax preference for equity? A very low effective equity tax rate would be required, and this seems inconsistent with the fact that a substantial share of equity earnings comes to investors as fully taxed dividends. One potential explanation is that many countries provide some form of dividend relief, either through reduced taxation at the corporate level (through a lower rate of corporate tax on distributed earnings or a deduction for dividends paid) or at the investor level (through a lower rate of tax on dividends received or a shareholder imputation credit for taxes paid at the corporate level). An alternative explanation comes from the ‘new view’ of dividend taxation (Auerbach, 1979; Bradford, 1981; King, 1977), under which the effective rate of individual tax on equity income may be the capital gains rate, adjusted for deferral – a very low rate – even if dividends are distributed, when retained earnings are the source of equity finance, as they are for most large corporations. This view stands in stark contrast to the ‘traditional’ view, under which the effective individual tax rate on equity earnings is a weighted average of the tax rates on dividends and capital gains, the weights reflecting the shares of corporate earnings distributed and retained.

The new view, which also attempts to explain why mature firms pay dividends, is based on the intuition that existing equity funds are ‘trapped’ within the firm, unable to get out easily without being subject to a tax rate on dividends. As a consequence, the dividend tax rate (or, more precisely, the excess of the dividend tax rate over the effective individual capital gains tax rate on retained earnings) will be capitalized into share values, having no effect on the incentive to distribute earnings or on the effective tax rate on equity-financed investment. Through the years, different empirical strategies have been used to test the relative validity of the traditional and new views of the impact of equity taxation. One approach, based on the *q*-theory investment model, appeared to provide strong support for the traditional view when based on UK data (Poterba and Summers, 1985) but equally strong support for the new view when based on US data (Desai and Goolsbee, 2004). Other approaches focusing on rates of return (Auerbach, 1984) and the source of investment funds (Auerbach and Hassett, 2003) have suggested the presence of firm heterogeneity, with the new view more relevant for ‘mature’ firms with ample internal funds.

Regardless of whether there are investors with a sufficiently low tax rate on equity, another serious challenge to the Miller model is that investors clearly do not specialize. Tax-exempt institutional entities invest substantially in equity, and higher-income individuals hold at least some corporate bonds. As discussed by Auerbach and King (1983), the Miller model breaks down when assets are risky and investors must balance the objectives of diversification and tax minimization. Tax preferences will

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
influence portfolios – those in higher brackets will still gravitate more towards assets, like equity, with more favourable individual tax treatment. This modification of the model implies that the incidence conclusions based on the simple Miller model are overly strong; while high-bracket investors suffer more from an increase in the corporate tax, even tax-exempt investors will also bear some of the burden.

Risk

Since the work of Domar and Musgrave (1944), economists have noted that taxes on capital income provide insurance as well as imposing burdens. As has been established in the literature, a proportional tax system that provides a full loss offset (that is, the same tax rate applies whether income is positive or negative) imposes a burden on investors only to the extent that the safe return is taxed. This result, combined with the empirical observation that the real, safe rate of return is very close to zero, led Gordon (1985) to suggest that the corporate income tax imposes few economic distortions and has little incidence, although it collects tax revenue on average (that is, in expected value).
Gordon's conclusion was challenged by Bulow and Summers (1984), who argued that, while the government shares in the income risk of a corporate investment, much of the investment risk is associated with fluctuations in the price of capital goods, and the corporate tax base excludes accruing capital gains and losses on fixed capital. Likewise, the corporate tax's risk-sharing is reduced by imperfect loss offsets, which also raise the effective tax rate. Thus, even if a pure, symmetric tax on accruing corporate income caused few distortions, this is unlikely to be true of actual corporate tax systems that consistently raise substantial revenues from corporations.

Imperfect competition

Harberger's conclusions on incidence were challenged by the econometric results of Krzyzaniak and Musgrave (1963), who, on the basis of time-series analyses of American manufacturing, reached the startling conclusion that *after*-tax profits rose in the short run in response to increases in the corporate tax rate. The Krzyzaniak–Musgrave contribution was criticized by a number of writers, but their empirical finding is possible under certain forms of imperfect competition. The study's methodology does not allow one to identify the nature of corporate responses, but corporations behaving in an oligopolistic manner need not maximize joint profits, and therefore might increase before-tax profits, and possibly even after-tax profits, by reducing output and hence increasing prices in response to an increase in the corporate tax.

The international economy

Unlike in the purely domestic context, there is a distinction between where income is earned and where its owner resides, and the concept of residence, itself, is applied not only to individuals but also to corporations. Countries may seek to tax corporate income on a source basis, a residence basis, or some combination of the two, and most countries follow this last approach, taxing at least some income at source at the corporate level, even if the corporation is owned abroad, and taxing at least some portfolio income of domestic residents on holdings of foreign assets.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As in the analysis underlying the Miller model, an equilibrium with individuals possessing different relative tax preferences for different assets leads to specialization of the highest-bracket investors in the most tax-favoured assets (Gordon, 1986), but the number of possible allocations of assets among investors is increased by the fact that individuals may hold foreign assets in many countries and in a variety of ways (for example, portfolio investment versus direct investment), and corporations (and, to a lesser extent, individuals) can change the location not only of their investments but also of their tax residence.

Among the key results in the international tax context is that a corporate income tax will be partially shifted to non-capital domestic factors of production, the more so the smaller is the taxing jurisdiction (Kotlikoff and Summers, 1987). Also, with many taxing jurisdictions, the possibility of 'tax competition' arises, with governments setting their corporate tax rates strategically in response to the tax polices of other countries. In this context, a 'race to the bottom' is a possible, though by no means a certain, outcome. But the reductions in corporate tax rates in recent decades provide some evidence for a strengthening of tax competition (Devereux, Griffith and Klemm, 2002).

The long run

One of the Harberger model's most important omissions is the impact of corporate income taxes on capital accumulation. We would expect an increase in the effective tax rate on new saving and investment to reduce capital accumulation. The resulting decline in the capital–labour ratio would increase before-tax returns to capital and lead to a fall in wages, thus partially shifting the tax burden from capital to labour, with much the same effect as capital flight in the open economy. This analysis would apply to the corporate tax as well, but only to the extent that the corporate income tax represents a tax on new saving and investment.

See Also

- dividend policy
- tax incidence
- tax competition

This is a revised version of the article by Peter Mieszkowski in the first edition of the dictionary.

Bibliography

Auerbach, A.J. 1979. Share valuation and corporate equity policy. *Journal of Public Economics* 11, 291–305.

Auerbach, A.J. 1983. Corporate taxation in the United States. *Brookings Papers on Economic Activity* 1983(2), 451–505.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Auerbach, A.J. 1984. Taxes, firm financial policy and the cost of capital: an empirical analysis. *Journal of Public Economics* 23, 27–57.

Auerbach, A.J. 1996. Capital allocation, efficiency and growth. In *Economic Effects of Fundamental Tax Reform*, ed. H. Aaron and W. Gale. Washington, DC: Brookings Institution.

Auerbach, A.J. and Hassett, K.A. 2003. On the marginal source of investment funds. *Journal of Public Economics* 87, 205–32.

Auerbach, A.J. and King, M.A. 1983. Taxation, portfolio choice and debt–equity ratios: a general equilibrium model. *Quarterly Journal of Economics* 98, 588–609.

Bank, S.A. 2006. A capital lock-in theory of the corporate income tax. *Georgetown Law Journal* 94, 889–947.

Bradford, D.F. 1981. The incidence and allocation effects of a tax on corporate distributions. *Journal of Public Economics* 15, 1–22.

Bulow, J.I. and Summers, L.H. 1984. The taxation of risky assets. *Journal of Political Economy* 92, 20–39.

Desai, M.A. and Goolsbee, A.D. 2004. Investment, overhang, and tax policy. *Brookings Papers on Economic Activity* 2004(2), 285–355.

Devereux, M.P., Griffith, R. and Klemm, A. 2002. Corporate income tax reforms and international tax competition. *Economic Policy* 17, 450–495.

Domar, E.D. and Musgrave, R.A. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58, 388–422.

Gordon, R.H. 1985. Taxation of corporate capital income: tax revenues versus tax distortions. *Quarterly Journal of Economics* 100, 1–27.

Gordon, R.H. 1986. Taxation of investment and savings in a world economy. *American Economic Review* 76, 1086–1102.

Harberger, A.C. 1962. The incidence of the corporation income tax. *Journal of Political Economy* 70, 215–40.

Harberger, A.C. 1966. Efficiency effects of taxes on income from capital. In *Effects of Corporation Income Tax*, ed. M. Krzyzaniak. Detroit: Wayne State University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hayashi, F. 1982. Tobin's marginal and average q: a neoclassical interpretation. *Econometrica* 50, 213–24.

King, M.A. 1977. *Public Policy and the Corporation*. London: Chapman & Hall.

Kotlikoff, L.J. and Summers, L.H. 1987. Tax incidence. In *Handbook of Public Economics*, vol. 2, ed. A. Auerbach and M. Feldstein. Amsterdam: North–Holland.

Krzyzaniak, M. and Musgrave, R.A. 1963. *The Shifting of the Corporation Tax*. Baltimore: Johns Hopkins Press.

Miller, M. 1977. Debt and taxes. *Journal of Finance* 32, 261–75.

OECD (Organization for Economic Cooperation and Development). 2002. *Revenue Statistics of OECD Member Countries 1965–2001*. Paris: OECD.

Poterba, J.M. and Summers, L.H. 1985. The economic effects of dividend taxation. In *Recent Advances in Corporate Finance*, ed. E.I. Altman and M.G. Subrahmanyan. Homewood, IL: Richard D. Irwin.

Shoven, J.B. 1976. The incidence and efficiency effects of taxes on income from capital. *Journal of Political Economy* 84, 1261–83.

Stiglitz, J.E. 1973. Taxation, corporate financial policy and the cost of capital. *Journal of Public Economics* 2, 1–34.

Summers, L.H. 1981. Taxation and investment: a q theory approach. *Brookings Papers on Economic Activity* 1981(1), 67–127.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1, 15–29.

How to cite this article

Auerbach, Alan J. "taxation of corporate profits." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000023> doi:10.1057/9780230226203.1681

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

taxation of foreign income

James R. Hines Jr.
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Taxation of foreign income entails the taxation by one country of income that its residents earn in another country. While most countries exempt active foreign business income from taxation, several large capital exporters subject foreign income to taxation but permit taxpayers to claim credits for taxes paid to foreign governments. There is extensive empirical evidence that the taxation of foreign income influences the magnitude of foreign investment and the tax avoidance activities of investors. Neutral taxation of foreign income entails considerations not only of the volume and location of investment, but also the effects of taxation on capital ownership.

Keywords

capital export neutrality; capital import neutrality; capital ownership neutrality; double taxation; foreign investment; national neutrality; tax credits; tax harmonization; taxable income; taxation of corporate profits; taxation of foreign income

Article

Taxation of foreign income entails the taxation by one country of income that its residents earn in another country.

Most countries subject some types of foreign income to taxation. Since this income is also typically taxed by the foreign countries in which it is earned, there is considerable scope for ruinous double taxation. For example, in the 1970s the corporate tax rate in the United States was 48 per cent while the corporate tax rate in Germany was 56 per cent; without some attenuation of double taxation, the combined tax rate of 104 per cent would probably have discouraged (profitable) American corporate investment in Germany.

International practice since the dawn of taxation is that countries tax income earned within their borders, whereas countries in which taxpayers are resident grant tax relief for foreign income in order to reduce

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
or eliminate double taxation. There are two primary alternative methods by which residence countries grant relief, the first being to exempt foreign income from taxation, and the second being to permit residents to claim credits for taxes paid to foreign governments. Many countries combine these systems, exempting active foreign business income from taxation while subjecting foreign personal income to taxation but permitting individuals to claim credits for income taxes paid to foreign jurisdictions.
The nature of international commerce is such that most foreign income is earned by businesses rather than individuals. Many countries largely exempt active foreign business income from taxation, though a number of major capital exporting nations, including the United States, the United Kingdom and Japan, tax foreign income while granting credits for taxes paid to foreign governments. With such a system of taxing foreign income, and a home-country corporate tax rate of 35 per cent, a corporation that earns 100 in a foreign country that imposes ten per cent tax rate pays taxes of 10 to the foreign government and 25 to its home government, since its home-country corporate tax liability of 35 is reduced to 25 by the foreign tax credit of ten. Since foreign tax credits are intended to alleviate international double taxation, credits are limited to home-country taxes due on foreign income; taxpayers are not permitted to use taxes paid to foreign governments to reduce home-country tax liability on domestic income. In addition, countries that tax active foreign income permit taxpayers to defer home-country taxation of certain business profits earned and reinvested abroad; that income is taxed only when repatriated to the country of residence.

Effects of taxing foreign income

The taxation of foreign income and the tax laws of other countries have the potential to influence a wide range of corporate and individual behaviour, including, most directly, the location and scope of international business activity. Studies of behavioural responses to international tax rules find that multinational firms invest less in high-tax countries than they do in otherwise-similar low-tax countries. This is most evident from the disproportionate shares of financial and real investment in tax haven countries (Hines, 2005), but also appears in cross-sectional econometric estimates of the determinants of foreign investment. Controlling for income levels and other observable characteristics of host countries, foreign direct investment levels are negatively associated with local corporate tax rates, the implied elasticity of investment with respect to the tax rate generally lying close to –0.6 in data covering the 1980s (Hines and Rice, 1994), and increasing in magnitude to –1 or greater in evidence data since the 1990s (Altshuler, Grubert and Newlon, 2001). High rates of local taxes other than corporate income taxes are likewise associated with reduced levels of foreign investment (Desai, Foley and Hines, 2004a).
There is extensive evidence that firms arrange financial flows and intrafirm sales to reallocate taxable income from high-tax countries to low-tax countries. This reallocation is commonly accomplished by concentrating corporate borrowing, and therefore interest deductions, in high-tax countries (Desai, Foley and Hines, 2004b), and by adjusting prices paid for intrafirm financial transactions and sales of goods and services to minimize income reported in high-tax countries (Clausing, 2003). As a consequence, multinational firms report significantly higher profit rates in low-tax countries than in high-tax countries (Desai, Foley and Hines, 2003), and the ability to reallocate taxable income only increases the attractiveness of investing in low-tax countries.
Taxation of foreign income, together with provision of foreign tax credits, dampens incentives to earn

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
income in low-tax countries, since lower foreign tax payments reduce available foreign tax credits and thereby create greater home-country tax obligations. Foreign investment in the United States is consistent with these incentives, in that investors from countries that exempt foreign income from taxation concentrate their investments more heavily in low-tax states than do investors from countries that tax foreign income (Hines, 1996). The taxation of foreign income restricts the attractiveness of investment in low-tax countries to situations either in which ample foreign tax credits are available, or in which investors can profitably defer home-country taxation. In practice, American firms are much more likely to reinvest foreign profits earned in low-tax locations, since immediately returning these profits to the United States would produce significant tax obligations (Desai, Foley and Hines, 2001).

The impact of home-country taxation is illustrated by the practice of granting ‘tax sparing’ credits for investments in certain developing countries, thereby permitting taxpayers to claim credits for normal rates of foreign taxes, whether or not these have been actually paid. Evidence indicates that Japanese investors are much more likely to receive local tax concessions in countries with which Japan has ‘tax sparing’ agreements than they are elsewhere, and that Japanese investment is concentrated in these countries as a result (Hines, 2001). Finally, the taxation of foreign income has even encouraged some individuals and multinational firms to expatriate, effectively changing their places of tax residence to avoid home-country taxation of lightly taxed foreign income (Desai and Hines, 2002).

Neutral taxation of foreign income

International tax rate differences may encourage inefficient allocation of economic activity; consequently, considerable effort has been devoted to understanding the properties of tax systems that create neutral incentives.

Capital export neutrality (CEN) is the doctrine that an investor's income should be taxed at the same total rate regardless of the location in which it is earned. If a home-country tax system satisfies CEN, then a firm seeking to maximize after-tax returns has an incentive to locate investments in a way that maximizes pre-tax returns. This allocation of investment promotes global economic efficiency under certain circumstances. The CEN concept is frequently invoked as a normative justification for taxation of foreign income with provision of foreign tax credits (Richman, 1963), though in practice, countries limit foreign tax credits and commonly defer taxation of unrepatriated active business income.

The same logic implies that governments acting on their own, without regard to world welfare, should want to tax the foreign incomes of their resident companies while permitting only deductions for foreign taxes paid. Such taxation satisfies what is known as national neutrality, discouraging foreign investment by imposing a form of double taxation, but doing so in the interest of the home country, which disregards the value of tax revenue collected by foreign governments. From the standpoint of the home country, foreign taxes are simply costs of doing business abroad, and therefore warrant the same treatment as other costs. This line of thinking suggests that countries fail to advance their own interests in permitting taxpayers to claim foreign credits, or worse, in exempting foreign income from taxation.

A third neutrality principle is capital import neutrality (CIN), the doctrine that the return to capital should be taxed at the same total rate regardless of the residence of the investor. Pure source-based taxation at rates that differ between locations can be consistent with CIN, since different investors are taxed at identical rates on the same income. In order for such a system to satisfy CIN, however, it is also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
necessary that individual income tax rates be harmonized, since CIN requires that the combined tax burden on saving and investment in each location should not differ between investors. While CEN is commonly thought to characterize tax systems that promote efficient production, CIN is thought to characterize tax systems that promote efficient saving (Horst, 1980).

The importance of ownership for productivity, and the reality that much foreign investment consists of acquisitions of existing assets by new owners, has prompted analysis of the features of tax systems that do not distort ownership of capital. Capital ownership neutrality (CON) is satisfied if every country taxes foreign income similarly, thereby avoiding tax-based ownership clienteles (Desai and Hines, 2003). From the standpoint of capital ownership, a country fails to advance world welfare by adopting a tax system that promotes CEN, if most capital exporters exempt foreign income from taxation.

The same circumstances that make CON desirable from the standpoint of world welfare also imply that countries acting on their own have incentives to exempt foreign income from taxation, regardless of what other countries do. The reason is that additional outbound foreign investment does not reduce domestic activity, since reduced home-country investment by domestic firms is offset by greater investment by foreign firms. Home-country welfare rises with the productivity of domestic factors, and is maximized by ownership patterns produced by exempting foreign income from taxation. Tax systems that exempt foreign income from taxation are therefore said to satisfy national ownership neutrality. Hence it is possible to understand why so many countries exempt active foreign business income from taxation, and it follows that, if every country did so, capital ownership would be allocated efficiently, to the benefit of global productivity.

See Also

- neutral taxation
- tax competition
- tax havens
- taxation of corporate profits

Bibliography

Altshuler, R., Grubert, H. and Newlon, T.S. 2001. Has U.S. investment abroad become more sensitive to tax rates? In *International Taxation and Multinational Activity*, ed. J.R. Hines Jr. Chicago: University of Chicago Press.

Clausing, K.A. 2003. Tax-motivated transfer pricing and US intrafirm trade prices. *Journal of Public Economics* 87, 2207–23.

Desai, M.A., Foley, C.F. and Hines, J.R., Jr. 2001. Repatriation taxes and dividend distortions. *National Tax Journal* 54, 829–51.

Desai, M.A., Foley, C.F. and Hines, J.R., Jr. 2003. Chains of ownership, tax competition, and the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

location decisions of multinational firms. In *Foreign Direct Investment in the Real and Financial Sector of Industrial Countries*, ed. H. Herrmann and R. Lipsey. Berlin: Springer.

Desai, M.A., Foley, C.F. and Hines, J.R., Jr. 2004a. Foreign direct investment in a world of multiple taxes. *Journal of Public Economics* 88, 2727–44.

Desai, M.A., Foley, C.F. and Hines, J.R., Jr. 2004b. A multinational perspective on capital structure choice and internal capital markets. *Journal of Finance* 59, 2451–87.

Desai, M.A. and Hines, J.R., Jr. 2002. Expectations and expatriations: tracing the causes and consequences of corporate inversions. *National Tax Journal* 55, 409–40.

Desai, M.A. and Hines, J.R., Jr. 2003. Evaluating international tax reform. *National Tax Journal* 56, 487–502.

Hines, J.R., Jr. 1996. Altered states: taxes and the location of foreign direct investment in America. *American Economic Review* 86, 1076–94.

Hines, J.R., Jr. 2001. Tax sparing and direct investment in developing countries. In *International Taxation and Multinational Activity*, ed. J.R. Hines Jr. Chicago: University of Chicago Press.

Hines, J.R., Jr. 2005. Do tax havens flourish? In *Tax Policy and the Economy*, vol. 19, ed. J.M. Poterba. Cambridge, MA: MIT Press.

Hines, J.R., Jr. and Rice, E.M. 1994. Fiscal paradise: foreign tax havens and American business. *Quarterly Journal of Economics* 109, 149–82.

Horst, T. 1980. A note on the optimal taxation of international investment income. *Quarterly Journal of Economics* 94, 793–8.

Richman, P.B. 1963. *Taxation of Foreign Investment Income: An Economic Analysis*. Baltimore: Johns Hopkins University Press.

How to cite this article

Hines, James R., Jr. "taxation of foreign income." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000247> doi:10.1057/9780230226203.1683

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

taxation of income

Alan J. Auerbach
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Income taxes are the single most important source of revenue for most countries, although there is an active debate about the relative attractiveness of alternatives such as broad-based consumption taxes. In practice, the income tax base deviates from a comprehensive income measure in several important respects, by excluding non-market activities, limiting refunds for losses, and including capital gains on realization rather than on accrual. Each deviation introduces additional distortions of taxpayer behaviour. Defining the family unit for purposes of income taxation remains a complex issue, as does the optimal degree of tax progressivity.

Keywords

capital gains taxation; consumption taxation; depreciation; direct taxation; double taxation; expenditure taxation; Haig–Simons measure of income; indexation of the income tax; indirect taxation; inflation; interpersonal utility comparisons; marginal tax rates; negative taxation; progressive and regressive taxation; realized vs unrealized capital gains; tax distortions; taxation of income

Article

Despite its current prominence as a government revenue source, the income tax's first appearance is relatively recent in the historical evolution of government finance.
Evidence of a serious national income tax is difficult to find before the end of the 18th century, when William Pitt achieved the passage in Great Britain of the Act of 1799, which imposed a comprehensive income tax, complete with exemptions and abatements for dependents, on all residents of Great Britain. Introduced to maintain the British government's solvency during the Napoleonic Wars, the income tax was dispatched once the French had been. Seligman (1911, p. 113) quotes a contemporaneous source as stating that the repeal of the tax by Parliament ‘was declared amidst the greatest cheering and the loudest exultation ever witnessed within the halls of the English Senate’. It was only decades later that the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
income tax reappeared in Britain.
This pattern of introduction during wartime, followed by repeal and eventual, permanent reinstitution is found in the experience of other countries, as well. In the United States, for example, the first income tax was introduced during the civil war in 1862, being abandoned in 1872. It reappeared in 1894 in similar form, but was almost immediately declared unconstitutional by the Supreme Court, which found it to be a ‘direct’ tax not apportioned among the states according to population. The 16th amendment to the constitution was required before the income tax could be imposed again in the United States, in 1913.
Over time, the income tax has grown in importance so that it now represents the single most important revenue source in most developed countries. In 2000 (according to OECD, 2002, Table 9), for example, taxes on income and profits among the G-7 countries (the United States, Japan, Germany, France, United Kingdom, Italy, and Canada) accounted for between roughly one quarter and one half of all revenues, with the United States relying the most on the income tax (51 per cent) and France the least (25 per cent).
Being a direct tax on individuals, rather than an indirect tax on transactions, the income tax requires a more developed government infrastructure than other revenue sources. This distinction also provides the key to understanding both why the income tax was seen as a fairer way to raise revenue and why it was so vehemently opposed. Through assessment of individuals, the income tax was better suited to the achievement of a progressive, broad-based tax structure than the agglomeration of indirect taxes and duties that preceded it. At the same time, this focus on individuals instead of transactions brought with it the perception of a challenge to individual liberty, both because of the exposure to the government of the individual's economic behaviour and the ability of government to levy arbitrarily high taxes on small groups of taxpayers (see, for example, Blum and Kalven, 1953).

The measurement of income

Dating back almost to the introduction of the income tax itself is the question of how income should be measured. What has come to be called the ‘Haig–Simons’ measure of income is now generally accepted as the appropriate base for an income tax (Haig, 1921; Simons, 1938). As expressed by Simons (1938, p. 50), ‘Personal income may be defined as the algebraic sum of (1) the market value of rights exercised in consumption and (2) the change in the value of the store of property rights between the beginning and end of the period in question.’ One may justify the Haig–Simons approach on grounds of both fairness and efficiency, the former because it treats individuals with different sources of income uniformly and the latter because it does not distort decisions of how to devote resources to the generation of income. Yet actual income taxes vary from this definition, for reasons of both administration and politics.

Market versus non-market activities

A range of activities generates imputed income that, by the Haig–Simons definition, should be included in the income tax base. Some sources of imputed income, such as imputed rent from owner-occupied housing, have been seriously considered for inclusion in the tax base. At the other extreme are various home production activities such as cleaning and home repair.
The inability and unwillingness of governments to tax income from non-market activities introduces a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
distortion of taxpayer choices, for it encourages the taxpayer to substitute non-market for market activities. Such substitution may be entirely legal, as in the purchase or cleaning of one's own home, or illegal, as in the establishment of professional cooperatives wherein members provide services to other members ‘for free’.

Realizations versus accruals

The Haig–Simons measure does not distinguish between realized and unrealized increases in wealth. Yet most income tax systems include capital gains in the tax base only when they are realized, if at all. This outcome is traceable in part to the difficulty of measuring unrealized gains, although this can hardly be a problem for the vast wealth held in marketable securities. A related difficulty, often ascribed to farmers and the owners of small businesses, involves illiquid assets which would have to be sold below their going-concern values were their owners subject to taxes on associated accrued gains. Finally, property rights to capital assets may be sufficiently vague or disputed that it is difficult even to attribute ownership until gains have actually been realized. A variety of proposals to tax such accrued gains retrospectively upon realization (for example, Vickrey, 1947; Auerbach, 1991) have attracted little more than academic attention.

This favourable treatment of capital gains facilitates the accumulation and transmission of wealth. It has therefore been viewed as mitigating the progressivity of the income tax system, leading some (for example, Kaldor, 1955) to favour an individual expenditure tax on grounds of equity. At the same time, others have argued that favourable capital gains treatment serves to encourage risk-taking, which is otherwise discriminated against by the income tax system (see the discussion below).

Nominal versus real income

Income has been viewed as an appropriate measure of the individual's ability to pay, but this ability is generally viewed in real rather than money terms, an individual being no better off with twice the income at twice the price level. Price-level indexation of the income tax requires two types of corrections: to the rate structure and to the base itself. The first is required because of the progressivity of marginal tax rates, the second because capital income is measured incorrectly in the presence of inflation.

When the income tax structure has marginal tax rates that rise with nominal income, increases in the price level raise both the marginal and the average tax burden on a given real income level. Correction for inflation involves indexing tax brackets to the price level, a practice implemented only in 1985 in the United States, after a period of relatively high inflation.

Capital income is generally mismeasured in an inflationary environment because changes in the real value of capital goods that are due to inflation are generally treated incorrectly by the tax system. This has led to four problems identified by the literature (see, for example, Aaron, 1976); the understatement of costs of goods sold from inventories, the understatement of the depreciation expenses associated with the use of durable capital goods, the overstatement of income received from bonds and other nominal commitments, and the overstatement of realized capital gains. In all cases, the problem arises from a failure to apply the Haig–Simons approach to changes in the value of capital assets. There have been

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
few attempts to ameliorate these distortions in practice. In fact, some economists have suggested that maintaining these apparently gratuitous distortions serves a positive purpose, namely, to weaken government's appetite to pursue inflationary policies (Fischer and Summers, 1989).

Losses

There is no logical reason why the income tax base for an individual cannot be negative, but treating losses as gains are treated would call for a negative tax payment, that is, a government refund. This outcome is rarely observed in practice, in part perhaps because restricting the use of losses imposes some limit on the extent to which taxpayers can fraudulently under-report income. Instead, individuals with currently negative tax bases are permitted to average the current base retrospectively or prospectively with the aim of achieving a net positive number. In the case of forward averaging (called ‘carrying forward’) this still amounts to the penalty of having to wait for the refund until future taxes are due without any interest to compensate for this deferral.

A closely related outcome occurs under a progressive tax structure when income is positive in every year, but fluctuates from year to year. Since marginal rates are higher in good years than bad (a milder version of what occurs when negative tax bases face a tax rate of zero), taxpayers face a higher tax in present value than if they received the same present value of income, but in a smooth stream over time. As with the treatment of losses, tax systems typically provide some imperfect form of averaging of incomes over several contiguous years to lessen this problem.

This treatment of losses and risky incomes has been viewed as discouraging the taking of risks (for example, Domar and Musgrave, 1944), and has been one of the more valid reasons for favouring the preferential treatment normally accorded capital gains. At the same time, there is no general presumption that income taxation, in itself, discourages the taking of risks, since it reduces not only the returns to risky investments but also the risks. In fact, the reduction in risk may increase private risk-taking (Domar and Musgrave, 1944; Tobin, 1958), but this outcome may be reversed if government cannot reduce the risk of its resulting revenue stream (Gordon, 1985).

Defining the unit of taxation

In addition to these problems of income measurement, ambiguities have arisen concerning the delineation of the taxpaying unit. Questions have concerned how broadly to define a unit at a given date, and over what time interval to measure the income accruing to that unit.

Tax treatment of the family

Tax systems vary in their treatment of related individuals. The method of treatment of family members affects the tax burden on a family because of the progressivity of the rate structure. Two individuals will generally be assessed a different tax bill if considered separately than if taxed jointly as a couple, regardless of how the rate structure is adjusted, since the total tax bill under separate taxation will depend on the distribution of taxable income between the individuals while under joint taxation it will not.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Even if families can be identified and grouped for purposes of taxation, the problem remains in deciding how to vary the tax schedule with family size. For this, one must have a measure of how to normalize income by family size to obtain a measure of the family's ability to pay. Such questions have been addressed but not often applied to the design of tax schedules.
Finally, how the family is grouped also matters if the tax-free transfer of resources through gifts and bequests is not allowed. The strict Haig–Simons approach would include gifts and bequests in the tax base of the recipient, but such transfers might not be observed if occurring within the unit of taxation.

Tax treatment over time

In Simons's own description of the appropriate measurement of income, the element of time plays a crucial role, since accretions to wealth must be defined over some interval. Indeed, the difference between a tax on income and a tax on expenditures amounts to a different choice of time interval over which to measure income for purposes of taxation. If, instead of annual income, we assessed taxes on lifetime income, then individuals would pay taxes in excess of lifetime consumption only to the extent that they accumulated resources over their lifetimes. Taking the further step of assessing families rather than individuals, we might ignore even the lifetime resource accumulation that would go to finance the consumption of subsequent generations.
This point has not been missed by advocates of the expenditure tax, who argue that, from a lifetime perspective, individual expenditures are a better measure of ability to pay than annual income. Under the annual income tax, individuals who consume their resources later in life face a heavier lifetime burden, paying taxes when income is initially earned and again when interest on the saved capital is received in later years. This outcome led to the charge that the annual income tax imposes unfair ‘double taxation’ of savings, an argument made by early proponents of a consumption tax (for example, Fisher, 1939; Kaldor, 1955). More recent arguments in favour of consumption taxation have focused on considerations of economic efficiency. The central point of this literature has been that, notwithstanding the government's inability to avoid all tax-induced distortions, a system including capital income taxation, which imposes greater and greater distortions on consumption as the time horizon lengthens, cannot be optimal (Chamley, 1986; Judd, 1985).

On the optimal progressivity of the income tax

As soon as the income tax became ensconced as a revenue source, economists began to consider how progressive it should be. Early researchers in the utilitarian tradition focused on how rapidly the tax burden should rise with income so as to exact an equal sacrifice from each individual given the particular utility function with which people were assumed to be endowed. The answer also depended on whether one was seeking equal absolute sacrifice, equal proportional sacrifice, or equal marginal sacrifice, all measured in units of the interpersonally comparable individual utilities (Musgrave, 1959, pp. 99–105).
Perhaps the most disquieting result from this line of investigation was that the achievement of equal marginal sacrifices required the equalization of incomes across individuals, if utility function were the same (Edgeworth, 1897). A missing element that would have altered this finding was the distortionary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
impact of taxes on economic behaviour. The high marginal tax rates needed to approach equality of after-tax incomes (100 per cent in the extreme case of complete equality) would undoubtedly become self-defeating, in that the after-tax incomes of all individuals would begin to fall as tax revenues ceased increasing with increases in marginal tax rates. Such an outcome would be inconsistent with any evaluation of social welfare that had Pareto efficiency as a necessary condition for an optimum.

There followed, eventually, a line of research that sought to reconcile the utilitarian aim of equal marginal sacrifice with the disincentive effects of marginal taxation. The seminal paper here is that of Mirrlees (1971), who found optimal marginal rates to be relatively low. Subsequent research has also shown that marginal tax rates should eventually approach zero at the highest incomes, a result that is not only dependent on various assumptions but also less politically controversial than one might first expect once it is recognized that it applies to marginal, not average, tax rates, and possibly only at very high incomes.

Further research on the distortions of the income tax has focused on its efficiency relative to other tax bases, such as labour income and consumption expenditures. A key result here is that of Atkinson and Stiglitz (1976), who showed that, under a relatively plausible restriction on preferences, government could not improve on a progressive tax on labour income using differential consumption taxes. If one interprets different commodities as consumption at different dates, then the implication is that a tax on labour income, or equivalently a tax on lifetime consumption, is the optimal progressive tax. Many complications limit the direct application of this result in the determination of actual tax policy, but the lesson has been helpful in what remains an active area of research. One notable complication is that tax policy evolves over time. This evolution makes comparisons of tax systems different from evaluations of a transition from one tax system to another, in which the treatment of assets accumulated under the previous tax system must be taken into account in performing welfare analysis (Auerbach and Kotlikoff, 1987).

See Also

- capital gains taxation
- consumption taxation
- optimal taxation
- public finance

Bibliography

Aaron, H., ed. 1976. *Inflation and the Income Tax*. Washington, DC: Brookings Institution.

Atkinson, A. and Stiglitz, J. 1976. The design of tax structure: direct versus indirect taxation. *Journal of Public Economics* 6, 55–75.

Auerbach, A. 1991. Retrospective capital gains taxation. *American Economic Review* 81, 167–78.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Auerbach, A. and Kotlikoff, L. 1987. *Dynamic Fiscal Policy*. Cambridge: Cambridge University Press.

Blum, W. and Kalven, H. 1953. *The Uneasy Case for Progressive Taxation*. Chicago: University of Chicago Press.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54, 607–22.

Domar, E. and Musgrave, R. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58, 388–422.

Edgeworth, F. 1897. The pure theory of taxation. *Economic Journal* 7, 550–71.

Fischer, S. and Summers, L. 1989. Should governments learn to live with inflation? *American Economic Review* 79, 382–88.

Fisher, I. 1939. The double taxation of savings. *American Economic Review* 29, 16–33.

Gordon, R. 1985. Taxation of corporate capital income: tax revenues versus tax distortions. *Quarterly Journal of Economics* 100, 1–27.

Haig, R. 1921. The concept of income: economic and legal aspects. In *The Federal Income Tax*, ed. R. Haig. New York: Columbia University Press.

Judd, K. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 28, 59–83.

Kaldor, N. 1955. *An Expenditure Tax*. London: Allen & Unwin.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Musgrave, R. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

OECD (Organization for Economic Cooperation and Development). 2002. *Revenue Statistics of OECD Member Countries 1965–2001*. Paris: OECD.

Seligman, E.R.A. 1911. *The Income Tax*. New York: Macmillan.

Simons, H. 1938. *Personal Income Taxation*. Chicago: University of Chicago Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, 65–86.

Vickrey, W. 1947. *Agenda for Progressive Taxation*. New York: Ronald Press.

How to cite this article

Auerbach, Alan J. "taxation of income." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000024>
doi:10.1057/9780230226203.1684

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

taxation of the family

James Alm
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

In administering an individual income tax, a country must decide what constitutes an ‘individual’. This choice has traditionally been seen as one between making either the family or the individual the ‘unit of taxation’. The choice between the family and individual as the unit of taxation in the income tax – indeed in any tax or transfer programme – is not clear-cut, and involves difficult trade-offs between competing and worthwhile goals. This article examines some of the issues that countries face in choosing the unit of taxation, or what is often referred to as ‘taxing the family’.

Keywords

community property laws; equity; horizontal equity; income splitting; labour supply; marriage and divorce; marriage tax; progressive and regressive taxation; tax compliance costs; taxation of income; taxation of the family; vertical equity; women's work and wages

Article

Nearly all countries around the world impose an individual income tax. In administering this tax, each country must decide what constitutes an ‘individual’; that is, each country must choose the ‘unit of taxation’. This choice has traditionally been seen as one between the family and the individual. In the former case, the incomes of all members of a family are aggregated, and the income tax (with all of its relevant provisions) is then imposed on total family income. In the latter case, each individual is taxed only on his or her own individual income, even if he or she is a member of a family unit in which other members have taxable income.
The choice between the family and individual as the unit of taxation is not clear-cut, and involves difficult trade-offs between competing and worthwhile goals. With the dramatic increase in recent years of different household ‘types’ – cohabiting but not legally married couples, extended families, same-sex couples, unrelated individuals living together – these issues have become even more complicated. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
presence of numerous other tax and transfer programmes whose magnitude is determined by family status complicates these issues still more.
This article examines some of the issues that countries face in choosing the unit of taxation, or what is often referred to as ‘taxing the family’.

Some goals and principles in taxing the family

Countries have a variety of goals in choosing the structure of the individual income tax. Such a tax is usually viewed as balancing the various desirable attributes of taxation: taxes must be raised (*adequacy*) in a way that treats individuals fairly (*equity*) and in a way that minimizes interference in economic decisions (*efficiency*). See Boskin and Sheshinski (1983) and Apps and Rees (1999) for an analysis of the optimal taxation of the family.
Defining equity is quite difficult. One notion of equity requires that taxpayers with greater income pay greater amounts of taxes. It is generally felt that a progressive rate structure is best able to achieve vertical equity, sometimes referred to as the *progressivity* goal of taxation.
Another notion requires that taxpayers who are equal in all relevant respects pay equal amounts of taxes. The difficulty here lies in defining ‘equals’. Equals can be defined in terms of married couples with equal income, but also as any ‘household’ with equal income. If a married couple is seen as the relevant household type for defining equals, then achieving the goal of *horizontal equity across households* requires that married couples with equal incomes pay equal taxes. However, if a household is defined more broadly, then achieving this goal requires that any households with equal income pay the same amount of taxes, requiring the additional and separate goal of *equal payments by singles and couples*. This goal can easily be broadened to apply to all household types.
Still another goal is *marriage neutrality*, which requires that a couple's combined income tax liability remain unchanged with marriage. However, there is substantial evidence that many couples pay more in taxes as a married couple than their combined taxes as single individuals; this is often referred to as a ‘marriage tax’ or ‘marriage penalty’. There is also evidence that marriage can reduce tax liabilities, in which case the reduction in taxes is called a ‘marriage subsidy’ or ‘marriage bonus’. As discussed later, it is well documented that the US individual income tax is not marriage neutral, and exhibits a large and variable marriage penalty – and marriage bonus – whose magnitude has changed over time. There is also evidence that the income tax is not marriage-neutral in many other countries.
In general, any income tax can create a marriage penalty or subsidy if two conditions are satisfied: the tax is based on household income, and the tax imposes different marginal tax rates at different levels of income (Steuerle, 1999). Many tax and transfer programmes meet these conditions, so that that marriage non-neutrality exists throughout the fiscal system in almost all countries. For the United States, the General Accounting Office (1996) identified 59 provisions in the individual income tax code that contribute to a marriage penalty or subsidy, and over 1,000 federal laws in which marital status is a factor in the determination of taxes or transfers, ranging from income tax and welfare provisions to programmes involving veterans’ payments, immigrant benefits, and other social insurance programmes.
It is now well-known that no individual income tax can achieve the simultaneous goals of *horizontal equity across families*, *equal payments by singles and couples*, *progressivity*, and *marriage neutrality*. Choosing the features of the individual income tax therefore requires that countries must face trade-offs

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in their pursuit of worthwhile goals. Countries have made very different choices in this regard.

Worldwide practice in taxing the family

In the case of the United States, the tax treatment of the family has varied over time (Bittker, 1975; Berliant and Rothstein, 2003). The basic unit of income taxation was initially the individual. However, after the Second World War a growing number of states instituted community property laws, which allowed married couples to divide their income equally and file separate tax returns, thereby giving a significant tax advantage to couples living in community property states. In response, the Revenue Act of 1948 changed the unit of taxation from the individual to the family with the adoption of ‘income splitting’ for married couples, in which all couples were allowed to aggregate and to divide in half their income for federal tax purposes. Due to the progressive nature of the individual income tax, a couple's joint tax liability fell with marriage (for example, a marriage bonus).

This marriage bonus grew over the next two decades. Public pressure to remedy this disparity let to the adoption of the Tax Reform Act of 1969, which established a new, separate tax schedule for single individuals that insured that single persons would incur a maximum tax liability of 120 per cent of a married couple with equal income. However, a side effect of this reform was the creation, for the first time, of a marriage tax or penalty for many married couples, especially for couples with similar earnings. Since then, various tax and demographic changes have markedly affected the potential for a marriage penalty or subsidy, as well as the magnitude of each. In the longer run, the size of the marriage penalty will be heavily influenced by the alternative minimum tax.

Other countries have made very different choices in taxing the family (Alm and Melnik, 2005). In most OECD countries (as of 2002 tax laws), the individual is the unit of taxation, and joint filing for couples is not permitted. Joint filing is required in only seven countries (Belgium, France, Greece, Luxembourg, Portugal, Switzerland, and the United States), while six countries allow couples to select the filing status (Germany, Iceland, Ireland, Norway, Poland, and Spain). A total of 17 OECD countries use only the individual as the unit of taxation, and, as noted, another six countries in which the taxpayer can choose between single or joint taxation. Not every country that permits or requires joint filing allows joint assessment (for example, income splitting between the spouses). For instance, Greece, Norway, and Spain all have provisions for joint filing, but income splitting does not apply, which means that joint filing is not meaningfully different from single filing, except when joint filing allows a couple to use different personal exemptions. Income splitting is present in some form in only nine of the 32 OECD countries. In most countries that allow income splitting, the income of the spouses is simply aggregated, so that the tax system does not differentiate between households with equal combined incomes based on how the income is distributed within the couple. However, there are exceptions to this as well. For example, Belgium allows only limited income splitting, which applies only to those couples in which there is a significant differential between the spouses’ incomes.

Income splitting in the presence of a progressive tax rate structure creates a tax benefit to couples when spouses earn different incomes, as evidenced quite clearly by the US experience. Furthermore, the tax benefit is a function of the difference in those incomes and the marginal tax rate structure. For instance, Luxembourg has narrowly defined marginal tax rate brackets, and a relatively small differential can translate into a significant tax saving for a married couple as opposed to two single taxpayers with similar incomes, as long as neither spouse falls into the highest income bracket. In contrast, in Iceland

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and Ireland a much larger difference in incomes may have no impact on tax liability due to the marginal tax rate structure.

Many OECD countries also have special tax provisions that apply to single-earner couples, in an attempt to provide some form of tax relief to these couples. The most common provision is some form of credit, deduction, allowance, or rebate (for example, Australia, Austria, Canada, the Czech Republic, Denmark, Iceland, Italy, Japan, the Republic of Korea, and the Slovak Republic). Another popular provision is income splitting, used in Belgium, France, Germany, Ireland, Luxembourg, Poland, Portugal, and the United States.

In general, the dominant practice of individual income taxation in OECD countries is to choose the individual rather than the family as the unit of taxation, and thereby to tax individuals on their own income even if they are married. As a result, the individual income tax is largely marriage neutral in these countries. This practice of taxing the individual is one that has tended to emerge since the mid-1970s in these countries. Even so, there remains much diversity in how OECD countries choose to tax the family.

Some effects of taxing the family

Defining the taxable unit as the family rather than the individual is controversial. The principal arguments revolve around equity issues. However, there are also efficiency issues, as well as revenue effects.

The basic economic model of marriage indicates that income taxes may affect the gains to marriage via two paths. First, differential income tax treatment of married couples may alter the total taxes paid by the couple relative to taxes paid as single individuals. If total taxes paid increase (decrease) with marriage, *ceteris paribus*, then the gains to marriage unambiguously fall (rise). Second, marriage may change the marginal tax rate faced by the couple relative to that faced as singles. A higher marginal tax rate with marriage increases the tax liability of the couple and so lowers the benefits of marriage; however, a higher marginal tax rate also lowers the after-tax wage rates of the individuals, thereby reducing the opportunity cost of household production work and increasing the gains from marriage. The tax system therefore both creates incentives for those marriages that involve traditional gender roles and discourages those marriages with two wage earners.

Much recent empirical work has attempted to disentangle the effects of income taxes on marital decisions, with most of this work focusing on the United States. See Whittington and Alm (2003) for a summary of many of these studies. This work has tended to find that the marriage tax has a small but statistically significant impact on marriage and divorce probabilities. The income tax may also affect the timing of the marriage decision, and several studies have found that couples in the United States and in Canada, England and Wales have timed their marriages to avoid one year of the tax penalty. There is some evidence that income tax influences the likelihood that individuals live together as a legally married versus as a cohabiting couple. Even so, in most instances the magnitude of the tax impact is relatively small. For example, Alm and Whittington (1999) find that at mean values a ten per cent rise in the marriage penalty leads to a 2.3 per cent reduction in the possibility of first marriage, and Alm and Whittington (1997) estimate that doubling the tax penalty increases the probability that a couple delays its marriage to the next tax year by one per cent but that the tax penalty subsidy has no impact on the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
timing of divorce.

Marriage penalties and subsidies also affect the labour supply decisions (both participation and hours) of married individuals. Consider the secondary earner of the family. With the family as the unit of taxation, the secondary earner in a couple will be taxed at the marginal tax rate faced by the family on its combined income, and this tax rate is likely to be much higher than the tax rate that the individual would face if single. Taxing the family thereby discourages both labour force participation and hours worked of the secondary earner. Recent estimates of labour supply elasticities indicate that female labour force participation is especially responsive to marginal tax rates, particularly for women in high-income couples. See Kniesner and Ziliak (2008) for a comprehensive survey.

In sum, this recent research clearly demonstrates that the marriage penalty/subsidy distorts some individual decisions, even if these effects are not always large.

A number of other studies have attempted to calculate the actual penalties or subsidies in the US individual income tax. Although the precise estimates differ, their broad outlines are generally the same. For example, Feenberg and Rosen (1995) use individual tax returns to compute the US marriage penalty for 1994. They find that 51 per cent of married couples paid an average marriage penalty of about $1,200, 38 per cent received an average marriage subsidy of $1,400, and 11 per cent were unaffected. Couples more likely to incur a marriage penalty were two-earner families (especially with similar incomes), families with children, families with higher family income, and older families; single-earner couples generally received a marriage bonus. However, there was much dispersion in the size of the penalty/bonus across households. In total, income taxes were $6 billion higher than otherwise. Studies by Alm and Whittington (1996), the Congressional Budget Office (1997), and Bull et al. (1998) give comparable results.

The marriage penalty/bonus affects families at different levels of income very differently. For example, Alm and Whittington (1996) find that families in the highest family income quintile had an absolute tax penalty that was unambiguously larger than that borne by low- or middle-income couples but that, as a percent of income, the penalties and subsidies were much larger for low-income families. The Congressional Budget Office (1997) estimates that the average tax penalty for a family with less than $20,000 in adjusted gross income (AGI) was 7.6 per cent, compared with 1.6 per cent for families with AGI greater than $50,000. However, the level of family income does not unambiguously determine the magnitude of the marriage penalty/bonus. Rather, it is the distribution of income between husband and wife that largely determines the magnitude and the sign of the change in taxes with marriage.

Overall, these differential tax treatments of individuals and families introduce large, variable, and capricious inequities due to unequal treatment of taxpayers based solely on their marital status:

- between married couples with one earner (who get a bonus relative to what they would pay as singles) and those with two earners (who pay a penalty), even if each type of couple pays the same taxes as a married couple;
- between married couples and cohabiting couples – individuals may choose to live together as unmarried cohabitors because of the income tax savings that unmarried status gives them;
- between married couples and single households (for example, the so-called ‘singles tax’); and
- between married couples and extended households (for example, non-marital cohabitation in same-sex households, in households with related individuals, in households with unrelated

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

individuals).

As Steuerle (1999) has noted, the marriage tax is almost like a voluntary tax, imposed on those who have decided voluntarily to marry.

Conclusions

What do we want an individual income tax to achieve? There is today an enormous, and increasing, diversity of family structures in many countries. In the 1950s, the 'traditional family' was typically a single-earner household with a stay-at-home spouse. Now, many individuals choose to live alone, two-earner families are the norm, non-marital cohabitation among opposite and same-sex couples is common, extended families are increasing in numbers, and there are widespread instances of unrelated individuals living together. These newer types of households are, by many definitions, 'families'. However, they are treated very differently, and often much less favourably, than the traditional households once envisioned as the norm by the tax codes in many countries.

If concern with the marriage penalty/bonus is an overriding issue, it is certainly possible to move the individual income tax toward *marriage neutrality*. One obvious method here is to make the individual the unit of taxation. However, it is also possible to move closer to neutrality even while retaining the family as the unit of taxation, by such piecemeal reform options as increasing the standard deduction for married couples, expanding the tax brackets facing married couples, establishing a secondary earner deduction, expanding the phase-out range of transfer programmes, flattening overall rate structures, or allowing optional individual filing. More fundamental reform options include eliminating progressivity (for example, a flat rate tax) or even eliminating the individual income tax and replacing it with a national sales tax or a value-added tax.

However, these reform options would only reduce the marriage penalties (and bonuses) in the individual income tax. There would still be marriage penalties and bonuses throughout other parts of the tax and transfer system.

Moreover, suppose the individual is made the unit of taxation, as many OECD countries have chosen to do. This choice is not without its own efficiency, equity, and adequacy problems. An important justification for the use of the family as the unit of taxation is the notion that families with equal family income should pay equal taxes. There is no question that making the individual the unit of taxation would violate this goal of *horizontal equity across families*, as well as *equal payments by singles and couples*. There are also significant administrative and compliance issues from individual taxation. How are itemized deductions split between partners? How is unearned (or capital) income split between partners? Who claims the tax benefits from children? How do the tax enforcement agencies verify the legitimacy of these declarations? What are the compliance costs of individual filing? Many other such issues naturally arise, and the ways in which these issues are resolved vary greatly across countries.

It may well be that the importance of the traditional family unit still justifies its favourable tax treatment. This is clearly the avenue that the United States has chosen, and it seems unlikely that this choice will change anytime soon. However, it may also be time to recognize that a diverse society can no longer treat one family structure so differently from the others. Making the individual the unit of taxation would eliminate the marriage tax/subsidy (and the singles tax), and would also re-establish the principle of horizontal equity, broadly defined to apply to individuals and not to families or to couples. Many

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
countries have in fact chosen to make the individual the unit of taxation.
There are no easy choices here, and it is inevitable that the goals of taxation are often conflicting. Taxing the family requires facing these difficult trade-offs directly.

See Also

- family decision making
- family economics
- horizontal and vertical equity
- marriage and divorce
- marriage markets
- taxation of income

Bibliography

Alm, J. and Melnik, M.I. 2005. Taxing the ‘family’ in the individual income tax – an international perspective. *Public Finance and Management* 5, 67–109.

Alm, J. and Whittington, L.A. 1996. The rise and fall and rise … of the marriage tax. *National Tax Journal* 49, 571–89.

Alm, J. and Whittington, L.A. 1997. Income taxes and the timing of marital decisions. *Journal of Public Economics* 64, 219–40.

Alm, J. and Whittington, L.A. 1999. For love or money? The impact of income taxes on marriage. *Economica* 66, 297–316.

Apps, P.F. and Rees, R. 1999. Individual versus joint taxation in models with household production. *Journal of Political Economy* 107, 393–403.

Berliant, M. and Rothstein, P. 2003. Possibility, impossibility, and history in the origins of the marriage tax. *National Tax Journal* 56, 303–17.

Bittker, B.I. 1975. Federal income taxation and the family. *Stanford Law Review* 27, 1388–463.

Boskin, M.J. and Sheshinski, E. 1983. Optimal tax treatment of the family. *Journal of Public Economics* 20, 281–97.

Bull, N., Holtzblatt, J., Nunns, J.R. and Rebelein, R. 1998. Assessing marriage penalties and bonuses. Working paper. Washington, DC: Office of Tax Analysis, U.S. Department of the Treasury.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Congressional Budget Office. 1997. *For Better or For Worse: Marriage and the Federal Income Tax*. Washington, DC: Congress of the United States.

Feenberg, D.R. and Rosen, H.S. 1995. Recent developments in the marriage tax. *National Tax Journal* 48, 91–101.

Kleven, H.J., Kreiner, C.T. and Saez, E. 2006. The optimal income taxation of couples. Working Paper No. 12685. Washington, DC: NBER.

Kniesner, T.J. and Ziliak, J.P. 2008. Evidence of tax-induced individual behavioral responses. In *Fundamental Tax Reform: Issues, Choices, and Implications*, ed. J.W. Diamond and G.R. Zodrow. Cambridge, MA: MIT Press.

Steuerle, C.E. 1999. Valuing marital commitment: the radical restructuring of our tax and transfer systems. *The Responsive Community* 9, 35–45.

US General Accounting Office. 1996. *Income Tax Treatment of Married and Single Individuals*. GAO/GGD-96-175. Washington, DC: US General Accounting Office.

Whittington, L.A. and Alm, J. 2003. The effects of public policy on marital status in the United States. In *Marriage and the Economy – Theory and Evidence from Advanced Industrial Societies*, ed. S. Grossbard-Shechtman. New York: Cambridge University Press.

How to cite this article

Alm, James. "taxation of the family." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000227> doi:10.1057/9780230226203.1682

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

taxation of wealth

Alan J. Auerbach
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

One of the oldest forms of taxation, wealth taxes may take the form of annual taxes based on property or net worth, or assessments collected at less regular intervals, on estates or inheritances or in the form of emergency capital levies. Wealth taxes are still prominent but have become less important than income taxes as a source of revenue. Although wealth taxes are related in structure to taxes on capital income, their economic effects depend on their form. In addition to explicit taxes on wealth, governments impose implicit capital levies through changes in tax policy.

Keywords

capital levies; dynamic inconsistency; estate taxes; George, H.; implicit wealth taxes; inheritance taxes; Land tax; local wealth taxes; net worth taxation; precautionary savings; property tax; single tax; taxation and risk-taking; taxation of capital income; taxation of wealth; Tiebout hypothesis

Article

Wealth taxation is one of the oldest methods of government revenue collection, having been used at least since the time of the ancient Greeks.
According to Seligman (1895, p. 34), Athens levied a general property tax not only on land and houses, but also on slaves, cattle, furniture, and money. The succeeding centuries have seen many types of wealth taxation tried and others proposed, with perhaps no other form of taxation being the subject of such heated debate.
Wealth taxes are in some sense taxes on capital income. All assets have value because of the returns they generate, though the returns need not be in explicit form (as with business assets) but may be implicit (as with owner-occupied housing or gold). Taxing wealth on an annual basis may therefore be viewed as equivalent to taxing its return at a rate sufficient to produce the same tax revenue. However, a number of factors make wealth taxation and capital income taxation different in practice.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
First, governments often levy taxes on very particular forms of wealth (such as land), while income taxation has more typically been broad-based. Second, taxes on income have usually been confined to income explicitly realized. This is, for example, the nearly universal approach to capital gains taxation. Wealth taxes, such as property taxes on owner-occupied real estate, have not followed the same principle. Thus, it has been possible for taxpayers with very little liquidity but substantial wealth to be burdened with taxes well in excess of their *explicit* income. Moreover, wealth is generally even more unevenly distributed among the population than realized capital income. Finally, while it is natural to expect that income taxes would not exceed 100 per cent of income, there is no comparable upper bound on wealth taxes short of the entire stock of wealth. The perception that wealth taxes threaten the rights of individual citizens through the possibility of discriminatory or unfair taking of assets is undoubtedly a factor in the controversy that has often surrounded wealth taxation.

As of 2000, wealth taxes accounted for five per cent of tax revenues for the average OECD country, although the share was ten per cent or more in several countries, including the United States, the United Kingdom, Canada and Japan (OECD, 2002, Table 23).

Types of wealth taxation

In discussing the history and economic effects of wealth taxation, it is useful to distinguish the primary forms of wealth taxation that have been used. There are four that may be considered important.

Property taxation

This is the oldest form of wealth taxation, dating from antiquity. It is characterized by a tax at regular intervals (for example, yearly) on particular forms of private wealth, most commonly land, but also other forms of property. Originating in an era when land ownership was a much better measure of one's ability to pay than is currently so, property taxes have gradually been replaced and supplemented by other forms of taxation. In the United States, for example, where property taxes are still relatively important, the fraction of government revenue coming from property taxation fell to ten percent or so in the years after the Second World War, compared with around 40 per cent during the first three decades of the 20th century (Carter et al., 2006, Figure Ea-c).

Estate and inheritance taxation

Taxes on bequests and inheritances first appeared long after general property taxes. They differ from property taxes in that they are assessed only once, at death, and typically apply to most assets bequeathed.

Estate taxes are typically quite complicated and not particularly successful either at revenue collection or wealth redistribution. Some have referred to the estate tax as a ‘voluntary tax’ because there are so many tax-planning devices available to reduce or eliminate the tax burden imposed on transfers to wealth.

Net worth taxation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As discussed above, net worth taxation, which applies to assets net of liabilities, is very similar in effect to a tax on capital income, differing primarily in its coverage of assets which do not generate substantial current realized income.

Whereas property taxes and taxes on estates or inheritances are quite common, only a handful of developed countries currently supplement their revenue collections with broad-based annual taxes on net worth. At the end of the 20th century, only two OECD countries, Iceland and Spain, reported non-negligible revenues from taxes on net worth, although even for these two countries such taxes still accounted for less than one half of one per cent of total revenues (OECD, 2002).

Capital levies

In wartime, countries need vast resources over short periods of time. The preferred method of raising such funds has been the issuance of national debt, but several countries resorted in the 20th century to the capital levy, a ‘one time only’ tax on existing wealth, more burdensome than annual net worth taxes but ostensibly temporary. Such levies were imposed in the First World War by Germany, Czechoslovakia, Austria and Hungary, and prior to the Second World War by Italy and Hungary (Hicks, Hicks and Rostas, 1941). Naturally, fully unanticipated, one-time wealth taxes are non-distortionary if not particularly fair, but the appearance of one country more than once in the above list indicates the difficulty of using capital levies unexpectedly, a problem popularly known as ‘dynamic inconsistency’.

The economic effects of wealth taxation

One may start an analysis of the effects of wealth taxation by noting its similarity to the taxation of capital income, with its coincident discouragement of saving. In practice, however, different forms of wealth taxation may have different or additional effects because of their design. For example, the capital levy may be less distortionary to the extent that it is unanticipated.

The ‘single tax’

In *Progress and Poverty* (1882), Henry George argued for the use of a tax on the rent of unimproved land as the chief source of government revenues. Though some contemporary authors disagreed, it is fairly clear that George's tax, in hitting the return to a productive factor in extremely inelastic supply, would have imposed minimal distortions of economic behaviour. Of course, the ‘single tax’ movement that followed for many years after was invested with a fervour based on much more than the desire for Pareto efficiency.

Taxation and risk-taking

Many authors (for example, Domar and Musgrave, 1944; Stiglitz, 1969) have noted the potentially different impact on risk-taking occurring under capital income and wealth taxes because of the failure of the latter to vary with the asset returns actually realized. But this distinction is more apparent than real once investors have taken the opportunity to scale their asset holdings up or down in response to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
taxation. As Tobin (1958) showed, a tax on risky capital income has no effect on individual welfare and private risk-taking when the safe rate of return is zero. It may further be shown (Gordon, 1985) that, when social risks are efficiently traded, neither does *social* risk-taking change when such a tax is levied. These results are quite easily extended to the case when the safe rate of return is positive and the tax is on capital income *in excess* of this safe return on assets. Thus, if we view a tax on all capital income as one on the safe return to capital and one on the excess returns that compensate for risk, only the former component has economic impact. Hence, a capital income tax may be seen as equivalent to a tax levied on asset values multiplied by a fixed, safe rate of return, in which case it is obviously equivalent to a tax on wealth (Kaplow, 1994). Note that this equivalence requires efficient risk-bearing. Otherwise, capital income taxation may be preferred because it allows the government to pool risks that individual investors have not been able to pool privately.

The Tiebout hypothesis

In the United States, property taxes are used primarily to pay for local public services, and are the dominant source of finance for such services. This connection means that, if communities differ in their property tax burdens, one cannot ignore the implied differences in public services if it is necessary to live in a community to partake of its public services, such as education, trash removal, and police and fire protection. In the absence of such a connection, one would expect property tax differentials to be reflected in land prices. However, with government using the taxes to pay for desired public outputs, one might expect a different result.

This notion was formalized by Tiebout (1956), who sketched a theory in which different communities levy different taxes and provide different bundles of public services, the result being a Pareto-optimal allocation with individuals choosing their place of residence according to the bundle of local public goods and services desired. Aside from the difference between such an entrance fee and the actual property tax, there are many other issues that arise in considering the validity of the Tiebout model (Mieszkowski and Zodrow, 1989). Nevertheless, local wealth taxes, with competing jurisdictions, are clearly different in their impact from national wealth taxes.

Bequests and the estate tax

If annual wealth taxes discourage saving in a way similar to annual capital income taxes, one might presume the same result for estate taxes, with the large anticipated one-time burden having an important impact on lifetime saving. However, this presupposes an economic model of bequests that is by no means well accepted, that is, that they are the manifestation of a desire to leave resources to heirs, and that they are influenced by the ‘price’ of an after-tax dollar in the heir's hands. At least one factor leading to bequests is the absence of efficient annuities markets, inducing the need for elderly individuals to engage in precautionary saving to provide for unanticipated health expenses or a longer than predicted lifetime. Such saving would not be influenced at all by taxes imposed after death. Hence, one might view estate taxes as being potentially less distortionary than wealth taxes on the living, but this benefit has not been enough to overcome administrative complexity and popular opposition to strong estate taxes.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Policy changes and implicit wealth taxes

In addition to any taxes they explicitly impose on wealth, governments also use implicit wealth taxes whenever tax policies alter the attractiveness of different assets. For example, under a switch in the individual tax base from income to consumption, investors receive a higher after-tax return on new investments but are saddled with unanticipated taxes on the decumulation of existing assets for consumption purposes (Auerbach and Kotlikoff, 1987). Implicit wealth taxes of this sort are potentially much larger in magnitude than formal wealth taxes themselves.

See Also

- estate and inheritance taxes
- property taxation
- redistribution of income and wealth
- single tax
- Tiebout hypothesis

Bibliography

Auerbach, A. and Kotlikoff, L. 1987. *Dynamic Fiscal Policy*. Cambridge: Cambridge University Press.

Carter, S., Gartner, S., Haines, M., Olmstead, A., Sutch, R. and Wright, G. 2006. *Historical Statistics of the United States: Millennial Edition*, vol. 5. Cambridge: Cambridge University Press.

Domar, E. and Musgrave, R. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58, 388–422.

George, H. 1882. *Progress and Poverty*. New York: Appleton.

Gordon, R. 1985. Taxation of corporate capital income: tax revenues versus tax distortions. *Quarterly Journal of Economics* 100, 1–27.

Hicks, J., Hicks, U. and Rostas, L. 1941. *The Taxation of War Wealth*. Oxford: Clarendon Press.

Kaplow, L. 1994. Taxation and risk taking: a general equilibrium perspective. *National Tax Journal* 47, 789–98.

Mieszkowski, P. and Zodrow, G. 1989. Taxation and the Tiebout model: the differential effects of head taxes, taxes on land rents, and property taxes. *Journal of Economic Literature* 27, 1098–146.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

OECD (Organization for Economic Cooperation and Development). 2002. *Revenue Statistics of OECD Member Countries 1965–2001*. Paris: OECD.

Seligman, E. 1895. *Essays in Taxation*. 10th edn. New York: Macmillan, 1925.

Stiglitz, J. 1969. The effects of income, wealth and capital gains taxation on risk-taking. *Quarterly Journal of Economics* 83, 263–83.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, 65–86.

How to cite this article

Auerbach, Alan J. "taxation of wealth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000025>
doi:10.1057/9780230226203.1685

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Taylor rules

Athanasios Orphanides
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Taylor rules are simple monetary policy rules that prescribe how a central bank should adjust its interest rate policy instrument in a systematic manner in response to developments in inflation and macroeconomic activity. This article reviews the development and characteristics of Taylor rules in relation to alternative monetary policy guides and discusses their role for positive and normative monetary policy analysis.

Keywords

central banks; difference rules; econometric policy evaluation; Federal Reserve System; forecasting; Friedman, M.; full employment; inflation; inflation targeting; interest rate rules; IS–LM models; macroeconometric models; monetary policy; monetary policy rules; monetary transmission mechanism; money stock; money supply; natural growth of output; natural rate and market rate of interest; natural rate of employment; nominal income; output gap; policy design; price stability; real output; stabilization policy; Taylor rules; time consistency; uncertainty; velocity of circulation; Wicksell, J. G. K.

Article

Taylor rules are simple monetary policy rules that prescribe how a central bank should adjust its interest rate policy instrument in a systematic manner in response to developments in inflation and macroeconomic activity. They provide a useful framework for the analysis of historical policy and for the econometric evaluation of specific alternative strategies that a central bank can use as the basis for its interest rate decisions.

A perennial question in monetary economics has been how the monetary authority should formulate and implement its policy decisions so as to best foster ultimate policy objectives such as price stability and full employment over time. It is widely accepted that well-designed monetary policy can counteract macroeconomic disturbances and dampen cyclical fluctuations in prices and employment, thereby

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
improving overall economic stability and welfare. In principle, when economic growth unexpectedly weakens below the economy's potential, accommodative monetary policy can stimulate aggregate demand and restore full employment. Likewise, when inflationary pressures develop, monetary restriction can restore the central bank's price stability objective. In practice, however, given the limited knowledge that economists have about the macroeconomy – for example, about macroeconomic dynamics, about the monetary transmission mechanism, and even about the measurement of fundamental concepts such as the natural rates of output, employment and interest – there is substantial disagreement about the scope of stabilization policy and about policy design.

One approach is to decide upon what seems to be the best policy on a period-by-period basis, without appeal to any specific policy guide. A seeming advantage of this approach is that it gives policymakers the discretion to use their judgement period by period. However, a basic tenet of modern research is that systematic policy – that is, policy based on a contingency plan or policy rule – has important advantages over a purely discretionary policy approach. By committing to follow a rule, policymakers can avoid the inefficiency associated with the time-inconsistency problem that arises when policy is formulated in a discretionary manner. Following a rule allows policymakers to communicate and explain their policy actions more effectively. Policy based on a well-understood rule enhances the accountability of the central bank and improves the credibility of future policy actions. Also, by making future policy decisions more predictable, rule-based policy facilitates forecasting by financial market participants, businesses, and households, thereby reducing uncertainty.

Various proposals for monetary policy rules have been made over time, and a vast literature continues to examine the relative advantages and drawbacks of alternatives in abstract theoretical terms, in the context of empirical macroeconometric models, and in terms of the practical experience accumulated from past policy practice. To appreciate the appeal and limitations of Taylor rules, it is useful to relate their development to other proposals for systematic monetary policy.

Development of monetary policy rules

Some proposals suggest postulating a rule in terms of the main objectives of monetary policy, for example ‘maintain economic stability’ or ‘maintain a constant aggregate price level’. (See Simons, 1936, for early arguments favouring price-level targeting over discretionary policy.) One important practical difficulty with these proposals, however, is that the concepts involved are not under the control of the central bank and thus the proposals are not operational. In essence, these proposals fail to draw a clear distinction between the objectives of monetary policy and the policy instruments that are at least under the approximate control of the central bank. As a result, the suggested rules are only implicit in nature and are difficult to monitor and to distinguish from discretionary policy in a meaningful manner.

To be useful in practice, policy rules must be simple and transparent to communicate, implement and verify. This requires a clear choice of what should serve as the policy instrument – for example the money supply, m, or the short-term interest rate, i – and clear guidance as to how any other information necessary to implement the rule – for instance recent readings or forecasts of inflation and economy activity – should be used to adjust the policy instrument.

Perhaps the simplest example of a policy rule is the proposal that the central bank maintain a constant rate of growth of the money supply – Milton Friedman's k-percent rule (Friedman, 1960). The rule draws

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

on the equation of exchange expressed in growth rates:

$$\Delta m + \Delta v = \pi + \Delta q \quad (1)$$

where $\pi \equiv \Delta p$ is the rate of inflation and *p*, *m*, *v*, and *q* are (the logarithms of), respectively, the price level, money stock, money velocity, and real output. Selecting the constant growth of money, *k*, to correspond to the sum of a desired inflation target, π^*, and the economy's potential growth rate, Δq^*, and adjusting for any secular trend in the velocity of money, Δv^*, suggests a simple rule that can achieve, on average, the desired inflation target, π^*:

$$\Delta m = \pi^* + \Delta q^* - \Delta v^*. \quad (2)$$

Further, if the velocity of money were fairly stable this simple rule would also yield a high degree of economic stability. An early illustration of this rule appeared in 1935 in the work of Carl Snyder, a statistician at the Federal Reserve Bank of New York. After estimating that the normal rate of growth of trade in the United States was about four per cent per year at the time and observing that the velocity of money was stable, Snyder argued that ‘the highest attainable degree of general industrial and economic stability will be gained by an expansion of currency and credit … at this rate [four per cent]’ (Snyder, 1935, p. 198). During the 1960s and early 1970s, Milton Friedman's recommendation that the Federal Reserve control the rate of money growth to equal four per cent per year was similarly based on the assumption that potential output growth in the Unites States roughly equalled four per cent – the prevailing estimate at that time.

Another way to interpret this policy rule is in terms of the growth of nominal income, $\Delta x = \pi + \Delta q$. With the economy's natural growth of nominal income defined as the sum of the natural growth rate of output and the central bank's inflation objective, $\Delta x^* = \pi^* + \Delta q^*$, a rule for constant money growth can be seen as targeting this natural growth rate. An advantage of a constant money growth rule is that very little information is required to implement it. If velocity does not exhibit a secular trend, the only required element for calibrating the rule is the economy's natural growth of output. In addition, while the calibration of this rule does not rest on the specification of any particular model, the rule is remarkably stable across alternative models of the economy. In this sense, the policy of maintaining a constant growth rate of money is arguably the ultimate example of a rule that is robust to possible model misspecification.

Simple modifications allowing for some automatic response of money growth to economic developments have also been proposed as simple rules that could deliver improved macroeconomic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

performance (see, for example, Cooper and Fischer, 1972). Among the simplest such alternatives is the rule associated with Bennett McCallum (1988; 1993):

$$\Delta m = \Delta x^* - \Delta v^* - \varphi_{\Delta x}(\Delta x - \Delta x^*). \quad (3)$$

McCallum showed that, if a rule such as this (for example, with $\varphi_{\Delta x} = 0.5$) had been followed, the performance of the US economy likely would have been considerably better than actual performance, especially during the 1930s and 1970s – the two periods of the worst monetary policy mistakes in the history of the Federal Reserve.

A factor that complicates the use of the money stock as a policy instrument is the potential for instability in the demand for money due either to temporary disturbances or to persistent changes resulting from financial innovation. In part for this reason, central banks generally prefer to adjust monetary policy using an interest rate instrument.

A policy rule quite as simple as Friedman's *k*-percent rule cannot be formulated with an interest rate instrument. As early as Wicksell's (1898) monumental treatise on *Interest and Prices*, it was recognized that attempting to peg the short-term nominal interest rate at a fixed value does not constitute a stable policy rule. (Indeed, this was one reason why Friedman, 1968, and others expressed a preference for rules with money as the policy instrument.) Wicksell argued that the central bank should aim to maintain price stability, which in theory could be achieved if the interest rate were always equal to the economy's natural rate of interest, r^*. Recognizing that the natural rate of interest is merely an abstract, unobservable concept, however, he noted: ‘This does not mean that the bank ought actually to *ascertain* the natural rate before fixing their own rates of interest. That would, of course, be impracticable, and would also be quite unnecessary.’ Rather, Wicksell pointed out that a simple policy rule that responded systematically to prices would be sufficient to achieve satisfactory, though imperfect, stability: ‘If prices rise, the rate of interest is to be raised; and if prices fall, the rate of interest is to be lowered; and the rate of interest is henceforth to be maintained at its new level until a further movement in prices calls for a further change in one direction or the other’ (Wicksell, 1898, p. 189, emphasis in the original). In algebraic terms, Wicksell proposed what is arguably the simplest reactive monetary rule with an interest rate instrument:

$$\Delta i = \theta \pi. \quad (4)$$

Wicksell's simple interest rate rule did not attract much attention in policy discussions, perhaps because of its exclusive focus on price stability and lack of explicit reference to developments in real economic activity.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The classic Taylor rule and its generalizations

The policy rules that are commonly referred to as Taylor rules are simple reactive rules that adjust the interest rate policy instrument in response to developments in both inflation and economic activity. An important advance in the development of these rules can be identified with the policy regime evaluation project reported in a volume published by the Brookings Institution (Bryant, Hooper and Mann, 1993). The objective of the project was to identify simple reactive interest rate rules that would deliver satisfactory economic performance for price stability and economic stability across a range of competing estimated models. The Brookings project examined rules that set deviations of the short-term nominal interest rate, i, from some baseline path, i^*, in proportion to deviations of target variables z, from their targets, z^*:

$$i - i^* = \theta(z - z^*). \tag{5}$$

The collective findings pointed to two alternatives as the most promising in delivering satisfactory economic performance across models. One targeted nominal income, while the other targeted inflation and real output:

$$i - i^* = \theta_\pi(\pi - \pi^*) + \theta_q(q - q^*). \tag{6}$$

The potential usefulness of this particular rule as a benchmark for setting monetary policy was further highlighted in the celebrated contribution by John B. Taylor at the Fall 1992 Carnegie-Rochester Conference on Public Policy. Taylor developed a ‘hypothetical but representative policy rule’ (1993, p. 214) by using the sum of the equilibrium or natural rate of interest, r^*, and inflation, π, for i^* and setting the inflation target and equilibrium real interest equal to two and the response parameters to one half. The result was what became known as the classic Taylor rule:

$$i = 2 + \pi + \frac{1}{2}(\pi - 2) + \frac{1}{2}(q - q^*). \tag{7}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Taylor noted that, if one used the deviation of real quarterly output from a linear trend to measure the output gap, $(q - q^*)$, and the year-over-year rate of change of the output deflator to measure inflation, π, this parameterization appeared to describe Federal Reserve behaviour well in the late 1980s and early 1990s.

The confluence of the econometric evaluation evidence supporting the stabilization properties of this rule and its usefulness for understanding historical monetary policy in a period generally accepted as having good policy performance generated tremendous interest, and numerous central banks began to monitor this policy rule or related variants to provide guidance in policy decisions. These developments also greatly influenced monetary policy research and teaching. By linking interest rate decisions directly to inflation and economic activity, Taylor rules offered a convenient tool for studying monetary policy while abstracting from a detailed analysis of the demand for and supply of money. This allowed the development of simpler models (see the survey in Clarida, Gali and Gertler, 1999, and papers in Taylor, 1999) and the replacement of the 'LM curve' with a Taylor rule in treatments of the Hicksian IS–LM apparatus. (It should be noted, however, that this abstraction is overly simplistic when the short-term interest rate approaches zero. At the zero bound, the stance of monetary policy can no longer be measured or communicated with a short-term interest rate instrument; see, for example, Orphanides and Wieland, 2000). Subsequent research (see Orphanides, 2003b, for a survey) suggested that a generalized form of Taylor's classic rule could provide a useful common basis both for econometric policy evaluation across diverse families of models and for historical monetary policy analysis over a broad range of experience:

$$i = (1 - \theta_i)(r^* + \pi^*) + \theta_i i_{-1} + \theta_\pi(\pi - \pi^*) + \theta_q(q - q^*) + \theta_{\Delta q}(\Delta q - \Delta q^*). \quad (8)$$

The generalized Taylor rule (8) nests rule (6) as a special case but introduces two additional elements. First, it allows for inertial behaviour in setting interest rates, $\theta_i > 0$, which proves particularly important for policy analysis in models with strong expectational channels (Woodford, 2003). Second, it allows the policy response to developments in economic activity to take two forms: a response to the level of the output gap, $(q - q^*)$, or its difference, which can also be restated as a response to the difference between output growth and its potential, $(\Delta q - \Delta q^*)$. The generalized Taylor rule also nests another simplification of special interest, $\theta_i = 1$ and $\theta_q = 0$, which yields a family of difference rules similar to Wicksell's original proposal:

$$\Delta i = \theta_\pi(\pi - \pi^*) + \theta_{\Delta q}(\Delta q - \Delta q^*). \quad (9)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

These difference rules are also of interest because, like money-growth rules, their implementation does not require estimates of the natural rate of interest or the level of potential output (and the output gap) but only of the growth rate of potential output. Indeed, these rules may be viewed as a reformulation of money-growth rules in terms of an interest rate instrument. To see the relationship of (9) to money growth targeting, note that, by substituting the money growth in rule (3) into the equation of exchange, that rule can be stated in terms of the velocity of money:

$$\Delta v - \Delta v^{*} = (1 + \varphi_{\Delta x})(\Delta x - \Delta x^{*}). \quad (10)$$

To reformulate this strategy in terms of an interest rate rule, consider the simplest formulation of money demand as a (log-) linear relationship between velocity deviations from its equilibrium and the rate of interest. In difference form this is

$$\Delta v - \Delta v^{*} = a\Delta i + e, \quad (11)$$

where $a > 0$ and *e* summarizes short-run money demand dynamics and temporary velocity disturbances. An interest-rate-based strategy that avoids the short-run velocity fluctuations, *e*, may be obtained by substituting the remaining part of (11) into (10). This yields

$$\Delta i = \theta((\pi - \pi^{*}) + (\Delta q - \Delta q^{*})) \quad (12)$$

for some $\theta > 0$, which, as can be readily seen, has the same form as rule (9).

In light of this flexibility in nesting a wide range of alternative monetary policy strategies and the relative simplicity of the form (8), Taylor rules have been used to discuss a variety of policy regimes, from money growth targeting (see, for example, Clarida and Gertler, 1997) to inflation targeting (see, for example, Orphanides and Williams, 2007).

A crucial element for the design and operational implementation of a Taylor rule is the detailed description of its inputs. This requires specificity regarding the measures of inflation and economic activity that the policy rule should respond to, whether forecasts or recent outcomes of these variables

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
are to be employed, and the source of these data or forecasts. In addition, the source of information and updating procedures regarding the unobservable concepts required for implementing the rule must be stipulated. Specificity in these dimensions is essential for practical analysis because there is often a multitude of competing alternatives and a lack of consensus about the appropriate concepts and sources of information that ought to be used for policy analysis. This situation is particularly vexing in regard to the treatment of unobservable concepts, such as the output gap. Unfortunately, econometric policy evaluation exercises suggest that inferences regarding the performance of a particular Taylor rule often depend sensitively on assumptions regarding the availability and reliability of these inputs. Differences in underlying assumptions complicate comparisons across studies and often explain differences in reported findings.

An illustrative example of this sensitivity relates to improper treatment of information regarding the current state of the economy. A common pitfall in theoretical policy evaluation exercises is to assume that the current state of the economy – for example, the current output gap – can be perfectly observed. Under this assumption, a Taylor rule with a vigorous response to the output gap is often recommended as ‘optimal’ in model-based policy evaluations. However, naive adoption of such recommendations would be counterproductive. Available real-time estimates of the output gap are imperfect, and historical experience suggests that the mismeasurement is often substantial. Under these circumstances, better stabilization outcomes would result if policy did not respond to the output gap at all or if it responded to output growth instead (Orphanides, 2003a). If the natural rate of interest is also unknown and its real-time estimates are subject to significant mismeasurement, the difference variant of the Taylor rule, (9), proves considerably more robust than the Brookings variant, (6), reversing the ranking of the two alternatives that is implied under perfect knowledge (Orphanides and Williams, 2002).

Another example of such sensitivity relates to the use of forecasts in the Taylor rule. Because of lags in the monetary policy transmission mechanism, pre-emptive policy reaction is generally recommended, especially with respect to inflation. But inferences regarding the performance of forecast-based policy are sensitive to the quality of the forecasts. In some models, Taylor rules responding to several-quarters-ahead forecasts of inflation appear more promising for stabilization than rules focusing only on near-term conditions. However, this conclusion is not robust and is overturned once the potential unreliability of longer-term forecasts due to model misspecification is factored into the analysis (Levin, Wieland and Williams, 2003).

As already noted, Taylor rules have proven valuable for historical policy analysis. Following Taylor (1993), numerous authors have examined historical monetary policy in the United States using either calibrated or estimated versions of Taylor rules (8). Studying the characteristics of policy in periods associated with good or bad economic performance helps identify aspects of policy that may be associated with such differences in performance. A complicating factor is the need for real-time data and forecasts for proper inference (Orphanides, 2001). The pitfall of using *ex post* revised data and retrospective estimates of unobserved concepts in estimating Taylor rules is not uncommon. However, interpretations of historical policy based on information that was unavailable to policymakers when policy decisions were made is of questionable value. Policy prescriptions from a fixed rule are distorted as the inputs to the rule are revised from those originally available to policymakers, and therefore counterfactual comparisons of alternative policy rules can be misleading when they are based on revised data.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Despite these challenges, some useful elements of policy design emerge from historical analysis of Taylor rules, (8). First, and arguably most important, good stabilization performance is associated with a strong reaction to inflation. Second, good performance is associated with policy rules that exhibit considerable inertia. Third, a strong reaction to mismeasured output gaps has historically proven counterproductive. Fourth, successful policy could still usefully incorporate information from real economic activity by focusing on the growth rate of the economy. To be sure, such broad principles provide insufficient guidance for identifying the precise policy rule that might be ideal in a specific context. But this is not the objective of policy design with Taylor rules. Rather, the goal is the identification of simple guides that are robust to misspecification and other sources of error experienced over history.

In summary, Taylor rules offer a simple and transparent framework with which to organize the discussion of systematic monetary policy. Their adoption as a tool for policy discussions has facilitated a welcome convergence between monetary policy practice and monetary policy research and proved an important advance for both positive and normative analysis.

See Also

- inflation targeting

Bibliography

Bryant, R.C., Hooper, P. and Mann, C., eds. 1993. *Evaluating Policy Regimes: New Research in Empirical Macroeconomics*. Washington DC: Brookings.

Clarida, R., Gali, J. and Gertler, M. 1999. The science of monetary policy. *Journal of Economic Literature* 37, 1661–707.

Clarida, R. and Gertler, M. 1997. How the Bundesbank conducts monetary policy. In *Reducing Inflation: Motivation and Strategy*, ed. C. Romer and D. Romer. Chicago: University of Chicago Press.

Cooper, J.P. and Fischer, S. 1972. Simulations of monetary rules in the FRB–MIT–Penn model. *Journal of Money, Credit and Banking* 4, 384–96.

Friedman, M. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Levin, A., Wieland, V. and Williams, J.C. 2003. The performance of forecast-based monetary policy rules under model uncertainty. *American Economic Review* 93, 622–45.

McCallum, B.T. 1988. Robustness properties of a rule for monetary policy. *Carnegie-Rochester*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Conference Series on Public Policy 29, 173–203.

McCallum, B.T. 1993. Specification and analysis of a monetary policy rule for Japan. *Bank of Japan Monetary and Economic Studies* 11(2), 1–45.

Orphanides, A. 2001. Monetary policy rules based on real-time data. *American Economic Review* 91, 964–85.

Orphanides, A. 2003a. Monetary policy evaluation with noisy information. *Journal of Monetary Economics* 50, 605–31.

Orphanides, A. 2003b. Historical monetary policy analysis and the Taylor rule. *Journal of Monetary Economics* 50, 983–1022.

Orphanides, A. and Wieland, V. 2000. Efficient monetary policy design near price stability. *Journal of the Japanese and International Economies* 14, 327–65.

Orphanides, A. and Williams, J.C. 2002. Robust monetary policy rules with unknown natural rates. *Brookings Papers on Economic Activity* 2002(2), 63–145.

Orphanides, A. and Williams, J.C. 2007. Inflation targeting under imperfect knowledge. In *Monetary Policy under Inflation Targeting*, ed. F. Mishkin and K. Schmidt-Hebbel. Santiago: Central Bank of Chile.

Simons, H.C. 1936. Rules vs authorities in monetary policy. *Journal of Political Economy* 44, 1–30.

Snyder, C. 1935. The problem of monetary and economic stability. *Quarterly Journal of Economics* 49, 173–205.

Taylor, J.B. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39, 195–214.

Taylor, J.B., ed. 1999. *Monetary Policy Rules*. Chicago: University of Chicago.

Wicksell, K. 1898. *Interest and Prices*. Trans. R.F. Kahn, London: Macmillan, 1936.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton: Princeton University Press.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Orphanides, Athanasios. "Taylor rules." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000215>
doi:10.1057/9780230226203.1686

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Taylor, Fred Manville (1855 – 1932)

Daniel R. Fusfeld
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

market socialism; Taylor, F. M.

Article

Taylor made his chief contribution to economic theory in his 1928 presidential address to the American Economic Association, in which he laid out the basic principles of market socialism (Taylor, 1929). He argued that rational allocation of resources could be achieved in a socialist state if three conditions are met: citizens obtain income from the state in exchange for services; income is freely spent on goods offered for sale by the state at given prices; prices are set at full costs of production. The third condition can be met through a trial-and-error method in which prices of factors of production are set at levels that clear the market. Given these costs and consumer demand, markets for finished products can be cleared by adjusting levels of output and inventories. Such a system could achieve results similar to those of a competitive private enterprise economy.

Taylor's doctorate was in philosophy from the University of Michigan (1888). He taught at Albion College 1879–92, and in the Department of Economics at Michigan from 1892 to 1929. A strong advocate of laissez-faire policies and the gold standard, Taylor was a noted expositor of economic theory, with emphasis on Marshallian partial equilibrium analysis, analytic rigour and a libertarian ideology. His *Principles* textbook (Taylor, 1911) went through nine editions from 1911 to 1925.

Selected works

1911. *Principles of Economics*. Ann Arbor: University of Michigan Press, 1918; New York: Ronald Press, 1921. 9th edn, 1925.

1929. The guidance of production in a socialist state. *American Economic Review* 19(1), 1–8 Reprinted in *On the Economic Theory of Socialism*, ed. B.E. Lipincott, New York: McGraw-Hill, 1938.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Fusfeld, Daniel R. "Taylor, Fred Manville (1855–1932)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000028>
doi:10.1057/9780230226203.1687

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Taylorism

A.L. Friedman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

scientific management; soldiering; Taylor, F. W.; Taylorism; time study

Article

Taylorism refers to the system of management developed by Frederick Winslow Taylor. Taylor called his system ‘scientific management’. Scientific management is clearly described in Taylor's two most famous works, *Shop Management* (1903) and *The Principles of Scientific Management* (1911). Scientific management is based on the following principles:

1. 1. Management gathers and systematizes all the workers’ traditional knowledge.
2. 2. All possible ‘brainwork’ is removed from the shop and centred in the planning or layout department.
3. 3. The work should be divided into its simplest constituent elements: the tasks. Management should try to limit individual ‘jobs’ to a single task as far as possible.
4. 4. Managers should specify the tasks to be done in complete detail. These tasks should be presented to the worker in written form. They should note not only what is to be done, but also how it is to be done and the exact time allowed for doing it.
5. 5. The work should be monitored closely.

Taylor's techniques for gathering information about work was time study, the measurement of elapsed time for each component operation of a work process. Taylor also recommended that the foreman's job should be divided into more simplified task collections. Shop-floor foremen should be divided into the setting-up boss, speed boss, quantity inspector and repair boss. However, the main division was to separate work-design and manning-level decisions away from shop-floor foremen and to the planning department.
The purpose of Taylor's system was to eliminate ‘soldiering’, or low worker effort. This could either take the form of natural soldiering, the natural instinct and tendency for men to take it easy, or

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
systematic soldiering, the calculated reduction of effort arising from actions and communication among groups of workers. The ultimate cause of both forms of soldiering for Taylor ‘lay in the ignorance of the management as to what really constitutes a proper day's work for a workman’ (1911, p. 53). Once this was determined ‘scientifically’, workers would be forced to comply with this standard by careful monitoring of their performance and by a differential piecework payment system. A target rate of work would be determined by work study. If workers exceeded this target, they would receive a bonus, but bonus payments would reach a ceiling between 30 per cent and 100 per cent of the standard work rate. If workers failed to meet their targets, they would lose earnings.

There is a wide range of opinions as to the importance of Taylorism. For some, Taylorism represents the dominant theory and practice of 20th-century management (Drucker, 1954; Braverman, 1974). For others, Taylorism is viewed as having widespread ideological impact, but not much influence on practice because it was successfully resisted by workers and was too expensive for managers (Edwards, 1979). Finally, there are those who view Taylorism as an expression of important changes in management practices, but that moves towards Taylorism have not been universal. The application of Taylorism is contingent on environmental factors (Friedman, 1977; Littler, 1982).

Bibliography

Braverman, H. 1974. *Labor and Monopoly Capital*. New York: Monthly Review Press.

Drucker, P. 1954. *The Practice of Management*. New York: Harper & Row.

Edwards, R. 1979. *Contested Terrain*. London: Heinemann.

Friedman, A.L. 1977. *Industry and Labour*. London: Macmillan.

Littler, C.R. 1982. *The Development of the Labour Process in Capitalist Societies*. London: Heinemann.

Taylor, F.W. 1903. Shop management. In *Scientific Management*, F.W. Taylor. London: Harper & Row, 1964.

Taylor, F.W. 1911. The principles of scientific management. In *Scientific Management*, F.W. Taylor. London: Harper & Row, 1964.

How to cite this article

Friedman, A. L. "Taylorism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000030> doi:10.1057/9780230226203.1688

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

teams

Roy Radner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A team consists of a number of decision-makers, with common interests and beliefs, but controlling different decision variables and basing their decisions on (possibly) different information. The economic theory of teams is concerned with (1) the allocation of decision variables and information among the team members, and (2) the characterization of efficient decision rules, given the allocation of tasks and information. The theory of teams thus addresses a middle ground between the theory of individual decision under uncertainty and the theory of games, and provides a natural framework for the analysis of mechanisms for decentralization, including market-like mechanisms.

Keywords

decisions under uncertainty; demand price; efficient allocation; game-playing; games; information costs; informational decentralization; optimal decision functions; probability; signalling; socialism; teams

Article

The economic theory of teams addresses a middle ground between the theory of individual decision under uncertainty and the theory of games. A *team* is made up of a number of decision-makers, with common interests and beliefs, but controlling different decision variables and basing their decisions on (possibly) different information. The theory of teams is concerned with (1) the allocation of decision variables (tasks) and information among the members of the team, and (2) the characterization of efficient decision rules, given the allocation of tasks and information.

For example, in the pre-computer age, airline companies had a number of ticket agents who were authorized to sell reservations on future flights with only partial information about what reservations had been booked by other agents. A team-theoretic issue would be the characterization of best rules for those agents to use under such circumstances, taking account of the joint probability distribution of demands for reservations at the different offices, the losses due to selling too many or too few reservations in total, and so forth. A second issue would be the calculation of the increase in expected profit that would

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
be obtained by providing each agent with better information about the status of reservations at all offices, or by increased centralization of the reservation process. To calculate this increase in expected profit – the *value* of the additional information – one of course needs to know something about the best decision rules with and without the additional information. However, providing this additional information would require additional communication, transmission, processing, and storage, all of which would be costly. The value of the information puts an upper bound on the additional cost that should be incurred. The value – and cost – of the information will depend on its structure and on the structure of the team's decision problem, and not just on some simple measure of the ‘quantity’ of information. (For a study of the airline reservation problem, and other models of sales organization, from a team-theoretic point of view, see Beckmann, 1958 and McGuire, 1961, respectively.)

In this entry we shall sketch a formal model of team theory, the characterization of optimal team decision functions, and the evaluation of information in a team. The theory will be illustrated with a discussion of decentralized resource allocation, followed by concluding remarks on the incentive problem.

A formal model

We consider a *team* and *M members*. Each member *m* controls an *action*, say a_m. The resulting utility to the team depends on the *team action*,

$$a = (a_1, \ldots, a_M),$$

and on the *state of the environment*. (Since the team members have common interests, there is a single utility for the whole team.) The state of the environment comprises all the variables about which team members may be uncertain before choosing their actions. It is determined exogenously, i.e. is not subject to the control or influence of the team members. If we denote the state of the environment by *s*, then we can denote the utility to the team by $u(a, s)$; the function *u* will be called the *payoff* function for the team. Before choosing an action, each team member *m* receives an information signal, y_m. This information signal is determined by the state of the environment, say $y_m = \eta_m(s)$. (This includes the case of ‘noisy’ information, if the description of the state of the environment includes a description of the noise.) We shall call η_m the *information* function for member *m*, and the *M*-tuple $\eta = (\eta_1, \ldots, \eta_M)$ will be called the *information structure* of the team.

Each team member *m* will choose his action on the basis of the information signal he receives, according to a decision function, say α_m Thus

$$a_m = \alpha_m(y_m) = \alpha_m(\eta_m[s]). \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

If we use the symbol α to denote the *team decision* function, i.e., the *M*-tuple of individual decision functions, then the utility to the team, in *state s*, of using the information structure η and decision function α can be expressed as

$$U(s) = u[\alpha(\eta[s]), s], \tag{2}$$

To express the team's uncertainty about the state of the environment, we suppose that *s* is determined according to some probability distribution, ϕ, on the set *S* of possible states. This probability distribution may be interpreted as ‘objective’ or ‘personal’; in the latter case it represents the beliefs of the team members (Savage, 1954). It is part of the definition of a team that its members have common beliefs, as well as common utility functions.

With the state *s* distributed according to the probability distribution ϕ, the utility $U(s)$ in (2) is a random variable. We shall assume that the team chooses its decision function so as to maximize the (mathematical) expectation of this utility,

$$E[U(s)] = \sum_s \varphi(s)U(s) \equiv \omega(\alpha, \eta, \varphi). \tag{3}$$

As a special case, suppose that the information functions of the team members are identical. In this case, the team decision problem is formally identical to a one-person decision problem in which the same person controls all of the actions. An alternative interpretation of this case is that the information is *centralized*. By contrast, if at least two team members have essentially different information functions, then we may say that the information is *decentralized*. With this definition of (informational) decentralization, we see that all organizations but the very smallest are likely to be decentralized to some extent.

The expected utility for the team depends on the team members’ decision function, the team information structure, and the probability distribution of states of the environment, as well as on the ‘structure’ of the decision problem, i.e. the way in which the utility (2) depends on the members’ action and the state of the environment. This is brought out by the notation in (3). If we want to compare the usefulness of two different information structures, we have to associate with them some corresponding decision function α and probability distribution ϕ. Since the probability distribution (sometimes called the ‘prior distribution’) represents either objective probabilities or the team members’ common beliefs before they receive further information, it is natural to take it as a datum of the problem. On the other hand, since the decision functions can be chosen by the team, it is natural to associate with each information structure

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the corresponding team decision function that maximizes its expected utility (given the information structure). Thus the optimization problem for the team may be posed in two stages: (1) for a given information structure, characterize the optimal team decision function(s); (2) optimize the information structure, taking account of the costs of – or constraints on – making the information available, and with the proviso that for each information structure the team uses an optimal decision function.
More will be said below about each of these stages of the problem. However, it should be emphasized here that the choice of information structure comprises most of the organizational design choices that are not concerned with conflicts of interests or beliefs among the organization's members. The information structure is, of course, affected by the pattern of observation and communication in the team. In addition, the allocation of tasks within the team is expressed by the information structure. To see this, suppose that each member of the team were assigned an information structure; then a reassignment of decision variables to team members would be formally equivalent to reassigning information functions to decision variables.

Optimal decision functions

We shall now consider the characterization of team decision functions that are optimal for a given structure of information. It will be useful to recall here the corresponding problem for a single-person decision problem (see, e.g., Marschak and Radner, 1972, ch. 2). We may use the same model and notation as in the previous section, but remembering that there is only one member of the team. The following statement provides a general characterization of the optimal decision function: *For each information signal, choose an action that maximizes the conditional expected utility given the particular signal*.
This characterization is easily derived from equations (2) and (3). In equation (3), group the terms in the sum according to the information signal associated with each state; this gives us

$$E[U(s)] = \sum_{y} \sum_{\eta(s)=y} \varphi(s) U(s). \tag{4}$$

From (2), if $\eta(s) = y$, then the resulting utility in that state is

$$U(s) = u[\alpha(y), s]. \tag{5}$$

For each signal *y*, the decision-maker can choose an action $a = \alpha(y)$. Hence, combining (4) and (5) we see that, for each signal *y*, the decision-maker should choose $a = \alpha(y)$ to maximize

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\sum_{\eta(s)=y} \varphi(s)u(a, s).$$
(6)

Let $\psi(y)$ denote the probability of y, and let $\phi(s|y)$ denote the conditional probability of s given y. By definition,

$$\psi(y) = \sum_{y=\eta(s)} \varphi(s),$$

and if $y = \eta(s)$,

$$\varphi(s|y) = \frac{\varphi(s)}{\psi(y)},$$

or

$$\varphi(s) = \psi(y)\varphi(s|y).$$

Hence (6) can be written as

$$\psi(y) \sum_{\eta(s)=y} \varphi(s|y)u(a, s),$$
(7)

so that maximizing (6) is equivalent to maximizing

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\sum_{\eta(s)=y} \varphi(s|y)\,u(a, s),$$

(8)

which we recognize as the conditional expected utility using the action *a*, given the signal *y*. This proves the above characterization of the best decision function. (Notice that we have implicitly assumed that the signal has positive probability. There is no loss of generality in doing so; we can simply exclude from consideration all signals that have zero probability, since they do not affect the expected utility.)

The characterization of optimal single-person decision functions can be extended to the case of a team, but in a restricted way. Consider a particular team member *i*. If a team decision function, say $\hat{\alpha}$, is optimal, then surely *i*'s decision function α_i is optimal given that each other member *j* uses $\hat{\alpha}_j$. Hence *i* is faced, so-to-speak, with a one person decision problem in which the other members' decision functions form part of *i*'s 'environment'. The following is therefore a *necessary* condition for a team decision function to be optimal:

Person-by-person-optimality condition

For each member *i*, and for each signal y_i with positive probability, the corresponding action $a_i = \alpha_i(y_i)$ maximizes the team's conditional expected utility given the signal, y_i and the decision functions of the other members.

Although person-by-person-optimality is necessary for optimality, it need not be sufficient. However, one can prove the following:

Theorem 1: If each member's action is a real finite-dimensional vector chosen from some open rectangle, and if for each state *s* the team's utility is a concave and differentiable function of the team action, then any team decision function that is person-by-person-optimal is also optimal. (For a proof of this theorem, and an example in which a person-by-person-optimal decision function fails to be optimal, see Marschak and Radner, 1972, ch. 10, s. 3.; for a more complete treatment, see Radner, 1962.)

The person-by-person-optimality condition can be applied to yield more detailed characterizations of optimal team decision functions for special cases, e.g., in which the utility function is quadratic or piecewise-linear (see Marschak and Radner, 1972, ch. 10). A few such applications are illustrated below.

The evaluation of information

As noted at the beginning of this discussion, many of the most interesting qsts in organizational design concern the comparison of alternative information structures. One information structure is better than another to the extent that it permits better decisions; on the other hand, this improvement can be obtained only at some additional cost.

We first consider the case in which the utility from the decisions is additively separable from the cost of the information structure; we shall call this the *separable case*. In this case, one is justified in defining

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the *gross value* of an information structure as the difference between (1) the expected utility derived from its best use and (2) the maximum utility obtainable using no information (beyond that contained in the prior probability distribution of states).

If the team has no information (the null information structure), then its decision function reduces to a single team action. The maximum expected utility that the team can obtain with the null information structure is

$$V_0(\varphi) = \max_a \sum_s u(a, s). \tag{9}$$

Hence in the separable case, the gross value of an information structure η is defined as

$$V(\eta, \varphi) \equiv \max_\alpha \omega(\alpha, \eta, \varphi) - V_0(\varphi). \tag{10}$$

(Cf. equation (3).)

Note that the value of an information structure depends on the prior distribution, as well as on the entire structure of the decision problem (available actions, utility function, etc.). This should make one suspect that there is no way to tell whether one information structure is more valuable than another just by examining the two information structures alone.

To examine this question more carefully, it is useful to introduce another representation of information. Consider again for the moment the single-person case; an information structure then consists of a single function from states to information signals. For any given signal, there is a set of states that give rise to that signal. This correspondence between signals and sets of states determines a partition of the set of states; each element of the partition is a set of all states that lead to a particular signal; denote this partition by (S_y). It is obvious that any two information structures that give rise to the same partition are equivalent from the point of view of the decision-maker, and in particular must have the same value. In other words, the names or labels of the signals are unimportant.

Suppose now that the set S of states of the environment, the number M of team members, and the set A of team actions are fixed. Consider the family of all team decision problems that can be formulated with the given triple (S, M, A). In other words, consider the set of all pairs (u, ϕ), where u is a utility function for the team, and ϕ is a prior distribution, compatible with (S, M, A). We shall say that one information structure *is as valuable* as another information structure if the value of the first is greater than or equal to the value of the second for all team decision problems compatible with (S, M, A). (Value is defined by equation (10).)

The following criterion provides a simple test for the relation ‘as valuable as’. Of two partitions of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
set *S*, we shall say the first is as *fine* as the second if every element of the first partition is a subset of some element of the second (the first can be obtained by 'refining' the second). Let $\eta = (\eta_1, \ldots, \eta_M)$ and $\chi = (\chi_1, \ldots, \chi_M)$ be two team information structures. We shall say that η is as fine as χ if for every team member *m*, η_m is as fine as χ_m. One can prove (see Marschak and Radner, 1972, Ch. 2, Sec. 6):

Theorem 2: Assume that every team member has at least three alternative actions; then the information structure η is at least as valuable as the information structure χ if and only if, for each member *m*, η_m is as fine as χ_m.

Theorem 2 can be extended to deal with 'noisy' information (see, e.g., McGuire, 1972).

Since two partitions of a set need not be ranked by the relation 'as fine as', it is clear from Theorem 2 that the relation 'as valuable as' is only a partial ordering of information structures. This implies, in particular, that there is no numerical measure of 'quantity of information' that can rank all information structures in order of value, independent of the decision problem in which the information is used.

If the utility of the team decision and the cost of information are not additively separable, then an alternative definition of value of information must be used. For example, suppose that the outcome of the team decision and the cost of the information structure are both measured in dollars, and the the team is not risk-neutral, so that the team utility is some (nonlinear) function of the outcome and the cost. Then we can define the value of the information structure as the 'demand price', i.e., the smallest cost that would make the team indifferent between using the information structure and having no information beyond the prior distribution. (For further discussion of the comparison of information structures, see McGuire, 1972. For more on the value of and demand for information, see Arrow, 1972.)

Decentralization

We have used the term informational decentralization to refer to a structure of information in which not all members have the same information function. In an economic organization the information structure is generated by processes of observation, communication, storage, and computation. For example, suppose that each team member *m* starts by observing a different random variable, say $\zeta_m(s)$. If there were no communication among the members before actions were taken, then each member's information would be the same as his observation – an extreme form of decentralization. On the other hand, if there were complete communication of their observations among the members, then their information functions would be identical, namely $\zeta = (\zeta_1, \ldots, \zeta_M)$. Alternatively, the latter information structure could be generated by having all members communicate their observations to a central agency, which would then compute the team action and communicate the corresponding individual action to each member. In the last two cases, we would say that the information structure is completely centralized, because all of the members' actions were based on the same information.

Rarely does one encounter in a real organization the extremes of no communication or complete communication just described. Rather, one finds that numerous devices are used to bring about a partial exchange of information. The usefulness of such devices is measured by the excess of additional value (expected utility) they contribute over the costs of installing and operating them. Examples of such devices are the dissemination of reports and instructions, the formation of committees and task forces, and 'management by exception'. Formal models and a comparative analysis of some of these devices are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
given in (Marschak and Radner, 1972, ch. 6). In particular, this methodology is used to elucidate the value of two different forms of management by exception.

Allocation of resources in a team

For many economists, the purely competitive market represents the ideal model of economic decentralization. Indeed, in some economic literature, ‘decentralization’ and ‘pure competition’ are synonymous. The potential usefulness of market-like mechanisms to decentralize economic decision-making in a socialist economy has also been discussed by students of socialism (Lange and Taylor, 1938; Lerner, 1944; Ward, 1967).

The theory of teams provides a natural framework for the analysis of market mechanisms as a device for decentralization. For example, consider the problem of allocating resources to productive enterprises. Suppose that some resources are initially held centrally by a ‘resource manager’. Before any exchange of information, the resource manager observes the supplies of centrally available resources, and each enterprise manager observes his respective local conditions of production: technology, supplies of local resources, etc. The action of the resource manager is to allocate the central resources among the enterprises. The action of an enterprise manager includes (say) the choice of techniques and the levels of inputs of local resources. The state of the environment comprises the total supplies of central resources and the local conditions.

At one extreme, the team action could be taken without any communication. In particular, the central resources would be allocated based only on the prior probability distribution of local conditions. Regarding the supplies of central resources, each enterprise manager would know only the prior probability distribution of such supplies, and the allocation rule to be used by the resource manager. We might call the resulting information structure ‘routine’.

At the other extreme (complete centralization), each enterprise manager might be required to report to the resource manager all of his information about local conditions. The resource manager would then compute both the optimal decisions of the enterprises and the optimal allocation of resources. Accordingly, the resources would be allocated and the enterprise managers would then be ‘instructed’ by the resource manager as to what actions they should take.

In a market mechanism, the resource manager would announce prices (of central resources), and the enterprise managers would respond with demands. In the literature on allocation and price-adjustment mechanisms it is usually assumed that this exchange of messages is iterated until an equilibrium of supply and demand is reached (this may require infinitely many iterations!). In a real application of such a mechanism, only a few iterations would typically be feasible, and equilibrium would not be reached. Thus one could not appeal to the theory of optimality of the equilibria of such processes. Nevertheless, the exchange of information produced by even a few iterations might be quite valuable, i.e., the information structure might be much more valuable than the ‘routine’ structure, and possibly close in value to that of complete centralization.

Indeed, research done to-date on models of such processes suggests that price and demand signals are strikingly efficient in conveying the information needed for good allocation decisions, even out of equilibrium (Radner, 1972; Groves and Radner, 1972; T.A. Marschak, 1959, 1972; Arrow and Radner, 1979; Groves and Hart, 1982; Groves, 1983; Hogan, 1971).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Incentives in teams

The model of a team assumes that the team members have identical interests and beliefs. Thus no special incentives are required to persuade the individual members to honestly implement the given information structure or to take the decisions prescribed by the optimal team decision function. A full-fledged theory of economic organization should, of course, take account of conflicting interests and beliefs, and the resulting problems of incentives.

This article is not the place to review the growing literature on this subject, but a few comments may be useful here. In general, it is not possible to solve the 'incentive problem' costlessly. (For exceptions to this generalization, see Groves, 1973; Green and Laffont, 1979.) Thus, in an economic organization, there will be two sources of efficiency loss: (1) decentralization of information, having the effect that individual actions will be based on information that is less complete than the information jointly available to the organization as a whole; (2) conflicts of interests and beliefs among the decision-makers, leading to distortions of information and action ('game-playing'). In fact these two sources are not so easily disentangled. For example, under conditions of uncertainty and limited information, it will typically be difficult for a supervisor (or organizer) to determine whether a particular decision-maker is providing correct information or following a prescribed decision rule, since to achieve this would require the supervisor to have all of the information that is available to the subordinate. In other words, informational decentralization leads to de facto decentralization of authority. (For references to the literature on incentives and decentralization in economic organizations see Arrow, 1974; Hurwicz, 1979; Radner, 1975, 1986; and Stiglitz, 1983.)

See Also

- efficient allocation
- signalling and screening

Bibliography

Arrow, K.J. 1972. The value of and demand for information. Ch. 6 of McGuire and Radner (1986).

Arrow, K.J. 1974. *The Limits of Organization*. New York: Norton.

Arrow, K.J. and Radner, R. 1979. Allocation of resources in large teams. *Econometrica* 47, 361–85.

Beckmann, M.J. 1958. Decision and team problems in airline reservations. *Econometrica* 26, 134–45.

Green, J. and Laffont, J.-J. 1979. *Incentives in Public Decision-Making*. Amsterdam: North-Holland.

Groves, T. 1973. Incentives in teams. *Econometrica* 41, 617–31.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Groves, T. 1983. The usefulness of demand forecasts for team resource allocation in a stochastic environment. *Review of Economic Studies* 50, 555–71.

Groves, T. and Radner, R. 1972. Allocation of resources in a team. *Journal of Economic Theory* 3, 415–44.

Groves, T. and Hart, S. 1982. Efficiency of resource allocation by uninformed demand. *Econometrica* 50, 1453–82.

Hogan, T.M. 1971. A comparison of information structures and convergence properties of several multisector economic planning procedures. Technical Report No. 10, Center for Research in Management Science, University of California, Berkeley.

Hurwicz, L. 1979. On the interaction between information and incentives in organizations. In *Communication and Control in Society*, ed. K. Krittendorf. New York: Gordon and Breach, 123–47.

Lange, O. and Taylor, F.M. 1938. *On the Economic Theory of Socialism*. Minneapolis: University of Minnesota Press.

Lerner, A. 1944. *The Economics of Control*. New York: Macmillan.

Marschak, J. and Radner, R. 1972. *Economic Theory of Teams*. New Haven: Yale University Press.

Marschak, T.A. 1972. Computation in organizations: the comparison of price mechanisms and other adjustment processes. Ch. 12 of McGuire and Radner (1986), 237–82.

Marschak, T.A. 1959. Centralization and decentralization in economic organizations. *Econometrica* 27, 399–430.

McGuire, C.B. 1961. Some team models of a sales organization. *Management Science* 7, 101–130.

McGuire, C.B. 1972. Comparisons of information structures. Ch. 5 of McGuire and Radner (1986), 101–30.

McGuire, C.B. and Radner, R. 1986. *Decision and Organization*. 2nd edn, Minneapolis: University of Minnesota Press; originally published Amsterdam: North-Holland, 1972.

Radner, R. 1972. Allocation of a scarce resource under uncertainty: an example of a team. Ch. 11 of McGuire and Radner (1986), 217–36.

Radner, R. 1962. Team decision problems. *Annals of Mathematical Statistics* 33, 857–881.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Radner, R. 1975. Economic Planning under uncertainty. Ch. 4 of *Economic Planning, East and West*, ed. M. Bornstein. Cambridge, Mass.: Ballinger, 93–118.

Radner, R. 1987. Decentralization and incentives. In *Information, Incentives, and Economic Mechanisms: Essays in Hornor of Leonid Hurwicz*, ed. T. Groves, R. Radner and S. Reiter. Minneapolis: University of Minnesota Press.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Stiglitz, J.E. 1983. Risk, incentives, and the pure theory of moral hazard. *The Geneva Papers on Risk and Insurance* 8, 4–33.

Ward, B. 1967. *The Socialist Economy*. New York: Random House.

How to cite this article

Radner, Roy. "teams." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000031> doi:10.1057/9780230226203.1689

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

technical change

S. Metcalfe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Successive transformations of economic society from agricultural to industrial form and beyond to the service economy have consolidated a process of economic change with its own inner logic of tremendous power, a logic which harnessed continual technical and organizational innovation to the pursuit of profit. The intertwining of emergent knowledge and economic adaptation within the instituted frame of modern capitalism is at the core of economic self-transformation. This article treats this topic in three parts: the relation between new knowledge and economic transformation; some consequences of technical change; and the residual productivity debate.

Keywords

additivity; adjustment costs; capital intensity; capitalism; capital–labour substitution; consumption growth frontier; diffusion of technology; division of labour; economic growth; efficient allocation; entrepreneurship; equilibrium; Industrial Revolution; innovation; input–output analysis; Internet; invention; knowledge; labour productivity; learning; leisure; market institutions; maturity; population growth; production functions; productivity growth; research and development; residual productivity; roundabout methods of production; specialization; standards of living; structural change; technical change; thrift; total factor productivity; trigger effects; wage–profit frontier

Article

The successive transformations of economic society since the 14th century from agricultural to industrial form and beyond to the service economy have consolidated a process of economic change with its own inner logic of tremendous power, a logic which harnessed continual technical and organizational innovation to the pursuit of profit. At the core of the new logic is the intertwining of emergent knowledge and economic adaptation to its hidden possibilities that has been the basis for the sustained increase of aggregate output per person employed – the chief proximate source of increased standards of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
living in the Western world – the progressive mechanization and automation of production methods, and the continuous development of the economic structure (Kuznets, 1977; Mokyr, 1990; 2002). The gains in material welfare, in length of human life, in life experience and functioning have been beyond anything achieved before the 18th century, yet knowledge-driven progress comes at a price. It is necessarily uneven in its incidence across space and time, and the ensuing disparities of performance can and do impose heavy human adjustment costs as old ways give ground to the new. Skills are devalued, capital assets lose the capacity to generate income, while there is little prospect of the losers receiving compensation. If the balance sheet speaks to progress, it does so in a tangled way. This is the ethic of competition, and nowhere is this unevenness more apparent than in the seemingly unavoidable differences in economic performance between advanced and developing countries. Knowledge-driven economic growth is never a smooth, balanced affair of proportional expansion with each activity advancing in step. Rather, as Schumpeter insisted, it involves disharmony and fierce competition between new and old activities and places, a diversity of growth rates and profit rates and continual reallocation of labour and capital between and within activities. It is occasionally useful to study such processes as if structural change were absent, but to do so courts the danger of missing the substance and therefore the process and significance of technical and economic change. For the very process of uneven adjustment is a powerful stimulus to the development of further knowledge. This is the Faustian bargain accepted by the Western world with its origins in Reformation and Renaissance and its consolidation in the 19th and 20th centuries.

What was the nature of this emergent combination of knowledge-generating system and market system of economic adaptation? All productive activity involves the transformation of materials and energy, in a purposeful, intelligent and information-dependent fashion to add value to the materials and energy. These transformations are of physical form, or of spatial location, or of availability in time; indeed, the history of technical change is a history of invention and innovation directed to providing new inanimate energy sources, to providing means to control the application of energy, and to providing new, synthetic materials on which to work. What economists see from one perspective as the substitution of capital for labour in more roundabout methods of production is from this perspective the substitution of non-human for human energy, a process that was well established through the use of water power in the early middle ages and went on to be revolutionized by innovations in steam power, internal combustion and ultimately electricity. All existing industries were affected by the new power sources, (textiles, iron and steel, mining in particular) in the first Industrial Revolution, but new industries emerged, too, to produce the machines of increasing sophistication and specialization to harness the new power sources and to extend their application to transport and communication. That much product innovation was required to exploit the new possibilities warns us that the relation between product and process technology is usually close. That new forms of business organization were needed to deal with increased capital intensity and new risks tells us that technical and organizational change were seldom far removed from each other. The increasing incentives to exploit the material base of the planet and the ability to synthesise materials not found as natural compounds gave further licence to the innovation process with oil, aluminum, plastics and pharmaceuticals, each becoming commonplace in the 20th century. But it would be a mistake to focus attention on the great, traditional industries alone, whether producing capital goods, consumer goods or intermediate products; trends of a different nature reflected the deeper ways in which new knowledge was working its transformative effects. Superficially this is seen in terms of the displacement of manual labour as an energy supply, but this misses the point that the human role was

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
moving increasingly into one of the management and coordination of information that is essential for any rational activity. Within business enterprises, within public bureaucracies and within markets, greater human effort was required to manage and coordinate the flows of information demanded by economic growth and its handmaiden – a richer and deeper division of labour. That this could be possible only in the presence of a productivity-enhancing revolution in information generating, storing and communicating activities should be obvious at least from our position in the early 21st century. In this regard, the innovation sequence and economic adaptations associated with the printing press proved to be of profound importance: for it eliminated a long-standing constraint on the exact reproduction of information, and made possible its transmission over generations (storage), and its more ready transport. Written communication increased in relative importance compared with face-to-face conversation, the costs associated with codification declined dramatically and, consequentially, the organization of a spatially distributed but deeply connected process of the growth of science and technology became possible (Eisenstein, 1979). The subsequent developments of telegraph and telephone, of information-processing machines, of television and radio, and now of the Internet are amplifications of the revolution begun by printing. Since all societies are knowledge based, that epithet counts for little: what matters is the form of information society that prevails, and the way in which inanimate energy and its harnessing machinery has been applied progressively to further transform the production, transport and storage of information is a development at least as significant as the discovery of economical steam power in Watt's day. The growth in the productivity of information activities made possible in manufacturing, transport and communication, and now service activities, is accepted by all commentators as immense, for it has economized on scarce mental capacity, not just on limited human strength. Quite remarkably, total employment has continued to rise even though its occupational composition, like the composition of economic output, is continually shifting in response to new technological possibilities, despite the prognostications of less confident observers of modern capitalism. Modern capitalism appears, by accident no doubt, to have evolved a set of institutional rules which not only promotes the efficient use of what is known to further specific human ends but serves also to greatly stimulate the production of further new knowledge. The solution of one problem through the speculative deployment of imagination serves only to open up further problems, while the future direction and outcomes of this process remain necessarily hidden from view. Knowledge and economy are deeply intertwined, and the restless nature of the one reinforces the restless nature of the other. This is the nature of that Faustian bargain between a knowledge system and the market process.

As the above sketch might warn the reader, the economic analysis of technical progress is not a straightforward matter. The familiar tools of equilibrium economics are best suited to discussing the long-run effects of new products and methods of production; they are not well suited to analysis of the disequilibrium processes by which new technologies are generated, improved and absorbed into the economic structure. All these processes take time, operate with different velocities and are subject to complicated interactions and feedback effects. It has been traditional to divide the analysis of technical change into three branches: *invention*, the creation of new products and processes; *innovation*, the transfer of invention to commercial application; and *diffusion*, the spread of innovation into the economic environment. Unfortunately, this has provided a somewhat fragmented approach to the study of technical change in which an understanding of interdependence and feedback between the stages has been hidden, together with important elements which emphasize the continuity of advance.

A more unified approach is possible if we place the growth of knowledge and its instantiation in novel

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
innovations in the context of a competitive market process, in which firms seek to differentiate themselves and gain commercial advantages by introducing new products/services and processes. That the market system is a defining instituted feature of modern capitalism goes a long way to explaining its experimental proclivities and thus the importance to it of enterprise. It is an open system in which, in principle, all existing constellations of resources are open to challenge by rival, entrepreneurial conjectures according to their anticipated profitability. Innovations (like breakthroughs in science) are always statements of disagreement, in this case about the efficacy of the prevailing allocation of resources. Thus there is an important boundary in play, to work the system needs order and agreement, to progress it needs disorder and disagreement, it is the institutions of capitalism that contribute to keeping the two in balance. Of course, firms are not the only important source of new knowledge; they are only one element in a more comprehensive system of public and private research institutes of many kinds; but with respect to innovation (as distinct from invention) they play a virtually dominant role as the one combinatorial agency bringing together the different kinds of problem solving and exploiting resources needed to create wealth from knowledge. The way in which they combine internal with external innovation resources is an important aspect of their innovative ability, as Marshall knew well; external organization, connecting to customers, suppliers and other sources of knowledge, is as much a part of a firm's productive capital as is its internal organization. The three questions which arise from this viewpoint are: (*a*) by what processes is technological variety generated?; (*b*) by what processes do different varieties acquire economic weight?; and, (*c*) by what mechanisms does the process of acquiring economic weight shape the development of technological variety?

To the first question belongs the study of a firm's strategy in changing its knowledge base and articulating new and improved products and processes. The role of science in modern invention, the organization of R&D activity, and the links between a firm and other knowledge-generating institutions each play a role with respect to variety and its generation. But no individual variety of product or process is significant until it acquires economic weight, and the greater the weight the greater the impacts of the new technology upon its environment. In regard to the second question, innovations acquire economic significance because they are superior either from the point of view of users or from the point of view of their producers or both. Clearly, however, the more profitable it is to use new products and processes and the more profitable it is to supply them, the more quickly will they acquire economic weight and displace existing products and processes. The dynamics of adjustment to new opportunities depend on how different the new technologies are from established forms, and how the economic environment evaluates those differences relative to the standards of value and cost. The third question contains some of the most complex questions of all, relating to the inducement mechanisms which generate and shape technological variety. Clearly there are important non-economic factors at work. However, different environments for market exploitation do make it profitable to develop a technology in different directions, and the experience of exploiting a technology in a given environment, more often than not, gives rise to important learning effects which indicate an agenda for subsequent development and applications in other areas of activity. Technologies do not emerge into the economic sphere fully fledged but typically in immature form, and evolve very much according to the bottlenecks and incentives to development which arise in their application. Progress thus tends to be localized around a canalized path of advance and to be contingent on factor prices and the values consumers place on different combinations of functional features. The same technological opportunity exploited in different environments would in all probability develop in different directions. By a similar token, a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
technology which is mature in one environment may be developing rapidly in another. Maturity is at root an economic concept applicable to situations where the expected benefits fall short of the expected costs of further advancing technology.

Within the development of economic thought, the study of technical change has never played a major role. Indeed, from Adam Smith onwards, and with the exceptions of Marx and Marshall, it was progressively written out of classical economic analysis. Thus, despite Smith's emphasis on the division of labour as a form of induced technological and organizational change, little of his remarkably productive insight survived in subsequent writings, apart from the maintained separation of the agricultural and manufacturing sectors as different loci of progress and increasing returns. Not surprisingly, no classical writer foresaw that technical progress in agricultural methods would dispel the niggardliness of nature and banish the spectre of the stationary state. By the time Robbins came to write his methodological characterization of the neoclassical scheme in 1932, not only had technical progress been handed over to the psychologists and engineers, but the very nature of the questions posed by economists had changed fundamentally. Gone was the emphasis on accumulation and progress and in its place stood the analysis of the allocation of given resources under given technical conditions and, moreover, subject to a definition of competition as a state of equilibrium quite incompatible with the increasing-returns implications of the division of labour. The analysis of an organic process became instead the search for the solution of a given jigsaw puzzle. Only Schumpeter (1911) provided a clear way forward. He insisted that technical progress be viewed as a transformation arising from *within* the capitalist system, that it was an integral part of the competitive process and that a key role was played by the entrepreneur and entrepreneurial profits in the process by which technologies acquire economic weight. Orthodox equilibrium theory, it will be noted, had found no room for the entrepreneur. It has been left to the post-1945 generation of economists to reassert the importance of technological change. So far they have done so in a piecemeal, empirical fashion with little attempt to reintegrate the phenomena back into a formal framework of accumulation and structural change. The writings of Pasinetti (1981) and of Nelson and Winter (for example, 1982) can be said, from quite different perspectives, to make this step and have stimulated many others to follow their lead and develop an evolutionary approach to technical change (Dosi, 2000; Nelson and Winter, 2002; Witt, 2003).

Some consequences of technical change

If the process of technical change remains difficult to handle, we can still make limited progress with an analysis of its consequences using long-period methods of analysis. Here, one of the most compelling features of production in modern industrial societies is its roundabout nature. The Industrial Revolution placed modern economies on a path of increasingly roundabout production arrangements in which resources are devoted to elaborate chains of production where raw materials (mineral or agricultural) are worked into intermediate commodities for further processing into final commodities and services with the aid of complex tools and machinery. Specialization and the division of labour are the natural features of such roundabout, mechanized methods, as Adam Smith made clear. When discussing technical progress it is particularly important to recognize that the majority of changes occur within the structure of input–output relations and not only in the activities producing final commodities for consumption (Pasinetti, 1981). A framework which treats this structure as a black box into which primary inputs flow

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and final outputs emerge will not be a useful foundation for the study of technical progress and its effects.

To illustrate some possibilities we employ the following analytic device. Consider a self-contained component of an economic system; we call it a subsystem, which produces a single consumption good, cloth, via three separate, constant returns to scale activities. A lathe is produced with inputs of labour and itself, a loom is produced with inputs of labour and the lathe, and final output, cloth, is produced with labour and the loom. The lathe and the loom are produced means of production; they are outputs of one activity and inputs into another productive activity (Kurz and Salvadori, 1995). Imagine this subsystem to be embedded in a competitive capitalist economy, and that it is analysed in long-period equilibrium conditions in which capital invested in each activity in the subsystem supports a common rate of profits, r, and grows at the common rate, g. There is no structural change taking place in the relative importance of the three activities. Given the profits rate, there will be a unique pattern of relative production prices of the three commodities and a unique level of the real wage, w (ratio of money wage to the price of cloth). Similarly, given the growth rate there is a unique pattern of employment within the subsystem and a unique level of consumption per worker, c (ratio of cloth output to total employment in the subsystem). Now it is well known that higher values of r are related to lower values of w, while higher values of g are related to lower values of c. The corresponding so-called wage–profit and consumption growth frontiers are downward sloping, satisfy the dual property that r=g when w=c, and have a common, finite maximum value for r and g, corresponding to zero w and zero c, respectively. These frontiers are a convenient vehicle with which to explore the effects of technical change.

Starting from a position in which only one production process is available in each activity, consider the long-run equilibrium effects of technical change. Two basic categories of change may be considered, in each case involving changes in one or more input–output coefficients in the subsystem: first, improvements, which imply no qualitative change to any output or input and require only that less of at least one existing input is used within at least one of the processes; and second, inventions, which do imply qualitative change, a physically different output (for example, a new lathe or loom) is produced by an entirely new process.

Whatever the precise changes in input–output coefficients, inventions and improvements can always be classified into three groups by comparing the long-period properties of the new bundle of processes with those of the existing bundle. Dominant technical changes are those which are economically superior over the entire range of profit rates consistent with the existing technology. At the ruling real wage and relative price structure associated with the ‘old’ method, the new process supports a higher rate of profit, and this is the basis for its superiority and – we must here conjecture – its adoption in the subsystem. By similar reasoning, redundant technical changes are those which are economically inferior for all possible wage and price constellations; they constitute failed inventions. Finally, conditional changes are those whose superiority or otherwise do depend on the prevailing relative price structure. For the invention or improvement to become an innovation, it must be economically superior when evaluated at the prevailing price structure. Only dominant changes and the superior set of conditional changes satisfy this condition and can have an economic effect – that is, become innovations.

The long-period effects of innovations depend on the nature of the change in technology and the position of the corresponding process in the input–output structure. In particular, changes in the machine producing processes have quite different consequences from changes within the cloth activity. The more important consequences may be summarized as follows. An improvement or invention in a machine

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
process will alter the entire relative price structure of the subsystem. At the ruling rate of profits, the price of the commodity whose method is improved is reduced relative to the price of all other produced commodities, while the price of all commodities which use the output of the improved process are reduced relative to the money wage. The further down the chain of input–output relations lay the improvements, the greater is the breadth of the consequences of the technical change. Consequently, the simplest case involves an improvement to the cloth activity: cloth falls in price but the relative prices of lathes and looms are unaffected. The corollary of these effects is that any technical change increases the real wage consistent with the ruling rate of profits. Corresponding to the changes in price relations are changes in the structure of employment within the subsystem. A labour-saving technical improvement in a machine activity reduces the proportion of total subsystem employment absorbed by that activity, but how it redistributes employment among the other activities depends on the particular nature and location of the change in question. Any improvement in cloth making, by contrast, has no effects on the equilibrium employment structure. All technical changes will increase the level of consumption per head consistent with the ruling growth rate. Naturally, the magnitude of these effects depends on the ruling values of the growth rate and rate of profits. It is often convenient to summarize the effects of technical change in terms of the associated differences in w–r and c–g frontiers before and after the technical change. In brief, all dominant changes give rise to new frontiers which lie above the ones associated with the old methods. In the case of conditional changes, the old and new frontiers intersect at least once. Nothing clear-cut can be established about the effects of technical change on the aggregate degree of mechanization, whether measured by the value capital–labour or the value capital–output ratio. Depending on the basis of valuing the capital stock, capital intensity may increase or decrease and the different measures may even move in opposite directions. The concept of neutral technical progress has traditionally been a focus of attention in relation to the effects of progress upon the distribution of income. As an example, the traditional case of Harrod-neutral technical progress (no effect on the value capital–output ratio at the ruling r and g) is achieved, trivially, with an improvement in labour productivity confined to the final consumption activity but, more generally, requires that labour productivity increase in equal proportionate amounts in each and every process. Such Harrod-neutral changes leave the structure of employment and relative commodity prices unchanged. There is little doubt that neutral progress of any kind is not to be expected in practice, nor is it a particularly interesting analytic category. Indeed, at given r and g values, Hicks's neutral technical progress (no effect on the capital–labour ratio) is logically impossible in an input–output subsystem of the kind discussed here (Steedman, 1985). More interesting in terms of technological interdependence are the induced technical changes known as trigger effects (Simon, 1951; Fujimoto, 1983). Where technical progress occurs in a machine activity, it may so alter the relative profitability of other activities in which that good is an input that it becomes profitable to adopt different processes within those other activities. In this way the effects of technical change in any machine activity may trigger changes in production methods far beyond the activity in question.

In summary, even under the hypotheses of long-run equilibrium conditions the consequences of technical progress are complex, and are associated with changes in relative prices, real incomes and physical patterns of employment of all inputs. Unless attention is confined to progress in consumption goods, the full ramifications of technical progress can be understood only within an input–output framework. A fortiori one can only understand the inducements to change technology within such a framework of technological interdependence.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The residual debate

A central focus for the literature on technical progress has been provided since the early 1950s by a debate on the measurement of total factor productivity and the implications which follow for our understanding of the growth process. Within the neoclassical tradition, the sources of economic expansion were considered to be population growth and thrift, with growth in labour productivity dependent upon the substitution of capital for labour. Despite the early protests of Schumpeter (1911) that these mechanisms were of negligible significance in explaining long-term growth of capitalist economies, it was not until a series of studies demonstrated the apparent independence of output growth from accumulation that debate could be engaged. The ingenious methods of Abramovitz (1956), Solow (1957) and Kendrick (1973) showed beyond reasonable doubt that the modern growth of the US economy was in proportionate terms at least three-quarters due to increased efficiency in the use of productive inputs and not to the growth in the quantity of resource inputs per se. The implication was quite devastating: an adequate explanation of economic growth appeared to lie outside the traditional concerns of economists, to constitute a residual hypothesis.

From these early studies followed a lengthy sequence of extensions and amendments (that continues unabated today) creating a rich tapestry of data on the growth of the major industrialized and developing nations, and their constituent activities. For our purposes it is the framework employed to identify the contributing sources of economic growth which is of primary interest. For the measured quantities of inputs and outputs are brittle constructs easily swayed by errors of measurement or aggregation, and particularly marked by a failure to allow for quality change in consumption and capital goods, the disamenities of modern growth and the valuation to be placed on enhanced leisure time. Despite the sophisticated efforts to refine measures of the productive input, taking detailed account of the effects of education on labour quality (Denison, 1962) and on the measurement of capital goods and their services (Jorgenson and Griliches, 1967), agreement on the size of the so-called residual element in growth remains as elusive as ever. Here lies a paradox: the importance of new skills and of inventions and quality improvements in capital and intermediate goods increases with the rate of technical progress, as the effects of the information revolution confirm; thus the faster the rate of progress the more the difficulty in measuring its contribution to economic growth. The rise of the service economy and its intangible outputs and inputs certainly adds to these difficulties, and it has become commonplace to suggest that the increased importance of services will reduce the possibilities for further growth in total factor productivity, let alone its accurate measurement (Griliches, 1992). Certainly these considerations increase the merits of attempts to measure productivity growth at the level of more finely defined activities and take advantage of new micro data-sets (Bartlesmann and Doms, 2000). We cannot explore this further here, other than to point to the fact that aggregate productivity growth is now to be treated as a combination of improvements within activities and the structural changes that reallocate output and resource inputs between activities. Productivity change in this frame takes on a more evolutionary hue, as sketched above and premised on the unevenness of progress and adaptation to it (Nelson, 1989).

The central organizing concept behind the early studies was the aggregate production function and the separation of observed growth in output per worker into two independent and additive elements: capital–labour substitution, reflected in movements around a given production function; and increased efficiency

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in resource use, as reflected by shifts in this function. To maintain additivity, the analysis had to be confined to marginal changes in output and input, and could not be applied cumulatively to longer periods without introducing an interaction term between capital substitution and increased efficiency. Within this framework all inputs, the factor services, stand on an equal footing, and constant returns alloyed with universal perfect competition allow marginal productivity pricing to identify the contribution which the growth or relative decline of each input makes to the growth of output per worker. To identify the growth of total factor productivity in a short time interval one need only subtract from the growth of output the growth in total factor input, itself a factor-price-weighted sum of the growth rates of the individual inputs. The sensitivity of such a procedure to errors of measurement in inputs, outputs and relative prices will be obvious.

Some difficulties, immediately apparent from the controversy over capital and distribution, now enter the picture. From the point of view of the long-run supply of productive services, all inputs do not stand on an equal footing. In particular, the flow of capital services depends on the stocks of usable capital instruments and thus on the ability of the economic system to maintain and augment such stocks in quality and quantity. But these capital instruments are produced and reproduced by productive activities which themselves are subject to technical progress over time. Thus, to treat independently increases in efficiency and increases in the stock of capital goods is at least misleading, unless one maintains that technical progress occurs only in consumption activities. The consequences of this for the measurement of total factor productivity are severe (Rymes, 1971). To illustrate, consider an economy growing at a constant rate over time with a constant saving ratio, with the rate of increase in labour productivity the same in all activities, and the capital–output ratio constant. In such an economy the rate of increase in efficiency due to technical progress is exactly measured by the rate of increase in productivity per worker, and not by the measured increase of total factor productivity; which is of a smaller magnitude since it wrongly deducts the effects of induced capital–labour substitution. Increased efficiency makes it easier to reproduce capital goods, such that all the observed rate of increase in the capital–labour ratio (equal to the growth of output per worker) is induced by the enhanced efficiency in the processes producing capital goods. There is no independent capital deepening to contribute to the growth of labour productivity. It is not surprising that when we identify labour as the only primary input then the natural measure of increased efficiency is the rate of increase of labour productivity. Capital goods are after all instruments made by labour too, indeed in some traditions of thought they are described and analysed as so much ‘stored-up labour’. All this, of course, leaves untouched a second aspect of the capital controversy, namely, the severe conditions which have to be imposed to generate an aggregate production function along which output per worker is positively associated with the quantity of capital per worker, and for which input prices may be claimed to measure the corresponding marginal products of factor services (Bliss, 1975; Harcourt, 1972). It is perhaps for this reason that studies of residual productivity have become more prominent at the industry level with as detailed a specification as possible of the relevant physical flows of factor services. But disaggregation does not avoid the problem and the fact that the capital inputs of one activity are derived from the outputs of other activities. The growth of labour productivity in any one activity depends not only upon its own increase in efficiency but upon increased efficiency in the activities supplying it with capital goods, materials and energy. Thus we are back in the world of input–output interdependence in which the results of enhanced efficiency are imported and exported between activities in the way outlined in the previous section. There can be no doubt as to the value of the residual productivity debate; it awakened interest in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
origins and effects of technical progress and stimulated several new lines of research. However, it never did attempt to answer the questions about the constitution and generation of residual productivity growth. These remain the dominant questions as we seek to further understand the complex mechanisms that link technical change to the growth of wealth in modern capitalism.

See Also

- creative destruction
- neo-Ricardian economics
- Sraffian economics
- structural change

Bibliography

Abramovitz, M. 1956. Resource and output trends in the United States since 1870. *American Economic Review, Papers and proceedings* 46, 5–23.

Bartlesmann, E. and Doms, M. 2000. Understanding productivity: lessons from longitudinal data. *Journal of Economic Literature* 38, 569–94.

Bliss, C.J. 1975. *Capital Theory and the Distribution of Income*. Amsterdam: North-Holland.

Denison, E.F. 1962. *The Sources of Economic Growth in the United States*. New York: Committee for Economic Development.

Dosi, G. 2000. *Innovation, Organisation and Economic Dynamics*. Cheltenham: Edward Elgar.

Eisenstein, E. 1979. *The Printing Press as an Agent of Change*. Cambridge: Cambridge University Press.

Fujimoto, T. 1983. Inventions and technical change: a curiosum. *Manchester School of Economics and Social Studies* 51, 16–20.

Griliches, Z., ed. 1992. *Output Measurement in the Service Sectors*. Chicago: Chicago University Press.

Harcourt, G.C. 1972. *Some Cambridge Controversies in the Theory of Capital*. Cambridge: Cambridge University Press.

Harrod, R. 1948. *Towards a Dynamic Economics*. London: Macmillan.

Jorgenson, D. and Griliches, Z. 1967. The explanation of productivity change. *Review of Economic*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Studies 34, 249–83.

Kendrick, J. 1973. *Postwar Productivity Trends in the United States, 1948–1969*. New York: NBER.

Kurz, H. and Salvadori, N. 1995. *Theory of Production: A Long-Period Analysis*. Cambridge: Cambridge University Press.

Kuznets, S. 1977. Two centuries of American economic growth: reflections on US experience. *American Economic Review, Papers and proceedings* 67, 1–14.

Mokyr, J. 1990. *The Lever of Riches*. Oxford: Oxford University Press.

Mokyr, J. 2002. *The Gifts of Athena*. Oxford: Oxford University Press.

Nelson, R.R. 1989. Industry growth accounts and production functions when techniques are idiosyncratic. *Journal of Economic Behavior & Organization* 11, 323–41.

Nelson, R. and Winter, S. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA and London: Belknap Press.

Nelson, R.R. and Winter, S. 2002. Evolutionary theorizing in economics. *Journal of Economic Perspectives* 16(2), 23–46.

Pasinetti, L.L. 1981. *Structural Change and Economic Growth*. Cambridge: Cambridge University Press.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Rymes, T. 1971. *On Concepts of Capital and Technical Change*. Cambridge: Cambridge University Press.

Schumpeter, J. 1911. *The Theory of Economic Development*. Oxford: Oxford University Press, 1934.

Simon, H. 1951. Effects of technical change in a linear model. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans. New York: John Wiley & Sons; London: Chapman & Hall.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

Steedman, I. 1985. On the ‘impossibility’ of Hicks’ neutral technical progress. *Economic Journal* 95, 746–58.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Witt, U. 2003. *The Evolving Economy: Essays on the Evolutionary Approach to Economics*. Cheltenham: Edward Elgar.

How to cite this article

Metcalfe, S. "technical change." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000034> doi:10.1057/9780230226203.1690

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

technology

Joel Mokyr
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Technology is the utilization of natural phenomena and regularities for human purposes. Propositional knowledge – sets of statements about natural regularities and phenomena – provides the epistemic base of technology, whose width largely determines society's ability to generate technology. The high social rate of return to technology leads to underinvestment in new technology, justifying some government subsidization. Only since the 18th century have producers and scientists been systematically linked, allowing technology to flourish. Technology is the main factor driving economic growth; the scope for technology transfer ensures it will continue to be so, as long as the appropriate institutions are in place.

Keywords

ancient Rome; China; competences; economic growth; endogenous technology; Enlightenment; Industrial Revolution; innovation and invention; institutions; intellectual property rights; Islam; isoquant; Japan; Judaism; patents; production function; propositional knowledge; rate of return; research; science; techniques; technology

Article

Technology may be defined as the utilization of natural phenomena and regularities for human purposes. It is a matter of debate whether the manipulation of regularities in human behaviour by firm management should be properly classified as technology, or whether technology should be defined more narrowly as the harnessing of physics, chemistry and biology. Either way, it has become a distinctive feature of *homo sapiens*, and the success of the species in expanding technology has had momentous consequences for our planet, both physically and biologically.

In standard neoclassical economics, technology is regarded as a mapping from inputs to outputs. However, such definitions inevitably assign technology to a ‘black box’ category (Rosenberg, 1994). Yet the economic approach illustrates the many aspects of technology relevant to social science. Much

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of this is captured in the concept of the isoquant, which is a summary description of the mapping. The isoquant displays the three important economic aspects of technology. One is the basic constraint that human knowledge imposes on what people can do. The lowest isoquant – that is, the fewest inputs that combine to produce a unit of output – denotes that limitation to human knowledge at any given moment. There is no implication that a more efficient production is not *possible* in some metaphysical sense, but rather that conditional on the state of knowledge in time *t* this is the best that can be done. Second, the isoquant map indicates that not all producers are necessarily producing at best-practice technology. The entire set above the lowest isoquant is feasible, and while these techniques are by definition less efficient than best practice, there may be many good reasons why average practice is often considerably below best practice. Finally, the fact that the isoquant is a curve and not a point indicates one of the fundamental features of technology, namely, that there are many ways to skin a cat. One of the deepest issues in economics is the choice of technique from the available menu and how that choice is affected by economic parameters. The shape of the isoquant, moreover, tells us a great deal about the nature of the technology available, the degree to which factors are substitutes for one another, the rate at which the marginal products of factors are declining, and so on.

The production function approach only implicitly allows recognition of technology's fundamental nature, namely, that it is, above all, knowledge. By writing the function to include a shift factor, we allow for the growth of knowledge to enable an economy to do things it could not do previously. That technology is first and foremost human knowledge is not always fully recognized. Needless to say, in order to result in production, this knowledge in the vast majority of cases requires some strongly complementary inputs (tools, materials and energy), which economists define as capital and intermediate inputs. Yet in the deepest sense technology exists in the knowledge defining how certain actions plus certain inputs lead to outcomes we deem desirable.

For that reason, the fundamental unit of technology can be regarded as the *technique*, a concept close to Nelson and Winter's ‘routine’ (Nelson and Winter, 1982). A technique is basically a set of instructions on how to produce, much like a simple recipe. Some techniques may, of course, be hugely complicated, with many conditional and nested statements, but their syntax remains prescriptive. Hence the term ‘prescriptive’ knowledge, which contains anything from baking a cake to driving instructions to engineering handbooks. The master-set of all techniques available in society is what Joan Robinson (1956) once called ‘the book of blueprints’ and it constitutes a monstrous menu from which firms and economists make selections. Two types of questions about this menu suggest themselves: how do agents really learn the contents of the menu and make selections, and how did the menu get written in the first place?

The full meaning of technology can be realized by adding the concept of *competence*. Competence concerns the execution of the instructions in the technique by agents. Instructions can be codified (in writing or orally), but no set of instructions is ever complete: they need to be read, interpreted, and carried out. If it were possible to write a *complete* set of instructions that would be wholly self-contained and self-explanatory, competence would be irrelevant and production could be entirely carried out by automatons. It is clear, however that all techniques contain implicit or ‘tacit’ components that require the agency of a human to interpret and carry out the instructions. The size of the ‘tacit’ component varies over time and from field to field, but can never be reduced to zero (Cowan and Foray, 1997). It consists of a certain savoir-faire that comes with experience or imitation, but is hard to learn from codified information.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Technology and knowledge

To understand why and how technology contains what it does at any given time, we need to consider more carefully where it comes from. Many standard definitions of technology refer to ‘science’ as an essential ingredient. Thus the Oxford English Dictionary defines technology as ‘1. the application of scientific knowledge for practical purposes. 2. the branch of knowledge concerned with applied sciences’. Such a definition is patently ahistorical: the close association between ‘science’ (in the modern sense of a consensual, formal, and analytical understanding of natural phenomena) and technology is a product of the past two centuries. For many centuries, people had been employing technology in a variety of fields, yet it is hard to think of a medieval blacksmith, a peasant in biblical Palestine, or a miner in ancient Roman Spain as relying on ‘science’. Technology is therefore part of production in whatever form we observe; science came into the picture only very recently.

As an alternative to the somewhat anachronistic emphasis on science, I have proposed for historical purposes the concept of *propositional knowledge* (Mokyr, 2002). Propositional knowledge is a set of statements about natural regularities and phenomena. These may be expressed in terms of firm regularities such as the laws of thermodynamics or purely in terms of description, measurement, and cataloguing. The distinction between propositional and prescriptive knowledge seems obvious: the planet Neptune and the structure of DNA were not ‘invented’; they were there prior to discovery, whether we knew it or not. The same cannot be said about diesel engines or aspartame. Polanyi (1962, p. 175) notes that the distinction is recognized by patent law, which permits the patenting of inventions (additions to prescriptive knowledge) but not of discoveries (additions to propositional knowledge). He points out that the difference boils down to the observation that prescriptive knowledge can be ‘right or wrong’ whereas ‘action can only be successful or unsuccessful’.

The main point is that this knowledge supports the prescriptive knowledge that is the essence of technology. This support is the *epistemic base* of technology. This base can be narrow or wide, depending on how much of the natural regularities of the technique is known. But it was common for things to be invented despite a narrow or negligible epistemic base. Through luck and serendipity, through dogged trial-and-error, or through an intuitive sense that defies precise analysis, inventors stumbled upon things that worked and worked well, without actually understanding *why and how* they worked. Such concepts are of course relative. It may seem to us, for example, that Alessandro Volta, who built the first working electrical battery in 1800, did not know quite how and why his ‘pile’ worked, but our own understanding of this, while broader than his, may still be quite limited compared with what may be known about the subatomic nature of electricity in the future.

The width of the epistemic base determines to a great extent the effectiveness of the process whereby society creates new technology. Hit-and-miss experimentation or a ‘try-every-bottle-on-the-shelf’ method may well yield new techniques that work, but they will tend to be one-off advances, which soon enough reach the upper bound of their capacity. Further adaptation and tweaking following an invention is far more effective if the basic modus operandi is understood. Moreover, if one does not understand why something works, it will be hard to know what does not work. Enormous amounts of human energy were misallocated, largely by highly talented individuals, in research on alchemy, astrology, attempts to build *perpetuum mobile* machines, and similar impossibilities. Advances in propositional knowledge

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
eventually terminated these programmes. In other words, the lack of propositional knowledge greatly increased the costs of research and development, and until about 1750 most new technical advances soon ran into diminishing returns in terms of their further improvement and development. Lack of an adequate epistemic base also often curtailed the effectiveness of existing technology. For instance, in agriculture, the knowledge that fertilizer increased yields had existed for thousands of years but, until 19th-century organic chemistry widened the epistemic base, basic distinctions between nitrates, phosphorus and potassium were not made, and thus often enough the quantities and kinds of fertilizers used were poor. Better understanding allowed these to be calibrated exactly, which brought about huge improvements in yields.

Technology as an economic good

Much of modern growth theory relies on ‘endogenous technology’ in which technology is being produced by inputs within the economy and responds to prices and costs (Aghion and Howitt, 1997). This literature has successfully dealt with many of the uncomfortable characteristics of technology in models of a growing economy. Of those, a number stand out. One is that like all forms of knowledge it is a purely non-rivalrous good. By giving it away for free, the original owner loses no knowledge of his own, but his capability to exploit this knowledge for commercial ends is reduced. Moreover, by giving it away he loses control over its diffusion since it is normally hard to prevent the new owner from giving it to a third. Second, new knowledge is the ultimate example of increasing returns, a good that is fixed and of no marginal cost to produce. Third, much technology is inappropriable in the sense that, once one person has it, others can often easily imitate or reverse-engineer the technique. Fourth, techniques are hard to exactly quantify, since they often have complex relations with other techniques, ranging from the purely complementary to the pure substitute. Hence, attempts to somehow ‘count’ the number of techniques in a society and to relate them to inputs seems ill-fated and to violate Einstein's dictum that some things that count cannot be counted and a more axiomatic way of measuring technology is needed (Olsson, 2000). Fifth, the process of technology generation is subject to far more uncertainty than any other economic activity. Moreover, this is uninsurable risk. Each invention is made only once, so that there is only a limited amount one can learn from the experience of other inventions. The risk is not only that the technique an inventor is trying to write may not be feasible (or at least not feasible for her), it is also that even if the search is successful someone else may have got there first or the technique may not be commercially exploitable (Rosenberg, 1996). Finally, much of the underlying propositional knowledge, the foundation and essential input of inventive activity, is available at no charge from scientific literature. Yet accessing it may be rather costly all the same.

The production of new technology has changed dramatically over past centuries. Until late in the 19th century, the lone inventor slaving away in a small workshop or lab was the paradigmatic creator of new technology. Some of the more successful ones may have worked on commission (such as the great Richard Roberts, the foremost mechanical engineer in Britain in the first half of the 19th century), but basically they were individuals working on their own account. Some of them were professional inventors who hoped – often in vain – to find the ‘killer ap’ that would make them rich. Others did not bother about the money, and made their inventions for their own satisfaction or for the benefit of mankind, and demonstratively refused to take out patents. Since the late 19th century an increasing share of inventive activity has become part of ‘corporate R&D’, an organized and often bureaucratized form

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of activity, often systematic and always driven by a corporate bottom line. The corporate research lab first emerged in the big German chemical concerns in the late 19th century, but the system was soon adopted in other industrialized nations. The individual inventor working in the proverbial garage has not been quite eliminated, and even today it happens that some lone wolf will come up with an idea that the huge research labs of large corporations did not think of. The agility and creativity of the single human mind may not altogether perish in the dark-suited world of corporate profits, but it is also clear that, when such an invention is successful, the road to riches usually leads to control by or a merger with a larger company with access to credit, marketing networks, and development facilities.

Despite the many market failures in the knowledge industry, however, there is a market for technology since it is clearly valuable (Arora, Fosfuri and Gambardella, 2004). One form this market takes is technical consulting, through which firms purchase expertise not otherwise available to them. This practice can be easily traced back to the early 18th century, and by the time the Industrial Revolution came along consulting engineers of a variety of kinds was common among technologically advanced firms. Another way the market for technology operated was through licensing. Licences were bought and sold commonly in countries in which the patent system worked effectively to protect intellectual property rights, and they are as close as we can get to seeing how technological knowledge was valued (Khan and Sokoloff, 2001). Patent protection of one form or another was quite common in 18th- and 19th-century America and Europe, and firms could buy technology owned by another firm, a practice that has continued into our own time. Once a patent expires, however, the technique becomes common access. For that and other reasons, some industries elected to protect their techniques by secrecy, of which the best-known example is the still secret recipe for Coca Cola, code-named ‘Merchandise 7X’, kept under lock and key in a vault in the Sun Trust Bank Building in Atlanta, Georgia.

From a social point of view, it still seems to be the consensus that most societies seriously underinvest in the creation of new technology. This largely reflects the high social rate of return, which is widely regarded to be higher than the private rate of return resulting from the difficulty of appropriating and exploiting all the benefits of new technology and spillovers from one industry to another as well as across different countries (Mansfield et al., 1977). Such rates of return are notoriously hard to compute, and differ substantially among industries, to say nothing of the difficulty in distinguishing between average and marginal rates. But overall these rates are significantly higher in innovation than in other investment projects. The production of ‘knowledge’ is thus widely regarded as a market requiring some form of government intervention through the subsidization of pure scientific research and the support of some technologies with high rates of spillover.

Technology as a historical force

Simple models that relate complex social systems to a single technological advance or to a number of them have been proposed by some historians, but have not found a large following. These models include Lynn White's suggestion that the feudal system followed from the adoption of the horse stirrup, and Karl Wittfogel's notion that oriental despotisms had their origins in hydraulic technology and the need to coordinate water control (Smith and Marx, 1994). Economists, too, have felt that at times technological developments did affect economic performance and that certain key inventions such as printing with moveable fonts or navigational techniques developed in the 15th century did have major

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
effects on history, although in most cases they are non-committal about what the real exogenous variable is. What is missing from the historical record before 1750 is much evidence that technology was ever powerful enough to bring about sustained economic growth such as to make a significant difference in living standards within a reasonable time. This is not to say that the episodes of high technological creativity in Song China or Renaissance Europe did not lead to major qualitative changes in the control that people had over their resources or their daily lives. But, in so far as they led to economic growth, the effects were limited in time and space, and were often offset by population responses.

Beyond that it should be noted that most societies that ever existed were not technologically creative (Mokyr, 1990). Even societies that can take credit for substantial cultural or artistic achievements, such as Greek, Hellenistic, or Jewish societies, were often not terribly inventive. Indeed, to take a global look at human history, the miracle is that technology actually changed as much as it did. Until quite recently, inventors were widely regarded as dangerous. This was in part because every act of invention is in some sense an act of rebellion and disrespect towards earlier generations and their know-how. For societies that held the wisdom of their forefathers in deep respect, such as Judaism and late Islam, inventors were little different from heretics. Moreover, inventions often threatened to reduce the value of existing human or physical capital by making it obsolete and in some cases redundant. Many governments saw disequilibrium caused by technological shocks as a threat to the status quo, and took a ‘make-no-waves’ approach toward new technology. Entrenched interests often took a Luddite attitude towards innovation, blocking it where they could. While Enlightenment Europe started to challenge every conventional wisdom and embarked on a new path, three literate and sophisticated empires – the Ottomans, Ch'ing China, and Tokugawa Japan – each in its own way closed off most innovation and chose stasis over progress. Second, in most societies, there was a deep social gap between, on the one hand, educated and informed people who studied nature and mathematics and, on the other, those who did the grunt work in the fields, mines, or workshops. Many improvements that were seemingly within reach of Roman society, such as casting iron and eyeglasses, were not achieved. The conventions and social class structure that prevented this kind of communication were slowly bridged in medieval Europe by monks, who were simultaneously the educated class and deeply interested in applying new technology such as windmills and mechanical clocks. But not until 18th-century Europe was there an organized and concerted effort to bring those who knew things and those who made things in direct contact with one another. Only after that happened could producers access and use the propositional knowledge of the natural philosophers (as they were called then), while at the same time the needs of manufacturers and farmers began to affect the research agendas of those in charge of expanding knowledge.

Modern science and technology

After 1820 or so, the connections between science and technology become slowly tighter, but it is unclear which of the two was the dog and which the tail. The relation between the two varied substantially from industry to industry and from technique to technique. In some areas the science came first and then informed the technology (which in turn may then have led to a further sophistication of the science). This was surely true in a field like telegraphy, in which Oersted's famous discovery of electromagnetism in 1819 led scientists to speculate that an electromagnetic telegraph was possible. It was equally true in medicine after 1870 when scientists demonstrated that bacteria were the cause of many infectious diseases, leading to a series of techniques in preventive hygiene. But in other areas the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
technology came first and science followed, often at a distance. Consider materials technology: steel had been made for centuries before some of Lavoisier's students finally realized in 1786 what gave it its special properties. Even after that, the great breakthroughs in steel-making of the 1850s and 1860s were only marginally informed by the metallurgy of the time. In energy technology the gap was even larger: it took the world almost a century and a half after Newcomen's first successful engine in 1712 to finally nail down the principles that made it work.

Any simple statement about the sequencing of science or technology is in any case likely to be false. The two complemented and reinforced one another, theory and practice working cheek by jowl. Technology often operated as a ‘focusing device’ for scientists, showing them a well-defined problem they could then try to solve. In many cases, technology confirmed or inspired theoretical work. Heinrich Hertz's work on oscillating sparks in the 1880s and the subsequent development of wireless communications by Oliver Lodge confirmed Maxwell's purely theoretical work on electromagnetic fields. The success of the Wright brothers at Kitty Hawk in 1903 resolved the dispute among physicists on whether heavier-than-air machines were feasible at all. Following their successful flight, Ludwig Prandtl published his magisterial work on how to compute airplane lift and drag using rigorous methods.

The simple ‘linear model’ in which pure science leads to applied science and from there to technology is further undermined by the important feedback from technology to science that has been called ‘artificial revelation’. In many fields science has been constrained by technology: astronomy depended on telescopes, microbiology on microscopes, chemistry on electrical batteries. In our own time, fast computers have become an indispensable tool for virtually every field of research. In many cases, significant scientific progress occurs when the tools to measure, to observe, or to analyse were significantly improved (Price, 1984). In that way, it could be argued, technology has become self-reinforcing, and the historical models in which technology shocks have no persistence and eventually asymptote off to a new equilibrium have become irrelevant. This is particularly true because modern technology increasingly has the capability to combine and hybridize techniques with other, seemingly unrelated, techniques. Some techniques, indeed, have had such strong and so many complementarities with others that they have been dubbed ‘general purpose technologies’ – steam, electricity, steel, and lasers all come to mind (Bresnahan and Trajtenberg, 1995). Such hybridizing technology can grow at a dazzling pace, even if no pure new knowledge is added, simply through a growing number of combinations and recombinations (Weitzman, 1996). Moreover, modern communications and access technology make it much easier for inventors to scan what is available and find the ‘right’ match to create ever more sophisticated hybrids.

Technology and growth

The exponential growth in technological capabilities is responsible for the emergence of modern growth. However, not all growth derives from improved technology: improved allocations and scale economies account for some proportion of it. But growth based on knowledge is different in some important respects from other forms of growth. One is that it seems much harder to reverse. Whereas wealth based on the gains from trade can be quickly lost due to political turmoil or war, knowledge is much like the proverbial genie that cannot be placed back in the bottle. Although there are historical cases of knowledge actually being ‘lost’, they tend to be rare and in a modern economy quite hard to imagine. More controversial is the question of whether the accumulation of knowledge will ever run into

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
diminishing returns through the exhaustion of technological opportunities or the proliferation of knowledge beyond our capability to contain and control it even with the very best access technology. Concerns that ‘everything that can be invented has been invented’ have been made repeatedly in the past and been held up to ridicule as often by historians of technology. Indeed, technological progress often requires more and better new technology simply because many techniques have unforeseen consequences that require modification or replacement. Internal combustion engines were one of the defining inventions of the 20th century, but their impact on the environment has increasingly emphasized the need for an alternative approach. Similar instances of technological ‘bite-back’ can be observed in a host of other modern techniques and, while not *all* of them necessarily have a technological ‘fix’, better knowledge surely is part of the solution to any problems caused by new techniques. Similarly, technological successes create new needs, which themselves create an endogenous demand for new techniques. Thus the unprecedented increase in life expectancy requires an entire new set of techniques catering to the needs and wishes of people in advanced age brackets, who were a negligible proportion of the population only a century ago.

What is striking about the history of technology is that change has as often as not been competence-reducing rather than competence-increasing. Much effort has gone into making modern technology easy to operate and maintain, with the ingenuity being frontloaded in the design. Once the design is perfected, it can be mass-produced and operated by workers of relatively low skill, and increasingly by automatons, a process Marx termed ‘deskilling’. While this is surely not true across the board, it is increasingly the case not just in manufacturing but in services as well. Such routinization of technology suggests a possible bifurcation in the demands for competence. On the one hand, new techniques will be devised by highly skilled scientists and engineers, whose access to the appropriate propositional knowledge is almost immediate, and whose ingenuity will drive continuous progress. However, their numbers are sufficiently small that their supply is not really a serious constraint. They can be drawn from the elite applicants to the top technological universities in the world, picked by their mathematical skills and creativity. It is a small elite of original, skilled, and driven minds that drives technological progress, as it always has. The fact that their designs can be implemented and maintained by workers whose skills may be quite limited means that the progress of technology is not really constrained much by the supply of human capital. Technology transfer is a concept that captures the very real possibility that knowledge invented in one society can in the end be utilized effectively in another, if the institutional parameters are properly lined up. There seems, hence, little reason to believe that technology-driven growth is a temporary phenomenon.

Technology and institutions

Technology is the key to worldwide economic growth, though it is clear that in many Third World nations the lack of good institutions stands in the way of the adoption of more productive new technology. The payoff structure, both for the generation of new and for the adoption of existing technology, is determined by institutions, and this sets the stage for technological outcomes. The analysis of economists has tried to segment growth between technological achievements that push the product possibility frontier out and institutional achievements that move the economy closer to that frontier. Yet the interactions between the two make such decompositions hazardous. Economists have long realized that inventors, just like everybody else, respond to incentives. So, for that matter, do the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
natural philosophers and mathematicians who provide them with the epistemic base for their new techniques. Yet the exact nature of the proper payoff for those who add to the stock of useful knowledge is the subject of some debate.

Whatever the rewards for successful invention, society needs above all to ensure that those who experiment and research are not penalized, even if their research appears absurd or offensive to most others. Penalizing people because their ideas are eccentric or ‘heretical’ has become rather rare in our age but, with the intensification in the resistance of such ideological organizations as animal-rights or anti-nuclear groups, and the rise of public concern about sensitive areas such as human cloning and stem cells, certain fields of research may be in jeopardy. In the absence of sticks, what the optimal carrots are is far from agreed upon. The most widely used reward is patents, but historically patents have been quite ambiguous as a tool to encourage technological progress (Jaffe and Lerner, 2004). The alternatives to patents all have advantages and drawbacks. Secrecy, of course, is the most costly since the social marginal costs of sharing information are zero. Moreover, any system of intellectual property rights based on secrecy discriminates against those techniques that lend themselves to reverse-engineering or obvious imitation. Awarding the three p's (prizes, pensions, patronage) to successful inventors has coexisted in many places with the fourth (patents). In *ancien régime* France, the Royal Academy was authorized to award such distinctions to inventions that benefited the realm. That such decisions were highly subjective at best and open to nepotism and corruption at worst seems obvious, but no creative society has ever been able to avoid them: Samuel Crompton and Edmund Cartwright, two of the most successful inventors of the British Industrial Revolution, were voted substantial awards by the British Parliament because they had failed to secure patent protection. The South Carolina legislature awarded Eli Whitney $50,000 for his invention in 1794 of the cotton gin, which was easy to imitate. Our own age awarded the Nobel Prize for physics to inventor Jack Kilby in 2000 for the research that led to the integrated circuit.

Yet it should be recognized that contributions to knowledge require more complex incentives than mere property rights. No academic economist should be surprised by a statement that the payoff to successful research is more than just a *financial* compensation or rewards correlated with it. While the 20th century was the age of profit-driven corporate research, it also witnessed unprecedented flourishing of open-source activity, in which participants were incentivized in ways that transcended simple profit-maximizing behaviour. This is not only the case in certain software-writing enterprises such as Linux or Mozilla. It holds for much of the university- or government-driven programme of scientific research, in which academic researchers increased the body of propositional knowledge by a huge multiple while rarely getting rich in the process. The centrality of university research, supported by government grants in the emergence of many of the major technologies of the late 20th century, should serve as evidence of the complexities of the motivations of those who add to the stock of propositional knowledge. In research, the name of the game is credit, not profit. Researchers want property rights to their work, but normally prefer peer recognition to a cheque. The results of their work, through complex interactions with technology, have been the cheapest lunch in human history.

Technology and institutions co-evolve, but they obviously follow very different evolutionary dynamics and selection processes; it cannot be expected that their joint evolution will ever result in an optimal environment for technological progress (Nelson, 1994). At the same time, the corporate and government sectors have emerged as key players in the creation of new technology. The government sets most of the rules of the game (patent laws and enforcement, antitrust, licensing) as well as some priorities (for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
example, military and space research, national institutes for health), while corporations in an oligopolistic market maximize profits subject to the institutional structure, and rely on the epistemic bases created mostly by people at universities or research institutions. The net result is imperfect on many levels (especially the now quite problematic patent system). And yet it has produced and will continue to produce a dynamic, innovative society in which technological progress and economic growth have become the rule rather than the exception (Baumol, 2002). From a long-term historical point of view, that is quite a miracle.

See Also

- growth and institutions
- patents

Bibliography

Aghion, P. and Howitt, P. 1997. *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Arora, A., Fosfuri, A. and Gambardella, A. 2004. *Markets for Technology : The Economics of Innovation and Corporate Strategy*. Cambridge, MA: MIT Press.

Baumol, W. 2002. *The Free-Market Innovation Machine*. Princeton, NJ: Princeton University Press.

Bresnahan, T. and Trajtenberg, M. 1995. General purpose technologies: engines of growth? *Journal of Econometrics* 65, 83–108.

Cowan, R. and Foray, D. 1997. The economics of codification and the diffusion of knowledge. *Industrial and Corporate Change* 6, 595–622.

Jaffe, A. and Lerner, J. 2004. *Innovation and its Discontents*. Princeton, NJ: Princeton University Press.

Khan, B. and Sokoloff, K. 2001. The early development of intellectual property institutions in the United States. *Journal of Economic Perspectives* 15(2), 1–15.

Mansfield, E., Rapoport, J., Romeo, A., Wagner, S. and Beardsley, G. 1977. Social and private rates of return from industrial innovations. *Quarterly Journal of Economics* 91, 221–40.

Mokyr, J. 1990. *The Lever of Riches: Technological Creativity and Economic Progress*. New York: Oxford University Press.

Mokyr, J. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, MJ: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Nelson, R. 1994. Economic growth through the co-evolution of technology and institutions. In *Evolutionary Economics and Chaos Theory: New Directions in Technology Studies*, ed. L. Leydesdorff and P. Van Den Besselaar. New York: St Martins Press.

Nelson, R. and Winter, S. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap Press/Harvard University Press.

Olsson, O. 2000. Knowledge as a set in idea space: an epistemological view on growth. *Journal of Economic Growth* 5, 253–76.

Polanyi, M. 1962. *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago: Chicago University Press.

Price, D. 1984. Notes towards a philosophy of the science/technology interaction. In *The Nature of Knowledge: Are Models of Scientific Change Relevant?*, ed. R. Laudan. Dordrecht: Kluwer.

Robinson, J. 1956. *The Accumulation of Capital*. Homewood, IL: Irwin.

Rosenberg, N. 1994. *Exploring the Black Box*. New York: Cambridge University Press.

Rosenberg, N. 1996. Uncertainty and technological change. In *Technology and Growth*, ed. J. Fuhrer and J. Sneddon Little. Conference Series No. 40. Boston: Federal Reserve Bank of Boston.

Smith, M. and Marx, L., eds. 1994. *Does Technology Drive History?* Cambridge, MA: MIT Press.

Weitzman, M. 1996. Hybridizing growth theory. *American Economic Review* 86, 207–13.

How to cite this article

Mokyr, Joel. "technology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000211> doi:10.1057/9780230226203.1691

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

temporary equilibrium

J.-M. Grandmont
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Temporary general equilibrium views the dynamic evolution of an economy as taking place sequentially in calendar time, with decisions being made and equilibrium being achieved at each date in the light of the traders' expectations about the future. The article surveys the contributions of the field to the microeconomic foundations of macroeconomics, in particular to the analysis of monetary phenomena, non-clearing markets, imperfect competition and the foundations of Keynesian unemployment, as well as the study of economic dynamics and (in)stability of self-fulfilling expectations under various learning schemes, in relation in particular with 'excess volatility' of financial markets.

Keywords

arbitrage; capital market imperfections; classical dichotomy; equilibrium theory; error learning models; E-stability; excess volatility; expectations; fiat money; futures markets; game theory; imperfect competition; intertemporal equilibrium; intertemporal substitution effects; Keynesian unemployment; liquidity trap; local stability; microfoundations; monopolistic competition; new Keynesian macroeconomics; non-clearing markets; oligopolistic competition; quantity theory of money; real balance effect; self-fulfilling expectations; speculation; staggered price and wage setting; sticky information; Say's Law; temporary equilibrium; uncertainty principle; Walras's Law; wealth effect

Article

The conceptual framework

The fact that trade and markets take place sequentially over time in actual economies is a trivial observation. It has nevertheless far-reaching implications. At any moment, economic units have to make decisions that call for immediate action, in the face of a future that is as yet unknown. Expectations about the unknown future play therefore an essential role in the determination of current economic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
variables. On the other hand, the expectations that traders hold at any time are determined by the information that they have at that date on the economy, in particular on its current and past states. Observed economic processes are thus the result of a strong and complex interaction between expectations of the traders involved and the actual realizations of economic variables.

Economists have long recognized that such an interaction should be at the heart of any satisfactory theory of economic dynamics. The temporary equilibrium approach was indeed designed quite a while ago by the Swedish school (Lindahl, 1939) and J.R. Hicks (1939; 1965), with the intent to establish a general conceptual framework that would enable economists to cope with the study of dynamical economic systems, and in particular to incorporate in their models the subtle interplay between expectations and actual realizations of economic variables that seems factually so important. Economic theorists have employed this framework in a systematic way since then, using in particular the powerful techniques of modern equilibrium and/or game theory; this effort has yielded important improvements of our understanding of the microeconomic foundations of macroeconomics.

Before reviewing briefly a few of these important advances, it may be worthwhile to make clear what the basic characteristics of the temporary equilibrium approach are, and to compare it with others. To fix ideas, let us assume that time is divided into an infinite, discrete sequence of dates. We may envision first a specific institutional set-up, that was called a *futures economy* by Hicks (1939), and later generalized by Arrow and Debreu. Let us assume that markets for exchanging commodities are opened at a single date, say date 0; assume further that at that date, markets exist for contracts to deliver commodities at each and every future date $t \geq 0$. The specification of a ‘commodity’ will then involve not only the physical characteristics of the good or service to be delivered, but also the location and the circumstances (‘state of nature’) of the delivery. One gets then what has been called a ‘complete’ set of futures markets at the initial date $t = 0$ (Debreu, 1959, ch. 7).

It is clear that this framework is essentially timeless. Once an equilibrium is reached at date 0 (this equilibrium may be Walrasian or the result of any other game theoretic equilibrium notion), production and trade do take place sequentially in calendar time. But the coordination of the decisions of all traders is achieved at a single date through futures markets. There is no sequence of markets over time, and no role for expectations, money, financial assets, or stock markets.

Let us consider next another, more dynamic, type of organization, in which markets do open in every period. In this framework, traders would exchange at every date commodities immediately available on spot markets, promises to deliver specific commodities at later dates on futures markets, as well as money, financial assets and/or stocks (of course markets must be ‘incomplete’ in the sense of Arrow–Debreu at every date, otherwise reopening markets would serve no purpose). To convey the following discussion most simply, let us assume away all sources of uncertainty and consider the case where the state of the economy at any date can be described by a single real number. To simplify matters further, let us assume that the state of the economy at *t*, say x_t, is completely determined by the forecasts $x^e_{i,t+1}$ made by all traders $i = 1, \ldots, m$ at date *t* about the future state, through the relation

$$x_t = f(x^e_{1,t+1}, \ldots, x^e_{i,t+1}, \ldots, x^e_{m,t+1}) \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The temporary equilibrium map *f* describes the result of the market equilibrating process at date *t* – be it Walrasian or not – for a given set of forecasts. Of course, in the study of any particular economy, the map *f* will be derived from the 'fundamental' characteristics of the economy: tastes, endowments, technologies, the rules of the game, the policies followed by the government.

The foregoing formulation does seem to take into account the observed fact that markets unfold sequentially in calendar time. It is, however, incomplete since no specification of the way in which forecasts are made at each date has been offered at this stage.

We must first discuss a concept that was introduced by Hicks himself, that of an *intertemporal equilibrium*, with self-fulfilling expectations, and that has been extensively used recently in a variety of contexts. Such an intertemporal equilibrium is defined formally, in the present framework, as an infinite sequence of states $\{x_t\}$ and of forecasts $\{x^e_{i,t+1}\}$ satisfying (1) and

$$x^e_{i,t+1} = x_{t+1} \tag{2}$$

for all dates. Although time appears explicitly in this formulation, it should be clear that this particular equilibrium concept is also intrinsically *timeless*. Indeed all elements of the sequences of equilibrium states $\{x_t\}$ and of equilibrium forecasts $\{x^e_{i,t+1}\}$ are determined simultaneously by an outside observer: present and future markets are equilibrated all at the same time.

The preceding discussion shows how we must proceed to describe a sequential adjustment of markets, in calendar time. We *must* add to the temporary equilibrium relationship (1) a specification of the way in which traders forecast the future at each date *as a function of their information on current and past states of the economy*. If we assume, for the simplicity of the exposition, that the information available to traders at date *t* is represented by the sequence $(x_t, x_{t-1}, \ldots)$, that means that we have to add to (1), *m expectations* functions of the form

$$x^e_{i,t+1} = \psi_i(x_t, x_{t-1}, \ldots) \tag{3}$$

The equations (1) and (3) describe then in a consistent way a sequential adjustment of markets – a sequence of *temporary equilibria* – in which time goes forward, as it should. Given past history $(x_{t-1}, x_{t-2}, \ldots)$, (1) and (3) determine the current temporary equilibrium state and forecasts. Once such a temporary equilibrium is reached, production and exchange takes place at date *t*, and the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economy can move forward to date $t+1$, where the equilibrating process is repeated.
The temporary equilibrium approach, as sketched in the formulation (1) plus (3) is the general formulation, in fact the *only* sort of formulation that is allowed, if one wishes to describe the evolution of the economy as a *sequence* of markets that adjust one after each other. One should expect accordingly the approach to include self-fulfilling expectations as a special case. Indeed choose a particular intertemporal equilibrium. Then the associated sequence of states, say $\{\bar{x}_t\}$, is a solution of the difference equation

$$\bar{x}_t = F(\bar{x}_{t+1}) \tag{4}$$

in which $F(x) = f(x, \ldots, x)$ for all *x*. Consider now the economy at date *t*, and assume that past states have been $(\bar{x}_{t-1}, \bar{x}_{t-2}, \ldots)$. Assume that the traders know the characteristics of the economy, or at least the map *F*, and further that the map *F* is invertible (we are voluntarily vague about the domain of definitions of the functions under consideration, to simplify the present methodological discussion, but these technical details can be fixed up). The traders are then able to infer that the recurrence satisfied by current and past states, that is, $x_t = F^{-1}(x_{t-1})$, will obtain in the future as well, their forecasting rule may be viewed as the result of iterating twice that relation, or of inverting $\bar{x}_{t+1} = F^2(\bar{x}_{t-1})$, for all $i = 1, \ldots, m$

$$\psi_i(\bar{x}_t, \bar{x}_{t-1}, \ldots) = F^{-2}(\bar{x}_{t-1}). \tag{5}$$

If this relation holds, $\bar{x}_t$ is indeed a temporary equilibrium state (that is, it solves (1) and (3)) at date *t*, given past history $(\bar{x}_{t-1}, \bar{x}_{t-2}, \ldots)$.
As we have just shown, the temporary equilibrium method includes self-fulfilling expectations as a special case. This shows incidentally that the opposition often made in the literature, between self-fulfilling expectations, that are claimed to be ‘forward looking’, and ‘backward looking’ expectations as in (3), is presumably misleading. The temporary equilibrium approach is indeed much more general, since it permits to incorporate in the analysis the fact that traders usually learn the dynamics laws of their environment only gradually, and thus to study in principle how convergence toward self-fulfilling expectations may or may not obtain in the long run.
The preceding discussion was carried out in a simple one-dimensional world operating under certainty. It should be clear nevertheless that the qualitative conclusions we obtained hold as well in a more complex, multidimensional world operating under uncertainty.
When the evolution of the economy is described as a sequence of temporary equilibria, at each date, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
current equilibrium states are determined by past history. In this framework, a number of issues arise naturally. First, one has to find the conditions under which the dynamic evolution of the economy is well defined. In other words, when does a temporary equilibrium exist? Second, does the corresponding dynamical system have long-run equilibrium states, such as deterministic stationary states or cycles, and/or stationary stochastic processes, along which expectations are self-fulfilling? Under which conditions, in particular on the formation of expectations, do the sequences of temporary equilibria so generated converge to such a long-run equilibrium? This is precisely the sort of questions that have attracted the attention of modern economic theorists working in temporary equilibrium theory.

Overview

We turn now to a brief appraisal of this research effort, referring the interested reader to more extensive and more technical surveys that already exist in the literature, see for example Grandmont (1977; 1987; 1998).

Money and assets in competitive markets

Considering a sequence of markets opens immediately the possibility for traders to hold money and more generally, assets of various kinds for saving, borrowing, transactions purposes and/or insurance motives. The application of the modern techniques of temporary equilibrium theory to the study of monetary phenomena has led to a major reappraisal, in the 1970s, of classical and neoclassical monetary theories in competitive environments. It has permitted in particular to solve an old problem that had puzzled economic theorists for some time (Hahn, 1965), namely, why fiat money, which has no intrinsic value, should have a positive value in exchange in competitive markets. The answer provided by traditional neoclassical theory relied essentially upon unit-elastic price expectations and the presence of real balance or wealth effects (Patinkin, 1965). Modern temporary equilibrium methods have shown that sort of answer to be surely incomplete and presumably mistaken: intertemporal substitution effects have to play an important role, and this can be achieved only by abandoning the hypothesis of unit-elastic expectations and by introducing some degree of inelasticity of expectations with respect to current observations (an example of such a condition was used in (5) above, where the forecast was made to depend on past states but not on the current state). The reappraisal of monetary theory by means of the temporary equilibrium method clarified greatly many confusing debates of the preceding literature: the relations between Walras's and Say's Law, the meaning and the validity of the classical dichotomy and the quantity theory of money, the possibility of monetary authorities to manipulate the interest rates or the money supply, the existence of a ‘liquidity trap’ (Grandmont, 1983). The introduction of cash-in-advance constraints in temporary competitive equilibrium models of money (Grandmont and Younès, 1972; 1973) yielded important insights into the relations between its respective roles as a store of value and as a medium of exchange, and time preference, and permitted to make precise the microeconomic foundations of Milton Friedman's theory of optimum cash balances (1969; see also Woodford, 1990). Such models of money using cash-in-advance constraints have been popular in modern macroeconomics, following the contribution of R.E. Lucas, Jr. (1980). Capital market imperfections that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
make explicit the essential role of money in exchange have since been central to the modern analysis of monetary theory and policy (see Wallace, 2001).

The introduction of assets of various kinds in competitive markets leads also to the possibility of speculation and arbitrage in capital markets. Different persons with different tastes or expectations will then be willing to trade such assets. An important question is to study the conditions ensuring the existence of a temporary equilibrium in that context. A neat answer to that problem was provided by J.R. Green (1973) and O.D. Hart (1974): there must be some agreement between the traders' expectations about future prices.

Temporary equilibria with non-clearing markets and imperfect competition

A temporary equilibrium need not be Walrasian. One may consider cases where prices and/or wages are set through monopolistic or oligopolistic competition at the beginning of each elementary period and remain temporarily fixed within that period. A temporary equilibrium corresponding to these prices is then achieved at each date by quantity rations that set upper or lower bounds on the traders' transactions. It had been known for some time that traditional Keynesian macroeconomic models of unemployment involved, explicitly or implicitly, the assumption of temporarily fixed prices and/or wages, as noted by Hicks himself (1965). The choice-theoretic structure of these models was rather unclear, however, which was a source of some confusion. The systematic study of temporary equilibrium models with quantity rationing undertaken in the 1970s produced deep insights on this issue, and unveiled the hidden but central role played by quantity signals, as perceived by the traders in addition to the price system, to achieve an equilibrium in such models.

One major outcome of this research programme was the discovery that different types of unemployment could obtain, and even co-exist. 'Keynesian unemployment' corresponds to a situation where there is an excess supply on the labour and the goods markets. In such a situation, firms perceive constraints on their sales because demand is too low. Keynesian policies aiming at increasing aggregate demand may work in such a case. But unemployment may co-exist with an excess demand on the goods markets. In such a regime, called 'classical unemployment' by Malinvaud (1977), the source of unemployment is rather the low profitability of productive activities. Keynesian policies may not work in that case; one has to resort to policies that restore profits, such as lowering real wages. In that respect, these results achieved a remarkable synthesis, within a unified and clear conceptual framework, between two paradigms that appeared fundamentally distinct beforehand.

The research on this topic proceeded very early on to endogenize prices and wages and yielded numerous insights of the connections between Keynesian models of unemployment and price or wage making in monopolistic or oligopolistic models of competition (see Barro and Grossman, 1976; Benassy, 1982; 1986; Grandmont and Laroque, 1976; Hart, 1982; Malinvaud, 1977; Negishi, 1979). It has since become a cornerstone building block of the modern reformulation of so-called 'new Keynesian macroeconomics' (Benassy, 2002; Dreze, 1991; Mankiw and Romer, 1991). While early formulations focused on temporary equilibria with non-clearing markets due to exogenously staggered price and wage setting (see Taylor, 1999, for a survey), more recent research seeks to explain such staggering of price and wage changes as the rational reaction of agents under the gradual diffusion of 'sticky information' (Mankiw and Reiss, 2003).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Learning and (in)stability

The temporary equilibrium approach includes self-fulfilling expectations as a particular case, and is in fact more general, since it can incorporate learning in the formation of the traders' expectations. An important issue, that has been early on the agenda of that research programme (Fuchs and Laroque, 1976), is then to know whether the sequences of temporary equilibria that are associated to given learning processes or expectations functions converge eventually to a long-run equilibrium along which forecasting mistakes vanish. The question arises of course for long-run equilibria that are simple, such as steady states, or more complex, such as deterministic cycles (Grandmont, 1985). The general lesson that seems to come out the research works done on the topic appears to be some kind of '*uncertainty principle*' (Grandmont, 1998). Learning generates local instability of self-fulfilling expectations whenever agents are on average uncertain about the local stability of the system, and thus ready to extrapolate a wide range of regularities (trends) out of past deviations from equilibrium, and when the influence of expectations on the dynamics is significant. On the other hand, learning may generate locally stable dynamics when either expectations do not matter much or traders extrapolate a restricted range of stable trends out of past deviations from equilibrium.
The above principle arises in a wide variety of learning processes, in particular in 'error learning' models, least squares and Bayesian learning. Of course, if one is willing to restrict the range of learning schemes, one may be able to produce sharper stability criteria (see in particular the concept of 'E-stability' developed by Evans and Honkapohja, 1999, for a particular class of learning processes). Local learning instability due to the above 'uncertainty principle' may explain why markets in which expectations are thought to play a significant role, such as markets for financial assets, durable goods, capital or inventories, display more volatility than others. In a similar vein, Brock and Hommes (1997) have shown local instability in a cobweb model, and convergence to more or less complex cyclical or chaotic long-run equilibria, when agents choose in variable proportions among different more or less efficient (and costly) learning schemes, and have sought to apply this approach to explain 'excess volatility' in financial markets.

See Also

- fiat money
- fixprice models
- general equilibrium
- non-clearing markets in general equilibrium
- new Keynesian macroeconomics
- sticky wages and staggered wage setting

Bibliography

Barro, R.J. and Grossman, H.I. 1976. *Money, Employment and Inflation*. Cambridge: Cambridge

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
University Press.

Benassy, J.P. 1982. *The Economics of Market Disequilibrium*. New York: Academic Press.

Benassy, J.P. 1986. *Macroeconomics: An introduction to the non-Walrasian approach*. New York: Academic Press.

Benassy, J.P. 2002. *The Macroeconomics of Imperfect Competition and Nonclearing Markets : a dynamic general equilibrium approach*. Cambridge, MA: MIT Press.

Brock, W.A. and Hommes, C. 1997. A rational route to randomness. *Econometrica* 65, 1059–95.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. Cowles Foundation Monograph No. 17. New York: Wiley.

Dreze, J. 1991. *Underemployment Equilibria*. Cambridge: Cambridge University Press.

Evans, G.W. and Honkapohja, S. 1999. Learning dynamics. In *Handbook of Macroeconomics*, vol. 1, ed. J.B. Taylor and M. Woodford. New York: North-Holland.

Friedman, M. 1969. *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

Fuchs, G. and Laroque, G. 1976. Dynamics of temporary equilibria and expectations. *Econometrica* 44, 1157–78.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45, 535–72.

Grandmont, J.M. 1983. *Money and Value: A Reconsideration of Classical and Neoclassical Monetary Theories*. Econometric Society Monograph No. 5. Cambridge: Cambridge University Press.

Grandmont, J.M. 1985. On endogenous competitive business cycles. *Econometrica* 53, 995–1045.

Grandmont, J.M., ed. 1987. *Temporary Equilibrium: Selected Readings*. New York: Academic Press.

Grandmont, J.M. 1998. Expectations formation and stability of large socioeconomic systems. *Econometrica* 66, 741–81.

Grandmont, J.M. and Laroque, G. 1976. On temporary Keynesian equilibria. *Review of Economic Studies* 43, 53–67.

Grandmont, J.M. and Younès, Y. 1972. On the role of money and the existence of a monetary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
equilibrium. *Review of Economic Studies* 39, 355–72.

Grandmont, J.M. and Younès, Y. 1973. On the efficiency of a monetary equilibrium. *Review of Economic Studies* 40, 149–65.

Green, J.R. 1973. Temporary general equilibrium in a sequential trading model with spot and future transactions. *Econometrica* 41, 1103–23.

Hahn, F.H. 1965. On some problems of proving the existence of an equilibrium in a monetary economy. In *The Theory of Interest Rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Hart, O.D. 1974. On the existence of equilibrium in a securities model. *Journal of Economic Theory* 9, 293–311.

Hart, O.D. 1982. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97, 109–38.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press; 2nd edn. 1946.

Hicks, J.R. 1965. *Capital and Growth*. Oxford: Clarendon Press.

Lindahl, E. 1939. *Theory of Money and Capital*. London: Allen and Unwin.

Lucas, R.E., Jr. 1980. Equilibrium in a pure currency economy. *Economic Inquiry* 18, 203–20. Also in *Models of Monetary Economies*, ed. J.H. Kareken and N. Wallace, Minneapolis: Federal Reserve Bank of Minneapolis, 1980.

Malinvaud, E. 1977. *The Theory of Unemployment Reconsidered*. Oxford: Basil Blackwell.

Mankiw, N.G. and Romer, D., eds. 1991. *New Keynesian Economics*, 2 vols. Cambridge, MA: MIT Press.

Mankiw, N.G. and Reiss, R. 2003. Sticky information: a model of monetary nonneutrality and structural slumps. In *Knowledge, Information and Expectations in Modern Macroeconomics*, ed. P. Aghion et al. Princeton: Princeton University Press.

Negishi, T. 1979. *Microeconomic Foundations of Keynesian Macroeconomics*. Amsterdam: North-Holland.

Patinkin, D. 1965. *Money, Interest and Prices*. 2nd edn. New York: Harper & Row.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Taylor, J.B. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of Macroeconomics*, vol. 1, ed. J.B. Taylor and M. Woodford. New York: North-Holland.

Wallace, N. 2001. Whither monetary economics? *International Economic Review* 42, 847–69.

Woodford, M. 1990. The optimum quantity of money. In *Handbook of Monetary Economics*, vol. 2, ed. B.M. Friedman and F.H. Hahn. New York: North-Holland.

How to cite this article

Grandmont, J.- M. "temporary equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000040>
doi:10.1057/9780230226203.1692

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

term structure of interest rates

Burton G. Malkiel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term structure of interest rates concerns the relationship among the yields of bonds that differ only with respect to their terms of maturity. This article explains the three traditional explanations of the term structure. 1. The expectations theory considers the long rate to be an average of current and future short rates. 2. The liquidity-preference theory posits that illiquid, risky long-terms bonds must yield a premium over expected short rates. 3. The hedging-pressure theory stresses the influence of the preferred habitats of different investors. A survey of empirical work on the term structure including affine yield models concludes.

Keywords

affine models of the term structure; bonds; central banks; expectations theory; expectations-forming mechanisms; Fisher, I.; Greenspan, A.; hedging-pressure theory; Hicks, J.; inflation; interest rates; liquidity premium; liquidity-preference theory; long-term interest rates; Lutz, F.; no-arbitrage condition; preferred habitat theory; risk premium; short-term interest rates; term structure of interest rates; yield curve

Article

The term structure of interest rates plays a critical role in the decisions of individuals and corporations and in the conduct of monetary policy. Individuals deciding between an adjustable and a fixed-rate mortgage, and corporations deciding whether to finance their operations with short- or long-term debt, can make sensible decisions only if they understand the factors that determine the relationship between short- and long-term rates. Central banks, which are considered to have substantial control over short-term interest rates, need to understand the likely effect on long rates from their activities in the short-term market.

During 2005, in his final year as Chairman of the Federal Reserve, Alan Greenspan referred to the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

behaviour of long-term rates as a ‘conundrum’ (Greenspan, 2005, p. 8). The US central bank had raised short-term rates at eleven consecutive meetings but the long-term government-bond rate actually declined over the period. This article examines the Greenspan conundrum and presents the traditional theories offered by economists to explain the relationship between short- and long-term interest rates.

Consider a one-year zero coupon bond that makes a single known payment at maturity, which is *F*, the face value of the bond. The bond's price is *P* and the investor's return is *R*,1, which is the bond's simple (one-year) yield to maturity:

$$P = \frac{F}{(1 + R, 1)}.$$

Assuming annual compounding, the yield to maturity of an *N*-year bond, we can find *R*,*N* by solving the equation

$$P = \frac{F}{(1 + R, N)^N}.$$

The term structure of interest rates concerns the relationship among the yields of (zero coupon) default-free securities that differ only with respect to their term to maturity. The relationship is more popularly known as the shape of the yield curve, which is pictured by plotting the various yields to maturity (*R*s) on the vertical axis against the different years to maturity (*N*s) on the horizontal axis. Explaining the shape of curve has been a subject of intense examination by economists for over 60 years. Historically, three competing theories have attracted the widest attention. These are known as the expectations, liquidity preference, and hedging-pressure (or preferred habitat) theories of the term structure.

The expectations theory

According to the expectations theory, the shape of the yield curve can be explained by investors’ expectations about future short-term interest rates. I use the term ‘interest rate’ to refer to the yield to maturity of a zero coupon bond of a specific term to maturity (*N*). This proposition dates back at least to Irving Fisher (1896), but the main development of the theory was done by Hicks (1939) and Lutz (1940). More recent versions of the theory have been developed by Malkiel (1966), Roll (1970; 1971), and Cox, Ingersoll and Ross (1981).

Suppose, for example, that investors believe that the prevailing level of interest rates is unsustainably high and that lower rates are more probable than higher ones in the future. Under such circumstances, long-term bonds will appear to investors as more attractive than shorter-term issues *if* both sell at equal yields. Long-term bonds will permit an investor to earn what is believed to be an unusually high rate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
over a longer period of time than short-term issues, whereas investors in shorter bonds subject themselves to the prospect of having to reinvest their funds later at the lower yields that are expected. Moreover, longer-term bonds are likely to appreciate in value if expectations of falling rates prove correct. Thus, if short and long securities sold at equal yields, investors and arbitrageurs would tend to bid up the prices (force down the yields) of long-term bonds while selling off short-term securities, causing their prices to fall (yields to rise). Thus, a descending yield curve with short issues yielding more than longer ones can be explained by expectations of lower future rates. Similarly, an ascending yield curve, with longer issues yielding more than shorter-term ones, can be explained by expectations of rising rates.

Under the assumptions of the perfect-certainty variant of the expectations theory, there are no transactions costs, and all investors make identical and accurate forecasts of future interest rates. The theory then implies a formal relationship between long- and short-term rates of interest. Specifically, the analysis leads to the conclusion that the long rate is an average of current and expected short rates. Consider the following simple two-period example, where only two securities exist (a one-year and a two-year zero coupon bond), and investors have funds at their disposal for one or two years. Let capital *R*s stand for actual market rates (yields), while lower-case *r*s stand for expected or forward rates. Prescripts represent the time periods for which the rates are applicable, while postscripts stand for the maturity of the bonds. Thus, $_t R_2$ indicates today's actual two-year rate, while $_{t+1} r_1$ stands for the expected one-year rate in period $t+1$.

If investors are profit maximizers, it follows that each investor will choose that security (or combination of securities) that maximizes his return for the period during which his funds are available. Consider the alternatives open to the investor who has funds available for two years. The two-year investor will have no incentive to move from one bond to another when he can make the same investment return from buying a combination of short issues or holding one long issue to maturity. If such an investor invests one dollar in a one-year security and then reinvests the proceeds at maturity (that is, $1 + {}_t R_1$) in a one-year issue next year, his total capital will grow to $(1 + {}_t R_1)(1 + {}_{t+1} r_1)$ at the end of the two-year period. Alternatively, if he invests his dollar in a two-year zero coupon issue, he will have at maturity $(1 + {}_t R_2)^2$. In equilibrium, where the investor has no incentive to switch from security to security, the two alternatives must offer the same overall yield, namely,

$$(1 + {}_t R_2)^2 = (1 + {}_t R_1)(1 + {}_{t+1} r_1). \tag{1}$$

Thus, the two-year rate can be expressed as a geometric average involving today's one-year rate and the one-year rate of interest anticipated next year:

$$(1 + {}_t R_2) = [(1 + {}_t R_1)(1 + {}_{t+1} r_1)]^{1/2}. \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

If eq. (2) holds, then the holding-period return for the one-year investor will also be the same whether he buys a one-year bond and holds it to maturity or buys a two-year bond and sells it after one year. If eq. (2) does not hold, say because the two-year rate was lower than the average of the current and prospective one-year rate, an arbitrageur could make a sure profit by selling the two-year issue short and purchasing a series of one-year securities. It is in this sense that the expectations theory fulfills a no-arbitrage condition.

In similar fashion, the rate on longer-term issues must turn out to be an average of the current and a whole series of future short-term rates of interest. Only when this is true can the pattern of short and long rates in the market be sustained. The long-term investor must expect to earn through successive investment in short-term securities the same return over his investment period that he would earn by holding a long-term bond to maturity. In general, the equilibrium relationship is,

$$(1 + t, R, N) = [(1 + t, R, 1)(1 + t + 1, r, 1) \ldots\ldots \times (1 + t + N - 1, r, 1)]^{1/N} \quad (3)$$

The expectations theory can be extended to a world of uncertainty and it can account for every sort of yield curve. If short-term rates are expected to be lower in the future, then the long rate, which we have seen must be an average of those rates and the current short rate, will lie below the short rate. Similarly, long rates will exceed the current short rate if rates are expected to be higher in the future.

Notice that the expectations theory is capable of explaining the Greenspan conundrum. Suppose that the central bank's aggressive raising of short-term rates was expected to lower future inflation and weaken future economic activity, thus leading to expectations of lower future short-term rates. In such a case, the long rate could fall because, while the current short rate was higher, the entire set of future short rates would be lower than was previously expected.

The liquidity-preference theory

The liquidity-preference theory, advanced by Hicks (1939), accepts that expectations are important in influencing the shape of the yield curve. Nevertheless, it argues that, in a world of uncertainty, short-term issues are more desirable to investors than longer-term issues because they are more liquid. Short-term issues can be converted into cash at short notice without appreciable loss in principal value, even if rates change unexpectedly. Long-term issues, however, will tend to fluctuate widely in price with unanticipated changes in interest rates and hence ought to yield more than shorts by the amount of a risk premium.

If no premium were offered for holding long-term bonds, it is argued that most individuals and institutions would prefer to hold short-term issues to minimize the variability of the money value of their portfolios. On the borrowing side, however, there is assumed to be an opposite propensity. Borrowers

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

can be expected to prefer to borrow at long term to assure themselves of a steady source of funds. This leaves an imbalance in the pattern of supply and demand for the different maturities – one that speculators might be expected to offset. Hence, the final step in the argument is the assertion that speculators are also averse to risk and must be paid a liquidity premium to induce them to hold long-term securities. The arbitrage described in the exposition of the expectations theory is not riskless. Thus, even if interest rates are expected to remain unchanged, the yield curve should be upward sloping, since the yields of long-term bonds will be augmented by risk premiums necessary to induce investors to hold them. While it is conceivable that short rates could exceed long rates, if investors think that rates will fall sharply in the future, the 'normal relationship' is assumed to be an ascending yield curve.

Formally, the liquidity premium is typically expressed as an amount that is to be added to the expected future rate in arriving at the equilibrium-yield relationships described in eqs. (1) through (3). If we let $L, 2$ stand for the liquidity premium that should be added to next year's forecasted one-year rate, we have

$$(1 + t, R, 2)^2 = (1 + t, R, 1)(1 + t + 1, r, 1 + L, 2) \tag{4}$$

and

$$(1 + t, R, 2) = [(1 + t, R, 1)(1 + t + 1, r, 1 + L, 2)]^{1/2}. \tag{5}$$

Thus, if $L, 2$ is positive (that is, if there is a liquidity premium), the two-year rate will be greater than the one-year rate even when no change in rates is expected. It has also been customary to assume that $L, 3$, the premium to be added to the one-year rate forecast for two years hence (that is, period $t + 2$), is even greater than $L, 2$, so that the three-year rate will exceed the two-year rate when no change is expected in short-term rates over the next three years. In general, the liquidity-premium model may be written as

$$(1 + t, R, N) = (1 + t, R, 1)(1 + t + 1, r, 1 + L, 2) \ldots\ldots (1 + t + N - 1, r, 1 + L, N)]^{1/N}. \tag{6}$$

On the assumption that $L, N > L, N - 1 > \ldots > L, 2 > 0$, the yield curve will be positively sloped even when no changes in rates are anticipated.

Another potential explanation of the Greenspan conundrum is consistent with the liquidity-preference theory. The prompt actions of the central bank to insure that future inflation is contained could engender

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
expectations of greater future economic stability and thus lower risk premiums throughout the economy.

The hedging-pressure or preferred habitat theory

Other critics of the expectations theory, including Culbertson (1957) and Modigliani and Sutch (1966), argue that liquidity considerations are far from the only additional influence on bond investors. While liquidity may be a critical consideration for a commercial banker considering an investment outlet for a temporary influx of deposits, it is not important for a life insurance company seeking to invest an influx of funds from the sale of long-term annuity contracts. Indeed, if the life insurance company wants to hedge against the risk of interest-rate fluctuations, it will prefer long, rather than short, maturities. Long-term investments will guarantee the insurance company a profit regardless of what happens to interest rates over the life of the contract.

Many pension funds and retirement savers find themselves in a wholly analogous situation. A retirement saver who has funds to invest in bonds for n periods will find an n-period pure discount (zero coupon) bond to be the safest investment. It is assumed that, if investors are risk averse, they can be tempted out of their preferred habitats only with the promise of a higher yield on a bond of any other maturity. Of course, other investors such as commercial banks or corporate investors will hedge against risk by confining their purchases to short-term issues. These investors will need higher yields on longer-term issues to induce them to invest in such securities. Under this hedging-pressure theory, however, there is no reason for term premiums to be necessarily positive or to be an increasing function of maturity.

Under an extreme (and somewhat implausible) form of the argument suggested by Culbertson, the short and long markets are effectively segmented, and short and long yields are determined by supply and demand in each of the segmented markets.

One of the popular explanations of the Greenspan conundrum was that many long-term investors, such as pension funds, were moving money out of the stock market and into the long-term bond market in order to ‘immunize’ their long-term liabilities. Such buying creates additional demand for long-term bonds, driving the yields of such securities down.

Empirical analysis of term structure theories

The chief obstacle to effective empirical analysis of the determinants of the term structure of interest rates has been the lack of independent evidence concerning expectations of future interest rates. Consequently, the first step in most empirical tests of the pure form of the expectations theory has been to set up some mechanism by which expectations may reasonably have been formed by market participants. Since people usually estimate the future by relying, at least in part, on historical information, this procedure has often involved the generation of forecasts of future interest rates from past values of these rates. Then investigators have sought to determine whether empirical yield curves have been consistent with these hypothetical forecasts and with the premise that investors, in fact, behave as the expectations theory claims. Thus, in essence, two theories were tested jointly: first, a theory of expectations formation, and second, a theory of the term structure. Of course, it is important to realize that any inability to confirm the expectations theory may be due to a failure to specify properly an expectations-forming mechanism rather than a failure of the theory to offer a correct explanation of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the shape of the yield curve. Nevertheless, the wide body of evidence we have does suggest a general conclusion.

The expectations-forming mechanisms utilized in empirical studies have been varied and inventive. They have included an error-learning mechanism (Meiselman, 1962); distributed lags on past rates (Modigliani and Sutch, 1966) or on inflation (Modigliani and Shiller, 1973; Fama, 1976); use of *ex post* data under an assumption that market efficiency and rationality require that *ex post* realizations do not differ systematically from ex ante views (Roll, 1970; 1971; Fama, 1984a; 1984b); and survey data assumed to reflect the actual expectations of market participants (Kane and Malkiel, 1967; Malkiel and Kane, 1969; Kane, 1983). While affirming the general importance of expectations in influencing the shape of the yield curve, empirical studies have generally rejected the pure form of the expectations hypothesis. There does appear to be an upward bias to the shape of the yield curve, indicating that term premiums do exist. But, contrary to the liquidity-preference theory, term premiums do not increase monotonically over the whole span of forward rates. Moreover, such term premiums vary over time. In addition, there appear to be seasonal patterns in the forward rates calculated from the short end of the yield curve.

Campbell (1995) points out that the pure expectations hypothesis implies that, whenever long yields exceed short yields, short yields should tend to rise in the future. It also implies that long yields must rise in the future so as to produce the capital losses that equate the short-term holding period returns between long and short maturities. He finds, instead, that mean excess returns on long bonds are positive. But excess returns on longer-term bonds do not rise throughout the maturity spectrum. The excess returns on zero coupon two-month bonds over one-month bills are positive, and excess returns over short bills rise with maturity at first, but after one year begin to decline and actually become negative for 10-year zero coupon bonds. And when the long-short yield spread is high, long yields have tended to fall, thus amplifying the yield differential between long and short bonds. Note that it is precisely this behaviour that Chairman Greenspan referred to as a ‘conundrum’ during 2005.

Affine yield models

More recent work on the term structure of interest rates has focused directly on how the shape of the term structure changes over time. The literature has evolved mostly in continuous time, and it is assumed that the future dynamics of the term structure depend on the evolution of some single factor or multiple factors that follow a stochastic process. The models are rooted in a framework consistent with the risk-adjusted expectations theory where arbitrage opportunities are not possible in equilibrium and where (log) bond yields are functions (which are often ‘affine’) of a single state variable that describes movements in future short-term rates or in a set of state variables related to the workings of the economy and the formation of expectations.

The pioneering models of this type were presented by Vasicek (1977) and Cox, Ingersoll and Ross (1985). These early papers developed single-factor models where the specific factor was the very short-term (instantaneous) default-free interest rate. Thus, all the information that is relevant for the determination of long-term rates was compressed into one stochastic process for very short rates. The process is a continuous time analogue to an autoregressive process where there is a fixed ‘normal’ short-term rate that can serve as an anchor for mean reversion. Once the diffusion process for short rates is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
specified, arbitrage is relied upon to explain the observed yields of bonds of different maturities. The models described above are special cases of the affine class of term structure models.

The models that followed have employed multiple factors but maintained the assumption that (log) bond prices are linear functions of the state variables. Duffie and Kan (1996) and Singleton and Dai (1999) presented models where the factors are the yields of *n* various fixed maturity bonds. Litterman and Scheinkman (1999) show that at least 95 per cent of the variation in yield changes can be explained by three latent factors, and interpret these factors in terms of the 'level' of yields, the 'slope' of the yield curve, and the 'curvature' of the curve. Further work has made significant strides in identifying and relaxing the restrictive assumptions of these models and in improving estimation techniques. Dewachter and Lyrio (2006) have jointly modelled the term structure as well as the dynamics of various macroeconomic variables. Their paper provides a macroeconomics interpretation of the Litterman–Scheinkman latent factors. They interpret the 'level' of yields factor as representing inflation expectations, the 'slope' of the yield curve factor as representing business-cycle conditions, and the 'curvature' factor as representing an independent monetary policy factor. Note that these factors are exactly those referenced earlier in an attempt to explain the Greenspan conundrum. A full discussion of affine yield models is contained in a survey article by Piazzesi (2006).

See Also

- affine term structure models
- bonds
- capital asset pricing model
- efficient markets hypothesis
- risk aversion

Bibliography

Campbell, J. 1995. Some lessons from the yield curve. *Journal of Economic Perspectives* 9(3), 129–52.

Cox, J., Ingersoll, J. and Ross, S. 1981. The relationship between forward prices and futures prices. *Journal of Financial Economics* 9, 321–46.

Cox, J., Ingersoll, J. and Ross, S. 1981. 1985. A theory of the term structure of interest rates. *Econometrica* 53, 385–407.

Culbertson, J. 1957. The term structure for interest rates. *Quarterly Journal of Economics* 71, 485–517.

Dewachter, H. and Lyrio, M. 2006. Macro factors and the term structure of interest rates. *Journal of Money, Credit, and Banking* 38, 119–40.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Duffie, D. and Kan, R. 1996. A yield factor model of interest rates. *Mathematical Finance* 6, 379–406.

Fama, E. 1976. Forward rates as predictors of future spot rates. *Journal of Financial Economics* 3, 361–77.

Fama, E. 1984a. The information in the term structure. *Journal of Financial Economics* 13, 509–28.

Fama, E. 1984b. Term premiums in bond returns. *Journal of Financial Economics* 13, 529–46.

Fisher, I. 1896. Appreciation and interest. *Economic Journal* 6, 567–70.

Greenspan, A. 2005. Statement of Alan Greenspan, Chairman, Board of Governors of the Federal Reserve System, before the Committee on Banking, Housing and Urban Affairs. United States Senate. February 16, 2005. Online. Available at http://banking.senate.gov/_files/ACF5A5C.pdf, accessed 20 January 2006.

Hicks, J. 1939. *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory*. Oxford: Clarendon Press, 1974.

Kane, E. 1983. Nested tests of alternative term-structure theories. *Review of Economics and Statistics* 65, 115–23.

Kane, E. and Malkiel, B. 1967. The term structure of interest rates: an analysis of a survey of interest-rate expectations. *Review of Economics and Statistics* 49, 343–55.

Litterman, R. and Scheinkman, J. 1999. Common factors affecting bond returns. In *The Debt Market*, ed. S. Ross. Northampton, MA: Edward Elgar.

Lutz, F. 1940. The structure of interest rates. *Quarterly Journal of Economics* 55, 36–63.

Malkiel, B. 1966. *The Term Structure of Interest Rates: Expectations and Behavior Patterns*. Princeton, NJ: Princeton University Press.

Malkiel, B. and Kane, E. 1969. Expectations and interest rates: a cross-sectional test of the error-learning hypothesis. *Journal of Political Economy* 77, 453–70.

Meiselman, D. 1962. *The Term Structure of Interest Rates*. Englewood Cliffs, NJ: Prentice Hall.

Modigliani, F. and Sutch, R. 1966. Innovations in interest rate policy. *American Economic Review* 56, 178–97.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Modigliani, F. and Shiller, R. 1973. Inflation, rational expectations and the term structure of interest rates. *Economica* 40, 12–43.

Piazzesi, M. 2006. Affine term structure models. In *Handbook of Financial Econometrics*, ed. Y. Ait-Sahalia and L. Hansen. Amsterdam: North-Holland.

Roll, R. 1970. *The Behavior of Interest Rates: An Application of the Efficient Market Model to U.S. Treasury Bills*. New York: Basic Books.

Roll, R. 1971. Investment diversification and bond maturity. *Journal of Finance* 26, 51–66.

Singleton, K. and Dai, Q. 2000. Specification analysis of affine term structure models. *Journal of Finance* 55, 1943–78.

Vasicek, O. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5, 177–88.

How to cite this article

Malkiel, Burton G. "term structure of interest rates." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000042>
doi:10.1057/9780230226203.1693

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

terms of trade

Ronald Findlay
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article describes the various approaches that have been made to the ‘terms of trade’, the question of determining and measuring the rate at which countries that engage in international trade divide the benefits between them. It traces the evolution of the relevant ideas, beginning with the pioneering contributions of David Ricardo, John Stuart Mill and Alfred Marshall.

Keywords

balance of trade; barter; commodity terms of trade; comparative advantage; double factoral terms of trade; gains from trade; gross barter terms of trade; immiserizing growth; income terms of trade; index numbers; international trade theory; Marshall, A.; Marshall–Lerner condition; Mill, J. S.; North–South economic relations; offer curve; optimal tariffs; Prebisch, R.; reciprocal supply and demand; Ricado, D.; shadow pricing; Singer, H.; single factoral terms of trade; stationary state; terms of trade; transfer problem

Article

The two most basic questions about international trade are, ‘What goods will each country export?’ and ‘What will be the ratios at which the exports of one country exchange for those of its trading partners?’ The first problem is that of ‘comparative advantage’; the second that of the ‘terms of trade’, which is the subject of the present article. David Ricardo, in Chapter 7 of the Principles (1817), gave a definitive answer to the first question and went a long way towards the solution of the second, though it was John Stuart Mill and Alfred Marshall who eventually gave the complete answer.
In the following discussion it will be convenient to assume, for simplicity, that there is only a single good exported and imported, and sometimes even that there is only a single factor of production, such as labour of a given quality. In practice, of course, we would have to use index numbers for unit values and physical volumes of exports and imports, giving rise to all the familiar problems (see index numbers).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Concepts and definitions

There are a number of alternative concepts and associated statistical measures of the terms of trade. The most prominent are listed below.

1. The commodity or net barter terms of trade is by far the most common meaning of the term, and is usually what is meant when the expression is used without any qualifying prefix. In principle, this is the relative price of the ‘exportable’ in terms of the ‘importable’, the number of units of the latter obtainable for each unit of the former. It has the dimensions of ‘nine waistcoat buttons for a copper disc’ in the words of Lewis Carroll's ‘Song of the Aged, Aged Man’ in *Through the Looking Glass*, words that D.H. Robertson (1952, ch. 13) used as the motto for a delightful essay on the terms of trade. In statistical practice, the commodity terms of trade are calculated as changes in the ratio of an export price index to an import price index, relative to a base year.

2. The *gross barter* terms of trade is a concept introduced by F.W. Taussig. It is the ratio of the *volume* of imports to the *volume* of exports. It coincides with the commodity terms of trade when trade is balanced, that is, there are no international loans or unrequited transfers. A deficit in the trade balance would cause the gross barter terms to be more favourable than the commodity or net barter terms, and vice versa. This should not, of course, be understood to mean that a trade deficit is necessarily preferable to balanced trade, since the additional imports now may have to be paid for by future trade surpluses.

3. The *income* terms of trade, sometimes also referred to as ‘the purchasing power of exports’ corresponds to the commodity terms of trade multiplied by the volume of exports. This is equal to the volume of imports under balanced trade, and exceeds or falls short of it if there is a surplus or deficit, respectively, in the balance of trade. In other words, it is the level of imports in real terms that can be sustained by current export earnings.

4. The single factoral terms of trade refers to the marginal or average productivity of a factor in the export sector, evaluated in terms of the imported good at the commodity terms of trade. The concept is meaningful for any single factor or production taken separately, though it is sometimes defined in a non-operational fashion in the literature as referring to ‘units of productive power’.

5. The double factoral terms of trade is an attempt to go behind the international exchange of commodities to the productive factors that are ‘embodied’ in them. Thus, if units are chosen, to the effect that a unit of labour in England produces a unit of cloth and a unit of labour in Portugal produces a unit of wine, commodity terms of trade of say five wine to one cloth would mean that a unit of English labour exchanges implicitly for five units of Portuguese labour in international trade.

The first three concepts of the terms of trade are all measurable in practice, subject to the usual index number problems. The commodity terms of trade are routinely calculated for most countries in the world by international agencies such as the United Nations (UN), the World Bank and the International Monetary Fund (IMF). The gross barter and income terms of trade have also been calculated for several countries.

The single factoral terms of trade, for any particular factor, can also be computed. Indeed it corresponds exactly to the concept of ‘shadow prices’ for primary inputs that has recently been developed in the literature on cost–benefit analysis in distorted open economies. Thus it could indicate what the value of a worker or an acre of land, engaged say in the coffee export sector, was worth in terms of imported food at the commodity terms of trade. This could serve as a valuable guide to resource allocation if a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
comparison were made with what these resources could produce in the domestic food sector.
The double factoral terms of trade, however, is either misleading if it is computed for any particular single factor in a world of more than one scarce input, or non-operational if defined amorphously as applying to units of ‘productive power’. This concept has been regarded by more than one economist as the fundamental one, and no less an authority than D.H. Robertson, in the essay referred to above, called it the ‘true’ terms of trade. Equally eminent authorities, such as Haberler (1955) and Viner (1937), have, however, been more sceptical, and rightly so.
The concept has recently come to the fore again, after many decades of neglect, in connection with the theories of A. Emmanuel (1972) on ‘unequal exchange’ in trade between high-wage and low-wage countries, which he regards as a form of ‘exploitation’ of the latter by the former. It is possible to interpret Emmanuel as saying that it is only when the double factoral terms of trade are equal to unity that there is no unequal exchange. As Emmanuel himself acknowledges, however, his argument requires equal capital intensity in the export sectors of the trading partners. We may all agree with Robert Burns that ‘A Man's a Man for a’ That’ in terms of dignity and spiritual worth. It is another thing, however, to say that skill or physical capital, both accumulated at some cost, should count for nothing, and that the only ‘fair’ exchange is one that takes place according to the simple labour theory of value.
Furthermore, it is clear that the commodity terms of trade can improve while the factoral terms worsen, and vice versa. Thus suppose that, initially, one day's labour in ‘North’ and ‘South’ produces a unit of steel and coffee, respectively, and that the commodity terms of trade was one unit of steel for one unit of coffee. Suppose now that one worker in the North produces three steel, while his counterpart in the South still produces only one coffee. Let the commodity terms of trade now be two units of steel for one of coffee. The commodity terms have doubled in favour of the South, while its factoral terms have deteriorated to two-thirds instead of unity. Which situation would the South prefer?

Fundamental determinants

Ricardo did not determine the terms of trade explicitly in his analysis in Chapter 7 of the *Principles*. He was only able to show that the equilibrium value would be between the comparative cost ratios of the two countries, specified by the linear technologies. It was John Stuart Mill (1844) who solved the problem by his numerical example of ‘reciprocal supply and demand’, later refined by Marshall (1930) through the geometric device of the ‘offer curves’ showing the excess supplies and demands of the two goods in each country as functions of the terms of trade, the equilibrium value of which would be determined by setting world excess supply equal to zero. Marshall demonstrated the possibility of multiple equilibria and also established a criterion for stability of equilibrium that is in use to this day, in the form of the so-called Marshall–Lerner condition that the sum of the import demand elasticities has to be greater than unity.
In modern terms it is the preferences of the consumers that have to be introduced to close the model. Once these are introduced the equilibrium value(s) of the terms of trade are determined as a function of these preferences, the labour endowments and the technical coefficients of production. The subsequent development of the literature has generalized Ricardo's analysis to any number of goods, factors and countries and to variable instead of fixed technical coefficients. The determination of the terms of trade is thus technically nothing other than that of finding the equilibrium vector(s) of relative prices for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
general equilibrium models in which there is a world market for tradable goods and internationally mobile factors, and national markets for non-traded goods and internationally immobile factors.
In addition to constituting a central problem for the theory of international trade in its ‘positive’ aspect, the terms of trade plays if anything an even more critical role in the ‘normative’ dimension of evaluating the ‘gains from trade’. It is crucial to keep these two facets of the terms of trade conceptually distinct, though of course they are both involved in almost every theoretical or policy problem. Another essential distinction is between the terms of trade as an exogenously determined parameter, as in the ‘small’ open economy models, and as an endogenously determined variable, the equilibrium value of which is altered by some change in circumstances or parameters, such as factor endowments, technology or tastes. Much confusion has been caused in the literature by failure to bear these basic distinctions in mind at all times.
In the realm of positive theory, the terms of trade generally appear in comparative static exercises as the key dependent variable, upon which the effect of some exogenous shock is sought. As an example, consider the effect of a switch in the composition of home demand in favour of the imported commodity. At constant terms of trade this would create an excess demand for the imported good. On the assumption of Walrasian stability, this must lead to a deterioration of the home country's terms of trade for the world market to return to equilibrium.
The famous ‘transfer problem’ is another example of this sort of comparative static exercise. The transfer of purchasing power at constant terms of trade would lead to an excess supply in the world market of the transferor's exportable, if the home propensity to consume this good is greater than that of the recipient country (the so-called ‘classical presumption’). Thus the terms of trade of the transferor would deteriorate, given Walrasian stability, imposing a ‘secondary burden’ on the transferor.
Finally, we may consider the effects of economic growth, in the form of exogenous changes in factor endowment or technical innovations in either sector, a literature that was stimulated by Hicks's (1953) inaugural lecture on the ‘dollar shortage’. Here again the analysis consists in finding the effect of the change on excess supply or demand at constant terms of trade, and thus obtaining the direction of movement in the terms of trade necessary to clear the market, assuming stability in the Walrasian sense.

Welfare effects

All these exercises in positive theory of course have welfare consequences for both trading partners. In the case of the two-country transfer problem the transferor is worse off, even if the terms of trade were to move in its favour, while the recipient is better off, even if the terms of trade were to turn against it. In the case of a shift in the composition of home demand towards imports the welfare of the trading partner will rise under normal conditions (see below) as a result of this improvement in its terms of trade. If growth in one country creates an excess demand for imports at constant terms of trade, its passive partner will also benefit from the resulting increase in the relative price of its export.
In the last two cases a country experiences an exogenous improvement in its terms of trade, with no alteration in its own preferences, technology or factor endowment. Must its welfare necessarily increase as a result? The answer in general is ‘yes’, unless there are domestic distortions such as monopoly or monopsony in product or factor markets, exogenous wage differentials or real factor–price rigidities. A simple example of how it is possible for a country to experience a loss in welfare as a result of an improvement in its commodity terms of trade can be constructed as follows. Suppose that domestic production is completely specialized on the export good and that the real wage is fixed in terms of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
imported good. At constant employment, and therefore constant marginal physical productivity of labour in terms of the exported good, the real wage would be lower in terms of the imported good because of the improvement in the terms of trade. This will induce a decline in employment and output until the marginal physical product of labour rises in the same proportion as the relative price of the imported good has fallen, so that the original level of the real wage is restored. The terms of trade improvement, given employment and output, increases welfare, but the contraction in these variables induced by the change in the terms of trade reduces welfare. This negative effect can clearly be sufficient to outweigh the positive effect of the terms of trade gain considered in isolation, since the counteraction can be very sharp if the marginal productivity of labour schedule is assumed to be sufficiently elastic.

Haberler (1955, p. 30), in a characteristically penetrating and judicious discussion of the subject, has stated that ‘other things being equal an improvement in the commodity terms of trade does imply an increase in real national income’. As our analysis of the example in the previous paragraphs shows, however, even such a cautious formulation needs to be interpreted with care. It would obviously be a mistake to compound the welfare effects of an exogenous shift in the terms of trade with the direct welfare effects of some independent shock. In our example, however, the terms of trade change was the sole shift in the data, the contraction of employment and output being induced by this very change in the terms of trade itself.

When the change in the terms of trade is a consequence of some exogenous shock, such as a change in tastes, technology or factor endowment, it is clearly erroneous to infer the total change in welfare solely from the direction of change in the terms of trade. Technological progress in the export sector that causes deterioration in the terms of trade can obviously leave a country better off in spite of the deterioration, even though it would of course have been still better off if the terms of trade had remained unchanged. It is this sort of consideration that has led to the introduction of concepts such as the factoral terms of trade, since these measures could show an improvement even when the commodity terms of trade deteriorate. In general, however, it is a mistake to expect any single concept of the terms of trade to be an unambiguous indicator of changes in the gains from trade when there are shifts in the fundamental determinants of tastes, technology and factor endowments.

The welfare effects of such changes can be broken into two parts: first, the effect at unchanged terms of trade and second, the effect of the associated change in the terms of trade. The net effect on welfare may thus be positive or negative and need not correspond with the direction of the change in the terms of trade. Bhagwati (1958) established the possibility that the net effect on welfare of the country experiencing economic growth can be negative, a phenomenon that he termed ‘immiserizing growth’.

Finally, we may consider the terms of trade as an objective of policy, when the country has some degree of monopoly power in international markets. The consideration of a rational policymaker, ignoring the possibility of retaliation, would be to restrict trade to such an extent as to equate at the margin the benefit resulting from the improvement in the terms of trade with the loss of welfare resulting from the decline in the volume of trade. This is the famous ‘optimum tariff’ argument, the level of which varies inversely with the elasticity of foreign demand for imports.

Secular tendencies

In addition to comparative static analyses of the type considered up to now, the literature also contains some more speculative hypotheses about secular tendencies in the terms of trade. In the Ricardian

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
tradition capital accumulation and technical progress lead to a steady expansion in the supply of manufactures, while the supply of primary products is always constrained by the limited availability of ‘land’ and other natural resources. Ricardo's theorem for a closed economy – that growth would raise the relative price of food and therefore the rent of land until a ‘stationary state’ is approached – has been extended to the world economy in the form of a presumption that there would be a tendency for the terms of trade to move against manufactures and in favour of primary products. Keynes, Beveridge, Robertson and E.A.G. Robinson all took part in a long-running debate on this issue. The story that W.S. Jevons kept enormous stocks of coal in his basement is a bizarre manifestation of this phobia. W.W. Rostow (1962, chs 8 and 9) gives a very interesting review and analysis of this literature, which foreshadows the views associated with the Club of Rome on the depletion of exhaustible natural resources.

Discussions of the secular tendencies of the terms of trade since the Second World War, however, have been dominated by the view of Raul Prebisch (1950) and Hans Singer (1950) that the historical record shows a long-run tendency for the commodity terms of trade of the less developed countries to deteriorate. The evidence was a series showing an apparent long-run improvement in Britain's terms of trade between 1870 and 1940. Theoretical reasons given for the alleged tendency have been lower income-elasticity of demand for primary products than for manufactures, technical progress that economizes on the use of imported raw materials and monopolistic market structures in the industrial countries combined with competitive conditions in the supply of primary products. The general consensus on the statistical debate that has arisen on this issue is that there has not been any discernible secular trend for the commodity terms of trade of the developing countries to deteriorate (see Spraos, 1980, for a summary and assessment of the evidence; Lewis, 1969, presents an interesting alternative theoretical and empirical analysis of this problem). Hadass and Williamson (2001) provide a useful recent summary and critique of the ongoing debate on the Prebisch–Singer thesis.

The Prebisch–Singer hypothesis and the more general concerns of the ongoing North–South dialogue have also spawned a number of so-called ‘North–South’ models, in which the interaction of an advanced industrial region with a less developed and structurally dissimilar, labour-abundant, primary producing region is studied in a dynamic context. The terms of trade play a key role in these models, since the growth rate of the South is linked to this variable through dependence on capital goods imported from the North. This and other analytical issues related to secular trends in the terms of trade are further discussed in Findlay (1980; 1984), Taylor (1981) and Darity (1990).

See Also

- comparative advantage
- Heckscher–Ohlin trade theory
- index numbers
- Prebisch, Raúl

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bhagwati, J. 1958. Immiserizing growth: a geometrical note. *Review of Economic Studies* 25, 201–5.

Darity, W. 1990. Fundamental determinants of the terms of trade reconsidered: long-run and long-period equilibrium. *American Economic Review* 80, 816–27.

Emmanuel, A. 1972. *Unequal Exchange*. New York: Monthly Review Press.

Findlay, R. 1980. The terms of trade and equilibrium growth in the world economy. *American Economic Review* 70, 291–99.

Findlay, R. 1984. Growth and development in trade models. In *Handbook of International Economics*, vol. 1, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.

Haberler, G. 1955. *A Survey of International Trade Theory*. Princeton: International Finance Section.

Hadass, Y. and Williamson, J.G. 2001. Terms of trade shocks and economic performance 1870–1940: Prebisch and Singer revisited. Working Paper No. 8188. Cambridge, MA: NBER.

Hicks, J.R. 1953. An inaugural lecture. *Oxford Economic Papers* 5, 117–35.

Lewis, W.A. 1969. *Aspects of Tropical Trade 1883–1965*. Stockholm: Almqvist and Wiksell.

Marshall, A. 1930. *The Pure Theory of Foreign Trade*. Clifton, NJ: Augustus M. Kelley, 1974. (Privately printed, 1879.)

Mill, J.S. 1844. Of the laws of interchange between nations; and the distribution of the gains of commerce among the countries of the commercial world. In *Essays on Some Unsettled Questions of Political Economy*. London: London School of Economics, 1948.

Prebisch, R. 1950. *The Economic Development of Latin America and its Principal Problems*. New York: United Nations.

Ricardo, D. 1817. *On the Principles of Political Economy and Taxation*. Cambridge: Cambridge University Press, 1951.

Robertson, D.H. 1952. *The Terms of Trade: Chapter 13 of Utility and All That*. New York: Macmillan.

Rostow, W.W. 1962. *The Process of Economic Growth*. New York: Norton.

Singer, H.W. 1950. The distribution of gains between investing and borrowing countries. *American Economic Review, Papers and Proceedings* 5, 473–85.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Spraos, J. 1980. The statistical debate on the net barter terms of trade between primary commodities and manufactures. *Economic Journal* 90, 107–28.

Taylor, L. 1981. North-south trade and southern growth: bleak prospects from a structuralist point of view. *Journal of International Economics* 11, 589–601.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper Bros.

How to cite this article

Findlay, Ronald. "terms of trade." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000044> doi:10.1057/9780230226203.1694

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

terrorism, economics of

S. Brock Blomberg and Gregory D. Hess

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides a comprehensive study of the economic determinants of domestic and transnational terrorism and the role that the economy plays in fostering a more peaceful world. We describe the research associated with the microfoundations of terrorist groups and how they organize. We also analyse models of conflict resolution to investigate the relative importance of macroeconomic factors for domestic and transnational terrorism. We describe a number of data-sets employed by researchers in the field and end by describing the most recent research which investigates the linkages between terrorism, democratization, globalization and development.

Keywords

civil conflict; democratization; development; foreign direct investment; globalization; gravity models; international trade; terrorism, economics of; war and economics

Article

Since the prominent post-2000 terrorist incidents in high-income cities such as New York, Madrid and London, and the persistent incidence of terrorism in Middle East countries such as Israel and Iraq, both academia and the media have become involved in a careful examination of its causes. Terrorism is, however, neither new nor novel – indeed, the very origin of the term points to a long history, dating back to the late 1700s. (The word ‘terrorism’ apparently first appeared in the English language in reference to the ‘reign of terror’ associated with the rule of France by the Jacobins from 1793–94. The first incidence was actually reported in first century bc when Jewish terrorists, Zealots-Sicarri, incited a riot which led to a mass insurrection against the Roman Empire. See Laqueur, 1977, pp. 7–8.) While terrorism has been present for longer than one might realize, research on the economics of terrorism has a shorter history. The literature has primarily focused on two areas: the microfoundations of terrorism – understanding why organizations employ terrorist tactics – and the macroeconomic causes and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
consequences of these tactics. The latest wave of research highlights the relationship between terrorism, globalization and democratization.

Microeconomics of terrorism

Microeconomic-based (or at least economic-sympathetic) research in terrorism generally concludes that terrorist organizations behave rationally (see, for example, Bueno de Mesquita, 2005). So, while at key junctures terrorists initiate futile, high-cost (to terrorists' own selves) insurgencies and lose sight of whether the tactics really bring them closer to their political goals, for example, the behaviour is likely more driven by the clandestine nature of organizations and the daily struggle to maintain operational security by distorting rebels' understanding of the world outside their organization (see Bell, 1990; 2002). This also points to the utility of propagating ideological movements in terrorism.

To whit, Hudson's review (1999) shows that for every study finding a purported psychological regularity, or social psychology, there exist a contradictory study. This is consistent with Krueger and Maleckova's (2003) finding that terrorists do not come from the poorer or uneducated segments of society. Given that organizations prefer to send competent terrorists, and thus select for the highest qualified, these results are best interpreted as showing that irrationality and individual oppression are not the underlying causes of terrorist incidents.

Moreover, if ideology plays a role in terrorist recruitment, its focus has clearly evolved. With respect to the spreading of revolutionary ideologies, there has been a change in the motive of terrorists since the November 1979 takeover of the US embassy in Tehran, the Iranian capital. Up to that point, terrorism had been primarily motivated by revolutionary and separatist ideologies – for example, the Red Army or Shining Path (see Wilkinson (2001). Since then, religious-based fundamentalism has played a more primary role. For example, the share of religious-based terrorist organizations had grown from four per cent to over 50 per cent by 1995 (Hoffman, 1997).

However, it is misguided to search for the causes of terrorism in ideologies such as radical Islam. Rather, researchers such as Mishal and Sela (2002), Wilhelmsen (2004) and Brynjar and Hegghammer (2004) and Pape (2005) document that rational calculations rather than fundamentalist fury guide terrorist organizations' decision-making. From the perspective of leadership, the evidence is strong that terrorists strategically choose targets to best realize their political agenda. (In turn, Enders and Sandler, 2002, analyse the government's response to the innovations in the supply of terrorism.)

Macroeconomics of terrorism

Given that terrorist behaviour can be described within a rational choice framework, many researchers have constructed models in order to evaluate the macroeconomic costs and consequences of terrorism based on Grossman (1991). Grossman provides the seminal economics paper investigating the integral linkages between civil conflict and the economy. He presents a general equilibrium model that treats insurrection and the suppression of insurrection as economic activities willingly undertaken by the participants. The ruler has to trade off higher taxes not only against the lower tax revenue that comes about when people devote less time to productive activities but also against the added cost of having to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
hire soldiering services to suppress insurrection. Grossman finds that economies in which the soldiering technology is effective can move themselves to no-conflict equilibria by devoting some resources to soldiering and keeping tax rates low.

With respect to the link between terrorism and the economic environment, Blomberg, Hess and Weerapana (2004a) present a model that describes how one factor – the state of the economy – can lead groups to resort to terrorist attacks. Other authors such as Bernholz (2004) and Wintrobe (2002) have studied the important influences of increased fundamentalism and group solidarity in driving terrorist activity. Still, it is important to note that economic conditions help identify the underlying determinants of these activities. (There is also an existing literature that analyses how economics influences conflict in general. However, most of the analysis to this point has considered the impact on conflicts such as war without considering alternative types of conflict such as terrorism. For example, Hess and Orphanides, 1995; 2001a; 2001b, estimate the probability of conflict for the United States doubles when the economy has recently been in a recession and the president is running for re-election. Abadie and Gardeazabal, 2003, also find a strong relation between the economy and terrorism.) The evidence is best summarized in Blomberg, Hess and Weerapana (2004b) who provide an analysis of the relationship between economic growth phases (for example, expansions and contractions) and transitions into and out of terrorism incidents.

Violence, however, knows no bounds or lack of imagination. As such, it is often challenging to type forms of conflict into discrete, unbending categories. Acknowledging this point, there is a literature analysing the economic impact of terrorism as opposed to other forms of violence. Blomberg, Hess and Orphanides (2004), investigate the impact of various forms of conflict such as terrorism, internal wars and external wars on a country's economic growth. They find that, on average, the incidence of transnational terrorism has a significantly negative effect on growth, albeit one that is considerably smaller and less persistent than that associated with either external wars or internal conflict. They also find that terrorism is associated with a redirection of economic activity away from investment spending and towards government spending.

Other authors have concentrated on analysing one country in particular or on isolating one economic channel through which terrorism harms growth. Eckstein and Tsiddon (2004) provide an analysis of the macroeconomic consequences of terrorism in Israel, and find a large impact of domestic terrorism on economic activity. Using bilateral trade data, Blomberg and Hess (2006) establish that terrorism has a diminishing effect on international trade, and Blomberg and Mody (2006) demonstrate that violence also has a negative impact on foreign direct investment. Glick and Taylor (2004) also provide an analysis of the effect of external conflict on international trade over a longer historical period.

Terrorism data and empirical regularities

To test theories of terrorism as well as to estimate the costs of terrorism, researchers need reliable measures of terrorism. Several competing international data-sets for terrorist incidents have been cataloging attacks since the late 1960s. The dynamics across the major data-sets are roughly similar. In each data-set, the number of events increases during the period 1969 to 1987 (see the discussion in Blomberg and Hess, 2007a; 2007b). The US State Department and ITERATE data-sets estimate a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
similar steady increase from approximately 100 to 200 incidents per year up to 500 to 600 incidents per year. The RAND data-set estimates a similar trend, though the levels are smaller (from a base of approximately 100 incidents per year). The difference in the levels of terrorism in these data-sets arises because RAND does not include terrorism from state actor to non-state actor within a country and so systematically underestimates the number of attacks as compared with the other data-sets.

These trends demonstrate that the number of terrorist incidents steadily increased and peaked in the mid- to late 1980s. For several years thereafter, the worldwide intensity of transnational violence – violence motivated by international political considerations – fell steadily. In the late 1980s, according to the ITERATE data-set (see Mickolus et al., 2004), approximately one-and-one-half transnational violent events occurred every day. This frequency declined to less than one-half of an event a day by 2000. The decline also indirectly implies that the number of countries affected by a violent event fell over that period. This point has been more seriously addressed in Enders and Sandler (2005), who demonstrate that there has been no increase in violence from terrorism since the Al-Qaeda attacks on the United States on 11 September 2001. They show, if anything, that terrorism has fallen. However, over the time period in question the number of violent episodes may have risen – particularly since 11 September 2001. The average number of deaths per incident was 0.83 during 1968 to 1993. In seven of the following ten years, the number of deaths per incident was higher. Over the entire sample (1968–2003) there has been about one death per incident, and since 2001 the average has been five times that rate. Blomberg and Hess (2006; 2007a; 2007b) demonstrate two additional points. First, the recent drop in terrorism is systematic across regions, governments, income classes, and degrees of openness. Second, the hot spots for terrorism, as measured by incidence per capita, appear to be richer democracies, economies more open to trade and Middle Eastern countries.

Even though there is less systematic data available, terrorism was prominent before the 1960s and has evolved since its inception. In the late 19th and early 20th centuries, terrorists targeted political figures, the most notable examples being the 1881 assassination of Alexander II in Russia and the 1934 assassination of Alexander I of Yugoslavia. Current weapons and computer technology have transformed terrorists into literal artillery weapons causing a large amount of damage with only a few conspirators. Finally, many of these targets are now civilians rather than actual political heads.

Terrorism, globalization, democratization and development

The changing dynamics of terrorism has led to an emerging set of papers investigating the relationship between globalization, democratization, development and terrorism. At its best, democracy provides a framework that aids the peaceful resolution of political conflicts (see Hess and Orphanides, 2001a, for a discussion of how democracy affects the likelihood of external conflict). It offers access to decision makers and political institutions for citizens, while it also provides checks and balances within competing branches of the government. It also makes political organization cheaper and lowers the costs of political action. In turn, democracies should make illegal activities more expensive than legitimate political activity. In expectation, therefore, there should be less terrorist violence in democracies. Alternatively, the key to the success of any terrorist act is recruitment and organization – both of which are made easier in free societies. Ironically, characteristics of democracies such as civil liberties and freedom of religion, association and movement could actually facilitate terrorist organization building.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Moreover, as terrorists often target innocent civilians, free speech and a free press (hallmarks of democracy) may be good channels for spreading fear. As a consequence, democracies could actually foster terrorist activities.

Is terrorism associated with the positive attributes of a democracy? Eubank and Weinberg (1994; 2001) find that terrorist groups are in fact more frequently hosted by democratic societies. Similarly, Li and Schaub (2004) find more incidents in democratic countries. Still, it may not be democracy that is the true driving force at work here; rather, the process of democratization may be the real culprit. For example, the experience of less democratic or newly democratizing countries such as Afghanistan and Iraq suggests that the transitional period between authoritarianism and democracy is a particularly susceptible one for terrorist activity (see Eubank and Weinberg, 1998). Thus, the linkages between terrorism incidence and the evolution of a country's institutional governance are an important area for further study.

Other evidence on the link between democracy and transnational terrorism is decidedly mixed. Li (2005) attempts to disaggregate the many dimensions of democracy; he finds that voter turnout reduces terrorist incidents, but that constraints on government authority increase incidents. Finally, he finds press freedom raises incidents in a country. Taken as a whole, therefore, the effect of democracy on terrorism is not straightforward.

Globalization also affects the economic environment in which terrorist organizations can operate. If terrorism emerges from an environment of economic deprivation, then globalization, in so far as it enhances economic growth, may offset terrorist tendencies. Alternatively, if globalization increases inequality across countries and groups, then we might expect globalization to lead to more violence. Furthermore, globalization's associated lowered barriers to flows of goods, factors of production (including labour) and finance could make a network of terrorist operations cheaper to operate. Overall, globalization, like democracy, affects the costs, benefits and resources constraints of terrorists in many ways. Learning whether or not globalization is a net contributor to terrorism is therefore an empirical matter. Still, there is ample theory to support either conjecture.

Krug and Reinmoeller (2004) argue that globalization is an important determinant of terrorism. They build a model to explain the internationalization of terrorism as a natural response to a globalizing economy. As countries become more economically integrated and market-oriented, there is no discrimination between what certain terrorist groups might see as bad products and good products or investments. Moreover, the same advances in technology that allow for easy access of goods and services also allow for easy access to military hardware and technology. In the short run, globalization may have the consequence of creating a series of winners and losers. These same losers will find it easier to retaliate in response to their losses, thereby multiplying the effect of globalization on terrorism. (In a pooled cross-section analysis of globalization and transnational terrorism, Li and Schaub, 2004, find that international trade and investment have little effect on the number of terrorist events.)

An alternative view put forth by Crenshaw (2001) is that it is naive to believe that globalization encourages international terrorism; while globalization and terrorism may seem to affect one another, there is something more complicated at work. She argues that the latest wave of terrorism should be seen as a series of civil wars which may be a strategically unified reaction to American power rather than directly to globalization.

In a series of papers, Blomberg and Hess (2006; 2007a) and Blomberg and Rosendorff (2007) are able to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
uncover several important linkages between globalization, democratization, development and terrorism by employing the workhorse model in the international trade literature – the gravity model. Terrorism is defined by the fact that 'its ramifications transcend national boundaries' (see Mickolus et al., 2004). Transnational terrorism requires, therefore, a flow of resources across borders, and these papers consider both the source and target countries' characteristics in determining terrorism. The characteristics of a country that might make it a likely target country may indeed be very different from the characteristics that make a country a likely source of international terrorism. The features of the polity that make a country a terrorist-producer may be different from the political structures, institutions and environment that make a state a terrorist target. The conclusion from these papers is that the advent of democratic institutions, high income and more openness in a *source* country significantly reduces conflict. However, the advent of these same positive developments in *targeted* countries actually increases conflict. *Ceteris paribus*, the impact of being a democracy or participating in the WTO/IMF for a *source* country decreases the number of terrorist strikes by about two or three per year, which is more than two standard deviations greater than the average number of strikes between any two countries in a given year. Terrorism is a nascent field in economics. Sadly, though realistically, it is also likely to be a growing field in economics. Conflict and the economy are intimately related at the domestic and at the international levels, and across various forms of governance. Moreover, innovations in the strategy of terrorism will continue (see, for example, Berman and Laitin's analysis, 2005, of suicide attacks). Terrorism therefore will remain part of the violence spectrum and affect our economic landscape for years to come.

Bibliography

Abadie, A. and Gardeazabal, J. 2003. The economic costs of conflict: a case study of the Basque country. *American Economic Review* 93, 113–32.

Bell, J.B. 1990. Revolutionary dynamics: the inherent inefficiencies of the underground. *Terrorism and Political Violence* 2, 193–211.

Bell, J.B. 2002. The organization of Islamic terror: the global jihad. *Journal of Management Inquiry* 11, 261–6.

Berman, E. and Laitin, D. 2005. Hard targets: theory and evidence on suicide attacks. Working Paper No. 11740. Cambridge, MA: NBER.

Bernholz, P. 2004. Supreme values as the basis for terror. *European Journal of Political Economy* 20, 317–33.

Blomberg, S.B. and Hess, G.D. 2006. How much does violence tax trade? *Review of Economics and Statistics* 88, 599–612.

Blomberg, S.B. and Hess, G.D. 2007a. The Lexus and the olive branch. In *Economic Consequences of*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Terrorism in Developed and Developing Countries, ed. P. Keefer and N. Loayza. Cambridge University Press.

Blomberg, S.B. and Hess, G.D. 2007b. From (no) butter to guns? Understanding the economic role in terrorism. In *Economic Consequences of Terrorism in Developed and Developing Countries*, ed. P. Keefer and N. Loayza. Cambridge University Press.

Blomberg, S.B., Hess, G.D. and Orphanides, A. 2004. The macroeconomic consequences of terrorism. *Journal of Monetary Economics* 51, 1004–30.

Blomberg, S.B., Hess, G.D. and Weerapana, A. 2004a. An economic model of terrorism. *Conflict Management and Peace Science* 21, 17–28.

Blomberg, S.B., Hess, G.D. and Weerapana, A. 2004b. Economic conditions and terrorism. *European Journal of Political Economy* 20, 463–78.

Blomberg, S.B. and Mody, A. 2006. How severely does violence deter international investment? Working paper, International Monetary Fund.

Blomberg, S.B. and Rosendorff, P. 2007. A gravity model of globalization, democracy and transnational terrorism. In *Guns and Butter: The Economic Causes and Consequences of Violent Conflict*, ed. G. Hess. Cambridge, MA: MIT Press.

Brynjar, L. and Hegghammer, T. 2004. Jihadi strategic studies: the alleged Al Qaida policy study preceding the Madrid bombings. *Studies in Conflict and Terrorism* 27, 355–75.

Bueno de Mesquita, E. 2005. The quality of terror. *American Journal of Political Science* 49, 515–30.

Crenshaw, M. 2001. Why America? The globalization of civil war. *Current History* 100, 425–32.

Eckstein, Z. and Tsiddon, D. 2004. Macroeconomic consequences of terror: theory and the case of Israel. *Journal of Monetary Economics* 51, 971–1002.

Enders, W. and Sandler, T. 2002. Patterns of transnational terrorism, 1970–1999: alternative time-series estimates. *International Studies Quarterly* 46, 145–65.

Enders, W. and Sandler, T. 2005. After 9–11: Is it all so different now? *Journal of Conflict Resolution* 49, 259–77.

Eubank, W. and Weinberg, L. 1994. Does democracy encourage terrorism? *Terrorism and Political Violence* 6, 417–43.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Eubank, W. and Weinberg, L. 1998. Terrorism and democracy: what recent events disclose. *Terrorism and Political Violence* 10, 108–18.

Eubank, W. and Weinberg, L. 2001. Terrorism and democracy: perpetrators and victims. *Terrorism and Political Violence* 13, 155–64.

Glick, R. and Taylor, A. 2004. Collateral damage: the economic impact of war. Working Paper No. 11565. Cambridge, MA: NBER.

Grossman, H.I. 1991. A general equilibrium model of insurrections. *American Economic Review* 81, 912–21.

Hess, G.D. and Orphanides, A. 1995. War politics: an economic, rational-voter framework. *American Economic Review* 85, 828–46.

Hess, G.D. and Orphanides, A. 2001a. Economic conditions, elections, and the magnitude of foreign conflicts. *Journal of Public Economics* 80, 121–40.

Hess, G.D. and Orphanides, A. 2001b. War and democracy. *Journal of Political Economy* 109, 776–810.

Hoffman, B. 1997. The confluence of international and domestic trends in terrorism. *Terrorism and Political Violence* 9, 1–15.

Hudson, R.A. 1999. *The Sociology and Psychology of Terrorism: Who Becomes a Terrorist and Why?* Washington, D.C.: Federal Research Division, Library of Congress.

Krueger, A.B. and Maleckova, J. 2003. Education, poverty and terrorism: is there a causal connection? *Journal of Economic Perspectives* 17(4), 119–44.

Krug, B. and Reinmoeller, P. 2004. The hidden cost of ubiquity: globalisation and terrorism. Research paper, Erasmus Research Institute of Management, RSM Erasmus University.

Laqueur, W. 1977. *Terrorism*. London: Weidenfeld and Nicolson.

Li, Q. 2005. Does democracy promote or reduce transnational terrorist incidents? *Journal of Conflict Resolution* 49, 278–97.

Li, Q. and Schaub, D. 2004. Economic globalization and transnational terrorism: a pooled time-series analysis. *Journal of Conflict Resolution* 48, 230–58.

Mickolus, E.F., Sandler, T., Murdock, J.M. and Flemming, P. 2004. *International Terrorism: Attributes*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Terrorist Events, 1968–2003 (ITERATE 5). Dunn Loring, VA: Vinyard Software.

Mishal, S. and Sela, A. 2002. Participation without presence: Hamas, the Palestinian authority and the politics of negotiated existence. *Middle Eastern Studies* 38(3), 1–26.

Pape, R. 2005. *Dying to Win*. New York: Random House.

Wilhelmsen, J. 2004. *When Separatists become Islamists: The Case of Chechnya*. Report No. 2004/00445. Kjeller: Norwegian Defence Research Establishment (FFI).

Wilkinson, P. 2001. *Terrorism Versus Democracy: the Liberal State Response*. London: Frank Cass.

Wintrobe, R. 2002. Can suicide bombers be rational? Unpublished manuscript, University of Western Ontario.

How to cite this article

Blomberg, S. Brock and Gregory D. Hess. "terrorism, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_E000268> doi:10.1057/9780230226203.1695

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

testing

Frank Kleibergen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Hypothesis testing is the customary instrument for analysing the empirical validity of an economic theory. Hypothesis testing is thus an important tool for conducting statistical inference in economic models. In this article we show how an economic theory is tested in a statistical model. We begin with the discussion of the basic results on hypothesis testing and then focus on some recent developments that have improved testing in commonly used economic models such as the linear instrumental variables regression model. We use a real economic example to illustrate the main findings.

Keywords

Anderson–Rubin statistic; bootstrap; generalized method of moments; hypothesis testing: *see* testing; Lagrange multipliers; least squares; likelihood ratios; limited information maximum likelihood; linear models; maximum likelihood; Neymann–Pearson Lemma; price elasticity; probability; statistical inference; testing; two-stage least squares; Wald statistics

Article

Hypothesis testing is the customary instrument for analysing the empirical validity of an economic theory. This theory is reduced to a hypothesis which is tested in a statistical model. Hypothesis testing is thus an important tool for conducting statistical inference in economic models. An impressive literature has emerged which discusses tests of economic hypotheses. Instead of providing an incomplete overview of this literature, we provide a somewhat hands on discussion of testing in which we show how an economic theory is tested in a statistical model. We therefore begin with the discussion of the basic results on hypothesis testing and later focus on some recent developments that have improved testing in commonly used economic models. We use a real economic example to illustrate the main findings.
When testing an economic hypothesis, we want our test results to hold generally and not to be affected by highly specific assumptions on the statistical model such as, for example, the distribution of the disturbances. Under these general conditions, the finite sample distributions of the involved test statistics are unknown. The realized values of the test statistics are then confronted with critical values that result from the large sample distributions of the statistics under the hypothesis of interest. Since this is currently the common approach to testing, our discussion is conducted solely from the large sample perspective.
We illustrate the tests of an economic theory by testing for a unit demand price elasticity for the demand for oranges. We use data on the demand for oranges in the United States during 1910–59. The data results from Nerlove and Waugh (1961) and is also used in Berndt (1991, pp. 417–20). The demand equation is specified as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$log(P_t) = \alpha + \gamma log(Q_t) + \beta log(RI_t) + \varepsilon_t, t = 1, \ldots, T, \quad (1)$$

with Q_t the traded quantity, P_t the price of oranges and RI_t the real disposable income. The other available series are current (AC_t) and past advertisement (AP_t, averaged over the last ten years). We test the hypothesis of a unit demand price elasticity $H_0: \gamma = -1$ against the alternative hypothesis $H_1: \gamma \neq -1$.

The trinity of tests

Most tests result from one of the three main principles for constructing a test statistic – Wald, likelihood ratio (LR) and Lagrange multiplier (LM) or score – which are often referred to as the trinity of tests (for example, Engle, 1984). When $p(y: \theta)$ is the joint density of the $T \times 1$ data vector y and we want to test the hypothesis $H_0: \theta = \theta_0$ on the $m \times 1$ vector of parameters θ against the alternative hypothesis $H_1: \theta \neq \theta_0$, these three statistics read:

$$\mathrm{Wald}(\theta_0) = T(\hat{\theta} - \theta_0)' V(\hat{\theta})^{-1} (\hat{\theta} - \theta_0) \mathrm{LR}(\theta_0) = 2\left[L(y, \hat{\theta}) - L(y, \theta_0)\right] \mathrm{LM}(\theta_0) = \frac{1}{T} s(y, \theta_0)' I(\theta_0)^{-1} s(y, \theta_0), \quad (2)$$

with $L(y; \theta)$ the logarithm of the likelihood or joint density $p(y; \theta)$, $L(y, \theta) = \log(p(y, \theta))$; $s(y; \theta)$ is the score, $s(y, \theta) = \frac{\partial}{\partial \theta} L(y, \theta)$; $\hat{\theta}$ is the maximum likelihood estimator under H_1 so $s(y, \hat{\theta}) = 0$, $V(\theta)$ is the covariance matrix of $\hat{\theta}$ and $I(\theta)$ is the information matrix $(V(\theta) = I(\theta)^{-1})$,

$$I(\theta) = -E\left[\frac{1}{T} \frac{\partial^2}{\partial \theta \partial \theta'} L(y, \theta)\right]. \quad (3)$$

The three statistics provide different measures of the relative distance between H_0 and H_1. Under sufficient regularity conditions which ensure the consistency and asymptotic normality of $\hat{\theta}$, all three statistics converge under H_0 to the same $\chi^2(m)$ distributed random variable when the sample size becomes large (for example, Newey and MacFadden, 1994, Th. 9.2). When these regularity conditions hold, usage of a specific statistic is typically a matter of computational ease. The Wald and LM statistics analyse the model only under H_1 or H_0 resp. so one could have a preference for either, given the computational effort to analyse the model under H_0 or H_1. The LR statistic involves the analysis of the model under H_0 and H_1 and is therefore more demanding to compute than the Wald and LM statistics. When one is conducting tests on only one element of θ, the so-called *t*-statistic is often used which equals the square root of the Wald statistic and has a large sample normal distribution under H_0.

Significance, size and power

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

When we test H_0, we specify a significance level of $100 \times (1-\alpha)\%$ which sets the probability that we reject H_0 while it is true equal to α. The critical value associated with this significance level is then such that the probability mass of the large sample distribution of the statistic under H_0 above the critical value equals α. We then reject H_0 with $(1-\alpha) \times 100\%$ significance when the realized value of the statistic exceeds the $(1-\alpha) \times 100\%$ critical value. Another manner to test H_0 with $100 \times (1-\alpha)\%$ significance is by using the *p*-value associated with the realized value of the statistic. The *p*-value equals the probability mass of the large sample distribution of the statistic under H_0 that lies above the realized value of the statistic. Hence, we reject H_0 with $100 \times (1-\alpha)\%$ significance when the *p*-value is less than α.

Tests of H_0: $\theta = \theta_0$ with $100 \times (1-\alpha)\%$ significance for a range of values of θ_0 can be used to obtain the $100 \times (1-\alpha)\%$ confidence set of θ. The $100 \times (1-\alpha)\%$ confidence set of θ contains all values of θ_0 for which H_0: $\theta = \theta_0$ is not rejected with $100 \times (1-\alpha)\%$ significance and therefore contains the true value of θ with probability $1-\alpha$.

Besides computational issues there are several other reasons to prefer a specific statistic, especially when it is unclear whether the regularity conditions, which imply the large sample distribution of the statistic under H_0, hold. Examples of such other reasons are invariance to transformations of the parameters, observed size of the statistics and discriminatory power.

Especially in models that are nonlinear in the parameters, it is appealing to use a statistic whose specification is invariant to nonlinear transformations of the parameters so it does not depend on the specification of the model. This property is violated by the Wald statistic but satisfied by the LR and LM statistics (for example, Dagenais and Dufour, 1991; Dufour, 1997). Hence it is better to use either the LR or LM statistic in such models.

The specification of the significance level of the test intends to control the *Type I error* or probability that we reject H_0 while it holds. The rejection frequency under H_0, to which we refer as the size of the test, should therefore coincide with α. Because we use the large sample distribution of the statistic under H_0 instead of the unknown finite sample distribution to obtain the critical value of the test, this is, however, not the case. The statistic whose size properties dominate those of the others is then typically preferred. The size properties of the different statistics can often be improved by computing the critical values using the bootstrap instead of the large sample distribution (for example, Horowitz, 2001).

The *Type II error* of the statistic is the probability of not rejecting H_0 while it is false. We thus prefer statistics that minimize the *Type II error*, or, put differently, maximize the discriminatory power while preserving an adequate size. When the likelihood function is known, the Neymann–Pearson Lemma implies that the LR statistic is the most powerful statistic for testing a point null hypothesis, like $H_0: \theta = \theta_0$, against a point alternative, like $H_2: \theta = \theta_2$. For composite alternatives, like $H_1: \theta \neq \theta_0$, there is typically no statistic that is the most powerful one in all cases.

Tests in the linear regression model

To construct test statistics for the linear demand eq. (1), we assume that the disturbances are independently and identically distributed so we can estimate the parameters using least squares:

$$\log(P_t) = -\underset{(1.66)}{6.19} - \underset{(0.11)}{0.79}\log(Q_t) + \underset{(0.23)}{0.92}\log(RI_t) + \hat{\varepsilon}_t. \quad (4)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(The standard errors are reported below the parameter estimates.) To compute the Wald, LR and LM statistics that test for a unit demand price elasticity, we further assume the disturbances to be independently identically normally distributed with mean zero. The specifications of the three tests then read

$$\mathrm{Wald}(\theta_0) = \frac{\mathrm{RSSR} - \mathrm{USSR}}{\mathrm{USSR}/T}, \mathrm{LR}(\theta_0) = T\log\left[\frac{\mathrm{RSSR}}{\mathrm{USSR}}\right] \text{and} \mathrm{LM}(\theta_0) = \frac{\mathrm{RSSR} - \mathrm{USSR}}{\mathrm{RSSR}/T}, \tag{5}$$

which test $H_0: \theta = \theta_0$ and where USSR is the unrestricted sum of squared residuals $\hat{\varepsilon}_t$, that is, the residuals under H_1, and RSSR is the restricted sum of squared residuals, that is, the residuals under H_0. Under H_0 and when the disturbances are independently identically distributed with mean zero and a finite fourth order moment all three statistics converge to the same $\chi^2(1)$ distributed random variable when the sample size gets large.
The expression of the LM statistic in (5) is such that we can compute it as well using an auxiliary regression of the restricted residuals under H_0 on all explanatory variables. The expression of LM(θ_0) is then such that it equals T times the R^2 of this regression.
The values of the statistics that test for a unit demand price elasticity that result from (5) read

$$\mathrm{Wald}(-1) = 3.62 > \mathrm{LR}(-1) = 3.50 > \mathrm{LM}(-1) = 3.38. \tag{6}$$

(The value of the Wald statistic in (6) is computed using (5) but could alternatively have been computed using the least squares estimator and standard error that are reported in (4) since $3.62 \approx \left(\frac{-0.79+1}{0.11}\right)^2$.) All three statistics are smaller than the 95 per cent critical value of 3.84 that results from the large sample distribution of the statistics under H_0, which is the $\chi^2(1)$ distribution, so we do not reject the unit demand price elasticity with 95 per cent significance. The Wald statistic that tests for a unit demand price elasticity exceeds the value of the LR statistic which again exceeds the value of the LM statistic. This result always holds for tests on the parameters of the linear regression model and is not a result of the involved data.

Specification tests

Estimation of the demand eq. (1) by least squares as in (4) presumes that the traded quantity is exogenous since least squares leads to an inconsistent estimator when the traded quantity is endogenous. When the traded quantity is endogenous, we need to use an estimator that remains consistent in that case, like, for example, the two-stage least squares (2SLS) estimator or the limited information maximum likelihood (LIML) estimator (see, for example, Theil, 1953; Hood and Koopmans, 1953).
Exogeneity or endogeneity of the traded quantity lead to different specifications of the statistical model for the demand for oranges. A test for the appropriate specification of the model is the Durbin–Wu–Hausman (DWH) statistic, which tests the difference between two estimators, $\hat{\theta}$ and $\ddot{\theta}$, one of which, $\hat{\theta}$, is efficient and consistent in the model under the null hypothesis but not in the model under the alternative hypothesis, while the other estimator, $\ddot{\theta}$, is consistent in both

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
models (see Durbin, 1954; Wu, 1973; Hausman, 1978):

$$\mathrm{DWH}(\theta) = T(\tilde{\theta} - \hat{\theta})'\left[V(\tilde{\theta}) - V(\hat{\theta})\right]^{-}(\tilde{\theta} - \hat{\theta}), \tag{7}$$

with $V(\tilde{\theta})$ and $V(\hat{\theta})$ the covariance matrices of $\tilde{\theta}$ and $\hat{\theta}$ and $[\ldots]^{-}$ the generalized inverse operator. Under sufficient regularity conditions, the DWH statistic converges under H_0 to a $\chi^2(m)$ distributed random variable with m the minimum of the number of elements of θ and the rank of the matrix $V(\tilde{\theta}) - V(\hat{\theta})$.
Using the current and past advertisement variables as instruments, we computed the DWH statistic to test the null hypothesis of exogeneity of the traded quantity against the alternative hypothesis of endogeneity using both the 2SLS and LIML estimators:

$$\begin{aligned}\mathrm{DWH}_{2SLS}(\theta) = {} & T(\tilde{\theta}_{2SLS} - \hat{\theta}_{LS})'\left[V(\tilde{\theta}_{2SLS}) - V(\hat{\theta}_{LS})\right]^{-} \\ & (\tilde{\theta}_{2SLS} - \hat{\theta}_{LS}) = 4.99, \\ \mathrm{DWH}_{LIML}(\theta) = {} & T(\tilde{\theta}_{LIML} - \hat{\theta}_{LS})'\left[V(\tilde{\theta}_{LIML}) - V(\hat{\theta}_{LS})\right]^{-} \\ & (\tilde{\theta}_{LIML} - \hat{\theta}_{LS}) = 4.96\,.\end{aligned} \tag{8}$$

Both statistics exceed the 95 per cent critical value of 3.84 of the large sample distribution of the DWH statistic under H_0, which is a $\chi^2(1)$ distribution. Hence, we reject with 95 per cent significance that the traded quantity is exogenous. This implies that we have to account for the endogeneity of the traded quantity when we test the unit demand price elasticity hypothesis.

Tests in the linear instrumental variables regression model

To accommodate the endogeneity of the traded quantity, we test the demand price elasticity in a linear instrumental variables regression model

$$y_t = x_t\beta + w_t'\gamma + \varepsilon_t x_t = z_t'\pi + w_t'\delta + v_t, \tag{9}$$

where y_t and x_t are the endogenous variables, w_t is a $k_w \times 1$ vector that contains the included exogenous variables, z_t is a $k_s \times 1$ vector that contains the instruments and ε_t and v_t are the disturbances. The instruments z_t are uncorrelated

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

with the structural disturbances ε_t. We assume that the vectors of disturbances $\begin{pmatrix}\varepsilon_t \\ v_t\end{pmatrix}$, $t = 1, \ldots, T$ are independently identically distributed. The variables for the demand for oranges are such that

$y_t = \log(P_t)$, $x_t = \log(Q_t)$, $w_t = \begin{pmatrix}1 \\ \log(RI_t)\end{pmatrix}$ and $z_t = \begin{pmatrix}\log(AC_t) \\ \log(AP_t)\end{pmatrix}$.

The structural parameter β is typically our parameter of interest and we therefore partial out w_t from the model by replacing all remaining variables with the residuals that result from regressing them on w_t:

$$\hat{y}_t = \hat{x}_t\beta + \varepsilon_t\hat{x}_t = \hat{z}_t'\pi + v_t, \tag{10}$$

with $\hat{y}_t$, $\hat{x}_t$ and $\hat{z}_t$ the residuals that result from regressing y_t, x_t and z_t resp. on w_t.

We want to test a hypothesis on the structural parameter β, H_0: $\beta = \beta_0$, like, for example, that of a unit demand price elasticity. We discuss some statistics that can be used for this purpose most of which belong to the trinity of tests.

Wald statistics

Using either the 2SLS or LIML estimator, we can test H_0 using a Wald statistic:

$$\mathrm{Wald}_{2SLS}(\beta_0) = T(\hat{\beta}_{2SLS} - \beta_0)'V(\hat{\beta}_{2SLS})^{-1}(\hat{\beta}_{2SLS} - \beta_0)\mathrm{Wald}_{LIML}(\beta_0) = T(\hat{\beta}_{LIML} - \beta_0)'V(\hat{\beta}_{LIML})^{-1}(\hat{\beta}_{LIML} - \beta_0), \tag{11}$$

with $\hat{\beta}_{2SLS}$ the 2SLS estimator: $\hat{\beta}_{2SLS} = (\hat{\pi}'\Sigma_{t=1}^T\hat{z}_t\hat{x}_t)^{-1}\hat{\pi}'\Sigma_{t=1}^T\hat{z}_t\hat{y}_t$, $\hat{\pi} = (\Sigma_{t=1}^T\hat{z}_t\hat{z}_t')^{-1}\Sigma_{t=1}^T\hat{z}_t\hat{x}_t$, and $\hat{\beta}_{LIML}$ the LIML estimator:

$$\hat{\beta}_{LIML} = \arg\min_\beta \frac{\left[\Sigma_{t=1}^T\hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]'\left[\Sigma_{t=1}^T\hat{z}_t\hat{z}_t'\right]^{-1}\left[\Sigma_{t=1}^T\hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]}{\Sigma_{t=1}^T(\hat{y}_t - \hat{x}_t\beta)^2 - \left[\Sigma_{t=1}^T\hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]'\left[\Sigma_{t=1}^T\hat{z}_t\hat{z}_t'\right]^{-1}\left[\Sigma_{t=1}^T\hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]}. \tag{12}$$

Both Wald statistics converge under H_0 and a number of regularity conditions which rule out zero values of π to a $\chi^2(1)$ distributed random variable when the sample size gets large (for example, Newey and MacFadden, 1994). The assumption of a non-zero value of π for the large sample distribution implies that finite sample distributions of both Wald statistics depend on the value of π. The actual size of the Wald statistics can therefore deviate considerably from the assumed *Type 1 error* which makes these statistics unreliable for usage in practice (for example, Nelson and Startz,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1990; Bound, Jaeger and Baker, 1995; Dufour, 1997; Staiger and Stock, 1997). The bootstrap cannot be used to overcome these size distortions (for example, Horowitz, 2001).

Anderson – Rubin statistic

Anderson and Rubin (1949) construct a test for H_0 by substituting the equation of $\hat{x}_t$ into the equation of $\hat{y}_t$:

$$\hat{y}_t - \hat{x}_t\beta_0 = \hat{z}_t'\phi + u_t, \tag{13}$$

with $\phi = \pi(\beta - \beta_0)$ and $u_t = \varepsilon_t + v_t(\beta - \beta_0)$. Under H_0, φ equals zero and a test for H_0 can therefore be conducted using a test for a zero value of φ. Anderson and Rubin (1949) proposed to use the *F*-statistic that tests for a zero value of φ in (13) for this purpose. This *F*-statistic is commonly referred to as the Anderson–Rubin (AR) statistic.
When the disturbances are independently and identically distributed with finite fourth-order moments, the AR statistic converges under H_0 to a $\chi^2(k_z)/k_z$ distributed random variable when the sample size gets large. This large sample distribution of the AR statistic does not depend on the value of π, which makes the AR statistic a more reliable statistic for practical purposes than the Wald statistics in (11). A disadvantage of the AR statistic is that its large sample distribution is proportional to a χ^2 distribution with a degrees of freedom parameter that equals the number of instruments while we conduct a test on only one parameter. This reduces the discriminatory power of the AR statistic when the number of instruments is large, which is often the case.

LM statistic

When we assume the vector of disturbances $\begin{pmatrix}\varepsilon_t \\ v_t\end{pmatrix}$ to be independently identically normally distributed with mean zero, we can construct the likelihood function and therefore also the LM statistic for testing H_0 (see Kleibergen, 2002):

$$\mathrm{LM}(\beta_0) = \frac{1}{\tilde{\sigma}_{\varepsilon\varepsilon}}\left(\sum_{t=1}^{T}\hat{z}_t\tilde{\varepsilon}_t\right)'\tilde{\pi}(\beta_0)\left[\tilde{\pi}(\beta_0)'\left(\sum_{t=1}^{T}\hat{z}_t\hat{z}_t'\right)\tilde{\pi}(\beta_0)\right]^{-1}\left(\sum_{t=1}^{T}\hat{z}_t\tilde{\varepsilon}_t\right), \tag{14}$$

with $\tilde{\varepsilon}_t = \hat{y}_t - \hat{x}_t\beta_0$,

$$\tilde{\pi}(\beta_0) = \left(\sum_{t=1}^{T}\hat{z}_t\hat{z}_t'\right)^{-1}\sum_{t=1}^{T}\hat{z}_t\left[\hat{x}_t - \tilde{\varepsilon}_t\frac{\tilde{\sigma}_{v\varepsilon}}{\tilde{\sigma}_{\varepsilon\varepsilon}}\right], \tilde{\sigma}_{\varepsilon\varepsilon} = \frac{1}{T-k_z}\sum_{t=1}^{T}(\hat{y}_t - \hat{x}_t\beta_0)^2, \tilde{\sigma}_{v\varepsilon} = \frac{1}{T-k_z}\sum_{t=1}^{T}\hat{x}_t(\hat{y}_t - \hat{x}_t\beta_0) \tag{15}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and where $\tilde{x}_t$ and $\tilde{y}_t$ are the residuals that result from regressing x_t and y_t on w_t and z_t. When the disturbances $\begin{pmatrix}\varepsilon_t \\ v_t\end{pmatrix}$ are independently identically distributed and have finite fourth order moments, the LM statistic converges to a $\chi^2(1)$ distributed random variable when the sample size increases. The convergence of the LM statistic does not depend on the value of π. The large sample distribution of the LM statistic under H_0 is therefore typically a rather accurate approximation of the finite sample distribution, and this approximation can even be further improved by using the bootstrap (see Kleibergen, 2004).

LR statistic

Under identically independently normal distributed disturbances, Moreira (2003) constructs the likelihood ratio statistic to test H_0:

$$\mathrm{LR}(\beta_0) = \frac{1}{2}\left[\mathrm{AR}(\beta_0) - \mathrm{r}(\beta_0) + \sqrt{(\mathrm{AR}(\beta_0) + \mathrm{r}(\beta_0))^2 - 4\mathrm{r}(\beta_0)(\mathrm{AR}(\beta_0) - \mathrm{LM}(\beta_0))}\right], \tag{16}$$

with AR(β_0) k_w times the AR statistic that tests H_0:

$$\mathrm{AR}(\beta_0) = \frac{1}{\tilde{\sigma}_{\varepsilon\varepsilon}}\left(\sum_{t=1}^{T}\tilde{z}_t\tilde{\varepsilon}_t\right)'\left(\sum_{t=1}^{T}\tilde{z}_t\tilde{z}_t'\right)^{-1}\left(\sum_{t=1}^{T}\tilde{z}_t\tilde{\varepsilon}_t\right), \tag{17}$$

and r(β_0) a statistic that tests for a zero value of π under H_0, so by using $\tilde{\pi}(\beta_0)$,

$$\mathrm{r}(\beta_0) = \frac{1}{\tilde{\sigma}_{vv.\varepsilon}}\tilde{\pi}(\beta_0)'\left(\sum_{t=1}^{T}\tilde{z}_t\tilde{z}_t'\right)\tilde{\pi}(\beta_0), \tag{18}$$

where $\tilde{\sigma}_{vv.\varepsilon} = \tilde{\sigma}_{vv} - \frac{\tilde{\sigma}_{v\varepsilon}^2}{\tilde{\sigma}_{.\varepsilon\varepsilon}}$ and $\tilde{\sigma}_{vv} = \frac{1}{T-k_z}\sum_{t=1}^{T}\tilde{x}_t^2$. Moreira (2003) shows that, when the disturbances are independently identically distributed with finite fourth-order moments, the large sample distribution of LR(β_0) under H_0 is conditional on the value of r(β_0). We thus need to use a different critical value to determine the significance of a realized value of the LR statistic for every value of r(β_0). When r(β_0) is zero, the large sample distribution of LR (β_0) is identical to a $\chi^2(k_w)$ distribution while it equals the $\chi^2(1)$ distribution for large values of r(β_0). Besides the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dependence on $r(\beta_0)$, the large sample distribution of $LR(\beta_0)$ does not depend on π, which makes $LR(\beta_0)$ a trustworthy statistic for practical purposes. Andrews, Moreira and Stock (2006) show that the LR statistic is the most powerful of those statistics whose large sample distributions do not depend on π and are invariant under transformations of the model.

The unit demand price elasticity

We test for a unit demand price elasticity using each of the above statistics.

$$\begin{gathered} \mathrm{Wald}_{2SLS}(-1) = 191 \quad \mathrm{Wald}_{LIML}(-1) = 178 \\ AR(-1) = 73.7 \quad LM(-1) = 67.3 \quad LR(-1) = 69.1 . \end{gathered} \tag{19}$$

The value of $r(\beta_0)$ is 174, which makes the large sample distribution of LR(−1) given $r(\beta_0)$ identical to the large sample distribution of LM(−1), which is a $\chi^2(1)$. The large sample distribution of AR(−1) is a $\chi^2(2)$ distribution, and the large sample distributions of the Wald statistics are $\chi^2(1)$ distributions while a non-zero value of π is assumed. All statistics reject the hypothesis of a unit demand price elasticity with 95 per cent significance. This shows the importance of accounting for the endogeneity of the traded quantity since this hypothesis was not rejected in the linear regression model that assumes the traded quantity to be exogenous. The values of the statistics whose large sample distributions are not influenced by the value of π, that is, AR(−1), LM(−1) and LR(−1), are all of the same order of magnitude, while the Wald statistics are much larger. This indicates the different behaviour of these statistics and we recommend that these Wald statistics not be used.

More general specifications

We discussed the trinity of tests in a linear model estimated using either least squares or instrumental variables. The Wald, LM and LR statistics extend to a large variety of models which are possibly nonlinear in the parameters and have unknown likelihood functions. The expression of the Wald statistic is such that it can be applied to any estimator which has a normal large sample distribution and for which a consistent estimator of the asymptotic variance exists. The LM test is applicable in any model where the estimators solve a first-order condition. The LR test is based on the difference of an objective function under H_0 and H_1, a specification that allows it to accommodate more general statistical models (for example, Engle, 1984; Newey and MacFadden, 1994). In these general settings, such as, for example, the generalized method of moments, the large sample distribution of the Wald statistic is often affected by nuisance parameters while the large sample distributions of the LM and LR statistics remain robust to the value of these nuisance parameters (for example, Kleibergen, 2005). Hence, the LM and LR statistics often provide more reliable tests.

See Also

- bootstrap
- central limit theorems
- endogeneity and exogeneity
- instrumental variables

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- simultaneous equations models
- statistical inference
- statistics and economics

Bibliography

Anderson, T. and Rubin, H. 1949. Estimators of the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics* 21, 570–82.

Andrews, D., Moreira, M. and Stock, J. 2006. Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* 74, 715–52.

Berndt, E. 1991. *The Practice of Econometrics, Classic and Contemporary*. Reading, MA: Addison-Wesley.

Bound, J., Jaeger, D. and Baker, R. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–50.

Dagenais, M. and Dufour, J.-M. 1991. Invariance, nonlinear models, and asymptotic tests. *Econometrica* 59, 1601–15.

Dufour, J.-M. 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65, 1365–388.

Durbin, J. 1954. Error in variables. *Review of the International Statistical Institute* 22, 23–32.

Engle, R. 1984. Wald likelihood ratio and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Hausman, J. 1978. Specification tests in econometrics. *Econometrica* 46, 1251–72.

Hood, W. and Koopmans, T. 1953. *Studies in Econometric Method*, vol. 14 of *Cowles Foundation Monograph*. New York: Wiley.

Horowitz, J. 2001. The bootstrap. In *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Kleibergen, F. 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70, 1781–803.

Kleibergen, F. 2004. Expansions of GMM statistics that indicate their properties under weak and/or many instruments and the bootstrap. Working paper, Brown University.

Kleibergen, F. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73, 1103–24.

Moreira, M. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71, 1027–48.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Nelson, C. and Startz, R. 1990. Some further results on the exact small sample properties of the instrumental variables estimator. *Econometrica* 58, 967–76.

Nerlove, M. and Waugh, F. 1961. Advertising without supply control: some implications of a study of the advertising of oranges. *Journal of Farm Economics* 43, 813–37.

Newey, W. and McFadden, D. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

Staiger, D. and Stock, J. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.

Theil, H. 1953. Estimation and simultaneous correlation in complete equation systems. Mimeo. The Hague: Central Planning Bureau.

Wu, D.-M. 1973. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41, 733–50.

How to cite this article

Kleibergen, Frank. "testing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000203> doi:10.1057/9780230226203.1696

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

the Baring crisis of 1890

Marc Weidenmier
From The New Palgrave Dictionary of Economics Online, , 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The Baring crisis of 1890 is one of the world's most famous financial crises over the last 200 years. The crisis is well known because the Bank of England put together a rescue fund to save the House of Baring. The investment banking firm was in financial trouble because it was the primary debt issuer for Argentina which was experiencing an economic and financial downturn.The Baring crisis is regarded as an early example of a central bank playing the role of a lender of last resort.

Keywords

Argentina; banking; Baring; banking crises; Latin America; lender of last resort

Article

The origins of the Baring crisis of 1890, the most famous financial crisis of the 19th century, can be traced to the world debt crisis of 1873 and ensuing recession, which had large economic effects on Argentina and Latin America. The region did not recover from the downturn until the early 1880s following a resurgence of foreign trade and capital flows from Europe. For Argentina, one of the chief obstacles to economic growth and development was the absence of a strong central government. Historically, the national authority shared power with the provincial governments, and it also faced an internal threat from indigenous people who lived on the pampas. The central government consolidated its power and expanded its borders by driving the indigenous people off the pampas in a series of wars during the late 1870s. With the election of Julio Roca as the country's president, the Indian War hero was able to broker an agreement with the ruling elites of the provinces, which centralized the power of the national government.

One of the primary goals of Roca's government was to employ foreign capital to construct railroads, public works, and to modernize Buenos Ayres (Marichal, 1989). The new leader's first major loan was a railway issue which completed two major trunk lines in the South American country. The construction of a transportation network throughout the country helped consolidate Roca's power as well as stimulate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic activity by opening up the market for commercial agriculture. Roca also transformed Buenos Aires into the ‘Paris of South America’ by constructing broad avenues, spacious parks, a well-functioning water supply and drainage system, and a modern port. The national and local government carried out a series of state-run infrastructure projects in Latin America (Marichal, 1989; Mitchener and Weidenmier, 2008).

Although the economic policies of the Roca administration stimulated short-run economic activity in Argentina, they posed serious dangers in the long run. The country's expanding debt could only be serviced if the country had sufficient tax revenues. Unfortunately, it would take years before the government would realize significant revenues from commercial activity stimulated by the infrastructure investments (Ford, 1956).

Roca finished his term as Argentina's president in 1886. His brother-in-law Miguel Celman became the country's leader following a fraudulent election. Rather than continuing the policies of his predecessor, Celman reduced the government's role in the administration of the railways. The newly elected president sold the Central Norte and Andino railways, two of the country's most important, to British capitalists. The funds from the sale were supposed to be used to reduce the country's rising debt level. Instead of restoring fiscal discipline, the country began issuing additional debt through state banks even though it stopped borrowing funds to finance railway projects.

From 1886 to 1890, Argentina passed a series of ‘banking reforms’ that fuelled the expansion of credit and paper money issues (Williams, 1920). National and provincial banking authorities ratified a Free Banking Law in 1887 which authorized any banking association to issue notes provided it purchased gold bonds to the full amount of the notes issued. There were several problems with the law. It permitted banks meeting minimum capital requirements to issue paper notes backed by government gold bonds. The bank notes, however, were not redeemable in gold, and since the bonds were new issues, they constituted a new liability on the government's balance sheet. The banks that participated in the note issuance scheme floated loans in Europe to finance the purchase of the domestic gold bonds. This scheme worked as long as foreign investors agreed to purchase the Argentine bonds and as long as additional note issuances were backed 100 per cent by specie. By 1890, Argentine provincial banks had issued more than 30 million pounds of external debt.

Argentina's loose monetary and fiscal policies led to a decline in the country's financial and macroeconomic conditions from the mid-1880s until the outbreak of the Baring crisis in 1890. High-powered money grew at an annual average rate of 18 per cent, inflation averaged 17 per cent, and the paper peso depreciated at an average rate of 19 per cent between 1884 and 1890 (della Paolera and Taylor, 2001). By 1890, nearly 40 per cent of the foreign borrowing was going towards debt service, and 60 per cent of imports were going toward consumption goods.

Weakening economic conditions in Argentina reduced the demand for Argentine securities on the London market. Domestic investors began to dump the country's paper peso. Although the government used specie to defend the exchange rate, the stock of gold at the Banco Nacional had declined to such an extent by December 1889 that the financial institution could no longer carry out this currency operation. Strikes, demonstrations and a failed coup by military leaders ensued in 1889 and 1890. Inflation reduced the real wages of Argentine workers. The country's lax monetary and fiscal policies drained the banking system of specie, provoked a series of banking crises, and ultimately the Baring crisis in 1890. Argentine real GDP declined by more than ten per cent between 1890- and 1891. Last-minute attempts to reform

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the country's poor economic policies failed, and the country entered into a decade long recession.
The Argentine crisis potentially had serious implications for global financial markets, especially London. Baring Brothers was the primary investment bank for the South American country. The firm purchased and issued debt for Argentina. The investment bank was heavily involved with the Buenos Aires Water Supply and Drainage Loan, a new debt issue the underwriter failed to sell on the London market (Eichengreen, 1999. The House of Baring secretly notified the Bank of England that it could not service its debt obligations in November 1890. The central bank then pooled financial resources to prevent the beleaguered investment banking firm from causing a larger meltdown on the London market. The Bank of England secured loans from the Bank of France, Russia's central bank, and British financial institutions to help Baring Brothers service its debt obligations and prevent a larger meltdown on the British market. The rescue operation succeeded and prevented a general financial collapse on European markets. Some scholars have argued that the Baring crisis provides one of the earliest examples of a central bank playing the role of a lender-of-last resort in financial markets (Mitchener and Weidenmier, 2008).
Although the Bank of England prevented a financial collapse in Europe, the central bank did little to assist Argentina and Latin America. Argentina experienced a deep recession for several years and did not fully recover from the crisis until the turn of the century. In the absence of macroeconomic data such as GDP, Mitchener and Weidenmier (2008) use interest rates to examine the effects of the Baring crisis on emerging markets. They find that interest rates in increased more than 1600 basis points in Latin America while interest rates in other emerging markets were flat. This suggests that the crisis had severe negative macroeconomic effects in Latin America. The evidence suggests that the Baring crisis was largely a regional financial crisis which had few economic effects outside Latin America.

See Also

- banking crises

Bibliography

Della Paolera, G. and Taylor, A. 2001. *Straining at the Anchor: The Argentine Currency Board and the Search for Macroeconomic Stability, 1880–1935*. Chicago: University of Chicago Press.

Eichengreen, B. 1999. The Baring crisis in a Mexican mirror. *International Political Science Review* 20, 249–70.

Ford, A. 1956. Argentina and the Baring crisis of 1890. *Oxford Economic Papers* 8, 127–50.

Marichal, C. 1989. *A Century of Debt Crises in Latin America*. Princeton, N.J.: Princeton University Press.

Mitchener, K. and Weidenmier, M. 2008. Baring crisis and the Great Latin American meltdown of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1890s. *Journal of Economic History* 68, 462–500.

Williams, J. 1920. *Argentine International Trade under Inconvertible Paper Currency, 1880–1900*. Cambridge, Mass.: Harvard University Press.

How to cite this article

Weidenmier, Marc. "the Baring crisis of 1890." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_B000344>
doi:10.1057/9780230226203.1892

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

the economy of ancient Greece

Paul Cartledge
From The New Palgrave Dictionary of Economics Online, 2008,
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

There were many ‘economies’ rather than a single ‘economy’ in ancient Greece (a culturally interlinked world, c. 800–300 BCE, stretching across the Mediterranean basin and around the Black Sea). Except in Athens, agriculture (cereals, olives, grapevines, and the raising of small-stock animals – sheep, goats, pigs) predominated over trade and industry as an economic driver. The Greeks did not invent coinage but spread it and embedded it, and although they were thoroughly familiar with the idea of markets and market prices, they did not develop a market economy.

Keywords

ancient economy; oikos; ‘primitivists’, ‘modernists’; Finley, Moses; Athens; Sparta; ‘proxy data’; Mediterranean triad; agriculture; grain; olive oil; wine; trade, local, regional and inter-regional; manufacture; technology; slavery; money, coined and non-coin; markets

Article

Definitions

‘Ancient Greece’ for the purposes of this entry will be taken to refer to the period from roughly the eighth century BCE to the end of the fourth century BCE: that is, from the rediscovery of literacy by means of the invention of the Greek alphabet, the renewal of intensive trade contacts with the near East, and the beginnings of large-scale permanent overseas emigration and establishment of Mediterranean-wide trade networks, down to the start of the new post-Alexander the Great (d. 323 BCE) ‘Hellenistic’ (mixed Greek-oriental) world.
‘Economy’ is more difficult to specify, or pin down. The word is of Greek derivation but in ancient Greek it meant primarily and literally the management of an individual household (*oikos*) not the management of a ‘city’ or ‘national’ economy. This led one school of modern interpreters (late 19th-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
century German), the so-called ‘primitivists’, to speak of the entire period under consideration here as one of ‘household’ not ‘national’ economy. That is a considerable and highly misleading exaggeration, but it does draw attention to the fact that an ancient Greek city did not have an economy, or practise economics, in anything like a post-Adam Smith, let alone post-Alfred Marshall, sense. Their opponents, the ‘modernists’ or ‘modernizers’, claimed no less excessively that ancient Greece was, economically speaking, pretty much similar to the modern ‘developed’ world (and similar too in its stages of development), except that it operated on an infinitely smaller scale and without the benefits – such as economies of scale – made possible only by the scientific and technological revolutions of early modernity. A sensible compromise was firmly advocated by Moses Finley (1973), who rightly drew attention especially to Greek ideology and terminology; but because he sometimes underestimated the quantity and sophistication of ancient Greek economic activity, he too was (mis)labelled a ‘primitivist’. (The debate is usefully summarized in Scheidel and von Reden, 2002; see esp. Andreau 2002, Cartledge, 2002, Meikle, 2002; also Manning and Morris, 2005.)

A second reason for being chary of the term ‘economy’ is that, after the wave of emigration noted above, there were at any one time between 600 and 300 BCE some 1,000 separate Greek political entities, radically self-differentiated politically but also often very different indeed economically speaking. This is why I have in the past written of ‘the economy (economies) of ancient Greece’ (Cartledge, 2002; cf. Davies, 1998). At one extreme, classical fifth- and fourth-century Athens was as ‘developed’ as any Greek city before Hellenistic Alexandria (founded 332). In the international port of Peiraieus it even had a ‘commercial centre’ separated physically as well as spiritually from the political centre (Garland, 2001), and its total population (civic centre plus surrounding territory) was far larger (c. 250–300,000) and far more diverse ethnically and occupationally than any other Greek city's. (In size of home territory, however, its c. 2,500•sq•km were exceeded by Sparta, c. 8,400, Syracuse, c. 4,000, and Panticapaeum in the Crimea, c. 3,000: Hansen and Nielsen, 2004, pp. 70–3). At the opposite extreme were fundamentally rural settlements, in Arcadia for example, cut off from the sea and long-distance trade, surviving, modestly, on ‘natural’, pastoral as well as agricultural, economy. In between, the modal Greek city had a population of some 2,000–8,000, occupied some 100•sq•km, and practised versions of ‘mixed economy’, in which agriculture and stock-raising always predominated over trade and manufacturing industry.

A third problem with doing ancient Greek economic history is that the contemporary ancient data accessible today are resolutely unstatistical. This is partly because the ancient Greeks did not think and so did not audit themselves statistically but also because the nature of their politically overdetermined economies did not generally encourage or lend itself to statistical computation. The figures we get in our sources – for instance, for the impossibly inflated aggregate numbers of slaves in a particular city at any one time – tend therefore to be at best extreme outliers, at worst rhetorical inventions, rarely something reliably identifiable and usable in-between. Hence the regular resort of scholars to ‘proxy data’ – modern data of climate, crop-yields, etc. – making assumptions of continuity and stability as between ancient and modern conditions that are often untestable, but still useful as models or thought-experiments and for setting workable parameters (Manning and Morris, 2005; on this and on all other matters discussed in this entry, see now Scheidel, Morris and Saller, 2007).

The Mediterranean triad: agriculture and trade

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The Mediterranean triad of dietary staples – grain, olive oil and wine – was established as such in the Greek sphere during the third millennium BCE (Renfrew, 1972). Not much in the way of improvements in seed-selection, or efficiency in growing or harvesting techniques, is detectable during our period, owing to the likely constraints of Greek soils and microclimates, and the certain constraints of technological backwardness (not even the wheelbarrow was known, apparently).

One huge exception was the massively profitable exploitation after 600 BCE of the black-earth soils of the Ukraine and Crimea (see Panticapaeum, above) for the achieving of – by old Greek standards – huge yields of bread wheat, more nutritious as well as more easily processed than the default Greek grain-crop, barley (up to five times more drought-resistant than wheat on average) (Sallares, 1991). Since the northern shore of the Black Sea cannot grow olives (because of winter frost), there was a considerable uplift in the production of olive oil further south (especially around Athens) for export to these deprived colonial Greeks, in exchange for which came, besides the bread wheat (especially again to Athens: Moreno, 2007), dried fish and slaves.

Olives and their by-products were culturally as well as economically vital, and universally employed – even if not universally manufactured – as unguent, medicament, lubricant, and source of energy as well as food (Foxhall, 2007). This was a standing incentive to extensification. In one small area near Athens, for example, terracing of marginal land is estimated to have extended the cultivated area by some 40 per cent. Regions and cities that experienced sharp population growth, such as Athens in the fifth century, also resorted to intensification, either by reducing the regularity or duration of fallowing, or by intercultivation of grain with olives, or by a combination of the two.

The grapevine flourishes in some soils and aspects more than others. The wine produced around Athens, for example, in sharp contrast to its local olive oil, was thought far less desirable than that produced on the northern Aegean island of Thasos, where legislation was introduced in the fifth century to control the highly lucrative export trade. Other islands that specialized in wine-production for export were Lesbos, Chios and Samos, and each production region generated its own distinctive shape of two-handled, pointed-base, pottery transport vessels (known as *amphorae*), further distinguished by the liberal application of amphora-stamps – a sort of ancient Greek equivalent of *appellation contrôlée*.

Long-distance, interregional trade in wine, grain and other commodities (especially metals, such as copper from Cyprus or iron from Elba) was sharply marked off institutionally and terminologically from smaller-scale, local wholesale and retail trading (Garnsey, Hopkins and Whittaker, 1983). Long-distance traders were *emporoi*, literally ‘passengers’ (on ships), and they traded typically in purpose-built, ‘round’, sail-driven merchant vessels to economize on crew and time. But such trade was in Mediterranean weather conditions always risky, not to mention the threat from oar-driven pirate pinnaces and galleys; and from the later fifth century the larger operators were encouraged to take out bottomry or maritime loans – high-risk, high-interest – as a form of insurance, in deals struck with a new breed of commercially minded bankers (Cohen, 1994). The owners of the banks as of the ships would be free, but the bankers and traders might just as likely be slaves as free citizens, and not rarely of non-Greek origin (Reed, 2003).

Labour and manufacture

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The ancient Greek world was almost entirely one of human labour, not labour-saving technological devices, and, insofar as production was undertaken beyond the scale of household subsistence (Mattingly and Salmon, 2001), it was a world of manufactories rather than factories. Wind power was of course exploited in navigation, but watermills were a thing of the pretty distant future so far as ancient Greece was concerned. Lifting devices and other forms of ‘engineering’ were most assiduously developed for religious not secular purposes, such as the construction of a temple (Landels, 1978). On the other hand, traditional craft skills in carpentry, metallurgy, ceramics and the weaving of cloth had operated at a high level since the eighth century, even though typically on an individual household or small workshop basis (Burford, 1972). One large exception were the gangs of workers employed in the silver-bearing lead mines belonging to the city of Athens, where possibly as many as 20,000 or even 30,000 may have been employed at any one time in digging, extracting and washing the ore (Lauffer, 1979). But these were not free citizens: they were slaves, performing a task classified as servile, fit only for less-than-human beings.

Slavery

Unfreedom is of hoary antiquity, cross-culturally, but it took the ingenuity of the Greeks of the sixth century BCE to transform various kinds of personal dependency into full-blown ownership of the ‘chattel’ slave variety (the kind practised in the American Old South, the Caribbean and Brazil from the 17th to 19th centuries) (Dal Largo and Katsari, 2007). At Athens, for example, all slaves were of the chattel variety, bought on the markets to which they had been brought by traders dealing with the countries of the Black Sea and western Anatolian regions (Scythians and Paphlagonians, for example). But in Sparta, although the same word (*douloi*) was used to describe them, the servile workforce of Helots (‘captives’) was created by enslaving local Greeks and by fair means or foul keeping them locked within a system of hereditary bondage for some four centuries or more (Luraghi and Alcock, 2003). Almost all the 50–100,000 Helots, like the majority of the 100,000 or so slaves at Athens, were somehow engaged in agriculture. Just how economically efficient that system of helotage was cannot be easily determined, but it delivered the goods in the sense that it was the basis of Sparta's status as a great Greek power for most of those 400 years, and the basis too of Sparta's extraordinary warrior-communalist lifestyle.

Money and coinage

Money has a number of functions and uses, some of which, but not all, can be handily fulfilled by coins (Howgego, 1995). Greeks did not invent the idea of stamping a fixed weight of precious metal (gold, silver, electrum) with a badge and slogan or some other authenticating device, but they did develop this Lydian invention phenomenally, for political as well as economic reasons, and did transmit it to otherwise alien neighbouring cultures such as that of the Persians. The first coins struck, in the first quarter of the sixth century, were of electrum (a gold-silver mix) or silver, and of relatively large denominations, not usable therefore as small change. But by the end of that century sometimes really small fractions were in quite general use, at least in the Aegean, and by the end of the fifth century a move had been made towards a fiduciary coinage of bronze. The Spartans, idiosyncratic as ever, refused

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to strike silver coins (until their third-century BCE ‘normalization’), but did operate some sort of ‘currency’ of iron (in the form perhaps of cooking spits). Such monetized spits were used elsewhere than in Sparta too, and offered as dedications to the gods and goddesses in sanctuaries, a nice reminder that the sacred and the profane were close partners in ancient Greece.

Markets

Finally, markets – and the issue of whether, and if so how far, any Greek city developed anything like a market economy: that is, not an economy with markets but an economy centrally defined by price-fixing markets. A famous passage of Aristotle's *Nicomachean Ethics*, written in the 330•s, has been read as making the intellectual breakthrough in embryo to a labour theory of value and/or a market theory of price, but it can just as well be read as an exercise in moral philosophy using economic illustrations. Elsewhere, in the *Politics*, Aristotle makes his preference for a ‘free’ Agora unambiguously plain: by ‘free’ he means one where sordid economic transactions were kept to a minimum, or at bay altogether – for example, by barring from the holding of political office any citizen who had traded in a commercial Agora within the past decade, as at Thebes (Austin and Vidal-Naquet, 1977).

There is also objective evidence for a certain conventionality of non-market price-fixing – some commodities turn up in widely different contexts and periods valued at a suspiciously identical exchange price. Likewise, the cost of labour purchased on the market, for instance for large-scale civic construction projects such as temples, seems inelastic to a degree that would be considered irrational in an (economically speaking) free labour-market situation.

Nevertheless there are hints and signs from earlier in the fourth century of an increasing and increasingly generalized marketization of commodity exchange, a process that ‘took off’exponentially under the conditions of the new world opened up by the middle Eastern conquests of Alexander the Great (334–323). But ‘Hellenistic’ economic globalization (as it were) is another topic, for another essay (see Cartledge, 1997 for an overall outline sketch; in detail, Archibald, Davies and Gabrielsen, 2005; Archibald, Davies, Gabrielsen and Oliver, 2001).

Bibliography

Andreau, J. 2002. Twenty years after Moses I. Finley's *The Ancient Economy*. In Scheidel and von Reden, 2002.

Archibald, Z., Davies, J.K. and Gabrielsen, V. (eds.) 2005. *Making, Moving and Managing: The New World of Ancient Economies*, 323-31 B.C. Oxford: Oxford University Press.

Archibald, Z., Davies, J.K., Gabrielsen, V. and Oliver, G.J. (eds.) 2001. *Hellenistic Economies*. London and New York: Routledge.

Austin, M. and Vidal-Naquet, P. 1977. *Economic and Social History of Ancient Greece: An Introduction*. London: Batsford.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Burford, A. 1972. *Craftsmen in Greek and Roman Society*. London and New York: Thames & Hudson.

Cartledge, P.A. 1997. Introduction. In *Hellenistic Constructs: Essays in Culture, History and Historiography*, ed. P. Cartledge, P. Garnsey and E. Gruen. California: University of California Press.

Cartledge, P.A. 2002. The economy (economies) of ancient Greece. In Scheidel and von Reden, 2002.

Cartledge, P.A., Cohen, E.E. and Foxhall, L. (eds.) 2002. *Money, Labour and Land. Approaches to the Economies of Ancient Greece*. London and New York: Routledge.

Cohen, E.E. 1994. *The Athenian Economy: A Banking Perspective*. Princeton: Princeton University Press.

Dal Largo, E. and Katsari, C. (eds.) 2007. *Slave Systems, Ancient and Modern*. Cambridge: Cambridge University Press.

Davies, J.K. 1998. Ancient economies: muddles and models. In Parkins and Smith, 1998.

Finley, M.I. 1973. *The Ancient Economy*. California: University of California Press. [Latest edition by I. Morris, 1999.]

Foxhall, L. 2007. *Olive Cultivation in Ancient Greece: Seeking the Ancient Economy*. Oxford: Oxford University Press.

Garland, R. 2001. *The Piraeus*, 2nd edn. Bristol: Bristol Classical Press.

Garnsey, P., Hopkins, K. and Whittaker, C.R. (eds.) 1983. *Trade in the Ancient Economy*. London: Chatto & Windus.

Hansen, M.H. and Nielsen, T.H. 2004. *An Inventory of Archaic and Classical Greek Poleis*. Oxford: Oxford University Press.

Howgego, C. 1995. *Ancient History from Coins*. London and New York: Routledge.

Landels, J.G. 1978. *Engineering in the Ancient World*. London: Chatto & Windus.

Lauffer, S. 1979. *Die Bergwerkssklaven von Laureion*, 2nd edn. Stuttgart: F. Steiner.

Luraghi, N. and Alcock, S.E. (eds.) 2003. *Helots and their Masters in Laconia and Messenia: Histories, Ideologies, Structures*. Washington, DC: Center for Hellenic Studies.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Manning, J.G. and Morris, I. (eds.) 2005. *The Ancient Economy: Evidence and Models*. Stanford: Stanford University Press.

Mattingly, D.J. and Salmon, J.B. (eds.) 2001. *Economies beyond Agriculture in the Classical World*. London and New York: Routledge.

Meikle, S. 2002. Modernism, economics and ancient history. In Scheidel and von Reden 2002.

Moreno, A. 2007. *Feeding the Democracy: The Athenian Grain Supply in the Fifth and Fourth Centuries BC*. Oxford: Oxford University Press.

Parkins, H. and Smith, C.J. (eds.) 1998. *Trade, Traders, and the Ancient City*. London and New York: Routledge.

Reed, C.M. 2003. *Maritime Traders in the Ancient Greek World*. Cambridge: Cambridge University Press.

Renfrew, C. 1972. *The Emergence of Civilization: The Aegean and the Cyclades in the Third Millennium B.C.* London: Methuen.

Sallares, R. 1991. *The Ecology of the Ancient Greek World*. London: Duckworth.

Scheidel, W., Morris, I. and Saller, R. (eds.) 2007. *The Cambridge Economic History of the Greco-Roman World*. Cambridge: Cambridge University Press.

Scheidel, W. and von Reden, S. (eds.) 2002. *The Ancient Economy*. Edinburgh: Edinburgh University Press.

How to cite this article

Cartledge, Paul. "the economy of ancient Greece." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_E000317> doi:10.1057/9780230226203.1882

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Theil, Henri (1924 – 2000)

Teun Kloek
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Henri Theil was a highly prolific writer of books and articles in the fields of econometric methods and applied econometrics. His best-known methodological contributions include the method of two-stage least squares, inequality coefficients (for evaluation of forecasting errors), and two measures for income inequality. Most of his other methodological contributions can be found in his *Principles of Econometrics* (1971). His best-known applied work is on economic forecasting, income inequality, applied demand analysis, and consumption patterns.

Keywords

consumer demand systems; Cowles Commission; distribution-free statistics; econometric equation systems; estimation; Florida model; forecasting; Gricliches, Z.; income inequality; inequality coefficient; information theory; Kendall–Theil estimator; limited information maximum likelihood method; Monte Carlo studies; outliers; rational random behaviour; Rotterdam model demand system; Siegel, A.; simultaneous equations; stabilization policy; statistical decomposition analysis; Theil, H.; Theil–Sen estimator; three-stage least squares; two-stage least squares

Article

Henri Theil was a highly prolific writer of books and articles in the fields of econometrics, statistics and applied economics. The best known of his many contributions are the method of two-stage least squares, two inequality coefficients for the evaluation of forecast errors, and two measures for income inequality based on information theory, along with extensive applied work on stabilization policy, demand systems and consumption patterns.
Theil was born in Amsterdam in 1924. His parents moved first to Gorinchem and later to Utrecht, where he attended the gymnasium and obtained his diploma in 1942. He started as a student of chemistry, physics and mathematics at Utrecht University. There he was exposed to, among other subjects, formal

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
calculus, which he disliked. In his later life he disdainfully used to call it ‘epsilontics’. When he was confronted with the option between a loyalty oath to the German Nazi occupying forces and working in Germany, he decided to go underground, as did the majority of his fellow students. After some time he was caught and sent to a concentration camp, where he was incarcerated for six months. When he became seriously ill with diphtheria, his parents managed to get him out through bribery.

After the war, in 1945 he enrolled as a student at the University of Amsterdam and studied economics. In addition, he attended lectures in mathematical statistics and was a research assistant of Professor Daniel van Dantzig, at that time the most prominent Dutch statistician. As these lectures emphasized foundations rather than applications, Theil taught himself the rest of statistics by studying the two volumes of Kendall (1943; 1946). In 1951 he obtained his Ph.D. and in the same year he married Eleonore Goldschmidt. From 1951 to 1955 he worked at the Central Planning Bureau in The Hague, the main institute of policy analysis in the Netherlands, and one of the important advisers to the Dutch government.

In 1953 Theil was appointed Professor of Econometrics at the Netherlands School of Economics at Rotterdam (now Erasmus University), at first on a part-time basis while he continued his job at the Central Planning Bureau. He spent the academic year 1955–56 as a visitor at the Cowles Commission in Chicago. There he learnt the rule: ‘publish or perish’. In September 1956 he obtained a full-time position in Rotterdam, where he became the first director of the Econometric Institute of the Netherlands School of Economics. This gave him the possibility to appoint research associates and assistant professors to work for him. In the ten years of his period in the Econometric Institute he attracted 27 foreign visitors, most of whom spent a year at the institute, the most prominent ones being A. S. Goldberger, A. Zellner, F. M. Fisher, M. Nerlove and Z. Griliches. In the same year Theil also started a programme in quantitative economics, where matrix algebra and mathematical statistics were prerequisites for his courses in econometric theory.

In 1965 Theil was appointed University Professor at the University of Chicago, one of the ten posts of this type at the university. In 1981 he moved to the University of Florida (at Gainesville), where he accepted the first Eminent Scholar Chair in Florida's State University System (the McKethan–Matherly Chair). He retired in 1994, but thereafter published several articles and in 1996 a book. He died in 2000. During his life he wrote at least 18 books and about 250 published articles. Several books and about half of the articles were co-authored by more than 80 colleagues, most of them his juniors. He also supervised 15 dissertations (for details, see Raj and Koerts, 1992). He could be very generous, but a smooth cooperation with him required either acceptance of his authority or outstanding diplomatic gifts. There were numerous rumours about frictions and conflicts with colleagues and others, but written evidence is usually lacking.

Theil wrote his first well-known paper (1950), a contribution to distribution-free and robust statistics, while he was a research assistant. Consider a set of points $(x_1, y_1), \ldots, (x_n, y_n)$ and consider a pair $(x_i, y_i), (x_j, y_j)$ with $x_i \neq x_j$. Then the slope of the line through such a pair of points is given by $(y_j - y_i) / (x_j - x_i)$. He proposed the median of all these slopes (for $x_i < x_j$) as a distribution-free estimator of the regression line through the *n* points. This is an outlier robust estimator in the sense that this median is based on two ‘good’ points (that is, points that are not outliers) provided that the fraction of outliers is less than 0.293. In addition, it is a quite efficient estimator. The method is sometimes called the Kendall–Theil estimator, since Theil's paper made use of a result of M. G. Kendall on rank

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

correlation. It was extended by Sen (1968) for the case that there are ties among the *x* observations. This method is known by the name 'Theil–Sen estimator'. Finally, Theil's approach was crucially improved by Siegel (1982), who proposed the repeated median estimator. Siegel's first step takes for each *i* the median across all *j* of the above ratios ($j \neq i$), and the second step takes the median of all medians obtained in the first step. This estimator does not even break down for a fraction of outliers close to 0.5 (see Rousseeuw and Leroy, 1987, for a more extensive treatment of this subject).

Theil's best known contribution to econometrics, the two-stage least squares (2SLS) method, dates from his period at the Central Planning Bureau. He proposed it in a paper (1953) that was unpublished at the time, but it was soon known at the Cowles Commission at the University of Chicago, in those days the most important centre for econometric theory. In fact, Radner and Bobkowski (1954) of the Cowles Commission wrote a short paper in which they explained 2SLS using Cowles Commission notation. The first published description of 2SLS can be found in Theil's monograph *Economic Forecasts and Policy* (1958), while the original 1953 paper has been reprinted in Raj and Koerts (1992). The problem of systems of simultaneous equations is that most of these equations contain current (that is, non-lagged) endogenous explanatory variables. As a rule, these are not independent of the disturbances of the equation, which leads to inconsistent parameter estimates if (ordinary) least squares is applied. The 2SLS method for estimating such an equation consists of two steps. First, the current endogenous explanatory variables are regressed on the exogenous and lagged endogenous variables of the system. Second, least squares is applied to the original equation where the current explanatory endogenous variables are replaced by the explained parts of the corresponding first stage regressions. The main virtue of the method was that it was a considerable simplification on limited information maximum likelihood, the method proposed four years earlier by Anderson and Rubin (1949). This method requires the computation of eigenvalues of a matrix, a considerable burden at a time when very few universities and other research institutions had computers. (For a more extensive discussion of 2SLS and related methods, see Davidson and MacKinnon, 1993.)

The Dutch Central Planning Bureau published each year a number of forecasts on several important macroeconomic variables. Theil's tasks included the evaluation of these forecasts. One of the criteria he developed for that purpose was the *inequality coefficient*, defined by

$$U_1 = \frac{\sqrt{\frac{1}{n}\Sigma (P_t - A_t)^2}}{\sqrt{\frac{1}{n}\Sigma P_t^2} + \sqrt{\frac{1}{n}\Sigma A_t^2}},$$

where P_t and A_t denote time series for the predicted and actual changes, respectively, of a certain variable (see Theil, 1958). An attractive property of the formula is that it is positive and does not exceed unity. The zero lower bound corresponds with perfect forecasts, the unit upper bound with the worst possible set of forecasts. Unfortunately this coefficient has a serious drawback. The denominator depends on the chosen forecasting procedure, while when different forecasting procedures are to be compared it is desirable that the denominator is the same. Granger and Newbold (1973) gave a more

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
formal criticism using a time series model. Several years earlier Theil (1955) had already proposed an alternative formula:

$$U_2 = \frac{\sqrt{\frac{1}{n}\Sigma(P_t - A_t)^2}}{\sqrt{\frac{1}{n}\Sigma A_t^2}},$$

which is not subject to the above criticisms. It exceptionally takes values greater than unity and it is exactly equal to unity in the case of no change extrapolation (P_t=0 for all t). In (1966) Theil stated why he preferred U_2 to U_1, but he used the same symbol (U) and the same term (inequality coefficient) for both U_1 and U_2 with the unfortunate consequence that if 'the' Theil inequality coefficient is cited one does not know whether the author has used U_1 or U_2 unless an explicit definition or reference is given. His criticism on U_1 was not generally noticed, as is demonstrated by the fact that U_1 survives in several places, for instance, in the well-known computer application Eviews.

Theil's most important research projects in his Rotterdam years concerned forecasting, stabilization policy, and information theory. The policy project was quite ambitious. The theory of consumer demand, where a consumer is assumed to maximize his utility function subject to his budget constraint, served as a paradigm. He assumed a government with a quadratic 'welfare function' to be maximized subject to the reduced form of a linear (or linearized) macroeconometric model, taking account of the uncertainty caused by the disturbances of the model. This implied linear decision rules for the government. While the theory was already largely developed in his 1958 book, Theil applied the theory in his 1964 book to the US economy in the 1930s and to the Dutch economy in the period 1957–9. After 1964 he left the subject to others.

Theil's book on economics and information theory appeared in 1967 but was entirely written before he left Rotterdam in September 1966. The idea came up in a discussion with Barten and the author in 1962 or 1963 on Barten's first version of the demand system that would later obtain the name 'Rotterdam model'. Barten (1964) had proposed a system of differenced demand equations. Theil amended this idea, first, by proposing premultiplication by the corresponding current budget shares w_i (before assuming constant coefficients), and later replaced the current budget shares by the averages of current and lagged budget shares, as was customary in the computation of the Törnqvist approximation to the growth rate of the Divisia quantity index. When commenting on this discussion in *Studies in Global Econometrics* Theil stated, after having mentioned Barten's research: 'In 1965 I modified this approach *slightly* …' (emphasis added). Not everybody was happy about the Rotterdam model, however. It was criticized at an early stage by McFadden (see Yoshihara, 1969).

When Theil, in exploring the consequences of the assumptions underlying the Rotterdam model, arrived at terms of the type $\sum w_i \log w_i$, he recognized this as an expression from information theory, and started to study the literature on that subject (see also Raj and Koerts, 1992, Vol. 1, pp. 25–6). Theil proceeded to develop several economic applications of information formulas. The best known is his measure for income inequality

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{x_i}\right)$$

where n is the number of subsets in a population, x_i the population share of subset i in the total population, and y_i its income share in total income. The formula has attractive decomposition properties, which explains its popularity among applied researchers. In the same book he also proposed a second measure, which is obtained from the above formula by interchanging the roles of y and x. In certain situations this is considered to be preferable.

For his lectures in Chicago Theil wrote a new version of his notes on econometric methods, which appeared in 1971 under the title *Principles of Econometrics*. This book was one of the most important reference texts in the field of econometric methods during the 1970s and later. It is also a useful reference to most of his other methodological contributions, such as his work on aggregation, on specification analysis, on the *k*-class, on mixed estimation (with A. S. Goldberger), on three-stage least squares (with A. Zellner), on the final form of econometric equation systems (with J. C. G. Boot), on efficient estimation of disturbances, and on several other subjects.

In his Chicago period, Theil also extended his work on statistical decomposition analysis (1972) and on consumer demand systems (1975; 1976). In the early 1970s he wrote three papers containing an economic theory of the second moments of disturbances of behavioural equations, also called rational random behaviour (see his 1975 and 1980 books for summaries and references). This was an ambitious project, which, however, met with little response in the profession.

The demand systems were relatively rich in parameters, with the consequence that asymptotic results were unreliable. Theil organized several Monte Carlo studies (as a rule carried out by his students or associates) in order to get more reliable test results. A survey of this work is given in Theil (1986).

In 1978 *Economics Letters*, a new journal, was launched. It specialized in short papers, only lightly refereed and published with minimum delay. Theil very much liked this formula and during the following 20 years he published (with several co-authors) a huge number of papers in the journal on a wide variety of subjects. His principal interest was now in world income inequality and international consumption patterns. For the latter purpose he developed a new type of demand system, later called the ‘Florida model’, and a new estimation method based on information theory. Surveys of most of this research can be found in Theil (1987; 1989; 1996).

Three of Theil's books became citation classics: *Economic Forecasts and Policy* (1958), *Economics and Information Theory* (1967), and *Principles of Econometrics* (1971). He obtained honorary degrees from the University of Chicago (1964), the Free University of Brussels (1974), Erasmus University Rotterdam (1983) and Hope College in Michigan (1985).

Given Theil's enormous productivity, the above survey of his work is necessarily very incomplete. The three volumes edited by Raj and Koerts (1992) contain considerably more information about his work up to 1992. Also of interest is a Festschrift edited by Bewley and Tran Van Hoa (1992) that appeared at about the same time. The interview by Bewley (2000) gives a good impression of Theil as a person and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

research worker. In all these publications one can also find more about his life (see also Kloek, 2001; 2002). In his final years he developed a taste for autobiographical writing (see Theil, 1996, pp. 1–6; 1997).

See Also

- econometrics
- forecasting
- Griliches, Zvi
- two-stage least squares and the *k*-class estimator
- wage inequality, changes in

The author acknowledges permission to reproduce copyright material from Kloek (2001, pp. 263–9).

Selected works

1950. A rank-invariant method of linear and polynomial regression analysis. *Proceedings of the Royal Netherlands Academy of Sciences* 53, 386–92, 521–5, 1397–412.

1953. Estimation and simultaneous correlation in complete equation systems. Mimeographed Memorandum of the Central Planning Bureau, The Hague. Reprinted in Raj and Koerts (1992), vol. 1, ch. 6.

1955. Who forecasts best? *International Economic Papers* 5, 194–9.

1958. *Economic Forecasts and Policy*. Amsterdam: North-Holland.

1964. *Optimal Decision Rules for Government and Industry*. Amsterdam: North-Holland.

1966. *Applied Economic Forecasting*. Amsterdam: North-Holland.

1967. *Economics and Information Theory*. Amsterdam: North-Holland.

1971. *Principles of Econometrics*. New York: Wiley.

1972. *Statistical Decomposition Analysis: With Applications in the Social and Administrative Sciences*. Amsterdam: North-Holland.

1975, 1976. Theory and Measurement of Consumer Demand, 2 vols. Amsterdam: North-Holland.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1980. *The System-Wide Approach to Microeconomics*. Oxford: Basil Blackwell.

1986. (With T. Taylor and J. Shonkwiler.) Monte Carlo testing in systems of equations. In *Advances in Econometrics*, vol. 5, ed. D. Slottje and G. Rhodes, Jr. Greenwich, CT: JAI Press.

1987. (With K. Clements) *Applied Demand Analysis: Results from System-Wide Approaches*. Cambridge, MA: Ballinger.

1989. (With C.-F.Chung and J. Seale, Jr.) *International Evidence on Consumption Patterns*. Greenwich, CT: JAI Press.

1996. *Studies in Global Econometrics*. Dordrecht: Kluwer.

1997. Rotterdamse Herinneringen [Rotterdam memories]. In *Kritisch en Constructief* [Critical and Constructive], ed. H. Van Dijk et al. Rotterdam: Liber Amicorum for T. Kloek, Econometric Institute, Erasmus University.

Bibliography

Anderson, T. and Rubin, H. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20, 46–63.

Barten, A. 1964. Consumer demand functions under conditions of almost additive preferences. *Econometrica* 32, 1–38.

Bewley, R. 2000. Mr Henri Theil: an interview with the International Journal of Forecasting. *International Journal of Forecasting* 16, 1–16.

Bewley, R. and Tran Van Hoa. 1992. *Contributions to Consumer Demand and Econometrics*. London: Macmillan.

Davidson, R. and MacKinnon, J. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Granger, C. and Newbold, P. 1973. Some comments on the evaluation of economic forecasts. *Applied Economics* 5, 35–47.

Kendall, M. 1943, 1946. The Advanced Theory of Statistics, 2 vols. London: Griffin.

Kloek, T. 2001. Obituary: Henri Theil, 1924–2000. *Statistica Neerlandica* 55, 263–69.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kloek, T. 2002. Henri Theil in Rotterdam, 1953–1966. Paper presented at the Theil Memorial Conference, Amsterdam.

Radner, R. and Bobkoski, F. 1954. Some recent work of H. Theil on estimation in systems of simultaneous equations. Unpublished discussion paper of the Cowles Commission. Reprinted in Raj and Koerts (1992), vol. 1, ch. 8.

Raj, B. and Koerts, J. 1992. *Henri Theil's Contributions to Economics and Econometrics*, 3 vols, Dordrecht: Kluwer.

Rousseeuw, P. and Leroy, A. 1987. *Robust Regression and Outlier Detection*. New York: Wiley.

Sen, P. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63, 1379–89.

Siegel, A. 1982. Robust regression using repeated medians. *Biometrika* 69, 242–4.

Yoshihara, K. 1969. Demand functions: an application to the Japanese expenditure pattern. *Econometrica* 37, 257–74.

How to cite this article

Kloek, Teun. "Theil, Henri (1924–2000)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000195> doi:10.1057/9780230226203.1697

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

theory appraisal

Ellery Eells and Daniel M. Hausman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists and other scientists appraise theories in terms of criteria such as evidential support, predictive accuracy, usefulness, and reliability in research and practice. This entry addresses three general problems concerning theory appraisal. 1. What is it for evidence to be confirmationally relevant to a theory? 2. How can evidential support be measured? 3. On what other criteria should theory appraisal in economics depend? The third question raises special problems, since economic models so often incorporate statements that appear to be false. But answers to general questions concerning confirmation matter to the conduct of economics, too.

Keywords

Bayes, T.; Bayesian confirmation theory; *ceteris paribus* ; confirmation; Duhem–Quine problem; experimental economics; Friedman, M.; habit; Hempel, C.; Hume, D.; hypothetico-deductive method; induction; likelihood; McCloskey, D.; Mill, J. S.; Popper, K.; prior probability; probability; rhetoric of economics; subjective probability; testing; theory appraisal

Article

There are several reasons why economists and others appraise theories and models. They may want to judge whether theories and models are worthy of study, whether one can rely on them in research and practice, or whether one can judge them to be true or false or predictively adequate. Different purposes may call for different appraisals.

All sciences face three general problems of theory appraisal. 1. What is it for evidence *e* to confirm, disconfirm, or be evidentially irrelevant to an hypothesis *h*, and how is confirmation possible? 2. How should scientists measure the extent to which *e* confirms *h* or how well confirmed *h* is overall? 3. How does the appraisal of a theory depend on the extent to which it is confirmed, and on what else may theory appraisal depend? The third question raises special problems with respect to economics, because

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the conclusions of economics are usually hard to test, whether experimentally or against market data. Economic models incorporate many statements that appear to be false, and the basic principles of economics typically contain, at least implicitly, vague and ineliminable *ceteris paribus* qualifications. These peculiarities are not unique to economics, though their combination and the particular form that questions of theory appraisal take in economics are distinctive.

Answers to general questions concerning confirmation and theory appraisal have immediate implications for the conduct of economics. For example, defenders of real business cycle theory have constructed computational ‘experiments’ in which technology shocks in a calibrated computer model of a simplified economy give rise to simulated business cycles. Are these appropriate methods for testing economic theories? Do the results of these ‘experiments’ justify a positive appraisal of real business cycle theory (Kydland and Prescott, 1996)?

The discussion of confirmation and its measurement, which occupies the next two sections, is quite general, while the discussion of theory appraisal in the final section emphasizes distinctive features of economics.

Confirmation

Evidence *e* confirms an hypothesis *h* when it provides some support for *h*; *e* disconfirms *h* when it provides some evidence against *h*. The theory of confirmation is not concerned with the limiting cases of conclusive proof or refutation. Unlike deductive inferences, in which it is impossible for the premises to be true and the conclusion false, confirmation is a matter of inductive inference, which is fallible. From the two premises, ‘All humans are mortal’ and ‘Bill Gates is human’, one can deduce that Bill Gates is mortal. Though infallible, a deductive inference such as this one is also ‘non-ampliative’ – the conclusion is already implicit in the premises, and drawing the conclusion arguably does not advance our knowledge. To infer that Bill Gates will die from the premises that he is human and that all humans born before 1850 have died is, in contrast, to take a risk. The conclusion goes beyond what is asserted in the premises. The conclusion that *all* humans are mortal obviously goes far beyond any data concerning past births and deaths.

David Hume (1748) issued a serious challenge to the rationality of inductive inferences and to the idea that there is such a thing as confirmation or rational support. As an empiricist, he insisted that all evidence derives from perception and thus can be stated as reports of observations. He then asked what argument employing only observation reports as premises could establish universal generalizations or conclusions concerning phenomena not yet observed. Such arguments could not be deductive, since they are fallible. There was nothing logically impossible about Europeans observing black swans in Australia, even though all swans Europeans previously observed had been white. To get from premises reporting observations to a generalization or a claim about something not yet observed requires as an additional premise or as a rule of inference some principle to the effect that nature is, in the relevant respect, uniform or regular. But Hume points out that we are entitled to rely on such a premise only if it is a logical truth or is itself established by experience. A principle of uniformity is not a logical truth, and it is no easier to establish on the basis of reports of past uniformities than the particular induction we are attempting to justify. Hume draws the extreme sceptical conclusion that humans have no good reason at all to believe the conclusions of inductive inferences, or, in other words, that observational evidence never provides any rational support for any hypotheses. What keeps people out of harm's way and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
enables them to survive is habit. Observing regularities, people cannot help expecting them to persist. Hume confesses that, once he leaves his study, he cannot take his sceptical conclusions seriously. Furthermore, treating confirmation as no more than a matter of conditioned responses to repetition leaves one with few ways to distinguish between sensible uses of evidence on the one hand and prejudice, superstition or even phobia on the other. Hume's problem of induction shows that the notion of justification that is relevant to science and everyday life is piecemeal rather than global. In inductive arguments, scientists can legitimately employ premises that are not observation reports and are hence not beyond questioning. (And, of course, observation reports are fallible, too.) In order to clarify and inform scientific practice, theories of confirmation need to capture the distinction between evidence that supports an hypothesis and evidence that does not, rather than to deny that this distinction exists.

In thinking about confirmation, it is helpful to draw three distinctions. First, one may distinguish between 'incremental' confirmation – confirmation as a relation between an hypothesis *h* and a particular piece of evidence *e* – and total or absolute confirmation – as a judgement about how well supported *h* is overall or as a relation between *h* and the total available evidence. Evidence *e* confirms *h* incrementally if *e* provides (or would provide) some additional support for *h*, whether or not *h* is well confirmed in the sense of total confirmation. Second, confirmation can be thought of either *qualitatively* or *quantitatively*. In the total sense, *h* is either qualitatively well confirmed or poorly confirmed, while quantitative confirmation theory attempts to quantify how well or poorly confirmed *h* is. In the incremental sense, evidence *e* may, qualitatively, either confirm, disconfirm, or be evidentially irrelevant to an hypothesis *h*, while quantitative theory attempts to measure how much *e* boosts or diminishes how well supported *h* is. Finally, confirmation can be either *comparative* or *non-comparative*. Non-comparative confirmation concerns just one hypothesis–evidence pair. In comparative confirmation, on the other hand, one assesses, for example, how well an *e* supports an *h* relative to how well a different e' supports the same *h* or how well the same *e* supports a different h'.

A simple and natural idea is that a general hypothesis of the form 'All *F*s are *G*s' is (incrementally) confirmed by 'positive instances' – that is, by objects that are both *F* and *G*. A negative instance – an object that is an *F* but not a *G* – does not merely disconfirm the general hypothesis; negative instances refute universal generalizations. This somewhat misleading asymmetry plays a large role in Karl Popper's insistence that science relies on confirmation rather than verification. For an example of instance confirmation, an increase in a firm's sales following a drop in the price of its product confirms the law of demand. If in addition to instance confirmation one assumes that logically equivalent hypotheses are confirmed by the same evidence, one is immediately faced with a paradoxical conclusion. A white shoe confirms the generalization 'Everything that is not black is not a raven', which is logically equivalent to the generalization 'All ravens are black'. But do reports of white shoes confirm 'All ravens are black' (Hempel, 1945)? The most common of the many responses to this paradox rely on a quantitative theory of confirmation and are discussed briefly below.

The positive instance criterion was offered as part of a sufficient condition for the confirmation of universal conditionals. Hempel (1945) generalized this criterion to his *satisfaction criterion* which applies to more complex logical structures for evidence and hypothesis, and which provides explicit definitions of confirmation, disconfirmation, and evidential irrelevance. In defending his satisfaction criterion, Hempel relied on what he regarded as intuitively obvious *conditions of adequacy* for definitions of confirmation. Besides the *equivalence condition* mentioned above, the *entailment condition* covers the limiting case where evidence logically entails an hypothesis: it says that, if *e* entails

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
h, then *e* confirms *h*. Hempel also assumed a more controversial *special consequence condition*, which says that, if *e* confirms *h* and *h* entails *h′*, then *e* confirms *h′*.

The criteria for confirmation discussed above apply when evidence reports and hypotheses are stated in the same language, which Hempel took to be an observational language. What about confirmation of *theories*, which typically contain theoretical as well as observational vocabulary? *Hypothetico-deductivism* (HD) or the *hypothetico-deductive* method is the idea that theories and hypotheses are confirmed by their observational consequences. So, for example, axioms concerning individual preference are confirmed by the same evidence that supports the law of demand, even though that evidence is not an instance of the preference axioms, because that evidence (and the law of demand itself) can be deduced from models which include axioms governing individual preferences. The positive instance and satisfaction criteria are formulations of the idea, roughly, that observations that are *logically consistent with* a hypothesis confirm the hypothesis, while HD says that *deductive consequences of* a theory or model confirm the theory or model. Of course, if the prediction fails (if an observational deductive consequence of a theory turns out to be false), then this is supposed to provide *disconfirmation* of the theory.

The HD method is subject to two serious problems. The first is the problem of irrelevant conjunction or of distributing credit. If an hypothesis *h* logically implies an observation report *e*, then so does the conjunction, *h*&*g*, where *g* can be any sentence whatsoever. If, as the HD method says, an observation report *e* confirms any hypothesis that implies *e*, then *e* confirms *h*&*g*. That seems bad enough. Worse still, since *h*&*g* implies *g*, the special consequence condition implies that *e* confirms *g*, or in other words any sentence whatsoever. A natural response would be to refine the basic HD idea and deny that *e* confirms *h* when *h* entails *e*, unless some further condition is met to the effect that there is no logically weaker hypothesis *h′* that also entails *e*. But these difficulties are not easily remedied.

The second problem, which is particularly poignant in economics, concerns the distribution of blame. Testable implications can rarely if ever be deduced from scientific theories without additional premises. Deductive logic tells us that, if those testable implications are false, then at least one of the premises from which they follow must be false, but deductive logic alone does not single out which is the culprit. This is the so-called ‘Duhem–Quine problem’ and is discussed later.

Quantitative confirmation: probabilistic approaches

The most influential probabilistic approach to confirmation is called ‘Bayesian confirmation theory’. The basic idea is that evidence *e* confirms hypothesis *h* if and only if the conditional probability $p(h/e)$ (defined as $p(h\&e)/p(e)$) is greater than the unconditional probability $p(h)$. Disconfirmation is defined by reversing the inequality, and evidential irrelevance is defined by changing the inequality to an equality. The function *p* measures an agent's ‘subjective probability’ or degree of belief. Given a set of axioms governing preference, which are stronger than those that must be satisfied in order to attribute ordinal utility functions to agents, one can prove that there exists a ‘cardinal’ utility representation, which is unique up to a positive affine transformation, and which assigns to each lottery a utility equal to the sum of the utilities of its prizes weighted by the agent's subjective probabilities. Such cardinal representation theorems (for example, Harsanyi, 1977, ch. 4) establish that the degrees of belief of agents who satisfy the axioms conform to the axioms of the probability calculus. (Important work in foundations of subjective probability includes Ramsey, 1931; de Finetti, 1937; Savage, 1972; Jeffrey, 1983; and Joyce,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1999.)

Suppose that $p(h/e)$, the agent's subjective conditional probability, is equal to what the agent's subjective degree of belief in *h* would be if the agent learned that *e* is true – that is, in the wake on an observation that *e* is true, the agent changes or updates *p*(*h*) to be equal to the current $p(h/e)$. On this assumption $p(h/e) > p(h)$ if and only if the agent's degree of belief in *h* would increase if *e* were learned. In that case, it is natural to say that, for this agent, *e* is positively evidentially relevant to *h*, even when *e* is not in fact learned. So the Bayesian maintains that *e* confirms *h* if and only if $p(h/e) > p(h)$, that *e* disconfirms *h* if and only if the inequality is reversed, and *e* is evidentially irrelevant if and only if $p(h/e) = p(h)$.

People's subjective probabilities will differ, even if they are equally rational, because their background knowledge differs. So, according to the Bayesian approach, confirmation is a relation among three things: evidence, hypothesis and background knowledge. The reason this approach is called ‘Bayesian’ confirmation theory is because of its frequent use of a mathematical theorem discovered by Thomas Bayes (1764), a simple version of which is: $p(h/e) = p(e/h)\,p(h)/p(e)$. This expression follows easily from the definition of conditional probability which tells us that *p*(*h*/*e*)=*p*(*h*&*e*)/*p*(*e*) and *p*(*e*/*h*)=*p*(*h*&*e*)/*p*(*h*). Bayes's theorem shows how the quantity of interest, $p(h/e)$, depends on the so-called ‘prior probability’ – *p*(*h*) – the ‘likelihood’ – $p(e/h)$ – and the subjective probability of the evidence. When *p*(*e*) is low – that is, when an observation is surprising – then one has much stronger evidence than if the observation was expected.

Bayesian confirmation theory also suggests *measures* of the *degree* of evidential support that *e* provides for *h*. The most common measure is the *difference measure*: $d(h, e) = p(h/e) - p(h)$, where confirmation, disconfirmation and evidential irrelevance correspond to whether this measure is positive, negative or zero, and degree goes by the magnitude of the difference, but there are other measures, too.

One application of the idea of degree of evidential support has been to the ravens paradox, discussed above. Let *h* be the hypothesis that all ravens are black; let *Ra* symbolize the statement that object *a* is a raven; and let *Ba* symbolize the statement that *a* is black. If *h* is probabilistically independent of *Ra* (that is, $p(h/Ra) = p(h)$), then a positive instance (or report of one), *Ra*&*Ba*, of *h* confirms *h* in the Bayesian sense (that is, *p*(*h*/*Ra*&*Ba*)>*p*(*h*)) if and only if $p(Ba/Ra) < 1$ (that is, if it is not *already* certain that *a* would be black if a raven). Given the same independence assumption, one can prove that a positive instance, *Ra*&*Ba*, confirms *h* more than a contrapositive instance, *not-Ba*&*not-Ra*, does, provided that one's degree of belief that a given raven will be black is less than one's degree of belief that something that isn't black isn't a raven. Since for most people the subjective probability that a particularly non-black thing such as a white shoe is not raven is much higher than the subjective probability that any particular raven is black, black ravens will provide much stronger confirmation of ‘All ravens are black’ than white shoes or pale economists.

Bayesian confirmation theory can also be used to assess Hempel's proposed conditions of adequacy for criteria of confirmation. In particular, Hempel's special consequence condition (unlike the equivalence conditions or the entailment condition) does not follow from the Bayesian definition of confirmation and the axioms of the probability calculus, and by attending to the circumstances in which the special consequence condition can fail, theorists have constructed intuitively compelling examples of an *h* entailing an *h*′, and an *e* confirming *h* while disconfirming *h*′. Such examples tell against the special consequence condition and also in favour of Bayesian confirmation theory (for example, Eells, 1982).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
One standard objection to Bayesian confirmation theory is that, if *e* is already known to be true, then $p(h/e)$ must be the same as *p*(*h*) and, by the Bayesian definition of confirmation, *e* cannot confirm *h*. This is called 'the problem of old evidence'. One possible Bayesian solution to the problem, suggested by Glymour (1980), would be to say that it is not the already known *e* that confirms *h*, but rather a newly discovered logical or explanatory relation between *e* and the particular *h*. Other solutions have been proposed and various versions of the problem have been distinguished (see Earman, 1992, for discussion). The problem remains one of lively debate.

A second objection to Bayesian confirmation theory is more general and leads to the main contemporary probabilistic alternative. Prior probabilities – the degree of belief agents have in hypotheses in advance of gathering particular evidence – and likelihoods – $p(e/h)$ – play crucial roles in Bayesian confirmation theory. But is it plausible to suppose that these are known? If *h* is a newly formulated physical hypothesis, for example, what would *justify* an assignment of probability to it prior to evidence? One response to the problem is to point to convergence theorems (as in de Finetti, 1937) which show that differences in priors will 'wash out' in the long run. But this response is not very satisfactory, since assessments of theories need to be made now, and if initial differences in priors are large enough, the long run must be long indeed.

Those who favour a *likelihood* approach to the evaluation of evidence – for example, Edwards (1972) and Royall (1997) – believe that there are important contexts in which likelihoods can be known, but that reliance on prior probabilities is rarely justifiable. According to one formulation of the likelihood account, *e* confirms *h* more than *e* confirms h' if and only if $p(e/h)$ is greater than $p(e/h')$. This is a comparative principle. It doesn't matter that both of these likelihoods may be tiny. The likelihood approach separates the question of which hypothesis it is more justified to believe given all the evidence from the question of which hypothesis is better supported by the particular piece of evidence. There are a variety of different measures of confirmation in terms of likelihoods (see Fitelson, 2001, and Forster and Sober, 2002, for recent discussion and references). Although those who emphasize likelihoods avoid relying on priors, they cannot of course avoid relying on likelihoods, and these too may be hard to know. Although evidence statements are often entailed by the conjunction of the hypothesis under test and a variety of other statements concerning the specific circumstances, measuring apparatus, the absence of interferences and so forth, the conditional probability of the evidence statement conditional on the particular hypothesis may be as hard to nail down as the prior probability.

Inexactness and theory appraisal

The appraisal of theories ought presumably to be strongly influenced by how well confirmed they are both absolutely and in comparison with alternatives. If there is very strong evidence in support of a theory – that is, if a theory is very well confirmed – then the risk of believing it and relying on it both in practice and for the purposes of gaining further knowledge should be small.

But, given how ambitious and wide-reaching the claims of theories often are, it is questionable whether even the best-supported of scientific theories is all that well supported. How could evidence from our little corner of the universe justify conclusions about all bodies and all forces? Furthermore, other features of theories should influence appraisal of them, in addition to how well supported they are. Some theories are more 'promising' than others. Some are easier to use or to teach. Some theories are more

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
compatible than others with strongly held metaphysical and religious beliefs or with other branches of science. A fruitful falsehood may be worth more than a tedious truth. Appraising theories and choosing among them depends on much more than how well confirmed they are.

Appraising economic theories is particularly challenging because of the limited possibilities for experimentation and because of the limited relevance of market data, which are influenced by many factors from which economic theories abstract. The Duhem–Quine problem is a huge practical problem rather than just a philosophical possibility. Consider, for example, Card and Kreuger's (1995) study of the effect of an increase in minimum wages in New Jersey in 1990, which found, in contradiction to the prediction of simple economic models, that unemployment among unskilled workers did not decline. To carry out this test, they compared employment in fast-food restaurants in New Jersey, where minimum wages increased, and next-door Pennsylvania, where minimum wages did not change. Their results might show, as they maintained, that relatively small changes in the minimum wage do not cause unemployment among unskilled workers. But the results might instead be due to peculiarities of fast-food restaurants, to other changes in the economy or popular culture of New Jersey or Pennsylvania, or to flaws in their data. The problem is ubiquitous: market data bear on economic theories through so many intermediary premises that successes of predictions give economists little reason for confidence in their models, and failures give economists little reason for dissatisfaction with their models. To bridge the gap between even the very best data and economic theories requires a motley assortment of simplifications and *ceteris paribus* clauses, which rarely inspire much confidence. When predictions fail, it is typically more sensible to blame the problem on the simplifications or *ceteris paribus* conditions than on the theory. That means that economists can hang on to their basic principles without much risk of refuting them but also without much prospect of improving them.

If one turns from predictions given by economic models to the propositions out of which the models are constructed, one finds that they rely heavily on premises such as ‘Commodities are infinitely divisible’, ‘All agents have perfect information concerning prices and quantities of all commodities’, or ‘Consumption decisions are taken by a single representative consumer’. No one thinks these claims are true, but many believe that their falsity does not matter. Perhaps they can be regarded as approximations. Perhaps they can be regarded as idealizations. Perhaps they can be regarded as abstracting from details that are, for some purposes, irrelevant. In addition, many of the basic principles or ‘laws’ of economics – claims such as ‘Agents’ preferences are transitive’ or ‘Consumers prefer more commodities to fewer’ – are false, at least if they are construed as universal generalizations. Yet they seem somehow to capture important facts that enable economists to build models that predict and explain important economic phenomena.

The traditional view of theory appraisal in economics receives its best exposition in the works of John Stuart Mill (1843, Book VI). Although an empiricist, Mill thought in terms of causation. Individual causes can be and ought to be studied by observation (if one is fortunate enough to find a domain where no other causal factors are interfering) or experimentally (if one can construct controlled circumstances which differ only with respect to the presence or absence of the particular cause one investigates). Mill's famous methods of induction guide inferences concerning individual causes (1843, Book III).

When one is studying phenomena that are influenced by many causes, like the phenomena that economists study, Mill maintains that direct inductive methods cannot be used. Instead one needs first to determine the laws of the main causes by studying other domains where their action can be isolated, and then one needs to deduce their effect when combined. Mill believes that it is possible to learn the basic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
behavioural generalizations governing economic behaviour by means of introspective psychology on the one hand and engineering on the other. Economics then takes the form of exploring the implications of models in which economists combine what they know of the most important causal factors with simplified descriptions of the circumstances to which the model is to be applied. So, for example, economists know that firms generally aim to maximize net revenue and that they have a choice of different production processes which employ relatively more or relatively less unskilled labour. When minimum wage laws increase, wages of unskilled labour increase (unless market rates are already higher than the new minimum wage), and firms will economize on the use of unskilled labour. Employment of unskilled labour will consequently decline. Mill urges in addition that empirical tests such as the Kreuger and Card study be carried out as checks on economists' deductions and to determine whether they have left out some significant factor. But predictive failures do not show any errors in one's basic principles, which have, in Mill's view, already been conclusively established by the results of introspective observation of the principles separately. The appraisal of basic theory depends on the quasi-experimental study of its individual basic 'laws' or principles, and it is largely disconnected from the appraisal of specific models, which depends on whether they serve the particular purposes for which they are constructed.

Although Mill's view that the basic principles of economics are proven truths is unsustainable, he was right to emphasize the difficulty of testing economics against uncontrolled market data and to stress the value of evidence concerning specific principles gathered from everyday experience. As the example of the employment effects of minimum-wage laws illustrates, economists rely on everyday experience to suggest and justify generalizations concerning the behaviour of firms including the flexibility of their employment policies. Causal experience cannot, of course, show how single-minded or relentless firms are in maximizing net returns, or whether this causal factor is one of a small set of factors that is of supreme importance with respect to virtually all economic phenomena. But, given the complexities and ambiguities in empirical studies such as Card and Krueger's, it is not unreasonable to continue to expect increases in unemployment among unskilled workers to result from increases in minimum wages, and it is not unreasonable to hang on to fundamental principles such as profit maximization.

In the second half of the 20th century, many economists became uneasy about espousing a view of theory appraisal that emphasized the extent to which the basic principles of economics were independently confirmed or disconfirmed, often by casual experience. One reason was that empirical work apparently showed that firms do not attempt to maximize profits (Hall and Hitch, 1939; Lester, 1946). Rather than addressing the details of these studies, Milton Friedman (1953) argues famously that such inquiries into the 'realism' of the 'assumptions' of economics reflected methodological confusion. All that matters is how successfully economic models are at predicting the market price and quantity data that are of interest to economists. Friedman's view appeared to economists to be in better accord with an up-to-date philosophy of science. It was empiricist, and, later on, many saw it as corresponding roughly to Karl Popper's insistence on falsification (Blaug, 1976, p. 149). Given how difficult it is to test economic theory against market data, Friedman's view of theory appraisal in fact served to insulate economics from empirical criticism, which may have been its main appeal.

During the second half of the 20th century methodologists flirted with a number of views of theory appraisal, many of which were constructed with the natural sciences or even mathematics in mind, and none of which has found general favour among economists or economic methodologists. The difficulties of theory appraisal have even led some influential voices to argue for an abandonment of the pursuit of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
any normative standards. McCloskey (1985) thus apparently argues that, apart from very unspecific requirements of honesty and open discussion, there are no rules constraining the appraisal economists offer of their theories, while Hands (2001) suggests that the only fruitful questions to ask concerning theory appraisal are questions about the sociological factors that lead communities of scientists to endorse one theory or another.

In our view, in contrast, it is impossible to evade normative questions about how theories and models ought to be appraised. Though these questions have no simple answers, the answers must turn in large part on matters of confirmation. Decisions to rely on particular claims of economics both for theoretical and for practical purposes are weighty indeed, and it would be massively irresponsible to base policy on some economic thesis without asking how well confirmed it is.

See Also

- assumptions controversy
- ceteris paribus
- experimental economics
- falsificationism
- Mill, John Stuart
- models
- philosophy and economics
- rhetoric of economics
- science, economics of
- testing

Bibliography

Bayes, T. 1764. An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London* 53, 370–418.

Blaug, M. 1976. Kuhn versus Lakatos *or* Paradigms versus research programmes in the history of economics. In *Method and Appraisal in Economics*, ed. S. Latsis. Cambridge: Cambridge University Press.

Card, D. and Krueger, A. 1995. *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton: Princeton University Press.

de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincaré* 7, 1–68. English translation in *Studies in Subjective Probability*, ed. H. Kyburg and H. Smokler. New York: John Wiley, 1964.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Duhem, P. 1954. *The Aim and Structure of Physical Theory*, trans. P. Wiener. Princeton: Princeton University Press, 1954.

Earman, J. 1992. *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA and London: MIT Press.

Edwards, A. 1972. *Likelihood*. Cambridge: Cambridge University Press.

Eells, E. 1982. *Rational Decision and Causality*. Cambridge and New York: Cambridge University Press.

Fitelson, B. 2001. Studies in Bayesian confirmation theory. Ph.D. dissertation, University of Wisconsin-Madison.

Forster, M. and Sober, E. 2002. Why likelihood? In *The Nature of Scientific Evidence*, ed. M. Taper and S. Lee. Chicago: University of Chicago Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago: University of Chicago Press.

Glymour, C. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Hall, R. and Hitch, C. 1939. Price theory and business behaviour. *Oxford Economic Papers* 2, 12–45.

Hands, D. 2001. *Reflection without Rules: Economic Methodology and Contemporary Science Theory*. Cambridge: Cambridge University Press.

Harsanyi, J. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Hempel, C. 1945. Studies in the logic of confirmation. *Mind* 54, 1–26, 97–121. Reprinted, with his Postscript, in *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: Free Press, 1965.

Hume, D. 1748. *An Inquiry concerning Human Understanding*. Indianapolis: Bobbs-Merrill, 1955.

Jeffrey, R. 1983. *The Logic of Decision*. 2nd edn. Chicago and London: University of Chicago Press.

Joyce, J. 1999. *The Foundations of Causal Decision Theory*. Cambridge and New York: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kydland, F. and Prescott, E. 1996. The computational experiment: an econometric tool. *Journal of Economic Perspectives* 10(1), 69–85.

Lester, R. 1946. Shortcomings of marginal analysis for wage-employment problems. *American Economic Review* 36, 62–82.

McCloskey, D. 1985. *The Rhetoric of Economics*. Madison: University of Wisconsin Press.

Mill, J. S. 1843. *A System of Logic*. London: Longman, Green & Co., 1949.

Ramsey, F. 1931. Truth and probability. In *The Foundations of Mathematics and other Logical Essays*, ed. R. Braithwaite. London: Routledge and Kegan Paul.

Royall, R. 1997. *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, FL: Chapman and Hall.

Savage, L. 1972. *The Foundations of Statistics*, 2nd edn. New York: Dover Publications, Inc.

How to cite this article

Eells, Ellery and Daniel M. Hausman. "theory appraisal." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000216>
doi:10.1057/9780230226203.1698

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

thin markets

Marzena Rostek and Marek Weretka
From The New Palgrave Dictionary of Economics Online, ,
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A thin market is a market with few buying or selling offers. The concept of market thinness, while general, is typically used in the context of financial markets. When the number of buying or selling offers is small, investors' trading positions are large relative to market size. Trading then requires price concessions and thus exerts an impact on prices. A thin market is characterized by low trading volume, high volatility and high bid–ask spreads. This article discusses the modelling of thin markets, some typical phenomena of such markets, and their implications for market design.

Keywords

asset pricing; blockage discount; inventory models; liquidity; market efficiency; market power; market structure; oligopoly; overshooting; predatory trading; thin markets

Article

A thin market is a market with few buying or selling offers. It is also known as a narrow market. The signature characteristic of a thin market is traders' price impact. When the number of buying or selling offers is small, investors' trading positions are large relative to market size. Trading then requires price concessions and thus exerts an impact on prices. A thin market is characterized by low trading volume, high volatility, and high bid–ask spreads. The concept of market thinness, while general, is typically used in the context of financial markets.
Market thinness is a particular source of market illiquidity. Liquidity is broadly defined as the ease of trading a security. Apart from market power, lack of liquidity can result from asymmetric information, transaction costs, search and bargaining frictions, imperfect financing ability, limited commitment and spatial considerations.

Price impact in financial markets

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Since transaction-level data became available in the early 1990s, it has been well understood that institutional investors (such as mutual funds, hedge funds, pension funds and investment banks), whose trade on the New York Stock Exchange (NYSE) accounts for more than 70 per cent of daily trading volume, exert a significant impact on prices and take this into account in their trading strategies. The seminal empirical studies include Chan and Lakonishok (1993, 1995), and Keim and Madhavan (1995, 1996, 1998). To mitigate the adverse effects of price impact, large traders do not place their orders at once; rather, they break them up into smaller blocks, which are then placed sequentially. For instance, at the NYSE, only about 20 per cent of the total trading value of all institutional purchases and sales is completed within a single trading day, while more than 50 per cent takes at least four days for execution. If traded at once, a typical institutional package would represent more than 60 per cent of the total trading volume. In financial slang, institutional investors are referred to as elephant traders and institutional trading blocks as iceberg orders. In fact, the trading costs associated with such price impact dominate the explicit costs of trade, such as commission, order processing and brokerage fees. Consequently, extensive resources are devoted to estimating price impact and designing best execution. Such techniques are available to institutional as well as retail investors in the form of software called market impact models.

Modelling thin markets

On the theory side, the presence of market power poses challenges to modelling. In particular, the competitive approach is not suitable for modelling thin markets since it assumes that no individual trader can affect the market price by their buying or selling orders.

A large body of literature has emerged to explain how price impact affects individual portfolio choices of investors and the equilibrium in financial markets. The theoretical mechanisms underlying these models can be grouped into three categories: asymmetric information, inventory effects and nonequilibrium mechanisms. Traditionally, the leading class of models with price impacts is based on asymmetric or private information (e.g. Glosten and Milgrom, 1985; Kyle, 1985, 1989; Easley and O'Hara, 1987; Back, 1992; Foster and Viswanathan, 1996; Holden and Subrahmanyam, 1996). In such models, price impact arises because high sales by an informed trader are interpreted by remaining traders as a signal of low asset value, and hence reduce the asset price.

For many market events that involve anticipated demand or supply shocks, such as preannounced inclusions of new stock into the S&P, the price impact component that derives from asymmetric information can only partially explain the observed magnitudes of price changes. Therefore, an alternative strand of the literature based on inventory effects has emerged. There, because of diversification concerns, risk-averse traders are willing to absorb large, risky orders only at price concessions (Ho and Stoll, 1981; Grossman and Miller, 1988; Vayanos, 2001; Attari, Mello and Ruckes, 2005; Brunnermeier and Pedersen, 2005; Pritsker, 2005; DeMarzo and Uroševi•, 2006 extended by Uroševi•, 2005). These papers capture price impact by building Cournot-type models with one or several large investors and a continuum of (infinitesimally) small price-taking traders.

Under Cournot market structure, large investors trade only with small competitive traders. In contrast, it is a stylized fact of financial markets that large investors trade effectively with one another in the sense that an order placed by a large investor is primarily absorbed by (a possibly small group of) other large

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
traders. Essentially, this is a market structure of bilateral oligopoly, except that all traders can buy and sell. The seminal papers embedding these features are Kyle (1989) and Vayanos (1999). An equilibrium model of symmetric-information market environments in which all buyers and sellers have price impact can be found in Weretka (2006). Rostek and Weretka (2007) study dynamic thin markets and establish implications of market thinness for asset pricing. Rostek and Weretka (2008) examine the implications of market thinness for information aggregation, efficiency and price discovery. Carvajal and Weretka (2007) show that the pricing kernel exists in thin markets.

Finally to study such markets, several nonequilibrium models with price impact have been proposed (e. g. Bertsimas and Lo, 1998; Almgren and Chriss, 2000; Subramanian and Jarrow, 2001; Dubil, 2002; Almgren, 2003; Huberman and Stanzl, 2004; Almgren et al., 2005; Engle and Ferstenberg, 2007). These models assume motivated empirically functional forms of price impact functions, one for every trader, which are then used to analyse market dynamics.

Thin market phenomena

Handling large orders through order break-up is but one difference between thin and competitive markets. Market thinness leads to a number of other empirical phenomena that are hard to reconcile with competitive equilibrium asset pricing models, such as CAPM of C-CAPM.

Pareto inefficiency

Because of the reduction of buying and selling orders in response to market power, traders do not fully diversify the idiosyncratic risk of their holdings. As a result, allocations are not efficient. This is optimal, since the benefits from diversification need to be balanced against the extra (with respect to the competitive, price-taking, setting) cost of price impact.

Response to liquidity shocks

As evidenced by the empirical literature, the exogenous shocks in asset supply, such as inclusions of new stocks into the S&P, weight changes in stock market indices, or forced liquidations, result in price overshooting. Even if the shock is preannounced, on the date of the actual event, the price drops below the new fundamental value to attain that value only in subsequent periods. This phenomenon, often referred to as price overshooting, cannot occur in the competitive model, where prices adjust to the new fundamental value immediately following the shock or its announcement. The overshooting effect is the equilibrium reaction of thin markets to anticipated and unanticipated shocks in asset supply. The overall observed price change can be decomposed into temporary and permanent components. The permanent effect, which occurs upon the announcement, represents the adjustment of the fundamental value that results from the changes in inventory. The effect is amplified by temporary price concessions demanded by liquidity providers to be willing to absorb the shock on the day of its occurrence whether or not the shock is anticipated. An alternative explanation of overshooting involves predatory trading (Brunnermeier and Pedersen, 2005): When a large trader needs to liquidate a portfolio quickly, other investors sell and subsequently buy back the asset. This strategy lowers the price at which they can

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
obtain the liquidated portfolio.

Excess volatility and volatility clustering

One of the consequences of price overshooting in thin markets is excess price volatility. That is, the presence of the price impact leads to excess return volatility and changes in volatility unrelated to changes in fundamentals. Since, in addition, the price impact varies over time, periods of high price impact feature high price volatility, thereby inducing volatility clustering.

Limits to arbitrage

Another novel feature of thin markets is the coexistence of anticipated price differentials and limits to arbitrage in equilibrium. According to the competitive theory of asset pricing, whenever there are anticipated price differentials, a trader can make infinite profit by taking unbounded positions. When a market is thin, however, price impact naturally limits the benefits from arbitrage for active traders and also reduces incentives to enter the market. Therefore, unlike in a competitive model, the profits from entering the market are bounded, and even small fixed entry costs may prevent outsiders from arbitraging the price differentials. These entry costs include explicit trading costs, such as transaction costs, but also the cost associated with learning the characteristics of the stocks. Empirically, it may take months for outside capital to bid prices back to their fundamental value (Mitchell, Pedersen and Pulvino, 2007).

Asset valuation

In the presence of price impact, the market value of a large block of shares no longer coincides with the cash value that could be obtained by liquidating the portfolio. To account for the difference, valuation specialists often apply a so-called blockage discount. A typical instance where blockage discounts are applied involves the transfer of a property in a case of divorce. It is in the interest of the divorcees to claim a large price impact (and blockage discount) which implies a large tax discount. The practical approach is based on the implementation shortfall (Perold, 1988), which measures the difference between the closing or arrival price and the final execution price.

Implications for market design

Market thinness has further prompted changes in market design towards automation of the trade execution. To ease competition through the trading cost of price impact, many exchanges have adopted an electronic trading system with posted orders (e.g. Nasdaq, NYSE, Euronext, and the stock exchanges in London, Toronto and Vancouver). In the presence of asymmetric information, market thinness can perversely reduce liquidity under continuous trading. Therefore, several markets have returned to more traditional trading systems.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- arbitrage
- countervailing power
- efficient markets hypothesis
- exchange
- liquidity constraints
- Pareto efficiency

Bibliography

Almgren, R. 2003. Optimal execution with non-linear impact functions and trading enhanced risk. *Applied Mathematical Finance* 10, 1–18.

Almgren, R. and Chriss, N. 2000. Optimal execution of portfolio transactions. *Journal of Risk* 3, 5–39.

Almgren, R., Thum, C., Hauptmann, E. and Li, H. 2005. Equity market impact. *Risk* 57–62.

Attari, M., Mello, A. and Ruckes, M. 2005. Arbitraging arbitrageurs. *Journal of Finance* 60(5), 2471–511.

Back, K. 1992. Insider trading in continuous time. *Review of Financial Studies* 5(3), 387–409.

Bertsimas, D. and Lo, A. 1998. Optimal control of execution costs. *Journal of Financial Markets* 1, 1–50.

Brunnermeier, M. and Pedersen, L. 2005. Predatory trading. *Journal of Finance* 4, 1825–63.

Carvajal, A. and Weretka, M. 2007. *State prices and arbitrage in thin financial markets*, University of Wisconsin-Madison.

Chan, L. and Lakonishok, J. 1993. Institutional traders and intraday stock price behavior. *Journal of Financial Economics* 33, 173–99.

Chan, L. and Lakonishok, J. 1995. The behavior of stock price around institutional trades. *Journal of Finance* 50, 1147–74.

DeMarzo, P.M. and Uroševi•, B. 2006. Ownership dynamics and asset pricing with a large shareholder. *Journal of Political Economy* 4, 145–74.

Dubil, R. 2002. Optimal liquidation of venture capital stakes. *Journal of Entrepreneurial Finance and*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Business Ventures 7(2), 56–81.

Easley, D. and O'Hara, M. 1987. Price, trade size, and information in securities markets. *Journal of Financial Economics* 19, 69–90.

Engle, R. and Ferstenberg, R. 2007. Execution risk. *Journal of Portfolio Management* 33, 34–45.

Foster, F.D. and Viswanathan, S. 1996. Strategic trading when agents forecast the forecasts of others. *Journal of Finance* 51(4), 1437–78.

Glosten, L. and Milgrom, P. 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 13, 71–100.

Grossman, S.J. and Miller, M.H. 1988. Liquidity and market structure. *Journal of Finance* 43(3), 617–37.

Ho, T. and Stoll, H.R. 1981. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics* 9, 47–73.

Holden, C.W. and Subrahmanyam, A. 1996. Risk aversion, liquidity, and endogenous short horizons. *Review of Financial Studies* 9(2), 691–722.

Huberman, G. and Stanzl, W. 2004. Price manipulation and quasi-arbitrage. *Econometrica* 74(4), 1247–76.

Keim, D. and Madhavan, A. 1995. The anatomy of the trading process: empirical evidence on the behavior of institutional traders. *Journal of Financial Economics* 37, 371–98.

Keim, D. and Madhavan, A. 1996. The upstairs markets for large-block transactions analyses. and measurement of price effects. *Review of Financial Studies* 9, 1–39.

Keim, D. and Madhavan, A. 1998. Execution costs and investment performance: an empirical analysis of institutional equity trades, Rodney L. White Center for Financial Research Working Paper, 09-95.

Kyle, A.S. 1985. Continuous auctions and insider trading. *Econometrica* 53, 1315–36.

Kyle, A.S. 1989. Informed speculation and imperfect competition. *Review of Economic Studies* 56, 517–56.

Mitchell, M., Pedersen, L.H. and Pulvino, T. 2007. Slow moving capital. *American Economic Review*, P&P, 97(2), 215–20.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Perold, A. 1988. The implementation shortfall: paper versus reality. *Journal of Portfolio Management* 14, 4–9.

Pritsker, M. 2005. Large investors: implications for equilibrium asset, returns, shock absorption, and liquidity, Finance and Economic Discussion Series 2005-36, Board of Governors of the Federal Reserve System.

Rostek, M. and Weretka, M. 2007. Frequent trading and price impact in thin markets, University of Wisconsin-Madison.

Rostek, M. and Weretka, M. 2008. Small double auctions, University of Wisconsin-Madison.

Subramanian, A. and Jarrow, R. 2001. The liquidity discount. *Mathematical Finance* 11, 447–74.

Uroševic, B. 2005. Moral hazard and dynamics of insider ownership stakes, Universitat Pompeu Fabra, Economics Working Papers, 787.

Vayanos, D. 1999. Strategic trading and welfare in a dynamic market. *Review of Economic Studies* 66, 219–54.

Vayanos, D. 2001. Strategic trading in a dynamic noisy market. *Journal of Finance* 56(1), 131–71.

Weretka, M. 2006. *Endogenous market power*, University of Wisconsin-Madison.

How to cite this article

Rostek, Marzena and Marek Weretka. "thin markets." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000249> doi:10.1057/9780230226203.1872

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

third world debt

François Bourguignon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Excessive dependence on foreign savings is a threat. But governments can lead by example by securing fiscal surpluses. As well, appropriate financial sector, tax and other microeconomic policies can help stimulate private domestic savings. Under these conditions, foreign borrowing can be a healthy complement to domestic savings.

Keywords

capital account liberalization; consumption smoothing; contract enforcement; debt crises; debt intolerance; debt overhang; debt relief; debt relief Laffer curve; dollarization; exchange rate policy; fixed exchange rates; floating exchange rates; heavily indebted poor countries; international reserves; liquidity constraint; solvency constraint; sovereign debt; Third World debt

Article

Many developing countries have encountered difficulties in managing their foreign debts in the 19th and 20th centuries. A number of crisis episodes are reviewed by Fishlow (1985). Some of them, such as the Baring crisis of 1890, threatened the stability of the international financial system. Others were confined more specifically to a single country with a dispersion of creditors on bonded debt, as was the recent case of the Argentine default of 2002. While there is a long history of ‘Third World’ debt accumulation and subsequent defaults, the frequency of debt crises in developing countries has increased dramatically since the 1982 Mexican debt crisis. The focus of this article is the episodes since the mid-1980s, and the economic literature that emerged to analyse these events.

Much of the attention of the international community on Third World debt during the 1980s and early 1990s was focused on middle-income countries. During the mid- to late 1990s, debt relief for highly indebted poor countries (HIPCs) increasingly occupied the attention of policymakers around the world, as debt relief became a cause célèbre for a number of international NGOs. An increasing share of overseas development assistance in the new millennium has been devoted to cancelling the official credits of bilateral and multilateral donors through the Highly Indebted Poor Country ‘HIPC’ Initiative.

Readers who want more detailed accounts should consult the volumes edited by Smith and Cuddington (1985), Sachs (1989) and Husain and Diwan (1989), as well as the books by Cohen (1991) and Cline (1995). On debt relief for low-income countries, a detailed discussion is provided by Birdsall and Williamson (2002), or one can consult the volume edited by Addison, Hansen and Tarp (2004). On recent analyses on the origins of debt crises as well as links to exchange rate policy, the reader is referred to Calvo and Reinhart (2002), Reinhart, Rogoff and Savastano (2003) and Eichengreen, Hausmann and Panizza (2003). Cline (1995, ch. 4) provides a critical review of the theory of sovereign borrowing.

General analytic framework

The neoclassical view of sovereign borrowing by developing countries is straightforward. Developing countries have lower capital stocks and resulting higher returns to capital than the high-income countries. These circumstances provide both borrowers and lenders with the incentive to engage in mutually beneficial debt transactions. The experience with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sovereign defaults over the last two centuries leads to other considerations. Debt contracts across international borders do not fall under a clear legal jurisdiction for enforcement, so the incentives for repayment on the part of the borrower must be understood.

Eaton and Gersovitz (1981) developed a model where developing countries borrow for consumption-smoothing purposes. Under certain circumstances, the desire for continued access to foreign credit for these purposes can be an adequate incentive for borrowers to pay, an incentive that is balanced against the one-off gains from defaulting. Creditors, understanding these incentives, will lend only up to the point at which the marginal costs and benefits of lending are equal. Clearly, this assumes that there is full information about these costs and benefits and coordinated action on the part of the creditors.

Cooper and Sachs (1985) provide another seminal analysis on foreign borrowing where the repayment question is divided into three components, or risks: illiquidity, insolvency and repudiation. The analysis looks at the optimal borrowing path for a debtor nation, taking into account constraints associated with the aforementioned three risks. For example, external solvency is limited by the present discounted value of future trade surpluses. If these surpluses happen to grow at a slower pace than the rate of interest, then the country faces a solvency constraint. The liquidity constraint may arise when, in a given year, the external trade surplus is insufficient to cover debt service, perhaps due to a short-run change in the terms of trade or other external factors. (Later analyses of solvency and liquidity would analyse explicitly an additional constraint: the internal transfer of resources from the domestic private sector to the government debtor. This issue is particularly troublesome in the context of devaluations to improve the trade balance – the external transfer constraint – which lead to a higher domestic currency tax bill that must be raised to effect the internal transfer from agents to the state; see Dornbusch, 1988.) Despite long-run solvency, regulatory issues (for example, limited exposure to a single borrower) or coordination problems among creditors may lead individual banks to refuse to extend loans to cover that debt service. Repudiation risk, similar to the Eaton and Gersovitz concept discussed above, is another reason for creditors to limit loans, even when a negotiated solution to repudiation is anticipated.

Indeed, another critical question from the theoretical literature is the point at which debt relief, or debt forgiveness, becomes mutually beneficial to debtors and creditors. Krugman (1989) provides a simple definition of debt overhang: the expected present value of the resource transfer available to service debt is less than the value of debt. This sounds generally like an insolvency condition, expressed in expected present value terms, and the ‘overhang’ is the degree of excess debt that cannot be repaid under current expected conditions. Incentive problems then arise for overcoming the debt overhang. First of all, the existence of the overhang implies that there is a high marginal tax rate on domestic investment, as much of the return to investment will have to be captured by the sovereign to close the debt overhang. This poses an incentive problem for the sovereign attempting to adjust taxes and exchange rates to increase the capacity to pay: much of the gain from adjustment will accrue to the creditors. Thus, from the debtor's point of view, there are strong incentives to lobby for a reduction in principal, or a reduction in interest rates, as the only effective means to restore growth and overcome the crisis.

Understanding the dilemma faced by the debtor, creditors may also benefit from debt reduction. At high enough levels of debt, the probability of outright default increases rapidly enough that the expected total repayment to creditors may begin to fall as the overall face value of debt increases. In other words, there is a debt relief Laffer curve (Krugman, 1989), as described in Figure 1. At very low levels of indebtedness, the sovereign is expected to pay in full, and there is no discount on secondary markets. This is represented by the 45 degree line in Figure 1: expected repayment equals the face value of the debt outstanding. As indebtedness levels increase, the probability of eventually facing a repayment problem increases. At point A, expected repayment is now below the face value of debt, and a secondary market discount would appear. However, at this stage it makes sense for creditors to continue lending – or even with ‘defensive’ lending in the face of an external shock – in that an increase in the stock of debt outstanding still leads to an increase in expected repayment. This is no longer the case beyond point C. At point B indebtedness levels have reached the point where additional lending reduces expected repayment. Debt forgiveness, reducing the face value of debt outstanding, would then actually lead to an increase in the expected value of repayment. Yet this raises an issue of coordination among creditors since the value to the collective creditor community of reducing the debt by one dollar is higher than the average price (expected value) which represents the value of that operation to the individual investor (Bulow and Rogoff, 1988). As a result some agreement must be reached among creditors on how to forgive debt, including ‘burden-sharing’ arrangements.

Figure 1

The debt relief Laffer curve

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

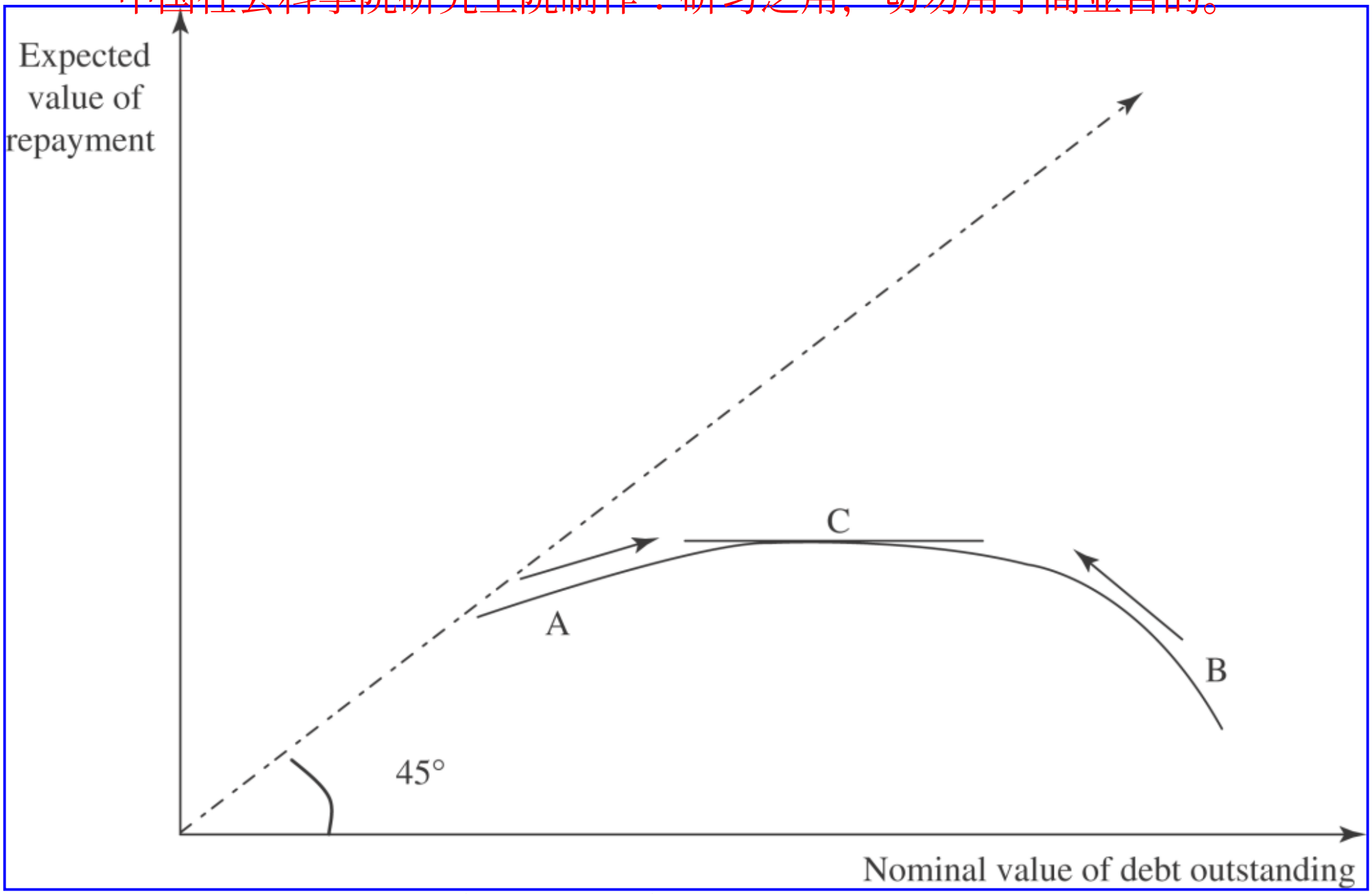

In summary, when debt becomes large enough, the prospects for simply growing out of the problem are dim. ‘Debt overhang’ can tax investment and limit debtors’ incentives to engage in policy reforms to restore solvency, or even limit their incentive to continue to honour their obligations. As secondary market prices fall, the price for purchasing a claim on a debtor deviates greatly from the price a debtor has to pay to retire a claim via regularly scheduled debt service. An alternative is to use international reserves, if available, to purchase debt on the secondary market, thus lowering the cost of retiring the claim. Bulow and Rogoff (1988) point out, however, that the price paid on the secondary market is still too high, in that it represents the average price of debt, and ignores the marginal increase in expected repayments.

This analytical framework helps one understand the key issues facing both policy makers and creditors who are trying to realize mutually beneficial exits from a debt crisis. One overriding complication that emerges in a real crisis is uncertainty about the values of the parameters that define whether a country is facing a liquidity or a solvency problem, or if it

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

is simply unwilling to pay. Krugman (1989) and others have asserted that, in an uncertain world, the distinction between liquidity and solvency is 'misleading': liquidity problems arise precisely because there is some reasonable probability that the country is insolvent. The brief recounting below of the history of recent debt crises reveals the difficulty in determining these values in the midst of a crisis. This uncertainty leads to divergent views on the best solution to a particular crisis episode.

This analytical framework continues to be useful. As noted below, the debt crises of the 1990s inspired a broader literature that examines how capital flow reversals, the currency structure of bank or corporate balance sheets and exchange rate policies all interact in causing the debt crises of this later period.

Evolution of external debt: 1980s to the present

Table 1 describes the evolution of the external debt of developing countries over the 1980–2003 period, in constant 2003 dollars. Two noteworthy trends are important for the discussion of the evolution of debt crisis episodes. First of all, we see a shift towards debt owed to 'public creditors' – that is, official bilateral and multilateral institutions – during the mid- to late 1980s, as the Baker and Brady Plans responded to the debt crises of the early 1980s. Over the same period, the share owed by the private sector in developing countries declined, as developing country governments assumed the liabilities of the private sector following the crisis. Later, during the early 1990s, we see an acceleration in the growth of the total debt stock, led by a greater role of both lending by private sector creditors and borrowing by private sector borrowers. This is not surprising, given that many developing countries pursued capital account liberalization at the start of the 1990s.

Total developing country external debt, 1980–2003

	1980	1985	1990	1995	2000	2003
*Total external debt**	1,093.1	1,421.4	1,747.1	2,379.7	2,436.6	2,553.0
•Short term	269.6	266.6	310.8	490.0	401.2	508.6
•Medium and long term	823.4	1154.8	1436.3	1889.8	2035.4	2044.4
•Owed by public sector	719.8	1085.0	1405.4	1698.6	1517.4	1556.2
•Owed by private sector	373.3	336.4	341.7	681.3	919.4	997.1
•Owed to public creditors	342.9	510.1	803.4	1046.8	896.4	933.0
•Owed to private creditors	750.2	911.3	943.8	1333.1	1540.4	1620.3
Memo items						
•Share owed to private creditors	0.69	0.64	0.54	0.56	0.63	0.63
•Share owed by private sector	0.34	0.24	0.20	0.29	0.38	0.39
•Total external debt/GDP	0.18	0.28	0.32	0.38	0.38	0.37
•Total external debt/exports (goods and services)	0.90	1.56	1.53	1.55	1.29	1.11

*Constant 2003 dollars, billions

Source: World Bank (2005a); author's calculations.

Debt crises of the 1980s

The five paragraphs immediately following have been taken with some modification from Kenen (1992).

Commercial banks and other private institutions did not lend extensively to developing countries in the 1950s and 1960s. The build-up of debt that led to the crisis of the 1980s began in 1974, after the increase in the price of oil which followed the Arab–Israel war in October 1973. Some of the oil-exporting countries ran huge current account surpluses and deposited the proceeds with foreign commercial banks, and many oil-importing countries ran huge deficits. The imbalance was heavily concentrated on the developing countries because the industrial countries moved into recession in 1974–5, reducing their total imports and focusing the global 'oil deficit' on the oil-importing developing countries.

The economic environment of the 1970s was very conducive to large borrowing. Nominal interest rates were low, and real interest rates were negative. Many developing countries were posting higher growth rates of output and exports than in earlier decades, which held down the ratios of debt to output and of debt-service payments to exports. As noted in the analytical framework above, under these conditions debt accumulation is bound to appear sustainable.

Furthermore, the banks had protected themselves from two of the three risks facing them. First, the deposits of the oil-exporting countries were denominated mainly in US dollars, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
currency in which they were paid for their oil, and most of the banks' own loans to developing countries were likewise dollar-dominated, so the banks were protected from exchange rate risks. Second, the interest rates charged on most of the loans were based on the London Inter-Bank Offered Rate (LIBOR), the interest rate paid on most of the deposits, and were adjusted frequently to maintain the spread over LIBOR, so the banks were protected from interest rate risks.

Yet the way that banks protected themselves from interest rate risk increased their exposure to default risk by linking the debtors' obligations to very volatile interest rates. Furthermore, the link to LIBOR, combined with the debtors' dependence on export earnings from products with cyclically sensitive prices, made it very likely that the debtor countries would prosper or suffer together, reducing the protection usually afforded by diversification. Finally, the risk of each bank's exposure to a particular country depended on the country's total debt, which depended in turn on the volume of loans made by other banks and on the amount of debt owned by all of the country's borrowers. In brief, there were externalities all over the place, but no one was paying attention to them.

The country which set off the crisis was just becoming an oil exporter rather than an oil importer. Mexico had borrowed heavily in the expectation of big oil exports, and much of its external debt bore floating interest rates. When interest rates rose sharply in the early 1980s, after the tightening of monetary policy in the United States, Mexico's debt-service payments soared. Together with capital flight, produced by expectations of devaluation, the increase in interest payments depleted Mexico's foreign-exchange reserves, and Mexico had to suspend those payments in August 1982.

The crisis was caused by a combination of the externalities on the creditor side, discussed above, and lax fiscal policies on the debtor side. In addition, countries with more closed trade regimes and overvalued exchange rates confronted the greatest difficulties in adjusting to the changes in the external environment.

By 1985, 15 countries were identified as highly indebted and requiring coordinated assistance from the international community, and these 15 were later extended to 17. In terms of magnitudes, the average debt–export ratio of this group peaked at 384 per cent in 1986 (Cline, 1995).

On the side of the creditors, there were serious concerns about a potential systemic crisis. Exposure rates reached remarkably high levels during the lending boom to developing countries. Sachs (1989) reports that at the end of 1982 the exposures of nine major banks to developing countries had reached a staggering 289 per cent of bank capital, with much of this exposure concentrated in several Latin American countries. Prudential regulations on the concentration of exposure to individual borrowers did not avert this problem, since banks listed individual public agencies or state enterprises within each country as separate borrowers.

Policy response: the Baker and Brady plans

The initial response to the debt distress of the early 1980s was to coordinate efforts for new lending. Creditors largely comprised a group of large international banks. Payment arrears were viewed as primarily a problem of illiquidity. International economic conditions had turned against a number of heavily indebted emerging markets: sharply rising interest rates, declining prices for a number of commodities amidst a global economic recession. Coordinated efforts to extend new financing would allow these countries to honour their obligations until the external situation improved and structural reforms implemented. Projection models (for example, Cline, 1985) suggested that, with feasible policy reforms and improving external conditions, most debtors could make good use of the new money and emerge as strong solvent sovereigns within a reasonable time period. The IMF would coordinate new lending.

These coordinated efforts were combined with targets for banks' and international financial institutions contributions under the 1985 'Baker Plan', named after James Baker, the US Treasury Secretary of that time. Private banks would provide $20 billion in loans, to be matched by a similar (gross) amount by the multilateral banks (Cline, 1995).

Under the Baker Plan, substantial lending occurred, but less than the original commitments, partially due to the absence of IMF agreements in some defaulting countries. Some of the new lending was used for buybacks of existing debt, policy conditions were not always met on the recipient side and, in many countries, arrears on new loans began to accumulate. Meanwhile, by the late 1980s many creditor banks had been able to accumulate provisions against their losses. As a result, debt reduction or forgiveness would not cause the same degree of stress on the financial system as had been feared during the early years of the crisis. By 1988, a number of banks were used a 'menu approach', accepting concessional exit bonds or contributing to the new lending strategy. Also, by the late 1980s, secondary market prices had declined enough that banks that chose to exit via secondary markets were accepting large losses. Any coordinated reduction plan that would offer a reduction less than the secondary market discount would be viewed as favourable. A number of countries did emerge from the period of the 'new money' strategy without resorting to debt reduction: South Korea, Indonesia, and Turkey, are several prominent cases. Chile was a notable example from the Latin American region.

Under the leadership of Nicholas Brady, the new US Secretary of the Treasury, the 'Brady Plan' (1989–94) was initiated with a focus on 'voluntary' debt reduction; however, still suggesting a menu of options including rescheduling for those banks/country cases where it might be of interest. The target for debt reduction was $70 billion. US government bonds were offered as collateral for new bonds used for debt reduction purposes, thus further attracting the creditors. Eventually, about $60 billion of debt was forgiven by 1994. In general, debt reduction averaged about 30 per cent for private bank debts restructured; however, since official debt was about half the debt outstanding, total debt reduction amounted to about 15 per cent of outstanding debt (Cline, 1995). Renewed economic growth in the early 1990s, accompanied by renewed market access, seemed to indicate that the programme was a resounding success.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Debt crises of the 1990s

By the end of the Brady Plan period, however, a new round of debt distress had appeared appear in emerging markets. During the late 1980s and early 1990s, many developing countries removed restrictions to capital inflows and liberalized their domestic financial sectors. As noted above, borrowing by the private sector led to the accumulation of debt, as did borrowing from private sector creditors. Another important change that emerged during the 1990s was the shift to bond issues, as opposed to commercial bank loans. This change obviously led to a greater dispersion and variety of creditors.

The new cycle of debt distress and crises again started with Mexico. In March 1994 an initial run on reserves occurred shortly following the assassination of the leading political candidate and a reversal of capital inflows which had been quite strong for a number of years. Mexico had been running large current account deficits during these years, in the context of a crawling band exchange rate policy, and international interest rates were also rising. In the aftermath of this first run on reserves, the central bank expanded domestic credit and it offered long-term bondholders new short-term, dollar-denominated *Tesobonos*. There was a fear of allowing interest rates to rise due to weaknesses in the banking system (Lustig, 2001). By December, during the last month of the Salinas administration, the government was encountering increasing difficulty in rolling over *Tesobonos*. An attempted ‘controlled devaluation’ of 15 per cent was followed by a renewed run on reserves. The monetary authorities were left with no option but to allow the currency to float.

Some authors (for example, Dornbusch and Werner, 1994) attribute the Mexican crisis to a traditional problem of an overvalued fixed exchange rate, with persistent current account imbalances and the corresponding accumulation of debt. With the increasing indebtedness, shorter terms and higher interest rates were demanded by the market, further increasing the likelihood of default. Devaluation would also increase indebtedness ratios relative to domestic resources. Others (Calvo and Mendoza, 1996) place greater emphasis on stock imbalances – namely, money balances, short-term debt and gross reserves – and sharp adjustment driven by global capital markets with ‘herding’ behaviour on the part of foreign bondholders. From this point of view, Mexico suffered ‘cruel punishment’ for the ‘petty crime’ of attempting a controlled devaluation to correct a modest balance of payments disequilibrium. This debate again reflected some of the difficulties in determining whether the country was insolvent or merely facing a liquidity problem that could be resolved by some coordination among creditors, as discussed in the theoretical framework above.

Indeed, the ‘punishment’ was heavy. GDP declined by more than six per cent in 1995, real wages declined sharply in the aftermath of the devaluation, and poverty rates increased by about seven or eight percentage points. Half a decade would pass before poverty rates returned to pre-crisis levels.

In early 1995, an international support programme to finance the gap in Mexico's balance of payments was arranged of the order of some $50 billion in commitments from the IMF and bilateral sources – mainly the United States. The Fund programme of about $18 billion was the largest lending programme in IMF history at that time. The floating of the currency, along with improved trade possibilities due to the implementation of NAFTA, and these official financing flows allowed Mexico to return to growth averaging about 5.5 per cent over the 1996–98 period.

Possibly through contagion from the Mexico crisis, but also for domestic reasons, Argentina suffered a sharp withdrawal of capital and loss of international reserves in early 1995. There, too, an international support package was arranged, and robust growth was restored in 1996.

Mexico's crisis was then followed by several other financial crises in the late 1990s and into the new millennium, all involving either outright default or near default on substantial external liabilities. In East Asia, even South Korea – considered to be virtually a ‘developed’ country – was unable to avoid a substantial crisis based on the accumulation of short-term external liabilities of the private corporate and banking sector rather than government debt. Starting in Thailand, in the summer of 1997 the Asian financial crisis quickly spread by the end of the year to affect heavily Indonesia, Malaysia, Korea and the Philippines. Japan, Taiwan and Singapore also suffered through substantial problems in their financial sectors.

Each country affected by the East Asian crisis had particular circumstances leading up to the crisis; however, the most heavily hit countries had a number of common problems. Fixed exchange rates and either implicit or explicit guarantees by the government motivated private companies and banks to inadequately account for the risk involved in borrowing heavily in foreign currency. (The common features mentioned in this section are drawn from Kawai, Newfarmer and Schmukler, 2001. Goldstein, 1998, also emphasizes a combination of factors, including external imbalance and financial sector fragility due to mismatches.) There were widening current account deficits financed by short-term capital flows. Relatively unregulated financial systems conducted a credit boom with investment in low-return activities, along with substantial currency and maturity mismatches in their balance sheets. Corporations borrowed heavily, leading to exposure to interest and exchange rate shocks. Political uncertainty led to doubts about the continuation of existing exchange rate and other policies.

As in the case of Mexico, economists differ in assigning a direct cause of the crisis. There are those that emphasized fundamentals – along the lines of the common characteristics across countries mentioned above. Others (for example, Radelet and Sachs, 1998) emphasize that the problem was one of illiquidity rather than insolvency, and foreign investor panic led to a withdrawal of the willingness to provide liquidity. Kaminsky and Schmukler (1999) provide empirical evidence to support the view that investors’ reactions to news events – as measured by changes in the stock market – can be characterized as ‘herding behaviour’.

The impact of the crisis was dramatic. The Thai and Indonesian economies shrank by 10.5 and 13.1 per cent respectively, in real terms. The Korean and Malaysian economies shrank by 6.9 and 7.4 per cent, respectively.

In response to the crisis, the IMF was working on numerous fronts to provide financial support to cover debt-servicing needs. Programmes were established in the main countries

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
affected, and the World Bank also provided substantial fast-disbursing funds. Korea returned to borrower status with the World Bank, securing the largest individual loans in World Bank history, despite previous ‘graduation’ from borrower status. Overall, including both multilateral and bilateral donors, $57 billion was committed for Korea, $35 billion for Indonesia and $17 billion for Thailand over the 1997–98 period (Radelet and Sachs, 1998).

In 1998, Russia was the next emerging market economy to encounter difficulties in servicing its external debt (Kharas, Pinto and Ulatov, 2001). In August, 1998, the authorities announced a 90-day moratorium on external debt, as well as their intention to restructure all obligations due through the end of 1999. Within three weeks, the central bank floated the ruble with ruble–dollar rate subsequently tripling. The Russian authorities had attempted to sustain the exchange rate to secure inflation and to avoid the economic collapses of Asian countries that had devalued in the previous year. One month prior to the collapse, the authorities had arranged for a $22.6 billion external financing package led by the IMF. The IMF programme included measures to restore long-term fiscal balance. The short-term financing was also expected to restore foreign investor confidence and assist in a voluntary swap of short-term government (‘GKO’) bonds for longer term Eurobonds.

The impact of the crisis on Russia was strong, but fairly short-lived. GDP declined by 5.3 per cent in 1998, but fully recovered in 1999. Headcount poverty (using the two dollar a day definition) rose from 23 per cent in 1996 to 36 per cent in 1998. By 2000 this measure of poverty had fallen to 24 per cent again (World Bank, 2005b).

It was feared at that time that the Russian crisis would contaminate other emerging countries. In effect, the lack of confidence of investors resulted in capital outflows in several countries. Combined with domestic events, the situation became somewhat serious in Brazil early in 1999. But the adoption of a floating exchange rate as well as tightening fiscal measures re-established confidence and reversed capital outflows.

Argentina, too, was contaminated by the Russian default. Its sovereign debt experienced a sharp rise in spreads, rising to nearly 15 per cent over equivalent US bonds in September 1998. By the end of 1998, bond spreads had declined to below ten per cent; however, the country had slipped into a period of stagnation and eventual recession that lasted through mid-2001. A combination of lingering doubts among investors about the hard currency peg (a quasi-currency board arrangement), national elections in 1999, and falling international commodity prices contributed to the malaise. By mid-2001 confidence had reached a low point, Argentines were pulling (mostly dollarized) deposits out of banks and the central bank was losing reserves at a rapid pace. In the middle of the year, a ‘mega swap’ was conducted to prolong maturities without any reduction in the stock outstanding. This ‘market-based’ approach failed to stem the loss of reserves. New resources of approximately $6 billion by the IMF in August also failed to make a difference. In December, the government declared a freeze on deposits in a last-ditch effort to stem the haemorrhaging. By the end of the month, facing escalating street protests, the president would resign. In early January 2002 an interim president declared unilateral suspension of payments on foreign public debt. The exchange rate would soar from the fixed 1:1 with the dollar (which had dated back to 1991) to over three ARS per dollar.

This latest crisis was particularly dramatic. The population had treated the 1:1 fixed exchange rate with the dollar as a national institution. The vast majority of deposits, loans and business contracts were set in dollars. A combination of terms of trade shocks and declining investors’ confidence in politicians’ willingness or ability to sustain the system would be accompanied by a painful, gradual, deflationary adjustment of the real exchange rate leading to the eventual collapse in early 2002.

Resolution of the debt default was also particularly problematic. There were many series of bond issues in various international legal jurisdictions and numerous issues that had been passed on by underwriters to the retail level. Relations between the government and multilateral lenders also reached a stand-off, with little or no ‘fresh money’ provided, and brief periods of suspension of payments. For private creditors, the government took a hard-line position: an offer of approximately 25 cents on the dollar, in present value terms – a level lower than the value implied by secondary markets at the time. Some 100 billion dollars were eligible to be exchanged for a menu of bonds, both par and discount and also in either domestic or foreign currency. By the time of the actual transaction international interest rates had declined, so that the final offer presented a present value of just over 30 cents to the dollar of face value, a level close to prevailing secondary market discounts. This transaction represented the sharpest discounts on sovereign debt exchange in history. Just over three-fourths of the defaulted bonds were eventually exchanged, leaving a fairly substantial community of ‘holdouts’.

GDP declined by an average of 2.9 per cent per year over the 1999–2001 period. In 2002 the economy shrank by nearly 11 per cent. Headcount poverty rates (national poverty line) more than doubled to around 50 per cent in 2002. Thanks to a robust recovery, GDP in real local currency terms would finally return to 1998 levels in mid-2005. The poverty rate would decline to around just under 40 per cent.

Debt and economic policies: analysis of the crises of the 1990s

As noted in the analytical framework above, indebtedness strategies cannot be viewed in isolation from other economic policies. The ability to pay is determined by the growth of exports and the effectiveness of the debtors’ tax system to secure resources for debt repayment. Clearly export growth is affected by both microeconomic and macroeconomic policies, and in particular, the exchange rate regime. Vulnerabilities in the financial system, due to lax regulation and supervision, can trigger bank withdrawals, capital flight, and exchange rate movements that dramatically shift a country from ‘apparent’ solvency to immediate insolvency.

A new nomenclature has developed in the recent literature. Many developing countries commit the ‘original sin’ (Eichengreen, Hausmann and Panizza, 2003) of borrowing in foreign currency. Dollar-denominated liabilities, or even broader dollarization of the domestic financial systems, result in a ‘fear to float’ (Calvo and Reinhart, 2002) the nominal exchange rate. Countries with fixed exchange rate regimes supported by large capital inflows face the risk of ‘sudden stop’ (Calvo, Izquierdo and Talvi, 2003) of these inflows, leading to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

distressed debt restructuring or default. The prevalence of debt crises in emerging economies with debt to GDP ratios below OECD averages has led some analysts to diagnose these countries as suffering from ‘debt intolerance’ (Reinhart, Rogoff and Savastano, 2003).

HIPCs

In parallel to these events involving countries borrowing from private markets, there was growing concern that the poor economic performance of many low-income countries was due to excessive indebtedness. A fundamental difference for these countries was that they had virtually no access to private markets for lending: debt stocks had been accumulated with official bilateral and multilateral lenders. There is no rollover risk and the behaviour of private lenders is not an issue. On the other hand, a number of economists in the 1980s began to emphasize debt overhang as a deterrent to growth in heavily indebted poor countries, or ‘HIPCs’, as they would come to be known. Others, supported by prominent international NGOs, emphasized that debt service detracted from the resources available for poor countries’ governments to provide basic services for their people.

Debt relief for poor countries dates back to the establishment of the ‘Paris Club’ in the 1950s – the coordinating institution for bilateral creditors to agree on debt relief. Since the early 1980s, however, multilateral efforts to provide debt relief have intensified. Easterly (1999) provides a review of debt relief for poor countries over the previous two decades. The 1977–79 meetings of the United Nations Conference on Trade and Development led to debt write-offs of some $6 billion for 45 low-income countries. In the early and mid-1980s, World Bank reports on Africa increasingly emphasized the debt servicing difficulties of low-income countries. By the late 1980s, debt relief for these countries had entered the agenda of the annual G7 summits. The summit of 1988 in Toronto established a menu of debt relief terms to be offered to these countries. The multilaterals established special programmes as well: the World Bank's Special Program of Assistance (SPA) to low-income Africa and the IMF's Enhanced Structural Adjustment Facility (ESAF). These facilities would provide fast-disbursing funds to assist in debt repayments.

During the 1990s low-income country debt relief gained further momentum. On the global political front, there was growing support for debt relief – including from religious leaders and high-profile figures in popular culture. The year 2000 marked a symbolic target date for debt relief and ‘Jubilee 2000’ political movements to promote debt relief spread across some 60 countries. According to Jubilee Research, a global petition for debt relief collected 24 million signatures. The ‘Toronto terms’ debt relief mentioned above was followed by a series of expansions of debt relief during international summits in the early 1990s. In 1996, the World Bank and the IMF established the Highly Indebted Poor Country (HIPC) Initiative, and in the same year, the Paris Club of bilateral creditors committed to 80 per cent debt relief in net present value terms (Easterly, 1999). In 1999, the HIPC initiative was ‘enhanced’ to provide faster and broader debt relief and to link debt relief with the elaboration of poverty reduction strategies (as countries were asked to prepare Poverty Reduction Strategy Papers, or PRSPs). These papers would be prepared in collaboration with, and with the support of, the World Bank and IMF. The IMF initiated a new lending facility called the Poverty Reduction and Growth Facility (PRGF) to provide interim financing to the HIPCs as they prepared these papers.

The technical criterion for a low-income country to be considered ‘highly indebted’ is either a net present value (NPV) of debt greater than 150 per cent of exports or an NPV of debt greater than 250 per cent of government revenues. HIPCs must pass through a process of several steps before receiving debt relief. A ‘Decision Point’ is reached when a country displays a minimum degree of macroeconomic stability, has cleared any arrears in debt service, and has prepared an Interim PRSP. The World Bank and IMF also prepare a debt sustainability analysis at this stage. At the decision point, an initial level of debt relief is granted ‘conditionally’ upon successful passage to the ‘Completion Point’ stage.

During the ‘interim’ period between the decision point and completion point, the IMF provides financing under a PRGF and the World Bank may provide policy-based credits through the International Development Association (IDA) as well. If a country has achieved satisfactory performance under the agreed measures of a PRGF and the PRSP has been implemented for at least one year, then the country may reach the Completion Point. At this point, debt relief is granted ‘irrevocably’. Additional debt relief – ‘topping off’ – may also be provided by multilateral and bilateral creditors.

As of mid-March 2005, 27 countries had reached or surpassed the Decision Point, of which 15 had reached the Completion Point, representing a total of approximately $32 billion (in NPV terms at the year of the Decision Point) of debt relief committed by multilateral and bilateral creditors (IDA–IMF, 2005). Another 11 countries were at the pre-Decision Point stage. In NPV terms of 2004, the total expected cost is $58 billion. For the 27 Decision Point (or beyond) countries, debt stocks were reduced by two-thirds (see World Bank, 2007).

The HIPC Initiative has generated both supporters and critics. A number of NGOs have complained that the process is too slow and onerous. Easterly (1999) emphasizes that debt relief is likely to be followed by a re-accumulation of debt unless countries change their long-run savings preferences. There have been complaints that the targets are too ‘ad hoc’ and the ‘criteria for debt relief too narrow’ (Addison, Hansen and Tarp, 2004). Others have stated that the programme has generated false expectations in that the real resource transfer involved is limited because these debts would have been ‘rolled over’ indefinitely (Cohen, 2001). On the other hand, there have been proposals to expand debt relief, with deeper relief for current HIPCs: broadening the scope to include more low-income countries, safeguarding countries from returning to unsustainable debt levels via a new contingency facility, and shifting the focus of future development assistance from loans to grants and with more multilateral coordination of these aid flows (Birdsall and Williamson, 2002). At the annual meetings of the IMF and World Bank in 2005, donor countries committed to additional debt relief, mostly via covering capital losses to the multilateral institutions if they cancel their own loans to HIPCs.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Looking ahead: the ‘new financial architecture’

Much of the focus of a new system for international debt flows is on how to make crisis resolution less costly for both debtors and creditors. The IMF proposed a new Sovereign Debt Restructuring Mechanism (SDRM) (Krueger, 2002), that would attempt to capture some of the features of corporate debt restructuring in developed countries. The main features of this proposal were fourfold: (*a*) ‘majority restructuring’ – a mechanism whereby a qualified majority of bondholders could commit the minority to a restructuring agreement; (*b*) ‘stay on creditor enforcement’ – a period during which creditors would not be able to pursue litigation to receive payment while attempts are made to reach a majority restructuring agreement; (*c*) protection of creditor interests – guarantees that debtors will not make payments to ‘non-priority’ creditors and that appropriate policies are pursued by the debtor so as not to reduce asset values (an IMF programme might be the means of supporting the latter); and (*d*) priority financing – a mechanism to reach agreement on ‘new money’ to facilitate the debt restructuring process, with new money receiving ‘senior status’ relative to old debt.

In principle, the SDRM addresses directly many of the incentive and coordination problems discussed above. To date, little progress has been made in implementing the proposal, with perhaps the exception of the first clause. Collective action clauses allow for a pre-defined majority or super-majority of bondholders to commit to a debt restructured with a distressed debtor. Collective action clauses are already common in many debt contracts issued under British law; however, this spontaneous market response does not address the aggregation issue. The clauses are attached and specified for individual bond issues, and thus they would not assure that all privately held debt could come under a single rule. Some other form of international agreement or institution would be needed to solve the aggregation issue.

Lessons from the crises: domestic debt and fiscal management

A combination of accumulating short-term external debt and heavily managed, or fixed, exchange rates can be fatal, leading to debt distress and subsequent economic strife. Many governments in developing countries are, as of the early 21st century, rebalancing their balance sheets towards debt denominated in domestic currency. By de-linking public debt stocks from the exchange rate, countries are gradually overcoming their ‘fear of floating’ the exchange rate. With more competitive exchange rates, a number of previous defaulters are also running balance of payments surpluses and improving their net foreign asset positions. All of these changes are occurring gradually, leaving them exposed to sudden shocks in the external environment.

Fiscal management remains an issue, however, in many of these countries, and it can limit growth prospects. The shift to borrowing locally crowds out domestic private borrowers, especially in those cases where domestic financial systems are still small following collapses during previous debt crisis episodes. In addition, a number of countries still have domestic financial systems with assets and liabilities denominated in foreign currency: another motivation for ‘fear of floating’.

In the end, an excessive dependence on foreign savings is a threat, as taught to us by the fundamentals of the intertemporal external balance constraints discussed earlier. In fact, many analysts have turned their attention to global imbalances involving the persistent current account deficits of the Unites States as a potential source of future disturbances that could affect those developing countries that are gradually trying to unwind their net external liabilities (see, for example, Goldstein, 2005). Clearly, governments can lead by example by securing fiscal surpluses (Gill and Pinto, 2005, conclude from their review that the first major lesson of the debt crises is ‘that paying attention to the government's intertemporal budget constraint . . is vital’). As well, appropriate financial sector, tax and other microeconomic policies can help stimulate private domestic savings (Cohen, 1992, concludes that a shift up in the capital stock needs to be accompanied by a level of saving consistent with this larger capital stock in order to sustain growth; he notes that the recipient country must also have an unusual combination: a relatively high endowment of human capital and relatively low physical capital). Under these conditions, foreign borrowing can be a healthy complement to domestic savings.

See Also

- financial structure and economic development

Bibliography

Addison, T., Hansen, H. and Tarp, F. 2004. *Debt Relief for Poor Countries*. Basingstoke: Palgrave Macmillan.

Birdsall, N. and Williamson, J. 2002. *Delivering on Debt Relief: From IMF Gold to a New Aid Architecture*. Washington, DC: Center for Global Development and the Institute for International Economics.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bulow, J. and Rogoff, K. 1988. The buyback boondoggle. *Brookings Papers on Economic Activity* 1988(2), 675–98.

Calvo, G.A. and Mendoza, E.G. 1996. Petty crime and cruel punishment: lessons from the Mexican debacle. *American Economic Review* 86, 170–5.

Calvo, G.A. and Reinhart, C. 2002. Fear of floating. *Quarterly Journal of Economics* 117, 379–408.

Calvo, G.A., Izquierdo, A. and Talvi, E. 2003. Sudden stops, the real exchange rate, and fiscal sustainability: Argentina's lessons. Working Paper No. 9828. Washington, DC: NBER.

Cline, W.R. 1985. International debt: from crisis to recovery. *American Economic Review* 75, 190–8.

Cline, W.R. 1995. *International Debt Reexamined*. Washington, DC: Institute for International Economics.

Cohen, D. 1991. *Private Lending to Sovereign States: A Theoretical Autopsy*. Cambridge, MA: MIT Press.

Cohen, D. 1992. The debt crisis: a post mortem. In *NBER Macroeconomics Annual*, vol. 7, ed. O. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Cohen, D. 2001. The HIPC initiative: true and false promises. *International Finance* 4, 363–80.

Cooper, R.N. and Sachs, J.D. 1985. Borrowing abroad: the debtor's perspective. In *International Debt and the Developing Countries*, ed. G.W. Smith and J.T. Cuddington. Washington, DC: World Bank.

Dornbusch, R. and Werner, A. 1994. Mexico: stabilization, reform and no growth. *Brookings Papers on Economic Activity* 1994(1), 253–315.

Dornbusch, R. 1988. Our LDC debts. In *The United States in the World Economy*, ed. M. Feldstein. Chicago: University of Chicago Press.

Easterly, W. 1999. How did highly indebted poor countries become highly indebted? Reviewing two decades of debt relief. Policy Research Working Paper No. 2225, World Bank.

Eaton, J. and Gersovitz, M. 1981. Debt with potential repudiation: theoretical and empirical analysis. *Review of Economic Studies* 48, 289–309.

Eichengreen, B., Hausmann, R. and Panizza, U. 2003. Currency mismatches, debt intolerance and original sin: why they are not the same and why it matters. Working Paper No. 10036. Cambridge, MA: NBER.

Fishlow, A. 1985. Lessons from the past: capital markets during the 19th century and the interwar period. *International Organization* 39, 383–439.

Gill, I. and Pinto, B. 2005. Public debt in developing countries: has the market-based model worked? Policy Research Working Paper No. 3674, World Bank.

Goldstein, M. 1998. *The Asian Financial Crisis: Causes, Crises and Systemic Implications*. Washington, DC: Institute for International Economics.

Goldstein, M. 2005. What might the next emerging-market financial crisis look like? Working Paper No. WP 05-7, Institute for International Economics.

Jubilee Research. Online. Available at http://www.jubileeresearch.org/about/about.htm, accessed 11 April 2007.

Husain, I. and Diwan, I. 1989. *Dealing with the Debt Crisis*. Washington, DC: World Bank.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

International Development Association and International Monetary Fund. 2005. Heavily Indebted Poor Countries (HIPC) Initiative – statistical update. 4 April 2005. Online. Available at http://siteresources.worldbank.org/INTDEBTDEPT/ProgressReports/20446696/HIPCStatUpdate200504042.pdf, accessed 11 April 2007.

Kaminsky, G.L. and Schmukler, S.L. 1999. What triggers market jitters? A chronicle of the Asian crisis. Policy Research Working Paper No. 2094, World Bank.

Kawai, M., Newfarmer, R. and Schmukler, S. 2001. Crisis and contagion in east Asia: nine lessons. Policy Research Working Paper No. 2610, World Bank.

Kenen, P. 1992. Third World debt. In *The New Palgrave Dictionary of Money and Finance*, ed. P. Newman, M. Milgate and J. Eatwell. London: Macmillan.

Kharas, H., Pinto, B. and Ulatov, S. 2001. An analysis of Russia's 1998 meltdown: fundamentals and market signals. *Brookings Papers on Economic Activity* 2001(1), 1–67.

Krueger, A.O. 2002. *A New Approach to Sovereign Debt Restructuring*. Washington, DC: International Monetary Fund.

Krugman, P. 1989. Market-based debt-reduction schemes. In *Analytics of International Debt*, ed. J. Frankel. Washington, DC: International Monetary Fund.

Lustig, N. 2001. Life is not easy: Mexico's quest for stability and growth. *Journal of Economic Perspectives* 15(1), 85–106.

Radelet, S. and Sachs, J. 1998. The East Asian financial crisis: diagnosis, remedies, prospects. *Brookings Papers on Economic Activity* 1998(1), 1–74.

Reinhart, C., Rogoff, K. and Savastano, M. 2003. Debt intolerance. Working Paper No. 9908. Washington, DC: NBER.

Sachs, J. 1989. *Developing Country Debt and Economic Performance, Volume I: The International Financial System*. Chicago: NBER and University of Chicago Press.

Smith, G.W. and Cuddington, J.T. 1985. International borrowing and lending: what have we learned from theory and experience? In *International Debt and the Developing Countries*, ed. G.W. Smith and J.T. Cuddington. Washington, DC: World Bank.

World Bank. 2005a. *Global Development Finance*. Washington DC: World Bank.

World Bank. 2005b. *World Development Indicators 2005*. Washington DC: World Bank.

World Bank. 2007. Debt issues. Online. Available at http:// www.worldbank.org/debt, accessed 11 April 2007.

How to cite this article

Bourguignon, François. "third world debt." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000048> doi:10.1057/9780230226203.1699

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Thompson, Thomas Perronet (1783 – 1869)

Murray Milgate and Alastair Levy
From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Corn Laws; currency; Ricardo, David P.; theory of rent; Thompson, Thomas P.; Westminster Review

Article

Appointed as the first Crown Governor of the British territory of Sierra Leone in 1808, Thompson was recalled under suspicion of financial impropriety in 1809. The real explanation for his departure, however, had more to do with the fact that the Sierra Leone Company (which had governed since 1790) found excessively disturbing Thompson's determination to rid the colony of an apprenticeship system whose features, as he saw it, were hardly different from those of slavery. The abuses which Thompson observed had developed, it should be noted, despite the fact that the Sierra Leone Company had been set up by anti-slavery philanthropists, including William Wilberforce and the economist Henry Thornton, with the intention of returning liberated slaves from the Americas to Africa (and, it was hoped, to illustrate the profitability of an African colonial trade not based on slavery).

On his return voyage to England, Thompson was the victim of an act of piracy. His vessel was boarded by the crew of a French corvette, and while its captain entertained Thompson, the British vessel was liberated of its cargo and provisions. Once safely back in England, Thompson applied to the Prime Minister (Lord Liverpool) for another official posting, but ‘in case no other situation should present itself’, he considered the possibility of single-handedly introducing the study of political economy into the University of Cambridge ‘in order to provide a living for myself’ (letter to E.P. Sells, January 1811, cited in Johnson, 1957, p. 70). This idea was not entirely fanciful, since Thompson was a graduate of Queen's College (BA, seventh wrangler, 1802) and a fellow of that college. It transpired, however, that the first regular lectures on the subject at Cambridge were given by George Pryme in 1816, and Thompson instead reactivated his commission in the army (into which service he had switched from the navy in 1806). His command of a defeat at Muscat led to his being court martialled but acquitted (though with a reprimand for ‘rashly undertaking the expedition with so small a detachment’) in 1820.

In 1822 Thompson played an active role in the founding the *Westminster Review* (financed by a £4,000 advance from Bentham), which aimed to provide a radical alternative to the Tory *Quarterly Review* and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the Whig *Edinburgh Review*. James Mill turned down the offer of its editorship, although he did contribute to its first number (in 1824) for which Thompson himself wrote an article on the ‘Instrument of exchange’. This constituted his first published work on economics. Thompson was sole proprietor of the *Review* from 1829 until 1836 when, on his election to the reformed House of Commons for the seat of Hull (where he had been born on 15 March 1783), he transferred its ownership to William Molesworth.

In 1826 Thompson published the first of his longer tracts on economics, *An Exposition of Fallacies on Rent, Tithes, &etc*. Ostensibly an attack on the Ricardian theory of rent (indeed, Thompson retitled its second edition *The True Theory of Rent in Opposition to Mr Ricardo and Others*), John Stuart Mill described it as ‘a striking exemplification of the mistakes of an ingenious, but not thoroughly informed mind’ and claimed that, in fact, Thompson's ‘theory of rent differs from that of Mr Ricardo only in the expression’ (1828, pp. 178–9). However, Mill's claim is open to question. Thompson argued that rent was determined by ‘the limited quantity of land in comparison with the competitors for its produce’ (1826, pp. 8–9) – quite how this could be said to be essentially Ricardo's theory ‘in different words’ is difficult to see. Not only does it fail to distinguish between extensive and intensive rent, but it ignores altogether the effect on the conditions of production of wage goods of restrictions on the importation of corn which is the key to Ricardo's argument. The only observation that needs to be made about Mill's claim is, perhaps, that it may tell us rather more about his own contribution to the decline of Ricardian economics than about Ricardo's theory of rent. That Mill could advance such a claim within five years of Ricardo's death makes it less difficult to understand why many felt that ‘little remained of Ricardo's theory’ by the end of that decade.

In 1827 followed the *Catechism of the Corn Laws* which Mill pronounced ‘one of the most useful works which have appeared in the present controversy’ (1828, p. 186); an interesting judgement given that its opening section was based on the *Exposition*. The *Catechism* presents a list of 120 (later increased to 365) ‘Proteus-like fallacies’ and answers, and has been referred to as ‘the arsenal whence the Anti-Corn Law League drew its best weapons’ (Allibone, 1871). There then followed, during the seven years he owned the *Westminster Review*, better than 100 articles for that periodical on subjects as diverse as the reform of the House of Lords and Catholic emancipation. Most of these were republished in his multi-volume *Exercises, Political and Other* in 1842. Also to be mentioned are his opinions on currency questions (for example, 1848), where he was an opponent of ‘inflationist’ proposals largely on the grounds of the redistribution against workers which he saw as part of the process. In the crisis of 1847, when the Bank Act of 1844 was again the subject of hot debate, Thompson wrote: ‘I hold to my opinion that there will be mischief on the Currency question. I receive more half-mad pamphlets from Birmingham’ (letter to J. Bowring, April 1847, quoted in Johnson, 1957, p. 265). In 1852 he attempted to introduce into Parliament measures which would protect the value of the currency against depreciation due to new gold discoveries, but these were defeated.

There is much more that could be said of Thompson's remarkable career. He was a moral-force, class-alliance chartist (and was invited to participate in writing the draft act of parliament which was to become the People's Charter); he voted consistently with the Radicals when a member of parliament; he constructed, and published, a non-axiomatic system of geometry (Euclid without the axioms); and he invented an enharmonic organ which was exhibited at the Great Exhibition of 1851, where it received an honourable mention. He died at Blackheath on 6 September 1869.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Selected works

1826. *An Exposition of Fallacies on Rent, Tithes, &etc*. London: Hatchard & Son. 2nd edn, retitled *The True Theory of Rent in Opposition to Mr Ricardo and Others*, 1826.

1827. *A Catechism on the Corn Laws; with a list of fallacies and answers*. London: J. Ridgway.

1842. *Exercises, Political and Other*, 6 vols. London: E. Wilson.

1848. *A Catechism on the Currency*. London: E. Wilson.

1859. *Catechism on the Ballot; or a list of fallacies and the answers*. London: G. Brown.

Bibliography

Allibone, S.A. 1859–71. *A Critical Dictionary of English Literature and British and American Authors, Living and Deceased*, 3 vols. Philadelphia: Childs & Peterson; London; N. Trübner & Co.

Johnson, L.G. 1957. *General T. Perronet Thompson: 1763–1869: His Military, Literary and Political Campaigns*. London: Allen & Unwin.

Mill, J.S. 1828. A review of the third edition of *Catechism on the Corn Laws*. *Westminster Review* 13, 169–87.

How to cite this article

Milgate, Murray and Alastair Levy. "Thompson, Thomas Perronet (1783–1869) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000252> doi:10.1057/9780230226203.1889

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Thornton, Henry (1760 – 1815)

David Laidler
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank of England; Bullion Committee; Bullionist controversies; central banking; classical theory of money; Convertibility; credit; Fisher, I.; forced saving; Hume, D.; inflation; lender of last resort; monetary economics; neutrality of money; quantity theory of money; real bills doctrine; Ricardo, D.; Specie; Thornton, H.; transfer problem; velocity of circulation; Wicksell, J. G. K.

Article

I. Henry Thornton was born in 1760, the youngest son of John Thornton, a London merchant prominent in the Russian trade. All three of John Thornton's sons were important in the business community and all three served as Members of Parliament. The eldest, Samuel, followed his father in the Russian trade, was director of the Bank of England, and its Governor between 1799 and 1801; Robert served as Governor of the East India Company for a time, but business reverses were eventually to lead to his emigration to the United States; and Henry become an extremely successful London banker. He died, probably from consumption, in 1815.

John Thornton had been and early member of the Evangelicals, as those followers of John Wesley who remained within the Church of England were called, and Henry too was among their leaders, the most famous of whom was his second cousin and close friend William Wilberforce. The movement became known as the Clapham Sect largely because their informal headquarters was Thornton's country house, located in that then outlying village. The Evangelical were also known as ‘the Party of Saints’ and what we would now regard as the conventional piety and respectability of the Victorian middle classes owe much to their influence. Nevertheless, their milieu was not Victorian but Georgian and Regency England, where their insistence that public policy be informed by the same high moral purpose as their private lives was profoundly radical. Their best-known accomplishment was ending Britain's participation in the slave trade in 1809, and in 1833 the abolition of slavery itself in the British Empire; but the role of their Sunday School Movement in promoting popular literacy in Britain, not to mention the influence of their British and Foreign Bible Society on 19th-century missionary activity throughout the world, is also noteworthy.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Henry Thornton was at the centre of all of these activities and many others as organizer, fund-raiser and donor. Before his marriage in 1796 he habitually devoted six-sevenths of his considerable income to charity, and perhaps a quarter thereafter. During his 33 years' service in Parliament, in addition to his work against the slave trade, he supported such progressive causes as peace with the American colonies, accommodation with France, and Catholic Emancipation. He also devoted considerable time and energy to religious writings and his great-great-grandson E.M. Forster (1951) records that his posthumously published volume of *Family Prayers* was something of a Victorian bestseller which was still earning royalties for his descendants at the end of the 19th century.

Among all of this activity, Henry Thornton found time to study monetary economics. As a prominent banker and Member of Parliament it was natural that he would take a practical interest in such matters, particularly given the financial turbulence associated with the French Wars of 1793–1815 and the suspension of the gold convertibility of Bank of England notes which accompanied them. He gave evidence to the Parliamentary Committees enquiring into the circumstances of the suspension in 1797, and he was an important member of both the Commons Committee which investigated Irish currency problems in 1804 and the famous 'Bullion Committee' of 1810. However, he was also and above all a great monetary theorist, and his outstanding treatise, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802), gives him a strong claim to be regarded as the most important contributor to monetary economics between David Hume (1752) and Knut Wicksell (1898). Only David Ricardo could seriously be regarded as his rival here.

II. The early 18th century had seen considerable progress in monetary economics, and David Hume's three essays of 1752 are rightly regarded as containing the core of classical monetary theory. They set out the quantity theory doctrine that, other things being equal, the price level varies with the quantity of money, and accompany this with an analysis of the way in which, under a commodity standard, balance of payments mechanisms operate so as to equalize price levels and distribute the precious metals among countries. Although allowing that monetary changes can have short-run effects on real output, they also develop the basic classical postulate that money is neutral in the long run, affecting only prices; and in particular they argue that the rate of interest is not a monetary phenomenon.

Banks are scarcely mentioned in Hume's analysis, and though Adam Smith (1776) paid considerable attention to them, his model was the 18th-century Scottish system. Scottish commercial banks held their reserves in claims upon London, not upon any Scottish central bank, and Scotland was a small, largely price-taking economy. Hence Smith's analysis of the interaction of bank behaviour, the price level and the balance of payments, though remarkably perceptive, was far from complete. It had little to say about the transmission mechanisms at work here and about the role of financial assets other than banknotes in the monetary system. Moreover it had nothing at all to say about central banking.

By the 1790s, the development of the English monetary system had far outstripped the growth of knowledge concerning the principles that underlay its operations, and the financial crisis which culminated in the suspension of February 1797 drew attention to this gap in most dramatic fashion. Thornton's *Paper Credit*, published in 1802 but perhaps begun as early as 1796, not only remedied this deficiency, but brought monetary theory to a level of sophistication that it was not to surpass until the end of the 19th century, as a brief sketch of its contributions will make quite evident.

III. *Paper Credit* begins with a detailed description of the contemporary English monetary system, showing how a rather wide variety of credit instruments had come to circulate as what we would now call money, alongside coin and banknotes, and it argues that the velocities of various components of this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
complex ‘circulating medium’ differ among instruments and fluctuate over time. In common with virtually every monetary economist before Irving Fisher (and many thereafter), Thornton regarded velocities of circulation as frequently unstable and he discussed in some detail how the Bank of England should behave, both to minimize the occurrence of monetary instability and to offset its consequences when it arose. Thornton was by no means the only contributor to the ‘Bullionist Controversy’, as the debates of the period are called, to recognize the crucial role and responsibilities of the Bank of England as a central bank, but there were many, not least among the directors of that institution, who refused to do so; and Thornton's exposition of the issues involved represents an important contribution to monetary economics.

No doubt drawing upon his own first-hand observations of the mechanisms at work during the turbulent 1790s, Thornton stressed both the crucial role and the volatility of the public's confidence in the banking system's ability to redeem its liabilities (in terms of Bank of England notes in the case of country and private London banks, and, under convertibility, in terms of specie in the case of the Bank of England). He understood that bank customers, who were confident that they could obtain Bank of England notes or specie when they required it, would not in fact seek such accommodation, and that only those who had doubts about the convertibility of their assets would demand their redemption. Hence he argued that any initial fall in confidence could lead to a self-reinforcing drain of reserves from the system if the Bank of England responded to it by reducing lending and hence cutting down the supply of the very central bank notes that the public were demanding from country and London banks. For Thornton, the right response to such an ‘internal’ (that is, within the country) drain of reserves from the banking system was for the Bank of England to lend freely to all solvent borrowers in order to restore and maintain the public confidence in the system. In short, the by now conventional textbook analysis of the central bank's ‘lender of last resort’ function found its first full statement in *Paper Credit*.

But Thornton understood well enough that an internal drain was not the only possible source of pressure on reserves. An external drain associated with what we would now call an adverse balance of payments was also a possibility, and here the required remedy might be different. He was clear that, to the extent that the drain stemmed from an uncompetitively high domestic price level, it could only be remedied by monetary contraction, and hence by the central bank scaling down its loans, including those made to the rest of the banking system. In the conventional wisdom of the later 19th century concerning sound central bank practice, an external drain was always appropriately to be met by such measures, but Thornton (unlike Ricardo, who is the true father of that conventional wisdom) was more subtle than this in his analysis.

For him, money wages were sticky and any sudden monetary contraction carried with it the danger of disrupting markets and causing real output and employment to fall, a danger to be avoided if at all possible. Hence when developing the implications for Bank of England policy of his pioneering analysis of what was later to be called ‘the transfer problem’, he advocated that temporary drains of specie abroad, associated with bad harvests or once and for all subsidy payments to allies, be accompanied by as little domestic monetary contraction as seemed to that institution to be prudent. Under arrangements prevailing after 1797, he was even willing to entertain temporary departures of sterling from par with specie in the face of temporary external drains rather than risk the domestic disruption that might accompany monetary contraction. Thornton was thus in *Paper Credit* far from being an advocate of an automatic gold standard, and his views have something in common with those of such later advocates of managed paper currency as Thomas Attwood – not to mention John Maynard Keynes, as certain

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
commentators, notably Hicks (1967) and Beaugrand (1981) have pointed out.

IV. McCulloch (1845), who confused Henry with his brother Samuel, regarded *Paper Credit* as being too partial to the Bank of England in its arguments, but though it may certainly be regarded as a defence of that institution's behaviour during the early years of the restriction, it is nevertheless a critical defence. Even so, by 1810, Henry Thornton was a prominent member of the Bullion Committee, and had become one of the Bank's sternest critics, advocating, both as a signatory to the Committee's Report (Cannan, 1919) and in two Commons speeches on the Report, that the obligation to redeem its notes in specie be reimposed upon it as soon as possible, a measure which was designed to narrow considerably the scope for discretion left to the Bank when confronted with an external drain.

Thornton's policy stance had changed between 1802 and 1810, but there is no evidence that his underlying analytic views were any different. First and foremost, and despite certain affinities, mentioned above, between his work and that of subsequent advocates of managed paper standards, Henry Thornton was always, as Hicks (1967) has put it, a ‘hard money’ man as far as long-run policy questions were concerned. He regarded the maintenance of the specie value of Bank of England liabilities as the proper overriding end of monetary policy. After 1797 he expected the Bank of England, subject to certain caveats about bad harvests and once and for all transfers, to manage its discounts so as to stabilize the exchange rate and the price of specie. In 1802 he believed that the Bank could be trusted to do so without the check of convertibility, but by 1810 he had changed his mind.

Though the actual conduct of monetary policy, particularly after 1811, shows that, luckily for Britain, they did not always practise what they preached, the directors of the Bank declared themselves firmly committed to the so-called ‘real bills doctrine’ in their evidence to the Bullion Committee, as they did in many other statements. This doctrine distinguishes between ‘real bills’, drawn to finance goods in the process of production and distribution, and ‘fictitious bills’ those which simply represent a debt with no corresponding real asset to back them. It then argues that a banking system in general, and a central bank in particular, which confines its activities to the discount of the former, cannot affect the price level. The quantity of money generated by following such practices will, so it is claimed, vary with the volume of output and adjust itself automatically and passively to the ‘needs of trade’.

Thornton had considered and comprehensively refuted this bundle of fallacies in *Paper Credit*. He had shown that, because there is no necessary relationship between the period for which commercial bills are discounted and the period of time that elapses between the beginning of the production of a particular unit of output and its final consumption, the distinction between ‘real’ and ‘fictitious’ bills was specious. Distinguishing between credit per se, and the role of credit instruments as components of the circulating medium, he had also shown how money, even if created against the security of good quality commercial bills, could influence the price level. Finally, and crucially, he had shown that the demand of manufacturers and merchants for bank credit would vary with the relationship between the banking system's lending rate and the expected rate of profit in such a way that, if the latter were high relative to the former, potentially unlimited monetary expansion and inflation could be generated by a banking system whose central authority took the real bills doctrine as its sole operating guide.

These arguments of Thornton's play a central role in the 1810 *Report* of the Bullion Committee and reflect his influence on that document. The explicit rejection of them by the directors of the Bank of England, not to mention widespread concern about inflation during 1809–10, was a crucial factor in persuading the committee in general, and Thornton in particular as one of its key members, to recommend that the constraint of specie convertibility be reimposed upon the Bank as soon as possible,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
a recommendation which was, of course, rejected by Parliament in 1811 along with the rest of the *Bullion Report*.

V. The reader familiar with the later literature of monetary economics will recognize the essentially Wicksellian (for example, 1898) flavour of Thornton's discussion of the relationship between bank lending policies and inflation. In a parliamentary speech of 1811 on the *Bullion Report* he elaborated on his earlier analysis by allowing for the influence of inflation expectations on the perceived real interest burden implied by any given nominal bank lending rate. This insight, which plays only an occasional and peripheral role in Wicksell's work, was of course central to the contributions of another great monetary theorist, Irving Fisher (1896). Moreover in his analysis of these matters, Thornton developed a version of what was later to be called the ‘forced saving’ doctrine which played an important role in early 20th-century business cycle theory. In the light of all this, it would be easy to jump to the conclusion that Thornton's work was well known to his successors. However, it was not.

Failing health and a relatively early death removed Henry Thornton from the centre of monetary controversy just as David Ricardo came to the height of his powers and influence. It was Ricardo and not Thornton who was destined to become the recognized authority to whom 19th-century monetary economists working within the classical tradition looked for guidance in matters of monetary theory. As Hutchison (1968) points out, J.S. Mill (1848) was the last important 19th-century author to recognize Thornton's contributions. Thereafter, his name faded from view, and was not even known to Wicksell; it is largely due to the efforts of Jacob Viner (1924; 1937) and particularly Friedrich von Hayek (1939) that his true stature came to be appreciated in the 20th century. Nevertheless, his ideas were well known to his contemporaries, not least to Ricardo, and as transmitted by them, not always without a certain loss of subtlety, they permeate 19th-century classical monetary theory. Thus if Henry Thornton's name was often forgotten by economists, his contributions to the subject were certainly not. For most men, this would be small consolation indeed, but one suspects that so benevolent and self-effacing a man as Henry Thornton might have been content with such an outcome.

See Also

- Hume, David
- quantity theory of money

Selected works

1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. Together with his evidence given before the Committees of Secrecy of the two Houses of Parliament in the Bank of England, March and April 1797, some manuscript notes, and his speeches on the Bullion Report, May 1811. Edited with an introduction by F.A. von Hayek, London: George Allen & Unwin, 1939. Reprinted, London: Frank Cass & Co.; New York: Augustus Kelley, 1962.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Beaugrand, P. 1981. *Henry Thornton: un precurseur de J.M. Keynes*. Paris: Presses Universitaires de France.

Cannan, E., ed. 1919. *The Paper Pound of 1797–1821: The Bullion Report*. London: P.S. King & Son. 2nd edn, 1921, reprinted New York: Augustus M. Kelley, 1969.

Fisher, I. 1896. Appreciation and interest. *AEA Publications* 3(11), 331–442.

Forster, E.M. 1951. Henry Thornton. In E.M. Forster, *Two Cheers for Democracy*, London: Abinger. Reprinted, London: Edward Arnold, 1972.

Hayek, F.A. von. 1939. Introduction to Thornton (1802).

Hicks, J.R. 1967. Thornton's *Paper Credit*. In *Critical Essays in Monetary Theory*, London: Oxford University Press.

Hume, D. 1752. Of money; Of the balance of trade; Of interest. In *Essays, Moral, Political and Literary*, ed. D. Hume. London. Reprinted London: Oxford University Press, 1962.

Hutchison, T.W. 1968. Thornton, Henry. In *International Encyclopaedia of the Social Sciences*, vol. 16. New York: Macmillan and Free Press.

McCulloch, J.R. 1845. *The Literature of Political Economy*. London.

Mill, J.S. 1848. *Principles of Political Economy with Some of Their Applications to Social Philosophy*. London: Parker.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 2 vols, ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Clarendon Press, 1976.

Viner, J. 1924. *Canada's Balance of International Indebtedness 1900–1913*. Cambridge, MA: Harvard University Press.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper Bros.

Wicksell, K. 1898. *Interest and Prices*. Trans. R.F. Kahn, London: Macmillan for the Royal Economic Society, 1936.

How to cite this article

Laidler, David. "Thornton, Henry (1760–1815)." The New Palgrave Dictionary of Economics. Second

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000051>
doi:10.1057/9780230226203.1700

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

threshold models

Timo Teräsvirta
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article contains a short account of threshold, smooth transition and Markov switching autoregressive models. Neural network models are highlighted as well. Linearity testing, parameter estimation and, more generally, modelling are considered. Forecasting with threshold models receives attention. Suggestions for further reading are supplied.

Keywords

ARIMA models; artificial neural network models; asymmetric behaviour; forecasting; identification; likelihood; linear models; logistic smooth transition regression models; Markov chains; Markov-switching models; maximum likelihood; modelling; purchasing power parity; regression error specification test; smooth transition regression models; switching regression models; threshold autoregressive models; threshold models

Article

The models

Stochastic nonlinear models have been widely used in economic applications. They may arise directly from economic theory. There also exist nonlinear models that have first been suggested by statisticians, engineers and time series analysts and then found application in economics. A broad class of these models, here called threshold models, has the property that the models are either piecewise linear or may be more generally considered as linear models with time-varying parameters. This category of nonlinear models includes switching regression or threshold autoregressive models, smooth transition models, Markov-switching or hidden Markov models. Artificial neural network models may also be included in this class of nonlinear models.
The switching regression (SR) model or, in its univariate form, the threshold autoregressive (TAR)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
model, is defined as follows:

$$y_t = \sum_{j=1}^{r} (\boldsymbol{\alpha}_j' \mathbf{z}_t + \varepsilon_{jt}) I(c_{j-1} < s_t \leq c_j)$$

(1)

where $\mathbf{z}_t = (\mathbf{w}_t', \mathbf{x}_t')'$ is a vector of explanatory variables, $\mathbf{w}_t = (1, y_{t-1}, \ldots, y_{t-p})'$ and $\mathbf{x}_t = (x_{1t}, \ldots, x_{kt})'$, s_t is an observable switch-variable, usually assumed to be a continuous stationary random variable, $c_0, c_1, \ldots, c_r$ are switch or threshold parameters, $c_0 = -\infty$, $c_r = M < \infty$, and $I(A)$ is an indicator function: $I(A)=1$ when event A occurs; zero otherwise. Furthermore, $\boldsymbol{\alpha}_j = (\alpha_{0j}, \alpha_{1j}, \ldots, \alpha_{mj})'$ such that $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}_j$ for $i \neq j$, where $m = p + k + 1$, $\varepsilon_{jt} = \sigma_j \varepsilon_t$ with $\{\varepsilon_t\} \sim \text{iid}(0, 1)$, and $\sigma_j > 0$, $j = 1, \ldots, r$. It is seen that (1) is a piecewise linear model whose switch-points, however, are generally unknown. The most popular choice in applications is r=2, that is, the model has two regimes. If s_t=t, eq. (1) is a linear model with r–1 breaks and $\boldsymbol{\alpha}_j \neq \boldsymbol{\alpha}_{j+1}$, $j = 1, \ldots, r-1$. These models have recently become quite popular in econometrics and there is now a substantial literature on how to determine the number of breaks and estimate the break points. For a generalization of the threshold autoregressive regression model to the vector case, see Tsay (1998).

TAR models have been used to characterize asymmetric behaviour in GNP or unemployment rates and to consider the purchasing power parity hypothesis. They have also been applied to modelling interest rate series as well as other financial time series.

One may substitute $I(s_t = j)$ for $I(c_{j-1} < s_t \leq c_j)$ in (1), where s_t is an unobservable regime indicator with a finite set of values $\{1, \ldots, r\}$. On the assumption that s_t follows a first-order Markov chain, that is, $\Pr\{s_t = i | s_t = j\} = p_{ij}$, $i, j = 1, \ldots, r$, (1) becomes a hidden Markov or Markov-switching (MS) model (see Lindgren, 1978). Higher-order Markov chains are possible but rarely used in econometric applications.

Equation (1) with an unobservable regime indicator is not, however, the most frequently applied hidden Markov model in econometrics. Consider the univariate model

$$y_t = \mu_{s_t} + \sum_{j=1}^{p} \alpha_j (y_{t-j} - \mu_{s_{t-j}}) + \varepsilon_t$$

(2)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where s_t follows a first-order Markov chain as before, and $\mu_{s_t} = \mu^{(i)}$ for $s_t=i$, such that $\mu^{(i)} \neq \mu^{(j)}$, $i \neq j$.

The stochastic intercept of this model, $\mu_{s_t} - \sum_{j=1}^{p} \alpha_j \mu_{s_{t-j}}$, can thus obtain r^{p+1} different values, which gives the model the desired flexibility. A comprehensive discussion of MS models can be found in Hamilton (1994, ch. 22). The model (2) has been frequently fitted, for example, to GNP series and interest rates. In the latter case, the model may be used for identifying changes in the monetary policy of the central bank.

SR and Markov-switching models contain a finite number of regimes. There is a class of models called smooth transition regression (STR) models, in which there are two 'extreme regimes', and the transition between them is smooth. A basic STR model is defined as follows:

$$y_t = \boldsymbol{\varphi}' \mathbf{z}_t + \boldsymbol{\psi}' \mathbf{z}_t G(\gamma, \mathbf{c}, s_t) + \varepsilon_t \tag{3}$$

where $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \ldots, \varphi_m)'$ and $\boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \psi_m)'$ are parameter vectors, $\mathbf{c} = (c_1, \ldots, c_K)'$ is a vector of location parameters, $c_1 \leq \ldots \leq c_K$, and $\varepsilon_t \sim \mathrm{iid}(0, \sigma^2)$. The transition function G(γ ,c,s_t) is a bounded function of s_t, continuous everywhere in the parameter space for any value of s_t. The logistic transition function has the general form

$$G(\gamma, \mathbf{c}, s_t) = \left(1 + \exp\left\{-\gamma \prod_{k=1}^{K} (s_t - c_k)\right\}\right)^{-1}, \gamma > 0 \tag{4}$$

where $\gamma > 0$ is an identifying restriction. Equation (3) jointly with (4) defines the logistic STR (LSTR) model. The most common choices in practice for K are K=1 and K=2. For K=1, the parameters $\boldsymbol{\varphi} + \boldsymbol{\psi} G(\gamma, \mathbf{c}, s_t)$ change monotonically as a function of s_t from ϕ to $\phi+\psi$. For K=2, they change symmetrically around the mid-point $(c_1+c_2)/2$ where this logistic function attains its minimum value. Slope parameter γ controls the slope and c_1 and c_2 the location of the transition function. When K=1 and $\gamma \to \infty$ in (4), the model (3) becomes an SR model with r=2.

The LSTR model with K=1 (LSTR1 model) is capable of characterizing asymmetric behaviour. As an example, suppose that s_t measures the phase of the business cycle. Then the LSTR1 model can describe processes whose dynamic properties are different in expansions from what they are in recessions, and the transition from one extreme regime to the other is smooth. The same is true for SR and the MS models with the difference that instead of a smooth transition there is an abrupt switch. The LSTR2

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

model is appropriate whenever the dynamic behaviour of the process is similar at both large and small values of s_t and different in the middle.

Yet another nonlinear model that is worth mentioning because it is related to threshold models is the artificial neural network (ANN) model. The simplest single-equation case is the so-called 'single hidden-layer' ANN model. It has the following form

$$y_t = \boldsymbol{\beta}_0' \mathbf{z}_t + \sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j' \mathbf{z}_t) + \varepsilon_t \tag{5}$$

where y_t is the output series, $\mathbf{z}_t$ is the vector of inputs, and $\boldsymbol{\beta}_0' \mathbf{z}_t$ is a linear unit with $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0,p+k})'$. Furthermore, β_j, $j=1,\ldots,q$, are parameters, called "connection strengths" in the neural network literature. Function G(·) is a bounded, asymptotically constant function called the 'squashing function' and γ_j, $j = 1, \ldots, q$, are parameter vectors. They form the hidden layer which the name of the model refers to. Typical squashing functions are monotonically increasing ones such as the logistic function and the hyperbolic tangent function. The errors ε_t are often assumed iid $(0, \sigma^2)$. Many neural network modellers assume $\boldsymbol{\beta}_0 = \beta_{00}$, where β_{00} is called the 'bias'.

A theoretical argument for the use of ANN models is that they are universal approximators. Suppose that $y_t = H(\mathbf{z}_t)$, that is, there exists a functional relationship between y_t and $\mathbf{z}_t$. Then, under mild regularity conditions for H, there is a positive integer $q \le q_0 < \infty$ such that for an arbitrary $\delta > 0$, $|H(\mathbf{z}_t) - \sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j' \mathbf{z}_t)| < \delta$. The importance of this result lies in the fact that q is finite, so that any unknown function H can be approximated arbitrarily accurately by a linear combination of squashing functions $G(\boldsymbol{\gamma}_j' \mathbf{z}_t)$. This has been discussed in several papers, including Cybenko (1989), Funahashi (1989) and Hornik, Stinchcombe and White (1989). Neural network models are very generously parameterized and are only locally identified. The log-likelihood typically contains a large amount of local maxima, which makes parameter estimation difficult.

Testing linearity

All threshold models nest a linear model but they are not identified when the data are generated from this linear model. For this reason, testing linearity before fitting a threshold model is necessary in order to avoid the estimation of an unidentified model whose parameters cannot be estimated consistently. In this case, linearity testing has to precede any nonlinear estimation.

There exist general misspecification tests that are linearity tests if the specification to be tested is linear; see Bierens (1990) and Stinchcombe and White (1998). There also exist parametric tests that have been

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
designed to be tests against an unspecified alternative but are not consistent against deviations from linearity. The popular Regression Error Specification Test (RESET) of Ramsey (1969) is such a test. Teräsvirta (1998) and van Dijk, Teräsvirta and Franses (2002) discuss tests against smooth transition regression models and Hansen (1999) surveys linearity testing against TAR models. Linearity testing in the Markov-switching framework is considered in Garcia (1998). Some recent econometrics textbooks discuss linearity tests against various threshold models.

As already mentioned, threshold models nest a linear model and are not identified if linearity holds. For example, the STR model (3) is not identified if $\gamma = 0$ in (4) or $\boldsymbol{\psi} = \mathbf{0}$. In the former case, ψ and **c** are not identified, and in the latter, the nuisance parameters are γ and **c**. Consequently, the standard asymptotic theory is not applicable in testing linearity. This problem may be solved following Davies (1977). Let γ be the vector of nuisance parameters. For example, $\boldsymbol{\gamma} = (\gamma, \mathbf{c}')'$ in (3) when the null hypothesis is $\boldsymbol{\psi} = \mathbf{0}$. When γ is known testing linearity is straightforward. Let $S_T(\gamma)$ be the corresponding test statistic whose large values are critical and define $\Gamma = \{\boldsymbol{\gamma}: \boldsymbol{\gamma} \in \Gamma\}$, the set of admissible values of γ. When γ is unknown, the statistic is not operational because it is a function of γ. The problem is solved by defining another statistic $S_T = \sup_{\boldsymbol{\gamma} \in \Gamma} S_T(\boldsymbol{\gamma})$ that is free of nuisance parameters γ. The asymptotic distribution of S_T under the null hypothesis does not generally have an analytic form, but Davies (1977) gives an approximation to it that holds under certain conditions, including the assumption that $S(\boldsymbol{\gamma}) = \operatorname{plim}_{T \to \infty} S_T(\boldsymbol{\gamma})$ has a derivative. Other choices of test statistic include the average:

$$S_T = \operatorname{ave} S_T(\boldsymbol{\gamma}) = \int_\Gamma S_T(\boldsymbol{\gamma})\, dW(\boldsymbol{\gamma}) \tag{6}$$

where $W(\gamma)$ is a weight function defined by the user such that $\int_\Gamma W(\boldsymbol{\gamma})\, d\boldsymbol{\gamma} = 1$, and the exponential

$$\exp S_T = \ln\left(\int_\Gamma \exp\{(1/2) S_T(\boldsymbol{\gamma})\}\, dW(\boldsymbol{\gamma})\right). \tag{7}$$

Andrews and Ploberger (1994) have recommended these tests and demonstrated their local asymptotic optimality properties. The statistics (6) and (7) are two special cases in the family of average exponential tests (for definitions and details, see Andrews and Ploberger, 1994). Hansen (1996) shows how to obtain asymptotic critical values for these statistics by simulation under rather general conditions. His method is computationally intensive but useful. It may be pointed out that it works for SR and STR models where s_t is observable. For MS models, the situation is more complicated, see Garcia (1998) for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
discussion.
A computationally simpler alternative is to circumvent the identification problem instead of directly solving it. It has been popular in testing linearity against smooth transition models, eq. (3). The idea is to replace the transition function (4) by its Taylor series approximation around the null hypothesis $\gamma=0$. This transforms the testing problem into one of testing a linear hypothesis in a linear auxiliary regression; see Luukkonen, Saikkonen and Teräsvirta (1988) or Teräsvirta (1998).

Parameter estimation

Parameters of threshold models have to be estimated numerically. This is because the objective function to be optimized is not quadratic in parameters, so an analytical solution to the problem does not exist. The easiest models to estimate are the switching regression or threshold autoregressive models. Their parameters are estimated conditionally by ordinary least squares (OLS), given the switch parameters $c_1, \ldots, c_r$, and the combination of $c_1, \ldots, c_r$ yielding the smallest sum of squared residuals gives the estimates of these and the other parameters. For example, when $r=1$ the OLS estimation is repeated for a set of c_1 values such that both regimes contain at least a certain minimum amount of observations, typically 10% or 15% of the total number. Under rather general conditions, including stationarity and ergodicity of the TAR process, the least squares estimators are $\sqrt{T}$-consistent and asymptotically normal. The threshold parameter estimators are super (T-) consistent.
Smooth transition models are estimated using standard maximum likelihood. The most efficient numerical method is the Newton–Raphson method that makes use of both the first (the score) and the second (the Hessian) partial derivatives of the log-likelihood function. It has many variants in which the Hessian is replaced by computationally simpler alternatives that either do not require second derivatives, such as the method of scoring or the so-called Berndt–Hall–Hall–Hausman (BHHH) algorithm, or avoid inverting the Hessian altogether. Examples of this include the steepest descent and variable metric methods. Of the latter, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm can be found in a number of modern software packages.
Hidden Markov models cannot be estimated using standard optimization algorithms because they contain the latent variable s_t. Their parameters are typically estimated using the expectation–maximization (EM) algorithm (see Cappé, Moulines and Rydén, 2005, and Hamilton, 1994, ch. 22).
Estimation of ANN models using maximum likelihood is often numerically demanding because the likelihood function can contain a large number of local maxima, due to a large number of parameters. This problem has been discussed in Medeiros, Teräsvirta and Rech (2006) and White (2006). Because of this difficulty, the literature on ANN models contains a wide variety of estimation methods of various kinds; see for example Fine (1999) and White (2006).

Modelling

Typically, economic theory does not uniquely determine the functional form of a threshold model. This means that the model builder has to select an appropriate model for the problem at hand. In this case, applying a consistent modelling strategy is helpful. The modelling approach of Box and Jenkins (1970)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for the ARIMA class of linear models is a case in point. When it comes to threshold models, modelling strategies consisting of stages of specification, estimation and evaluation have been worked out and applied for TAR or, more generally, SR classes of models as well as STR models. For the former, see Tsay (1989) (univariate models) and Tsay (1998) (multivariate models) and for the latter, Teräsvirta (1998) or Teräsvirta (2004). Medeiros, Teräsvirta and Rech (2006) suggest a similar procedure for ANN models. An essential first stage in all these strategies is testing linearity. If linearity is not rejected, the task of the model builder is considerably simplified.

Forecasting

The main purpose of univariate nonlinear models is forecasting. Multivariate models may also be useful for policy analysis. Forecasts are typically conditional means in which the conditioning set consists of a subset of the information available at the time of making the forecast. In nonlinear models such as threshold models, a typical situation is that making forecasts for more than one period ahead requires numerical methods. This is due to the fact that for a random variable X, generally $Eg(X) \neq g(EX)$. Equality holds if g is a linear function of X.
To illustrate, assume an information set $\mathcal{F}_T$ at time T. The optimal one-period mean forecast $f^y_{T,1} = E\{y_{T+1}|\mathcal{F}_T\}$, the conditional mean of y_{T+1}, given $\mathcal{F}_T$. For example, consider the simple bivariate model

$$y_t = g(x_{t-1}) + \varepsilon_t \tag{8}$$

where

$$x_t = \beta x_{t-1} + \eta_t \tag{9}$$

with $|\beta| < 1$, and $\{\eta_t\}$ is a sequence of independent, identically distributed random variables with zero mean. Function $g(\cdot)$ may define an SR, STR or ANN model. The forecast for y_{T+1} equals $f^y_{T,1} = g(x_T)$ as $E\{\varepsilon_{T+1}|\mathcal{F}_T\} = 0$. Thus, if one knows the function $g(\cdot)$, one-step forecasts can be obtained with no difficulty.
The optimum two-step forecast is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$f_{T,2}^{y} = E\left\{y_{T+2}|\mathcal{F}_T\right\} = E\left\{g(x_{T+1})|\mathcal{F}_T\right\}.$$

As x_{T+1} is not usually known at time T, it has to be forecast from its autoregressive equation. This gives a one-step OLS forecast $f_{T,1}^{x} = \beta x_T$. The two-step forecast equals

$$f_{T,2}^{y} = E\left\{g(f_{T,1}^{x} + \eta_{T+1})|\mathcal{F}_T\right\}. \quad (10)$$

The exact forecast equals

$$fe_{T,2}^{y} = \int_{-\infty}^{\infty} g(f_{T,1}^{x} + z)\mathrm{d}D(z)$$

where $D(z)$ is the cumulative distribution function of z. The integral has to be calculated numerically. It may, however, also be approximated by simulation or by bootstrapping the residuals of the estimated model; see Granger and Teräsvirta (1993) or Teräsvirta (2006a). This alternative becomes even more practical when the forecast horizon exceeds two periods. Yet another alternative is to ignore the error η_{T+1}, but the "naive" forecast $fn_{T,2}^{y} = g(f_{T,1}^{x})$ is biased. In practice, the function $g(\cdot)$ is not known and has to be specified and estimated from the data before forecasting.
One may also obtain the forecast directly as

$$fd_{T,2}^{y} = E(y_{T+2}|\mathcal{F}_T)$$

so that $y_{t+2} = g_2(x_t, y_t) + \varepsilon_t^{*}$, say, and the function $g_2(\cdot)$ has to be determined and estimated separately, rather than derived from the one-step representation (8). A difficulty with this approach is that the errors will not necessarily be white noise. A separate forecast function is needed for each forecast horizon.
All forecasts from hidden Markov models can be obtained analytically by a sequence of linear

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

operations. This is a direct consequence of the fact that the regimes in (1) when $I(c_{j-1} < s_t \leq c_j)$ is replaced by $I(s_t = j)$, where s_t is a latent discrete variable, are linear in parameters. This is discussed for example in Hamilton (1993) or Teräsvirta (2006a).

Experiences from large empirical studies in which macroeconomic variables are forecast with threshold models, are mixed. No model dominates the others, and in several cases nonlinear threshold models do not improve the accuracy of point forecasts compared to linear models. Recent studies of this type include Stock and Watson (1999), Marcellino (2004) and Teräsvirta, van Dijk and Medeiros (2005).

Further reading

Many statistics and econometrics monographs contain accounts of threshold models, among them Franses and van Dijk (2000), Granger and Teräsvirta (1993) and Guégan (1994). Tong (1990) focuses on TAR models. There are also useful book chapters and review articles such as Bauwens, Lubrano and Richard (2000) offering a Bayesian perspective, Brock and Potter (1993), Teräsvirta (2006a,b) and Tsay (2002). For hidden Markov models, see Cappé, Moulines and Rydén (2005) and Hamilton (1993, 1994, ch. 22). The latter reference concentrates on the autoregressive model (2). Several thorough treatments of ANN models exist; see for example Fine (1999) or Haykin (1999).

See Also

- forecasting
- identification
- model selection
- nonlinear time series analysis
- statistical inference
- testing

Bibliography

Andrews, D.W.K. and Ploberger, W. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383–414.

Bauwens, L., Lubrano, M. and Richard, J.-F. 2000. *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Bierens, H.J. 1990. A consistent conditional moment test of functional form. *Econometrica* 58, 1443–58.

Box, G.E.P. and Jenkins, G.M. 1970. *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Brock, W.A. and Potter, S.M. 1993. Nonlinear time series and macroeconometrics. In *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao and H.D. Vinod. Amsterdam: North-Holland.

Cappé, O., Moulines, E. and Rydén, T. 2005. *Inference in Hidden Markov Models*. New York: Springer.

Cybenko, G. 1989. Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals, and Systems* 2, 303–314.

Davies, R.B. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–54.

Fine, T.L. 1999. *Feedforward Neural Network Methodology*. Berlin: Springer.

Franses, P.H. and van Dijk, D. 2000. *Non-Linear Time Series Models in Empirical Finance*. Cambridge: Cambridge University Press.

Funahashi, K. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks* 2, 183–92.

Garcia, R. 1998. Asymptotic null distribution of the likelihood ratio test in Markov switching models. *International Economic Review* 39, 763–88.

Granger, C.W.J. and Teräsvirta, T. 1993. *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Guégan, D. 1994. *Séries chronologiques non linéaires à temps discret*. Paris: Economica.

Hamilton, J.D. 1993. Estimation, inference and forecasting of time series subject to changes in regime. In *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao and H.R. Vinod. Amsterdam: North-Holland.

Hamilton, J.D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, B.E. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64, 413–30.

Hansen, B.E. 1999. Testing for linearity. *Journal of Economic Surveys* 13, 551–76.

Haykin, S. 1999. *Neural Networks. A Comprehensive Foundation*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Hornik, K., Stinchcombe, M. and White, H. 1989. Multi-layer feedforward networks are universal approximators. *Neural Networks* 2, 359–66.

Lindgren, G. 1978. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics* 5, 81–91.

Luukkonen, R., Saikkonen, P. and Teräsvirta, T. 1988. Testing linearity against smooth transition autoregressive models. *Biometrika* 75, 491–99.

Marcellino, M. 2004. Forecasting EMU macroeconomic variables. *International Journal of Forecasting* 20, 359–72.

Medeiros, M.C., Teräsvirta, T. and Rech, G. 2006. Building neural network models for time series: a statistical approach. *Journal of Forecasting* 25, 49–75.

Ramsey, J.B. 1969. Tests for specification errors in classical least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31, 350–71.

Stinchcombe, M.B. and White, H. 1998. Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* 14, 295–325.

Stock, J.H. and Watson, M.W. 1999. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, ed. R.F. Engle and H. White. Oxford: Oxford University Press.

Teräsvirta, T. 1998. Modeling economic relationships with smooth transition regressions. In *Handbook of Applied Economic Statistics*, ed. A. Ullah and D.E. Giles. New York: Dekker.

Teräsvirta, T. 2004. Smooth transition regression modeling. In *Applied Time Series Econometrics*, ed. H. Lütkepohl and M. Krätzig. Cambridge: Cambridge University Press.

Teräsvirta, T. 2006a. Forecasting economic variables with nonlinear models. In *Handbook of Economic Forecasting*, vol. 1, ed. G. Elliott, C.W.J. Granger and A. Timmermann. Amsterdam: North-Holland.

Teräsvirta, T. 2006b. Univariate nonlinear time series. In *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*, ed. T.C. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Teräsvirta, T., van Dijk, D. and Medeiros, M.C. 2005. Smooth transition autoregressions, neural networks, and linear models in forecasting macroeconomic time series: a re-examination. *International Journal of Forecasting* 21, 755–74.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Tong, H. 1990. *Non-Linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.

Tsay, R.S. 1989. Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association* 84, 231–40.

Tsay, R.S. 1998. Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* 93, 1188–202.

Tsay, R.S. 2002. *Analysis of Financial Time Series*. New York: Wiley.

van Dijk, D., Teräsvirta, T. and Franses, P.H. 2002. Smooth transition autoregressive models – a survey of recent developments. *Econometric Reviews* 21, 1–47.

White, H. 2006. Approximate nonlinear forecasting methods. In *Handbook of Economic Forecasting*, ed. G. Elliott, C.W.J. Granger and A. Timmermann. Amsterdam: North-Holland.

How to cite this article

Teräsvirta, Timo. "threshold models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_N000162> doi:10.1057/9780230226203.1701

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Th ü nen, Johann Heinrich von (1783 – 1850)

Jürg Niehans
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Thünen was born in Canarienhausen (Oldenburg) on 24 June 1783. He died on his estate Tellow (Mecklenburg), near Rostock, on 22 September 1850. His paternal ancestors were farmers; despite the ‘von’, they did not belong to the aristocracy.

After his father's early death, Thünen's mother married a timber merchant. The boy grew up in a small town on the northern seaboard, where he obtained a good, but short, high-school education. As an apprentice he got to know hard manual labour on a farm. There followed academic studies on all aspects of agronomy, including natural sciences, mathematics and economics, at the agricultural colleges of Gross-Flottbeck and Celle (where he heard Thaer) and at the University of Göttingen (where he read Adam Smith). Nevertheless, Thünen remained essentially a scientifically gifted autodidact. During this period (around 1803), he seems to have conceived the idea of his ‘isolated state’.

Newly married to the daughter of a respected landowner, Thünen first operated a rented estate. In 1809, with the inheritance from his father, he bought from his brother-in-law the rather run-down estate of Tellow with about 1,200 acres of land. Though his heart was in his intellectual pursuits rather than in practical farming, he succeeded in gradually paying off his initial debt and in raising the value of his property, leaving to his four children a prosperous estate with ample liquid funds.

Like Quesnay, who came from a similar background, Thünen made the farm his economic paradigm. With the Physiocrats and Thaer, he belongs to those representatives of the Enlightenment who regarded improvements in agriculture as the key to economic progress. For his estate he kept meticulous accounts, which he used to compute optimal solutions to management problems. He was a model employer with philanthropic, if somewhat paternalistic, ideas on social policy, who established a profit-sharing plan for his employees.

Of Thünen's *magnum opus*, ‘The Isolated State with Respect to Agriculture and Political Economy’, the first part, including the analysis of rent, location and resource allocation, appeared in 1826 after more than 20 years of work. The second part, containing the marginal productivity theory of distribution, only appeared in 1850. Additional papers, including important contributions on forestry, were published in 1863 by Thünen's biographer, H. Schumacher. All of this material is united in the third edition of 1875, but Waentig's later edition and also the English translations are limited to Part I and the first (and more important) half of Part II. Additional material was published by Braeuer in his volume of selected works, which also includes a bibliography of Thünen's writings. The literary remains, including unpublished manuscripts, are preserved in the Thünen-Archiv at the University of Rostock.

It has been said that Thünen was a prophet with little honour in any country, and even less in his own. This is inaccurate. It is true that he was at first disappointed about the reception of his book. Nevertheless, by 1827 he was an internationally known authority on agriculture, and the first edition was sold out within seven years. Tellow became a mecca for agronomists, attracting visitors from all over Europe. In 1830, Thünen was made a *doctor philosophiae honoris causa* by the University of Rostock. Politically a progressive liberal, he was elected to the National Assembly in Frankfurt in 1848, but could not attend because of his declining health. In the same year, the town of Teterow, with flags flying and bands playing, made him an honorary burgher. Like Quesnay, he died revered as a sage.

Thünen's scientific achievements are at different levels. In agronomy he made important contributions to the ‘statics’ of the soil, which are concerned with the steady state where fertility, by suitable crop rotation and fertilization, is maintained at an optimal level. In economics his most fundamental contribution is the method of deriving economic propositions from explicit optimizing models. By 1824 (as Braeuer reports) this had led him to the differential calculus, which he may thus have been the first to apply to economic problems. At a time when German economists liked to criticize Adam Smith for his ‘rationalism’, Thünen criticized him for his lack of an explicit theory, which he undertook to provide. In mathematical elegance his contribution falls far short of Cournot's, but it exceeds the latter in breadth and depth. It makes Thünen one of the patron saints of modern economics.

A more specific contribution is Thünen's theory of rent, location and resource allocation. In terms of modern economics, its elements can be summarized as follows. Suppose rye is sold in a central city at a given market price P. Production takes place on an unlimited plain of uniform fertility. Transportation to the market over s miles costs ts per bushel. The producer price, p, therefore, declines with increasing distance according to $p=P-ts$, as illustrated in the NE quadrant of Figure 1. Output per acre depends on labour per acre

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

according to $q=q(a)$. At each producer price, the manager selects his method of production in such a way that the marginal product of labour, $q'(a)$, evaluated at the producer price, equals the given (and uniform) wage rate, as described in the NW quadrant. The curve in the SW quadrant expresses the decline in the marginal product of labour as increasingly more labour-intensive methods are applied. The area 'below' the marginal product curve is total product. While the rectangle $q'(a)$ goes to wages, the shaded residual represents land rent. The curve in the SE quadrant, finally, shows labour intensity as a diminishing function of distance.

Figure 1

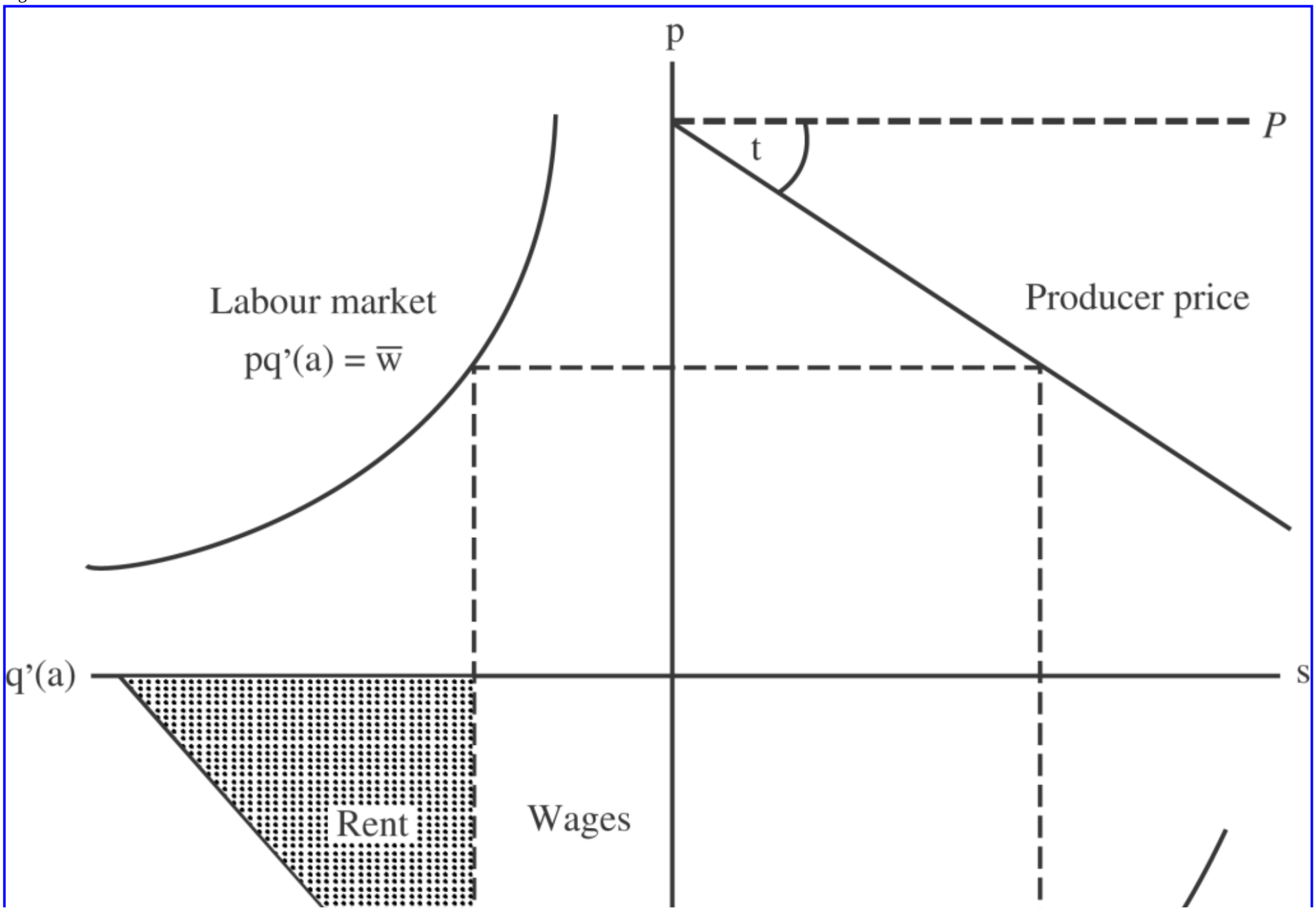

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

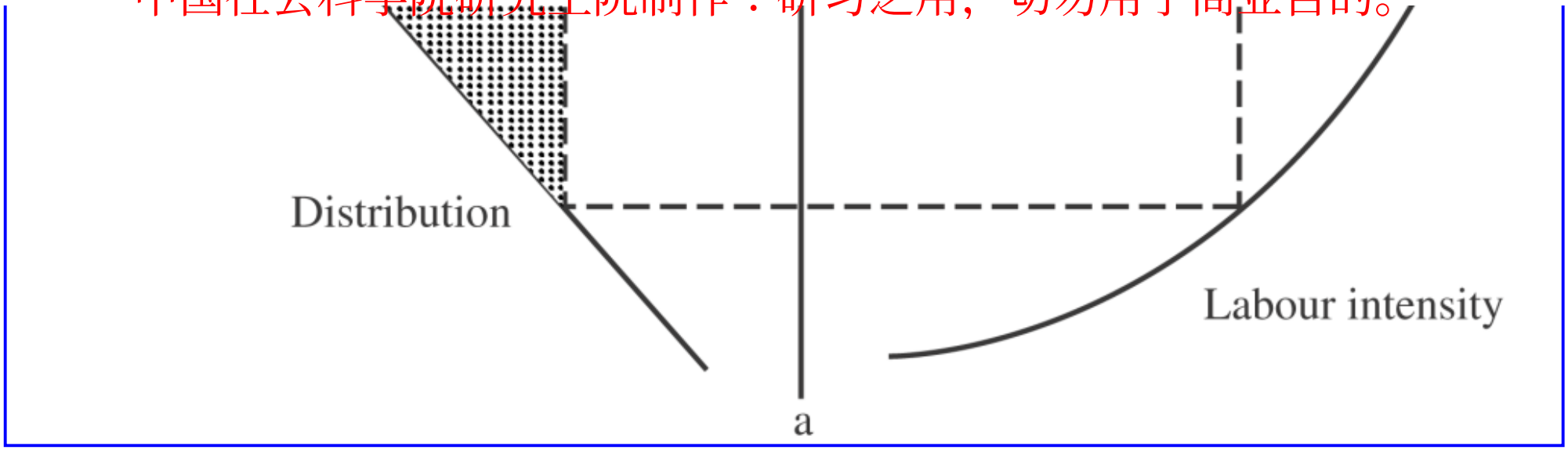

The solid curve in Figure 2 graphs land rent from rye production as a diminishing function of distance. Similar rent–distance curves can be constructed for other products like vegetables or lumber. They are represented by, respectively, the dotted and the broken curve. At each distance the farmer will plant the product promising the highest rent. This results in Thünen's famous rings. For a given product there may also be rings of different technologies.

Figure 2

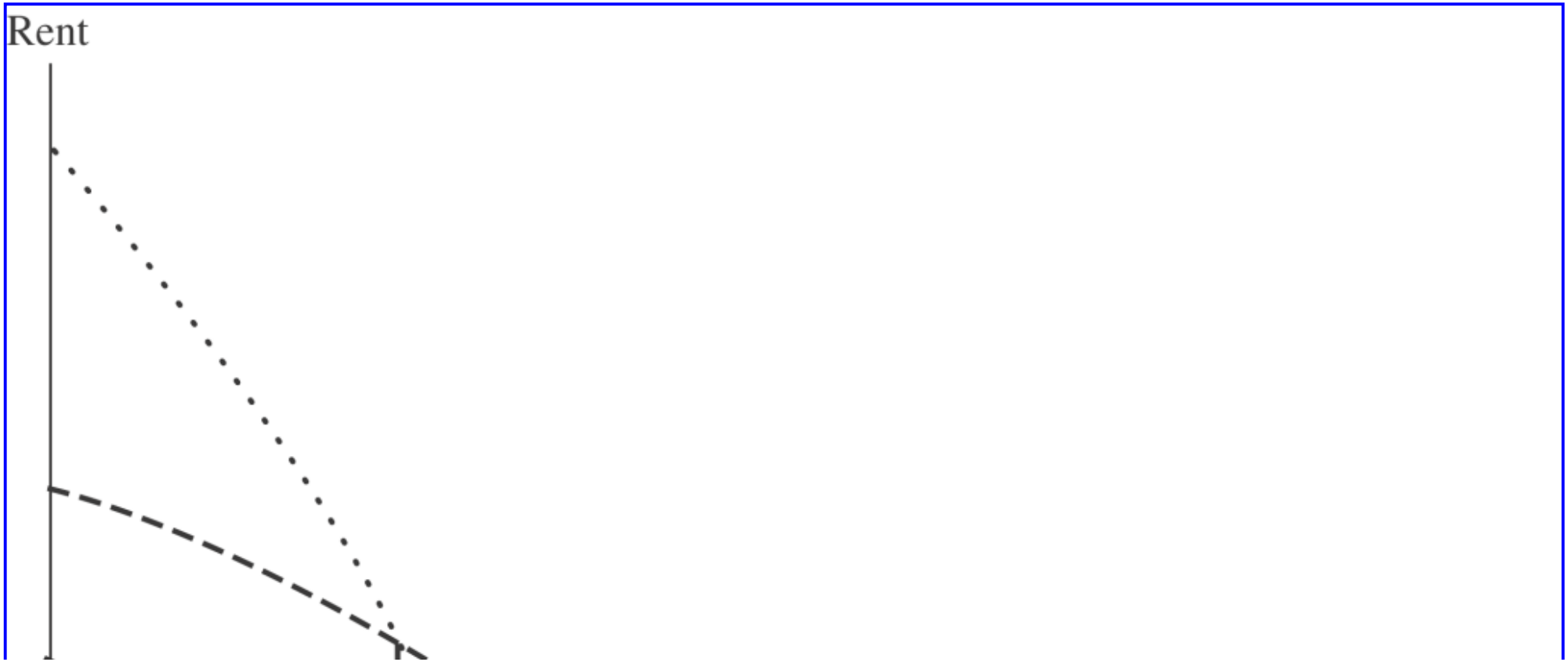

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

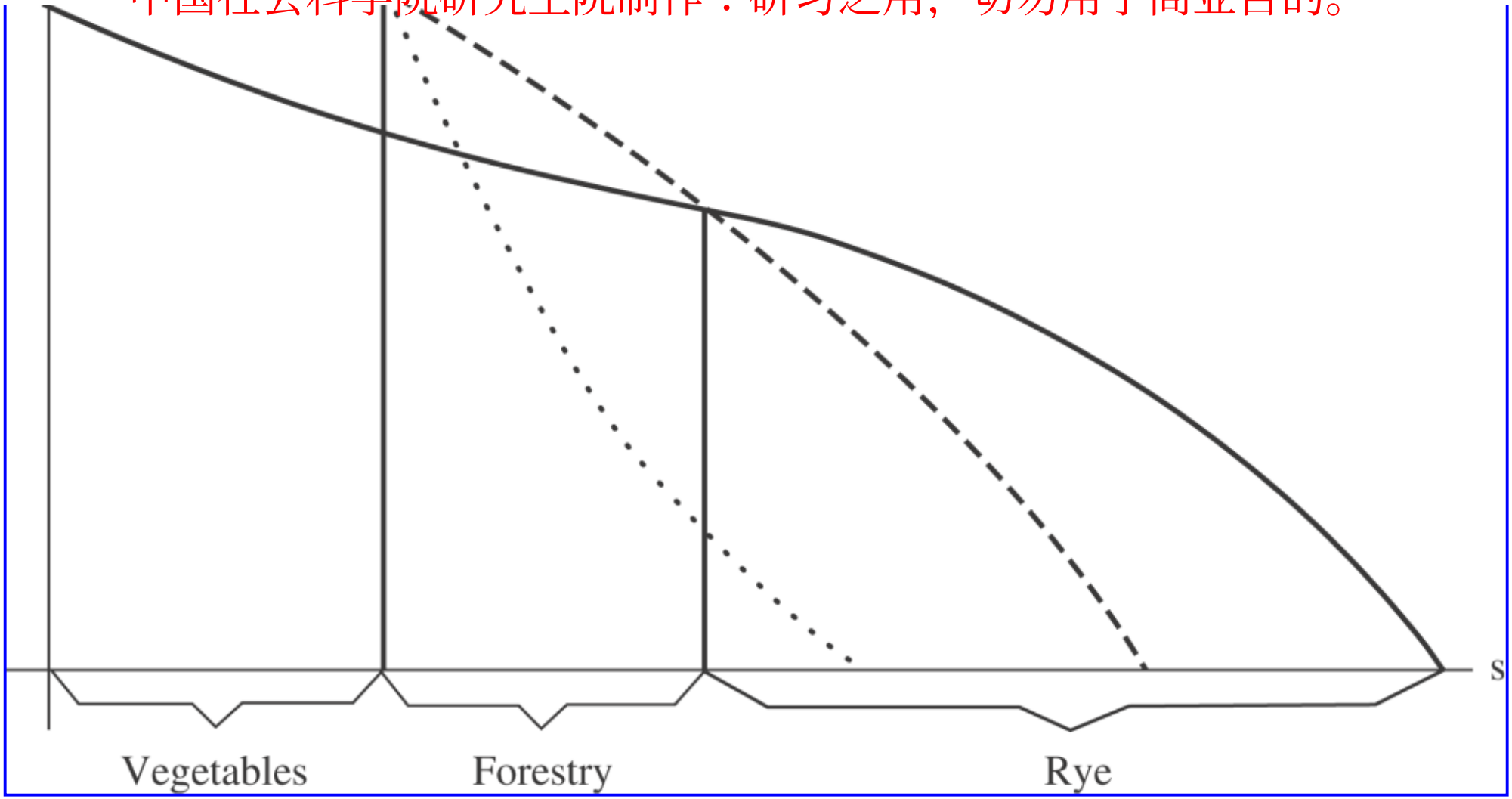

In analysing the comparative statics of the ‘isolated state’, Thünen shows that lower transportation costs and more rapidly diminishing returns tend to increase the distance from the city at which a good is produced or a technology is used. It is important to note that Thünen provides not only a theory of location, but also of factor intensities. That the relative efficiency of different technologies depends on market conditions is one of the main propositions he wanted to demonstrate.

The basic model is extended by Thünen in numerous directions. If the required quantities are given, the model determines their market prices. Since the rural workers do not generally pay the given city prices, their money wages will not actually be uniform. Freight costs may not be proportionate to distance. Substitutes and joint products are discussed. To the flows of agricultural products to the market centre, Thünen adds the reverse flows of consumer goods and means of production (like manure) and he pays attention to the unequal quality of the soil. The problem of the spatial distribution of several cities is raised, though not solved. It is finally shown that agricultural protection, by reducing the efficiency of land use, makes both parties worse off and that land taxes do not distort allocation. Despite its richness, Thünen's analysis remains partial in the sense that it does not determine a general spatial equilibrium. His notions about the price mechanism are crude. The long-winded discourse is replete with empirical calculations, relating Thünen's analysis to his account books down to the most minute details.

By applying his optimizing approach to factor inputs, Thünen became one of the originators of the marginal productivity theory of distribution. Using the Ricardian subterfuge of a rentless margin of cultivation, he explains his basic idea in the following words:

> Output *p* is the joint product of labour and capital. How should the share of each factor in the joint product be measured? We measured the effectiveness of capital by the increment in the output per worker due to an increase in the capital he works with. In this context, labour is constant, but capital is a variable magnitude. Suppose

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

now that this procedure is continued, but in the reverse sense of considering capital as constant and labour as growing. In this case, in a large-scale operation, the effectiveness of labour (the contribution of the worker to output) is recognized from the increment in total output due to the augmentation of workers by one (II, §19).

As in Turgot, the output increments, from both capital and labour, are postulated to decline with an increasing factor input. The profit-maximizing entrepreneur will determine each factor input in such a way that the sales proceeds from the last unit are equal to the given factor price. This implies that at the point of minimum cost the ratio of factor prices is equal to the ratio of what today would be called their marginal products. The word ‘marginal’ does not occur, but the expressions ‘margin’ or ‘limit’ are constantly used.

From the laws that govern actual distribution, Thünen, deeply concerned with the ‘social problem’, proceeded to the laws that ought to govern it. This led him to the most controversial of his achievements, his famous ‘natural wage’ formula. In Thünen's economy, per capita output, p (measured in rye), depends on capital per worker, q (measured in terms of the tools a worker can make in a year). Output is divided between the wage, w, and the rental on capital, r, according to $p(q)=w+rq$. In sharp contrast to later notions, savings are supposed to come out of wages while property income is consumed. Specifically, savings are the excess of wages over some subsistence minimum, a. The economy is growing by the construction of new farms at the rent-less margin of cultivation.

With interest rate $(p-w)/wq$, the return on savings is

$$R = \frac{p-w}{wq}(w-a).$$

Thunen's ‘natural wage’ maximizes R on the assumption of fixed q (and thus p). By equating dR/dw to zero, the natural wage is easily determined as the geometric mean of p and a, $w = \sqrt{pa}$.

Thünen's cumbersome exposition has given rise to many misunderstandings. Some (including Marshall) argued that the correct interest rate would have been $(p-w)/q$. In this case, the natural wage turns out to be the arithmetic mean $w = \frac{1}{2}(p+a)$ (as already suggested by Knapp). This criticism would be valid for a one-sector economy in which q is simply a stock of rye. Actually Thünen (as noted by Samuelson) considers a (rudimentary) two-sector economy in which capital goods are produced by labour only (at constant cost). Thünen is right, therefore, in valuing q at the wage rate w.

Another objection (raised, among others, by Wicksell and strongly reiterated by Samuelson) concerns the postulated constancy of q. After all, an increase in w presumably leads to an increase in q (and thus in p). Thünen anticipated this objection, for he supplemented his mathematical derivation, both verbally and by numerical examples, with a cogent explanation of how the overall maximum of R is to be found by searching over different q (and thus p). In fact, if output and marginal productivity wages are allowed to adjust to changes in q, the necessary condition for a maximum, as Dorfman (1986) showed, is again Thünen's square-root formula.

Many have thought Thünen's natural wage to be inconsistent with his own marginal productivity theory. If wages correspond to the marginal product of labour, how can they at the same time be expected to conform to some particular social ideal? This objection, however, loses its force once it is realized that Thünen (as observed by Dickinson) determined the capital/labour ratio at which the marginal productivity wage happens to be equal to his natural wage.

The fundamental objection to the natural wage formula is that it makes no sense for workers to be interested in the returns on their savings only. What Thünen seems to have been groping for, more than a century before Phelps et al., was a ‘Golden Rule’ of capital accumulation leading to some sort of optimal growth path. (In a one-sector model, the arithmetic variant of the natural wage has indeed such properties; they are analysed in Samuelson, 1986.) He never got it right; in such an optimization problem, the savings parameter, a, can hardly be treated as given. Thünen regarded his formula as important enough to have it engraved on his tombstone in the churchyard of Belitz. It commemorates a brilliant failure.

Part III of the ‘Isolated State’ is concerned with the efficiency of forest management, thereby extending the incomplete treatment in Part I. The detailed analysis of the optimal spacing of trees is of interest mainly to forest engineers. In analysing the optimal rotation period, however, Thünen makes another important contribution to economic theory. He had already pointed out in Part I that the value of a forest should not be measured by the sales value of the timber if the trees are cut today, but rather by the present value of the timber if the trees are cut and sold at the end of the optimal rotation period. In an efficient operation, the latter exceeds the former; if not, the trees should be cut at once. Efficient forest management is thus interpreted as a problem of capital and interest, providing economic theory with one of its most fruitful paradigms.

Thünen's optimality criterion, in contrast to Wicksell and Fisher, is not the equality of the marginal product of capital and the rate of interest, which, by disregarding the value of land, results in cutting trees too late. As Manz (1986) has shown, Thünen was probably the first to use the correct criterion of maximal land rent, which shortly afterwards was so brilliantly developed by Faustmann. The formula derived in Part III is flawed by incorrect discounting, and the exposition is clumsy. Nevertheless, with respect to substantive content, the capital theory implied in Thünen's forest model is superior to Böhm-Bawerk's and it was not surpassed in economic science before Wicksell.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- location theory
- marginal productivity theory
- monocentric versus polycentric models in urban economics

Selected works

1826–63. Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie. Pt I: Untersuchungen über den Einfluss, den die Getreidepreise, der Reichtum des Bodens und die Abgaben auf den Ackerbau ausüben. Hamburg: Perthes, 1826. 2nd edn, Rostock: Leopold, 1842. Pt II: Der naturgemässe Arbeitslohn und dessen Verhältniss zum Zinsfuss und zur Landrente. Rostock: Leopold; 1st section 1850; 2nd section 1863. Pt III: Grundsätze zur Bestimmung der Bodenrente, der vorteilhaftesten Umtriebszeit und des Werts der Holzbestände von verschiedenem Alter für Kieferwaldungen, Rostock: Leopold, 1863.

1951. *Ausgewählte Texte*. Ed. Walter Braeuer, Die grossen Sozialökonomen, Vol. VII. Meisenheim: Hain, 1951. English translations: Pt I: *Von Thünen's Isolated State*. Ed. Peter Hall, trans. Carla M. Wartenberg, Oxford: Pergamon Press, 1966. Pt II, section 1, in Bernard W. Dempsey, *The Frontier Wage*, Chicago: Loyola University Press, 1960.

Bibliography

Braeuer, W. 1950. Der Mathematiker-Oekonom. Zur Erinnerung an Johann Heinrich von Thünen. *Kyklos* 4(2/3), 150–71.

Buhr, W. 1983. Mikroökonomische Modelle der von Thünenschen Standorttheorie. *Zeitschrift für Wirtschafts- und Sozialwissenschaften* 103(3), 587–627.

Bulow, F. 1958. Johann Heinrich von Thülow, F. 1958. Johann Heinrich von Thünen als forstwirtschaftlicher Denker. Zur Erinnerung an den 175. Geburtstag Johann Heinrich von Thünens am 24. Juni 1783. *Weltwirtschaftliches Archiv* 80(2), 183–233.

Dickinson, H.D. 1969. Von Thünen's economics. *Economic Journal* 79(316), December, 894–902.

Dorfman, R. 1986. Comment: P.A. Samuelson, ‘Thünen at two hundred’. *Journal of Economic Literature* 24.

Ehrenberg, R. 1905a. Johann Heinrich von Thünen. *Thünen-Archiv. Organ für exakte Wirtschaftsforschung* Vol. 1.

Ehrenberg, R. 1905b. Thünen's erste wirtschaftswissenschaftliche Studien. *Thünen-Archiv. Organ für exakte Wirtschaftsforschung* Vol. 1.

Ehrenberg, R. 1905c. Thünen und Thaer. *Thünen-Archiv Organ für exakte Wirtschaftsforschung* Vol. 1.

Ehrenberg, R. 1909. Entstehung und Wesen der wissenschaftlichen Methode Johann Heinrich von Thünen's. *Thünen-Archiv. Organ für exakte Wirtschaftsforschung* Vol. 2.

Franz, G., ed. 1958. Johann Heinrich von Thünen. *Zeitschrift für Agrargeschichte und Agrarsoziologie* (Special issue) 6, Frankfurt: DLG.

Knapp, G.F. 1865. *Zur Prüfung der Untersuchungen Thünen's über Lohn und Zinsfuss im isolierten Staate*. Braunschweig: Vieweg.

Krzymowski, R. 1928. Graphical presentation of Thünen's theory of intensity. *Journal of Farm Economics* 10 (October), 461–82.

Leigh, A.H. 1946. Von Thünen's theory of distribution and the advent of marginal analysis. *Journal of Political Economy* 54 (December), 481–502.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Manz, P. 1986. Forestry economics in the steady-state. The contribution of JH von Thünen. *History of Political Economy* 18(2), Summer, 281–90.

Moore, H.L. 1895. Von Thünen's theory of natural wages. *Quarterly Journal of Economics* 9, April, 291–304; July, 388–408.

Salin, E. 1926. Der isolierte Staat 1826–1926. *Zeitschrift für die gesamte Staatswissenschaft* 81(3), 410–31.

Samuelson, P.A. 1983. Thünen at two hundred. *Journal of Economic Literature* 21(4), December, 1468–88.

Samuelson, P.A. 1986. Yes to Robert Dorfman's vindication of Thünen's natural-wage derivation. *Journal of Economic Literature* 24.

Schneider, E. 1959. Johann Heinrich von Thünen und die Wirtschaftstheorie der Gegenwart. *Schriften des Vereins für Sozialpolitik* 14, 14–28.

Schumacher, H. 1868. *Johann Heinrich von Thünen. Ein Forscherleben*. Rostock: Leopold.

Seedorf, W. and Seraphim, H.-J., eds. 1933. *Johann Heinrich von Thünen zum 150. Geburtstag. Versuch der Würdigung einer Forscherpersönlichkeit*. Rostock: Hinstorffs.

Woermann, E. 1959. Johann Heinrich von Thünen und die landwirtschaftliche Betriebslehre der Gegenwart. *Schriften des Vereins für Sozialpolitik* 34, 28–45.

How to cite this article

Niehans, Jürg. "Thünen, Johann Heinrich von (1783–1850)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000054> doi:10.1057/9780230226203.1702

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tiebout hypothesis

Dennis Epple
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tiebout (1956) argued that efficient local public good provision would emerge as households choose among communities offering different local public goods bundles. This article highlights research linking the Tiebout hypothesis to real-world local political jurisdictions, the first such link having been forged by the pioneering contribution of Oates (1969). Subsequent research has studied voting over tax and expenditure policies within municipalities in a metropolitan area, and sorting of the metropolitan population both within and across those municipalities. This research has provided the foundation for econometric analysis and policy applications.

Keywords

clubs; collective choice; demography; equity vs efficiency; exit and voice; fiscal zoning; income stratification; jurisdictional competition; land use planning; local government; local public goods; multi-community equilibrium; neighbourhood effects; peer effects; school choice; school vouchers; Tiebout hypothesis; Tiebout, C.

Article

In his famous paper, Charles Tiebout (1956) argued that there were realistic conditions under which local public goods would be provided efficiently; an efficient allocation would emerge as each household selected the community providing the public good levels most closely aligned to its preferences. This has come to be known as the ‘Tiebout hypothesis’ and the related Tiebout community-choice mechanism has been dubbed ‘voting with your feet’. This article focuses on research linking the Tiebout hypothesis to local political jurisdictions.

Oates (1969) gave Tiebout's hypothesis empirical content. He reasoned that if households selected among communities in the way Tiebout conjectured, ‘capitalization’ should result. That is, *ceteribus paribus*, housing prices should be higher in communities with high levels of public good provision and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
lower in communities with high tax rates. Oates tested and found support for these predictions using data for municipalities in New Jersey. His paper led to an explosion of research on capitalization and launched research into a variety of related aspects of the Tiebout hypothesis.

Choice within jurisdictions

Tiebout was largely silent about how communities would settle on their levels of public good provision, though he emphasized parallels with market provision. Study of this market-based approach is largely the domain of the theory of clubs. Another approach, initiated by Barr and Davis (1966), focuses on collective choice within communities. In the terms of Albert Hirschman (1970), Tiebout emphasized ‘exit’ while Barr and Davis emphasized ‘voice’. Much research followed. Bergstrom and Goodman (1973) formalized estimation of demand for local public goods. Romer and Rosenthal (1979) investigated the role of agenda setters. Micro-level estimation of demand for local public goods was undertaken by Bergstrom, Rubinfeld and Shapiro (1982).

Goldstein and Pauly (1981) observed that neglect of self-selection of households into communities would potentially bias estimates of demands for local public goods. Work followed linking intra-community choice and inter-community choice, beginning with Rubinfeld, Shapiro and Roberts (1987), and continuing with research on models of multi-community equilibrium.

Equilibrium among jurisdictions

Not surprisingly, households with higher incomes tend to prefer communities with high levels of local public goods. It is natural to ask whether income stratification can be sustained in equilibrium. Building on the work of Ellickson (1971), Westhoff (1977) proves existence of equilibrium in a model with sorting across communities and voting within communities. Westhoff's model is extended to incorporate housing markets by Epple, Filimon and Romer (1984), who demonstrate that income-stratified equilibria can be sustained by differentials across communities in the price per unit housing of services. Fernandez and Rogerson (1998) add an important dynamic feature, with community education spending by each generation affecting incomes of the succeeding generation.

While households tend to sort by income across communities, there is much income variation within jurisdictions, even within small neighbourhoods (Hardman and Ioannides, 2004). Several approaches seek to capture this intra-community heterogeneity as an outcome in multi-community equilibrium. One approach emphasizes heterogeneity in preferences as well as incomes. Structural estimation of multi-community equilibrium models embodying such heterogeneity is undertaken by Epple and Sieg (1999) and Epple, Romer and Sieg (2001). This framework is applied to study large-scale policy change by Sieg et al. (2004).

An alternative approach emphasizing heterogeneity and durability of housing is developed by Nechyba (1997). Nechyba has extended and applied this framework to study important policy issues, with particular emphasis on school choice and vouchers (Nechyba, 2000). Structural estimation taking Nechyba's model as a point of departure is undertaken by Ferreyra (2005) who extends the model to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
include heterogeneity in household tastes, including tastes for sectarian and non-sectarian schools. Still another approach, by Bayer, McMillan and Rueben (2004), permits detailed investigation of the way a household's own demographic characteristics affect preferences with respect to the demographic composition of communities and the quality of local public goods. Bayer, Ferreira, and McMillan (2005) apply this framework to estimate household preferences for community composition and to investigate the extent to which sorting within a metropolitan area is driven by preferences for education.
Residents of a jurisdiction may affect the public goods provided therein via ‘neighbourhood effects’ and ‘peer effects’. Peer effects are introduced into a multi-jurisdictional model by deBartolome (1990). While research on the Tiebout model emphasizes choice among municipalities in a metropolitan area, there is also population sorting within municipalities, especially central cities. Benabou (1996) and Durlauf (1996) study how peer effects influence such sorting, and the economic consequences of such sorting. Multi-community models increasingly emphasize peer effects (Nechyba, 2000; Epple and Romano, 2003; Bayer, McMillan and Rueben, 2004; Rothstein, 2006; Sethi and Somanathan, 2004; Ferreyra, 2005).

Equity and efficiency

Local governments impose many restrictions on land use (Fischel, 1985). Hamilton (1975) emphasizes the potential efficiency-enhancing role of zoning. Other research investigates ‘fiscal zoning’, the allegation that jurisdictions use zoning to restrict entry by households who would contribute less in taxes than the cost of public services they would consume. Early contributions are in Mills and Oates (1975). Multi-community models with community residents choosing zoning by majority rule are developed by Fernandez and Rogerson (1997) and Calabrese, Epple and Romano (2005). Computational results in the latter reveal that zoning can enhance efficiency, but also support critics who argue that fiscal zoning benefits wealthy households at the expense of poorer households. Henderson (1985) emphasizes the role of the private sector in community development. Henderson and Thisse (2001) focus on the role of developers in determining the character of the housing stock in communities. Glaeser and Gyourko (2002) conclude that land use restrictions play a major role in driving up housing prices in some areas of the United States, particularly California and some eastern cities.
The essence of the Tiebout hypothesis is that localities provide differing public good bundles to reflect variation within the population in tastes and incomes. Education is arguably the most important locally provided good, and efficiency arguments favouring decentralized provision in Tiebout equilibrium lie in uneasy juxtaposition with equity arguments favouring more centralized provision to increase equality of educational opportunity. US Courts have mandated intervention in many states to achieve greater equality of spending (Evans, Murray and Schwab, 1998). There is growing emphasis in research (Duncombe and Yinger, 1998) and the courts on policies designed to yield greater equality of educational outcomes. There is also increasing emphasis on incentive systems that might stimulate efficient provision of public education (Ladd, 1996) and increasing recognition that interventions intended to achieve greater equity may affect both political support for public education and effectiveness of provision. Oates (2006, p. 42) puts the matter succinctly: ‘There seems to be an inevitable tension here.’ Policies promoting equity need to be designed to harness local incentives for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
effective provision while also recognizing their impact on political support for public provision.

See Also

- clubs
- educational finance
- local public finance
- social interactions (empirics)
- social interactions (theory)
- urban economics
- urban political economy

In writing this article, I have benefited greatly from Oates (2006) and Fischel (2006). More extensive treatment is provided in Ross and Yinger (1998), Scotchmer (2002) and Epple and Nechyba (2004).

Bibliography

Barr, J. and Davis, O. 1966. An elementary political and economic theory of expenditures of state and local governments. *Southern Economic Journal* 33, 149–65.

Bayer, P., Ferreira, F. and McMillan, R. 2005. Tiebout sorting, social multipliers and the demand for school quality. Working paper, Yale University.

Bayer, P., McMillan, R. and Rueben, K. 2004. An equilibrium model of sorting in an urban housing market. Working Paper No. 10865. Cambridge, MA: NBER.

Benabou, R. 1996. Equity and efficiency in human capital investment: the local connection. *Review of Economic Studies* 63, 237–64.

Bergstrom, T. and Goodman, R. 1973. Private demands for public goods. *American Economic Review* 63, 280–96.

Bergstrom, T., Rubinfield, D. and Shapiro, P. 1982. Micro-based estimates of demand functions for local school expenditures. *Econometrica* 50, 1183–205.

Calabrese, S., Epple, D. and Romano, R. 2005. On the political economy of zoning. Working paper, Carnegie Mellon University.

deBartolome, C. 1990. Equilibrium and inefficiency in a community model with peer group effects. *Journal of Political Economy* 98, 110–33.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Duncombe, W. and Yinger, J. 1998. School finance reform: aid formulas and equity objectives. *National Tax Journal* 51, 239–62.

Durlauf, S. 1996. A theory of persistent income inequality. *Journal of Economic Growth* 1, 75–93.

Ellickson, B. 1971. Jurisdictional fragmentation and residential choice. *American Economic Review Papers and Proceedings* 61, 334–9.

Epple, D., Filimon, R. and Romer, T. 1984. Equilibrium among local jurisdictions: toward an integrated treatment of voting and residential choice. *Journal of Public Economics* 24, 281–308.

Epple, D. and Nechyba, T. 2004. Fiscal decentralization. In *Handbook of Regional and Urban Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Epple, D. and Romano, R. 2003. Neighborhood schools, choice, and the distribution of educational benefits. In *The Economics of School Choice*, ed. C. Hoxby. Cambridge, MA: NBER.

Epple, D. and Sieg, H. 1999. Estimating equilibrium models of local jurisdictions. *Journal of Political Economy* 107, 645–81.

Epple, D., Romer, T. and Sieg, H. 2001. Interjurisdictional sorting and majority rule: an empirical analysis. *Econometrica* 69, 1437–65.

Evans, W., Murray, S. and Schwab, R. 1998. Education finance reform and the distribution of education resources. *American Economic Review* 88, 789–812.

Fernandez, R. and Rogerson, R. 1997. Keeping people out: income distribution, zoning, and the quality of public education. *International Economic Review* 38, 23–42.

Fernandez, R. and Rogerson, R. 1998. Public education and income distribution: a dynamic quantitative evaluation of education-finance reform. *American Economic Review* 88, 813–33.

Ferreyra, M. 2005. Estimating the effects of private school vouchers in multi-district economies. Working paper, Carnegie Mellon University.

Fischel, W. 1985. *The Economics of Zoning Laws: A Property Rights Approach to American Land Use Controls*. Baltimore: Johns Hopkins University Press.

Fischel, W. 2006. Footloose at fifty: an introduction to the Tiebout anniversary essays. In *The Tiebout Model at Fifty: Essays in Public Economics in Honor of Wallace Oates*, ed. W. Fischel. Cambridge, MA: Lincoln Institute of Land Policy.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Glaeser, E. and Gyourko, J. 2002. The impact of zoning on housing affordability. NBER Working Paper No. 8835. Cambridge, MA: NBER.

Goldstein, G. and Pauly, M. 1981. Tiebout bias on the demand for local public goods. *Journal of Public Economics* 16, 131–43.

Hamilton, B. 1975. Zoning and property taxation in a system of local governments. *Urban Studies* 12, 205–11.

Hardman, A. and Ioannides, Y. 2004. Neighbors' income distribution: economic segregation and mixing in US urban neighborhoods. *Journal of Housing Economics* 13, 368–82.

Henderson, J. 1985. The Tiebout model: bring back the entrepreneurs. *Journal of Political Economy* 93, 248–64.

Henderson, J. and Thisse, J.-F. 2001. On strategic community development. *Journal of Political Economy* 109, 546–69.

Hirschman, A. 1970. *Exit, Voice, and Loyalty*. Cambridge, MA: Harvard University Press.

Ladd, H. 1996. *Holding Schools Accountable: Performance-Based Reform in Education*. Washington, DC: Brookings Institution.

Mills, E. and Oates, W., eds. 1975. *Fiscal Zoning and Land Use Controls*. Lexington, MA: Heath-Lexington Books.

Nechyba, T. 1997. Existence of equilibrium and stratification in local and hierarchical public good economies with property taxes and voting. *Economic Theory* 10, 277–304.

Nechyba, T. 2000. Mobility, targeting and private school vouchers. *American Economic Review* 90, 130–46.

Oates, W. 1969. The effects of property taxes and local public spending on property values: an empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy* 77, 957–71.

Oates, W. 2006. The many faces of Tiebout. In *The Tiebout Model at Fifty: Essays in Public Economics in Honor of Wallace Oates*, ed. W. Fischel. Cambridge, MA: Lincoln Institute of Land Policy.

Romer, T. and Rosenthal, H. 1979. Bureaucrats versus voters: on the political economy of resource allocation by direct democracy. *Quarterly Journal of Economics* 93, 563–87.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ross, S. and Yinger, J. 1998. Sorting and voting: a review of the literature on urban public finance. In *Handbook of Regional and Urban Economics*, vol. 3, ed. P. Cheshire and E. Mills. Amsterdam: North-Holland.

Rothstein, J. 2006. Good principals or good peers? Parental valuation of school characteristics, Tiebout equilibrium, and the incentive effects of competition among juridisctions. *American Economic Review* 96, 1333–50.

Rubinfeld, D., Shapiro, P. and Roberts, J. 1987. Tiebout bias and the demand for local public schooling. *Review of Economics and Statistics* 69, 426–37.

Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of Public Economics*, vol. 4, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Sethi, R. and Somanathan, R. 2004. Inequality and segregation. *Journal of Political Economy* 112, 1296–321.

Sieg, H., Smith, V., Banzhaf, H. and Walsh, R. 2004. Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *International Economic Review* 45, 1047–77.

Tiebout, C. 1956. A pure theory of local public expenditures. *Journal of Political Economy* 64, 416–24.

Westhoff, F. 1977. Existence of equilibrium in economies with a local public good. *Journal of Economic Theory* 14, 84–112.

How to cite this article

Epple, Dennis. "Tiebout hypothesis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000056> doi:10.1057/9780230226203.1703

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tiebout, Charles Mills (1924 – 1968)

Murray Milgate
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

foreign-trade multiplier; local public finance; public goods; regional development; Tiebout, C. M.; urban economics

Article

Tiebout was born in Norwalk, Connecticut, took his BA from Wesleyan University in 1950, and his Ph. D. from the University of Michigan in 1957. After holding appointments at Northwestern (1954–8) and the University of California at Los Angeles (1958–62), he became professor of economics and business administration at the University of Washington at Seattle in 1962. He died on 16 January 1968. By far his most important work was his ‘Pure Theory of Local Public Expenditure’, which appeared in the October number of the *Journal of Political Economy* for 1956, from which is derived the so-called Tiebout hypothesis, which is the subject of a separate article in this Dictionary. However, his work on problems in regional and urban economics was more extensive than this one theoretical article on the local provision of public goods might suggest.

For example, in the volume of the *Journal of Political Economy* that carried his essay on public goods, there also appeared a paper which examined the effects of export growth on the pattern of regional economic development. This analysis represents an attempt to apply a Keynesian model of income determination to regional development. Tiebout argued that exports are only one of a number of sources that act to determine the growth of regional income, and through what appears to be the first application of the foreign-trade multiplier to regional analysis he attempts to reach conclusions as to the relative significance of regional exports vis-à-vis regional demand as a source of income generation. Principal among these is that export-led regional growth is likely to be most effective when the regional base is small.

This paper was followed by work on the construction and use of regional and inter-regional input–output models (1957), which itself produced empirical investigations into the regional distribution of economic activity in the American states of California (1963) and of Washington (1969). The results of these studies were still appearing after Tiebout's early death – especially to be noted in this regard is his inter-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

regional input–output model of the linkages between the economies of Washington and California (1970). Add to this his work on the regional impact of the federal government's dispensation of its defence and space budgets (1964), and it becomes fairly clear that to record Tiebout's name solely in regard to his contribution to the pure theory of public goods would be to present a rather one-sided picture of his scientific interests.

Selected works

1956a. A pure theory of local public expenditures. *Journal of Political Economy* 64, 416–24.

1956b. Exports and regional economic growth. *Journal of Political Economy* 64, 160–4.

1957. Regional and inter-regional input–output models: an appraisal. *Southern Economic Journal* 24, 140–7.

1960a. Community income multipliers: a population growth model. *Journal of Regional Science* 2, 75–84.

1960b. Economies of scale and metropolitan governments. *Review of Economics and Statistics* 42, 442–4.

1963. (With W.L. Hansen.) An intersectoral flows analysis of the Californian economy. *Review of Economics and Statistics* 45, 409–18.

1964. (With R.S. Peterson.) Measuring the impact of regional defence-space expenditures. *Review of Economics and Statistics* 46, 421–8.

1969. An empirical regional input-output model: the state of Washington. *Review of Economics and Statistics* 51, 334–40.

1970. Inter-regional input-output: an empirical California–Washington model. *Journal of Regional Science* 10, 133–52.

How to cite this article

Milgate, Murray. "Tiebout, Charles Mills (1924–1968)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000057>
doi:10.1057/9780230226203.1704

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

time consistency of monetary and fiscal policy

Paul Klein
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Why do even benevolent policymakers frequently break their promises? Kydland and Prescott (1977) discovered that, when outcomes depend on expectations, rational policy choices typically depend on whether (*a*) the policymaker takes into account the constraint that the expected policy is the actual policy or (*b*) she takes expectations as given. A government that *commits* itself to a policy takes this constraint into account, a government that acts at its discretion does not. Since the commitment policy leads to a better outcome, there is the temptation to announce it and then to abandon this policy. This is the time inconsistency problem.

Keywords

asymmetric information; central bank independence; commitment; expectations; Friedman rule; government budget constraint; inflation; inflationary expectations; Markov perfect equilibrium; Phillips curve; public debt; Ramsey equilibrium; rational expectations; reputation; second best; sticky prices; sustainable equilibrium; time consistency of monetary and fiscal policy; time inconsistency

Article

A (possibly time- and state-contingent) strategy is said to be time inconsistent if an agent finds it optimal from the point of view of some initial period 0 but finds it suboptimal in some subsequent period *t*. Time inconsistency can obviously arise if the government has time-varying preferences because of alternations of government, as shown in Persson and Svensson (1989). However, as Kydland and Prescott (1977) discovered, the time-inconsistency problem is a pervasive feature of environments with a single benevolent policymaker taking decisions over time. This happens even though the policymaker has stable preferences and even in situations where there is no apparent conflict of interest – though the emphasis there should perhaps be on the word ‘apparent’. This means that everyone in the economy can often be made better off if the policymaker gains access to a commitment technology – a mechanism

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that forces him or her to keep his or her promises.

As pointed out by Fischer (1980), what these environments have in common is (*a*) that the Pareto optimal allocation is not implementable, (*b*) that the behaviour of the private sector depends on its expectations about future government behaviour, and (*c*) that the government and a typical individual do not share the same preferences. The third of these sounds stronger than it is. It is consistent with the benevolence of the policymaker, that is, that she maximizes the utility of a representative individual. All that it requires is a minimal amount of selfishness on the part of a typical individual – for example, she doesn't internalize the government's budget constraint.

It is important to note that either (*a*) or (*b*) on its own is not problematic. If the Pareto optimum can be achieved and I declare a policy at the beginning of time that is consistent with this Pareto optimum, I have no reason to deviate in the future since nothing better is feasible. On the other hand, suppose the Pareto optimum cannot be achieved but current behaviour does not depend on future policy. Then the policymaker will make the right trade-off in each period and the second best will be achieved. However, if both features are present at the same time, then the policy that achieves the second best will typically be time inconsistent: when choosing policy for period *t*, the policymaker faces different incentives in period 0 from the incentives faced in period *t*. In period 0 she rationally takes into account the effects on expectations; in period *t* it is no longer rational to do that, since expectations in the past are bygones. The temptation to bring the economy closer to the first-best renders the second-best solution time inconsistent, and rational expectations force the economy into a Pareto inferior third-best equilibrium.

The Phillips curve and inflation bias

The central example in Kydland and Prescott (1977) is a central bank setting inflation in an environment with an expectations-augmented Phillips curve. This environment has the feature that inflation surprises lead to deviations of output from its ‘natural’ rate. The authors then assume that a positive inflation surprise is good, since the ‘natural’ rate of output is suboptimal because of various (unspecified) distortions. However, in any rational expectations equilibrium, inflation expectations are fulfilled and output equals its natural rate. If inflation is bad in itself (other things being equal), then the best rational expectations equilibrium features zero expected and actual inflation. This is the second best: the best equilibrium that can be achieved subject to the constraint of rational expectations. The first-best would be for output to be at some ideal level greater than its natural rate but with inflation maintained at zero. However, this ideal outcome is not consistent with rational expectations. Now suppose inflation policy can be revised after expectations are formed. Then the zero-inflation policy that gives rise to the best rational expectations equilibrium is not time consistent: when it is time to set inflation, the central bank would like to set an inflation rate above zero since a small rise in inflation has no first-order effect on the welfare cost of inflation but does have a first-order effect on the welfare cost of output being less than its ideal level. The result, under discretionary monetary policy, is a tendency for inflation to be above its desired level (inflation bias) with no (positive) effect on output or employment.

Overtaxation of capital and liquidity

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Another early contribution is that of Fischer (1980), who discussed a situation where a fiscal authority decides on the levels of taxation on labour and capital and of public expenditure. There, the problem is that tomorrow's capital stock depends on today's investment. Meanwhile, today's investment depends on expectations about future capital income taxes. The result is that a government that sets taxes sequentially will typically overtax capital income. Similarly, Calvo (1978) described the time consistency problem in a monetary economy with money in the utility function. He found that, when lump-sum taxation is not available, the Friedman rule of optimal monetary policy is not time consistent; in general, the government wants to expand the money supply to relax the government budget constraint. This is because monetary expansion, like capital taxation, is distortionary only *ex ante*; it is the *expectation* of monetary expansion (and the consequent inflation) that leads people to economize on liquidity, a socially free resource.

Relation with game theoretic concepts

The relationship between time consistency, time inconsistency and various concepts in game theory have been much discussed. A common view, but by no means a consensus, has emerged, asserting that the best way to think about the situation is that the (Ramsey 1927) optimal policy, ‘second best’ or ‘commitment solution’ (these phrases are used interchangeably) is an equilibrium of one game, the time-consistent policy the equilibrium of another. The first game lets the government move before time starts, choosing a time- and state-contingent policy for the indefinite future. Thereafter it does not move again. The second game has the government moving sequentially, setting policy in each period as it arrives. When the policies implied by these equilibria differ, then we say that the Ramsey policy is time inconsistent. On the other hand, any equilibrium of the second game is time-consistent by construction. This view of course leaves open what the correct solution concept is for these various games. Chari and Kehoe (1990) and several successors discuss the appropriate solution concept for the second type of game. From this literature has emerged the concept of ‘sustainable equilibrium’ which roughly corresponds to the sequential equilibrium concept of Kreps and Wilson (1982) but modified to apply to economies with one large agent and many ‘small’ agents. A recent formulation can be found in Phelan and Stacchetti (2001).

Solving the time inconsistency problem

The literature on time consistency may usefully be divided into two parts: one attempts to characterize the equilibrium of the sequential-move game, the other tries to solve the time-inconsistency problem by somehow erasing the difference between the Ramsey policy and the time-consistent policy. A celebrated paper in the second category is Lucas and Stokey (1983), which discusses a dynamic monetary economy without capital. A price-taking representative agent chooses labour supply and the government sets labour taxes so as to minimize distortions subject to a government budget constraint. Government expenditure is exogenous. The government can issue state-contingent debt which it is committed to honouring. The main finding of the paper is that, if public debt has a sufficiently rich maturity structure, then the Ramsey optimal policy is time consistent. However, if only one-period state-contingent bonds

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
are available, then the optimal policy is typically not time consistent. Persson, Persson and Svensson (1987) extend this result to monetary economies, showing how the government can render the optimal policy time consistent by accumulating nominal assets equal to the stock of money so that the outstanding stock of net nominal claims is zero. A minor mistake in that paper was pointed out by Calvo and Obstfeld (1990), but the basic result stands and applies quite generally, as explained in Alvarez, Kehoe and Neumayer (2001). The latter paper establishes a link between the optimality of the Friedman rule (zero nominal interest rates) and the possibility of rendering the Ramsey optimal policy time-consistent: for a wide class of economies they show that the Ramsey optimal policy can be made time-consistent if and only if the Friedman rule is optimal. Domínguez (2007) extends this result further for an economy with capital, showing that, if capital taxes are set one period in advance, then a sufficiently rich maturity structure of public debt is sufficient to render the optimal policy time consistent. Another sub-literature in this category looks at reputational mechanisms that might render the optimal policy time consistent, or at least bring the time-consistent solution closer to the second-best optimum. In monetary policy, a key early contribution is Barro and Gordon (1983). In an environment with an expectations-augmented Phillips curve, it shows, using the well-known folk theorem from game theory, that the optimal monetary policy in the environment described by Kydland and Prescott (1977) can be sustained as a time-consistent equilibrium provided the policymaker is patient enough. A paper that analyses fiscal policy with a reputational-style approach is Kotlikoff, Persson and Svensson (1988). In this paper it is shown how the optimal tax scheme can be sustained in a two-period overlapping generations environment by threats of moving to the third-best equilibrium if any generation deviates. Yet another set of solutions to the time-consistency problem of monetary policy is found in Rogoff (1985) and Persson and Tabellini (1993). The first, using an idea from industrial organization first published by Vickers (1985), shows that a monetary policymaker will typically be better off delegating monetary policy to a central banker that cares more about low inflation and less about output or employment than he or she does. Rogoff's result is aptly described as delegation to a ‘conservative central banker’. This delegation improves welfare but does not achieve the second-best optimum. By contrast, the key result in Persson and Tabellini (1993) is that the second best can be achieved by signing a performance contract with the central banker.

Analysing what happens when the time inconsistency problem cannot be solved

The literature on characterizing time-consistent policy also divides into two parts: one focusing on a solution concept (‘sustainable equilibrium’) that is nearly always set-valued and the other on a refinement (‘Markov perfect equilibrium’) that often (but certainly not always) yields a unique equilibrium. The concept of Markov perfect equilibrium, whose purpose is essentially to rule out any reputational mechanisms, is defined in a game-theoretic setting in Maskin and Tirole (2001) and in a macroeconomic setting by, among several others, Klein, Krusell and Ríos-Rull (2006). In the latter paper, the authors find that in the Markov perfect equilibrium labour tends to be under-taxed even in an environment where no other taxes are available and the only other endogenous variable is government spending. That is, a government acting sequentially tends to exaggerate the distortionary effects of taxation. This is in marked contrast to the case of capital and inflation taxes. The reason is that the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
current labour income tax *encourages* labour supply in the previous period, by intertemporal substitution. This effect is ignored by a sequentially moving government that thus neglects a beneficial effect of raising the current labour income tax rate.

Other papers studying Markov perfect equilibrium in a fiscal policy setting include Klein and Ríos-Rull (2003), who look at time-consistent capital and labour income taxation where capital taxes are set one period in advance and the budget has to balance in each period. The main finding is that a calibrated model can replicate the capital and labour taxes that we observe in, say, the United States, reasonably well. Krusell, Martin and Ríos-Rull (2004) consider public debt policy in an economy without capital and find that for positive initial debt there is a unique equilibrium but infinitely many steady states. For negative initial debt there are infinitely many equilibria, each associated with infinitely many steady states.

On the other hand, Phelan and Stacchetti (2001) study the set of sustainable equilibria in an economy with capital but without public debt. The methods used are similar to those in Abreu, Pearce and Stacchetti (1990). Fernandez-Villaverde and Tsyvinski (2002) consider all the sustainable equilibria in a stochastic environment with capital. They compare the best (from a welfare point of view) in that class with the Markov perfect equilibrium and the Ramsey equilibrium. Also, a literature is emerging on time-consistent policy under asymmetric information. Recent contributions include Sleet (2003) and Sleet and Yeltekin (2004).

A new departure in the study of time consistency of monetary policy is the consideration of the role of sticky prices. This introduces a new channel through which time inconsistency may arise: when prices are sticky and firms are bound to produce whatever is demanded at the given price, a surprise monetary expansion raises output. This is typically *ex post* welfare-improving in an economy suffering from some distortion, typically monopolistic rather than perfect competition. Important contributions include Albanesi, Chari and Christiano (2003a), who show that without commitment the economy can get stuck in an ‘expectation trap’ in the following sense. There are multiple, Pareto-ranked, equilibria. In the lower-ranked equilibria, the private sector expects high inflation. The measures the private sector takes to protect itself from high inflation create incentives for the policymaker to accommodate these expectations. On the other hand, in Albanesi, Chari and Christiano (2003b) the same authors show that optimal monetary policy is time consistent – and the Markov perfect equilibrium unique – in a wide class of models.

See Also

- central bank independence
- Phillips curve

Bibliography

Abreu, D., Pearce, D. and Stacchetti, E. 1990. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica* 58, 1041–63.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Albanesi, S., Chari, V.V. and Christiano, L. 2003a. Expectation traps and monetary policy. *Review of Economic Studies* 70, 715–41.

Albanesi, S., Chari, V.V. and Christiano, L. 2003b. How severe is the time inconsistency problem in monetary policy? *Federal Reserve Bank of Minneapolis Quarterly Review* 27(3), 17–33.

Alvarez, F., Kehoe, P.J. and Neumayer, P.A. 2001. The time consistency of fiscal and monetary policies. Staff Report No. 305, Federal Reserve Bank of Minneapolis.

Barro, R.J. and Gordon, D.B. 1983. Rules, discretion and reputation in a model of monetary policy. *Journal of Monetary Economics* 12, 101–21.

Calvo, G. 1978. On the time consistency of optimal policy in a monetary economy. *Econometrica* 46, 1411–28.

Calvo, G.A. and Obstfeld, M. 1990. Time consistency of fiscal and monetary policy: a comment. *Econometrica* 58, 1245–7.

Chari, V.V. and Kehoe, P.J. 1990. Sustainable plans. *Journal of Political Economy* 98, 784–802.

Domínguez, B. 2007. On the time-consistency of optimal capital taxes. *Journal of Monetary Economics* 54, 686–705.

Fernandez-Villaverde, J. and Tsyvinski, A. 2002. *Optimal fiscal policy in a business cycle model without commitment*. Mimeo: University of Pennsylvania.

Fischer, S. 1980. Dynamic inconsistency, cooperation and the benevolent dissembling government. *Journal of Economic Dynamics and Control* 2, 93–107.

Klein, P., Krusell, P. and Ríos-Rull, J.-V. 2006. *Time-consistent public expenditure*. Mimeo: Princeton University.

Klein, P. and Ríos-Rull, J.-V. 2003. Time-consistent optimal fiscal policy. *International Economic Review* 44, 1217–45.

Kotlikoff, L.J., Persson, T. and Svensson, L.E. 1988. Social contracts as assets: a possible solution to the time-consistency problem. *American Economic Review* 78, 662–77.

Kreps, D.M. and Wilson, R.B. 1982. Sequential equilibria. *Econometrica* 50, 863–94.

Krusell, P., Martin, F. and Ríos-Rull, J.-V. 2004. *On the determination of government debt*. Mimeo:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Simon Fraser University.

Kydland, F.E. and Prescott, E.C. 1977. Rules rather than discretion: the inconsistency of optimal plans. *Journal of Political Economy* 85, 473–91.

Lucas, R.E. and Stokey, N.L. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12, 55–93.

Maskin, E. and Tirole, J. 2001. Markov perfect equilibrium. *Journal of Economic Theory* 100, 191–219.

Persson, M., Persson, T. and Svensson, L. 1987. Time consistency of fiscal and monetary policy. *Econometrica* 55, 1419–31.

Persson, T. and Svensson, L.E. 1989. Why a stubborn conservative would run a deficit: policy with time-inconsistent preferences. *Quarterly Journal of Economics* 104, 325–45.

Persson, T. and Tabellini, G. 1993. Designing institutions for monetary stability. *Carnegie-Rochester Conference Series on Public Policy* 39, 53–84.

Phelan, C. and Stacchetti, E. 2001. Sequential equilibria in a Ramsey tax model. *Econometrica* 69, 1491–518.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37(145), 47–61.

Rogoff, K. 1985. The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100, 1169–90.

Sleet, C. 2003. Optimal taxation with private government information. *Review of Economic Studies* 71, 1217–39.

Sleet, C. and Yeltekin, S. 2004. *Credible social insurance*. Mimeo: University of Iowa.

Vickers, J. 1985. Delegation and the theory of the firm. *Economic Journal Supplement* 95, 138–47.

How to cite this article

Klein, Paul. "time consistency of monetary and fiscal policy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000219> doi:10.1057/9780230226203.1705

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

time preference

Murray N. Rothbard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Azpilcueta Navarrus, M. de; Bailey, S.; Böhm-Bawerk, E. von; capital theory; capitalization; Fetter, F. A.; Fisher, I.; Galiani, F.; interest rates; interest theory; Longfield, M.; Lottini da Volterra, G. F.; marginal productivity; Menger, C.; Mises, L. E. von; Rae, J. (1786–1872); Summenhart, C.; time preference; Turgot, A. R. J.; usury; value and distribution theory; value theory

Article

Time preference is the insight that people prefer ‘present goods’ (goods available for use at present) to ‘future goods’ (present expectations of goods becoming available at some date in the future), and that the social rate of time preference, the result of the interactions of individual time preference schedules, will determine and be equal to the pure rate of interest in a society. The economy is pervaded by a time market for present as against future goods, not only in the market for loans (in which creditors trade present money for the right to receive money in the future), but also as a ‘natural rate’ in all processes of production. For capitalists pay out present money to buy or rent land, capital goods, and raw materials, and to hire labour (as well as buying labour outright in a system of slavery), thereby purchasing expectations of future revenue from the eventual sales of product. Long-run profit rates and rates of return on capital are therefore forms of interest rate. As businessmen seek to gain profits and avoid losses, the economy will tend toward a general equilibrium, in which all interest rates and rates of return will be equal, and hence there will be no pure entrepreneurial profits or losses.

In centuries of wrestling with the vexed question of the justification of interest, the Catholic scholastic philosophers arrived at highly sophisticated explanations and justifications of return on capital, including risk and the opportunity cost of profit forgone. But they had extreme difficulty with the interest on a riskless loan, and hence denounced all such interest as sinful and usurious.

Some of the later scholastics, however, in their more favourable view of usury, began to approach a time preference explanation of interest. During a comprehensive demolition of the standard arguments for the prohibition of usury in his *Treatise on Contracts* (1499), Conrad Summenhart (1465–1511), theologian at the University of Tübingen, used time preference to justify the purchase of a discounted debt, even if

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the debt be newly created. When someone pays $100 for the right to obtain $110 at a future date, the buyer (lender) doesn't profit usuriously from the loan because both he and the seller (borrower) value the future $110 as being worth $100 at the present time (Noonan, 1957).

A half-century later, the distinguished Dominican canon lawyer and monetary theorist at the University of Salamanca, Martin de Azpilcueta Navarrus (1493–1586) clearly set forth the concept of time preference, but failed to apply it to a defence of usury. In his *Commentary on Usury* (1556), Azpilcueta pointed out that a present good, such as money, will naturally be worth more on the market than future goods, that is, claims to money in the future. As Azpilcueta put it:

> a claim on something is worth less than the thing itself, and•…•it is plain that that which is not usable for a year is less valuable than something of the same quality which is usable at once. (Gordon, 1975, p. 215)

At about the same time, the Italian humanist and politician Gian Francesco Lottini da Volterra, in his handbook of advice to princes, *Avvedimenti civili* (1574), discovered time preference. Unfortunately, Lottini also inaugurated the tradition of moralistically deploring time preference as an overestimation of a present that can be grasped immediately by the senses (Kauder, 1965, pp. 19–22).

Two centuries later, the Neapolitan abbé, Ferdinando Galiani (1728–87), revived the rudiments of time-preference in his *Della Moneta* (1751) (Monroe, 1924). Galiani pointed out that just as the exchange rate of two currencies equates the value of a present and a spatially distant money, so the rate of interest equates present with future, or temporally distant, money. What is being equated is not physical properties, but subjective values in the minds of individuals.

These scattered hints scarcely prepare one for the remarkable development of a full-scale time preference theory of interest by the French statesman, Anne Robert Jacques Turgot (1727–81), who, in a relatively few hastily written contributions, anticipated almost completely the later Austrian theory of capital and interest (Turgot, 1977). In the course of a paper defending usury, Turgot asked: why are borrowers willing to pay an interest premium for the use of money? The focus should not be on the amount of metal repaid but on the usefulness of the money to the lender and borrower. In particular, Turgot compares the ‘difference in usefulness which exists at the date of borrowing between a sum currently owned and an equal sum which is to be received at a distant date’, and notes the well-known motto, ‘a bird in the hand is better than two in the bush’. Since the sum of money owned now ‘is preferable to the assurance of receiving a similar sum in one or several years’ time’, returning the same principal means that the lender ‘gives the money and receives only an assurance’. Therefore, interest compensates for this difference in value by a sum proportionate to the length of the delay. Turgot added that what must be compared in a loan transaction is not the value of money lent with the value repaid, but rather the ‘value of the *promise* of a sum of money compared to the value of money available now’ (Turgot, 1977, pp. 158–9).

In addition, Turgot was apparently the first to arrive at the concept of *capitalization*, a corollary to time preference, which holds that the present capital value of any durable good will tend to equal the sum of its expected annual rents, or returns, discounted by the market rate of time preference, or rate of interest. Turgot also pioneered in analysing the relation between the quantity of money and interest rates. If an increased supply of money goes to low time-preference people, then the increased proportion of savings

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to consumption lowers time preference and hence interest rates fall while prices rise. But if an increased quantity goes into the hands of high time-preference people, the opposite would happen and interest rates would rise along with prices. Generally, over recent centuries, he noted, the spirit of thrift has been growing in Europe and hence time preference rates and interest rates have tended to fall.

One of the notable injustices in the historiography of economic thought was Böhm-Bawerk's brusque dismissal in 1884 of Turgot's anticipation of his own time-preference theory of interest as merely a ‘land fructification theory’ (Böhm-Bawerk, vol. 1, 1884–9). Partly this dismissal stemmed from Böhm's methodology of clearing the ground for his own positive theory of interest by demolishing, and hence sometimes doing injustice to, his own forerunners (Wicksell, 1911, p. 177). The unfairness is particularly glaring in the case of Turgot, because we now know that in 1876, only eight years before the publication of his history of theories of interest, Böhm-Bawerk wrote a glowing tribute to Turgot's theory of interest in an as yet unpublished paper in Karl Knies's seminar at the University of Heidelberg (Turgot, 1977, pp. xxix–xxx).

In the course of his demolition of the Ricardo–James Mill labour theory of value on behalf of a subjective utility theory, Samuel Bailey (1825) clearly set forth the concept of time preference. Rebutting Mill's statement that time, as a ‘mere abstract word’, could not add to value, Bailey declared that ‘we generally prefer a present pleasure or enjoyment to a distant one’, and therefore prefer present goods to waiting for goods to arrive in the future. Bailey, however, did not go on to apply his insight to interest.

In the mid-1830s, the Irish economist Samuel Mountifort Longfield worked out the later Austrian theory of capital as performing the service for workers of supplying money at present instead of waiting for the future when the product will be sold. In turn the capitalist receives from the workers a time discount from their productivity. As Longfield put it, the capitalist

> pays the wages immediately, and in return receives the value of [the worker's] labour,… [which] is greater than the wages of that labour. The difference is the profit made by the capitalist for his advances•…•as it were, the discount which the labourer pays for prompt payment. (Longfield, 1971)

The ‘pre-Austrian’ time analysis of capital and interest was most fully worked out, in the same year 1834, by the Scottish and Canadian eccentric John Rae (1786–1872). In the course of attempting an anti-Smithian defence of the protective tariff, Rae, in his *Some New Principles on the Subject of Political Economy* (1834), developed the Böhm-Bawerkian time analysis of capital, pointing out that investment lengthens the time involved in the processes of production. Rae noted that the capitalist must weigh the greater productivity of longer production processes against waiting for them to come to fruition. Capitalists will sacrifice present money for a greater return in the future, the difference – the interest return – reflecting the social rate of time preference. Rae saw that people's time preference rates reflect their cultural and psychological willingness to take a shorter or longer view of the future. His moral preferences were clearly with the low time-preference thrifty as against the high time-preference people who suffer from a ‘defect of the imagination’. Rae's analysis had little impact on economics until resurrected at the turn of the 20th century, whereupon it was generously hailed in the later editions of Böhm-Bawerk's history of interest theories (Böhm-Bawerk, vol. 1, 1959).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Time preference, as a concept and as a foundation for the explanation of interest, has been an outstanding feature of the Austrian School of economics. Its founder, Carl Menger (1840–1921), enunciated the concept of time preference in 1871, pointing out that satisfying the immediate needs of life and health are necessarily prerequisites for satisfying more remote future needs. In addition, Menger declared, ‘all experience teaches that we humans consider a present pleasure, or one expected in the near future, more important than one of the same intensity which is not expected to occur until some more distant time’ (Wicksell, 1924, p. 195; Menger, 1871, pp. 153–4). But Menger never extended time preference from his value theory to a theory of interest; and when his follower Böhm-Bawerk did so, he peevishly deleted this discussion from the second edition of his *Principles of Economics* (Wicksell, 1924, pp. 195–6).

Böhm-Bawerk's *Capital and Interest* (1884) is the *locus classicus* of the time preference theory of interest. In his first, historical volume, he demolished all other theories, in particular the productivity theory of interest; but five years later, in his *Positive Theory of Capital* (1889), Böhm brought back the productivity theory in an attempt to combine it with a time preference explanation of interest (Böhm-Bawerk, vols 1 and 2, 1959). In his ‘three grounds’ for the explanation of interest, time preference constituted two, and the greater productivity of longer processes of production the third, Böhm ironically placing greatest importance upon the third ground. Influenced strongly by Böhm-Bawerk, Irving Fisher increasingly took the same path of stressing the marginal productivity of capital as the main determinant of interest (Fisher, 1907; 1930).

With the work of Böhm-Bawerk and Fisher, the modern theory of interest was set squarely on the path of placing time preference in a subordinate role in the explanation of interest, determining only the rate of consumer loans and the supply of consumer savings, while the alleged productivity of capital determines the more important demand for loans and for savings. Hence, modern interest theory fails to integrate interest on consumer loans and producers’ returns into a coherent explanation.

In contrast, Frank A. Fetter, building on Böhm-Bawerk, completely discarded productivity as an explanation of interest and constructed an integrated theory of value and distribution in which interest is determined solely by time preference, while marginal productivity determines the ‘rental prices’ of the factors of production (Fetter, 1915; 1977). In his outstanding critique of Böhm-Bawerk, Fetter pointed out a fundamental error of the third ground in trying to explain the return on capital as ‘present goods’ earning a return for their productivity in the future; instead, capital goods are *future* goods, since they are only valuable in the expectation of being used to produce goods that will be sold to the consumer at a future date (Fetter, 1902). One way of seeing the fallacy of a productivity explanation of interest is to look at the typical practice of any current microeconomics text: after explaining marginal productivity as determining the demand curve for factors with wage rates on the *y*-axis, the textbook airily shifts to interest rates on the *y*-axis to illustrate the marginal productivity determination of interest. But the analog on the *y*-axis should not be interest, which is a ratio and not a price, but rather the *rental price* (price per unit time) of a capital good. Thus, interest remains totally unexplained. In short, as Fetter pointed out, marginal productivity determines rental prices, and time preference determines the rate of interest, while the capital value of a factor of production is the expected sum of future rents from a durable factor discounted by the rate of time preference or interest.

The leading economist adopting Fetter's pure time preference view of interest was Ludwig von Mises, in his *Human Action* (Mises, 1949). Mises amended the theory in two important ways. First, he rid the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
concept of its moralistic tone, which had been continued by Böhm-Bawerk, implicitly criticizing people for 'under'-estimating the future. Mises made clear that a positive time preference rate is an essential attribute of human nature. Secondly, and as a corollary, whereas Fetter believed that people could have either positive or negative rates of time preference, Mises demonstrated that a positive rate is deducible from the fact of human action, since by the very nature of a goal or an end people wish to achieve that goal as soon as possible.

See Also

- Böhm-Bawerk, Eugen von
- Fisher, Irving

Bibliography

Bailey, S. 1825. *A Critical Dissertation on the Nature, Measure, and Causes of Value*. New York: Kelley, 1967.

Böhm-Bawerk, E. von. 1884–9. *Capital and Interest*, vols 1 and 2, 4th edn, South Holland, IL: Libertarian Press, 1959.

Fetter, F.A. 1902. The 'roundabout process' in the interest theory. *Quarterly Journal of Economics* 17 (November), 163–80. Repr. in Fetter (1977).

Fetter, F.A. 1915. *Economic Principles*, vol. 1. New York: The Century Co.

Fetter, F.A. 1977. *Capital, Interest, and Rent: Essays in the Theory of Distribution*, ed. M. Rothbard. Kansas City: Sheed Andrews and McMeel.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Fisher, I. 1930. *The Theory of Interest*. New York: Kelley & Millman, 1954.

Gordon, B. 1975. *Economic Analysis Before Adam Smith: Hesiod to Lessius*. New York: Barnes & Noble.

Kauder, E. 1965. *A History of Marginal Utility Theory*. Princeton: Princeton University Press.

Longfield, S.M. 1971. *The Economic Writings of Mountifort Longfield*, ed. R.D.C. Black. Clifton, NJ: Kelley.

Menger, C. 1871. *Principles of Economics*. Ed. J. Dingwall and B. Hoselitz, Glencoe, IL: Free Press,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1950.

Mises, L. von. 1949. *Human Action: A Treatise on Economics*. 3rd revised edn, Chicago: Regnery, 1966.

Monroe, A., ed. 1924. *Early Economic Thought*. Cambridge, MA: Harvard University Press.

Noonan, J.T., Jr. 1957. *The Scholastic Analysis of Usury*. Cambridge, MA: Harvard University Press.

Rae, J. 1834. Some new principles on the subject of political economy. In *John Rae: Political Economist*, ed. R.W. James. Toronto: University of Toronto Press, 1965.

Turgot, A.R.J. 1977. *The Economics of A.R.J. Turgot*, ed. P.D. Groenewegen. The Hague: Martinus Nijhoff.

Wicksell, K. 1911. Böhm-Bawerk's theory of interest. In K. Wicksell, *Selected Papers on Economic Theory*, ed. E. Lindahl. Cambridge, MA: Harvard University Press, 1958.

Wicksell, K. 1924. The new edition of Menger's *Grundsatze*. In K. Wicksell, *Selected Papers on Economic Theory*, ed. E. Lindahl. Cambridge, MA: Harvard University Press, 1958.

How to cite this article

Rothbard, Murray N. "time preference." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000063> doi:10.1057/9780230226203.1706

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

time series analysis

Francis X. Diebold, Lutz Kilian and Marc Nerlove
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The analysis of economic time series is central to a wide range of applications, including business cycle measurement, financial risk management, policy analysis based on structural dynamic econometric models, and forecasting. This article provides an overview of the problems of specification, estimation and inference in linear stationary and ergodic time series models as well as non-stationary models, the prediction of future values of a time series and the extraction of its underlying components. Particular attention is devoted to recent advances in multiple time series modelling, the pitfalls and opportunities of working with highly persistent data, and models of nonlinear dependence.

Keywords

ARFIMA models; ARIMA models; ARMA models; ARMAX models; autocovariance generating functions; autoregressive conditional heteroskedasticity (ARCH); band-pass filter; co-integration; common factors; Cournot, A.; ergodicity and non-ergodicity; estimation; factor model forecasts; forecasting; generalized autoregressive conditionally heteroskedastic (GARCH); generalized method of moments; Granger, C.; heteroskedasticity; Hodrick–Prescott (HP) filter; inference; Jevons, W.; Kalman filter; least squares; linear processes; long memory models; Markov chain methods; maximum likelihood; minimum mean-square error (MMSE) criterion; multiple time series analysis; noise; nonlinear time series analysis; optimal prediction and extraction theory; prediction; principal components analysis; regime switching models; seasonal adjustment; simultaneous-equations model; Slutsky, E.; smooth transition regression models; specification; spectral analysis; spurious regressions; state-space methods; stationarity; stochastic volatility models; structural vector autoregressions; time domain analysis; time series analysis; unit root; unobserved components models; vector autoregressions; volatility dynamics; Wiener–Kolmogorov theory; Wold decomposition theorem; Yule, G.

Article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Any series of observations ordered along a single dimension, such as time, may be thought of as a time series. The emphasis in time series analysis is on studying the dependence among observations at different points in time. What distinguishes time series analysis from general multivariate analysis is precisely the temporal order imposed on the observations. Many economic variables, such as GNP and its components, price indices, sales, and stock returns are observed over time. In addition to being interested in the contemporaneous relationships among such variables, we are often concerned with relationships between their current and past values, that is, relationships over time.

The study of time series of, for example, astronomical observations predates recorded history. Early writers on economic subjects occasionally made explicit reference to astronomy as the source of their ideas. For example, Cournot (1838) stressed that, as in astronomy, it is necessary to recognize the *secular* variations which are independent of the periodic variations. Similarly, Jevons (1884) remarked that his study of short-term fluctuations used the methods of astronomy and meteorology. During the 19th century interest in, and analysis of, social and economic time series evolved into a new field of study independent of developments in astronomy and meteorology (see Nerlove, Grether and Carvalho, 1979, pp. 1–21, for a historical survey).

Harmonic analysis is one of the earliest methods of analysing time series thought to exhibit some form of periodicity. In this type of analysis, the time series, or some simple transformation of it, is assumed to be the result of the superposition of sine and cosine waves of different frequencies. However, since summing a finite number of such strictly periodic functions always results in a perfectly periodic series, which is seldom observed in practice, one usually allows for an additive stochastic component, sometimes called ‘noise’. Thus, an observer must confront the problem of searching for ‘hidden periodicities’ in the data, that is, the unknown frequencies and amplitudes of sinusoidal fluctuations hidden amidst noise. An early method for this purpose is *periodogram analysis*, suggested by Stokes (1879) and used by Schuster (1898) to analyse sunspot data and later by others, principally William Beveridge (1921; 1922), to analyse economic time series.

Spectral analysis is a modernized version of periodogram analysis modified to take account of the stochastic nature of the entire time series, not just the noise component. If it is assumed that economic time series are fully stochastic, it follows that the older periodogram technique is inappropriate and that considerable difficulties in the interpretation of the periodograms of economic series may be encountered.

At the time when harmonic analysis proved to be inadequate for the analysis of economic and social time series, another way of characterizing such series was suggested by the Russian statistician and economist, Eugen Slutsky (1927), and by the British statistician, G.U. Yule (1921; 1926; 1927). Slutsky and Yule showed that, if we begin with a series of purely random numbers and then take sums or differences, weighted or unweighted, of such numbers, the new series so produced has many of the apparent cyclic properties that were thought at the time to characterize economic and other time series. Such sums or differences of purely random numbers and sums or differences of the resulting series form the basis for the class of autoregressive moving-average (ARMA) processes which are used for modelling many kinds of time series. ARMA models are examples of time domain representations of time series. Although the latter may look very different from spectral representations of time series, there is a one-to-one mapping between time domain analysis and spectral analysis. Which approach is preferred in practice is a matter only of convenience. The choice is often determined by the transparency

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

with which a given question can be answered. The remainder of this article explores these two complementary approaches to the analysis of economic time series.

1 Basic theory

1.1 Stationarity and ergodicity of time series processes

Consider a random variable x_t where $t \in N$, the set of integers; the infinite vector $\{x_t, t \in N\}$ is called a discrete time series. Let M denote a subset of T consecutive elements of N. The distribution of the finite dimensional vector $\{x_t, t \in M\}$ is a well-defined multivariate distribution function, $F_M(\cdot)$. The time series $\{x_t, t \in N\}$ is said to be *strictly stationary* if, for any finite subset M of N and any integer τ, the distribution function of $\{x_t, t \in M + \tau\}$ is the same as the distribution function of $\{x_t, t \in M\}$. In other words, the joint distribution function of the finite vector of observations on x_t is invariant with respect to the origin from which time is measured. All the unconditional moments of the distribution function, if they exist, are independent of the index t; in particular,

$$E(x_t) = \mu$$
$$\gamma(\tau) = E[x_t - \mu][x_{t+\tau} - \mu], \quad (1)$$

where $\gamma(\tau)$ is the autocovariance function and depends only on the difference in indices, τ. Time-series processes for which (1) holds, but which are not necessarily strictly stationary according to the definition above, are said to be weakly *stationary*, *covariance stationary*, or *stationary to the second order*. Time-series processes for which $F_M(\cdot)$ is multivariate normal for any subset M of N are called *Gaussian processes*. For Gaussian processes covariance stationarity implies strict stationarity.
In practice, we usually observe only one realization of a finite subset of the time series of interest, corresponding to one of the many possible draws of length T from $F_M(\cdot)$. The question is whether the moments of x_t may be inferred from one such realization; for example, from the time averages of sums (or sums of products) of the observed values of a time series. If the process is what is known as *ergodic*, time averages of functions of the observations on the time series at T time points converge in mean square to the corresponding population expectations of x_t across alternative draws, as $T \to \infty$ (Priestley, 1981, pp. 340–3; Doob, 1953, p. 465). It is possible for a process to be stationary, yet not ergodic.
Consider, for example, the process $x_t^{(i)} = \eta^{(i)} + \varepsilon_t$, where $x_t^{(i)}$ denotes the i*th* draw for observation x_t from the universe of all possible draws for x_t. Suppose that $\eta^{(i)} \sim N(0, \lambda^2)$ is the mean of the i*th* draw and that $\varepsilon_t \sim N(0, \sigma^2)$ is independent of $\eta^{(i)}$. This process is clearly stationary in that the probability limit of the ensemble average is zero, yet the time average $\Sigma_{t=1}^{T} x_t^{(i)} / T = \eta^{(i)} + \Sigma_{t=1}^{T} \varepsilon_t / T$ converges

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

to $n^{(l)}$ rather than zero, thus violating ergodicity.

1.2 The Wold decomposition and general linear processes

Let $\{\varepsilon_t\}$ be one element of a time series of serially uncorrelated, identically distributed random variables with zero mean and variance σ^2. Then the infinite, one-sided moving average (MA) process

$$x_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}, \tag{2}$$

where $b_0 = 1$ and $\Sigma_{j=0}^{\infty} b_j^2 < \infty$, is also a well-defined stationary process with mean 0 and variance $\sigma^2 \Sigma_0^{\infty} b_j^2$. Processes of this form and, more generally, processes based on an infinite two-sided MA of the same form are called *linear processes*, are always ergodic, and play a key role in time series analysis (Hannan, 1970).

The importance of the process (2) is underscored by the Wold decomposition theorem (Wold, 1938), which states that any weakly stationary process may be decomposed into two mutually uncorrelated component processes, one an infinite one-sided MA of the form (2) and the other a so-called linearly deterministic process, future values of which can be predicted exactly by some linear function of past observations. The linearly deterministic component is non-ergodic.

2 Linear processes in time and frequency domains

2.1 Autocovariance and autocovariance generating functions

The autocovariance function of a stationary process, defined in (1) above, or its matrix generalization for vector processes, provides the basic representation of time dependence for weakly stationary processes. For the stationary process defined in (2), it is

$$\gamma(\tau) = \sigma^2 \sum_{j=0}^{\infty} b_j b_{j+\tau}. \tag{3}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Let *z* denote a complex scalar. Then the autocovariance generating transform is defined as

$$g(z) = \sum_{-\infty}^{\infty} \gamma(\tau) z^{\tau} \tag{4}$$

in whatever region of the complex plane the series on the right-hand side converges. If the series $\{x_t\}$ is covariance stationary, convergence will occur in an annulus about the unit circle. The autocovariance generating transform for the one-sided MA process defined in (2) is

$$g(z) = \sigma^2 B(z) B\left(z^{-1}\right) \tag{5}$$

where

$$B(z) = \sum_{k=0}^{\infty} b_k z^k.$$

If *B*(*z*) has no zeros on the unit circle, the process defined in (2) is invertible and also has an infinite-order autoregressive (AR) representation as

$$A(L) x_t = \varepsilon_t, \tag{6}$$

where *L* is the lag operator such that $L^j x_t = x_{t-j}$ and $A(L) = a_0 + a_1 L + a_2 L^2 + \ldots$.
So-called ARMA processes have an autocovariance generating transform which is a rational function of z. If the ARMA process is both stationary and invertible, *g*(*z*) may be written as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$G(z) = \frac{P(z)P\left(z^{-1}\right)}{Q(z)Q\left(z^{-1}\right)} = \sigma^2 \frac{\Pi_{k=1}^{m}(1-\beta_k z)\left(1-\beta_k z^{-1}\right)}{\Pi_{j=1}^{n}(1-\alpha_j z)\left(1-\alpha_j z^{-1}\right)}$$
(7)

where $|\beta_k|, |\alpha_j| < 1 \ \forall j, k.$ Then the corresponding ARMA model is

$$Q(L)x_t = P(L)\varepsilon_t,$$
(8)

where

$$Q(L) = \prod_{j=1}^{n}(1-\alpha_j L) \text{and} P(L) = \prod_{k=1}^{m}(1-\beta_k L).$$

2.2 Spectral density functions

If the value of *z* lies on the complex unit circle, it follows that $z = e^{-i\lambda}$, where $i = \sqrt{-1}$ and $-\pi \le \lambda \le \pi$. Substituting for *z* in the autocovariance generating transform (5) and dividing by 2π, we obtain the *spectral density function* of a linearly non-deterministic stationary process $\{x_t\}$ in terms of the frequency λ:

$$f(\lambda) = (1/2\pi)g\left(e^{i\lambda}\right) = \left(\sigma^2/2\pi\right)B\left(e^{i\lambda}\right)B\left(e^{-i\lambda}\right) = (1/2\pi)\sum_{-\infty}^{\infty}\gamma(\tau)e^{-i\lambda\tau}, \qquad -\pi \le \lambda < \pi.$$
(9)

Thus, the spectral density function is the Fourier transform of the autocovariance function. It can be shown that for a process with absolutely summable autocovariances the spectral density function exists and can be used to compute all of the autocovariances, so the same time series can be characterized equivalently in terms of the autocovariance function in the time-domain or in terms of the spectral density function in the frequency domain.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The spectral density function for a linearly non-deterministic, stationary, real-valued time series is a real-valued, non-negative function, symmetric about the origin, defined in the interval $[-\pi, \pi]$:

$$f(\lambda) = (1/2\pi)\left[\gamma(0) + 2\sum_{\tau=1}^{\infty}\gamma(\tau)\cos\lambda\tau\right]. \tag{10}$$

Moreover,

$$E(x_t - \mu)^2 = \int_{-\pi}^{\pi} f(\lambda)\mathrm{d}\lambda, \tag{11}$$

so that the spectral density function is a frequency-band decomposition of the variance of $\{x_t\}$. When the process generating $\{x_t\}$ is merely stationary, that is, when $\{x_t\}$ may have a linearly deterministic component, the spectral density function is

$$f(\lambda) = \int_{-\pi}^{\pi} e^{i\lambda\tau}\mathrm{d}F(\lambda), \tag{12}$$

where $F(\lambda)$ is a distribution function (Doob, 1953, p. 488). Note that deterministic seasonal effects, for example, may cause a jump in the spectral distribution function.

The autocovariance function, its generating transform and the spectral distribution function all have natural generalizations to the multivariate case, in which $\{x_t\}$ can be thought of as a vector of time-series processes.

The estimation and analysis of spectral density and distribution functions play an important role in all forms of time-series analysis. More detailed treatments are Doob (1953), Fishman (1969), Koopmans (1974), Fuller (1976), Nerlove, Grether and Carvalho (1979, ch. 3) and Priestley (1981).

2.3 Unobserved components (UC) models

In the statistical literature dealing with the analysis of economic time series it is common practice to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
classify the types of movements that characterize a time series as trend, cyclical, seasonal, and irregular components. The idea that a time series may best be viewed as being composed of several unobserved components is by no means universal, but it plays a fundamental role in many applications, for example, the choice of methods for seasonal adjustment. Nerlove, Grether and Carvalho (1979, ch. 1) review the history of the idea of unobserved components in economics from its origin early in the 19th century.

In the 1960s, Nerlove (1964; 1965; 1967) and Granger (1966) suggested that the typical spectral shape of many economic time series could be accounted for by the superposition of two or more independent components with specified properties. There are basically two approaches to the formulation of UC models: Theil and Wage (1964) and Nerlove and Wage (1964), Nerlove (1967) and Grether and Nerlove (1970) choose the form of components in such a way as to replicate the typical spectral shape of the series which represents their superposition. For example, let T_t represent the trend component, C_t the cyclical, S_t the seasonal, and I_t the irregular of a monthly time series; then the observed series can be represented as

$$y_t = T_t + C_t + S_t + I_t, \tag{13}$$

where

$$T_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p,$$

$$C_t = \frac{1 + \beta_1 L + \beta_2 L^2}{(1 - \alpha_1 L)(1 - \alpha_2 L)} \varepsilon_{1t},$$

$$S_t = \frac{1 + \beta_3 L + \beta_4 L^2}{1 - \gamma L^{12}} \varepsilon_{2t},$$

$$I_t = \varepsilon_{3t},$$

and ε_{1t}, ε_{2t}, and ε_{3t} are i.i.d. normal variables with variances σ_{11}, σ_{22}, and σ_{33}, respectively. This approach has been carried forward by Harvey (1984), Harvey and Peters (1990) and Harvey and Todd (1984).

An alternative approach is to derive the components of the UC model from a well-fitting ARMA model (obtained after suitably transforming the data), given sufficient a priori identifying restrictions on the spectral properties of the components. See Box, Hillmer and Tiao (1978), Pierce (1978; 1979), Burman (1980), Hillmer and Tiao (1982), Hillmer, Bell and Tiao (1983), Bell and Hillmer (1984), Burridge and Wallis (1985), and Maravall (1981; 1984). The basis of this procedure is the fact that every stationary

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

UC model, or the stationary part of every UC model, has an equivalent ARMA form, the so-called canonical form of the UC model (Nerlove and Wage, 1964; Nerlove, Grether and Carvalho, 1979, ch. 4).

3 Specification, estimation, inference and prediction

3.1 Autocovariance and spectral density functions

Suppose we have a finite number of observations of a realization of the process generating the time series, say $x_1, \ldots, x_T$. For expository purposes it is assumed that all deterministic components of x_t have been removed. If μ is unknown, this may be accomplished by subtracting the sample mean of the time series observations from the data prior to the analysis. For a zero mean series x_t there are basically two ways of estimating $\gamma(\tau)$ defined in (1): the first is the biased estimator

$$c(\tau) = (1/T)\sum_{t=1}^{T-|\tau|} x_t x_{t+|\tau|}, \quad \tau = 0, \pm 1, \ldots, \pm M, \; M \leq (T-1). \tag{14}$$

The second is the unbiased estimator

$$\tilde{c}(\tau) = [1/(T-|\tau|)]\sum_{t=1}^{T-|\tau|} x_t x_{t+|\tau|}, \quad \tau = 0, \pm 1, \ldots, \pm M, \quad M \leq T-1. \tag{15}$$

Although $c(\tau)$ is biased in finite samples, it is asymptotically unbiased. The key difference between $c(\tau)$ and $\tilde{c}(\tau)$ is that $c(\tau)$ is a positive definite function of τ whereas $\tilde{c}(\tau)$ is not (Parzen, 1961, p. 981). The variance and covariances of the estimated autocovariances are derived, *inter alia*, by Hannan (1960), and Anderson (1971). As $T \to \infty$, both tend to zero, as the estimates are asymptotically uncorrelated and consistent. However,

$$E[c(\tau) - Ec(\tau)]^2 / E[c(\tau)] \to \infty \quad \text{as} \quad \tau / T \to 1. \tag{16}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
This property accounts for the failure of the estimated autocorrelation function

$$r(\tau) = c(\tau) / c(0) \tag{17}$$

to damp down as $\tau \to \infty$, as it should for a stationary, linearly non-deterministic process (Hannan, 1960, p. 43).

A ‘natural’ estimator of the spectral density function is obtained by replacing $\gamma(\tau)$ in (10) by $c(\tau)$ or $\tilde{c}(\tau)$. The resulting estimator is proportional, at each frequency, to a sample quantity called the periodogram:

$$I_T(\lambda) = (2 / T)\left|\sum_{1}^{T} e^{i\lambda t} x_t\right|^2 \tag{18}$$

usually evaluated at the equi-spaced frequencies

$$\lambda = 2k\pi / T, \quad k = 1, 2, \ldots, [T / 2] \tag{19}$$

in the interval $[0, \pi]$. Although, for a stationary, nonlinearly deterministic process, the periodogram ordinates are asymptotically unbiased estimates of the spectral densities at the corresponding frequencies, they are not consistent estimates; moreover, the correlation between adjacent periodogram ordinates tends to zero with increasing sample size. The result is that the periodogram presents a jagged appearance which is increasingly difficult to interpret as more data become available.

In order to obtain consistent estimates of the spectral density function at specific frequencies, it is common practice to weight the periodogram ordinates over the frequency range or to form weighted averages of the autocovariances at different lags. There is a substantial literature on the subject. The weights are called a ‘spectral window’. Essentially the idea is to reduce the variance of the estimate of an average spectral density around a particular frequency by averaging periodogram ordinates which are asymptotically unbiased and independently distributed estimates of the corresponding ordinates of the spectral density function. Related weights can also be applied to the estimated autocovariances which are substituted in (10); this weighting system is called a ‘lag window’.

Naturally the sampling properties of the spectral estimates depend on the nature of the ‘window’ used to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

obtain consistency (see Priestley, 1981, pp. 432–94 for further discussion). Regardless of the choice of window, the 'bandwidth' used in constructing the window must decrease at a suitable rate as the sample size grows. In the spectral window approach, this means that the window width must decrease at a slower rate than the sample size. In the lag window approach, this means that the number of included autocovariances must increase at a slower rate than the sample size.

3.2 ARMA models

The autocovariance function and the spectral density function for a time series represent nonparametric approaches to describing the data. An alternative approach is to specify and estimate a parametric ARMA model for x_t This approach involves choosing the orders of the polynomials P and Q in (7) and (8) and perhaps also specifying that one or more coefficients are zero or placing other restrictions on P and Q. The problem then becomes one of estimating the parameters of the model.

Despite the poor statistical properties of the estimated autocovariance function and a related function called the partial autocorrelation function, these are sometimes used to specify the orders of the polynomials P and Q. An alternative approach is to select the model that minimizes the value of information-theoretic criteria of the form

$$\mathrm{IC}(i) = \log(\hat{\sigma}_i^2) + k_i c_T, \tag{20}$$

where k_i refers to the number of estimated parameters in the candidate models $i = 1, \ldots, M$, and $\hat{\sigma}_i^2$ to the corresponding maximum likelihood estimate of the residual variance. Such criteria incorporate a trade-off between the fit of a model and its degree of parsimony. That trade-off depends on the penalty term c_t (Akaike, 1970; 1974; Schwarz, 1978). There is no universally accepted choice for c_T. For $c_T = 2/T$ expression (20) reduces to the Akaike information criterion (AIC), for example, and for $c_T = \ln(T)/T$ to the Schwarz information criterion (SIC). The asymptotic properties of alternative criteria will depend on the objective of the user and the class of models considered.

Given the orders of the AR and MA components, a variety of maximum likelihood or approximate maximum likelihood methods are available to estimate the model parameters. Newbold (1974) shows that, if x_t is characterized by (8) with $\varepsilon_t \sim NID(0, \sigma^2)$, then the exact likelihood function for the parameters of $P(\cdot)$ and $Q(\cdot)$ is such that the maximum likelihood estimates of the parameters and the least-squares (LS) estimates (in general highly nonlinear) are asymptotically identical. Only in the case of a pure AR model are the estimates linear conditional on the initial observations. Several approximations have been discussed (Box and Jenkins, 1970; Granger and Newbold, 1977; Nerlove, Grether and Carvalho, 1979, pp. 121–5).

Exact maximum likelihood estimation of ARMA models has been discussed by, inter alia, Newbold

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(1974), Anderson (1977), Ansley (1979), and Harvey (1981). Following Schweppe (1965), Harvey suggests the use of the Kalman filter to obtain the value of the exact-likelihood function, which may be maximized by numerical methods. The Kalman filter approach is easily adapted to the estimation of UC models in the time domain.

An alternative to exact or approximate maximum-likelihood estimation in the time domain was suggested by Hannan (1969). Estimates may be obtained by maximizing an approximate likelihood function based on the asymptotic distribution of the periodogram ordinates defined in (18). These are asymptotically independently distributed (Brillinger, 1975, p. 95), and the random variables $2I_T(\lambda)/f(\lambda)$ have an asymptotic χ^2 distribution with two degrees of freedom (Koopmans, 1974, pp. 260–5). This means that the asymptotic distribution of the observations, $\{x_1, \ldots, x_T\}$ is proportional to

$$\prod_{j=0}^{[T/2]} [1/f(\lambda_j)]\exp[-I(\lambda_j)/f(\lambda_j)] \tag{21}$$

where $\lambda_j = 2j\pi/T,\ j = 0, \ldots, [T/2]$, are the equi-spaced frequencies in the interval $[0, \pi]$ at which the periodogram is evaluated (Nerlove, Grether and Carvalho, 1979, pp. 132–6). Since the true spectral density $f(\lambda)$ depends on the parameters characterizing the process, this asymptotic distribution may be interpreted as a likelihood function. Frequency domain methods, as these are called, may easily be applied in the case of UC models.

Whether approximate or exact maximum-likelihood estimation methods are employed, inference may be based on the usual criteria related to the likelihood function. Unfortunately, serious difficulties may be encountered in applying the asymptotic theory, since the small sample distribution of the maximum likelihood estimator may differ greatly from the limiting distribution in important cases (Sargan and Bhargava, 1983; Anderson and Takemura, 1986).

3.3 Prediction and extraction

The problem of prediction is essentially the estimation of an unknown future value of the time series itself; the problem of extraction, best viewed in the context of UC models described in section 2.3, is to estimate the value of one of the unobserved components at a particular point in time, not necessarily in the future. Problems of trend extraction and seasonal adjustment may be viewed in this way (Grether and Nerlove, 1970). How the prediction (or extraction) problem is approached depends on whether we are assumed to have an infinite past history and, if not, whether the parameters of the process generating the time series are assumed to be known. In practice, of course, an infinite past history is never available, but a very long history is nearly equivalent if the process is stationary or can be transformed to stationarity. It is common, as well, to restrict attention to linear predictors, which involves no loss of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
generality if the processes considered are Gaussian and little loss if merely linear. To devise a theory of optimal prediction or extraction requires some criterion by which to measure the accuracy of a particular candidate. The most common choice is the minimum mean-square error (MMSE) criterion, which is also the conditional expectation of the unknown quantity. For a discussion of alternative loss functions see Granger (1969) and Christoffersen and Diebold (1996; 1997).

The theory of optimal prediction and extraction due to Kolmogorov (1941) and Wiener (1949) and elaborated by Whittle (1963) for discrete processes assumes a possibly infinite past history and known parameters. As a special case of the Wiener–Kolmogorov theory for non-deterministic, stationary processes, consider the linear process defined by (2). Since the ε_t are i.i.d. with zero mean and variance σ^2, it is apparent that the conditional expectation of x_{t+v}, given all innovations from the infinite past to *t*, is

$$\hat{x}_{t+v} = b_v \varepsilon_t + b_{v+1} \varepsilon_{t+1} + \ldots \tag{22}$$

Of course, even if the parameters $b_j,\ j = 0, 1, \ldots,$ are assumed to be known, the series $\{\varepsilon_t\}$ is not directly observable. The ε_t's are sometimes called the *innovations* of the process, since it is easy to show that $\varepsilon_{t+1} = x_{t+1} - \hat{x}_{t+1}$ is the one-step ahead prediction error. If the process is invertible, it has the autoregressive representation (6) and so can be expressed solely in terms of the, generally infinite-order, autoregression

$$\hat{x}_{t+v} = D(L)\, x_t, \tag{23}$$

where the generating transform of the coefficients of *D* is

$$D(z) = \frac{1}{B(z)} \left[\frac{B(z)}{z^v} \right]_+ .$$

The operator $[\cdot]_+$ eliminates terms involving negative powers of z.

The problem of extraction is best viewed in the context of multiple time series; in general we wish to ‘predict’ one time series $\{y_t\}$ from another related series $\{x_t\}$. It is not necessary that the series $\{y_t\}$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

actually be observed as long as its relationship to an observed series $\{x_t\}$ can be described (Nerlove, Grether and Carvalho, 1979, ch. 5).

The Kalman filter approach to prediction and extraction (Kalman, 1960) is both more special and more general than the Wiener–Kolmogorov theory: attention is restricted to finite-dimensional parameter spaces and linear processes, but these processes need not be stationary. The parameters may vary with time, and we do not require an infinite past. This approach represents a powerful tool of practical time-series analysis and may be easily extended to multiple time series. A full discussion, however, requires a discussion of 'state-space representation' of time series processes and is beyond the scope of this entry (Harvey, 1989)

4 Multiple time series analysis

A general treatment of multiple time series analysis is contained in Hannan (1970). The two-variable case will serve to illustrate the matter. Two stationary time series $\{x_t\}$ and $\{y_t\}$ are said to be jointly stationary if their joint distribution function does not depend on the origin from which time is measured. Joint stationarity implies, but is not in general implied by, weak or covariance joint stationarity; that is, $cov(x_t, y_s)$ is a function of $s - t$ only. In this case the cross-covariance function is

$$\gamma_{yx}(\tau) = E[y_t - \mu_y][x_{t-\tau} - \mu_x], \tag{24}$$

where $\mu_x = Ex_t$ and $\mu_y = Ey_t$. Note that $\gamma_{yx}(\tau)$ and $\gamma_{xy}(\tau)$ are, in general, different. The cross-covariance generating function is defined as

$$g_{yx}(z) = \sum_{-\infty}^{\infty} \gamma_{yx}(\tau) z^{\tau} \tag{25}$$

in that region of the complex plane in which the right-hand side of (25) converges. For two jointly stationary series this occurs in an annulus containing the unit circle. In this case, the cross-spectral density function is defined as

$$f_{yx}(\lambda) = (1/2\pi) g_{yx}\left(e^{i\lambda}\right). \tag{26}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Since $\gamma_{yx}(\tau)$ and $\gamma_{xy}(\tau)$ are not equal, the cross-spectral density function is complex valued and can be decomposed into a real part (the co-spectral density) and a complex part (the quadrature spectral density):

$$f_{yx}(\lambda) = c_{yx}(\lambda) + iq_{yx}(\lambda). \tag{27}$$

In polar form, the cross-spectral density may be written as

$$f_{yx}(\lambda) = \alpha_{yx}(\lambda)\exp[i\varphi_{yx}(\lambda)], \tag{28}$$

where $\alpha_{yx}(\lambda) = [c_{yx}^2(\lambda) + q_{yx}^2(\lambda)]^{1/2}$ is called the amplitude or gain, and where $\varphi_{yx}(\lambda) = \arctan\{-q_{yx}(\lambda)/c_{yx}(\lambda)\}$ is called the phase. Another useful magnitude is the coherence between the two series, defined as

$$\rho_{yx}(\lambda) = \frac{|f_{yx}(\lambda)|^2}{f_{xx}(\lambda)f_{yy}(\lambda)}, \tag{29}$$

which measures the squared correlation between *y* and *x* at a frequency λ. Clearly, $\rho_{yx}(\lambda) = \rho_{xy}(\lambda)$. Estimation of cross-spectral density functions and related quantities is discussed in Priestley (1981, pp. 692–712).

Often it is convenient to impose additional parametric structure in modelling multiple time series. The workhorse multiple time series model in econometrics has been the covariance-stationary *K*-dimensional vector autoregressive model, which may be viewed as a natural generalization of the univariate AR model discussed earlier:

$$A(L)x_t = \varepsilon_t$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $A(L) = I_K - A_1 L - \ldots - A_p L^p$. Here each variable in x_t is regressed on its own lags as well as lags of all other variables in x_t up to some pre-specified lag order p. This *vector autoregression* (VAR) can also be viewed as an approximation to a general linear process x_t, and may be estimated by LS. Similarly, the formulation of ARMA and UC models discussed earlier may be extended to the multivariate case by interpreting the polynomials in the lag operator as matrix polynomials and by replacing the scalar random variables by vectors. Although these vector ARMA and UC models bear a superficial resemblance to the corresponding univariate ones, their structure is, in fact, much more complicated and gives rise to difficult identification problems. In the univariate case, we can formulate simple conditions under which a given covariance function identifies a unique ARMA or UC model, but in the multivariate case these conditions are no longer sufficient. Hannan (1970; 1971) gives a complete treatment. State-space methods have also been employed to study the structure of multivariate ARMA models (Hannan, 1976; and, especially, 1979).

5 Unit roots, co-integration and long memory

Standard tools for time series analysis have been developed for processes that are covariance stationary or have been suitably transformed to achieve covariance stationarity by removing (or explicitly modelling) deterministic trends, structural breaks, and seasonal effects. The presence of a *unit root* in the autoregressive lag order polynomial of an ARMA process also violates the assumption of stationarity. Processes with a unit root are also called integrated of order one (or $I(1)$ for short) because they become covariance-stationary only upon being differenced once. In general, $I(d)$ processes must be differenced d times to render the process covariance-stationary.

The presence of unit roots has important implications for estimation and inference. When the scalar process x_t is $I(1)$ the variance of x_t will be unbounded, model innovations will have permanent effects on the level of x_t, the autocorrelation function does not die out, and x_t will not revert to a long-run mean. Moreover, coefficients of $I(1)$ regressors will have nonstandard asymptotic distributions, invalidating standard tools of inference.

The simplest example of an autoregressive integrated moving-average (ARIMA) process is the random walk process: $x_t = x_{t-1} + \varepsilon_t$. The potential pitfalls of regression analysis with $I(1)$ data are best illustrated by the problem of regressing one independent random walk on another. In that case, it can be shown that R^2 and $\hat{\beta}$ will be random and that the usual t-statistic will diverge, giving rise to seemingly significant correlations between variables that are unrelated by construction. This *spurious regression* problem was first discussed by Yule (1926), further illustrated by Granger and Newbold (1974), and formally analyzed by Phillips (1986) and Phillips and Durlauf (1986). Similar problems arise in deterministically detrending $I(1)$ series (Nelson and Kang, 1981; Durlauf and Phillips, 1988). Unbalanced regressions, that is, regressions in which the regressand is not of the same order of integration as the regressor, may also result in spurious inference. An exception to this rule is inference

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

on coefficients of mean zero *I*(0) variables in regressions that include a constant term (Sims, Stock and Watson, 1991).

The standard response to dealing with *I*(1) data is to difference the data prior to the analysis. There is one important exception to this rule. There are situations in which several variables are individually *I*(1), but share a common unit root component. In that case, a linear combination of these variables will be *I*(0):

$$c' x_t = u_t \sim I(0),\ c \neq 0 \tag{31}$$

where x_t denotes a *K*-dimensional vector of *I*(1) variables and *c* is a $(K \times 1)$ parameter vector. In other words, these variables share a common stochastic trend. This phenomenon is known as *co-integration* (Granger, 1981; Engle and Granger, 1987) and *c* is known as the co-integrating vector. Clearly, *c* is not unique. It is common to normalize one element of *c* to unity. The LS estimator of *c* in (31) is consistent, but corrections for omitted dynamics are recommended (Stock and Watson, 1993; Phillips and Hansen, 1990). Co-integrating relationships have been used extensively in modelling long-run equilibrium relationships in economic data (Engle and Granger, 1991).

Variables that are co-integrated are linked by an *error correction* mechanism that prevents the integrated variables from drifting apart without bound. Specifically, by the Granger representation theorem of Engle and Granger (1987), under some regularity conditions, any *K*-dimensional vector of co-integrated variables x_t can be represented as a vector error correction (VEC) model of the form:

$$\Delta x_t = \sum_{i=1}^{p-1} \Gamma_i \Delta x_{t-i} - \Pi x_{t-p} \tag{32}$$

where $\Gamma_i,\ i = 1, \ldots, p-1,$ and $\Pi \equiv BC$ are conformable coefficient matrices and Δ denotes the first-difference operator. Model (32) allows for up to *r* co-integrating relationships where *r* is the rank of Π. For $r = 0,$ the error correction term in model (32) drops out and the model reduces to a difference-stationary VAR. For $r = K,$ all variables are *I*(0) and model (32) is equivalent to a stationary VAR in levels. Otherwise, there are $0 < r < K$ common trends. If the $(r \times K)$ matrix of co-integrating vectors, *C*, is known, the model in (32) reduces to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\Delta x_t = \sum_{i=1}^{p-1} \Gamma_i \Delta x_{t-i} - B z_{t-p}$$
(32′)

where $z_{t-p} \equiv C x_{t-p}$, and the model may be estimated by LS; if only the rank *r* is known, the VEC model in (32) is commonly estimated by full information maximum likelihood methods (Johansen, 1995).

Starting with Nelson and Plosser (1982), a large literature has dealt with the problem of statistically discriminating between *I*(1) and *I*(0) models for economic data. Notwithstanding these efforts, it has remained difficult to detect reliably the existence of a unit root (or of co-integration). The problem is that in small samples highly persistent, yet stationary processes are observationally equivalent to exact unit root processes. It may seem that not much could hinge on this distinction then, but it can be shown that *I*(1) and *I*(0) specifications that fit the data about equally well may have very different statistical properties and economic implications (Rudebusch, 1993).

For processes with roots near unity in many cases neither the traditional asymptotic theory for *I*(0) processes nor the alternative asymptotic theory for exact *I*(1) processes will provide a good small-sample approximation to the distribution of estimators and test statistics. An alternative approach is to model the dominant root, ρ, of the autoregressive lag order polynomial as *local-to-unity* in the sense that $\rho = 1 - c/T,\ c > 0$. This asymptotic thought experiment gives rise to an alternative asymptotic approximation that in many cases provides a better small-sample approximation than imposing the order of integration or relying on unit root pretests (Stock, 1991; Elliott, 1998).

Stationary ARMA processes are ‘short memory’ processes in that their autocorrelation function dies out quickly. For large τ, ARMA autocorrelations decay approximately geometrically, that is, $\rho(\tau) \approx r^{\tau}$, where *r* is a constant such that $|r| < 1$. In many applied contexts including volatility dynamics in asset returns, there is evidence that the autocorrelation function dies out much more slowly. This observation has motivated the development of the class of *fractionally integrated* ARMA (ARFIMA) models:

$$Q(L)(1-L)^{d} x_t = P(L)\varepsilon_t$$
(33)

where *d* is a real number, as opposed to an integer (Baillie, 1996). Stationarity and invertibility require $|d| < 0.5$, which can always be achieved by taking a suitable number of differences. The autocorrelation function of an ARFIMA process decays at a hyperbolic rate. For large τ, we have $\rho(\tau) \approx \tau^{2d-1}$, where $d < 1/2$ and $d \neq 0$. Such ‘*long memory*’ models may be estimated by the two-step procedure of Geweke and Porter-Hudak (1983) or by maximum likelihood (Sowell, 1992; Baillie, Bollerslev and Mikkelsen, 1996). A detailed discussion including extensions to the notion of fractional co-integration is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

provided by Baillie (1996). Long memory may arise, for example, from infrequent stochastic regime changes (Diebold and Inoue, 2001) or from the aggregation of economic data (Granger, 1980; Chambers, 1998). Perhaps the most successful application of long-memory processes in economics has been work on modelling the volatility of asset prices and powers of asset returns, yielding new insights into the behaviour of markets and the pricing of financial risk.

6 Nonlinear time series models

The behaviour of many economic time series appears to change distinctly at irregular intervals, consistent with economic models that suggest the existence of floors and ceilings, buffer stocks and regime switches in the data. This observation has given rise to a large literature dealing with nonlinear time series models. Nonlinear time series models still have a Wold representation with linearly unpredictable innovations, but these innovations are nevertheless dependent over time. This has important implications for forecasting and for the dynamic properties of the model. For example, the effects of innovations in nonlinear models will depend on the path of the time series and the size of the innovation, and may be asymmetric.

6.1 Nonlinear dynamics in the conditional mean

The increasing importance of nonlinear time series models in econometrics is best illustrated by two examples: hidden Markov chain models and smooth transition regression models of the conditional mean.
The idea of hidden Markov chains first attracted attention in econometrics in the context of *regime switching models* (Hamilton, 1989). The original motivation was that many economic time series appear to follow a different process during recession phases of the business cycle than during economic expansions. This type of regime-switching behaviour may be modelled in terms of an unobserved discrete-valued state variable (for example, 1 for a recession and 0 for an expansion) that is driven by a Markov chain. The transition from one state to another is governed by a matrix of transition probabilities that may be estimated from past data. The essence of this method thus is that the future will in some sense be like the past. A simple example of this idea is the regime-switching AR(1) model:

$$x_t = a_{1st} x_{t-1} + \varepsilon_t, \ \varepsilon_t \sim NID(0, \sigma^2) \tag{34}$$

where the regime s_t is the outcome of an unobserved two-state Markov chain with s_t independent of ε_τ for all t and τ. In this model, the time-varying slope parameter will take on different values depending on the state s. Once the model has been estimated by maximum likelihood methods, it is possible to infer how likely a given regime is to have generated the observed data at date t. An excellent review of the literature on hidden Markov models is provided by Cappé, Moulines and Ryden (2005); for a general

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

treatment of state space representations of nonlinear models, also see Durbin and Koopman (2001). The idea of *smooth transition regression models* is based on the observation that many economic variables are sluggish and will not move until some state variable exceeds a certain threshold. For example, price arbitrage in markets will only set in once the expected profit of a trade exceeds the transaction cost. This observation has led to the development of models with fixed thresholds that depend on some observable state variable. Smooth transition models allow for the possibility that this transition occurs not all of a sudden at a fixed threshold but gradually, as one would expect in time series data that have been aggregated across many market participants. A simple example is the smooth-transition AR(1) model:

$$x_t = \Phi(z_{t-1}, \ldots, z_{t-d}, \Gamma) x_{t-1} + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma^2) \tag{35}$$

where $\Phi(\cdot)$ denotes the transition function, z_t is a zero mean state variable denoting the current deviation of x_t from a (possibly time-varying) equilibrium level and Γ is the vector of transition parameters. Common choices for the transition function are the logistic or the exponential function. For example, we may specify $\Phi(\cdot) = (\exp\{\gamma(z_{t-1})^2\})$ with $\gamma < 0$. If $z_{t-1} = 0$, $\Phi(\cdot) = 1$ and the model in (35) reduces to a random walk model; otherwise, $\Phi(\cdot) < 1$ and the model in (35) reduces to a stationary AR(1). The degree of mean reversion is increasing in the deviation from equilibrium. For further discussion see Granger and Teräsvirta (1993).

6.2 Nonlinear dynamics in the conditional variance

While the preceding examples focused on nonlinear dynamics in the conditional mean, nonlinearities may also arise in higher moments. The leading example is the conditional variance. Many economic and financial time series are characterized by volatility clustering. Often interest centres on predicting these *volatility dynamics* rather than the conditional mean. The basic idea of modelling and forecasting volatility was set out in Engle's (1982) path-breaking paper on autoregressive conditional heteroskedasticity (ARCH). Subsequently, Bollerslev (1986) introduced the class of generalized autoregressive conditionally heteroskedastic (GARCH). Consider a decomposition of x_t into the one-step ahead conditional mean, $\mu_{t|t-1} \equiv E(x_t|\Omega_{t-1})$, and conditional variance, $\sigma^2_{t|t-1} \equiv Var(x_t|\Omega_{t-1})$, where Ω_{t-1} denotes the information set at $t-1$:

$$x_t = \mu_{t|t-1} + \sigma_{t|t-1} v_t \quad v_t \sim NID(0, 1) \tag{36}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The leading example of a GARCH model of the conditional variance is the GARCH(1,1) model, which is defined by the recursive relationship

$$\sigma^2_{t|t-1} = \omega + \alpha\varepsilon^2_{t-1} + \beta\sigma^2_{t-1|t-2} \tag{37}$$

where $\varepsilon_t \equiv \sigma_{t|t-1} v_t$, and the parameter restrictions $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ ensure that the conditional variance remains positive for all realizations of v_t. The standard estimation method is maximum likelihood. The basic GARCH(1,1) model may be extended to include higher-order lags, to allow the distribution of v_t to have fat tails, to allow for asymmetries in the volatility dynamics, to permit the conditional variance to affect the conditional mean, and to allow volatility shocks to have permanent effects or volatility to have long memory. It may also be extended to the multivariate case.
It follows directly from the formulation of the GARCH(1,1) model that the optimal, in the MMSE sense, one-step-ahead forecast equals $\sigma^2_{t+1|t}$. Similar expressions for longer horizons may be obtained by recursive updating. There is a direct link from the arrival of news to volatility measures and from volatility forecasts to risk assessments. These and alternative volatility models and the uses of volatility forecasts are surveyed in Andersen et al. (2006). For a comparison of GARCH models with the related and complementary class of stochastic volatility models, see Andersen, Bollerslev and Diebold (2006) and Shephard (2005).

7 Applications

Time series analytic methods have many applications in economics. Here we consider five: (1) analysis of the cyclic properties of economic time series, (2) description of seasonality and seasonal adjustment, (3) forecasting, (4) dynamic econometric modelling, and (5) structural vector autoregressions.

7.1 Analysis of the cyclic properties of economic time series

Suppose that the time series $\{x_t\}$ is a linearly non-deterministic stationary series and that the series $\{y_t\}$ is formed from $\{x_t\}$ by the linear operator

$$y_t = \sum_{j=m}^{n} w_j x_{t-j}, \quad \sum_{m}^{n} w_j^2 < \infty . \tag{38}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Such an operator is called a time-invariant linear filter. Analysis of the properties of such filters plays an important role in time series analysis since many methods of trend estimation or removal and seasonal adjustment may be represented or approximated by such filters. An interesting example that illustrates the potential pitfalls of using such filters is provided by Adelman (1965), who showed that the 20-year long swings in various economic series found by Kuznets (1961) may well have been the result of the trend filtering operations used in preliminary processing of the data. For a fuller treatment see Nerlove, Grether and Carvalho (1979, pp. 53–7).

Since the 1980s, there has been increased interest in the use of nonlinear filters for extracting the business cycle component of macroeconomic time series. Examples include the band-pass filter (Christiano and Fitzgerald, 2003) and the Hodrick–Prescott (HP) filter (Hodrick and Prescott, 1997; Ravn and Uhlig, 2002). The latter approach postulates that $y_t = \tau_t + c_t$, where τ_t denotes the trend component and c_t the deviation from trend or 'cyclical' component of the time series y_t. The trend component is chosen to minimize the loss function:

$$\sum_{t=1}^{T} c_t^2 + \lambda \sum_{t=1}^{T} \left[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}) \right]^2 \tag{39}$$

where $c_t = y_t - \tau_t$ and λ is a pre-specified parameter that depends on the frequency of the observations. The trade-off in this optimization problem is between the degree to which the trend component fits the data and the smoothness of the trend.

7.2 Description of seasonality and seasonal adjustment

Many economic time series exhibit fluctuations which are periodic within a year or a fraction thereof. The proper treatment of such seasonality, whether stochastic or deterministic, is the subject of a large literature, summarized rather selectively in Nerlove, Grether and Carvalho (1979, ch. 1). More recent treatments can be found in Hylleberg (1992), Franses (1996) and Ghysels and Osborn (2001). Seasonality may be modelled and its presence detected using spectral analysis (Nerlove, 1964) or using time domain methods. Deterministic seasonality, in the form of model parameters that vary deterministically with the season, offers no great conceptual problems but many practical ones. Stochastic seasonality is often modelled in the form of seasonal unit roots. In that case, seasonal differencing of the data removes the unit root component. Multiple time series may exhibit seasonal co-integration. Sometimes it is convenient to specify stochastic seasonality in the form of an UC model (Grether and Nerlove, 1970). Appropriate UC models may be determined directly or by fitting an ARIMA model and deriving a related UC model by imposing sufficient a priori restrictions (Hillmer and Tiao, 1982; Bell and Hillmer, 1984).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

7.3 Forecasting

One of the simplest forecasting procedures for time series is exponential smoothing based on the relationship

$$\hat{x}_{t+1|t} = (1-\theta)x_t + \theta\hat{x}_{t|t-1} \tag{40}$$

where x_t, is the observed series and $\hat{x}_{j|k}$ is the forecast of the series at time *j* made on the basis of information available up to time *k*. Muth (1960) showed that (40) provides an MMSE forecast if the model generating the time series is $x_t - x_{t-1} = \varepsilon_t - \theta\varepsilon_{t-1}$. Holt (1957) and Winters (1960) generalized the exponential smoothing approach to models containing more complex trend and seasonal components. Further generalization and proofs of optimality are contained in Theil and Wage (1964) and Nerlove and Wage (1964).

Perhaps the most popular approach to forecasting time series is based on ARIMA models of time series processes (Box and Jenkins, 1970). The developments discussed in the preceding paragraph led to the development of UC models, which give rise to restricted ARIMA model forms (Nerlove, Grether and Carvalho, 1979). State-space representations of these models permit the application of the Kalman filter to both estimation and forecasting. Harvey (1984) presents a unified synthesis of the various methods.

More recently, the focus has shifted from traditional forecasting methods towards methods that exploit the increased availability of a large number of potential predictors. Consider the problem of forecasting $y_{t+h|t}$ based on its own current and past values as well as those of *N* additional variables, x_t. Of particular interest is the case in which the number of predictors, *N*, exceeds the number of time series observations, *T*. In that case, principal components analysis provides a convenient way of extracting a low-dimensional vector of common factors from the original data-set x_t (Stock and Watson, 2002a; 2002b). Forecasts that incorporate estimated common factors have proved successful in many cases in reducing forecast errors relative to traditional time series forecasting methods. Boivin and Ng (2005) provide a systematic comparison of alternative *factor model forecasts*. Another promising forecasting method is *Bayesian model averaging* across alternative forecasting models (Raftery, Madigan and Hoeting, 1997). The latter method builds on the literature on forecast combinations (Bates and Granger, 1969).

7.4 Dynamic econometric modelling

There is a close connection between multivariate time-series models and the structural, reduced and final forms of dynamic econometric models; the standard simultaneous-equations model (SEM) is a specific and restricted case.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Suppose that a vector of observed variables y_t may be subdivided into two classes of variables, 'exogenous', $\{x_t\}$, and endogenous, $\{z_t\}$. A dynamic, multivariate simultaneous linear system may be written.

$$\begin{bmatrix} \Psi_{11}(L) & \Psi_{12}(L) \\ 0 & \Psi_{22}(L) \end{bmatrix} \begin{pmatrix} z_t \\ x_t \end{pmatrix} = \begin{bmatrix} \Theta_{11}(L) & 0 \\ 0 & \Theta_{22}(L) \end{bmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \tag{41}$$

where $\Psi_{ij}(\cdot)$ and $\Theta_{ij}(\cdot)$, $i, j = 1, 2$ are matrix polynomials in the lag operator L. Such systems are known as vector ARMAX models and conditions for their identification are given by Hatanaka (1975). The reduced form of the system is obtained by expressing z_t as a function of lagged endogenous and current and lagged exogenous variables. The final form is then obtained by eliminating the lagged endogenous variables (see Zellner and Palm, 1974; Wallis, 1977).

7.5 Structural vector autoregressions

An important special case of the dynamic SEM is the structural vector autoregressive model in which all variables are presumed endogenous, the lag structure is unrestricted up to some order p, and identification of the structural from is achieved by imposing restrictions on the correlation structure of the structural innovations (Sims, 1980). The most common form of the structural VAR(p) model imposes restrictions on the contemporaneous interaction of structural innovations. Consider the structural form for a K-dimensional vector $\{x_t\}$, $t = 1, \ldots, T$:

$$B_0 x_t = \sum_{i=1}^{p} B_i x_{t-i} + \eta_t, \tag{42}$$

where $\eta_t \sim (0, \Sigma_\eta)$ denotes the $(K \times 1)$ vector of serially uncorrelated structural innovations (or shocks) and B_i, $i = 0, \ldots, p$, the $(K \times K)$ coefficient matrices. Without loss of generality, let $\Sigma_\eta = I$. The corresponding reduced form is

$$x_t = \sum_{i=1}^{p} B_0^{-1} B_i x_{t-i} + B_0^{-1} \eta_t = \sum_{i=1}^{p} A_i x_{t-i} + \varepsilon_t \tag{43}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $\varepsilon_t \sim (0, \Sigma_\varepsilon)$. Since $\varepsilon_t = B_0^{-1} \eta_t$, it follows that $\Sigma_\varepsilon = B_0^{-1} B_0^{-1'}$. Given a consistent estimate of the reduced form parameters A_i, $i = 1, \ldots, p$, and Σ_ε, the elements of B_0^{-1} will be exactly identified after imposing $K(K-1)/2$ restrictions on the parameters of B_0^{-1} that reflect the presumed structure of the economy. Given estimates of B_0^{-1} and A_i, $i = 1, \ldots, p$, estimates of the remaining structural parameters may be recovered from $B_i = B_0 A_i$.

In practice, the number of restrictions that can be economically motivated may be smaller or larger than $K(K-1)/2$. Alternative estimation strategies that remain valid in the over-identified case include the generalized method of moments (Bernanke, 1986) and maximum likelihood (Sims, 1986). An instrumental variable interpretation of VAR estimation is discussed in Shapiro and Watson (1988). Semi-structural VAR models that are only partially identified have been proposed by Bernanke and Mihov (1998).

Alternative identification strategies may involve putting restrictions on the long-run behaviour of economic variables (Blanchard and Quah, 1989; King et al., 1991) or on the sign and/or shape of the impulse responses (Faust, 1998). Other possibilities include identification via heteroskedasticity (Rigobon, 2003) or the use of high-frequency data (Faust, Swanson and Wright, 2004).

The estimates of the structural VAR form may be used to compute the dynamic responses of the endogenous variables to a given structural shock, variance decompositions that measure the average contribution of each structural shock to the overall variability of the data, and historical decompositions of the path of x_t based on the contribution of each structural shock.

8 Conclusions

The literature on time series analysis has made considerable strides since the 1980s. The advances have been conceptual, theoretical and methodological. The increased availability of inexpensive personal computers in particular has revolutionized the implementation of time series techniques by shifting the emphasis from closed-form analytic solutions towards numerical and simulation methods. The ongoing improvements in information technology, broadly defined to include not only processing speed but also data collection and storage capabilities, are likely to transform the field even further. For example, the increased availability of large cross-sections of time series data, the introduction of ultra high-frequency data, the electronic collection of micro-level time series data (such as web-based data or scanner data), and the increased availability of data in real time all are creating new applications and spurring interest in the development of new methods of time series analysis. These developments already have brought together the fields of empirical finance and time series econometrics, resulting in the emergence of the new and fertile field of financial econometrics. As the use of time series methods becomes more widespread in applied fields, there will be increasing interest in the development of methods that can be adapted to the specific objectives of the end user. Another question of growing importance is how to deal with rapidly evolving economic environments in the form of structural breaks and other model

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
instabilities. Finally, the improvement of structural time series models for macroeconomic policy analysis will remain a central task if time series analysis is to retain its importance for economic policymaking.

See Also

- cointegration
- factor models
- forecasting
- long memory models
- maximum likelihood
- nonlinear time series analysis
- seasonal adjustment
- spectral analysis
- spurious regressions
- trend/cycle decomposition
- vector autoregressions

Bibliography

Adelman, I. 1965. Long cycles – fact or artifact? *American Economic Review* 60, 443–63.

Akaike, H. 1970. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22, 203–17.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716–23.

Anderson, T. 1971. *The Statistical Analysis of Time Series*. New York: John Wiley.

Anderson, T. 1977. Estimation for autoregressive moving average models in the time and frequency domains. *Annals of Statistics* 5, 842–65.

Anderson, T. and Takemura, A. 1986. Why do noninvertible moving averages occur? *Journal of Time Series Analysis* 7, 235–54.

Andersen, T., Bollerslev, T. and Diebold, F. 2006. Parametric and nonparametric volatility measurement. In *Handbook of Financial Economics*, ed. L. Hansen and Y. Ait-Sahalia. Amsterdam: North-Holland.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Andersen, T., Bollerslev, T., Christoffersen, P. and Diebold, F. 2006. Volatility and correlation forecasting. In *Handbook of Economic Forecasting*, ed. G. Elliott, C. Granger and A. Zimmermann. Amsterdam: North-Holland.

Ansley, C. 1979. An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika* 66, 59–65.

Baillie, R. 1996. Long memory processes and fractional integration in econometrics. *Journal of Econometrics* 73, 5–59.

Baillie, R., Bollerslev, T. and Mikkelsen, H.-O. 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 74, 3–30.

Bates, J. and Granger, C. 1969. The combination of forecasts. *Operational Research Quarterly* 20, 451–68.

Bell, W. and Hillmer, S. 1984. Issues involved with seasonal analysis of economic time series. *Journal of Business and Economic Statistics* 2, 291–349.

Bernanke, B. 1986. Alternative explanations of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy* 25, 49–100.

Bernanke, B. and Mihov, I. 1998. Measuring monetary policy. *Quarterly Journal of Economics* 113, 869–902.

Beveridge, W. 1921. Weather and harvest cycles. *Economic Journal* 31, 429–52.

Beveridge, W. 1922. Wheat prices and rainfall in western Europe. *Journal of the Royal Statistical Society* 85, 412–59.

Blanchard, O. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.

Boivin, J. and Ng, S. 2005. Understanding and comparing factor-based forecasts. *International Journal of Central Banking* 1(3), 117–52.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–27.

Box, G., Hillmer, S. and Tiao, G. 1978. Analysis and modeling of seasonal time series. In *Seasonal Analysis of Economic Time Series*. ed. A. Zellner. Washington, DC: Bureau of the Census, Department

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Commerce.

Box, G. and Jenkins, G. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Brillinger, D. 1975. *Time Series: Data Analysis and Theory*. New York: Holt.

Burman, J. 1980. Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society. Series A* 143, 321–37.

Burridge, P. and Wallis, K. 1985. Calculating the variance of seasonally adjusted series. *Journal of the American Statistical Association* 80, 541–52.

Cappé, O., Moulines, E. and Ryden, T. 2005. *Inference in Hidden Markov Models*. New York: Springer-Verlag.

Chambers, M. 1998. Long memory and aggregation in macroeconomic time series. *International Economic Review* 39, 1053–72.

Christiano, L. and Fitzgerald, T. 2003. The band pass filter. *International Economic Review* 44, 435–65.

Christoffersen, P. and Diebold, F. 1996. Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics* 11, 561–71.

Christoffersen, P. and Diebold, F. 1997. Optimal Prediction under asymmetric loss. *Econometric Theory* 13, 808–17.

Cournot, A. 1838. Researches into the Mathematical Principles of the Theory of Wealth, trans. N. Bacon. New York: Macmillan, 1927.

Diebold, F. and Inoue, A. 2001. Long memory and regime switching. *Journal of Econometrics* 105, 131–59.

Doob, J. 1953. *Stochastic Processes*. New York: John Wiley.

Durbin, J. and Koopman, S. 2001. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Durlauf, S. and Phillips, P. 1988. Trends versus random walks in time series analysis. *Econometrica* 56, 1333–54.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Elliott, G. 1998. The robustness of co-integration methods when regressors almost have unit roots. *Econometrica* 66, 49–58.

Engle, R. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* 50, 987–1008.

Engle, R. and Granger, C. 1987. Co-integration and error correction: Representation, estimation and testing. *Econometrica* 55, 251–76.

Engle, R. and Granger, C. 1991. *Long Run Economic Relations: Readings in Co-integration*. Oxford: Oxford University Press.

Faust, J. 1998. The robustness of identified VAR conclusions about money. *Carnegie-Rochester Conference Series on Public Policy* 49, 207–44.

Faust, J., Swanson, E. and Wright, J. 2004. Identifying VARs based on high frequency futures data. *Journal of Monetary Economics* 51, 1107–31.

Fishman, G. 1969. *Spectral Methods in Econometrics*. Cambridge: Harvard University Press.

Franses, P. 1996. *Periodicity and Stochastic Trends in Economic Time Series*. Oxford: Oxford University Press.

Fuller, W. 1976. *Introduction to Statistical Time Series*. New York: John Wiley.

Geweke, J. and Porter-Hudak, S. 1983. The estimation and application of long memory series models. *Journal of Time Series Analysis* 4, 221–38.

Ghysels, E. and Osborn, D. 2001. *The Econometric Analysis of Seasonal Time Series*. Cambridge: Cambridge University Press.

Granger, C. 1966. The typical spectral shape of an economic variable. *Econometrica* 34, 150–61.

Granger, C. 1969. Prediction with a generalized cost of error function. *Operations Research Quarterly* 20, 199–207.

Granger, C. 1980. Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* 14, 227–38.

Granger, C. 1981. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–30.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Granger, C. and Newbold, P. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.

Granger, C. and Newbold, P. 1977. *Forecasting Economic Time Series*. New York: Academic Press.

Granger, C. and Teräsvirta, T. 1993. *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Grether, D. and Nerlove, M. 1970. Some properties of ‘optimal’ seasonal adjustment. *Econometrica* 38, 682–703.

Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–84.

Hannan, E. 1960. *Time Series Analysis*. London: Methuen.

Hannan, E. 1969. The estimation of mixed moving average autoregressive systems. *Biometrika* 56, 223–5.

Hannan, E. 1970. *Multiple Time Series*. New York: John Wiley.

Hannan, E. 1971. The identification problem for multiple equation systems with moving average errors. *Econometrica* 39, 751–65.

Hannan, E. 1976. The identification and parameterization of ARMAX and state space forms. *Econometrica* 44, 713–23.

Hannan, E. 1979. The statistical theory of linear systems. In *Developments in Statistics*, ed. P. Krishnaiah. New York: Academic Press.

Harvey, A. 1981. *Time Series Models*. Oxford: Allan.

Harvey, A. 1984. A unified view of statistical forecasting procedures. *Journal of Forecasting* 3, 245–75.

Harvey, A. 1989. *Forecasting. Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. and Peters, S. 1990. Estimation procedures for structural time series models. *Journal of Forecasting* 9, 89–108.

Harvey, A. and Todd, P. 1984. Forecasting economic time series with structural and Box–Jenkins

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
models: a case study (with discussion). *Journal of Business and Economic Statistics* 1, 299–315.

Hatanaka, M. 1975. On the global identification of the dynamic simultaneous equations model with stationary disturbances. *International Economic Review* 16, 545–54.

Hillmer, S., Bell, W. and Tiao, G. 1983. Modeling considerations in the seasonal analysis of economic time series. In *Applied Time Series Analysis of Economic Data*, ed. A. Zellner. Washington, DC: Bureau of the Census, Department of Commerce.

Hillmer, S. and Tiao, G. 1982. An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* 77, 63–70.

Hodrick, R. and Prescott, E. 1997. Postwar US business cycles: an empirical investigation. *Journal of Money, Credit and Banking* 29, 1–16.

Holt, C. 1957. Forecasting seasonals and trends by exponentially weighted moving averages. ONR Research Memorandum No. 52, Carnegie Institute of Technology.

Hylleberg, S. 1992. *Modelling Seasonality*. Oxford: Oxford University Press.

Jevons, W. 1884. *Investigations in Currency and Finance*. London: Macmillan.

Johansen, S. 1995. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Kalman, R. 1960. A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers – Journal of Basic Engineering* D 82, 35–45.

King, R., Plosser, C., Stock, J. and Watson, M. 1991. Stochastic trends and economic fluctuations. *American Economic Review* 81, 819–40.

Kolmogorov, A. 1941. Interpolation und Extrapolation von stationären zufälligen Folgen. *Bulletin of the Academy Science (Nauk), USSR, Mathematical Series* 5, 3–14.

Koopmans, L. 1974. *The Spectral Analysis of Time Series*. New York: Academic Press.

Kuznets, S. 1961. *Capital and the American Economy: Its Formation and Financing*. New York: Princeton University Press for the National Bureau of Economic Research.

Maravall, A. 1981. Desestacionalization y Politica Monetaria. Economic Studies No. 19. Madrid: Bank of Spain.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Maravall, A. 1984. Model-based treatment of a manic depressive series. Working paper, Bank of Spain.

Muth, J. 1960. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55, 299–305.

Nelson, C. and Kang, H. 1981. Spurious periodicity in inappropriately detrended time series. *Econometrica* 49, 741–52.

Nelson, C. and Plosser, C. 1982. Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics* 10, 139–62.

Nerlove, M. 1964. Spectral analysis of seasonal adjustment procedures. *Econometrica* 32, 241–86.

Nerlove, M. 1965. A comparison of a modified Hannan and the BLS seasonal adjustment filters. *Journal of the American Statistical Association* 60, 442–91.

Nerlove, M. 1967. Distributed lags and unobserved components in economic time series. In *Ten Economic Essays in the Tradition of Irving Fisher*, ed. W. Fellner et al. New York: John Wiley.

Nerlove, M., Grether, D. and Carvalho, J. 1979. *Analysis of Economic Time Series*. New York: Academic Press.

Nerlove, M. and Wage, S. 1964. On the optimality of adaptive forecasting. *Management Science* 10, 207–24.

Newbold, P. 1974. The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika* 61, 423–6.

Parzen, E. 1961. An approach to time series analysis. *Annals of Mathematical Statistics* 32, 951–89.

Phillips, P. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–40.

Phillips, P. and Durlauf, S. 1986. Multiple time series regression with integrated processes. *Review of Economic Studies* 53, 473–95.

Phillips, P. and Hansen, B. 1990. Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies* 57, 99–125.

Pierce, D. 1978. Seasonal adjustment when both deterministic and stochastic seasonality are present. In *Seasonal Analysis of Economic Time Series*, ed. A. Zellner. Washington, DC: Bureau of the Census,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Department of Commerce.

Pierce, D. 1979. Signal extraction error in nonstationary time series. *Annals of Statistics* 7, 1303–20.

Priestley, M. 1981. *Spectral Analysis and Time Series*. New York: Academic Press.

Raftery, A., Madigan, D. and Hoeting, J. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92, 179–91.

Ravn, M. and Uhlig, H. 2002. On adjusting the HP-filter for the frequency of observations. *Review of Economics and Statistics* 84, 371–76.

Rigobon, R. 2003. Identification through heteroskedasticity. *Review of Economics and Statistics* 85, 777–92.

Rudebusch, G. 1993. The uncertain unit root in real GNP. *American Economic Review* 83, 264–72.

Sargan, J. and Bhargava, A. 1983. Maximum likelihood estimation of regression models with moving average errors when the root lies on the unit circle. *Econometrica* 51, 799–820.

Schuster, A. 1898. On the investigation of hidden periodicities with application to the supposed 26-day period of meteorological phenomena. *Terrestrial Magnetism and Atmospheric Electricity* [now *Journal of Geophysical Research*] 3, 13–41.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.

Schweppe, F. 1965. Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory* 11, 61–70.

Shapiro, M. and Watson, M. 1988. Sources of business cycle fluctuations. *NBER Macroeconomics Annual* 3, 111–56.

Shephard, N. 2005. *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Sims, C. 1986. Are forecasting models usable for policy analysis? *Quarterly Review, Federal Reserve Bank of Minneapolis* 10, 2–16.

Sims, C., Stock, J. and Watson, M. 1991. Inference in linear time series models with some unit roots. *Econometrica* 58, 113–44.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Slutsky, E. 1927. The summation of random causes as the source of cyclic processes. *Econometrica* 5, April 1937, 105–46.

Sowell, F. 1992. Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics* 53, 165–88.

Stock, J. 1991. Confidence intervals for the largest autoregressive root in US economic time series. *Journal of Monetary Economics* 28, 435–60.

Stock, J. and Watson, M. 1993. A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* 61, 783–820.

Stock, J. and Watson, M. 2002a. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–79.

Stock, J. and Watson, M. 2002b. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20, 147–62.

Stokes, G. 1879. Note on searching for hidden periodicities. *Proceedings of the Royal Society* 29, 122–5.

Theil, H. and Wage, S. 1964. Some observations on adaptive forecasting. *Management Science* 10, 198–206.

Wallis, K. 1977. Multiple time series analysis and the final form of econometric models. *Econometrica* 45, 1481–97.

Whittle, P. 1963. *Prediction and Regulation by Linear Least-squares Methods*. London: English Universities Press.

Wiener, N. 1949. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. New York: John Wiley.

Winters, P. 1960. Forecasting sales by exponentially weighted moving averages. *Management Science* 6, 324–42.

Wold, H. 1938. *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksell.

Yule, G. 1921. On the time-correlation problem, with special reference to the variate-difference correlation method. *Journal of the Royal Statistical Society* 84, 497–526.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Yule, G. 1926. Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89, 1–64.

Yule, G. 1927. On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London A* 226, 267–98.

Zellner, A. and Palm, F. 1974. Time series analysis and simultaneous equation econometric models. *Journal of Econometrics* 2, 17–54.

How to cite this article

Diebold, Francis X., Lutz Kilian and Marc Nerlove. "time series analysis." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000064> doi:10.1057/9780230226203.1707

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

time use

Duncan Ironmonger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Time is a finite, irreplaceable resource, and unlike money is equally distributed. Statistics of time use and money use can be combined, giving two-dimensional pictures of individual economic behaviour and the national economy. A framework for the *microeconomics* of time use and household production was established in the 1960s. In coming decades, a *macroeconomics* of time use and household production will arise. Within a production, consumption and investment framework, this will employ continuous national time accounts and satellite accounts of the household economy. Consequently the household's true economic role and its powerful interactions with the market will be revealed.

Keywords

American Time Use Survey (ATUS); Bureau of Labor Statistics (BLS); child care; consumption activities; diary-based surveys of time use; education; Gross Economic Product (GEP); Gross Household Product (GHP); Gross Market Product (GMP); household economy; household production; household productivity; household technology; human capital activities; investment in human capital; labour statistics; leisure; macroeconomic modelling of the household economy; microeconomics of household production and consumption; Multinational Time Use Study (MTUS); national time accounts; Paid work; production activities; Reid, M.; Robbins, L.; satellite accounts of the household economy; social interaction and recreation; System of National Accounts (SNA); third-person test; time income and expenditure; Time Resources and Time Use Account; time use; work-leisure trade-off

Article

Time is a finite, irreplaceable resource available to every man, woman and child in equal amounts of 168 hours per week over the course of life. Time use refers to the allocation of time to alternative uses such as sleep, leisure or work.
Time is perhaps the most fundamental scarce resource; unlike money income or wealth, it is equally

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
distributed; and how well or wastefully it is used largely determines the progress, achievement and well-being of individuals, families, communities and societies.

One way to assess progress and well-being is to measure aggregate changes in the uses of time rather than, or in addition to, the usual monetary statistics of national income and expenditure. National time accounts are more comprehensive than money accounts as they simultaneously measure the productive time spent in both the market and the household, as well as the time spent in consumption of outputs from both.

The advent of radio astronomy in 1957, using a wider spectrum than that provided by the visible light frequencies, opened up new views of the astronomical universe. Similarly, official statistical organizations are beginning to provide new views of the economic and social universe by observing the world through a ‘time frequency’ rather than the more visible ‘money frequency’.

Researching time use and human behaviour

Time use economics is the scientific study of human behaviour concerning the allocation of time between alternative uses (cf. Robbins, 1932, p. 16).

Time use can be usefully considered under the three comprehensive categories of production, consumption and investment. These accord with the theoretical framework of macroeconomics. Time use data show how individuals and households allocate their time between production (paid work and unpaid work), consumption (meals, television, social interaction and recreation) and human capital formation and maintenance (education, self-care and sleep).

Margaret Reid provided the well-known ‘third person’ test for separating production from consumption: ‘If an activity is of such character that it might be delegated to a paid worker, then that activity shall be deemed productive’ (Reid, 1934, p. 11).

To facilitate analysis, macroeconomics makes a further distinction between resources for immediate use – consumption – and those for use over a longer time period – investment. Accordingly, time spent learning a skill or gaining knowledge in education is clearly investment in human capital. Similarly, time spent in sleep and self-care can be regarded as a necessary daily investment to maintain functioning minds and bodies.

When people spend time, in economic terms they are effectively allocating time between market production, household production, consumption and investment. Modelling of household time-allocation decisions goes beyond understanding the simple work-leisure trade-off. It provides knowledge of the detailed interactions between production, consumption and investment activities. It also provides a framework for the analysis of the derived demands for market commodities implicit in household production and consumption.

In a much-quoted article in *Scientific American*, Vanek (1974) surveys the time spent in household production in the United States over a period of 40 years. Juster and Stafford (1991) provides a comprehensive appraisal of the importance of time allocation as an analytic construct and a review of what had been learned from time allocation data in modelling economic behaviour and the dynamics of economic change. Robinson and Godbey (1999) give an account of the American data from 1965, 1975, 1985 and 1995, and Gershuny (2000) comprehensively surveys the time use data for Britain.

Becker (1965) offers a theoretical framework for the microeconomics of time use by including the cost

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of time on the same footing as the cost of market goods and by assuming households ‘combine time and market goods to produce more basic commodities that directly enter their utility functions’ (1965, p. 494). This neoclassical approach to a theory has been criticized on many grounds, summarized recently by Folbre (2004).

While neoclassical micro-theory provides a useful framework for a *microeconomics of household production and consumption*, it completely disregards the full macroeconomic magnitude of household production. In most developed countries more labour is required for household production than for market production (Goldschmidt-Clermont and Pagnossin-Aligisakis, 1995). Indeed, if the unpaid time spent in caring for children is fully measured, including the ‘parallel’ or ‘secondary’ time when other activities are simultaneously undertaken, total work in household production is considerably greater than market work (Ironmonger, 2004).

Existing official labour statistics are misleading indicators of total work as they ignore labour inputs to household production. They measure only that time spent on activities within the production boundary of the System of National Accounts (SNA). A broader definition of work includes unpaid work outside the SNA boundary but within the general production boundary. This non-SNA work is employment in household production – the provision of meals, clean clothes, accommodation, care and transport by households for themselves or other households without remuneration (Ironmonger, 2001).

Time use economics is much more than the microeconomics of choice. It includes the study of the large non-monetary economy where households combine time, *their own capital* and intermediate inputs from the market to produce services that compete with market services. Including the use of household capital, the Gross Household Product (GHP) of the household economy is comparable to the Gross Market Product (GMP) of the market (Ironmonger, 1996a).

Time use data will provide the raw materials for a new *macroeconomics of the household economy*. A principal objective of the recent Eurostat and United States initiatives to collect more time use data has been not only to value unpaid work but to produce satellite accounts giving monetary valuations of the household economy (Varjonen et al., 1999). To be effective as a new set of tools for macroeconomic modelling of the household economy, these accounts and the national time accounts will need to be a continuous time series of annual or even quarterly data covering a run of years.

Considering its enormous magnitude, the performance and rate of growth of the household economy deserve as much attention as that given to the market economy. Macroeconomic modelling of the household economy could help explain the rate of change of household production, the impact of changes in household technology on productivity, and the effects of monetary and fiscal policies on GHP and its broad components.

Modelling of the total economic system – especially the dynamic interactions between the market and household economies – will be greatly improved by using a continuous record of time use taking account of changing economic, social and technological factors. Booms and slumps affect the mix between paid and unpaid work in the market and household economies; new products and technologies can quickly affect the relative productivity of these sectors, and tax policies can alter the relative attractiveness of paid work, unpaid work and leisure.

Using national time accounts and satellite accounts of household production, a macroeconomic model of the household economy could be constructed and linked with a macroeconomic model of the market (Ironmonger, 1994; 1996b). Such a comprehensive model could investigate the cyclical and long-run

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
relationships between market and household production. One hypothesis worth verification is that the two spheres of production vary in a counter-cyclical pattern (Ironmonger, 1989). If this is so, the amplitude of the fluctuation of Gross Economic Product (GEP, the sum of GHP and GMP) will be less than either component.

Measuring time use

The accurate scientific measurement of how people use time began with independent surveys in a number of countries, particularly in the USSR and the United States in the 1920s. A major advance was made in 1965 when Alexander Szalai directed internationally comparable, diary-based surveys in 12 countries under the sponsorship of UNESCO and the International Social Science Council (Szalai, 1972). Subsequent official measurement of time use included Norway at ten-year intervals from 1971, the Netherlands at five-year intervals from 1975, Canada in 1981, 1986, 1992 and 1998, and Australia in 1987, 1992 and 1997. Several developing countries – for example, India in 1998–9 and South Africa in 2001 – have now conducted official time use surveys.

This development in scientific measurement culminated in the harmonized surveys across some 20 European countries in 1998–2003. They used what is regarded as the most valid technique to measure time use – a diary recording the chronology of various time uses over one or more days from a representative sample of the population. The European surveys collected one weekday and one weekend day from people aged ten upwards; the Indian surveyed one day from age six upwards. Although the 1965 surveys were in one city in one month of the year, the subsequent official surveys cover the whole population, urban and rural, and all seasons of the year.

Research on time use by universities, government agencies and private business has been greatly facilitated by access to the unit record files of individual diary days. The Multinational Time Use Study (MTUS) contains a growing collection of these files covering 78 surveys in 27 countries, including the original 12 from the 1965 cross-national time budget study. The ‘World’ files of MTUS are a particularly rich resource for international comparison, as data have been made as comparable as possible by providing common categories of time use and standard definitions of demographic and household variables.

In addition to the diary-based surveys of time use, some official statistical organizations make estimates, on a continuous basis, of time used in the market economy. For example, in Australia, since 1966, the monthly household survey of employment and unemployment also provides estimates of average hours of market work. In this survey the interviewer's question is: ‘How many hours did you work last week in (all) your job(s)?’ This is a ‘stylized’ question subject to biases that differ from the diary-based method. Each method – detailed diary or stylized question – has its own bias against ‘reality’.

With the inauguration of the American Time Use Survey (ATUS) in January 2003, the Bureau of Labor Statistics and the US Census Bureau took a giant stride in the scientific measurement of time use. In the world's first continuous survey, diary data on time use are collected each month from a representative sample of adults aged 15 or more years.

This groundbreaking response to fill the biggest single gap in the Federal Statistical System of the United States was 12 years in development. It traces its origin to 1991 when a bill introduced in the 102nd Congress called for the Bureau of Labor Statistics to ‘conduct time-use surveys of unremunerated

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
work performed in the United States and to calculate the monetary value of such work' (Horrigan and Herz, 2004).

Table 1 shows the Time Resources and Time Use Account for the United States of America for 2003 derived from the first results of ATUS and arranged according to the macroeconomic categories of production, consumption and investment. Estimates of the time use of children under 15 based on other surveys have been included to complete the table.

National time accounts: Time Resources and Time Use Account, United States of America, 2003

	Women	Men	Children	Women	Men	Children	Total
Population (million)				118.1	111.9	60.7	290.7
	Average hours per week			Million hours per week			
Time Resources							
Total time income	**168.0**	**168.0**	**168.0**	**19,841**	**18,799**	**10,198**	**48,838**
Time Use							
Time expenditure							
Production activities							
Household economy (Non-SNA production)	32.6	20.3	*5.0*	3,850	2,272	*304*	6,425*
Market economy (SNA production)	20.1	32.0	-	2,374	3,581	-	5,955
Total production activity	**52.7**	**52.3**	*5.0*	**6,224**	**5,952**	*304*	**12,380**
Consumption activities							
Eating and drinking	8.3	8.7	*10.0*	980	974	*607*	2,561
Watching TV	16.8	19.3	*26.0*	1,984	2,160	*1,578*	5,722
Social and recreation	16.9	18.6	*24.0*	1,996	2,081	*1,457*	5,534
Total consumption activity	**42.0**	**46.6**	*60.0*	**4,960**	**5,215**	*3,642*	**13,817**
Human capital activities							
Education	3.5	3.2	*22.0*	413	358	*1,335*	2,107
Self-care	6.2	4.6	*5.0*	732	515	*304*	1,550
Sleep	60.6	59.3	*74.0*	7,157	6,636	*4,492*	18,284
Total human capital activity	**70.3**	**67.1**	*101.0*	**8,302**	**7,508**	*6,131*	**21,942**
Telephone, etc.	3.0	2.0	*2.0*	354	224	*121*	700
Total time expenditure	**168.0**	**168.0**	**168.0**	**19,841**	**18,799**	**10,198**	**48,838**

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

* ATUS provides estimates of secondary child care, but restricted to times when respondents are awake and children are not in bed. If these estimates are used, this figure would increase by about 19 per cent to 7,600 million hours per week if, without double-counting other household work, the hours spent in caring for children younger than 13 years as a secondary activity is included. Most of his extra time occurred simultaneously with consumption activity but some overlapped with market work, education and self-care.

Source: Estimates of the Households Research Unit, Department of Economics, University of Melbourne based on Bureau of Labor Statistics, Time-Use Survey – First Results Announced by BLS, 14 September 2004, http://www.bls.gov/tus/home.htm and Population Division, US Census Bureau.

As the way children spend time is a critical issue for many research and policy purposes, time use data should be extended to include children. An innovative feature of the new longitudinal study of Australian children is a time-use diary where parents record details of what their child did in two 24-hour periods (Australian Institute of Family Studies, 2005).

Conclusion

The development of national time accounts and satellite accounts of the household economy should be an interactive process between researchers, model builders, policymakers and official statisticians. The national money income and expenditure accounts developed this way.

At the beginning of the 21st century statistical measurements are starting to provide the raw materials for a *macroeconomics of the household economy*. New continuous time use observations will provide regular national time accounts and satellite accounts of household production.

The new data will enable the building of more relevant economy-wide models to explore the continuing interactions between the household and the formal market and public sectors of the economic system. The true macroeconomic role of the household will become clear, not only as the place for consumption and leisure but as the largest user of labour time in economic production. The new household-based models should provide better understanding of issues of work and leisure, and produce policies conducive to a more equitable and satisfactory allocation of time.

See Also

- Becker, Gary S.
- childcare
- household production and public goods
- household surveys
- human capital
- intrahousehold welfare
- leisure
- national accounting, history of
- Robbins, Lionel Charles

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- value of time
- women's work and wages

Bibliography

Australian Institute of Family Studies. 2005. *Growing Up in Australia: The Longitudinal Study of Australian Children: 2004 Annual Report*. Canberra: AIFS.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75, 493–517.

Folbre, N. 2004. A theory of the misallocation of time. In *Family Time: The Social Organization of Care*, ed. N. Folbre and M. Bittman. London: Routledge.

Gershuny, J. 2000. *Changing Times*. Oxford: Oxford University Press.

Goldschmidt-Clermont, L. and Pagnossin-Aligisakis, E. 1995. *Measures of Unrecorded Economic Activities in Fourteen Countries*. Occasional Paper No. 20. New York: Human Development Report Office, UNDP.

Horrigan, M. and Herz, D. 2004. Planning, designing, and executing the BLS American time-use survey. *Monthly Labor Review* 127(10), 3–19.

Ironmonger, D. 1989. Households and the household economy. In *Households Work: Productive Activities, Women and Income in the Household Economy*, ed. D. Ironmonger. Sydney: Allen and Unwin.

Ironmonger, D. 1994. Modeling the household economy. In *Economics, Econometrics and the LINK: Essays in Honor of Lawrence R. Klein*, ed. M. Dutta. Amsterdam: North Holland.

Ironmonger, D. 1996a. Counting outputs, capital inputs and caring labor: estimating Gross Household Product. *Feminist Economics* 2, 37–64.

Ironmonger, D. 1996b. Time use and satellite accounts for modelling the household economy. Paper presented at the IARIW 24th General Conference, Lillehammer, Norway, August.

Ironmonger, D. 2001. Household Production. In *International Encyclopedia of the Social & Behavioral Sciences*, vol 10, ed. N. Smelser and P. Baltes. Oxford: Elsevier Science.

Ironmonger, D. 2004. Bringing up Bobby and Betty: the inputs and outputs of childcare time. In *Family Time: The Social Organization of Care*, ed. N. Folbre and M. Bittman. London: Routledge.

Juster, F. and Stafford, F. 1991. The allocation of time: empirical findings, behavioural models, and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
problems of measurement. *Journal of Economic Literature* 29, 471–522.

Reid, M. 1934. *Economics of Household Production*. New York: John Wiley.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Robinson, J. and Godbey, G. 1999. *Time for Life: The Surprising Ways Americans Spend Their Time*. University Park: Penn State Press.

Szalai, A., ed. 1972. *The Uses of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries*. The Hague: Mouton.

Vanek, J. 1974. Time spent in housework. *Scientific American* 231, 116–20.

Varjonen, J., Niemi, I., Hamunen, E., Paakkonen, H. and Sandstrom, T. 1999. *Proposal for a Satellite Account of Household Production*. Working Paper No. 9/1999/A4/11. Brussels: Eurostat.

How to cite this article

Ironmonger, Duncan. "time use." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000212> doi:10.1057/9780230226203.1708

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Timlin, Mabel Frances (1891 – 1976)

Robert W. Dimand
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

American Economic Association; Canada, economics in; general equilibrium; Keynesianism; Timlin, M. F.

Article

The Keynesian economist Mabel Timlin was the first tenured woman among Canadian economists, the first woman elected president of the Canadian Political Science Association (which then covered all social sciences, including economics), the first woman outside the natural sciences elected a Fellow of the Royal Society of Canada (1951), and one of the first ten women to serve on the executive committee of the American Economic Association (1958–60), despite becoming an assistant professor only in her 50th year, after a long career as an academic secretary. She was born in Forest Junction, Wisconsin, on 6 December 1891, and, after studying at the Milwaukee State Normal School, taught in Wisconsin and rural Saskatchewan. She became a secretary at the University of Saskatchewan in 1921, while studying for a BA there. At first Timlin intended to study economics, but after seeing the Department of Economics and Political Science at Saskatchewan she decided (probably correctly) that she could learn more economics on her own. She took a BA with great distinction in English in 1929, and then directed the university's correspondence courses in economics. Mabel Timlin became an instructor in economics at the University of Saskatchewan in 1935, after completing graduate course work in economics at the University of Washington during summers and a six-month leave. Her doctoral dissertation at the University of Washington, supervised by the much younger Raymond Mikesell, was accepted in 1940 and published as *Keynesian Economics* (1942). In 1941, Timlin became an assistant professor of economics at the University of Saskatchewan (associate professor 1946, full professor 1950) and a member of the executive committee of the Canadian Political Science Association (Vice-President 1953–5, President 1959–60).

Keynesian Economics did more than introduce Keynesian theory into Canadian academic life. Timlin offered one of the early general equilibrium interpretations of John Maynard Keynes's *General Theory*, and was particularly noteworthy in treating it as a system of shifting equilibrium, presented with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
innovative diagrams on which she collaborated with the eminent geometer H.S.M. Coxeter. Timlin began work on *Keynesian Economics* in 1935, before Keynes published his *General Theory*: Benjamin Higgins had come to Saskatoon from the London School of Economics in 1935 for a one-year appointment, carrying a copy of the summary of Keynes's Cambridge lectures that Robert Bryce had presented in Friedrich Hayek's LSE seminar.

Beyond her work on Keynes, Timlin also expounded international developments in welfare economics and general equilibrium analysis to a Canadian audience more used to historical and institutional economics than to formal theory (for example, Timlin, 1946). Timlin (1953) sharply criticized the Bank of Canada for failing to follow Keynesian countercyclical stabilization policies during the Korean War inflation. Much of her later work (for example, Timlin, 1951; 1958; 1960) concerned immigration policy, emphasizing the economic benefits of freer immigration.

Mabel Timlin never married. Generations of former students were her extended family. She remained active as a scholar long after her official retirement in 1959, publishing a major report on the social sciences in Canada in 1968. She remained devoted to the University of Saskatchewan despite job offers from such institutions as the University of Toronto, and died in Saskatoon on 19 September 1976.

See Also

- Canada, economics in
- Keynesian revolution
- Keynesianism

Selected works

1942. *Keynesian Economics*. Toronto: University of Toronto Press.

1946. General equilibrium analysis and public policy. *Canadian Journal of Economics and Political Science* 12, 483–95.

1947. John Maynard Keynes. *Canadian Journal of Economics and Political Science* 13, 363–5.

1951. *Does Canada Need More People*? Toronto: Oxford University Press.

1953. Recent developments in Canadian monetary policy. *American Economic Review: Papers and Proceedings* 43, 42–53.

1955. Monetary stabilization policies and Keynesian theory. In *Post-Keynesian Economics*, ed. K.R. Kurihara. London: George Allen & Unwin.

1958. Canadian immigration with special reference to the post-war period. In *International Migration*. International Economic Association. London: Macmillan.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1960. Presidential address: Canada's immigration policy, 1896–1910. *Canadian Journal of Economics and Political Science* 26, 517–32.

1968. The social sciences in Canada: retrospect and prospect. In *The Social Sciences in Canada: Two Studies*, ed. M.F. Timlin and A. Faucher. Ottawa: Social Science Research Council of Canada.

1977. (With biographical note by A.E. Safarian and introduction by L. Tarshis.) *Keynesian Economics*. Toronto: McClelland and Stewart, Carleton Library.

Bibliography

Ainley, M.G. 1999. Mabel F. Timlin, 1891–1976: A woman economist in the world of men. *Atlantis: A Women's Studies Journal* 23, 28–38.

Spafford, S. 2000. *No Ordinary Academics: Economics and Political Science at the University of Saskatchewan, 1910–1960*. Toronto: University of Toronto Press.

How to cite this article

Dimand, Robert W. "Timlin, Mabel Frances (1891–1976)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000246>
doi:10.1057/9780230226203.1709

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tinbergen, Jan (1903 – 1994)

Peter A. Cornelisse and Herman K. van Dijk
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Jan Tinbergen was the first Nobel Laureate in economics in 1969. This article presents a brief survey of his many contributions to economics, in particular to macroeconometric modelling, business cycle analysis, economic policymaking, development economics, income distribution, international economic integration and the optimal regime. It further emphasizes his desire to contribute to the solution of urgent socio-economic problems and his passion for a more humane world.

Keywords

business cycles; capability tax; Club of Rome; cobweb model; computable general equilibrium models; development economics; econometrics; economic integration; educational attainment; foreign aid; Frisch, R.; gravitation model; Great Depression; Haavelmo, T.; Haberler, G.; Harvard Barometer; income distribution; inequality; inequality between nations; innovation; international division of labour; international external effects; Keynes, J. M.; lump sum taxes; Mitchell, W.; models; Netherlands Bureau for Economic Policy Analysis; optimal distribution of income; Persons, W.; planning in stages; positional-exchange criterion; rent seeking; Slutzky, E.; social welfare function; statistical mechanics; targets and instruments; tax shifting; taxation of income; Tinbergen rule; Tinbergen, J.; velocity of integration; Yule, G.

Article

Jan Tinbergen was born in The Hague, The Netherlands, on 12 April 1903, the first of five children in an intellectually stimulating family with a love of foreign languages. Eventually two of the children would win a Nobel Prize: Jan in economics (in 1969) and Niko, an ethologist, in physiology or medicine (in 1973).

Jan Tinbergen enrolled as a student of mathematical physics at Leiden University in 1921, where he obtained his doctorate in 1929. By that time he had already decided to switch to economics. From 1926

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to 1928 Tinbergen worked as a conscientious objector to national military service, first in a convict prison and later, and of greater import to his subsequent career, at the Central Bureau of Statistics. He continued to work there until 1945. In 1933 he became extraordinary professor of statistics, mathematical economics and econometrics at the Netherlands School of Economics in Rotterdam. As a result of his quantitative approach to the study of economic dynamics, he was invited to the League of Nations in Geneva during the period 1936–8 in order to carry out statistical tests of business cycles theories. In 1945, at the end of the Second World War, Tinbergen was appointed as the first director of the Central Planning Bureau at The Hague. He held this position until 1955, when he became full professor of mathematical economics and development planning at the Netherlands School of Economics, later Erasmus University, Rotterdam. Throughout the 1960s and a part of the 1970s he acted as adviser to various international organizations and to governments of a considerable number of less-developed countries. He was elected chairman of the United Nations Committee on Development Planning in 1965 and held this position until 1972. In 1969 he was awarded, together with Ragnar Frisch, the first Nobel Prize in Economics. After his retirement as full professor in 1973 he held the Cleveringa Chair in Leiden for two years. He continued to be involved in various research projects at old age. Jan Tinbergen died on 9 June 1994.

Personal motivation

Already at an early age Tinbergen was profoundly impressed by the horrors of the Great War – subsequently numbered as the First World War – partly because of the fate of the Austrian refugee children his parents had lodged. Later, in Leiden as a student, when he was invited by his postman to join him on his rounds, he was appalled by the conditions of poverty in which the local population lived. Wishing to contribute to the struggle against such social evils, he decided to become an economist. This decision was characteristic of Tinbergen and his attitude towards economic science in his later life: his scientific contributions would always be inspired by the desire to tackle the social problems he observed. Paul Ehrenfest, professor of theoretical physics and Tinbergen's mentor in Leiden, was not unsympathetic towards the switch from physics to economics. Having made important contributions to statistical mechanics together with his wife Tatyana Afanasyeva, he called Tinbergen's attention to the possibilities that a mathematical representation of economic problems would offer. The dissertation on minimum problems in physics and economics that Tinbergen defended in 1929 bridged the two disciplines.

Econometric modelling and business cycle research

In 1969 Tinbergen was awarded, together with R. Frisch, the first Nobel Prize in Economics ‘for having developed and applied dynamic models for the analysis of economic processes’, as the Nobel Prize committee described it.
The desire to combat the socio-economic consequences of the Great Depression of the 1930s was Tinbergen's most important motivation for studying business cycles. In his inaugural address as extraordinary professor in 1933 he summarized his project as ‘statistics and mathematics in the service of business cycle research’. His approach contrasted with previous approaches to business cycle research

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(for more details, see, for example, Morgan, 1990 and Jolink, 2003). After a 19th-century undertaking by Juglar (1862) ascribing the recurrent business crises in Europe and North America to credit crises, and Jevons's (1884) study pointing to agricultural production cycles connected with sunspot numbers, several research projects in the early 20th century were devoted to the construction of so-called business cycle barometers. The purpose was to measure economic fluctuations through a particular index (or set of indices) with the aim of giving warning signals for turning points that would lead to a depression. An example was the Harvard Index of Business Conditions, informally known as the Harvard Barometer, constructed by a team led by Persons (1919). Another well-known descriptive approach to the business cycle during this period had been initiated by Mitchell (1913). His work was followed by that of Yule (1927) and Slutzky (1927), who suggested that the cumulative effect of random shocks could be the cause of cyclical patterns in economic variables. Frisch (1933), co-recipient of the 1969 Nobel Prize, applied these ideas introducing econometric models in which impulse propagation mechanisms led to business cycles.

However useful it could be as a starting point, Tinbergen criticized descriptive analysis as being too vague for use in policy preparation, and started a quantitatively oriented research programme to explore the possible economic causes of the periodic upswings and downswings in economic activity. In an earlier theoretical study Aftalion (1927) had argued that lags in an economic model could generate cyclical variation in economic activity. Following up this argument, Tinbergen specified a first simple case using a system of difference equations to express lagged responses of supply to price changes in a market for a single good. He noted that the systematic fluctuations that could arise in such a system had been observed in an empirical study of the pork market by the German economist Hanau (1928), a phenomenon that became known as the ‘cobweb model’ (Tinbergen, 1979, presents additional relevant literature).

Tinbergen subsequently generalized the specification of dynamic equations with lagged adjustment processes to macroeconomic settings, arguing that fluctuations in components of national product, such as investment and consumption expenditures, would lead to business cycle fluctuations in general economic activity. In 1936 he published the first applied macroeconometric model (for the Netherlands). It was a dynamic model, consisting of 22 equations in 31 variables. Employing what we now see as basic statistical techniques like correlation and regression analysis, it was meant to be used for the analysis of the particularly pressing unemployment problem. The specification of consumption and employment in this model anticipated elements of Keynes's theory (1936). This modelling exercise resulted in a strong policy recommendation in favour of a devaluation of the Dutch guilder to tackle unemployment. But its importance for the economics profession was far more profound: for the first time the economic-policy debate had been based on empirically tested, quantitative economic analysis and not on rather informally stated economic theory, the so-called verbal approach. Thus, according to Solow (2004, p. 159), Tinbergen's work during this period ‘was a major force in the transformation of economics from a discursive discipline into a model-building discipline’.

In 1936 Haberler had published a survey of theories on business cycles for the League of Nations. As a follow-up, and in reaction to the dynamic model for the Netherlands Tinbergen had published in that year, the same institution invited him to examine statistically which factors could be considered to contribute most to macroeconomic fluctuations. This project resulted in his two-volume book *Statistical Testing of Business Cycles Theories* (1939). The first volume contained a description of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

methodology applied, while the second volume presented a dynamic macroeconometric model for the United States with the aim of studying business cycles in that country after the First World War. This model was not only considerably larger than the one for The Netherlands; as imports and exports were much less important for the United States, it also allowed a relatively undisturbed view of internal dynamic mechanisms. Subsequently, the US model was much refined and enlarged by Klein (1950) and Duesenberry et al. (1965). Tinbergen presented his views on the dynamics of business cycles and on objectives and instruments of business-cycle policy for a wider audience in Tinbergen (1943) and Tinbergen and Polak (1950).

Discussion with Keynes

The formulation of some relations in Tinbergen's 1936 model showed some resemblance to Keynes's theory. Nevertheless, in an article in the *Economic Journal* of 1939, Keynes was remarkably sceptical of Tinbergen's work. Keynes labelled Tinbergen's method of estimating the parameters of an econometric model and computing quantitative policy scenarios as ‘statistical alchemy’, arguing that this approach ‘… is a means of giving quantitative precision to what, in qualitative terms, we know already as the result of a complete theoretical analysis’ (Keynes, 1939, p. 560). Their widely diverging views on the relevance of quantitative economic analysis were also illustrated by Keynes's reaction to Tinbergen's estimate of the price elasticity of demand for exports. When, in 1919, Keynes had strongly criticized the excessive war indemnity payments enforced upon Germany after the First World War, his argument had depended critically on the value of this elasticity. Tinbergen empirically found this value to be minus 2, precisely the value that Keynes had assumed a priori in his study. When informed about this Keynes replied: ‘How nice that you found the correct figure’ (Kol and de Wolff, 1993, p. 8.)

Keynes's critical attitude towards macroeconometric modelling and analysis originated from his view that the underlying economic theory should be complete in the sense that it should include all relevant variables and set out in detail its causal and dynamic structure. Econometrics could be used only for measuring the relations (‘curve fitting’ was the term used); it could not refute economic hypotheses or evaluate economic models. Tinbergen, on the other hand, argued that economic theories cannot be complete. Econometric research could be useful for scrutinizing elements of economic theories and for examining whether one theory describes reality better than another. Further, it could provide the numerical values of the coefficients in dynamic models that determine the cyclical and stability properties of a model, and, by applying a testing procedure of trial and error, it could yield suggestions for an improved specification of dynamic lags.

In this controversy Tinbergen's approach soon gained the upper hand as increasing numbers of economists, especially in the United States, noted its practical results in terms of model construction and verification, including forecasting. However, Keynes's comments on the role of expectations and uncertainty in macroeconometrics and on specification and simultaneous equation biases remained relevant. Haavelmo (1943) advocated the use of probability theory in bridging the gap between theory and data in business cycle analysis. Later these issues would become the subject of intensive debate and research.

Theory and practice of economic policy

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In 1945 Tinbergen was appointed director of the newly established Central Planning Bureau (CPB), an institution occupied with forecasting the effects of economic policy and advising the government on related matters (tasks which are more adequately captured by its present-day English name: Netherlands Bureau for Economic Policy Analysis). In the aftermath of the Second World War work at the CPB concentrated on the nation's pressing macroeconomic problems: a depleted capital stock, severe inflationary pressure, low levels of employment and an extreme shortage of foreign exchange.

In the economics discipline macroeconometric modelling had rapidly become accepted as a useful tool with the publication of such studies as by Klein (1950), Leontief (1950) and Klein and Goldberger (1955). But Tinbergen, having gained experience with the practice of policy preparation, felt the need for a systematic discussion of the logic of economic policy and of the use of models for policy purposes. It led to several monographs on the theory of economic policy (1952; 1956). He distinguished, among other things, between reforms (changes in the foundations of society), qualitative policy (changes in the structure of economic and social organization) and quantitative policy (changes in the instruments of economic policy). The latter could help to avoid the shortcomings of the traditional approach by offering a systematic policy where trial and error had been practised, by taking account of interdependence between instruments and by providing a quantitative indication of effects. Further, building on earlier work by Frisch distinguishing between various types of variables in relation to their role in policy models, Tinbergen demonstrated the connection between the analytical, or explanatory version and the policy, or normative version of economic models. In the analytical version, the policy targets were explained by other endogenous variables and by exogenous variables, which included the policy instruments. In the policy version the position of targets and instruments would be reversed (targets becoming exogenous and instruments endogenous variables) such that, in a well-behaved linear system, a solution requires only equality of the numbers of targets and instruments. This conclusion, which became known as the ‘Tinbergen rule’, brought an end to the popular misconception of a one-to-one correspondence between targets and instruments.

Development economics

In reaction to his experiences during a trip to India in 1951, Tinbergen left the Central Planning Bureau in 1955 and moved to the field of development economics, more specifically the planning of the socio-economic development of low-income countries. Much earlier he had published a mathematical-statistical study of the theory of long-term economic growth (1942), but this had related only to industrialized countries. In the model technological progress had explicitly been included and the statistical tests (with data for Great Britain, France, Germany and the United States from the decades before the First World War) already suggested that capital and labour growth could explain only a relatively small portion of the growth of production.

Characteristically, Tinbergen applied a quantitative, systematic policy approach to the development problem. This approach, which became known as ‘planning-in-stages’, distinguished macro, middle and micro stages, dealing with policy problems of private and public decision makers at the national, sectoral and project level, respectively (1967). In view of the difficult transportation conditions and the scarcity of skilled labour in developing countries, he subsequently added spatial and educational dimensions to the backbone of the planning-in-stages approach. He greatly simplified the calculation procedure for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
project evaluation by devising the semi-input-output method. This method was based on the notion that only the indirect effects emanating from sectors producing non-tradable (national) goods needed to be incorporated. At a time when computer capacity was still very limited, such a simplification was most useful. However, consistency between the micro stage and the other two levels was achieved only with the advent of computable general equilibrium models.

Tinbergen acted as adviser on matters related to economic development to the governments of Egypt, Turkey, Venezuela, Surinam, Indonesia and Pakistan, and he wrote studies for international organizations such as UNESCO and the OECD. As Chairman of the UN Committee on Development Planning from 1965 to 1972, he was involved with, among other things, the preparation of the UN Second Development Decade (1971–80).

Income distribution

Tinbergen revisited the field of income distribution after his retirement as full professor (1972a; 1975). His approach, then as much as before, was inspired to a considerable extent by the positional-exchange criterion that had emerged from discussions in his student days with Paul Ehrenfest. According to this criterion a distribution of welfare could be considered fair when no one would wish to take another person's position. It was, for example, expressed in the individual welfare function Tinbergen proposed, which depended negatively on the difference (positive or negative) between the level of schooling required for a job and the actual schooling obtained by the person on this job. The notion that an income distribution is the outcome of a confrontation of demand and supply factors was another characteristic element of his approach. Thus, the development of a country's income distribution would be governed to a large extent by the process of technological innovation (a demand factor) and the rise of educational attainment levels (a supply factor). On the basis of material from the United States and The Netherlands from 1900 onwards, he found that this ‘race’ was mostly won by the rise in education, which resulted in more equitable distributions.

In his contributions to the field of income distribution – which concentrated on the remuneration of labour categories – he aimed to examine the effect of some unorthodox propositions. One such proposition was to consider the applicability of a capability tax which, as a lump sum tax, would be preferable to the familiar income tax. (Remarkably, this proposal ran counter to his finding that tax changes have a very slight impact on primary incomes, such that tax shifting would hardly be a problem.) Further, and true to his conviction that scientific progress and practical applications depend on quantitative tests of hypotheses, he treated welfare as measurable on the assumption that further progress in this area would be feasible. Assuming that workers move freely from one job to another so utility would be equalized, he derived an empirical relation expressing the connection between wage income on the one hand and attained schooling and the difference between attained and required schooling on the other. He then used this relation to compute an optimal or just distribution of income, tentatively relating to the situation in The Netherlands in the early 1960s. It would require very considerable shifts in income as compared with the actual situation.

International economic integration

Tinbergen's earliest work on international economic relations was still connected with national

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
policymaking. Thus, his estimates of price elasticities of trade packages were meant to examine the effectiveness of a devaluation policy, where he emphasized the need to use long-term rather than short-term elasticities. His gravitation model (1962, Appendix VI) was a Newtonian approach to the explanation of bilateral trade flows which appeared to depend positively on the GNPs of the trade partners and negatively on the shipping distance separating them. It could be used to identify, among other things, the magnitude of potential trade lost to higher-than-average trade barriers, which impeded the efficient international division of labour he advocated in a number of studies written in the 1960s. Tinbergen (1954) applauded the international economic integration movement as it could remove trade barriers (which he dubbed negative economic integration) and could even result in new institutions for coordinated and centralized policymaking (positive economic integration). But he attached particular importance to the fact that economic integration would effectively reduce the probability of armed conflicts. From historical processes in Europe he derived a ‘velocity of integration’ which he hoped would remain positive until full integration at the regional and indeed the world level were achieved (1991a).

The optimal regime

His lifelong concern for (inter)national policymaking and, in that context, his special concern for the underdog resulted in a number of publications on the optimal economic order. In a deviation from his usual approach, Tinbergen emphasized in his Nobel Prize acceptance speech (1969) that the problem here consisted not of establishing the right mix of values of economic variables but of finding the proper set of institutions regarding the size and content of the public sector, the extent and content of (de) centralization of socio-economic decision making and therefore also of market regulation. He developed his ideas on the optimal order within a welfare-economic framework concerned with identifying the conditions that must be fulfilled to achieve maximum social welfare subject to the restrictions, such as production technologies, that apply in human society (1972b). In such a setting it would be useful to select the social welfare function at the beginning so as to limit the ethical possibilities in the subsequent analysis. The activities of the institutions would be described by a number of behaviour equations, the total of which should coincide with the conditions for optimal welfare. Tinbergen argued against rigidities, privileges, monopolies and insider-determined remunerations that bore no relation to marginal productivities, but he also rejected excessively generous social security systems that invited rent seeking. In Tinbergen's view the interests of developing countries deserved separate attention in discussions on the optimal economic order. No country would accept within its borders an income inequality between groups of rich and poor citizens as could be found between rich and poor countries in the world. Not only must obstacles to exports from developing countries be removed; it would also be necessary to support these countries’ development efforts by providing technical and financial aid. Tinbergen urged replacing the arbitrary UN target for international aid of 0.7 per cent of GNP of rich countries by the volume of aid that would be required for a harmonization of incomes within a predetermined number of years. He coordinated a study (1977) for the Club of Rome offering views on the international order, development aid, food production, the international division of labour, energy sources and raw materials, technological development, the environment and the arms race, among other things.
With the help of the theory of the optimal regime Tinbergen further sought to rid the confrontation between the Communist East and the Capitalist West of the dogmatic character that dominated world

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

politics before the fall of Communism in 1989. Horrified by the prospect of nuclear warfare, he devoted a large part of his later years to a plea for a rational debate on the pros and cons of both systems and for a stronger role of a reformed United Nations taking decisions that would incorporate international external effects (1990).

In conclusion

Tinbergen's contribution to the economics discipline lies in his pioneering work in a number of different economic fields. He would not consider himself an expert even in these areas, would gladly admit that others who had come in after him had meanwhile gained a better understanding, and would move on to another area where another pressing social problem needed to be addressed. In his own words (1991b), ‘solving the most urgent problems first’ is what moved him most in his intellectual agenda.

He had little patience with studies lacking applicability to practical problems, and was not much impressed by scientific elegance for its own sake. His work discipline, punctuality and efficiency were exemplary. For an appointment, students and assistants he supervised would get seven minutes on the watch he would keep nearby. Still, Tinbergen also gave innumerable lectures for organizations and social action groups even of humble status.

His intense desire for a more humane world led him to put great trust in the benevolence and effectiveness of governments and international organizations, realizing that policies to overcome social problems would nearly always require the participation of public institutions. The latter's serious shortcomings in terms of management and governance were just another problem to be solved. He nursed a strong hope that people would behave more sensibly over time and learn to avoid the terrible conflicts that had caused so much suffering and devastation in the 20th century. It was for all these characteristics that Samuelson (2004, p. 153) described Tinbergen as ‘a humanist saint’. Naturally, during his long life Tinbergen was often deeply disappointed. Still, his optimism never left him, if only because, as he said at an advanced age: ‘I cannot afford to be pessimistic’.

See Also

- development economics
- econometrics
- foreign aid
- Frisch, Ragnar Anton Kittel
- gravity models
- Keynes, John Maynard
- redistribution of income and wealth

The authors gratefully acknowledge pertinent comments offered by the editors and Marcel Boumans.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1933. *Statistiek en Wiskunde in Dienst van het Konjunktuuronderzoek* [Statistics and mathematics in the service of business cycle research]. Inaugural Lecture, Netherlands School of Economics, Rotterdam, 4 October. Amsterdam: Arbeiderspers.

1936. An economic policy for 1936. In *Jan Tinbergen: Selected Papers*, ed. L. Klaassen, L. Koyck and H. Witteveen. Amsterdam: North-Holland, 1959 (original in Dutch).

1939. *Statistical Testing of Business Cycle Theories. I: A Method and Its Application to Investment Activity. II: Business Cycles in the United States of America, 1919–1932*. Geneva: League of Nations.

1940. On a method of statistical business cycle research. A reply. *Economic Journal* 50, 141–54.

1942. Zur Theorie der Langfristigen Wirtschaftsentwicklung [On the theory of long-term economic growth]. *Weltwirtschaftliches Archiv* 55, 511–49.

1943. *Economische Bewegingsleer* [Economic dynamics]. Amsterdam: Noord-Hollandsche.

1950. (With J. Polak.) *The Dynamics of Business Cycles: A Study in Economic Fluctuations*. Chicago: University of Chicago Press.

1952. *On the Theory of Economic Policy*. Amsterdam: North-Holland.

1954. *International Economic Integration*. Amsterdam: Elsevier.

1956. *Economic Policy: Principles and Design*. Amsterdam: North-Holland.

1962. *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: The Twentieth Century Fund.

1967. *Development Planning*. London: World University Library.

1969. The use of models: experiences and prospects. In *Nobel Lectures, Economics 1969–1980*, ed. A. Lindbeck. Singapore: World Scientific Publishing Co., 1992. Online. Available at http://nobelprize.org/economics/laureates/1969/tinbergen-lecture.html, accessed 14 June 2006.

1972a. An Interdisciplinary Approach to the Measurement of Utility or Welfare. Fifth Geary Lecture for The Economic and Social Research Institute, Dublin.

1972b. Some features of the optimum regime. In *Optimum Social Welfare and Productivity*, ed. J. Tinbergen et al. New York: New York University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1975. *Income Distribution: Analysis and Policies*. Amsterdam: North-Holland.

1977. *RIO: Reshaping the International Order: A Report to the Club of Rome*. New York: New American Library.

1979. Recollections of professional experiences. *Banca Nazionale del Lavoro Quarterly Review* 131, 331–60.

1990. *World Security and Equity*. Aldershot, UK: Edward Elgar.

1991a. The velocity of integration. *De Economist* 139, 1–11.

1991b. Solving the most urgent problems first. In *Eminent Economists: Their Life Philosophies*, ed. M. Szedberg. Cambridge and New York: Cambridge University Press.

Bibliography

Aftalion, A. 1927. The theory of economic cycles based on the capitalistic technique of production. *Review of Economic Statistics* 9, 165–70.

Duesenberry, J., Fromm, G., Klein, L. and Kuh, E. 1965. *The Brookings Quarterly Econometric Model of the United States*. Chicago: Rand McNally.

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic Essays in Honour of Gustav Cassel*. London: Allen and Unwin.

Haberler, G. von. 1936. *Prosperity and Depression: A Theoretical Analysis of Cyclical Fluctuations*. Geneva: League of Nations.

Hanau, A. 1928. Die Prognose der Schweinepreise [Forecasting the prices of pork], *Vierteljahreshefte zur Konjunkturforschung*, Sonderheft 18. Berlin: Institut für Konjunkturforschung, 1930.

Haavelmo, T. 1943. Statistical testing of business-cycle theories. *Review of Economic Statistics* 25, 13–18.

Jevons, W. 1884. *Investigations in Currency and Finance*. London: Macmillan.

Juglar, C. 1862. *Des Crises Commerciales et de Leur Retour Périodique en France, en Angleterre et aux Etats-Unies*. Paris: Guillaumin.

Jolink, A. 2003. *Jan Tinbergen: The Statistical Turn in Economics, 1903–1955*. Rotterdam: CHIMES.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Keynes, J.M. 1919. *The Economic Consequences of the Peace*. London: Macmillan.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1939. Professor Tinbergen's method. *Economic Journal* 49, 558–68.

Keynes, J.M. 1940. Comment. *Economic Journal* 50, 154–6.

Klein, L. 1950. *Economic Fluctuations in the United States, 1921–1941*. New York: Wiley.

Klein, L. and Goldberger, A. 1955. *An Econometric Model of the United States, 1929–1952*. Amsterdam: North-Holland.

Kol, J. and de Wolff, P. 1993. Tinbergen's work: change and continuity. *De Economist* 141, 1–28.

Leontief, W. 1950. *The Structure of American Economy*. New York: Oxford University Press.

Mitchell, W. 1913. *Business Cycles and Their Causes*. Berkeley, California. University Memoirs, Vol. III.

Morgan, M. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Persons, W. 1919. An index of general business conditions. *Review of Economic Statistics* 1, 111–205.

Pronk, J. 2003. Tinbergen, idealist en inspirator. Lecture given on the occasion of the centennial of Tinbergen's birth. Erasmus University, Rotterdam, 9 April. Online. Available at http://www.janpronk.nl/index62.html, accessed 14 June 2006.

Samuelson, P. 2004. Homage to Jan Tinbergen. In *Economics with a Purpose: Tinbergen Centennial Issue*, ed. P. Cornelisse, H. van Dijk and H. Don. *De Economist* 152(2).

Solow, R. 2004. Progress in economics since Tinbergen. In *Economics with a Purpose: Tinbergen Centennial Issue*, ed. P. Cornelisse, H. van Dijk and H. Don. *De Economist* 152(2).

Slutzky, E. 1927. The summation of random causes as the source of cyclic processes. *The Problem of Economic Conditions* 3, 34–64 (English summary, 156–61).

Yule, G. 1927. On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London, A* 226, 267–98.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Cornelisse, Peter A. and Herman K. van Dijk. "Tinbergen, Jan (1903–1994)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000065> doi:10.1057/9780230226203.1710

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tintner, Gerhard (1907 – 1983)

Karl A. Fox
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

calculus of variations; dynamic economic theory; linear programming; probability; statistical inference; stochastic programming; technological risk; Tintner, G.; uncertainty; variate difference method

Article

Gerhard Tintner was born in Nuremberg, Germany, of Austrian parents, and educated in Vienna, completing his doctorate in economics, statistics and law at the University of Vienna in 1929. Tintner was much ahead of his time in important respects. First, he made early and significant contributions toward the development of a theory of behaviour under uncertainty (Tintner, 1941a; 1941b; 1942a; 1942b; 1942c). Second, he consistently stressed the need for a broad view of probability in the behavioural sciences and economics (Tintner, 1960; 1968). His seminal article ‘Foundations of Probability and Statistical Inference’ (1949) started from Carnap's view of probability as *degree of confirmation* and raised issues some of which are now being debated in current reformulations of econometric methodology (Harper and Hooker, 1976; Koch and Spizzichino, 1982). Third, he firmly believed that the tools of modern disciplines such as cybernetics and system theory should be adapted and used to gain insight into individual and social behaviour, which is the basis of all applied economic models (Tintner and Sengupta, 1972).

Tintner's first book (1935) was written as part of the programme of the Austrian Institute for Trade Cycle Research. In it Tintner applied Anderson's (1927) variate difference method to some 300 series of commodity prices from 1845 to 1914. Under certain assumptions, this method eliminated (most of) the random component from each series, leaving the systematic component for further study; Tintner (1940) presented a more complete statement of the method.

By 1935, Tintner had become enthusiastic about the work of the American mathematicians G.C. Evans (1922; 1924; 1930) and C.F. Roos (1925; 1934), who were applying calculus of variations to theoretical problems in economic dynamics. It appears that Tintner hoped to make a major breakthrough by extending the Evans–Roos approach. From 1936 to 1942 he published a series of brilliant articles (1936;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1937; 1938a; 1938b; 1939; 1941a; 1941b; 1942a; 1942b; 1942c) on such topics as maximization of utility over time, the theoretical derivation of dynamic demand curves, and the pure theory of production under technological risk and uncertainty; his last article of this type was ‘A Note on Welfare Economics’ (1946). Apparently these articles attracted little attention under the disturbed conditions of the time and were not consulted by the young economists who applied similar methods in the 1950s and 1960s to the theory of economic growth and uncertainty. Tintner's work on dynamic economic theory deserves a thorough reappraisal.

The early literature on linear programming dealt exclusively with the deterministic case. Tintner (1955) and Charnes and Cooper (1959) were the first to develop theories and methods for dealing with the various stochastic cases in which inputs, outputs, technical coefficients and/or constraints are subject to random disturbances. Tintner's development of an active approach to stochastic programming (as opposed to a passive approach) pointed the way to current research on self-tuning control combining both estimation and regulation (Sengupta, 1985). Tintner's students and others also made important contributions to stochastic programming; by the late 1970s its literature included several hundred articles and a number of books – see, for example, Kolbin (1977), Tintner and Sengupta (1972), van Moeseke (1965), and Sengupta (1972; 1982).

A selected bibliography of Tintner's publications through 1967 is included in Fox, Sengupta and Narasimham (1969). Tintner spent the bulk of his career at Iowa State University (1937–62) and the University of Southern California (1963–73). From 1973 until shortly before his death (in Vienna, 13 November 1983) he was professor of econometrics at the Technische Universität in Vienna and Honorary Professor at the University of Vienna. His textbooks (Tintner 1952; 1953) had considerable influence on the teaching of econometrics and he also published important articles on multivariate analysis, time series analysis and homogeneous systems in mathematical economics.

Selected works

1935. *Prices in the Trade Cycle*. Vienna: Julius Springer.

1936. A note on distribution of income over time. *Econometrica* 4, 60–6.

1937. Monopoly over time. *Econometrica* 5, 160–70.

1938a. The maximization of utility over time. *Econometrica* 6, 154–8.

1938b. The theoretical derivation of dynamic demand curves. *Econometrica* 6, 375–80.

1939. Elasticities of expenditure in the dynamic theory of demand. *Econometrica* 7, 266–70.

1940. *The Variate Difference Method*. Cowles Commission Monograph No. 5. Bloomington, IN: Principia Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1941a. The theory of choice under subjective risk and uncertainty. *Econometrica* 9, 298–304.

1941b. The pure theory of production under technological risk and uncertainty. *Econometrica* 9, 305–12.

1942a. A contribution to the non-static theory of choice. *Quarterly Journal of Economics* 56(February), 274–306.

1942b. A contribution to the non-static theory of production. In *Studies in Mathematical Economics and Econometrics: In Memory of Henry Schultz*, ed. O. Lange. et al. Chicago: University of Chicago Press.

1942c. The theory of production under nonstatic conditions. *Journal of Political Economy* 50, 645–67.

1946. A note on welfare economics. *Econometrica* 14, 69–78.

1949. Foundations of probability and statistical inference. *Journal of the Royal Statistical Society*, Series A (General) 112, Part III, 251–79.

1952. *Econometrics*. New York: Wiley; London: Chapman & Hall.

1953. *Mathematics and Statistics for Economists*. New York: Rinehart.

1955. Stochastic linear programming with applications to agricultural economics. In *Symposium on Linear Programming*, vol. 1. Washington, DC: National Bureau of Standards.

1960. *Handbuch der Ökonometrie*. Berlin: Springer-Verlag.

1968. *Methodology of Mathematical Economics and Econometrics*. Chicago: University of Chicago Press. Also issued as Vol. 2, No. 6, of the *International Encyclopedia of Unified Science*, Chicago: University of Chicago Press.

1972. (With J.K. Sengupta.) *Stochastic Economics: Stochastic Processes, Control, and Programming*. New York: Academic Press.

Bibliography

Anderson, O. 1927. On the logic of the decomposition of statistical series into separate components. *Journal of the Royal Statistical Society* 90, Part III, 548–69.

Charnes, A. and Cooper, W.W. 1959. Chance-constrained programming. *Management Science* 6(1), 73–9.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Evans, G.C. 1922. A simple theory of competition. *American Mathematical Monthly* 29, 371–80.

Evans, G.C. 1924. The dynamics of monopoly. *American Mathematical Monthly* 31, 77–83.

Evans, G.C. 1930. *Mathematical Introduction to Economics*. New York: McGraw-Hill.

Fox, K.A., Sengupta, J.K. and Narasimham, G.V.L., eds. 1969. *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*. Berlin and New York: Springer-Verlag.

Harper, W.L. and Hooker, C.A., eds. 1976. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, vol. 2. Dordrecht: D. Reidel.

Koch, G. and Spizzichino, F., eds. 1982. *Exchangeability in Probability and Statistics*. Amsterdam: North-Holland.

Kolbin, V.V. 1977. *Stochastic Programming*. Dordrecht: D. Reidel.

Moeseke, P. van. 1965. Stochastic linear programming. *Yale University Economic Essays* 5(1), 197–253.

Roos, C.F. 1925. A mathematical theory of competition. *American Journal of Mathematics* 47, 163–75.

Roos, C.F. 1934. *Dynamic Economics*. Cowles Commission Monograph No. 1. Bloomington, IN: Principia Press.

Sengupta, J.K. 1972. *Stochastic Programming: Methods and Applications*. Amsterdam: North-Holland.

Sengupta, J.K. 1982. *Decision Models in Stochastic Programming*. Amsterdam: North-Holland.

Sengupta, J.K. 1985. *Information and Efficiency in Economic Decision*. Dordrecht: Martinus Nijhoff.

How to cite this article

Fox, Karl A. "Tintner, Gerhard (1907–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000066>
doi:10.1057/9780230226203.1711

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Titmuss, Richard Morris (1907 – 1973)

Philippe Fontaine
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Arrow, K.; social policy; blood procurement; Buchanan, J.; ethics and economics; Friedman, M.; gifts; Hayek, F.; Institute of Economic Affairs; invisible hand; Olson, M.; socialism; Solow, R.; Titmuss, R.

Article

A professor of social administration at the London School of Economics (LSE) from 1950 to his death, Richard Titmuss has often been depicted as an inept critic of economics. With *The Gift Relationship* (1970), however, he managed to attract the attention of leading economists. Robert Solow (1971) and Kenneth Arrow (1972), for instance, took pains to write lengthy review articles of what they and a number of their prominent peers, such as James Buchanan and Milton Friedman, regarded as a highly significant book. In a subject which has a solid tradition of confining ethical matters to its periphery and which has frequently resisted ideas emanating from other social sciences, it is paradoxical that a book of strong ethical inspiration and uncertain disciplinary origins attracted so much interest. One way out this paradox is perhaps to note that, starting with Mancur Olson's *The Logic of Collective Action* (1965), and accompanying the economic difficulties of the late 1960s, economists began to question the power of the invisible hand of the market in bringing about social cohesion. By the early 1970s the time was ripe for reconsidering the virtues of alternative coordinating mechanisms.

The context played a role in the reception of *The Gift Relationship*, but it was above all its subject that made the difference. Long before the book was published in early 1971, Titmuss (1959; 1963; 1968) had been a major figure in the debate over the welfare state, and in this capacity had criticized economists for their incomplete notion of what holds a society together and the unfortunate policy prescriptions they derived from it. With his new essay, however, he presented his own conception of social cohesion and social policy by contrasting the British system of blood procurement and distribution, based on free giving, to the partly commercialized US system. Metaphorically, the gift of blood illustrated the consolidation of the social bond, while its sale stood for social collapse. In other words, Titmuss pointed to the dangers of society's increasing commercialization.

The son of a small farmer and his wife of less humble origins, Titmuss lacked formal schooling beyond

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the age of 14, and when he joined the Eugenics Society in 1937 it was hardly expected he would end up in academe. Yet, after participating in several of the society's research projects, including those of the Population Investigation Committee, then headed by Alexander Carr-Saunders, Director of the LSE, he made a name for himself in academic circles. Subsequently, the historian Keith Hancock approached him to work as historian of the Cabinet office, as a result of which Titmuss produced the monumental *Problems of Social Policy* (1950). Not only did the book secure him the chair in Social Administration at the LSE but it also illustrated the importance of the experience of the Second World War in shaping his 'vision of good conduct as generalised obligation' (see Reisman, 2004). Based on values of solidarity and social duty, this vision may well have been suitable to describe the wartime and immediate post-war British society. From the mid-1960s, however, it became clear that it was incomplete, as was the vision of those economists who regarded the market as the main, if not exclusive, coordinating mechanism in society.

It is these economists' ideas and the applications suggested by their allies at the Institute of Economic Affairs (IEA), the London-based think tank inspired by Hayek, that Titmuss fought from the late 1960s to 1973 (see Fontaine, 2002). He was of the opinion that 'social growth' was more important than its economic counterpart, and that a 'socialist' social policy – not the invisible hand of the market – was essential to social cohesion. These two ideas were intimately connected. Titmuss used the former to show that the economists' social indicators were inadequate: the economy could grow economically and still regress socially because negative externalities supplanted positive ones. Likewise, 'socialist' social policies stimulated ethical behaviour, which generated positive externalities and averted negative externalities, whereas 'private' social policies, as envisaged by the IEA, favoured commercialism, which neglected positive externalities and underestimated negative externalities.

While Titmuss's criticism that economists relied too much on the invisible hand of the market may have actually applied only to a few of them (Solow, 1971), his thesis that too much commercialism undermines the social bond concerned them all. Most economists remained unconvinced, but, faced with Titmuss's emphasis on the role of the gift in so unusual a setting as impersonal interactions, where they typically saw selfishness as reigning supreme, the economists were forced to contemplate the possibility that 'a world of giving may actually increase efficiency in the operation of the economic system' (Arrow, 1972, p. 351).

See Also

- altruism, history of the concept

Selected works

1950. *Problems of Social Policy*. London: HMSO.

1959. *Essays on The Welfare State*. New Haven, CT: Yale University Press.

1963. Ethics and economics of medical care. *Medical Care* 1, 16–22.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1968. *Commitment to Welfare*. London: Allen and Unwin.

1970. *The Gift Relationship: From Human Blood to Social Policy*. London: Allen and Unwin.

Bibliography

Arrow, K. 1972. Gifts and exchanges. *Philosophy and Public Affairs* 1, 343–62.

Fontaine, P. 2002. Blood, politics and social science: Richard Titmuss and the Institute of Economic Affairs, 1957–1973. *Isis* 93, 401–34.

Olson, M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Reisman, D. 2004. Richard Titmuss: welfare as good conduct. *European Journal of Political Economy* 20, 771–94.

Solow, R. 1971. Blood and thunder. *Yale Law Journal* 80, 1696–711.

How to cite this article

Fontaine, Philippe. "Titmuss, Richard Morris (1907–1973)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000172> doi:10.1057/9780230226203.1712

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tobin, James (1918 – 2002)

Donald D. Hester
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

James Tobin was a brilliant economist and the leading proponent of Keynesian economics in the second half of the 20th century. He greatly advanced understanding of financial institutions and monetary theory and policy. He stressed the importance of asset holdings and wealth on consumer spending. He also argued that ‘q’, the ratio of a firm's market value to the replacement value of its assets, was an important determinant of its investment decisions. He made major contributions to econometric methods, international economics, the theory of growth and business cycles, and policies designed to improve the welfare of minorities and the poor.

Keywords

absolute income hypothesis; Allais, M.; American Economic Association; animal spirits; banking industry; Baumol, W.; bonds; economic growth; financial intermediation; fiscal policy; Fleming, J.; floating exchange rate; Friedman, M.; government debt; household portfolios; income elasticity of demand; income velocity of money; interest rate elasticity; investment decisions; IS–LM model; Keynes, J. M.; Keynesianism; leisure; liquidity preference; Lucas, R.; maximum likelihood; measure of economic welfare (MEW); monetarism; monetary policy; monetary theory; monetary transmission mechanism; multicollinearity; Mundell, R.; national income accounts; negative income tax; non-accelerating inflationary rate of unemployment (NAIRU); permanent-income hypothesis; Phillips curve; pollution; portfolio balance; portfolio demand for money; rational expectations hypothesis; rationing; relative income hypothesis; Ricardian equivalence theorem; Solow, R.; stocks and flows; Swan, T.; technology; traditional vs modern; Tobin tax; Tobin, J.; Tobin's q; tobit model; transactions demand for money; wage rigidity

Article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

James Tobin was a major contributor to economic science and macroeconomic analysis who received the Nobel Memorial Prize in Economic Science in 1981 for ‘his analysis of financial markets and their relations to expenditure decisions, employment, production, and prices’. He received a BA in 1939, an MA in 1940, and a Ph.D. in 1947 from Harvard University. (Between 1941 and 1945 Tobin was an officer on a destroyer in the US Navy.) He made many early penetrating theoretical and empirical explorations of the underpinnings of Keynes's *General Theory of Employment, Interest, and Money* (1936). Throughout his long career he was a vigorous participant in the design of fiscal and monetary policies and in actively promoting their implementation. He was on the faculty of Yale University from 1950, and from 1957 until his death he held a prestigious Sterling professorship there.

He published a large number of influential papers in professional journals and other places, and his impact on economics can only be hinted at in this unavoidably incomplete article. Buiter (2003) has published a much longer but still incomplete survey of Tobin's work, which consists of about 500 articles over a period of 60 years. Many of his papers are accessible in a series of volumes of his collected works (1971; 1975; 1982; 1996a) in addition to their original places of publication.

Apart from his dissertation, his earliest empirical evaluations of the Keynesian approach (1947; 1951), were influential informal studies that respectively examined the interest elasticity of the demand for money and the underpinnings of the consumption function. The latter showed that evidence for Duesenberry's (1949) ‘relative income hypothesis’ was not persuasive when compared with a wealth-augmented variation of Keynes's absolute income hypothesis, when individuals were broken down into savers and dissavers and variations in the cost of living were taken into account. The importance of wealth and portfolio composition on behaviour would be a major theme in Tobin's work in subsequent decades.

Perhaps his most important applied econometrics article (1950) examined the demand for food in the United States. It refined techniques that had been used in his dissertation, which analysed the consumption function. By combining estimates of income elasticities obtained from cross-sectional budget data with price elasticities estimated from time series, Tobin attempted to avoid the problem of multicollinearity that often arises when one uses time series data exclusively. He focused on food rather than aggregate consumption in this paper in order to avoid problems associated with purchases of durable goods. This paper has stood the test of time remarkably well and was celebrated in a special issue of the *Journal of Applied Econometrics* (1997).

Shortly after finishing this paper, Tobin (1952a) and Tobin and H. S. Houthakker, (1951; 1952) completed an elegant series of papers on the theory of rationing and its implications for econometric analysis.

Household balance sheets and spending

Upon arriving at Yale, Tobin developed a distinctive empirical approach to household spending and portfolio behaviour that was foreshadowed by an illuminating survey article, ‘Asset Holdings and Spending Decisions’ (1952b). He critically reviewed attempts to include money, liquid assets, and government debt holdings by households as predictors of consumption in both macroeconomic and microeconomic studies. He argued that such asset holdings are best viewed as part of a dynamic optimization process; people accumulate liquid assets in preparation for executing spending

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

programmes. Liquid assets cannot sensibly be interpreted as exogenous determinants of consumption. An early empirical study (1957) analysed the relation between consumer holdings of debt, liquid assets and durable goods and decisions to acquire more of each of them. He reported that high levels of debt discouraged acquisition of more debt and durable goods, and in an illuminating diagram suggested that individuals with different attributes tend to have different desired mixtures of debt and liquid assets, which can be interpreted as target levels of ‘portfolio balance’. A second empirical study, coauthored with H. W. Watts (1960) extended this portfolio balance approach to broad collections of assets and liabilities.

Tobin made a very important econometric methods contribution that was a consequence of his work with household assets. He noted that many household assets took on zero values until a certain threshold was crossed. Linear regressions attempting to represent relations, say, between a household's income and the value of its car would be severely misspecified if some observations in a sample had no car. In (1958a) he developed an ingenious probit-regression technique (subsequently named ‘tobit’ by A. S. Goldberger), which allowed maximum likelihood techniques to be used to obtain consistent estimates of such relationships. The technique continues to be widely used.

A dynamic aggregative model and the effects of money on economic growth

Portfolio balance assumed a central role in a very innovative model (1955) where Tobin managed to combine endogenous economic fluctuations and economic growth. This paper, which Tobin viewed as his favourite (see Breit and Spencer, 1986, p. 128), is Keynesian in that money wages may be inflexible and may lead to unemployed labour. An important innovation in the model was a linear homogeneous production function that allowed for substitution between capital and labour, as in the nearly contemporaneous economic growth models of Solow (1956) and Swan (1956). However, its major new contribution was the formal introduction of portfolio balance and money. Changes in money are equal to the government deficit in this model. Wealth consists of physical capital and the real money supply. Momentary price equilibrium exists when the price level allows wealth holders to be satisfied with the mix of capital and real money holdings in their portfolios. Because investors are assumed to be risk averse, portfolio equilibrium does not require that the negative expected rate of change of price equals the expected marginal productivity of capital. Depending upon price expectations, monetary expansion may affect the rate of capital formation positively or negatively. The paper briefly explores how technical progress relates to price changes and growth.

Introducing inflexible money wage rates leads to a variety of results that include inflation, deflation, secular economic growth or stagnation, and economic cycles, which can be seen when two summary relations describing labour market balance and portfolio balance are examined. Recovery from cyclical slumps may occur if extreme conditions cause money wage rates to become flexible or if capital depreciates sufficiently.

In this model and in its specializations to economic growth (1965b; 1968), money and government debt are indistinguishable. In an important unpublished manuscript dating from 1958, which eventually was published in 1998, Tobin addressed this distinction and presented an elegant analysis of how monetary policy worked through banks and how changes in the composition of government debt affected capital and growth. In Tobin (1961) and in his contribution to the Commission on Money and Credit (1963a), Tobin drew on this manuscript to analyse debt and monetary policies and their relation to the return on

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
capital. The reasons for the long delay in publication of the manuscript are unclear, but surely part of the explanation was that Tobin was a member of the Kennedy administration's Council of Economic Advisors for 18 months beginning in January 1961.

The demand for money

Tobin published two papers that were narrowly focused on the demand for money (1956; 1958b). The first was an inventory theoretic model of the transactions demand for money that had close antecedents in Allais (1947) and Baumol (1952). These three papers cast doubt on the constancy of the income velocity of money, Y/M, by arguing that an individual's income elasticity of demand for money is likely to be about 0.5 and the interest rate elasticity about – 0.5. As the theory predicted, the post-war income velocity of money in the United States rose as per capita income and interest rates rose.
The second paper was a path-breaking effort that proposed a novel explanation for the portfolio demand for money. It was written at about the time Markowitz was writing his monograph (1959) on portfolio selection at Yale, but had a very different orientation. For simplicity Tobin assumed cash had a riskless rate of return. He introduced uncertainty about the rate of return on a second asset, and explored how an investor should split his portfolio between cash and the risky asset over a single planning period. His goal was to eliminate an unsatisfactory assumption about interest rate expectations of investors that underlay Keynes's exposition of liquidity preference. He assumed either that investors' utility functions were quadratic or that the distribution of the risky rate of return was normal, and derived the optimal mixture of the risky asset and cash. Although his discussion was not error-free, as was pointed out by Borch (1969) and Feldstein (1969), it provided a foundation from which a large literature in finance developed. Initially Tobin assumed the second asset was a consol, in the spirit of Keynes's discussion. Then he showed that the second asset could be a linear combination of a set of risky assets and that investors' portfolio problem could often be reduced to a mixture of that combination and cash. He was the first to recognize 'two fund' separation, which would play a major role in the theory of finance.

Financial intermediation and policy

Tobin's unpublished manuscript included a discussion of commercial banks and their role in the transmission of monetary policy, topics that he extensively developed in the 1960s. An early unpublished paper (1959) examined how financial intermediaries in general responded to monetary controls; it would be extensively revised and appear with the same title in a rigorous discussion of monetary policy, coauthored with William Brainard (1963). Tobin and Brainard argued that a monetary policy action's effect can be measured with the required rate of return on real capital; restrictive (easing) monetary policy will raise (lower) this rate. They assumed that all assets and liabilities are gross substitutes. Thus, for assets an increase in the rate of return on one asset increases the demand for it and does not increase the demand for any other asset, and analogously for liabilities. They analysed how a monetary policy-induced change in currency affects the required rate of return in regimes with controlled and uncontrolled intermediaries, where controls consist of reserve requirements and interest rate ceilings on deposit liabilities. The controls may strengthen the efficacy of monetary policy, but it works when they are absent. Tobin (1963b) provided an accessible intuitive discussion of the responses

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of banks and other intermediaries to monetary policy actions.

Tobin (1969) presented a general equilibrium interpretation of the overall approach and emphasized the importance of ‘q’, the ratio of the market value of a firm to the replacement value of its assets, a central focus of his approach to monetary policy and a prominent variable in subsequent empirical studies of the investment decision. Intuitively, when q is greater (less) than unity, stockholders gain (lose) when a firm undertakes net new investment. *Ceteris paribus*, a high real rate of interest (and thus restrictive monetary policy) is likely to reduce market values and thus discourage new investment. Tobin and Brainard (1977) reported a very ambitious attempt to measure average q using panel data and introduced the important distinction between marginal and average q, a topic that was subsequently thoroughly developed by Hayashi (1982). While acknowledging that q was an endogenous variable partly determined by animal spirits and investor expectations, they argued that an estimate of q for firms should be included among the variables guiding monetary policy.

Tobin and Brainard (1968) reported simulation experiments using a hypothetical model that illustrated the central role of q and the importance of taking into account cross equation restrictions. Drawing on well-known arguments from the theory of consumer demand, they suggested that empirical models of financial markets often were misspecified because investigators ignored adding-up constraints on parameters across equations on interest rate, income and lagged variables. They argued that such requirements were especially important in dynamic specifications, where often the only lag in an asset demand equation was the asset's value in a previous period; in principle, all lagged asset variables in a well-specified portfolio adjustment model must appear in each asset demand equation. Implicitly, these two Tobin and Brainard articles indicate the enormous sensitivity of models to arbitrary accounting conventions and the unavoidable errors in variables that such conventions lead to.

Monetary theory and policy

Beginning with his review (1965a) of Friedman and Schwartz's *Monetary History of the United States*, Tobin increasingly assumed the role of defending the Keynesian approach to macroeconomics against an emerging monetarist formulation that was led by Milton Friedman. Tobin had been a major contributor to the highly successful policies of the Kennedy and Johnson administration policies of the early 1960s, but was very critical of the latter's financing of the Vietnamese war (1981, pp. 357, 360). As the decade wore on, the Keynesian approach came increasingly under attack. For example, Milton Friedman's presidential address (1968) to the American Economic Association, which had been preceded by Phelps (1967) with the same idea, questioned the existence of a stationary Phillips curve. Friedman argued that workers would increasingly seek redress from inflation if unemployment were below the non-accelerating inflationary rate of unemployment (NAIRU, where the price-inflation Phillips curve crosses the abscissa). If a non-stationary Phillips curve led to accelerating rates of inflation, prices would no longer be sticky as the Keynesian approach assumed. Friedman's alternative approach argued that money demand was a function of permanent income and that empirical support for the monetarist model came from the fact that in time series money led income, which suggested that money was a causal element. Tobin and Swan (1969) reported empirical relations between permanent income and money that did not support the monetarist model. Tobin (1970) reported simple theoretical models that illustrated that one could not infer from lead–lag relationships in time series of money and income whether a model was

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
fundamentally Keynesian or monetarist. Space limitations prevent a full and fair exposition of the debate, which continued with varying levels of intensity for many years; see, for example, Tobin (1993b; 1995).

The controversies led to ambitious efforts to see whether fiscal and monetary policy effects in the static IS–LM model were likely to carry over in a dynamic long-run framework. Tobin and Buiter (1976) and Tobin (1979) examined this question formally using elegant techniques and were not able to conclude in general that the effects of policy were either transitory or long-lasting in models with a stationary state. They were able to describe some model specifications with predictable effects in the long run and discussed their intuitive plausibility. Tobin and Buiter (1980) examined the same issue in the context of a growth model, with similarly inconclusive results. All the same, these papers are valuable examples of how to analyse such questions.

Tobin provided a relatively complete statement of his views on the theory of macroeconomic policy in his three Yrjö Jahnsson lectures and a related Paish Lecture given in England that appear in Tobin (1980). His first Jahnsson lecture considered how price level changes affected economic activity, focusing on the Pigou effect, the Keynes effect, and Fisher's debt-deflation hypothesis. He believed that the Fisher hypothesis won over the Pigou effect in the short and medium run. He denied that Keynes had demonstrated an underemployment equilibrium, but accepted the IS–LM model as a reasonable guide to policy in the short run.

His second Jahnsson lecture attacked both Friedman's views about a shifting Phillips curve that are discussed above and Lucas's (1972) rational expectations hypothesis. He argued that the latter was more radical than Friedman's discussion because it denied even the short-run efficacy of monetary policy. Tobin explained that, while everyone agrees that expectations are important, expectations are highly diverse in an economy and based on different information sets; therefore, they are not likely to be unbiased. Further, he denied the Lucas contention that labour and product markets are always being cleared at existing wages and prices, and criticized as ad hoc Lucas's specification ‘about the information available to buyers and sellers’ (Tobin, 1980, p. 42).

His third Jahnsson lecture addressed an obvious deficiency of the short-run IS–LM model, namely, that the capital stock and other measures of wealth are assumed to be constant even though investment is occurring and debt is being issued. He sketched out a model that integrates flow-of-funds accounts with variables in the IS–LM model, which derived from his papers with Buiter. An especially ambitious attempt at constructing such a model is reported in Backus, Brainard, Smith and Tobin (1980).

The Paish lecture examined the plausibility of Ricardian equivalence – the idea that there is no difference resulting from decisions to finance an increase in government spending by printing money and/or raising taxes, and/or issuing government debt. Robert Barro (1974) had resurrected and developed this thesis, which denies that government deficits have an effect because households take into account the fact that increased issues of debt would require an increase in future taxes, which would be needed to service and retire the debt. While the Barro argument might be true in a simple hypothetical world where household dynasties are infinitely lived and each household in a generation behaves so as to fully take into account effects on its counterparts in other generations, Tobin argued that demographic events like no heirs, generational myopia, and financial conditions such as liquidity constraints, possibly resulting from imperfect capital markets, and wealth and other distorting taxes invalidated it.

Economic development and the measurement of economic activity

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As a visiting professor at the University of Nairobi in the 1972–3 academic year, Tobin wrote two instructive papers that use the linear activity analysis model. The first (1974a) examined a model with two technologies, traditional and modern (more productive), where unemployment was possible in an economy with a constant saving rate and where capital in each of the technologies had possibly different depreciation rates. Depending upon parameter values, he showed that it could be optimal to invest capital in the traditional rather than the modern technology, in both the short and the long runs. In (1974b) he studied the effects of expulsion of alien residents and expropriation of their assets. Depending upon the compensation, if any, for expropriating the assets of the former residents, the linear model yielded predictions about how remaining citizens would fare. Tobin did not analyse the ethical justification, if any, for expulsion and expropriation, but nevertheless provided a very insightful discussion about who might gain and lose in situations where expropriation occurred.

William Nordhaus and Tobin in a very ambitious paper entitled ‘Is Growth Obsolete?’ (1972) proposed a primitive measure of economic welfare (MEW) for the United States. They argued that the national income accounts measure production and are not appropriate for analysing welfare, for several reasons. The accounts do not measure the flow of services from the stock of durable goods or the value of human capital acquired through education and training, nor do they account for the depletion of minerals, environmental degradation, and the disamenities of urbanization. The national income accounts include all consumer and government expenditures. Nordhaus and Tobin viewed many of these as ‘instrumental’ expenditures that are regrettably necessary for welfare. Examples include commuting costs, police services, national defence and sanitation. They subtracted such expenditures from their measure, which is expressed on a per capita basis, and made an imputation for the flow of leisure. The authors argued that technical progress and capital formation have more than compensated for the depletion of natural resources so far and are guardedly optimistic that this will continue. However, they stated that the effects of pollutants on melting polar ice caps warrant a higher priority in research and that pollution is a problem because it is not priced to reflect social costs. While the possibility of catastrophic global disturbances cannot be excluded, their conclusion was that growth is not obsolete.

International economics

Tobin and Braga de Macedo (1980) modified the standard paradigm of Mundell (1961) and Fleming (1962) of how monetary policy works in a floating exchange rate system by introducing the exchange rate in asset demand functions and assuming that all assets are gross substitutes. There are three assets – cash, foreign holdings and bonds – because bonds and capital are conventionally assumed to be perfect substitutes. In the case of a small open economy they showed that with these modifications fiscal policy affects the economy's level of real income, although the effect cannot be signed without information about some parameters in this variation of a standard IS–LM model. This contrasts with the Mundell and Fleming models where fiscal policy is impotent in a pure floating exchange rate system.

The authors then expanded the small open-economy model to have four assets, by assuming that bonds and capital are not perfect substitutes. They continued to assume that all assets are gross substitutes and employed a discrete time specification to analyse some of the consequences of asset accumulation. This second model describes a single period, but beginning-of-period stocks and within-period flows yield complex effects that unfortunately cannot be summarized adequately in the present article. However, it

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

is usually the case that fiscal policy again has non-zero effects in a flexible exchange regime. They also examined fiscal policy in a two-country, flexible-exchange world, where neither is small and in each country bonds and capital are perfect substitutes. The rich analytical framework of this paper is applied insightfully in Tobin (1993a) and partly underlies a simulation exercise by Tobin and Brainard (1992). Tobin (1978) argued that countries should impose a tax on purchases of financial instruments denominated in another country's currency in order to allow some autonomy in setting domestic stabilization policies. This controversial ‘Tobin tax’ proposal to throw sand into the gears of international finance first appeared in Tobin (1974c, pp. 88–92), achieved wide approval outside the United States, and was analysed in ul Haq, Kaul and Grunberg (1996).

Overview and summary

As stated at the outset, one cannot do justice to Tobin's extraordinary career in a short article. However, before closing, it is important to acknowledge his extensive contributions to public debate on policy that appear in Tobin (1966; 1996b), and especially on welfare and inequality that are reprinted in part in Tobin (1982, pp. 497–624). He was a forceful advocate of the negative income tax, reducing inequality, and improving the economic condition of minorities in the United States, especially during the turbulent 1960s and 1970s.

See Also

- IS–LM
- Keynesian revolution
- monetary and fiscal policy overview
- monetary transmission mechanism
- natural rate of unemployment
- rationing
- Ricardian equivalence theorem
- Tobin's q
- Tobit model

Selected works

1947. Liquidity preference and monetary policy. *Review of Economics and Statistics* 29(2), 124–31.

1950. A statistical demand function for food in the U.S.A. *Journal of the Royal Statistical Society* 113, Series A, Part II, 113–41.

1951. Relative income, absolute income, and saving. In *Money, Trade, and Economic Growth (Essays in Honor of John Henry Williams)*. New York: Macmillan.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1951. (With H. Houthakker.) The effects of rationing on demand elasticities. *Review of Economic Studies* 18(3), 1–14.

1952a. A survey of the theory of rationing. *Econometrica* 20, 521–53.

1952b. Asset holdings and spending decisions. *American Economic Review* 42(2), 109–23.

1952. (With H. Houthakker.) Estimates of the free demand for rationed foodstuffs. *Economic Journal* 62, 103–18.

1955. A dynamic aggregative model. *Journal of Political Economy* 63(2), 103–15.

1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38(3), 241–47.

1957. Consumer debt and spending: some evidence from analysis of a survey. *Consumer Installment Credit*. Part II, vol. 1. Washington, DC: US Government Printing Office.

1958a. Estimation of relationships for limited dependent variables. *Econometrica* 26(1), 24–36.

1958b. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25, 65–86.

1959. Financial intermediaries and the effectiveness of monetary controls. Discussion Paper No. 63 (January), Cowles Foundation, Yale University.

1960. (With H. Watts.) Consumer expenditures and the capital account. In *Proceedings of the Conference on Consumption and Saving*, ed. I. Friend and R. Jones. Philadelphia: University of Pennsylvania Press.

1961. Money, capital, and other stores of value. *American Economic Review* 51(2), 26–37.

1963a. An essay on the principles of debt management. In *Fiscal and Debt Management Policies*. Englewood Cliffs, NJ: Prentice Hall.

1963b. Commercial banks as creators of ‘money’. In *Banking and Monetary Studies*, ed. D. Carson. Homewood, IL: Richard D. Irwin.

1963. (With W. Brainard.) Financial intermediaries and the effectiveness of monetary controls. *American Economic Review* 53, 383–400.

1965a. The monetary interpretation of history. *American Economic Review* 55, 464–85.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1965b. Money and economic growth. *Econometrica* 35, 671–84.

1966. *National Economic Policy*. New Haven, CT: Yale University Press.

1968. Notes on optimal monetary growth. *Journal of Political Economy* 76, 833–59.

1968. (With W. Brainard.) Pitfalls in financial model building. *American Economic Review* 58(2), 99–122.

1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1(1), 15–29.

1969. (With C. Swan.) Money and permanent income: some empirical tests. *American Economic Review* 59, 285–95.

1970. Money and income: post hoc ergo propter hoc? *Quarterly Journal of Economics* 84, 301–17.

1971. *Essays in Economics: Macroeconomics*. Chicago: Markham Publishing.

1972. (With W. Nordhaus.) Is growth obsolete? In *Economic Growth: Fiftieth Anniversary Colloquium V*, National Bureau of Economic Research. New York: Columbia University Press.

1974a. Technological development and employment. *Eastern Africa Economic Review* 6(1), 1–26.

1974b. Notes on the economic theory of expulsion and expropriation. *Journal of Development Economics* 1(1), 7–18.

1974c. *The New Economics, One Decade Older*. Princeton, NJ: Princeton University Press.

1975. *Essays in Economics: Consumption and Econometrics*. Amsterdam: North-Holland.

1976. (With W. Buiter.) Long-run effects of fiscal and monetary policy on aggregate demand. In *Monetarism: Studies in Monetary Economics*, vol. 1, ed. J. Stein. Amsterdam: North-Holland.

1977. (With W. Brainard.) Asset markets and the cost of capital. In *Economic Progress, Private Values and Public Policy*, ed. R. Nelson and B. Balassa. Amsterdam: North-Holland.

1978. A proposal for international monetary reform. *Eastern Economic Review* 4(3–4), 153–9.

1979. Deficit spending and crowding out in shorter and longer runs. In *Theory for Economic Efficiency:*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Essays in Honor of Abba P. Lerner, ed. H. Greenfield et al. Cambridge MA: MIT Press.

1980. Asset Accumulation and Economic Activity: *Reflections on Contemporary Macroeconomic Theory*. Yrjö Jahnsson Lectures. Chicago: University of Chicago Press.

1980. (With D. Backus, W. Brainard, and G. Smith.) A model of US financial and nonfinancial behavior. *Journal of Money, Credit, and Banking* 12, 259–93.

1980. (With W. Buiter.) Fiscal and monetary policies, capital formation, and economic activity. In *The Government and Capital Formation*, ed. G. von Furstenberg. Cambridge, MA: Ballinger.

1980. (With J. Braga de Macedo.) The short-run macroeconomics of floating exchange rates: an exposition. In *Flexible Exchange Rates and the Balance of Payments: Essays in Memory of Egon Sohmen*, ed. J. Chipman and C. Kindleberger. Amsterdam: North-Holland.

1981. Reflections inspired by proposed constitutional restrictions on fiscal policy. In *Economic Regulation: Essays in Honor of James R. Nelson*, ed. K. Boyer and W. Shepherd. East Lansing, MI: Michigan State University.

1982. *Essays in Economics: Theory and Policy*. Cambridge, MA: MIT Press.

1992. (With W. Brainard.) On the internationalization of portfolios. *Oxford Economic Papers* 44, 553–65.

1993a. Policies and exchange rates: a simple analytical framework. In *Japan, Europe, and International Financial Markets: Analytical and Empirical Perspectives*, ed. R. Sato, R. Levich and R. Ramachandran. Cambridge: Cambridge University Press.

1993b. Price flexibility and output stability: an old Keynesian view. *Journal of Economic Perspectives* 7 (1), 45–65.

1995. The natural rate as new classical macroeconomics. In *The Natural Rate of Unemployment: Reflections on 25 Years of the hypothesis*, ed. R. Cross. Cambridge: Cambridge University Press.

1996a. *Essays in Economics: National and International*. Cambridge, MA: MIT Press.

1996b. *Full Employment and Growth: Further Keynesian Essays on Policy*. Cheltenham: Edward Elgar.

1998. (With S. Golub.) *Money, Credit, and Capital*. Boston: Irwin/McGraw-Hill.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Allais, M. 1947. *Economie et intérêt*. Paris: Impremerie Nationale.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6), 1095–117.

Baumol, W. 1952. The transactions demand for cash: an inventory theoretic approach. *Quarterly Journal of Economics* 66, 545–56.

Borch, K. 1969. A note on uncertainty and indifference curves. *Review of Economic Studies* 36, 1–4.

Breit, W. and Spencer, R., eds. 1986. Tobin. In *Lives of the Laureates*. Cambridge, MA: MIT Press.

Buiter, W. 2003. James Tobin: an appreciation of his contribution to economics. *Economic Journal* 113, 585–631.

Duesenberry, J. 1949. *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.

Feldstein, M. 1969. Mean-variance analysis in the theory of liquidity preference and portfolio selection. *Review of Economic Studies* 36, 5–12.

Fleming, J. 1962. Domestic financial policies under fixed and under floating exchange rates. *International Monetary Fund Staff Papers* 9, 369–79.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1), 1–17.

Hayashi, F. 1982. Tobin's marginal q and average q: a neoclassical interpretation. *Econometrica* 50, 213–24.

Lucas, R. 1972. Econometric testing of the natural rate hypothesis. In *The Economics of Price Determination Conference*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.

Keynes, J. M. 1936. *General Theory of Employment, Interest, and Money*. New York: Macmillan.

Magnus, J. and Morgan, M., eds. 1997. The experiment in applied econometrics. *Journal of Applied Econometrics* 12(5), 459–662.

Markowitz, H. 1959. Portfolio Selection: Efficient Diversification of Investments. Cowles Foundation Monograph 16. New York: John Wiley and Sons.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mundell, R. 1961. Flexible exchange rates and employment policy. *Canadian Journal of Economics and Political Science* 27, 509–17.

Phelps, E. 1967. Phillips curves, expectations of inflation, and optimal unemployment over time. *Economica* 34, 254–81.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70, 65–94.

Swan, T. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

ul Haq, M., Kaul, I. and Grunberg, I., eds. 1996. *The Tobin Tax: Coping with Financial Volatility*. New York: Oxford University Press.

How to cite this article

Hester, Donald D. "Tobin, James (1918–2002)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000069> doi:10.1057/9780230226203.1713

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tobin's q

Gary Smith
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tobin's q is the ratio of the market value of a firm to the replacement cost of its assets, a statistic that depends on the firm's profitability and financial markets' required rate of return. Although there are a variety of measurement issues, including the distinction between marginal and average q, Tobin's q can be used to predict investment spending or to control for a firm's current and future profitability in empirical studies of corporate structure and behaviour.

Keywords

accounting conventions; adjustment costs; corporate finance; depreciation; heterogeneous capital; inflation; investment decisions; Keynes, J. M.; monopoly rents; rate of return; takeovers; technological progress; Tobin, J.; Tobin's q

Article

A standard tenet of corporate finance is that the retention of earnings to finance expansion raises a stock's price if the rate of return on these investments, ρ, is larger than shareholders' required return on their stock, R. For example, an investment that costs $1 million is worth more than $1 million to stockholders if $\rho > R$. This principle suggests that whether a firm's stock sells for a premium or a discount relative to the cost of its assets depends on ρ versus R and, further, that a firm's investment decisions ought to depend on a comparison of ρ with R.
The shareholders' required return is determined in financial markets by shareholders pricing stock to yield an anticipated return that is competitive with comparable investments they might make. This determination of required returns is one of the primary ways in which financial markets affect real economic activity. When interest rates fall, required stock returns decline, making it more likely that $\rho > R$, so that the profits from prospective investments are sufficient to make these investments attractive for firms that care about their shareholders.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

But where do firms (or economic forecasters) see these shareholder required returns? It is straightforward to calculate the yield to maturity on a bond by determining the discount rate that equates the present value of the promised cash flow to the market price. There are no comparable calculations for stocks because the cash flow to investors is unknown. An ingenious alternative, proposed by James Tobin (Brainard and Tobin, 1968; Tobin, 1969), is to look at stock prices. Specifically, Tobin argues that we should look at how financial markets value a firm relative to the replacement cost of the firm's assets:

$$q = \frac{\text{market value}}{\text{replacement cost}}$$

The numerator and denominator of Tobin's q can be aggregate market value and replacement cost or, equivalently, price per share and assets per share. If a firm has debts, these can be included in the numerator.

What determines q?

How can assets be valued in financial markets at other than their replacement cost? Assets are of value to shareholders only to the extent that they generate profits. It matters not at all that a factory cost $1 billion to build if it doesn't make a cent of profits. For a factory to be worth what it costs shareholders, it must earn the shareholder's required rate of return.

For a simple example, suppose that a hamburger chain has no debt and pays out all earnings as dividends. Assume also that a new restaurant costs $1 million to build and is expected to earn a constant 20 per cent profit ($\rho = 0.20$), $200,000 a year for ever. The value that financial markets place on the $200,000 annual cash flow depends on how highly hamburger earnings are valued. If Treasury bonds yield five per cent, perhaps stock in risky hamburger restaurants is priced to yield ten per cent ($R = 0.10$). Because the anticipated dividends are a constant $200,000, the market value of the restaurant is

$$V = \frac{\$200,000}{0.10} = \$2,000,000$$

Valued at $2 million, the $200,000 annual dividend provides stockholders their requisite ten per cent return.

In this case, the market value of the restaurant is twice its construction cost: q=2. Stockholders welcome this investment, because the use of $1 million in potential dividends to build the restaurant provides $2 million in market value. The underlying reason is that the restaurant's 20 per cent profit rate is larger than the market's ten per cent required rate of return.

If, on the other hand, shareholders’ required rate of return is 25 per cent, then

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$V = \frac{\$200,000}{0.25} = \$800,000$$

Now $q = 0.8$ and the construction of the restaurant will be to the detriment of shareholders. Because the restaurant earns only 20 per cent on its cost, the stock must be valued at less than the asset's cost in order to provide shareholders their 25 per cent required return.

Implications for investment decisions

A connection between business investment spending and market value relative to cost was pointed out many years ago by John Maynard Keynes (1936, p. 151):

> [T]he daily revaluations of the Stock Exchange, though they are primarily made to facilitate transfers of old investments between one individual and another, inevitably exert a decisive influence on the rate of current investment. For there is no sense in building up a new enterprise at a cost greater than that at which a similar existing enterprise can be purchased; whilst there is an inducement to spend on a new project what may seem an extravagant sum, if it can be floated on the stock exchange at an immediate profit.

Similarly, Tobin argues that a firm should invest in new buildings and equipment if the stock market will value the project at more than its cost (that is, if the project's q is greater than 1). If the market value is larger than the cost, shareholders prefer that the firm make this investment rather than distribute its cost as dividends, gladly giving up a dollar of dividends in exchange for a two-dollar increase in the value of their stock. Put more plainly, the appropriate question a firm should ask is whether, if it were to sell shares in its new venture, it could raise enough money to cover the project's cost. It can if the value of Tobin's q is larger than 1, but not otherwise. Thus Tobin's q provides a barometer of the incentives for business investment.
Similarly, the firm should compare the price it can get for selling its existing assets with the value that financial markets place on these assets. If the market value is less than the sale price ($q<1$), the firm is worth more dead than alive, and it should sell off its assets and distribute the proceeds either through dividends or share repurchases. A low market value relative to replacement cost may also motivate takeover bids, since an outside group may profit by purchasing enough stock to gain control of a company and then liquidating its assets.

Marginal and average q

Why don't firms immediately exploit any differences between market value and replacement costs, thereby causing q to return to 1 instantaneously? Three, sometimes related, explanations involve convex

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
adjustment costs, monopoly rents and heterogeneous capital (Lucas and Prescott, 1971; Mussa, 1977; Smith, 1981; Hayashi, 1982; Abel, 1983; Erickson and Whited, 2000; Gomes, 2001). A consideration of these issues requires a distinction between average q, the aggregate market value and replacement cost of a firm's assets, and marginal q, the change in a firm's market value resulting from a specific investment relative to the cost of that investment.

An observed average q that is greater than 1 might accurately reflect a company's substantial profits, but further investment may be restrained by convex adjustment costs that cause the marginal q for a large expansion to be less than 1. It might be prohibitively expensive for a hamburger chain to triple its size in a year. A related strand of literature (summarized by Hubbard, 1998) investigates how investment by financially constrained firms may be less sensitive to q and more sensitive to the firm's cash flow.

Similarly, a firm might earn monopoly rents that cause average q to exceed 1, but have a marginal q that is less than 1 because new investments would erode these rents. A hamburger restaurant with a patented secret formula might not want to open a competing restaurant next door that would lure away customers. The grower of an exotic fruit may not want to flood the market with produce that could only be sold by lowering prices.

Finally, prospective ventures might be quite different from the firm's existing operations, with different rates of return and with risks that command different required returns. A tobacco company acquires a snack food company; a yogurt maker enters the bottled water market; a software company enters the video game market. In each case, marginal q might be quite different from average q.

A further complication is that observed market values presumably take into account not only existing assets but also future investments anticipated by the market. Suppose, for example, that a firm currently has assets with a replacement cost and market value that both equal $100 million and is planning to make an investment that will cost $20 million and have a market value of $30 million. Average q for the firm's current assets is 1 and marginal q for this investment is 1.5. If the stock market takes into account this projected investment, the current market value is increased by $10 million (discounted somewhat to the extent the value added will occur in the future and if there is uncertainty about whether the investment will be made). Thus observed average q reflects both the profitability of the current capital stock and the perceived profitability of the firm's future opportunities, overstating the former and understating the latter.

Estimates of q

The book value reported by firms is a rough proxy, a starting point, for estimating the replacement cost of a firm's assets. Accounting conventions value assets at historical cost, with no adjustments for subsequent cost increases, and depreciate assets according to accounting conventions rather than true economic depreciation. Inflation can cause book values to understate the replacement cost of assets (think of real estate); technological progress can cause the reverse (think of computers). The market value of a firm's equity is commonly estimated by multiplying the market price of a firm's stock by the number of shares outstanding; data on the market value of a firm's preferred stock and debt are not so easily obtained because databases generally record the book values reported by firms in their balance sheets.

A variety of often complex procedures have been employed to estimate the market value of a firm's

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
debts and the replacement cost of its assets (Brainard, Shoven and Weiss, 1980; Lindenberg and Ross, 1981; Lewellen and Badrinath, 1997; Lee and Tompkins, 1999), while other authors argue that more readily available book values provide sufficiently accurate approximations (Chung and Pruitt, 1994; Perfect and Wiles, 1994).

Because of these measurement errors, when q is used as an explanatory variable in a regression model, least squares estimates of the coefficient of q will be biased towards zero and estimates of the coefficients of other explanatory variables may be biased towards zero or away from zero. For example, in a model that uses a firm's q and current cash flow to predict investment spending, the coefficient of q will be biased downward (towards zero) and the coefficient of cash flow may be biased upward. Erickson and Whited (2000) propose sophisticated estimators to deal with measurement error and obtain relatively large estimates of the relationship between q and investment.

Another issue is whether we should use the market's valuation or the firm's internal valuation of prospective investments, since the firm may have better information about the projected cash flows and speculative stock market noise can causes market prices to wander from fundamental values (Morck, Shleifer and Vishny, 1990; Blanchard, Rhee and Summers, 1993). Several authors have proposed creative ways of estimating marginal q from information available to managers (Abel and Blanchard, 1986; Gilchrist and Himmelberg, 1995) or from stock analysts' earnings forecasts (Cummins, Hasset and Oliner, 2006; Bond and Cummins, 2000). Gentry and Mayer (2006) apply the q model to real estate investment trusts (REITs) and find that the use of appraised value in place of accounting-based replacement cost increases the estimated empirical relationship between REIT investment and q.

Uses of q

Empirical studies using Tobin's q initially focused on either explaining q (Lindenberg and Ross, 1981; Salinger, 1984) or using q to predict investment spending (von Furstenberg, 1977; Summers, 1981; Hayashi, 1982), but have since broadened to include many issues in corporate finance that hold investment opportunities constant. For example, if we want to use cross-section data to see whether dividend policy affects the value of a firm, we need to control for each firm's profitability. Thus q has been used in studies of the effects of managerial equity ownership (Morck, Shleifer and Vishny, 1988; McConnell and Servaes, 1990), the size of a company's board of directors (Yermack, 1996), corporate diversification (Berger and Ofek, 1995; Rajan, Servaes and Zingales, 2000); and dividend changes (Lang and Litzenberger, 1989; Denis, Denis and Sarin, 1994). For similar reasons, Tobin's q has been used to hold investment opportunities constant while investigating the determinants of capital structure (Titman and Wessels, 1988), leveraged buyouts (Opler and Titman, 1993), and takeovers (Lang, Stulz and Walkling, 1989; Servaes, 1991).

Tobin's q will no doubt be used in many other empirical studies of corporate structure and behaviour because it circumvents the unresolved issue of how to estimate shareholders' risk-adjusted required return by looking directly at observable market prices, which incorporate both the cash flow expectations of investors and the required returns they use to discount this anticipated cash flow.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- monetary transmission mechanism
- profit and profit theory
- Tobin, James

Bibliography

Abel, A. 1983. Optimal investment under uncertainty. *American Economic Review* 73, 228–33.

Abel, A. and Blanchard, O. 1986. The present value of profits and cyclical movements in investment. *Econometrica* 54, 249–73.

Berger, P. and Ofek, E. 1995. Diversification's effect on firm value. *Journal of Financial Economics* 37, 39–65.

Blanchard, O., Rhee, C. and Summers, L. 1993. The stock market, profit, and investment. *Quarterly Journal of Economics* 108, 115–36.

Bond, S. and Cummins, J. 2000. The stock market and investment in the new economy: some tangible facts and intangible fictions. *Brookings Papers on Economic Activity* 2000(1), 61–124.

Brainard, W., Shoven, J. and Weiss, L. 1980. The financial valuation of the return to capital. *Brookings Papers on Economic Activity* 1980(2), 453–502.

Brainard, W. and Tobin, J. 1968. Pitfalls in financial model-building. *American Economic Review* 58, 99–122.

Chung, K. and Pruitt, S. 1994. A simple approximation of Tobin's q. *Financial Management* 23, 70–4.

Cummins, J., Hassett, K. and Oliner, S. 2006. Investment behavior, observable expectations, internal funds. *American Economic Review* 96, 796–810.

Denis, D., Denis, D. and Sarin, A. 1994. The information content of dividend changes: cash flow signaling, overinvestment, and dividend clienteles. *Journal of Financial and Quantitative Analysis* 29, 567–87.

Erickson, T. and Whited, T. 2000. Measurement error and the relationship between investment and q. *Journal of Political Economy* 108, 1027–57.

Gentry, W. and Mayer, C. 2006. What can we learn about the sensitivity of investment to stock prices

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
with a better measure of Tobin's q? Working paper, Columbia University.

Gilchrist, S. and Himmelberg, C. 1995. Evidence on the role of cash flow in reduced-form investment equations. *Journal of Monetary Economics* 36, 541–72.

Gomes, J. 2001. Financing investment. *American Economic Review* 91, 1263–85.

Hayashi, F. 1982. Tobin's marginal q and average q: a neoclassical interpretation. *Econometrica* 50, 213–24.

Hubbard, R. 1998. Capital market imperfections and investment. *Journal of Economic Literature* 36, 193–225.

Keynes, J. M. 1936. *The General Theory of Employment, Interest, and Money*. London: Macmillan.

Lang, L. and Litzenberger, R. 1989. Dividend announcements: cash flow signaling vs. free cash flow hypothesis? *Journal of Financial Economics* 24, 181–91.

Lang, L., Stulz, R. and Walkling, R. 1989. Managerial performance, Tobin's q, and the gains from successful tender offers. *Journal of Financial Economics* 24, 137–54.

Lee, D. and Tompkins, J. 1999. A modified version of the Lewellen and Badrinath measure of Tobin's q. *Financial Management* 28, 20–31.

Lewellen, W. and Badrinath, S. 1997. On the measurement of Tobin's q. *Journal of Financial Economics* 44, 77–122.

Lindenberg, E. and Ross, S. 1981. Tobin's q ratio and industrial organization. *Journal of Business* 54, 1–32.

Lucas, R., Jr. and Prescott, E. 1971. Investment under uncertainty. *Econometrica* 39, 659–81.

McConnell, J. and Servaes, H. 1990. Additional evidence on equity ownership and corporate value. *Journal of Financial Economics* 27, 595–612.

Morck, R., Shleifer, A. and Vishny, R. 1988. Management ownership and market valuation: an empirical analysis. *Journal of Financial Economics* 20, 293–316.

Morck, R., Shleifer, A. and Vishny, R. 1990. The stock market and investment: is the market a sideshow? *Brookings Papers on Economic Activity* 1990(2), 157–202.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mussa, M. 1977. External and internal adjustment costs and the theory of aggregate and firm investment. *Economica* 44, 163–78.

Opler, T. and Titman, S. 1993. The determinants of leveraged buyout activity: free cash flow vs. financial distress costs. *Journal of Finance* 48, 1985–99.

Perfect, S. and Wiles, K. 1994. Alternative constructions of Tobin's q: an empirical comparison. *Journal of Empirical Finance* 1, 313–41.

Rajan, R., Servaes, H. and Zingales, L. 2000. The diversification discount and inefficient investment. *Journal of Finance* 55, 35–80.

Salinger, M. 1984. Tobin's q, unionization, and the concentration-profits relationship. *RAND Journal of Economics* 15, 159–70.

Servaes, H. 1991. Tobin's q and the gains from takeovers. *Journal of Finance* 46, 409–19.

Smith, G. 1981. Investment and q in a stock valuation model. *Southern Economic Journal* 47, 1007–20.

Summers, L. 1981. Taxation and corporate investment: a q-theory approach. *Brookings Papers on Economic Activity* 1981(1), 67–127.

Titman, S. and Wessels, R. 1988. The determinants of capital structure choice. *Journal of Finance* 43, 1–19.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money Credit and Banking* 1, 15–29.

von Furstenberg, G. 1977. Corporate investment: does market valuation matter in the aggregate? *Brookings Papers on Economic Activity* 1977(2), 347–97.

Yermack, D. 1996. Higher market valuation of companies with a small board of directors. *Journal of Financial Economics* 40, 185–212.

How to cite this article

Smith, Gary. "Tobin's q." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000070> doi:10.1057/9780230226203.1714

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tobit model

Jean-Marc Robin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tobit models are used to model variables subject to exogenous censoring. For example, duration data cannot be observed longer than the survey period; hours of work cannot be observed negative although an individual might be better off consuming more leisure time than is available. This article reviews a list of econometric techniques to estimate Tobit models. Maximum likelihood, Heckman's two-stage estimator, and Powell's trimmed least squares are successively addressed.

Keywords

censored regression model; endogenous regressors; heteroskedasticity; labour supply; maximum likelihood; non-normal errors; ordinary least squares; probit model; Tobit model; trimmed least squares; two-stage estimation

Article

The Tobit model, or censored regression model, is useful to learn about the conditional distribution of a variable y^* given a vector of regressors x, when y^* is observed only if it is above or below some known threshold (censoring). In the original model of Tobin (1958), for example, the dependent variable was expenditures on durables, and values below zero are not observed.

Censoring models state that the observed dependent variable y follows from the latent variable y^* as

$$y = \max\{y^*, 0\},$$

where we have assumed a censoring of the form $y^*>0$ without loss of generality because, for any given top or bottom threshold a, it is always possible to change y^* into $\pm(y^*-a)$.

Censoring may either be a property of the sample or a property of the population. For example, top-coding of earnings in the Current Population Survey (CPS) generates censoring in a way that is independent of individual decisions. In contradistinction, the zero purchases of Tobin's households *are* individual decisions. This type of censoring is usually modelled as a corner solution of a decision-theoretic model. For example, the labour supply model predicts that the number of hours worked by a person is equal to the interior solution of the consumption-and-leisure utility maximization problem, if it is greater than zero; it is zero otherwise. (See Pudney, 1989, for a survey of the economics and econometrics of corner solutions.)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The relationship between the latent variable y^* and regressors x is assumed linear:

$$y^* = x^{\mathrm{T}}\beta + u,$$

where β is a vector of parameters, $x^{\mathrm{T}}\beta$ denotes the scalar product of x and β (^{T} is the transpose operator), and u is a residual component with cumulative distribution function (cdf) F conditional on x. We assume that the distribution of u given x, that is F, is continuous. It hence has a density $f = F'$.

The Tobit model corresponds to the particular case of $F(u) = \Phi\left(\frac{u}{\sigma}\right)$, where Φ denotes the cdf of the standard normal distribution $N(0,1)$, and σ^2 is the variance of u (that is $u \sim N(0, \sigma^2)$.

The distribution of y given x

Let $G(y|x)$ denote the cdf of the observation y given x. The distribution of y is not continuous. It has a mass point at 0. The probability mass at 0 is

$$G(0|x) \equiv \Pr\left\{y = 0|x\right\} = \Pr\left\{y^* \leq 0|x\right\} = \Pr\left\{u \leq -x^{\mathrm{T}}\beta|x\right\} = F(-x^{\mathrm{T}}\beta).$$

Notice that $F(-x^{\mathrm{T}}\beta) = 1 - F(x^{\mathrm{T}}\beta)$ if u has a symmetric distribution.

Any positive observation $y>0$ is necessarily such that $y=x^{\mathrm{T}}\beta+u$. Therefore, the cdf of the observed outcome at $y>0$ given x is equal to the cdf of u at $y-x^{\mathrm{T}}\beta$:

$$G(y|x) = F(y - x^{\mathrm{T}}\beta), \ \forall y > 0.$$

The density of any observation $y>0$ given x is

$$g(y|x) = \frac{\partial G(y|x)}{\partial y} = f(y - x^{\mathrm{T}}\beta).$$

Notice that, since 0 is a mass point of the distribution of y, its density at 0 can be defined as

$$g(0|x) = G(0|x) - G(0^-|x) = G(0|x),$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where $G(0^{-}|x) = \lim_{y \to 0^{-}} G(y|x) = 0$. (The probability density function, pdf, of a distribution or random variable is defined relative to a particular measure. Continuous variables admit a density with respect to the Lebesgue measure. Discrete distributions admit a density with respect to the counting measure. One can also define a density function for mixed discrete-continuous distributions with respect to mixtures of the Lebesgue measure and the counting measure.)

In the case of the Tobit model, $f(u) = \frac{1}{\sigma}\phi\left(\frac{u}{\sigma}\right)$, where $\phi(v) = \Phi'(v) = \frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}$ is the density of the standard normal distribution. So,

$$g(0|x) = G(0|x) = F(-x^{\mathrm{T}}\beta) = \Phi\left(\frac{x_i^{\mathrm{T}}\beta}{\sigma}\right), g(y|x) = f(y - x^{\mathrm{T}}\beta) = \frac{1}{\sigma}\phi\left(\frac{y - x^{\mathrm{T}}\beta}{\sigma}\right), \ \forall y > 0.$$

Moments of y given x

Two conditional moments of *y* are of particular interest: $\mathbb{E}(y|x)$ and $\mathbb{E}(y|x, y > 0)$. First, notice that

$$y = \max\left\{y^*, 0\right\} \geq y^*$$

implies that

$$\mathbb{E}[y|x] \geq \mathbb{E}[y^*|x]$$

and

$$\mathbb{E}[y|x, y > 0] = \frac{\mathbb{E}[y|x]}{\Pr\{y > 0|x\}} \geq \mathbb{E}[y|x].$$

So both $\mathbb{E}(y|x)$ and $\mathbb{E}(y|x, y > 0)$ overestimate the first moment of the variable of interest, that is $\mathbb{E}(y^*|x)$. Specifically,

$$\mathbb{E}[y|x] = \mathbb{E}[\max\{y^*, 0\}|x] = \mathbb{E}[\max\{x^{\mathrm{T}}\beta + u, 0\}|x] = \int_{-x^{\mathrm{T}}\beta}^{+\infty}(x^{\mathrm{T}}\beta + u)f(u)\mathrm{d}u = x^{\mathrm{T}}\beta(1 - F(-x^{\mathrm{T}}\beta)) + \int_{-x^{\mathrm{T}}\beta}^{+\infty} uf(u)\mathrm{d}u$$

and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\mathbb{E}[y|x,\ y>0] = \frac{\mathbb{E}[y|x]}{\Pr\{y>0|x\}} = x^{\mathrm{T}}\beta + \lambda(x^{\mathrm{T}}\beta)$$

with

$$\lambda(z) = \frac{\int_z^{+\infty} uf(u)\mathrm{d}u}{1-F(z)}.$$

In the particular case of the Tobit model, $\phi'(v) = -v\phi(v)$. It thus follows that $\lambda(v) = \sigma\frac{\phi\left(\frac{v}{\sigma}\right)}{\Phi\left(\frac{v}{\sigma}\right)}$. Notice that $\frac{\phi}{\Phi}$ is the inverse Mills ratio of the standard normal distribution.

Ordinary least squares

Let $\{(y_i,\ x_i),\ i=1,\ \ldots,\ N\}$ be an i.i.d. random sample of observations. Regressing y_i on x_i for the uncensored observations i such that $y_i>0$ does not yield a consistent estimator of β because of the omitted variable $\lambda(x_i^{\mathrm{T}}\beta)$ which is correlated with the regressors x_i.
For the Tobit model, a two-stage estimation procedure can be devised.

1. 1. First, estimate a Probit model for $d_i \equiv 1\{y_i>0\}$ (=•1•if $y_i>0$ and=0 otherwise):

$$\Pr\left\{d_i = 1|x_i\right\} = \Phi(x_i^{\mathrm{T}}c),$$

   with $c = \frac{\beta}{\sigma}$. Let $\hat{c}$ be the Probit estimator of c.

2. 2. Regress y_i on x_i and the inverse Mills ratio $\frac{\phi(x_i^{\mathrm{T}}\hat{c})}{\Phi(x_i^{\mathrm{T}}\hat{c})}$ by OLS. This yields a consistent estimator of β and σ.

Two remarks are in order. First, as any multi-stage estimation procedure, the OLS estimator of the second stage has $\frac{\phi(x_i^{\mathrm{T}}\hat{c})}{\Phi(x_i^{\mathrm{T}}\hat{c})}$ instead of $\frac{\phi(x_i^{\mathrm{T}}c)}{\Phi(x_i^{\mathrm{T}}c)}$. The measurement error $\frac{\phi(x_i^{\mathrm{T}}\hat{c})}{\Phi(x_i^{\mathrm{T}}\hat{c})} - \frac{\phi(x_i^{\mathrm{T}}c)}{\Phi(x_i^{\mathrm{T}}c)}$ tends to 0 when N tends to infinity. So the OLS estimator of (β, σ) is asymptotically unbiased and consistent. However, its asymptotic variance has to be corrected for the statistical error on parameter c.
Second, the first stage requires knowing the entire distribution of u_i. It is therefore not clear why one would want to use this two-stage procedure instead of maximum likelihood (ML), which is efficient.

Maximum likelihood

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The likelihood of one observation y_i conditional on x_i is $g(y_i|x_i)$. The conditional sample log-likelihood is then

$$L_N = \sum_{i=1}^{N} \ln g(y_i|x_i) = \sum_{i=1}^{N} \left\{ d_i \ln f(y_i - x_i^T\beta) + (1 - d_i) \ln F(-x_i^T\beta) \right\}.$$

Under standard regularity conditions, the values of β and any other parameters of F that maximize the log-likelihood L_N are root-N consistent and asymptotically normal and efficient.
For the Tobit model, we obtain

$$L_N = \sum_{i=1}^{N} (1 - d_i) \ln \left(1 - \Phi \left(\frac{x_i^T\beta}{\sigma} \right) \right) - N_+ \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{N} d_i (y_i - x_i^T\beta)^2$$

where $N_+ = \Sigma_{i=1}^{N} (1 - d_i)$ is the number of uncensored observations.
It is useful to change (β , σ) into $(c, s) = (\frac{\beta}{\sigma}, \frac{1}{\sigma})$ because

$$L_N = \sum_{i=1}^{N} (1 - d_i) \ln (1 - \Phi(x_i^T c)) + N_+ \ln s - \frac{1}{2} \sum_{i=1}^{N} d_i (s y_i - x_i^T c)^2$$

is strictly concave with respect to (c,s). Maximizing L_N with respect to (c,s) is easy and fast using standard gradient algorithms. One can then use the delta method to recover an estimate of the asymptotic variance of $(\hat{\beta}, \hat{\sigma}) = (\frac{\hat{c}}{\hat{s}}, \frac{1}{\hat{s}})$.
Consistency of ML obviously rests on the model being well specified. Non-normal errors and heteroskedasticity (when homoskedastic, normal errors are assumed) lead to inconsistent estimates.

Trimmed least squares

Powell's (1986) symmetrically trimmed least squares is a simple consistent estimator that is consistent under the assumption that the distribution of u_i is symmetric. It can yet be non-normal or heteroskedastic.

The idea is to replace y_i by $2x_i'\beta$ when $y_i \geq 2x_i'\beta$, if $x_i'\beta > 0$, and drop all observations such that $x_i'\beta \leq 0$ from the sample, as no symmetric trimming is possible in this case. In effect, let

$$\tilde{y}_i = \min \left\{ y_i, 2x_i^T\beta \right\} = x_i^T\beta + \tilde{u}_i$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where

$$\tilde{y}_i = \begin{cases} -x_i^T\beta & \text{if } u_i \le -x_i^T\beta \\ u_i & \text{if } -x_i^T\beta < u_i \le x_i^T\beta \\ x_i^T\beta & \text{if } x_i^T\beta < u_i \end{cases}$$

As u_i has a symmetric distribution conditional on x_i, then so does $\tilde{y}_i$. Hence,

$$\mathbb{E}[\tilde{y}_i|x_i] = x_i^T\beta.$$

The trimmed least squares estimator is obtained by iterating the following sequential procedure until convergence. Start with an initial value β_0 for β. For example, regress y_i on x_i on the uncensored sample. If, at iteration p, one has obtained a value β_p for β, then compute β_{p+1} by regressing $\tilde{y}_i(\beta_p) = \min\left\{y_i, 2x_i^T\beta_p\right\}$ on x_i using the sample of observations i such that $x_i'\beta_p > 0$.

Endogenous regressors

In the standard labour supply model, the observed dependent variable y_i is the actual number of hours worked by individual i, and the latent variable y_i^* is the interior solution to a utility maximization problem. This interior solution depends on the individual's wage, w_i, and other variables x_i such as non-labour income or education and age:

$$y_i^* = x_i^T\beta + \alpha w_i + u_i.$$

The residual u_i captures unobserved heterogeneity factors influencing the trade-off between consumption and leisure. It is usually understood that wages w_i and unobserved taste shifters u_i are correlated across individuals: $\mathrm{Cov}(w_i, u_i) \neq 0$. Suppose that w_i and x_i are both observed when y_i=0. The following simple control-function procedure can apply to solve the endogeneity problem. Suppose that there exists a vector z_i of instruments such that

$$w_i = z_i^T\gamma + v_i,$$

with $\mathrm{Cov}(z_i, v_i) = 0$. Suppose also that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$u_i = \rho v_i + \varepsilon_i,$$

with ε_i normal $N(0, \sigma^2)$ conditional on x_i, z_i and v_i. This will be the case in particular if u_i and v_i are jointly normal conditional on x_i and z_i.
Then, the following two-stage procedure produces consistent estimators of β, α and ρ.

1. 1. Regress w_i on z_i by OLS and compute residuals $\hat{v}_i$.
2. 2. Estimate the Tobit model

$$y_i = \max\left\{x_i^T\beta + \alpha w_i + \rho\hat{v}_i + \eta_i, 0\right\},$$

assuming $\eta_i = \varepsilon_i - \rho(\hat{v}_i - v_i)$ normally distributed, by ML or other appropriate method.

One can test for the exogeneity of w_i by testing for $\rho = 0$ with a standard t-test. If the null hypothesis is rejected, then this two-stage procedure yields consistent estimates, but correct asymptotic standard errors, accounting for the approximation of v_i by $\hat{v}_i$, require a specific calculation.
Finally, if w_i is not observed when y_i=0, which is the case for wages of not employed individuals, this procedure does not work. Heckman (1974) assumed joint normality of (u_i, v_i) and applied maximum likelihood to (y_i, w_i), i=1,…,N, conditional on exogenous variables.

See Also

- endogeneity and exogeneity
- logit models of individual choice
- maximum likelihood
- selection bias and self-selection
- selection bias and self-selection
- two-stage least squares and the k-class estimator

Bibliography

Blundell, R.W. and Smith, R.J. 1986. An exogeneity test for a simultaneous equation Tobit model with an application to labor supply. *Econometrica* 54, 679–86.

Blundell, R.W. and Smith, R.J. 1989. Estimation in a class of simultaneous equation limited dependent variable models. *Review of Economic Studies* 56, 37–57.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–94.

Powell, J.L. 1986. Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* 54, 1435–60.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Pudney, S. 1989. *Modelling Individual Choice: the Econometrics of Corners, Kinks and Holes*. Oxford: Blackwell.

Tobin, J. 1958. Estimation for relationships with limited dependent variables. *Econometrica* 26(2), 24–36.

How to cite this article

Robin, Jean-Marc. "Tobit model." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000207> doi:10.1057/9780230226203.1715

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tocqueville, Alexis Charles Henri Cl é rel de (1805 – 1859)

J. Goodwin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

civil society; democracy; equality of conditions; individualism; liberty; Physiocracy; Tocqueville, A. C. H. C. de

Article

Alexis de Tocqueville was born at Verneuil, in Normandy, France, on 29 July 1805. In 1831 he journeyed to the United States with his friend Gustave de Beaumont to study the American penal system. He then wrote *Democracy in America*, the first volume of which appeared in 1835, the second in 1840. Tocqueville was a member of the French Chamber of Deputies and served briefly as Minister of Foreign Affairs in the republic established after the Revolution of 1848. The events of this period are recounted in his *Recollections* (1893). Tocqueville was among those arrested during the *coup d'état* of Louis Napoleon on 2 December 1851, and he subsequently retired from public life. Tocqueville devoted his last years to a major study of the French Revolution, although he completed only the first volume before his death. This appeared as *The Old Régime and the French Revolution* in 1856. He died at Cannes on 16 April 1859.

Tocqueville was interested in the political, cultural, and, to a lesser extent, economic consequences of 'democracy', by which he meant not representative government or political arrangements of any sort, but 'equality of conditions'. (John Stuart Mill would argue that Tocqueville had confounded the effects of 'democracy' with the tendencies of modern commercial society.) By equality of conditions, Tocqueville meant neither the absence of classes nor mere equality of opportunity, but something like rough social equality, including, especially, the absence of the legally prescribed hierarchy of social groups characteristic of 'aristocratic' societies.

Tocqueville's major intellectual, not to say political, preoccupation was discovering how 'liberty' might be preserved under democratic conditions. By 'liberty' Tocqueville meant above all the local control and administration of a community's common affairs by a politically engaged and civic-minded populace. He was thus a strong critic of both the administrative centralization of the state and the narrow, self-interested 'individualism' of bourgeois society. This explains Tocqueville's appeal to those on both the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
right and left of the political spectrum.

Tocqueville's search for the institutional and ideological supports of liberty under democratic conditions was the ulterior purpose of his journey to the United States, a country which had managed to combine democracy and liberty. In *Democracy in America*, Tocqueville analysed a number of factors which he believed helped to maintain political liberty in the United States, including administrative decentralization, the profusion of voluntary associations, and what he termed ‘self-interest properly understood’, that is, a disposition to devote part of one's time and wealth to the good of the community. *The Old Régime and the French Revolution*, by contrast, explored the failure of France's revolutionary transition to democracy to produce a stable liberal regime; this Tocqueville attributed principally to the immense administrative centralization of the pre-revolutionary period and the consequent degradation of French political culture.

Tocqueville's reflections on economic matters are few. In fact, he once ‘confessed’ to Nassau Senior that he ‘was insufficiently informed on this important portion of human science’. In *The Old Régime*, however, Tocqueville did not hesitate to criticize the Physiocrats, whom he believed perhaps best represented the abstract and utopian type of intellectual nourished by the illiberal environment of pre-revolutionary France. Tocqueville thought that the Physiocrats lacked any concern for political, as opposed to economic, liberty, offering only the ‘intellectual panacea’ of universal education.

They were for abolishing all hierarchies, all class distinctions, all differences or rank, and the nation was to be composed of individuals almost exactly alike and unconditionally equal. In this undiscriminated mass was to reside, theoretically, the sovereign power; yet it was to be carefully deprived of any means of controlling or even supervising the activities of its own government.

Tocqueville's reflections on economic matters in *Democracy in America* comprise only a few pages of that voluminous work. Tocqueville argued that rents tend to rise and the terms of leases to shorten in democracies owing to the dissolution of the close, customary relationship between landlord and tenant and its replacement by the impersonal contract. He also thought that democratic conditions made it easier for workmen to combine and pressure their employers for higher wages; he thus argued that ‘a slow, progressive rise in wages is one of the general laws characteristic of democratic societies’. At the same time, Tocqueville feared that the very richest industrialists could wait out strikes and force permanently lower wages on their workers. In fact, Tocqueville believed that a dangerous business or industrial ‘aristocracy’ might arise within the womb of democratic society. However, this potential aristocracy was not to be greatly feared, Tocqueville thought, since industrialists seldom look beyond their own interests and share no common traditions or corporate spirit; still, Tocqueville warned, ‘if ever again permanent inequality of conditions and aristocracy make their way into the world, it will have been by that door that they entered’.

Selected works

1835. *Democracy in America*, vol. 1. Ed. J.P. Mayer, trans. G. Lawrence, New York: Doubleday, 1969.

1840. *Democracy in America*, vol. 2. Ed. J.P. Mayer, trans. G. Lawrence, New York: Doubleday, 1969.

1856. *The Old Régime and the French Revolution*. Trans. S. Gilbert from the 4th French edn of 1858,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
New York: Doubleday, 1955.

1872. *Correspondence and Conversations of Alexis de Tocqueville with Nassau William Senior, 1834–1859*. 2nd edn., ed. M.C.M. Simpson. Reprinted, New York: A.M. Kelley, 1968.

1893. *Recollections*. Ed. J.P. Mayer and A.P. Kerr, trans. G. Lawrence, New York: Doubleday, 1970.

How to cite this article

Goodwin, J. "Tocqueville, Alexis Charles Henri Clérel de (1805–1859)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000073> doi:10.1057/9780230226203.1716

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tooke, Thomas (1774 – 1858)

Massimo Pivetti
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank Charter Act (1844); Bank of England; Banking School; convertibility; cost of production; Currency School; endogenous money; free trade movement; Political Economy Club; price level; rate of interest; Tooke, T.; Torrens, R.; Wicksell, J. G. K.

Article

Thomas Tooke, the leading member of the Banking School, was born at St Petersburg in 1774, the eldest son of William Tooke, historian of Russia and man of letters, at that time chaplain to the English church at St Petersburg. Not a professional scientist but an active man of business of comfortable social standing, Thomas was successively a partner in the London firms of Stephen Thornton & Co. and Astell, Tooke & Thornton, Russian merchants, and was governor of the Royal Exchange Corporation and chairman of the St Katharine's Dock Company. In 1802 he married Priscilla Combe, by whom he had three sons.

As an early supporter of the free trade movement, he drew up the Merchants' Petition of the City of London, which contained the statement of the principles of free trade and was presented to the House of Commons in May 1820. He gave evidence on monetary questions before several parliamentary committees, from the Resumption Committees of 1819 to the Committees on Bank Acts in the 1850s. Tooke was elected Fellow of the Royal Society in March 1821; shortly afterwards, with Ricardo, Malthus, James Mill and others, he founded the Political Economy Club and took a prominent part in its discussions until very late in his life. He died in London on 26 February 1858. A few days later, in a letter to Engels, Marx wrote: 'Friend Thomas Tooke has died, and with him the last English economist of any value' (Letter of 5 March 1858, in Marx and Engels, 1983, p. 284).

Tooke's writings may be divided into two groups, two phases of his work which it is useful to distinguish for a better appraisal of his contribution. The first phase consists essentially in a systematic attempt to collect and analyse as much historical material and statistical information as possible, connected with price changes in England from 1793 onwards: a thorough observation of facts, aimed at understanding the determinants of fluctuations in the domestic price level. This phase is represented by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
his writings from 1823 to 1838 – from Tooke's first pamphlet *Thoughts and Details on the High and Low Prices of the Thirty Years from 1793 to 1822* to the first two volumes of his *History of Prices*, that is to say, with his *Considerations on the State of the Currency* (1826) *and* the two *Letters* to Lord Grenville (1829) in between. In the second group of writings, Tooke finally brings into focus and elaborates a few significant general principles which he gradually became certain could be derived from his observation of facts; moreover, he fully perceives the conflict between those principles and the prevailing notions, and copes with the arguments of his critics. This second phase of Tooke's work covers the writings from Volume 3 of the *History* (1840) to Volumes 5 and 6 (1857, a year before Tooke's death) and comprises his *Inquiry into the Currency Principle* (1844) – the most representative and outstanding piece, together with Volume 4 of the *History* (1848), of Tooke's voluminous work.
The main result emerging from Tooke's observation of facts in the first group of writings can be summarized as follows: the great fluctuations of prices that occurred in the 45 years following 1792 must be attributed to circumstances affecting the conditions of supply of commodities, rather than to the alterations in the system of the currency – the latter being represented by the suspension of convertibility from 1793 and its resumption after 1819 (by the Resumption Act of 1819). The prevailing view was that the value of the currency had been depreciated by the suspension, and enhanced by the contraction in the amount of circulating medium that the resumption of convertibility was alleged to have brought about. According to Tooke, ‘the most extensive induction of facts’ made it apparent that the phenomena of high prices from 1792 to 1819 and of the comparatively low prices after 1819, did not originate in the variations in the quantity of money (independently of whether the latter proceeded from the alterations in the system of the currency or from any other cause). The great fluctuations of prices originated instead from alterations in the cost of production and from other ‘accidents’ affecting supply: the character of the seasons (more unfavourable on the average from 1793 to 1818 than from 1818 to 1837); marked variations in the cost of imported commodities, as well as in the existence and removal of various obstacles (revolutions, wars) from the several sources of foreign supply; significant improvements in machinery and sciences generally, all tending to reduce the cost of production of numerous commodities (or to provide cheaper substitutes). In Volume 2 of the *History*, the rate of interest is listed for the first time amongst the causes of the high and of the low prices in the period under consideration – ‘a higher rate of interest constituting an increased cost of production’ and ‘a reduction of the general rate of interest’ leading ‘to reproduction at a diminished cost’ (1838, pp. 847 and 849). As we shall see, this is in our view the crucial point upon which hinge those aspects of Tooke's contribution that are most relevant for the modern scholar of capitalism.
The connection between money and prices occupies the centre of the stage in the second group of Tooke's writings: ‘the prepossession or prejudice’, as he puts it, that the quantity of money must have a direct influence on the prices of commodities. By ‘money’ must be understood, Tooke insists in pointing out to the supporters of the Currency Theory, not only coin and paper money (banknotes), but also cheques, bills of exchange, settlements and whatever form of paper credit which may come to be a component part of the circulating medium, performing the functions of money in daily transactions. By 1844 he was fully convinced

> that the prices of commodities do not depend upon the quantity of money indicated by the amount of bank notes, nor upon the amount of the whole of the circulating medium; but that, on the contrary, the amount of the circulating medium is the consequence of prices.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(1844, p. 123)

Tooke's evidence for this conclusion was ultimately the fact that the banks, including the Bank of England, did not appear to have the power to add to the quantity of money in circulation – unless other independent circumstances, such as an extension of trade and a rise in prices, were ‘coincidently’ in progress (1844, p. 66); nor did the banks appear to have the power to diminish the total amount of the circulation. Banks may withhold loans and discounts, and may refuse any longer to issue their own notes, but those loans, discounts and notes will be replaced in due course, Tooke argues, ‘by other expedients calculated to answer the same purpose’ (1844, p. 122). Only compulsory paper money issued directly by a government in payment for goods and services (like the French *assignats*) constitutes a fresh source of demand, so that alterations in its quantity act directly as an originating cause on prices (see 1844, pp. 68–78; 1848, pp. 183–97).

The power of the banks to expand and contract the quantity of the circulating medium at pleasure, was taken for granted by the Currency School and by most writers. It was a challenging task, for Tooke and the Banking School, to convince those writers of the lack of such a power, and, in consequence, of the fact that such alterations in the quantity of money as do actually occur are the *effect* of increased transactions and prices, and not the *cause* of them. Tooke has the great merit of having succeeded in bringing into focus the heart of the matter: the question of the effects of changes in the rate of interest on the inducement to purchase commodities. ‘Abundance of money’ – that is, a high disposition on the part of the banks to make advances in the way of loan or discount – results in the first place in a high price of securities and a low rate of interest; thus the power of the banks to add to the amount of the circulating medium, and hence to act as an originating cause on trade and prices, will ultimately depend on whether a low rate of interest supplies the *stimulus* to purchase commodities. Tooke points out that actual experience does not validate the notion that the facility of borrowing at a low rate of interest, not only confers the power of purchasing commodities, but also affords the motive and inducement to do it. ‘The error’, he says, ‘is in supposing the *disposition* or *will* to be co-extensive with the power’ (1844, p. 79). No relation of cause and effect between variations in the rate of interest and variations in the demand for commodities can be inferred from trustworthy evidence (compare 1857, vol. 5, p. 345).

The questions of the connection between money and prices and between the rate of interest and the price level are thus clearly seen as two sides of the same coin. Arguing against the dominant opinion that a low rate of interest raises prices and that a high rate depresses them, Tooke actually maintained that a persistent reduction in the rate of interest constitutes a reduction in the cost of production, which could not fail, by the competition of the producers, to bring about a fall of prices (compare 1844, p. 81). He went so far as to state that it is difficult to find evidence of facts more in contrast with the influence ascribed to a low rate of interest in raising prices and vice versa: ‘The theory is not only not true, but the reverse of the truth’ (1844, p. 84).

It is important to notice that Tooke's conception of the relation between the rate of interest and the price level, and the connected notion of ‘endogenous money’ (as we would now call his view of the relation between money and prices), are in no way contingent upon the particular currency system of his day, with the relevant part played in it by precious metals. Rather, it is the denial of any power on the part of the banks to regulate at will the amount of the circulating medium, together with the emphasis on the circumstances affecting supply, that provide Tooke with the basis for his criticism of the idea that every influx or efflux of the precious metals must cause a rise and fall of prices, independently of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
circumstances connected with the cost of production of commodities. On that same basis he opposes the prevailing view that the discovery of a gold mine within the premises of the Bank of England – Ricardo's famous assumption in his first pamphlet (1811) – would necessarily raise the prices of commodities. And he argues that for an increased production of gold to be associated with a permanent rise in the prices of commodities, measured in gold, the increased production must be the consequence either of the discovery of more fertile mines, or of improved methods of working the existing ones (1848, pp. 199 ff.; 1857, vol. 6, pp. 413–4).

Monetary policy questions, naturally, permeate all of Tooke's writings. The chief place amongst them is occupied by the Bank Charter Act of 1844 and the controversies that both preceded and followed its implementation – controversies centred upon the idea of a separation of the business of the note issue from the banking business of the Bank of England. This idea, opposed by Tooke, was given statutory effect by the Act, and brought about in due time many of the shortcomings that had been foreseen by Tooke (for an extensive critical account of Tooke's views on banking policy, see Gregory, 1928; 1929, vol. 1).

On several policy issues that are still relevant today, the modern orthodox scholar of monetary questions and central banking policy is likely to find himself more in agreement with Tooke's views than with those of the supporters of the Currency School. This hardly applies, however, to those views of Tooke's which are more strictly connected with his conception of the relation between money and prices, and of the influence of the rate of interest on the price level. As an important example of one such view, one may refer to Tooke's contention that, as the Bank of England and the banks collectively cannot arbitrarily change the amount of the circulating medium, nor operate through that medium on the prices of commodities, the only ‘infallible means’ they have to influence foreign exchanges – ‘so as to arrest a drain, or to resist an excessive influx’ – is by a forcible operation on securities: a great advance in the rate of interest on the one hand, or a great reduction of it in the other. Now, the articulated line of argument laid down by Tooke in discussing the power of the central bank to influence foreign exchanges (cf. 1844, pp. 123–4), appears on the whole no less alien from today's quantity of money approach to problems of general prices than it was in Tooke's day.

Besides Marx (see above, and also 1857–8; 1859), Tooke's most outstanding contemporaries who praised his work and ideas were Malthus (1823) and J.S. Mill (1844; 1852, ch. 24, pp. 203–4). (Malthus's appreciation, however, must not be overrated: he tends to understand Tooke's early contribution merely as a confirmation of his own views on value against those of Ricardo – namely, ‘that everything must be attributed to supply and demand’, rather than simply to ‘labour and the costs of production’; 1823, p. 218.) The most strenuous opponent of Tooke's ideas and policy recommendations was Robert Torrens (1840; 1844; 1848). This author's criticisms grew increasingly severe as Tooke's work advanced with the development of more general principles from the empirical analyses. By 1844, Tooke's thesis that the prices of commodities do not depend upon the quantity of money is referred to and criticized as ‘the most astonishing of the many astonishing fallacies’ (Torrens, 1844, p. 43). Torrens's criticisms are extensively dealt with by Tooke in Volume 4 of the *History of Prices* (1848), and by Fullarton (1844) and Wilson (1847).

If Torrens was the most outstanding critic of Tooke amongst classical economists, Knut Wicksell, the father of 20th-century monetary theory, has been his most outstanding critic since the inception of marginalism. Wicksell's conceptions ultimately constitute the main reference-point of this century's (not

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
so large) literature in which Tooke's work and ideas are somehow taken into consideration, starting from Gregory's *Introduction* to the *History of Prices* (1928). In fact, we can look today at Tooke and Wicksell as the chief exponents of two alternative ways of reasoning about the connection between money and prices. Wicksell's criticisms of Tooke's view are somewhat vitiated by their being mostly based upon the interest elasticity of the demand for loan capital, as postulated by the marginalist theory (see 1898, ch. 7; 1906, pp. 175–208). There is, however, one important criticism which does not reflect Wicksell's tendency to superimpose upon Tooke's view his own theory. He criticizes Tooke's reasoning about the effect of the rate of interest on the cost of production and commodities prices, as entailing that every persistent move in either direction would cause a progressive divergence of both interest and prices from their initial levels: a persistent reduction in the rate of interest

> would lead to a reduction … in the demand for loans by business people, money would flow into the banks and would cause a further reduction of interest rates, and so on, until the rate fell to nil – In other words, the money rate of interest would be in a state of unstable equilibrium. (Wicksell, 1906, p. 187)

This conclusion actually follows, not from Tooke's view of the influence of the rate of interest on prices, but from his conception of the rate of interest as a magnitude ‘entirely governed by the supply of and demand for monied capital’, on which the central bank can exercise only a *temporary* influence (1826, sect. I; 1857, pp. 556–7; see also Newmarch, 1857, pp. 66–72).

Not to have acknowledged that the monetary authorities do have the power of determining the rate of interest – albeit a power exercised under a wide range of constraints – constitutes, in our opinion, the main shortcoming of Thomas Tooke and the Banking School.

Selected works

1823. *Thoughts and Details on the High and Low Prices of the Last Thirty Years from 1793 to 1822*. London: John Murray.

1826. *Considerations on the State of the Currency*. London: John Murray.

1829a. *A Letter to Lord Grenville, on the effects ascribed to the resumption of cash payments on the value of the currency*. London: John Murray.

1829b. *A Second Letter to Lord Grenville, on the Currency in Connexion with the Corn Trade and on the Corn Laws, to which is added a postscript on the present commercial stagnation*. London: John Murray.

1838. *A History of Prices and of the State of the Circulation from 1793 to 1837*, 2 vols. London: Longman, Orme, Brown, Green & Longmans.

1840. *A History of Prices and of the State of the Circulation in 1838 and 1839, with remarks on the corn*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

laws and some of the alterations in our banking system. London: Longman, Orme, Brown, Green & Longmans.

1844. *An Inquiry into the Currency Principle; the connection of the currency with prices and the expediency of a separation of issue from banking*, 2nd edn. Series of Reprints of Scarce Works on Political Economy No. 15, London: London School of Economics and Political Sciences, 1959.

1848. *A History of Prices and of the State of the Currency from 1839 to 1847 inclusive: with a general review of the currency question and remarks on the operation of the Act 7 & 8 Vict. 32*. London: Longman, Brown, Green & Longmans.

1856. *On the Bank Charter Act of 1844, its Principles and Operations; with suggestions for an improved administration of the Bank of England*. London: Longman, Brown, Green & Longmans.

1857. (With W. Newmarch.) *A History of Prices and of the State of the Circulation During the Nine Years 1848–1856* (in two volumes; forming the fifth and sixth volumes of the history of prices from 1792 to the present time). London: Longman, Brown, Green, Longmans & Roberts.

Bibliography

Fullarton, J. 1844. *On the Regulation of Currencies; being an examination of the principles on which it is proposed to restrict, within certain fixed limits, the future issues on credit of the Bank of England, and of the other banking establishments throughout the country*. London: John Murray.

Gregory, T.E. 1928. *An Introduction to Tooke and Newmarch's A History of Prices and of the State of the Circulation from 1792 to 1856*. Series of Reprints of Scarce Works on Political Economy No. 16. London: London School of Economics and Political Sciences, 1962.

Gregory, T.E. 1929. *British Banking Statutes and Reports, 1832–1928*, 2 vols. London: Oxford University Press.

Johnson, A. 1856. *Currency Principles Versus Banking Principles; being strictures on Mr Tooke's pamphlet on the Bank Charter Act of 1844*. London: Richardson Brothers.

Malthus, T.R. 1823. Review of Tooke's ‘Thoughts and Details on the High and Low Prices of the Last Thirty Years’. *Quarterly Review* 29(April), 214–39.

Marx, K. [1857–8]. *Grundrisse* [Rough Draft]: *Foundations of the Critique of Political Economy*. Harmondsworth: Penguin, 1973.

Marx, K. 1859. *A Contribution to the Critique of Political Economy*. Moscow: Progress Publishers, 1978.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Marx, K. and Engels, F. 1983. *Collected Works*, vol. 40. London: Lawrence & Wishart.

Mill, J.S. 1844. Review of Tooke's ‘An Inquiry into the Currency Principle’. *Westminster Review*, March–June.

Mill, J.S. 1852. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*, 3rd edn. London: John W. Parker and Son.

Newmarch, W. 1857. Evidence before the Select Committee of the House of Commons on Bank Acts, 5th June. In Gregory (1929, vol. 2).

Ricardo, D. 1811. The High Price of Bullion, a Proof of the Depreciation of Bank Notes. In *The Works and Correspondence of David Ricardo*, 4th edn. vol. 3, ed. P. Sraffa. Cambridge: Cambridge University Press, 1966.

Ricardo, D. 1811. Reply to Mr Bosanquet's Practical Observations on the Report of the Bullion Committee. In *The Works and Correspondence of David Ricardo*, vol. 3, ed. P. Sraffa. Cambridge: Cambridge University Press, 1966.

Torrens, R. 1840. *A Letter to Thomas Tooke, esq. in reply to his objections against the separation of the business of the Bank into a department of issue, and a department of deposit and discount: with a plan of bank reform*. London: Longman, Orme, Brown, Green & Longmans.

Torrens, R. 1844. *An Inquiry into the Practical Working of the Proposed Arrangements for the Renewal of the Charter of the Bank of England, and the Regulation of the Currency: with a refutation of the fallacies advanced by Mr Tooke*. London: Smith, Elder & Co.

Torrens, R. 1848. *The Principles and Practical Operation of Sir Robert Peel's Bill of 1844 Explained and Defended against the Objections of Tooke, Fullarton, and Wilson*. London: Longman, Brown, Green & Longmans.

Wicksell, K. 1898. *Interest and Prices*. London: Macmillan, 1936.

Wicksell, K. 1906. *Lectures on Political Economy*. Vol. 2: *Money*. London: Routledge & Kegan Paul, 1962.

Wilson, J. 1847. *Capital, Currency, and Banking; being a collection of a series of articles published in the ‘Economist’ in 1845, on the principles of the Bank Act of 1844, and in 1847, on the recent monetarial and commercial crisis; concluding with a plan for a secure and economical currency*. London: *The Economist*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Pivetti, Massimo. "Tooke, Thomas (1774–1858)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000078>
doi:10.1057/9780230226203.1717

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Torrens, Robert (1780 – 1864)

B.A. Corry
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

absolute advantage; Bank Charter Act (1844); Banking School; capital theory of value; colonization; comparative advantage; convertibility; corn-ratio theory of profits; Currency School; invariable standard of value; labour theory of value; real bills doctrine; reciprocal tariffs; Ricardo, D.; Sraffa, P.; terms of trade; Torrens, R.

Article

Torrens, if not in the top rank of the classical economists, or in the class for example of Ricardo, Senior or John Stuart Mill, certainly was of the second rank and the equal of, or even above, James Mill or McCulloch in terms of originality, theoretical reasoning and the range of economic topics that he considered. His work was almost completely neglected in the years after his death in 1864 and his re-emergence to his rightful place as an important member of the Classical School was initially due to Seligman in his famous article ‘On Some Neglected British Economists’ (1903) and later to the definitive study by Lionel Robbins (1958). In recent years Torrens has also come to the fore again because of the debates surrounding the Sraffa interpretation of Ricardo.

Robert Torrens was a most prolific writer and produced a vast quantity of books and pamphlets on all sorts of economic matters for over 50 years. His first publication appeared in 1808 (*The Economists Refuted*) and his last in 1858 (*Lord Overstone on Metal and Paper Currency*). He managed all of this against the background of an extremely busy life that included several different careers. He was a professional soldier – a colonel in the Marines – and was decorated for gallantry at the battle of Anholt. Subsequently he became the proprietor of the *Globe* newspaper, a Member of Parliament, the planning genius behind the colonization and development of New South Wales, a founder member of the Political Economy Club and many other things besides. He even found time to write two never-read novels, the *Hermit of Killarney* and *Coelibia in Search of Husband*, both of which contain hefty chunks of economic discourse.

His specific contributions to economics may be dealt with under the general headings of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
microeconomics, theory of money and banking, commercial policy and colonization.
Torrens's main contributions to microeconomics concerned the Ricardian system. He objected to the search for an absolute, invariant measure of value and also tried to replace the labour quantity theory of value with a capital theory of value – where relative commodity values are determined by relative capital inputs. He did not however fully realize that his definition of capital combining wages and materials was different from Ricardo's which was really only wages.
On the other hand, and somewhat inconsistently, it is now clear (see Langer, 1982; de Vivo, 1985) that Torrens fully understood the corn-ratio theory of profits that Sraffa ascribed to Ricardo, and that he (Torrens) derived this from Ricardo. He also saw, following Ricardo that given the agricultural rate of profit, the price of manufactured goods relative to corn, was given. All of this is clearly spelt out in the second edition of *An Essay on the External Corn Trade* (1820) and must therefore lead one to doubt Hollander's argument (1979) that Ricardo did not mean the corn-ratio theory of profit and the key role of the agriculture sector in his analysis to be taken too seriously.
In the first edition (1815) of the *Essay* Torrens has a clear statement of the principle of comparative advantage well before its more popularly ascribed origins in Ricardo's *Principles* (1819). He makes a clear distinction between absolute and comparative advantage and indeed he actually hints at this distinction in his earlier *Economist Refuted* (1808).
In the field of money and banking Torrens is best known for his championing of the Currency School in their debate with the Banking School. Essentially the currency principle was that a mixed currency, that is, a currency consisting of notes and coins, should be regulated so that movements in it were the same as under a purely metallic currency. Unlike the bullionists, however, the Currency School did not believe that convertibility alone would achieve the conformability of a mixed currency to a metallic one. To this end, Torrens may claim to have been the originator of the plan, activated in the Bank Charter Act of 1844, to separate the issue and banking department of the Bank of England. This he did in his *Letter to Lord Melbourne* (1837) and he later vigorously defined the legislation in his *Principles and Practical Operation of Sir Robert Peel's Bill of 1844* (1848).
Students of Torrens find an inconsistency with this aspect of his monetary thinking and his earlier espousing of the anti-bullionist position; in particular his *Essay on Money and Paper Currency* (1812) is a strong plea for a paper currency without convertibility and relying on the real bills doctrine to prevent excess issue. The reasons for abandoning this extreme anti-bullionist position and his switch to the Ricardian line are explained in his *On the Means of Establishing a Cheap, Secure and Uniform Currency* (1828).
Torrens's main contribution to the theory of commercial policy was to suggest a modification of the general classical case for free trade. He pointed out, and was amongst the first to do so, that a country might alter its terms of trade in its favour by use of an import tariff. In a series of letters to Lord John Russell (published in 1844 as *The Budget*) he argued the case for what he termed ‘reciprocity’, that is, if some countries had tariffs unilateral free trade was a mistaken policy and in these cases reciprocal tariffs should be adopted. Against the change that he was abandoning the central classical (Ricardian) belief in free trade, Torrens replied that he was just applying the logic of the Ricardian analysis.
Finally, Torrens had a significant influence on the theory and practice of colonization. Along with most of the later classical economists he rejected the Smithian view that colonies were of no economic benefit to the colonial power. Much of the later classical case for colonies was based on the view that colonies would provide profitable investment outlets to offset a declining rate of profit at home. Torrens used this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
argument in some of his later writings but his main argument was that colonies were an ideal solution to the Malthusian overpopulation problem. In this view he was undoubtedly influenced by his interpretation of the causes of Irish poverty – a country incidentally where Torrens was born and in general its problems had a profound effect on his thinking.

In terms of colonization policy Torrens, like Wakefield, was opposed to the movement of labour on to free land on the grounds that this would lead to a dispersed population and land holdings of suboptimal size. He advocated systematic colonization with the price of land set sufficiently high that large units of capital would have to be amassed before the immigrant labourers became independent farmers.

Selected works

1820. *An Essay on the External Corn Trade*. London: Longman, Rees, Orme, Brown & Green.

1821. *An Essay on the Production of Wealth*. London: Longman, Hurst, Rees, Orme, & Brown.

1835. *Colonization of South Australia*. London: Longman, Rees, Orme, Brown, & Green.

1837. *A Letter to the Right Honourable Lord Viscount Melbourne*. London: Longman, Rees, Orme, Brown & Green.

1844. *The Budget*. London: Smith, Elder & Co.

1847. *On the Operation of the Bank Charter Act of 1844*. London: Smith & Elder.

Bibliography

De Vivo, G. 1985. Robert Torrens and Ricardo's ‘corn-ratio’ theory of profits. *Cambridge Journal of Economics* 9, 89–92.

Langer, G.F. 1982. Further evidence for Sraffa's interpretation of Ricardo. *Cambridge Journal of Economics* 6, 397–400.

Robbins, L.C. 1958. *Robert Torrens and the Evolution of Classical Economics*. London: Macmillan & Co. (This book includes brief summaries of all of Torrens's writings.)

Seligman, E.R.A. 1903. On some neglected British economists. *Economic Journal* 13, 335–63, 511–35.

How to cite this article

Corry, B. A. "Torrens, Robert (1780–1864)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000080>
doi:10.1057/9780230226203.1718

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

total factor productivity

Diego Comin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Total factor productivity (TFP) is the portion of output not explained by the amount of inputs used in production. This article sets out the measurement and importance of TFP for growth, fluctuations and development as well as likely future directions of research.

Keywords

endogenous growth; innovation; patents; real business cycles; research and development; Solow residual; technology; total factor productivity

Article

Total factor productivity (TFP) is the portion of output not explained by the amount of inputs used in production. As such, its level is determined by how efficiently and intensely the inputs are utilized in production.

TFP growth is usually measured by the Solow residual since Solow (1957). Let g_Y denote the growth rate of aggregate output, g_K the growth rate of aggregate capital, g_L the growth rate of aggregate labour, and alpha the capital share. The Solow residual is then defined as $g_Y - \alpha * g_K - (1 - \alpha) * g_L$. The Solow residual accurately measures TFP growth if (a) the production function is Cobb–Douglas, (b) there is perfect competition in factor markets, and (c) the growth rates of output and the inputs are measured accurately.

TFP plays a critical role on economic fluctuations, economic growth and cross-country per capita income differences. At business cycle frequencies, TFP is strongly correlated with output and hours worked. Based on this observation, Kydland and Prescott (1982) initiated the real business cycle (RBC) literature. In the standard business cycle model, shocks to TFP are propagated by pro-cyclical labour supply and investment, thereby generating fluctuations in output and labour productivity at business cycle frequencies with an amplitude that resembles the US data. Subsequent work has introduced pro-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
cyclical fluctuations in measured TFP by incorporating unmeasured labour hoarding and/or capacity utilization in the standard framework (see, for example, Burnside, Eichenbaum and Rebelo, 1995; Basu, 1996; King and Rebelo, 1999). In this way, TFP fluctuations can be driven by shocks to aggregate demand in addition to the standard interpretation that attributes them to aggregate supply shocks.

As shown in the landmark article by Robert Solow (1956), long-run growth in income per capita in an economy with an aggregate neoclassical production function must be driven by growth in TFP. For over 30 years, the conceptual difficulty when trying to endogenize TFP growth was how to pay for the fixed costs of innovation in a perfectly competitive economy with constant returns to scale in capital and labour. In this context, all output is exhausted by paying capital and labour their marginal products; therefore, no resources are left to pay for the innovation costs. Romer (1990) and Aghion and Howitt (1992) solved this problem by granting the innovator monopolistic rights over his innovation, which are sustainable through the patent system. In this way, innovators can recoup the initial fixed costs of innovation through the profit margin they make from commercializing their patent.

By linking the TFP growth rate to innovation, endogenous growth models shed light on the determinants of TFP growth. R&D subsidies and an abundance of skilled labour reduce the marginal cost of conducting R&D and increase the rate of innovation development and, therefore, the TFP growth rate. Expanding markets increase the innovators' revenues, leading to more innovation and higher TFP growth.

Solow (1956) also demonstrated that cross-country differences in technology may generate important cross-country differences in income per capita. Klenow and Rodriguez-Clare (1997) and Hall and Jones (1999) have confirmed that most of the gap in income per capita between rich and poor countries is associated with large cross-country differences in TFP. Cross-country differences in TFP can be due to differences in the physical technology used by countries or in the efficiency with which technologies are used. To explore the relative importance of these factors, it is necessary to have data on direct measures of technology. Comin, Hobijn and Rovito (2006) put together direct measures of technology adoption for approximately 75 different technologies and show that the cross-country differences in technology are approximately four times larger than cross-country differences in income per capita. Further, technology is positively correlated to income per capita. Thus, cross-country variation in TFP is, to a large extent, determined by the cross-country variation in physical technology.

Likely future directions

Economic fluctuations

Recognizing that a large portion of TFP growth is caused by endogenous innovation decisions has significant implications for the business cycle. This is likely to be an important research topic. Comin and Gertler (2006) show that low-persistence, non-technological shocks generate pro-cyclical fluctuations in the market value of innovations. Agents arbitrage these innovation opportunities and generate a pro-cyclical rate of innovation development and, hence, of TFP growth. The model-induced fluctuations in TFP are as large and persistent as in the data. More important, by linking a component of TFP to innovation activity, TFP becomes a mechanism that propagates low-persistence shocks, thus

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
increasing its persistence, rather than a source of disturbances as in standard RBC models. This same logic can be extended to other processes that determine the endogenous level of technology, such as endogenous technology adoption processes, which are more relevant in developing economies. This may be an important ingredient to understanding high and medium-term fluctuations in developing economies.

Long-run growth

A significant fraction of innovations are not patented. For some, this is because they are not embodied in any new good or are not a recipe for a new chemical process and, therefore, are not patentable. Others are not patented because innovators simply decide not to apply for a patent. Three important areas of research are to understand (*a*) how important patents are for innovation activity, (*b*) the determinants of non-patentable innovations and (*c*) how they interact with the patentable R&D type of innovations that fit the properties of the Romer (1990) and Aghion and Howitt (1992) models.

Two papers have argued that patents are not necessary for the innovators to recoup the innovation costs. Innovators in Hellwig and Irmen (2001) can obtain rents to cover innovation costs despite being perfectly competitive because they face an increasing marginal cost of producing the intermediate goods that embody their innovations. Boldrin and Levine (2000) model innovation in perfectly competitive settings. In their model, to copy an innovation it is necessary to purchase one unit of the good that embodies it. Hence, the innovator is the monopolist of the first unit produced, and the revenues he extracts from selling it may cover the innovation costs, making up for a lack of patent protection.

Comin and Mulani (2006) model the development of disembodied innovations such as managerial and organizational techniques, personnel, accounting and work practices, and financial innovations. These are very different from embodied innovations in that the rents extracted by the innovators are not associated with selling the innovation per se. This has some interesting implications. First, the revenues accrued by the innovator–producer originate from the increased efficiency in producing his good or service with the innovation. If the innovator–producer has some monopolistic power in the market for his good or service, the increased efficiency from using the innovation in production yields an increase in profits that may cover the innovating costs. Second, since the innovator–producer's gain from innovating comes from the increased efficiency of production, the marginal private value of developing disembodied innovations is increasing in the value of the firm. This has important cross-sectional and time-series implications. In the cross-section, firms with higher values (resulting from larger sizes or ability to charge higher markups) have more incentives to develop disembodied innovations. In the time series, shocks that reduce the value of the firm reduce its incentives to develop disembodied innovations. One such shock may be an increase in the probability that a competitor steals the market. If the occurrence of this shock requires the development of a new patentable product, the model implies the possibility of an aggregate trade-off between investments in developing disembodied innovations and embodied innovations. A complete understanding of the determinants of these different types of innovation may be critical for explaining secular TFP dynamics.

Development

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Understanding the determinants of technology adoption is key to explaining cross-country variation in TFP. On the theory side, an increasing number of theories link the adoption of technologies to the role of institutions (Acemoglu, Antras and Helpman, 2007), financial markets (Alfaro et al., 2006; Aghion, Comin and Howitt, 2006), endowments (Caselli and Coleman, 2006) and policies (Holmes and Schmitz, 2001). The challenge is to bring these theories to the data and assess their empirical relevance. The new country-level data on measures of micro technologies must be an important input towards this goal.

See Also

- endogenous growth theory
- real business cycles
- technology

I thank Steven Durlauf for helpful comments.

Bibliography

Acemoglu, D., Antras, P. and Helpman, E. 2007. Contracts and technology adoption. *American Economic Review* 97, 916–43.

Aghion, P., Comin, D. and Howitt, P. 2006. When does domestic saving matter for economic growth? Working Paper No. 12275. Cambridge, MA: NBER.

Aghion, P. and Howitt, P. 1992. A model of growth through creative destruction. *Econometrica* 60, 323–51.

Alfaro, L., Chanda, A., Kalemi-Ozcan, S. and Sayek, S. 2006. How does foreign direct investment promote economic growth? Exploring the effects of financial markets on linkages. Working Paper No. 12522. Cambridge, MA: NBER.

Basu, S. 1996. Procyclical productivity: increasing returns or cyclical utilization? *Quarterly Journal of Economics* 111, 719–51.

Boldrin, M. and Levine, D. 2000. Growth under perfect competition. Mimeo, UCLA.

Burnside, C., Eichenbaum, M. and Rebelo, S. 1995. Capital utilization and returns to scale. In *NBER Macroeconomics Annual*, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Caselli, F. and Coleman, J. 2006. The world technology frontier. *American Economic Review* 96, 499–522.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Comin, D. and Gertler, M. 2006. Medium term business cycles. *American Economic Review* 96, 523–51.

Comin, D., Hobijn, B. and Rovito, E. 2006. Five facts you need to know about technology diffusion. Working Paper No. 11928. Cambridge, MA: NBER.

Comin, D. and Mulani, S. 2006. A theory of growth and volatility at the aggregate and firm level. Working Paper No. 11503. Cambridge, MA: NBER.

Hall, R. and Jones, C. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114, 83–116.

Hellwig, M. and Irmen, A. 2001. Endogenous technical change in a competitive economy. *Journal of Economic Theory* 101, 1–39.

Holmes, T.J. and Schmitz, J.A., Jr. 2001. A gain from trade: from unproductive to productive entrepreneurship. *Journal of Monetary Economics* 47, 417–46.

King, R. and Rebelo, S. 1999. Resuscitating real business cycles. In *Handbook of Macroeconomics*, vol. 1B, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Klenow, P. and Rodriguez-Clare, A. 1997. The neoclassical revival in growth economics: has it gone too far. In *NBER Macroeconomics Annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Kydland, F. and Prescott, E. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, 1345–70.

Romer, P. 1990. Endogenous technological change. *Journal of Political Economy* 98(5), S71–S102.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70 (1), 65–94.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39, 312–20.

How to cite this article

Comin, Diego. "total factor productivity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000081>
doi:10.1057/9780230226203.1719

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tournaments

Tor Eriksson
From The New Palgrave Dictionary of Economics Online, , 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Tournament theory is a theory of promotion-based incentives which also contributes to the understanding of firms' wage structures and individual earnings. In tournaments wage-rank differentials act as an incentive scheme when firms cannot directly observe employees' effort. Empirical evidence is mainly from sports and lab experiments while there are fewer studies of businesses and organizations.

Keywords

pay; promotions; raises; relative performance; wage structures

Article

A significant proportion of employees work together with other workers, some of whom are performing the same tasks. Moreover, for many jobs it is difficult to observe individual performance. Under such conditions relative performance schemes are frequently used a mechanism for motivating employees. These can take the form of competition for promotions or schemes that award tenure when performance exceeds a certain standard. Because of the similarity of these reward schemes to those found in professional sports competitions, scholars have named them tournaments. Tournament theory aims at explaining promotions and raises associated with them. As a considerable share of pay increases occurs through job (title) changes, the theory complements other explanations of individual earnings differentials. Tournament models also contribute to our understanding of firms' wage structures and differences therein.

The basic model

The basic tournament model developed in Lazear and Rosen (1981) has risk-neutral agents, a firm with two identical contestants (employees), operating in a competitive industry. The employee's utility is a function of her net income, that is, the difference between income from work and the cost of effort.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Output is the sum of the two employees' effort and two stochastic disturbances: a common and an individual-specific component. The firm does not directly observe the employees' efforts. The payment scheme is fixed in advance and has a winner's and a loser's prize to the better and poorer performer of the contestants, respectively. The prizes are independent of levels as well as of differences in employees' performance. The worker's problem is to choose her optimal level of effort taking into account the firm's compensation scheme. In other words she maximises her expected utility, which depends on the probability of winning the two prizes and her cost of effort.

As the contestants are assumed to be identical ex ante, the probability of winning is 0.5 and they choose the same equilibrium level of effort, which depends on two factors. The first is the difference between the winner's and the loser's prize. A larger prize spread induces employees to compete harder for earning the promotion and hence to exert more effort. Second, the more important individual-specific random components in output are, the less effort the contestants will provide.

The firm also needs to set wage levels high enough to attract workers to participate in the tournament. The solution is that the firm chooses a wage spread that induces employees to put forth effort up to the point where the marginal cost of effort equals the marginal benefit of it to the firm. Thus, the tournament compensation scheme is efficient and brings about first-best level of effort. Luck also affects the firm's decision. To maintain a given level of effort, an increase in the random component in output has to be offset by an increase in the wage spread.

Extensions and predictions

The Lazear-Rosen article contained analyses where some of the assumptions of the basic tournament model were relaxed, and this line of research continued in a series of articles (Green and Stokey, 1983; Nalebuff and Stiglitz, 1983; O'Keeffe, Viscusi and Zeckhauser, 1984) which offered several extensions; see also the survey by McLaughlin (1988). These papers examined a number of additional variations on tournaments such as winning by a gap, rewards depending on distance from the loser's performance and multiple prizes. Empirically, these are of rather limited interest and have not been followed up in the recent literature.

Assumptions that turned out to affect equilibrium outcomes qualitatively were risk-neutrality, only two contestants, homogeneous agents and the one-stage tournament. The key result from relaxing the assumption of risk-neutrality is that the optimal level of effort is lower than the first-best level as risk-averse contestants want income insurance. Thus, risk aversion narrows the wage spread which yields lower effort. Does the size of the tournament matter? Intuitively, the larger the number of contestants, the lower the expected probability of winning, so in order to induce same level of effort, the firm has to widen the wage spread. Consequently, the spread increases in tournament size. With risk-averse agents this result is fragile (McLaughlin, 1988), but for relevant tournament ranges the spread is plausibly increasing.

Allowance for heterogeneous contestants has been modelled by assuming that ability differentials equal differences in marginal cost of effort. Two information structures for heterogeneity in ability been studied: (i) agents know their own but not their contestant's ability, and (ii) full knowledge. In both cases mixing high and low ability agents gives rise to inefficiency; in (i) because of no sorting and in (ii) because the tournament is not attractive to low performers. In the asymmetric information case, more

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
knowledge (such as entry credentials) is needed. Alternatively a larger pay spread can induce self-sorting. In the full knowledge case, segregating low and high performers also leads to inefficiency (in both leagues). Handicaps may overcome some of the problems: typically at the expense of efficiency, however.

In many firms there are usually several rounds of promotion competitions over workers' careers. Rosen (1986) analyses the multistage tournament, more precisely a sequential, elimination tournament like in professional tennis. In this setting promoted employees earn not only a raise but also the expected value of continued competition. With a constant spread, effort would decline through rounds (as the tournament shrinks in size) and would be lowest at the top. Rosen shows that in order to keep effort unchanged, the wage spread has to grow linearly with rank, and will make a jump at the top to compensate for the absence of further competition. An implication of the analysis is that the firm uses a convex pay-rank schedule as a means for providing incentives to all its employees.

A key prediction of tournament theory is that increased wage spread yields higher effort and output. However, the participation constraint puts an upper bound on the optimal wage spread. When cooperation among employees is important, or when jobs are strongly interdependent, relative performance games may give rise to too strong incentives for uncooperative behaviour. Extending the basic model to allow for employees to behave strategically against their rivals shows that sabotaging behaviour lowers equilibrium effort (Lazear, 1989). Thus, firms might also compress their internal wage structures on efficiency grounds. Another form of strategic employee behaviour that can lower effort is collusion among employees. This (and sabotaging) is more likely when there are few contestants. Mitigating harmful effects by increasing the tournament size is costly. As the contestant pool grows, it becomes more heterogeneous giving rise to increased inefficiency. Tournament models also predict that firms will favour insiders over outsiders (Chan, 1996).

Several papers compare (the efficiency of) tournaments with other individual incentive pay schemes, in particular piece rates. The question that has not been addressed much is why do (especially larger) firms use tournament structures? The principal advantage of tournaments is that contestants are insulated from common risks. Thus, tournaments are predicted to be more common in risky environments. The other main advantage is that it is less costly to measure relative than absolute performance. A third, but less well understood, possibility is that promotions also can serve other functions, notably sorting of employees.

Empirical tests

Tournament theory yields quite a number of testable predictions. These have been tested on data from sports, laboratory experiments and businesses. The studies examine whether observed empirical patterns are consistent with tournament models, but none has to my knowledge tested tournaments against the other theoretical explanation for promotions: promotions as signals (Waldman, 1984).

Following Ehrenberg and Bognanno's (1990) study of the effects of prize structures on professional golfers’ performance, a large empirical sports literature has built up, providing in the main evidence in support of several of tournament theory's predictions. Sports data are rich and of high quality, but to what extent the results are generalizable to firms remains an open question.

Almost from the beginning tournaments have been studied by means of lab experiments that allow

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
researchers to control for a host of parameters while examining a number of treatments, and also to compare with other incentive schemes. Beginning with Bull et al. (1987), several experimental studies have confirmed the predictions about the impact of wage spread on performance but have found a high variance in subjects' effort. In general many of the other predictions like those concerning tournament size, number of rounds and degree of uncertainty have been supported by the laboratory evidence. Again, generalizability to businesses is an issue. Some key concepts like cost of effort or attitudes to risk are very difficult to observe outside the lab. However, as has been shown in some recent experiments, endogenous sorting can have profound effects on results. Generalizing from the lab to the upper echelons of organizations can in particular be associated with external validity problems.

Evidence outside sports and the lab is rather scarce. The evidence that exists is often based on data from a single firm, specific industries or firms' managerial personnel. Many studies have tested the prediction that a larger wage spread increases employees' effort and hence firm performance. The hypothesis finds support in many, but not all, studies. Another implication that has been tested is the convexity of firms' wage structures. These results are, however, consistent not only with tournament theory but also with other theories. A problem in testing tournaments is that there are potential alternative explanations for individual predictions. Thus, a positive relation between pay spread and firm performance, and convex within-firm wage structures, are also consistent with convex returns to ability due to magnification effects (in hierarchies bosses' decisions matter more) and with the promotions-as-signals hypothesis. More direct tests concern the hypotheses of a positive relation between number of contestants and the magnitude of the raise from promotion, and that tournament organizations favour insiders. One way to arrive at more convincing evidence is to test several hypotheses on the same data set. Only a few studies (e.g. Eriksson, 1999; Knoeber and Thurman, 1994) have done this. A number of single-firm studies have documented more facts about wage and promotion dynamics within firms. Although one should be cautious in treating them as stylized facts, a fruitful avenue for further research on tournaments is to extend tournament models to account for patterns observed in these studies.

See Also

- hierarchy
- incentive compatibility
- payment systems
- sliding scales (wages)

Bibliography

Bull, C., Schotter, A. and Weigelt, K. 1987. Tournaments and piece-rates: An experimental study. *Journal of Political Economy* 95, 1–33.

Chan, W. 1996. External recruitment versus internal promotion. *Journal of Labor Economics* 14, 555–70.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ehrenberg, R. and Bognanno, M. 1990. Do tournaments have incentive effects? *Journal of Political Economy* 98, 1307–24.

Eriksson, T. 1999. Executive compensation and tournament theory: empirical tests on Danish data. *Journal of Labor Economics* 17, 262–80.

Green, J. and Stokey, N. 1983. A comparison of tournaments and contracts. *Journal of Political Economy* 91, 349–64.

Knoeber, C. and Thurman, W. 1994. Testing the theory of tournaments: an empirical analysis of broiler production. *Journal of Labor Economics* 12, 155–79.

Lazear, E. 1989. Pay equality and industrial politics. *Journal of Political Economy* 97, 561–80.

Lazear, E. and Rosen, S. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89, 841–64.

McLaughlin, K. 1988. Aspects of tournament models: a survey. In *Research in Labor Economics*, Vol. 9, pp. 225–56. ed. R. Ehrenberg. Greenwich, Conn.: JAI Press.

Nalebuff, B. and Stiglitz, J. 1983. Prizes and incentives: towards a general theory of compensation and competition. *Bell Journal of Economics* 14, 21–43.

O'Keeffe, M., Viscusi, K. and Zeckhauser, R. 1984. Economic contests: comparative reward schemes. *Journal of Labor Economics* 2, 27–56.

Rosen, S. 1986. Prizes and incentives in elimination tournaments. *American Economic Review* 76, 701–15.

Waldman, M. 1984. Job assignments, signalling, and efficiency. *Rand Journal of Economics* 15, 255–67.

How to cite this article

Eriksson, Tor. "tournaments." The New Palgrave Dictionary of Economics Online. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 05 March 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000250> doi:10.1057/9780230226203.1898

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Townshend, Hugh (1890 – 1974)

Victoria Chick

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

liquidity preference; loanable funds; Townshend, H.; Walras's Law

Article

Townshend is an anomaly amongst economists, for he owes his reputation almost entirely to one brilliant article. He took a First in mathematics at Cambridge in 1912 and stayed on to prepare for the civil service examinations under Keynes's supervision. He served with the Post Office, where his duties included economic forecasting.

His correspondence with Keynes over the just-published *General Theory* (Keynes, 1979) reveals both the extent of Townshend's intellectual grasp of that complex and difficult book and something, whether derived from his studies with Keynes or innate in his temperament, which allowed him to accept aspects of *The General Theory* which others resisted. These qualities bore fruit in the famous 12-page article, published as a note in the *Economic Journal* (1937a). This note takes issue with Hicks's attempt, in his review of the *General Theory* (Hicks, 1936; Keynes, 1936), to transform the theory of liquidity preference into a mirror image of loanable funds theory by Walras's Law. Townshend saw that this was an attempt to retain the link between prices and the flow concepts of cost and demand. In contrast, he argued, it was in the nature of Keynes's liquidity preference theory that expectations of the future could change the value of assets overnight and be reflected in market prices of those assets even in the absence of actual trading. Thus current prices could be determined by subjective as well as objective factors and future prices were indeterminate.

Townshend's achievement was to ‘follow liquidity preference theory where it led: to the destruction of determinate price’ (Shackle, 1967). Keynes had left this implicit.

Townshend's restatement required the courage to go against established modes of thought: whether by temperament or because they are imbued with outdated conceptions of science, most economists are determinists. Townshend's stock can only rise as the methodological change that has occurred in science becomes known across the divide between science and the arts.

Townshend also wrote four book reviews for the *Economic Journal* (1937b; 1938; 1939; 1940) which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

show a breadth of conception and keenness of intellect from which one wishes economics had benefited more.

Selected works

1937a. Liquidity-premium and the theory of value. *Economic Journal* 47, 157–69.

1937b. Review of R.G. Hawtrey, *Capital and Employment. Economic Journal* 47, 321–6.

1938. Review of G.L.S. Shackle, *Expectations, Investment and Income. Economic Journal* 48, 520–23.

1939. Review of H. Munro, *Principles of Monetary-Industrial Stability. Economic Journal* 49, 102–5.

1940. Review of F. Hayek, *Profits, Interests and Investment. Economic Journal* 50, 99–103.

Bibliography

Hicks, J.R. 1936. Mr. Keynes's theory of employment (review article). *Economic Journal* 46, 238–53.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, J.M. 1973, 1979. *Collected Writings of J.M. Keynes*, vols 14 and 29, ed. D.E. Moggridge. London: Macmillan.

Shackle, G.L.S. 1967. *The Years of High Theory: Invention and Tradition in Economic Thought, 1926–1939*. Cambridge: Cambridge University Press.

How to cite this article

Chick, Victoria. "Townshend, Hugh (1890–1974)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000085> doi:10.1057/9780230226203.1720

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Toynbee, Arnold (1852 – 1883)

A. Kadish
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic history; Industrial Revolution; liberalism; New Liberalism; Toynbee, A.

Article

Arnold Toynbee, best known for his lectures on the Industrial Revolution (published posthumously in May 1884), was during the late 1870s and early 1880s a major influence on the shape and direction of the interest at Oxford in socio-economic questions and their history.

Born in London on 23 August 1852, Arnold Toynbee was the fourth child and second son of Dr Joseph Toynbee, FRS, a philanthropist and a successful aural surgeon. Initial plans for his education at Rugby were first delayed by an accident at the age of 13 or 14 which resulted in severe concussion and, in the long run, in recurring migraines, impeding prolonged mental exertion. These plans were finally shelved for financial reasons following Joseph Toynbee's death in a laboratory accident. After two unprofitable years in a military preparatory school, and some classes at King's College London, Toynbee's education was reduced to long periods of solitary reading.

He developed an independent if unsystematic bent, coupled with overconfidence in his capability to master on his own any subject which might catch his attention.

Having come into a modest inheritance from his father's estate at the age of 21, Toynbee entered Pembroke College, Oxford, in January 1873 with the intention of reading for Greats. He shortly afterwards migrated to Balliol, which he found socially and intellectually more attractive. For health reasons he eventually settled for a Pass degree, obtained in 1878. However, he had impressed the college sufficiently to be offered a tutorship in charge of the candidates for the Indian civil service. In 1881 he was appointed Senior Bursar, and at the time of his death was about to be elected Fellow. Toynbee had meanwhile become involved in a number of public causes, all of which may be seen as related to efforts to revive liberalism as a radical-reformist movement. These included Church reform aimed at the democratization of Church of England government on the parish level, adult education through the cooperative movement, rural reform, Irish land reform, and municipal politics. In 1883 he stood, unsuccessfully, as a liberal candidate for one of the north ward seats on the Oxford City Council

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(represented by T.H. Green until his death in 1882) and Toynbee may well have contemplated the possibility of a political career. His academic interests reflect his search for a scientifically reasoned programme for comprehensive reform, while his inquiries were greatly influenced by a fear of the consequences of rising working-class radicalism.

The Industrial Revolution lectures, delivered in 1881–2, offered a liberal interpretation of industrialization and its political, social and economic consequences as an alternative to both socialist and laissez-faire views of industrial society. Despite their ideological bias and fragmentary form they constituted an important departure in English economic history. They demonstrated the usefulness of an historical approach to the study of industrial society, thereby suggesting an alternative to economic theory, which at the time was increasingly regarded as either morally or ideologically unacceptable or as inapplicable to current conditions. Hence Toynbee directly contributed to the Oxford inclination towards an empirical and historical approach to the study of socio-economic questions. In addition, Toynbee outlined the possibility of an historical and, thereby, a relativist consideration of economic theories, seen as reflecting historical circumstances in which they were formed and, while of a limited general application, of considerable interest to the historian. Finally, and perhaps most importantly, the Industrial Revolution lectures suggested an autonomous approach to the study of economic history, based on its own type of primary sources, in which economic circumstances were not placed in causal subservience to political developments (as in the work of W. Cunningham). Toynbee's approach was ideologically and philosophically acceptable to the next generation of economic historians. It was not materialistic yet it tended to regard economic change in terms of general impersonal trends rather than attributing it to the conscious action of narrow interest groups (as in the work of J.E.T. Rogers).

Following an attempt to convince a hostile London audience of the futility of land nationalization and the desirability and attainability of liberal alternatives, Toynbee suffered a nervous breakdown. While convalescing he contracted meningitis and died at the age of 30, widely hailed as a martyr to the cause of social harmony and to the type of reformism which became known as New Liberalism. His reputation and martyrdom contributed to the popularity of the university settlements, beginning with Toynbee Hall, and a general interest at Oxford and Cambridge in social and economic reform. His importance as an economic historian, however, emerged only with the development of the study some years after his death.

See Also

- Industrial Revolution

Selected works

A partial collection of Toynbee's work including a reconstruction of his Industrial Revolution course based on his and on some of his students' notes was edited by his friend Alfred Milner and published posthumously as *Lectures on the Industrial Revolution in England. Popular addresses, notes and other fragments*, London: Rivingtons, 1884. *‘Progress and Poverty’: a criticism of Mr H. George: being two lectures delivered in … London* (London: Kegan Paul, 1883) was published in pamphlet form and appended to the 2nd edition (1887) of *Lectures on the Industrial Revolution in England*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Kadish, A. 1986. *Apostle Arnold. The Life and Death of Arnold Toynbee (1852–1883)*. Durham, NC: Duke University Press.

Montague, F.C. 1889. *Arnold Toynbee*. Baltimore: Johns Hopkins University Studies in Historical and Political Science.

How to cite this article

Kadish, A. "Toynbee, Arnold (1852–1883)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000086> doi:10.1057/9780230226203.1721

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tozer, John Edward (1806 – 1877)

G. Campanelli
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

fixed capital; mathematics and economics; Tozer, J. E.; Whewell, W.

Article

Tozer was born at Woolwich in 1806 and died in London in 1877. He was privately educated and at the age of 26 was admitted as a ‘Pensioner’ of Caius College, Cambridge. He immediately showed his ability in mathematics by carrying off the first prize in 1833 and 1834. In the four years after his graduation in 1836, Tozer wrote two essays on mathematical economics, the first on ‘Machinery’ (1838), and the second on ‘Landlords’ (1840). However, soon after the publication of these two papers he abandoned economics and went into law, in which he achieved distinction. Later he went into university administration.

Tozer's two papers on economics represent a systematic application of mathematical reasoning to political economy. In that period Whewell, at Trinity College, Cambridge, was trying to introduce mathematical analysis into economics and Tozer adopted his method. Like Whewell, Tozer believed that mathematics, because it turned economics into a ‘science’ characterized by a series of propositions leading to ‘axiomatic truths’, was not only appropriate but necessary to the subject (Tozer, 1838, pp. 1–2).

Of the two essays on economics, that on ‘Machinery’ is by far the more interesting. In this paper, Tozer wants to provide a mathematical basis for the idea that the employment of machinery always increases the wealth of community. His final conclusion is that the capitalist is ‘not only … unable to secure his own advantage at the expense of any other class, he cannot even prevent a general participation in the benefit’ (Tozer, 1838, p. 10).

However, the most original feature of this paper is Tozer's mathematical treatment of the problem of machinery. In particular the calculation of the annuity and the algebraic formulation of the construction period of machinery can undoubtedly be regarded as a sophisticated contribution to the fixed-capital debate of that time.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- machinery question

Selected works

1838. *Mathematical Investigation of the Effect of Machinery on the Wealth of a Community in which it is Employed and on the Fund for the Payment of Wages*. Cambridge: Cambridge Philosophical Society, Transaction 6.

1840. *On the Effect of the Non-Residence of Landlords on the Wealth of a Community*. Cambridge: Cambridge Philosophical Society, Transaction 7.

How to cite this article

Campanelli, G. "Tozer, John Edward (1806–1877)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000087>
doi:10.1057/9780230226203.1722

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tradable and non-tradable commodities

A.D. Woodland
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The distinction between internationally tradable and non-tradable commodities lies at the heart of the reason for the development of the theory of international trade as an area of economics distinct from the general theory of value. The existence of non-tradable commodities as well as tradable commodities implies that some markets are domestic while others are international. Connections between the prices of these two sets of commodities have been the subject of the famous factor-price equalization and Stopler–Samueslon theorems. Non-tradable goods play a prominent role in the analysis of many problems, such as tariff reform, exchange rates and international transfers.

Keywords

Bastable, C. F.; complementarities; factor-price equalization; globalization; Heckscher–Ohlin–Vanek model; interest rate differentials; international portfolio choice; international trade theory; Mill, J. S.; project evaluation; purchasing power parity; real exchange rates; reduced form analysis; Ricardo, D.; Rybczynski theorem; shadow pricing; sticky prices; Stolper–Samuelson theorem; substitutes and complements; tariffs; terms of trade; Torrens, R.; tradable and non-tradable commodities

Article

The distinction between internationally tradable and non-tradable commodities lies at the heart of the reason for the development of the theory of international trade as an area of economics distinct from the general theory of value. If the theory of value approach were adopted, nations would simply be a collection of production and consumption units, each with its own monetary and political system. The international trade literature has, however, imposed the assumption that certain classes of commodities are non-tradable. Accordingly, there exist purely domestic markets as well as international markets, and this is arguably the most important distinguishing feature of the international trade literature.
The early classical economists made a clear distinction between products which were assumed to be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
internationally tradable, and factors of production such as land, labour and capital, which were assumed to be non-tradable internationally but perfectly mobile within each nation. This distinction, which is central to the writings of such eminent classical theorists as Ricardo, Torrens, J.S. Mill and Bastable, was carried on by early 20th-century writers such as Taussig (1927), Yntema (1932), Ohlin (1933), Haberler (1936) and Mosak (1944). The modern literature, initiated perhaps by Samuelson (1953), has continued with the same basic model structure.

The traditional international trade model, therefore, contains non-tradable commodities. However, they are in the form of non-producible factors of production rather than products, and this provides important structure to the traditional model. Moreover, and perhaps more importantly, further structure is typically imposed by assuming that the non-tradable factors are in fixed supply. As a result, consumer preferences do not play a role in the market for non-tradable commodities. It therefore follows that the traditional model of trade encompasses non-tradable commodities in a very special way.

A more general, and more useful, way to deal with non-tradable commodities is to define a set of commodities that are non-tradable, and allow this set to include products as well as factors. While the existence and importance of non-tradable commodities in the form of products was recognized as early as, for example, Ohlin (1933, ch. 8), Haberler (1936, pp. 34–5) and Taussig (1927, ch. 5), it was not until much later that non-tradable products were explicitly incorporated into international trade models. For earlier surveys of this literature, see McDougall (1970) and Woodland (1982, ch. 8).

There is the fundamental question of why certain commodities are not traded internationally. Some may be non-traded because of their intrinsic nature; they are simply not transportable. Others may be transportable and hence tradable but are not traded because it is unprofitable to do so due to the costs of transportation or other expenses such as tariffs. Finally, products may be tradable, but trade in them may be illegal – an extreme form of trade quota.

While it has long been recognized that there is a difference between a commodity being tradable and being traded, and that the difference arises as a result of the profitability of trade, few models actually deal with non-traded commodities in this way. Rather, it is typically assumed that transport costs are zero for some commodities (tradable) and infinite, or, at least, sufficiently large for others (non-tradable) so as to preclude their trade. Exceptions do exist. Hadley and Kemp (1966) and Woodland (1968) explicitly model transportation and thus endogenize the division into traded and non-traded commodities. Xu (2003), using a continuum of goods with transport costs and tariffs, has the boundary between traded and non-traded goods endogenous. Melitz (2003) models a continuum of firms with different levels of productivity producing a continuum of varieties in which only the more productive firms export, the remainder producing varieties only for the domestic market.

Relationships between domestic and international markets

The existence of non-tradable commodities as well as tradable commodities implies that some markets are domestic while others are international. However, this does not mean that the domestic markets operate in isolation from international markets. On the contrary, the prices of domestic (non-tradable) commodities will be influenced by activities in the international markets. Though of less apparent interest, the reverse is also true: the prices of internationally traded commodities are influenced by activities in domestic markets. The various national markets for non-tradable commodities are, of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

course, connected only indirectly via the international markets for tradable commodities.

Let $p = (p_t, p_n)$ denote the partition of the price vector for a nation into tradable and non-tradable commodities, and let $X(p) = (X_t(p), X_n(p))$ denote the vector of excess supply functions correspondingly partitioned. If the foreign nation's functions and variables are distinguished by an *, the perfectly competitive equilibrium conditions for internationally tradable commodities and the non-tradable commodities of the home and foreign nations may be written as

$$X_t(p_t, p_n) + X_t^*(p_t, p_n^*) = 0 \quad (1)$$

$$X_n(p_t, p_n) = 0 \quad (2)$$

$$X_n^*(p_t, p_n^*) = 0. \quad (3)$$

Under reasonable regularity conditions the market equilibrium conditions for the domestic commodities in the home nation, (2), may be solved for p_n as a function of p_t as $p_n = P_n(p_t)$. Similarly, (3) may be solved as $p_n^* = P_n^*(p_t)$. Thus, the equilibrium conditions for domestic commodities provide the connection between the prices of tradable commodities and the prices of domestic or non-tradable commodities. An elegantly simple analysis of the connection between the markets for tradable and non-tradable commodities is provided by Jones (1974).

A significant amount of the international trade literature has been devoted to the relationship between prices of tradable and non-tradable commodities. Within the context of factors being the only non-tradable commodities, there are several famous propositions that emerge. The Stolper–Samuelson (1941) theorem indicates, within a two-product, two-factor model, that if the relative price of one product increases then one factor price will increase proportionately more and the other factor price will fall. The factor whose price increases, and whose real income therefore increases, is the one used relatively intensively by the product whose price increases. Jones (2006) provides a historical perspective on this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
theorem and the literature it generated. Second, the factor-price equalization theorem provides conditions under which the factor-price vectors, here p_n and p_n^*, are the same in the two nations despite differences in factor endowments. The sufficient conditions are that each nation should have the same production technology and that their endowment vectors be sufficiently close so that each nation is diversified and produces the same set of traded goods (Samuelson, 1953; McKenzie, 1955; Dixit and Norman, 1980). Recent contributions include Blackorby, Schworm and Venables (1993) and Chakrabarti (2006), who provide necessary and sufficiency conditions.

The role of non-tradable commodities

The reason non-tradable commodities complicate the analysis of many problems in international trade theory is that, whenever there is a disturbance to equilibrium, there will be an effect on the markets for non-tradable goods. The price of non-tradable goods will have to adjust to restore equilibrium. This adjustment of prices will, in general, affect the variable of interest and may yield different qualitative results than are obtained from a model without non-tradable commodities (factors or products). Much of the literature is concerned with the question of how the introduction of non-tradable products affects results obtained from models with only tradable products. However, some recent literature has gone the other way, enquiring whether trade in one or more factors alters results obtained on the assumption that all factors are non-tradable.

As an example of the role that non-tradable products can play, consider an increase in the international price for a small open economy's import good. If there are just two traded products, a single consumer, and all non-tradable commodities are factors in fixed supply, then it is well known that the quantity of imports will fall. The introduction of a third product that is produced and consumed but not traded internationally can upset this result if there are sufficient complementarities in production or consumption. The rise in the price of the imported product causes imports to change directly, and indirectly via the consequent change in the price of the non-tradable product. At the initial price of the non-tradable good, the import price increase induces higher home production and lower demand since the income effect is unambiguously negative. This direct effect implies lower imports, just as when non-tradable commodities don't exist. However, if imports and the non-tradable good are net complements, an excess supply of the non-tradable good ensues from the import price increase (ruling out inferiority). Its price then falls to clear the domestic market. This fall in price causes a reduction in the net supplies of both the imported and the non-tradable goods since they are net complements. If this indirect reduction in the net supply of the importable good is sufficiently strong to outweigh the direct positive effect of the increase in its price, the quantity of imports will rise. Thus arises the paradoxical case where an increase in the price of the imported product causes the level of imports to rise. This case occurs because of the assumed net complementarity between the imported and non-tradable products.

It is noteworthy that many of the difficulties that occur when non-tradable products are included in a model arise in the consumption sector. In the example of the previous paragraph, if there is a fixed demand for the non-tradable product the indirect income effect vanishes and so the quantity of imports falls in response to a rise in their price. Alternatively, one can ensure this result by assuming that the non-tradable and imported products are net substitutes.

For some problems the existence of non-tradable commodities has no substantial influence upon the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
formal solution. The prices of non-tradable products can be eliminated from the equilibrium conditions by first solving for their prices in terms of the prices for tradable goods, and then substituting into the excess supply functions for tradable goods. This yields the excess supply functions expressed in terms of tradable goods' prices, and the equilibrium conditions reduce to

$$\ddot{X}(p_t) = X_t(p_t, P_n(p_t)) = 0. \quad (4)$$

These resulting 'reduced form' excess supply functions can then be used directly to analyse various problems. For example, the usual stability conditions for international equilibrium can be applied to the reduced form excess supply functions on the assumption that domestic markets clear instantly. The problem of the effect of a transfer of income upon the terms of trade can be similarly handled. For details on the reduced form approach to non-tradable goods see Dixit and Norman (1980, pp. 89–92) and Woodland (1982, pp. 172–3, 218–22). Those that prefer to deal with the structural form include Komiya (1967), McDougall (1970) and Jones (1974).

Situations where the role of non-tradable commodities is prominent

For other problems, or where interest centres on the variables in the structural model relating to non-traded goods, the existence of non-tradable commodities has to be explicitly taken into account. Some specific instances are as follows:

1. 1. In the case of shadow pricing of commodities for the purpose of evaluating the welfare effects of a public project in the presence of tariffs, the existence of non-tradable commodities provides special complications. While tradable commodities should be evaluated using world prices, the appropriate shadow prices for non-tradable commodities depend in a complex way upon technology and taste conditions (Warr, 1982; Dinwiddy and Teal, 1987).
2. 2. Several analyses of technological advances in the production of a traded product upon the output levels of other traded goods, and upon the real exchange rate, have focused attention upon the role played by non-tradable commodities. In general, the effect of the 'boom' upon the other tradable good's production is ambiguous, stemming partly from the adjustments in the market for non-tradable commodities. For details, see Corden and Neary (1982) and references therein.
3. 3. The effect of the introduction of non-tradable products upon the Stolper–Samuelson, Rybczynski and factor-price equalization theorems was thoroughly analysed by Ethier (1972). In the case of the Stolper–Samuelson theorem, a change in the relative price of the non-tradable products induces a change in the price of the non-tradable product. This has a further effect upon factor prices. The question of whether a particular factor gains in real income depends upon the directions of these price changes, and can be answered only from knowledge of the technology and preferences. In the case of the factor-price equalization theorem, Mainwaring (1978)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

demonstrates that non-traded goods may result in the non-equalization of national interest rates, while Deardorff and Courant (1990) argue that non-traded goods reduce the likelihood of factor-price equalization by reducing the size of the diversification cone.

4. 4. The ‘purchasing power parity’ theory of exchange rates states that the relative value of currencies equals the relative purchasing power of each currency in its domestic markets. Clearly, the existence of non-tradable commodities with different prices across countries will cause the purchasing power of each currency to be different independently of the value of the exchange rate.
5. 5. The responses to currency devaluation are also potentially affected by the existence of non-tradable products. An example of such an analysis is Dornbusch (1973), who concludes that his basic results hold up when a non-tradable product is introduced into the model. Neary (1980) models sticky prices in domestic labour and product markets and shows that short-run policy (for example, devaluation) responses are affected. More generally, non-traded goods affect macroeconomic variables in dynamic models. They explain persistent deviations from purchasing power parity and interest rate differentials (Backus and Smith, 1993) and affect the responses to an oil price shock (Marion, 1984), while productivity shocks in traded and non-traded goods sectors have different effects (Murphy, 1986).
6. 6. The welfare implications of tariff reform depend upon the existence of non-tradable products. Following early work by McDougall (1970) and Dornbusch (1974) on the role of non-traded goods for tariff reform, Hatta (1977) and Fukushima (1979) have shown that the policy of reduction of the highest rate of import duty to the next highest is welfare improving if non-tradable and tradable products are net substitutes. If they are sufficiently complementary, the policy may reduce welfare. Non-traded goods are incorporated, and play important roles, in the analyses by Diewert, Turunen-Red and Woodland (1989; 1991) of tariff reform and in the trade restrictiveness index of Anderson, Bannister and Neary (1995). Xu (2003) shows that increased wage inequality can arise from the shifting boundary between traded and non-traded goods as a result of a tariff reform.
7. 7. The effect of a transfer of income from abroad upon the welfare of a small open economy is affected by adjustments in the non-traded goods markets and may lead to the transfer paradox whereby the recipient's welfare declines, as shown by Yano and Nugent (1999). Schweinberger (2002) emphasizes the role of complementarities between non-traded and traded goods in this context.
8. 8. Non-traded goods play a prominent role in the modelling of international portfolio choice and explanations for home country bias (Tesar, 1993; Baxter, Jermann and King, 1998).
9. 9. Davis and Weinstein (2001) show that, while the Heckscher–Ohlin–Vanek model is empirically rejected, a model that accounts for non-traded goods is consistent with the data for trade by ten OECD countries.

Concluding comments

It is somewhat surprising that non-tradable products (as opposed to factors) have only relatively recently

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
attracted explicit attention in the literature, given that most economic activity is in the markets for domestic products. Perhaps more future studies will introduce non-tradable products automatically, except where a reduced form analysis is applicable, and give more attention to the endogeneity of the division of commodities into traded and non-traded groupings. Globalization, whereby the costs of undertaking international trade are being reduced by technological improvements in transport and communications and trade policy liberalization, will continue to change the boundary between traded and non-traded commodities. The trend towards more outsourcing of services and intermediate inputs from foreign countries provides a clear example.

See Also

- factor prices in general equilibrium
- globalization
- tariffs
- terms of trade
- trade costs

Bibliography

Anderson, J.E., Bannister, G.I. and Neary, J.P. 1995. Domestic distortions and international trade. *International Economic Review* 36, 139–57.

Backus, D.K. and Smith, G.W. 1993. Consumption and real exchange rates in dynamic economies with non–traded goods. *Journal of International Economics* 35, 297–316.

Baxter, M., Jermann, U.J. and King, R.G. 1998. Nontraded goods, nontraded factors, and international non-diversification. *Journal of International Economics* 44, 211–29.

Blackorby, C., Schworm, W. and Venables, A. 1993. Necessary and sufficient conditions for factor price equalization. *Review of Economic Studies* 60, 413–34.

Chakrabarti, A. 2006. Factor price equalization beyond a ‘cubic’ world. *Economic Theory* 27, 483–91.

Corden, W.M. and Neary, J.P. 1982. Booming sector and deindustrialization in a small open economy. *Economic Journal* 92, 825–48.

Davis, D.R. and Weinstein, D.E. 2001. An account of global factor trade. *American Economic Review* 91, 1423–53.

Deardorff, A.V. and Courant, P.N. 1990. On the likelihood of factor price equalization with nontraded goods. *International Economic Review* 31, 589–96.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Diewert, W.E., Turunen-Red, A.H. and Woodland, A.D. 1989. Productivity- and Pareto-improving changes in taxes and tariffs. *Review of Economic Studies* 56, 199–215.

Diewert, W.E., Turunen-Red, A.H. and Woodland, A.D. 1991. Tariff reform in a small open multi-household economy with domestic distortions and nontraded goods. *International Economic Review* 32, 937–57.

Dinwiddy, C. and Teal, F. 1987. Shadow prices for non-traded goods in a tax-distorted economy: formulas and values. *Journal of Public Economics* 33, 207–21.

Dixit, A.K. and Norman, V. 1980. *Theory of International Trade*. Cambridge: Cambridge University Press.

Dornbusch, R. 1973. Devaluation, money, and nontraded goods. *American Economic Review* 63, 871–80.

Dornbusch, R. 1974. Tariffs and nontraded goods. *Journal of International Economics* 4, 177–85.

Ethier, W. 1972. Nontraded goods and the Heckscher–Ohlin model. *International Economic Review* 13, 132–47.

Fukushima, T. 1979. Tariff structure, non-traded goods and theory of piecemeal policy recommendations. *International Economic Review* 20, 427–35.

Haberler, G. 1936. *The Theory of International Trade with its Applications to Commercial Policy*. Trans. from the 1933 German edition by A. Stonier and F. Benham. London: William Hodge.

Hadley, G. and Kemp, M.C. 1966. Equilibrium and efficiency in international trade. *Metroeconomica* 18 (2), 125–41.

Hatta, T. 1977. A recommendation for a better tariff structure. *Econometrica* 45, 1859–69.

Jones, R.W. 1974. Trade with non-traded goods: the anatomy of inter-connected markets. *Economica* 41, 121–38.

Jones, R.W. 2006. Protection and real wages: the history of an idea. *Japanese Economic Review* 57, 457–66.

Komiya, R. 1967. Non-traded goods and the pure theory of international trade. *International Economic Review* 8(2), 132–52.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mainwaring, L. 1978. Interest rate equalization theorem with nontraded goods. *Journal of International Economics* 8, 11–9.

Marion, N.P. 1984. Nontraded goods, oil price increases and the current account. *Journal of International Economics* 16, 29–44.

McDougall, I.A. 1970. Non-traded commodities and the pure theory of international trade. In *Studies in International Economics*, ed. I.A. McDougall and R.H. Snape. Amsterdam: North-Holland.

McKenzie, L.W. 1955. Equality of factor prices in world trade. *Econometrica* 23, 239–57.

Melitz, M.J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71, 1695–725.

Mosak, J.L. 1944. *General Equilibrium Theory in International Trade*. Bloomington, IN: Principia Press.

Murphy, R.G. 1986. Productivity shocks, non-traded goods and optimal capital accumulation. *European Economic Review* 30, 1081–95.

Neary, J.P. 1980. Nontraded goods and the balance of trade in a neo-Keynesian temporary equilibrium. *Quarterly Journal of Economics* 95, 403–29.

Ohlin, B. 1933. *Interregional and International Trade*. Cambridge, MA: Harvard University Press. Reprinted 1952.

Samuelson, P.A. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21, 1–20.

Schweinberger, A.G. 2002. Foreign aid, tariffs and nontraded private or public goods. *Journal of Development Economics* 69, 255–75.

Stolper, W. and Samuelson, P.A. 1941. Protection and real wages. *Review of Economic Studies* 9, 58–73.

Taussig, F.W. 1927. *International Trade*. New York: Kelley, 1966.

Tesar, L.L. 1993. International risk-sharing and non-traded goods. *Journal of International Economics* 35, 69–89.

Warr, P.G. 1982. Shadow pricing rules for non–traded commodities. *Oxford Economic Papers* 34, 305–25.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Woodland, A.D. 1968. Transportation in international trade. *Metroeconomica* 20(2), 130–5.

Woodland, A.D. 1982. *International Trade and Resource Allocation*. Amsterdam: North-Holland.

Xu, B. 2003. Trade liberalization, wage inequality, and endogenously determined non-traded goods. *Journal of International Economics* 60, 417–31.

Yano, M. and Nugent, J.B. 1999. Aid, nontraded goods, and the transfer paradox in small countries. *American Economic Review* 89, 431–49.

Yntema, T.O. 1932. *A Mathematical Reformulation of the General Theory of International Trade*. Chicago: University of Chicago Press.

How to cite this article

Woodland, A. D. "tradable and non-tradable commodities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000095> doi:10.1057/9780230226203.1723

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

trade and poverty

Nina Pavcnik
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the relationship between trade and poverty, with specific focus on this link in developing countries. Since the 1980s, our understanding of the link between trade and poverty has been increasingly informed by empirical evidence. We first review the literature that emphasizes the dynamic effects of trade on poverty via growth. We next consider the literature on the static relationship between trade and poverty through changes in relative prices of goods and wages of the less educated. We conclude with a discussion of trade and child labour.

Keywords

child labour; foreign direct investment; Hecksher–Ohlin trade theory; import substitution; labour mobility; Mercosur; outsourcing; poverty; poverty alleviation; returns to schooling; skill-biased technical change; Stolper–Samuelson theorem; trade and poverty; trade liberalization; wage inequality

Article

Trade liberalization is one of the most common policy prescriptions offered to initiate the process of poverty eradication in poor countries. Since the 1980s, our understanding of the link between trade and poverty in developing countries has been increasingly informed by empirical evidence. One part of the literature considers dynamic effects of trade on poverty via growth, while other parts emphasize the more static relationship between trade and poverty through changes in relative prices of goods and wages of the less educated.

Growth is potentially the most important channel through which trade might affect poverty. The usual argument goes as follows: trade promotes growth and growth leads to lower poverty. While many economists believe that these dynamic effects provide an important channel towards poverty reduction, the link between trade and poverty via growth has been empirically elusive. Many cross-country studies on trade and growth, perhaps best exemplified by Frankel and Romer (1999), find a positive association

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
between trade and growth. But others, most notably Rodriguez and Rodrik (2001), have questioned the robustness of these findings and whether the positive association is indicative of good domestic policies and institutions in countries with economically sound trade policy. Growth could lower poverty by expanding employment and earnings opportunities of the poor, but growth could also bypass the poor (see Ravallion, 2001). Two influential empirical studies by Dollar and Kraay (2002; 2004) conclude that growth is good for the poor by showing that average incomes of the poor move in tandem with average national incomes, and that trade via growth is good for the poor by showing that countries with bigger tariff cuts on average observed greater declines in poverty. Yet these findings remain a topic of heated academic debate, outlined in Deaton (2005) and Ravallion (2001).

Much of the academic debate on trade and poverty has focused on the static relationship between trade and poverty through changes in relative prices and wages of less educated workers. The stylized version of the Hecksher–Ohlin model predicts that trade should benefit the poor in a less developed country that is relatively well endowed with less educated labour and has a comparative advantage in unskilled labour-intensive goods. According to the Stolper–Samuelson theorem, trade liberalization will increase the return to the factor of production that is relatively abundant in a country and decrease the return to the scarce factor. Consequently, trade liberalization should reduce inequality between educated and less educated workers and lift the poor out of poverty in developing countries. In an influential project on trade and employment in developing countries, Krueger (1983) used this insight to argue in favour of outward-oriented trade regimes over import-substitution strategies adopted by many poor countries at that time.

Many developing countries liberalized trade during the 1980s and 1990s. As detailed micro-surveys of workers, households, and firms became more readily available, the researchers began to empirically examine the consequences of these reforms. Motivated by the academic debate on whether trade with poor countries in part contributed towards growing wage inequality in the OECD countries during the 1980s, these studies initially focused on the consequences of trade for wage inequality between more and less educated workers rather than for poverty. As reviewed in Wood (1999) and Goldberg and Pavcnik (2007), trade reforms were associated with increases (rather than the expected decreases) in inequality. Studies reviewed in Goldberg and Pavcnik (2007) concluded that the increases in the relative earnings of educated workers in poor countries could have been in part caused by an increase in the relative demand for educated workers in the aftermath of trade reforms due to trade-induced skill-biased technological change, outsourcing, or the higher skill intensity of exporting firms relative to non-exporters. But the studies remained silent on how trade reforms affected those at the bottom of the income distribution through economy-wide changes in the *absolute* demand for less skilled workers and through other channels ranging from consumption, industry wages, and unemployment to compliance with labour market standards. Only recently have studies focused on these issues.

Porto (2006) embeds heterogeneous households in a general equilibrium model of trade, where trade-induced relative price changes influence household welfare through changes in labour income (via the usual Hecksher–Ohlin mechanisms) and consumption. He explicitly accounts for the fact that households at the bottom of the welfare distribution are relatively less endowed with educated labour *and* spend a higher share of their household budget on basic items such as food than richer households. The model, combined with the estimates of key parameters from micro surveys, is used to simulate the effects of Mercosur-induced changes in prices on welfare and poverty in Argentina. Although poor

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Argentine households were worse off after Mercosur because they tended to consume a relatively high share of the now more expensive unskilled-labour-intensive goods, these negative consumption effects were substantially smaller than the trade-induced increases in labour income of less educated workers.
A series of papers in the *Globalization and Poverty* project organized by Harrison (2007) consider the link between trade and poverty in a setting where labour is not perfectly mobile across industries and/or regions within a country. These studies examine whether individuals living in areas or industries that were more exposed to trade experienced smaller or bigger changes in poverty than less exposed areas or industries. The Ricardo–Viner model with industry-specific labour would predict that industry tariff cuts are associated with declines in relative industry wages. Since tariff declines in many developing countries were concentrated in sectors with lower wages and higher shares of unskilled workers before the reforms, this channel could in the short run increase poverty in areas with a higher pre-reform concentration of protected industries relative to the national trend. A study by Topalova (2007) shows that individuals living in rural Indian districts, where pre-reform employment was heavily concentrated in industries that experienced larger declines in tariffs, suffered increases in poverty relative to the national trade of declining poverty during the 1990s. Topalova conjectures that these individuals fare relatively worse because immobility, in part stemming from inflexible labour regulations, precludes them from reallocating from sectors that have been hit hardest by declines in tariffs toward sectors or areas that benefit from export expansion. A related study by Hanson (2007) finds that the poor living in Mexican states with higher concentration of export-oriented industries and inflows of foreign direct investment (FDI) fared better during the 1990s than the poor in areas with lower export and FDI exposure.
Overall, the studies of trade and poverty based on micro evidence suggest that there is substantial heterogeneity in how trade affects the poor within developing countries. Future work on trade and poverty needs to further examine barriers that inhibit the less educated from moving away from industries, areas, or firms that have been hit hardest by tariffs cuts to industries and firms benefiting from new exporting opportunities.
An important component of the policy debate on the effect of trade on the world's poor is the link between trade and child labour. Many are concerned that, because less developed countries are assumed to specialize in exports of low-skill products, high-income countries foster child labour in low-income countries by raising the demand for the products intensive in unskilled labour. In fact, as discussed in Edmonds and Pavcnik (2005a), the link between trade and child labour is far more complicated and depends on how trade affects family incomes and poverty, the availability of substitutes or complements for the child's work, the returns to education, and consumption prices in addition to how trade affects the demand for child labour. Empirically, the association between trade and living standards seems to be the dominant factor in how trade affects child labour. Cross-country evidence in Edmonds and Pavcnik (2006) provides no support for the claim that trade perpetuates high levels of child labour in poor countries via the demand channel. Similarly, Edmonds and Pavcnik (2005b) show that child labour declined in Vietnam following the rice market liberalization that relaxed rice export quota and improved the standard of living of many net-rice producing households, even though the employment opportunities in the rice sector increased. There are many reasons for a connection between child labour and family incomes or poverty. Ranjan (2001) in particular emphasizes the relaxation of credit constraints as important for understanding why growing trade might be associated with declining child

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
labour despite rising demand for unskilled labour. In this vein, Edmonds, Pavcnik and Topalova (2007) argue that, in the Indian context, the child's economic contribution to the household through the avoidance of schooling costs is important in understanding the interconnections among trade policy changes, child labour and schooling.

See Also

- child labour
- Heckscher–Ohlin trade theory
- inequality (measurement)
- poverty
- poverty alleviation programmes
- trade costs

Bibliography

Deaton, A. 2005. Measuring poverty in growing world (or measuring growth in a poor world). *Review of Economic Statistics* 87, 1–19.

Dollar, D. and Kraay, A. 2002. Growth is good for the poor. *Journal of Economic Growth* 7, 195–225.

Dollar, D. and Kraay, A. 2004. Trade, growth and poverty. *Economic Journal* 114, F22–F49.

Edmonds, E. and Pavcnik, N. 2005a. Child labor in the global economy. *Journal of Economic Perspectives* 18(1), 199–220.

Edmonds, E. and Pavcnik, N. 2005b. The effect of trade liberalization on child labor. *Journal of International Economics* 65, 401–41.

Edmonds, E. and Pavcnik, N. 2006. International trade and child labor: cross-country evidence. *Journal of International Economics* 68, 115–40.

Edmonds, E., Pavcnik, N. and Topalova, P. 2007. Trade adjustment and human capital investments: evidence from Indian tariff reform. Working paper No. 2884. Cambridge, MA: NBER.

Frankel, J.A. and Romer, D. 1999. Does trade cause growth? *American Economic Review* 89, 279–399.

Goldberg, P. and Pavcnik, N. 2007. Distributional effects of globalization in developing countries. *Journal of Economic Literature* 45, 39–82.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hanson, G. 2007. Globalization, labor income, and poverty in Mexico. In Harrison (2007).

Harrison, A., ed. 2007. *Globalization and Poverty*. Chicago: University of Chicago Press and NBER.

Krueger, A. 1983. *Trade and Employment in Developing Countries, Vol. 3: Synthesis and Conclusions*. Chicago: University of Chicago Press.

Porto, G. 2006. Using survey data to assess the distributional effects of trade policy. *Journal of International Economics* 70, 140–60.

Ranjan, P. 2001. Credit constraints and the phenomenon of child labor. *Journal of Development Economics* 64, 81–102.

Ravallion, M. 2001. Growth, inequality, and poverty: looking beyond averages. *World Development* 29, 1803–15.

Rodriguez, F. and Rodrik, D. 2001. Trade policy and economic growth: a skeptic's guide to cross-national evidence. In *NBER Macro Annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press for NBER.

Topalova, P. 2007. Trade liberalization, poverty, and inequality: evidence from Indian districts. In Harrison (2007).

Wood, A. 1999. Openness and wage inequality in developing countries: the Latin American challenge to East Asian conventional wisdom. In *Market Integration, Regionalism and the Global Economy*, ed. R. Baldwin et al. Cambridge: Cambridge University Press.

How to cite this article

Pavcnik, Nina. "trade and poverty." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000243> doi:10.1057/9780230226203.1724

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

trade costs

Jeffrey H. Bergstrand
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Trade costs refer to the additional costs paid potentially by the final consumer of a good or service beyond the price at which a producer sells a good. In international trade, such costs may include the transport cost from origin to destination, taxes (or tariffs) imposed by importing nations' governments, the costs of infrastructure to facilitate trade, the costs of communications, and foreign exchange costs.

Keywords

communication costs; currency unions; distance; economic geography; exchange rate volatility; factor content of trade; foreign direct investment; foreign exchange costs; gravity equation; information costs; outsourcing; portfolio flows; regional trade agreements; rent seeking; sticky prices; tariffs; trade costs

Article

'Trade costs' refer to the costs above and beyond the 'mill price' that the final consumer of a good (or service) pays. If a product is sold by producer *i* to consumer *j* at (mill) price p_i (in dollars), the consumer pays p_i+T_{ij}, where T_{ij} denotes the 'trade costs' (also in dollars). Such costs may cover the opportunity costs of resources, but some trade costs are just rent seeking barriers. International trade provides a fertile ground for studying the various types and economic effects of trade costs. International trade flows travel across large distances and empirical estimates of the costs of transporting goods between two economic centres are available. However, such flows also face less obvious trade costs. A broad interpretation of trade costs includes – beyond transport costs – information-gathering costs for a consumer to locate a foreign producer, financial and legal costs of negotiating contracts, policy-related barriers, and costs of final distribution in the importing country. As discussed in Anderson and van Wincoop (2004), the total trade costs associated with exporting a good from producer *i* to consumer *j* may be an average *ad valorem* add on of 170 per cent to the (mill) price of a good. In this article, we discuss some of the different types of international trade costs (subsets of which form national and local

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
trade costs), and how such costs can influence the volume of trade between two nations and the relative prices of nations' goods.

Except for economic size, trade costs are probably the most important factor determining the volume of trade between a pair of countries (see Anderson, 1979; Bergstrand, 1985; and Anderson and van Wincoop, 2003). Trade costs play a critical role in understanding outsourcing, the factor content of trade, the field of economic geography, foreign direct investment and foreign affiliate sales of multinational enterprises, and the proliferation of regional trade agreements in the post-war period. Obstfeld and Rogoff (2000) argue that trade costs in goods markets provide the critical common element that also explains at least *six major puzzles* in international macroeconomics.

The trade cost that comes immediately to mind is the cost of transporting a good from producer *i* to consumer *j*. EXW ('ex works') refers to the price of a good at the point of origin; mill price is a synonym for the ex works price ('mill' refers to where the good was produced). FOB ('free on board') refers to the price of a good delivered to and put 'on board' an overseas vessel. CIF ('cost, insurance, freight') refers to the price of a good to a named overseas port, including insurance costs. Empirical researchers have used both FOB export data and CIF import data but prefer the latter because import data measured at customs points is more accurate.

A common measure of international transport costs is consequently the difference between the CIF and FOB values of a trade flow. The International Monetary Fund (IMF) provides data on average 'CIF/FOB factors' [100 × (CIF value–FOB value)/FOB value] for countries. Baier and Bergstrand (2001) report that average CIF/FOB factors for 16 Organisation for Economic Co-operation and Development (OECD) countries in 1958 and 1988 were 8.2 per cent and 4.3 per cent, respectively. David Hummels (2001) finds that freight rates vary dramatically across countries with average transport costs ranging from 3.8 per cent of EXW price p_i for the United States to 13.3 per cent for land-locked Paraguay in 1994, varying even more across commodities within countries. Hummels (1999) finds evidence that inflation-adjusted tramp shipping rates have declined between 40 and 70 per cent from 1950 to 1995, but also finds evidence suggesting ocean shipping rates have not declined. It is common to express transport costs on an *ad valorem* (or rate) basis. Hence, the price faced by consumer *j* for producer *i*'s product, p_{ij}, can be expressed as $p_{ij}=p_i\ (1+tc_{ij})$ where tc_{ij} is the (CIF – FOB)/FOB factor (for example, 0.04).

Another important trade cost is that associated with policy-related barriers imposed by national (or perhaps sub-national) governments. The trade cost most often envisioned here is a 'tariff', a tax imposed at customs points on imported goods. *Specific* tariffs are expressed in an amount of the home currency per unit imported good; T_{ij} used earlier denoted a specific trade cost. *Ad valorem* tariffs are expressed as a fraction of the value of the good; hence, an *ad valorem* tariff of ta_{ij} would cause the imported price of producer *i*'s product for consumer *j* to be $p_{ij}=p_i\ (1+ta_{ij})$ when the tariff is imposed on the FOB value and $p_{ij}=p_i\ (1+tc_{ij})(1+ta_{ij})$ when the tariff is imposed on the CIF value. Other entries in this dictionary address tariffs in more detail.

Transport costs and tariff barriers are arguably the easiest trade costs to measure *directly*. Because of difficulty in measuring other types of trade costs directly, empirical economists have turned to *indirect* methods to estimate trade costs. Indirect methods fall into two basic categories: inferring trade costs from differences in trade volumes between pairs of countries and inferring trade costs from differences in prices between pairs of countries.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

International trade economists have long used and increasingly applied the ‘gravity equation’ to explain empirically international trade flows (see Feenstra, 2004, ch. 5; gravity equation). The gravity equation typically explains bilateral trade flows between country pairs using cross-sectional or panel data on pairs’ gross domestic products (GDPs), bilateral distances between country pairs’ economic centres, and several other variables representing bilateral trade costs to infer the effects of such costs on members’ trade. Distance has long been a central variable explaining trade volumes, and typically has been interpreted as a measure of transport costs. However, the effect of distance probably measures more than transport costs, such as information costs. For instance, empirical work explaining bilateral foreign direct investment (FDI) flows also finds that distance has an economically significant effect on deterring such flows, even though theory suggests that measures of trade costs and FDI costs should have opposite effects on each others’ flows (see Markusen, 2002). Portes and Rey (2005) find empirically that distance also has a significant negative effect on portfolio flows, for which the transaction cost should be minimal. Other sources of trade costs (which consume resources) include infrastructure, communication and foreign exchange costs. Limao and Venables (2001) find that infrastructure has a significant effect on trade volumes, with a decline in the level of infrastructure investment from the median level to the 75th percentile equivalent to a 2,166-mile (3,466-km) increase in sea distance travelled. Tang (2006), using various measures of information technology, finds that communication costs have a significant effect on the volume of trade. Economists have long thought that exchange rate variability and its associated uncertainty should impose a significant trade cost and deterrent to trade. However, a survey of studies reveals that the trade-volume effects are probably small (see Cote, 1994).

With the use of the gravity equation in international trade, indirect estimates of the trade costs associated with national trade policies have proliferated. While a few empirical studies have looked at the effects of tariff rates explicitly on trade flows, the vast bulk have measured the presence or absence of (typically regional) economic integration agreements and currency unions on trade between country pairs. Many earlier studies using standard gravity equations found surprisingly small estimated effects on trade of arguably important trade agreements such as the Treaty of Rome (see Frankel, 1997). However, more recent studies incorporating modern theoretical foundations for the gravity equation and econometric techniques suggest that such small estimates are probably due to a bias introduced by self-selection of countries into such agreements (see Baier and Bergstrand, 2004; 2007). By contrast, work by Rose (2000) indicates that a currency union may have a very strong impact on trade between country pairs.

Estimates of trade costs have also been inferred indirectly using discrepancies in prices between countries. Engel and Rogers (1996) demonstrated that price variability of similar goods between US and Canadian cities is much greater than that between equidistant cities in the same country. Engel and Rogers (2001) showed that a ‘real barriers’ effect owing to incomplete market integration is present, but also that some dispersion could be explained by exchange rate variability and sticky price behaviour. Parsley and Wei (2001) showed that distance, unit-shipping costs, and exchange rate variability all contribute to dispersion of relative tradable goods prices across 96 cities in Japan and the United States. However, Crucini, Telmer and Zachariadis (2005) used price data from European cities to show that goods markets – at least, those in Europe – may be much more integrated than earlier work showed.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- currency unions
- factor content of trade
- foreign direct investment
- gravity equation
- international outsourcing
- regional and preferential trade agreements

Bibliography

Anderson, J.E. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69, 106–16.

Anderson, J.E. and van Wincoop, E. 2003. Gravity with gravitas: a solution to the border puzzle. *American Economic Review* 93, 170–92.

Anderson, J.E. and van Wincoop, E. 2004. Trade costs. *Journal of Economic Literature* 62, 691–751.

Baier, S.L. and Bergstrand, J.H. 2001. The growth of world trade. *Journal of International Economics* 53, 1–27.

Baier, S.L. and Bergstrand, J.H. 2004. Economic determinants of free trade agreements. *Journal of International Economics* 64, 29–64.

Baier, S.L. and Bergstrand, J.H. 2007. Do free trade agreements actually increase members' international trade? *Journal of International Economics* 71, 72–95.

Bergstrand, J.H. 1985. The gravity equation in international trade: some microeconomic foundations and empirical evidence. *Review of Economics and Statistics* 67, 474–81.

Cote, A. 1994. Exchange rate volatility and trade: a survey. Working Paper, No. 94–5, Bank of Canada.

Crucini, M., Telmer, C.I. and Zachariadis, M. 2005. Understanding European real exchange rates. *American Economic Review* 95, 724–38.

Engel, C. and Rogers, J.H. 1996. How wide is the border? *American Economic Review* 86, 1112–25.

Engel, C. and Rogers, J.H. 2001. Deviations from purchasing power parity: causes and welfare costs. *Journal of International Economics* 55, 29–58.

Feenstra, R.C. 2004. *Advanced International Trade: Theory and Evidence*. Princeton, NJ: Princeton

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
University Press.

Frankel, J.A. 1997. *Regional Trading Blocs*. Washington, DC: Institute for International Economics.

Hummels, D. 1999. Have international transportation costs declined? Unpublished manuscript, Purdue University.

Hummels, D. 2001. Toward a geography of trade costs. Unpublished manuscript, Purdue University.

Limao, N. and Venables, A.J. 2001. Infrastructure, geographical disadvantage, transport costs and trade. *World Bank Economic Review* 15, 451–79.

Markusen, J.R. 2002. *Multinational Firms and the Theory of International Trade*. Cambridge, MA: MIT Press.

Obstfeld, M. and Rogoff, K. 2000. The six major puzzles in international macroeconomics: is there a common cause? In *NBER Macroeconomics Annual*, ed. B.S. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Parsley, D.C. and Wei, S.-J. 2001. Explaining the border effect: the role of exchange rate variability, shipping costs, and geography. *Journal of International Economics* 55, 87–106.

Portes, R. and Rey, H. 2005. The determinants of cross-border equity flows. *Journal of International Economics* 65, 269–96.

Rose, A. 2000. One money, one market: estimating the effect of common currencies on trade. *Economic Policy* 30, 7–45.

Tang, L. 2006. Communication costs and trade of differentiated goods. *Review of International Economics* 14, 54–68.

How to cite this article

Bergstrand, Jeffrey H. "trade costs." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000244> doi:10.1057/9780230226203.1725

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

trade cycle

A. Medio
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycle; capital intensity; chaos; classical theory of the cycle; entrepreneurship; equilibrium; equilibrium business cycle theory; forced saving; Hawtrey, R. G.; Hayek, F. A. von; Hobson, J. A.; innovation; inventory investment; Juglar cycles; Kitchin cycles; Kitchin, J.; Kondratieff cycles; long run and short run; marginal efficiency of capital; monetary theory of the cycle; multiple equilibria; multiplier–accelerator theory; overinvestment theory of the cycle; population growth; stability of equilibrium; technical change; trade cycle; underconsumption theory of the cycle; uniqueness of equilibrium; voluntary saving; Walras's Law; white noise.

Article

The dynamics of capitalist economies are characterized by two facts: sustained growth of production and employment and wide oscillations of these magnitudes and the level of prices as well. This oscillatory behaviour of economic activity as a whole is indeed the subject of trade cycle theory. The use of the word ‘cycle’, besides pointing to alternation of ups and downs, also suggests the idea that oscillations are somewhat regular.

Concerning the occurrence and regularity of the cycle there exist two entirely different positions. According to one line of thought, it is possible to explain it exogenously. The economic system by itself would not display any tendency to fluctuate regularly but for the influence of external cyclical impulses such as the alternation of seasons. A more sophisticated version of this principle recognizes that external impulses do not even have to be cyclical in order to induce regular fluctuations of the system. This approach, which is very old and recurrent among economists, reflects an essentially pessimistic view about their ability to explain a prominent feature of modern economies.

The opposite opinion holds that the generation and the persistence of cycles are totally or mainly endogenous to the economic system. This is an idea rather difficult to argue cogently. One can fairly easily represent a dynamic system characterized by a succession of expansions and contractions. It is a much harder task to define rigorously a system whose oscillations persist indefinitely, independent of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
external impulses, with amplitude and frequency determined solely by the structural parameters. Only recently has the economics profession acquired the necessary mathematical tools to solve this problem satisfactorily.

A possible third alternative – which possesses some elements of the other two – is to postulate that oscillations are normally damped and would therefore eventually die out, were they not continually revived by erratic shocks that resupply the system with the energy needed to sustain the cyclical motion. We shall return to this idea later since its discussion merges into that of some recent developments in economic dynamics.

The very existence of specific (and measurable) cycles of different lengths has been the object of heated discussions among economists. Joseph Schumpeter, the author of a monumental historical investigation on business cycles (1939), detected three main types of cycles. The shortest ones, named after the economist Joseph Kitchin, would last approximately three years. The intermediate ones (the Juglar cycles) would comprise three Kitchins and last approximately ten years. Finally, the ‘long waves’ (Kondratieff) would reflect major technological innovations and extend over 50–60 years. Subsequent empirical research failed to find conclusive evidence of the actual existence of such regularities in the ups and downs of capitalist economies. And yet the recession of the 1970s has prompted a renewed interest in Schumpeter's work and one hears economists speak once again of the Kondratieff cycle.

Two considerations are here in point. First of all, the (courageous) formulations of hypotheses concerning the exact shape and duration of cycles and the subsequent discussions thereof provided fresh evidence as well as new and better tools of analysis, which greatly improved our knowledge of the subject, even when those hypotheses were eventually discarded. Secondly, since the 1930s economic agents’ improved understanding of fluctuations, and the vastly increased public intervention to counteract their most negative consequences, have certainly modified both the economic mechanisms that produce those fluctuations and people's expectations about them. On this point too we shall have to return.

The discussion of long waves naturally leads one to consider trend in relation to cycle. Indeed the rising phase of a very long cycle may not be distinguishable over a certain period of time from sustained growth. The presence of a trend itself raises a number of problems. First of all, we would like to understand the economic forces that determine it. In most analyses of growth and cycle one assumes that trend is determined mainly by such long-run factors as population growth and technical progress. The latter, however, are usually represented by given functions of time, which is tantamount to admitting that we do not know much about them.

On the other hand, the causal relation may well run in the opposite direction: that is to say, in an economy characterized by sustained expansion, both population growth and productivity might be stimulated by general prosperity, the latter being determined by other factors. In most economic systems there obviously exist ‘hereditary factors’ owing to which the ‘long run’ is resolved into a chain of ‘short run’ events. For instance, technical progress – at least that of the ‘learning by doing’ type – depends on the cumulative levels of production (or investment) over a more or less distant past. Moreover, both productivity and desired consumption display important ‘ratchet effects’, that is, they move more easily upwards than downwards. All these considerations suggest a dependence of trend on the cyclical path actually followed by the economy. However, a rigorous analysis of the process through which this influence is exerted is still lacking.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
There also exists an inverse causal relation running from trend to cycle; that is, the characteristics of a cycle, in particular its amplitude and duration, are markedly different according whether the economy is in a phase of stagnation or prosperity.

Unfortunately, the question of the relation between cycle and trend remains to this day one of the many obscure points of economic theory. Most models tend to overcome this difficulty by assuming that short run (cycle) and long run (trend) may be dealt with separately and then re-combined in a ‘cyclical growth model’. This procedure, however, requires rather formidable (and usually hidden) hypotheses; in particular, the relevant dynamic equations of the model must be linear, so that one may apply the principle of superposition of solutions of a dynamic model. The different authors’ explanations of the trade cycle are better investigated in the wider context of their economic theories as a whole and nothing close to a comprehensive survey may be attempted here. We shall only try to put the contemporary discussions into a historical prospective, without which it would be difficult to understand what the various theories state and why.

Broadly speaking, we can identify three main phases in the development of cycle theory.

The first phase – the classical one – comprises the works of economists who wrote in the 18th century and in the first half of the 19th century. These writers did not provide a true scientific explanation of the cycle, but addressed certain basic questions whose understanding would prove essential to subsequent developments. In particular, classical economists debated the question of the stability of economic systems with a view to establishing whether capitalist economies possess an inherent tendency to equilibrium, that is, whether they are able to generate and maintain prices and quantities consistent with one another as well as with the structural parameters of the system. This problem, formulated by Adam Smith in terms of the ‘invisible hand’, was discussed in the first half of the 19th century mainly in relation to the possibility (or probability) of general crises of overproduction. In this context, Say (1803), Ricardo (1817) and James Mill (1821) shared the opinion that production always creates its own demand and no general overproduction is therefore possible. Lauderdale (1804), Sismondi (1819) and especially Malthus (1820) dissented and pointed out that accumulation (saving) is not simply a redistribution of expenditure between consumption and investment goods. If incentive to invest is lacking or too weak, saving may well result in a general lack of expenditure and a consequent decline of production and employment.

The idea that capitalist economies have an intrinsic tendency to disequilibrium was also argued, in a very different context and with different overtones, by Karl Marx. Marx never produced a rigorous explanation of the cycle but his ideas provided inspiration to contemporary writers, in particular Kalecki and Goodwin. Marx, too, opposed the so-called Say's Law and pointed out that in a market economy, where purchases and sales are disjoint operations connected by the intermediation of money, discrepancies between demand and supply are always possible, not only in each individual sector but also in the economy as a whole. In Volume 1 of *Das Kapital* (1867), Marx came close to formulating a self-consistent model of the cycle. Accumulation of capital, Marx argues, by reducing the rate of unemployment (the ‘reserve industrial army’), pushes up wages (down profits), thus discouraging further investment. The ensuing recession brings about higher unemployment, leading to lower wages (higher profits) and therefore re-establishing the profitability of accumulation. This cyclical mechanism would be reformulated rigorously one century later by Richard Goodwin (1967) and others, to provide a ‘classical theory of the cycle’ based on the interaction of capital accumulation and distribution of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
income. A second phase in the theory of the cycle, which we can locate between 1850 and 1930, may be termed ‘modern’ and among its forerunners we shall mention Juglar (1862), Jevons (1884) and Tugan-Baranowsky (1901). In this phase the cycle as such became an object of specific investigation. Economists then attempted to define hypotheses and construct theories to explain the ‘true and fundamental’ causes of economic fluctuations.

Broadly speaking, we can distinguish three main ‘explanations’ of the cycle, namely: (i) a monetary theory; (ii) an overinvestment theory, in its turn subdivided into a monetary and a real variation; (iii) an underconsumption theory. We shall briefly present the main ideas of the leading exponents of each group of theories. Separate consideration will be given to Keynes's contribution, which marks a watershed between modern and contemporary cycle theories.

The monetary theory of the trade cycle has been most forcefully argued by Ralph Hawtrey (1919). According to this author, the rising phase of the cycle is caused by credit expansion, realized mainly through a reduction of the rate of interest. This induces inventory accumulation by dealers, whose increased demand in turn stimulates producers’ expenditure. The rise in demand for investment and consumption goods brings about a undesired *reduction* of the inventory–sales ratio to which dealers respond with further accumulation. A self-sustaining expansionary process ensues – possibly reinforced by secondary speculative waves – and will continue as long as monetary expansion goes on.

However, Hawtrey argues, insofar as the monetary system is constrained by a link between global liquidity and a real asset whose quantity is limited, monetary expansion must come to an end. Monetary flows, moreover, tend to generate fluctuations of real variables, owing to the lagged response of the demand for money to changes in income. Thus, in the ascending phase of the cycle, demand for money does not grow *pari passu* with expenditure and consumers’ income. Consequently the banking system experiences net monetary inflows, which permit it to pursue an expansionary policy. However, as soon as demand for money catches up with income, banks’ liquidity deteriorates, forcing them sooner or later to increase the rate of interest and squeeze credit. This initiates a downwards cumulative process. In a recession the lagged response of the demand for money will again act as a brake, slowing down the decline of the real output and employment. Eventually excess liquidity will reappear, the rate of interest will be reduced and the cycle will start anew.

The overinvestment theory of the cycle in its monetary version is perhaps best represented by Hayek's so-called ‘concertina effect’, as described in his book *Prices and Production* (1931).

Two ideas here play a crucial role. The first one is a typically neo-Austrian proposition stating that capital intensity (assumed to be unambiguously measurable) is an inverse function of the rate of interest. The second idea is based on the distinction between voluntary and forced saving. An increase in voluntary saving is accompanied by a reduction in the rate of interest and by an increase in capital intensity. While these adjustments take place, the system remains in equilibrium and no fluctuations arise. On the contrary, when saving is forced by an excessive credit expansion, investment is no longer constrained by *ex ante* saving. Owing to the unduly low rate of interest, the production process becomes too ‘indirect’ that is, there will be an excessive development of those stages of production which are more removed from the final, consumption stage.

In sum, according to Hayek's theory, the structure of demand and in particular the distribution between investment and consumption goods is determined by the propensity to save, whereas the structure of production is a function of the rate of interest. In equilibrium the rate of interest is fixed so as to make those two structures consistent. Excess credit causes an overproduction of capital goods, which must

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
eventually manifest itself through increased profitability in the consumption good sector and corresponding losses in the investment sector. The latter will therefore experience a crisis that will turn the boom into a recession.

It is interesting to observe that in both Hawtrey's and Hayek's theories the culprit for the recession is the banking system, which keeps the rate of interest too low during expansion. However, for Hayek this leads to an undue ‘lengthening’ of the production process, bound to be reversed when the supply of money ceases to grow at an excessive rate. For Hawtrey, a low rate of interest provokes an excessive rise in global demand vis-à-vis the available stock of money.

A real version of the overinvestment theory of the cycle was put forward by Wicksell (1907) and Schumpeter, whose ideas on this subject are best summarized in the already quoted treatise (1939). These authors shared the view that growth and cycle are intrinsically related, and the main theoretical problem in this context is to explain why economic expansion does not take place smoothly, ‘as trees grow’, but why it occurs in leaps and bounds. Wicksell and Schumpeter also agreed that the oscillatory behaviour of capitalist economies is related to the process of innovation through which the employment of limited quantities of primary factors (labour and land) results in increasingly greater amounts of consumption and investment goods. Essentially the trade cycle depends on the fact that innovations, that is, the introduction of new techniques, new products and new markets, are not distributed uniformly in time, but take place in a discontinuous manner, in groups or ‘swarms’.

Schumpeter especially emphasized the role of entrepreneurial activity in the generation of business cycles. Since innovation implies a break of routine, it cannot occur smoothly but requires a minimum critical amount of energy to overcome inertia. Such energy is provided by exceptional individuals who possess the courage, strength and imagination necessary to ‘do things differently’. Once these few pioneers have opened the way, many others will follow to share in the extra-defeating process, made available by innovation. This is a self-defeating process, though. Insofar as the new methods products have been absorbed by the market, prices and profits will fall, terminating the boom and reversing the direction of the cycle. Secondary waves many amplify the oscillations, bringing about overoptimistic booms followed by severe slumps.

For Schumpeter and Wicksell alike, however, not everything about recession is bad, provided the worst consequences of deep depressions can be avoided. Recession is indeed a phase of adjustment during which innovations are ‘digested’ and leads the economy to an intrinsically superior stage. Recession, so to speak, realizes what the boom had promised: a permanently increased flow of commodities, reduced costs, entrepreneurial profits transformed into higher incomes of the other social classes. Schumpeter insisted upon the Darwinian function of (normal) recessions, during which ‘lame ducks’ are eliminated and only the stronger and more efficient survive. Similar arguments would be employed in the 1980s on both sides of the Atlantic Ocean to justify severe anti-inflationary policies.

The best-known exponent of underconsumption theories of the cycle was undoubtedly Hobson (1922). His theory of investment did not differ much from that of Wicksell and Schumpeter and he shared their view that investment opportunities are basically determined by the needs of development, which are in turn governed by technical changes and population growth.

According to Hobson, though, depressions are caused by insufficient expenditure, which in capitalist economies arises from a skew distribution of income. Increases in income are accompanied by more than proportional increases in saving, leading to overinvestment first and overproduction later. Hobson did not believe in the efficacy of the traditional remedies to overproduction, namely a reduction in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
rate of interest and in the level of prices. As concerns the former, he was of the opinion that saving responded little, if at all, to changes in the interest rate. On the other hand, changes in prices are too sluggish to counteract undesired changes in real variables. Instead, actual economies eliminate excess saving in the most inefficient and painful way, that is, through depression and unemployment.
Hobson's arguments are clear anticipation of certain Keynesian ideas which would become very popular a few decades later.
It may be noticed that both the overinvestment and the underconsumption theories locate the origin of the crises in a ‘vertical’ imbalance between the investment and consumption sectors, an idea which had already been discussed by Marx, Rosa Luxemburg and Tugan-Baranowsky. Those theories differ from one another concerning which of the two sectors is overexpanded, the investment sector according to the former, the consumption sector to the latter. Both of them, moreover, could be comprised under the more general label of overcapitalization theories, but the one (underconsumption) argues that there are too many capital goods (and consequently too many consumption goods) vis-à-vis global demand, the other (overinvestment) maintains that there is excess accumulation vis-à-vis saving. It follows that policy recommendations suggested by the supporters of these theories conflict: a reduction of consumption according to overinvestment; a redistribution of income leading to greater consumption according to underconsumption theory.
Keynes's own theory of the cycle – as distinguished from Keynesian theories – constitutes a *trait d'union* between the modern and the contemporary phase. Even if all the elements necessary to build a true model of the cycle exist in Keynes's work, this task was left to his followers. We shall describe here the essence of Keynes's argument as it appears in the *General Theory* (1936).
At an abstract level, Keynes was fully aware that cyclical motion must result from alternations in the relative strength of expansive and contractive forces, to wit that the cycle must be nonlinear. The changes in the balance between expansive and contractive forces may take place smoothly or abruptly, the latter case being more frequently in a boom, the former in a slump. These ideas would be further developed by younger economists inspired by Keynes, in particular by Kaldor and Goodwin.
For Keynes the trade cycle is a complex mechanism but its most important manifestations are the fluctuations of the marginal efficiency of capital, which is determined by psychological as well as economic considerations, and primarily depends on abundance or scarcity of capital goods, their cost and expectations about their future returns. Most often a recession is precipitated by a turnaround of expectations, which, to Keynes, is a much more decisive factor than the increase in the rate of interest sometimes associated with it.
In the last part of the boom, entrepreneurs’ optimism offsets all the other unfavourable circumstances: excess capital goods, increasing costs and a high rate of interest. Estimates of future returns to assets are distorted and exaggerated by ignorance, speculation, and vested interests of financial intermediaries and asset-owners. When finally the overoptimistic forecasts are falsified by facts, there follow excessive and even catastrophic readjustments. The increase in liquidity preference that usually accompanies the decrease in the marginal efficiency of capital aggravates the situation and nullifies the effects of expansive monetary measures.
At this point, Keynes contrasts his own theory with that based on overinvestment and observes that the latter is an ambiguous term. If overinvestment means that capital goods in general are so abundant that no investment project can be found whose expected return could justify the cost, then, according to Keynes, this is a rare occurrence even at the peak of a boom. Instead – Keynes continues – investment is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
excessive either in the sense that actual returns are lower than those on the expectations of which the decision to invest has been taken, or in the sense that severe unemployment makes investment superfluous. It follows that in a recession the right remedy is to reduce, not to increase, the rate of interest.

On the other hand, Keynes thought that the essential propositions of the underconsumption theory were substantially correct. Under capitalist rules of the game, the volume of investment is not controlled, let alone planned, and depends on the vagaries of the marginal efficiency of capital, with a rate of interest systematically kept above a conventional minimum. In this situation, in order to maintain high rates of employment it may well be necessary to stimulate consumption. Keynes's only criticism of the underconsumption theory is its neglect of the possibility of stimulating investment directly. But this – according to Keynes – is a matter of expediency rather than theory.

The modern phase of the cycle theory was characterized by a rich theoretical investigation which led to a deeper understanding of the dynamics of capitalist economies. It did not result, however, in the production of true models, that is, sets of rigorously formulated propositions containing the (necessary and) sufficient conditions to represent in idealized form the cyclical behaviour of the economy. The ‘contemporary’ phase of cycle theory – from the 1930s onwards – does not consist so much in the search for new explanations of the ‘enigma of the cycle’ (Wicksell, 1907) as in the attempt to analyse in a rigorous manner concepts and ideas put forward by earlier writers in an intuitive form.

The foundations of the mathematical theory of the trade cycle were laid down mostly in the 1930s, and the early works in this field (Frisch, 1933; Tinbergen, 1959: Kalecki, 1935; 1943; and Samuelson, 1939) were characterized by a distinctly Keynesian flavour. Keynes's own writings – and of course even more so the world crisis – had convinced a growing number of economists that the free functioning of the market was more likely to lead to an oscillatory behaviour of income, employment and prices, than to full employment equilibrium.

Multiplier–accelerator theory, in different versions and with different refinements, rapidly became predominant in the decades immediately before and after the Second World War.

In the 1960s, however, the economics profession's attention turned away from the problem of the cycle. Several reasons contributed to this change, but four of them seem to stand out as crucial.

First of all, economic events after the Second World War seemed to suggest that business cycles had become obsolete. After a phase of settlement following the turmoil of the war, the world economy (or at least that of the major industrialized countries) seemed to have entered an epoch of sustained growth without (or with only minor) fluctuations. Economists, and especially the younger ones, were quick to respond to the current social and political mood.

Secondly, the prevailing models of the cycle, those of Keynesian inspiration, suffered from a major drawback. The solution of such models consists of fluctuations, perfectly regular in the sense that a certain periodic orbit, once established, repeats itself over and over again.

In such a situation, economic agents would sooner or later notice the periodic character of the dynamics of the system and learn to calculate the amplitude and frequency of the cycles. This in turn would lead to a revision of their expectations. The behavioural hypotheses of the model – on which the cyclical motion of the system depends – would no longer be tenable and the model itself would have to be reformulated. Incidentally, this criticism of the deterministic models of the cycle is perhaps the most important element of truth in the theory of rational expectations.

Thirdly, historical data on the main economic variables do not confirm the periodic regularity predicted

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
by the model.
Finally, and partly in response to the theoretical difficulties mentioned above, there has recently been a revival of what we would call the ‘static prejudice’ in economics, whose most explicit expression is perhaps ‘equilibrium business cycle theory’. In a nutshell, this theory describes the working of the economy by means of a model whose deterministic part is characterized by a unique, stable equilibrium. A stochastic part is added to it which depends on imperfect information of economic agents, whose decisions are therefore mistaken. With rational expectations these errors are ‘white noise’, that is, they are normally distributed above and below the optimal (equilibrium) values. The resulting fluctuations of the system are non-periodic but bounded and their amplitude and frequency can be estimated statistically. This approach is reminiscent of certain early ideas first developed in the 1920s and 1930s by Slutsky (1937), Hotelling (1927), Yule (1927), Frisch (1933) and later by Kalecki (1954), to explain the persistence of the cycle. However – as Hicks commented long ago (1950) – when the random factors ‘explain’ a substantial part of the deviations of a system from its equilibrium position, the proposed theory amounts to a confession of ignorance.
Mentioning a ‘static prejudice’ in economics raises a few methodological questions, discussion of which is perhaps the best way to introduce the recent developments in the theory of the cycle (and, more generally in economic dynamics) and to conclude this essay.
In spite of the great practical importance of the problems discussed and the high intellectual quality of the results obtained, cycle theory has generally been regarded as an interesting but as a whole marginal branch of economics, whose exponents as such rarely managed to hold centre stage in professional debates.
The great theoretical systems developed in the 18th and 19th centuries, for example, those of Ricardo and of Walras, were more suited to the explanation of interdependence of economic variables, rather than their evolution in time. This problem is better formulated by means of a system of algebraic equations, whose solution (not necessarily unique) is a set of values of the variables – typically prices of commodities and quantities produced – which is consistent with the postulated relationships and the exogenously fixed parameters. The equations themselves define certain equilibrium conditions, for example, equality of demand and supply, or uniformity of the rate of profits. Uniqueness and stability of equilibrium are deemed to be desirable properties of the model, in a descriptive as well as in normative sense. Indeed only stable systems are observable. Multiple equilibria, on the other hand, are not satisfactory for two reasons. First of all, in the general case they are alternatively stable and unstable; secondly, the multiplicity of equilibria introduces a certain amount of relativism as far as their optimality properties are concerned.
Discussion of stability must bring (and historically has brought) dynamic considerations into the picture. Early writers provided intuitive, but not very cogent stories arguing that the system ‘tended to’ or ‘gravitated towards the natural or equilibrium values and prices’. Walras took a step forward, providing a principle (Walras's Law) which shows that, under the postulated maximizing behaviour, disequilibria of notional demands and supplies cannot be entirely haphazardous, but most take place according to certain rules (that is, the value of total excess demand must be equal to zero). This led to rather optimistic considerations as far as stability of competitive equilibrium was concerned, but was a far cry from a rigorous statement of the problem.
A deeper understanding of the intricacies of non-equilibrium behaviour of dynamical systems and more powerful mathematical techniques were required. Fundamental progress along this line was made in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
interwar period (Hicks, 1939) and more so during and after the Second World War, when certain necessary mathematical results became available to economists. (The list of important contributions is too long to be quoted exhaustively here: we shall only mention Samuelson, 1941; 1942; Metzler, 1945; Morishima, 1952; Arrow and Hurwicz, 1958.)

The conclusions were rather dismaying though, since a rigorous analysis showed that sufficient stability conditions entailed the most heroic assumptions on the specifications of the system.

Economists' overwhelming preoccupation with equilibrium explains their neglect of cycle theory. Indeed cycle theory must by definition concern itself with off-equilibrium states of the system.

Empirical observation suggests that the economy is not normally in equilibrium, but fluctuates without ever coming to rest or, if we except relatively rare historical occurrences, without 'exploding'. Is this restlessness only the result of random external shocks, or is it a structural characteristic of the system, owing to the operation of endogenous mechanisms? In dealing with this problem the cycle theorist is naturally led to posing typically dynamic questions such as 'what are economic agents' reactions to non-equilibrium, and therefore unsatisfactory situations?'; or, more generally, 'what laws of motion govern the system, starting from a given, generally off-equilibrium state?'

An exact formulation of this problem required an even more sophisticated analytical apparatus than was the case with static theory. Its natural mathematical set-up is a system of differential (or difference) equations, whose solution is a set of functions of time which, given the initial conditions, describes, so to speak, the history of the variables under consideration. The fact that the relevant part of theory of differential equations became available to economists at a rather late date is perhaps a further explanation of their preference for static rather than dynamic problems.

In discussing the problem of fluctuations, economists must soon run up against the question of linearity. Most systems, as described by economists' models, are linear; that is, the coefficients that appear in those models do not depend on the variables or their derivatives with respect to time. The crucial importance of this assumption can be appreciated by considering how it affects the analysis of stability. From the point of view of linear analysis, equilibrium is either stable or unstable. A disturbance is therefore followed by a return to rest in the former case, or by an indefinite increase in the magnitude of the deviation in the latter. But this is a very rough description of the behaviour of an economic system. Linear analysis, by its very nature, completely disregards the effect of the size of disturbances. Indeed an economy may be stable with respect to small deviations from equilibrium, but not so when those deviations are large. On the other hand, an economy may be locally unstable, in the sense than it does not show any tendency to return to its equilibrium position when subjected to small shocks, but it may possess self-correcting mechanisms which only operate far from equilibrium. In either case, the conclusions reached by means of a linearization around the equilibrium point would be misleading.

The linear approximation is unsatisfactory not only because it provides a simplified and therefore distorted picture of reality. A much more fundamental shortcoming is that there exist certain phenomena (economic or otherwise) which cannot be idealized at all by means of linear models. In particular, it is well known that a system of linear differential (or difference) equations, if structurally stable, cannot describe sustained oscillations, that is, oscillations that do not expire or explode.

The discovery that certain basic questions of economic dynamics, in particular persistent cycles, could not be tackled effectively by means of linear models led an increasing number of economists to make use of nonlinear methods of analysis. In so doing, they experienced what Poincaré – to whom modern dynamic analysis owes more than to anybody else – had described several decades earlier. The study of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the cycle had proved precious to the analyst, all the more so since it constituted a first necessary step into the mysterious and hitherto inaccessible realm of non-equilibrium dynamics.

At that point the phrase ‘cycle theory’ became an elliptical expression designating not a particular problem, but economic dynamics as a whole or, more appropriately, a method for investigating dynamic processes in economics. But from the analysis of fluctuations economists have obtained much more than a number of new methods and mathematical tools. The most aware of them have undergone a true cultural revolution that Frisch already in the early 1930s (1933) thought would be as important to economics as the transition from classical to quantum mechanics had been to physics.

Equilibrium states, stable and unstable and even limit cycles have now been revealed as rather special configurations in a much more complex and morphologically rich theoretical universe. As soon as the linearity assumption has been dropped, even a simple model may exhibit a very complicated behaviour. The objection may be raised that this is very nice, but what does it have to do with real systems, in our case real economies? On the contrary, recent developments in dynamic theory have finally provided economists with tools of analysis such that theoretical results are, at least qualitatively, comparable with empirical observations. Take, for example, so-called chaotic behaviour. This characterizes a large class of dynamic models, economic or otherwise and, in common parlance, describes irregular fluctuations of a type that has so far been exclusively associated with random (and therefore essentially unexplained) disturbances. After all, what could be more realistic than irregular but bounded fluctuations of income, employment and prices, and what more unrealistic than a stable unique equilibrium point?

Clearly cycle theory and, more generally, economic dynamics is in a state of transition. To produce theoretically meaningful and socially relevant models of real systems, economists must fulfil two main requirements:

1. (i) to define mechanisms of adjustment that realistically describe economic agents’ behaviour in disequilibrium;
2. (ii) to employ techniques of analysis suitable to study the dynamic systems resulting from the operation of those mechanisms.

Recent developments have contributed substantially to solve some of the problems in (i). Instead, off-equilibrium behaviour of economic agents remains to this day a rather fuzzy area of economic investigation.

See Also

- aggregate demand theory

Bibliography

Arrow, K.J. and Hurwicz, L. 1958. On the stability of the competitive equilibrium. *Econometrica* 26, 522–52.

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Essays in Honour of Gustav Cassel. London: G. Allen & Unwin.

Goodwin, R.M. 1950. A non-linear theory of the cycle. *Review of Economics and Statistics* 32, 316–20.

Goodwin, R.M. 1951. The non-linear accelerator and the persistence of business cycles. *Econometrica* 19, 1–17.

Goodwin, R.M. 1953. Econometrics in business cycle analysis. In *Readings in Business Cycles and National Income*, ed. R.V. Clemence and A.H. Hansen. New York: Norton.

Goodwin, R.M. 1955. A model of cyclical growth. In *The Business Cycle in the Post-War World*, ed. E. Lundberg. London: Macmillan.

Goodwin, R.M. 1967. A growth cycle. In *Socialism, Capitalism and Economic Growth*, ed. C.H. Feinstein. Cambridge: Cambridge University Press.

Hawtrey, R.G. 1919. *Currency and Credit*. London: Longmans.

Hayek, F. 1931. *Prices and Production*. London: G. Routledge & Sons.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Hicks, J.R. 1950. *A Contribution to the Theory of the Trade Cycle*. Oxford: Clarendon Press.

Hobson, J.A. 1922. *The Economics of Unemployment*. London: G. Allen & Unwin.

Hotelling, H. 1927. Differential equations subject to error, and population estimates. *Journal of the American Statistical Association* 22, 283–314.

Jevons, W.S. 1884. *Investigations in Currency and Finance*. London: Macmillan.

Juglar, C. 1862. *Des crises commerciales, et leur retour périodique en France, en Angleterre, et aux Etats-Unis*. Paris: Guillaumin.

Kaldor, N. 1940. A model of the trade cycle. *Economic Journal* 50, 78–92.

Kalecki, M. 1935. A macrodynamic theory of business cycles. *Econometrica* 3, 327–44.

Kalecki, M. 1943. *Studies in Economic Dynamics*. London: G. Allen & Unwin.

Kalecki, M. 1954. *Theory of Economic Dynamics*. London: G. Allen & Unwin.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kalecki, M. 1968. Trend and business cycle reconsidered. *Economic Journal* 78, June, 263–76; corrigendum, September, 729.

Kalecki, M. 1971. *Selected Essays on the Dynamics of the Capitalist Economy* 1933–1970. Cambridge: Cambridge University Press.

Lauderdale, J.M. 1804. *An Inquiry into the Nature and Origin of Public Wealth and into the Means and Causes of its Increase*. Edinburgh: A. Constable & Co. London: Hurst, Robinson & Co.

Lucas, R. 1981. *Studies in Business Cycle Theory*. Oxford: Basil Blackwell.

Marx, K.H. 1867. *Capital*, vol. 1. New York: Penguin Books, 1976.

Metzler, L.A. 1945. The stability of multiple markets: the Hicks conditions. *Econometrica* 13, 277–92.

Malthus, T.R. 1820. *Principles of Political Economy*. London: John Murray.

Mill, J. 1821. *Elements of Political Economy*. London: Baldwin, Cradock & Joy.

Morishima, M. 1952. On the laws of change of price-system in an economy which contains complementary commodities. *Osaka Economic Papers* 1, May, 101–13.

Ricardo, D. 1817. *On the Principles of Political Economy and Taxation*. London: John Murray.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics* 21, May, 75–8.

Samuelson, P.A. 1941–1942. The stability of equilibrium. Pt. I: comparative statics and dynamics; Pt. II: linear and non-linear systems. *Econometrica* 9 (April 1941), 97–120; 10 (January 1942), 1–25.

Say, J.B. 1803. *Traité d'économie politique*. Paris: Déterville.

Schumpeter, J.A. 1939. *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process*. New York and London: McGraw-Hill.

Sismondi, J.C.L. 1819. *Nouveaux principes d'économie politique*. Paris: Delaunay.

Slutsky, E. 1937. The summation of random causes as the source of cyclical processes. *Econometrica* 5, April, 105–46.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Tinbergen, J. 1959. *Selected Papers*, ed. L.H. Klaassen. Amsterdam: North-Holland.

Tugan-Baranowsky, M. 1901. *Studien zur Theorie und Geschichte der Handelskrisen in England*. Jena.

Wicksell, K. 1907. Krisornas gata. *Statskonomisk Tidsskrift* 21, 255–84. Trans. C.G. Uhr as ‘The Enigma of Business Cycles’, *International Economic Papers* No. 3 (1953), 58–74.

Yule, G.U. 1927. On a method of investigating periodicities in disturbed series. *Philosophical Transactions of the Royal Society of London* Series A 226, April, 267–98.

How to cite this article

Medio, A. "trade cycle." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000090> doi:10.1057/9780230226203.1726

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

trade policy, political economy of

Devashish Mitra
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This area of research tries, through the introduction of politics in economic models, to explain the existence and the extent of anti-trade bias in trade policy. The two main approaches, namely, the median-voter approach and the special-interest approach are surveyed. Certain applications of these approaches to policy issues, such as trade agreements, the issue of reciprocity versus unilateralism in trade policy, regionalism versus multilateralism, hysteresis in trade policy and the choice of policy instruments, are discussed. Finally, the empirical literature on the political economy of trade policy is surveyed. The new literature that employs a more ‘structural’ approach is emphasized.

Keywords

congestion problem; Cournot oligopoly; customs unions; deadweight loss; free trade; free trade areas; free-rider problem; Heckscher–Ohlin trade theory; hysteresis in trade policy; intermediate goods; international trade; lobby formation; lobbying; majority rule; median voter model; monopolistic competition; multilateralism in trade policy; non-tariff barriers; optimal obfuscation principle; political competition; political economy; political support function; political-contributions model; progressive and regressive taxation; proportional representation; protection; reciprocity in trade policy; regional and preferential trade agreements; risk sharing; single-peaked preferences; special-interest politics; specific-factors trade theory; tariff-formation function; tariffs; tariffs vs. subsidies; terms of trade; trade agreements; trade policy, political economy of; trade-diverting vs. trade-creating bilateral agreements; unilateralism in trade policy; World Trade Organization

Article

While economists clearly understand the benefits of free trade, they have always found it difficult to explain departures from it in the real world. Most of these departures are in the direction of limiting the volume of trade. In trying to explain the existence and the extent of this anti-trade bias in trade policy,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
trade economists have introduced politics in their economic models. Parts of this political-economy literature have also tried to explain why policy instruments that are more efficient than trade policy and can achieve the same political and economic objectives are not often used. An important empirical contribution of the political-economy literature has been to uncover the main determinants of cross-country and cross-industry variations in protection.

Modelling approaches

Median-voter approach

Political economy models of trade are of two main types. One of them adopts the majority voting approach. Such models are called ‘median voter’ models in the literature. Preferences are assumed to be ‘single peaked’ and conditions are imposed such that the most preferred policy of each individual is monotonic in a certain characteristic. Then, with other individual characteristics held constant across the population, the tariff chosen under two-candidate electoral competition is the median voter's most preferred tariff. The median voter here is the median individual in the economy when ranked according to the characteristic under consideration. Mayer (1984) applies this median-voter principle to the Heckscher–Ohlin and specific-factors trade models. In the Heckscher–Ohlin case, the political economy equilibrium tariff is the most-preferred tariff of the median individual in the economy-wide ranking of the ratio of capital to labour ownership. If this median individual's capital to labour ratio is less than the economy's overall capital to labour ratio – that is, if the asset distribution in the economy is unequal – the equilibrium trade policy is different from free trade and is one that redistributes income from capital to labour – pro-trade in a labour-abundant economy and anti-trade in a capital-abundant economy.

Special-interest politics

The other type of political economy model in the trade literature focuses on ‘special-interest’ politics. The first papers to model lobbying explicitly in the trade arena were by Findlay and Wellisz (1982) and Feenstra and Bhagwati (1982). The Findlay–Wellisz model is a two-sector model in which production in each sector is carried out using a factor of production specific to that sector – land for food production and capital for manufactures – and an economy-wide general factor, namely, labour. Both types of specific factor owners are fully organized politically and they lobby against each other. This simple model shows the existence of an equilibrium tariff determined through the Nash interaction between the two opposing groups. The government is modelled very indirectly through a tariff formation function which is increasing in the amount of labour devoted to lobbying by the import-competing specific factor and decreasing in labour used in lobbying by the specific-factor owners in the export sector.
While only labour is used as an input in lobbying in the Findlay–Wellisz model, both capital and labour are used as inputs into lobbying in the Feenstra–Bhagwati model. However, only one sector is assumed to be politically active in the model. Unlike in the Findlay–Wellisz model, the government in the Feenstra–Bhagwati model is not a monolithic entity but has a two-layered structure. While one layer is a clearing house for lobbies, the other cares about social welfare. The tariff is determined through an interaction between the two layers.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Another approach to modelling ‘special-interest’ trade politics is the ‘political support function’ approach pioneered by Hillman (1989). (Some of the classification terminology here is borrowed from Rodrik, 1995, to which the interested reader is referred for a detailed typology of political-economy models. See also Helpman, 2002, for an analytical survey within a unified framework.) Under such an approach, the government's objective function, also called the political support function, incorporates its preferential treatment of each organized industry as well as the cost of protecting this industry given by the excess burden on society. Van Long and Vousden (1991) use a specific form of Hillman's political support function which is linear in the welfare levels of different types of specific-factor owners, with different weights being assigned to different factors.

Magee, Brock and Young (1989) explicitly model electoral competition. They use a two-sector, two-factor Heckscher–Ohlin set-up with two political parties – one pro-trade and another pro-protection – and two lobbies – one representing capital and the other labour. Lobbies contribute to their respective favoured political parties to maximize their chances of winning elections. Policy platforms here are chosen prior to decisions on campaign contributions.

The special-interest approach has evolved from the simple Findlay–Wellisz ‘tariff-formation function’ approach to the state-of-the-art Grossman and Helpman (1994) ‘political-contributions’ model. The latter is path-breaking for several reasons. First, it is multi-sectoral. Second, it provides micro-foundations for the behaviour of organized lobbies and politicians. A ‘menu-auctions’ approach is used in modelling policy bidding by interest groups. Multiple principals, namely, the various organized lobbies, try to influence the common agent, namely the government. The government's objective function is a weighted sum of political contributions and aggregate welfare, while each lobby maximizes its welfare net of political contributions. Most importantly, especially from the empirical angle, the level of protection for each industry is derived as an estimable function of industry characteristics and other political and economic factors. Protection to organized sectors is negatively related to import penetration and the (absolute value of) import demand elasticity, while protection to unorganized sectors is positively related to these two variables. With everything else held constant, organized sectors are granted higher protection than unorganized sectors.

While Grossman and Helpman in their models take the existence of organized lobbies as given, Mitra (1999) extends their framework to endogenize lobby formation. He shows that we are closest to free trade when the government cares too little or too much about aggregate welfare relative to political contributions. While the former leads to the formation of a large number of mutually opposing lobbies, the latter situation is one where hardly any lobbies get formed. Mitra also shows that a higher concentration of asset ownership in the economy leads to the formation of a larger number of organized lobbies representing sectors that are heavily protected. Magee (2002) analyses a single lobby's organization problem in the context of the collection of political contributions in a repeated game setting. (See also Pecorino, 1998, for an analysis of the same issue with a tariff-formation function approach.)

Theoretical applications

Trade agreements

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The first important theoretical application we discuss is the issue of trade agreements. Using their political-contributions approach, Grossman and Helpman (1995a) analyse trade policy in a setting with two large countries, where they show an additional terms-of-trade component in the tariff expression in a non-cooperative setting. This component gets eliminated in a cooperative setting of international trade negotiations, and the relative size of protection in any sector in the two countries then depends on the relative political power of the same industry in the two countries. Thus there is a rationale for 'trade talks' as opposed to 'trade wars'. Using what is very close to a 'political-support' function approach, Bagwell and Staiger (1996; 1999) show that, even when political economy considerations are taken into account, the only rationale for (reciprocal) trade agreements is the elimination of terms-of-trade externalities. They use this approach to develop a rationale for the General Agreement on Tariffs and Trade (GATT)/World Trade Organization (WTO) and its different rules. (See Bagwell and Staiger, 2002, for an in-depth discussion.)

The next natural question then is whether free trade agreements are of any value to countries whose actions have no impact on the international terms of trade. Maggi and Rodriguez-Clare (1998) have a political economy explanation for the unilateral commitment to free trade agreements by small countries. Their setting is one in which owners of capital first decide in which sector to invest, and then those who invest in a particular sector (the import-competing sector) lobby the government for protection. The lobbying is modelled as a Nash bargaining game between the lobby and the government. While the lobby at least compensates the government for the deadweight losses generated in the second stage, it may not compensate the government for the welfare loss through the inter-sectoral misallocation of capital in the first stage in the expectation of protection in the second stage. In such a situation, it is possible that a government will commit to a free trade agreement in a prior stage 'zero'.

Mitra (2002) builds on the Maggi–Rodriguez-Clare version of the Grossman–Helpman framework, augmenting it with the decision to incur fixed costs (to build relationships with politicians in power and/or to form a lobby) prior to the actual lobbying, but, importantly, not providing room for any capital mobility. However, the main result of the Maggi–Rodriguez-Clare model goes through even in this newly modified set-up. This is the result that generally governments with low bargaining power with respect to domestic lobbies are the ones that precommit to free trade agreements.

Grossman and Helpman (1995b) have provided a detailed analysis of political-economy factors responsible for the emergence of free trade agreements. Using their 'political-contributions' approach, they show that such agreements between two countries are impossible if in every sector one country has a higher tariff than the other. These agreements might be politically feasible only when tariffs on some goods are higher in one country while other tariffs are higher in its partner country. The possibility of exclusion of certain sectors from the trade agreement also raises the chances that the agreement will be signed.

Reciprocity and unilateralism in trade policy

I now move to the issue of reciprocity and unilateralism in trade policy. Bagwell and Staiger (1996; 1999; 2002) have analysed the issue of reciprocal trade liberalization in considerable detail in both bilateral and multilateral settings (see also Hillman and Moser, 1996). In their models, reciprocity is a way of eliminating terms-of-trade externalities in the setting of trade policy. While considerable work

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
has been done on the role of reciprocity in trade policy, the causal interaction between unilateral and reciprocal trade liberalization has been a somewhat neglected issue. Krishna and Mitra (2005) modify the Mitra (1999) lobby-formation framework to study exactly this link. (See Bhagwati, 1990, for an early informal discussion of this idea. See also Coates and Ludema, 2001, for an alternative channel based on risk sharing through which unilateralism induces reciprocity.) While reciprocal reduction in trade barriers can reasonably be expected to occur in contexts involving trade negotiations between countries, Krishna and Mitra examine instead the question of whether unilateral trade liberalization by one country can induce reciprocal liberalization by its partner *in the absence of any communication or negotiations* between the two countries. In this context, they show that unilateral liberalization by one country can affect the political economy equilibrium in the partner country through the formation of an export lobby there, in a manner that induces it to liberalize trade.

The political economy of regionalism versus multilateralism

An important question raised by Bhagwati (1993; 1994) in several of his writings is whether regionalism is a ‘stumbling block’ or a ‘stepping stone’ to multilateralism. (For a purely economic answer that relies on coordination failure based on sector-specific sunk costs and ‘friction’ in trade negotiation, see McLaren, 2002.) Levy (1997) uses a Heckscher–Ohlin set-up with monopolistic competition and a median-voter approach to address this issue. He finds that bilateral agreements between countries similar in factor endowments result in the subsequent blocking of multilateral trade agreements. He also finds that bilateral agreements can never increase the political support for multilateralism. Krishna (1998) addresses the same issue in a political economy set-up where profits get a much greater weight than other components of welfare in the government's objective function (political-support function approach). The set-up is one of Cournot oligopoly. He finds greater political support for trade-diverting bilateral agreements (regionalism) than for trade-creating ones. Such agreements can also make previously feasible multilateral agreements politically infeasible. This effect turns out to be increasing in the magnitude of the trade diversion that takes place under bilateralism.

Free trade areas versus customs unions

Next I move to the determinants of the actual shape or form a preferential trading arrangement will take. Panagariya and Findlay (1996) study the choice between a customs union and a free-trade area in the context of how they affect lobbying activity and the structure of external tariffs. Using a tariff-formation function approach, they focus on the free-rider problem in lobbying in the case of customs union arising from the requirement of a common external tariff. Richardson (1994) is similar in spirit and finds the same free-rider effect under a customs union, due to which free-trade areas are preferred by import-competing producers.
McLaren (2004) takes a different approach to the choice between a free trade area and a customs union. He analyses what he calls the ‘dynamics of political influence’. As the external tariff is common across members of a customs union, it has to be set jointly by all the members, which can be done only if an agreement is reached among them. This makes the external tariff relatively less reversible under a customs union than under a free-trade area. It is this relative irreversibility that McLaren focuses on,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
even though he abstracts from the uniformity aspect. Using a political contributions approach with Nash bargaining between the capitalists and the government, he finds that a customs union is more likely when the government has a short lifespan and firms have a long lifespan. This is because the more permanent nature of trade policy under a customs union requires the upfront payment of contributions, which is the government's share in the present value of the stream of surpluses generated over time.

The choice of policy instruments

The next important application concerns the choice of policy instruments. One of the major propositions of the theory of commercial policy is that, if distortions or policy goals are not directly trade-related, then direct subsidies are more efficient than tariffs (Bhagwati and Ramaswami, 1963; Johnson, 1965; Bhagwati, 1971). One simple explanation for the existence of tariffs despite their low efficiency is that they generate revenues while subsidies use them up (see for instance Bhagwati and Ramaswami, 1963). However, in a Bhagwati and Srinivasan (1980; 1982) framework of fully competitive revenue seeking, subsidies can be preferable to tariffs as they are not subject to revenue seeking.

Rodrik (1986) is the first author to look at this issue by endogenizing policy. He does this by using a simplified version of the Findlay–Wellisz model. He argues that, since tariffs are general to an industry and subsidies can, in principle, be firm-specific, the free-rider problem in lobbying for tariffs may result in a smaller level of endogenous tariffs than endogenous subsidies, thereby possibly reversing the conventional welfare ranking of tariffs and subsidies. Mitra (2000) argues that, even from the point of view of the import-competing firms, tariffs may be preferable to subsidies since lobbying in the latter may face a congestion problem while in the former the free-rider problem may offset the congestion problem.

This issue of the choice of instruments is also addressed by Grossman and Helpman (1994). They argue that, when the policy instrument used for redistribution is more efficient, it creates greater competition among lobbies and thus results in a larger proportion of the surplus in the hands of the government. Therefore, lobbies themselves may want to tie the hands of the government to using relatively inefficient instruments. Wilson (1990) makes a similar argument in a model with electoral competition where he shows that a higher efficiency of redistributive instruments leads to more contributions and more transfers in equilibrium.

The choice between tariffs and subsidies has also been considered in a voting framework in a series of papers by Mayer and Riezman (1987; 1989; 1990). They show that tariffs can be chosen in equilibrium outcome when voters differ along dimensions other than factor endowments, such as tastes and preferences, treatment under income taxes, and so forth. Besides, income tax progressivity might mean that the cost of financing subsidies is borne unevenly, which might lead some individuals to prefer tariffs whose costs are more evenly distributed.

Feenstra and Lewis (1991) argue that tariffs are informationally more efficient than subsidies. When the world price of importables falls, a tariff equivalent to this decline will compensate the losers without making others worse off relative to the initial situation. This will be possible without any knowledge on the part of the government of individual production and consumption levels. Magee, Brock and Young (1989) propose another information-based explanation, which they call the principle of ‘optimal obfuscation’. Indirect policies such as tariffs are less observable by those who bear its costs.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Staiger and Tabellini (1987) argue that, when governments provide surprise protection to those hurt by world price fluctuations, the time-inconsistency problem might be less severe with more inefficient policies.

Hysteresis in trade policy

A model that helps us understand status quo bias in trade policy is Fernandez and Rodrik (1991). They consider a two-sector economy that initially has a certain given tariff on its imports. Eliminating this tariff will result in a movement of workers from the import-competing sector to the export sector. What is *ex ante* unknown is which of the workers initially in the import-competing sector will be successful in moving to the export sector. All the workers who are in the export sector right from the beginning will gain, while those who are always in the import-competing sector and remain there after the reforms will lose. Another group that gains is the group of movers from the contracting import-competing sector to the expanding export sector. Suppose 30 per cent of the population is in the export sector and 70 per cent in the import-competing sector to start with. After the reforms, let us suppose that this split is 60 per cent and 40 per cent respectively. This means that 60 per cent of the population will gain *ex post* from the reform. While 30 per cent who are initially in the export sector know for sure *ex ante* they are going to benefit, the remaining 70 per cent do not know which 30 per cent out of them will lose and which 40 per cent will gain. If they know for sure that the loss incurred by the losing 40 per cent is greater than the gain to the remaining 30 per cent, then all the voters who are initially in the import-competing sector will vote against the reform. Due to the individual-specific uncertainty faced by workers in the import-competing sector, each of them will vote on the basis of an expected loss, arising from the fact that losers in this sector lose much more than gainers in that sector gain. Thus, even though *ex post* a majority gain from the reforms, *ex ante* a majority of the workers vote against the trade reforms. However, if a dictator or an international financial institution forces a reform upon these people, it will not be reversed since as we know in this case *ex post* there is going to be majority support for the reforms.

Beyond the monolithic government

In the existing political economy literature on trade policy determination, a single, monolithic policymaker is often assumed. Until very recently, the paper by Feenstra and Bhagwati (1982) was the only exception. McLaren and Karabay (2004) make a departure from such a simple structure to study trade policy setting in the presence of parliamentary or congressional institutions. They also incorporate electoral competition between political parties and show that in their setting the equilibrium tariff is the optimum of the median voter in the median district. They find that the relationship between the likelihood of import protection and the geographical concentration of import-competing interests is non-monotonic, with a maximum occurring at moderate levels of concentration. Too much concentration leads to a control of too few seats, while too much dispersion leads to no control of any seats.
A paper by Grossman and Helpman (2005) allows actual policy formation to be the interplay between the policy platform announced by the party leadership and the actions of individual legislators who want to maximize political success. Maximization of political success involves resolving the trade-off

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
between conforming to the party platform and making one's constituents happy, thereby resulting in a deviation of ‘policy reality’ from ‘policy rhetoric’. The authors find a protectionist bias when the legislature operates under majority rule. This bias is increasing in the geographical concentration of assets and capital market imperfections, and is decreasing in party discipline.

Empirical evidence

The old empirical literature

The empirical literature in this area has evolved from being highly ‘reduced form’ and atheoretical to being fairly ‘structural’ and guided by tight theoretical models. Important papers in the earlier literature include Caves (1976), Saunders (1980), Ray (1981), Marvel and Ray (1983), Ray (1991) and Trefler (1993). (See Rodrik, 1995, for a detailed survey of this literature.) The main finding of this early empirical literature is that protection is higher for sectors that are labour-intensive, low-skill and low-wage, for consumer-goods industries, for industries facing high import penetration, where geographical concentration of production is high but that of consumers is low, and in sectors with low levels of intra-industry trade. (For an examination of the cross-national variation in average protection levels across industrialized countries, see Mansfield and Busch, 1995. They find that non-tariff barriers are increasing in country size, unemployment rate and number of parliamentary constituencies, and are higher for countries that use proportional representation as their electoral system.)

The new empirical literature

In the Heckscher–Ohlin version of the Mayer median-voter model, a simple comparative static exercise produces the result that a rise in asset inequality will make trade policy more pro-trade in a labour-abundant economy and more protectionist in a capital-abundant economy. Dutt and Mitra (2002) find strong support for this result using cross-country data on inequality, capital-abundance and diverse measures of protection. (In this context, it is also important to mention Milner and Kubota, 2005, who use a median-voter approach to empirically investigate the relationship between democratization and trade reforms in developing countries.)

Dutt and Mitra (2005) also perform a cross-country empirical investigation of the role of political ideology in trade policy determination. They use a political-support function approach within a two-sector, two-factor Heckscher–Ohlin model (see Milner and Judkins, 2004, on this issue. Hiscox (2001) studies six Western nations to look at how historically the nature and structure of partisanship on trade issues change over time and depend on the extent of inter-sectoral factor mobility. Hiscox (2002) looks at the same question exclusively for the United States, analysing major pieces of congressional trade legislation between 1824 and 1994.

Two empirical papers, Goldberg and Maggi (1999) and Gawande and Bandyopadhyay (2000), estimate the Grossman–Helpman ‘Protection for Sale’ tariff expressions using industry-level data from the United States. The two papers are similar in the questions they address, but are somewhat different in the details of their approaches. While Goldberg and Maggi restrict their focus to the protection expressions, Gawande and Bandyopadhyay concentrate more on the lobbying aspects and the determinants of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
magnitude of contributions. Goldberg and Maggi use the basic Grossman–Helpman framework, while Gawande and Bandyopadhyay introduce intermediate goods. The econometric specifications are therefore somewhat different in the two papers. However, the results in the two papers are very similar. Both confirm empirically the Grossman–Helpman prediction regarding the relationship of protection to import penetration and import-demand elasticity. With everything else held constant, organized sectors are granted higher protection than unorganized sectors. Both papers find that the weight on aggregate welfare in the government's objective function is several times higher than that on contributions. Also, the estimates of the proportion of the population organized are very high in both papers.

Mitra, Thomakos and Ulubasoglu (2002) and McCalman (2004) obtain similarly high parameter estimates of the Grossman–Helpman model for Turkey and Australia respectively. An interesting result that comes out of the empirical exercise by Mitra, Thomakos and Ulubasoglu is that the relative weight on aggregate welfare was higher in the democratic regimes than under the dictatorial regimes in Turkey.

Gawande, Krishna and Robbins (2006) empirically investigate ‘the susceptibility of government policies to lobbying by foreigners’. Using a new data-set on foreign political activity in the United States, they investigate the empirical relationship between trade protection and lobbying activity. Their theoretical framework is an extension of the ‘Protection for Sale’ model to include foreign lobbies. They find that foreign lobbying activity has significantly affected US trade barriers in a negative direction, as predicted by their model. They conclude: ‘If the policy outcome absent any lobbying by foreigners is characterized by welfare-reducing trade barriers, lobbying by foreigners may result in reductions in such barriers and raise consumer surplus (and possibly improve welfare).’

In another empirical application through an extension of the Grossman–Helpman model, Gawande and Krishna (2005) investigate the effects of lobbying competition between upstream and downstream producers for US trade policy. Their parameter estimates are a significant improvement over those in the earlier literature even though they do not completely resolve the puzzle.

Thus, we see that the political economy literature on trade policy has evolved a great deal on both the theoretical and the empirical sides, as well as in terms of the complexity of applications it can handle.

See Also

- political competition
- policy reform, political economy of
- ‘political economy’
- political institutions, economic approaches to

Bibliography

Bagwell, K. and Staiger, R. 1996. Reciprocal trade liberalization. Working Paper No. 5488. Cambridge, MA: NBER.

Bagwell, K. and Staiger, R. 1999. An economic theory of GATT. *American Economic Review* 89, 215–48.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bagwell, K. and Staiger, R. 2002. *The Economics of the World Trading System*. Cambridge, MA, and London: MIT Press.

Bhagwati, J. 1971. The generalized theory of distortions and welfare. In *Trade, Balance of Payments and Growth*, ed. J. Bhagwati et al. Amsterdam: North-Holland.

Bhagwati, J. 1990. Aggressive unilateralism. In *Aggressive Unilateralism*, ed. J. Bhagwati and H. Patrick. Ann Arbor: University of Michigan Press.

Bhagwati, J. 1993. Regionalism and multilateralism: an overview. In *New Dimensions in Regional Integration*, ed. A. Panagariya and J. De Melo. Washington, DC: World Bank.

Bhagwati, J. 1994. Threats to the world trading system: income distribution and the selfish hegemon. *Journal of International Affairs* 48, 279–85.

Bhagwati, J. and Ramaswami, V. 1963. Domestic distortions, tariffs and the theory of the optimum subsidy. *Journal of Political Economy* 71, 44–50.

Bhagwati, J. and Srinivasan, T. 1980. Revenue seeking: a generalization of the theory of tariffs. *Journal of Political Economy* 88, 1069–87.

Bhagwati, J. and Srinivasan, T. 1982. The welfare consequences of directly unproductive profit-seeking (DUP) lobbying activities: price versus quantity distortions. *Journal of International Economics* 13, 33–44.

Caves, R. 1976. Economic models of political choice: Canada's tariff structure. *Canadian Journal of Economics* 9, 278–300.

Coates, D. and Ludema, R. 2001. A theory of trade policy leadership. *Journal of Development Economics* 65, 1–29.

Dutt, P. and Mitra, D. 2002. Endogenous trade policy through majority voting: an empirical investigation. *Journal of International Economics* 58, 107–33.

Dutt, P. and Mitra, D. 2005. Political ideology and endogenous trade policy: an empirical investigation. *Review of Economics and Statistics* 87, 59–72.

Feenstra, R. and Bhagwati, J. 1982. Tariff seeking and the efficient tariff. In *Import Competition and Response*, ed. J. Bhagwati. Chicago and London: University of Chicago Press.

Feenstra, R. and Lewis, T. 1991. Distributing the gains from trade with incomplete information.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Economics and Politics 3, 29–40.

Fernandez, R. and Rodrik, D. 1991. Resistance to reform: status-quo bias in the presence of individual-specific uncertainty. *American Economic Review* 81, 1146–54.

Findlay, R. and Wellisz, S. 1982. Endogenous tariffs, the political economy of trade restrictions and welfare. In *Import Competition and Response*, ed. J. Bhagwati. Chicago and London: University of Chicago Press.

Gawande, K. and Bandyopadhyay, S. 2000. Is protection for sale? A test of the Grossman–Helpman theory of endogenous protection. *Review of Economics and Statistics* 82, 139–52.

Gawande, K. and Krishna, P. 2003. The political economy of trade policy: empirical approaches. In *Handbook of International Trade*, ed. J. Harrigan and E. Kwan Choi. Malden, MA: Basil Blackwell.

Gawande, K. and Krishna, P. 2005. Lobbying competition over US trade policy. Working Paper No. 11371. Cambridge, MA: NBER.

Gawande, K., Krishna, P. and Robbins, M. 2006. Foreign lobbies and US trade policy. *Review of Economics and Statistics* (forthcoming).

Goldberg, P. and Maggi, G. 1999. Protection for sale: an empirical investigation. *American Economic Review* 89, 1135–55.

Grossman, G. and Helpman, E. 1994. Protection for sale. *American Economic Review* 84, 833–50.

Grossman, G. and Helpman, E. 1995a. Trade wars and trade talks. *Journal of Political Economy* 103, 675–708.

Grossman, G. and Helpman, E. 1995b. The politics of free trade agreements. *American Economic Review* 85, 667–90.

Grossman, G. and Helpman, E. 2005. A protectionist bias in majoritarian politics. *Quarterly Journal of Economics* 120, 1239–82.

Helpman, E. 2002. Politics and trade policy. In *Interest Groups and Trade Policy*, ed. G. Grossman and E. Helpman. Princeton, NJ: Princeton University Press.

Hillman, A. 1989. *The Political Economy of Protection*. Chur: Harwood Academic Publishers.

Hillman, A. and Moser, P. 1996. Trade liberalization as politically optimal exchange of market access.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In *The New Transatlantic Economy*, ed. M. Canzoneri, W. Ethier and V. Grilli. Cambridge: Cambridge University Press.

Hiscox, M. 2001. *International Trade and Political Conflict: Commerce, Coalitions and Mobility*. Princeton, NJ: Princeton University Press.

Hiscox, M. 2002. Commerce, coalitions, and factor mobility: evidence from congressional votes on trade legislation. *American Political Science Review* 96, 593–608.

Johnson, H. 1965. Optimal trade interventions in the presence of domestic distortions. In *Trade Growth and Balance of Payments*, ed. R. Caves, H. Johnson and P. Amsterdam: North-Holland.

Krishna, P. 1998. Regionalism and multilateralism: a political economy approach. *Quarterly Journal of Economics* 113, 227–51.

Krishna, P. and Mitra, D. 2005. Reciprocated unilateralism in trade policy. *Journal of International Economics* 65, 461–87.

Levy, P. 1997. A political-economic analysis of free-trade agreements. *American Economic Review* 87, 506–19.

Magee, C. 2002. Endogenous trade policy and lobby formation: an application to the free-rider problem. *Journal of International Economics* 57, 449–71.

Magee, S., Brock, W. and Young, L. 1989. *Black Hole Tariffs and Endogenous Policy Theory*. Cambridge and New York: Cambridge University Press.

Maggi, G. and Rodriguez-Clare, A. 1998. The value of trade agreements in the presence of political pressures. *Journal of Political Economy* 106, 574–601.

Mansfield, E. and Busch, M. 1995. The political economy of trade barriers: a cross-national analysis. *International Organization* 49, 723–49.

Marvel, H. and Ray, E. 1983. The Kennedy Round: evidence on the regulation of trade in the US. *American Economic Review* 73, 190–7.

Mayer, W. 1984. Endogenous tariff formation. *American Economic Review* 74, 970–85.

Mayer, W. and Riezman, R. 1987. Endogenous choice of trade policy instruments. *Journal of International Economics* 23, 377–81.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mayer, W. and Riezman, R. 1989. Tariff formation in a multidimensional voting model. *Economics and Politics* 1, 61–79.

Mayer, W. and Riezman, R. 1990. Voter preferences for trade policy instruments. *Economics and Politics* 2, 259–73.

McCalman, P. 2004. Protection for sale and trade liberalization: an empirical investigation. *Review of International Economics* 12, 81–94.

McLaren, J. 2002. A theory of insidious regionalism. *Quarterly Journal of Economics* 117, 571–608.

McLaren, J. 2004. Free trade agreements, customs unions and the dynamics of political influence. Mimeo, University of Virginia.

McLaren, J. and Karabay, B. 2004. Trade policy making by an assembly. In *Political Economy of Trade, Aid and Foreign Investment Policies*, ed. D. Mitra and A. Panagariya. Amsterdam: Elsevier.

Milner, H. and Judkins, B. 2004. Partisanship, Trade policy, and globalization: is there a left–right divide on trade policy? *International Studies Quarterly* 48, 95–119.

Milner, H. and Kubota, K. 2005. Why the move to free trade? Democracy and trade policy in the developing countries. *International Organization* 59, 107–43.

Mitra, D. 1999. Endogenous lobby formation and endogenous protection: a long run model of trade policy determination. *American Economic Review* 89, 1116–34.

Mitra, D. 2000. On the endogenous choice between protection and promotion. *Economics and Politics* 12, 33–52.

Mitra, D. 2002. Endogenous political organization and the value of trade agreements. *Journal of International Economics* 57, 473–85.

Mitra, D., Thomakos, D. and Ulubasoglu, M. 2002. ‘Protection for sale’ in a developing country: democracy vs. dictatorship. *Review of Economics and Statistics* 84, 497–508.

Panagariya, A. and Findlay, R. 1996. A political-economy analysis of free trade areas and customs union. In *The Political Economy of Trade Reform: Essays in Honor of Jagdish Bhagwati*, ed. R. Feenstra, D. Irvin and G. Grossman. Cambridge, MA: MIT Press.

Pecorino, P. 1998. Is there a free-rider problem in lobbying? Endogenous tariffs, trigger strategies and the number of firms. *American Economic Review* 88, 652–60.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ray, E. 1981. The determinants of tariff and non-tariff restriction in the United States. *Journal of Political Economy* 89, 105–21.

Ray, E. 1991. Protection of manufactures in the United States. In *Global Protectionism: Is The US Playing on a Level Field?*, ed. D. Greenaway. London: Macmillan.

Richardson, M. 1994. Why a free trade area? The tariff also rises. *Economics and Politics* 6, 79–96.

Rodrik, D. 1986. Tariffs, subsidies and welfare with endogenous policy. *Journal of International Economics* 21, 285–96.

Rodrik, D. 1995. Political economy of trade policy. In *Handbook of International Economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: North-Holland.

Saunders, R. 1980. The political economy of effective protection in Canada's manufacturing sector. *Canadian Journal of Economics* 13, 340–8.

Staiger, R. and Tabellini, G. 1987. Discretionary trade policy and excessive protection. *American Economic Review* 77, 823–37.

Trefler, D. 1993. Trade liberalization and the theory of endogenous protection. *Journal of Political Economy* 101, 138–60.

Van Long, N. and Vousden, N. 1991. Protectionist responses and declining industries. *Journal of International Economics* 30, 87–103.

Wilson, J. 1990. Are efficiency improvements in government transfer policies self defeating in equilibrium? *Economics and Politics* 2, 241–58.

How to cite this article

Mitra, Devashish. "trade policy, political economy of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000213> doi:10.1057/9780230226203.1727

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

trade, technology diffusion and growth

Jonathan Eaton and Samuel Kortum
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Invention is a fundamental driver of growth in living standards around the world. Trade and technology diffusion are two ways the benefit of an invention spreads to other countries. Impediments to diffusion give rise to differences in living standards and to comparative advantage in production, providing an incentive to trade. We review some basic facts. We then provide a simple model of the connections among these processes, showing how it can explain these facts.

Keywords

balanced growth; capital goods trade; Cobb–Douglas functions; diffusion of technology; economic growth; economic growth in the very long run; endogenous growth; international trade; invention; patents; product life cycle; Schmookler, J.; technical change

Article

‘That the creation and diffusion of technological knowledge is at the heart of modern economic growth is now widely accepted’ (Jacob Schmookler, 1966). While technological advances drive world growth, differences in the ability of countries to exploit inventions create wide disparities in income. Understanding how such disparities arise requires identifying how the benefits of technological progress spread. A country might benefit from a new technique through trade, importing goods embodying that technique, or through diffusion, learning the technique and using it.

Impediments to technology diffusion mean that some countries have techniques that others lack, generating comparative advantage in production, as in Ricardo's model of trade. Over time, diffusion can eliminate or even reverse comparative advantage. As Vernon's (1966) product cycle posited, if diffusion is slow more inventive countries will export goods associated with recent inventions while importing goods embodying older technology. More recent work has modelled how technologies arise and spread across countries, driving world growth.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Before turning to how economists have modelled trade, diffusion and growth, we fill in this picture with some observations.

An overview of the evidence

When we look at trade in goods, several features stand out: (1) Over the second half of the 20th century, trade has grown as a share of GDP. (2) Nevertheless, geography continues to matter: countries buy much more from themselves and their neighbours than their shares in world production would imply. (3) Rich countries trade more among themselves and with poor countries than poor countries trade among each other.

Our ability to quantify invention is limited to observations on resources devoted to it (for example, research scientists and engineers) or on activities related to inventive output (for example, patents). These two measures paint a very similar picture: invention is concentrated in a small number of rich countries, although many rich countries are not particularly inventive. The patent data, like the trade data, have a further bilateral dimension as inventors from one country seek protection at home and elsewhere. As with merchandise trade, most patenting is done at home or nearby. Yet, most countries issue a majority of their patents to foreign inventors.

Income measures for the post-war era come from national accounts. Sifting through various pieces of evidence, economic historians, in particular Maddison (2003), have constructed measures going back several centuries. The basic picture here is that many countries have experienced sustained growth over a long period of time, with only infrequent switches in relative position.

Other evidence on invention and diffusion comes from longer ago. Diamond (1997) describes archaeological findings on when and where great innovations in agriculture, the domestication of plants and animals for farming, occurred, and how they diffused across continents. Sheep and wheat, for example, originated in south-western Asia around 8,500 bc and within a few millennia had diffused throughout the Eurasian land mass. Corn and turkeys had originated in Mesoamerica by 3,500 and slowly found their way through much of the Americas. Before Columbus, however, the two sets of technologies remained confined to their hemisphere of origin.

A model

We now turn to a model of international technology diffusion, international trade and economic growth that is consistent with these observations.

A world of ideas

We think about technology as a set of ideas that can be applied to production. New ideas generate technological advances if they lead to new goods and services or improve existing ones.

Ideas are inherently non-rival. An idea can be used in many places at the same time. Yet it may take a long time for an idea to spread. We sketch a simple formulation of this process, building on Nelson and Phelps (1966) and Krugman (1979).

We distinguish among three classes of a country's technology at any moment: (i) the measure of ideas T_i

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

that country *i* invented itself, (ii) the measure of ideas T_i^A *available* to *i*, and (iii) the measure T_{ni}^E of *exclusive* ideas invented in *i* that are not yet available in *n*. Defining T^W as the set of ideas in the world:

$$T^W = \sum_i T_i.$$

New ideas arrive to country *i* at rate $\dot{T}_i$. Initially they are exclusive to *i*. They become known in country $n \neq i$, thus transiting from T_{ni}^E to T_n^A, with a hazard ε_{ni}.
Two sets of dynamic equations describe the evolution of technology. The first applies to the increase in available technology in country *n*:

$$\dot{T}_n^A = \dot{T}_n + \sum_{i \neq n} \varepsilon_{ni} T_{ni}^E.$$

The second applies to the change in its exclusive technologies:

$$\dot{T}_{ni}^E = \dot{T}_i - \sum_{n \neq i} \varepsilon_{ni} T_{ni}^E. \quad (1)$$

The literature has related invention to human effort and to research spillovers, acknowledging that inventors stand on the shoulders of others. A simple formulation highlighting research spillovers, used by Krugman (1979), is

$$\dot{T}_i = \iota_i T_i^A,$$

where $\iota_i \geq 0$ is an exogenous rate of invention in *i*. Romer (1990) models the determination of ι_i, turning it into an endogenous growth model. A simple formulation, highlighting human input, in the spirit of the semi-endogenous growth model of Jones (1995), is:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\dot{T}_i = \iota_i L_i.$$

Here L_i is the labour force in country i, which grows at a constant rate $g_L>0$ in each country.
With an arbitrary number of countries, under general assumptions about the parameters of invention (the ι's) and diffusion (the ε's), patterns of the distribution of technology can be complex. Nevertheless, if these parameters remain constant, with the matrix of ε's indecomposable (ensuring that every proper subset of countries can absorb technology from outside), the world will evolve toward a balanced growth path with all T's growing at rate g_T. In the research spillover case g_T is given by an eigenvalue (the Frobenius root) of the matrix of ε's and ι's, and is increasing in both sets of parameters. In the human input case g_T equals the population growth rate g_L. In either case more inventive countries and those quicker to adopt technologies from elsewhere will have more ideas at any moment. Countries that are slow to adopt eventually fall far enough behind that they can draw from a stock of unknown technologies large enough to pull them along at the same growth rate as the most advanced (Eaton and Kortum, 1999). Diamond's archaeological examples illustrate the phenomenon nicely.
A two-country example, using the human input specification of invention (so that $g_L=g_T>0$) provides some basic results. With two countries we can simplify $T^E_{ni} = T^E_i$ and $\varepsilon_{ni}=\varepsilon_i$, and define the measure of ideas that have spilled out of either country's exclusive technology as T^C. It is straightforward to calculate the various technology measures along a balanced growth path:

$$T_i = \frac{\iota_i}{g_L} L_i T^E_i = \frac{g_L}{g_L+\varepsilon_i} T_i T^C = \frac{\varepsilon_1}{g_L+\varepsilon_1} T_1 + \frac{\varepsilon_2}{g_L+\varepsilon_2} T_2 T^A_i = T_i + \frac{\varepsilon_n}{g_L+\varepsilon_n} T_n \; n \neq i.$$

These equations show us how the various measures of technology, each growing at rate g_L, relate to the underlying parameters of invention and diffusion. In particular, a higher rate of diffusion ε_1 out of country 1 leads to more technology available in country 2.
For an idea to affect economic welfare it has to be connected with a good or service that people enjoy. We can then ask how the distribution of ideas relates to production and trade, and thus how ideas ultimately affect welfare.
To make this connection we need to be concrete about preferences. We assume that utility can be represented as a constant-elasticity-of-substitution function of consumption over a wide variety of differentiated goods indexed by j. The implied price index is:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$p = \left[\int_0^J p(j)^{-(\sigma-1)} dj\right]^{-1/(\sigma-1)},$$

(2)

where $p(j)$ is the price of good *j*, *J* is the measure of goods, and $\sigma > 0$ is the elasticity parameter.

Ideas as new goods

A particularly simple case equates an idea to a new good, with one worker required to produce one unit, as in Krugman (1979) and Romer (1990). We let $J \to \infty$ in (2) so as not to limit growth. A good not yet invented has an implicit price of infinity, hence to bound utility we require $\sigma > 1$.

Without trade countries can consume only the goods they know how to produce themselves. Under perfect competition a unit of any good that is produced will cost the local wage. Since T_i^A is the measure of goods produced in country *i*, the real wage in country *i*, which is the inverse of the price index there, is simply:

$$\frac{w_i}{p_i} = \left(T_i^A\right)^{1/(\sigma-1)}.$$

An increase in country *i*'s available technology raises welfare by increasing the variety of available goods. More rapid technology diffusion out of country 1 helps country 2 while doing no harm to country 1. The dynamics of technology diffusion lead to parallel growth in living standards around the world, at rate $g_T/(\sigma-1)$.

International trade allows a country to consume goods that it doesn't know how to make itself. In the two-country case with costless trade there are two cases to consider.

In the first, the countries' relative labour forces are in line with their access to technology. More precisely:

$$\frac{T_1^E}{T^W} < \frac{L_1}{L^W} < \frac{T_1^E + T^C}{T^W},$$

where $L^W = L_1 + L_2$. In this case $w_1 = w_2 = w$. Both countries will produce some goods using the common technology and all goods sell at a price equal to the common wage. The common real wage is:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\frac{w}{p} = \left(T^{W}\right)^{1/(\sigma-1)}.$$

Comparing this wage to the autarky wage above, trade perfectly substitutes for diffusion. Faster technology diffusion would require less trade, but leave welfare in both countries unchanged.
The other case emerges if one country, say country 1, is advanced technologically relative to the size of its labour force, or if:

$$\frac{T_1^E}{T^W} > \frac{L_1}{L^W}.$$

Now $w_1 > w_2$, and country 1 produces only goods associated with T_1^E, while 2 produces goods associated with all its available technology, which includes T_2^E and T^C. Country 1 exports its exclusive goods in exchange for both country 2's exclusive goods and the goods that both could in principle make. The demand for any good produced in country 1 relative to any good produced in country 2 is:

$$\frac{c_1}{c_2} = \left(\frac{w_1}{w_2}\right)^{-\sigma}.$$

Since total demand for country 1 relative to country 2 labour is:

$$\frac{L_1}{L_2} = \frac{T_1^E c_1}{T_2^A c_2},$$

the relative wage satisfies:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\frac{w_1}{w_2} = \left(\frac{L_1}{L_2}\right)^{-1/\sigma}\left(\frac{T_1^E}{T_2^A}\right)^{1/\sigma}.$$

Country 1 has a higher standard of living the smaller it is and the *higher* the ratio of its exclusive ideas to the ideas available to country 2. The real wage in country 1 is

$$\frac{w_1}{p} = \left\{T^W + \left[\left(\frac{w_1}{w_2}\right)^{\sigma-1} - 1\right]T_2^A\right\}^{1/(\sigma-1)},$$

which exceeds $\left(T^W\right)^{1/(\sigma-1)}$.

The two cases with international trade illustrate basic points about the relationship between technology diffusion, trade, and welfare. In the case of equal wages and costless trade the rate of diffusion does not matter for income or welfare. With unequal wages, more diffusion necessarily benefits country 2, while the effect on country 1 is ambiguous. If T_2^A starts off very small relative to T^W, an increase in diffusion benefits country 1 by allowing it to import more goods at a low price. But at higher values of T_2^A the gain from being able to import more goods at a low price is offset by the increase in the price it pays for all imports. At some point T_2^A gets so large that $w_1=w_2$ and we are back in the previous case (bounding country 1's loss from 2's access to its technology).

In the equal wage case technology diffusion can act as a substitute for trade: once a good enters the common technology it may no longer be traded. In the second case, diffusion from country 1 to country 2 reverses the direction of trade. What 1 once exported it now imports.

A straightforward extension allows for multiple inventive, high-wage countries and multiple low-wage countries relying on the common technology. Innovative countries exchange goods associated with their exclusive technologies among themselves. They export these goods to low-wage countries in exchange for goods in the common technology. But low-wage countries have no incentive to trade goods in the common technology among themselves.

Ideas as better goods

Alternatively, an idea may not be a new good but rather a new technique for producing an existing good. This approach appears in Grossman and Helpman (1991), Aghion and Howitt (1992), Kortum (1997), and Eaton and Kortum (1999). Since there are no new goods, we set J=1 in (2) and drop the restriction that the substitution elasticity exceeds 1.

Let's begin with a stark case of international trade, along the lines of Armington (1969), in which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

country 1 produces only goods $j \leq 1/2$ and country 2 only goods $j > 1/2$. The efficiency with which country *i* produces its range of goods is T_i. The demand for any good produced in country 1 relative to any good produced in country 2 is:

$$\frac{c_1}{c_2} = \left(\frac{w_1 / T_1}{w_2 / T_2}\right)^{-\sigma},$$

Since the total demand for country 1 relative to country 2 labour is:

$$\frac{L_1}{L_2} = \frac{c_1 / T_1}{c_2 / T_2},$$

the relative wage satisfies:

$$\frac{w_1}{w_2} = \left(\frac{L_1}{L_2}\right)^{-1/\sigma}\left(\frac{T_1}{T_2}\right)^{(\sigma-1)/\sigma}.$$

As above, country 1's relative wage is higher the smaller it is, but its technological lead has an ambiguous effect, depending on σ.

Consider technological stagnation in country 2 so that T_2=1 for ever while T_1 grows at rate g_T>0 (with relative labor forces not changing). If σ<1 the stagnant country 2 actually grows faster while σ > 1 means faster growth for country 1. With Cobb–Douglas preferences (σ=1) wages grow at the same rate: country 1's greater inventiveness is exactly offset by its worsening terms of trade. Oil-exporting countries, for example, can grow faster than the rest of the world even though they do not rank highly in standard technology indicators.

If we remove the Armington assumption, so that either country can in principle produce anything, the possibilities for parallel growth expand. Suppose that to produce a unit of any good in country 2 requires one worker, as does producing any good *j*>1/2 in country 1. Country 1 can produce any good $j \leq 1/2$ with efficiency T_1>1. If σ<1, the stagnant country 2 cannot grow faster for ever. Once w_2 hits w_1, thenceforth w_2=w_1 as country 1 produces an ever greater range of goods *j*>1/2, approaching its share of the world labour force asymptotically. In this case with inelastic demand, even with no technology diffusion trade spreads the benefits of 1's technological advances evenly. If σ>1, on the other hand, then as country 1's technology improves it remains specialized in goods $j \leq 1/2$ and its relative wage

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

grows for ever.

While this two-country example illustrates some basic points about innovation, diffusion and trade very neatly, it fails to account for a number of the basic facts discussed earlier. In particular, the world consists of many countries displaying varying degrees of innovative activity. While countries trade with each other, barriers to international trade remain substantial.

Eaton and Kortum (1999; 2002) develop a framework in which the same basic forces drive growth and trade among many heterogeneous countries separated by trade barriers. They treat an idea as a way to produce some good $j \in [0,1]$ with some efficiency $q(j)$ drawn from a Pareto distribution with parameter θ. The implication is that, if country *i* has access to a measure T_i^A of ideas, its best technology for making good *j*, $z_i(j)$, is a realization drawn from:

$$G(z, T_i^A) = \Pr[Z_i \leq z] = e^{-T_i^A z^{-\theta}},$$

a type II extreme value distribution. Looking across the unit continuum of goods, $e^{-T_i z^{-\theta}}$ is the fraction that country *i* can produce with an efficiency no more than *z*.

With no international trade, the price of good *j* in country *i* is $p_i(j) = w / z_i(j)$. The real wage in country *i* as well as the standard of living there, is thus

$$\frac{w_i}{p_i} = \left[\int_0^\infty z^{\sigma-1} dG(z; T_i^A)\right]^{1/(\sigma-1)} = \gamma (T_i^A)^{1/\theta}, \tag{3}$$

(where γ is constant that depends on θ and σ). Trade provides the potential to exploit comparative advantage. An iceberg transport costs d_{ni} separates country *n* from country *i* (meaning that delivering 1 unit of a good to country *n* requires shipping $d_{ni} \geq 1$ units from *i*). In the case of no diffusion $T_i^A = T_i^E$. The real wage in country *i* then becomes:

$$\frac{w_i}{p_i} = \gamma \frac{w_i}{\left[\sum_{k=1}^N T_k^E (w_k d_{ik})^{-\theta}\right]^{-1/\theta}} \geq \gamma (T_i^E)^{1/\theta}.$$

Trade thus gives consumers in country *i* access to everyone's technologies, appropriately weighted by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

input and transportation costs.

With two countries we get the Ricardian model of Dornbusch, Fischer and Samuelson (1977). Ordering goods in decreasing order of country 1's relative efficiency $A(j) = z_1(j) / z_2(j)$ yields:

$$A(j) = \begin{cases} \left(\dfrac{1-j}{j}\dfrac{T_1^E}{T_2^A}\right)^{1/\theta} & j < \dfrac{T_1^E}{T^W} \\ 1 & \dfrac{T_1^E}{T^W} \le j \le \dfrac{T_1^A}{T^W} \\ \left(\dfrac{1-j}{j}\dfrac{T_1^A}{T_2^E}\right)^{1/\theta} & j > \dfrac{T_1^A}{T^W} \end{cases}.$$

Note that the range of goods over which $A(j) = 1$ is governed by the extent of technology diffusion. With no transport costs and symmetric Cobb–Douglas preferences the equilibrium is characterized by a cutoff good j^* such that country 1 produces goods $j \in [0,j^*)$ while country 2 produces $j \in (j^*,1]$. The relative wage w_1/w_2 and j^* satisfy the conditions for labour market clearing and comparative advantage:

$$\frac{w_1}{w_2} = \frac{L_2}{L_1}\frac{j^*}{1-j^*}\frac{w_1}{w_2} = A(j^*).$$

This two-country case is analyzed in Eaton and Kortum (2007a) with research effort determined endogenously. Rodriguez-Clare (2007) derives quantitative implications of an N-country case with research exogenous.

Ideas embodied in inputs

Our discussion so far has considered two ways in which the benefits of invention spread around the world: through trade in final goods that embody the technology, and through the diffusion of the disembodied idea itself. Ideas embodied in goods used for production provides a third conduit. Trade in capital goods is an important example. As Eaton and Kortum (2001) document, the production and export of capital goods is concentrated in a small number of research-intensive countries. For example, many countries have airlines that fly wide-bodied aircraft, but much of the technology for producing

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
them is limited to Seattle and Toulouse.

Macroeconomists attribute a significant share of income growth to advances in capital equipment, causing its relative price to decline over time. Costless trade would imply the same relative price of capital goods everywhere, allowing all countries to benefit equally from inventions embodied in capital goods.

In fact, trade and price data suggest enormous geographic fragmentation of markets, with poor countries facing a systematically higher relative price of capital. A simple growth model translates differences in the relative price of capital goods to differences in per worker output. If capital has a share α in a Cobb–Douglas production function and depreciates at rate δ, country *i* with a savings rate s_i will have a steady-state level of output per worker y_{it} given by:

$$y_{it} = \left(\frac{s_i}{\left(\delta + \frac{g^K}{1-\alpha}\right)(P_{it}^K / P_i^C)} \right)^{\alpha/(1-\alpha)}.$$

Here P_{it}^K / P_i^C is the relative price of capital goods in country *i* at time *t* and g^K is the rate at which the price of capital goods declines. With constant iceberg trade barriers, g^K is the same everywhere in the world, but P_{it}^K / P_i^C is lower in countries with better access to capital goods. The formulation is thus consistent with a common world growth rate, but with persistent differences in levels over time.

Summary

We have sketched a simple model of trade, diffusion and growth motivated by some basic facts. A body of work has sought to quantify various models of this type. Eaton and Kortum (2007b) review some of this work.

See Also

- balanced growth
- comparative advantage
- diffusion of technology
- endogenous growth theory
- growth and international trade
- inequality between nations
- Perron–Frobenius theorem

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- product life cycle
- Ricardian trade theory
- Schmookler, Jacob
- Schumpeterian growth and growth policy design
- technology
- transfer of technology

Bibliography

Aghion, P. and Howitt, P. 1992. A model of growth through creative destruction. *Econometrica* 60, 323–51.

Armington, P.S. 1969. A theory of demand for products distinguished by place of production. *IMF Staff Papers* 16, 159–78.

Diamond, J. 1997. *Guns, Germs, and Steel*. New York: W.W. Norton.

Dornbusch, R., Fischer, S. and Samuelson, P.A. 1977. Comparative advantage, trade, and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67, 823–39.

Eaton, J. and Kortum, S. 1999. International technology diffusion: theory and measurement. *International Economic Review* 40, 537–70.

Eaton, J. and Kortum, S. 2001. Trade in capital goods. *European Economic Review* 45, 1195–235.

Eaton, J. and Kortum, S. 2002. Technology, geography, and trade. *Econometrica* 70, 1741–80.

Eaton, J. and Kortum, S. 2007a. Innovation, diffusion, and trade. In *Entrepreneurship, Innovation, and the Growth Mechanism of the Free-Enterprise Economies*, ed. E. Sheshinski, R. Strom and W. Baumol. Princeton University Press, 276–299.

Eaton, J. and Kortum, S. 2007b. Technology and the global economy: a framework for quantitative analysis. Manuscript. New York University and University of Chicago.

Grossman, G.M. and Helpman, E. 1991. *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.

Jones, C.I. 1995. R&D-based models of economic growth. *Journal of Political Economy* 103, 759–84.

Kortum, S. 1997. Research, patenting, and technological change. *Econometrica* 65, 1389–419.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Krugman, P.R. 1979. A model of innovation, technology transfer, and the world distribution of income. *Journal of Political Economy* 87, 253–66.

Maddison, A. 2003. *The World Economy: Historical Statistics* (CD ROM). Paris: OECD.

Nelson, R.R. and Phelps, E.S. 1966. Investment in humans, technological diffusion, and economic growth. *American Economic Review* 56, 69–75.

Rodriguez-Clare, A. 2007. Trade, diffusion, and the gains from openness. Mimeo, Pennsylvania State University.

Romer, P.M. 1990. Endogenous technical change. *Journal of Political Economy* 98(5), S71–S102.

Schmookler, J. 1966. *Invention and Economic Growth*. Cambridge, MA: Harvard University Press.

Vernon, R. 1966. International investment and trade in the product cycle. *Quarterly Journal of Economics* 80, 190–207.

How to cite this article

Eaton, Jonathan and Samuel Kortum. "trade, technology diffusion and growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000233> doi:10.1057/9780230226203.1728

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tragedy of the commons

Elinor Ostrom
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘The tragedy of the commons’ arises when it is difficult and costly to exclude potential users from common-pool resources that yield finite flows of benefits, as a result of which those resources will be exhausted by rational, utility-maximizing individuals rather than conserved for the benefit of all. Pessimism about the possibility of users voluntarily cooperating to prevent overuse has led to widespread central control of common-pool resources. But such control has itself frequently resulted in resource overuse. In practice, especially where they can communicate, users often develop rules that limit resource use and conserve resources.

Keywords

collective action; common-pool resources; common property; free rider problem; Hardin, G.; open-access resources; private property; social dilemmas; state property; tragedy of the commons

Article

The term ‘the tragedy of the commons’ was first introduced by Garrett Hardin (1968) in an important article in *Science*. Hardin asked us to envision a pasture ‘open to all’ in which each herder received large benefits from selling his or her own animals while facing only small costs of overgrazing. When the number of animals exceeds the capacity of the pasture, each herder is still motivated to add more animals since the herder receives all of the proceeds from the sale of animals and only a partial share of the cost of overgrazing. Hardin (1968, p. 1244) concluded:

> Therein is the tragedy. Each man is locked into a system that compels him to increase his herd without limit – in a world that is limited. Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hardin's article is one of the most cited publications of recent times as well as among the most influential for ecologists and environmental policy researchers. Almost all textbooks on environmental policy cite Hardin's article and discuss the problem that Hardin so graphically identified.

Hardin's article deals in general with a broad class of resources that are referred to in the more technical literature as ‘common-pool resources’. Common-pool resources yield finite flows of benefits (such as firewood, fish and water) where it is difficult and costly to exclude potential users (Ostrom, Gardner and Walker, 1994). Each person's use of a resource system subtracts resource units from the quantity of units available to others, as Hardin so dramatically described. The initial theoretical studies of common-pool resources tended to analyse simple systems. It has frequently been assumed that the resource generates a predictable, finite supply of one type of resource unit (for example, cubic feet of water or tons of fish) in each time period. Users are assumed to be short-term, profit-maximizing actors who have complete information and are homogeneous in terms of their assets, skills, discount rates and cultural views. In this theory, *anyone* can enter a resource and take resource units.

Hardin thought of users as being trapped in this situation – largely because he did not envision that users could self-organize and devise institutions to extract themselves from tragic overuse. In the conventional textbook theory (Clark, 1976), scholars have tended to agree with Hardin that the users could not extract themselves from this situation. Organizing so as to create rules that specify who is an authorized user and the rights and duties of authorized users creates a public good for those involved. All users benefit from this public good, whether they contribute or not (Olson, 1965). Thus, getting ‘out of the trap’ is itself a second-level dilemma. Since much of the initial problem exists because the individuals are in a dilemma whereby they impose negative externalities on one another, it is not consistent with the conventional theory that individuals can solve a second-level dilemma when they are already predicted to be unable to solve the initial social dilemma. Thus, extensive free-riding is predicted in most efforts to self-organize and govern a resource as a community of users.

Because of these predictions and because many open-access resources have indeed resulted in tragic levels of overuse and sometimes destruction, many scholars and public officials have relied upon the conventional analysis to justify the need for centralized control of all common-pool resources. National legislation has been passed in many countries, and administrative responsibilities for managing natural resources have been turned over to centralized agencies. Unfortunately, the results of many of these efforts have been the opposite of what was hoped. Evidence has now been amassed that central regulation has frequently accelerated resource deterioration, complicated by several problems of corruption and inefficiency. In-depth case analyses have documented the accelerated overharvesting of forests that occurred after national governments declared themselves to be the owners of forested land (National Research Council, 1986; Ascher, 1995). Similar problems have occurred with inshore fisheries when national agencies presumed that they had exclusive jurisdiction over all coastal waters (Finlayson and McCay, 1998).

Policy analysts tend to look for certainty and want to know whether the tragedy of the commons theory is either right or wrong. A more productive approach is to ask under what conditions it is correct and when it makes the wrong predictions. In settings where there is a large group, no one communicates, and where no rights to the resource exist, Hardin's theory is supported by considerable evidence. There are many settings in the world where the tragedy of the commons has occurred and continues to occur – ocean fisheries and the atmosphere being the most obvious.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Contrary to the conventional theory, however, multiple studies have demonstrated that users have overcome social dilemmas to craft institutions to govern their own resources (National Research Council, 1986; 2002; McCay and Acheson, 1987; Ostrom, 1990; 2005). The possibility, however, that the users would find ways to organize themselves was not mentioned in basic economic textbooks on environmental problems until recently (compare Clark, 1976, with Hackett, 1998). The design principles that characterize robust, long-lasting, institutional arrangements for the governance of common-pool resources have been identified (Ostrom, 1990) and supported by further testing (Guillet, 1992; Morrow and Hull, 1996; Weinstein, 2000).

A recent National Research Council (2002) report provides an excellent overview of the substantial research showing that many common-pool resources are governed successfully by non-state provision units and that some government and private arrangements also succeed. No simple governance system has been shown to be successful in all settings (Dietz, Ostrom and Stern, 2003). Many of the robust resource governance systems documented in the above-cited research do not resemble the textbook versions of either a government or a strictly private for-profit firm, especially when participants have constituted self-governing units. Scholars who draw on traditional conceptions of ‘the market’ and ‘the state’ have not recognized these self-organized systems as potentially viable forms of organization and have either called for their removal or ignored their existence. It is paradoxical that many vibrant, self-governed institutions have been wrongly classified or ignored in an era that many observers consider to be one of ever greater democratization.

Careful laboratory experiments have also shown that when a group of individuals are given unrestricted access to harvest from a common-pool resource, they substantially overuse it. What is rather striking is that in the laboratory using exactly the same parameters, but changing only one variable, namely, the capacity to communicate with one another, individuals can come to agreements and keep them to harvest very close to an optimal level (Ostrom, Gardner and Walker, 1994). This result has been replicated many times (see, for example, Casari and Plott, 2003).

Thus, Hardin opened a discourse on a fascinating and difficult puzzle of why individuals in some settings can overcome the threat to long-term sustainable use of a resource whereas other resources are so threatened. Scholars from multiple disciplines have wrestled with this question for several decades, including the creation of the International Association for the Study of Common Property (IASCP), the Scientific Committee on Problems of the Environment (SCOPE) (see Burger et al., 2001), considerable research in the field and in the experimental laboratory, and the development of sophisticated agent-based models of human-environmental relationships (Janssen, 2003).

In the decades since Hardin's article appeared, we have learned that the type of resource must be analysed separately from the type of property arrangement. Common-pool resources exist wherever natural resources or human-made facilities exist and where excluding users is costly and consumption by some subtracts from the benefits available to others. Many types of property arrangements exist in relationship to these kinds of resources, including government ownership, private property and common property. Hardin incorrectly presumed that most common-pool resources were open-access resources where property rights had not been well-defined.

It is now known that the users of a common-pool resource will:

- expend considerable time and energy devising workable institutions for governing

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- and managing common-pool resources;
- follow costly rules so long as they believe that others also follow these rules;
- monitor each other's conformity with these rules; and
- impose sanctions on each other at a cost to themselves.

The likelihood that resource users themselves will develop effective institutions for regulating the use of common-pool resources is increased by the following factors:

- low discount rates (most resource users have secure tenure, and plan on using the resource for a long time into the future);
- homogeneous interests (most resource users share similar technologies, skills, and cultural views of the resource);
- the cost of communication among individuals is low; and
- the cost of reaching binding and enforceable agreements is relatively low.

Thus, in field settings where there are relatively small- to moderate-sized groups, and where there is autonomy to make their own agreements and authority to do so, many user groups have self-organized to extract themselves from the tragedy.

Large groups have more difficulty governing common-pool resources, but usually because size is negatively associated with the factors listed above. In relatively homogeneous groups in which mechanisms exist for reaching binding agreements on methods of government and management resource use, even quite large groups are able to arrive at effective rules to limit the use of their resource. Further, when large groups are composed of smaller groups that focus on specific parts of a larger problem, such as how to regulate water distribution on a branch of an irrigation canal, smaller groups can be clustered into ever larger aggregations that may be able to address problems that affect all participants.

One of the key findings of empirical field research on collective action and common-pool resources is the multiplicity of specific rules-in-use found in successful common-pool resource regimes around the world. One of the most important types of rules is *boundary* rules, which determine who has rights and responsibilities and what territory is covered by a particular governance unit. Many different boundary rules are used successfully to control common-pool resources around the world, but an important aspect of these rules is the match between the organization of users and the resource rather than the specific rule used. The 35th anniversary of the publication of Hardin's original article was celebrated with a special issue of *Science* (Dietz, Ostrom and Stern, 2003), demonstrating that all forms of ownership could succeed or fail and that more critical than the form of ownership was the establishment of legitimate and agreed-upon boundaries that were effectively enforced.

Some governance units face considerable biophysical constraints in dealing with a natural common-pool resource such as a groundwater basin, a river or an air shed. Such resources have their own geographic boundaries, and creating a match between the boundary of those who are authorized users and the resource itself is a challenge. On the other hand, the biophysical world does not have as strong an impact on the efficacy of using diverse boundaries for governing and managing forest resources. More important is the agreement of those involved about who is to be included and the appropriate physical boundaries. Rules specifying duties as well as rules for sharing benefits are also crucial. No resource system functions well over time if all that users do is harvest from it with no investment to increase the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
productivity of the resource itself. Once basic rules – defining who is a legitimate beneficiary, who must contribute to the maintenance of the resource, and the actions that must or may be taken or are forbidden – have been accepted as legitimate by the users, many users will follow rules so long as they believe others are doing so.

Another lesson learned is that any effort to develop new rules for governing and managing complex resources is likely to generate unexpected results and be subject to initial errors. Thus, all technological and institutional interventions need to be approached as an adaptive process that helps generate information about errors so that those involved and others can learn from errors rather than continue to make them. No panaceas exist. Wholesale solutions imposed on many different resources in a large terrain are more likely to be ineffective than efforts that enhance the institutional environment that encourages responsible self-governance, self-monitoring, and self-enforcement.

Thus, a modified theory of the commons is slowly evolving that has identified the factors that are repeatedly mentioned in empirical studies of diverse common-pool resources.

See Also

- access to land and development
- collective action
- common property resources
- Olson, Mancur
- property law, economics and
- public goods

Bibliography

Ascher, W. 1995. *Communities and Sustainable Forestry in Developing Countries*. San Francisco: ICS Press.

Burger, J., Ostrom, E., Norgaard, R., Policansky, D. and Goldstein, B. 2001. *Protecting the Commons: A Framework for Resource Management in the Americas*. Washington, DC: Island Press.

Casari, M. and Plott, C. 2003. Decentralized management of common property resources: experiments with a centuries-old institution. *Journal of Economic Behavior and Organization* 51, 217–47.

Clark, C. 1976. *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*. New York: Wiley.

Dietz, T., Ostrom, E. and Stern, P. 2003. The struggle to govern the commons. *Science* 302, 1907–12.

Finlayson, A. and McCay, B.J. 1998. Crossing the threshold of ecosystem resilience: the commercial extinction of northern cod. In *Linking Social and Ecological Systems: Management Practices and Social*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mechanisms for Building Resilience, ed. F. Berkes and C. Folke. New York: Cambridge University Press.

Guillet, D. 1992. *Covering Ground: Communal Water Management and the State in the Peruvian Highlands*. Ann Arbor: University of Michigan Press.

Hackett, S. 1998. *Environmental and Natural Resources Economics: Theory, Policy, and the Sustainable Society*. Armonk, NY: M. E. Sharpe.

Hardin, G. 1968. The tragedy of the commons. *Science* 162, 1243–8.

Janssen, M., ed. 2003. *Complexity and Ecosystem Management: The Theory and Practice of Multi-Agent Systems*. Northampton, MA: Edward Elgar.

McCay, B. and Acheson, J. 1987. *The Question of the Commons: The Culture and Ecology of Communal Resources*. Tucson: University of Arizona Press.

Morrow, C. and Hull, R. 1996. Donor-initiated common pool resource institutions: the case of the Yanesha Forestry Cooperative. *World Development* 24, 1641–57.

National Research Council. 1986. *Proceedings of the Conference on Common Property Resource Management*. Washington, DC: National Academy Press.

National Research Council. 2002. *The Drama of the Commons*, Committee on the Human Dimensions of Global Change, ed. E. Ostrom, T. Dietz, N. Dolšak, P. Stern, S. Stonich and E. Weber. Washington, DC: National Academy Press.

Olson, M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Ostrom, E. 2005. *Understanding Institutional Diversity*. Princeton, NJ: Princeton University Press.

Ostrom, E., Gardner, R. and Walker, J. 1994. *Rules, Games, and Common-Pool Resources*. Ann Arbor: University of Michigan Press.

Weinstein, M. 2000. Pieces of the puzzle: solutions for community-based fisheries management from native Canadians, Japanese cooperatives, and common property researchers. *Georgetown International Environmental Law Review* 12, 375–412.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Ostrom, Elinor. "tragedy of the commons." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000193>
doi:10.1057/9780230226203.1729

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transaction costs, history of

M. Klaes
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

While the basic insight that underlies the transaction cost concept is probably as old as human reflection on economic issues itself, it became associated in the 19th century with the notion of economic friction, which was subsequently expressed as a cost. Historically, the transaction cost concept has developed from narrow interpretations typical of the monetary and general equilibrium literature towards relational interpretations, based on particular market microstructural models of how economic agents interact with each other, and finally with institutional interpretations embracing a more general analysis of economic institutions, including market and non-market forms of coordination.

Keywords

cash balances; circulation cost; Coase, R. H.; contract theory; coordination problem; firm, theory of; frictions; institutional economics; institutional transaction cost economics; law and economics; link costs; modularity; money; new institutional economics; non-market coordination; organization, economics of; productive and unproductive labour; property rights; transaction cost economics; transactions costs; transactions demand for money; Marschak, J.; Williamson, O. E.

Article

As a concept, transaction costs are used in numerous ways in economics, from simply referring to the fees charged by a financial broker to a much broader concept encompassing the comparative efficiency of alternative modes of resource allocation and economic coordination.
At the most general level, transaction costs are the costs that arise beyond the point of production of a good to effect its allocation. From there on, the literature is fragmented regarding the various facets of the concept. The distinction between production and allocation may not be meaningful in all instances. Transaction costs may or may not include transport costs, may or may not refer only to market exchange, may or may not be reduced to a single alternative category such as information costs or the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
cost of time. Some authors measure transaction costs in monetary terms, others as departures from first-best outcomes, or just on the basis of qualitative comparative rankings of feasible institutional alternatives. Indeed, whether transaction costs should be regarded as costs at all has been subject to controversy, too. Hence, the term can be and has been applied in virtually every conceivable economic and social scientific context. Its wide diffusion has been attributed to the systematic ambiguity inherent in its unqualified application (Klaes, 2001), and its usefulness for serious analysis has been questioned on these grounds (for example, Dixit, 1996, p. 35).

Although often used as a catch-all expression, thinking in transaction cost terms has nevertheless yielded a rich array of models and analytical frameworks that have helped redefine how economists look at economic exchange and coordination. Circumspect definition specific to the particular context in which one seeks to use the concept should help to avoid semantic pitfalls. Furthermore, not only are systematically ambiguous notions common in economics (Clower, 1995), they actually serve a distinct and valuable purpose in the coordination of research (Klaes, 2006). A sceptical stance towards the aggregate ambiguity of the transaction cost concept does thus not necessarily amount to a criticism of particular applications of the term, although a certain tendency can be observed outside of what has become known as a ‘new’ institutional economics (see below) towards alternative expressions such as ‘friction’ (for example, Luttmer, 1996) or ‘link costs’ (for example, Kranton and Minehart, 2001, p. 492).

Monetary, relational and institutional interpretations

The various interpretations of the transaction cost concept have traditionally been classified into two general categories, juxtaposing a narrow interpretation typically associated and compatible with a ‘neoclassical’ tradition, and a broader and more institutionally minded interpretation, located in the theory of the firm and the economics of property rights, and calling for more or less radical revisions of this tradition (Coase, 1972; cf. Dahlman, 1979; Allen, 2000). While careful analysis allows one to distinguish between more than two categorically different kinds of transaction cost, the range of extant applications of the concept is best thought of as forming a spectrum of broadening scope: (*a*) narrow interpretations typical of the monetary and general equilibrium literature; (*b*) relational interpretations that are based on particular market micro-structural conceptions or models of how economic agents interact with each other beyond the traditional economic dimensions of price and quantity signals; (*c*) institutional interpretations, which formulate transaction costs as part of a more general analysis of economic institutions, including market and non-market forms of coordination. Institutional interpretations of transaction costs are the result of applying relational interpretations of transaction costs – originally defined on the basis of exchange within markets – to non-market settings such as networks, firms or clans, with the aim of expressing the comparative economic performance of alternative institutional solutions to the coordination problem in transaction cost terms.

In monetary conceptions, transaction costs are the direct costs that an economic agent incurs when engaging in a market transaction, if we leave most or all of the microstructural details of the exchange context unspecified. At the most basic level, these costs are expressed as a reduction in the value of a transaction, technically equivalent to a transaction tax. In more developed interpretations of monetary transaction costs, these costs are conceptualized as the direct monetary costs incurred when engaging in a particular market transaction, resulting from the use of intermediary and adjunct services (brokerage,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
transport).

Relational transaction cost interpretations rely on a more explicit conceptualization than monetary transaction cost interpretations of how economic agents interact with one another when they engage in market exchange. To some extent, one may subsume economic contract theory under this part of the transaction cost literature, although explicit transaction cost conceptions as such play at best a subordinate role in those approaches (for example, Grossman and Hart, 1986; Holmstrom and Milgrom, 1994).

A relational interpretation of transaction costs has played a more pronounced role in Coase's (1937; 1960) contributions to the theory of the firm and the economics of property rights. While he himself referred to ‘marketing costs’ or the ‘costs of market transactions’ in those seminal papers and did not embrace the term ‘transaction costs’ until relatively late (Coase, 1974, p. 494), his approach to defining these costs heavily influenced the conceptualization of transaction costs in what has become known as a ‘new’ institutional economics (Eggertsson, 1990; Furubotn and Richter, 2005; Ménard and Shirley, 2005), a movement best thought of as part of a renewed interest in the institutionalist traditions in economics during the last decades of the 20th century (Rutherford, 1994). Faced with the considerable theoretical challenge of providing a comprehensive micro-structural theory of exchange, a frequent strategy in this literature has been to follow Coase in decomposing relational transaction costs heuristically according to the different steps involved in concluding a market transaction. While classifications of individual authors differ, most heuristics can be accommodated within a framework that distinguishes between: (*a*) the costs of locating and attracting potential trading partners and of pre-sale inspection; (*b*) contracting and fulfilment costs; (*c*) policing and enforcement costs.

In most instances, relational conceptions of transaction costs still proceed from a contractual (in the legal sense) and therefore market-based understanding of ‘transaction’. Institutional interpretations, by contrast, further broadened the scope of the transaction cost notion by applying it not just to contractual settings but also to alternative forms of economic coordination. The distinguishing characteristic of this third interpretation of transaction costs is not the comparative character of the underlying analysis, since both monetary and relational conceptions allow comparative assessments of alternative solutions to a given coordination problem. Rather, the difference results from the endeavour of applying Coase's ‘marketing’ costs to non-market settings, comparing market coordination alongside non-market forms within a given array of alternative institutional forms. However they are defined on the micro-structural level, once they are institutionally interpreted transaction costs therefore reflect the costs of economic coordination more generally.

Some economists regard the transaction cost concept as inseparably wedded to a framework of analysis that attempts to reduce institutional features of the economy to the core neoclassical notion of cost, thereby failing to develop conceptual tools more attuned to the complexities of economic institutions. In the eyes of these critics, building an analysis of the institutional features of economic coordination on a concept that invites its own minimization undermines the very starting point of any seriously institutional approach to economics, since a world of zero transaction costs would constitute an institutional void. In other words, one may argue that an important aspect of transaction costs is their reflection of investment in institutional capital. In a way, this point mirrors the various critiques of reducing labour to an economic cost. It has had limited impact on those who have sought to reform and expand economic analysis in the name of the economics of property rights, the field of law and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economics, transaction cost economics, the theory of the firm, the economics of organization, and the study of long-term economic change (see Coase, 1988; Williamson, 2000; Alchian, 1991; Demsetz, 1988; North, 1990; Langlois, 2002).

Origins

The basic insight that underlies the transaction cost concept, which one finds embodied in pre-historical emergence of generally accepted media of exchange, is probably as old as human reflection on economic issues itself. It is thus not difficult to identify numerous precursors to the notion of transaction costs as one finds it presently developed in the economic literature, in particular in its monetary branches. When Aristotle (1932, pp. 13–14) wrote on the origin of money for example, he observed that once villages grew and combined into city-states they would require a medium of exchange that was portable and easy to handle. He also noted that impressing a stamp on a piece of metal would help avoiding repeat measurement of its embodied value.

Aristotle's insights have been reiterated time and again whenever subsequent economic writers discussed the origin of money. Crucial steps for the further entrenchment of the concept were the depiction of the various impediments to exchange as an economic cost, followed by the crystallization of the term ‘transaction(s) costs’ itself (probably Scitovsky, 1940, p. 307; cf. Hardt, 2006), which entered the economic literature from the financial markets where it was in popular usage in the 1930s (anon., 1936).

Towards the end of the 19th century, economists tended to address the impediments to exchange as ‘frictions’ in the economic system (cf. Davidson, 1896). On the metaphorical level, this interpretation resulted from a general post-Enlightenment tendency to look at the economic system in mechanistic terms, illustrated by recurring metaphors such as Hume's (1752, II, iii, 1) ‘wheels of trade’ or Mill's (1848, III, xxvi, 1) ‘machinery of exchange’. One should not discount the economic cost of the physical friction that a medium of exchange is exposed to either (Say, 1803, I.XXI.xi, 1–3).

A prominent early attempt to describe economic friction as a cost can be found in Menger (1871, pp. 170–1), who notes that every transaction requires economic sacrifices. At the very minimum, these consist of a loss of time, but may also include transport and storage costs, sales costs, taxes, commissions, communication costs, and more generally all the costs associated with intermediaries and the monetary system. While clearly describing these various sacrifices in opportunity-cost terms, Menger does not address them explicitly as a cost and he refers to them merely to define the boundary conditions of the validity of his theory of exchange.

It was not until Marx's (1893, pp. 123–46) analysis of *Zirkulationskosten*, as the ‘costs of circulation’ that result from the continuous exchange of money into goods and back into money again, that one finds an extended conceptual analysis of the costs associated with exchange. Marx regarded these costs as ‘*faux frais*’, costs which are necessary to sustain the circulation of capital but are nevertheless unproductive in that they do not contribute to the creation of value. ‘Pure circulation costs’, the most important component of circulation cost, refers to bargaining costs, accounting costs, and, following in Say's footsteps, the costs resulting from the wear and tear of the medium of exchange. One may note in passing that Marx's discussion of the costs of exchange in terms of the classical notion of unproductive labour provides one of the most immediate links between classical political economy and modern

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
debates on, for example, the relationship between production, transaction costs, and the question whether the latter are best regarded as a cost in the first instance (for example, Barzel, 1985; Goldberg, 1985).

Conceptual development

In the 1930s, a number of authors explored the implications of ‘selling’ or ‘marketing’ costs as part of the growing literature on advertising, monopolistic competition and the theory of the firm (Braithwaite, 1928; Chamberlin, 1933; Coase, 1937). While the theory of the firm experienced a strong revival three decades later within the emerging new institutional economics (Malmgren, 1961; Cheung, 1969; Williamson, 1970; Alchian and Demsetz, 1972), it was Hicks's (1935) explanation of the holding of cash balances on the basis of the costs one incurs when converting assets into cash (Klaes, 2000) that provided the strongest immediate impetus to the coining and further differentiation of ‘transaction costs’, notably via the post-war neo-Keynesian literature, its inventory approach to the transactions demand for money, and the general question of cash balances in general equilibrium theory (Makower and Marschak, 1938; Marschak, 1950; Baumol, 1952; Tobin, 1956; Patinkin, 1965; Foley, 1970; cf. Ulph and Ulph, 1975).

The monetary economics literature proceeded largely on the basis of a narrow price-based understanding of transaction costs, although comparative issues that pointed towards broader interpretations became a pressing technical concern in its mature phase (for example, Hahn, 1973). Arrow (1965; 1969) provided a crucial conceptual link between the general equilibrium literature on transaction costs and the emerging literature in the new institutional economics by moving from a monetary to a comparative institutional interpretation of transaction costs. By contrast, the economics of property rights and the emerging field of law and economics developed an increasing variety of relational transaction cost interpretations, typically expressed in contractual terms (Coase, 1960; Alchian, 1965; Demsetz, 1967; Calabresi, 1968; Posner, 1972; Macneil, 1981). The overall thrust of the new institutional economics, notably in economic history (Davis and North, 1970; North, 1985) and through the formulation of a ‘transaction cost economics’ by Williamson (1975; 1985), has, however, been the spelling out of the details of a comprehensive institutional analysis of institutional arrangements on the basis of comparative institutional interpretations of transaction costs (see also Aoki, 2001; Langlois, 2002; Greif, 2006).

Empirical studies, even on the basis of monetary transaction costs, which one might have expected to be more amenable to direct measurement than broader interpretations, have displayed considerable divergence regarding the level of observed transaction costs. Once one moves beyond narrow monetary transaction cost conceptions, for example to account for observed bid–ask spreads or for violations of the law of one price, one is forced to delimit those dimensions of transaction costs that are not empirically accessible as such and have therefore to be inferred indirectly. In the absence of a commonly agreed empirical definition of transaction costs this will inevitably lead to results that are sensitive to the particular definition that one employs (Gould and Galai, 1974; Fama, 1991).

Alternative approaches, largely found in the comparative institutional transaction cost literature, have included attempts to infer macroeconomic transaction costs on the basis of the size of the transaction

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sector of an economy (Wallis and North, 1986), or to derive proxy measures for transaction costs on the basis of transaction characteristics such as asset specificity (Williamson, 1991; Nooteboom, 1993). The irony of sectoral measures is that economies with less well-developed transaction sectors appear to exhibit lower levels of transaction costs if those costs are measured in terms of sector size, whereas micro-structurally those economies in fact suffer from higher levels of transaction costs due to significant barriers to smooth exchange and coordination of economic activity. Proxy measures, in turn, are at risk of running into difficulty once they seek to embrace institutional transaction cost interpretations, because they have to proceed on the assumption that transactions can be defined in such a way that there is a core to them that remains unaffected across alternative institutional settings. Once one moves to non-market forms of economic coordination, this may become problematic (Masten, 1996), although the modularity literature has begun addressing this issue (Langlois, 2006).

See Also

- adjustment costs
- financial intermediation
- firm, theory of the
- law, economic analysis of
- new institutional economics
- property rights
- switching costs
- trade costs
- vertical integration

Bibliography

Alchian, A.A. 1965. Some economics of property rights. In *Economic Forces at Work*, ed. A.A. Alchian. Indianapolis, IN: Liberty Press, 1977.

Alchian, A.A. 1991. Development of economic theory and antitrust: a view from the theory of the firm. *Journal of Institutional and Theoretical Economics* 147, 232–4.

Alchian, A.A. and Demsetz, H. 1972. Production, information costs, and economic organization. *American Economic Review* 62, 777–95.

Allen, D.W. 2000. Transaction costs. In *Encyclopedia of Law and Economics*, vol. 1, ed. B. Bouckaert and G. De Geest. Cheltenham: Edward Elgar.

anon. 1936. Born of necessity. *Arcadia Tribune*, 12 September, 2. Arcadia, CA: Arcadia Publishing Company.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Aoki, M. 2001. *Toward a Comparative Institutional Analysis*. Cambridge, MA: MIT Press.

Aristotle. 1932. *Politics*, trans. H. Rackham. London and Cambridge, MA: Heinemann and Harvard University Press, 1967.

Arrow, K.J. 1965. Uncertainty and the welfare economics of medical care: reply (the implications of transaction costs and adjustment lags). *American Economic Review* 55, 154–8.

Arrow, K.J. 1969. The organization of economic activity: issues pertinent to the choice of market versus nonmarket allocation. In *Collected Papers*, vol. 2, ed. K.J. Arrow. Oxford: Blackwell, 1983.

Barzel, Y. 1985. Transaction costs: are they just costs? *Journal of Institutional and Theoretical Economics* 141, 4–16.

Baumol, W.J. 1952. The transactions demand for cash: an inventory theoretic approach. *Quarterly Journal of Economics* 66, 545–56.

Braithwaite, D. 1928. The economic effects of advertising. *Economic Journal* 38, 16–37.

Calabresi, G. 1968. Transaction costs, resource allocation and liability rules: a comment. *Journal of Law and Economics* 11, 67–73.

Chamberlin, E.H. 1933. *Theory of Monopolistic Competition*. 6th edn. Cambridge, MA: Harvard University Press, 1948.

Cheung, S.N. 1969. *The Theory of Share Tenancy*. Chicago: University of Chicago Press.

Clower, R.W. 1995. Axiomatics in economics. *Southern Economic Journal* 62, 307–19.

Coase, R.H. 1937. The nature of the firm. *Economica* NS 4, 386–405.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Coase, R.H. 1972. *Industrial Organization: A Proposal for Research*. Chicago and London: University of Chicago Press.

Coase, R.H. 1974. The choice of the institutional framework: a comment. *Journal of Law and Economics* 17, 493–6.

Coase, R.H. 1988. *The Firm, the Market, and the Law*. Chicago and London: Chicago University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dahlman, C.J. 1979. The problem of externality. *Journal of Law and Economics* 22, 141–62.

Davidson, M.G. 1896. Friction in economics. In *Dictionary of Political Economy*, vol. 2, ed. R.H.I. Palgrave. London: Macmillan.

Davis, L.E. and North, D.C. 1970. Institutional change and American economic growth: a first step towards a theory of institutional innovation. *Journal of Economic History* 30, 131–49.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57, 347–59.

Demsetz, H. 1988. *Ownership, Control and the Firm: The Organization of Economic Activity*. Oxford: Blackwell.

Dixit, A.K. 1996. *The Making of Economic Policy*. Cambridge, MA: MIT Press.

Eggertsson, T. 1990. *Economic Behaviour and Institutions*. Cambridge: Cambridge University Press.

Fama, E.F. 1991. Efficient capital markets: II. *Journal of Finance* 46, 1575–617.

Foley, D.K. 1970. Economic equilibrium with costly marketing. *Journal of Economic Theory* 2, 276–91.

Furubotn, E.G. and Richter, R. 2005. *Institutions and Economic Theory*, 2nd edn. Ann Arbor, MI: University of Michigan Press.

Goldberg, V.P. 1985. Production functions, transactions costs and the new institutionalism. In *Issues in Contemporary Microeconomics and Welfare*, ed. G. Feiwel. Albany, NY: State University of New York Press.

Gould, J.P. and Galai, D. 1974. Transactions costs and the relationship between put and call prices. *Journal of Financial Economics* 1, 105–29.

Greif, A. 2006. *Institutions and the Path to the Modern Economy*. Cambridge: Cambridge University Press.

Grossman, S.J. and Hart, O.D. 1986. The cost and benefit of ownership: a theory of vertical and lateral integration. *Journal of Political Economy* 94, 691–719.

Hahn, F.H. 1973. On transaction costs, inessential sequence economies and money. *Review of Economic Studies* 40, 449–61.

Hardt, •. 2006. Transaction cost economics as a three dimensional externally driven research program.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Studia Ekonomiczne 48–9, 7–31.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* NS 2, 1–19.

Holmstrom, B. and Milgrom, P. 1994. The firm as an incentive system. *American Economic Review* 84, 972–91.

Hume, D. 1752. Of money. In *Essays: Moral, Political, Literary*, ed. E.F. Miller. Indianapolis, IN: Liberty Fund, 1987.

Klaes, M. 2000. The history of the concept of transaction costs: neglected aspects. *Journal of the History of Economic Thought* 22, 191–216.

Klaes, M. 2001. *Begriffsgeschichte*: between the Scylla of conceptual and the Charybdis of institutional history of economics. *Journal of the History of Economic Thought* 23, 153–79.

Klaes, M. 2006. Founding economic concepts. *Storia del Pensiero Economico* NS 3, 23–39.

Kranton, R.E. and Minehart, D.F. 2001. A theory of buyer–seller networks. *American Economic Review* 91, 485–508.

Langlois, R.N. 2002. Modularity in technology and organization. *Journal of Economic Behavior and Organization* 49, 19–37.

Langlois, R.N. 2006. The secret life of mundane transaction costs. *Organization Studies* 27, 1389–410.

Luttmer, E.G. 1996. Asset pricing in economies with frictions. *Econometrica* 64, 1439–67.

Macneil, I.R. 1981. Economic analysis of contractual relations: its shortfalls and the need for a ‘rich classificatory apparatus’. *Northwestern University Law Review* 75, 1018–63.

Makower, H. and Marschak, J. 1938. Assets, prices and monetary theory. *Economica* NS 5, 261–88.

Malmgren, H.B. 1961. Information, expectations and the theory of the firm. *Quarterly Journal of Economics* 75, 399–421.

Marschak, J. 1950. The rationale of the demand for money and of ‘money illusion’. *Metroeconomica* 2, 71–100.

Marx, K. 1893. *Das Kapital*, vol. 2. Berlin: Dietz, 1953.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Masten, S.E. 1996. Empirical research in transaction cost economics: challenges, progress, directions. In *Transaction Cost Economics and Beyond*, ed. J. Groenewegen. Boston: Kluwer.

Ménard, C. and Shirley, M.M. 2005. *Handbook of New Institutional Economics*. Dordrecht: Springer.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Wien: Braumüller.

Mill, J.S. 1848. *Principles of Political Economy*, 7th edn. London: Longmans, Green, 1909.

Nooteboom, B. 1993. An analysis of specificity in transaction cost economics. *Organization Studies* 14, 443–51.

North, D.C. 1985. Transaction costs in history. *Journal of European Economic History* 14, 557–76.

North, D.C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

Patinkin, D. 1965. *Money, Interest, and Prices*, 2nd edn. New York and Tokyo: Harper & Row and Weatherhill.

Posner, R.A. 1972. *Economic Analysis of Law*. Boston: Little, Brown.

Rutherford, M. 1994. *Institutions in Economics: The Old and the New Institutionalism*. Cambridge: Cambridge University Press.

Say, J.B. 1803. *A Treatise on Political Economy*, 4th edn. trans. C. R. Prinsep, Philadelphia: Lippincott, Grambo & Co, 1855.

Scitovsky, T. 1940. A study of interest and capital. *Economica* NS 7, 293–317.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38, 241–7.

Ulph, A.M. and Ulph, D.T. 1975. Transaction costs in general equilibrium theory: a survey. *Economica* 42, 355–72.

Wallis, J.J. and North, D.C. 1986. Measuring the transaction sector in the American economy, 1870–1970. In *Long Term Factors in American Economic Growth*, ed. S.L. Engerman and R.E. Gallman. Chicago and London: University of Chicago Press.

Williamson, O.E. 1970. *Corporate Control and Business Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Williamson, O.E. 1975. *Markets and Hierarchies*. New York: Free Press.

Williamson, O.E. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

Williamson, O.E. 1991. Comparative economic organization: the analysis of discrete structural alternatives. *Administrative Science Quarterly* 36, 269–96.

Williamson, O.E. 2000. The new institutional economics: taking stock, looking ahead. *Journal of Economic Literature* 38, 595–613.

How to cite this article

Klaes, M. "transaction costs, history of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000239>
doi:10.1057/9780230226203.1730

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transfer of technology

Wolfgang Keller
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article discusses the transfer of technological knowledge with a focus on firms in different countries. The dominant form of such technology transfer is not market transactions at arm's length but technological learning externalities. The article reviews the evidence for technological learning from importing, exporting, and foreign direct investment activities.

Keywords

asymmetric information; foreign direct investment; foreign direct investment spillovers; growth and international trade; innovation; learning externalities; productivity growth; research and development; technology spillovers; trade costs; transfer of technology

Article

Asymmetric information often makes the buying, selling or licensing of technology unfeasible. Instead, international technology transfer tends to occur through non-market channels. In the United States, for example, the figures on trade in services in the balance of payments are much smaller than estimates on US technological externalities (McNeil and Fraumeni, 2005). These externalities are called technology spillovers.

A good starting point is the model of international technology transfer by Howitt (2000). There are many intermediate good sectors each characterized by its own level of technology. Different countries and sectors employ different technologies based on domestic innovation and technology transfer from abroad. At any given time *t*, the technology frontier, A_t^{max}, denotes the highest technology level across all countries and sectors. The technology frontier is growing because innovations worldwide push it out over time. An innovation in a particular sector brings this sector's productivity up to the technology frontier. This means that the model includes international and inter-sectoral technology spillovers, since if there had been many innovations in other countries or sectors, the jump to the technology frontier for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

sector *i* is larger than if there had been few innovations elsewhere.

With productivity in successfully innovating sectors jumping to the frontier level, a country's average productivity across sectors (denoted by A_t) rises. In Howitt's model, average productivity changes according to

$$\dot{A}_t = \lambda r_t (A_t^{max} - A_t), \quad (1)$$

where r_t is a measure of domestic R&D investment, and $\lambda > 0$ is a parameter. In (1), the change in A_t is positively related to domestic R&D, and it is positively related to the average technology gap $(A_t^{max} - A_t)$. Equation (1) leads to a common long-run growth rate shared by all countries; however, those investing more in R&D will enjoy relatively high productivity levels, all else being equal.

The literature emphasizes that technology transfer is facilitated by firms' international activity, though to date this has not been comprehensively articulated at the theoretical level. International trade and foreign direct investment have long been discussed as some of the most important channels, and the econometric evidence is discussed in the following (see also Keller, 2004).

First, imports may lead to technology transfer. Coe and Helpman (1995) test the prediction of the trade and growth models of Grossman and Helpman (1991) and Rivera-Batiz and Romer (1991), in which foreign R&D creates new intermediate inputs and perhaps generates spillovers for the home country through importing activity. Output is produced with labour and differentiated capital goods that enter in a constant elasticity of substitution (CES) function. The intermediate product range in each country is expanded through R&D, and countries can benefit from other countries' R&D by importing foreign-developed intermediate goods. Under certain assumptions, a country's productivity *f* is given by

$$\ln f = \ln B + \alpha \ln n^e, \quad (2)$$

where n^e is the range of intermediate goods employed in the country, and $B>0$ is a parameter. According to (2), productivity is positively related to the intermediate products range employed in this country.

A country's demand for intermediates will depend on bilateral trade barriers and transport costs. Coe and Helpman (1995) distinguish between foreign and domestic products, since trade costs are often very different between these:

$$\ln f_c = \alpha_c + \beta^d \ln n_c + \beta^f \ln n_c^f + \varepsilon_c,$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where n_c^f is defined as the bilateral import share weighted R&D of country *c*'s trade partners: $n_c^f = \Sigma_{c' \neq c} m_{cc'} r_{c'}$. This captures the prediction that if a country imports primarily from high-R&D countries, it is likely to benefit more from foreign technology than if it imports primarily from low-R&D countries.

While Coe and Helpman (1995) estimate a positive and quantitatively large effect from import-weighted foreign R&D, subsequent work by Keller (1998) shows that this per se cannot be taken as evidence for imports-related international technology transfer. Instead of using data on actual bilateral imports, Keller (1998) conducts robustness checks with two alternative foreign variables based on randomly created shares as well as no shares at all: $\tilde{n}_c^f = \Sigma_{c' \neq c} \mu_{cc'} r_{c'}$ and $\tilde{\tilde{n}}_c^f = \Sigma_{c' \neq c} r_{c'}$. Since these alternative variables yield similar or even stronger results, as in Coe and Helpman (1995), the observed import patterns cannot explain the estimated effects.

By making progress on a number of fronts, further work has produced robust evidence for imports-related technology transfer. First, Xu and Wang (1999) and Keller (2000) note that as a matter of theory, foreign technology spillovers are the result of capital goods trade, and not aggregate trade, which Coe and Helpman (1995) use to construct their import shares. Xu and Wang (1999) show that if the foreign variable n_c^f is based on bilateral capital goods trade shares, it performs better than both Coe and Helpman's (1995) original and Keller's (1998) alternative variables. Moreover, since most R&D is conducted in a relatively small part of manufacturing, imports-related technology transfer effects are relatively difficult to estimate with country-level data, as in Coe and Helpman; in a study among Organisation for Economic Co-operation and Development (OECD) countries at the two- and three-digit industry level, Keller (2002) finds robust evidence that imports are a channel for international technology transfer.

Second, the major question regarding exports is whether firms learn about foreign technology through exporting experience. There is abundant evidence that in a given cross-section exporters are on average more productive than non-exporters (Bernard and Jensen, 1999). This does not address the question whether exporting firms become more productive because of learning effects associated with exporting, or whether firms that are more productive to begin with export more. According to much anecdotal evidence, firms do benefit from interacting with the foreign customer, for instance because the latter imposes higher product quality standards than the domestic customer, while at the same time providing information on how to meet the higher standards. The econometric evidence is more mixed, however. While learning-by-exporting has been emphasized primarily for low- and middle-income countries' firms, there is in principle no reason why it is limited to these countries. Bernard and Jensen (1999) analyse learning-by-exporting using data on US firms. In studying the performance of four different sets of firms – exporters, non-exporters, starters and quitters – separately, Bernard and Jensen (1999) do not model export market participation explicitly. They find that labour productivity growth is about 0.8 per

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

cent higher among exporters than non-exporters. This estimate is fairly small, and it becomes even smaller (and insignificant) for longer time horizons. At the same time, this is conditional on plant survival. Bernard and Jensen also show that exporters are ten per cent more likely to survive than non-exporters. This difference is indicative of higher productivity growth for exporters than non-exporters, because low productivity growth is the primary reason for plants failing. Thus, there may be learning-from-exporting effects that amount to more than 0.8 per cent, although it is not clear whether they are substantial.

The paper by Clerides, Lach and Tybout (1998) provides evidence on learning externalities from exporting using micro data from Columbia, Morocco and Mexico. By estimating simultaneously a dynamic discrete choice equation that determines export market participation, these authors take into account the consideration that it is on average the already-productive firms that self-select into the export market. The export market participation decision is given by

$$y_{it} = \begin{cases} 1 & \text{if} \quad 0 \le \beta^{x} X_{it} + \beta^{e} e_{it} + \sum_{j=1}^{J} \beta_j^{c} \ln(AVC_{it-j}) \\ & \qquad + \sum_{j=1}^{J} (F^{0} - F^{j}) y_{it-j} + \eta_{it} \\ 0 & \text{otherwise,} \end{cases}$$

(4)

and any learning from exporting effects are uncovered by simultaneously estimating an autoregressive cost function

$$\ln(AVC_{it}) = \gamma_0 + \sum_{j=1}^{J} \gamma_j^{k} \ln(K_{it-1}) + \gamma^{e} \ln(e_t) + \sum_{j=1}^{J} \gamma_j^{c} \ln(AVC_{it-j}) + \sum_{j=1}^{J} \gamma_j^{y} y_{it-j} + v_{it}$$

(5)

In eq. (4), y_{it} is the export indicator of plant i in period t, X_{it} is a vector of exogenous plant characteristics, e_t is the exchange rate, AVC_{it} are average costs, K_{it} is capital, and F^0 and F^j are sunk costs of export market participation.

Equation (4) states that one only sees a plant exporting if the profits from doing so are greater than from not exporting (the latent threshold is expressed in terms of observables). Equation (5) asks whether past exporting experience reduces current cost (captured by the parameters γ_j^{y}), conditional on past costs and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
size (proxied by capital). Clerides, Lach and Tybout (1998) show results for the three countries separately, and also by major industry, using maximum likelihood (MLE) and generalized method of moments (GMM) methods. These estimations show no significant positive effects from past exporting experience on current cost. These authors' descriptive plots of average cost before and after export market entry support this conclusion. Thus, exporting does not facilitate technology transfer. According to the Clerides, Lach and Tybout's (1998) analysis, exporters are more productive, but that is because they self-select themselves into the export market.

Using similar methods, van Biesebroeck (2005) has revisited the issue by studying productivity dynamics of firms in nine African countries. In contrast to Clerides, Lach and Tybout, he estimates that the firm starting to export boosts productivity by about 25 per cent on average in his sample. Van Biesebroeck (2005) also estimates that the higher productivity growth of exporters versus non-exporters is sustained. By employing instrumental-variable and semi-parametric techniques as alternative ways to deal with the selection issue, Van Biesebroeck's analysis is more comprehensive than most. His analysis generally supports the notion that exporting leads to the transfer of technological knowledge. In trying to reconcile his findings with some of the earlier results, Van Biesebroeck shows that part of the difference in productivity growth between exporters and non-exporters appears to be due to unexploited scale economies for the latter. This suggests that at least in part his results are due to constraints imposed by demand, and not due to technology transfer in the sense of an outward shift of the production possibility frontier at all levels of production. We need richer data to make further progress on distinguishing these hypotheses.

Third, foreign direct investment (FDI) has long been considered as an important channel for technology transfer. Among the possible mechanisms are knowledge spillovers, labour turnover, linkages, and advanced specialized inputs. Also, multinational companies are well known to be more productive and do more R&D than purely domestic firms, so they are likely sources for such productivity benefits. Moreover, governments all over the world spend large amounts of resources to attract affiliates of multinationals to their jurisdiction. If this is rational economic policy, there ought to be large technology transfers associated with FDI.

Numerous studies have estimated FDI spillovers since 1970. Recently, authors focus on panel data analysis with micro data, since this reduces problems resulting from unobserved heterogeneity across firms and sectors. Typically, a general relationship between productivity growth of domestically owned firms (Δf) and a measure of the change in inward FDI (ΔFI) is specified in order to uncover evidence for FDI spillovers:

$$\Delta f_{ist} = \beta X' + \gamma \Delta FI_{ist} + u_{ist}, \quad (6)$$

Here, X is a vector of control variables, u is a regression error, and i, s and t are firm, industry and time subscripts, respectively. The spillover parameter γ is estimated positive if productivity growth of firms in industries that have experienced large increases in FDI exceeds that of firms in industries where FDI

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
has grown little.

Until about 2002, many authors concluded, by and large, that there is no evidence for FDI spillovers. This is also reflected in a number of surveys (Lipsey and Sjöholm, 2005; Görg and Greenaway, 2004; Hanson, 2001). The paper by Aitken and Harrison (1999) even estimates a *negative* relationship between FDI and productivity for a sample of Venezuelan plants. Since technology learning spillovers can hardly be negative, the analysis probably picks up something else. One possibility, first suggested by Aitken and Harrison (1999), is that the negative coefficient is due to increased competition for local plants through foreign entry. Alternatively, it could be due to endogeneity, if FDI flows to sectors in which firms are relatively weak.

The paucity of evidence has led some to look elsewhere: if there are no spillovers to domestic firms in the same industry, perhaps they exist for domestic suppliers of multinational firms? Contractual relations between foreign-owned affiliates and their domestic suppliers suggest that the technology transfer could, in principle, be specified and paid for – in which case these are not externalities. However, there could be learning effects on top of this. The paper by Smarzynska Javorcik (2004) finds evidence consistent with vertical spillovers between firms in different industries in Lithuania, but no within-industry spillovers.

Haskel, Pereira and Slaughter (2002) and Keller and Yeaple (2003) have returned to the original question of spillovers from FDI in a given industry. The former estimate an equation like (6) for FDI into the United Kingdom, and the latter for FDI into the United States. Haskel, Pereira and Slaughter (2002) estimate positive spillovers, which, however, are relatively small, as the authors note. More importantly, these authors, as is the case for Smarzynska Javorcik (2004), do not fully address the possibility that FDI inflows may be endogenous.

The first paper to show that multinationals can cause economically large productivity benefits to domestically owned firms is Keller and Yeaple (2003). These authors deal with endogeneity concerns using instrumental variable techniques, and they employ the by now well-known Olley and Pakes (1996) method of computing firm productivity. The resulting estimates imply an influence of FDI on productivity growth that is much larger than in existing studies. Using data on about 1,300 US manufacturing firms for the years 1987–96, they estimate that FDI spillovers explain about 11 per cent of productivity growth during this time.

Keller and Yeaple (2003) also reconcile their results with earlier studies that have found no evidence for FDI spillovers. For one, FDI spillovers are heterogeneous, with much stronger effects in the relatively high-technology industries. Secondly, large FDI spillovers are estimated only with the high-quality data on FDI by industry they employ. If, instead, Keller and Yeaple (2003) use FDI data similar to that more commonly available in other studies, they too estimate only a small or zero effect.

Thus, in contrast to the earlier literature, the most recent micro productivity studies tend to estimate positive, and in some cases also economically large spillovers associated with FDI. As one would expect, the effects are heterogeneous across industries, with stronger effects in relatively high-tech industries. It is not clear yet whether strong FDI spillovers occur only in relatively rich but not in relatively poor countries. Another of Keller and Yeaple's findings, that relatively weak firms benefit more from FDI than stronger firms, suggests that FDI spillovers are not limited to rich countries, where firm productivity tends to be relatively high.

To conclude, recent experience with research on the channels of technology transfer in all three areas,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
imports, exports and FDI, clearly shows that it is crucial to have access to detailed information or at least proxy variables on the technology being transferred. Given that much of technology transfer is associated with externalities, this may be the single most important issue that future work needs to address.

See Also

- countertrade
- diffusion of technology
- foreign direct investment
- trade, technology diffusion and growth

Bibliography

Aitken, B. and Harrison, A. 1999. Do domestic firms benefit from FDI? Evidence from Venezuela. *American Economic Review* 89, 605–18.

Bernard, A. and Jensen, B. 1999. Exceptional exporter performance: cause, effect, or both? *Journal of International Economics* 47, 1–26.

Clerides, S., Lach, S. and Tybout, J. 1998. Is learning by exporting important? Micro-dynamic evidence from Columbia, Mexico, and Morocco. *Quarterly Journal of Economics* 113, 903–47.

Coe, D. and Helpman, E. 1995. International R&D spillovers. *European Economic Review* 39, 859–87.

Görg, H. and Greenaway, D. 2004. Much ado about nothing? Do domestic firms really benefit from FDI? *World Bank Research Observer* 19, 171–97.

Grossman, G. and Helpman, E. 1991. *Innovation and Growth in the Global Economy*. Cambridge, MA: MIT Press.

Hanson, G. 2001. Should countries promote FDI? G-24 discussion paper series. New York and Geneva: United Nations.

Haskel, J., Pereira, S. and Slaughter, M. 2002. Does inward FDI boost the productivity of domestic firms? Working Paper No. 8724. Cambridge, MA: NBER.

Howitt, P. 2000. Endogenous growth and cross-country income differences. *American Economic Review* 90, 829–44.

Keller, W. 1998. Are international R&D spillovers trade-related? Analyzing spillovers among randomly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

matched trade partners. *European Economic Review* 42, 1469–81.

Keller, W. 2000. Do trade patterns and technology flows affect productivity growth? *World Bank Economic Review* 14, 17–47.

Keller, W. 2002. Trade and the transmission of technology. *Journal of Economic Growth* 7, 5–24.

Keller, W. 2004. International technology diffusion. *Journal of Economic Literature* 42, 752–82.

Keller, W. and Yeaple, S. 2003. Multinational enterprises, international trade, and productivity growth: firm-level evidence from the United States. Working Paper No. 9504. Cambridge, MA: NBER.

Lipsey, R. and Sjöholm, F. 2005. The impact of inward FDI on host countries: why so different answers? In *Does Foreign Direct Investment Promote Development?*, ed. T.H. Moran, E.M. Graham and M. Blomström. Washington, DC: Institute for International Economics.

McNeil, L. and Fraumeni, B. 2005. International trade and economic growth: a possible methodology for estimating cross-border R&D spillovers. BEA Working Paper 2005-03, Washington, DC.

Olley, S. and Pakes, A. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64, 1263–97.

Rivera-Batiz, L. and Romer, P. 1991. Economic integration and endogenous growth. *Quarterly Journal of Economics* 106, 531–55.

Smarzynska Javorcik, B. 2004. Does FDI increase the productivity of domestic firms? In search of spillovers through backward linkages. *American Economic Review* 94, 605–27.

Van Biesebroeck, J. 2005. Exporting raises productivity in Sub-Saharan African manufacturing firms. *Journal of International Economics* 67, 373–91.

Xu, B. and Wang, J. 1999. Capital goods trade and R&D spillovers in the OECD. *Canadian Journal of Economics* 32, 1258–74.

How to cite this article

Keller, Wolfgang. "transfer of technology." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000107> doi:10.1057/9780230226203.1731

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transfer pricing

Jack M. Mintz
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Governments establish tax rules for setting transfer prices for non-arm's length transactions made by multinationals, following guidelines established by the OECD (1995) under Article 9 of the OECD Model Tax Convention. Various methodologies have been established; the first preference is determining comparable uncontrolled prices according to the arm's length principle. Given the difficulties of achieving comparable transactions in determining price, margins or profitability, other methods, such as allocating profits among members of a corporate group according to a formula, have instead been relied upon for multi-jurisdictional corporate income taxation in some circumstances.

Keywords

arm's length prices; capital intensity; comparable uncontrolled price; competent authority; cost-plus margin; double taxation; formulary apportionment; resale-minus margin; tax treaties; transactional net margin method; transfer pricing

Article

Transfer prices are established for goods and services sold between related parties among members of a multinational group. With the growth of cross-border transactions by multinational businesses, tax authorities increasingly deal with issues related to the proper assessment of transfer prices to measure corporate income, since multinationals may use transfer prices to reduce worldwide payment of taxes by shifting income from high-tax to low-tax jurisdictions.
For example, by charging a high price for goods sold to an affiliate purchaser in a high-tax country, the purchaser's reported profits are reduced while higher profits are reported by the vendor affiliate in the low-tax country. The taxes saved by reporting higher costs in the high-tax country are in excess of the additional tax paid on extra income earned in the low-tax country. Similarly, reporting a lower price on goods sold by a company operating in a high-tax jurisdiction to related parties’ purchasers located in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
low-tax jurisdictions also reduces worldwide taxes.

Legal framework of transfer pricing

Governments enact legislation to assess profits earned by multinational companies operating in their jurisdiction based on the ‘arm's length principle’, which is a price that would be charged by unrelated businesses undertaking a transaction under similar facts and circumstances. As a basis for national legislation, assessment and adjudication, legislation may draw from the guidelines established by the Organisation for Economic Co-operation (1995) for transfer pricing. Each government is ultimately responsible for the development of its tax policy and process, and many require companies to document contemporaneously their pricing methodologies when determining their taxable profits in a jurisdiction. When a multinational is reassessed by a country for its transfer prices, it could be faced with double taxation of income if other countries choose a different transfer pricing methodology or disagree on the quantum of the inter-company transaction. Many governments have entered into tax treaties that allow corporations to seek an alternative process for double taxation relief (competent authority). In many countries, when taxpayers enter the competent authority programme for double taxation relief they may be obliged to give up their right of appeal and access to tax courts through domestic channels.

Economic value of transfer prices

Even without taxation, transfer prices would be set by multinationals with the objective to improve business performance by efficiently allocating resources amongst the various competing segments of a firm's value chain, and to appropriately reward employees and shareholders of the multinational group. The pricing strategy could therefore distort resource allocation decisions and profits along with determining managerial compensation or the amount of income earned by minority shareholders owning shares in the parent company or affiliates. To help align the interests of the managers with shareholders to maximize shareholder value and to protect the interests of minority shareholders, transfer prices may be set to reflect market conditions, including a reasonable approximation of the arm's length price that would be established between two unrelated parties. Some multinationals have kept two sets of books – one for accounting and another for tax purposes – so that control can be separated from tax motives for establishing transfer prices.

Determining the appropriate price for transfer prices is not an easy task, since prices of comparable transactions can be difficult to obtain. Many transactions involve intangible assets and intellectual property related to research and development, marketing and trademarks, which may be unique to a specific firm and embedded and intertwined with service and tangible goods transfers. Risks also need to be compensated through pricing. Costs are influenced by the size of the transaction and market conditions, including the degree of competitiveness, and regulations and consumer preferences and factor endowments also affect pricing. Finding comparable prices is a difficult matter at best, so it is not uncommon for tax authorities to challenge the transfer price methodology used by multinational companies in assessing taxable income earned in a jurisdiction.

Methods for determining transfer prices

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Countries have generally agreed upon a ranking of methodologies to determine transfer prices using the OECD guidelines (1995) although the United States has adopted a ‘best method’ approach that allows taxpayers to select the most appropriate method. The OECD guidelines identified three transactional and two profit-based methods. The transactional methods include the comparable uncontrolled price method (CUP), the resale-minus price method (RPM), and the cost-plus method (CPLM). The two profit based methods are the profit splits method (PSM) and the transactional-net-margin method (TNMM). The CUP method as observed on transactions between unrelated parties is viewed as the most appropriate methodology for measuring transfer prices for non-arm's length transactions made among members of the multinational group. However, to employ a CUP, corporations must show that the transaction that they have chosen has similar terms and conditions to those of a related-party transaction including quality and reliability, availability of volume of supply, provision of services, licensing, type and characteristics of property (patent, trademark or know-how), functions undertaken by the enterprise, business strategies, risks and market conditions.

It may be quite difficult to find CUPs that can be adjusted to reflect all potential differences. Therefore, other methodologies might be used instead.

Two other transactional-based methods include the resale-minus and cost-plus approach. The resale-minus method is typically used for distribution companies in that an inter-company price is determined by subtracting a gross profit margin from the product's resale price. An alternative method is the cost-plus method, which would be appropriate for a manufacturer to consider. A profit margin would be added to manufacturing and other costs to determine the cost-plus margin on goods and services sold from one affiliate to another in the multinational group.

With both the resale-minus and cost-plus approach, the multinational's margin on related party sales or costs can be compared with similar transactions made between unrelated parties. However, as with the CUP methodology, care must be taken to put transactions on a comparable basis. For example, companies might have different capital–labour ratios in producing goods and services. Therefore, it would not be surprising if resale-minus or cost-plus margins, reflecting the ratio of profits to costs, tend to be higher for companies relying on capital-intensive techniques of production. For comparability, an adjustment for capital intensity would be needed.

Given that corporate income is a payment made to shareholders after the deduction of borrowing costs, profit-based methods are used as an approach when it is not appropriate to use transactional approaches. The split-profit method would result in a transfer price for a transactions where both parties to a transaction own valuable and significant intangible assets. The transactional net margin method (TNMM) determines the price of an inter-company transfer of goods and services by comparing an associate affiliate's operating profits, in relation to an appropriate base (sales, costs or assets for example), to profits earned by uncontrolled firms. To establish comparability, adjustments are needed for differences in the age of capital (in part due to the distorting impact of inflation on profits) and risk.

Economic aspects of transfer pricing

When businesses choose transfer prices that vary from the ‘true’ price, they trade off the benefits of tax reduction with non-tax costs, including distorting managerial behaviour or greater expected cost of reassessment by tax authorities (Haufler, 2001). Thus, the greater the tax savings from shifting income

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
from high to low transactions, the more the transfer prices will be distorted for tax purposes. Governments may counteract transfer pricing by cutting corporate income tax rates or pursuing more aggressively transfer pricing practices of multinationals through the legal process.
With tighter transfer pricing, companies find that other income-shifting approaches for reducing worldwide taxes can be simpler, such as shifting debt, leasing, insurance, licensing and other deductible expenses to high-tax countries with income report by affiliates operating in tax havens. Transfer pricing litigation may result if the interest rates, fees and royalties charged are not justified at market rates.

Allocation or apportionment methods as a substitute

Given the constraints in assessing comparable prices, margins and profits, tax authorities will sometimes rely on other approaches to assessing corporation taxes rather than assess transfer prices to determine accounting income earned in a country. One approach is to assess a share of the worldwide income earned by a multinational group that would be allocated or apportioned to a specific jurisdiction (formulary apportionment). The share of profits apportioned to a jurisdiction could be based on one or several factors, including payroll, capital and sales as a portion of worldwide amounts. The allocation or apportionment method (Martens-Weiner, 2006) is used in some federal states, including Canada, Switzerland and the United States, to avoid the necessity of determining accounting profits using the transaction approach under separate accounting. It is also used by California to assess its corporate income tax on multinationals operating in the state (Californian income is assessed by multiplying the tax rate by income, which is a percentage of worldwide income; the percentage is determined by a formula). Some experts have advocated the use of the apportionment method at the international level. At the present time, in 2007, member states of the European Union have been debating proposals to consolidate corporate income tax bases with an apportionment method that would allocate profits to each member state.

See Also

- tax havens
- tax treaties

Bibliography

Haufler, A. 2001. *Taxation in a Global Economy*. Cambridge: Cambridge University Press.

Martens-Weiner, J. 2006. *Company Tax Reform in the European Union*. New York: Springer.

OECD (Organisation for Economic Co-operation and Development). 1995. *Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations*. Paris: OECD.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mintz, Jack M. "transfer pricing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000226> doi:10.1057/9780230226203.1732

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transfer problem

Philip Brock
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A financial transfer of wealth between countries necessitates adjustments in expenditure, production, and relative prices that collectively comprise the transfer problem. Since the 1920s the transfer problem arising from war reparations and other unrequited transfers has occupied the attention of Keynes, several Nobel laureates, and other distinguished economists. The early literature centred on static two-country models of transfers, while more recent research has highlighted the intertemporal dimension of the transfer problem.

Keywords

infinite horizons; intertemporal models; intertemporal terms of trade; overlapping generations model; reparations; representative agent; terms of trade; transfer paradox; transfer problem; Walras's Law

Article

Development of the theoretical literature on the transfer problem mirrors the historical context within which the issue first arose. In 1919, as part of the Treaty of Versailles that followed the First World War, Germany was required to make reparations payments to the European powers to which it surrendered (see Eichengreen, 1986). Initially, much discussion of Germany's capacity to pay proceeded on the basis of constant international prices and assumed that governments could automatically engineer the required changes in spending on traded goods at home and abroad. But Keynes (1929), in an article which introduced the phrase ‘transfer problem’ into the professional literature, argued that a country required to make a fixed transfer of purchasing power to another would suffer a secondary burden in the form of a further decline in its purchasing power due to an induced deterioration in its international terms of trade. Ohlin (1929) argued in response that a secondary benefit – or terms of trade improvement – was as likely to occur due to expenditure effects and the presence of non-traded goods (see especially Mundell, 2002).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ohlin's central insight can be stated simply, following Pigou (1932), using a two-country, two-commodity model of international trade. To highlight the central point, assume initially that production of both goods is exogenously given and that all income is devoted to consumption. If the markets for both commodities clear, then by Walras's Law we need only consider one, for example the good exported by the home country, where that good is denoted *x*. The total supply of *x* must equal the sum of domestic and foreign demands:

$$S + S^* = D(y, p) + D^*(y^*, p)$$

where *S* and *S** designate domestic and foreign supplies, taken as exogenous for the moment (with asterisks denoting foreign values throughout), and *D* and *D** domestic and foreign demands (each of which depends on real income and relative prices). Now assume that an amount *T* of purchasing power is transferred from the home to the foreign country. Domestic demand for *x* falls by $D_y T$ where D_y is the home country's marginal propensity to consume *x* out of income, while foreign demand for this good rises by $D^*_{y^*} T$ where $D^*_{y^*}$ is the foreign country's marginal propensity to consume *x*, also its marginal propensity to import. Equilibrium in the market for *x* at the initial prices requires that an export surplus in the amount *T* results from income effects alone – in other words, that $\left(D_y + D^*_{y^*}\right)T = T$, or $D_y + D^*_{y^*} = 1$. If $D_y + D^*_{y^*} < 1$, the combined marginal propensities to spend on *x* out of income are too small to generate a sufficient surplus at initial relative prices. This is the ‘orthodox’ case in which the transfer-making country's terms of trade deteriorate, creating a secondary burden (Samuelson, 1952; 1971). If $D_y + D^*_{y^*} > 1$, then to the contrary the transfer-making country's terms of trade improve. Once adjustments in production are introduced, the terms of trade will deteriorate when the bias in tastes in each country towards consumption of the exportable good is greater than the bias in production due to international differences in factor endowments or technologies (Jones, 1975). Transport costs, by increasing the correlation within countries of patterns of production and consumption, reinforce the orthodox presumption of a secondary burden (Samuelson, 1952). Introducing a third country adds an additional set of supply and demand elasticities and the possibility of complementarity in production and consumption. A number of cases arise – such as a ‘transfer paradox’ in which a transfer immiserizes the recipient country – whose existence is inconsistent with market stability in two-country two-commodity models (on the stability question, see Samuelson, 1947; 1971; for the case of more than two countries, see Gale, 1974; Bhagwati, Brecher and Hatta, 1983). Extensive literature reviews can be found in Eaton (1989) and Brakman and van Marrewijk (1998).

Research on the transfer problem since 1980 has moved toward intertemporal issues. The analytical tools used are the representative agent and overlapping generations frameworks in either a two-period or infinite-horizon setting. Two-country analyses (for example, Djaji•, Lahiri and Raimondos-Moller,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1998) have focused on the adjustment of the intertemporal terms of trade (the world interest rate) to a transfer. The two-country, overlapping-generations model has attracted particular interest because the transfer paradox can occur with just two countries, since the competitive equilibrium need not be Pareto efficient (see Galor and Polemarchakis, 1987; Haaparanta, 1989).

Much recent research concerns the financing of a transfer. As noted by Gavin (1992) and Devereux and Smith (2005), France borrowed an amount equivalent to almost one quarter of its GDP in order to finance its reparations payments to Germany during 1872 and 1873. In contrast with static analyses, intertemporal models allow the initial payment of a transfer to be partially financed by international borrowing (Sachs, 1981; Obstfeld and Rogoff, 1995). As a simple illustration, consider a small open economy with fixed endowment income ($\bar{y}$) and initial foreign debt (b_{t-1}) that must pay a one-period reparation (τ_t). With a given world interest rate (r), the current account at time t will be $CA_t \equiv \bar{y} - c_t - rb_{t-1} - \tau_t$. Assume perfect foresight and that consumption (c_t) can be approximated by permanent income ($\tilde{y}_t$):

$$\tilde{y}_t \equiv -rb_{t-1} + \frac{r}{1+r}\left[\bar{y} - \tau_t + \frac{\bar{y} - \tau_{t+1}}{1+r} + \ldots\right] \equiv \bar{y} - rb_{t-1} - \frac{r}{1+r}\left[\tau_t + \frac{\tau_{t+1}}{1+r} + \ldots\right],$$

where $\sum_{n=0}^{\infty}\left(\frac{1}{1+r}\right)^n = \frac{1+r}{r}$. Then the current account will be determined by the timing and magnitude of current and future reparations payments:

$$CA_t = -\tau_t + \frac{r}{1+r}\left[\tau_t + \frac{\tau_{t+1}}{1+r} + \ldots\right].$$

Given an initially balanced current account, a reparations payment in period t only will result in a current account deficit of $\left(\frac{1}{1+r}\right)\tau_t$, thereby creating a debt obligation that spreads out the necessary accompanying decline in consumption by $\left(\frac{r}{1+r}\right)\tau_t$ across all periods from time t onward. The imposition of a uniform reparations payment in all time periods $(\tau_t = \tau_{t+1} = \ldots = \bar{\tau})$ will lower consumption by $\bar{\tau}$ so that the improvement in the trade account ($TA_t \equiv \bar{y} - c_t$) exactly offsets the reparations payment in period t (a condition that is imposed in a static analysis of the transfer problem). In settings involving reproducible capital, reparations payments (or other transfers) may have an additional impact on current account dynamics by setting in motion a gradual adjustment of the capital stock to a new equilibrium (Brock, 1996; Chatterjee, Sakoulis and Turnovsky, 2003).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Beginning with Krugman (1999), the term ‘transfer problem’ has been applied to an abrupt reduction in capital flows during an economic crisis. The formal similarity between adjustment to a capital flow reduction and a reparations payment can be seen from the following trade account identity:

$$TA_t \equiv CA_t + rb_{t-1} + \tau_t$$

where $CA_t \equiv b_{t-1} - b_t$. If we hold the current account constant, an increase in reparations payments ($\tau_t \uparrow$) requires an accompanying improvement in the trade account. A reduction of capital flows ($b_t \downarrow$) requires a similar improvement in the current and trade accounts. However, since there is a binding borrowing constraint and no direct wealth effect associated with an abrupt reduction in capital flows, this second transfer problem is conceptually distinct from the classical transfer problem, thus supporting Nurkse's (1961) admonition not to apply indiscriminately the analysis of an unrequited transfer to problems involving international capital movements.

Despite the large literature on the static and intertemporal dimensions of unrequited transfers, there is relatively little empirical research on the transfer problem. Papers by Yano and Nugent (1999), Lane and Milesi-Ferretti (2004), Devereux and Smith (2005), and Rajan and Subramanian (2005) are notable promising exceptions.

See Also

- Keynes, John Maynard
- Ohlin, Bertil Gotthard
- terms of trade

This is a revision and extension of the article in the first edition of the New Palgrave dictionary by Barry Eichengreen.

Bibliography

Bhagwati, J.M., Brecher, R.A. and Hatta, T. 1983. The generalized theory of transfers and welfare. *American Economic Review* 73, 606–18.

Brakman, S. and van Marrewijk, C. 1998. *The Economics of International Transfers*. Cambridge, UK: Cambridge University Press.

Brock, P. 1996. International transfers, the relative price of non-traded goods, and the current account. *Canadian Journal of Economics* 29, 163–80.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Chatterjee, S., Sakoulis, G. and Turnovsky, S. 2003. Unilateral capital transfers, public investment, and economic growth. *European Economic Review* 47, 1077–103.

Devereux, M. and Smith, G. 2005. Transfer problem dynamics: macroeconomics of the Franco-Prussian war indemnity. Working Paper No. 1025, Department of Economics Queens University, Kingston, Ontario, Canada.

Djaji•, S., Lahiri, S. and Raimondos-Moller, P. 1998. The transfer problem and the intertemporal terms of trade. *Canadian Journal of Economics* 31, 327–36.

Eaton, J. 1989. Foreign public capital flows. In *Handbook of Development Economics*, vol. 2, ed. H. Chenery and T.N. Srinivasan. Amsterdam: North-Holland.

Eichengreen, B. 1986. Macroeconomics and history. In *The Future of Economic History*, ed. A. Field. Boston: Martinus Nijhoff.

Gale, D. 1974. Exchange equilibrium and coalitions. *Journal of Mathematical Economics* 1, 63–6.

Galor, O. and Polemarchakis, H. 1987. Intertemporal equilibrium and the transfer paradox. *Review of Economic Studies* 54, 147–56.

Gavin, M. 1992. Intertemporal dimensions of international economic adjustment. *American Economic Review* 82, 174–79.

Haaparanta, P. 1989. The intertemporal effects of international transfers. *Journal of International Economics* 26, 371–82.

Jones, R. 1975. Presumption and the transfer problem. *Journal of International Economics* 5, 263–74.

Keynes, J.M. 1929. The German transfer problem. *Economic Journal* 39, 11–17.

Krugman, P. 1999. Balance sheets, the transfer problem, and financial crises. *International Tax and Public Finance* 6, 459–72.

Lane, P. and Milesi-Ferretti, G. 2004. The transfer problem revisited. *Review of Economics and Statistics* 86, 841–57.

Mundell, R.A. 2002. Keynes and Ohlin on the transfer problem. In *Bertil Ohlin: A Centennial Celebration (1899–1999)*, ed. R. Findlay, L. Jonung and M. Lundahl. Cambridge, MA: MIT Press.

Nurkse, R. 1961. Causes and effects of capital movements (1933). In *Equilibrium and Growth in the*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

World Economy, ed. G. Haberler and R. Stern. Cambridge, MA: Harvard University Press.

Obstfeld, M. and Rogoff, K. 1995. The intertemporal approach to the current account. In *Handbook of International Economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: North-Holland.

Ohlin, B. 1929. The reparation problem: a discussion. *Economic Journal* 39, 172–83.

Pigou, A.C. 1932. The effects of reparations on the ratio of international exchange. *Economic Journal* 42, 532–43.

Rajan, R. and Subramanian, A. 2005. What undermines aid's impact on economic growth? Working Paper No. 05/126, International Monetary Fund, Washington, DC.

Sachs, J. 1981. The current account and macroeconomic adjustment in the 1970s. *Brookings Papers on Economic Activity* 1981(1), 201–82.

Samuelson, P. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P. 1952. The transfer problem and transport costs. *Economic Journal* 62, 278–304.

Samuelson, P. 1971. On the trail of conventional beliefs about the transfer problem. In *Trade, Balance of Payments and Growth*, ed. J. Bhagwati et al. Amsterdam: North-Holland.

Yano, M. and Nugent, J. 1999. Aid, nontraded goods, and the transfer paradox in small countries. *American Economic Review* 89, 431–49.

How to cite this article

Brock, Philip. "transfer problem." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000111> doi:10.1057/9780230226203.1733

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transition and institutions

Peter Murrell
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article examines the evolution of institutions and economists' thinking on institutions during transition. Early in transition, institutions were virtually ignored in the majority of normative prescriptions, but were central in the evolutionary institutional approach. Later, after events influenced intellectual developments, institutions were at the centre of analysis. Growth is strongly related to institutional construction. Transition countries built institutions speedily but with marked variation across countries. Legal systems and independent governmental agencies were sources of institutional growth, while government bureaucracies and informal mechanisms detracted from institutional growth. In China, reforms addressed problems that institutions usually do, but in unusual ways.

Keywords

China, economics in; colonialism; corruption; creative destruction; dual track liberalization; evolutionary-institutional view; institutional development; institutions; liberalization; market institutions; privatization; resource curse; rule of law; shock therapy; social capital; stabilization; transition

Article

'Transition' is the widely accepted term for the thoroughgoing political and economic changes that followed the fall of communism in Eastern Europe (EE) and the Soviet Union. Some 29 countries are involved in this continuing process, which began in 1989–91 and involves the types of transformations that usually took a century or more in today's developed countries. A related, but distinct, process has been under way in China since 1978.

Transition has been coterminous with a remarkable change in emphasis within economics. In 1989, to highlight the importance of institutions was a distinctly minority activity. Now, institutions are at the heart of both research and policy discussions. Similarly, while many early influential analyses of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
transition virtually ignored institutions, current discussions place them at the centre. Developments in transition countries made an important contribution to general trends within economics (Roland, 2000).

This article focuses on institutions in the transition process and in economists' deliberations on that process. It begins with early normative prescriptions, in which institutions were virtually ignored by the majority of contributors, and then examines changing views on institutions, showing how events on the ground influenced intellectual developments. We then provide basic facts on institutional development, describing the impressive progress that has been made, which suggests modification of the standard assumption that institutional construction must be slow. Nevertheless, there is marked divergence across countries. This article examines the sources of institutional growth and the 'great divide' between the successful and the unsuccessful institution builders (Berglof and Bolton, 2002). It closes by considering the seemingly anomalous case of China, showing that the anomaly is more apparent than real. China's reforms addressed the problems that institutions address everywhere, but in ways that are not recognizable to those using a first-best institutional template.

Ideas and institutions in the earliest phase of transition

Economists began to deal with the transition with a deluge of normative prescriptions. The majority view in its most stark incarnation came to be called shock therapy, the notion that the best way forward was as fast as possible on all fronts, taking advantage of a political window of opportunity. These types of reforms were certainly the aim of the first post-communist governments and their Western advisers in Poland, Russia and many other countries. According to shock therapy's proponents, the soon-to-be-observed dissonance between objectives and follow-through was variously due to the absence of a clear vision, lack of willpower, and a nefarious political opposition.

Institutions were ignored within the shock therapy approach for a variety of reasons. They could be built so easily that they did not require much attention (Sachs, 1991). They were not deemed important enough to mention (Blanchard et al., 1991). They would take so long to develop that other elements of policy came first (Fischer and Gelb, 1991). Or, they could not be built without first creating the actors who would demand them within the political process (Boycko, Shleifer and Vishny, 1995).

In shock therapy analysis, political economy considerations led to emphasis on the destruction of the old institutions and trumped any concerns about the dangers of an institutional void. Macroeconomics governed microeconomic institutional change, as exemplified by the International Monetary Fund's short-term focus on raising taxes in Russia, while largely ignoring sensible tax reforms (Black, Kraakman and Tarassova, 2000). Rapid liberalization was advocated, while downplaying its effects on the governance of contractual relations. The transaction costs of ownership change and corporate governance after privatization were deemed of secondary importance.

When economic performance in the early years of transition proved disappointing, diagnoses followed the earlier analyses: strong, but necessary, stabilization programmes had led to recessions (Blanchard, Froot and Sachs, 1994); that is, in the early 1990s, the most influential analyses did not associate the steep, sometimes catastrophic, recessions with institutional problems. For example, such analyses led to the conclusion that liberalization, privatization and stabilization should move even faster in Russia in 1992 than they had in Eastern Europe two years earlier.

Kogut and Spicer (2004; 2005) use numerical citation analyses to analyse patterns in the early

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economics literature on transition. They document the links among a core group of economists subscribing to the shock therapy approach. This group had strong connections to the international financial institutions and the US government, and were able to spread their views in reforming countries under the auspices of these powerful organizations. Kogut and Spicer also identify dissenters from this point of view, in particular Murrell (1992), Dewatripont and Roland (1992), and McKinnon (1991).

Early in transition the dissenting view was labelled evolutionary or gradualist, but was later given the much more felicitous name, evolutionary-institutional (Roland, 2000).

The evolutionary-institutional view emphasized the importance of institutions, suggesting, for example, that the nature and timing of liberalization, privatization and stabilization depended critically on the existing institutional framework. Some institutions were prerequisites for a functioning market economy, and the absence of these might necessitate the slowing of reforms. Because new market economy institutions were hard to create, it might be better to use crude second-best institutions, even some of the old ones, while maintaining a focus on building new ones. This might lead to a two-sector approach, where a nascent private sector was governed by new institutions, while some of the old mechanisms of governance prevented convulsions in the old state sector, negatively affecting the development of the new capitalism. This approach was particularly congenial for those who thought that the growth of the new private sector was crucial (Kornai, 1990) or whose advice reflected elements of Chinese reforms (McMillan and Naughton, 1992).

The suggestion that economic reform should be gradual was a conclusion, rather than a starting point. It grew out of analyses that were standard in the literature (North, 1990). Because ideas and organizations adapt to an institutional framework, there is no certainty of an immediate functional response to new institutions. Difficulties in creating new institutions suggest a wariness of quick reforms when their success depends on functioning institutions. Instead, a nascent private sector produces the most nimble response in a new environment of fast-changing institutions.

Evolving ideas on the role of institutions

Events changed ideas. All transition countries experienced deep recession. Recovery began after several years, with its inception unrelated to any specific policy initiatives. If anything, recovery began on retreat from the earlier policies. The degree of adherence to standard policy advice could not explain the cross-country pattern of recession and growth.

Although these facts were consistent with the evolutionary-institutional view, the most influential contribution in changing the terms of debate was a paper co-authored by one of shock therapy's main proponents (Blanchard and Kremer, 1997). Undoubtedly, this paper had such a large effect because one of its authors was an influential economist who had previously attributed little importance to institutions (see Blanchard et al., 1991; Blanchard, Froot and Sachs, 1994; and the review of the latter in Murrell, 1995). The paper formalizes ideas already present in the earlier evolutionary-institutional literature in a simple, but powerful model. The model highlights the incentives to break agreements in the absence of effective governance, leading to a loss of production. Output decline comes later but is larger where the complexity of old production relations is greater. If opportunities improve over time, the model generates a U-shaped path for production. These predictions are consistent with the comparative profiles of growth in the transition countries, with recession initially steeper but ultimately shallower in Eastern

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Europe than in the former Soviet Union (FSU).

There is much to learn about the relationship between institutional change and production decline, but there is general agreement in some areas. Pre-transition institutions contributed to enterprise productivity. These institutions offered credibility in the negotiating of agreements, contract enforcement, specification of control rights over assets, mechanisms for the generation and allocation of working and investment capital, and many other services. When the communist systems fell apart and market institutions were still on the drawing board, these crucial services were no longer supplied. The lack of institutional support was particularly critical at the beginning of transition for several reasons: socialist firms were large, implying a need for sophisticated governance mechanisms; inter-firm relationships were highly particularized, implying great potential for hold-up problems; and necessary adjustments were enormous, implying the need for effective financial markets.

Even without effective institutions, production rebounded due to the spontaneous growth of private sector opportunities. Nascent small businesses could take advantage of these opportunities if they received a minimal amount of institutional support, that is, protection from extreme criminality, prevention of civil chaos, and the benign neglect of the state. Such businesses develop their own self-enforcing agreements and do not need sophisticated courts or contract law. Physical possession solves many concerns about property rights. Closely held firms that are self-financed do not need corporate governance institutions.

But to rebound from recession is not the same as sustained growth. The latter requires more than the benign neglect of the state: it requires a set of institutions that support non-self-enforcing agreements, secure property without possession, enable firms to expand beyond the limits of self-finance, and undertake many other activities that are not feasible without effective rules of the game. While such ideas seem commonplace now, they were not to the fore in the debates at the start of transition, except in the evolutionary-institutional perspective.

In addition to the institutional interpretation of the causes of collapse and recovery, two further factors contributed to economists' changing views. First, econometric studies showed that differences in the application of the standard policies did not explain differences in economic performance (for example, de Melo et al., 2001; Falcetti, Raiser and Sanfey, 2002). Second, variations in performance became more noticeable in the trajectories out of recession. Countries appeared to be sorting themselves into two groups. Those in EE were generally performing better than those in the FSU, but there were enough exceptions (for example, the Baltics, Serbia) to suggest that the EE–FSU distinction was not the key. As Berglof and Bolton (2002, p. 77) noted, 'A growing and deepening divide has opened up between transition countries where economic development has taken off and those caught in a vicious cycle of institutional backwardness and macroeconomic instability'.

Beck and Laeven (2005) were the first to test this new institutional paradigm of growth in transition in a rigorous framework, although their study is naturally characterized by a paucity of data points. They find that there is very large divergence in the performance of transition countries and that institutional development is the key factor in explaining the divergence. Moving from Russia's level of institutional development in 1996 to Poland's level would lead to a growth rate increase of 4.4 per cent a year. In contrast, differences in policies are unimportant. Papers studying privatization, agricultural markets and foreign direct investment contain results on economic performance at a more disaggregated level that complement those of Beck and Laeven. While relative neglect of institutions characterized the early stages of transition, the centrality of institutions is now conventional wisdom.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Evolving institutions

One reason why institutions were not emphasized in early transition was the widely held assumption that institutional construction would be very slow. The transition countries provided an ideal testing ground for this assumption. Having rejected a set of old institutions and turned to creating new ones, how fast and how successful could institutional construction be? The answer, for some countries only, is: surprisingly quick and successful. Transition experience refutes one element of conventional wisdom, that institutional development is inevitably very slow, while bolstering another, that failure is commonplace.

Murrell (2003) concluded that there had been widespread, large, continuing improvements in institutional quality from 1990 to 2000. An updating especially for this article extends this analysis to 2004, using the popular institutional measures developed by Kaufmann, Kraay and Mastruzzi (2005). This updating shows that institutional scores for transition countries as a whole are no better and no worse than one would expect given levels of economic development. This is remarkable, since it implies that in less than 15 years the transition countries built institutions that match those in countries that have had capitalist systems for much longer. For example, on the rule of law, Hungary, Slovenia and Estonia are comparable to Chile, Israel, Greece, Italy, Spain and Taiwan. On regulatory quality, Estonia ranks above Sweden, while Hungary, Lithuania, Slovakia, Latvia and the Czech Republic are grouped with the United States, Japan, Italy and Spain.

These results, which are based on expert opinions and surveys, are supported by studies examining the micro details of institutional development. Djankov et al. (2002) collect data on highly specific aspects of the functioning of legal systems, such as collecting on a bad cheque. They find that the ex-socialist countries fare better than both French-legal-origin and German-legal-origin countries. Pistor, Raiser and Gelfer (2000) examine the quality of laws on shareholder and creditor rights, finding the transition countries superior to many developed economies.

The second distinctive feature in institutional development is the divergence between one group of countries whose institutions are at a comparatively high level and improving and another group that has not crossed the great divide and is even losing some of the gains from the 1990s. By 2004, the EE–Baltic group has institutional scores higher than expected, given general levels of economic development on all six of the Kaufmann, Kraay and Mastruzzi (2005) indicators, and these scores improved dramatically in the preceding decade. The Commonwealth of Independent States (FSU minus the Baltics) scores below expected levels and has been regressing from 1996 to 2004, after showing remarkable signs of institutional improvement in the early 1990s.

It is difficult to exaggerate the importance of this empirical evidence on basic hypotheses on institutional development. Before transition, the assumption was that modern institutional development is a very long process, fraught with the possibility of failure. The first element of this assumption has been refuted. In the years before 1990, capitalism and democracy were absent in EE and the USSR. Then there was a mammoth fall in national income, due to institutional lacunae. Yet now a large group of countries seems set on the road to sustainable institutional development. In contrast, the second element of standard assumptions has been verified. In a significant number of transition countries, slow initial progress on institutional development has been followed by severe regression.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The sources of institutional development

There are two alternative perspectives to take when viewing the sources of institutional development. First, one can analyse which country-level factors best explain aggregate institutional outcomes. Second, one can ask which particular mechanisms or organizations inside a country contributed most to institutional performance. Evidence on both is only currently being generated, and is very scant. This is true both of transition and in general.

Beck and Laeven (2005) have carried out the most systematic study of the causes of aggregate institutional development in transition. They find two principal determinants: the strength of the incumbent socialist elite and the importance of natural resources (the resource curse). Both are negatively related to improvement in institutions. They also confirm the analysis of Black, Kraakman and Tarassova (2000) that certain types of privatization might have been inimical to institutional development. Early macroeconomic policies, belonging to the FSU and being eligible for the European Union do not affect institution building. These negative results are important since they reject prominent hypotheses. One popular theory not explored by Beck and Laeven is that colonial heritage might have influenced institutional development, particularly in the case of countries influenced by the Austrian, Ottoman or Russian empires.

One can also ask which particular mechanisms contributed most to institutional performance (Murrell, 2003). Formal institutions have played a more beneficial role than informal institutions, such as culture or pertinent elements of social capital. Of the formal institutions, political and legal structures and independent governmental agencies contributed relatively more to institutional development. State administrative bodies detracted from institutional performance, changing slowly and contributing to relatively high levels of corruption. These facts are generally consistent with the old Schumpeterian message of creative destruction, but applied to non-market organizations.

One very surprising feature of transition is the relatively strong role of some legal institutions. A series of empirical observations on the courts suggests a divergence from prevailing views on the role of the legal system (see the essays in Murrell, 2001 for example). The legal system has never been identified as playing a strong role in developing countries, and transition was not conducive to the effectiveness of the law. Yet current evidence suggests that it is easier than usually assumed to fashion a legal system that facilitates economic processes, even when that system is far from the standards of developed countries.

Institutions in the Chinese reforms

On the surface, Chinese reforms might be cast as a refutation of the above. China began its reforms with a basic constraint on institutional change – movement from the existing system could not be too great or too fast. This meant that new institutions would not be best practice, but had to be incremental variations on existing ones. Hence, China does not fare well when matched against standard criteria for judging institutions. This stands in contrast to the astounding success of the Chinese economy.

Nevertheless, China's reforms can be interpreted as bolstering the basic conclusion of the centrality of institutions in transition. China created successful, transitional institutions (Qian, 2003). By experiment,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
by confining itself to incremental changes that could be easily understood, and by implementing Pareto-improving changes in the early years of reform, China pursued a deft, but previously untrodden path of institutional change.

Qian (2003) provides examples of these transitional institutions. China implemented a dual-track approach to liberalization, which led to markets in above-plan production, but kept quotas and controlled prices on the levels of production that had existed before reforms. This promoted efficiency at the margin, while endorsing the existing set of informal rights to infra-marginal production, thus protecting the welfare of those who otherwise might have lost heavily from reforms. A highly distinctive ownership form appeared, township and village enterprises (TVEs), which played a significant role in China's growth in the first two decades of reform. TVEs can be interpreted as a mechanism for protecting decentralized property rights when the state is unable to guarantee more formal ones for private owners. Anonymous banking served as a commitment device, limiting government predation by reducing information flows. This arrangement can be understood as a crude substitute for the protection of financial property rights when the independence of the legal system is not a real possibility in the short-run.

Therefore, China constructed mechanisms to address the problems that institutions address in successful countries. However, those mechanisms would not look familiar when matched against best practice in developed countries. As the evolutionary-institutional perspective emphasized, it is fruitless to try to imitate best practices when human capital and institutional capability are not sufficient. In such situations, it might be best to deploy a set of transitional institutions, much more suited to the particular circumstances of a country and its capabilities. This observation resonates with the experience of EE and the FSU reviewed above. Countries with less benign starting points for creating best-practice institutions were doomed to fail in the process, while others could succeed given the right political and human capital preconditions.

Of course, there was a reason why in early transition best-practice institutions were advocated for all countries. It was feared that a country might find it hard to replace transitional institutions once they were set in place, becoming trapped at a low level of development. Whether this fear was ultimately justified will be addressed by Chinese experience in the coming decades.

See Also

- command economy
- dual track liberalization
- Great Divide
- market institutions
- new institutional economics
- output fall – transformational recession

Bibliography

Beck, T. and Laeven, L. 2005. Institution building and growth in transition economies. Policy Research

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Working Paper 3657. Washington, DC: World Bank.

Berglof, E. and Bolton, P. 2002. The great divide and beyond: financial architecture in transition. *Journal of Economic Perspectives* 16(1), 77–100.

Black, B., Kraakman, R. and Tarassova, A. 2000. Russian privatization and corporate governance: what went wrong? *Stanford Law Review* 52, 1731–808.

Blanchard, O., Dornbusch, R., Krugman, P., Layard, R. and Summers, L. 1991. *Reform in Eastern Europe*. Cambridge, MA: MIT Press.

Blanchard, O., Froot, K.A. and Sachs, J.D. 1994. *The Transition in Eastern Europe*. Chicago and London: University of Chicago Press.

Blanchard, O. and Kremer, M. 1997. Disorganization. *Quarterly Journal of Economics* 112, 1091–126.

Boycko, M., Shleifer, A. and Vishny, R. 1995. *Privatizing Russia*. Cambridge, MA: MIT Press.

de Melo, M., Denizer, C., Gelb, A. and Tenev, S. 2001. Circumstances and choice: the role of initial conditions and policies in transition economies. *World Bank Economic Review* 15, 1–31.

Dewatripont, M. and Roland, G. 1992. The virtues of gradualism and legitimacy in the transition to a market economy. *Economic Journal* 102, 291–300.

Djankov, S., Porta, R.L., Lopez-de-Silanes, F. and Shleifer, A. 2002. Courts: the Lex Mundi project. Working Paper No. 8890. Cambridge, MA: NBER.

Falcetti, E., Raiser, M. and Sanfey, P. 2002. Defying the odds: initial conditions, reforms, and growth in the first decade of transition. *Journal of Comparative Economics* 30, 229–50.

Fischer, S. and Gelb, A. 1991. The process of socialist economic transformation. *Journal of Economic Perspectives* 5(4), 91–105.

Kaufmann, D., Kraay, A. and Mastruzzi, M. 2005. *Governance Matters IV: Governance Indicators for 1996–2004*. Washington, DC: World Bank.

Kogut, B. and Spicer, A. 2004. Critical and alternative perspectives on international assistance to post-communist countries: a review and analysis. The World Bank, Operations Evaluation Department. Online Available at http://www.worldbank.org/oed/transitioneconomies/docs/literature_review.pdf, accessed 31 May 2007.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kogut, B. and Spicer, A. 2005. *Taking Account of Accountability: Academics, Transition Economics, and Russia*. Paris: INSEAD.

Kornai, J. 1990. *The Road to a Free Economy*. New York: Norton.

McKinnon, R. 1991. *The Order of Economic Liberalization: Financial Control in the Transition to a Market Economy*. Baltimore, MD: Johns Hopkins University Press.

McMillan, J. and Naughton, B. 1992. How to reform a planned economy: lessons from China. *Oxford Review of Economic Policy* 8, 130–43.

Murrell, P. 1992. Evolution in economics and in the economic reform of the centrally planned economies. In *Emerging Market Economies in Eastern Europe*, ed. C.C. Clague and G. Rausser. Cambridge, MA: Basil Blackwell.

Murrell, P. 1995. The transition according to Cambridge, Mass. *Journal of Economic Literature* 33, 164–78.

Murrell, P. 2001. *Assessing the Value of Law in Transition Economies*. Ann Arbor: University of Michigan Press.

Murrell, P. 2003. The relative levels and the character of institutional development in transition economies. In *Political Economy of Transition and Development: Institutions, Politics and Policies*, ed. N. Campos and J. Fidrmuc. Boston, Dordrecht and London: Kluwer.

North, D. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.

Pistor, K., Raiser, M. and Gelfer, S. 2000. Law and finance in transition economies. *Economics of Transition* 8, 325–68.

Qian, Y. 2003. How reform worked in China. In *In Search of Prosperity: Analytic Narratives on Economic Growth*, ed. D. Rodrik. Princeton: Princeton University Press.

Roland, G. 2000. *Transition and Economics: Politics, Markets, and Firms*. Cambridge, MA: MIT Press.

Sachs, J. 1991. Poland and Eastern Europe: what is to be done? In *Foreign Economic Liberalization: Transformation in Socialist and Market Economies*, ed. A. Koves and P. Marer. Boulder, CO: Westview Press.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Murrell, Peter. "transition and institutions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_I000264>
doi:10.1057/9780230226203.1734

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transitivity

Wayne Shafer
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

convexity; integrability; law of demand; principle of persistent nonpreference; strong axiom of revealed preference; transitivity; weak axiom of revealed preference

Article

Transitivity is formally just a property that a binary relation might possess, and thus one could discuss the concept in any context in economics in which an ordering relation is used. Here, however, the discussion of transitivity will be limited to its role in describing an individual agent's choice behaviour. In this context transitivity means roughly that if an agent chooses *A* over *B*, and *B* over *C*, that agent ought to choose *A* over *C*, or at least be indifferent. On the surface this seems reasonable, even ‘rational’, but this ignores how complicated an agent's decision making process can be. For an excellent discussion of this issue see May (1954). Given a model of agent behaviour, transitivity can be imposed as a direct assumption, or can be an implication of the model for choice behaviour. The standard model of agent behaviour in economics is that the agent orders prospects by means of a utility function, which in effect assumes transitivity. With appropriate continuity and convexity restrictions on utility functions, the model allows one to demonstrate that: (1) Individual demand functions are well defined, continuous, and satisfy the comparative static restriction, the strong axiom of revealed preference (SARP). (In the smooth case, this corresponds to the negative semi-definiteness and symmetry of the Slutsky matrix.) (2) Given a finite collection of such agents with initial endowments of goods, a competitive equilibrium exists. What will be discussed in the remaining part of this article is to what extent one can obtain results analogous to (1) and (2) above while using a model of agent behaviour which does not assume or imply transitive behaviour. To keep the discussion as simple as possible, we will only consider the situation in which the agent's set of feasible commodity vectors is the non-negative orthant of *n*-dimensional Euclidean space, and the agent's problem is to choose a commodity vector *x* when faced with positive prices and income. A vector *p* in the positive orthant of Euclidean *n*-space will denote the vector of price-income ratios, or a ‘price’ system.

Two models of agent behaviour which have a long history in economics will now be described. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
first, which will be called the ‘local’ theory, takes as its primitive the assumption that if an agent is currently consuming at a vector *x*, he is able to determine if an infinitesimal change in *x*, say *x* to *x*+*dx*, is a change for better or worse. This idea is represented by a function $x \to g(x)$, mapping each vector *x* into a *n*-vector *g*(*x*) such that a small movement from *x* in the direction of *y* is an improvement if $g(x)(y-x) > 0$, and not an improvement otherwise. Given a price system *p*, an affordable *x* is an equilibrium point for the agent if $g(x)(y-x) \leq 0$, for all affordable *y*. That is, no small movement from *x* in the direction of an affordable *y* is an improvement. A basic question is whether for every *p*, an equilibrium *x* exists. This approach goes back at least to Pareto, and most economists, including Pareto, concerned themselves with the ‘integrability’ problem: when is there a quasi-concave utility function such that *g*(*x*) is a positive scalar multiple of the vector of marginal utilities, for each *x*? Note that if an agent has a differentiable quasi-concave utility function u, and one defines *g* by $g(x) = \lambda(x)Du(x)$ for any $\lambda(x) > 0$ then $g(x)(y-x) > 0$ is equivalent to *Du*(*x*)(*y*–*x*)>0, and this implies *u*(*x*+*t*(*y*–*x*))>*u*(*x*) for *t* positive and sufficiently small. Thus if *g* is ‘integrable’, the agent acts as if he maximizes a utility function, and thus results (1) or (2) above will be satisfied. Some economists, however, believed that the local theory could be used to describe agent behaviour without assuming the integrability conditions. Most notable is the work of Allen (1932), Georgescu-Roegen (1936; 1954) and Katzner (1971). Without the integrability conditions and the implied utility function (and thus implied transitivity) the existence of an equilibrium *x* given any *p* is nontrivial. This problem was solved by Georgescu-Roegen (1954), who showed that if *g* is continuous and *g* satisfied the ‘principle of persistent nonpreference’ (PPN), that is, *g*(*x*)(*y*–*x*)<0 implies *g*(*y*)(*x*–*y*)>0, then an equilibrium point will exist in any budget set. It should be noted that the integrability problem mentioned above requires PPN (for quasi-concave utility), as well as the Frobenius conditions for mathematical integrability. It is easy to show that Georgescu-Roegen's assumptions imply that the resulting demand correspondence will be upper-hemicontinuous.

The second basic approach to modelling agent behaviour will be called the ‘global’ theory. In this approach, the primitive of the theory is a binary relation *R* on the commodity space with *xRy* having the interpretation ‘*x* is at least as good as *y*’. Define the strict preference relation *P* by *xPy* is equivalent to not *yRx*. (*P* could also be taken as the primitive.) Given a price system *p*, an affordable *x* in an equilibrium point if *yPx* implies *py*>1, that is, any vector *y* preferred to *x* is not affordable. A basic question is whether such an equilibrium point will exist. This approach dates back to Frisch (1926), and the usual approach was to specify conditions on *R* which imply *R* has a representation by a continuous utility function, that is, $u(x) \geq u(y)$ equivalent to *xRy*. This problem was solved by Debreu (1954), who showed that *R* must be reflexive, complete, transitive and continuous. With the addition of appropriate convexity conditions, this approach yields the results (1) and (2) above. However, in a remarkable paper, Sonnenschein (1971) showed that one could remove transitivity from the list of standard assumptions and still have a well defined demand correspondence which is upper-hemicontinuous. Specifically, he demonstrated that if *R* is continuous, reflexive, and $P(x) = \{y: yPx\}$ is convex for all *x*, then an equilibrium point will exist in any budget set, and the resulting demand correspondence is upper-hemicontinuous. (He also assumed *R* is complete, but that assumption was used only to show that an equilibrium point is comparable to every affordable *y*.) Note that if *g* represents local theory, and *g* is continuous, then the *R* defined by *xRy* is equivalent to $g(x)(y-x) \leq 0$, satisfies Sonnenschein's conditions, and an equilibrium *x* for *R* is an equilibrium point for *g*, in any budget. Thus Sonnenschein demonstrated that Georgescu-Roegen's condition PPN is not necessary for the existence of equilibrium

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

points in a budget set.

In order to resolve question (1) above, a theory must predict a unique equilibrium point in each budget set, in order to get a well defined demand function. In the local theory, if one assumes $g(x) \neq 0$ for all *x* and strengthens PPN to SPPN: $g(x)(y - x) \leq 0,\ x \neq y$ implies $g(y)(x - y) > 0,$ then the equilibrium point *x* will be unique in any budget, and the resulting demand function will be continuous and satisfy the weak axiom of revealed preference (WARP). Thus the local theory, without assuming mathematical integrability (implied transitivity), yields a theory of individual demand functions satisfying WARP. On the other hand, given a continuous demand function *h* satisfying WARP, if *h* has a continuous inverse, then $g = h^{-1}$ yields a local theory with *g* satisfying SPPN and generating *h*. Now consider the global theory. If *R* is represented by a continuous, strictly quasi-concave nonsatiated utility function, then *R* will be reflexive, complete, transitive, strongly convex and nonsatiated. If one simply removes the assumption of transitivity from this list, then Sonnenschein's result implies that an equilibrium point will exist, and the remaining assumptions imply that this point will be unique, and that the resulting demand function will be continuous and satisfy WARP (see Shafer, 1974). Furthermore, Kim and Richter (1986) showed that, with a slight variation in the assumptions on *R*, any continuous demand function *h* satisfying a modified version of WARP can be generated by such an *R*. Thus, from the point of view of having single valued demand functions, the absence of transitivity, either assumed as in the global theory or implied as in the local theory, is essentially equivalent to Samuelson's theory of observed demand satisfying WARP. Since WARP includes the ‘law of demand’, that is, normal goods have downward sloping demand, in my view little is lost by not assuming transitivity.

Now question (2) above, the problem of existence of competitive equilibrium will be discussed. Again, Sonnenschein observed that if one took the standard assumptions on *R* normally used in proofs of existence of a competitive equilibrium, and removed the transitivity assumption, then demand correspondences would be well defined and convex valued, and the standard proof techniques would still work, so equilibrium would exist. Thus transitivity is irrelevant to demonstrating the internal consistency of the competitive model. Note, however, that the assumptions needed by Sonnenschein to demonstrate that individual demand correspondences are well defined and upper-hemicontinuous, namely continuity and convex preferred sets, are too weak to ensure convex valued demand correspondences. Nevertheless, Mas-Colell (1974) demonstrated that with only these assumptions on preferences, competitive equilibria will exist. Thus the only properties of individual preferences which are important to the existence of competitive equilibrium are continuity and convexity.

See Also

- integrability of demand
- orderings

Bibliography

Allen, R.G.D. 1932. The foundations of a mathematical theory of exchange. *Economica* 12, 197–226.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision Processes*, ed. R.M. Thrall, C.H. Combs and R.L. Davis. New York: Wiley.

Frisch, R. 1926. Sur un problème d'économie pure. *Norsk matematisk forenings skrifter* 16, 1–40.

Georgescu-Roegen, N. 1936. The pure theory of consumer's behavior. *Quarterly Journal of Economics* 50, 545–93.

Georgescu-Roegen, N. 1954. Choice and revealed preference. *Southern Economic Journal* 21(October), 119–30.

Katzner, D. 1971. Demand and exchange analysis in the absence of integrability conditions. In *Preferences, Utility, and Demand*, ed. J. Chipman. et al. New York: Harcourt, Brace, Jovanovich.

Kim, T. and Richter, M. 1986. Nontransitive–nontotal consumer theory. *Journal of Economic Theory* 38, 324–63.

Mas-Colell, A. 1974. An equilibrium existence theorem without complete or transitive preferences. *Journal of Mathematical Economics* 1, 237–46.

May, K. 1954. Intransitivity, utility, and aggregation in preference patterns. *Econometrica* 22, 1–13.

Shafer, W. 1974. The nontransitive consumer. *Econometrica* 42, 913–19.

Sonnenschein, H. 1971. Demand theory without transitive preferences, with applications to the theory of competitive equilibrium. In *Preferences, Utility and Demand*, ed. J. Chipman. et al. New York: Harcourt, Brace, Jovanovich.

How to cite this article

Shafer, Wayne. "transitivity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000118> doi:10.1057/9780230226203.1735

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transversality condition

Robert A. Becker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The transversality condition for an infinite horizon dynamic optimization problem is the boundary condition determining a solution to the problem's first-order conditions together with the initial condition. The transversality condition requires the present value of the state variables to converge to zero as the planning horizon recedes towards infinity. The first-order and transversality conditions are sufficient to identify an optimum in a concave optimization problem. Given an optimal path, the necessity of the transversality condition reflects the impossibility of finding an alternative feasible path for which each state variable deviates from the optimum at each time and increases discounted utility.

Keywords

arbitrage; asset pricing models; bubbles; capital accumulation programmes; competitive equilibrium; depreciation; Euler equations; infinite horizons; optimal growth paths; Ramsey model; transversality condition

Article

The transversality condition for an infinite horizon dynamic optimization problem acts as the boundary condition determining a solution to the problem's first-order conditions together with the initial condition. Malinvaud (1953) introduced the transversality condition as part of the sufficient conditions for intertemporally efficient capital accumulation programmes. He required the present value of the capital stock to converge to zero as the planning horizon tended towards infinity. An *efficient programme* is a feasible path of capital stocks starting from a given initial stock, together with a consumption path, having the property that no other feasible programme from the same starting stock provides as much consumption in every period and more consumption in at least one period. He showed that it was possible for a programme to be efficient for any finite planning horizon yet be inefficient when the program was considered over the entire infinite horizon. These inefficient programmes

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
'overaccumulated' capital and failed to satisfy the transversality condition.
Current theories emphasize the transversality condition's necessity. This is illustrated for the discrete time one-sector discounted Ramsey optimal growth model. There is a single all purpose consumption good, c_t, produced using capital goods, k_{t-1}, carried over from the previous period, and fixed labour. The planner decides how much to consume in the current period and how much to save for next period's production. Capital depreciates entirely within the period. The planner's initial stock of capital produces goods available in the first period. The planner obtains utility, $u(c_t)$, from consumption at time t and maximizes the discounted sum of future utilities. The discount factor, δ, is a given constant.
The planner's problem is:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u(c_t) \text{ by choice of } \left\{c_t, k_{t-1}\right\}_{t=1}^{\infty}, \text{ subject to:}$$
(1)

$$c_t + k_t \leq f(k_{t-1}) \text{ for } t = 1, 2, \ldots;$$
$$c_t \geq 0,\ k_{t-1} \geq 0 \text{ all } t;\ \ k_0 \leq k, \text{ where } k > 0 \text{ is given.}$$
(2)

Feasible programs are sequences $\{c_t, k_{t-1}\}_{t=1}^{\infty}$ which satisfy (2). Assume $u: [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $u(0) = 0$, and satisfies the Inada condition: $\lim_{c \to 0+} u'(c) = \infty$. The production function $f: [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $f(0) = 0$, satisfies $\lim_{k \to 0+} f'(k) = \infty$, and $\lim_{k \to 0+} f'(k) < 1$. There is a maximum sustainable stock, $b > 0$,with $f(b) = b$ and $0 < k < b$. The discount factor satisfies $0 < \delta < 1$. There is a unique optimal program, $\{\bar{c}_t, \bar{k}_{t-1}\}_{t=1}^{\infty}$.
The optimal program satisfies $(\bar{c}_t, \bar{k}_{t-1}) > 0$ for each t. The Kuhn–Tucker necessary conditions for an optimum, known as the *Euler*, or *no-arbitrage conditions*, are:

$$\delta f'(\bar{k}_t) u'(\bar{c}_{t-1}) = u'(\bar{c}_t), \quad \text{for each } t.$$
(3)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

If the planner's horizon is a finite period, *T*, then (3) and the complementary slackness condition $\delta^{T-1}u'(\bar{c}_T)\bar{k}_T = 0$ obtain. The latter condition states capital's terminal value is zero. For the infinite horizon case of interest, it is natural to conjecture the *transversality* condition holds as a necessary condition for optimality:

$$\lim_{T\to\infty} \delta^{T-1}u'(\bar{c}_T)\bar{k}_T = 0. \tag{4}$$

The optimal path of the infinite horizon problem converges monotonically to the stationary optimal programme (c^*, k^*),with $c^* = f(k^*) - k^*$ and $\delta f'(k^*) = 1$. Hence, the transversality condition holds. The economic intuition underlying the necessity of the transversality condition is independent of this convergence result.

Equation (3) expresses the unprofitability of the one-period reversed arbitrages developed below. An *arbitrage* represents a feasible change in the optimal path. *Reversed arbitrages* perturb the optimum for finitely many consecutive periods. *Unreversed arbitrages* change the optimal path permanently from some given time on to infinity. A necessary condition for an optimal path is that no arbitrage increase the discounted sum of future utilities above the optimal discounted utility. The necessity of the transversality condition can be interpreted as a type of no-arbitrage condition for unreversed arbitrages.

Assume $\{\bar{c}_t, \bar{k}_{t-1}\}_{t=1}^{\infty}$ is optimal. Suppose the planner decides to increase the first period's consumption and forgoes one unit of capital to be used in next period's production. The marginal gain is $u'(\bar{c}_1)$ in units of utility at time 1. A *T-period reversed arbitrage* occurs if at time $T+1$ the planner reacquires the unit of capital forgone at time one. After time $T+1$, the arbitrage no longer affects the path.

Two costs are incurred by the acquisition at time *T*+1. First, there is the *direct cost* or *repurchase cost* of forgone consumption, which arises from converting a unit of consumption at time $T+1$ to a unit of capital to be saved for the next period's production. This direct cost equals $\delta^T u'(\bar{c}_{T+1})$ in period 1 utils. The indirect cost arises because the net marginal product of that unit of capital is lost to the planner in every period between $t = 2$ and $t = T+1$. The indirect cost at time *t* in utils of time 1 is $\delta^{t-1}u'(\bar{c}_t)[f'(\bar{k}_{t-1}) - 1]$; adding over $t = 2, \ldots, T+1$ yields the present value (focal date one) of the indirect cost. The total cost equals the sum of the direct and indirect costs:

$$\sum_{t=2}^{T+1} \delta^{t-1}u'(\bar{c}_t)[f'(\bar{k}_{t-1}) - 1] + \delta^T u'(\bar{c}_{T+1}). \tag{5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

A necessary condition for the optimality is, for any *T*, the marginal benefit of a *T*-period reversed arbitrage is equal to its marginal (discounted) cost. Thus,

$$u'(\bar{c}_1) = \sum_{t=2}^{T+1} \delta^{t-1} u'(\bar{c}_t)[f'(\bar{k}_{t-1}) - 1] + \delta^T u'(\bar{c}_{T+1}).$$

(6)

Equation (6) applied to a one-period reversed arbitrage reduces to (3), evaluated at $T = 1$. Equation (3) follows from the same reasoning when a reversed arbitrage starts at time *t*.

Equation (6) contains no further information. However, the infinite horizon means the planner can also contemplate the profitability of an *unreversed* arbitrage in which the unit of capital is permanently sacrificed at time $t = 1$. There is no repurchase cost associated with an unreversed arbitrage, hence the conditions for the unprofitability of an unreversed arbitrage must be

$$u'(\bar{c}_1) = \sum_{t=2}^{\infty} \delta^{t-1} u'(\bar{c}_t)[f'(\bar{k}_{t-1}) - 1].$$

(7)

But (6) and (7) can hold as $T \to \infty$ only if $\lim_{T \to \infty} \delta^T u'(\bar{c}_{T+1}) = 0$, which implies the transversality condition since the capital stocks are bounded. An unreversed arbitrage over-accumulates capital in comparison with the optimal programme. This cannot be optimal. Thus, the transversality condition expresses the zero marginal profit condition for the open-ended arbitrages which are available only in the infinite horizon context. The rate of capital accumulation is thereby limited.

The Euler equations and the transversality conditions are necessary for optimality. For concave optimal growth models – the case where utility and production functions are concave – the Euler equations and transversality conditions are also sufficient to identify an optimal programme. Indeed, transversality conditions are necessary (and sufficient) together with the Euler equations in other dynamic models where the state variables need not be capital stocks.

The transversality condition's necessity is important in connecting dynamic equilibrium paths and optimal growth paths in infinitely lived representative agent economies. The basic idea is to match the Euler and transversality conditions in the equilibrium and optimal growth settings. This results in an equivalence principle: an optimal growth path is a competitive equilibrium, and vice versa.

The necessity of the transversality condition can also be used to exclude bubbles from occurring in asset

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

pricing models. The asset's fundamental is the present discounted value of its future dividend payouts. A bubble exists if the asset's price differs from its fundamental value. For example, if the asset is a perpetuity and offers a constant dividend stream, then the no-arbitrage conditions state the capital gain yield plus the dividend yield equals the interest rate at each time. Formally,

$$\frac{p_{t+1} - p_t}{p_t} + \frac{r}{p_t} = \rho, \quad t = 1, 2, \ldots. \tag{8}$$

where p_t is the asset's current price at time t, r is the asset's return (or dividend) in each period, and ρ is the constant interest rate.

Equation (8) is a first-order difference equation. Its solution is

$$p_t = (1 + \rho)^{t-1}\left[p_1 - \left(\frac{r}{\rho}\right)\right] + \left(\frac{r}{\rho}\right). \tag{9}$$

For each choice of p_1, there is a sequence, $\{p_t\}_{t=1}^{\infty}$, calculated from (9). Thus, there are an infinite number of price systems satisfying this asset's no-arbitrage equation.

Efficient markets would imply the absence of arbitrage opportunities for all time and single out the solution $p_t = (r/\rho)$, which occurs if and only if $p_1 = (r/\rho)$. Initial prices with $p_1 > (r/\rho)$ would create a bubble where p_t exceeds its *fundamental value*, (r/ρ). Prices continue to rise simply because investors expect them to do so. There is a *negative bubble* if $p_1 < (r/\rho)$. The asset's price becomes negative in finite time, an impossibility as prices must always remain non-negative. Hence, $p_1 \geq (r/\rho)$.

The transversality condition takes the form

$$\lim_{t \to \infty} p_t\left(\frac{1}{1+\rho}\right)^{t-1} = 0. \tag{10}$$

If this is an equilibrium condition, then the initial price must be $p_1 = (r/\rho)$ and the asset's market price equals its fundamental at each time. Imposition of the transversality condition picks out the only

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
solution of (8) that is not a bubble.

Many deterministic models have stochastic counterparts. For example, technology shocks in the Ramsey problem lead to stochastic Euler equations expressing no-arbitrage opportunities in expectations, or on average, when the planner's objective is the expected discounted sum of future utilities. The corresponding transversality condition also holds in expectations. There can exist particular realizations of the shocks for which a bubble persists. However, on average, there are no unprofitable unreversed arbitrages.

The argument for the necessity of the transversality condition given above is heuristic. Weitzman (2003) presents an analogous intuitive rationale for the transversality condition in continuous time. The terms ‘reversed’ and ‘unreversed’ arbitrage originate in Gray and Salant (1983). Rigorous arguments are found in the references below.

See Also

- arbitrage pricing theory
- bubbles
- efficient markets hypothesis
- government budget constraint
- tulipmania

Bibliography

Becker, R.A. and Boyd, J.H. 1997. *Capital Theory, Equilibrium Analysis, and Recursive Utility*. Malden, MA: Blackwell Publishers.

Benveniste, L.M. and Schienkman, J.A. 1982. Duality theory for dynamic optimization models of economics: the continuous time case. *Journal of Economic Theory* 27, 1–19.

Gray, J.A. and Salant, S.W. 1983. Transversality conditions in infinite horizon models. Working paper, Washington State University.

Kamihigashi, T. 2005. Necessity of the transversality condition for stochastic models with bounded or CRRA utility. *Journal of Economic Dynamics and Control* 29, 1313–29.

LeRoy, S.F. 2004. Rational exuberance. *Journal of Economic Literature* 41, 783–804.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21, 233–68.

Michel, P. 1982. On the transversality condition in infinite horizon optimal problems. *Econometrica* 50, 975–86.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mirman, L.J. and Zilcha, I. 1975. Optimal growth under uncertainty. *Journal of Economic Theory* 11, 329–39.

Weitzman, M.L. 1973. Duality theory for infinite horizon convex models. *Management Science* 19, 783–9.

Weitzman, M.L. 2003. *Income, Wealth, and the Maximum Principle*. Cambridge, MA: Harvard University Press.

How to cite this article

Becker, Robert A. "transversality condition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000218>
doi:10.1057/9780230226203.1736

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

transversality conditions and dynamic economic behaviour

Takashi Kamihigashi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Transversality conditions are optimality conditions often used along with Euler equations to characterize the optimal paths of dynamic economic models. This article illustrates the role of transversality conditions in characterizing optimal paths as well as in ruling out economic phenomena such as asset bubbles and hyperdeflations in infinite-horizon models.

Keywords

bubbles; calculus of variations; dynamic models; dynamic optimization; Euler equations; hyperdeflation; infinite horizons; optimality; overlapping generations models; Ponzi games; Ramsey model; transversality condition; transversality conditions and dynamic economic behaviour

Article

Transversality conditions are optimality conditions often used along with Euler equations to characterize the optimal paths of dynamic economic models. The purpose of this article is to illustrate the role of transversality conditions in characterizing optimal paths as well as in ruling out economic phenomena such as asset bubbles and hyperdeflations in infinite-horizon models. See transversality condition for mathematical foundations.

A geometric example

A simple geometric example best illustrates the mathematical roles of an Euler equation and a transversality condition. What is the shortest path from a point A to a straight line L infinitely long in both directions? The answer is, of course, the straight line from point A to line L that is perpendicular to line L. There are two conditions involved here. The first condition is that the shortest path be a straight line: one cannot make the path shorter by deviating from it and eventually returning to it. This is the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
implication of the Euler equation for this problem. But there are infinitely many straight lines from point *A* to line *L*. In fact, a straight line from point *A* to line *L* can be arbitrarily long; even very bad choices satisfy the Euler equation. This is why one needs the second condition, that the shortest path be perpendicular to line *L*. This additional condition ensures that one cannot make the path shorter by deviating from it and never returning to it.

The condition of perpendicularity in this example and similar conditions on end points in other problems are called transversality conditions in dynamic optimization theory (Hestenes, 1966, p. 87). According to Bolza (1904, p. 106), the term was first introduced by Kneser (1900). Both Euler equations and transversality conditions were initially developed for calculus-of-variations problems. In economics, in particular in macroeconomics, both types of conditions are used mainly for infinite-horizon models in discrete time as well as in continuous time. In what follows, we focus on discrete-time models, which are technically easier to deal with.

A general discrete-time problem

Consider the following maximization problem:

$$\max_{\{x_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t v(x_t, x_{t+1}) \tag{1}$$

$$\text{s.t.} \quad x_{t+1} \geq 0, \quad t = 0, 1, 2, \ldots, \tag{2}$$

$$x_0 \text{ given}, \tag{3}$$

where $\beta \in (0,1)$ is called the discount factor, v is called the return function, and x_t is an n-dimensional vector. There may be other constraints, but we assume that they are never binding at the optimum. We

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

also assume that the non-negativity constraint is never binding at the optimum, and that the return function v is differentiable and concave. Though the return function may depend on t, we do not assume so here for notational simplicity. To be concrete, we interpret x_t as the stock of wealth (or capital) at the beginning of period t. In most economic problems, it is costly to accumulate wealth. Hence we assume that $v_2(x,y) \leq 0$ for all x, y.

The Euler equation for this problem is simply the first-order condition with respect to x_{t+1}:

$$v_2(x_t, x_{t+1}) + \beta v_1(x_{t+1}, x_{t+2}) = 0. \tag{4}$$

This condition means that no gain can be achieved by deviating from an optimal path for one period. The transversality condition is given by

$$\lim_{T \to \infty} \beta^T [-v_2(x_T, x_{T+1})] x_{T+1} = 0. \tag{5}$$

Though this is the typical transversality condition in economics, other transversality conditions are possible depending on the specification of constraints. Condition (5) can be interpreted as saying that the present discounted value of wealth at infinity must be zero, or wealth (x_{T+1}) should not grow too fast compared with its discounted marginal value ($\beta^T[-v_2(x_T, x_{T+1})]$). In other words, the transversality condition (5) rules out overaccumulation of wealth. The idea is that, if one saves too much and spends too little, then one is not behaving optimally.

It is well known that the Euler equation (4) and the transversality condition (5) are sufficient for optimality (Stokey and Lucas, 1989, p. 89). This result is often credited to Mangasarian (1966), who showed a finite-horizon version of the result.

Since the Euler equation is simply the first-order condition with respect to x_{t+1}, it is a necessary condition for optimality. On the other hand, necessity of the transversality condition is often considered to be a difficult issue. But there are two simple ways to prove its necessity if the objective function is assumed to be finite for all feasible paths (Kamihigashi, 2002; 2005). If this assumption fails, one can try the following test. Shift the entire optimal path downward by a small fixed proportion. Does it reduce the value of the objective function by only a finite amount? If so, the transversality condition is necessary. See Kamihigashi (2001; 2003) for precise assumptions and statements. See Weitzman (1973), Benveniste and Scheinkman (1982), and Michele (1982) for earlier results and arguments.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The Ramsey model

To see how the transversality condition can be used in practice, consider the basic Ramsey model:

$$\max_{\{c_t, x_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{6}$$

$$\text{s.t.} \quad c_t + x_{t+1} = f(x_t), \quad c_t, x_{t+1} \geq 0, \quad t = 0, 1, 2, \ldots, \tag{7}$$

$$x_0 \text{ given}, \tag{8}$$

where c_t is consumption and x_t is the stock of capital at the beginning of period t. We assume that the utility function u and the production function f are continuously differentiable, strictly increasing, and strictly concave. We also assume that $f(0)=0$, $\lim_{c \to 0} u'(c) = \lim_{x \to 0} f'(x) = \infty$, and $\lim_{x \to \infty} f'(x) < 1$. The model here is a special case of the general problem described above with $v(x_t, x_{t+1}) = u(f(x_t) - x_{t+1})$. The Euler equation and the transversality condition are

$$u'(c_t) = \beta u'(c_{t+1}) f'(x_{t+1}), \tag{9}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\lim_{T \to \infty} \beta^T u'(c_T) x_{T+1} = 0. \tag{10}$$

Given any initial capital stock x_0, there are three types of paths obeying (7) and (9). Specifically, there is a unique level of consumption $\hat{c}_0$ such that a path $\{c_t, x_{t+1}\}$ satisfying (7) and (9) converges to the unique steady state (which is determined by the strictly positive capital stock x^* such that $\beta f'(x^*) = 1$) if and only if $c_0 = \hat{c}_0$. If $c_0 > \hat{c}_0$, then (7) is eventually violated; such paths are ruled out on feasibility grounds. If $c_0 < \hat{c}_0$, then c_t converges to zero and x_t converges to the capital stock $\bar{x}$ given by $f(\bar{x}) = \bar{x}$. It can be shown that such paths violate the transversality condition and thus are not optimal. The path converging to the steady state satisfies the transversality condition since c_t and x_t converge to their steady state values; hence this is the optimal path.

The preceding argument shows that when one restricts attention to the dynamical system defined by (7) and (9), most paths do not converge to the steady state. This is an example of the Hahn problem (Hahn, 1987). The Hahn problem disappears here when one takes the transversality condition into account, since only the path converging to the steady state satisfies the transversality condition as well as the feasibility requirements.

Asset bubbles

Transversality conditions are often used to rule out asset bubbles. To be specific, consider a deterministic version of the Lucas (1978) asset pricing model. There are many homogeneous agents, a single good, and a single asset that pays a dividend of d_t units of the good in each period t. The population of agents is normalized to one; so is the supply of the asset. Each agent solves

$$\max_{\{c_t, x_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{11}$$

$$\text{s.t.} \quad c_t + p_t x_{t+1} = (p_t + d_t) x_t, \quad c_t, x_{t+1} \geq 0, \quad t = 0, 1, 2, \ldots, \tag{12}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$x_0 = 1 \text{ given,} \tag{13}$$

where c_t is consumption, p_t is the price of the asset, and x_t is shares in the asset at the beginning of period t. In equilibrium, c_t=d_t and x_t=1. We assume that the utility function u is concave, differentiable and strictly increasing. The Euler equation and the transversality condition in equilibrium are

$$u'(d_t)\,p_t = \beta u'(d_{t+1})(p_{t+1} + d_{t+1}), \tag{14}$$

$$\lim_{T \to \infty} \beta^T u(d_T)\,p_T = 0. \tag{15}$$

It is easy to see that the sequence $\{p_t^*\}$ given by

$$p_t^* = \sum_{i=1}^{\infty} \beta^i \frac{u'(d_{t+i})}{u'(d_t)} d_{t+i} \tag{16}$$

satisfies the Euler equation (14). The right-hand side of (16) is called the fundamental value of the asset. Let $\{b_t\}$ be any nonnegative sequence satisfying

$$u'(d_t)b_t = \beta u'(d_{t+1})b_{t+1}.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(17)

Then the sequence $\{p_t^* + b_t\}$ also satisfies the Euler equation. Hence there are infinitely many paths satisfying the Euler equation. The extra component b_t, which grows at a gross rate of $u'(d_t)/[\beta u'(d_{t+1})]$, is interpreted as a bubble.

Notice that the bubble component b_t, if strictly positive, violates the transversality condition (15) (with $p_T=b_T$). Therefore, if the transversality condition is necessary for optimality, the bubble component must vanish, so that the price must always be equal to the fundamental value. This is indeed the case in this model (Kamihigashi, 2001, p. 1007).

In stochastic models, bubbles can be ruled out by the same argument under standard assumptions, but there are pathological cases in which bubbles are possible (Kamihigashi, 1998; Montrucchio and Privileggi, 2001). Bubbles arise more easily in models with heterogenous agents such as overlapping generations models, where there is no economy-wide transversality condition. See speculative bubbles.

Hyperdeflations

Transversality conditions are often used to rule out hyperdeflations in money-in-the-utility-function models of the type studied by Brock (1974) and Obstfeld and Rogoff (1986). In these models, agents derive utility from real money balances in addition to consumption. As in the Lucas asset pricing model, there are many paths satisfying the Euler equation. Along a path satisfying the Euler equation with a positive bubble, the value of real balances grows unboundedly or, equivalently, the nominal price level keeps declining towards zero. Such paths are often interpreted as exhibiting hyperdeflations. Under reasonable assumptions, hyperdeflationary paths are ruled out by the transversality condition, which once again rules out overaccumulation of wealth, or real balances.

However, there are cases in which the transversality condition does not rule out hyperdeflationary paths (Obstfeld and Rogoff, 1986, p. 356). This is because agents, who derive utility from real balances, benefit directly from accumulating wealth.

No-Ponzi-game conditions

In formulating a consumer's maximization problem, one must include some constraint on debt, since otherwise the consumer would never pay back his debt, letting it grow unboundedly. One way to rule out this behaviour is to prohibit debt entirely, that is, to require wealth to be always non-negative (as in (12)). A more lenient way is to require only the present discounted value of wealth at infinity to be non-negative. This type of condition is known as a no-Ponzi-game condition (Blanchard and Fischer, 1989, p. 49), but often called a transversality condition as well. A no-Ponzi-game condition is a constraint that prevents overaccumulation of debt, while a typical transversality condition is an optimality condition that rules out overaccumulation of wealth. They place opposite restrictions, and should not be confused.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- bubbles
- calculus of variations
- Euler equations
- Ramsey model
- speculative bubbles
- transversality condition

Bibliography

Benveniste, L.M. and Scheinkman, J.A. 1982. Duality theory for dynamic optimization models of economics: the continuous time case. *Journal of Economic Theory* 27, 1–19.

Blanchard, O.J. and Fischer, S. 1989. *Lectures on Macroeconomics*. Cambridge, MA: MIT Press.

Bolza, O. 1904. *Lectures on the Calculus of Variations*. Chicago: University of Chicago Press.

Brock, W.A. 1974. Money and growth: the case of long run perfect foresight. *International Economic Review* 15, 750–77.

Hahn, F.H. 1987. ‘Hahn problem’. In *The New Palgrave: A Dictionary of Economics*, vol. 2, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Hestenes, M.R. 1966. *Calculus of Variations and Optimal Control Theory*. New York: Wiley.

Kamihigashi, T. 1998. Uniqueness of asset prices in an exchange economy with unbounded utility. *Economic Theory* 12, 103–22.

Kamihigashi, T. 2001. Necessity of transversality conditions for infinite horizon problems. *Econometrica* 69, 995–1012.

Kamihigashi, T. 2002. A simple proof of the necessity of the transversality condition. *Economic Theory* 20, 427–33.

Kamihigashi, T. 2003. Necessity of transversality conditions for stochastic problems. *Journal of Economic Theory* 109, 140–49.

Kamihigashi, T. 2005. Necessity of the transversality condition for stochastic models with bounded or CRRA utility. *Journal of Economic Dynamics and Control* 29, 1313–29.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kneser, A. 1900. *Lehrbuch der Variationsrechnung*. Braunschweig: F. Vieweg und Sohn.

Lucas, R.E., Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46, 1429–445.

Mangasarian, O.L. 1966. Sufficient conditions for the optimal control of nonlinear systems. *SIAM Journal of Control* 4, 139–52.

Michele, P. 1982. On the transversality condition in infinite-horizon problems. *Econometrica* 50, 975–85.

Montrucchio, L. and Privileggi, F. 2001. On fragility of bubbles in equilibrium asset pricing models of Lucas-type. *Journal of Economic Theory* 101, 158–88.

Obstfeld, M. and Rogoff, K. 1986. Ruling out divergent speculative bubbles. *Journal of Monetary Economics* 17, 349–62.

Stokey, N. and Lucas, R.E., Jr. 1989. *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

Weitzman, M.L. 1973. Duality theory for infinite horizon convex models. *Management Science* 19, 783–9.

How to cite this article

Kamihigashi, Takashi. "transversality conditions and dynamic economic behaviour." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000217> doi:10.1057/9780230226203.1737

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

treatment effect

Joshua D. Angrist
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The term ‘treatment effect’ refers to the causal effect of a binary (0–1) variable on an outcome variable of scientific or policy interest. Economics examples include the effects of government programmes and policies, such as those that subsidize training for disadvantaged workers, and the effects of individual choices like college attendance. The principal econometric problem in the estimation of treatment effects is selection bias, which arises from the fact that treated individuals differ from the non-treated for reasons other than treatment status per se. Treatment effects can be estimated using social experiments, regression models, matching estimators, and instrumental variables.

Keywords

average treatment effect; constant-effects models; identifying assumptions; instrumental variables (IV) methods; law of large numbers; local average treatment effect; matching estimators; monotonicity; omitted variables bias: see selection bias; potential-outcomes framework; propensity-score matching; regression models; selection bias; switching regressions model; treatment effect; two-stage least squares; two-step estimators; Wald estimator

Article

A ‘treatment effect’ is the average causal effect of a binary (0–1) variable on an outcome variable of scientific or policy interest. The term ‘treatment effect’ originates in a medical literature concerned with the causal effects of binary, yes-or-no ‘treatments’, such as an experimental drug or a new surgical procedure. But the term is now used much more generally. The causal effect of a subsidized training programme is probably the mostly widely analysed treatment effect in economics (see, for example, Ashenfelter, 1978, for one of the first examples, or Heckman and Robb, 1985 for an early survey). Given a data-set describing the labour market circumstances of trainees and a non-trainee comparison group, we can compare the earnings of those who did participate in the programme and those who did not. Any empirical study of treatment effects would typically start with such simple comparisons. We might also use regression methods or matching to control for demographic or background characteristics.

In practice, simple comparisons or even regression-adjusted comparisons may provide misleading estimates of causal effects. For example, participants in subsidized training programmes are often observed to earn less than ostensibly comparable controls, even after adjusting for observed differences (see, for example, Ashenfelter and Card, 1985). This may reflect some sort of omitted variables bias, that is, a bias arising from unobserved and uncontrolled differences in earnings potential between the two groups being compared. In general, omitted variables bias (also known as selection bias) is the most serious econometric concern that arises in the estimation of treatment effects. The link between omitted variables bias, causality, and treatment effects can be seen most clearly using the potential-outcomes framework.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Causality and potential outcomes

The notion of a causal effect can be made more precise using a conceptual framework that postulates a set of potential outcomes that could be observed in alternative states of the world. Originally introduced by statisticians in the 1920s as a way to discuss treatment effects in randomized experiments, the potential outcomes framework has become the conceptual workhouse for non-experimental as well as experimental studies in many fields (see Holland, 1986, for a survey and Rubin, 1974; 1977, for influential early contributions). Potential outcomes models are essentially the same as the econometric *switching regressions* model (Quandt, 1958), though the latter is usually tied to a linear regression framework. Heckman (1976; 1979) developed simple two-step estimators for this model.

Average causal effects

Except in the realm of science fiction, where parallel universes are sometimes imagined to be observable, it is impossible to measure causal effects at the individual level. Researchers therefore focus on average causal effects. To make the idea of an average causal effect concrete, suppose again that we are interested in the effects of a training programme on the post-training earnings of trainees. Let Y_{1i} denote the potential earnings of individual i if he were to receive training and let Y_{0_i} denote the potential earnings of individual i if not. Denote training status by a dummy variable, D_i. For each individual, we observe $Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$, that is, we observe Y_{1i} for trainees and Y_{0i} for everyone else. Let $E[\cdot]$ denote the mathematical expectation operator, i.e., the population average of a random variable. For continuous random variables, $E[Y_i] = \int y f(y) dy$, where $f\bullet(y)$ is the density of Y_i. By the law of large numbers, sample averages converge to population averages so we can think of $E[\cdot]$ as giving the sample average in very large samples. The two most widely studied average causal effects in the treatment effects context are the average treatment effect (ATE), $E[Y_{1i} - Y_{0i}]$, and the average treatment effect on the treated (ATET), $E[Y_{1i} - Y_{0i}|D_i = 1]$. Note that the ATET can be rewritten

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1].$$

This expression highlights the counter-factual nature of a causal effect. The first term is the average earnings in the population of trainees, a potentially observable quantity. The second term is the average earnings of trainees had they not been trained. This cannot be observed, though we may have a control group or econometric modelling strategy that provides a consistent estimate.

Selection bias and social experiments

As noted above, simply comparing those who are and are not treated may provide a misleading estimate of a treatment effect. Since the omitted variables problem is unrelated to sampling variance or statistical inference, but rather concerned with population quantities, it too can be efficiently described by using mathematical expectation notation to denote population averages. The contrast in average outcomes by observed treatment status is

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = E[Y_{1i} - Y_{0i}|D_i = 1] + \{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\}$$

Thus, the naive contrast can be written as the sum of two components, ATET, plus selection bias due to the fact that the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

average earnings of non-trainees, $E[Y_{0i}|D_i = 0]$, need not be a good stand-in for the earnings of trainees had they not been trained, $E[Y_{0i}|D_i = 1]$.

The problem of selection bias motivates the use of random assignment to estimate treatment effects in social experiments. Random assignment ensures that the potential earnings of trainees had they not been trained – an unobservable quantity – are well-represented by the randomly selected control group. Formally, when D_i is randomly assigned, $E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}]$. Replacing $E[Y_i|D_i = 1]$ and $E[Y_i|D_i = 0]$ with the corresponding sample analogs provides a consistent estimate of ATE.

Regression and matching

Although it is increasingly common for randomized trials to be used to estimate treatment effects, most economic research still uses observational data. In the absence of a randomized experiment, researchers rely on a variety of statistical control strategies and/or natural experiments to reduce omitted variables bias. The most commonly used statistical techniques in this context are regression, matching and instrumental variables.

Regression estimates of causal effects can be motivated most easily by postulating a constant-effects model, where $Y_{1i} - Y_{0i} = \alpha$ (a constant). The constant-effects assumption is not strictly necessary for regression to estimate an average causal effect, but it simplifies things to postpone a discussion of this point. More importantly, the only source of omitted-variables bias is assumed to come from a vector of observed covariates, X_i, that may be correlated with D_i. The key assumption that facilitates causal inference in regression models (sometimes called an identifying assumption), is that

$$E[Y_{0i}|X_i, D_i] = X_i'\beta, \tag{1}$$

where β is a vector of regression coefficients. This selection-on-observables assumption has two parts. First, Y_{0_i} (and hence Y_{1i}, given the constant-effects assumption) is mean-independent of D_i conditional on X_i. Second, the conditional mean function for Y_{0i} given X_i is linear. Given eq. (1), it is straightforward to show that

$$E\{Y_i(D_i - R[D_i|X_i])\} / E\{D_i(D_i - R[D_i|X_i])\} = \alpha, \tag{2}$$

where $R[D_i/X_i]$ are the fitted values from a regression of D_i on X_i. This is the coefficient on D_i from the population regression of Y_i on D_i and X_i (that is, the regression coefficient in an infinite sample). Again, the law of large numbers ensures that sample regression coefficients estimate this population regression coefficient consistently.

Matching is similar to regression in that it is motivated by the assumption that the only source of omitted variables or selection bias is the set of observed covariates, X_i. Unlike regression, however, matching estimates of treatment effects are constructed by matching individuals with the same covariates instead of through a linear model for the effect of covariates. Instead of (1), the selection-on-observables assumption becomes

$$E[Y_{ji}|X_i, D_i] = E[Y_{ji}|X_i], \text{ for } j = 0, 1. \tag{3}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This implies

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E\{E[Y_{1i}|X_i, D_i = 1] - [Y_{0i}|X_i, D_i = 1]|D_i = 1\} = E\{E[Y_{1i}|X_i, D_i = 1] - [Y_{0i}|X_i, D_i = 0]|D_i = 1\}$$
(4a)

and, likewise,

$$E[Y_{1i} - Y_{0i}] = E\{E[Y_{1i}|X_i, D_i = 1] - [Y_{0i}|X_i, D_i = 0]\}$$
(4b)

In other words, we can construct ATET or ATE by averaging *X*-specific treatment-control contrasts, and then reweighting these *X*-specific contrasts using the distribution of X_i for the treated (for ATET) or using the marginal distribution of X_i (for ATE). Since these expressions involve observable quantities, it is straightforward to construct consistent estimators from their sample analogs.

The conditional independence assumption that motivates the use of regression and matching is most plausible when researchers have extensive knowledge of the process determining treatment status. An example in this spirit is the Angrist (1998) study of the effect of voluntary military service on the civilian earnings of soldiers after discharge, discussed further below.

Regression and matching details

In practice, regression can be understood as a type of weighted matching estimator. If, for example, $E[D_i|X_i]$ is a linear function of X_i (as it might be if the covariates are all discrete), then it is possible to show that eq. (2) is equivalent to a matching estimator that weights cell-by-cell treatment-control contrasts by the conditional variance of treatment in each cell (Angrist, 1998). This equivalence highlights the fact that the most important econometric issue in a study that relies on selection-on-observables assumptions to identify causal effects is the validity of these conditional independence assumptions, not whether regression or matching is used to implement them.

A computational difficulty that sometimes arises in matching models is how to find good matches for each possible value of the covariates when the covariates take on many values. For example, beginning with Ashenfelter (1978), many studies of the effect of training programmes have shown that trainees typically experience a period of declining earnings before they go into training. Because lagged earnings is both continuous and multidimensional (since more than one period's earnings seem to matter), it may be hard to match trainees and controls with exactly the same pattern of lagged earnings. A possible solution in this case is to match trainees and controls on the *propensity score*, the conditional probability of treatment given covariates. Propensity-score matching relies on the fact that, if conditioning on X_i eliminates selection bias, then so does conditioning on $P[D_i = 1|X_i]$, as first noted by Rosenbaum and Rubin (1983). Use of the propensity score reduces the dimensionality of the matching problem since the propensity score is a scalar, though in practice it must still be estimated. See Dehejia and Wahba (1999) for an illustration.

Regression and matching example

Between 1989 and 1992, the size of the military declined sharply because of increasing enlistment standards.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Policymakers would like to know whether the people – many of them black men – who would have served under the old rules but were unable to enlist under the new rules were hurt by the lost opportunity for service. The Angrist (1998) study attempts to answer this question. The regression and matching assumptions seem plausible in this context because soldiers are selected on the basis of a few well-documented criteria related to age, schooling and test scores and because the control group used in the study also applied to enter the military.

Naive comparisons clearly overestimate the benefit of military service. This can be seen in Table 1, which reports differences-in-means, matching and regression estimates of the effect of voluntary military service on the 1988–91 Social Security-taxable earnings of men who applied to join the military between 1979 and 1982. The matching estimates were constructed from the sample analog of (4a), that is, from covariate-value-specific differences in earnings, weighted to form a single estimate using the distribution of covariates among veterans. The covariates in this case are the age, schooling and test-score variables used to select soldiers from the pool of applicants. Although white veterans earn $1,233 more than non-veterans, this difference becomes negative once the adjustment for differences in covariates is made. Similarly, while non-white veterans earn $2,449 more than non-veterans, controlling for covariates reduces this to $840.

Matching and regression estimates of the effects of voluntary military service in the United States

Race	Average earnings in 1988–91	Differences in means	Matching estimates	Regression estimates	Regression minus matching
	(1)	(2)	(3)	(4)	(5)
Whites	14,537	1,233.4	–197.2	–88.8	108.4
	(60.3)	(70.5)	(62.5)	(28.5)	
Non-whites	11,664	2,449.1	839.7	1,074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

Notes: Figures are in nominal US dollars. The table shows estimates of the effect of voluntary military service on the 1988–91 Social Security-taxable earnings of men who applied to enter the armed forces during 1979–82. The matching and regression estimates control for applicants' year of birth, education at the time of application, and Armed Forces Qualification Test (AFQT) score. There are 128,968 whites and 175,262 non-whites in the sample. Standard errors are reported in parentheses.

Source: Adapted from Angrist (1998, Tables II and V).

Table 1 also shows regression estimates of the effect of voluntary military service, controlling for the same covariates used for matching. These are estimates of α_r in the equation

$$Y_i = \sum_X d_{iX}\beta_X + \alpha_r D_i + e_i,$$

where β_X is a regression-effect for $X_i = X$ and α_r is the regression parameter. This corresponds to a saturated model for discrete X_i. The regression estimates are larger than (and significantly different from) the matching estimates. But the regression and matching estimates are not very different economically, both pointing to a small earnings loss for White veterans and a modest gain for Non-whites.

Instrumental variables estimates of treatment effects

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The assumptions required for regression or matching to identify a treatment effect are often implausible. Many of the necessary control variables are typically unmeasured or simply unknown. Instrumental variables (IV) methods solve the problem of missing or unknown controls, much as a randomized trial also obviates the need for regression or matching. To see how this is possible, begin again with a constant effects model without covariates, so $Y_{1i} - Y_{0i} = \alpha$. Also, let $Y_{0i} = \beta + \varepsilon_i$, where $\beta \equiv E[Y_{0i}]$. The potential outcomes model can now be written

$$Y_i = \beta + \alpha D_i + \varepsilon_i, \tag{5}$$

where α is the treatment effect of interest. Because D_i is likely to be correlated with ε_i, regression estimates of eq. (5) do not estimate α consistently.

Now suppose that in addition to Y_i and D_i there is a third variable, Z_i, that is correlated with D_i, but unrelated to Y_i for any other reason. In a constant-effects world, this is equivalent to saying Y_{0i} and Z_i are independent. It therefore follows that

$$E[\varepsilon_i | Z_i] = 0, \tag{6}$$

a conditional independence restriction on the relation between Z_i and Y_{0i}, instead of between D_i and Y_{0i} as required for regression or matching strategies. The variable Z_i is said to be an IV or just ‘an instrument’ for the causal effect of D_i on Y_i.

Suppose that Z_i is also a 0–1 variable. Taking expectations of (5) with Z_i switched off and on, we immediately obtain a simple formula for the treatment effect of interest:

$$\frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]} = \alpha. \tag{7}$$

The sample analog of this equation is sometimes called the Wald estimator, since it first appear in a paper by Wald (1940) on errors-in-variables problems. There are other more complicated IV estimators involving continuous, multi-valued, or multiple instruments. For example, with a multi-valued instrument, we might use the sample analog of Cov(Z_i, Y_i)/Cov(D_i, Y_i). This simplifies to the Wald estimator when Z_i is 0–1. The Wald estimator captures the main idea behind most IV estimation strategies since more complicated estimators can usually be written as a linear combination of Wald estimators (Angrist, 1991).

IV example

To see how IV works in practice, it helps to use an example, in this case the effect of Vietnam-era military service on the earnings of veterans later in life (Angrist, 1990). In the 1960s and early 1970s, young men were at risk of being drafted

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for military service. Concerns about fairness led to the institution of a draft lottery in 1970 that was used to determine priority for conscription in cohorts of 19-year-olds. A natural instrumental variable for the Vietnam veteran treatment effect is draft-eligibility status, since this was determined by a lottery over birthdays. In particular, in each year from 1970 to 1972, random sequence numbers (RSNs) were randomly assigned to each birth date in cohorts of 19-year-olds. Men with lottery numbers below an eligibility ceiling were eligible for the draft, while men with numbers above the ceiling could not be drafted. In practice, many draft-eligible men were still exempted from service for health or other reasons, while many men who were draft-exempt nevertheless volunteered for service. So veteran status was not completely determined by randomized draft eligibility; eligibility and veteran status are merely correlated.

For white men who were at risk of being drafted in the 1970–71 draft lotteries, draft-eligibility is clearly associated with lower earnings in years after the lottery. This can be seen in Table 2, which reports the effect of randomized draft-eligibility status on average Social Security-taxable earnings in column (2). Column (1) shows average annual earnings for purposes of comparison. For men born in 1950, there are significant negative effects of eligibility status on earnings in 1970, when these men were being drafted, and in 1981, ten years later. For example, the 1981 estimate for whites is –436 dollars. In contrast, there is no evidence of an association between eligibility status and earnings in 1969, the year the lottery drawing for men born in 1950 was held but before anyone born in 1950 was actually drafted.

Instrumental variables estimates of the effects of military service on US white men born 1950

Earnings year	Earnings		Veteran status		Wald estimate of veteran effect
	Mean	Eligibility effect	Mean	Eligibility effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	–435.8	0.267	0.159	–2,741
	(210.5)		(.040)	(1,324)	
1970	2,758	–233.8			–1,470
	(39.7)			(250)	
1969	2,299	–2.0			
		(34.5)			

Notes: Figures are in nominal US dollars. There are about 13,500 observations with earnings in each cohort. Standard errors are shown in parentheses.

Sources: Adapted from Angrist (1990, Tables 2 and 3), and unpublished author tabulations. Earnings data are from Social Security administrative records. Veteran status data are from the Survey of Income and Program Participation.

Because eligibility status was randomly assigned, the claim that the estimates in column (2) represent the causal effect of *draft eligibility* on earnings seems uncontroversial. An additional assumption embodied in equation (6) is that the only reason eligibility affects earnings is military service. Given this, the only information required to go from draft-eligibility effects to veteran-status effects is the denominator of the Wald estimator, which is the effect of draft-eligibility on the probability of serving in the military. This information is reported in column (4) of Table 2, which shows that draft-eligible men were 0.16 more likely to have served in the Vietnam era. For earnings in 1981, long after most Vietnam-era servicemen were discharged from the military, the Wald estimates of the effect of military service reported in column (5) amount to about 15 percent of earnings. Effects were even larger in 1970, when affected soldiers were still in the army.

IV with heterogeneous treatment effects

The constant-effects assumption is clearly unrealistic. We'd like to allow for the fact that some men may have benefited from military service while others were undoubtedly hurt by it. In general, however, IV methods fail to capture either ATE or ATET in a model with heterogeneous treatment effects. Intuitively, this is because only a subset of the population is affected by any particular instrumental variable. In the draft lottery example, many men with high lottery numbers volunteered for service anyway (indeed, most Vietnam veterans were volunteers), while many draft-eligible

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
men nevertheless avoided service. The draft lottery instrument is not informative about the effects of military service on men who were unaffected by their draft-eligibility status. On the other hand, there is a sub-population who served solely because they were draft-eligible, but would not have served otherwise. Angrist, Imbens and Rubin (1996) call the population of men whose treatment status can be manipulated by an instrumental variable the set of *compliers*. This term comes from an analogy to a medical trial with imperfect compliance. The set of compliers are those who 'take their medicine', that is, they serve in the military when draft-eligible but they do not serve otherwise.

Under reasonably general assumptions, IV methods can be relied on to capture the causal effect of treatment on compliers. The average causal effect for this group is called a local average treatment effect (LATE), and was first discussed by Imbens and Angrist (1994). A formal description of LATE requires one more bit of notation. Define potential treatment assignments D_{0i} and D_{1i} to be individual i's treatment status when Z_i equals 0 or 1. One of D_{0i} or D_{1i} is counterfactual since observed treatment status is

$$D_i = D_{0i} + Z_i(D_{1i} - D_{0i}).$$

The key identifying assumptions in this setup are (*a*) conditional independence, that is, that the joint distribution of $\{Y_{1i}, Y_{0i}, D_{1i}, D_{0i}\}$ is independent of Z_i; and (*b*) monotonicity, which requires that either $D_{1i} D_{0i}$ for all i or vice versa. Monotonicity requires that, while the instrument might have no effect on some individuals, all of those who are affected should be affected in the same way (for example, draft eligibility can only make military service more likely, not less). Assume without loss of generality that monotonicity holds with $D_{1i} \geq D_{0i}$. Given these two assumptions, the Wald estimator consistently estimates LATE, written formally as $E[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}]$. In the draft lottery example, this is the effect of military service on those veterans who served because they were draft eligible but would not have served otherwise. In general, LATE compliers are a subset of the treated. An important special case where LATE=ATET is when D_{0i} equals zero for everyone. This happens in a social experiment with imperfect compliance in the treated group and no one treated in the control group.

IV Details

Typically, covariates play a role in IV models, either because the IV identification assumptions are more plausible conditional on covariates or because of statistical efficiency gains. Linear IV models with covariates can be estimated most easily by two-stage least squares (2SLS), which can also be used to estimate models with multi-valued, continuous, or multiple instruments. See Angrist and Imbens (1995) or Angrist and Krueger (2001) for details and additional references.

See Also

- instrumental variables
- matching estimators
- regression-discontinuity analysis
- Rubin causal model
- selection bias and self-selection
- two-stage least squares and the *k*-class estimator

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Angrist, J. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review* 80, 313–35.

Angrist, J. 1991. Grouped-data estimation and testing in simple labor-supply models. *Journal of Econometrics* 47, 243–266.

Angrist, J. 1998. Estimating the labor market impact of voluntary military service using Social Security data on military applicants. *Econometrica* 66, 249–88.

Angrist, J. and Imbens, G. 1995. Two-stage least squares estimates of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 90, 431–42.

Angrist, J., Imbens, G. and Rubin, D. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91, 444–55.

Angrist, J. and Krueger, A. 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives* 15(4), 69–85.

Ashenfelter, O. 1978. Estimating the effect of training programs on earnings. *Review of Economics and Statistics* 6, 47–57.

Ashenfelter, O. and Card, D. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics* 67, 648–60.

Dehejia, R. and Wahba, S. 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 1053–62.

Heckman, J. 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475–92.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–61.

Heckman, James J. and Robb, R. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman and B. Singer. New York: Cambridge University Press.

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–70.

Imbens, G. and Angrist, J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 467–75.

Quandt, R. 1958. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* 53, 873–80.

Rosenbaum, P. and Rubin, D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66, 688–701.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Rubin, D. 1977. Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* 2, 1–26.

Wald, A. 1940. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* 11, 284–300.

How to cite this article

Angrist, Joshua D. "treatment effect." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000206> doi:10.1057/9780230226203.1738

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

trend/cycle decomposition

Charles R. Nelson

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The decomposition of economic time series is motivated by the idea that distinct forces account for long-term growth, for variation over a time frame associated with the business cycle, and though the seasons. While the latter is typically suppressed by ‘seasonal adjustment’, the issue of how to separate trend from cycle in series such as GDP has been hotly debated since the 1970s and remains unsettled. Surprisingly varied patterns follow from alternative approaches, some placing the bulk of variation into the cycle/trend following a smooth line, others attributing shifts in level to ‘permanent shocks’ to the trend.

Keywords

ARMA models; business cycles; ergodicity; filtering; identification; Kalman filter; random walks; trend/cycle decomposition; unobserved components models

Article

Macroeconomists distinguish between the forces that cause long-term growth and those that cause temporary fluctuations such as recessions. The former include population growth, capital accumulation, and productivity change, and their effect on the economy is permanent. The latter are generally monetary shocks such as shifts in central bank policy that affect the real economy through price rigidities that cause output to deviate temporarily from its long-run path. This conceptual dichotomy motivates the decomposition of aggregate output, real GDP, into two components: the *trend* which accounts for long-term change, and the *cycle* which is a short-term deviation from trend. While economists no longer believe the ‘business cycle’ to be deterministically periodic, that terminology remains. Seasonal variation could be a third component, though it has been suppressed in ‘seasonally adjusted’ data such as GDP.
This suggests we may express the natural log of GDP (or any other ‘trending’ time series), denoting the observation at time t by ‘y_t’, as follows:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$y_t = \tau_t + c_t.$$

Here τ_t denotes the value of the trend and c_t the cycle at time t, neither of which is observed directly. Since this single equation cannot be solved directly for the unknown trend and cycle, additional assumptions are required for 'identification', a procedure which allows estimates of them to be calculated from the GDP data. The fundamental identifying assumption is that the cycle component is temporary, that it dies out after a sufficiently long time. However, this assumption of 'stationarity' or 'ergodicity', which distinguishes it from trend, which is permanent, does not suffice by itself to achieve identification. More has to be said about the nature of the trend.

The simplest specification of trend is to make τ_t a linear function of time where the slope is the long-term growth rate. A second identifying assumption is that trend should account for as much of the variation in the data as possible, minimizing the amplitude of the implied cycle. This is achieved by least squares regression of y_t on time and the estimated trend is $\hat{\tau}_t = a + b\ time$ where a and b are estimates of intercept and slope respectively. The implied cycle component is then $\hat{c}_t \equiv y_t - \hat{\tau}_t$. Though successful in accounting for a large fraction of the change in GDP over long periods, this approach implies cycles of extraordinary length, well beyond the roughly seven years between recession dates identified by the National Bureau of Economic Research for the United States, and the pattern is contrary to economic intuition (for the United States the 1970s, a decade of poor economic performance, were well above the trend line while the 1990s, a decade of prosperity, were well below trend). A more flexible trend function is clearly called for, but quadratic or higher-order polynomials in time imply unstable paths when extrapolated into the future. Perron (1989) suggested a segmented trend function allowing for an occasional change level or slope to be captured by dummy variables.

A general approach to estimating a flexible and adaptive trend is filtering, where estimated trend is a weighted average of adjacent observations. Here it is the weighting scheme which identifies the components. For example, $\hat{\tau}_t = .25 \cdot y_{t-1} + .50 \cdot y_t + .25 \cdot y_{t+1}$ applies symmetric though unequal weights to the current observations and its immediate neighbours. No filter is perfect in the sense of revealing the actual trend, but a desirable filter is one that extracts as much of the trend as possible from the data. A criterion for choosing a filter would be that it produces cycles having characteristics that match our notions of the business cycle, for example that recessions occur on average about every seven years. A widely used filter that does this is the Hodrick and Prescott (1980), filter which penalizes deviations from trend and changes in trend through a loss function.

The distinction between trend and cycle implies that the forecast of GDP far in the future must be the trend, since the cycle will die away. The approach to trend/cycle decomposition proposed by Beveridge and Nelson (1981) turns this conclusion on its head by proposing that the trend at a date in time be defined as the forecast of the distant future (adjusted for average growth). Specifically, they estimate an autoregressive moving average (ARMA) time series model for the growth rate and compute the forecast of the level into the distant future, adjusting for average growth. The resulting measure of trend shows whether actual GDP is above or below its forecast growth path, the difference being the cycle. Since

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
parameters of the ARMA model are identified, and computation of forecasts is straightforward, the Beveridge–Nelson decomposition is identified. It turns out that the trend component is a random walk with drift regardless of the specific ARMA model, and this accords with the intuition that only unexpected shocks can affect a long horizon forecast.
To obtain the general expressions for the components we rearrange the ARMA model as:

$$\varphi(L)\Delta y_t = \theta(L)\varepsilon_t$$

$$\Delta y_t = \psi(L)\varepsilon_t$$

where the average growth rate has been suppressed, the statistical shock ε_t is serially random, Δ denoted first difference, and L is the lag operator, and $\psi(L)=\theta(L)/\phi(L)$. The growth rate of GDP can be thought of as a weighted history of all past shocks where the coefficient of ε_{t-k} is ψ_k plus the expected average growth rate μ. It may be shown that an algebraically equivalent expression is

$$y_t = \psi(1)\sum_{k=0}^{\infty}\varepsilon_{t-k} - \tilde{\psi}(L)\varepsilon_t\tilde{\psi}_k = \sum_{j=k+1}^{\infty}\psi_j.$$

Note that the first term is the sum of all past shocks each with weight equal to the total effect of all past shocks. The second term may be shown to be a stationary time series with mean zero. Thus the trend is always a random walk regardless of the ARMA model.
For example, growth in US GDP is roughly an AR(1) process with coefficient .25, so the effect of a shock on the trend is $\psi(1)=1/(1-.25)=1.33$. This illustrates the surprising implication that the trend component may be highly variable; indeed, the results obtained by Beveridge and Nelson imply that variation in observed GDP is largely the result of variation in the trend component and is therefore permanent.
Unobserved components models identify trend and cycle by specifying a separate and specific stochastic process for each. The trend is generally assumed to be a random walk with drift, allowing it to account for long-term growth while permitting it to be shifted by stochastic shocks. The cycle is assumed to be a process that is stationary in the sense of reverting to a mean over time. (The mean of the cycle is zero for symmetric variation around trend, but evidence exits for asymmetric cycles with a negative mean.) This approach was introduced to economics by Harvey (1985) and Clark (1987). An example would be the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
following:

$$\tau_t = \tau_{t-1} + \mu + \eta_t c_t = \varphi \cdot c_{t-1} + \varepsilon_t.$$

The parameter μ is the long-term growth rate, the shock η is random and may be positive or negative, the parameter ϕ measures the persistence of the cycle, and shocks ε drive the cycle. The two shocks are often assumed to be uncorrelated, which reduces the number of parameters to be estimated by one but may also place an unwarranted restriction on the relation between the two components. More generally the cycle process may have a higher-order ARMA specification. Identification of the parameters depends on whether a specific model implies a sufficient number of estimable parameters in the corresponding ARMA reduced form representation of Δy_t (corresponding to identification of simultaneous equation models). Given an identified model and parameter estimates, the estimated trend and cycle may be computed using the Kalman filter.
A useful result is that the random walk trend in the unobserved components model is identified even if its parameters are not identified. Morley, Nelson and Zivot (2003) show that the Beveridge–Nelson trend is always the conditional expectation of the trend component given past data. What identifies the trend is the random walk specification for the trend along with the assumption that the cycle process does not persist indefinitely. Thus, the long-horizon forecast reflects only the trend, and such forecasts can always be computed from the reduced form ARMA model.

See Also

- business cycle measurement
- data filters
- state space models
- time series analysis
- unit roots

Bibliography

Beveridge, S. and Nelson, C.R. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* 7, 151–74.

Clark, P.K. 1987. The cyclical component of U.S. economic activity. *Quarterly Journal of Economics* 102, 797–814.

Harvey, A.C. 1985. Trends and cycles in macroeconomic time series. *Journal of Business and Economic*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Statistics 3, 216–27.

Hodrick, R.J. and Prescott, E.C. 1980. Postwar US business cycles: an empirical investigation. Working paper, Carnegie-Mellon University.

Hodrick, R.J. and Prescott, E.C. 1997. Postwar U.S. business cycles: an empirical investigation. *Journal of Money Credit and Banking* 29, 1–16.

Morley, J.C., Nelson, C.R. and Zivot, E. 2003. Why are the Beveridge–Nelson and unobserved-components decompositions of GDP so different? *Review of Economics and Statistics* 85, 235–43.

Perron, P. 1989. The great crash, the oil price shock, and the unit root hypothesis. *Econometrica* 57, 1361–401.

How to cite this article

Nelson, Charles R. "trend/cycle decomposition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000221>
doi:10.1057/9780230226203.1739

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Triffin, Robert (1911 – 1993)

Albert Kervyn
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

European Monetary System; European Payment Union; gold-exchange standard; International Monetary Fund; international reserves; Special Drawing Rights; Triffin Plan; Triffin, R.

Article

Triffin was born in 1911 in the pleasant Belgian village of Flobecq. After a brilliant school career, a scholarship enabled him to study law and economics at the university of Louvain. Another scholarship sent him to Harvard, where he got a thorough grounding in theory with Schumpeter and Leontief. His 1938 dissertation, ‘Monopolistic Competition and General Equilibrium Theory’, earned a Wells Prize and was published in 1940.

After a brief return to Belgium, he was appointed instructor at Harvard, and was soon cut off from Europe by the outbreak of the war. In 1942 he joined the Federal Reserve Board to organize a research section on Latin America. This launched him on his parallel career of advising central banks on reform of monetary and exchange arrangements. He rapidly developed what was to remain a main characteristic of his work in this area: a flair for practical suggestions and an imagination that provided alternatives in case of political objections to the first proposals. His success in a number of Latin American countries led to his appointment in 1946 as head of the exchange control division in the newly-created International Monetary Fund (IMF). Moving to Europe as IMF chief representative, he developed a proposal for the European Clearing Union. Having later transferred to the State Department, he succeeded in negotiating his proposal through OEEC, and became the recognized father of the European Payment Union of 1950.

Triffin left the State Department after a policy disagreement, and went to Yale, where he stayed from 1951 to 1977. There he published his two classic works: *Europe and the Money Muddle* in 1957; *Gold and the Dollar Crisis* in 1959 and 1960. The first book reviews the European monetary experience, using an integrated analysis of the money supply and the balance of payments. It proclaims the end of the ‘dollar shortage’ and explains the dilemma of the gold-exchange standard in the absence of an adequate supply of gold: either the key-currency country maintains equilibrium in its balance of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
payments, and other countries will experience a shortage of the reserves needed to support an expansion of trade and transactions, with an attendant brake on growth; or the desirable growth of world reserves will be preserved only through persistent increases in the liabilities of the key-currency country, raising increasing doubts about its ability to redeem such liabilities, especially when they begin to exceed the country's dwindling gold reserves. The second book, shorter and less analytic, goes further: it contains a bold prophecy of a dollar glut, bound in time to bring down the gold-exchange standard. The fate experienced by sterling in the interwar years was bound to catch up with the dollar as well.

The second part of the book contained the famous ‘Triffin Plan’ to obviate these dangers: on the one hand, the controlled creation of an international reserve instrument by the IMF; on the other hand, regional monetary arrangements, with emphasis on European integration.

The announced dollar glut did indeed develop over the following years with the predicted consequences; the gold parity rate was first abandoned, then the dollar became inconvertible. The first policy prescription was timidly followed with the creation of Special Drawing Rights, but without the essential element of conversion of dollar balances. Indeed the excessive accumulation of such balances was later denounced by Triffin as the main factor in the inflationary development of the 1970s, and as an unwarranted capital inflow into the USA, which on the contrary should be a leading exporter of capital.

The other prescription was more successful, and Triffin moved back to Europe (and the University of Louvain) in order to participate more fully in the emergence and development of the European Monetary System.

He was throughout a most active participant in the debate on money and exchanges. In a continuous stream of studies, papers and memoranda, he pressed forward for advances on the two fronts of international monetary reform and European integration.

His reputation as an analyst and skilled deviser of techniques led to his being called in as a consultant all over the world, on domestic as well as regional monetary issues.

Later in life, Triffin remained a much concerned ‘citizen of the world’ and applied his policy-oriented brand of ‘economics of persuasion’ not only to his particular area of expertise, but also to the issues of development and disarmament.

Selected works

As a policy-oriented persuader, Triffin has been a prolific writer. In his attempt to reach a wide public for his ideas, he gives the impression of never having refused a contribution. As a result a complete bibliography would include well over 300 items – many in less accessible publications. The following list is therefore highly selective.

1940. *Monopolistic Competition and General Equilibrium Theory*. Cambridge, MA: Harvard University Press.

1957. *Europe and the Money Muddle*. New Haven: Yale University Press.

1960a. *Statistics of Sources and Uses of Finance, 1948–58*. Paris: OEEC.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1960b. *Gold and the Dollar Crisis*. New Haven: Yale University Press.

1964. *The Evolution of the International Monetary System*. Princeton Studies in International Finance No. 12. Princeton: Princeton University Press.

1966. *The World Money Maze*. New Haven: Yale University Press.

1969. The thrust of history in international monetary reform. *Foreign Affairs* 47, 477–92.

1971. The use of SDR finance for collectively agreed purposes. *Banca Nazionale del Lavoro Quarterly Review* 96(March), 3–12.

1973. The role of developing European monetary union. In *Europe and the Evolution of the International Monetary System*, ed. A.K. Swoboda. Leiden: Sijthoff.

1975. The Community and the disruption of the world monetary system. *Banca Nazionale del Lavoro Quarterly Review* 112(March), 3–35.

1976. Size, sources and beneficiaries of international reserve creation, 1970–74. In *Economic Progress, Private Values and Public Policy*, ed. B. Balassa and R. Nelson. New Haven: Yale University Press.

1978a. ‘Europe and the money muddle’ revisited. *Banca Nazionale del Lavoro Quarterly Review* 124 (March), 49–65.

1978b. *Gold and the Dollar Crisis: Yesterday and Today*. Princeton Essays in International Finance No. 132. Princeton: Princeton University Press.

1978c. The international role and fate of the dollar. *Foreign Affairs* 57, 269–86.

1981a. Le système monétaire Européen dans le cadre du système monétaire mondial. *Banque* 55, 535–40.

1981b. An economist's career: What? Why? How? *Banca Nazionale del Lavoro Quarterly Review* 139 (September), 239–59.

1981c The first two years of FECOM transactions. Economic Papers No. 2. Brussels: Directorate-General for Economic and Financial Affairs, European Commission, July.

1984a. How to end the world ‘infession’. In *Europe's Money*, ed. R.S. Masera and R. Triffin. Oxford: Clarendon Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1984b. Une tardive autopsie du Plan Keynes de 1943. *Revue d'économie politique*.

1986. Correcting the world monetary scandal. *Challenge* 28(January/February), 4–14.

How to cite this article

Kervyn, Albert. "Triffin, Robert (1911–1993)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000126>
doi:10.1057/9780230226203.1740

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Trotsky, Lev Davidovitch (1879 – 1940)

Richard B. Day
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bukharin, N. I.; collectivization; comparative coefficients; industrialization; planning; socialism; Stalin, J. V.; trade monopoly; trade unions; Trotsky, L. D.

Article

Born in 1879, the son of Jewish farmers living near the Black Sea, Trotsky became an important political figure by the time of the Second Congress of the Russian Social Democratic Party in 1903. Disagreeing with Lenin's centralizing view of party organization, Trotsky either favoured the Mensheviks or attempted to mediate between them and the Bolsheviks until making his peace with Lenin in 1917. In the 1905 Revolution he served as chairman of the St Petersburg Soviet, drawing upon that experience to develop the theory of ‘permanent revolution’ in his book *Results and Prospects*. In the 1917 Revolution Trotsky ranked second only to Lenin among Bolshevik party leaders. He orchestrated the seizure of power and subsequently organized and led the Red Army in the civil war. During the early 1920s Trotsky's political influence waned, and by the middle of the decade he became the political leader and intellectual mentor of the Left Opposition to Stalin. Defeated by Stalin in the intra-party struggle, in 1929 Trotsky was deported from the Soviet Union. In exile he edited *Biulleten’ Oppozitsii* (Bulletin of the Opposition) and published numerous other writings critical of Stalinist policy, the most important being *The Revolution Betrayed*. Unable to answer Trotsky's criticisms on intellectual grounds, in August 1940 Stalin replied in the only way he knew: he had Trotsky assassinated in Mexico, his last place of exile.

In *Results and Prospects* (first published in 1906), Trotsky predicted that Russian backwardness would guarantee the revolution in permanence. Surrounded by stronger enemies, the Russian state had prevented the nobility from becoming politically independent. The nobility were mere tax collectors, extracting revenue from the peasants in order to promote development; and the bourgeoisie, likewise, were weaker than their Western counterparts, for much of the economy was built with foreign loans, serviced by grain exports. The proletariat, in contrast, enjoyed disproportionate strength. Few in number, Russian workers were concentrated in large factories organized around foreign technology. Trotsky

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
predicted that the proletariat would overthrow the autocracy, by-passing the bourgeois revolution, but would then confront a counter-revolutionary alliance when it implemented its programme. The counter-revolution would be supported by Germany, Austria and France, who would be anxious to prevent the revolution's spread and to safeguard their investments. When these countries mobilized, however, they would drive their own workers to revolt, thereby making the revolution permanent both domestically and internationally.

Aware of Russia's historical dependence on the world economy, Trotsky characteristically viewed economic issues in an international context. Modern industry, he believed, had become so capital intensive that production could only be profitable through specialization in service of the world market. It was in the nature of socialism to emancipate the productive forces from the fetters of the nation state. A victory of the proletariat in the leading countries would mean ‘a radical restructuring of the very economic foundation in correspondence with a more productive international division of labour, which is alone capable of creating a genuine foundation for a socialist order’ (Trotsky Archives No. T-3148). When the international revolution did not come to Soviet Russia's aid as Trotsky had expected, he continued to insist that industrialization must draw upon the resources of the world market. Opposing Stalin's notion of an isolated socialist state (Socialism in One Country), he argued that ‘a properly regulated growth of export and import with the capitalist countries prepares the elements of the future commodity and product exchange [which will prevail] when the European proletariat assumes power and controls production’ (Trotsky Archives No. T-3034). Soviet Russia's relation to the West would involve a dialectic of cooperation and struggle in which the Soviet state would regulate its ‘dependence’ on capitalism through its monopoly of foreign trade. The alternative, the Stalinist vision of autarky, would mean reliance ‘on the curbed and domesticated productive forces, that is … on the technology of backwardness’ (Trotsky, 1941, p. 53).

Uppermost in Trotsky's mind throughout the 1920s was the need not only to preserve access to foreign technology, but also to reduce domestic prices in order to maintain the trade monopoly. In 1923 he warned the party that ‘Contraband is inevitable if the difference between external and internal prices goes beyond a certain limit … contraband, comrades … undermines and washes away the monopoly’ (*Dvenadstatyi s'ezd RKP* (b) (1923), p. 372; 12th Congress of the Russian Communist Party (Bolsheviks)). Without this protection for new Soviet industries, planned growth would be impossible. For the promotion of new industrial construction, Trotsky proposed to supplement domestic tax revenues by taking advantage of Europe's need for foreign markets and by pursuing all manner of credits:

> What does foreign credit do for our economic development? Capitalism makes advances to us against our savings which do not yet exist … As a result, the foundations of our development are extended … The dialectics of historical development have resulted in capitalism becoming for a time the creditor of socialism. Well, has not capitalism been nourished at the breasts of feudalism? History has honoured the debt. (*Pravda*, 20 September 1925)

In addition to making use of foreign credits, Trotsky hoped to resume the tsarist pattern of exporting grain in exchange for finished goods. In 1925 he predicted that the Soviet economy would be unable to satisfy more than a fraction of its need for new equipment:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> We must not … forget for a moment the great mutual dependence which existed between the economies of tsarist Russia and world capital. We must just bring to mind the fact that nearly two-thirds of the technical equipment in our works and factories used to be imported from abroad. This dependence has hardly decreased in our own time, which means that it will scarcely be economically profitable for us in the next few years to produce at home the machinery we require, at any rate, more than two-fifths of the quantity, or at best more than half of it. (*Pravda*, 20 September 1925)

Trotsky hoped to reconcile a high level of foreign trade with socialist protectionism through strict determination of priorities. Soviet industries should economize on scarce capital, specialize in those products in greatest demand, standardize output and reduce costs, while leaving the remaining needs to be met by low-cost imports. A system of comparative coefficients should be devised by the planners, comparing the cost and quality of Soviet products with foreign competition. A poor coefficient would then signal the advisability of imports in the short run and of re-equipment in the long run, as new resources became available. ‘A comparative coefficient is the same for us as a pressure gauge for a mechanic on a locomotive. The pressure of foreign production is for us the basic factor of our economic existence. If our relation to this production is [unsatisfactory], then foreign production will sooner or later pierce the trade monopoly’ (*Ekonomicheskaia Zhizn’*, 18 August 1925).

In spite of his balanced approach to industrialization, official Soviet historiography insists that Trotsky was a ‘super-industrializer’, determined to plunder the peasantry. In reality he attempted more systematically than any of his contemporaries to avert the crisis of forced industrialization by balancing the needs of the peasantry against those of industry through a policy of ‘commodity intervention’. To the extent that export-oriented growth clearly depended upon the peasants bringing grain to market, Trotsky was quite aware that the most urgent consumer needs would also have to be satisfied through imports. The world market was to function as a ‘reserve’ for both light and heavy industry. The ‘goods famine’, or the chronic shortage of consumer goods, was ‘obvious and incontestable proof that the distribution of national economic resources between state industry and the rest of the economy has … acquired the necessary proportionality’ (Trotsky Archives No. T-2983). The real enemies of the peasantry, in Trotsky's view, were the authors of Socialism in One Country – Stalin, who saw only the needs of the machine-building industries, and Bukharin, who urged the peasant to ‘enrich’ himself without seriously considering the need to provide consumer goods upon which these savings might be spent.

It was Trotsky's concern for the legitimate needs of workers and peasants alike which led him in the 1930s to reconsider the role of market forces, for a time at least, in socialist planning. As early as 1925 he had warned that it was ‘impossible to push industrialization forward with the aid of unreal credits’ (Trotsky, 1955, p. 186). During the first five-year plan he called for restraints upon the inflationary financing of heavy industry and ‘strict financial discipline’, even at the expense of closing down enterprises. A stable currency, in turn, would provide an instrument whereby the masses themselves could democratically control production decisions from below. ‘The innumerable living participants in the economy’, Trotsky wrote in 1932,

> state and private, collective and individual, must announce their needs and their respective intensities not only through the statistical calculations of the planning commissions, but

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> also by the direct pressure of supply and demand. The plan … [must be] verified, and in an important measure must be achieved through the market. (*Biulleten' Oppozitsii* 31 (1932), p. 8)

A planned market, free trade unions, and restoration of Soviet democracy: these were the three elements without which any talk of socialism was a mockery.

> If there existed the universal mind described in the scientific fantasy of Laplace – a mind which might simultaneously register all the processes of nature and society, measure the dynamic of their movement and forecast the results of their interactions – then, of course, such a mind could *a priori* draw up a faultless and exhaustive economic plan, beginning with the number of hectares of wheat and ending with buttons on a waistcoat. True, it often appears to the bureaucracy that it possesses just such a mind: and that is why it so easily emancipates itself from control by the market and by soviet democracy. The reality is that the bureaucracy is cruelly mistaken in its appraisal of its own spiritual resources. (*Biulleten' Oppozitsii* 31 (1932), p. 8)

In *The Revolution Betrayed* (1937), his most thorough critique of Stalinist 'planomania', Trotsky concluded that the real basis of bureaucratic power had nothing to do with Stalin's pompous claims of industrial triumphs; the horrible truth was that the whole bureaucratic edifice had come to rest upon nothing more profound or despicable than an ability to manufacture poverty. Queues were the foundation of Soviet power and the innermost secret of the police state:

> The basis of bureaucratic rule is the poverty of society in objects of consumption. When there are enough goods in a store, the purchasers can come whenever they want to. When there are few goods, the purchasers are compelled to stand in line. When the lines are very long, it is necessary to appoint a policeman to keep order. Such is the starting point of the Soviet bureaucracy. It 'knows' who is to get something and who has to wait. (Trotsky, 1937, p. 112)

Historians will continue to debate whether Trotsky's policies might have avoided forced collectivization and the excesses of Stalin's five-year plans. On one point, however, there can be no dispute: Trotsky was perfectly correct to conclude that Stalin's pursuit of autarky had more in common with the ideals of Hitler than with those of Marx. The Russian revolution, confined to a single backward country, did not lead to the emancipation of the proletariat. Trotsky attempted to reinterpret and apply Marxism to the unexpected conditions of an isolated revolutionary experiment. He did not win the battle against Stalin. He did, however, help to explain and attempt to avert one of the great tragedies of the 20th century.

Selected works

1921. *Terrorism and Communism*. Ann Arbor: University of Michigan Press, 1963.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1923. *The New Course*. Ann Arbor: University of Michigan Press, 1965.

1926. *Towards Socialism or Capitalism*? London: Methuen & Co.

1927. *The Platform of the Left Opposition*. London: New Park Publications, 1963.

1930. *My Life*. New York: Grosset & Dunlap, 1960.

1930. *The History of the Russian Revolution*. Trans. M. Eastman, London: Victor Gollancz, 1965.

1937. *The Revolution Betrayed*. New York: Pioneer Publishers, 1945.

1941. *Stalin: An Appraisal of the Man and His Influence*. London: Hollis & Carter, 1947.

1962. *The Permanent Revolution and Results and Prospects*, trans. J.G. Wright and B. Pearce. London: New Park Publications.

1973. *Biulleten' Oppozitsii* (Bulletin of the Opposition). New York: Monad.

Bibliography

Day, R.B. 1973. *Leon Trotsky and the Politics of Economic Isolation*. London: Cambridge University Press.

Deutscher, I. 1954. *The Prophet Armed. Trotsky: 1879–1921*. London: Oxford University Press.

Deutscher, I. 1959. *The Prophet Unarmed. Trotsky: 1921–1929*. London: Oxford University Press.

Deutscher, I. 1963. *The Prophet Outcast. Trotsky: 1929–1940*. London: Oxford University Press.

Figes, O. 1996. *A People's Tragedy: A History of the Russian Revolution*. New York: Viking Press.

Howe, I. 1978. *Leon Trotsky*. New York: Viking.

Knez-Paiz, B. 1978. *The Social and Political Thought of Leon Trotsky*. Oxford: Clarendon Press.

Thatcher, I. 2003. *Trotsky*. London: Routledge.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Day, Richard B. "Trotsky, Lev Davidovitch (1879–1940)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000129>
doi:10.1057/9780230226203.1741

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

trust in experiments

Iris Bohnet
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Trust is the willingness to make oneself vulnerable to another person's actions, based on beliefs about that person's trustworthiness. This article focuses on interpersonal trust and trustworthiness between two people, a trustor and a trustee, as measured in laboratory experiments. A trustee behaves trustworthily if he voluntarily refrains from taking advantage of the trustor's vulnerability. Trust applies to all transactions where the outcome is partly under the control of another person and not fully contractible. The article discusses measurement issues, the motives for and influences on trust and trustworthiness (incentives, repetition and demographic variables) and questions of external validity.

Keywords

cooperation; fairness; folk theorem; investment game; public goods games; reciprocity; repeated games; social preferences; trust; trust in experiments; trust games; ultimatum games

Article

Trust is the willingness to make oneself vulnerable to another person's actions, based on beliefs about that person's trustworthiness.
This article focuses on interpersonal trust between two people, a trustor and a trustee. A trustee behaves trustworthily if he voluntarily refrains from taking advantage of the trustor's vulnerability. Trust applies to all transactions where the outcome is partly under the control of another person and not fully contractible, for example, between employers and employees or patients and doctors. Trust and trustworthiness are typically measured in surveys or laboratory experiments. We shortly discuss some survey evidence but focus on behavioural measures of trust.

Measurement

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Trust attitudes have typically been measured by the following survey question (for example, used in the General Social Survey and the World Values Survey): ‘Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?’ Based on this question, trust has declined dramatically across the world since the 1960s. There are noticeable cross-country differences, with Scandinavians most and Latin Americans least likely to trust others. An empirical literature building on Knack and Keefer (1997) shows that trust attitudes are positively correlated with various measures of a country's economic performance.

Around 1990, two seminal papers started a new wave in the economic research on trust. In 1988, Camerer and Weigelt employed a binary-choice trust game, and in 1995 Berg, Dickhaut and McCabe the ‘investment game’ to study trust. In the binary-choice trust game, the trustor decides between a sure outcome and trust. If she chooses the sure thing, she and her trustee both receive (S, S). If she is willing to trust, both either end up with a moderate payoff exceeding S (M, M), or the trustor receives a lower payoff than if she had not trusted, and the trustee the highest possibly payoff, (L, H). Thus, for the trustor, $M > S > L$, and for the trustee, $H > M > S$. In the investment game, a trustor and a trustee are endowed with a certain amount of money, A (in some experiments, only trustors are endowed). The trustor can send any amount, $X \leq A$, to the trustee. X is multiplied by k>1 by the experimenter. In most experiments, k=3. Trustees receive kX and then decide how much of it, $Y \leq A + kX$, to return to their trustor. The final payoffs are $A - X + Y$ for the trustor and $A + kX - Y$ for the trustee. X is commonly referred to as trust and Y, or more precisely, Y/X, measures trustworthiness for X>0. In both games, the equilibrium prediction based on selfish money-maximization and rationality is zero trustworthiness and zero trust.

The relationship between trust attitudes, as measured in surveys, and trust behaviour, as measured in experiments, is not clear. Some have found that they are related (for example, Fehr and Schmidt, 2002), others that they are not (for example, Glaeser et al., 2000). While the investment game and the binary-choice trust game have turned out to be the most widely used games to study trust experimentally, related games include the ‘gift exchange game’, the ‘moonlighting game,’ and standard public goods games (for a review, see Camerer, 2003).

What motivates trust and trustworthiness?

Trust is based on preferences, namely, the willingness to be vulnerable to someone else, and on expectations, namely, the belief about someone else's trustworthiness. A person's willingness to be vulnerable may be related to her attitudes to risk (for example, Eckel and Wilson, 2004), her social preferences (for example, Cox, 2004), and her willingness to accept the risk of betrayal (Bohnet and Zeckhauser, 2004). Bohnet and Zeckhauser introduced an analytical framework to disentangle the various motives, and show that people dislike making themselves vulnerable to the actions of another person more than to natural circumstances. This suggests betrayal aversion: people care not only about outcomes but also about how outcomes come to be. This finding was supported by neuroscientific evidence (Kosfeld et al. 2005).

The relevance of expectations of trustworthiness for trust has typically been measured by including a question about trustors’ beliefs. While this measure is not perfect, generally the relationship between expectations of trustworthiness and trust is very strong. For example, using a within-subject design with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
behavioural controls for risk and social preferences, Ashraf, Bohnet and Piankov (2006) found that expectations of trustworthiness explain most of the variance in trust in an investment game but that social preferences also matter.

Trustworthiness is based on trustees' social preferences, which may be related either to outcomes (for a survey, see Fehr and Schmidt, 2002) or to what the trustors' actions reveal about their intentions. In a seminal paper, Rabin (1993) introduced a theoretical model of intention-based preferences, reciprocity, into the literature. A large number of empirical studies suggests the importance of reciprocity in trust interactions (for example, Fehr, Gächter and Kirchsteiger, 1997) although outcome-based social preferences also play an important role for trustworthiness (for example, Cox, 2004; Ashraf, Bohnet and Piankov, 2006).

What influences trust and trustworthiness?

Incentives

According to most models, trustors should be more likely to trust the higher the expected returns are from trusting. Bohnet, Herrmann and Zeckhauser (2006) measured the elasticity of trust and found that trust is responsive both to changes in the likelihood and to the cost of betrayal in Western countries. However, this does not necessarily apply in other parts of the world. For example, in Persian Gulf countries people hardly responded to such changes. Instead, many basically demanded a guarantee of trustworthiness before trusting, suggesting substantial aversion to betrayal. In addition, incentives may also not work as predicted by theory if they not only affect behaviour directly but also exhibit an influence on preferences, thus either fostering or undermining people's willingness to accept vulnerability and be trustworthy voluntarily (Bohnet, Frey and Huck, 2001).

Repetition

Generally, people are more likely to trust and be trustworthy in repeated than in one-shot interactions. Theoretically, this result is expected in a traditional model when interactions are indefinitely repeated (folk theorem) but not in finitely repeated games. In support of the theory, experimental evidence suggests that trust and trustworthiness rates are generally higher in indefinitely than in finitely repeated games but they are also higher in the latter than in one-shot interactions. The equilibrium prediction of no trust and trustworthiness is generally refuted, although trust and trustworthiness rates typically drop substantially as the end of the game draws nearer (for example, Gächter and Falk, 2002).

Demographic variables

Generally, the evidence is not as conclusive as we might expect or wish. While in theory variables such as gender, race or country of origin should be easy to control for, experiments produce different results precisely because of the different sets of control variables and the different subject pools used. The most promising approaches include those identifying overarching frameworks able to account for a variety of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
studies. We discuss three such frameworks here: history of discrimination, societal organization and market integration.

Groups that historically have been discriminated against, such as women and minorities, are generally less likely to trust. At the same time, often these groups are more trustworthy (for example, Alesina and LaFerrara 2002; Buchan, Croson and Solnick, 2003; Eckel and Wilson, 2003).

Group-based societal organization based on long-standing relationships and repeated interactions within groups can substantially reduce the social uncertainty involved in trust. It is often referred to as ‘collectivist’ in contrast to the Western ‘individualist’ model of organization, which produces trust through more anonymous, institutional arrangements such as contracts and insurance. Trust in strangers has often been found to be higher in individualist (for example, the United States or Switzerland) than in collectivist countries (for example, Japan or the Persian Gulf countries), although the rather small number of studies and sample sizes does not allow any definite conclusions at this point (for example, Bohnet, Herrmann and Zeckhauser, 2006; but see also Croson and Buchan, 1999).

The degree of market integration is related to norms of cooperation and fairness in public goods and ultimatum games. Similarly, the norms of reciprocity typically found in trust experiments in developed countries seem to apply more strongly in societies in which goods and services are exchanged in the market rather than in informal reciprocal-exchange arrangements. Greig and Bohnet's survey of the evidence (2006) suggested that the positive relationship between trust and trustworthiness, normally taken to indicate reciprocity, is more pronounced in developed than in developing countries.

External validity

Experiments allow for maximum internal control. Concerns typically arising in field settings such as lack of randomization, selection and endogeneity can easily be addressed by experimental design. To address concerns about the subject pools experimentalists typically use, that is, North American or European students, experiments are now run with representative samples (for example, Fehr et al., 2002 in Germany) and with student and non-student subjects in other parts of the world (for example, Cardenas and Carpenter, 2005, for a survey). To directly test the external validity of trust experiments, Karlan (2005) ran investment games with members of a group lending association in Peru, and compared trustworthiness in the experiment with repayment rates. The more trustworthy subjects indeed were significantly more likely to repay their loans a year later.

See Also

- altruism in experiments
- behavioural game theory
- experimental economics
- public goods experiments
- reciprocity and collective action
- risk aversion
- social capital

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Alesina, A. and LaFerrara, E. 2002. Who trusts others? *Journal of Public Economics* 85, 207–34.

Ashraf, N., Bohnet, I. and Piankov, N. 2006. Decomposing trust and trustworthiness. *Experimental Economics* 9, 193–208.

Berg, J., Dickhaut, J. and McCabe, K.A. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10, 290–307.

Bohnet, I., Frey, B.S. and Huck, S. 2001. More order with less law: on contract enforcement, trust and crowding. *American Political Science Review* 95, 131–44.

Bohnet, I., Herrmann, B. and Zeckhauser, R. 2006. The requirements for trust in Gulf and Western countries. *Working paper*.

Bohnet, I. and Zeckhauser, R. 2004. Trust, risk and betrayal. *Journal of Economic Behavior and Organization* 55, 467–84.

Buchan, N., Croson, R. and Solnick, S. 2003. Trust and gender: an examination of behavior, biases, and beliefs in the investment game. Working paper, Wharton School, University of Pennsylvania.

Camerer, C.F. 2003. *Behavioral Game Theory*. Princeton: Princeton University Press.

Camerer, C.F. and Weigelt, K. 1988. Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56, 1–36.

Cardenas, J.C. and Carpenter, J. 2005. Experiments and economic development: lessons from field labs in the developing world. Working paper, Middlebury College.

Cox, J.C. 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46, 260–81.

Croson, R. and Buchan, N. 1999. Gender and culture: international experimental evidence from trust games. *American Economic Review* 89, 386–91.

Eckel, C.C. and Wilson, R.K. 2003. Conditional trust: sex, race and facial expressions in a trust game. Working paper, Rice University.

Eckel, C.C. and Wilson, R.K. 2004. Is trust a risky decision? *Journal of Economic Behavior and Organization* 55, 447–66.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fehr, E. and Schmidt, K. 2002. Theories of fairness and reciprocity – evidence and economic applications. In *Advances in Economics and Econometrics*, ed. M. Dewatripont, L. Hansen and S. Turnovsky. Cambridge: Cambridge University Press.

Fehr, E., Gächter, S. and Kirchsteiger, G. 1997. Reciprocity as a contract enforcement device: experimental evidence. *Econometrica* 64, 833–60.

Fehr, E., Fischbacher, U., von Rosenbladt, B., Schupp, J. and Wagner, G. 2002. A nation-wide laboratory-examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Schmollers Jahrbuch* 122, 519–42.

Gächter, S. and Falk, A. 2002. Reputation and reciprocity: consequences for the labour relation. *Scandinavian Journal of Economics* 104, 1–26.

Glaeser, E.L., Laibson, D.I., Scheinkman, J.A. and Soutter, C.L. 2000. Measuring trust. *Quarterly Journal of Economics* 115, 811–46.

Greig, F. and Bohnet, I. 2006. Is there reciprocity in a reciprocal-exchange economy? Evidence of gendered norms from a slum in Nairobi, Kenya. Working paper, Kennedy School of Government, Harvard University.

Karlan, D. 2005. Using experimental economics to measure social capital and predict financial decisions. *American Economic Review* 95, 1688–99.

Knack, S. and Keefer, P. 1997. Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics* 112, 1251–88.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U. and Fehr, E. 2005. Oxytocin increases trust in humans. *Nature* 435, 673–6.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83, 1281–302.

How to cite this article

Bohnet, Iris. "trust in experiments." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000241> doi:10.1057/9780230226203.1742

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tsuru, Shigeto (1912 – 2006)

Tsuneo Nakauchi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

business cycles; Japan, economics in; Marxian methodology; Tsuru, S.

Article

Tsuru was born on 6 March 1912 in Oita and brought up in Nagoya. Political disapproval of his involvement in a socialist study group led to his leaving Japan for the United States in 1913.
After receiving a Ph.D. at Harvard in 1940, he taught there briefly as a lecturer. During that period, he married Masako Wada, the niece of Marquis Koichi Kido. They left the United States for Japan in 1942 on board a ship exchanging citizens during wartime.
During the reconstruction of Japan after the Second World War, he served first in the Ministry of Foreign Affairs and later as the vice minister of the Economic Stabilization Board, where he took part in the preparation of the first issue of the Board's Economic White Paper.
In 1948, he was appointed Professor of Economics at Hitotsubashi University, where he later served for nine years as the director of the Institute of Economic Research and for three years as president of the university from 1972 until retirement. After retirement, he served as an adviser to the Asahi newspaper. He then assumed a professorship at Meijigakuin University.
Tsuru's analytical works in economics are based on his wide background in the area of both Marxian and modern economics. His principal studies have been incorporated in the *Collected Works of Shigeto Tsuru* (1976). These constitute 13 volumes, and the last one, in English, is entitled *Towards a New Political Economy*.
The main areas of the author's interests which developed from his Harvard days encompassed Marxian methodology, business cycle theories and their application to Japan's economic development. A continuing emphasis in the studies was on aspects of the development of capitalism in Japan. His book *Has Capitalism Changed*? (1961) reflects this particular interest.
Tsuru was one of the first economists in Japan to have drawn the attention of the general public to environmental problems by applying his unusual skills in putting academic concepts into the language of ordinary people.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Selected works

1961. *Has Capitalism Changed*? Tokyo: Iwanami Shoten Publishers.

1976. *Collected Works of Shigeto Tsuru*, 13 vols. Tokyo: Kodansha. Vol. 13, *Towards a New Political Economy*, is in English.

How to cite this article

Nakauchi, Tsuneo. "Tsuru, Shigeto (1912–2006)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000132>
doi:10.1057/9780230226203.1743

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tucker, Josiah (1713 – 1799)

G. Shelton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

free trade; Hume, D.; self-interest; Tucker, J.

Article

Born in Laugharne, Carmarthenshire, Tucker was Dean of Gloucester from 1758 until his death, and was also a rector in Bristol for over 50 years. Although his career as an ecclesiastic was a long and honourable one, he was best known in his own day for his active part in many contemporary controversies. Whether the subject was the naturalization of foreign Protestants and Jews, the undesirable effect of low-priced liquors, or the cruel custom of cock-throwing on Shrove Tuesday, his pen was always ready. He was responsible for the earliest study of the Methodist movement and the first substantial critique of Locke's political philosophy. The themes which recurred most often were his opposition to monopolies and his hatred of war. His interest in political affairs was not confined to the press: he participated in several Bristol elections as the local Whig agent.

Tucker's period of greatest notoriety came during the American Revolution. In a steady stream of publications he rejected both the conciliation policy of Burke and that of war. Although he had no sympathy for the ideas espoused by the more radical Americans and their supporters in Britain, he saw no economic reason for attempting to retain the colonies by force, since he was convinced that they would willingly trade with her as long as it was in their interest to do so.

In his *Essay on Trade* (1749) Tucker recognized the need for a scientific study of what is now called economics but only the first part of what to be his ‘great work’ on the subject was ever printed, and then only for circulation among friends. However, his other works, which contained the bulk of his ideas, were known to Quesnay and Turgot (who translated one of them) well before the Physiocrats’ first writings appeared, and several of his books were to be found in Adam Smith's library. These ideas included: self-love as a socially useful drive, labour as the true source of wealth, and the importance of machinery as a means of increasing that wealth. His aim was to encourage high productivity which would lead, in turn, to lower prices, increased demand, and more jobs. Anything which obstructed the free circulation of labour and capital, especially regulations supporting vested interests, should be

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
eliminated. On the other hand, Tucker did not expect that self-interest and the public good would always coincide; some legislation was necessary to encourage that happy outcome by making what was socially desirable also profitable.

Tucker's most significant contribution may have been his argument against the notion put forward by David Hume that rich countries were likely over time to lose their wealth to poorer ones. Tucker eventually convinced Hume that the factors which made a nation rich in the first place tended to give it a practically insurmountable advantage over its less wealthy neighbours. Since Britain enjoyed such an advantage, once Tucker's reasoning was accepted, as it was by Pitt in the 1780s, opposition to free trade could be disarmed and the way cleared for its triumph in the 19th century.

Selected works

1749. *A Brief Essay on the Advantages and Disadvantages which Respectively Attend France and Great Britain with Regard to Trade* [*The Essay on Trade*].

1751–2. *Reflections on the Expediency of a Law for the Naturalisation of Foreign Protestants*. Part I (1751), Part II (1752).

1753. *Letters to a Friend Concerning Naturalisation*.

1755. *The Elements of Commerce and Theory of Taxes*. In Schuyler (1931).

1757. *Instructions for Travellers*. In Schuyler (1931).

1774. *Four Tracts Together With Two Sermons on Political and Commercial Subjects*. Gloucester and London: Raikes & Rivington.

1775. *A Letter to Edmund Burke*.

1781. *A Treatise Concerning Civil Government*. London: T. Cadell.

Bibliography

Clark, W.E. 1903. *Josiah Tucker, Economist*. New York: Columbia University Press.

Schuyler, R.L., ed. 1931. *Josiah Tucker: A Selection from his Economic and Political Writings*. New York: Columbia University Press.

Semmel, B. 1965. The Hume-Tucker debate and Pitt's trade proposals. *Economic Journal* 75, 759–90.

Shelton, G. 1981. *Dean Tucker and Eighteenth-Century Economic and Political Thought*. London:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Macmillan.

How to cite this article

Shelton, G. "Tucker, Josiah (1713–1799)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000134>
doi:10.1057/9780230226203.1744

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tugan-Baranovsky, Mikhail Ivanovich (1865 – 1919)

Alec Nove
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

exploitation; labour theory of value; Marxist political economy; serfdom; subjective theory of value; surplus value; Tugan-Baranovsky, M. I.; Underconsumptionism

Article

Of mixed Ukrainian-Tartar origin, Tugan-Baranovsky was born in the Kharkov province, and graduated from Kharkov university in 1888. His *Magister* dissertation for Moscow University was on industrial cycles in Great Britain, and he spent six months of his research time in London in 1892. There could scarcely have been a more masterly master's thesis. It was published in 1894. While criticizing crude underconsumptionist theories, and pointing out that ‘the process of production creates its own market’, especially for producers’ goods, he went on to stress that the simple model derived from J.-B. Say assumes that ‘that entrepreneur, before beginning production, has a wholly correct and accurate knowledge of the requirements of the market and of the output of every branch of industry’. He cited Moffat's phrase ‘the continuous struggle between the requirements of unknown demand and the fluctuations of unknown supply’. He contrasted the ‘propensity to save’ with the output of capital goods of various types, and with the opportunities to invest, which can and do get out of line with one another. He collected much empirical data. In the words of Alvin Hansen, ‘he began a new way of thinking about the problem’ of business cycles.

His doctoral dissertation was another masterpiece, full of original research, *The Russian Factory, Past and Present* (*Russkaya fabrika v proshlom i nastoyashchem*) (1898), which has appeared in English translation. This was a major contribution to economic history. In vivid and well-written pages, Tugan-Baranovsky shows the great importance of the state and of serfdom, and the subsequent growth of market-orientated industries based on free labour (though some workers were serfs on quit-rent, a few of whom became serf millionaires). He also made stimulating observations concerning ‘natural’ and ‘artificial’ industrialization, relevant to today's concerns with economic development.

His major contribution to economic theory was *Osnovy politicheskoi ekonomii* (1917), which went through many editions, and represented an attempt at a synthesis between Marxist political economy (the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
labour theory of value) and subjective value theory. He considered that the marginalists ignored ‘the objective conditions of production’, while Marxists failed to recognize that not only objective factors but also subjective valuations were an integral part of a theory of value. He argued that Marx confused value (*Wert*) with cost (*Kosten*). He basically supported Marx's theory of exploitation, but defined ‘surplus value’ as equal to the value of the *products* acquired (consumed) by the capitalists, which earned him criticism from Kondratiev and Struve. He retained from his early Marxism the belief that economists should regard man as not just another factor of production. If horses could write economics, there would be a horse theory of value.

Tugan-Baranovsky to the end of his life retained a particular interest in agriculture and in (voluntary) cooperation. One of his last articles drew attention, prophetically, to the effect of the egalitarian land redistribution of 1917–18 on the marketing of foodstuffs.

His academic career was mainly in the University of St Petersburg, though he was dismissed in 1899 for ‘political unreliability’ and only reinstated in 1905, as a *privatdozent*. His election to the chair of political economy in 1913 was vetoed by the Minister of Education. Re-elected in 1917, he did not take up his appointment, but returned to his native Ukraine. He became Academician, dean of the Faculty of Law of Kiev, chairman of the Ukrainian cooperatives, president of the Ukrainian economic association, and for a short period Minister of Finance, amid turmoil and civil war. He died in 1919, on his way to Odessa to board a ship for France. He must be seen as the most original of the Russian economists of his generation. Alas, in the Soviet Union he was known chiefly as a ‘legal-marxist’ opponent of Lenin, and few had the opportunity to study his works, though *The Russian Factory* was reprinted.

Selected works

1894. *Promyshlennye Krizisy v sovremennoi Anglii* [Industrial crises in contemporary Britain]. St Petersburg. 2nd Russian edn. trans. into French by J. Schapiro as *Les crises industrielles en Angleterre*, Paris: M. Giard & E. Briere, 1913.

1898. *Russkaia fabrika v proshlom i nastoiashchem* [The Russian factory, past and present]. St Petersburg. 3rd Russian edn. trans. by A. Levin and C.S. Levin, under the supervision of G. Grossman, as *The Russian Factory*, Homewood, IL: R.D. Irwin, for the American Economic Association, 1970.

1905. *Teoreticheskie osnovy marksizma* [The theoretical foundations of Marxism]. St Petersburg. Trans. into German as *Theoretische Grundlagen der Marxismus*, Leipzig: Duncker & Humblot, 1905.

1906. *Souremennyi sotsializm v svoem istoricheskom razuitii*. Trans. M.I. Redmount as *Modern Socialism in its Historical Development*, London: S. Sonnenschein & Co., 1910. Reprinted New York: Russell & Russell, 1966.

1914a. *Ekonomicheskaia priroda kooperativov i ikh klassifikatsiia* [The economic nature of cooperatives and their classification]. Moscow.

1914b. *Ocherki iz noveishei istorii proliticheskoi ekonomii i sotsializma* [Outlines of the recent history

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of political economy and of socialism]. St. Petersburg.

1917. *Osnovy politicheskoi ekonomii* [Foundations of political economy]. Petrograd.

1918. *Sotsializm kak polozhitelnoe uchenie* [Socialism as a positive subject]. Petrograd.

How to cite this article

Nove, Alec. "Tugan-Baranovsky, Mikhail Ivanovich (1865–1919)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000135> doi:10.1057/9780230226203.1745

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

tulipmania

Peter Garber
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This curious speculation in tulip bulbs near four centuries ago has become a modern synonym for and warning about the irrational speculation that may break out in any asset market. Nevertheless, a look at the forces that actually drove it provides an alternative, fundamental explanation and a warning to observers of asset markets not to accept so quickly unprovable psychological explanations of asset prices.

Keywords

asset pricing; bubbles; East India Company; forward contracts; multiple equilibria; Mississippi Bubble; South Sea Bubble; speculative bubbles; tulipmania

Article

The Netherlands of 1634–7 was the scene of a curious speculation in tulip bulbs that has come to be known as the Dutch tulipmania. Single bulbs of rare and prized varieties such as Semper Augustus or Viceroy became worth a middle-sized fortune. In its most extreme final phase in January-February 1637, prices of even common varieties such as Switsers or Witte Kroone soared twentyfold within a month and then crashed back to their original values. That these were prices of easily reproducible horticultural products has added to the bemusement of generations of historians and economists.
In the succeeding 370 years, the *historical* tulipmania became, in itself, an obscure footnote to the *conceptual* tulipmania of economics and finance, a word warning of the obvious, delusional speculative excess that human behaviour in financial markets can create (see, for example, Kindleberger, 1996). It is interchangeable with words like ‘bubble’ or ‘mania’, which also arose from historically distant events such as the Mississippi or South Sea Bubbles or the more recent ‘irrational exuberance’. These words have been used by economic theorists to emphasize an historical basis for the salience of unstable multiple equilibria in forward-looking financial and macroeconomic theories. They have also been used

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to justify ignoring financial market outcomes that contradict favoured asset pricing theories by means of an arbitrary invocation of the existence of a bubble.

The traditional image of tulipmania

Modern references to the tulipmania usually depend on the brief description in Charles Mackay's *Extraordinary Popular Delusions and the Madness of Crowds* (1852). The tulip originated in Turkey and spread into western Europe in the mid-16th century. The tulip was immediately accepted by the wealthy as a beautiful and rare flower, appropriate for the most stylish gardens. The market was for durable bulbs, not flowers. The Dutch dominated the market for tulips, initiating the development of methods to create new flower varieties. The bulbs that commanded high prices produced unique, beautifully patterned flowers; common tulips were sold at much lower prices.

Beginning in 1634, non-professionals entered the tulip trade in large numbers. According to Mackay, individual bulb prices reached astronomic levels. For example, a single Semper Augustus bulb was sold at the height of the speculation for 5,500 guilders, a weight of gold equal to $66,000 evaluated at $600/oz. Mackay provided neither the sources of these bulb prices nor the dates on which they were observed, however.

Finally, and unexplained by Mackay, the frenzy suddenly terminated. According to Mackay, even rare bulbs could find no buyers at ten per cent of their previous prices, creating long-term economic distress. Mackay presented no evidence of immediate post-collapse transaction prices of the rare bulbs. Instead, he cited prices from bulb sales of 60 years, 130 years, or 200 years later as indicators of the magnitude of the collapse and of the obvious misalignment of prices at the peak of the speculation. Moreover, Mackay provided no evidence of the general economic context from which the speculation emerged.

The fundamentals of the tulipmania

Unfortunately, the fundamentals of markets in rare bulbs present a much more prosaic picture. The bulk of the speculation concerned highly prized tulips that were infected with mosaic virus. Mosaic virus had the effect of producing unique feathery patterns in the flower that could be reproduced only through propagation by budding, not by seeds. Hence, the rate of reproduction was much more limited that one might expect. Such bulbs were traded primarily among professionals. Their prices were supported by a strong demand by flower fanciers, not only in the rapidly growing Netherlands of the golden age but also by the wealthy nobility and merchants of surrounding countries. During the period of the tulipmania, 1634–7, the already high prices of such bulbs doubled or tripled. Over the course of decades or centuries, prices for these varieties converged to the low cost of reproduction, and this has been taken as evidence of folly.

However, an examination of the pricing of prized flower varieties throughout history reveals a similar pattern: prices of the prototype are very high, perhaps even representing a medium fortune. Then, as they are reproduced through succeeding generations, they become common. Just as for the value of a prized racehorse put out to stud, the high initial price represents the present discounted value of the valuation above cost of successive, expanding generations, wherein the value of any individual exemplar is bound to fall.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The more frenzied phase of the tulipmania described by Mackay took place from mid-1636 to February 1637, but especially in January, 1637. At this time trading, especially among the non-professionals, took place in newly organized ‘colleges’, which were located in taverns. The trading was not for actual bulbs but for contracts for forward delivery. Since bulbs had to remain in the ground through the winter, none were actually delivered on these contracts before the speculation ended. Contracts were not marked to market, and margin was not posted. A small, fixed amount of ‘wine money’ had to be delivered by the buyer, which provides the flavour of, if not the fuel for, what was happening during the frenzied trading in these taverns.

When this part of the speculation collapsed in February 1637, some city governments proposed winding up outstanding contracts with a ten per cent payment on contracted amounts if the buyer refused to accept delivery. Perhaps this is where Mackay got the notion that bulbs could not be sold at ten per cent of their previous value, even though a buyer might refuse the deal if prices had fallen only to 90 per cent of the contracted amount. There were very few takers even on this offer, but short sales were in any case unenforceable contracts under Dutch law.

When one looks at notarized contracts for actual bulbs, however, the picture is quite different. Some rare bulbs that were auctioned for high prices at the very peak of the speculation in February 1637 still sold for high, albeit much lower, prices in 1642. For example, an Admirael Liefkens bulb was sold for 1,345 guilders at the peak and for 220 guilders in 1642, an annual percentage decline in value of 36 per cent. This rate of decline is comparable to the typical pattern of price behaviour for valued varieties in successive historical periods and does not indicate anything unusual in the mania.

It is the rare bulb price behaviour during the tulipmania that has been emphasized historically. But at the very end common bulbs sold in bulk shot up twentyfold and soon collapsed back to one-twentieth of the peak. It is this usually ignored bit of the episode that remains a puzzle.

An historical background

The tulip market was introduced into the Netherlands during the Eighty Years’ War of independence between the Dutch and the Spanish, and the tulipmania occurred in the middle of the Thirty Years’ War as the two conflicts merged. The Spanish were thwarted in their attempts to subjugate the Netherlands, which consolidated its territory and eventually seized control of most of international shipping. The Thirty Years’ War of 1618–48 was particularly destructive of the populations and economies of central Europe, with many principalities in the Holy Roman Empire losing one-third of their populations.

In every year of the war, the Dutch fielded large armies and supported large fleets, though the population of the Netherlands was no more than 1.5 million. The Dutch provided much of the strategic planning and finance for the Protestant effort, along with France, negotiating and financing the successive interventions of Denmark and Sweden on the Protestant side in the 1620s and 1630s.

From 1620 to 1645, the Dutch established near-monopolies on European trade with the East Indies and Japan, conquered most of Brazil, took possession of the Dutch Caribbean islands, and founded New York. In 1635 the Dutch formed a military alliance with Richelieu's France, which eventually placed the Spanish Netherlands in a precarious position. In 1639 the Dutch completely destroyed a second Spanish Armada of a size comparable to that of 1588. As a result of the war, Spain ceased to be the dominant power in Europe, and the Netherlands, though small in population and resources, temporarily became a major power centre because of its complete control over international trade and international finance.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The Dutch were to 17th-century trade and finance as the British were to 19th-century trade and finance. Sophisticated finance mechanisms evolved with the establishment of its trade and finance dominance. Amsterdam became the leading market for short- and long-term credit; and markets in stocks, commodity futures, and options materialized early in the 17th century. Trading of national loans of many countries centred on Amsterdam, as did a market in the shares of joint stock companies. The East India Company, founded in 1602, gradually gained control over east Asian trade and consistently paid out large dividends. Interest rates on Dutch markets were remarkably low for the times; for example, the East India Company paid no more than five per cent on advances during the 17th century.

There were some dark periods during this golden age, and it should be carefully noted that these occurred during the years of the tulipmania. From 1635 to 1637, bubonic plague ravaged the Netherlands. In July 1634 the Holy Roman Empire completely defeated Swedish forces in the Battle of Nordlingen, forcing a treaty on the German Protestant principalities in the May 1635 Peace of Prague and releasing Spanish resources for the war against the Dutch. Along with the growing war weariness in the Netherlands, these events forced France to enter the Thirty Years' War militarily with the Dutch alliance in 1635. Initially unprepared, the French suffered major setbacks, culminating in an imperial invasion of northern France in August 1636.

How should we interpret the tulipmania?

The tulipmania is an obscure event from distant history that provides a cornucopia of concepts even now. One can take one's pick from the following views, depending on personal taste:

- It was an outburst of speculative fever that serves to the present day as a warning of the dangers of market speculation.
- It was a curious event limited to the Dutch winter of 1636–7 in the middle of an outbreak of bubonic plague and at the time of the greatest success of the Catholic armies of the Empire in the Thirty Years War against the Protestants.
- It was a drinking game in which people without wealth made the equivalent of million euro bets with each other, with no intention or possibility of paying.
- It was a swing of fashion in the most wealthy society of the era, which caused the most exquisite of tulips to have a higher price than Rembrandt's *Night Watch*.
- It was a reasonable and well-calculated investment that still causes the most wonderful outburst of colour every Dutch springtime.

See Also

- speculative bubbles

Bibliography

Cooper, P. 1970. *New Cambridge Modern History, Volume 4: The Decline of Spain and the Thirty Years' War*. Cambridge: Cambridge University Press, 1970.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Garber, P. 2000. *First Bubbles: The Fundamentals of Early Manias*. Cambridge, MA: MIT Press.

Kindleberger, C. 1996. *Manias, Panics and Crashes: A History of Financial Crises*, 3rd edn. New York: Wiley.

Mackay, C. 1852. *Extraordinary Popular Delusions and the Madness of Crowds*, vol. 2, 2nd edn. London: Office of the National Illustrated Library.

Posthumus, N. 1929. The tulip mania in Holland in the years 1636 and 1637. *Journal of Economic and Business History* 1, 434–66.

Schama, S. 1987. *The Embarrassment of Riches*. New York: Alfred Knopf.

How to cite this article

Garber, Peter. "tulipmania." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000137> doi:10.1057/9780230226203.1746

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Turgot, Anne Robert Jacques, Baron de L'Aulne (1727 – 1781)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

advances; Böhm-Bawerk, E. von; capital accumulation; capital and interest theory; class; Condorcet, Marquis de; division of labour; Du Pont de Nemours, P. S.; Gournay, Marquis de; inequality; interest; laissez-faire; law of variable proportions; loanable funds; marginal revolution; market price; natural price; Physiocracy; Quesnay, F.; saving equals investment; sharecropping; single tax; slavery; Smith, A.; stages theory of progress; surplus; thrift; Turgot, A. R. J.

Article

Economist, philosopher and administrator, Turgot was born in Paris, in 1727, the third son of a well-established Norman family with a long tradition of public service in the magistrature. Destined originally for a career in the Church, his education was extensive. Because of shyness, his education commenced at home with a private tutor; it continued at the Collèges Duplessis and Bourgogne where, among other things, he studied the philosophical systems of Newton and Locke. In October 1746 he entered the Seminary of Saint-Sulpice in preparation for the priesthood. From June 1749 to early 1751 he was resident student at the Maison de Sorbonne, an annex of the theological faculty of the University of Paris. His already considerable academic distinction led to his election to the office of prior in 1750. This honorary position inspired two of his earliest works, of which the second, *Philosophical Review of the Successive Advances of the Human Mind* (Turgot, 1750a) contained a demonstration of the importance of economic surplus for the development of civilization as part of his four stages theory of human progress:

> Tillage … is able to feed more men than are employed in it … Hence towns, trade, the useful arts and accomplishments, the division of occupations, the differences in education, and the increased inequality in the conditions of life. Hence leisure … (and) the cultivation of the arts. (Turgot, 1750a, p. 43)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
His father's death in early 1751 possibly saved Turgot from having to take his final vows, since his inheritance provided sufficient income to commence the administrative career he desired. He gained appointment to some judicial positions, including that of Master of Requests in early 1753, the stepping stone to a career as provincial intendant. During the 1750s, Turgot's prolonged residence in Paris allowed immense intellectual activity but left time for extensive travels through France when accompanying Gournay on his official tours of inspection of French industry. His contributions to the *Encyclopédie* (Turgot, 1756; 1757a; 1757b), ranging from articles on Etymology, Existence and Expansibility to Fairs and Foundations, spread his fame as a philosopher and gained him the friendship of Voltaire, whom he visited in 1760 during his only journey abroad.

Although Turgot's interest in economics had commenced as early as 1749 when he wrote a critique of Law's system of paper currency, and such interest was maintained in the Sorbonne orations and other early writings, it was considerably stimulated by Gournay's friendship. Gournay's death inspired Turgot's famous eulogy (Turgot, 1759) and earlier he had encouraged Turgot's translation of Tucker (Turgot, 1755), Turgot's comments on Gournay's notes to the translations of Child (Turgot, 1753–4), and most probably, aspects of the content of Turgot's two economic articles for the *Encylopédie* (1757a; 1757b). Gournay's friendship was particularly important because it brought Turgot's economics under more substantial English influence as compared with Physiocracy (Groenewegen, 1977, p. xiv). Turgot's first meeting with Quesnay cannot be precisely dated. It may not have occurred till 1756 or 1757 when their mutual association with the *Encyclopédie* may have brought them together. Turgot (1759, p. 26) cites Quesnay's contributions with considerable approval, indicating that his generally good relations with the Physiocrats must by then have been well established. His presence at Quesnay's meeting in the Entresol at Versailles as a ‘handsome young Master of Requests’ was in any case recorded by Mme de Hausset (n. d., pp. 117–19). A lifelong friendship with Du Pont de Nemours, which began in 1763, must have strengthened good relations with the Physiocrats even further. Another enduring friendship was made with the philosopher and mathematician, Condorcet. Both friends produced memoirs of Turgot's life after his death (Du Pont, 1782; Condorcet, 1786).

In 1761 Turgot was appointed Intendant of Limoges, a large district containing most of the provinces of Limousin and Angoumois, and this position he filled with distinction for 13 years. The task of the 18th-century intendant, a post compared by Morley (1886, p. 112) to that of Chief Commissioner in a large district of the former British Empire in India, were many:

> He had to collect direct taxation, rectify justice, promote the arts of agriculture, encourage industry and commerce … Everything came within the scope – sanitation and public order, morality and poor relief, the recruiting and billeting of soldiers, military equipment, rations and transport, religious processions and the pairs of churches, colleges and libraries, parochial and municipal finance. (Dakin, 1939, pp. 27–8, the standard source for details of Turgot's administrative career as intendant)

Despite this cumbersome and heavy administrative load, Turgot managed to introduce some reforms. These included changes to the assessment and collection of the taille, transmutation of the *corvée* and the *milice* into money payments, and establishing public workshops to alleviate hardships suffered by the population of his province during the long and severe famine of 1769 to 1772. Many of his better

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
known economic writings date from this period: first of all, his *Reflections on the Production and Distribution of Wealth* (Turgot, 1766a, but not published till 1769–70 in serial form in the *Ephémérides*), a draft for a paper on value and money (Turgot, 1769), observations on two winning entries in a prize competition he had organized on the subject of taxation (Turgot, 1767a; 1767b), and a series of memoranda connected with the ministration of his province in which he pleaded with the central authorities for reforms on the basis of carefully elucidated theoretical principles. The more important of these deal with taxation in general (Turgot, 1763), mines and quarries (Turgot, 1767a), the grain trade (Turgot, 1770a), the rate of interest (Turgot, 1770b) and the trade mark on iron products (Turgot, 1773). His administrative work permitted regular but infrequent visits to Paris to see friends and attend the salons of Mme de Graffigny, Mme de Geoffrin and later, Mlle de Lespinasse. Apart from the French intelligentsia, he there became acquainted with foreign notables like Hume, Adam Smith, Franklin and Gibbon. The exile imposed by his administrative position also inspired a substantial correspondence with Du Pont de Nemours, Condorcet, and his personal secretary, Caillard, making up a large part of the five volume edition of his works as edited by Schelle (1913–23). Schumpeter (1954, p. 248) notes a further significant aspect about Turgot's administrative career: ‘nearly all his creative work must have been done between 18 and 34 because during the 13 years at Limoges, Turgot can have had but scanty leisure, during his nearly 2 years of ministerial office, practically none.’

Louis XVI's succession to the throne in 1774 marks the next stage in Turgot's career; his membership of the Royal Council, first as Minister of the Navy (from 20 July 1774), then as Minister of Finance (from 24 August 1774 to 12 May 1776, the date of his dismissal). While lamenting the fact that so much more could have been done, Du Pont de Nemours (1782) summarized Turgot's career as minister in terms of the reforms accomplished. These included restoration of the domestic free trade in grain, abolition of many small, local duties and other constraints on trade, and the January 1776 measures, of which partial suppression of the guilds and replacing the *corvée* with a more general land tax were the more controversial measures. These last, now generally known as the six Edicts, ultimately caused his downfall even though he did secure royal support for their forcible registration at a famous *Lit de Justice*. As an ‘experience in economic politics, an exception to the general rule that French ministers of finance are financiers rather than economists’, Faure (1961) gives a detailed account of Turgot's ministerial experience and the opposition it encountered almost from the start during the grain riots of early 1775 and the intrigues surrounding the campaign to prevent registration of his 1776 Edicts. The reforms Turgot had accomplished were reversed within six months from his downfall, and ‘leisure and complete freedom as the principal net product from my two years in the ministry’ was how he himself sarcastically summed up his achievements in a letter to Caillard (Schelle, 1913–23, vol. 5, p. 488). The period of retirement in the five years which remained of his life were not years of inactivity.

> The sciences which he had formerly cultivated, easily filled up his time; he studied mathematics, he sought to bring the thermometer to greater precision, he searched with l'Abbé Rochon, after various expeditious convenient, and cheap methods of multiplying copies of writing to supply the place of printing, … he preserved all his passion for literature and poetry … (Condorcet, 1786, pp. 255–62)

In 1778 he was elected President of the Académie des Inscriptions et des Belles Lettres. He died in Paris

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in March 1781 from gout, a family illness that had steadily wrecked his health, worsening particularly during the last decade of life.

Although Turgot is now largely remembered as a very important 18th-century French economist and a pre-revolutionary reformist finance minister, such an assessment fails to reflect his youthful ambitions and work. Meek (1973, pp. 1–2) indicates that

> Turgot set out from the beginning with the conscious intention of becoming a polymath rather than a specialist …. A list of works to be written … begins with 'The Barcimedes', a tragedy, and ends with 'On Luxury, Political Reflections', and in between these are forty-eight others, including works on universal history, the origin of languages, love and marriage, political geography, natural theology, morality and economics, as well as numerous translations from foreign languages, literary works, and treatises on scientific subjects … What is (especially) remarkable is … that Turgot managed, during his short life of only fifty-four years, to make some contribution to so many of them, or at least to retain an active and intelligent interest in them.

Turgot was therefore considerably more than an economist, and some of the qualities Meek listed needed to be highlighted to underscore that fact. In the first place, he was a superb linguist, 'reading seven languages – Greek, Latin, Hebrew, Spanish, German, Italian, and English, the last three of which he spoke fluently' (Dakin, 1939, pp. 10–11) and from some of which he published poetry translations (Turgot, 1760; 1762; 1778). This linguistic skill is reflected in his magnificent library, the catalogue of which (Tsuda, 1974) demonstrates his ability to profit from the economics writings of other countries. Secondly, his wider interests influence the interpretation to be given to his economics. Turgot's contributions cannot be simply assessed in terms of his importance in fashioning certain parts of the marginalists' toolbox, as done, for example, by Schumpeter (1954). He is far more correctly depicted as 'an author of transition between the Physiocrats at the end of the eighteenth century and the English classical economists at the start of the nineteenth' (Bordes, 1981, p. xvi), that is, the true contemporary of those like Smith, Steuart, Condillac, Verri and Beccaria producing economic treatises building on Physiocracy in that quarter century ending in 1776 during which political economy emerged as the science of the reproduction, circulation and distribution of wealth (see Groenewegen, 1983a). Although, apart from the skeleton form of his *Reflections* (Turgot, 1766a), Turgot never completed such a treatise, this skeleton combined with his youthful views on social progress allows his economics to be depicted as something essential to the understanding of historical stages (see Finzi, 1981). The reduced emphasis on the economics developed by him as a by-product of his administrative career this implies prevents his depiction as a 19th-century liberal (see Morley, 1886; Bourrinet, 1965) or as a general precursor of equilibrium theory (Nogaro, 1944, pp. 26–7; Bordes, 1981, pp. xxvi–xxviii).

Schelle (1913–23, vol. 1, pp. 29–30, 79) draws attention to the fact that the young Turgot was interested above all in sociology, shown by his attempts at analysing causes of progress and decay in taste, science and the arts, and that this analytical interest was enhanced by studying the formation of languages, because etymology provides valuable clues not only to the progression of ideas but to the needs from which ideas originate. Turgot's early work on language formation and social progress appears to have suggested the importance of economic factors in explaining this process, and that the means by which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
peoples gain their subsistence, determining as it does their access to economic surplus, is particularly important to explain the manner in which societies, morals, laws, the arts and the sciences gradually develop. Turgot (1750b, p. 172) explicitly relates certain characteristics in the formation of languages to stages of hunters, shepherds and husbandmen with their different requirements for communication. The notion of progress between these stages is implied in his critique (Turgot, 1751a, pp. 242–3) of the alleged virtues of equality in the savage state where he shows that, by preventing the division of labour and the accumulation of capital on which abundance and secure subsistence depend, such equalities also prevent progress in the sciences and arts. The subsequent fragment *On Universal History* (Turgot, 1751b) combines these elements into ‘a quite advanced statement of the four stages theory – or at any rate a three stages theory, with a distinct hint of the fourth stage’ (Meek, 1977, p. 22).

Although Turgot's *On Universal History* was not completed, its basic notions stayed with him for the rest of his life and in a number of aspects received further development. The systematic attempt to explain general progress by stages from hunters, shepherds, farmers to a commercial society to a large extent provides the basis on which Turgot constructed his analysis of the production and distribution of wealth in the *Reflections*, as is explicitly recognized in his discussion of cattle as a form of moveable wealth (Turgot, 1766a, pp. 66–7). Moreover, the whole of the *Reflections* is imbued with Turgot's sociological concerns with nature of progress and historical development, thereby reinforcing the need to interpret its contents in terms of stadial development. Such a view is also appropriate for its alleged original purpose as providing explanations to accompany an extensive questionnaire on the Chinese economy and society which he had prepared for two young Chinese students in 1766 (see Groenewegen, 1977, pp. xvii–xix). This aspect of the *Reflections* may be demonstrated from a summary of its contents, a process facilitated by the parts into which Du Pont divided it for publication in *Ephémérides*.

Under this subdivision (Turgot, 1766a, pp. 43–56); the first part of the *Reflections* analyses the basic features of the production and distribution of wealth within an agricultural society. Although capital advances are used in such a society, the distributional aspect of such use is ignored at this stage except for its final sections dealing with what for Turgot were contemporary manifestations of agricultural production (1766a, pp. 56–6). Turgot argues at the outset that such an agricultural society presumes a division of labour, a natural consequence of its inequality of property ownership. Hence it presumes a specific set of class relations, that is, division of society between a proprietors’ class owning land and living from its surplus produce without a need to work, and working classes without property earning their living from their labour. Within this working class, the division of labour divides those cultivating the soil to produce food and raw material or products of prime necessity from artisans who transform those primary materials into forms more suitable for people's use. Because artisans depend on those working in agriculture for their livelihood, Turgot calls them a stipendiary class. Because they only transform existing wealth without generating a surplus, Turgot calls them a sterile class to contrast their work with that in agriculture which produces such a surplus and thereby generates new wealth. In this way Turgot demonstrates the appropriateness of Physiocratic class analysis for understanding agricultural society.

At the end of this discussion Turgot suggests that the relationship between proprietors and the working classes in agriculture is itself subject to change with respect to the manner in which proprietors draw the surplus from the land through the organization of production. The method springing first to mind, landlords hiring wage labour for themselves, Turgot views as an unlikely candidate to be first from the perspective of actual historical development. Slavery appears to have been first in this regard. Though

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Turgot saw slavery persisting in colonial societies, elsewhere economic circumstances, combined with humanity and landlord's convenience gradually transformed slavery first into bondage of the soil and then vassalage, where former serfs become tenants and surplus product rent and other stipulated dues. One such tenancy, the dominant form in Turgot's France, was sharecropping or *métayage* in which the landlord made the advances in return for a fixed part of the produce; another, more advanced form existed where a capitalist farmer, or entrepreneur (Turgot, 1766b, pp. 28–9) rented the land from the proprietor for a specified rent and period of time, himself providing the necessary advances for cultivation. Capitalist farming or *la grande culture* had begun to emerge in France during Turgot's time, and the farmer/entrepreneur class it created ‘has a quite distinct status from that of the ploughman/sharecropper. He does not earn his living by the sweat of his brow like labourers but by employing his capital in a lucrative manner like the shipowners of Nantes and Bordeaux employ theirs in maritime trade’ (Turgot, 1766b, p. 29). As nations become more wealthy, and capital accumulates, a new proprietors’ class is created who live without working from the revenue of money or capital. The section which opens the second part of the *Reflections* draws attention to this feature by dealing with ‘capitals in general and the revenue of money’.

Explanations of the origin and use of capital, and its impact on ‘the system of distribution of wealth which I have just outlined’ (Turgot, 1766a, p. 56) requires an elementary acquaintance with the theory of money, commerce, exchange and value and hence some retracing of steps. After this digression, which contains little that is new, Turgot presents a fascinating analysis on both the uses of capital and its origins through accumulation and thrift. Turgot discusses accumulation and thrift both historically and analytically. Historically, accumulation is associated with slavery and surplus product from land: analytically, prudence and a desire for self-improvement are seen as major motives for thrift. Turgot argues that the savings process is greatly facilitated by the introduction of money but that this raises new complications such as a need to distinguish saving, hoarding and investment. Turgot's saving–investment analysis denied the possibility that money savings were able to induce substantial leakages from the circular flow because hoarding was seen as irrational and money had only a limited role as a store of value. Turgot argued that savings were immediately transformed into investment (see Groenewegen, 1971).

Turgot's analysis of the productive use of capital and its social implications is presented in the second and third parts of the *Reflections*. These reveal the degree to which his economics had departed from Physiocracy and anticipated views subsequently developed by Adam Smith. First of all, Turgot's exposition extends the use of capital to all sectors of industry thereby not confining it to agriculture as Quesnay had done. Secondly, Turgot, like Smith, links an increasing need for capital in production with extensions of the division of labour and a consequent lengthening of the time period of production. Thirdly, Turgot associates the provision of capital to industry with a new class of society, the capitalist/entrepreneur as owners of moveable wealth, who invest these resources to reap a return. Hence the working classes of agriculture, manufacturing and trade ‘may be divided into two orders of men, that of the Entrepreneurs of Capitalists, who make all the advances, and that of the mere wage earning workmen’ (Turgot, 1766a, pp. 72–3). Of special significance for analysing distribution, this new class appropriates the resources by which it can live without labour through the creation of interest and profit as a new income type. Profit in this context, is clearly associated by Turgot with a return on productive investment, comprising an interest component, a premium for risk and remuneration for the time and trouble of the entrepreneur in supervising the investment. Part of the *Reflections* therefore suggests that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the quantitative changes of gradual capital accumulation (perhaps first experienced within agriculture) by a qualitative leap create a new stage of society, the commercial or capitalist stage (see Meek, 1973, pp. 21–6).

However, this view is partially contradicted in some of the *Reflections*' later sections. These reveal the new class as mere lenders of money and show Turgot equivocating on whether interest and profit have the same disposable status as the net product of land. Such aspects of Turgot's work show that for him 'commercial society' perhaps remained 'incorporated into the agricultural state [never becoming] a separate stage, characterised and led by an internal logic of its own' (Finzi, 1982, p. 116), and reinforce the position that Turgot's analytical schema is a transition from the Physiocrats to subsequent classical political economy retaining that ambiguous relationship between capital and land, rent and interest, not really resolved analytically until Ricardo's distribution analysis (see Cartelier, 1981).

Despite this ambivalence in depicting the final stage of social progress, these sections of the *Reflections* also contain some of Turgot's most analytically significant contributions to economics. Having shown that 'capitals are the indispensable foundation of all lucrative enterprises' and that the continual reproduction of these capitals 'with a steady profit' constitutes 'the true idea of the circulation of money' the disturbance of which may cause economic decline (1766a, pp. 75–6), Turgot analyses the mutual interrelationship of the returns on various types of investment and the rate of interest. Interest itself is shown by Turgot to be determined by the demand for and supply of loanable funds, the demand arising from both consumption and investment needs. These investment needs, or employments of capital as Turgot calls them, are described as: purchasing a landed estate, which yields least; lending a capital at interest the return of which is greater; and investing in agricultural, manufacturing or commercial enterprises, the return of which is greatest. Irrespective of these inequalities in yield to the various employments of capital, Turgot argues that competition combined with capital mobility causes a tendency to equilibrium between them.

> As soon as the profits resulting from an employment of money, whatever it may be, increase or diminish, capitals turn in that direction or withdraw from other employments, or withdraw and turn towards other employments, and this necessarily alters in each of these employments, the relation between the capital and the annual product. (1766a, p. 87)

This investment analysis must be seen as a substantial advance on the earlier literature, and hence as a major contribution to economics.

Turgot's other economic writings can be seem as supplementing the analytical framework of the *Reflections*. This can be particularly illustrated from his theory of value, the outlines of which had been developed by the early 1750s (Schelle, 1913–23, vol. 1, p. 385). Its foundation rested on a relationship between current (market) price and fundamental value dependent on competition and resources mobility. Subsequently (Turgot, 1767b, p. 120 n.) this proposition is elaborated to demonstrate that the market price of a commodity 'ruled as it is by supply and demand' and liable to 'very sudden fluctuations' though 'not in any essential proportion to the fundamental value,… has a tendency to approach it continually, and can never move far away from it permanently'. Turgot therefore developed the classical position which saw 'natural prices' as the centres of gravitation for market prices. Elaborations on the elementary theory of wages of the *Reflections* (1766a, pp. 45–6) are made within this value framework.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Turgot did this in a letter to Hume where on the argument that taxes increase ‘the fundamental price of labour’ or ‘the cost of his subsistence’, a tax on wages must be rapidly absorbed in market wage rates (Schelle, 1913–23, vol. 2, pp. 662–3). The *Reflections* was confined to brief explanations of the current or market price; the unfinished ‘Value and Money’, with its unsuccessful attempts at determining value in various exchange situations, appears as an elaboration of the underlying competitive theory rather than as a new departure towards a more subjective value theory. Turgot's famous analysis of the ‘law of variable proportions’ (Schumpeter, 1954, pp. 260–1) may also be noted here. This arose in criticism of a common Physiocratic assumption that product was invariably proportional to advances. Turgot (1767b, pp. 111–12) argued instead that as ‘advances are gradually increased up to the point where they yield nothing, each increase would be less and less productive’, thereby clearly recognizing the possibility of diminishing returns.

On the basis of these contributions, Schumpeter (1954, pp. 260–1, 307, 332) argued the Turgot was a writer in advance of his time by anticipating much of what became important after the ‘marginal revolution’. Turgot's analysis of the market mechanism is reminiscent of Böhm-Bawerk and Menger; his ‘interest and capital theory … clearly foreshadowed much of the best thought in the last decades of the nineteenth century’. However, it seems more reasonable to conclude ‘that the resemblance between Turgot's economics and that of post-1870 writers is superficial’ and that both in temperament and thrust his economics is part of the classical tradition (Groenewegen, 1982). His development of the Physiocratic notion of reproduction (Turgot, 1763; 1766a, pp. 75–6) and his emphasis on the principle of competition as regulator of the rate of interest, wages and values in general, are firmly within that ‘classical tradition rehabilitated by Sraffa’ (Ravix and Romani, 1984, p. 145).

Turgot's strong laissez-faire position, which turned him into the patron saint of the French liberal economics tradition of the middle of the 19th century, was most systematically expounded in his eulogy of Gournay (1759). This rested on the principle that unrestrained self-interest yields the best results in economic activity, a principle he applied wherever he could during his administrative career. It justified his pleas (1770b) and subsequent imposition of domestic free trade in grain, his criticism of the prevalent regulation of lending at interest (1770a), and the suppression of the guilds in one of his famous 1776 Edicts. More important is his discussion of taxation principles. Turgot's major paper on the subject (Turgot, 1763), after setting out some general principles, defends the concept of the single tax on net product on the basis of Physiocratic theory. However, he identifies difficulties in its implementation. These need detailed examination if the benefits of the policy are to be achieved. More generally, it can be said of his policy implementation that though based on broad principles, these were in practice always modified to cater for actual circumstances.

Turgot's work and its importance in the history of economics have occasionally been vigorously debated, most notably in the controversies over the degree of influence he exerted on Adam Smith and Böhm-Bawerk's interpretation of his capital and interest theory. An assessment of the evidence (Groenewegen, 1969) suggests that Turgot influenced Smith on only a few fairly specific points and that the broad similarities (and differences) in their economic systems are largely explained by their common heritage of British and French predecessors. The quarrel over Böhm-Bawerk's interpretation of Turgot's interest theory (involving Cassel, Wicksell and Marshall) is more instructive for the light it sheds on the participants than for discovering Turgot's views on the subject. For example, it can be suggested that Böhm-Bawerk's position may have been influenced by his considerable youthful debts to Turgot's theory

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
while Marshall's involvement may be explained by antipathy to the Austrian economists and some striking similarities between his and Turgot's interest theory (see Groenewegen, 1983b). This debate highlights his analytical contributions to interest and capital theory.

The doctrine of social progress, which played such an important part in establishing Turgot's vision, was also applied by him to his history of ideas. In his fragment *On Universal History* (1751b, pp. 95–6) a cumulative notion of intellectual progress is presented, in which ideas are seen to develop necessarily from the systems of predecessors; each scientist, as it were, standing on the shoulders of those who came before. Two decades later, Turgot applied this doctrine to the history of economics when defending Melon, the financial economist, against Du Pont's charge that Melon's work was historically unimportant because it was wrong. ‘Someone entering the world after Montesquieu, Hume, Cantillon, Quesnay, M. de Gournay, etc. is less struck by the merit arising from Melon's priority because he does not appreciate it; for him it is no more than a date, and when he reads him, he knew already more than his book’ (Turgot to Caillard, 1 January 1771, in Schelle, 1913–23, vol. 3, p. 500). This line of thought can be applied to Turgot himself. He built on the work of Montesquieu, Hume, Cantillon, Quesnay and Gournay, thereby becoming a major participant in constructing 18th-century classical political economy with noteworthy contributions of his own particularly to the theory of value, capital and interest, production and distribution.

Turgot's works were collected on three occasions: by his friend Du Pont (Turgot, 1808–11), by Daire and Dussard (Turgot, 1844), and by Schelle (1913–23) together with a biography and associated material. Few of his writings were published in his lifetime, but from 1788 to 1792 some of his major economic writings were republished by his friends Condorcet and Du Pont. Comparison of these texts, manuscript versions and the text of the collected works suggests differences attributable to Du Pont, who edited the text for ideological and occasionally political reasons (see Groenewegen, 1977, pp. xxxiv–xxxvi). Schelle first drew attention to, and then removed, many of these corrections, but was not completely successful in this. For this reason, and because of its omissions, particularly of subsequently discovered items from Turgot's voluminous correspondence, Schelle's edition can no longer be described as definitive. Preparing such an edition of Turgot's works awaits both the generous financing required for the task and the services of a devoted editor.

See Also

- physiocracy
- Smith, Adam

Selected works

1749. *Letter to M. l'abbé de Cicé on the replacing of Money by Paper, also known as the ‘Letter on paper money’*. Trans. in Groenewegen (1977).

1750a. *Philosophical Review of the Successive Advances of the Human Mind*. Trans. in Meek (1973).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1750b. Remarques critiques sur les Réflexions Philosophiques de Maupertuis *Sur l'Origine des langues et la signification des mots*. In Schelle (1913–23, vol. 1).

1751a. Lettre á Madame de Graffigny sur les *Lettres d'une Péruvienne*. In Schelle (1913–23, vol. 1).

1751b. *On Universal History*. Trans. in Meek (1973).

1753–4. *Remarks on the Notes to the Translation of Josiah Child* [by Gournay]. Trans. in Groenewegen (1977).

1755. *Reflections on the Expediency of a Law for the Naturalisation of Foreign Protestants*, by Josiah Tucker, translated into French with notes by Turgot. London and Paris.

1756. *Etymologie, Existence, Expansibilité. In Encyclopédie ou Dictionnaire Raisonné des Sciences, des Arts, et Des Métiers*, vol. 6. Paris.

1757a. *Fairs and Markets*. Trans. in Groenewegen (1977).

1757b. *Foundations*. Trans. as Article I in the Appendix to Condorcet, *Life of M. Turgot*, London, 1787.

1759. *In Praise of Gournay*. Trans. in Groenewegen (1977).

1760. *Salomon Gessner, la Mort d'Abel, Poéme*. Traduit par M. Huber [et Turgot]. Paris.

1762. *Salomon Gessner, Idylles et Poémes Champêtres*. Traduit par M. Huber [et Turgot]. Lyon.

1763. *Plan for a Paper on Taxation in General*. Trans. in Groenewegen (1977).

1766a. *Reflections on the Production and Distribution of Wealth*. Trans. in Groenewegen (1977).

1766b. *On the Characteristics of La Grande and La Petite Culture*. In *Quesnay, Farmers 1756 and Turgot, Sur la Grande et la petitite Culture*, ed. P. Groenewegen, Reprints of Economic Classics, Series 2, No. 2, Sydney: University of Sydney, 1983.

1767a. Extract d'un mémoire de M.C. qui contient les principes de l'administration politique, sur la propriété des carrières et des mines, et sur les règles de leur exploitation. In *Ephémérides du Citoyen*, vol. 7. Paris.

1767b. *Observations on a Paper by Saint-Péravy*. Trans. in Groenewegen (1977).

1767c. *Observations on a Paper by Graslin*. Trans. in Groenewegen (1977).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1769. Value and money. Trans. in Groenewegen (1977).

1770a. *Letters on the Grain Trade*. Extracts trans. in Groenewegen (1977).

1770b. *Paper on Lending at Interest*. Extracts trans. in Groenewegen (1977).

1773. *Letter to l'abbé Ternay on the 'Marque des Fers'*. Trans. in Groenewegen (1977).

1778. *Virgile, Didon, Poéme en vers métrieque hexamètres, divisés en trois chants, traduit du 4e livre de l'Enéide*. n.p.

1808–11. *Oeuvres de Turgot précédées et accompagnées de mémoires et de notes sur son vie, son administration et ses ouvrages*, ed. P.S. du Pont de Nemours. Paris.

1844. *Oeuvres de Turgot, nouvelle édition*, ed. E. Daire and H. Dussard. Paris.

Bibliography

Bordes, C. 1981. Présentation. In *Turgot, économiste et administrateur*, ed. C. Bordes and S. Morange. Paris: Presses Universitaires de France.

Bourrinet, J. 1965. Turgot, théoricien de l'individualisme libéral. *Revue d'histoire économique et sociale* 43, 465–89.

Cartelier, J. 1981. La contradiction terre/capital-argent chez Turgot. In *Turgot, économiste et administrateur*, ed. C. Bordes and J. Morange. Paris: Presses Universitaires de France.

Condorcet, J.A.N.Caritat de. 1786. *The Life of M. Turgot, Controller-General of the Finances of France in the Years 1774, 1775 and 1776*. Trans. from the French with an Appendix, London, 1787.

Dakin, D. 1939. *Turgot and the Ancien Régime in France*. London: Methuen.

Du Pont, P.S. 1782. *Mémoires sur la vie et les ouvrages de M. Turgot*. Philadelphia and Paris.

Faure, E. 1961. *La disgrâce de Turgot, 12 Mai 1776*. Paris: Gallimard.

Finzi, R. 1981. Turgot: l'histoire et l'économie: 'Nécessité' de l'économie politique? 'Historicite' des lois économiques? In *Turgot, économiste et administrateur*, ed. C. Bordes and J. Morange. Paris: Presses Universitaires de France.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Finzi, R. 1982. The theory of historical stages in Turgot and Quesnay. *Kenzei, Kenkyu* 33(2), 109–18.

Groenewegen, P.D. 1969. Turgot and Adam Smith. *Scottish Journal of Political Economy* 16(3), 271–87.

Groenewegen, P.D. 1971. A re-interpretation of Turgot's theory of capital and interest. *Economic Journal* 81, 327–40.

Groenewegen, P.D. 1977. *The Economics of A.R.J. Turgot*. The Hague: Martinus Nijhoff.

Groenewegen, P.D. 1982. Turgot: forerunner of neo-classical economics? *Kenzei, Kenkyu* 33(2), 119–33.

Groenewegen, P.D. 1983a. Turgot, Beccaria and Smith. In *Italian Economics, Past and Present*, ed. P. Groenewegen and J. Halevi. Sydney: Frederick May Foundation.

Groenewegen, P.D. 1983b. Turgot's place in the history of economic thought: a bi-centenary estimate. *History of Political Economy* 15, 585–616.

Hausset, Mme de. (n.d.) *Secret Memoirs of the Court of Louis XV and XVI, taken from the Memoir of Madame of Hausset, Lady's maid to Madame de Pompadour and from the Journal of the Princess Lamballe*. London: Grolier Society.

Meek, R.L., ed. 1973. *Turgot on Progress, Sociology and Economics*. Cambridge: Cambridge University Press.

Meek, R.L. 1977. Smith, Turgot and the four stages theory. In R.L. Meek, *Smith, Marx and After*. London: Chapman & Hall.

Morley, J. 1886. Turgot. In *Critical Miscellanies*, vol. 2. London: Macmillan.

Nogaro, B. 1944. *Le développement de la pensée économique*. Paris: Pichon and Durand-Auzias.

Ravix, J.T. and Romani, P.M. 1984. Argent, ‘Capital’ et reproduction chez Turgot. In *Production, circulation et monnaie*, ed. R. Arena et al. Paris: Presses Universitaires de France.

Schelle, G. 1913–23. *Oeuvres de Turgot et documents le concernant*. Paris: Félix Alcan.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press; London: Allen & Unwin.

Tsuda, T. 1974. *Catalogue des livres de la bibliothèque de Turgot*. Tokyo: Hitotsubashi University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Groenewegen, Peter. "Turgot, Anne Robert Jacques, Baron de L'Aulne (1727–1781)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000139> doi:10.1057/9780230226203.1747

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

turnpike trusts

Dan Bogart

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Turnpike trusts were private organizations that built and operated toll roads in Britain and the United States during the 18th and 19th centuries. They emerged in 17th century Britain because local governments were unwilling to invest in roads. They issued bonds to finance investment and imposed tolls on road users. In Britain, travel times and freight charges declined by over 40 per cent during 1750–1800; they fell also in the United States. Turnpike trusts also raised land values and promoted urbanization. They show how changes in institutional arrangements can encourage infrastructure investment and promote economic development.

Keywords

land values; turnpike bonds; turnpike trusts; urbanization

Article

Turnpike trusts were widely employed private organizations that built and operated toll roads in Britain and the United States during the 18th and 19th centuries.

Britain and the United States relied heavily on road transport during their early stages of economic development. They faced a problem, however, because their existing road network was ill-suited for the rising volume of traffic, in particular the growing use of large wagons and carriages. In both economies the demand for road improvements was ultimately satisfied through an institutional innovation known as the ‘turnpike trust’ or the ‘turnpike company.’

Britain had a large network of roads and pathways as early as the 16th century. Although the network was called the ‘Kings Highway’, responsibility for maintenance was placed upon local governments known as parishes. Parishes financed road improvements by forcing their residents to work without pay and by levying property taxes. This method of financing was satisfactory in a pre-industrial economy, in which road improvement costs were low and traffic was largely internal to the parish. Conditions changed during the 17th and 18th centuries, when wages increased and inter-regional trade and travel

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
began to grow. Each of these factors contributed to a divergence between the road expenditure that parishes were willing to provide and the amount needed for an improved network. Turnpike trusts emerged as a solution to this problem. Trusts were promoted by landowners and merchants, who lobbied for an act of parliament. Each act transferred authority from parishes to a body of trustees composed of the promoters and other local property owners. Trustees were given the right to finance improvements along a particular road by levying tolls and issuing bonds. The tolls were efficacious because they forced road-users to contribute to the costs of improvement, whereas the bonds helped to mobilize funds for initial investment. The act also placed a number of restrictions on trustees. For example, they could not charge tolls above a maximum schedule, and they could not earn direct profits. Instead, it was expected that trustees would benefit indirectly through higher property values (Albert, 1972).

The first turnpike act was passed in 1663, and applied to a short section of the Great North Road connecting London with Leeds, York, and Newcastle. The second turnpike act was not passed until 1695, and it was not until the 1720s that trusts became common along the major highways leading into London. Between 1750 and 1770 turnpike trusts diffused throughout much of the road network, especially in the industrializing areas of the West Midlands and the North. After 1770, the network continued to expand, even as canals were being built. By 1840 there were around 1,000 turnpike trusts managing 20,000 miles (Pawson, 1977).

The British colonies of North America inherited the original system, in which roads were free and local governments – parishes, towns, or counties – were responsible for maintenance and improvement. A similar problem emerged where local governments were unwilling to pay for road improvements, despite an increasing need for investment. Significant institutional changes did not occur until after the American Revolution, when states began passing legislative acts creating turnpike companies (Durrenberger, 1931).

US turnpikes companies were similar to British turnpike trusts, except they were corporations and financed most of their investments by issuing stock. Turnpike companies were widely established in New England and the Middle Atlantic states between 1792 and 1845. The early companies were adopted along roads linking major cities such as Boston, Philadelphia, and New York with smaller cities in their western hinterlands. The later turnpike companies generally built and operated roads that led to other turnpikes and canals. By 1845 there were over 800 turnpike companies managing approximately 15,000 miles (Klein and Majewski, 2004).

In Britain and the US the official rationale for creating turnpikes was that the ‘ordinary’ laws for repairing highways needed to be amended if the roads were to be improved. Did turnpike trusts and turnpike companies meet these expectations? In Britain turnpike trusts were generally successful in increasing road maintenance and investment. On average, they spent between 10 and 20 times more than the parishes they replaced. Most trusts purchased land and materials in order to widen their roads and improve the surface. Many trusts also spent substantial sums on maintenance, as they had to deal with the growing volume of traffic in the 18th century (Bogart, 2005a). US turnpike companies were most successful in raising road investment. The amount of capital raised through stock issues was particular striking given that dividends were rarely paid (Klein, 1990). This contrasts with British turnpike bonds, which usually yielded a return of between four and five per cent (Albert, 1972). American turnpike companies had more difficulties paying for maintenance, in part because traffic volumes were lower, but also because companies had difficulties collecting tolls (Klein, 1990).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How were road-users affected by the rise of turnpike trusts and turnpike companies? The evidence for British transport costs shows that the benefits from improved roads substantially outweighed the burden of the tolls, as travel times and freight charges declined by over 40 per cent between 1750 and 1800 (Pawson, 1977; Bogart, 2005b). The evidence from the United States suggests a similar pattern, in which travel times and freight charges fell after turnpike companies improved the road (Durrenberger, 1931). The accounts of contemporaries also suggest that turnpike trusts raised land values and contributed to urbanization. These indirect benefits were especially important because they provided added motivation for landowners and merchants to promote turnpikes and purchase their stocks and bonds.

Turnpikes are often viewed alongside the canal companies and railroads that superseded them in the second quarter of the 19th century. Improving a road was far less expensive, and therefore the turnpike movement did not lead to domestic and international capital flows as with canals and railways. The benefits of turnpikes were also smaller given the natural limits of horse-drawn transport. That said, one should recognize that turnpikes generated substantial benefits in their era (Pawson, 1997; Bogart, 2005b). At a time when local and central governments were largely ineffective, these organizations provided a mechanism by which transport investment could be implemented. Their history also provides an illustration of how changes in institutional arrangements can encourage infrastructure investment and promote economic development.

See Also

- growth and institutions
- Industrial Revolution
- tragedy of the commons

Bibliography

Albert, W. 1972. *The Turnpike Road System in England 1663–1840*. Cambridge: Cambridge University Press.

Bogart, D. 2005a. Did turnpike trusts increase transportation investment in eighteenth-century England?. *Journal of Economic History* 65, 439–68.

Bogart, D. 2005b. Turnpike trusts and the transportation revolution in eighteenth century England. *Explorations in Economic History* 42, 479–508.

Durrenberger, J. 1931. *Turnpikes: A Study of the Toll Road Movement in the Middle Atlantic States and Maryland*. Cos Cob, CT: John E. Edwards, 1968.

Klein, D. 1990. The voluntary provision of public goods? The turnpike companies of early America. *Economic Inquiry* 28, 788–812.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Klein, D. and Majewski, J. 2004. Turnpikes and toll roads in nineteenth century America. EH. Net Encyclopedia, ed. R. Whaples. Online. Available at http://eh.net/encyclopeida/article/Klein.Majewski.Turnpikes, consulted 3 November 2005.

Pawson, E. 1977. *Transport and Economy: The Turnpike Roads of Eighteenth Century Britain*. New York: Academic Press.

How to cite this article

Bogart, Dan. "turnpike trusts." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000196> doi:10.1057/9780230226203.1748

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Tversky, Amos (1937 – 1996)

Eldar Shafir
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Amos Tversky (1937–1996), a cognitive psychologist, is regarded as a giant in the study of human judgment and decision making, and one of the founders of behavioural economics. His early work in mathematical psychology focused on choice, similarity and measurement. With Daniel Kahneman, he collaborated on a highly influential study of judgmental heuristics and biases, and later published a seminal paper on prospect theory, a descriptive theory of individual choice. These projects have had a revolutionary impact on the study of judgment and decision making. Tversky's work has been influential across many disciplines; he won many awards for diverse accomplishments.

Keywords

behavioural economics; decision making; intransitive preferences; judgment; Kahneman, D.; mathematical psychology; measurement; normative theory; preference reversals; prospect theory; risk aversion; similarity models; Tversky, A.; utility maximization; risk seeking; endowment effects; support theory; probability judgments; Econometric Society

Article

Amos Tversky, a cognitive psychologist, is regarded as a giant in the study of human judgment and decision making, and one of the founders of behavioural economics. Born on 16 March 1937 in Haifa, Israel, his father was a veterinarian and his mother was a social worker and member of the first Israeli Parliament and those following, for some 15 years. Tversky received his BA from the Hebrew University in Jerusalem in 1961, majoring in philosophy and psychology, and a Ph.D. in psychology from the University of Michigan in 1965.

The early work

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Tversky's early work in mathematical psychology focused on the study of individual choice behaviour and the analysis of psychological measurement, exploring almost from the beginning the surprising implications of simple and intuitively compelling psychological assumptions. In one early work, Tversky (1969) showed how a series of pair-wise choices could yield intransitive patterns of preference. To do this, he created a set of options such that differences on an important dimension were negligible between adjacent alternatives, but proved to be consequential once compounded across a number of options, yielding a reversal of preference between the first and the last. This pattern not only contradicted a fundamental assumption of utility theory; it also provided a revealing glimpse into the psychological processes involved in decisions of this kind.

For another example, Tversky's (1977) highly influential model of similarity made a number of simple psychological assumptions: items are mentally represented as collections of features, with the similarity between items an increasing function of the features that they have in common, and a decreasing function of their distinct features. Feature weights are task-dependent, such that, for example, the features of the subject of comparison loom larger than the referent's, and common features matter more in judgments of similarity, whereas distinctive features receive greater attention in judgments of dissimilarity. This simple and elegant theory was able to explain observed asymmetries in similarity judgments (A is more similar to B than B is to A), and the fact that item A may be perceived as quite similar to item B and item B quite similar to item C, but items A and C may be perceived as highly dissimilar. Foreshadowing the immensely elegant work to come, these early papers were predicated on the technical mastery of relevant normative theories, and explored simple and compelling psychological principles until their unexpected, and often striking, theoretical implications became apparent.

Another impressive project concerned the mathematical and axiomatic foundations of measurement, in the physical sciences, but especially in the study of behaviour. Although fundamental to modern science, measurement was long considered unproblematic. In fact, it represents non-trivial issues concerning the assignment of numbers to objects in terms of their structural correspondence. Our measurement models, for example, are often are not determined by the data. Tversky's involvement in this project would stretch over two decades and result in three massive volumes (co-authored with Krantz, Luce, and Suppes, 1971; 1989; 1990).

The collaboration with Daniel Kahneman

Tversky's long and extraordinarily influential collaboration with Daniel Kahneman began in 1969 and spanned the fields of judgment and decision making. Having recognized that intuitive predictions and likelihood estimates tend not to follow the principles of statistics or the laws of probability, Tversky and Kahneman (1974) embarked on the study of biases as a method for investigating judgmental heuristics. The beauty of the work was most apparent in the interplay of psychological intuition with normative theory, accompanied by memorable demonstrations. The research showed that judgments often violate basic normative principles despite the fact that people are quite sensitive to these principles' normative appeal. An important theme in this work is a rejection of the claim that people are not able to grasp the relevant normative considerations. Rather, recurrent and systematic errors are attributed to people's reliance on intuitive judgment and heuristic processes in situations where the applicability of normative criteria is not immediately apparent. The experimental demonstrations are noteworthy not only because they violate normative theory, but also because they contradict people's own assumptions about how

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
they make decisions.
Two early examples of judgmental heuristics illustrate this tension:

1. 1. When presented with a description of Linda, a young, single, outspoken and very bright woman, who majored in philosophy, had participated in anti-nuclear demonstrations, and is concerned with issues of discrimination and social justice, most people think Linda is more likely to be a feminist bank teller than a bank teller – even though, of course, the likelihood of the latter must be greater than the former (since all feminist bank tellers are bank tellers).
2. 2. When asked to estimate the number of seven-letter words on a typical page of English text, people are inclined to guess that there are fewer words whose penultimate letter is N than end in ING – even though the latter are necessarily a subset of the former.

In both cases, a heuristic judgment leads to what is known as the conjunction fallacy. In the first, people rely on the fact that Linda is more similar to a feminist bank teller than to a prototypical bank teller; in the second, frequency is judged via the ease with which examples can be brought to mind. In both cases, the reliance on intuitive heuristics leads people to ignore simple normative constraints that, upon reflection, they readily endorse.
In 1979, Kahneman and Tversky published their seminal paper on prospect theory. Although the theory is formally confined to the analysis of individual choice between binary monetary gambles, it incorporates fundamental insights that have revolutionized current theorizing about decision making more generally. Contrary to the notion of utility maximization, which focuses on final assets, the psychological carriers of value in prospect theory are gains and losses relative to some reference point, which is often the status quo. Diminishing sensitivity to greater amounts leads prospect theory's value function to be concave for gains and convex for losses (that is, above and below the reference point, respectively), yielding risk aversion for gains and risk seeking for losses (except for very low probabilities, where these trends can reverse). Because prospects can often be framed as gains or as losses relative to some reference point, this can generate ‘framing effects’, wherein alternative descriptions trigger opposing risk attitudes and elicit discrepant preferences regarding the same final outcomes. For example, imagine being $300 richer than you are and having a choice between $100 for sure and an equal chance at $200 or nothing. Alternatively, imagine being $500 richer and having to choose between a sure $100 loss and an equal chance to lose $200 or nothing. Although the two scenarios offer the same final outcomes ($400 versus an equal chance at $300 or $500), people tend to prefer the certain $100 gain in the first and the chance of a greater loss or nothing in the second, thus expressing opposing preferences.
According to prospect theory, people are loss averse: the loss associated with giving up a good is greater than the pleasure associated with obtaining it. Loss aversion yields ‘endowment effects’ wherein the mere possession of a good can lead to higher valuation of it than if it were not in one's possession (Kahneman, Knetsch and Thaler, 1990), and it can create a general reluctance to negotiate or trade because the disadvantages of departing from the status quo loom larger than the advantages presented by possible alternatives (Samuelson and Zeckhauser, 1988). Furthermore, the impact of probabilities in prospect theory is not linear; rather, it consists of a transformation of the relevant probabilities into ‘decision weights’ which capture the impact on decision makers, exhibited most clearly at the extremes of certainty and impossibility. For example, a reduction in the likelihood of a threatening outcome

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
from .02 to 0 has a much greater impact on people (as exhibited, say, in their willingness to pay) than a comparable change in likelihoods from .67 to 65.

Later work

Tversky returned to the study of judgment and in his work on support theory (Tversky and Koehler, 1994), a theory of probabilistic judgment that formally distinguishes between events in the world and the manner in which they are mentally represented. Probabilities in support theory are attached not to events, as in standard models, but rather to descriptions of events, called hypotheses. Probability judgments are based on the support (strength of evidence) of the focal hypothesis relative to that of alternative, or residual, hypotheses. The theory distinguishes between *explicit disjunctions*, which are hypotheses that list their individual components (for example, ‘a car crash due to oil spill, or due to driver fatigue, or due to break failure’), and *implicit disjunctions* that do not (‘a car crash’). According to the theory, unpacking the description of an event from an implicit to an explicit disjunction generally increases its support and, hence, the perceived likelihood. As a result, alternative descriptions of an event can give rise to substantially different judgments.

A fundamental assumption underlying normative theories is the extensionality principle: options that are extensionally equivalent are assigned the same value, and extensionally equivalent events are assigned the same probability. Normative theories are concerned with options and events in the world: different descriptions of the same states are similarly evaluated. According to Tversky's analyses, on the other hand, judgments and decisions are constructed, not merely revealed, during their elicitation, and their construction depends on the framing of the problem, the method of elicitation, and the valuations and attitudes that these trigger. The extensionality principle is deemed descriptively invalid because alternative decision contexts and alternative descriptions of options or events often produce systematically different judgments and preferences.

Behaviour, Tversky's research made clear, is the outcome of normative ideals that people endorse upon reflection, combined with psychological processes that intrude upon and shape behaviour independently of any deliberative intent. These insights led to dramatic and memorable studies concerning, among others, the hot hand in basketball (Tversky and Gilovich, 1989), the perceived relationship between weather and rheumatism (Redelmeier and Tversky, 1996), money illusion (Shafir, Diamond and Tversky, 1997), self-deception (Quattrone and Tversky, 1984), overconfidence (Griffin and Tversky, 1992), and a variety of other economic, medical and political decisions. Tversky was an intellectual giant whose work had an exceptionally broad appeal, to economists, philosophers, statisticians, physicians, political scientists, sociologists and legal theorists, among others.

Tversky taught at Hebrew University (1966–78) and at Stanford University (1978–96), where he was the inaugural Davis–Brack Professor of Behavioral Sciences and Principal Investigator at the Stanford Center on Conflict and Negotiation. He spent leave periods at Harvard University, the Center for Advanced Studies in the Behavioral Sciences, the Center for Advanced Study at Hebrew University, and the Oregon Research Institute. After 1992 he held an appointment as Senior Visiting Professor of Economics and Psychology and Permanent Fellow of the Sackler Institute of Advanced Studies at Tel Aviv University.

Tversky won many awards for diverse accomplishments. As a young officer in 1956, he earned Israel's highest honour for bravery for rescuing a soldier who had frozen in panic after lighting an explosive

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
charge. His dissertation, under the supervision of Clyde Coombs, won the University of Michigan's Marquis Award. He won the Distinguished Scientific Contribution Award of the American Psychological Association in 1982, a MacArthur Prize in 1984, and the Warren Medal from the Society of Experimental Psychologists in 1995. He was a foreign member of the National Academy of Sciences, and a member of the Econometric Society and the American Academy of Arts and Sciences. He was awarded honorary doctorates by the University of Göteborg, the State University of New York at Buffalo, the University of Chicago, and Yale University.

Tversky was in the midst of an enormously productive time when he died of metastatic melanoma on 2 June 1996, at his home in Stanford, California. For a selection of his writings, as well as a complete bibliography, see Shafir (2004); for excellent collections of papers influenced by Tversky's work on judgment and choice, respectively, see Gilovich, Griffin and Kahneman (2001), and Kahneman and Tversky (2000).

When it awarded Daniel Kahneman the 2002 Nobel Memorial Prize in Economic Sciences ‘for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty’, the Royal Swedish Academy of Sciences, which does not award prizes posthumously, took the unusual step of acknowledging Tversky in its Nobel citation, explaining that his joint work with Kahneman formulated alternative theories that better account for observed behaviour. Two months later, Tversky also posthumously won with Kahneman the prestigious 2003 Grawemeyer Award, which recognizes powerful ideas in the arts and sciences. The citation noted that it was ‘difficult to identify a more influential idea than that of Kahneman and Tversky in the human sciences’.

See Also

- Kahneman, Daniel
- prospect theory

Selected works

1969. The intransitivity of preferences. *Psychological Review* 76, 31–48.

1977. Features of similarity. *Psychological Review* 84, 327–52.

1971. (With D.H. Krantz, R.D. Luce and P. Suppes.) *Foundations of Measurement: Vol. 1. Additive and Polynomial Representations*. San Diego: Academic Press.

1974. (With D. Kahneman.) Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–31.

1979. (With D. Kahneman.) Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1984. (With G.A. Quattrone.) Causal versus diagnostic contingencies: on self-deception and on the voter's illusion. *Journal of Personality and Social Psychology* 46, 237–48.

1989. (With T. Gilovich.) The cold facts about the 'hot hand' in basketball. *Chance* 2(1), 16–21.

1989. (With D.H. Krantz, R.D. Luce and P. Suppes.) *Foundations of Measurement: Vol. 2. Geometrical, Threshold, and Probabilistic Representations*. San Diego: Academic Press.

1990. (With D.H. Krantz, R.D. Luce and P. Suppes.) *Foundations of Measurement: Vol. 3. Representation, Axiomatization, and Invariance*. San Diego: Academic Press.

1992. (With D. Griffin.) The weighing of evidence and the determinants of confidence. *Cognitive Psychology* 24, 411–35.

1994. (With D.J. Koehler.) Support theory: a nonextensional representation of subjective probability. *Psychological Review* 101, 547–67.

1996. (With D.A. Redelmeier.) On the belief that arthritis pain is related to the weather. *Proceedings of the National Academy of Sciences* 93, 2895–6.

1997. (With E. Shafir and P. Diamond.) On money illusion. *Quarterly Journal of Economics* 112, 341–74.

2000. (With D. Kahneman, eds.) *Choices, Values, and Frames*. New York: Cambridge University Press and Russell Sage Foundation.

Bibliography

Gilovich, T.D., Griffin, D.W. and Kahneman, D., eds. 2001. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge Univ. Press.

Kahneman, D., Knetsch, J.L. and Thaler, R. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98, 1325–48.

Samuelson, W. and Zeckhauser, R. 1988. Status quo bias in decision making. *Journal of Risk and Uncertainty* 1, 7–59.

Shafir, E., ed. 2004. *Preference, Belief, and Similarity: The Selected Writings of Amos Tversky*. Cambridge, MA: MIT Press.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Shafir, Eldar. "Tversky, Amos (1937–1996)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000187>
doi:10.1057/9780230226203.1749

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Twiss, Travers (1809 – 1897)

Murray Milgate
From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Keywords

Twiss, Travers; political economy

Article

In economics, Twiss's reputation rests primarily upon two contributions: one on the machinery question and the other a *View of the Progress of Political Economy in Europe since the Sixteenth Century* (1847) of some 300 pages. Both of these works originated in lectures during his tenure as Drummond Professor of Political Economy at Oxford (1942–7). The latter numbers with McCulloch's much shorter *Historical Sketch of the Rise and Progress of the Science of Political Economy* (1926) as being among the first significant histories of the discipline published in English. The only works of comparable significance in the area which predate it appeared in French: Blanqui's *Historie de l'économie politique en Europe* (1837–8) and Jean Paul Alban de Villeneuve-Bargemon's *Historie de l'économie politique* (1836–8 and 1841). Twiss acknowledges his debt to the abovementioned authors, but has been criticized (for example, by Cossa) for a tendency to rely too heavily upon second-hand sources in the construction of his argument.

The published versions of his lectures at Oxford are all that Twiss left to the literature of economics. Twiss was born in London on 19 March 1809 and died there on 14 January 1897, and was educated at University College, Oxford, taking his BA (in mathematics and classics) in 1830. From 1830 until 1863 he was a fellow of that college. In 1835 he commenced the study of law in Lincoln's Inn and was admitted to the Bar in 1840. Following his term as Drummond Professor (in which he succeeded Merivale), he turned more and more to the study of international law, and in 1852 he was elected to the chair in that field at King's College, London. In 1855 he moved to Oxford as Regius Professor of Civil Law, where he remained until 1870. In 1867 he became the Queen's advocate-general, and was knighted in 1868.

At this point occurred ‘the catastrophe which put an end to his official career’, as the original edition of this *Dictionary* put it. It seems that in 1872, Twiss instituted an action for malicious libel with intent to extort against a solicitor who had put about statements impugning the moral propriety of Twiss's wife.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As the case proceeded, Lady Twiss was called to testify. However, an arduous cross-examination proved to be too much for her, and she departed London before its conclusion, thus causing Twiss's case to collapse and precipitating his resignation from all offices. Of course, it is not surprising (given the climate of the times) that Lady Twiss's breakdown should have been interpreted as telling evidence against her – but from what we now know of these extraordinary Victorian public rituals over sexual behaviour and preference, and of the pressures placed on the principal actors in such notorious trials, a rather different verdict might just as plausibly be drawn from the episode. From the point of view of individual and social psychology, however, even more interesting is the question of just why these kinds of cases were voluntarily brought before the courts in the first place.

Selected works

1845. *On Certain Tests of a Thriving Population*. London.

1847. *View of the Progress of Political Economy in Europe Since the Sixteenth Century*. London: Longmans, Brown, Green and Longmans.

1861. *Law of Nations considered as Independent Political Communities*. Oxford: Oxford University Press.

How to cite this article

Milgate, Murray. "Twiss, Travers (1809–1897) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000068> doi:10.1057/9780230226203.1888

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

two-sector models

H. Uzawa
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

marginal rate of substitution; Rybczynski theorem; two-sector models

Article

Problems of interest in economic theory, from both the theoretical and policy points of view, occur only when there exists a multitude of goods and services, each of which is either produced by different technologies or utilized for different purposes in consumption. However, it is often the case that an economic system with a multitude of goods and services is too complicated to analyse effectively and to derive conclusions of any practical use. two-sector models enable us to bring forth essential elements of the economic mechanisms in a more complicated real world while still making it possible to analyse graphically the basic structure of equilibrium and to understand the policy implications within the framework of the two-sector analysis. The two-sector analysis plays a particularly important role in trade theory and in growth theory.

A typical two-sector model concerns itself with an economy in which there exist two productive sectors, to be referred to as sector 1 and sector 2, respectively. In the context of growth theory, one sector produces consumption goods and the other investment goods. Both goods are assumed to be composed of homogeneous quantities and to be produced by two factors of production, capital and labour. Both capital and labour are also assumed to be composed of homogeneous quantities.

In each sector, production is assumed to be subject to constant returns to scale and diminishing marginal rates of substitution between capital and labour. Joint products are excluded and external (dis-)economies do not exist. The output in each sector is determined by the quantities of capital and labour allocated to that sector. In sector j, let Y_j be the quantity of good j produced by the input of capital and labour by the quantities K_j and L_j, respectively, then we may write

$$Y_j = F_j(K_j, L_j), \quad j = 1, 2. \tag{1}$$

For each j, the production function $F_j(K_j,\bullet L_j)$ is linear homogeneous and continuously differentiable, so that the marginal rate of substitution between capital and labour is well defined.

Let K and L be the quantities of capital and labour which exist in the economy at a particular moment of time. If both capital and labour are assumed to be freely transferred from one sector to another and both are fully employed, then we have

$$K_1 + K_2 = K, \quad L_1 + L_2 = L. \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In a typical two-sector model, it is often assumed that the allocation of two factors of production is perfectly competitive, so that in each sector the wage w is equal to the marginal product of labour and the rentals r of capital goods to the marginal product of capital:

$$w = p_j \frac{\partial F_j}{\partial L_j}, \quad r = p_j \frac{\partial F_j}{\partial K_j}, \tag{3}$$

where p_j is the price of good j.
In what follows, good 1 is taken as the numéraire, so that $p_1=1$ and $p_2=p$.
Since production is assumed to be subject to constant returns to scale, the model is reduced to one involving per capita quantities only. Let us introduce the following notation:

- $k=K/L$: the capital–labour ratio in the economy as a whole,
- $k_j=K_j/L_j$: the capital–labour ratio in sector j,
- $Y_j=Y_j/L$: output of good j per capita,
- $v_j=L_j/L$: the proportion of labour allocation in sector j,
- $\omega=w/r$: the wage–rental ratio.

The relations (1)–(3) are then reduced to the following:

$$y_j = f_j(k_j)v_j, \tag{4}$$

where $f_j(k_j)=F_j(k_j, 1)$,

$$v_1 = \frac{k - k_2}{k_1 - k_2}, \quad v_2 = \frac{k_1 - k}{k_1 - k_2}, \tag{5}$$

$$\omega = \frac{f_j(k_j)}{f_j'(k_j)} - k_j, \tag{6}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$p = \frac{f_1'(k_1)}{f_2'(k_2)}, \tag{7}$$

$$y_1 = f_1(k_1)\frac{k - k_2}{k_1 - k_2}, \quad y_2 = f_2(k_2)\frac{k_1 - k}{k_1 - k_2}. \tag{8}$$

The relation (6) means that the wage–rentals ratio ω is equal to the marginal rate of substitution between capital and labour. The capital–labour ratio k_j which satisfies (6) is uniquely determined for given wage–rentals ratio ω; it may be written $k_j=k_j(\omega)$, which is referred to as the optimum capital–labour ratio corresponding to the wage–rentals ratio ω. It is easily seen that the optimum capital–labour ratio $k_j(\omega)$ is an increasing function of the wage–rentals ratio ω. In fact, by differentiating (6) with respect to ω, we get

$$\frac{dk_j}{d\omega} = \frac{\left[f_j'(k_j)\right]^2}{f_j(k_j)f_j''(k_j)} > 0, \tag{9}$$

because of the diminishing marginal rate of substitution condition: $f'_j(k_j)>0$ and $f_j''(k_j) < 0$.
The relationships between the price ratio p and the wage–rentals ratio ω may be obtained by differentiating (7) logarithmically, and noting (9):

$$\frac{1}{p}\frac{dp}{d\omega} = \frac{1}{\omega + k_2(\omega)} - \frac{1}{\omega + k_1(\omega)}. \tag{10}$$

Hence, we have the following proposition:
The relative price p of good 2 is an increasing or decreasing function of wage–rental ratio ω according to whether good 1 is more or less capital-intensive than good 2.
In particular, if good 1 is always more capital-intensive than good 2, then the relative price p of good 2 (with respect to the price of good 1) is an increasing function of the wage–

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

rentals ratio ω. The latter is indeed nothing but the essence of the factor–price–equalization theorem. In this case, we can see from (8) that, as the endowment ratio k is increased, the output of good 1 is increased and that of good 2 is decreased, provided that the wage–rentals ratio ω or relative price p remains constant.

The wage–rentals ratio ω or price ratio p is determined once the demand conditions are specified.

The allocation of capital and labour between two sectors in the perfectly competitive situation, as described above, may be viewed from another point of view. It is easily seen that the allocation of capital and labour which satisfies (1)–(3) is nothing but the solution of the following optimum problem:

Find the allocation (K_1, K_2, L_1, L_2) which maximizes the national product

$$Y = p_1 Y_1 + p_2 Y_2$$

subject to the constraints (1) and

$$K_1 + K_2 \leq K, \quad L_1 + L_2 \leq L. \qquad (2')$$

In fact, wage w and rentals r are the Lagrange multipliers associated with the constraints (2′).

The set of all combinations (Y_1, Y_2) of two goods satisfying (1) and (2′) then is a convex set, of all possible combinations of the quantities of two goods which can be produced from the given endowments of capital and labour, K and L. The competitive allocations of capital and labour then result in those combinations of two goods for which the national product, evaluated at prices p_1 and p_2, is maximized.

These observations lead us to the following conclusion. Namely, if the demand conditions are those obtained by an optimization of a certain community preference ordering, then the equilibrium prices and outputs are uniquely determined.

The analysis may be carried out in terms of a geometric presentation. For the given techniques of production and the given factors of production, the set of all possible combinations of two goods produced in the two-sector economy is represented by the production possibility set, as shown by the shaded area in Figure 1. In Figure 1, the quantities of good 1 and good 2 are measured along the abscissa and ordinate, respectively, and the boundary curve of the production possibility set is the transformation curve F_1F_2, showing the maximum quantity of one good that can be produced, given a specific quantity of the other to be produced. The transformation curve is concave toward the origin and the tangent at each point on the transformation curve has a slope equal to the relative price of two goods, as shown in Figure 1. It is possible to prove these properties by using the contract box, as in Figure 2. In Figure 2, the endowments of capital and labour are measured along the sides of the box, and the allocations of capital and labour between two sectors are entered in the box from opposite corners. An efficient allocation of factors of production is realized only at a point at which two isoquants are tangent to each other. The efficient locus in the contract box corresponds to the transformation curve in Figure 1. The configuration described in Figure 2 represents the case where good 1 is more capital-intensive than good 2. Let the point A in Figure 1 correspond to the point A in the contract box in Figure 2. Suppose the quantity of good 1 to be produced is reduced by ΔY_1 and the production of good 2 is increased by ΔY_2, resulting in a shift from A to A' along the efficiency locus. At point A, the isoquants in two sectors have a common tangent; let B_1 and B_2 be the points on the labour-side at which the tangent line at A intersects. The two distances, O_1B_1 and O_2B_2, measure the values of the two goods produced, p_1Y_1/w and p_2Y_2/w, respectively. Let C be the point at which O_1A intersects with the isoquant passing through A', and let B_1' and B_2' be the points on the labour-side at which the tangent line at C intersects. Then $B_1'B_1 = p_1\Delta Y_1 / w$ and $B_2'B_2 \simeq p_2\Delta Y_2 / w$, where the symbol $\simeq$ indicates that both sides are approximately equal, converging to the equality as ΔY_1 approaches to 0. Hence, $\Delta Y_1 / \Delta Y_2 \simeq p_2 / p_1$.

Figure 1

The transformation curve

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

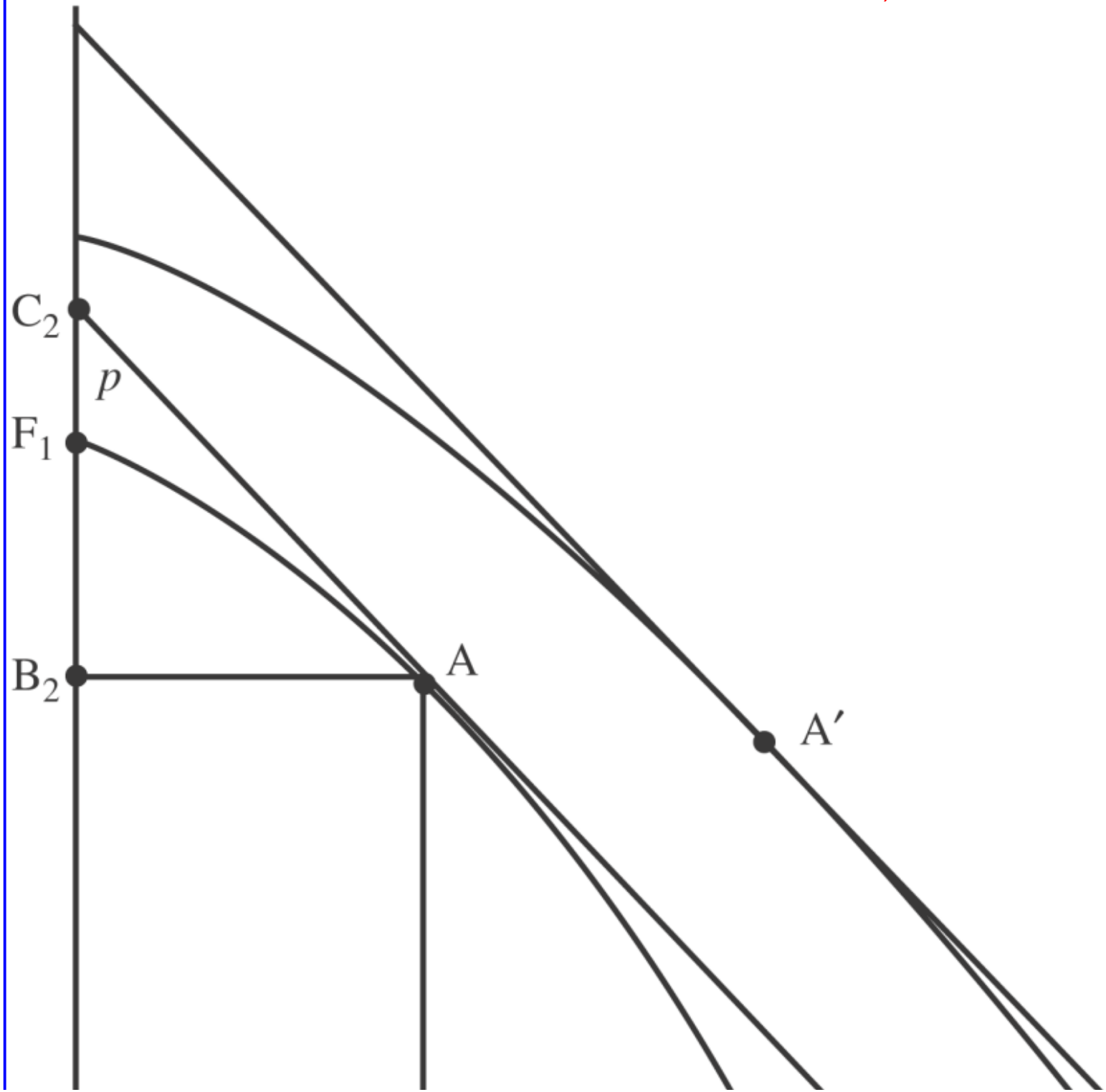

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

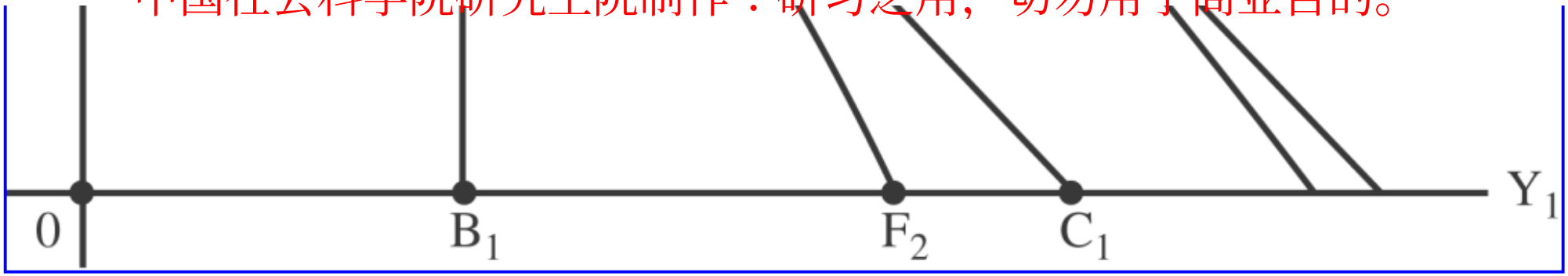

Figure 2
The contract box

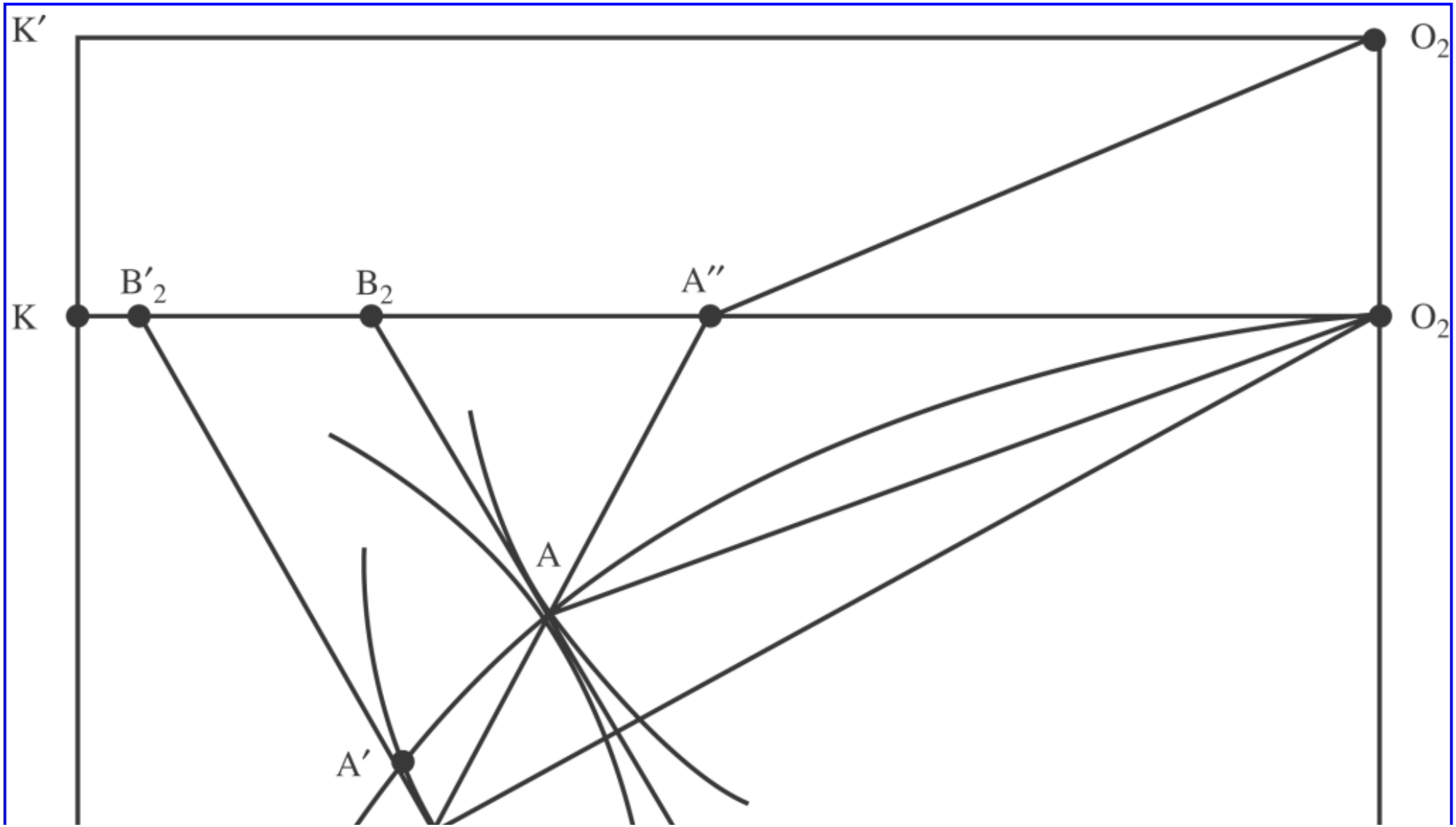

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

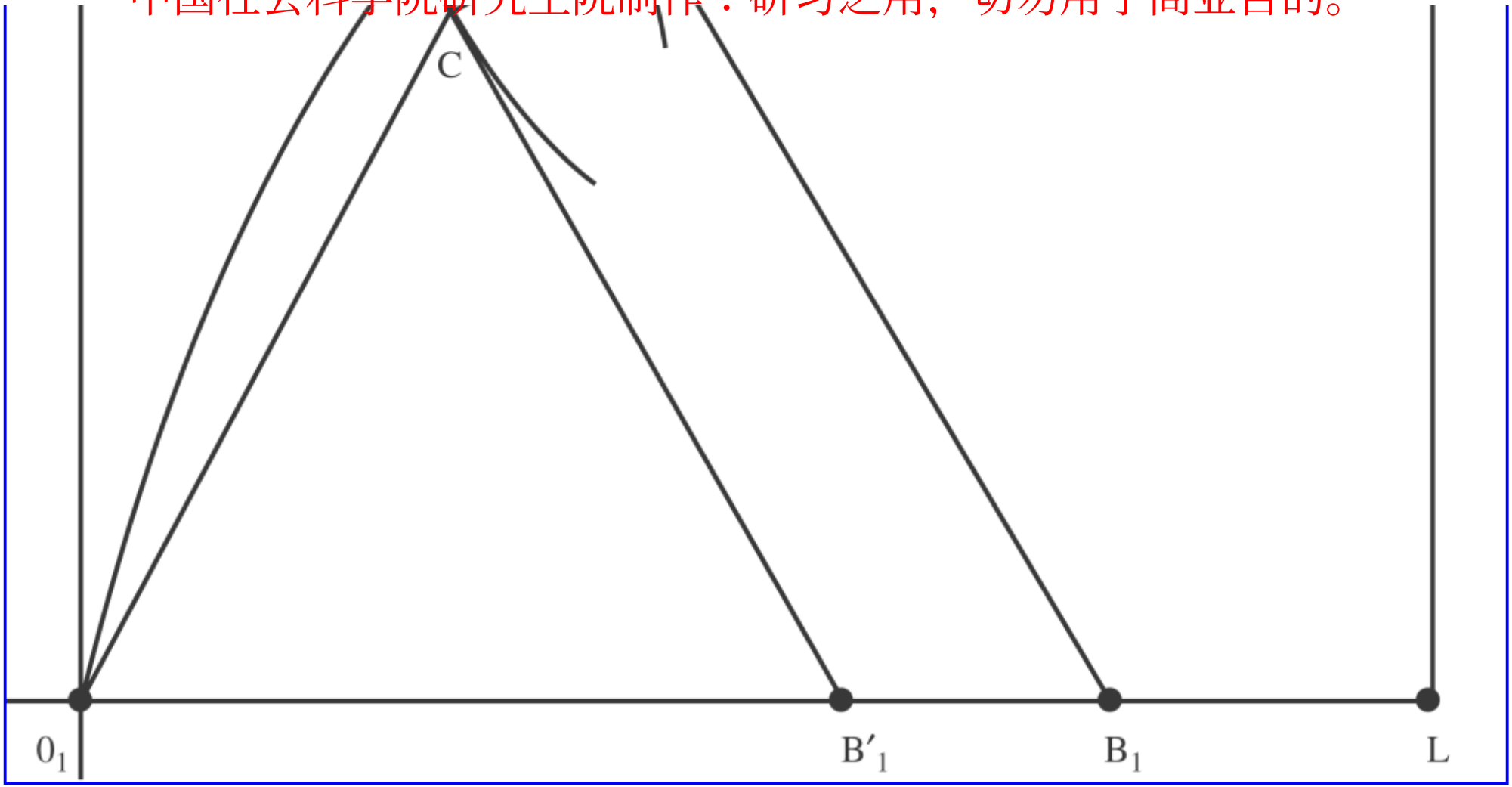

As the output of good 1 is reduced, the point *A* moves towards O_1 along the efficiency locus. Then the optimum capital–labour ratio, which is represented by LAO_1B_1, is increased, and the tangent line at A' is steeper than that at *A*, indicating an increase in the price ratio p_2/p_1. The transformation curve thus is shown to be concave towards the origin and the tangent at each point on it is equal to the price ratio.

The relationships between the price ratio and the wage-rentals ratio may also be discussed in terms of a two-dimensional diagram, as in Figure 3. Suppose good 1 is more capital-intensive than good 2. For the given wage–rentals ratio ω, the unit by which good 2 is measured is so adjusted that $p_2/p_1=1$ and the unit-isoquants for both goods share the same cost line *CC*, as shown in Figure 3. The distance *OC* along the abscissa measure $p_1/w=p_2/w$. Suppose the wage–rentals ratio is increased from ω to ω'. Then at the new wage–rentals ratio ω', the configuration of the cost lines in two sectors must be of the form described in Figure 3. Namely, $OB_1<OB_2$; hence, $p_1'/w' < p_2'/w'$, implying $p_2'/p_1' < 1 = p_2/p_1$. Thus we have proved that, as the wage–rentals ratio is increased, the price of a good which is more labour-intensive has been increased relative to that of a less labour-intensive good.

Figure 3

The relationships between price ratio *p* and wage–rentals ratio ω

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

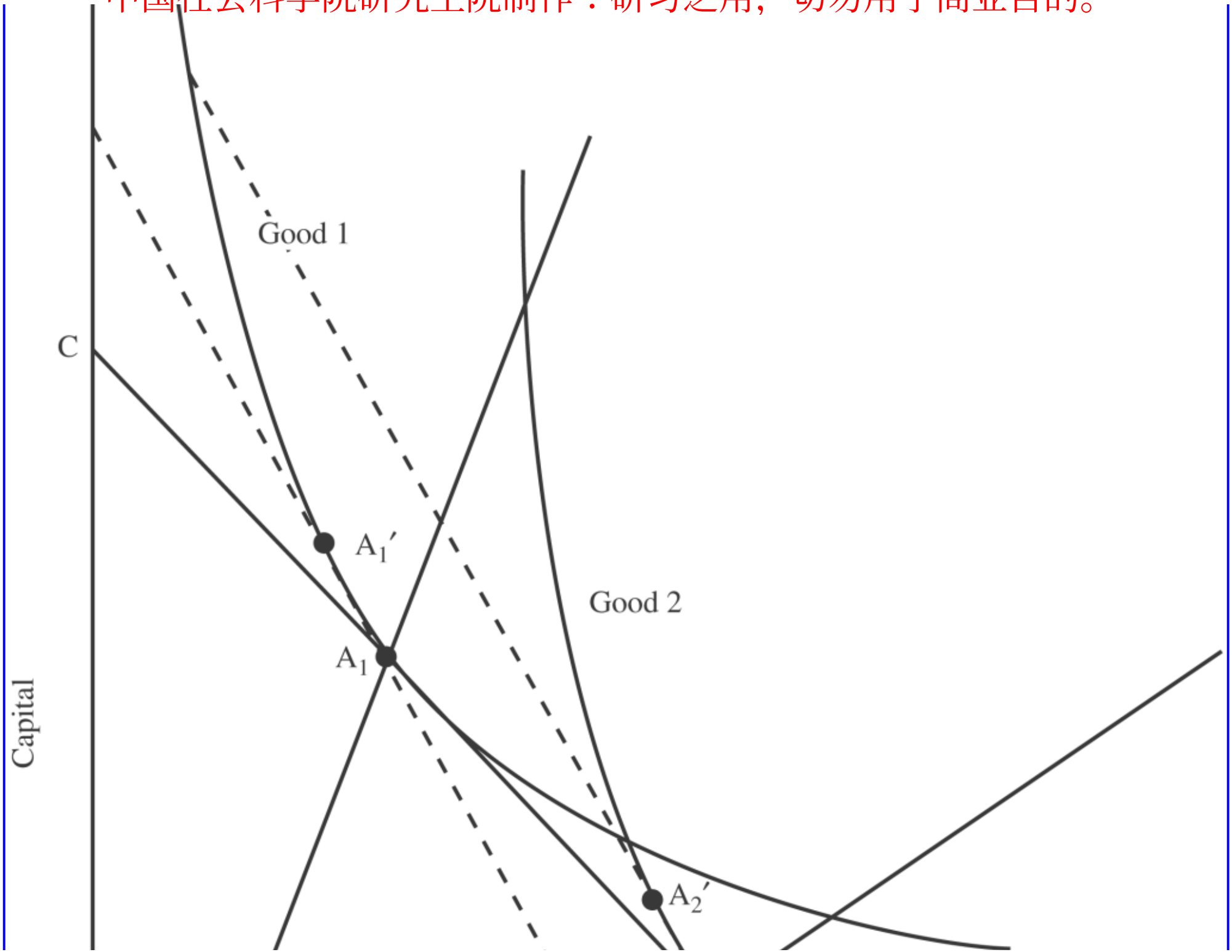

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

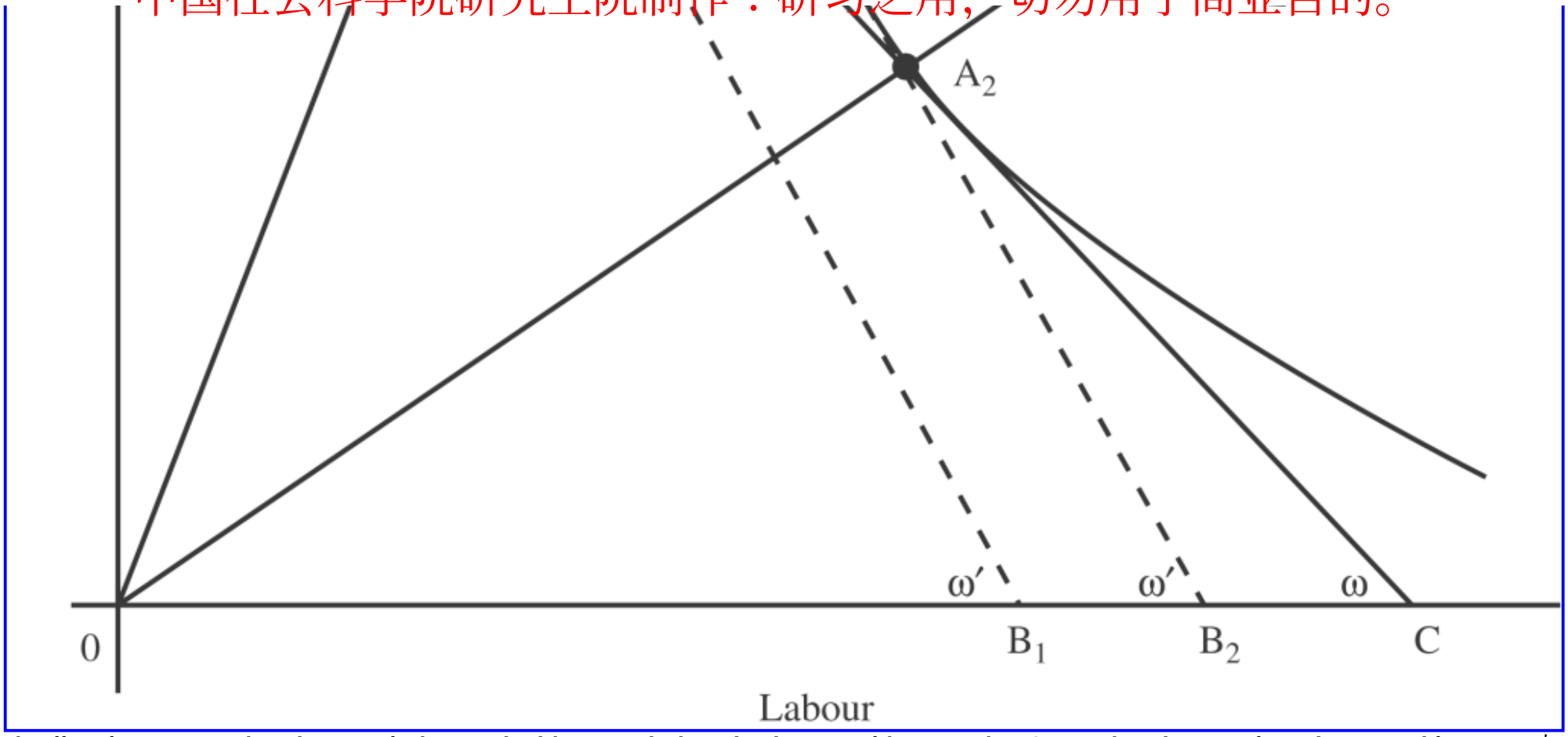

The effect of an increase in the endowment of either capital or labour may also be analysed in terms of the contract box. Suppose the endowment of capital is increased from K to K' so that the new contract box is indicated by $O_1LO_2'K'$ in Figure 2. If the relative price $p=p_2/p_1$ remains unchanged, then the factor price ratio $\omega=w/r$ also remains unchanged. Let A'' be the point at which the extension of O_1A intersects with the line originating from O_2' which is parallel to O_2A. Then at the new configuration, the output of good 1 is increased by AA'', while the output of good 2 is decreased from O_2A to $O_2'A'$. Thus we have shown the proposition, known as the Rybczynski theorem: An increase in the endowment of capital increases the output of a good which is more capital-intensive than the other, while decreasing the output of another good which is less capital-intensive, provided the price ratio of the two goods remains constant. The resulting shift in the transformation curve is described in Figure 3, where the efficient point A moves to A' in the new environment.

See Also

- factor price frontier

How to cite this article

Uzawa, H. "two-sector models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000146>

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
doi:10.1057/9780230226203.1750

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

two-sided markets

M. Armstrong and J. Wright
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A growing number of industries are organized as so-called two-sided markets in which platforms enable interactions between two groups of users, each of which cares about the size and attributes of the other group on the same platform. The literature to date examines how platforms set prices to the two sides and whether the resulting price structure results in market failure. The answers to these questions depend on the nature of cross-group and own-group externalities, the types of fees possible (membership or per-transaction), and whether one or both sides multihome.

Keywords

access pricing; advertising; competitive bottlenecks; complementary goods; cross-group externalities; matching; multihoming; network effects; platforms; price discrimination; shopping malls; singlehoming; two-sided markets; user fees

Article

There are many examples of markets where two or more groups of participants interact via ‘platforms’. Of course, there are countless examples where firms compete to supply two or more groups. However, in a set of interesting cases, cross-group network effects are important, and the benefit enjoyed by a member of one group on a platform depends upon how well that platform does in attracting custom from the other group. For instance, a general purpose credit card scheme cannot offer a valuable service to either side unless it persuades a large number of consumers to carry its card and a large number of retailers to accept its card. The literature on two-sided markets investigates such markets.
Many examples of two-sided markets involve platforms that mediate transactions between consumers and retailers (or sellers). Examples include shopping malls, supermarkets, and debit and credit card payment schemes. Another set of examples includes matching agencies, such as real estate agencies that facilitate search and trade between home buyers and home sellers (or landlords and tenants). Advertising

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
media including Yellow Pages, newspapers, television, and Internet portals also help match potential buyers with sellers, although in a less directed way. Software platforms such as video games, computer operating systems and word processors that connect users and application developers (or readers and writers) provide a further set of examples.

Early theoretical works on two-sided markets tended to focus on specific industries, such as Baxter's (1983) normative analysis of the structure of fees in a credit card network (see credit card industry). More general theoretical frameworks to analyse two-sided markets have been offered by Armstrong (2006), Armstrong and Wright (2007), Caillaud and Jullien (2001; 2003), Hagiu (2005), and Rochet and Tirole (2003), among others. These models extend the earlier literature on network externalities to incorporate heterogeneous agents (the two sides of the market), as well as allowing for price discrimination across the two types of users. A central question is: What determines the structure of prices in two-sided markets? For instance, why is it that shopping malls offer free parking to shoppers and recover the cost from retailers, and why does American Express charge merchants but provide rebates to cardholders? A second question concerns whether the resulting price structure causes any form of market failure. We consider both questions.

Effects of cross-group externalities

Consider a generic two-sided market with buyers and sellers. Positive cross-group externalities have two effects. First, like network effects in one-sided markets, they make demand more sensitive to price. Second, like pricing for complementary goods, they make platforms charge less to one group if this increases the demand from the other group and the other group generates a positive margin. This implies that platforms will charge buyers less and sellers more when sellers value buyers more than vice versa (Armstrong, 2006). By attracting buyers (with a discount), platforms can then attract the more lucrative sellers. Consistent with this idea, Yellow Pages directories are typically given away to readers for free, and profits are made entirely from charging advertisers. This assumes, of course, that the two sides cannot easily internalize the externalities between themselves – a precondition for the structure of prices to be non-neutral in a two-sided market (Rochet and Tirole, 2006). It also assumes that agents make decisions about which of the competing platforms to join. This latter assumption is explored in the following two sections.

Membership fees versus usage charges

Platforms may charge for their services on a ‘lump-sum’ basis: magazines set cover prices and advertisement rates, nightclubs set entry fees for men and women. Alternatively, charges might be levied on a ‘per-transaction’ basis: typically credit card holders receive a percentage rebate on the amount they spend, and retailers pay a percentage of the revenue they collect; real estate agents’ fees are levied only in the event of a sale; charges for telecommunications service are levied on a per-call basis. And sometimes a combination of the two approaches is used: video game platforms charge consumers a fixed charge for their consoles, but game developers pay royalties for their sales.

The analysis of two-sided markets can be quite sensitive to the nature of pricing. With lump-sum charges, profits are the sum of profits obtained from each side, and it is possible to think of one side as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
subsidizing the other. When cross-group network effects are strong, there are often multiple consistent demand outcomes for a given set of prices, which means that some method (such as specifying consumers' beliefs about other consumers' choices) is needed to pin down demand as a function of prices uniquely (Caillaud and Jullien, 2001; 2003).

With per-transaction charges, by contrast, an agent's decision whether to join a platform is less sensitive to his beliefs about the number of agents from the other side who join the platform. In this case, the equilibrium structure of prices will primarily reflect the need to balance the two sides of the market (Schmalensee, 2002; Wright, 2004). A matching market with lots of potential sellers but few potential buyers will not generate many successful transactions, and this suggests that more buyers need to be attracted to the platform by charging sellers rather than buyers for successful transactions.

A crucial difference between the two forms of tariff is that cross-group externalities are less important with usage charges. Charging on a usage basis is a good strategy for an entrant. If an agent has to pay a new platform only in the event of a successful transaction, then the agent does not have to worry about how well the entrant does in its dealings with the other side of the market. With per-transaction charging, to attract one side a new platform does not have to first get the other side 'on board'.

Singlehoming versus multihoming

When an agent chooses to use only one platform, that agent is said to 'singlehome' and when he uses several platforms he 'multihomes'. For instance, while shoppers may tend to only shop at their nearby shopping mall (singlehome), retailers may locate in several shopping malls (multihome) in order to gain access to the full range of local shoppers. Similarly, people might read a single newspaper each day, while advertisers have to place adverts in several newspapers to reach the whole readership. This pattern, where 'buyers' singlehome and 'sellers' multihome, characterizes a number of two-sided markets. Armstrong and Wright (2007) show this pattern arises endogenously when only buyers view the platforms as differentiated (as may be true in the two examples just mentioned). This leads to a 'competitive bottleneck' – platforms compete aggressively to sign up buyers, charging them less than cost (perhaps nothing), and then make their profits from sellers who want to reach these buyers and do not have a choice of which platform to join in order to reach them. Platforms need to compete to attract the singlehoming buyers but they hold a monopoly position when they deal with sellers. As charges to sellers will be too high, there will be too few sellers from a social welfare point of view. The same logic is also seen for mobile telephony (Armstrong, 2002; Wright, 2002), in which fixed-line callers are charged high fees to call mobile subscribers, who join a single mobile network and receive handset subsidies for doing so. As a result, there may well be too few calls made to each mobile subscriber. To counter this market failure, fixed-to-mobile termination charges are regulated in several countries.

Negative own-group externalities

In many examples of two-sided markets, agents not only like more agents from the other side, but they dislike more agents from the same side (for example, firms generally dislike advertising by rival firms in the same directory). Negative own-group externalities will justify higher prices, reflecting the 'pollution effect' of attracting additional agents. One case where negative own-group externalities arise

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
endogenously is when sellers join the platform to better compete for buyers on the other side. Consider, for example, the case of payment schemes that attract cardholders and merchants. To the extent merchants accept cards to attract customers from each other, their private willingness to accept cards includes the surplus their customers get from using cards. As a result, card schemes will charge merchants more and cardholders less (Wright, 2004). In addition, since the surplus of cardholders is over-represented in the profits of the card schemes, merchants will tend to be charged too much and cardholders too little.

Negative own-group externalities can also explain agreements to exclude rival agents from the same platform. For instance, a shopping mall may restrict the number of competing retailers (such as bookstores). If the platform finds it difficult (or costly) to recover revenue from the consumer side, this may be a way to drive up revenue from retailers. If a television channel cannot charge viewers, it may maximize profits by promising one car maker that it will not show an advert from a rival car maker in the same advertising slot. More generally, platforms act somewhat like regulators (Hagiu, 2005; Rochet and Tirole, 2006). They impose rules, conditions and prices for the platform that help solve various inefficiencies, at the cost perhaps of introducing others.

See Also

- credit card industry

Bibliography

Armstrong, M. 2002. The theory of access pricing and interconnection. In *Handbook of Telecommunications Economics*, ed. M. Cave, S. Majumdar and I. Vogelsang. Amsterdam: North-Holland.

Armstrong, M. 2006. Competition in two-sided markets. *RAND Journal of Economics* 37, 668–91.

Armstrong, M. and Wright, J. 2007. Two-sided markets, competitive bottlenecks and exclusive contracts. *Economic Theory* 32, 353–80.

Baxter, W.F. 1983. Bank interchange of transactional paper: legal perspectives. *Journal of Law and Economics* 26, 541–88.

Caillaud, B. and Jullien, B. 2001. Competing cybermediaries. *European Economic Review* 45, 797–808.

Caillaud, B. and Jullien, B. 2003. Chicken & egg: competition among intermediation service providers. *RAND Journal of Economics* 34, 309–28.

Hagiu, A. 2005. Two-sided platforms: pricing and social efficiency. Harvard Business School mimeo.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Rochet, J.-C. and Tirole, J. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1, 990–1029.

Rochet, J.-C. and Tirole, J. 2006. Two-sided markets: a progress report. *RAND Journal of Economics* 37, 645–67.

Schmalensee, R. 2002. Payment systems and interchange fees. *Journal of Industrial Economics* 50, 103–22.

Wright, J. 2002. Access pricing under competition: an application to cellular networks. *Journal of Industrial Economics* 50, 289–315.

Wright, J. 2004. Determinants of optimal interchange fees in payment systems. *Journal of Industrial Economics* 52, 1–26.

How to cite this article

Armstrong, M. and J. Wright. "two-sided markets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000201>
doi:10.1057/9780230226203.1751

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

two-stage least squares and the k-class estimator

N.E. Savin

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Two-stage least squares has been a widely used method of estimating the parameters of a single structural equation in a system of linear simultaneous equations. This article first considers the estimation of a full system of equations. This provides a context for understanding the place of two-stage least squares in simultaneous-equation estimation. The article concludes with some comments on the lasting contribution of the two-stage least squares approach and more generally the future of the identification and estimation of simultaneous-equations models.

Keywords

asymptotic distribution; Bayesian method-of-moments approach; Cowles Foundation; full and limited information methods; full-information maximum likelihood; generalized least squares; generalized method of moments; heteroskedasticity and autocorrelation; homoskedasticity; identification; indirect least squares; instrumental variables; *k*-class estimators; limited information maximum likelihood; linear models; maximum likelihood; ordinary least squares; reduced-form equations; simultaneous equations models; structural parameters; two-stage least squares (2SLS); two-stage least squares estimator and the *k*-class estimator; vector autoregressions

Article

Two-stage least squares (2SLS) was originally proposed as a method of estimating the parameters of a single structural equation in a system of linear simultaneous equations. It was introduced more or less independently by Theil (1953a; 1953b; 1961), Basmann (1957) and Sargan (1958). The early work on simultaneous equations estimation was carried out by a group of econometricians at the Cowles Foundation. This work was based on the method of maximum likelihood. In particular, Anderson and Rubin (1949; 1950) developed the limited information maximum likelihood (LIML) estimator for the parameters of a single structural equation. Anderson (2005) gives the history of 2SLS a revisionist twist

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

by pointing out that Anderson and Rubin (1950) indirectly includes the 2SLS estimator and its asymptotic distribution. The notation of that paper is difficult and the exposition is somewhat obscure, which may explain why few econometricians are aware of its contents. See Farebrother (1999) for additional insights into the precursors of 2SLS.

2SLS was by far the most widely used method in the 1960s and the early 1970s. The explanation involves both the state of statistical knowledge among applied econometricians and the state of computer technology. The classic treatment of maximum likelihood methods of estimation is presented in two Cowles Commission monographs: Koopmans (1950), *Statistical Inference in Dynamic Economic Models*, and Hood and Koopmans (1953), *Studies in Econometric Method*, which was directed at a wider audience. Among applied econometricians, relatively few had the statistical training to master the papers in these monographs, especially Koopmans (1950). By the end of the 1950s computer programs for ordinary least squares were available. These programs were simpler to use and much less costly to run than the programs for calculating LIML estimates. Owing to advances in computer technology, and, perhaps, also the statistical background of applied econometricians, the popularity of 2SLS started to wane towards the end of the 1970s. In particular, the difficulty of calculating LIML estimates was no longer an important constraint.

This article first considers the estimation of a full system of equations and then focuses on 2SLS. This approach provides a context for understanding the place of 2SLS in simultaneous-equation estimation. The article is organized as follows. A two-equation structural form model with normal errors and no lagged dependent variables is introduced in section 1. Section 2 reviews the properties of the ordinary least squares estimator of the parameters of a structural equation. The indirect least squares estimator is introduced in section 3. In section 4 presents the indirect feasible generalized least squares estimator, and briefly discusses maximum likelihood methods. Section 5 develops two rationales for the 2SLS procedure, and the *k*-class family of estimators is defined in section 6. Finite sample results on the comparisons of estimators are reported in section 7, and the concluding comments are in section 8. (Our exposition of structural-form estimation draws heavily on the treatment by Goldberger, 1991. For the presentation of GMM and more recent methods of simulation-equation estimation, see Mittelhammer, Judge and Miller, 2000.)

1 The model

In the spirit of Goldberger (1991), we consider a two-equation demand and supply model to fix ideas and notation. The endogenous variables are y_1 (quantity) and y_2 (price), the exogenous variable is x (income), and the disturbances are u_1 (demand shock) and u_2 (supply shock). For convenience the intercepts are suppressed in both equations.The *structural form* of the model is

$$\text{Demand}\quad y_1 = \alpha_1 y_2 + \alpha_2 x + u_1, \tag{1.1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\text{Supply}\ y_2 = \alpha_3 y_1 + u_2. \tag{1.2}$$

With the terms in y_1 and y_2 transferred to the left-hand side, the matrix representation of structural form is

$$(y_1, y_2)\begin{pmatrix} 1 & -\alpha_3 \\ -\alpha_1 & 1 \end{pmatrix} = x(\alpha_2, 0) + (u_1, u_2),$$

or

$$\mathbf{y}'\mathbf{\Gamma} = \mathbf{x}'\mathbf{B} + \mathbf{u}'$$

In the structural-form coefficient matrices **Γ** and **B**, the columns refer to equations, while the rows refer to variables.
Each endogenous variable can be solved for in terms of the exogenous variables and structural shocks to get the *reduced form* of the model:

$$\text{Quantity}.\ y_1 = \pi_{11}x + v_1, \tag{1.3}$$

$$\text{Price}.\ y_2 = \pi_{12}x + v_2. \tag{1.4}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In matrix form,

$$(y_1, y_2) = x(\pi_{11}, \pi_{12}) + (v_1, v_2),$$

or

$$\mathbf{y}' = \mathbf{x}'\mathbf{B}\Gamma^{-1} + \mathbf{u}'\Gamma^{-1} = \mathbf{x}'\Pi + \mathbf{v}'$$

The reduced form is derived by post-multiplying the structural form by Γ^{-1}, where $\Pi = \mathbf{B}\Gamma^{-1}$ is the reduced-form coefficient matrix and $\mathbf{v}' = \mathbf{u}'\Gamma^{-1}$ is the reduced-form disturbance vector.
Next we consider the statistical specification of a linear simultaneous-equation model for the general case of a $m \times 1$ endogenous-variable vector **y**, the $k \times 1$ exogenous-variable vector **x** and the $m \times 1$ structural-disturbance vector **u**. The specification is the following:

$$\mathbf{y}'\Gamma = \mathbf{x}'\mathbf{B} + \mathbf{u}', \tag{A1}$$

$$\Gamma \text{ nonsingular}, \tag{A2}$$

$$E(\mathbf{u}|\mathbf{x}) = \mathbf{0}, \tag{A3}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$V(\mathbf{u}|\mathbf{x}) = \Sigma \text{positive definite}. \tag{A4}$$

Here $\mathbf{\Gamma}$ is $m \times m$, $\mathbf{B}$ is $k \times m$, $\mathbf{\Sigma}$ is $m \times m$. Assumption (A1) gives the system of *m* structural equations in *m* endogenous variables. Assumption (A2) says that the system is complete in the sense that **y** is uniquely determined by **x** and **u**. (A3) says that **x** is exogenous in the sense that the conditional expectation of **u** given **x** is zero for all values of **x**. Assumption (A4) is a homoskedasticity requirement, and positive definiteness rules out exact linear dependency among the structural disturbances.
The implications of the specification (A1)–(A4) are the following:

$$\mathbf{y}' = \mathbf{x}'\mathbf{\Pi} + \mathbf{v}', \mathbf{v}' = \mathbf{u}'\mathbf{\Gamma}^{-1}, \tag{B1}$$

$$E(\mathbf{v}|\mathbf{x}) = \mathbf{0}, \tag{B2}$$

$$V(\mathbf{v}|\mathbf{x}) = (\mathbf{\Gamma}^{-1})\mathbf{\Sigma}\mathbf{\Gamma}^{-1} = \mathbf{\Omega} \text{positive definite}. \tag{B3}$$

The reduced-form disturbance vector **v** is mean-independent of, and homoskedastic with respect to, the exogenous variable vector **x**.
Next we turn from the population to the sample. We suppose that a sample of *n* observations from the multivariate distribution of **x** and **y** is obtained by stratified sampling: *n* values of **x** are selected, forming the rows of the $n \times k$ observed matrix **X** with rank $(\mathbf{X}) = k$. For each observation, a random drawing is made from the relevant conditional distribution of **y** given **x**, giving the rows of the $n \times m$ observed matrix **Y**, where the successive drawings are independent. The statements about asymptotic properties of the estimators rely on the additional assumption that the matrix $\mathbf{X}'\mathbf{X} / n$ has a positive definite limit. If

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
instead sampling is random from the joint distribution of **x** and **y**, there is no substantial change in the results.

2 Ordinary least squares

In simultaneous equations models, the parameters of interest are the structural parameter, the $\alpha's$ in the demand-supply example and the elements of **Γ**and **B** and in general case, rather than the reduced form parameters, the $\pi's$ or **Π**. Ordinary least squares (OLS) estimation of the structural parameters is not appropriate because the structural parameters are not coefficients of the conditional expectation functions among the observable variables. We now illustrate this point for the supply equation of the demand and supply model.
The reduced form of the demand and supply model expressed explicitly in terms of the structural parameters:

$$y_1 = (\alpha_2 x + \alpha_1 u_2 + u_1) / (1 - \alpha_1 \alpha_3), \tag{2.1}$$

$$y_2 = (\alpha_2 \alpha_3 x + \alpha_3 u_1 + u_2) / (1 - \alpha_1 \alpha_3). \tag{2.2}$$

For convenience, suppose that x, u_1 and u_2 are trivariate-normally distributed with zero means, variances $\sigma_x^2, \sigma_1^2, \sigma_2^2,$ and zero correlations. Then y_2 and y_1 are bivariate normal, so the conditional expectation of y_2 given y_1 is

$$E(y_2 | y_1) = \alpha^* y_1$$

with

$$\alpha^* = C(y_1, y_2) / V(y_1).$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

If $\alpha^* = \alpha_3$, then the sample least squares regression of y_2 on y_1 will provide a unbiased minimum variance estimator of α_3. If $\alpha^* \neq \alpha_3$, then least squares is not appropriate for the estimation of α_3. From equations (2.1) and (2.2) we calculate

$$C(y_1, y_2) = (\alpha_2\alpha_3\sigma_x^2 + \alpha_3\sigma_1^2 + \alpha_1\sigma_2^2) / (1 - \alpha_1\alpha_3)^2, V(y_1) = (\alpha_2^2\sigma_x^2 + \sigma_1^2 + \alpha_1^2\sigma_2^2) / (1 - \alpha_1\alpha_3)^2.$$

Let $\theta = \alpha_2^2\sigma_x^2 + \sigma_1^2$. Then

$$\alpha^* = \alpha_3 \frac{\theta}{(\theta + \alpha_1^2\sigma_2^2)} + \frac{\alpha_1\sigma_2^2}{\theta + \alpha_1^2\sigma_2^2}.$$

Clearly the parameter of interest α_3 is not the slope of the conditional expectation function of y_2 given y_1. This result is usually described by saying that OLS gives a biased estimator of the structural parameter α_3. Another description is that OLS gives a unbiased estimator of slope of the conditional expectation function, which happens to differ from the slope of the structural equation. Observe that $\alpha^* = \alpha_3$ in the special case with $\alpha_1 = 0$; in this case, y_1 is a function of x and u_1 only so that $E(u_2|y_1) = 0$.

The problem with OLS can be illustrated without relying on normality. From (1.2) we get

$$E(y_2|y_1) = \alpha_3 y_1 + E(u_2|y_1).$$

From eq. (2.2),

$$C(y_1, u_2) = C(\alpha_2 x + \alpha_1 u_2 + u_1, u_2) / (1 - \alpha_1\alpha_3) = \alpha_1\sigma_2^2 / (1 - \alpha_1\alpha_3).$$

Because y_1 and u_2 are correlated, we see that $E(u_2|y_1) \neq E(u_2) = 0$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

3 Indirect least squares

The next method we consider uses OLS to estimate the reduced-form parameters, and then converts the OLS reduced-form estimates into estimates of the structural-form parameters. This method, called 'indirect least squares' (ILS), produces estimates that are consistent, although not unbiased. Koopmans and Hood (1953) attribute ILS to M.A. Girshick. Again see Farebrother (1999) for precursors.

The key to ILS is the relation that relates the reduced-form coefficients to the structural-form coefficients, namely, $\mathbf{\Pi} = \mathbf{B}\mathbf{\Gamma}^{-1}$, which can be rewritten as $\mathbf{\Pi}\mathbf{\Gamma} = \mathbf{B}$. Suppose $\mathbf{\Pi}$ is known along with the prior knowledge that certain elements of $\mathbf{\Gamma}$ and $\mathbf{B}$ are zero. The question is whether we can solve $\mathbf{\Pi}\mathbf{\Gamma} = \mathbf{B}$ uniquely for the remaining unknown elements of $\mathbf{\Gamma}$ and $\mathbf{B}$. When a structural parameter is uniquely determined, we say that the parameter is *identified in terms of* $\mathbf{\Pi}$ or, more simply, that is identified. This suggests that the identified structural-form parameters may be estimated via OLS estimates of the reduced-form coefficients.

The relation between reduced-form and structural coefficients for the demand and supply model is the following:

$$(\pi_{11} \quad \pi_{12})\begin{pmatrix} 1 & -\alpha_3 \\ -\alpha_1 & 1 \end{pmatrix} = (\alpha_2 \quad 0).$$

There are two equations in three unknowns:

$$\pi_{11} - \alpha_1\pi_{12} = \alpha_2\pi_{12} - \alpha_3\pi_{11} = 0. \tag{3.1}$$

On the right-hand-side of (3.1), solve the equation for $\alpha_3 = \pi_{12} / \pi_{11}$. We conclude that the slope coefficient of the supply equation is identified. With respect to estimation, the ILS estimate of α_3 is obtained by replacing π_{11} and π_{12} by their OLS counterparts.

The ILS estimator of α_3 is consistent since the equation-by-equation OLS estimators of π_{11} and π_{12} are consistent. Moreover, the equation-by-equation OLS estimates are the same as the generalized least squares (GLS) estimates, that is, the OLS and GLS estimates coincide in every sample. This is because the explanatory variables are identical in the two reduced-form equations. A consequence is that the ILS estimator is asymptotically efficient.

4 Indirect feasible generalized least squares

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

For some simultaneous-equation models, prior knowledge that certain elements of **Γ** and **B** are zero implies restrictions on **Π**. In this situation, equation-by-equation OLS estimates of the π's are not optimal, and ILS does not yield a unique estimate of the structural parameters. We now illustrate the case with restrictions on **Π** using a modification of the original structural model.

The modified model has three exogenous variables, x_1 (income), x_2 (wage rate) and x_3 (interest rate). The modification consists of allowing the three exogenous variables to enter the supply equation:

$$\text{Demand}: \; y_1 = \alpha_1 y_2 + \alpha_2 x_1 + u_1, \tag{4.1}$$

$$\text{Supply}: \; y_2 = \alpha_3 y_1 + \alpha_4 x_1 + \alpha_5 x_2 + \alpha_6 x_3 + u_2. \tag{4.2}$$

The reduced-form of the modified structural-form system is

$$\text{Quantity}: \; y_1 = \pi_{11} x_1 + \pi_{21} x_2 + \pi_{31} x_3 + v_1, \tag{4.3}$$

$$\text{Price}: \; y_2 = \pi_{12} x_1 + \pi_{22} x_2 + \pi_{32} x_3 + v_2 \tag{4.4}$$

In the **ΠΓ** = **B** format, the relation between the reduced-form and structural coefficients is:

$$\begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \end{pmatrix} \begin{pmatrix} 1 & -\alpha_3 \\ -\alpha_1 & 1 \end{pmatrix} = \begin{pmatrix} \alpha_2 & \alpha_4 \\ 0 & \alpha_5 \\ 0 & \alpha_6 \end{pmatrix}.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

There are now six equations in six unknowns:

$$\begin{aligned} \pi_{11} - \alpha_1\pi_{12} &= \alpha_2 & \pi_{12} - \alpha_3\pi_{11} &= \alpha_4 \\ \pi_{21} - \alpha_1\pi_{22} &= 0 & \pi_{22} - \alpha_3\pi_{21} &= \alpha_5 \\ \pi_{31} - \alpha_1\pi_{32} &= 0 & \pi_{32} - \alpha_3\pi_{31} &= \alpha_6 . \end{aligned} \tag{4.5}$$

The system on the left of (4.5) determines the parameters of the demand equation. Solve either of the equations that has 0 on its right-hand side for $\alpha_1 = \pi_{31} / \pi_{32} = \pi_{21} / \pi_{22}$, and then get α_2 from the remaining equation. Clearly, the coefficients of the demand equation are identified in terms of **Π**. Furthermore, there is a restriction on the π's, namely $\pi_{31} / \pi_{32} = \pi_{21} / \pi_{22}$, because on the left of (4.5) there are three equations in two unknowns.

The system on the right-hand side of (4.5), which refers to the supply equation, consists of three equations in four unknowns. Once a value is assigned to α_3, the equations can be solved for $\alpha_4, \alpha_5, \alpha_6$. A different arbitrary value for α_3 generates different values for $\alpha_4, \alpha_5, \alpha_6$. The solution is not unique. Hence, the coefficients of the supply equation are not identified in terms of **Π**.

With respect to estimation, ILS using the equation-by-equation OLS estimates of **Π** will not give unique estimates of the structural parameters of the supply equation. The result is two different ILS estimates of α_1. This problem can be overcome by estimating the reduced-form subject to the restriction $\pi_{31} / \pi_{32} = \pi_{21} / \pi_{22}$. The restricted estimates of the π ′ *s* can be converted into unique estimates of the α's using the sample counterpart of the system (4.5).

Suppose there are restrictions on **Π**. Then the fact that the explanatory variables are identical in every reduced-form equation does not imply that the OLS and GLS estimates of the π's are the same. In other words, OLS estimation of the reduced form will not be optimal. If the variance matrix of the reduced-form disturbance vector **Ω** is known, then GLS subject to the restrictions on **Π** is the natural (nonlinear) estimation procedure. The conversion of the GLS estimates of the π's into estimates of the α's can be described as ‘indirect GLS’. Since the GLS estimator is consistent and asymptotically efficient, the indirect-GLS estimators of **Γ** and **B** are also consistent and asymptotically efficient.

When **Ω** is unknown, as is true in practice, feasible GLS is the natural estimation procedure for **Π**. Feasible GLS is similar to GLS except that an estimator $\widehat{\Omega}$ is used in place of **Ω**. The estimator $\widehat{\Omega}$ comes from the residuals of the equation-by-equation OLS reduced-form regressions. The resulting estimates of the α's are referred to as ‘indirect-FGLS’ estimates because the FGLS estimates of the π's are converted into estimates of the α's. Because the FGLS estimator of **Π** is consistent and asymptotically efficient, the indirect-FGLS estimators of **Γ** and **B** are also consistent and asymptotically efficient.

Indirect GLS and indirect FGLS are referred to as ‘full-information’ methods because they use all the restrictions on **Π** at once. Estimation of a single structural equation using only the restrictions on **Π** for that equation alone is often called ‘limited information’ estimation. If all the restrictions are correctly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

specified, then full-information estimators are more efficient than limited-information estimators.

In some variants of the simultaneous-equation model it is assumed that $\mathbf{u}|\mathbf{x}$ is multivariate normal. The addition of the normality assumption enables the estimation of Π by maximum likelihood. The resulting estimator of the structural parameters is known as 'full-information maximum likelihood', or FIML. If Ω is known, then FIML coincides with indirect-GLS. If Ω is unknown, FIML differs from indirect-FGLS, but the estimators have the same asymptotic distribution.

The difference between FIML and indirect FGLS can be clarified by briefly turning from the population to the sample. Let $\widehat{\mathbf{V}} = \mathbf{Y} - \mathbf{XP}$, where $\mathbf{P} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the estimator of Π obtained by equation-by-equation OLS. The estimator of Ω used in FGLS is $\widehat{\Omega} = (1/n)\widehat{\mathbf{V}}'\widehat{\mathbf{V}}$. The criterion minimized by FGLS is $\mathrm{tr}(\widehat{\Omega}^{-1}\mathbf{VV})$, where $\mathbf{V} = \mathbf{Y} - \mathbf{X}\Pi$. FIML proceeds by inserting $\widehat{\Omega} = (1/n)\widehat{\mathbf{V}}'\widehat{\mathbf{V}}$ (as a conditional solution) into the log-likelihood function to obtain the log-likelihood concentrated on $|\mathbf{V}'\mathbf{V}|$. The consequence is that the criterion minimized by FIML is $|\mathbf{V}'\mathbf{V}|$. The difference in the criteria explains the difference in the estimators.

The maximum likelihood estimation of a single structural-form equation that uses only the restrictions on Π for that equation alone is referred to as 'limited-information maximum likelihood', or LIML. We next consider another limited-information estimation method.

5 Two-stage least squares

The 2SLS estimator uses the unrestricted reduced-form estimate **P**, the equation-by-equation OLS estimates of the π's, which accounts for its popularity. The mechanics of the 2SLS method can be described simply. In the first stage, the right-hand-side endogenous variables of the structural equation are regressed on all the exogenous variables in the reduced form, and the fitted values are obtained. In the second stage, the right-hand-side endogenous variables are replaced by their fitted values, and the left-hand-side endogenous variable of the equation is regressed on the right-hand-side fitted values and the exogenous variables included in the equation.

Two rationales for the 2SLS procedure are now developed using the demand equation of the modified structural model. The starting point for the first rationale is the expectation of the demand equation conditional on x_1, x_2, and x_3. Taking expectations gives

$$E(y_1|x_1, x_2, x_3) = \alpha_1 E(y_2|x_1, x_2, x_3) + \alpha_2 x_1 + E(u_1|x_1, x_2, x_3),$$

or

$$E(y_1|x_1, x_2, x_3) = \alpha_1 y_2^* + \alpha_2 x_1.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

From the reduced-form eq. (4.4), $y_2^* = E(y_2|x_1, x_2, x_3) = \pi_{12}x_1 + \pi_{22}x_2 + \pi_{32}x_3$. Because y_2^* is linear function of the exogenous variables, it is exogenous. If y_2^* were observed, then y_1 could be regressed on y_2^* and x_1 to get unbiased estimates of α_1 and α_2. But y_2^* is unobservable because π_{12}, π_{22} and π_{32} are unknown. However, an unbiased and consistent estimate $\hat{y}_2$ can be obtained by replacing the unknown π's by their OLS estimates. Then, making the replacement of $\hat{y}_2$ for y_2^* produces consistent estimates of the structural parameters.
The second rationale exploits the fact that in the population the following moment conditions hold:

$$E(u_1) = 0, \quad C(x_1, u_1) = C(x_2, u_1) = C(x_3, u_1) = 0.$$

These imply two orthogonality conditions:

$$E(y_2^* u_1) = 0, \quad E(x_1 u_1) = 0.$$

If we let $u_1 = y_1 - (a_1 y_2^* + a_2 x_1)$, then α_1 and α_2 are the values for a_1 and a_2 that make $E(y_2^* u_1) = 0$ and $E(x_1 u_1) = 0$. 2SLS chooses the estimates that make the analogous sample quantities zero, that is, $\Sigma_i \hat{y}_{2i} u_{1i} = 0$ and $\Sigma_i x_{1i} u_{1i} = 0$, $(i = 1, \ldots, n)$. This illustrates that 2SLS has an *instrumental-variable* (IV) interpretation.
The IV interpretation can be illustrated more explicitly by writing the demand equation in terms of the observations for a sample size of n:

$$\mathbf{y_1} = \mathbf{y_2}\alpha_1 + \mathbf{x_1}\alpha_2 + \mathbf{u_1} = (\mathbf{y_2} \;\; \mathbf{x_1}) \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \mathbf{u_1} = \mathbf{Z_1}\boldsymbol{\alpha} + \mathbf{u_1},$$

where in the context of the demand equation $\mathbf{y_1}$ and $\mathbf{y_2}$ are the columns of $\mathbf{Y}$ and $\mathbf{x_1}$ is the first column of $\mathbf{X}$. As we have shown, regressing $\mathbf{y_1}$ on $\mathbf{Z_1}$ will not give a consistent estimator for $\boldsymbol{\alpha_1}$. Instead replace $\mathbf{Z_1}$ by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\hat{Z}_1 = NZ_1 = N(y_2, x_1) = (Ny_2, Nx_1) = (\hat{y}_2, x_1),$$

where **N** is the idempotent matrix $X(X'X^{-1})X'$. Regressing y_1 on $\hat{Z}_1$ gives the normal equations,

$$\hat{Z}_1'\hat{Z}_1 a_1^* = \hat{Z}_1' y_1,$$

the solution to which is the 2SLS estimator.

The 2SLS normal equations are equivalent to a set of orthogonality conditions: $\hat{Z}_1' u_1 = 0$, where $u_1 = y_1 - Z_1 a_1^*$. The equivalence follows from an algebraic fact:

$$\hat{Z}_1' Z_1 = Z_1' N' Z_1 = Z_1' N' N Z_1 = \hat{Z}_1' \hat{Z}_1.$$

The variables in $\hat{Z}_1$ are legitimate instruments because they are, at least asymptotically, uncorrelated with the disturbance. The IV interpretation implies that the 2SLS estimator is consistent. In fact, it is the optimal feasible IV estimator.

We also note that the 2SLS estimator can be interpreted as a general-method-of-moments (GMM) estimator. In the above example, this follows from the fact that a_1^* minimizes the quadratic form

$$(y_1 - Z_1 a_1^*)' X(X'X)^{-1} X'(y_1 - Z_1 a_1^*).$$

It can be shown that 2SLS is the optimal feasible GMM estimator. An advantage of the GMM approach is that heteroskedasticity and autocorrelation can be accommodated by an appropriate redefinition of the optimal weighting matrix in the definition of the GMM estimator (see Ruud, 2000, pp. 718–21)

We conclude this section with some remarks on estimation in the simultaneous-equations model.

1. 1. If all the structural equations are identified, and there are no restrictions on **Π**, then ILS, indirect-FGLS, FIML, LIML and 2SLS all produce the same estimates.
2. 2. If there are restrictions on **Π**, then LIML and 2SLS produce different estimates in the sample,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

but the estimators have the same asymptotic distribution, and similarly for indirect FGLS and FIML.

3. 3. If a parameter is not identified, then there is no method to estimate it consistently.
4. 4. We have confined our attention to the case in which the prior information used for identification consists of normalizations and exclusions (zero restrictions). If other information is available (for example, Σ is diagonal, or a coefficient in one structural equation is equal to a coefficient in another), then some modifications are needed in the description of the estimators and their statistical properties.

6 The k-class family

The *k*-class family of estimators of the coefficients of a single structural equation is illustrated for the demand equation of the modified structural model. For this equation, the estimator is

$$\mathbf{a}_k^* = \left(\mathbf{Z}_1'(\mathbf{I} - k\mathbf{M})\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'(\mathbf{I} - k\mathbf{M})\mathbf{y}_1,$$

where $\mathbf{M} = \mathbf{I} - \mathbf{N}$. This family was introduced by Theil (1953b; 1961). It includes the OLS estimator $(k = 0)$ and the 2SLS estimator $(k = 1)$.
A remarkable fact is that the *k*-class family includes LIML. The LIML estimator is obtained by setting $k = \lambda^*$, where λ^* the smallest root of

$$|\mathbf{W}_1 - \lambda\mathbf{W}| = 0.$$

In the determinantal equation, $\mathbf{W}_1$ is the cross-product matrix of residuals from the OLS regression of $(\mathbf{y}_1, \mathbf{Y}_2)$ on $\mathbf{x}_1$ (the included endogenous variables on the included exogenous variable), and $\mathbf{W}$ is the cross-product matrix of the residuals from the OLS regression of $(\mathbf{y}_1, \mathbf{Y}_2)$ on $\mathbf{X}$ (the included endogenous on all the exogenous variables). Moreover, the LIML estimator of α_1 is the value of a_1 that minimizes the variance ratio:

$$\frac{(1, -a_1)\mathbf{W}_1(1, -a_1)'}{(1, -a_1)\mathbf{W}(1, -a_1)'},$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Accordingly, the LIML estimator is also sometimes referred to as the 'least-variance ratio' estimator. The *k*-class estimator is consistent if $(k-1)$ converges in probability to 0 and has the same limiting distribution as 2SLS if $n^{1/2}(k-1)$ converges in probability to 0. These conditions are clearly not satisfied when $k = 0$ and hence by OLS. A proof that these conditions are satisfied for LIML is given in Amemiya (1985, pp. 237–8). Zellner (1998) shows that a Bayesian method-of-moments approach justifies certain other members of the *k*-class (and double *k*-class) family of estimators.

7 Finite sample distributions

There has been debate about whether the LIML estimator is better than 2SLS. The reason for this debate is that the finite sample distributions of the estimators differ when there are restrictions on Π. Hence we now limit our attention to the case where restrictions are present.

A key difference between the estimators is the existence of moments. The 2SLS estimator has moments up to certain order. By contrast, the LIML estimator has no moments. This result holds for an arbitrary number of included endogenous variables. Mariano (2001) reviews the moment existence results.

In addition to the moment existence results, closed form expressions for the moments and probability densities of 2SLS, LIML and *k*-class estimators have been derived. These expressions are complicated; see Phillips (1983) for specific references.

The finite sample results have mostly come from the study of a model with two included endogenous variables. Moreover, it is often assumed that the disturbances are normal. The finite sample results are obtained using analytical and simulation methods. A survey of the results and their practical implications is given in Mariano (2001). One of the results is that the LIML distribution is far more symmetric than 2SLS, though more spread out, and it approaches normality faster. Similarly, Anderson (1982) concludes that for many cases that occur in practice the standard the normal theory is inadequate for 2SLS, but provides a fairly good approximation to the actual distribution of the LIML estimator. The symmetry result is not surprising because the approximate distribution of the LIML estimator (obtained from large sample asymptotic expansions) is median unbiased. Median unbiasedness does not hold in general for 2SLS.

It is helpful to put the debate over the relative merits of LIML and 2SLS in perspective. On the assumption that the model is correctly specified, the presumption is that maximum likelihood will have better properties in some well-defined sense. This is because it uses more information than GMM. See Anderson, Kunitomo and Morimume (1986) and Takeuchi and Morimume (1985) for results on the second-order efficiency of LIML. An advantage of GMM is that it is robust to misspecifications of the disturbance distribution, although this advantage was not the original motivation for introducing 2SLS.

8 Epilogue

Keynesian economics initially played a key role in propelling research into the estimation of simultaneous-equations models. Interest in the estimation of linear simultaneous-equations models began to wane from about the late 1970s. Historically, this decline paralleled the decline of the Keynesian paradigm as a result of the monetarist counter-revolution and later rational expectations. At

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the same time, there was growing awareness that linear simultaneous-equations models often suffered from potentially serious misspecification. On the one hand, even if one takes for granted the statistical specification of the model (A1–A4), economic theory did not provide a satisfactory basis for deciding what endogenous and exogenous variables should be excluded from individual structural equations. As a consequence, the identification of structural parameters was open to question. On the other hand, the statistical specification of the linear structural-form model was itself questionable, due to either the nature of the data or considerations of economic theory or both. The issue of identification was highlighted by Sims (1980) in a well-known paper aptly titled 'Macroeconomics and Reality'.

Simultaneous-equations models have been generalized by the introduction of nonlinear structural equations. Amemiya (1974) generalized the 2SLS method to nonlinear models. His nonlinear 2SLS estimator is a GMM estimator. However, unlike in the case of linear models, it cannot be thought of as being obtained in two steps where the first consists of running a least squares regression. The GMM is a two-step estimator, but the first step consists of choosing a weighting matrix. More generally, GMM and IV estimators can be thought of as the descendants of the 2SLS approach. They constitute the contemporary basis for much of the estimation of structural parameters macroeconomics. It is in this sense that 2SLS lives on as a structural estimation method.

Does the identification and estimation of simultaneous-equations models have a future? There are some positive signs. Although indirectly, these issues continue to play a role in the identified vector autoregression literature, for example, Bernanke and Blinder (1992), Gordon and Leeper (1995), Cushman and Zha (1997). More recently, there has been a revival of interest in simultaneous-equations models formulated as nonlinear dynamic stochastic general equilibrium models. This formulation has the advantage that the resulting models are viewed as more firmly anchored in economic theory. Estimation of such models presents serious challenges. Some examples where these challenges are addressed include DeJong, Ingram and Whiteman (2000) and Fernandez-Villaverde and Rubio-Ramirez (2005).

See Also

- seemingly unrelated regressions
- simultaneous equations models

Bibliography

Amemiya, T. 1974. The nonlinear two-stage least-stage estimator. *Journal of Econometrics* 2, 105–10.

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Anderson, T. 1982. Some recent developments on the distribution of single-equation estimators. In *Advances in Econometrics*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.

Anderson, T. 2005. Origins of the limited information maximum likelihood and two-stage least squares

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
estimators. *Journal of Econometrics* 127, 1–16.

Anderson, T., Kunitomo, N. and Morimune, K. 1986. Comparing single equation estimators in a simultaneous equation system. *Econometric Theory* 2, 1–32.

Anderson, T. and Rubin, H. 1949. Estimator of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20, 46–63.

Anderson, T. and Rubin, H. 1950. The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 21, 570–82.

Basmann, R. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25, 77–83.

Bernanke, B. and Blinder, A. 1992. The federal funds rate and the channels of monetary transmission. *American Economic Review* 82, 901–21.

Cushman, D. and Zha, T. 1997. Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary Economics* 39, 433–48.

DeJong, D., Ingram, B. and Whiteman, C. 2000. A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* 98, 203–23.

Farebrother, R. 1999. *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. New York: Springer.

Fernandez-Villaverde, J. and Rubio-Ramirez, J. 2005. Estimating dynamic equilibrium economies: linear versus nonlinear likelihood. *Journal of Applied Econometrics* 20, 891–910.

Goldberger, A. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Gordon, D. and Leeper, E. 1995. The dynamic impacts of monetary policy: an exercise in tentative identification. *Journal of Political Economy* 102, 1228–47.

Hood, W. and Koopmans, T. 1953. *Studies in Econometric Method*. Cowles Foundation Monograph 14. New Haven: Yale University Press.

Koopmans, T. 1950. *Statistical Inference in Dynamic Economic Models*. Cowles Commission Monograph 10. New York: John Wiley and Sons.

Koopmans, T. and Hood, W. 1953. The estimation of simultaneous linear economic relationships. In

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Studies in Econometric Method, ed. W. Hood and T. Koopmans. Cowles Foundation Monograph 14. New Haven: Yale University Press.

Mariano, R. 2001. Simultaneous equation model estimators: statistical properties and practical implications. In *A Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Mittelhammer, R., Judge, G. and Miller, D. J. 2000. *Econometric Foundations*. Cambridge: Cambridge University Press.

Phillips, P. 1983. Exact small sample theory in the simultaneous equation model. In *Handbook of Econometrics*, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Ruud, P. 2000. *An Introduction to Classical Econometric Theory*. Oxford: Oxford University Press.

Sargan, J. 1958. Estimation of economic relationships using instrumental variables. *Econometrica* 67, 557–86.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Takeuchi, K. and Morimune, K. 1985. Third-order efficiency of the extended maximum likelihood estimator in a simultaneous equation system. *Econometrica* 53, 177–200.

Theil, H. 1953a. Repeated least-squares applied to a complete equation systems. Mimeo. The Hague: Central Planning Bureau.

Theil, H. 1953b. Estimation and simultaneous correlation in complete equation systems. Mimeo. The Hague: Central Planning Bureau.

Theil, H. 1961. *Economic Forecasts and Policy*, 2nd edn. Amsterdam: North-Holland.

Zellner, A. 1998. The finite sample properties of simultaneous equations' estimates and estimators: Bayesian and non-Bayesian approaches. *Journal of Econometrics* 83, 185–212.

How to cite this article

Savin, N. E. "two-stage least squares and the k-class estimator." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000147> doi:10.1057/9780230226203.1752

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

uncertainty and general equilibrium

Mukul Majumdar and Roy Radner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews alternative approaches to incorporating uncertainty in Walrasian models. It begins with a sketch of the Arrow–Debreu model of complete markets. An extension of this framework allowing for economic agents to have different information about the environment is followed by a critique. When markets are incomplete and trades take place sequentially, several types of equilibrium concept arise according to the hypotheses we make about the way traders form their expectations. We present conditions for the existence of equilibria for two such equilibrium concepts, and discuss the possible failure to attain Paretian welfare optima.

Keywords

Arrow–Debreu model; bounded rationality; budget constraints; competitive equilibrium; conditional probability; consumption possibility set; equilibrium; existence of competitive equilibrium; expectation formation; expected utility hypothesis; general equilibrium; incomplete information; incomplete markets; indicative planning; inside information; limited liability; moral hazard; nonprice information; optimality of competitive equilibrium; Pareto efficiency; perfect foresight; production possibility set; pseudo-equilibrium; rational expectations; rational expectations equilibrium; risk; sequential trading; steady state; temporary (or momentary) equilibrium; uncertainty; Walras's Law

Article

One of the notable intellectual achievements of economic theory during the second half of the 20th century has been the rigorous elaboration of the Walras–Pareto theory of value; that is, the theory of the existence and optimality of competitive equilibrium. Although many economists and mathematicians contributed to this development, the resulting edifice owes so much to the pioneering and influential work of Arrow and Debreu that in this paper we shall refer to it as the ‘Arrow–Debreu theory’. (For comprehensive treatments, together with references to previous work, see Debreu, 1959; Arrow and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Hahn, 1971.)

The Arrow–Debreu theory was not originally put forward for the case of uncertainty, but an ingenious device introduced by Arrow (1953), and further elaborated by Debreu (1953), enabled the theory to be reinterpreted to cover the case of uncertainty about the availability of resources and about consumption and production possibilities. (See Debreu, 1959, ch. 7, for a unified treatment of time and uncertainty.)

Subsequent research has extended the Arrow–Debreu theory to take account of (*a*) differences in information available to different economic agents, and the 'production' of information, (*b*) the incompleteness of markets, and (*c*) the sequential nature of markets. The consideration of these complications has stimulated the developments of new concepts of equilibrium, two of which will be elaborated in this article under the headings: (*a*) equilibrium of plans, prices, and price expectations (EPPPE) and (*b*) rational expectations equilibrium (REE). The exploration of these features of real-world markets has also made possible a general-equilibrium analysis of money and securities markets, institutions about which the original Arrow–Debreu theory could provide only limited insights. It has also led to a better understanding of the limits to the ability of the 'invisible hand' in attaining a Pareto optimal allocation of resources.

Review of the Arrow – Debreu model of a complete market for present and future contingent delivery

In this section, we review the approach of Arrow (1953) and Debreu (1959) to incorporating uncertainty about the environment into a Walrasian model of competitive equilibrium. The basic idea is that commodities are to be distinguished, not only by their physical characteristics and by the location and dates of their availability and/or use, but *also by the environmental event in which they are made available and/or used*. For example, ice cream made available (at a particular location on a particular date) if the weather is hot may be considered to be a different commodity from the same kind of ice cream made available (at the same location and date) if the weather is cold. We are thus led to consider a list of 'commodities' that is greatly expanded by comparison with the corresponding case of certainty about the environment. The standard arguments of the theory of competitive equilibrium, applied to an economy with this expanded list of commodities, then require that we envisage a 'price' for each commodity, the resulting set of price ratios specifying the market rate of exchange between each pair of commodities.

Just what institutions could, or do, effect such exchanges is a matter of interpretation that is, strictly speaking, outside the model. We shall present one straightforward inpt, and then comment briefly on an alternative inpt.

First, however, it will be useful to give a more precise account of concepts of environment and event that we shall be employing. The description of the 'physical world' is decomposed into three sets of variables: (*a*) decision variables, which are controlled (chosen) by economic agents; (*b*) environmental variables, which are not controlled by any economic agent; and (*c*) all other variables, which are completely determined (possibly jointly) by decisions and environmental variables. A *state* of the environment is a complete specification (history) of the environmental variables from the beginning to the end of the economic system in qst. An *event* is a set of states; for example, the event 'the weather is hot in New York on 1 July 1970' is the set of all possible histories of the environment in which the temperature in New York during the day of 1 July 1970 reaches a high of at least (say) 75°F. Granted

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that we cannot know the future with certainty, at any given date, there will be a family of *elementary observable* (knowable) *events*, which can be represented by a partition of the set of all possible states (histories) into a family of mutually exclusive subsets. It is natural to assume that the partitions corresponding to successive dates are successively finer, which represents the accumulation of information about the environment.

We shall imagine that a ‘market’ is organized before the beginning of the physical history of the economic system. An elementary contract in this market will consist of the purchase (or sale) of some specified number of units of a specified commodity to be delivered at a specified location and date, if and only if a specified elementary event occurs. Payment for this purchase is to be made now (at the beginning), in ‘units of account’, at a specified price quoted for that commodity–location–date–event combination. Delivery of the commodity in more than one elementary event is obtained by combining a suitable set of elementary contracts. For example, if delivery of one quart of ice cream (at a specified location and date) in hot weather costs $1.50 (now) and delivery of one-quart in non-hot weather costs $1.10, then sure delivery of one quart (that is, whatever the weather) costs $1.50+$1.10=$2.60.

There are two groups of economic agents in the economy: *producers* and *consumers*. A producer chooses a production plan, which determines his input and/or output of each commodity at each date in each elementary event (we shall henceforth suppress explicit reference to location, it being understood that the location is specified in the term ‘commodity’). For a given list of prices, the present value of a production plan is the sum of the values of outputs minus the sum of the values of inputs. Each producer is characterized by a set of production plans that are (given the technological know-how) feasible for him: his production possibility set.

A consumer chooses a consumption plan, which specifies his consumption of each commodity at each date in each elementary event. Each consumer is characterized by: (*a*) a set of consumption plans that are (physically, psychologically, and so on) feasible for him: his consumption possibility set; (*b*) preferences among the alternative plans that are feasible for him; (*c*) his endowment of physical resources, that is, a specification of the quantity of each commodity, for example, labour, at each date in each event, with which he is exogenously endowed; and (*d*) his shares in each producer, that is, the fraction of the present value of each producer's production plan that will be credited to the consumer's account. (For any one producer, the sum of the consumers’ shares is unity.) For given prices and given production plans of all the producers, the present net worth of a consumer is the total value of his resources plus the total value of his shares of the present values of producers’ production plans.

An equilibrium of the economy is a list of prices, a set of production plans (one for each producer), and a set of consumption plans (one for each consumer), such that (*a*) each producer's plan has maximum present value in his production possibility set; (*b*) each consumer's plan maximizes his preferences within his consumption possibility set, subject to the additional (budget) constraint that the present cost of his consumption plan not exceed his present net worth; (*c*) for each commodity at each date in each elementary event, the total demand equals the total supply: that is, the total planned consumption equals the sum of the total resource endowments and the total planned net output (where inputs are counted as negative outputs).

Notice that (*a*) producers and consumers are ‘price takers’; (*b*) for given prices there is no uncertainty about the present value of a production plan or of given resource endowments, nor about the present cost of a consumption plan; (*c*) therefore, for given prices and given producers’ plans, there is no uncertainty about a given consumer's present net worth; (*d*) since a consumption plan may specify that, for a given

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
commodity at a given date, the quantity consumed is to vary according to the event that actually occurs, a consumer's preferences among plans will reflect not only his ‘taste’ but also his subjective beliefs about the likelihoods of different events and his attitude towards risk (Savage, 1954).

It follows that beliefs and attitudes towards risk play no role in the assumed behaviour of producers. On the other hand, beliefs and attitudes towards risk do play a role in the assumed behaviour of consumers, although for given prices and production plans each consumer knows his (single) budget constraint with certainty.

We shall call the model just described an ‘Arrow–Debreu’ economy. One can demonstrate, under ‘standard conditions’: (*a*) the existence of an equilibrium, (*b*) the Pareto optimality of an equilibrium, and (*c*) that, roughly speaking, every Pareto optimal choice of production and consumption plans is an equilibrium relative to some price system for some distribution of resource endowments and shares (Debreu, 1959). In another direction of research initiated by Debreu (1970), the focus was to identify properties (like local uniqueness, finiteness) of Walrasian equilibria that were *generic* (typical or robust in a given context or in a class models). In what follows we shall use the term ‘generic’ informally and invite the reader to verify the exact definition from the original reference.

In the above interpretation of the Arrow–Debreu economy, all accounts are settled before the history of the economy begins, and there is no incentive to revise plans, reopen the markets or trade in shares. There is an alternative inpt, which will be of interest in connection with the rest of this article, but which corresponds to exactly the same formal model. In this second inpt, there is a single commodity at each date – let us call it ‘gold’ – that is taken as a numeraire at that date. A ‘price system’ has two parts: (1) for each date and each elementary event at that date, there is a price, to be paid in gold at the beginning date, for one unit of gold to be delivered at the specified date and event; (2) for each commodity, date, and event at that date, a price, to be paid in gold at that date and event, for one unit of the commodity to be delivered at that same date and event. The first part of the price system can be interpreted as ‘insurance premiums’ and the second part as ‘spot prices’ at the given date and event. The insurance interpretation is to be made with some reservation, however, since there is no real object being insured and no limit to the amount of insurance that an individual may take out against the occurrence of a given event. For this reason, the first part of the price system might be better interpreted as reflecting a combination of betting odds and interest rates.

Although the second part of the price system might be interpreted as spot prices, it would be a mistake to think of the determination of the equilibrium values of these prices as being deferred in real time to the dates to which they refer. The definition of equilibrium requires that the agents have access to the complete system of prices when choosing their plans. In effect, this requires that at the beginning of time all agents have available a (common) forecast of the equilibrium spot prices that will prevail at every future date and event.

Extension of the Arrow – Debreu model to the case in which different agents have different information

In an Arrow–Debreu economy, at any one date each agent may have incomplete information about the state of the environment, but all the agents will have the *same* information. This last assumption is not tenable if we are to take good account of the effects of uncertainty in an economy. We shall now sketch how, by a simple reinpt of the concepts of production possibility set and consumption possibility set, we can extend the theory of the Arrow–Debreu economy to allow for differences in information among the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economic agents.
For each date, the information that will be available to a given agent at that date may be characterized by a partition of the set of states of the environment. To be consistent with our previous terminology, we should assume that each such information partition must be at least as coarse as the partition that describes the elementary events at that date; that is, each set in the information partition must contain a set in the elementary event partition for the same date. For example, each set in the elementary event partition at a given date might specify the high temperature at that date, whereas each set in a given agent's information partition might specify only whether this temperature was higher than 75°F, or not.
An agent's information restricts his set of feasible plans in the following manner. Suppose that at a given date the agent knows only that the state of the environment lies in a specified set A (one of the sets in his information partition at that date), and suppose (as would be typical) that the set A contains several of the elementary events that are in principle observable at that date. Then any action that the agent takes at that date must necessarily be the same for all elementary events in the set A. In particular, if the agent is a consumer, then his consumption of any specified commodity must be the same in all elementary events contained in the information set A; if the agent is a producer, then his input or output of any specified commodity must be the same for all events in A. (We are assuming that consumers know what they consume and producers what they produce at any given date.)
Let us call the sequence of information partitions for a given agent his information structure and let us say that this structure is fixed if it is given independent of the actions of himself or any other agent. Furthermore, in the case of a fixed information structure, let us say that a given plan (consumption or production) is compatible with that structure if it satisfies the conditions described in the previous paragraph, at each date.
Suppose that consumption and production possibility sets of the Arrow–Debreu economy are interpreted as characterizing, for each agent, those plans that would be feasible if he had ‘full information’ (that is, if his information partition at each date coincided with the elementary event partition at that date). The set of feasible plans for any agent with a fixed information structure can then be obtained by restricting him to those plans in the full information possibility set that are also compatible with his given information structure.
From this point on, all of the machinery of the Arrow–Debreu economy (with some minor technical modifications) can be brought to bear on the present model. In particular, we get a theory of existence and optimality of competitive equilibrium relative to fixed structures of information for the economic agents. We shall call this the ‘extended Arrow–Debreu economy’. *We should add that differences among information structures of the agents may lead to a significant reduction of the number of active markets*. (For a fuller treatment, see Radner, 1968; 1982.)

Choice of information

There is no difficulty in principle in incorporating the choice of information structure into the extended Arrow–Debreu economy. We doubt, however, that it is reasonable to assume that the technological conditions for the acquisition and use of information generally satisfy the hypotheses of the standard theorems on the existence and optimality of competitive equilibrium.
The acquisition and use of information about the environment typically require the expenditure of goods and services, that is, of commodities. If one production plan requires more information for its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
implementation than another (that is, requires a finer information partition at one or more dates), then the list of (commodity) inputs should reflect the increased inputs for information. In this manner a set of feasible production plans can reflect the possibility of choice among alternative information structures. Unfortunately, the acquisition of information often involves a ‘set-up cost’, that is, the resources needed to obtain the information may be independent of the scale of the production process in which the information is used. This set-up cost will introduce a non-convexity in the production possibility set, and thus one of the standard conditions in the theory of the Arrow–Debreu economy will not be satisfied (Radner, 1968).

Even without set-up costs, *there is a general tendency for the value of information to exhibit ‘increasing returns’, at least at low levels*, provided that the structure of information varies smoothly with its cost. This striking phenomenon leads to discontinuities in the demand for information. (For a precise statement, see Radner and Stiglitz, 1984).

Critique of the extended Arrow – Debreu economy

If the Arrow–Debreu model is given a literal inpt, then it clearly requires that the economic agents possess capabilities of imagination and calculation that exceed reality by many orders of magnitude. Related to this is the observation that the theory requires in principle a complete system of insurance and futures markets, which system appears to be too complex, detailed, and refined to have practical significance. A further obstacle to the achievement of a complete insurance market is the phenomenon of ‘moral hazard’ (Arrow, 1965).

A second line of criticism is that the theory does not take account of at least three important institutional features of modern capitalist economies: money, the stock market, and active markets at every date. These two lines of criticism have an important connection, which suggests how the Arrow–Debreu theory might be improved. If, as in the Arrow–Debreu model, each production plan has a sure unambiguous present value at the beginning of time, then consumers have no interest in trading in shares, and there is no point in a stock market. If all accounts can be settled at the beginning of time, then there is no need for money during the subsequent life of the economy; in any case, the standard motives for holding money are not applicable.

On the other hand, once we recognize explicitly that there is a sequence of markets, one for each date, and not one of them complete (in the Arrow–Debreu sense), then certain phenomena and institutions not accounted for in the Arrow–Debreu model become reasonable.

First, there is uncertainty about the prices that will hold in future markets, as well as uncertainty about the environment.

Second, producers do not have a clear-cut natural way of comparing net revenues at different dates and states. Stockholders have an incentive to establish a stock exchange since it enables them to change the way their future revenues depend on the states of the environment. As an alternative to selling his shares in a particular enterprise, a stockholder may try to influence the management of the enterprise in order to make the production plan conform better to his own subjective probabilities and attitude towards risk.

Third, consumers will typically not be able to discount all of their ‘wealth’ at the beginning of time, because (*a*) their shares of producers’ future (uncertain) net revenues cannot be so discounted and (*b*) they cannot discount all of their future resource endowments. Consumers will be subject to a sequence

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of budget constraints, one for each date (rather than to a single budget constraint relating present cost of his consumption plan to present net worth, as in the Arrow–Debreu economy).
Fourth, economic agents may have an incentive to speculate on the prices in future markets, by storing goods, hedging, and so on. Instead of storing goods, an agent may be interested in saving part of one date's income, in units of account, for use on a subsequent date, if there is an institution that makes this possible. There will thus be a demand for ‘money’ in the form of demand deposits.
Fifth, agents will be interested in forecasting the prices in markets at future dates. These prices will be functions of both the state of the environment and the decisions of (in principle, all) economic agents up to the date in qst.
Sixth, if traders have different information at a particular date, then the equilibrium prices at that date will reflect the pooled information of the traders, albeit in a possibly complicated way. Hence, traders who have a good model of the market process will be able to infer something about other traders’ information from the market prices.

Expectations and equilibrium in a sequence of markets

Consider now a sequence of markets at successive dates. Suppose that no market at any one date is complete in the Arrow–Debreu sense: that is, at every date and for every commodity there will be some future dates and some events at those future dates for which it will not be possible to make current contracts for future delivery contingent on those events. In such a model, several types of ‘equilibrium’ concept suggest themselves, according to the hypotheses we make about the way traders form their expectations.
Let us place ourselves at a particular date–event pair; the excess supply correspondence at that date–event pair reflects the traders’ information about past prices and about the history of the environment up through that date. If a given trader's excess supply correspondence is generated by preference satisfaction, then the relevant preferences will be conditional upon the information available. If, furthermore, the trader's preferences can be scaled in terms of utility and subjective probability, and conform to the expected utility hypothesis, then the relevant probabilities are the conditional probabilities given the available information. These conditional probabilities express the trader's expectations regarding the future. Although a general theoretical treatment of our problem does not necessarily require us to assume that traders’ preferences conform to the expected utility hypothesis, it will be helpful in the following heuristic discussion to keep in mind this particular interpretation of expectations.
A trader's expectations concern both future environmental events and future prices. Regarding expectations about future environmental events, there is no conceptual problem. According to the expected utility hypothesis, each trader is characterized by a subjective probability measure on the set of complete histories of the environment. Since, by definition, the evolution of the environment is exogenous, a trader's conditional subjective probability of a future event, given the information to date, is well defined.
It is not so obvious how to proceed with regard to traders’ expectations about future prices. We shall contrast two possible approaches. In the first, which we shall call the *perfect foresight* approach, let us assume that the behaviour of traders is such as to determine, for each complete history of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

environment, a unique corresponding sequence of price systems, say $\varphi_t^*(e_t)$, where e_t is the particular event at date t. If the 'laws' governing the economic system are known to all, then every trader can calculate the sequence of functions φ_t^*. In this case, at any date–event pair a trader's expectations regarding future prices are well defined in terms of the functions φ_t^* and his conditional subjective probability measures on histories of the environment, given his current information. Traders need not agree on the probabilities of future environmental events, and therefore they need not agree on the probability distribution of future prices, *but they must agree on which future prices are associated with which events*. We shall call this last type of agreement the condition of *common price expectation functions*.

Thus, the perfect foresight approach implies that, in equilibrium, traders have common price expectation functions. These price expectation functions indicate, for each date–event pair, what the equilibrium price system would be in the corresponding market at that date–event pair. Pursuing this line of thought, it follows that, in equilibrium, the traders would have strategies (plans) such that, if these strategies were carried out, the markets would be cleared at each date–event pair. Call such plans *consistent*. A set of common price expectations and corresponding consistent plans is called an *equilibrium of plans, prices and price expectations (EPPPE)*.

This model of equilibrium can be extended to cover the case in which different traders have different information, just as the Arrow–Debreu model was so extended. In particular, one could express in this way the hypothesis that a trader cannot observe the individual preferences and resource endowments of other traders. Indeed, one can also introduce into the description of the state of the environment variables that, for each trader, represent his alternative hypotheses about the 'true laws' of the economic system. In this way the condition of common price expectation functions can lose much of its apparent restrictiveness.

The situation in which traders enter the market with different nonprice information presents an opportunity for agents to learn about the environment from prices, since current market prices reflect, in a possibly complicated manner, the nonprice information signals received by the various agents. To take an extreme example, the 'inside information' of a trader in a securities market may lead him to bid up the price to a level higher than it otherwise would have been. In this case, an astute market observer might be able to infer that an insider has obtained some favourable information, just by careful observation of the price movement. More generally, *an economic agent who has a good understanding of the market is in a position to use market prices to make inferences about the (nonprice) information received by other agents*.

These inferences are derived, explicitly or implicitly, from an individual's 'model' of the relationship between the nonprice information received by market participants and the market prices. On the other hand, the true relationship is determined by the individual agents' behaviour, and hence by their individual models. Furthermore, economic agents have the opportunity to revise their individual models in the light of observations and published data. Hence, there is a feedback from the true relationship to the individual models. An equilibrium of this system, in which the individual models are identical with the true model, is called *rational expectations equilibrium (REE)*.

This concept of equilibrium is more subtle, of course, than the ordinary concept of the equilibrium of supply and demand. In a rational expectations equilibrium, not only are prices determined so as to equate supply and demand, but individual economic agents correctly perceive the true relationship

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
between the nonprice information received by the market participants and the resulting equilibrium market prices. This contrasts with the ordinary concept of equilibrium in which the agents respond to prices but do not attempt to infer other agents' nonprice information from the actual market prices. Although it is capable of describing a richer set of institutions and behaviour than is the Arrow–Debreu model, the perfect foresight approach is contrary to the spirit of much of competitive market theory in that it postulates that individual traders must be able to forecast, in some sense, the equilibrium prices that will prevail in the future under all alternative states of the environment. Even if one grants the extenuating circumstances mentioned in previous paragraphs, this approach still seems to require of the traders a capacity for imagination and computation far beyond what is realistic. An equilibrium of plans and price expectations might be appropriate as a conceptualization of the ideal goal of indicative planning, or of a long-run steady state towards which the economy might tend in a stationary environment.

These last considerations lead us in a different direction, which we shall call the *bounded rationality* approach. This approach is much less well defined, but expresses itself in terms of various retreats from the hypothesis of 'fully rational' behaviour by traders, for example, by assuming that the trader's planning horizons are severely limited, or that their expectation formation follows some simple rules-of-thumb. An example of the bounded-rationality approach is the theory of *temporary (or momentary) equilibrium*.

In the evolution of a sequence of temporary equilibria, each agent's expectations will be successively revised in the light of new information about the environment and about current prices (see Grandmont, 1987). Therefore, the evolution of the economy will depend upon the rules or processes of expectation formation and revision used by the agents. In particular, there might be interesting conditions under which such a sequence of temporary equilibria would converge, in some sense, to a (stochastic) steady state. This steady state, for example, a stationary probability distribution of prices, would constitute a fourth concept of equilibrium (see Bhattacharya and Majumdar, 2007).

Of the four concepts of equilibrium, the first two are perhaps the closest in the spirit to the Arrow–Debreu theory. How far do some of the conclusions of the Arrow–Debreu theory extend to this new situation? We turn now to this qst. The literature subsequent to the publication of Radner (1967; 1968; 1972) is already voluminous. The interested reader is referred to the reviews by Magill and Shafer (1991), Geanakopolos (1990), Shafer (1998), and the books by Duffie (1988), Magill and Quinzii (1996), LeRoy and Werner (2001).

Equilibrium of plans, prices and price expectations

Consider now the model of perfect-foresight equilibrium sketched above, in which the agents have common information at every date–event pair (for a precise description of the model, see Radner, 1972). Three features of the situation are different from the Arrow–Debreu model: (1) there is a sequence of markets (or rather a 'tree' of markets), one for each date–event pair, no one of which is complete: (2) for each agent, there is a separate budget constraint corresponding to each date–event pair; (3) even if there is a natural bound on consumption and production, there is no single natural bound on the *positions* that traders can take in the markets for securities, if short sales are permitted, (4) there is no obvious objective for each firm to pursue, since each firm's profit is defined only for each date–event pair.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To deal with points (3) and (4), consider the following assumptions. Regarding (3), although there is no *single* natural bound on traders' positions, *some* bound is natural; for example, a commitment to deliver a quantity of a commodity vastly greater than the total supply would not be credible to moderately well-informed traders. Regarding (4), assume that the manager of each firm has preferences on the sequence of net revenues that can be represented by a continuous, strictly concave utility function. We elaborate on variations these and other assumptions. In other respects, we make the 'standard' assumptions of the Arrow–Debreu model.

A canonical model of sequential trading

For ease of exposition we sketch a matchbox model of sequential trading, variations and extensions of which have provided useful building blocks in the formal development beyond the Arrow–Debreu framework. Our exposition draws upon the excellent introduction to the literature by Shafer (1998). There are two periods, 0 and 1, with S states of nature in period 1. In each period, ℓ commodities are traded, with the trades in period 1 being contingent on the realized state s: thus, there, are $L \equiv \ell(S+1)$ contingent commodities in the model (that is, the commodity space is R^L). We have $I \geq 2$ agents, each characterized by a utility function u_i (representing the preferences) and an endowment vector w_i (in R^L_{++}, the set of all strictly positive vectors in R^L). Denote by $w=(w_1,\ldots,w_I)$ the list of endowment vectors. The utility functions are assumed to have the appropriate smoothness and boundary conditions, strict monotonicity and strict quasiconcavity properties. Each agent is supposed to know his own characteristics and each observes the true state when it occurs in period 1. We write a vector y in R^L in the form $y=(y(0),y(1),\ldots,y(S))$ with each $y(s)$ in R^ℓ. A *spot price system* is a vector p in R^L_{++}. See Magill and Shafer (1991) for a more complete description and examples.

We first review the Arrow–Debreu competitive equilibrium in this model in which all trades and prices are decided in period 0. We denote a list of prices by $P=(P(0), P(s))$; $P(0)$ is the vector of prices for goods consumed in period 0, and $P(s)(s \geq 1)$ is the price vector in period 0 for delivery of goods in period 1 contingent on state s being the realized state. The ith agent's optimization problem is

$$\underset{x}{\text{maximize}}\, u_i(x)$$
$$\text{subject to: } P(0)(x(0) - w_i(0)) + \sum_{s \geq 1} P(s)[x(s) - w_i(s)] = 0 .$$

Note that the agent *faces a single budget constraint*. A *competitive equilibrium* is a collection $((\bar{x}_i)_{i \geq 1}, \bar{P})$ such that

1. (i) $\bar{x}_i$ solves agent i's optimization problem at $\bar{P}$; and
2. (ii) $\sum_i [\bar{x}_i(s) - w_i(s)] = 0$ for $s=0,\ldots, S$.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To introduce the model of *sequential trading*, we suppose there are *J assets* (or *securities*) which are traded in period 0 and return dividends in period 1. A unit of asset j will cost q_j units of account payable in 0 ($q=(q_1,\ldots,q_J)$) in R^J) and return $\nu_j(s)$ units of account in period 1 in state s. An asset is called *nominal* if the returns are given exogenously; it is called *real* if the return at state s is the market value of a commodity vector, that is, if $\nu_j(s)=p(s)a_j(s)$ for some vector $a_j(s)$ in R^{ℓ}. Of course, mixtures are possible, but, for our matchbox model, we consider only the pure real asset case or the pure nominal asset case. Denote by V the $S\times J$ matrix of returns that has in row s and column j the dividend, $\nu_j(s)$, of asset j in state s, and let $\nu(s)$ denote the vector of the returns in state s. In the real asset case, the list of vectors $a=(a_j(s))$ parameterizes the asset structure, and the returns matrix $V(p)$ is a function of p. In the nominal case the returns matrix V itself parametrizes the asset structure. Denote by z_j the amount purchased of asset j, with $z=(z_1,\ldots,z_J)$ being the portfolio of assets. The amount z_j may be positive or negative; the assets are considered to be in zero net supply.
We now apply to this model the concept of an *equilibrium of plans, prices, and price expectations*. At a spot price system p and an asset price system q, define an agent's optimization problem as

$$\begin{aligned}
&\underset{x,z}{\text{maximize}} \quad u_i(x)\\
&\text{subject to:} \quad p(0)x(0) - w_i(0) = -qz,\\
&p(s)(x(s) - w_i(s)) = \sum_j v_j(s) z_j,\ s = 1, \ldots, S.
\end{aligned}$$

We should stress that the agent faces *a multiplicity of budget constraints*. The first constraint listed is that the net expenditure on goods plus the cost of the portfolio of assets must sum to zero. The constraints for $s \geq 1$ indicate that if s is the realized state in period 1, the *net* expenditure on goods must equal the dividends of the asset portfolio. The purchase of the assets in period 0 and their dividends in period 1 provides a means both for transferring income between period 0 and period 1 and for transferring income across the potential states in period 1. Note that these constraints preclude the agent from planning bankruptcy in any state; implicit in the constraints is an infinite penalty for bankruptcy.
An *equilibrium of plans, prices, and price expectations* is a list $((\bar{x}_i, \bar{z}_i)_{i\geq 1}, (\bar{p}, \bar{q}))$ such that:

1. (i) $(\bar{x}_i, \bar{z}_i)$ solves agent i's optimization problem at $(\bar{p}, \bar{q})$;
2. (ii) $\sum_i(\bar{x}_i(s) - w_i(s)) = 0$, $s=0,\ldots,S$; and
3. (iii) $\sum_i \bar{z}_i = 0$.

The interpretation of the EPPPE concept is as follows. In period 0, each agent i observes the current spot prices $\bar{p}(0)$ and the asset prices $\bar{q}$, the ‘prices’. Then, based on ‘price expectations’ about spot prices in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

period 1, say $p^e(s)$, $s = 1, \ldots, S$, the agent solves the optimization problem, forming the 'plans' $(\bar{x}_i, \bar{z}_i)$. If it turns out that the price expectations of all the agents are the *same* and the *common* expectation $\bar{p}(s)(s = 1, \ldots, S)$ is such that, together with the observed prices $\bar{p}(0)$ and $\bar{q}$, all markets clear, then we are in an EPPPE equilibrium.

The information requirements of this model are quite strict. Each agent is fully informed in the sense that he or she will be able to verify the true state once it occurs. In addition, each agent knows exactly the distribution of returns of each asset across the states (that is, each agent knows each $a_j(\cdot)$ in the real asset case and each $\nu_j(\cdot)$ in the nominal case).

We now define *completeness of markets* in this model. The formal definition is that, for any possible vector of units of account across the *S* states of nature, an agent can form a portfolio of assets that gives this distribution of returns. That is, for any *S*-vector *y* of net expenditures on goods in period 1 $(y_s = p(s)(x(s) - w(s)))$, there is some portfolio *z* such that *y*=*Vz*. In our model, for the nominal case this is equivalent to the *returns matrix V* having the rank *S*, so that the *column vectors* of *V span all* of R^S. In particular, there must be $J \geq S$ assets. In the case of real assets the rank of the returns matrix *V*(*p*) is a function of *p*, and is thus endogenous to the model. However, *since V*(·) *is linear in p, it has a 'generic' rank* (see Magill and Shafer, 1990, for a precise formulation), and this rank is the maximum rank the returns matrix can take on at any *p*. The real asset structure is defined to be complete if *this generic rank is S*. Again *this requires* $J \geq S$. We note for later reference that, if one imposes restrictions on the size of trades an agent can make in the asset market, then these markets cannot be complete regardless of the number of assets, since by restricting asset trades *z* we cannot in general expect to express every vector in R^S in the form *Vz*.

A very useful implication of completeness will now be stated. Suppose the market is complete. Then given a competitive equilibrium $((\bar{x}_i)_{i \geq 1}, \bar{p})$ one can construct a canonical model with an appropriate EPPPE equilibrium $((\bar{x}_i, z_i)_{i \geq 1}, \bar{p}, \bar{q})$: it should be stressed that the *allocation* $(\bar{x}_i)_{i \geq 1}$ *is the same*. Conversely, given an EPPPE $((\bar{x}_i, \bar{z}_i)_{i \geq 1}, \bar{p}, q)$, one can construct a competitive equilibrium $((\bar{x}_i)_{i \geq 1}, \bar{p})$. This 'translation' of the frameworks has been exploited in the formal analysis, and has clear implications for the *optimality* of an EPPPE (see Magill and Shafer, 1991; Shafer, 1998).

Radner (1972) demonstrated that an equilibrium exists in a more general model provided we impose bounds on the size of trades in the asset markets.

Existence Theorem 1: In the canonical model described above, if each agent's optimization problem is modified with an additional constraint of the form $z \geq b$, *then an equilibrium of plans, prices, and price expectations exists.*

Reasonable or not, the 'ad hoc' constraint in this result posed an intellectual challenge for subsequent researchers. Moreover, as Hart (1975) discovered, without exogenous bounds on the size of trades in the asset markets, equilibria may fail to exist. The main characteristic of Hart's example is that the *returns matrix changes rank with p*, and one approach has been to restrict attention to asset structures that do not exhibit this behaviour. Cass (2006) and Werner (1985) showed that, if one restricts attention to *pure nominal asset structures*, then equilibria exist *without* imposing lower bounds. Similarly, Geanakoplos and Polemarchakis (1986) observed that, in the *real asset case*, if all assets are denominated in terms of the market value of a single good, then the returns matrix *V*(*p*) will have constant rank and – just as in the nominal case – equilibria always exist. In general, then, we have the following theorem.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Existence Theorem 2: In the canonical model, if the asset structure is such that the matrix of returns has constant rank, then an equilibrium of plans, prices, and price expectations exists.

To get some insights into the non-existence issue, we provide a simple example. Consider the Radner model with one state in period 1, one asset, two consumers, and two goods, with the following data. Endowments are $w_i(s) = (1, 1)$ for s=0, 1. Utility functions are $u_1(x) = v_1(x(0)) + v_1(x(1))$ with $v_1(x) = (1/3)\ln x_1 + (2/3)\ln x_2$ for agent 1 and $u_2(x) = v_2(x(0)) + (1/2)v_2(x(1))$ with $v_2(x) = (2/3)\ln x_1 + (1/3)\ln x_2$ for agent 2. The competitive equilibrium prices can be easily computed in this log-linear economy; they are $p_1(0) = 11/36$, $p_2(0) = 10/36$, $p_1(1) = 7/36$, and $p_2(1) = 8/36$. Now consider a Radner version of the model, with one real asset given by a_1=(8, –7). The return on this asset in state 1 is $V = p_1(1)8 - p_2(1)7$, so investing in this asset is essentially a bet that the relative price of good 1 in terms of good 2 is greater than 7/8. Since there is one state and one asset, this is the complete markets case. Note that the 1×1 returns matrix drops rank precisely in the case the relative price is 7/8 in state 1.

We now show that this model does not have an EPPPE equilibrium. First, we try for an equilibrium with the return $V \neq 0$. If such an equilibrium existed, it would have to coincide with the competitive equilibrium since V has rank 1, but in the unique competitive equilibrium the period-1 ratio is 7/8, so V=0, a contradiction. Second, consider the possibility of an equilibrium with V=0. Then there would be no transfers of income between periods 0 and 1, so the period-1 equilibrium would have to coincide with the static competitive equilibrium with the utility functions υ_1 and υ_2. But it is easy to see from the symmetry of the functions and the equal endowments that the relative price ratio in this case would be 1, and thus $V \neq 0$, again a contradiction. Thus no equilibrium exists. Note, however, that this example is *not* robust: alter the asset a small amount so that $V \neq 0$ at the price ratio 7/8 and then the complete markets case will work; or alter endowments or utility parameters a little so that the period-1 price ratio is no longer 7/8, and the complete markets case again works.

This example gives a clue on how to proceed for the existence problem with real assets when markets are complete. One aims at generic results. In this case, remember, $V(p)$ has constant rank S on an open set of full measure in the space of prices. As we have mentioned above, an EPPPE at which $V(p)$ has rank S is equivalent to a competitive equilibrium with a complete set of contingent commodity contracts. That is, the allocations are the same, and there is an easy correspondence between competitive equilibrium prices and the corresponding EPPPE prices. Thus a natural approach is first to obtain a competitive equilibrium, which always exists in our model, and then to construct the corresponding EPPPE prices. If at these prices $V(p)$ has rank S, then we have an EPPPE.

Kreps (1982) made the critical observation that, if the rank of the returns matrix is less that S, then a perturbation of the returns structure a will restore $V(p)$ to full rank (as in our preceding ex), and thus we will have an EPPPE. That is, *generically* in a, an EPPPE exists. Similarly, Magill and Shafer (1990) observed that a small perturbation of endowment would cause the competitive equilibrium prices to move into the region where $V(p)$ has full rank, and thus an equilibrium exists generically in endowments.

Generic Existence Theorem 1: In the canonical model:

1. (1) *if* $I \geq S$ *then, for each* $w \in R^L_{++}$, *an equilibrium of plans, prices, and price expectations exists for almost all (asset structure) a in* R^{LJS};

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

2. (2) *for each asset structure a for which V(·)has generic rank S, an equilibrium of plans, prices, and price expectations exists for almost all endowment lists w in* R^{LI}_{++}.

In the case where both $V(p)$ can change rank with p and markets are not complete (in particular, if $J<S$), the trick of first obtaining a competitive equilibrium and then converting it to an EPPPE equilibrium is no longer available. Nevertheless, by defining a 'pseudo' equilibrium concept that replaces the competitive equilibrium in the argument for the complete market case, Duffie and Shafer (1985) were able to show that an EPPPE equilibrium exists generically in both *a* and *w*.

Generic Existence Theorem 2: In the canonical model with all real assets, an equilibrium of plans, prices, and price expectations exists for almost all (a,w) in $R^{\ell J S} \times R^{L}_{++}$.

We now look at the issue of *local uniqueness*. In what follows, in 'counting' equilibria we are counting the number of equilibrium allocations, since there may be certain redundancies in equilibrium prices. In the complete market case this is fairly straightforward, requiring only an adaptation of Debreu's (1970) argument, since competitive equilibria and EPPPE equilibria coincide when the returns matrix has full rank. In the case of incomplete markets and all real assets, an argument similar to Debreu's applied to the 'pseudo' equilibrium also works.

Local Uniqueness Theorem 1: In the canonical model, if the asset structure is such that markets are complete, or if all assets are real, then for almost all w in the space of endowment lists there exist a finite number of EPPPE, and each equilibrium is locally a smooth function of endowment lists w and asset structures a or V.

The situation with nominal assets and incomplete markets is, however, completely different. In this case there is 'serious indeterminacy' as the following result suggests.

Local Uniqueness Theorem 2: In the canonical model with nominal assets, let the returns matrix V satisfy J<S and I>J. Then, for almost all w in the space of endowment lists, the set of allocations of an EPPPE contains a set homeomorphic to R^{S-1}. (See Geanakoplos and Mas-Colell, 1989, and a similar result by Balasko and Cass, 1989.)

Next, we turn to Pareto optimality. At an EPPPE equilibrium at which the returns matrix has rank *S*, the resulting equilibrium allocation will also be a competitive equilibrium allocation, and thus fully Pareto efficient. This leads to the following theorem.

Pareto Optimality Theorem 1: For the canonical model:

1. (1) *in the nominal asset case, if V has rank S then every EPPPE equilibrium allocation is Pareto efficient*;
2. (2) *in the real asset case, if the asset structure is such that the generic rank of V(·) is S then, for almost all w in the space of endowment lists, EPPPE equilibrium allocations are Pareto efficient.*

In case of incomplete markets, one certainly does not expect full allocative efficiency. The following result (Genakopolos and Polemarchakis, 1986) emphasizes the failure of the first fundamental theorem.

Non-Optimality Theorem 1: For the canonical model, if J<S then, generically in (w,a) in R^{L}_{++}, *EPPPE equilibrium allocations are not Pareto efficient.*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
One might hope that, in an appropriate sense, EPPPE equilibria are constrained efficient. There is still no generally accepted notion of what the correct definition of ‘constrained efficiency’ might be in this case; some argue that the concept cannot be properly defined unless the reasons for incompleteness of markets are endogenously embedded into the model. Nevertheless, we can discuss certain efficiency properties of the equilibria. First, there are *robust* examples of the model with multiple equilibria in which two of the EPPPE equilibria have the property that one Pareto dominates the second (Shafer, 1998). As a consequence of such robust examples, any efficiency property that the incomplete market EPPPE equilibria may possess must be ‘weak’. One approach to constrained efficiency is to follow the Lange–Lerner tradition: if a central planner were permitted to choose the asset portfolios for the agents, and then allow agents to trade freely on competitive markets for commodities, could the planner improve upon an EPPPE equilibrium? The answer is, in an appropriate generic sense, ‘yes’ (Geanakopolos and Polemarchakis, 1986). This is, of course, not possible if markets are complete.

Another ‘natural’ question to ask: is there a connection between how inefficient the EPPPE equilibria are and how incomplete the markets may be? One measure of incompleteness is $S–J$, assuming the J assets give the returns matrix a generic rank J. By introducing a new asset, which reduces the incompleteness in this sense, does efficiency improve? The answer is ‘no’, again due to an example of Hart (1975), in which a new asset is introduced but the new EPPPE equilibrium allocation is Pareto dominated by the original EPPPE equilibrium allocation. This suggests that perhaps this notion of ‘almost complete’ is at fault.

Production

We first discuss the question of existence of equilibrium, but before paraphrasing the existence theorem we must define what we shall call a *pseudo-equilibrium*.

The definition of pseudo-equilibrium is obtained from the definition of equilibrium by replacing the requirement of consistency of plans by the condition that each date and each event the difference between total saving and total investment (by consumers) is smaller at the pseudo-equilibrium prices than at any other prices.

One can prove (Radner, 1972) that under assumptions about technology and consumer preferences similar to those used in the Arrow–Debreu theory, and with the additional assumptions sketched above: (*a*) there exists a pseudo-equilibrium; (*b*) if in a pseudo-equilibrium the current and future prices on the stock market are all strictly positive, then the pseudo-equilibrium is an equilibrium.

The crucial difference between this theorem and the corresponding one in the Arrow–Debreu theory seems to be due to the form taken by Walras's Law, which in this model can be paraphrased by saying that saving must be at least equal to investment at each date in each event. This form derives from the replacement of a single budget constraint (in terms of present value) by a sequence of budget constraints, one for each date–event pair.

In the above model with production, the ‘shareholders’ have unlimited liability, and therefore have a status more like that of partners than of shareholders, as these terms are usually understood. One way to formulate limited liability for shareholders is to impose the constraint on producers that their net revenues be non-negative at each date–event pair. However, in this case producers’ correspondences may not be upper semicontinuous. This is analogous to the problem that arises when, for a given price

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
system, the consumer's budget constraints force him to be on the boundary of his consumption set. In the case of the consumer, this situation is avoided by some assumption (see Debreu, 1959, notes to ch. 5, pp. 88–9; Debreu, 1962). However, for the case of the producer, it is not considered unusual in the standard theory of the firm that, especially in equilibrium, the maximum profit achievable at the given price system could be zero (for example, in the case of constant returns to scale).

What are conditions on the producers and consumers that would directly guarantee the existence of an equilibrium, not just a pseudo-equilibrium? In other words, under what conditions would the share markets be cleared at every date–event pair? Notice that, if there is an excess supply of shares of a given producer j at a date–event pair (t, e), then at date (t+1) only part of the producer's revenue will be ‘distributed’. One would expect this situation to arise only if his revenue is to be negative in at least one event at date t+1; thus, at such a date–event pair the producer would have a deficit covered neither by ‘loans (that is, not offset by forward contracts) nor by shareholders’ contributions. In other words, the producer would be ‘bankrupt’ at that point.

One approach might be to eliminate from a pseudo-equilibrium all producers for whom the excess supply of shares is not zero at some date–event pair, and then search for an equilibrium with the smaller set of producers, and so on, successively reducing the set of producers until an equilibrium is found. This procedure has the trivial consequence that an equilibrium always exists, since it exists for the case of pure exchange (the set of producers is empty)! This may not be the most satisfactory resolution of the problem, but it does point up the desirability of having some formulation of the possibility of ‘exit’ for producers who are not doing well.

Although the above model with production does not allow for ‘exit’ of producers (except with the modification described in the preceding paragraph), it does allow for ‘entrance’ in the following limited sense. A producer may have zero production up to some date, but plans to produce thereafter; this is not inconsistent with a positive demand for shares at preceding dates.

The creation of new ‘equity’ in an enterprise is also allowed for in a limited sense. A producer may plan for a large investment at a given date–event pair, with a negative revenue. If the total supply of shares at the preceding date–event pair is nevertheless taken up by the market, this investment may be said to have been ‘financed’ by shareholders.

The above assumptions describe a model of producer behaviour that is not influenced by the shareholders or (directly) by the prices of shares. A common alternative hypothesis is that a producer tries to maximize the current market value of this enterprise. There seems to us to be at least two difficulties with this hypothesis. First, there are different market values at different date–event pairs, so it is not clear how these can be maximized simultaneously. Second, the market value of an enterprise at any date–event pair is a price, which is supposed to be determined, along with other prices, by an equilibrium of supply and demand. The ‘market-value-maximizing’ hypothesis would seem to require the producer to predict, in some sense, the effect of a change in his plan on a price *equilibrium*: in this case, the producers would no longer be price-takers, and one would need some sort of theory of general equilibrium for monopolistic competition.

There is one circumstance in which the value of the firm can be defined unambiguously, given the system of present prices and common expectations about future prices. Call a price system *arbitrage-free* if it is not possible to make a sure, positive cash flow from trading, without a positive investment. An equilibrium price system is, a fortiori, arbitrage-free. One can show (see Radner, 1967; Harrison and Kreps, 1979; Duffie and Shafer, 1985 that an arbitrage-free price system implicitly determines a system

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of ‘insurance premiums’ for a corresponding family of events. This means that by suitable trading one can insure oneself against the occurrence of any of these events. If these events include all of the uncertain events that may affect the (uncertain) revenues of the firm, then they can be used in a natural way to define a present value of the firm at any date–event pair, for any production plan of the firm, and no probability judgements are needed to calculate the value. On the other hand, if the family of ‘insurable events’ is not rich enough, then the value is a random variable, and stockholders may not agree on its probability distribution.

A survey of results on the generic existence of equilibrium with production (and stock markets) is given in Magill and Shafer (1991) (see also Duffie, 1988, ch. 2).

Rational expectations equilibrium

The formal study of rational expectations equilibrium was introduced by Radner (1967); it was taken up independently by Lucas (1972) and Green (1973), and further investigated by Grossman, Allen, Jordan, and others. We should emphasize that we are concerned here with the aspect of ‘rational expectations’ in which traders make inferences from market prices about other traders’ information, a phenomenon that is of interest only when traders do not all have the same nonprice information. The term ‘rational expectations equilibrium’ (REE) has also been used to describe a situation in which traders correctly forecast (in some sense or other) the probability distribution of future prices. (See Radner, 1982, for references to the work of Muth and others on this topic.)

The concept of REE has been used to make a number of interesting predictions about the behaviour of markets: see, for example, Futia (1979; 1981) and the references cited there. A sound foundation for such applications requires the investigation of conditions that would ensure the existence and stability of REE.

We adopt the convention that the future utility of the commodities to each trader depends on the *state of the environment*. With this convention, we can model the inferences that a trader makes from the market prices and his own nonprice information signal by a family of conditional probability distributions of the environment given the market prices and his own nonprice information. We shall call such a family of conditional distributions the trader's *market model*. Given such a market model, the market prices will influence a trader's demand in two ways: first, through his budget constraint, and second, through his conditional expected utility function. It is this second feature, of course, that distinguished theories of rational expectations equilibrium from earlier models of market equilibrium.

Given the traders’ market models, the equilibrium prices will be determined by the equality of supply and demand in the usual way, and thus will be a deterministic function of the joint nonprice information that the traders bring to the market. In order for the market models of the traders to be ‘rational’, they must be consistent with that function. To make this idea precise, it will be useful to have some formal notation. Let p denote the vector of market prices, e denote the (utility-relevant) state of the environment, and s_i denote traders i's nonprice information signal (I=1,…,I). The joint nonprice information of all traders together will be denoted by $s=(s_1,\ldots,s_I)$. We shall call s the ‘joint signal’. (The term ‘state of information’ is also commonly applied to this array.) Trader i's market model, say m_i, is a family of conditional probability distributions of e, given s_i and p. Given the traders’ market models, the equilibrium price vector will be some (measurable) function of the joint nonprice information, say

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$p = \varphi(s)$.

To model the required rationality of the traders' models, suppose that, for each *i*, trader *i* has (subjective) prior beliefs about the environment and the information signals that are expressed by a joint probability distribution, say Q_i, of *e* and *s*. These prior beliefs need not, of course, be the same for all traders. Given the price function ϕ, a *rational* market model for trader *i* would be the family of conditional probability distributions of *e*, given s_i and *p*, that are derived from the distribution Q_i and the price function ϕ; thus (supposing *e* and *s* to be discrete variables),

$$m_i(e' \mid s_i', p') = \text{Prob}_{Q_1}(e = e' \mid s_i' \text{ and } \varphi(s) = p'). \quad (1)$$

A given price function ϕ, together with the rationality condition (1), would determine the total market excess supply for each price vector *p* and each joint information signal *s*, say $Z(p,s,\phi)$. Note that the excess supply for any *p* and *s* *depends also on the price function* ϕ, since (in principle) the entire price function is used to calculate the conditional distribution in (1). We can now define a *rational expectations equilibrium* (REE) to be a price function ϕ^* such that, for (almost) every *s*, excess supply is zero at the price vector $\phi^*(s)$, that is,

$$Z(\varphi^*(s), s, \varphi^*) = 0, \text{ for almost every } s. \quad (2)$$

If markets are incomplete, the existence of REE is not assured by the 'classical' conditions of ordinary general equilibrium analysis. Even under such conditions, if traders condition their expected utilities on market prices, then their demands can be discontinuous in the price function. Specific examples of the nonexistence of REE due to such discontinuities were given by Kreps (1977), Green (1977), and others. These examples naturally led theorists to question whether the absence of REE is pervasive or is confirmed to a 'negligible' set of such examples. The work of Radner, Allen, and Jordan (see Jordan and Radner, 1982; Allen, 1986, for references) provided – in a certain context – an essentially complete answer, which can be loosely summarized in the statement that *REE exists generically except when the dimension of the space of private information is equal to the dimension of the price space*. (Recall that REE exists *generically* in a given model if, for any vector of parameters values for which REE does not exist, there are arbitrarily small perturbations of the parameters for which REE does exist.) Furthermore, if the dimension of the space of private information is strictly *less* than the dimension of the price space, then generically there is a REE that is *fully revealing*, that is, in which the price reveals to each trader all the nonprice information used by *all* traders (Radner, 1979; Allen, 1981).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Equilibrium and learning with imperfect price models

A more stringent ‘rational expectations’ requirement would concern the opportunities that traders might have for learning from experience. For example, suppose that there is a market at each of a succession of dates *t*, and that the successive exogenous vectors (e_t,s_t) are independent and identically distributed. Suppose further that at the beginning of date *t* trader *i* knows the *past* history of environments, prices, and his own nonprice information. On the basis of this history he updates his initial market model to form a current market model. These current market models, together with the nonprice information signals at date *t*, then determine an equilibrium price at date *t*, say p_t^*, as above. The updating of models constitutes the *learning* process of the traders. For a given learning process, one might ask whether the process converges in any useful sense, and if so, whether the models are asymptotically consistent with the (endogenously determined) actual relationship between signals and equilibrium prices, that is whether they converge to a REE. In this case one would say that the REE is *stable* (relative to the learning process).

Thus far, answers to this question are only fragmentary. Bray (1982) has studied a simpler linear asset-market model in which, at each date, each trader *i* updates his model by calculating an ordinary least-squares estimate of the regression of *e* on *p* and s_i, using all the past values (e_i, p_t^*, s_{it}). For this example, Bray proves stability. On the other hand, Blume and Easley (1982) present a somewhat less optimistic view of the possibility of learning rational expectations. They define a class of learning procedures by which traders use successive observations to form their subjective models, where the term model for trader *i* means a conditional distribution of *s*, given s_i and *p*. They show that rational expectations equilibria are at least ‘locally stable’ under learning, but that learning possesses may also get stuck at a profile of subjective models that is not an REE. The learning procedures defined by Blume and Easley are applied to a fairly general class of stochastic exchange environments that do not possess the special linear structure of the above example. However, to accommodate this additional generality, Blume and Easley constrain traders to choose their subjective models from a fixed finite set of models and convex combinations thereof. Hence, for some profiles of subjective models, market clearing may result in a ‘true’ model that lies outside the admissible set. It is then intuitively plausible that a natural learning procedure could get stuck at a profile of subjective models that differs from the resulting true model but is in some sense the best admissible approximation to the true model, even if the admissible set contains an REE model. This phenomenon is illustrated in Section 5 of their paper. For a review of a number of themes related to learning, see Blume and Easley (1998).

Bibliography

Allen, B. 1981. Generic existence of completely revealing equilibria for economies with uncertainty when prices convey information. *Econometrica* 49, 1173–99.

Allen, B. 1986. General equilibrium with rational expectations. In *Contributions to Mathematical Economics: Essays in Honor of Gerard Debreu*, ed. A. Mas-Colell and W. Hildenbrand. Amsterdam:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
North-Holland.

Arrow, K.J. 1953. Le rôle de valeurs boursières pour la repartition la meilleure des risques. *Econometrie* 11, 41–8. Trans. as ‘The role of securities in the optimal allocation of risk-bearing’, *Review of Economic Studies* 31 (1964), 91–6.

Arrow, K.J. 1965. *Aspects of the Theory of Risk-Bearing*. Helsinki: Yrjo Johansson Foundation.

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Balasko, Y. and Cass, D. 1989. The structure of financial equilibrium I: exogenous yields and unrestricted participation. *Econometrica* 57, 135–62.

Bhattacharya, R. and Majumdar, M.K. 2007. *Random Dynamical Systems: Theory and Applications*. Cambridge: Cambridge University Press.

Blume, L.E. and Easley, D. 1982. Learning to be rational. *Journal of Economic Theory* 26, 318–39.

Blume, L.E. and Easley, D. 1998. Rational expectations and rational learning. In *Organizations with Incomplete Information*, ed. M. Majumdar. Cambridge: Cambridge University Press.

Bray, M. 1982. Learning, estimation and the stability of rational expectations. *Journal of Economic Theory* 26, 318–39.

Burke, J. 1986. Existence of equilibrium for incomplete market economies with production and stock trading. Unpublished manuscript, Texas A&M University.

Cass, D. 2006. Competitive equilibrium with incomplete financial markets. *Journal of Mathematical Economics* 42, 384–405.

Debreu, G. 1953. Une économie de l’incertain. Mimeo, Electricité de France, Paris. Trans. as ‘Economics under Uncertainty’, in G. Debreu, *Mathematical Economics, Twenty Papers of Gerard Debreu*. Cambridge: Cambridge University Press, 1983.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Debreu, G. 1962. New concepts and techniques for equilibrium analysis. *International Economic Review* 3, 257–73.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38, 387–92.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Duffie, D. 1988. *Security Markets: Stochastic Models*. Boston: Academic Press.

Duffie, D. and Shafer, W. 1985. Equilibrium in incomplete markets I: a basic model of generic existence. *Journal of Mathematical Economics* 14, 285–300.

Duffie, D. and Shafer, W. 1987. Equilibrium in incomplete markets II: generic existence in stochastic economies. *Journal of Mathematical Economics* 15, 199–216.

Futia, C.A. 1979. Stochastic business cycles. Unpublished manuscript. Murray Hill, NJ: AT&T Bell Laboratories.

Futia, C.A. 1981. Rational expectations in stationary linear models. *Econometrica* 49, 171–92.

Geanakopolos, J. 1990. An introduction to general equilibrium with incomplete asset markets. *Journal of Mathematical Economics* 19, 1–38.

Geanakoplos, J. and Mas-Colell, A. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47, 22–38.

Geanakoplos, J. and Polemarchakis, H. 1986. Existence, regularity, and constrained suboptimality of competitive allocations when the asset market is incomplete. In *Uncertainty, Information and Communication: Essays in Honor of Kenneth Arrow*, vol. 3, ed. W.P. Heller, R.M. Starr and D.A. Starrett. Cambridge: Cambridge University Press.

Grandmont, J.-M. 1987. Temporary equilibrium. In *The New Palgrave*, vol. 4, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Green, J. 1973. Information, inefficiency, and equilibrium. Discussion Paper No. 284, Harvard Institute of Economic Research.

Green, J. 1977. The nonexistence of informational equilibria. *Review of Economic Studies* 44, 451–63.

Harrison, J.M. and Kreps, D.M. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20, 381–408.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11, 418–43.

Jordan, J.S. and Radner, R. 1982. Rational expectations in microeconomic models: an overview. *Journal of Economic Theory* 26, 201–23.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kreps, D.M. 1977. A note on ‘fulfilled expectations’ equilibria. *Journal of Economic Theory* 14, 32–43.

Kreps, D.M. 1982. Multi-period securities and the efficient allocation of risk: a comment on the Black–Scholes option pricing model. In *The Economics of Uncertainty and Information*, ed. J. McCall. Chicago: University of Chicago Press.

LeRoy, S.E. and Werner, J. 2001. *Principles of Financial Economics*. Cambridge: Cambridge University Press.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–24.

Magill, M. and Quinzii, M. 1996. *Theory of Incomplete Markets*. Cambridge, MA: MIT Press.

Magill, M. and Shafer, W. 1990. Characterization of generically complete real asset structures. *Journal of Mathematical Economics* 19, 167–94.

Magill, M. and Shafer, W. 1991. Incomplete markets. In *Handbook of Mathematical Economics*, vol. 4, ed. W. Hildenbrand and H. Sonnenschein. New York: Elsevier.

Radner, R. 1967. Equilibrie des marchés à terme et au comptant en cas d’incertitude. *Cahiers d’ Econometrie* 4, 35–52. Paris: CNRS.

Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36, 31–58.

Radner, R. 1972. Existence of equilibrium of plans, prices, and price expectations in a sequence of markets. *Econometrica* 40, 289–304.

Radner, R. 1979. Rational expectations equilibrium: generic existence and the information revealed by prices. *Econometrica* 47, 655–7.

Radner, R. 1982. Equilibrium under uncertainty. In *Handbook of Mathematical Economics*, vol. 2, ed. K. J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Radner, R. and Stiglitz, J.E. 1984. A nonconcavity in the value of information. In *Bayesian Models of Economic Theory*, ed. M. Boyer and R.E. Kihlstrom. Amsterdam: North-Holland.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Wiley.

Shafer, W. 1998. Equilibrium with incomplete markets in a sequence economy. In *Organizations with Incomplete Information*, ed. M. Majumdar. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Werner, J. 1985. Equilibrium in economies with incomplete financial markets. *Journal of Economic Theory* 36, 110–9.

Younes, Y. 1985. Competitive equilibrium and incomplete market structures. Unpublished manuscript, Department of Economics, University of Pennsylvania (revised October 1986).

How to cite this article

Majumdar, Mukul and Roy Radner. "uncertainty and general equilibrium." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000007> doi:10.1057/9780230226203.1754

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

uncertainty

Peter P. Wakker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article deals with individual decision making under uncertainty (unknown probabilities). Risk (known probabilities) is not treated as a separate case, but as a sub-case of uncertainty. Many results from risk naturally extend to uncertainty. The Allais paradox, commonly applied to risk, also reveals empirical deficiencies of expected utility for uncertainty. The Ellsberg paradox reveals deviations from expected utility in a relative, not an absolute, sense, giving *within-person* comparisons: for some events (ambiguous or otherwise) subjects deviate more from expected utility than for other events. Besides aversion, many other attitudes towards ambiguity are empirically relevant.

Keywords

additive probabilities; Allais paradox; Allais, M.; ambiguity and ambiguity aversion; ambiguity attitudes; Bayesian statistics; Bernoulli, D.; betweenness models; concave utility; convex analysis; convex probability weighting; convex utility; cumulative prospect theory; de Finetti, B.; decision theory; decision under risk; disappointment aversion theory; Ellsberg paradox; existence of equilibria; expected utility; gambling; Greenspan, A.; insurance; Kahneman, D.; Keynes, J. M.; Knight, F.; known probabilities; likelihood; loss aversion; moral hazard; multiple priors model; neuroeconomics; non-expected utility; objective probability; pessimism; probabilistic risk attitudes; probabilistic sophistication; probability; prospect theory; Ramsay, F.; rank dependence; rank-dependent models; rank-dependent utility; reference dependence; revealed preference; risk; risk aversion; Savage, L.; sensitivity; separability; source preference; state spaces; stochastic dominance; subjective probability; sure-thing principle; Tversky, A.; uncertainty; unknown probabilities; Wald, A.; weighted utility; within-person comparison

Article

In most economic decisions where agents face uncertainties, no probabilities are available. This point was first emphasized by Keynes (1921) and Knight (1921). It was recently reiterated by Greenspan (2004, p. 38):

> … how … the economy might respond to a monetary policy initiative may need to be drawn from evidence about past behavior during a period only roughly comparable to the current situation. … In pursuing a risk-management approach to policy, we must confront the fact that only a limited number of risks can be quantified with any confidence.

Indeed, we often have no clear statistics available. Knight went so far as to call probabilities unmeasurable in such cases. Soon after Knight's suggestion, Ramsey (1931), de Finetti (1931) and Savage (1954) showed that probabilities can be defined in the absence of statistics after all, by relating them to observable choice. For example, $P(E) = 0.5$ can be derived from an observed indifference between receiving a prize under event *E* and receiving it under not-*E* (the complement to *E*). Although widely understood today, the idea that something as intangible as a subjective degree of belief can be made observable through choice behaviour, and can even be quantified precisely, was a major intellectual advance. Ramsey, de Finetti and Savage assumed that the agent, after having determined the probabilities subjectively (as required by some imposed rationality axioms), proceeds as under

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

expected utility for given objective probabilities. The Allais (1953) paradox (explained later) revealed a descriptive difficulty: for known probabilities, people often do not satisfy expected utility. Hence, we need to generalize expected utility. Another, more fundamental, difficulty was revealed by the Ellsberg (1961) paradox (also explained later): for unknown probabilities, people behave in ways that cannot be reconciled with any assignment of subjective probabilities at all, so that further generalizations are needed. (The term 'subjective probability' always refers to additive probabilities in this article.)

Despite the importance and prevalence of unknown probabilities, understood since 1921, and the impossibility of modelling these through subjective probabilities, understood since Ellsberg (1961), decision theorists continued to confine their attention to decision under risk with given probabilities until the late 1980s. Wald's (1950) multiple priors model did account for unknown probabilities, but attracted little attention outside statistics.

As a result of an idea of David Schmeidler (1989, first version 1982), the situation changed in the 1980s. Schmeidler introduced the first theoretically sound decision model for unknown probabilities without subjective probabilities, called rank-dependent utility or Choquet expected utility. At the same time, Wald's multiple priors model was revived when Gilboa and Schmeidler (1989) established its theoretical soundness; a similar result was obtained independently by Chateauneuf (1991, Theorem 2). These discoveries provided the basis for non-expected utility with unknown probabilities that had been sorely missing since 1921. Since the late 1980s, the table has turned in decision theory. Nowadays, most studies concern unknown probabilities. Gilboa (2004) contains recent papers and applications. This article concentrates on conceptual issues of individual decisions in the possible absence of known probabilities.

Theoretical studies of non-expected utility have usually assumed risk aversion for known probabilities (leading to concave utility and convex probability weighting), and ambiguity aversion for unknown probabilities (Camerer and Weber, 1992, section 2.3). These phenomena best fit with the existence of equilibria and can be handled using conventional tools of convex analysis (Mukerji and Tallon, 2001). Empirically, however, a more complex fourfold pattern has been found. For gains with moderate and high likelihoods, and for losses with low likelihoods, risk aversion is prevalent indeed, but for gains with low likelihoods and for losses with high likelihoods the opposite, risk seeking, is prevalent.

The fourfold pattern resolves the classical paradox of the coexistence of gambling and insurance, and leads, for instance, to new views on insurance. Whereas all classical studies of insurance explain insurance purchasing through concave utility, empirical measurements of utility have suggested that utility is not very concave for losses, often exhibiting more convexity than concavity (surveyed by Köbberling, Schwieren and Wakker, 2006). This finding is diametrically opposite to what has been assumed throughout the insurance literature. According to modern decision theories, insurance is primarily driven by consumers' overweighting of small probabilities rather than by marginal utility.

The fourfold pattern found for risk has similarly been found for unknown probabilities, and usually to a more pronounced degree. Central qsts in uncertainty today concern how to analyse not only classical marginal utility but also new concepts such as probabilistic risk attitudes (how people process known probabilities), loss aversion and reference dependence (the framing of outcomes as gains and losses), and, further, states of belief and decision attitudes regarding unknown probabilities ('ambiguity attitudes').

We end this introduction with some notation and definitions. *Decision under uncertainty* concerns choices between *prospects* such as $(E_1: x_1, \ldots, E_n: x_n)$, yielding *outcome* x_j if *event* E_j obtains, $j = 1, \ldots, n$. Outcomes are monetary. The E_js are events of which an agent does not know for sure whether they will obtain, such as who of n candidates will win an election. The E_js are mutually exclusive and exhaustive. No probabilities of the events need to be given. Because the agent is uncertain about which event obtains, he is uncertain about which outcome will result from the prospect, and has to make decisions under uncertainty.

Decision under risk and non-expected utility through rank dependence

Because risk is a special and simple subcase of uncertainty (as explained later), this chapter on uncertainty begins with a discussion of *decision under risk*, where the probability $p_j = P(E_j)$ is given for each event E_j. We can then write a prospect as $(p_1: x_1, \ldots, p_n: x_n)$, yielding x_j with probability p_j, $j = 1, \ldots, n$. Empirical violations of expected-value maximization because of risk aversion (prospects being less preferred than their expected value) led Bernoulli (1738) to propose expected utility, $\Sigma_{j=1}^{n} p_j U(x_j)$, to evaluate prospects, where U is the utility function. Then risk aversion is equivalent to concavity of U.

Several authors have argued that it is intuitively unsatisfactory that risk attitude be modelled through the utility of money (Lopes, 1987, p. 283). It would be more satisfactory if risk attitude were also related to the way people feel about probabilities. Economists often react very negatively to such arguments, based as they are on introspection and having no clear link to revealed preference. Arguments against expected utility that are based on revealed preference were put forward by Allais (1953).

Figure 1 displays preferences commonly found, with K denoting $1,000:

(0.06:25K, 0.07:25K, *0.87:0*) $\prec$ (0.06:75K, 0.07:0, *0.87:0*) and

(0.06:25K, *0.87:25K*, 0.07:25K) $\succ$ (0.06:75K, *0.87:25K*, 0.07:0).

Figure 1

A version of the Allais paradox for risk

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

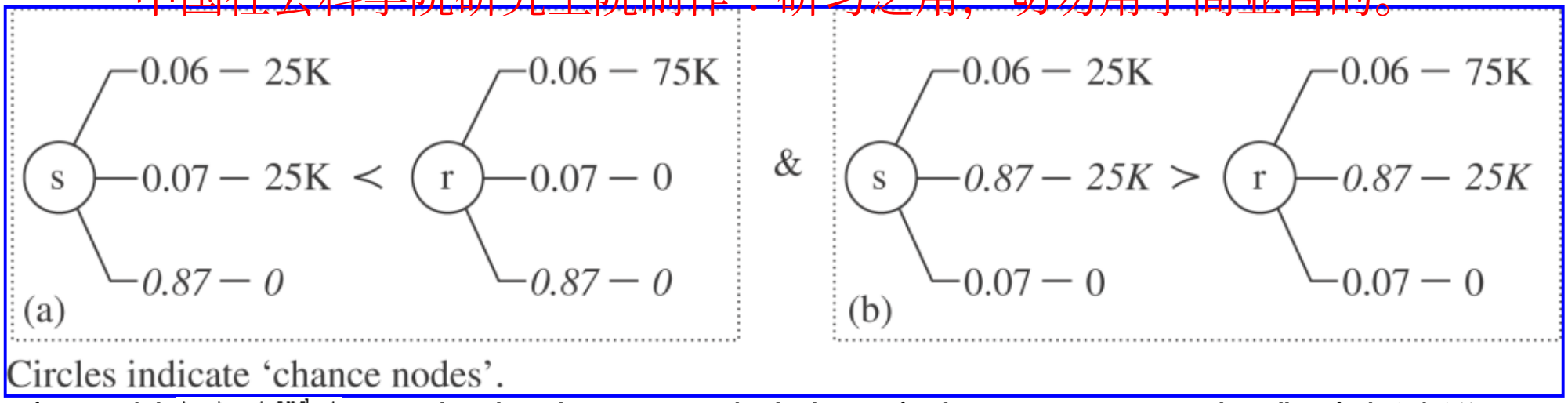

Preference symbols $\succcurlyeq$, $\succ$, $\prec$ and $\preccurlyeq$ are as usual. We denote the outcomes in a rank-ordered manner, from best to worst. In Figure 1a, people usually prefer the 'risky' (r) prospect because the high payment of 75K is attractive. In Figure 1b, people usually prefer the 'safe' (s) certainty of 25K for sure. These preferences violate expected utility because, after dropping the common (italicized) term 0.87U(0) from the expected-utility inequality for Figure 1a and dropping the common term 0.87U(25K) from the expected-utility inequality for Figure 1b, the two inequalities become the same. Hence, under expected utility either both preferences should be for the safe prospect or both preferences should be for the risky one, and they cannot switch as in Figure 1. The special preference for safety in the second choice (the *certainty effect*) cannot be captured in terms of utility. Hence, alternative, non-expected utility models have been developed.

Based on the valuable intuition that risk attitude should have something to do with how people feel about probabilities, Quiggin (1982) introduced rank-dependent utility theory for risk. The same theory was discovered independently for the broader and more subtle context of uncertainty by Schmeidler (1989, first version 1982), a contribution that will be discussed later. A *probability weighting function w*: [0,1] → [0,1] satisfies $w(0) = 0$, $w(1) = 1$, and is strictly increasing and continuous. It reflects the (in)sensitivity of people towards probability. Assume that the outcomes of a prospect $(p_1: x_1, \ldots, p_n: x_n)$ are rank-ordered, $x_1 \geq \cdots \geq x_n$. Then its *rank-dependent utility* (*RDU*) is $\sum_{j=1}^{n} \pi_j U(x_j)$, where *utility U* is as before, and π_j, the *decision weight* of outcome x_j, is $w(p_1 + \cdots + p_j) - w(p_1 + \cdots + p_{j-1})$ (which is $w(p_1)$ for $j = 1$).

Tversky and Kahneman (1992) adapted their widely used original prospect theory (Kahneman and Tversky, 1979) by incorporating the rank dependence of Quiggin and Schmeidler. Prospect theory generalizes rank dependence by allowing a different treatment of gains from that of losses, which is desirable for empirical purposes. In this article on uncertainty, I focus on gains, in which case prospect theory in its modern version, sometimes called cumulative prospect theory, coincides with RDU.

With rank dependence, we can capture psychological (mis)perceptions of unfavourable outcomes being more likely to arise, in agreement with Lopes's (1987) intuition. We can also capture decision attitudes of deliberately paying more attention to bad outcomes. An extreme example of the latter pessimism concerns worst-case analysis, where all weight is given to the most unfavourable outcome. Rank dependence can explain the Allais paradox because the weight of the 0.07 branch in Figure 1b may exceed that in Figure 1a:

$$w(1) - w(0.93) \geq w(0.13) - w(0.06). \tag{1}$$

This inequality holds for *w*-functions that are steeper near 1 than in the middle region, a shape that is empirically prevailing indeed.

The following figures depict some probability weighting functions. Figure 2a concerns expected utility, and Figure 2b a *convex w*, which means that

$$w(p + r) - w(r) \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

is increasing in r for all $p \geq 0$. This is equivalent to w' being increasing, or w'' being positive. Equation (1) illustrates this property. Equation (2) gives the decision weight of an outcome occurring with probability p if there is an r probability of better outcomes. Under convexity, outcomes receive more weight as they are ranked worse (that is, r is larger), reflecting *pessimism*. It implies low evaluations of prospects relative to sure outcomes, enhancing risk aversion.

Figure 2

(a) Expected utility: linearity; (b) Aversion to risk: convexity; (c) Insensitivity: inverse-S (d) Prevailing empirical finding; (e) Extreme insensitivity; (f) Extreme aversion and insensitivity.

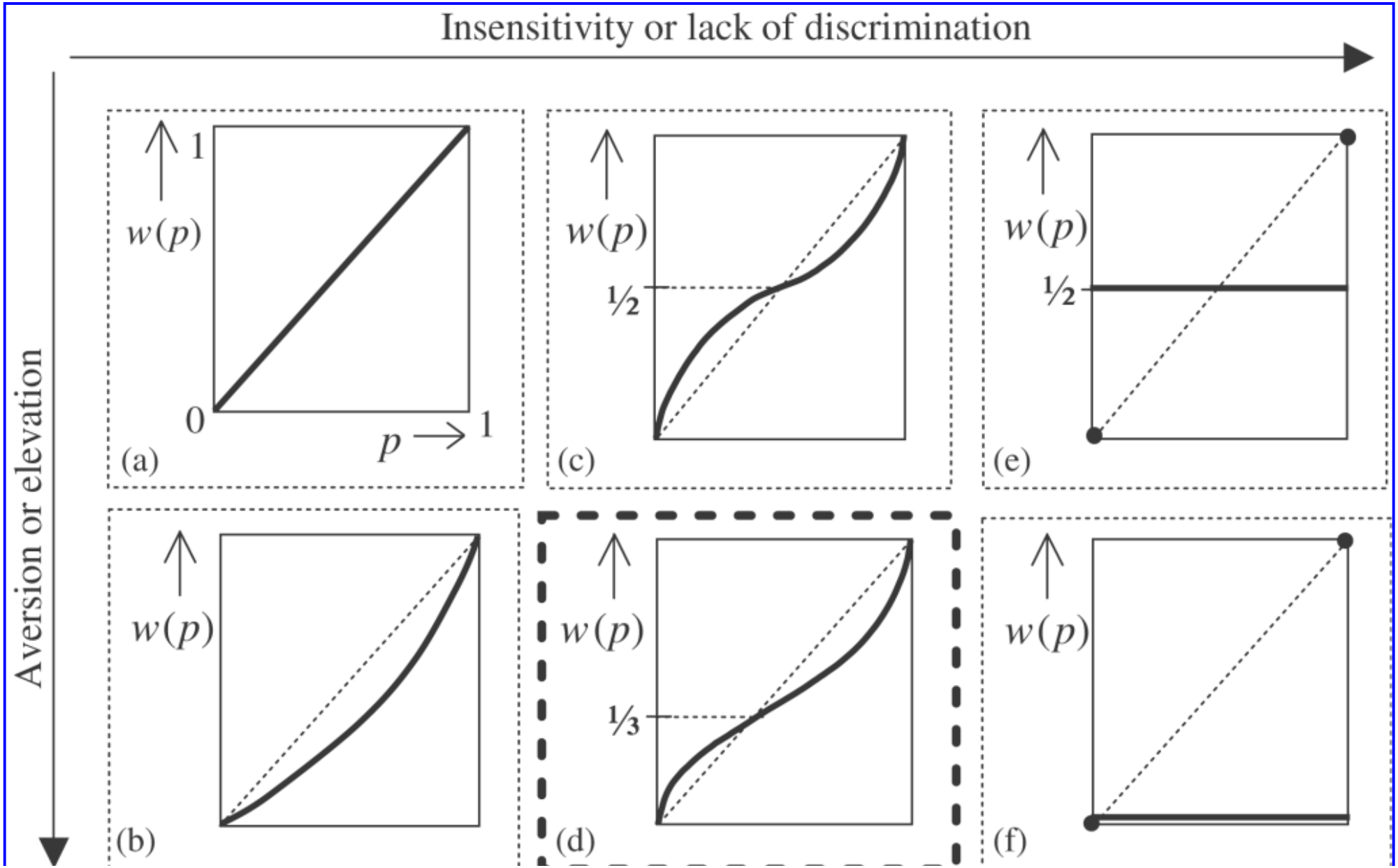

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Empirical studies have found that usually $w(p) > p$ for small *p*, contrary to what convexity would imply, and that $w(p) < p$ only for moderate and high probabilities *p* (inverse-S; Abdellaoui, 2000; Bleichrodt and Pinto, 2000; Gonzalez and Wu, 1999; Tversky and Kahneman, 1992), as in Figures 2c and 2d. It leads to extremity-oriented behaviour with both best and worst outcomes overweighted. The curves in Figures 2c and 2d also satisfy eq. (1) and also accommodate the Allais paradox. They predict risk seeking for prospects that with a small probability generate a high gain, such as in public lotteries. The inverse-S shape suggests a cognitive insensitivity to probability, generating insufficient response to intermediate variations of probability and then, as a consequence, overreactions to changes from impossible to possible and from possible to certain. These phenomena arise prior to any 'motivational' (value-based) preference or dispreference for risk. Extreme cases of such behaviour are in Figure 2e and 2f (where we relaxed the continuity requirement for *w*).
Starmer (2000) surveyed non-expected utility for risk. The main alternatives to the rank-dependent models are the betweenness models (Chew, 1983; Dekel, 1986), with Gul's (1991) disappointment aversion theory as an appealing special case. Betweenness models are less popular today than the rank-dependent models. An important reason, besides their worse empirical performance (Starmer, 2000), is that models alternative to the rank-dependent ones did not provide concepts as intuitive as the sensitivity to probability/information modelled through the probability weighting *w* of the rank-dependent models. For example, consider a popular special case of betweenness, called weighted utility. The value of a prospect is

$$\frac{\sum_{i=1}^{n} p_i f(x_i) U(x_i)}{\sum_{j=1}^{n} p_j f(x_j)} \tag{3}$$

for a function $f : \mathbb{R} \to \mathbb{R}^+$. This new parameter *f* can, similar to rank dependence, capture pessimistic attitudes of overweighting bad outcomes by assigning high values to bad outcomes. It, however, applies to outcomes and not to probabilities. Therefore, it captures less extra variance of the data in the presence of utility than *w*, because utility also applies to outcomes. For example, for fixed outcomes, eq. (3) cannot capture the varying sensitivity to small, intermediate and high probabilities found empirically. Both pessimism and marginal utility are entirely specified by the range of outcomes considered without regard to the probabilities involved. It seems more interesting if new concepts, besides marginal utility, concern the probabilities and the state of information of the decision maker rather than outcomes and their valuation. This may explain the success of rank-dependent theories and prospect theory.

Phenomena under uncertainty that naturally extend phenomena under risk

The first approach to deal with uncertainty was the Bayesian approach, based on de Finetti (1931), Ramsey (1931) and Savage (1954). It assumes that people assign, as well as possible, subjective probabilities $P(E_j)$ to uncertain events E_j. They then evaluate prospects $(E_1 : x_1, \ldots, E_n : x_n)$ through their (subjective) expected utility $\sum_{j=1}^{n} P(E_j) U(x_j)$. This model was the basis of Bayesian statistics and of much of the economics of uncertainty (Greenspan, 2004). The empirical measurement of subjective probabilities has been studied extensively (Fishburn, 1986; Manski, 2004; McClelland and Bolger, 1994). We confine our attention in what follows to models that have been introduced since the mid-1980s, models that deviate from Bayesianism. To Bayesians (including this author) such models are of interest for descriptive purposes.
Machina and Schmeidler (1992) characterized *probabilistic sophistication*, where a decision maker assigns subjective probabilities to events with unknown probabilities and then proceeds as for known probabilities. The decision maker may, however, deviate from expected utility for known probabilities, contrary to the Bayesian approach, and Allais-type behaviour can be accommodated.
The difference between objective, exogenous probabilities and subjective, endogenous probabilities is important. The former are stable, and readily available for analyses, empirical tests and communication in group decisions. The latter can be volatile and can change at any time by mere further thinking by the agent. For descriptive studies, subjective probabilities may become observable only after complex measurement procedures. Hence, I prefer not to lump objective and subjective probabilities together into one category, as has been done in several economic works (Ellsberg, 1961, p. 645; Epstein, 1999). In this article, the term risk refers exclusively to exogenous objective probabilities. Such probabilities can be considered a limiting case of subjective probabilities, in the same way as decision under risk can be considered a limiting case of decision under uncertainty (Greenspan, 2004, pp. 36–7). Under a differentiability assumption for state spaces, Machina (2004) formalized this inpt. Risk, while not occurring very frequently, is especially suited for applications of decision theory.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The Allais paradox is as relevant to uncertainty as it is to risk (MacCrimmon and Larsson, 1979, pp. 364–5; Wu and Gonzalez, 1999). Figure 3 presents a demonstration by Tversky and Kahneman (1992, section 1.3). The analogy with Figure 1 should be apparent. The authors conducted the following within-subjects experiment. Let *d* denote the difference between the closing value of the Dow Jones on the day of the experiment and on the day after, where we consider the events $\mathrm{H(igh)}: d > 35$, $\mathrm{M(iddle)}: 35 \geq d \geq 30$, $\mathrm{L(ow)}: 30 > d$. The total Dow Jones value at the time of the experiment was about 3000. The right prospect in Figure 3b is (H:75K, L:25K, M:0), and the other prospects are denoted similarly. Of 156 money managers during a workshop, 77 per cent preferred the risky prospect *r* in Figure 3a, but 68 per cent preferred the safe prospect *s* in Figure 3b. The majority preferences violate expected utility, just as they do under risk: after dropping the common terms *P*(L)*U*(0) and *P*(L)*U*(25K) (*P* denotes subjective probabilities), the same expected-utility inequality results for Figure 3a as for Figure 3b. Hence, either both preferences should be for the safe prospect, or both preferences should be for the risky one, and they cannot switch as in Figure 3. This reasoning holds irrespective of what the subjective probabilities *P*(H), *P*(M) and *P*(L) are. (The condition of expected utility that is falsified here, Savage's (1954) ‘sure-thing principle’, can be related to the separability preference condition of consumer theory.)

Figure 3

The certainty effect (Allais paradox) for uncertainty

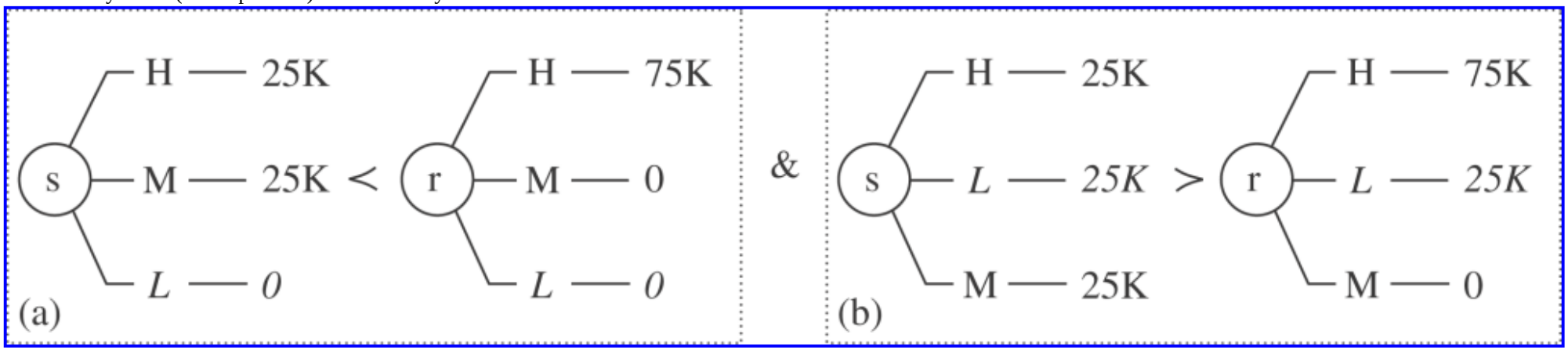

Schmeidler's (1989) *rank-dependent utility* (*RDU*) can accommodate the Allais paradox for uncertainty. We consider a *weighting function* (or non-additive probability or capacity) *W* that assigns value 0 to the vacuous (empty) event ∅, value 1 to the universal event, and satisfies monotonicity with respect to set-inclusion (if $A \supset B$ then $W(A) \geq W(B)$). Probabilities are special cases of weighting functions that satisfy *additivity*: $W(A \cup B) = W(A) + W(B)$ for disjoint events *A* and *B*. General weighting functions need not satisfy additivity. Assume that the outcomes of a prospect $(E_1: x_1, \ldots, E_n: x_n)$ are rank-ordered, $x_1 \geq \cdots \geq x_n$. Then the prospect's *rank-dependent utility* (*RDU*) is $\sum_{j=1}^{n} \pi_j U(x_j)$ where *utility U* is as before, and π_j, the *decision weight* of outcome x_j, is $W(E_1 \cup \cdots \cup E_j) - W(E_1 \cup \cdots \cup E_{j-1})(\pi_1 = W(E_1))$. The decision weight of x_j is the marginal *W* contribution of E_j to the event of receiving a better outcome.

Quiggin's RDU for risk is the special case with probabilities $p_j = P(E_j)$ given for all events, and $W(E_j) = w(P(E_j))$ with *w* the probability weighting function. Tversky and Kahneman (1992) improved their 1979 prospect theory not only by avoiding violations of stochastic dominance, but also, and more importantly, by extending their theory from risk to uncertainty, by incorporating Schmeidler's RDU.

Figure 3 can, just as in the case of risk, be explained by a larger decision weight for the M branches in Figure 3b than in Figure 3a:

$$W(\mathrm{M} \cup \mathrm{H} \cup \mathrm{L}) - W(\mathrm{H} \cup \mathrm{L}) \geq W(\mathrm{M} \cup \mathrm{H}) - W(\mathrm{H}). \tag{4}$$

This inequality occurs for *W*-functions that are more sensitive to changes of events near the certain universal event $\mathrm{M} \cup \mathrm{H} \cup \mathrm{L}$ than for events of moderate likelihood such as $\mathrm{M} \cup \mathrm{H}$. Although for uncertainty we cannot easily draw graphs of *W* functions, their properties are natural analogs of those depicted in Figures 2a–f. *W* is *convex* if the marginal *W* contribution of an event *E* to a disjoint event *R* is increasing in *R*, that is,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$W(E \cup R) - W(R) \quad (5)$$

is increasing in R (with respect to set inclusion) for all E. This agrees with eq. (4), where increasing R from H to $H \cup L$ leads to a larger decision weight for $E = M$. Our definition of convexity is equivalent to other definitions in the literature such as $W(A \cup B) + W(A \cap B) \geq W(A) + W(B)$. (Take $E = A - B$, and compare $R = A \cap B$ with the larger $R = A$.)
For probabilistic sophistication $(W(\cdot) = w(P(\cdot)))$, convexity of W is equivalent to convexity of w under usual richness conditions, illustrating once more the close similarity between risk and uncertainty. Equation (5) gives the decision weight of an outcome occurring under event E if better outcomes occur under event R. Under convexity, outcomes receive more weight as they are ranked worse (that is, R is larger), reflecting *pessimism*. Theoretical economic studies usually assume convex W's, implying low evaluations of prospects relative to sure outcomes.
Empirical studies have suggested that weighting functions W for uncertainty exhibit patterns similar to Figure 2d, with unlikely events overweighted rather than, as convexity would have it, underweighted (Einhorn and Hogarth, 1986; Tversky and Fox, 1995; Wu and Gonzalez, 1999). As under risk, we get extremity orientedness, with best and worst outcomes overweighted and lack of sensitivity towards intermediate outcomes (Chateauneuf, Eichberger and Grant, 2005; Tversky and Wakker, 1995).

Phenomena for uncertainty that do not show up for risk: the Ellsberg paradox

Empirical studies have suggested that phenomena found for risk hold for uncertainty as well, and do so to a more pronounced degree (Fellner, 1961, p. 684; Kahn and Sarin, 1988, p. 270; Kahneman and Tversky, 1979, p. 281), in particular regarding the empirically prevailing inverse-S shape and its extension to uncertainty (Abdellaoui, Vossmann and Weber, 2005; Hogarth and Kunreuther, 1989; Kilka and Weber, 2001; Weber, 1994, pp. 237–8). It is plausible, for example, that the absence of known probabilities adds to the inability of people to sufficiently distinguish between various degrees of likelihood not very close to impossibility and certainty. In such cases, inverse-S shapes will be more pronounced for uncertainty than for risk. This observation entails a within-person comparison of attitudes for different sources of uncertainty, and such comparisons are the main topic of this section.
For Ellsberg's paradox, imagine an urn K with a known composition of 50 red balls and 50 black balls, and an ambiguous urn A with 100 red and black balls in unknown proportion. A ball is drawn at random from each urn, with R_k denoting the event of a red ball from the known urn, and the events B_k, R_a and B_a defined similarly. People prefer to bet on the known urn rather than on the ambiguous urn, and common preferences are:

$$(B_k: 100,\ R_k: 0) \succ (B_a: 100,\ R_a: 0) \text{ and } (B_k: 0,\ R_k: 100) \succ (B_a: 0,\ R_a: 100).$$

Such preferences are also found if people can themselves choose the colour to bet on so that there is no reason for suspecting an unfavourable composition of the unknown urn. Under probabilistic sophistication with probability measure P, the two preferences would imply $P(B_k) > P(B_a)$ and $P(R_k) > P(R_a)$. However, $P(B_k) + P(R_k) = 1 = P(B_a) + P(R_a)$ yields a contradiction, because two big numbers cannot give the same sum as two small numbers. Ellsberg's paradox consequently violates probabilistic sophistication and, a fortiori, expected utility. Keynes (1921, p. 75) discussed the difference between the above two urns before Ellsberg did, but did not put forward the choice paradox and deviation from probabilistic sophistication as Ellsberg did. We now analyse the example assuming RDU.
In many studies of uncertainty, such as Schmeidler (1989), expected utility is assumed for risk, primarily for the sake of simplicity. Then, $W(B_k) = W(R_k) = 0.5$ in the above example, with these W values reflecting objective probabilities. Under RDU, the above preferences imply $W(B_a) = W(R_a) < 0.5$, in agreement with convex (or eventwise dominance, or inverse-S; for simplicity of presentation, I focus on convexity hereafter) weighting functions W. This finding led to the widespread misunderstanding that it is primarily the Ellsberg paradox that implies convex weighting functions for unknown probabilities, a condition that was sometimes called ‘ambiguity aversion’. I have argued above that it is the Allais paradox, and not the Ellsberg paradox, that implies these conclusions, and I propose another interpretation of the Ellsberg paradox hereafter, following works by Amos Tversky in the early 1990s.
First, it is more realistic not to commit to expected utility under risk when studying uncertainty. Assume, therefore, that $W(B_k) = W(R_k) = w(P(B_k)) = w(P(R_k)) = w(0.5)$ for a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

nonlinear probability weighting function. It follows from the Ellsberg paradox that $w(B_a) = w(R_a) < w(0.5)$. This suggests:

Hypothesis.: In the Ellsberg paradox, the weighting function is more convex for the unknown urn than for the known urn. □

Thus, the Ellsberg paradox itself does not speak to convexity in an absolute sense, and does not claim convexity for known or for unknown probabilities. It speaks to convexity in a relative (within-person) sense, suggesting *more* convexity for unknown probabilities than for known probabilities. It is, for instance, possible that the weighting function is concave, and not convex, for both known and unknown probabilities, but is less concave (and thus more convex) for the unknown probabilities (Wakker, 2001, section 6; cf. Epstein, 1999, pp. 589–90, or Ghirardato and Marinacci, 2002, example 25).

With information only about observed behaviour, and without additional information about the compositions of the urns or the agent's knowledge thereof, we cannot conclude which of the urns is ambiguous and which is not. It would then be conceivable that urn K were ambiguous and urn A were unambiguous, and that the agent satisfied expected utility for A and was optimistic or ambiguity seeking (concave weighting function, eq. (5) decreasing in R) for K, in full agreement with the Ellsberg preferences. Which of the urns is ambiguous and which is not is based on extraneous information, being our knowledge about the composition of the urns and about the agent's knowledge thereof. This point suggests that no endogenous definition of (un)ambiguity is possible.

The Ellsberg paradox entails a comparison of attitudes of one agent with respect to different sources of uncertainty. It constitutes a *within-agent* comparison. Whereas the Allais paradox concerns violations of expected utility in an absolute sense, the Ellsberg paradox concerns a relative aspect of such violations, finding more convexity (or eventwise dominance, or inverse-S) for the unknown urn than for the known urn. Such a phenomenon cannot show up if we study only risk, because risk is essentially only one source of uncertainty. Apart from some volatile psychological effects (Kirkpatrick and Epstein, 1992; Piaget and Inhelder, 1975), it seems plausible that people do not distinguish between different ways of generating objective known probabilities.

Uncertain events of particular kinds can be grouped together into sources of uncertainty. Formally, let *sources* be particular algebras of events, which means that sources are closed under complementation and union, and contain the vacuous and universal events. For example, source $\mathcal{A}$ may concern the performance of the Dow Jones stock index tomorrow, and source $\mathcal{B}$ the performance of the Nikkei stock index tomorrow. Assume that A from source $\mathcal{A}$ designates the event that the Dow Jones index goes up tomorrow, and B from source $\mathcal{B}$ the event that the Nikkei index goes up tomorrow. If we prefer (A:100, not-A,0) to (B:100, not-B:0), then this may be caused by a special source preference for $\mathcal{A}$ over $\mathcal{B}$, say, if $\mathcal{A}$ comprises less ambiguity for us than $\mathcal{B}$. However, it may also occur simply because we think that event A is more likely to occur than event B. To examine ambiguity attitudes we have to find a way to 'correct' for differences in perceived levels of likelihood.

One way to detect (*strong*) *source preference* for $\mathcal{A}$ over $\mathcal{B}$ is to find an $\mathcal{A}$-partition $(A_1, \ldots, A_n)$ and a $\mathcal{B}$-partition $(B_1, \ldots, B_n)$ of the universal event such that for each j, (A_j:100, not-A_j,0)•(B_j:100, not-B_j:0) (Nehring, 2001, definition 4; Tversky and Fox, 1995; Tversky and Wakker, 1995). Because both partitions span the whole universal event, we cannot have stronger belief in every A_j than B_j (under some plausible assumptions about beliefs), and hence there must be a preference for dealing with $\mathcal{A}$ events beyond belief. Formally, the condition requires that a similar preference of $\mathcal{B}$ over $\mathcal{A}$ is never detected. The Ellsberg paradox is a special case of this procedure.

Under the above approach to source preference, there is a special role for probabilistic sophistication. For a source $\mathcal{A}$ for which not some of its events are more ambiguous than others, it is plausible that $\mathcal{A}$ exhibits source indifference with respect to itself. This condition can be seen to amount to the additivity axiom of qualitative probability (if A_1 is as likely as A_3, and A_2 is as likely as A_4, then $A_1 \cup A_2$ is as likely as $A_3 \cup A_4$ whenever $A_1 \cap A_2 = A_3 \cap A_4 = \emptyset$), which, under sufficient richness, implies probabilistic sophistication for $\mathcal{A}$ under RDU, and does so in general (without RDU assumed) under an extra dominance condition (Fishburn, 1986; Sarin and Wakker, 2000). The condition also comprises source sensitivity (Tversky and Wakker, 1995). Probabilistic sophistication, then, entails a *uniform degree* of ambiguity of a source.

In theoretical economic studies it has usually been assumed that people are averse to ambiguity, corresponding with convex weighting functions. Empirical studies, mostly by psychologists, have suggested a more varied pattern, where different sources of ambiguity can arouse all kinds of emotions. For example, Tversky and Fox (1995) found that basketball fans exhibit source preference for ambiguous uncertain events related to basketball over events with known probabilities, which entails ambiguity seeking. This finding is not surprising in an empirical sense, but its conceptual implication is important: attitudes towards ambiguity depend on many ad hoc emotional aspects, such as a general aversion to deliberate secrecy about compositions of urns, or a general liking of basketball. Uncertainty is a large domain, and fewer regularities can be expected to hold universally for uncertainty than for risk, in the same way as fewer regularities will hold universally for the utility of non-monetary outcomes (hours of listening to music, amounts of milk to be drunk, life duration, and so on) than for the utility of monetary outcomes. It means that there is much yet to be discovered about uncertainty.

Models for uncertainty other than rank-dependence

Multiple priors

An interesting model of ambiguity by Jaffray (1989), with a separation of ambiguity beliefs and ambiguity attitudes, unfortunately has received little attention as yet. A surprising case of unknown probabilities can arise when the expected utility model perfectly well describes behaviour, but utility is state-dependent. The (im)possibility of defining probability

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in such cases has been widely discussed (Drèze, 1987; Grant and Karni, 2005; Nau, 2006).

The most popular alternative to Schmeidler's RDU is the multiple priors model introduced by Wald (1950). It assumes a set $\mathcal{P}$ of probability measures plus a utility function U, and evaluates each prospect through its minimal expected utility with respect to the probability distributions contained in $\mathcal{P}$. The model has an overlap with RDU: if W is convex, then RDU is the minimal expected utility over $\mathcal{P}$ where $\mathcal{P}$ is the CORE of W, that is, the set of probability measures that eventwise dominate W. Drèze (1961; 1987) independently developed a remarkable analog of the multiple priors model, where the maximal expected utility is taken over $\mathcal{P}$, and $\mathcal{P}$ reflects moral hazard instead of ambiguity. Drèze also provided a preference foundation. Similar functionals appear in studies of robustness against model misspecification in macroeconomics (Hansen and Sargent, 2001).

Variations of multiple priors, combining pessimism and optimism, employ convex combinations of the expected utility minimized over $\mathcal{P}$ and the expected utility maximized over $\mathcal{P}$ (Ghirardato, Maccheroni and Marinacci, 2004, proposition 19). Such models can account for extremity orientedness, as with inverse-S weighting functions and RDU. Arrow and Hurwicz (1972) proposed a similar model where a prospect is evaluated through a convex combination of the minimal and maximal utility of its outcomes (corresponding with $\mathcal{P}$ being the set of all probability measures). This includes maximin and maximax as special cases. Their approach entails a level of ambiguity so extreme that no levels of belief other than ‘sure-to-happen’, ‘sure-not-to-happen’ and ‘don't know’ play a role, similar to Figure 2e and 2f, and suggesting a three-valued logic. Other non-belief-based approaches, including minimax regret, are in Manski (2000) and Savage (1954), with a survey in Barberà, Bossert and Pattanaik (2004).

Other authors proposed models where for each single event a separate interval of probability values is specified (Budescu and Wallsten, 1987; Kyburg, 1983; Manski, 2004). Such interval-probability models are mathematically different from multiple priors because there is no unique relation between sets of probability measures over the whole event space and intervals of probabilities separately for each event. The latter models are more tractable than multiple priors because probability intervals for some relevant event are easier to specify than probability measures over the whole space, but these models did not receive a preference foundation and never became popular in economics. Similar models of imprecise probabilities received attention in the statistics field (Walley, 1991).

Wald's multiple priors model did receive a preference axiomatization (Gilboa and Schmeidler, 1989), and consequently became the most popular alternative to RDU for unknown probabilities. The evaluating formula is easier to understand at first than RDU. The flexibility of not having to specify precisely what ‘the’ probability measure is, while usually perceived as an advantage at first acquaintance, can turn into a disadvantage when applying the model. We then have to specify exactly what ‘the’ set of probability distributions is, which is more complex than exactly specifying only one probability measure (cf. Lindley, 1996).

The simple distinction between probability measures that are either possible (contained in $\mathcal{P}$) or impossible (not contained in $\mathcal{P}$), on the one hand adds to the tractability of the model, but on the other hand cannot capture cognitive states where different probability measures are plausible to different degrees. To the best of my knowledge, the multiple priors model cannot yet be used in quantitative empirical measurements today, and there are no empirical assessments of sets of priors available in the literature to date. Multiple priors are, however, well suited for general theoretical analyses where only general properties of the model are needed. Such analyses are considered in many theoretical economic studies, where the multiple priors model is very useful.

The multiple priors model does not allow deviations from expected utility under risk, and a desirable extension would obviously be to combine the model with non-expected utility for risk. Promising directions for resolving the difficulties of the multiple priors model are being explored today (Maccheroni, Marinacci and Rustichini, 2005).

Model-free approaches to ambiguity

Dekel, Lipman and Rustichini (2001) considered models where outcomes of prospects are observed but the state space has not been completely specified, as relevant to incomplete contracts. Similar approaches with ambiguity about the underlying states and events appeared in psychology in repeated-choice experiments by Hertwig et al. (2003), and in support theory (Tversky and Koehler, 1994). This section discusses two advanced attempts to define ambiguity in a model-free way that have received much attention in the economic literature.

In a deep paper, Epstein (1999) initiated one such approach, continued in Epstein and Zhang (2001). Epstein sought to avoid any use of known probabilities and tried to endogenize (un)ambiguity and the use of probabilities. (He often used the term uncertainty as equivalent to ambiguity.) For example, he did not define risk neutrality with respect to known probabilities, as we did above, but with respect to subjective probabilities derived from preferences as in probabilistic sophistication (Epstein, 1999, eq. (2)). He qualified probabilistic sophistication as ambiguity neutrality (not uniformity as done above). Ghirardato and Marinacci (2002) used another approach that is similar to Epstein's. They identified absence of ambiguity not with probabilistic sophistication, as did Epstein, but, more restrictively, with expected utility.

The above authors defined an agent as ambiguity averse if there exists another, hypothetical, agent who behaves the same way for unambiguous events, but who is ambiguity neutral for ambiguous events, and such that the real agent has a stronger preference than the hypothetical agent for sure outcomes (or unambiguous prospects, but these can be replaced by their certainty equivalents) over ambiguous prospects. This definition concerns traditional between-agent within-source comparisons as in Yaari (1969). The stronger preferences for certainty are, under rank-dependent models, equivalent to eventwise dominance of weighting functions, leading to non-emptiness of the CORE (Epstein, 1999, lemma 3.4; Ghirardato

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and Marinacci, 2002, corollary 13). These definitions of ambiguity aversion are not very tractable because of the 'there exists' clause. It is difficult to establish which ambiguity neutral agent to take for the comparisons. To mitigate this problem, Epstein (1999, section 4) proposed eventwise derivatives as models of local probabilistic sophistication. Such derivatives exist only for continua of events with a linear structure, and are difficult to elicit. They serve their purpose only under restrictive circumstances (ambiguity aversion throughout plus constancy of the local derivative, called coherence; see Epstein's Theorem 4.3).

In both above approaches, ambiguity and ambiguity aversion are inextricably linked, making it hard to model attitudes towards ambiguity other than aversion or seeking (such other attitudes include insensitivity), or to distinguish between ambiguity-neutrality or -absence (Epstein, 1999, p. 584, 1st para; Epstein and Zhang, 2001, p. 283; Ghirardato and Marinacci, 2002, p. 256, 2nd para). Both approaches have difficulties distinguishing between the two Ellsberg urns. Each urn in isolation can be taken as probabilistically sophisticated with, in our inpt, a uniform degree of ambiguity, and Epstein's definition cannot distinguish which of these is ambiguity neutral (cf. Ghirardato and Marinacci, 2002, middle of p. 281). Ghirardato and Marinacci's definition does so, but only because it selects expected utility (and the urn generating such preferences) as the only ambiguity-neutral version of probabilistic sophistication. Any other form of probabilistic sophistication, that is, any non-expected utility behaviour under risk, is then either mismodelled as ambiguity attitude (Ghirardato and Marinacci, 2002, pp. 256–7), or must be assumed not to exist.

We next discuss in more detail a definition of (un)ambiguity by Epstein and Zhang (2001), whose aim was to make (un)ambiguity endogenously observable by expressing it directly in terms of a preference condition. They called an event E *unambiguous* if

$$(E: c,\ E_2: \gamma, E_3: \beta, E_4: x_4, \ldots, E_n: x_n) \succcurlyeq (E: c,\ E_2: \beta, E_3: \gamma, E_4: x_4, \ldots, E_n: x_n) \text{ implies } (E: c',\ E_2: \gamma, E_3: \beta, E_4: x_4, \ldots, E_n: x_n) \succcurlyeq (E: c',\ E_2: \beta, E_3: \gamma, E_4: x_4, \ldots, E_n: x_n) \quad (6)$$

for all partitions $E_2, \ldots, E_n$ of not-*E*, and all outcomes $c, c', x_4, \ldots, x_n, \gamma \succcurlyeq \beta$, with a similar condition imposed on not-*E*. In words, changing a common outcome *c* into another common outcome c' under *E* does not affect preference, but this is imposed only if the preference concerns nothing other than to which event (E_2 or E_3) a good outcome γ is to be allocated instead of a worse outcome β. Together with some other axioms, eq. (6) implies that probabilistic sophistication holds on the set of events satisfying this condition, which in the interpretation of the authors designates absence of ambiguity (rather than uniformity). As we will see next, it is not clear why eq. 6 would capture the absence of ambiguity.

Example.: Assume that events are subsets of [0,1), $E = [0, 0.5)$, $\text{not} - E = [0.5, 1)$, and *E* has unknown probability π. Every subset *A* of *E* has probability $2\pi\lambda(A)$ (λ is the usual Lebesgue measure, that is, the uniform distribution over [0,1)) and every subset *B* of not-*E* has probability $2(1-\pi)\lambda(B)$. Then it seems plausible that event *E* and its complement not-*E* are ambiguous, but conditional on these events ('within them') we have probabilistic sophistication with respect to the conditional Lebesgue measure and without any ambiguity. In Schmeidler (1989), the ambiguous events *E* and not-*E* are called horse events, and the unambiguous events conditional on them are called roulette events. Yet, according to eq. (6), events *E* and not-*E* themselves are unambiguous, both preferences in eq. (6) being determined by whether $\lambda(E_2)$ exceeds $\lambda(E_3)$. □

In the example, the definition in eq. (6) erroneously ascribes the unambiguity that holds for events conditional on *E*, so 'within *E*', to *E* as a whole. Similar examples can be devised where *E* and not-*E* themselves are unambiguous, there is 'non-uniform' ambiguity conditional on *E*, this ambiguity is influenced by outcomes conditional on not-*E* through non-separable interactions typical of non-expected utility, and eq. (6) erroneously ascribes the ambiguity that holds within *E* to *E* as a whole.

A further difficulty with eq. (6) is that it is not violated in the Ellsberg example with urns A and K as above (nor if the uncertainty regarding each urn is extended to a 'uniform' continuum as in Example 5.8ii of Abdellaoui and Wakker, 2005), and cannot detect which of the urns is ambiguous. The probabilistic sophistication that is obtained in Epstein and Zhang (2001, Theorem 5.2) for events satisfying eq. (6), and that rules out the two-urn Ellsberg paradox and its continuous extension of Abdellaoui and Wakker (2005), is mostly driven by their Axioms 4 and 6 (the latter is not satisfied by all rank-dependent utility maximizers contrary to the authors' claim at the end of their Section 4; their footnote 18 is incorrect) and the necessity to consider also intersections of different-urn events (see their Appendix E). This imposes, in my terminology, a uniformity of ambiguity over the events satisfying eq. (6) that, rather than eq. (6) itself, rules out the above counterexs.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Multi-stage approaches to ambiguity

Several authors have considered two-stage approaches with intersections of first-stage events A_i, $i = 1, \ldots, \ell$ and second-stage events K_j, $j = 1, \ldots, k$, so that $n = \ell k$ events A_iK_j result, and prospects $(A_iK_j : x_{ij})_{i=1}^{\ell}{}_{j=1}^{k}$ are considered. It can be imagined that in a first stage it is determined which event A_i obtains, and then in a second stage, conditional on A_i, which event K_j obtains. Many authors considered such two-stage models with probabilities given for the events in both stages, the probabilities of the first stage interpreted as ambiguity about the probabilities of the second stage, and non-Bayesian evaluations used (Levi, 1980; Segal, 1990; Yates and Zukowski, 1976).
Other authors considered representations

$$\sum_{i=1}^{\ell} Q(A_i)\,\phi\left(\sum_{j=1}^{k} P(K_j)\,U(x_{ij})\right) \tag{7}$$

for probability measures P and Q, a utility function U, and an increasing transformation φ. For φ the identity or, equivalently, φ linear, traditional expected utility with backwards induction results. Nonlinear φ's give new models. Kreps and Porteus (1979) considered eq. (7) for intertemporal choice, interpreting nonlinear φ's as non-neutrality towards the timing of the resolution of uncertainty. Ergin and Gul (2004) and Nau (2006) reinterpreted the formula, where now the second-stage events are from a source of different ambiguity than the first-stage events. A concave φ, for instance, suggests stronger preference for certainty, and more ambiguity aversion, for the first-stage uncertainty than for the second.
Klibanoff, Marinacci and Mukerji (2005) considered cases where the decomposition into A- and K-events is endogenous rather than exogenous. This approach greatly enlarges the scope of application, but their second-order acts, that is, prospects with outcomes contingent on aspects of preferences, are hard to implement or observe if those aspects cannot be related to exogenous observables.
Equation (7) has a drawback similar to eq. (3). All extra mileage is to come from the outcomes, to which also utility applies, so that there will not be a great improvement in descriptive performance or new concepts to be developed.

Conclusion

The Allais paradox reveals violations of expected utility in an absolute sense, leading to convex or inverse-S weighting functions for risk and, more generally, for uncertainty. The Ellsberg paradox reveals deviations from expected utility in a relative sense, showing that an agent can deviate more from expected utility for one source of uncertainty (say one with unknown probabilities) than for another (say, one with known probabilities). It demonstrates the importance of within-subject between-source comparisons.
The most popular models for analysing uncertainty today are based on rank dependence, with multiple priors a popular alternative in theoretical studies. The most frequently studied phenomenon is ambiguity aversion. Uncertainty is, however, a rich empirical domain with a wide variety of phenomena, where ambiguity aversion and ambiguity insensitivity (inverse-S) are prevailing but are not universal patterns. The possibility of relating the properties of weighting functions for uncertainty to cognitive inpts such as insensitivity to likelihood-information makes RDU and prospect theory well suited for links with other fields such as psychology, artificial intelligence (Shafer, 1976) and neuroeconomics (Camerer, Loewenstein and Prelec, 2004).

See Also

- Allais, Maurice
- Allais paradox
- ambiguity and ambiguity aversion
- Bernoulli, Daniel

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- certainty equivalence
- de Finetti, Bruno
- decision theory in econometrics
- expected utility hypothesis
- extremal quantiles and value-at-risk
- Kahneman, Daniel
- non-expected utility theory
- rational behaviour
- rational expectations
- revealed preference theory
- risk
- risk aversion
- Savage, Leonard J. (Jimmie)
- Savage's subjective expected utility model
- separability
- statistical decision theory
- stochastic dominance
- Tversky, Amos
- utility

Han Bleichrodt, Chew Soo Hong, Edi Karni, Jacob Sagi and Stefan Trautmann made useful comments.

Bibliography

Abdellaoui, M. 2000. Parameter-free elicitation of utilities and probability weighting functions. *Management Science* 46, 1497–512.

Abdellaoui, M., Vossmann, F. and Weber, M. 2005. Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. *Management Science* 51, 1384–99.

Abdellaoui, M. and Wakker, P.P. 2005. The likelihood method for decision under uncertainty. *Theory and Decision* 58, 3–76.

Allais, M. 1953. Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’école Américaine. *Econometrica* 21, 503–46.

Arrow, K. and Hurwicz, L. 1972. An optimality criterion for decision making under ignorance. In *Uncertainty and Expectations in Economics: Essays in Honour of G. L. S. Shackle*, ed. C. Carter and J. Ford. Oxford: Basil Blackwell.

Barberà, S., Bossert, W. and Pattanaik, P. 2004. Ranking sets of objects. In *Handbook of Utility Theory, Vol. 2, Extensions*, ed. S. Barberà, P. Hammond and C. Seidl. Dordrecht: Kluwer Academic Publishers.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5, 175–92.

Bleichrodt, H. and Pinto, J. 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science* 46, 1485–96.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Budescu, D. and Wallsten, T. 1987. Subjective estimation of precise and vague uncertainties. In *Judgmental Forecasting*, ed. G. Wright and P. Ayton. New York: Wiley.

Camerer, C., Loewenstein, G. and Prelec, D. 2004. Neuroeconomics: why economics needs brains. *Scandinavian Journal of Economics* 106, 555–79.

Camerer, C. and Weber, M. 1992. Recent developments in modelling preferences: uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5, 325–70.

Chateauneuf, A. 1991. On the use of capacities in modeling uncertainty aversion and risk aversion. *Journal of Mathematical Economics* 20, 343–69.

Chateauneuf, A., Eichberger, J. and Grant, S. 2005. Choice under uncertainty with the best and worst in mind: NEO-additive capacities. CERMSEM, CEM, University of Paris I.

Chew, S.H. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* 51, 1065–92.

de Finetti, B. 1931. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae* 17, 298–329.

Dekel, E. 1986. An axiomatic characterization of preferences under uncertainty: weakening the independence axiom. *Journal of Economic Theory* 40, 304–18.

Dekel, E., Lipman, B. and Rustichini, A. 2001. Representing preferences with a subjective state space. *Econometrica* 69, 891–934.

Drèze, J. 1961. Les fondements logiques de l'utilité cardinale et de la probabilité subjective. *La Décision, Colloques Internationaux du CNRS*, 73–83.

Drèze, J. 1987. *Essays on Economic Decision under Uncertainty*. Cambridge: Cambridge University Press.

Einhorn, H. and Hogarth, R. 1986. Decision making under ambiguity. *Journal of Business* 59, S225–S250.

Ellsberg, D. 1961. Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics* 75, 643–69.

Epstein, L. 1999. A definition of uncertainty aversion. *Review of Economic Studies* 66, 579–608.

Epstein, L. and Zhang, J. 2001. Subjective probabilities on subjectively unambiguous events. *Econometrica* 69, 265–306.

Ergin, H. and Gul, F. 2004. A subjective theory of compound lotteries. Working paper, MIT.

Fellner, W. 1961. Distortion of subjective probabilities as a reaction to uncertainty. *Quarterly Journal of Economics* 75, 670–89.

Fishburn, P. 1986. The axioms of subjective probability. *Statistical Science* 1, 335–58.

Ghirardato, P., Maccheroni, F. and Marinacci, M. 2004. Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* 118, 133–73.

Ghirardato, P. and Marinacci, M. 2002. Ambiguity made precise: a comparative foundation. *Journal of Economic Theory* 102, 251–89.

Gilboa, I., ed. 2004. *Uncertainty in Economic Theory: Essays in Honor of David Schmeidler's 65th Birthday*. London: Routledge.

Gilboa, I. and Schmeidler, D. 1989. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18, 141–53.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gonzalez, R. and Wu, G. 1999. On the shape of the probability weighting function. *Cognitive Psychology* 38, 129–66.

Grant, S. and Karni, E. 2005. Why does it matter that beliefs and valuations be correctly represented? *International Economic Review* 46, 917–34.

Greenspan, A. 2004. Innovations and issues in monetary policy: the last fifteen years. *American Economic Review, Papers and Proceedings* 94, 33–40.

Gul, F. 1991. A theory of disappointment aversion. *Econometrica* 59, 667–86.

Hansen, L. and Sargent, T. 2001. Robust control and model uncertainty. *American Economic Review, Papers and Proceedings* 91, 60–6.

Hertwig, R., Barron, G., Weber, E. and Erev, I. 2003. Decisions from experience and the effect of rare events in risky choice. *Psychological Science* 15, 534–9.

Hogarth, R. and Kunreuther, H. 1989. Risk, ambiguity, and insurance. *Journal of Risk and Uncertainty* 2, 5–35.

Jaffray, J.-Y. 1989. Linear utility theory for belief functions. *Operations Research Letters* 8, 107–12.

Kahn, B. and Sarin, R. 1988. Modeling ambiguity in decisions under uncertainty. *Journal of Consumer Research* 15, 265–72.

Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–91.

Keynes, J.M. 1921. *A Treatise on Probability*. London: Macmillan.

Kilka, M. and Weber, M. 2001. What determines the shape of the probability weighting function under uncertainty. *Management Science* 47, 1712–26.

Kirkpatrick, L. and Epstein, S. 1992. Cognitive-experiential self-theory and subjective probability: further evidence for two conceptual systems. *Journal of Personality and Social Psychology* 63, 534–44.

Klibanoff, P., Marinacci, M. and Mukerji, S. 2005. A smooth model of decision making under ambiguity. *Econometrica* 73, 1849–12.

Knight, F. 1921. *Risk, Uncertainty, and Profit*. New York: Houghton Mifflin.

Köbberling, V., Schwieren, C. and Wakker, P.P. 2006. Prospect-theory's diminishing sensitivity versus economics' intrinsic utility of money: how the introduction of the euro can be used to disentangle the two empirically. Mimeo, Department of Economics, University of Maastricht.

Kreps, D. and Porteus, E. 1979. Dynamic choice theory and dynamic programming. *Econometrica* 47, 91–100.

Kyburg, H., Jr. 1983. *Epistemology and Inference*. Minneapolis: University of Minnesota Press.

Levi, I. 1980. *The Enterprise of Knowledge*. Cambridge, MA: MIT Press.

Lindley, D. 1996. Discussion of Walley (1996). *Journal of the Royal Statistical Society B* 58, 47–8.

Lopes, L. 1987. Between hope and fear: the psychology of risk. *Advances in Experimental Psychology* 20, 255–95.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

MacCrimmon, K. and Larsson, S. 1979. Utility theory: axioms versus ‘paradoxes’. In *Expected Utility Hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen. Dordrecht: Reidel.

Machina, M. 2004. Almost-objective uncertainty. *Economic Theory* 24, 1–54.

Machina, M. and Schmeidler, D. 1992. A more robust definition of subjective probability. *Econometrica* 60, 745–80.

Maccheroni, F., Marinacci, M. and Rustichini, A. 2005. Ambiguity aversion, malevolent nature, and the variational representation of preferences. Mimeo, Istituto di Metodi Quantitativi, Università Bocconi, Milan.

Manski, C. 2000. Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics* 95, 415–42.

Manski, C. 2004. Measuring expectations. *Econometrica* 72, 1329–76.

McClelland, A. and Bolger, F. 1994. The calibration of subjective probabilities: theories and models 1980–1994. In *Subjective Probability*, ed. G. Wright and P. Ayton. New York: Wiley.

Mukerji, S. and Tallon, J.-M. 2001. Ambiguity aversion and incompleteness of financial markets. *Review of Economic Studies* 68, 883–904.

Nau, R. 2006. Uncertainty aversion with second-order utilities and probabilities. *Management Science* 52, 136–45.

Nehring, K. 2001. Ambiguity in the context of probabilistic beliefs. Mimeo, UC Davis.

Piaget, J. and Inhelder, B. 1975. *The Origin of the Idea of Chance in Children*. New York: Norton.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behaviour and Organization* 3, 323–43.

Ramsey, F. 1931. Truth and probability. In *The Foundations of Mathematics and other Logical Essays*. London: Routledge and Kegan Paul.

Sarin, R. and Wakker, P.P. 2000. Cumulative dominance and probabilistic sophistication. *Mathematical Social Sciences* 40, 191–6.

Savage, L. 1954. *The Foundations of Statistics*. New York: Wiley.

Schmeidler, D. 1982. Subjective probability without additivity. Working paper, Foerder Institute of Economic Research, Tel Aviv University.

Schmeidler, D. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57, 571–87.

Segal, U. 1990. Two-stage lotteries without the reduction axiom. *Econometrica* 58, 349–77.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

Starmer, C. 2000. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38, 332–82.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Tversky, A. and Fox, C. 1995. Weighing risk and uncertainty. *Psychological Review* 102, 269–83.

Tversky, A. and Kahneman, D. 1992. Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5, 297–323.

Tversky, A. and Koehler, D. 1994. Support theory: a nonextensional representation of subjective probability. *Psychological Review* 101, 547–67.

Tversky, A. and Wakker, P.P. 1995. Risk attitudes and decision weights. *Econometrica* 63, 1255–80.

Wakker, P.P. 2001. Testing and characterizing properties of nonadditive measures through violations of the sure-thing principle. *Econometrica* 69, 1039–59.

Wald, A. 1950. *Statistical Decision Functions*. New York: Wiley.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Weber, E. 1994. From subjective probabilities to decision weights: the effects of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin* 115, 228–42.

Wu, G. and Gonzalez, R. 1999. Nonlinear decision weights in choice under uncertainty. *Management Science* 45, 74–85.

Yaari, M. 1969. Some remarks on measures of risk aversion and on their uses. *Journal of Economic Theory* 1, 315–29.

Yates, J. and Zukowski, L. 1976. Characterization of ambiguity in decision making. *Behavioral Science* 21, 19–25.

How to cite this article

Wakker, Peter P. "uncertainty." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000005> doi:10.1057/9780230226203.1753

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

uncovered interest parity

Peter Isard
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article provides an overview of the uncovered interest parity assumption. It traces the history of the concept, summarizes evidence on the empirical validity of uncovered interest parity, and discusses different interpretations of the evidence and the implications for macroeconomic analysis. The uncovered interest parity assumption has been an important building block in multi-period models of open economies and, although its validity is strongly challenged by the empirical evidence, at least at short time horizons, its retention in macroeconomic models is supported on pragmatic grounds by the lack of much empirical support for existing models of the exchange risk premium.

Keywords

arbitrage; capital controls; covered interest parity (CIP); credit risk; exchange market intervention; exchange rate dynamics; exchange rate expectations; exchange risk premium; forward exchange; incomplete information; inflation expectations; interest rate differentials; Jensen's inequality; Keynes, J. M.; peso problem; portfolio balance models; prediction bias; rational bubbles; rational expectations; rational learning; risk premium; spot exchange; unbiasedness hypothesis; uncovered interest parity (UIP)

Article

The assumption of uncovered interest parity (UIP) is an important building block for macroeconomic analysis of open economies. It provides a simple relationship between the interest rate on an asset denominated in any one country's currency unit, the interest rate on a similar asset denominated in another country's currency, and the expected rate of change in the spot exchange rate between the two currencies.

The theory of interest parity received prominence from expositions by Keynes (for example, 1923, pp. 115–39), whose attention had been captured by the rapid expansion of organized trading in forward exchange following the First World War (Einzig, 1962, pp. 239–41, 275). Although an understanding of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the forward exchange market must have developed within various banking circles during the second half of the 19th century, apart from an isolated exposition by a German economist, Walther Lotz (1889), the 19th-century literature on foreign exchange theory apparently dealt only with spot exchange rates (Einzig, 1962, pp. 214–15). Forward exchange trading gave rise to the notion of covered interest parity (CIP), which related the differential between domestic and foreign interest rates to the percentage difference between forward and spot exchange rates. Since it was clear that forward rates also reflected perceptions about future spot rates, it was a short step to the assumption of UIP, which builds on the theory of CIP by essentially postulating that market forces drive the forward exchange rate into equality with the expected future spot exchange rate.

Basic concepts

The concept of interest parity recognizes that portfolio investors at any time t have the choice of holding assets denominated in domestic currency, offering the own rate of interest r_t between times t and $t+1$, or of holding assets denominated in foreign currency, offering the own rate of interest r_t^*. Thus, an investor starting with one unit of domestic currency should compare the option of accumulating $1+r_t$ units with the option of converting at the spot exchange rate into s_t units of foreign currency, investing in foreign assets to accumulate $s_t(1+r_t^*)$ units of foreign currency at time $t+1$, and then reconverting into domestic currency. If the domestic and foreign assets differ only in their currencies of denomination, and if investors have the opportunity to cover against exchange rate uncertainty by arranging at time t to reconvert from foreign to domestic currency one period later at the forward exchange rate f_t (in units of foreign currency per unit of domestic currency), then market equilibrium requires the condition of CIP:

$$1 + r_t = s_t(1 + r_t^*) / f_t. \quad (1)$$

If condition (1) did not hold, profitable market arbitrage opportunities could be exploited without incurring any risks.

Investors also have the opportunity to leave their foreign currency positions uncovered at time t and to wait until time $t+1$ to make arrangements to reconvert into domestic currency at the spot exchange rate s_{t+1}. Unlike f_t, the value of s_{t+1} is unknown at time t, and so the attractiveness of holding an uncovered position must be assessed in terms of the probabilities of different outcomes for s_{t+1}. The assumption of UIP postulates that markets will equilibrate the return on the domestic currency asset with the expected value at time t (E_t) of the yield on an uncovered position in foreign currency:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$1 + r_t = E_t[s_t(1 + r_t^*) / s_{t+1}] = s_t(1 + r_t^*)E_t(1 / s_{t+1}).$$
(2)

This is essentially equivalent to combining the CIP condition with the assumption that exchange rates are driven, at the margin, by risk-neutral market participants who stand ready to take uncovered spot or forward positions whenever the forward rate deviates from the expected future spot rate.
By manipulating condition (1), it is easily seen that CIP implies

$$\frac{f_t - s_t}{s_t} = \frac{1 + r_t^*}{1 + r_t} - 1$$
(3)

Hence, as a first approximation (for values of $1 + r_t$ in the vicinity of 1):

$$r_t^* - r_t \approx (f_t - s_t) / s_t$$
(4)

In addition, when Jensen's inequality – that is, the difference between $E_t(1 / s_{t+1})$ and $1 / E_t(1 / s_{t+1})$ – is ignored, the assumption of UIP can be approximated as

$$r_t^* - r \approx E_t[(s_{t+1} - s_t) / s_t] = (E_t s_{t+1} - s_t) / s_t.$$
(5)

The assumption of UIP adds an element of dynamics to the CIP condition by hypothesizing a relationship between the observed values of variables at time *t* and the value of the spot exchange rate that market participants expect at time *t* to prevail at time $t + 1$. As such, UIP has been embedded in many multi-period models of open economies. The CIP and UIP conditions can be written for any duration of the time period between *t* and $t + 1$. Thus, if the UIP assumption was valid at all horizons, the observed values of the spot exchange rate and the term structures of domestic and foreign interest rates could be used to infer the expected future time path of the spot exchange rate (Porter, 1971).
In addition to playing an important role in the development of multi-period models of open economies,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the UIP condition has been a central focal point in the policy debate over the effectiveness of official intervention in exchange markets (Henderson and Sampson, 1983). To the extent that UIP was valid at short time horizons, official intervention could not succeed in changing the spot exchange rate relative to the expected future spot rate unless the authorities chose to allow interest rates to change. In this sense, exchange market intervention could not be viewed as providing the authorities with an effective policy instrument in addition to interest rates. Thus, the case for intervention has been considered by some to depend on whether the empirical evidence rejects UIP.

Empirical evidence

The theory leading to the CIP condition – and hence also to the UIP assumption – abstracts entirely from any credit risks, capital controls, or explicit taxes on domestic and foreign currency investments. Keynes (1923, pp. 126–7) was well aware that investor choices between foreign and domestic assets do not depend on interest rates and exchange rates alone:

> … the various uncertainties of financial and political risk … introduce a further element which sometimes quite transcends the factor of relative interest. The possibility of financial trouble or political disturbance, and the quite appreciable probability of a moratorium in the event of any difficulties arising, or of the sudden introduction of exchange regulations which would interfere with the movement of balances out of the country, and even sometimes the contingency of a drastic demonetization, – all these factors deter … [market participants], even when the exchange risk proper is eliminated, from maintaining large … balances at certain foreign centres.

In those circumstances where it is valid to abstract from the types of considerations cited by Keynes, the CIP condition has been generally confirmed. As one source of evidence, interviews at large banks have established that the CIP condition is used as a formula for determining the exchange rates and interest rates at which trading is actually conducted. Foreign exchange traders use Eurocurrency interest rate differentials to determine the forward exchange rates (in relation to spot rates) that they quote to customers, while traders in Eurocurrency deposits use the spreads between forward and spot exchange rates to set the spreads between the interest rates that their banks offer on domestic and foreign currency deposits (Herring and Marston, 1976; Levich, 1985). As additional evidence, Taylor (1989) has constructed a database of the bid and offer rates quoted contemporaneously for exchange rates and interest rates by foreign exchange and money market brokers, as recorded on the ‘pad’ of the chief dealer at the Bank of England. The data include observations on one- two-, three-, six-, and twelve-month maturities during selected intervals between 1967 and 1987. Taylor's study found no evidence of unexploited profit opportunities during relatively calm periods in foreign exchange and money markets, although potentially exploitable profitable arbitrage opportunities did ‘occasionally occur’ during periods of market turbulence, where the frequency, size and persistence of such opportunities were positively related to length of maturity. Consistently, in circumstances when it is not valid to abstract from capital controls and risks, empirical research has confirmed that deviations from CIP can be related systematically to the effective taxes imposed by capital controls and to non-currency-specific risk

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
premiums associated with prospective controls (Dooley and Isard, 1980).

The UIP assumption is more difficult to test than the CIP condition, since market expectations of future exchange rates are not directly observable. Accordingly, UIP has generally been tested jointly with the assumption that exchange market participants form rational expectations, such that future realizations of the exchange rate will equal the value expected at time *t* plus an error term that is uncorrelated with all information known at time *t*. Together the two assumptions imply that

$$s_{t+1} = f_t + u_{t+1} \tag{6}$$

and hence

$$s_{t+1} - s_t = r_t - r_t^* + u_{t+1} \tag{7}$$

where *u* represents a prediction error. This has led economists to assess the UIP assumption empirically by estimating the values of the *a* and *b* coefficients in the specification forms

$$s_{t+1} = a_0 + a_1 f_t + u_{t+1} \tag{8}$$

and

$$s_{t+1} - s_t = b_0 + b_1(r_t - r_t^*) + u_{t+1} \tag{9}$$

where it is assumed that the error terms have zero means and are serially uncorrelated.

Empirical assessments of UIP as a framework for predicting the future spot exchange rate have distinguished two issues: the size of the prediction errors, and the question of whether the predictions are systematically biased. On the first issue, it has become widely acknowledged that interest differentials

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
explain only a small proportion of subsequent changes in exchange rates (Isard, 1978; Mussa, 1979; Frenkel, 1981). This finding has been generally interpreted as implying that observed changes in exchange rates are predominantly the result of unexpected information or 'news' about economic developments, policies or other relevant factors.

The issue of whether predictions are systematically biased can be assessed by testing the hypothesis of unbiasedness – namely, that $(a_0, a_1) = (0, 1)$ in eq. (8) or $(b_0, b_1) = (0, 1)$ in eq. (9). Notably, the test that the slope coefficient is unity receives strong support from studies based on (8) but is soundly rejected by studies based on (9) – at least for prediction horizons of a year or less. However, the apparent conflict between the two sets of regression evidence has been resolved in favour of the latter finding, as it is now accepted that (8) is not a legitimate regression equation (Meese, 1989). The explanation is based on the fact that the sample variances of the spot rate and forward rate are essentially equal.

Although the empirical evidence strongly rejects the unbiasedness hypothesis at prediction horizons of up to one year, the evidence is much more favourable to unbiasedness at horizons of 5 to 20 years. In particular, when data for industrial countries are pooled, and when annual exchange rate changes and interest differentials (for each country relative to a numeraire country) are averaged over non-overlapping 5- to 20-year periods, the slope coefficients in eq. (9) become insignificantly different from unity (Flood and Taylor, 1997, who note that the average one-year change over n years is equivalent to the change over n years multiplied by a scale factor; see also Chinn and Meredith, 2004).

Does prediction bias refute the UIP assumption?

Economists have not resolved how to interpret the strong rejection of the unbiasedness hypothesis at short prediction horizons. Several possible explanations have been suggested, with different implications for UIP.

One interpretation rejects the UIP hypothesis but not the rational expectations assumption. According to this view, the finding of systematic prediction bias suggests that market participants are risk averse and require risk premiums to hold uncovered foreign currency positions. The prediction bias is thus perceived as an omitted variable problem that can be addressed, in concept, by extending the right-hand side of eq. (9) to include an expression for the risk premium. A second interpretation of prediction bias abandons the assumption that market participants are fully rational.

Other possible explanations do not require rejection of either UIP or the rational expectations hypothesis. These include explanations based on the 'peso problem', simultaneity bias, incomplete information with rational learning, and self-fulfilling prophecies or rational 'bubbles'.

The suggestion that prediction bias reflects a 'peso problem' is generally attributed to Rogoff (1980) and Krasker (1980), who drew attention to an episode in which the Mexican peso sold at a forward discount for a prolonged period prior to its widely anticipated devaluation in 1976. Although market expectations eventually proved correct and may well have been rational *ex ante*, the fact that the devaluation did not occur immediately after it became anticipated made the forward rate a biased predictor over finite data samples that included the pre-devaluation period. The general point is that, even if market participants are risk neutral and form rational expectations, the forward rate can be biased as a predictor of the future spot rate – and the interest rate differential biased as a predictor of the change in the spot rate – whenever market participants repeatedly expect the spot rate to change in response to a policy action or

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
some other event that fails to materialize over a relatively long series of observations.
The suggestion that rejection of the unbiasedness hypothesis reflects simultaneity bias was alluded to by Isard (1988) and later emphasized by McCallum (1994). In particular, given that the monetary authorities in most countries rely on a short-term interest rate as a policy instrument that they are prepared to adjust, *inter alia*, in response to undesired exchange rate movements, the estimates of b_1 may be biased by the failure to estimate (9) simultaneously with a second relationship between the interest rate differential and the change in the exchange rate.
As suggested by Lewis (1988; 1989), prediction bias can also emerge under UIP and rational expectations if market participants lack complete information but engage in a process of rational learning. This explanation is analogous to the peso problem in so far as it provides an interpretation in which market participants are risk neutral and fully rational but prone to make repeated mistakes.
Yet another possibility consistent with UIP is the conjecture that prediction bias arises from the self-fulfilling prophecies of rational, risk-neutral market participants. Such prophecies, which are often referred to as ‘rational bubbles’, have received attention as logical possibilities; but few economists, if any, consider them to have much plausibility as empirical phenomena (Mussa, 1990).

Where things stand

Because the validity of the UIP hypothesis cannot be tested directly and is not resolved by the rejection of the unbiasedness hypothesis, economists have resorted to indirect tests as a means of obtaining suggestive evidence. In particular, survey data on exchange rate expectations have been collected by several different sources since the early 1980s, and a number of studies have shown that exchange rate expectations, as measured by the average forecasts of sample respondents, deviate considerably from prevailing forward exchange rates (Frankel and Froot, 1987; Takagi, 1991; Chinn and Frenkel, 2002). To the extent that survey measures of average expectations are meaningful, this would appear to be strong evidence against UIP.
That said, it also needs to be recognized that intertemporal models of open-economy macroeconomics require equations that link current spot exchange rates to expected future exchange rates. Thus, on pragmatic grounds, the case for abandoning the UIP hypothesis depends on how well economists can model the deviation from UIP – namely, the difference between the forward exchange rate and the expected future spot rate, which is generally referred to as the exchange risk premium.
Behavioural hypotheses about the exchange risk premium can be tested by embedding them in models of observable exchange rates. The first conceptual models of the exchange risk premium were based on a portfolio balance framework in which financial claims were distinguished by currencies of denomination but not by the countries obligated to meet the claims (see, for example, Dooley and Isard, 1983). Empirical tests of this class of portfolio balance model have explained at most a small portion of the variation over time in the exchange risk premium (Tryon, 1983; Boughton, 1987). More sophisticated behavioural hypotheses have recognized – in the spirit of the quotation above from Keynes – that exchange risks and credit risks are interrelated, and that the magnitudes of these risks reflect the relative macroeconomic and political conditions, prospects, and uncertainties of the countries that have issued the portfolio claims (Dooley and Isard, 1983; Isard, 1988). While casual evidence suggests that this type of hypothesis is broadly capable of explaining the empirical behaviour of exchange rates (Dooley and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Isard, 1991), formal empirical tests that capture the many factors contributing to exchange rate risk are difficult to design, and economists have not yet provided a well-specified replacement for the UIP assumption.

Accordingly, many intertemporal open-economy macroeconomic models continue to impose the UIP assumption – or the assumption of UIP adjusted by an exogenous exchange risk premium that provides a mechanism for analyzing the effects of 'exogenous' changes in risk perceptions or asset preferences. However, consistent with the evidence that rejects the unbiasedness hypothesis, it has proved difficult to mimic the observed behaviour of key macroeconomic variables with models that impose the UIP assumption and also treat exchange rate expectations as fully model-consistent. Thus, models that impose the UIP assumption tend to treat exchange rate expectations as not completely rational. One fairly common practice, for example, is to treat exchange rate expectations (and inflation expectations) as having both forward-looking (model-consistent) and backward-looking components.

Quite apart from ongoing debates over the validity of the UIP assumption as an *ex ante* hypothesis, and the usefulness of incorporating the UIP assumption into macroeconomic models, there is abundant evidence, as noted above, that the changes in spot exchange rates that are expected *ex ante* are generally dominated by unexpected changes. Thus, regardless of the usefulness of UIP as an *ex ante* hypothesis for macroeconomic modelling, it is quite clear that UIP by itself provides a very inaccurate framework for predicting the changes in exchange rates that are observed *ex post*.

See Also

- exchange rate dynamics

This article draws extensively on Isard (1991; 1995). *The views expressed are those of the author and do not necessarily reflect those of the International Monetary Fund*.

Bibliography

Boughton, J. 1987. Tests of the performance of reduced-form exchange rate models. *Journal of International Economics* 23, 41–56.

Chinn, M. and Frenkel, J. 2002. Survey data on exchange rate expectations: more currencies, more horizons, more tests. In *Monetary Policy, Capital Flows, and Financial Market Developments in an Era of Financial Globalization: Essays in Honour of Max Fry*, ed. W. Allen and D. Dickinson. London: Routledge.

Chinn, M. and Meredith, G. 2004. Monetary policy and long horizon uncovered interest parity. *IMF Staff Papers* 51, 409–30.

Dooley, M. and Isard, P. 1980. Capital controls, political risk and deviations from interest-rate parity. *Journal of Political Economy* 88, 370–84.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dooley, M. and Isard, P. 1983. The portfolio-balance model of exchange rates and some structural estimates of the risk premium. *IMF Staff Papers* 30, 683–702.

Dooley, M. and Isard, P. 1991. A note on fiscal policy, investment location decisions, and exchange rates. *Journal of International Money and Finance* 10, 161–8.

Einzig, P. 1962. *The History of Foreign Exchange*. London: Macmillan.

Flood, R. and Taylor, M. 1997. Exchange rate economics: what's wrong with the conventional macro approach? In *The Microstructure of Foreign Exchange Markets*, ed. J. Frankel, G. Galli and A. Giovannini. Chicago: University of Chicago Press.

Frankel, J. and Froot, K. 1987. Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review* 77, 133–53.

Frenkel, J. 1981. Flexible exchange rates, prices and the role of ‘news’: lessons from the 1970s. *Journal of Political Economy* 89, 665–705.

Henderson, D. and Sampson, S. 1983. Intervention in foreign exchange markets: a summary of ten staff studies. *Federal Reserve Bulletin* 69, 830–36.

Herring, R. and Marston, R. 1976. The forward market and interest rates in the Eurocurrency and national money markets. In *Eurocurrencies and the International Monetary System*, ed. C. Stern, J. Makin and D. Logue. Washington, DC: American Enterprise Institute.

Isard, P. 1978. *Exchange-Rate Determination: A Survey of Popular Views and Recent Models*. Princeton Studies in International Finance No. 42. Princeton: International Finance Section, Department of Economics, Princeton University.

Isard, P. 1988. Exchange rate modeling: an assessment of alternative approaches. In *Empirical Macroeconomics for Interdependent Economies*, ed. R. Bryant et al. Washington, DC: Brookings Institution.

Isard, P. 1991. Uncovered interest parity. In *The New Palgrave Dictionary of Money and Finance*, 1st edn, ed. P. Newman, M. Milgate and J. Eatwell. Basingstoke: Palgrave Macmillan. Also issued as Working Paper WP/91/51, Washington, DC: International Monetary Fund, 1991.

Isard, P. 1995. *Exchange Rate Economics*. Cambridge: Cambridge University Press.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. London: Macmillan.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Krasker, W. 1980. The peso problem in testing the efficiency of forward exchange markets. *Journal of Monetary Economics* 6, 269–76.

Levich, R. 1985. Empirical studies of exchange rates: price behavior, rate determination and market efficiency. In *Handbook of International Economics*, vol. 2, ed. R. Jones and P. Kenen. Amsterdam: North-Holland.

Lewis, K. 1988. The persistence of the ‘peso problem’ when policy is noisy. *Journal of International Money and Finance* 7, 5–21.

Lewis, K. 1989. Changing beliefs and systematic rational forecast errors with evidence from foreign exchange. *American Economic Review* 79, 621–36.

Lotz, W. 1889. Die Währungsfrage in Österreich-Ungarn. *Schmollers Jahrbuch* 13, 34–5.

McCallum, B. 1994. A reconsideration of the uncovered interest parity relationship. *Journal of Monetary Economics* 33, 105–32.

Meese, R. 1989. Empirical assessment of foreign currency risk premiums. In *Financial Risk: Theory, Evidence and Implications*, ed. C. Stone. Boston: Kluwer.

Mussa, M. 1979. Empirical regularities in the behavior of exchange rates and theories of the foreign exchange market. *Carnegie-Rochester Conference Series on Public Policy* 11, 9–57.

Mussa, M. 1990. *Exchange Rates in Theory and Reality*. Essays in International Finance No. 179. Princeton: International Finance Section, Department of Economics, Princeton University.

Porter, M. 1971. A theoretical and empirical framework for analyzing the term structure of exchange rate expectations. *IMF Staff Papers* 18, 613–42.

Rogoff, K. 1980. An empirical investigation of the martingale property of foreign exchange futures prices. Chapter 1 of Essays on expectations and exchange rate volatility. Doctoral dissertation, Cambridge, MA: MIT. Online. Available at http://post.economics.harvard.edu/faculty/rogoff/papers/ChapterOne.pdf, accessed 4 May 2006.

Takagi, S. 1991. Exchange rate expectations: a survey of survey studies. *IMF Staff Papers* 38, 156–83.

Taylor, M. 1989. Covered interest arbitrage and market turbulence. *Economic Journal* 99, 376–91.

Tryon, R. 1983. Small empirical models of exchange market intervention: a review of the literature. Staff Studies No. 134. Washington, DC: Board of Governors of the Federal Reserve System.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Isard, Peter. "uncovered interest parity." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000008>
doi:10.1057/9780230226203.1755

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

underconsumptionism

Michael Schneider
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

acceleration principle; aggregate demand; Baran, P. A.; Catchings, W.; Clark, J. B.; disproportionate production; Emmanuel, A.; Harrod–Domar growth model; Hoarding; Hobson, J. A.; imperialism; Kautsky, K.; Keynesianism; Lauderdale, Eighth Earl of; Lenin, V. I.; Luxemburg, R.; Malthus, T. R.; Marx, K. H.; redistribution of income; Rodbertus, J. K.; saving equals investment; Sismondi, J. C. L. S. de; Sweezy, P.M.; underconsumption; underconsumptionism; wages fund

Article

‘Underconsumption’ is the label given to theories which attribute the failure of the total output of an economy to continue to be sold at its cost of production (including normal profit) to too low a ratio of consumption to output. According to underconsumption theories, such deficient consumption leads either to goods being able to be sold only at below-normal rates of profit, or to goods not being able to be sold at all. These effects are seen as leading in turn to cutbacks in production and increases in unemployment. Underconsumption theories are thus amongst those which seek to explain cyclical or secular declines in the rate of economic growth.

Where underconsumption exists in the sense that the ratio of consumption to output is below the optimum level, it follows that the ratio of ‘unconsumed’ output to total output must be too high. For the period in which underconsumption breaks out, underconsumptionists in general both identify this ‘unconsumed’ output with saving, and equate saving with investment. Thus Haberler, to whose 1937 analysis of underconsumption and related theories the reader should turn for the best extended treatment of the subject, wrote that in ‘its best-reasoned form …, the under-consumption theory uses “under-consumption” to mean “over-saving”•’, and that in ‘the under-consumption or over-saving theory … savings are, as a rule, invested …’ (Haberler, 1937, pp. 115 and 117).

While the theories advanced by underconsumptionists overlap with some other macroeconomic theories in certain ways, their basic characteristics make them distinct in other respects. Underconsumption theories share with Keynesian theories, for example, the characteristic that they are ‘demand-side’ (as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
opposed to ‘supply side’) theories. However, there is a fundamental difference between the two, in that Keynesian theories attribute the failure of total output to reach the full employment level to a deficiency of *aggregate* demand. The two types of theory consequently have different implications. As Robbins succinctly put it, with reference to the underconsumptionist J.A. Hobson, for ‘Mr Keynes, one way out of the slump would be a revival of investment; for Mr Hobson, this would simply make matters worse’ (Robbins, 1932, p. 420). A further difference between the two lies in the fact that, by contrast with Keynesian theories, hoarding plays no part in underconsumption theories. Underconsumptionists in general confine their analyses to the real sector of the economy, and where monetary factors are discussed at all they are treated as secondary.

There are also some connections between underconsumption theories and the accelerator theory of investment. As Haberler pointed out, the acceleration principle can be used ‘in support of a special type of the under-consumption theory of the business cycle’ (Haberler, 1937, p. 30). More importantly, a variant of the principle can be seen as underlying all underconsumption theories. When first expounded by J.M. Clark, the acceleration principle was used to explain the level of activity in the investment goods sector of an economy by changes in the demand for finished goods. In essence, underconsumptionists base their theories on the idea that changes in the demand for consumption goods determine the future level of activity in the investment goods sector. By this means they draw the conclusion that the level of activity in the economy as a whole is wholly determined by consumption demand.

In a review of Harrod's *Towards a Dynamic Economics*, Joan Robinson suggested that ‘Mr Harrod's analysis provides the missing link between Keynes and Hobson’ (Robinson, 1949, p. 80). The resemblance of underconsumption theories to growth models of the Harrod–Domar type is in fact greater than their resemblance to Keynesian theories. As Domar pointed out in the *American Economic Review* article (1947) expounding his growth model, he shared with Hobson a concern with the capacity-creating effect of investment, a question which Keynes hardly touched on in the *General Theory*. The essential features of underconsumption theories can in fact be captured by a growth model of the Harrod–Domar type, in which however the driving force is provided not by the rate of growth of investment but by the rate of growth of consumption. Such a model is particularly appropriate in the case of those theories which treat underconsumption as a secular rather than a cyclical phenomenon.

There are connections too between underconsumption theories and explanations of ‘economic crises’ in terms of ‘disproportionate production’, to use Marx's terminology. By ‘disproportionate production’ Marx meant an allocation of labour time between sectors or industries other than that required to satisfy social need as reflected in demand. Now underconsumption involves an allocation of too few resources to the consumption goods sector and too many resources to the investment goods sector. But as Haberler pointed out, such ‘vertical disproportion’ should be distinguished from ‘horizontal disproportion’. And unlike cases of horizontal disproportion (if optimal stock levels are ignored), vertical disproportion, involving industries not equidistant from consumption goods industries, cannot be rectified immediately by a return to ‘proportionate’ production. For the excessive production of investment goods consequent upon underconsumption leaves a legacy in the form of excessive productive capacity.

Underconsumption theories thus should be distinguished from the more general category of ‘disproportionality’ theories; the disproportionality element they incorporate is specific and has distinctive consequences.

Over-investment theories provide a different example of vertical disproportion. As they are defined by Haberler, over-investment theories offer an explanation of the excessive aggregate demand characteristic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of an economy during the upswing of a trade cycle. Therefore they also belong to a different category from underconsumption theories, even though the deficiency of consumption characteristic of the latter is accompanied by excessive investment.

Despite their basic similarities, underconsumption theories differ as to the cause of, and hence remedies for, underconsumption. A view to be found in the writings of some underconsumptionists, especially the less well-known ones, is that underconsumption is due to total purchasing power falling short of the value of output. Since all the value of output accrues to the owner of one factor of production or another, this proposition as it stands cannot be sustained. This view as to the cause of underconsumption should not be confused, however, with a superficially similar view relating to its effects, which is at least implicit in all underconsumptionist thinking. This is the idea that income is generated not by production but by purchases of what is produced. In underconsumptionist writings, by contrast for example with Keynesian writings, income may fall short of the value of output.

How this may be so is perhaps best seen in terms of period analysis. An outbreak of underconsumption will lead in the first period to excessive saving accompanied by excessive investment. In the second period the resulting additional capacity will be used, and unless there is an increase in consumption the level of output will exceed the demand for it; hence in this period the income generated by purchases will fall short of the value of output, while at the same time saving, if it is defined as that part of income (as opposed to output) not consumed, will just match investment demand. The deficient demand in the second period will lead in the third period to actual output falling short of potential output, that is to excess capacity, with saving however continuing to equal investment.

One underconsumptionist who clearly did not attribute underconsumption to lack of purchasing power is Malthus. In his correspondence with Ricardo, Malthus instead took the position that ‘a nation must certainly have the power of purchasing all that it produces, but I can easily conceive it not to have the will’ (Sraffa, 1952, p. 132). Like some other underconsumptionists, notably Sismondi and Hobson, Malthus believed that one cause of underconsumption is to be found in the limited capacity of human beings to expand their wants, at least in the short run. It was Malthus's view that men have a tendency towards indolence once their needs for necessaries are satisfied. If in the face of such limited growth in human wants capital accumulation continues apace, the resulting increase in productive capacity will fail to be matched by an equal increase in consumer demand. The remedy for this state of affairs, suggested Malthus, is an increase in commerce, both domestic and foreign, so as to stimulate tastes by exposing the population to new products.

Most commonly, however, underconsumptionists find the cause of underconsumption in a maldistribution of income. The underlying argument is simple. If different economic classes have different propensities to consume, the distribution of an excessive share of income to classes with a relatively low propensity to consume will result in underconsumption. Underconsumptionists agree that a remedy cannot be found by a redistribution of income towards the capitalist class, which they see as having a relatively high propensity to save. They differ, however, on the question of the class to which income should in cases of underconsumption be redistributed. The earliest underconsumptionists ruled out a redistribution of income towards workers, perhaps partly because it was incompatible with their adherence to the wages fund doctrine, according to which total wages are fixed by the capital set aside in advance to pay them. They advocated rather a redistribution of income towards landlords. The first underconsumptionist to advocate a redistribution of income towards workers was Sismondi, whose example was followed by most later underconsumptionists.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Underconsumption theories were first put forward in the 19th century. While some 17th- and 18th-century writers, most notably Mandeville and the Physiocrats, advocated an increase in expenditure on consumption goods, none of them linked this with a corresponding reduction in investment. Therefore although they may be seen as predecessors of Keynes, they should not be classified as underconsumptionists. The first to advance an underconsumption theory in the sense outlined above was Lauderdale, in *An Inquiry into the Nature and Origin of Public Wealth* (1804). Perhaps the best known of the subsequent underconsumptionists are Malthus, Sismondi, Rodbertus, Hobson and Rosa Luxemburg. For a fuller (and sometimes different) account of the theories of these and other underconsumptionists than is possible here, the reader should turn to Bleaney (1976) or Nemmers (1956). Further, additional light has been shed on the overall nature of underconsumptionist theories by the several attempts that have been made to express the theory put forward by Malthus in the form of a model, notably by Eagly (1974), Eltis (1980) and Costabile and Rowthorn (1985).

While some underconsumption theories were largely prompted by current or expected economic events, in other cases the inspiration was mainly intellectual. Both factors seem to have been important in the case of Lauderdale, the earliest underconsumptionist. Lauderdale was in part reacting against the praise of parsimony by Adam Smith, but he was also alarmed at the prospect of the British government using its revenue after the end of the Napoleonic wars for the purpose of capital accumulation in place of wartime consumption. More generally, as a precaution against underconsumption, Lauderdale advocated a lessening of the current inequality of wealth, as Malthus was also to do. By contrast Spence, in *Britain Independent of Commerce* (1807), developed an underconsumption theory on the basis of Physiocratic ideas. His solution for underconsumption was encouragement of consumption by landlords, so as to restore the income of the manufacturing class to its former level.

His correspondence with Ricardo shows that Malthus had developed underconsumptionist views by 1814. This fact is doubly significant. It proves both that Malthus's underconsumptionism preceded the depressed economic conditions which followed the ending of the Napoleonic wars in 1815, rather than being a response to them, and that Marx's charge that Malthus plagiarized Sismondi is unfounded. The underconsumptionist elements in Malthus's thinking are to be found not only in his correspondence with Ricardo, but also in his *Principles of Political Economy Considered with a View to their Practical Application* (1820). The latter had an influence on the underconsumption theory put forward in Chalmers' *Political Economy* (1832). It may also have provided a stimulus for the underconsumption theory advanced in a pamphlet entitled *Considerations on the Accumulation of Capital* (1822). Published anonymously, this pamphlet was written by Cazenove, the friend of Malthus who was later to edit (also anonymously) the second edition of Malthus's *Principles*.

Like Lauderdale, Sismondi reacted against Adam Smith's views on parsimony, and like Malthus he had become an underconsumptionist by the end of the Napoleonic wars, as is evidenced by the material contained in the article entitled 'Political Economy' which Sismondi wrote in 1815 for Brewster's *Edinburgh Encyclopaedia*. A complete account of Sismondi's underconsumption theory is only to be found, however, in his *Nouveaux principes d'économie politique* (1819). Here Sismondi argued that where producers supply a large anonymous market, competition for profits leads each of them on the one hand to overestimate the demand for the commodity he produces and over-accumulate capital accordingly, and on the other hand so to depress wages that they grow at a slower rate than profits. Sismondi's remedies for underconsumption include organization of industry on a local basis, and a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
redistribution of income towards wages.

For a discussion of possible sources of the underconsumptionist elements in the writings of Robert Owen and the Ricardian Socialists the reader is referred to King (1981). A more comprehensive underconsumption theory than in those writings is to be found in Rodbertus's ‘second letter’ to von Kirchmann, published in 1850–51. Rodbertus was reacting against the ideas of Jean-Baptiste Say and his followers. His own view was that in a laissez-faire economy underconsumption must inevitably emerge and worsen, because ‘natural’ laws will ensure that an ever increasing productivity of labour will be accompanied by an ever decreasing share of income going to wages. His remedy was ‘rational’ intervention in the economy to counteract these ‘natural’ laws.

The emphasis in Marx's economic theory on the necessity in a capitalist economy for value not only to be generated in production but also realized by sale makes that theory well adapted to use to the development of an underconsumption theory. Marx himself gave substantial praise to Sismondi for his exposition of such a theory, and there are several passages in Marx's own writings which put forward an underconsumptionist view. On the other hand, there is a well-known passage in Volume 2 of *Capital* which condemns underconsumption theories in no uncertain terms, and in any case there are other elements in his economic theory which are so much more important to Marx that he is not usually classified as an underconsumptionist. Many, though by no means all, of his followers have in fact condemned underconsumption theories. Examples of such condemnation are to be found in some of Lenin's writings, notably his pamphlet entitled *A Characterisation of Economic Romanticism* (*Sismondi and our Native Sismondists*), written in 1897. This pamphlet was particularly directed at the underconsumptionist views of the Russian ‘Populists’, or ‘Narodniks’, who had argued that capitalism could not survive in Russia without the consumer markets provided by its then dwindling peasant economy. It was Lenin's view that for the development of capitalism expansion of the market for investment goods is more important than expansion of the market for consumption goods.

Amongst Marx's earlier followers, those who most strongly supported the underconsumptionist element in Marx's thinking were Kautsky and Rosa Luxemburg. Rosa Luxemburg's main arguments were set out in *The Accumulation of Capital* (1913). Contrasting the over-growing generation of value in a capitalist economy with the inability of workers and unwillingness of capitalists to realize that value by increasing their consumption, she crossed swords with Tugan Baranovski, who had argued that capitalists ‘see to it that ever more machines are built for the sake of building – with their help – ever more machines’ (Luxemburg, 1913, p. 335). Rosa Luxemburg took the same view as that advanced by J.B. Clark, in his introduction to the English translation of Rodbertus's ‘second letter’ to von Kirchmann, namely that ‘this case presents no glut: but it is an unreal case’ (Rodbertus, 1898, p. 15). She concluded that because it was inevitably faced by increasing underconsumption, a capitalist economy could only survive as long as it was able to dispose of its surplus to non-capitalist consumers, either at home or abroad, the latter accounting in her view for policies of imperialism. Apart from Rosa Luxemburg, others who have both drawn on Marx's ideas and made use of underconsumption theory include Sweezy in *The Theory of Capitalist Development* (1942), Baran and Sweezy in *Monopoly Capital* (1966) and Emmanuel in *Unequal Exchange* (1969).

A causal connection between underconsumption and policies of imperialism was also argued to exist by the non-Marxist writer J.A. Hobson, in *Imperialism: A Study* (1902). Jointly with A.F. Mummery, Hobson had reacted to the depression in trade in the 1880s by putting forward an underconsumption

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
theory in *The Physiology of Industry* (1889), which was the first underconsumptionist work actually to use the term ‘underconsumption’. In this book Mummery and Hobson argued that the sole source of demand for investment goods is demand for consumption goods. From this they drew the conclusion, as Malthus had done, that there exists an optimum ratio between saving (investment) and spending (consumption). Like Sismondi, they stressed the role of competition in causing supply to exceed demand. They went beyond the earlier underconsumptionists, however, in specifically arguing that neither a fall in the rate of interest nor a fall in the price level could remedy a state of depression brought about by underconsumption. Hobson's subsequent restatement of this theory, with various amplifications, made him the most influential 20th-century exponent of underconsumption theories.

In *The Physiology of Industry* Mummery and Hobson drew the policy conclusion that ‘where Under-consumption exists, Savings should be taxed’ (Mummery and Hobson, 1889, p. 205). In his later works, however, from *The Problem of the Unemployed* (1896) on, Hobson laid most stress on a redistribution of income from what he called ‘unearned income’ (income unrelated to effort) to wages as the main remedy for underconsumption. The most comprehensive expositions of Hobson's underconsumption theory are to be found in *The Industrial System* (1909), which is characterized by a more extensive treatment of underconsumption in a growing economy, *The Economics of Unemployment* (1922), and *Rationalisation and Unemployment* (1930).

Other 20th-century exponents of underconsumption theories include Foster and Catchings, in a number of jointly written books. The theories of Major Douglas, however, with their lack of reference to over-investment and their emphasis on the role of money and credit, do not fit well into the underconsumptionist category.

Underconsumption theories have never been acceptable to orthodox economists, perhaps partly because underconsumptionists in general have lacked rigour in the exposition of their ideas, and partly because underconsumption theories have been seen as a threat to the saving necessary for economic growth in particular, and to capitalism in general. They have also attracted less attention since 1936 than before, because Keynes's *General Theory* satisfied the needs of many of those whose intuitions led them to seek a ‘demand-side’ explanation of economic depression. However, underconsumption theories can be argued still to provide a useful supplement to Keynesian theories, as a reminder that there is a limit to the extent to which employment can be increased by increases in investment alone. There is perhaps some recognition of this in the distinction which is now commonly made as to whether the current need is for an ‘investment-led’ or a ‘consumption-led’ recovery.

See Also

- Hobson, John Atkinson
- Keynes, John Maynard

Bibliography

Baran, P.A. and Sweezy, P.M. 1966. *Monopoly Capital*. New York: Monthly Review Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Bleaney, M. 1976. *Underconsumption Theories: A History and Critical Analysis*. London: Lawrence & Wishart.

Cazenove, J. 1822. *Considerations on the Accumulation of Capital and its Effects on Profits and on Exchangeable Value*. London: J.M. Richardson. Published anonymously.

Chalmers, T. 1832. *On Political Economy, in Connexion with the Moral State and Moral Prospects of Society*. Glasgow: William Collins.

Costabile, L. and Rowthorn, R.E. 1985. Malthus's theory of wages and growth. *Economic Journal* 95, 418–37.

Domar, E.D. 1947. Expansion and employment. *American Economic Review* 37, 34–55.

Eagly, R.V. 1974. *The Structure of Classical Economic Theory*. Oxford: Oxford University Press.

Eltis, W.A. 1980. Malthus's theory of effective demand and growth. *Oxford Economic Papers* 32, 19–56.

Emmanuel, A. 1969. *Unequal Exchange: A Study of the Imperialism of Trade*. New York: Monthly Review Press, 1972.

Haberler, G. 1937. *Prosperity and Depression*. Geneva: League of Nations.

Hobson, J.A. 1896. *The Problem of the Unemployed*. London: Methuen.

Hobson, J.A. 1902. *Imperialism: A Study*. London: Nisbet.

Hobson, J.A. 1909. *The Industrial System*. New York: Longmans.

Hobson, J.A. 1922. *The Economics of Unemployment*. London: Allen & Unwin.

Hobson, J.A. 1930. *Rationalisation and Unemployment*. London: Allen & Unwin.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

King, J.E. 1981. Perish commerce! Free trade and underconsumption in early British radical economics. *Australian Economic Papers* 20(37), 235–57.

Lauderdale, Eighth Earl of. 1804. *An Inquiry into the Nature and Origin of Public Wealth*. Edinburgh. Reprinted, New York: Augustus M. Kelley, 1962.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lenin, V.I. 1897. A characterisation of economic romanticism (Sismondi and our native Sismondists). In V.I. Lenin, *Collected Works*, vol. 2, Moscow: Foreign Languages Publishing House, 1962.

Luxemburg, R. 1913. *The Accumulation of Capital*. Trans. A., Schwarzschild, with an introduction by J., Robinson, London: Routledge & Kegan Paul, 1951.

Malthus, T.R. 1820. *Principles of Political Economy Considered with a View to their Practical Application*. London: Murray.

Malthus, T.R. 1836. *Principles of Political Economy Considered with a View to their Practical Application*. 2nd edition. London: William Pickering.

Marx, K. 1885. *Capital*, vol. 2. Moscow: Foreign Languages Publishing House, 1957.

Mummery, A.F. and Hobson, J.A. 1889. *The Physiology of Industry*. London: Murray.

Nemmers, E.E. 1956. *Hobson and Underconsumption*. Amsterdam: North-Holland.

Robbins, L. 1932. Consumption and the trade cycle. *Economica* 12, 413–30.

Robinson, J. 1949. Mr Harrod's dynamics. *Economic Journal* 59, 68–85.

Rodbertus, K. 1898. *Overproduction and Crises*. London: Swan Sonnenschein.

Sismondi, J.C.L. 1815. *Political Economy*. New York: Kelley, 1966.

Sismondi, J.C.L. 1819. *Nouveaux principles d'économie politique*. Paris: Delaunay.

Spence, W. 1807. Britain independent of commerce. In *Tracts on Political Economy*, London: Longman, Hurst, Orme & Brown, 1822.

Sraffa, P., ed. 1952. *Works and Correspondence of David Ricardo*, vol. 6. Cambridge: Cambridge University Press.

Sweezy, P.M. 1942. *The Theory of Capitalist Development*. New York: Monthly Review Press.

How to cite this article

Schneider, Michael. "underconsumptionism." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000010>
doi:10.1057/9780230226203.1756

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

underemployment equilibria

Jean-Jacques P. Herings
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The standard model of general equilibrium is extended by allowing for expectations about supply opportunities by households and firms. In this framework there is typically a 1-dimensional continuum of underemployment equilibria that range from equilibria with arbitrarily pessimistic expectations to equilibria with rather optimistic expectations. An example illustrates the model and highlights some features of underemployment equilibria. The multiplicity of equilibria has a natural interpretation as being the result of coordination failures. The results in this framework are compared with those of the fixprice literature. Extensions to a monetary economy are discussed.

Keywords

agent optimization; animal spirits; Arrow–Debreu model of general equilibrium; Cobb–Douglas functions; competitive equilibrium; coordination failures; economics of general disequilibrium; excess capacity; excess demand; fixprice models; game theory; general equilibrium; general equilibrium models of coordination failures; incomplete markets; involuntary unemployment; Keynes, J.M.; Keynesianism; market clearing; neoclassical model; non-market clearing prices; Pareto optimality; path connectedness; positive externalities; price rigidities; rational expectations; rationing; seigniorage; self-justifying expectations; spillover effects; strategic complementarities; supply opportunities; temporary equilibrium; underemployment equilibria; wage rigidities; Walras's law

Article

Underemployment of resources refers to the situation where an increase in the resource utilization rate could lead to a Pareto improvement. Typical examples are involuntary unemployment and idle production capacities. There are two quite distinct views on the underemployment of resources. In the standard neoclassical world of the Arrow–Debreu model, underemployment of resources cannot occur. In the competitive equilibrium, involuntary unemployment does not exist, and production capacities are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
left idle when only such is Pareto optimal.

The Keynesian tradition, in contrast, builds on wage and price rigidities in its explanation of underemployment of resources. Indeed, Keynes's contribution has been reinterpreted by Clower (1965) and Barro and Grossman (1971) as the economics of general disequilibrium, in which price rigidities lead to quantity constraints for households and firms, which generally have spillover effects in other markets. This lead has been further developed in the fixprice literature, originating in the work of Bénassy (1975), Drèze (1975), and Younès (1975), and in general equilibrium theories on temporary equilibrium (see Grandmont, 1977, for a survey).

Although the fixprice literature stresses wage and price rigidities, Keynes himself postulates that it is possible to encounter self-justifying expectations, beliefs which are individually rational but which may lead to socially irrational outcomes (Keynes, 1936, ch. 12). We therefore would like to address the question whether underemployment of resources is possible when expectations are rational, agents optimize, and trade takes place at competitive prices. The underlying reason for underutilization of resources comes from coordination failures, self-justifying expectations which are individually rational but socially suboptimal.

In the literature, one may distinguish three classes of models with coordination failures. The first class consists of rather abstract game-theoretic models following the seminal work of Bryant (1983); see Cooper, 1999, for a state-of-the art account. An important message coming from this stream of the literature is that coordination failures may occur when there are strategic complementarities and positive externalities. However, these models typically lack a coordinating role for the price mechanism. Strategic models with a coordinating role for prices constitute the second class. Roberts (1987) presents a model that meets these criteria. It is a simple model of a closed economy that allows for coordination failures in a strategic setting. However, outcomes in the second class of models are often not robust to slightly different specifications of the model (Jones and Manuelli, 1992).

The third class of models consists of general equilibrium models of coordination failures. We refer to Citanna et al. (2001) for the most general presentation of these ideas. The third class leads to results that are robust and general. The methodological assumptions are shared with those of the neoclassical model: agent optimization, market clearing, and rational expectations. Underemployment of resources occurs as a consequence of self-confirming, pessimistic expectations about supply opportunities.

Competitive equilibria

Consider the classical general equilibrium model with *H* households, *F* firms and *L* commodities as described in Debreu (1959). A household *h* is characterized by its consumption set, for the sake of simplicity equal to $\mathbb{R}^L_+$, a utility function $u^h: \mathbb{R}^L_+ \to \mathbb{R}$, and initial endowments $e^h \in \mathbb{R}^L_+$. The feasible production plans of firm *f* are described by the production possibility set Y^f. If firm *f* chooses production plan y^f and the prices at which trade takes place are $p \in \mathbb{R}^L$, then the firm's profits equal $p \cdot y^f$. Household *h* receives a share θ^{fh} of the profits of firm *f*.

We assume both households and firms to be price takers. If trade takes place at prices *p*, then firm *f* faces the following profit maximization problem:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\max_{y^f \in Y^f} p \cdot y^f.$$

Under standard assumptions the firm's profit maximization problem is well-defined. For the remainder we assume the profit maximizing production bundle to be unique. This can also be shown to hold under standard assumptions, mainly requiring the strong assumption of decreasing returns to scale.
The utility maximization problem of household *h* reads as follows:

$$\max_{x^h \in \mathbb{R}^L_+} u^h(x^h) \text{ s.t. } p \cdot x^h \leq w^h,$$

where w^h equals the value of the household *h*'s initial endowments, $p \cdot e^h$, plus the household's share in the firms' profits, $\sum_f \theta^{fh} p \cdot y^{*f}$, with y^{*f} the profit maximizing production bundle chosen by firm *f*. Under standard assumptions, the maximization problem of the household has a unique solution x^{*h}.
Under the usual microeconomic methodological premises of agent optimization and market clearing, together with rational expectations, one defines a *competitive equilibrium* as a price system p^* and an allocation $(x^*, y^*) = (x^{*1}, \ldots, x^{*H}, y^{*1}, \ldots, y^{*F})$ such that at prices p^* households maximize utility by choosing the consumption bundle x^{*h} and firms maximize profits by choosing the production plan y^{*f}.

Rationing

The puzzle as to how competitive equilibria are achieved in real-world economies remains substantial. First, it is well-known that price adjustment processes need not converge to an equilibrium (Debreu, 1974; Saari and Simon, 1978; Saari, 1985). Blad (1978) stresses that convergence, even if it takes place, can take quite some time. Second, in many situations some agents, or coalitions of agents, set prices at levels not compatible with competitive equilibrium. Drèze (1989) models unions that set wages above competitive levels. Herings (1997) and Tuinstra (2000) show that political interference in the market mechanism can be rational from a partisan point of view and might be responsible for sustained deviations from prices that clear markets. Third, Drèze and Gollier (1993), Drèze (1997), and Herings and Polemarchakis (2005) argue that certain price rigidities are a welfare-improving response to market incompleteness. This argument is particularly valid for the two forms of underemployment most frequently encountered, namely, unemployed labour and excess capacities, two examples of commodities for which future markets are hardly developed.
For the moment we maintain the assumption that trade takes place at given prices *p*. Here, *p* may or may

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
not be competitive. We are not focusing on a specific theory of non-market clearing prices, but rather are interested in knowing how agents make decisions given that trade takes place at prices *p*. We deviate from the standard framework and assume that it is not common knowledge whether these prices are competitive or not. Even when all agents know whether prices are competitive or not, it is not necessarily the case that all agents know that all agents know whether prices are competitive or not, and it is even less likely that all agents know that all other agents know that all other agents know whether prices are competitive or not, and so on. Common knowledge of whether prices are competitive requires structural knowledge about the economy, very much at odds with the standard general equilibrium paradigm whereby in a decentralized economy agents only have to maximize utility given the prices that are quoted in the marketplace.

Our price system *p* may or may not be competitive. Since this fact is not common knowledge, it no longer makes sense for households and firms to express their unconstrained demands and supplies, and they should form expectations about supply and demand opportunities. One possible choice for these expectations is optimistic expectations: all households and firms do not expect to be constrained in either supply or demand. When prices are competitive, we are back in the situation of competitive equilibrium. The question is: are these the only possible expectations compatible with the microeconomic methodological premises of agent optimization and market clearing, together with rational expectations? Motivated by the empirical regularity that constraints on the supply side are more common than those on the demand side, as is suggested by unemployment in labour markets or unused capacities in production processes, we follow van der Laan (1980) and Kurz (1982) and restrict attention to constraints on the supply side. Moreover, for the sake of simplicity, we consider point expectations about supply opportunities.

If trade takes place at prices *p* and firm *f* expects supply opportunities of at most $\bar{y}^f \in \mathbb{R}^L_+$, then firm *f*•'s profit maximization problem becomes:

$$\max_{y^f \in Y^f} p \cdot y^f \text{ s.t. } y^f \leq \bar{y}^f.$$

At this point it should be noted that, if a firm *f* does not produce a particular commodity *l*, the value of $\bar{y}_l^f$ is entirely inconsequential. Again, under standard assumptions the firm's profit maximization problem is well-defined. In fact, the constraints related to the expected supply opportunities ensure that the firm's profits are bounded from above, a property that does not hold in general for the competitive model. At prices *p* and expected supply opportunities of $\bar{y}^f$, the supply of firm *f*, that is, the profit maximizing production plan of firm *f*, is denoted by $s^f(p, \bar{y}^f)$ and the firm's profits by $\pi^f(p, \bar{y}^f)$. The wealth of household *h* is then equal to $w^h = p \cdot e^h + \Sigma_f \theta^{fh} \pi^f(p, \bar{y}^f)$. The utility maximization problem of household *h* that trades at prices *p*, expects supply opportunities equal to $\bar{z}^h$, and has budget $w^{\bullet h}$ equals:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\max_{x^h \in \mathbb{R}^L_+} u^h(x^h) \text{ s.t. } p \cdot x^h \leq w^h \text{ and } \underline{z}^h \leq x^h - e^h.$$

Under standard assumptions, the maximization problem of the household has a unique solution $d^h(p, \underline{z}^h, w^h)$.
Since supply may not equal demand, one needs a rule to address discrepancies, called a rationing mechanism. Expected supply opportunities should be related to the rationing mechanism, which determines the allocation in case of excess supplies. For labour markets, one can think of a priority system that determines which worker is the first to become unemployed, who is next, and so on and so forth. Another rationing mechanism would share the burden of unemployment equally among workers, for instance by the imposition of an upper bound on the number of hours worked per week.
For notational convenience we assume the latter rationing mechanism in all markets, implying that in equilibrium rational agents face the same expected supply opportunities, $\bar{y}^1 = \cdots = \bar{y}^F = -\underline{z}^1 = \cdots = -\underline{z}^H$. We denote the commonly expected supply opportunities by $r \in \mathbb{R}^L_+$.
At expected supply opportunities *r*, every firm *f* faces the constraint $y^f \leq r$ and every household *h* optimizes utility subject to $-r \leq x^h - e^h$. All the results remain true with appropriate modifications for general rationing systems; see Herings (1996b) for a survey of rationing systems encountered in the literature.
A firm or household is said to be rationed in the market of commodity l' if the expected supply opportunities in this market are binding. More precisely, a firm *f* is *rationed* in the market of commodity l' at prices *p* and expected supply opportunities *r* if $\pi^f(p, \bar{r}) > \pi^f(p, r)$, where $\bar{r}_{l'} = +\infty$ and, for $l \neq l'$, $\bar{r}_l = r_l$. A household *h* is *rationed* in the market of commodity l' at prices *p* and expected supply opportunities *r* if $u^h(d^h(p, -r, w^h)) < u^h(d^h(p, -\bar{r}, w^h))$, where $\bar{r}$ is related to *r* as before. There is *rationing* on the market of commodity l' if at least one firm or at least one household is rationed on the market of commodity l'.

Definition: . An *underemployment equilibrium* of the economy $\mathcal{E} = ((u^h, e^h)_h, ((Y^f, (\theta^{fh})_h)_f)$ at prices *p* are expected supply opportunities r^* and an allocation (x^*, y^*) such that

1. (a) for every firm *f*, $y^{*f} = s^f(p, r^*)$,
2. (b) for every household *h*, $x^{*h} = d^h(p, -r^*, w^{*h})$, where $w^{*h} = p \cdot e^h + \sum_f \theta^{fh} \pi^f(p, r^*)$,
3. (c) $\sum_h x^{*h} = \sum_h e^h + \sum_f y^{*f}$.

An example

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As a simple example, let us consider an economy with one household, one firm, and two commodities. Let us interpret the commodities as leisure and an aggregate consumption good, and suppose that the household owns initially one unit of leisure and nothing of the consumption good, $e = (1, 0)$. The household's utility function is Cobb–Douglas, $u(x) = x_1, x_2$. The firm transforms the labour input into output by the production function $y_2 = \frac{2}{3}\sqrt{-3y_1}$, where $y_1 \leq 0$. When we normalize the price of output to be 1, turning the wage rate into the real wage rate, it is not hard to verify that the competitive equilibrium is given by $p^* = (1, 1)$, $x^* = (\frac{2}{3}, \frac{2}{3})$, and $y^* = (-\frac{1}{3}, \frac{2}{3})$.

Now suppose that it is possible to trade at the competitive equilibrium prices, so $p = (1, 1)$, but it is not common knowledge that these prices are competitive, and as a consequence firms and households form point expectations on supply opportunities $r = (r_1, r_2)$. We want to verify whether such expectations can be self-confirming. It is easily verified that for each $r_1^* \in [0, \frac{1}{3})$ there is an underemployment equilibrium with expected supply opportunities r^* given by $r_2^* = \frac{2}{3}\sqrt{3r_1^*}$, $x^* = (1 - r_1^*, r_2^*)$, and $y^* = (-r_1^*, r_2^*)$. The household expects a constraint on labor supply equal to $-r_1^*$ yielding labour income r_1^*. The firm expects a constraint on the supply of output equal to $r_2^* = \frac{2}{3}\sqrt{3r_1^*}$. It optimally demands an amount of labour equal to r_1^*, leading to profits $\frac{2}{3}\sqrt{3r_1^*} - r_1^*$. Notice that the optimal labour demand by the firm equals the constraint on labour supply anticipated by the household. The household's capital income is equal to $\frac{2}{3}\sqrt{3r_1^*} - r_1^*$, leading to total income of $\frac{2}{3}\sqrt{3r_1^*}$, to be spent on the aggregate consumption good. The optimal demand of the household for the aggregate consumption good equals the supply opportunities expected by the firm, thereby confirming those expectations. There is rationing in both markets. The household is rationed in the labour market and the firm in its output market.

Finally, every (r_1^*, r_2^*) with $r_1^* \geq \frac{1}{3}$ and $r_2^* \geq \frac{2}{3}$ sustains an underemployment equilibrium that coincides with a competitive equilibrium in terms of the allocation reached, $x^* = (\frac{2}{3}, \frac{2}{3})$, $y^* = (-\frac{1}{3}, \frac{2}{3})$. In this case, there is no market with rationing.

In the example, two extreme underemployment equilibria stand out. One is the underemployment equilibrium with completely pessimistic expectations about supply opportunities, $r^* = (0, 0)$, $x^* = (1, 0)$, $y^* = (0, 0)$. The other is the underemployment equilibrium with expectations about supply opportunities that are sufficiently optimistic to obtain the competitive allocation; the minimally optimistic expectations to achieve this are $r^* = (\frac{1}{3}, \frac{2}{3})$. These extreme underemployment equilibria are connected by a set of underemployment equilibria with more moderate expectations on supply opportunities.

In the example, trade was supposed to take place at competitive prices to highlight underemployment caused by mis-coordination of expectations and not by relative prices that are incompatible with competitive equilibrium. One may argue that it is a probability zero event that trade takes place at competitive prices. The crucial features of the example remain unchanged when trade takes place at non-competitive prices. Suppose that trade takes place at a real wage p_1 above the competitive wage rate of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1. It can be verified that there is still a no-trade equilibrium sustained by completely pessimistic expectations on expected supply opportunities. Although the competitive allocation is no longer feasible when the real wage is above the competitive level, it can be verified that there is also an underemployment equilibrium where the firm does not face rationing and the household observes rationing of its labour supply; the minimally optimistic expectations on supply opportunities that sustain this equilibrium are given by

$$r^* = \left(\frac{1}{3(p_1)^2}, \frac{2}{3p_1}\right)$$

leading to consumption and production

$$x^* = \left(1 - \frac{1}{3(p_1)^2}, \frac{2}{3p_1}\right), \quad y^* = \left(\frac{-1}{3(p_1)^2}, \frac{2}{3p_1}\right).$$

The same underemployment allocation is sustained by more optimistic expectations

$$r_1^* = \frac{1}{3(p_1)^2} \text{ and } r_2^* \geq \frac{2}{3p_1}.$$

Again, the two extreme equilibria are connected by a continuum of underemployment equilibria, with expectations ranging from completely pessimistic to expectations that sustain an underemployment equilibrium without rationing of the firm and with rationing of the labour supply of the household. When the real wage p_1 is below the competitive level, there is still an underemployment equilibrium with completely pessimistic expectations about supply opportunities and no trade. There is also an underemployment equilibrium without rationing of the supply of the household but with rationing of the firm's supply of output. Let $\bar{r}_2$ be equal to $4(\sqrt{1+3(p_1)^2})/6p_1$. Notice that $\bar{r}_2$ is below 2/3 when the real wage p_1 is below 1. The minimally optimistic expectations that sustain an equilibrium without rationing of the household are given by $r^* = (1 - \bar{r}_2/p_1, \bar{r}_2)$ leading to an allocation $x^* = (\bar{r}_2/p_1, \bar{r}_2)$, $y^* = (\bar{r}_2/p_1 - 1, \bar{r}_2)$. The same underemployment equilibrium allocation is sustained by the more optimistic expectations $r_1^* \geq 1 - \bar{r}_2/p_1$ and $r_2^* = \bar{r}_2$. The two extreme

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
underemployment equilibria are connected by a continuum of underemployment equilibria with more moderate expectations on supply opportunities.

Animal spirits

The example suggests that in general there is a 1-dimensional continuum of equilibria, ranging from a no-trade equilibrium with completely pessimistic expectations to an equilibrium with rather optimistic expectations and without rationing in at least one market. This result is almost true, except that the case with completely pessimistic supply expectations leads to zero income for the households, a case that is well-known to be problematic for equilibrium existence. Inspired by preliminary results in van der Laan (1982), Herings (1996a; 1998) and Citanna et al. (2001) provide conditions under which the following result holds.

Theorem: . Under standard assumptions, the economy $\mathcal{E} = ((u^h, e^h)_h, (Y^f, (\theta^{fh})_h)_f)$ *where trade takes place at prices p possesses a connected set of underemployment equilibria E such that for every* $\rho \in (0, \infty)$, *there is an equilibrium* (r^*, x^*, y^*) *in E with* $\max_l r_l^* = \rho$. (A set is path-connected if any two points in the set can be connected by a path that does not leave the set. Path-connectedness implies connectedness, a slightly weaker topological property of sets, which loosely speaking means that the set consists of one piece.)

The theorem gives general equilibrium underpinnings to the Keynesian ideas that changes in expectations, also referred to as animal spirits, can affect equilibrium economic activity, in particular the level of output and employment. The theorem rules out the case with completely pessimistic expectations, corresponding to $\max_l r_l^* = 0$. It shows that the set *E* links equilibria with arbitrarily pessimistic expectations ($\max_l r_l^*$ arbitrarily small, but positive) to equilibria with rather optimistic expectations ($\max_l r_l^*$ arbitrarily large). Notice that the condition $\max_l r_l^*$ arbitrarily large only implies that for one market expectations are sufficiently optimistic to rule out rationing. The expectations on supply opportunities on other markets might still be completely pessimistic.

In the absence of production, the statement of the theorem can be simplified somewhat. It is shown in Herings (1998), under standard assumptions, that the economy $\mathcal{E} = ((u^h, e^h)_h)$ where trade takes place at prices *p* possesses a connected set of underemployment equilibria *E* such that for every $\rho \in [0, \infty)$ there is an equilibrium (r^*, x^*, y^*) in *E* with $\max_l r_l^* = \rho$. In exchange economies, the connected set includes an underemployment equilibrium with completely pessimistic expectations.

The theorem demonstrates that the set of equilibria is at least 1-dimensional. In general, one should expect the dimension of the set of equilibria to be exactly equal to 1. The reason is that the model postulates *L* free variables corresponding to the expected supply opportunities *r* and *L* market clearing conditions. Let *E* be an economy where trade takes place at prices *p* and let $z: R_+^L \to R^L$ denote the excess demand function of the economy, a function of expected supply opportunities *r*. Because of Walras's law, it holds that for every $r \in R_+^L$, $p \cdot z(r) = 0$. This implies that there are only $L - 1$ independent market clearing conditions. Since there are *L* free variables, this leaves a 1-dimensional

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
solution set.
At this point it should be observed that the same reasoning also applies to the standard competitive model. And indeed, in general the set of competitive equilibria is 1-dimensional too. Whenever (p^*, x^*, y^*) constitutes a competitive equilibrium, so does $(\lambda p^*, x^*, y^*)$ for λ positive. However, in the standard competitive model, the excess demand function is homogeneous of degree 1, implying that the same allocation is sustained by p^* and λp^*. The homogeneity property holds for prices but not for expectations. In general, the excess demand function z is not homogeneous of any degree, and it is not the case that λr^* is part of an underemployment equilibrium when r^* is. Generically, the set E of the theorem contains a 1-dimensional set of distinct equilibrium allocations.

Coordination failures

The theorem makes clear that a multiplicity of equilibria results even when prices are competitive. The interpretation of the multiplicity of equilibria as coordination failures and the link to the macroeconomic literature on coordination failures were made in Drèze (1997). It would be tempting to conclude that, when trade takes place at competitive prices, then the connected set of underemployment equilibria contains the competitive equilibrium allocation. Although it is true that the competitive equilibrium allocation is an underemployment equilibrium allocation sustained by trade at competitive prices coupled with sufficiently optimistic expectations, it is possible to produce examples where it is outside the connected set of the theorem (Citanna et al., 2001).
Under an additional assumption akin to gross substitutability, the following result holds. If trade takes place at competitive equilibrium prices p, then the economy $\mathcal{E} = ((u^h, e^h)_h, (Y^f, (\theta^{fh})_h)_f)$ possesses a connected set of underemployment equilibria E such that for every $\rho \in (0, \rightarrow)$ there is an equilibrium (r^*, x^*, y^*) in E with $\max_l r_l^* = \rho$, and for every $\rho \in (0, \rightarrow)$ there is an equilibrium (r^*, x^*, y^*) in E with $\min_l r_l^* = \rho$. More precisely, the following assumption on the aggregate excess demand function suffices. If $r, \bar{r} \in \mathbb{R}^L_+$ with $r \leq \bar{r}$ and $r_{l'} = \bar{r}_{l'}$, then $z_{l'}(r) \leq z_{l'}(\bar{r})$. The interpretation of the assumption is the following. A weak increase in expected supply opportunities in markets different from l' should lead to a weak increase in the excess demand for commodity l'.
This assumption, though strong, is not unreasonable. On the household side, a household may lower its supply of commodity l' in exchange for more supply of another commodity, for instance if the household switches to a more attractive job. A household may also increase its demand for commodity l' as a consequence of higher income. Indeed, more expected supply opportunities of commodities different from l' weakly increase the household's income, which will lead to more demand of commodity l' if it is a normal good. On the producer side, if l' is an output for some firm, then increased supply opportunities of other goods, lead to a weakly lower supply of commodity l', when the firm directs more inputs to the production of the other goods. If l' is an input for a firm, then increased supply opportunities of other goods naturally lead to more production, and thereby an increased input demand, in particular for commodity l'. Notice that the assumption needs to hold only in the aggregate.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Efficiency

It is not hard to argue that underemployment equilibria are not Pareto optimal in general. As soon as there are two commodities, l and l', and two households h and h', such that household h' is rationed in the market for commodity l' but not in the market for commodity l, whereas household h is not rationed in the market for commodity l', then it follows almost immediately that the marginal rate of substitution of commodity l for l' is not the same for households h and h'. This contradicts Pareto optimality.

It has been argued before that price rigidities may emerge for efficiency reasons. The argument of the previous paragraph makes clear that such is not the case in a complete markets setting when coordination failures are absent. In an incomplete market setting, however, Drèze and Gollier (1993) and Herings and Polemarchakis (2005) show that price rigidities may lead to equilibria that are Pareto superior to competitive equilibria. In general, it will depend on the magnitude of the coordination failures, whether or not welfare improvements result.

Extensions

For illustrative purposes, we have so far considered the simplest case where trade takes place at predetermined prices for all commodities. It is not hard to generalize this set-up substantially, and allow for general lower bounds $\underline{p}$ and upper bounds $\overline{p}$ that define the set of admissible prices at which trade may take place. The notion of underemployment equilibrium should then be extended by the requirement that only when the price of a commodity l equals its lower bound is rationing of the supply of commodity l allowed for. In such a more general setting, it may be interesting to also allow for demand rationing when the price of a commodity equals its upper bound. Allowing for demand rationing will enlarge the set of equilibria. By taking all $\underline{p}_l$ equal to $-\infty$ and all $\overline{p}_l$ to ∞, we obtain the notion of competitive equilibrium as a special case. The results we mentioned before remain true in this more general set-up.

The existence of a continuum of underemployment equilibria is a robust phenomenon. This seems to be in striking contrast with the conclusion of the fixprice literature, where equilibria are typically locally unique. The reason for this apparent disparity is that the fixprice literature puts one additional constraint on the equilibrium set. It is assumed that there is no rationing in the market of an a priori determined numeraire commodity, called money. Not only is the interpretation of one of the commodities as money controversial, it is also misleading as far as the indeterminacy of equilibrium is concerned.

Suppose we follow Drèze and Polemarchakis (2001) and extend the model by a model of a central bank, which produces money at no cost. Households and firms need money to make transactions, a particular example being a cash-in-advance transactions technology. The central bank sets the nominal interest rate at which it accommodates all money demand. The central bank redistributes the revenues from seigniorage to the households in the form of dividends. Money does not enter into the utility function of the households.

This model fits exactly into the framework discussed so far. Without loss of generality, commodity L may serve as money. The central bank can be modelled as firm F, which can produce the output money in arbitrary amounts without using inputs. The profits of firm F are equal to the seigniorage of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
central bank. The price of money is equal to the interest rate. Since the central bank accommodates money demand, the money supply of the central bank is equal to that of a profit-maximizing firm *F* that expects a constraint on the supply of commodity *L* equal to the aggregate demand for commodity *L* (for strictly positive interest rates). Our theorem on the existence of a connected set of underemployment equilibria therefore applies to the case where money is explicitly introduced, thereby contradicting the determinacy results obtained in the fixprice literature.

We have shown how underemployment equilibria result in a general equilibrium model where agents are allowed to form expectations on expected supply opportunities. We analyse whether such expectations can be self-confirming and argue that, even at competitive prices, a continuum of equilibria results, including an equilibrium with approximately no trade and a competitive equilibrium. Such equilibria also arise at other price systems, but are then a consequence of both self-confirming pessimistic expectations and of prices incompatible with competitive equilibrium. Expected supply opportunities in underemployment equilibria bear a close resemblance to the self-justifying expectations of Keynes (1936), beliefs which are individually rational but socially suboptimal. The further study of underemployment equilibria in models with time and uncertainty, incomplete markets, price-setting agents, and a monetary authority features prominently on the research agenda, as it would allow for explicit links with the modern macroeconomic literature on inflation, output and unemployment.

See Also

- determinacy and indeterminacy of equilibria
- existence of general equilibrium
- fixprice models
- general equilibrium (new developments)
- involuntary unemployment
- money and general equilibrium
- rationing

Bibliography

Barro, R. and Grossman, H. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

Bénassy, J.-P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42, 503–23.

Blad, M. 1978. On the speed of adjustment in the classical tatonnement process: a limit result. *Journal of Economic Theory* 19, 186–91.

Bryant, J. 1983. A simple rational expectations Keynes-type model. *Quarterly Journal of Economics* 98, 525–8.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Citanna, A., Crès, H., Drèze, J., Herings, P. and Villanacci, A. 2001. Continua of underemployment equilibria reflecting coordination failures, also at Walrasian prices. *Journal of Mathematical Economics* 36, 169–200.

Clower, R. 1965. The Keynesian counter-revolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cooper, R. 1999. *Coordination Games – Complementarities and Macroeconomics*. Cambridge: Cambridge University Press.

Debreu, G. 1959. *Theory of Value*. New Haven, CT: Yale University Press.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–21.

Drèze, J. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16, 301–320.

Drèze, J. 1989. *Labour Management, Contracts and Capital Markets: A General Equilibrium Approach*. Oxford: Basil Blackwell.

Drèze, J. 1997. Walras–Keynes equilibria-coordination and macroeconomics. *European Economic Review* 41, 1735–62.

Drèze, J. and Gollier, C. 1993. Risk sharing on the labour market. *European Economic Review* 37, 1457–82.

Drèze, J. and Polemarchakis, H. 2001. Monetary equilibria. In *Economics Essays: A Festschrift for Werner Hildenbrand*, ed. G. Debreu, W. Neuefeind and W. Trockel. Heidelberg: Springer-Verlag.

Grandmont, J.-M. 1977. Temporary general equilibrium theory. *Econometrica* 45, 535–72.

Herings, P. 1996a. Equilibrium existence results for economies with price rigidities. *Economic Theory* 7, 63–80.

Herings, P. 1996b. *Static and Dynamic Aspects of General Disequilibrium Theory*. Theory and Decision Library Series C: Game Theory, Mathematical Programming and Operations Research. Dordrecht: Kluwer Academic Publishers.

Herings, P. 1997. Endogenously determined price rigidities. *Economic Theory* 9, 471–98.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Herings, P. 1998. On the existence of a continuum of constrained equilibria. *Journal of Mathematical Economics* 30, 257–73.

Herings, P. and Polemarchakis, H. 2005. Pareto improving price regulation when the asset market is incomplete. *Economic Theory* 25, 135–54.

Jones, L. and Manuelli, R. 1992. The coordination problem and equilibrium theories of recessions. *American Economic Review* 82, 451–71.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kurz, M. 1982. Unemployment equilibrium in an economy with linked prices. *Journal of Economic Theory* 26, 100–23.

Roberts, J. 1987. An equilibrium model with involuntary unemployment at flexible, competitive prices and wages. *American Economic Review* 77, 856–74.

Saari, D. 1985. Iterative price mechanisms. *Econometrica* 53, 1117–31.

Saari, D. and Simon, C. 1978. Effective price mechanisms. *Econometrica* 46, 1097–125.

Tuinstra, J. 2000. The emergence of political business cycles in a two-sector general equilibrium model. *European Journal of Political Economy* 16, 509–34.

van der Laan, G. 1980. Equilibrium under rigid prices with compensation for the consumers. *International Economic Review* 21, 63–73.

van der Laan, G. 1982. Simplicial approximation of unemployment equilibria. *Journal of Mathematical Economics* 9, 83–97.

Younès, Y. 1975. On the role of money in the process of exchange and the existence of a non-Walrasian equilibrium. *Review of Economic Studies* 42, 489–501.

How to cite this article

Herings, Jean-Jacques P. "underemployment equilibria." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000072>
doi:10.1057/9780230226203.1757

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

unemployment and hours of work, cross country differences

Richard Rogerson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since the 1960s labour market outcomes among the world's richest economies have changed dramatically, especially in terms of unemployment rates and time devoted to market work. This article summarizes the evidence regarding these changes and discusses some of the explanations that have been proposed for why these labour market outcomes have evolved so differently across economies.

Keywords

barriers to entry; common shock; cross-country differences in unemployment and hours worked; employment protection; home production; hours worked; income support; labour market regulation; layoffs; leisure; neoclassical growth theory; product market regulation; productivity growth; skill-biased technical change; technical change; time use; unemployment; unemployment insurance; wage dispersion; wage rigidity; wage setting institutions

Article

Understanding the forces that determine resource allocations in decentralized economies is one of the fundamental objectives of economics. Because countries often differ greatly in terms of economic policies and institutions, examining how resource allocations differ across countries provides a promising opportunity to learn about these forces. Conversely, when we see large differences in allocations across countries, we are presented with an important opportunity to learn about what factors can generate these large differences. One prominent case in point is the large differences in labour market outcomes – specifically in terms of unemployment rates and hours of work – that exist across rich industrialized economies. A large literature has emerged that documents the nature of these differences and seeks to determine which factors can account for them. This article provides a brief overview of this literature.

Cross-country differences in unemployment

The literature on cross-country differences in unemployment was motivated by a simple real world development: a large and persistent rise in unemployment in several continental European countries relative to the United States, starting in the mid-1970s. Figure 1 displays this fact by showing the evolution of unemployment rates since 1956 for the United States and the average of four economies from Continental Europe: Belgium, France, Germany and Italy.

Figure 1
Unemployment in the United States and Continental Europe. *Source*: OECD Labor Market Database.

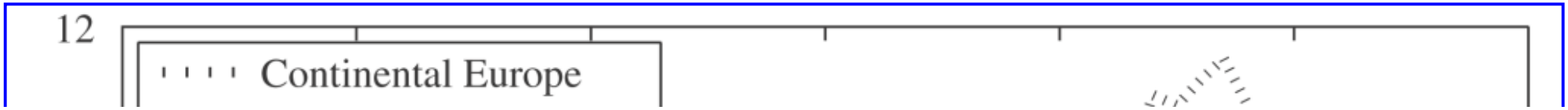

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

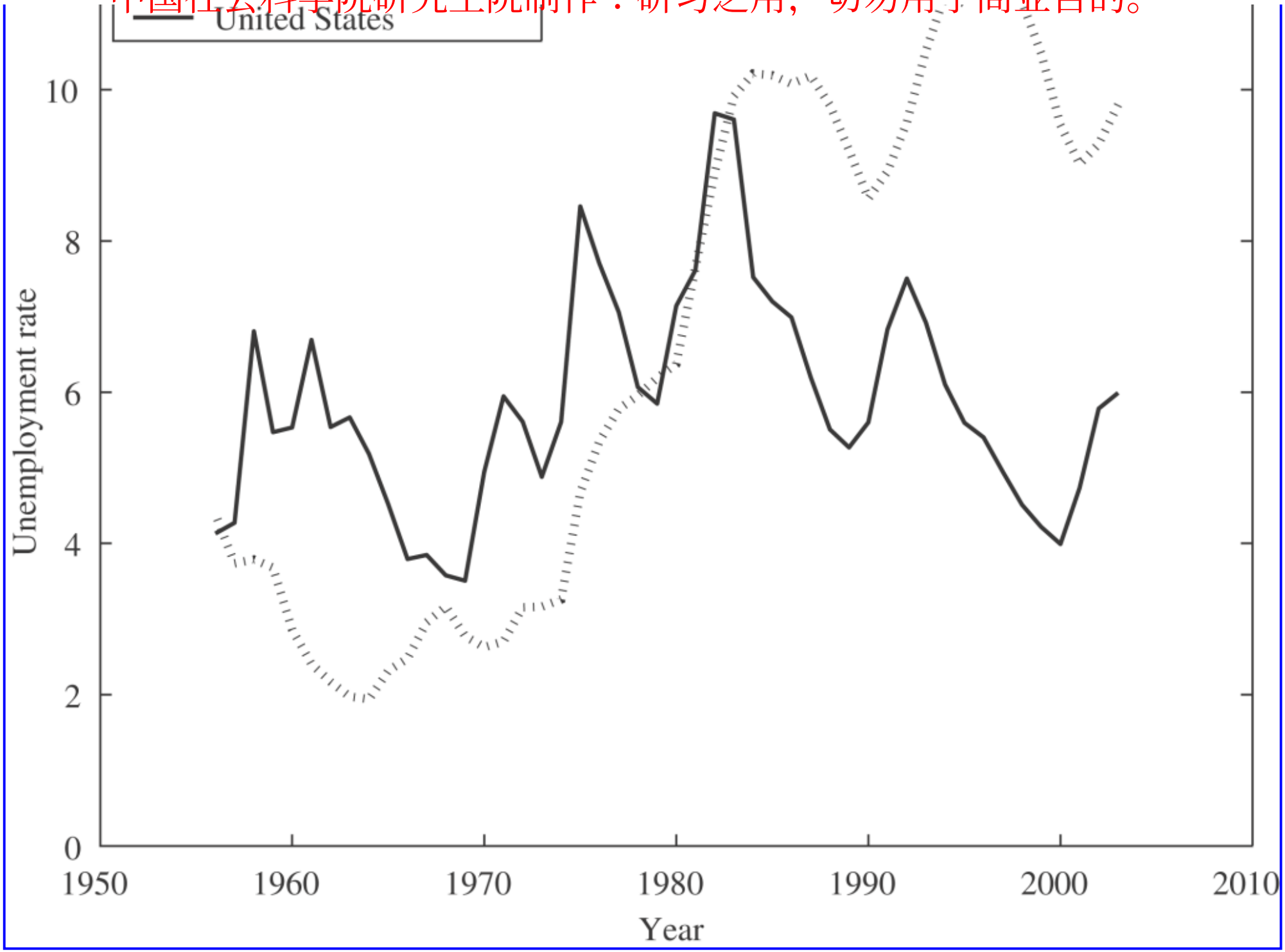

As the figure reveals, prior to 1970 the unemployment rate in these European countries averaged around three per cent, while since 1990 it has averaged more than ten per cent. This dramatic increase is concentrated in the period 1975–85. In contrast, while the United States also experienced an increase in average unemployment during the 1975–85 period, this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

increase is only transitory; the US unemployment rate has averaged close to five per cent in both the early and later time periods. Faced with this dramatic change in relative unemployment rates across countries, economists were naturally led to ask why. Based on the time series plots in Figure 1, an answer to this question had to address two important issues:

1. 1. Why did the increase occur in the 1975–85 period?
2. 2. Why did it occur in some countries and not in others?

At a general level, one can imagine two classes of explanations. One class postulates that something changed in one set of economies during the period 1975–85 but not in the others. A second class postulates that something changed in all economies during the 1975–85 period, but that economies responded in different ways to this change due to differences in their institutions or policies.

Krugman (1994) was the first to suggest that an explanation of the second type seemed promising. His story emphasized that the process of skill-biased technological change became more prominent beginning in the 1970s, thereby creating an economic force tending to increase the dispersion in individual wages. Although all economies were subjected to this underlying change in technological progress, this change was propagated differently across economies. In the United States, wage setting institutions were ‘flexible’ and the result was increased wage dispersion and little change in unemployment. In the economies of Continental Europe, wage setting institutions were ‘rigid’ and did not allow wages to become more dispersed, so the result was instead an increase in unemployment. While subsequent work (see for example, Card, Kramarz and Lemieux, 1999) did not provide support for this story, at least in its simplest form, the contribution was important because it suggested an important class of explanations.

The issue of rigorously distinguishing between the two classes of explanations was subsequently taken up in a paper by Blanchard and Wolfers (2000). Based on statistical analysis, these authors argued that the ‘common shock’ story was most promising. The force of this conclusion is tempered somewhat by two features of the analysis. First, the result that the ‘different shocks’ explanation does not provide a good account of the data is very much dependent on what shocks are explicitly incorporated in the analysis. In particular, Blanchard and Wolfers did not incorporate the fact that taxes changed considerably over this time period, a point we will return to below. Second, their analysis did not attempt to identify what the important common shock(s) were, and did not construct an explicit model to quantify how various institutions affected the propagation of these shocks. However, making use of advances in general equilibrium modelling of unemployment (such as the Diamond–Mortensen–Pissarides matching model or the Lucas–Prescott island model), subsequent work has sought to remedy this limitation by quantitatively evaluating particular candidates from the ‘common shock’ class of explanations in the context of fully specified models.

An early example in this literature was Bertola and Ichino (1995). They argued that the common shock was a permanent increase in the transient nature of production opportunities. The key differences in economic institutions were wage setting institutions that compressed wages and employment protection policies that made layoffs prohibitively costly. Several alternative analyses have since followed. Ljungqvist and Sargent (1998) argue that the common shock was a permanent increase in the amount of ‘turbulence’ for workers, and that the key institutional difference is generosity of income support for displaced workers. Mortensen and Pissarides (1999) and Marimon and Zilibotti (1999) both construct models in which the common shock was skill-biased technological change. While Mortensen and Pissarides stress differences in unemployment insurance (UI) benefits and employment protection as the key institutional differences, Marimon and Zilibotti simply stress differences in UI benefits. Closely related, Hornstein, Krusell and Violante (2007) argue that the common shock is an increase in the rate of capital embodied technological change and that the key institutional differences are taxes, income support programmes, and employment protection.

While explanations of the ‘common shock’ variety have become popular in the literature, some researchers have argued against them. Using purely statistical methods, Nickell, Nunziata and Ochel (2006) challenge the Blanchard and Wolfers finding. In terms of model based analyses, Daveri and Tabellini (2000) argue that differences in the changes in tax rates between Continental Europe and the United States were central to understanding the different evolutions in unemployment rates. They also argue that the impact of higher taxes is very much influenced by wage setting institutions, thereby explaining why some other European countries that also experienced large increases in tax rates did not experience sharp increases in unemployment. This last point – that the effects of individual policies and institutions are not additive – has recently been emphasized by both Blanchard and Giavazzi (2002) and Pries and Rogerson (2005). Another model-based analysis is contained in Pissarides (2007). He argues that a significant part of the relative increase in European unemployment is associated with the slowing of productivity growth that came as European productivity converged to US levels.

The literature has made important headway in evaluating specific combinations of driving forces and propagation mechanisms, but there is still much more work to be done. First, much of the work to date has contrasted the behaviour of the United States with an average European country. Given the substantial heterogeneity within Europe, in terms of both policies and outcomes, it is desirable to push these exercises to consider outcomes on a country-by-country basis. Second, as noted above, the literature has focused almost exclusively on accounting for the rise in unemployment in a handful of European economies since 1970. There are also many other interesting episodes in the data. For example, Spain, the United Kingdom and the Netherlands all experienced dramatic increases in unemployment similar to those documented earlier, but each of these countries has subsequently experienced a sharp decrease in unemployment. Understanding the sources of these dynamics should also prove to be very valuable.

While the above discussion has focused on the efforts to understand the sharply different evolutions of unemployment across a small set of countries since the 1970s, the broader

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

research issue is to understand the quantitative consequences of various policy and institutional features on aggregate unemployment. Alvarez and Veracierto (1999) is one example of work that fits with this more general objective. As motivation for this general question one need only look at the distribution of unemployment rates across countries at any point in time. For example, Table 1 shows the distribution of unemployment rates in 2005.

Unemployment rates (2005)

$u<4.5$	$4.5<u<6$	$6<u<8$	$u>8$
NZ (3.7)	Norway (4.6)	Canada (6.8)	Belgium (8.1)
Ireland (4.3)	UK (4.6)	Portugal (7.6)	Finland (8.4)
Japan (4.4)	Denmark (4.8)	Italy (7.7)	Spain (9.2)
Switzerland (4.5)	Australia (5.1)	Sweden (7.7)	France (9.8)
	US (5.1)		Germany (11.2)
	Netherlands (5.2)		
	Austria (5.2)		

Source: OECD Labor Market Database.

As the reader can see, the dispersion of unemployment rates across countries is large. Understanding the forces that shape this distribution of outcomes remains an open and challenging research issue.

Other measures of labour market outcomes

If one thinks more carefully about characterizing resource allocations across countries, and specifically as this pertains to the labour market, it becomes clear that differences in unemployment rates may not be the best summary measure of differences in labour market allocations. The benchmark conceptual framework for modern thinking about aggregate resource allocation is the one-sector neoclassical growth model. This model stresses two margins that economists believe to be of first-order importance in thinking about aggregate allocations: the fraction of available time that is devoted to market work, and the fraction of output that is invested rather than consumed. Viewed from this perspective, the unemployment rate is the natural summary statistic to focus on only if it is a good measure of time devoted to market work. Historically, the framework of traditional Keynesian models assumed that this was indeed the case: the simple versions of these models assumed that labour supply is represented as some constant volume of available hours that was unaffected by any aspects of the economic environment. The only reason that observed hours of work would differ from this given value was unemployment. In such a conceptual framework, unemployment and total hours of work provide exactly the same information. But is this conceptual framework adequate to examine labour allocations in modern industrialized economies? Developments such as the rise of female labour force participation, the trend towards early retirement, the changing workweek, and the expansion of part-time work suggest that to view labour supply as a fixed volume of work determined only by the size of the population is to neglect some important economic forces.

To pursue this issue further it is instructive to take a closer look at the data. We look at data for 2005, but the basic messages of the analysis are not affected by the choice of year. Table 2 reports hours of work across countries, organizing the countries into four groups based on their hours worked. For each country, aggregate hours of work are computed as the product of two series from the OECD Labor Market Database: total civilian employment and annual hours of work per person in employment. It is important to note that the measure of annual hours of work per person in employment in this data-set attempts to take into account not only the length of the standard workweek but also the number of statutory holidays, sick days and vacation days. To compare aggregate hours of work across countries one has to make some normalization based on population. One could imagine different normalizations, such as the entire population, the adult population (those aged 15 and over), or the working-age population (those aged 15–64). The resulting patterns are not much affected by this choice, and the numbers reported in Table 2 are based on dividing total hours by the size of the working age population.

Annual hours worked per person aged 15–64 (2005)

$h<1{,}000$	$1{,}000<h<1{,}150$	$1{,}150<h<1{,}300$	$h>1{,}300$
Belgium (941)	Norway (1,044)	Finland (1,167)	Australia (1,323)
Germany (954)	Italy (1,046)	Denmark (1,191)	Japan (1,333)
France (961)	Ireland (1,122)	Sweden (1,193)	US (1,339)
Netherlands (979)	Austria (1,134)	Portugal (1,213)	NZ (1,386)
	Spain (1,145)	UK (1,240)	

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Canada (1,284)

Switzerland (1,286)

Source: OECD Labor Market Database.

The correlation between unemployment rates and hours of work in the 2005 cross section is –0.58, indicating a fairly sizable negative correlation. This negative correlation is reflected in the fact that Germany and France are both among the highest unemployment countries as well as the lowest hours-worked countries, while New Zealand is the lowest unemployment country and the highest hours-worked country. However, there are also several counter-examples to this pattern. The Netherlands and Norway, for example, have hours worked and unemployment rates that are both substantially below the average, while Canada has unemployment and hours of work that are both substantially above average. We conclude from this that even at a qualitative level differences in hours of work and differences in unemployment are sometimes quite distinct.

But more importantly, even when differences in hours of work and differences in unemployment describe similar situations qualitatively, the quantitative differences are dramatically different. For example, consider the following question. If the unemployment rate in a country such as France were reduced to the same level as the United States by placing some currently unemployed French people into employment, and having them work the same amount as those French people who are currently working, by how much would the gap in hours worked between France and the United States drop? The answer is that the gap would drop from its current value of 378 to 343, a decrease of less than ten per cent. From a pure accounting perspective, differences in unemployment rates are an almost insignificant component of differences in aggregate hours of work. Put somewhat differently, even if we completely understood the factors that give rise to observed differences in unemployment rates, we would understand practically none of the differences in hours of work.

Cross-country differences in hours of work

The previous section suggests that, if one examines cross-country labour market outcomes from the perspective of differences in resource allocations across economies, it will be useful to look at differences in hours of work rather than unemployment rates. An interesting starting point is to look at the evolution of hours worked for the two sets of economies depicted earlier in Figure 1. Figure 2 presents this information, where the reader is reminded that in this figure Continental Europe reflects the simple average for the economies of Belgium, France, Germany and Italy.

Figure 2

Hours of work in the United States and Continental Europe. *Sources*: OECD Labor Market Database; GGDC Database.

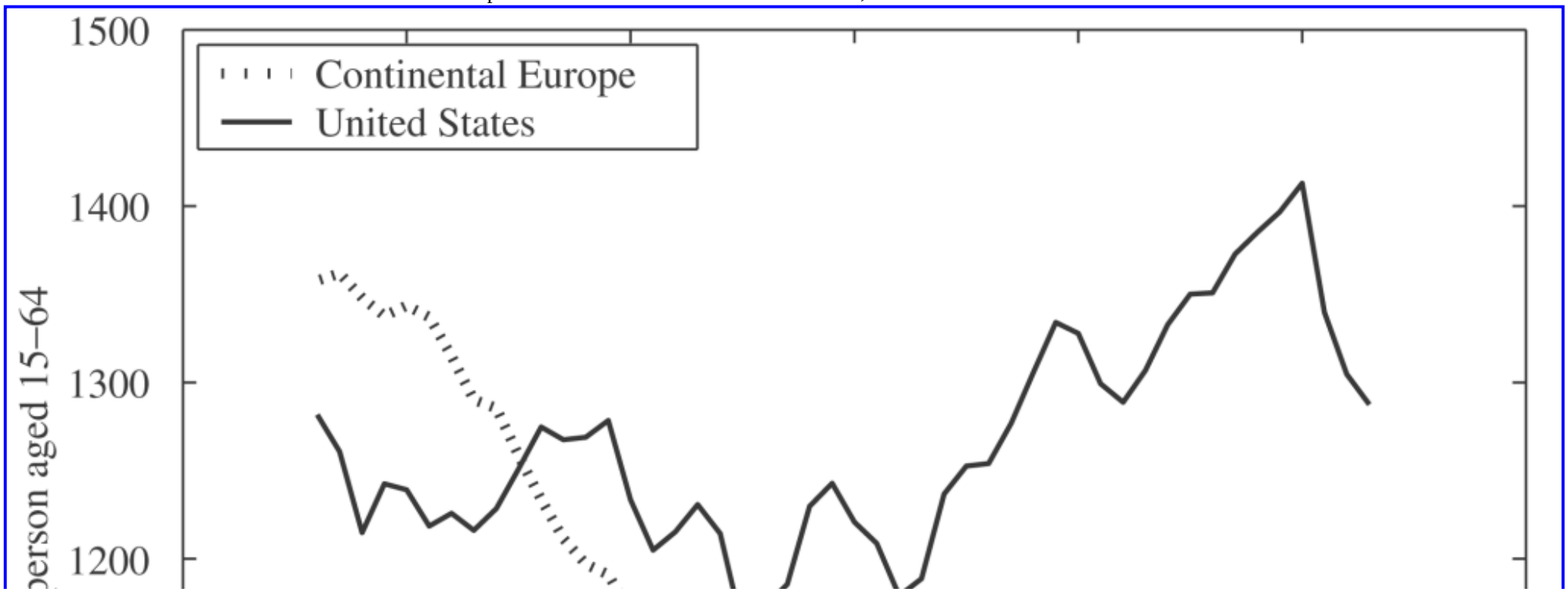

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

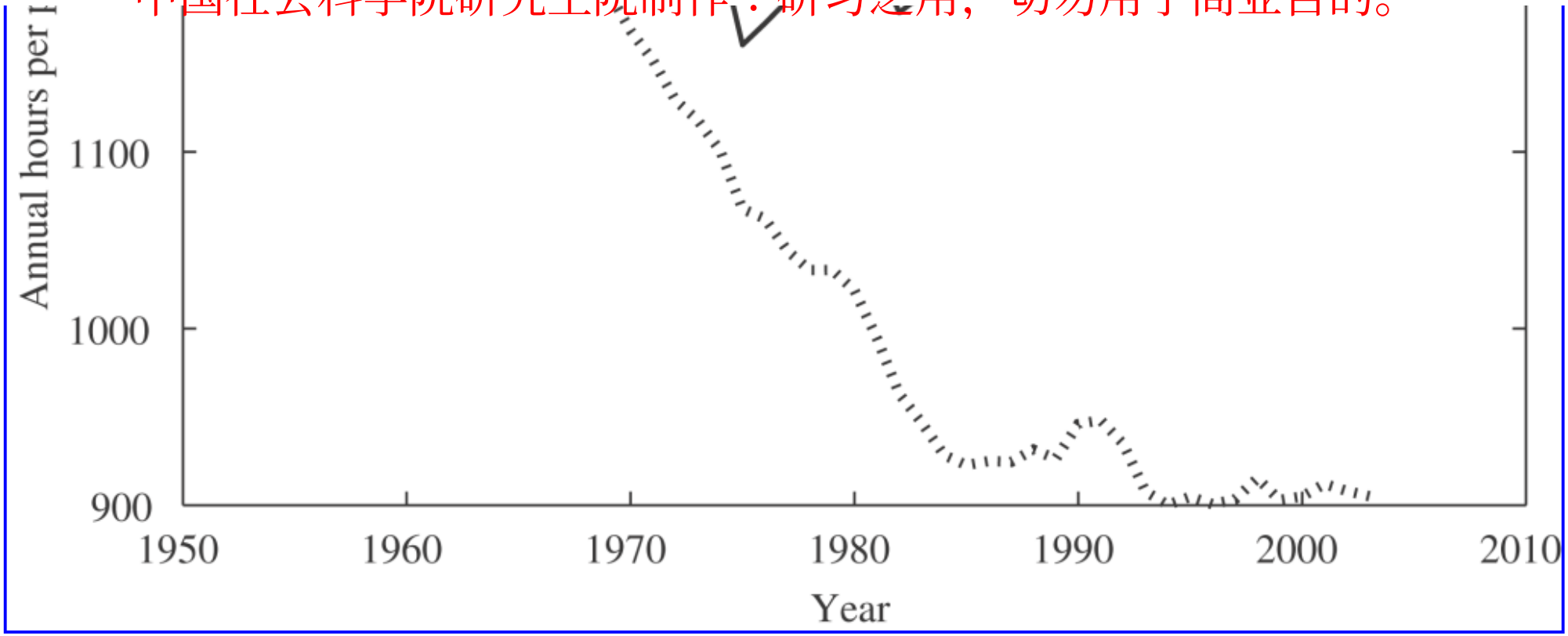

Several features are worth noting. Hours of work in Continental Europe decrease at a fairly constant rate from 1956 through to the mid-1980s, at which point they flatten out. The magnitude of this decrease in hours worked in Europe is enormous – a drop of over 35 per cent. In contrast, hours worked in the United States are virtually the same in 2005 as they were in 1956, though there are some low-frequency movements in the series during this time period. In contrast with the unemployment rate evolutions, it is of particular interest to note that there is nothing distinctive about the period 1975–85 from the perspective of the decline of hours in Europe.

Given the dramatic change in relative hours worked across these two sets of economies, it is not surprising that a literature has emerged that seeks to understand this change. Here too, one can imagine two different classes of explanations, one based on changes in some feature of the economic environment in Continental Europe relative to the United States, and the other based on a common change in the economic environments that has been propagated differently in the two sets of economies. There are two key differences relative to the literature on unemployment rates: timing and magnitude. Because we see changes beginning in the mid-1950s, and continuing at a fairly constant rate through to the mid-1980s, we presumably need to identify changes that exhibit this general time series pattern. The second difference is that we are talking about changes that are roughly an order of magnitude larger in terms of implications for hours of work.

Interestingly, whereas the unemployment literature has for the most part pursued the ‘common shocks’ explanation, the hours of work literature has instead focused on differences in policy changes across countries. This view has been heavily influenced by the contribution of Prescott (2004). He argues that, once cross-country differences in taxes are incorporated into the standard growth model, hours of work in the United States and several European economies in both the early 1970s and early 1990s are very close to those predicted by the model that assumes no other differences between the different economies. This general finding is further supported by Davis and Henrekson (2005), Ohanian, Raffo and Rogerson (2006), and Rogerson (2007).

One argument against the tax explanation, as noted by Alesina, Glaeser and Sacerdote (2005), is that it requires an aggregate labour supply elasticity that is larger than the individual labour supply elasticities typically found in estimation exercises using micro data. Recent work suggests that this critique is misplaced. Chang and Kim (2006) and Rogerson and Wallenius (2007) both argue that, when non-convexities are relevant for individual level labour supply decisions, the tight connection between micro and macro elasticities is broken. In particular, both papers argue that reasonable calibrations imply that small micro elasticities are consistent with large macro elasticities. Additional discussion can be found in Prescott (2006) and Rogerson (2006).

While an explanation based on taxes has been the one most developed to date, there are competing explanations. In work that is closely related but distinct from the Prescott analysis,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ljungqvist and Sargent (2007) argue that it is the nature of benefit programmes in Europe rather than the high tax rates per se that are responsible for the large differences in employment rates between the United States and several European economies. Interestingly, Ljunqvist and Sargent (2007) argue that large aggregate elasticities are inconsistent with the observed cross-country differences once one properly models benefit programmes. On a very different note, Blanchard (2005) has argued that the dominant factor is differences in preferences across countries. In particular, he suggests that the income effect on leisure is larger in European countries than in the United States, and as European productivity has increased to near US levels since the 1960s, Europeans have responded with a larger increase in leisure than that which occurred earlier in the United States. An interesting connection between this explanation and the one based on tax rates is that it is the income effect that is central to generating the tax effects in the analysis of Prescott.

While the discussion has thus far focused on understanding the reasons for the very different evolutions of hours worked between the United States and Continental Europe, the data in Table 2 reveal a host of other differences that are also of interest. Rogerson (2006) describes some additional patterns of interest that are found when one disaggregates the data by age, gender and sector as well as along the employment and hours per worker margins. At a general level, the issue is to understand both qualitatively and quantitatively how various policies or institutions influence hours worked in an economy. Prominent examples of the policies and institutions of interest include such things as labour market regulations (for example, minimum wages, employment protection), product market regulation (for example, entry barriers), tax and transfer policies (for example, unemployment insurance, social security, disability), wage setting practices (for example, importance of unions, centralized versus decentralized wage negotiations). In addition to the papers mentioned earlier, other examples of work that address some of these issues are Bertrand and Kramarz (2002) and Messina (2006) for entry barriers, and Bentolila and Bertola (1990) and Hopenhayn and Rogerson (1993) for firing taxes.

Economists have also recently begun to examine differences in time allocations across countries. If it is the case that individuals in one country spend much more time in market work than individuals in another country, where does this show up in terms of other uses of time? Specifically, to what extent do these differences reflect differences in leisure versus differences in time devoted to home production? In a provocative paper, Freeman and Schettkat (2001) analysed time use data for American and German couples in the 1990s and found that total working time was similar across the two economies, but that there was a systematic difference in the composition of this total working time: couples in the United States devoted more time to market work than German couples, whereas German couples devoted more time to home production than American couples. This finding, coupled with the fact that additional time use surveys have been initiated both in Europe and the United States, has spawned a growing literature on the general issue of time allocations across countries, including Freeman and Schettkat (2005) and Hamermesh, Burda and Weil (2007). Olovsson (2004), Ragan (2005) and Rogerson (2007) all argue that modelling home production is important in understanding the differences in hours of market work across countries.

Conclusions

Labour market outcomes differ dramatically across industrialized countries along several dimensions, including hours of work and unemployment. These differences have changed dramatically over time. Understanding the source of these differences and their evolution over time is a key challenge for economists, and will likely have important consequences for policymakers as well as yield important insights regarding the role of various factors in shaping labour market outcomes. This article has provided a brief introduction to this line of research. While much progress has been made to date, much more work remains to be done.

See Also

- hours worked (long run trends)
- labour market institutions
- labour supply
- search models of unemployment
- unemployment

Bibliography

Alesina, A., Glaeser, E. and Sacerdote, B. 2005. Work and leisure in the US and Europe: why so different? In *NBER Macroeconomics Annual*, vol. 20, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Alvarez, F. and Veracierto, M. 1999. Labor market policies in an equilibrium search model. In *NBER Macroeconomics Annual*, vol. 14, ed. B.S. Bernanke and J.J. Rotemberg.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Cambridge, MA: MIT Press.

Bentolila, S. and Bertola, G. 1990. Firing costs and labor demand: how bad is eurosclerosis? *Review of Economic Studies* 57, 381–402.

Bertola, G. and Ichino, A. 1995. Wage inequality and unemployment: United States vs. Europe. In *NBER Macroeconomics Annual*, vol. 10, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge MA: MIT Press.

Bertrand, M. and Kramarz, F. 2002. Does entry regulation hinder job creation? Evidence from the French retail industry. *Quarterly Journal of Economics* 117, 1369–413.

Blanchard, O. 2005. The economic future of Europe. *Journal of Economic Perspectives* 18(4), 3–26.

Blanchard, O. and Giavazzi, F. 2002. Macroeconomic effects of regulation and deregulation in goods and labor markets. *Quarterly Journal of Economics* 117, 879–907.

Blanchard, O. and Wolfers, J. 2000. The role of shocks and institutions in the rise of European unemployment: the aggregate evidence. *Economic Journal* 110, 1–33.

Card, D., Kramarz, F. and Lemieux, T. 1999. Changes in relative structure of wages and employment: a comparison of the United States, Canada and France. *Canadian Journal of Economics* 32, 843–77.

Chang, Y. and Kim, S. 2006. From individual to aggregate labor supply: a quantitative analysis based on a heterogeneous agent macroeconomy. *International Economic Review* 47, 1–27.

Daveri, F. and Tabellini, G. 2000. Unemployment, growth and taxation in industrial countries. *Economic Policy* 15, 47–104.

Davis, S. and Henrekson, M. 2005. Tax effects on work activity, industry mix and shadow economy size: evidence from rich country comparisons. In *Labour Supply and Incentives to Work in Europe*, ed. R. Goméz-Salvador, A. Lamo, B. Petrongolo, M. Ward and E. Wasmer. Northampton, MA: Edward Elgar.

Freeman, R. and Schettkat, R. 2001. Marketization of production and the US–Europe employment gap. *Oxford Bulletin of Economics and Statistics* 63, 647–70.

Freeman, R. and Schettkat, R. 2005. Marketization of household production and the EU–US gap in work. *Economic Policy* 20, 6–50.

Hamermesh, D., Burda, M. and Weil, P. 2007. The distribution of total work in the EU and the US. In *Are Europeans Lazy or Americans Crazy?* ed. T. Boeri. Oxford: Oxford University Press.

Hopenhayn, H. and Rogerson, R. 1993. Job turnover and policy evaluation: a general equilibrium analysis. *Journal of Political Economy* 101, 915–38.

Hornstein, A., Krusell, P. and Violante, G. 2007. Technology–policy interactions in frictional labor markets. *Review of Economic Studies* 74, 1089–124.

Krugman, P. 1994. Past and prospective causes of high unemployment. *Economic Review of the Federal Reserve Bank of Kansas City* 79(4th quarter), 23–43.

Ljungqvist, L. and Sargent, T. 1998. The European unemployment dilemma. *Journal of Political Economy* 106, 514–50.

Ljungqvist, L. and Sargent, T. 2007. Do taxes explain European unemployment? Indivisible labor, human capital, lotteries and savings. Working paper, New York University.

Marimon, R. and Zilibotti, F. 1999. Unemployment versus mismatch of talents: reconsidering unemployment benefits. *Economic Journal* 109, 266–91.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Messina, J. 2006. The role of product market regulations in the process of structural change. *European Economic Review* 50, 1863–90.

Mortensen, D. and Pissarides, C. 1999. Unemployment responses to skill-biased technology shocks. *Economic Journal* 109, 242–65.

Nickell, S., Nunziata, L. and Ochel, W. 2006. Unemployment in the OECD since the 1960s. What do we know? *Economic Journal* 115, 1–27.

Ohanian, L., Raffo, A. and Rogerson, R. 2006. Long-term changes in labor supply and taxes: evidence from OECD countries, 1956–2004. Working Paper No. 12786. Cambridge, MA: NBER.

Olovsson, C. 2004. Why do Europeans work so little? Mimeo, Stockholm School of Economics.

Pissarides, C. 2007. Unemployment and hours of work: the North Atlantic divide revisited. *International Economic Review* 48, 1–36.

Prescott, E. 2004. Why do Americans work so much more than Europeans? *Quarterly Review of the Federal Reserve Bank of Minneapolis* 28(1), 2–13.

Prescott, E. 2006. The transformation of macroeconomic policy and research, 2004 Nobel Prize address. *Journal of Political Economy* 114, 203–36.

Pries, M. and Rogerson, R. 2005. Hiring policies, labor market institutions and labor market flows. *Journal of Political Economy* 113, 811–39.

Ragan, K. 2005. Taxes, transfers and time use: fiscal policy in a household production model. Mimeo, University of Chicago.

Rogerson, R. 2006. Understanding differences in hours worked. *Review of Economic Dynamics* 9, 365–409.

Rogerson, R. 2007. Structural transformation and the deterioration of European labor market outcomes. Working Paper No. 12889. Cambridge, MA: NBER.

Rogerson, R. and Wallenius, J. 2007. Micro and macro elasticities in a life cycle model with taxes. Working Paper No. 13017. Cambridge, MA: NBER.

How to cite this article

Rogerson, Richard. "unemployment and hours of work, cross country differences." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_C000586> doi:10.1057/9780230226203.1759

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

unemployment insurance

Patricia M. Anderson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Unemployment insurance (UI) is a social insurance programme in which compensation is paid to unemployed workers. Much of the research on UI has focused on the inherent disincentives. For example, higher benefits have been found to increase unemployment durations, with little clear positive impact on the quality of new jobs. Additionally, financing UI through payroll taxes that are not completely experience-rated provides an incentive for firms to lay off workers. Thus, while UI is an important safety net for unemployed workers, it may also increase unemployment overall.

Keywords

adjustment costs; consumption smoothing; labour demand; labour supply; layoffs; reservation wage; search models of unemployment; tax incidence; unemployment; unemployment insurance

Article

Unemployment insurance (UI) is a social insurance programme whereby compensation is paid to unemployed workers. The federal–state UI programme in the United States dates from the Social Security Act of 1935, while many European countries began national programmes even earlier. For example, the National Insurance Act of 1911 established UI in the United Kingdom, while Italy and Germany established programmes in 1919 and 1927, respectively.

Institutional aspects of UI

While specific institutional details differ across countries, a typical UI programme is limited to workers who are unemployed through no fault of their own (that is, who have neither quit nor been fired for cause), who are actively looking for work, and who have a demonstrated attachment to the labour force, with the benefits being paid for a limited period of time (see for example, Atkinson and Micklewright,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1991). Typically, the weekly benefit amount (WBA) is based on previous earnings, using a replacement rate schedule that is subject to a minimum and maximum WBA.

Financing of UI programmes also differs across countries. In some cases it is funded out of general revenues, while in most countries it is funded by a flat tax levied on employers and sometimes on employees (see Storey and Niesner, 1997 for a complete overview of UI programmes in the G-7 nations). Empirical evidence from the United States, though, shows that even when the tax is levied solely on the employer, the incidence is largely on the employee (Anderson and Meyer, 1997; 2000). In the United States, the employer tax is also experience-rated. That is, a firm's tax rate depends on the use of the UI system by its previous employees, with new firms typically charged a rate based on industry experience for the first few years. While each state has its own institutions, a typical system can be characterized by thinking of each firm as having an account with the state UI authority. Taxes are paid into this account, and benefits are paid out of it when the firm lays off employees. A schedule then relates the balance in this account to a tax rate, subject to minimum and maximum rates. A high account balance will merit a lower tax rate, while a lower (possibly even negative) balance will merit a higher tax rate. This characterization is a gross simplification, but captures the basic components of an experience-rated system.

Economic incentives of UI

With experience rating, a firm which lays off a worker today can expect to pay some fraction of that worker's benefits through higher tax payments in the future (for early theoretical work, see Feldstein, 1976; Baily, 1976; Brechling, 1977a; 1977b). For the most common systems used in the United States, one can calculate this marginal tax cost (MTC) of a layoff (see for example, Topel, 1983; this derivation follows Anderson and Meyer, 1993). Let θ be the growth of employment (that is, $N_{t+1} = \theta N_t$), γ be the growth in the taxable wage base (that is, $W_{t+1} = \gamma W_t$), and approximate the tax schedule as a linear relationship with slope η. The tax bill can then be expressed in terms of benefits paid, N, W and the parameters γ and η. For interest rate *i*, when benefits paid increase by a dollar, the present value of the change in this tax bill is

$$\sum_{t=1}^{\infty} \frac{\theta^2\gamma^2\eta}{1-\theta^2\gamma^2\eta}\left(\frac{1-\theta^2\gamma^2\eta}{1+i}\right)^t.$$

This sum converges to

$$\frac{\theta^2\gamma^2\eta}{i+\theta^2\gamma^2\eta},$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which is referred to as the marginal tax cost (MTC) of a layoff. For firms at the maximum (or minimum) tax rate, $\eta = 0$ and this MTC will be zero, implying that benefits to this firm's workers are completely subsidized. Alternatively, the steeper the slope of the tax schedule the closer this marginal tax cost is to 1, and the more perfectly experience-rated is the system.

The subsidy inherent from incomplete experience rating provides an incentive to lay off workers. In fact, Topel (1983) estimates that incomplete experience rating in the United States could be responsible for about one-third of temporary-layoff unemployment spells. More broadly, the MTC can be thought of as a simple adjustment cost, implying that tighter experience rating would not only reduce layoffs in a downturn, but also reduce hiring in a boom, resulting in decreased employment fluctuations (see for example, Anderson, 1993; Card and Levine, 1994; Anderson and Meyer, 1994).

More work exists on the employee disincentives of UI than the firm disincentives. Two simple models imply that higher benefit levels will result in longer unemployment durations. First, as shown in Moffitt and Nicholson (1982), incorporating UI into the budget constraint of a labour supply model results in income and substitution effects which both imply fewer weeks worked. Similarly, a simple job-search model implies that higher UI benefits lower the cost of unemployment, thus increasing the reservation wage and lowering the probability of accepting a new job. The result is again an increase in the duration of unemployment.

The simple job-search model is extended in Mortensen (1977) to incorporate realistic UI programme features such as minimum work requirements for initial qualification, and limited duration of benefits. In this model, for the unemployed who are not currently qualified to receive UI, a new job that could lead to future UI qualification is more valuable the higher the benefit level. In this case, a negative relationship between duration and benefit level would be expected. Allowing benefits to be of limited duration has additional implications as well. First, the level of benefits should have no effect on the reservation wage after they run out. However, search intensity may increase around exhaustion, implying that potential duration of benefits may have a direct effect on unemployment duration.

The economic effects of UI are not all negative. For example, a job-search model also implies that increased duration should result in higher-quality jobs being found. Additionally, UI benefits can allow individuals to smooth consumption during unemployment spells. Finally, this consumption smoothing benefit implies UI can also play an important role as a macroeconomic stabilizer by helping maintain aggregate spending. More broadly, there is a growing literature on optimal UI which takes into account the need to balance costs and benefits (see for example, Baily, 1978; Chetty, 2006, for partial equilibrium and Hopenhayn and Nicolini, 1997; and Acemoglu and Shimer, 1999 for general equilibrium analyses).

Empirical evidence on the effects of UI

Most empirical work on UI focuses on the costs, with studies generally confirming a positive relationship between duration and benefit levels. For example, studies of UI recipients in the United Kingdom have found elasticities of around 0.3, while those in the United States have found elasticities in the range of 0.4 to 1. Studies in other OECD countries have typically found relatively low elasticities,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

although they tend to be measured without much precision (see Atkinson and Micklewright, 1991 for a review of these studies). Additionally, empirical studies that allow for the exit rate out of unemployment to 'spike' around the time of benefit exhaustion have found just such an effect both overall (for example, Meyer, 1990) and for new jobs and recalls separately (for example, Katz and Meyer, 1990).

Additionally, studies in the United States have tended to find that a one-week increase in potential duration results in between a 0.1 and 0.2 week increase in unemployment, with Canadian studies finding slightly larger effects (Atkinson and Micklewright, 1991).

There are fewer empirical studies of the benefits of UI. A notable exception is Gruber (1997), which finds a large consumption smoothing effect of higher benefits. Finally, while a job-search model implies that higher-quality jobs should be found, empirical evidence is largely mixed on this effect (see Decker, 1997 for a review of US studies of benefit effects). In particular, in the United States, several re-employment bonus experiments took place in which UI claimants were offered a cash bonus if they found a new job within a specific amount of time (see Meyer, 1995, for a review). The early experiments found significant reductions in unemployment durations, but no real decline in post-unemployment earnings as would be expected if benefits were significantly subsidizing search. Overall, then, while UI is an important safety net for unemployed workers, it may also increase unemployment.

See Also

- adjustment costs
- layoffs
- social insurance
- unemployment

Bibliography

Acemoglu, D. and Shimer, R. 1999. Efficient unemployment insurance. *Journal of Political Economy* 107, 893–928.

Anderson, P.M. 1993. Linear adjustment costs and seasonal labor demand: evidence from retail trade firms. *Quarterly Journal of Economics* 108, 1015–42.

Anderson, P.M. and Meyer, B.D. 1993. Unemployment insurance in the United States: layoff incentives and cross subsidies. *Journal of Labor Economics* 11(1, Part II), S70–S95.

Anderson, P.M. and Meyer, B.D. 1994. The effects of unemployment insurance taxes and benefits on layoffs using firm and individual data. Working Paper No. 4960. Cambridge, MA: NBER.

Anderson, P.M. and Meyer, B.D. 1997. The effects of firm specific taxes and government mandates with an application to the U.S. unemployment insurance program. *Journal of Public Economics* 65, 119–45.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Anderson, P.M. and Meyer, B.D. 2000. The effects of the unemployment insurance payroll tax on wages, employment, claims and denials. *Journal of Public Economics* 78, 81–106.

Atkinson, A.B. and Micklewright, J. 1991. Unemployment compensation and labor market transitions: a critical review. *Journal of Economic Literature* 29, 1679–727.

Baily, M. 1976. On the theory of layoffs and unemployment. *Econometrica* 45, 1043–63.

Baily, M.N. 1978. Some aspects of optimal unemployment insurance. *Journal of Public Economics* 10, 379–402.

Brechling, F. 1977a. The incentive effects of the U.S. unemployment insurance tax. In *Research in Labor Economics*, ed. R. Ehrenberg. Greenwich, CT: JAI Press.

Brechling, F. 1977b. Unemployment insurance and labor turnover: summary of theoretical findings. *Industrial and Labor Relations Review* 30, 483–94.

Card, D. and Levine, P.B. 1994. Unemployment insurance taxes and the cyclical and seasonal properties of unemployment. *Journal of Public Economics* 53, 1–29.

Chetty, R. 2006. A general formula for the optimal level of social insurance. *Journal of Public Economics* 90, 1879–901.

Decker, P.T. 1997. Work incentives and disincentives. In *Unemployment Insurance in the United States: Analysis of Policy Issues*, ed. C.J. O'Leary and S.A. Wandner. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Feldstein, M.S. 1976. Temporary layoffs in the theory of unemployment. *Journal of Political Economy* 84, 937–57.

Gruber, J. 1997. The consumption smoothing benefits of unemployment insurance. *American Economic Review* 87, 192–205.

Hopenhayn, H.A. and Nicolini, J.P. 1997. Optimal unemployment insurance. *Journal of Political Economy* 105, 412–38.

Katz, L.F. and Meyer, B.D. 1990. Unemployment insurance, recall expectations, and unemployment outcomes. *Quarterly Journal of Economics* 105, 973–1002.

Meyer, B.D. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58, 757–82.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Meyer, B.D. 1995. Lessons from the U.S. unemployment insurance experiments. *Journal of Economic Literature* 33, 91–131.

Moffitt, R. and Nicholson, W. 1982. The effect of unemployment insurance on unemployment: the case of federal supplemental benefits. *Review of Economics and Statistics* 64, 1–11.

Mortensen, D.T. 1977. Unemployment insurance and job search decisions. *Industrial and Labor Relations Review* 30, 505–17.

Storey, J.R. and Niesner, J.A. 1997. Unemployment compensation in the group of seven nations. In *Unemployment Insurance in the United States: Analysis of Policy Issues*, ed. C.J. O'Leary and S.A. Wandner. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Topel, R.H. 1983. On layoffs and unemployment insurance. *American Economic Review* 73, 541–59.

How to cite this article

Anderson, Patricia M. "unemployment insurance." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000065>
doi:10.1057/9780230226203.1760

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

unemployment measurement

Katharine Bradbury
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Measures of unemployment tally people without a job who are looking for one. For measurement purposes, the critical question is what constitutes ‘looking’. This article summarizes how unemployment is measured in the United States and Europe, and describes recent research investigating the permeability of the dividing line between the unemployed and ‘marginally attached’ subgroups of those out of the labour market. A continuum between unemployed and entirely inactive individuals indicates that additional measures beyond unemployment may be useful in judging the state of the labour market.

Keywords

labour force participation rate; labour market search; non-employment; unemployment; unemployment measurement

Article

Measures of unemployment attempt to count individuals who do not have a job but are looking for one. While the concept is reasonably straightforward, various measurement approaches are used to distinguish those out of work who are looking from those who are not, generally based on the specific methods these individuals use to search for employment, how intensively they search, and how long it has been since they ‘actively’ searched.

Official measure in the United States

In the United States, unemployment is gauged by comparing the number of unemployed individuals with the size of the labour force, as determined by a monthly survey of households. The civilian labour force is defined as individuals aged 16 and older who are either employed or unemployed, but not on active duty in the armed forces. (See U.S. Bureau of Labor Statistics, 2006; Jacobs, 2006, pp. 4–8.)

Individuals are considered unemployed if (*a*) they lack a job, (*b*) are available to work, and (*c*) have actively sought employment in the four weeks preceding the survey or are awaiting call-back to an existing job (even if they did not actively seek employment). Active job search includes contacting employers or employment agencies, sending out résumés, or placing or answering advertisements; simply reading want ads is not considered active job search. Individuals are employed if they worked at least one hour as paid employees during the reference week, or worked in their own business or farm, or worked unpaid for 15 hours or more in a family business, or had a job but were temporarily away from it due to illness, vacation, labour management disputes, parental leave, job training, or other personal or family related reasons.

Anyone in the civilian non-institutional population (ages 16 and older and neither in the military nor in an institution such as a prison or mental hospital) who is not employed and not unemployed is considered to be out of the labour force. The U.S. Bureau of Labor Statistics (BLS) collects information on those out of the labour force to assess the degree to which they may be ‘marginally attached’ to the labour force, by asking about their interest in and availability for work.

Based on these questions, the BLS defines a set of ‘alternative measures of labor underutilization’ that either subtract from the official unemployment rate (for example, by counting only the long-term unemployed) or add to it. The nature of the questions and the alternative measures were revised as of January 1994. Figure 1 plots two of these ‘alternative

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

measures' and a third concept, along with the official unemployment rate, over the 1994–2006 period. The line marked 'discouraged' adds to the official unemployed all discouraged workers, defined as those who have given a job market-related reason for not currently looking for a job, including people who think that no work is available, or who could not find work, or who lack schooling or training, or who say potential employers think they are too young or old, or who believe they have been subject to other types of discrimination. The next measure adds all other marginally attached workers, defined as persons who currently are neither working nor looking for work but indicate that they want, and are available for, a job and have looked for work at some time in the preceding 12 months.

Figure 1

US measures of labour underutilization, 1994–2006. *Notes*: Not seasonally adjusted, 12-month centred moving averages. Discouraged workers are a subset of the marginally attached; the marginally attached are a subset of those who say they want a job. Each measure adds the noted group to 'officially' unemployed individuals and expresses that sum as a per cent of the labour force plus the noted group. *Sources*: US Bureau of Statistics and author's calculations.

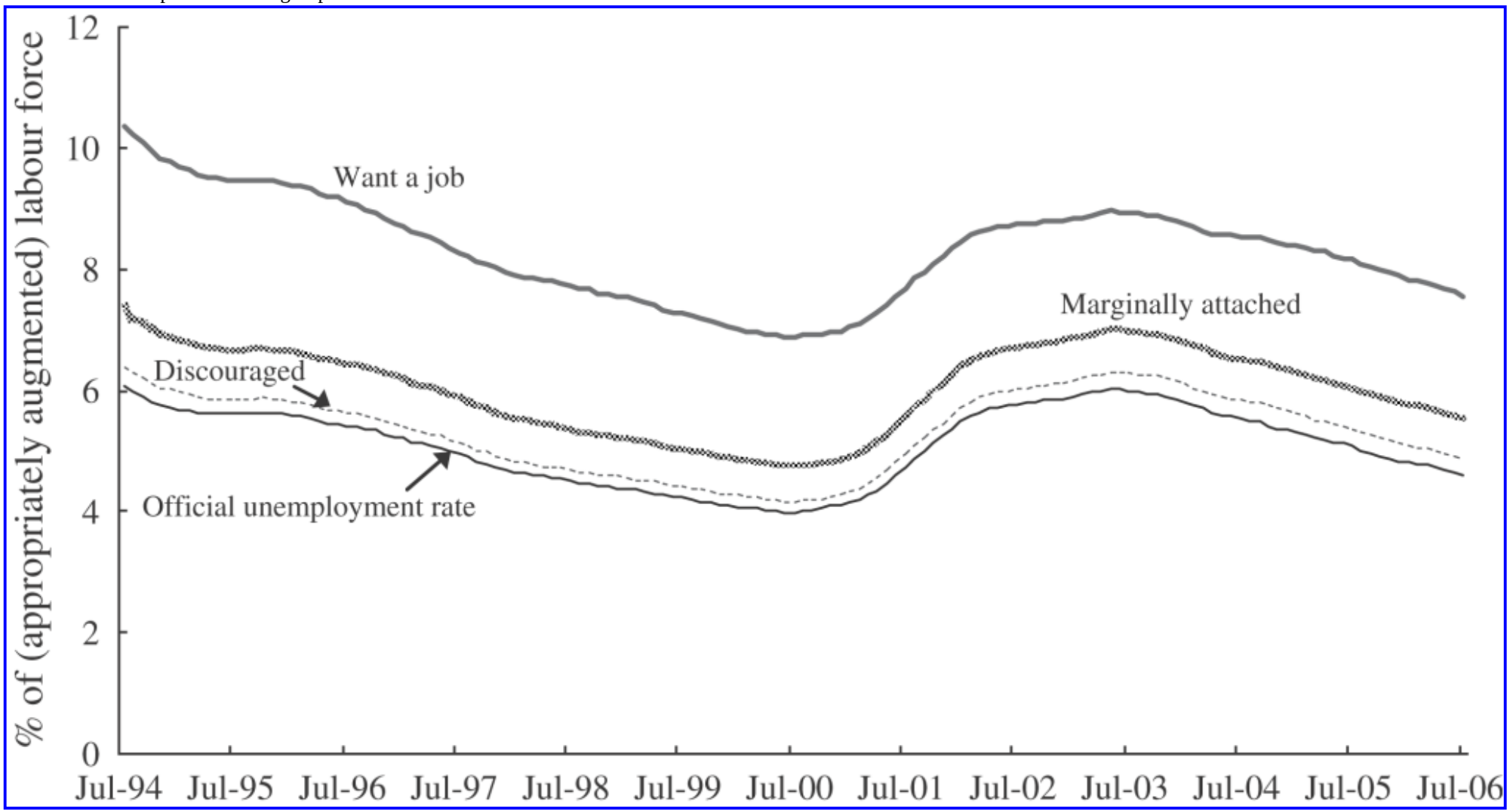

Beyond what the BLS defines as 'marginally attached' is a group who has not looked for work in the last 12 months but answers 'yes' to the question 'do you currently want a job?' This line is labelled 'want a job' in the figure. The alternative measures add from a few tenths of a percentage point (discouraged workers) to a percentage point (marginally attached) to several percentage points (want a job) to the official unemployment rate. While too small to be seen clearly in the figure, the number of individuals in these marginal categories varied cyclically over the 1994–2005 period in a manner similar to the number unemployed.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Unemployment measures elsewhere in the industrialized world

In most of the industrialized world, the concept of unemployment is the same and the definition for measurement purposes is quite similar. The dimensions along which measures differ include the type of job search activities that distinguish the unemployed from the marginally attached, the definition of 'currently available' for work, and the treatment of individuals on layoff or waiting to start a new job. In addition, age cut-offs may vary. (For more detailed discussion of inter-country definitional differences, see Sorrentino, 2000.)

The measurement definition chosen by the Statistical Office of the European Communities (Eurostat) in September 2000 is used to calculate 'standardized' unemployment rates for member nations. Eurostat conducts the European Union Labor Force Survey (EU LFS) on a quarterly basis and also releases 'harmonized' monthly estimates based on data from member states. The EU LFS divides the population of working age (defined as ages 15 and older) into three groups: employed, unemployed, and inactive. The economically active population (like the labour force in the United States) comprises employed and unemployed persons.

According to Eurostat, the unemployed are those aged 15–74 (or 16–74 in a few nations), who (*a*) were without work during the reference week, (*b*) were available to start work within two weeks, and (*c*) had either actively sought work in the past four weeks or had already found a job to start within the next three months. The specific steps that qualify as actively seeking work include any of the following: being in contact with a public employment office or a private agency to find work, applying to employers directly, asking among friends, relatives, unions, and so on, to find work, placing or answering job advertisements, studying job advertisements, taking a recruitment test or examination or being interviewed, or undertaking various activities to set up a business.

Thus, the EU includes those who only study job advertisements as unemployed (if they pass the other screens) – this is true in Canada as well – while the United States does not consider reading ads as active job search. In addition, persons waiting to start a new job are considered unemployed in Europe, but they are not considered unemployed in the United States unless they have actively searched for work within the previous four weeks. Persons on temporary layoff are considered unemployed in the United States even if they do not seek work, but in Europe they may be counted as employed, unemployed, or inactive, depending on search activity and the strength of their attachment to their job (based on pay and/or a definite recall date within three months).

The dividing line between being unemployed and out of the labour force

Because the distinction is necessarily arbitrary for measurement purposes, a research literature has investigated the dividing line between the unemployed and those out of the labour force. One strand of this literature focuses on transition probabilities among labour market states, investigating the degree to which those out of the labour force (or a marginally attached subset of them) are as likely to move into employment as those labelled unemployed. The issue is that measured unemployment might miss some signals of labour market tightness or slack if those out of the labour force behave similarly to the unemployed. Clark and Summers reported that 'many of those classified as not in the labor force are functionally indistinguishable from the unemployed' (1979, p. 31). Flinn and Heckman (1983), by contrast, examined young workers' transition probabilities into employment and rejected the hypothesis that the distinction between unemployment and being out of the labour force is behaviourally meaningless.

Several recent papers – including Jones and Riddell (1999; 2006) focusing on Canada, Garrido and Toharia (2004) for Spain, Brandolini, Cipollone, and Viviano (2006) for Italy, and Schweitzer (2003) for the UK – describe selected groups of individuals who are officially out of the labour force but might be considered close to the unemployed, as they are attached to the labour force in various ways. These authors test hypotheses regarding behavioral differences – most notably transition probabilities to employment over various time horizons – between the unemployed and subgroups of those out of the labour force. For example, Jones and Riddell consider those who say they want a job; Brandolini, Cipollone and Viviano examine those who searched for employment between five and eight weeks before the survey; Garrido and Toharia examine 'passive' job searchers; Schweitzer subdivides those available for or wanting a job according to their primary non-labour market activity. Jones and Riddell also subdivide the unemployed into several categories along similar dimensions, and Schweitzer distinguishes the long-term and short-term unemployed; they examine rates of transition to employment of these unemployed subgroups as well.

These researchers find that most of the marginally attached categories lie between the unemployed and the remainder of the 'inactive' group; that is, their transitions into employment are higher than the completely inactive, but still generally lower than those of the unemployed. Jones and Riddell and Schweitzer also find heterogeneity within the ranks of the unemployed and note that some marginally attached categories have higher transition rates into employment than selected subcategories of the unemployed.

Measuring labour market slack

These measurement issues are of more than academic interest because unemployment is the most widely used indicator of the degree of tightness or slack in the labour market and, by extension, in the overall economy; as a consequence, it is used by policymakers as a key signal of potential inflationary pressures. The research discussed above points to a continuum of labour market attachment among the non-employed, from those classified as unemployed through various marginally attached groups to people who expressly do not want a job. Some of the research authors argue that unemployment should be defined and measured more inclusively than it is currently. More generally, the arbitrariness of the dividing line between the states of being unemployed and out of the labour force, together with heterogeneity among subgroups within the inactive population, suggest that policymakers might

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

gain useful information by looking at a range of measures – along with the official unemployment rate – in judging the state of the labour market.

Because the relationship between the measured unemployment rate and 'true' economic slack and hence inflation may vary, depending on the specific definitions used in measuring unemployment, potential labour market entrants, the age and gender composition of the population, and labour market institutions, researchers have developed and investigated a variety of alternative indicators of labour market slack. One set of alternative measures sidesteps the difficulty of choosing a dividing line between the unemployed and inactive population by concentrating on the distinction between employment and non-employment. For example, the European Council revised its labour market targets in 2000, replacing the goal of reducing unemployment with the goal of increasing employment rates (employment–population ratios) (European Parliament, 2000). Similarly, Juhn, Murphy, and Topel (1991; 2002), and Murphy and Topel (1997) analyse non-employment and argue that 'the unemployment rate has become progressively less informative about the state of the labor market' (1997, p. 295). Others consider the labour force participation rate an indicator of interest along with the unemployment rate (for example, Anderson, Barrow, and Butcher, 2005; Bradbury, 2005). Complementary approaches consider a variety of direct measures of labour market tightness, either individually (for example, Shimer's 2005 job-finding rate among the unemployed) or in combination (for example, a composite measure of US labour market tightness compiled by Barnes, Chahrour, Olivei and Tang, 2007).

See Also

- labour supply
- natural rate of unemployment
- unemployment

Bibliography

Anderson, K., Barrow, L. and Butcher, K.F. 2005. Implications of changes in men's and women's labor force participation for real compensation growth and inflation. *Topics in Economic Analysis & Policy* 5(1) (Article 7).

Barnes, M., Chahrour, R., Olivei, G. and Tang, G. 2007. A principal components approach to estimating labour market pressure and its implicaitons for inflation. Federal Reserve Bank of Boston Public Policy Brief Series, No. 07-2.

Bradbury, K. 2005. Additional slack in the economy: the poor recovery in labor force participation during this business cycle. Public Policy Brief No. 05–2. Boston, MA: Federal Reserve Bank of Boston.

Brandolini, A., Cipollone, P. and Viviano, E. 2006. Does the ILO definition capture all unemployment? *Journal of the European Economic Association* 4, 153–79.

Clark, K.B. and Summers, L.H. 1979. Labor market dynamics and unemployment: a reconsideration. *Brookings Papers on Economic Activity* 1979(1), 13–72.

European Parliament. 2000. Lisbon European Council 23 and 24 March 2000 presidency conclusions. Online. Available at http://www.europarl.europa.eu/summits/lis1_en.htm, accessed 20 December 2006.

Flinn, C.J. and Heckman, J.J. 1983. Are unemployment and out of the labor force behaviorally distinct labor force states? *Journal of Labor Economics* 1, 28–42.

Garrido, L. and Toharia, L. 2004. What does it take to be (counted as) unemployed: the case of Spain. *Labour Economics* 11, 507–23.

Jacobs, E.E., ed. 2006. *Handbook of U.S. Labor Statistics*, 9th edn. Lanham, MD: Bernan Press.

Jones, S.R.G. and Riddell, W.C. 1999. The measurement of unemployment: an empirical approach. *Econometrica* 67, 147–61.

Jones, S.R.G. and Riddell, W.C. 2006. Unemployment and nonemployment: heterogeneities in labor market states. *Review of Economics and Statistics* 88, 314–23.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Juhn, C., Murphy, K.M. and Topel, R. 1991. Why has the natural rate of unemployment increased over time? *Brookings Papers on Economic Activity* 1991(2), 75–142.

Juhn, C., Murphy, K.M. and Topel, R. 2002. Current unemployment, historically contemplated. *Brookings Papers on Economic Activity* 2002(1), 79–116.

Murphy, K.M. and Topel, R. 1997. Unemployment and nonemployment. *American Economic Review* 87, 295–300.

Schweitzer, M. 2003. Ready, willing, and able? Measuring labour availability in the UK. Working Paper No. 03-03. Cleveland, OH: Federal Reserve Bank of Cleveland.

Shimer, R. 2005. Reassessing the ins and outs of unemployment. Mimeo, Department of Economics, University of Chicago.

Sorrentino, C. 2000. International unemployment rates: how comparable are they? *Monthly Labor Review* 123, 3–20.

U.S. Bureau of Labor Statistics. 2006. How the government measures unemployment. Online. Available at http://ww.bls.gov/cps/cps_htgm.htm, accessed 20 December 2006.

How to cite this article

Bradbury, Katharine. "unemployment measurement." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000074> doi:10.1057/9780230226203.1761

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

unemployment

Robert Topel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The *unemployed* are individuals who are without work but who are actively seeking employment. The *unemployment rate* is the percentage of the labour force – the total number of people either working or seeking work – that is unemployed. The evidence suggests that the ‘natural rate’ of unemployment (or non-employment) is not a constant towards which the labour market converges; rather, it varies with labour market fundamentals.

Keywords

disability insurance; labour market institutions; layoffs; natural rate of unemployment; Phillips curve; search models of unemployment; unemployment; unemployment insurance

Article

The *unemployed* are individuals who are without work but who are actively seeking employment. The *unemployment rate* is the percentage of the labour force – the total number of people either working or seeking work – that is unemployed. Economists and others are interested in unemployment because it says something – we are not sure exactly what – about the economic conditions generally and the success or failure of economic policy.

Some amount of unemployment is both inevitable and efficient because economic fundamentals are stochastic and information is costly. This point was memorialized in the *natural rate* hypothesis of Friedman (1968), Phelps (1974) and Alchian (1969). As Friedman put it:

> ‘The natural rate of unemployment’… is the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is embedded in them the actual structural characteristics of labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on.

This point may seem obvious today, but its origins are fairly modern – the 1960s were not so long ago in the history of economic theory. The contributions of Friedman, Phelps and Alchian were reactions to the place of unemployment in Keynesian models, which posited a stable trade-off between unemployment and inflation – the ‘Phillips curve’. But their broader impact was to establish that unemployment is an equilibrium phenomenon that occurs for the reasons stated above. This view has framed virtually all subsequent research on unemployment, and its formalization is a continuing research endeavour.

Formalization began with the search theories of McCall (1970), Mortensen (1970) and Gronau (1971); see Rogerson, Shimer and Wright (2005) for a modern survey. The subsequent ‘islands’ metaphor of Lucas and Prescott (1974) established a formal analysis of equilibrium unemployment. See search models of unemployment.

Data on unemployment are collected by government statistical agencies, based on household surveys that follow more or less uniform standards and definitions in developed countries. For example, in the United States unemployment statistics are collected monthly as part of the Current Population Survey, administered by the Bureau of Labor Statistics, which is a rotating sample of roughly 60,000 households that records (among other things) respondents’ self-reported labour market activities during the previous week. Jobless persons who have engaged in some effort to find employment in the past four weeks, or who are awaiting recall to a previous job, are recorded as unemployed. Jobless persons who

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
do not report search activities are recorded as ‘out of the labour force’.

In a dynamic economy people may be unemployed for many reasons. Young, new entrants to the labour force seek employment, and, as in marriage, it is typically not a good strategy to take the first opportunity that comes along. Other unemployed individuals may have left their previous job to look for something better, or they may have been permanently laid off from a previous job because of changing market conditions. Still others may be on temporary layoff, anticipating recall by their previous employer. These examples demonstrate that unemployment (and other labour force ‘states’) is inherently dynamic: the *stock* of unemployed is ever-changing, and is determined by labour market *flows*. New individuals are constantly joining the ranks of the unemployed via quits, layoffs, or entry to the labour force – the *inflow* to unemployment – while other unemployed job seekers locate and accept new jobs, or choose to stop looking – the *outflow* from unemployment. Changes in either flow affect the level of unemployment.

No brief overview can do justice to the vast literature on unemployment, nor can it evaluate the myriad social polices – such as unemployment insurance, public employment agencies or ‘active’ labour market programmes – that are meant to reduce unemployment or soften its impact on individuals. (Layard, Nickell and Jackman, 1991, is a slightly dated survey of key issues.) So my aims are more modest. I will summarize key facts about unemployment in the United States, and the factors that have affected the evolution of unemployment, while drawing parallels with other developed economies. Following the arguments in Juhn, Murphy and Topel (JMT) (1991; 2002) and Murphy and Topel (1987; 1997), I provide evidence of a long-term decline in the relative demands for less-skilled workers, so the rewards to employment have declined for marginal workers. This would raise the ‘natural rate’ of unemployment, but the story is complicated by the fact that some of the unemployed eventually leave the labour force. Over the long run, these changes in economic fundamentals increase joblessness – the total of unemployment and non-participation. This means that current unemployment data have a much different interpretation than in the past. For example, the US unemployment rates of 1974, 1997 and 2006 were about equal, at 4.9 per cent of the labour force. Yet non-participation among prime-aged men rose from 5.2 per cent in 1974 to 8.2 per cent in 1987 and 9.4 per cent in 2006. The reason is that many among the least-skilled had given up searching for work – they were no longer counted as unemployed, but changing labour market forces had left them jobless.

Labour market flows and the evolution of unemployment

Figures 1 and 2 show the evolution of the male unemployment rate in the United States and several other developed economies since 1965, using comparable definitions. (I present evidence on men because women's labour force participation is typically more varied.)

Figure 1

Unemployment rates in the United States and selected countries, 1965–2005. *Source*: produced by author.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

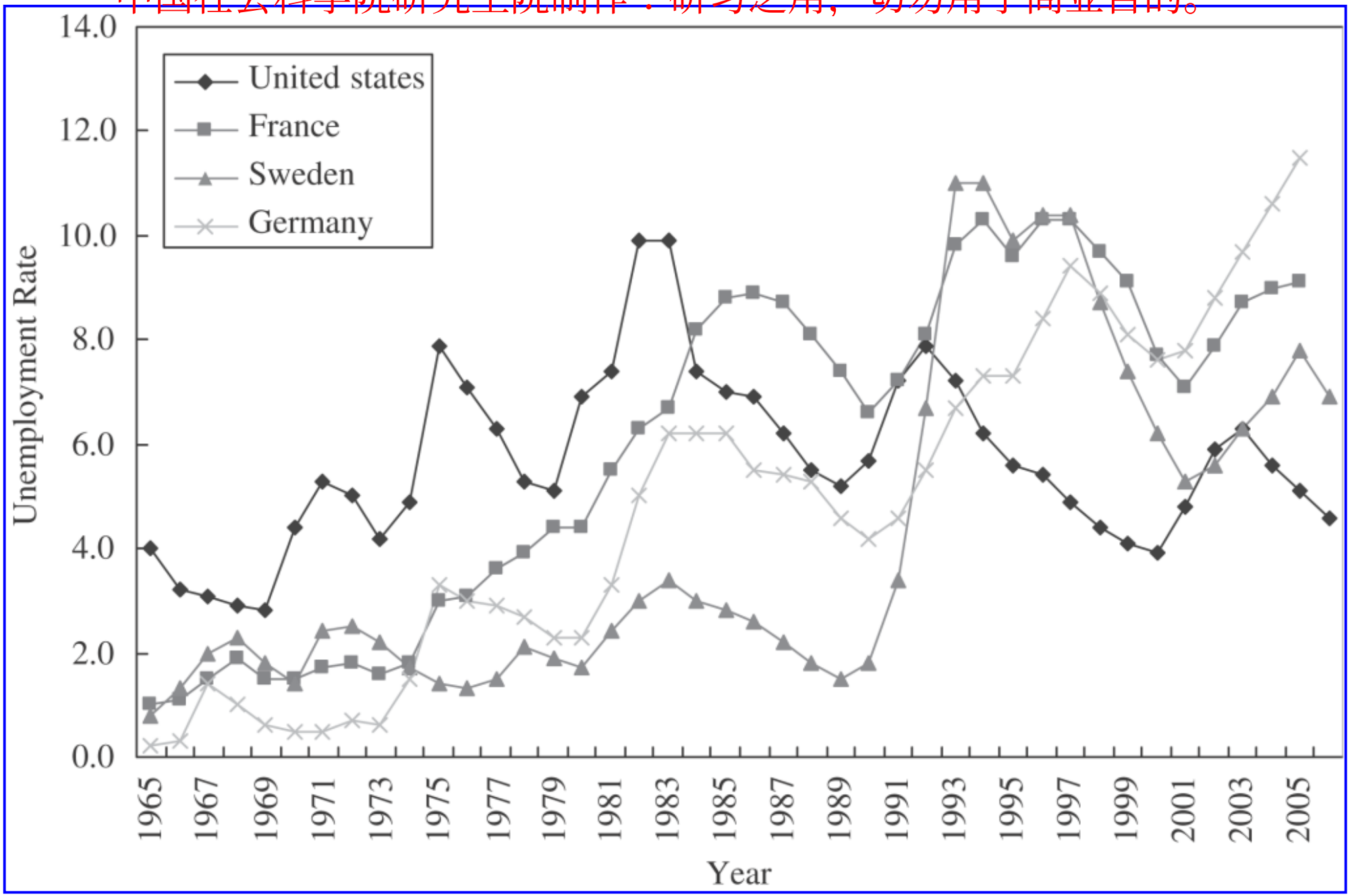

Figure 2
Unemployment rates in the United States and selected countries, 1965–2005. *Source*: produced by author.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

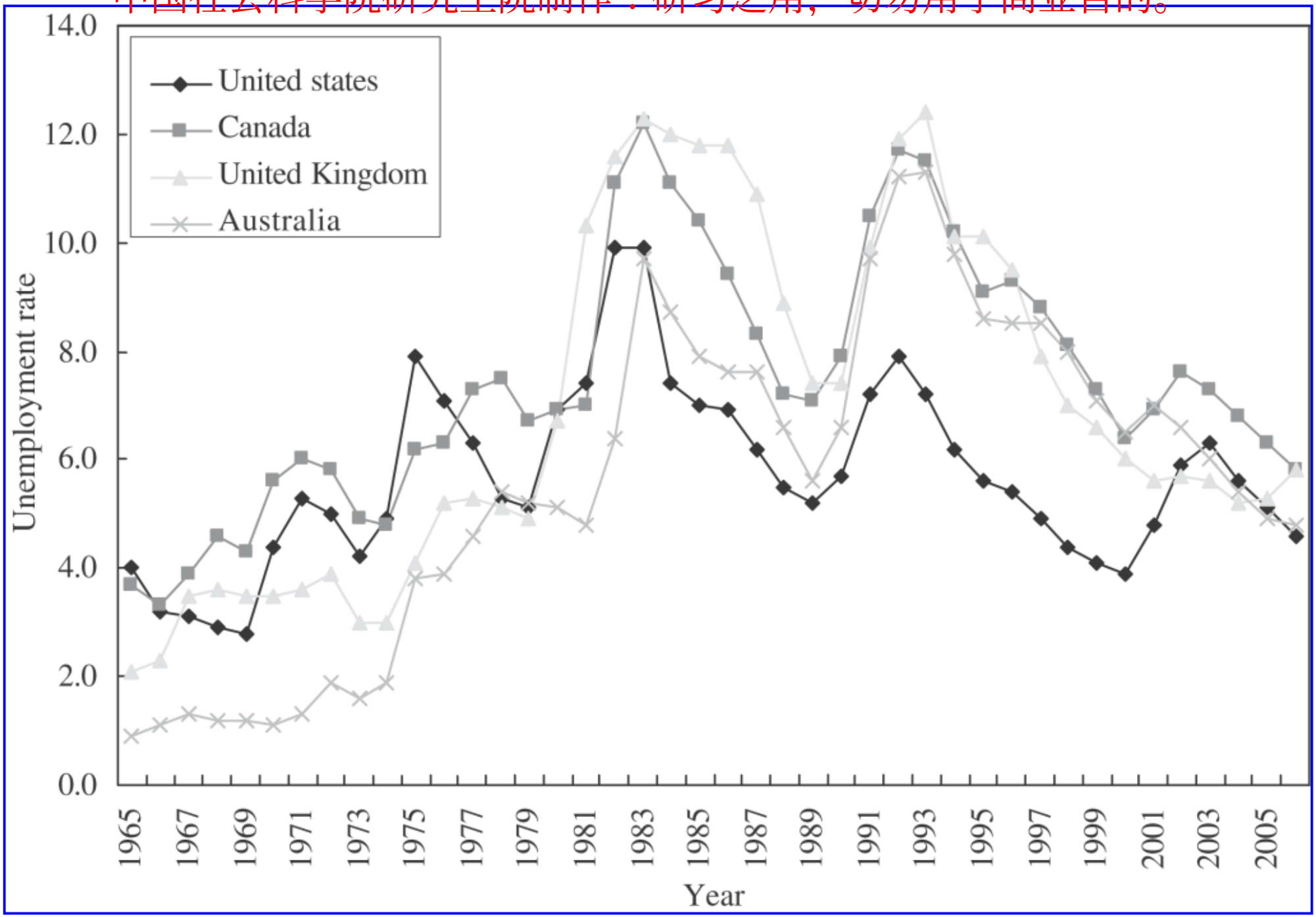

Focusing on long-term changes, the key message is that unemployment drifted upward in most industrial countries, but especially in Western Europe. The United States is an exception – in comparison with Western Europe, the United States had relatively high unemployment in the 1960s, but substantially lower relative unemployment after about 1990. Figure 1 also highlights the recent return to ‘low’ unemployment in the United States. By 2000 the US unemployment rate had reached its lowest level in 30 years, and unemployment rates in 1999–2000 were close to the extremely low rates seen during the late 1960s. This is the culmination of a long downward trend in unemployment: the peak unemployment

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
rates in the recessions of 1982–3, 1991–2 and 2001–2 were progressively lower over time, reversing the trend of rising peaks between the 1970–1, 1974–5 and 1982–3 recessions. (The recession of 1980 did not fit this pattern but as, seen in the figure, did not represent much of a peak in terms of unemployment rates.) It appears that US unemployment has come full circle: unemployment rose for 15 years (from 1968 to 1983), and then fell over the next 17 years (from 1983 to 2000), with intervening cyclical swings. One might conclude from these data that the labour market conditions of the late 1960s and late 1990s were comparable. But in fact the decline in unemployment masks more fundamental changes in labour market flows, driven largely by changes in labour demand that have affected less skilled workers.

The level of unemployment is determined by labour market flows in and out joblessness. One reason for the divergence of US and European unemployment rates is the importance of very long unemployment spells in Europe. According to data collected by the OECD, the *average* duration of unemployment spells in France, Germany or Sweden is over one year, compared with only four months in the United States. (For Sweden, I count individuals enrolled in active labour market programmes, which are required of persons who have not found employment within a fixed number of months.) For OECD Europe as a whole, about 45 per cent of all unemployment spells last more than one year, compared with only 12 per cent in the United States. This means that transitions *out of* unemployment in Europe occur more slowly, which (other transitions equal) raises the unemployment rate.

Suppose there are only two labour market 'states,' employment (*E*) and unemployment (*U*). Denote by λ_{EU} the instantaneous transition (hazard) rate from *E* to *U* – the inflow to unemployment – and let λ_{EU} be the corresponding hazard for transitions from *U* to *E* – the outflow from unemployment. If these transition rates are constant over time and across individuals, then the probability that an individual is unemployed at any date is simply:

$$\tilde{u} = \frac{\lambda_{EU}}{\lambda_{EU} + \lambda_{UE}} \quad (1)$$

Under these assumptions, eq. (1) is also the *unemployment rate* in a large population of identical individuals, with corresponding employment rate $e = 1 - \tilde{u}$. Equation (1) accords with the intuition about labour market flows stated earlier: the unemployment rate will be higher the greater the rate of inflow to unemployment, λ_{EU}, or the smaller the rate of outflow from unemployment, λ_{EU}. In this simple set-up the expected duration of an unemployment spell is $D = \lambda_{UE}^{-1}$, so policies, institutions or events that increase the duration of spells will increase measured unemployment. In an accounting sense this is why unemployment in Europe is so high – the unemployed remain so for a very long time. *Why* are unemployment durations so long in Europe and relatively short in the United States? I return to this issue below.

Equation (1) demonstrates the key elements of labour market flows, but it isn't very satisfactory as an empirical tool, for (at least) two reasons. First, as noted above, some jobless individuals are not actively seeking employment, at least by the definitions of labour market surveys, and are categorized as 'out of the labour force' (*O*). Yet many of these 'non-participants' do take jobs, and they may join the ranks of the unemployed by initiating job search. We can accommodate these facts by adding '*O*' as a third labour market state, which also adds more transition possibilities (λ_{OE}, λ_{EO}, λ_{UO} and so on). Second, labour market flows are obviously not constant – they vary over time and generate corresponding fluctuations in employment, unemployment and labour force participation. So let transition rates be time-varying (for example, λ_{EU}•(*t*) is the hazard rate from *E* to *U* at time *t*). Define *e*(*t*), *u*(*t*) and *o*(*t*) as the fractions of the relevant *population* that are employed, unemployed or out of the labour force at date *t*. (The unemployment *rate* is the fraction of the labour force that is unemployed, or $\tilde{u}(t) = \frac{u(t)}{1 - o(t)}$.) Then the law of motion for *u*(*t*) is

$$\frac{du(t)}{dt} = e(t)\lambda_{EU}(t) + o(t)\lambda_{OU}(t) - u(t)\left[\lambda_{UE}(t) + \lambda_{UO}(t)\right] \quad (2)$$

As above, changes in unemployment are driven by labour-market flows. Other things the same, the fraction of the population that is unemployed increases when transitions *to* unemployment rise. These newly unemployed individuals may have been employed (*E*) or they may be previous non-participants (*O*) who have begun to search for work. Similarly, unemployment will fall if transition rates *from* unemployment rise. One usually thinks of this in terms of greater 'job finding' (λ_{UE}•(*t*)), but (2) makes clear that transitions to non-participation – say because of deteriorating labour market opportunities that reduce the return to continued search – will also reduce unemployment.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

JMT (2002) build on (2) to calculate average entry rates and durations of spells for both unemployment (U) and non-employment ($N=U+O$) among American men from 1967 through 2000. Figure 3 shows their results for unemployment. Through the late 1980s entry rates and durations of unemployment spells showed a common pattern, rising in recessions and falling in recoveries, with some evidence of a secular increase in both components. But this tight correspondence was broken in the 1990s – increased incidence of spells played a minor role in the recession of 1991–2, while durations soared. The ensuing decline in unemployment during the 1990s expansion was driven almost entirely by a reduction in the incidence of unemployment spells, while durations of unemployment remained high – in fact, flows into unemployment fell below their levels in low-unemployment 1960s, while durations were roughly twice as long. With fewer but longer spells, the population distribution of unemployment became much more concentrated than before.

Figure 3

Entry rates and durations for unemployment, 1967–2000. *Source*: produced by author.

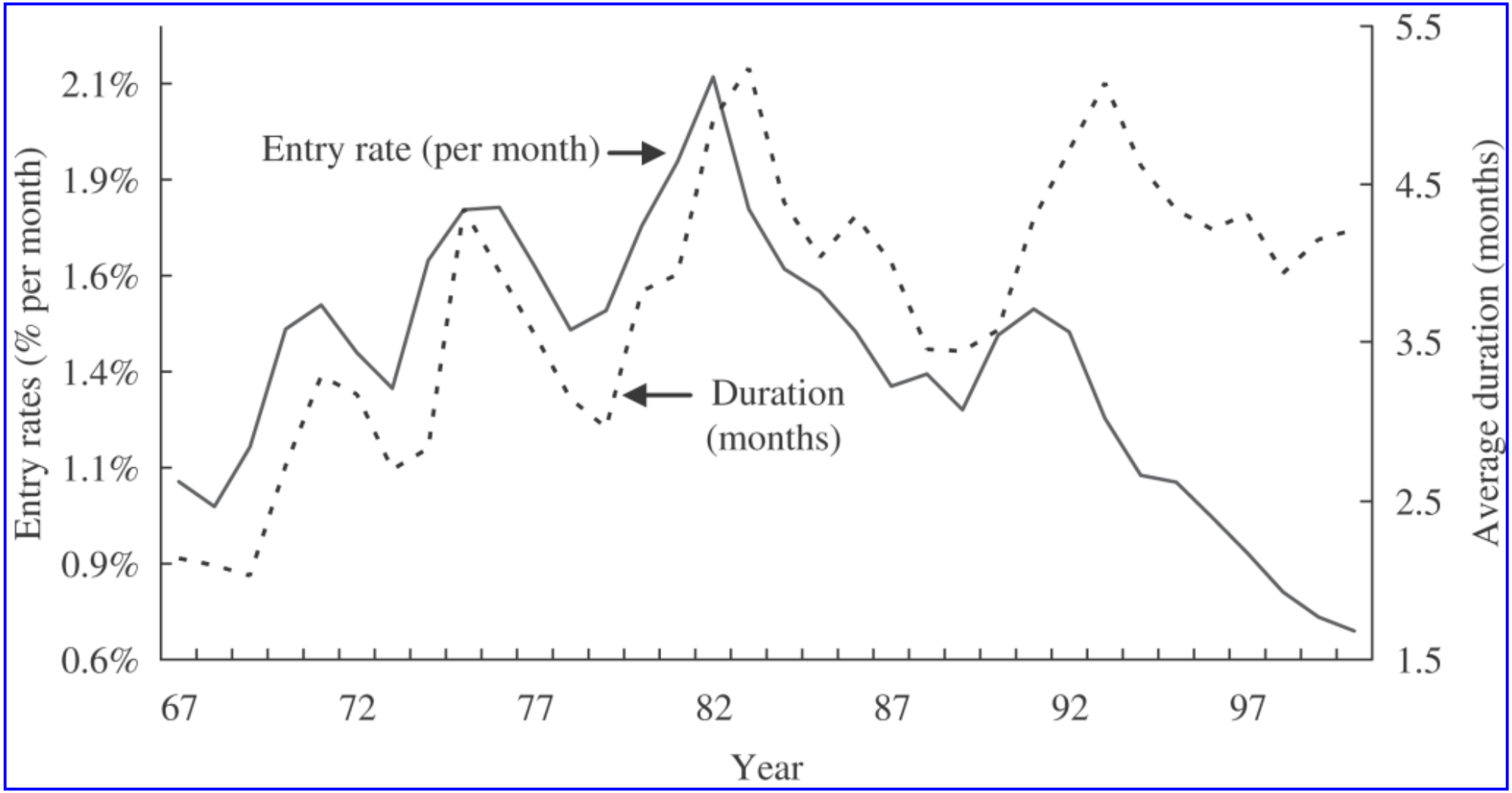

The dichotomy between incidence and duration is more extreme for non-employment, N, which measures total joblessness in the population without regard to whether individuals are searching for a job. The incidence (entry rate) of jobless spells roughly corresponds with the incidence of unemployment – compare Figure 3 – but durations of non-employment show a steady increase, more than doubling (to 15 months) since the 1960s. The sharp increase in average durations in the 1990s is especially noteworthy, reflecting the increased proportion of American men who have simply withdrawn from the labour force. Why did this occur? In an accounting sense it is because a large fraction of labour-force ‘withdrawals’ were temporary in earlier decades, but by the 1990s many men had become ‘full-time’ non-participants. They had left the labour force and made no efforts to find work, so that average transition rates from non-employment to employment plummeted (Figure 4).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Figure 4
Entry rates and durations for non-employment, 1967–2000. *Source*: produced by author.

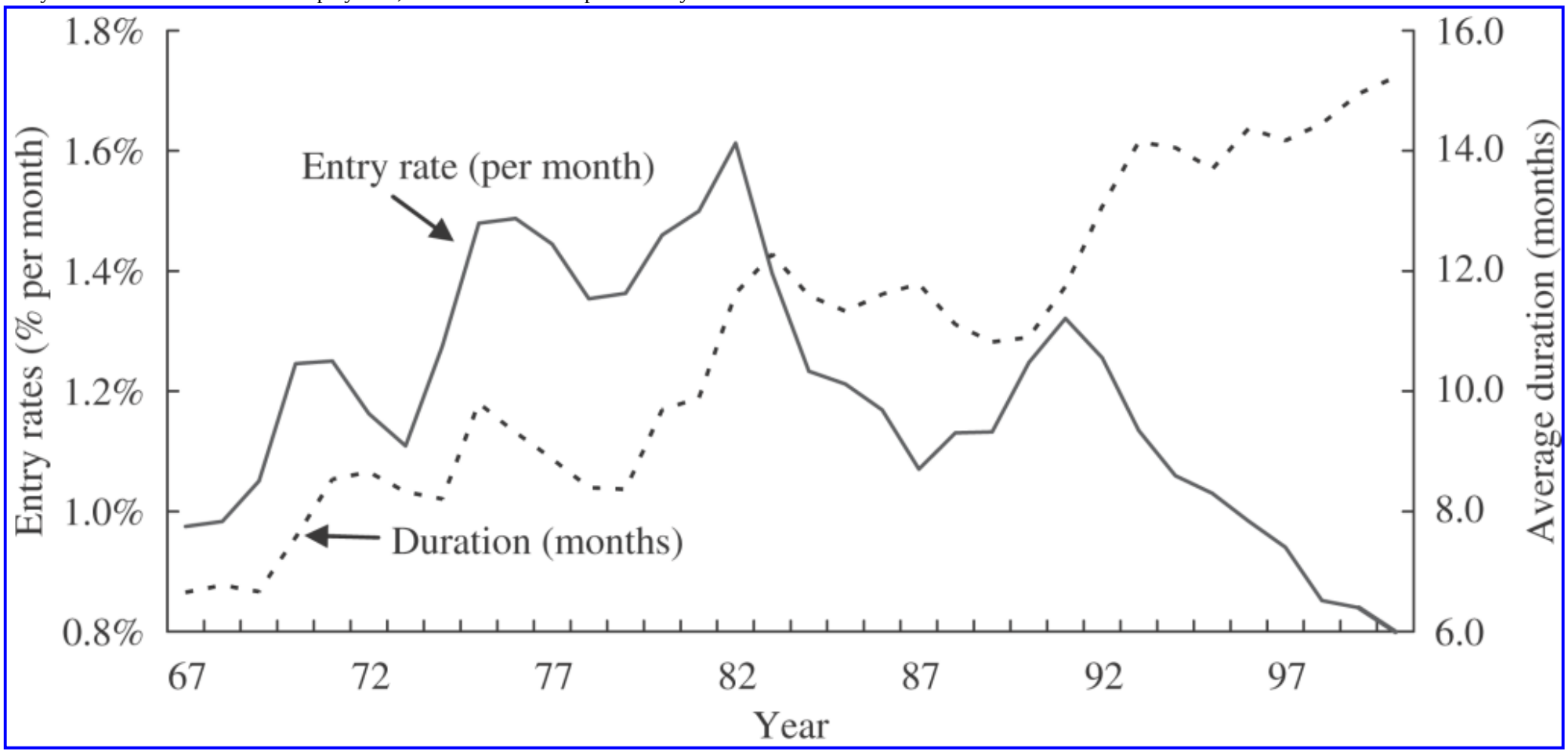

Who are the jobless?

To get a handle on why this occurred, it is worth examining the characteristics of those without jobs. The most basic fact is that unemployment and overall joblessness are much more common among the least skilled. Measuring skill by years of completed schooling, unemployment rates are higher among those with fewer years of schooling, and they also increase more during recessions. JMT (2002) examine a broader definition of skill, based on an individual's position in the overall wage distribution. (JMT impute wages for year-round non-workers from the wage distribution of those who work very few weeks. See JMT, 2002, for a description of their methods.)

Figures 5A and 5B show percentages of the population who are unemployed and out of the labour force (OLF) by percentage intervals of the wage distribution between 1967 and 2000. As with education, both unemployment and non-participation are more common among the less skilled, and in each recession their unemployment rates rise most sharply. Up through the early 1990s, both components of joblessness showed a secular increase that was concentrated among low-wage individuals. For unemployment this trend was reversed in the 1990s but non-participation continued to rise, especially among the least skilled. By 2000 about 20 per cent of men whose skills would put them in the bottom decile of the wage distribution were out of the labour force, which is more than double the fraction of non-participants in the 1960s. Adding the unemployed, by the end of the 1990s nearly 30 per cent of these men were jobless. By comparison, men whose skills put them above the 60th percentile of the wage distribution showed virtually no long-term increase in either unemployment or non-participation. In other words, to understand rising joblessness in the United States, we must focus on changes in economic fundamentals that have affected mainly the lower end of the skill distribution.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Figure 5A
Unemployment rates by wage percentile groups, 1967–2000. *Source*: produced by author.

Figure 5A Unemployment rates by wage percentile groups, 1967 – 2000. Source: produced by author.

Figure 5B
Out of labour force rates by wage percentile groups, 1967–2000. *Source*: produced by author.

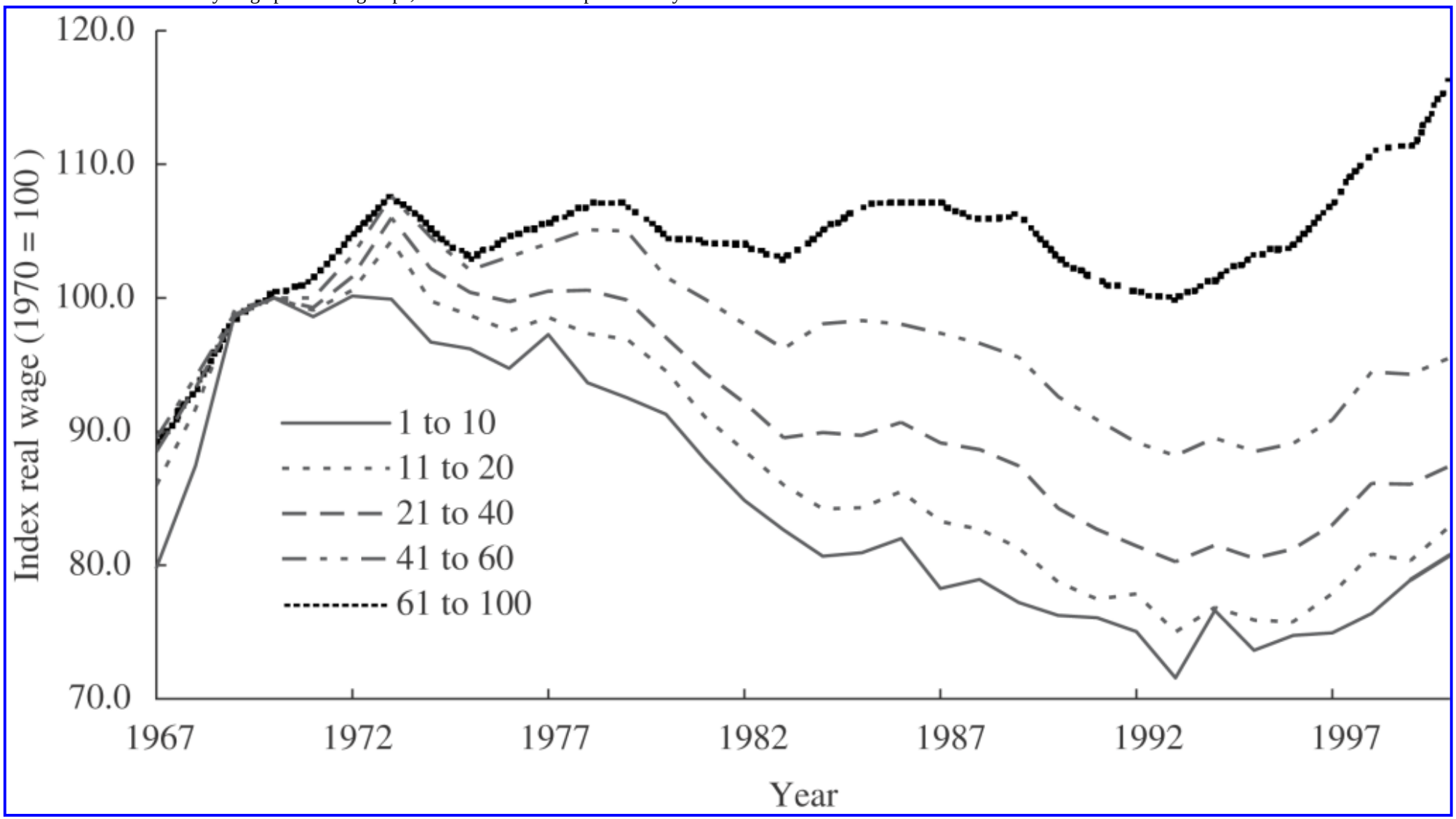

The most obvious explanation is that shifts in relative labour demands have reduced labour market opportunities available to the least skilled, so they end up working less. Evidence consistent with this is the well-documented decline in relative wages among low-skilled workers, shown in Figure 6. Between 1970 and 1993, real wages of men in the first decile of the wage distribution declined by over 25 per cent, with smaller though still important declines for all workers below the median of the wage distribution. The post-1993 growth in real wages, which affected the entire skill distribution, corresponds to a convergence of relative fractions of time spent unemployed (Figure 5A), but non-participation remained quite

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

high among low-wage workers. The continue rise in non-participation while real wage grew in the 1990s suggests a shift in labour supply. Autor and Duggan (2003; 2006) point to Disability Insurance (DI), which became relatively more attractive to low-skill workers who faced a long-term deterioration of labour market opportunities. They document that participation continued to fall because DI subsidized non-work, which was most attractive to low-skilled workers. In earlier times these individuals would have spent transitory periods of unemployment or non-participation, but they would have remained attached to the labour force over the long run. By 2000, many less skilled individuals – faced with declining working opportunities – had simply withdrawn.

Figure 6

Indexed real wages by percentile group, 1967–2000. *Source*: produced by author.

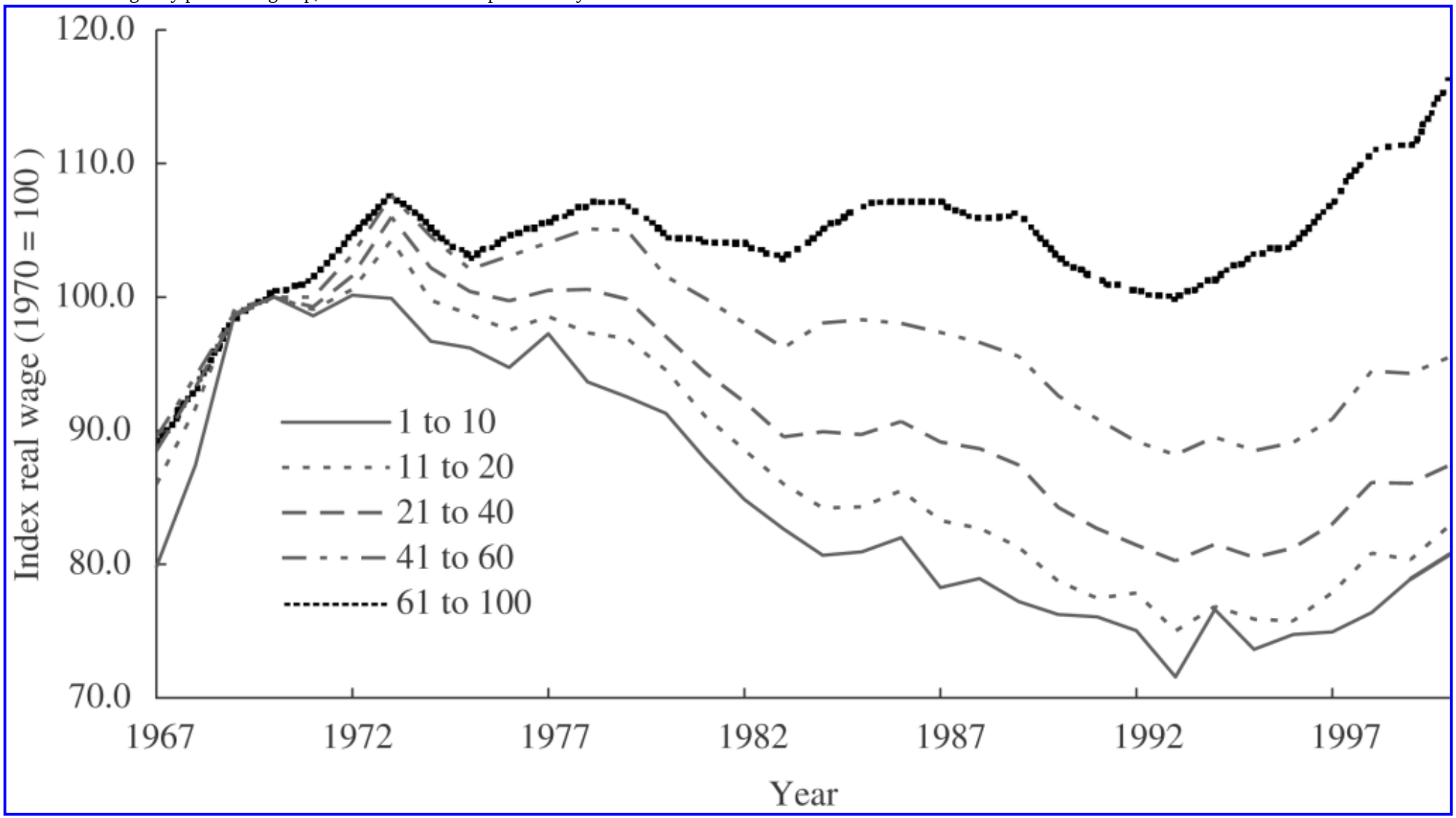

Conclusion: US and European unemployment revisited

The broad message of the above evidence is that the ‘natural rate’ of unemployment (or non-employment) is not a constant towards which the labour market converges; rather it varies with labour market fundamentals – as originally suggested by Phelps (1974). But why have unemployment rates evolved so differently in the United States and Europe? A

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

compelling interpretation is that differences in labour market institutions generate differences in measured unemployment despite similarities in labour market fundamentals. In the United States, the maximum duration of unemployment insurance (UI) is typically six months, and the fraction of earnings replaced by UI is about half. Aside from DI, income support for the long-term jobless is comparatively low. These features may accelerate job-finding among those with marketable skills but weaken labour force attachment among the least skilled, whose opportunities have deteriorated. Then long jobless durations among the least skilled show up in labour force withdrawal. In Europe, UI coverage is much more liberal in terms of both the level and duration of benefits, so those who would leave the labour force in the United States are counted as long-term unemployed. The forces at work and the employment prospects of affected workers are not as different as standard unemployment statistics may suggest.

See Also

- labour market search
- labour supply
- search models of unemployment
- unemployment and hours of work, cross country differences
- unemployment insurance
- unemployment measurement

Bibliography

Alchian, A. 1969. Information costs, pricing, and resource utilization. *Western Economic Journal* 7, 109–28.

Autor, D.H. and Duggan, Mark G. 2003. The rise in the disability rolls and the decline in unemployment. *Quarterly Journal of Economics* 118, 157–205.

Autor, H. and Duggan, Mark G. 2006. The growth in the Social Security disability rolls: a fiscal crisis unfolding. *Journal of Economic Perspectives* 20(3), 71–96.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58, 1–17.

Gronau, R. 1971. Information and frictional unemployment. *American Economic Review* 61, 290–301.

Juhn, C., Murphy, K.M. and Topel, R.H. 1991. Why has the natural rate of unemployment increased over time? *Brookings Papers on Economic Activity* 1991(1), 75–126.

Juhn, C., Murphy, K.M. and Topel, R.H. 2002. Current unemployment, historically contemplated. *Brookings Papers on Economic Activity* 2002(1), 79–136.

Layard, R., Nickell, S. and Jackman, R. 1991. *Unemployment: Macroeconomic Performance and the Labour Market*. Oxford and New York: Oxford University Press.

Ljungqvist, L. and Sargent, T.J. 1998. The European unemployment dilemma. *Journal of Political Economy* 106, 514–50.

Lucas, R. and Prescott, E. 1974. Equilibrium search and unemployment. *Journal of Economic Theory* 4, 103–24.

McCall, J. 1970. Economics of information and job search. *Quarterly Journal of Economics* 84, 113–26.

Mortensen, D. 1970. A theory of wage and employment dynamics. In *Microeconomic Foundations of Employment and Inflation Theory*, ed. E.S. Phelps et al. New York: W.W. Norton.

Murphy, K.M. and Topel, R.H. 1987. The evolution of unemployment in the United States. In *NBER Macroeconomics Annual 1987*, ed. S. Fischer. Cambridge, MA: MIT Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Murphy, K.M. and Topel, R.H. 1997. Unemployment and nonemployment. *American Economic Review* 187, 295–300.

Phelps, E.S. 1974. Economic policy and unemployment in the 1960s. *Public Interest* 34(Winter), 30–46.

Rogerson, R., Shimer, R. and Wright, R. 2005. Search-theoretic models of the labor market: a survey. *Journal of Economic Literature* 43, 959–88.

Topel, R.H. 1993. What have we learned from empirical studies of unemployment and turnover? *American Economic Review* 83, 110–5.

How to cite this article

Topel, Robert. "unemployment." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 14 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000016> doi:10.1057/9780230226203.1758

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

uneven development

Donald J. Harris
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A striking characteristic of capitalist development is the phenomenon of uneven development, defined as persistent differences in levels and rates of economic development between different sectors of the economy. However, much existing economic theory predicts that many observed features of differentiation would tend to wash out as a result of competitive market forces. This article seeks to bridge this gap. It proposes a strategy for the analysis of uneven development that advances toward a historically and empirically relevant theory. The analysis draws in part on elements of the emerging paradigm of neo-Schumpeterian evolutionary theory and on some documented empirical regularities.

Keywords

aggregation (theory); bounded rationality; competition; concentration; creative accumulation; creative destruction; differentiation among firms; diffusion of technology; endogenous growth; evolutionary economics; firm, theory of; first-mover advantages; growth centres; Harrod–Domar growth model; industry evolution; innovation; invention; irreversible investment; learning; life-cycle of industry; market failure; saturation effect; Schumpeterian competition; shift effect; technical change; underdevelopment; uneven development

Article

In examining the general character of the process of capitalist development as it has appeared historically across many different countries over a long period of time, it emerges that one of its most striking characteristics is the phenomenon of uneven development. Specifically, the process is marked by persistent differences in levels and rates of economic development between different sectors of the economy.

This differentiation appears at many levels and in terms of a multiplicity of quantitative and qualitative indices (Kuznets, 1966; Maddison, 1982; Mueller, 1990; Pritchett, 1997; Salter, 1966). Relevant measures that sharply identify the phenomenon include the level of labour productivity in different sectors, the level of wages, occupational and skill composition of the labour force, the degree of mechanization and vintage of production techniques, rates of profit, rates of growth, and the size structure of firms. The phenomenon appears regardless of the level of aggregation or disaggregation of the economy, except for the extreme case of complete aggregation – in which case, structural properties of the economy are made to disappear. For example, it appears at the level of comparing the broad aggregates of manufacturing industry and agriculture, at the level of individual industries within the manufacturing sector, and at the level of individual firms in an industry. It appears on a regional level within national economies as well as on a global scale between different national economies. In this latter context, one form taken is the continued differentiation between underdeveloped and advanced economies, usually identified as the problem of underdevelopment.

These disparities appear from observation of the economy as a whole at any given moment and over long periods of time. While the relative position of particular sectors may change from one period to another, there is, nevertheless, always a definite pattern of such differentiation. We may say, therefore, and certainly it is an implication of these observations, that these disparities are continually reproduced by the process of development. Uneven development, in this sense, is an intrinsic or inherent property of the economic process. Far from being merely transitory, it appears to be a pervasive and permanent condition.

Now, it is an equally striking fact that, when we examine the theoretical literature on economic growth, we find the completely opposite picture. In particular, the dominant conception of the growth process that has characterized the post-Second World War literature is constructed in terms of uniform rates of expansion in output, productivity and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
employment in all sectors of the economy. In this sense, it is largely a literature of steady-state growth, whether presented in multi-sectoral or aggregative models (Burmeister and Dobell, 1970; Harris, 1978). Some notable and relevant exceptions, including Haavelmo (1954), Leon (1967), Nelson and Winter (1982), Pasinetti (1981), Salter (1965), explicitly examine aspects of the problem of persistent differentiation posed here. The recent flurry of work in endogenous growth theory seeks to incorporate some relevant elements of the problem into the neoclassical conception of the growth process (Aghion and Howitt, 1998). However, much of existing economic theory predicts that, given enough time, many of the features of differentiation which we observe empirically would tend to wash out as a result of the operation of competitive market forces (Harris, 1988). Such differentiation should therefore be viewed only as a transitory feature of the economic process.

Thus, on the one side, we find a historical picture of uneven development as a persistent phenomenon, and on the other, a theory that essentially negates and denies this fact. It is possible to go some of the way towards bridging this gap. Accordingly, I consider here a strategy for analysis of uneven development that breaks through the narrow limits of the existing steady-state theory and advances towards a historically and empirically relevant theory.

The analytics of uneven development

It is necessary to start by recognizing the intrinsic character of the individual firm as an expansionary unit of capital with a complex organization. Various efforts have been made to develop a theory of the firm on this basis. (See, for instance, Penrose, 1959; Baumol, 1967; Marris, 1967; Winter, 2006). In this conception, growth is the strategic objective on the part of the firm. This urge to expand is not a matter of choice. Rather, it is a necessity enforced upon the firm by its market position and by its existence within a world of firms where each must grow in order to survive. It is reinforced also by sociological factors. It is this character of the firm that constitutes the driving force behind the process of expansion of the economy.

In the aggregate, the global economy is conceived to consist of an ordered system of firms (an interlocking network of individual circuits of capital) and its sectors (classified variously as industries, regions, national economies) likewise to be clusters of the firms that are the component units of this system. In this system, it is the firms that compete, not industries or regions, national economies or ‘North’ versus ‘South’. The state sets the rules and jointly determines the external conditions (externalities) within which the firms operate.

This is a crucial starting point because it establishes the idea of growth as the outcome of a process driven by active agents, not by exogenous factors. In particular, in the context of the capitalist economy, growth is the outcome of the self-directed and self-organizing activity of firms, each of which seeks to expand and improve its competitive position in relation to the rest. Once this principle is recognized it becomes possible to move towards an understanding of the problem of uneven development.

The imperative of growth impels the firm constantly to seek new investment opportunities wherever they are to be found. Such opportunities may lie within a wide range – in existing product lines, in new products and processes, in new geographical spaces and natural resource frontiers, or in the take-over of existing firms. However, at the core of this movement, viewed historically over the long term, are the invention, innovation and diffusion of new technologies that give rise to new products and services (Freeman, 1982; Landes, 1969; 1999; Marx, 1906, ch. 15; Mokyr, 1990; 2002).

The emergence of growth centres or leading sectors is a reflection of this underlying process. It is a consequence of the effort on the part of many firms to create or to rush into those spheres where a margin of profitability exists that allows them to capture new profit and growth opportunities. It may be conceived to take the form of a ‘swarm’ (Schumpeter, 1934, p. 223) or ‘contagion’ (Baumol, 1967, p. 101), marked by both entry and exit of firms. Such spheres are opened up, typically through complementary ‘macroinventions’ and ‘microinventions’ (Mokyr, 1990, p. 13) and in a sporadic and discontinuous pattern, as a consequence of the ongoing investment and innovative activity of firms and the competitive interactions among them. It is this constant flux, consisting of the emergence of new growth centres, their rapid expansion relative to existing sectors, and the relative decline of others, that shows up in the economy as a whole as uneven development.

The process of industry evolution

The form of this process, as it appears at the level of particular industries and products, has been identified in terms of certain empirical regularities, though there are also significant variations across industries and products. Studies show that, with some exceptions, the growth of many new industries and products follows a life cycle pattern (Gold, 1964; Gort and Klepper, 1982; Klepper, 1997; Klepper and Graddy, 1990; Mullor-Sebastian, 1983; Wells, 1972). It may be represented schematically by an S-shaped curve of the time-path of output as in Figure 1. (For simplicity, no distinction is made here between products and processes, an industry is assumed to produce a single product, and short-term turbulence in the path of output is ignored.) Accordingly, we may distinguish three phases of expansion: I, the initial phase, where total output is a minute share of aggregate output and grows at a low rate; II, a phase of rapid growth in which output expands rapidly and its share of aggregate output grows; III, the sector reaches a threshold beyond which its growth rate tends to level off and perhaps to decline.

Figure 1

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Life cycle of an industry

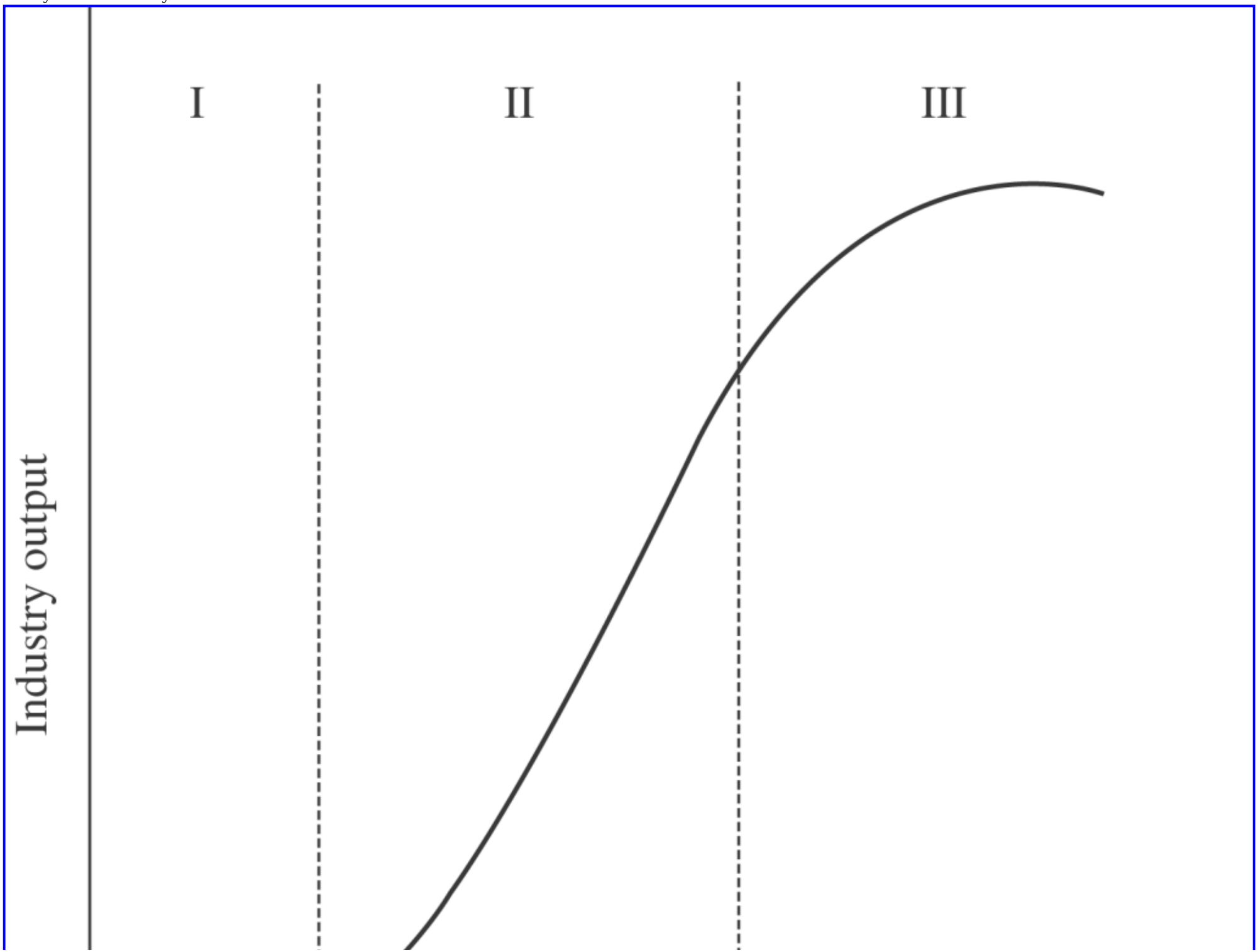

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

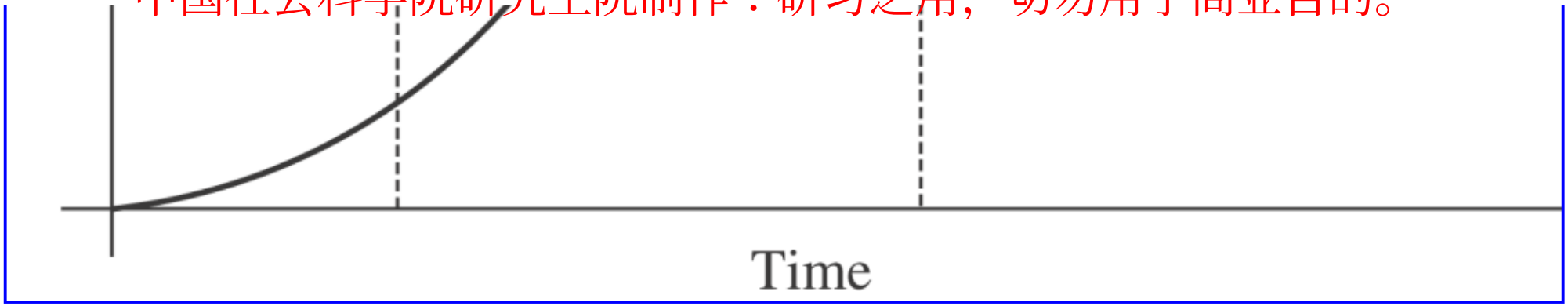

To characterize the associated pattern of technological innovation, Kuznets (1979) identifies a sequence of four distinct phases constituting the life cycle of ‘major’ innovations: (1) the *pre-conception* phase in which necessary scientific and technological preconditions are laid; (2) the phase of *initial application* involving the first successful commercial application of the innovation; (3) the *diffusion* phase marked by spread in adoption and use of the innovation along with continued improvements in quality and cost; (4) the phase of *slowdown* and *obsolescence* in which further potential of the innovation is more or less exhausted and some contraction may occur. This taxonomy is not all-embracing, and there are others that emphasize other features, but it is suggestive in pointing to a certain internal logic of the innovation process.

The process of industry evolution is also typically associated with a changing firm-structure of the industry. In many industries, there is a proliferation of small firms in phase I. As the diffusion of the product occurs and growth speeds up, there is a ‘shaking out’ process by which many of the smaller firms disappear (exit) and the available market is concentrated in the remaining firms. When the industry reaches ‘maturity’, in phase III, there is likely to be a high degree of concentration. This association between industry life cycle and changing firm-structure (commonly called ‘co-evolution’) suggests that the dynamic of expansion through innovation is simultaneously a process of the concentration of capital.

This sequence of a single product-cycle, schematically described here, is but a small segment of the time sequence characterizing the historical evolution of the economy. Given that firms are growing, making profits, and seeking to continue to grow, it must be supposed that at least some of them, having entered into phase III, would seek to launch into new investment opportunities. They will therefore actively seek new products that will initiate a corresponding new sequence. Alternatively, the new sequence could come from entry of new start-up firms.

It follows that we can map out the dynamic evolution of the economy as *a sequential process* that is discontinuous, punctuated and stochastic, with varying and overlapping time-scales of the different product-cycles, where the overall growth is accountable for on the basis of (1) the individual growth of particular new products coming on stream, (2) the growth of pre-existing products, each of which is growing at a different rate depending on the particular phase reached in its life cycle, and (3) over time the irregular accretion of new products as the innovation process continues.

In this context, the relative position of any firm-cluster (region or national economy) at any time on a relevant index of development may be seen as a matter of the particular products it has managed to capture as a result of the previous pattern of accumulation, the ongoing activity of firms operating within it and the particular timing of their entry into the life cycle of new products.

The causes that produce and sustain the observed patterns of differentiation must then be found within the internal dynamics of this process, leaving aside such historically contingent factors as wars, colonial control, ‘foreign’ intervention, that may also be considered relevant and important.

What role is to be assigned to demand as a factor in this process? At the level of individual consumer products or industries, a common conception is that demand acts as an autonomous factor with a definite influence on the life-cycle pattern of evolution of the product. That influence is exerted in the early phase of introduction of a new product because of an element of resistance due to ‘habit’ formed in a customary pattern of consumption. It is exerted also in the maturity phase because of the operation of ‘saturation effects’ in consumption. But there are reasons to doubt the strength and effectiveness of such factors, as well as their supposed autonomy.

First, in an economy undergoing regular and rapid change, it is not evident what role there is for habit except for the habit of change itself. The experience of, and adaptation to, change may create a high degree of receptivity to change. What then becomes decisive in the evolution of demand (for consumer goods) is the growth of income, and the changing relative price and quality of products. Income and price elasticities of demand are an imperfect, proximate expression of this dynamic effect.

Second, in so far as these latter factors are crucial to the formation of demand, it may be argued that there is a certain self-fulfilling aspect of the expansionary process at the level of industry demand. In particular, investment generates the demand that provides the market for the new products which investment itself creates. This occurs in two ways. First structural interdependence in the economy at the level of both production and expenditure patterns allows for the possibility of a certain mutual provisioning of markets when expansion takes place on a broad front. Second, as a new product unfolds through the stages of the innovation process, it undergoes both improvements in quality and a decline in price relative to other products. This development provides a substantive basis for making inroads into the market for existing closely related products and hence promotes demand through a shift from ‘old’ to ‘new’ products. It is perhaps this shift effect which is mistakenly identified as a saturation effect by adopting a one-sided and static view of a dynamic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and interdependent process.
Each and every individual firm must of course secure a place in the market for its product. Its success in this regard is dependent on its own efforts and capabilities.

Competition, firm capabilities, entry/exit conditions, and the social environment

Analytical treatment (including formal modelling) of the process of industry evolution has flourished since the 1980s in tandem with an outpouring of empirical studies covering different industries, countries and time periods. Much of this work is done within the frame of an emerging paradigm in the Schumpeterian tradition of evolutionary dynamics (Futia, 1980; Iwai, 1984a, 1984b; Nelson and Winter, 1982; Dosi, 1984) and there are other theoretical approaches (Loury, 1979; Dasgupta and Stiglitz, 1980; Durlauf, 1993). For a review of the current state of the art and challenges for research, focusing on the evolutionary approach, see Malerba (2006). Relevant for the present purposes are the significant insights provided so far by this work into the mechanisms and causal factors that govern the process of industry evolution and account for the persistence of differentiation among firms.
The neo-Schumpeterian approach develops an explicit formulation of ‘Schumpeterian competition’ in which firms innovate to win super-normal profits, profits are reinvested to provide further growth through innovation and market expansion, and there are winners and losers due to the operation of selection mechanisms and learning mechanisms. Decisions are typically based on bounded rationality. It is shown that such competitive behaviour under specified conditions gives rise to persistent differentiation among firms in terms of size, productivity, costs of production, product characteristics, profitability and growth and may breed long-term sustainable market concentration among surviving firms, with or without entry. Economies of scale and scope are not a necessary part of the story; a key factor is increasing returns to knowledge and learning. Though there exists a strong tendency to concentration, it is not inevitable, and depends on industry characteristics that vary across industries. There also exist dual tendencies of ‘creative destruction’ and ‘creative accumulation’.
A distinctive feature of this approach is the conception of the firm itself as an organizational unit. The firm is conceived as the embodiment of a set of strategic assets (competences or capabilities), tangible and intangible, consisting of knowledge, skills, and routines, gained through path-dependent experience and learning, that are specific to each firm and non-tradeable. These assets evolve over time (through ‘competence accumulation’) with the ongoing process of evolution of the industry and through interaction with the changing environment. Consequently, diversity among firms is not only a characteristic of the system of firms, it is also reproduced by the evolutionary dynamics of the competitive process.
Some key factors determining the evolutionary path of industry structure in terms of firm composition are the following. (1) First- (second-, third-) mover advantages arising from a combination of unique internal attributes of the mover, product characteristics, network effects among users, and random chance events. (2) Non-pecuniary network externalities associated with cues and information gained from interacting with the ‘local’ social environment of firms, users of the product, and institutions involved in knowledge creation and information dissemination (on the national level, the ‘national system of innovation’). (3) Spillover effects among firms and across industries, which may be both positive and negative. (4) Increasing returns to knowledge and learning within the firm. (5) A firm may become ‘locked-in’ to its own trajectory of technology development and reap increasing returns therefrom, but eventually suffer a disadvantage from generating irreversibilities and inertia causing inability to adjust to change (‘success breeds failure’). (6) The very same factors that confer advantages upon early entrants and incumbent firms may create barriers to entry for ‘latecomers’, depending on the stage of industry evolution and timing of entry.
Some relatively neglected factors that need to be integrated into a more comprehensive analysis include: (1) the role of market demand, as related to the mutual interaction between producers and users (consumers, other firms, and the state); (2) the role of the financial system (Schumpeter had assigned a crucial role to the granting of credit ‘as an order on the economic system to accommodate itself to the purposes of the entrepreneur’ (1934, p. 107)); (3) workplace and labour market interactions, lightly touched upon by Mansfield (1968, ch. 5) and vividly described in historical detail by Braverman (1974); (4) the system of governance by the state, that sets and enforces the rules and norms, including property rights, governing conduct by firms.
Within this extended framework of analysis, it is possible to explain not only how some firms (or firm-clusters) come to capture the position of leaders (and may eventually lose it to others), but equally how some are left behind, others drop out altogether (exit), and still others remain on the ‘periphery’ (so to speak) lacking the internal and external capabilities to enter. In this regard, the explanatory power of this analysis is readily applicable to commonly discussed empirical and historical phenomena such as ‘deindustrialization’, ‘catching up’ (convergence), and ‘falling behind’ (divergence).
What emerges from this analysis also is an understanding of the critical role of public policy and programmes to foster economic development. Because of the pervasiveness of externalities and various forms of coordination problems, market failures are intrinsic to the process, calling thereby for collective intervention to achieve efficiency and socially optimal results.

The aggregation problem

All the preceding analysis concerns the pattern of industrial growth viewed at the level of an individual industry and the firms (or firm-clusters interpreted as, say, regions or countries) that compose it. There is nothing in this analysis to indicate how the pattern of growth of different industries translates into aggregate expansion at the level of the economy as a whole, or how the various industrial patterns fit together to form a complete whole. This is a substantive problem requiring further analytical treatment on its own terms. Its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
significance derives from recognition that the economy as a whole is not just the sum of its parts. Hence, the motion of the economy cannot simply be deduced from the movement of its parts. The usual methodological device of the ‘representative firm’ necessarily fails in the present context.

A related aspect of the problem is associated with the manifold and complex ways in which growth in one sector (however defined) mutually conditions and is conditioned by growth in other sectors. Such mutual interaction is a necessary consequence of economic interdependence in both production and exchange. (Hence, models of international trade that claim to show uneven development arising uniquely from exchange of products give a one-sided representation of the problem.) The existence of such interaction implies that there is a certain cumulative effect intrinsic in the growth process. Understanding the exact mechanisms through which this effect operates is one of the central analytical problems for the analysis of uneven development.

There is no guarantee that in the aggregate there is always sufficient demand for all products. It is here that the analysis comes full circle, back to the problem of overall effective demand that motivated the early post-war growth theory initiated by Harrod (1948) and Domar (1957). This problem was a central focus of the analysis in the Keynesian and Post Keynesian tradition, less so in the case of the neoclassical tradition (as detailed in Harris, 1985). It appears now that it cannot be escaped in making the transition to the analysis of uneven development.

The analytical framework presented here lays the groundwork for addressing this larger set of problems.

See Also

- development economics
- economic growth, empirical regularities in
- endogenous growth theory
- Schumpeterian growth and growth policy design

Bibliography

Aghion, P. and Howitt, P. 1998. *Endogenous Growth Theory*. Cambridge, MA: MIT Press.

Baumol, W.J. 1967. *Business Behaviour, Value and Growth*, revised edn. New York: Harcourt, Brace and World.

Braverman, H. 1974. *Labor and Monopoly Capital, the Degradation of Work in the Twentieth Century*. New York: Monthly Review Press.

Burmesier, E. and Dobell, A.R. 1970. *Mathematical Theories of Economic Growth*. London: Macmillan.

Dasgupta, P. and Stiglitz, J. 1980. Industrial structure and the nature of innovative activity. *Economic Journal* 90, 266–93.

Domar, E.D. 1957. *Essays in the Theory of Economic Growth*. New York: Oxford University Press.

Dosi, G. 1984. *Technical Change and Industrial Transformation*. London: Macmillan.

Durlauf, S.N. 1993. Nonergodic economic growth. *Review of Economic Studies* 60, 349–66.

Freeman, C. 1982. *The Economics of Industrial Innovation*, 2nd edn. Cambridge, MA: MIT Press.

Futia, C.A. 1980. Schumpeterian competition. *Quarterly Journal of Economics* 94, 675–95.

Gold, B. 1964. Industry growth patterns: theory and empirical results. *Journal of Industrial Economics* 13, 53–73.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gort, M. and Klepper, S. 1982. Time paths in the diffusion of product innovations. *Economic Journal* 92, 630–53.

Haavelmo, T. 1954. *A Study in the Theory of Economic Evolution*. Amsterdam: North-Holland.

Harris, D.J. 1978. *Capital Accumulation and Income Distribution*. Stanford: Stanford University Press.

Harris, D.J. 1985. The theory of economic growth: from steady states to uneven development. In *Contemporary Issues in Macroeconomics and Distribution*, ed. G. Feiwel. London: Macmillan.

Harris, D.J. 1988. On the classical theory of competition. *Cambridge Journal of Economics* 12, 139–67.

Harrod, R.F. 1948. *Towards a Dynamic Economics*. London: Macmillan.

Iwai, K. 1984a. Schumpeterian dynamics, Part I: an evolutionary model of selection and imitation. *Journal of Economic Behavior and Organization* 5, 159–90.

Iwai, K. 1984b. Schumpeterian dynamics, Part II: technological progress, firm growth and ‘economic selection’. *Journal of Economic Behavior and Organization* 5, 321–51.

Klepper, S. 1997. Industry life cycles. *Industrial and Corporate Change* 6, 145–81.

Klepper, S. and Graddy, E. 1990. The evolution of new industries and the determinants of market structure. *RAND Journal of Economics* 21, 27–44.

Kuznets, S. 1966. *Modern Economic Growth, Rate, Structure, and Spread*. New Haven: Yale University Press.

Kuznets, S. 1979. Technological innovations and economic growth. In *Growth, Population, and Income Distribution, Selected Essays*. New York: Norton.

Landes, D.S. 1969. *The Unbound Prometheus*. Cambridge: Cambridge University Press.

Landes, D.S. 1999. *The Wealth and Poverty of Nations*. New York: Norton.

Leon, P. 1967. *Structural Change and Growth in Capitalism*. Baltimore: Johns Hopkins Press.

Loury, G.C. 1979. Market structure and innovation. *Quarterly Journal of Economics* 93, 395–410.

Maddison, A. 1982. *Phases of Capitalist Development*. Oxford: Oxford University Press.

Malerba, F. 2006. Innovation and the evolution of industries. *Journal of Evolutionary Economics* 16, 3–23.

Mansfield, E. 1968. *The Economics of Technological Change*. New York: Norton.

Marris, R. 1967. *The Economic Theory of ‘Managerial’ Capitalism*. London: Macmillan.

Marx, K. 1906. *Capital*, vol. I. Chicago: Charles H. Kerr.

Mokyr, J. 1990. *The Lever of Riches, Technological Creativity and Economic Progress*. Oxford: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mokyr, J. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton: Princeton University Press.

Mueller, D.C. 1990. *The Dynamics of Company Profits: An International Comparison*. Cambridge: Cambridge University Press.

Mullor-Sebastian, A. 1983. The product life cycle theory: empirical evidence. *Journal of International Business Studies* 14, 95–105.

Nelson, R.R. and Winter, S.G. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, MA: Harvard University Press.

Pasinetti, L.L. 1981. *Structural Change and Economic Growth*. Cambridge: Cambridge University Press.

Penrose, E.T. 1959. *The Theory of the Growth of the Firm*. Oxford: Blackwell.

Pritchett, L. 1997. Divergence, big time. *Journal of Economic Perspectives* 11(3), 3–17.

Salter, W.E.G. 1965. Productivity growth and accumulation as historical processes. In *Problems in Economic Development*, ed. E.A.G. Robinson. London: Macmillan.

Salter, W.E.G. 1966. *Productivity and Technical Change*. Cambridge: Cambridge University Press.

Schumpeter, J.A. 1934. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Wells, L.T., Jr. 1972. *The Product Life Cycle and International Trade*. Boston: Harvard University Press.

Winter, S.G. 2006. Toward a neo-Schumpeterian theory of the firm. *Industrial and Corporate Change* 15, 125–41.

How to cite this article

Harris, Donald J. "uneven development." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000019> doi:10.1057/9780230226203.1762

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

unforeseen contingencies

Barton L. Lipman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

While unforeseen contingencies – possible events that agents do not think of when planning or contracting – are often said to greatly affect the nature of contracting, we lack useful formal models. Most of the existing models boil down to assuming that agents give zero probability to some events that might actually occur, an approach which is not particularly useful for studying the effects of unforeseen contingencies on contracting.

Keywords

control rights; expected utility; incomplete contracts; long-term and short-term contracts; probability; rationality, bounded; short-term contracts; uncertainty; unforeseen contingencies

Article

Many writers have suggested that the nature of contracting, firm structure, and even political constitutions cannot be well understood without taking account of the role of *unforeseen contingencies*. As I explain in more detail below, many definitions are possible, but I will define unforeseen contingencies to be possibilities that the agent does not 'think about' or recognize as possibilities at the time he makes a decision. In virtually any reasonably complex situation, real people do not consider all of the many possible situations that may arise. Because of this, contracts, for example, typically assign broad categories of rights and obligations rather than calling for very specific actions as a function of what might occur. Similarly, firms are designed to figure out what to do rather than simply being programmed to implement some given set of actions. Finally, laws, especially sweeping ones such as constitutions, are intentionally left vague to allow adaptation to circumstances as they arise.

Unfortunately, while it is easy to find eloquent statements in the economics literature regarding the importance of unforeseen contingencies for understanding the nature of economic and political institutions – see, for example, Hayek (1960); Williamson (1975); or Hart (1995) – there is no agreed formal model. I sketch a few known approaches below, but none of them provides a model that can be used to study these

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
issues.

To make this point concretely, I focus on a particular example of an aspect of contracting that we would want a model of unforeseen contingencies to help us understand, namely, the choice between long-term and short-term contracts. It seems obvious that one of the main advantages of a series of short-term contracts is that it is easier to anticipate the relevant contingencies for the near future than for the distant future. Hence in environments with many unforeseen contingencies relative to the value of long-term contracting, we should expect to see more short-term contracting. I will argue below that none of the models of unforeseen contingencies in the literature can be used to illustrate this simple idea.

Before discussing the approaches taken, it is important to clarify what I mean by unforeseen contingencies. First, as I use the term, unforeseen contingencies are *not* events that the agent has considered but assigned zero probability. This notion is something standard models deal with perfectly well. More importantly, the existence of such events seems to have little to do with the features of economic and political institutions we believe to be related to unforeseen contingencies. To be concrete, consider the trade-off discussed above between long-term contracts and short-term contracts. If the only sense in which some contingencies in the distant future are not foreseen is that they are given zero probability, then the agents will perceive zero costs to excluding them. Hence they will see no ‘foreseeability’ advantage to short-term contracts, so a model of unforeseen contingencies based on such a definition cannot say anything interesting about the trade-off.

It is also important to note that the use of the term ‘unforeseen’ in law is often closer to the zero probability definition than the definition I use here. In particular, legal usage often seems to suggest that a contingency is ‘unforeseen’ by an agent if it occurred even though the agent gave it ‘low’ probability *ex ante*. For example, a 1997 US tax law allows a person who sells his home to exclude some of the capital gains from taxation under certain conditions if ‘unforeseen circumstances’ precipitated the move. The Internal Revenue Service (2006, p. 16) recently issued regulations listing events that would ‘count’ as such unforeseen circumstances, including divorce, job loss, or multiple births from a single pregnancy. Surely, most homeowners would not be startled to learn that couples sometimes divorce, that job losses can occur, or that a pregnancy could yield triplets, so these circumstances are not ‘unforeseen’ in the sense used here. Instead, such events are ‘unforeseen’ in the sense that they were very unlikely *ex ante*, too unlikely to influence the home purchase decision.

While this use of ‘unforeseen’ is evidently valuable for some purposes, it does not seem to be appropriate to the issues of interest here, though it is closer than the zero probability definition. To see this, consider again the trade-off between long-term and short-term contracts. If ‘unforeseen’ contingencies are recognized but given low probability, they can still be incorporated into the contract and, in the absence of costs to doing so, will be. Hence, in the absence of contracting costs, again, short-term contracts will have no foreseeability advantages if this is what we mean by foreseeability. On the other hand, if there are costs of writing ‘long’ contracts, it may be optimal to exclude contingencies with low probability. Hence a sequence of short-term contracts (which delays some of the writing costs) may be better. On the other hand, it is not clear that the advantage of short-term contracts is a gain in foreseeability so much as it is a delay in writing. See Al Najjar, Anderlini and Felli (2006) for a particularly interesting related model.

Turning to models, the idea of how unforeseen contingencies are represented is common to most of the models in the area, though with many variations. (A different approach, which I do not discuss, involves an explicit logic rather than focusing on a state space. See Halpern and Rêgo, 2005 for a good example of this approach and an overview of much of this literature.) In standard models of uncertainty without unforeseen contingencies, there is a set of states of the world, say Ω, which represents the uncertainty. A

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

state $\omega \in \Omega$ should be thought of as a specification of every possible circumstance conceivably relevant to the agent's situation. For example, for a firm, a state might specify input prices, demand conditions, technological possibilities, what is going on with its rivals, and so on. Part of what is meant by the phrase ‘every relevant circumstance’ is that, if we know the state, then we know the exact consequence (profits or utility) the agent receives as a function of whatever course of action he might choose.

To give a concrete example, suppose there are two relevant sources of uncertainty: whether it rains and whether there is a revolution in country X. This gives us four possible states of the world:

$$\Omega = \{(\text{rain, revolution}),\ (\text{rain, no revolution}),\ (\text{no rain, revolution}),\ (\text{no rain, no revolution})\}.$$

Consider an agent who has never considered the possibility of revolution. This agent sees only two possibilities: rain or no rain. That is, the agent has a *subjective state space S*, describing the possibilities as he perceives them, given by

$$S = \{(\text{rain}),\ (\text{no rain})\}.$$

We can think of the ‘state’ (rain) as the event {(rain, revolution), (rain, no revolution)} and think of the ‘state’ (no rain) as analogous. Thus the state space as seen by the agent is actually a partition of the true state space. The variation within an event of this partition is variation that the agent has simply not thought of.

This basic idea appears in numerous forms in the literature. This partition description appears in Ghirardato (2001) and Dekel, Lipman and Rustichini (2001), among others. A more complex form appears in Li (2006a) and in Heifetz, Meier and Schipper (2006a), both of which allow the possibilities recognized by the agent to vary with the true state of the world. For example, it might be that when the true state is (rain, revolution), the agent recognizes the possibility of the revolution, while if it is (rain, no revolution), he does not.

While this idea for representing knowledge is almost uniformly used in the literature, there is greater variation in the way decision-making is represented. Continuing with the rain and revolution example, suppose the agent has a certain amount of money and can either use it to buy an umbrella or invest it in country X or simply save it. Suppose Table 1 gives the true, objective payoff of the agent as a function of his choice and the real state.

Objective payoffs

Objective state	Payoff if buy	Payoff if save	Payoff if invest
(rain, revolution)	5	0	–100
(rain, no revolution)	5	0	10
(no rain, revolution)	6	8	–100

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(no rain, no revolution)	6	8	10

Turning to the agent's perceptions, continue to assume that he sees the possibilities only as rain versus no rain. Intuitively, the consequences of buying the umbrella or saving money are unambiguous since they only depend on this. Putting it differently, these acts are measurable with respect to the agent's awareness, so there seems to be no problem. On the other hand, how does the agent perceive the payoffs to investment?

A number of papers in the literature use models that treat the payoffs to investing in ‘states’ (rain) and (no rain) as exogenously given (see, for example, Heifetz, Meier and Schipper, 2006b; Li, 2006b). In a somewhat more restrictive version of the same idea, Modica, Rustichini and Tallon (1998) assume that if the agent does not foresee some possibility, this means he implicitly assumes some particular resolution of this uncertainty. For example, perhaps the agent implicitly assumes there will not be a revolution, so he sees the payoff to investing in each ‘state’ as ten. In other words, these approaches come down to treating the agent's view of the available actions as given by Table 2 for some numbers x and x'.

Perceived payoffs, version 1

Subjective state	Perceived payoff if buy	Perceived payoff if save	Perceived payoff if invest
(rain)	5	0	x
(no rain)	6	8	x'

With more than one agent, these models generally do allow the agents to perceive different possibilities and to assign payoffs differently. For example, if two agents have to jointly agree on what to do with their money in the example above, we could assume that one of them perceives only rain versus no rain, while the other perceives only revolution versus no revolution.

While such a model can be useful for some purposes, it does not appear useful for studying the kinds of issues mentioned in my opening paragraph. To see this, note that the default approach of Modica, Rustichini and Tallon (1998) is identical to a model where the states (rain, revolution) and (no rain, revolution) have zero probability. While other values of x and x' are not as directly interpreted, again, the model is identical to one where the agent believes a revolution is impossible (and may have ‘incorrect’ beliefs about his payoffs). As argued above, if ‘unforeseen’ is taken to mean ‘zero probability’, then short-term contracts do not have foreseeability advantages over long-term contracts. Hence, at least for the purposes of studying the trade-off between short-term and long-term contracts, this model of decision-making does not appear to be useful.

Part of what this approach omits is recognition by an agent that his conception of the world is incomplete. Intuitively, what we need to understand the postulated trade-off between short-term and long-term contracts is a recognition by the agent that his conception of what the world will be like in 2016 will be clearer in 2015 than in 2006.

To state this more concretely in the context of the example, the agent might perceive only the possibility of rain versus no rain, but understand that this omits many currently unforeseen possibilities. How might we represent such a situation? One approach is to separate the agent's uncertainty about what events may occur in the world from his uncertainty about what his payoff will be given a particular action. In the story above, we said that the agent's payoff to investing depends on whether it rains and whether there is revolution. Presumably, the agent does not care about rain or revolution per se but instead cares about what utility or payoff he receives. That is, the ‘states’ as the agent perceives them may be more usefully thought of statements about what payoff the agent gets from each action. In the example, then, any vector

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of three numbers (giving the payoffs to the three actions in some order) could be a ‘state’. Table 3 shows one possible representation along this line.

Perceived payoffs, version 2

Subjective state	Perceived payoff if buy	Perceived payoff if save	Perceived payoff if invest
(rain, 1)	5	0	10
(rain, 2)	5	0	–20
(no rain, 1)	6	8	20
(no rain, 2)	6	8	–10

The subjective states (rain, 1) and (rain, 2) represent the agent's uncertainty about the payoffs to investing when the only objective uncertainty he can think of (rain versus no rain) is resolved in favour of rain.

This idea also appears in various forms in the literature. Fishburn (1970) includes an early statement of the idea and more involved treatments appear in Ghirardato (2001), Kreps (1979; 1992), Dekel, Lipman and Rustichini (2001), Halpern and Rêgo (2006) (embodied in their ‘virtual moves’), and Epstein, Marinacci, and Seo (2007), among others. Some of these models (for example, Kreps or Dekel, Lipman, and Rustichini) use expected utility over these ‘states’, while others (for example, Ghirardato or Epstein, Marinacci, and Seo) use models of agents who are uncertainty averse with respect to what the ‘right’ payoff is.

At first glance, this approach appears to be capable of generating a model we could use for the purposes of interest. In fact, this model looks very similar to the ‘observable but unverifiable’ uncertainty story used by Grossman and Hart (1986), Hart and Moore (1988), and Hart (1995) to model incomplete contracts. For brevity, I henceforth refer to this as the *GHM approach*. This approach assumes that some of the variables relevant to the contracting parties are observed by them but cannot be ‘shown’ to a court. As a result, the parties cannot, according to this approach, contract on these variables because a dispute about their realizations cannot be settled by the court. (These papers often also use the assumption that certain actions are indescribable, but this aspect of the GHM approach is not relevant here.) Hence contracts can only allocate control rights – that is, assign the rights to make various decisions *ex post*. Intuitively, if some variables cannot be contracted on, one has to rely on the parties to choose appropriately in the relevant contingencies.

Similarly, it seems natural to assume that these subjective states cannot be contracted over. If a ‘state’ is simply a specification of a utility function for each agent as a function of the actions, how can such a state be verified? Thus the GHM approach appears to fit naturally with this approach to modelling unforeseen contingencies.

To be more concrete, consider again the trade-off between short-term and long-term contracts. It seems natural to assume that the set of subjective states for a period far in the future is larger than the set for a period closer to the present. In this sense, any contract regarding a distant period cannot be as ‘fine tuned’ as a contract for a close period. This will naturally give a trade-off between the value of contracting far in advance and the value of contracting once the parties know more, and so can contract better.

To be still more explicit, suppose the table above gives the subjective state space. Suppose that if the agents write a long-term contract today, it can only specify an outcome as a function of whether it rains or not. On the other hand, assume that they expect that if they wait and write a contract at a later date, they will learn whether the ‘correct’ state space is {(rain, 1), (no rain, 1)} or {(rain, 2), (no rain, 2)}. Thus a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
contract written later will not share risk as well, but can specify the outcomes given rain and no rain more efficiently.

Unfortunately, the work of Maskin and Tirole (1999) calls this conclusion into doubt. They show that observable but unverifiable variables (and indescribable actions) do not justify a departure from standard contract theory. More specifically, even with observable but unverifiable variables, a mechanism can be designed that will induce the parties to reveal to the court what the values of the unverifiable variables are. Thus the fact that they cannot *prove* these facts to the court is not a problem since the court knows they will tell the truth.

Given the similarity to GHM, this suggests that the above model of unforeseen contingencies may not yield results different from standard contract theory either. In terms of the example above, the agents might be able to set up a mechanism that effectively enables them to write a contract specifying different outcomes in the subjective states (rain, 1) and (rain, 2). If so, there cannot be any gain in waiting.

In principle, there are ways one could introduce more realism to the Maskin–Tirole framework and overturn their conclusion. For example, they suggest that bounded rationality may imply that the agents do not understand and so cannot use the complex mechanisms needed to enforce truth telling. On the other hand, it seems surprising that considerations other than unforeseen contingencies would be needed to generate something different from standard contract theory. Maskin and Tirole do not allow the kind of aversion to payoff uncertainty present in Epstein, Marinacci, and Seo, so this is another direction that may be fruitful.

In short, unforeseen contingencies appear to be important to understanding economic and political institutions but, as yet, economic theory lacks a formal model of the phenomenon that can be used to study these issues.

See Also

- incomplete contracts
- long run and short run

The author would like to thank Eddie Dekel, Jing Li, Aldo Rustichini, and Marie-Odile Yanelle for discussions and comments.

Bibliography

Al Najjar, N., Anderlini, L. and Felli, L. 2006. Undescribable events. *Review of Economic Studies* 73, 849–68.

Dekel, E., Lipman, B. and Rustichini, A. 2001. Representing preferences with a unique subjective state space. *Econometrica* 69, 891–934.

Epstein, L., Marinacci, M. and Seo, K. 2007. Coarse contingencies. Working paper. University of Rochester.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Fishburn, P. 1970. *Utility Theory for Decision Making*. Publications in Operations Research, No. 18, New York: Wiley.

Ghirardato, P. 2001. Coping with ignorance: unforeseen contingencies and non–additive uncertainty. *Economic Theory* 17, 247–76.

Grossman, S. and Hart, O. 1986. The costs and benefits of ownership: a theory of vertical and lateral integration. *Journal of Political Economy* 94, 691–719.

Halpern, J. and Rêgo, L. 2005. Interactive awareness revisited. In *Proceedings of Tenth Conference on Theoretical Aspects of Rationality and Knowledge*, ed. R. van der Meyden. Singapore: National University of Singapore.

Halpern, J. and Rêgo, L. 2006. Extensive games with possibly unaware players. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, NY: ACM Press.

Hart, O. 1995. *Firms, Contracts, and Financial Structure*. Oxford: Clarendon Press.

Hart, O. and Moore, J. 1988. Incomplete contracts and renegotiation. *Econometrica* 56, 755–86.

Hayek, F. 1960. *The Constitution of Liberty*. Chicago: University of Chicago Press.

Heifetz, A., Meier, M. and Schipper, B. 2006a. Interactive unawareness. *Journal of Economic Theory* 130, 78–94.

Heifetz, A., Meier, M. and Schipper, B. 2006b. Unawareness, beliefs, and games. Working paper, University of California–Davis.

Internal Revenue Service. 2006. *Selling Your Home*. Publication 523. Washington, DC: Internal Revenue Service, US Treasury Department.

Kreps, D. 1979. A representation theorem for ‘preference for flexibility’. *Econometrica* 47, 565–76.

Kreps, D. 1992. Static choice and unforeseen contingencies. In *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn*, ed. P. Dasgupta, D. Gale, O. Hart and E. Maskin. Cambridge: MIT Press.

Li, J. 2006a. Information structures with unawareness. Working paper, University of Pennsylvania.

Li, J. 2006b. Dynamic games with unawareness. Working paper, University of Pennsylvania.

Maskin, E. and Tirole, J. 1999. Unforeseen contingencies and incomplete contracts. *Review of Economic*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Studies 66, 83–114.

Modica, S., Rustichini, A. and Tallon, J.-M. 1998. Unawareness and bankruptcy: a general equilibrium model. *Economic Theory* 12, 259–92.

Williamson, O. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

How to cite this article

Lipman, Barton L. "unforeseen contingencies." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000075>
doi:10.1057/9780230226203.1763

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

unit roots

Peter C.B. Phillips
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Models with autoregressive unit roots play a major role in modern time series analysis and are especially important in macroeconomics, where questions of shock persistence arise, and in finance, where martingale concepts figure prominently in the study of efficient markets. The literature on unit roots is vast and applications of unit root testing span the social, environmental and natural sciences. The present article overviews the theory and concepts that underpin this large field of research and traces the originating ideas and econometric methods that have become central to empirical practice.

Keywords

ARCH models; ARMA time-series processes; Bayesian inference; Bayesian statistics; Bayesian time series analysis; bias reduction; break point analysis; classical statistics; cointegration; confidence intervals; efficient markets hypothesis; forecasting; functional central limit theorem; GARCH models; integrated conditional heteroskedasticity models; Lagrange multipliers; long-run variance; martingales; model selection; nonstationarity; polynomials; present value; probability; rational expectations business cycle models; real business cycles; spurious regressions; statistical inference; stochastic trends; term structure of interest rates; unit root distributions; unit roots; Wiener process

Article

Economic and financial time series have frequently been successfully modelled by autoregressive moving-average (ARMA) schemes of the type

$$a(L)y_t = b(L)\varepsilon_t, \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where ε_t is an orthogonal sequence (that is, $E(\varepsilon_t)=0$, $E(\varepsilon_t \varepsilon_s)=0$ for all $t \neq s$), L is the backshift operator for which $Ly_t=y_{t-1}$ and $a(L)$, $b(L)$ are finite-order lag polynomials

$$a(L) = \sum_{i=0}^{p} a_i L^i, \quad b(L) = \sum_{j=0}^{q} b_j L^j,$$

whose leading coefficients are $a_0=b_0=1$. Parsimonious schemes (often with $p+q \leq 3$) are usually selected in practice either by informal 'model identification' processes such as those described in the text by Box and Jenkins (1976) or more formal order-selection criteria which penalize choices of large p and/or q. Model (1) is assumed to be irreducible, so that $a(L)$ and $b(L)$ have no common factors. The model (1) and the time series y_t are said to have an *autoregressive unit root* if $a(L)$ factors as $(1-L)a_1(L)$ and a *moving-average unit root* if $b(L)$ factors as $(1-L)b_1(L)$.

Since the early 1980s, much attention has been focused on models with autoregressive unit roots. In part, this interest is motivated by theoretical considerations such as the importance of martingale models of efficient markets in finance and the dynamic consumption behaviour of representative economic agents in macroeconomics; and, in part, the attention is driven by empirical applications, which have confirmed the importance of random walk phenomena in practical work in economics, in finance, in marketing and business, in social sciences like political studies and communications, and in certain natural sciences. In mathematics and theoretical probability and statistics, unit roots have also attracted attention because they offer new and important applications of functional limit laws and weak convergence to stochastic integrals. The unit root field has therefore drawn in participants from an excitingly wide range of disciplines.

If (1) has an autoregressive unit root, then we may write the model in difference form as

$$\Delta y_t = u_t = a_1(L)^{-1} b(L) \varepsilon_t, \tag{2}$$

where the polynomial $a_1(L)$ has all its zeros outside the unit circle. This formulation suggests more general nonparametric models where, for instance, u_t may be formulated in linear process (or Wold representation) form as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$u_t = c(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \text{ with } \sum_{j=0}^{\infty} c_j^2 < \infty,$$

(3)

or as a general stationary process with spectrum $f_u(\lambda)$. If we solve (2) with an initial state y_0 at t=0, we have the important partial sum representation

$$y_t = \sum_{j=1}^{t} u_j + y_0 = S_t + y_0,$$

(4)

showing that S_t and hence y_t are 'accumulated' or 'integrated' processes proceeding from a certain initialization y_0. A time series y_t that satisfies (2) or (4) is therefore said to be integrated of order one (or a unit root process or an I(1) process) provided $f_u(0)>0$. The latter condition rules out the possibility of a moving-average unit root in the model for u_t that would cancel the effect of the autoregressive unit root (for example, if $b(L)=(1-L)b_1(L)$ then model (2) is $\Delta y_t = \Delta a_1(L)^{-1} b_1(L)\,\varepsilon_t$ or, after cancellation, just $y_t = a_1(L)^{-1} b_1(L)\,\varepsilon_t$, which is not I(1)). Note that this possibility is also explicitly ruled out in the ARMA case by the requirement that $a(L)$ and $b(L)$ have no common factors. Alternatively, we may require that $u_t \neq \Delta v_t$ for some weakly stationary time series v_t, as in Leeb and Pötscher (2001) who provide a systematic discussion of I(1) behaviour. The partial sum process S_t in (4) is often described as a *stochastic trend*.

The representation (4) is especially important because it shows that the effect of the random shocks u_j on y_t does not die out as the time distance between j and t grows large. The shocks u_j then have a *persistent* effect on y_t in this model, in contrast to stationary systems. Whether actual economic time series have this characteristic or not is, of course, an empirical issue. The question can be addressed through statistical tests for the presence of a unit root in the series, a subject which has grown to be of major importance since the mid-1980s and which will be discussed later in this article. From the perspective of economic modelling the issue of persistence is also important because, if macroeconomic variables like real GNP have a unit root, then shocks to real GNP have permanent effects, whereas in traditional business cycle theory the effect of shocks on real GNP is usually considered to be only temporary. In more recent real business cycle theory, variables like real GNP are modelled in such a way that over the long run their paths are determined by supply side shocks that can be ascribed to technological and demographic forces from outside the model. Such economic models are more compatible with the statistical model (4) or close approximations to it in which the roots are local to unity in a sense that is described later in this essay.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Permanent and transitory effects in (4) can be distinguished by decomposing the process u_t in (3) as follows

$$u_t = \{C(1) + (L-1)\breve{C}(L)\}\varepsilon_t = C(1)\varepsilon_t + \breve{\varepsilon}_{t-1} - \breve{\varepsilon}_t, \tag{5}$$

where $\breve{\varepsilon}_t = \breve{C}(L)\varepsilon_t$, $\breve{C}(L) = \Sigma_0^\infty \breve{c}_j L^j$ and $\breve{c}_j = \Sigma_{j+1}^\infty c_s$. The decomposition (5) is valid algebraically if

$$\sum_{j=0}^{\infty} j^{1/2}|c_j| < \infty, \tag{6}$$

as shown in Phillips and Solo (1992), where validity conditions are systematically explored. Equation (5) is sometimes called the Beveridge–Nelson (1981) or BN decomposition of u_t, although both specialized and more general versions of it were known and used beforehand. The properties of the decomposition were formally investigated and used for the development of laws of large numbers and central limit theory and invariance principles in the paper by Phillips and Solo (1992). When the decomposition is applied to (4) it yields the representation

$$y_t = C(1)\sum_1^t \varepsilon_j + \tilde{\varepsilon}_0 - \tilde{\varepsilon}_t + y_0 = C(1)\sum_1^t \varepsilon_j + \xi_t + y_0, \text{ say}, \tag{7}$$

where $\xi_t = \breve{\varepsilon}_0 - \breve{\varepsilon}_t$. The right side of (7) decomposes y_t into three components: the first is a martingale component, $Y_t = C(1)\Sigma_1^t \varepsilon_j$, where the effects of the shocks ε_j are permanent; the second is a stationary component, where the effects of shocks are transitory, viz. $\xi_t = \breve{\varepsilon}_0 - \breve{\varepsilon}_t$, since the process $\breve{\varepsilon}_t$ is stationary with valid Wold representation $\breve{\varepsilon}_t = \breve{C}(L)\varepsilon_t$ under (6) when ε_t is stationary with variance σ^2; and the third being the initial condition y_0. The relative strength of the martingale component is measured by the magnitude of the (infinite dimensional) coefficient $C(1) = \Sigma_{j=0}^\infty c_j$, which plays a large role in the measurement of long-run effects in applications. Accordingly, the decomposition (7) is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

sometimes called the martingale decomposition (cf., Hall and Heyde, 1980) where it was used in various forms in the probability literature prior to its use in economics.

The leading martingale term $Y_t = C(1)\Sigma_{s=1}^{t}\varepsilon_s$ in (7) is a partial sum process or stochastic trend and, under weak conditions on ε_t (see Phillips and Solo, 1992, for details) this term satisfies a functional central limit theorem whereby the scaled process

$$n^{-1/2}Y_{[nr]} \Rightarrow B(r), \tag{8}$$

a Brownian motion with variance $\omega^2 = C(1)^2\sigma^2 = 2\pi f_u(0)$, a parameter which is called the long-run variance of u_t, and where [·] signifies the integer part of its argument. Correspondingly,

$$n^{-1/2}Y_{[nr]} \Rightarrow B(r), \tag{9}$$

provided $Y_0 = o_p(\sqrt{n})$ A related result of great significance is based on the limit

$$n^{-1}\sum_{t=1}^{[nr]} Y_{t-1}\varepsilon_t C(1) \Rightarrow \int_0^r BdB \tag{10}$$

of the sample covariance of Y_{t-1} and its forward increment, $C(1)\,\varepsilon_t$. The limit process $M(r) = \int_0^r BdB$ is represented here as an Ito (stochastic) integral and is a continuous time martingale. The result may be proved directly (Solo, 1984; Phillips, 1987a; Chan and Wei, 1988) or by means of martingale convergence methods (Ibragimov and Phillips, 2004) which take advantage of the fact that $\Sigma_{t=1}^{k} Y_{t-1}\varepsilon_t$ is a martingale. The limit theory given (9) and (10) was extended in Phillips (1987b; 1988a) and Chan and Wei (1987) to cases where the model (2) has an autoregressive root in the vicinity of unity ($\rho = 1 + \frac{c}{n}$, for some fixed c) rather than precisely at unity, in which case the limiting process is a linear diffusion (or Ornstein–Uhlenbeck process) with parameter c. This limit theory has proved particularly useful in the analysis of asymptotic local power functions of unit root tests (Phillips, 1987b) and the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
construction of confidence intervals (Stock, 1991). Phillips and Magdalinos (2007) considered moderate deviations from unity of the form $\rho = 1 + \frac{c}{k}$, where $k \to \infty$ but $\frac{k}{n} \to 0$, so that the roots are local but further away from unity, showing that central limit laws rather than functional laws apply in this case (see also Giraitis and Phillips, 2006). This theory is applicable to mildly explosive processes (where $c>0$) and therefore assists in bridging the gap between the limit theory for the stationary, unit root and explosive cases.

Both (8) and (10) have important multivariate generalizations that play a critical role in the study of spurious regressions (Phillips, 1986) and cointegration limit theory (Phillips and Durlauf, 1986; Engle and Granger, 1987; Johansen, 1988; Phillips, 1988a; Park and Phillips, 1988; 1989). In particular, if $y_t = (y'_{at}, y'_{bt})'$, $u_t = (u'_{at}, u'_{bt})'$ and $\varepsilon_t = (\varepsilon'_{at}, \varepsilon'_{bt})'$ are vector processes and $E(\varepsilon_t \varepsilon'_t) = \Sigma$, then: (i) the decomposition (5) continues to hold under (6), where $|c_j|$ is interpreted as a matrix norm; (ii) the functional law (8) holds and the limit process is vector Brownian motion $B = (B'_a, B'_b)'$ with covariance matrix $\Omega = C(1)\Sigma C(1)'$; and (iii) sample covariances converge weakly to stochastic processes with drift, as in

$$n^{-1}\sum_{t=1}^{[nr]} y_{at-1} u'_{bt} \Rightarrow \int_0^r B_a dB'_b + \lambda_{ab} r, \tag{11}$$

where $\lambda_{ab} = \Sigma_{k=1}^{\infty} E(u_{a0} u'_{bk})$ is a one sided long-run covariance matrix. The limit process on the right side of (11) is a semimartingale (incorporating a deterministic drift function $\lambda_{ab} r$) rather than a martingale when $\lambda_{ab} \neq 0$.

The decomposition (7) plays an additional role in the study of cointegration (Engle and Granger, 1987). When the coefficient matrix $C(1)$ is singular and β spans the null space of $C(1)'$, then $\beta' C(1) = 0$ and (7) leads directly to the relationship

$$\beta' Y_t = 0, \quad \text{a.s.},$$

which may be interpreted as a long run equilibrium (cointegrating) relationship between the stochastic trends (Y_t) of y_t. Correspondingly, we have the empirical cointegrating relationship

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\beta' y_t = v_t,$$

among the observed series y_t with a residual $v_t = \beta'(\xi_t + y_0)$ that is stationary. The columns of β span what is called the cointegration space.

The above discussion presumes that the initialization y_0 has no impact on the limit theory, which will be so if y_0 is small relative to the sample size, specifically, if $y_0 = o_p(\sqrt{n})$. However, if $y_0 = O_p(\sqrt{n})$, for example if $y_0 = y_{0\theta_n}$ is indexed to depend on past shocks u_{-j} (satisfying a process of the form (3)) to some point in the distant past θ_n which is measured in terms of the sample size n, then the results can differ substantially. Thus, if $\theta_n=[\kappa n]$, for some fixed parameter $\kappa>0$, then $y_{0\theta_n} = \sum_1^{[\kappa n]} u_{-j}$, and $n^{-1/2} y_{0\theta_n} \Rightarrow B_0(\kappa)$, for some Brownian motion $B_0(\kappa)$ with covariance matrix Ω_{00} given by the long-run variance matrix of u_{-j}. Under such an initialization, (9) and (11) are replaced by

$$n^{-1/2} y_{[nr]} \Rightarrow B(r) + B_0(\kappa) := B(r, \kappa), \quad \text{say} \tag{12}$$

and

$$n^{-1} \sum_{t=1}^{[nr]} y_{at-1} u_{bt}' \Rightarrow \int_0^r B_a(s, \kappa)\, dB_b(s)' + \lambda_{ab} r,$$

so that initializations play a role in the limit theory. This role becomes dominant when κ becomes very large, as is apparent from (12). The effect of initial conditions on unit root limit theory was examined in simulations by Evans and Savin (1981; 1984), by continuous record asymptotics by Phillips (1987a), in the context of power analysis by Müller and Elliott (2003), for models with moderate deviations from unity by Andrews and Guggenberger (2006), and for cases of large κ by Phillips (2006).

Model (4) is of special interest to economists working in finance because its output, y_t, behaves as if it has no fixed mean and this is a characteristic of many financial time series. If the components u_j are independent and identically distributed (i.i.d.) then y_t is a random walk. More generally, if u_j is a martingale difference sequence (mds) (that is orthogonal to its own past history so that $E_{j-1}(u_j) = E(u_j|u_{j-1}, u_{j-2}, \ldots, u_1) = 0$) then y_t is a martingale. Martingales are the essential

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
mathematical elements in the development of a theory of fair games and they now play a key role in the mathematical theory of finance, exchange rate determination and securities markets. Duffie (1988) provides a modern treatment of finance that makes extensive use of this theory.

In empirical finance much attention has recently been given to models where the conditional variance $E(u_j^2|u_{j-1}, u_{j-2}, \ldots, u_1) = \sigma_j^2$ is permitted to be time varying. Such models have been found to fit financial data well and many different parametric schemes for σ_j^2 have been devised, of which the ARCH (autoregressive conditional heteroskedasticity) and GARCH (generalized ARCH) models are the most common in practical work. These models come within the general class of models like (1) with mds errors. Some models of this kind also allow for the possibility of a unit root in the determining mechanism of the conditional variance σ_j^2 and these are called integrated conditional heteroskedasticity models. The IGARCH (integrated GARCH) model of Engle and Bollerslev (1986) is an example, where for certain parameters $\omega \geq 0$, $\beta \geq 0$, and $\alpha > 0$, we have the specification

$$\sigma_j^2 = \omega + \beta\sigma_{j-1}^2 + \alpha u_{j-1}^2, \tag{13}$$

with $\alpha + \beta = 1$ and $u_j = \sigma_j z_j$, where the z_j are i.i.d. innovations with $E(z_j)=0$ and $E(z_j^2) = 1$. Under these conditions, the specification (13) has the alternative form

$$\sigma_j^2 = \omega + \sigma_{j-1}^2 + \alpha\sigma_{j-1}^2(z_{j-1}^2 - 1), \tag{14}$$

from which it is apparent that σ_j^2 has an autoregressive unit root. Indeed, since

$$E(\sigma_j^2|\sigma_{j-1}^2) = \omega + \sigma_{j-1}^2,$$

σ_j^2 is a martingale when $\omega = 0$. It is also apparent from (14) that shocks as manifested in the deviation $z_{j-1}^2 - 1$ are persistent in σ_j^2. Thus, σ_j^2 shares some of the characteristics of an I(1) integrated process.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

But in other ways, σ_j^2 is very different. For instance, when $\omega=0$ then $\sigma_j^2 \to 0$ almost surely as $j\to\infty$ and, when $\omega>0$, σ_j^2 is asymptotically equivalent to a strictly stationary and ergodic process. These and other features of models like (13) for conditional variance processes with a unit root are studied in Nelson (1990).

In macroeconomic theory also, models such as (2) play a central role in modern treatments. In a highly influential paper, R. Hall (1978) showed that under some general conditions consumption is well modelled as a martingale, so that consumption in the current period is the best predictor of future consumption, thereby providing a macroeconomic version of the efficient markets hypothesis. Much attention has been given to this idea in subsequent empirical work.

One generic class of economic model where unit roots play a special role is the 'present value model' of Campbell and Shiller (1988). This model is based on agents' forecasting behaviour and takes the form of a relationship between one variable Y_t and the discounted, present value of rational expectations of future realizations of another variable $X_{t+i}(i=0,1,2,\ldots)$. More specifically, for some stationary sequence c_t (possibly a constant) we have

$$Y_t = \theta(1-\delta)\sum_{i=0}^{\infty}\delta^i E_t(X_{t+i}) + c_t. \tag{15}$$

When X_t is a martingale, $E_t(X_{t+i})=X_t$ and (15) becomes

$$Y_t = \theta X_t + c_t, \tag{16}$$

so that Y_t and X_t are cointegrated in the sense of Engle and Granger (1987). More generally, when X_t is I (1) we have

$$Y_t = \theta X_t + \bar{c}_t, \tag{17}$$

where $\bar{c}_t = c_t + \theta\sum_{k=1}^{\infty}\delta^k E_t(\Delta X_{t+k})$, so that Y_t and X_t are also cointegrated in this general case.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Models of this type arise naturally in the study of the term structure of interest rates, stock prices and dividends and linear-quadratic intertemporal optimization problems.

An important feature of these models is that they result in parametric linear cointegrating relations such as (16) and (17). This linearity in the relationship between Y_t and X_t accords with the linear nature of the partial sum process that determines X_t itself, as seen in (4), and has been extensively studied since the mid-1980s. However, in more general models, economic variables may be determined in terms of certain nonlinear functions of fundamentals. When these fundamentals are unit root processes like (4), then the resulting model has the form of a nonlinear cointegrating relationship. Such models are relevant, for instance, in studying market interventions by monetary and fiscal authorities (Park and Phillips, 2000; Hu and Phillips, 2004) and some of the asymptotic theory for analysing parametric models of this type and for statistical inference in such models is given in Park and Phillips (1999; 2001), de Jong (2004), Berkes and Horváth (2006) and Pötscher (2004). More complex models of this type are nonparametric and different methods of inference are typically required with very different limit theories and typically slower convergence rates (Karlsen, Myklebust, and Tjøstheim, 2007; Wang and Phillips, 2006). Testing for the presence of such nonlinearities can therefore be important in empirical practice (Hong and Phillips, 2005; Kasparis, 2004).

Statistical tests for the presence of a unit root fall into the general categories of classical and Bayesian, corresponding to the mode of inference that is employed. Classical procedures have been intensively studied and now occupy a vast literature. Most empirical work to date has used classical methods but some attention has been given to Bayesian alternatives and direct model selection methods. These approaches will be outlined in what follows.

Although some tests are known to have certain limited (asymptotic) point optimality properties, there is no known procedure which uniformly dominates others, even asymptotically. Ploberger (2004) provides an analysis of the class of asymptotically admissable tests in problems that include the simplest unit root test, showing that the conventional likelihood ratio (LR) test (or Dickey–Fuller, 1979; 1981, *t* test) is not within this class, so that the LR test, while it may have certain point optimal properties, is either inadmissible or must be modified so that it belongs to the class. This fundamental difficulty, together with the nonstandard nature of the limit theory and the more complex nature of the asymptotic likelihood in unit root cases partly explains why there is such a proliferation of test procedures and simulation studies analysing performance characteristics in the literature.

Classical tests for a unit root may be classified into parametric, semiparametric and nonparametric categories. Parametric tests usually rely on augmented regressions of the type

$$\Delta y_t = a y_{t-1} + \sum_{i=1}^{k-1} \phi_i \Delta y_{t-i} + e_t, \tag{18}$$

where the lagged variables are included to model the stationary error u_t in (2). Under the null hypothesis

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of a unit root, we have a=0 in (18) whereas when y_t is stationary we have a<0. Thus, a simple test for the presence of a unit root against a stationary alternative in (18) is based on a one-sided t-ratio test of $\mathcal{H}_0: a = 0$ against $\mathcal{H}_1: a < 0$. This test is popularly known as the ADF (or augmented Dickey–Fuller) test (Said and Dickey, 1984) and follows the work of Dickey and Fuller (1979; 1981) for testing Gaussian random walks. It has been extensively used in empirical econometric work since the Nelson and Plosser (1982) study, where it was applied to 14 historical time series for the USA leading to the conclusion that unit roots could not be rejected for 13 of these series (all but the unemployment rate). In that study, the alternative hypothesis was that the series were stationary about a deterministic trend (that is, trend stationary) and therefore model (18) was further augmented to include a linear trend, viz.

$$\Delta y_t = \mu + \beta t + a y_{t-1} + \sum_{i=1}^{k-1} \phi_i \Delta y_{t-i} + e_t. \tag{19}$$

When y_t is trend stationary we have a<0 and $\beta \neq 0$ in (19), so the null hypothesis of a difference stationary process is a=0 and β=0. This null hypothesis allows for the presence of a non-zero drift in the process when the parameter $\mu \neq 0$. In this case a joint test of the null hypothesis $\mathcal{H}_0: a = 0$, β=0 can be mounted using a regression F-test. ADF tests of a=0 can also be mounted directly using the coefficient estimate from (18) or (19), rather than its t ratio (Xiao and Phillips, 1998).

What distinguishes both these and other unit root tests is that critical values for the tests are not the same as those for conventional regression F- and t-tests, even in large samples. Under the null, the limit theory for these tests is nonstandard and involves functionals of a Wiener process. Typically, the critical values for five or one per cent level tests are much further out than those of the standard normal or chi-squared distributions. Specific forms for the limits of the ADF t-test (ADF_t) and coefficient (ADF_a) test are

$$ADF_t \Rightarrow \frac{\int_0^1 W dW}{\left(\int_0^1 W^2\right)^{1/2}}, \quad ADF_a \Rightarrow \frac{\int_0^1 W dW}{\int_0^1 W^2}, \tag{20}$$

where W is a standard Wiener process or Brownian motion with variance unity. The limit distributions represented by the functionals (20) are known as unit root distributions. The limit theory was first explored for models with Gaussian errors, although not in the Wiener process form and not using functional limit laws, by Dickey (1976), Fuller (1976) and Dickey and Fuller (1979; 1981), who also provided tabulations. For this reason, the distributions are sometimes known as Dickey–Fuller

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

distributions. Later work by Said and Dickey (1984) showed that, if the lag number k in (18) is allowed to increase as the sample size increases with a condition on the divergence rate that $k = o(n^{1/3})$, then the ADF test is asymptotically valid in models of the form (2) where u_t is not necessarily autoregressive. Several other parametric procedures have been suggested, including Von Neumann ratio statistics (Sargan and Bhargava, 1983; Bhargava, 1986; Stock, 1994a), instrumental variable methods (Hall, 1989; Phillips and Hansen, 1990) and variable addition methods (Park, 1990). The latter also allow a null hypothesis of trend stationarity to be tested directly, rather than as an alternative to difference stationarity. Another approach that provides a test of a null of trend stationarity is based on the unobserved components representation

$$y_t = \mu + \beta t + r_t + u_t, \quad r_t = r_{t-1} + v_t, \tag{21}$$

which decomposes a time series y_t into a deterministic trend, an integrated process or random walk (r_t) and a stationary residual (u_t). The presence of the integrated process component in y_t can then be tested by testing whether the variance (σ_v^2) of the innovation v_t is zero. The null hypothesis is then $\mathcal{H}_0: \sigma_v^2 = 0$, which corresponds to a null of trend stationarity. This hypothesis can be tested in a very simple way using the Lagrange multiplier (LM) principle, as shown in Kwiatkowski et al. (1992), leading to a commonly used test known as the KPSS test. If $\hat{e}_t$ denotes the residual from a regression of y_t on a deterministic trend (a simple linear trend in the case of (21) above) and $\tilde{\omega}_e^2$ is a HAC (heteroskedastic and autocorrelation consistent) estimate constructed from $\hat{e}_t$, then the KPSS statistic has the simple form

$$LM = \frac{n^{-2}\sum_{t=1}^{n} S_t^2}{\tilde{\omega}_e^2},$$

where S_t is the partial sum process of the residuals $\sum_{j=1}^{t} \hat{e}_j$. Under the null hypothesis of stationarity, this LM statistic converges to $\int_0^1 V_X^2$, where V_X is a generalized Brownian bridge process whose construction depends on the form (X) of the deterministic trend function. Power analysis indicates that test power depends importantly on the choice of bandwidth parameter in HAC estimation and some recent contributions to this subject are Sul, Phillips and Choi (2006) and Müller (2005) and Harris, Leybourne and McCabe (2007). Other general approaches to testing I(0) versus I(1) have been

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
considered in Stock (1994a, 1999).

By combining r_t and u_t in (21) the components model may also be written as

$$y_t = \mu + \beta t + x_t, \quad \Delta x_t = a x_{t-1} + \eta_t. \tag{22}$$

In this format it is easy to construct an LM test of the null hypothesis that y_t has a stochastic trend component by testing whether a=0 in (22). When a=0, (22) reduces to

$$\Delta y_t = \beta + \eta_t, \quad \text{or} \quad y_t = \beta t + \sum_1^t \eta_i + y_0, \tag{23}$$

and so the parameter μ is irrelevant (or surplus) under the null. However, the parameter β retains the same meaning as the deterministic trend term coefficient under both the null and the alternative hypothesis. This approach has formed the basis of several tests for a unit root that have been developed (see Bhargava, 1986; Schmidt and Phillips, 1992) and the parameter economy of this model gives these tests some advantage in terms of power over procedures like the ADF in the neighbourhood of the null. This power advantage may be further exploited by considering point optimal alternatives in the construction of the test and in the process of differencing (or detrending) that leads to (23), as pursued by Elliott, Rothenberg and Stock (1995). In particular, note that (23) involves detrending under the null hypothesis of a unit root, which amounts to first differencing, whereas if the root were local to unity, the appropriate procedure would be to use quasi differencing. However, since the value of the coefficient in the locality of unity is unknown (otherwise, there would be no need for a test), it can only be estimated or guessed. The procedure suggested by Elliott, Rothenberg and Stock (1995) is to use a value of the localizing coefficient in the quasi-differencing process for which asymptotic power is calculated by simulation to be around 50 per cent, a setting which depends on the precise model for estimation that is being used. This procedure, which is commonly known as generalized least squares (GLS) detrending (although the terminology is a misnomer because quasi-differencing not full GLS is used to accomplish trend elimination) is then asymptotically approximately point optimal in the sense that its power function touches the asymptotic power envelope at that value. Simulations show that this method has some advantage in finite samples, but it is rarely used in empirical work in practice, partly because of the inconvenience of using specialized tables for the critical values of the resulting test and partly because settings for the localizing coefficient are arbitrary and depend on the form of the empirical model.

Some unit root tests based on standard limit distribution theory have been developed. Phillips and Han

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(2008), for example, give an autoregressive coefficient estimator whose limit distribution is standard normal for all stationary, unit root and local to unity values of the autoregressive coefficient. This estimator may be used to construct tests and valid confidence intervals, but tests suffer power loss because the rate of convergence of the estimator is $\sqrt{n}$ uniformly over these parameter values. So and Shin (1999) and Phillips, Park and Chang (2004) showed that certain nonlinear instrumental variable estimators, such as the Cauchy estimator, also lead to *t*-tests for a unit root which have an asymptotic standard normal distribution. Again, these procedures suffer power loss from reduced convergence rates (in this case, $n^{1/4}$), but have the advantage of uniformity and low bias. Bias is a well known problem in autoregressive estimation and many procedures for addressing the problem have been considered. It seems that bias reduction is particularly advantageous in the case of unit root tests in panel data, where cross-section averaging exacerbates bias effects when the time dimension is small. Some simulation and indirect inference procedures for bias removal have been successfully used both in autoregressions (Andrews, 1993; Gouriéroux, Renault and Touzi, 2000) and in panel dynamic models (Gouriéroux, Phillips and Yu, 2006).

Semiparametric unit root tests are among the most commonly used unit root tests in practical work and are appealing in terms of their generality and ease of use. Tests in this class employ nonparametric methods to model and estimate the contribution from the error process u_t in (2), allowing for both autocorrelation and heterogeneity. These tests and the use of functional limit theory methods in econometrics, leading to the limit formulae (20), were introduced in Phillips (1987a). Direct least squares regression on

$$\Delta y_t = a y_{t-1} + u_t \tag{24}$$

gives an estimate of the coefficient and its *t*-ratio in this equation. These two statistics are then corrected to deal with serial correlation in u_t by employing an estimate of the variance of u_t and its long-run variance. The latter estimate may be obtained by a variety of kernel-type HAC or other spectral estimates (such as autoregressive spectral estimates) using the residuals $\hat{u}_t$ of the OLS regression on (24). Automated methods of bandwidth selection (or order selection in the case of autoregressive spectral estimates) may be employed in computing these HAC estimates and these methods typically help to reduce size distortion in unit root testing (Lee and Phillips, 1994; Stock, 1994a; Ng and Perron, 1995; 2001). However, care needs to be exercised in the use of automated procedures in the context of stationarity tests such as the KPSS procedure to avoid test inconsistency (see Lee, 1996; Sul, Phillips and Choi, 2006).

This semiparametric approach leads to two test statistics, one based on the coefficient estimate, called the $Z(a)$ test, the other based on its *t*-ratio, called the $Z(t)$ test. The limit distributions of these statistics are the same as those given in (20) for the ADF coefficient and *t*-ratio tests, so the tests are asymptotically equivalent to the corresponding ADF tests. Moreover, the local power functions are also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

equivalent to those of the Dickey–Fuller and ADF tests, so that there is no loss in asymptotic power from the use of nonparametric methods to address autocorrelation and heterogeneity (Phillips, 1987b). Similar semiparametric corrections can be applied to the components models (21) and (22) leading to generally applicable LM tests of stationarity ($\sigma_v^2 = 0$) and stochastic trends (a=0).

The Z tests were extended in Phillips and Perron (1988) and Ouliaris, Park and Phillips (1989) to models with drift, and by Perron (1989) and Park and Sung (1994) to models with structural breaks in the drift or deterministic component. An important example of the latter is the trend function

$$h_t = \sum_{j=0}^{p} f_j t^j + \sum_{j=0}^{p} f_{m,j} t_m^j, \quad \text{where } t_m^j = \begin{cases} 0 & t \in \{1, \ldots, m\} \\ (t-m)^j & t \in \{m+1, \ldots, n\} \end{cases} \tag{25}$$

which allows for the presence of a break in the polynomial trend at the data point t=m+1. Collecting the individual trend regressors in (25) into the vector x_t, there exists a continuous function $X(r) = (1, r, \ldots, r^p)'$ such that $D_n^{-1} x_{[nr]} \to X(r)$ as $n \to \infty$ uniformly in $r \in [0,1]$, where D_n=diag(1,n,…, n^p). If $\mu = \lim_{n\to\infty}(m/n) > 0$ is the limit of the fraction of the sample where the structural change occurs, then the limiting trend function $X_\mu(r)$ corresponding to (25) has a similar break at the point μ. All the unit root tests discussed above continue to apply as given for such broken trend functions with appropriate modifications to the limit theory to incorporate the limit function $X_\mu(r)$. Indeed, (25) may be extended further to allow for multiple break points in the sample and in the limit process. The tests may be interpreted as tests for the presence of a unit root in models where broken trends may be present in the data. The alternative hypothesis in this case is that the data are stationary about a broken deterministic trend of degree p.

In order to construct unit root tests that allow for breaking trends like (25) it is necessary to specify the break point m. (Correspondingly, the limit theory depends on $X_\mu(r)$ and therefore on μ.) In effect, the break point is exogenously determined. Perron (1989) considered linear trends with single break points intended to capture the 1929 stock market crash and the 1974 oil price shock in this way. An alternative perspective is that any break points that occur are endogenous to the data and unit root tests should take account of this fact. In this case, alternative unit root tests have been suggested (for example, Banerjee, Lumsdaine and Stock, 1992; Zivot and Andrews, 1992) that endogenize the break point by choosing the value of m that gives the least favourable view of the unit root hypothesis. Thus, if $ADF(m)$ denotes the ADF statistic given by the t-ratio for α in the ADF regression (19) with a broken trend function like (25), then the trend break ADF statistic is

$$ADF(\hat{m}) = \min_{\underline{m} \le m \le \overline{m}} ADF(m), \quad \text{where } \underline{m} = [n\underline{\mu}], \quad \overline{m} = [n\overline{\mu}],$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(26)

for some $0 < \underline{\mu} < \bar{\mu} < 1$. The limit theory for this trend break *ADF* statistic is given by

$$ADF(\hat{m}) \Rightarrow \inf_{\mu \in [\underline{\mu}, \bar{\mu}]} \left[\int_0^1 W_{X_\mu} dW \right] \left[\int_0^1 W_{X_\mu}^2 \right]^{-1/2}, \tag{27}$$

where W_X is detrended standard Brownian motion defined by $W_X(r) = W(r) - \left[\int_0^1 WX\right]\left[\int_0^1 XX\right]^{-1} X(r)$. The limit process $X_\mu(r)$ that appears in the functional W_{X_μ} is dependent on the trend break point μ over which the functional is minimized. Similar extensions to trend breaks are possible for other unit root tests and to multiple breaks (Bai, 1997; Bai and Perron, 1998; 2006; Kapetanios, 2005). Critical values of the limiting test statistic (27) are naturally further out in the tail than those of the exogenous trend break statistic, so it is harder to reject the null hypothesis of a unit root when the break point is considered to be endogenous.

Asymptotic and finite sample critical values for the endogenized trend break *ADF* unit root test are given in Zivot and Andrews (1992). Simulations studies indicate that the introduction of trend break functions leads to further reductions in the power of unit root tests and to substantial finite sample size distortion in the tests. Sample trajectories of a random walk are often similar to those of a process that is stationary about a broken trend for some particular breakpoint (and even more so when several break points are permitted in the trend). So continuing reductions in the power of unit root tests against competing models of this type is to be expected and discriminatory power between such different time series models is typically low. In fact, the limit Brownian motion process in (9) can itself be represented as an infinite linear random combination of deterministic functions of time, as discussed in Phillips (1998), so there are good theoretical reasons for anticipating this outcome. Carefully chosen trend stationary models can always be expected to provide reasonable representations of given random walk or unit root data, but such models are certain to fail in post-sample projections as the post-sample data drifts away from any given trend or broken trend line. Phillips (1998; 2001) explores the impact of these considerations in a systematic way.

From a practical standpoint, models with structural breaks attach unit weight and hence persistence to the effects of innovations at particular times in the sample period. In effect, break models simply dummy out the effects of certain observations by parameterizing them as persistent effects. To the extent that persistent shocks of this type occur intermittently throughout the entire history of a process, these models are therefore similar to models with a stochastic trend. However, if only one or a small number of such breaks occur then the process does have different characteristics from that of a stochastic trend.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In such cases, it is often of interest to identify the break points endogenously and relate such points to institutional events or particular external shocks that are know to have occurred.

More general nonparametric tests for a unit root are also possible. These rely on frequency domain regressions on (24) over all frequency bands (Choi and Phillips, 1993). They may be regarded as fully nonparametric because they test in a general way for coherency between the series y_t and its first difference Δy_t. Other frequency domain procedures involve the estimation of a fractional differencing parameter and the use of tests and confidence intervals based on the estimate. The time series y_t is fractionally integrated with memory parameter d if $(1-L)^d y_t = u_t$ and u_t is a stationary process with spectrum $f_u(\lambda)$ that is continuous at the origin with $f_u(0)>0$, or a (possibly mildly heterogeneous) process of the form given in (3). Under some rather weak regularity conditions, it is possible to estimate d consistently by semiparametric methods irrespective of the value of d. Shimotsu and Phillips (2005) suggest an exact local Whittle estimator $\hat{d}$ that is consistent for all d and for which $\sqrt{n}(\hat{d}-d) \Rightarrow N(0, \frac{1}{4})$, extending earlier work by Robinson (1995) on local Whittle estimation in the stationary case where $|d| < 1$. These methods are narrow band procedures focusing on frequencies close to the origin, so that long run behaviour is captured. The Shimotsu–Phillips estimator may be used to test the unit root hypothesis $\mathcal{H}_0 : d = 1$ against alternatives such as $\mathcal{H}_1 : d < 1$. The limit theory may also be used to construct valid confidence intervals for d.

The $Z(a)$, $Z(t)$ and ADF tests are the most commonly used unit root tests in empirical research. Extensive simulations have been conducted to evaluate the performance of the tests. It is known that the $Z(a)$, $Z(t)$ and ADF tests all perform satisfactorily except when the error process u_t displays strong negative serial correlation. The $Z(a)$ test generally has greater power than the other two tests but also suffers from more serious size distortion. All of these tests can be used to test for the presence of cointegration by using the residuals from a cointegrating regression. Modification of the critical values used in these tests is then required, for which case the limit theory and tables were provided in Phillips and Ouliaris (1990) and updated in MacKinnon (1994).

While the Z tests and other semiparametric procedures are designed to cope with mildly heterogeneous processes, some further modifications are required when there is systematic time-varying heterogeneity in the error variances. One form of systematic variation that allows for jumps in the variance has the form $E(\epsilon_t^2) = \sigma_t^2 = \sigma^2 g(\frac{t}{n})$, where the variance evolution function $g(\frac{t}{n})$ may be smooth except for simple jump discontinuities at a finite number of points. Such formulations introduce systematic time variation into the errors, so that we may write $\epsilon_t = g(\frac{t}{n})\eta_t$, where ζ_t is a martingale difference sequence with variance $E\zeta_t^2 = \sigma^2$. These evolutionary changes then have persistent effects on partial sums of ε_t, thereby leading to alternate functional laws of the form

$$n^{-1/2}\gamma_{[nr]} \Rightarrow B_g(r) = \int_0^r g(s)\, dB(s),$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in place of (8). Accordingly, the limit theory for unit root tests changes and some nonparametric modification of the usual tests is needed to ensure that existing asymptotic theory applies (Beare, 2006) or to make appropriate corrections in the limit theory (Cavaliere, 2004; Cavaliere and Taylor, 2007) so that there is less size distortion in the tests.

An extension of the theory that is relevant in the case of quarterly data is to the seasonal unit root model

$$(1 - L^4) y_t = u_t. \tag{28}$$

Here, the polynomial $1–L^4$ can be factored as $(1–L)(1+L)(1+L^2)$, so that the unit roots (or roots on the unit circle) in (28) occur at 1, –1, i, and $–i$, corresponding to the annual (L=1) frequency, the semi-annual (L=–1) frequency, and the quarter and three quarter annual ($L=i,–i$) frequency respectively. Quarterly differencing, as in (28), is used as a seasonal adjustment device, and it is of interest to test whether the data supports the implied hypothesis of the presence of unit roots at these seasonal frequencies. Other types of seasonal processes, say monthly data, can be analysed in the same way.

Tests for seasonal unit roots within the particular context of (28) were studied by Hylleberg et al. (1990), who extended the parametric *ADF* test to the case of seasonal unit roots. In order to accommodate fourth differencing, the autoregressive model is written in the new form

$$\Delta_4 y_t = \alpha_1 y_{1t-1} + \alpha_2 y_{2t-1} + \alpha_3 y_{3t-2} + \alpha_4 y_{3t-1} + \sum_{i=1}^{p} \phi_i \Delta_4 y_{t-i} + \varepsilon_t, \tag{29}$$

where $\Delta_4=1–L^4$, $y_{1t}=(1+L)(1+L^2)y_t$, $y_{2t}=–(1–L)(1+L^2)y_t$, and $y_{3t}=–(1–L^2)y_t$. The transformed data y_{1t}, y_{2t}, y_{3t} retain the unit root at the zero frequency (long run), the semi-annual frequency (two cycles per year), and the annual frequency (one cycle per year). When $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0$, there are unit roots at the zero and seasonal frequencies. To test the hypothesis of a unit root (L=1) in this seasonal model, a t-ratio test of $\alpha_1=0$ is used. Similarly, the test for a semi-annual root (L=–1) is based on a t-ratio test of $\alpha_2=0$, and the test for an annual root on the t-ratios for $\alpha_3=0$ or $\alpha_4=0$. If each of the α's is different from zero, then the series has no unit roots at all and is stationary. Details of the implementation of this procedure are given in Hylleberg et al. (1990), the limit theory for the tests is developed in Chan and Wei (1988), and Ghysels and Osborne (2001) provide extensive discussion and applications.

Most empirical work on unit roots has relied on classical tests of the type described above. But Bayesian methods are also available and appear to offer certain advantages like an exact finite sample analysis (under specific distributional assumptions) and mass point posterior probabilities for break point

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
analysis. In addressing the problem of trend determination, traditional Bayes methods may be employed such as the computation of Bayesian confidence sets and the use of posterior odds tests. In both cases prior distributions on the parameters of the model need to be defined and posteriors can be calculated either by analytical methods or by numerical integration. If (18) is rewritten as

$$y_t = \rho y_{t-1} + \sum_{1}^{k-1} \phi_i \Delta y_{t-i} + e_t \tag{30}$$

then the posterior probability of the nonstationary set $\{\rho \geq 1\}$ is of special interest in assessing the evidence in support of the presence of a stochastic trend in the data. Posterior odds tests typically proceed with 'spike and slab' prior distributions (π) that assign an atom of mass such as $\pi(\rho = 1) = \theta$ to the unit-root null and a continuous distribution with mass $1-\theta$ to the stationary alternative, so that $\pi(-1 < \rho < 1) = 1 - \theta$. The posterior odds then show how the prior odds ratio $\theta/(1-\theta)$ in favour of the unit root is updated by the data.

The input of information via the prior distribution, whether deliberate or unwitting, is a major reason for potential divergence between Bayesian and classical statistical analyses. Methods of setting an objective correlative in Bayesian analysis through the use of model-based, impartial reference priors that accommodate nonstationarity are therefore of substantial interest. These were explored in Phillips (1991a), where many aspects of the subject are discussed. The subject is controversial, as the attendant commentary on that paper and the response (Phillips, 1991b) reveal. The simple example of a Gaussian autoregression with a uniform prior on the autoregressive coefficient ρ and with an error variance σ^2 that is known illustrates one central point of controversy between Bayesian and classical inference procedures. In this case, when the prior on ρ is uniform, the posterior for ρ is Gaussian and symmetric about the maximum likelihood estimate $\hat{\rho}$ (Sims and Uhlig, 1991), whereas the sampling distribution of $\hat{\rho}$ is biased downwards and skewed with a long left-hand tail. Hence, if the calculated value of $\hat{\rho}$ were found to be $\hat{\rho} = 1$, then Bayesian inference effectively assigns a 50 per cent posterior probability to stationarity $\{|\rho| < 1\}$, whereas classical methods, which take into account the substantial downward bias in the estimate $\hat{\rho}$, indicate that the true value of ρ is much more likely to be in the explosive region $\{\rho > 1\}$.

Another major point of difference is that the Bayesian posterior distribution is asymptotically Gaussian under very weak conditions, which include cases where there are unit roots ($\rho = 1$), whereas classical asymptotics for $\hat{\rho}$ are non-standard, as in (20). These differences are explored in Kim (1994), Phillips and Ploberger (1996) and Phillips (1996). The unit root case is one of very few instances where Bayesian and classical asymptotic theory differ. The reason for the difference in the unit root case is that Bayesian asymptotics rely on the local quadratic shape of the likelihood and condition on a given trajectory, whereas classical asymptotics rely on functional laws such as (9), which take into account the persistence in unit root data which manifest in the limiting trajectory.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Empirical illustrations of the use of Bayesian methods of trend determination for various macroeconomic and financial time series are given in DeJong and Whiteman (1991a; 1991b), Schotman and van Dijk (1991) and Phillips (1991a; 1992), the latter implementing an objective model-based approach. Phillips and Ploberger (1994; 1996) develop Bayes tests, including an asymptotic information criterion PIC (posterior information criterion) that extends the Schwarz (1978) criterion BIC (Bayesian information criterion) by allowing for potential nonstationarity in the data (see also Wei, 1992). This approach takes account of the fact that Bayesian time series analysis is conducted conditionally on the realized history of the process. The mathematical effect of such conditioning is to translate models such as (30) to a 'Bayes model' with time-varying and data-dependent coefficients, that is,

$$y_{t+1} = \hat{\rho}_t y_t + \sum_{1}^{k-1} \hat{\Phi}_{it} \Delta y_{t-i} + e_t, \tag{31}$$

where $(\hat{\rho}_t, \hat{\Phi}_{it}, i = 1, \ldots, k-1)$ are the latest best estimates of the coefficients from the data available to point '*t*' in the trajectory. The 'Bayes model' (31) and its probability measure can be used to construct likelihood ratio tests of hypotheses such as the unit root null $\rho = 1$, which relate to the model selection criterion PIC. Empirical illustrations of this approach are given in Phillips (1994; 1995).

Nonstationarity is certainly one of the most dominant and enduring characteristics of macroeconomic and financial time series. It therefore seems appropriate that this feature of the data be seriously addressed both in econometric methodology and in empirical practice. However, until the 1980s this was not the case. Before 1980 it was standard empirical practice in econometrics to treat observed trends as simple deterministic functions of time. Nelson and Plosser (1982) challenged this practice and showed that observed trends can be better modelled if one allows for stochastic trends even when there is some deterministic drift. Since their work there has been a continuing reappraisal of trend behaviour in economic time series and substantial development in the econometric methods of nonstationary time series. But the general conclusion that stochastic trends are present as a component of many economic and financial time series has withstood extensive empirical study.

This article has touched only a part of this large research field and traced only the main ideas involved in unit root modelling and statistical testing. This overview also does not cover the large and growing field of panel unit root testing and panel stationarity tests. The reader may consult the following review articles devoted to various aspects of the field for additional coverage and sources: (*a*) on unit roots: Phillips (1988b), Diebold and Nerlove (1990), Dolado, Jenkinson and Sosvilla-Rivero (1990), Campbell and Perron (1991), Stock (1994b), Phillips and Xiao (1998), and Byrne and Perman (2006); (*b*) on panel unit roots: Phillips and Moon (1999), Baltagi and Kao (2000), Choi (2001), Hlouskova and Wagner (2006); and (*c*) special journal issues of the *Oxford Bulletin of Economics and Statistics* (1986; 1992), *the Journal of Economic Dynamics and Control* (1988), *Advances in Econometrics* (1990), *Econometric Reviews* (1992), and *Econometric Theory* (1994).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- ARCH models
- Bayesian time series analysis
- cointegration
- Markov processes
- martingales
- present value
- statistical inference

Bibliography

Andrews, D.W.K. 1993. Exactly median-unbiased estimation of first-order autoregressive/unit root models. *Econometrica* 61, 139–66.

Andrews, D.W.K. and Guggenberger, P. 2006. Asymptotics for stationary very nearly unit root processes. Mimeo, Yale University.

Bai, J. 1997. Estimating multiple break one at a time. *Econometric Theory* 13, 315–52.

Bai, L. and Perron, P. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66, 47–78.

Bai, L. and Perron, P. 2006. Multiple structural change models: a simulation analysis. In *Econometric Theory and Practice*, ed. Corbae, B.E. Hansen and S.N. Durlauf. New York: Cambridge University Press.

Baltagi, B.H. and Kao, C. 2000. Nonstationary panels, cointegration in panels and dynamic panels: a survey. *Advances in Econometrics* 15, 7–51.

Banerjee, A., Lumsdaine, R. and Stock, J. 1992. Recursive and sequential tests of the unit root and trend break hypotheses: theory and international evidence. *Journal of Business and Economic Statistics* 10, 271–87.

Beare, B.K. 2006. Unit root testing with unstable volatility. Mimeo, Yale University.

Berkes, I. and Horváth, L. 2006. Convergence of integral functionals of stochastic processes. *Econometric Theory* 22, 304–22.

Beveridge, S. and Nelson, C.R. 1981. A new approach to decomposition of economic time series into

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
permanent and transitory components with particular attention to measurement of the ‘business cycle’. *Journal of Monetary Economics* 7, 151–74.

Bhargava, A. 1986. On the theory of testing for unit roots in observed time series. *Review of Economic Studies* 52, 369–84.

Box, G.E.P. and Jenkins, G.M. 1976. *Time Series Analysis: Forecasting and Control*, rev. edn. San Francisco: Holden Day.

Byrne, J.P. and Perman, R. 2006. Unit roots and structural breaks: a survey of the literature. Mimeo, University of Strathclyde.

Campbell, J.Y. and Perron, P. 1991. Pitfalls and opportunities: what macroeconomists should know about unit roots (with discussion). In *NBER Macroeconomics Annual 1991*, ed. O.J. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Campbell, J.Y. and Shiller, R.J. 1988. Interpreting cointegrated models. *Journal of Economic Dynamics and Control* 12, 505–22.

Cavaliere, G. 2004. Unit root tests under time-varying variances. *Econometric Reviews* 23, 259–92.

Cavaliere, G. and Taylor, A.M.R. 2007. Testing for unit roots in time series models with non-stationary volatility. *Journal of Econometrics* 140(2), 919–47.

Chan, N.H. and Wei, C.Z. 1987. Asymptotic inference for nearly nonstationary AR(1) processes. *Annals of Statistics* 15, 1050–63.

Chan, N.H. and Wei, C.Z. 1988. Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* 16, 367–401.

Choi, I. 2001. Unit roots for panel data. *Journal of International Money and Finance* 20, 249–72.

Choi, I. and Phillips, P.C.B. 1993. Testing for a unit root by frequency domain regression. *Journal of Econometrics* 59, 263–86.

Christiano, L.J. 1992. Searching for a break in GNP. *Journal of Business & Economic Statistics* 10, 237–50.

de Jong, R. 2004. Addendum to ‘asymptotics for nonlinear transformations of integrated time series’. *Econometric Theory* 20, 623–35.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
DeJong, D.N. and Whiteman, C.H. 1991a. Reconsidering trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 28, 221–54.

DeJong, D.N. and Whiteman, C.H. 1991b. The temporal stability of dividends and stock prices: evidence from the likelihood function. *American Economic Review* 81, 600–17.

Dickey, D.A. 1976. Estimation and hypothesis testing in nonstationary time series. Ph.D. thesis, Iowa State University.

Dickey, D.A. and Fuller, W.A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.

Dickey, D.A. and Fuller, W.A. 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057–72.

Diebold, F.X. and Nerlove, M. 1990. Unit roots in economic time series. *Advances in Econometrics* 8, 3–70.

Dolado, J.J., Jenkinson, T. and Sosvilla-Rivero, S. 1990. Cointegration and unit roots. *Journal of Economic Surveys* 4, 249–73.

Duffie, D. 1988. *Security Markets: Stochastic Models*. San Diego: Academic Press.

Elliott, G., Rothenberg, T.J. and Stock, J.H. 1995. Efficient tests of an autoregressive unit root. *Econometrica* 64, 813–36.

Engle, R.F. and Bollerslev, T. 1986. Modeling the persistence of conditional variances. *Econometric Reviews* 5, 1–50.

Engle, R.F. and Granger, C.W.J. 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica* 55, 251–76.

Evans, G.B.A. and Savin, N.E. 1981. Testing for unit roots: 1. *Econometrica* 49, 753–79.

Evans, G.B.A. and Savin, N.E. 1984. Testing for unit roots: 2. *Econometrica* 52, 1241–69.

Fuller, W.A. 1976. *Introduction of Statistical Time Series*. New York: Wiley.

Ghysels, E. and Osborn, D.R. 2001. *Econometric Analysis of Seasonal Time Series*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Giraitis, L. and Phillips, P.C.B. 2006. Uniform limit theory for stationary autoregression. *Journal of Time Series Analysis* 27, 51–60.

Gouriéroux, C., Phillips, P.C.B. and Yu, J. 2007. Indirect inference for dynamic panel models. *Discussion Paper No. 1550*, Cowles Foundation, Yale University.

Gouriéroux, C., Renault, E. and Touzi, N. 2000. Calibration by simulation for small sample bias correction. In *Simulation-Based Inference in Econometrics: Methods and Applications*, ed. R.S. Mariano, T. Schuermann and M. Weeks. Cambridge: Cambridge University Press.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis. *Journal of Political Economy* 86, 971–87.

Hall, A. 1989. Testing for a unit root in the presence of moving average errors. *Biometrika* 76, 49–56.

Hall, P. and Heyde, C.C. 1980. *Martingale Limit Theory and its Application*. New York: Academic Press.

Harris, D., Leybourne, S. and McCabe, B. 2007. Modified KPSS tests for near integration. *Econometric Theory* 23, 355–63.

Hlouskova, J. and Wagner, M. 2006. The performance of panel unit root and stationarity tests: results from a large scale simulation study. *Econometric Reviews* 25, 85–116.

Hong, S.H. and Phillips, P.C.B. 2005. Testing linearity in cointegrating relations with an application to purchasing power parity. Discussion Paper No. 1541, Cowles Foundation, Yale University.

Hu, L. and Phillips, P.C.B. 2004. Dynamics of the federal funds target rate: a nonstationary discrete choice approach. *Journal of Applied Econometrics* 19, 851–67.

Hylleberg, S., Engle, R.F., Granger, C.W.J. and Yoo, S. 1990. Seasonal integration and cointegration. *Journal of Econometrics* 44, 215–38.

Ibragimov, R. and Phillips, P.C.B. 2007. Regression asymptotics using martingale convergence methods. *Discussion Paper No. 1473*, Cowles Foundation, Yale University.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–54.

Kapetanios, G. 2005. Unit root testing against the alternative hypothesis of up to m structural breaks. *Journal of Time Series Analysis* 26, 123–33.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Karlsen, H.A., Myklebust, T. and Tjøstheim, D. 2007. Nonparametric estimation in a nonlinear cointegration model. *Annals of Statistics* 35(1).

Kasparis, I. 2004. Detection of functional form misspecification in cointegrating relations. Mimeo, University of Nottingham.

Kim, J.Y. 1994. Bayesian asymptotic theory in a time series model with a possible nonstationary process. *Econometric Theory* 10, 764–73.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. 1992. Testing the null of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics* 54, 159–78.

Lee, C.C. and Phillips, P.C.B. 1994. An ARMA prewhitened long-run variance estimator. Mimeo, Yale University.

Lee, J.S. 1996. On the power of stationary tests using optimal bandwidth estimates. *Economics Letters* 51, 131–7.

Leeb, H. and Potscher, B. 2001. The variance of an integrated process need not diverge to infinity and related results on partial sums of stationary processes. *Econometric Theory* 17, 671–85.

MacKinnon, J.G. 1994. Approximate asymptotic distribution functions for unit-root and cointegration tests. *Journal of Business and Economic Statistics* 12, 167–76.

Müller, U. 2005. Size and power of tests for stationarity in highly autocorrelated time series. *Journal of Econometrics* 128, 195–213.

Müller, U. and Elliott, G. 2003. Tests for unit roots and the initial condition. *Econometrica* 71, 1269–86.

Nelson, D.B. 1990. Stationarity and persistence in the GARCH (1, 1) model. *Econometric Theory* 6, 318–34.

Nelson, C.R. and Plosser, C. 1982. Trends and random walks in macroeconomic time series: some evidence and implications. *Journal of Monetary Econometrics* 10, 139–62.

Ng, S. and Perron, P. 1995. Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90, 268–81.

Ng, S. and Perron, P. 2001. Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69, 1519–54.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ouliaris, S., Park, J.Y. and Phillips, P.C.B. 1989. Testing for a unit root in the presence of a maintained trend. In *Advances in Econometrics and Modelling*, ed. B. Raj. Norwell, MA: Kluwer.

Park, J.Y. 1990. Testing for unit roots and cointegration by variable addition. *Advances in Econometrics* 8, 107–33.

Park, J.Y. and Phillips, P.C.B. 1988. Statistical inference in regressions with integrated processes: Part I. *Econometric Theory* 4, 468–97.

Park, J.Y. and Phillips, P.C.B. 1989. Statistical inference in regressions with integrated processes: Part II. *Econometric Theory* 5, 95–131.

Park, J.Y. and Phillips, P.C.B. 1999. Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory* 15, 269–98.

Park, J.Y. and Phillips, P.C.B. 2000. Nonstationary binary choice. *Econometrica* 68, 1249–80.

Park, J.Y. and Phillips, P.C.B. 2001. Nonlinear regressions with integrated time series. *Econometrica* 69, 1452–98.

Park, J.Y. and Sung, J. 1994. Testing for unit roots in models with structural change. *Econometric Theory* 10, 917–36.

Perron, P. 1989. The great crash, the oil price shock and the unit root hypothesis. *Econometrica* 57, 1361–401.

Phillips, P.C.B. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–40.

Phillips, P.C.B. 1987a. Time series regression with a unit root. *Econometrica* 55, 277–301.

Phillips, P.C.B. 1987b. Towards a unified asymptotic theory of autoregression. *Biometrika* 74, 535–47.

Phillips, P.C.B. 1988a. Regression theory for near-integrated time series. *Econometrica* 56, 1021–44.

Phillips, P.C.B. 1988b. Multiple regression with integrated processes. In *Statistical Inference from Stochastic Processes*, ed. N.U. Prabhu. *Contemporary Mathematics* 80, 79–106.

Phillips, P.C.B. 1991a. To criticize the critics: an objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics* 6, 333–64.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Phillips, P.C.B. 1991b. Bayesian routes and unit roots: de rebus prioribus semper est disputandum. *Journal of Applied Econometrics* 6, 435–74.

Phillips, P.C.B. 1992. The long-run Australian consumption function reexamined: an empirical exercise in Bayesian inference. In *Long Run Equilibrium and Macroeconomic Modelling*, ed. C. Hargreaves. Cheltenham: Edward Elgar.

Phillips, P.C.B. 1994. Model determination and macroeconomic activity. Fisher–Schultz Lecture to the European Meetings of the Econometric Society, Maastricht. Discussion Paper No. 1083, Cowles Foundation, Yale University.

Phillips, P.C.B. 1995. Bayesian model selection and prediction with empirical applications. *Journal of Econometrics* 69, 289–332.

Phillips, P.C.B. 1996. Econometric model determination. *Econometrica* 64, 763–812.

Phillips, P.C.B. 1998. New tools for understanding spurious regressions. *Econometrica* 66, 1299–326.

Phillips, P.C.B. 2001. New unit root asymptotics in the presence of deterministic trends. *Journal of Econometrics* 11, 323–53.

Phillips, P.C.B. 2006. When the tail wags the unit root limit distribution. Mimeo, Yale University.

Phillips, P.C.B. and Durlauf, S.N. 1986. Multiple time series regression with integrated processes. *Review of Economic Studies* 53, 473–96.

Phillips, P.C.B. and Han, C. 2007. Gaussian inference in AR(1) time series with or without a unit root. *Econometric Theory* 24(3).

Phillips, P.C.B. and Hansen, B.E. 1990. Statistical inference in instrumental variables regression with I (1) processes. *Review of Economic Studies* 57, 99–125.

Phillips, P.C.B. and Magdalinos, T. 2007. Limit theory for moderate deviations from a unit root. *Journal of Econometrics* 136, 115–30.

Phillips, P.C.B. and Moon, H.R. 1999. Linear regression limit theory for nonstationary panel data. *Econometrica* 67, 1057–111.

Phillips, P.C.B. and Moon, H.R. 2000. Nonstationary panel data analysis: an overview of some recent developments. *Econometric Reviews* 19, 263–86.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Phillips, P.C.B. and Ouliaris, S. 1990. Asymptotic properties of residual based tests for cointegration. *Econometrica* 58, 165–93.

Phillips, P.C.B., Park, J.Y. and Chang, Y. 2004. Nonlinear instrumental variable estimation of an autoregression. *Journal of Econometrics* 118, 219–46.

Phillips, P.C.B. and Perron, P. 1988. Testing for a unit root in time series regression. *Biometrika* 75, 335–46.

Phillips, P.C.B. and Ploberger, W. 1994. Posterior odds testing for a unit root with data-based model selection. *Econometric Theory* 10, 774–808.

Phillips, P.C.B. and Ploberger, W. 1996. An asymptotic theory of Bayesian inference for time series. *Econometrica* 64, 381–413.

Phillips, P.C.B. and Solo, V. 1992. Asymptotics for linear processes. *Annals of Statistics* 20, 971–1001.

Phillips, P.C.B. and Xiao, Z. 1998. A primer on unit root testing. *Journal of Economic Surveys* 12, 423–69.

Ploberger, W. 2004. A complete class of tests when the likelihood is locally asymptotically quadratic. *Journal of Econometrics* 118, 67–94.

Pötscher, B.M. 2004. Nonlinear functions and convergence to Brownian motion: beyond the continuous mapping theorem. *Econometric Theory* 20, 1–22.

Robinson, P.M. 1995. Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* 23, 1630–61.

Said, S.E. and Dickey, D.A. 1984. Testing for unit roots in autoregressive moving average models of unknown order. *Biometrika* 71, 599–608.

Sargan, J.D. and Bhargava, A. 1983. Testing residuals from least squares regression for being generated by the Gaussian random walk. *Econometrica* 51, 153–74.

Schmidt, P. and Phillips, P.C.B. 1992. LM tests for a unit root in the presence of deterministic trends. *Oxford Bulletin of Economics and Statistics* 54, 257–87.

Shimotsu, K. and Phillips, P.C.B. 2005. Exact local whittle estimation of fractional integration. *Annals of Statistics* 33, 1890–933.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Schotman, P. and van Dijk, H.K. 1991. A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics* 49, 195–238.

Sims, C.A. and Uhlig, H. 1991. Understanding unit rooters: a helicopter tour. *Econometrica* 59, 1591–9.

So, B.S. and Shin, D.W. 1999. Cauchy estimators for autoregressive processes with applications to unit root tests and confidence intervals. *Econometric Theory* 15, 165–76.

Solo, V. 1984. The order of differencing in ARIMA models. *Journal of the American Statistical Association* 79, 916–21.

Stock, J.H. 1991. Confidence intervals for the largest autoregressive root in US macroeconomic time series. *Journal of Monetary Economics* 28, 435–59.

Stock, J. 1994a. Deciding between I(1) and I(0). *Journal of Econometrics* 63, 105–31.

Stock, J.H. 1994b. Unit roots, structural breaks and trends. In *Handbook of Econometrics*, vol. 4, ed. R. F. Engle and D. McFadden. Amsterdam: North-Holland.

Stock, J.H. 1999. A class of tests for integration and cointegration. In *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, ed. R.F. Engle and H. White. Oxford: Oxford University Press.

Stock, J.H. and Watson, M.W. 1988. Variable trends in economic time series. *Journal of Economic Perspectives* 2(3), 147–74.

Sul, D., Phillips, P.C.B. and Choi, C.-Y. 2006. Prewhitening bias in HAC estimation. *Oxford Bulletin of Economics and Statistics* 67, 517–46.

Wang, Q. and Phillips, P.C.B. 2006. Asymptotic theory for local time density estimation and nonparametric cointegrating regression. Discussion Paper no. 1594, Cowles Foundation, Yale University.

Wei, C.Z. 1992. On predictive least squares principles. *Annals of Statistics* 20, 1–42.

Xiao, Z. and Phillips, P.C.B. 1998. An ADF coefficient test for a unit root in ARMA models of unknown order with empirical applications to the U.S. economy. *Econometrics Journal* 1, 27–43.

Zivot, E. and Andrews, D.W.K. 1992. Further evidence on the great crash, the oil price shock and the unit root hypothesis. *Journal of Business and Economic Statistics* 10, 251–70.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Phillips, Peter C. B. "unit roots." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000024> doi:10.1057/9780230226203.1764

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

United States, economics in (1776 – 1885)

Stephen Meardon
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economics in the United States before 1885 was not a discipline of credentialled professionals. Its debates were published in political pamphlets and newspaper editorials as well as college textbooks and scholarly treatises. Its principal project was to set the boundaries of a doctrinal system of economic liberalism and determine the appropriate functions of government. The system was characterized by the sanctity of private property, the celebration of individual labour, and the harmony of the economic and moral orders. Those parts of the system that were imported from abroad were adapted, sometimes ingeniously, to economic and political circumstances at home.

Keywords

American Economic Association; American Free Trade League; banknotes; Bastiat, F.; Bryant, W. C.; Cardozo, J. N.; Carey, H. C.; Carey, M.; catallactics; combinations; economic liberalism; Ely, R. T.; Enlightenment; Finney, C.; Franklin, B.; free banking; free labour doctrine; free trade; George, H.; Great Awakening; Hamilton, A.; James, E.; Jefferson, T.; land tax; Leavitt, J.; Liberty Party; List, F.; Madison, Bishop J.; Malthus's theory of population; McCulloch, J. R.; Physiocracy; Paine, T.; paper money; Patten, S. N.; Perry, A. L.; Petty, W.; political economy; protectionism; public works; Raymond, D.; rent; Ricardo, D.; Say, J.-B.; Second Great Awakening; Slavery; Smith, A.; Tucker, G.; Turgot, A.R. J.; United States, economics in; Vattel, E.; Wayland, F.; wealth; Wells, D. A.

Article

Pre-independence to 1820

Before US independence, the middle Atlantic colonies offered more fertile ground than New England or the South for ideas of political economy from Britain and the European Continent to take root. Eighteenth-century Philadelphia was rivalled only by New York and Boston in its population, which

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
grew from over 4,000 in 1700 to over 28,000 in 1790. It was unrivalled in its cosmopolitanism. The religious diversity of Pennsylvania promoted tolerance of notions that, especially from the perspective of the New England theocracy, were tainted by deism or secularism. One consequence was a thriving press. Philadelphia, therefore, was the destination of Benjamin Franklin (1706–1790) when he ran away from his printer's apprenticeship in Boston at the age of 17. The many ideas he published in his new situation included his adaptation of the work of Sir William Petty to American commercial policy. Franklin's *A Modest Enquiry Into the Nature and Necessity of a Paper Currency* (1729), following Petty's *Discourse on Political Arithmetic* (1690), emphasized the importance of the quantity of specie in circulation to commercial prosperity, and a favourable balance of trade to the quantity of specie. Beyond Petty, Franklin argued that issuance of paper money would at once substitute for the specie that was then lacking and, by stimulating production and exports, improve the balance of trade – and thereby garner more specie. Paper money, though commonly disdained, could effect a virtuous circle of commercial vitality and specie accumulation.

In 1757, when Franklin was renowned in America and abroad as a scientist and statesman, he was appointed the Pennsylvania Assembly's emissary to the Crown. The governor of the colony had vetoed the issuance of paper money and other fiscal measures desired by the colonists; Franklin was to convey their complaints. Following the onerous internal taxes established by the Stamp Act of 1765 and the import duties of the Townshend Acts of 1767, they would have more cause to complain. The popular response was ‘No taxation without representation’ coupled with the colonists’ mutual agreement to refuse the importation of British goods. Franklin's own response, more nuanced but to the same effect, drew upon the Physiocratic notion of the pre-eminence of agriculture that had been impressed upon him in France by Anne Robert Jacques Turgot. In his *Positions to be Examined, concerning National Wealth* (1769), Franklin explained that trade is fair when both parties know the value of the traded goods, meaning the value of labour embodied in them; and value is most knowable when the goods consist of agriculture, not manufactures, because agriculture is derived more directly from labour. The implication was that by creating a captive market in America for British manufactured goods, Britain was attempting to hoodwink the colonists. The colonists’ eschewal of legal British trade and their substitution of smuggled foreign goods, therefore, were justified. Besides, Franklin added, the frugality necessitated by the boycott would make domestic farmers and tradesmen more productive (Dorfman, 1966, vol. 1, 191–3).

While Franklin engaged the political-economic ideas of the Enlightenment in public argument, he also sought to disseminate them in higher education. His Academy of Philadelphia, later to become the University of Pennsylvania, commenced classes in 1751 with a non-sectarian board of trustees, a progressive Scottish cleric as president, and a curriculum including governance and commerce. The institution's secular orientation and curriculum stood in contrast to those of Harvard (1636), the College of William and Mary (1693), and Yale (1701), which were aligned closely with the Congregational or Anglican churches, directed to the training of ministers and missionaries, and uncongenial to the innovations in moral philosophy and legal and commercial thought from Britain and Europe. Even Princeton (1746), Columbia (1754), and Brown (1764), which were intended to embody a new spirit of religious tolerance, succumbed ultimately to a more constrained orthodoxy. Indeed, so too did Franklin's Academy. But the Revolutionary war shook the orthodoxy. Franklin's curricular plan made inroads at William and Mary in 1779 (O’Connor, 1944, pp. 64–6). Bishop James Madison (1749–1812), the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
College's president, cousin of the fourth US president, and a man of radical political views notwithstanding his prominence in the Episcopal Church, became America's first teacher of political economy (O'Connor, 1944, pp. 20–1).

Madison is believed to have relied upon Emer de Vattel's *The Law of Nations, or the Principles of Natural Law Applied to the Conduct and to the Affairs of Nations and of Sovereigns* (1758). The book stood out for its maxims of sovereignty more than commercial policy. To Vattel, the sovereignty of the state is inalienable. While in some times and places the head of state may be a prince, nevertheless, 'the State is not, and can not be, a patrimony' (1758, pp. 29, 33). The sovereign is obliged to serve the state's interests, not his own. Although they are peripheral, the book also includes chapters on general commerce, the financing of public roads and canals, money and exchange, and commerce between nations. Vattel makes brief but notable mention of the desirability of defraying the expense of roads and canals from the general revenue of the entire nation, and of the ruler's obligation to encourage trade that is beneficial but to restrict that which is hurtful. In the latter category is that 'ruinous' trade which results in a negative balance, entailing more gold and silver leaving the country than entering it (1758, pp. 44, 43).

In Vattel's work, the political thought of the enlightenment is bonded to mercantile economic thought. Neither was particularly avant-garde for the 1770s. Thomas Paine (1737–1809) did not steel the colonists' patriotic will with 25 editions and 150,000 copies of his *Common Sense* (1776) by arguing merely that King George had not fulfilled adequately a sovereign's obligations to his subjects. Paine remonstrated that monarchy yields tyranny and war; only republican government would serve the cause of liberty. The differences of Paine from Vattel in political economy are equally significant. 'Our plan is commerce,' Paine declared. It was no time to ponder which particular avenues and products and circumstances of trade were hurtful and which were not; it was time to throw off the commercial shackles of Britain altogether and trade with all the world. 'Trade flourishes best when it is free', he continued a short while later, 'and it is weak policy to attempt to fetter it' (Paine, 1778, p. 153).

Paine wrote polemics for the literate multitudes. His work was not the stuff of college curricula; Adam Smith's *Wealth of Nations* was fitter for that purpose. Smith exploded the 'absurd' balance-of-trade doctrine (Smith, 1789, p. 456), illuminating in more scholarly fashion than did Paine the virtues of commercial freedom, whether international or domestic, and nudging England towards the 'natural system of perfect liberty and justice' in the administration of its colonial trade (1789, p. 572). Regarding roads and canals and other means of facilitating commerce, Smith was disinclined to see them financed with the general revenue of the state. More consistent with his system was their financing by tolls or by local or provincial taxes (1789, pp. 682–3). The *Wealth of Nations* was taken up at William and Mary sometime between the mid-1780s and the 1790s. It was read, debated, and absorbed in America several years earlier, however: an edition was published in Philadelphia within a few years of the Revolution (O'Connor, 1944, pp. 21–2).

Questions of foreign trade policy, and the role of the federal government in undertaking public works and institutions for promoting commerce, dominated the economic debate of the early republic. They were argued passionately because it was believed that upon their answers depended not only the nation's prosperity, but its character and survival. Thomas Jefferson (1743–1826) sought free trade between America and Europe less for the goods that Americans could obtain than for the occupations they would undertake in consequence. In his *Notes on the State of Virginia* (1785), Jefferson observed that, if trade

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
had not been interrupted and manufactures stimulated by the Revolutionary War, Americans in greater numbers would work the land and buy their textiles from abroad. The advantage of making a living from the land lay in the republican virtues that husbandry instilled in the smallholder. ‘Let our workshops remain in Europe’, Jefferson offered; ‘the loss by the transportation of commodities across the Atlantic will be made up in happiness and permanence of government’ (Spiegel, 1960, pp. 42–3).

Alexander Hamilton (1755–1804), Jefferson's ideological rival, did not repudiate the Virginian's vision of the free trade of American agriculture for European manufactures. He just denied that, under the circumstances, it was realizable. The European powers regularly imposed barriers to the importation of American products, and the several American states had little power to persuade or compel them to do otherwise. Hamilton's answer was to change the circumstances of American government. The unlikelihood of the Jeffersonian ideal became an argument for a strong central authority. In *Federalist* No. 11 (1788), Hamilton argued that adoption of the Constitution would create a central government with the power to bargain for commercial privileges abroad. If instead disunion persisted, ‘we should then be compelled to content ourselves with the first price of our commodities, and to see the profits of our trade snatched from us to enrich our enemies and persecutors.’

Hamilton envisioned a larger manufacturing sector than did Jefferson, but he did not admit that the growth of manufactures would retard agriculture. In his famous *Report on Manufactures* (1791), composed during his term as the nation's first Secretary of the Treasury, he offered two reasons. First, even if manufactures slowed the extensive cultivation of land, they would ‘promote a more steady and vigorous cultivation of the lands occupied’. Second, manufactures probably would not slow the extensive cultivation of land, anyway. Immigrants drawn to America by manufacturing jobs would eventually leave them for agriculture, attracted by the promise of independent proprietorship of land (Spiegel, 1960, p. 33).

The question that followed was, ‘Will manufacture develop by itself?’ Hamilton thought not. The scarcity of labour, high wages, and lack of capital conspired against American manufacturers. To survive against European competition would be difficult – but not hopeless. Raw materials were more abundant in the United States and, although wages were indeed higher, the greater expense of labour was counterbalanced by the expenses of transportation and customs duties for those who would import foreign goods. The scarcity of capital could be addressed by fostering a banking system and by the Treasury's issuance of debt, which would serve as a form of money. To Hamilton, as to Paine, the business of America was commerce; but Hamilton was not as squeamish as Paine and his ilk, including Jefferson, about using the powers of the newly constituted federal government to build the institutions and infrastructure of commerce and to encourage industries that might otherwise languish.

Hamilton's name was invoked often in ensuing controversies that exercised Americans and their representatives. Among them were questions of the federal government's role in financing ‘internal improvements’, including the national turnpike stretching from the Potomac to the Ohio River, authorized in 1806; the First Bank of the United States, the charter of which expired in 1811; the Second Bank of the United States, the fate of which was decided in the presidential election of 1832; and the ‘American System’ of tariff protection for domestic manufactures, which sparked debate through most of the 19th century. One of the many whom Hamilton inspired was Daniel Raymond (1786–1849), who became in 1820 the first American author of a systematic treatise on political economy.

From 1820 to the Civil War

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Raymond penned his *Thoughts on Political Economy* in his free time as an underemployed Baltimore lawyer (Spiegel, 1960, p. 55). His careful definition of national wealth as ‘a capacity for acquiring the necessaries and comforts of life’, not the value of tangible assets, was intended to combat the idea that the United States was wealthy by virtue mainly of its climate, soil, and nearly limitless territory. Raymond argued that other things mattered too, and more: population density, the distribution of private wealth, the capital and technology applied to agriculture, transport, and other industries, and above all ‘the industrious habits of the people’ (Spiegel, 1960, p. 60). The list corresponded closely to the priorities for government action envisioned by Hamilton.

Raymond's distinction in American economics is that he was first in a class, not that he was influential. More credit for influence is due to Mathew Carey (1760–1839), an Irish publisher and controversialist who emigrated to Philadelphia in 1784 and soon re-established himself in the same vocations. On political economy, he wrote prolifically – but not systematically in the manner of Raymond, whose work he sponsored (Spiegel, 1960, p. 55). Carey was concerned mainly with one issue: the tariff. His *Addresses of the Philadelphia Society for the Promotion of National Industry* (1819) responded to the country's financial crisis of 1819 with a call for further tariff protection for domestic industry. Although the tariff bill of 1820 did not pass, the protectionists’ champion in Congress, Henry Clay, acknowledged Carey's importance to the ferment of the tariff of 1824 (Dorfman, 1966, vol. 1, pp. 384, 389–90). Carey's influence was also exercised importantly, if indirectly, through the writings of Friedrich List (1789–1846), the German thinker who was renowned posthumously in his country and abroad for his protectionist ‘national system’ of political economy. List devised his system during a five-year sojourn in Pennsylvania from 1825 to 1830, in which time he became acquainted with Carey's ideas, lent them his own voice and authority, and propagated them through the protectionist periodical, the *National Gazette* (Dorfman, 1966, vol. 2, p. 577).

The protectionist trend in trade policy that gathered strength in the mid-Atlantic states caused alarm in the South. Jacob Newton Cardozo (1786–1873), publisher of Charleston's *Southern Patriot*, addressed the tariff and other questions in light of David Ricardo's system. An American edition of Ricardo's *Principles of Political Economy and Taxation* was published in 1819; another avenue of his large influence was the *Encyclopædia Britannica* article on ‘Political Economy’ by John Ramsay McCulloch, Ricardo's chief expositor in Britain. The article was republished in 1825 with introduction, summary, and annotations by Columbia College's Reverend John McVickar (O’Connor, 1944, p. 136).

Cardozo's *Notes on Political Economy* appeared in 1826. To Cardozo, the unenlightened social arrangements of Europe – above all, impediments to the alienation of land – had obscured to Ricardo what was evident in the United States. In the absence of unwise customs and legislation, rent was never pernicious. Indeed, some payments that were nominally rent, but were really the returns to the ingenuity of the cultivator, were positively benign. Conversely, where rent was pernicious, it was due to legislation that fomented, in one way or another, ‘monopoly’. In the United States, such legislation was the protective tariff. The genius of Cardozo's argument was that indicted protectionism even more forcefully than did Ricardo's system, thereby serving the interests of the cotton-exporting South and the entrepôt of Charleston, while casting Southern landowners in a productive rather than a parasitic role (Dorfman, 1966, vol. 1, pp. 551–66).

Bank regulation and the sustenance of slavery preoccupied the South as much as the bane of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
protectionism. The question of greater urgency varied with the occasion. The 1828 ‘Tariff of Abominations’ and the crisis stemming from South Carolina's ‘nullification’ of the law in 1832 were occasions of substantive political–economic debate – but so was President Jackson's veto, in 1832, of the proposed re-charter of the Second Bank of the United States. While the Democratic and Whig parties agreed that the currency should be convertible for gold, there were large differences between and even within them regarding the appropriate degrees competition and regulation of banks, especially those involved in issuing currency. They also differed on the function of the Second Bank of the United States as competitor, usurper, and regulator in the domain of private and state banks. Jackson's objection to the Second Bank of the United States was that its shareholders enjoyed a monopoly of the Federal government's bank transactions. Cardozo shared Jackson's anti-monopoly convictions, but appreciated the Bank's potential as a regulator of the note issuance of other banks, and thus of the recurring cycles of credit expansion, inflation, and contraction. Banks of deposit and discount – ‘banking in its legitimate meaning’ – should be subject to free competition, according to Cardozo, but banks of issue should be subject to severe restriction (Leiman, 1966, p. 112). This exception to perfect freedom served, by means of a stable currency, to protect the value of property and promote trade. Cardozo's justified similarly his support for slavery. Abolitionism was a ‘conspiracy against property’, while slavery heightened the United States’ comparative advantage in agriculture and its gains from free trade (Leiman, 1966, p. 176–7).

By the 1820s and 30s, the inviolability of property and free trade were touchstones of economic thought not only in the South. In the North, too, notwithstanding its traditional allegiance to Hamiltonian federalism, the ideas had wide currency. Jefferson's Embargo Act of 1807, which prohibited exports in the (disappointed) expectation of compelling Britain and France to rescind their restraints on American trade, provoked ferocious opposition in the Northeast. Ideas of free trade, and economic liberalism more generally, sank their roots, notwithstanding the irony of their having been planted in opposition to one of their leading advocates. William Cullen Bryant (1794–1878), who began in 1826 his life's work as editor of the *New York Evening Post*, and concurrently one of the nation's most renowned poets and most prominent spokesmen for unfettered commercial freedom, cut his literary and economic teeth during the embargo. The first poem he published as a boy in rural Massachusetts was a widely circulated satire on the subject. ‘In vain Mechanics ply their curious art, / And bootless mourn the interdicted mart’, complained the 13-year-old Bryant. Jefferson's diplomacy, he mocked, ‘His grand ‘*restrictive energies*’ employs, / And wisely regulating trade – destroys’ (Bryant, 1808).

To gain acceptance in Northern colleges, however, economic liberalism had to be disassociated from French political radicalism and garbed in religious piety. The first event was already accomplished in the mid-1790s, as the enthusiasm of American partisans of the cause of the French Revolution diminished. The second was the outcome of a longer process that started at the same time and gathered momentum as political economy made its way into college curricula as a subject in its own right. Most colleges introduced political economy as such, and explicitly, in the 1820s: although Harvard, Columbia, and Princeton did so between 1817 and 1819, they were followed by Dickinson, Pittsburgh, Bowdoin, Amherst, Yale, Rutgers, the US Military Academy, Geneva (later renamed Hobart), Williams, Dartmouth, Brown, Union, and Hamilton, all in the years 1822 to 1827 (O’Connor, 1944, p. 100). In most cases the favoured text was Jean Baptiste Say's *Treatise on Political Economy*. Because Say disapproved of several of the social changes in France since the Revolution, he passed the political test. Because he was a protestant Huguenot, he passed the religious one – although it was a close call

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(O'Connor, 1944, pp. 120, 133–4).

The religious test became more stringent in the 1830s. Religious revivals had recurred with varying intensity since the Great Awakening of the 1730s to 1750s, but none since had swept through the country with as much intensity as those commencing in western New York in 1826 and spreading outward, nationwide, through roughly the decade that followed (Cole, 1954, pp. 75–6). 'Harvests' of new souls were most bountiful among the middle class and young in small towns and rural areas (Cole, 1954, p. 80). Colleges were especially fertile ground: the demographics were right, and in most cases so too was the geography. College presidents, besides, saw revivals as a way of boosting their institutions' visibility and prestige. Students and faculty were particularly amenable to the particular character of the Second Great Awakening. In past eras evangelism sought the salvation of the individual; now it sought that of the nation, even the world. The most popular evangelist of the period, Charles Finney, called upon his listeners to strive for temperance, moral reform in general, and the abolition of slavery (Cole, 1954, p. 77). The last of these causes, even within the flock, was controversial. That the awakened moral sensibilities demanded an end to slavery, there was consensus; that they required the North to impose abolition upon an unwilling South, there was not. As evangelical religion asserted itself in the same citadels where political economy was establishing its foothold, the subject required a text reflecting at least the sensibilities, if not always the objectives, of its teachers and students.

That text was *Elements of Political Economy* (1837) by Reverend Francis Wayland (1796–1865), president of Brown University. Wayland, who had already published his *Elements of Moral Science* in 1835, was the exemplary teacher of political economy for two generations of antebellum men. Political economy partook of moral philosophy; the two subjects were taught proximately, but separately, in the students' junior or senior year. In the *Moral Science* students learned the nature of their moral constitution, including the power and limits of their unaided conscience to discern right from wrong in their private and public conduct. Inasmuch as conscience was limited, there was recourse first to natural religion and finally to revealed religion. Revealed religion was the study of God's word as manifested in the Bible; natural religion was the study of God's design as manifested in the consequences of human behaviour. Taking intemperance as an example, Wayland illustrated that one could survey the vice and poverty that resulted from the consumption of liquor, consider that God has a design relating all causes to their effects, and thereby infer that God's design forbids intemperance 'as though He had said so by a voice from heaven' (Wayland, 1837, p. 120). Of course, one could study other sorts of behaviour and their consequences – production, exchange, distribution, and individual and public consumption – and infer God's design therein, too. To do so required another volume: that was the thrust of Wayland's *Political Economy*.

Wayland taught that the economic role of government is 'to construct the arrangements of society as to give free scope to the laws of Divine Providence' (O'Connor, 1944, p. 189). Man must be free to produce what he will and dispose of the product as he pleases. International trade should be unencumbered by protective tariffs. Although the incorporation of banks should be made conditional on arrangements to ensure the convertibility of their notes into specie, nobody who is willing to accept the conditions should be refused a bank charter. Poor laws supporting those capable of work were founded on the premise the rich were obliged to support the poor by law instead of by charity, and as such were a violation of property. Combinations of workers, like combinations of capital owners, were oppressive, even tyrannical. All that man required was the right to sell his labour and invest his capital freely.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Legislation granting one party or the other any additional privileges was ‘impolicy’ (Wayland, 1841, pp. 108–23).

As a treatise in political economy, Wayland's book was rivalled, after 1848, by John Stuart Mill's *Principles of Political Economy*. Mill conveyed the requisite ethical tone, if not a religious one, and allowed sufficient qualifications to the cases for free trade and strict constraints on banknote issuance to appeal to a broader swathe of political opinion than did Wayland. But Mill's book was too lengthy to serve as a college text (Dorfman, 1966, vol. 2, p. 710). In that role, Wayland's was unrivalled. The Elements of Political Economy went through no fewer than 24 printings in Boston and New York between 1837 and 1875, with three more in London and a translation to Hawaiian (O’Connor, 1944, p. 174).

For his large readership and his tightly connected system of economic liberty, Wayland has been designated the ‘Ricardo of evangelists’ (Cole, 1954, p. 178). The designation is not, notably, the ‘evangelist of Ricardo’. The last would not apply, for in one crucial respect Wayland was not the system's most fervent expositor. Those who adhered to the doctrine that man's liberty to reap the fruits of his own industry was sacrosanct, and who followed its implications wherever they might lead, found their way in short order to the question of slavery and determined that it could not be countenanced. ‘Free labor’ was their slogan. Wayland did not travel with them so far: he was silent on the slavery question for the better part of a generation. He believed that the Constitution left slavery a matter for states to decide (Dorfman, 1966, vol. 1, p. 760) – but even as a matter of state law, Wayland was not exercised by the subject.

The Reverend Joshua Leavitt (1794–1873) proselytized a more thoroughgoing interpretation of the free labour doctrine. In the same year that Wayland first published the *Elements of Political Economy*, Leavitt became editor of the *Emancipator*, an organ urging free trade, cheap postage, temperance, and above all, the demise of slavery by political action. The *Emancipator* also advocated the election of James G. Birney of the newly formed Liberty Party as President of the United States in 1840 and 1844 (Cole, 1954, p. 40). Although the Liberty Party was ineffectual and Birney unsuccessful, Leavitt's periodical and the presidential campaigns it championed marked the beginnings of a longer effort. Its objectives were at once to intensify opposition to slavery and to peel the opponents’ support away from the dominant Whig and Democratic parties.

The effort began to show effects once the Liberty Party was eclipsed by the Free Soil Party in 1848 and 1852. Northern ‘Barnburner’ Democrats, who sided with the majority of their party in opposition to the protective tariff, national bank, and internal improvements, but disfavoured its pro-slavery orientation, drifted towards the Free Soilers. So did radical ‘Conscience Whigs’, who could not abide their own party's concessions to the Slave Power. In the run-up to the election of 1852, while Democrats were able temporarily to quell their internal conflict, the cleavage among Whigs widened. The radicals demanded political abolition; the conservative ‘Cotton Whigs’, who represented the party's manufacturing constituency, were loath to disrupt commerce (and the Northern mills’ supply of southern cotton) by fuelling the sectional conflict. The rancour contributed to the Whigs’ loss of the presidency. At that moment, the urgency for conservatives to answer the articulation of the free labour doctrine to political abolition was manifest.

The task was delicate: the Whig party cast itself as a friend of labour and opponent of slavery. The tariff, a central plank in its platform, had long been promoted by Whig icons such as Henry Clay and Daniel

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Webster as a way of arming workers in ‘an unequal contest with the pauper labor of Europe’ (Eckes, 1995, p. 24). Advocacy of labour did not require compromise of the interests of manufacturers: the party line maintained that in general, as in tariff legislation, the interests of labour and capital were in harmony (Foner, 1970, pp. 20–1). Nor did opposition to slavery require concrete support for its abolition: inheritors of the mantle of Clay and Webster detested slavery but did not see fit to involve the federal government in the fundamental question of its existence, except in new states and territories where prior law did not apply. In order for the conservatives to save the Whig party without sacrificing their convictions, the free labour doctrine and the presumption of harmony of interests had to be retained, but both had to be interpreted to support the protective tariff and to oppose political abolition.

The task was taken up by 19th-century America's most original economic thinker, Henry Charles Carey (1793–1879), son of Mathew Carey. By 1852 the younger Carey was already an important framer of Whig doctrine. The title of his first major book, *The Harmony of Nature as Exhibited in the Laws which Regulate the Increase of Population and of the Means of Subsistence: and in the Identity of the Interests of the Sovereign and the Subject; the Landlord and the Tenant; the Capitalist and the Workman; the Planter and the Slave* (1836) suffices to describe its contents. His second, a two-volume *Principles of Political Economy* (1840), developed more fully his system: its centrepiece was a law of the progression of civilization that entailed a new refutation of the Ricardian theory of rent. His third, The Past, the Present, and the Future (1848), demonstrated that the same law prescribed tariff protection for the United States – an argument that he propounded relentlessly to the end of his life. What remained was to show what the law implied for slavery.

That Carey did in *The Slave Trade, Domestic and Foreign: Why it Exists and How it May be Extinguished* (1853). For the unacquainted reader he reiterated the law of progression. Contrary to Ricardo's teaching, Carey showed that the land that is cultivated first is not the most fertile, which is at the bottom of river valleys where the vegetation is dense. The first settlers to an area do not have the numbers or the technology to clear and till the bottom land: they settle instead at higher altitudes and cultivate less fertile land. As they eke out a living and their numbers grow, they need not confine their work to agriculture. Their employments diversify, and some produce manufactures. Manufacturers devise machines and techniques that are useful in agriculture, allowing the population to inhabit and cultivate more fertile land. They move down the hillside; they produce more food; their numbers grow, and they diversify further their employments; their manufacturers invent better machines and techniques; they move further down the hillside and cultivate better land; and so on. Thus progress requires the development of technology that passes into agricultural use, whether by design or happenstance. Development of technology depends on the diversification of industry – or as Carey put it more vividly, ‘the natural tendency of the loom and the anvil to seek to take their place by the side of the plough and the harrow’ (Carey, 1853, p. 50).

The natural tendency is interrupted at a country's peril. The United States inflicted upon itself such an interruption, according to Carey, in so far as it relaxed its impediments to foreign trade, impelling settlers to go West and farm more extensively for export rather than planting roots and farming more intensively for the home market. But that was not the only consequence of excessive foreign trade. Amongst a geographically dispersed population with few and small population centres, land tenure was characterized by larger parcels, which were more congenial to the cultivation of crops by slave labour, and correspondingly the demand for free labour was small (Carey, 1853, p. 51).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The extinction of slavery, then, was not to be achieved by political abolition. The sudden reversal of the relationship between master and slave would produce indolence in the latter and cause the ruin of both (Carey, 1853, p. 23–4). It was to be achieved by abolishing the economic conditions that fostered slavery. The conditions were reducible to the small demand for free labour. To increase the demand, it was necessary to ensure that land was cultivated intensively, in small plots, and for sale in a home market that also comprised manufacturers. To ensure that outcome, it was necessary to restrain foreign trade with a protective tariff. Only with the traditional programme of the Whig Party, not the radical one of the Free Soil Party, would both master and slave be set free: the master from his dependence on distant markets, the slave from his shackles.

Carey's system ultimately failed to be the cement to hold together the Whig Party, but not because it was unpersuasive under the circumstances when he wrote it. Circumstances changed. The Missouri Compromise of 1820 had prohibited the future admission of slave states above latitude 36°30′. Because the compromise rendered the slavery question irrelevant in most of the territory into which the United States would expand, moderate opponents of slavery were willing to leave the question up to state legislatures in the remaining areas. At least one professed opponent of slavery even argued, paradoxically, for its *expansion* into those areas and beyond. George Tucker (1775–1861), who introduced political economy to curriculum of the University of Virginia in 1825, had earlier served his district in the US House. There, in 1820, he borrowed Malthusian population theory to demonstrate that as the nation expanded westward the means of subsistence would increase, so the number of slaves could be expected to grow. If slaves were confined to the South while only whites migrated to the new areas, the stage would be set for a violent confrontation. The slaves, growing more populous relative to whites in the South if not elsewhere, would rise up against their masters. If slavery were allowed instead to spread westward, then the value of white labour would fall, the value of slave labour would follow, the price of slaves would fall below the cost of their maintenance, and the slaves would eventually be freed voluntarily (Dorfman, 1966, vol. 2, p. 544). Carey's *The Slave Trade* was but a more comprehensive (and more plausible) justification for maintaining the same inclination manifested for so long by Tucker and many others: to oppose slavery while conciliating slaveholders. At last, just one year after the publication of Carey's tract, the position was clearly untenable. The Kansas–Nebraska Act, signed into law by President Franklin Pierce in 1854, allowed the slavery question to be put to a vote in the two newly organized territories, both of which were north of 36°30′, and to admit them to the Union as slave states or free according to the voters’ wishes. In effect, the Act repealed the Missouri Compromise; and Tucker's decades-old notion that such a result would promote, not for ever postpone, slavery's demise could no longer be believed. Erstwhile conservative Whigs, with Carey in the front ranks, dissolved their party and joined the Conscience Whigs and Barnburner Democrats to organize the Republican Party.

Yet Carey's free-labour protectionism had become obsolete only in part. Former protectionist Whigs dominated the new Republican Party in financial matters, including tariff policy. After the outbreak of war, they instituted an irredeemable currency and a succession of tariff increases, mainly for war finance. Those who understood the free labour doctrine to have the opposite implications for money and tariffs did not oppose the new legislation: considering the stakes, they acknowledged its expediency. But this marriage of convenience between adherents to otherwise rival political–economic doctrines lasted only as long as the war.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The slavery debate was settled with the war's end, but the money and tariff controversies erupted anew. Popular discontentment was aroused immediately by the myriad internal and external war taxes. The Republican Congress established a revenue commission to study the tax system and recommend an overhaul; its chairman (and after the first year, the sole commissioner) was David A. Wells (1828–1898), known to be a disciple of Henry Carey. From 1865 through 1867, in the first three of his five annual reports, Wells took care not to antagonize his mentor or his patrons. The dominant view in the Republican Party was that tariffs, necessary for revenue in wartime, were opportune for protection in peacetime. Wells recommended accordingly that the internal excises should be dismantled, but the tariffs, which now yielded revenues of over 45 per cent of the value of imports, should be maintained. Wells's report for 1868 marked a change of his thinking, and more. Favourable and unfavourable reviewers alike read it as an insider's repudiation of the American System. By attempting ‘indiscriminate or universal protection’, Wells determined, the protective tariff rendered ‘all protection a nullity’ because the iron and steel industries’ output was the textile manufactures’ input (Wells, 1869, p. 3). The problem could not be solved simply by raising further the textile tariff: any modification of the tariff law would incite a general scramble for more protection. Given the political reality, Wells reasoned that it was better to determine tariffs by a simple and invariant principle than to attempt ad hoc changes. The principle was a tariff for revenue, not protection. Wells's stand for it, which only hardened in his reports for 1869 and 1870, widened the divide within the Republican coalition. Arthur Latham Perry (1830–1905) of Williams College, an evangelist for free trade in the mould of Leavitt, congratulated Wells for having written ‘our Bible in onslaughts against the monopolists’; Henry Carey published a missive likening Wells to Judas Iscariot.

Although Wells arrived at the revenue tariff position by way of expediency, he occupied it thereafter as a sincere and doctrinaire exponent of free trade. There he joined William Cullen Bryant, who appended in 1866 the presidency of the American Free Trade League (AFTL) to his duties as newspaper editor and publisher; and Perry, who barnstormed for the AFTL after completing in the same year his *Elements of Political Economy*, the successor to Wayland's work as the principal American textbook on the subject. Bryant, a septuagenarian at the war's end, was of an older generation whose powers were waning; Wells and Perry, in their thirties, were more representative of the spirit of the post-bellum decade. Free trade was the cause that they championed most actively, but they were led to it by attitudes and methods of broader significance.

Both Wells and Perry were schooled in the clerical system of Wayland and were moved by the evangelical impulse of reform, but they were also young enough to partake of the fascination with science that characterized the American mind of the 1840s and 1850s. The fascination was stoked by the appearance of Darwin's *On the Origin of Species* in 1859, but, like the inquiry into biological adaptation to which Darwin contributed, was present much earlier. The famed Swiss geologist and zoologist Louis Agassiz arrived in the United States in 1846, drew thousands to his public lectures in Boston, and inspired textile magnate Abbott Lawrence to establish a scientific school at Harvard for him to lead. Wells was among the first four graduates of the Lawrence Scientific School. There he learned the talents of empirical observation and statistical argument that he displayed as Commissioner of the Revenue. Perry's scientific inclinations were manifested differently from Wells's but were consonant with them. Perry himself served up few statistics, but he sought to purge political economy of metaphysical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
presumptions that were unanswerable by statistical measurement. The goal could be achieved, he thought, by circumscribing the science, which had been stymied by preoccupation with ‘wealth’. Because the word admitted so many meanings, none of them precise, whatever thing it named was hard to measure. Because the thing was hard to measure, the word was ‘the bog whence most of the mists have arisen which have beclouded the whole subject’ and was ‘totally unfit for any scientific purpose whatever’ (Perry, 1866, p. 29). A better word than ‘wealth’ was ‘value’. Value was no sooner determined than it was measurable; one had only to relinquish any hope of finding an invariant standard of it. The value of a thing was always relative to the other thing for which it was exchanged, and was determined only when the exchange was made. The amount of money exchanged depended on the exchanging parties’ estimates of their desires and the efforts required for their satisfaction. Refashioned thus as the science of exchanges, political economy would concern itself only with things that are measurable in units of money. ‘Catallactics’, or perhaps ‘economics’, was the more accurate name for such a science, Perry allowed – but in this particular question of terminology he was less fastidious. Regardless of the science's name, once metaphysics was ostensibly exiled from it, one groped with difficulty within its scope for justification of government constraints of exchange. Therein lay Perry's enthusiasm for free trade, his unsurpassed appreciation of Frédéric Bastiat, and his more general presumption in opposition to commercial legislation, foreign or domestic.

While Wells's affinity for statistics and Perry's redefinition of political economy reflected the scientific aspirations of the post-bellum generation, the American Social Science Association (ASSA) reflected the union of those aspirations with the generation's reformist impulse. The impulse drew urgency from the tremendous economic transformation that the war had merely paused, not reversed. Fewer than 30 railroad miles were in operation in the United States in 1830; in 1860, the number exceeded 30,000. The railroad hastened migration from the eastern states to the West, but it also changes patterns of life and work within states. Urbanization, which encompassed only nine per cent of the population in 1830, swelled to 20 per cent in 1860 and continued upwards. From its founding in 1865, the ASSA directed itself to inquiry into the attendant problems: sanitary conditions, relief of the poor, prevention of crime, reform of criminal law, prison discipline, treatment of the insane, in addition to other matters in the domain of ‘social science’ (Haskell, 1977, p. 98). Its organization, following that of the British association upon which it was modelled, consisted originally of four departments: Education, Public Health, Jurisprudence, and Economy, Trade, and Finance – the latter of which was concerned with issues ranging from the hours of work to prostitution and intemperance, public libraries, tariffs and other taxes, the national debt, regulation of markets, and the currency (pp. 104–5). Because participants generally shared the classical liberal assumptions of Wells and Perry – indeed, Wells was an early head of the Economy, Trade, and Finance Department and later president of the Association, and Perry was a regular contributor – their inquiries into these subjects tended not to yield proposals for ambitious programmes of government subsidization or regulation. Instead they were aimed (as the Association's constitution put it) at the diffusion of ‘sound principles’ and at bringing people together ‘for the purpose of obtaining by discussion the real elements of Truth’ (p. 101).

In matters of foreign trade, sound principles implied rejection of protectionism; in the currency, the resumption of specie payments. In the ‘labor question’, which Wells and Perry discussed at a round table in 1866 and again in 1867 (pp. 113–14), sound principles called perhaps for the collection of ample data regarding hours, wages, and conditions of work, but not their regulation on behalf of able-bodied and able-minded adults. Because the interests of workers and employers were in harmony, where poor work

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
conditions existed they were the result either of a general lack of material progress or lack of knowledge by the actors. The solution to lack of material progress was to clear away any legislative impediments to it. The solution to lack of knowledge was for the ASSA to produce and disseminate it. In neither case was the solution to draw up redistributive ‘class legislation’ or other assaults on the prerogatives of property.

The panic of 1873, the depression in its wake, and the great railroad strike of 1877 undermined confidence in the solutions of the liberal reformers and made room for challengers. The labour question had metastasized into ‘the social question’. The most original and widely read author to address it was Henry George (1839–1897), an autodidact working as a journalist in California when the publication of *Progress and Poverty* made him famous in 1879. George admired David Wells, and corresponded with him as early as 1871. He joined Wells in advocating free trade and denouncing ‘monopoly’ (Dorfman, 1969, p. 142). But George was less deferential to property than was Wells, in whose circles he was viewed less of an ally than a danger.

George attributed the persistent poverty of labour amidst material progress to the unearned rewards of one particular kind of property. Land was the free gift of nature: although improvements on it were the work of man and the returns to improvements were earned, the returns to land were not. Yet the returns to land grew inexorably as population and mechanical invention increased and cultivation was extended. The solution that George proposed was a single tax on land of 100 per cent of its annual value. Under such a system it would matter little if one held title to land, leased it to another, and paid a 100 per cent tax on the rent; or if instead the state held formal title to the land and leased it to (presumably) the same person who would rent it from a private owner. Private property in land could be formally retained, but in effect land would be appropriated by the state for the benefit of all. At the same time, the right to all other property would be respected more scrupulously than before, because the single tax on land would obviate the need for all other taxes.

The political economists associated with liberal reform caught the scent of socialism from George's proposal, and they responded with alarm. Yet Henry George shared most of the values and even some of the legislative prescriptions with which the liberal reformers were most closely identified. He departed from them importantly only in his assumption of the illegitimacy of private ownership of one kind of possession. The challenges that would follow in the middle of the turbulent 1880s, from a new generation of economists including Edmund James (1855–1925), Simon Nelson Patten (1852–1922), and Richard T. Ely (1854–1943), were of another order entirely. The professional credentials of the young economists, which included doctoral studies in Germany and university positions in the United States, were different from Wells's and Perry's (let alone George's); and rather than sharing the values and prescriptions of their elders, they repudiated them thoroughly, defining themselves by opposition to economic liberalism. These economists were the founders in 1885 of the American Economic Association.

See Also

- American Economic Association
- Carey, Henry Charles
- catallactics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- free banking era
- George, Henry
- land tax
- List, Friedrich
- Mill, John Stuart
- slavery
- tariffs
- United States, economics in (1885–1945)
- United States, economics in (1945 to present)
- Wells, David Ames

Bibliography

Bryant, W.C. 1808. *The Embargo, or, Sketches of the Times: A Satire*. Boston.

Cardozo, J.N. 1826. *Notes on Political Economy*. Charleston: A.E. Miller.

Carey, H.C. 1836. *The Harmony of Nature as Exhibited in the Laws which Regulate the Increase of Population and of the Means of Subsistence: and in the Identity of the Interests of the Sovereign and the Subject; the Landlord and the Tenant; the Capitalist and the Workman; the Planter and the Slave*. Philadelphia: Carey, Lea & Blanchard.

Carey, H.C. 1840. *Principles of Political Economy*. Philadelphia: Carey, Lea & Blanchard.

Carey, H.C. 1848. *The Past, the Present, and the Future*. Philadelphia: Carey & Hart.

Carey, H.C. 1853. *The Slave Trade, Domestic and Foreign: Why it Exists, and How it May be Extinguished*. Philadelphia: A. Hart.

Carey, M. 1819. *Addresses of the Philadelphia Society for the Promotion of National Industry*. Philadelphia: M. Carey and Son.

Cole, C.C., Jr. 1954. *The Social Ideas of the Northern Evangelists, 1826–1860*. New York: Columbia University Press.

Dorfman, J. 1966. *The Economic Mind in American Civilization, 1606–1865*, 2 vols. New York: Augustus M. Kelley.

Dorfman, J. 1969. *The Economic Mind in American Civilization, 1865–1918*. New York: Augustus M. Kelley.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Eckes, A.E., Jr. 1995. *Opening America's Market: U.S. Foreign Trade Policy Since 1776*. Chapel Hill: University of North Carolina Press.

Foner, E. 1970. *Free Soil, Free Labor, Free Men: The Ideology of the Republican Party Before the Civil War*. New York: Oxford University Press.

Foner, E. 1976. *Tom Paine and Revolutionary America*. New York: Oxford University Press.

Franklin, B. 1729. A modest enquiry into the nature and necessity of a paper currency. In *The Writings of Benjamin Franklin*, vol. 2, ed. A.H. Smyth. New York: Macmillan, 1907.

Franklin, B. 1769. Positions to be examined, concerning national wealth. In *The Writings of Benjamin Franklin*, vol. 5, ed. A.H. Smyth. New York: Macmillan, 1907.

George, H. 1879. *Progress and Poverty: An Inquiry Into the Cause of Industrial Depressions, and of Increase of Want With Increase of Wealth – The Remedy*. San Francisco: W. M. Hinton.

Hamilton, A. 1791. Report on manufactures. In *Encyclopedia of Tariffs and Trade in U.S. History*, vol. 2, ed. C.C. Northrup and E.C. Turney. Westport, CT: Greenwood Press, 2003.

Hamilton, A., Jay, J. and Jefferson, T. 1787–88. *The Federalist*. Online. Available http://www.foundingfathers.info/federalistpapers, accessed 3 September 2006.

Haskell, T.L. 1977. *The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth-Century Crisis of Authority*. Baltimore, MD: the Johns Hopkins University Press.

Jefferson, T. 1785. *Notes on the State of Virginia*. Philadelphia: Prichard and Hall, 1788.

Leiman, M.M. 1966. *Jacob N. Cardozo: Economic Thought in the Antebellum South*. New York: Columbia University Press.

McVickar, J. 1825. *Outlines of Political Economy*. New York: Wilder & Campbell.

O'Connor, M.J.L. 1944. *Origins of Academic Economics in the United States*. New York: Garland Publishing, Inc., 1974.

Paine, T. 1776. Common sense: addressed to the inhabitants of America, etc. In *The Complete Writings of Thomas Paine*, vol. 1, ed. P.S. Foner. New York: Citadel Press, 1945.

Paine, T. 1778. The American crisis: VII. In *The Complete Writings of Thomas Paine*, vol. 1, ed. P.S.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Foner. New York: Citadel Press, 1945.

Perry, A.L. 1866. *Elements of Political Economy*. New York: Charles Scribner's Sons.

Petty, Sir W. 1690. *Discourse on Political Arithmetic*. Reprinted in *Essays on Mankind and Political Arithmetic*. London: Cassell, 1888.

Raymond, D. 1820. *Thoughts on Political Economy*. Baltimore: F. Lucas.

Say, J.-B. 1821. *A Treatise on Political Economy; or the Production, Distribution, and Consumption of Wealth*, trans. from the 4th edn by C.R. Prinsep. Boston: Wells and Lilly.

Smith, A. 1789. *An Inquiry into the Nature and Causes of the Wealth of Nations*, 5th edn, ed. E. Cannan. New York: Modern Library, 1937.

Spiegel, H.W. 1960. *The Rise of American Economic Thought*. Philadelphia: Chilton Company.

Vattel, E. de. 1758. *The Law of Nations, or the Principles of Natural Law Applied to the Conduct and to the Affairs of Nations and of Sovereigns, trans. Charles G. Fenwick*. Vol. 3 of *The Classics of International Law: Le Droit des Gens*, ed. J.B. Scott. Washington, DC: Carnegie Institution, 1916.

Wayland, F. 1835. *Elements of Moral Science*. New York: Cooke and Co.

Wayland, F. 1837. *Elements of Moral Science*, 4th edn. Boston: Gould and Lincoln, 1859.

Wayland, F. 1837. *Elements of Political Economy*. New York: Leavitt, Lord & Co.

Wayland, F. 1841. *Elements of Political Economy*, 4th edn. Boston: Gould and Lincoln, 1873.

Wells, D.A. 1869. *Report of the Special Commissioner of the Revenue for the Year 1868*. In House Ex. Doc. No. 16, 40th Cong., 3rd Session. Washington, DC: Government Printing Office.

How to cite this article

Meardon, Stephen. "United States, economics in (1776–1885)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000071> doi:10.1057/9780230226203.1765

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

United States, economics in (1885 – 1945)

Bradley W. Bateman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The history of American economics following the founding of the American Economic Association in 1885 is not a simple linear narrative of the triumph of neoclassical economics over historical economics. On the contrary, American economics remained a highly plural enterprise until the 1930s. Although there was strife in the 1880s over the proper method of doing economic research, this strife quickly gave way to a long period of détente. Only following the secularization of economics in the 1920s and the advent of the synthesis of neoclassical and Keynesian economics in the 1940s did this pluralism end.

Keywords

Adams, H. C.; American Economic Association; Anderson, B.; Arrow, K. J.; Bain, J.; Catchings, W.; Chamberlin, E.H.; Clark, J. B.; Clark, J. M.; Commons, J. R.; Cowles Commission; Cowles, A.; Currie, L.; Dantzig, G. B.; Douglas, P.; Dunbar, C.; Econometric Society; Elym, R. T.; Fetter, F.; Fisher, I.; Foster, W. T; German Historical School; Gibbs, W.; Gilbert, M.; Great Depression; Haavelmo, T.; Hadley, A. T.; Hamilton, W.; Hotelling, H.; Institutionalism; Knight, F. H.; Koopmans, T. C.; Kuznets, S.; Laissez-faire; Laughlin, J. L.; Marginal School; Mason, E.; Mitchell, W. C.; National Bureau of Economic Research; neoclassical economics; New Deal; Newcomb, S.; Patten, S. N.; Pollak Foundation for Economic Research (USA); probabilistic revolution; Progressivism; real-bills doctrine; Samuelson, P. A.; Seligman, E. R. A.; Snyder, C.; Sumner, W. G.; Sumner, W. G.; Taussig, F. W.; Underconsumptionism; United States, economics in; Veblen, T. I.; Walker, F. A.; Warburton, C.; Willis, H.P.; Young, A. A.

Article

The 60-year period from the founding of the American Economic Association (AEA) in 1885 to 1945 marks the period when American economics was first professionalized and then came to take its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
characteristic, modern form. However, this story is not a simple linear narrative, as the events during this period were influenced by a series of historical contingencies and social forces that meant the outcome remained unpredictable until sometime during the last decade of the period.

Historians and economists alike have tended to believe that the story is somewhat more straightforward than it actually was. The historian Dorothy Ross (1991), for instance, has argued that the rise to prominence of John Bates Clark as America's premier (and first native) economic theorist at the turn to the 20th century marks the triumph of neoclassicism in American economics. More recently, Nancy Cohen (2002) has argued that neoclassicism came to dominate even earlier than Ross suggests, by as much as a decade. Both of these arguments are based, in part, on the well worn idea that American economics was characterized by a '*Methodenstreit*', or 'war of methods' in which the Marginalist School defeated the previously strong Historical School by the end of the 19th century. But while it is true that there was a crisis in American economics during the decade after the founding of the AEA, it had a somewhat different nature from that usually understood: it was much more about the purpose of economics than about method.

American economics, circa 1885

The first doctorate in political economy issued by an American university was in 1886 when Henry Carter Adams received his degree from Johns Hopkins University. Before that time, if an American wanted a doctorate in economics, it was necessary to travel to Europe to study. The vast majority who chose this route studied in Germany, to which 9,000 Americans travelled for this purpose between 1820 and 1920 (Herbst, 1965). The most prominent of these young Americans went on to help found the AEA after their return, including Richard T. Ely, J.B. Clark and Simon Nelson Patten.

In Germany, these young economists found a profession that was characterized by scepticism towards Adam Smith's system of 'perfect liberty' and his arguments for laissez-faire. It now seems fair to say that the economic ideas that the Germans were most disturbed by were in fact a caricature of Smith: the 'Manchesterism' that they found so offensive was more the product of David Ricardo than Adam Smith, but their objection was deep and profound to what they took to be an argument unsuited for all nations at all times. They believed that Smith had made universalizing assumptions about human nature and human behaviour that were not universally correct. In the place of these assumptions, they intended to build a more empirically accurate economics based on careful historical study of how the individual behaviour and economic institutions of different societies had come about. Thus, rather than assuming that all people always weighed their options and made their choices so as to maximize their individual well-being, the leading German economists in the second half of the 19th century believed that it was necessary to study how cultural customs had evolved and how these shaped individual behaviour. Known as the 'Historical School', the leading German economists at this time espoused the idea of '*Nationalökonomie*', or the study of how each nation had come to its current position. These economists taught their students contemporary methods of historical study: careful archival work, the study of laws and customs, the collection of data, and the slow, inductive accumulation of knowledge. However, this did not preclude the study of marginal analysis. Karl Knies, at Heidelberg, who taught marginal techniques to the Austrians Friedrich von Wieser and Eugen Böhm-Bawerk, similarly taught the two most important Americans in Germany, John Bates Clark, the pioneer American marginalist, and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Richard T. Ely, one of the leading representatives of American historical economics.
When Clark returned from Germany in 1875 and Ely in 1880 they found the American economics profession unprepared for much of what they had to offer. American economists were anxious about their own lack of theoretical sophistication. Writing on the occasion of the nation's centennial in 1876, Charles Dunbar, Professor of Economics at Harvard, had complained of the unoriginal and derivative character of American economics. Dunbar ‘observed that American scholarship as yet contributed nothing to fundamental economic knowledge. In his reading, American economics to date had been derivative, stagnant, and sterile’ (Barber, 2003, p. 231). This may be have been an exaggeration, but it was the case that American economics texts at this time, mostly written by college presidents for undergraduate moral philosophy classes, lacked theoretical sophistication, and the young economists returning from Europe found an open field for the possibility of producing a new American economics. Another reason for this lack of theoretical sophistication was that texts were written to propagate the virtues of free enterprise, hard work, and republican democracy, a troika of American virtues that posed a problem for the returning young economists. The crux of the problem was that the elders who controlled the newly created academic positions that the young economists hoped to enter expected anyone they hired to espouse the same beliefs regarding laissez-faire. Francis Amasa Walker, who taught at the Massachusetts Institute of Technology (MIT) and who served as the first president of the AEA, would reflect that laissez-faire, ‘was not made the test of economic orthodoxy, merely. It was used to decide whether a man were an economist at all’ (in Coats, 1988, p. 362).
Unfortunately for men like Clark and Ely, this made employment in academe difficult, if not impossible, for on their immediate return from Germany they almost all adhered to Christian socialism. In part, this commitment reflected the enthusiasm they had acquired in Germany for social reform. But it also reflected an effort on their part to build an American ‘*Nationalökonomie*’. Trained to believe that each country had a unique culture and unique institutions, these young men latched onto evangelical Protestant Christianity, and tried to use it to fashion a native argument for economic reform.
Following the Civil War, after an era when the importance of hard work, free enterprise and the rights of capital had been universally preached (Bateman, 2005). American Protestantism had begun to split into two ‘parties’ (cf. Marty, 1986). The ‘private party’ focused on individual salvation and did not view the emerging economic conditions as a matter for Christian concern. The ‘public party’, on the other hand, focused on social reform while de-emphasizing the older evangelical concern with individual piety. In the new economic order of the post-bellum world these two strands of evangelical thought became emblematic of the two most common (and opposing) responses to the growth in industrial production, increasing urbanization, rapid immigration and growing inequality. Each group took the Bible as its primary text, but their purposes and understanding of the world could not have been more different. Whereas the private party Christians focused on individual piety, the public party wanted to build ‘the Kingdom of God on this earth’ through collective action and with the aid of the state.

The progressive era

At the centre of the public party of American Protestantism were the young economists who had returned from Germany. They had learned a new scientific ethic of empirical study and had made initial contact with marginal reasoning. But balanced against this, was the fact, uncomfortable for their elders,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that almost to a person these returning young economists were interested in changing the balance of power between working men and the owners of the new, vertically integrated enterprises that employed them. Thus, the young German-trained economists had some things that the profession craved, such as technical prowess, but they also manifested a political attitude that was unacceptable to much of the profession.

Laissez-faire in 19th-century America was not a technical argument for the efficiency of free markets, but rather a manifestation of a broader Protestant ethos. It meant a limited role for the state in the economy, but it also meant the unfettered right of capitalists to employ workers on the capitalists' terms; at the time of the Civil War, for instance, virtually throughout the United States it was still illegal to strike or to form a labour union. Thus, when the young, German-trained economists challenged laissez-faire, they were as likely to be arguing for the right of workers to strike as they were to be arguing for the municipal provision of water. In the last decades of the 19th century, then, laissez-faire meant both a limited role for the state and privilege for the prerogatives of capital. Men like Ely and Patten meant to challenge these ideas head-on and they took their warrant to do so as much from Christian scripture as they did from economic theory: they saw their work as economic theorists as an expression of their Christian commitments.

However, they generally met scorn from the older generation. Under the leadership of Ely, Patten, and Clark, the young advocates of labour and the possibility of beneficial state intervention formed the American Economic Association (AEA) in 1885, in part as an effort to provide support for one another and to seek an alternative form of professional recognition. But not all the AEA's members were economists; about a quarter of the members at the first meeting in Saratoga Springs, New York, were Protestant ministers. The young economists with doctorates could perform empirical research showing the extent of poverty, inequality and industrial dislocation in the economy; the ministers could preach the young economists' findings from their pulpits on Sunday morning to help motivate their parishioners to act to help build the Kingdom.

The formation of the AEA led to one of the most well-known exchanges between the older advocates of laissez-faire and the young economists who hoped to establish a role for the state in the functioning of the economy. William Graham Sumner at Yale University was the most high-profile advocate of laissez-faire at the time of the founding of the AEA, but one of Ely's colleagues at Johns Hopkins, Simon Newcomb, was also widely recognized. Newcomb was an acclaimed astronomer and mathematician who occasionally lectured on economics at Johns Hopkins and he wrote widely on economics in the popular press. Following the establishment of the AEA in 1885, *Science* magazine asked members of the old and new schools to debate their positions in a series of exchanges (reprinted in Adams, 1886). The exchange was often vitriolic, with Newcomb accusing Ely of being a socialist, no small charge in the immediate aftermath of the Haymarket riots. Often unnoticed in the debate, however, is the fact that both sides called on the authority of Alfred Marshall and William Stanley Jevons to establish their own points. More than anything in the debate, this last fact perhaps illustrates the extent to which the American economics profession at this time was animated by political differences over laissez-faire and the role of the state, rather than primarily by an argument over methods.

Much to the chagrin of Ely, the AEA did not long remain a haven for advocates of the state. With the pressure on the young economists to demonstrate they were not radicals who advocated violence, they felt it necessary to show themselves open to their older colleagues. And since the Association was also billing itself as the national organization for all economists, and because the young economists depended

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
on the older generation for jobs and support, they hardly felt that they could deny the older economists membership in the new organization when they asked to join. Thus, by 1892, virtually all the older advocates of laissez-faire were members of the organization and the founding statement of principles in favour of a role for the state and amelioration of economic dislocation, which Ely had helped draft in 1885 at the time of the founding, had been abandoned. Ely's response was to boycott the annual meeting that year.

At this point, it is undeniable there was great tension in the profession. It is important to realize, however, that, John Bates Clark, the premier marginalist, and Richard T. Ely, the premier historical economist, had worked together to found the AEA. It is true that Clark has dropped his interest in Christian socialism by 1885, but he had not turned completely against historical analysis. Nor, strictly speaking, was Ely an anti-marginalist; Ely had introduced marginal utility into his textbook writing as early as 1893. E.R.A. Seligman, another of the AEA founders, represents a good case study of the way that many economists at this time balanced marginalist and historical techniques in their work. The real conflict within the profession at this time was about the possible role of the state and whether the purpose of economics was to defend laissez-faire, rather than whether marginalism should be used in economic analysis (cf. Yonay, 1998).

The diversity within the profession is also clearly manifest in the leading figures of the old school that emerged during the last two decades of the 19th century: Arthur Hadley, Frank Taussig, and J. Laurence Laughlin. These three men became the leaders respectively at Yale, Harvard and Chicago in the era when departments were emerging as the dominant institutions in the formation of the American economics profession. The work of these men represents three of the main economic problems facing the nation at the end of the 19th century (international trade and tariffs, railroads and monetary policy) and their theoretical eclecticism, while adhering to some form of the older cost-based theories of classical economics, demonstrates the evolving nature of the profession.

J. Laurence Laughlin was the most rigidly classical of these three ‘young traditionalists’ (Dorfman, 1949). Despite writing a dissertation in history under Henry Adams, Laughlin gave no credence to the Historical School and was the last of the major figures in the old school to join the AEA (in 1904). Although not a prolific writer, his *History of Bimetallism in the United States* (1885) is a landmark study of the history of late 19th-century fights over currency and monetary policy. Ironically, however, after founding the *Journal of Political Economy* (1892), Laughlin opened the journal to economists of all stripes, and turned the book reviewing over to Thorstein Veblen for several years.

Frank Taussig is probably the most well remembered of the three, not least for his influence on the generation of students who enroled in his ‘EC 11’ graduate seminar at Harvard. Widely respected as a teacher, he was a gregarious person who was the first of the old school economists to join the AEA (in 1886). Although himself an adherent of classical economics, particularly of Ricardo and John Stuart Mill, Taussig was capacious in his analysis of economic problems and was often willing to see the legitimacy of other points of view. There is no better example of this than in his *Tariff History of the United States* (1888), an adaptation of his dissertation that went through many revisions in its subsequent editions; although he was a dedicated free trader, Taussig had a subtle understanding of the use of tariffs for revenue collection and appreciated, for instance, the arguments for sometimes protecting infant industries. Like Laughlin, Taussig was also an eclectic and important early journal editor. Charles Dunbar had founded the *Quarterly Journal of Economics* at Harvard (1886), but when Taussig became its editor, he opened his pages to economists of all views.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Arthur Hadley tackled one of the major economic issues in late 19th-century American capitalism, the railroads. His early reputation was based on his *Railroad Transportation* (1885), in which he argued that businesses with large fixed costs would not necessarily shut down when prices fell below the cost of production. Instead, he pointed out that as long as the firm could cover what are now known as variable costs and still make some contribution towards paying fixed costs such as interest payments, that they would stay in business. Hadley represents an intriguing figure for he acknowledged that his results on the role of fixed cost contradicted some of Ricardo's arguments and he also accepted the basic validity of marginal utility reasoning. Thus, he seemed to have transcended classical economics in many regards. Yet his outlook was unmistakably laissez-faire, and he believed that there was no reason for further work in marginal utility theory, since its validity had been shown and that more work represented an unnecessary foray into psychology. In the end, his economics was driven largely by the study of costs, as in classical economics, and his conclusions did not stray far from the classic dogmas.

Members of the older generation were not alone in forming strong departments during the last decade of the 19th century. Johns Hopkins, which was founded in 1876 to establish higher education, especially graduate education, in the German style in the United States, was the first notable American graduate course in economics, producing the first Ph.D. granted in economics in the United States. The department's two notables were Simon Newcomb and Richard T. Ely. However, after his falling out with Newcomb, Ely left Johns Hopkins in 1892 for the University of Wisconsin, where he quickly assembled one of the top economics departments in the nation. Initially, Ely's move to Wisconsin was seen to be a possible precursor for regional factionalism in American economics. After Ely failed to attend the annual meeting of the AEA in 1892, there was concern among many Eastern economists that he would lead a movement from the Midwest to create a new professional organization that would promote his original challenge to laissez-faire.

However, following Ely's academic trial in Wisconsin in 1894 on charges of entertaining a union organizer in his home and of advocating socialism in his lectures, people from both camps began to look for common ground. Ely self-consciously chose to lower his profile after he was acquitted in his trial, and the advocates of laissez-faire began to realize that it was not good for the profession's credibility to have high-profile public disagreement. What followed at the turn of the century was the emergence of a kind of détente in American economics: leading figures in the profession continued to build strong departments around the country, but economists were granted a wide berth to examine social and economic conditions and to use the tools they saw best suited for the specific question at hand. Support for laissez-faire and government intervention were both accepted; what was expected was a rigorous approach to one's position.

One basis for this détente was undoubtedly the common Protestant background of most American economists at this time. Although Ely, Patten, John R. Commons, and Henry Carter Adams were asking American Protestants to turn away from their traditional position in favour of the prerogatives of the owners of capital, they were making the appeal on biblical grounds and were self-consciously appealing to the emerging public party of Protestantism. Since prominent members of the group of younger economists (for example, Ely and Adams) made it clear in their academic trials that they were not advocating the overthrow of the state, but rather the empirical study of the conditions of labourers, it became harder to demonize them for their impulse to seek what were clearly Christian ends. For their part, the older advocates on laissez-faire were willing to begin the difficult work of absorbing new theoretical techniques and accepted the call of the younger economists to undertake more empirical

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
work and to let its results inform their understanding of the true effects of relatively untrammelled markets.

This détente between the younger and older economists, and between advocates of laissez-faire and advocates of more rights for labourers, created a fertile ground for American economics. During this period, there was not a single orthodoxy in American economics; one could work with marginalist ideas, historical ideas, or with both. New School economists such as Ely, Clark, and Seligman, as well as Old School economists such as Taussig and Hadley, all employed both techniques in their work. All that was required to be taken seriously in this world of plural methods was a dedication to the examination of the contemporary economic issues that were arising as America became an industrialized, urbanized nation. That the *American Economic Review*, which was founded in 1911, showed no marked tendency towards any school in its published articles is strong evidence that there was no single dominant school at this time.

Perhaps the most notable dissent from this détente was the *sui generis* Thorstein Veblen. Like the young economists returning from Germany, Veblen was interested in a more ‘scientific’ economics. But he was never much interested in large-scale empirical work such as the social survey movement fostered by Ely and Commons at the turn of the century. Instead, Veblen wanted economics to be rebuilt as an evolutionary science based on Darwinian principles of natural selection (Hodgson, 2004). Veblen's greatest theoretical advances would come during the first decade of the 20th century, but they were offered as part of a sardonic and biting criticism of Clark's work and so not only alienated Clark and his followers but also the bulk of the profession who saw their toolbox as containing many different techniques, one of which might sometimes be marginal reasoning. Veblen agreed that American economists should examine the newly emerging American capitalism, but he was not interested in pluralism of techniques.

Perhaps the most visible sign of the emerging détente among most American economists at this time was Ely's election to the presidency of the AEA in 1900. Not only had his feared defection been averted, but he had been successfully pulled back into the centre of the organization. Ely and Veblen never shared a close personal or professional relationship and Veblen would later harbour a bitter resentment that his brilliance and originality had been ignored by the AEA in the pivotal years when it would have helped his professional stature. Despite the fact that Ely and Veblen shared an interest in historical analysis, Veblen's style and his self-certainty in the Darwinian method kept him apart from the mainstream.

The year 1900 also marked the beginning of the Progressive Era, a profound shift in American society that would propel economists like Ely, Adams and Commons into the mainstream of American thought. With the emergence of progressivism during the first two decades of the 20th century, the cultural, political, and religious centre of American society would move away from 19th-century ideals of laissez-faire and towards an ethos of active civic engagement in trying to ameliorate the many social dislocations of the new industrialism, emerging urban poverty and concentrations of power in large corporations. The public party of Protestantism was the central force of progressivism in the first decade of the new century and economists like Ely and Adams, who had been subjected to political pressure to mitigate their views in the late 19th century, now found themselves in great demand and at the head of progressive projects such as regulatory commissions and social survey projects (Furner, 1975; Bateman 2001). Just as the laissez-faire economists in the 19th century had benefited from the dominant Protestant ethos, so too would the young, so-called ‘ethical economists’ benefit from the rise of public party Protestantism early in the 20th century. This public recognition, and the university positions that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
often came with it, undoubtedly made it easier for the ethical economists to ignore Veblen's criticisms. This is not to say that the economics profession as a whole was receiving the attention it believed it deserved. On the contrary, for a significant part of the 20th century, the AEA was actively concerned with measures that would secure it an appropriate public profile (see Bernstein, 2001).

The First World War and the end of the idyll

At the same time that the young economists who had founded the AEA were rising to popular prominence, another generation of marginalist economists was also beginning to emerge. One of the most distinguished of the new generation of marginalists was trained by Ely. Allyn Young received his doctorate working under Ely at Wisconsin and became the first of many to become revisers and co-authors of Ely's best-selling *Principles* text. Some of the new generation of marginalists, such as Frank Fetter, examined the psychological dimensions of utility and sought to more clearly articulate the welfare implications of marginal utility reasoning. Perhaps the most innovative marginalist thinker to emerge after Clark was Irving Fisher. Fisher's mentors at Yale were William Graham Sumner and Willard Gibbs. Gibbs was one of the most prominent mathematicians of his generation and he influenced Fisher to develop marginal economics in a more technical, mathematical form that would later come to characterize American neoclassical economics.

While Fisher's work was recognized and lauded, it was not, however, representative of the mainstream in the first two decades of the 20th century. Progressivism defined the centre of American political and intellectual life from roughly the turn of the century to the end of the First World War; but during the second decade of the century it shed some of its public rhetoric of *Protestant* reform as it began to pull in Jewish writers like Walter Lippman and Herbert Croly. Progressivism also became more focused on industrial efficiency as many embraced Frederick Taylor's time and motion studies as a means to achieve greater productivity and raise the standard of living. But the moralism of Protestant reform still suffused much of progressive thinking and would lead to the movement's quick demise after the war.

The problem for progressivism after 1918 was that one of its greatest proponents, Woodrow Wilson, had led the nation into war using the rhetoric of reform and democracy. He had been supported by the Protestant clergy, many of whom had used their pulpits to preach the justness and necessity of the war when America had entered the conflict in 1917. Thus, when the war was over, and the atrocity of the trench fighting was driven home to people, there was a quick and sudden turn against progressivism and especially against the moralism and rhetoric of moral improvement that had underpinned it (Danbom, 1987). The hope that had supported the progressive movement was now widely seen to have been based on an unrealistic understanding of human nature.

This shift away from progressive moralism had a quick and deep effect upon all the American social sciences that had been pioneered by public party Protestants; sociology, political science, and economics all made a sudden turn to a more ‘scientific’ and less ‘moralistic’ basis. Dorothy Ross (1991) has called this ‘the advent of scientism’. Of course, all three social sciences had considered themselves scientific before the war (Furner, 1975), and all had been interested in empirical survey work of urban and rural populations; but they had all operated on an implicit belief that if this survey work revealed social pathologies such as poverty and child labour, that good Christian women and men would surely act to alleviate them when faced with the evidence. After the war, such a reliance on an idea that people were

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
well motivated and altruistic was abandoned. So, too, was the idea that people who lived in squalid social conditions would experience moral improvement if the conditions were changed.

Thus, in the years immediately before 1920, progressive social science quickly unravelled. Realizing that they had lost the sympathy of the larger populace, and faced with the need for serious soul-searching on their own part, all three social sciences made an explicit effort to eschew the optimistic, moralistic rhetoric of progressivism and embraced a new kind of 'scientific' endeavour. Entwined in their decisions to embrace a more value-neutral approach to social enquiry was a clear understanding that both public and private funding hinged on catching the tone of the times.

The name of this movement in economics was 'institutionalism' (see institutionalism, old; Rutherford, 2000; Bateman, 2004). The term was coined in 1918, at exactly the moment when the break from liberal Protestantism was happening in all three of these social sciences. The men who formed the nucleus of this emerging group, men like Walton Hamilton and Wesley Mitchell, self-consciously endeavoured to set up an empirical, data-generated research project that would be appealing to funding agencies such as the Carnegie and Rockefeller Foundations. The founding of the National Bureau of Economic Research (NBER), with Mitchell as the director, was one of the signal achievements of the early institutionalists.

One effect of the effort towards a more empirical basis for economic research was that subjects that had been at the centre of American economics for at least four decades, such as poverty and philanthropy, were dropped from the research agenda of almost all institutionalists. Instead, intense attention was paid to the cost structure of American industry, the business cycle, and the working of the financial system; these seemed to be the proper objects of serious economic scientists. Ultimately, the object of the institutionalists was to find a more scientific basis for 'social control', thus eradicating the need for the moralism of the progressives.

The most notable break with the past, however, was that for the first time since the founding of the AEA, there was now a *group* of American economists who were attempting to establish an institutional and historical approach to economics and who did not want a détente with marginal analysis. The *Methodenstreit* that many people now project back on to the late 19th century was actually beginning to emerge. The institutionalists were interested in developing a behaviouralist basis for individual behaviour and they eschewed the idea that marginal decision-making was the driving force behind most economic activity.

The institutionalists were also the first *group* of American economists to work to establish an explicitly secular economics. This reflected their desire to distance themselves from the moralizing rhetoric of the Christian economists who had founded the AEA and drew from the work of one of the institutionalists' main influences in pragmatic philosophy, John Dewey. While not every individual American economist had been a Protestant before the First World War, the ethos of Protestantism had suffused and stabilized the détente that held for most of the three decades after Ely had been exonerated in his academic trial. Liberal Protestant economists had believed that a kind of moral Darwinism was at work in American society and their common purpose in exploring social questions to determine where and when state intervention might (or might not) be appropriate had been made possible by their shared ethos. Now, however, the men who stepped forward to found institutionalism were making an explicit argument that science alone should inform their work; they staked their future on the idea that they could do better, more empirical science than the a priori marginalists who they believed depended on untested and incorrect assumptions about human behaviour.

The marginalists were fully up to the fight, however, and engaged the institutionalists after 1930 in an

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
increasingly pointed dispute. In the first decade of institutionalism's rise, the détente held reasonably well. And during the 1920s, institutionalism was at least as well represented in the top graduate schools as marginalism. While marginalist thinkers such as Irving Fisher went about their work, the instituionalists controlled two of the top four graduate courses (Columbia and Wisconsin) and produced a large plurality of American Ph.D.s (Bowen, 1953; Backhouse, 1998; Biddle, 1998).

The Great Depression and the New Deal

Contrary to the idea that American economics in the 1920s and 1930s was characterized by a final struggle between some latter-day historicists (that is, institutionalists) against the dominant neoclassicists, the pluralism of this period was much richer and represented a wide range of possibilities regarding the future direction of the profession. It might more correctly be said that no school of thought in American economics was completely dominant at this time. A nice example of this diversity would be to consider Harvard in the 1930s. Frank Taussig was still on the faculty and remained one of the leading authorities on international trade. Joseph Schumpeter would join the department in 1932, bringing a Continental influence, if not exactly an Austrian one. Edward Chamberlin would finish his work on imperfect competition at Harvard during the 1920s, under the twin influences of Marshall and Allyn Young. Young, who died in 1929, had been one of America's top theorists, but as noted above, he was a student of Ely's who had later tacked hard to the marginalist tradition. One could not define this department as simply neoclassical, but neither was it simply institutionalist. The full secularization of American economics, however, had prepared the ground for a more strident return to laissez-faire arguments, much like the ones that had existed before 1885.

For by the 1930s, leading marginalist thinkers such as Frank Knight were prepared to engage in what Yuval Yonay (1998) has termed ‘the fight for the soul of economics’. Despite their control of many of the top graduate courses (Backhouse, 1998), and their initial success at fundraising, the institutionalists were hit hard by the kind of criticism that Knight levelled against Sumner Slichter in his 1932 review of Slichter's new institutionalist textbook. Knight did not like Slichter's methods of analysis, but his real bête noire was Slichter's focus on intervention to improve the performance of the economy. By this time, the institutionalists had clearly staked out their position of advocacy for ‘controlling’ the economy, and the advocates of marginalism, such as Knight and Jacob Viner, strongly disagreed. However, though some took clear sides, there were others who straddled both camps: John Maurice Clark (J.B. Clark's son) drew on many neoclassical tools, such as externalities, but his work on the control of business went further in directions favoured by institutionalists, turned away from the cause in the 1930s, arguing that many of the instituionalists’ concerns were best handled by treating them as externalities within the neoclassical model; before his untimely death in 1929, Allyn Young had done his pioneering work in the economies of scale and supervised Chamberlin's doctoral thesis. The divide between institutionalism and neoclassicism was thus still blurred in this period.

Equally destructive of the myth that American economics was essentially a linear narrative of the development of neoclassical economics after 1890 is the rich American tradition of work on money, banking and the business cycle during the inter-war years (Laidler, 1999). The work in these areas drew from the pre-First World War work of major figures such as Laughlin, Fisher, and Mitchell, and represented a wide range of theoretical development, as well as a full range of policy options.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Perhaps the most well remembered work from this period is Irving Fisher's work on the quantity theory and the relationships between money growth, the price level, and economic activity. Fisher (1911; 1923; 1925) was not alone in positing a close relationship between money and prices. Carl Snyder (1924) of the Federal Reserve Bank of New York drew on Fisher's work to suggest that the close relationship between money and prices supported a ‘money growth guideline’ (Laidler, 1999).

There was no orthodoxy in monetary and cycle theory, however. For instance, Irving Fisher (1925) came to believe that there was no business cycle, simply fluctuations around the mean values of prices. This outlook contrasted sharply with Mitchell's careful empirical search for the factors that underlie what he believed were the regular oscillations of the economy. Both Mitchell and Alvin Hansen worked in the 1920s to develop versions of the accelerator principle, a concept previously developed by J.M. Clark in 1917, trying to uncover the ways that a growing economy could pick up momentum and how that same pattern of growth could ultimately lead to a downturn.

Likewise, just as some economists did not agree with Fisher's conclusions about the nature of the cycle, there were many who did not agree with his conclusions regarding the relationship between money and prices. In the inter-war years, there developed what David Laidler has termed an American version of the British Banking School. H. Parker Willis, who worked for the Federal Reserve in the second decade of the 20th century and later taught at Columbia's School of Business, and Benjamin Anderson, who had a position at Harvard before becoming the chief economist at Chase Bank, both argued against Fisher's use of the quantity theory. Willis had studied under Laughlin at Chicago and he followed in Laughlin footsteps in denying the necessary connection between money and prices.

In this rich mix of work regarding money and the business cycle, there was, not surprisingly, widespread disagreement about the possibilities for stabilization policy. Economists as diverse as Fisher, Mitchell, and Allyn Young supported different kinds of stabilization policy. Others, like Willis, adhered to a version of the real-bills doctrine, arguing that stability would come only through a prudent effort on the part of the Federal Reserve to limit its lending to those with high-quality, short-term commercial paper. Although they did not have academic appointments, William Trufant Foster and Waddill Catchings (1923; 1925; 1928) used the Pollak Foundation for Economic Research as an effective platform to publicize a version of underconsumptionist theory. They argued that there was a need for monetary and fiscal expansion to sustain consumption and avoid recession. Paul Douglas (1927) at Chicago, who became a US senator from Illinois, made these ideas popular with his call for large-scale public works.

While there were, thus, many forms of arguments for fiscal and monetary efforts to sustain prosperity, it might seem that the institutionalists’ predisposition for controlling the economy would have been popular after 1929, but neither the popular nor the professional tide turned toward the institutionalists after 1929. In retrospect, it was, perhaps, their unique bad luck to have played a role in developing parts of Franklin Delano Roosevelt's initial response to the Great Depression, his first New Deal. Many institutionalists joined Rexford Tugwell from Columbia University in Roosevelt's first term of administration, but they failed to provide a recognizably successful policy for combating the depression. Although the marginalist economists were not offering a popular plan for recovery, the institutionalists’ efforts in the New Deal did not provide them with a set of successes upon which to build their legacy (Barber, 1996).

It was also the bad luck of the institutionalists that by the end of the 1930s, when John Maynard Keynes's *General Theory of Employment, Money, and Interest* (1936) came to be widely seen as

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
providing a theoretical underpinning for recovery, that the emerging neoclassical theorists like Paul Samuelson were already working to give his theory a neoclassical underpinning. The efforts to cast the *General Theory* in a general equilibrium model effected a remarkable and unexpected transformation in the future prospects for American economics, overshadowing the fact that the foundations of what came to be thought of as Keynesian fiscal policy were laid by institutionalists (Barber, 1996; Rutherford and DesRoches, 2008). Before the Great Depression, economists of all stripes had argued about possible monetary and fiscal interventions (see Barber, 1985; Laidler, 1999). Although the pure idea of 'social control' was an institutionalist construction, the potential for using fiscal and monetary policy came from almost every corner of the profession. Keynes's work provided a common analytical framework for examining such macroeconomic interventions, and Samuelson's work linked that analytical framework to marginalist, neoclassical ideas. Likewise, Alvin Hansen's embrace of Keynes's theoretical framework after his initial resistance lent important impetus to an alternative form for analysis of the business cycle, even though he turned to Keynesian economics only when he saw that it could be used to defend ideas he already held. Hansen built his graduate seminar on fiscal policy at Harvard around his reinterpretation of Keynes's *General Theory*, and in the 1940s and early 1950s wrote a series of texts that embodied much of the institutionalist concern with the business cycle and economic stabilization (see Mehrling, 1997, chs 7–8). The project that institutionalists had undertaken in the 1920s to provide a new psychological basis for economic behaviour had never led to any substantive advances, which helped to make the mathematical elegance of Samuelson's work all the more attractive. It did not hurt that the tools of neoclassicism, in their Keynesian guise, were now seen to be as amenable to intervention as they were to laissez-faire. Edward Chamberlin's (1933) work analysing imperfect competition in the 1930s also started to lend a new sense of realism and possibility to the emerging neoclassical framework. In the 1930s, much empirical work had been undertaken by institutionalists, but it was Edward Mason and Joe Bain, who, drawing on Chamberlin's theory, developed the framework that was to dominate empirical work on industrial economics in the 1950s. Harold Hotelling's theoretical breakthroughs in formulating mathematical models of resource depletion in the 1930s added even more lustre to neoclassicism. A new, sharp sense of what it meant to be an economic theorist was emerging, and institutionalism did not appear to have a ready response.

The econometric movement and the Second World War

The event that symbolized this confidence in more formal theorizing than the institutionalists had generally favoured was the foundation of the Econometric Society in 1930. It was an international society, but Fisher, Schumpeter and other American economists were influential in its formation. Its importance lay in providing a focus for mathematical theories, covering both the cycle and microeconomics, and statistical research. An influential figure was Alfred Cowles, who not only supported the Econometric Society and its new journal, but also established the Cowles Commission. This was an economic research organization, set up in 1932, the heyday of which was from 1939 to 1955 when it was based in Chicago. Located outside the economics department, it provided a focus for the development and propagation of neoclassical theory (with particular emphasis on Walrasian general equilibrium theory) and statistical methods for testing and applying the theory. It was here, in the early 1940s, that Tjalling Koopmans and Trygve Haavelmo developed the methods that led to what has been

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
called the probabilistic revolution in econometrics (Morgan, 1990).

Two further developments were important in the transformation of American economics that took place in this period. One was the influx of a large number of émigrés from Europe. The United States had always been a home for such people, and in the 1920s many economists arrived from Russia and Eastern Europe, but this increased dramatically in the 1930s and 1940s. By the mid-1940s, almost half the authors of articles in the *American Economic Review* had been born outside the United States, most of these being affiliated to American universities (Backhouse, 1998).

The other influence was the Second World War itself. This was, like no previous war, an economists' war (Bernstein, 2001). Economists were recruited en masse into government. Many were employed to tackle what were clearly economic problems relating to domestic economic activity or to the estimation of enemy economic capacity. The most notable outcome of such work was national income analysis. Official estimates of US national income had first been calculated in 1933, in response to the onset of the depression, when Simon Kuznets was seconded to the Bureau of Foreign and Domestic Commerce from the NBER, where he had been working on the problem (elsewhere Clark Warburton, at the Brookings Institution, and Laughlin Currie had been working on similar lines). Under Robert Nathan this work was developed, and monthly figures were produced by 1938. But it was only during the war that these estimates were developed, under Martin Gilbert, into a system of accounts. One reason for this was that national income proved indispensable to the war effort, its main achievement being to calculate what Roosevelt could promise in his Victory Program.

However, the significance of the war went further than this, for economists also became involved in activities not traditionally associated with their subject. Operations Research, initiated in Britain in the 1930s, was taken up by the American armed forces, through the Office of Strategic Services (a forerunner of the Central Intelligence Agency) where economists were employed alongside mathematicians, statisticians and physicists to solve problems related to military strategy and tactics. Out of this arose techniques that later proved influential, such as linear programming, with which members of the Cowles Commission (Koopmans and George Dantzig) were heavily involved. Economists achieved a high reputation as general problem-solvers. Most important, however, was the effect on the way economics was conceived. Much of this work was focused on optimization and was highly technical. Economics came to be seen as akin to engineering. In the 1920s and for much of the 1930s it had been institutionalism that was associated with quantitative work; statistical work related to neoclassical theory did exist (for example work on measuring demand functions) but there was no parallel with the work being done at the NBER and by the institutionalists. In contrast, by the 1940s, there was in place a serious research programme, with techniques that were perceived to rival those of the hard sciences, in which theory and data interacted in a way that was different from that found in inter-war institutionalism. The transition was a very slow process: for instance, when Kenneth Arrow entered Columbia as an undergraduate in the 1940s, he was still not taught modern price theory (Colander, Holt and Rosser, 2004, ch. 10), but rather the institutional economics of the 1920s and 1930s. The scene was set for the disputes that were, in the late 1940s and early 1950s, to determine the way economics was to evolve after the Second World War.

Conclusions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
At the beginning of the 20th century, America's economic heritage was still tied up with the cultural influence of Protestantism. By the 1930s, that legacy had disappeared. The significance of institutionalism was not that it continued the earlier historical economics acquired by Ely, J.B. Clark and their contemporaries in Germany but that it was an organized movement for a purely secular and scientific economics (Rutherford, 1999). Some 50 years after the founding of the AEA, the profession was still split deeply on the proper role of the state in the economy, but everyone in the discussion now believed that the role of the state was a scientific question and was actively engaged in the development of the tools to answer the question. Neoclassical methods, the forerunners of those that dominated the profession in the post-war era, were being developed, but there was an immense variety, which the labels neoclassical and institutionalist fail to capture. It was a period of genuine pluralism in economics (Morgan and Rutherford, 1998).

In retrospect, it is possible to discern the advance of neoclassical and more technical economics on a broad front. Young, Chamberlin, Hotelling, Samuelson and others were establishing a theoretical framework that could animate both microeconomic and macroeconomic work. However, rather than see this process as inevitable it is important to see the importance of external factors in determining the outcome of inter-war pluralism. The Great Depression exerted a profound effect; institutionalist planning was tainted by the failure of Roosevelt's first New Deal, in which planners such as Rexford Tugwell were heavily involved. The Second World War was perhaps even more important in helping change perceptions of what economics was and ideas about the position of economists in society. The rise of the Nazi party and their policies in Europe not only removed a major rival to the supremacy of Anglo-American economics, but caused an influx of economists into the United States who proved highly influential.

In addition, it is important not to exaggerate the extent of any neoclassical victory. Institutionalists remained strong in many fields. The NBER's work in establishing a statistical basis on which empirical analysis could be based was unrivalled. Furthermore, even where there would appear to be evidence for the conversion of institutionalists to other approaches, significant elements of institutionalism remained, as in J.M. Clark's work on the control of business or Hansen's use of Keynesian methods for analysing the business cycle. The legacy of institutionalism was widespread. There was a swing away from institutionalism, which can be documented in many ways (see, for example, Backhouse, 1998; Biddle, 1998) but the story was not linear, and in the 1930s and 1940s it was highly dependent on external events.

See Also

- American Economic Association
- Clark, John Bates
- Ely, Richard Theodore
- institutionalism, old
- neoclassical synthesis
- Young, Allyn Abbott

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Backhouse, R.E. 1998. The transformation of US economics 1920–1960: viewed through a survey of journal articles. *History of Political Economy* 30(Suppl.), 85–107. In *Interwar Pluralism to Postwar Neoclassicism*, ed. M.S. Morgan and M. Rutherford. Durham, NC: Duke University Press.

Barber, W.J. 1985. *From New Era to New Deal: Herbert Hoover, the Economists, and American Economic Policy, 1921–1933*. Cambridge: Cambridge University Press.

Barber, W.J. 1988. *Breaking the Academic Mould: Economists and American Higher Learning in the Nineteenth Century*. Middletown, CT: Wesleyan University Press.

Barber, W.J. 1996. *Designs within Disorder: Franklin D. Roosevelt, the Economists, and the Shaping of American Economic Policy, 1933–1945*. Cambridge: Cambridge University Press.

Barber, W.J. 2003. American economics to 1900. In *A Companion to the History of Economic Thought*, ed. W.J. Samuels, J.E. Biddle and J.B. Davis. Oxford: Blackwell.

Bateman, B.W. 1998. Clearing the ground: the demise of the social gospel movement and the rise of neoclassicism in American economics. *History of Political Economy* 30(Suppl.), 29–52. In *Interwar Pluralism to Postwar Neoclassicism*, ed. M.S. Morgan and M. Rutherford. Durham, NC: Duke University Press.

Bateman, B.W. 2001. Make a righteous number: social surveys, the men and religion forward movement, and quantification in American economics. *History of Political Economy* 33(Suppl.), 57–85. In *The Age of Economic Measurement*, ed. J.L. Klein and M.S. Morgan. Durham, NC: Duke University Press.

Bateman, B.W. 2004. Why institutional economics matters as a category of historical analysis. *Research in the History of Economic Thought and Methodology* 22A, 219–56.

Bateman, B.W. 2005. Bringing in the state?: The life and times of laissez-faire in the 19th century United States. *History of Political Economy* 37(Suppl.) In *The Role of Government in the History of Political Economy*, ed. Steven Medema and Peter Boettke. Durham, NC: Duke University Press.

Bateman, B.W. and Kapstein, E. 1999. Retrospectives: between God and the market: the religious roots of the American economic association. *Journal of Economic Perspectives* 13(4), 249–58.

Bernstein, M. 2001. *A Perilous Progress: Economists and Public Purpose in Twentieth Century America*. Princeton, NJ: Princeton University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Biddle, J. 1998. Institutional economics: a case of reproductive failure? *History of Political Economy* 30 (Suppl.), 108–33. In *Interwar Pluralism to Postwar Neoclassicism*, ed. M.S. Morgan and M. Rutherford. Durham, NC: Duke University Press.

Bowen, H.R. 1953. Graduate education in economics. *American Economic Review* 43, 1–223.

Chamberlin, E.S. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Coats, A.W. 1988. The educational revolution and the professionalization of American economics. In *Breaking the Academic Mould: Economists and American Higher Learning in the Nineteenth Century*, ed. William J. Barber. Middletown, CT: Wesleyan University Press.

Coats, A.W. 1992a. *British and American Economic Essays, vol. 1: On the History of Economic Thought*. London: Routledge.

Coats, A.W. 1992b. *British and American Economic Essays, vol. 2: The Sociology and Professionalization of Economics*. London: Routledge.

Cohen, N. 2002. *The Reconstruction of American Liberalism, 1865–1914*. Chapel Hill, NC: University of North Carolina Press.

Colander, D., Holt, R. and Rosser, J.B., eds. 2004. *The Changing Face of Economics: Conversations with Cutting Edge Economists*. Ann Arbor: University of Michigan Press.

Danbom, D. 1987. *The World of Hope: Progressives and the Struggle for an Ethical Public Life*. Philadelphia, PA: Temple University Press.

Dorfman, J. 1949. *The Economic Mind in American Civilization, vol. 3: 1865–1918*. New York: Viking Press.

Douglas, P.H. 1927. The modern technique of mass production and its relation to wages. *Proceedings of the Academy of Political Science* 12, 17–42.

Fisher, I. 1911. *The Purchasing Power of Money*. New York: Macmillan.

Fisher, I. 1923. The business cycle largely ‘a dance of the dollar’. *Journal of the American Statistical Association* 18, 1024–28.

Fisher, I. 1925. Our unstable dollar and the so-called business cycle. *Journal of the American Statistical Association* 20, 179–202.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Foster, W.T. and Catchings, W. 1923. *Money*. Boston: Houghton Mifflin.

Foster, W.T. and Catchings, W. 1925. *Profits*. Boston: Houghton Mifflin.

Foster, W.T. and Catchings, W. 1928. *The Road to Plenty*. Boston: Houghton Mifflin.

Furner, M.O. 1975. *Advocacy and Objectivity: A Crisis in the Professionalization of American Social Sciences, 1865–1905*. Lexington, KY: University of Kentucky Press.

Hadley, A. 1885. *Railroad Transportation*. New York and London: G.P. Putnam.

Herbst, J. 1965. *The German Historical School in American Scholarship: A Study in the Transfer of Scholarship*. Ithaca, NY: Cornell University Press.

Hodgson, G. 2004. *The Evolution of Institutional Economics: Agency, Structure and Darwinism in American Institutionalism*. London and New York: Routledge.

Laidler, D.W. 1999. *Fabricating the Keynesian Revolution: Studies of the Inter-War Literature on Money, the Cycle, and Unemployment*. Cambridge: Cambridge University Press.

Laughlin, J.L. 1885. *The History of Bimetallism in the United States*. 4th edn, New York: Appleton, 1898.

Marty, M. 1986. *Protestantism in the United States: Righteous Empire*. New York: Scribners.

Mehrling, P. 1997. *The Money Interest and the Public Interest: American Monetary Thought, 1920–1970*. Cambridge, MA: Harvard University Press.

Morgan, M.S. 1990. *A History of Econometric Analysis*. Cambridge: Cambridge University Press.

Morgan, M.S. and Rutherford, M. 1998. *Interwar Pluralism to Postwar Neoclassicism*. Durham, NC: Duke University Press.

Ross, D. 1991. *The Origins of American Social Science*. Cambridge: Cambridge University Press.

Rutherford, M. 1999. Institutionalism as ‘scientific’ economics. In *From Classical Economics to the Theory of the Firm: Essays in Honour of D.P. O’Brien*, ed. R.E. Backhouse and J. Creedy. Cheltenham: Edward Elgar.

Rutherford, M. 2000. Understanding institutional economics: 1918–1929. *Journal of the History of Economic Thought* 22, 277–308.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Rutherford, M. and DesRoches, C.T. 2008. The institutionalist reaction to Keynesian economics. *Journal of the History of Economic Thought*.

Snyder, C. 1924. New measures in the equation of exchange. *American Economic Review* 14, 699–713.

Taussig, F. 1888. *Tariff History of the United States*. New York and London: G.P. Putnam & Sons.

Yonay, Y. 1998. *The Struggle for the Soul of Economics: Institutionalist and Neoclassical Economists in America between the Wars*. Princeton, NJ: Princeton University Press.

How to cite this article

Bateman, Bradley W. "United States, economics in (1885–1945)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000069> doi:10.1057/9780230226203.1766

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

United States, economics in (1945 to present)

Roger E. Backhouse

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

After 1945, American economics was transformed as radically as in the previous half century. Economists' involvement in the war effort compounded changes that originated in the 1930s to produce profound effects on the profession, and many of these were continued through institutions that developed during the Cold War. This article traces the way the institutions of the profession interacted with the content of economics to produce the technical economics centred on a core of economic theory and econometric methods that dominate it today. Attention is also drawn to the broader role of American profession in economics outside the United States.

Keywords

Allied Social Science Association; American Economic Association; American Enterprise Institute; American Finance Association; Austrian economics; axiomatics; Banking School; behavioural economics; bounded rationality; Brookings Institution; Chicago School; complexity theory; Council of Economic Advisers (USA); Cowles Commission; Currency School; development economics; Econometric Society; econometrics; evolutionary game theory; experimental economics; Federal Reserve System; formalism; Foundation for Economic Education; Friedman, M.; game theory; general equilibrium; Harvard University; Heritage Foundation; heterodox economics; history of economic thought; industrialism; institutional economics; International Monetary Fund; IS–LM model; Keynesian revolution; Keynesianism; Koopmans, T. C.; Liberty Fund; macroeconometrics; macroeconomics, origins and history of; marginalist controversy; Massachusetts Institute of Technology; mathematical economics; mathematics and economics; microeconomics; microfoundations; Models; monopolistic competition; Mont Pèlerin Society; Nash, J.; national accounting; old institutionalism; operations research; perfect competition; positive economics; Post Keynesian economics; preference reversals; probability distributions; public choice; radical economics; RAND Corporation; rational choice; Rockefeller Foundation; Schultz, H.; simultaneous equations; statistics and economics; Stiglitz, J.;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
systems analysis; Union of Radical Political Economy; United States, economics in; Vining, R.; von Neumann, J.; World Bank

Article

The effect of the Second World War

The Second World War is more than a conventional dividing line, for it profoundly affected the course of economics in the United States. The interwar period had been one of pluralism within economics: there was a variety of competing approaches to the subject, none of which was dominant. It was also comparatively easy to discern distinctively ‘American’ trends in economics, which could be related either to the intellectual environment (notably the pragmatism of C.S. Peirce, William James and John Dewey) or to economic circumstances (such as the recent establishment of the Federal Reserve System). Within a decade of the Second World War, if not earlier, this had changed dramatically. Economics was becoming more technical, the foundations of an orthodoxy were being laid, and the position of the United States in relation to other countries was changing. The conventional explanation is the Keynesian revolution, reinforced by the rise of mathematical economics, but there is much more to the story than that.

The key to this picture is the so-called ‘old’ institutionalism. In the interwar period, Institutionalism was a very broad movement aimed at making economics more scientific through placing it on firmer empirical foundations. Though its boundaries were very blurred, it is reasonable to see it as covering economists as diverse in their empirical work as Wesley Mitchell, Simon Kuznets, Gardner Means, John Commons and John Maurice Clark. Though they had connections with economists in Europe, it was a distinctively American movement. Mathematical economics is inherently less culturally specific, but even here there were distinctive American approaches to the subject: Paul Samuelson's early work, under E.B. Wilson, drew on a type of mathematics very different from that used by Europeans. The same could be said about the early econometricians, from Henry Ludwell Moore to Henry Schultz: there were important European parallels, but they were pursuing research in a way that was distinctive. Monetary economics illustrates both the distinctiveness of American economics and the blurred boundaries between different approaches to the subject. Because the Federal Reserve System had been established much later than the major European central banks, there was much more lively debate over the principles on which it should be run. The result was a rich mixture of arguments spanning the divides between neoclassical and institutionalist, Harvard and Chicago, Banking School and Currency School.

The Second World War was important for several reasons. First, economics became tied up with the war effort. Economists were clearly involved in places such as the Office of Price Administration, the Treasury or the War Production Board. It was in the last of these that national income statistics, first calculated in 1933, were developed into a system of national accounts, providing the basis for planning the massive shift of resources from civilian to military production that took place after 1941. However, perhaps more significantly, economists became involved in fighting the war, primarily through the Office of Strategic Services (OSS), forerunner of the Central Intelligence Agency which employed around 50 economists under Edward Mason in its Research and Analysis division (Leonard, 1991; Katz, 1989). In the OSS, economists and other social scientists were employed alongside physicists and other

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
scientists in tasks where economics shaded imperceptibly into statistics and engineering. They analysed intelligence, and became intimately involved in questions of military strategy and tactics, emerging from the war with an enhanced reputation: not only was economic analysis important, but many economists had proved themselves useful as general problem solvers. Operations research, a set of techniques centered on optimization subject to constraints, came to be much more central to economics.

At the end of the war, the Servicemen's Readjustment Act (1944), the so-called G.I. Bill, offered financial support to US ex-servicemen who wanted to continue their education. This fuelled a dramatic increase in the university system. The number of bachelor's degrees awarded in US higher education institutions, which had never risen above 187,000 before the war, rose to 271,000 in 1947–8 (317,00 if higher degrees are included too) and 432,000 in 1949–50, many of these choosing to study economics. This accelerated the generational shift that was taking place, providing academic openings for economists returning from government service to civilian life, many of these being in institutions that had not been prominent before the war. While some economics departments continued as before, there was a shift in the profession's center of gravity away from places such as Wisconsin (the leading centre for Institutionalism) towards ones like MIT, Berkeley and Stanford – other places, such as Columbia, Harvard, Chicago and Yale were important before and after the war (Barber, 1997; Backhouse, 1998).

The subject began to be taught using textbooks written by young economists (Kenneth Boulding, Lorie Tarshis and Paul Samuelson) during the 1940s, in place of ones that had their origins nearer the turn of the century.

The Cowles Commission and RAND

A particularly important center for quantitative work in the 1940s was the Cowles Commission, which had moved to Chicago in 1938, and of which Jacob Marshak became Research Director in 1943. He laid out a programme of research focusing on the development of new methods to take account of the specific features of economic data, perceived to be simultaneity, the importance of random disturbances and the prevalence of aggregate time-series data. This programme proved to be one that attracted American economists, many of whom were involved in the war effort, and many of the highly technical European emigrés such as Trygve Haavelmo, Tjalling Koopmans and Abraham Wald. Not surprisingly, the operations research side of economics was dominant here, not simply in obvious ways, such as the work by Koopmans and George Dantzig on linear programming and the simplex method, but in the broader conception of economics as engineering. This was not confined to the Cowles Commission (one can see such an influence at other places such as Massachusetts Institute of Technology, MIT) but such work clearly centred on Cowles.

The important idea that emerged from this phase of the Cowles Commission's work was that the economic system could be analysed as a probability distribution, the task of economics being to identify the properties of that distribution. General equilibrium theory, embracing individual optimization within a system of simultaneous equations, provided an account of the structural relationships. Statistical methods, pioneered by Haavelmo and Koopmans, provided the means for relating that theory to data that exhibited random shocks in addition to any systematic relationships between the data, not just estimating coefficients but also testing the theory. This has been called a probabilistic revolution (Morgan, 1991).

Controversy over these new methods erupted in 1947 when Koopmans (1947) challenged, head on, what had previously been considered the scientific way to do empirical economics – the National Bureau of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Economic Research's (NBER) meticulous data-gathering and comparatively informal data-analysis – represented by *Measuring Business Cycles*, by Arthur Burns and Wesley Mitchell (1946). As Rutledge Vining (1949), replying for the NBER, justly claimed, Koopmans had written a manifesto for the Cowles Commission's new methods, privileging the testing of theory over the search for hypotheses. In addition to the techniques mentioned above, its fruits ranged from the monetary theory that formed the heart of Don Patinkin's *Money, Interest and Prices* (1956), the leading exposition of macroeconomic theory till the 1970s, and Lawrence Klein's models of the US economy, from which developed much of macroeconometrics.

In the late 1940s, the RAND Corporation, in Santa Monica, California, emerged as a new focus for technical economic analysis. RAND was initially a division of the Douglas Aircraft Company, but from 1948 it became a non-profit organization, funded at first by the US Air Force, and later by other bodies, of which the Ford Foundation was the most important. It was an interdisciplinary environment where economists worked alongside scientists, mathematicians, engineers and other social scientists. It was established by senior figures in the US Air Force and was motivated by the Soviet threat and the emerging Cold War, and embodied lessons learned in the Second World War. Through H. Rowan Gaither, Chairman of RAND's Board of Trustees and, from 1953, President of the Ford Foundation, RAND became closely linked to Ford: its overall product was ‘systems analysis’, a broad umbrella under which a range of mathematical work could be sponsored, linked primarily by certain sets of mathematical techniques and a vision of economics centered on rational choice. Though it was associated with much else, from Kenneth Arrow's *Social Choice and Individual Values* (1951) to *Linear Programming and Economic Analysis* by Dorfman, Samuelson and Solow (1958), its main significance was in game theory, bringing together economists from Cowles (such as Arrow) with economists and mathematicians from Princeton (which included John Nash), the major academic centre of research into game theory during the 1950s.

The precise significance of the military involvement in economics is not yet clear. The Office of Naval Research (ONR) provided much funding, especially for game theory research, and the US Air Force was behind RAND. Clearly some projects were directly driven by military imperatives, such as working out a strategy for responding to (or anticipating) a Soviet nuclear strike. There are also clear links from systems analysis/operations research, and the techniques associated with these, to military requirements. Against this, those involved emphasize that researchers at RAND were given great freedom and military sponsorship had little or no effect on what they did (see Mirowski, 2002). However, even if researchers did have a high degree of freedom, there was certainly selection bias in the types of projects and researchers who received support from these sources, and the scale of such funding makes it plausible to argue that may have had a significant effect on the way the profession developed.

Mathematics, technique and the ‘core’

To say that economics has become more mathematical in the post-war period is too obvious to need justification. However, the significance of this process and the way it came about are far less obvious and need disentangling.

Mark Blaug (1999) has labelled what happened to economics after the 1950s ‘the formalist revolution’. However, within this lie a number of very different developments. One is the incursion into economics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of formalist mathematics. At the outset of *The Theory of Value*, Debreu (1959) wrote that he was approaching his subject with the degree of rigour associated with the contemporary formalist school of mathematics. His work formed part of a broader movement towards placing economic theory on an axiomatic foundation, and comprising the literature on existence, uniqueness and stability of general equilibrium (see Weintraub 2002). Even here, however, it is possible to discern strands that on closer inspection are very different. Ingrao and Israel (1990) distinguished the formal and interpretive branches, associating the former with Debreu and the latter with Arrow. Others have traced differences to disputes over formalism in mathematics (Weintraub, 2002). The most eminent mathematician to engage with economics, John von Neumann, was not only a critic of formalism (in the sense of Hilbert): his interest in economics stemmed from a broader concern with artificial intelligence that, Mirowski (2002) has argued, differentiated his views sharply from economists at the Cowles Commission and others pursuing general equilibrium analysis.

More significant than this is the fact that most economics, as Solow (1997) has observed, is not formalistic in this sense. Rather, what has happened is that economics has become more ‘technical’: he was probably right to argue that axiomatics was of no interest to most economists. Perhaps the most influential exponent of mathematics in economics, Paul Samuelson, whose *Foundations of Economic* Analysis (1947), written at Harvard and the basis for the style of economics he established at MIT, amounted to a manifesto for mathematical economics. His work, which arose from a mathematical tradition very different from the European traditions out of which von Neumann and Debreu came, sought to be rigorous without being based on axiomatization. Even further from formalism, but equally influential, was the Chicago School, dominated from the 1940s to the 1970s by Milton Friedman. Friedman favoured simpler models and was more sceptical about complex mathematical reasoning. Thus Hands and Mirowski (1998) have distinguished three schools in post-war neoclassical price theory – Stanford (Arrow), MIT (Samuelson) and Chicago (Friedman). Whether or not one accepts the claim made by Hands and Mirowski that these represent three responses to the failure of Henry Schultz's attempt to quantify demand theory before his death in 1938, this provides a useful way to represent the variety of ways in which a common theoretical core was developed.

However, becoming more technical is not synonymous with using mathematics. Another dimension is the separation of theory and application. Though the distinction between theory and applied work is taken for granted by most contemporary economists, the situation was very different before the Second World War. There was much work where it is impossible to draw any distinction between statements that are intended to describe the world and ones that are at the level of theory. In what we would now consider applied work, the practice of clearly separating theory and application is something that emerged only after the Second World War (see Backhouse, 1998). This change is reflected in the language economists began to use: they began to talk in terms of models. Though the idea of a model has deeper roots, talking in terms of models took off only from 1939, having been very rare before that. As the discipline changed, so did the curriculum, something in which the American Economic Association (AEA) became involved. During the 1940s, partly in response to demands of wartime, and partly because of broader uncertainty about how economics should be taught, the AEA established committees on undergraduate education, the main outcome being a report in 1950. Out of this rose the suggestion to review graduate education, resulting in a report by Howard Bowen, sponsored by the AEA and funded by Rockefeller, which appeared in 1953. On the grounds that ‘technical knowledge’ of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economics would be useful for those working as economists in government, business and education, this argued that ‘there should be a “common core” for all students who are to be awarded advanced degrees in economics’ (Bowen, 1953, p. 2). This core consisted ‘primarily of economic theory including value, distribution, money, employment, and at least a nodding acquaintance with some of the more esoteric subjects such as dynamics, theory of games and mathematical economics’. No one, it was argued, had claim to an economics Ph.D. without ‘rigorous initiation’ into these and economic history, history of economic thought, statistics and research methods (Bowen, 1953, p. 43). Interestingly, mathematics was placed alongside Russian, German and Chinese: it was important to have some economists with knowledge of it, but it was not necessary for all to do so.

In Bowen's report, the core was still very broad – a statement of the range of knowledge – that economics Ph.D's should be expected to have. Over the following two decades it came to be used more narrowly. For example, Richard Ruggles (1962, p. 487) wrote of the function of graduate training being ‘to provide a common core of basic economic theory’ that would be used elsewhere in the programme, and observed that ‘at a great many universities’ training in mathematics was required. However, this was still discussed alongside language requirements. Though such questions had been raised as early as the Bowen Report, it was in the 1960s that the AEA meetings hosted debates over the role of economic history and history of economic thought in the graduate curriculum. Gordon (1965) conducted a survey implying, as Bowen had found a decade earlier, that though most graduate schools still offered the subject, history of economic thought was declining, and that there was pressure for it to decline further, particularly from younger faculty. In the survey by Nancy Ruggles (1970), the subject was defined in the now familiar way of a unifying core of micro and macro theory, quantitative methods (interestingly, econometrics, simulation, survey methods and operations research) and a range of applied fields that did not include any history.

These trends continued to the end of the century. They are best summed up by saying that economists were increasingly being trained, at Ph.D. level, as technicians rather than as scholars in the traditional sense of the term. In the 1940s, when concerns were raised about this in AEA meetings, it was still plausible to respond the demands of scholarship, and breadth of education, were compatible with mastering the necessary technical skills; but by the 1970s this was becoming more and more difficult. The demands of technique were pushing courses that provided breadth, symbolized by history of economic thought, out of the curriculum. By the end of the 1980s, this had gone so far that some Liberal Arts professors claimed that Ph.D.'s from the leading graduate schools were no longer equipped to teach at undergraduate level: not only did they know too little of the past and present literature on economics, but they did not know enough about the institutions of contemporary market economies. There were even signs that Ph.D. students themselves were sceptical about the value of the hurdles over which they were jumping (Colander and Klamer, 1987; 1990). In response to these concerns, the AEA established a Commission on Graduate Education in Economics, which reported in 1991 (Krueger, 1991; Hansen, 1991; see Coats, 1992a, for a comparison of this and Bowen's report). Though some changes were recommended, these were minor and had little effect (Colander, 1992). When Colander (2005) repeated his earlier survey a decade and a half later, he found much lower levels of dissatisfaction, though concluded that the students had adjusted to the more technical syllabus, not the other way round.

Economic analysis

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The way the use of mathematics spread within economics was inextricably linked with developments in economic analysis. It was not simply a matter of making earlier, less rigorous, analysis more precise. To be able to use the mathematical techniques in the way they did, the basis on which economics rested had to change. For the institutionalists, very broadly interpreted, being scientific meant basing economics firmly on evidence about how the world worked. It was because he believed that empirical research would establish accounts of human behavior that were more complex than those offered by economic theorists that Mitchell (1925) had predicted that economists would lose interest in an abstract, artificial man. The result was that the 1930s and 1940s saw a wealth of empirical work on industrial organization, pricing, labour markets and many other aspects of economic behaviour. But mathematical theory, given the techniques then available to economists, necessitated working with simpler assumptions, in which agents were maximizers operating in markets where competitive structures were precisely defined, and if possible were perfectly competitive. It is perhaps because of the strong institutionalist element in American economics that the debates through which these simplifying assumptions were established were dominated by American economists.

The clash between institutionalism and new, technical approaches explicitly came to the surface in Koopmans's review (1947) of *Measuring Business Cycles*, by Arthur Burns and Wesley Mitchell (1946). Koopmans presented his approach as building on the work of those like Burns and Mitchell who simply measured: it was necessary to pass beyond that to the ‘Newton stage’ in economic theorizing, where theory and data analysis informed each other. Data would be used to test theory. Representing the old view, Rutledge Vining (1949) pointed out that the Cowles Commission methods involved more than this: that they presumed a specific type of theory and empirical methods. If one did not know what theory was suitable, different empirical methods were required. Burns and Mitchell, Vining argued, were concerned with discovery as much as with testing, for, in the absence of empirical work, economists did not know what form theory should take.

Substantially the same issue arose in the so-called ‘marginalist’ controversy, provoked by Richard Lester's article in the *American Economic Review* (1946). Though Lester was portrayed by critics as drawing naive conclusions from surveys, and as presenting a radical challenge to profit maximization, he is better seen as arguing that economics needed to be based firmly on the mass of evidence that had been accumulated during the previous decade or more on how firms behaved and on how labour markets worked. Controversy here was more prolonged and more complex, spilling over into discussion of industrial organization, where Lester's critics included economists on both sides of the divide between Harvard (dominated by Mason and Chamberlin) and Chicago (where Friedman and Stigler were; see Lee, 1984; Mongin, 1992). Fritz Machlup changed the debate into one about marginal analysis and, together with Friedman, established the principle that economics was about explaining behavior, not explaining how decisions were made. Though he did not intend it that way, Friedman's (1953) methodology of positive economics, with its emphasis on testing predictions, not assumptions, could be taken as vindicating economic theory's neglect of its empirical foundations.

It is no coincidence that these controversies took place in the pages of US journals, as did a less prominent one a few years later on the role of mathematics in economics (Novick, 1954, and ensuing discussion; see Mirowski, 2002, pp. 402–5). There were parallels in other countries, but it was in the United States, where in the 1930s institutionalists and neoclassicals had vied with each other, that the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
most marked cleansing of approaches that could not be formalized so easily was taking place. In the 1950s and 1960s, formal modelling based on maximization and, increasingly, competitive markets spread throughout the discipline. Price theory became more formal and increasingly dominated what came to be known as microeconomics. Keynes's behavioural microfoundations (based on ‘propensities’ and imprecise generalizations from observed behaviour such as ‘animal spirits’) were replaced with optimizing ones and macroeconomics came to be seen through the lens of utility maximizing agents, as in Don Patinkin's *Money, Interest and Prices* (1956). General equilibrium analysis, summed up by Debreu's *Theory of* Value (1959), though never more than a minority activity, came to be seen as the fundamental theory on which more workaday theorizing rested.

During this period, however, there were limits to the application of formal theory. Though formal microfoundations could be provided for many of the functions, macroeconomics was seen as separate, not entirely reducible to a single, consistent microeconomic theory. In microeconomics, strategy and industrial structure remained outside the purview of formal theory, empirical work dominating work on industrial organization. Development economics offers another example of a field that stood apart from other fields, reflecting the assumption, held widely though not universally, that people in different societies behaved in different ways. Thus, although economists later came to see the rise of formal theory and mathematical methods as the key development during the period, its progress was slow, and it was anything but pervasive as late as the 1960s. The way in which less formal approaches, based on assumptions that ran directly counter to those underlying what later became dominant, is nicely illustrated by a project entitled ‘The Inter-University Study of Labor Problems and Economic Development’, undertaken by John Dunlop, Clark Kerr, Frederick Harbison and Charles Myers (see Cochrane, 1979). Its thesis, that industrialism required the development of a new type of man, ensured that, as the assumptions underlying modern theory became established, it came to be seen as a quaint relic of the past. However, its significance rests in its being a large project, lasting over 20 years, receiving $855,000 from the Ford Foundation and $200,000 from Carnegie, producing around 40 books and, in Cochrane's view, helping to define labour economics as that field existed in the late 1970s. Though its final report came as late as 1975, its objectives were framed against the background of thinking on labour questions that the older institutionalist economists would have understood; its analysis drew on sociology and industrial relations as well as on what would now be recognized as properly economic analysis.

However, from the 1970s things changed. Formal methods, based on individual optimization, were used to analyse problems of uncertainty and information, the most prominent exponent of this being Joseph Stiglitz. These ideas were applied to labour markets, finance, and many other fields. Macroeconomics turned away from what Lucas called ‘free parameter’ models – ones containing parameters that were not based on optimizing behaviour. Public choice theory, which emerged at the boundaries of economics and political science, brought government and much organizational behaviour within the scope of rational choice (see Medema, 2000). Initially, this was not widely considered to be economics, with the result that public choice scholars found it hard to publish in the major journals, leading James Buchanan, Gordon Tullock and their associates to establish the Public Choice Society, and to develop their own journals. However, fairly soon the main economics journals opened up to such work. Methods were found to build models of general equilibrium with monopolistic competition, these enabling trade theory to move away from assumptions of perfect competition that were thought unrealistic in many contexts. Game theory was introduced to analyse problems of strategy, first transforming industrial organization

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
and later being extended to almost every other field of economics. Development economics, like macroeconomics, ceased to be considered as resting on principles different from those in the rest of the discipline. Rational choice methods were even applied by economists with clear radical sympathies, such as in the rational choice Marxism of John Roemer and Jon Elster.

Most of these developments were international in their reach, but all were centred, squarely, in the United States. In none of the cases just mentioned would it be conceivable to write the story without discussing the role of American economists and economists based in the United States, whereas it would be possible (if not always the full picture) to do so without mentioning work in the rest of the world. The collective effect of these developments was to transform a discipline in which, though rational, maximizing behaviour was central, numerous exceptions and special cases existed, to one where it could plausibly be argued that economic theory was simply working out the implications of maximizing behaviour. Economic theory could be seen as resting on a single behavioural postulate.

Faced with this scenario, in which economic theory in the United States became, methodologically, narrower, some economists rebelled. Radical economists, stimulated by the Vietnam War, began to argue in the late 1960s, that economists were systematically ignoring questions such as power, class, and income distribution. Frustrated by their inability to persuade their more orthodox colleagues to take their ideas sufficiently seriously, they formed the Union of Radical Political Economy, setting up networks, conferences and a journal (Coats, 1992b; 2001). Shortly afterwards, inspired by Joan Robinson's Ely Lecture at the AEA in 1971, Alfred Eichner, Jan Kregel and others organized what developed into the grouping known as Post Keynesian economics (Lee, 2000). ‘Austrian economists’, influenced by the work of Ludwig von Mises, Friedrich Hayek, and Ludwig Lachmann, encouraged by Hayek's being awarded the Nobel Memorial Prize in 1974, also began to organize themselves. In all cases, the organization of these groups was motivated by the sense of exclusion they felt from the mainstream, represented by the meetings of the American Economic Association and the leading economics journals, who regarded their work as generally of low quality. These movements remained small, with strengths in particular institutions (Austrians at New York and Auburn; Post Keynesians at Rutgers and Tennessee; Public Choice in Virginia; Radicals at Amherst and the New School).

These self-consciously ‘heterodox’ groups were but part of a wider fragmentation of the discipline. Technological changes meant that economic print runs for books and journals fell during the period, and the costs of travel and communications fell. Together with the increased size of the economics profession, these developments made it easier for sub-fields of economics to organize, represented most clearly by the rapid rise in the number of specialist journals. The changing character of the profession was reflected in the Allied Social Science Association (ASSA), the main professional meeting of American economists, organized by the AEA. By 1998, though the AEA, the American Finance Association and the Econometric Society dominated the meetings, there were 52 societies affiliated with the ASSA (compared with 34 in 1980), and the AEA was having to restrict the number of sessions these societies were organizing, something it had not done two decades before.

The 1970s and 1980s were arguably years of integration, when the American economics became more homogeneous, the core of microeconomics based on individual optimizing behaviour being applied to more and more. More than ever before, and in dramatic contrast with the situation before the Second World War, there was an orthodoxy. However, this was questioned and developed at both empirical and theoretical levels. One reason was that developments in data collection and in computing meant that economists were able to analyse the behaviour of real individuals in a way that economists of earlier

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

generations (see, for example, Mitchell 1925) could do little more than dream about:

'microeconometrics' was recognized with the 2000 Nobel Memorial Prize for James Heckman and Daniel McFadden. It became possible to engage in quantitative analysis of microeconomics as never before. Another reason was that economists turned to experimentation as a source of data: experimental economics, considered esoteric as late as the 1970s, was given respectability with the debates over preference reversals in the pages of the *American Economic Review* and the award of the 2002 Nobel Memorial Prize to Daniel Kahneman and Vernon Smith. This rapidly spread throughout the profession. When 'behavioural' economics started being taken seriously in finance, a field where predictive power was always paramount, it was a sign that alternatives to conventional views of rationality were being taken very seriously. Bounded rationality, on which Herbert Simon had been working since the 1950s at Carnegie Mellon, and for which he got the Nobel Memorial Prize in 1978, moved from being something idiosyncratic, if respected, to being a mainstream technique. Evolutionary game theory and complexity theory offered new ways to think about economic change that expanded the boundaries of what was accepted in the subject. By the end of the century, though rational choice models remained immensely strong, it became much harder to describe economics as dominated by an orthodoxy. Once again, though these developments were international in their scope, they were centred on the United States, just as were the developments of the 1970s. Ideas whose main supporters were European, such as the competing views of consumer theory associated with Werner Hildenbrand (who derived demand functions from assumptions about distributions of characteristics across individual consumers), had far less influence.

Economists, ideology and policy

Ideology was never far from the surface. In the 1940s concern with 'Reds' was common in the United States, though economists might consider the problem only to dismiss it. After 1945, as the Cold War developed, these concerns with Communism grew, reaching their peak with Joseph McCarthy's search for Communist sympathizers. Economists had frequently been viewed with suspicion amongst businessmen, some of whom were important patrons of higher education, but the stakes were raised. Planning was suspect, a legacy from the days of the New Deal, and Keynes provided a convenient focus, for he was a more real threat than Marx: according to the *Chicago Tribune*, he was the Englishman who ruled America (see C.D.W. Goodwin, 1998, on these episodes). Influential figures argued that Keynesianism was tantamount to Communism. Textbooks, such as those of Lorie Tarshis and Samuelson, that discussed Keynesian theory were attacked and sometimes removed from syllabi under pressure from aggrieved sponsors (Colander and Landreth, 1996; 1998).

The cases where economists were forced out of academic positions because of real or alleged Communist sympathies are comparatively easy to document (Goodwin, 1998; Lee, 2004). What is much harder to prove is the effect this had on how economics pursued their work. There were certainly great pressures to be technical, for arcane communications between specialists were much less likely to be considered suspect than ideas that reached out beyond academia. Using an evolutionary model, Goodwin (1998, p. 79) distinguishes between 'conceptual variation' and 'intellectual selection', arguing that the attitudes of economists' patrons must have influenced the latter. However, doubts about its closeness to communism did not prevent Keynesianism from becoming widely accepted in academia, though that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
may have contributed to its being expressed in more careful, technical language than might otherwise have been the case (see Colander and Landreth, 1996, p. 172).

This bias towards becoming more technical chimed with another pressure – to be seen as doing science. When the National Science Foundation was established in 1950, the inclusion of social science was controversial and did not take place till 1956. If economists were to obtain support, they had to ensure that their work was seen as scientific. Given prevalent beliefs about science at the time, this favoured narrower, more technical work, and worked against the pluralistic interdisciplinarity that had been more common before the war (Goodwin, 1998, pp. 65–7). Similar issues arose in the context of support by philanthropic foundations, of which Sloan, Russell Sage, Rockefeller and Ford were the most important. Here, concern with being rigorous was intertwined with suspicion of planning and doubts about Keynesianism.

Similar considerations affected the body that brought economists into the heart of the US government, the Council of Economic Advisers (CEA) established by the Employment Act of 1946. This was intended to be a conservative institution, providing expert advice with minimal government interference. Its first chair, Edwin G. Nourse, shared this view: unlike his colleagues on the CEA, he was careful to avoid being seen as an advocate for White House policy (Bernstein, 2001, pp. 110–11). Unlike his successor, Leon Keyserling, he viewed economics as providing technical, disinterested expertise. Despite criticism, the CEA survived, achieving its greatest influence in the Kennedy administration, when Walter Heller, Kermit Gordon and James Tobin applied Keynesian demand-management policy to the problem of reducing unemployment.

Given that President Lyndon Johnson would not let it compromise his Great Society program, the escalating war in Vietnam led to rising federal deficits. CEA members warned the President about the consequences of this, but the CEA's Keynesian policies were nonetheless blamed for the inflation and dislocation that followed during the 1970s. After 1979, alongside the decline of Keynesianism in academia, influence on stabilization policy rapidly shifted to the Federal Reserve under Paul Volcker and later Alan Greenspan. This shift from the CEA to the Fed marked not a decline in the influence exerted by economists, but a change in its structure: there was a convergence between research done in academia and in central banks and other agencies (see McCallum, 2000, p. 123), and a shift of emphasis towards microeconomic policy. Economists increasingly saw their role, not as engineers advising on how to operate fiscal and monetary levers, but as designers of institutions and of systems that would achieve desired outcomes in a world where policymakers were seen as part of the system rather than outsiders manipulating it. Frequently this involved creating new markets, or ‘reinventing the bazaar’ (McMillan, 2002).

There were also more conscious attempts to impose an ideological agenda on economics. RAND, the most influential think tank in the 1950s, became closely involved with the Ford Foundation (these arguments are developed in Amadae, 2003). It was explicitly a non-political organization, directed towards impartial research. However, under the chair of its board of trustees, H. Rowan Gaither, also president of the Ford Foundation after 1953, RAND focused on ‘systems analysis’, based on principles of rational action. Rational choice, central to the work of RAND since its inception, could be seen as providing, though its focus on the independent individual, a justification for a free society, and an alternative to Communist collectivism (see Amadae, 2003). RAND's ideology, like that of Ford, was one of technocratic management, by experts using rigorous quantitative techniques. This ideology became

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
prominent in government in the 1960s when applied by Robert McNamara, who came from the Ford Motor Company, as Secretary of Defense.

Others had a more explicit ideological agenda. The American Enterprise Institute (established 1943), The Foundation for Economic Education (1946) and the Liberty Fund (1960) were established specifically to propagate free-market ideas. There were followed in the 1970s by a series of think tanks specifically to develop and apply such ideas to policy. The Heritage Foundation (1973) was specifically seen as providing a counterweight to the Brookings Institution (established 1927), which had come to be seen as part of finely tuned liberal policymaking machine (liberal being understood in the American sense). The aim of its president, Edwin Feulner, was to create ‘a new conservative coalition that would replace the New Deal coalition which had dominated American politics for half a century’ (L. Edwards, 2005, p. 371). When Ronald Reagan took office, the Heritage Fund provided policy ideas ready to put before the new administration. Hayek, who had moved to Chicago in 1950, played a particularly influential role in stimulating such organizations, within the United States and elsewhere, being part of an influential network centred on the Mont Pèlerin Society, an international group of libertarian thinkers established in 1947, whose founders included four Chicago economists and representatives from the Foundation for Economic Education.

Businessmen and conservative foundations also sought to stimulate free market thinking within economics, many of them effectively targeting specific institutions and programs. Though tiny compared with the big foundations such as Rockefeller and Ford, the Volker Fund (which supported Hayek and Mises), the Earhart Foundation (with a programme of one-year fellowships), the Scaife, Bradley and Olin foundations (which between them targeted support at, *inter alia*, Chicago's Law and Economics programme, and various centres of public choice theory in Virginia) managed to achieve influence out of proportion to their size (see Backhouse, 2005).

The international dimension

After the Second World War, the United States dominated the economics profession. Not only had German economics been devastated by the Nazi Party, but the resulting emigration contributed enormously to the expansion of American economics. The United States was not the only home for German and other European exiles, many moving to Britain, but it received more than any other country. Britain experienced no such loss, but its university system was too small for it to be a serious rival. The result was that many ideas that had originated in Europe rapidly came to be associated with the United States. The clearest examples of this are general equilibrium theory and econometrics, where European emigrés, led by Jacob Marshak, were instrumental in developing ideas that rapidly lost any close connection with their European origins. For example, in the 1930s, general equilibrium analysis had been an almost exclusively Viennese phenomenon (in Karl Menger's seminar), whereas by the 1950s, not only had those who worked on it there (Wald and von Neumann) moved to the United States, but its leading practitioners were an American (Arrow) and a Frenchman (Debreu), but both based in the United States.

Another clear example of this process is Keynesian economics, central to the evolution of American economics from the 1940s to at least the 1980s. This clearly originated in Britain, and British economists such as John Hicks and James Meade played important parts in the subsequent Keynesian

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
revolution. However, Keynesianism was rapidly Americanized. The key figure here was Alvin Hansen, the force behind Harvard's fiscal policy seminar, and later author of the influential *A Guide to Keynes* (1953). As has been argued by Mehrling (1998), Hansen's ‘conversion’ did not involve a rejection of his earlier ideas; rather, Keynesianism provided a vehicle through which his ideas on policy, rooted in the American institutionalist tradition, could be developed. Lawrence Klein (1947) provided another interpretation of Keynesian economics, relating it to with the econometric approach emanating from the Cowles Commission. Samuelson (1948) integrated Keynesian ideas into a textbook aimed at American students. During the 1950s and 1960s, the most influential work on macroeconomics was, with few exceptions, undertaken in the United States. Friedman's work on the consumption function provides another example of Keynesian ideas being assimilated into an American tradition (the empirical studies of the NBER).

What was happening here is that economics was becoming more international, but centred on the United States, a development made possible by the openness of the American system at a time when the profession was expanding and opportunities for immigrants were great. The United States dominated, not simply because of its size, but because of its resources. During the interwar period, the Rockefeller Foundation had been instrumental in building up economics in key European institutions in Britain, Scandinavia and many other countries (cf. Goodwin, 1998). After 1945, given the close American involvement in Europe that resulted from the war and reconstruction, this influence increased, accelerated by the reduced cost of international travel. In country after country, the economics profession changed in several ways. Academic systems became more open and competitive, with increased emphasis on publication in journals. There was a movement away from publication in the native language towards publication in English. Journals moved away from being national organs to ones that published articles by economists from a wide range of countries. Graduate education moved towards the American model, away from the traditional European model of a major thesis, publishable as a book, towards a Ph.D. comprising advanced coursework and a short thesis that could be the basis for three journal articles. The mathematical demands made of students rose progressively. Many economists either undertook postgraduate study in the United States or spent sabbaticals in US universities.

The speed and extent of these changes varied enormously (see the case studies in Coats, 1997). For example, in the UK, the proportion of staff with a degree from an American university rose steadily from 1950 to the 1990s. The highest proportion was at the London School of Economics, where it reached 45 per cent by the mid 1990s, whereas in other universities it was only five per cent. In Belgium, CORE at the Université Catholique de Louvain was an important centre for economists with strong US connections. Similarly, there was variability in the speed with which Ph.D. requirements changed, some British universities adopting the American model in the 1950s and 1960s, while others did not require any coursework beyond undergraduate level till the 1990s. In Continental Europe, there was the complication of language, and in many cases of academic systems that were much more rigid and less rapid to change, but many of these changes still took place. Outside Europe, there was the further factor of decolonization. At the end of the Second World War, many countries were still closely linked to former colonial powers, and the changes involved a switch from those to the United States. There is dispute over whether this process should be labelled ‘Americanization’ or simply ‘internationalization’ (see Coats, 1997, pp. 395–9). The process certainly did involve

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
internationalization, and it was arguable that many changes (such as the move towards advanced coursework) were necessitated by the rising technical demands made by the subject. As has already been explained, many of the ideas on which the period's economics was based were European, not American in origin. Arguably, the United States appeared to dominate what was primarily an international system simply because of its size. However, there are strong reasons for considering the process as involving Americanization. In many cases, in Europe and elsewhere, the United States provided an example that was deliberately copied. In other cases, changes were brought about through connections with US universities. Harry Johnson, Canadian, but a professor at Chicago and Geneva, was important in bringing about changes at LSE where he also held a chair in the 1960s. Chicago economists developed close links with Latin American countries, consciously exporting Chicago economics to Chile: Chilean students studied in Chicago, and Chicago staff taught at the Catholic University of Chile (see A. Harberger, 1997; Valdes, 1995). Similar developments took place in Brazil, though involving a much wider range of universities: Chicago, Berkeley, Harvard, Yale, Michigan, Illinois and Vanderbilt (see M. R. Loureiro, 1997). The US Agency for International Development and the Ford Foundation provided a significant role in funding several of these inter-university agreements.

Similar remarks could be made about the US role in the international organizations that emerged after 1945, notably the International Monetary Fund (IMF) and the World Bank: they were vehicles for the internationalization of economics, along a model dominated by the United States.

Conclusions

Since the Second World War, economics in the United States has been transformed as dramatically as in its previous half-century. It is inevitable that economists looking at these changes focus on the economic ideas themselves, usually telling the story as one of progress. However, this transformation involved changes in the structure of the profession (notably in the nature of graduate education) and its place in society as well as changes in economic analysis. It is natural for historians to focus on connections between economic ideas and the institutions out of which they arose. To do this raises the question of whether these external factors influenced the course of economic ideas: of whether things could have been different. The difficulty here is that it is hard to construct a plausible account of how things might have been different because, even if it was the result of adaptation to chance events, what actually happened generally looks inevitable in retrospect. However, what can be done is to sketch some of the possible routes that could have been taken but were not.

Different paths were open for American economics in the 1930s. During the New Deal, economists such as Gardner Means had built up an enormous body of statistical data on how product and labour markets operated. From this starting point, economists could have chosen to build models that were less general but more securely rooted in specific institutional detail than was Walrasian general equilibrium theory. The route the discipline actually took was determined, *inter alia*, by the Second World War and the Cold War and the encouragement it gave to certain types of theorizing and certain types of empirical work.

Macroeconomics offers a second account of alternative paths that might have been taken. The interwar literature contained discussions of rational expectations, dynamics, intertemporal equilibrium and credibility of policy regimes (see Backhouse and Laidler, 2004). However economists did not pursue such ideas but developed a macroeconomics centred on a static equilibrium framework (the IS–LM

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

model); much of the dynamics that had been lost was then 'rediscovered' in the 1970s. Had they developed a different set of ideas from the interwar literature (even from Keynes's own work), macroeconomics could not have developed as it did.

To argue in this way is not to claim that one path was right and the other wrong, or even that economists were aware of the directions in which their own theoretical choices (aimed at solving specific problems) would lead. Instead, historical accounts generally rest on two pillars. First, the standards by which economists judge their work – their standards of scientific rationality – reflect the intellectual climate in which they are working. In some cases, likely factors can be identified. For example, Weintraub (1998) has argued that conceptions of what it meant to be rigorous changed dramatically as a result of developments in quantum mechanics. But many of the influences on the criteria economists use to assess their work are harder to identify and have to be disentangled, cautiously, out of the historical record. Second, evolution requires not simply a mechanism for generating new ideas but also a selection process. To understand the way economic ideas have developed since the Second World War, it is necessary to consider the demand for economic ideas as well as the supply: thus, even if the economists are resolutely impartial, applying high scientific standards to their work, the identities and view of their patrons may serve, through favouring some types of inquiry rather than others, to affect the evolution of the subject. We may be too close properly to understand many of the connections, but economics during this period, just as much as the economics of earlier centuries, cannot be divorced from the institutional setting in which it developed.

See Also

- United States, economics in (1776–1885)
- United States, economics in (1885–1945)

Bibliography

Amadae, S.M. 2003. *Rationalizing Capitalist Democracy*. Chicago: Chicago University Press.

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Backhouse, R.E. 1998. The transformation of U.S. economics, 1920–1960, viewed through a survey of journal articles. In Morgan and Rutherford (1998), pp. 85–107.

Backhouse, R.E. 2005. The rise of free-market economics: economists and the role of the state since 1970. In *The Role of Government in the History of Economic Thought*, ed. S.G. Medema and P. Boettke. Durham, NC: Duke University Press. *History of Political Economy* 37 (Annual Supplement).

Backhouse, R.E. and Laidler, D.E.W. 2004. What was lost with ISLM. In *The IS–LM Model: Its Rise, Fall, and Strange Persistence*, ed. K.D. Hoover and M. De Vroey. Durham, NC: Duke University Press. *History of Political Economy* 36 (Annual Supplement).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Barber, W.J. 1997. Postwar changes in American graduate education in economics. In Coats (1997).

Bernstein, M. 2001. *A Perilous Progress: Economists and Public Purpose in Twentieth Century America*. Princeton, NJ: Princeton University Press.

Blaug, M. 1999. The formalist revolution or what happened to orthodox economics after World War II? In *From Classical Economics to the Theory of the Firm: Essays in Honour of D.P.O'Brien*, ed. R.E. Backhouse and J. Creedy. Cheltenham: Edward Elgar.

Bowen, H.R. 1953. Graduate education in economics. *American Economic Review* 43(4, part 2), 1–223.

Burns, A. and Mitchell, W.C. 1946. *Measuring Business Cycles*. New York: NBER.

Coats, A.W. 1992a. Changing perceptions of American graduate education in economics, 1953–1991. *Journal of Economic Education* 23, 341–52.

Coats, A.W. 1992b. Economics in the United States, 1920–1970. In *On the History of Economic Thought: British and American Economic Essays*, vol. 1, ed. A.W. Coats. London: Routledge.

Coats, A.W., ed. 1997. *The Post-1945 Internationalization of Economics*. Durham, NC: Duke University Press. *History of Political Economy* 28 (Annual Supplement).

Coats, A.W. 2001. The AEA and the radical challenge to social science. In *Economics Broadly Considered: Essays in Honor of Warren J. Samuels*, ed. J.E. Biddle, J.B. Davis and S.G. Medema. London: Routledge.

Cochrane, J.L. 1979. *Industrialism and Industrial Man in Retrospect: A Critical Review of the Ford Foundation's Support for the Inter-University Study of Labor*. Ann Arbor, MI: Ford Foundation, distributed by University Microfilms International.

Colander, D. 1992. The sounds of silence: the profession's response to the COGEE report. *American Journal of Agricultural Economics* 80, 600–7.

Colander, D. 2005. The making of an economist redux. *Journal of Economic Perspectives* 19(1), 175–98.

Colander, D. and Klamer, A. 1987. The making of an economist. *Journal of Economic Perspectives* 12 (3), 95–111.

Colander, D. and Klamer, A. 1990. *The Making of an Economist*. Boulder, CO: Westview Press.

Colander, D. and Landreth, H. 1996. *The Coming of Keynesianism to America: Conversations with the*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Founders of Keynesian Economics. Cheltenham: Edward Elgar.

Colander, D. and Landreth, H. 1998. Political influence on the textbook Keynesian revolution: God, man and Lorie Tarshis at Yale. In *Keynesianism and the Keynesian Revolution in America: A Memorial Volume in Honour of Lorie Tarshis*, ed. O.F. Hamouda and B.B. Price. Cheltenham: Edward Elgar.

Debreu, G. 1959. *The Theory of Value*. London: Wiley.

Dorfman, R., Samuelson, P.A. and Solow, R.M. 1958. *Linear Programming and Economic Analysis*. New York: McGraw Hill.

Edwards, L. 2005. *The Power of Ideas: The Heritage Foundation at Twenty-Five Years*. Ottawa, IL: Jameson Books.

Friedman, M. 1953. The methodology of positive economics. In *Essays in Positive Economics*. Chicago: Chicago University Press.

Goodwin, C.D.W. 1998. The patrons of economics in a time of transformation. In Morgan and Rutherford (1998).

Gordon, D.F. 1965. The role of the history of economic thought in the understanding of modern economic theory. *American Economic Review* 55, 119–27.

Hands, D.W. and Mirowski, P. 1998. A paradox of budgets: the postwar stabilization of American neoclassical demand theory. In Morgan and Rutherford (1998).

Hansen, A. 1953. *A Guide to Keynes*. London: McGraw Hill.

Hansen, W.L. 1991. The education and training of economics doctorates: major findings of the Executive Secretary of the American Economic Association's commission on graduate education in economics. *Journal of Economic Literature* 29, 1054–87.

Harberger, A. 1997. Good economics comes to Latin America, 1955–95. In Coats (1997).

Ingrao, B. and Israel, G. 1990. *The Invisible Hand: Economic Equilibrium in the History of Science*, trans. I. McGilvray. Cambridge, MA: MIT Press.

Katz, B. 1989. *Foreign Intelligence*. Cambridge, MA: Harvard University Press.

Klein, L.R. 1947. *The Keynesian Revolution*, 2nd ed. London: Macmillan, 1968.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Koopmans, T.C. 1947. Measurement without theory. *Review of Economics and Statistics* 29(3), 161–72.

Krueger, A.O. 1991. Report of the Commission on Graduate Education in Economics. *Journal of Economic Literature* 29, 1035–53.

Lee, F.S. 1984. The marginalist controversy and the demise of full-cost pricing. *Journal of Economic Issues* 18(4), 1107–32.

Lee, F.S. 2000. On the genesis of post Keynesian economics: Alfred S. Eichner, Joan Robinson and the founding of post Keynesian economics. *Research in the History of Economic Thought and Methodology* 18-C, 3–258.

Lee, F.S. 2004. To be a heterodox economist: the contested landscape of American economics, 1960s and 1970s. *Journal of Economic Issues* 38, 747–63.

Lester, R.A. 1946. Shortcomings of marginal analysis for wage-employment policies. *American Economic Review* 36, 63–82.

Leonard, R. 1991. War as a ‘simple economic problem’: the rise of an economics of defense. In *Economics and National Security: A History of their Interaction*, ed. C.D. Goodwin. Durham, NC: Duke University Press. *History of Political Economy 23* (Annual Supplement).

Loureiro, M.R. 1997. The professional and political impacts of the internationalization of economics in Brazil. In Coats (1997).

McCallum, B.T. 2000. Recent developments in monetary policy analysis: the roles of theory and evidence. In *Macroeconomics and the Real World. Volume 1: Econometric Techniques and Macroeconomics*, ed. R.E. Backhouse and A. Salanti. Oxford: Oxford University Press.

McMillan, J. 2002. *Reinventing the Bazaar: A Natural History of Markets*. New York: W.W. Norton.

Medema, S.G. 2000. ‘Related disciplines’: the professionalization of public choice analysis. In *Toward a History of Applied Economics*, ed. R.E. Backhouse and J. Biddle. Durham, NC: Duke University Press. *History of Political Economy 32* (Annual Supplement).

Mehrling, P.G. 1998. *The Money Interest and the Public Interest: American Monetary Thought, 1920–1970*. Cambridge, MA: Harvard University Press.

Mirowski, P. 2002. *Machine Dreams: Economics Becomes a Cyborg Science*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mitchell, W.C. 1925. Quantitative analysis in economic theory. *American Economic Review* 15(1), 1–12.

Mongin, P. 1992. The ‘full-cost’ controversy of the 1940s and 1950s: a methodological assessment. *History of Political Economy* 24, 311–56.

Morgan, M. 1991. *A History of Econometric Ideas*. Cambridge: Cambridge University Press.

Morgan, M. and Rutherford, M., eds. 1998. *From Interwar Pluralism to Postwar Neoclassicism*. Durham, NC: Duke University Press. *History of Political Economy 30* (Annual Supplement).

Novick, D. 1954. Mathematics: logic, quantity and method. *Review of Economics and Statistics* 36, 357–8.

Patinkin, D. 1956. *Money, Interest and Prices*. Evanston, IL: Row, Peterson.

Ruggles, N., ed. 1970. *Economics*. Englewood, Cliffs NJ: Prentice Hall.

Ruggles, R. 1962. Relation of the undergraduate major to graduate economics. *American Economic Review* 52, 483–89.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. 1948. *Economics: An Introductory Analysis*. New York: McGraw-Hill.

Solow, R.M. 1997. How did economics get that way, and what way did it get? *Daedalus* 126, 39–58.

Valdes, J.G. 1995. *Pinochet's Economists: The Chicago School in Chile*. Cambridge: Cambridge University Press.

Vining, R. 1949. Koopmans on the choice of variables to be studied and of methods of measurement. *Review of Economics and Statistics* 31(2), 77–86.

Weintraub, E.R. 1998. Axiomatisches missverstaendniss. *Economic Journal* 108, 1837–47.

Weintraub, E.R. 2002. *How Economics became a Mathematical Science*. Durham, NC: Duke University Press.

How to cite this article

Backhouse, Roger E. "United States, economics in (1945 to present)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009
<http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000076>
doi:10.1057/9780230226203.1767

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Uno, Kozo (1897 – 1977)

T. Sekine

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Marxian value analysis; stages theory of development; Uno, K.; use value

Article

A prominent Japanese Marxian economist known especially for his rigorous and systematic reformulation of Marx's *Capital*. Born in Kurashiki in western Japan in a year of intense social unrest, Uno early took an interest in anarcho-syndicalism and Marxism. Not being of an activist temperament, however, he strictly disciplined himself to remain, throughout his life, within the bounds of independent academic work. For this deliberate separation of theory (science) from practice (ideology) he was frequently criticized. After studying in Tokyo and Berlin in the early 1920s, Uno taught at Tohoku University (1924–38), the University of Tokyo (1947–58) and Hosei University (1958–68). During most of the war years he kept away from academic institutions. He authored many controversial books, especially after the war. His eleven-volume *Collected Works* were published by Iwanami-Shoten in 1973–4.

The problem with Marx's *Capital*, according to Uno, is that it mixes the theory and history of capitalism in a haphazard fashion (described as ‘chemical’ by Schumpeter) without cogently establishing their interrelation. Uno's methodological innovation lies in propounding a stages theory of capitalist development (referring to the stages of mercantilism, liberalism, and imperialism) and using it as a mediation between the two.

Capitalism is a global market economy in which all socially needed commodities tend to be produced as value (that is, indifferently to their use-values) by capital. This tendency is never consummated since many use-values in fact fail to conform to this requirement. Only in theory, which synthesizes ‘pure’ capitalism, can one legitimately envision a complete triumph of value over use-values. The inevitable gap between history, in which use-values appear in their raw forms, and pure theory, in which they are already idealized as merely distinct objects for use, must be bridged by stages theory, which structures itself around use-values of given types (as ‘wool’, ‘cotton’, and ‘steel’ respectively typify the use-values of the three stages).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Uno's emphasis on ‘pure’ capitalism as the theoretical object has invited many uniformed criticisms. His synthesis of a purely capitalist society as a self-contained logical system follows the genuine tradition of the Hegelian dialectic, and is quite different from axiomatically contrived neoclassical ‘pure’ theory. Unlike the latter which takes the capitalist market for granted, Uno's theory logically generates it by step-by-step syntheses of the ever-present contradiction between value and use-values. The pure theory of capitalism is thus divided into the three doctrines of circulation, production, and distribution according to the way in which this contradiction is settled. By specifically articulating the abiding dialectic of value and use-values, already present in *Capital*, Uno has given Marxian economic theory its most systematic formulation, a formulation which militates against the two commonest Marxist errors known as voluntarism and economism.

Uno's approach is not dissimilar to Karl Polanyi's in appreciating the tension between the substantive (use-value) and the formal (value) aspect of the capitalist economy. Unlike Polanyi, however, Uno ascribes more than relative importance to capitalism, in the full comprehension of which he sees the key to the clarification of both pre-capitalist and post-capitalist societies. Thus Uno's approach reaffirms and exemplifies the teaching of Hegel (and Marx) that one should ‘learn the general through the particular’, and not the other way round.

Bibliography

Albritton, R. 1984. The dialectic of capital: a Japanese contribution. *Capital and Class* 22, 157–76.

Albritton, R. 1985. *A Japanese Reconstruction of Marxist Theory*. London: Macmillan.

Itoh, M. 1980. *Value and Crisis: Essays in Marxian Economics in Japan*. New York: Monthly Review Press.

Sekine, T.T. 1975. Uno-Riron: a Japanese contribution to Marxian political economy. *Journal of Economic Literature* 13, 847–77.

Sekine, T.T. 1984. *The Dialectic of Capital: A Study of the Inner Logic of Capitalism*. Tokyo: Toshindo Press.

Uno, K. 1980. *Principles of Political Economy: Theory of a Purely Capitalist Society*. Trans. from the Japanese by T.T., Sekine. Brighton: Harvester Press.

How to cite this article

Sekine, T. "Uno, Kozo (1897–1977)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000030>
doi:10.1057/9780230226203.1768

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban agglomeration

William C. Strange
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Urban agglomeration is the spatial concentration of economic activity in cities. It can also take the form of concentration in industry clusters or in employment centres within a city. One reason that agglomeration takes place is that there exist external increasing returns, also known as agglomeration economies. Evidence indicates that there exist both urbanization economies, associated with city size, and localization economies, associated with the clustering of industry. Both effects attenuate geographically. Theoretical research has identified many sources of agglomeration economies, including labour market pooling, input sharing, and knowledge spillovers. Empirical research has offered evidence consistent with each of these.

Keywords

input sharing; knowledge spillovers; labour market pooling; localization economies; migration; new economic geography; production functions; productivity; rent seeking; systems of cities; urban agglomeration; urban wage premium; urbanization economies

Article

Urban agglomeration is the spatial concentration of economic activity in cities. It can also take the form of concentration in industry clusters or in centres of employment within a city.
That both kinds of concentration exist is not debatable. Cities contain roughly 80 per cent of the US population, and urban population densities are approximately four times the national average. It is not just aggregate activity that is agglomerated; individual industries are concentrated too. There are many examples. Computer software is well-known for its spatial concentration, especially in the Silicon Valley. Automobile manufacturing, finance, business services, and the production of films and television programmes are other notable examples of industrial clustering. Agglomeration also takes place within cities in the form of densely developed downtowns and sub-centres. These patterns are not

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
unique to the United States. Capital and labour are highly agglomerated in every developed country, and they are increasingly agglomerated in the developing world.

Localization and urbanization economies

There are two sorts of agglomeration economies. Urbanization economies are associated with city size or diversity. Localization economies are associated with the concentration of particular industries. The idea that a city's size or diversity contributes to agglomeration economies is often attributed to Jacobs (1969), while the idea that industrial localization increases productivity goes back to Marshall (1890).

Economists have looked for evidence of these effects in a number of ways. Since agglomeration economies by definition enhance productivity, one natural approach is to estimate a production function. Estimating the production function requires establishment level data on inputs, including employment, land, capital, and materials. Data on labour is the easiest to obtain, although even the most detailed data-sets are incomplete, for instance by omitting experience in a particular occupation. Although data on purchased materials are sometimes available, data on internally sourced inputs typically are not. Measuring a firm's capital presents serious problems, including accounting for depreciation. Finally, even with good input data it is necessary to control for endogeneity of input use.

Despite the difficulties inherent in estimating a production function, a substantial body of research has estimated the impact of agglomeration on productivity. The very rough consensus is that doubling city size increases productivity by an amount that ranges from two to eight per cent. Some estimates are lower. The diversity of the local environment, another aspect of urbanization, has also been shown to be positively related to productivity. In addition, there is evidence of localization economies of roughly similar magnitude (see Rosenthal and Strange, 2004, for a review).

There are other ways to look for evidence of localization and urbanization economies. Glaeser and Mare (2001) identify the existence of an urban wage premium, with workers in cities of over a million residents earning roughly a third more in nominal wages than workers in cities of fewer than 100,000. Even after controlling for the selection of highly productive workers into cities, a significant premium remains. This is evidence of agglomeration economies because firms would not be willing to pay such premium in the absence of a corresponding productivity advantage. Rosenthal and Strange (2003) consider the arrival of new business establishments and find that diversity encourages arrivals and that localization economies are more important than urbanization economies for the industries examined. Finally, Henderson, Kuncoro and Turner (1995) analyse employment growth in the United States during the 1970–87 period. For mature industries, the specialization of employment at the metropolitan level is positively associated with growth. These papers are fairly typical of those that have measured the existence of agglomeration economies. The broad conclusion is that both urbanization and localization economies are present.

The sources of agglomeration economies: why do cities and clusters exist?

Marshall (1890) identifies three forces that can explain industry clustering: input sharing, labour market pooling, and knowledge spillovers. Input sharing exists when, for example, a clothing manufacturer in New York is able to purchase a great variety of relatively inexpensive buttons from a nearby company

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
that specializes in button manufacturing. Correspondingly, the button manufacturer also benefits because there will be many nearby clothing manufacturers to whom it can sell buttons. The process is, therefore, a circular one. Labour market pooling exists when a film production company in the Los Angeles area can quickly fill a position by hiring one of the many specialized production workers already present locally. Similarly, a specialized worker in Los Angeles can more easily find a new position without having to relocate. In both instances, labour pooling reduces search costs and improves match quality, providing valuable benefits for employers and workers. Knowledge spillovers exist when industrial engineers can learn the tricks of the trade from random interactions with other programmers in the same location. Any of these forces can explain industry clustering. They can also give rise to cities. The sharing of business service inputs can, for example, lead firms in very different industries to benefit from locating in close proximity to each other. Similar sorts of stories can be told about labour market pooling and knowledge spillovers.

Marshall's list is, of course, incomplete. Many other forces can lead to agglomeration. First, there is greater availability of consumer amenities in large cities. A major league sports franchise, for instance, requires a significant fan base in order to be economical. Second, natural advantage can explain both urbanization and localization. For instance, heavy manufacturing has historically developed near sources of minerals and where water transportation was possible. Third, internal economies of scale coupled with transactions costs can lead to self-reinforcing agglomeration. This explanation is the heart of the New Economic Geography (NEG, Fujita, Krugman and Venables, 1999), and it has received much attention in recent years.

Various approaches have been adopted in modelling agglomeration economies. Perhaps the simplest is to assume that there is some sort of public good that can be shared more economically in a larger city or cluster (Arnott, 1979). This force operates on both the production and consumption sides. Productivity is enhanced by infrastructure, and utility is increasing in public goods. An alternative is to assume that there are local externalities, with agents directly making their neighbours better off (that is, more knowledgeable). Agglomeration economies can also arise from thick market effects in search or matching (Helsley and Strange, 1990). The important common element from all these explanations is that agglomeration is associated with situations where market outcomes are not guaranteed to be efficient. Duranton and Puga (2004) provide a more detailed survey of the sharing, matching, and spillover micro-foundations of agglomeration.

The discussion thus far has suggested that agglomeration is always a positive outcome, at least as a second-best solution to market failures. This need not always be the case. Another reason to agglomerate is rent seeking. Ades and Glaeser (1995) show that there are many situations where urbanization can allow a city's residents to claim the output of other agents. They argue that imperial Rome supported a population of more than one million at least in part because the rewards of empire were distributed to residents of Rome as ‘bread and circuses’ in order to preserve domestic order. In this case, a city may exist, not because it adds to productivity, but because it allows redistribution.

Empirical work has provided evidence of the presence of Marshall's forces. Jaffe, Trajtenberg and Henderson (1993) provide direct evidence, showing that patent citations are geographically localized. Holmes (1999) considers local input sharing. He shows that more concentrated industries have a higher value of purchased input intensity, equal to purchased inputs divided by sales. This is consistent with the presence of input sharing. Costa and Kahn (2001) consider one aspect of labour pooling: matching

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
between workers and employers. The key result is that ‘power couples’, where both partners have at least a college degree, have disproportionately and increasingly located in large metropolitan areas. After considering other possible explanations, they conclude that power couples have become increasingly urbanized at least in part because it easier for both individuals to find good matches for their specialized skills. There is also evidence of the non-Marshallian agglomeration economies. Glaeser, Kolko and Saiz (2001) provide evidence of consumption effects. Kim (1995) documents the importance of natural advantage. Evidence of effects predicted by NEG is reviewed by Head and Mayer (2004). Looking across industries at the sorts of industry characteristics associated with agglomeration, Rosenthal and Strange (2001) find that all of the factors discussed above contribute to industrial agglomeration. The evidence is strongest for labour market pooling.

The geography of agglomeration economies: cities and neighbourhoods

It is common to consider agglomeration economies at the city level. This is because it is significantly easier to carry out estimation of production functions, wage premiums, births, or growth in that sort of aggregate analysis. In the previous section's analysis, however, it is clear that agglomeration economies depend on the distance between agents rather than on the political boundaries of cities. Thus, it makes sense to consider the degree to which agglomeration economies are at work at different levels of geography. Are they a neighbourhood effect or do they operate at the city level? In a sense, this question about the boundary of an agglomeration is parallel to asking about the boundary of a firm. The canonical question in the theory of the firm literature is: make or buy? Should an activity be carried out internally or through market transactions? The parallel agglomeration question is: near or far? Should one activity take place in close proximity to another or at a great distance?
Rosenthal and Strange (2003) address this issue by using geo-coding software to measure total employment and own-industry employment within a certain distance of an employer. Using these measures, the paper calculates the effects of the local environment on the number of firm births and on these new firms’ employment levels for six industries (computer software, apparel, food processing, printing and publishing, machinery, and fabricated metals). The key result is that agglomeration economies attenuate with distance. The effect of additional employment beyond five miles is shown to be roughly one-quarter to one-half of the effect of additional employment within a firm's own zipcode (postal code). This result is consistent with both the concentration of employment downtown and in sub-centres.

Agglomeration economies in a system of cities or regions

There has been considerable theoretical work on the implications of agglomeration economies for a system of cities or regions. Fujita and Thisse (2002) review and synthesize the literature. With apologies for oversimplification, the analysis in the literature proceeds as follows. First, it is assumed that there exists some sort of agglomeration economy. The specification may be of a reduced form shifting of the production function or of a particular agglomerative force. In the NEG literature (Fujita, Krugman and Venables, 1999), for instance, agglomeration arises from backward linkages between firms and input suppliers and forward linkages between firms and consumers. Second, equilibrium in the system is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
characterized. The equilibration always involves the individual location decisions taken by firms and households. It may also involve some coordination by a large agent, either a local government or a profit-maximizing city developer. Finally, the dynamic properties of the equilibrium are considered.
The systems of cities literature obtains several results. First, the characteristics of cities in an equilibrium system depend crucially on the sorts of agglomeration economies at work. If there are only localization economies, then the system will feature cities that specialize by industry. If there are in addition, or instead, urbanization economies, then diverse cities can arise. Second, there can be multiple equilibria. This means, for example, that there is no guarantee that the largest city in the middle of North America would be where Chicago is. Third, history matters. Agglomeration economies can be a conservative force in that they make it difficult for firms and workers to change their locations. Fourth, there is potential for catastrophic change, with a small change in parameters inducing a large change in outcomes. The attraction of other firms can cause an agglomeration to persist beyond the point at which it would have arisen from *de novo* location decisions. Eventually, however, when the attractiveness of other locations becomes sufficiently great, the agglomeration collapses suddenly. Fifth, equilibrium is not likely to be efficient. This result arises most starkly in a model where city sizes are determined solely by individual migration decisions. This ignores the existence of private developers and governments, which are both rewarded from realizing more efficient cities. However, in order to realize an efficient allocation, a developer would require unlimited control of city formation, a condition that is unlikely to obtain (Helsley and Strange, 1997).

See Also

- new economic geography
- spatial economics
- systems of cities
- urbanization

Bibliography

Ades, A. and Glaeser, E. 1995. Trade and circuses: explaining urban giants. *Quarterly Journal of Economics* 110, 195–227.

Arnott, R. 1979. Optimal city size in a spatial economy. *Journal of Urban Economics* 61, 65–89.

Costa, D. and Kahn, M. 2001. Power couples. *Quarterly Journal of Economics* 116, 1287–315.

Duranton, G. and Puga, D. 2004. Micro-foundations of urban agglomeration economies. In *Handbook of Urban and Regional Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland, 2063–2118.

Fujita, M. and Thisse, J. 2002. *The Economics of Agglomeration*. Cambridge: Cambridge University

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Press.

Fujita, M., Krugman, P. and Venables, A. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge: Cambridge University Press.

Glaeser, E. and Mare, D. 2001. Cities and skills. *Journal of Labor Economics* 192, 316–42.

Glaeser, E., Kolko, J. and Saiz, A. 2001. Consumer city. *Journal of Economic Geography* 1, 27–50.

Head, K. and Mayer, T. 2004. The empirics of agglomeration and trade. In *Handbook of Urban and Regional Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: Elsevier, 2609–2670.

Helsley, R. and Strange, W. 1990. Agglomeration economies and matching in a system of cities. *Regional Science and Urban Economics* 20, 189–212.

Helsley, R. and Strange, W. 1997. Limited developers. *Canadian Journal of Economics* 30, 329–48.

Henderson, J., Kuncoro, A. and Turner, M. 1995. Industrial development in cities. *Journal of Political Economy* 103, 1067–85.

Holmes, T. 1999. Localization of industry and vertical disintegration. *Review of Economics and Statistics* 81, 314–25.

Jacobs, J. 1969. *The Economy of Cities*. New York: Vintage.

Jaffe, A., Trajtenberg, M. and Henderson, R. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108, 577–98.

Kim, S. 1995. Expansion of markets and the geographic distribution of economic activities: the trends in U.S. regional manufacturing structure, 1860–1987. *Quarterly Journal of Economics* 110, 881–908.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Rosenthal, S. and Strange, W. 2001. The determinants of agglomeration. *Journal of Urban Economics* 50, 191–229.

Rosenthal, S. and Strange, W. 2003. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics* 85, 377–93.

Rosenthal, S. and Strange, W. 2004. Evidence on the nature and sources of agglomeration economies. In *Handbook of Urban and Regional Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Elsevier, 2119–2172.

How to cite this article

Strange, William C. "urban agglomeration." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000064>
doi:10.1057/9780230226203.1769

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban economics

John M. Quigley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Urban economics emphasizes: the spatial arrangements of households, firms, and capital in metropolitan areas; the externalities which arise from the proximity of households and land uses; and the public policy issues which arise from the interplay of these economic forces.

Keywords

cities; congestion; density gradient; diversity (ethnic and racial); endogenous growth; external economies; housing markets; human capital; labour markets; land markets; land use; monopolistic competition; patents; pollution; price discrimination; property taxation; regional economics; residential segregation; road pricing; Schelling, T.; social externalities; spatial competition; suburbanization; tipping point models; transport costs; urban agglomeration; urban consumption externalities; urban economics; urban production externalities; von Thünen, J.; zoning

Article

Cities exist because production or consumption advantages arise from higher densities and spatially concentrated location. After all, spatial competition forces firms and consumers to pay higher land rents – rents that they would not be willing to pay if spatially concentrated economic activity did not yield cost savings or utility gains. Economists have long studied the forces leading to these proximities in location, focusing first and foremost upon the importance of transport costs.

Early theorists (for example, von Thünen, as early as 1826; see Hall, 1966) considered land use and densities in an agrarian town where crops were shipped to a central market. Early models of location deduced that land closer to the market would be devoted to producing crops with higher transport costs and higher output per acre. Cities in the 19th century at this time were characterized by high transport costs for both goods and people, and manufactured goods were produced in close proximity to a central node – a port or a railway from which goods could be exported to world markets. The high costs of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
transporting people also meant that workers' residences were located close to employment sites. Transport improvements in the late 19th century meant that urban workers could commute cheaply by streetcar, thereby facilitating the suburbanization of population into areas surrounding the central worksite. More radical technical change in the first decades of the 20th century greatly reduced the cost of transporting materials and finished goods. The substitution of the truck for the horse and wagon finally freed production from locations adjacent to the export node. The introduction of the private auto a decade later further spurred the decentralization of US metropolitan areas.

Spatial forces

The seminal literature in urban economics provides positive models of the competitive forces and transport conditions which give rise to the spatial structure of modern cities. These models emphasize the trade-off between the transport costs of workers, the housing prices they face, and the housing expenditures they choose to make. Relatively simple models can explain the basic features of city structure – for example, the gradient in land prices with distance to the urban core; the house price gradient; the relationship between land and housing price gradients; the intensity of land use; and the spatial distribution of households by income (see Breuckner, 1987, for a review).
Empirical investigations of these phenomena reveal clearly that these gradients have been decreasing over time. Indeed, the flattening of price and density gradients over time has been observed in the United States since as long ago as the 1880s. (Early work is reported in Mills, 1972.) In interpreting these trends, it is important to sort out the underlying causes. The stylized model described above emphasizes the roles of transport cost declines (in part, as a result of technical change and the role of the private auto), increases in household income, and population growth in explaining suburbanization. These models also rely upon the stylized fact that the income elasticity of housing demand exceeds the income elasticity of marginal transport costs. The alternative, largely ad hoc, explanations stress specific causes, for example the importance of tax policies which subsidize low-density owner-occupied housing, the importance of neighbourhood externalities which vary between cities and suburbs, or the role of variations in the provision of local public goods. There is a variety of empirical analyses of the determinants of the variations in density gradients over time and space. A general finding is that levels and intertemporal variations in real incomes and transport costs are sufficient to explain a great deal of the observed patterns of suburbanization.

Durable capital

But, of course, variations in many of these other factors are highly correlated with secularly rising incomes and declining commuting costs, so any parcelling out of root causes is problematic. The elegant and parsimonious models of urban form have proven easy to generalize in some dimensions – for example, to incorporate stylized external effects and variations in income distributions across urban areas. It has proven to be substantially harder to recognize the durability of capital in tractable equilibrium models. The original models assumed that residential capital is infinitely malleable, and that variations in income or transport costs would be manifest in the capital intensity of housing over space in a direct and immediate way. The decline in land rents with distance from the urban centre means that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
developers' choices of inputs vary with capital-to-land ratios – declining with distance to the core. Dwellings are small near the urban core and large at the suburban fringe. Tall buildings are constructed near the urban centre, and more compact buildings are constructed in peripheral areas. But, of course, these structures and housing units are extremely durable, with useful lives exceeding 40 years or more. Thus, insights derived from the perspective in which the capital stock adjusts instantly to its long-run equilibrium position in response to changed economic conditions are limited.

Incorporating durable housing into models of residential location and urban form implies some recognition of the fact that 'history matters' in the structure and form of urban areas. Cities with the same distribution of income and demographics and with identical transport technologies may be quite different in their spatial structures, depending upon their historical patterns of development. Extensions of these simple models analyse the form of urban areas when developers have myopic or perfect foresight and when development is irreversible. With myopic developers, land is developed at each distance from the centre to the same density as it would have been built with malleable capital, but, once built, its capital intensity is frozen. Thus, with increasing opportunity costs of land over time, population and structural densities may increase with distance from the urban core.

With perfect foresight, the developer maximizes the present value of urban rents per acre, which vary with the timing of urban development. The present value of a parcel today is its opportunity cost in 'agriculture' until development plus its market value after conversion (minus construction costs). With perfect foresight, developers choose the timing of the conversion of land to urban use as a function of distance to the urban core, and development proceeds in an orderly fashion over time. Locations are developed according to their distance from the centre.

Of course, durable residential capital also implies that structures may depreciate or become obsolete. In particular, a historical pattern of development along concentric rings from the urban core, together with rising incomes, means that the most depreciated and obsolete dwellings are the more centrally located. But embedded in each of these parcels of real estate is the option to redevelop it in some other configuration. Obsolete and depreciated dwellings commanding low prices are those for which the option to exercise redevelopment is less costly.

Models of development with perfect foresight in which residential capital depreciates imply that the timing of initial redevelopment of residential parcels depends only on their distance from the urban core (since that indexes their vintage of development). These models imply that the capital intensity of land use does not exhibit the smooth and continuous decline with distance from the core. Capital intensity does decline with distance, on average, but the relationship is not monotonic.

With uncertainty, developers take into account their imperfect knowledge of future prices in making land use decisions today. But this means that developers may make mistakes by developing land too soon. As a consequence, land development may often proceed in a leapfrog pattern. Landowners may withhold some interior land from development in anticipation of higher rents and profitable development later on (see Capozza and Helsley, 1990, for a unified treatment).

The key point in these modern models of urban form which incorporate durable residential capital is that the timing as well as the location of development affect the choices made by housing suppliers. History 'matters' in these models, just as it does in the decisions of housing suppliers in urban areas.

Externalities

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Theory

Recent work has greatly extended these urban models to address explicitly the production and consumption externalities which give rise to cities. The basic models combine Marshallian notions of ‘economics of localized industry’ and Jacobs's (1969) notions of ‘urbanization economies’ with the perspective on monopolistic competition and product diversity introduced by Dixit and Stiglitz (1977).

On the consumption side, the general form of these models assumes that household utility depends on consumption of traded goods, housing, and the variety of local goods. The markets for traded goods and housing are competitive, while the differentiated local goods are sold in a monopolistically competitive market. If there is less differentiation among local goods, then variety loses its impact on utility; greater differentiation means that variety has a greater effect on utility. Under reasonable assumptions, the utility of a household in the city will be positively related to the aggregate quantity of local goods it consumes and the *number* of types of these goods which are available in the economy (see Quigley, 2001, for examples).

On the production side of the economy, the importance of a variety of locally produced inputs can be represented in a parallel fashion. For example, suppose that the aggregate production function includes labour, space and a set of specialized inputs. Again, the markets for labour and space can be taken as competitive, while the differentiated local inputs are purchased in a monopolistically competitive market. If there is less differentiation among inputs, then variety loses its impact on output; greater differentiation means that variety has a greater effect on output. For example, a general counsel may operate alone. However, she may be more productive if assisted by a general practice law firm, and even better served by firms specializing in contracts, regulation and mergers. Again, under reasonable conditions, output in the city will be related to quantities of labour, space, and specialized inputs utilized and also to the number of different producer inputs available in that city.

The theoretical models built along these lines yield a remarkable conclusion: diversity and variety in consumer goods or in producer inputs can yield external scale economies, even though all individual competitors and firms earn normal profits. In these models, the size of the city and its labour force will determine the number of specialized local consumer goods and the number of specialized producer inputs, given the degree of substitutability among the specialized local goods in consumption and among specialized inputs in production. A larger city will have a greater variety of consumer products and producer inputs. Since the greater variety adds to utility and to output, in these models larger cities are more productive, and the well-being of those living in cities increases with their size. This will hold true even though the competitive and monopolistically competitive firms in these models each earn a normal rate of profit (see Fujita and Thisse, 2002, for a comprehensive treatment).

Applications: pollution and transport

As emphasized above, however, the advantages of urban production and consumption are limited. Explicit recognition of the land and housing markets and the necessity of commuting suggests that, at some point, the increased costs of larger cities – higher rents arising from the competition for space, and higher commuting costs to more distant residences – will offset the production and consumption advantages of diversity. Other costs like air and noise pollution no doubt increase with size as well.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Nevertheless, even when these costs are considered in a more general model, the optimal city size will be larger when the effects of diversity in production and consumption are properly reckoned. Urban output will be larger and productivity will also be greater (see Quigley, 1998).

The empirical evidence assembled to support and test these theoretical insights about the regional economy is potentially very valuable. Hitherto, much of the discussion about the sources of economic growth was framed at that national level, and most of the aggregate empirical evidence – time series data across a sample of countries – was inherently difficult to interpret. By framing these theoretical propositions at the level of the region, it is possible to investigate empirically the sources of endogenous economic growth by using much richer bodies of data within a common set of national institutions. Geographical considerations of labour market matching and efficiency (Helsley and Strange, 1990), of the concentration of human capital (Rauch, 1993), and of patent activity (Jaffe, Trajtenberg and Henderson, 1993) have all been studied at the metropolitan and regional levels, and considerable effort is under way to use regional economic data to identify and measure more fully the sources of American economic growth. These are major research activities exploring urban externalities in urban economies throughout the developed world. This research programme is still in its infancy.

Of course, specialization, diversity and agglomeration are not the only externalities arising in cities. High densities and close contact over space reinforce the importance of many externalities in modern cities. Among the most salient are the external effects of urban transport – congestion and pollution. Most work trips in urban areas are undertaken by private auto. (Indeed, in 2000, less than four per cent of commuting was by public transit; see Small and Gomez-Ibanez, 1999.) In most US cities, automobiles are the dominant technology for commuting from dispersed origins to concentrated worksites. This technology is even more efficient for commuting from dispersed residences to dispersed worksites in metropolitan areas. Since commuting is concentrated in morning and evening hours, roads may be congested during peak periods, and idle during off-peak periods. Road users pay the average costs of travel when they commute during peak periods. They take into account the out-of-pocket and time costs of work trips, and in this sense commuters consider the average level of congestion in their trip-making behaviour. But commuters cannot be expected to account for the incremental congestion costs their travel imposes on other commuters. This divergence between the marginal costs of commuting and the average costs of commuting may be large during peak periods on arterial roadways. The imposition of congestion tolls, increasing the average costs paid by commuters to approximate the marginal costs they impose on others, would clearly improve resource allocation. In the absence of efficient road pricing, the rent gradients in metropolitan areas are flatter, and the patterns of residential location are more centralized than they would otherwise be. Land markets are distorted and the market price of land close to the urban core is less than its social value.

The obstacles to improved efficiency are technological as well as political. Until recently, mechanisms for charging road prices were expensive and cumbersome. But modern technology (for example, transponders to record tolls electronically) makes road pricing easy on bridges, tunnels and other bottlenecks to the central business district. Regular commuters affix a device to their autos, a device which can automatically debit the traveller's account. It would be a simple matter to vary these charges by time of day or intensity of road use and to make the schedule of these changes easily available to commuters. So far, at least in the United States, about the only form of price discrimination on bridges, tunnels and bottlenecks has been by vehicle occupancy, not by time of day and intensity of road use. It is surely possible to profit from the experience of other countries (such as Singapore) in pricing scarce

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
roadways.
Political resistance is a major factor inhibiting the diffusion of road pricing. Typically, tolls are imposed in new facilities and the proceeds are pledged to retire debt incurred in construction. Paradoxically, tolls are thus imposed on new uncongested roads. Later on, when the roads become congested, the initial debt has been retired, and there is political support for removing the toll. (After all, 'the investment in the bridge has been repaid.') This is surely an instance where economics can better inform public policy.

Applications: social, spatial and neighbourhood

Urban areas have always been characterized by social externalities as well. The close contact of diverse racial and ethnic groups in cities gives rise to much of the variety in products and services which enrich consumption. But the city is also characterized by the concentration of poverty and by the high levels of segregation by race and class.
The spatial concentration of households by income is, of course, predicted by the models of residential housing choice described above. A central question is the extent to which poverty concentrations give rise to externalities which disadvantage low-income households *relative* to their deprived circumstances in the absence of concentration. A great deal of qualitative research by other social scientists suggests that this is the case. Quite recent econometric research, however, suggests that this proposition is quite hard to demonstrate quantitatively by reliance on non-experimental data (see Manski, 1995.)
Nevertheless, the view that concentrations of disadvantaged households lead to more serious social consequences simply because of concentration is widely shared. For example, in low-wage labour markets most jobs are found through informal local contacts. If unemployed workers are spatially concentrated, it follows that informal contacts will produce fewer leads to employment.
Economic models of residential location also suggest that households will be segregated by race – to the extent that race and income are correlated. Yet research clearly indicates that the segregation of black households in urban areas is far greater than can be explained by differences in incomes and demographic characteristics.
Until quite recently, these patterns of segregation could be explained by explicitly discriminatory policies in the housing market. During the period of black migration to northern cities, rents were substantially higher in black neighbourhoods than in white neighbourhoods. As levels of migration tapered off in the 1970s, price differentials declined. The patterns of residence by race may be explicable by the tipping point models of Thomas Schelling (1971). In these models, there is a distribution of tolerance among the population, reflecting the maximum fraction of neighbours of a different race tolerated by any household. In this formulation, the race of each household provides an externality to all neighbouring households. It is easy to show that the racial integration of neighbourhoods may be impossible to achieve under many circumstances.
Despite this, there is widespread evidence of conscious discrimination in the markets for rental and owner-occupied housing (Ross and Yinger, 2002), four decades after passage of the first Fair Housing legislation.
Racial segregation in housing markets may have particularly important welfare implications as jobs continue to suburbanize at a rapid rate. Racial barriers to opening up the suburbs for residence may lead to higher unemployment rates among minority workers (see Glaeser, Hanushek and Quigley, 2004.)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The barriers to the integration of the suburbs by race and income are also related to the fiscal externalities which are conferred by one category of residents upon another category. Most local tax structures emphasize *ad valorem* property taxes, and in most urban areas towns are free to vary property tax rates to finance locally chosen levels of public expenditure. If local tax revenues are proportional to house value, and if local public expenditures are proportional to the number of households served, local governments have strong incentives to increase the property value per household in their jurisdictions. To achieve this outcome, local governments may simply use zoning regulations to prohibit construction of housing appropriate to the budgets of lower-income households. The prohibition of high-density housing and multi-family construction, the imposition of minimum lot-size restrictions and the imposition of development fees can all be used as devices to increase property tax revenue per household. Importantly, these rules also increase the price of low-income housing. Many of these regulations can also be cloaked in terms of ecological balance and environmental protection. The inability of higher levels of government to achieve balance and equity in new residential development in US urban areas is quite costly.

Summary

The field of urban economics emphasizes the spatial arrangements of households, firms, and capital in metropolitan areas, the externalities which arise from the proximity of households and land uses, and the policy issues which arise from the interplay of these economic forces.

See Also

- urban agglomeration
- urban production externalities
- urban transportation economics

Bibliography

Brueckner, J. 1987. The structure of urban equilibria: a unified treatment of the Muth–Mills Model. In *Handbook of Regional and Urban Economics*, vol. 2, ed. E. Mills. Amsterdam: North-Holland.

Capozza, D. and Helsley, R. 1990. The stochastic city. *Journal of Urban Economics* 28, 187–203.

Dixit, A. and Stiglitz, J. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Fujita, M. and Thisse, J.-F. 2001. *Economics of Agglomeration*. Cambridge: Cambridge University Press.

Glaeser, E., Hanushek, E. and Quigley, J. 2004. Opportunities, race, and urban location: the legacy of John Kain. *Journal of Urban Economics* 56, 70–9.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hall, P., ed. 1966. *Von Thünen's Isolated State*, trans. C. Wartenberg. Oxford: Pergamon Press.

Helsley, R. and Strange, W. 1990. Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* 20, 189–212.

Jacobs, Jane. 1969. *The Economy of Cities*. New York: Random House.

Jaffe, A., Trajtenberg, M. and Henderson, R. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108, 577–98.

Manski, C. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Mills, E. 1972. *Studies in the Structure of Urban Economy*. Baltimore, MD: John Hopkins Press.

Quigley, J. 1998. Urban diversity and economic growth. *Journal of Economic Perspectives* 12(2), 127–38.

Quigley, J. 2001. The renaissance in regional research. *Annals of Regional Science* 35, 167–78.

Rauch, J. 1993. Productivity gains from geographic concentration of human capital. *Journal of Urban Economics* 34, 380–400.

Ross, S. and Yinger, J. 2002. *The Color of Credit*. Cambridge, MA: MIT Press.

Schelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1, 143–86.

Small, K. and Gomez-Ibanez, J. 1999. Urban Transportation. In *Handbook of Regional and Urban Economics*, vol. 3, ed. Paul C. Cheshire and Edwin S. Mills. Amsterdam: North-Holland.

How to cite this article

Quigley, John M. "urban economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000035> doi:10.1057/9780230226203.1770

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban environment and quality of life

Matthew E. Kahn
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Urban quality of life is a key determinant of where the educated choose to live and work. Recognizing the importance of attracting the high-skilled, many cities are investing to transform themselves into ‘consumer cities’. This essay examines the supply and demand for non-market urban quality of life. It provides an overview of hedonic methods often used to estimate how much households pay for non-market urban attributes such as temperate climate and clean air.

Keywords

compensating differentials; congestion; crime; hedonic functions; internal migration; pollution; population density; quality of life; urban environment and quality of life; urban growth; urbanization

Article

Soon a majority of the world's population will live in cities. In 1950, 30 per cent of the world's population lived in cities. By 2000, this fraction had grown to 47 per cent. It is predicted to rise to 60 per cent by 2030 (United Nations Population Division, 2004). While the popular media focus on the growth of ‘mega-cities’, much urbanization occurs through the development of new cities and the growth of smaller metro areas (Henderson and Wang, 2004).

People have migrated to cities in pursuit of better economic opportunities than are available in the countryside (Harris and Todaro, 1970). In the past, urbanization has been viewed as representing a trade-off. Urban workers earned higher wages than rural residents but suffered from a lower quality of life. In the 1880s, the average urbanite in the United States had a life expectancy ten years lower than that of the average rural resident (Haines, 2001). Frederick Engels bemoaned the density and the squalor in Britain's manufacturing cities in the mid-19th century. Urban historians have provided indelible descriptions of US cities in the past. In the 19th century, dead horses littered the streets of New York City, and thousands of tenement-dwellers were exposed to stinking water, smoky skies, and ear-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
shattering din (Melosi, 1982; 2001). During the 19th and early 20th centuries, the skies above such major cities as Chicago and Pittsburgh were dark with smoke from steel smelters, heavy industrial plants, and burning coal.

Since the early 20th century, many major cities in the developed world have experienced sharp improvements in quality of life. By 1940, the urban mortality premium had vanished (Haines, 2001). Starting in the early 1970s, air pollution, water pollution and noise pollution have sharply fallen in many major US cities. While there are several causes of this progress, ranging from effective regulation to industrial transition from manufacturing to services and technological advance, the net result of this trend is that past ‘production cities’ are transforming themselves into ‘consumer cities’.

Cities that have high quality of life will have greater success at attracting the footloose highly educated to live there. Empirical studies have documented that a location's stock of educated people plays an important role in generating urban growth (Glaeser, Scheinkman and Shleifer, 1995).

First, I sketch how a city ‘produces’ quality of life. I then discuss the demand for urban quality of life using a household production function framework. While urban quality of life is a valued ‘commodity’, there are no explicit markets where it can be purchased. Utility maximizing households face a trade-off in choosing where to locate. In cities with higher quality of life, home prices are higher. Measuring this price premium for quality continues to be a major focus of much environmental and urban empirical research.

The supply of urban quality of life

Each city can be thought of as a differentiated product. Its attributes include some exogenous factors such as climate and risk of natural disasters, and endogenous factors such as average commuting times, pollution and crime. Some of these endogenous attributes are by-products of economic activity. A city of 10,000 bike-riding lawyers would have much cleaner air than another city with 500,000 old-car-driving steel workers. None of the steel workers driving old cars intends to pollute local air. Pollution represents an unintended consequence of their daily commuting mode and of local industrial production. This example highlights the importance of scale, composition and technique effects in determining local environmental quality. In the above example, scale refers to whether the city has 10,000 people or 500,000 people. Composition effects focus on consumption patterns (such as bike versus car) and industrial patterns (such as law firms versus steel plants). If one controls for a city's scale and composition, urban environmental quality can be high due to technique effects brought about by government regulation or the free market designing new capital with low emissions (for example, hybrid cars such as the Toyota Prius).

Early research in urban economics emphasized scale effects such that the biggest cities suffered more quality-of-life degradation as they expanded (Tolley, 1974). Anyone can migrate to a big city without paying an ‘entry fee’. When an extra person moves to a big city from a smaller city, this migration causes net social damage (due to extra congestion and pollution). Migrants will ignore the fact that their choice degrades local public goods in the destination city, but a benevolent planner would not. In the absence of a big city entry fee, the big city grows beyond its efficient size.

Cross-city empirical research has documented that such urban challenges as crime, pollution and congestion are all greater in big cities than in smaller cities (Glaeser, 1998; Henderson, 2002). But this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
‘cost’ of city bigness is declining over time. In the 1990s, crime fell fastest in the largest US cities (Levitt, 2004). Ambient air pollution is improving in many major cities despite a continued increase in population (Glaeser and Kahn, 2004). The suburbanization of employment in all major US metropolitan areas has meant that that population ‘sprawl’ has not increased commute times.

City size is not a sufficient statistic for determining a city's quality of life. Other relevant factors are the city's geography, industrial and demographic composition, and government policy. A city's geography determines its climate and its capacity for handling local pollution. Put simply, some cities have it and some cities don’t. As Billy Graham once said, ‘The San Francisco Bay Area is so beautiful, I hesitate to preach about heaven while I'm here.’

Cities differ in their ability to absorb growth without suffering urban quality-of-life degradation. World Bank researchers have recently documented the importance of climate and topological features of the city in determining how much air pollution is caused by economic growth (see Dasgupta et. al., 2004). Windier cities and cities that receive more rainfall suffer less ambient pollution from a given amount of emissions.

The composition of city economic activity also plays a key role in determining the supply of quality of life. All else equal, a city that specializes in manufacturing relative to services will have a lower quality of life. Such a city will have greater levels of ambient particulate and sulphur dioxide pollution. Water pollution will be greater, and more hazardous waste sites will be created. The rise and decline of manufacturing in the US rust belt over the 20th century provides dramatic evidence documenting these effects (Kahn, 1999). A similar ‘natural-experiment’ has played out as communism died. In major cities in the Czech Republic, Hungry and Poland air pollution improved in the 1990s because the phase out of energy subsidies contributed to the shutdown of communist era industrial plants (Kahn 2003). As major cities such as New York and London and Chicago have experienced an industrial transition from manufacturing to finance and services, more people work in the service and tourist industries, and these workers have a financial stake in keeping the city's quality of life high.

A city's demographics also play a role in determining its quality of life. A city filled with senior citizens will offer a different set of restaurants and cultural opportunities from a city filled with immigrants and young parents. If a city can attract the highly educated, then a virtuous circle can be set off. Since more highly educated people earn more income, this will attract better restaurants and other commercial amenities.

Government policy plays a role in determining a city's quality of life. Boston's Big Dig project has cost over US$14 billion and is intended to beautify Boston by submerging its ugly highways connecting the city centre to the waterfront and increasing the supply of green parks. Successful Clean Air Act regulation has sharply reduced vehicle emissions in Los Angeles. Rudy Giuliani, Mayor of New York City, achieved wide acclaim for improved policing that some have argued contributed to the sharp decline in the city's crime rate in the 1990s.

The supply of urban quality of life varies across cities and within cities. Some variation such as proximity to a major park or body of water is exogenously determined, but public policy can also have differential effects on quality of life across a city's neighbourhoods. The Clean Air Act has reduced Los Angeles’ smog by much more in inland Hispanic communities than along the Pacific Ocean (Kahn, 2001). Economists are just starting to investigate the general equilibrium impacts of regulations that differentially improve urban quality of life in some parts of a city relative to other parts of the same city

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(Sieg et al., 2004). If the improvements in quality of life were unexpected, then homeowners in such areas will receive a windfall. Long-standing renters in communities that have experienced regulation-induced improvements in local public goods will pay higher rents and may no longer be able to afford to live in their old community.

Demand for urban quality of life

The household production function approach offers a framework for modelling the demand for non-market local public goods such as climate, street safety and local environmental quality. A person gains utility from being healthy, safe and comfortable. To achieve these goals, one purchases market goods such as doctor visits, home alarm systems and home entertainment systems. In addition, this person might choose a city and a residential community within this city featuring a temperate climate, low smog levels and safe streets.

Each household must choose a city and a community within that city to live in.

Households that value quality of life face a trade-off in that each city represents a bundle of non-market attributes and economic opportunities. Some cities such as San Francisco are beautiful but home prices are very high. Other cities such as Houston offer warm winter weather and cheap housing but its residents face severe summer humidity. Market products can offset such city's disamenities. Before the advent and diffusion of cheap air conditioning, humid cities would feature much lower home prices to compensate households for summer humidity. The diffusion of the air conditioner has allowed households to enjoy the benefits of living in warmer cities such as Houston during winter without suffering from humidity in summer (Rappaport, 2003). This market product has increased the demand for living in humid cities.

Households may reveal different willingness to trade off non-market goods depending on the household's age, income and demographic circumstances. A household with children may place greater weight on communities with good schools. Households may differ in their demand for urban attributes. Asthmatics will avoid highly polluted cities and skiers will not mind the cold New England winters. Household demand may also hinge on idiosyncratic factors; for example, an individual who grew up in a specific city may want to remain living near his childhood friends.

The hedonic equilibrium approach for valuing urban quality of life

The theory of compensating differentials says that it will be more costly to live in ‘nicer’ cities (Rosen, 2002). This theory is really a ‘no arbitrage’ result. If migration costs are low across urban areas and if potential buyers are fully informed about the differences in non-market urban attributes bundles, then real estate prices will adjust such that homes in cities with higher quality of life will sell for a premium. An enormous empirical literature has estimated cross-city and within-city hedonic price functions to estimate the implicit compensating differentials for non-market goods. In these studies, the dependent variable is the price of home *i* in city *j* in community *m* in year *t*. Define X_{it} as home *i*'s physical attributes in year *t*. A_{jt} represents city *j*'s attributes in year *t* and A_{mjt} represents the attributes of community m located in city *j* in the year *t*. Given this notation, a standard real estate hedonic regression will take the form:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\text{Price}_{ijmt} = \beta_0 + \beta_1^{*}X_{it} + \beta_2^{*}A_{jt} + \beta_3^{*}A_{mjt} + \varepsilon_{ijmt} \quad (1)$$

Multivariate regression estimates of this regression yield estimates of the compensating differentials for city level local public goods (based on β_2) and community-level local public goods (based on β_3). These coefficients represent the marginal implicit prices for small increases in the consumption of local public goods. Studies that control for a vector of local public goods are able to pinpoint the relative importance of different features of cities and communities ranging from climate to air pollution to urban crime. Equation (1) highlights the fact that households face a rich set of choices both across cities and across communities within the same city.

Environmental studies have used this hedonic framework to estimate compensating differentials for a myriad of different environmental local characteristics. For example, Costa and Kahn (2003) examine the compensating differential for living in nice climate in 1970 and in the year 1990. In 1970, a person would have to pay $1,288 (1990 dollars) in higher home prices per year to purchase San Francisco's climate over Chicago's climate. In 1990, this yearly price differential increased by $6,259 (1990 dollars) to $7,547. Chay and Greenstone (2005) use 1980 and 1990 data for all US counties find that a ten per cent reduction in ambient total suspended particulates increased home prices by three per cent. While much of the urban quality of life literature has focused on US city data, a promising research trend is examining international evidence.

Conclusion

Urban economic development policymakers have pursued very different growth strategies. Some cities subsidize sports stadiums while others build airports or downtown cultural centres. Such targeted investment is unlikely to yield the key urban anchor. This essay has argued that cities than can provide and enhance urban quality of life will attract the high-skilled. An end result of attracting this group is a more vibrant, diversified local economy. As per-capita incomes continue to rise, the demand for living and working in high quality-of-life cities will increase. The empirical literature continues to inquire into what the key components of quality of life are.

See Also

- city and economic development
- environmental Kuznets curve
- urban agglomeration
- urban growth
- urbanization

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Chay, K. and Greenstone, M. 2005. Does air quality matter? Evidence from the housing market. *Journal of Political Economy* 113, 376–424.

Costa, D. and Kahn, M. 2003. The rising price of nonmarket goods. *American Economic Review* 93, 227–32.

Dasgupta, S., Hamilton, K., Pandey, K. and Wheeler, D. 2004. Air pollution during growth: accounting for governance and vulnerability. World Bank Policy Research Working Paper 3383. Washington, DC: World Bank.

Glaeser, E. 1998. Are cities dying? *Journal of Economic Perspectives* 12(2), 139–60.

Glaeser, E. and Kahn, M. 2004. Sprawl and urban growth. In *Handbook of Urban Economics*, vol. 4, ed. V. Henderson and J. Thisse. Amsterdam: North Holland.

Glaeser, E., Scheinkman, J. and Shleifer, A. 1995. Economic growth in a cross-section of cities. *Journal of Monetary Economics* 36, 117–43.

Haines, M. 2001. The urban mortality transition in the United States 1800–1940. Historical Paper 134. Cambridge. MA: NBER.

Harris, J. and Todaro, M. 1970. Migration, unemployment & development: a two-sector analysis. *American Economic Review* 60, 126–42.

Henderson, V. 2002. Urban primacy, external costs, and quality of life. *Resource and Energy Economics* 24(1/2), 95–106.

Henderson, V. and Hyoung Gun Wang 2004. Urbanization and city growth. Online. Available at http://www.econ.brown.edu/faculty/henderson/papers.html, accessed 29 June 2005.

Kahn, M. 1999. The silver lining of rust belt manufacturing decline. *Journal of Urban Economics* 46, 360–76.

Kahn, M. 2001. The beneficiaries of Clean Air Act legislation. *Regulation* 24, 34–9.

Kahn, M. 2003. New evidence on Eastern Europe's pollution progress. *Topics in Economic Analysis & Policy* 3(1), article 4. Online. Available at http://www.bepress.com/bejeap/topics/vol3/iss1/art4, accessed 28 June 2005.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Levitt, S. 2004. Understanding why crime fell in the 1990s: four factors that explain the decline and six that do not. *Journal of Economic Perspectives* 18(1), 167–90.

Melosi, M. 1982. *Garbage in the Cities: Refuse, Reform and the Environment: 1880–1980*. College Station: Texas A&M Press.

Melosi, M. 2001. *Effluent America: Cities, Industry, Energy and The Environment*. Pittsburgh, PA: University of Pittsburgh Press.

Rappaport, J. 2003. Moving to nice weather. Research Working Paper 03–07. Kansas: Federal Reserve Bank of Kansas City.

Rosen, S. 2002. Markets and diversity. *American Economic Review* 92, 1–15.

Sieg, H., Smith, V., Banzhaf, H. and Walsh, R. 2004. Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *International Economic Review* 45, 1047–77.

Tolley, G. 1974. The welfare economics of city bigness. *Journal of Urban Economics* 1, 324–45.

United Nations Population Division. 2004. *World Population Prospects: The 2004 Revision Population Database*. Online. Available at http://esa.un.org/unpp, accessed 28 June 2005.

How to cite this article

Kahn, Matthew E. "urban environment and quality of life." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000059>
doi:10.1057/9780230226203.1771

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban growth

Yannis M. Ioannides and Esteban Rossi-Hansberg
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

‘Urban growth’ refers to the process of growth and decline of economic agglomerations. The pattern of concentration of economic activity and its evolution have been found to be an important determinant, and in some cases the result, of urbanization, the structure of cities, the organization of economic activity, and national economic growth. The size distribution of cities is the result of the patterns of urbanization, which result in city growth and city creation. The evolution of the size distribution of cities is in turn closely linked to national economic growth.

Keywords

agricultural development; cities; congestion; demography; economic growth; neoclassical model; industrial revolution; internal migration; knowledge spillovers; new economic geography; population density; production externalities; quality ladder model of economic growth; returns to scale; rural growth; spatial distribution; symmetry breaking; systems of cities; urban agglomeration; urban economic growth; vs national economic growth; urban economics; urban growth; urbanization; Zipf's Law

Article

Urban growth – the growth and decline of urban areas – as an economic phenomenon is inextricably linked with the process of urbanization. Urbanization itself has punctuated economic development. The spatial distribution of economic activity, measured in terms of population, output and income, is concentrated. The patterns of such concentrations and their relationship to measured economic and demographic variables constitute some of the most intriguing phenomena in urban economics. They have important implications for the economic role and size distribution of cities, the efficiency of production in an economy, and overall economic growth. As Paul Bairoch's magisterial work (1988) has established, increasingly concentrated population densities have been closely linked since the dawn of history with the development of agriculture and transportation. Yet, as economies move from those of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
traditional societies to their modern stage, the role of the urban sector changes from merely providing services to leading in innovation and serving as engines of growth.

Measurement of urban growth rests on the definition of ‘urban area’, which is not standard throughout the world and differs even within the same country depending upon the nature of local jurisdictions and how they might have changed over time (this is true even for the United States). Legal boundaries might not indicate the areas covered by urban service-providers. Economic variables commonly used include population, area, employment, density or output measures, and occasionally several of them at once, not all of which are consistently available for all countries. Commuting patterns and density measures may be used to define metropolitan statistical areas in the USA as economic entities, but major urban agglomerations may involve a multitude of definitions.

The study of urban growth has proceeded in a number of different directions. One direction has emphasized historical aspects of urbanization. Massive population movements from rural to urban areas have fuelled urban growth throughout the world. Yet it is fair to say that economics has yet to achieve a thorough understanding of the intricate relationships between demographic transition, agricultural development and the forces underlying the Industrial Revolution. Innovations were clearly facilitated by urban concentrations and associated technological improvements. A related direction focuses on the physical structure of cities and how it may change as cities grow. It also focuses on how changes in commuting costs, as well as the industrial composition of national output and other technological changes, have affected the growth of cities. Another direction has focused on understanding the evolution of systems of cities – that is, how cities of different sizes interact, accommodate and share different functions as the economy develops and what the properties of the size distribution of urban areas are for economies at different stages of development. Do the properties of the system of cities and of city size distribution persist while national population is growing? Finally, there is a literature that studies the link between urban growth and economic growth. What restrictions does urban growth impose on economic growth? What economic functions are allocated to cities of different sizes in a growing economy? Of course, all of these lines of inquiry are closely related and none of them may be fully understood, theoretically and empirically, on its own. In what follows we address each in turn.

Urbanization and the size distribution of cities

The concentration of population and economic activity in urban areas may increase either because agents migrate from rural to urban areas (urbanization) or because economies grow in term of both population and output, which results in urban as well as rural growth. Urban centres may not be sustained unless agricultural productivity has increased sufficiently to allow people to move away from the land and devote themselves to non-food producing activities. Such ‘symmetry breaking’ in the uniform distribution of economic activity is an important factor in understanding urban development (Papageorgiou and Smith, 1983). Research on the process of urbanization spans the early modern era (the case of Europe having been most thoroughly studied; De Vries, 1984) to recent studies that have applied modern tools to study urbanization in East Asia (Fujita et al., 2004). The ‘New Economic Geography’ literature has emphasized how an economy can become ‘differentiated’ into an industrialized core (urban sector) and an agricultural ‘periphery’ (Krugman, 1991). That is, urban concentration is beneficial because the population benefits from the greater variety of goods produced

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(forward linkages) and may be sustained because a larger population in turn generates greater demand for those goods (backward linkages). This process exploits the increasing returns to scale that characterize goods production but does not always lead to concentration of economic activity. The range of different possibilities is explored extensively in Fujita, Krugman and Venables (1999). These ideas have generated new lines of research; see several related papers in Henderson and Thisse (2004).

The process of urban growth is closely related to the size distribution of cities. As the urban population grows, will it be accommodated in a large number of small cities, or in a small number of large cities, or in a variety of city sizes? While cities have performed different functions in the course of economic development, a puzzling fact persists for a wide cross-section of countries and different time periods. The size distribution of cities is Pareto-distributed, is ‘scale-free’. Gabaix (1999) established this relationship formally. He showed that, if city growth is scale independent (the mean and variance of city growth rates do not depend on city size: Gibrat's Law) and the growth process has a reflective barrier at some level arbitrarily close to zero, the invariant distribution of city sizes is a Pareto distribution with coefficient arbitrarily close to 1: Zipf's Law. (Empirical evidence on the urban growth process as well as Zipf's Law is surveyed by Gabaix and Ioannides, 2004.)

These results imply that the size distribution of cities and the process of urban growth are closely related. Eeckhout (2004) extends the empirical investigation by examining in depth all *urban places* in the United States and finds that the inclusion of the lower end of the sample leads to a log-normal size distribution. Duranton (2004) refines the theory by means of a quality-ladder model of economic growth that allows him to model the growth and decline of cities as cities win or lose industries following technological innovations. Ultimately, the movements of cities up and down the hierarchy balance out so as to produce a stable, skewed size distribution. This theory is sufficiently rich to accommodate subtle differences across countries (in particular the United States and France) that constitute systematic differences from Zipf's Law. Rossi-Hansberg and Wright (2004) use a neoclassical growth model that is also consistent with observed systematic deviations from Zipf's Law: in particular, the actual size distribution of cities shows fewer smaller and larger cities than the Pareto distribution, and the coefficient of the Pareto distribution has been found to be different from 1 although centred on it. They identify the standard deviation of the industry productivity shocks as the key factor behind these deviations from Zipf's Law. The evident similarity of the conclusions of those two papers clearly suggests that the literature is closer than ever before to resolving the Zipf's Law ‘puzzle.’

Urban growth and city structure

Understanding urbanization and economic growth requires understanding the variety of factors that can affect city size and therefore its short-term dynamics. All of them lead to the basic forces that generate the real and pecuniary externalities that are exploited by urban agglomeration, on one hand, and congestion, which follows from agglomeration, on the other. Three basic types of agglomeration forces have been used, in different varieties, to explain the existence of urban agglomerations (all of them were initially proposed in Marshall, 1920): (*a*) knowledge spillovers, that is, the more biomedical research there is in an urban area, the more productive a new research laboratory will be; (*b*) thick markets for specialized inputs: the more firms that hire specialized programmers, the larger the pool from which an additional firm can hire when the others may be laying off workers; and (*c*) backward and forward

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
linkages. Local amenities and public goods can themselves be relevant agglomeration forces.
The size of urban agglomerations is the result of a trade-off between the relevant agglomeration and congestion forces. Urban growth can therefore be the result of any city- specific or economy-wide change that augments the strength or scope of agglomeration forces or reduces the importance of congestion forces. One example that has been widely used in the literature is reductions in commuting costs that lead to larger cities in terms of area, population, and in most models also output (Chatterjee and Carlino, 1999). Another example is the adoption of information and communication technologies that may increase the geographical scope of production externalities, therefore increasing the size of cities.
Changes of underlying economic factors cause cities to grow or decline as they adjust to their new equilibrium sizes. Another more subtle factor is changes in the patterns of specialization that are associated with equilibrium city sizes. That is, the coexistence of dirty industry with high-tech industry generates too much congestion, and therefore cities specialize in one or the other industry. Adjustments in city sizes and patterns of specialization in turn may be slow, since urban infrastructure, as well as business structures and housing are durable, and new construction takes time (Glaeser and Gyourko, 2005). However, this type of change lead only to transitional urban growth, as city growth or decline eventually dies out in the absence of other city-specific or economy-wide shocks. Even when any of the economy-wide variables, such as population, grows continuously, the growth rate of a specific city will dwindle because of new city creation (Ioannides, 1994; Rossi-Hansberg and Wright, 2004).
Much attention has also been devoted to the effect that this type of urban growth has on urban structure. Lower commuting costs may eliminate the link between housing location choices and workplace location. This results in more concentration of business areas, increased productivity because of, say, knowledge spillovers, and lower housing costs in the periphery of the city. Urban growth can therefore lead to suburbanization as well as multiple business centres, as in Fujita and Ogawa (1982) or Lucas and Rossi-Hansberg (2002). Those phenomena become increasingly important because of the decline in transport and commuting costs brought about by the automobile along with public infrastructure investments. In other words, urban growth is associated with sprawl (Anas, Arnott and Small, 1998).

Urban and national economic growth

Most economic activity occurs in cities. This fact links national and urban growth. An economy can grow only if cities, or the number of cities, grow. In fact, Jacobs (1969) and Lucas (1988) underscore knowledge spillovers at the city level as a main engine of growth. The growth literature has also argued that, in order for an economy to exhibit permanent growth, the aggregate technology has to exhibit asymptotically constant returns to scale (Jones, 1999). If not, the growth rate in an economy will either explode or converge to zero. How is this consistent with the presence of scale effects at the city level? Eaton and Eckstein (1997), motivated by empirical evidence on the French and Japanese urban systems, study the possibility of parallel city growth, which is assumed to depend critically on intercity knowledge flows together with the accumulation of partly city-specific human capital across a given number of cities. Rossi-Hansberg and Wright (2004) propose a theory where scale effects and congestion forces at the city level balance out in equilibrium to determine the size of cities. Thus, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economy exhibits constant returns to scale through the number of cities increasing along with the scale of the economy. Hence, economic growth is the result of growth in the size and the number of cities. If balanced growth is the result of the interplay between urban scale effects and congestion costs, these theories have important implications for the size distribution of cities and the urban growth process. These implications turn out to be consistent with the empirical size distribution of cities, that is, Zipf's Law, and with observed systematic deviations from Zipf's Law.

To summarize: urban growth affects the efficiency of production and economic growth, and the way agents interact and live in cities. Understanding its implications and causes has captured the interest of economists in the past and deserves to continue doing so in the future.

See Also

- city and economic development
- endogenous growth theory
- location theory
- new economic geography
- power laws
- spatial economics
- symmetry breaking
- systems of cities
- urban agglomeration
- urban economics
- urban environment and quality of life
- urban production externalities
- urbanization

Bibliography

Anas, A., Arnott, R. and Small, K. 1998. Urban spatial structure. *Journal of Economic Literature* 36, 1426–64.

Bairoch, P. 1988. *Cities and Economic Development*. Chicago: University of Chicago Press.

Chatterjee, S. and Carlino, G. 1999. Aggregate metropolitan employment growth and the deconcentration of metropolitan employment. Working paper. Philadelphia: Federal Reserve Bank of Philadelphia.

De Vries, J. 1984. *European Urbanization: 1500–1800*. Cambridge, MA: Harvard University Press.

Duranton, G. 2004. Urban evolutions: the still, the fast, and the slow. Working paper. London:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Department of Geography and Environment, London School of Economics.

Eaton, J. and Eckstein, Z. 1997. Cities and growth: theory and evidence from France and Japan. *Regional Science and Urban Economics* 27, 443–74.

Eeckhout, J. 2004. Gibrat's Law for (all) cities. *American Economic Review* 94, 1429–51.

Fujita, M. and Ogawa, H. 1982. Multiple equilibria and structural transition of nonmonocentric urban configurations. *Regional Science and Urban Economics* 12, 161–96.

Fujita, M., Krugman, P. and Venables, A. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Fujita, M., Mori, T., Henderson, J. and Kanemoto, Y. 2004. Spatial distribution of economic activities in Japan and China. In *Handbook of Regional and Urban Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Gabaix, X. 1999. Zipf's Law for cities: an explanation. *Quarterly Journal of Economics* 114, 739–67.

Gabaix, X. and Ioannides, Y. 2004. The evolution of city size distributions. In *Handbook of Regional and Urban Economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Glaeser, E. and Gyourko, J. 2005. Urban decline and durable housing. *Journal of Political Economy* 113, 345–75.

Henderson, J. and Thisse, J.-F., eds. 2004. *Handbook of Regional and Urban Economics*, vol. 4. Amsterdam: North-Holland.

Ioannides, Y. 1994. Product differentiation and economic growth in a system of cities. *Regional Science and Urban Economics* 24, 461–84.

Jacobs, J. 1969. *The Economy of Cities*. New York: Random House.

Jones, C. 1999. Growth: with and without scale effects. *American Economic Review P&P* 89(2), 139–44.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99, 483–99.

Lucas, R., Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22(1), 3–42.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lucas, R. and Rossi-Hansberg, E. 2002. On the internal structure of cities. *Econometrica* 70, 1445–76.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Papageorgiou, Y. and Smith, T. 1983. Agglomeration as local instability of spatially uniform steady-states. *Econometrica* 51, 1109–19.

Rossi-Hansberg, E. and Wright, M. 2004. Urban structure and growth. Working paper. Stanford: Department of Economics, Stanford University.

How to cite this article

Ioannides, Yannis M. and Esteban Rossi-Hansberg. "urban growth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000060> doi:10.1057/9780230226203.1772

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban housing demand

Todd Sinai
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Urban housing demand is a reflection of households' desire to live in cities. In this article, I discuss possible reasons why US households have exhibited an increasing taste for urban living, including employment, urban amenities, and consumption opportunities. Next, I explain how growing urban housing demand led to rising house prices and a sorting of households across cities by income. That dynamic generated a divergence across housing markets in the value of the tax subsidy to owner-occupied housing as well as housing market risk. Those factors, in turn, had a feedback effect on urban housing demand.

Keywords

housing markets; housing supply; housing tax subsidies; Internet, economics of; monocentric city models; population growth; preference externalities; productivity growth; superstar cities; urban housing demand; willingness-to-pay

Article

At its core, the demand for urban housing is just the manifestation of the demand for living in urban areas. On net, residence patterns suggest that most people want to live in or near cities, and that desire is increasing over time. In fact, by 1999, 75 per cent of US households lived in cities (Rosenthal and Strange, 2003). Today, urban America is where housing demand is most likely to exceed housing supply and generate rising house prices, where the tax system provides the greatest subsidy to owner-occupied housing, and where the housing market is the most volatile. In this article I discuss some of the causes and consequences of urban housing demand, and the supporting evidence.

Location

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
The classic explanation for the concentration of households in cities is that people want to live close to their jobs. That notion, developed in the monocentric city model of Alonso (1964), Mills (1972) and Muth (1969), leads to a prediction that is rarely as evident in reality as it is in theory: housing costs should rise as the distance to the employment centre falls since households would be willing to pay more in order to save time getting to work. Instead, households often settle for a longer commute in exchange for other positive qualities of a non-urban community, such as the density of development, the calibre of the school system, local taxes and amenities, and the similarity of the other residents to themselves. Since the 1960s, the patterns of where people live have begun to shift back to cities, even though people are now less likely to work in the downtown areas. According to Glaeser, Kolko and Sarz (2001), between 1960 and 1990 the rate of growth of commutes where the household lives in the city increased while the growth rate of commutes originating in the suburbs fell. Within cities, the high-income population has been moving closer to the central downtown area. Glaeser et al. argue that nowadays thriving cities are ‘consumer cities’, ones that attract highly educated households through appealing cultural amenities, such as museums, restaurants and opera. In fact, between 1977 and 1995, a temperate and dry climate, a coastal location, and more live performance venues and restaurants per capita predicted future population growth. By contrast, having more bowling alleys was correlated with population decline.

Indeed, the very congestion that urban economists typically point to as a reason that cities become unattractive may lead to an availability and quality of goods and services that are appealing. In a city, the large number of residents living in close proximity makes it feasible for even niche markets to be served since a critical mass of potential customers exists. Joel Waldfogel (2003), who in a series of papers termed this phenomenon ‘preference externalities’, found empirical support in the markets for broadcast radio, newspapers, and restaurants. For example, when there is a larger local consumer base for a certain format of radio station, calibre of newspaper, or style of restaurant, the more of them exist in a city. By revealed preference, that greater variety increases city dwellers’ welfare, because the more options there are for residents that share a particular set of tastes the more they consume.

Even the advent of the Internet has not dampened the consumption appeal of living in cities. Since the Internet makes information and goods universally available to anyone, no matter where they live, it substitutes for living in a city. However, Sinai and Waldfogel (2004) find that the number and variety of websites focused on a city increases with the city's population. By enhancing the welfare benefits of living in cities – perhaps by mitigating the effects of congestion or facilitating communication and connection among city residents – these sites have an offsetting positive effect on urban housing demand.

Urban housing demand and house prices

Two measures of the intensity of urban housing demand are house prices and the rate of house price growth. In some cities, housing is in inelastic supply because there is little or no open land and local regulations either restrict development or make it prohibitively expensive or slow. In that case, demand for a location leads to bidding up of the price of land in order to equilibrate housing demand with the available supply. Indeed, when one compares house prices across cities and town, areas that presumably have higher demand because they offer better amenities and fiscal conditions exhibit higher house prices (Roback, 1982). Another indication of high demand for a city is population growth, which occurs when

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
housing development is easy. I focus on high house prices because they can change the character of a city, which then has a further effect on urban housing demand.

Since the 1950s, a handful of metropolitan areas experienced real house price growth that significantly exceeded the national average, leading to a widening gap across locations in average house prices. For example, in 1950 the average house price in San Francisco was 37 per cent higher than the average across all metropolitan areas. By 2000 the gap had grown to 218 per cent. In order for land prices to continually grow in one location relative to another, the demand for that location must be growing as well. One possible explanation is that productivity growth in a handful of cities has exceeded the national average, and residents pay more to live in productive cities because their wage rises with their productivity. Another potential rationalization is that some cities are becoming more appealing over time and residents are paying more for increasingly higher quality.

Another possibility is that the rapid growth in the number and earnings of high-income households in the United States has led to an increased willingness-to-pay for scarce locations. Since some cities are in such limited supply, households have to outbid each other to live there, leading to land prices that grow with the aggregate spending power of the clientele that prefers that particular city. In ‘Superstar Cities’, Gyourko, Mayer and Sinai (2006) show that inelastically supplied, high-demand cities have income distributions that are shifted to the right: low-income families can live there only if they have a very strong preference for the city, while high-income families can live there even if they only modestly prefer it. As the national high-income population grows, the greater number of high-income families outbid relatively low-income families (as well as some high-income families) who are unwilling or unable to pay a higher premium to live in their preferred location. Gyourko, Mayer and Sinai find that such superstar locations experience supra-normal house price growth and a shift of their income distributions to the right as they experience inflows of high-income households and outflows of their lowest-income residents. This pattern has been intensified as cities have begun to ‘fill up’ due to the growing national population. For example, in 1960 only Los Angeles and San Francisco were demonstrably inelastically supplied. By 2000 more than 20 cities were. Gyourko, Mayer and Sinai show that cities that ‘fill up’ experience a right-shift in their income distributions and higher price growth after their transition into superstar city status.

These findings imply that there must be something unique and attractive about superstar cities, otherwise potential residents would turn to cheaper locations and superstar cities would not be able to sustain excess price growth. A niche-market appeal may be due to particular amenities, or the kind of preference externalities described by Waldfogel. As preference agglomerations form, the highest willingness-to-pay households are those that share the same preferences. If such sorting is along income lines, rising house prices can lead to high-income homogeneity, which itself makes an area more desirable to high-income residents. That dynamic implies that certain urban markets will evolve into luxury areas and grow increasingly unaffordable for the average household.

The inelasticity of housing supply also leads to price changes, and a correlated change in the demand for urban housing, in cities that are experiencing *declining* demand. Glaeser and Gyourko (2005) point out that, since housing does not quickly depreciate once built, if the demand for a city declines then house prices must fall since quantity cannot easily adjust downwards. That decline in prices can spur demand by low-income households that cannot afford to live anywhere else, leading to sorting into low-income enclaves rather than high-income ones.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The tax subsidy to owner-occupied housing

Differences in house prices among cities also affect the benefits homeowners obtain from their houses, which in turn affect the demand for urban housing. One such benefit in the United States that often is especially valuable in cities is the favourable Federal income tax treatment for owner-occupied housing, worth a total of $420 billion in 1999. The nature of these tax benefits is well-documented in this edition of the *New Palgrave* (housing policy in the United States). Gyourko and Sinai (2004) note that two conditions are necessary in order to receive a high value of this tax subsidy: a high-priced home, so that the subsidy operates on a larger base, and a high tax rate, which in the progressive US tax system follows from having a high income.

Because of this, the very same superstar city dynamic discussed earlier leads to an unequal distribution of the housing subsidy across the country. Superstar cities experience both house price growth and relatively high-income residents, and thus should also have the highest tax subsidies, further increasing the demand for urban housing in hot markets. Indeed, the tax subsidy is highly concentrated in a handful of cities, with just five metropolitan areas receiving more than 85 per cent of the total tax benefits in 1990. Between 1980 and 2000, the rise in house prices in superstar cities more than offset declining marginal tax rates, leading to a greater concentration of tax benefits in a handful of metropolitan areas. This tax subsidy has been shown to lead to higher house prices, either because the subsidy induces households to consume a larger quantity of housing or simply because house prices capitalize the present value of the future tax savings. Recent estimates of the after-tax price elasticity of housing demand cluster around –0.5, and the income elasticity around 0.25. Urban areas tend to exhibit relatively high demand elasticities, as demand is more readily capitalized into land prices rather than the limited new supply. By contrast, rural areas have much lower measured elasticities of housing demand.

Risk and the demand for urban housing

Since urban housing markets tend to have inelastic supply, they are more volatile as shocks to housing demand are transmitted more completely into rents and prices. That higher risk may deter households from living in urban areas since they would face more uncertainty over housing costs, whether they rented or owned. Also, house price volatility generates an additional cost because it distorts other investment decisions (Flavin and Yamashita, 2002). A mitigating factor, demonstrated in Sinai and Souleles (2005), is that long length-of-stay households can reduce their effective risk by owning their houses, in essence prepaying their housing costs. Other research suggests that uncertainty over house price growth simply may lead households to purchase housing in a city sooner than they otherwise would have in order to prevent housing costs from outpacing their income growth.

See Also

- housing policy in the United States
- housing supply
- residential real estate and finance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- residential segregation
- Tiebout hypothesis
- urban agglomeration
- urban economics
- urban environment and quality of life
- urban political economy
- urbanization

Bibliography

Alonso, W. 1964. *Location and Land Use*. Cambridge, MA: Harvard University Press.

Flavin, M. and Yamashita, T. 2002. Owner-occupied housing and the composition of the household portfolio over the life cycle. *American Economic Review* 92, 345–62.

Glaeser, E. and Gyourko, J. 2005. Urban decline and durable housing. *Journal of Political Economy* 113, 345–75.

Glaeser, E., Kolko, J. and Sarz, A. 2001. Consumer city. *Journal of Economic Geography* 1, 27–50.

Gyourko, J., Mayer, C. and Sinai, T. 2006. Superstar cities. Working Paper No. 12355. Cambridge, MA: NBER.

Gyourko, J. and Sinai, T. 2004. The (un)changing geographical distribution of housing tax benefits: 1980 to 2000. In *Tax Policy and the Economy*, vol. 18, ed. J. Poterba. Cambridge, MA: MIT Press.

Mills, E. 1972. *Studies in the Structure of the Urban Economy*. Baltimore, MD: Johns Hopkins University Press.

Muth, R. 1969. *Cities and Housing*. Chicago: University of Chicago Press.

Roback, J. 1982. Wages, rents, and the quality of life. *Journal of Political Economy* 90, 1257–78.

Rosenthal, S.S. and Strange, W.C. 2003. Evidence on the nature and sources of agglomeration economies. In *Handbook of Urban and Regional Economics*, ed. J.V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Sinai, T. and Souleles, N. 2005. Owner-occupied housing as a hedge against rent risk. *Quarterly Journal of Economics* 120, 763–89.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sinai, T. and Waldfogel, J. 2004. Geography and the Internet: is the internet a substitute or complement for cities? *Journal of Urban Economics* 56, 1–24.

Waldfogel, J. 2003. Preference externalities: an empirical study of who benefits whom in differentiated product markets. *RAND Journal of Economics* 34, 557–68.

How to cite this article

Sinai, Todd. "urban housing demand." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000063>
doi:10.1057/9780230226203.1773

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban political economy

Robert W. Helsley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Models of local public finance generally emphasize the roles of household mobility and community heterogeneity in the provision of local public services. In contrast, the emerging field of urban political economy examines how economic and political institutions influence the formation of local public policies. Key issues include the strength of the local executive, whether local politicians are elected ‘at large’ or to serve the interests of particular wards, the norms that govern behaviour and decision-making within city councils, and institutional innovation, especially the growth of so-called ‘private governments’.

Keywords

common pool problem; local public finance; minimum winning coalitions; parliamentary systems; presidential systems; principal and agent; private government; Tiebout hypothesis; urban political economy

Article

Economic models of local government applied to the United States generally suppress the roles of politics and political institutions. Indeed, the dominant model of local government, the Tiebout (1956) model of the provision of local public services to mobile residents, can be seen as an explicit attempt to eliminate the need for politics in cities. If individuals are highly mobile and communities offer a diverse menu of local taxes and expenditures, then there is no need for political expression – households can satisfy their demands for local public services by choosing to live in the community that provides their optimal bundle. Households ‘vote with their feet’. When local politics are explicitly considered, the political process is usually treated as an idealized form of majority rule in which residents vote directly on tax and spending programmes, and the political outcome corresponds to the most preferred policy of the median voter. This ‘institutionless’ view of local public finance has in fact been quite successful in,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for example, characterizing the demand for local public goods (Rubinfeld, 1987).

Local political institutions

However, most local policy choices are not made directly by residents. According to the International City/Council Management Association (ICMA), 43.7 per cent of US municipalities with populations over 2,500 were governed by the combination of a mayor and a city council in 2000, while 48.3 per cent were governed by the combination of a city council and a city manager. Thus, 90 per cent of US municipalities were governed at least in part by a representative body. Council members may be elected ‘at-large’, that is, from the entire city, or by wards or districts within the city. Some cities adopt a mixed system, in which the council contains both at-large and ward representatives.

Mayors (or their offices) are traditionally classified as being either ‘strong’ or ‘weak’. Strong mayors have broad powers, including a veto over some city council decisions. Strong mayors also prepare the city's budget, and have hiring and firing authority over the heads of city departments. In weak mayoral systems, most executive and legislative authority rests with the city council; the mayor performs largely ceremonial and organizational functions. Strong mayors are generally elected independently from members of the city council, and are more common in mayor–council systems. Baqir (2002), based on a sample of roughly 2,000 US municipalities in 1990, reports that 98 per cent of mayors in mayor–council systems were independently elected, compared to 65 per cent of mayors in council–manager systems. Strong mayors are generally associated with fiscal discipline, and there is some support for this view in other branches in the political economics literature. For example, the literature on comparative politics suggests that presidential systems have greater accountability to voters and less collusion within and between the branches of government than parliamentary systems (Persson, Roland and Tabellini, 1997; 1998; 2000). Persson, Roland and Tabellini (2000) show that presidential systems have lower levels of government spending as a share of national product. Inman and Fitts (1990) show that between 1795 and 1988 ‘strong’ presidents (those with ‘independent political strength’, identified from a survey of historians) were associated with lower levels of federal spending in the United States. Baqir (2002) suggests that a strong mayor may have a similar disciplinary effect on local government spending.

Many studies of local political institutions in North America examine the impacts of the reform movement of the early 20th century. The reform movement brought a number of changes in local government structure that were allegedly designed to limit the exercise of private interest and patronage in city politics and promote the pursuit of public interests and professional management. Some of the specific institutional changes that followed included the introduction of at-large and non-partisan elections for city council (a change that has since been partially reversed), the council–manager form of local government, civil service exams as a basis for appointment and promotion in the bureaucracy, and, in some areas, the replacement of the mayor–council form with a group of city ‘commissioners’, each of whom had executive and legislative responsibility for a different city department.

Early studies of reform governments expressed the hope that managerial expertise and autonomy in personnel matters could lead to lower costs for the delivery of local public services, and in particular, lower labour costs for municipal governments. However, subsequent empirical studies provide little support for this view: public expenditure levels and patterns in US cities seem to have been largely unaffected by the adoption of city managers, at-large representation and non-partisan elections.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The most compelling study of the reform movement in the recent economics literature is Rauch (1995). Rauch's hypothesis is that by creating a population of career bureaucrats in city government, the reform movement put in place incentives that encouraged investment in infrastructure and other ‘long-gestation-period’ projects. Rauch views the relationship between the city council and the bureaucracy as a principal–agent problem. Before reform, the agent, that is, the bureaucracy, is assumed to act as a political appointee who shares the council's immediate focus on retaining office. After reform, the bureaucracy is professionalized and the agent is assumed to have some job security and therefore a longer time horizon. The agent may then use his ‘powers of information collection and expenditure oversight’, in combination with costly or imperfect monitoring by the principal to direct resources towards longer-term projects that may further the agent's career. The implication is that this type of reform should increase the share of expenditures devoted to investment, as opposed to current public consumption. Using a panel of 144 cities over 23 years, Rauch regresses the infrastructure share of municipal expenditure on dummy variables for the use of civil service exams, the presence of a city manager, and the adoption of a commission form of local government. After accounting for the inertia generated by the durability of infrastructure investment, use of the civil service is found to have a positive impact on the share of expenditure devoted to infrastructure. Interestingly, in the cases where they are statistically significant, the presence of a city manager and the adoption of a commission form of government are both associated with lower levels of infrastructure spending.

The common pool problem in city councils

City councils are, in effect, local legislatures. One way to model the operation of a city council is by analogy with models of other legislative institutions. In that spirit, imagine a city council in which each councillor represents a well-defined local constituency. If councillors are elected by ward or district, then the constituencies will be geographic, as in most national, state and provincial legislatures. Councillors elected at-large may have non-geographic constituencies that are defined by a common ideology or policy initiative. Suppose that each councillor is motivated by holding office and that this gives him or her an incentive to pursue programmes and policies that provide net benefits to his or her constituents. It is generally assumed that the policies and programmes that are chosen by legislatures are ‘distributive’ in the sense that their costs are more widely distributed than their benefits. For example, benefits may be restricted to a particular district or group, while the supporting tax payments are made by residents of the entire city. Spending and tax choices are made by a majority vote of council members. The literature on legislative decision-making discusses a number of issues that relate to the efficiency of the policy choices that will emerge in this context. First, there is an incentive for ‘minimum winning coalitions’ within the legislature to form for the purpose of approving distributive policies (Riker, 1962). A minimum winning coalition is the smallest set of legislators that can guarantee passage of a proposal under majority voting. If proposals or projects have spillover costs and benefits, as distributive policies generally do, then the exclusion of the interests of delegates outside of a winning coalition will lead to inefficient choices. Second, minimum winning coalitions should be highly unstable, since excluded delegates have strong incentives to alter the coalition structure. Each member of the legislature faces some probability that he or she will be excluded from the minimum winning coalition for any particular policy proposal. Third, Weingast, Shepsle and Johnsen (1981), Shepsle and Weingast (1984), and others suggest that the resulting uncertainty helps explain the practice of ‘universalism’, in which the size of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
coalitions and the set of approved projects exceed the minimum winning size. In its extreme form, universalism involves a ‘norm of reciprocity’ in which each delegate supports the project of every other, and so a project for every delegate or constituency is approved. Finally, Weingast, Shepsle and Johnsen (1981) argue that politicians in such a setting have an incentive to count the resource costs of geographically earmarked programmes as benefits. They refer to this as the ‘Robert Moses’ effect: ‘pecuniary gains in the form of increased jobs, profits, and local tax revenues go to named individuals, firms, and localities from whom the legislator may claim credit and exact tribute’ (1981, p. 648). Such ‘political cost-accounting’ will obviously encourage individual representatives to support higher than efficient levels of public spending.

More formally, following Persson and Tabellini (2000, section 7.1), imagine that there are *M* seats on the city council and that the fixed population of each constituency is *N*. Thus, the aggregate population of the city is *MN*. If council members are elected by district or ward, so the constituencies are geographic, then the assumption of fixed constituencies implies that the population is immobile. Suppose that all residents are identical and have quasi-linear preferences of the form $U(g) + x$, where *g* is per capita consumption of a publicly provided good, *x* is the numeraire and the sub-utility function *U* (·) is increasing and strictly concave. All residents have the same exogenous income *y*. Public services are financed through lump-sum taxes that balance the city's budget. Each councillor is assumed to be a perfect representative of his or her constituent group.

If one takes utilitarianism as a normative benchmark, the efficient provision of public services in this symmetric setting maximizes aggregate utility $M(U(g) + x)$ subject to the resource constraint $MN(y - x - g) = 0$. The first-order condition for this problem implies $U'(g) = 1$: the marginal benefit of the public service should equal its marginal cost in every constituency. Represent this efficient level of provision by g^*.

In contrast, under extreme universalism, or with decentralized provision and centralized finance, each delegate chooses a level of the public service to maximize the utility of a representative constituent, taking the levels chosen by other delegates as fixed. If one lets g^0 represent the conjectured level chosen by others, the balanced budget requirement implies that the lump-sum tax τ for any group satisfies $\tau M = g + (M - 1)g^0$. Thus, an individual delegate chooses *g* to maximize

$$y - \frac{g + (M - 1)g^0}{M} + U(g). \tag{1}$$

The first-order condition for this problem implies $U'(g) = 1/M$. Each member of the legislature perceives that they pay only a fraction 1/*M* of the costs of the public services that they acquire. This is known as the common pool problem. If one lets g^U represent the level of provision under this extreme form of universalism, the concavity of $U(\cdot)$ implies $g^U > g^*$. The common pool problem thus leads to overprovision. Persson and Tabellini (2000, p. 163) summarize the nature of the distortion as follows:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
‘The problem here lies in the collective choice procedure, in which the tax rate is residually determined once all spending decisions have been made in a decentralized fashion. Concentration of benefits and dispersion of costs lead to excessive spending when such spending is residually financed out of a common pool of tax revenue.’

The first-order condition for g^U implies

$$\frac{dg^U}{dM} = -\frac{1}{MU''(g)} > 0 \qquad (2)$$

by concavity. Thus, the level of overprovision increases as the constituencies become smaller, *ceteris paribus*. Finally, if one allows $G^U = Mg^U$ to represent aggregate spending, we have

$$\frac{dG^U}{dM} = g^U + M\frac{dg^U}{dM} > 0. \qquad (3)$$

This is an instance of Weingast, Shepsle and Johnsen's (1981) ‘law of 1/*n*’: aggregate spending, and therefore the inefficiency of excessive spending increases with the number of constituencies or the size of the legislature.

This implication of the common pool problem seems to be supported by the evidence. Landbein, Crewson and Brasher (1996), based on a sample of 192 cities in 1980, all of which have a council–manager form of government and a weak mayor, find that local government expenditure per capita is positively related to the number of elected members of the city council. Baqir (2002) finds that the size of US local governments (measured by expenditures or employment per capita or expenditures as a share of total income) increases with the size of the city council. Baqir also finds that expenditures (per capita or as a share of total income) are not significantly different in councils where a majority of members are elected at-large, but that local government employment per capita is lower when at-large councillors are in the majority. However, evaluated at the sample means, employment per capita is actually higher where a majority of councillors are elected at large. This is consistent with the hypothesis that at-large councillors serve their (non-geographic) constituencies in much the same manner that ward councillors serve the interests of their wards. The positive relationship between the size of the government and the size of the council is unaffected by the presence of at-large elections. Baqir also examines the impact of a strong city executive, and finds that expenditures do not increase with council size when the city has a strong mayor with the power to veto city council decisions. As noted above, this is consistent with recent models and results from the literature on comparative politics.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Private government

Private governments are voluntary, exclusive organizations that supplement services provided by the public sector. There are two broad classes of private governments in the United States. Residential private governments, sometimes called residential community associations (RCAs), common interest developments (CIDs), or homeowner associations (HOAs), exist to further the interests of residential property owners. Commercial private governments, sometimes called business improvement districts (BIDs) or business investment areas (BIAs), exist to further the interests of their member firms. Private governments are highly controversial. Garreau (1991) labels them ‘shadow governments’, and argues that they are undemocratic, discriminatory, and operate outside of the constitutional restrictions that public governments face.

Residential private governments are an increasingly important component of housing markets and local government systems in North America. Garreau (1991, p. 189) estimates that there may have been as many as 130,000 RCAs in the United States in 1988. McKenzie (1996) reports that the number of CIDs in the United States grew from a few hundred in 1960 to 150,000 in 1993 and that they then housed 32 million people. The Community Associations Institute (an industry trade association) maintains that there were 231,000 RCAs in the United States in 2002, housing 57 million people. The 2001 American Housing Survey from the US Bureau of the Census reports that 28 per cent of all new-housing residents paid community association fees in 2001. Residential private governments provide security and sanitation services, and manage and maintain common facilities, including recreational facilities and infrastructure. They also regulate property use and individual conduct through covenants, codes, and restrictions in property deeds.

There are fewer commercial private governments, but their impacts are also substantial. Pack (1992) estimates that there were 400 BIDs in the United States in 1992, while Mitchell's (2001) survey found 404 independently managed BIDs in the United States in 1999. BIDs typically provide security, marketing and sanitation services. Mitchell reports that 94 per cent of BIDs engage in marketing, 85 per cent provide maintenance and sanitation services, and 68 per cent provide security. Mitchell's survey also found that 88 per cent of BIDs engaged in some form of policy advocacy, like lobbying governments on behalf of business interests. BIDs have become a key component of downtown revitalization strategies in many, if not most, major North American cities.

Private governments raise a number of interesting economic issues. First, they may have significant impacts on the traditional public sector. To the degree that private and public spending are substitutes, private governments provide a mechanism for the public sector to withdraw from certain activities. Helsley and Strange (1998; 2000a) show that such ‘strategic downloading’ is a characteristic of equilibrium in a game where public and private governments simultaneously choose levels of provision to maximize the welfare of their citizens and members, respectively. Cheung (2004) finds evidence of strategic downloading in a sample of California communities. Second, membership in private governments may be inefficient. Helsley and Strange (1999) argue that one of the essential features of gated communities is that they divert crime to other areas. This increases the incentive for other communities to engage in similar private policing activities (the activities are strategic complements), and may lead to excessive gated community development. Third, private governments have complex

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
welfare effects. Citizens with high demands for public services are generally made better off by this form of privatization. By joining the private government, they can supplement what is for them an inadequate level of public provision. If strategic downloading occurs, citizens with low demands, who choose not to join a private government, are also better off, since the level of public provision falls towards their optimal level. However, citizens in the middle of the distribution – whose demands were relatively well served by the traditional public sector – are made worse off.

The field of urban political economy is in its infancy. The roles that economic and political institutions play in the formation of local public policies are clearly deserving of further study.

See Also

- local public finance
- systems of cities
- Tiebout hypothesis
- urban economics
- urban environment and quality of life

Bibliography

Baqir, R. 2002. Districting and government overspending. *Journal of Political Economy* 110, 1318–54.

Caro, R.A. 1974. *The Power Broker: Robert Moses and the Fall of New York*. New York: Knopf.

Cheung, R. 2004. The interaction between public and private governments: an empirical analysis. Mimeo, Department of Economics, Florida State University.

Garreau, J. 1991. *Edge Cities: Life on the New Frontier*. New York: Doubleday.

Helsley, R.W. 2004. Urban political economics. In *Handbook of Regional and Urban Economics*, vol. 4, ed. J.V. Henderson and J.F. Thisse. Amsterdam: North-Holland.

Helsley, R.W. and Strange, W.C. 1998. Private government. *Journal of Public Economics* 69, 281–304.

Helsley, R.W. and Strange, W.C. 1999. Gated communities and the economic geography of crime. *Journal of Urban Economics* 46, 80–105.

Helsley, R.W. and Strange, W.C. 2000a. Potential competition and public sector performance. *Regional Science and Urban Economics* 30, 405–28.

Helsley, R.W. and Strange, W.C. 2000b. Social interactions and the institutions of local government. *American Economic Review* 90, 1477–90.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Inman, R.P. and Fitts, M.A. 1990. Political institutions and public policy: evidence from the U.S. historical record. *Journal of Law, Economics and Organization* 6, 79–132.

Landbein, L.I., Crewson, P. and Brasher, C.N. 1996. Rethinking ward and at-large elections in cities. *Public Choice* 88, 275–93.

McKenzie, E. 1996. Homeowner association private governments in the American political system. *Papers in Political Economy*, 75, University of Western Ontario.

Mitchell, J. 2001. Business improvement districts and the new revitalization of downtowns. *Economic Development Quarterly* 15, 115–23.

Pack, J.R. 1992. BIDs, DIDs, SIDs, SADs: private government in urban America. *The Brookings Review* 10, 18–21.

Persson, T. and Tabellini, G. 2000. *Political Economics*. Cambridge, MA: MIT Press.

Persson, T., Roland, G. and Tabellini, G. 1997. Separation of powers and political accountability. *Quarterly Journal of Economics* 112, 1163–202.

Persson, T., Roland, G. and Tabellini, G. 1998. Towards micropolitical foundations of public finance. *European Economic Review* 42, 685–94.

Persson, T., Roland, G. and Tabellini, G. 2000. Comparative politics and public finance. *Journal of Political Economy* 108, 1121–61.

Rauch, J.E. 1995. Bureaucracy, infrastructure, and economic growth: evidence from U.S. cities during the progressive era. *American Economic Review* 85, 968–79.

Riker, W.H. 1962. *The Theory of Political Coalitions*. New Haven: Yale University Press.

Rubinfeld, D.L. 1987. The economics of the local public sector. In *Handbook of Public Economics*, vol. 2, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Shepsle, K.A. and Weingast, B.R. 1984. Political solutions to market problems. *American Political Science Review* 78, 417–34.

Tiebout, C. 1956. A pure theory of local expenditure. *Journal of Political Economy* 64, 416–24.

Weingast, B.R., Shepsle, K.A. and Johnsen, C. 1981. The political economy of benefits and costs: a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

neoclassical approach to distributive politics. *Journal of Political Economy* 89, 642–64.

How to cite this article

Helsley, Robert W. "urban political economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000067>
doi:10.1057/9780230226203.1774

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban production externalities

Antonio Ciccone

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Urban production externalities (agglomeration effects) are external effects among producers in areas with a high density of economic activity. Such external effects are the main explanation for why productivity is usually highest in those areas of a country where economic activity is densest. There is some disagreement about the strength of urban production externalities. What is clear, however, is that even weak urban production externalities can explain large spatial differences in productivity.

Keywords

congestion; increasing returns; instrumental variables; intermediate inputs; knowledge spillovers; labour productivity; local technological externalities; localization economies; Mincerian wage regression; non-transportable input sharing; non-transportable intermediate inputs; outsourcing; schooling externalities; spatial externalities; skill-biased technical change; urban agglomeration; urban production externalities; urbanization economies; spatial wage differentials

Article

Urban production externalities (agglomeration effects) are external effects among producers in areas with a high density of economic activity. Such external effects are the main explanation for why productivity is usually highest in those areas of a country where economic activity is densest. The best understood urban production externalities are technological externalities due to knowledge spillovers and non-transportable input sharing, both of which are already discussed by Marshall (1920).

That local technological externalities translate into increasing returns at the city level is demonstrated formally by Henderson (1974). Building on the analysis of Chipman (1970), he also shows that such increasing returns are compatible with perfect competition. Abdel-Rahman (1988), Fujita (1988; 1989), and Rivera-Batiz (1988) present a rigorous analysis of decentralized market equilibria with increasing returns at the city level due to intermediate input sharing. These contributions build on the formalization

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of monopolistic competition of Spence (1976) and Dixit and Stiglitz (1977) to show how increasing returns to city size emerge when some intermediate inputs are non-transportable and produced subject to increasing returns at the plant level.

There is some disagreement about the strength of increasing returns to the density (or scale) of local economic activity and therefore about the strength of urban production externalities. This is partly because the best approach to estimation is still unclear. What is clear, however, is that even weak urban production externalities can explain much of the large spatial differences in productivity observed in many countries. This is because spatial differences in the density of economic activity are very large, so that even a small degree of increasing returns to density can explain sizable spatial productivity differences. Moreover, mobile physical capital and tradable intermediate inputs imply that the strength of increasing returns to density exceeds the strength of urban production externalities (by approximately a factor of two).

The remainder of this article first illustrates the link between the strength of urban production externalities and the strength of increasing returns to the density of economic activity (or increasing returns at the city level). It then turns to the advantages and drawbacks of different empirical approaches to urban production externalities.

A model of urban production externalities and increasing returns

The link between urban production externalities and increasing returns to the density of economic activity is easily illustrated using the technology-spillover model of Ciccone and Hall (1996) extended to include costlessly tradable intermediate inputs. This extension is important for understanding why the strength of increasing returns to density is approximately twice the strength of urban production externalities. Including urban production externalities due to non-transportable intermediate-input sharing in the model would be straightforward (see Ciccone and Hall, 1996) but not change any of the relevant conclusions.

Model set-up

Let $f(n_f, k_f, m_f; Q_c, A_c)$ be the production function that describes the amount of output produced by firm f on an acre of space when employing n workers, m units of costlessly tradable intermediate inputs, and k units of capital (lower-case letters denote per-acre quantities). The acre is embedded in county c with total output Q and total acreage A (upper-case letters denote total county-level quantities). The simplest production function to deal with is one where the externality depends multiplicatively on the density of economic activity Q/A, and the elasticity of $f(n, k, m; Q, A)$ with respect to all its arguments is constant. In this case,

$$q_f = \left(n_f^{\alpha} k_f^{\beta} m_f^{1-\alpha-\beta}\right)^{1-\rho}\left(\frac{Q_c}{A_c}\right)^{\lambda} \quad (1)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$\lambda \geq 0$ captures the strength of urban production externalities (agglomeration effects); for example, $\lambda = 2\%$ means that a doubling of the density of economic activity is associated with a two per cent increase in the output of the firm (for a given amount of inputs used by the firm). $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ determine the relative importance of labour, capital and intermediate inputs in production. And $0 \leq \rho \leq 1$ captures possible decreasing returns to labour, capital and intermediate inputs when holding the amount of land used in production constant (congestion effects).

Input demand and value added

Firms maximize profits taking factor prices and aggregate output in each county as given. Profit maximization implies that firms employ capital up to the point where its marginal product is equal to the national rental price of capital *R* (measured in units of output), which gives rise to a demand for capital equal to $k_f = \beta(1-\rho)q_f / R$. The demand for intermediate inputs can be determined analogously as $m_f = (1-\alpha-\beta)(1-\rho)q_f$, where we have assumed that one unit of intermediate inputs can be produced with one unit of output. Substituting these factor demand functions in (1) and solving for output at the firm level yields that q_f is proportional to $(n_f)^{\alpha(1-\rho)/(1-(1-\rho)(1-\alpha))}(Q_c / A_c)^{\lambda/(1-(1-\rho)(1-\alpha))}$. Moreover, the demand for intermediate inputs implies that value added at the firm level (*y*) and county level (*Y*) are a fraction $1-(1-\alpha-\beta)(1-\rho)$ of the total value of production at the firm and county level respectively, that is, $y_f = q_f - m_f = (1-(1-\alpha-\beta)(1-\rho))q_f$ and $Y_c = Q_c - M_c = (1-(1-\alpha-\beta)(1-\rho))Q_c$. Hence, value added at the firm level is linked to firm-level employment and county-level value added by

$$y_f = (\gamma n_f)^{\frac{\alpha(1-\rho)}{1-(1-\rho)(1-\alpha)}}\left(\frac{Y_c}{A_c}\right)^{\frac{\lambda}{1-(1-\rho)(1-\alpha)}}, \tag{2}$$

where γ is an unimportant constant.

Increasing returns to density

The amount of labour *N* employed in a county is taken to be distributed uniformly in space; $n_f = N_c / A_c$ for all firms *f* in the county. Substituted in (2), this yields that output per acre in a county is linked to employment per acre by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\frac{Y_c}{A_c} = \left(Y\frac{N_c}{A_c}\right)^{1+\theta}, \tag{3}$$

where the strength of increasing returns to the density of economic activity θ is given by

$$\theta = \frac{(\lambda - \rho)}{\alpha(1-\rho) - (\lambda - \rho)}. \tag{4}$$

As expected, increasing returns to density are stronger when urban production externalities λ are strong and congestion effects ρ are weak. A necessary condition for productivity to be greater in areas with dense economic activity is that urban production externalities (agglomeration effects) more than offset congestion effects, $\theta > 0$.

From increasing returns to density to urban production externalities

The relationship between increasing returns to the density of economic activity θ and the strength of net agglomeration effects $\lambda - \rho$ in (4) depends on $\alpha(1-\rho)$, the elasticity of output with respect to labour. In equilibrium, this elasticity equals the share of labour in the total value of production. In the United States, the share of labour in value added is around two thirds (for example, Gollin, 2002) and the share of intermediate inputs in value added around one half (for example, Basu, 1995), which implies a share of labour in the total value of production of around one third. To see that this implies that urban production externalities are magnified, note that for small values of $\lambda - \rho$ (4) implies

$$\theta \cong \frac{\lambda - \rho}{\alpha(1-\rho)} = 3(\lambda - \rho), \tag{5}$$

where we have made use of $\alpha(1-\rho) = 1/3$. A one-point increase in the strength of urban production externalities therefore implies a three-point increase in the strength of increasing returns to the density of economic activity. Much of this magnification is due to the presence of intermediate inputs. In a model without intermediate inputs where physical capital earns one third of value added, the factor of magnification would have been (only) 3/2.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Empirical approaches and results

Increasing returns to city or industry size

Early empirical studies of urban scale effects by Sveikauskas (1975), Segal (1976), Moomaw (1981; 1985), and Tabuchi (1986) focused on the link between city size and productivity at the city and the city-industry level. The empirical results indicate that doubling city size increases productivity by between three and eight per cent. Nakamura (1985) and Henderson (1986; 2003) extend the analysis by distinguishing between urbanization economies and localization economies. Localization economies are increasing returns related to the size of city industries, while urbanization economies refer to increasing returns to overall city size. Henderson concludes that scale effects are mostly at the industry level, but Nakamura finds evidence for both urbanization and localization economies.

Most studies of the strength of agglomeration economies measure output as the value of production or value added from the U.S. Census Bureau's Census of Manufacturers. This data-set does not contain information on the value of services that plants purchase in the market or obtain from headquarters. Census of Manufacturers data will therefore overstate the value added of city industries. This bias is likely to be greater in larger cities, for two reasons. First, there is more service outsourcing in larger cities, due to the larger variety of services available (Holmes, 1999; Ono, 2007). Second, headquarter services are more likely to be counted twice in larger cities, as such cities are more likely to contain both a plant and its headquarters. The total value of production from the Census of Manufacturers has the additional disadvantage of counting twice all intermediate inputs traded within and across industries located in the same city.

Increasing returns to density and the productivity of US states

In the United States, the finest level of geographical detail with reliable data on value added is the state level. Ciccone and Hall (1996) therefore estimate increasing returns to the density of economic activity by combining state-level value added data with the model in (3). Aggregating county-level value added to the state level yields that labour productivity in state s, Y_s / N_s, is equal to

$$\frac{Y_s}{N_s} = D_c(\theta) \equiv \sum_{c=1}^{C_s} \left(\frac{YN_c}{A_c} \right)^{1+\theta} \frac{N_c}{N_s}, \tag{6}$$

where C_s is the number of counties in the state. Hence, the strength of increasing returns to the county-level density of economic activity can be estimated by combining cross-state variation in labour productivity and the state-level density index $D_c(\theta)$, which depends on county-level employment density and the distribution of employment across counties. Ciccone and Hall find θ to be just above

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
five per cent, using a least-squares approach. Because of large differences in the density of economic activity, this limited degree of increasing returns to density can explain more than half of the sizable differences in output per worker across US states.

Ciccone and Hall's work is about the degree of increasing returns to the density of economic activity, not about the strength of urban production externalities. Going from one to the other is rather straightforward, however. Using (5) yields that θ equal to five per cent corresponds to a net agglomeration effect $\lambda - \rho$ of 1.7 per cent. According to the *Flow of Funds Accounts of the United States, 1982–1990* prepared by the Board of Governors of the Federal Reserve System (1997), the share of land in the total value of production ρ in the private sector outside of agriculture and mining is around 0.5 per cent. Hence, λ is between 2 and 2.5 per cent, which implies that a doubling of the density of economic activity in a county is associated with a 2–2.5 per cent increase in the output firms produce with a given amount of inputs (see (1)). Mobile physical capital and tradable intermediate inputs therefore imply that the strength of increasing returns to density exceeds the strength of urban production externalities by a factor of two. Hence, more than half of the differences in output per worker across US states can be explained by rather weak urban production externalities.

An important concern when estimating agglomeration economies is potential feedback from productivity to the density of economic activity. To address this possibility, Ciccone and Hall (1996) use an instrumental variables approach. The instruments for the state-level density index used are the population and population density of US states between 1850 and 1880, as well as the presence or absence of a railroad in each state in 1860 and the distance of states from the eastern seaboard. The identifying assumption is that the original sources of agglomeration in the United States have remaining influences only through the preferences of people about where to live. The instrumental variables estimates of θ are between 5.5 and 6.1 per cent, and therefore very similar to the least squares estimates.

Agglomeration effects in Europe

For many European countries, value added data is available at a much finer level of geographic detail than for the United States. Employing such data for France, Germany, Italy, Spain and the UK, Ciccone (2002) finds an average degree of increasing returns to the local density of economic activity of between four and five per cent, only slightly below estimates for the United States. Rice, Venables and Patacchini (2006) find a similar result using geographically detailed earnings data for the UK. They also take into account the scale of production in neighbouring locations weighted by travel times, and find that productivity benefits diminish quickly with travel distance.

Human capital externalities?

An open question is whether there are agglomeration economies associated with the geographic concentration of human capital. Rauch (1993) examines this issue by augmenting a standard Mincerian wage regression (for example, Card, 1999) with data on the characteristics of cities where people live. His empirical model relates wages of individuals i in cities c, w_{ic} to relevant individual characteristics (for example, education, experience), X_{ic}, and to the average level of schooling of the city, S_c, and other

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

city characteristics, Z_c,

$$\log w_{ic} = aX_{ic} + bS_c + cZ_c + \varepsilon_{ic} \quad (7)$$

where ε_{ic} summarizes all other (unobserved) factors affecting individual wages across cities. Least-squares estimation of (7) using US data for 1980 yields a positive and significant coefficient on average schooling in the city (*b*), and Rauch therefore concludes that there are human capital externalities at the city level.

A drawback of Rauch's approach is that it cannot account for time-invariant unobserved city characteristics that increase both schooling and wages. Another drawback is that city-level schooling is taken to be exogenous. Acemoglu and Angrist (2001) address these drawbacks by taking US states, rather than cities, to be the relevant aggregate unit in (7). In this case, the data allow for an analysis of the effects of increases in average state-level schooling on individual wages. Moreover, Acemoglu and Angrist show that changes in average schooling at the state level can be instrumented by state-level changes in compulsory-schooling and child-labour laws. Their approach yields no evidence of significant schooling externalities between 1960 and 1980.

Ciccone and Peri (2006) show that a positive effect of average schooling in a Mincerian wage equation like (7) may not reflect human capital externalities but a downward sloping demand function for human capital. They therefore propose an alternative approach, which exploits the fact that the wage differential between more and less educated workers reflects differences in marginal social products of the two worker types when human capital externalities are absent. This approach does not yield evidence of significant human capital externalities at the level of US cities or states between 1960 and 1990.

Moretti (2004a) finds that US cities where the labour force share of college graduates increased most between 1980 and 1990 also saw the largest wage increase for college graduates. Using Census of Manufacturers plant-level data, Moretti (2004b) finds that the output of plants in high-tech city industries rises with levels of schooling in other high-tech industries in the same city. This evidence is consistent with human capital externalities. An alternative explanation could be that skill-biased technological progress translated into increases in the productivity and wages of college graduates in high-tech industries. Cities that specialized in industries experiencing rapid productivity growth would in this case see faster output growth and attract more college graduates. This alternative hypothesis is especially plausible for the 1980–90 period, which was marked by rising college wage premia due to skill-biased technological progress (for example, Katz and Murphy, 1992).

See Also

- externalities
- new economic geography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- urban agglomeration

Bibliography

Abdel-Rahman, H.M. 1988. Product differentiation, monopolistic competition and city size. *Regional Science and Urban Economics* 18, 69–86.

Acemoglu, D. and Angrist, J. 2001. How large are the social returns to education: evidence from compulsory schooling laws. In *NBER Macroeconomic Annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Basu, S. 1995. Intermediate goods and business cycles: implications for productivity and welfare. *American Economic Review* 85, 512–31.

Board of Governors of the Federal Reserve System. 1997. *Flow of Funds Accounts of the United States, 1982–1990*. Washington, DC: Federal Reserve.

Card, D. 1999. The causal effect of education on earnings. In *Handbook of Labor economics*, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Chipman, J.S. 1970. External economies of scale and competitive equilibrium. *Quarterly Journal of Economics* 84, 347–85.

Ciccone, A. 2002. Agglomeration effects in Europe. *European Economic Review* 46, 213–27.

Ciccone, A. and Hall, R.E. 1996. Productivity and the density of economic activity. *American Economic Review* 86, 54–70.

Ciccone, A. and Peri, G. 2006. Identifying human capital externalities: theory with applications. *Review of Economic Studies* 73, 381–412.

Dixit, A.K. and Stiglitz, J.E. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67, 297–308.

Fujita, M. 1988. A monopolistic competition model of spatial agglomeration: differentiated product approach. *Regional Science and Urban Economics* 18, 87–124.

Fujita, M. 1989. *Urban Economic Theory: Land Use and City Size*. Cambridge: Cambridge University Press.

Gollin, D. 2002. Getting income shares right. *Journal of Political Economy* 110, 458–74.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Henderson, J.V. 1974. The sizes and types of cities. *American Economic Review* 64, 640–56.

Henderson, J.V. 1986. Efficiency of resource usage and city size. *Journal of Urban Economics* 19, 47–70.

Henderson, J.V. 2003. Marshall's scale economies. *Journal of Urban Economics* 53, 1–28.

Holmes, T. 1999. Localization of industry and vertical disintegration. *Review of Economics and Statistics* 81, 314–25.

Katz, L.F. and Murphy, K.M. 1992. Changes in relative wages, 1963–1987: supply and demand factors. *Quarterly Journal of Economics* 107, 35–78.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Moomaw, R.L. 1981. Productivity and city size: a critique of the evidence. *Quarterly Journal of Economics* 96, 675–88.

Moomaw, R.L. 1985. Firm location and city size: reduced productivity advantages as a factor in the decline of manufacturing in urban areas. *Journal of Urban Economics* 17, 73–89.

Moretti, E. 2004a. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics* 121, 175–212.

Moretti, E. 2004b. Workers education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review* 94, 656–90.

Nakamura, R. 1985. Agglomeration economies in urban manufacturing industries: a case of Japanese cities. *Journal of Urban Economics* 17, 108–24.

Ono, Y. 2007. Market thickness and outsourcing services. *Regional Science and Urban Economics* 37, 220–38.

Rice, P.,Venables, A.J. and Patacchini, E. 2006. Spatial determinants of productivity: analysis for the regions of Great Britain. *Regional Science and Urban Economics* 36, 727–52.

Rauch, J.E. 1993. Productivity gains from geographic concentration of human capital: evidence from the cities. *Journal of Urban Economics* 34, 380–400.

Rivera-Batiz, F.L. 1988. Increasing returns, monopolistic competition, and agglomeration economies in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

consumption and production. *Regional Science and Urban Economics* 18, 125–53.

Segal, D. 1976. Are there returns to scale in city size? *Review of Economics and Statistics* 58, 339–50.

Spence, M. 1976. Product selection, fixed costs, and monopolistic competition. *Review of Economic Studies* 43, 217–35.

Sveikauskas, L. 1975. The productivity of cities. *Quarterly Journal of Economics* 89, 393–413.

Tabuchi, T. 1986. Urban agglomeration, capital augmenting technology, and labor market equilibrium. *Journal of Urban Economics* 20, 211–28.

How to cite this article

Ciccone, Antonio. "urban production externalities." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000061> doi:10.1057/9780230226203.1775

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urban transportation economics

Alex Anas
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Advances by economists in understanding the demand, capacity and supply, pricing, finance and performance of urban transportation systems is reviewed. The economics of urban transportation has emphasized externalities such as traffic congestion. The effects of transport systems and travel behaviour on real estate prices, urban land use and density and urban expansion as well as the reciprocal effects of urban form on the nature and utilization of transport systems are studied by economists.

Keywords

congestion; congestion tolls; derived demand; George, H.; labour market discrimination; land markets; land taxes; land use; land values; land-use zoning; Pigouvian taxes; residence location choice; Samuelson, P.; spatial mismatch; static assignment models; stationary state models; transport externalities; transport system performance; urban economics; urban sprawl; urban transportation economics; value of time; work location choice

Article

The study of transportation in urban areas relates to urban economics and to public economics and finance. The development of cities and their land use patterns cannot be understood without studying the transportation systems that shape them, nor can urban transportation systems be understood independently of the urban economy.
Unique aspects of urban transportation economics relate to demand, capacity and supply, the performance of urban transportation systems, and pricing and finance. We provide a discussion of the key conceptual issues and knowledge in each of these areas of the field and point out some challenges that remain. (For reviews of transportation economics focused less on its relationship to urban economics and more on technical issues internal to transportation, see Arnott and Kraus, 2003, and Small and Verhoef, 2006.)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Demand

The demand for transport is ‘derived demand’. Travel provides utility mostly because it is a means to an end, be it a consumer purchase, getting to work or to recreation. The travel itself usually has a disutility which varies according to the quality, reliability and safety of the transport system or the particular trip. Hence, virtually all transport choices involve a trade-off between the inconvenience and cost of a trip on the one hand and the frequency with which that trip is chosen relative to other trips on the other.
Beginning with the emergence of the telephone, the demands for travel and for communication have become increasingly interlinked in an urban setting. While travel and communication are substitutes because a phone call, fax or e-mail (or a messenger or letter in the pre-telephone days) may reduce the need for a trip, they are also complements because cheaper communication generates higher demand for goods, services and personal contacts. From this higher demand more travel is subsequently derived.
An important aspect of urban travel is the fact that the out-of-pocket cost of travel can be low relative to the value of time expended in that travel. As such, travel competes with leisure and with work as a key activity to which time must be allocated. The dominance of time-cost means that market prices are less important than full opportunity costs in the explanation and measurement of travel behaviour. Values of time vary greatly among consumers since wage rates vary but also because of other factors. Thus, consumers who undertake similar trips frequently incur vastly different opportunity costs.
The demand for urban travel by consumers is derived from a complex set of hierarchically linked choices. At the top of the hierarchy and slowest to change are decisions relating to where to work and where to reside. Lower in the hierarchy and more malleable are choices about the number and type of cars to own including the possibility of dispensing with cars and relying on walking or public transport for some trips. Yet lower on the hierarchy are choices about the destinations and frequency of discretionary trips, the frequency of commuting (to the extent that work arrangements do not require daily commuting), the destination of the commute being implicit in the residence–workplace choice, and the mode of transport (private automobile, taxi, public transit or walking) that will be used on each such round trip. There is the choice of the time of day during which particular trips are made and the trip chaining decision about whether several trips may be combined into a tour. In a tour, the trips are linked sequentially, beginning and ending at the point of origin (for example, the consumer's home). Finally, also important are the choices of particular routes (of the highway network, for example) on which trips occur. Firms make a similar set of choices. At the top of the hierarchy is the location of the plant vis-à-vis suppliers and markets, followed by the choice of a vehicle fleet and associated decisions of the modes (barge, rail, truck and so on) to use to get products to market or procure them from suppliers.
The study of travel demand using econometric techniques has not yet advanced to the point where a unified theory of travel can be tested that deals simultaneously with all or even most of the levels in the hierarchy. Led by McFadden (1973), transport economists have mostly focused on mode choice: the split of the demand for travel between competing modes such as auto, urban rail or bus and car pooling on a particular trip or round trip. This has resulted in the widespread use of a rich variety of discrete choice models (such as logit, nested logit or probit) that are designed to predict the probability that a randomly selected traveller will choose a particular mode to commute to work. Modelling the choice of residence or job locations has not received much attention from transportation economists. Virtually all studies in this area can be found in the urban economics literature, and some have emphasized the joint

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
choice of residence location and mode of commuting (Anas and Duann, 1985). Most travel demand studies focus exclusively on commuting, ignoring the fact that discretionary non-work travel is continually increasing with incomes, car ownership and suburbanization. And the trade off between commuting and discretionary trip-making under a time budget constraint has remained relatively unexamined.

Another area of demand that has received attention is the choice of travel route on a congested highway network. The work of Beckmann, McGuire and Winsten (1956) counterfactually conceived traffic on a network as a stationary state process of steady flow, rather than as a system of queues and bottlenecks causing complex flow dynamics. Despite this, the simplicity of the model led to the prolific development of static assignment models by operations researchers. These models use system optimization principles to simulate how travellers choose the least costly route on a congested network. Stochastic cost perceptions have been introduced into this type of stationary state models (Daganzo and Sheffi, 1977). More recently, a variety of dynamic simulation models that recognize the queue and bottleneck nature of traffic are under development based on the principle that travellers choose not only a route but also departure–arrival times (Arnott, de Palma and Lindsey, 1990).

Capacity and supply

One aspect of supply is that most transport is made possible in large part by consumer effort, time, and by inputs purchased by the consumer such as car, gasoline, vehicle maintenance and garage. Viewed this way supply becomes virtually inseparable from demand. As such it would make sense to model a part of the supply decision within a household production context.

Another aspect of supply is that the public sector is involved in the planning, provision and operation of most travel infrastructure. This includes highways as well as buses and urban rail. A key decision variable is capacity, measured as the throughput of passengers per hour that can be transported in a particular direction at a given time during the day on a particular facility. This throughput determines the user's travel time. Also important, however, are the safety, privacy, reliability and quality of the travel time and its components such as in-vehicle time, waiting at a station, searching for parking and time walking to and from stations and parking lots.

The key supply decision is the quantity of streets and highways and public transit rights-of-way. Road capacity relative to demand determines, in part, the level of traffic congestion in an area. Since land is the prime input in roads, more road building reduces the land available for other uses such as housing or production, raising the market price of land in such uses. In turn, the price of land chiefly determines how much land is allocated to create road capacity in an area. Thus, the most congested areas are also the ones where land is the most expensive. With extremely expensive land as in Tokyo, London, Paris or downtown New York, the substitution of capital for land results in tunnelling for transit systems (subways) and even for some roads.

An important supply question is whether economies of scale exist in congested highway traffic flow. Congestion occurs when the vehicles sharing the same road segment at the same time reach a critical value relative to road capacity. The addition of one more vehicle then begins to delay the other vehicles. The total cost experienced by the vehicle stream increases by a marginal cost that is higher than the cost privately born by the marginal vehicle's passengers. The difference between this social marginal cost and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the private average cost is the monetary value of the sum of delays the marginal vehicle imposes on all the vehicles travelling with it on the road segment. The evidence seems to suggest that this congestion process exhibits no economies of scale at least at a crude level. Scaling up (or down) road capacity and the volume of traffic in the same proportion, would not increase the average cost of travel. Capital costs of highways, on the other hand, were found to exhibit significant scale economies by the engineering estimates of Meyer, Kain and Wohl (1965), but since then Keeler and Small (1977) and others have found statistical evidence of weak or virtually no scale economies.

In contrast to highways, rail-based public transit systems are subject to scale economies and, more importantly, to economies of density. As more passengers use these systems (for example by reducing headways between successive trains), the per-passenger total average cost comes down because of the high fixed costs involved in system construction and maintenance. This is the chief reason why such rail infrastructure is uneconomical in US cities below some critical size such as one million or more people, or in suburban areas of low land use densities where the passengers' time-costs of accessing transit stations can be high (Kim, 1979).

System performance

The reconciliation of supply and demand results in system performance. Unlike other markets in which price is the only salient outcome of market performance, in transport the outcomes include travel time, the level of congestion or travel delay, air pollution from car traffic, accidents, system reliability (that is, the variability of travel time from day to day or hour to hour), and pecuniary and non-pecuniary externalities caused by the transport system. While travel time, congestion and reliability costs are primarily born by the travellers, air pollution and accidents have costs that are born by travellers as well as by non-travellers. Thus, the economic performance of a transport system cannot be measured completely without evaluating the social costs and benefits created by these external effects.

The purely pecuniary externalities of transport are pervasive. For example, as noted, the creation of transport capacity has a direct effect on the supply and price of land available for other uses and can thus cause land scarcity. But this is only the direct effect of capacity provision. The indirect effect on land values and land use is quite different and works at both the extensive and intensive margins. At the extensive margin, cities endowed with more road and transit capacity can expand to land areas that were previously inaccessible. At the intensive margin, transport systems work by changing the *relative* accessibility of land parcels. Areas that are made relatively more accessible than before gain value, while areas made relatively less accessible lose value. As a result of these shifts, windfall gains and losses in land markets should be among the chief measures of transport system performance evaluation. The aggregate change in land values can be positive or negative. Since most land is owned by consumers (such as homeowners) transport system changes play an important role in redistributing private wealth and public revenues from *ad valorem* property taxes by changing an existing pattern of accessibility.

Transportation, land use and land prices are the central foci in urban economics and a variety of models have been developed. Virtually all of these assume that all jobs are located at a predetermined centre, an anachronism given that current downtowns in US cities contain no more than ten per cent of the jobs. Versions of this basic model based on linear programming have been developed to model road capacity

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
provision and transit investment in congested cities (Mills, 1972; Kim, 1979).

Other pecuniary externalities centre on the improved discretionary mobility enabled by transport systems. Such mobility improvements have received praise as well as criticism. Improved mobility enables easier, cheaper and more frequent contacts among firms and among firms and consumers. This should result in positive social benefits enhancing productivity and boosting economic growth. It has been noted in the large literature on *spatial mismatch* that central-city minorities in the United States who are less-mobility enabled, are at a disadvantage competing for suburban employment. While discrimination and suburban land use exclusion cause minorities to be clustered and socially cloistered in central cities, lower car ownership may also hinder their ability to compete for distant suburban jobs. Improved mobility induces economic agents to locate in a more spread out pattern, substituting cheaper outlying land for more expensive, centrally located land. The resulting urban land use pattern, common in the United States, has been dubbed ‘urban sprawl’. Sprawl has been blamed for a variety of ills stemming from the increased dependence on cars and reduced pedestrian mobility that sprawled land use promotes. Among such perceived ills, for example, is the alleged demise of social and neighbourhood cohesion and the rising obesity of American children and adults.

Pricing and finance

In practice, urban roads and transit systems are subsidized. In the United States a large part of the cost of highways and roads comes from general income taxes. The rest of the cost comes from taxes on gasoline and taxes on real property. Urban rail systems are also heavily subsidized with fares covering only about half of the operating and maintenance costs. Hence, for all forms of urban transport with the possible exception of unregulated taxis and jitneys, market-based user fees and marginal cost prices do not play the role they do in other markets.

What does economic theory tell us about how urban transport systems should be priced and financed? The answer will be different for highway and rail systems, primarily because the latter are subject to economies of scale.

The congestion externality is key in highway pricing and investment (Vickrey, 1969). Economic efficiency requires that each traveller pay his full marginal social cost on each road segment that he uses. As we saw earlier, the full marginal social cost includes the monetary value of the delay each traveller imposes on his co-travellers. This is higher where congestion is high, falling to zero where congestion is not present. It has been shown that if congestion tolls can be properly calculated and levied on travellers, then with no economies of scale in roads, the tolls collected from the vehicles using a particular road segment would in the long run cover the amortized costs of road construction and maintenance. The only requirement is for road planners to build more (less) road capacity where toll revenue exceeds (falls short) of these amortized costs.

The congestion toll has three coincident theoretical interpretations. First, it is a Pigouvian tax (Pigou, 1947) because it levies, on the source of a negative externality, a tax that closes the gap between the social marginal cost and the private average cost. In this role, the toll causes travellers to economize on travel by internalizing the negative externality they create. Second, because a road can be viewed as a (congested) public good, the congestion toll in the long run serves to equate the marginal benefit of road capacity with the marginal cost of supplying it, the Samuelson rule for the optimal finance of a public

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
good (Samuelson, 1954). The toll itself is a marginal benefit since it measures the reduction in total travel cost that would be realized if one more unit of road capacity were to be added, while the marginal cost of the capacity is the cost of purchasing the additional capacity. Third, since the aggregate toll revenue from the road segment is equal to the land rent the road would fetch in an alternative use, the aggregate toll is equivalent to a confiscatory tax on the owners of the land, the Henry George rule (George, 1879). On the view that the land used for roads is privately owned and operated by competitive or contestable firms, the Pigouvian pricing described above would be the outcome of profit maximization, and the aggregate toll revenue would confiscate the profits of these private road owners. On the alternative view that the land used for roads is owned by society, the congestion tolls are the fees travellers pay society for the right to use the road, and in the long run these fees add up to the rent on land, provided land markets are competitive.

Keeler and Small (1977) empirically estimated what congestion tolls should be in the San Francisco Bay Area on the assumption of fixed land use. The effects of tolls on urban form have been studied within the naïve theoretical urban model that assumes all jobs are at a central point (Arnott and MacKinnon, 1978) or a central point and a suburban ring (Sullivan, 1983). Simple simulations based on such models show small efficiency gains of up to one per cent of income for reasonably congested cities. Recent studies, based on modern assumptions of completely dispersed employment, show similar efficiency gains (Anas and Xu, 1999). All of these studies show that congestion tolls could significantly reduce travel times. But the welfare benefits of tolls would come mostly from changes in travel mode and the timing of travel during the day, rather than from land use adjustments.

Congestion tolls have become more popular in recent years and have seen such prominent implementation as in central London. But the correct calculation of first-best tolls is a quagmire. Chief among the difficulties is the fact that one must know how the value of travel time is distributed among travellers using the same road segment. If I share the road with higher (lower) income drivers, the toll on me should be higher (lower). Without knowledge of the distribution, accurate first-best tolls cannot be computed because values of time vary so widely among people. A second difficulty is that road use varies enormously throughout the day, requiring that first-best tolls should similarly vary. The problem is simplified somewhat by dividing the day into peak and off-peak periods. A third difficulty is that the technology used to detect congestion and calculate tolls should not be so obtrusive on travel as to create more congestion than the tolls would alleviate. Automatic vehicle identification by several means is feasible and not expensive. This may contribute to a wider use of tolls in the future.

Although the calculation of first-best tolls is highly daunting, a number of second-bests are available. Tolls levied on major roads but not on local roads may be effective second-bests. A tax on the market price of parking in heavily congested destinations such as the downtowns of major cities would achieve some of the efficiencies of first-best tolls. Taxes on gasoline are not nearly as effective, because gasoline usage is not closely related to the congestion created on a trip. Such taxes heavily penalize driving on congestion-free roads, for example.

Unlike highways, rail transit should be priced as a regulated natural monopoly. Since marginal cost is below average cost at any scale, marginal cost pricing ensures efficiency but requires a subsidy to the transit operator to cover fixed costs. Thus for transit systems, theory tells us that fares should be set to cover variable operating costs, while other taxes should be used to purchase the fixed inputs, including land (right-of-way). The debate then, should be about what these other taxes should be. Considerable

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
evidence exists showing that land around transit stations appreciates in value after a transit investment is announced or constructed. Anas and Duann (1985) used an empirically estimated general equilibrium model to predict prior to construction that residential property values around the proposed stations of the Chicago Midway line would increase, with the increase sharply tapering off with distance from the stations. They estimated that the aggregate increase could pay for about 40 per cent of the construction cost. McMillen and McDonald (2004) used *ex post* data on housing sales and confirmed that these predictions were accurate. Taxing such windfall gains is one source of revenue for fixed facilities, although there are practical complications about how to accurately measure and document the land value appreciation in a legal-administrative context.

Transportation as a tool to shape land use

It has been observed that the underpricing of road travel, especially as it relates to the unpriced congestion externality makes travel cheaper than its marginal cost. This not only causes excessive urban expansion but also induces planners to use faulty cost–benefit measures and thus invest in too much road building as argued by Kraus, Mohring and Pinfold (1976). Excessive road capacity in turn reinforces the excessive urban expansion.

In view of the many pecuniary externalities of transportation, and since perfect pricing is not possible, a combination of judicious capacity provision and land-use zoning to ensure better accessibility to main roads and rail lines could have significant benefits. Such economies of transport–land use interdependence may be possible to exploit in urban planning and urban design at the level of smaller areas and neighbourhoods. Boarnet and Crane (2000) have examined whether land use policy and urban form can significantly affect travel behaviour in such settings. Similar concerns exist at the macro urban level (Gordon, Kumar and Richardson, 1989). In the future, planners could use such knowledge when major decisions are made on how much capacity to supply, where to supply it and how much to restrict development around it. More often than not, however, when urban planners intervene with land use controls they may fail to find the golden rule, causing distortions in land markets that could outweigh the efficiencies that can be gained by influencing travel.

See Also

- Coase theorem
- congestion
- external economies
- George, Henry
- land markets
- Pigouvian taxes
- time use
- urban economics
- value of time

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Anas, A. and Duann, L. 1985. Dynamic forecasting of travel demand, residential location and land development: policy simulations with the Chicago area transportation/land use analysis system. *Papers of the Regional Science Association* 56, 38–58.

Anas, A. and Xu, R. 1999. Congestion, land use and job dispersion: a general equilibrium model. *Journal of Urban Economics* 45, 451–73.

Arnott, R., de Palma, A. and Lindsey, R. 1990. Departure time and route choice for the morning commute. *Transportation Research B* 24, 209–28.

Arnott, R. and Kraus, M. 2003. Principles of transport economics. In *Handbook of Transportation Science*, 2nd edn., ed. R. Hall. Boston: Kluwer.

Arnott, R. and MacKinnon, J. 1978. Market and shadow land rents with congestion. *American Economic Review* 68, 588–600.

Beckmann, M., McGuire, C. and Winsten, C. 1956. *Studies in the Economics of Transportation*. New Haven, CT: Yale University Press.

Boarnet, M. and Crane, R. 2000. *Travel by Design: The Influence of Urban Form on Travel*. New York: Oxford University Press.

Daganzo, C. and Sheffi, Y. 1977. On stochastic models of traffic assignment. *Transportation Science* 11, 253–74.

George, H. 1879. *Progress and Poverty*. New York: Robert Shalckenbach Foundation, 1955.

Gordon, P., Kumar, A. and Richardson, H. 1989. The influence of metropolitan spatial structure on commuting time. *Journal of Urban Economics* 26, 138–51.

Keeler, T. and Small, K. 1977. Optimal peak-load pricing, investment, and service levels on urban expressways. *Journal of Political Economy* 85, 1–25.

Kim, T. 1979. Alternative transportation modes in an urban land use model: a general equilibrium approach. *Journal of Urban Economics* 6, 197–215.

Kraus, M., Mohring, H. and Pinfold, T. 1976. The welfare costs of non-optimum pricing and investment policies for freeway transportation. *American Economic Review* 66, 532–47.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

McFadden, D. 1973. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McMillen, D. and McDonald, J. 2004. Reaction of house prices to a new rapid transit line: Chicago's Midway line, 1983–1999. *Real Estate Economics* 32, 462–86.

Meyer, J., Kain, J. and Wohl, M. 1965. *The Urban Transportation Problem*. Cambridge, MA: Harvard University Press.

Mills, E. 1972. Markets and efficient resource allocation in urban areas. *Swedish Journal of Economics* 74, 100–13.

Pigou, A. 1947. *A Study in Public Finance*, 3rd edn. London: Macmillan.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36, 387–9.

Small, K. and Verhoef, E. 2006. *Urban Transportation Economics*, 2nd edn. London: Routledge.

Sullivan, A. 1983. The general equilibrium effects of congestion externalities. *Journal of Urban Economics* 14, 80–104.

Vickrey, W. 1969. Congestion theory and transport investment. *American Economic Review, Papers and Proceedings* 59, 251–60.

How to cite this article

Anas, Alex. "urban transportation economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000062>
doi:10.1057/9780230226203.1776

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

urbanization

Sukkoo Kim
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Cities first arose in the Fertile Crescent a few thousand years after the discovery of agriculture. Yet the history of urbanization is not one of steady progress. Pre-industrial urbanization rose with technological advances in agriculture and transportation which fostered population growth and trade, but fell with famine and disease. Just as important, cities rose and fell with the military fortunes of city states, territorial empires and nation states. With the Industrial Revolution, urbanization rose dramatically. As population shifted out of agriculture into manufacturing and services, cities became the dominant landscape of human civilization.

Keywords

amenities; civilization; division of labour; early industrialization; face-to-face interaction; feudalism; globalization; Industrial Revolution; labour markets; labour productivity; Marshallian externalities; Middle East; North, D.; transaction costs; transportation costs; transportation revolution; urbanization

Article

The rise of cities in the ancient Middle East

The first city in human history is believed to have emerged around 3200 bc in Sumer, Mesopotamia, between the Tigris and Euphrates rivers, as a consequence of the Neolithic Revolution which saw a shift in food production from hunting and gathering to agriculture based on domesticated plants and animals (Childe, 1950). The emergence of cities in Sumer marked the beginning of an ‘urban revolution’, but the revolution was an exceedingly slow one. Since agriculture began in the Fertile Crescent around 8500 bc, the first city emerged several thousand years after the discovery of agriculture. Moreover, the emergence of cities was not unique to Mesopotamia. Interestingly, cities emerged independently in at least two other places, China and the New World, places where major domestication of plants and animals arose

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
independently.
While it is extremely difficult to determine the causes of the emergence of cities in ancient times, scholars such as Childe (1950) and Bairoch (1988) believe that agriculture caused cities to form because it increased population growth and provided surplus food for a non-agricultural population. Since demand for food is believed to be income inelastic, an increase in income from a rise in agricultural productivity will increase demand for secondary and tertiary products (Wrigley, 1987). The urban concentration of secondary and tertiary employment, such as crafts and commerce, enabled the exploitation of the division of labour, fostered technological innovations in many areas of the economy from irrigation, transportation, metallurgy to writing, and lowered the costs of coordinating long-distance trade.
Even more importantly, cities were centres of states before the rise of territorial empires and nation states. A city state was composed of a governing city and its food-producing hinterlands. It was a distinct geographical, political and economic unit of organization. By establishing a body of formal and informal rules of property rights, city states provided their citizens with the incentives to improve productivity or, more fundamentally, to acquire more knowledge (North, 1981). Cities probably became centres of government because close face-to-face interactions between the ruling elite and its administration increased the efficiency of governing by lowering the costs of generating, collecting and processing information. In addition, dense, walled cities provided effective defence against raiders.
The cities that arose in Sumer, Mesopotamia, were independent city states. Archaeological evidence indicates that there may have been as many as 15 city states by 3000 bc. A typical city state may have contained population of 25,000 with rural population of about 500,000. Hammond (1972) suggests that the impetus for city states in Sumer may have been the need to coordinate the provision of public goods such as large-scale irrigation, drainage, communal storage, and defence against other city states. Over the following millennium, the number and size of cities grew in this region, some reaching populations of 100,000 or more.
When the Sumerian cities were conquered by Babylon by 1800 bc, the era of early city states gave way to the era of territorial empires such as those of Babylon, Egypt and Canaan. Scholars generally believe that urbanization suffered under these empires except for cities, like Babylon, that served as political and military centres. Babylon at its height may have reached populations of 200,000–300,000. While the exact causes of the rise of these territorial empires are not clear, among them may have been the growing benefits of trade over longer distances and changes in military technology which allowed for the control of larger areas.
Of the major empires, it was in Egypt that cities declined most significantly. Indeed, many scholars would argue that Egypt was an empire without cities. To the extent that cities existed in Egypt, they were centres of religion and administration. The lack of cities in Egypt is often attributed to the Pharaoh's centralized control of irrigation, trade and all other facets of the economy (Hammond, 1972).
Cities also declined in the other territorial empires, but probably less, as many remained relatively independent. Despite the general decline of urbanization during this period, a new type of coastal city emerged in Phoenicia. These cities, which grew in numbers around 1200 bc, arose principally to trade goods throughout the Aegean and the Mediterranean.
Cities first emerged in the Middle East, but reached their greatest heights in the Mediterranean in the ancient era. Most likely, cities arose later in the Mediterranean because agriculture arrived in this region a century and a half after its discovery in the Fertile Crescent. It arose first in the Fertile Crescent

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
because of favourable geography and climate, which provided an abundance of indigenous species suitable for domestication (Diamond, 1997). From there, agriculture spread about 0.7 miles per year and reached the Mediterranean region sometime between 7000 and 6000 bc. By 2000–1450 bc, Aegean civilization seems to have constructed small city states with relatively limited agricultural hinterland and trade. However, urbanization flourished under the Greek and Roman civilizations, as the Mediterranean Sea became the highway of transport and communications.

The Greeks formed small, independent city states composed of coastal cities and their adjoining farmlands between 800 and 146 bc. Their largest city, Athens, may have reached 100,000 in population, but most other cities rarely exceeded 40,000. Bairoch (1988) estimates that perhaps as much as 15–25 per cent of the Greek population lived in cities with more than 5,000 inhabitants. Because the Greek soil was relatively poor, a large portion of the population was engaged in local and long-distance trade in the Mediterranean. The Greek invention of coined money facilitated market exchanges. The Greek cities are also known for their political innovations. While most city states and territorial empires, with the exception of the Phoenician city states, were ruled by monarchy or aristocracy, democracy arose in many Greek cities.

The formation of the Roman Empire between 146 bc and ad 300 represented the largest political and economic integration of territory in the ancient world. Rome, as the military and administrative centre, grew to an astonishing size, surpassing 800,000 inhabitants and perhaps reaching a million by ad 2. Unlike in the earlier territorial empires such as Egypt, cities prospered under the Roman Empire. Politically, cities became military and administrative centres and collected taxes from the surrounding countryside for Rome; economically, cities acted much like independent city states. Under conditions of peace and uniform law a great numbers of cities emerged as commercial activity increased. Bairoch (1988) estimates that about 8–15 per cent of the population resided in cities. While most cities were small, 20 or more may have reached 20,000 or more inhabitants.

When the Roman Empire disintegrated around ad 476, it signalled the decline of the Mediterranean world. In the resulting so-called Dark Ages between ad 500 and 800, urbanization fell as frequent wars and invasions contributed to economic insecurity. But the two centuries following this period were a period of urban renaissance. By this time, Europe was divided into two regions. The southern Mediterranean part was conquered by Arab Muslims while the northern part was composed of Christian Europe. In Muslim Spain, the urban population rose rapidly. In Christian Europe, urbanization grew in some places for defensive reasons, but in Italy commercial city states rose to great heights (Bairoch, 1988). Ruled by merchant elites, Italian city states engaged in extensive long-distance trade using newly developed technology such as the compass, but cities like Venice also grew because their naval powers provided protection for their ships in the Adriatic Sea and beyond (Lane, 1973).

The growth of cities in western Europe

The period called the Middle Ages, between ad 1000 and 1500, marked the beginning of the rise of western Europe. Because this region was further away from, and possessed a very different climate from the Fertile Crescent, agriculture arrived between 6000 and 2500 bc, much later than in the Mediterranean (Diamond, 1997). In the Middle Ages, technological advances such as the heavier plough, horse collar, and the three-field system increased the agricultural productivity of western Europe significantly.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Between 1000 and 1340, there was a strong growth in urbanization based on feudal city states. Under feudalism, a lord provided protection using a castle and knights; in exchange for protection, slaves, serfs and free labourers offered free labour.

The rise of western Europe was interrupted between 1300 and 1490 as famine, disease and wars led to major losses of population and de-urbanization. Famines between 1315 and 1317, the plague and the Black Death between 1330 and 1340 and between 1380 and 1400, and series of wars, civil wars, feudal rebellion and banditry afflicted much of Europe and contributed to the decline of the medieval economy (Palmer and Colton, 1965). Urbanization fell as feudal city states crumbled. North (1981) argues that a series of major technological advances in military warfare, such as the introduction of the pike, longbow, cannon, and eventually the musket, as well as a growing market for mercenaries, contributed to the decline of feudalism since a feudal lord could no longer provide adequate protection for his manor against these new military developments.

The modern period in history begins with the year 1500. Between 1500 and 1800, the opening of Atlantic commerce and the formation of nation states fundamentally transformed Europe and the world. Advances in ocean shipping, which led to the discovery of maritime routes to the Americas and Asia, ushered in a new era of international trade. Although cities in the Mediterranean remained important, the focus of urbanization shifted toward the Atlantic as urbanization grew rapidly in nations with easy access to the open ocean such as Portugal, Spain, the United Kingdom and the Netherlands. In addition, nation states based on monarchies arose throughout Europe and established colonies in Africa and the New World. These new nation states were supported by an unprecedented growth in military and civil administration financed by taxes and debt (Brewer, 1990).

The rise of western Europe and globalization were accompanied by a significant growth in urbanization. In Europe, while the upward trend was not uniform over time, de Vries (1984) finds that the number of cities with populations of at least 10,000 rose from 154 to 364 between 1500 and 1800. The growth in urbanization was concentrated in the very largest cities, whose main functions were to serve either as a government capital or as a port city (de Vries, 1984; Bairoch, 1988). The concentration of merchants in cities lowered the costs of financing and coordinating trade around the globe. Likewise, governments became concentrated in cities since the efficiency of military and government operations involved the collection and processing of an enormous amount of information (Brewer, 1990).

The pre-industrial cities in the Middle East and Europe possessed a variety of political regimes across space and over time. Most often, city states were ruled by kings with absolute power, but in some instances they were ruled by merchant elites. While the rulers of city states provided order and stability, they also imposed heavy tax burdens on their subjects. Between 1000 and 1800, city states that were ruled by absolutist governments grew much less significantly than those governed by merchants or assemblies (De Long and Shleifer, 1993).

Industrialization and urbanization

The Industrial Revolution, which began in Britain around 1700, transformed the modern world in a short period of time. In pre-industrial times, which spanned thousands of years from the ancient and medieval periods to the first two centuries of the modern era, cities became important centres of government, trade and artisan manufacturing, but most of the population lived in rural farms and villages. Yet, within a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
little over a century after the onset of the Industrial Revolution, the majority of people in Britain lived in cities. As industrialization spread to other European nations and the United States in the 19th century, and eventually to an ever growing list of nations in the 20th century, world urbanization rose rapidly.
The Industrial Revolution's transformation of the urban order began in the countryside. The early factories arose in rural villages and towns rather than in established major urban centres, and less urbanized places industrialized more rapidly in Europe and in the United States (Bairoch, 1988).
However, as industrialization matured it was significantly correlated with urbanization. The rise of the manufacturing sector was not only responsible for the emergence of new industrial cities, but also contributed to the growth of traditional urban centres. Between 1800 and 1900, the share of the urban population in industrialized countries almost tripled, from 11 to 30 per cent (Bairoch, 1988).
It remains unclear why early industrialization was rural or why late industrialization was urban. Scholars generally believe that early factories chose rural locations because of the availability of water power, coal or unskilled labour (Bairoch, 1988). Because early factories used women and child labour intensely, Goldin and Sokoloff (1984) believe that, in the United States, early industrialization arose in rural New England rather than in the rural South because the relative labour productivity of women and children was less than that of men in the former region. For Rosenberg and Trajtenberg (2004), industrialization caused urbanization as firms adopted the Corliss steam engine as their primary power source.
Kim (2005a; 2005b) suggests that explanations of why industrialization led to urbanization are likely to rest on the rise of division of labour and the labour market. Prior to industrialization, goods were produced by self-employed artisans who made the entire product. With industrialization, factory owners hired workers in a labour market and employed them in specialized tasks. Because early industrialization was concentrated in a limited number of industries and was limited in scale, firms located in rural places since the costs of recruiting workers were relatively modest. However, as industrialization rose in scale and spread to numerous industries, the agglomeration of workers and firms in cities deepened the extent of the division of labour and lowered labour market transaction costs.
One of the major developments associated with the Industrial Revolution was a transportation revolution. In the pre-industrial period, the bulk of long-distance trade occurred over bodies of water such as canals, rivers, lakes, seas and oceans. With the introduction of the railroad and later trucks and airplanes, overland transportation costs fell dramatically. In the United States, the integration of regional economies led to a significant increase and then decrease in regional specialization (Kim, 1995). The rise of US regional trade led to the rise of many large inland cities like Chicago, which emerged to coordinate the increase in domestic trade.
In the second half of the 20th century, although the rate of urbanization slowed in industrialized nations, cities remain a vital component of the modern economy. Scholars believe that one or more factors, such as Marshallian externalities (technological spillovers, non-traded industry specific inputs, and labour market pooling), market size and natural advantages, cause the formation and growth of cities (Henderson and Thisse, 2004). Moreover, while economic factors are much more significant in modern than in pre-modern times, political factors, such as tariffs and dictatorships, remain important for urbanization (Ades and Glaeser, 1995).

Patterns of world urbanization

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Urbanization was not confined to the Middle East and Europe. Cities arose indigenously in India (1000–400 bc), China (700–400 bc) and the New World (100 bc). Cities diffused to Korea (108 bc–ad 313) and Japan (ad 650–700) from China, and to south-west Asia (ad 700–800) from China and India. In the New World, cities arose independently in Mesoamerica in Mexico and the Andes in South America. With the arrival of maize agriculture from Mesoamerica, urbanization reached the south-western and eastern parts of North America. Whereas China and Japan contained some of the largest pre-industrial cities, and had urbanization rates comparable to, or perhaps even higher than, pre-industrial European societies, the cities in the New World were smaller, fewer in number, and less stable.

In Africa, there is considerable uncertainty as to whether agriculture and cities arose independently. While there is evidence of domesticated agriculture in Sahel (5000 bc) and tropical West Africa (3000 bc), scholars remain unsure whether founder crops arrived from elsewhere (Diamond, 1997). There is evidence of cities in Africa as early as 1000 bc, but it is also not clear whether these cities resulted from outside influences (Bairoch, 1988). In Australia, no cities arose indigenously as the aboriginal population remained hunters and gatherers.

The coming of the modern era in 1500 not only transformed western Europe but also decisively altered the path of development around the world. As European nations colonized the New World and parts of Africa and Asia, they transplanted their technologies, agriculture, germs and political institutions to their colonies. From the colonies, the Europeans extracted new plants and resources and traded them around the globe. The demography of the New World colonies was fundamentally altered as the native population suffered and, to a varying extent, was supplanted by European immigrants and African slaves. In general, colonization and globalization were accompanied by a rapid growth in urbanization in Europe and the colonies.

In the 19th and the 20th centuries, the uneven diffusion of the Industrial Revolution around the globe determined the patterns of world urbanization. While there is no general consensus on the causes of uneven economic development, explanations usually rest on one or more factors related to geography, technology, trade and institutions (see Aghion and Durlauf, 2005). In nations that developed, industrialization led to a rapid growth in urbanization; in those that did not, urbanization remained relatively low. In addition, most of the urban population in poor nations became concentrated in a handful of very large cities.

Summary

Cities in history arose with the advent of agriculture as they became centres of governments, crafts, religion and universities. As markets and trade developed, cities became centres of finance and commerce. While the patterns of urbanization differed greatly over space and time, the causes of urbanization were the same around the world. Pre-industrial urbanization rose with technological advances in agriculture, artisan manufacturing, transportation and trade, but fell with environmental degradation, famine and disease. Just as importantly, cities rose and fell with military and administrative efficiency of city states, territorial empires and nation states.

With the Industrial Revolution, cities became centres of factories and offices. Although early industrialization began in rural areas, the shift in the employment composition from agriculture to manufacturing and services led to rapid urbanization. The formation and growth of industrial or modern cities are attributed to benefits from a variety of market and non-market factors such as the division of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
labour, lower search costs of matching specialized workers and firms, information spillovers, market size, and non-traded intermediate inputs (Henderson and Thisse, 2004). With the rise in disposable incomes, cities became centres of arts, entertainment and other amenities.

The history of civilization is the history of urbanization. Without cities, the pillars of civilization – literature, science and the arts – would not exist. But as the industrial era gives way to the information era, will cities disappear? Leamer and Storper (2001) believe that face-to-face interactions in cities are likely to remain important for some time to come. For these scholars, the coordination of complex, unfamiliar and innovative activities depends on the successful transfer of uncodifiable messages and requires long-term relationships, trust, closeness and agglomerations. Wherever the new innovative activity may arise, be it in commerce, finance, politics, arts or science, the future of civilization is likely to rest on the success of its citizens.

See Also

- central place theory
- location theory
- marketplaces
- new economic geography
- regional development, geography of
- spatial economics
- systems of cities
- urban agglomeration
- urban economics
- urban growth
- urban political economy
- urban production externalities

Bibliography

Ades, A. and Glaeser, E. 1995. Trade and circuses: explaining urban giants. *Quarterly Journal of Economics* 110, 195–227.

Aghion, P. and Durlauf, S., eds. 2005. *Handbook of Economic Growth*. Amsterdam: Elsevier.

Bairoch, P. 1988. *Cities and Economic Development*. Chicago: University of Chicago Press.

Brewer, J. 1990. *The Sinews of Power: War, Money and the English State, 1688–1783*. Cambridge, MA: Harvard University Press.

Childe, V. 1950. The urban revolution. *Town Planning Review* 21, 3–17.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

De Long, J. and Shleifer, A. 1993. Princes and merchants: European city growth before the Industrial Revolution. *Journal of Law and Economics* 36, 671–702.

de Vries, J., 1984. *European Urbanization 1500–1800*. Cambridge, MA: Harvard University Press.

Diamond, J. 1997. *Guns, Germs, and Steel*. New York: W.W. Norton.

Goldin, C. and Sokoloff, K. 1984. The relative productivity hypothesis of industrialization: the American case, 1820–1850. *Quarterly Journal of Economics* 99, 461–88.

Hammond, M. 1972. *The City in the Ancient World*. Cambridge, MA: Harvard University Press.

Henderson, J. and Thisse, J.-F., eds. 2004. *Handbook of Regional and Urban Economics*, vol. 4. Amsterdam: Elsevier.

Kim, S. 1995. Expansion of markets and the geographic distribution of economic activities: the trends in U.S. regional manufacturing structure, 1860–1987. *Quarterly Journal of Economics* 110, 881–908.

Kim, S. 2005a. Industrialization and urbanization: did the steam engine contribute to the growth of cities? *Explorations in Economic History* 42, 586–98.

Kim, S. 2005b. Division of labor and the rise of cities: evidence from U.S. industrialization, 1850–1880. Mimeo. St. Louis: Washington University.

Lane, F. 1973. *Venice: A Maritime Republic*. Baltimore, MD: Johns Hopkins University Press.

Leamer, E. and Storper, M. 2001. The economic geography of the internet age. *Journal of International Business Studies* 32, 641–65.

North, D. 1981. *Structure and Change in Economic History*. New York: W.W. Norton.

Palmer, R. and Colton, J. 1965. *The History of the Modern World*. New York: Alfred Knoft.

Rosenberg, N. and Trajtenberg, M. 2004. A general purpose technology at work: the Corliss steam engine in the late-nineteenth-century United States. *Journal of Economic History* 64, 61–99.

Wrigley, E. 1987. *People, Cities and Wealth*. Oxford: Basil Blackwell.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Kim, Sukkoo. "urbanization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000066> doi:10.1057/9780230226203.1777

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Ure, Andrew (1778 – 1857)

William Lazonick
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

manufacturing; Marx, Karl; technology; Ure, Andrew

Article

Andrew Ure, MD, was professor of chemistry and natural science at Anderson's College, Glasgow from 1804 to 1830. In 1830, he introduced the word ‘thermostat’ into the English language in conjunction with a patent that he secured (Standfort, 1982, p. 659). At about the same time, he moved to London to serve as a consultant in analytical chemistry to the Board of Customs. From 1832 to 1834, his major research assignment was to ascertain the wastage rate of raw material in sugar refining in order to determine the rebates on raw sugar import duties that British refiners could legitimately claim. Ure (1843, p. iv) complained that his research saved the exchequer £300,000 but yielded him only £800 in remuneration and cost him his health.

To recuperate, he ‘spent several months in wandering through the factory districts of Lancashire, Cheshire, Derbyshire, &c., with the happiest results to his health; having everywhere experienced the utmost kindness and liberality from the mill-proprietors’ (Ure, 1835, p. viii). Two important books were the result. *The Philosophy of Manufactures* (1835) and *The Cotton Manufacture of Great Britain* (1836) are detailed technical treatises on the industry at the heart of Britain's industrial revolution, interlaced with commentary on the salutary moral, intellectual and physical effects of factory life on the workers.

Ironically, it was Karl Marx who established Ure's place in the history of economics. Ure's blatantly pro-capitalist stance, combined with his obvious technical expertise, made him the perfect ‘horse's mouth’ in Marx's attempt to show how capitalists used technology to throw adult males out of work and turn women and children into mere appendages of the machine. Marx (1867 [1977], pp. 560, 563–4) invoked the authority of Ure, who, in the conflict-ridden 1830s, argued that the diffusion of more automated technology would ‘put an end … to the folly of trades' unions’, proving that ‘when capital enlists science into her service, the refractory hand of labour will be taught docility’ (Ure, 1835, pp. 23, 368).

Writing some three decades later, however, Marx failed to distinguish pro-capitalist ideology from ongoing reality. Contrary to Ure and Marx, adult male workers had not been definitively humbled, even

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in the presence of mechanization. Rather, certain groups of workers had maintained substantial control of work organization and had built up considerable union power (Lazonick, 1979). In effect, Marx's uncritical use of Ure provided the ‘evidence’ needed to confirm that, in their confrontation with capitalists armed with technology, workers had ‘nothing to lose but their chains’. Theory and history were parting company in Marx's theory of capitalist development (Lazonick, 1986).

See Also

- Taylorism

Selected works

1835. *The Philosophy of Manufactures*. London: Knight.

1836. *The Cotton Manufacture of Great Britain*. London: Knight.

1843. *The Revenue in Jeopardy from Spurious Chemistry*. London: Ridgway.

Bibliography

Lazonick, W. 1979. Industrial relations and technical change: the case of the self-acting mule. *Cambridge Journal of Economics* 3(September), 231–62.

Lazonick, W. 1986. Theory and history in Marxian economics. In *The Future of Economic History*, ed. A.J. Field. The Hague: Kluwer-Nijhoff.

Marx, K. 1867. *Capital*, Volume I, New York: Vintage, 1977.

Standfort, J.T. 1982. Thermostat. In *Encyclopedia Americana*, Vol. 26, New York: Grolier.

How to cite this article

Lazonick, William. "Ure, Andrew (1778–1857)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000037> doi:10.1057/9780230226203.1869

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

user fees

Edwin S. Mills
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

benefit taxes; Mills, E. S.; Tiebout hypothesis; user charges: *see* user fees; user fees

Article

The genealogy of the term ‘user fees’ (or, synonymously, ‘user charges’) is neither long nor coherent. Neither Marshall nor Pigou appears to have used the term. The term was in common usage in the USA during the early post-World War II years; see e.g. Stockfisch (1960). Throughout its short history, the term seems to have been employed much more frequently in the United States than elsewhere. Since about 1970, the term has appeared in the indexes of most US public finance textbooks.

No writer has provided a careful definition of the term or distinguished it from similar terms. Rosen (1985) defines a user fee as a price charged for a commodity or service produced by a government. Some writers appear to restrict the term to charges for services produced by governments. To add confusion, many writers apply the term ‘user fee’ to charges levied by a government for the discharge of wastes to the air and water environment. In this usage, the term is synonymous with ‘effluent fee’. Although most economists believe that governments should protect the environment by fees or regulations, since the environment has the characteristics of a public good, the environment is not in any reasonable sense a commodity or service produced by a government.

How is a user fee distinguished from two related concepts, a benefit tax and a price? There is a legal and constitutional difference between fees and taxes levied by governments, but the issue here is the economic content of the terms.

A benefit tax is any tax levied proportionately to benefits received by the taxpayer from a commodity or service provided by a government. The appropriate distinction is that a fee is paid only if the consumer decides freely to consume the commodity or service, whereas the taxpayer may be forced to pay a benefit tax even though he or she is not free to decide whether to consume the commodity or service. If the term ‘benefit tax’ is restricted to taxes whose amounts are no greater than the value to the taxpayer of the commodity or service consumed, the important distinction disappears. No rational consumer would refuse to pay a tax which is less than the benefit which the consumer receives from commodities or

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
services financed by the tax. Thus, the distinction between voluntary and involuntary payment becomes unimportant. In practice, governments levy many taxes in the name of benefits even though they are larger than the benefits derived from the commodity or service provided. Most ostensibly benefit taxes are only approximations to fees for the commodity or service consumed. A gasoline tax is an approximation to a fee for road consumption or congestion and pollution externalities. The Tiebout theory (1956) implies that all local government taxes can be viewed as benefit taxes.

There seems to be no important distinction between a user fee and a price except that the term ‘user fee’ is used when government is the supplier. Public finance economists frequently use the term ‘user fee’ when reference is to a service, such as electricity, provided by government, even though the service is sometimes provided by private suppliers and the charge is then referred to as a price.

Why make the distinction between a user fee and a price? There seems to be no justification except to identify the supplier. Yet that typically is, and always could be, clear from the context. It appears to be unjustified to coin a different, and indeed clumsy, term merely to identify the supplier. One suspects that some intellectual product differentiation is behind the distinction.

See Also

- environmental economics
- externalities

Bibliography

Rosen, H. 1985. *Public Finance*. Homewood, IL: Richard D. Irwin.

Stockfisch, J.A. 1960. Fees and service charges as a source of city revenue: a case study of Los Angeles. *National Tax Journal* 13(2), 97–121.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64, 416–24.

How to cite this article

Mills, Edwin S. "user fees." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000041> doi:10.1057/9780230226203.1778

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Usher, Abbot Payson (1884 – 1965)

William N. Parker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Clapham, J. H.; economic history; inventions; Kondratieff cycles; measurement; Schumpeter, J. A.; technical change; Usher, A. P.

Article

Usher occupied the chair of European economic history at Harvard University from 1936 to 1949 and was surely the most productive and original scholar to occupy this post. For economists of later decades, his most significant book was *A History of Mechanical Inventions* (1st edn, 1929; 2nd edn, 1954). In it he identified invention as a four-stage process in which the individual inventor, being seized of a problem in the presence of the intellectual and physical elements for a solution, achieves the primary insight (called by Usher the ‘saltatory act’ and by his students the ‘ah-ha!’ or ‘Eureka’ moment) and completes the invention through a stage of ‘critical revision’. Usher's work here became noticed by economists when it was taken up by J.A. Schumpeter to form the historical basis of his descriptive and theoretical work on *Business Cycles* (1939) and also through its relation to the Kondratieff ‘long waves’ based on the clustering of a few major inventions at discrete points in and around the 19th century (1770–80, 1840–60, 1890–1910). At a time when economists treated technological change as an element as exogenous to economics as physical geography, Usher alone thought it worth examining as a complex socio-economic ‘thread’ in history. In this he was the forerunner of such modern students as Schmookler, Mansfield, Ruttan, Nelson and Rosenberg, though his book largely emphasized the technical (supply-side) aspects of the process.

The identification of Usher with the study of technological change is unfortunate, since his many monographs and articles, his two textbooks, and his classroom teaching reveal a comprehensive grasp of the experience of the West in economic life and organization in all its major aspects. It is perhaps fair to say that his mind was fascinated with those points where societies face nature. Population growth, geographical resource patterns, transport, industrial location, technology, physical costs and physical constraints on social action and organization were the themes around which his view of economic history was organized. His insights were those of the engineer, not those of the sociologist.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This bias in Usher's work undoubtedly derived from a deep-seated liberal ideology which regulated both his topics and his methods of research. He looked on economic history not simply as an interesting outlet for scientific curiosity but as an instrument to help societies achieve a rational control over their environment. But, himself the most modest and unpolitical of men, he evidently saw little use to studies where the social control to which they could lead impinged on the individual's personal and private life and values. And methodologically Usher was a committed empiricist. He evinced, and often reiterated, a deep distrust of what he called the idealistic formulations of Marx, Weber and Parsons. Yet his admiration of the British school typified by Clapham was moderated by an uneasiness over its commitment to a purely literary or descriptive methodology. He was most at home in the study of specific limited topics in which a quantifiable trend could be observed over a long period and where measurement and economic theory of a Marshallian variety could be employed. His concrete applications of the German theories and models of industrial location were particularly powerful, and inspired the later work of E. Hoover, W. Isard and others. He must be accounted, along with S. Kuznets and A. Gerschenkron as a patriarch of the so-called ‘new’ economic history in the United States, and of those three, Usher's grasp of the relation between theory, measurement and the phenomena of historical change must be accounted to have been the most philosophical and careful, and the best exemplified in concrete historical studies.

Selected works

A bibliography of Usher's writings is contained in Lambie (1956). This volume also contains an essay on Usher's thought and writings. See also the article by John Dales in the *International Encyclopedia of the Social Sciences*, vol. 16 (1968), and the generous memorial tribute by A. Gerschenkron, retained in the files of the Harvard Department of Economics.
Among Usher's books and articles, *A History of Mechanical Inventions* (1929; 1954), and his two advanced level texts, *An Introduction to the Industrial History of England* (1920) and (with W. Bowden and M. Karpovich) *An Economic History of Europe since 1750* (1937), were the most durable and highly valued by students. His major contributions to early modern European economic history are *The History of the Grain Trade in France, 1400–1710* (1913) and *The Early History of Deposit Banking in Mediterranean Europe* (1943). Usher's attitudes toward economic history and methodology are best stated in three articles, (1932, 1949 and 1951) and in Chapter 4 of *A History of Mechanical Inventions*. His attitude toward economics and economic policy is well stated in his 1934 address to the American Economic Association.

1913. *The History of the Grain Trade in France, 1400–1710*. Cambridge, MA: Harvard University Press.

1920. *An Introduction to the Industrial History of England*. Boston: Houghton Miffin.

1929. *A History of Mechanical Inventions*. 2nd edn, Cambridge, MA: Harvard University Press, 1954.

1932. The application of the quantitative method to economic history. *Journal of Political Economy* 40, 186–209.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1934. A liberal theory of constructive statecraft (Address to the American Economic Association). *American Economic Review, Papers and Proceedings* 24, 1–10.

1937. (With W. Bowden and M. Karpovich.) *An Economic History of Europe since 1750*. New York: American Book Company.

1943. *The Early History of Deposit Banking in Mediterranean Europe*. Cambridge, MA: Harvard University Press.

1949. The significance of modern empiricism for history and economics. *Journal of Economic History* 9, 137–55.

1951. Sir John Howard Clapham and the empirical reaction in economic history. *Journal of Economic History* 11, 148–53.

Bibliography

Dales, J. 1968. Usher, Abbot Payson. In *International Encyclopedia of the Social Sciences*, vol. 16. New York: Macmillan.

Gerschenkron, A. 1965. Abbot Payson Usher: a memorial tribute. Files of the Department of Economics, Harvard University.

Lambie, J., ed. 1956. *Architects and Craftsmen in History. Festschrift für Abbot Payson Usher*. Veröffentlichungen der List Gesellschaft, vol. 2. Tübingen: J.C.B. Mohr (Paul Siebeck).

Schumpeter, J.A. 1939. *Business Cycles*. 2 vols. New York: McGraw-Hill.

How to cite this article

Parker, William N. "Usher, Abbot Payson (1884–1965)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000042> doi:10.1057/9780230226203.1779

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

utilitarianism and economic theory

Jonathan Riley
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Utilitarianism is a family of moral and political philosophies according to which general utility or social welfare is ultimately the sole ethical value or good to be maximized. Normative economics endorsed a hedonistic version of utilitarianism from the latter part of the 18th century well into the 20th century. Despite the ordinalist revolution, some version of utilitarianism continues implicitly to serve as the ethical basis for economic policy judgements. While there are signs that this may be changing, economic theory has not yet moved decisively beyond utilitarianism, nor is it clear that it should.

Keywords

abundance; bargaining; Bentham, J.; Binmore, K.; cardinal utility; checks and balances; classical economics; consequentialism; contract curve; contractualism; distributive justice; Edgeworth, F. Y.; enlightened self-interest; equality; ethical pluralism; Harsanyi, J. C.; hedonistic utilitarianism; Helvetius, C. A.; Hume, D.; individuality; interpersonal utility comparisons; Jevons, W. S.; leximin; liberal democracy; majority rule; Marshall, A.; Maximin; Mill, J. S.; Mirrlees, J.; Moore, G. E.; natural justice; naturalistic fallacy; normative economics; optimal taxation; ordinal utility; outcome utilitarianism; proportional bargaining; rational choice utilitarianism; Rawls, J.; representative democracy; revealed preference theory; security; Sen, A.; Sidgwick, H.; social contract; Spencer, H.; subsistence; utilitarianism; utilitarianism and economic theory; utility; value; veil of ignorance; von Neumann and Morgenstern; welfarism

Article

Utilitarianism is a family of moral and political philosophies according to which general utility or social welfare is ultimately the sole ethical value or good to be maximized.
Amartya Sen (1979) suggests that members of the family typically combine ‘outcome utilitarianism’,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
which, at least if population is not a variable, identifies the goodness of an outcome with the sum of individual utilities at that outcome, and some version of ‘consequentialism’, which focuses on selected means of achieving outcomes, such as acts, rules, dispositions, or some combination of these, and identifies the right means in any situation as those which result in an optimal feasible outcome according to outcome utilitarianism. (For analysis of the issues raised when population is a variable, see Blackorby, Bossert and Donaldson, 2005.) Outcomes can be defined to include the acts, rules, or other means that lead to them. Thus, for any set of possible outcomes, a typical version of utilitarianism calculates social welfare $W(x)$ at any outcome × by adding together the individual utilities at x:

$$\text{for all } x\text{: } W(x) = \sum u_i(x). \tag{1}$$

The doctrine then prescribes as best an outcome x^* at which the sum of utilities is maximized, that is, $W(x^*) = \max \Sigma u_i(x)$.

As Sen also points out, outcome utilitarianism can be factorized into ‘sum-ranking’, the claim that the proper way to aggregate individual utilities is to add them, and ‘welfarism’, the principle that the goodness of an outcome is an increasing function of the set of individual utilities so that non-utility information relating to any outcome is of no ethical import. Welfarism implies ‘Paretianism’, the claim that one outcome is better than another if (but not only if) at least somebody has more utility whereas nobody has less utility in the one outcome than in the other. Full axiomatizations of outcome utilitarianism have been provided by d’Aspremont and Gevers (1977) and Maskin (1978).

Different versions of utilitarianism may vary not only in terms of their consequentialist structures but also in terms of their conceptions of utility. Hedonistic utilitarianism conceives of utility as pleasure including freedom from pain, for example, whereas rational choice utilitarianism conceives of utility as a numerical representation of a preference revealed by consistent choice behaviour which is in principle observable. Sen himself arguably takes a quite restricted view of the utilitarian family because he seems to take for granted that utility, although capable of bearing plural interpretations, can be seen only as a simple phenomenon that remains the same whatever its sources and intended objects. As a result, welfarism and Paretianism cannot accommodate the idea of different aspects or kinds of utility, some of which may be intrinsically more valuable than others, where each kind is inseparably associated with a distinctive mix of human capacities, resources and/or institutional activities. Thus, as an alternative to utilitarianism, Sen (1980) introduces the family of ‘utility-supported moralities’, in which the goodness of an outcome still depends on utility but non-utility information is needed to distinguish among different kinds of utility and assign them distinct intrinsic values. Sen's move, although reflecting conventional wisdom, appears to be unnecessary and to that extent encourages an overly hasty dismissal of utilitarianism's potential normative appeal (Edwards, 1979; Riley, 1988; 2001; 2006a; 2007b; 2007c; 2009).

Normative economics, or that portion of economic theory that evaluates institutions such as markets and policies such as tax proposals with a view to offering prescriptions for society, is not necessarily linked

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
to any version of utilitarianism. Nevertheless, as an historical matter, economists generally endorsed hedonistic utilitarianism from the latter part of the 18th century, when both classical economics (or political economy) and classical hedonistic utilitarianism first began to emerge as systematic bodies of thought, well into the 20th century, long after the marginalist revolution of the 1870s when classical economics evolved into neoclassical economics. Even today, economists typically appeal to some version of utilitarianism, although utility is rarely defined in hedonistic terms. As Kenneth Arrow says: ‘The implicit ethical basis of economic policy judgement is some version of utilitarianism’ (1973, p. 246).

Jeremy Bentham (1789) and his followers, including, among others, James Mill, David Ricardo, and, for a time in his youth, John Stuart Mill, constituted the original school whose hedonistic version of utilitarianism remained dominant for so long within economics, although the doctrine was modified to some extent and combined with other ingredients largely as a result of Henry Sidgwick's influence on the early neoclassical economists such as W. Stanley Jevons, Alfred Marshall and, especially, Francis Y. Edgeworth. But the Benthamites were certainly not the first to preach a doctrine of general utility. Sidgwick traced the doctrine to Richard Cumberland's *De Legibus Naturae* (1672), for example, although Cumberland was not a hedonist. Bentham admitted that his hedonistic utilitarianism owed much to the writings of Claude Adrien Helvétius and David Hume.

Hume, like his contemporaries Francis Hutcheson and Adam Smith, asserted that human sentiments are approved of as prudent, virtuous or just only if they are seen as useful or agreeable either to the individual or to others. To be sure, none of these thinkers claimed that moral sentiments such as the sentiment of justice originate in the individual's idea of general utility. Despite various differences in their accounts, they seem to agree that an innate moral sense immediately feels a peculiar pleasure at the propriety of virtuous acts and dispositions towards other people, and that this pleasure can be made a powerful motive by suitably encouraging and educating the individual's natural sympathy for other human beings. But Bentham and Mill did not claim that moral sentiments originated in the individual's conception of general welfare either, although they rejected the notion of an innate moral sense. Bentham seems to have believed that people are predominantly motivated by self-love but that self-interest can be made to harmonize with the general welfare by giving the egoist incentives to comply with a utilitarian legal code. The sentiment of justice reduces to cooperating with others in terms of the code, which any rational egoist will do provided he is threatened with sufficient punishment by others for non-compliance. Mill did not rely exclusively on enlightened self-interest and external punishment but instead argued (rather as Hume did, especially in the *Enquiry Concerning the Principles of Morals*, 1751) that natural sympathy for fellow human beings can be raised to such a pitch through moral education that the individual comes to feel far more pleasure when he sacrifices his self-love to the extent required for conscientious mutual cooperation in terms of utilitarian laws and customs that distribute reciprocal rights and duties.

Smith, Hume and Hutcheson may not have been fully fledged hedonistic utilitarians like Bentham and Mill, who were prepared to assert that general welfare can demand the radical reform of rules commonly alleged to be dictates of an innate moral sense. But they may be fairly depicted as what Sidgwick called ‘conservative’ utilitarians, who sought to explain that common sense morality is generally compatible with public utility. In any case, Hume and Hutcheson appear to have supplied crucial ingredients for Mill's utilitarianism, the one through his account of the moral sentiment of justice, the other through his distinction among different kinds of pleasure, some of which are more intrinsically valuable than others.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Benthamite utilitarianism

Bentham's assumption that individuals are primarily self-interested fits well with economic theory. He took for granted that pleasure, including freedom from pain, is a single kind of agreeable feeling whose properties are invariant across different sentient beings. Any competent person can estimate the amount of pleasure which he or she experiences in any situation, he argued, by considering factors such as intensity, duration, certainty, propinquity, fecundity and purity. Yet he apparently did not intend to claim that much precision is possible. Indeed, as Mitchell (1918) documents, Bentham frankly admitted at times that intensity of pleasure cannot be measured and that interpersonal comparisons of pleasure cannot be based on the facts. Nevertheless, he insisted that alternative institutions and policies must be evaluated in terms of their consequences for each and every sentient being's pleasure and pain. Thus, rough estimates of aggregate net pleasure must be constructed before moral and political reasoning can even get started.

To obtain such estimates, Bentham seems to have assumed that everyone shares certain vital concerns, including security of expectations, subsistence, abundance and equality, and that different persons ought to be counted as if they are the same person when society chooses how to distribute the means of attaining these vital ingredients of anyone's happiness. The upshot is that the maximization of aggregate net pleasure becomes inseparably associated with a legal code that distributes equal rights and correlative duties for the purpose of providing the greatest total amount of security (in effect, equal marginal security) for every individual's vital concerns (Rosen, 1983; Kelly, 1990). These legal rights must include, among others, a right to subsistence and, consistently with that, rights to keep or trade the fruits of one's own labour and saving so as to foster abundance. Bentham's implicit premise, apparently, is that such legal rights are any rational egoist's source of his greatest amount of net pleasure, especially when duration and fecundity are taken into account, and that, to ensure universal compliance with the law, the egoist will endorse external punishment for violations of the correlative duties. Other animals can also be afforded some rights to protect their vital interests to some degree, although the power to exercise such claims must rest with humans.

At the same time, Bentham argued that the laws of property should be designed to promote an egalitarian distribution of income and wealth so far as is consistent with maximizing general security under an optimal code of equal rights. He gave priority to security, including a guarantee of subsistence as well as rights to own the means of production, over perfect equality of income and wealth as a source of pleasure for any rational individual. But he also endorsed the assumption that any individual experiences declining increments of pleasure from additional units of money or other material assets after subsistence is assured.

According to Benthamite utilitarianism, rational hedonistic egoists must be given incentives to act as they would act if they were aiming to maximize the general good conceived in terms of security, subsistence, abundance and equality. Admittedly, egoists will not face the threat of legal penalties with respect to private conduct that is left unregulated by the public authorities. Even so, as Sidgwick (1877; 1886) concluded, Bentham seems to have maintained that enlightened self-interest is always in harmony with virtue, not only in an ideal world but also in the world of actual experience, so that vicious conduct always involves a miscalculation. If Sidgwick's reading is correct, then Bentham prescribed external

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sanctions to discourage miscalculations of genuine self-interest, apparently on the assumption that such mistakes are likely to be observed to any considerable degree only when the egoist holds power over others, inadequate checks exist against the abuse of power, and the unchecked power corrupts the egoist's judgement.

J.S. Mill's qualitative utilitarianism

Although he said that he found much of permanent value in Bentham's doctrine, Mill made clear that he wished to ‘enlarge’ Benthamite utilitarianism by making room for higher moral and aesthetic sentiments which Bentham had largely ignored as a result of his focus on self-interest. Unlike Bentham, Mill argued that human nature is highly plastic and that many people in civil societies are observed to have developed characters in which sympathy for others and a conscientious desire to do right are powerful motives that would restrain self-interest even in the absence of any threat of external punishment. He agreed that pleasure is the only thing of intrinsic value but maintained that there are plural kinds such that higher kinds are intrinsically more valuable than lower kinds, with the implication that competent people who experience both will not give up even a bit of the higher for any amount of the lower ‘which their nature is capable of•’ (1861a, p. 211).

Mill apparently classified among the highest kinds of pleasurable feeling the complex moral sentiment of justice as he understood it. As he explains in the final chapter of *Utilitarianism* (1861a), this complex feeling of security (or what others might call the feeling of freedom) can be fully experienced only by cooperating with others in terms of an optimal code that distributes equal rights and correlative duties to all individuals. In his view, this higher kind of pleasure can become such a powerful motive that the just individual rarely if ever even considers pursuing his self-interest to the point of violating others’ rights. If so, Mill's pluralistic utilitarian doctrine is able to provide a more stable foundation than Bentham's purely quantitative doctrine can provide for a liberal system of weighty rights and duties. At the same time, individuals can still be permitted to freely pursue their selfish concerns in competitive markets, provided they comply with the code of justice.

Despite his qualitative gloss, which was probably suggested to him by Hutcheson's discussion of different kinds of pleasure in *A System of Moral Philosophy* (1755), Mill's doctrine remains rather similar to Bentham's, at least in broad outline. To be sure, there are important differences. In *On Liberty* (1859), Mill emphasized the importance of individuality as an ingredient of happiness, for instance, and he argued that a right to complete liberty of ‘purely self-regarding’ conduct is essential to promote individuality. He also went beyond Bentham by taking seriously the possibility of a decentralized socialism. Yet they agree on the need to maximize security by giving suitable priority to an optimal code of equal rights, and they also agree that institutions and policies should promote an egalitarian distribution of income and wealth so far as is consistent with the rights distributed by the code.

Moreover, like Bentham, Mill seems to have despaired of the possibility of ever acquiring data sufficiently precise to permit factual calculations of aggregate net pleasure. Rather, Mill argued that the test of quantity as well as quality of any two pleasures or pains was the unanimous judgement of competent persons who had experienced both, or, in case of disagreement, the majority judgement of such persons, where judgements are apparently assumed to be nothing more than preference orderings defined over the relevant net pleasures. By implication, for any pair of possible outcomes x and y, any

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

competent individual i who estimates that the quantity of net pleasure $u_i(x)$ which he will experience at x is at least as great as the quantity of the same kind of net pleasure $u_i(y)$ which he can expect at y, forms a weak judgement or preference R_i defined over pleasures which then, by virtue of hedonism, determines his preference over outcomes:

$$\text{for all } x, y: u_i(x)\, R_i\, u_i(y) \rightarrow x\, R_i\, y. \tag{2}$$

where R_i includes an asymmetric factor P_i denoting that i forms a strict preference if he estimates that $u_i(x) > u_i(y)$, and a symmetric factor I_i denoting that he is indifferent when he estimates that $u_i(x) = u_i(y)$.

The information about quantities of pleasure is no more precise than that contained in the fallible individuals' estimates, even if any competent person makes use of Bentham's guidelines to consider intensity, duration and so forth, and also accepts such common psychological generalizations as declining marginal pleasure of income beyond some threshold of subsistence. Given the Benthamite dictum that 'everybody is to count for one', aggregating over the individual judgements largely boils down to majority rule, which is consistent with the strongly majoritarian forms of representative democracy defended by Bentham in his *Constitutional Code* (1830) and by James Mill in his *Essay on Government* (1820). Yet majority rule does not rely on interpersonal comparisons of pleasure. Strictly speaking, each person's judgement might be counted as one by being represented by the same interpersonally comparable ordinal utility function as everyone else's in the modern economic sense of utility. In this case, even though any positive monotonic transformation of the single utility function used to represent each person's estimates of pleasure is permissible, adding the utility numbers to form hedonistic utilitarian judgements will usually (but not always) select majority winners if such outcomes exist and always resolve majority preference cycles when they occur (Riley, 2007c).

J.S. Mill also argued that any competent individual who judges that the kind of pleasure which he will experience at x is higher in quality than the kind of pleasure which he can expect at y, would never sacrifice even a bit of $u_i(x)$ for any amount of $u_i(y)$. In effect, the two kinds are incommensurable and cannot be traded off against one another in terms of the same scale of value. If the pleasure of the moral sentiment of justice is higher in quality than the enjoyable feelings of ordinary expediency, for instance, then a competent individual refuses to trade off even a little of his or anyone else's security of equal rights for, say, any quantity of enjoyment associated with ill-gotten income. The kind of painful insecurity associated with violating rights always outweighs the amount of merely expedient pleasure that might be gained by means of the violation. Because he feared that the popular majority might not be sufficiently educated to be competently acquainted with the pleasures of equal justice, Mill was less inclined than James Mill, his father, and Bentham were to support majoritarian democracy. Indeed, he seems to have been impressed by Thomas Macauley's objections that the Benthamites' methodology was flawed in so far as they attempted to deduce political conclusions on the assumption that individuals are rational self-interested agents without emotional ties to existing institutions (Lively and Rees, 1978).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
A more historical approach was needed, one that recognized the danger of majority tyranny and did not ignore as irrational those traditional institutions, attitudes and practices that might help to combat it. Thus, in *Considerations on Representative Government* (1861b), Mill argued for a limited form of representative democracy in which a distinctive system of checks and balances is designed to give more power to highly educated minorities, whose special expertise is generally acquired in traditional ways outside the democratic political system, to promote competent government and security of equal rights (Riley, 2007a).

Nevertheless, Mill's qualitative hedonism was generally dismissed as incoherent, and his model of liberal democracy was rejected as undemocratic. Under the influence of leading philosophers and economists, utilitarianism took a turn towards false quantitative precision, and the links with democracy were obscured.

Sidgwick and the early neoclassical economists

Sidgwick (1874), who billed himself as a hedonistic utilitarian, played a central role in all of this. He was in fact a rational intuitionist who argued that utilitarianism presupposed certain rational intuitions such as the axiom that an ethical agent must be impartial between one person's pleasures and another's. His argument muddied the hedonistic utilitarian tradition, which held with Hume that intrinsic value is not a function of a priori reason but rather of feelings of pleasure wherever located, whose origins and properties can in principle be inferred by reason only on the basis of experience. Moreover, although he claimed to be a strong admirer of J.S. Mill, Sidgwick asserted that a consistent hedonism must be purely quantitative because qualitative hedonism implicitly relies on some intrinsic value besides pleasure to distinguish among different qualities of pleasure. His assertion was apparently accepted as gospel by his friends Jevons (1874), Edgeworth (1877; 1881) and Marshall (1884). His student Moore later repeated it in the course of accusing Mill of spreading ‘contemptible nonsense’ (1903, p. 72). Yet Sidgwick and his followers merely beg the question against Mill because they assume that pleasure is a simple agreeable feeling that always exhibits the same properties. They never consider the possibility that pleasure is a term that covers a family of agreeable feelings, some of which are intrinsically more valuable than others.

Sidgwick also raised serious doubts about the rationality of a purely quantitative hedonistic utilitarianism by arguing that practical reason is ‘divided against itself’ because it cannot resolve basic conflicts between rational egoism and rational benevolence in some situations. This ‘dualism of practical reason’ is said to be manifested when a reasonable wish to preserve one's own life or other vital interests that ought to be protected by rights apparently collides with reasonable utilitarian duties to promote others’ welfare by sacrificing one's own life and enjoyments. Some such ethical dualism or pluralism is now a staple of the philosophical literature. Many argue that genuine maximizing utilitarianism conflicts with individual rights so fundamentally that an impartial rational resolution of the conflict is impossible. Moore went even further than Sidgwick and insisted that a ‘naturalistic fallacy’ is committed if intrinsic goodness is defined to be synonymous with pleasure or desire-satisfaction or any other natural property. Rather, goodness is said to be an indefinable non-natural quality that emanates in mysterious fashion from certain complex ‘organic wholes’ consisting of plural natural ingredients including, perhaps, pleasure as an essential ingredient, although Moore struggled to make up on his mind on this point (see,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for example, Edwards, 1979). But Sidgwick rejected Moore's anti-hedonistic view that ideals of goodness are directly intuited by anyone capable of appreciating the relevant ‘organic unities’. Indeed, without defining goodness to mean pleasure, Sidgwick accepted that pleasure including freedom from pain might ultimately be found to be the only thing of intrinsic value. He endorsed Bentham's ‘empirical method’ of ascertaining quantities of pleasure and pain, albeit reluctantly given the difficulties which he saw in applying that method. He also argued that quantitative hedonistic utilitarian reasoning accords with the bulk of common moral intuitions, upon which it can thus legitimately rely as rules for calculating right and wrong actions, and called for formal models of a utilitarian calculus under ideal conditions to help clarify its implications.

Early neoclassical economists, especially Edgeworth, answered the call and used the tools of mathematical calculus to formulate ideal versions of quantitative hedonistic utilitarianism. Unlike Bentham or Mill, Edgeworth assumed that natural units (‘just perceivable increments’) of pleasure and pain can in principle be ascertained and aggregated over varying populations and time horizons. Rather than consider individuals’ judgements of the quantities of net pleasure to be expected from feasible options as in (2) above, he imagined an ideal situation in which the quantities are definitely known such that, for any individual *i*, a unique natural utility function can be specified which, by virtue of hedonism, determines what *i*'s preferences over outcomes ought to be:

$$\text{for all } x, y: u_i(x) \geq u_i(y) \rightarrow x R_i y. \tag{3}$$

There is no longer any need for *i*'s estimates of pleasure because the utility numbers are assumed to be already known with fantastic precision. There is not even any need for *i* to form or express his preferences over outcomes. Person *i*'s natural utility function indicates the exact amount of interpersonally comparable natural pleasure which *i* can reasonably expect to experience at any given option, whether or not *i* appreciates this fact. Any competent ‘impersonal observer’ with the requisite individual utility information can then determine a best option by adding up the unique individual utility numbers at each option to see which option has the greatest sum total of pleasure net of pain. A utilitarian calculus is thus no longer necessarily linked to majoritarian aggregation procedures as it was with Bentham and Mill. Indeed, an authoritarian elite might perform and enforce the utilitarian calculations. Sidgwick and Edgeworth seem to have been surprisingly open to the possibility of some such utilitarian elite.

The marginalists also found more or less ingenious ways to overcome Sidgwick's ‘dualism of practical reason’. They imported the idea of an evolutionary process as they understood it, whereby ignorant and selfish individuals might eventually evolve into intelligent and virtuous ones through cultural and (as Herbert Spencer suggested) even biological transmission of the relevant concepts and dispositions. Jevons went so far as to endorse the Hegelian notion that this evolutionary solution was under divine direction. Edgeworth did not go so far. Yet, like Jevons, he took for granted that the more highly evolved members of a refined minority have greater capacities than the masses do for pleasure (the minority can enjoy ‘higher pleasures’ in the sense of larger quantities of pleasure viewed as a single kind

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of agreeable feeling), and he emphasized that utilitarianism does not necessarily imply equal distribution of the 'means of pleasure' or of the work required to produce the means.
Moreover, Edgeworth was specific about the way in which Sidgwick's 'dualism of practical reason' could be overcome in the economic arena. After calculating the 'contract curve' of Pareto-efficient allocations that self-interested traders could willingly negotiate and enforce as contracts, and showing that the contract curve converges on a perfectly competitive equilibrium only as the number of bargainers goes to infinity, he stressed the indeterminacy faced by any finite number of bargainers in selecting among efficient contracts. He then suggested that under certain conditions, including equal prior probabilities of any particular efficient contract being selected, the selfish bargainers would agree to accept a utilitarian contract as a just compromise because it is one of the efficient options on the contract curve, ignoring instances where the utilitarian bargain might make an individual worse off than his initial endowment (Creedy, 1986, pp. 79–92; Newman 2003, pp. xxxvii–xlvii). This contractualist argument for utilitarianism is interesting not only because it anticipates John Harsanyi (1977; 1992) but also because Edgeworth was well aware that a utilitarian bargain is typically distinct from a competitive equilibrium and thus calls for redistributive measures.

The ordinalist revolution

By the turn of the 20th century, hedonism was under siege in both psychology and philosophy as evolutionary psychology and behaviourism came to the fore along with philosophical idealism, pragmatism and ethical pluralism. Hedonism was largely abandoned within economics during the ordinalist revolution of the 1930s and 1940s, whose leading figures included Lionel Robbins, John Hicks, R.G.D. Allen, Abram Bergson and Paul Samuelson. The ordinalists, who registered doubts about meaningful interpersonal comparability as well as cardinal measurability of utility, recognized that the analysis of efficient allocations does not require either a hedonistic theory of motivation or rich interpersonally comparable cardinal utility information. They redefined the concept of utility to denote not pleasure but rather a formal numerical representation of any preference ordering revealed by consistent choice behaviour, without reference to the motivations or reasons underlying the revealed preference. In effect, utility merely represents what the agent is disposed to choose, independently of any psychological explanation or ethical justification for the given dispositions. Thus, in contrast to (2) or (3) above, any individual *i*'s utility function is independent of hedonism and simply recapitulates the information which is contained in *i*'s revealed preference ordering over outcomes:

$$\text{for all } x, y: u_i(x) \geq u_i(y) \leftrightarrow x R_i y. \tag{4}$$

Moreover, any positive monotonic transformation of the utility function is also an admissible utility function because such transformations preserve the information in the preference ordering which is being represented.
Given the restriction to such impoverished individual utility information, purely ordinal and bereft of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
any ethical standards, it is hardly surprising that there was an impulse to push aside issues of distributive justice, relegating them to other disciplines such as political philosophy. Normative economics could then concentrate on relatively uncontroversial Pareto efficiency considerations, including clarification of the conditions under which a general competitive equilibrium is efficient, for instance, and of the different conditions under which an efficient allocation can be achieved as a competitive equilibrium (cf. the ‘fundamental theorems’ of welfare economics).

Arrow's (1951) ‘general impossibility theorem’ may initially have reinforced the impulse to sidestep distributive issues because the theorem shows in effect that a social welfare function must reflect the preferences of a dictator if it is required to rely exclusively on purely ordinal utility information to generate rational social or moral choices from any set of distinct individual preferences. Yet normative economics cannot plausibly ignore distributive justice and the consequent need for interpersonal comparisons. In this regard, Arrow's negative result has also stimulated a large literature in which social choice theorists and game theorists have clarified many different forms of social decision functions and games, including their various informational requirements, and thereby clarified alternative theories of justice and morality which might be employed within normative economics. In addition to Arrow himself, Amartya Sen (1970; 1982; 2002; 2007) has played a leading role in this literature. But numerous others, including John C. Harsanyi (1955; 1977; 1992) and Kenneth Binmore (1994; 1998; 2005), have made noteworthy contributions.

It remains unclear what impact this ongoing literature will ultimately have on normative economics. Perhaps, as has been suggested (Mongin, 2006; see also Mongin and d’Aspremont, 1998), another revolution is under way, what might be called a non-utility revolution in so far as the thrust of it is to argue that a social welfare procedure ought to rely at least in part on information about the outcomes which is not reflected in individual preferences. Yet economists continue to defend and employ versions of utilitarianism. Indeed, Ng (1975; 1985) has made a case for redeploying within normative economics a so-called Benthamite doctrine that brings back Edgeworth's notion of ‘just perceivable units’ of pleasurable feeling. But economists have tended to adopt rational choice versions of utilitarianism which, unlike the forms of utilitarianism typically discussed by modern philosophers such as Hare (1981), Ng and Singer (1981), Gibbard (1987; 1990) and Brandt (1992), do not tie the idea of utility to pleasure, desire-satisfaction, norm-expression, or any other motivation. Harsanyi's doctrine is a leading example (for a cogent summary of its main features, see Hammond, 1987).

Harsanyi's rational choice utilitarianism

Harsanyi defines utility as merely a numerical indicator of any preference revealed by rational choice behaviour but he also argues that utility functions can be viewed as cardinal and interpersonally comparable rather than purely ordinal. He builds various conditions into his idea of what constitutes fully rational and moral behaviour so that meaningful interpersonal comparisons can be made of the gains or losses of utility that represent relative intensities of revealed preferences, that is, how much one outcome is preferred or opposed relative to another. Technically, this implies that a person's utility function can be subjected to any positive linear (but not nonlinear) transformations but, if any person's utility function is transformed, then every other person's must also be transformed in the same way. Such interpersonally comparable cardinal utility information is needed to operate a rational choice utilitarian

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
calculus in anything like the manner imagined by neoclassical economists such as Edgeworth.
Like John Rawls (1971; 1993), Harsanyi assumes that rational moral agents will imagine themselves in an original position under a veil of uncertainty about their particular social circumstances in order to calculate a fair social contract, that is, a conception of justice in terms of which they mutually agree to cooperate despite their different personal preferences. Unlike Rawls, however, Harsanyi assumes that rational choice behaviour under risk and uncertainty conforms to standard expected utility theory. He argues that individual attitudes towards risk should be used to infer personal preference intensities over outcomes, in which case the von Neumann–Morgenstern method of cardinalization becomes appropriate and a cardinal expected utility function represents an individual's revealed preferences under risk and uncertainty. Also, a moral agent must become an impersonal observer by forgetting his actual circumstances and imagining that he has an equal chance of occupying any person's social position with that person's preference intensities over outcomes. To avoid double counting, any impersonal observer must also ignore what Ronald Dworkin (1977, p. 234) calls ‘other-oriented’ preferences defined over other persons’ positions. All impersonal observers are guaranteed to make identical interpersonal comparisons and moral choices because human beings supposedly share a fundamentally similar psychology, which is left unspecified, though hedonism is rejected as naive. Finally, impersonal observers will jointly choose to constrain themselves, Harsanyi claims, by making a binding commitment to the same optimal code of moral rules. A person whose revealed preferences satisfy all these conditions is behaving as if he were a rule utilitarian, whatever his desires, feelings or other motivations might be.
Harsanyi's theory is vulnerable to various serious objections. It is not clear that revealed attitudes towards risk measure intensity of subjective feelings or that they would have much normative significance even if they did, for instance, even when individuals can be assumed to be concerned exclusively with outcomes and to ignore the process of assigning probabilities (‘gambling’) per se. Also, as Diamond (1967) has pointed out, Harsanyi's claim that moral and social utility must be a linear function of the individual utilities seems overly rigid because it ignores the distribution of the utilities. Rather than define morality so that it always demands the simple addition or averaging of individual utilities, a more appealing approach would insist that the process of moral and social decision-making should give everyone a ‘fair shake’. Fairness might be thought, for instance, to require a quasi-Rawlsian concern to maximize the utility of those individuals or groups with the worst utility levels in comparison to others. (For further discussion of this quasi-Rawlsian approach known in the literature as ‘leximin’, see, for example, Hammond 1976; 1977; and Deschamps and Gevers, 1978. Unlike Rawls's maximin theory, which works in terms of ‘primary goods’ that every rational person is presumed to want, the leximin theory works in terms of utility levels.)
Doubts about the von Neumann–Morgenstern method of cardinalization and the fairness of sum-ranking have prompted some leading economists, including Arrow, Diamond and James Mirrlees, to take seriously an ordinalist variant of rational choice utilitarianism.

Ordinalist utilitarianism

Ordinalist utilitarianism presupposes that interpersonally comparable ordinal utility information is available. In other words, it must be assumed that meaningful interpersonal comparisons can be made of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the levels of utility which represent revealed preference orderings. Technically, this implies that a person's utility function can be subjected to any positive monotonic transformations, but if any person's utility function is transformed, then every other person's must be similarly transformed.

An ordinalist utilitarian calculus is quite flexible in terms of its functional form. In contrast to (1) above, social welfare is calculated as the sum of utilities with a particular concave transformation *f* {.} applied to each person's utility function:

$$\text{for all } x\text{: } W(x) = \sum f\{u_i(x)\} = f\{\sum u_i(x)\}. \quad (5)$$

But *f* {.} may be subjected to any positive monotonic transformations without affecting *W* because the aggregation process is meaningful only for ordinal (rather than cardinal) comparable utility information. Arrow (1973) has shown that this ordinalist utilitarian approach subsumes the quasi-Rawlsian leximin theory of distributive justice as a special case, namely, the case in which the concavity of *f* (.) is extreme because each person is assumed to be highly risk-averse. Another important application is in the modern theory of optimal taxation (Mirrlees, 1971; 1982; Diamond and Mirrlees, 1974; Stiglitz, 1987). Indeed, optimal tax theory can be viewed as the further development and refinement of a body of thought that includes Edgeworth (1897) as well as Mill's and even Bentham's recommendation that a tax system ought to satisfy a principle of equal marginal sacrifice above some threshold of subsistence guaranteed for all.

Nevertheless, ordinalist utilitarianism also seems vulnerable to serious objections. Arrow remains reluctant to accept interpersonal comparisons of ordinal utility, for example, because he fears that they imply a denial of individuality: ‘the autonomy of individuals, an element of mutual incommensurability among people seems denied by the possibility of interpersonal comparisons’ (1977, p. 225). Individual rights to freely pursue one's own good in one's own way, of the sort defended by Mill, seem to be of peculiar moral importance, and should not be overridden by considerations of general utility based on putative interpersonal comparisons. Moreover, as Arrow also remarks, it is disappointing that, even if meaningful ordinalist comparisons are assumed possible, seemingly mild conditions (including a weak equity axiom) confine us to the quasi-Rawlsian ‘leximin’ theory of justice. But leximin is arguably too extreme because it absolutely forbids institutions and policies that fail to maximize the utility level of the worst-off, even if those measures result in massive utility gains for everyone else. Finally, as Gibbard (1987) suggests, ordinalist utilitarianism, like any other version of rational choice utilitarianism, needs an ‘empirically adequate’ psychology as well as a convincing theory of ethical deliberation, even if hedonism continues to be rejected as both a psychology and an ethical theory. Grounds for right action cannot simply be equated with what people are disposed to choose. Rather, utility, and thus welfarism and Paretianism, must be tied to a normative theory that identifies any individual's morally significant interests in any given social context and also justifies which choice dispositions ought to be formed to achieve the relevant interests in this or that situation. Such a normative theory must in turn be tied to a psychology that supplies empirically adequate psychic concepts and explains how dispositions to choose are formed in terms of the concepts.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In light of such objections to utilitarianism even in its ordinalist guise, some leading economists and philosophers see the need for a non-welfarist theory that makes use of non-utility information to evaluate possible outcomes. Although there are many different ways to go beyond utilitarianism and welfarism, two of the most interesting ways have been proposed by Binmore and Sen, respectively.

Beyond utilitarianism?

Binmore, inspired by his reading of Hume, presents a provocative proportional bargaining theory of ‘natural justice’, which, like Harsanyi's utilitarian theory, relies on cardinal comparable utility functions to represent the preferences revealed by moral choices. Unlike Harsanyi, though, Binmore assumes that moral agents remain predominantly selfish and are inclined to cheat on bargains unless they are threatened with sufficient punishment by others for non-compliance. He builds inequalities of bargaining power into his theory of moral behaviour by assuming that even bargainers who place themselves in an original position under a veil of ignorance will rely on cultural standards of interpersonal comparison which are associated with a given Nash bargaining equilibrium. His theory thus relies in part on non-utility information. The relevant Nash bargaining outcome supplies the cultural norms which agents in the original position use to calculate a proportional bargaining equilibrium, for instance, and it has a privileged status as a fallback position if these agents fail to agree to coordinate on the proportional bargaining solution. As a result, Binmore concludes that, in general, utilitarian bargains would not be willingly enforced by rational expected utility maximizers seeking justice.

Binmore apparently assumes that people are predominantly selfish because human behaviour is ultimately constrained in accord with the selfish gene paradigm. But there is no compelling scientific evidence for that paradigm. Rather, human nature appears to be highly plastic. If so, rational agents might eventually be moulded by cultural forces into social and moral actors who effectively believe that they are the same person – no different from anyone else – when it comes to certain vital personal interests that ought to be treated as rights. In this context, a utilitarian bargain, involving some code of justice that distributes equal rights and correlative duties whose content is held by the culture to be extremely valuable for every person's well-being, is an efficient and fair Nash equilibrium point. Even in the absence of any threat of punishment by others, compliance with the rules is enforced by the actor's own conscience, a powerful internal ‘judicious spectator’ which threatens to inflict harsh punishment in the form of intense feelings of guilt for cheating. Indeed, there is textual evidence that Hume himself took seriously this possibility of utilitarian justice (Riley, 2006b).

Sen also advocates the use of non-utility information in the course of proposing a pluralistic theory of ethical evaluation and distributive justice that involves protecting a list of basic human functionings and capabilities or ‘freedoms’ for each member of society (Sen, 1985; 2002; 2007). Moreover, his Paretian liberal impossibility result (Sen, 1970) shows that a social welfare function cannot in general simultaneously satisfy utility-based Paretian values and non-utility-based liberal rights as he conceives them. His work contains deep insights into the sort of moral and political philosophy which normative economics seems to require, and it also casts serious doubt on whether any version of welfarism can accommodate such insights.

Nevertheless, it is not entirely obvious that a move beyond utilitarianism is required, especially if the utilitarian family is defined (as it arguably should be) to include what Sen calls ‘utility-supported moralities’. A case can be made, for example, that genuine maximizing utilitarianism involves an

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
optimal code that distributes weighty equal rights and correlative duties, as many leading utilitarians, including Bentham and Mill as well as Brandt, Hare, and Harsanyi, have insisted (Riley, 2006a).

Moreover, it may even be possible to combine welfarism with a normative theory that interprets any individual's morally significant interests in terms of functionings and capabilities, the most vital of which ought to be protected by strong rights. A moral individual is presumably able to form dispositions to make the relevant moral choices, at least after engaging in a process of impartial deliberation about the functionings and capabilities of the different people implicated in any choice situation. Consistent ethical behaviour may reveal preference orderings that can be represented by utility functions. If so, an ethical version of welfarism could perhaps subsume Sen's ethical theory. The more general point is that, by itself, rational choice welfarism is essentially a formal shell that needs to be filled in with a substantive psychology and ethical theory. It can be filled in with various theories besides hedonism, to which it has no essential tie (Riley, 2001).

Normative economists evidently face numerous competing candidates when attempting to select a ‘reasonable’ social welfare procedure that is both efficient and fair. Yet ordinalist utilitarianism perhaps has more appeal than has so far appeared. Objections similar to Arrow's are often voiced against any version of utilitarianism (see, for example, Smart and Williams, 1977; Sen and Williams, 1982; Scheffler, 1982). Yet the objection that even ordinalist utilitarianism cannot accommodate reasonable principles of distributive justice arguably turns on an improper formulation of interpersonal comparisons of utility. As usually formulated in terms of extended sympathy, where one person places himself in another's position and imagines that he experiences the same psychic phenomena as the other experiences in that position, interpersonal comparisons involve the sort of double counting against which Harsanyi and Dworkin caution (see also Gibbard, 1987, p. 144). Reformulating interpersonal comparisons so that they do not involve double counting of any person's utility can liberate ordinalist utilitarianism from being chained to ‘leximin’. Arrow's other worry – that the possibility of interpersonal comparisons seems to deny individuality and personal integrity – might also be met by limiting the permissible scope of interpersonal comparisons so as to preserve weighty rights of individuality or self-development. Such an ethical limitation may be endorsed by utilitarians if a code that distributes such rights and correlative duties is deemed essential to the maximization of general utility.

Gibbard rightly objects that ordinalist utilitarianism by itself can hardly provide an acceptable ethical theory if utility merely represents given dispositions to choose, independently of any ethical justification for the dispositions. He advises that ethical thinking must rely on ‘rough quantitative’ estimates of which outcomes are ‘more worth wanting’ than others: ‘we should settle for some vagueness and indecision, epistemological and normative’ (1987, p. 148). This is wise advice, although it may be possible to combine ordinalist utilitarianism with an appealing ethical theory such that ‘some vagueness and indecision’ surrounds the ‘psychic magnitudes’ used in ethical thinking whereas utility functions represent revealed preferences that reflect the relevant ethical thinking. Indeed, a form of qualitative ordinalist utilitarianism along Millian lines deserves further study (Edwards, 1979; Riley, 1988; 2007b; 2007c; 2009). Such a qualitative ordinalist utilitarianism might incorporate what is taken to be Mill's theory of ethical thinking, including his suggested order of priority among intrinsically different kinds of ethical judgements, without making any commitment to psychological or ethical hedonism. Eventually, though, the time may come when some version of hedonism is again taken seriously (see, for example, Kahneman, Diener and Schwarz, 1999; Feldman, 2004).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- Arrow, Kenneth Joseph
- Bentham, Jeremy
- Edgeworth, Francis Ysidro
- happiness, economics of
- Harsanyi, John C.
- interpersonal utility comparisons
- Mill, John Stuart
- Mirrlees, James
- optimal taxation
- Sen, Amartya
- Sidgwick, Henry
- utility
- welfare economics

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*, 2nd rev. edn., New Haven: Yale University Press, 1963.

Arrow, K.J. 1973. Some ordinalist-utilitarian notes on Rawls's theory of justice. *Journal of Philosophy* 70, 245–63.

Arrow, K.J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review* 67, 219–25.

d'Aspremont, C. and Gevers, L. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 46, 199–210.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*, ed. J.H. Burns and H.L. A. Hart. London: Athlone Press, 1970.

Bentham, J. 1830. *Constitutional Code*, vol. 1, ed. F. Rosen and J.H. Burns. Oxford: Oxford University Press, 1983.

Binmore, K. 1994. *Playing Fair*. Cambridge, MA: MIT Press.

Binmore, K. 1998. *Just Playing*. Cambridge, MA: MIT Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Binmore, K. 2005. *Natural Justice*. Oxford: Oxford University Press.

Blackorby, C., Bossert, W. and Donaldson, D. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.

Brandt, R. 1992. *Morality, Utilitarianism, and Rights*. Cambridge: Cambridge University Press.

Creedy, J. 1986. *Edgeworth and the Development of Neoclassical Economics*. Oxford: Blackwell.

Cumberland, R. 1672. *De Legibus Naturae: A Treatise of the Laws of the Nature*, ed. J. Parkin. Indianapolis: Liberty Fund, 2005.

Deschamps, R. and Gevers, L. 1978. Leximin and utilitarian rules: a joint characterization. *Journal of Economic Theory* 17, 143–63.

Diamond, P.A. 1967. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: comment. *Journal of Political Economy* 75, 765–66.

Diamond, P.A. and Mirrlees, J.A. 1974. Optimal taxation and public production, I: production efficiency, and II. tax rules. *American Economic Review* 61, 8–27, 261–78.

Dworkin, R. 1977. *Taking Rights Seriously*. London: Duckworth.

Edgeworth, F.Y. 1877. *New and Old Methods of Ethics, or 'Physical Ethics' and 'Methods of Ethics'*. Oxford and London: James Parker. Repr. in Newman (2003).

Edgeworth, F.Y. 1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: Kegan Paul. Repr. in Newman (2003).

Edgeworth, F.Y. 1897. The pure theory of taxation. *Economic Journal* 7, 46–70, 226–38, 550–71.

Edwards, R.B. 1979. *Pleasures and Pains: A Theory of Qualitative Hedonism*. Ithaca and London: Cornell University Press.

Feiwel, G.R., ed. 1987. *Arrow and the Foundations of the Theory of Economic Policy*. London: Macmillan.

Feldman, F. 2004. *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford: Clarendon Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gibbard, A.F. 1987. Ordinal utilitarianism. In Feiwel (1987).

Gibbard, A.F. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Hammond, P.J. 1976. Equity, Arrow's conditions and Rawls' difference principle. *Econometrica* 44, 793–804.

Hammond, P.J. 1977. Dual interpersonal comparisons of utility and the welfare economics of income distribution. *Journal of Public Economics* 7, 51–71.

Hammond, P.J. 1987. On reconciling Arrow's theory of social choice with Harsanyi's fundamental utilitarianism. In Feiwel (1987).

Hare, R.M. 1981. *Moral Thinking: Its Method, Levels and Point*. Oxford: Clarendon Press.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Harsanyi, J.C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Harsanyi, J.C. 1992. Game and decision theoretic models in ethics. In *Handbook of Game Theory*, vol. 1, ed. R.J. Aumann and S. Hart. Amsterdam: Elsevier.

Hume, D. 1751. *An Enquiry Concerning the Principles of Morals*, ed. T.L. Beauchamp. Oxford: Oxford University Press, 1998.

Hutcheson, F. 1755. *A System of Moral Philosophy*, ed. D. Carey. London and New York: Continuum, 2000.

Jevons, W.S. 1874. Utilitarianism. In *Pure Logic and Other Minor Works of W.S. Jevons*, ed. R. Adamson and H.A. Jevons. London: Macmillan, 1890.

Kahneman, D., Diener, E. and Schwarz, N. 1999. *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation.

Kelly, P.J. 1990. *Utilitarianism and Distributive Justice*. Oxford: Clarendon Press.

Lively, J. and Rees, J. 1978. *Utilitarian Logic and Politics*. Oxford: Clarendon Press.

Marshall, A. 1884. On utilitarianism: a summum bonum. In *The Early Economic Writings of Alfred*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Marshall, 1867–1890, vol. 2, ed. J.K. Whitaker. New York: Free Press, 1975.

Maskin, E.S. 1978. A theorem on utilitarianism. *Review of Economic Studies* 45, 93–6.

Mill, J. 1820. An essay on government. In *James Mill: Political Writings*, ed. T. Ball. Cambridge: Cambridge University Press, 1992.

Mill, J.S. 1859. On liberty. In *Collected Works*, vol. 18, ed. J.M. Robson. Toronto: University of Toronto Press; London: Routledge, 1977.

Mill, J.S. 1861a. Utilitarianism. In *Collected Works*, vol. 10, ed. J.M. Robson. Toronto: University of Toronto Press; London: Routledge, 1969.

Mill, J.S. 1861b. Considerations on representative government. In *Collected Works*, vol. 19, ed. J.M. Robson. Toronto: University of Toronto Press; London: Routledge, 1977.

Mirrlees, J.A. 1971. An exploration in the theory of optimal income taxation. *Review of Economic Studies* 38, 175–208.

Mirrlees, J.A. 1982. The economic uses of utilitarianism. In Sen and Williams (1982).

Mitchell, W.C. 1918. Bentham's felicific calculus. *Political Science Quarterly* 33, 161–83.

Mongin, P. 2006. A concept of progress for normative economics. *Economics and Philosophy* 22, 19–54.

Mongin, P. and d'Aspremont, C. 1998. Utility theory and ethics. In *Handbook of Utility Theory*, vol. 1, ed. S. Barbera, P. Hammond and C. Seidl. Dordrecht: Kluwer.

Moore, G.E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press, 1959.

Moore, G.E. 1912. *Ethics*. Oxford: Oxford University Press, 1958.

Newman, P. 2003. *F.Y. Edgeworth: Mathematical Psychics and Further Papers on Political Economy*. Oxford: Oxford University Press.

Ng, Y.-K. 1975. Bentham or Bergson? Finite sensibility, utility functions, and social welfare functions. *Review of Economic Studies* 42, 545–70.

Ng, Y.-K. 1985. Some fundamental issues in social welfare. In *Issues in Contemporary Microeconomics and Welfare*, ed. G.R. Feiwell. London: Macmillan.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ng, Y.-K. and Singer, P. 1981. An argument for utilitarianism. *Canadian Journal of Philosophy* 11, 229–39.

Pigou, A.C. 1920. *The Economics of Welfare*, 4th rev. edn. London: Macmillan, 1932.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 1993. *Political Liberalism*, 2nd edn. New York: Columbia University Press, 2002.

Riley, J. 1988. *Liberal Utilitarianism*. Cambridge: Cambridge University Press.

Riley, J. 2001. Welfare (philosophical aspects). In *International Encyclopedia of the Social and Behavioral Sciences*, vol. 25, ed. N.J. Smelser and P.B. Bates. London: Pergamon.

Riley, J. 2006a. Liberal rights in a Pareto-optimal code. *Utilitas* 18, 61–79.

Riley, J. 2006b. Genes, memes and justice. *Analyse & Kritik* 28, 32–56.

Riley, J. 2007a. Mill's neo-Athenian model of liberal democracy. In *J.S. Mill's Political Thought: A Bicentennial Reassessment*, ed. N. Urbinati and A. Zakaras. New York: Cambridge University Press.

Riley, J. 2007b. *Mill's Radical Liberalism*. London: Routledge.

Riley, J. 2007c. Classical ordinalist utilitarianism. Unpublished working paper.

Riley, J. 2009. The interpretation of maximizing utilitarianism, and distributive justice. *Social Philosophy and Policy* 26(Winter).

Rosen, F. 1983. *Jeremy Bentham and Representative Democracy: A Study of the Constitutional Code*. Oxford: Clarendon Press.

Scheffler, S. 1982. *The Rejection of Consequentialism*, Rev. edn. Oxford: Oxford University Press. 1994.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Sen, A.K. 1979. Utilitarianism and welfarism. *Journal of Philosophy* 76, 463–89.

Sen, A.K. 1980. Plural utility. *Proceedings of the Aristotelian Society* 81, 193–215.

Sen, A.K 1982. *Choice, Welfare and Measurement*. Oxford: Blackwell. Repr. Harvard University Press, 1997.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sen, A.K. 1985. *Commodities and Capabilities*. Amsterdam: North-Holland. Repr. Oxford University Press, 1999.

Sen, A.K. 2002. *Rationality and Freedom*. Cambridge, MA: Harvard University Press.

Sen, A.K. 2007. *Freedom and Justice*. Cambridge, MA: Harvard University Press.

Sen, A.K. and Williams, B.A.O., eds. 1982. *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Sidgwick, H. 1874. *The Methods of Ethics*, 7th edn. London: Macmillan, 1907.

Sidgwick, H. 1877. Bentham and Benthamism in politics and ethics. In *Essays on Ethics and Method: Henry Sidgwick*, ed. M.G. Singer. Oxford: Clarendon Press, 2000.

Sidgwick, H. 1886. *Outlines of the History of Ethics*, 5th edn. London: Macmillan, 1902.

Smart, J.J.C. and Williams, B.A.O. 1977. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Stiglitz, J.E. 1987. Pareto efficient and optimal taxation and the new welfare economics. In *Handbook of Public Economics*, vol. 2, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

How to cite this article

Riley, Jonathan. "utilitarianism and economic theory." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000073> doi:10.1057/9780230226203.1780

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

utility

R.D. Collison Black
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

cardinal utility; classical theory of value; consumer surplus; cost of production; demand; diminishing marginal utility; Dupuit, A.-J.-E.J.; Edgeworth, F. Y.; exchange; exchange value; Gossen, H. H.; Hicks, J. R.; interpersonal utility comparisons; Jevons, W. S.; marginal revolution; marginal utility; market price; Marshall, A.; Menger, C.; natural price; ordinal utility; Pareto, V.; Pigou, A. C.; price theory; revealed preference theory; use value; utility; value; Walras, L.

Article

Utility is a term which has a long history in connection with the attempts of philosophers and political economists to explain the phenomenon of value. It has most frequently been given the connotation of ‘desiredness’, or the capacity of a good or service to satisfy a want, of whatever kind. Its use with that meaning can be traced back at least to Gershom Carmichael's 1724 edition of Pufendorf's *De Officio Hominis et Civis Iuxta Legam Naturalem*, and arguably came down to him through the medieval schoolmen from Aristotle's *Politics*.

Utility in the sense of desiredness is a purely subjective concept, clearly distinct from usefulness or fitness for a purpose – the more normal everyday sense of the word and the first meaning given for it by the *Oxford English Dictionary*.

While most political economists of the 18th and 19th centuries used the term in this subjective sense, the distinction was not always kept clear, most notably in the writings of Adam Smith. In a famous passage in the *Wealth of Nations* Smith wrote:

> The word VALUE, it is to be observed, has two different meanings, and sometimes expresses the utility of some particular object, and sometimes the power of purchasing other goods which the possession of that object conveys. The one may be called ‘value in use’; the other, ‘value in exchange’. The things which have the greatest value in use have frequently little or no value in exchange; and, on the contrary, those which have the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> greatest value in exchange have frequently little or no value in use. Nothing is more useful than water; but it will purchase scarce any thing; scarce any thing can be had in exchange for it. A diamond, on the contrary, has scarce any value in use; but a very great quantity of other goods may frequently be had in exchange for it. (1776, Book I, ch. IV)

Smith has sometimes been accused, because of the wording of this passage, of falling into the error of claiming that things which have *no* value in use can have value in exchange, which is tantamount to saying that utility is not a necessary condition for a good to have value. It would appear, however, that Smith was not here using the theme ‘value in use’, or utility, in the subjective sense of desiredness but in the normal objective sense of usefulness (cf. Bowley, 1973, p. 137; O’Brien, 1975, pp. 80 and 97). Most other classical economists and even Smith himself in his *Lectures on Jurisprudence* used the term in its subjective sense, but the passage in the *Wealth of Nations* gave rise to considerable confusion and misinterpretation. Nor was this the only source of confusion in the early writing on the subject: even those who used the term utility in its subjective sense were not always clear as to whether it should be considered a feeling in the mind of the user or a property of the good or service used. Thomas De Quincey, for example, referred to the ‘intrinsic utility’ of commodities (*Logic of Political Economy*, 1844, p. 14).

Most classical economists, however, were not greatly concerned with the subtleties of meaning which the term utility might contain. Generally they used it in the broad sense of desiredness, and Ricardo employed it in a typically classical way when he wrote

> Utility then is not the measure of exchangeable value, although it is absolutely essential to it. If a commodity were in no way useful – in other words, if it could in no way contribute to our gratification – it would be destitute of exchangeable value, however scarce it might be, or whatever quantity of labour might be necessary to procure it. (*Principles of Political Economy and Taxation*, 1817, ch. 1, sect. I)

‘Useful’ here is interpreted as ‘contributing to gratification’ but the very word carries an echo of Smith's confusion.

For Ricardo and others in the mainstream classical tradition down to J.S. Mill and Cairnes utility became a necessary but not a sufficient condition for a good to possess value. In this context, the utility referred to was generally the total utility of the good to the purchaser, or the utility of a specific quantity which is all that is available in the circumstances of the example – for example, the utility of a single item of food to a starving person.

As a result of this approach it followed, in the words of J.S. Mill, that ‘the utility of a thing in the estimation of the purchaser is the extreme limit of its exchange value: higher the value cannot ascend; peculiar circumstances are required to raise it so high’ (*Principles of Political Economy*, 1848, Book III, ch. II, §. 1). Classical economists like Mill accepted the view put forward by J.B. Say that ‘labour is not creative of objects, but of utilities’ but could see the weakness in Say's contention that price measured utility. Clearly in the case of competitively produced commodities it did not and it was cost of production and not utility (in the total sense) which determined value.

Since the classical economists were mainly interested in ‘natural’ rather than ‘market’ price, that is, in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
long-run normal values which were mainly determined by supply and cost, the fact that they had no theory to explain fully the relationships between utility, demand and market price was not a matter of concern to most of them. Nevertheless in the period from about 1830 to 1870 a number of attempts were made to work out these relationships more fully or to clarify aspects of them. Some of these attempts took place in Britain, within the framework of the classical system, but not surprisingly some of the best work was done at this time in France, where the tradition of demand analysis was stronger.

The full explanation of the relation between utility and demand requires the distinction between total utility and increments of utility, and the recognition of the principle that consumption of successive increments of a commodity yields not equal but diminishing increments of satisfaction or utility to the consumer. A number of writers in the mid-19th century showed an understanding of this point, but only a few stated it explicitly and correctly. Among those in Britain who did so were William Foster Lloyd (*A Lecture on the Notion of Value, delivered before the University of Oxford in Michaelmas Term*, 1833) and Nassau Senior (*Outline of the Science of Political Economy*, 1836), but neither proceeded to develop his insights into a complete theory of the relationship between utility, demand and market values.

The French engineer A.J. Dupuit was the first to present an analysis which clearly explained the concept of marginal utility and related it to a demand curve, in his paper ‘On the Measurement of the Utility of Public Works’ (*Annales des Ponts et Chaussées*, vol. 8, 1844; English translation in *International Economic Papers*, No. 2, 1952, pp. 83–110). Dupuit also extended his analysis to show that the total area under the demand curve represents the total utility derived from the commodity; deducting from this the total receipts of the producer he arrived at the ‘utility remaining to consumers’ or what was later to be termed ‘consumers’ surplus’.

The significance of Dupuit's contribution is now well recognized, but at the time of appearance it had little impact. The same is even more true of the work of Hermann Heinrich Gossen, one of the few German contributors to utility theory in this period. His book *Entwicklung der Gesetze des Menschlichen Verkehrs*, published in 1854, contained not only a statement of the ‘law of satiable wants’, or diminishing marginal utility, but also of the proposition that to maximize satisfaction from any good capable of satisfying various wants it must be allocated between those uses so as to equalize its marginal utilities in all of them.

Gossen's analysis of the principles of utility maximization was thus more complete than any which had preceded it. Yet his one book, which foreshadowed many features of general equilibrium as well as utility theory, received virtually no attention until 1878, 20 years after the author's death, when Robert Adamson, W.S. Jevons's successor as Professor of Philosophy and Political Economy at Manchester, obtained a copy of it and drew it to the attention of Jevons himself.

By that time the whole character of utility analysis and its place in economic theory had begun to change significantly. This change is usually dated from the very nearly simultaneous publication of Jevons's *Theory of Political Economy* in England and Menger's *Grundsätze der Volkswirtschaftslehre* Austria, both in 1871, and Walras's *Eléments d’économie politique pure* in Switzerland in 1874. All these works contained a treatment of the theory of value in which the analysis of diminishing marginal utility (under a variety of other names) played a considerable part, but each of the three authors seems to have arrived independently at the main ideas of his theory without indebtedness to the others or to the predecessors already mentioned above.

This remarkable example of multiple discovery in the history of ideas has come to be known as ‘the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Marginal Revolution'. Discussion of its causes and character lies outside the scope of this article, but it is generally accepted that, as Sir John Hicks has said, 'the essential novelty in the work of these economists was that instead of basing their economics on production and distribution, they based it on exchange' (Hicks, 1976, p. 212).

A major element in this 'shift in attention' undoubtedly was a change from the classical concept of value in use, or total utility, as a necessary but not sufficient condition to explain the normal values of freely reproducible commodities, to the concept of what Jevons called 'the degree of utility' and of adjustments in it, through exchange of quantities of goods held or consumed, in order to maximize satisfaction. Marginal analysis can, however, be applied to questions of production and distribution as well as consumption, and hence the 'Marginal Revolution' involved more than a new stage in the development of utility theory.

Although all three pioneers of the Marginal Revolution did contribute to that development they also contributed in other ways to the theory of pricing and exchange. Perhaps only for Jevons was the theory of utility genuinely central to the structure of his economic work. On the opening page of his *Theory of Political Economy* he emphatically asserted that '*value depends entirely upon utility*' and he went on to say that 'Political Economy must be founded upon a full and accurate investigation of the conditions of utility' (1871, p. 46). Jevons indeed appears to have shared with his classical predecessors the view that a theory of value must go beyond the phenomena of demand and supply to some more fundamental explanation which for him was to be found in utility rather than in labour. 'Labour is found often to determine value, but only in an indirect manner, by varying the degree of utility of the commodity through an increase or limitation of supply' (Jevons, 1871, p. 2).

Apart from differences of terminology, Walras's treatment of utility in relation to the problem of exchange had substantial similarities with that of Jevons; but Walras saw the problem in a different context.

> His whole attention was focused on market phenomena and not on consumption … while the driving force in the theory of exchange is, as Walras saw it, the endeavour of all traders to maximise their several satisfactions, it is marketplace satisfactions rather than dining-room satisfactions which Walras had in mind. (Jaffé, 1973, pp. 118–19)

For Menger, as for Walras, the concepts of utility theory formed only a part of a much large analytical structure (concerned in his case not so much with equilibrium as with development), but unlike both Walras and Jevons he refused to state his theories in mathematical terms.

Menger developed a theory of economizing behaviour showing how the individual would seek to satisfy his subjectively felt needs in the most efficient manner. In the process he elaborated the essential propositions of a theory of maximizing behaviour for the consumer, but he expressed them in terms of the satisfaction of needs by the consumption of successive units of goods. In his discussion of this process Menger used the same phrases – use-value and exchange-value – which Smith had used almost a century earlier, and with similar connotations. Use-value he defined as 'the importance that goods acquire for us because they *directly* assure us the satisfaction of needs that would not be provided for if we did not have the goods at our command. Exchange value is the importance that goods acquire for us because their possession assures the same result *indirectly*' (Menger, 1871, p. 228). Menger did use the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
term ‘utility’, but not as a synonym for use-value; he viewed it as an abstract relation between a species of goods and a human need, akin to the general term ‘usefulness’. As such it constituted a prerequisite for a good to have economic character, but had no quantitative relationship to value.

The three pioneers of the Marginal Revolution thus saw the problem of the relationship of utility to exchange value in different contexts and expressed their solutions to it in different ways. Inevitably also their first solutions were incomplete in various respects. For example, the precise relationships between the individual's utility function and demand function, the market demand function and the market price were not clearly specified in some of the earlier formulations; it remained for later contributors such as Marshall, Wicksteed and Edgeworth to deal with these points.

Nevertheless, despite their differences of terminology and approach, the writings of the pioneers did contain a common core which gradually gained wider acceptance, and by 1890 with the appearance of Marshall's *Principles of Economics* it seemed that the new analysis of market values had been effectively integrated with an analysis of supply and cost which served to explain long-run ‘normal’ values. It did provide some things which the classical system had not contained, among them a consistent theory of consumer behaviour, expressed in terms of utility.

So in 1899 it was possible for Edgeworth to write that:

> the relation of utility to value, which exercised the older economists, is thus simply explained by the mathematical school. The value in use of a certain quantity of commodity corresponds to its total utility; the value in exchange to its marginal utility (multiplied by the quantity). The former is greater than the latter, since the utility of the final increment of commodity is less than that of every other increment. (Edgeworth, 1899, p. 602)

At this stage utility analysis appeared to have evolved to something approaching finality, and in 1925 Jacob Viner could still say:

> In its developed form it is to be found sympathetically treated and playing a prominent role in the exposition of value theory in most of the current authoritative treatises on economic theory by American, English, Austrian, and Italian writers.

Yet Viner immediately went on to add:

> In the scientific periodicals, however, in contrast with the standard treatises, sympathetic expositions of the utility theory of value have become somewhat rare. In their stead are found an unintermittent series of slashing criticisms of the utility economics. (Viner, 1925, p. 179)

The principal criticisms which Viner noted were the apparent involvement of utility theory with hedonistic psychology and the problems of measuring welfare in terms of utility. In later years questions of the measurement and summation of utility came to trouble economists more and more.

The two basic problems involved here are whether utility can be measured cardinally or simply

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
ordinally, and whether interpersonal comparisons of utility are possible. The pioneers of the Marginal Revolution were not unaware of these problems; Jevons nowhere attempted to define a unit of utility, and indeed said that ‘a unit of pleasure or pain is difficult even to conceive’, but he went on to say that ‘it is from the quantitative effects of the feelings that we must estimate their comparative amounts’ (Jevons, 1871, p. 14). However they may be estimated, Jevons did not hesitate to refer to ‘quantity of utility’, and his whole analysis proceeds by treating utility as if it could be measured. The question was not examined in detail by Walras or Menger, but both their analyses treat utility as cardinally measurable.

On the question of interpersonal comparisons of utility, Menger and Walras seemed to find no difficulty, and Walras was prepared to speak of a ‘maximum of utility’ for society (Walras, 1874, p. 256). Jevons on the other hand declared that ‘every mind is … inscrutable to every other mind, and no common denominator of feeling seems possible’ (Jevons, 1871, p. 211) – but this did not always prevent him from comparing and aggregating utilities.

In the early editions of his *Principles of Economics* Marshall fully accepted the idea of utility as cardinally measurable and allowed the possibility if not of interpersonal certainly of inter-group comparisons of utility (1890, pp. 151 and 152). In later years he became more reticent and defensive on these points, and he was always more concerned than Jevons with the effects of feelings rather than the feelings themselves; yet cardinal utility always remained the basis of Marshall's demand theory.

Now, as Sir John Hicks said:

> if one starts from a theory of demand like that of Marshall and his contemporaries, it is exceedingly natural to look for a welfare application. If the general aim of the economic system is the satisfaction of consumer wants, and if the satisfaction of individual wants is to be conceived of as a maximising of utility, cannot the aim of the system be itself conceived of as a maximising of utility – universal utility, as Edgeworth called it? If this could be done and some measure of universal utility could be found, the economist's function could be widened out, from the understanding of cause and effect to the judgement of the effects – whether, from the point of view of want-satisfaction, they are to be judged as successful or unsuccessful, good or bad. (Hicks, 1956, p. 6)

This was, in effect, the task which was undertaken by Marshall's successor, A.C. Pigou, in his *Economics of Welfare* (1920; earlier version published under the title *Wealth and Welfare*, 1912). Pigou made no attempt to establish a measure of universal utility; instead he took what Marshall had called ‘the national dividend’, aggregate real income, as the ‘objective counterpart’ of economic welfare. Pigou argued that economic welfare would be greater when aggregate real income increased, when fluctuations in its amount were reduced, and when it was more equally distributed among persons. It was in the context of this last point that interpersonal utility comparisons were most evident; Pigou argued that

> the old ‘law of diminishing utility’ thus leads securely to the proposition: Any cause which increases the absolute share of real income in the hands of the poor, provided that it does not lead to a contraction in the size of the national dividend … will in general, increase economic welfare. (1920, p. 89)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In the 1930s most economists became increasingly uncomfortable with the idea of measurement and interpersonal or intergroup comparisons of utility. In 1934, in a famous article entitled ‘A Reconsideration of the Theory of Value’, Hicks and Allen used the technique of indifference curves originated by Edgeworth and developed by Walras's successor at Lausanne, Vilfredo Pareto, in presenting a theory of consumer behaviour involving only ordinal comparisons of satisfaction. A few years later a further step towards eliminating what were now considered dubious psychological assumptions from that theory was taken by treating consumer behaviour solely on the basis of revealed preference.

Accompanying these changes there was a movement away from the type of welfare economics developed on the basis of utility theory by Marshall and Pigou towards that based on Pareto's concept of an economic optimum as a position from which it is impossible to improve anyone's welfare without damaging that of another.

Indifference analysis and revealed preference theory are now standard features of microeconomic theory; but the utility concept has not disappeared; the most widely used introductory economics texts still tend to begin their treatments of household behaviour with an account of utility theory.

See Also

- Gossen, Hermann Heinrich

Bibliography

Bowley, M. 1973. Utility, the paradox of value and ‘all that’ and classical economics. In *Studies in the History of Economic Theory before 1870*, M. Bowley. London: Macmillan.

De Quincey, T. 1844. *The Logic of Political Economy*. Edinburgh: Blackwood.

Dupuit, A.J. 1844. On the measurement of the utility of public works. *Annales des Ponts et Chaussées*, 2nd Series, vol. 8. Trans. in *International Economic Papers* No. 2, London: Macmillan, 1952.

Edgeworth, F.Y. 1899. Utility. In *Dictionary of Political Economy*, vol. 3 , ed. R.H.I. Palgrave. London: Macmillan.

Gossen, H.H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs und der daraus Fliessenden Regeln für menschliches Handeln*. Brunswick: Viewig. Translated as *The Laws of Human Relations and the Rules of Human Action Derived Therefrom*, Cambridge, MA: MIT Press, 1983.

Hicks, J.R. 1956. *A Revision of Demand Theory*. Oxford: Clarendon Press.

Hicks, J.R. 1976. ‘Revolutions’ in economics. In *Method and Appraisal in Economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hicks, J.R. and Allen, R.G.D. 1934. A reconsideration of the theory of value. *Economica*, NS 1, 52–76, 196–219.

Howey, R.S. 1960. *The Rise of the Marginal Utility School, 1870–1889*. Lawrence: University of Kansas Press.

Jaffé, W. 1973. Léon Walras's role in the ‘marginal revolution’ of the 1870s. In *The Marginal Revolution in Economics*, ed. R.D. Collison Black, A.W. Coats and C.D. Goodwin. Durham, NC: Duke University Press.

Jevons, W.S. 1871. *The Theory of Political Economy*. London: Macmillan.

Kauder, E. 1965. *A History of Marginal Utility Theory*. Princeton: Princeton University Press.

Lloyd, W.F. 1833. A lecture on the notion of value, delivered before the University of Oxford in Michaelmas Term 1833. Reprinted in *Economic History*, supplement to the *Economic Journal*, No. 2 (1927), 168–83.

Marshall, A. 1890. *Principles of Economics*. 9th (variorum) edn, ed. C.W. Guillebaud, London: Macmillan, 1961.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Trans. J. Dingwall and B.F. Hoselitz as *Principles of Economics*, Glencoe, IL: Free Press, 1951.

Mill, J.S. 1848. *Principles of Political Economy*. Ed. W. J. Ashley, London: Longmans, 1909.

O’Brien, D.P. 1975. *The Classical Economists*. Oxford: Clarendon Press.

Pigou, A.C. 1912. *Wealth and Welfare*. Expanded and republished as *The Economics of Welfare*, London: Macmillan, 1920.

Ricardo, D. 1817. *The Principles of Political Economy and Taxation*. Vol. 1 of *The Works and Correspondence of David Ricardo*, ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

Senior, N.W. 1836. *An Outline of the Science of Political Economy*. London: W. Clowes. Reprinted, London: Allen & Unwin, 1938.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. R.H. Campbell and A.S. Skinner, Oxford: Clarendon Press, 1976.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Stigler, G.J. 1950. The development of utility theory. Pts. I and II. *Journal of Political Economy* 58, 307–27; 373–96. Reprinted in G.J. Stigler, *Essays in the History of Economics*, Chicago: University of Chicago Press, 1965.

Viner, J. 1925. The utility concept in value theory and its critics. *Journal of Political Economy* 33, 369–87 Reprinted in J. Viner, *The Long View and the Short*, Glencoe, IL: Free Press, 1958.

Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz et Cie. Trans. W. Jaffé as *Elements of Pure Economics*, London: Allen & Unwin, 1954.

How to cite this article

Black, R. D. Collison. "utility." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_U000047> doi:10.1057/9780230226203.1781

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

value elicitation

Glenn W. Harrison
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Economists are interested in eliciting values at the level of the individual because market values do not provide the information needed to measure consumer surplus, value new products, or value goods that have no market. Direct and indirect procedures have been developed to elicit values, and each has some strengths and weaknesses. The evidence points to several recommendations for best practice in the reliable elicitation of values, trading off transparency and rigour.

Keywords

auctions; bias correction; cheap talk; conjoint choice; contingent valuation; cost–benefit analysis; English auction; latent choice models; maximum likelihood; multiple price lists; revealed preference theory; second-price auction; value elicitation; Vickrey auction

Article

Why elicit values? The prices observed on a market reflect, on a good competitive day, the equilibrium of marginal valuations and costs. They do not quantitatively reflect the infra-marginal or extra-marginal values, other than in a severely censored sense. We know that infra-marginal values are weakly higher, and extra-marginal values are weakly lower, but beyond that one must rely on functional forms to extrapolate. For policy purposes this is generally insufficient to undertake cost–benefit calculations. When producers are contemplating a new product or innovation they have to make some judgement about the value that will be placed on it. New drugs, and the R&D underlying them, provide an important example. Unless one can heroically tie the new product to existing products in terms of shared characteristics, and somehow elicit values on those characteristics, there is no way to know what price the market will bear. Value elicitation experiments can help fill that void, complementing traditional marketing techniques (see Hoffman et al., 1993).

Many goods and services effectively have no market, either because they exhibit characteristics of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
public goods or it is impossible to credibly deliver them on an individual basis. These non-market goods have traditionally been valued using surveys, where people are asked to state a valuation 'contingent on a market existing for the good'. The problem is that these surveys are hypothetical in terms of the deliverability of the good and the economic consequences of the response, and this understandably generates controversy about their reliability (Harrison, 2006).

Procedures

Direct methods for value elicitation include auctions, auction-like procedures and 'multiple price lists'. Sealed-bid auctions require the individual to state a valuation for the product in a private manner, and then award the product following certain rules. For single-object auctions, the second-price (or Vickrey) auction awards the product to the highest bidder but sets the price equal to the highest rejected bid. It is easy to show, to students of economics at least, that the bidder has a dominant strategy to bid his true value: any bid higher or lower can only end up hurting the bidder in expectation. But these incentives are not obvious to inexperienced subjects. A real-time counterpart of the second-price auction is the English (or ascending bid) auction, in which an auctioneer starts the price out low and then bidders increase the price to become the winner of the product. Bidders seem to realize the dominant strategy property of the English auction more quickly than in comparable second-price sealed-bid auctions, no doubt due to the real-time feedback on the opportunity costs of deviations from that strategy (see Rutström, 1998; Harstad, 2000). Familiarity with the institution is also surely a factor in the superior performance of the English auction: first encounters with the second-price auction rules lead many non-economists to assume that there must be some 'trick'.

Related schemes collapse the logic of the second-price auction into an auction-like procedure due to Becker, DeGroot and Marschak (1964). The basic idea is to endow the subject with the product, and to ask for a 'selling price'. The subject is told that a 'buying price' will be picked at random, and that, if the buying price that is picked exceeds the stated selling price, the product will be sold at that price and the subject will receive that buying price. If the buying price equals or is lower than the selling price, the subject keeps the lottery and plays it out. Again, it is relatively transparent to *economists* that this auction procedure provides a formal incentive for the subject to truthfully reveal the certainty-equivalent of the lottery. One must ensure that the buyout range exceeds the highest price that the subject would reasonably state, but this is not normally a major problem. One must also ensure that the subject realizes that the choice of a buying price does not depend on the stated selling price; a surprising number of respondents appear not to understand this independence, even if they are told that a physical randomizing device is being used.

Multiple price lists present individuals with an ordered menu of prices at which they may choose to buy the product or not. In this manner the list resembles a menu, akin to the price comparison websites available online for many products. For any given price, the choice is a simple 'take it or leave it' posted offer, familiar from retail markets. The set of responses for the entire list is incentivized by picking one at random for implementation, so the subject can readily see that misrepresentation can only hurt for the usual revealed preference reasons. Refinements to the intervals of prices can be implemented, to improve the accuracy of the values elicited (see Andersen et al., 2006). These methods have been widely used to elicit risk preferences and discount rates, as well as values for products (see Holt and Laury,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
2002; Harrison, Lau and Williams, 2002; Andersen et al., 2007).

Indirect methods work by presenting individuals with simple choices and using a latent structural model to infer valuations. The canonical example comes from the theory of revealed preference, and confronts the decision-maker with a series of purchase opportunities from a budget line and asks him to pick one. By varying the budget lines one can 'trap' latent indifference curves and place nonparametric or parametric bounds on valuations. The same methods extend naturally to variations in the non-price characteristics of products, and merge with the marketing literature on 'conjoint choice' (for example, Louviere, Hensher and Swait, 2000; Lusk and Schroeder, 2004). Access to scanner data from the massive volume of retail transactions made every day promises rich characterizations of underlying utility functions, particularly when merged with experimental methods that introduce exogenous variation in characteristics in order to statistically condition and 'enrich' the data (Hensher, Louviere and Swait, 1999). One of the attractions of indirect methods is that one can employ choice tasks which are familiar to the subject, such as binary 'take it or leave it' choices or rank orderings. The lack of precision in that type of qualitative data requires some latent structure before one can infer values, but behavioural responses are much easier to explain and motivate for respondents.

One major advantage of undertaking structural estimation of a latent choice model is that valuations can be elicited in a more fundamental manner, explicitly recognizing the decision process underlying a stated valuation. A structural model can control for risk attitudes when choices are being made in a stochastic setting, which is almost always the case in practical settings. Thus one can hope to tease apart the underlying deterministic valuation from the assessment of risk. Likewise, non-standard models of choice posit a myriad of alternative factors that might confound inference about valuation: respondents might distort preferences from their true values, they might exhibit loss aversion in certain frames, and they might bring their own home-grown reference points or aspiration levels to the valuation task. Only with a structural model can one hope to identify these potential confounds to the valuation process. Quite apart from wanting to identify the primitives of the underlying valuation free of confounds, normative applications will often require that some of these distortions be corrected for. That is only possible if one has a complete structural model of the valuation process.

A structural model also provides an antidote to those that claim that valuations are so contextual as to be an unreliable will-o'-the-wisp. If someone is concerned about framing, endowment effects, loss aversion, preference distortions, social preferences, and any number of related behavioural notions, it is impossible to generate a scientific dialogue without being able to write out a structural model and jointly estimate it.

Lessons and concerns

The most important lesson that has been learned from decades of experimental research into the behavioural properties of these procedures to elicit values is: keep it simple. This refers primarily to the nature of the task given to respondents. It can be dangerous to rely on fancy rules that ensure incentives to truthfully reveal valuations only if everyone sees a complete chain of logic, even if that logic is apparent to trained economists. Of course, one can use 'cheap talk' and just tell people to reveal the truth since it is in their best interests, but one cannot be sure that such admonitions work reliably. Cultural familiarity with institutions counts for a lot when subjects are otherwise placed in an artefactual

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
valuation task.

The desire to keep it simple has a corollary: the use of more rigorous statistical techniques to infer valuations. This implication follows from the need to make inferences about valuations on a cardinal scale when responses are often between subject and qualitative. Progress has been made in the use of numerical simulation methods for the maximum likelihood estimation of random utility models that allow extraordinary flexibility (for example, Train, 2003).

We also have a better understanding now of the manner in which valuations may be biased by being hypothetical, due to procedural devices in the institution being employed, and because of field context (for example, Harrison, Harstad and Rutström, 2004). More constructively, methods have been developed to undertake *ex ante* 'instrument calibration' to remove biases using controlled experiments, and to implement *ex post* 'statistical calibration' to filter out any remaining systematic biases (see Harrison, 2006).

Finally, the manner in which valuations change with states of nature is starting to be understood. Insights here again come from thinking about valuation as a latent, structural decision process. If we observe the same person state a different value for the same product at two different times, is it because he has a shift in his utility function, a change in some argument of his utility function, a change in his perceived opportunity set, or something else? If valuation is viewed as a process we can begin to design procedures that can help us identify answers to these questions, and better understand the valuations that are observed.

See Also

- auctions (experiments)
- demand theory
- experimental economics
- experimental methods in economics
- revealed preference theory
- utility

Bibliography

Andersen, S., Harrison, G.W., Lau, M.I. and Rutström, E.E. 2006. Elicitation using multiple price lists. *Experimental Economics* 4, 383–405.

Andersen, S., Harrison, G.W., Lau, M.I. and Rutström, E.E. 2007. Valuation using multiple price list formats. *Applied Economics* 39 675–82.

Becker, G.M., DeGroot, M.H. and Marschak, J. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9, 226–32.

Harrison, G.W. 2006. Experimental evidence on alternative environmental valuation methods.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Environmental and Resource Economics 34, 125–62.

Harrison, G.W., Harstad, R.M. and Rutström, E.E. 2004. Experimental methods and elicitation of values. *Experimental Economics* 2, 123–40.

Harrison, G.W., Lau, M.I. and Williams, M.B. 2002. Estimating individual discount rates for Denmark: a field experiment. *American Economic Review* 5, 1606–17.

Harstad, R.M. 2000. Dominant strategy adoption and Bidders' experience with pricing rules. *Experimental Economics* 3, 261–80.

Hensher, D., Louviere, J. and Swait, J.D. 1999. Combining sources of preference data. *Journal of Econometrics* 89, 197–221.

Hoffman, E., Menkhaus, D.J., Chakravarti, D., Field, R.A. and Whipple, G.D. 1993. Using laboratory experimental auctions in marketing research: a case study of new packaging for fresh beef. *Marketing Science* 3, 318–38.

Holt, C.A. and Laury, S.K. 2002. Risk aversion and incentive effects. *American Economic Review* 5, 1644–55.

Louviere, J.J., Hensher, D.A. and Swait, J.D. 2000. *Stated Choice Methods: Analysis and Application*. New York: Cambridge University Press.

Lusk, J.L. and Schroeder, T.C. 2004. Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics* 2, 467–82.

Rutström, E.E. 1998. Home-grown values and the design of incentive compatible auctions. *International Journal of Game Theory* 3, 427–41.

Train, K.E. 2003. *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.

How to cite this article

Harrison, Glenn W. "value elicitation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000075>
doi:10.1057/9780230226203.1782

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

value judgements

John C. Harsanyi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Harsanyi, J. C.; preferences; public choice; value judgements

Article

The claim of objective validity

One may define value judgements as judgements of approval or disapproval claiming objective validity. Many of our judgements of approval and disapproval do not involve such claims. When I say that I like a particular dish, I do not mean to imply that other people ought to like it too or that those disliking it are making a mistake. All I am doing is expressing my personal preference and my personal taste. (But an expert chef or an expert food critic may very well claim that his judgements about food have some degree of objective validity – in the sense that other gastronomic experts would tend to agree with his judgements. Of course, it is an empirical question whether his claim would be justified and, more generally, how much agreement there is in fact among expert judges of food.) Yet when I say that Hitler's murder of many millions of innocent people was a moral outrage, I do mean to do more than express my personal moral attitudes and do mean to imply that anybody who tried to defend Hitler's actions would be morally wrong.

In claiming objective validity, value judgements resemble factual judgements (both those dealing with empirical facts and those dealing with logical–mathematical facts). But they resemble judgements of personal preference in expressing human attitudes (those of approval or disapproval) rather than expressing beliefs about matters of fact, as factual judgements do. But this immediately poses a difficult philosophical problem: We can understand what it means for factual judgements to be objectively valid, that is, to be true, or to be objectively invalid, that is, to be false. They will be true if they describe the relevant facts as these facts actually are, and will be false if they fail to do so. But in what sense can judgements expressing human attitudes be objectively valid or invalid?

It seems to me that this can happen in at least two different ways. Such judgements can be objectively invalid either because they are contrary to the facts or because they are based on the wrong value

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
perspective. Value judgements can be contrary to the facts in the following sense: When we form our attitudes, we do so on the basis of some specific factual assumptions so that our attitudes and our judgements expressing these attitudes will be contrary to the facts if they are based on false factual assumptions. Mistaken factual assumptions may vitiate both our value judgements about instrumental values and those about intrinsic values. Thus, if I approve of using A as a means to achieve some end B, I will do this on the assumption that A is causally effective in achieving B. Hence, my approval will be mistaken if this assumption is incorrect. Likewise, if I approve of A as an intrinsically desirable goal, I will do this on the assumption that A has some qualities I find intrinsically attractive. My approval will be mistaken if in fact A does not possess these qualities.

Another way of value judgement may be objectively invalid is by being based on a value perspective different from the one it claims to have. For example, I may claim that my support for some government policy is based on its being in the public interest, even though actually it is based on its being in my own personal interest. Or, I may praise a very undistinguished novel as a great work of art merely because it supports my own political point of view. When a person claims to base his value judgement on one value perspective though actually he bases it on another, he may be simply lying, being fully aware of not telling the truth. Another possibility is that he is unaware, or only half aware, of using a value perspective different from the one he claims to use. (Likewise, when a person is making a value judgement based on false factual assumptions, he may or may not be fully aware of the falsity of these assumptions.)

Disagreements in value judgements

As we all know, disagreements in value judgements are extremely common and in many cases are very hard, or even impossible, to resolve. It seems to me that in most cases careful analysis would show that these disagreements about values are based on disagreements about the facts. Yet they may be very hard to resolve because these factual disagreements may be about very subtle facts about which reliable information is very hard, or even impossible, to obtain. For instance, our value judgements about a person's behaviour will often crucially depend on what we think his motives are. Some observers may attribute very noble motives to him, while others may do the opposite. Yet the available evidence might be consistent with either assumption. Other value judgements we make may hinge on our predictions about future facts. Thus, different economists may advocate very different economic policies because they have very different expectations on the likely effects of specific policies – even if their ultimate policy objectives are much the same. Yet, at the present stage of our knowledge about the economic system, we may be unable to tell with any degree of confidence which predictions are right and which are wrong.

Of course, we could avoid most of these disagreements if we refrained from making value judgements until we could ascertain with some assurance that the factual assumptions underlying the value judgements we want to make are correct. But this would require more intellectual self-discipline than most of us can muster. We have to act one way or another; and it is psychologically much easier for us to act if we can manage to entertain value judgements justifying our actions – even if the factual assumptions underlying these value judgements go far beyond, or are even clearly inconsistent with, the available evidence.

Let me add that most disagreements in value judgements are not disagreements about what the basic

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
values of human life actually are. Rather, most disagreements are about the relative weights and the relative priorities to be assigned to different basic values. Some individuals and some societies will learn from their experience – possibly based on a very idiosyncratic personal or national history – that things tend to work out best if value A is given far greater weight than value B is. Other individuals and other societies will reach very much the opposite conclusion on the basis of their experience. Once a given ranking of these two values has been adopted, it may be retained for a long time even when conditions change and make this ranking utterly inappropriate. For instance, an individual or a society that suffered a good deal from lack of individual freedom may be so preoccupied with political liberty as to neglect the need for social discipline – even under conditions that would make the need for social discipline paramount.

Besides disagreements about the facts, another source of value conflicts is philosophical disagreements about the correct value perspectives to be used in making various classes of value judgements. For instance, even if two people agree about all the relevant facts, they may still make conflicting moral value judgements if they disagree about the nature of morality and, therefore, disagree about the nature of the moral perspective to be used in making moral value judgements. (For instance, one individual may favour a utilitarian interpretation of morality – see, for example, Harsanyi, 1977 – while the other may favour an entitlement interpretation – see Nozick, 1974.) In the same way, disagreements about the nature of the aesthetic perspective to be used in making aesthetic value judgements may lead to disagreements about the artistic quality of various works of art.

Value judgements in economics

There was a time when many economists wanted to ensure the objectivity of economic analysis by excluding value judgements, and even the study of value judgements, from economics. (A very influential advocate of this position has been Robbins, 1932.) Luckily, they have not succeeded; and we now know that economics would have been that much poorer if they had.

After some important preliminary work in the 1930s and the 1940s, mainly in welfare economics, a new era in the study of economically relevant value judgements, has started with Arrow's *Social Choice and Individual Values* (1951). This book has shown how to express alternative value judgements in the form of precisely stated formal axioms, how to investigate their logical implications in a rigorous manner, and how to examine their mutual consistency or inconsistency. Arrow's book and the research inspired by it have greatly enriched economic theory not only in welfare economics but also in several other fields, including the theory of competitive equilibrium. It has given rise to a new sub-discipline called *public choice theory*, which is a rigorous study of voting and of alternative voting systems and which has made important contributions to the study of alternative political systems and of alternative moral codes and, more indirectly, to the study of alternative economic systems as mechanisms of social choice.

Of course, value judgements often play an important role in economics even when they are not the main subjects of investigation. They influence the policy recommendations made by economists and their judgements about the merits of alternative systems of economic organization. But this need not impair the social utility of the work done by economists as long as it is work of high intellectual quality and as long as the economists concerned *know* what they are doing, *know* the qualifications their conclusions are subject to, and *tell* their readers what these qualifications are. In particular, intellectual honesty

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

requires economists to *state* their political and moral value judgements and to make clear how their conclusions differ from those that economists of different points of view would tend to reach on the problems under discussion. What is no less important, they should make clear how *uncertain* many of their empirical claims and their predictions actually are. This is particularly important in publications addressed mainly to people outside the economist profession.

See Also

- interpersonal utility comparisons
- philosophy and economics
- positivism

Bibliography

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Harsanyi, J.C. 1977. Morality and the theory of rational behavior. *Social Research* 44, 623–56.

Nozick, R. 1974. *Anarchy, State and Utopia*. Oxford: Blackwell.

Putnam, H. 2004. *The Collapse of the Fact/Value Dichotomy and Other Essays*. Cambridge, MA: Harvard University Press.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

How to cite this article

Harsanyi, John C. "value judgements." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000004> doi:10.1057/9780230226203.1783

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

value of life

W. Kip Viscusi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The economic approach to valuing risks to life focuses on risk–money trade-offs for very small risks of death, or the value of statistical life (VSL). These VSL levels will generally exceed the optimal insurance amounts. A substantial literature has estimated the wage–fatality risk trade-offs, implying a median VSL of $7 million for US workers. International evidence often indicates a lower VSL, which is consistent with the lower income levels in less developed countries. Preference heterogeneity also generates different trade-off rates across the population as people who are more willing to bear risk will exhibit lower wage–risk trade-offs.

Keywords

compensating differentials; contingent valuation; deterrence; hedonic models; hedonic prices; human capital; India; life insurance; risk to life; Schelling, T.; Smith, A.; South Korea; Taiwan; value of life; value of statistical life

Article

Issues pertaining to the value of life and risks to life are among the most sensitive and controversial in economics. Much of the controversy stems from a misunderstanding of what is meant by this terminology. There are two principal value-of-life concepts – the amount that is optimal from the standpoint of insurance, and the value needed for deterrence. These concepts address quite different questions that are pertinent to promoting different economic objectives.

The insurance value received the greater attention in the literature until recent decades. The basic principle for optimal insurance purchases is that it is desirable to continue to transfer income to the post-accident state until the marginal utility of income in that state equals the marginal utility of income when healthy. In the case of property damage, it is desirable to have the same level of utility and marginal utility of income after the accident as before. In contrast, fatalities and serious injuries affect one's utility function, decreasing both the level of utility and the marginal utility for any given level of income, making a lower income level after a fatality desirable from an insurance standpoint. Thus, the value of life and limb from the standpoint of insurance may be relatively modest.

The second approach to valuing life is the optimal deterrence amount. What value for a fatality sets the appropriate incentives for those avoiding the accident? In the case of financial losses, the optimal insurance amount, the optimal deterrence amount, and the ‘make whole’ amount are identical; however, for severe health outcomes, such as fatalities, the optimal deterrence amount will exceed the optimal level of compensation.

The economic measure for the optimal deterrence amount is the risk–money trade-off for very small risks of death. Since the concern is with small probabilities, not the certainty of death, these values are referred to as the value of statistical life (VSL). Economic estimates of the VSL amounts have included evidence from market decisions that reveal the implicit values reflected in behaviour as well as the use of survey approaches to elicit these money–risk trade-offs directly. Government regulators in turn have used these VSL estimates to value the benefits associated with risk reduction policies. Because of the central role of VSL estimates in the economics literature, those analyses will be the focus here rather than income replacement for accident victims.

Valuing risks to life

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Although economics has devoted substantial attention to issues generally termed the ‘value of life’, this designation is in many respects a misnomer. What is at issue is usually not the value of life itself but rather the value of small risks to life. As Schelling (1968) observed, the key question is how much people are willing to pay to prevent a small risk of death. For small changes in risk, this amount will be approximately the same as the amount of money that they should be compensated to incur such a small risk. This risk–money trade-off provides an appropriate measure of deterrence in that it indicates the individual's private valuation of small changes in the risk. It thus serves as a measure of the deterrence amount for the value to the individual at risk of preventing accidents and as a reference point for the amount the government should spend to prevent small statistical risks. Because the concern is with statistical lives, not identified lives, analyses of government regulations now use these VSL levels to monetize risk reduction benefits.

Suppose that the amount people are willing to pay to eliminate a risk of death of 1/10,000 is $700. This amount can be converted into a value of statistical life estimate in one of two ways. First, consider a group of 10,000 individuals facing that risk level. If each of them were willing to contribute $700 to eliminate the risk, then one could raise a total amount to prevent the statistical death equal to 10,000 people multiplied by $700 per person, or $7 million. An alternative approach to conceptualizing the risk is to think of the amount that is being paid per unit risk. If we divide the willingness to pay amount of $700 by the risk probability of one in 10,000, then one obtains the value per unit risk. The value per statistical life is $7 million using this approach as well.

Posing hypothetical interview questions to ascertain the willingness-to-pay amount has been a frequent survey technique in the literature on the value of life. Such studies are often classified as ‘contingent valuation surveys’ or ‘stated preference surveys’, in that they seek information regarding respondents’ decisions given hypothetical scenarios (see Jones-Lee, 1989; Viscusi, 1992). Survey evidence is most useful in addressing issues that cannot be assessed using market data. How, for example, do people value death from cancer compared with acute accidental fatalities? Would people be interested in purchasing pain-and-suffering compensation, and does such an interest vary with the nature of the accident? Potentially, survey methods can yield insights into these issues.

Evidence from actual decisions that people make is potentially more informative than trade-offs based on hypothetical situations if suitable market data exists. Actual decision-makers are either paying money to reduce a risk or receiving actual compensation to face a risk, which may be a quite different enterprise from dealing with hypothetical interview money. In addition, the risks to them are real so that they do not have to engage in the thought experiment of imagining that they face a risk. It is also important, however, that individuals accurately perceive the risks they face. Surveys can present respondents with information that is accurate. Biased risk perceptions may bias estimates of the money–risk trade-off in the market. Random errors in perceptions will bias estimates of the trade-off downward. The reason for this result can be traced to the standard errors-in-variables problem. A regression of the wage rate on the risk level, which is measured with error, will generate a risk variable coefficient that will be biased downward if the error is random. The estimated wage–risk trade-off will consequently understate its true value.

Empirical evidence on the value of statistical life

A large literature has documented significant trade-offs between income received and fatality risks. Most of these studies have examined wage–risk trade–offs but many studies have focused on product and housing risks as well. The wage-risk studies have utilized data from the United States as well as many other countries throughout the world. The primary implication of these results is that estimates of the value of life in the United States are clustered in the $4 million to $10 million range, with an average value of life in the vicinity of $7 million.

Since the time of Adam Smith (1776), economists have observed that workers will require a ‘compensating differential’ to work on jobs that pose extra risk. These wage premiums in turn can be used to assess risk–money trade-offs and the value of life. The underlying methodology used for this analysis derives from the hedonic price and wage literature, which focuses on ‘hedonic’ or ‘quality-adjusted’ prices and wages. Rosen (1986) and Smith (1979), among others, review this methodology.

To see how the hedonic model works, let us begin with the supply side of the market. The worker's risk decision is to choose the job with the fatality risk p that provides the highest level of expected utility (*EU*). The worker faces a market offer curve $w(p)$ that is the outer envelope of the individual firms’ market offer curves. Let there be two states of the world: good health with utility $U(w)$ and death with utility $V(w)$, where this term is the worker's bequest function. The utility function has the property that good health is preferable to ill health, and workers are risk-averse or risk-neutral, or $U(w) > V(w)$; $U', V' > 0$; and $U'', V' \leq 0$. The job choice is to

$$\underset{p}{MAX}\ EU = (1 - p)U(w(p)) + pV(w(p)),$$

leading to the result

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\frac{dw}{dp} = \frac{U(w) - V(w)}{(1-p)U'(w) + pV'(w)}.$$

The wage-risk trade-off *dw*/*dp* based on the worker's choice of a wage–risk combination for a job is the value of statistical life, which equals the difference in utility between the two health states divided by the expected marginal utility of consumption.

What trade-off rate *dw*/*dp* the worker will select will depend not only on worker preferences but also on the shape of the market offer curve. The best available market opportunities will be those that offer the highest wage for any given level of risk, or the outer envelope of the offer curves for the individual firms. Each individual firm will offer a wage that is a decreasing function of the level of safety. The cost function for producing safety increases with the level of safety, so the wage decline associated with incremental improvements in safety must be increasingly great to keep the firm on its isoprofit curve.

Figure 1 illustrates the nature of the hedonic labour market equilibrium. The curves OC_1 and OC_2 represent two possible market offer curves from firms with risky jobs. As the risk level is reduced, firms will offer lower wages. EU_1 and EU_2 are expected utility loci of two workers, both of whom have selected their optimal job risk from available market opportunities. The curve $w(p)$ represents the locus of market equilibria, which consists of the points at which worker indifference curves are tangent to the market offers. Thus, the empirical estimation of the hedonic labour market equilibrium focuses on the joint influence of demand and supply.

Figure 1

Market process for determining compensating differentials

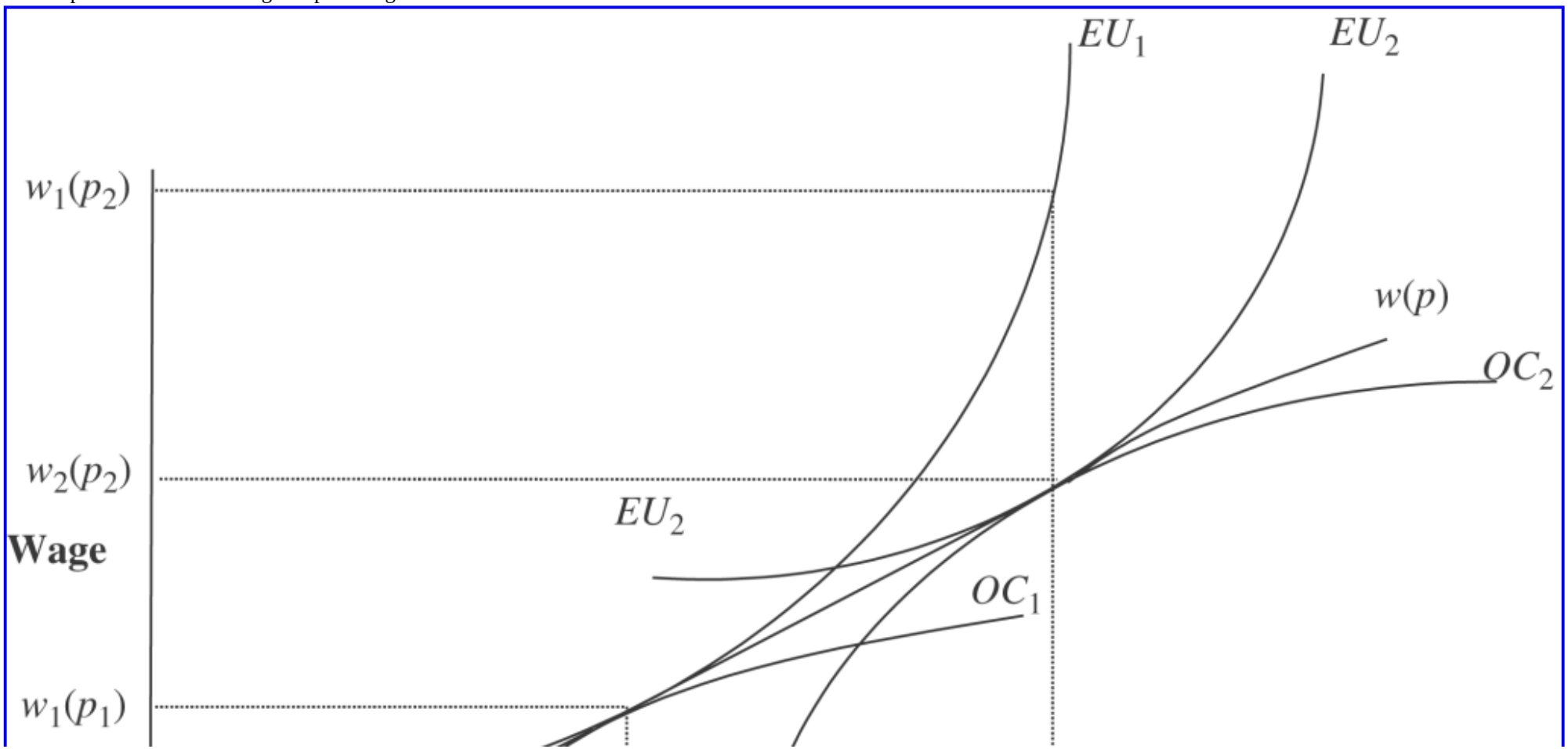

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

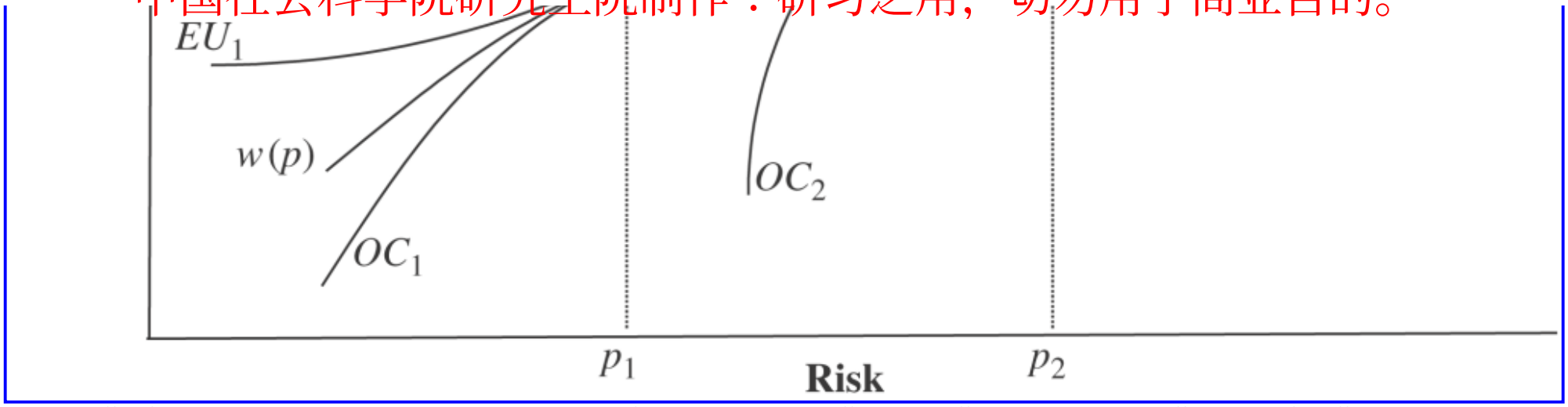

The trade-offs reflected in market equilibria do not represent a schedule of individual VSL trade-off values at different risks, but rather different VSLs for different workers. Worker 1 chooses risk p_1 with associated wage $w(p_1)$, and worker 2 chooses risk p_2 for wage $w(p_2)$. However, worker 1 would not accept risk p_2 for $w(p_2)$ even when that is the point on the hedonic equilibrium curve. Rather, worker 1 will require wage $w_1(p_2) > w_2(p_2)$ to accept this risk.

The canonical hedonic wage equation is

$$\ln w_i = \alpha + X_i'\beta + \gamma_1 p_i + \gamma_2 q_i + \gamma_3 WC_i + \varepsilon_i,$$

where w_i is worker i's wage, X_i is a vector of personal characteristics and job characteristics, p_i is the worker's fatality risk, q_i is the nonfatal injury and illness risk, and WC_i is a measure of the worker's compensation benefits. Not all labour market studies of VSL include the q_i and WC_i terms. Moreover, there are some differences in the form of the workers' compensation benefit term that is included. The most common is the expected workers' compensation replacement rate, which is the product of the injury risk and the benefit level divided by the wage rate. These differences in the empirical specification account for some of the differences across studies in the estimated VSL.

As a practical matter, there are many systematic differences that have becomes apparent in these studies. Workers at very high-risk jobs tend to have lower values of life on average since they have self-selected themselves into the very risky occupation. Through their job choices these individuals have revealed their greater willingness to endanger their lives. Workers at lower-risk jobs typically have greater reluctance to risk their lives, which accounts for their selection into these safer pursuits. Such differences are apparent in practice, as the estimated values of life for workers in the average risk jobs tend to be several times greater than those for workers in very risky jobs.

Other differences correlated with worker affluence are also evident. Health status is a normal economic good, and individuals' willingness to pay to preserve their health increases with income. Blue-collar workers, for example, have a lower value of life than do white-collar workers. In addition, there is a positive income elasticity of the estimated values of risks to life and health. Based on a sample of 50 wage–risk studies from ten countries, Viscusi and Aldy (2003) estimate that VSL has an income elasticity of 0.5 to 0.6.

These differences by income level in the VSL amounts are also borne out in the international evidence on wage-risk trade-offs, such as the study of Australia and Japan by Kniesner and Leeth (1991). Table 1 summarizes representative VSL studies from throughout the world. More affluent countries such as Japan and Canada tend to have higher revealed VSL levels than countries such as South Korea, India and Taiwan. The major international anomaly is the United Kingdom, for which labour market estimates have been very unstable across studies and sometimes quite high. Deficiencies of the UK fatality risk data or correlation of these values with other unobservables may account for this pattern. Because of these limitations, the benefit assessments for risk reductions in the UK are based on stated preference values rather than labour market values, which is the approach taken by US regulatory agencies.

Labour market estimates of value of statistical life throughout the world

Study/Country	Value of statistical life ($ millions)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Median value from 30 US studies	7.0
Australia	4.2
Austria	3.9 – 6.5
Canada	3.9 – 4.7
Hong Kong	1.7
India	1.2 – 1.5
Japan	9.7
South Korea	0.8
Switzerland	6.3 – 8.6
Taiwan	0.2 – 0.9
United Kingdom	4.2

Note: All estimates are in year 2000 US dollars.*Source*: Viscusi and Aldy (2003). For concreteness, single representative studies are drawn from their Table 4.

Because of individual heterogeneity in preferences and resources, it is not surprising that estimated values of life often differ considerably across empirical studies. These differences are not a sign that such studies are necessarily in error. These samples often consist of workers with quite different risk levels and who are situated differently. International comparisons, for example, consistently reveal differences across countries, not only because of the aforementioned aspects of heterogeneity, but because of the differences in the social insurance and workers' compensation arrangements that may be present in these countries.

The role of heterogeneity is evidenced in estimates for the implicit value for non-fatal job injuries for different worker groups. This analysis follows the same general methodological approach as does the literature on the implicit value of life. The difference is that the focus is on non-fatal job risks rather than fatalities. On average, workers value non-fatal loss injuries on the job at values ranging from $20,000 to $70,000 per expected job injury. Thus, for example, a worker at the high end of this range would require $2,000 to face a one chance in 25 of being injured that year.

The estimates of the implicit values of injuries for other labour market groups who have different attitudes towards risk vary substantially from this amount. Interestingly, women often work at hazardous jobs and appear to have wage–risk trade-offs similar to those of men. Other personal characteristics generate more evidence of heterogeneity in preferences. Cigarette smokers and people who don use seat belts in their automobiles work on risky jobs for less per expected injury than do people who don't smoke and who use seat belts in their automobiles. What is noteworthy is that these results are not hypothetical willingness-to-pay values that these groups have expressed with respect to risks. Rather, they represent actual differences in compensation based on observed patterns of decisions in the marketplace. Markets work as expected in that they match workers to the jobs that are most appropriate for their preferences. This is a constructive role of market sorting that promotes a more efficient match-up than if, for example, all individuals were constrained to have the same job riskiness.

Preference heterogeneity has additional implications. Recall from Figure 1 that workers may settle along different points of the available market opportunities. However, if workers face the same opportunities locus, then the worker choosing the higher risk p_2 must always be paid a wage $w(p_2) > w(p_1)$ if $p_2 > p_1$. Interestingly, that pattern does not always hold. As shown by Viscusi and Hersch (2001), smokers choose jobs that are riskier than non-smokers' jobs but offer less additional wage compensation for incurring the risks. Smokers and non-smokers face different market offer curves and, most important, these offer curves provide for a flatter wage–risk gradient for smokers. There may be an efficiency-based rationale for these differences, as smokers are more prone to job accidents, so that there safety-related productivity is less.

Studies of the money–risk trade-offs are not restricted to the labour market. There have been a number of efforts to assess price–risk trade-offs for a variety of commodities. The contexts analysed by economists include the choice of highway speed, seat belt use, installation of smoke detectors, property values in polluted areas, and prices of automobiles. The most reliable of these studies outside the labour market are those pertaining to automobile prices in that they follow the same kind of approach as is used in the wage–risk literature. In particular, the analysts obtain price information on a wide variety of automobile models. Using regression analysis, they assess the incremental contribution of the safety characteristics per se to the product price, controlling for other product attributes. The results of these studies suggest a value of life around $5 million.

The duration and quality of life

The value-of-life terminology is misleading to the extent that risk reduction efforts do not confer immortality but simply extend life. Because of that, the major concern should not be with the value of life but with the value of extending life for different periods. In the case of preventing the risk of death to a young person, the increase in life expectancy that will be generated will exceed that for preventing a risk of death to older people. Some kind of age adjustment may be appropriate. The quantity of life matters, but which years of life matter

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

most? Is a year of life at age 45 more valuable than a year of life at age 5 or age 70? How do various health impairments correlated with age affect the value one should attach to such years of life, and should the fact that very young children have not yet received the value of the education and rearing by their parents matter? The total 'human capital', which is the set of personal attributes such as education and training that affect one's income, will be greater for older children who are further along in their development. Resolving such questions remains highly problematic.

Considerable attention has been devoted to economic analysis of age effects, including studies by Shepard and Zeckhauser (1984) and Johansson (2002). If capital markets were perfect, then VSL would steadily decline with age, reflecting the shortening of life expectancy. If, however, there are capital market imperfections, then VSL will display an inverted U-shaped relationship with age. A similar pattern is exhibited empirically by lifetime consumption patterns, which some theoretical models have linked to VSL levels over the life cycle. Although empirical estimates of the age effects are still being refined, the available evidence from survey data and market-based studies suggests that there is an inverted-U-shaped relation. The main empirical controversies concern the tails of the age distribution. To what extent is there a flattening of the VSL–age relation for the very old age groups, and how should VSL levels be assigned to children?

The quality of the life of the years saved clearly matters as well. Life years in deteriorating health may be less valuable to the individual than years in good health. Some analysts have suggested that the measure should focus on quality-adjusted life years. Making these quality adjustments has yet to receive widespread empirical implementation and is often controversial. There may be quite legitimate fears of government efforts to target expenditures by denying health care to those whose life quality is deemed to be low. People often adapt to changes in health status so that external observers may overstate the decline in well-being that occurs with serious illnesses.

Conclusion

Economic estimates of the trade-offs people make between risk and either prices or wages serve a variety of functions. First, they provide evidence on how people make decisions involving risk in labour market and product market contexts. The fact that there are probabilistic health effects does not imply that markets cease to function. Second, these estimates have proved useful in providing a reference point for how the government should value the benefits associated with regulations and other policies that reduce risk. Third, the existence of these estimates and economists' continuing efforts to refine the values has served to highlight many of the fundamental ethical issues involved, such as how society should value reducing risks to people in different age groups.

See Also

- compensating differentials
- hedonic prices

Bibliography

Hersch, J. 1998. Compensating differentials for gender-specific job injury risks. *American Economic Review* 88, 598–627.

Johansson, P.-O. 2002. On the definition and age-dependency of the vale of statistical life. *Journal of Risk andUncertainty* 25, 251–63.

Jones-Lee, M. 1989. *The Economics of Safety and Physical Risk*. Oxford: Basil Blackwell.

Kniesner, T. and Leeth, J. 1991. Compensating wage differentials for fatal injury risk in Australia, Japan, and the United States. *Journal of Risk and Uncertainty* 4, 75–90.

Rosen, S. 1986. The theory of equalizing differences. In *Handbook of Labor Economics*, ed. O. Ashenfelter and R. Layard.Amsterdam: North-Holland.

Schelling, T. 1968. The life you save may be your own. In *Problems in Public Expenditure Analysis*, ed. S. Chase. Washington, DC: Brookings Institution.

Shepard, D. and Zeckhauser, R. 1984. Survival versus consumption. *Management Science* 30, 423–39.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Smith, A. 1776. *The Wealth of Nations*. Ed. E. Cannan. New York: Modern Library, 1937.

Smith, R. 1979. Compensating differentials and public policy: a review. *Industrial and Labor Relations Review* 32, 339–52.

Viscusi, W. 1992. *Fatal Tradeoffs: Public and Private Responsibilities for Risk*. New York: Oxford University Press.

Viscusi, W. and Aldy, J. 2003. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27, 5–76.

Viscusi, W. and Hersch, J. 2001. Cigarette smokers as job risk takers. *Review of Economics and Statistics* 83, 269–80.

Zeckhauser, R. and Shepard, D. 1976. Where now for saving lives? *Law and Contemporary Problems* 39, 5–45.

How to cite this article

Viscusi, W. Kip. "value of life." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000005> doi:10.1057/9780230226203.1784

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

value of time

Reuben Gronau
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Time is a scarce resource or, to use a popular adage – ‘time is money’. The value of time depends on its usage and the complementary resources used with it. Firms pay for their workers’ time according to the workers’ value or marginal product. Households naturally place a value on their time when they sell it in the market, but they also assign a value to the time they use in the home sector. This value determines (and is sometimes determined by) the optimum combination of activities a person engages in, and the optimum combination of goods and time used in each activity. It affects the supply of labour and the demand for goods.

The recognition of the importance of time for many economic decisions related and unrelated to the labour market (e.g., schooling; transportation) is not new. The generalization of the model is associated with Becker's (1965) theory of home production. Becker (following Mincer, 1963) reformulates traditional consumption theory by shifting the focus of analysis from goods to activities (‘commodities’, in his terms). By this approach the source of the household's welfare is its activities, which in turn, are a combination of goods and time. Welfare is maximized subject to home technology, the budget constraint, and the time constraint. Formally, the welfare function depends on the activity levels (Z_i)

$$U = U(Z_1, \ldots, Z_n)$$

where each activity is ‘produced’ through a combination of goods (X_i) and time (T_i)

$$Z_i = d_i(X_i, T_i).$$

The consumer's welfare is maximized subject to the budget constraint

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\sum P_i X_i = W(Z_n) + V$$

and the time constraint

$$\sum T_i = T,$$

where P_i denote prices, $W(Z_n)$ is labour income (Z_n denoting the activity work in the market), V is non-labour income, and T is the total time available.
The maximization of welfare subject to the home production technology and the time and budget constraints yields the optimum allocation of activities:

$$\partial U / \partial Z_i = \lambda \hat{\Pi}_i,$$

and the optimum combination of inputs in the production of each activity

$$(\partial Z_i / \partial T_i) / (\partial Z_i / \partial X_i) = \hat{W} / P_i,$$

where λ denotes the marginal value of income, $\hat{\Pi}_i$ is the shadow price of activity i, and $\hat{W}$ is the shadow price of time. The shadow price of the activity equals its marginal cost of production

$$\hat{\Pi}_i = P_i(\partial X_i / \partial Z_i) + \hat{W}(\partial T_i / \partial Z_i).$$

Thus an increase in the shadow price of time leads to substitution of time in favour of goods and a substitution from time-intensive to goods-intensive activities.
When there are no external constraints on hours of market work the value people place on their time depends on their marginal wage rate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\hat{w} = w + (u_n / \lambda),$$

where w is the marginal wage rate (the change in earnings as a result of a change in market work net of taxes and any expenditures associated with work) and u_n denotes the marginal utility of labour.

However, even when one is not free to change one's working hours the shadow price of time increases with wages and with income because of the increase in time scarcity.

The importance of the value of time to allocative decisions has been shown in a wide range of contexts: fertility (Becker, 1960; Willis, 1973; Schultz, 1975), health (Grossman, 1972) and most notably, labour supply and transportation. Thus, women with higher wages have higher opportunity costs of raising children and therefore tend to reduce fertility, substituting ‘quality’ for ‘quantity’. Travellers who place a high value on their time prefer faster but more expensive modes of transport to slower and cheaper modes. Married women with young children or with high earning husbands place higher value on their time and are, therefore, more reluctant to participate in the labour force.

Theory predicts that the shadow price of time changes with the person's wage rate. It does not imply that the two are equal; they differ if the marginal net wage differs from the average wage, when labour involves direct utility (or disutility), or when it is assumed that the utility generated by an activity depends on the time inputs involved (Bruzelius, 1979).

The value of time saving is a major component of the benefits of the investment in many transportation projects (Beesley, 1965; Tipping, 1968). To evaluate the shadow price of time transportation economists studied the trade-off between time and money implicit in the choice of modes of transport, choice of route, location decision, and demand for travel. Studying commuter choices it is found that the value placed by commuters on their time is only 1/5 to 1/2 of their wage rate. The value of walking and waiting time is found to be 2.5–3.0 times greater than the value of in-vehicle time. Differences in convenience, comfort, effort etc. are reflected in estimates of time value in bus travel that are higher than travel by car, and values that tend to increase with the length of the trip. Finally, differences between the gross and the net wage and constrained working hours result in estimates that are higher for business travel than for personal travel. (For a recent discussion of the estimating methods and results see Bruzelius, 1979.)

A second source for the study of the value of time at home is labour-force-participation behaviour. A person is supposed to participate in the labour force if the wage he is offered exceeds the value of his marginal productivity at home – that is, his value of home time. Studying the labour force participation patterns of US married women Gronau (1973) found that the value of time of these women increases with their schooling (most noticeably with college attendance). It is little affected by the husband's schooling and income and by age, and increase sharply when the family has children. Having a child under 3 years of age increases the value of its mother's time at home by over 25 per cent (in particular if she has a college education), but this effect diminishes as the child grows older.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- family economics
- gender roles and division of labour
- household production and public goods
- leisure
- women's work and wages

Bibliography

Becker, G.S. 1960. An economic analysis of fertility. In *Demographic and Economic Change in Developed Countries*, University-National Bureau Conference Series, No. 11, Princeton: Princeton University Press.

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75(September), 493–517.

Beesley, M.E. 1965. The value of time spent in travelling; some new evidence. *Economica* 45(May), 174–85.

Bruzelius, N. 1979. *The Value of Travel Time*. London: Croom Helm.

Gronau, R. 1973. The effect of children on the housewife's value of time. *Journal of Political Economy* 81(2), March–April, Supplement, 168–99.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80(2), March–April, 223–55.

Mincer, J. 1963. Market prices, opportunity costs, and income effects. In *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, ed. C. Christ et al., Stanford: Stanford University Press.

Schultz, T.W. (ed.) 1975. *Economics of the Family: marriage, children, and human capital*. London and Chicago: Chicago University Press for the National Bureau of Economic Research.

Tipping, D.G. 1968. Time savings in transport studies. *Economic Journal* 78(December), 843–54.

Willis, R.J. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81(2), March–April, Supplement, 14–64.

How to cite this article

Gronau, Reuben. "value of time." The New Palgrave Dictionary of Economics. Second Edition. Eds.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000006> doi:10.1057/9780230226203.1785

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

value-added tax

Gilbert E. Metcalf
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

A value-added tax (VAT) is a tax on the value created in goods or services during production, distribution, and sales. VATs are generally constructed as consumption taxes. In principle, a consumption VAT is neutral in its treatment of savings and consumption. In practice, VATs are often designed with exemptions and zero-ratings that generate consumption distortions. An important issue for VATs is their implementation in a federal system. A number of modifications of the basic VAT structure have been proposed to strike a balance between the efficiency of a common rate across political units and the desire for country-specific VAT rates.

Keywords

compensating VAT (CVAT); consumption taxation; depreciation; tax evasion; tax neutrality; value-added tax; vertical integration; viable integrated VAT (VIVAT)

Article

A value-added tax (VAT) is a tax on the value created in a good or service by a business at any stage of production, distribution or sales.

Definitions and equivalencies

Value-added is simply the difference between the value of the goods and services sold and the value of goods and services purchased as intermediate inputs. Consider the general cash flow equation for a firm.

$$S + K^{+} = L + M + K^{-}. \tag{1}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The equation states that cash comes into a firm from capital inflows (K^+) – new equity and borrowing – and proceeds from sales. Cash is used for payments for labour (L) and intermediate goods (M). Capital purchases are generally included in M. This makes a VAT a consumption tax. If capital depreciation is included in M, then this would be an income-type VAT. If no capital deduction of any form is allowed, then this would be a gross output VAT. For the remainder of the article, I focus on a consumption-type VAT. In addition, cash is retained or used for dividend and interest payments as well as any retirement of debt and equity (K^-).
Value added was defined above as the difference between revenue from sales and the cost of inputs:

$$VA = S - M = L + (K^- - K^+). \quad (2)$$

Equation (2) demonstrates that there are several ways to impose a VAT. We could tax gross sales net of intermediate input purchases at each stage of production. This forms the basis for a ‘subtraction method’ VAT. Alternatively, we could tax gross sales and allow a credit for taxes paid by registered suppliers of intermediate inputs to the firm. The ‘credit method’ VAT works in this fashion. A third method is to tax the factor payments to capital and labour. This forms the basis for an ‘addition method’ VAT.
Value-added taxes are common throughout the world with the notable exception of the United States. Most countries use the credit method, arguing that this method is self-enforcing since the ability to take a credit for VAT paid at an earlier stage of production requires suppliers to provide an invoice detailing their VAT payments.

Tax neutrality

As described at this most general level, a VAT shares all the attributes of a broad-based consumption tax. If comprehensively applied, it is neutral across all forms of purchased consumption. And since capital purchases are expensed (immediately deducted from the tax base), the rate of return on capital is unaffected by the tax. As with all consumption taxes, the efficiency gains from a switch from income to consumption taxation depend significantly on the tax treatment of old capital; on this point, see Auerbach and Kotlikoff (1987). In practice, VATs are not neutral for a number of reasons. First, if capital is not expensed, returns to capital are affected by the tax. Most VATs are consumption-type VATs so this is not a significant problem. Second, as noted by Cnossen (1998), some countries extend the VAT up through the manufacturing or wholesale stage but not through the retail stage. This creates distortions across consumption given the different ensuing tax rates on different commodities.
Finally, many VATs exempt certain sectors from the tax, ‘zero-rate’ sectors or commodities, or apply a reduced rate to certain commodities (for example, food for home preparation). Zero-rating in a credit method VAT means that firms apply a zero rate to their tax base but receive a credit for all VAT paid by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
suppliers. Zero-rating has no impact if applied at an intermediate stage of production since any taxes forgone at one stage are made up at the next stage. Zero-rating at the retail stage means the commodity is untaxed by a VAT. Exempting sectors from the VAT process may create peculiar outcomes. If an intermediate sector is exempted from taxation, downstream stages of production will pay a VAT not only on their value added but on the value added created in sectors upstream from the exempt sector for which no credit was received. The result is that exemptions at an intermediate stage of production can lead to the effective VAT rate being *higher* than the nominal rate. For this reason, many countries that exempt certain sectors (generally small businesses) allow voluntary participation in the VAT system. Note too that exemptions at the retail stage create incentives for vertical integration to increase the proportion of value added exempt from taxation.

Design issues

A VAT can be levied on an ‘origin’ or ‘destination’ basis. An origin VAT taxes value added in the country in which the value added is produced, while a destination VAT taxes value added where it is consumed. Most countries employ a destination VAT and use a border tax adjustment whereby a VAT is applied to the value of imports and a rebate provided for the value of exports. While it is popularly believed that border tax adjustments favour export industries, a flexible exchange rate in general leads to the same trade balance whether the VAT is applied on an origin or destination basis. Grossman (1980) demonstrates that this proposition fails in a world with trade in intermediate goods.

Border tax adjustments are commonly applied by customs authorities, and this has given rise to special problems for the European Union with its abolition of border controls in 1992. Keen and Smith (1996) note conflicts between two important goals: maximum autonomy for individual countries to set their own tax rates and a system of country VAT structures that does not impede the creation of a single European market. Keen and Smith propose a ‘viable integrated VAT’ (VIVAT) to address this problem. The VIVAT applies a harmonized VAT rate to intermediate producers in all European countries and a different rate for final consumption sales. The rate on final sales would vary across countries based on individual country preferences. The VIVAT can be thought of as a harmonized EU-wide VAT and a system of individual country retail sales taxes, a point that reminds us of the close connection between a VAT and a retail sales tax.

McLure (2000) notes that the VIVAT requires firms to charge different rates to different classes of customers, a non-trivial burden. He also notes that a destination-based system of VATs in a sub-national system can give rise to tax evasion by households or unregistered firms importing goods (which are zero-rated at the exporting country's border). McLure proposes a compensating VAT (CVAT), essentially an additional federal-level tax to guarantee the tax revenues that might otherwise be lost to cross-border tax evasion. The key point here is that considerable administrative complexity comes into play when a VAT is implemented by a group of countries (or states) within a common trading system (or federal government).

As with any other consumption tax framework, taxing housing and financial services is problematic with a VAT. One approach for treating housing services follows from an arbitrage argument that the present value of the stream of future consumption services from housing is equal to its purchase price. With this assumption, a tax-prepayment approach levies the VAT on the first sale of a house (but not subsequent

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sales) as well as additions or maintenance. With constant tax rates, this tax payment is equal to the present value of the taxes that would be paid on the housing services enjoyed by occupants of the house. If tax rates rise (fall) in the future, the tax prepayment approach raises less (more) revenue than if the housing services were taxed directly. Alternatively, the sale of all residential housing and rental income are subject to tax while the purchase of residential housing is deductible. This approach treats housing like any other capital asset which produces services (housing). Measuring and taxing the imputed rental income on owner occupied housing is a significant problem for this approach. For this reason, most consumption taxes favour the tax prepayment approach.

Financial services are even more difficult to handle under consumption taxes. One approach is to tax the net cash flow from financial services. In the terminology of Meade (1978), this would be an R+F (real plus financial) consumption tax. As Auerbach and Gordon (2002) point out, this creates considerable administrative problems since other transactions are treated on an R basis, thus giving rise to arbitrage opportunities to avoid the tax. In the European Union financial services are exempt from VAT, though Huizanga (2002) has argued that it is increasingly feasible to bring this sector into the VAT system. This sanguine perspective is not shared by all economists.

Tax incidence and impacts on saving and labour supply

Because a VAT in its purest form is a consumption tax, its distributional impact as well as behavioural impacts are the same as those of any broad-based consumption tax. To the extent that the VAT is implemented in non-neutral ways (exemptions and zero-rating of sectors, multiple tax rates, and so forth) consumption distortions will arise similar to those of any differential rate commodity tax system. Many countries apply a VAT tax structure with lower rates on perceived necessities (food, for example) on distributional grounds. Most economic analyses of VAT proposals recommend a uniform tax rate on all commodities to avoid consumption distortions, and recommend using an income tax to effect desired income redistribution. Cnossen (1998), however, recommends a dual rate system for developing countries on the grounds that income taxes are administratively unfeasible in these countries. While reducing tax rates on food and other necessities provides benefits to low-income households, this is a blunt instrument for redistribution given the resulting reduction in taxes to wealthy people's purchase of food (and other low or zero-rated commodities).

See Also

- consumption taxation
- fiscal federalism
- optimal taxation
- tax competition

Bibliography

Auerbach, A.J. and Gordon, R.H. 2002. Taxation of financial services under a VAT. *American*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Economic Review 92, 411–16.

Auerbach, A.J. and Kotlikoff, L.J. 1987. *Dynamic Fiscal Policy*. New York: Cambridge University Press.

Cnossen, S. 1998. Global trends and issues in value added taxation. *International Tax and Public Finance* 5, 399–428.

Grossman, G. 1980. Border tax adjustments: do they distort trade? *Journal of International Economics* 10, 117–28.

Huizanga, H. 2002. A European VAT on financial services? *Economic Policy* 35, 499– 534.

Keen, M. and Smith, S. 1996. The future of value-added tax in the European Union. *Economic Policy* 23, 375–420.

McLure, C. 2000. Implementing subnational value added taxes on internal trade: the compensating VAT (CVAT). *International Tax and Public Finance* 7, 723–40.

Meade, J.E. 1978. *The Structure and Reform of Direct Taxation*. London: Allen & Unwin.

How to cite this article

Metcalf, Gilbert E. "value-added tax." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000053>
doi:10.1057/9780230226203.1786

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

variance decomposition

Helmut Lütkepohl
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bayesian methods; bootstrap; Choleski decompositions; cointegrated variables; cointegration; factor analysis; forecasting; least squares; maximum likelihood; multivariate analysis; principal components; simultaneous equations models; structural vector autoregressions; variance decomposition; vector autoregressions

Article

Variance decomposition is a classical statistical method in multivariate analysis for uncovering simplifying structures in a large set of variables (for example, Anderson, 2003). For example, *factor analysis* or *principal components* are tools that are in widespread use. Factor analytic methods have, for instance, been used extensively in economic forecasting (see for example, Forni et al. 2000; Stock and Watson, 2002). In macroeconomic analysis the term ‘variance decomposition’ or, more precisely, ‘forecast error variance decomposition’ is used more narrowly for a specific tool for interpreting the relations between variables described by vector autoregressive (VAR) models. These models were advocated by Sims (1980) and used since then by many economists and econometricians as alternatives to classical simultaneous equations models. Sims criticized the way the latter models were specified, and questioned in particular the exogeneity assumptions common in simultaneous equations modelling. VAR models have the form

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{1}$$

where $y_t = (y_{1t}, \ldots, y_{Kt})'$ (the prime denotes the transpose) is a vector of *K* observed variables of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

interest, the A_i's are $(K \times K)$ parameter matrices, p is the lag order and u_t is a zero mean error process which is assumed to be white noise, that is, $E(u_t) = 0$, the covariance matrix, $E(u_t u_t') = \Sigma_u$, is time invariant and the u_t's are serially uncorrelated or independent. Here deterministic terms such as constants, seasonal dummies or polynomial trends are neglected because they are of no interest in the following. In the VAR model (1) all the variables are a priori endogenous. It is usually difficult to disentangle the relations between the variables directly from the coefficient matrices. Therefore it is useful to have special tools which help with the interpretation of VAR models. Forecast error variance decompositions are such tools. They are presented in the following.

An h steps ahead forecast or briefly h-step forecast at origin t can be obtained from (1) recursively for $h = 1, 2, \ldots$, as

$$y_{t+h|t} = A_1 y_{t+h-1|t} + \cdots + A_p y_{t+h-p|t}. \tag{2}$$

Here $y_{t+j|t} = y_{t+j}$ for $j \leq 0$. The forecast error turns out to be

$$y_{t+h} - y_{t+h|t} = u_{t+h} + \sum_{i=1}^{h-1} \Phi_i u_{t+h-i} \sim \left(0, \Sigma_h = \Sigma_u + \sum_{i=1}^{h-1} \Phi_i \Sigma_u \Phi_i'\right),$$

that is, the forecast errors have mean zero and covariance matrices Σ_h. Here the Φ_i's are the coefficient matrices of the power series expansion $(I_K - A_1 z - \cdots - A_p z^p)^{-1} = I_K + \Sigma_{i=1}^{\infty} \Phi_i z^i$. Note that the inverse exists in a neighbourhood of z=0 even if the VAR process contains integrated and cointegrated variables. (For an introductory exposition of forecasting VARs, see Lütkepohl, 2005.)

If the residual vector u_t can be decomposed in instantaneously uncorrelated innovations with economically meaningful interpretation, say, $u_t = B\varepsilon_t$ with $\varepsilon_t \sim (0, I_K)$, then $\Sigma_u = BB'$ and the forecast error variance can be written as $\Sigma_h = \Sigma_{i=0}^{h-1} \Theta_i \Theta_i'$, where $\Theta_0 = B$ and $\Theta_i = \Phi_i B, \quad i = 1, 2, \ldots$. Denoting the (n,m)th element of Θ_j by $\theta_{nm,j}$, the forecast error variance of the kth element of the forecast error vector is seen to be

$$\sigma_k^2(h) = \sum_{j=0}^{h-1} (\theta_{k1,j}^2 + \cdots + \theta_{kK,j}^2) = \sum_{j=1}^{K} (\theta_{kj,0}^2 + \cdots + \theta_{kj,h-1}^2).$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The term $(\theta_{kj,0}^2 + \cdots + \theta_{kj,h-1}^2)$ may be interpreted as the contribution of the *j*th innovation to the *h*-step forecast error variance of variable *k*. Dividing the term by $\sigma_k^2(h)$ gives the percentage contribution of innovation *j* to the *h*-step forecast error variance of variable *k*. This quantity is denoted by $\omega_{kj,h}$ in the following. The $\omega_{kj,h}$, $j = 1, \ldots, K$, decompose the *h*-step ahead forecast error variance of variable *k* in the contributions of the ε_t innovations. They were proposed by Sims (1980) and are often reported and interpreted for various forecast horizons.

For such an interpretation to make sense it is important to have economically meaningful innovations. In other words, a suitable transformation matrix *B* for the reduced form residuals has to be found. Clearly, *B* has to satisfy $\Sigma_u = BB'$. These relations do not uniquely determine *B*, however. Thus, restrictions from subject matter theory are needed to obtain a unique *B* matrix and, hence, unique innovations ε_t. A number of different possible sets of restrictions and approaches for specifying restrictions have been proposed in the literature in the framework of *structural VAR models*. A popular example is the choice of a lower-triangular matrix *B* obtained by a Choleski decomposition of Σ_u (for example, Sims, 1980). Such a choice amounts to setting up a system in recursive form where shocks in the first variable have potentially instantaneous effects also on all the other variables, shocks to the second variable can also affect the third to last variable instantaneously, and so on. In recursive systems it may be possible to associate the innovations with variables, that is, the *j*th component of ε_t is primarily viewed as a shock to the *j*th observed variable. Generally, the innovations can also be associated with unobserved variables, factors or forces and they may be named accordingly. For example, Blanchard and Quah (1989) consider a bivariate model for output and the unemployment rate, and they investigate effects of supply and demand shocks. Generally, if economically meaningful innovations can be found, forecast error variance decompositions provide information about the relative importance of different shocks for the variables described by the VAR model.

Estimation of reduced form and structural form parameters of VAR processes is usually done by least squares, maximum likelihood or Bayesian methods. Estimates of the forecast error variance components, $\omega_{kj,h}$, are then obtained from the VAR parameter estimates. Suppose the VAR coefficients are contained in a vector α, then $\omega_{kj,h}$ is a function of α, $\omega_{kj,h} = \omega_{kj,h}(\alpha)$. Denoting the estimator of α by $\hat{\alpha}$, $\omega_{kj,h}$ may be estimated as $\hat{\omega}_{kj,h} = \omega_{kj,h}(\hat{\alpha})$. If $\hat{\alpha}$ is asymptotically normal, that is, $\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\hat{\alpha}})$, then, under general conditions, $\hat{\omega}_{kj,h}$ is also asymptotically normally distributed, $\sqrt{T}(\hat{\omega}_{kj,h} - \omega_{kj,h}) \xrightarrow{d} \mathcal{N}(0, \sigma_{kj,h}^2 = \frac{\partial \omega_{kj,h}}{\partial \alpha'} \Sigma_{\hat{\alpha}} \frac{\partial \omega_{kj,h}}{\partial \alpha})$, provided the variance of the asymptotic distribution is non-zero. Here $\partial \omega_{kj,h} / \partial \alpha$ denotes the vector of first-order partial derivatives of $\omega_{kj,h}$ with respect to the elements of α (see Lütkepohl, 1990, for the specific form of the partial derivatives). Unfortunately, $\sigma_{kj,h}^2$ is zero even for cases of particular interest, for example, if $\omega_{kj,h} = 0$ and, hence, the *j*th innovation does not contribute to the *h*-step forecast error variance of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

variable k (see Lütkepohl, 2005, section 3.7.1, for a more detailed discussion). The problem can also not easily be solved by using bootstrap techniques (cf. Benkwitz, Lütkepohl and Neumann, 2000). Thus, standard statistical techniques such as setting up confidence intervals are problematic for the forecast error variance components. They can at best give rough indications of sampling uncertainty. The estimated $\omega_{kj,h}$'s are perhaps best viewed as descriptive statistics.

See Also

- impulse response function
- structural vector autoregressions
- vector autoregressions

Bibliography

Anderson, T. 2003. *An Introduction to Multivariate Statistical Analysis*, 3rd edn. New York: John Wiley.

Benkwitz, A., Lütkepohl, H. and Neumann, M. 2000. Problems related to bootstrapping impulse responses of autoregressive processes. *Econometric Reviews* 19, 69–103.

Blanchard, O. and Quah, D. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. 2000. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* 82, 540–52.

Lütkepohl, H. 1990. Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics* 72, 116–25.

Lütkepohl, H. 2005. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–48.

Stock, J. and Watson, M. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–79.

How to cite this article

Lütkepohl, Helmut. "variance decomposition." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000070>
doi:10.1057/9780230226203.1788

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

variance, analysis of

Andrew Gelman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Analysis of variance (ANOVA) is a statistical procedure for summarizing a classical linear model – a decomposition of sum of squares into a component for each source of variation in the model – along with an associated test (the *F*-test) of the hypothesis that any given source of variation in the model is zero. More generally, the variance decomposition in ANOVA can be extended to obtain inference for the variances of batches of parameters (sources of variation) in multilevel regressions. ANOVA is a useful addition to regression in that it structures inferences about batches of parameters.

Keywords

analysis of variance (ANOVA); balanced and unbalanced data; Bayesian inference; classical linear models; classical method of moments; contrast analysis; experimental economics; finite-population standard deviation; fixed effects and random effects; generalized linear models; linear models; linear regression; multilevel models; non-exchangeable models; probability; super-population standard deviation; variance decomposition

Article

1 Introduction

Analysis of variance (ANOVA) represents a set of models that can be fit to data, and also a set of methods for summarizing an existing fitted model. We first consider ANOVA as it applies to classical linear models (the context for which it was originally devised; Fisher, 1925) and then discuss how ANOVA has been extended to generalized linear models and multilevel models. Analysis of variance is particularly effective for analysing highly structured experimental data (in agriculture, multiple treatments applied to different batches of animals or crops; in psychology, multi-factorial experiments manipulating several independent experimental conditions and applied to groups of people; industrial experiments in which multiple factors can be altered at different times and in different locations).
At the end of this article, we compare ANOVA with simple linear regression.

2 Analysis of variance for classical linear models

2.1 ANOVA as a family of statistical methods

When formulated as a statistical model, analysis of variance refers to an additive decomposition of data into a grand mean, main effects, possible interactions and an error term. For example, Gawron et al. (2003) describe a flight-simulator experiment that we summarize as a 5×8 array of measurements under five treatment conditions and eight different airports. The corresponding two-way ANOVA model is $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$. The data as described here have no replication, and so the two-way interaction becomes part of the error term.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(If, for example, each treatment x airport condition were replicated three times, then the 120 data points could be modelled as $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, with two sets of main effects, a two-way interaction, and an error term.)

This is a linear model with $1 + 4 + 7$ coefficients, which is typically identified by constraining the $\Sigma_{i=1}^{5} \alpha_i = 0$ and $\Sigma_{j=1}^{8} \beta_j = 0$. The corresponding ANOVA display is shown in Table 1:

1. 1. For each source of variation, the degrees of freedom represent the number of effects at that level, minus the number of constraints (the five treatment effects sum to zero, the eight airport effects sum to zero, and each row and column of the 40 residuals sums to zero).
2. 2. The total sum of squares – that is, $\Sigma_{i=1}^{5} \Sigma_{j=1}^{8} (y_{ij} - \bar{y}_{..})^2$ – is $0.078 + 3.944 + 1.417$, which can be decomposed into these three terms corresponding to variance described by treatment, variance described by airport, and residuals.
3. 3. The mean square for each row is the sum of squares divided by degrees of freedom. Under the null hypothesis of zero row and column effects, their mean squares would, in expectation, simply equal the mean square of the residuals.
4. 4. The *F*-ratio for each row (except for the last) is the mean square, divided by the residual mean square. This ratio should be approximately 1 (in expectation) if the corresponding effects are zero; otherwise we would generally expect the *F*-ratio to exceed 1. We would expect the *F*-ratio to be less than 1 only in unusual models with negative within-group correlations (for example, if the data *y* have been renormalized in some way, and this had not been accounted for in the data analysis).
5. 5. The *p*-value gives the statistical significance of the *F*-ratio with reference to the F_{ν_1, ν_2}, where ν_1 and ν_2 are the numerator and denominator degrees of freedom, respectively. (Thus, the two *F*-ratios in Figure 1 are being compared to $F_{4,28}$ and $F_{7,28}$ distributions, respectively.) In this example, the treatment mean square is lower than expected (an *F*-ratio of less than 1), but the difference from 1 is not statistically significant (a *p*-value of 82 per cent), hence it is reasonable to judge this difference as explainable by chance, and consistent with zero treatment effects. The airport mean square is much higher than would be expected by chance, with an *F*-ratio that is highly statistically significantly larger than 1; hence we can confidently reject the hypothesis of zero airport effects.

Classical two-way analysis of variance for data on five treatments and eight airports with no replication

Source	Degrees of freedom	Sum of squares	Mean square	*F*-ratio	*p*-value
Treatment	4	0.078	0.020	0.39	0.816
Airport	7	3.944	0.563	11.13	< 0.001
Residual	28	1.417	0.051		

Note: The treatment-level variation is not statistically distinguishable from noise, but the airport effects are statistically significant.

Sources for all examples in this article: Gelman (2005) and Gelman and Hill (2006).

Figure 1

ANOVA display for two logistic regression models of the probability that a survey respondent prefers the Republican candidate for the 1988 US presidential election. *Notes*: Point estimates and error bars show median estimates, 50% intervals and 95% intervals of the standard deviation of each batch of coefficients. The large coefficients for ethnicity, region and state suggest that it might make sense to include interactions, hence the inclusion of ethnicity×region and ethnicity×state interactions in the second model. *Source*: data from seven CBS News polls.

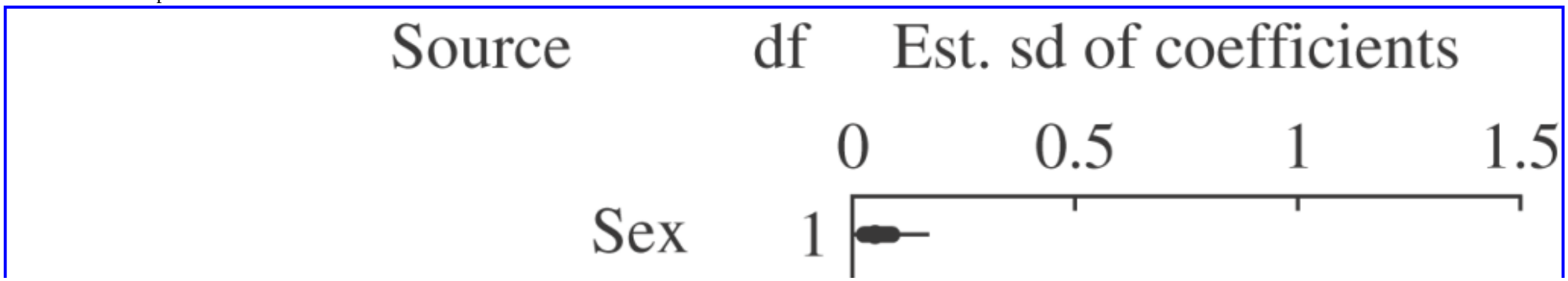

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

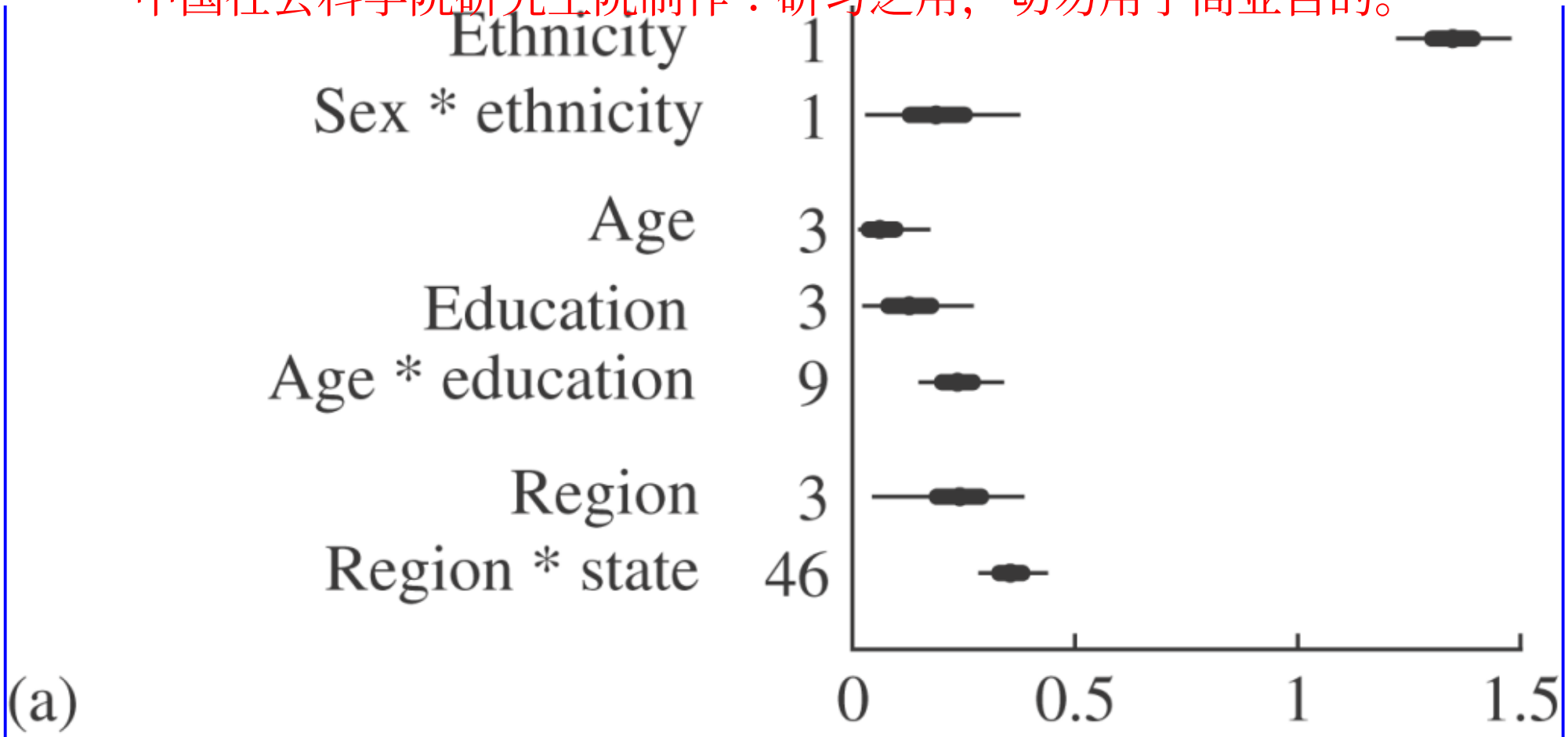

(a)

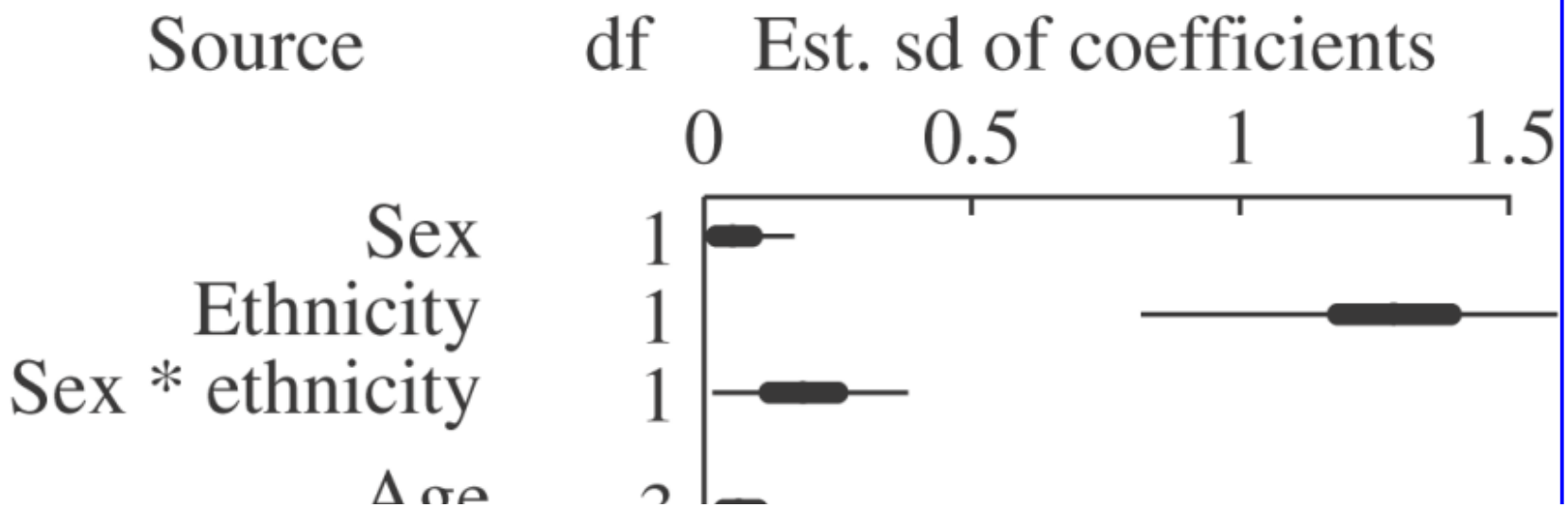

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

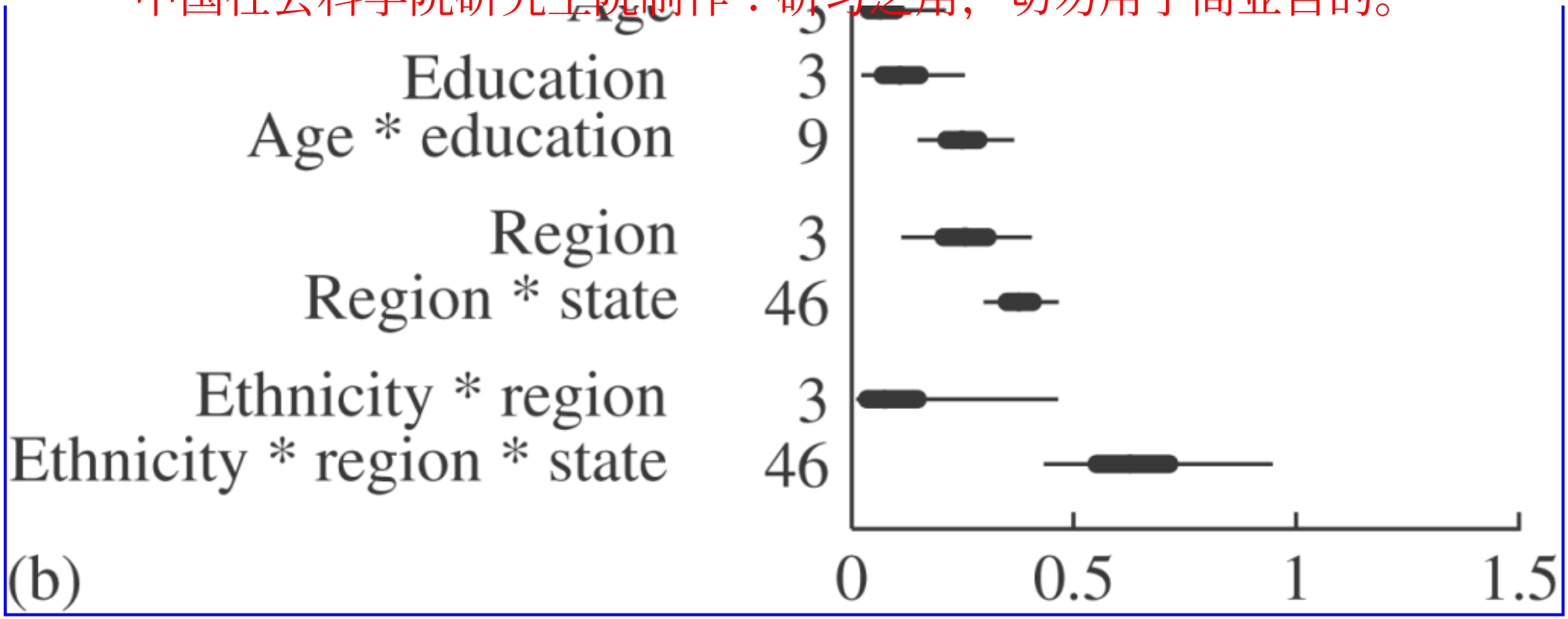

More complicated designs have correspondingly complicated ANOVA models, and complexities arise with multiple error terms. We do not intend to explain such hierarchical designs and analyses here, but we wish to alert the reader to such complications. Textbooks such as Snedecor and Cochran (1989) and Kirk (1995) provide examples of analysis of variance for a wide range of designs.

2.2 ANOVA to summarize a model that has already been fitted

We have just demonstrated ANOVA as a method of analysing highly structured data by decomposing variance into different sources, and comparing the explained variance at each level with what would be expected by chance alone. Any classical analysis of variance corresponds to a linear model (that is, a regression model, possibly with multiple error terms); conversely, ANOVA tools can be used to summarize an existing linear model.

The key is the idea of ‘sources of variation’, each of which corresponds to a batch of coefficients in a regression. Thus, with the model $Y = X\beta + \varepsilon$, the columns of X can often be batched in a reasonable way (for example, in Table 1, a constant term, four treatment indicators, and seven airport indicators) and the mean squares and F-tests then provide information about the amount of variance explained by each batch.

Such models could be fitted without any reference to ANOVA, but ANOVA tools could then be used to make some sense of the fitted models, and to test hypotheses about batches of coefficients.

2.3 Balanced and unbalanced data

In general, the amount of variance explained by a batch of predictors in a regression depends on which other variables have already been included in the model. With *balanced data*, however, in which all groups have the same number of observations (for example, each treatment applied exactly eight times, and each airport used for exactly five observations), the variance decomposition does not depend on the order in which the variables are entered. ANOVA is thus particularly easy to interpret with balanced data. The analysis of variance can also be applied to unbalanced data, but then the sums of squares, mean squares, and F-ratios will depend on the order in which the sources of variation are considered.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

3 ANOVA for more general models

Analysis of variance represents a way of summarizing regressions with large numbers of predictors that can be arranged in batches, and a way of testing hypotheses about batches of coefficients. Both these ideas can be applied in settings more general than linear models with balanced data.

3.1 F-tests

In a classical balanced design (as in the example in Table 1), each *F*-ratio compares a particular batch of effects to zero, testing the hypothesis that this particular source of variation is not necessary to fit the data.

More generally, the *F*-test can compare two nested models, testing the hypothesis that the smaller model fits the data adequately (so that the larger model is unnecessary). In a linear model, the *F*-ratio is $\frac{(SS_2 - SS_1)/(df_2 - df_1)}{SS_1/df_1}$, where SS_1, df_1 and SS_2, df_2 are the residual sums of squares and degrees of freedom from fitting the larger and smaller models, respectively.

For generalized linear models, formulas exist using the *deviance* (the log-likelihood multiplied by –2) that are asymptotically equivalent to *F*-ratios. In general, such models are not balanced, and the test for including another batch of coefficients depends on which other sources of variation have already been included in the model.

3.2 Inference for variance parameters

A different sort of generalization interprets the ANOVA display as inference about the variance of each batch of coefficients, which we can think of as the relative importance of each source of variation in predicting the data. Even in a classical balanced ANOVA, the sums of squares and mean squares do not exactly do this, but the information contained therein can be used to estimate the variance components (Cornfield and Tukey, 1956; Searle, Casella and McCulloch, 1992). Bayesian simulation can then be used to obtain confidence intervals for the variance parameters. As illustrated in this article we display inferences for standard deviations (rather than variances) because these are more directly interpretable. Compared with the classical ANOVA display, our plots emphasize the estimated variance parameters rather than testing the hypothesis that they are zero.

3.3 Generalized linear models

The idea of estimating variance parameters applies directly to generalized linear models as well as unbalanced data-sets. All that is needed is that the parameters of a regression model are batched into ‘sources of variation’. Figure 1 illustrates with a multilevel logistic regression model, predicting vote preference given a set of demographic and geographic variables.

3.4 Multilevel models and Bayesian inference

Analysis of variance is closely tied to multilevel (hierarchical) modelling, with each source of variation in the ANOVA table corresponding to a variance component in a multilevel model (see Gelman, 2005). In practice, this can mean that we perform ANOVA by fitting a multilevel model, or that we use ANOVA ideas to summarize multilevel inferences. Multilevel modelling is inherently Bayesian in that it involves a potentially large number of parameters that are modelled with probability distributions (see, for example, Goldstein, 1995; Kreft and De Leeuw, 1998; Snijders and Bosker, 1999). The differences between Bayesian and non-Bayesian multilevel models are typically minor except in settings with many sources of variation and little information on each, in which case some benefit can be gained from a fully Bayesian approach which models the variance parameters.

4 Related topics

4.1 Finite population and super-population variances

So far in this article we have considered, at each level (that is, each source of variation) of a model, the standard deviation of the corresponding set of coefficients. We call this the *finite-population* standard deviation. Another quantity of potential interest is the standard deviation of the hypothetical *super-population* from which these particular coefficients were drawn. The point estimates of these two variance parameters are similar – with the classical method of moments, the estimates are identical, because the super-population variance is the expected value of the finite-population variance – but they will have different uncertainties. The inferences for the finite-population standard deviations are more precise, as they

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

correspond to effects for which we actually have data.

Figure 2 illustrates the finite-population and super-population inferences at each level of the model for the flight-simulator example. We know much more about the five treatments and eight airports in our data-set than for the general populations of treatments and airports. (We similarly know more about the standard deviation of the 40 particular errors in out data-set than about their hypothetical super-population, but the differences here are not so large because the super-population distribution is fairly well estimated from the 28 degrees of freedom available from these data.)

Figure 2

Median estimates, 50% intervals and 95% intervals for (a) finite population and (b) super-population standard deviations of the treatment-level, airport-level and data-level errors in the flight-simulator example from Table 1. *Note*: The two sorts of standard deviation parameters have essentially the same estimates, but the finite-population quantities are estimated much more precisely. (We follow the general practice in statistical notation, using Greek and Roman letters for population and sample quantities, respectively.)

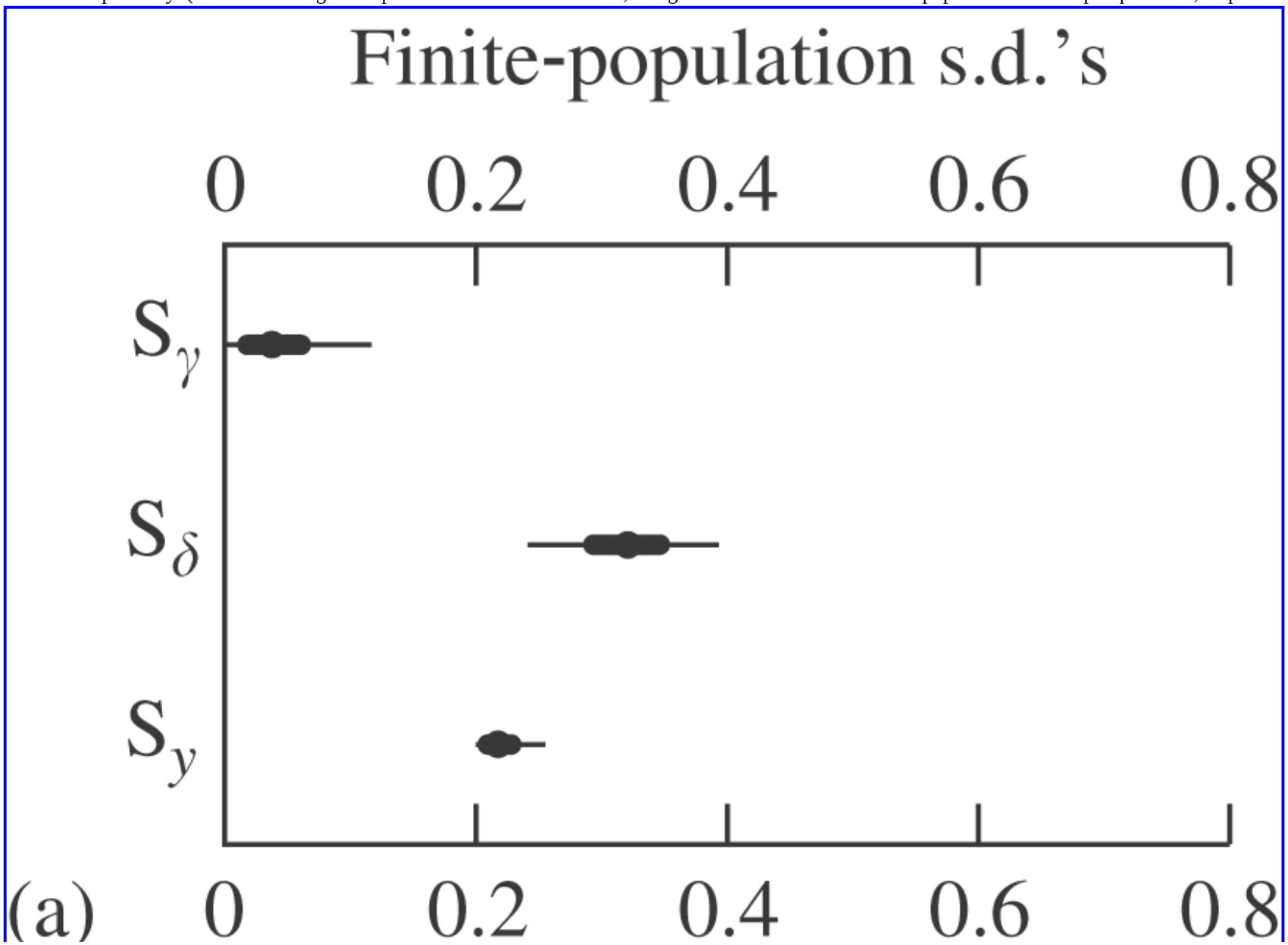

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

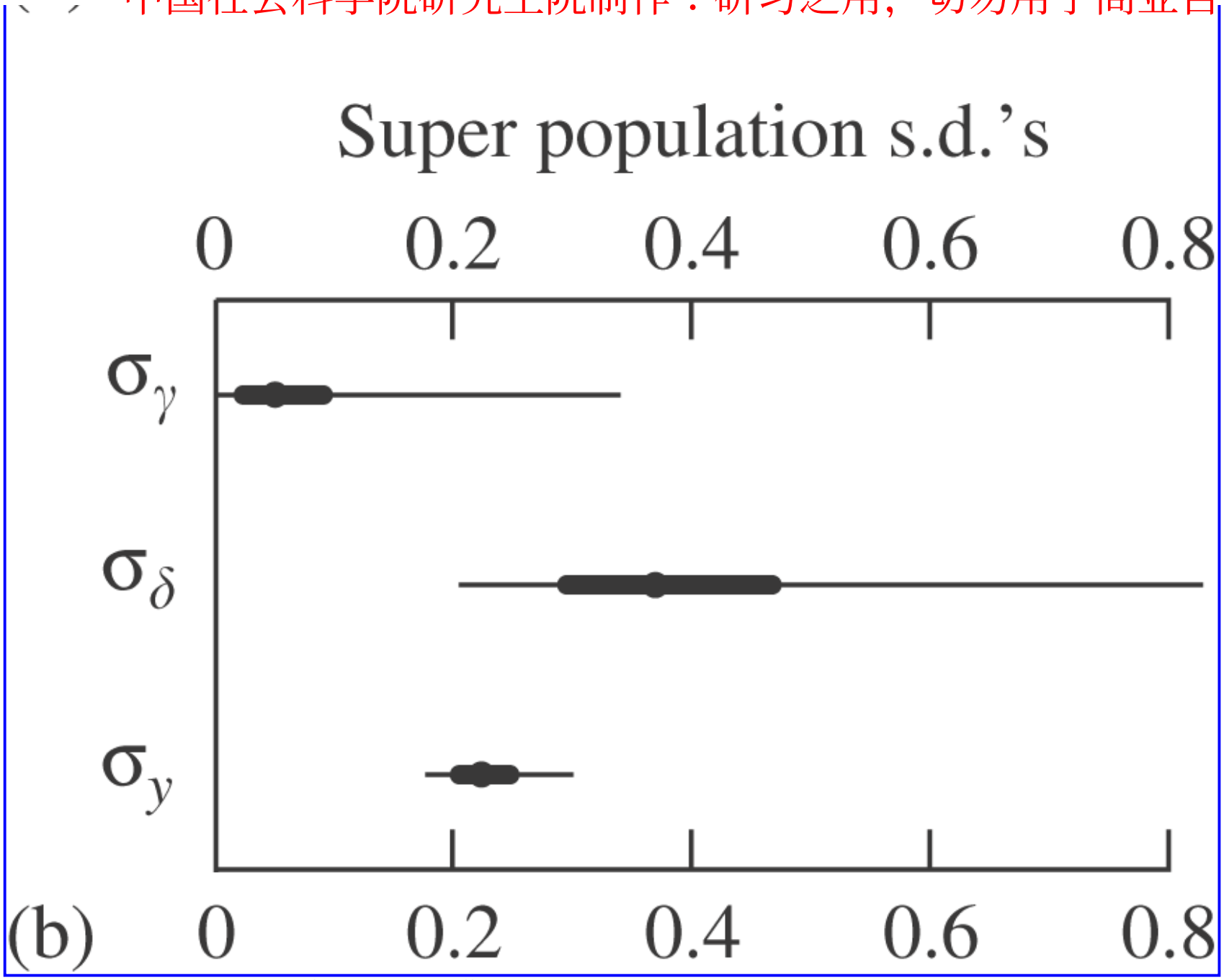

There has been much discussion about fixed and random effects in the statistical literature (see Eisenhart, 1947; Green and Tukey, 1960; Plackett, 1960; Yates, 1967; LaMotte, 1983; and Nelder, 1977; 1994, for a range of viewpoints), and unfortunately the terminology used in these discussions is incoherent (see Gelman, 2005, sec. 6). Our resolution to some of these difficulties is to always fit a multilevel model but to summarize it with the appropriate class of estimand – super-population or finite population – depending on the context of the problem. Sometimes we are interested in the particular groups at hand; at other times they are a sample from a larger population of interest. A change of focus should not require a change in the model, only a change in the inferential summaries.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

4.2 Contrast analysis

Contrasts are a way to structuring the effects within a source of variation. In a multilevel modelling context, a contrast is simply a group-level coefficient. Introducing contrasts into an ANOVA allows a further decomposition of variance. Figure 3 illustrates for a 5×5 latin square experiment (this time, not a split plot): the left plot in the figure shows the standard ANOVA, and the right plot shows a contrast analysis including linear trends for the row, column and treatment effects. The linear trends for the columns and treatments are large, explaining most of the variation at each of these levels, but there is no evidence for a linear trend in the row effects.

Figure 3

ANOVA displays for a 5×5 latin square experiment (an example of a crossed three-way structure): (a) with no group-level predictors, (b) contrast analysis including linear trends for rows, columns and treatments. *Note*: See also the plots of coefficient estimates and trends in Figure 4.

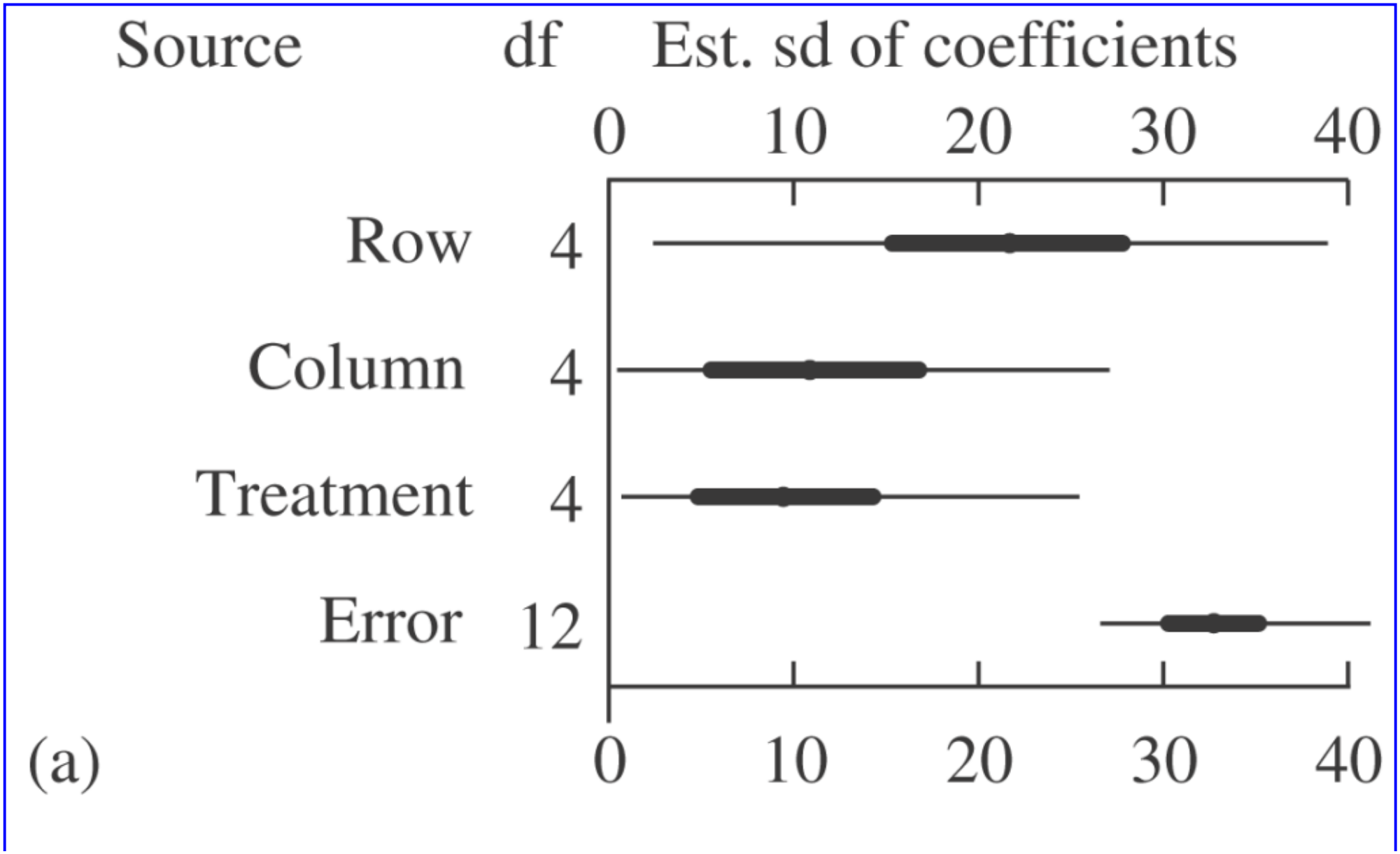

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

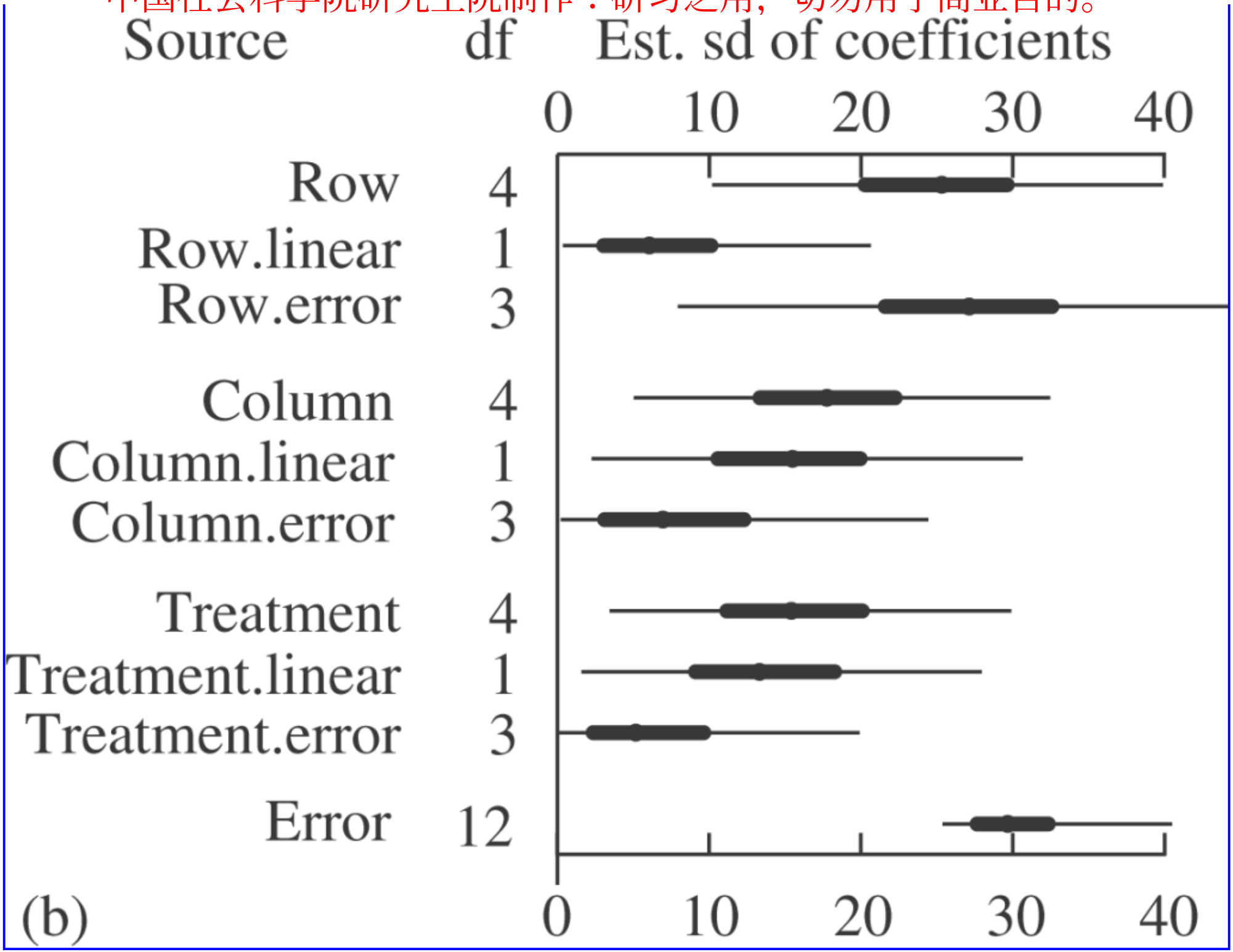

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Figure 4 shows the estimated effects and linear trends at each level (along with the raw data from the study), as estimated from a multilevel model. This plot shows in a different way that the variation among columns and treatments, but not among rows, is well explained by linear trends.

Figure 4

Estimates ±1 standard error for the row, column, and treatment effects for the latin square experiment summarized in Figure 3. *Note*: The five levels of each factor are ordered, and the lines display the estimated linear trends.

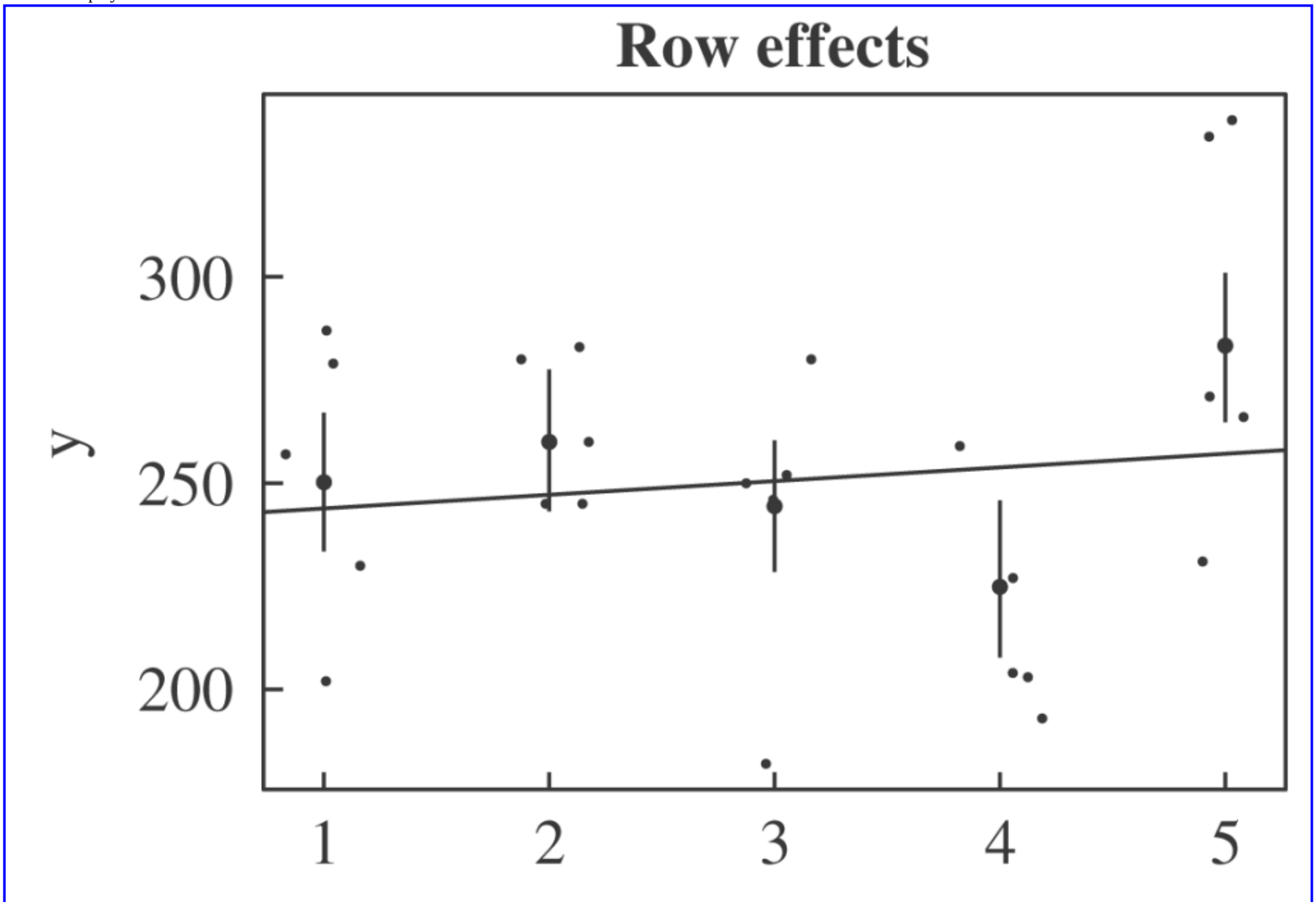

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

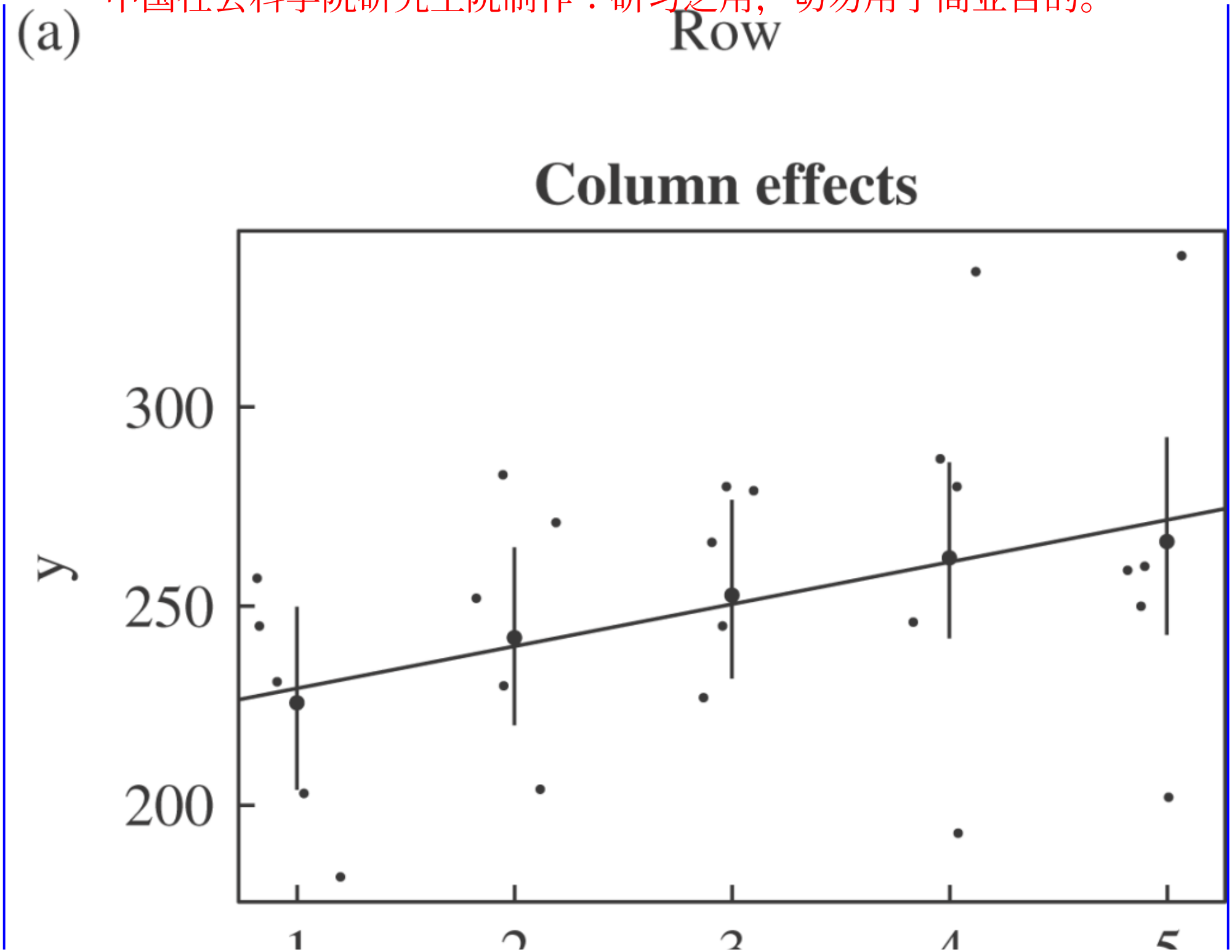

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

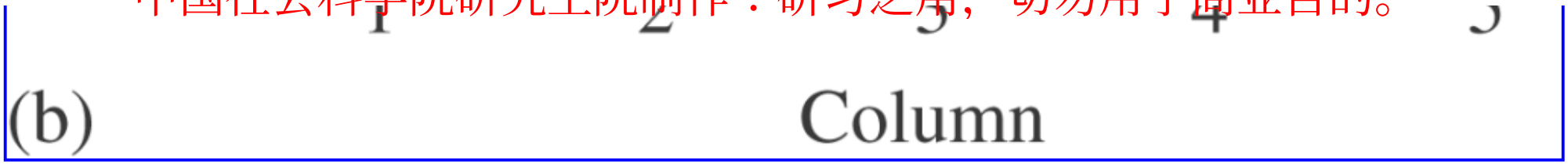

4.3 Non-exchangeable models

In all the ANOVA models we have discussed so far, the effects within any batch (source of variation) are modelled exchangeably, as a set of coefficients with mean 0 and some variance. An important direction of generalization is to non-exchangeable models, such as in time series, spatial structures (Besag and Higdon, 1999), correlations that arise in particular application areas such as genetics (McCullagh, 2005), and dependence in multi-way structures (Aldous, 1981; Hodges et al., 2005). In these settings, both the hypothesis-testing and variance-estimating extensions of ANOVA become more elaborate. The central idea of clustering effects into batches remains, however. In this sense, ‘analysis of variance’ represents all efforts to summarize the relative importance of different components of a complex model.

5 ANOVA compared with linear regression

The analysis of variance is often understood by economists in relation to linear regression (for example, Goldberger, 1964). From the perspective of linear (or generalized linear) models, we identify ANOVA with the structuring of coefficients into batches, with each batch corresponding to a ‘source of variation’ (in ANOVA terminology).

As discussed by Gelman (2005), the relevant inferences from ANOVA can be reproduced by using regression – but not always least-squares regression. Multilevel models are needed for analysing hierarchical data structures such as ‘split-plot designs’, where between-group effects are compared with group-level errors, and within-group effects are compared with data-level errors.

Given that we can already fit regression models, what do we gain by thinking about ANOVA? To start with, the display of the importance of different sources of variation is a helpful exploratory summary. For example, the two plots in Figure 1 allow us to quickly understand and compare two multilevel logistic regressions, without getting overwhelmed with dozens of coefficient estimates.

More generally, we think of the analysis of variance as a way of understanding and structuring multilevel models – not as an alternative to regression but as a tool for summarizing complex high-dimensional inferences, as can be seen, for example, in Figure 2 (finite-population and super-population standard deviations) and Figures 3 and 4 (group-level coefficients and trends).

See Also

- Bayesian statistics
- Fisher, Ronald Aylmer
- linear models
- two-stage least squares and the *k*-class estimator

We thank Jack Needleman, Matthew Rafferty, David Pattison, Marc Shivers, Gregor Gorjanc, and several anonymous commenters for helpful suggestions and the National Science Foundation for financial support.

Bibliography

Aldous, D. 1981. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 581–98.

Besag, J. and Higdon, D. 1999. Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society B*, 691–746.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Cochran, W. and Cox, G. 1957. *Experimental Designs*, 2nd edn. New York: Wiley.

Cornfield, J. and Tukey, J. 1956. Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 907–49.

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. *Biometrics* 3, 1–21.

Fisher, R.A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Gawron, V., Berman, B., Dismukes, R. and Peer, J. 2003. New airline pilots may not receive sufficient training to cope with airplane upsets. *Flight Safety Digest* (July–August), 19–32.

Gelman, A. 2005. Analysis of variance: why it is more important than ever (with discussion). *Annals of Statistics* 33, 1–53.

Gelman, A. and Hill, J. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gelman, A., Pasarica, C. and Dodhia, R. 2002. Let's practice what we preach: using graphs instead of tables. *American Statistician* 56, 121–30.

Goldberger, A. 1964. *Econometric Theory*. New York: Wiley.

Goldstein, H. 1995. *Multilevel Statistical Models*, 2nd edn. London: Edward Arnold.

Green, B. and Tukey, J. 1960. Complex analyses of variance: general problems. *Psychometrika* 25, 127–52.

Hodges, J., Cui, Y., Sargent, D. and Carlin, B. 2005. Smoothed ANOVA. Technical report, Department of Biostatistics, University of Minnesota.

Kirk, R. 1995. *Experimental Design: Procedures for the Behavioral Sciences*, 3rd edn. Pacific Grove, CA: Brooks/Cole.

Kreft, I. and De Leeuw, J. 1998. *Introducing Multilevel Modeling*. London: Sage.

LaMotte, L. 1983. Fixed-, random-, and mixed-effects models. In *Encyclopedia of Statistical Sciences*, ed. S. Kotz, N. Johnson and C. Read. New York: Wiley.

McCullagh, P. 2005. Discussion of Gelman (2005). *Annals of Statistics* 33, 33–8.

Nelder, J. 1977. A reformulation of linear models (with discussion). *Journal of the Royal Statistical Society A* 140, 48–76.

Nelder, J. 1994. The statistics of linear models: back to basics. *Statistics and Computing* 4, 221–34.

Plackett, R. 1960. Models in the analysis of variance (with discussion). *Journal of the Royal Statistical Society B* 22, 195–217.

Searle, S., Casella, G. and McCulloch, C. 1992. *Variance Components*. New York: Wiley.

Snedecor, G. and Cochran, W. 1989. *Statistical Methods*, 8th edn. Ames: Iowa State University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Snijders, T. and Bosker, R. 1999. *Multilevel Analysis*. London: Sage.

Yates, F. 1967. A fresh look at the basic principles of the design and analysis of experiments. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4, 777–90.

How to cite this article

Gelman, Andrew. "variance, analysis of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_A000098> doi:10.1057/9780230226203.1787

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

varying coefficient models

Andros Kourtellos and Thanasis Stengos
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Varying coefficient models offer a compromise between fully nonparametric and parametric models by allowing for the desired flexibility of the response coefficients of standard regression models to uncover hidden structures in the data without running into the serious curse of the dimensionality issue.

Keywords

functional coefficient models; heteroskedasticity; least squares; linear regression models; maximum likelihood; nonparametric estimation; parameter heterogeneity; random coefficients model; smooth coefficient models; tuning variables; varying coefficient models

Article

One of the most interesting forms of nonlinear regression models is the varying coefficient model (VCM). Unlike the linear regression model, VCMs were introduced by Hastie and Tibshirani (1993) to allow the regression coefficients to vary systematically and smoothly in more than one dimension. It is worth noting the distinction between the VCM and the so-called random coefficients model, which assumes that the coefficients vary non-systematically (randomly). Versions of the VCM are encountered in the literature as functional coefficient models (see Cai, Fan and Yao, 2000) and smooth coefficient models (see Li et al., 2002).

VCMs are very useful tools in applied work in economics as they can be used to model parameter heterogeneity in a general way. For example, Durlauf, Kourtellos and Minkin (2001) estimate a version of the Solow model that allows the parameters for each country to vary as functions of initial income. This work is extended in Kourtellos (2005), who finds parameter dependence on initial literacy, initial life expectancy, expropriation risk and ethnolinguistic fractionalization. Li et al. (2002) use the above smooth coefficient model to estimate the production function of the non-metal mineral industry in China. Stengos and Zacharias (2006) use the same model to examine an intertemporal hedonic model of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the personal computer market, where the coefficients of the hedonic regression were unknown functions of time. Hong and Lee (2003) forecast the nonlinearity in the conditional mean of exchange rate changes using a VCM that allows the autoregressive coefficients to vary with investment positions. Ahmad, Leelahanon and Li (2005) apply the VCM in the estimation of a production function in China's manufacturing industry to show that the marginal productivity of labour and capital depends on the firm's R&D values. Mamuneas, Savvides and Stengos (2006) study the effect of human capital on total factor productivity in an empirical growth framework. In what follows we present the basic structure of the standard VCM specification as it appears in the literature and then proceed to discuss certain aspects of estimation and some of its recent generalizations.

Basic specification

Consider the following VCM

$$y_i = \beta(z_i)' X_i + u_i \tag{1}$$

with $E(u_i|X_i) = 0$, where $X_i = (1, x_{i2}, \ldots, x_{ip})'$ is a p dimensional vector of slope regressors and $\beta(z_i)' = (\beta_1(z_{i1}), \beta_2(z_{i2}), \ldots, \beta_p(z_{ip}))$ is a p dimensional vector of varying coefficients, which take the form of unknown smooth functions of $z_{i1}, z_{i2}, \ldots, z_{ip}$, respectively. Notice that $\beta_1(z_i)$ is a varying intercept that measures the direct relationship between the tuning variable z_i and the dependent variable in a nonparametric way. We refer to the variables z_i's as tuning variables, and they can be one-dimensional or multidimensional. These functions map the tuning variables into a set of local regression coefficient estimates that imply that the effect of X_i on y_i will not be constant but rather it will vary smoothly with the tuning variables. These tuning variables could take the form of a scalar like time or it could be a vector of dimension q. A common situation in the literature arises when the z_j is the same for all j.

It is worth is noting that the VCM (1) is a very flexible and rich family of models. One of the reasons is that the general additive separable structure of (1) offers also a very useful compromise to the general high-dimensional nonparametric regression that is known to suffer from the curse of dimensionality. This allows for nonparametric estimation even when the conditioning regressor space is in high dimensional. Another is that it nests many well-known models as a special case. For instance, consider the following cases. If $\beta_j(z_{ij}) = \beta_j$, for all j then we are dealing with the usual linear model. If $\beta_j(z_{ij}) = \beta_j z_{ij}$ for some variable j, we simply have the interaction term $\beta_j x_{ij} z_{ij}$ entering the regression function. If $x_i = c$ (a constant) or if $z_{ij} = x_{ij}$ for all $j = 1, \ldots p$, then the model takes the generalized additive form where the additive components are unknown functions (see Hastie and Tibshirani, 1990;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Linton and Nielsen 1995).

We now set out some estimation issues. A popular estimation approach is based on local polynomial regression, as illustrated by Fan (1992), Fan and Gijbels (1996), and Fan and Zhang (1999), which we present in the context of a random sample design. Given a random sample $\{(z_i, X_i, Y_i)\}_{i=1}^{n}$, the estimation procedure solves a simple local least squares problem. To be precise, for each given point z_0 the functions $\beta_j(z)$, $j = 1 \ldots p$ are approximated by local linear polynomials $\beta_j(z) \approx c_{j0} + c_{j1}(z - z_0)$ for z in a neighborhood of z_0. This leads to the following weighted local least squares problem:

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\{c_{j0} + c_{j1}(z - z_0)\}X_{ij}\right]^2 K_h(z_i - z_0) \tag{2}$$

for a given kernel function K and bandwidth h, where $K_h(\cdot) = K(\cdot / h) / h$. While this method is simple, it is implicitly assumed that the functions $\beta_j(z)$ possess the same degrees of smoothness and hence can be approximated equally well in the same interval. Fan and Zhang (1999) allow for different degrees of smoothness for different coefficient functions by proposing a two-stage method. This is similar in spirit to what Huang and Shen (2004) do for global smoothers using regression splines but allowing each coefficient function to have different (global) smoothing parameters.

An attractive alternative to local polynomial estimation is a global smoothing method based on general series methods such as polynomial splines and trigonometric approximation (see Ahmad, Leelahanon and Li, 2005; Huang, Wu and Zhou, 2004; Huang and Shen, 2004; Xue and Yang, 2006a). All these papers emphasize the computational savings from having to solve only one minimization problem. Ahmad, Leelahanon and Li stress the efficiency gains of the series approach over a kernel-based approach when one allows for conditional heteroskedasticity. We should note that the inference for the estimated coefficients will differ for different choices of approximation, and the asymptotic properties of such estimators are generally not easy to obtain.

Although the model was initially developed for i.i.d. data, it has been extended for time series data by Chen and Tsay (1993), Cai, Fan, and Yao (2000), Huang and Shen (2004), and Cai (2007) for strictly stationary processes with different mixing conditions. The coefficient functions typically now become functions of time and/or lagged values of the dependent variable. It is worth noting that estimation issues such as bandwidth selection are similar, as in the i.i.d. data case (see Cai, 2007). The varying coefficient model has also been employed to analyse longitudinal data (see Brumback and Rice (1998), Hoover et al. (1998), and Huang, Wu and Zhou (2004).

The partially linear varying coefficient model

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

An interesting special case of eq. (1), where the unknown coefficient functions depend on a common z_i, is the partially linear VCM. Here some of the coefficients are constants (independent of z_i). In that case, eq. (1) can be rewritten as

$$y_i = \alpha' W_i + \beta(z_i)' X_i + u_i \quad (3)$$

where W_i is the ith observation on a $(1\times q)$ vector of additional regressors that enter the regression function linearly. The estimation of this model requires some special treatment as the partially linear structure may allow for efficiency gains, since the linear part can be estimated at a much faster rate, namely, $\sqrt{n}$.

The partially linear VCM has been studied by Zhang, Lee and Song (2002), Xia, Zhang and Tong (2004), Ahmad, Leelahanon and Li (2005), and Fan and Huang (2005). Zhang, Lee and Song (2002) suggest a two-step procedure where the coefficients of the linear part are estimated in the first step using polynomial fitting with an initial small bandwidth using cross validation (see Hoover et al., 1998). In other words, the approach is based on under-smoothing in the first stage. Then these estimates are averaged to yield the final first-step linear part estimates which are then used to redefine the dependent variable and return to the environment of eq. (1), where local smoothers can be applied as described above. Alternatively, Xia, Zhang and Tong (2004) separate the estimation of γ from that of $\beta(z_i)$ by noting that the former can be estimated globally, but the latter locally. This is what they call a 'semi-local least squares procedure', and they achieve a more efficient estimate of γ without under-smoothing using standard bandwidth selection methods. Once γ has been estimated, then again the linear part can be used to redefine the dependent variable and return to the environment of eq. (3).

More recently, Fan and Huang (2005) use a profile least squares estimation approach to provide a simple and useful method for (3). More precisely, they construct a Wald test and a profile likelihood ratio test for the parametric component that share similar sampling properties. More importantly, they show that the asymptotic distribution of the profile likelihood ratio test under the null is independent of nuisance parameters, and follows an asymptotic χ^2 distribution. They also propose a generalized likelihood ratio test statistic to test whether certain parametric functions fit the nonparametric varying coefficients. This hypothesis test includes testing for the significance of the slope variables X (zero coefficients) and the homogeneity of the model (constant coefficients). Other work on specification testing includes Li et al. (2002), Cai, Fan and Yao (2000), Cai (2007), Yang et al. (2006) that mainly rely on bootstrapping in their implementation.

Generalizations and extensions

A useful generalization of (1) is to allow the dependent variable to be related to the regression function

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

nonlinearly $m(X_i, Z_i) = \beta(z_i)' X_i$ via some given link function $g(\cdot)$

$$y_i = g(\beta(z_i)' X_i) + u_i \quad (4)$$

This generalization is known as the generalized varying coefficient model and was originally proposed by Hastie and Tibshirani (1993). Cai, Fan, and Li (2000) study this model using local polynomial techniques and propose an efficient one-step local maximum likelihood estimator. Notice that if $g(\cdot)$ is the normal CDF then (4) generalizes the standard tool of the discrete choice literature, namely the probit model.

Another strand of the literature allowed for a multivariate tuning variable z_l, $l = 1, 2, \ldots, q$. Although Hastie and Tibshirani (1993) proposed a back-fitting algorithm to estimate the varying coefficient functions, they did not provide any asymptotic justification. The most notable advance in this context has been by Xue and Yang (2006a), who propose a generalization of the VCM as in (1) that allows the varying coefficients to have an additive coefficient structure on regression coefficients to avoid the curse of dimensionality

$$\beta_j(z) = \gamma_{j0} + \gamma_{j1}(z_1) + \cdots + \gamma_{jq}(z_q) \text{ for all } j.$$

Under mixing conditions, Xue and Yang (2006a) propose local polynomial marginal integration estimators, while Xue and Yang (2006b) study this model using polynomial splines.

Finally, Cai et al. (2006) have shifted the discussion to consider a structural VCM. They examine the case of endogenous slope regressors, and propose a two-stage IV procedure based on local linear estimation procedures in both stages. We believe that this line of research is fruitful for economic applications.

Conclusion

VCMs have increasingly been employed as useful tools that allow for a compromise between fully nonparametric and parametric models. This compromise allows for the desired flexibility to uncover hidden structures that underlie the response coefficients of standard regression models without running into the serious curse of the dimensionality issue. More importantly, the structure of the VCM that allows the regression coefficients to vary with a tuning variable is very appealing in many economic applications, for it has a natural interpretation of non-constant marginal effects.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- economic growth nonlinearities
- nonparametric structural models

Bibliography

Ahmad, I., Leelahanon, S. and Li, Q. 2005. Efficient estimation of a semiparametric partially varying linear model. *Annals of Statistics* 33, 258–83.

Brumback, B. and Rice, J. 1998. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 93, 961–76.

Cai, Z. 2007. Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics* 136, 163–88.

Cai, Z., Fan, J. and Li, R. 2000. Efficient estimation and inferences for varying-coefficient models. *Journal of American Statistical Association* 95, 888–902.

Cai, Z., Fan, J. and Yao, Q. 2000. Functional coefficient regression models for nonlinear time series models. *Journal of American Statistical Association* 95, 941–56.

Cai, Z., Das, M., Xiong, H. and Wu, Z. 2006. Functional coefficient instrumental variables models. *Journal of Econometrics* 133, 207–41.

Chen, R. and Tsay, R. 1993. Functional coefficient autoregressive models. *Journal of the American Statistical Association* 88, 298–308.

Durlauf, S., Kourtellos, A. and Minkin, A. 2001. The local Solow growth model. *European Economic Review* 45, 928–40.

Fan, J. 1992. Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87, 998–1004.

Fan, J. and Gijbels, I. 1996. *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fan, J. and Huang, T. 2005. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* 11, 1031–57.

Fan, J. and Zhang, W. 1999. Statistical estimation in varying-coefficient models. *Annals of Statistics* 27,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1491–518.

Hastie, T. and Tibshirani, R. 1990. *Generalized Additive Models*. New York: Chapman and Hall.

Hastie, T. and Tibshirani, R. 1993. Varying coefficient models. *Journal of the Royal Statistical Society*, Series B 55, 757–796.

Hong, Y. and Lee, T.-H. 2003. Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* 85, 1048–62.

Hoover, D., Rice, C., Wu, C. and Yang, L. 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85, 809–22.

Huang, J. and Shen, H. 2004. Functional coefficient regression models for nonlinear time series: a polynomial spline approach. *Scandinavian Journal of Statistics* 31, 515–34.

Huang, J., Wu, C. and Zhou, L. 2004. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14, 763–88.

Kourtellos, A. 2005. Modeling parameter heterogeneity in cross-country growth regression models. Mimeo, Department of Economics, University of Cyprus.

Li, Q., Huang, C., Li, D. and Fu, T. 2002. Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics* 20, 412–22.

Linton, O. and Nielsen, J. 1995. A kernel method of estimating structural nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.

Mamuneas, T., Savvides, A. and Stengos, T. 2006. Economic development and the return to human capital: a smooth coefficient semiparametric approach. *Journal of Applied Econometrics* 21, 111–32.

Stengos, T. and Zacharias, E. 2006. Intertemporal pricing and price discrimination: a semiparametric hedonic analysis of the personal computer market. *Journal of Applied Econometrics* 21, 371–86.

Stone, C. 1977. Consistent nonparametric regression. *Annals of Statistics* 5, 595–620.

Xia, Y., Zhang, W. and Tong, H. 2004. Efficient estimation for semivarying-coefficient models. *Biometrika* 91, 661–81.

Xue, L. and Yang, L. 2006a. Estimation of semiparametric additive coefficient model. *Journal of Statistical Planning and Inference* 136, 2506–34.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Xue, L. and Yang, L. 2006b. Additive coefficient modeling via polynomial spline. *Statistica Sinica* 16, 1423–46.

Yang, L., Park, B., Xue, L. and Härdle, W. 2006. Estimation and testing for varying coefficients in additive models with marginal integration. *Journal of the American Statistical Association* 101, 1212–227.

Zhang, W., Lee, S.-Y. and Song, X. 2002. Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis* 82, 166–88.

How to cite this article

Kourtellos, Andros and Thanasis Stengos. "varying coefficient models." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000068> doi:10.1057/9780230226203.1789

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Veblen goods

B. Curtis Eaton
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

conspicuous consumption; leisure; leisure class; relative consumption trap; status and economics; upward-sloping demand curves; Veblen goods; Veblen, T.; well-being

Article

The utility that an individual derives from a Veblen good is an increasing function of the individual's consumption of the good *relative* to the consumption of others.

In *The Theory of the Leisure Class*, Thorstein Veblen observed that people value status, and further that in modern societies one's status is determined primarily by one's relative consumption of highly visible goods. ‘In order to gain and hold the esteem of men it is not sufficient merely to posses wealth … The wealth … must be put in evidence, for esteem is awarded only on evidence’ (1899, p. 36). The evidence consists of the conspicuous consumption of certain costly goods as prescribed by ‘the accredited cannons of [conspicuous] consumption, the effect of which is to hold the consumer up to a standard of expensiveness and wastefulness in his consumption of goods’ (1899, p. 116). Veblen was certainly not the first person to articulate the view that esteem can be achieved by conspicuous displays of wealth, but he saw more clearly than others the futility and wastefulness of this form of status seeking.

Following Leibenstein (1950), much of the literature on Veblen goods has focused on the possibility that the demand curve might be upward sloping. The inefficiency or wastefulness associated with Veblen goods is perhaps a more serious matter – see Hopkins and Kornienko (2004) for a theoretical analysis.

Veblen seems to have thought that beyond some modest level of affluence societies get caught in what might be called the *relative consumption trap* in which all added productivity is soaked up by the wasteful consumption of Veblen goods with no effect on well-being: ‘The need of conspicuous waste … stands ready to absorb any increase in the community's industrial efficiency or output of goods, after the most elementary physical wants have been provided for’ (1899, p. 110).

The recent literature on perceived well-being suggests that affluent societies may in fact be caught in this trap. A number of studies have shown that the correlation of average well-being and per capita income in affluent societies is very weak, in some cases non-existent. Much of the evidence is surveyed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

in Robert Frank's (1999) provocative book, *Luxury Fever*. Others have shown that an individual's well-being is negatively associated with the incomes of one's neighbours, and further that the effects on one's well-being of an increase in one's own income and an increase of the same magnitude in the average income of one's neighbours are approximately offsetting (see Luttmer, 2005, for example).

With the aid of a simple representative agent model, we can readily see how affluent societies can get stuck in the relative consumption trap. There is a continuum of agents, all of whom have identical preferences and budgets. Preferences of a representative person are captured in the following utility function:

$$U_r(x_r, y_r, v_r) = D(v_r - v) + F(x_r) + G(y_r),$$

where v_r, x_r and y_r are, respectively, quantities of a *pure Veblen good*, leisure, and a standard consumption good, and v is average consumption of the Veblen good. The Veblen good is *pure* in the sense that the utility derived from it, $D(v_r - v)$, is dependent only on relative consumption, $v_r - v$. The functions D, F and G are strictly increasing and concave. Leisure and the standard good are essential, but the Veblen good is not ($D'(0)$ is finite). Each individual is endowed with 1 unit of time to be allocated to leisure and work, and with asset income a. The wage rate is w, and the prices of the Veblen and standard goods are both 1.

For an interior solution to the individual's choice problem, the following marginal conditions must hold:

$$\frac{F'(x_r)}{w} = G'(y_r) = D'(v_r - v).$$

In addition the budget constraint, $wx_r + y_r + v_r = w + a$, will be satisfied.

Since everyone is identical, in equilibrium $v_r = v$, so the conditions that characterize an interior equilibrium are

$$\frac{F'(x^*)}{w} = G'(y^*) = D'(0),$$

and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$wx^* + y^* + v^* = w + a.$$

Notice that, in equilibrium, the marginal value of the Veblen good, $D'(0)$, is independent of w and a, and since $G'(y^*) = D'(0)$, so too is the equilibrium quantity of the standard good.

What happens as a increases? Clearly, y^* doesn't change, and neither does x^*, since $F'(x^*)/w = G'(y^*)$ and w hasn't changed. So all of the added purchasing power is devoted to the Veblen good, and, since no one's relative consumption of the Veblen good has changed, there will be no change in equilibrium utility.

What happens as w increases? As in the first scenario, y^* doesn't change, but w having increased, x^* must decrease to satisfy $F'(x^*)/w = G'(y^*)$. But this implies that the increase in expenditure on the Veblen good (dv^*) exceeds the increase in full income (dw^*), so in this case *more* than all of the added purchasing power is soaked up by the Veblen good. In addition, since neither y^* nor equilibrium relative consumption of the Veblen good changes, and x^* decreases, equilibrium utility decreases.

So, in this model, if the equilibrium is interior, then

$$\frac{dy^*}{da} = 0,\ \frac{dx^*}{da} = 0,\ \frac{dv^*}{da} = 1,\ \frac{du^*}{da} = 0,\ \frac{dy^*}{dw} = 0,\ \frac{dx^*}{dw} < 0,\ \frac{dv^*}{dw} > 1,\ \frac{du^*}{dw} < 0.$$

Of course, the equilibrium is not necessarily interior. In particular, since $D'(0)$ is finite, unless the society is sufficiently affluent, in equilibrium nothing is spent on the Veblen good ($v^*=0$). But once the society is affluent enough so that it begins to squander its resources on the wasteful Veblen good, it is stuck in the relative consumption trap.

See Also

- conspicuous consumption
- consumption externalities
- happiness, economics of
- leisure class
- social status, economics and
- Veblen, Thorstein Bunde

Bibliography

Frank, R. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess*. New York: Free Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Hopkins, E. and Kornienko, T. 2004. Running to keep in the same place: consumer choice as a game of status. *American Economic Review* 94, 1085–107.

Leibenstein, H. 1950. Bandwagon, snob, and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64, 183–207.

Luttmer, E.F.P. 2005. Neighbors as negatives: relative earnings and well-being. *Quarterly Journal of Economics* 120, 963–1002.

Veblen, T. 1899. *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: The Modern Library, 1934.

How to cite this article

Eaton, B. Curtis. "Veblen goods." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000076> doi:10.1057/9780230226203.1790

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Veblen, Thorstein Bunde (1857 – 1929)

Geoffrey M. Hodgson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article outlines the work of the American institutional economist Thorstein Veblen (1857–1929), stressing his critique of neoclassical economics and his development of an alternative, evolutionary approach to the analysis of social, economic and technological change. Veblen's analytical approach to both technology and institutions is discussed, as well as his explicit application of the evolutionary ideas from Darwinian biology to economics.

Keywords

Ayres, C. E.; Clark, J. M.; Commons, J. R.; conspicuous consumption; consumer sovereignty; customs; Darwin, C.; economic man; evolutionary economics; explanation; habit; hedonism; innovation; institutional economics; James, W.; Mark, K. H.; McDougall, W.; Mitchell, W. C.; natural selection; neoclassical economics; old institutionalism; Peirce, C. S.; pragmatism; preferences; Spencer, H.; Sumner, W. G.; technical change; utililitarianism; Veblen, T. B.

Article

Thorstein Veblen was one of the most influential economists of the early 20th century and one of the founders of the American school of institutional economics.
Veblen was the fourth son and sixth child of Norwegian immigrants who settled in eastern Minnesota in United States. Educated at Carleton College, Johns Hopkins University, Yale University and Cornell University, he took various university posts at Chicago, Stanford, Missouri and New York. At Johns Hopkins he came in contact with the pragmatist philosopher Charles Sanders Peirce, and at Yale he was influenced by William Graham Sumner. Veblen read widely in biology, psychology and philosophy, as well as the social sciences. The works of Immanuel Kant, Charles Darwin, William James, Karl Marx, William McDougall and Herbert Spencer also made an enduring mark. Despite the popularity of his ideas, Veblen's career was marred by scandal and he never held a senior academic post (Jorgensen and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Jorgensen, 1999). He died in meagre circumstances in California.

Several of his most important theoretical works date from the 1890s, when he was at the University of Chicago. In 1898 he published his classic article ‘Why is Economics Not an Evolutionary Science?’ in the *Quarterly Journal of Economics*. The following year saw the appearance of his first book *The Theory of the Leisure Class*. Although this is an original and sophisticated theoretical work, its mockery of the wasteful rich turned it into a bestseller. Other academic articles followed in the *Quarterly Journal of Economics*, the *Journal of Political Economy* and elsewhere, the most important of which have been collected in *The Place of Science in Modern Civilization and Other Essays* (1919). These articles provided a critique of ‘neoclassical’ economics (a term he coined to refer to equilibrium-oriented approaches involving individual utility maximization) and suggestions of a new approach to economics on ‘evolutionary’ and ‘Darwinian’ lines.

He is remembered today as the founder of the school of ‘institutional economics’ that prospered in the United States between the first and second world wars. This school involved leading American economists such as John Maurice Clark, John Rogers Commons and Veblen's student, Wesley Clair Mitchell.

However, Veblen and his followers did not construct an integrated system of economic theory. This is partly because the original foundations of Veblenian institutionalism were challenged. Pragmatist philosophy, instinct-habit psychology and evolutionary ideas had been foundational for Veblen's thought. However, by the 1920s they had lost much of their former popularity. Thus, at the high point of its influence, American institutionalism faced fundamental philosophical and theoretical difficulties. After 1940, the ‘old’ institutional economics lost ground to the rising generation of formal and mathematically inclined theorists. By the 1960s the American institutional school was confined to a small minority of adherents. However, in economics in recent years there has been a revival of interest in both evolutionary ideas and the legacy of the ‘old’ institutional school.

Veblen (1919, p. 73) argued that neoclassical economics adopted a faulty and ‘hedonistic’ psychology, involving ‘a passive and substantially inert and immutably given human nature’. He criticized the idea of the individual as a given ‘globule of desire’, lambasting the neoclassical picture of the optimizing and omniscient economic agent as ‘a lightning calculator of pleasures and pains’. He saw this ‘economic man’ as having ‘neither antecedent nor consequent’, lacking an account of how human wants are formed and portraying humans as utility-maximizing automata. Veblen (1914) proposed an alternative theory of human agency, in which ‘instincts’ such as ‘workmanship’, ‘emulation’, ‘predatoriness’ and ‘idle curiosity’ play a major role. Habit and instinct replaced the utilitarian pleasure-pain principle.

Following the pragmatist philosophy of Peirce and James, Veblen rejected the Cartesian notion of the supremely rational and calculating individual, instead seeing agents as propelled in the main by habits and customs. Habits of thought provide the point of view from which facts and events are interpreted. When they are shared and reinforced within a society or group, individual habits assume the form of socio-economic institutions. In turn, institutions create and reinforce habits of action and thought: ‘The situation of today shapes the institutions of tomorrow through a selective, coercive process, by acting upon men's habitual view of things, and so altering or fortifying a point of view or a mental attitude handed down from the past’ (Veblen, 1899, pp. 190–1).

In *The Theory of the Leisure Class* and elsewhere, he argued that consumption is a ‘conspicuous’ and social process. Through consumption, humans signal status and social position, and thereby stimulate the desires of others. Accordingly, individual tastes are malleable and the idea of unalloyed ‘consumer

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
sovereignty' is a myth.

Veblen saw conventions, customs and institutions as repositories of social knowledge. Institutional adaptations and behavioural norms were stored in individual habits and could be passed on by education or imitation to succeeding generations. His explanations of economic growth privileged knowledge and institutions, rather than the accumulation of physical assets.

Veblen addressed the 'evolutionary' processes of innovation and transformation in a modern economy. Neoclassical theory is defective in this respect because it indicated 'the conditions of survival to which any innovation is subject, supposing the innovation to have taken place, not the conditions of variational growth' (Veblen, 1919, pp. 176–7). He saw it as important to consider why innovations take place, and not merely to dwell over equilibrium conditions with given technological possibilities. The question for him was not how things stabilize themselves in a 'static state', but how they endlessly grow and change.

Veblen saw Darwinian evolutionary principles as crucial to the understanding of the processes of institutional and technological development in a capitalist economy. He was the first economist to argue at length that Darwinian evolutionary principles should be applied to economics. He upheld that economics should become an 'evolutionary' and 'post-Darwinian' science. There is a current revival in 'evolutionary' approaches in economics but the Veblenian precedent for this type of approach is not always acknowledged.

Darwinian evolution involves three essential features. First, there must be sustained variation among the members of a species or population. Variations may be random or purposive in their origin, but without them, as Darwin insisted, natural selection cannot operate. Second, there must be some principle of heredity or continuity involving some mechanism through which individual characteristics are passed on to succeeding generations. Third, natural selection operates because better-adapted organisms leave increased numbers of offspring, or because the variations that are preserved bestow advantage in the struggle for survival.

Veblen applied the same three Darwinian principles to economic evolution. He recognized the role of creativity and novelty with his concept of 'idle curiosity'. Habits and institutions were regarded as relatively durable heritable traits. Concerning selection, Veblen (1899, p. 188) wrote: 'The life of man in society, just as the life of other species, is a struggle for existence, and therefore it is a process of selective adaptation. The evolution of social structure has been a process of natural selection of institutions.' This did not mean that social phenomena were to be explained wholly or largely in biological terms, but that Darwinian principles could be applied to social and economic units and processes.

Veblen saw Darwinian evolutionary processes as open-ended and suboptimal. Unlike advocates of laissez faire, he did not use Darwinian principles to justify market competition. He was critical of apologetic tendencies in social science which regard existing institutions as necessarily efficient or optimal. He described particularly regressive or disserviceable institutions as 'archaic', 'ceremonial' or even 'imbecile'. Furthermore, he used Darwinian ideas to rebut of Marx's teleological suggestions that history was leading inevitably to a communist future.

Like Darwin, Veblen emphasized the importance of processual, causal explanation. Although he did not use the word, he had an appreciation of Darwinian evolution as an 'algorithmic' process. Veblen used phrases such as 'cumulative causation', 'theory of a process, of an unfolding sequence' and 'impersonal sequence of cause and effect' to connote the same idea. This focus on algorithmic processes is revolutionary and modern; it directs attention to ongoing processes rather than static equilibria alone.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Consequently, rather than taking individual reasons or preferences as themselves sufficient to understand motivations, Veblen pointed to the need for causal explanations of reasons or preferences themselves. He did not underestimate the importance of human intentionality – but it had to be explained rather than assumed. Such explanations involved the evolution of social institutions and their interplay with biological and psychological characteristics. He thus acknowledged processes of dual inheritance or coevolution (again to use modern terms) where there was evolution and transmission at both the instinctive and the cultural levels.

Along with the assumption of fixed preference functions, Veblen also criticized the widespread assumption in economic theory of a fixed set of technological possibilities. Technological change can challenge established institutions and vested interests. In *The Theory of Business Enterprise* and elsewhere Veblen distinguished between industry (making goods) and business (making money). This dichotomy parallels the earlier suggestion in *The Theory of the Leisure Class* that there is a distinction between serviceable consumption to satisfy human need and conspicuous consumption for status and display. Subsequently, institutionalists such as Clarence E. Ayres elevated the different conflict between technology and institutions into a universal principle, and dubbed it the ‘Veblenian dichotomy’. This is misleading, because Veblen never saw such a conflict as universal, and he saw institutions as the indispensable fabric of economic life (Hodgson, 2004).

In the last two decades of the 20th century, evolutionary and institutional ideas again become prominent in economics. Pragmatism has again become fashionable in philosophy and the concept of habit has returned to psychology. Many of Veblen's ideas, including those on institutional evolution and the role of knowledge in economic growth, now seem strikingly modern. The conditions exist for a deeper appreciation of his contribution to economics and social science.

See Also

- Ayres, Clarence Edwin
- Clark, John Maurice
- Commons, John Rogers
- evolutionary economics
- institutionalism, old
- Mitchell, Wesley Clair
- ‘neoclassical’

Selected works

1898. Why is economics not an evolutionary science? *Quarterly Journal of Economics* 12, 373–97. Reprinted in Veblen (1919).

1899. *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: Macmillan.

1904. *The Theory of Business Enterprise*. New York: Charles Scribners.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1915. *Imperial Germany and the Industrial Revolution*. New York: Macmillan.

1914. *The Instinct of Workmanship, and the State of the Industrial Arts*. New York: Macmillan.

1918. *The Higher Learning in America: A Memorandum on the Conduct of Universities by Business Men*. New York: Huebsch.

1919. *The Place of Science in Modern Civilization and Other Essays*. New York: Huebsch.

1921. *The Engineers and the Price System*. New York: Harcourt Brace and World.

1923. *Absentee Ownership and Business Enterprise in Recent Times*. New York: Huebsch.

1934. *Essays on Our Changing Order*. New York: Viking Press.

Bibliography

Daugert, S. 1950. *The Philosophy of Thorstein Veblen*. New York: Columbia University Press.

Dorfman, J. 1934. *Thorstein Veblen and His America*. New York: Viking Press.

Hodgson, G. 2004. *The Evolution of Institutional Economics: Agency, Structure and Darwinism in American Institutionalism*. London and New York: Routledge.

Jorgensen, E. and Jorgensen, H. 1999. *Thorstein Veblen: Victorian Firebrand*. Armonk, NY: M. E. Sharpe.

Rutherford, M. 1994. *Institutions in Economics: The Old and the New Institutionalism*. Cambridge: Cambridge University Press.

Tilman, R. 1996. *The Intellectual Legacy of Thorstein Veblen: Unresolved Issues*. Westport, CT: Greenwood Press.

How to cite this article

Hodgson, Geoffrey M. "Veblen, Thorstein Bunde (1857–1929)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000014> doi:10.1057/9780230226203.1791

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

vector autoregressions

Tao Zha
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Vector autoregressions are a class of dynamic multivariate models introduced by Sims (1980) to macroeconomics. These models have been primarily used to bring empirical regularities out of the time series data, to provide forecasting and policy analysis, and to serve as a benchmark for model comparison. Economic applications often impose more restrictions on vector autoregressions than originally thought necessary. Recent econometric developments have made it feasible to handle vector autoregressions with a wide class of restrictions and have narrowed the gap between these models and dynamic stochastic general equilibrium models.

Keywords

Bayesian econometrics; Bayesian priors; Cowles Commission; dynamic multivariate models; dynamic stochastic general equilibrium models; Gibbs samplers; identification; impulse responses; likelihood function; marginal data density; Markov chain Monte Carlo method; Markov-switching vector autoregressions; probability density function; recursive identification; structural shocks; structural vector autoregressions; vector autoregressions

Article

Vector autoregressions (VARs) are a class of dynamic multivariate models introduced by Sims (1980) to macroeconomics. These models arise mainly as a response to the ‘incredible’ identifying assumptions embedded in traditional large-scale econometric models of the Cowles Commission. The traditional approach uses predetermined or exogenous variables, coupled with many strong exclusion restrictions, to identify each structural equation. VARs, by contrast, explicitly recognize that all economic variables are interdependent and thus should be treated endogenously. The philosophy of VAR modelling begins with a multivariate time series model that has minimal restrictions, and gradually introduces identifying information, with emphasis always placed on the model's fit to data.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
While the traditional econometric approach allows disturbances or shocks to structural equations to be correlated, the VAR methodology insists that structural shocks ought to be independent of one another. The independence assumption plays an essential role in achieving unambiguous economic interpretations about structural shocks such as technology and policy shocks; it can be tested using recently developed econometric tools (Leeper and Zha, 2003). The bulk of VAR work has focused on identifying structural shocks as a way to specify the contemporaneous relationships among economic variables. With most dynamic relationships unrestricted, the intent of such an identifying strategy is to construct models that have both economic interpretability and superior fit to data. Dynamic responses to a particular shock, called impulse responses, are often used as economic interpretations to the model. They summarize the properties of all systematic components of the system and have become a major tool in modern economic analysis.

Modelling policy shocks explicitly is important in addressing the practical importance of the Lucas critique. If policy switches regime, such a change may be viewed as a sequence of random shocks from the public's viewpoint (Sims, 1982). If this sequence displays a persistent pattern, the public will adjust its expectations formation accordingly and the Lucas critique may be consequential. For the practice of monetary policy, however, it is an empirical question how significant this adjustment is. Leeper and Zha (2003) construct an econometric measure from the sequence of policy shocks implied by regime switches to gauge whether the public's behaviour could be well approximated by a linear model. This measure is particularly useful if counterfactual exercises regarding the effects of policy changes are conducted with respect to the Lucas critique.

VARs have also been used for other tasks. Armed with a Bayesian prior, VARs have been known to produce out-of-sample forecasts of economic variables as well as, or even better than, those from commercial forecasting firms (Litterman, 1986; Geweke and Whiteman, 2006). Because of their ability to forecast, VARs have given researchers a convenient diagnostic tool to assess the feasibility or plausibility of real-time policy projections of other economic models (Sims, 1982). VARs have been increasingly used for policy analysis and as a benchmark for comparing different dynamic stochastic general equilibrium (DSGE) models. Restrictions on lagged coefficients have been gradually introduced to give more economic interpretations to individual equations. All these developments are positive and help narrow the gap between statistical and economic models.

This article discusses these and other aspects of VARs, summarizes some key theoretical results for the reader to consult without searching for different sources, and provides a perspective on where future research in this area will be headed.

General framework

Structural form

VARs are generally represented in a structural form of which the reduced form is simply a byproduct. The general form is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\mathbf{y}_t'\mathbf{A} = \sum_{l=1}^{p} \mathbf{y}_{t-l}'\mathbf{A}_l + \mathbf{z}_t'\mathbf{D} + \varepsilon_t', \tag{1}$$

where $\mathbf{y}_t$ is an $n \times 1$ column vector of endogenous variables, $\mathbf{A}$ and $\mathbf{A}_l$ are $n \times n$ parameter matrices, $\mathbf{z}_t$ is an $h \times 1$ column vector of exogenous variables, **D** is an $h \times n$ parameter matrix, p is the lag length, and ε_t is an $n \times 1$ column vector of structural shocks. The parameters of individual equations in (1) correspond to the columns of **A**, $\mathbf{A}_l$, and **D**. The structural shocks are assumed to be i.i.d. and independent of one another:

$$E(\varepsilon_t|\mathbf{y}_{t-s}, s > 0) = \underset{n\times 1}{\mathbf{0}}, \quad E(\varepsilon_t\varepsilon_t'|\mathbf{y}_{t-s}, s > 0) = \underset{n\times n}{\mathbf{I}},$$

where $\underset{n\times n}{\mathbf{0}}$ is the $n \times n$ matrix of zeros and $\underset{n\times n}{\mathbf{I}}$ is the $n \times n$ identity matrix. It follows that the reduced form of (1) is

$$\mathbf{y}_t' = \sum_{l=1}^{p} \mathbf{y}_{t-l}'\mathbf{B}_l + \mathbf{z}_t'\mathbf{C} + u_t', \tag{2}$$

where $\mathbf{B}_l = \mathbf{A}_l\mathbf{A}^{-1}$, $\mathbf{C} = \mathbf{D}\mathbf{A}^{-1}$, and $u_t' = \varepsilon_t'\mathbf{A}^{-1}$. The covariance matrix of u_t is $\Sigma = \left(\mathbf{A}\mathbf{A}'\right)^{-1}$.
In contrast to the traditional econometric approach, the VAR approach puts emphasis almost exclusively on the dynamic properties of endogenous variables $\mathbf{y}_t$ rather than exogenous variables $\mathbf{z}_t$. In most VAR applications, $\mathbf{z}_t$ simply contains the constant terms.

Identification

One main objective in the VAR literature is to obtain economically meaningful impulse responses to structural shocks ε_t. To achieve this objective, it is necessary to impose at least $n(n-1)/2$ identifying restrictions, often on the contemporaneous coefficients represented by **A** in the structural system (1). In his original work, Sims (1980) makes the contemporaneous coefficient matrix **A** triangular for identification. The triangular system, often called the recursive identification, has a ‘Wold chain causal’

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
interpretation which is based on the timing of how shocks affect variables contemporaneously. It assumes that some shocks may influence only a subset of variables within the current period. This identification is still popular because it is straightforward to use and can yield some results that match widely held views. Christiano, Eichenbaum and Evans (1999) discuss extensively how recursive identification can be used in policy analysis.

There are fundamental economic applications that require identification under alternative assumptions rather than the recursive system. One familiar example is the determination of price and quantity as discussed in Sims (1986) and Gordon and Leeper (1994). Both variables are often determined *simultaneously* by the supply and demand equations in equilibrium; this simultaneity is inconsistent with recursive identification. Bernanke (1986) and Blanchard and Watson (1986) pioneered other applications of non-recursive identified VARs. Estimation of non-recursive VARs presents technical difficulties that are absent in recursive systems. These difficulties help explain the use of recursive VARs even if this maintained assumption is implausible. Recent developments in Bayesian econometrics, however, have made it feasible to estimate non-recursive VARs.

All of these works focus on the contemporaneous coefficient matrix. There are other ways to achieve identification. Blanchard and Quah (1993) and Gali (1992) propose using identifying restrictions directly on short-run and long-run impulse responses, which have been used in quantifying the effects of technology shocks and various nominal shocks, although the unreliable statistical properties of long-run restrictions are documented by Faust and Leeper (1997).

Many VAR applications rely on exact identification: the number of identifying restrictions equals $n(n-1)/2$. This counting condition is necessary but *not* sufficient for identification. To see this point, consider a three-variable VAR with the following restrictions

$$\mathbf{A} = \begin{bmatrix} * & * & 0 \\ 0 & * & * \\ * & 0 & * \end{bmatrix}$$

where *'s indicate unrestricted coefficients and 0's indicate exclusion restrictions. This VAR is *not* identified because in general there exist two distinct sets of structural parameters that deliver the same dynamics of $\mathbf{Y}_t$. For larger and more complicated systems with both short-run and long-run restrictions, there has been, until recently, no practical guidance as to whether the model is identified. The paper by Rubio-Ramirez, Waggoner and Zha (2005) develops a theorem for a necessary and sufficient condition for a VAR to be exactly identified. This theorem applies to a wide range of identified VARs, including those used in the literature. The basic idea is to transform the original structural parameters to the $(np+n)\times n$ matrix F (which is a function of $\mathbf{A}, \mathbf{A}_1, \ldots \mathbf{A}_p, \mathbf{D}$) so that linear restrictions can be applied to each column of F. The linear restrictions for the ith column of F can be summarized by the matrix Q_i of rank q_i, where q_i is the number of restrictions. According to their theorem, the VAR model is exactly identified if and only if $q_i = n - i$ for $1 \le i \le n$. This result gives the researcher a practical way to determine whether a VAR model is identified.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

When the number of identifying restrictions is greater than $n(n-1)/2$, a VAR is over-identified. Allowing for over-identification is important since economic theory often implies more than $n(n-1)/2$ restrictions. Moreover, many economic applications call for restrictions on the model's parameters beyond the contemporaneous coefficients (Cushman and Zha, 1997). Restrictions on the lag structure, such as block recursions, offer an effective way to handle over-parameterization when the lag length is long (Zha, 1999). Classical or Bayesian econometric procedures can be used to test over-identifying restrictions. A review of theoretical results for Bayesian estimation and inference for both exactly identified and over-identified VARs is discussed below.

Impulse responses

Impulse responses are most commonly used in the VAR literature and are defined as $\partial \mathbf{y}_{t+s} / \partial \varepsilon_t'$ for $s \geq 0$. Let Φ_s be the $n \times n$ impulse response matrix at step *s* and the *i*th row of Φ_s be responses of the *n* endogenous variables to the *i*th one-standard-deviation structural shock. One can show that the impulse responses can be recursively updated as

$$\Phi_s = \Phi_{s-1}\mathbf{B}_1 + \ldots + \Phi_{s-p}\mathbf{B}_p \tag{3}$$

with the convention that $\Phi_0 = \mathbf{A}^{-1}$ and $\Phi_v = \underset{n \times n}{\mathbf{0}}$ for $v < 0$.

The concept of impulse response is economically appealing and is used in strands of literature other than VAR work. For example, impulse responses to technology shocks or monetary policy shocks in a DSGE have been often compared to those in a VAR model. In empirical monetary economics, impulse responses of various macroeconomic variables to policy shocks have been a focal point in the recent debate on the effectiveness of monetary policy. These shocks can be thought of as shifts (deviations) from the systematic part of monetary policy that are hard to predict from the viewpoint of the public.

It is sometimes argued that identified VARs are unreliable because certain conclusions are sensitive to the specific identifying assumptions. This argument is a sophism. All economic models, DSGE model and VARs alike, are founded on ‘controversial’ assumptions, and the results can be sensitive to these assumptions. What researchers should do is to select a class of models based on how well they fit to the data, analyse how reasonable the underlying assumptions are, and examine whether there are robust conclusions across models.

Christiano, Eichenbaum, and Evans (1999) and Rubio-Ramirez, Waggoner and Zha (2005) show some important robust results across different VAR models that have reasonable assumptions and fit to the data equally well. One prominent example is the robust conclusion that a large fraction of the variation in policy instruments, such as the short-term interest rate, can be attributed to the systematic response of policy to shocks originating from the private economy. Such a conclusion is expected of good monetary policy, but it also explains the subtle and difficult task of identifying monetary policy shocks separately

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
from the other shocks affecting the economy.

Estimation and inference

Bayesian prior

When one estimates a VAR model for macroeconomic time series data, there is a trade-off between using short and long lags. A VAR with a short lag is prone to mis-specification, and a VAR with a long lag length is likely to suffer from the over-fitting problem. The Bayesian prior proposed by Sims and Zha (1998) is designed to eliminate the over-fitting problem without reducing the dimension of the model. It applies to not only reduced-form but also identified VARs.
To describe this prior simply, let $\mathbf{z}_t$ contain only a constant term and thus **D** is a $1 \times n$ vector of parameters. Rewrite the structural system (1) in the compact form of $\mathbf{y}_t'\mathbf{A} = \mathbf{x}_t'\mathbf{F} + \varepsilon_t'$, where

$$\underset{1\times k}{\mathbf{x}_t'} = \left[\mathbf{y}_{t-1}' \wedge \mathbf{y}_{t-p}' \ \mathbf{z}_t'\right], \quad \underset{n\times k}{\mathbf{F}'} = \left[\mathbf{A}' \wedge \mathbf{A}_p' \ \mathbf{D}'\right],$$

and $k = np + h$. For $1 \le j \le n$, let $\mathbf{a}_j$ be the *j*th column of **A** and $\mathbf{f}_j$ be the *j*th column of **F**. The first component of the prior is that $\mathbf{a}_j$ and $\mathbf{f}_j$ have Gaussian distribution

$$\mathbf{a}_j \sim N(\mathbf{0}, \mathbf{S}) \ \text{ and } \ \mathbf{f}_j \,|\, \mathbf{a}_j \sim N(\mathbf{P}\mathbf{a}_j, \mathbf{H}), \tag{4}$$

where $\underset{n\times k}{\mathbf{P}'} = \begin{bmatrix} \underset{n\times n}{\mathbf{I}} & \underset{n\times n}{\mathbf{0}} & \cdots & \underset{n\times n}{\mathbf{0}} & \underset{n\times 1}{\mathbf{0}} \end{bmatrix}$, which is consistent with the reduced-form random walk prior of Litterman (1986). The covariance matrices **S** and **H** are assumed to be diagonal matrices and are treated as hyperparameters. In principle, one could estimate these hyperparameters or integrate them out in a hierarchical framework. In practice, the values of these hyperparameters are specified before estimation. The *i*th diagonal element of **S** is λ_0 / σ_i. The diagonal element of **H** that corresponds to the coefficient on lag l of variable *i* in equation *j* is $(\lambda_0\lambda_1\lambda_2^{\delta(i,j)}) / (\sigma_i l^{\lambda_3})$, where $\delta(i, j)$ equals 1 if $i = j$ and 0 otherwise. The diagonal element of **H** corresponding to the constant term is the square of $\lambda_0\lambda_4$. The hyperparameter λ_0 controls the overall tightness of belief about the random walk feature, as well as tightness on the prior of **A** itself; λ_1 further controls the tightness of belief on random walk and the relative tightness on the prior of lagged coefficients; λ_2 controls the influence of variable *i* in equation *j*; λ_3 controls the rate at which the influence of lag decreases as its length increases; and λ_4 controls the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

relative tightness on the zero value of the constant term. The hyperparameters σ_i are scale factors to make the units uniform across variables, and are chosen at the sample standard deviations of residuals from univariate autoregressive models fitted to the individual time series in the sample (Litterman, 1986).

A VAR with many variables and a long lag is likely to produce relatively large coefficient estimates on distant lags and thus volatile sampling errors. The prior described here is designed to reduce the influence of distant lags and the unreasonable degree of explosiveness embedded in the system. It is essential for ensuring reasonable small-sample properties of the model, especially when there are relatively few degrees of freedom in a large VAR.

The aforementioned prior, however, does not take into account the features of unit roots and cointegration relationships embedded in many time series. For this reason, Sims and Zha (1998) add another component to their prior. This component uses Litterman's idea of dummy observations to express beliefs on unit roots and cointegration. Specifically, there are $n+1$ dummy observations added to the original system, which can be written as

$$\mathbf{Y}_d\mathbf{A} = \mathbf{X}_d\mathbf{F} + \mathbf{E}, \tag{5}$$

where **E** is a matrix of random shocks,

$$\underset{(n+1)\times n}{\mathbf{Y}_d} = \begin{bmatrix} \mu_5\bar{y}_1^0 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mu_5\bar{y}_n^0 \\ \mu_6\bar{y}_1^0 & \cdots & \mu_6\bar{y}_n^0 \end{bmatrix}, \quad \underset{(n+1)\times 1}{\mathbf{c}_d} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mu_6 \end{bmatrix},$$

$$\underset{(n+1)\times(np+1)}{\mathbf{X}_d} = [\mathbf{Y}_d \ \cdots \ \mathbf{Y}_d \ \ \mathbf{c}_d],$$

and $\bar{y}_i^0$ is the sample average of the *p* initial conditions for the *i*th variable of $\mathbf{Y}_t$ and μ_5 and μ_6 are hyperparameters. The first $n+1$ dummy-observation equations in (5) express beliefs that all variables are stationary with means equal to $\bar{y}_i^0$'s or cointegration is present. The larger the values of μ_5 and μ_6, the stronger these beliefs. Since the values of λ's and μ's move in opposite directions to increase or loosen the tightness of the prior, the two symbols λ and μ are kept distinct. In applied work, the values of the hyperparameters for quarterly data are typically set to $\lambda_0 = 1$, $\lambda_1 = 0.2$, and $\lambda_2 = \lambda_3 = \lambda_4 = \mu_5 = \mu_6 = 1.0$. For monthly data, $\lambda_0 = 0.6$, $\lambda_1 = 0.1$, $\lambda_2 = 1.0$, $\lambda_4 = 0.1$, and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
$\mu_5 = \mu_6 = 5.0$, while the choice of the lag decay weight λ_3 is somewhat complicated and is elaborated in Robertson and Tallman (1999).

By taking into account the cointegration relationships among macroeconomic variables, this additional component of the prior helps improve out-of-sample forecasting, reduces the difference in forecasting accuracy between using the vintage and final data, and produces robust impulse responses to monetary policy shocks across VARs with different identification assumptions (Robertson and Tallman, 1999; 2001). Furthermore, Leeper, Sims and Zha (1996) demonstrate that with this prior it is feasible to estimate VAR models with as many as 18 variables – far more than the current DSGE models can handle. Because the prior proposed by Sims and Zha (1998) reflects *widely held* beliefs in the behaviour of macroeconomic time series, it has been often used as a base line prior in the Bayesian estimation and inference of VAR models.

Marginal data density

If a model is used as a candidate for the ‘true’ data-generating mechanism, it is imperative that the model's fit to the data is superior to those of alternative models. Recent developments in Bayesian econometrics have made it feasible to compare nested and non-nested models for their fits to the data (Geweke, 1999). With a proper Bayesian prior, one can numerically compute the marginal data density (MDD) defined as

$$\int_{\Theta} L(\mathbf{Y}_T|\varphi)\, p(\varphi)\, d\varphi, \tag{6}$$

where ϕ is a collection of all the model's parameters, Θ is the domain of ϕ, $\mathbf{Y}_T$ is all the data up to T, and $L(\mathbf{Y}_T|\varphi)$ is the proper likelihood function. To determine the goodness of fit of a DSGE model, for example, one can compare its MDD with that of a VAR model (Smets and Wouters, 2003; Del Negro and Schorfheide, 2004).

As a VAR is often used as a benchmark for comparing different models, it is important that one compute its MDD efficiently and accurately. For an unrestricted reduced-form VAR as specified in (2), there is a standard closed-form expression for (6) so that no Markov chain Monte Carlo (MCMC) method is needed to obtain the MDD. For restricted (tightly parameterized) VARs implied by a growing number of economic applications, there is in general no closed-form solution to (6), and a numerical approximation to (6) is needed. Because of a high dimension in the VAR parameter space and possible simultaneity in an identified model, popular MCMC approaches such as importance sampling and modified harmonic mean methods require a long sequence of posterior draws to achieve numerical reliability in approximating (6), and thus are computationally very demanding.

Chib (1995) offers a procedure for accurate evaluations of the MDD that requires the existence of a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gibbs sampler by partitioning ϕ into a few blocks. One can sample alternately from the conditional posterior distribution of one block of parameters given other blocks. While sampling between blocks entails additional simulations, the Chib algorithm can be far more efficient than other methods because each conditional posterior probability density function (PDF) can be evaluated in closed form. The objects needed to complete this algorithm are the closed-form prior PDF and the conditional posterior PDF for each block.

Because the prior discussed so far includes the dummy observations component, there is a question as to whether this overall prior has a standard PDF. To answer this question, it can be shown from (4) and (5) that the overall prior PDF is

$$\mathbf{a}_j \sim N(\mathbf{0}, \bar{\mathbf{S}}) \text{ and } \mathbf{f}_j | \mathbf{a}_j \sim N(\mathbf{P}\mathbf{a}_j, \bar{\mathbf{H}}), \tag{7}$$

where $\bar{\mathbf{S}} = \mathbf{S}$ and $\bar{\mathbf{H}} = \left(\mathbf{X}_d'\mathbf{X}_d + \mathbf{H}^{-1}\right)^{-1}$. The result (7) follows from the two claims:

$$\left(\mathbf{X}_d'\mathbf{X}_d + \mathbf{H}^{-1}\right)^{-1}\left(\mathbf{X}_d'\mathbf{Y} + \mathbf{H}^{-1}\mathbf{P}\right) = \mathbf{P};$$

$$\mathbf{Y}_d'\mathbf{Y}_d + \mathbf{P}'\mathbf{H}^{-1}\mathbf{P} = \left(\mathbf{Y}_d'\mathbf{X}_d + \mathbf{P}'\mathbf{H}^{-1}\right)\mathbf{P}.$$

Given the prior (7), Waggoner and Zha (2003a) develop a Gibbs sampler for identified VARs with the linear restrictions studied in the VAR literature. These restrictions can be summarized as

$$\underset{n\times n}{\mathbf{Q}_j}\,\mathbf{a}_j = \underset{n\times 1}{\mathbf{0}}, \quad \underset{n\times k}{\mathbf{R}_j}\,\mathbf{f}_j = \underset{n\times 1}{\mathbf{0}}; \quad j = 1, \ldots n. \tag{8}$$

If there are q_j restrictions on $\mathbf{a}_j$ and r_j restrictions on $\mathbf{f}_j$, the ranks of $\mathbf{Q}_j$ and $\mathbf{R}_j$ are q_j and r_j

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

respectively. Let $\mathbf{U}_j$ ($\mathbf{R}_j$) be an $n \times q_j$ ($n \times r_j$) matrix whose columns form an orthonormal basis for the null space of $\mathbf{Q}_j$ ($\mathbf{R}_j$). The conditions in (8) are satisfied if and only if there exist a $q_j \times 1$ vector $\mathbf{b}_j$ and an $r_j \times 1$ vector $\mathbf{g}_j$ such that $\mathbf{a}_j = \mathbf{U}_j\mathbf{b}_j$ and $\mathbf{f}_j = \mathbf{V}_j\mathbf{g}_j$. The vectors $\mathbf{b}_j$ and $\mathbf{g}_j$ are the free parameters of $\mathbf{a}_j$ and $\mathbf{f}_j$ dictated by the conditions in (8). It follows from (7) that the prior distribution of $\mathbf{b}_j$ and $\mathbf{g}_j$ is jointly normal.

As for the conditional posterior PDFs, it can be shown that the posterior distribution of $\mathbf{g}_j$ conditional on $\mathbf{b}_j$ is normal and that the posterior distribution of $\mathbf{b}_j$ conditional on $\mathbf{b}_i$'s for $i \neq j$ has a closed-form PDF and can be simulated from it exactly. These results enable one to use the efficient method of Chib (1995). The MDD calculated this way is reliable and requires little computing time. For example, it takes less than one minute to obtain a very reliable estimate of the MDD for a large VAR with 13 lags and 10 variables. Such accuracy and speed make it feasible to compare a large number of identified VARs with different degrees of restriction.

Error bands

Because impulse responses are of central interest in interpreting dynamic multivariate models and helping guide the directions for new economic theory to be developed (Christiano, Eichenbaum and Evans, 2005), it is essential that measures of the statistical reliability of estimated impulse responses be presented as part of the process of evaluating models. The Bayesian methods reviewed so far in this essay make it feasible to construct the error bands around impulse responses. The error bands can contain any probability and are typically expressed in both .68 and .90 probability bands to characterize the shapes of the likelihood implied by the model.

The error bands of impulse responses reported in most VAR works are constructed as follows. One begins with the Gibbs sampler to draw $\mathbf{b}_j$ and $\mathbf{g}_j$ for $j = 1, \ldots n$. For each posterior draw, the free parameters $\mathbf{b}_j$'s and $\mathbf{g}_j$'s are transformed to the original structural parameters **A**, $\mathbf{A}_l$ ($l = 1, \ldots p$), and **D**; then the impulse responses are computed according to (3). The empirical distribution for each element of the impulse responses is formed and the equal-tail .68 and .90 probability intervals around each element are computed. The probability intervals have exact small-sample properties from a Bayesian point of view; and .90 or .95 probability intervals have been used in the empirical literature to approximate classical small-sample confidence intervals when the high dimensional parameter space and a large number of nuisance parameters make it difficult or impossible to obtain exact classical inferences.

One issue related to the error bands around impulse responses, whose importance is beginning to be recognized, is normalization. A normalization rule selects the sign of each draw of impulse responses from the posterior distribution. If there is no restriction imposed on the sign of each column of the contemporaneous coefficient matrix **A**, then the likelihood or the posterior function remains the same when the sign of a column of **A** is reversed. Without any sign restriction, the error bands for impulse responses would be symmetric around zero and thus the estimated responses would be determined to be imprecise.

The conventional normalization is to keep the diagonal of **A** always positive, based on the notion that a choice of normalization cannot have substantive effects on the results. But this notion is mistaken. If an

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
identified VAR is non-recursive, normalization can generate ill-determined or unreasonably wide error bands around some impulse responses because some coefficients on the diagonal may be insignificantly different from zero.

Waggoner and Zha (2003b) show that normalized likelihoods can be different across normalization rules and that inappropriate normalization tends to produce a multi-modal likelihood. They propose a normalization rule designed to prevent the normalized likelihood from being spuriously multi-modal and thus avoid unreasonably wide error bands caused by the multi-modal likelihood. The algorithm for their normalization is straightforward to implement: for each posterior draw of $\mathbf{a}_j$, keep $\mathbf{a}_j$ if $\mathbf{e}_j'\mathbf{A}^{-1}\hat{\mathbf{a}}_j > 0$ and replace $\mathbf{a}_j$ with $-\mathbf{a}_j$ if $\mathbf{e}_j'\mathbf{A}^{-1}\hat{\mathbf{a}}_j < 0$, where $\mathbf{e}_j$ is the *j*th column of the $n \times n$ identity matrix. This algorithm works for not only short-run but also long-run restrictions (Evans and Marshall, 2002).

Another important issue related to error bands, not addressed until recently, is the characterization of the uncertainty around estimated impulse responses not only at one particular point but also around the shape of the responses as a whole. Let $\Phi_s(i, j)$ be the *s*-step impulse response of the *j*th variable to the *i*th structural shock. The associated error band is only pointwise. It is very unlikely in economic applications, however, that uncertainty about $\Phi_s(i, j)$ is independent across *j* or *s*. For example, the response of output to a policy shock is likely to be negatively correlated with the response of unemployment, and the response of inflation this period is likely to be positively correlated with the previous and next responses.

The procedure proposed by Sims and Zha (1999) takes into account these possible correlations across variables and across time. To use this procedure, one can simply stack all the relevant impulse responses into a column vector denoted by $\tilde{\mathbf{c}}$, where the tilde refers to a posterior draw. From a large number of posterior draws, the mean $\bar{\mathbf{c}}$ and covariance matrix $\bar{\Omega}$ of $\tilde{\mathbf{c}}$ are computed. For each posterior draw $\tilde{\mathbf{c}}$, the *k*th component $\tilde{Y}_k = (\tilde{\mathbf{c}} - \bar{\mathbf{c}})'\bar{\mathbf{w}}_k$ is calculated, where $\bar{\mathbf{w}}_k$ is the eigenvector corresponding to the *k*th largest eigenvalue of $\bar{\Omega}$. From the empirical distribution of $\tilde{Y}_k$, one can tabulate different quantiles such as $Y_{k,.16}$ and $Y_{k,.84}$. Thus, the .68 probability error bands explained by the *k*th component of variation in the group of impulse responses can be computed as $\mathbf{c}_{.16} = \bar{\mathbf{c}} + Y_{k,.16}\bar{\mathbf{w}}_k$ and $\mathbf{c}_{.84} = \bar{\mathbf{c}} + Y_{k,.84}\bar{\mathbf{w}}_k$. For a particular economic application, if it turns out that only one to three eigenvalues dominate the covariance matrix of $\tilde{\mathbf{c}}$, these kinds of connecting-dots error bands can be useful in understanding the magnitudes and directions of uncertainty among a group of interrelated impulse responses. This method has proven to be particularly useful in economic applications that characterize the uncertainty around the entire paths, not just points one at a time (Cogley and Sargent, 2005; Nason and Rogers, 2006).

Markov-switching VARs

The class of VARs discussed thus far assumes that the parameters are constant over time. This assumption is made mainly for the technical constraint on estimation and inference, however. Many macroeconomic time series display patterns that seem impossible to capture by constant-parameter VARs. One prominent example is changes in volatility over time. In the VAR framework, volatility changes mean that the reduced-form covariance matrix Σ is not constant. In policy analysis, there is a serious debate on whether the coefficients in the policy rule have changed over time, or whether the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
variances of shocks in the private sector have changed over time, or both. Time-varying VARs are designed to answer these kinds of questions. Stock and Watson (2003) use the reduced-form VAR framework to show that fluctuations in US business cycles can be largely explained by changes in Σ. Sims and Zha (2006b) identify the behaviour of monetary policy from the rest of the VAR system and show that changes in the coefficients in monetary policy are, at most, modest and the variance changes in shocks originating from the private sector dominate aggregate fluctuations.

There have been a number of studies on time-varying VARs that allow the coefficients or the covariance matrix of residuals or both to change over time. These models typically let all the coefficients drift as a random walk or persistent process. To the extent that this kind of modelling tries to capture possible changes in the model's parameters, the model tends to over-fit because the dimension of time variation embedded in the data is much lower than the model's specification. Conceptually, there is a problem of distinguishing shocks to the residuals from shocks to the coefficients. The inability to distinguish among these shocks makes it difficult to interpret the effects of, say, monetary policy shocks.

The Markov-switching VAR introduced by Sims and Zha (2006a) is designed to overcome the over-fitting problems present in the other time-varying VARs and, at the same time, maintain clear interpretation of structural shocks. It builds on the Markov-switching model of Hamilton (1989), but emphasizes ways to restrict the degree of time variation allowed in the VAR. It has a capability to approximate parameter drifts arbitrarily well with the growing number of states, while restricting the transition matrix to be concentrated on the diagonal. This feature also allows discontinuous jumps from one state to another, which appears to matter for aggregate fluctuations.

To see how this method works, suppose that the parameter z_t drifts according to the process $z_t = \rho z_{t-1} + v_t$ where $v_t \sim N(0, \sigma^2)$. By discretizing this autoregressive process, one can let the probability of the transition from state *j* to *i* be proportional to

$$\Pr\left[z_t \in \left(\frac{\tau_i \sigma}{\sqrt{1-\rho^2}}, \frac{\tau_{i+1}\sigma}{\sqrt{1-\rho^2}}\right) \middle| z_{t-1} = \frac{\tau_j + \tau_{j+1}}{2}\frac{\rho\sigma}{\sqrt{1-\rho^2}}\right]$$
$$= \Psi\left(\frac{\tau_{i+1}}{\sqrt{1-\rho^2}} - \frac{\tau_j + \tau_{j+1}}{2}\frac{\rho}{\sqrt{1-\rho^2}}\right) - \Psi\left(\frac{\tau_i}{\sqrt{1-\rho^2}} - \frac{\tau_j + \tau_{j+1}}{2}\frac{\rho}{\sqrt{1-\rho^2}}\right),$$

where $\Psi(\,)$ is the standard normal cumulative probability function. The values of τ divide up the interval between –2 and 2 (two standard deviations). For nine states, for example, one has $\tau_1 = -2$, $\tau_2 = -1.5$, $\tau_3 = -1$, …, $\tau_8 = 1.5$, and $\tau_9 = 2$. Careful restrictions on the degree of time variation, as well as on the constant parameters themselves, will put VARs a step closer to DSGE modelling. Recent work by Davig and Leeper (2005) shows an example of how to use a DSGE model to restrict a VAR on monetary and fiscal policy.

Conclusion

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

There is a tension between models that have clear economic interpretations but offer a poor fit to data and models that fit well but have few a priori assumptions and are therefore less interpretable (Ingram and Whiteman, 1994; Del Negro and Schorfheide, 2004). The original philosophy motivating VARs assumes that the economy is sufficiently complex and that simplified theoretical models, while useful in organizing thought about how the economy works, generally abstract from important aspects of the economy. VAR modelling begins with the minimal restrictions on dynamic time-series models, explores empirical regularities that have been ignored by simple models, and insists on the model's fit to data. The emphasis on fit has begun to bear fruit, as an increasing array of dynamic stochastic general equilibrium models have been tested and compared with VARs (Christiano, Eichenbaum and Evans, 2005; Smets and Wouters, 2003). Markov-switching VARs go a step further in bringing VARs even closer to the data and thus provide a new benchmark for model comparison.

At the same time, considerable progress has been made to narrow the gap between VARs and DSGE models. Some results from VARs have provided empirical support to the key assumption made by real business cycle (RBC) models that monetary policy shocks play insignificant roles in generating business fluctuations. Nason and Cogley (1994) and Cogley and Nason (1995) discuss similar results from both VAR and RBC approaches. Fernandez-Villaverde, Rubio-Ramirez and Sargent (2005) provide conditions and examples under which there exists the VAR representation of a DSGE model. Sims and Zha (2006a) display a close connection between an identified VAR and a DSGE model, and provide a measure for determining whether the ‘invertibility problem’ is a serious issue.

Undoubtedly there are payoffs in moving beyond the original VAR philosophy by imposing more restrictions on both contemporaneous relationships and lag structure while the restrictions are guided carefully by economic theory. Although moving in this direction is desirable, it is essential to maintain the spirit of VAR analysis as originally proposed by Sims (1980). This requires that heavily restricted VARs be subject to careful evaluation in terms of fit. Recent advances in Bayesian estimation and inference methods of restricted VARs make it feasible to compute the MDD accurately and efficiently and, therefore, to determine whether the restrictions have compromised the fit. These methods, however, still fall short of handling VARs with cross-equation restrictions implied by DSGE models. Thus, the challenge ahead of us is to develop new tools for VARs with possible cross-equation restrictions.

See Also

- Bayesian econometrics
- Bayesian methods in macroeconometrics
- Markov chain Monte Carlo methods
- structural vector autoregressions

Bibliography

Bernanke, B. 1986. Alternative exploration of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy* 25, 49–99.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Blanchard, O. and Quah, D. 1993. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 83, 655–73.

Blanchard, O. and Watson, M. 1986. Are business cycles all alike? In *The American Business Cycle: Continuity and Change*, ed. R. Gordon. Chicago: University of Chicago Press.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 1313–21.

Christiano, L., Eichenbaum, M. and Evans, C. 1999. Monetary policy shocks: what have we learned and to what end? In *Handbook of Macroeconomics*, vol. 1A, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Christiano, L., Eichenbaum, M. and Evans, C. 2005. Nominal rigidities and the dynamics effects of a shock to monetary policy. *Journal of Political Economy* 113, 1–45.

Cogley, T. and Nason, J. 1995. Output dynamics in real business cycle models. *American Economic Review* 85, 492–511.

Cogley, T. and Sargent, T. 2005. Drifts and volatilities: monetary policies and outcomes in the post WWII U.S. *Review of Economic Dynamics* 8, 262–302.

Cushman, D. and Zha, T. 1997. Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary Economics* 39, 433–48.

Davig, T. and Leeper, E. 2005. Fluctuating macro policies and the fiscal theory. Working Paper No. 11212. Cambridge, MA: NBER.

Del Negro, M. and Schorfheide, F. 2004. Priors from general equilibrium models for VARs. *International Economic Review* 45, 643–73.

Evans, C. and Marshall, D. 2002. Economic determinants of the nominal treasury yield curve. Working paper. Federal Reserve Bank of Chicago.

Faust, J. and Leeper, E. 1997. When do long-run identifying restrictions give reliable results? *Journal of Business and Economic Statistics* 15, 345–53.

Fernandez-Villaverde, J., Rubio-Ramirez, J. and Sargent, T. 2005. A, B, C's (and D's) for understanding VARs. Working Paper No. 2005-9. Federal Reserve Bank of Atlanta.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gali, J. 1992. How well does the IS-LM model fit postwar U.S. data? *Quarterly Journal of Economics* 107, 709–38.

Geweke, J. 1999. Using simulation methods for Bayesian econometric models: inference, development, and communication. *Econometric Reviews* 18, 1–73.

Geweke, J. and Whiteman, C. 2006. Bayesian forecasting. In *The Handbook of Economic Forecasting*, ed. G. Elliott, C. Granger and A. Timmermann. Amsterdam: North-Holland.

Gordon, D. and Leeper, E. 1994. The dynamic impacts of monetary policy: an exercise in tentative identification. *Journal of Political Economy* 102, 1228–47.

Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–84.

Ingram, B. and Whiteman, C. 1994. Supplanting the Minnesota prior: forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics* 34, 497–510.

Leeper, E., Sims, C. and Zha, T. 1996. What does monetary policy do? *Brookings Papers on Economic Activity* 2, 1–78.

Leeper, E. and Zha, T. 2003. Modest policy interventions. *Journal of Monetary Economics* 50, 1673–700.

Litterman, R. 1986. Forecasting with Bayesian vector autoregressions – five years of experience. *Journal of Business and Economic Statistics* 4, 25–38.

Nason, J. and Cogley, T. 1994. Testing the implications of long-run neutrality for monetary business cycle models. *Journal of Applied Econometrics* 9, S37–S70.

Nason, J. and Rogers, J. 2006. The present-value model of the current account has been rejected: round up the usual suspects. *Journal of International Economics* 68, 159–87.

Robertson, J. and Tallman, E. 1999. Vector autoregressions: forecasting and reality. *Federal Reserve Bank of Atlanta Economic Review* 84(1), 4–18.

Robertson, J. and Tallman, E. 2001. Improving federal-funds rate forecasts in VAR models used for policy analysis. *Journal of Business and Economic Statistics* 19, 324–30.

Rubio-Ramirez, J., Waggoner, D. and Zha, T. 2005. Markov-switching structural vector autoregressions: theory and applications. Working Paper No. 2005–27. Federal Reserve Bank of Atlanta.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48, 1–47.

Sims, C. 1982. Policy analysis with econometric models. *Brookings Papers on Economic Activity* 1, 107–52.

Sims, C. 1986. Are forecasting models usable for policy analysis. *Federal Reserve Bank of Minneapolis Quarterly Review* 10(1), 2–16.

Sims, C. and Zha, T. 1998. Bayesian methods for dynamic multivariate models. *International Economic Review* 39, 949–68.

Sims, C. and Zha, T. 1999. Error bands for impulse responses. *Econometrica* 67, 1113–55.

Sims, C. and Zha, T. 2006a. Does monetary policy generate recessions? *Macroeconomic Dynamics* 10 (2) 231–272.

Sims, C. and Zha, T. 2006b. Were there regime switches in US monetary policy? *American Economic Review* 96, 54–81.

Smets, F. and Wouters, R. 2003. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association* 1, 1123–75.

Stock, J. and Watson, M. 2003. Has the business cycle changed? Evidence and explanations. Prepared for the Federal Reserve Bank of Kansas City Symposium ‘Monetary Policy and Uncertainty: Adapting to a Changing Economy’, Jackson Hole, Wyoming, 28–30 August.

Waggoner, D. and Zha, T. 2003a. Likelihood-preserving normalization in multiple equation models. *Journal of Econometrics* 114, 329–47.

Waggoner, D. and Zha, T. 2003b. A Gibbs simulator for structural vector autoregressions. *Journal of Economic Dynamics & Control* 28, 349–66.

Zha, T. 1999. Block recursion and structural vector autoregressions. *Journal of Econometrics* 90, 291–316.

How to cite this article

Zha, Tao. "vector autoregressions." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000066> doi:10.1057/9780230226203.1792

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

venture capital

Josh Lerner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Venture capital is independently managed, dedicated capital focusing on equity or equity-linked investments in privately held, high-growth companies. Research into venture capital has focused on the structure and financing of venture partnerships, the financial and operational interactions of venture capitalists with portfolio firms, and the exiting of venture capital investments. Major areas needed further research include the internationalization of venture capital, the impact of public policy, and the real economic effects of these funds.

Keywords

agency conflicts; asymmetric information; capital gains taxation; corporate governance; covenants; entrepreneurship; financial intermediaries; high-risk assets; initial public offerings; investment; limited liability; limited partnerships; liquidity constraints; maximum likelihood; monitoring; patents; pay-for-performance incentives; reputation; research and development; venture capital

Article

Venture capital is independently managed, dedicated capital focusing on equity or equity-linked investments in privately held, high-growth companies. The first venture firm, American Research and Development, was formed in 1946 and invested in companies commercializing technology developed during the Second World War. Because institutions were reluctant to invest, it was structured as a publicly traded closed-end fund and marketed mostly to individuals, a structure emulated by its successors.

By 1978 limited partnerships had become the dominant investment structure. Limited partnerships have an important advantage: capital gains taxes are not paid by the limited partnership. Instead, only the taxable investors in the fund pay taxes. Venture partnerships have predetermined, finite lifetimes. To maintain limited liability, investors must not become involved in the management of the fund.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Activity in the venture industry increased dramatically in early 1980s. Much of the growth stemmed from the US Department of Labor's clarification of Employee Retirement Income Security Act's ‘prudent man’ rule in 1979, which had prohibited pension funds from investing substantial amounts of money into venture capital or high-risk asset classes. The rule clarification explicitly allowed pension managers to invest in high-risk assets, including venture capital.

The subsequent years saw both very good and trying times for venture capitalists. Venture capitalists backed many successful companies, including Apple Computer, Cisco, Genentech, Google, Netscape, Starbucks, and Yahoo! But commitments to the venture capital industry were very uneven, creating a great deal of instability. The annual flow of money into venture funds increased by a factor of ten during the early 1980s. From 1987 through 1991, however, fund-raising steadily declined as returns fell. Between 1996 and 2003, this pattern was repeated.

Venture capital investing can be viewed as a cycle. In this article, I follow the cycle of venture capital activity. I begin with the formation of venture funds. I then consider the process by which such capital is invested in portfolio firms, and the exiting of such investments. I end with a discussion of open research questions, including those relating to internationalization and the real effects of venture activity.

Fund-raising

Research into the formation of venture funds has focused on two topics. First, the commitments to the venture capital industry have been highly variable since the mid-1970s. Understanding the determinants of this variability has been a topic of continuing interest to researchers. Second, the structure of venture partnerships has attracted increasing attention.

First, Poterba (1987; 1989) notes that the fluctuations could arise from changes in either the supply of or the demand for venture capital. It is very likely, he argues, that decreases in capital gains tax rates increase commitments to venture funds, even though the bulk of the funds are from tax-exempt investors. The drop in the tax rate may spur corporate employees to become entrepreneurs, thereby increasing the need for venture capital. The increase in demand due to greater entrepreneurial activity leads to more venture fund-raising.

Gompers and Lerner (1998b) find empirical support for Poterba's claim: lower capital gains taxes have particularly strong effects on venture capital supplied by tax-exempt investors. This suggests that the primary mechanism by which capital gains tax cuts affect venture fund-raising is the higher demand of entrepreneurs for capital. The authors also find that a number of other factors influence venture fund-raising, such as regulatory changes and the returns of venture funds.

A second line of research has examined the contracts that govern the relationship between investors (limited partners) and the venture capitalist (general partner). Gompers and Lerner (1999) find that compensation for older and larger venture capital organizations is more sensitive to performance than that of other venture groups. Also, the cross-sectional variation in compensation terms for younger, smaller venture organizations is considerably lower. The fixed component of compensation is higher for smaller, younger funds and funds focusing on high-technology or early stage investments. Finally, Gompers and Lerner do not find any relationship between the incentive compensation and performance. The authors argue that these results are consistent with a learning model in which neither the venture capitalist nor the investor knows the venture capitalist's ability. With his early funds, the venture capitalist will work hard even without explicit pay-for-performance incentives: if he can establish a good

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
reputation, he can raise subsequent funds. These reputation concerns lead to lower pay for performance for smaller and younger venture organizations. Once a reputation has been established, explicit incentive compensation is needed to induce the proper effort.

Covenants also play an important role in limiting conflicts in venture partnerships. Their use may be explained by two hypotheses. First, because negotiating and monitoring covenants are costly, they will be employed when monitoring is easier and the potential for opportunistic behaviour is greater. Second, in the short run the supply of venture capital services may be fixed, with a modest number of funds of carefully limited size raised each year. Increases in demand may lead to higher prices when contracts are written. Higher prices may include not only increases in monetary compensation, but also greater consumption of private benefits through fewer covenants.

Gompers and Lerner (1996) show that both supply and demand conditions and costly contracting are important in determining contractual provisions. Fewer restrictions are found in funds established during years with greater capital inflows and funds, when general partners enjoy higher compensation. The evidence illustrates the importance of general market conditions on the restrictiveness of venture partnerships. In periods when venture capitalists have relatively more bargaining power, the venture capitalists are able to raise money with fewer stings attached.

Lerner and Schoar (2004) examine rationales for constraints on liquidity. Venture groups often impose severe restrictions on transfers of partnership interests beyond what is required by securities law. They argue that these curbs allow general partners to screen for long-term investors. A limited partner who expects many liquidity shocks would find these restrictions especially onerous. Thus, the limited partners investing will be highly liquid, facilitating fund-raising in follow-on funds. The authors show that restrictions on liquidity are less common in later funds organized by the same venture group, when information problems are presumably less severe.

Investing

A second broad area of research has focused on the ties between venture capitalists and the firms in which they invest.

This literature emphasizes the informational asymmetries that characterize young firms, particularly in high-technology industries. These problems make it difficult for investors to assess firms, and permit opportunistic behaviour by entrepreneurs after finance is received. Specialized financial intermediaries, such as venture capitalists, address these problems by intensively scrutinizing firms before providing capital and monitoring them afterwards.

Economic theory examines the role that venture capitalists play in mitigating agency conflicts between entrepreneurs and investors. The improvement in efficiency might be due to the active monitoring and advice that is provided (Cornelli and Yosha, 2003; Hellmann, 1998; Marx, 1994), the screening mechanisms employed (Chan, 1983), the incentives to exit (Berglöf, 1994), the proper syndication of the investment (Admati and Pfleiderer, 1994), or investment staging (Bergemann and Hege, 1998; Sahlman, 1990).

Staged capital infusion is the most potent control mechanism a venture capitalist can employ. The shorter the duration of an individual round of financing, the more frequently the venture capitalist monitors the entrepreneur's progress. The duration of funding should decline and the frequency of re-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
evaluation increase when the venture capitalist believes that conflicts with the entrepreneur are likely. If monitoring and information gathering are important – as models such as those of Amit, Glosten and Muller (1990) and Chan (1983) suggest – venture capitalists should invest in firms where asymmetric problems are likely, such as early stage and high-technology firms with intangible assets. The capital constraints faced by these companies will be large and these investors will address them.

Gompers (1995) shows that venture capitalists concentrate investments in early stage companies and high-technology industries where informational asymmetries are significant and monitoring is valuable. He finds that early stage firms receive significantly less money per round. Increases in asset tangibility are associated with longer financing duration and reduce monitoring intensity.

In a related paper, Kaplan and Strömberg (2003) document how venture capitalists allocate control and ownership rights contingent on financial and non-financial performance. If a portfolio company performs poorly, venture capitalists obtain full control. As performance improves, the entrepreneur obtains more control. If the firm does well, the venture capitalists relinquish most of their control rights but retain their equity stake.

Related evidence comes from Hsu (2004), who studies the price entrepreneurs pay to be associated with reputable venture capitalists. He analyses firms which received financing offers from multiple venture capitalists. Hsu shows that high investor experience is associated with a substantial discount in firm valuation.

Venture capitalists usually make investments with peers. The lead venture firm involves other venture firms. One critical rationale for syndication in the venture industry is that peers provide a second opinion on the investment opportunity and limit the danger of funding bad deals.

Lerner (1994a) finds that in the early investment rounds experienced venture capitalists tend to syndicate only with venture firms that have similar experience. He argues that, if a venture capitalist were looking for a second opinion, then he would want to get one from someone of similar or greater ability, certainly not from someone of lesser ability.

The advice and support provided by venture capitalists is often embodied in their role on the firm's board of directors. Lerner (1995) examines whether venture capitalists' representation on the boards of the private firms in their portfolios is greater when the need for oversight is larger, looking at changes in board membership around the replacement of CEOs. He finds that an average of 1.75 venture capitalists are added to the board between financing rounds when a firm's CEO is replaced in the interval; between other rounds 0.24 venture directors are added. No differences are found in the addition of other outside directors.

Hochberg (2005) studies the influence of venture capitalists on the governance of a firm following its initial public offering (IPO). Venture-backed firms manage earnings less in the IPO year, as measured by discretionary accounting accruals. Venture-backed firms also experience a stronger wealth effect when they adopt a poison pill, which implies that investors are less worried that the poison pill will entrench management at the expense of shareholders. Finally, venture-backed firms more frequently have independent boards and audit and compensation committees, as well as separate CEOs and chairmen.

It is natural to ask why other financial intermediaries (such as banks) cannot duplicate these features of the venture capitalists, and undertake the same sort of monitoring. Economists have suggested several explanations for the apparent superiority of venture funds in this regard. First, because regulations limit

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
banks' ability to hold shares, they cannot freely use equity. Second, banks may not have the necessary skills to evaluate projects with few collateralizable assets and significant uncertainty. Finally, venture funds' high-powered compensation schemes give venture capitalists incentives to monitor firms closely. Banks sponsoring venture funds without high-powered incentives have found it difficult to retain personnel.

So far, this section has highlighted the ways in which venture capitalists can successfully address agency problems in portfolio firms. During periods when the amount of money flowing into the industry grows dramatically, however, competition between venture groups can introduce distortions.

Gompers and Lerner (2000) examine the relation between the valuation of venture deals and inflows into venture funds. Doubling inflows leads to a 7–21 per cent increase in valuation levels. But success rates don't differ significantly between investments made during periods of low inflows and valuations on the one hand and those made in booms on the other. The results indicate that the price increases reflect increasing competition for investment.

Exiting

A third major area of research has been the process whereby venture funds exit investments. This topic is important because, in order to make money on their investments, venture capitalists must sell their equity stakes.

Initial research into the exiting of venture investments focused on IPOs. This reflects the fact that typically the most profitable exit opportunity is an IPO. Barry et al. (1990) and Megginson and Weiss (1991) document that venture capitalists hold significant equity stakes and board positions in the firms they take public, which they continue to hold a year after the IPO. They argue that this pattern reflects the certification they provide to investors that the firms they bring to market are not overvalued. Moreover, they show that venture-backed IPOs have less of a positive return on their first trading day, a finding that has been subsequently challenged (Lee and Wahal, 2004; Kraus, 2002). The authors suggest that investors need a smaller discount because the venture capitalist has certified the offering's quality.

Subsequent research has examined the timing of the exit decision. Several potential factors affect when venture capitalists choose to bring firms public. Lerner (1994b) examines how the valuation of public securities affects when venture capitalists choose to finance companies in another private round in preference to taking the firm public. He shows that investors take firms public when market values are high, relying on private financings when valuations are lower. Seasoned venture capitalists appear more proficient at timing IPOs.

Another consideration may be the venture capitalist's reputation. Gompers (1996) argues that young venture firms have incentives to 'grandstand', or take actions that signal their ability to potential investors. Specifically, young venture firms bring companies public earlier than older one to establish a reputation and successfully raise new funds. Gompers shows that the effect of recent IPOs on the amount of capital raised is stronger for young venture firms, providing them with greater incentives to bring companies public earlier.

Lee and Wahal (2004) propose a variant of the 'grandstanding' hypothesis: they posit that venture firms have an incentive to underprice IPOs. The publicity surrounding a successful offering will enable the venture group to raise more capital than it could otherwise. Lee and Wahal confirm this hypothesis by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
showing a positive relationship between first-day returns and subsequent fund-raising by venture firms. The typical venture firm, however, does not sell its equity at the time of the IPO. After some time, venture capitalists usually return money to their limited partners by distributing their shares. Gompers and Lerner (1998a) examine distributions. After significant increases in stock prices prior to distribution, abnormal returns around the distribution are negative. Cumulative excess returns for the 12 months following the distribution also appear to be negative. While the overall level of venture capital returns does not exhibit abnormal returns relative to the market (Brav and Gompers, 1997), there is a distinct rise and fall around the time of the stock distribution. The results are consistent with venture capitalists possessing inside information and with the (partial) adjustment of the market to that information.

A related research area is venture-fund performance. Kaplan and Schoar (2005) show substantial persistence across consecutive venture funds. General partners that outperform the industry in one fund are likely to outperform in the next fund, while those who underperform in one fund are likely to underperform with the next fund. These results contrast with those of mutual funds, where persistence is difficult to identify.

Cochrane (2005) estimates the returns of venture capital investments. He notes that many analyses of returns focus only on investments that go public, get acquired, or go out of business. Such calculations may produce biased returns by concentrating only on the portfolio's ‘winners’ and outright failures. Cochrane develops a maximum likelihood estimate that uses existing data, but adjusts for these selection biases. While these papers – as well as Gompers and Lerner (1997) and Jones and Rhodes-Kropf (2003) – represent a first step towards understanding these issues, much more work remains to be done.

Future research

While financial economists know much more about venture capital than they did a decade ago, there are many unresolved issues. I highlight here three promising areas.

The rapid growth in the US venture capital market has led institutional investors to look abroad. In a pioneering study, Jeng and Wells (2000) examine the factors that influence venture fund-raising internationally. They find that the strength of the IPO market is an important determinant of venture commitments, supporting Black and Gilson's (1998) hypothesis that the key to a successful venture industry is the existence of robust IPO markets. Jeng and Wells find, however, that the IPO market does not influence commitments to early-stage funds as much as those to later-stage ones. Much more remains to be explored regarding the internationalization of venture capital.

One provocative finding from Jeng and Wells's analysis is that government policy can dramatically affect the health of the venture sector. Researchers have only begun to examine the ways in which policymakers can catalyse the growth of venture capital and the companies in which they invest (Irwin and Klenow, 1996; Lerner, 1999; Wallsten, 2000). Clearly, much more needs to be done in this arena.

A final area is the thorniest: the impact of venture capital on the economy. Demonstrating a causal relationship between innovation, job growth and venture activity is a challenging empirical problem. Kortum and Lerner (2000) examine the influence of venture capital on patented inventions in the United States over three decades, finding that increases in venture capital activity in an industry are associated with significantly higher patenting rates. One dollar of venture capital, they suggest, is three times more

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
likely than one dollar of corporate R&D to stimulate patenting. (Hellmann and Puri, 2000, also explore the impact of venture capital on innovation.) Many research opportunities remain in this arena.

See Also

- entrepreneurship

Bibliography

Admati, A. and Pfleiderer, P. 1994. Robust financial contracting and the role for venture capitalists. *Journal of Finance* 49, 371–402.

Amit, R., Glosten, L. and Muller, E. 1990. Entrepreneurial ability, venture investments, and risk sharing. *Management Science* 36, 1232–45.

Barry, C., Muscarella, C., Peavy III, J. and Vetsuypens, M. 1990. The role of venture capital in the creation of public companies: evidence from the going public process. *Journal of Financial Economics* 27, 447–71.

Berglöf, E. 1994. A control theory of venture capital finance. *Journal of Law, Economics, and Organizations* 10, 247–67.

Bergemann, D. and Hege, U. 1998. Venture capital financing, moral hazard, and learning, *Journal of Banking and Finance* 22, 703–35.

Black, B. and Gilson, R. 1998. Venture capital and the structure of capital markets: banks versus stock markets. *Journal of Financial Economics* 47, 243–77.

Brav, A. and Gompers, P. 1997. Myth or reality? Long-run underperformance of initial public offerings; evidence from venture capital and nonventure capital-backed IPOs. *Journal of Finance* 52, 1791–821.

Chan, Y. 1983. On the positive role of financial intermediation in allocation of venture capital in a market with imperfect information. *Journal of Finance* 38, 1543–68.

Cochrane, J. 2005. The risk and return of venture capital. *Journal of Financial Economics* 75, 3–52.

Cornelli, F. and Yosha, O. 2003. Stage financing and the role of convertible debt. *Review of Economic Studies* 70, 1–32.

Gompers, P. 1995. Optimal investment, monitoring, and the staging of venture capital. *Journal of Finance* 50, 1461–89.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gompers, P. 1996. Grandstanding in the venture capital industry. *Journal of Financial Economics* 42, 133–56.

Gompers, P. and Lerner, J. 1996. The use of covenants: an empirical analysis of venture partnership agreements. *Journal of Law and Economics* 39, 463–98.

Gompers, P. and Lerner, J. 1997. Risk and reward in private equity investments: the challenge of performance assessment. *Journal of Private Equity* 1, 5–12.

Gompers, P. and Lerner, J. 1998a. Venture capital distributions: short- and long-run reactions. *Journal of Finance* 53, 2161–83.

Gompers, P. and Lerner, J. 1998b. What drives venture fundraising? *Brookings Papers on Economic Activity—Microeconomics*, 149–92.

Gompers, P. and Lerner, J. 1999. An analysis of compensation in the U.S. venture capital partnership. *Journal of Financial Economics* 51, 3–44.

Gompers, P. and Lerner, J. 2000. Money chasing deals? The impact of fund inflows on private equity valuations. *Journal of Financial Economics* 55, 281–325.

Hellmann, T. 1998. The allocation of control rights in venture capital contracts. *RAND Journal of Economics* 29, 57–76.

Hellmann, T. and Puri, M. 2000. The interaction between product market and financing strategy: the role of venture capital. *Review of Financial Studies* 13, 959–84.

Hochberg, Y. 2005. Venture capital and corporate governance in the newly public firm. Working paper, Northwestern University.

Hsu, D. 2004. What do entrepreneurs pay for venture capital affiliation? *Journal of Finance* 59, 1805–44.

Irwin, D. and Klenow, P. 1996. High tech R&D subsidies: estimating the effects of Sematech. *Journal of International Economics* 40, 323–44.

Jeng, L. and Wells, P. 2000. The determinants of venture funding: evidence across countries. *Journal of Corporate Finance* 6, 241–89.

Jones, C. and Rhodes-Kropf, M. 2003. The price of diversifiable risk in VC and private equity. Working paper, Columbia University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Kaplan, S. and Strömberg, P. 2003. Financial contract theory meets the real world: an empirical analysis of venture capital contracts. *Review of Economic Studies* 70, 281–315.

Kaplan, S. and Schoar, A. 2005. Private equity performance: returns, persistence, and capital. *Journal of Finance* 60, 1791–823.

Kortum, S. and Lerner, J. 2000. Assessing the contribution of venture capital to innovation. *RAND Journal of Economics* 31, 674–92.

Kraus, T. 2002. Underpricing of IPOs and the certification role of venture capitalists: evidence from Germany's Neuer Markt. Working paper, University of Munich.

Lee, P. and Wahal, S. 2004. Grandstanding, certification and the underpricing of venture capital backed IPOs. *Journal of Financial Economics* 73, 375–407.

Lerner, J. 1994a. The syndication of venture capital investments. *Financial Management* 23, 16–27.

Lerner, J. 1994b. Venture capitalists and the decision to go public. *Journal of Financial Economics* 35, 293–316.

Lerner, J. 1995. Venture capitalists and the oversight of private firms. *Journal of Finance* 50, 301–18.

Lerner, J. 1997. An empirical exploration of a technology race. *RAND Journal of Economics* 28, 228–47.

Lerner, J. 1999. The government as venture capitalist: the long-run effects of the SBIR program. *Journal of Business* 72, 285–318.

Lerner, J. and Schoar, A. 2004. The illiquidity puzzle: theory and evidence from private equity. *Journal of Financial Economics* 72, 3–40.

Marx, L. 1994. Negotiation and renegotiation of venture capital contracts. Working paper, University of Rochester.

Megginson, W. and Weiss, K. 1991. Venture capital certification in initial public offerings. *Journal of Finance* 46, 879–93.

Poterba, J. 1987. How burdensome are capital gains taxes? Evidence from the United States. *Journal of Public Economics* 33, 157–72.

Poterba, J. 1989. Venture capital and capital gains taxation. In *Tax Policy and the Economy*, ed. L. Summers. Cambridge, MA: MIT Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sahlman, W. 1990. The structure and governance of venture capital organizations. *Journal of Financial Economics* 27, 473–524.

Wallsten, S. 2000. The effects of government-industry R&D programs on private R&D: the case of the Small Business Innovation Research program. *RAND Journal of Economics* 31, 82–100.

How to cite this article

Lerner, Josh. "venture capital." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000025> doi:10.1057/9780230226203.1793

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Verri, Pietro (1728 – 1797)

Peter Groenewegen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

balance of production and consumption; consumption taxation; marginalism; mathematical economics; tax incidence; Verri, P.

Article

Italian economist, administrator and philosopher, Verri was born in Milan in 1728, educated in Rome and Parma, served with Austria in the Seven Years War and at this time was introduced to the study of economics by General Henry Lloyd (Venturi, 1978; 1979). His economic writings of the 1760s, such as *Elementi di Commercio* (1760) and the dialogues on monetary disorders in the State of Milan (1762), led to his appointment to a number of positions in the Austrian civil service in Milan. His administrative achievements include the abolition of tax farming (1770) and lowering and simplifying the tariff (1786). From 1764 to 1766 he edited with his brother Alessandro the periodical *Il Caffè*, which attracted contributions on economics from Beccaria and Frisi as well as himself (Verri, 1764). His most important economic publication, *Reflections on Political Economy*, appeared in 1771, went through numerous editions and was translated into French, German and Dutch and more recently into English. Other economic works on monetary and trade questions, including his 1769 pamphlet advocating freedom of the domestic corn trade, contribute to his reputation as a most important 18th-century Italian economist (MuCulloch, 1845, pp. 26–7). More recently he has been noted for inspiring early developments in mathematical economics (Theocharis, 1961, pp. 27–34). He died in 1797.

Verri's *Reflections* is a complete treatise on political economy, reminiscent of Turgot's *Reflections on the Production and Distribution of Wealth* (1766) with its tight, logical framework and division into fairly short sections. Although these cover a wide range of subjects, they are interconnected by the basic theme of the work, the increase in annual reproduction of the nation through trade of surplus product which Verri related to the balance of production and consumption. This ratio or balance is the key concept in Verri's economic analysis, since it not only influences economic growth but also value (it approximates the ratio of sellers to buyers at home and abroad), the rate of interest (it represents thriftiness conditions) and, via its influence on the balance of trade, it also determines national money supply. An excess of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
production over consumption lowers the price level and the rate of interest, expands the money supply, animates industry and facilitates the collection of taxes. Some features of this analysis may be specifically noted. Verri does not appear to have been aware of the importance of capital, as is demonstrated in his general discussion of production (sections 26–8) and his treatment of the interest rate as a monetary phenomenon (sections 14–15). Secondly, his emphasis on supply and demand (used to determine all prices including the rate of interest) combined with references to utility and scarcity in the context of value (section 4) explains why this part of his work has been linked with marginalist economics. The last 11 sections discuss taxation and public finance, including a presentation of five canons of taxation (section 30), a tax incidence analysis arguing against the Physiocratic view that all taxes fall on the landlord (sections 32–3) and a plea for indirect consumption taxation as a fair and administratively easy way to raise revenue. Anti-Physiocratic elements in his economics are not confined to tax issues, but apply to his discussion of special classes (section 24), the importance of agriculture (section 28) and are apparent in his view that free trade should be largely confined to domestic activity (section 40). Verri's *Reflections* were highly regarded when they appeared, and could be found, for example, in Smith's library. His work, though now largely ignored, may therefore have exerted greater influence than is generally believed.

Selected works

1760. *Degli elementi di commercio*. In *Scrittori classici italiani di economia politica*, parte moderna, vol. 17, ed. P. Custodi, Milan, 1804.

1762. *Dialogo sui disordine delle monete nello stato di Milano nel 1762*. In *Scrittori classici italiani di economia polìtica*, parte moderna, vol. 16, ed. P. Custodi, Milan, 1804.

1764. Considerazioni sul lusso. In *Scrittori classici italiani di economia politica*, parte moderna, vol. 17, ed. P. Custodi, Milan, 1804.

1771. *Reflections on Political Economy*. Trans. B. McGilvray and ed. P. Groenewegen, Reprints of Economic Classics, Series 2, No. 4, Sydney: University of Sydney, 1986.

Bibliography

McCulloch, J.R. 1845. *The Literature of Political Economy*. London: LSE reprint, 1938.

Theocharis, R.D. 1961. *Early Developments in Mathematical Economics*. London: Macmillan.

Venturi, F. 1978. Le ‘Meditazioni sulla economica politica’ di Pietro Verri: edizioni, echi e discussioni. *Rivista storica italiana* 90, 530–94.

Venturi, F. 1979. Le avventure del generale Henry Lloyd. *Rivista storica italiana* 91, 369–433.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Groenewegen, Peter. "Verri, Pietro (1728–1797)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000028>
doi:10.1057/9780230226203.1794

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

vertical integration

Michael H. Riordan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Modern economics takes a two-way approach to vertical integration. The theory of the firm approach focuses on how the unified control of successive production and distribution processes changes investment incentives, while the industrial organization approach studies how vertical integration affects the exercise of market power.

Keywords

asset specificity; backward integration; bargaining costs; bilateral vertical contracts; Chicago School; commitment; control rights; double markups; enforceable contracts; exclusive dealing; firm, theory of; foreclosure; forward integration; free-rider problem; hold-up problem; imperfect information; incomplete contracts; industrial organization; market exchange; market power; quasi-rent; raising rivals' costs; relationship-specific assets; transaction costs; variable proportions distortions; vertical integration

Article

Vertical integration is the unified ownership and operation of successive production and distribution processes by a single firm. Backward integration occurs when a manufacturer controls the production of inputs, and forward integration occurs when the manufacturer controls distribution. The alternative (market exchange) is to procure inputs and distribution services from independent suppliers. Vertical integration is a matter of degree, as firms often are only partially integrated in one direction or the other. Vertical integration raises issues for business strategy and public policy. A major theme in the theory of the firm literature is that vertical integration remedies underinvestment in relationship-specific assets due to opportunistic bargaining when contracts are incomplete. Accordingly, vertical integration enhances operational efficiency by improving investment incentives and reducing bargaining costs. Major themes of the industrial organization literature are that vertical integration reduces a firm's procurement or distribution costs, or raises those of its rivals. Accordingly, vertical integration is a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
strategy for competitive advantage. Policy issues concern whether the prevention or regulation of vertical integration improves consumer and social welfare.

Theory of the firm approach

The neoclassical theory of the firm assumes managers choose inputs and outputs to maximize profits subject to a production function, on the assumption that the governance of transactions is costless. The modern theory, in contrast, focuses explicitly on transaction costs, including efficiency losses, arising within and between firms.

The transaction-cost approach views vertical integration as a response to difficulties negotiating and executing market contracts (Coase, 1937; Klein, Crawford and Alchian, 1978; Williamson, 1975; 1985; 1996). The transaction-cost advantages of vertical integration over market exchange are most pronounced when contracts are incomplete, and uncertain future transactions require prior investments in relationship-specific assets for operational efficiency. In these circumstances, market exchange runs afoul of the hold-up problem. Relationship-specific assets by definition are strictly more valuable in a particular transactional relationship than in alternative uses; the difference in use value is called a quasi-rent. Thus asset specificity locks investors into bilateral relationships, while contractual incompleteness exposes them to costly bargaining over quasi-rents. Bargaining costs include failures to adapt transactions efficiently to unfolding circumstances and the direct costs of dispute resolution. Vertical integration improves operational efficiency by replacing dysfunctional bargaining with centralized authority over transactions, but adds bureaucratic costs, including efficiency losses from low-powered managerial incentives. A key hypothesis is that bargaining costs of market exchange rise with asset specificity faster than the bureaucratic costs of vertical integration, leading to two propositions: first, vertical integration is more likely the more important asset specificity is for efficiency; second, vertical integration supports more investment in relationship-specific assets than market exchange (Riordan and Williamson, 1985). Empirical research generally bears out the implied positive correlation between vertical integration and the level of asset specificity (Shelanski and Klein, 1995).

The more formal property-rights approach studies how *ex post* bargaining when contracts are incomplete distorts *ex ante* relationship-specific investments (Grossman and Hart, 1986; Hart and Moore, 1990; Hart, 1995). Ownership confers control rights over the use of non-human assets used in production. While some specific control rights may be contracted away, the residual rights are held by the owners. Furthermore, managers make non-contractable relationship-specific investments to increase the value of these assets. The hold-up problem is manifest because *ex post* bargaining distributes the returns from these investments. Owner-managers who control the non-human assets of a firm undertake relationship-specific investments to the extent that the hold-up problem does not discourage them. Employee-managers, however, have less incentive because owners of the complementary non-human assets appropriate much of the investment returns. Thus vertical integration has mixed effects on managerial incentives. By eliminating the hold-up problem of market exchange, vertical integration improves investment incentives of owner-managers, while converting owner-managers into employees diminishes their incentives. Accordingly, the direction of vertical integration matters. Backward integration enhances investment incentives of downstream managers and degrades managerial incentives upstream, while forward integration has opposite effects. Optimal vertical integration depends on the importance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of relationship-specific investments by managers at each stage of production and distribution. For example, forward integration is predicted when upstream managerial effort is particularly important for operational efficiency. Predictions of the property-rights approach depend sensitively on managers' investment opportunities to improve efficiency, and are difficult to test empirically (Whinston, 2003).

Vertical integration also improves efficiency by reducing information imperfections at the root of bargaining costs (Arrow, 1975; Riordan, 1990). At the same time, the changed information structure diminishes investment incentives of employee-managers by aggravating the hold-up problem. This perspective reconciles with the property-rights approach by interpreting business information as an asset for which the owner has control rights. An open question is how the change in information structure derives from primitive conditions (Hart, 1995).

Vertical integration might be motivated by the pursuit of greater bargaining leverage, rather than just greater efficiency (Bolton and Whinston, 1993). A vertically integrated supplier with scarce capacity over-invests in its downstream unit in order to negotiate better terms from independent customers. The unfortunate effect is to discourage independents' investments in relationship-specific assets. Consequently, vertical integration tends to be excessive from a social welfare perspective.

Industrial organization approach

While the theory of the firm deals mainly with the reasons for vertical integration, industrial organization is more concerned with its effects on competition. Building on the neoclassical theory of the firm, industrial organization studies how market power distorts transactions. Much of this literature presupposes that transactions are governed by uniform prices for inputs and outputs. In this context, the Chicago School approach identifies efficiencies of vertical integration arising from a more profitable exercise of market power, including output expansion resulting from the elimination of 'double markups' when vertically related firms each exercise market power, the correction of 'variable proportions distortions' when independent downstream firms substitute towards more competitively supplied inputs, and the prevention of free-riding on point-of-sale services (Perry, 1989).

The post-Chicago approach, by contrast, studies how foreclosure resulting from vertical integration reduces competition and raises rivals' costs (Ordover, Saloner and Salop, 1990; Riordan, 1998; Salinger, 1988; Salop and Scheffman, 1987). Foreclosure might drive up procurement costs or deny scale economies. Accordingly, an appropriate policy analysis weighs efficiencies against possible anti-competitive effects of vertical integration (Riordan and Salop, 1995). The post-Chicago approach demonstrates conditions for anti-competitive foreclosure more rigorously than the traditional foreclosure doctrine attacked by the Chicago School (Bork, 1978).

Another recent approach to vertical foreclosure studies the commitment problem of a supplier with market power who deals with customers bilaterally (Hart and Tirole, 1990; Rey and Tirole, 2007). Multilateral contracts involving a supplier and several downstream rival customers might be prevented by antitrust policy, or be unenforceable due to monitoring problems. Allowing more sophisticated contracting than just uniform pricing, the privacy of bilateral vertical contracts nevertheless fosters opportunism. A supplier has an adverse incentive to negotiate individual contracts that disadvantage other rival customers. Consequently, equilibrium supply contracts with favourable terms result in more downstream competition than would maximize total profits. Vertical integration restores monopoly

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
power because the vertically integrated supplier is loath to set terms that hurt its own downstream division. The vertically integrated supplier offers less favourable terms to downstream rivals, raising their variable costs and causing higher downstream prices. Enforceable contracts with multilateral elements, such as exclusive dealing, also improve profits. Moreover, a vertically integrated firm has a greater incentive to enter into exclusive supply deals that foreclose upstream competitors and effectively cartelize a downstream industry (Chen and Riordan, 2007). Such non-efficiency motives for vertical integration sometimes are contrary to consumer and social welfare, but are inconsequential if market power is absent.

See Also

- firm boundaries (empirical studies)
- hold-up problem
- incomplete contracts

Bibliography

Arrow, K.J. 1975. Vertical integration and communication. *Bell Journal of Economics* 6, 173–83.

Bolton, P. and Whinston, M.D. 1993. Incomplete contracts, vertical integration, and supply assurance. *Review of Economic Studies* 60, 121–48.

Bork, R.H. 1978. *The Antitrust Paradox: A Policy at War with Itself*. New York: Basic Books.

Chen, Y. and Riordan, M.H. 2007. Vertical integration, exclusive dealing, and *ex post* cartelization. *RAND Journal of Economics* 38, 1–21.

Coase, R.H. 1937. The nature of the firm. *Economica* 4, 386–405.

Grossman, S.J. and Hart, O.D. 1986. The costs and benefits of ownership: a theory of vertical and lateral integration. *Journal of Political Economy* 94, 691–719.

Hart, O. 1995. *Firms, Contracts and Financial Structure*. New York: Oxford University Press.

Hart, O. and Moore, J. 1990. Property rights and the nature of the firm. *Journal of Political Economy* 98, 1119–58.

Hart, O. and Tirole, J. 1990. Vertical integration and market foreclosure. *Brookings Papers on Economic Activity: Microeconomics*, 205–76.

Klein, B., Crawford, R.A. and Alchian, A.A. 1978. Vertical integration, appropriable rents, and the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
competitive contracting process. *Journal of Law and Economics* 21, 297–326.

Ordover, J.A., Saloner, G. and Salop, S.C. 1990. Equilibrium vertical foreclosure. *American Economic Review* 80, 127–42.

Perry, M.K. 1989. Vertical integration: determinants and effects. In *Handbook of Industrial Organization*, vol. 1, ed. R. Schmalensee and R. Willig. Amstersdam: North-Holland.

Rey, P. and Tirole, J. 2007. A primer on foreclosure. In *Handbook of Industrial Organization*, vol. 3, ed. M. Armstrong and R.H. Porter. Amsterdam: North-Holland.

Riordan, M.H. 1990. What is vertical integration? In *The Firm as a Nexus of Treaties*, ed. M. Aoki, B. Gustafsson and O.E. Williamson. London: Sage.

Riordan, M.H. 1998. Anticompetitive vertical integration by a dominant firm. *American Economic Review* 88, 1232–48.

Riordan, M.H. and Salop, S.C. 1995. Evaluating vertical mergers: a post-Chicago approach. *Antitrust Law Journal* 63, 513–68.

Riordan, M.H. and Williamson, O.E. 1985. Asset specificity and economic organization. *International Journal of Industrial Organization* 3, 365–78.

Salinger, M.A. 1988. Vertical mergers and market foreclosure. *Quarterly Journal of Economics* 103, 345–56.

Salop, S.C. and Scheffman, D.T. 1987. Cost-raising strategies. *Journal of Industrial Economics* 36, 19–34.

Shelanski, H.A. and Klein, P.G. 1995. Empirical research in transaction cost economics: review and assessment. *Journal of Law, Economics, and Organization* 11, 335–61.

Whinston, M.D. 2003. Transaction cost determinants of vertical integration. *Journal of Law, Economics, and Organization* 19, 1–23.

Williamson, O.E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

Williamson, O.E. 1985. *The Economic Institutions of Capital*. New York: Free Press.

Williamson, O.E. 1996. *The Mechanisms of Governance*. New York: Oxford University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Riordan, Michael H. "vertical integration." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000029>
doi:10.1057/9780230226203.1795

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Vickrey, William Spencer (1914 – 1996)

Richard Arnott
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

William Spencer Vickrey was awarded the 1996 Nobel Prize in Economics ‘for his fundamental contributions to the economic theory of incentives under asymmetric information’. While best known as the father of auction theory, he made important contributions on a broad range of topics including social choice, marginal cost pricing, the design of tax systems, transportation economics, urban economics, and macroeconomics.

Keywords

asymmetric information; auctions (theory); congestion pricing; cumulative averaging taxation; efficient allocation; estate taxes; Gibbard–Satterthwaite theorem; Groves–Clark–Vickrey mechanisms; land use; land value taxation; marginal cost pricing; Mirrlees, J. A.; natural rate of unemployment; optimal income tax; philanthropy; philosophy and economics; Ramsey pricing; responsive pricing; revelation of preferences; second best; Shoup, C. S.; site value taxation; social choice; strategic misrepresentation; transport economics; urban economics; user fees; veil of ignorance; Vickrey second-price auction; Vickrey, W. S.

Article

William Spencer Vickrey was awarded the 1996 Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, jointly with James A. Mirrlees, ‘for [his] fundamental contributions to the economic theory of incentives under asymmetric information’ (Royal Swedish Academy of Sciences, 1996). His most influential papers, as well as a comprehensive bibliography, are contained in *Public Economics: Selected Papers by William Vickrey* (Arnott et al., 1994). Articles that pay tribute to his contributions include Arnott (1998), Drèze (1995), Harriss (2000), and Laffont (2003).
Vickrey was born in Victoria, Canada, where his maternal grandfather had built a group of department

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
stores. Towards the end of the First World War, his father, a Congregational minister, became actively involved in the relief of Armenian and Greek refugees from the Ottoman Empire. In conjunction with this work, the family moved to New York and later Switzerland. Vickrey attended high school in Scarsdale, NY, and Phillips Andover Academy. He received his bachelor's degree in engineering and mathematics from Yale University in 1935. He then moved to Columbia University for graduate studies in economics. After two years of course work, he went to Washington, DC, to work for the National Resources Committee, in a group pioneering studies on the structure of the US economy. During the Second World War, as a conscientious objector, he served in the Division of Tax Research of the Department of the Treasury, working on broad macroeconomic issues related to war finance and more specific issues of tax structure. In 1946 he returned to Columbia to teach and to complete his doctoral dissertation, published as *An Agenda for Progressive Taxation* (1947). Apart from sabbaticals and missions abroad, he remained at Columbia until his death.

Throughout his career, Vickrey had considerable practical policy experience. As part of Carl Shoup's team, he helped design several countries' tax systems, including Japan's after the Second World War. He also advised many public utilities, and even introduced skip-stop scheduling to the Indian rail service. But he was never a major player on the policy front. His legacy is his body of publications. He published eight longer works, including graduate textbooks in microeconomics and macroeconomics, three technical monographs, and two co-authored country tax system studies, as well as his thesis. Apart from his thesis, however, he is best known for his some 200 papers.

Though covering an unusually broad range of topics, his papers are consistent in theme and style. While his choice of topics stemmed from social and moral concerns, his treatments of them stressed improvements in resource allocation. 'Greater efficiency for the common good' would be an appropriate slogan for his work. His style of writing and reasoning is idiosyncratic and paradoxical. Most of his papers advocate specific policy innovations. To reach a broader audience, he developed his ideas primarily verbally, and with literary flair, but the precision and sophistication of his economic reasoning largely defeated this purpose. He also tended to emphasize practical issues of policy implementation, while presenting in an offhand manner the novel theoretical and conceptual insights for which the papers have been remembered. Many of his policy recommendations, though derived with impeccable logic, were impractical at the time he proposed them. (Technological advances have since rendered some, such as auto congestion pricing and land value taxation, more practical.) These smaller paradoxes can be resolved by understanding Vickrey as a social crusader with a theorist's cast of mind. The larger paradox is that the tension between crusader and theorist was the source first of his creativity, and then of the neglect for many years of much of his work, and ultimately of the distinction of his intellectual legacy. Many of his ideas were overlooked until they were independently discovered many years later, while others lay dormant until their time had come.

Vickrey's major contributions lie in four areas: social choice and resource allocation mechanisms, taxation, marginal cost pricing, and urban transportation.

Vickrey is best known as the father of auction theory, due to his seminal paper, 'Counterspeculation, Auctions, and Competitive Sealed Tenders' (1961). The question posed by Vickrey in that paper is how to achieve efficiency in resource allocation with a small number of buyers or sellers under asymmetric information. He presented and analysed two classes of mechanisms that circumvent the strategic misrepresentation of costs and preferences. The first is auctions. Consider the simplest auction in which there is a single item for sale, for which bidders have different private valuations. Efficiency entails the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
item being sold to the bidder with the highest valuation. If the item is sold to the highest bidder at his bid, then each bidder has an incentive to bid less than his valuation, since if he bids his valuation he gains no surplus from purchase of the item. In deciding on his bid, each bidder must guess others' valuations, and there is no guarantee that the item will be sold to the bidder with the highest valuation. Suppose instead that the item is sold to the highest bidder but at the *second* highest bid (the Vickrey second-price auction). Whatever other bidders do, if a particular bidder bids more than his valuation he will win more often but only when the second-highest bid, and therefore the price he pays, exceeds his valuation, while if he bids less than his valuation he will win less often and only when the winning bid falls short of his valuation. Since bidding one's private valuation is the dominant strategy, the item should go to the bidder with the highest valuation, achieving efficiency. Auction theory has developed from this insight. Auctions are now extensively used in the allocation of goods with a small number of buyers; timber and drilling rights, bandwidth, Treasury bills and sealed bid tenders are well-known examples. Vickrey's paper also investigated a class of demand-revealing mechanisms – now known as Groves–Clark–Vickrey mechanisms – that elicit truthful revelation of preferences, for pure public goods for example.

Vickrey made several other contributions to the literature on social choice and resource allocation mechanisms. In 'Measuring Marginal Utility by Reactions to Risk' (1945), he provided the first statement of social choice based on the maximization of expected utility behind the veil of ignorance, which was independently stated by Harsanyi (1955) a decade later, and also the first formulation of the optimal income tax problem, which was not solved until a quarter-century later by James Mirrlees (1971). 'Utility, Strategy, and Social Decision Rules' (1960) provides a masterful survey of social choice theory as of that date and conjectures what is now known as the Gibbard–Satterthwaite theorem.

The efficiency of marginal cost pricing in general and of short-run marginal-cost pricing in particular were well understood when Vickrey was a graduate student. Vickrey's contributions were in communicating the breadth of application of the principles and in conceiving ingenious technological schemes for their implementation. From the early 1950s he was a strong advocate of responsive marginal cost pricing, whereby the current price reflects the current realization of stochastic demand and supply; for instance, he proposed varying the parking meter rate on a city block according to the meters' realized occupancy rate (1959) and dealing with airline overbooking via responsive pricing (1972). His crusading for congestion pricing in transportation, for site value taxation (1970), and for the extended application of user fees to finance local public services (1963) are examples of his advocacy of marginal cost pricing in novel contexts.

His major contributions to the theory of taxation derived from his experience in the Department of the Treasury during the Second World War. All are contained in his thesis, which was a tour de force. The goal of the thesis was the comprehensive design of a practical, coherent, fair, and efficient tax system. At the time, a steeply progressive income tax was the primary source of federal tax revenue. Rates had been increased sharply to generate the revenue needed to finance the war effort. The combination of steep progression and high rates encouraged taxpayers to devote considerable effort to altering the timing of expenditures and receipts in order to average income. To eliminate this waste, Vickrey proposed cumulative averaging, a method of taxing individuals on their discounted lifetime incomes, with a minimum of accounting. A steeply progressive estate tax was also in place, which wealthy individuals avoided through generation-skipping trusts. Vickrey proposed an integrated successions tax, which retained the progressivity of the tax while reducing the incentives for complex tax avoidance

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
schemes. Since then, income and estate tax rates have been lowered and made less progressive, mitigating the problems that Vickrey's proposed reforms addressed. His contributions lie not so much in his specific proposals, however, as in his conception of redesigning the tax system from basic principles. How much influence his book had on the Carter Commission in Canada, the Meade Committee in the UK, or the Reagan tax reforms is hard to say, but they are in the same spirit.

While best known for his auctions paper, Vickrey was also the pre-eminent transport economic theorist of his generation. As a transport economist, he is famous for his 50-year-long advocacy of auto congestion pricing (of charging drivers for the externality cost each imposes on other drivers by slowing them down), and in North America at least is known as the father of congestion pricing. After many years of political resistance, urban auto congestion pricing is slowly being adopted, first in Singapore and more recently in London and Stockholm. His first work on the subject (1955) was a proposed fare structure for New York City's subway system, based on marginal cost pricing principles. His empirical research on the project entailed travelling the subway system, stopwatch in hand, while his discussion of deviations from first-best marginal cost pricing to take into account equity concerns and the transit authority's budgetary constraints anticipates the theory of the second best. His second work (1959) detailed an automobile congestion-pricing scheme for downtown Washington, DC. A schedule gives the prices of traversing major intersections by time of day. Each car is equipped with a transponder. When the car enters an intersection, its transponder sends a signal to a roadside receptor and is conveyed to a central computer, which adds the appropriate charge to the car's bill. The theoretical work on the project included an independent derivation of Ramsey pricing with a marginal cost of public funds. His later work in urban transport economics is noteworthy in two respects. He, more than any other urban transport economist, grappled with the complex physics of auto congestion. He also introduced the bottleneck model of traffic congestion (1969), the first analytically tractable model of equilibrium rush-hour traffic dynamics. Each commuter decides when to leave home in the morning, trading off schedule inconvenience against congestion delay. Congestion takes the form of a queue behind a bottleneck of fixed flow capacity, with the queue length (and hence the departure time distribution) evolving to achieve equilibrium.

His other work spans a diversity of topics. Viewing unemployment as a tragic waste of human resources, he wrote many macroeconomic papers arguing against a natural rate of unemployment and for Keynesian macroeconomic stabilization. He also made important contributions to urban economics, most noteworthy of which are pioneering papers on traffic congestion and land use (Solow and Vickrey, 1971) and on the Henry George theorem (1977) – which states that efficient spatial economies can be decentralized via marginal cost pricing, with land rents being used to cover the deficits deriving from the economies of scale underlying agglomeration. His miscellaneous papers cover such topics as game theory, student loans, gerrymandering, international dispute resolution, cost-of-living indices, equivalence scales and sorting theory. One paper on the economics of traffic accidents (1968), another on the economics of philanthropy (1962), and another on economics and philosophy (1950) have been influential. The last of these papers provides insight into the moral purpose underlying Vickrey's work.

Prior to his receipt of the Nobel Prize, Vickrey's work, apart from his justly celebrated auctions paper, was not well known to most economists. But Vickrey is much more than just a ‘one-paper economist’. The same intellectual qualities that spawned the auctions paper are evident in the rest of his work. Perhaps the whole of the rest of his work is greater than the sum of the individual papers, demonstrating the wealth of ideas that are generated when a brilliant economic theorist applies his creativity to devising

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
solutions to practical public policy problems.

See Also

- auctions (theory)
- congestion
- land tax
- marginal and average cost pricing
- progressive and regressive taxation
- utilitarianism and economic theory

Selected works

1945. Measuring marginal utility by reactions to risk. *Econometrica* 13, 319–33.

1947. *An Agenda for Progressive Taxation*. New York: Ronald Press.

1950. Ethics and economics: an exchange of questions between economics and philosophy. In *Goals of Economic Life*, ed. A.D. Ward. New York: National Council of Churches.

1955. A proposal for revising New York's subway fare structure. *Journal of Operations Research Society of America* 3, 38–68.

1959. Statement on the pricing of urban street use. *Hearings: US Congress, Joint Committee on Metropolitan Washington Problems* 11(November), 466–77.

1960. Utility, strategy, and social decision rules. *Quarterly Journal of Economics* 74, 507–35.

1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16, 8–37.

1962. One economist's view of philanthropy. In *Philanthropy and Public Policy*, ed. F.G. Dickinson. New York: NBER.

1963. General and specific financing of urban services. In *Public Expenditure Decisions in the Urban Community*, ed. H.G. Schaller. Washington, DC: Resources for the Future.

1968. Automobile accidents, tort law, externalities, and insurance: an economist's critique. *Safety: Law and Contemporary Problems* 33, 464–87.

1969. Congestion theory and transport investment. *American Economic Review* 59, 251–60.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1970. Defining land value for taxation purposes. In *The Assessment of Land Value*, ed. D.H. Holland. Madison, WI: University of Wisconsin Press.

1971. (With R. M. Solow.) Land use in a long, narrow city. *Journal of Economic Theory* 3, 403–47.

1972. Airline overbooking: some further solutions. *Journal of Transport Economics and Policy* 6, 257–70.

1977. The city as a firm. In *The Economics of Public Services*, ed. M.S. Feldstein and R.F. Inman. New York: Wiley.

Bibliography

Arnott, R. 1998. William Vickrey: Contributions to public policy. *International Tax and Public Finance* 5, 93–113.

Arnott, R., Arrow, K., Atkinson, A. and Drèze, J., eds. 1994. *Public Economics: Selected Papers by William Vickrey*. Cambridge: Cambridge University Press.

Drèze, J.H. 1995. Forty years of public economics: a personal perspective. *Journal of Economic Perspectives* 9(2), 111–30.

Harriss, C.L. 2000. William Spencer Vickrey, 1914–1996: Nobel Laureate in Economics. *Economic Journal* 110, 708–19.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63, 309–21.

Laffont, J.-J. 2003. William Vickrey: A pioneer in the economics of incentives. In *Nobel Lectures, Economics, 1996–2000*, ed. T. Persson. Singapore: World Scientific Publishing Co.

Mirrlees, J.A. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38, 175–208.

Royal Swedish Academy of Sciences. 1996. Press Release, 8 October. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1996/press.html, accessed 18 January 2007.

How to cite this article

Arnott, Richard. "Vickrey, William Spencer (1914–1996)." The New Palgrave Dictionary of Economics.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000073>
doi:10.1057/9780230226203.1796

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Vind, Karl (1933 – 2004)

Hans Keiding
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

control theory; cores; exchange equilibrium; expected utility hypothesis; fairness; general equilibrium; Lyapunov theorem; mathematical economics; optimal control; statistics; utility representations; Vind, K.; Walrasian equilibria

Article

Karl Vind was born in a small provincial town in Denmark on 3 April 1933. His mother died when he was only a few years old, and his father was often absent from home, so Karl and his two brothers were taken care of by relatives. The family reports early interest and skill in economics and mathematics.
Karl Vind finished his school years in 1951. He studied economics at the University of Copenhagen, and attended lectures in mathematics by Werner Fenchel (known for his contributions to the theory of convexity). He graduated in 1958, and after finishing military service was employed at the Faculty of Social Sciences; he also had a position as scientific researcher at what was later known as the Institute of Economics. His future scientific orientation was formed in the years 1962–3, which he spent as a Rockefeller Fellow at the University of California (Berkeley), where he was inspired by the highly fertile research environment around Gerard Debreu. He returned to Berkeley as visiting associate professor from 1964 to 1966. While at Berkeley Karl Vind was married in 1962 to Anni (Mortensen); they had sons named Lars and Jacob and adopted a daughter named Dorthe.
After his return to Copenhagen in 1966 Karl Vind obtained a position as professor in economics, later changed to a chair in mathematical economics, which he held until his retirement in April 2003 at the age of 70. He spent several long periods in Berkeley as well as at the Center for Operations Research and Analysis (CORE) in Louvain. After retirement, he retained his office at the institute and participated in its everyday activities until his death in July 2004 after a short illness.
Karl Vind's published research covers many fields – international trade theory, control theory, general equilibrium, game theory, the theory of choice under uncertainty – so that his publication record is consistent with his own interpretation of mathematical economics as ‘the derivative of economic theory’ – what is mathematical economics today becomes economic theory tomorrow. However, most of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
topics he studied engaged him over long periods; some of the results appearing in the later years were at least partially obtained in the early years of his career.

In the years after graduating, Karl Vind had been interested in mathematical statistics and control theory, something which is witnessed by his work on optimal control with jumps in the state variables (1967). He never returned to control theory, but it usefully inspired him to apply Lyapunov's theorem, which at that time was known to control theorists but not to researchers in general equilibrium. Vind demonstrated that it could be used to show the equivalence of core and Walrasian equilibria in large economies (1964). This was a major breakthrough at the time, achieved independently by Aumann (1964). The approach using Lyapunov's theorem was innovative and offered a new approach to modelling large economies.

Also from this period is the short piece on the core of an exchange economy (1965), which pioneers the extension of the results obtained for economies with infinitely many agents to economies with a finite number of agents. Vind's result does not go all the way to establishing a connection between core and equilibrium, something which was achieved only several years later. This later development might perhaps have been simpler and faster if researchers had followed Vind's early approach; he had the bad luck of being ahead of his time.

In the following years, Vind's published research dealt with extensions of the general equilibrium model in several directions. An example is the paper written with David Schmeidler (1972) on fair net trades, proposing a new approach to the concept of fairness as well as an elegant formalism. In much of his work from this period Karl Vind was concerned with the structural properties of exchanges. His concept of an exchange equilibrium was intended to capture the essential properties of trade in markets. He was, however, not quite satisfied with the initial formulations of the exchange equilibrium, which were never published. After several reformulations the concept appeared in 1983 as ‘equilibrium with coordination’.

In the late 1960s Karl Vind started on another project, dealing with utility representations of preferences, which remained at the draft stage until it was finally published as a monograph in 2003. His work on the so-called mean groupoids was inspired by the need to extend the general equilibrium model to include time and uncertainty, which at that time seemed to call for specific functional forms of utility representations. Vind realized that there was a common structure behind utility representations over time and under uncertainty, related to the operation of taking mixtures of consumption programmes, and this led Vind's theory of mean groupoids. The work had already taken shape as a draft for a research monograph around 1970, but its final publication was considerably delayed, partly for practical reasons and partly due to the emergence of new results from Vind himself and others. Some of these had to do with the extension of the expected utility hypothesis to preferences that are not necessarily complete, one of Karl Vind's later research projects.

As a researcher, Karl Vind remained active after retirement as an organizer of and participant in scientific meetings and seminars. His influence on the mathematical economics profession goes beyond his published work, since he took great pleasure in following the work of other researchers, in particular young ones, who received valuable suggestions from a person genuinely interested in their work. Due to this aspect of his scientific activity, he has had a lasting influence on the development of the field.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- mathematics and economics

Selected works

1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5, 165–77.

1965. A theorem on the core of an economy. *Review of Economic Studies* 32, 47–8.

1967. Control systems with jumps in the state variables. *Econometrica* 35, 273–7.

1972. (With D. Schmeidler.) Fair net trade. *Econometrica* 40, 637–42.

1983. Equilibrium with coordination. *Journal of Mathematical Economics* 12, 275–85.

2003. *Independence, Additivity, Uncertainty*. Berlin: Springer Verlag.

Bibliography

Aumann, R. 1964. Markets with a continuum of traders. *Econometrica* 32, 265–90.

How to cite this article

Keiding, Hans. "Vind, Karl (1933–2004)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000063> doi:10.1057/9780230226203.1797

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Viner, Jacob (1892 – 1970)

Henry W. Spiegel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bentham, J.; Chicago School.; cost theory; dumping; economic development; envelope curve; history of economic thought; international trade theory; kinked demand curve; Knight, F. H.; Marshall, A.; Mill, J. S.; monopolistic competition; partial equilibrium; religion and economics; Taussig, F. W.; trade creation; trade diversion; utilitarianism; Viner, J.

Article

Jacob Viner, the economic theorist and historian of economic thought, was born and raised in Montreal, the son of immigrant parents from eastern Europe. As an undergraduate he attended McGill University, where he was taught economics by Stephen Leacock, the famous humorist. Leacock used texts by Mill and Walker, Milk and Water, as the students referred to them, showing ‘good judgment’ according to an account that Viner gave later in life. For graduate work he went to Harvard, where he earned a Ph.D. in 1922. He was a student and eventually became a close friend of Frank W. Taussig, the well-known authority on economic theory and international economics. At that time and during the earlier part of Viner's career he and Taussig were rare specimens in what was, except for a very few others, essentially a ‘wasp’ establishment. But in other respects their background was quite different. Viner was a self-made man who had emancipated himself from the immigrant quarter of Montreal, while Taussig was born into a patrician family with wealth and native culture.

Taussig's specialities were the fields to which Viner himself was drawn and in which he earned great distinction, in addition to his perhaps even more distinguished work in the history of economics, where his accomplishments were almost without rival.

During the two world wars, during the Great Depression, and on and off at other times, Viner did consulting and other work in Washington, but he was foremost an academic, who taught at the University of Chicago in 1916–17 and from 1919 to 1946, when he went to Princeton and taught there until his retirement in 1960. Viner advanced rapidly at Chicago, where the department then was headed by J.M. Clark, and he became a full professor at age 32. A few decades earlier, in the same department, Veblen had risen to the rank of assistant professor only at age 43. But Veblen had defied convention

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
both in his writings and personal life.

Viner's tenure at Chicago coincided in part with his editorship of the *Journal of Political Economy* for a period of 18 years. Most of the time the post was held jointly with Frank H. Knight, who, after having earlier spent two years at Chicago, returned to it in 1927. Both men imprinted on the journal the mark of their own great gifts.

Viner's contributions to economic theory and the history of economic thought are embodied in periodical articles that were reprinted in book form in 1958 under the title *The Long View and the Short*. His contributions to general theory consist principally of two remarkable articles, one published in 1921 and the other ten years later. Of the two, the second on ‘Cost Curves and Supply Curves’ (Viner, 1958, pp. 50–78) made an immediate and powerful impact on the profession. Written, as it was, by a then well-established scholar, it contained virtually the whole of the modern exposition, graphic and otherwise, of the theory of cost, including the envelope curve, about which Viner had a legendary dispute with his mathematically more proficient Chinese draftsman. It also contained, perhaps for the first time in print, the words ‘marginal revenue’. All this matter eventually entered into the elementary textbooks. Viner's accomplishment paralleled that of Knight, whose graphic portrayal of the theory of production in *Risk, Uncertainty and Profit* (1921) likewise entered into the mainstream of economic theory and became the basis for the textbook treatment of the matter. Among the two great scholars there was forged a substantial portion of partial equilibrium analysis as it evolved during the first half of the 20th century.

Viner's earlier article, ‘Price Policies: The Determination of Market Price’, published in 1921 and covering barely five pages in the reprint of 1958, was in some respects an even more dazzling achievement than the later and much better-known one. Five years ahead of Sraffa, six years ahead of the publication of Joan Robinson's and Chamberlin's books on the subject, Viner developed here, in a short paragraph, the outlines of the theory of monopolistic competition. He writes of inflexible prices, ‘differentiation’ of products, advertising, non-price competition and other characteristics of markets that are neither fully competitive nor completely monopolistic. In such markets producers may succeed in creating a special demand for their products. They can then to some extent determine prices independently of the prices charged by their competitors and still maintain their sales (Viner, 1958, pp. 5–6). In the same context Viner also developed, in a few sentences, the theory of what became later known as the kinky demand curve, 18 years ahead of Sweezy's article on the subject.

These were indeed path-breaking contributions, but their existence was virtually ignored until Viner's article was reprinted in 1958. The place of the original contribution – L.C. Marshall, ed., *Business Administration*, University of Chicago Press, 1921 – was not exactly obscure but elusive nevertheless from the standpoint of a reader looking for innovations in economic theory. Chamberlin did not mention Viner in the bibliographies that he appended to successive editions of his book and which eventually listed around 1,500 items. As regards the kinky demand curve, there is no reference to Viner in Sweezy's article in the *Journal of Political Economy* for 1939, of which Viner then was the co-editor, nor in Stigler's critique published in the same journal in 1947. All this is an unresolved puzzle. No one knows why Viner never developed more fully the ideas sketched in his brief article of 1921 and why he, who in other contexts did not shy away from announcing his priority, remained silent about this one. The ideas surely were his own and not derived from an oral tradition at Harvard, whose only potential fount, Allyn Young, came to Harvard only in 1920, when Viner was already teaching at Chicago. He could not very well have anticipated unfriendly criticism, because the Chicago of the 1920s, where J.M. Clark had a senior position, was not the Chicago of the later so-called Chicago School, all of whose leaders,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
beginning in the mid-1940s, voiced disapproval of the theory of monopolistic competition.

Viner not only had an analytical mind that was stocked with original ideas, but combined with this a stupendous book learning that within the scope of the humanities and social sciences, and especially their history, was virtually universal and gave special depth to his studies in the history of economics. He was perhaps not as scintillating a writer as Schumpeter, nor did he turn out, as did Schumpeter, a comprehensive treatise on the subject, but his work, scattered in periodical articles, contains far more reliable and judicious interpretations of such matters as utilitarianism, and classical and Marshallian economics. The most important of Viner's articles on the history of economics were reprinted in Part II of the collection published in 1958. Their coverage extends all the way from the mercantilists to Marshall and Schumpeter. The essay on mercantilist thought shows the mercantilists in pursuit both of power and wealth as ultimate ends of national policy. Another on Adam Smith demonstrates, among other matters, that Smith was not a doctrinaire advocate of laissez-faire, a quality that he shares with Viner. Smith was a favoured subject of Viner's studies, and in 1965 he contributed an introduction of 145 pages to a new edition of Rae's *Life of Adam Smith*, the standard biography. An essay about the utility concept in value theory defends the concept against its critics. Writing about Bentham and J.S. Mill (1949), Viner clarifies the meaning of the former's hedonic calculus and by restricting it to comparisons between pain and pleasure contributes to the rehabilitation of this concept, for which, he believes, an idea of Benjamin Franklin's may have been the inspiration. Mill and Marshall are both viewed in their Victorian setting. The former's *Principles*, a combination of ‘hard-headed rules and utopian aspirations’, was ‘exactly the doctrine that Victorians of goodwill yearned for’. Marshall fitted into the Victorian age that was complacent about the present and optimistic with respect to future progress.

Except for the collection of his articles published in 1958 and two posthumous publications, all of Viner's books are about international economics, with a collection of his articles in this field, titled *International Economics*, published in 1951. His work in international economics covers virtually all its phases – theory, history of thought, and policy – with occasional use of empirical material. His earliest book, *Dumping* (1923), contained the first comprehensive and systematic study of this subject. It was followed a year later by Viner's doctoral dissertation on *Canada's Balance of International Indebtedness, 1900–1913*, which was written on the suggestion of Taussig, who directed a number of related empirical studies designed to demonstrate the operation of the balance-of-payments adjustment process. In 1937 there was published Viner's masterwork, *Studies in the Theory of International Trade*, which blends in an inimitable manner theoretical analysis and erudite doctrinal history. Its aim was to trace the evolution of the modern theory of international trade. It starts out with the mercantilists and continues with the bullionist controversies, the currency school–banking school controversy, the international mechanism of adjustment, and the doctrine of the gains from trade. In Viner's view, the comparative-cost doctrine is dependent on a real-cost theory of value rather than on opportunity cost. While this view was not in tune with the time, there are many forward-looking sections in the book, including references to a lecture given by Viner in 1931 in which later models of Lerner, Leontief and Hicks were anticipated.

In 1950 Viner published *The Customs Union Issue*, which contained the distinction between trade creation and trade diversion, the starting point of later discussions of the matter. Viner's articles on *International Economics*, collected in 1951, start with one on the most-favoured-nation clause and end with an essay on the economic foundations of international organizations. Many of the articles are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
indispensable for the study of the policy issues of the time. In 1952 Viner made his contribution to the emerging field of economic development in a book on *International Trade and Economic Development*. In this work he took a far less favourable view of a number of public policies designed to accelerate economic development than was commonly held at that time. He refused to identify agriculture with poverty, stressed that industrialization was more often a consequence than a cause of prosperity, and placed the main burden of promoting development on the underdeveloped country itself.

Viner had for long been interested in theological ideas, especially of the more remote past, and after his retirement he started out on a project designed to explore the relationship between religious and economic thought. This great project proved open-ended. After Viner's death only two fragments were published, one on *The Role of Providence in the Social Order* (1972), and the other on *Religious Thought and Economic Society* (1978). The first of these works is an original accomplishment that traces the derivation of a number of economic ideas from theological precedents, for example, the theory of international trade that is grounded in differences in factor endowments, Smith's invisible hand, and the providential origin of social inequality that was claimed in the past. The second work is written along more conventional lines and reviews the economic doctrines of the Fathers of the Church, of the Scholastics, secularizing tendencies in later Catholic social thought, and Protestantism and the rise of capitalism. This last chapter contains a critical analysis of the Weber–Tawney thesis of the Calvinist origin of capitalism.

To place Viner's work into its proper historical setting, a word is in order about his relation to the Chicago School. A common conception takes his membership or leadership in this school for granted, but this view is mistaken. Viner himself said that much in a remarkable letter to Patinkin written shortly before his death (Patinkin, 1981, p. 266; the letter is also reproduced by Reder, 1982, p. 7). It must be remembered that at the time when Viner taught at Chicago, the designation ‘Chicago School’ was not yet a commonly used term. To be sure, Viner's views about laissez-faire, Keynesian economics and government intervention had something in common with the views held by representatives of the Chicago School, but on the whole he was a more pragmatic thinker and more aware of the need for qualification and consideration of circumstances of time and place. Moreover, Viner, from whom stems the definition ‘ economics is what economists do’, would not have felt comfortable within the confines of a school, especially of one that at times has come close to defining economics as the study of competitive markets. The early leaders of what later became known as the Chicago School were Henry Simons and Knight, not Viner. Like no one else, Knight had a charismatic appeal that yielded conversions to libertarianism in his classroom – James Buchanan has testified to this – and that made him the more likely founder of a school. It is significant also that Viner, and, for that matter, Knight too, urged deficit spending during the Great Depression. Viner called the plea for an annual balanced budget a mouldy fallacy (Viner, 1933, p. 129). He was critical of Hayek's libertarianism (Viner, 1961). He denied that competition was both a norm and normal, pointing out instead that

> monopoly is so prevalent in the markets of the western world today that discussions of the merits of the free competitive market as if that were what we are living with or were at all likely to have the good fortune to live with in the future seem to me academic in the only pejorative sense of that adjective.

He also insisted that ‘no modern people will have zeal for the free market unless it operates in a setting

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

of “distributive justice” with which they are tolerably content’ (Viner, 1960, pp. 66, 68). (The article in which Viner developed these ideas was ostensibly an exposition on the *rhetoric* of laissez-faire, an early exercise in an approach that D.N. McCloskey was to apply on a wider scale more than a quarter century later.) Against Friedman Viner supported discretionary monetary management rather than conduct in conformity with a ‘rule’ (Viner, 1962). And, last but not least, it was Viner who created the substance of the theory of monopolistic competition, which in a peculiar dialectic was later to become the target of the Chicago School.

See Also

- Chicago School

Selected works

1923. *Dumping: A Problem in International Trade*. Chicago: University of Chicago Press.

1924. *Canada's Balance of International Indebtedness, 1900–1913: An Inductive Study in the Theory of International Trade*. Cambridge, MA: Harvard University Press.

1930. *Lectures in Price and Distribution Theory*. Economics 301, University of Chicago, Summer Quarter, 1930; ed. M.D. Ketchum, mimeo, 1931.

1933. Inflation as a remedy for depression. In *Proceedings of the Institute of Public Affairs*, Seventh Annual Session, Athens: University of Georgia.

1937. *Studies in the Theory of International Trade*. New York: Harper.

1949. Bentham and J.S. Mill: the utilitarian background. *American Economic Review* 39, 360–82.

1950. *The Customs Union Issue*. New York: Carnegie Endowment for International Peace; London: Stevens.

1951. *International Economics: Studies*. Glencoe, IL: Free Press.

1952. *International Trade and Economic Development: Lectures Delivered at the National University of Brazil*. Glencoe, IL: Free Press; Oxford: Clarendon Press.

1958. *The Long View and the Short: Studies in Economic Theory and Policy*. Glencoe, IL: Free Press. (With bibliography.)

1960. The intellectual history of laissez faire. *Journal of Law and Economics* 3(October), 45–69.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1961. Hayek on freedom and coercion. *Southern Economic Journal* 2(3), 230–6.

1962. The necessary and the desirable range of discretion to be allowed to a monetary authority. In *In Search of a Monetary Constitution*, ed. L.B. Yeager. Cambridge, MA: Harvard University Press.

1963. Review article on C.B. Macpherson's *Political Theory of Possessive Individualism: Hobbes to Locke*. Canadian Journal of Economics and Political Science 29, 548–59.

1965. Guide to John Rae's ‘Life of Adam Smith’. Introduction to J. Rae, *Life of Adam Smith*. New York: Kelley.

1972. *The Role of Providence in the Social Order: An Essay in Intellectual History*. Philadelphia: American Philosophical Society.

1978. *Religious Thought and Economic Society: Four Chapters of an Unfinished Work*, ed. J. Melitz and D. Winch. Durham, NC: Duke University Press. Also published in *History of Political Economy* 10 (1978), 9–189.

Bibliography

Baumol, W.J. and Seiler, E.V. 1979. Jacob Viner. *International Encyclopedia of the Social Sciences*, vol. 18, Biographical Supplement. New York: Free Press.

Davis, J.R. 1971. *The New Economics and the Old Economists*. Ames: Iowa State University Press.

Machlup, F. 1972. What was left on Viner's desk. *Journal of Political Economy* 80, 353–64.

Machlup, F., Samuelson, P.A. and Baumol, W.J. 1972. In Memoriam, Jacob Viner (1892–1970). *Journal of Political Economy* 80, 1–15.

Patinkin, D. 1981. *Essays on and in the Chicago Tradition*. Durham, NC: Duke University Press.

Reder, M.W. 1982. Chicago economics: permanence and change. *Journal of Economic Literature* 20, 1–38.

Robbins, L. 1970. *Jacob Viner: A Tribute*. Princeton: Princeton University Press.

Samuels, W.J., ed. 1976. *The Chicago School of Political Economy*. Published jointly by the Association for Evolutionary Economics and Division of Research, Graduate School of Business Administration, Michigan State University, East Lansing.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Winch, D. 1981. Jacob Viner. *American Scholar* 50, 519–25.

How to cite this article

Spiegel, Henry W. "Viner, Jacob (1892–1970)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000031>
doi:10.1057/9780230226203.1798

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

vintage capital

Raouf Boucekkine, David de la Croix and Omar Licandro
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article reviews the early vintage capital literature of the 1960s, and identifies the factors behind the revival of topic from the 1990s. The fundamental properties of the seminal vintage capital growth models are disentangled, and the origins of the associated controversy on the importance of embodied technical progress are evoked. The recent revival of this literature is analysed with special emphasis on the rising support for the Solowian view of investment following Gordon's 1990 fundamental work on the price of durable goods, and the emergence of a new vintage human capital literature devoted to some fundamental economic growth issues.

Keywords

capital–labour substitution; Cobb–Douglas technology; cohort size; consumption smoothing; demographic transition; demography; Denison, E.; depreciation; diffusion of technology; embodiment hypothesis; fertility; human capital; inequality (explanations); information technology revolution; innovation; investment cycles; Leontief technology; life expectancy; neoclassical growth theory; neoclassical production functions; obsolescence; overlapping generations models; Phelps, E.; population growth; productivity slowdown; quasi-rent; replacement echoes; Solow, R.; technical progress; transition dynamics; vintage capital; vintage human capital

Article

In neoclassical growth theory capital is assumed to be homogeneous and technical progress disembodied, meaning that all capital units benefit equally from any technological improvement. The disembodied nature of technical progress looks unrealistic, as acknowledged by Solow (1960, p. 91):

> This conflicts with the casual observation that many if not most innovations need to be embodied in new kinds of durable equipment before they can be made effective.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> Improvements in technology affect output only to the extent that they are carried into practice either by net capital formation or by the replacement of old-fashioned equipment by the latest models…

Accounting for the age distribution of capital is a way to cope with this criticism. It actually inspired an important stream of the growth literature of the 1950s and 1960s, giving birth to vintage capital theory. An economy is said to have a *vintage capital* structure if machines and equipment belonging to separate generations have different productivity, or face different depreciation schedules as in Benhabib and Rustichini (1991). Let us denote by $I(v)$ the number of machines of vintage v. With zero physical depreciation, vintage technology v is

$$Y(v, t) = F(I(v), L(v, t), e^{\gamma v}),$$

where $Y(v, t)$ is the output of vintage v at time $t \geq v$ and $L(v, t)$ is the amount of labour assigned to this vintage. Parameter $\gamma > 0$ designates the rate of technical progress, which is said to be *embodied* since it benefits only vintage v. $F(.)$ has the properties of a neoclassical production function. Vintages produce the same final good

$$Y(t) = \int_{t-T(t)}^{t} Y(v, t)\ dv,$$

where $Y(t)$ is total production and $T(t)$ is the lifetime of the oldest operative vintage.

Besides realism, vintage capital models were initially thought to be able to generate quite different long-run properties and short-term dynamics from neoclassical growth models. Because the productivity gap between new and old vintages is increasing over time, the latter need not be operated for ever, and, contrary to the neoclassical growth theory, the lifetime of capital goods might well be finite (Johansen, 1959). Such a property was thought to involve non-monotonic transition dynamics governed by the replacement of scrapped goods, known as the replacement echoes principle, which again sharply departs from neoclassical growth models.

On more general ground, vintage capital models were at the heart of the embodiment controversy, which opposed Solow to some leading growth theorists and empiricists, among them Phelps (1962) and Denison (1964). While the former argued that accounting for the fraction of technological progress which is exclusively conveyed by capital accumulation (that is, embodied technical progress) is important to accounting for growth, Phelps argued that the decomposition of technical progress is irrelevant in the long run. Recent studies notably by Gordon (1990) have resuscitated this controversy, as we shall see. Before developing all these themes, it should be noted that, whereas early vintage capital

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

theory primarily focused on physical capital accumulation, recent contributions have taken the same view of human capital accumulation (see Chari and Hopenhayn, 1991). Vintage human capital is generated either by successive vintages of technologies which induce specific skills or by demographic conditions. Such contributions have brought out a new and quite appealing understanding of the mechanisms behind technology diffusion and demographic transitions, for example. We briefly review them also.

The lifetime of capital

In Johansen (1959), technology is 'putty-clay', meaning that capital–labour substitution is permitted *ex ante* but not once capital is installed. Technological progress is assumed to be labour-saving. Because factor proportions are fixed *ex post*,

$$Y(v, t) = F\left(I(v), e^{\gamma v} L(v, t)\right) = g(\lambda(v))\ I(v),$$

where the labour–capital ratio $\lambda(v)$ and the size of the capital stock $I(v)$ are both decided at the time of installation, and employment is $L(v, t) = \lambda(v) e^{\gamma v} I(v)$. In Johansen, *obsolescence* determines the range of active vintages. Quasi-rents of vintage v at date t are proportional to $g(\lambda(v)) - \lambda(v)\ e^{\gamma v} w(t)$, where $w(t)$ is the equilibrium wage. Since wages are permanently growing as a direct consequence of technical progress, quasi-rents are decreasing. Machines of vintage v are operated as long as their quasi-rents remain positive. Consequently, the scrapping age is defined by $T = t^* - v$ where $g(\lambda(v)) = \lambda(v)\ e^{\gamma v} w\left(t^*\right)$. Therefore, Johansen's framework leads to an endogenous, finite lifetime of capital.

Replacement echoes

If capital lifetime is finite, there might be a room for replacement echoes, as mentioned above. Solow et al. (1966) examine this question in the simpler case of a Leontief technology, when factor substitution is not allowed either *ex ante* or *ex post*. In such a case, $Y(v, t) = Y(v) = I(v) = e^{\gamma v}\ L(v)$, for all $t \geq v$. One unit of vintage capital v produces one unit of output once combined with $e^{-\gamma v}$ units of labour. Technical progress is embodied and takes the form of a decreasing labour requirement. For the same reasons as in Johansen, capital goods are scrapped in finite time. Using in addition a constant saving rate, and some technical assumptions, Solow et al. show convergence to a unique balanced growth path, delivering the same qualitative asymptotic behaviour as the neoclassical growth model. This was quite disappointing, since under finite lifetime one would have expected an investment burst from time to time, giving rise to replacement echoes.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Let us normalize the labour supply to unity. From labour market clearing, $\int_{t-T(t)}^{t} L(v)\, dv = 1$. Under constant lifetime, time differentiation of the equilibrium condition yields $L(t) = L(t-T)$, implying that investment is mainly driven by replacement activities. When obsolete capital is destroyed, new investments are needed to replace the scrapped machines, creating enough jobs to clear the labour market. As a direct consequence, job creation and investment have a periodic behaviour, implying that investment cycles are reproduced again and again in the future.

Solow et al. (1966) did not find echoes because of the constant saving rate assumption, which completely decouples investment from replacement. In an optimal growth model with linear utility and the same technological assumptions, Boucekkine, Germain and Licandro (1997) show (finite time) convergence to a constant lifetime, letting replacement echoes operate and generate everlasting fluctuations in investment, output and consumption. Under strictly concave preferences, fluctuations do arise in the short run but get dampened in the long run by consumption smoothing (see Boucekkine et al., 1998). Therefore, the short-run dynamics of vintage capital models differ strikingly from the neoclassical growth model, provided capital and labour are to some extent complementary, consistently with the observed dynamics of investment at both the plant level (Doms and Dunne, 1998) and the aggregate level (Cooper, Haltiwanger and Power, 1999). Non-monotonic behaviour has also been shown by Benhabib and Rustichini (1991) for vintage models with non-geometric depreciation.

The embodiment hypothesis

A crucial property of vintage capital models is the embodied nature of technological progress: the incorporation of innovations into the production process cannot be achieved without the acquisition of the new vintages which are their exclusive material support. According to Solow (1960), embodiment can have crucial implications for growth accounting. To make the point, he considers a Cobb–Douglas vintage technology

$$Y(v,t) = [e^{\gamma v} I(v)]^{1-\alpha} L(v,t)^{\alpha},$$

and the capital–labour ratio adjusts continuously. The embodiment hypothesis takes the form of quality adjustments, with capital's quality growing at rate γ. In sharp contrast to Johansen, capital lifetime needs not be finite, since under Cobb–Douglas technology any wage cost could be covered by assigning arbitrarily small amounts of labour.

A striking outcome of Solow's model is its aggregation properties. Denote by $L(t)$ the total labour supply, and define quality adjusted capital as

$$K(t) = \int_{-\infty}^{t} e^{\gamma v} I(v)\, dv.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(1)

Since marginal labour productivity equalizes across vintages, aggregate output becomes

$$Y(t) = K(t)^{1-\alpha} L(t)^{\alpha}.$$

Aggregate vintage technology in Solow (1960) degenerates into a neoclassical production function. However, by differentiating (1), the motion law for capital is slightly different

$$K'(t) = e^{\gamma t} I(t),$$

reflecting embodied technical change. Since $e^{-\gamma t}$ measures the relative price of investment goods at equilibrium, the value of capital is by definition $A(t) = e^{-\gamma t} K(t)$, and evolves following

$$A'(t) = I(t) - \gamma A(t).$$

Technological progress operates as a steady improvement in equipment quality, which in turn implies obsolescence of the previously installed capital. In Solow, obsolescence does not show up through finite time scrapping but through labour reallocation reflecting a declining value of capital.
This important point has been at the heart of recent literature on the *productivity slowdown* and the *information technology revolution* (see Whelan, 2002). Indeed, the potential implications for growth of embodied technical progress were tremendously controversial in the 1960s. In a famous statement, Denison (1964) claimed ‘the embodied question is unimportant’. His argument was merely quantitative and restricted embodiment to changes in the average age of capital in a one-sector growth accounting exercise. In particular, his reasoning omits *de facto* the relative price of capital channel. Greenwood, Hercowitz and Krusell (1997), by using Gordon's (1990) estimates of the relative price of equipment, quantitatively evaluate the two-sector Solow model, claiming that around 60 per cent of US per-capita growth is due to embodied technical change. As pointed out by Hercowitz (1998), Gordon's series have been good news for the Solowian view.

Vintage human capital

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The vintage capital growth literature typically considers labour as a homogenous good. However, just as physical capital is heterogenous, so too is the labour force. The concept of vintage human capital was explicitly used in the 1990s to treat some specific issues related to technology diffusion, inequality and economic demography.

In a world with a continuous pace of innovations, a representative individual faces the typical question of whether to stick to an established technology or to move to a new and better one. The trade-off is the following: switching to the new technique would allow him to employ a more advanced technology but he would lose the expertise, the *specific human capital*, accumulated on the old technique. In Chari and Hopenhayn (1991) and Parente (1994), individuals face exactly this dilemma. In such frameworks the generated vintage human capital distributions essentially mimic the vintage distribution of technologies, the time sequence of innovations being generally exogenously given. Chari and Hopenhayn (1991) consider a two-period overlapping generations model where different vintage technologies, operated by skilled and unskilled workers, coexist. Old workers are experts in the specific vintage technology they have run when young. The degree of complementary between skilled and unskilled labour affects negatively the velocity of technological diffusion, since young individuals have strong incentives to invest in old technologies when their unskilled labour endowment is highly complementary to the skilled labour of the old.

Jovanovic (1998) argues that vintage capital models are particularly well suited to explain income disparities across individuals and across countries. The main mechanism behind them is the following. Under the assumption that machines' quality and labour's skill are complementary, the best machines are assigned to the best-skilled individuals, exacerbating inequality. If reassignment is frictionless, then the best-skilled workers are immediately assigned to the frontier technology, the second-best go to the machines just below the frontier, and so on. Even though it is at odds with Chari and Hopenhayn, where adoption costs induce a much slower switching of technologies, frictionless reassignment has the virtue, consistent with cross-country evidence, of implying persistent inequality, in contrast to Parente (1994), which bears leapfrogging.

On the theoretical side, Jovanovic makes an important contribution to the vintage capital literature to the extent that he addresses the hard problem of combining vintage physical capital and vintage human capital in a framework where the vintage distributions of both assets are endogenous. Jovanovic uses an assignment model *à la* Sattinger (1975) to solve this difficult problem. Firms combine machines and workers in fixed proportions, say one machine for one worker, at every instant. Because labour resources are fixed, the latter fixed-proportions assumption implies that old machines become unprofitable at a finite time, as in Johansen. Vintage human capital comes from human capital accumulation *à la* Lucas (1988): the growth of the stock of human capital determines the maximal quality of human capital available: if the worker has human capital, h, and works a fraction of time u (in production), then her skill is given by $s=u\,h$. The typical assignment problem of a firm having acquired capital of a given vintage is to find the optimal vintage human capital or skill of the associated worker (via profit maximization), which makes it possible to achieve the pairing of skills and machines on the basis of the persistent inequality mechanism outlined above.

Demographics

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

One likely channel through which demographics affect growth is the size, quality and composition of the workforce. From this perspective, generations of workers can be understood as being vintages of human capital. In a continuous-time overlapping generations framework, Boucekkine, de la Croix and Licandro (2002) model the vintage specificity of human capital from schooling decisions. Individuals optimally decide how many years to spend at school as well as their retirement age; life expectancy has a positive effect on both because of its beneficial impact on the return to education. In such a framework, the vintage specificity of human capital depends, not on technological vintages as in Chari and Hopenhayn (1991), but on cohort-specific demographic characteristics, including education.

The observed relation between demographic variables, such as mortality, fertility and cohort sizes, and growth is anything but linear. Since a key element is between-generation differences in human capital, these nonlinearities may be modelled by the mean of a vintage structure of population. Boucekkine et al. (1998) generate nonlinear relationships between economic growth and both population growth and life expectancy. A longer life, for example, has several conflicting effects. On the one hand it increases the incentives to acquire education and reduces the depreciation rate of aggregate human capital. But on the other, an older population, which finished its schooling a long time ago, is harmful for economic growth.

Conclusion

After a relatively long stagnation, the vintage capital literature, which was a fundamental growth area in the 1960s, has been experiencing a revival since the early 1990s. This revival is due to several factors, among them the rising support for the Solowian view of investment following Gordon's fundamental work on the price of durable goods, the emergence of a new vintage capital growth theory led by Benhabib and Rustichini (1991) relying on a novel and appropriate mathematical set-up, and notably the increasingly common view that some fundamental economic growth issues (like technology diffusion, for example) do require the vintage structure to be better appraised. Of course, many tasks within this new literature remain to be addressed. In particular, much work is needed to bring the vintage models closer to the data. The work of Gilchrist and Williams (2000) is fundamental is this respect.

See Also

- diffusion of technology

Bibliography

Benhabib, J. and Rustichini, A. 1991. Vintage capital, investment, and growth. *Journal of Economic Theory* 55, 323–39.

Boucekkine, R., de la Croix, D. and Licandro, O. 2002. Vintage human capital, demographic trends and growth. *Journal of Economic Theory* 104, 340–75.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Boucekkine, R., Germain, M. and Licandro, O. 1997. Replacement echoes in the vintage capital growth model. *Journal of Economic Theory* 74, 333–48.

Boucekkine, R., Germain, M., Licandro, O. and Magnus, A. 1998. Creative destruction, investment volatility and the average age of capital. *Journal of Economic Growth* 3, 361–84.

Chari, V. and Hopenhayn, H. 1991. Vintage human capital, growth, and the diffusion of new technology. *Journal of Political Economy* 99, 1142–65.

Cooper, R., Haltiwanger, J. and Power, L. 1999. Machine replacement and the business cycle: lumps and bumps. *American Economic Review* 84, 921–46.

Denison, E. 1964. The unimportance of the embodied question. *American Economic Review – Papers and Proceedings* 54, 90–4.

Doms, M. and Dunne, T. 1998. Capital adjustment patterns in manufacturing plants. *Review of Economic Dynamics* 1, 409–29.

Gilchrist, S. and Williams, J. 2000. Putty-clay and investment: a business cycle analysis. *Journal of Political Economy* 108, 928–60.

Gordon, R. J. 1990. *The Measurement of Durable Goods Prices*. Chicago: University of Chicago Press.

Greenwood, J., Hercowitz, Z. and Krusell, P. 1997. Long-run implications of investment specific technological change. *American Economic Review* 87, 342–62.

Hercowitz, Z. 1998. The embodiment controversy: a review essay. *Journal of Monetary Economics* 41, 217–24.

Johansen, L. 1959. Substitution versus fixed production coefficients in the theory of economic growth. *Econometrica* 27, 157–76.

Jovanovic, B. 1998. Vintage capital and inequality. *Review of Economic Dynamics* 1, 497–530.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Parente, S. 1994. Technology adoption, learning-by-doing, and economic growth. *Journal of Economic Theory* 63, 346–69.

Phelps, E. 1962. The new view of investment: a neoclassical analysis. *Quarterly Journal of Economics* 76, 548–67.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Sattinger, M. 1975. Comparative advantage and the distribution of earnings and abilities. *Econometrica* 43, 455–68.

Solow, R. 1960. Investment and technological progress. In *Mathematical Methods in Social Sciences 1959*, ed. K. Arrow, S. Karlin and P. Suppes. Stanford: Stanford University Press.

Solow, R., Tobin, J., Von Weizsacker, C. and Yaari, M. 1966. Neoclassical growth with fixed factor proportions. *Review of Economic Studies* 33, 79–115.

Whelan, K. 2002. Computers, obsolescence and productivity. *Review of Economics and Statistics* 84, 445–61.

How to cite this article

Boucekkine, Raouf, David de la Croix and Omar Licandro. "vintage capital." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000072> doi:10.1057/9780230226203.1799

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

vintages

Dale W. Jorgenson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Vintages are durable goods acquired at different points of time. The acquisition prices for capital goods of each vintage at each point of time together with investments of all vintages at each point of time constitute the basic data on quantities and prices. These data can be employed in generating the complete vintage accounting system.

Keywords

capital measurement; capital services; depreciation; durable goods model of production; investment; Jorgenson, D. W.; price–quantity duality; production functions; rate of return; vintage accounting; vintages

Article

Investment represents the acquisition of capital goods at a given point of time. The quantity of investment is measured in the same way as the durable goods themselves. For example, investment in equipment is the number of machines of a given specification and investment in structures is the number of buildings of a particular description. The price of acquisition of a durable good is the unit cost of acquiring a piece of equipment or a structure.
By contrast with investment, capital services are measured in terms of the use of a durable good for a stipulated period of time. For example, a building can be leased for a period of years, an automobile can be rented for a number of days or weeks, and computer time can be purchased in seconds or minutes. The prices of the services of a durable good is the unit cost of using the good for a specified period.

Aggregation over vintages

We can refer to durable goods acquired at different points of time as different *vintages* of capital. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
flow of capital services is a quantity index of capital inputs from durable goods of different vintages. Under perfect substitutability among the services of durable goods of different vintages, the flow of capital services is a weighted sum of past investments. The weights correspond to the relative efficiencies of the different vintages of capital.

The durable goods model of production is characterized by price-quantity duality. The rental price of capital input is a price index corresponding to the quantity index given by the flow of capital services. The rental prices for all vintages of capital are proportional to the price index for capital input. The constants of proportionality are given by the relative efficiencies of the different vintages of capital.

We develop notation appropriate for the intertemporal theory of production by attaching time subscripts to the variables that occur in the theory. We can denote the quantity of output at time t by y_t and the quantities of J inputs at time t by $x_{jt} (j = 1, 2, \ldots, J)$. Similarly, we can denote the price of output at time t by q_t and the prices of the J inputs at time t by $p_{jt} (j = 1, 2, \ldots, J)$.

In order to characterize capital as a factor of production, we require the following additional notation:

- A_t – quantity of capital goods acquired at time t.
- $K_{t,\tau}$ – quantity of capital services from capital goods of age τ at time t.
- $p_{A,t}$ – price of acquisition of new capital goods at time t.
- $p_{Kt,\tau}$ – rental price of capital services from capital goods of age τ at time t.

To present the durable goods model of production we first assume that the production function, say F, is homothetically separable in the services of different vintages of capital:

$$y_t = F\left[G\left(K_{t,0}, K_{t,1} \ldots K_{t,\tau} \ldots\right), x_{2t} \ldots x_{Jt}\right]. \quad (1)$$

Where K_t is the flow of capital services, we can represent this quantity index of capital input as follows:

$$K_t = G,$$

where the function G is homogeneous of degree one in the services from capital goods of different ages. If we assume that the quantity index of capital input K_t is characterized by perfect substitutability among the services of different vintages of capital, we can write this index as the sum of these services:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$K_t = \sum_{\tau=0}^{\infty} K_{t,\tau}.$$

Under the additional assumption that the services provided by a durable good are proportional to initial investment in this good, we can express the quantity index of capital input in the form:

$$K_t = \sum_{\tau=0}^{\infty} d_\tau A_{t-\tau}. \quad (2)$$

The flow of capital services is a weighted sum of past investments with weights given by the relative efficiencies $\{d_\tau\}$ of capital goods at different ages.

Under constant returns to scale we can express the price of output as a function, say Q, of the prices of all inputs. The price function Q is homothetically separable in the rental prices of different vintages of capital:

$$q_t = Q\left[P\left(p_{K,t,0}, p_{K,t,1} \cdots p_{K,t,\tau} \cdots\right), p_{2t} \cdots p_{Jt}\right]. \quad (3)$$

Where $p_{K,t}$ is a price index of capital services, we can represent this index as follows:

$$p_{K,t} = P,$$

where the function P is homogeneous of degree one in the rental prices of capital goods of different ages. Under perfect substitutability among the services of different vintages of capital, we can write the price index of capital input P as the price of the services of a new capital good:

$$p_{K,t} = p_{K,t,0}.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Under the additional assumption that the services provided by a durable good are proportional to the initial investment, we can express the rental prices of capital goods of different ages in the form:

$$p_{K,t,\tau} = d_\tau p_{K,t}, \ (\tau = 0, 1, \ldots). \tag{4}$$

The rental prices are proportional to the rental price of capital input with constants of proportionality given by the relative efficiencies $\{d_\tau\}$ of capital goods of different ages.
Given the quantity of capital input K_t, representing the flow of capital services, and the price of capital inputs $p_{K,t}$, representing the rental price, capital input plays the same role in production as any other input. We next derive the prices and quantities of capital inputs from the prices and quantities for acquisition of durable goods $p_{A,t}$ and A_t.

Vintage accounting

We begin our description of the measurement of capital input with the quantities estimated by the perpetual inventory method. Taking the first difference of the expression for capital stock in terms of past investments (2), we obtain:

$$K_t - K_{t-1} = A_t + \sum_{\tau=1}^{\infty} (d_\tau - d_{\tau-1}) A_{t-\tau}, = A_t - R_t,$$

where R_t is the level of replacement requirements in period t. The change in capital stock from period to period is equal to the acquisition of investment goods less replacement requirements.
We turn next to a description of the price data required for the measurement of the price of capital input. There is a one-to-one correspondence between the vintage quantities that appear in the perpetual inventory method and the prices that appear in our vintage price accounts. To bring out this correspondence we use a system of present or discounted prices. Taking the present as time zero, the discounted price of a commodity, say q_t, multiplied by a discount factor:

$$q_t = \prod_{s=1}^{t} \frac{1}{1 + r_s} p_t.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The notational convenience of present or discounted prices results from dispensing with explicit discount factors in expressing prices for different time periods.

In the correspondence between the perpetual inventory method and its dual or price counterpart the price of acquisition of a capital good is analogous to capital stock. The price of acquisition, say $q_{A,t}$ is the sum of future rental prices of capital services, say $q_{K,t}$, weighted by the relative efficiencies of capital goods in all future periods:

$$q_{A,t} = \sum_{\tau=0}^{\infty} d_\tau q_{K,t+\tau+1} \tag{5}$$

This expression may be compared with the corresponding expression (2) giving capital stock as a weighted sum of past investments.

Taking the first difference of the expression for the acquisition price of capital goods in terms of future rentals (5), we obtain:

$$q_{A,t} - q_{A,t-1} = -q_{K,t} - \sum_{\tau=1}^{\infty} (d_\tau - d_{\tau-1}) q_{K,t+\tau} = -q_{K,t} + q_{D,t},$$

where $q_{D,t}$ is depreciation on a capital good in period t. The period-to-period change in the price of acquisition of a capital good is equal to depreciation less the rental price of capital. Postponing the purchase of a capital good makes it necessary to forgo one period's rental and makes it possible to avoid one period's depreciation. In the correspondence between the perpetual inventory method and its price counterpart, investment corresponds to the rental price of capital and replacement corresponds to depreciation.

We can rewrite the expression for the first difference of the acquisition price of capital goods in terms of undiscounted prices and the period-to-period discount rate:

$$p_{K,t} = p_{A,t-1} r_t + p_{D,t} - (p_{A,t} - p_{A,t-1}), \tag{6}$$

where $p_{A,t}$ is the undiscounted price of acquisition of capital goods, $p_{K,t}$ the price of capital services, $p_{D,t}$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
depreciation, and r_t the rate of return, all in period t. The price of capital services $p_{K,t}$ is the sum of return per unit of capital $p_{A,t-1}r_t$, depreciation $p_{D,t}$, and the negative of revaluation, $p_{A,t} - p_{A,t-1}$.
To apply this formula we require a series of undiscounted acquisition prices for capital goods $p_{A,t}$, rates of return r_t, depreciation on new capital goods, $p_{D,t}$, and revaluation of existing capital goods $p_{A,t} - p_{A,t-1}$.
To calculate the rate of return in each period we set the formula for the rental price $p_{K,t}$ times the quantity of capital K_{t-1} equal to property compensation. All of the variables entering this equation – current and past acquisition prices for capital goods, depreciation, revaluation, capital stock and property compensation – except for the rate of return, are directly observable. Replacing these variables by the corresponding data we solve this equation for the rate of return. To obtain the capital service price itself we substitute the rate of return into the original formula along with the other data. This completes the calculation of the service price.
In the perpetual inventory method data on the quantity of investment goods of every vintage are used to estimate capital formation, replacement requirements and capital stock. In the price counterpart of the perpetual inventory method data on the acquisition prices of investment goods of every vintage is required. In the full price–quantity duality that characterizes the vintage accounts, capital stock corresponds to the acquisition price of durable goods and investment corresponds to the rental price of capital services.

Conclusion

The distinguishing feature of capital as a factor of production is that durable goods contribute capital services to production at different points of time. The services provided by a given durable good are proportional to the initial investment. In addition, the services provided by different durable goods at the same point of time are perfect substitutes. The weights correspond to the relative efficiencies of the different vintages of capital. The durable goods model of production was originated by Walras (1954) and is discussed in greater detail by Jorgenson (1973) and Diewert (1980).
The durable goods model is characterized by price–quantity duality. The rental price of capital input is a price index corresponding to the quantity index given by the flow of capital services. The rental prices for all vintages of capital are proportional to the price index for capital input. The constants of proportionality are given by the relative efficiencies of the different vintages of capital. The dual to the durable good model of production was introduced by Hotelling (1925) and Haavelmo (1960). The dual to this model has been further developed by Arrow (1964) and Hall (1968).
The acquisition prices for capital goods of each vintage at each point of time together with investments of all vintages at each point of time constitute the basic data on quantities and prices. These data can be employed in generating the complete vintage accounting system originated by Christensen and Jorgenson (1973) and described by Jorgenson (1980). Price and quantity data that we have described for a single durable good are required for each durable good in the system. These data are used to derive price and quantity indexes for capital input in the theory of production presented in the entry on production functions.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- production functions
- technical change

Bibliography

Arrow, K.J. 1964. Optimal capital policy, the cost of capital, and myopic decision rules. *Annals of the Institute of Statistical Mathematics* 16, 16–30.

Christensen, L.R. and Jorgenson, D.W. 1973. Measuring economic performance in the private sector. In *The Measurement of Economic and Social Performance*, ed. M. Mess. New York: Columbia University Press for the National Bureau of Economic Research.

Diewert, W.E. 1980. Aggregation problems in the measurement of capital. In *The Measurement of Capital*, ed. D. Usher. Chicago: University of Chicago Press.

Haavelmo, T. 1960. *A Study in the Theory of Investment*. Chicago: University of Chicago Press.

Hall, R.E. 1968. Technical change and capital from the point of view of the dual. *Review of Economic Studies* 35(1), 35–46.

Hotelling, H.S. 1925. A general mathematical theory of depreciation. *Journal of the American Statistical Association* 20, 340–53.

Jorgenson, D.W. 1973. The economic theory of replacement and depreciation. In *Econometrics and Economic Theory*, ed. W. Sellekaerts. New York: Macmillan, 189–221.

Jorgenson, D.W. 1980. Accounting for capital. In *Capital, Efficiency, and Growth*, ed. G. von Furstenberg. Cambridge: Ballinger, 251–319.

Walras, L. 1874. *Elements of Pure Economics*. Trans. W. Jaffé, Homewood: Irwin, 1954.

How to cite this article

Jorgenson, Dale W. "vintages." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000032> doi:10.1057/9780230226203.1800

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

virtual economy

Clifford G. Gaddy
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The virtual economy was the system of informal rent distribution that arose in post-Soviet Russia in the 1990s as nonviable Soviet-era manufacturing industries sought to protect themselves from the discipline of the market. Enterprise directors and their allies throughout the economy (including government officials) colluded to use nonmarket prices and various forms of nonmonetary exchange such as barter to transfer value from resource sectors to manufacturing industry. The article discusses the system's historical roots, describes some of its characteristic phenomena, and outlines a model for behaviour of enterprises.

Keywords

arrears; barter; barter in transition; command economy; corruption; non-market prices; non-monetary exchange; relational capital; rent sharing; Soviet Union, economics in; tax offsets; virtual economy

Article

The virtual economy was the name given to the system of informal rent sharing or value distribution that prevailed in Russia in the 1990s. Featuring widespread use of nonmonetary exchange and nonmarket prices to conceal transfers of value, especially from resource sectors to manufacturing industry, the virtual economy reached a peak in the run-up to the country's financial crisis in August 1998.
The strategies used by enterprise directors to participate in this nonmonetary economy fundamentally changed the behaviour of hundreds and thousands of noncompetitive manufacturing enterprises in Russia during the transition process. The behavioural adaptation permitted enterprises to survive in the transition environment where they ought to have failed. The expectation had been that when the old Soviet industrial structure was shocked by the sudden collapse of central planning and the subsequent launching of radical market reforms – including mass privatization and elimination of overt subsidies – economic agents would be forced to change their behaviour to become competitive in a market economy. The transition was thus intended as a Darwinian process whereby only those enterprises that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
could transform themselves into competitive operations would survive. But in the case of Russia, the dinosaurs survived – without restructuring. They did change, but instead of adapting *to* the market, they changed to protect themselves *from* the market.

In essence the virtual economy was a peculiar system of rent distribution in which the primary vehicle through which agents laid claim to rents was *production*. The virtual economy was the set of informal institutions that facilitated the production of goods that were value-subtracting, that is, worth less than the value of the inputs used to produce them. Enterprises were able to engage in such production because they had recipients who were willing to accept fictitious (nonmarket) pricing of the goods at levels that masked their lack of profitability. Buyers and sellers colluded to hide the fictitious nature of the pricing. In the classic form of the virtual economy, they did so by avoiding money, instead using barter and other forms of nonmonetary exchange, as well as even more intricate subterfuges.

Since value was being destroyed as the system operated, there had to be a source of value. The ultimate ‘value pump’ in Russia was the fuel and energy sector, above all one single company, Gazprom – Russia's natural gas monopoly. In exchange for the rights to keep what it earned from exports, Gazprom pumped value into the system by supplying gas without being paid for it (or, more generally, at a cost that was low enough to keep enterprises operating). Gazprom subsidies, which then led to arrears to the government, were the primary way in which unprofitable activity was supported in Russia.

The virtual economy evolved and persisted because it met the needs of so many actors in the economy. Workers and managers at industrial dinosaurs benefited because the virtual economy postponed the ultimate reckoning for loss-making firms. Government, especially at the sub-national level, where much of the important action took place, benefited because the virtual economy system maintained employment and the provision of social services. Gazprom also benefited, since the value transfers it made to the virtual economy were the price it paid to appropriate the massive rents from exports.

The roots of the virtual economy mechanisms lay in the Soviet system, especially the production relationships that had developed under the Soviet command economy. These relationships represented a peculiar type of asset, ‘relational capital’, which supplemented the enterprise's conventional physical and human capital. Thanks to relational capital, market reform policies did not necessarily compel the enterprise to restructure in order to be able to compete in the market environment. Enterprises chose between whether to become more competitive in the market, by investing in physical and human capital, or to be better protected from the market, by investing in relational capital.

The term

The term ‘virtual economy’ was coined in 1998 by Gaddy and Ickes, building on terminology in a Russian government report from 1997. In early 1996, alarmed by the extent of tax delinquency in the country, President Boris Yeltsin appointed a special blue-ribbon panel to investigate the low rate of collection of taxes in Russia. Presenting its findings after an 18-month investigation, the panel reported that the country's largest companies conducted 73 per cent of all their business in the form of barter and other nonmonetary forms of settlement. Especially alarming was the extent of nonmonetary payments of taxes. During the period of review, these large enterprises paid less than eight per cent of their tax bills in actual cash. They simply failed to meet 29 per cent of their obligations at all, while ‘paying’ the remaining 63 per cent in the form of offsets and barter goods. The market value of the goods delivered was far below the nominal price used in the offsets, leaving the government with substantially less in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
real revenues than officially accounted for. In summing up their own conclusions about the contemporary Russian economy, the investigatory commission wrote:

> An economy is emerging where prices are charged which no one pays in cash; where no one pays anything on time; where huge mutual debts are created that also can't be paid off in reasonable periods of time; where wages are declared and not paid; and so on. … [This creates] illusory, or virtual earnings, which in turn lead to unpaid, or virtual fiscal obligations, [with business conducted at] nonmarket, or virtual prices. (Karpov, 1997)

Gaddy and Ickes (1998) suggested that the entire system be called a virtual economy ‘because it [was] based on illusion, or pretense, about almost every important parameter of the economy: prices, sales, wages, taxes, and budgets’. The pretence that had become the norm was as characteristic of the virtual economy as were the colourful forms of nonmonetary exchange.

The nonmonetary economy

The nonmonetary means of payment that characterized the virtual economy spanned a wide range. They included direct exchanges of goods (true barter), either bilaterally or through ‘chains’ with multiple participants, offsets (where debts accrued by one party were later paid off not in money but in goods), and promissory notes called *veksels*. *Veksels* – the name is derived from the German *Wechsel* (‘promissory note’) – were a widespread nonmonetary payment mechanism that ranged from being a substitute for money to essentially a form of barter.

There were several key nodes in the barter chains, above all the major natural monopolies known popularly as the ‘three fat boys’ (*tri tolstyaka*) – Gazprom (the natural gas monopoly), RAO UES (the electricity monopoly), and MPS (the state railways). All three frequently complained that they collected as little as ten per cent of their revenues in cash. Almost all enterprises in Russia were consumers of the output of these three companies, rail freight transport, gas and electricity. The three monopolies also accounted for about 25 per cent of taxes due to the federal budget. The fact that everyone needed to purchase services from the ‘fat boys’ meant that there was a ready demand for the *veksels* (IOUs) of these companies. It was this special position that put them at the core of the non-payments system in Russia.

The other key player in the barter economy was the government, or rather, governments at all levels. Here again was an agent to whom nearly everyone had an obligation. The volume of accrued unpaid taxes, plus the huge fines and penalties levied for nonpayment, presented governments with an almost inexhaustible supply of debts. And, in turn, governments themselves owed many others. They were, like the natural monopolies, a key node for barter.

One particularly important phenomenon was tax offsets. An enterprise owed taxes to the government, and concluded an agreement whereby those tax obligations were settled by delivering goods or performing services for the government. Of all the forms of nonmonetary transactions observed in Russia in the 1990s, the mechanism of tax offsets was the most characteristic of the virtual economy. Russian governments at all levels grew increasingly willing to offset enterprises’ tax obligations against goods or services delivered to the government. By the end of 1997, the accumulated tax debt was

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
enormous. Industrial enterprises were particularly egregious delinquents. The sum owed by the enterprises at the end of 1997 was equal to 46 per cent of the amount they actually remitted in taxes for the whole of 1997. These enormous debts gave impetus to the practice of tax offsets.

Consider, for example, an enterprise that was able to supply the local government with services in lieu of taxes. The enterprise could have paid its tax liability in money, but that would have required selling its output for cash. Alternatively, the enterprise could negotiate with the government to supply some service as an offset for taxes. If the enterprise had resources that were not fully utilized, the latter alternative was likely to reduce the effective tax burden on the enterprise. Moreover, once the government showed itself to be willing to engage in tax offsets, the options open to enterprises expanded. The enterprise could now potentially pay its taxes not only with its own products but also with products it received in barter deals from other enterprises. This greatly reduced the cost to the seller of accepting goods rather than cash.

The motivation for governments to join in the barter economy was simple. They reasoned that if they could not get cash, it was better to reach some sort of settlement than receive nothing at all. In some cases, especially at the local level, an enterprise could offer to deliver goods or services to the city or regional government in lieu of taxes. At the federal level, it was more common for the government to cancel tax arrears or taxes due by writing off the government's own debt to the enterprise in question for state orders. Once the practice was established with respect to past arrears, there was an anticipatory factor: enterprises began to feel confident that they could henceforth ship off products to the government, knowing that later they would be allowed to offset their taxes in an equivalent amount. Less than 60 per cent of all federal taxes collected in 1997 were paid in cash; the rest were in the form of offsets.

The federal government was particularly victimized by these schemes. Enterprises frequently colluded with regional and local officials to hide income and hence keep revenues away from the federal government for taxes whose revenues were split between local and national authorities. In other cases, local governments demanded that enterprises pay their taxes in the form of goods and services that could be used only locally and not be shared with the federal government (for instance, by providing road construction or repairs of buildings). Often, if the federal government received anything at all in these schemes, it was only what the regional governments did not want.

In one notorious case reported in the Russian press in the spring of 1998, the oblast (province) government of Samara had permitted enterprises to pay their regional taxes in the form of goods. One of the items offered turned out to be ten tons of toxic chemicals from a local chemical plant. Although the plant claimed (and was given) credit for 400 million rubles (80,000 dollars) in taxes, auditors later determined that the chemicals were worthless (and indeed dangerous). The Samara government never suffered from this curious deal, however, since it had previously sought and received permission from the federal ministry of labour to fulfil its obligations to the federal unemployment compensation fund by delivering goods instead of money. Among the goods it offered were the ten tons of toxic chemicals (Gaddy and Ickes, 2002, p. 176).

As a result of these practices, the Russian budget ran massive deficits. Even using the inflated prices used in the offset deals, federal revenues plummeted – from 16.2 per cent of GDP in 1995 to 12.4 per cent in 1998. To finance its deficits, the government had resorted to extensive borrowing outside and inside Russia at increasing and unsustainably high costs, thus digging itself even deeper in debt. Finally, on 17 August 1998, the government defaulted on about 40 billion dollars' worth of its own ruble-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
denominated debt instruments (so-called GKOs), some 17 billion dollars of which were held by foreigners.

The Soviet roots

The roots of the virtual economy lay in the structure and institutions bequeathed to Russia by its Soviet predecessor. Some parts of the economy, notably the resource industries, were value-adding. But most of the vast manufacturing sector that Russia had inherited could not compete in a market setting. In fact, by the final years of the Soviet era, the manufacturing sector was in poor condition even on the terms of the planned economy. By official Soviet standards, more than one-third of equipment in Russian industry was physically obsolete. Soviet planning practice, which emphasized output over costs, set physical, rather than economic, obsolescence as the criterion for removing a machine from the factory. As long as the machine could produce anything at all, it was kept in production. The result was very low replacement rates for capital equipment.

The location of industry in the Soviet economy was another problem. Not only did Soviet location policy ignore transport costs but it also failed to take into account the costs associated with the cold Russian climate – in terms of energy use, health maintenance and many other factors. By being placed in some of Russia's coldest and most remote regions, the manufacturing enterprises were rendered even less competitive and less attractive for foreign investment.

Equally important as the structure of the Soviet economy and its lack of competitiveness was the fact that this reality was hidden. As the market transition began, past history and performance gave no information about which sectors, or enterprises, were value-adding and which were value-destroying. The culprit was distorted Soviet pricing.

Soviet pricing and the ‘circus mirror’ effect

Soviet prices were not based on opportunity cost, or value; rather, they were simply an accounting instrument to measure plan fulfilment. Although Soviet prices were set arbitrarily, they were not set randomly. They were determined by specific rules of the system, which produced some systematic biases. First, the planners underpriced raw material inputs, especially energy. They based raw materials prices only on the operating costs of extraction, while ignoring rent. In so doing, they disregarded the opportunity cost of using the resources now rather than in the future. The planners’ overriding goal was to increase today's output. Scarcity pricing might have induced more conservation, but it would have militated against maximizing current production. This bias in raw materials prices fed into the system of industrial prices. Heavy consumers of energy were, in effect, subsidized. So, too, were heavy users of capital, thanks to the absence of interest charges. In short, costs of production were calculated on the basis of an incomplete enumeration of costs. This led to lower prices for inputs, especially resource inputs, than for final uses and thus an understatement of the share of gross output used in production and, hence, an overstatement of net output.

In addition to incomplete cost-based pricing, the Soviet system was explicitly biased towards certain users. The Soviet leadership assigned priority in the economy to heavy industry, especially the defence industry, and it was important that it appear that these sectors were producing value. This non-scarcity-based pricing was like a distorting mirror at the carnival. It created the illusion that many enterprises

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
were value-producing when in fact they were value-destroying.

The ‘loot chain’

A further factor contributing to the opaqueness of the Soviet economy and its post-Soviet successor was the way in which income from control of assets was passed down as payoffs through what Gregory Grossman (1998) referred to as the loot chain. In the USSR, wealth diverted from the official state economy into private hands was shared among networks of individuals in the form of payoffs, bribes, and other schemes. Over time an ever greater proportion of people's incomes depended on the chain of corruption and side payments.

The virtual economy perpetuated the loot chain in post-Soviet Russia. The living standards of a huge number of people depended on the chain of production and distribution of goods and services in the virtual economy system, where value redistribution, in contrast to looting pure and simple, occurred in a form that paralleled and was intertwined with actual productive economic activity. This made it especially difficult for agents to discern what their own value and the value of their assets would have been in a well-developed and transparent economy. Basic ideas of a market economy, such as the relationship between individual effort and reward, became almost impossibly obscure. One's static position in the production process – for instance, membership in the workforce of a particular enterprise – was more important for success than individual skills and abilities. The Soviet system separated ‘what you get’ from ‘what you do’. The reality was that the effort–reward nexus was random. Instead of ‘from each according to his ability, to each according to his needs (or ability)’, it was ‘to each according to some unknowable, random criterion’. The durability of the misperception depended on its opaqueness. There was no alternative, competing information about the real relationship. This meant that the loot chain was also a constraint on the future evolution of the economy. Individuals were dependent on the prevailing system and they could not know what an alternative system would offer. The uncertainty caused them to resist abandoning the prevailing system.

Impermissibility of true reform

While there was no accurate information about the economic importance of the large Soviet manufacturing sector, its social and political importance was unavoidable. Many of the least competitive enterprises – the so-called dinosaurs of Soviet industry – were socially the most important. They employed millions of workers and provided for tens of millions of their family members. Entire cities depended on them. The sheer size of this sector – as shown by employment – operated to maintain its social and political importance, and the illusion of its economic performance. In a sense, then, the importance of the manufacturing sector in Russia was an illusion economically but continued to be a political and social reality.

This latter reality constrained serious market reform policies. Russia did not formally reject the policies themselves; instead, it continued with a pretence of market reform. Policymakers launched one measure after another in their attempt to transform Russia into a market economy. But very few of those measures were allowed to play themselves out to their full extent. The consequences of complete and proper implementation would have been politically intolerable. Thus, while the nation's leadership

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
proclaimed reform policies, enterprises and other agents continued to behave in ways that rendered the policies ineffective.

The behavioural implications

The range of behavioural options in the virtual economy was broad. The ability to use nonmonetary mechanisms to pay taxes to governments and bills to the natural monopolies fundamentally changed the range of opportunities for action available to Russian enterprise directors. By allowing enterprises to settle their obligations by delivering goods for which there was no effective demand, the governments and the monopolies offered an incentive to avoid restructuring. For many enterprises it was easier to produce such goods than to restructure and earn additional monetary income to pay bills in cash. Producing those goods allowed for the use of idle capital and labour. In short, offsets and barter permitted some enterprises to survive without restructuring. To represent the full range of choice, not only market-oriented activity but also behaviour characteristic of the virtual economy, Gaddy and Ickes (2002) employed the notion of a two-dimensional space, called *r-d* space. The following sections outline their model.

Market distance, d

The impact of liberalization on the Soviet economy can be expressed with a spatial metaphor: liberalization revealed the distance that a Russian enterprise would have to travel to compete in the world economy. Let *d* designate the enterprise's ‘distance to the market’ at the start of transition. Clearly, *d* depends on the enterprise's initial endowments of the things that matter for market viability – physical and human capital, as well as the enterprise's marketing structure and organizational behaviour, but also the characteristics of the good that the enterprise produces (its quality and cost of production). Formally, define an enterprise's *d* as the amount of capital expenditure needed to enable the enterprise to produce a product that is competitive in the market. The fundamental reason for measuring *d* in terms of the investment cost is that transition causes a divergence between the value of existing (inherited) capital and that of newly installed capital.

One may begin to grasp this point by recalling what happened to traditional models of investment in market economies during the energy crisis of the 1970s. Those models predicted that investment would decline, given the tremendous increase in the price of energy. In fact, however, spending on new equipment and buildings soared. The reason for this discrepancy between model and reality was the divergence between the value of installed capital that was energy intensive and new capital that was energy saving. The conventional model ignored the sharp decline in the economic value of the existing capital stock as a result of the 1973 energy crisis. Installed capital had been the result of investment decisions based on low energy prices; hence, its value fell dramatically once energy prices quadrupled. This in turn only increased the demand for new investment in energy-saving equipment. The result was a divergence between the value of installed capital (which lost value) and that of new capital (which had full economic value). In the Russian context, measuring market distance *d* by the need for new capital investment is a way of capturing the cost of filling the gap between the value of inherited (Soviet) capital and new (market) capital.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Distribution of d

The level of *d* differs widely among enterprises in the economy. An enterprise that already produces a product it can sell in world markets at a price above cost will have a value of *d* equal to 0. A completely noncompetitive enterprise will have an enormously large *d*. Everyone else will be somewhere between. For example, an oil-producing enterprise will have a very low *d*. Its product is already right for the market. It may need only relatively small investments in marketing, and so on. A Soviet-style machine tool producer, in contrast, is likely to have a long distance to travel.

The distribution of *d*'s in transition economies differs in two respects from that in market economies. In transition economies the range of *d*'s is greater and the distribution is more skewed. Both differences stem from the dissimilarity in the process of entry and exit in market and planned economies. In a market economy, whether or not a new firm attempts to enter an industry depends on the founders' expectations about the new firm's competitiveness. They will enter if they expect the firm's potential costs to be lower (its productivity to be higher) than those of existing firms. No firm enters an industry in which it expects it will be noncompetitive. Over time the competitiveness of some firms declines, so *d* increases. But if a firm in a market economy has too high a level of *d*, it will be forced to close. Competition and hard budget constraints cause high-*d* enterprises to shut down.

In a transition economy, by contrast, some enterprises have very high *d*'s that would not be observed in a market economy. There are several reasons for these high-*d* enterprises. First, in socialist economies entry was not determined by expectations of profitability or competitiveness but rather by the need to fulfil plan targets. Second, insulation from the world economy meant that enterprises were created that produced goods for which the country might not have had a comparative advantage. Third, especially in the case of Russia, the priority given to defence production led to a proliferation of enterprises that produced goods whose market collapsed with the end of the Soviet Union. Fourth, since the geographic location of industry in the Soviet period was based on ignoring transport costs (as well as the costs associated with extraordinarily cold temperatures), the location of enterprises was also a factor in increasing the *d* in many cases. For all these reasons, the distribution of the *d*'s in Russia at the onset of the transition had a much higher mean and was more skewed to the right than in a mature market economy. This extra mass of high-*d* enterprises was the burden of the Soviet legacy. And it was this burden that was the essence of the restructuring problem: so many enterprises had to radically reduce their distance to the market at the same time.

One way to think of the purpose of economic reform is to reduce the average distance in the economy. This occurs through three means: (1) exit of high-*d* enterprises; (2) entry of new low-*d* enterprises; (3) and reduction of the *d* of surviving enterprises. In an ideal market world, market distance would be the only condition that characterized the state of an enterprise. If the only important difference in enterprises were their initial level of *d*, then policies that put pressure on existing high-*d* enterprises and encourage creation of new low-*d* enterprises would have the effect of pushing the distribution in the direction of the market.

Relational capital

The conventional view of restructuring, whereby reform means reducing *d*, assumes that each enterprise

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
has one set of resources – its physical and human capital – which it must use ever more efficiently in order to survive. The virtual economy view, by contrast, posits that some enterprises have another resource, relational capital, which they can draw on to enhance their chances for survival. Relational capital is the stock of goodwill that an enterprise can use to avoid the strictures of the budget constraint. An enterprise that has high relational capital can undertake transactions (bartering, using tax offsets, delaying payment) that other enterprises, with low amounts of relational capital, cannot get away with. To put it another way, relational capital is goodwill that can be translated into the ability to continue to engage in production and exchange without reducing the distance to the market. It is therefore the existence of this second dimension that can explain the persistent survival of high-*d* enterprises in the Russia of the 1990s.

At the onset of transition enterprises differed in their inherited relational capital – call it *r* – just as they differed in their *d*. Some enterprises (or their directors) had very good relations with local and/or federal officials. Relations with other enterprises also varied.

Origins of relational capital

The relational capital of Russian enterprises was initially accumulated in the Soviet system. Enterprise directors relied heavily on the accumulation and use of personal connections. Relational capital was passed forward to the post-Soviet system in a deceptively simple manner: it was spontaneously privatized. And here lies an important aspect of economic transition in Russia. As Hewett (1988) described, plan fulfilment in the Soviet economy required enterprise directors to use informal skills. Their ability to accomplish this, and their position in the economic hierarchy, was critical to their incomes. While directors earned income from these positions, they did not legally own the source of these incomes. The demise of the planning system, which had already begun with Mikhail Gorbachev's reforms in the late *perestroika* period, had the effect of increasing the autonomy of enterprise directors. With the start of economic reform and privatization, the role of the enterprise director increased; other mediating actors (planners, party officials) played less and less of a formal role in economic allocation. Directors used this opportunity to appropriate the returns to the relationships they had developed and cultivated under the previous system. However, in order for directors to appropriate these returns, the enterprises had to continue to operate. Much of the relational capital was both enterprise-specific and person-specific. To the extent that it was enterprise-specific, the director could not cash out the relational capital. The primary form of these connections was relationships with directors of other enterprises, often in related lines of activity, and with ministerial officials and local government officials. The relational capital was worthless to the incumbent director unless he remained in that particular enterprise. He could not leave the enterprise and take the relational capital with him. Furthermore, because it was person-specific, he could not sell it to someone else. Instead, in order to appropriate the rents accruing to his relational capital, he had to remain in the enterprise and keep it operating. The privatization of relational capital is thus an important part of the explanation of why directors fought to keep open enterprises that had few prospects in the market economy.

r – d space

The concept of relational capital can be used to revise the spatial representation of the Russian transition

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economy. There are now two state variables that describe the nature of an enterprise. In addition to the dimension of market distance, enterprises can be arrayed in terms of their level of relational capital. The initial conditions of an enterprise can thus be described by a two-dimensional space, *r–d* space, in which each enterprise has its own location.

Whether one views the enterprise sector in a single (*d*) dimension or in the two dimensions of *r–d* space is critical for how reform policy is understood. The conventional, one-dimensional view assumes that economic reform measures will have the greatest impact on those enterprises that have the highest level of *d*. According to this assumption, for example, if budget constraints are tightened, enterprises that are farthest from the market will be under greatest competitive pressure. Similarly, it is assumed that if the economy is opened to international competition, the greatest impact will be on those enterprises that are most in need of restructuring. In the two-dimensional *r–d* space environment, the effects of market-type reforms need not have this property at all. Tightening the budget constraint will not necessarily put the most pressure on the enterprise that is most inefficient (with the highest *d*). If the enterprise has been endowed with high *r*, it may be insulated against the impact of this policy; it can use relations to evade the budget constraint. And if tight budget constraints are enforced against enterprises that are lower in *r*, then the policy may, in fact, have greater impact on low-*d* enterprises than high-*d* enterprises.

It is not just the initial levels of either *r* or *d* that matter, of course. An enterprise's location in *r–d* space is not the immutable relic of its past; it depends on the path of enterprise investment decisions. If the enterprise has invested in *r*, it will improve its resistance to policies of tight budget constraints. The enterprise director's problem is to decide how much to invest in reducing distance and how much to invest in relational capital.

See Also

- barter
- barter in transition
- Soviet Union, economics in

Bibliography

Gaddy, C. and Ickes, B. 1998. Russia's virtual economy. *Foreign Affairs* 77, 53–67.

Gaddy, C. and Ickes, B. 2002. *Russia's Virtual Economy*. Washington, DC: Brookings Institution Press.

Grossman, G. 1998. Subverted sovereignty: historic role of the Soviet underground. In *The Tunnel at the End of the Light: Privatization, Business Networks, and Economic Transformation in Russia*, ed. S.S. Cohen, A. Schwartz and J. Zysman. Berkeley: University of California Press.

Hewett, E. 1988. *Reforming the Soviet Economy*. Washington, DC: Brookings Institution Press.

Karpov, P. 1997. On the causes of the low rate of tax collection (nonpayments in the fiscal system),

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

general causes of the ‘payments crisis’, and the possibility of restoring the solvency of Russian enterprises. Report of the Inter-Agency Balance-Sheet Commission, chaired by P.A. Karpov. Moscow (December). [In Russian]

How to cite this article

Gaddy, Clifford G. "virtual economy." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000067>
doi:10.1057/9780230226203.1801

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

voluntary contribution model of public goods

Richard Cornes
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article surveys the literature on the model of voluntary contributions to public goods that has developed since the early 1980s. This literature draws explicitly on noncooperative game theory. We present a recent novel statement of the problem, based on ‘replacement functions’, which is both more powerful and more transparent than the traditional formulation that uses players' best response functions. We survey existence, uniqueness and comparative static properties of the basic model, and also sketch some of the extensions of that model – impure public goods, weakest link and best shot – that have been proposed and applied to such problems as global public goods and the global commons. We also draw attention to recent attempts to dynamize the model.

Keywords

pure public goods; impure public goods; noncooperative games; neutrality; free riding; underprovision; weakest and weaker links; best and better shots; replacement functions; lotteries and public good provision

Article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Two classic papers by Samuelson (1954, 1955) played a major role in provoking interest in the problem of public good provision. However, they did not provide an explicit model of decentralized provision. His formal analysis focused on necessary conditions for their optimal provision. Elements of a positive model of decentralized provision – hereafter the standard model – were gradually developed during the following decades, and more complete formal analyses were provided by Cornes and Sandler (1985) and by Bergstrom, Blume and Varian (1986).

1 Introduction

Consider a community with an exogenous number, n, of members. They have preferences over a private good and a public good. Player i's consumption of the private good is y_i, and the total provision of the public good is G. Preferences, resource constraints, and the technology that converts individual contributions into the total available public good are summarized, respectively, by the following assumptions:

- **Preferences** Player i's preferences are represented by a utility function, $u_i(y_i, G)$, which is strictly increasing in both arguments and quasiconcave. Both goods are normal.
- **Resource constraint** $y_i+c_ig_i \leq m_i$, where player i's unit cost as a contributor c_i, and money income m_i, are exogenously given. g_i is player i's contribution to the public good.
- Technology of public good provision $G = \sum_{j=1}^{n} g_j$.

The model considers the Nash equilibrium of the static noncooperative game containing these elements when each player is choosing her best response, $\hat{g}_i$, to the choices made by all others, $G_{-i} = \sum_{j=1, j\neq i}^{n} g_j$.
This formulation slightly generalizes the standard model in that we allow unit costs to differ across contributors. This extension, initially explored by Ihori (1996), has interesting implications.

2 A graphical treatment

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Analyses typically derive a best response function for each player. This determines the player's most preferred choice of contribution as a function of the choices made by all other players: $\hat{g}_i = b_i(G_{-i})$, where $\hat{g}_i$ is player $\bar{i}$ '$-i$'s utility-maximizing response. A Nash noncooperative equilibrium is an allocation at which every player chooses her best response. Formally, it is a solution to the n equations provided by the individual best response functions in the n unknowns, $g_1, g_2, \ldots, g_n$. Questions about existence, uniqueness and other properties of equilibrium become questions about the existence, uniqueness and other properties of solutions to this set of equations. Such an approach, though naturally suggested by noncooperative game theory, is not the most helpful or transparent method of tackling these issues. We shall briefly sketch an alternative approach, suggested by Cornes and Hartley (2007a), which provides both a rigorous and powerful tool of analysis, and a simple and transparent geometric representation.

Individual behaviour

Figure 1(a) shows player i's preferences, constraints and choices. Suppose her income is m_i. If the sum of all other players' contributions is G'_{-i}, player i can devote all her money income to private goods consumption and enjoy the public good provided by others. This allocation is the point E'. Each unit of private good consumption given up by i augments total public good provision by the amount $1/c_i$. Thus her budget constraint is the line $E'F'$. Her most preferred choice is the point of tangency, P'. By varying G_{-i} parametrically, we can trace out the income expansion path II, which summarizes the player's behaviour. The locus II is everywhere continuous, and slopes upwards at all points at which player i is at an interior point, choosing strictly positive values of both y_i and g_i. If there is a finite value of G_{-i} at which $\hat{y}_i = m_i$, at that point the locus become horizontal.

Figure 1
A player's replacement function

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

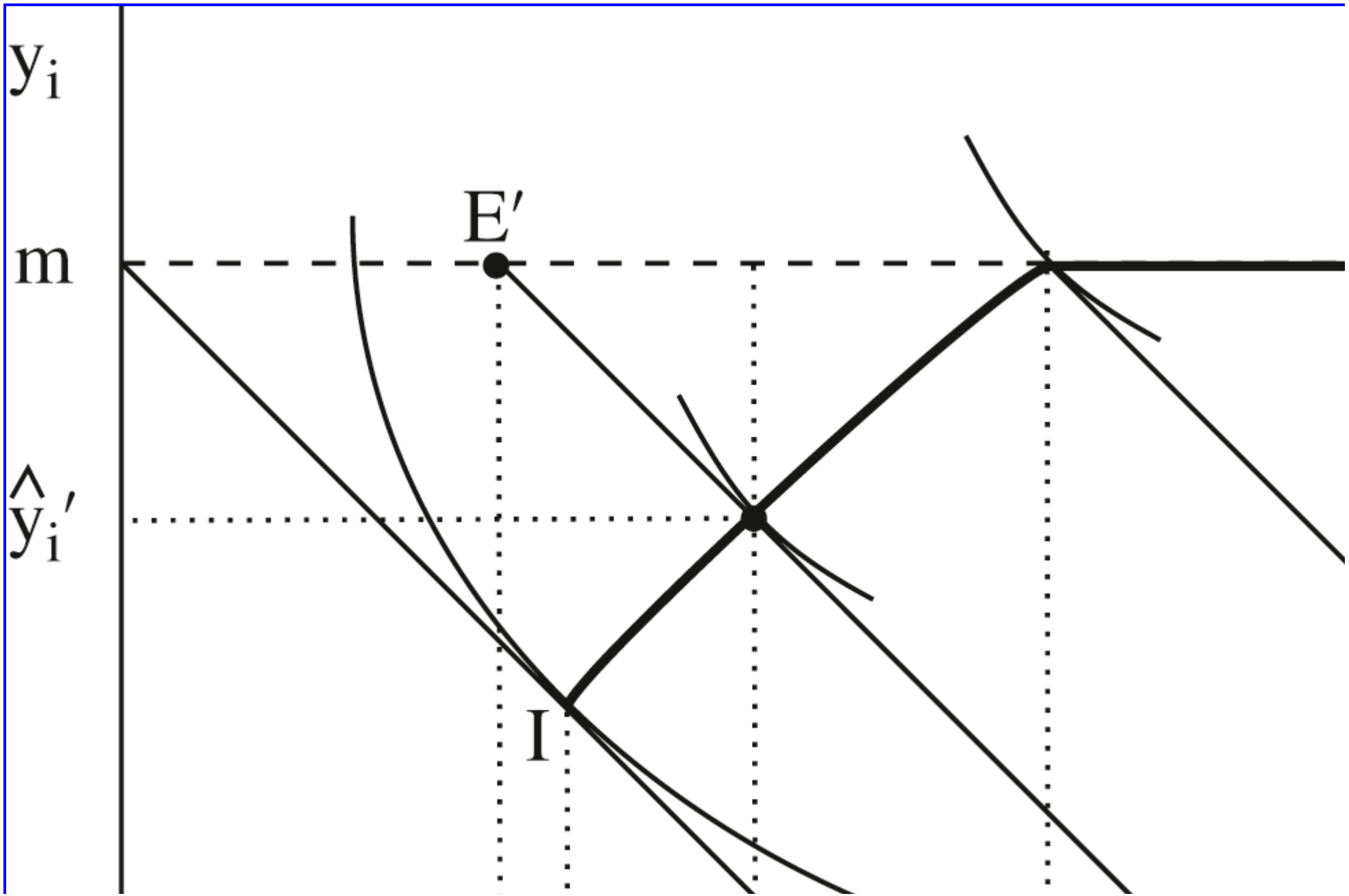

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Note that, for any given value of G, the vertical distance between the income expansion path and the locus of the income constraint mm measures the implied value of expenditure by i on the public good, $c_i\hat{g}_i = m_i - \hat{y}_i$. If c_i=1, this same distance measures the quantity of the public good. If $c_i \neq 1$, then a simple scaling up or down of the vertical axis in panel (b) allows us to depict the quantity g_i. In any event, under our assumptions, to any given level of total public good provision G above a certain value there corresponds a unique level of contribution by player i, $\hat{g}_i$, that is consistent with that observed level, in the sense that $\hat{g}_i$ is a best response to the quantity $G_{-i} = G - \hat{g}_i$. We write the implied functional relationship as $\hat{g}_i = r_i(G)$ and call this player i's replacement function. The figure suggests that every player has a replacement function that is continuous, everywhere non-increasing, and strictly decreasing in G wherever the replacement value itself is positive.

One further property of an individual's replacement function is significant. Suppose that, at a given level of G, player i is a strictly positive contributor. Consider the consequence of an increase of Δm_i in i's money income. At that given level of G, Figure 1 panel (a) shows that her chosen allocation is unchanged. She consumes an unchanged quantity of the private good. Thus, her contribution to the public good changes by the amount

$$\Delta\hat{g}_i = \frac{\Delta m_i}{c_i}.$$

Geometrically, the graph of (the positive section of) her replacement function rises vertically by an amount that, appropriately deflated by the cost parameter, equals the income change. This property plays a crucial role in comparative static analysis.

Nash equilibrium

Figure 2 shows the graphs of players' replacement functions in a three-player game. Equilibrium is an allocation at

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which the aggregate quantity of the public good is consistent with the replacement values to which it gives rise. In an *n*-player voluntary contribution game, it is an allocation at which

$$R(G) \equiv \sum_{j=1}^{n} r_j(G) = G.$$

The 'aggregate replacement function' $R(G)$ is shown as the thick line in Figure 2. It is simply the vertical sum of the individual graphs. A Nash equilibrium may be depicted graphically as a point where the graph of $R(G)$ intersects the 45° ray through the origin in Figure 2. This relationship describes a Nash equilibrium in the form of a single equation in a single unknown, G, regardless of how many players there are, and how they differ with respect to preferences, unit costs and money incomes. Armed with the properties already sketched above of the individual replacement functions, scrutiny of this equation is sufficient to provide a complete positive analysis of the model. First, however, note the following simple points. First, the sum of two continuous functions is continuous. Second, the sum of two monotonic functions is itself monotonic.

Figure 2

Nash equilibrium

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

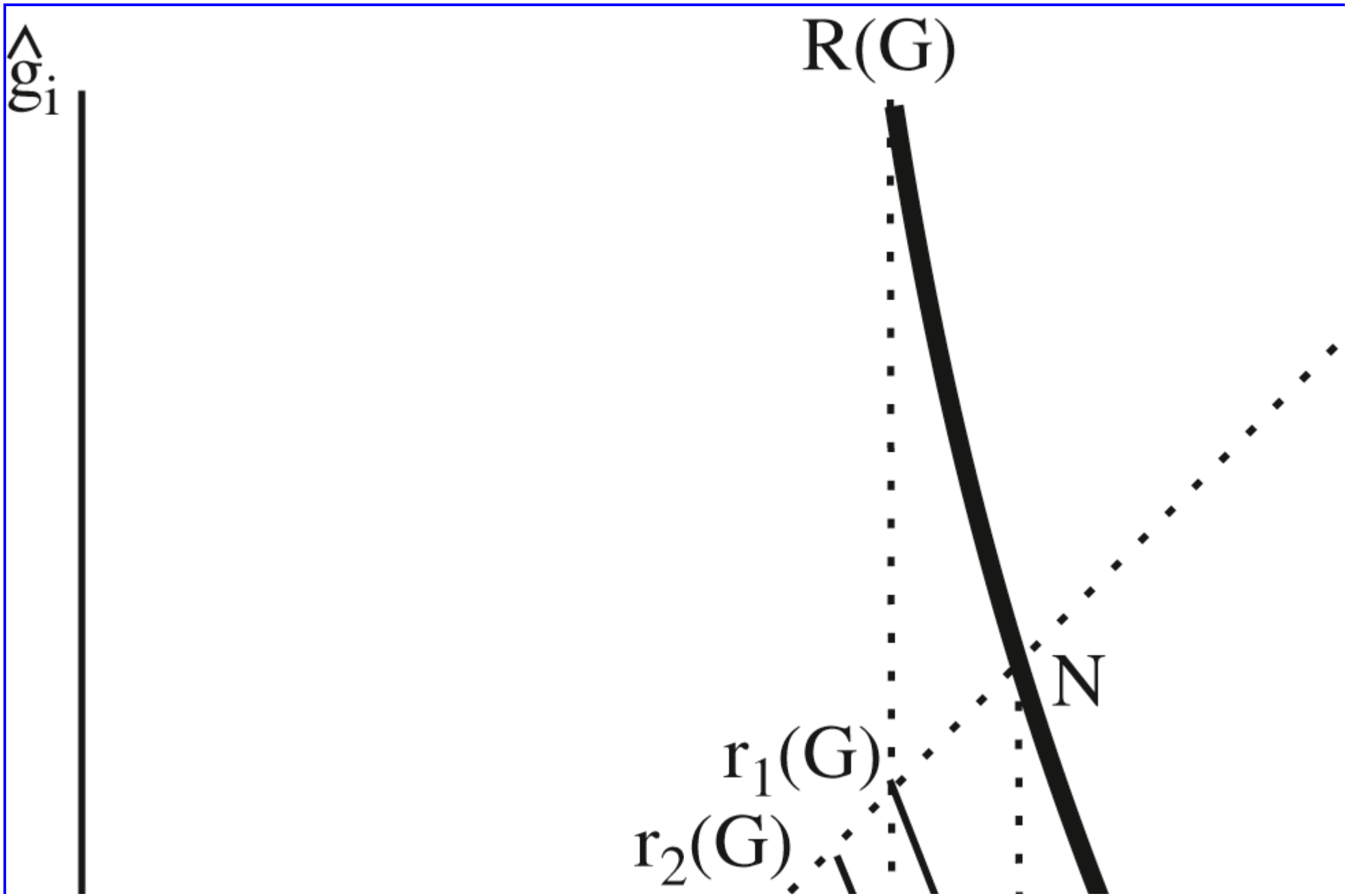

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

3 Properties of the equilibrium

We now have all the ingredients for a rigorous analysis of the equilibrium properties of the model, which we now investigate.

Existence of equilibrium

Consider the player whose replacement graph reaches the 45° ray furthest from the origin in (G, $R(G)$) space. It is possible that, at that level of G, all other players are choosing to contribute zero. In this case, we have found an equilibrium, at which the chosen player is the sole contributor. Alternatively, there may be other players whose replacement values are positive. In this case, we have found a value of G at which $R(G) \equiv \sum_{j=1}^{n} r_j(G) > G$. Then monotonicity implies that, as G rises, the left-hand side of this inequality falls, while the right-hand side rises. Continuity implies that there must be a finite value of G at which the equilibrium condition holds. Either way, an equilibrium certainly exists. In Figure 2, this is the point G^N, at which the sum of all players' contribution levels that are individually consistent with G^N is also collectively consistent.

Uniqueness of equilibrium

Monotonicity implies that $R(G)$ is everywhere nonincreasing. Clearly, G is a strictly increasing function of itself. Thus, there can only be one value of G at which $R(G)=G$.

Presumptive inefficiency of equilibrium

In the basic model, in which a common unit cost is assumed across players, there is a general presumption that too little of the public good is provided at equilibrium, in the sense that Pareto-superior allocations can be obtained by increasing the level of public good provision. This may be confirmed by a simple envelope argument. Suppose that, at an equilibrium, players j and k are both positive contributors. Starting from the equilibrium, a small increase in player j's contribution imposes a second-order cost on player j, but generates a first-order benefit for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

player k. Similarly, a small increase in player k's contribution imposes a second-order cost for player k and a first-order utility gain for player j. Thus it is possible for both to be made better off if both raise their contributions slightly above their equilibrium levels. Furthermore, such a move will not hurt other players, and will generally benefit them. Thus it is Pareto-improving.

In the current model, in which unit costs are allowed to differ across players, this remains true. There is also, however, a second source of inefficiency. This arises from the fact that the 'wrong' people contribute at equilibrium. Consider an equilibrium at which both a high-cost and a low-cost contributor are making positive contributions. An initial transfer of income from the high to the low-cost player shifts the replacement function of the high-cost player down, and that of the low-cost player up, in the neighbourhood of the equilibrium value of G. But the latter shift is quantitatively greater, so that the aggregate replacement function shifts upwards. The equilibrium provision therefore must rise, and contemplation of Figure 1(a) makes it clear that all players are better off in the new equilibrium.

Note that we talk of presumptive, not necessary, underprovision. This is for two reasons. First, as Cornes and Sandler (1996, p. 160) point out, if every player prefers to consume the private and public good in fixed proportions, so that their indifference curves are L-shaped, then the equilibrium is Pareto efficient. This possibility disappears if we allow some substitutability between the private and public goods. A second possibility, which certainly needs to be taken more seriously in policy discussions of public good provision than is sometimes done, is that the equilibrium involves zero total provision and that, even when provision is zero, the sum of all player's marginal valuations is less than the minimum cost of producing an increment of the public good. In this case, the public good neither is, nor should be, provided.

Neutrality

Suppose that two players – say i and j – have the same value for the cost parameter. Consider an equilibrium, G^*, at which both are strictly positive contributors. Now transfer an amount of income, Δm, from one to the other. In the neighborhood of G^*, the recipient's replacement graph shifts upwards by the amount Δm. The donor's graph shifts downwards by the amount Δm. Thus, if both remain positive contributors at G^*, that value remains the sole

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

equilibrium public good provision level. Nothing real has changed – equilibrium levels of private good consumption and of total public good provision, and therefore equilibrium utility levels, are unaffected by the income transfer. This is the famous neutrality property of the standard model, which assumes a common value of the cost parameter for all players. Often attributed to Warr (1983), it was foreshadowed in earlier work by Shibata (1971).

Non-neutrality

The reasoning that led to the neutrality result allows us to understand easily the circumstances under which neutrality fails to hold. First, suppose that the source of the income transfer is initially choosing to contribute zero. Then, at the initial level of G, the reduction in her income cannot shift the relevant portion of her replacement function downwards – she is already contributing zero. The recipient's replacement graph shifts upwards. Therefore the aggregate replacement graph shifts upwards, and the equilibrium provision of the public good must now be higher. Transfers between existing contributors and noncontributors will have real consequences, leading to changes in both the equilibrium total public good provision and also in individual equilibrium utility levels. It is even possible, as Cornes and Sandler (2000) point out, that transfers from each of several noncontributors to contributors leads to a Pareto-superior allocation.
Second, our discussion of the presumptive inefficiency of equilibrium has already shown that an income transfer from a high-unit-cost contributor to a low-unit-cost contributor will lead to a higher level of equilibrium provision and to a Pareto improvement.

Implications of a cost change

Suppose that player i is initially a positive contributor, and that she enjoys an exogenous reduction in her unit cost. Consideration of Figure 1 shows that the level of her preferred contribution that is associated with the initial equilibrium value of G must now be higher. In the absence of any other shocks, the equilibrium level of total provision must rise. Thus, every player except for i enjoys a higher equilibrium utility. However, player i herself may be either better or worse off – on the one hand, total provision is higher, but on the other hand she is now

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

contributing a higher share of that total, since her fellow contributors have reduced their contributions.

Limiting behaviour as *n* gets large

The implications of adding players to the community are very easy to trace using our suggested approach. Suppose a fourth player joins the group of three depicted in Figure 2. To identify the new equilibrium, we merely add the new player's replacement graph to the existing ones. There are two possibilities. It is possible that, at the equilibrium of the three-player community the fourth player would choose to contribute zero. This will be the case if the extra player's replacement graph hits the horizontal axis in Figure 2 at a point to the left of G^N. The equilibrium level of total provision, and the choices and utilities of the three initial players, are unchanged. The fourth player chooses to contribute nothing, and enjoys the existing level of public good, while allocating all of his money income to private good consumption. Alternatively, the replacement value of the new player is positive at the existing equilibrium. In this case, the graph of the aggregate replacement function shifts upwards in the neighborhood of the initial equilibrium. The new equilibrium involves a higher total provision level. Existing contributors will reduce their individual contributions, and all are advantaged by the addition of the extra player. In the presence of a large number of potential contributors who may differ in terms of incomes, preferences or unit costs, the diagram strongly suggests the conclusion reached by Andreoni (1988) – namely, that when n is large, the proportion of players who make strictly positive equilibrium contributions may be vanishingly small. Almost all players choose zero contributions.

4 Extensions

Early attempts to apply the voluntary contributions model – for example, to charitable giving, in which the aggregate G is the total quantity subscribed to some good cause – suggest that the very strong implications of the simple model – neutrality when unit costs are the same across contributors, and its implication that, when n is large, the number of strictly positive contributors will be a very small fraction of n – are difficult to square with empirical evidence. In addition, recent concerns with global and regional public goods have led to an interest in situations that naturally seem to involve public good technologies other than the summation one described above.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

We now briefly review some of the recent extensions and modifications of the model.

Technology of public good provision

Hirshleifer (1983) suggested two types of public good which are not captured by the summation technology and which, he argued, may be of empirical significance. They are characterized by different public good provision technologies. Best-shot and weakest-link public goods are captured, respectively by the following technologies:

$$\text{Best-shot}: G = Max\{g_1, g_2, \ldots, g_n\}$$

$$\text{Weakest-link}: G = Min\{g_1, g_2, \ldots, g_n\}.$$

Hirshleifer's example of a best-shot public good involves defensive guns ringing a city, each trying to shoot down an approaching missile. What matters to the city's inhabitants is the accuracy of the single most accurate shot. His example of a weakest-link involves a group of farmers, each owning a pie-shaped slice of land within a circular area surrounded by sea. Each is responsible for the maintenance of his part of the perimeter dyke. In the event of a storm that threatens to breach the dyke, it is the level of maintenance of the least well-maintained stretch of wall that determines the level of security enjoyed by all. Sandler (2004) suggests a wide range of situations involving regional or global public goods that are better captured by one or other of these formulations than by the standard summation formulation.

These formulations have very distinctive equilibrium properties. For example, consider a two-player model with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

the weakest link technology in which there is an equilibrium at which both contribute, say, ten units to the public good. Then any allocation at which each is contributing x units, where x lies between zero and ten, is also an equilibrium. After all, if the other player is contributing x units, it does not pay you to contribute any more than x, since the total provision is defined by the smaller individual contribution. This game there can have a continuum of equilibria. Hirshleifer himself suggested that the players may be expected to choose the Pareto-dominant equilibrium. However, experimental evidence suggests that players find it surprisingly hard to coordinate on the Pareto dominant equilibrium.

Cornes (1993) and Cornes and Hartley (2007b) consider the class of games in which the total level of a public good is generated by individual contributions according to a constant returns to scale CES production process: $G = \left[\sum_{i=1}^{n} g_i^{v}\right]^{\frac{1}{v}}$. The summation model is obtained by putting $v=1$: $v \to +\infty$ generates the best shot, and $v \to -\infty$ generates the weakest link. They show that, if $-\infty < v < 1$, the resulting weaker link model has a unique equilibrium. It is only at the limit, when the isoquants associated with the production technology are L-shaped, that Hirshleifer's problem of multiple equilibria arises. Moreover, if player i is contributing less than player j at an equilibrium – perhaps because her income is lower, or because she has less interest in the public good – then player i has a higher marginal product as a contributor. Hence, neutrality with respect to income transfers breaks down, and a transfer from player j to player i may lead to a higher equilibrium level of public good provision and may be Pareto improving.

Situations involving $v>1$ are better-shot games. Here, the production technology is inherently nonconvex, and again multiple equilibria may arise. For finite values of v, an equilibrium may involve positive contributions by each of a team of positive contributors, while the rest make zero contributions. However, there may be many such equilibria, each involving a different team of contributors. In Hirshleifer's best-shot case, if there are n players, there may be n equilibria, each of which involves a single ‘champion’, or ‘dragon-slayer’, who is the sole positive contributor, while all others make zero contributions. Again, achieving an equilibrium requires the players to resolve a tricky coordination problem.

Preferences

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Cornes and Sandler (1984, 1994) extend the basic model by modifying the individual preferences. They included player *i*'s own contribution as an argument of her own utility function, in addition to the aggregate quantity *G*:

$$u_i(\,.\,) = u_i(\gamma_i, g_i, G).$$

They suggest this formulation as a model of charitable giving, a suggestion explored by Andreoni (1988). Donor *i* not only cares about the total amount given to the charitable giving, *G*, but also experiences a 'warm glow' of satisfaction from her own contribution, g_i. If the standard resource constraint and public good technology are retained, this modification is sufficient to produce very rich comparative static possibilities: neutrality does not generally hold, and an increase in player *i*'s money income alone may either increase or reduce the equilibrium level of *G*. Finally, note that if the utility function is assumed to take the Cobb-Douglas form – $u_i(\,.\,) = \gamma_i^{\alpha} g_i^{\beta} G^{\gamma}$ – then at any equilibrium every player will make a positive contribution to the public good. A proliferation of noncontributors as the number of players increases is no longer implied.

This extension significantly broadens the range of potential applications of the model. First, there is nothing to stop us from considering situations in which player *i* regards *G* as a public bad – $\frac{\partial u_i(.)}{\partial G} < 0$. Thus the model may be interpreted as one involving congestion or pollution. Each may still be a positive contributor at equilibrium, the pollution or congestion being an incidental by-product that is jointly generated alongside the private good g_i. Kotchen (2006) and Ruebbelke (2002) have explored such models.

Morgan (2000) and Duncan (2002) have used a slight modification of this model to investigate the potential role that lotteries, or raffles, may play in raising the public good level above that implied by the voluntary contribution model. The basic idea is simple. The presence of the public good by itself involves a positive externality, and will tend to be underprovided. If individuals buy lottery tickets, each of which partially contributes to the public good and also gives its purchaser a probability of winning a money prize, a negative externality is thereby added – by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

buying a ticket, and increasing my chance of winning the prize, I inflict a negative externality on other ticket holders. There are two externalities, one beneficial and one harmful. The resulting equilibrium, at which these externalities tend to counteract each other, may involve a higher level of public good provision than if it were provided simply by individual contributions in the absence of the lottery.

Dynamic models

Up to this point, our discussion has remained firmly within the context of a one-shot static game. It is natural to wonder how the properties of equilibrium – in particular its presumptive inefficiency – are affected if we allow the contribution game to be played over many time periods. Schelling (1960, p. 45) suggested that such a setting may allow each player to make a small contribution, then wait to see whether others follow suit, before deciding whether to make a further small contribution. His suggestion has been analysed more formally by others, notably by Admati and Perry (1991) and Marx and Matthews (2000).

The last-named authors, whose analysis includes a useful discussion of the difference between their model and that of Admati and Perry, allow every player to choose a contribution level in each time period – any non-negative contribution, however large or small, is admissible. The properties of equilibria depend on (i) the degree of heterogeneity of players' valuations of the public good, (ii) the rate at which future costs and benefits are discounted, and (iii) whether or not there is a significant step in the benefit function – for example, a bridge generates no benefits until it is completed, thus representing an extreme example of a benefit function with a discrete step. They provide good news and bad news. The good news is that, if contributions can be made in small increments over time, equilibria can be attained that are more efficient than the equilibrium associated with the one-shot game. They argue that, if players' valuations are similar, and the rate of discount low, then nearly efficient equilibria exist. Furthermore, the presence of a significant benefit jump helps the prospects of successful completion of a project. An efficient equilibrium of the dynamic game may exist even in situations in which the only equilibrium of the static game involves zero contributions. The bad news is that, in common with many other dynamic games, there also exist other equilibria involving zero contributions.

Duffy, Ochs and Vesterlund (2007) have investigated the properties of such dynamic models experimentally.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

They confirm that contributions do indeed tend to be higher in dynamic games of the kind proposed by Marx and Matthews, but their results cast doubt on the claimed importance of jumps in the benefit function.

See Also

- non-cooperative games (equilibrium existence)
- public goods

Bibliography

Admati, A. and Perry, M. 1991. Joint projects without commitment. *Review of Economic Studies* 58, 259–76.

Andreoni, J. 1988. Privately provided public goods in a large economy: the limits of altruism. *Journal of Public Economics* 35, 57–73.

Bergstrom, T.C., Blume, L. and Varian, H. 1986. On the private provision of public goods. *Journal of Public Economics* 29, 25–49.

Cornes, R.C. 1993. Dyke maintenance and other stories: some neglected types of public good. *Quarterly Journal of Economics* 107, 259–71.

Cornes, R.C. and Hartley, R. 2007a. Aggregative public good games. *Journal of Public Economic Theory* 9, 201–19.

Cornes, R.C. and Hartley, R. 2007b. Weak links, good shots and other public good games: building on BBV. *Journal of Public Economics* 91, 1684–707.

Cornes, R.C. and Sandler, T. 1984. Easy riders, joint production and collective action. *Economic Journal* 94, 580–98.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Cornes, R.C. and Sandler, T. 1985. The simple analytics of pure public good provision. *Economica* 52, 103–16.

Cornes, R.C. and Sandler, T. 1994. Comparative static properties of the impure public good model. *Journal of Public Economics* 54, 403–21.

Cornes, R.C. and Sandler, T. 1996. *The Theory of Externalities, Public Goods and Club Goods*, 2nd edn. New York: Cambridge University Press.

Cornes, R.C. and Sandler, T. 2000. Pareto-improving redistribution in the pure public good model. *German Economic Review* 1, 169–86.

Duffy, J., Ochs, J. and Vesterlund, L. 2007. Giving little by little: dynamic voluntary contribution games. *Journal of Public Economics* 91, 1708–30.

Duncan, B. 2002. Pumpkin pies and public goods: the raffle fundraising strategy. *Public Choice* 111, 49–71.

Hirshleifer, J. 1983. From weakest-link to best-shot: the voluntary provision of public goods. *Public Choice* 41, 371–86.

Ihori, T. 1996. International public goods and contribution productivity differentials. *Journal of Public Economics* 61, 139–54.

Kotchen, M. 2006. Green markets and private provision of public goods. *Journal of Political Economy* 114, 816–34.

Marx, L. and Matthews, S. 2000. Dynamic voluntary contribution to a public project. *Review of Economic Studies* 67, 327–58.

Morgan, J. 2000. Financing public goods by means of lotteries. *Review of Economic Studies* 67, 761–84.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Ruebbelke, D. 2002. *International Climate Policy to Combat Global Warming: An Analysis of the Ancillary Benefits of Reducing Carbon Emissions*. Cheltenham: Edward Elgar.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36, 387–9.

Samuelson, P.A. 1955. A diagrammatic exposition of a theory of public expenditure. *Review of Economics and Statistics* 37, 350–6.

Sandler, T. 2004. *Global Collective Action*. New York: Cambridge University Press.

Schelling, T. 1960. *The Strategy of Conflict*. Oxford: Oxford University Press.

Shibata, H. 1971. A bargaining model of the pure theory of public expenditure. *Journal of Political Economy* 79, 1–29.

Warr, P.G. 1983. The private provision of a public good is independent of the distribution of income. *Economics Letters* 13, 207–11.

How to cite this article

Cornes, Richard. "voluntary contribution model of public goods." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 24 June 2009 <http://www.dictionaryofeconomics.com.ezproxy.bu.edu/article?id=pde2009_P000373>
doi:10.1057/9780230226203.1904

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

von Neumann, John (1903 – 1957)

Gerald L. Thompson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

His life

Jansci (John) von Neumann was born to Max and Margaret Neumann on 28 December 1903 in Budapest, Hungary. He showed an early talent for mental calculation, reading and languages. In 1914, at the age of ten, he entered the Lutheran Gymnasium for boys. Although his great intellectual (especially mathematical) abilities were recognized early, he never skipped a grade and instead stayed with his peers. An early teacher, Laslo Ratz, recommended that he be given advanced mathematics tutoring, and a young mathematician Michael Fekete was employed for this purpose. One of the results of these lessons was von Neumann's first mathematical publication (joint with Fekete) when he was 18. Besides his native Hungarian, Jansci (or Johnny, as he was universally known in his later life) spoke German with his parents and a nurse and learned Latin and Greek as well as French and English in school. In 1921 he enrolled in mathematics at the University of Budapest but promptly left for Berlin, where he studied with Erhard Schmidt. Each semester he returned to Budapest to take examinations without ever having attended classes. While in Berlin he frequently took a three-hour train trip to Göttingen, where he spent considerable time talking to David Hilbert, who was then the most outstanding mathematician of Germany. One of Hilbert's main goals at that time was the axiomatization of all of mathematics so that it could be mechanized and solved in a routine manner. This interested Johnny and led to his famous 1928 paper on the axiomatization of set theory. The goal of Hilbert was later shown to be impossible by Kurt Gödel's work, based on an axiom system similar to von Neumann's, which resulted in a theorem, published in 1930, to the effect that every axiomatic system sufficiently rich to contain the positive integers must necessarily contain undecidable propositions. After leaving Berlin in 1923 at the age of 20, von Neumann studied at the Eidgenossische Technische Hochschüle in Zurich, Switzerland, while continuing to maintain his enrolment at the University of Budapest. In Zurich he came into contact with the famous German mathematician, Hermann Weyl, and also the equally famous Hungarian mathematician, George Polya. He obtained a degree in Chemical Engineering from the Hochschüle in Zurich in 1925, and completed his doctorate in mathematics from the University of Budapest in 1926. In 1927 he became a privatdozent at the University of Berlin and in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1929 transferred to the same position at the University of Hamburg. His first trip to America was in 1930 to visit as a lecturer at Princeton University, which turned into a visiting professorship, and in 1931 a professorship. In 1933 he was invited to join the Institute for Advanced Study in Princeton as a professor, the youngest permanent member of that institution, at which Albert Einstein was also a permanent professor. Von Neumann held this position until he took a leave of absence in 1954 to become a member of the Atomic Energy Commission.

Von Neumann was married in 1930 to Marietta Kovesi, and his daughter Marina (who became a vice-president of General Motors) was born in 1935. The marriage ended in a divorce in 1937. Johnny's second marriage in 1938 was to Klara Dan, whom he met on a trip to Hungary. They maintained a very hospitable home in Princeton and entertained, on an almost weekly basis, numerous local and visiting scientists. Klara later became one of the first programmers of mathematical problems for electronic computers, during the time that von Neumann was its principal designer.

In 1938 Oskar Morgenstern came to Princeton University. His previous work had included books and papers on economic forecasting and competition. He had heard of von Neumann's 1928 paper on the theory of games and was eager to talk to him about connections between game theory and economics. In 1940 they started work on a joint paper which grew into their monumental book, *Theory of Games and Economic Behavior* published in 1944. Their collaboration is described in Morgenstern (1976).

Von Neumann became heavily involved in defence-related consulting activities for the United States and Britain during World War II. In 1944 he became a consultant to the group developing the first electronic computer, the ENIAC, at the University of Pennsylvania. Here he was associated with John Eckert, John Mauchly, Arthur Burks and Herman Goldstine. These five were instrumental in making the logical design decisions for the computer, for example, that it be a binary machine, that it have only a limited set of instructions that are performed by the hardware, and most important of all, that it run an internally stored program. It was acknowledged by the others in the group that the most important design ideas came from von Neumann. The best account of these years is Goldstine (1972). After the war von Neumann and Goldstine worked at the Institute of Advanced study where they developed (with others) the JONIAC computer, a successor to the ENIAC, which used principles some of which are still being used in current computer designs.

In 1943 von Neumann became a consultant to the Manhattan Project which was developing the atomic bomb in Los Alamos, New Mexico. This work is still classified but it is known that Johnny performed superbly as a mathematician, an applied physicist, and an expert in computations. His work continued after the war on the hydrogen bomb, with Edward Teller and others. Because of this work he received a presidential appointment to the Atomic Energy Commission in 1955. He took leave from the Institute for Advanced Study and moved to Washington. In the summer of 1955 he fell and hurt his left shoulder. Examination of that injury led to a diagnosis of bone cancer which was already very advanced. He continued to work very hard at his AEC job, and prepared the Silliman lectures (von Neumann, 1958), but was unable to deliver them. He died on 8 February 1957 at the age of 53 in the Walter Reed Hospital, Washington, DC.

The theory of games

Without question one of von Neumann's most original contributions was the theory of games, with which it is possible to formulate and solve complex situations involving psychological, economic,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
strategic and mathematical questions. Before his great paper on this subject in 1928 there had been only a handful of predecessors: a paper by Zermelo in 1912 on the solution in pure strategies of chess; and three short notes by the famous French mathematician E. Borel. Borel had formulated some simple symmetric two-person games in these notes, but was not able to provide a method of solution for the general case, and in fact conjectured that there was no solution concept applicable to the general case. For a commentary on the priorities involved in these two men's work see the notes by Maurice Frechet, translations (by L.J. Savage) of the three Borel papers, and a commentary by von Neumann, all of which appeared with von Neumann (1953a).

The three main results of von Neumann's 1928 paper were: the formulation of a restricted version of the extensive form of a game in which each player either knows nothing or everything about previous moves of other players; the proof of the minimax theorem for two-person zero-sum games; and the definition of the characteristic function for and the solution of three-person zero-sum games in normal form. Von Neumann also carried out an extensive study of simplified versions of poker during this time, but they were not published until later.

The *extensive* form of a game is the definition of a game by stating its rules, that is, listing all of the possible legal moves that a player can make for each possible situation he can find himself in during a play of the game. A *pure strategy* in a game is a much more complicated idea – a listing of a complete set of decisions for each possible situation in which the player can find himself. A complete enumeration of all possible strategies shows that the number of such strategies is equal to the product of the number of legal moves for each situation, which implies that there is an astronomical number of possible strategies for any non-trivial game such as chess. Most of these are bad, and would never be used by a skilful player, but they must be considered to find its solution. The *normalized* form of a game is obtained by replacing the definition of a game as a statement of its rules, as is done in its extensive form, by a listing of all of the possible pure strategies for each player. To complete the normalized form of the game, imagine that each player has made a choice of one of his pure strategies. When pitted against another a unique (expected) outcome of the game will result. For the moment we will imagine that the outcome of the game is monetary, and therefore each player gets a ‘payoff’ at the end of the game which is actually money. (Later we will replace money by ‘utility’.) If the sum of the payments to all players is zero the game is said to be *zero-sum*; otherwise it is a *non-zero-sum* game.

The normalized form of a game is also called a *matrix game*, and any real $m \times n$ matrix can be considered a two-person zero-sum game. The row player has m pure strategies, $i=1, \ldots, m$, and the column player has n pure strategies, $j=1, \ldots, n$. If the row player chooses i and the column player chooses j then the payoff $a(i, j)$ is exchanged between them, where $a(i, j)>0$ means that the row player receives $a(i, j)$ from the column player, while a negative payoff means that the column player receives the absolute value of that amount from the row player.

The importance of the careful analysis of the extensive and normalized forms of a game is that it separates out the concept of strategy and psychology in any discussion of a game. As an example, in poker bidding high when having a weak hand is commonly called ‘bluffing’, and considered an aggressive form of play. As a result of this formulation, and the solution of simplified versions of the game von Neumann showed that in order to play poker ‘optimally’ it is necessary to bluff part of the time, i.e., it is a required part of the strategy of any good poker player. A similar analysis for simplified bridge shows that a required part of an optimal bridge strategy is to signal, via the way one discards low cards in a suit, whether the player holds higher cards in that suit.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The analysis of special kinds of games shows that some of them can be solved by using pure strategies. This class includes the games of ‘perfect information’ such as the board games of chess and checkers. However, even such a simple game as matching pennies shows that an additional strategic concept is needed, namely, that of a ‘mixed strategy’. This concept appeared first in the context of symmetric two-person games in Borel's 1921 paper. Briefly, a mixed strategy for either player is a finite probability function on his set of pure strategies. For matching pennies the common strategy of flipping the penny to choose whether to play heads or tails is a mixed strategy that chooses both alternatives with equal probability (1/2), and is, in fact, an optimal strategy for that game.

We now discuss the way that von Neumann made precise the definition of a solution to a matrix game. Let A be an arbitrary $m \times n$ matrix with real number entries. Let x be an m-component row vector, and let f be an m-component column vector all of whose components are ones. Then x is a *mixed strategy* vector for the row player in the matrix game A if it satisfies: $xf=1$ and $x \geq 0$. Similarly, let y be an n-component column vector, and let e be an n-component row vector of all whose components are ones. Then y is a *mixed strategy* vector for the column player in the matrix game A if it satisfies: $ey=1$ and $y \geq 0$. Mixed strategy vectors are also called *probability* vectors because they have non-negative components that sum to one, and hence could be used to make a random choice of a pure strategy by spinning a pointer, choosing a random number, etc. To complete the definition of the solution to a game, we need a real number v, called the *value of the game*. The solution to the matrix game A is now a triple, a mixed strategy x for the row player, a mixed strategy y for the column player, and a value v for the game: these quantities must solve the following pair of (vector) inequalities:

$$xA \geq v \text{ and } Ay \leq vf.$$

Because these are linear inequalities, one might suspect (and would be correct) that the optimal x, y and v can be found by using a linear programming code and a computer.

However, in the 1920s it was not clear that such a solution existed. In fact, Borel conjectured that it did not. The most decisive result of von Neumann's 1928 paper was to establish, using an argument involving a fixed point theorem, his famous *minimax theorem* to the effect that for an arbitrary real matrix A there exists a real number v and probability vectors x and y such that

$$\underset{x}{\text{Maximum}}\,\underset{y}{\text{Minimum}}\,xAy = \underset{y}{\text{Maximum}}\,\underset{x}{\text{Maximum}}\,xAy$$

This theorem became the keystone not only for the theory of two-person matrix games, but also for n-persons games via the characteristic function (to be discussed later).

We now discuss the major differences between von Neumann and Morgenstern (1944) and von Neumann's 1928 paper. The information available to each player was assumed, in the 1928 paper, to be the following: when required to move, each player knows either everything about the previous moves of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
his opponents (as in chess), or nothing (as in matching pennies). By using information trees, and partitioning the nodes of such trees into information sets, in 1944 this concept was extended to games in which players have only partial information about previous moves when they are required to make a move. This was a difficult but major extension, which has not been substantially improved upon since its exposition in the 1944 treatise.

A second major change in the basic theory of games was in the treatment of payoff functions. In the 1928 paper payoffs were treated as if they were monetary, and it was implicitly assumed that money was regarded as equally important by each of the players. In order to take into account the well-known objections, such as those of Daniel Bernoulli, to the assumption that a dollar is equally important to a poor man as a rich man, a monetary outcome to a player was replaced by the *utility* of the outcome. Although Bernoulli had suggested that the utility of x dollars should be the natural logarithm of x, so that the addition of a dollar to a rich man's fortune would be valued less than the addition of a dollar to a poor man's fortune, this specific utility concept was never universally accepted by economists. So utility remained a fuzzy, intuitive concept. Von Neumann and Morgenstern made the absolutely decisive step of axiomatizing utility theory, making it unambiguous and they can properly be said to have started the modern theory of utility, not only for game theory, but for all of economics and the social sciences.

Almost two-thirds of the 1944 treatise consists of the theory of n-person constant-sum games, of which only a small part, the three person zero-sum case, appears in the 1928 paper. When $n>2$, there are opportunities for cooperation and collusion as well as competition among the players, so that there arises the problem of finding a way to evaluate numerically the position of each player in the game. In 1928 von Neumann handled this problem for the zero-sum case by introducing the idea of the *characteristic* function of a game defined as follows: For each *coalition*, that is, subset S of players, let $v(S)$ be the minimax value that S is assured in a zero-sum two-person game played between S and its complementary set of players.

To describe the possible division of the total gain available among the players the concept of an imputation, which is a vector $(x(1),\ldots, x(n))$ where $x(i)$ represents the amount the player i obtains, was introduced. For a coalition C in a constant-sum game $v(C)$ is the minimum amount that the coalition C should be willing to accept in any imputation, since by playing alone against all the other players, C can achieve that amount for itself. Except for this restriction there is no other constraint on the possible imputations that can become part of a solution. An imputation vector x is said to *dominate* imputation vector y if there exists a coalition C such that (1) $x(i)\geq y(i)$ for all i in C, and (2) the sum of $x(i)$ for i in C does not exceed $v(C)$. The idea is that that the coalition C can ‘enforce’ the imputation x by simply threatening to ‘go it alone’, since it can do no worse by itself.

One might think, or hope, that a single imputation could be taken as the definition of a solution to an n-person constant-sum game. However, a more complicated concept is needed. By a von Neumann–Morgenstern solution to an n-person game is meant a set S of imputations such that (1) if x and y are two imputations in S then neither dominates the other; and (2) if z is an imputation not in S, then there exists an imputation x in S that dominates z. Von Neumann and Morgenstern were unable (for good reasons, see below) to prove that every n-person game had a solution, even though they were able to solve every specific game they considered, frequently finding a huge number of solutions.

At the very end of the 1944 book there appears a chapter of about 80 pages on general non-zero-sum games. These were formally reduced to the zero-sum case by the technique of introducing a fictitious player, who was entirely neutral in terms of the game's strategic play, but who either consumed any

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
excess, or supplied any deficiency so that the resulting n+1 person game was zero-sum. This artifice helped but did not suffice for a completely adequate treatment of the non-zero-sum case. This is unfortunate because such games are the most likely to be found useful in practice.
About 25 years after the treatise appeared, William Lucas (1969) provided as a counter-example, a general sum game that did not have a von Neumann–Morgenstern solution. Other solution concepts have been considered since, such as the Shapley value, and the core of a game.
One of the most interesting non zero-sum games considered in that chapter was the so-called *market game*. The first example of a market game (though it was not called that) was the famous horse auction of Böhm-Bawerk, published in 1881. The horses had identical characteristics, each of 10 buyers had a maximum price he was willing to bid, and each of 8 sellers had a minimum price he was willing to accept. Böhm-Bawerk's solution was to find the ‘marginal pairs’ of prices, which turned out to be included in the von Neumann–Morgenstern solution to this kind of game. Later work on this problem was done by Shapley and Shubik (1972) and Thompson (1980, 1981).

The expanding economy model

Another of von Neumann's original contribution to economics was von Neumann (1937), which contained an expanding economy model unlike any other economic model that preceded it. When von Neumann gave a seminar to the Princeton economics department in 1932 on the model, which was stated in terms of linear inequalities not equations, and whose existence proof depended upon a fixed point theorem more sophisticated than any published in the mathematics literature of the time, it is little wonder that he made no impression on that group. He repeated his talk on the subject at Karl Menger's mathematical seminar in Vienna in 1936, and published his paper in German in 1937 in the seminar proceedings. The paper became more widely known after it was translated into English and published in *The Review of Economic Studies* in 1945 together with a commentary by Champernowne.
Von Neumann's model consists of a closed production economy in which there are m processes and n goods. In order to describe it we use the vectors e and f previously defined together with the following notation:

- x is the $m\times 1$ intensity vector with $xf=1$ and $x\geq 0$.
- y is the $1\times$price vector with $ey=1$ and $y\geq 0$.
- $\alpha=1+a/100$ is the expansion factor, where a is the expansion rate.
- $\beta=1+b/100$ is the interest factor, where b is the interest rate. The model satisfies the following axioms:
- Axiom 1.• $xB\geq \alpha xA$ or $x(B-\alpha A)\geq 0$.
- Axiom 2.• $By\leq \beta Ay$ or $(B-\beta A)y\leq 0$.
- Axiom 3.• $x(B-\alpha A)y=0$.
- Axiom 4.• $x(B-\beta A)y=0$.
- Axiom 5.• $xBy>0$.

Axiom 1 makes the model closed, i.e., the inputs for a given period are the outputs of the previous. Axiom 2 makes the interest rate be such that the economy is *profitless*. Axiom 3 requires that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

overproduced goods be *free*. Axiom 4 forces inefficient processes not to be used. And Axiom 5 requires the total value of all goods produced to be positive.

In order to demonstrate that for any pair of nonnegative matrices *A* and *B*, solutions consisting of vectors *x* and *y* and numbers α and β exist, an additional assumption was needed:

$$\text{Assumption V.} \quad A + B > 0.$$

This assumption means that every process requires as an input or produces as an output some amount, no matter how small, of every good. With this assumption, and the assumption that natural resources needed for expansion were available in unlimited quantities, von Neumann showed that necessarily $\alpha = \beta$, that is, that the expansion and interest factors were equal. In his paper, von Neumann proved a sophisticated fixed point theorem and used it to prove the existence theorem for the EEM.

D.G. Champernowne (1945) provided the first acknowledgement that the economics profession had seen the article, and also provided its first criticisms. We mention three:

1. (1) Assumption V which requires that every process must have positive inputs or outputs of every other good was economically unrealistic.
2. (2) The fact that the model has no consumption, so that labour could receive only subsistence amounts of goods as necessary inputs for production processes, also seems unrealistic.
3. (3) The consequence of Axiom 3 that overproduced good should be free is too unrealistic.

Criticisms 1 and 2 were removed by Kemeny, Morgenstern, and Thompson (1956), who replace Assumption V by:

Assumption KMT-1. Every row of *A* has at least one positive entry.

Assumption KMT-2. Every column of *B* has at least one positive entry. The interpretation of KMT-1 is that every process must use at least one good as an input. And the interpretation of KMT-2 is that every good must be produced by some process. With these assumptions they were able to show that there were a finite number of possible expansion factors for which intensity and price vectors existed satisfying the axioms. They also showed how consumption could be added into the model, which responded to criticism 2.

An alternative way of handling these criticisms appears in Gale (1956).

In Morgenstern and Thompson (1969, 1976), the third criticism above was answered by generalizing the model to become an ‘open economy’. In such an economy the price of an overproduced good cannot fall below its export price, and it cannot rise above its import price. Generalizations of the open model have been made by Los (1974) and Moeschlin (1974).

Von Neumann's influence on economics

Although von Neumann has only three publications that can directly be called contributions to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
economics, namely, his 1928 paper on the theory of games, his 1937 paper (translated in 1945) on the expanding economy model and his 1944 treatise (with Morgenstern) on the theory of games, he had enormous influence on the subject. The small *number* of contributions is deceptive because each one consists of several different topics, each being important. We discuss these separately.

The expanding economy model, von Neumann (1937) consisted of two parts: the first input–output equilibrium model that permits expansion; and second the fixed point theorem. The linear input–output model is a precursor of the Leontief model, of linear programming as developed by Kantorovich and Dantzig, and of Koopman's activity analysis. This paper, together with A. Wald (1935) raised the level of mathematical sophistication used in economics enormously. Many current younger economists are high-powered applied mathematicians, in part, because of the stimulus of von Neumann's work.

The theory of games, von Neumann (1928) and von Neumann and Morgenstern (1944), was an enormous contribution consisting of several different parts: (1) the axiomatic theory of utility; (2) the careful treatment of the extensive form of games; (3) the minimax theorem; (4) the concept of a solution to a constant-sum n-person game; (5) the foundations of non-zero-sum games; (6) market games. Each of these topics could have been broken into a series of papers, had von Neumann taken the time to do so. And he could have forged a brilliant career in economics by publishing them. However, he found that making an exposition of the results that he had worked out in notes or in his head was less interesting to him than investigating still other new ideas.

Von Neumann's indirect contributions, such as the theory of duality in linear programming, computational methods for matrix games and linear programming, combinatorial solution methods for the assignment problem, the logical design of electronic computers, contributions to statistical theory, etc. are equally, important to the future of economics. Each of his contributions, direct or indirect, was monumental and decisive. We should be grateful that he was able to do so much in his short life. His influence will persist for decades and even centuries in economics.

Selected works

1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320.

1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematische Kolloquiums* 8, ed. Karl Menger. Trans. as ‘A model of general equilibrium’, *Review of Economic Studies* 13, (1945–6), 1–9.

1944. (With O. Morgenstern.) *Theory of Games and Economic Behavior*. Princeton: Princeton University Press. 2nd edn, 1947; 3rd edn, 1953.

1947. Discussion of a maximum problem. Unpublished working paper, Princeton, November, 9 pp.

1948. A numerical method for determining the value and the best strategies of a zero-sum two-person game with large numbers of strategies. Mimeographed, May, 23 pp.

1953a. Communications on the Borel notes. *Econometrica* 21, 124–5.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1953b. (With G.W. Brown.) Solutions of games by differential equations. In *Contributions to the Theory of Games*, Vol. 1, ed. H.W. Kuhn and A.W. Tucker. Annals of Mathematics Studies No. 28, Princeton: Princeton University Press.

1953c. (With D.B. Gillies and J.P. Mayberry.) Two variants of poker. In *Contributions to the Theory of Games*, Vol. 1, ed. H.W. Kuhn and A.W. Tucker. Annals of Mathematics Studies No. 28, Princeton: Princeton University Press.

1954. A numerical method to determine optimum strategy. *Naval Research Logistics Quarterly* 1, 109–15.

1958. *The Computer and the Brain*. New Haven: Yale University Press.

1963. *Collected Works, Vols* I–VI. New York: Macmillan.

Bibliography

Champernowne, D.G. 1945–6. A note on J. von Neumann's article. *Review of Economic Studies* 13, 10–18.

Debreu, G. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. Cowles Foundation Monograph No. 17, New York: Wiley.

Gale, D. 1956. The closed linear model of production. In *Linear Inequalities and Related Systems*, ed. H. W. Kuhn and A.W. Tucker, Princeton: Princeton University Press. A counter-example showing that optimal prices need not exist in Gale's original model was published by J. Hulsman and V. Steinmitz in *Econometrica* 40, (1972), 387–90. Proof of the existence of optimal prices in a modified Gale model was given by A. Soyster in *Econometrica* 42 (1974), 199–205.

Goldstine, H.H. 1972. *The Computer from Pascal to von Neumann*. Cambridge, Mass: MIT Press.

Heims, S.J. 1980. *John von Neumann and Norbet Wiener*. Cambridge, Mass: MIT Press.

Kemeny, J.G., Morgenstern, O. and Thompson, G.L. 1956. A generalization of von Neumann's model of an expanding economy. *Econometrica* 24, 115–35.

Los, J. 1974. The existence of equilibrium in an open expanding economy model (generalization of the Morgenstern–Thompson Model). In *Mathematical Models in Economics*, ed. J. and M.W. Los. Amsterdam and New York: North-Holland Publishing Co.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lucas, W. 1969. The proof that a game may not have a solution. *Transactions of the American Mathematics Society* 137, 219–29.

Luce, R.D. and Raiffa, H. 1957. *Games and Decisions: Introduction and Critical Survey*. New York: John Wiley & Sons.

Moeschlin, O. 1974. A generalization of the open expanding economy model. *Econometrica* 45, 1767–76.

Morgenstern, O. 1958. Obituary, John von Neumann, 1903–57. *Economic Journal* 68, 170–74.

Morgenstern, O. 1976. The collaboration between Oskar Morgenstern and John von Neumann on the theory of games. *Journal of Economic Literature* 14, 805–16.

Morgenstern, O. and Thompson, G.L. 1969. An open expanding economy model. *Naval Research Logistics Quarterly* 16, 443–57.

Morgenstern, O. and Thompson, G.L. 1976. *Mathematical Theory of Expanding and Contracting Economies*. Boston: Health–Lexington.

Oxtoby, J.C., Pettis, B.J. and Price, G.B., eds. 1958. John von Neumann 1903–1957. *Bulletin of the American Mathematical Society* 64(3), Part 2.

Shapley, L.S. 1953. A value for n-person games. In *Contributions to the Theory of Games* II, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Shapley, L.S. and Shubik, M. 1972. The assignment game. I: the core. *International Journal of Game Theory* 1, 111–30.

Shubik, M. 1982. *Game Theory in the Social Sciences: Concepts and Solutions*. Cambridge, Mass.: MIT Press.

Shubik, M. 1985. *A Game Theoretic Approach to Political Economy*. Cambridge, Mass.: MIT Press.

Thompson, G.L. 1956. On the solution of a game-theoretic problem. In *Linear Inequalities and Related Systems*, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Thompson, G.L. 1980. Computing the core of a market game. In *Extremal Methods and Systems Analysis*, ed. A.V. Fiacco and K.O. Kortanek. Berlin: Springer-Verlag.

Thompson, G.J. 1981. Auctions and market games. In *Essays in Game Theory and Mathematical*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Economics in Honor of Oskar Morgenstern, ed. R.J. Aumann. et al., Mannheim: Bibliographisches Institut Mannheim.

Wald, A. 1935. Über die eindeutige positive Losbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines Mathematischen Kolloquiums*, ed. K. Menger 6, 12–20.

How to cite this article

Thompson, Gerald L. "von Neumann, John (1903–1957)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000043>
doi:10.1057/9780230226203.1802

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Vorob'ev, Nikolai N. (1925 – 1995)

Leon A. Petrosyan
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

coalitional games; game theory; Markov, A.; Morgenstern, O.; probability theory; von Neumann, J.; Vorob'ev, N.

Article

Nikolay Vorob'ev is commonly regarded as the founder and the leader of game-theoretic school in the former Soviet Union.
Vorob'ev was born on 18 September 1925 in Leningrad (now St Petersburg). His father was a lawyer and his mother a surgeon. Beginning his education at technical institutes in Izevsk and Moscow, he returned to Leningrad in 1944 and become a student at the Leningrad Shipbuilding Institute. In 1946 he began study at the Faculty of Mathematics and Mechanics of the Leningrad State University. In 1948 he left the Shipbuilding Institute and graduated from the university. In 1947 Vorob'ev published his first paper in semigroup theory.
In 1948 Vorob'ev started a postgraduate programme at the Leningrad branch of the Steklov Mathematical Institute. His supervisor was Professor A.A. Markov, under whose influence he studied constructive mathematical logic, which was rapidly developing at that time. His Candidate of Science thesis in mathematics was devoted to logical deduction rules in systems with strong negation. He received his Candidate of Science degree in 1952. In the same year he joined the Steklov Mathematical Institute as a junior research associate. Here he once more changed his scientific interests and started research concerned with both algebra and probability theory.
Axiomatic training in algebra and logic, along with studies in probability theory, permitted Vorobe'ev to make a transition to the study of game theory. His paper ‘Controlled Processes and Game Theory’ (1955) was the first paper in game theory published in the former Soviet Union. His 1959 review article ‘Finite Noncooperative Games’ served for many years as a primary Russian language source for understanding game theory. In the next five years Vorob'ev made an attempt to develop the theory of coalitional games, that is, games in which players belonging to one coalition are acting as one player, and therefore mixed strategies have to be defined as correlated families of measures. To prove

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the existence of stable outcomes in such games, he solved some non-standard problems from combinatorial topology and probability theory, thus combining ideas and methods from various branches of mathematics. At that time he also made interesting generalizations of H. Kuhn's equivalence theorem about behaviour strategies in extensive games with perfect recall, proposed an algorithm on enumerating equilibrium points in bimatrix games and studied games with forbidden situations. These results constituted the basis of his Doctor of Science thesis, which he defended in 1961. In the same year he organized the Soviet Union's first laboratory for game theory and operations research at the Steklov Mathematical Institute of the Academy of Sciences. Under his supervision more than 30 students obtained candidate and doctoral degrees in game theory. In 1968 Vorob'ev organized the first all-Union game theory conference in Erevan (Armenia) and the second in 1971 in Vilnius (Lithuania). He was the main speaker at both conferences, which each attracted more than 100 participants. His addresses focused on methodological and philosophical aspects of game theory as well as areas of applications. He forecast the application of game theory in economics, military affairs, biology, law, ethics, sociology, medicine and literature.

In 1975 his laboratory moved to the Institute for Socio-Economic Problems. Unfortunately, the administration of the institute considered any application of mathematical methods in social sciences as inconsistent with prevailing Marxist-Leninist dogmas. Game theory was no exception, which was why the laboratory was forced to concentrate on mathematical problems arising in game theory. Vorob'ev wrote an interesting monograph *Foundations of Game Theory: Noncooperative Games* (published in English translation in 1994) and considered it the first volume in a planned series of books on game theory. The second volume, ‘Cooperative Games’ was not completed. He also wanted to write a volume titled ‘Dynamic Games’.

Vorob'ev was a brilliant lecturer. He taught part-time at the Leningrad State University and many other universities in Russia and elsewhere, delivering courses in game theory, algebra, probability theory and number theory. He wrote many textbooks, the most popular among which is *Game Theory for Economists and System Scientists* (published in English translation in 1977). He edited most of the translations of the principal Western scientific monographs into Russian, including the famous *Theory of Games and Economic Behavior* by J. von Neumann and O. Morgenstern (1944). He also edited two bibliographic indices on game theory literature up to 1974. They contain about 5,000 summaries of game-theory books and papers from all over the world.

See Also

- Aumann, Robert J.
- coalitions
- game theory
- Harsanyi, John C.
- Morgenstern, Oskar
- Nash, John Forbes
- von Neumann, John

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Selected works

1955. Controlled processes and game theory. *Viestnik of Leningrad University* 4(11), 49.

1958. Equilibrium points in bi-matrix games. *Theory of Probability and its Applications* 3, 297–309.

1959. Finite non-cooperative games. *Russian Mathematical Surveys* 14(4), 21–56.

1960. About partitioned strategies. *Probability Theory and Applications* 5, 457–9.

1970. The present state of the theory of games. *Russian Mathematical Surveys* 25(2), 77–36.

1976. *Game Theory: Bibliographic Index*. Leningrad: Nauka.

1977. *Game Theory for Economists and System Scientists*. New York: Springer-Verlag.

1980. *Game Theory: Bibliographic Index*. Leningrad: Nauka.

1994. *Foundations of Game Theory: Noncooperative Games*. Boston: Birkhauser.

Bibliography

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

How to cite this article

Petrosyan, Leon A. "Vorob'ev, Nikolai N. (1925–1995)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_N000154>
doi:10.1057/9780230226203.1803

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

voting paradoxes

Donald G. Saari
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

After using an example to motivate why voting theory is so central to the social sciences, this survey describes some of the more recent (and, surprisingly, benign) interpretations of Arrow's Impossibility Theorem as well as explanations of the wide selection of voting paradoxes that drive this academic area. As described, it now is possible to explain all positional voting paradoxes while creating any number of illustrating examples.

Keywords

anti-plurality system; approval voting; Arrow, K.; axiomatic approach to decision rule choice; Borda Count; Condorcet, M.; cumulative voting; elections; Gibbard–Satterthwaite theorem; impossibility theorem; independence of irrelevant alternatives; Kruskal–Wallis test; Luce, D.; plurality vote; Sonnenshein–Mantel–Debreu theorem; transitivity; voting paradoxes; voting rules; Walras's Law

Article

Almost daily, news articles describe important elections being held somewhere in the world. The newsworthiness of these events is obvious: election outcomes can change the political, societal and economic directions of a city, a state, or even a country. Elections, in fact, are everywhere; their use ranges from legislative bodies busily determining laws to a kindergarten class selecting a recess treat ‘with a show of hands’. As elections are important, we impose safeguards such as the secret ballot. But a strong message coming from voting theory is that the choice of a voting rule can do more to frustrate the ‘will of the voters’ than any scheming, cigar-smoking political boss.
To illustrate this comment, consider the following three-candidate example where A>B>C means a voter prefers A to B to C. Let four voters prefer A>B>C, three prefer A>C>B, two prefer C>A>B, two prefer C>B>A, and six prefer B>C>A. With the:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- *plurality vote*, or ‘vote for one’, *A wins* with the A>B>C ranking;
- *Borda Count*, where 2, 1, 0 points are assigned, respectively, to a voter's first, second and third ranked candidate, *B wins* with the B>C>A ranking;
- *anti-plurality*, or ‘vote for two’, system, which is equivalent to voting against a candidate, *C wins* where its C>B>A ranking happens to reverse the plurality ranking.

Not all candidates reflect the ‘will of these voters’, yet each ‘wins’ by selecting an appropriate voting rule. Pairwise majority votes offer no help with their A>B, B>C, C>A cycle. The message is that, rather than capturing the views of the voters, an election outcome may more accurately reflect the *choice of the voting rule*.

More general rules include *n*-candidate positional methods defined by *n* weights $w_1, w_2, \ldots, w_n = 0$; $w_1 > 0$ and $w_j \geq w_{j+1}$ where w_j points are assigned to a voter's *j*th ranked candidate; candidates are ranked by the sums of assigned points. While (1, 0, 0), (2, 1, 0) and (1, 1, 0) represent the above rules, (8, 3, 0) is still another choice. Different weights, however, may generate other election outcomes. Indeed, the above example allows *seven* different positional election rankings. For instance, the (8, 3, 0) outcome is a fourth strict ranking B>A>C; the three remaining rankings involve ties.

One probable reason for the many different election rules is that inventing new ones is limited only by one's imagination; for example, positional methods define run-off rules whereby, after the bottom-ranked candidates are dropped, the remaining two are reordered. With our example, the plurality, Borda, and anti-plurality run-offs elect, respectively, A, B and B. Other approaches allow each *voter* to select a positional method to tally his ballot. With *cumulative voting*, for instance, a voter splits, say, three points in any integer manner; for example, she may use (3, 0, 0), or (2, 1, 0). *Approval voting* (AV) allows a voter to vote for (approve) any number of candidates; for example, he could select (1, 0, 0) or (1, 1, 0). But, whenever voters can determine how to tally their own ballots, we must anticipate that a single profile (that is, listing of voters’ preferences) can admit many different outcomes. Indeed, while changing positional methods generates seven different rankings for our example, *all 13* ways to rank three candidates are admissible cumulative or AV outcomes. Some theorists view this flexibility as a virtue (for example, Brams, Fishburn and Merrill, 1988); others treat this extreme indeterminacy as a serious failing (for example, Saari and Van Newenhizen, 1988).

As our example demonstrates, selecting an inappropriate voting or decision rule could inadvertently cause inferior outcomes – with negative concomitant consequences. This is not an isolated phenomenon: with conservative assumptions, about 69 per cent of contested three-candidate elections allow election rankings to change with different positional methods (Saari and Tataru, 1999). The percentage significantly increases with more candidates.

Further underscoring the complexity is Arrow's (1951) seminal impossibility theorem. He first requires voters to have complete (all pairs are ranked), transitive (a voter preferring A>B and B>C prefers A>C) preferences without restrictions, and the societal outcomes to be complete transitive rankings. Then Arrow characterizes all rules satisfying two basic properties. The first (Pareto) is a unanimity condition whereby, if everyone ranks a pair of candidates in the same manner, that is the societal ranking.

To motivate the second, ‘independence of irrelevant alternatives’ (IIA), condition with a reoccurring phenomenon in the judging of figure-skating, suppose a committee's ranking is Susan>Barb>Jeannie. Imagine Barb's anguish if, told that had more judges liked Jeannie, Barb would have ranked over Susan.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Why should the judges' opinion of Jeannie affect the (Susan, Barb) ranking? Arrow's 'independence of irrelevant alternatives' (IIA) condition prohibits this difficulty. Essentially, IIA requires each pair's ranking to depend only on each voter's relative ranking of this pair.

With these minimal conditions, Arrow proves that, for three or more candidates, the only admissible rule is a *dictator* – a specified voter whereby the societal outcome *always* agrees with her preferences independent of what other voters want. Understandably, Arrow's result is often interpreted to mean 'no voting rule is fair'. An alternative, significantly more benign explanation is given below.

The overpowering message is that the choice of a decision rule is crucial. Indeed, determining which rules are 'optimal' is the primary concern of voting theory, where finding axiomatic characterizations of rules, or discovering paradoxical examples, seems to dominate. Another approach (Luce, 1959) imposes structure on the outcomes; this structure determines what voting rules are admitted and what restrictions must be imposed on voter choices. A third, recent emphasis examines the data structure – voter preferences – to determine what the voters want and then which voting rules deliver the appropriate outcome (Saari, 2000).

For a template, treat a voting rule as a mapping from the domain (space of individual preferences) to the range (space of societal outcomes). The axiomatic approach emphasizes properties of the mapping, Luce's approach emphasizes the structure of the range, and my recent approach emphasizes the structure of the domain. All three approaches are briefly described.

Axiomatic approach and paradoxes

Borrowed from mathematics, a standard justification for the 'axiomatic approach' is that 'it tells us what we are getting'. After all, axioms are intended to form the fundamental building blocks of a theory, so axiomatic characterizations should specify what to expect from different voting rules. But this expectation requires the conditions to be true axioms; most often they are not. Instead, many results *uniquely identify* a rule in terms of special, perhaps idiosyncratic, properties rather than characterizing the rule. As an analogy, it is easy to envision settings where certain properties uniquely identify 'John' as a studious, well-behaved student, while different properties uniquely identify 'John' as a street-wise juvenile delinquent. By concentrating on particular traits, both sets of properties uniquely identify John, but neither completely describes nor characterizes him.

Similarly, many so-called 'axiomatic characterizations' of voting rules are, in reality, properties that inadvertently emphasize *special profiles*, so while they uniquely *identify* certain rules, they do not characterize them. As an example, certain technical assumptions plus the condition 'a candidate top-ranked by most voters wins' uniquely *identifies* the plurality vote. Alternatively, the same technical conditions accompanied with the 'with *n*-candidates, a candidate may win even if bottom-ranked by all but one more than 1/*n* of the voters' property also uniquely *identifies* the plurality vote. Neither is an axiomatic characterization: by depending on special profiles, neither really 'tells us what we are getting'.

This literature, however, identifies valued voting rule properties. Another widely used approach with the same objective is to find 'voting paradoxes', that is, unexpected outcomes. Indeed, the origin of this field derives from a 1770 example (published in Borda, 1781) that Borda constructed to question the plurality vote: with his example the C>B>A plurality outcome conflicts with the pairwise rankings that are consistent with A>B>C; his (2, 1, 0) Borda Count conclusion agrees with the pairwise rankings.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In contrast, Condorcet (1785) believed we should decide via pairwise comparisons: a Condorcet winner (loser) is the candidate who beats (loses to) all other candidates in majority pairwise votes. To distinguish his approach from Borda's, he constructed an example whereby the Condorcet winner is not top-ranked by the Borda Count – or any positional rule. The controversy over whether Borda's or Condorcet's method is superior continues: comments on this debate are given below.

With examples, Condorcet illustrated that his method can fail; for example, the *Condorcet triplet* A>B>C, B>C>A, C>A>B defines the pairwise cycle A>B, B>C, C>A where neither a Condorcet winner nor loser exists. Later I explain why Condorcet's example remains central to voting theory.

Others continuing Condorcet's philosophy explored ways to handle cyclic outcomes; for example, Dodgson's (1876) (Lewis Carroll from ‘Alice in Wonderland’) method finds the ‘closest’ Condorcet winner (that is, over all possible lists of pairwise rankings, find the list with a Condorcet winner that is ‘closest’ to the actual election tallies), while Kemeny's Rule finds the ‘closest’ transitive ranking. Surprisingly, as Ratliff (2001) proved, the Dodgson winner need not be Kemeny top-ranked; it can be anywhere within the Kemeny ranking. As Ratliff (2003) also proved with examples, if Dodgson's method is extended to select the top two, or top three, candidates, the outcomes need not be consistent; that is, examples exist where the Dodgson winner is not a Dodgson top-two candidate, and none of them is in the Dodgson top three. Voting behaviour is very complex.

‘Paradoxes’, then, identify new properties of voting rules. Nurmi (1999; 2002), for instance, creates several examples illustrating how major voting rules disagree over a wide selection of desirable properties. His work suggests it may be futile to select voting rules based on specified properties because no rule may satisfy all of them, and most surely there are other valued properties that we have yet to recognize. Fishburn creates many fascinating examples; one (1981) has a plurality ranking of A>B>C>D, but, if D drops out, the same voters have the plurality ranking of C>B>A; Fishburn's example illustrates an unexpected reversal property of the plurality vote.

Examples disclose subtle properties of voting rules, so a way to find all such properties is to find *everything* that can happen: that is, a profile defines a list – an election ranking for each possible subset of candidates. The goal is to find *all* lists that can be created with all possible choices of positional rules and all possible profiles. Call this collection of lists a ‘dictionary’. Entries in a dictionary, then, describe all possible ranking properties for all positional rules and even for methods, such as AV and run-offs, based on positional and pairwise rules. Even entries outside the dictionary describe properties; for example, lists of the (A>B>C, B>A, C>A, C>B) type, where some profile allows the pairwise rankings to reverse the positional ranking, never are in the Borda Dictionary, so, by being a missing listing, it describes a Borda consistency property.

Such dictionaries exist (for example, Saari, 1989; Saari and Merlin, 2000) showing, for instance, that most positional rules allow *anything* to happen. For instance, rank seven candidates in any desired manner. Next, re-rank the seven six-candidate subsets (created by dropping someone) in any desired manner; for example, if you wish, reverse the original ranking, or select them randomly. Continue doing so with each subset of five, four, three and two candidates. While the choices could be chaotic, a profile exists where the voters’ plurality ranking for each subset is the selected one. (The same conclusion holds for most choices of positional rules over the different subsets.) What provides hope from these dictionaries is that the Borda Count – defined by (n–1, n–2, …, 1, 0) – is the unique rule (when used with every subset of candidates) that significantly minimizes the number and kinds of allowed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
paradoxes. Thus, the Borda Count enjoys the maximum number of positive properties; for example, only Borda always ranks a Condorcet winner over a Condorcet loser.
A related ‘dictionary’ result (Saari, 1992a) proves that a ten-candidate profile exists where 9(9!) (recall, $9! = (9)(8)(7)\ldots(2)(1)$, so 9(9!) is over three million) different election rankings without ties result from changing the positional method; each candidate is top ranked with some rules and bottom ranked with others. (For *n*-candidates, up to (*n*–1)[(*n*–1)!] different strict election rankings can emerge from changes in positional methods.)

Luce's approach

Arrow (1951) proved that with three or more candidates no voting rule satisfying his conditions always has transitive outcomes. Luce (1959) adopted a different approach; he imposed constraints on admissible election outcomes. His conditions, which are described in terms of probabilities to reflect his interest in individual decisions, are stricter than Arrow's. Expressed in terms of voting, Luce requires a candidate's vote percentage to remain consistent over all subsets of candidates. For instance, if A, B, and C receive, respectively, 1/3, 1/2, and 1/6 of the vote, then in a pairwise comparison B beats A by receiving $(1/2)/[(1/3)+(1/2)] = 3/5$ of the vote. Luce's conditions, then, capture settings where a candidate's support is intrinsic; relative to other candidates, the support remains fixed over all sets of candidates even should new ones join.
The accompanying voting rule and admissible profiles are not specified; they are selected to be consistent with Luce's conditions. But, even with his strong conditions, the accompanying profile restriction with the plurality vote is surprisingly relaxed. Only limited extensions of this approach have been explored for voting theory, but more is possible for settings where candidates have intrinsic support.

Emphasizing the data

So far I have sampled ways to analyse voting rules through properties of the rules and by imposing restrictions on admissible election outcomes. It remains to explore how the domain structure – the individual preferences – sheds light on these rules. The approach mimics how we might determine whether an election outcome reflects the ‘will of the voters’: one way is to compare the outcome with what the voters say they want. To develop methodology, reverse the order: first determine what the voters want, and then determine which voting rules respect these outcomes.
To indicate how to determine what the voters want, consider tallying an Alice>Barb ‘22:20’ election outcome. One tallying approach combines an Alice and a Barb vote – a tie. After counting the 20 ties, Alice breaks the tie as she has two extra supporters. For more candidates, the approach is to determine configurations of preferences that arguably constitute ties. This provides a filter; if a voting rule fails to deliver a tie, expect it to introduce a bias in election outcomes. While this is the motivation, the technical objective is to find a coordinate system for the space of profiles. Different coordinates represent how portions of profiles influence different voting rules.
Such a coordinate system for profiles has been established for any number of candidates (Saari, 1999; 2000). For intuition about how this is done and the kinds of available results, the three-alternative setting (Saari, 1999) is outlined. The space of profiles is divided into three distinct coordinates, or subspaces,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
capturing:

- profiles that cause *all possible* positional method problems, but with no effect on pairwise rankings;
- profiles that cause *all* problems with pairwise majority votes, but with no effect on positional rankings; and
- profiles where no problems arise with any positional or majority vote rule.

The power of such a coordinate decomposition is apparent. As a sample:

- The coordinates allow us to explain properties of election rules. For instance, positional rules failing to have a tie for the first class of profiles can seriously disagree with pairwise majority vote outcomes.
- The second class of profiles explains problems dating to the 1780s about conflicts between pairwise and positional methods as well as agendas, tournaments and so forth.
- Conflicts associated with any profile, such as our initial one, can be explained; for example, finding the portions of a profile in each of these directions identifies why different rules have different election outcomes.
- Examples illustrating any possible paradox can be constructed. Start with a profile in the last class where there is complete agreement among all rules. To introduce a conflict with positional methods, add a profile portion from the first class; to create conflict with pairwise outcomes, add a profile portion from the second class.

To determine the first coordinate direction, we must find all profiles affecting only positional outcomes. While this is done mathematically, for an intuitive explanation combine a ranking with its reversal, for example, (A>B>C, C>B>A): it is arguable that the outcome should be a tie. It is a tie for majority votes over pairs. But with positional rules (w_1, w_2, 0), the A:B:C tallies are w_1:$2w_2$:w_1.where a tie occurs if and only if (iff) $w_1 = 2w_2$; that is, the desired tie occurs iff the Borda Count is used. If this configuration is used as a filter, then beware of a non-Borda rule. This is because, instead of a tie, rules with $w_1>2w_2$ (for example, the plurality vote) have an $A = C > B$ outcome, while rules with $w_1<2w_2$ (for example, the anti-plurality vote) have a $B > A = C$ outcome. Consequently, profiles exist where non-Borda positional rankings must differ from majority vote outcomes.
Surprisingly, *all possible differences* among three-candidate positional election rankings reflect how different rules handle these *reversal* profile components. Indeed, to create the initial example, I started with one voter with the B>C>A preference. To generate differences in positional outcomes, add x reversal units of (A>B>C, C>B>A) and y of (A>C>B, B>C>A). As the plurality and anti-plurality tallies for A:B:C are, respectively, $x+y$:y:x and $x+y$:$2x+y$:$x+2y$, algebra yields my $x = 2$, $y = 3$ choices creating the desired positional outcomes – and conflicts. (Borda is not affected by reversal terms, so its ranking remains the starting B>C>A.) As all possible positional differences are generated by reversal terms, any justification for one positional rule (for example, properties that uniquely identify one rule over others) must reduce to analysing the reversal component (A>B>C, C>B>A) tally.
The second coordinate direction, capturing all conflict among pairwise majority votes, is the Condorcet

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
triplet with its resulting cycle. This component is responsible for all pairwise voting mysteries, including the majority vote cycles, differences in Dodgson's and Kemeny's methods, problems with agendas, tournaments and so forth. This assertion holds for any number of candidates. To create a Condorcet *n*-tuple, start with an *n*-candidate ranking, say A>B>C>D>E. For the next ranking, place the top candidate on the bottom, creating B>C>D>E>A. Continue until each candidate is in first, second, …, last place precisely once. This configuration should define a tie, and it does for all positional methods. But the profile also creates majority vote cycles. Surprisingly, these profile coordinate components cause all possible pairwise problems.

To illustrate with our initial example, start with the B>C>A preference. Adding *z* units of (A>B>C, B>C>A, C>B>A) results in A:B, B:C, C:A pairwise votes of, respectively, 2*z*:1+*z*, 2*z*+1:*z*, 2*z*:1:*z*. So $z = 2$ creates the desired cycle. Adding these reversal and Condorcet terms to the starting ranking yields the initial example.

The remaining coordinate directions, where nothing goes wrong, are called *Basic* directions. For candidate A, it consists of two preferring A>B>C, two preferring A>C>B, one preferring B>A>C, one preferring C>A>B; that is, two for each ranking where A is top-ranked, one for each where A is second-ranked. More generally with *n*-candidates, candidate X's Basic direction has (*n*–*j*) voters with each ranking where X is *j*th ranked. While not intuitive, these coordinate directions come from mathematics. The important point is that no conflict occurs in this profile space; for example, the tallies for *any* voting rule for all candidates identifies the tally for *all* voting rules over any subset of candidates. Nothing goes wrong. These three kinds of directions span the six dimensions of profile space, so they complete the three-alternative analysis. (A profile, of course, normally has only parts in each direction.)

Explaining all differences

All possible differences among three-candidate standard voting rules, then, reflect how voting rules react to reversal and Condorcet profile components. The many desirable properties of the Borda Count, for instance, arise because it is the only rule based on positional and majority votes that always delivers a tie for these components.

I indicated how all positional differences reflect how positional rules treat reversal terms, so it remains to describe the Condorcet components. For motivation, suppose three voters must vote for one of two candidates from each of three schools. Suppose the candidates are [Anne, Bob], [Connie, Dave], [Ellen, Fred]. Does a [Bob, Dave, Fred] outcome, each by 2:1, reflect the voters' views? To answer this question without knowing the actual preferences, all supporting preferences must be listed.

Four of the five profiles have two voters selecting different candidates from each school; this causes a tie. Breaking the tie is the last voter's [Bob, Dave, Fred] preference. The fifth profile has the preferences [Anne, Dave, Fred], [Bob, Connie, Fred], [Bob, Dave, Ellen]. It is difficult to argue against the outcome for the first four profiles as a tie is broken. At least statistically, then, the outcome respects most supporting profiles. But it is difficult to justify the fifth 'outlier' profile other than pointing to the 2:1 votes.

While most profiles justify the conclusion, suppose the fifth 'outlier' profile is the actual one where each voter wanted to elect a woman and a man. The profile reflects their wishes; the outcome does not. The reason is clear: the majority vote strictly emphasizes information about specific pairs; it ignores information – even intended relationships – among pairs. Consequently, rather than recognizing the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
added ‘balanced gender’ condition, the majority vote must ignore it.
To connect this example with the Condorcet triplet, identify Anne = B > A, Bob = A > B; Connie = C > B, Dave = B > C; Ellen = A > C, Fred = C > A: the Condorcet triplet becomes the outlier ‘fifth profile’, and the ‘balanced gender condition’ is equivalent to ‘transitivity’. Because any argument applied to one setting transfers to the other, it follows that the cyclic outcome for the Condorcet triplet (the ‘paradox of voting’) occurs because (*a*) this outcome reflects most supporting profiles (even though, by involving cyclic preferences, they are not admitted), and (*b*) the majority vote strips all connecting information, *including transitivity*, from the profile. (*c*) While majority pairwise voting may suffice if candidates have ‘intrinsic support’, it can distort outcomes for usual cases.
In general:

- Pairwise outcomes reflect the average over *all possible* supporting profiles; paradoxes, such as with the Condorcet triplet, indicate that the actual profile is an outlier relative to the average.
- Majority votes strip away all intended relationships, including transitivity, from the profile.
- Whenever intended relations are dropped, they come from profile portions based on Condorcet *n*-tuples.

Explaining mysteries

The above structure explains several mysteries. The ones described here compare the Borda and Condorcet rules, briefly discuss all rules based on pairwise outcomes, and explain Arrow's Impossibility Theorem.
As indicated, for any number of candidates all possible differences between the Borda and pairwise rankings manifest the majority vote's reaction to Condorcet *n*-tuples, which introduce cyclic affects. As an illustrating example, with two preferring A>B>C, and one preferring B>A>C, both the Borda and pairwise rankings reflect A>B>C. Adding x units of the Condorcet [B>A>C, A>C>B, C>B>A] never affects the Borda ranking, but its cyclic effect changes the A:B, B:C, C:A pairwise tallies to $2+x:2x+1$, $x+3:2x$, $2:2x+3$ where $x = 2$ makes B the Condorcet winner, $x \geq 4$ creates a cycle.
Any difference between the Borda and Condorcet winners, then, reflects Condorcet profile components. Thus, any argument supporting Condorcet over Borda must justify something other than a tie for a Condorcet triplet or *n*-tuple.
Voting rules relying on majority vote pairwise rankings, such as Kemeny's and Dodgson's rules, inherit the majority vote difficulties caused by Condorcet *n*-tuples. As these rules are primarily intended to handle cyclic behaviour, their value presumably emerges when the Condorcet component is dominant. But the stripping action of the majority vote over these components means that, unexpectedly, the rule cannot use information about the voters’ transitive preferences. Consequently, if the transitivity of voter preferences is valued, such rules should not be used. If transitivity is not valued, we must question using rules that impose transitivity on the outcomes.
A similar analysis holds for Arrow's Theorem (Saari, 2001). An unexpected feature of IIA, as with the majority vote, is to strip from the decision rule all information that individuals have transitive preferences. But, if the rule cannot use the transitivity of individual preferences, then transitive societal outcomes cannot be expected unless profiles are severely restricted; that is, the societal outcome reflects the imposed data structure rather than properties of the rule. One severe restriction is to use the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
preferences of a single voter; this explains Arrow's dictator.
As Arrow's negative result is strictly caused by IIA unintentionally stripping away valued information about individual preferences, resolutions must modify IIA to allow the rule to use this information. To illustrate, a transitive ranking, say A>B>C, separates some alternatives from others. Listing these separations as [A>B, 0], [B>C, 0], [A>C, 1] provides information about the transitive individual preferences. Let IIIA (*Intensity IIA*) be where a pair's societal ranking is determined by how each voter ranks the pair *and the number of separating alternatives*. By replacing IIA with IIIA in Arrow's conditions, Arrow's dictator is replaced with the Borda Count, and rules based on the Borda Count.

Strategic behaviour

Beyond the above ‘single-profile’ problems, multiple-profile concerns catalogue interesting changes in outcomes by changing a profile. They include the seminal Gibbard (1973)–Satterthwaite (1975) theorem asserting that, with three or more alternatives, no decision rule is immune from strategic behaviour: that is, with any rule, situations exist where some voter ensures a personally better outcome by voting according to other than her true preferences. There is, in fact, a host of related behaviour; see, for example, Nurmi (1999; 2002). Some rules, for instance, can cause a winning candidate to *lose* by attracting more supporting voters. Similarly, Fishburn and Brams (1983) discovered the ‘no-show’ paradox where, with the plurality run-off, a voter obtains a personally better outcome by *not* voting. These results reflect the higher dimensionality of profiles that accompanies added alternatives. With two candidates, a voter can vote for, or against, her favorite. With more alternatives, beyond her top and bottom choice, a voter can consider intermediate options. As suggested by the ‘don't waste your vote’ cry for strategic voting, situations exist where, by voting strategically, some voters can block personally lower-ranked candidates from winning. The Gibbard–Satterthewaite result proves this happens for all realistic rules.
A common source of problems, such as the no-show paradox, or where two subcommittees elect ‘A’ but the combined committee does not, and so forth, is when the rule loses monotonicity. Positional methods are monotonic; that is, with added support a candidate has higher tallies. But difficulties occur with rules involving several subsets of candidates; for example, a run-off involves {all *n*-candidates} and {top two}. What causes problems is that the first election determines who is advanced to the second. Consequently, added support for a winning candidate could also advance a stronger opponent to the run-off.

Implications for economics

Voting rules are aggregation methods: voters’ preference rankings are aggregated into a societal ranking. But as much of economics, and the social sciences, also involves aggregation rules, we must anticipate that the behaviour of voting rules predicts behaviour elsewhere in economics and other disciplines. This happens. As illustrations, the above result allowing 9(9!) different positional election rankings for a single ten-candidate profile, where almost any specified outcome can occur, has a parallel with the Sonnenshein (1972)–Mantel (1972)–Debreu (1974) Theorem asserting that any continuous function satisfying Walras's Laws can be (up to minor technical conditions on prices) the aggregate excess

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
demand function for some exchange economy. As another example, recall the voting result stating that, even if the rankings for the different subsets of candidates are selected in an arbitrary manner, a supporting profile can be found. The same behaviour arises in economics. The voting result allowing a ranking to be selected for each subset of candidates, and a profile can be found so that the selected ranking is the actual election ranking also has an economic parallel: that is, the Sonnenshein–Mantel–Debreu Theorem extends to where a different function can be selected for each subset of commodities, and an economy (initial endowment and utility function for each agent) can be found so that (with the same technical condition) the aggregate excess demand for each subset is the selected one (Saari, 1992b).

Voting results have parallels in non-parametric statistics, namely, select rankings for each subset of alternatives: for most non-parametric rules, a data-set can be found so that each set's actual ranking is the selected one. In voting, the positional rule most immune from the ‘anything can happen’ difficulty is the Borda Count. In nonparametric statistics, the Kruskal–Wallis test has similar properties (Haunsperger, 1992).

See Also

- democratic paradoxes
- paradoxes and anomalies
- rational choice and political science

Bibliography

Arrow, K. 1951. *Social Choice and Individual Values*. Cowles Foundation Monograph No. 17. New York: Wiley.

Borda, J. 1781. *Mémoire sur les élections au scrutin*. Paris: Histoire de l'Académie Royale des Sciences.

Brams, S., Fishburn, P. and Merrill, S. 1988. The responsiveness of approval voting: comments on Saari and Van Newenhizen. *Public Choice* 59, 112–31.

Condorcet, M. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1, 15–23.

Dodgson, C. 1876. A method for taking votes on more than two issues. In *Classics of Social Choice*, ed. I. McLean and A. Urken. Ann Arbor: University of Michigan Press, 1995.

Fishburn, P. 1981. Inverted orders for monotone scoring rules. *Discrete Applied Mathematics* 3, 27–36.

Fishburn, P. and Brams, S. 1983. Paradoxes of preferential voting. *Mathematics Magazine* 56, 207–14.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Haunsperger, D., 1992. Dictionaries of paradoxes for statistical tests on k-samples. *Journal American Statistical Association* 87, 149–55.

Luce, D. 1959. *Individual Choice Behavior*. Mineloa, NY: Dover Publications, 2005.

Mantel, R. 1972. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7, 348–53.

Nurmi, H. 1999. *Voting Paradoxes and How to Deal with Them*. New York: Springer-Verlag.

Nurmi, H. 2002. *Voting Procedures Under Uncertainty*. New York: Springer-Verlag.

Ratliff, T. 2001. A comparison of Dodgson's Method and Kemeny's Rule. *Social Choice & Welfare* 18, 79–89.

Ratliff, T. 2003. Some startling paradoxes when electing committees, *Social Choice & Welfare* 21, 433–54.

Saari, D. 1989. A dictionary for voting paradoxes. *Journal of Economic Theory* 48, 443–75.

Saari, D. 1992a. Millions of election rankings from a single profile. *Social Choice & Welfare* 9, 277–306.

Saari, D. 1992b. The aggregate excess demand function and other aggregation procedures. *Economic Theory* 2, 359–88.

Saari, D. 1999. Explaining all three alternative voting outcomes. *Journal of Economic Theory* 8, 313–55.

Saari, D. 2000. Mathematical structure of voting paradoxes. *Economic Theory* 15, 1–101.

Saari, D. 2001. *Decisions and Elections*. New York: Cambridge University Press.

Saari, D. and Merlin, V. 2000. A geometric examination of Kemeny's rule. *Social Choice & Welfare* 17, 403–38.

Saari, D. and Tataru, M. 1999. The likelihood of dubious election outcomes. *Economic Theory* 13, 345–63.

Saari, D. and Van Newenhizen, J. 1988. Is approval voting an ‘unmitigated evil?’ *Public Choice* 59,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
133–47.

Satterthwaite, M. 1975. Strategyproofness and Arrow's conditions. *Journal of Economic Theory* 10, 187–217.

Sonnenshein, H. 1972. Market excess demand functions. *Econometrica* 40, 649–63.

How to cite this article

Saari, Donald G. "voting paradoxes." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_V000069> doi:10.1057/9780230226203.1804

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wade, John (1788 – 1875)

Daniele Besomi
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

John Wade was born in London to working class parents. He worked for more than a decade as a journeyman wool sorter, then he ‘wrote his way out of obscurity’ (Harling, 2004). He was encouraged by Francis Place to engage in journalism: his first venture was a penny newspaper, *The Gorgon*, published in 1818–19 on money lent by Bentham and Henry Bickersteth (later Baron Langdale). Wade's articles are reputed to be well informed and detailed, so that *The Gorgon*'s influence surpassed its limited circulation. It attempted to find a junction point between, on the one hand, radical reformers and trade unionists, to which group Wade belonged, and, on the other hand, moderate reformers, with particular reference to the possible use of utilitarian doctrines to improve the condition of the labouring classes.

Keywords

Bentham, J.; capital–labour relations; cobweb theorem; Place, F.; speculation; technological progress; utilitarianism; Wade, J.

Article

In 1819 Wade published anonymously in two-penny sheets *The Black Book*, where he documented in detail the revenues and privileges of the ‘parasitic’ classes: clergy, aristocracy, and anyone connected with the government. The book (1820), later qualified as a handbook of radical agitators, was very successful, with over 50,000 copies sold in its various editions. In 1828 Wade joined the staff of the newly founded *Spectator*, and in the course of his life he wrote several inquiries and a number of compendia on such topics as *Manchester Massacre* (1819), *A Political Dictionary … Chiefly Designed for the Use of Members Of Parliament …* (1821), *A Popular Digest of the Laws of England* (1826, 24th edition in 1869), *Digest of Facts and Principles on Banking* (1826), *British History, Chronologically Arranged* (1839), *Principles of Money, with their Application to the Reform of the Currency and of Banking* (1842), the last book on *Harlotry and Concubinage* being dated 1859. His juvenile positions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
became more widely accepted in later years, and in 1862 he was granted a £50 annual pension by Lord Palmerston. He died in Chelsea in 1875.

Wade's *History of the Middle and Working Classes* (1833) offers interesting insights into the nature of capital and its relation to labour, but especially on periodic fluctuations in prices and activity. While most of Wade's contemporaries focused on crises, conceived either as accidental events or, more rarely, as periodically returning, Wade offered one of the earliest, if not *the* earliest, dynamic models of endogenous cycles in individual industries. Having recognized the existence of a ‘commercial cycle’ showing some ‘periodic regularity’ and recurring every five to seven years (Wade, 1833, pp. 211, 255), he offered an explanation consisting essentially in a cobweb-like mechanism. Price movements trigger changes in both demand and production, which in turn react against the original movement. When prices rise, demand falls while production increases. Supply thus outpaces demand, and prices fall, setting off the opposite movement (1833, p. 254). The cycle results from the system's tendency to correct price fluctuations. The assumption, implicit in this part of the argument, that reactions are either slow or lagged was made explicit a few pages later where Wade observed that the introduction of new machinery takes some time to fully develop its consequences on production and prices (1833, p. 257). On top of this mechanism, Wade considered a number of relieving or aggravating circumstances, including foreign competition, changes in fashion or technological progress. Notably, he saw ‘illusive speculation’ of an ‘epidemic character’, addressed ‘to the passion and not to the reason of mankind’, as the generator of crises. This argument was common at the time, but while his contemporaries saw in this the root of the problem, Wade was adamant that this was an aggravating cause only, superimposing onto the main mechanism.

See Also

- Bentham, Jeremy
- cobweb theorem
- utilitarianism and economic theory

Selected works

1820. *The Black Book, or, Corruption Unmasked!*…. London: Fairburn.

1833. *History of the Middle and Working Classes*. London: Effingham Wilson. Reprinted: New York: Kelly, 1966.

Bibliography

Besomi, D. John Wade's early endogenous dynamic model: ‘commercial cycle’ and theories of crises. *European Journal of the History of Economic Thought*, 15:4, December 2008.

Edmund, J.C. and Frow, R. 1994. John Wade and the Black Book. *Labour History Review* 59(2), 55.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Harling, P. 2004. Wade, John (1788–1875). *Oxford Dictionary of National Biography*. Oxford: Oxford University Press.

von Bergmann, E. 1895. *Die Wirtschaftskrisen: Geschichte der nationalökonomischen Krisentheorien*. Stuttgart: W. Kohlhammer.

Zegger, R.E. 1979. Wade, John. In *Biographical Dictionary of Modern British Radicals*, vol. 1, ed. J.O. Baylen and N.J. Gossman. Hassocks: Harvester Press.

How to cite this article

Besomi, Daniele. "Wade, John (1788–1875)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000137>
doi:10.1057/9780230226203.1805

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

wage curve

David G. Blanchflower and Andrew J. Oswald
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article summarizes evidence for the existence of a wage curve – a downward-sloping relationship between the level of pay and the local unemployment rate – in modern micro data. At the time of writing, the curve has been found in 40 nations. Its elasticity is approximately –0.1.

Keywords

labour economics; microfoundations; Phillips curve; Phillips, A. W.; power laws; unemployment; wage curve

Article

The wage curve is a statistical regularity or empirical ‘law’ of economics. It traces out, as in Figure 1, a downward-sloping relationship between wages and local unemployment. Its elasticity is approximately – 0.1. Although this kind of downward-sloping shape has since been replicated in many other nations and by many other authors, Figure 1 here is reproduced from work on US data by Blanchflower and Oswald (1994, p. 134). The *y* axis, here labelled ‘antilog’, is a measure of the level of pay.

Figure 1
United States wage curve, 1963–1987. *Source*: Blanchflower and Oswald (1994, p. 134).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

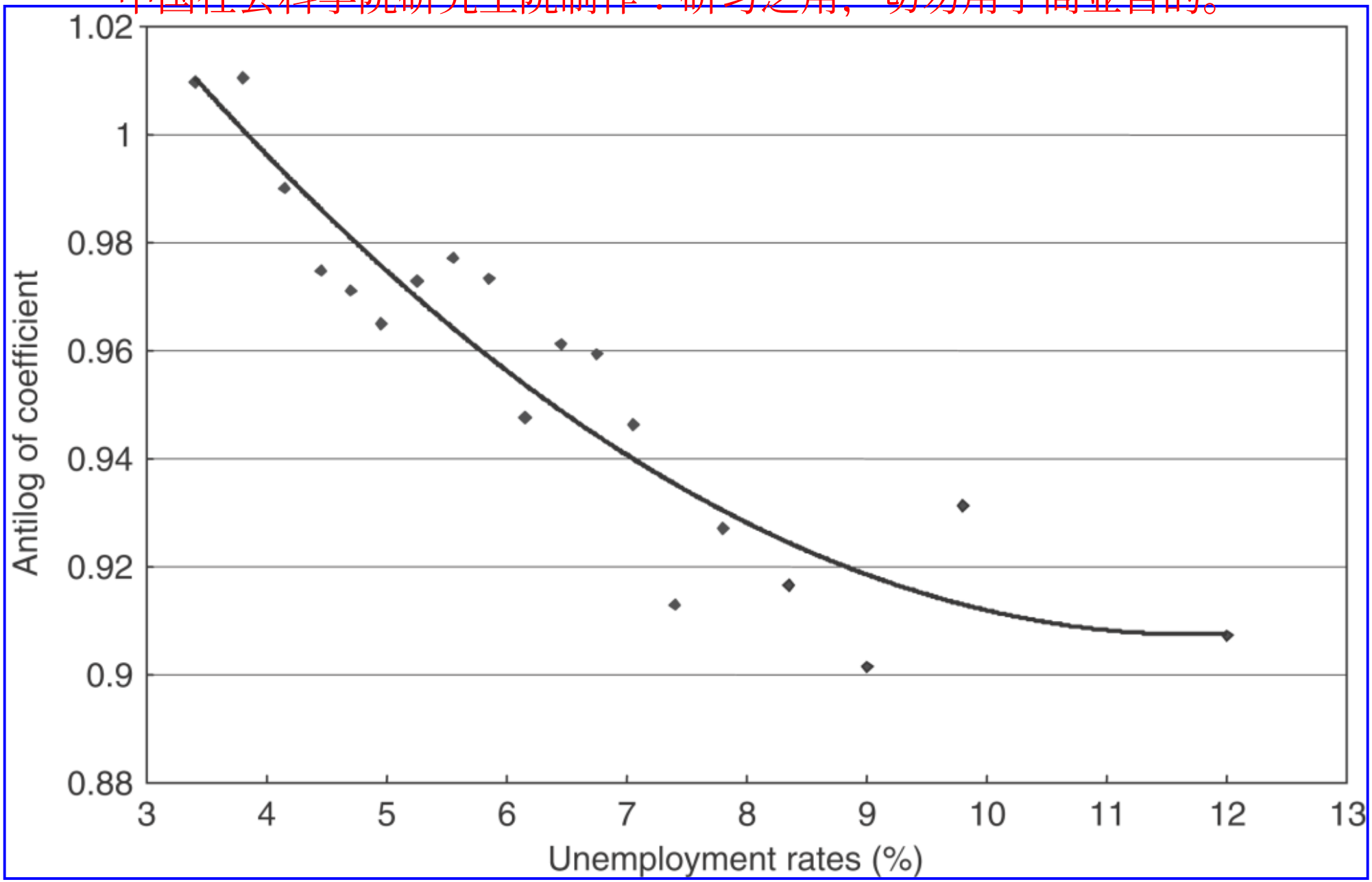

As an example, consider two regions within a country. Assume Region A's unemployment rate is double that in Region B. The wage-curve finding states that a worker's wage will then be ten per cent lower in Region A than the wage of an identical worker in Region B.

To understand the wage curve's place in intellectual history, it is useful to go back to one of the oldest questions in economics, namely, that of how the price of labour is affected by the unemployment rate. Following an empirical tradition begun by the New Zealand economist A.W. Phillips (1958), this issue has traditionally been studied with aggregate time-series methods. Although its robustness is still questioned, the Phillips curve, which is a relationship between wage growth and unemployment, has become part of the bedrock of macroeconomics textbooks. Sargan (1964) pointed out that it was possible to view the Phillips curve as being consistent with a steady-state solution where the level of pay depends on the level of unemployment.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Blanchflower and Oswald (for example, 1994; 1995) were among the first to argue instead for the use of microeconomic data in such a setting. Their book does not study the Phillips curve, nor does it use aggregate data. Instead, using samples of individual workers, the authors document the existence of a logarithmic curve – what physicists would call a power law – linking the level of the wage to the unemployment rate in the local area. They conclude that in 16 nations, including the United States, the data are well described by a wage curve with an unemployment elasticity of approximately –0.1.

Since then, those conclusions have been checked, and largely replicated, by other researchers and on different nations' data. Examples include Hoddinott (1996) for the Côte d'Ivoire; Janssens and Konings (1998) for Belgium; Sabin (1999) for China; Bellmann and Blien (2001) for Germany; and Garcia-Mainar and Montuenga-Gomez (2003) for Spain. A recent study by Sanz-de-Galdeano and Turunen (2006) has used a large longitudinal data-set on workers across the Eurozone and, once again, obtained a similar elasticity.

Evidence for a wage-curve pattern has been found in more than 40 countries. Its existence in the United States, however, is currently viewed as somewhat more controversial. One reason is that Blanchard and Katz (1997) argue for a Phillips curve, rather than a wage curve, in United States data. Staiger, Stock and Watson (2002) and Card and Hyslop (1997) also report a high level of autoregression in US wages. In contrast, Hines, Hoynes and Krueger (2001) conclude that a wage curve specification has a more natural theoretical interpretation and fits the data (hours as well as wages) better than the Phillips curve specification. The authors produce evidence of wage curves using annual and hourly earnings from the 1977–2000 March Current Population Survey files. The authors also uncover wage curves in the Panel Study of Income Dynamics (PSID). Using the PSID, Hines, Hoynes and Krueger suggest that a three percentage point decline in the unemployment rate is associated with a four per cent increase in real wages, which translates into an elasticity similar to the Blanchflower–Oswald number. Recently, Blanchflower and Oswald (2005) returned to the topic of the wage curve, and, in modern US data, argued that the United States has a long-run wage curve with the usual elasticity of –0.1 but that their 1994 book should have paid more attention to the high degree of autoregression in US state wages.

The wage curve seems relevant beyond its implications for labour economics. First, macroeconomic analysis has for some decades stressed the need for microeconomic foundations. Second, some macroeconomics textbooks make extensive theoretical use of a wage curve (at the aggregate level), but do not provide evidence for it.

Wage curves have been reported for Argentina, Australia, Austria, Belarus, Belgium, Brazil, Bulgaria, Burkina Faso, Canada, Chile, China, Côte d'Ivoire, Czech Republic, Denmark, East Germany, Estonia, Finland, France, Great Britain/UK, Holland, Hungary, India, Ireland, Italy, Japan, Latvia, New Zealand, Norway, Poland, Portugal, Romania, Russia, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Taiwan, Turkey, USA and West Germany. These studies are summarized in Blanchflower and Oswald (2005). A meta-analysis – on a sample of 208 wage–unemployment wage curve elasticities from the literature – by Nijkamp and Poot (2005, p. 445) concludes that

> the wage curve is a robust empirical phenomenon … but there is … evidence of publication bias. There is indeed an uncorrected mean estimate of about –0.1 for the elasticity. After controlling for publication bias by means of two different methods, we estimate that the 'true' wage curve elasticity at the means of study characteristics is about –0.07.

Why the wage curve exists, however, is not so well understood. One way to rationalize such a pattern is to appeal to non-competitive theories of the labour market – for example, to the idea of a no-shirking condition or a Nash bargaining-power locus. According to this kind of analytical framework, high local unemployment makes life tougher for workers (because, for example, they will find it harder to obtain work if laid off by their current employer), and therefore it is not necessary for employers to remunerate them so generously. The wage curve is then potentially an important element of a theory of equilibrium in the labour market such as in Shapiro and Stiglitz (1984) or Pissarides (2000).

Whatever the correct theoretical interpretation, new empirical results continue to emerge. Even in South Africa, where unemployment rates have run as high as 30 per cent, Kingdon and Knight (2006) conclude that there is a wage curve with an elasticity of –0.1. Although its conceptual foundations will go on being debated, and more research, especially for the United States, is required, the wage curve appears to be a pattern that holds in many nations.

See Also

- efficiency wages
- Phillips curve (new views)

We thank Bruce Weinberg for helpful suggestions.

Bibliography

Bellmann, L. and Blien, U. 2001. Wage curve analyses of establishment data from Western Germany. *Industrial and Labor Relations Review* 54, 851–63.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Blanchard, O. and Katz, L.F. 1997. What we know and do not know about the natural rate of unemployment. *Journal of Economic Perspectives* 11(1), 51–72.

Blanchflower, D.G. and Oswald, A.J. 1994. *The Wage Curve*. Cambridge, MA: MIT Press.

Blanchflower, D.G. and Oswald, A.J. 1995. An introduction to the wage curve. *Journal of Economic Perspectives* 9(3), 153–67.

Blanchflower, D.G. and Oswald, A.J. 2005. The wage curve reloaded. Working Paper No. 11338. Cambridge, MA: NBER.

Card, D. and Hyslop, D. 1997. Does inflation ‘grease the wheels’ of the labor market? In *Reducing Inflation: Motivation and Strategy*, ed. C.D. Romer and D.H. Romer. NBER Studies in Business Cycles, vol. 30. Chicago and London: University of Chicago Press.

Garcia-Mainar, I. and Montuenga-Gomez, V. 2003. The Spanish wage curve: 1994–1996. *Regional Studies* 37, 929–45.

Hines, J., Hoynes, H. and Krueger, A.B. 2001. Another look at whether a rising tide lifts all boats. In *The Roaring Nineties: Can Full Employment Be Sustained?* ed. A.B. Krueger and R.M. Solow. New York: Russell Sage Foundation.

Hoddinott, J. 1996. Wages and unemployment in an urban African labour market. *Economic Journal* 106, 1610–26.

Janssens, S. and Konings, J. 1998. One more wage curve: the case of Belgium. *Economics Letters* 60, 223–27.

Kingdon, G.G. and Knight, J. 2006. How flexible are wages in response to local unemployment in South Africa? *Industrial and Labor Relations Review* 59, 471–95.

Nijkamp, P. and Poot, J. 2005. The last word on the wage curve? A meta-analytic assessment. *Journal of Economic Surveys* 19, 421–50.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25, 283–99.

Pissarides, C.A. 2000. *Equilibrium Unemployment Theory*, 2nd edn. Cambridge, MA: MIT Press.

Sabin, L. 1999. The development of urban labour markets: China's urban wage curve, 1980–92. *Journal of Development Studies* 35(3), 134–52.

Sanz-de-Galdeano, A. and Turunen, J. 2006. The euro area wage curve. *Economics Letters* 92, 93–8.

Sargan, J.D. 1964. Wages and prices in the United Kingdom: a study in econometric methodology. In *Econometrics and Quantitative Economics*, ed. D.F. Hendry and K.F. Wallis. New York: Basil Blackwell, 1984.

Shapiro, C. and Stiglitz, J. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74, 433–44.

Staiger, D., Stock, J.H. and Watson, M.W. 2002. Prices, wages and the U.S. NAIRU in the 1990s. In *The Roaring Nineties: Can Full Employment be Sustained?* ed. A. Krueger and R. Solow. New York: Russell Sage Foundation.

How to cite this article

Blanchflower, David G. and Andrew J. Oswald. "wage curve." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000135> doi:10.1057/9780230226203.1806

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

wage indexation

Joshua Aizenman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

consumer price index; contingency contracts; inflation; monetary policy rules; nominal wage contracts; rational expectation; staggered wage setting; wage flexibility; wage indexation

Article

Wage indexation is a mechanism designed to adjust wages to information that cannot be foreseen when the wage contract is negotiated. A wage contract with indexation clauses will specify the wage base (that is, the money wage applicable in the absence of new information), the indexation formula that will be used to update wages, and how often updating will occur. Most traditional discussion has focused on wage indexation to the price level as a mechanism to stabilize real wages in the presence of inflation. More recently, however, attention has shifted to indexation to a wider set of indicators. These indicators include both richer price information (such as the value added price deflator and the terms of trade) and rules designed to index wages to indicators measuring the level of nominal activity (such as nominal GNP). Concurrently, growing attention has been given to the potential role of wage indexation in affecting the will and ability to reduce inflation.

The economic evaluation of the role and desirability of wage indexation is inherently tied to the assessment of the functioning of the labour market and the role of wage contracts. If labour markets are cleared continuously, as in an auction market (that is, if wages are adjusted continuously to equate the demand and the supply of labour), wage contracts serve no purpose and indexation clauses are either redundant or diminish welfare. On the other hand, the potential role of various wage indexation schemes grows the further we move away from an auction labour market. Consequently, an analysis regarding the role of wage indexation invites a specification of the nature of the deviations from an auction labour market and of the disequilibrium mechanism applied in that market. Indeed, challengers of the usefulness and relevance of wage indexation have remarked on the lack of rigorous understanding of the postulated deviations from auction market behaviour (see Barro, 1977). At the same time, a growing body of research has proceeded on the assumption that the existence of nominal contracts with limited degrees of indexation provides enough evidence to reject an auction labour-market clearing hypothesis

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(see Fischer, 1977b). This assumption has justified studies of the economics of wage indexation in models that lack a rigorous general equilibrium framework, but still provide insights into complicated economic environments. We start with a review of analytical studies on wage indexation, continue with overview of experience with wage indexation in various countries, and close with some interpretative remarks.

Analytical aspects of wage indexation

The usefulness of indexing wages to the price level has been the subject of considerable research, and perceptive comments on the topic can be found in publications going back to Keynes's *General Theory of Employment, Interest and Money* (1936). A renewed interest in the question was generated by Gray (1976; 1978) and Fischer (1977a), who integrated the rational expectation hypothesis with nominal contracts. They considered an economy where nominal contracts preset the contract wage before the realization of stochastic shocks. The rational expectation hypothesis is invoked to determine the contract wage, which is set at a level that is expected to clear the labour market. The contract agreement also specifies the degree of wage indexation to unanticipated inflation. A complete indexation implies real wage rigidity, whereas the absence of indexation entails nominal wage rigidity in which changes in the price level directly affect the real wage. The contract specifies also the determination of employment rule, which is assumed to be demand determined. Consequently, in general employment will deviate from the flexible equilibrium level (that is, from the employment level that will prevail in an economy where the wage is set as to clear the labour market continuously). The optimal degree of indexation is designed so as to minimize the expected squared output deviations from its market clearing level. This can be shown to be equivalent to minimizing the deadweight loss in the labour market for risk-neutral agents (see Aizenman and Frenkel, 1985). The optimal degree of wage indexation is a compromise between two opposing forces: the wish to neutralize the potential output consequences of monetary (nominal) shocks by keeping real wages stable, and the wish to reduce real wages in the presence of adverse real shocks. The first goal is accomplished by complete wage indexation to prices, and the second by partial indexation. Optimal indexation balances between these two forces, implying that greater importance of monetary relative to real shocks will be associated with higher indexation. Such an indexation scheme implies that the real sector is not insulated from monetary variability (see Gray, 1976; Fischer, 1977a). As a result, optimal indexation will tend to stabilize output around its full equilibrium level while it will tend to increase the volatility of prices.

Subsequent research had raised several important questions, for instance why wages are contingent only on prices and not on other relevant information. As Barro (1977) and Karni (1983) have pointed out, optimal contingencies will allow wages to adjust to all relevant information, thereby clearing the labour market continuously and eliminating the output effects of monetary policy. The fact that we find no contracts with rich sets of contingencies suggests, however, that it will be very costly to collect and process all the information needed to write and enforce full contingency contracts (see Fischer, 1977b; Blanchard, 1979). Another related question is the underlying justification of the disequilibrium hypothesis. As demonstrated by Cukierman (1980), the indexation derived by Gray is affected by the disequilibrium hypothesis. It can be shown, however, that this issue becomes inconsequential once we approach a full contingency contract because such a contract will clear the market independently of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
disequilibrium hypothesis (see Aizenman and Frenkel, 1985).

Further developments regarding wage indexation have extended the analysis to open economies. Flood and Marion (1982) showed that optimal indexation is determined by the exchange rate regime, whereas Aizenman and Frenkel (1985) demonstrated that optimal indexation is only one among many potential policies, and that there is a close linkage between wage indexation, monetary policy and exchange rate policies.

A relevant consideration in these discussions is the set of indicators to which the wage is indexed. Most of the above studies derived optimal indexation rules in terms of the underlying structural parameters (like the elasticities of demand in the money and labour market). While these results are informative, their usefulness is limited by the degree of availability of information regarding the underlying structure. In an environment with limited and costly information, indexation rules that use easily available data, without relying on the structural parameters, should have natural advantage. Such rules were studied by Marston and Turnovsky (1985) who investigated the usefulness of wage indexation to the GNP price deflator and to the GNP in the context of energy shocks. Aizenman and Frenkel (1986) pursued related research, showing that if the elasticity of demand for labour exceeds the elasticity of supply, then indexing nominal wages to nominal GNP is preferable to indexing to the value added price index, and this, in turn, is preferable to indexation to the CPI (this ranking is reversed when the elasticity of the supply of labour exceeds the elasticity of demand). Similar results are applicable for ranking the usefulness of targeting monetary policy to the above indicators. Taking another research direction, Fethke and Policano (1984) addressed the usefulness of coordinating the timing of wage negotiations in a multisectoral economy. They concluded that when disturbances are driven primarily by relative shocks (that is, shocks that hit the two sectors differentially) staggered negotiation is optimal.

Once we place wage indexation in its proper perspective as a macroeconomic policy instrument, a natural question arises regarding the linkages between wage indexation and other policies such as taxes and assets indexation (see Friedman, 1974), the risk-sharing effects of indexation (see Azariadis, 1978) and wage renegotiation (see Gray, 1978 and Aizenman, 1984). Further analysis and references regarding these important topics can be found in a useful conference volume (Dornbusch and Simonsen, 1983).

Experience with wage indexation

The experience with indexation of the last decades has been mostly with various degrees of wage indexation to the CPI. The precise indexation policy differs across countries considerably, depending on the centralization of the wage negotiation process and the degree to which wage indexation is viewed as income policy instead of as an instrument to enhance the efficiency of the labour market. For example, in the United States wage indexation is allowed, but there are no guidelines and the details of the indexation schemes are left for the contract negotiation. In Europe, Latin America and Israel labour negotiation tends to be more centralized, and the indexation provisions tend to be dictated by a centralized policymaker. Some countries (for instance, Italy and Brazil) have applied wage indexation as an implicit income policy. This was done by imposing a rigid base wage and a high degree of wage indexation (and in some cases with a cap at high income levels). Such a policy is a poor substitute for direct income policy because it generates distortions in the labour market. These distortive effects increase in periods associated with real shocks, such as changes in input prices and in aggregate demand.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Other countries have attempted to design partial indexation as a device to allow real wage flexibility in the presence of terms of trade shocks (see Brenner and Patinkin, 1977).

While experience with indexation differs across countries, several observations appear to be common to them all. *First*, the degree of indexation to the price level and the frequency of wage adjustment tend to go up with the level and volatility of inflation (see Ehrenberg, Danziger and San, 1983, and Kleiman, 1977). *Second*, a higher indexation rate tends to reduce linkages between excess demand forces in the labour market and wages (see Sachs, 1983). *Third*, limited indexation seems not to be a controversial issue for countries with stable and relatively low inflation rates. For countries with high and volatile inflation the desirability of wage indexation is an important policy issue when attention shifts to curbing that inflation. In various countries in the last decades we have observed the adoption of indexation at low and moderate inflation rates. Once inflation has risen to intolerable levels, however, the policymaker has tended to couple abrupt disinflationary policies with disindexation policies (for example in Iceland in 1983 and in Israel in 1985). This tendency is related to the fact that a typical indexation scheme adjusts wages to lagged inflation, implying that it builds in inertia, thereby a policy of disinflation will tend to raise real wages during the transition, generating unemployment (see Simonsen, 1983, and Fischer, 1984).

Concluding remarks

The role of nominal contracts and the potential role of wage indexation and macro policies are major research topics in macroeconomics. In recent years we have witnessed considerable development in this area, achieved by integrating the rational expectation hypothesis into models where transaction costs prevent continuous auction market clearing. The present state of theoretical research is, however, far from satisfactory. On the one hand, the theoretical papers reviewed above do not offer a framework that will satisfy ‘purists’, although they allow assessment of important policy issues in the presence of realistic contracts. On the other hand, ‘purists’ have not so far been able to explain the existence of nominal contracts of the type observed in various segments of the labour market. Interesting research directions that may provide further clues are frameworks that will recognize and model economic environments where decisions are costly. These costs stem from the observation that data gathering and screening are not free, and that resources are lost in the negotiation process. Such a framework will put a premium on simpler decision rules requiring less frequent negotiation, and nominal wage contracts may be one important example of such rules. The research into nominal contracts reviewed above is, we may hope, a step in that challenging direction.

The experience with wage indexation to prices suggests that greater attention should be given to the design of tractable indexation rules that will generate real wage flexibility in the presence of real shocks while retaining the purchasing power of wages in the presence of nominal disturbances. Such rules should be based upon widely available information. A candidate that deserves further exploration is wage indexation to nominal GNP. Simple-minded rules for indexation to the CPI have several potential disadvantages. They prevent real wage and employment adjustment in the presence of real shocks, thereby causing suboptimal employment. In the presence of nominal shocks and inflation, indexation to prices can generate dynamic inconsistencies – in the short and intermediate run it mitigates the losses associated with unanticipated inflation, but it thereby reduces the will to follow policies that are prudent

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
with regard to inflation, causing higher inflation in the long run. Once the policymaker attempts to disinflate, the indexation scheme may exacerbate the welfare costs associated with the transition to lower inflation. Thus, a policy device that is viewed as useful in the short run can be harmful in the long run.

Consequently, indexation rules are not a substitute for prudent macro-policies. Rules that index wages to nominal income or to the GDP deflator can serve a useful role as part of macro-policies that recognize the need to undergo real adjustment in the presence of real shocks. At the same time, they are deceptive and harmful if they are used as income policy tools in an attempt to maintain the purchasing power of wages in economies exposed to productivity and terms-of-trade shocks.

Bibliography

Aizenman, J. 1984. Optimal wage re-negotiation. *Journal of Monetary Economics* 13, 251–62.

Aizenman, J. and Frenkel, J.A. 1985. Optimal wage indexation, foreign exchange intervention, and monetary policy. *American Economic Review* 75, 402–23.

Aizenman, J. and Frenkel, J.A. 1986. Supply shocks, wage indexation, and monetary accommodation. *Journal of Money, Credit and Banking* 18, 304–22.

Azariadis, C. 1978. Escalation clauses and the allocation of cyclical risks. *Journal of Economic Theory* 18(1), 119–55.

Barro, R.J. 1977. Long-term contracting, sticky prices, and monetary policy. *Journal of Monetary Economics* 3, 305–16.

Blanchard, O.J. 1979. Wage indexation rules and the behavior of the economy. *Journal of Political Economy* 87, 798–815.

Brenner, R. and Patinkin, D. 1977. Indexation in Israel. In *Inflation Theory and Anti-Inflation Policy*, ed. E. Lunberg. London: Macmillan.

Cukierman, A. 1980. The effects of wage indexation on macroeconomic fluctuations. *Journal of Monetary Economics* 6, 147–70.

Dornbusch, R. and Simonsen, M.H., eds. 1983. *Inflation, Debt and Indexation*. Cambridge, MA: MIT Press.

Ehrenberg, R., Danziger, L. and San, G. 1983. Cost of living adjustment clauses in union contracts: a summary of results. *Journal of Labor Economics* 1, 215–45.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fethke, G. and Policano, A. 1984. Wage contingencies, the pattern of negotiation and aggregate implications of alternative contract structure. *Journal of Monetary Economics* 14(2), 151–70.

Fischer, S. 1977a. Wage indexation and macro-economic stability. In *Stabilization of Domestic and International Economy*, ed. K. Brunner and A.H. Meltzer. *Journal of Monetary Economics*, Supplementary 5, 107–47.

Fischer, S. 1977b. Long term contracting, sticky prices, and monetary policy: a comment. *Journal of Monetary Economics* 3, 317–23.

Fischer, S. 1984. Real balances, the exchange rate and indexation: real variables in disinflation. Working Paper No. 1497. Cambridge, MA: NBER.

Flood, R.P. and Marion, N.P. 1982. The transmission of disturbances under alternative exchange-rate regimes with optimal indexing. *Quarterly Journal of Economics* 97, 43–66.

Friedman, M. 1974. *Monetary Correction*. Washington, DC: American Enterprise Institute.

Gray, J.A. 1976. Wage indexation: a macro-economic approach. *Journal of Monetary Economics* 2, 221–35.

Gray, J.A. 1978. On indexation and contract length. *Journal of Political Economy* 86, 1–18.

Karni, E. 1983. On optimal wage indexation. *Journal of Political Economy* 91, 282–92.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kleiman, E. 1977. Monetary correction and indexation: the Brazilian and Israeli experience. *Explorations in Economic Research* 4(1), 141–76.

Marston, R.C. and Turnovsky, J.S. 1985. Imported material prices, wage policy, and macroeconomic stabilization. *Canadian Journal of Economics* 18, 273–84.

Sachs, J. 1983. Real wages and unemployment in the OECD countries. *Brookings Papers on Economic Activity* 1983(1), 255–304.

Simonsen, M.H. 1983. Current theory and the Brazilian experience. In *Inflation, Debt, and Indexation*, ed. R. Dornbusch and M.H. Simonsen. Cambridge, MA: MIT Press.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Aizenman, Joshua. "wage indexation." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000004>
doi:10.1057/9780230226203.1807

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

wage inequality, changes in

Stephen Machin and John Van Reenen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

We examine trends in wage inequality in the United States and other countries since the 1960s. We show that there has been a secular increase in the 90–50 wage differential in the United States and the United Kingdom since the late 1970s. By contrast the 50–10 wage differential rose mainly in the 1980s and flattened or fell in the 1990s and 2000s. We conclude that a version of the skill-biased technical change hypothesis combined with institutional changes (the decline in the minimum wage and trade unions) continues to offer the best explanation for the observed patterns of change.

Keywords

decentralization; elasticity of substitution; Heckscher–Ohlin trade theory; imperfect competition; international trade; labour market institutions; minimum wages; skill-biased technical change; Stolper–Samuelson theorem; technical change; trade unions; unemployment; wage differentials; wage inequality

Article

1 Introduction

Study of the structure of wages has been a preoccupation of economists for a long time and dates back at least as far as Adam Smith. Until the early 1990s economists commented on the remarkable stability of the wage structure in the post-war period. But then many empirical studies (for example, Bound and Johnson, 1992) noticed that wage inequality in America had risen dramatically since the late 1970s. Related empirical research (notably by Goldin and Katz, 1999; 2001) went back further in time uncovering other periods of changing wage structures in American history. Other countries, notably the United Kingdom, also saw a significant increase in wage inequality at about the same time as the recent US changes (Machin, 1996). These observations kick-started what has become a huge empirical and theoretical literature seeking to measure and explain changes in wage inequality (see the survey of Katz and Autor, 1999). Since wages are a major part of people's income and economic well-being, the increase in wage inequality feeds through to income, consumption and poverty rates. So understanding the patterns of wage inequality is important from a normative as well as a positive perspective.
In this article we examine what has happened to the wage distribution since the 1960s, looking principally at the United States, where the bulk of the economic research has focused, but where possible also examining other countries. Section 2 describes the observed changes in the structure of wages (although we fully acknowledge there are some contentious, and as of yet unresolved, issues about the observed patterns of change). Section 3 looks at the main explanations of the observed changes that have emerged from the large body of work in this area. Section 4 offers some conclusions.

2 What has happened to the wage distribution?

2.1 Overall trends in US wage inequality

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To set the scene, Figure 1 plots out the salient features of the US full-time weekly wage distribution from 1963 through to 2003. At least three things stand out from Figure 1. First, educational wage differentials – measured as the gap in pay between college and high school educated workers – have risen consistently since 1979 (after falling somewhat in the 1970s and rising somewhat in the 1960s). The rate of increase was more rapid in the 1980s than after 1992. (This ongoing secular rise in educational wage premia is also seen in the hourly wage series from March outgoing rotation group of the Current Population Survey, CPS; see Lemieux, 2006.) Second, the 90–10 wage differential – defined as the difference in weekly pay for those at the 90th and 10th percentiles of the overall wage distribution – has been rising since 1976 (and maybe even earlier). Third, the 'residual' 90–10 wage differential – the difference between those at the 90th and 10th percentiles of the overall wage distribution after controlling for education, experience and gender – has risen consistently since 1967, especially after the mid-1970s (see Juhn, Murphy and Pierce, 1993). This increase in 'within group' wage inequality has also generated much excitement and interest from theorists, but is particularly hard to interpret in the light of compositional changes (Lemieux, 2006).

Figure 1

Changes in US wage inequality, 1963–2003. *Note*: based on full-time weekly earnings for all workers in the March *Current Population Survey*. *Source*: Autor, Katz and Kearney (2005).

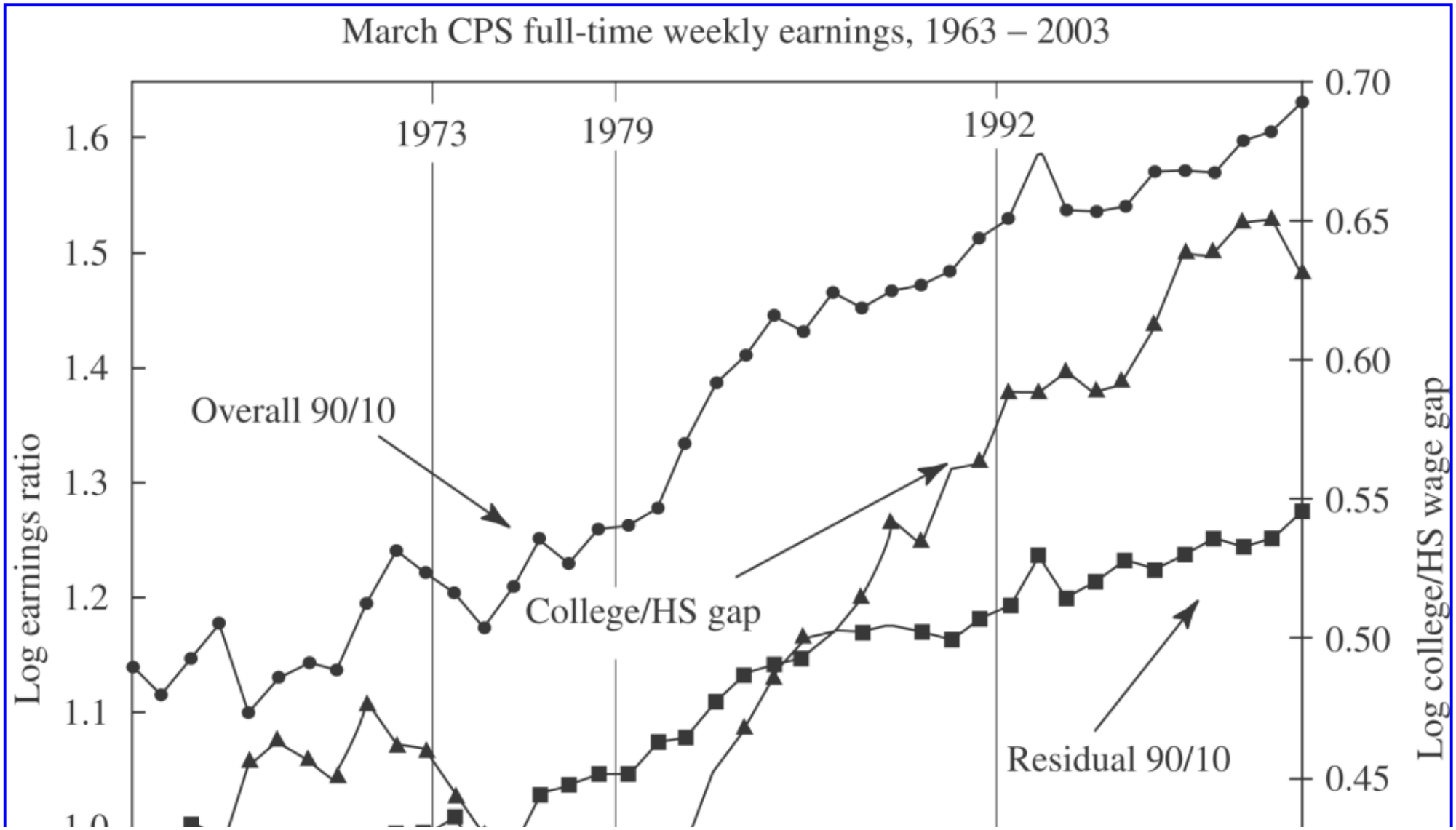

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

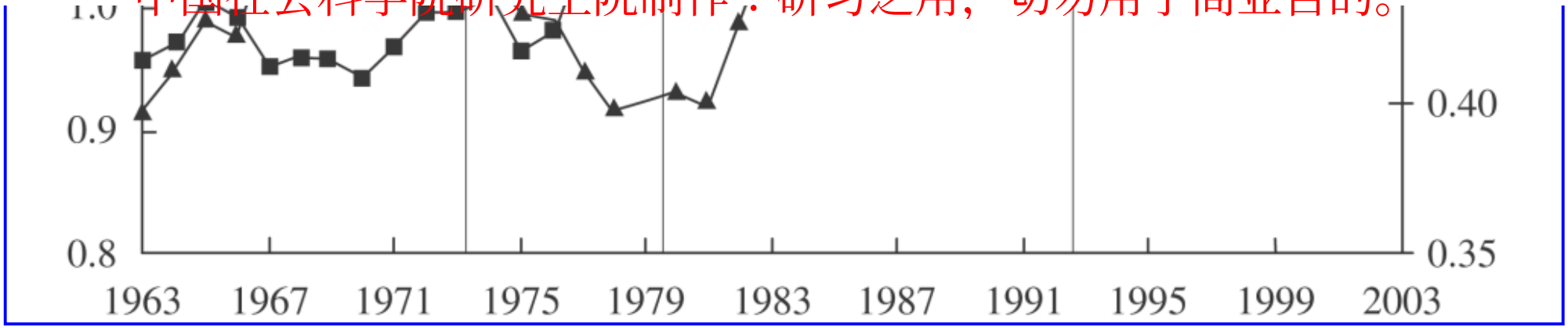

Even though different data-sets show some differences and there are some variations in inequality measures across data sources, the overall picture is one of a dramatic increase in American wage inequality since 1979.

2.2 Comparing the United Kingdom with the United States

The United Kingdom is another country where wage inequality has risen dramatically. Comparison of the United States and United Kingdom is useful to pin down certain issues to do with the rise in wage inequality. One important point is that since 1980 there are marked decadal differences in the opening up of the wage structure. Analysis of US and UK micro-data uncovers a clear picture for the 1980s in both countries: wage growth was more pronounced at higher points of the distribution, and faster wage growth higher up the distribution is almost monotonic in both counties, leading to large increases in wage inequality. An important difference, however, is that in the United Kingdom there was positive wage growth throughout the distribution whereas in the United States workers in the bottom quartile actually experienced zero or negative wage growth.

The picture becomes more complex post-1990. In both countries the 90–50 continues to diverge (‘upper tail inequality’) whereas the 50–10 (‘lower tail inequality’) in the United States actually shrinks, indicating some wage compression. In the United Kingdom the 50–10 is stable (increasing a bit in the 1990s and shrinking a bit in the 2000s). Overall then, the increase in wage inequality has been stronger in the upper tail than the lower tail taking the period as a whole, and has been more pronounced in the 1980s than post-1990.

A marked and important similarity between the two countries is the continuous and rapid growth of wages at the very top of the distribution. In the United Kingdom, wage growth at the 95th percentile (and above) is greater than at other percentiles of the wage distribution in the 1980s, 1990s and 2000s. This is also true for the United States (except for the 10th percentile in the 1990s). So within the picture of overall rising inequality the very rich have done particularly well.

The other key feature of the changing wage distributions in the United Kingdom and the United States (and elsewhere) has been the polarization of work into ‘good jobs’ and ‘bad jobs’ (defined as high-wage and low-wage jobs). While there has been significant growth in well-paid ‘good jobs’ at the upper tail of the distribution (like lawyers, senior managers and consultants) there has been an increase in low-paid ‘bad jobs’ in the lower tail of the distribution (like cleaners, hairdressers, shop assistants and burger flippers). In the 1990s especially it seems that the middle of the distribution seemed to do somewhat worse than those at the top or bottom. These findings have been reported on in the United States (Autor, Katz and Kearney, 2006), United Kingdom (Goos and Manning, 2007) and Germany (Spitz-Oener, 2006).

2.3 The experience of other countries

There is less systematic evidence for the evolution of the wage distribution outside of the United States and the United Kingdom, especially for more recent years. Table 1 uses OECD data to show 90–10 male wage ratio for a range of countries between 1980 and 2000. Broadly speaking, the 1980s rise in inequality was seen only in the United Kingdom and the United States and in specific countries where particular episodes to move to a much more market-oriented economy occurred (notably New Zealand). Elsewhere wage inequality did not alter much. The 1990s is a little different, with evidence of widening wage structures starting to occur in places previously characterized by stable wage structures – Germany is a very good example of this. Moreover, as we discuss below, the Continental European countries did experience a larger increase in unemployment, which may be due to the same underlying forces that have pushed up wage inequality in Britain and America.

Male 90–10 wage ratios across countries, 1980–2000

	Male 90–10 wage ratios		
	1980	1990	2000

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Australia	2.73	2.71	3.16
Finland	2.44	2.57	2.47[f]
France	3.38	3.46	3.28[e]
Germany	2.53[b]	2.44	2.86[e]
Italy	2.09[b]	2.38	2.44[c]
Japan	2.60	2.84	2.74[f]
Netherlands	2.32[a]	2.48	2.83[f]
New Zealand	2.72	3.08	3.55[d]
Sweden	2.11	2.07	2.35[e]
UK	2.63[b]	3.24	3.40
US	3.58	4.41	4.76

Note: Data is from different years where indicated by the following superscripts: a – 1985; b – 1986; c – 1996; d – 1997; e – 1998; f – 1999.

Source: OECD data website http://www.oecd.org.

3 Explanations of changes in wage inequality

A natural place to begin to analyse the observed changes in the wage structure is to consider a model of changes in supply and demand. We then need to incorporate institutional features (such as minimum wages and trade unions) into the model.

3.1 Sources of skill premia: supply and demand

Rising wage inequality has been accompanied by an increase in the relative demand for skilled or educated workers. This is evident since, despite the increase in the relative supply of more skilled workers in many countries, their relative wage has also gone up, suggesting that relative demand for skilled workers has been rising faster than relative supply. In Table 2, for example, the proportion of graduates grew from 20.8 per cent of the population in 1980 to 34.2 per cent in 2004 in the United States. (In the United States the graduate measure is having a bachelor's degree or higher – that is, excluding people with some college who do not get a degree.) The equivalent figures from the United Kingdom were even more dramatic – the growth in graduates was from 5 per cent to 21 per cent over the same time period. However, at the same time the relative wages of graduates compared with those of non-graduates increased. In a competitive model of the labour market with skilled and unskilled workers, these facts can be reconciled by an increase in the relative demand for skilled workers.

Aggregate trends in graduate/non-graduate employment and wages, UK and USA, 1980–2004

	UK		USA	
	% graduate share of employment	Relative weekly wage (full-time)	% graduate share of employment	Relative weekly wage (full-time)
1980	5.0	1.48	20.8	1.41
1985	9.8	1.50	24.2	1.53
1990	10.2	1.60	25.7	1.60
1995	14.0	1.60	31.8	1.65
2000	17.2	1.64	31.8	1.69
2004	21.0	1.64	34.2	1.66
Changes:				

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1980–2004	16.0	.16	13.4	.25
1980–90	5.2	.12	4.9	.19
1990–2000	7.0	.04	6.1	.09
2000–4	3.8	.00	2.4	–.02

Notes: Sample is all people aged 18–64 in work and earning, except for relative wages, which are defined for full-time workers. The relative wage ratios are derived from coefficient estimates on a graduate dummy variable in semi-log earnings equations controlling for age, age squared and gender (they are the exponent of the coefficient on the graduate dummy).

Sources: UK: derived from General Household Survey (GHS) and Labour Force Survey (LFS); updated from Machin and Vignoles (2005). US: derived from Current Population Survey data.

A simple way to formalize this, following Katz and Murphy (1992), is in the context of a constant elasticity of substitution (CES) production function with two labour inputs:

$$Q_t = [\alpha_t (a_t N_s)_t^{\rho} + (1 - \alpha_t)(b_t N_u)_t^{\rho}]^{1/\rho} \quad (1)$$

In eq. (1) aggregate output is Q and is produced with college educated-equivalent skilled labour (N_s) and high school-educated equivalent unskilled labour (N_u) in period t. The parameters a and b represent skilled and unskilled augmenting technical change, α indexes the share of work activities of skilled labour and ρ is a parameter that determines the elasticity of substitution between skilled and unskilled labour ($\sigma = \frac{1}{1-\rho}$). Skill-biased technological changes involve increases in a/b or α.

Assuming college and high school equivalents are paid their marginal product we can use eq. (1) to solve for the ratio of marginal products of the two types of labour, W_s/W_u, and relative supplies of labour, N_s/N_u, in year t as:

$$\ln(W_s / W_u)_t = (1/\sigma)[D_t - \ln(N_s / N_u)_t] \quad (2)$$

where

$$D_t = \sigma[\ln(\alpha_t / (1 - \alpha_t)) + \rho \ln(a / b)_t] \quad (3)$$

is a relative demand index of shifts favouring college equivalents and is measured in log quantity units. The impact of changes in relative skill supplies (N_s/N_u) depends on the elasticity of substitution, σ. The larger this parameter is, the bigger will be the effects of supply changes on relative wages. Equation (3) shows that changes in D can arise from (disembodied) skill-biased technical change, non-neutral changes in relative prices or quantities of non-labour inputs and shifts in product demand.

Katz and Murphy (1992) implemented an empirical version of eq. (2), replacing D with a linear time trend (‘*trend*’) for US data between 1963 and 1987. They estimate:

$$\ln(W_s / W_u)_t = \gamma_0 + \gamma_1 trend + \gamma_2 \ln(N_s / N_u)_t + v_t \quad (4)$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

finding $\hat{\gamma}_2$ to be significantly negative (equal to –0.709), implying an elasticity of substitution of about 1.4 ($\sigma = -1/\hat{\gamma}_2 = 1.41$), with a significant trend increase in the college premium of 3.3 per cent per annum ($\hat{\gamma}_1 = .033$). In the literature that has followed the estimates of the elasticity of substitution are typically in the 1.4 to 1.6 range (see, for example, the study of Autor, Katz and Kearney, 2005). The main point to take away from these estimates is that there appears to be a systematic demand shift towards more skilled workers throughout the four last decades of the 20th century.

This is not to suggest that supply side changes are unimportant. Deviations of relative skill supplies from the trend are negatively associated with deviations of the relative wage from trend as suggested by $\gamma_2<0$. The *slowdown* of the growth of education in more recent cohorts is certainly one factor accounting for the increase in inequality as shown by Card and Lemieux (2001). But the most important factor over the longer run in accounting for the growth in educational wage differentials appears to be the trend demand shift towards the more skilled. The critical question then becomes: what could account for this change?

3.2 The cause of relative demand shifts: technology or trade?

To date, the two main explanations for the demand shift towards the more skilled are skill-biased technological change (SBTC) and increased international trade. We examine each of these in turn.

3.2.1 Skill-biased technological change

Equation (3) above directly relates the change in the skill premia to SBTC. The idea is that new technologies such as information and communication technologies (ICTs) are complementary with the skills of more-educated workers. More-educated workers may find it easier to cope with the uncertainty surrounding new technologies in general, or may have a particular advantage in using ICT effectively. Rapid falls in the quality-adjusted prices of ICT or a more rapid investment in new technologies (for example, from higher R&D intensities) could therefore have shifted demand towards more-skilled workers.

There is now abundant empirical evidence that suggests that SBTC *is* an important and international phenomenon (for example, see the survey in Bond and Van Reenen, 2007). A typical analysis estimates the following cost share equation (usually for industries or workplaces):

$$\Delta SHARE = \beta_1 TECH + \beta_2 \Delta \ln(K/Y) + \beta_3 \ln(W_s/W_u) + e \quad (5)$$

where *SHARE* is the wage bill share of skilled workers, *TECH* is a measure of technical change, *K* is the capital stock, *Y* is value added, W_s/W_u is relative wages, Δ the difference operator and *e* an error term. This relationship can be derived from the stochastic form of a translog short-run variable cost function with labour as the two variables and physical capital and technological capital as the two fixed factors (for example, Berman, Bound and Griliches, 1994). The test of skill-biased technical change is whether $\beta_1>0$, and the overwhelming preponderance of econometric evidence supports this finding.

An example of the genre is Machin and Van Reenen (1998), who examine this relationship using manufacturing data across many industries in seven OECD countries (the United States, the United Kingdom, France, Japan, Germany, Denmark and Sweden) in the 1970s and 1980s. In all of the countries examined they found that demand was shifting more quickly towards skilled workers in the more technologically advanced industries (that is, $\beta_1>0$ in eq (5)). This was robust to using either occupation or education as a measure of skills, using either R&D intensity or computer use as a measure of technology, and instrumenting own R&D with frontier (US) R&D. In most countries they also found evidence of capital–skill complementarity ($\beta_2>0$). Estimating versions of eq. (5) in other countries, in non-manufacturing sectors (for example, Autor, Katz and Krueger, 1998) and on more disaggregated plant-level data (for example, Doms, Dunne and Troske, 1997) also appears to uncover evidence of SBTC.

There are several other sources of evidence on SBTC. Berman, Bound and Machin (1998) report evidence of faster skill demand shifts occurring in the same sorts of industries in different countries, and one may view this as informing the SBTC hypothesis (to the extent that similar industries in different countries utilize similar technologies). A less-used alternative to test for SBTC is to regress the adoption of technologies on skill prices (that is, when skilled workers' wages rise relative to those of unskilled workers this should

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
depress the incentive to adopt new technologies) or skilled labour supply – some evidence for this method is in Caroli and Van Reenen (2001) and Doms and Lewis (2006), and also supports SBTC. A third method is to directly estimate the production function or the cost function underlying the factor demand eq. (5). This has also tended to uncover evidence of skill–technology complementarity (for example, Bresnahan, Brynjolfsson and Hitt, 2002). Finally, some authors have directly regressed individual wages on computer use or controlling for other factors (for example, Krueger, 1993). In our view, this is a rather unsatisfactory test of SBTC, however, as computers are likely to be allocated to more productive workers, as has been found by several studies (Chennells and Van Reenen, 1997; DiNardo and Pischke, 1997).

Although we have stated the SBTC hypothesis in quite a blunt fashion, the influence of technical change almost certainly acts in a more subtle ways to affect outcomes as detailed case studies suggest (Blanchard, 2004). For example, some econometric studies suggest that technical change operates through organizational changes (for example, through decentralizing or delayering hierarchies) that are typically associated with increased demand for skilled workers (Caroli and Van Reenen, 2001; Bresnahan, Brynjolfsson and Hitt, 2002). Moreover, computerization does not simply involve increasing all skill demand, but it substitutes for different types of tasks. Autor, Levy and Murnane (2003) offer a more nuanced version of the SBTC hypothesis, arguing that computerization reduces the demand for routine tasks (for manual and non-manual workers) but results in an increase in demand for analytic or non-routine skills. Thus, routine non-manual tasks (for example, clerical work) may be replaced by computers, whilst some non-routine tasks done by manual workers (like cleaning) are largely unaffected by IT. The evidence on polarization of work referred to above where the ‘middle’ of the wage distribution has suffered at the expense of the bottom as well as the top is in line with this. Building on upon these empirical observations, Autor, Katz and Kearney (2006) develop a model where IT replaces routine tasks to rationalize the experience of the 1990s when polarization of jobs occurred in the United States.

Overall then, there is strong support for the importance of SBTC. Some critics (most strongly expressed in Card and DiNardo, 2002) argue that SBTC cannot be the reason for increased inequality because technical change is continuous whereas the change in wage inequality is episodic. Regardless of whether one agrees with the characterization of technical change, this misses the point that SBTC is meant to account for the longer-run pressure to increase skill demand (the *D* in eq. (2)) and not necessarily the ‘twist’ in the wage structure in the 1980s. Similarly, the fact that inequality growth slowed down post-1995 whereas productivity growth accelerated does not disprove the SBTC argument, as the speed of technical change is not the same as the bias of technical change.

3.2.2 Increased international trade

At first glance, the simple Heckscher–Ohlin model of trade offers a seemingly cogent explanation of why unskilled workers have faired badly in recent decades. Less-developed countries such as China and India have become integrated into the global economy as trade barriers and transportation and communication costs have fallen. Unskilled workers in the OECD counties now have to compete not only with workers at home but also with a large number of workers overseas. The influx of cheap goods produced with low-skill labour puts downward pressure on the wages and employment opportunities of unskilled workers in the West, and is responsible for the observed shifts in relative labour demand.

To model this we explicitly consider two regions: ‘North,’ which is skill-abundant and ‘South’ which is unskilled-abundant. There are four industries: tradable high-skill intensive, tradable low-skill intensive, non-tradable high-skill intensive and non-tradable low-skill intensive. The Stolper–Samuelson theorem establishes that relative wages in each country will depend on relative output prices of the tradable industries: the higher the relative price of the skill-intensive good, the higher the relative wage of the skilled workers. What happens when a small open economy in the North moves from autarky to free trade? The removal of trade barriers increases the relative price of the skill-intensive good and this means the skill premium rises in the North.

Although this model is coherent, it also offers several other predictions which turn out to be at odds with the data (see Desjonqueres, Machin and Van Reenen, 1999, for extensive discussion of these predictions). First, the increasing specialization of the North in skill-intensive goods under free trade means that employment should shift *between* industries to skill-intensive industries. But because relative skill prices have risen we should expect to see that employment *within* industries shifts towards (the cheaper) unskilled workers. Decompositions of the increase in the aggregate employment share of skilled workers, however, almost all show that within industries there has been a strong shift towards skilled workers. This might be because the level of aggregation of industries is too high, but more disaggregated industries and even firm-level studies suggest that a sizable proportion is ‘within’. Even more convincingly, Desjonqueres, Machin and Van Reenen (1999) show that non-traded sectors – such as hotels and wholesale outlets – also show a shift towards skilled workers (and an increase in the educational wage premium). This pattern of within-industry shifts is consistent with general SBTC, but inconsistent with the basic trade theory. Second, we should observe that relative prices of the unskilled-intensive sectors should fall rapidly in the North. There is some evidence for this in the United States but there is no significant relationship for any other country (at least until the mid-1990s). Even in the United States the evidence from Krueger (1996) suggests that this relationship was only apparent after 1989, when wage inequality grew slowly. Finally, naive regressions that include import penetration and other trade variables in eq. (5) generally find no role for these trade variables (for example, Machin and Van Reenen, 1998). This does not take into account the general equilibrium effects underlying the Heckscher–Ohlin model, of course.

Overall there is little support for the trade-based explanation of demand shifts. There are two caveats to this conclusion. First, most of these studies were based on data prior to the early 1990s when China started to become more of a major exporter. Second, trade might induce some of the skill biased technological change discussed in the previous section as suggested by Acemoglu (2002).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

3.3 Labour market institutions

Research trying to reconcile cross-country differences in change in wage inequality has emphasized the role of labour market institutions that affect wages differently in different places. There are several features of this work, ranging from studies that look in detail across countries to those that focus on the role played by particular labour market institutions like minimum wages or trade unions.

3.3.1 Cross-country evidence

As discussed in Section 2, there has been considerable heterogeneity in the evolution of relative wages across OECD countries since the 1970s. The rise in inequality was much stronger in the Anglo–Saxon countries (for example, the United States and the United Kingdom) than elsewhere (for example, France, Germany and Japan). Although the technology and/or trade shocks discussed in the previous subsections should be global events, the Continental European and Japanese economies have experienced a much greater increase in unemployment than the United States since the late 1970s. One view is that European unemployment and American inequality are ‘two sides of the same coin’ – institutional rigidities (and perhaps generous welfare benefits) placed a floor under the wages of unskilled workers in Continental Europe, resulting in increased unemployment rather than greater inequality. (There is a also a new, growing body of work arguing that tastes and social norms are important for explaining cross-country patterns of change; see, amongst others, Bénabou and Tirole, 2006.)

This is probably too crude. Nickell and Bell (1995) have shown that relative unemployment rates between skilled and unskilled workers did not rise by as much as would be expected in this simple model. Similarly, the cross-country correlation between the growth in unemployment and earnings inequality is not very strong (for example, Burniaux, Padrini and Brandt, 2006). Finally, European countries may have been better at keeping up the growth of supply of the quantity and quality of skills than in the United States and United Kingdom (although Table 2 shows that skill expansion in the United Kingdom was very rapid).

At the very least, the fact that wage inequality has not risen in the countries where minimum wages and/or union power remained strong suggests that institutions do have an important role to play.

3.3.2 Minimum wages

There is much evidence that minimum wages compress wage differentials (DiNardo, Fortin and Lemieux, 1996). In the United States the real value of the Federal minimum wage fell significantly during the 1980s, and some authors argue that this can account for all of the change in wage inequality (for example, Lee, 1999). By the same token the uprating of the minimum wage in the 1990s helps explain the slowdown in wage inequality. As Card and DiNardo (2002) emphasize, the time series pattern is very strong – see Figure 2.

Figure 2

The time series relationship between the US federal minimum wage and wage inequality. *Source*: Autor, Katz and Kearney (2005).

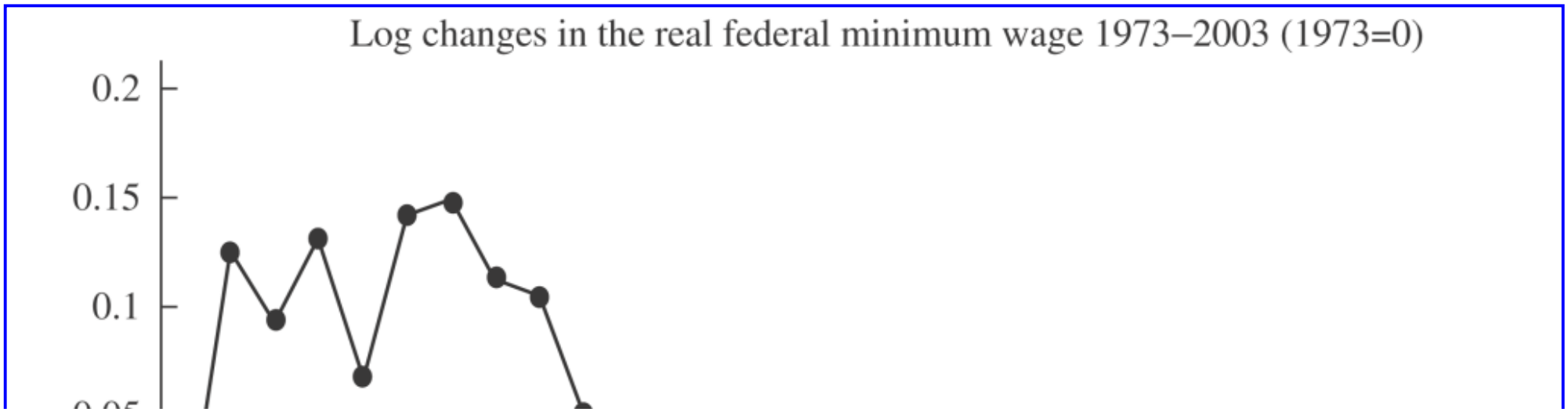

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

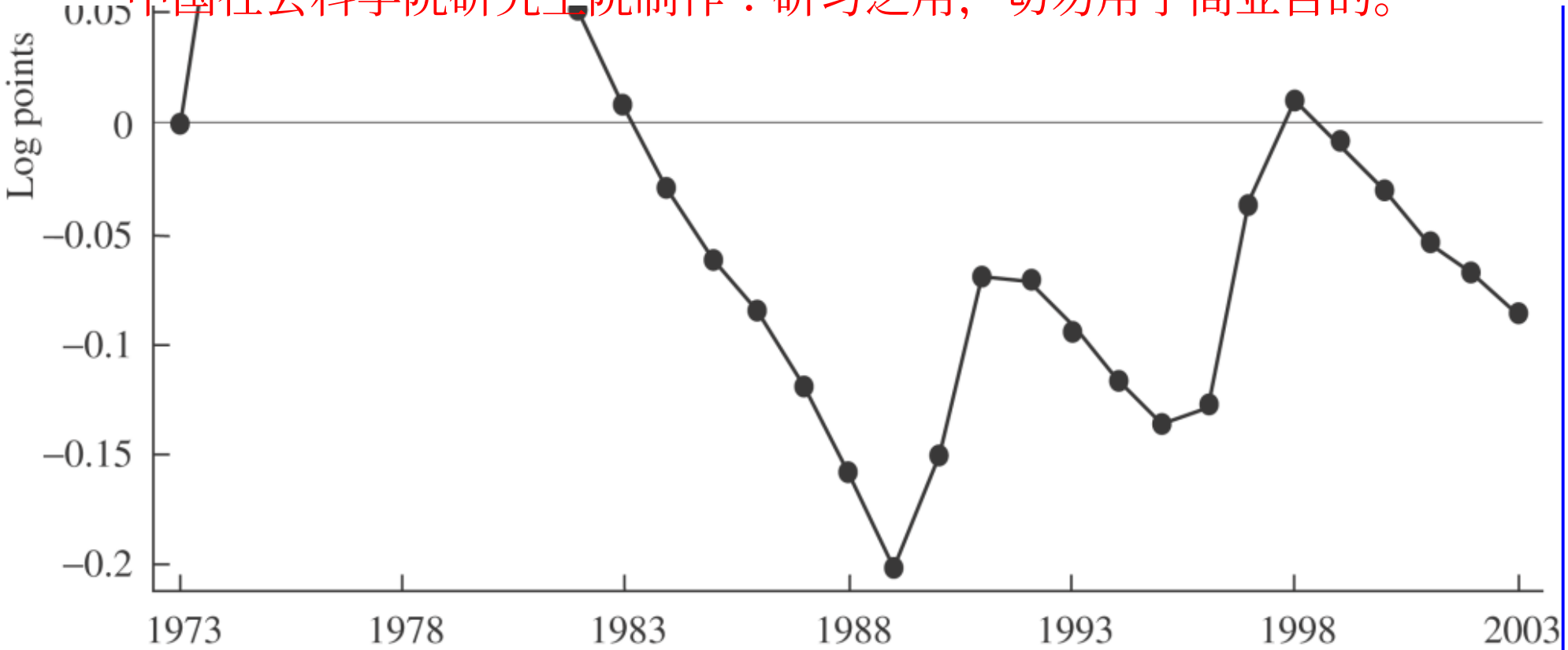

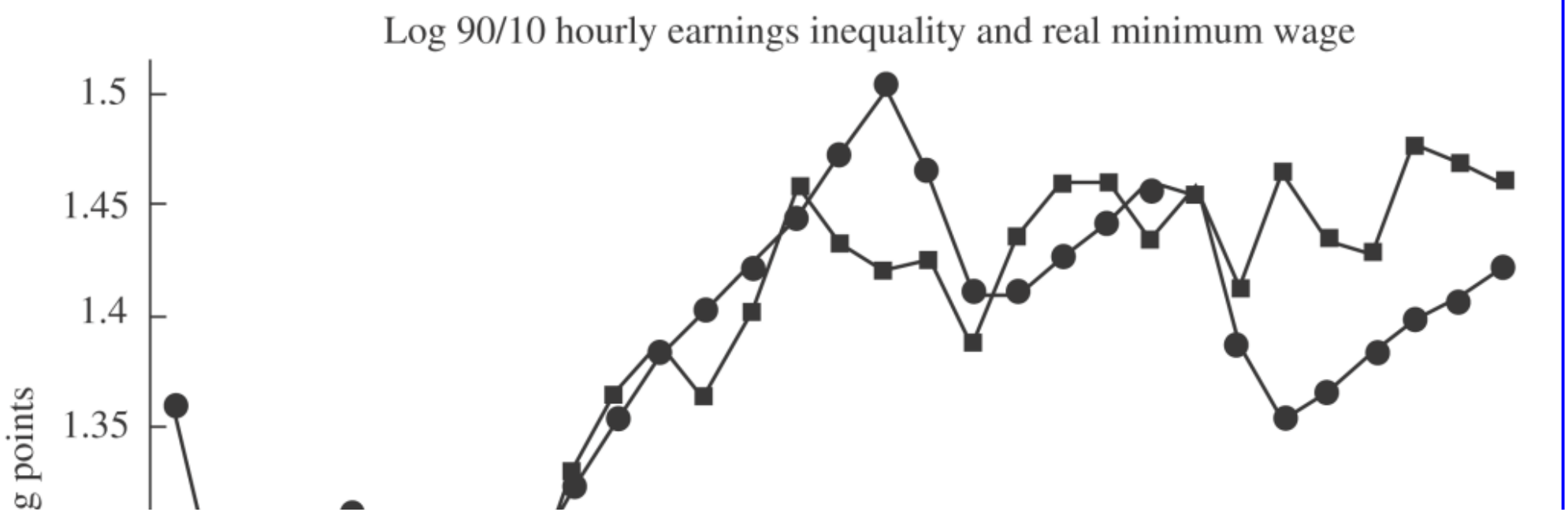

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

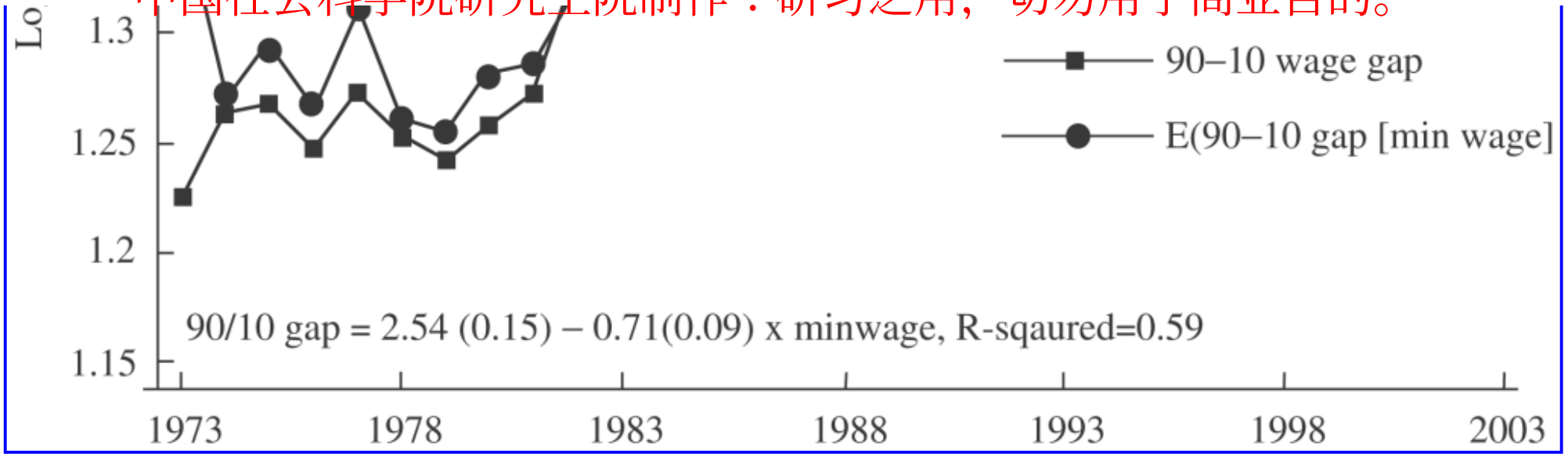

A problem with the ‘purely institutional’ argument, however, is that it seems highly unlikely that the minimum wage can explain what is happening in the top half of the wage distribution. Analysis of the minimum wage suggests that the impact on workers above median wages is close to zero. Nevertheless, the most striking finding of the analysis in Section 2 was that there appeared to be a secular increase in the 90–50 wage ratio since the late 1970s in the United States (and the United Kingdom). It is hard to reconcile these facts with the minimum wage-explains-all story. Similarly, when Autor, Katz and Kearney (2005) add the minimum wage to eq. (4), although it has the expected negative sign it does little to reduce the long-run unexplained relative demand shift towards higher education wage differentials.

Where the institutional story does better is in accounting for the dramatic increase in residual wage inequality in the bottom half of the wage distribution in the 1980s. This residual wage change was more episodic, and most of the change is plausibly accounted for by the minimum wage (and compositional effects – see below).

Another problem with the pure minimum wage explanation is that wage floors changed much less in other countries where wage inequality also rose. For example in the United Kingdom, the minimum wage system that operated at the time when wage inequality rose (the ‘Wage Councils’) only covered a relatively small proportion of the workforce (around 12 per cent at the time of abolition in 1993). Furthermore, during the 1993–9 time period when all non-agricultural minimum wages were abolished in the United Kingdom, wage inequality at the lower end actually started to stabilize (Dickens, Machin and Manning, 1999; Machin and Manning, 1994).

3.3.3 Trade unions and imperfect competition

As with minimum wages there is robust evidence that unions act to compress wage differentials (for example, Freeman, 1980; Card, 1996). Since unions have declined in the United States and the United Kingdom, this may be another institutional mechanism putting upwards pressure on wage inequality. Unionization rates fell from 25 per cent to 15 per cent between 1979 and 1998 in the United States, and from 53 per cent to 31 per cent in the United Kingdom over the same period. Gosling and Lemieux (2004) argue that union decline can account for over a third of the increase in male wage inequality in both countries over the 1983–98 period.

As with the minimum wage explanation, it is rather difficult to evaluate these statistical decompositions as they are not based on an underlying economic model. But it does seem rather implausible that unions could be the major explanation in the United States for the ongoing increase in the 90–50 ratio since (*a*) they comprise such a small part of the workforce and (*b*) their membership is mainly drawn from the bottom half of the wage distribution.

An alternative set of theories has emerged that emphasizes rents derived from imperfect competition (albeit from a different source from unions). This approach has frictions in the labour market that generate heterogeneous wages even for identical workers. Some more productive or technologically advanced firms may share quasi-rents to workers who are matched with them (for example, Van Reenen, 1996). If the dispersion of these wage premia has increased over time, this could lie behind the increased wage inequality. For example, in Caselli (1999) firms experiment with the uncertain new technology, and some of those that are successful obtain higher productivity, resulting in higher wages for the workers with whom they are matched. To date, there is little hard empirical evidence on these theories, although Faggio, Van Reenen and Salvanes (2006) offer some evidence that firm productivity heterogeneity has increased and this is linked to firm wage inequality as Caselli's model would suggest.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

4 Conclusions

There has been a dramatic increase in wage inequality since the late 1970s in the United States, the United Kingdom and other anglophone countries. A significant part of this is due to the growth of wage differentials between educational groups. We have argued that the fundamental reason for this is a long-run growth in the relative demand for skills driven by technology change (rather than trade). Changes in skill supply and institutional changes have affected the timing of how skill-biased technical change impacts upon the wage structure. The increase in inequality in the United States and the United Kingdom slowed down after 1990, but has continued to grow in the upper tail of the wage distribution, and wage inequality has started to rise in places previously characterized by stable wage structures (like Germany), indicating that explaining changing patterns of wage inequality remains high on the research agenda of empirical economists.

See Also

- inequality (international evidence)
- minimum wages
- skill-biased technical change
- technology
- trade and poverty

The Economic and Social Research Council has given financial support for this research through the Centre for Economic Performance. The usual disclaimer applies.

Bibliography

Acemoglu, D. 2002. Technical change, inequality and the labour market. *Journal of Economic Literature* 40, 7–72.

Autor, D., Katz, L. and Kearney, M. 2005. Trends in U.S. wage inequality: re-assessing the revisionists. Working Paper No. 11627. Cambridge, MA: NBER.

Autor, D., Katz, L. and Kearney, M. 2006. The polarization of the U.S. labor market. *American Economic Review* 96, 189–94.

Autor, D., Katz, L. and Krueger, A. 1998. Computing inequality: have computers changed the labor market? *Quarterly Journal of Economics* 113, 1169–214.

Autor, D., Levy, F. and Murnane, R. 2003. The skill content of recent technological change: an empirical investigation. *Quarterly Journal of Economics* 118, 1279–333.

Bénabou, R. and Tirole, J. 2006. Belief in a just world and redistributive politics. *Quarterly Journal of Economics* 121, 699–746.

Berman, E., Bound, J. and Griliches, Z. 1994. Changes in the demand for skilled labor within U.S. manufacturing industries: evidence from the Annual Survey of Manufacturing. *Quarterly Journal of Economics* 109, 367–98.

Berman, E., Bound, J. and Machin, S. 1998. Implications of skill-biased technological change: international evidence. *Quarterly Journal of Economics* 113, 1245–80.

Blanchard, O. 2004. The economic future of Europe. *Journal of Economic Perspectives* 18(4), 3–26.

Bond, S. and Van Reenen, J. 2007. Micro-econometric models of investment and employment. In *Handbook of Econometrics*, vol. 6, ed. J. Heckman and E. Leamer. Amsterdam: North Holland.

Bound, J. and Johnson, G. 1992. Changes in the structure of wages in the 1980s: an evaluation of alternative explanations. *American Economic Review* 83, 371–92.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bresnahan, T., Brynjolfsson, E. and Hitt, L. 2002. Information technology, workplace organization and the demand for skilled labour: firm-level evidence. *Quarterly Journal of Economics* 117, 339–76.

Burniaux, J., Padrini, F. and Brandt, N. 2006. Labour market performance, income inequality and poverty in OECD countries. Working Paper No. 500, Economics Department, OECD.

Card, D. 1996. The effects of unions on the structure of wages. *Econometrica* 64, 957–79.

Card, D. and DiNardo, J. 2002. Skill-biased technological change and rising wage inequality: some problems and puzzles. *Journal of Labor Economics* 20, 733–83.

Card, D. and Lemieux, T. 2001. Can falling supply explain the rising return to college for younger men? *Quarterly Journal of Economics* 116, 705–46.

Caroli, E. and Van Reenen, J. 2001. Skill-biased organizational change? Evidence from a panel of British and French establishments. *Quarterly Journal of Economics* 116, 1449–92.

Caselli, F. 1999. Technological revolutions. *American Economic Review* 89, 78–102.

Chennells, L. and Van Reenen, J. 1997. Technical change and earnings in British establishments. *Economica* 64, 587–604.

Desjonqueres, T., Machin, S. and Van Reenen, J. 1999. Another nail in the coffin? Or can the trade based explanation of changing skill structures be resurrected? *Scandinavian Journal of Economics* 101, 533–54.

Dickens, R., Machin, S. and Manning, A. 1999. The effects of minimum wages on employment: theory and evidence from Britain. *Journal of Labor Economics* 17, 1–22.

DiNardo, J., Fortin, N. and Lemieux, T. 1996. Labor market institutions, and the distribution of wages, 1973–1992: a semi-parametric approach. *Econometrica* 64, 1001–44.

DiNardo, J. and Pischke, S. 1997. The returns to computer use revisited: did pencils change the wage structure too? *Quarterly Journal of Economics* 112, 291–303.

Doms, M. and Lewis, E. 2006. Labor supply and personal computer adoption, Research Paper No. 06–10, Federal Reserve Bank of Philadelphia.

Doms, M., Dunne, T. and Troske, K. 1997. Workers, wages and technology. *Quarterly Journal of Economics* 112, 253–89.

Faggio, G., Van Reenen, J. and Salvanes, K. 2006. Increasing inequality of firm productivity. Mimeo, Centre for Economic Performance, London School of Economics.

Freeman, R. 1980. Unionism and the dispersion of wages. *Industrial and Labor Relations Review* 34, 3–23.

Goldin, C. and Katz, L. 1999. The returns to skill in the United States across the twentieth century. Working Paper No. 7126. Cambridge, MA: NBER.

Goldin, C. and Katz, L. 2001. Decreasing (and then increasing) inequality in America: a tale of two half centuries. In *The Causes and Consequences of Increasing Inequality*, ed. F. Welch. Chicago: University of Chicago Press.

Goos, M. and Manning, A. 2007. Lousy and lovely jobs: the rising polarization of work in Britain. *Review of Economics and Statistics* 89, 118–33.

Gosling, A. and Lemieux, T. 2004. Labor market reforms and changes in wage inequality. In *Seeking a Premier League Economy*, ed. D. Card, R. Blundell and R. Freeman. Chicago:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Chicago University Press.

Juhn, C., Murphy, K. and Pierce, B. 1993. Wage inequality and the rise in returns to skill. *Journal of Political Economy* 101, 410–42.

Katz, L. and Autor, D. 1999. Changes in the wage structure and earnings inequality. In *Handbook of Labor Economics*, vol. 3, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Katz, L. and Murphy, K. 1992. Changes in relative wages, 1963–87: supply and demand factors. *Quarterly Journal of Economics* 107, 35–78.

Krueger, A. 1993. How computers have changed the wage structure: evidence from microdata, 1984–1989. *Quarterly Journal of Economics* 108, 33–60.

Krueger, A. 1996. Skill intensity and industrial price growth. *Industrial Relations Research Association Proceedings* 48, 11–8.

Lee, D. 1999. Wage inequality in the United States during the 1980s: rising inequality or falling minimum wages. *Quarterly Journal of Economics* 114, 997–1023.

Lemieux, T. 2006. Increasing residual wage inequality: composition effects, noisy data, or rising demand for skill? *American Economic Review* 96, 461–98.

Machin, S. 1996. Wage inequality in the UK. *Oxford Review of Economic Policy* 7, 49–62.

Machin, S. and Manning, A. 1994. Minimum wages, wage dispersion and employment: evidence from the U.K. wages councils. *Industrial and Labor Relations Review* 47, 319–29.

Machin, S. and Van Reenen, J. 1998. Technology and changes in skill structure: evidence from seven OECD countries. *Quarterly Journal of Economics* 113, 1215–44.

Machin, S. and Vignoles, A. 2005. *What's the Good of Education? The Economics of Education in the UK*. Princeton: Princeton University Press.

Nickell, S. and Bell, B. 1995. The collapse in demand for the unskilled and unemployment across the OECD. *Oxford Review of Economic Policy* 11, 40–62.

Spitz-Oener, A. 2006. Technical change, job tasks and rising educational demands: looking outside the wage structure. *Journal of Labor Economics* 24, 235–70.

Van Reenen, J. 1996. The creation and capture of economic rents: wages and innovation in a panel of UK companies. *Quarterly Journal of Economics* 111, 195–226.

How to cite this article

Machin, Stephen and John Van Reenen. "wage inequality, changes in." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_C000551> doi:10.1057/9780230226203.1808

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

wages fund

John Vint
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The wages fund doctrine was an important element in the classical analysis of the labour market – elements of the wages fund doctrine are to be found in the *Wealth of Nations* (1776) – and articles attempting to defend it were still being produced over a hundred years later. Its longevity was due to its success in generating a wide range of economic predictions. It is noted for John Stuart Mill's recantation in 1869 – a rare case of an important doctrine being explicitly rejected by a major political economist.

Keywords

avances; Malthus, T. R.; Marcet, J.; marginal productivity theory; Martineau, H.; McCullough, J. R.; Mill, J. S.; Sidgwick, H.; Smith, A.; strikes; wages fund; Walker, F. A.

Article

The wages fund doctrine was an important element in the classical analysis of the labour market – elements of the wages fund doctrine are to be found in the *Wealth of Nations* in 1776 – and articles attempting to defend it were still being produced over a hundred years later. The wages fund doctrine began to take shape in the work of the early classical writers and became more rigid later. On the one hand, it was popularized by Marcet (1816) and Martineau (1832) among others, and employed in more vulgar forms in political debates; on the other, it was used much more carefully and technically by the classical economists to produce a wide range of economic predictions. It is noted for John Stuart Mill's recantation in 1869 – a rare case of an important doctrine being explicitly rejected by a major political economist.

The approach to the wages fund doctrine adopted by the classical writers varied over the period but by the middle of the 19th century there was a generally accepted analysis, from which Mill dissented. In this article I will begin by examining this recantation view of the doctrine and then outline the various phases in its development and decline.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In his book *On Labour*, which appeared in 1869, W.T. Thornton presented a critique of supply and demand analysis and the wages fund doctrine. John Stuart Mill reviewed the book in the *Fortnightly Review* for May and June 1869, and in this review he made his famous recantation from the wages fund doctrine. Mill sketched what he understood the accepted doctrine to be:

> There is supposed to be, at any given instant, a sum of wealth, which is unconditionally devoted to the payment of wages of labour. This sum is not regarded as unalterable, for it is augmented by saving, and increases with the progress of wealth; but it is reasoned upon as at any given moment a predetermined amount. More than that amount it is assumed that the wages-receiving class cannot possibly divide among them; that amount, and no less, they cannot but obtain. So that, the sum to be divided being fixed, the wages of each depend solely on the divisor, the number of participants. In this doctrine it is by implication affirmed, that the demand for labour not only increases with cheapness, but increases in exact proportion to it, the same aggregate sum being paid for labour whatever its price may be. (1869, p. 515)

The predetermined fund is usually conceived of as a stock of wage goods or necessaries, and if this stock is changed from one production period to the next, while the labour force remains constant, the real wage rate would also change. Some classical economists discussed the consequences of a change in the money fund under circumstances where the fund of wage goods remained the same. One line of argument was that the workers could consume only wage goods and not luxuries, and that therefore the increased expenditure on the fixed stock of goods would lead to an increase in prices, leaving the real wage rate unchanged. This can be regarded as the ‘rigid’ version of the theory. The fact that not all classical writers subscribed to this version all the time was ultimately an important feature in its demise.
The origins of the doctrine in British political economy are to be found in the work of Adam Smith, although he was heavily influenced by the Physiocrats’ notion of *avances*. In the *Wealth of Nations* Smith clearly proposed the notion that wages are advanced to workers by capitalists from capital, although there is nothing to suggest that these advances are predetermined or fixed. Moreover, while the wages fund theory presents a competitive solution to the determination of the wage rate, Smith argued for an uncompetitive solution ‘upon all ordinary occasions’ (1776, p. 82) due to the superior bargaining position of the employers.
In the first edition of the *Essay on the Principles of Population* 1798, Malthus put forward all the elements of the wages fund doctrine, including the argument that increased money payments would be offset by increased prices, leaving the real wage rate unchanged. However, in the first edition of the Malthus's *Principles of Political Economy* (1820), the wages fund was not well developed. Labour demand was loosely related to growth in capital or resources. By contrast, in the second edition of the *Principles of Political Economy* (1836), Malthus's discussion of wages is much fuller and much closer to a wages fund approach.
There is no simple, consistent statement of the wages fund doctrine as outlined above to be found in Ricardo's *Principles of Political Economy* (1817), or any of his other published works or correspondence. In the chapter on machinery, Ricardo does present an example which makes use of the logic of the wages fund, but there are other passages in the *Principles* and in his correspondence where

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Ricardo's approach to wages runs directly counter to the doctrine. For example, in correspondence with Trower, Ricardo (1951, VIII, p. 258) argued that under certain circumstances workers can consume luxuries.

The mature phase, when the accepted version fully emerged, is associated with the work of Marcet and McCulloch, and the doctrine was applied to a range of economic issues by Torrens, Senior and J.S. Mill. McCulloch has been regarded by prominent historians of economic thought such as Bonar (1885, p. 272) and Cannan (1893, p. 263) as the author of the analysis which had originated in Marcet's *Conversations on Political Economy* (1816), although it had, in fact, been originally expounded by Malthus in 1798 (see Vint, 1994, pp. 77–88).

McCulloch argues in the *Principles of Political Economy* (1825) that a country's ability to employ workers depended on the existence of an ‘amount of the accumulated produce of previous labour’. The wage rate depended on the ‘proportion which the whole capital bears to the whole amount of the labouring population’ (1825, p.173). He argued that wages do not depend on the amount of money allocated to labourers – if the mount of money halved but the quantity of wage goods remained the same the labourer ‘would carry a smaller quantity of pieces of gold and silver to market than formerly; but he would obtain the same quantity of commodities in exchange for them (1825, p. 174). This was the rigid fund of Malthus.

It was the Malthus–McCulloch rigid version of the wages fund doctrine which underpinned some of the later popularizations used to argue against the efficacy of strikes and the role of trades unions. But as Taussig pointed out (1896, pp. 239–40), the classical economists themselves, up until Mill's *Principles* in 1848, did not use the doctrine in this way. McCulloch himself, for example, eschewing the wages fund doctrine in this context, made a powerful case *for* trades unions in 1824; Torrens (1834) made use of the wages fund analysis to show that unions could act to raise wages. Mill did use the wages fund doctrine to deny the effectiveness of union action in 1848 but his case was heavily qualified.

Among the popularisers, Harriet Martineau made use of the rigid fund in ‘Manchester Strike’, one of the tales from the *Illustrations of Political Economy* (1832), to argue that strikes were futile; William Ellis argued that in a combination to raise wages ‘success is impossible’ because ‘the capital out of which the increased wages are to be drawn does not exist’ (1854, p. 224).

There were clearly potential weaknesses in the rigid version of the doctrine – were wages actually paid from capital; was the fund predetermined; could workers consume only wage goods? Despite these potential flaws the doctrine lasted for almost a century, and from the mid-1820s was an accepted part of classical theory. One potential explanation for its longevity is that it was used successfully by the major political economists such as Senior (1830) and J.S. Mill to produce a wide range of predictions relating, for example, to the effects of the introduction of machinery on the wage rate, the impact of various kinds of war loans on wages and the effects of landlord absenteeism (see Vint, 1994, pp. 124–75).

Mill's recantation came at the end of a decade of change for the British trade union movement. Events on the ground were important – there was an intense political debate concerning the legal status and role of trades unions, culminating in the appointment of Royal Commission of Inquiry in 1867. Alongside these events there was an important theoretical discussion concerning the wages fund doctrine in the work of Fawcett (1860), Longe (1866) and most importantly Thornton (1869). Longe argued that wages were not paid from capital, and both he and Thornton argued that the wages fund was not predetermined. In the ‘recantation’ Mill attacked the theory, arguing that the demand for labour does *not* increase with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
cheapness, that the wages fund is not predetermined and that the money funds in the hands of employers were flexible and could be bargained for by workers. He then went on to produce his famous recantation statement:

> The doctrine hitherto taught by all or most economists [including myself], which denied it to be possible that trade combinations can raise wages, or which limited their operation in that respect to the somewhat earlier attainment of a rise which the competition of the market would have produced without them, – this doctrine is deprived of its scientific foundation, and must be thrown aside. The right and wrong of the proceedings of Trades' Unions becomes a common question of prudence and social duty, not one which is peremptorily decided by unbending necessities of political economy. (1869, pp. 517–18)

Thus in a few short paragraphs Mill apparently disposed of a central tenet of classical economics and one which had a long history. There were attempts to revive the doctrine by Cairnes (1874) and others, but these failed. Cairnes's arguments were thoroughly demolished by F.A. Walker (1875) in the first comprehensive review of the debate. In 1879 Sidgwick also reconsidered the controversy in the *Fortnightly Review*. After agreeing that 'Professor Walker's argument gives a *coup de grâce* to the old wages-fund theory', Sidgwick threw out a challenge to Walker and other economists to put something in its place (1879, p. 411). This challenge and the response of Walker and others to it generated what Gordon referred to as the 'second round debate' (1973, pp. 23–31) and which resulted in the eventual development of marginal productivity theory. Until Sidgwick's watershed contribution, the point of reference for the discussion of wage theory was the debate which began in the 1860s, but thereafter subsequent writers made fewer references to the earlier controversy. There continued to be defences of the doctrine, but often the focal point was Walker's work, and none of these efforts succeeded in undoing the damage done earlier.

See Also

- Cairnes, John Elliott
- machinery question
- McCulloch, John Ramsay
- Mill, John Stuart
- Senior, Nassau William
- strikes
- Walker, Francis Amasa

Bibliography

Bonar, J. 1885. *Malthus and His Work*. New York: Allen and Unwin.

Cairnes, J.E. 1874. *Some Leading Principles of Political Economy Newly Expounded*. London:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Macmillan.

Cannan, E. 1893. *A History of the Theories of Production and Distribution*. London: Percival.

Ellis, W. 1854. *Lessons on the Phenomena of Industrial Life, and the Conditions of Industrial Success*, ed. Rev. R. Dawes. London: Groombridge.

Fawcett, H. 1860. Strikes: their tendencies and remedies. *Westminster Review* NS 18(1 July), 1–23.

Gordon, S. 1973. The wages fund controversy: the second round. *Quarterly Journal of Economics* 93, 501–22.

Longe, F.D. 1866. *A Refutation of the Wages Fund Theory of Modern Political Economy as enunciated by Mr. Mill and Mr. Fawcett, M.P*. London: Longmans, Green.

Malthus, T.R. 1798. *An Essay on the Principles of Population as it affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*. Repr. by Royal Economic Society as *First Essay on Population*. London: J. Johnson, 1926.

Malthus, T.R. 1820/1836. *Principles of Political Economy*, Variorum edition, ed. J. Pullen. London: Cambridge University Press for the Royal Economic Society, 1989.

Marcet, J. 1816. *Conversations on Political Economy, in which the Elements of that Science are Familiarly Explained*. London: Longman, Hurst, Rees, Orme and Brown.

Martineau, H. 1832. The Manchester strike. *Illustrations of Political Economy*, No. 7. London: Charles Fox.

McCulloch, J.R. 1825. *Principles of Political Economy*. London: Ward, Lock and Co.

Mill, J.S. 1848. *Principles of Political Economy*, vols. 2 and 3 of *Collected Works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Mill, J.S. 1869. Thornton on labour and its claims. *Fortnightly Review*, May/June, 505–18, 680–700.

Ricardo, D. 1817. Principles of Political Economy. In *The Works and Correspondence of David Ricardo*, vol. 1, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1951. *The Works and Correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Senior, N. 1830. *Three Lectures on the Rate of Wages*. London: John Murray.

Sidgwick, H. 1879. The wages fund doctrine. *Fortnightly Review*, September, 401–13.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Oxford University Press, 1976.

Taussig, F.W. 1896. *Wages and Capital*. London: Macmillan.

Thornton, W.T. 1869. *On Labour: Its Wrongful Claims and Rightful Dues, Its Actual Present and Possible Future*, London: Macmillan.

Torrens, R. 1834. *On Wages and Combinations*. London: Longman, Rees, Orme, Brown, Green and Longman.

Vint, J. 1994. *Capital and Wages: A Lakatosian History of the Wages Fund Doctrine*. London: Edward Elgar.

Walker, F.A. 1875. The wage-fund theory. *North American Review* 120(246), 84–19.

How to cite this article

Vint, John. "wages fund." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000134> doi:10.1057/9780230226203.1809

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wakefield, Edward Gibbon (1796 – 1862)

Donald Winch

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

colonization; exploitation; imperialism; international migration; Wakefield, E. G.

Article

Wakefield was born in London on 20 March 1796, the eldest son of Edward Wakefield, a radical Quaker philanthropist, statistician, and author of a standard work on Ireland which was highly regarded by Ricardo, James Mill and other members of the philosophic radical circle. His son was to become one of the more colourful characters to inhabit the margins of the history of economic debate, and can be variously described as a publicist, politician and author. Apart from his practical and frequently controversial contributions to the development of Australia, Canada and New Zealand, he left a distinctive mark in the annals of classical political economy during the middle third of the 19th century. After a chequered education at Westminster School and Edinburgh High School, from which he was expelled in 1811, Wakefield first read for the Bar and later served as secretary to the British envoy to the Court of Turin (1814–20). In 1816 he successfully eloped with a 16-year-old Ward-in-Chancery who died in childbirth in 1820. From 1820 to 1825 he served with the British legation in Paris and entertained ambitions of entering Parliament. In 1826 he made an attempt to acquire a rich wife by the most direct means available: he abducted the daughter of a wealthy family from her school and married her at Gretna Green. He was apprehended by her family at Calais and subsequently given a three-year sentence, which he spent studying capital punishment and transportation, writing a powerful pamphlet condemning the former, and turning the latter into what was to become a lifetime's preoccupation with colonization. His first work on the subject, *A Letter from Sydney* (1829), purporting to be the reflections of a disillusioned settler on the poor prospects for Australian social and economic development, was actually written from Newgate prison. After his release Wakefield produced a spate of books, articles and prospectuses on the subject of colonization which led to the formation of the National Colonization Society in 1830 – a society which obtained the support of a number of Members of Parliament and of the youthful John Stuart Mill. Although most of his writings dealt with colonization in one form or another, his work on *England and America: A Comparison of the Social and Political State of Both Nations*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(1833) is of wider interest for its diagnosis of the cause of the ‘uneasiness of the middle classes’ and for its economic interpretation of slavery. Wakefield also produced an edition of the *Wealth of Nations* (1835–9) which has some interesting editorial comments.

Wakefield's views on colonization were based on a dual analysis of Britain's need for an outlet for its surplus capital and population and a diagnosis of the causes of weak economic development in colonies of new settlement enjoying access to abundant land. His own schemes for ‘systematic colonization’ were intended as an almost self-regulating solution to both of these problems. Making use of ideas derived from the work of Robert Gourlay, Wakefield advanced a theory of growth in new countries which was designed to support a plan of optimal development. Contrary to the received view, he maintained that access to free or cheap land was responsible for population dispersion, scarcity of labour for hire, and consequent inability to reap the benefits of economies of scale through market concentration and the combined efforts of capital and labour. Under these circumstances the ‘natural’ pattern of development led to stagnation. Convict labour in Australia and slavery in the American South were both unsatisfactory expedients adopted to deal with a problem that could only be overcome by charging a ‘sufficient price’ for public or waste land which would deter premature dispersion, stabilize a revolving wage-labour force, and create a fund that could be used to subsidize immigration. The price was defined as one that was high enough to delay land acquisition by newly arrived immigrants without capital of their own, and low enough not to discourage voluntary immigration by reducing real wages and the return on capital.

Colonization on this plan required a new beginning in a colony that was not contaminated by convict labour; and for this purpose Wakefield initially chose South Australia, forming an association for this purpose in 1834. When his proposals were diluted in operation by the founders of the colony (among them another political economist, Robert Torrens), Wakefield turned his attention to New Zealand, serving as the Director of the New Zealand Colonization Company from 1839 to 1846. In 1838 he accompanied Lord Durham on his mission to Canada and wrote the appendix on land disposal to the resulting Durham report.

Wakefield's ideas are of interest for a number of reasons. He belongs to the non-Ricardian underworld by virtue of his attack on Say's Law, the wage-fund doctrine, and the associated idea that capital and labour could never be in surplus together – a mirror image of the problem in colonies where both were scarce. Yet his success in convincing John Stuart Mill and other economists of the correctness of his diagnosis of British and colonial problems gave new significance to the export of capital and labour to colonies and hence to the whole subject of colonization and the development of new countries as a topic within orthodox political economy.

Wakefield also plays a part in the Marxian tradition, or rather its demonology, as a result of Marx's decision to devote a chapter of *Capital* (vol. 1, ch. 23) to showing how Wakefield, under colonial conditions of labour scarcity, had been forced to reveal the underlying logic of capitalist exploitation. What could be achieved quite naturally under European conditions had to be created artificially in new colonies, with the additional subtlety that having served a term of exploitation, the wage-labourer had to pay for his replacement. One could also claim that Wakefield, less unwittingly, anticipated Hobson and Lenin in providing an economic interpretation of imperialism as a necessary response to stagnation in mature capitalist economies.

In 1853 Wakefield finally practised what he had been preaching by emigrating to New Zealand, where he died in 1862.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Selected works

1968. *The Collected Works of Edward Gibbon Wakefield*, ed. M.F. Lloyd Prichard. London: Collins.

Bibliography

Mills, R.C. 1915. *The Colonization of Australia,1829–42: The Wakefield Experiment in Expire Building*. London: Sidgwick & Jackson.

Winch, D. 1965. *Classical Political Economy and Colonies*. London: Bell & Sons.

How to cite this article

Winch, Donald. "Wakefield, Edward Gibbon (1796–1862)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000012> doi:10.1057/9780230226203.1810

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wald, Abraham (1902 – 1950)

E. Roy Weintraub
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Cassel, G.; general equilibrium; Menger, K.; Schlesinger, K.; sequential analysis; statistical decision theory; Wald, A.

Article

Born in Cluj, Rumania, Wald came to Vienna in 1927 to study mathematics with Karl Menger, the geometer and son of the economist Carl Menger. Menger introduced Wald to the active mathematical group in Vienna, and secured for him a position as mathematical tutor to the economist Karl Schlesinger. This led to Wald's producing the first proofs of existence for models of general equilibrium; his analysis was based on Cassel's restatement of the Walrasian model, as modified by Schlesinger's treatment of free goods. These works were published in the proceedings of Menger's mathematical colloquium, and a summary was published in the *Zeitschrift für Nationalökonomie* in 1936. These papers were remarkable for their time and, with von Neumann's paper on equilibrium in a model of an expanding economy, are the first significant contributions to the mathematical analysis of general equilibrium models in economics. Wald is the link between the early work by Walras and the later work by Kenneth Arrow, Gerard Debreu and Lionel McKenzie on the existence of competitive equilibria.

A fine mathematician, Wald was nevertheless prevented from gaining a regular academic position because of Viennese anti-Semitism. Menger helped Wald secure a consultancy position with Oskar Morgenstern who directed the Institut für Konjunkturforschung, where Wald took an interest in the statistical problems that were associated with the analysis of business cycles. Wald's book on seasonal adjustment of time series was a result of his work at Morgenstern's Institut.

Wald was able to escape from Vienna when the Nazis arrived, and made his way to the United States where he initially secured a fellowship, in 1938, at the Cowles Commission which was then at Colorado Springs. When the Commission moved to Chicago, Wald obtained a position, on a Carnegie grant, as Harold Hotelling's assistant at Columbia University. He moved to a faculty post at Columbia in 1941, and was promoted to Associate Professor in 1943 and Professor in 1944.

Wald's contributions to statistics are immense. His most significant paper appeared in 1939 in the *Annals*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Mathematical Statistics as ‘Contributions to the Theory of Statistical Estimation and Testing Hypotheses’ (in Wald, 1955). This paper, written before modern decision theory was developed, contained notions of decision space, weight and risk functions, and minimax solution (based on von Neumann's 1928 paper on game theory). Wald's paper was not appreciated at the time, much as was the case with his papers on general equilibrium theory. He did not return to statistical decision theory until 1946, after von Neumann and Morgenstern had presented the theory of games.

During the Second World War, Wald worked with the Statistical Research Group and developed much of the theory of sequential analysis. Although he did not create the idea of taking observations sequentially, Wald did invent the sequential probability ratio test. This original material was published in 1947 after wartime restrictions were lifted.

In 1950, at the height of his powers, Wald and his wife died in a plane crash in India.

See Also

- existence of general equilibrium
- revealed preference theory
- statistical decision theory

selected works

In 1952 *The Annals of Mathematical Statistics* devoted the first part of its Volume 23 to a memorial to Wald. Articles on Wald by Jacob Wolfowitz, Karl Menger and Gerhard Tintner were followed by a complete bibliography of Wald's writings. Wald's professional correspondence, and papers from his Viennese days, cannot be located, though it is possible that Karl Menger's archives, currently closed to examination, may contain some material on Wald.

1934. Über die eindeutige positive Lösbarkeit der neuen Produktions gleichungen I. In *Ergebnisse eines mathematischen Kolloquiums, 1933–34*, ed. K. Menger. Trans. W. Baumol as ‘On the unique non-negative solvability of the new production equations, Part I’, in *Precursors in Mathematical Economics*, ed. W.J. Baumol and S.M. Goldfeld, London School of Economics Series of Reprints of Scarce Works on Political Economy No. 19, London: London School of Economics, 1968.

1935. Über die Produktionsgleichungen der ökonomischen Wertlehre II. In *Ergebnisse eines mathematischen Kolloquiums*, ed. K. Menger, 1934–35. Trans. W. Baumol as ‘On the production equations of economic value theory, Part II’, in *Precursors in Mathematical Economics*, ed. W.J. Baumol and S.M. Goldfeld, London School of Economics Series of Reprints of Scarce Works on Political Economy No. 19, London: London School of Economics, 1968.

1936. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie*. Trans. O. Eckstein as ‘On some systems of equations in mathematical economics’, *Econometrica* 19 (1951), 368–403.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1947. *Sequential Analysis*. New York: John Wiley.

1950. *Statistical Decision Functions*. New York: John Wiley.

1955. *Selected Papers in Statistics and Probability*. New York: McGraw-Hill.

How to cite this article

Weintraub, E. Roy. "Wald, Abraham (1902–1950)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000013>
doi:10.1057/9780230226203.1811

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Walker, Francis Amasa (1840 – 1897)

A.W. Coats
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

bimetallism; residual claimant theory of wages; wages fund; Walker, F. A.

Article

Internationally the most widely known and esteemed American economist of his generation, Walker had a varied and distinguished public career. After obtaining his AB at Amherst in 1860, he studied law for one year before joining the Northern army and was successively a Civil War general, deputy to David A. Wells in the Budget Office, chief of the US Treasury's Bureau of Statistics, Superintendent of the Census of 1870 and 1880, Professor of Political Economy and History at Yale's Sheffield Scientific School, and also occasionally at Johns Hopkins, and President of the Massachusetts Institute of Technology, 1881–97. At home Walker was primarily known as an outstanding educational administrator and statistician, for he permanently raised the standards of government statistics, helped to create a permanent Bureau of the Census, and served as President of the American Statistical Association from 1882–97. Abroad, he was recognized more as an economic theorist, especially for his work on wages, money and currency policy.

His attack on the wages fund and formulation of a residual claimant theory of wages attracted widespread attention, though it gained few adherents. His writings on money, and a textbook on political economy, were also well regarded, and his support for bimetallism, which involved the monetization of silver, represented an important contribution to a highly controversial current policy debate. In 1878 Walker was appointed US Commissioner to the Paris International Monetary Conference, but in later years he refused comparable invitations as he became disenchanted with the slow progress of international negotiations.

A moderate critic of the ruling classical laissez-faire orthodoxy, Walker responded sympathetically to the rising young generation of German-trained American economists, hence he was both an obvious and in practice ideal choice as first President of the American Economic Association, from 1885 to 1892. His presidential addresses provide revealing insights into the condition of the subject and the emerging economics profession during those critical years. Walker was an open-minded man, forthright in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
expression but fair in controversy. He believed in competition while recognizing its imperfections. An undoctrinaire free trader, he was concerned about the growth of immigration and the decline in the native birth rate. An advocate of moderate reductions in hours of work, he was one of the first American economists to recognize entrepreneurial gains as rents of ability. Fifty years after his death, Walker's eminence was acknowledged when the American Economic Association selected his name for its most distinguished award, the Walker Medal.

Selected works

1876. *The Wages Question: A Treatise on Wages and the Wages Class*. New York: H. Holt, 1906.

1878. *Money*. London: Macmillan; New York: H. Holt, 1891.

1879. *Money in its Relations to Trade and Industry*. New York: H. Holt, 1907.

1883. *Land and its Rent*. Boston: Little, Brown, 1891.

1883. *Political Economy*. New York: H. Holt. 3rd revised and enlarged edn, 1911. Abridged, 1912.

1896. *International Bimetallism*. London: Macmillan; New York: H. Holt, 1897.

1899. *Discussions in Economics and Statistics*. 2 vols, ed. D.R. Dewey. New York: H. Holt.

1899. *Discussions in Education*, ed. J.P. Munroe. New York: H. Holt.

Bibliography

Newton, B. 1968. *The Economics of Francis Amasa Walker: American Economics in Transition*. New York: Kelley.

How to cite this article

Coats, A. W. "Walker, Francis Amasa (1840–1897)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000014>
doi:10.1057/9780230226203.1812

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Walras, Antoine Auguste (1801 – 1866)

Donald A. Walker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

land nationalization; numeraire; property rights; scarcity; Walras, A. A.; Walras, L.

Article

Amateur economist whose writings have received some limited attention, chiefly because some of his views and economic concepts influenced his son, Léon. Auguste Walras was born in Montpellier, France, on 1 February 1801, and died in Pau on 18 April 1866. He studied at the Ecole Normale of Paris (1820–23); was a tutor in Paris (1823–31); a secondary school teacher (1823, 1831–5); a professor of philosophy first at the Royal College in Lille (1839) and then at the Royal College in Caen (1840–47); and a regional school superintendent (1847–62).

Believing that an understanding of property requires a sound theory of value, Auguste Walras developed the unoriginal and unsatisfactory thesis, primarily on the basis of admittedly metaphysical considerations, that economic value depends upon scarcity (*rareté*). This he defined as the relation between the quantity of a commodity and the number of people that have need for it. He concluded that only scarce things are appropriated and constitute property (Walras, 1831). He then argued that natural law dictates that the state, like the individual, has the right to own property, and that land in particular should belong exclusively to society as a whole. Developing an explanation of the current ownership of land, he pointed out that it is a consequence of social institutions and historical events. During the feudal era, it was placed by the king under the suzerainty of individuals in return for their military services, and their descendants subsequently ruled it as public officials. Since the need for their feudal functions has disappeared, so also has their right to the use of land, and they have become parasites who benefit from economic growth without contributing to it. The class struggle is therefore between landowners and the rest of society, and social justice requires that it be resolved in favour of society as a whole. Believing in conciliation and rejecting revolutionary action, he argued that the state should acquire all land by purchasing it, and should rent it to private users. During the period before complete nationalization the increments in pure land rent arising from the progress of society should be taxed away, and there should be heavy taxes on transfers of land. Since individuals have the right to own what they make, taxation of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
produced wealth, as distinct from rent, is unjust. It is therefore an advantage of land nationalization that the state would be supported by the rent it earns and taxation could be eliminated (Walras, 1848). He regarded this proposal as being founded upon scientific analysis and described himself as a socialist, but it is clear that his interpretation and solution of 'the social problem' – the problem of the poverty of the working class during the 19th century – was highly coloured by his normative views and was bourgeois in character.

Auguste Walras also studied the function of precious metals in the growth of social wealth, in the measurement of value, and in exchange; argued that the increase of wealth is the object of economic science; made a distinction between capital and income, and between the market for services and the market for products; introduced the entrepreneur as a person who buys factors of production and sells products; and devised the concept of a *numéraire* (Walras, 1849). He did not fully develop these ideas nor integrate them into a theory of economic behaviour. His main concern in all his work was to buttress his theory of property and his solution to the social problem.

Selected works

1831. *De la nature de la richesse et de l'origine de la valeur*. Paris: Alexandre Johanneau; Evreux: Imprimérie d' Ancelle Fils. Ed. G. Leduc, Paris: Alcan, 1938.

1848. *La vérité sociale, par un travailleur*. Extracts of MS in Leroy (1928).

1849. *Théorie de la richesse sociale our résumé des principes fondamentaux de l'économie politique*. Paris: Guillaumin.

Bibliography

Jaffé, W. 1984. The antecedents and early life of Léon Walras. *History of Political Economy* (ed. D.A. Walker) 16, 1–57.

Leroy, L.M. 1928. *Auguste Walras, économiste; sa vie, son oeuvre*. Paris: Librairie Générale de Droit et de Jurisprudence.

Walras, L. 1908. Un initiateur en économie politique, A.A. Walras. *Revue du Mois* 6, August, 170–83.

How to cite this article

Walker, Donald A. "Walras, Antoine Auguste (1801–1866)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000018> doi:10.1057/9780230226203.1813

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Walras, L é on (1834 – 1910)

Donald A. Walker
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Léon Walras was the initiator of models of purely competitive general economic equilibration and equilibrium, of mathematical treatments of them, and of many aspects of microeconomic theory. In his period of maturity as a theoretician, he developed a comprehensive model that includes exchange, production of non-durable goods, production of capital goods, and monetary behaviour. That model features irrevocable disequilibrium behaviour and capital accumulation, and its equilibrium is therefore path dependent. His last theoretical effort, which was a failure but nevertheless very influential, was to try to develop a virtual and therefore path-independent model that would justify his static equation system.

Keywords

American Economic Association; Amstein, H.; applied economics; Barone, E.; bimetallism; Bortkiewicz, L. von; capital accumulation; capital theory; cash-balances theory of money; Cassel, K. G.; certainty; circular flow; comparative statics; competition; Condillac, E. B. de; consumers; convergence; Cournot, A. A.; Cousin, V.; demand function; durable goods; dynamic equation system; E commodity; economic growth; Edgeworth, F. Y.; entrepreneur; exchange; excess demand; existence of equilibrium; fiat money; Fisher, I.; Galiani, F.; general equilibrium; human nature; immaterial goods; interrelated markets; investment; Isnard, A. N.; Jevons, W. S.; justice; landlords; Lexis, W.; losses; marginal productivity theory; marginal utility; marginal utility of money; Marshall, A.; mathematical economics; mature comprehensive model; methodology; Mill, J. S.; monetary equilibrium; monopoly; Moore, H. L.; moving equilibrium; multi-good exchange; nationalization of land; natural monopoly, nationalization of; non-durable goods; non-virtual model; normative economics; Pantaleoni, M.; Pareto, V.; path dependency; parameters of the models; period analysis, Walrasian; Poinsot, L.; positive economics; preferences; price stabilization; production model; profit; Proudhon, P.-J.; Quesnay, F.; rareté; rate of interest; real balances; saving; Say, J.-B.; Schultz, H.; Schumpeter, J.; services; simultaneous equations;

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Smith, A.; social economics; stability of equilibrium; static equation system; stationary equilibrium; supply function; tatonnement; taxation, injustice of; Turgot, A.R.J.; uncertainty; uniqueness of equilibrium; utility; utility maximization; veil of money; virtual model; wages of management; Walras, A. A.; Walrasian equilibrium; Walrasian pricing; Walras, L.; Walras's Law; Wicksell, J. G. K.; Wicksteed, P. H.; workers; written pledges

Article

Léon Walras was the founder of models of general economic equilibrium.

1 Biography

Walras was born on 16 December 1834 in Evreux, which is in the Department of Eure in France, and christened Marie Esprit Léon. His father was Antoine Auguste Walras, a secondary school administrator and an amateur economist; his mother was Louise Aline de Sainte Beuve, the daughter of an Evreux notary. After studying at the College of Caen from 1844 to 1850, he entered the Lycée of Douai, where he received the bachelier-ès-lettres in 1851 and the bachelier-ès-sciences in 1853. He entered the School of Mines of Paris in 1854, but, finding the course of preparation of an engineer not to his liking, he gradually abandoned his academic studies in order to cultivate literature, philosophy and social science. Although those efforts resulted in a short story and a novel, *Francis Sauveur* (1858), it rapidly became apparent to him that his true interests lay with social science. Accordingly, in 1858 he agreed to his father's request to devote himself to economics and promised to continue his father's investigations (1965, vol. 1, pp. 1–2).

During his youth in Paris, Walras became a journalist for the *Journal des Economistes* and *La Presse* from 1859 to 1862, the author of a refutation on philosophical grounds of the normative economic doctrines of P.-J. Proudhon (Walras, 1860), an employee of the directors of the Northern Railway in 1862, and managing director of a cooperative association bank in 1865. He gave public lectures on cooperative associations in 1865; was co-editor and publisher with Léon Say of the journal *Le Travail*, a review devoted largely to the cooperative movement, from 1866 to 1868; and, during those years, gave public lectures on social topics (Walras, 1868) in which he advocated Victor Cousin's doctrine of compromise between economic classes. After the failure of the association bank in 1868, he found employment with a private bank until 1870 (1965, vol. 1, pp. 3–4). During the 1860s he tried intermittently to obtain an academic appointment in France, but he lacked the necessary educational credentials, and the 11 economics positions in higher education in France were monopolized by orthodox economists who, he complained, passed their chairs on to their relatives (1965, vol. 1, p. 3). His fortunes ultimately changed as a result of his participation in 1860 in an international congress on taxation in Lausanne, for that drew him to the attention of Louis Ruchonnet, a Swiss politician who secured his appointment in 1870 to an untenured professorship of economics at the Academy (subsequently University) of Lausanne in Switzerland. He was made a tenured professor there in 1871, and held that position throughout his teaching career.

Walras's personal life was initially unconventional. He and Célestine Aline Ferbach (1834–79) formed a common law union in the late 1850s. She had a son, Georges, by a previous liaison, and she and Walras

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
had twin daughters in 1863, one of whom died in infancy. In 1869 he married Célestine, thereby legitimizing their daughter, Marie Aline, and adopted Célestine's son. A long illness of Celestine's and the meagreness of Walras's salary made life very difficult for him for several years. His time and energy were sorely taxed not only by the need to care for his wife but by the need to supplement his salary by teaching extra classes, contributing to the *Gazette de Lausanne* and the *Bibliothèque Universelle*, and working as a consultant for La Suisse insurance company. Five years after Célestine's death in 1879, Walras married Léonide Désirée Mailly (1826–1900). The marriage was a happy one. Her annuity relieved his financial distress, and his situation was further improved in 1892 by an inheritance of 100,000 francs from his mother, which enabled him to pay debts incurred in publishing and disseminating his works, and to buy an annuity of 800 francs.

2 Influences upon his thought

Walras's professional life was devoted to research and teaching. He frequently asserted that his research was a development of his father's and that was true in some respects. It was under the influence of his father's classification of economic studies that Léon, as early as 1862, planned the division of his life's work into the study of pure theory, economic policies and normative goals (Walras to Jules du Mesnil Marigny, 23 December 1862, L 81; the ‘L’ stands for ‘letter’, and, like all correspondence cited in this article, the letter is in Walras, 1965), the areas of study that were ultimately set forth respectively in the *Eléments d’économie politique pure* (1874; 1877b; 1889; 1896a; 1900; 1926; 1954), the *Etudes d’économie social* (1896b) and the *Etudes d’économie politique appliquée* (1898). Léon adopted his father's classification of the factors of production into the services of labour, land and capital goods, regarding the source of each service as a type of capital. He adopted his father's definitions of capital as wealth that can be used more than once and of income as wealth that can be used only once, and modified his father's vague term ‘extensive utility’, clarifying it by defining it as the quantity-axis intercept of a market demand curve. The topic of utility had been treated in French thought by writers such as F. Galiani (a Neapolitan diplomat at Versailles) and E.B. de Condillac, and it was given further development under the name *rareté* by Auguste Walras, who thus bequeathed to Léon an interest in the concept of utility in relation to the value of goods and an awareness of its dependence upon scarcity, an interest that ultimately led him to define *rareté* as marginal utility. Auguste used the word ‘numeraire’ to mean an abstract unit of account, and Léon adapted the meaning of the word to his purposes. Auguste's philosophy of social justice and his belief in the desirability of nationalizing land were advocated by Léon throughout his adult life. Léon's major economic theories, however, were derived from his own original inspiration and from sources other than his father. Auguste's greatest contributions to Léon's development as an economist were to encourage him to study economics, to suggest that it should be a mathematical science (A.A. Walras, 1831, ch. 18; Walras, 1965, vol. 1, p. 493), and to give him access to a library of books on economics.

In that library was A.A. Cournot's *Recherches sur les principes mathématiques de la théorie des richesses* (1838), which Léon Walras credited with having demonstrated that economics could and should be expressed in mathematical form (Walras to Cournot, 20 March 1874, L 253; Walras to H.L. Moore, 2 January 1906, L 1614; Walras, 1905a). Cournot's work introduced Walras to the mathematical formulation of exchange between two locations, the theory of monopoly and the associated conditions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for profit maximization, the analysis of how prices are repeatedly changed in a search for equilibrium in a purely competitive market, and the demonstration of the effect of large numbers of traders upon the determinacy of price, all topics that Walras developed in his own work (1954, pp. 370–72, 434–40, 443).
The first demand curve that Walras beheld was Cournot's and he found it immensely suggestive. He was critical of it, however, because he perceived that Cournot's postulate that the quantity demanded of a good is a function only of its own price is inaccurate if more than two goods are exchanged, and that Cournot did not provide a theoretical rationale for the demand function. Those perceptions, Walras observed, were the starting point for his own inquiries (1965, vol. 1, p. 5; 1905a).
Other ingredients that went into the composition of Walras's theories were provided by Adam Smith, John Stuart Mill, François Quesnay, A.R.J. Turgot and Jean-Baptiste Say. Smith had revealed many of the consequences of unfettered competition and had formulated the concept of normal value. Mill had provided a supplement to and reinforcement of Cournot's and Smith's analyses of competitive pricing (Walras to Ladislaus von Bortkiewicz, 27 February 1891, L 999), and also an extension and grand synthesis of classical doctrines that served Walras as a catalyst for critical studies (Walras, 1954, pp. 404–5, 411, 419, 423). Quesnay, in his *Tableau économique*, had expressed the concept of a circular flow of income and of the interdependence of the various parts of the economy. Turgot had clearly delineated the idea of the simultaneous and mutually determined general equilibrium of those parts. Say (1836) had suggested the distinction between the capitalist and the entrepreneur had portrayed the entrepreneur as an intermediary between the market for productive services and the market for outputs, and, in that analysis and in his law of markets, had adumbrated the interdependence between the incomes of the factors of production and the demand for goods. Walras sharpened those ideas and made them a fundamental part of his general equilibrium model.
A.N. Isnard's *Traité des richesses* (1781), a book that Léon owned and that may have been in his father's library, was probably an important source of some of Walras's constructions (Jaffé, 1969; Klotz, 1994). Like Walras, Isnard was interested in determining equilibrium price ratios, set up a system of simultaneous equations of exchange showing the dependence of the value of each good upon the values of the others, stressed the necessity of having as many independent equations as unknowns, and perceived that the use of a numeraire rendered his system determinate. Anticipating Walras's treatment of production, Isnard assumed given ratios of the inputs in a mathematical model and expressed the costs of production in equation form. Also like Walras, Isnard studied the allocation of capital among different uses, coming to the conclusion, as did Walras, that in equilibrium the net rate of income of different capital goods is the same.
Finally, Louis Poinsot's *Eléments de statique* (1803) exerted a powerful influence upon Walras. He first read that book when he was 19 years old and kept it at his bedside for decades (Walras to Melle Dick May, 23 May 1901, L 1483). Poinsot painted a picture of the mutual interdependence of a vast number of variables, of how the dynamic forces in physical systems eventuate in an equilibrium in which each object is sustained in its path and relative position. Electrified by the implications of Poinsot's work, Walras conceived a magnificent project. He would emulate Poinsot's vision and analysis in reference to the general equilibrium of the economic universe! That he carried out that plan can be inferred from the striking similarity of the form of his work to Poinsot's, with its careful delineation of functional dependences and parameters, its sets of simultaneous equations and its equilibrium conditions.
Equipped, therefore, with ideas that he could take as building blocks and points of departure, with enough geometry and algebra to put together mathematical statements of economic relationships and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
conditions – his use of calculus in the *Eléments* came after the first edition – and with the explicit objective of developing a mathematical theory of general equilibrium, Walras began his scholarly activity in Lausanne in 1870. In a period of great creativity that lasted until 1878, he developed most of the foundations of the theory of general equilibrium that appeared in the first edition of the *Eléments*. Walras insisted to his publisher that the first part appear in 1874, before the second part (1877b) was completed, because he learned in May of that year that W.S. Jevons had published a mathematical theory of utility and exchange that was similar to his own (J. d'Aulnis de Bourouill to Walras, 4 May, 1874, L 267), and he was anxious to establish the independence of his discoveries and his priority in regard to most of them. For those same reasons, he published four brilliantly original memoirs containing the heart of his theory of general equilibrium during 1874, 1875 and 1876 (Walras, 1877a), paid for the costs of publication of his books, and sent copies of them and of his articles to his many correspondents. From 1878 to 1889, Walras significantly extended and refined his theory of general equilibrium (Section 3).

Walras was an extremely conscientious teacher, but he was an uninspiring lecturer (Walras, 1965, vol. 2, p. 560), and the students at Lausanne were interested in careers in law, not in economics, so he failed to develop disciples among them. Moreover, he was with increasing frequency afflicted by bouts of mental exhaustion and irritability that made it difficult for him to lecture and to read and write (see Walker, 2006a, pp. 183–7). In 1892 he took a leave of absence to regenerate his strength in order to be able to continue teaching, but soon realized he would find the strain of returning to his tasks insupportable and retired in that year, being at that time 58 years of age.

Subsequently Walras's powers waned rapidly. In 1899 and 1900, he tried unsuccessfully to develop a virtual model of general equilibration and equilibrium (Section 4). After 1900 he completely ceased theoretical construction (Walker, 2006a, p. 191), but he wrote a few articles in which he restated earlier ideas. In late 1901 and 1902, he made some inconsequential changes to the *Eléments* which were ultimately put into the text of the fourth edition (1900) to produce the 1926 edition, both of them unfortunately called the ‘definitive edition’ (1900, p. v; 1926, title page). The latter was chosen for translation by William Jaffé (Walras, 1954) and thus became the edition that is known in the anglophone world. Walras died on 5 January 1910 in Clarens, Switzerland.

3 The mature comprehensive model of general economic equilibrium

Walras's subject matter

Walras recognized that there were imperfectly competitive market structures and developed a theory of monopoly to take account of an important class of such phenomena (1954, lesson 41). Realizing, however, that the incorporation of non-competitive elements into his general equilibrium model was beyond his powers (1954, p. 256) and believing that a high degree of competition was ‘almost universal’ and deserved to be treated as the general case (Walras to Ladislaus von Bortkiewicz, 27 February 1891, L999), he devoted most of his energies to working out a comprehensive model of interrelated ‘freely competitive’ markets, the aspect of his theoretical work with which this entry is concerned. Competition is most effective, he noted, in organized markets, and he assumed that markets are of that type (1954, pp. 83–4), but he also regarded his analysis as applicable in a general way to less highly organized

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
competitive markets (1954, p. 84).

During his period of maturity as a theoretician, Walras modified and extended the model of competitive general equilibrium that he had presented in the first edition of the *Eléments*, constructing what will be called his mature comprehensive model. He presented it in 1889 in the second and greatest edition (and in the third, (1896a), identical to it in regard to the main body of the text). In the following exposition of that model, all the references to Walras (1954) are to passages that appeared in the 1889 edition.

The model is comprehensive in the sense that it deals with exchange, production, capital formation and credit, and monetary behaviour. It is non-virtual: it deals with irrevocable exchange at prices that are disequilibrium ones from the point of view of the state of the entire set of markets, and with the non-virtual dynamics of production, consumption, saving and investment. Those irrevocable economic activities occur during the course of the equilibrating process and are part of it (1889, pp. 235, 280). The sub-models included in the comprehensive model, such as the models of consumer demand, of the firm, of the entrepreneur, of exchange, of production, and so on, will sometimes be called theories, because Walras had reference to the behaviour of the real economy rather than purely hypothetical schemes.

Each major sub-model has four parts: structure, equilibration, equilibrium conditions and comparative statics.

Regarding the structure of each market, Walras assumed that preferences, the number of economically active individuals, the amounts of natural resources and technology are constant. He identified consumers, workers, landlords, capitalists and entrepreneurs, their economic characteristics, and their objectives and how they try to achieve them. He specified the types of goods that are traded, the institutional features and rules of the market, and the individual and market supply and demand functions for goods (material and immaterial).

Regarding the dynamic equilibrating processes by which the markets undergo adjustments when in disequilibrium, Walras called them ‘tatonnements’, which means ‘gropings’, to emphasize that the equilibrium magnitudes of prices and quantities are not known by the participants during the disequilibrium phase but are found by repeated trial and error experiments. Walras considered the exposition of tatonnement to be ‘the object and proper goal of pure economics’ because he believed that the real economy is stable (Walras to Bortkiewicz, 17 October 1889, L 927; Walras to Charles Gide, 3 November 1889, L 933). Walras gave a verbal demonstration of the stability of his model. He recognized that the dynamic functioning of markets depends on the economic agents, institutions and conditions identified in the first part of each model, and in order to portray the disequilibrium behaviour that he perceived in the real economy he accordingly discussed the activities and interactions among diverse economic agents in trade and production, the generation and elimination of profits and losses, the operation of the stock market and many other details of behaviour drawn from economic life. Most of his presentation of the model is concerned with its stability, that is, its behaviour in disequilibrium.

Thus the allegation, perpetuated by generations of commentators (for example, Jaffé, 1971, p. 281; 1981, pp. 252–61), that Walras devoted his attention almost exclusively to the conditions of static equilibrium in an abstract model devoid of institutional detail, economic facts and dynamic behaviour is a misrepresentation of his work.

Walras was partially responsible for that misrepresentation, because in 1900 he referred to his general equilibrium model as ‘static’ without qualification, and contrasted it with what he called ‘the dynamic point of view’, by which he sometimes meant the view taken in considering economic growth (Walras, 1954, p. 318). On the other hand, he also stated on many occasions that a dynamic theory is contained in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
his mature comprehensive model, and his usage will be followed in this article. The ‘static theory of exchange’, he wrote, ‘may be defined as the exposition of the equilibrium formula’. The ‘dynamic theory’, in contrast, which Walras claimed to have been the first to explore, is ‘the demonstration of the attainment of that equilibrium through the play of the raising and lowering of prices’ (1895, in 1965, vol. 2, p. 630). Similarly, in responding to Irving Fisher's criticism that he had not considered time, Walras pointed out that that was true only of his exposition of the conditions of static equilibrium, and that he gave a dynamic treatment of production in lesson 20 (1889) of the *Eléments* (Walras to Fisher, 28 July 1892, L 1064).

Theory of exchange

Walras was concerned in this theory with the determination of the equilibrium prices of goods and the quantities of goods exchanged. Setting forth the structure of exchange markets, he assumed that the preferences of the traders and the aggregate amounts of the goods they hold in each market are given. He first assumed that goods (including services) are exchanged directly for each other and then that they are exchanged for money. The participants include brokers, professional traders, retailers, wholesalers, the owners of the factors of production in their capacities as demanders of consumer goods and capital goods properly speaking, and entrepreneurs, who supply and demand goods. The supply and demand functions are reciprocally related (Walras, 1954, pp. 96–7). Given a trader's demand curve for A, its price times the related number of units he wants to buy is his supply of B expressed as a function of the price of A in terms of B. Observing what happens to the areas of the rectangles under the demand curve for A as its price rises, Walras deduced that the quantity supplied of B initially rises and then falls. In the same way, a trader's supply of A can be derived from his demand for B. Walras summed the individual demand and supply curves respectively in the market for A to obtain the market curves, and similarly for B. It will be seen that he adapted and extended this analysis of the dependence of the supply of one good upon the demand for another when he took up the question of multi-good exchange. Walras also assumed that in each market the rule is enforced that disequilibrium transactions are not allowed (1880a, p. 461; 1880b, p. 78; 1954, p. 85). He described that as being true of the 19th-century Paris bourse, but in fact disequilibrium transactions occurred there most of the time (Walker, 2000; 2001), in recognition of which he allowed late in his career that his description is in actuality ‘a hypothesis that no scientific spirit would hesitate to concede to the theoretician’ (Walras, 1895, in 1965, vol. 2, p. 630).

To explain demand and infuse his early model of exchange (1869–70) with purposive action, Walras developed a theory of preferences shortly before 1872 in which he assumed that traders want to maximize utility, that utilities are independent and additive, and that the marginal utility of a good is a decreasing function of the quantity acquired or consumed. Nevertheless, he was floundering in his attempts to relate utility to market behaviour, so he appealed for help to Antoine Paul Piccard, a professor of industrial mechanics at the Academy of Lausanne, who responded in 1872 by developing a model of utility maximization and deriving the individual demand function within it (1965, vol. 1, pp. 308–11), thus meriting a part of the credit that has previously been given exclusively to Walras for that achievement. Everything then fell into place for Walras, and he proceeded to develop the view of economizing and maximizing behaviour that he imprinted on Continental neoclassical economics. He extended the technique shown in Piccard's model, making utility maximization the driving force in each of his models, and obtaining the equilibrium conditions of the participants in a multi-good system (1954,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
lesson 12).
The dynamic behaviour of Walras's exchange model is a tatonnement process in the sense that the path of the price in each market is the unplanned outcome of market forces. The process depends upon human nature (see Walker, 2006a, pp. 114–39) and on the rules, institutions and conditions devised and enforced by market authorities and by government (Walras, 1880a; 1880b; 1895 in 1965, vol. 2, p. 632; and see 1954, p. 474). A price is initially cried at random (1877b, p. 127; 1954, p. 169) by any of the traders, and the suppliers and demanders subsequently follow the Walrasian pricing rule: that is, they change the price in the same direction as the sign of the market excess demand for the good. Suppliers call out progressively lower prices if it is negative, and demanders call out progressively higher prices if it is positive. Preferences are constant, and the rule against disequilibrium transactions ensures that the asset distribution remains unchanged during the equilibrating process. Therefore, the initial supply and demand functions and, consequently, the particular-equilibrium price on any given day in the temporarily isolated market, are not affected by the disequilibrium behaviour of the traders. That price equates the supply and demand quantities; it is quoted sooner or later, and the equilibrium amounts of the good are exchanged (1954, p. 106, lessons 6, 9).
Markets are not isolated, however, so Walras introduced the central feature of his contribution to economic science, namely an account, in his theory of exchange and in the other parts of his mature comprehensive model, of the interrelationships among the markets for different goods (including services). If a trader has a good that he wants to trade for several others, the amount that he offers in any market is related to the amounts that he offers in the other markets, so the amount that he wishes to purchase or sell of any good is seen to be a function not only of his preferences, his income and the price of that good but also of the prices of other goods. Consequently, the market supply and demand quantities and the price in any market are dependent in part upon the prices in other markets (1954, lesson 12).
Moreover, Walras explained that the sum of the values of a trader's quantities demanded must equal the sum of the values of his quantities supplied. That relation is one way of stating the individual budget equation, and it is a version, on the individual level (1954, p. 165), of what has come to be known as Walras's Law, a fundamental statement of the way that markets are interrelated. Walras was able to identify the law in part by reasoning an individual cannot demand any commodity without offering in return commodities (or money) having the same total value, so, if some of his excess demands are positive, others of them must be negative, and in part because of the device of the numeraire. The latter, a good in terms of which the values of all goods are expressed (1954, p. 161), made clear to him, as it had to Isnard, that there is exactly the right number of excess demands: in a system with n goods, there are only n–1 independent market equations involving n–1 price ratios, but also only n–1 unknowns, inasmuch as the price of the numeraire, the nth good, in terms of itself is unity (1954, pp. 161–2, 241). With reference to the market level in multi-commodity exchange, Walras affirmed that the sum of the positive or negative market excess demand quantities for each good multiplied by its price is zero (1954, p. 170), and he stated a version of Walras's Law for the market excess demand quantities of productive services (1954, p. 248). In a Walrasian equilibrium, supply equals demand for every good, so each excess demand quantity is zero. Each excess demand quantity, multiplied by the price of the good, must therefore be zero, so the sum of the excess demands each multiplied by the price must be zero. In the case of an individual, Walras stated only, regarding the variables, that ‘there will be between them all’ the relationship that indicates that their sum is equal to zero, without addressing explicitly whether the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
equation is true in disequilibrium as well as in equilibrium (1889, p. 143, which differs from Walras, 1954, p. 165). In the case of multi-commodity exchange, he probably implied that it is always true even though in disequilibrium the market supply and demand quantities of every commodity are not simultaneously equal (1954, pp. 169–70). In the productive services formulation, he indicated explicitly in the *Eléments* that it is true only in equilibrium; the functioning of the market mechanism is necessary when the economy is not in equilibrium, he stated, to drive the excess demands to zero and thus solve the equation (1954, pp. 248–9; 1889, pp. 242–6). He later declared, however, that it holds in both disequilibrium as well as in equilibrium (1898, pp. 277–8; Walker, 2006a, pp. 152–4). His implicit reasoning appears to be that the *sum* of the excess demand quantities, each multiplied by its price, is zero in disequilibrium even though some or all excess demand quantities are not zero, thus implying that the law is an identity.

Walras asserted regarding his mature comprehensive model (and hence regarding its sub-models) that equilibrium exists, on the grounds that the number of independent equations equals the number of unknowns (prices and quantities). He was, of course, mistaken in that belief. Wilhelm Lexis had pointed out in 1881 that Walras's equations might nevertheless not have real positive solutions or any solution at all (see Walras, 1965, vol. 1, p. 747), a fact that only since the late 1920s and early 1930s became well-known (see Weintraub, 1983; Van Daal, Henderiks and Vorst, 1985).

The interdependence of markets, Walras explained, gives rise to the major problem of general equilibrium analysis, which is the question of the stability of the model, implicitly containing the question of the existence of equilibrium. Will a system of freely competitive markets that is initially in disequilibrium converge to a position of equilibrium? After any market reaches temporary equilibrium through the exchange of the equalized market supply and demand quantities, the traders note what has happened to the prices in other markets. Their reaction is manifested in a shift of the market demand curve, which puts the market once more into disequilibrium and initiates another series of quoted prices leading to a new market-day equilibrium. Will its readjustment aid or impede the equilibrating process taking place in other markets? Does the series of market-day prices in the set of markets move closer to an equilibrium of the entire system or further from it? Walras claimed that he had shown that the answer to those questions is that the market system converges to general equilibrium as a result of the ways that markets are interrelated and of the operation of the Walrasian pricing rule in each market (Walras, 1954, pp. 172, 179–80).

Walras then specified the conditions that prevail in the static equilibrium of exchange of a multi-market system. The ratio of the raretés, or marginal utilities, of any two goods is equal to the ratio of their prices, and the price of any good in terms of another good is equal to the ratio of the prices of those two goods in terms of any third good (1954, p. 157). Those conditions are satisfied when the quantities supplied and demanded of each good are equal (1954, p. 172).

Finally, Walras briefly examined some features of the comparative statics of the exchange model (1954, pp. 147–9). He shifted the utility curves for a good and determined that its equilibrium price changes in the same direction as the shift in the curves. He then successively increased and decreased the traders' endowments of a good and determined that its equilibrium price successively decreases and increases.

Theory of production

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In his model of production, Walras was concerned with the determination of the equilibrium prices of productive services and the equilibrium rates of output of the quantities of non-durable goods. Walras did not present this model directly and without modifications as part of his mature comprehensive model. The latter deals with durable goods as well as non-durables and therefore contains a wider model of production. Nevertheless, Walras carried over the processes of pricing and aspects of the tatonnement in the non-durables model into the comprehensive model.

Setting forth the structure of the production model, Walras first identified the markets for productive services, in which he assumed that the amounts of economic resources and therefore the maximum possible amounts of their services are given. The demanders of productive services are the entrepreneurs. Their ultimate aim is to maximize utility, which they achieve through maximizing profits. In their capacities as entrepreneurs, they combine productive services and materials in proportions that are determined by what Walras called the technical coefficients of production. The coefficients, which he assumed to be fixed in much of his general equilibrium theorization, indicate the amount of each of the inputs that is used to make a unit of output. With fixed coefficients and given prices of the productive services, the average cost is constant as the firm's output varies. If any of those prices change, the average cost curve shifts in the same direction.

The suppliers of productive services are workers, who own personal faculties; landlords, who own natural resources; and capitalists, who own capital goods or provide capital funds (Walras, 1877b, p. 218; 1954, pp. 214–15). Their aim is to maximize utility, which motivates them to offer services to the entrepreneurs in exchange for income.

Walras then identified the market for consumer goods (material and immaterial). These goods (in the production model) are consumed immediately after being produced; they are used only once and are used up in that process. The suppliers of them are the entrepreneurs. The demanders are the workers, landlords, and capitalists acting in their roles as consumers, motivated in their purchases by the desire to maximize utility. They pay for them with the incomes that they have been paid by the entrepreneurs. The only type of capital goods produced in the model is non-durables, that is, variable capital goods like raw materials. Those goods, like consumer goods, are used up in a single application as soon as they are purchased. Of course, that is true of the services of all types of economic resources. In the model, there is no saving. The durable capital goods that are used in production do not depreciate or become obsolete, nor are they subject to accidents. There are no markets for them or for land.

The tatonnement in the production sub-model, and in the capital goods and monetary sub-models, and in the comprehensive model as a whole, is in considerable measure the outcome of the actions of entrepreneurs. Walras assumed that all resources are highly mobile. Entrepreneurs have good knowledge (but not perfect foresight) of the profitability or unprofitability of producing any particular good and accordingly enter or leave an industry. The tatonnement that occurs in the markets for inputs is a process of groping for the equilibrium amounts of resources employed in different industries. The entrepreneurs hire the factors of production, combining them in technologically determined proportions or experimenting to find optimum proportions if the coefficients are variable (1896a, pp. 490–1), and sell services and finished goods to consumers (1954, lesson 21, and pp. 426–7; Walker, 1996, ch. 13). The entrepreneurs hire and use disequilibrium quantities of productive services during the tatonnement, and produce disequilibrium quantities of goods (1889, pp. 234–5, 240–1, 249–50). The payment that the entrepreneurs receive in disequilibrium for their entrepreneurial activity is profit, which Walras defined on a per unit basis as the price of output minus its average cost, with the latter including the wages of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
management. An entrepreneur may undertake the functions of other factors of production – he may also, for example, be a capitalist or a manager of the firm – and ordinarily he has to do so, but his multifaceted role as entrepreneur is a distinct one (1954, p. 222).

The tatonnements in the markets for productive services and for consumer goods are interrelated. If the consumers' demand for a good increases, the price is bid up in accordance with the Walrasian pricing rule. The quantities demanded and supplied become equal at a high price because the supply function does not initially change. The price of the product then exceeds its cost of production, so the entrepreneurs in the industry make profits. Attracted by the prospect of doing the same, other entrepreneurs enter it, and existing firms increase their output. As the market supply of output function changes so that more output would be offered at each possible price, the price in the exchange market for the good is lowered by the entrepreneurs in an effort to dispose of the entire flow of output. As the demands for productive services increase, their prices are bid up, which raises the average cost of production. Thus the price falls and the average cost rises (1954, p. 253). If demand for a good decreases, its price falls below the average cost of production and the entrepreneurs make losses. This leads some of them to leave the industry and some of them to diminish the output of their firms. The prices of the productive services fall as the demand for them decreases, which lowers the average cost of production. As less of a good is offered, its price is forced up. Thus the average cost falls and the price rises (1954, p. 253). It will be noted that these are all non-virtual processes in the model. Walras concluded that the average cost of production and the price of the good become equal, whereupon equilibrium is reached and the tatonnement ends. It follows that in equilibrium the entrepreneur obtains neither profit nor loss (1877b, p. 232; 1954, p. 225).

Pursuing still further the question of the stability of the model, Walras reasoned that the interactions of traders in different markets aid the equilibrating process. The change in the output of a product, he argued, has an effect on its price that is unidirectional, whereas the unidirectional changes that it induces in the outputs of other products have only indirect repercussions on its price, and the latter more or less cancel each other because some of them change in one direction and some in another (1954, p. 246). In the non-virtual adjustment, 'The [resulting] system of new quantities of manufactured products and of new selling prices is thus closer to equilibrium than the old one; and we have only to continue the process of groping to approach still more closely to equilibrium' (1954, pp. 246–7). In other words, once again, ultimately the tatonnement leads to the simultaneous equalization of supply and demand in every market.

The equilibrium, Walras stressed, is the normal state of the market in the sense that it is the one to which the variables in disequilibrium perpetually and automatically tend in a regime of free competition (1954, p. 224). Since it contains the equilibrium of exchange (1954, p. 224), it is characterized by the conditions that the quantities supplied and demanded are simultaneously equal regarding each consumer good, each productive service, and each primary material. A stable circular flow is established in which the total cost equals the total revenue in each firm, the incomes received from the entrepreneurs by the owners of the factors of production equal the revenues earned by the firms, and those incomes are spent on consumer goods by the owners of the factors of production. Walras's theory of production therefore showed ways in which input and output markets for non-durable goods are linked together.

Sales at a disequilibrium price of the items that are produced do not alter the assets held by the participants because the quantities exchanged do not have an existence after the exchange; they are used up immediately. The equilibrium of the production model is therefore not path dependent. That is of no

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

significance, however, because the model is a hypothetical special case. It does not take account of the production of capital goods or of consumer durables which occurs in the real economy and in Walras's mature comprehensive model.
Walras then considered variations in some of the parameters of the production model. If the marginal utility curves for a good shift up, he reasoned, its price in terms of the numeraire increases. If the marginal utility curves shift down, the opposite occurs. If the quantity of a product or service possessed by the holders changes, its price changes in the opposite direction (1954, p. 260).

Theory of capital formation and credit

Walras referred to the three sources of services (labour, land, and capital) as different types of capital because they all endure through time and produce a flow of services, but in this article the unqualified word ‘capital’ or the term ‘capital goods’ will mean durable, man-made, inanimate instruments of production. Walras first examined the determination of the prices of land and personal faculties, as distinct from the prices of their services. The aggregate supply of land, a given condition, is perfectly inelastic, and its price is its gross income divided by the rate of net income (1954, pp. 270, 309). The number of workers is also a given condition. The price of a worker is equal to his gross income minus the cost of replacing and insuring him, divided by the rate of net income. Workers are not bought and sold, however, so their prices are virtual (1954, pp. 271, 311).
Walras's theory of capital is concerned with the determination of the amounts and prices of capital goods, as distinct from their services, and the determination of their rate of net income. Capital goods are specific items of real capital; capital funds, raised by the sale of shares on the bourse, constitute fluid and mobile purchasing power which can be used to acquire economic resources to construct different kinds of physical capital (1954, pp. 270, 311). Capital is formed by capitalists saving money and, most commonly, lending it to entrepreneurs (1954, p. 270). The net saving of the capitalists as a group equals aggregate income minus aggregate consumption, minus the depreciation and insurance costs of capital goods. The entrepreneurs purchase or rent capital goods, earn revenue from their use, and repay any sums they have borrowed (1954, p. 290, §§ 190, 235). Walras's identification of that process explains why he inserted the word ‘credit’ (1954, p. 270) into the name of his capital-goods model, but he did not develop a general theory of credit within it, because he did not introduce loans made by banks. Some capitalists prefer to own capital goods, so Walras assumed occasionally that they acquire them directly in physical form (1954, p. 289), and assumed frequently that they acquire them through buying stock certificates (for example, 1954, p. 289). In each case, the physical capital is used by entrepreneurs, so ‘the demand for new capital goods comes from entrepreneurs who manufacture products and not from capitalists who create savings’ (1954, p. 270). The entrepreneurs purchase the particular kinds of capital goods that are profitable, with the result that the kinds that are produced and used reflect the structure of demands for consumer goods (1954, pp. 225, 303).
Describing the tatonnement regarding both non-durable products and durable goods, that is, summing up the non-virtual character of the tatonnement in the mature comprehensive model, Walras explained that

> After a certain rate of net income and certain prices of services have been cried and after certain quantities of products and new capital goods have been manufactured, if this rate, these prices and these quantities do not satisfy the conditions of general equilibrium, it

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

> will be necessary not only to cry a new rate and new prices, but also to manufacture revised quantities of products and new capital goods. (1889, p. 280; 1954, p. 282; and see 1889, pp. 284–94; 1954, § 258, pp. 293–4)

One aspect of the tatonnement takes place in the stock market and another in the course of the production of capital goods. In the stock market, which is the market for new capital goods, each capitalist attempts to maximize utility by saving and acquiring more stocks that have relatively high yields and less of those with lower yields (1954, p. 289), with the result that the total value of new capital goods and the excess of income over consumption both move in the same direction as all prices. It follows, Walras maintained, that the tendency of the change in prices to destroy the equality between the total value of new capital goods and the excess of income over consumption is weaker than the tendency of the change in the rate of net income to bring the total value of new capital goods and the excess of income over consumption into equality with each other. Moreover, ‘in these conditions, it is evident that the price and the average cost of the capital good (K) will have been little altered from equality by the increases in the quantities produced of the capital goods (K′), (K″)…’ (1889, pp. 292–3). ‘Thus the system involving the new rate of net income and the new prices will be closer to equilibrium than the old system; and it is only necessary to continue the process of groping for the system to move still more closely to equilibrium’ (1954, p. 288).

In the course of the production of capital goods proper, entrepreneurs acquire more capital goods that yield relatively high returns, and diminish their use of capital goods that yield lower returns. As a consequence, the rate of net income from each capital good proper tends toward the same value (1954, p. 309). If profits are being made from the production of capital goods in an industry, new entrepreneurs enter it, and those already in it increase their rate of production. That drives up the prices of productive services, which causes the average cost to rise towards equality with the price of the capital good. If losses are incurred, entrepreneurs diminish production. That drives down the prices of productive services, which causes the average cost to fall towards equality with the price of the capital good (1889, p. 293). It is ‘probable’, or, ‘it is to be presumed’, Walras maintained, that the effects of changes in the output of a new capital good that tend to cause equality between its average cost and its price will be stronger than the contrary effect of interrelated changes in the output of other capital goods, so the process converges to equilibrium (1954, p. 293). Referring again to the non-virtual character of the tatonnement, Walras explained that ‘The new system of revised quantities manufactured and of revised costs of production and selling prices of new capital goods is thus nearer equilibrium than the original system’, and it is only necessary to continue the tatonnement to approach it more and more closely (1889, p. 293; 1954, p. 293).

The tatonnement in the mature comprehensive model involving both non-durable and durable goods represents and explains what happens in the real economy:

> Now, this tatonnement is precisely that which occurs by its own forces on the market for products [non-durables], under the regime of free competition, while the entrepreneurs who produce new capital goods, just like the entrepreneurs who produce products, increase their production or diminish it according to whether profits or losses are made. (1889, p. 294; 1954, p. 294)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Walras expressed the equilibrium conditions in the formation of new capital goods in the lengthy analysis that he called the theorem of the maximum utility of new capital goods, which he described as crowning and confirming his entire theoretical system (Walras to H.S. Foxwell, 16 December 1888, L 859; Walker, 1984b). Although he assumed initially that new capital goods do not require amortization or insurance, Walras then made the realistic assumption that capital goods wear out and are subject to accidents. The rate of net income generated by a particular capital good is then given by the gross income it earns minus amortization and insurance costs, divided by the price of the capital good. In equilibrium each trader maximizes utility by holding the quantities of capital goods that make the ratio of the marginal utility of each capital good to its price equal for all his capital goods. Because of the adjustment of yields and capital good prices, the rate of net income derived from every type of capital good is the same (1954, p. 281). There is a single price for each type of capital good, equal to its average cost (1954, p. 280). Furthermore, the price of any given type of well-maintained old capital goods is equal to the price of the same kind of new capital goods, so the equilibrium prices of all capital goods ‘are equal to the ratios of their net incomes to the rate of net income’ (1954, p. 309). The latter is the rate of interest. Its equilibrium value equates aggregate saving and investment (1954, p. 276).

Walras believed that through this analysis he had seen behind the veil of money or numeraire and discovered the real determinants of that rate. It is manifested in the banking system, he argued, but it is determined in the stock market. It is the ratio of net profit to the price of a share of stock, which in equilibrium equals the common ratio of the net price of capital services to the price of the good that yields them (1954, p. 290).

Walras then turned to the comparative statics of the capital goods market. If the price that has to be paid for the services of a capital good increases or decreases as a result of a parametric change, the price of the capital good itself increases or decreases. Its price also varies inversely with the rate of depreciation and with the rate of the insurance premium. If the rate of net income changes, the prices of all capital goods change in the same direction (1954, pp. 309–12).

Theory of money and circulating capital

Walras wanted to reduce the monetary mechanism ‘to its essential elements’ (1889, p. 379). He therefore carefully specified the structural and behavioural features of his mature fiat money model, drawing upon his direct experience as a young man with real financial matters and his extensive knowledge, accumulated throughout his career, of empirical monetary arrangements and problems. He explained that fiat money, like pieces of paper with an engraved image, has no utility of its own. Economic agents hold it because it enables them to purchase goods that have utility (1889, p. 378). Entrepreneurs and some consumers have net demands for cash balances because of the non-synchronization of payments and receipts (1886, pp. 40–4). The size of a consumer's desired cash balance depends upon desires to consume and save, which depend upon his character and habits, income, the value of the commodities he wants to buy, and the rate of interest (1889, pp. 377, 268–71). The size of an entrepreneur's desired cash balance depends upon the nature and state of his business, and on his character, habits, and the rate of interest (1889, p. 377). These are desired real balances because they represent the demand for a specific bundle of goods. The aggregate real desired cash balance is the nominal one divided by the price level, which is the inverse of the price of money. That aggregate is the demand for future consumer and capital goods (1889, pp. 378–9).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Walras incorporated a market for loans into his model. First, there are the demanders. They are consumers, entrepreneurs and savers, who go to the market and borrow. The first two groups buy the goods and services they need. A curious temporary assumption is that savers, the third group, borrow to obtain the funds they lend. Thus ‘we have the daily demand for money which is exercised on the market for money capital’ (1889, p. 381). What is happening behind the veil of money, Walras explained, is that the consumers and entrepreneurs are actually borrowing the fixed and circulating capital on which they spend the money, and the interest borrowers pay is generated by the fixed and circulating capital that the borrowed money finances. Second, there are the suppliers of funds in the loan market. In one group of them are entrepreneurs who have receipts from the previous day from sales to consumers and from sales of new fixed and circulating capital good to other entrepreneurs, and in the other group are landlords, workers and capitalists who have receipts from the previous day from sales of services (1889, p. 381)
Walras then specified how the mechanism of free competition operates in regard to monetary circulation in disequilibrium of the money market. It will be noted that the entire behaviour relating to money in the mature comprehensive model is non-virtual. Mainly explicitly, he indicated the tatonnement that was presented later by the Cambridge cash balances theorists. He created the temporal framework for the period analysis in his model by assuming first that production and consumption ‘extend over all the moments of the entire year’ (1889, p. 316). Workers, capital goods, and money are used up and are simultaneously reproduced and replaced. He then assumed that markets function every day and that different types of related behaviour occur sequentially on a series of days (1889, pp. 381–3). Thus he developed a continuous-production and periodic-market model.
A change, say a decrease, in the quantity of money causes disequilibrium. The Walrasian period analysis then indicates that the process of equilibration takes three ‘days’. On the day that the decrease occurs, the old rate of interest still rules and the quantity of cash balances demanded exceeds the quantity supplied. The immediate result is that the rate of interest increases. On ‘the next day on the market’, a temporary equilibrium is reached, ‘at a new higher rate of interest at which the desired cash balance would be reduced’ (1889, p. 383). During that day, the prices of all goods fall proportionately to the decrease in the quantity of money and the aggregate *nominal* desired cash balance remains lower, but the real balance is ‘able to become what it was before’ as a result of the fall of prices. Then, ‘on the day after that’, the third day, permanent equilibrium is attained, the rate of interest falling back to its old level, which is equality with the rate of net income from capital goods (1889, p. 383). The same sequence occurs if the aggregate desired cash balance function increases. If the parameters change in the opposite direction, the opposite sequence of adjustments occurs. There are many more details, situations and variations of assumptions that Walras considered regarding the model. He was able to sum up what he had done in this way: if the quantity of money increases or the desired cash balance decreases, prices rise proportionately, and the reverse. ‘This law extends to money the principle by virtue of which value increases with utility and decreases with the quantity’ and therefore integrates the determination of the value of money into his mature comprehensive model and therefore into his general theory of value (1889, p. 383).
When there is monetary equilibrium, that model is in equilibrium in all respects. Walras summarized its aspects in the following way. There is ‘the equilibrium of exchange’ in which prices are proportional to marginal utilities and consumer satisfaction is maximized; there is ‘the equilibrium of production’ with prices equal to average costs and zero profits; there is ‘the equilibrium of capital formation’ with prices of land, human faculties and capital goods proper proportional to the prices of their services; and finally

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
there is ‘the equilibrium of circulation, in view of the fact that the exchangers would have the cash balances that they desire at the announced rate of interest’ (1889, p. 379).

Walras therefore made significant theoretical innovations in his mature theory of money. In it, he raised basic questions about the nature of a true money, its role, the valuation of money, its place in preference functions and hence in demand and supply functions, the sequence of adjustments that occur after a change in its quantity, and the impact upon equilibrium prices and the rate of interest. His modelling of cash balance behaviour and dynamic period analysis anticipated some of the theoretical techniques used during the 1920s and 1930s by J.M. Keynes, D.H. Robertson, and J.R. Hicks, none of whom acknowledged his contributions. That was probably because the insightful analysis and potentially fruitful constructions that Walras put into the mature money model are in the forgotten 1889 edition.

Economic growth

Walras did not develop a complete model of economic growth, but he examined some aspects of the topic in connection with his mature comprehensive model. One was endogenous variations in its parameters. He was led to speculate about that subject by the consideration that the equilibrium conditions identified in the mature comprehensive model are never reached in the real economy because tatonnement takes time, and consequently parameters such as preferences and the amount of labour change before equilibrium is reached (1954, p. 380). In order to take account of this situation, ‘we shall now suppose that the annual production and consumption, which we had hitherto represented as a constant magnitude for every moment of the year under consideration, change from instant to instant along with the basic data of the problem’ (1954, p. 380). Services and goods are used up and are produced. Net new capital goods come into existence and are put to use, and circulating capital is borrowed by entrepreneurs from the capitalists in the form of money loans (1954, p. 379). Walras did not analyse in any further detail that moving equilibrium or ‘continuous market’ economy.

A second aspect of growth that he examined was ‘the laws of the variation of prices in a progressive economy’ (1954, p. 382), that is, some of the features of alternatives paths of economic growth. For this task he first defined economic progress as the substitution of capital services in place of land services in given production functions (1954, p. 383). The substitution implies variable coefficients of production, and to introduce these Walras used the theory of marginal productivity. He did not claim to have originated that theory although he anticipated some of its features. In fact, Hermann Amstein, a mathematician at Lausanne, worked out its major principles in 1877 (Amstein to Walras, 6 January 1877, L 364; translated in Jaffé, 1983, pp. 205–6). Walras did not understand or use Amstein's work, however, and the major credit for the theory of marginal productivity that first appeared in the *Eléments* in 1896 (Appendix III) must be given to Enrico Barone (1895).

Walras defined technical progress as changes in production functions, including the introduction of entirely new processes, but he did not analyse it, beyond concluding that it contributes, along with economic progress, to ensuring that output increases without limit in a progressive economy (Walras, 1954, p. 387). He also discussed, in a highly general way, how the prices of products and services vary with different amounts of capital and sizes of the population (1954, pp. 389–91). His principal conclusion was that the rate of net income falls as the stock of capital grows, the proximate causes of the process being rising rents and falling prices of capital services.

Walras was well aware that capital accumulation means economic growth and requires a different

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
characterization of equilibrium, noting that ‘In order to have a supply, a demand, and prices of capital goods, it is necessary to substitute for the conception of a stationary economic state that of a progressive economic state’ (1889, p. 264). His way of dealing with this problem in 1889 was to assume that new capital goods are not used until what he represented to be the end of the tatonnement, thinking that would preserve the static equilibrium. That did not, however, remedy the problem, as will now be shown.
Walras attempted to give a mathematical proof of the stability of the tatonnement of the mature comprehensive model, spread throughout the pages of his treatise. He believed that he showed that the model is stable by working with his system of static equations of general equilibrium. He posited a disequilibrium state and then varied the prices in the equations in accordance with the Walrasian pricing rule, and claimed that equilibrium is the result of the tatonnement. That claim is logically flawed, for two reasons.
The first is that the tatonnement in the mature comprehensive model, unlike the model of the production of non-durable goods and services, is path dependent even though the new capital goods are not used during any of the phases of adjustment before equilibrium is ostensibly reached. Positive net investment has the result that individual holdings of capital goods and many of the other nominal parameters and all the variables of the model are altered. Each different disequilibrium rate of production and sales of capital goods occurring within each phase of the tatonnement changes their prices and average costs, profits, and the rate of net income. Consequently, Walras's attempted proof was not rigorous and could not have been valid, given the static equations that he used. They have the endowments of assets in the initial disequilibrium as parameters. Their solutions therefore depend on those constants. The model, however, is not a virtual system, so the individual holdings of assets and their total amounts vary during the course of the tatonnement. The result is that the variables of the model do not converge to the solutions of Walras's static equation system. Any equilibrium to which the prices and quantities converge cannot be the one indicated by his equations because they do not describe his model.
The second reason is that the ‘equilibrium’ that Walras asserted exists at the end of the tatonnement is factitious and cannot materialize, even if there were no problem of path dependency. The supposed equilibrium could exist only transitorily while the model is held in a state of arbitrarily suspended animation by the postulate that the additions to the capital stock are not used – a *deus ex machina* that interrupts the incomplete workings of its endogenous processes. The instant that Walras removed the postulate, that is, the instant that the net new capital goods are put into use, the ‘equilibrium’ is ruptured, and through dynamic processes many of the nominal parameters and all the variables of the model change, in the way just indicated in the discussion of the consequences of the use of net new capital goods. If they continue to be produced, as Walras assumed, the system follows a path of growth. The equilibrium path and any stationary equilibrium that the system may eventually reach is quite unlike the solutions to the static equations of general equilibrium that he presented in the 1889 edition of the *Eléments* and subsequently.

4 Walras's last theoretical work

The written pledges sketch

In 1899 Walras changed his work in two major ways, and put the changes into the *Eléments* in 1900. One was to devise a new model of money and circulating capital (see below). The other was to try to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
construct a virtual model. The motive for the latter was that Walras had come to realize by 1899 that his equation system is compatible only with such a model. Instead of trying to develop a different equation system, however, one that would represent the non-virtual mature comprehensive model, he chose to abandon the latter, to retain his static equations, and to try to construct a virtual model that would serve as their foundation and justification. In the 1900 revision, he eliminated much of each of two forceful and lengthy statement of the non-virtual tatonnement (Walras, 1889, pp. 234–5, 280), which consequently appear in Jaffé's translation only in very abbreviated form (1954, pp. 242, 282). He retained, however, crucial parts of those statements and retained elsewhere throughout the revision most of the other language describing the non-virtual behaviour of the economy and of the mature comprehensive model. That explains why the reference ‘Walras, 1954’ can be cited to document many the features of the 1889 mature comprehensive model. It is also one of the principal causes of the 1900 and 1926 editions (and therefore Jaffé's translation of the latter) being a chaotic mixture of incompatible language and sub-models.

To construct a virtual model, Walras conceived the device of written pledges (*engagements écrits*, as they were and are called in the Paris bourse). He asserted that the model has three phases, made identifiable, he believed, ‘by means of the hypothesis of written pledges’. First, there is ‘the phase of preliminary gropings towards the establishment of equilibrium in principle’, the purely virtual phase (1954, p. 319). When a price is cried in any market, suppliers of goods and services write out the amounts that they pledge to produce and sell at that price, but only if it turns out to be the equilibrium value, that is, the one that is part of the set of prices that equates the market supply and demand quantities simultaneously in every market (1899, p. 103; 1900, pp. 215, 260; 1954, pp. 242, 282). Thus Walras adopted the device in order to eliminate changes in the holdings of assets before the entire system of markets has reached equilibrium, changes which would otherwise occur as a result of trade occurring in a market when it reaches market-day equilibrium while other markets are still in disequilibrium, or as a result of disequilibrium production, which changes the aggregate amounts of goods held before general equilibrium has been reached.

In the first phase, entrepreneurs are supposed to plan to move from unprofitable to profitable industries and to plan to create firms or to expand or contract their existing firms, and to predict accurately the financial results of their plans, without actually moving or creating or hiring or spending or producing at all. Owners of productive services are imagined to offer their services repeatedly at disequilibrium prices without actually earning any income or consuming any goods or services. The entire system of interrelated markets is imagined to go through complex costless processes of information acquisition, price changes and changes in the supply quantities that are pledged, all without anyone being allowed to agree to a single actual transaction or undertake any production or consumption, until the equilibrium set of prices has been found.

It is symptomatic of Walras's significantly diminished ability to concentrate and pursue lengthy chains of reasoning after 1898 that his words ‘the hypothesis of written pledges’, although followed by dozens of pages of modelling and theorization, including his immediately following account of the three phases which he introduced into the fourth edition of the *Eléments* at the same time as writing those words, is the last sentence in which Walras mentioned them in any of his writings. His references to written pledges had become fewer and fewer in the successive pages of the 1900 revision. Finally he either forgot about them or decided they were not an idea worth pursuing any further, abandoned his attempts to change the older constructions in his treatise to accord with their use, and introduced new sub-models

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in which they are not used.

Second, there is ‘the static phase in which equilibrium is effectively established *ab ovo* as regards the quantity of productive services and products made available during the period considered, under the stipulated conditions, and without any changes in the data of the problem’ (1954, p. 319). This means that the economy ‘remains [for the time being] *static* because of the fact that the new capital goods play no part in the economy until later in a period subsequent to the one under consideration’ (1954, p. 283). In this postulated static equilibrium, services and non-durable consumer commodities are produced, sold and used. Walras was asserting that the result of the tatonnement in the sketch is that the market supply and demand quantities are equated in every market simultaneously, whereupon the non-virtual activities of exchange, production, consumption, saving and investment take place. He therefore believed that none of the parameters (‘the data of the problem’) of his system of equations of general equilibrium undergoes endogenously induced changes during the tatonnement. He believed that a static equilibrium is consequently the one given by the solutions to his static equation system, and that his new version of tatonnement converges to that equilibrium for the same general reasons as he had adduced in 1889.

Third, continuing to write as though the sketch were a complete functioning system, Walras indicated that it undergoes ‘a dynamic phase in which equilibrium is constantly being disturbed by changes in the data and is constantly being re-established’ (1954, p. 319; 1900, p. 302). The new capital goods, both fixed and circulating, Walras wrote, ‘are made available during the second phase’ but ‘are not put to use until the third phase’. When they are used, however, ‘the first change in the data of our problem’ occurs (1954, p. 319). The result of changes in the data is that the ‘*fixed* equilibrium will then be transformed into a *variable* or *moving* equilibrium, which re-establishes itself automatically as soon as it is disturbed’(1954, p. 318). The use of the new capital goods generates economic growth.

Of course, Walras *asserted* that the three phases and all the behaviours and outcomes that he wanted the sketch to have, such as the equalization of supply and demand, do in fact occur in it, but in actuality the sketch does not support his claims. Those accounts are not descriptions of the sketch. They are simply postulates; they cannot be deduced from its structure and procedures, which is why Walras's work on virtuality is properly described as a sketch rather than a model. He made the mistake of assuming that potential demanders, whether consumers or entrepreneurs, do not write pledges to buy, so they have no way of making their desired demands known. For that reason alone (although there are many others; Walker, 1996, section I), the sketch cannot function. There are no individual or market demand functions, Walrasian pricing, transactions or production. Equilibrium does not exist because the number of unknowns (prices, the quantities of goods and services) exceeds the number of independent equations (the market supply functions). Moreover, the device of written pledges is so inherently flawed that there are no conceivable revisions that can transform it into a functioning system (Walker, 1996, section I). Finally, it is evident that the sketch has no explanatory value in reference to the real economy.

Despite the sketch's shortcomings, its aim of excluding irrevocable disequilibrium behaviour from a general equilibrium model, achieved simply as a postulate, nevertheless became a central and indispensable part of Walras's intellectual legacy for certain of his successors, including Gustav Cassel, Abraham Wald, John von Neumann, Kenneth Arrow, Frank Hahn and Gérard Debreu (Walker, 2006a, pp. 288–312). It is a pity that they adopted his virtual notion and were unaware of or disregarded his robust and more realistic mature comprehensive model, for through the development of many of its characteristics lies the way to a more useful general equilibrium theory, as recent research has shown (Walker, 2006a, pp. 334–6; 2006b).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Commodity E in the last model of the production and pricing of capital goods

In 1900 Walras invented a fictional good (E) constituted of perpetual annuity shares, another example of the deterioration of the quality of his modelling after 1899. It was by means of that concept that he chose to deal with a positive, zero or negative excess of aggregate income over aggregate consumption in his capital-goods model. It appears that he wished to express aggregate savings as a single homogenous quantity – the demand for E. Each perpetual share pays one unit of numeraire per year, and its price, determined by supply and demand, is the reciprocal of the rate of interest (Walras, 1954, pp. 274–6, 308–9). If people want an additional amount of interest income, they provide savings through purchasing new perpetual shares. The numeraire-capital that capitalists pay for units of E is used by entrepreneurs to buy productive services and materials that are transformed into new capital goods). Walras viewed the capital-goods markets as reaching equilibrium through adjustments of the goods' rates of return (1954, pp. 275–6, 308).

Walras's device of the E commodity, which he frankly described as imaginary (1954, p. 274), has not been adopted by economists who succeeded him, because of its remoteness from the realities of capital accumulation and the distortions that it creates in a model of that process. He did not realize that it is incompatible with the written pledges procedure. If the latter is assumed to occur, the capitalists have no way of expressing their demand for E. Moreover, as it happens, Walras retained his references to the purchasing of stock certificates and private and government bonds which appear in the mature capital goods model (for example, 1954, §§ 255, 270), and he introduced short-term loans into his new money model, although the incompatibility of E with those financial instruments increases the incoherence of the last two editions of the *Eléments*. His treatment of saving, investment and the capital goods market in the mature comprehensive model is superior to his final thoughts on the subject.

Walras's last model of money and circulating capital

In his revision of the *Eléments* in 1900, Walras stated that he wanted to design the structure of this model on 'exactly the same terms and in precisely the same way' as in the 1900 models of exchange, production and capital formation (1900, p. 42). He did not in fact do that because he mentioned written pledges only twice in the first half of the exposition of the model and not at all in the second half of it. In fact, he constructed the new model of money and real circulating capital before he thought of written pledges. He first presented it in an article in 1899. After the last page of the article, he added an afterthought, a note of 37 lines introducing the device of written pledges (1899, p. 103). He subsequently inserted most of the note into the text of the article, and inserted the result almost verbatim into the 1900 edition of the *Eléments* (1954, lessons 29, 30), completely eliminating the treatment of money that he had presented in the mature comprehensive model.

In contrast with his mature comprehensive model, Walras described the functions of money and the formation of money prices in his last model of general equilibrium on the assumption that there is no uncertainty in equilibrium, and consequently that the dates and monetary value of future purchases and sales are known (1954, pp. 317–18). Money is one type of circulating capital, he explained; the other is circulating physical capital. Replacing his formulation of an equation of exchange that had anticipated Irving Fisher's (1877b, pp. 180–81), Walras asserted that circulating physical capital yields utility from

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
its ‘service of availability’ – that is, by being readily available – and money provides, by proxy, the same service of availability as the good that it is destined to purchase and yields the same utility as that service. All economic agents try to hold utility-maximizing amounts of money and circulating physical capital (1954, pp. 320–1). The latter, held by consumers and entrepreneurs, is acquired with money, so the essential concern of Walras's model of circulating capital reduces to the question of the demand for and supply of money and its price.

Savers make some of their balances available as loans through buying or perpetual annuity shares (1954, pp. 318–20). The aggregate gross supply of money is the total stock issued by the monetary authority in the case of a fiat money economy, and is the amount of circulating coin in the case of a commodity-money economy (1954, pp. 321–4). The price of money, the numeraire, is unity and the price of its service is the rate of interest (1954, pp. 320, 327). Given the flows of receipts and purchases, the individual gross and net demand for cash balances and the individual net supply of them are functions of the rate of interest. The sum of the individual net demands for money is the aggregate demand function, and the sum of the individual net supplies of money is the aggregate supply function (1954, pp. 320–1). The tatonnement in the money market, Walras contended, explains how the rate of interest and the equilibrium aggregate net quantities of cash balances supplied and demanded are determined (1954, pp. 325, 327). The rate of interest changes according to the Walrasian pricing rule. When the excess quantity demanded of cash balances is positive, the rise in the rate decreases the quantity demanded of cash balances by consumers and entrepreneurs by increasing the cost of the service of availability that money provides, and also decreases the quantity demanded by entrepreneurs by causing a fall in profits and hence in the desired rate of production. The rise in the rate of interest also causes the net quantity of cash balances that savers want to supply to increase. If the desired supply of cash exceeds the desired demand at the current rate of interest, the opposite effects occur (1954, p. 333). The tatonnement continues until the equilibrium price of the service of availability of money is found – namely, the price that equates the net and therefore the gross quantities supplied and demanded of cash balances – whereupon the money market reaches equilibrium (1899, p. 96; 1900, pp. 297–319; 1954, pp. 315–33).

The equilibrium prices of all goods in terms of money are given by its role as the numeraire and by the workings of the entire model that determine the ratio of exchange between each good and the numeraire. In general equilibrium, the price of the service of all money held by different individuals for different purposes is the same (1954, p. 326). Moreover, because an underlying influence upon the rate of interest on money is the value productivity of physical capital, an influence exerted through variations in the volume of funds invested, the equilibrium rate on money is the same as the equilibrium rate of net income determined in the market for capital. There is therefore equality in the rate of net income from all capital goods and real and monetary circulating capital (1954, p. 323).

Walras then considered the comparative statics of the model. He changed the utility functions for the service of money and deduced that the marginal utility and value of the service of money changes in the same direction. He changed the quantity of money and deduced that the marginal utility and value of the service of money changes in the opposite direction, and that all prices change in the same direction without any alterations in relative prices (1954, p. 333). He noted that, if the utility curves for net income shift up or down, the rate of net income changes in the opposite direction. If the quantity of net income varies, the rate of net income varies in the same direction. If utility functions and the quantity of net income both vary in such a way that the marginal utilities remain unchanged, the rate of net income also remains unchanged (1954, p. 307).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
An aura of unreality is imparted to Walras's 1900 edition of the *Eléments* by his abstracting from money through much of his exposition of exchange, production and capital formation, and then by introducing it in such a way that it does not change their characteristics (1954, pp. 319–24). In particular, by postulating that there is no uncertainty in his last model of money and circulating capital, without the slightest explanation of how that would be possible, he eliminated consideration of much of the behaviour associated with money in the real economy. Money does, however, influence a great deal of real economic behaviour in special ways associated with uncertainty, a fact of which Walras's extensive writings on real monetary arrangements, problems and policies reveal him to have been perfectly cognizant. Moreover, his concept of fictional perpetual annuity shares is a superfluity that further detracts from the verisimilitude of his models of capital formation and money. He should instead have retained his mature model of the money market in which he dealt with the behaviour related to some of the major financial assets in which people actually invest.

5 Economic policies

Walras developed all his policy proposals during the years prior to his last theoretical efforts. He never mentioned the latter formulations in connection with real empirical matters. In particular, the written pledges sketch did not influence him to modify or innovate policy proposals, necessarily so because the functioning and hence the problems of the economy are not virtual.

Walras was greatly interested in the economic problems of his day and in socio-economic reform, guided in his major policy proposals by his normative convictions, which were derived from his father's philosophy of society and justice. Those convictions were a mixture of conventional 19th-century liberalism and notions of the rightness and efficacy of state interventionism (1896b). Like many scholars, each with different views, Walras bestowed the title of ‘natural law’ upon the principles of justice that he considered desirable, and so he might be called a natural-law philosopher or casuist. Nevertheless, he was not a natural-law economist. He did not believe that there is, behind observable facts, a structure of economic laws that are divinely ordained, or that are peculiarly in tune with the structure of the universe and human aspirations, or are ceaselessly at work so that violations of them can only result in chaos or frictions. Nor did he construct his economic model with the conscious intention of expressing his normative views. Sharply distinguishing normative and positive economics, he stated that he designed his theories for the purpose of understanding economic reality (Walker, 1984a; 2006a) and presented his normative work explicitly as such and carefully segregated it (Walras, 1896b) from the publications presenting his economic theories.

Walras's policy recommendations ranged over natural monopolies, which he believed should be nationalized; prices, which he believed should be stabilized by a monetary authority; bimetallism, which he believed had both advantages and disadvantages; the stock market, which he believed should be regulated by the state in order to improve its organization and ensure its integrity; taxes, which he believed were unjust and confiscatory and should be abolished; and land, which he believed should be purchased by the state and rented to private users, thereby providing it with revenue (1905b, pp. 272–3). Arguing that his advocacy of nationalization of land and natural monopolies was based upon a scientific understanding of the functioning of the economy, Walras called himself a ‘scientific socialist’.

6 Contributions

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Criticisms of Walras's work cannot obscure the greatness of his contributions. When he began his investigations in 1868, economics on the Continent was hardly a scientific pursuit but rather a mixture of normative prescriptions, classical theories expressed alongside protectionist doctrines, and commercial law. In England it was in the state exemplified by the work of J.S. Mill – with much that could be used as a basis for future investigations, but also without a clear view of the relationships of distribution and production, limited by a cost-of-production theory of value, and lacking a theory of supply and demand in multiple markets. The attitude of most of Walras's contemporaries was that, since economic behaviour involves preferences and the human will, it cannot be expressed in a rigid and deterministic set of algebraic relations. Walras changed all that, transforming economics and propelling it forward in a gigantic intellectual leap.

His contribution can be divided into two interrelated parts. One is that, in his mature comprehensive model, he constructed or refined or adapted to his purposes many of what became the fundamental building blocks of modern economic theory. In this effort he accomplished an enormous amount of highly creative economic analysis, brilliantly analysing the structure of economic reality to bring many of its essential features into clear relief, in eight major original contributions. First, he went far beyond the work of the other developers of the marginal utility theory by using it to analyse the disequilibrium and equilibrium behaviour of a variety of participants undertaking different economic functions in multiple markets, rather than confining the theory to the investigation of consumer demand and of exchange in an isolated market. Second, he had clear priority in constructing the theory of exchange in multiple competitive markets. In that regard, his work was greatly in advance of his predecessors' and was replete with fruitful constructions, theorems and postulates, like the reciprocal relation of supply and demand, the device of a numeraire, the individual budget equation, Walras's Law, the theorem of equivalent distributions, and the laws of change of prices. Third, he constructed a theory of the firm and of market supply in which he appears to have developed independently the modern idea of a firm's production function, derived the equation for a firm's average cost, expressed the firm's offer of output mathematically, and aggregated the firm's supply functions to obtain the market supply in a particular industry. Fourth, he was the first to examine the question of the existence of equilibrium in a competitive multi-market system of exchange and production. Fifth, in his work on tatonnement he initiated the study of the stability of competitive general equilibrium and contributed significantly to its understanding, with his most successful theorizing on the topic relating to his mature comprehensive model. There is nothing in the literature before Walras's time or until the time that his work was used by others that is even remotely like or on the level of his reasoning regarding the process of convergence to equilibrium of a non-virtual competitive multi-market system. Sixth, he developed a theory of the entrepreneur, of profits, and of the allocation of resources that became the basis of Continental work on those topics (Pareto, 1896–7, passim; Pareto, 1906, passim; Barone, 1896, p. 145; Schumpeter, 1912/1926, p. 76; Schumpeter, 1954, p. 893; Walker, 1986). Seventh, Walras created a fruitful theory of capital, achieving an early formulation of the conditions for a Pareto optimum in capital markets. As in a number of his other investigations, his characteristic contribution in that regard was not to be the first to think of the problem but to be the first to offer an account of those markets' disequilibrium interrelationships and equilibrium conditions in a model of the general equilibrium of an economy. Eighth, he developed a cash-balances theory of money in his period of maturity as a theoretician which had great originality and has stimulated much valuable research (Marget, 1931; 1935; and see Walker,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1970, p. 696; 1996, pp. 235–55). Those eight areas of analysis were the core of neoclassical microeconomic theory and thus constituted much of the structure of knowledge that was the starting place for 20th-century economics.

The second part of Walras's contribution was not the idea of the general equilibrium of a freely competitive economic system, which other economists had suggested; it was his implementation of that idea. Other economists had helped in fashioning the building blocks that Walras used. His achievement, however, was not only to develop them but also, through incisive analysis and constructive thought, to weave them into an account of the equilibrating processes of a complex, non-virtual, multi-market economy. Those building blocks dealt with real economic behaviour, and it is his use of building blocks with that essential quality that gives his work its richness, robustness and relevance. Walras was also the first economist to try to set up a system of equations to describe the conditions of static equilibrium of a general equilibrium model.

Walras thus accomplished by the mid-1870s far more than any other economist had done in regard to constructing a model of an economic system as a whole, and more single-handedly in that regard than any other economist in the history of the discipline.

7 Influence

Walras's work was not given the recognition that it merited in France during the 25 years after 1874, and his centennial, in 1934, elicited no conference on his work there. By the 1950s, however, the French attitude toward Walras had changed, as was ultimately symbolized by the creation in 1984 of the Centre Auguste et Léon Walras (but no longer symbolized in that way, for the research group has a new name; it is now a part of the organization known as ‘TRIANGLE, Unité mixte de Recherche 5206 du Centre national de la Recherche scientifique’) at the université Lumière Lyon 2. With the English, Walras's experience was also disappointing. His initial cordiality towards W.S. Jevons, as a fellow pioneer in mathematical economics, was dissipated by Jevons's failure to recognize Walras's contributions to the theory of exchange and to the construction of a relatively complete model of a competitive economy, and eventually Walras, quite unreasonably, came to regard Jevons as a plagiarist of his work (Walras to M. Pantaleoni, 17 August 1889, L 909). Similarly, Walras's relations with P.H. Wicksteed began well (Wicksteed to Walras, 1 December 1884, L 619) but deteriorated sharply when Wicksteed failed to give credit to those whom Walras considered to be the true originators of the theory of marginal productivity (Walras, 1965, L 1220, n. 3; 1896a, pp. 490–2). Walras justly felt neglected by Alfred Marshall, who mentioned him only thrice in the briefest of comments in the *Principles* (Marshall, 1890; 1920) and wrote not a word about Walras's development of general equilibrium theory. Walras also came to dislike Edgeworth for criticizing his theories of tatonnement, capital goods and the entrepreneur (Walras to Gide, 3 November 1889, L 933, and 11 April 1891, L 1000; Walras to Pantaleoni, 5 January 1890, L 953). In general, Walras believed, the English had closed their minds to his theories and had become spiteful in their treatment of them (see Walker, 1970, pp. 699–70).

The extremity of the language with which Walras characterized the English was unjustified, because, although he had reason for disappointment with their neglect of his general equilibrium theory, Jevons (1879, preface) and Edgeworth (1889) had recognized valuable elements in his work, and he was the only living economist included in the first edition of Palgrave's *Dictionary of Political Economy*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
(Sanger, 1899). The fact is that Walras grew hypersensitive about the motives of his critics, the failure of the majority of economists to recognize the value and priority of his contributions, and the possibility of plagiarism of his ideas during the 1880s and 1890s. There had been two periods in his life, he complained, ‘one during which I was a madman, and one during which everyone made my discoveries before me’ (Walras, undated, in Jaffé, 1983, p. 203, n.54).

This account of Walras's disappointments should be balanced by a realization that his scientific labours had afforded him ‘up to a certain point, pleasures and joys like those that religion provides to the faithful’ (Walras to Marie de Sainte Beuve, 15 December 1899, L 1432), and a recognition of the professional satisfactions that he increasingly experienced in the last two decades of his life. Maffeo Pantaleoni (1889), Enrico Barone (1895, in Jaffé, 1983, p. 186; 1896), and Vilfredo Pareto (Pareto to Walras, 15 October 1892, L 1077) contributed greatly toward giving Walras's work a secure place in Continental economics and thus ultimately in economics everywhere. In 1895•Pareto's appointment as Walras's successor to the chair of economics at Lausanne assured Walras that his doctrines, expressed in his treatment of a non-virtual competitive economy, would be perpetuated and developed, and the accessible literary presentations of Walras's ideas in Pareto's books (1896–7; 1906) began their widespread dissemination. Pareto borrowed most of the ideas of Walras that have been mentioned in this article, using them as the basis for his contributions to the theories of non-virtual general equilibrium, the monopolistic entrepreneur, and welfare economics. Wilhelm Lexis, Ladislaus von Bortkiewicz and Eugen von Böhm-Bawerk gave Walras's models serious attention. Knut Wicksell based his theory of price determination squarely upon Walras's work (Wicksell to Walras, 6 November 1893, L 1168), and Karl Gustav Cassel was inspired in the construction of his models (1903; 1918) by Walras's equation system and idea of virtuality expressed in the written pledges sketch. Walras was given recognition in the United States: in 1892 he was made an honorary member of the American Economic Association, Irving Fisher praised his work (Fisher, 1892, p. 45; 1896), H.L. Moore became his avowed disciple and explicator (Moore to Walras, 19 May 1909, enclosure to L 1747; Moore, 1929), and Henry Schultz taught Walras's economics and, like Moore, undertook theoretical and econometric studies of general equilibrium relationships (for example, Schultz, 1929; 1932; 1933). His mature comprehensive model was the starting point for the work of Joseph Schumpeter, who, throughout his entire career, devoted himself to the study of disequilibrium and general equilibration of non-virtual economic phenomena. The manifestations of acceptance led Walras to believe he would ultimately triumph, and that enabled him to achieve a mental calmness (Walras to Marie de Sainte Beuve, 15 December 1899, L 1432; Walras to A. Aupetit, 28 May 1901, L 1485). ‘Be assured of my serenity’, he wrote to old friends in 1904, ‘I have not the least doubt about the future of my method and even of my doctrine; but I know that success of this sort does not become clearly apparent until after the death of the author’ (Walras to G. and L. Renard, 4 June 1904, L 1574). Walras's prediction of great success was accurate. An indication of what the future would hold for his theories was given by the celebration of his jubilee in 1909 by the University of Lausanne, in the course of which he was honoured as the first economist to establish the conditions of general equilibrium, thus founding the School of Lausanne (Walras, 1965, L 1696, n. 5). His achievements were praised in a statement signed by 15 leading French scholars, including Charles Gide, Charles Rist, Georges Renard, Alfred Bonnet, Albert Aupetit and François Simiand (Walras, 1965, enclosure to L 1747), and in communications from many others (Pareto to the Dean of the Faculty of Law of the University of Lausanne, 6 June 1909, L 1755; Schumpeter to Walras, 7 June 1909, L 1756).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Walras's contributions inspired and provided a substantial beginning for the branches of general equilibrium theory and applications as they have developed since the 19th century. Indeed, the filiations of his ideas have become so numerous and dense as to be an integral and central part of the mainstream of modern economics. His achievement of developing particular theories and binding them together in a model of an entire economic system has given his work an influence on the verbal, mathematical and econometric study of the interrelationships of all parts of economic systems that has been durable and immense. For sheer genius and intuitive power in penetrating the veil of the chaos of immediately perceived experience and divining the underlying structure of fundamental economic relationships and their extensive interdependencies and consequences, Walras has been surpassed by no one.

See Also

- general equilibrium
- tâtonnement and recontracting

Selected works

1858. *Francis Sauveur*. Paris: E. Dentu.

1860. *L'Economie politique et la justice*. Paris: Guillaumin.

1868. *Recherche de l'idéal social: Leçons publiques faites à Paris*. Paris: Guillaumin; and in Walras (1896a).

1869–70. 2e tentative, 1869–1870. In Walras (1965), vol. 1, 218–19.

1871. *Discours d'installation. Séance académique du 20 Octobre*. Lausanne: Académie de Lausanne.

1877a. *Théorie mathématique de la richesse sociale: quatre mémoires*. Paris: Guillaumin.

1874, 1877b. *Eléments d'économie politique pure; ou théorie de la richesse sociale*. 2 parts. Lausanne: L. Corbaz; Paris: Guillaumin; Bâsle: H. Georg.

1880a, 1880b. La bourse, la spéculation et l'agiotage. *Bibliothéque Universelle et Revue Suisse* 5, March 1880a, 452–76; 6, April 1880b, 66–94.

1886. *Théorie de la monnaie*. Lausanne: Corbaz.

1889. *Eléments d'économie politique pure*, 2nd edn. Lausanne: F. Rouge; Paris: Guillaumin; Leipzig: Duncker & Humblot.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1893, 1904, 1909. Notice autobiographique. In Walras (1965), vol. 1, 1–15.

1895. Enclosure by Léon Walras to Vilfredo Pareto. In Walras (1965), vol. 2, 628–32.

1896a. *Eléments d'économie politique pure*, 3rd edn. Lausanne: F. Rouge; Paris: F. Pichon; Leipzig: Duncker & Humblot.

1896b. *Etudes d'économie sociale (Théorie de la réparation de la richesse sociale)*. Lausanne: F. Rouge; Paris: F. Pichon.

1898. *Etudes d'économie politique appliquée (Théorie de la production de la richesse sociale)*. Lausanne: F. Rouge; Paris: F. Pichon.

1899. Equations de la circulation. *Bulletin de la Société Vaudoise des Sciences Naturelles*, June, 85–102 and note, 103.

1900. *Eléments d'économie politique pure*, 4th edn. Lausanne: F. Rouge; Paris: F. Pichon.

1905a. Cournot et l'économique mathématique. *Gazette de Lausanne*, 13 July; and, except for a few deletions, in *Revue d'Economie Politique*, January–February, 1962.

1905b. Draft of a recommendation for a Nobel Peace Prize. In Walras (1965), vol. 3, 270–4.

1926. *Eléments d'économie politique pure*. Definitive edn. Paris: R. Picon & R. Durand-Auzias; Lausanne: F. Rouge.

1954. *Elements of Pure Economics*, trans. and annotated by William Jaffé. Homewood, IL: Richard D. Irwin; London: Allen & Unwin.

1965. *Correspondence of Léon Walras and Related Papers*, 3 vols. ed. W. Jaffé. North-Holland: Amsterdam.

Bibliography

Barone, E. 1895. On a recent book by Wicksteed. Manuscript translated into English by Walker in Jaffé (1983), 182–6.

Barone, E. 1896. Studi sulla distribuzione. *Giornale degli Economisti* 12, February, 107–55, 235–52.

Bertrand, J. 1883. Théorie des richesses. *Journal des Savants* 67(September), 499–508.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Cassel, G. 1903. *Nature and Necessity of Interest*. London: Macmillan.

Cassel, G. 1918. *Theory of Social Economy*. New York: Harcourt, 1932.

Cournot, A.A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Edgeworth, F.Y. 1889. The mathematical theory of political economy. [Review of] *Eléments d'économie politique pure* by Léon Walras, *Nature* 40, 434–6.

Fisher, I. 1892. Translator's note to Léon Walras, Geometrical theory of the determination of prices. *Annals of the American Academy of Political and Social Science* 3(July), 45–7.

Fisher, I. 1896. [Review of] *Eléments d'économie politique pure. Troisième edition. Par Léon Walras*.… [and of] *An Essay on the Co-ordination of the Laws of Distribution. By Philip H Wicksteed*…. *Yale Review* 5(1), 222–3.

Hicks, J.R. 1934. Léon Walras. *Econometrica* 2, 338–48.

Isnard, A.N. 1781. *Traité des richesses, contenant l'analyse de l'usage des richesses en général et de leurs valeurs*. London and Lausanne: Grasset.

Jaffé, W. 1964. New light on an old quarrel: Barone's unpublished review of Wicksteed's Essay on the coordination of the laws of distribution and related documents. *Cahiers Vilfredo Pareto* 3, 61–102; where necessary translated into English by Walker in Jaffé (1983).

Jaffé, W. 1969. A.N. Isnard, progenitor of the Walrasian general equilibrium model. *History of Political Economy* 1, 19–43; and in Jaffé (1983).

Jaffé, W. 1971. Reflections on the importance of Léon Walras. In the P. Hennipman festschrift, *Schaarste en Welvaart*, ed. A. Heertje et al. Amsterdam: Stenfert Kroese; and in Jaffé (1983).

Jaffé, W. 1981. Another look at Léon Walras's theory of *tâtonnement*. *History of Political Economy* 13, 313–36; and in Jaffé (1983).

Jaffé, W. 1983. *William Jaffé's Essays on Walras*, ed. D.A. Walker. Cambridge, New York: Cambridge University Press.

Jevons, W.S. 1879. *The Theory of Political Economy*, 2nd edn. London: Macmillan.

Klotz, G. 1994. Achylle Nicolas Isnard, précurseur de Léon Walras? *Economies et Sociétés*, Série

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Œconomia, Histoire de la Pensée économique 28(10–11), 29–52.

Lexis, W. 1881. Zur mathematisch-ökonomischen Literatur. *Jahrbücher für Nationalökonomie und Statistik* 3, n.s. 427–34.

Marget, A.W. 1931. Léon Walras and the ‘cash-balance approach’ to the problem of the value of money. *Journal of Political Economy* 39, 569–600.

Marget, A.W. 1935. The monetary aspects of the Walrasian system. *Journal of Political Economy* 43, 145–86.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan. 8th edn, 1920.

Moore, H.L. 1929. *Synthetic Economics*. New York: Macmillan.

Pantaleoni, M. 1889. *Principii de economia pura*. Florence: Barbèra.

Pareto, V. 1896–7. *Cours d’économie politique*. 2 vols. Lausanne: F. Rouge. Ed. G.-H. Bousquet and G. Busino, Geneva: Librairie Droz, 1964.

Pareto, V. 1906. *Manuale di economia politica*. Milan: Societá Editrice Libraria. Trans. A. Bonnet from the Italian as *Manuel d’économie politique*. Paris: V. Giard and E. Brière, 1909.

Poinsot, L. 1803. *Eléments de statique*. 8th edn. Paris: Bachelier, 1842.

Sanger, C.P. 1899. Walras, Marie Esprit Léon. In *Dictionary of Political Economy*, vol. 3, ed. R.H. Inglis Palgrave. London: Macmillan.

Say, J.-B. 1836. *Cours complet d’économie politique pratique, suivi des Melanges, Correspondance et Catéchisme d’économie politique*. 3rd. edn. Bruxelles: H. Dumont.

Schumpeter, J.A. 1912, 1926. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Trans. R. Opie, New York: Oxford University Press, 1961.

Schumpeter, J.H. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Schultz, H. 1929. Marginal productivity and the general pricing process, *Journal of Political Economy* 37, 505–51.

Schultz, H. 1932. Marginal productivity and the Lausanne School. *Economica* 12, 285–96.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Schultz, H. 1933. Interrelations of demand. *Journal of Political Economy* 41, 468–512.

Van Daal, J., Henderiks, R.E.D. and Vorst, A.C.F. 1985. On Walras' model of general economic equilibrium. *Zeitschrift für Nationalökonomie* 45(3), 219–44.

Walker, D.A. 1970. Léon Walras in the light of his correspondence and related papers. *Journal of Political Economy* 78(4), Part I, 685–701.

Walker, D.A. 1984a. Is Walras's theory of general equilibrium a normative scheme? *History of Political Economy* 16, 445–69.

Walker, D.A. 1984b. Walras and his critics on the maximum utility of new capital goods. *History of Political Economy* 16, 529–54.

Walker, D.A. 1986. Walras's theory of the entrepreneur. *De Economist* 134(1), 1–24.

Walker, D.A. 1996. *Walras's Market Models*. Cambridge, New York, Melbourne: Cambridge University Press.

Walker, D.A. 2000. La relation entre la Bourse au XIXe siècle et le modèle de Léon Walras d'un marché organisé. In *Les traditions économiques françaises 1848–1939*, ed. P. Dockès et al. Paris: Éditions du CNRS.

Walker, D.A. 2001. A factual account of the functioning of the nineteenth-century Paris Bourse. *European Journal of the History of Economic Thought* 8, 186–207.

Walker, D.A. 2006a. *Walrasian Economics*. Cambridge, New York: Cambridge University Press.

Walker, D.A. 2006b. Virtualité et non virtualité dans les modèles de l'équilibre général. Colloque Walras, Association Internationale Walras, University of Lausanne, 22 September 22.

Walras, A.A. 1831. *De la nature de la richesse et de l'origine de la valeur*. Paris: Johanneau.

Weintraub, E.R. 1983. On the existence of a competitive equilibrium: 1930–1954. *Journal of Economic Literature* 21, 1–39.

How to cite this article

Walker, Donald A. "Walras, Léon (1834–1910)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000019>
doi:10.1057/9780230226203.1814

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Walras's Law

Don Patinkin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Walras's Law is an expression of the interdependence among the excess-demand equations of a general-equilibrium system that stems from the budget constraint. It is so called because Walras made use of this interdependence from the first edition of his *Eléments d'économie politique pure* through the fourth, which for all practical purposes is identical with the definitive edition. Walras presents the argument for an exchange economy and extends his analysis to deal first with a simple production economy and then with one in which capital formation also takes place. Here, Walras's Law is derived in a more general way than usual.

Keywords

barter; dual decision hypothesis; excess demand and supply; general equilibrium; Hicks, J. R.; Keynes, J. M.; kinked demand curve; kinked supply curve; Lange, O. R.; liquidity preference; loanable funds; Modigliani, F.; Patinkin, D.; perfect competition; price rigidity; real-balance effect; Ricardo, D.; Say, J.-B.; Say's Law; tâtonnement; wage rigidity; Walras's Law

Article

Walras's Law (so named by Lange, 1942) is an expression of the interdependence among the excess-demand equations of a general-equilibrium system that stems from the budget constraint. Its name reflects the fact that Walras, the father of general-equilibrium economics, himself made use of this interdependence from the first edition of his *Eléments d'économie politique pure* (1874, §122) through the fourth (1900, §116), which edition is for all practical purposes identical with the definitive one (1926). I have cited §116 of this edition because it is the one cited by Lange (1942, p. 51, n. 2), though in a broader context than Walras's own discussion there (see below). In this section, Walras presents the argument for an exchange economy. In accordance with his usual expository technique (cf. his treatment of the *tâtonnement*), he repeats the argument as he successively extends his analysis to deal first with a simple production economy and then with one in which capital formation also takes place (ibid., §§206 and 250, respectively).

For reasons that will become clear later, I shall derive Walras's Law in a more general – and more cumbersome – way than it usually has been. Basically, however, the derivation follows that of Arrow and Hahn (1971, pp. 17–21), with an admixture of Lange (1942) and Patinkin (1956, chs I–III and Mathematical Appendix 3:a).

Let x_i^h be the decision of household *h* with respect to good $i (i = 1, \ldots, n)$, where 'goods' also include services and financial assets (securities and money). If $x_i^h \geq 0$, it is a good purchased by the household; if $x_i^h < 0$, it is a good (mainly, labour or some other factor-service) sold. Similarly, let y_i^f be the decision of firm *f* with respect to good *i*; if $y_i^f \geq 0$, it is a good produced and sold by the firm (i.e., a product-output); if $y_i^f < 0$, it is a factor-input.

Assume that firm *f* has certain initial conditions (say, quantities of fixed factors of production) represented by the vector $\mathbf{k}^f$ and operates in accordance with a certain production function. Following Patinkin (1956), let us conduct the conceptual individual-experiment of confronting the firm with the vector of variables **v** (the nature of which will be discussed below) while keeping $\mathbf{k}^f$ constant and asking it to designate (subject to its production function) the amounts that it will sell or buy of the various goods and services. By repeating this conceptual experiment with different values of the respective elements of **v**, we obtain the behaviour functions of firm *f*,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$y_i^f = y_i^f(\mathbf{v}; \mathbf{k}^f)\ (i = 1, \ldots, n). \tag{1}$$

For $y_i \geq 0$, this is a supply function; for $y_i < 0$, it is a demand function for the services of factors of production. Profits (positive or negative) of firm *f* are then

$$R^f = \sum_i p_i y_i^f\left(\mathbf{v}; \mathbf{k}^f\right), \tag{2}$$

Let d^{hf} represent the proportion of the profits of firm *f* received by the household *h*. Its total profits received are then $\Sigma_f d^{hf} R^f$ and its budget constraint is accordingly

$$\sum_i p_i x_i^h = \sum_f d^{hf} R^f, \tag{3}$$

which assumes that households correctly estimate the profits of firms (cf. Buiter, 1980, p. 7; I shall return to this point below). As with the firm, let us, *mutatis mutandis*, conduct individual-experiments with household *h* (with its given tastes), subject to its budget constraint (3) by varying the elements of **v**, while keeping its initial endowment (represented by the vector $\mathbf{e}^h$ constant. This yields the behaviour functions

$$x_i^h = x_i^h\left(\mathbf{v}; \mathbf{e}^h\right)\ (i = 1, \ldots, n). \tag{4}$$

For $x_i \geq 0$, this is a demand function for goods; for $x_i < 0$, it is a supply function (e.g., of factor-services).
Substituting from (2) and (4) into (3) then yields

$$\sum_i p_i x_i^h\left(\mathbf{v}; \mathbf{e}^h\right) = \sum_f d^{hf} \sum_i p_i y_i^f\left(\mathbf{v}; \mathbf{k}^f\right), \tag{5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

which holds identically for all **v**, $\mathbf{e}^h$, $\mathbf{k}^f$ and p_i. Summing up over all households then yields

$$\sum_h \sum_i p_i x_i^h\left(\mathbf{v}; \mathbf{e}^h\right) = \sum_h \sum_f d^{hf} \sum_i p_i y_i^f\left(\mathbf{v}; \mathbf{k}^f\right), \tag{6}$$

which we rewrite as

$$\sum_i p_i \sum_h x_i^h\left(\mathbf{v}; \mathbf{e}^h\right) = \sum_i p_i \sum_f \left(\sum_h d^{hf}\right) y_i^f\left(\mathbf{v}; \mathbf{k}^f\right). \tag{7}$$

On the assumption that firm *f* distributes all its profits,

$$\sum_h d^{hf} = 1 \text{ for all } f, \tag{8}$$

so that (7) reduces to

$$\sum_i p_i [X_i(\mathbf{v}; \mathbf{E}) - Y_i(\mathbf{v}; \mathbf{K})] = 0 \tag{9}$$

identically in all **v**, **E**, **K** and p_i, where

$$X_i(\mathbf{v}; \mathbf{E}) = \sum_h x_i^h\left(\mathbf{v}; \mathbf{e}^h\right) \text{ and } Y_i(\mathbf{v}; \mathbf{K}) = \sum_f y_i^f\left(\mathbf{v}; \mathbf{k}^f\right) \tag{10}$$

represent the aggregate demand and supply functions, respectively, for good *i*; **E** is a vector containing all the $\mathbf{e}^h$; and **K** a vector containing all the $\mathbf{k}^f$. If $X_i(\mathbf{v}; \mathbf{E}) - Y_i(\mathbf{v}; \mathbf{K}) > 0$, an

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

excess demand is said to exist in the market; if $X_i(\mathbf{v}, \mathbf{E}) - Y_i(\mathbf{v}, \mathbf{K}) < 0$, an excess supply; and if $X_i(\mathbf{v}, \mathbf{E}) = Y_i(\mathbf{v}, \mathbf{K})$, equilibrium.

Equation (9) is a general statement of Walras's Law. Its most frequent application in the literature has been (as in Walras's *Eléments* itself) to the general-equilibrium analysis of a system of perfect competition, in which the behaviour functions of firms are derived from the assumption that they maximize profits subject to their production function; and those of households are derived from the assumption that they maximize utility subject to their budget constraint. In this context, the vector **v** is the price vector $(p_1, \ldots, p_{n-1})$, with the *n*th good being money and serving as numéraire (i.e., $p_n = 1$), so that there are only $n - 1$ prices to be determined. Ignoring for simplicity vectors **E** and **K**, which remain constant in the conceptual market-experiment, equation (9) then becomes

$$\sum_{i=1}^{n} p_i[X_i(p_1, \ldots, p_{n-1}) - Y_i(p_1, \ldots, p_{n-1})] = 0 \text{ identically in the } p_i. \tag{11}$$

(Though it does not bear on the present subject, I should note that under the foregoing assumptions, and in the absence of money illusion, each of the demand and supply functions is homogeneous of degree zero in $p_1, \ldots, p_{n-1}$ and in whatever nominal financial assets are included in **E** and **K** (e.g., initial money holdings).) Thus Walras's Law states that no matter what the p_i, the aggregate value of excess demands in the system equals the aggregate value of excess supplies. This is the statement implicit in Lange's presentation (1942, p. 50).

Walras himself, however, sufficed with a particular and narrower application of this statement, and was followed in this by, *inter alia*, Hicks (1939, chs. IV: 3 and XII: 4–5), Modigliani (1944, pp. 215–16) and Patinkin (1956, ch. III: 1–3). Assume that it has been shown that a certain price vector $(p_1^0, \ldots, p_{n-1}^0)$ equilibrates all markets but the *j*th. Since (11) holds identically in the p_i, it must hold for this price vector too. Hence substituting the $n - 1$ equilibrium conditions into (11) reduces it to

$$p_j^0\left[X_j\left(p_1^0, \ldots, p_{n-1}^0\right) - Y_j\left(p_1^0, \ldots, p_{n-1}^0\right)\right] = 0. \tag{12}$$

Thus if $p_j^0 > 0$, the price vector $\left(p_1^0, \ldots, p_{n-1}^0\right)$ must also equilibrate the *j*th market, which means that only $n - 1$ of the equilibrium equations are independent. In this way Walras (and those who followed him) established the equality between the number of independent equations and the number of price-variables to be determined. (Though such an equality is not a sufficient condition for the existence of a unique solution with positive prices, it is a necessary – though not sufficient –condition for the peace-of-mind of those of us who do not aspire to the rigour of mathematical economists.)

It should however be noted that at the end of §126 of Walras's *Eléments* (1926), there is a hint of Lange's broader statement of the Law: for there Walras states that if at a certain set of prices ‘the total demand for some commodities is greater (or smaller) than their offer, then the offer of some of the other commodities must be greater (or smaller) than the demand for them’; what is missing here is the quantitative statement that the respective aggregate values of these excesses must be equal.

Since the contrary impression might be gained from some of the earlier literature (cf., e.g., Modigliani, 1944, pp. 215–16), it should be emphasized that no substantive difference can arise from the choice of the equation to be ‘dropped’ or ‘eliminated’ from a general-equilibrium system by virtue of Walras's Law. For identity (11) can be rewritten as

$$X_j(p_1, \ldots, p_{n-1}) - Y_j(p_1, \ldots, p_{n-1}) = -\frac{1}{p_j}\sum_{\substack{i=1 \\ i \neq j}}^{n} p_i[X_i(p_1, \ldots, p_{n-1}) - Y_i(p_1, \ldots, p_{n-1})] \text{ identically in the } p_i.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Thus the properties of the ‘eliminated’ equation are completely reflected in the remaining ones. Correspondingly, no matter what equation is ‘eliminated’, the solution for the equilibrium set of prices obtained from the remaining equations must be the same. (From this it is also clear that the heated ‘loanable-funds versus liquidity-preference’ debate that occupied the profession for many years after the appearance of the *General Theory*, was largely misguided; see Hicks, 1939, pp. 157–62; see also Patinkin, 1956, ch. XV: 3, and 1958, pp. 300–302, 316–17.)

In his influential article, Lange (1942, pp. 52–53) also distinguished between Walras's Law and what he called Say's Law, and I digress briefly to discuss this. As before, let the first $n - 1$ goods represent commodities and the nth good money. Then Say's Law according to Lange is

$$\sum_{i=1}^{n-1} p_i X_i(p_1, \ldots, p_{n-1}) = \sum_{i=1}^{n-1} p_i Y_i(p_1, \ldots, p_{n-1}) \text{identically in the } p_i. \tag{14}$$

That is, the aggregate value of commodities supplied at any price vector $(p_1, \ldots, p_{n-1})$ must equal the aggregate value demanded: supply always creates its own demand.

On both theoretical and doctrinal grounds, however, I must reject Lange's treatment of Say's Law. First of all, Lange himself demonstrates (ibid., p. 62) that identity (14) implies that money prices are indeterminate. In particular, subtracting (14) from (11) yields

$$X_n(p_1, \ldots, p_{n-1}) = Y_n(p_1, \ldots, p_{n-1}) \text{ for all } p_i. \tag{15}$$

That is, no matter what the price vector, the excess-demand equation for money must be satisfied, which in turn implies that money prices cannot be determined by market forces. But it is not very meaningful to speak of a money economy whose money prices are indeterminate even for fixed initial conditions as represented by the vectors **E** and **K**. So if we rule out this possibility, we can say that Say's Law in Lange's sense implies the existence of a barter economy. Conversely, in a barter economy (i.e., one in which there exist only the $n - 1$ commodities) Say's Law is simply a statement of Walras's Law. Thus from the above viewpoint, a necessary and sufficient condition for the existence of Say's Law in Lange's sense is that the economy in question be a barter economy: it has no place in a money economy (Patinkin, 1956, ch. VIII: 7).

Insofar as the doctrinal aspect is concerned, identity (14) cannot be accepted as a representation of Say's actual contention. For Say's concern was (in today's terminology) not the short-run viewpoint implicit in this identity, but the viewpoint which denied that in the long run inadequacy of demand would set a limit to the expansion of output. In brief, and again in today's terminology, Say's concern was to deny the possibility of secular stagnation, not that of cyclical depression and unemployment. Thus, writing in the first quarter of the 19th century, Say (1821a, p. 137) adduces evidence in support of his thesis from the fact ‘that there should now be bought and sold in France five or six times as many commodities, as in the miserable reign of Charles VI – four centuries earlier. Again, in his *Letters to Malthus* (1821b, pp. 4–5) Say argues that the enactment of the Elizabethan Poor Laws (codified at the end of the 16th century) proves that ‘*there was* no employ in a country which since then has been able to furnish enough for a double and triple number of labourers’ (italics in original). Similarly, Ricardo, the leading contemporary advocate of Say's *loi des débouchés*, discusses this law in chapter 21 of his *Principles* (1821), entitled ‘Effects of Accumulation of Profits and Interest’; on the other hand, he clearly recognizes the short-run ‘distress’ that can be generated by ‘Sudden Changes in the Channels of Trade’ (title of ch. 19 of his *Principles*. For further discussion see Patinkin, 1956, Supplementary Note L.)

Let me return now to the general statement of Walras's Law presented in equation (9) above. This statement holds for any vector **v** and not only for that appropriate to perfect competition. *In particular, Walras's Law holds also for the case in which households and/or firms are subject to quantity* constraints. In order to bring this out, consider the analysis of a disequilibrium economy presented in chapter XIII:2 of Patinkin (1956) and illustrated by Figure 1. In this figure, w is the money wage-rate, p the price level, N the quantity of labour, $N^d = Q(w/p, K_0)$ the firms’ demand curve for labour as derived from profit-maximization as of a given stock of physical capital K_0; and $N^s = R(w/p)$ is the supply curve of labour as derived from utility maximization subject to the budget constraint (these perfect-competition curves are what Clower (1965, p. 119) subsequently denoted as ‘notional curves’). Assume that because of the firms’ awareness that at the real wage rate $(w/p)_1$ they face a quantity constraint and will not be able to sell all of the output corresponding to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

their profit-maximizing input of labour N_1, they demand only N_2 units of labour, represented by point P in Figure 1. This constraint also operates on workers, who can sell only the foregoing quantity of labour instead of their optimal one N_3, represented by point H. In brief, at point P, both firms and workers are off their notional curves. In order to depict this situation, the notional curves must accordingly be replaced by quantity-constrained ones; namely, the kinked demand curve TAN_2 and kinked supply curve OUE. Note that for levels of employment before they become kinked, the curves coincide respectively with the notional ones (but see Patinkin, 1956, ch. XIII: 2, n. 9, for a basic analytical problem that arises with respect to the kinked demand curve TAN_2).

Figure 1

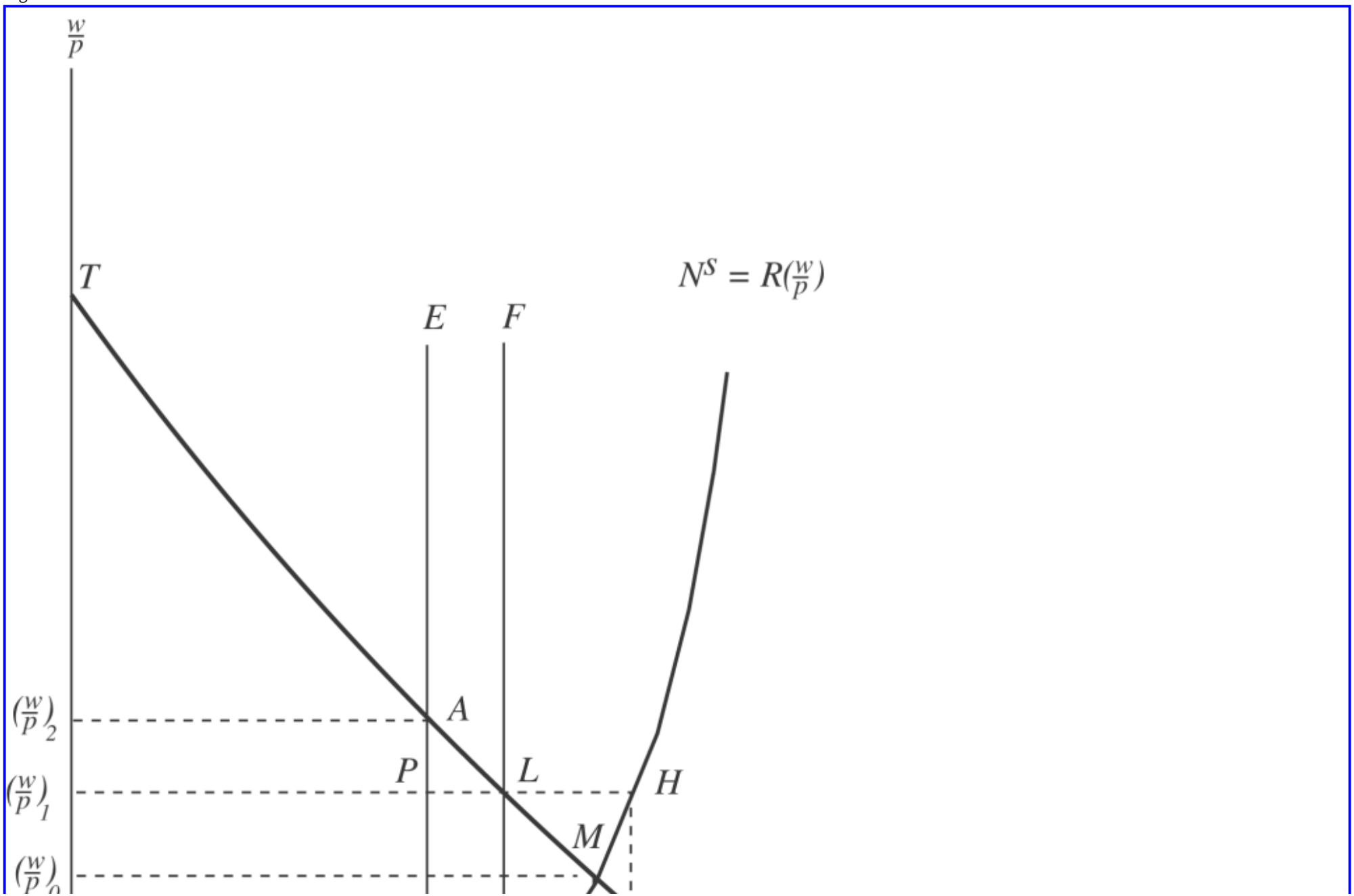

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

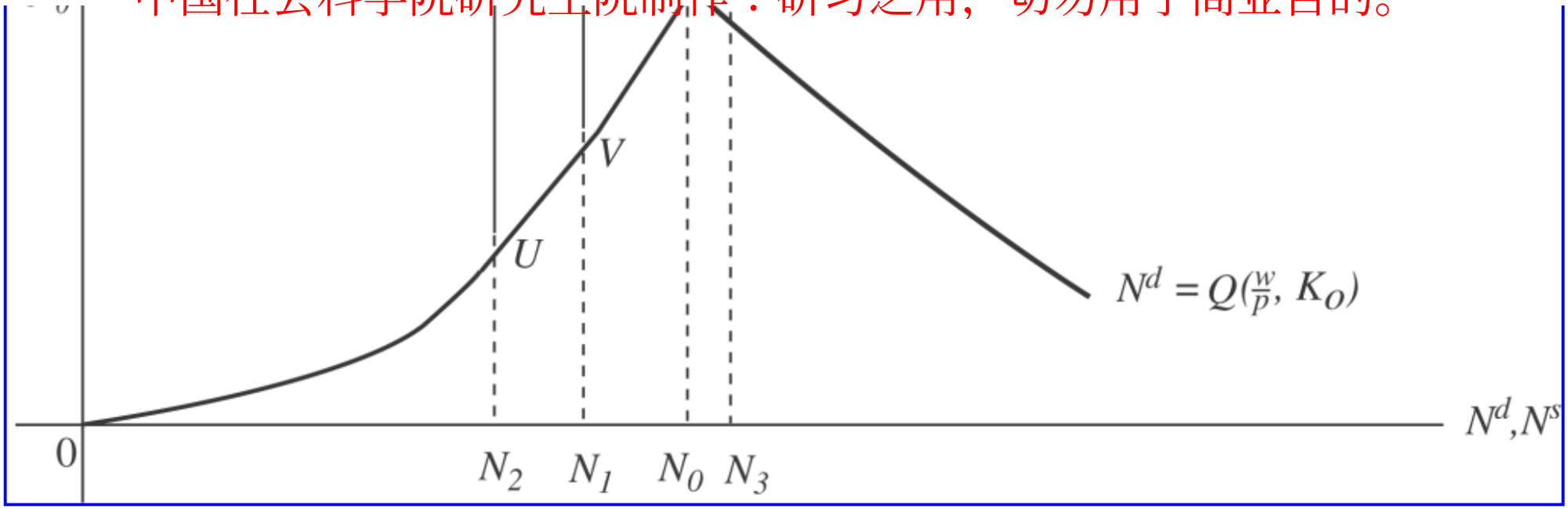

The obverse side of these constraints in the labour market are corresponding constraints in the commodity market. In particular, as Clower (1965, pp. 118–21) has emphasized, the demands of workers in this market are determined by their constrained incomes. Clower also emphasizes that it is this quantity constraint which rationalizes the consumption function of Keynes's *General Theory*, in which income appears as an independent variable. For in the absence of such a constraint, the individual's income is also a dependent variable, determined by the optimum quantity of labour he decides to sell at the given real wage rate in accordance with the labour-supply function $N^s = R(w/p)$ in Figure 1; and he makes this decision simultaneously with the one with respect to the optimum quantities of commodities to buy. If, however, his income is determined by a quantity constraint which prevents him from selling his optimum quantity of labour, the individual can decide on his demands for commodities only after his income is first determined. This is the so-called ‘dual decision hypothesis’ (Clower, ibid.). To this I would add (and its significance will become clear below) that the quantity constraint also rationalizes the form of Keynes's liquidity preference function, for this too depends on income (*General Theory*, p. 199). Furthermore, if the behaviour functions in the markets for labour, commodities and money balances are thus quantity-constrained, so too (by the budget constraint) will be that for bonds – the fourth market implicitly (and frequently explicitly) present in the Keynesian system. (The theory of the determination of equilibrium under quantity constraints – in brief, disequilibrium theory – has been the subject of a growing literature, most of it highly technical; for critical surveys of this literature, see Grandmont, 1977; Drazen, 1980; Fitoussi, 1983; and Gale, 1983, ch. 1.)

In the *General Theory* (ch. 2), Keynes accepted the ‘first classical postulate’ that the real wage is equal to the marginal product of labour, but rejected the second one, that it always also measures the marginal disutility of labour. In terms of Figure 1 this means that while firms are always on their demand curve $N^d = Q(w/p, K_0)$, workers are not always on their supply curve $N^s = R(w/p)$. Thus, for example, at the level of employment N_2, the labour market will be at point A on the labour-demand curve, corresponding to the real wage rate $(w/p)_2$; but the marginal utility of the quantity of commodities that workers then buy with their real-income $(w/p)_2$. N_2 is greater than the marginal disutility of that level of employment. And Keynes emphasizes that only in a situation of full-employment equilibrium – represented by intersection point *M* in Figure 1 – will both classical postulates be satisfied.

Consider now the commodity market as depicted in the usual Keynesian-cross diagram (Figure 2). The 45° line represents the amounts of commodities which firms produce and supply as they move along their labour-demand curve from point *T* to *M*. Thus *Y* represents the output (in real terms) of N_0 units of labour. Note too that the negative slope of the labour-demand curve implies that the real wage declines as we move rightwards along the 45° line.

Figure 2

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

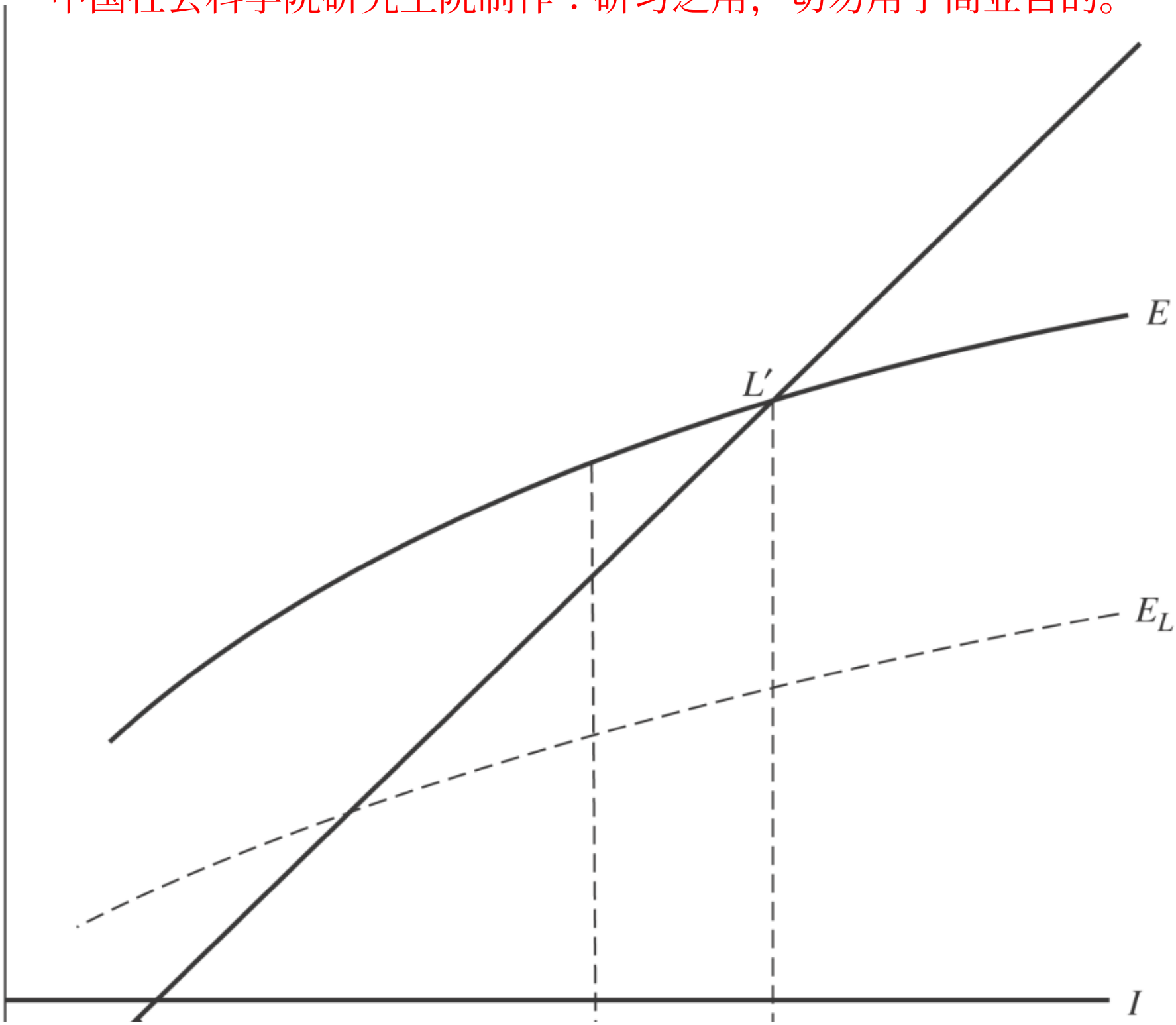

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

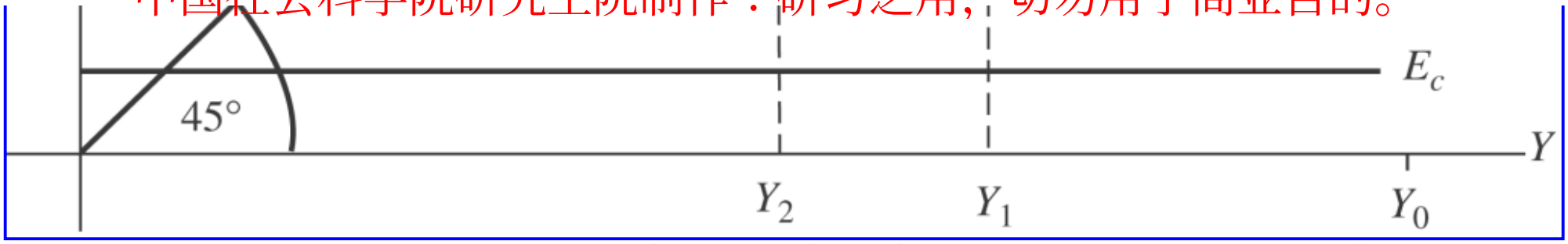

Curve E represents the aggregate demand curve, which is the vertical sum of the consumption function of workers (E_L) and capitalists (E_c), respectively, and of the investment function (I). For simplicity, these last two are assumed to be constant. The fact that curve E does not coincide with the 45° line reflects Keynes's assumption that in a monetary economy, Say's Law (in Keynes's sense, which is the macroeconomic counterpart of Lange's subsequent formulation) does not hold (ibid., pp. 25–6).

Consider now the consumption function of workers. The income which they have at their disposal is their constrained income as determined by the labour-demand curve in Figure 1. Thus assume that Y_1 and Y_2 in Figure 2 are the outputs corresponding to the levels of employment N_1 and N_2, respectively. The corresponding incomes of workers at these levels are $(w/p)_1 \cdot\cdot N_1$ and $(w/p)_2 \cdot\cdot N_2$. On the assumption that the elasticity of demand for labour is greater than unity, the higher the level of employment the greater the income of workers and hence their consumption expenditures. From Figure 2 we see that at income Y_2 there is an excess demand for commodities. This causes firms to expand their output to, say, Y_1, and hence their labour-input to, say, N_1, thus causing the constrained labour-supply curve to shift to the right to the kinked curve *OUVLF*. By construction, Y_1 is the equilibrium level of output.

What must now be emphasized is that Walras's Law holds in this situation too – *provided we relate this Law to excess-demand* functions *of the same type*. Thus if within our four-good Keynesian model we *consistently* consider notional behaviour functions, the excess supply of labour *LH* in Figure 1 corresponds to an excess demand for commodities which is generated by workers' planned consumption expenditures at the real wage rate $(w/p)_1$ and level of employment N_3 as compared with firms' planned output at that wage rate and lower level of employment N_2; and there will generally also exist a net excess planned demand for bonds and money. Alternatively, if we *consistently* consider constrained functions, then constrained equilibria exist in the labour market (point L), the commodity market (point L'), the bond market and the money market. Similarly, the broader form of Walras's Law states that a constrained (say) excess supply in the commodity market corresponds to a constrained net excess demand in the bond and money markets, while the labour market is in constrained equilibrium. In brief, a sufficient condition for the validity of Walras's Law is that the individual's demand and supply functions on which it is ultimately based are all derived from the same budget constraint, whether quantity-constrained or not. (This is the implicit assumption of Patinkin's (1956, p. 229; 1958, pp. 314–16) application of Walras's Law to a disequilibrium economy with unemployment, and the same is true for Grossman (1971) and Barro and Grossman (1971; 1976, p. 58).)

I must admit that the validity of Walras's Law in this Keynesian model depends on our regarding the kinked curve *OUVLF* as a labour-supply curve, and that this is not completely consistent with the usual meaning of a supply curve or function. For such a function usually describes the behaviour of an agent under constraints which leave him some degree of freedom to choose an optimum, whereas no such freedom exists in the vertical part of *OUVLF*. However, I prefer this inconsistency to what I would consider to be the logical – and hence more serious – inconsistency that lies at the base of the rejection of Walras's Law, and which consists of lumping together behaviour functions derived from different budget constraints.

It is thus clear that the foregoing constrained equilibrium in the labour market is not an equilibrium in the literal sense of representing a balance of opposing market forces, but simply the reflection of the passive adjustment by workers of the amount of labour they supply to the amount demanded by firms (cf. Patinkin, 1958, pp. 314–15). From this viewpoint, the constrained equilibrium in the labour market always exists and simply expresses the fact that, by definition, every *ex post* purchase is also an *ex post* sale. In contrast, as we have seen in the discussion of Figure 2 above, the corresponding constrained equilibrium in the commodity market is a true one; for, in accordance with the usual Keynesian analysis, were the level of Y to deviate from Y_1, automatic market forces of excess demand or supply would be generated that would return it to Y_1. And a similar statement holds, *mutatis mutandis*, for the constrained equilibria in the bond and money markets.

Note, however, that in the commodity market too there is an *ex post* element. This element is a basic, if inadequately recognized, aspect of the household behaviour implied by Clower's 'dual decision hypothesis': namely, that households' constrained decisions on the amount of money to spend on commodities is based on their *ex post* knowledge of the amount of money received from the constrained sale of their factor services. And to this I again add that a similar statement holds for their constrained decisions with reference to the amounts of bonds and money balances, respectively, that they will want to hold. (Note that an analysis in terms of constrained decisions can also be applied to the case in which households do not correctly estimate firms' profits in equation (3) above, and are thus forced to base their effective (say) consumption decisions on the *ex post* knowledge of these profits.)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

In this treatment of an economy with constrained functions, Clower (1965, pp. 122–3) has claimed that under these conditions Walras's Law does not hold. This is not true for the Law as hitherto discussed. What Clower seems to have in mind, however, is that though the excess supply of labour *LH* in Figure 1 is notional, it nevertheless exerts pressure on workers to reduce their money wages; in contrast, the notional excess demand for commodities corresponding to *LH* (see above) cannot – because of their constrained incomes – lead households to exert expansionary pressures on the commodity market. Thus there exists no *effective* excess demand for commodities to match the *effective* excess supply of labour. Accordingly, no ‘signal’ to the market is generated that will lead to the expansion of output and consequent reduction of unemployment (cf. also Leijonhufvud, 1968, pp. 81–91). And it is the absence of such a ‘signal’ that Leijonhufvud (1981, ch. 6) subsequently denoted as ‘effective demand failure’.

This ‘failure’, however – and correspondingly the failure of Walras's Law to hold in Clower's sense – is not an absolute one: for though there is no direct signal to the commodity market, an indirect one may well be generated. In particular, the very fact that the constrained equilibrium in the labour market does not represent a balancing of market forces means that the unemployed workers in this market are a potential source of a downward pressure on the money wage rate. And if this pressure is to some extent effective, the resulting decline in money wages will generate an increase in the real quantity of money, hence a decrease in the rate of interest, hence an increase in investment and consequently in aggregate demand – and this process may be reinforced by a positive real-balance effect (see chapter 19 of the *General Theory*, which, however, also emphasizes how many weak – and possibly perverse – links there are in this causal chain). Thus a sufficient condition for absolute ‘effective demand failure’ is the traditional classical one of absolute rigidity of money wages and prices.

An analogy (though from a completely different field) may be of help in clarifying the nature of the foregoing equilibrium in the labour market. Consider a cartel of (say) oil-producing firms, operating by means of a Central Executive for the Production of Oil (CEPO) which sets production quotas for each firm. The total quantity-constrained supply so determined, in conjunction with the demand conditions in the market, will then determine the equilibrium price for crude oil, and that equilibrium position is the relevant one for Walras's Law. But this will not be an equilibrium in the full sense of the term, for it coexists with market pressures to disturb it. In particular, the monopolistic price resulting from CEPO's policy is necessarily higher than the marginal cost of any individual member of the cartel. Hence it is to he interest of every firm in the cartel that all other firms adhere to their respective quotas and thus ‘hold an umbrella’ over the price, while it itself surreptitiously exceeds the quantity constraint imposed by its quota and thus moves closer to its notional supply curve as represented by its marginal-cost curve. And since in the course of time there will be some firms who will succumb to this temptation, a temptation that increases inversely with the ratio of its quota to the total set by CEPO, actual industry output will exceed this total, with a consequent decline in the price of oil. Indeed, if such violations of cartel discipline should become widespread, its very existence would be threatened.

This analogy is, of course, not perfect. First of all, unlike workers in the labour market, the member-firms of CEPO have themselves had a voice in determining the quantity constraints. Second, and more important, any individual firm knows that by ‘chiselling’ and offering to sell even slightly below the cartel price, it can readily increase its sales. But analogies are never perfect: that is why they remain only analogies.

A final observation: the discussion until now has implicitly dealt with models with discrete time periods. In models with continuous time, there are two Walras's Laws: one for stocks and one for flows: one for the instantaneous planned (or constrained) purchases and sales of assets (primarily financial assets) and one for the planned (or constrained) purchases and sales of commodity flows (cf. May, 1970; Foley and Sidrauski, 1971, pp. 89–91; Sargent, 1979, pp. 67–69; Buiter, 1980). On the other hand, in a discrete-time intertemporal model, in which there exists a market for each period, there is only one Walras's Law: for in such a model, all variables have the time-dimension of a stock (see Patinkin, 1972, ch. 1).

See Also

- general equilibrium
- temporary equilibrium

Bibliography

Arrow, K.J. and Hahn, F.H. 1971. *General Competitive Analysis*. San Francisco: Holden-Day; Edinburgh: Oliver and Boyd.

Barro, R. and Grossman, H.I. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61, 82–93.

Barro, R. and Grossman, H.I. 1976. *Money, Employment and Inflation*. Cambridge: Cambridge University Press.

Buiter, W.H. 1980. Walras's law and all that: budget constraints and balance sheet constraints in period models and continuous time models. *International Economic Review* 21, 1–16.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Clower, R. 1965. The Keynesian counterrevolution: a theoretical appraisal. In *The Theory of Interest Rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan, 103–25.

Drazen, A. 1980. Recent developments in macroeconomic disequilibrium theory. *Econometrica* 48, 283–306.

Fitoussi, J.P. 1983. Modern macroeconomic theory: an overview. In *Modern Macroeconomic Theory* ed. J.P. Fitoussi. Oxford: Basil Blackwell, 1–46.

Foley, D.K. and Sidrauski, M. 1971. *Monetary and Fiscal Policy in a Growing Economy*. London: Macmillan.

Gale, D. 1983. *Money: in Disequilibrium*. Cambridge: Cambridge University Press.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45, 535–72.

Grossman, H.I. 1971. Money, interest, and prices in market disequilibrium. *Journal of Political Economy* 79, 943–61.

Hicks, J.R. 1939. *Value and Capital*. Oxford: Clarendon Press.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Lange, O. 1942. Say's Law: A restatement and criticism. In *Studies in Mathematical Economics and Econometrics*, ed. Oscar Lange et al., Chicago: University of Chicago Press, 49–68.

Leijonhufvud, A. 1968. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press.

Leijonhufvud, A. 1981. *Information and Coordination: Essays in Macroeconomic Theory*. Oxford: Oxford University Press.

May, J. 1970. Period analysis and continuous analysis in Patinkin's macroeconomic model. *Journal of Economic Theory* 2, 1–9.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12, January, 45–88. As reprinted in *Readings in Monetary Theory*, ed. F.A. Lutz and L. W. Mints, Philadelphia: Blakiston, for the American Economic Association, 1951, 186–240.

Patinkin, D. 1956. *Money, Interest, and Prices*. Evanston, IL: Row, Peterson. (The material referred to appears unchanged in the second, 1965 edition.)

Patinkin, D. 1958. Liquidity preference and loanable funds: stock and flow analysis. *Economica* 25, 300–18.

Patinkin, D. 1972. *Studies in Monetary Economics*. New York: Harper & Row.

Ricardo, D. 1821. *On the Principles of Political Economy and Taxation*, 3rd edn. As reprinted in The Works and Correspondence of David Ricardo, Vol. I, ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

Sargent, T.J. 1979. *Macroeconomic Theory*. New York: Academic Press.

Say, J.B. 1821a. *Traité d'économie politique*, 4th edn, Paris: Deterville. As translated by C.R. Prinsep under the title *A Treatise on Political Economy*, Philadelphia, Grigg R. Elliot: 1834.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Say, J.B. 1821b. *Letters to Thomas Robert Malthus on Political Economy and Stagnation of Commerce*. London. As reprinted with an Historical Preface by H.J. Laski, London: George Harding's Bookshop, 1936; Wheeler Economic and Historical Reprints No. 2.

Walras, L. 1874. *Eléments d'économie politique pure*. Paris: Guillaumin. (Sections I–III of the work).

Walras, L. 1900. *Eléments d'économie politique pure*. 4th edn, Paris: F. Pichon.

Walras, L. 1926. *Eléments d'économie politique pure*. Definitive edition. Paris: F. Pichon (for our purposes, identical with 4th edition). As trans. by William Jaffé under the title *Elements of Pure Economics*, London: George Allen and Unwin, 1954.

How to cite this article

Patinkin, Don. "Walras's Law." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000020> doi:10.1057/9780230226203.1815

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

war and economics

Craufurd D. Goodwin
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

War has always intrigued economic thinkers: how economic interdependence affects the likelihood of conflict, the costs and benefits of aggression, how to fight a successful war and achieve a stable peace. Many prominent early economists had useful things to say, for example Smith, J.B. Clark, Pigou, Veblen and J.M. Keynes. Economists came into their element in the Second World War and the Cold War; welcomed into government, they distinguished themselves as macroeconomists, microeconomists, broad-based advisers, public administrators, and public figures. By the 1990s economists concerned with war and peace had established their own sub-discipline, with journals, textbooks and college courses.

Keywords

Angell, N.; autarky; Baran, P. A.; Cassel, G.; Clark, J. B.; Clark, J. M.; classical economics; Colm, G.; colonialism; Council of Economic Advisers (USA); Currie, L.; defence economics; Deterrence; Enlightenment; Galbraith, J. K.; game theory; heterodox economics; Hobson, J. A.; Keynes, J. M.; Kindleberger, C.; Lloyd, H.; Malthus, T. R.; marginal revolution; Mill, J. S.; operations research; Parsons, T.; Pigou, A. C.; Rostow, W. W.; rule of law; Schelling, T.; Smith, A.; Sweezy, P. M.; Veblen, T.; war and economics

Article

Prehistory

Political economy emerged from moral philosophy as a distinct field of study in the 18th century. Before that economic ideas were part of a continuum developing from the speculation of Greek philosophers about how to achieve the good life, the advice from medieval schoolmen about how to achieve salvation, and the recommendations of 17th century merchants and public administrators about how to strengthen the emerging nation-states. In this early literature economic ideas were entwined with concerns about

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
communal security. No matter whether the unit of concern was a nation, a city-state, or a manor, conquest by an external enemy was seen as a catastrophe. Conversely, great economic benefit was seen to come from military victory.

Hesiod in eighth-century bc Greece suggested that economic isolation and self-sufficiency might be the best protection against war because for voracious enemies out-of-sight was out-of-mind. Indeed, the ideal communities imagined by Plato in Atlantis, Sir Thomas More in Utopia, and Francis Bacon in New Atlantis were all remote from their neighbours (Lowry, 1991, p. 6). Security, they implied, lay in isolation and economic autarky. In economic interdependence there was danger. Economic interchange requires human interaction and from such contacts emerge frictions, cupidity, and ultimately physical violence. The state dealt with domestic violence through the police and the courts. Correspondingly international trade and flows of resources across borders should be constrained by the state to reduce tensions that could lead to war. The ancients recognized that, if war should begin, the structure of the economy could profoundly affect the capacity of a state to defend itself. The early Greeks appreciated that the highly efficient hoplite military formation, as well as other military tactics, depended upon the strength and high morale of their small-scale farming population. As an effective fighting machine a labour force of docile day labourers just would not do. Similarly, the citizen-based merchant marine of Athens was the foundation upon which her powerful navy rested (Lowry, 1991, p. 7).

Warfare was treated by the Greek philosophers and their admirers in the Middle Ages as simply one among many possible economic activities. Trade, colonization and conquest of one's neighbours were alternative uses of national resources. The task of leaders was to identify the best among them. This calculation required the kind of hardheaded cost–benefit analysis that became the hallmark of later economics. Erasmus wrote in 1516 in a tract for the future Charles V of Spain: ‘when the prince has put away all personal feelings, let him take a rational estimate long enough to reckon what the war will cost and whether the final end to be gained is worth that much – even if victory is certain’ (Lowry, 1991, p. 17). Erasmus warned that any responsible leader must guard against the self-interested advice of private persons who might gain from war while the community at large would lose. ‘It too often happens that nobles, who are more lavish than their means allow, when the opportunity is presented stir up war in order to replenish their resources even at home by the plunder of their peoples’ (1991, p. 17). This advice sounds like an early warning against a military–industrial complex!

War and classical political economy

To a degree, the classical political economy that appeared in the 18th century was an extension of the earlier ‘mirror for princes’ literature to which Erasmus contributed. It, too, was intended to guide statesmen in the preparation and implementation of public policy. An important difference, however, was that the political economists thought that they brought much greater impartiality and detachment to their work. Policy implications, they believed, emerged ineluctably from their theory, and ad hoc reasoning could be replaced by principle (Silberner, 1946). Adam Smith took the same cost–benefit stance towards warfare as Erasmus, but he developed its implications more fully, and he drew from them optimistic prospects for the future (Smith, 1776, pp. 689–708; and Goodwin, 1991, pp. 24–8). In the tradition of Enlightenment thinking, Smith presumed good political leaders would be rational, and the challenge, therefore, was to construct circumstances in which they would select peace over war.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Happily, Smith observed, conditions were likely to emerge from growth in ‘the necessaries, comforts and conveniences of life’ that would make nations ever more peaceful. A consequence of economic growth was that, over time, the opportunity costs of warfare rose – reflecting both the direct costs of munitions and supplies and the production forgone from military servicemen in the prime of their productive lives. On the other hand, the value of potential spoils of war to an advancing country would no more than remain constant. As the costs rose and the potential gains remained fixed, the likelihood of benefiting from aggression declined, and the warlike tendencies of rational nations fell steadily as they grew economically. The military preparedness of a prosperous nation, therefore, would seldom be for offence. However, preparedness remained critical for defence, especially against less prosperous neighbours whose cost–benefit calculations of aggression against them might for some time remain positive. The main task of the military establishment in any prosperous nation was to make the costs of aggression against it higher than any plausible gain. In this way the notion of rational deterrence was given a precise exposition by Smith.

An intriguing yet minor Enlightenment thinker who proposed an economic doctrine while applying it also to military strategy and tactics was the Welshman Henry Lloyd, a major-general and professional mercenary in the wars of the 18th century, parodied by Gilbert and Sullivan in *Pirates of Penzance* for his commitment to social theory (Speelman, 2002).

Thomas Robert Malthus among the classical economists provided a reason rooted in demographic growth to explain why the costs of conflict should always be kept clearly before potential aggressors (Malthus, 1826, p. 47; and Goodwin, 1991, p. 26). When an increasing population pressed hard upon natural resources a suffering nation was likely to look abroad for relief and conflict would ensue. Indeed, the slaughter of warfare became one of the powerful ‘positive’ checks to population growth. Deterrence costs should always be kept sufficient to make certain that aggression would be carried out by overpopulated countries against someone else. John Stuart Mill concluded in 1848 that deterrence had by that date been so successful that war had become likely only in colonies where ‘savages’ prepared unreasonable cost-benefit calculations (Mill, 1848, p. 707).

Economists, beginning with Smith, were quick to see an analogy between the security needs of all nations taken together and those of small communities that formed themselves into nation-states to benefit from the rule of law. Costs of protection could be reduced globally through international agreements. Smith favoured for Britain a kind of ‘commonwealth’ that would include former colonies. Later economists were enthusiastic supporters of such institutions of world government as the International Court of Justice at The Hague, the League of Nations and the United Nations.

Economic thinkers have often puzzled about just what is required to fight a successful war. Some pamphleteers in the 17th century thought that a stock of ‘treasure’ was needed. Smith thought, to the contrary, that a satisfactory flow of product was necessary; a vibrant and robust economy would allow the state to ‘draw from their subjects extraordinary aids upon extraordinary occasions’ (Smith, 1776, p. 446). Costs to the economy, he thought, were necessarily different in a small war from a large one. For example, soldiers who were not seriously challenged by the prospect of a slight skirmish in a minor war might be persuaded to volunteer at wages below market rates, led on by their ‘youthful fancies’ and ‘romantic lapses’. The taxpayers in this situation, moreover, might not complain about modest levies because ‘this amusement compensates the small difference between the taxes which they pay on account of the war, and those which they are accustomed to pay in time of peace’ (1776, p. 920). In a full-scale war, on the other hand, soldiers fearing personal injury demanded payment at top market rates, taxpayers

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
balked at the levies upon them, and it became necessary to issue public debt as the only way to raise revenue. Happily the net costs of war usually were low because productive capital was seldom damaged and, as the classical economist John Rae observed, the turmoil of conflict seemed to stimulate the powers of invention (Rae, 1834, pp. 222–3).

Marginal economics applied to war

An important turn in the evolution of modern economics occurred in the 1870s when the mainstream accepted utility maximization as its norm and incremental analysis as its method (Black, Coats and Goodwin, 1973, p. 19). This shift took place in the middle of the Pax Britannica, one of the most peaceful periods in European history, and the new analytical tools had seldom to be used to explain warfare. Under the marginal paradigm all destructive conflict was treated as socially irrational and contributing to net social disutility. War was mentioned in relation to public finance where it was usually categorized as a ‘bad’, like crime and disease, the opposite of a ‘good’. The main questions were how many public resources should be allocated to deterrence and how should these resources be deployed. Two writers who epitomized the marginalist approach to conflict in the years leading up to the First World War were the English economist/journalist Norman Angell, and the pioneering American price theorist John Bates Clark (Barber, 1991a, pp. 61–84). Angell used economic reasoning in a book on war entitled *The Great Illusion* (1910), which reached a large and influential readership. He claimed that because of widespread international economic integration war had become everywhere futile and irrational. No one could gain from a fight. The economic dependence of virtually all countries on their neighbours for markets, products and materials made it impossible for them to achieve a net gain through conquest. This was not to say that war was impossible, only that it was unlikely by the economist's logic. This position, of course, was the exact opposite of the venerable idea developed by Plato that economic integration was dangerous and economic autarky could be the road to world peace. In Angell's view, Adam Smith's desired condition for deterrence had been achieved in which all powerful nations that made a disinterested study must conclude that the costs of war exceeded the benefits. Even indemnity payments to winners – if they could be enforced, as they had been after the Franco–Prussian War – were likely to cause serious market dislocations in the recipient nation as well as in the payer and were of little net gain to anyone.

John Bates Clark of Columbia University, the most prominent American marginal economist, studied war as part-time Director of the Division of Economics and History of the Carnegie Endowment for International Peace, founded by Andrew Carnegie in 1910. In 1911 Clark mobilized in Bern most of the celebrated economists of the time from around the world to explore how economics could help achieve permanent world peace. Those in attendance were conscious of the current threat of war and agreed to turn their talents to how it could be avoided. Clark's hope was to identify adjustments that might be made to incentives and rewards in the global economy to reduce the likelihood of conflict. There was a prevailing sense among the economists that tensions grew out of structural errors and misunderstandings, and in this Clark saw an analogy with industrial relations. Mediation and arbitration that worked in one sphere might work also in the other. One task for economists could be to propose equitable middle ground between potential combatants.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Some dissenting economists ’ voices

Despite Clark's and Angell's optimism about the capacity of economics to show the way to a stable world peace, heterodox economists in the socialist, historical and emerging American Institutionalist traditions continued to worry that economic forces within the industrial nations would lead inevitably to conflict. Thorstein Veblen, John A. Hobson and others warned of tensions based on class division and the proclivity of one social group to exploit another (Biddle and Samuels, 1991; and Davis, 1991). When these tensions spread overseas through colonial expansion and an aggressive search for profits the possibility of international conflict became serious.

No matter whether economists writing about conflict in the years before the First World War were in the respectable mainstream or in the professional underworld, their reflections were mainly those of outsiders. War was a subject about which their theory seemed to have something to say, and so they said it. But there is little evidence that their observations had a measurable impact on the thought or actions of those who decided upon peace and war.

The First World War threw many prominent economists nearly into despair: the war was so costly, so barbarous, and so irrational! It seemed even to cast doubt on the controlling force of reason in society. Francis Hirst, editor of *The Economist*, in a personal manifesto in 1915 entitled ‘The Political Economy of War’ saw ahead only ‘social and economic ruin’ (Barber 1991a, p. 72). Gustav Cassel in Sweden and the young John Maynard Keynes in Britain were deeply discouraged to find that despite the global integration that Angell and others had described the belligerents seemed able and willing to continue the conflict for a prolonged period, far after any rational leaders would have been expected to make peace (Barber, 1991a, pp. 76–9). After the war finally came to a close Keynes wrote in his celebrated polemic against the peace treaty, *The Economic Consequences of the Peace* (1920), that sheer vindictiveness had overcome good sense. Later, Keynes, John Bates Clark, and other economists were outraged that the American Senate refused to ratify American participation in the League of Nations. It seemed to them that once again, prejudice and political opportunism had overwhelmed reason on vital matters related to peace and war. So what use was economics in the search for peace? After a century of giving positive answers to this question economists had now to admit ‘perhaps not very much’!

Economists in the years between the two world wars, apparently humbled by the seeming irrelevancy of their efforts in the years surrounding the first war, were generally reluctant to engage with subjects related to peace, war and national security. They had lost their self-confidence and they responded by withdrawing from the field. War for most economists simply was no longer a respectable area for attention. It became, instead, a condition exogenous to their models and left to be analysed by other disciplines. Symbolically, the Carnegie Endowment for International Peace restructured its efforts in search of peace away from the abortive multi-national study of the economics of war to a 132-volume economic and social history of the First World War under the direction of the distinguished Columbia historian James T. Shotwell. The clear message of the Carnegie switch was that since the deductive methods of the marginal economists had not shown the way to peace it was time to give inductive historians a chance.

One of the few exceptions to the interwar inattention by economists to peace and war came from an unlikely source, the Cambridge theorist and successor to Alfred Marshall, A.C. Pigou. In his *Political Economy of War* (1921) Pigou set out to shed light on ‘the anatomy and physiology’ of ‘a strained and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
stressed economy' (Barber, 1991b, p. 131). Another exception was John Maurice Clark who, with his father John Bates Clark, attempted to estimate the full economic cost of the First World War, including the value of lost human as well as physical capital (Clark, 1931).

It appeared to many professional economists in the years between the wars that the dragon they had to slay in order to achieve peace was not in fact war itself but rather the conditions that led to war – at that time, economic depression and unemployment. So long as workers could not find jobs, and bankruptcies destroyed entrepreneurs, civil insurrection was likely. And when economic distress led to the rise of demagogues then world peace was truly at risk. Enquiry into the root causes of macroeconomic crisis proceeded at several levels. In Europe, business cycle institutes looked for disturbing patterns in worldwide economic activity and for means of coping cooperatively with distress (De Marchi, 1991). In the United States, economists inside and out of the Roosevelt administration experimented with a variety of new stimulative and regulatory mechanisms. In Britain, Keynes and his students at the University of Cambridge developed and propagated a new macroeconomics focused on the adequacy of aggregate effective demand, and they pointed to fiscal policy as the means to achieve economic stability. Keynes recalled T.R. Malthus's observation that wars of aggression could emerge from nations that had been unsuccessful in solving their internal economic problems, but the crucial short-run internal challenge now, he noted, was unemployment rather than overpopulation. The causal sequence, nevertheless, was the same. The route to world peace lay in internal macroeconomic reform and in international economic cooperation.

Victory for the economists in the Second World War

The economics discipline entered the Second World War without having thought deeply over the prior two decades about how to respond to conflict. Certainly it had no sub-discipline, as it does today, concerned with the economics of war, peace, defence and security. Moreover, economics was held in relatively low repute, at least among those in power. It was perceived, correctly, as an eclectic multi-paradigmatic field that spoke still with many voices: marginalist, Institutionalist, Keynesian, socialist, historical and others. Moreover, several of these voices were to a degree threatening to the established order. They came in America from advisers to US President Franklin D. Roosevelt, friends of organized labour and farmers, from architects of a welfare state – and worse. On the American campus in the 1930s academic freedom cases frequently involved economists, the result of the authorities attempting to silence troublemakers (Hofstadter and Metzger, 1955). It was ironic that by 1945 the Second World War had given to economics a much improved reputation and a perception among outsiders of a social value that had not been recognized before. Arguably, it gave as much to economics as economics gave to it. Even though few economists had made a study of war before 1939 they seemed to know what to do when the war began. Undoubtedly the memory of poor administrative performance during the First World War caused governments to welcome economists with relatively open arms 25 years later. In the first war resources and finished goods had been squandered by the bureaucrats and business persons recruited for the occasion. War finance had stumbled, prices had risen unacceptably, and profiteering occurred to a scandalous degree. Now, the economists claimed, they could do it all better.

It turned out that in the Second World War economists of all stripes could find useful work. Moreover, they got along rather well with the public servants and a new set of volunteers from the private sector.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Institutionalists like John Kenneth Galbraith supervised price control and war mobilization (Galbraith, 1981). Keynesians like Lauchlin Currie and Gerhard Colm designed non-inflationary monetary and fiscal policy (Sandilands, 1990; and Stein, 1969). Microeconomists like Charles Kindleberger (Kindleberger, 1991) demonstrated the use of optimizing techniques in selecting bombing targets, and economic historians like Walt Rostow helped to create the first high-powered intelligence service (Lodewijks, 1991). Even smart young Marxists like Paul Baran and Paul Sweezy found useful roles. A consequence of the impressive wartime performance by economists was that, when the war ended, the discipline had gained new respect. Indeed, politicians and bureaucrats alike found the economist's perspective so useful that they searched for ways to embed the discipline in government in peacetime. The Council of Economic Advisers, set up in the White House, and the semi-public RAND Corporation were both established with this objective in mind. After the war economists were sprinkled liberally throughout both the executive and legislative branches and the Federal Reserve, and the economist's approach was seen in the consideration of problems both of peace and war. The placement by Secretary of Defense McNamara of ‘whiz kids’ – economists and ‘systems analysts’ – firmly by his side when he entered the Kennedy administration in 1960 was only the latest step in the infiltration by economists that had started 20 years before.

When the National Science Foundation was created after the Second World War in response to Vannevar Bush's persuasive picture, *Science, the Endless Frontier* (1945) it was symbolic that the case for inclusion of economics was made not on the basis of its theoretical innovations or contributions over two centuries to human welfare; instead, a list was presented of the contributions to the recent war effort. The sociologist/economist Talcott Parsons, who was charged by the Social Science Research Council to prepare a lobbying document for use with Congress, described the participation of economists in such new agencies as the Office of Strategic Services and the Foreign Economic Administration, as well as the strengthening of older ones such as the Bureau of Agricultural Economics and the Treasury. The most impressive contribution of economists, Parsons concluded, came in the application of the Keynesian theory of income determination, primarily by Simon Kuznets, to achieve wartime full employment without inflation (Klausner and Lidz, 1986, p. 100).

The discussion in this entry thus far has been mainly about how economics has addressed war throughout its history, but after 1945 the relationship ran in both directions. Economics was itself changed fundamentally by its involvement with war though contacts with new people, new problems, new professional circumstances and new funding sources such as the Office of Naval Research. The most revolutionary new tool brought into economic analysis during the war was game theory, emerging from the collaboration of the mathematician John Von Neumann with the economist Oscar Morgenstern (Weintraub, 1992). Used originally to model parties in conflict it was extended to models of interactions between all kinds of actors, including those in cooperation. Operations research was another field with deep military roots to which economics contributed, and from which it brought back tools to the discipline (Mirowski, 2002).

New global breadth was the third contribution of war to economics. Not only were conventional issues of international trade and finance given new urgency by the challenge of rebuilding the world economy, but for both economic and strategic reasons it became necessary to understand friends and foes more thoroughly. This gave rise to new fields called ‘comparative economic systems’ and ‘Soviet studies’. When it became necessary to organize recovery from the Second World War and construct strong

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
nations as bulwarks against Communism, ‘development economics’ appeared as a new sub-discipline, with generous funding from government and private foundations.

The triumph of economics in the Cold War

The rise to prominence of economics in the study of war began during the Second World War. The discipline had shown that it could be useful in many ways during full mobilization. But the complete edifice of modern defence economics was constructed only later, during the Cold War (1946–89). There were three reasons. Perhaps of greatest significance was the steadily increasing cost of weapons systems. More items than just nuclear weapons and space exploration bore big price tags – such things as carrier battle groups, strategic bombers and foreign basing of forces. Clearly, hard choices had to be made among alternatives. Yet the questions remained: who should make the choices and how should they be made? Conflict of interest might become rampant. Decision-making by politicians was not appealing, since they were likely to put the prospect of re-election ahead of the public interest, and maybe even of human survival. Military forces protected the republic, but from the perspective of a legislator they also resembled pork. The military contractors in the private sector were just as problematic: profit rather than national interest usually dominated their incentive structures, even when on leave in government. Despite legislation that established a unified command under the Joint Chiefs of Staff, leaders of the separate forces still engaged in wasteful competition for resources and resisted cost-saving cooperation. When the warrior president Dwight Eisenhower echoed Erasmus and warned of the dangers from a rising ‘military–industrial complex’ there was an unassailable cause for concern.

So enter the economists. They announced that optimization subject to constraints was their specialty. Scarcity was their stock in trade. Pure reason was their method – and they were incorruptible. They accepted neither ideology nor any sentimental appeals to service loyalty or norms other than the public welfare. They were trained to look for rent seekers under every bed. Voters accepted the necessity of an adequate defence in the Cold War but they wanted as little of it as possible to get the job done, and they wanted it at the lowest cost. Economists made a plausible case that they were best equipped to get the job done that way.

A second reason for the rise to power of economists during the Cold War was their evident command of analytical techniques critical to the management of nuclear confrontation. The skills of the economist in minimizing the costs of defence were certainly comforting, but their avowed capacity to manage weapons of mass destruction was valued even more. The essential consideration was that if nuclear weapons were ever used again this would constitute profound policy failure. Bluff, restraint, and an absolute reluctance to engage simply had to be the main characteristics of nuclear policy. But surely these were not the characteristics that were ingrained in the professional soldier, who was trained to fight – precisely the action that could not be countenanced when the weapons employed would kill friend and foe alike? Indeed, the Cold War was the economist's war par excellence, fought at last with the condition wished for by Adam Smith 200 years before that the powerful nations of the world would reach a stage when conflict of any kind would be ‘irrational’ and the role of strategy was exclusively to arrange for war never to occur. Deterrence had to be on a massive scale and every step taken to see that there could not be a miscalculation. The first mistake could be fatal. The memory of Sarajevo, and world war by accident, was very vivid still. Economists made a strong case for their role in military strategy and operations because, above all, they seemed to recognize the absolute necessity of peace during the Cold

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
War.
The third reason for the rise to prominence of economists in the Cold War, and for their fingers rather than those of more conventional warriors to be on the trigger, was the alternatives. In the post-war years an image emerged of high-placed military commanders that was deeply troubling; in important cases they seemed arrogant, vain, and sometimes out of control. As events unfolded and memoirs were published, Generals Patton, De Gaulle, MacArthur, Montgomery and others all appeared to be in varying degrees emotional prima donnas, insubordinate to civilian control. These men might be much beloved of their troops, and they might have tactical skills and ‘leadership qualities’ necessary to triumph in conventional battle, but could they be trusted with the ticket to Armageddon? Many voters said emphatically ‘No!’ Closer in time, by the 1950s, than the mythic warriors of the Second World War was General Curtis Lemay, chief of the United States Strategic Air Command, who talked of bombing the enemy back to the Stone Age. Worried citizens might be pardoned for wondering whether, if ever he took such action the enemy might not find the means to bring everyone else along with them. The fictional extrapolation of General Lemay, represented in the film *Dr Strangelove*, sent many a shiver down American spines. The Cuban missile crisis in 1962 also provided a real-life example of why the fate of the world should rest on cool heads and calculating minds. Rightly or wrongly, as early as the 1950s many inside and outside of government had concluded that military affairs had become too dangerous to be left to the military. But who else was there? The political leaders did not seem able to rise to the challenge. The brass might grumble at the prospect of military decision-making falling into the hands of bloodless civilians, but to others this location promised relief. Among the civilians the economists, appreciated now for their dispassion and detachment, seemed the most attractive candidates. In constructing policy toward Cuba the rumpled game theorist and Harvard professor Thomas Schelling (Nobel Prize 2005) seemed a much preferred alternative to Dr Strangelove.
Economists entered the highest reaches of military strategy by several routes. After the National Security Act of 1948 in the United States provided for establishment of a National Security Council, economists began to take important positions on the Council staff. The Harvard economist Carl Kaysen became Deputy National Security Adviser in the Kennedy administration and Walt Rostow from the Massachusetts Institute of Technology (MIT) became National Security Adviser to President Johnson. Robert McNamara, Secretary of Defense under both Kennedy and Johnson, while not a professional economist himself had a high regard for the economist's approach. He set up a systems analysis unit in the Pentagon under the economist Alain Enthoven that was intended to carry the economist's way of thinking throughout the military. The RAND Corporation was an important Cold War resource with advice flowing at the highest levels from such well-known economists as Henry Rowen, James Schlesinger, Stephen Enke and Charles Wolf. The academic discipline of the time also contributed a galaxy of stars who served as consultants and informal advisors to various parts of the legislative and executive branches. Thomas Schelling, Charles Hitch, Roland McKean and Martin Shubik are only a few of the most prominent. None of these considered himself a specialist exclusively on war but rather a general economist who could turn his attention, and his tools, to the particular problems presented by war. The appointments of James Schlesinger as Secretary of Defense and director of the Central Intelligence Agency, both in 1973, signalled the zenith of economics in the US defence establishment.
An intriguing characteristic of the economists involved with the Cold War was their unwillingness to hive off into a regularized subsection of the economics discipline, such as health or agricultural economics. Indeed, they often led a life of glamour and glory unknown to other economists before or

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
since. They operated near the apex of power, were cleared at the highest level of security, and dealt directly with urgent matters of life and death to the nation. Moreover, they were compelled to become multidisciplinary if they were to operate effectively and to become part of the informal community that contained government officials of all kinds, business leaders, scientists, and members of the other disciplines concerned with war, including history, law, philosophy, political science, sociology and others. They had to keep abreast of arcane technological developments and new weapons, learn about treaties, and comprehend the relations of the armed forces with society. They became part of the field known as ‘strategic studies’ (served by the London-based International Institute of Strategic Studies), students of international relations, and regular contributors to the mass media. They might see more of leaders in other defence-related fields, such as Bernard Brodie, Alistair Buchan, Hedley Bull, Harold Brown, Richard Garwin and Albert Wohlstetter, than they did of their disciplinary kin.

While some economists rose to positions of great influence within the defence and foreign policy establishment during the Cold War, others became involved in what was known sometimes as peace and conflict studies, or peace science. Prominent economists in this movement included Kenneth Boulding and Walter Isard. Their focus was particularly on costs of defence, causes of conflict, arms races, and arms control and disarmament. Their influence seems to have been confined mainly to their own community.

The landmark study in the defence economics literature of the cold war is *The Economics of Defense in the Nuclear Age* (1960) by C.J. Hitch and Roland McKean. This work codified much that was known already, made original contributions itself and pointed to research opportunities ahead. Two directions into which other economists moved were towards an understanding of arms races, led by the pioneering work of Lewis Richardson (*Arms and Insecurity*, 1960), Walter Isard and others, and towards an appreciation of the impact of defence expenditures on economic growth, led by Emile Benoit's *Defense and Economic Growth in Developing Countries* (1973).

To many in the home discipline of economics the defence economists of the 1950s, 1960s and 1970s must have seemed a strange breed, dancing to a different drummer from those with whom they had studied in graduate school. In fact the circumstances of these defence economists have had few parallels in the history of the discipline, and they did not last for long. Perhaps the debacle in Vietnam and the wind-down of the Cold War explain their loss of influence. Or perhaps they were not well equipped for the challenges of the 1980s when the Reagan administration in the United States and the Thatcher government in Britain talked of defeating the enemy rather than avoiding conflict. Moreover, the warfare that loomed was embedded in ethnic, religious and historical grievances about which economists knew rather little.

Birth of a sub-discipline

As economists retreated from the pinnacle of power in the 1980s, down from cabinet and sub-cabinet positions to advisors, analysts and consultants, a new sub-discipline of ‘defence economics’ emerged. In the words of two of its founders it was concerned with such topics as ‘the analysis of alliance burden sharing, the effects of contract design on procurement, the impact of defence expenditures on economic growth, and the economic consequences of arms control treaties’ (Sandler and Hartley, 1995, p. 1). This new applied sub-discipline was comparable to the older agricultural and labour economics, and to such

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
newer ones as health and environmental economics. Defence economists adhered to the core theory accepted in the mother discipline, and they applied approved analytical tools to problems that were perceived to be susceptible to the economists' methods. What distinguished these applied fields was knowledge of institutional facts in particular areas and some attention to policy issues prominent there. Sandler and Hartley, authors of a survey of the literature in the mid-1990s, described the field as follows: 'Defence economics involves the application of economic reasoning and methods to the study of defence-related issues. Defence economics differs from other fields of economics in at least three ways: (1) the set of agents studied (e.g. defence contractors, branches of the military); (2) the institutional arrangements of the defence establishment (e.g. procurement procedures); and (3) the set of issues investigated' (Sandler and Hartley, 1995, p. xi). They suggest that economists explore four 'basic economic problems': allocative efficiency, public choice considerations (why elected officials and bureaucrats behave the way they do), the distributive implications of defence decisions, and stability issues, including paths after shocks. A specialized journal, *Defense and Peace Economics*, was founded in 1990 to weld together two strands that during the cold war had gone in rather different directions. Sandler and Hartley recognize that recent events have shifted attention away from long-standing concentration on superpower confrontation to a variety of new issues such as the consequences of the break-up of the socialist world, the increased number of regional conflicts such as the Gulf War of 1991 and subsequent Iraq War that followed the relaxation of superpower hegemony, the proliferation of relatively inexpensive conventional weapons (some second-hand), the responsibilities of the arms exporters, the design of arms control treaties, and the macroeconomic implications of military downsizing. In the 1990s there was even talk of a peace dividend – an idea that seems merely poignant in the 21st century. Terrorism, guerilla warfare, the security implications of economic coalitions, and burden sharing within alliances remained topics on which economists found they had interesting things to say. Some issues, like procurement practices, are common to all parts of the public sector but receive disproportionate attention in defence economics because of the magnitude of defence expenditures.

A textbook for an undergraduate course on 'The Economics of War' published in 2006 (Poast, 2006) lists aspects of international conflict that may be understood better using the economist's tools: how to achieve mobilization or disarmament; the special problems presented by recruitment of a military labour force and weapons procurement; conflict in developing countries, the small-arms trade, peacekeeping; and the dilemmas of terrorism and the proliferation of the weapons of mass destruction in the 21st century.

John Maynard Keynes remarked in 1930 that he looked forward to the day when economists would be 'thought of as humble, competent people, on a level with dentists' (Keynes 1930, p. 332). It appears that in the 21st century economists concerned with the study of war have gained this status. It is now well accepted that war, like all human activity, requires the recognition of scarcity and the need to make choices based on forgone opportunities. This is the domain of economics. Defence economists stand ready to advise on these allocative decisions and to remind policymakers of the applicability of such concepts as externalities and public goods. The heroic years of defence economics are almost certainly gone for ever; the economists are today, as they say, on tap but not on top. Nevertheless, their usefulness remains, even if at a more modest level than before. The study of war is now an accepted part of economics, assigned to its own sub-field and dependent heavily on the tools and methods of public economics. In its current posture economics is less likely to find a cure for conflict than to make it more efficient and its prevention less costly. In a world full of shortages and sufferings this is no small

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
accomplishment.

See Also

- defence economics

Bibliography

Angell, N. 1910. *The Great Illusion: A Study of the Relation of Military Power in Nations to their Economic and Social Advantage*, 3rd edn. New York and London: G.P. Purnam's Sons. 1912.

Barber, W. 1991a. British and American economists and attempts to comprehend the nature of war, 1919–20. In Goodwin (1991).

Barber, W. 1991b. From the economics of welfare to the economics of warfare (and back) in the thought of A.C. Pigou. In Goodwin (1991).

Benoit, E. 1973. *Defense and Economic Growth in Developing Countries*. Boston: D.C. Heath.

Biddle, J.E. and Samuels, W.J. 1991. Thorstein Veblen on war, peace, and national security. In Goodwin (1991).

Black, R.D.C., Coats, A.W. and Goodwin, C.D. 1973. *The Marginal Revolution in Economics: Interpretation and Evaluation*. Durham, NC: Duke University Press.

Bush, V. 1945. *Science, the Endless Frontier*. Washington: US Government Printing Office.

Clark, J.M. 1931. *The Costs of the World War to the American People*. New Haven: Yale University Press.

Davis, C.M. 1991. Marxist and Soviet defense economics, 1848–1927. In Goodwin (1991).

De Marchi, N. 1991. League of Nations economists and the ideal of peaceful change in the decade of the thirties. In Goodwin (1991).

Galbraith, J.K. 1981. *A Life in our Times: Memoirs*. Boston: Houghton Mifflin.

Goodwin, C.D. (ed.) 1991. *Economics and National Security: A History of their Interaction*. Durham, NC: Duke University Press.

Hitch, C.J. and McKean, R.N. 1960. *The Economics of Defense in the Nuclear Age*. Cambridge, MA:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Harvard University Press.

Hofstadter, R. and Metzger, W.P. 1955. *The Development of Academic Freedom in the United States*. New York: Columbia University Press.

Keynes, J.M. 1920. *The Economic Consequences of the Peace*. New York: Harcourt Brace.

Keynes, J.M. 1930. Economic possibilities for our grandchildren. In *The Collected Writings of John Maynard Keynes*, ed. D. Moggridge. New York: Macmillan, 1972.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kindleberger, C.P. 1991. *The Life of an Economist: An Autobiography*. Cambridge, MA: Basil Blackwell.

Klausner, S.Z. and Lidz, V.M. 1986. *The Nationalization of the Social Sciences*. Philadelphia: University of Pennsylvania Press.

Leonard, R. 1991. War as a ‘simple economic problem’: the rise of an economics of defense. In Goodwin (1991).

Lodewijks, J. 1991. Developing economies, and national security policy. In Goodwin (1991).

Lowry, S.T. 1991. Preclassical perceptions of economy and security. In Goodwin (1991).

Malthus, T.R. 1826. *An Essay on the Principle of Population*, 6th edn. Vols 2 and 3 of *The Works of Thomas Robert Malthus*, ed. E.A. Wrigley and D. Souden. London: Pickering and Chatto, 1986.

Mill, J.S. 1848. *Principles of Political Economy*. Vols 2 and 3 of The Collected Works of John Stuart Mill, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Mirowski, P. 2002. *Machine Dreams: Economics Becomes a Cyborg Science*. New York: Cambridge University Press.

Pigou, A.C. 1921. *The Political Economy of War*. London: Macmillan.

Poast, P. 2006. *The Economics of War*. New York: McGraw Hill.

Rae, J. 1834. *Statement of Some New Principles on the Subject of Political Economy*. ed. R. Warren James. Toronto: University of Toronto Press, 1965.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Richardson, L. 1960. *Arms and Insecurity*. Pittsburgh: Boxwood Press.

Sandilands, R. 1990. *The Life and Political Economy of Lauchlin Currie*. Durham, NC: Duke University Press.

Sandler, T. and Hartley, K. 1995. *The Economics of Defense*. Cambridge: Cambridge University Press.

Silberner, E. 1946. *The Problem of War in Nineteenth Century Economic Thought*. Princeton: Princeton University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Clarendon Press, 1976.

Speelman, P.J. 2002. *Henry Lloyd and the Military Enlightenment of Eighteenth-Century Europe*. Westport, CT: Greenwood Press.

Stein, H. 1969. *The Fiscal Revolution in America*. Chicago: University of Chicago Press.

Weintraub, E.R. 1992. *Toward a History of Game Theory*. Durham, NC: Duke University Press.

How to cite this article

Goodwin, Craufurd D. "war and economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000136>
doi:10.1057/9780230226203.1816

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Warburton, Clark (1896 – 1979)

L. Yeager

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

monetarism; monetary disequilibrium; quantity theory of money; real business cycles; Warburton, C.

Article

American economist; pioneer, before Milton Friedman, in research later labelled ‘monetarist’, and a critic of Keynesianism during the years when that doctrine was crowding out attention to money.
Warburton was born on 27 January 1896 near Buffalo, New York, and died on 18 September 1979 in Fairfax, Virginia. After overseas military service during the First World War, he earned bachelor's and master's degrees from Cornell University. He published his 1932 Columbia Ph.D. dissertation as *The Economic Results of Prohibition*. He held teaching positions in India and the United States in the 1920s and early 1930s and worked at the Brookings Institution from 1932 to 1934, coauthoring *America's Capacity to Consume*. He then joined the newly organized Federal Deposit Insurance Corporation. Although routine FDIC work consumed much of his time (as his files reveal), he still managed to publish over 30 papers on monetary economics, most of them empirically oriented, from 1943 to 1953. Altered FDIC policy then impeded his research and publication until about 1962, when he took a brief leave to serve with the Banking and Currency Committee of the US House of Representatives. He was elected President of the Southern Economic Association for 1963–4. After retiring from the FDIC in 1965, he taught briefly at the University of California, Davis.
Warburton originally accepted a ‘real’ theory of the business cycle, but scrutiny of statistical and qualitative history changed his views. Using quarterly as well as annual data, he found that deviations from trend of the quantity of money generally preceded turning points in business conditions (and velocity deviations followed). While accepting a quantity-of-money theory of the price level in the long run, he recognized how elements of wage and price stickiness cause monetary disturbances to impinge on output first; he espoused a ‘monetary disequilibrium theory’ (which, despite its venerability, has ironically been mislabelled ‘Keynesian’ in recent years). He understood that disequilibrium does not necessarily imply irrational behaviour by individuals.
Warburton emphasized the role of money and inadequate monetary policy in the Great Depression of the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1930s. He continued to criticize the Federal Reserve for deficiencies in its economic theory and research and, in particular, for relying on interest rates in deciding and implementing policy. He believed that pure fiscal policy, unsupported by changes in the quantity of money, is ineffective as a tool of demand management. Sceptical of the authorities' ability to fine-tune the economy, he recommended a policy of steady growth in the quantity of money at a moderate rate appropriate to trends in the labour force, productivity and velocity.

For Warburton, monetarism was an interpretation reached inductively, not a comprehensive ideology. (So far as any ideology came across in conversations, it was a rather conventional New Deal reformism or liberalism with humanitarian underpinnings.)

Nineteen of Warburton's papers dating from 1945 to 1953 are reprinted, along with a new introduction, in *Depression, Inflation, and Monetary Policy* (1966). Up to his death, Warburton pursued research not only in substantive economics but also in the history of monetary doctrines. These continuing interests are manifest in his last article (published posthumously in *History of Political Economy*, 1981) and in voluminous manuscripts now deposited in the library of George Mason University, Fairfax, Virginia. Plans exist for editing and publishing much of this material.

See Also

- quantity theory of money

Selected works

1934. (With M. Leven and G. Moulton.) *America's Capacity to Consume*. Washington, DC: Brookings Institution.

1966. *Depression, Inflation, and Monetary Policy: Selected Papers, 1945–1953*. Baltimore: Johns Hopkins Press.

1981. Monetary disequilibrium theory in the first half of the twentieth century. *History of Political Economy* 13, 285–99.

How to cite this article

Yeager, L. "Warburton, Clark (1896–1979)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000027>
doi:10.1057/9780230226203.1817

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Warming, Jens (1873 – 1939)

P. Nørregaard Rasmussen

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

multiplier; Warming, J.

Article

Jens Warming was born on 9 December 1873 and died on 8 September 1939.
After graduating in law, Warming took up economics. In 1919 he became Professor in Applied Statistics at the University of Copenhagen, following more informal attachments to the university. Along with his teaching he produced a number of books describing empirically a wide variety of aspects of the Danish economy. In a way he created the field ‘applied statistics’ as an academic discipline in Denmark. He not only presented figures, but he surrounded them with reasoning, sometimes naive and not very well articulated, often full of wisdom. One example is his warning of the danger of overfishing because no rent is collected (Warming, 1931b). Another most important example is his discovery of the multiplier process, which he presented as early as 1928 and again in 1929–30 and 1931. These important contributions in economic theory were quite often formulated in a somewhat odd way, and they certainly did not attract his fellow economists in Scandinavia.
Warming's formulation of the multiplier runs as follows: assume a closed economy (an assumption he later modified) and consider an investment of, say, 100 units in a railway. This creates an income of equal size, part of it appearing as an increase in savings, but another part as consumption, that latter creating new incomes. This process, he argues, will go on until voluntary savings will increase, so that the newly-constructed railway ‘gets an owner’ (1929–30). This clearly means that the total voluntary savings in the end equal the impulse, that is, 100 units. An investment will ‘finance itself’, as he argued at length. However, it does not seem as if Warming was considering the multiplier as part of a more general theory of employment.

Selected works

1924. *Valutaspørgsmålet* [The problem of foreign exchange].

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1928. Beskæftigelsesproblemet [The employment problem]. *Gads Danske Magasin* 22, 609–721.

1929–30. *Danmarks Erhvervs- og Samfundsliv* [Denmark's economic and social life]. Copenhagen.

1931a. Tilpasning [Adaptation]. *Gads Danske Magasin* 25, 481–505.

1931b. Aalegaardsretten. *Nationalkonomisk Tidsskrift* 69, 151–62.

1932. International difficulties arising out of financing of public works during depression. *Economic Journal* 42, 211–24.

1939. *Danmarks Erhvervs- og Samfundsliv* [Denmark's economic and social life]. 2nd edn. Copenhagen: G.E.C. Gad. (This edition was never completed.)

Bibliography

Boserup, M. 1969. A note on the prehistory of the Kahn multiplier. *Economic Journal* 79, 667–9.

Gelting, J.H. 1964. Jens Warming. *Nationalkonomisk Tidsskrift*.

Hegeland, H. 1954. *The Multiplier Theory*. Lund: Gleerup.

Kahn, R.F. 1931. The relation of home investment to unemployment. *Economic Journal* 41, June, 173–98.

How to cite this article

Rasmussen, P. Nørregaard. "Warming, Jens (1873–1939)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000029>
doi:10.1057/9780230226203.1818

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Washington Consensus

John Williamson
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The original purpose of the term ‘Washington Consensus’ was to distinguish between a number of policies that remain the subject of partisan controversy and a group of policies that were thought to be consensual in the post-1989 world. After its creation in 1989 the term acquired at least two more meanings. Some used it to describe the policies of the IMF and World Bank, which came to embrace not only institutional reform and a concern with governance but also the two-corners solution for exchange rates and capital account convertibility. Others used the term as a synonym for laissez-faire.

Keywords

balance of payments; capital account liberalization; consensual policies; corruption; demographic transition; deregulation; development economics; East Asia; financial liberalization; fiscal discipline; foreign direct investment; globalization; governance; import-substituting industrialization; industrial policy; inflation; informal economy; institutions; intermediate exchange-rate regime; International Monetary Fund; laissez-faire; Mont Pèlerin Society; neoliberalism; partisan policies; privatization; property rights; public expenditure; redistribution of income; tax reform; trade liberalization; two-corners solution; Washington Consensus; World Bank

Article

The term ‘Washington Consensus’ was coined in 1989. The first written usage was in my background paper (Williamson, 1990) for a conference that the Institute for International Economics convened in order to examine the extent to which the old ideas of development economics that had governed Latin American economic policy since the 1950s were being swept aside by the set of ideas that had long been accepted as appropriate within the Organisation for Economic Co-operation and Development (OECD). In order to try to ensure that the country papers for that conference dealt with a common set of issues, I made a list of ten policies that I thought more or less everyone in Washington would agree were needed

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
more or less everywhere in Latin America, and labelled this the ‘Washington Consensus’. Little did it occur to me that 17 years later I would be asked to write about a term that had become the centre of fierce ideological controversy. On the contrary, I thought of the Washington Consensus as distinguishing between (*a*) what had become consensual and (*b*) what was likely to remain partisan, issues such as income distribution, capital account convertibility, the usefulness of incomes policy, the need to eliminate indexation, the size of the public sector, and the priority to be given to population control and environmental preservation.
The set of ‘consensual’ policies was:

1. 1. *Fiscal discipline*. This was in the context of a region where almost all the countries had run large deficits, which had led to balance of payments crises and high inflation that hit mainly the poor because the rich could park their money abroad.
2. 2. *Reordering public expenditure priorities*. This suggested switching expenditure in a pro-poor way, from things like indiscriminate subsidies to basic health and education.
3. 3. *Tax reform*: constructing a tax system that would combine a broad tax base with moderate marginal tax rates. This need not imply redistribution of income to the rich if the broadening of the tax base focuses on eliminating loopholes that are exploited by those who can afford to employ tax lawyers.
4. 4. *Liberalizing interest rates*. This was subsequently formulated in a broader way as financial liberalization, to cover also policies like bank privatization and allowing financial institutions to determine the allocation of credit.
5. 5. *A competitive exchange rate*. It was asserted (though this may not have been accurate reporting of the Washington scene) that there was a consensus in favour of ensuring that the exchange rate would be competitive.
6. 6. *Trade liberalization*. It was acknowledged that there was a difference of view about how fast trade should be liberalized, but it was asserted to be widely held that trade needed to be liberalized and countries stood to gain by outward-oriented policies.
7. 7. *Liberalization of inward foreign direct investment*. This specifically did not include comprehensive capital account liberalization, which did not command a consensus in Washington.
8. 8. *Privatization*. This was the one area in which what had originated as a neoliberal (Thatcherite) idea won broad acceptance. We have since been made very conscious that it matters a great deal how privatization is undertaken: it can be a highly corrupt process that transfers assets to a privileged elite for a fraction of their true value, but the evidence that it ultimately brings net benefits is quite strong.
9. 9. *Deregulation*. This focused specifically on easing barriers to entry and exit, not on abolishing regulations designed for safety or environmental reasons.
10. 10. *Property rights*. This was primarily about providing the informal sector with the ability to gain property rights at an acceptable cost.

The term ‘Washington Consensus’ proved controversial right from the start. Both reformers and critics took it to imply a belief that policy reforms had been imposed by Washington institutions like the International Monetary Fund (IMF) and the World Bank. The reformers resented the implication that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
they had been following orders rather than implementing policies that they had concluded were needed in their countries. The critics took it as confirmation of their darkest fears. It might have been better if the revolution in economic policy for development had been called something else, but the fact that there was such a revolution is clear. In the 1960s mainstream development thinking held that inflation was at worst harmless and at best a method of extracting resources for investment. Orthodox thought advocated import-substituting industrialization rather than export expansion through globalization. Governments sought to expand the industrial sector by creating more state firms rather than by creating a market-friendly environment. All this changed in the late 1980s, perhaps aided by the collapse of the Communist alternative model, which was widely welcomed in Washington. That was what the Washington Consensus was about.

As time progressed the term came to mean at least three different things. Some people stuck to the original concept. Others used it to mean the evolving policies of the IMF and the World Bank, presumably on the ground that originally it had pretty much described the policies they advocated in 1989. Policy evolved importantly in at least four dimensions. First, in common with development thinking in general, Washington came to understand the importance of institutions in promoting development. Thus, first Naím (1994) and subsequently Burki and Perry (1998) and Kuczynski and Williamson (2003) argued that one needed to supplement the prescriptions of the Washington Consensus with institutional development. Second, the Washington institutions came to place a great emphasis on governance, especially on avoiding corruption. Third, in the IMF in particular opinion moved strongly away from the concern with a competitive exchange rate and the implication of the desirability of an intermediate regime toward the ‘two-corners solution’ (which argues that to avoid speculation one should have either a firmly fixed or a freely floating exchange rate, but nothing in between). Fourth, again especially in the IMF, there emerged a strong preference for rapid capital account liberalization. Some of those (like the author) who had been very attached to the original concept of the Washington Consensus found these last two doctrines alien.

The third concept of the term involved an even more radical change from the original, though this is the version that appears to have been widely used by critics. It refers to a programme often described as ‘neoliberal’ (a term that was at one time used to describe the agenda of the Mont Pèlerin Society, of which Milton Friedman and Friedrich von Hayek were prominent members). In addition to the items in the original list, the programme includes low taxes, a minimal state that denies having any responsibility for income distribution, either a currency board or freely floating exchange rates, and rapid liberalization of the capital account. It has even been suggested that the Washington Consensus implies a belief that all markets are perfectly competitive so that neoclassical economic analysis is literally true always and everywhere and government action is always a mistake (Stiglitz, 2006, p. 24). Despite the name, those who used the term in this way regarded it as unnecessary to establish that these attitudes characterized large parts of Washington. Personally I do not recall having met anyone in Washington who subscribes to these rather bizarre beliefs.

The most controversial elements of the Washington Consensus proved to concern microeconomic liberalization. Points 4 and 6–9 of my original list pointed in that direction, as did especially the last alternative interpretation of the phrase alluded to above. In the wake of the collapse of communism I overestimated the extent to which the desire for an active government role in managing the economy would be replaced by the acceptance of market mechanisms. The experience of East Asia was frequently invoked to question whether a commitment to laissez-faire had really been the key to the region's fast

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

growth. On the contrary, many critics of the Washington Consensus argued that the success of the East Asian countries was due to their governments having implemented industrial policies. This runs into the objection that at least one of the East Asian success stories (Hong Kong) had developed under the closest to a laissez-faire system that the world has ever seen. Why should one attribute the success of Korea and Taiwan to what distinguishes them from Hong Kong and to some extent Singapore rather than to what they have in common (for example, macro stability, high savings, export orientation, good education and an early demographic transition)? The old questions of market versus state, or even the merits of industrial policy, have not been settled by these arguments.

See Also

- globalization

Bibliography

Burki, J. and Perry, G.E. 1998. *Beyond the Washington Consensus: Institutions Matter*. Washington: World Bank.

Kuczynski, P.-P. and Williamson, J. 2003. *After the Washington Consensus: Restarting Growth and Reform in Latin America*. Washington: Institute for International Economics.

Naím, M. 1994. Latin America: the second stage of reform. *Journal of Democracy* 5, 32–48.

Stiglitz, J.E. 2006. *Making Globalization Work*. New York: Norton.

Williamson, J. 1990. What Washington means by policy reform. In *Latin American Adjustment: How Much Has Happened?* ed. J. Williamson. Washington: Institute for International Economics.

How to cite this article

Williamson, John. "Washington Consensus." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000138>
doi:10.1057/9780230226203.1819

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

wavelets

James B. Ramsey
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Wavelets provide a flexible basis for representing a signal that can be regarded as a generalization of Fourier analysis to non-stationary processes, or as a filter bank that can represent complex functions that might include abrupt changes in functional form, or signals with time varying frequency and amplitude. Of greatest import for economic analysis is the orthogonal deconstruction of a signal into time scale components that allow economic relationships to be analysed time scale by time scale and then re-synthesized.

Keywords

capital asset pricing model; denoising; discrete wavelet transform; filter banks; forecasting; Fourier analysis; Fourier transforms; Heisenberg uncertainty principle; multiresolution analysis; projection pursuit; regime shifts; serial correlation; shrinkage; threshold models; waveform dictionaries; wavelet packets; wavelets

Article

Wavelets provide a powerful tool of analysis for economics and finance, as well as for scientists in a wide variety of fields, such as signal processing, medical imaging, data compression and geology. One interpretation of wavelets is that they are a collection of functions that provide a basis for the representation of test functions that may be complicated with localized shocks, have abrupt changes in functional form, or are signals with time varying frequency and amplitude.

Another interpretation is that wavelets are a generalization of Fourier analysis in which stationarity of the time series is no longer critical and localization of a signal can be achieved. In this light, to borrow a great insight by Strang (see Strang and Nguyen, 1996), Fourier analysis is best at representing functions that are composed of linear combinations of stationary inputs, but wavelets are like musical notation in that each note is characterized by its frequency, its position in time, and its duration.

A further interpretation of wavelets is that of a filter bank, so that different classes of wavelet functions are generated by prescribing different banks of filters. Filter banks can achieve results that are not possible with a single filter (Strang and Nguyen, 1996). Yet another interpretation is that of a decomposition of a signal in terms of different time scales, an interpretation that is at the heart of much economic analysis as represented by the long-standing notions of the ‘short, medium and long runs’ and is fundamental to the concept of the ‘business cycle’.

At this time there is a vast literature on wavelets in mathematics, statistics, and various branches of engineering, but relatively little in economics, although that situation is changing fast. An introduction to the economic literature that is highly recommended is Gencay, Selcut and Whitcher (2002). This is the most comprehensive and detailed coverage with numerous descriptions of economic applications and discussions of the statistical properties of the wavelet estimators. Bruce and Gao (1996) discuss the properties of wavelets and give instructions for calculating wavelets using S-Plus; Chui (1992), Percival and Walden (2000), and Strang and Nguyen (1996) develop the mathematics at a moderate level of difficulty and discuss the statistical properties of wavelets. Silverman and Vassilicos (1999) provide interesting examples of the applications of wavelets and further discussions of the statistical properties (see also Ramsey, 1999a). Two lower-level introductions for economists are Crowley (2005) and Schleicher (2002).

Two informative examples

Let **X** represent an N dimensional vector of observations on a time series or a function $f(\cdot)$ evaluated at the discrete points t=1,2,…N. One may consider an orthonormal transformation from **X** to an N dimensional vector **W**, with elements, W_n, n=0,1,…N–1, where **W** is generated from an N×N dimensional orthonormal matrix W; $\mathbf{W}=W\mathbf{X}$ and $I_N=W^TW$. Let $N=2^J$ for some integer J. This assumption, while inessential, is analytically convenient and is useful for defining an efficient algorithm for evaluating the wavelet coefficients. The N^2 elements of the transformation matrix W are the filter elements to be defined below and the N elements of the vector **W** are the wavelet coefficients that are given by the inner product $W\mathbf{X}$.

The wavelet defined here is known as the discrete wavelet transform (DWT). By choosing an orthonormal transform it is immediate that the modulus of **W** is the same as that of **X**; that is, $||\mathbf{X}^2||=||\mathbf{W}^2||$, where $||X^2||$ is given by $\mathbf{X}^T\mathbf{X}=\sum x_i^2$, and where w_n^2 is the energy contributed by the nth wavelet coefficient.

The power of wavelet analysis lies in the choices made for the components of W. Two examples illustrate the choices that can be made and indicate the scope that a wavelet analysis offers. (I am indebted to Percival and Walden, 2000, for the examples to follow.) The first example is the oldest and simplest function used to generate a wavelet transformation. For the sake of simplicity of exposition, choose $N=2^J=2^4=16$. The notation, $W_{i,*}^T$ indicates the ith row of the matrix W, i=0,1,…N–1. The function used to generate the elements of W for the DWT is the Haar function.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$W_{0,*}^T = \left\{-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0_1, \ldots 0_{14}\right\},\ W_{8,*}^T = \left\{-\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0_1, \ldots 0_{12}\right\} W_{12,*}^T = \left\{-\frac{1}{\sqrt{8}}{}_1, \ldots -\frac{1}{\sqrt{8}}{}_4, \frac{1}{\sqrt{8}}{}_1, \ldots \frac{1}{\sqrt{8}}{}_4, 0_1, \ldots 0_8\right\} W_{14,*}^T = \left\{-\frac{1}{4}{}_1, \ldots -\frac{1}{4}{}_8, \frac{1}{4}{}_1, \ldots \frac{1}{4}{}_8\right\},\ W_{15,*}^T = \left\{\frac{1}{4}{}_1, \ldots \frac{1}{4}{}_{16}\right\}$$

(1)

The unshifted filters are $W_{0,*}^T, W_{8,*}^T, W_{12,*}^T, W_{14,*}^T, W_{15,*}^T$. The remaining rows of *W*, are circularly shifted versions of the filter terms defined above. We have:

$$W_{1,*}^T = T^2 W_{0,*}^T = \left\{0, 0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0_1, \ldots 0_{12}\right\} W_{2,*}^T = T^4 W_{0,*}^T = \left\{0, 0, 0, 0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0_1, \ldots 0_{10}\right\} W_{9,*}^T = T^4 W_{8,*}^T = \left\{0, 0, 0, 0, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0_1, \ldots 0_8\right\} W_{13,*}^T = T^8 W_{12,*}^T = \left\{0_1, \ldots 0_8, -\frac{1}{\sqrt{8}}{}_1, \ldots -\frac{1}{\sqrt{8}}{}_4, \frac{1}{\sqrt{8}}{}_1, \ldots \frac{1}{\sqrt{8}}{}_4\right\}$$

(2)

The filter $W_{15,*}^T$ yields the 'scaling coefficient', and the remaining filters, $W_{i,*}^T$, yield the wavelet coefficients. From the above equations one sees that the Haar filters involve first differences between scale averages at non-overlapping intervals. At the lowest scale, 2^{j-1} for j=1, the first differences are between adjacent observations, at the next scale level, 2^{j-1} for j=2, the differences are between adjacent pairs of observations, at the scale 2^{j-1} for j=3 the differences are between groups of four terms, and so on. At the highest scale, J=4, we have two filters of the data, a first difference between the first and last eight of the observations, and an average over the full set of observations. The latter is a 'father' wavelet transform and the remaining rows of *W* are the 'mother' wavelet transforms.

In order to gain further insight, consider a Daubchies wavelet designated as D(4) (see Daubchies, 1992), which is a member of a sequence of discrete wavelet filters in which the Haar is D(2). Define $Y_t \equiv aX_t + bX_{t-1}$ and form the backward second discrete difference, $Y_t^{(2)}$ by:

$$Y_t^{(2)} \equiv Y_t^{(1)} - Y_{t-1}^{(1)} = Y_t - 2Y_{t-1} + Y_{t-2}.$$

(3)

For a particular choice of 'a' and 'b', the nth D(4) wavelet can be written as:

$$W_n = Y_{2n+1} - 2Y_{2n} + Y_{2n-1} = aX_{2n+1} + (b-2a)X_{2n} + (a-2b)X_{2n-1} + bX_{2n-2} = h_0 X_{2n+1} + h_1 X_{2n} + h_2 X_{2n-1} + h_3 X_{2n-2}.$$

(4)

Imposing orthonormality {and some other conditions to ensure uniqueness}, enables one to derive the values for h_i, i=0, 1, 2, 3, namely:

$$h_0 = \frac{1-\sqrt{3}}{4\sqrt{2}},\ h_1 = \frac{-3+\sqrt{3}}{4\sqrt{2}},\ h_2 = \frac{3+\sqrt{3}}{4\sqrt{2}},\ h_3 = \frac{-1-\sqrt{3}}{4\sqrt{2}}.$$

Repeating the exercise above with N=2^J=16 observations, we have:

$$W_{0,*}^T = \{h_1, h_0, 0_1, \ldots 0_{12}, h_3, h_2\},\quad W_{1,*}^T = \{h_3, h_2, h_1, h_0, 0_1, \ldots 0_{12},\}$$

where $W_{1,*}^T = T^2 W_{0,*}^T$, orthonormality requires that $\|W_{0,*}\|^2 = h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1$ and $\left\{W_{0,*}^T W_{1,*}\right\} = h_0 h_2 + h_1 h_3 = 0$. $W_{15,*}^T$ applied to the time series yields the mean.

A wavelet filter of length L must satisfy at a minimum for all nonzero integers n the following conditions.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\sum_{l=0}^{L-1} h_l = 0, \quad \sum_{l=0}^{L-1} h_l^2 = 1$$
(5)

$$\sum_{l=0}^{L-1} h_l h_{l+2n} = \sum_{l=-\infty}^{\infty} h_l h_{l+2n} = 0.$$

For the Haar wavelet, D(2), and the Daubchies D(4) wavelet L is respectively 2 and 4.

Wavelets and multiresolution analysis (MRA)

One can approach the definition of wavelets from a related perspective that indicates the similarities to and differences from Fourier transforms. In both cases, one is considering a projection of a signal onto a set of basis functions for the space containing the test function. In the case of the Fourier transform the basis functions are rescalings of the fundamental frequency, for example, $\left\{e^{-in\omega_0}\right\}$ where ω_0 is the fundamental frequency and *n* provides the scaling. In this expression there is no resolution in the time domain, and the signal is assumed to be stationary. In contrast, the wavelet-generating functions are defined over very general spaces and each function is compact. One has the recursive relationship:

$$g_{s,k}(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-k}{s}\right)$$
(6)

where *s* indicates the scale of the function, *k* is the location index; the term $\sqrt{s}$ ensures that the norm of $g(\cdot)$ is 1. The projection of the signal onto the scalable function $g(\cdot)$ depends on two parameters; *s*, which defines the time scale and implicitly designates a relevant range of frequencies, and *k*, which indicates the centre of location of the projection. The compactness of $g(\cdot)$ together with the time index *k* implies that the analysis of a time series is essentially local at each scale, whereas the Fourier analysis is essentially global. There is considerable latitude in the choice of the function $g(\cdot)$, or in the choice of the filters that generate the functions $g(\cdot)$. Desirable criteria include symmetry, smoothness and orthogonality. Whether one begins by specifying the properties of a basis function $g(\cdot)$ or one begins by specifying the properties of a filter $\{h_i\}$, the process generates two related classes of wavelet transforms, the 'father' wavelet that yields the 'scaling' coefficients and the 'mother' wavelets that yield the detail coefficients.
One can link the filter coefficients to the definition of the father and mother wavelets and link the father and mother transforms themselves by noting that, for a given sequence of low pass filter coefficients, $l(k)$, and the corresponding high pass filter coefficients, $h(k)$, one solves for $\Phi(t)$, father, and $\psi(t)$, mother, from:

$$\Phi(t) = \sqrt{2}\sum_{k=0}^{N} l(k)\Phi(2t-k)$$
(7)

$$\Psi(t) = \sqrt{2}\sum_{k=0}^{N} h(k)\Phi(2t-k).$$

For the Haar example above, the filter coefficients are: $l(k) = \left\{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\}$ and $h(k) = \left\{-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\}$.
From these equations, one derives the scaling, or 'smooth' coefficients and the wavelet or 'detail' coefficients of the function $f(\cdot)$ by the integrals,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$s_{J,k} = \int f(t)\Phi_{J,k}dt$$
(8)

$$d_{j,k} = \int f(t)\Psi_{j,k}dt, \quad j = 1, 2, \ldots J$$
(9)

where $\Phi_{J,k}$ and $\psi_{J,k}$ are the scaled and translated versions of Φ and ψ defined in equation (7). The function $f(t)$ can be synthesized by the equations:

$$f(t) = \sum_k s_{J,k}\Phi_{J,k}(t) + \sum_k d_{J,k}\Psi_{J,k}(t) + \ldots \sum_k d_{j,k}\Psi_{j,k}(t) \ldots + \sum_k d_{1,k}\Psi_{1,k}(t)\, f(t) = S_J + D_J + D_{J-1} + \ldots D_j + \ldots D_1 S_J = \sum_k s_{J,k}\Phi_{J,k}, \quad D_j = \sum_k d_{j,k}\Psi_{j,k}, \; j = 1, 2, \ldots J\, f(t) \approx S_{j-1} = S_j + D_j.$$
(10)

An easy way to visualize the above scale and locational decomposition of the signal is as a series of maps of ever greater detail as elements of D_j are added; S_J provides a smooth outline, D_J adds the highest scale detail; and the D_j add ever more detail as j decreases. One can approximate the function $f(t)$ by truncating the expansion at some j, $1<j<J$. This is known as a multiresolution analysis (MRA), which can yield enormous data compression by representing the function $f(t)$ with relatively few coefficients.

There are many choices for the basis function, or alternatively choices for the filter banks that provide a great deal of richness for the wavelet approach. Two generalizations of note are wavelet packets and an exploratory technique known as waveform dictionaries. The former analyses signals by basis functions that differ by location and scale as for wavelets, but also by an oscillation index; wavelet packets are most useful in representing time series that have short term, localized oscillations (Bruce and Gao, 1996). Waveform dictionaries (Mallat and Zhang, 1993) provide a modification to wavelet analysis. The basic function providing the basis is a function $g(.)$ defined by:

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{i\omega t}, \quad \gamma = (s, u, \omega).$$
(11)

The function $g_\gamma(t)$ has norm one, has scale s, and the time scale energy is centred at u and proportional to s. The Fourier transform of $g_\gamma(t)$ has its frequency energy centred at ω and is proportional to $1/s$. The dictionary of functions $g_\gamma(t)$ illustrate a very important principle of these transforms. Improved resolution in the time domain reduces resolution in the frequency domain and vice versa; this is a version of the Heisenberg uncertainty principle.

Applications in economics and finance

All analytical procedures can be assessed on the basis of their contribution in four categories: provide estimators in novel situations; improve efficiency or reduce bias; enhance robustness to modelling errors; or provide new insights into the data-generating processes. Wavelets have provided benefits in all these categories.

One advantage of the waveform dictionary approach indicated in eq. (11) is that the researcher need not prejudge the presence of frequency components as well as the occurrence of short-term shocks. The process is exploratory and projection pursuit methods can be utilized to isolate local and global characteristics. Waveform dictionaries have been used as an exploratory tool in the analysis of financial and foreign exchange data (Ramsey and Zhang, 1996; 1997). In the analysis of daily stock-price data and tic-by-tic exchange rate data, there was no strong structural evidence for any frequency, but there was weak evidence for frequencies that appeared and disappeared or that waxed and waned in strength. Most of the power was summarized in terms of time-localized bursts of activity. The results in both papers indicated that, while for any given time period surprisingly few wavelet coefficients were needed to fit the data, the relevant coefficients varied randomly from period to period. Each burst was characterized by a rapid increase in amplitude and fast oscillation in frequency; in short, market adjustment processes seem to be characterized by a rapid increase in oscillation amplitude and frequency followed by a decay in frequency and amplitude; adjustment is neither smooth nor fast.

For a deep analysis of the scaling properties of volatilities and the relationship between risk and time scales, see Gencay, Selcut and Whitcher (2001; 2003). Another example in these references is the estimation of time varying Betas in the capital asset pricing model (CAPM). The analysis indicated that in the cases examined beta coefficients varied substantially over time, thereby modifying the structure of optimal investment strategies.

Wavelets have been instrumental in improving the robustness and efficiency of estimation in numerous examples (see Jensen, 1999; 2000) for efficiency gains and enhanced robustness of estimates for the fractional differencing parameter in long memory processes (see also Gencay, Selcut, and Whitcher, 2002). This reference is also useful for examples of estimation of covariance matrices and providing confidence intervals.

Wavelets have been successfully employed in situations not amenable to standard approaches – for example, forming estimators in testing for serial correlation of unknown form in panel models. As Hong and Kao (2004) state in their abstract: ‘This paper proposes a new class of generally applicable wavelet-based tests for serial correlation of unknown form in the estimated residuals of a panel regression model, where error components can be one-way or two-way, individual and time effects can be fixed or random, and regressors may contain lagged dependent variables or deterministic/stochastic trending variables.’

Ramsey and Lampart (1998a; 1998b) discovered that the relationship between economic variables – for example, between money and income, or between consumption and income – can be decomposed into relationships at separate

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

scales with interesting implications for economic theory. The insights gained include reconciling a number of anomalies in the consumer and monetary theory literature as well as introducing the notion that observed time series are an aggregation of behaviour at different time scales; for example, the time path of consumption totals represents the actions by consumers operating on a variety of time scales. In the same papers the authors also discovered that at certain time scales the relationship between economic variables may be subject to variations in the delay. These results have been confirmed by other researchers (for example, Gencay, Selcut, and Whitcher, 2002). For an alternative approach to testing for causality in the frequency domain using wavelets, see Dalkir (2004).

Yet another insight provided by wavelet analysis is the distinction between 'smoothing' and denoising. The former, traditional in econometrics, is based on the assumption that the signal is smooth relative to the noise, whereas the latter allows for the signal to be as irregular as the noise, but with greater amplitude. For smooth signals subject to noise, the obvious approach in order to minimize the effect of the noise is to average in some manner. However, if the signal is not smooth, averaging is not a suitable approach in that the averaging process distorts the signal itself. One can claim that denoising is often more relevant to economic and financial analysis than is smoothing (see Ramsey, 2004). These remarks are particularly relevant in the context of estimating relationships involving regime shifts, threshold models, and other non-differential changes in variable values. In an important series of papers Donoho, Johnstone and coauthors explored the use of wavelets and the concept of shrinkage whereby the size of the wavelet coefficient estimates is reduced to allow for the presence of noise (see Donoho et al., 1995). Further, shrinkage can be applied differentially across scales thereby refining the technique (see Ramsey, 2004, for more recent references and Gencay, Selcut, and Whitcher, 2002, for a thorough development of wavelet denoising).

Forecasting is an important topic; see Fryzlewicz, van Bellegem and von Sachs (2002) and Li and Hinich (2002), who demonstrate how the wavelet approach disentangles the variation in forecastability over time scales; that is, the ability to forecast varies across time scales. At the simplest level a given time series can be decomposed into trend, business cycle and seasonal components by wavelets and individually structured forecasting methods applied to each component separately before synthesizing the entire signal in order to produce forecasts for the whole series (see Ramsey, 2004, for a brief review of the literature on forecasting using wavelets).

See Also

- forecasting
- long memory models
- seasonal adjustment
- spectral analysis
- structural change
- time series analysis

Bibliography

Brillinger, D.R. 1994. Some uses of cumulants in wavelet analysis. *Journal of Nonparametric Statistics* 6, 93–114.

Bruce, A. and Gao, H. 1996. *Applied Wavelet Analysis with S-Plus*. New York: Springer.

Chui, C.K. 1992. *An Introduction to Wavelets*. New York: Academic Press.

Crowley, P. 2005. An intuitive guide to wavelets for economists. Discussion Paper No. 1/2005, Research Department, Bank of Finland.

Dalkir, M. 2004. A new approach to causality in the frequency domain. *Economics Bulletin* 3(44), 1–14.

Daubchies, I. 1992. *Ten Lectures on Wavelets*. Philadelphia: SIAM.

Donoho, D.L., Johnstone, I.M., Keryachariam, G. and Picard, D. 1995. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society* 57, 301–69.

Fryzlewicz, P., van Bellegem, S. and von Sachs, R. 2002. Forecasting non-stationary time series by wavelet process modeling. Technical Report No. 0208, Institut de Statistique, Université Catholique de Louvain.

Gencay, R., Selcut, F. and Whitcher, B. 2001. Scaling properties of foreign exchange volatility. *Physica A* 289, 249–66.

Gencay, R., Selcut, F. and Whitcher, B. 2002. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*. New York: Academic Press.

Gencay, R., Selcut, F. and Whitcher, B. 2003. Systematic risk and time scales. *Quantitative Finance* 3, 108–16.

Hong, Y. and Kao, C. 2004. Wavelet testing for serial correlation of unknown form in panel models. *Econometrica* 72, 1519–63.

Jensen, M.J. 1999. Using wavelets to obtain a consistent ordinary least squares estimator of the long-memory parameter. *Journal of Forecasting* 18, 17–32.

Jensen, M.J. 2000. An alternative maximum likelihood estimator of long-memory processes using compactly supported wavelets. *Journal of Economic Dynamics and Control* 24, 361–87.

Li, T.H. and Hinich, M. 2002. A filter bank approach for modeling and forecasting seasonal patterns. *Technometrics* 44(1), 1–14.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mallat, S. and Zhang, Z. 1993. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing* 41, 3397–415.

Percival, D.B. and Walden, A.T. 2000. *Wavelet Methods for Time Series Analysis*. Cambridge: Cambridge University Press.

Ramsey, J.B. 1999a. The contribution of wavelets to the analysis of economic and financial data. In *Philosophical Transactions of the Royal Society, Series A: Mathematical and Physical Sciences*. Repr. in *Wavelets*, ed. B.W. Silverman and J.C. Vassilicos. Oxford: Oxford University Press.

Ramsey, J.B. 1999b. Regression over time scale decompositions: a sampling analysis of distributional properties. *Economic Systems Research* 11(2), 163–83.

Ramsey, J.B. 2004. Wavelets in economics and finance: past and future. *Studies in Nonlinear Dynamics and Econometrics* 6(3), Article 1.

Ramsey, J.B. and Lampart, C. 1998a. The decomposition of economic relationships by time scale using wavelets: money and income. *Macroeconomic Dynamics* 2, 49–71.

Ramsey, J.B. and Lampart, C. 1998b. The decomposition of economic relationships by time scale using wavelets: expenditure and income. *Studies in Nonlinear Dynamics and Econometrics* 3(4), 23–42.

Ramsey, J.B. and Zhang, Z. 1996. The application of waveform dictionaries to stock market index data. In *Predictability of Complex Dynamical Systems*, ed. Y.A. Kravtsov and J. Kadtke. New York: Springer.

Ramsey, J.B. and Zhang, Z. 1997. The analysis of foreign exchange rates using waveform dictionaries. *Journal of Empirical Finance* 4, 341–72.

Schleicher, C. 2002. An introduction to wavelets for economists. Working Paper No. 2002-3, Bank of Canada.

Silverman, B.W. and Vassilicos, J.C. 1999. *Wavelets: The Key to Intermittent Information*? Oxford: Oxford University Press.

Strang, G. and Nguyen, T. 1996. *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press.

How to cite this article

Ramsey, James B. "wavelets." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000133> doi:10.1057/9780230226203.1820

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

wealth

Robert L. Heilbroner
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

abstract labour; circulating capital; cost of production; exchange value; labour; marginal revolution; power; productive and unproductive labour; property; riches; scarcity; social stratification; stocks and flows; utility; value; wage differentials; wealth

Article

Wealth is a fundamental concept in economics – indeed, perhaps the conceptual starting point for the discipline. Despite its centrality, however, the concept of wealth has never been a matter of general consensus. Although wealth has not become a focus of heated controversy comparable to that of value (despite the fact that the two terms are inextricably conjoined, as we shall see), conceptions of wealth have clashed profoundly and even irreconcilably. The result has been a continuing discussion of deep importance for economics – not only for its intrinsic interest, but because it calls into question the very scope and content of the discipline itself.

At the root of the long history of disagreement about wealth lie two conflicting conceptions of what the word implies. One of these, far more ancient than the formal study of economics and still very much in general use, is the idea of wealth as tangible possessions. For over a century, however, this conception has been challenged by another, which has identified the nature of wealth in the pleasures or ‘utilities’ generated by tangible goods, rather than the goods themselves. In the differing implications arising from these ‘objective’ and ‘subjective’ conceptions of wealth lie consequences of great significance for a discipline that has traditionally considered itself to be concerned with the study of wealth.

The objective conception of wealth is as old as written history, but the economist has not been interested in records of slaves and land and gold, other than to remark (usually as an economic anthropologist) on the extraordinary variety of objects that have been utilized as embodiments of wealth. The analytic problem to which economists have been drawn has been the attempt to establish a common denominator in which to sum up the value represented by a heterogenous collection of objects. ‘The entire study of wealth is, indeed, meaningless unless there be a unit for measuring it;’ wrote J.B. Clark, ‘for the questions to be answered are quantitative. How great is the wealth of a nation?’ (Clark, 1899, p. 375).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In ordinary discourse, this common denominator has always been money, and we will later consider the cogency of this common sense rule. For the economist, however, the challenge has been to discover some metric less arbitrary and unstable than a monetary sum. Thus the idea of objective wealth becomes inextricably entwined with the need to discover a standard – an embodiment of ‘value’ – by which its extent can be calculated. In the late mercantilist period this measure of extent was conceived by Petty and Cantillon to be the ‘amounts’ of land and labour that entered into the production of things – a considerable advance over earlier ideas that gold and silver possessed intrinsic value. This dual standard was subsequently reduced by Adam Smith to labour alone. ‘It was not by gold or silver, but by labour that all the wealth of the world was originally purchased’, he wrote in *The Wealth of Nations*, ‘and its value, to those who possess it and who want to exchange it for some new productions, is precisely equal to the quantity of labour which it can enable them to purchase or command’ (Smith, 1776, p. 48).

The choice of an objective standard of wealth – in Smith's case the labour ‘commanded’ by goods – focused the discipline of economics on the processes by which these embodiments of wealth were amassed. By the 17th century, the rise of a market organization of trade and production had already brought to the fore the distinctively ‘economic’ problem of wealth – namely, the need to explain its accumulation as the outcome of impersonal processes rather than as the spoils of power. From Smith's Physiocratic predecessors through John Stuart Mill, the principal aim of political economy was accordingly to investigate the consequences of a competitive struggle for wealth, with respect both to its distribution among individuals and social classes, and to its effect on the development of the system as a whole.

Almost from the outset, however, the conception of wealth as an objective element in the economic process posed troublesome questions. One of these was the appropriate treatment of labour that produced services rather than tangible goods. Because services are flows, they cannot be included in wealth, if the latter is defined as a tangible stock. The difference, as Cassel explained, involves *time*: stocks are present in their entirety at a moment in time; flows require the passage of time (Cassel, 1918, p. 31). A second difficulty concerned the classification of different kinds of labour. Smith, for example, differentiated between productive and unproductive labour, calling ‘productive’ only the labour whose product could be sold to replenish the working capital of the manufacturer, and designating as ‘unproductive’ all services – ‘how honourable, how useful, or how necessary soever’ – because these activities consumed, but did not renew, the fund of circulating capital whence they derived their subsistence (Smith, 1776, p. 331).

In addition, Smith and Ricardo recognized that labour was itself a heterogenous rather that a simple ‘substance’, and that some means would have to be found to reduce its complexity to a uniform basis. Both consigned the solution of the problem to the workings of the market. This may have been adequate for a rough and ready explanation of wage differentials originally established by market considerations and subsequently perpetuated by social inertia, but it concealed the deeper problem of reducing a spectrum of labour skills to a common denominator of ‘simple’ labour without recourse to market forces – that is, to supply and demand.

Finally, as Marx was to point out, the classical economists did not perceive that labour was a concrete activity – the labour of Ricardo's deer hunter not being substitutable for that of his salmon fisherman – so that a level of ‘abstract’ labour had to be posited if labour was to serve as a universal equivalent, or measuring rod, for wealth. Although the full difficulty of reducing labour to its abstract essence escaped Marx himself, for all these reasons the concept of labour as a simple and self-evident metric became

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
increasingly difficult to accept.

In accounting for the decline of the objective view of wealth, however, it is likely that the difficulties enumerated above did not play so important a role as another, quite separate, objection. This was the awareness that wealth as an objective entity did not express the attribute of goods that seemingly endowed them with desirability, namely their capacity to yield pleasure or utility to their possessors or beneficiaries. Oddly enough, we can also trace this view of wealth to Smith, who declared that ‘Every man is rich or poor according to the degree in which he can afford the necessaries, conveniencies, and amusements of human life’ (Smith 1776, p. 47).

It was Ricardo who first pointed out the inconsistency in Smith's views, in that the subjective enjoyments yielded by wealth – its ‘riches’ – were not the same as the expenditure of labour power required for its creation – its ‘value’. Thus for Ricardo, two countries might be equally ‘rich’ in necessaries and conveniences, but the value of the riches of one would be larger than that of the second if they required more labour to produce (Ricardo, 1821, ch. 20).

Ricardo's distinction between riches and value marks a sharp distinction between subjective (enjoyment) and objective (embodiment) conceptions of wealth, but Ricardo himself did not pursue the analytic and conceptual horizons opened up by the subjective view. That was to be the work of the post-classical period, culminating in the marginalist ‘revolution’ of the 1870s. Although this episode is famous for its shift in the concept of value from labour to utility, it is apparent that this shift entailed an equally deep-seated and far-reaching change in the conception of wealth, and as a consequence, in the study of economics. The works of Gossen, Menger, Jevons and Walras – the pioneers in this redirection of economics – display considerable variations in their internal details but not in their underlying depiction of the task of economics. This was now seen as an examination of the conditions for the optimization of enjoyments (utilities), not for the maximization of tangible wealth (capital). Thus Jevons wrote in *The Theory of Political Economy*, ‘The problem of economics may, as it seems to me, be stated thus: Given, a certain population, with various needs and powers of production, in possession of certain lands and other sources of material: required, the mode of employing their labour which will maximize the utility of the produce’ (Jevons, 1871, p. 254, original in italics).

A striking consequence of this shift was the necessary divorce of economics from any quantitative estimation of the extent of wealth. Utility in the post-classical sense was not the same as the ‘use-values’ that had always been recognized by Smith or Ricardo or Marx as the prerequisites of exchangeability. Their use-values referred to *objective* attributes of goods – the hardness of diamonds, the softness of cloth – from which was derived the capacity of commodities to yield subjective satisfactions. The utilities of the marginalists, on the other hand, referred exclusively to the states of mind induced by the possession or use of objects. Unlike use-values, therefore, utilities were subject to continual, possibly radical shifts, induced by changes in tastes or income or the relative scarcities of objects – in all cases, changes in the relation between possessors and objects, and not changes in the physical character of the goods themselves.

From this perspective, utility therefore had no objective existence whatsoever. ‘We can never say absolutely that some objects have utility and others have not’, Jevons wrote; and following in that line, Robbins declared in *The Nature and Significance of Economic Science* (1932, p. 47) that ‘wealth is not wealth because of its substantial qualities. It is wealth because it is scarce’.

The emphasis on the psychological element of wealth and on the role of scarcity in conferring desirability to goods clarified many questions, for example the ancient water–diamonds paradox. In

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
addition the utility approach appeared to resolve the problem of valuation at a level of greater generality than labour. It could be used, for example, to explain exchange value in the case of goods that required little or no labour, such as Ricardo's ‘rare statues and pictures’, within the same analytic framework as in the case of goods in which labour constituted a major element of cost. Thus the rise of a subjective orientation to wealth and value – we can by now surely appreciate their inextricable association – seemed an immense liberation to economists who had struggled within the constraints of an objective theory of wealth and value, whether exclusively denominated in labour or not.

The new orientation was not, however, without its problems. In so far as marginal utility is normally a direct function of scarcity, its adoption as the metric of wealth entailed the awkward conclusion that wealth as a sum of enjoyments and conveniences might well *increase* as a consequence of the *diminution* of material abundance. Some of the marginalists accepted this result; others, such as Menger, called it only an ‘apparent’ paradox, on the grounds that the continual augmentation of goods would gradually remove them from the category of ‘economic’ goods, thereby excluding them from consideration as wealth (Menger, 1871, 111). This seems a question-begging resolution. In addition, the replacement of an objective by a subjective standard of wealth led to the even more awkward conclusion that the aggregation of the wealth of individuals was impossible on the same grounds as the aggregation of their feelings or experiences. It was such considerations that led Robbins to declare in his influential essay mentioned above that ‘in any rigid determination of Economics, the term wealth should be avoided’ (Robbins, 1932, p. 47n).

All attempts to define wealth have therefore led to difficulties and even paradoxes. The conceptual and mensurational problems of an objective approach denominated in labour have been equalled, perhaps even surpassed, by those of a subjective approach denominated in utilities. Notwithstanding Robbins's reservations, however, economists have not abandoned the use of wealth as a fundamental constitutive element of economics, nor have they given up attempts to measure it. Here we can trace the general line of development once more to Adam Smith, this time to his famous abandonment of the category of labour as the measuring rod of value and his substitution of a cost of production measure which simply added up the income flows – wages, rents and profits – accruing to the three major classes.

From Ricardo on, Smith has been accused of circularity or inconsistency in this choice of an ‘adding-up’ approach to value, in which no attempt was made to discover a common denominator of wealth. But as a practical solution to the problem of measuring a concept that was universally regarded as real and important, whatever its intrinsic difficulties, Smith's approach was not without merit. The cost of production, or adding-up, approach to national wealth provided a common sense basis for the representation of national power or collective well-being, regardless of the unexamined problems behind these representations.

At all events, in modern times the measurement of wealth has become a major preoccupation for virtually all advanced nations. In *The Statistical Abstract of the United States*, for example, we find time series of various stocks and flows that have been selected as being of particular significance for the measurement of national wealth. The stocks include such items as estimates of financial and real assets, business and residential capital, consumers’ stocks of durables, land and selected government assets; while the flows concentrate on gross national product and its components. These items have been selected partly on the basis of the availability of data and partly on the basis of their importance for national economic policy. They are neither a complete nor a consistent set of accounts, a number of important stocks and flows being absent, such as the stock of human capital, or the flow of unpaid

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
labour in household work. The method of valuing both stocks and flows also differs from the public to the private sector, since the standard valuation is that of 'market values', which cannot apply to public goods or services. This is not the place to discuss the problems of national accounting, but it is worth noting that the same standard of practicality is applied as we find in Adam Smith, as well as the same absence of any firm conceptual foundation.

There remains another aspect to the concept of wealth. It is expressed with his usual vigour by John Bates Clark in *The Distribution of Wealth*: 'Amounts of wealth are usually stated in money … The thought in the minds of men who use money as a standard of value runs forward to the power that resides in the coins. The intuitions that are at the basis of this popular mode of speech are nearer to absolute truth than much of economic analysis. They discern a power of things over men …' (1908, p. 376).

The aspect of wealth to which Clark directs attention is once more anticipated by Smith, who writes, 'Wealth, as Mr. Hobbes says, is power … the power of purchasing; a certain command over all the labour, or over all the produce of labour which is then in the market' (Smith, 1776, p. 48). This definition contains an insight of great significance. To the extent that wealth is a form of power, its inadequate denomination in terms of labour commanded or utilities generated becomes explicable by virtue of the inapplicability of either metric to the 'substance' in which power must be measured. What might that substance be? Smith and other early investigators of the nature of wealth-seeking society assumed it to be the expression of a universal desire to be admired. 'The rich man glories in his riches, because he feels that they naturally draw upon him the attention of the world … and he is fonder of his wealth, upon this account, than for all the other advantages it procures him', Smith wrote in *The Theory of Moral Sentiments* (1759, pp. 50–1).

What was unknown to Smith, or to others, like Senior, who followed his general lead in the psychology of wealth, is that prestige and wealth do not seem to be universally conjoined. Contemporary anthropologists emphasize that wealth differs in a crucial respect from prestige in that the defining characteristic of wealth is its ability to confer social power on its possessors, whereas the enjoyment of prestige carries no such intrinsic rights. As a consequence, we find that in primitive societies, where there is universal access to the resources needed for subsistence, wealth does not exist as a social category, in that no individual or group enjoys command over the labour or the product of others, save for the claims conferred by relations of kinship or communal obligation (Sahlins, 1972; Fried, 1967).

From the anthropological viewpoint, then, primitive societies enjoy Ricardian riches, but no Smithian or Marxian value. From this vantage point, 'wealth' ceases to appear as an eternal attribute of human society, whether as tangible goods or the utilities enjoyed by their beneficiaries. Rather, the crucial element in the conception of wealth, and in the constitution of economics as its study, lies in the historical advent of the institution of property, construed as the right to exclude others from the material or other resources to which legal title has been gained. From this perspective, the fundamental problem posed by wealth is that of tracing the evolution of the social stratification characteristic of all post-primitive societies. Wealth is the economic face of that political stratification, lodged in the hands of a class whose ability to grant or deny access to resources becomes the 'economic' basis for both prestige and power.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- capitalism
- socialism

Bibliography

Cassel, G. 1918. *Theory of Social Economy*. New York: Harcourt Brace & Co., 1932.

Clark, J.B. 1899. *The Distribution of Wealth*. Boston: Ginn & Company, 1908.

Fried, M. 1967. *The Evolution of Political Society*. New York: Random House.

Jevons, W.S. 1871. *The Theory of Political Economy*. Ed. R.D. Collison Black, Harmondsworth: Penguin, 1970.

Menger, C. 1871. *Principles of Economics*. New York: New York University Press, 1981.

Ricardo, D. 1821. *On the Principles of Political Economy and Taxation*. Ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

Robbins, L. 1932. *An Essay on the Nature of Significance of Economic Science*. London: Macmillan.

Sahlins, M. 1972. *Stone Age Economics*. Chicago: Aldine.

Smith, A. 1759. *The Theory of Moral Sentiments*. Ed. D.D. Raphael and A.L. Macfie, Oxford: Clarendon Press, 1976.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Ed. R.H. Campbell, A. S. Skinner and W.B. Todd, Oxford: Clarendon Press, 1976.

How to cite this article

Heilbroner, Robert L. "wealth." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000038> doi:10.1057/9780230226203.1821

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Weber, Alfred (1868 – 1958)

Martin Beckmann
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Launhardt, C. F. W.; location theory; transport costs; Weber, A.; Weber, M.

Article

German economist and cultural sociologist, born 30 July 1868 in Erfurt; died 2 May 1958 in Heidelberg. He studied in Berlin and was Privatdozent there 1900–1904, Professor in Prague 1904–7 and in Heidelberg from 1907 until his death, interrupted by voluntary retirement between 1933 and 1945.
Weber's *Reine Theorie des Standorts* (1909) established him as the leading location theorist since Von Thünen, although his theoretical model has been anticipated by Wilhelm Launhardt (1882). Weber's interest in the location of the emerging modern industries arose from his earlier study of home industries, their struggle for survival, and the resulting social conditions of those working under the putting out system. Only Part I was published, an intended empirical Part II never appeared. Under Alfred Weber's guidance several theses were written on particular industries. Alfred Weber's article in the *Grundriss der Sozialoekonomik*, Part VI (1922) is a restatement of his book.
After the First World War, Weber turned to ‘cultural sociology as cultural history’ (1935). This work, while overshadowed by that of his more famous brother, Max (1864–1920), and lacking its precision, provides a fresh perspective of the development of Western civilization.
Alfred Weber, although not an impressive speaker, was a highly influential teacher. Together with Karl Jaspers he re-established academic traditions of excellence at the University of Heidelberg in the post-Second World War years.
The location of industries according to Weber is governed by cost minimization. When production costs are independent of location, this means the minimization of transportation costs. In the case of two resource deposits and one market, the optimal location may fall inside the triangle spanned by the three given locations. Economies of joint location, based on the exchange of intermediate goods or the joint use of indivisible facilities induce ‘agglomeration’ of industrial activities in large centres.

See Also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- location theory
- Thünen, Johann Heinrich von

Selected works

1909. *Reine Theorie des Standorts*. Trans. C.J. Friedrichs as *Theory of the Location of Industries*, Chicago: University of Chicago Press, 1929.

1935. *Kulturgeschichte als Kultursoziologie*. 2nd edn, Munich: Piper, 1950.

Bibliography

Salin, E., 1958. Obituary. *Kyklos* 11, 318–40.

How to cite this article

Beckmann, Martin. "Weber, Alfred (1868–1958)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000046> doi:10.1057/9780230226203.1822

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Weber, Max (1864 – 1920)

David Beetham
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Max Weber was one of the leading historical political economists in the Germany of the 1890s. Weber's early work in political economy reflected the distinctive concerns of a younger generation of the Historical School, which sought to demonstrate the theoretical character of the concepts used in historical economics and the historical presuppositions of theory. Over time, Weber's research became increasingly wide-ranging and theoretical, involving an elucidation of the character of Western rationalism as applied to the basic structures of economy and society, and reflecting a shift in his disciplinary focus of interest from political economy to sociology.

Keywords

bureaucracy; capital accumulation; class; class conflict; conspicuous consumption; division of labour; exploitation; factor demand; historical materialism; Historical School, German; individualism; labour theory of value; managerial revolution; Menger, C.; Modernization; planning; private property; Protestant ethic thesis; surplus value; value judgements; Weber, M.

Article

Max Weber was born in 1864 at Erfurt and died in 1920 at Munich. After early studies in the history of commercial law, he established himself as one of the leading figures in a new generation of historical political economists in the Germany of the 1890s. He was appointed to chairs in political economy at Freiburg in 1894 and at Heidelberg in 1896. A nervous breakdown commencing in 1898 led to his withdrawal from academic teaching, but did little to impair the flow of his writing. In 1904 he took over the editorship of the *Archiv für Sozialwissenschaft und Sozialpolitik*, the leading academic journal in ‘social economics’, devoted to the exploration of the interrelationship between economy on the one hand, and law, politics and culture on the other. This interconnection formed the main site of Weber's own research, whose focus became increasingly wide-ranging and theoretical, involving an elucidation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of the character and presuppositions of modern Western rationalism, as applied to the basic structures of economy and society. Weber was also actively and often controversially involved in the political issues of Wilhelmine Germany, from a progressive national-liberal standpoint, and during the war was one of the leading polemicists for a democratization of the constitution. Such involvement gave particular sharpness to his discussions of social science methodology, the role of value judgements and the relation between academic analysis and political practice. It was only comparatively late in his life that he came to think of his work as ‘sociology’, though it is as one of the ‘founding fathers’ of sociology that he is now known. He resumed full-time teaching activity as professor at Munich in 1919 only shortly before his death.

Weber's early work in political economy can best be understood as reflecting the distinctive concerns of a younger generation of the Historical School (including Schulze-Gävernitz, Sombart, Max and Alfred Weber). At the methodological level they sought to resolve the controversy between the theoretical and historical schools by demonstrating the theoretical character of the concepts used in historical economics on the one hand, and the historical presuppositions of theory on the other. An important element in this resolution was to secure the acceptability of the Marxian concept of ‘capitalism’ as a valid concept for economic analysis, despite the untenability (as they saw it) of the labour theory of value, and exaggerated claims made for the materialist conception of history.

The recognition of the conflict between labour and capital as a systemic property of capitalism was central to Weber's early work. In his study of the impact of capitalist organization on the agricultural estates east of the Elbe, it supported his conclusion that class conflict had permanently undermined the economic basis of Junker political dominance in the Reich (Weber, 1892). From it he also derived a position on social policy which was critical of the paternalism of the ‘Kathedersozialisten’, arguing instead that trades unions should be given a secure legal status so that they could bargain for themselves on a more equal footing with capital. The distinctive Weberian conception of class conflict under capitalism was theorized neither in the Marxist terms of ‘exploitation’ nor in the neoclassical terms of ‘factor demand’, but in terms of a systematic competition for the social product on the basis of a power relation between the classes that was adjusted and underwritten at the political level.

If the incorporation of Marxian insights into the mainstream of social economics required that the analysis of class conflict be freed from the doctrine of surplus value, it also required a critique of the one-sided assumptions of historical materialism. This Weber offered in his most famous study *The Protestant Ethic and the Spirit of Capitalism* (1904). The argument of this work was that the profit-maximizing behaviour so characteristic of the bourgeoisie, which could be explained under fully developed capitalist conditions by its sheer necessity to survival in the face of competition, could not be so explained under the earlier phases of capitalist development. It was the product of an autonomous impulse to accumulate far beyond the needs of personal consumption, an impulse which was historically unique. Weber traced its source to the ‘worldly asceticism’ of reformed Christianity, with its twin imperatives to methodical work as the chief duty of life, and to the limited enjoyment of its product. The unintended consequence of this ethic, which was enforced by the social and psychological pressures on the believer to prove (but not earn) his salvation, was the accumulation of capital for investment.

Early critics of Weber's thesis misunderstood it as a purely cultural explanation for capitalism, as if ‘a Siberian Baptist or a Calvinist inhabitant of the Sahara’ must inevitably become a successful entrepreneur. Weber was in fact well aware both of the material preconditions for capitalist development, and of the social interests that are needed to support the dissemination of new ideas. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
crucial question about his thesis is whether the employment of wage labour that made unlimited accumulation possible in principle, also made it inevitable in practice; whether, that is, the Protestant ethic should be seen as providing a necessary motivation for capitalist accumulation, or rather a legitimation for it in the face of prevalent values favouring conspicuous consumption on the part of a leisured class. Weber himself saw it as both. His work is the most sophisticated in a long tradition of exploration of the cultural preconditions for capitalist accumulation, from Adam Smith's celebration of ‘parsimony’, to recent explanations of Britain's economic decline in terms of the gentrification of its entrepreneurial spirit.

At one level the ‘Protestant ethic thesis’ can thus be read as a critique of historical materialism, with its explanation of capitalist development as the necessary outcome of the feudal order, rather than as the result of a unique conjunction of favourable historical conditions, cultural and political as well as economic. At another level it can be read as an extended critique of the ahistorical theorizing of Carl Menger and the Austrian school. In Weber's view the methodical, calculating, welfare-maximizing behaviour of the neoclassical models was not a universal characteristic of human rationality as such, but a product of modern Western rationalism. His subsequent studies of the economic ethic of the major world religions (Confucianism, Hinduism, Buddhism, ancient Judaism; collected in Weber, 1921) were designed to elucidate this distinctive cultural complex. They showed that, while instrumental rationality was a universal category of social action, only in the modern West had the goal-maximizing calculation of the most efficient means to given ends become generalized. And while other cultures had attempted to make the world intelligible through the development of elaborate theodicies, or to create internally consistent systems of ethics or law, the distinctive features of Western rationalism were the scientific assumption that all things could be comprehended by reason, together with the attitude of practical mastery which sought to subject the world to human control rather than merely adjust oneself to it. In Weber's major unfinished theoretical work, *Economy and Society* (1922), capitalism was shown to be simply one expression, rather than the unique locus, of this ‘rationalization’ process. The work is structured around the antithesis between ‘traditional’ and ‘rationalized’ forms of action and organization in all spheres of social life, and the transition between the two provides the key to the Weberian theory of modernization.

The conclusion of Weber's mature work, that capitalism was to be understood as part of a wider ‘rationalization’ process, coincided with his analysis of its most advanced forms in contemporary Germany. According to this analysis, the distinctive feature of capitalist concentration was the change in its internal mode of organization: the adoption of a complex technical division of labour and a hierarchical structure of administration that increasingly resembled the bureaucratic type already established in the political sphere. For Weber, the bureaucratic model of administration was becoming generalized throughout all sectors of contemporary society, because of its efficiency in performing complex organizational tasks. Along with it went the emergence of a new middle class, whose distinctive position in the class structure depended neither on property ownership (capital) nor its absence (labour), but on the possession of technical and organizational skills, and on its authority position within a bureaucratic hierarchy.

Some commentators have seen Weber as an early forerunner of the ‘managerial revolution’ thesis. Certainly he was among the earliest to identify technical knowledge and organization as crucial sources of social power in modern societies. But to Weber no manager could be a substitute for the risk-taking entrepreneur who stood at the head of large capitalist organizations. The bureaucratic system, with its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
secure career and promotion prospects, represented a conservative principle in social life, in contrast to the dynamic principle of the market. Like his junior colleague, Joseph Schumpeter, Weber saw the major source of economic innovation to be provided by the captains of industry, ready to chance their judgement in the competition of the market place. This was directly paralleled in the political sphere by his theory of competitive leadership democracy, according to which the leaders of mass parties with their bureaucratic machines competed for support in the electoral market place. If the creative force of individualism, deriving from the Protestant ethic, had itself unintentionally produced the age of organization, in which competition at the individual level was eliminated, nevertheless the role of individualism was reasserted in Weberian social theory at the head of organizations, in the form of ‘charismatic’ leadership.

Weber's theory of bureaucracy also provided the basis for a thoroughgoing critique of socialist planning, as prefigured in the wartime German economy. Weber was quick to echo von Mises's contention that a coherent system of allocation was impossible without market indicators, since it confirmed his own historical analysis of the preconditions for rational economic calculation. However, his distinctive criticisms of socialist planning derived from the massive extension of bureaucratic coordination he believed it would entail. Without market competition, he argued, the economy would simply stagnate. Yet the workers would remain subordinate to the same hierarchy of authority at the work place, since this was determined by the technical requirements of production, not by the particular system of ownership. Indeed, their subordination would become a new form of slavery, since the separate hierarchies of the economic, legal and political spheres would be fused into a single, all-embracing structure of power. It was the dictatorship of the official, he concluded, not of the worker, that was on the march (Weber, 1918).

Overall, the progressive widening in the focus of Weber's theoretical concerns, from the conditions for economic rationality to the general theme of ‘rationalization’, and the subsumption of capitalism itself into the wider category of bureaucratic organization, reflected a shift in his disciplinary focus of interest from political economy to sociology. This was not just a personal development of Weber's, but one typical of the period in which he lived. With the narrowing of theoretical focus represented by neoclassical economics, it was left to the nascent discipline of sociology to take over some of the wider concerns of political economy. The rich tradition of the German Historical School, and the methodological debates which it had aroused, made German sociology particularly well placed for this enterprise. It was also particularly urgent in a country where the claims of Marxism to provide a convincing overall theory of society were widely accepted within the labour movement. It was no accident, therefore, that the most sustained rebuttal of these claims should come from the same context. As suggested above, however, Weber's approach to Marxism was not one of outright rejection, but of incorporating its insights into a different theoretical framework which left the validity of private property ownership intact. If the general presuppositions of liberalism had been thematized in the form of political philosophy in the 18th century, and of political economy in the 19th, they can be said to have received their distinctive 20th-century expression in the form of Weberian sociology.

See Also

- bureaucracy

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

- Historical School, German

Selected works

1892. Die Lage der Landarbeiter im ostelbischen Deutschland. In *Max Weber Gesamtausgabe*, vol. 1/3, ed. M. Riesebrodt. Tübingen: J.C.B. Mohr, 1984.

1904. *The Protestant Ethic and the Spirit of Capitalism*. London: Allen & Unwin, 1930.

1918. Socialism. In *Weber Selections in Translation*, ed. W.G. Runciman. Cambridge: Cambridge University Press, 1978.

1921. *Gesammelte Aufsätze zur Religionssoziologie*. Tübingen: J.C.B. Mohr.

1922. *Economy and Society*. New York: Bedminster Press, 1968.

Bibliography

Camic, C., Gorski, P. and Trubek, D., eds. 2005. *Max Weber's Economy and Society: A Critical Companion*. Stanford: Stanford University Press.

Ringer, F. 2004. *Max Weber: An Intellectual Biography*. Chicago: University of Chicago Press, 2004.

Swatos, W. and Kaelbar, L., eds. 2005. *The Protestant Ethic Turns 100: Essays on the Centenary of the Weber Thesis*. Boulder, CO: Paradigm Publishing.

Swedberg, R. 2005. *The Max Weber Dictionary: Key Words and Central Concepts*. Stanford: Stanford University Press.

Turner, S., ed. 2000. *The Cambridge Companion to Weber*. Cambridge: Cambridge University Press.

How to cite this article

Beetham, David. "Weber, Max (1864–1920)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000047>
doi:10.1057/9780230226203.1823

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Weintraub, Sidney (1914 – 1983)

J.A. Kregel
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand and supply; neoclassical synthesis; price theory; tax-based incomes policy; Weintraub, S.

Article

Weintraub was born in Brooklyn, New York, on 28 April 1914 and died in Philadelphia, Pennsylvania, on 19 June 1983. Professor at the University of Pennsylvania from 1950 until his death, Weintraub was widely known as the originator (with Henry Wallich) of tax-based incomes policy (TIP); his professional reputation is based on his early criticism (1959b) of the ‘neoclassical synthesis’ of Keynes's macroeconomics and Walrasian general equilibrium and his own highly original attempts to produce a microeconomics compatible with Keynes's macroeconomic theory.

A postgraduate year (1938–9) at the London School of Economics convinced him of the implications of Keynes's *General Theory* for price theory. His Ph.D. dissertation (‘Monopoly and the Economic System’, St Johns, 1941) and a series of articles on the formulation of demand in conditions of imperfect competition and imperfect information produced his innovative *Price Theory* (1949).

After the war, Weintraub concentrated on producing a micro-theory compatible with Keynes's theory of the endogeneous determination of the equilibrium level of output at less than full employment. His earlier work was extended to the demand for labour and the microfoundations of the aggregate demand and supply curves. Although this work, summarized in *An Approach to the Theory of Income Distribution* (1958), reached similar conclusions to the aggregate distribution theories then being worked out by Kaldor and Robinson, its inspiration was the formulation of a ‘Keynesian’ microeconomics rather than growth theory.

Evidence of the stability of the markup of prices over wage costs presented in *A General Theory of the Price Level* (1959a) led to a ‘watchtower approach’ to wage policy which preceded the widely discussed, but never applied, TIP proposal (1971a; 1971b).

Weintraub's prolific writing and lecturing activities were complemented in 1978 by the founding and editing (with Paul Davidson) of *The Journal of Post Keynesian Economics*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

See Also

- Post Keynesian economics

Selected works

1942. The foundations of the demand curve. *American Economic Review* 32, 538–52.

1946. Monopoly pricing and unemployment. *Quarterly Journal of Economics* 61, 108–24.

1949. *Price Theory*. New York: Pitman.

1956. A macroeconomic approach to the theory of wages. *American Economic Review* 46, 845–56.

1957. The micro-foundations of aggregate demand and supply. *Economic Journal* 67, 455–70.

1958. *An Approach to the Theory of Income Distribution*. Philadelphia: Chilton.

1959a. *A General Theory of the Price Level*. Philadelphia: Chilton.

1959b. *Classical Keynesianism, Monetary Theory and the Price Level*. Philadelphia: Chilton.

1966. *A Keynesian Theory of Employment Growth and Income Distribution*. Philadelphia: Chilton.

1971a. An incomes policy to stop inflation. *Lloyds Bank Review* 99(January), 1–12.

1971b. (With H.C. Wallich.) A tax-based incomes policy. *Journal of Economic Issues* 5(2), 1–19.

1978. *Keynes, Keynesians and Monetarists*. Philadelphia: University of Pennsylvania Press.

How to cite this article

Kregel, J. A. "Weintraub, Sidney (1914–1983)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000049>
doi:10.1057/9780230226203.1824

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

welfare costs of business cycles

Ayse Imrohoroglu
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The welfare cost of business cycles measures the benefits that would be obtained by individuals from eliminating all the macroeconomic instability in a given economy. In a seminal paper, Lucas (1985) argued that these benefits are almost certain to be trivially small, especially when they are compared with the benefits that can be achieved with more growth for the post-war US economy.

Keywords

business cycles; Great Depression; National Bureau of Economic Research; risk aversion; stabilization policy; unemployment insurance; welfare cost of business cycles

Article

The welfare cost of business cycles measures the benefits in terms of additional consumption that would be obtained by individuals from eliminating all the macroeconomic instability in a given economy. Hodrick and Prescott (1997) and Lucas (1977) define business cycles as recurrent fluctuations of output about trend and the co-movements among other aggregate time series. These fluctuations are typically represented as expansions and recessions in economic activity. The National Bureau of Economic Research, a private non-profit organization that is responsible for updating the business cycle chronology, defines a recession as ‘a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real gross domestic product (GDP), real income, employment, industrial production, and wholesale-retail sales’ (NBER, 2003, p. 1). One of the prevailing views in macroeconomics is that business cycles are welfare reducing and governments should try to stabilize the economy by using fiscal or monetary policies.

In his seminal work, Lucas (1985) proposes a simple framework to think about how to compute the cost of economic instability, and challenges the paradigm that business cycles have large welfare costs. His measure of the welfare cost for the United States turns out to be trivially small, which disputes the need

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
for developing more advanced policies that would eliminate fluctuations in the United States. The following section examines his work and the subsequent research.

Lucas proposes that in order to understand the welfare cost of instability we need to start with the preferences of a hypothetical consumer who is faced with a sequence of consumption goods over time labelled $\{c_t\}$. The expected utility of such a sequence can be calculated by,

$$E\left\{\sum_{t=0}^{\infty}\beta^t\frac{1}{1-\sigma}(c_t^{1-\sigma}-1)\right\}$$

where $U(c_t) = (c_t^{1-\sigma} - 1) / (1 - \sigma)$ is the period utility function, E is the expectations operator, β is the subjective discount factor and $\sigma > 0$ is the coefficient of relative risk aversion. An important property of this utility function is that consumers would prefer smooth consumption streams to fluctuating ones or that they would prefer a deterministic consumption path to a risky path with the same mean.

In this construct, in order to understand how consumers may feel about economic instability, we can simply ask them to evaluate their lifetime utility under two different scenarios. In particular, suppose the consumers are asked to compare the lifetime utility of a perfectly smooth consumption path with a consumption stream that increases in good times and decreases in bad times while maintaining the same average level over time. The latter consumption stream is the one that results in the case of business cycles. Surely, consumers who care about smoothing consumption over time will rank the utility generated by such a stream lower than the one from the smooth consumption stream. In fact, the higher the value of σ, the lower the utility of a fluctuating consumption stream will be. With this in mind we can ask a second question. What would it cost to compensate all individuals in terms of extra consumption, uniform across time and different shocks, so that they will be indifferent between the smooth and the fluctuating consumption paths? This turns out to be a fairly easy calculation where the following equation provides a quantitative answer:

$$\lambda = \frac{1}{2}\sigma\mu^2$$

where λ is the compensation parameter, σ is the coefficient of relative risk aversion and μ measures the standard deviation in consumption.

Our hypothetical example can be made concrete by examining the properties of personal expenditures on consumption in a particular economy. Lucas (2003) uses US data for the period 1947–2001 and calculates the standard deviation of the log of real per capita consumption about the linear trend to be 0.032. Using this estimate, we can arrive at several measures of the cost of instability based on different assumptions on the coefficient of relative risk aversion. The amazing part of the findings is that the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
magnitude of these estimates range between one 20th of one per cent to one or two tenths of a per cent of consumption for risk aversion parameters between one and four. (Risk aversion coefficients in this range are considered to be consistent with many observations in an economy. However, much higher values are needed for some other observations such as the equity premium, which is discussed shortly.) For example, if $\sigma = 2$ and $u = 0.032$, then the consumption compensation that is required to make an individual indifferent between a fluctuating versus a constant consumption stream is about 0.001. For the US economy that would suggest that an annual consumption compensation as low as $28.96 per person would be sufficient to make individuals indifferent between a fluctuating and a smooth consumption stream. (Personal consumption expenditures in the United States in 2004 were $8.6 trillion. One-tenth of a per cent results in a total consumption compensation of $8.6 billion. Using the 2004 population of 297 million people results in consumption compensation per person of $28.96.) Such a welfare cost is negligible not only in an absolute sense but also when compared with other welfare cost measures. For example, Lucas (2000) calculates the welfare loss of a one per cent reduction in the growth rate of the economy to be as high as 20 per cent of consumption and the welfare cost of ten per cent inflation to be one per cent of income annually. Both of these estimates are more than an order of magnitude higher than the welfare cost of economic instability.

Lucas proposes to take the low cost findings seriously as giving a range of estimates for the size of the potential gains from developing policies that would eliminate fluctuations in the United States. Taking these results seriously is exactly what the profession did. Twenty years after Lucas's (1985) study, many economists continue to work on this subject, investigating whether the conclusions reached in his framework are valid under more complicated and sometimes more realistic frameworks.

Many of the assumptions in the original framework have been challenged. One of the main assumptions is that all agents are identical and have access to fully developed capital markets. One can easily imagine that, while the costs of instability may be low for some consumers, such as those with large savings, they may be devastating for some others, who may not have the means to insure themselves against these shocks. Several papers have investigated the welfare costs of instability for heterogeneous agents with limited access to capital markets. (Starting with Imrohoroglu, 1989, papers that have introduced incomplete markets and examined the role of idiosyncratic risk include Atkeson and Phelan, 1994; Gomes, Greenwood and Rebelo, 2001; Krusell and Smith, 1999; 2002; and Krebs, 2003.) Krusell and Smith (1999) examine an economy with substantial heterogeneity where individuals face idiosyncratic and aggregate risk and can smooth their consumption only through private savings. Their economy generates a wealth distribution that resembles US wealth distribution reasonably well. They investigate whether the welfare costs of cycles may be very high for some members of the society such as the unemployed even if in aggregate the costs are relatively low. Their findings indicate that while the welfare effects of eliminating cycles do differ across consumers they are extremely small for almost all consumers. Only for a very few individuals with almost zero consumption are welfare losses found to be as high as two per cent of average consumption.

Some of the papers in this area have highlighted the importance of understanding the interaction between aggregate and individual shocks in an economy. For example, how long-lasting are the effects of a bad shock? Do aggregate shocks compound the effects of individual shocks? Storesletten, Telmer and Yaron (2001) show that, in an environment where small aggregate shocks can have a long-lasting impact on individuals' earnings, the welfare cost of business cycles can be much higher than the original

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
estimates. (Beaudry and Pages, 2001, also study idiosyncratic wage risk that worsens in recessions, and obtain high estimates. However, they do not allow for savings to help smooth consumption in the economy with fluctuations.) Atkeson and Phelan (1994), on the other hand, discuss the connection between aggregate and idiosyncratic risk, and suggest as a serious possibility that the elimination of aggregate risk does not affect individual risk at all. In their framework welfare cost estimates are close to zero. However, if the effects of a bad shock are assumed to be permanent, as in Krebs (2003), then the welfare costs of business cycles can be as high as 7.5 per cent of consumption. In such a framework, even if credit markets are perfect, individuals will not borrow to smooth the negative shocks they face since the effect of those shocks will persist for ever.

Another set of papers have introduced different preferences or have implicitly or explicitly used higher risk-aversion coefficients in examining the welfare cost of business cycles. While higher costs are obtained in some of these environments, there are questions about the soundness of using very high risk-aversion coefficients. For example, Tallarini (2000) finds much larger costs in a model with Epstein–Zin type preferences where preference parameters are chosen to be consistent with observed asset market data. However, the main factor behind this finding is the use of a high risk-aversion parameter to be consistent with asset price determination. (Similarly, Alvarez and Jermann, 2004, find large welfare costs of economic instability in a framework that uses high risk aversion to match the six per cent equity premium in asset markets. See also Dolmas, 1998; Obstfeld, 1994.) Otrok (2001), on the other hand, suggests that in a model that allows for potential time-non-separabilities in preferences, which is calibrated to be consistent with observed fluctuations in a general equilibrium model of business cycles, the welfare cost of business cycles turns out to be quite low.

It might also be possible to obtain a higher cost of fluctuations if there are links between economic growth and fluctuations. For example, Ramey and Ramey (1995) demonstrate a strong negative relationship between volatility and growth in a panel of 92 countries. However, in examining the welfare cost of instability, Epaulard and Pommeret (2003) find the volatility in the US economy to be too small to generate large benefits from stabilization policies even if reductions in volatility induce growth. Jones, Manuelli and Stacchetti (1999) demonstrate that the relationship between volatility in fundamentals and mean growth can be positive or negative. Their quantitative results indicate that the size of this effect is not large enough to generate large welfare costs of instability. Barlevy (2004a), on the other hand, proposes a set-up where eliminating fluctuations reallocates investment from periods of high investment to periods of low investment. This mechanism results in achieving higher growth rates without necessarily requiring higher investment levels. In such a framework, he finds the welfare cost of instability to be substantially higher than in the original Lucas estimates. The key to obtaining such large costs in his model is the presence of diminishing returns to investment, for which there is some, but not overwhelming, evidence.

It may be important to point out that the way Lucas, and Hodrick and Prescott have defined business cycles, namely, as fluctuations around a trend, has an important implication for the welfare cost calculations. If instead recessions were viewed as inefficient declines in output, as in the Keynesian view, and stabilization policies were seen as policies that would prevent economic activity from falling below its maximum potential, then the welfare cost measure could be higher. This is the case in DeLong, and Summers (1988) and Cohen (2000), who obtain welfare costs of stabilization of around 1.6 per cent and one per cent respectively. In their frameworks stabilization increases the average level of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
consumption.

It is important to stress that the estimates that have been discussed so far have been for the post-war US economy. The low cost estimate that is obtained in many of these papers is partly due to the relative stability of the US economy since the 1950s. Welfare costs of business cycles are higher in economies that are faced with larger fluctuations in consumption. Using the volatility of consumption in the United States prior to the Second World War, or the fluctuations in consumption that are observed in many developing countries, results in significantly higher welfare cost measures (see, for example, Pallage and Robe, 2003). In addition, in the post-war period the US economy had a well-developed unemployment insurance system that may have helped reduce the volatility in consumption. Economies with less-developed welfare systems also yield higher welfare costs of instability. (Chatterjee and Corbae, 2007, find that the potential benefit of reducing the likelihood of economic crises such as a Great Depression-style collapse of economic activity can range between 1.05 and 6.59 per cent of annual consumption. They also find that uninsured unemployment risk contributes significantly to the size of these gains.) Although there is still some debate over the size of the welfare costs of business cycles, the weight of the evidence seem to suggest that they may not be too high for the US economy. (See also Barlevy, 2004b, for a survey of the literature on the welfare cost of business cycles.)

See Also

- growth and cycles
- liquidity constraints
- neoclassical growth theory
- real business cycles

Bibliography

Alvarez, F. and Jermann, U. 2004. Using asset prices to measure the cost of business cycles. *Journal of Political Economy* 112, 1223–56.

Atkeson, A. and Phelan, C. 1994. Reconsidering the cost of business cycles with incomplete markets. *NBER Macroeconomic Annual*. Cambridge, MA: MIT Press.

Barlevy, G. 2004a. The cost of business cycles under endogenous growth. *American Economic Review* 94, 964–90.

Barlevy, G. 2004b. The cost of business cycles and the benefits of stabilization: a survey. Working Paper No. 10926. Cambridge, MA: NBER.

Beaudry, P. and Pages, C. 2001. The cost of business cycles and the stabilization value of unemployment insurance. *European Economic Review* 45, 1545–72.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Chatterjee, S. and Corbae, D. 2007. On the aggregate welfare cost of Great Depression unemployment. *Journal of Monetary Economics* 54, 1529–44.

Cohen, D. 2000. A quantitative defense of stabilization policy. FEDS Working Paper No. 2000–34. Board of Governors of the Federal Reserve System.

DeLong, B. and Summers, L. 1988. How does macroeconomic policy affect output? *Brookings Papers on Economic Activity* 1988(2), 433–80.

Dolmas, J. 1998. Risk preferences and the welfare cost of business cycles. *Review of Economic Dynamics* 1, 646–76.

Epaulard, A. and Pommeret, A. 2003. Recursive utility, growth, and the welfare cost of volatility. *Review of Economic Dynamics* 6, 672–84.

Gomez, J., Greenwood, J. and Rebelo, S. 2001. Equilibrium unemployment. *Journal of Monetary Economics* 48, 109–52.

Hodrick, R. and Prescott, E. 1997. Post-war U.S. business cycles: a descriptive empirical investigation. *Journal of Money, Credit, and Banking* 29, 1–16. Originally a 1978 working paper at Carnegie-Mellon University.

Imrohoroglu, A. 1989. The cost of business cycles with indivisibilities and liquidity constraints. *Journal of Political Economy* 97, 1364–83.

Jones, L., Manuelli, R. and Stacchetti, E. 1999. Technology (and policy) shocks in models of endogenous growth. Working Paper No. 7063. Cambridge, MA: NBER.

Krusell, P. and Smith, A. 1999. On the welfare effects of eliminating business cycles. *Review of Economic Dynamics* 2, 245–72.

Krusell, P. and Smith, A. 2002. Revisiting the welfare effects of eliminating business cycles. Working paper, Carnegie-Mellon University.

Krebs, T. 2003. Growth and welfare effects of business cycles in economies with idiosyncratic human capital risk. *Review of Economic Dynamics* 6, 846–68.

Lucas, R. 1977. Understanding business cycles. In *Stabilization of the Domestic and International Economy*, ed. K. Brunner and A. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol 5. Amsterdam: North-Holland.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Lucas, R. 1985. *Models of Business Cycles*. Yrjo Jahnsson Lectures. Oxford: Basil Blackwell.

Lucas, R. 2000. Inflation and welfare. *Econometrica* 68, 247–74.

Lucas, R. 2003. Macroeconomic priorities. *American Economic Review* 93, 1–14.

NBER (National Bureau of Economic Research). 2003. The NBER's recession dating procedure. Business Cycle Dating Committee, NBER, 21 October. Online. Available at http://0-www.nber.org.library.lemoyne.edu/cycles/recessions.pdf, accessed 27 September 2006.

Obstfeld, M. 1994. Evaluating risky consumption paths: the role of intertemporal substitutability. *European Economic Review* 38, 1471–86.

Otrok, C. 2001. On measuring the welfare cost of business cycles. *Journal of Monetary Economics* 47, 61–92.

Pallage, S. and Robe, M. 2003. On the welfare cost of economic fluctuations in developing countries. *International Economic Review* 44, 668–77.

Ramey, G. and Ramey, V. 1995. Cross-country evidence on the link between volatility and growth. *American Economic Review* 85, 1138–51.

Storesletten, K., Telmer, C. and Yaron, A. 2001. The welfare cost of business cycles revisited: finite lives and cyclical variation in idiosyncratic risk. *European Economic Review* 45, 1311–39.

Tallarini, T. 2000. Risk sensitive real business cycles. *Journal of Monetary Economics* 45, 507–32.

How to cite this article

Imrohoroglu, Ayse. "welfare costs of business cycles." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000130>
doi:10.1057/9780230226203.1825

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

welfare economics

Allan M. Feldman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Welfare economics attempts to define and measure the ‘welfare’ of society as a whole. It tries to identify which economic policies lead to optimal outcomes, and, where necessary, to choose among multiple optima. This article answers three fundamental qsts with three fundamental theorems. In a competitive economy, will an equilibrium outcome be optimal? Can any optimal outcome be achieved by a modified market mechanism? Is there a reliable way to measure social welfare, or to derive the preferences of society from the preferences of individuals? The negative answer to the third question is partly overcome by the theory of implementation.

Keywords

Arrow social welfare function; Arrow's th; Bergson, A.; central planning; command economy; common good; compensation tests; competitive equilibrium; Condorcet, Marquis de; consumer surplus; Corn Laws; decentralized socialism; externalities; first fundamental theorem of welfare economics; implementation; industrial policy; interpersonal utility comparisons; Kaldor, N.; laissez-faire; Lange, O. R.; lump-sum taxes; lump-sum transfers; majority rule; mechanism design; Mises, L. E. von; Nash strategy; Pareto efficiency; partial equilibrium; Pigou, A. C.; Pigouvian taxes; plurality voting; preference misrepresentation; preference profiles; production plans; Scitovsky, T.; second fundamental theorem of welfare economics; Smith, A.; social choice function; socialism; strategic behaviour; third fundamental theorem of welfare economics; voting; voting cycles; voting paradoxes; welfare economics

Article

In 1776, the same year as the American Declaration of Independence, Adam Smith published *The Wealth of Nations*. Smith laid out an argument that is now familiar to all economics students: (*a*) the principal human motive is self-interest; (*b*) the invisible hand of competition automatically transforms the self-interest of many into the common good; (*c*) therefore, the best government policy for the growth

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of a nation's wealth is that policy which governs least.
Smith's arguments were at the time directed against the mercantilists, who promoted active government intervention in the economy, particularly in regard to (ill-conceived) trade policies. Since his time, his arguments have been used and reused by proponents of laissez-faire throughout the 19th and 20th centuries. Arguments of Smith and his opponents are still very much alive today: the pro-Smithians are those who place their faith in the market, who maintain that the provision of goods and services in society ought to be done, by and large, by private buyers and sellers acting in competition with each other. One can see the spirit of Adam Smith in economic policies involving deregulation, tax reduction, denationalizing industries, and reduction in government growth in Western countries; and in the deliberate restoration of private markets in China, the former Soviet Union and other eastern European countries. The anti-Smithians are also still alive and well; mercantilists are now called industrial policy advocates, and there are intellectuals and policymakers who believe that: (*a*) economic planning is superior to laissez-faire; (*b*) markets are often monopolized in the absence of government intervention, crippling the invisible hand of competition; (*c*) even if markets are competitive, the existence of external effects, public goods, information asymmetries and other market failures ensure that laissez-faire will not bring about the common good; (*d*) and in any case, laissez-faire may produce an intolerable degree of inequality.
The branch of economics called welfare economics is an outgrowth of the fundamental debate that can be traced back to Adam Smith, if not before. It is the economic theory of measuring and promoting social welfare.
This entry is largely organized around three propositions. The first answers this qst: in an economy with competitive buyers and sellers, will the outcome be for the common good? The second addresses the issue of distributional equity, and answers this qst: in an economy where distributional decisions are made by an enlightened sovereign, can the common good be achieved by a slightly modified market mechanism, or must the market be abandoned? The third focuses on the general issue of defining social welfare, or the common good, whether via the market, via a centralized political process, or via a voting process. It answers this qst: does there exist a reliable way to derive the true interests of society, regarding, for example, alternative distributions of income or wealth, from the preferences of individuals?
This entry focuses on theoretical welfare economics. There are related topics in practical welfare economics which are only mentioned here. A reader interested in the practical problems of evaluating policy alternatives can refer to entries on consumer surplus, cost–benefit analysis and compensation principle, to name a few.

The first fundamental theorem, or, laissez-faire leads to the common good

> ‘The greatest meliorator of the world is selfish, huckstering trade.’ (R.W. Emerson, *Work and Days*)

In *The Wealth of Nations*, Book IV, Smith wrote: ‘Every individual necessarily labours to render the annual revenue of the society as great as he can. He generally indeed neither intends to promote the public interest, nor knows how much he is promoting it…. He intends only his own gain, and he is in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention.' The first fundamental theorem of welfare economics can be traced back to these words of Smith. Like much of modern economic theory, the first theorem is set in the context of a Walrasian general equilibrium model, developed almost a hundred years after *The Wealth of Nations*. Since Smith wrote long before the modern mathematical language of economics was invented, he never rigorously stated, let alone proved, any version of the first theorem. That was first done by Lerner (1934), Lange (1942) and Arrow (1951).

To establish the first theorem, we need to sketch a general equilibrium model of an economy. Assume all individuals and firms in the economy are price takers: none is big enough, or motivated enough, to act like a monopolist. Assume each individual chooses his consumption bundle to maximize his utility subject to his budget constraint. Assume each firm chooses its production vector, or input–output vector, to maximize its profits subject to some production constraint. Note that we assume *self-interest*, or the absence of *externalities*: an individual cares only about his own utility, which depends only on his own consumption. A firm cares only about its own profits, which depend only on its own production vector. The invisible hand of competition acts through prices; they contain the information about desire and scarcity that coordinate actions of self-interested agents. In the general equilibrium model, prices adjust to bring about equilibrium in the market for each and every good. That is, prices adjust until supply equals demand. When that has occurred, and all individuals and firms are maximizing utilities and profits, respectively, we have a competitive equilibrium.

The first theorem establishes that a competitive equilibrium is for the common good. But how is the common good defined? The traditional definition looks to a measure of total value of goods and services produced in the economy. In Smith, the 'annual revenue of the society' is maximized. In Pigou (1920), following Smith, the 'free play of self-interest' leads to the greatest 'national dividend'.

However, the modern interpretation of 'common good' typically involves Pareto optimality, rather than maximized gross national product. When ultimate consumers appear in the model, a situation is said to be Pareto optimal if there is no feasible alternative that makes everyone better off. Pareto optimality is thus a dominance concept based on comparisons of vectors of utilities. It rejects the notion that utilities of different individuals can be compared, or that utilities of different individuals can be summed up and two alternative situations compared by looking at summed utilities. When ultimate consumers do not appear in the model, as in the pure production framework to be described below, a situation is said to be Pareto optimal if there is no alternative that results in the production of more of some output, or the use of less of some input, all else equal. Obviously saying that a situation is Pareto optimal is not the same as saying it maximizes GNP, or that it is best in some unique sense. There are generally many Pareto optima. However, optimality is a common good concept that can get common assent: No one would argue that society should settle for a situation that is not optimal, because if A is not optimal, there exists a B that all prefer.

In spite of the multiplicity of optima in a general equilibrium model, most states are non-optimal. If the economy were a dart board and consumption and production decisions were made by throwing darts, the chance of hitting an optimum would be zero. Therefore, to say that the market mechanism leads an economy to an optimal outcome is to say a lot. And now we can turn to a modern formulation of the first theorem:

First fundamental theorem of welfare economics: Assume that all individuals and firms are self-interested price takers. Then a competitive equilibrium is Pareto optimal.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

To illustrate the theorem, we focus on one simple version of it, set in a pure production economy. For a general versions of the theorem, with both production and exchange, the reader can refer to Mas-Colell, Whinston and Green (1995).

In a general equilibrium production economy model, there are *K* firms and *m* goods, but, for simplicity, no consumers. We write $k = 1, 2, \ldots, K$ for the firms, and $j = 1, 2, \ldots, m$ for the goods. Given a list of market prices, each firm chooses a feasible input–output vector y_k so as to maximize its profits. We adopt the usual sign convention for a firm's input–output vector y_k: $y_{kj} < 0$ means firm *k* is a net *user* of good *j*, and $y_{kj} > 0$ means firm *k* is a net *producer* of good *j*. When we add the amounts of good *j* over all the firms, $y_{1j} + y_{2j} + \ldots + y_{Kj}$, we get the aggregate net amount of good *j* produced in the economy, if positive, and an aggregate net amount of good *j* used, if negative. What is feasible for firm *k* is defined by some fixed production possibility set Y_k. Under the sign convention on the input–output vector, if *p* is a vector of prices, firm *k*'s profits are given by

$$\pi_k = p \cdot y_k.$$

A list of feasible input–output vectors $y = (y_1, y_2, \ldots, y_K)$ is called a *production plan* for the economy. A *competitive equilibrium* is a production plan $\hat{y}$ and a price vector *p* such that, for every *k*, $\hat{y}_k$ maximizes π_k subject to y_k's being feasible. (Since the production model abstracts from the ultimate consumers of outputs and providers of inputs, the supply equals demand requirement for an equilibrium is moot.)

If $y = (y_1, y_2, \ldots, y_K)$ and $z = (z_1, z_2, \ldots, z_K)$ are alternative production plans for the economy, *z* is said to *dominate y* if the following vector inequality holds:

$$\sum_k z_k \geq \sum_k y_k.$$

The production plan *y* is said to be *Pareto optimal* if there is other production plan that dominates it. (Note that for two vectors *a* and *b*, $a \geq b$ means $a_j \geq b_j$ for every good *j*, with the strict inequality holding for at least one good.)

We now have the apparatus to state and prove the first theorem in the context of the pure production model:

First fundamental theorem of welfare economics, production version: Assume that all prices are positive, and that $\hat{y}$, *p* is a competitive equilibrium. Then $\hat{y}$ is Pareto optimal.

To see why, suppose to the contrary that a competitive equilibrium production plan $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_K)$ is not optimal. Then there exists a production plan $z = (z_1, z_2, \ldots, z_K)$ that dominates it. Therefore

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\sum_k z_k \geq \sum_k \hat{y}_k .$$

Taking the dot product of both sides with the positive price vector *p* gives

$$p \cdot \sum_k z_k > p \cdot \sum_k \hat{y}_k .$$

But this implies that, for at least one firm *k*,

$$p \cdot z_k > p \cdot \hat{y}_k ,$$

which contradicts the assumption that $\hat{y}_k$ maximizes firm *k*'s profits. Q.E.D.

First fundamental theorem drawbacks, and the second fundamental th

The first theorem of welfare economics is mathematically true but nevertheless open to objections. Here are the commonest.

First, the theorem is an abstraction that ignores the facts. Preferences of consumers are not given, they are created by advertising. The real economy is never in equilibrium, most markets are characterized by excess supply or excess demand, and are in a constant state of flux. The economy is dynamic, tastes and technology are constantly changing, whereas the model assumes they are fixed. The cast of characters in the real economy is constantly changing, the model assumes it fixed.

Second, the theorem assumes competitive behaviour, whereas the real world is full of monopoly and market power.

Third, the theorem assumes there are no externalities. In fact, if in an exchange economy person l's utility depends on person 2's consumption as well as his own, the theorem does not hold. Similarly, if in a production economy firm *k*'s production possibility set depends on the production vector of some other firm, the theorem breaks down. In a similar vein, the theorem assumes there are no public goods, that is, goods like national defence, judicial systems or lighthouses, that are necessarily non-exclusive in use. If such goods are privately provided (as they would be in a completely laissez-faire economy), then their level of production will be suboptimal.

Fourth, the theorem ignores distribution. Laissez-faire may produce a Pareto optimal outcome, but there are many different Pareto optima, and some are fairer than others. Some people are endowed with

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
resources that make them extremely rich, while others, through no fault of their own, are extremely poor. The first and second objections to the first theorem are beyond the scope of this article. The third, regarding externalities and public goods, is one that economists have always acknowledged. The standard remedies for these market failures involve various modifications of the market mechanism, including Pigouvian taxes (Pigou, 1920) on harmful externalities, or appropriate Coasian legal entitlements to, for example, clean air (Coase, 1960).

The important contribution of Pigou is set in a partial equilibrium framework, in which the costs and benefits of a negative externality can be measured in money terms. Suppose that a factory produces gadgets to sell at some market-determined price, and suppose that, as part of its production process, the factory emits smoke which damages another factory located downwind. In order to maximize its profits, the upwind factory will expand its output until its marginal cost equals price. But each additional gadget it produces causes harm to the downwind factory – the marginal external cost of its activity. If the factory manager ignores that marginal external cost, he will create a situation that is non-optimal in the sense that the aggregate net value of both firms' production decisions will not be as great as it could be. That is, what Pigou calls 'social net product' will not be maximized, although 'trade net product' for the polluting firm will be. Pigou's remedy was for the state to eliminate the divergence between trade and social net product by imposing appropriate taxes (or, in the case of beneficial externalities, bounties). The Pigouvian tax would be set equal to marginal external cost, and with it in place the gap between the polluting firm's view of cost and society's view would be closed. Optimality would be re-established.

Coase's contribution was to emphasize the reciprocal nature of externalities and to suggest remedies based on common law doctrines. In his view the polluter damages the pollutee only because of their proximity; for example, the smoking factory harms the other only if it happens to locate close downwind. Coase rejects the notion that the state must step in and tax the polluter. The common law of nuisance can be used instead. If the law provides a clear right for the upwind factory to emit smoke, the downwind factory can contract with the upwind factory to reduce its output, and if there are no impediments to bargaining, the two firms acting together will negotiate an optimal outcome. Alternatively, if the law establishes a clear right for the downwind factory to recover for smoke damages, it will collect external costs from the polluter, and thereby motivate the polluter to reduce its output to the optimal level. In short, a legal system that grants clear rights to the air to either the polluter or pollutee will set the stage for an optimal outcome, provided that bargaining is costless. If bargaining is costly, then the law should be designed with an eye towards minimizing social costs created by the externality.

With respect to public goods, since Samuelson (1954) derived formal optimality conditions for their provision, the issue has received much attention from economists; one especially notable theoretical question has to do with discovering the strengths of people's preferences for a public good. If the government supplies a public judicial system, for instance, how much should it spend on it (and tax for it)? At least since Samuelson, it has been known that financing schemes like those proposed by Lindahl (1919), where an individual's tax is set equal to his marginal benefit, provide perverse incentives for people to misrepresent their preferences. Schemes that are immune to such misrepresentations (in certain circumstances) have been developed (Clarke, 1971; Groves and Loeb, 1975).

But it is the fourth objection to the first theorem that may be most fundamental. What about distribution? There are two polar approaches to rectifying the distributional inequities of laissez-faire. The first is the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
command economy approach: a central bureaucracy makes detailed decisions about the consumption decisions of all individuals and production decisions of all producers. The main theoretical problem with the command approach is that it fails to create appropriate incentives for individuals and firms. On the empirical side, the experience of the late Soviet and Maoist command economies establish that highly centralized economic decision making leaves much to be desired, to put it mildly.

The second polar approach to solving distribution problems is to transfer income or purchasing power among individuals, and then to let the market work. The only kind of purchasing power transfer that does not cause incentive-related losses is the lump-sum money transfer. Enter at this point the standard remedy for distribution problems, as put forward by market-oriented economists, and our second major theorem.

The second fundamental theorem of welfare economics establishes that the market mechanism, modified by the addition of lump-sum transfers, can achieve virtually *any* desired optimal distribution. Under more stringent conditions than are necessary for the first theorem, including assumptions regarding quasi-concavity of utility functions and convexity of production possibility sets, the second theorem gives the following:

Second fundamental theorem of welfare economics: Assume that all individuals and producers are self-interested price takers. Then almost any Pareto optimal equilibrium can be achieved via the competitive mechanism, provided appropriate lump-sum taxes and transfers are imposed on individuals and firms.

One version of the second theorem, restricted to a pure production economy, is particularly relevant to an old debate about the feasibility of socialism; see particularly Lange and Taylor (1939) and Lerner (1944). Anti-socialists including von Mises (1922) argued that informational problems would make it impossible to coordinate production in a socialist economy; while pro-socialists, particularly Lange, argued that those problems could be overcome by a central planning board, which limited its role to merely announcing a price vector. This was called ‘decentralized socialism’. Given the prices, managers of production units would act like their capitalist counterparts; in essence, they would maximize profits. By choosing the price vectors appropriately, the central planning board could achieve any optimal production plan it wished.

In terms of the production model given above, the production version of the second theorem is as follows:

Second fundamental theorem of welfare economics, production version: Let $\hat{y}$ be any optimal production plan for the economy. Then there exists a price vector p such that $\hat{y}$, p is a competitive equilibrium. That is, for every k, $\hat{y}_k$ maximizes $\pi_k = p \cdot y_k$ subject to y_k being feasible.

The proof of the second theorem will not be presented here.

Adjusting the economy and voting

We rarely choose between a laissez-faire economy and a command economy. Our choices are almost always more modest. When choosing among alternative tax policies, or trade and tariff policies, or development policies, or anti-monopoly policies, or labour policies, or transfer policies, what shall guide the choice? The applied welfare economist's advice is usually based on some notion of increasing total output in the economy. The practical political decision, in a democracy, is normally based on voting.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Applied welfare economics

The applied welfare economist usually focuses on ways to increase total output, ‘the size of the pie’, or at least to measure changes in the size of the pie. Unfortunately, theory suggests that the pie cannot be easily measured. This is so for a number of reasons. To start, any measure of total output is a scalar, that is, a single number. If the number is found by adding up utility levels for different individuals, illegitimate interpersonal utility comparisons are being made. If the number is found by adding up the values of aggregate net outputs of all goods, there is an index number problem. The value of a production plan will depend on the price vector at which it is evaluated. But, in a general equilibrium context, the price vector will depend on the aggregate net output vector, which will in turn depend on the distribution of ownership or wealth among individuals.

An early and crucial contribution to the analysis of whether or not the economic pie has increased in size was made by Kaldor (1939, p. 550), who argued that the repeal of the Corn Laws in England could be justified on the grounds that the winners might in theory compensate the losers: ‘it is quite sufficient [for the economist] to show that even if all those who suffer as a result are fully compensated for their loss, the rest of the community will still be better off than before’. Unfortunately, Scitovsky (1941) quickly pointed out that Kaldor's compensation criterion (as well as one proposed around the same time by Hicks) was inconsistent. Consider a move from situation A to situation B. It is possible to judge B Kaldor superior to A (the move is an improvement) and simultaneously judge A Kaldor superior to B (the move back would also be an improvement). This Scitovsky paradox can be avoided via a two-edged compensation test, according to which B is judged better than A if (a) the potential gainers in the move from A to B could compensate the potential losers, and still remain better off, and (b) the potential losers could not bribe the gainers to forgo the move.

However, while Scitovsky's two-edged criterion has some logical appeal, it still has a major drawback: it ignores distribution. Therefore, it can make no judgement about alternative distributions of the same size pie. Even worse, both the Kaldor and the Scitovsky criteria would approve of a change that makes the wealthiest man in society richer by $1 billion, while making each of the million poorest people worse off by $999. This is a judgement that many people would reject as wrong or immoral.

Another important tool for measuring changes in economic welfare is the concept of consumer's surplus, which Marshall (1920, book 3, ch. 6) defined as the difference between what an individual would be willing to pay for an object, at most, and what he actually does pay. With a little faith, the economic analyst can measure aggregate consumers’ surplus (note the new position of the apostrophe) by calculating an area under a demand curve, and this is in fact commonly done in order to evaluate changes in economic policy. The applied welfare economist attempts to judge whether the pie would grow in a move from A to B by examining the change in consumers’ surplus (plus profits, if they enter the analysis). Some faith is required because consumers’ surplus, like the Kaldor criterion, is theoretically inconsistent; see for example Boadway (1974).

Under certain circumstances, however, consumers’ surplus inconsistencies can be ruled out. In particular, if individual utility functions are all quasilinear, of the form $u_i(x_i) = v_i(x_{ij, j \neq m}) + x_{im}$, then consumers’ surplus paradoxes disappear. (Here $u_i(x_i)$ is person i's utility, as a function of his consumption bundle $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, and the utility function can be separated into two parts, the first one of which is a function $v_i(\cdot)$ which depends on quantities of all the goods *except* the mth, and the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
second of which is simply the quantity of the *m*th good. The mth good can be interpreted as the ‘money’ good; all the individuals like it, and value it the same way, with the same marginal utility, of one.) The assumption of quasilinear preferences is a very strong one if we think about ‘real’ commodities like wine and bread, but it has a certain intuitive appeal if we are inclined to believe in utility from ‘money’. To sum up this section, although the tools of applied welfare economics are widely used and very important in practice, in theory they should be viewed with some skepticism.

Voting

In many cases, interesting decisions about economic policies are made either by government bureaucracies that are controlled by legislative bodies, or by legislative bodies themselves, or by elected executives: in short, either directly or indirectly, by voting. The second theorem itself raises qsts about distribution that many would view as essentially political. How should society choose the Pareto-optimal allocation of goods that is to be reached via the modified competitive mechanism? How should the distribution of income be chosen? How can the best distribution of income be chosen from among many Pareto optimal ones? Majority rule is a commonly used method of choice in a democracy, both for political choices and economic ones, and we now turn our attention to it.

The practical objections to voting, the fraud, the deception, the accidents of weather, are well known. To quote Boss Tweed, the infamous 19th century chief of New York's Tammany Hall: ‘As long as I count the votes, what are you going to do about it?’ We will examine the theoretical problems.

The central theoretical problem with majority voting has been known since the time of Condorcet's *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, published in 1785: Voting may be logically inconsistent. The now standard Condorcet voting paradox assumes three individuals 1, 2 and 3, and three alternatives *x*, *y* and *z*, where the three voters have the following preferences:

1:	x	y	z
2:	y	z	x
3:	z	x	y

(Following an individual's number the alternatives are listed in his order of preference, from left to right.) Majority voting between pairs of alternatives will reveal that *x* beats *y*, *y* beats *z*, and, paradoxically, *z* beats *x*.

It is now clear that such voting cycles are not peculiar; they are generic, particularly when the alternatives have a spatial aspect with two or more dimensions (Plott, 1967; Kramer, 1973). This can be illustrated by taking the alternatives to be different distributions of one economic pie. Suppose, in other words, that the distributional issues raised by the first and second theorems are to be ‘solved’ by majority voting, and assume for simplicity that what is to be divided is a fixed total of wealth, say 100 units.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Now let x be 50 units for person 1, 30 units for person 2 and 20 units for person 3. That is, let $x = (50, 20, 30)$. Similarly, let $y = (30, 50, 20)$ and $z = (20, 30, 50)$. The result is that our three individuals have precisely the voting paradox preferences. Nor is this result contrived, it turns out that *all* the distributions of 100 units of wealth are connected by endless voting cycles (see McKelvey, 1976). The reader can easily confirm that for any distributions u and v, that he may choose, there exists a voting sequence from u to v, and another back from v to u!

The reality of voting cycles should give pause to the economist who recommends legislation about economic choices, especially choices among alternative distributions of income or wealth.

Social welfare and the third fundamental th

How then can economic choices be made; how, for example, might the distribution problem be solved? One potential answer is to assert the existence of a Bergson (1938) economic welfare function $E(\cdot)$, that depends on the amounts of non-labour factors of production employed by each producing unit, the amounts of labour supplied by each individual, and the amounts of produced goods consumed by each individual. Then solve the problem by maximizing $E(\cdot)$. If necessary conditions for Pareto optimality are derived that must hold for any $E(\cdot)$, this exercise is harmless enough; but if a *particular* $E(\cdot)$ is assumed and distributional implications are derived from it, then an objection can be raised: why that Bergson function $E(\cdot)$, and not a different one?

At first, in his modestly titled ‘A Difficulty in the Concept of Social Welfare’ (1950), and later, in his classic monograph *Social Choice and Individual Values* (1963), Kenneth Arrow brought together both the economic and political streams of thought sketched above. Arrow's theorem can be viewed in several ways: it is a statement about the distributional qsts raised by the first and second theorems; it is an extension of the Condorcet voting paradox; it is a statement about the logic of voting; and it is a statement about the logic of Bergson welfare functions, compensation tests, consumers’ surplus tests, and indeed all the tools of the applied welfare economist. Because of its central importance, Arrow's theorem can be justifiably called the third fundamental theorem of welfare economics.

Arrow's analysis is at a high level of abstraction, and requires some additional model building. From this point onward we assume a given set of at least three distinct alternatives, which might be allocations in an exchange economy, distributions of wealth, tax bills in a legislature, or candidates in an election. The alternatives are written x, y, z and so on. We assume a fixed society of individuals, numbered 1, 2,…, n. Let R_i represent the preference relation of individual i, so xR_iy means person i likes x as well as or better than y. (Strict preference is shown with a P_i, and indifference with a I_i.) A *preference profile* for society is a specification of preferences for each and every individual, or symbolically, $R = (R_1, R_2, \ldots, R_n)$. We write R_S for *society’s* preference relation, arrived at in a way yet to be specified.

Arrow was concerned with the logic of how individual preferences are transformed into social preferences. That is, how is R_S found? Symbolically we can represent the transformation this way:

$$R_1, R_2, \ldots, R_n \rightarrow R_S.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Now if society is to make decisions regarding things like distributions of wealth, it must 'know' when one alternative is as good as or better than another, even if both are Pareto optimal. To ensure it can make such decisions, Arrow requires that R_S be *complete*. That is, for any alternatives x and y, either xR_Sy or yR_Sx (or both, if society is indifferent between the two). If society is to avoid the illogic of voting cycles, its preferences ought to be *transitive*. That is, for any alternatives x, y and z, if xR_Sy and yR_Sz, then xR_Sz. Following Sen (1970), we call a transformation of preference profiles into complete and transitive social preference relations an *Arrow social welfare function*, or more briefly, an Arrow SWF. Anyone can make up an Arrow SWF, just as anyone can make up a Bergson function, or for that matter a simple moral judgement about when one distribution of wealth is better than another. But arbitrary judgments are unsatisfactory and so are arbitrary Arrow functions. Therefore, Arrow imposed some reasonable conditions on his function. Following Sen's (1970) version of Arrow's theorem, there are four conditions:

1. 1. Universality. The function should always work, no matter what individual preferences might be. It would not be satisfactory, for example, to require unanimous agreement among all the individuals before determining social preferences.
2. 2. Pareto consistency. Social preferences should be consistent with the *Pareto criterion*. That is, if everyone prefers x to y, then the social preference is x over y.
3. 3. Independence. Suppose there are two alternative preference profiles for individuals in society, but suppose individual preferences regarding x and y are exactly the same under the two alternatives. Then the social preference regarding x and y must be exactly the same under the two alternatives. In particular, if individuals change their minds about a third 'irrelevant' alternative, this should not affect the social preference regarding x and y.
4. 4. Non-dictatorship. There should not be a dictator. In Arrow's abstract model, person i is a *dictator* if society always prefers what he prefers; that is, if i prefers x to y, then the social preference is x over y.

An economist or policymaker who wants an ultimate answer to qsts involving distributions, or qsts involving choices among alternatives that are not comparable under the Pareto criterion, could use an Arrow SWF for guidance. Unfortunately, Arrow showed that imposing conditions 1 to 4 guarantees that Arrow functions *do not exist*:
Third fundamental theorem of welfare economics: There is no Arrow social welfare function that satisfies the conditions of universality, Pareto consistency, independence, non-dictatorship.
In order to illustrate the logic of the theorem, we will use a stronger assumption than independence. This assumption is called *neutrality–independence–monotonicity* (NIM), defined as follows. Suppose for some group of individuals V, some preference profile, and some pair of alternatives x and y, all members of V prefer x to y, all individuals *not* in V prefer y to x, and the social preference is x over y. Then for *any* preference profile and *any* pair of alternatives w and z, if all people in V prefer w to z, the social preference must be w over z. In short, if V gets its way in one instance, when everyone opposes it, then it must have the power to do it again, when the opposition may be weaker.
A group of individuals V is said to be decisive if for all alternatives x and y, whenever all the people in V

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

prefer *x* to *y*, society prefers *x* to *y*. Note that if person *i* is a dictator, $\{i\}$ is decisive, and conversely. Assumption NIM asserts that, if *V* prevails when it is opposed by everyone else, it must be *decisive*. If the social choice procedure is majority rule, for example, any group of $(n+1)/2$ members, for n odd, or $(n/2)+1$ members, for n even, is decisive. Moreover, it is clear that majority rule satisfies the NIM assumption, since if *V* prevails for a particular *x* and *y* when everyone outside of *V* prefers *y* to *x*, then *V* must be a majority, and must always prevail. (Majority rule is just one example of a procedure that satisfies NIM; there are many other procedures that also do so.)

Now we are ready to turn to a short and simple version of the third theorem.

Third fundamental theorem of welfare economics, short version: There is no Arrow SWF that satisfies the conditions of universality, Pareto consistency, neutrality–independence–monotonicity, and non-dictatorship.

The proof goes as follows. First, there must exist decisive groups of individuals, since by the Pareto consistency requirement the set of all individuals is one. Now let *V* be a decisive group of minimal size. If there is just one person in *V*, he is a dictator. Suppose then that *V* includes more than one person. We show this leads to a contradiction.

If there are two or more people in *V*, we can divide it into non-empty subsets *V*1 and *V*2. Let *V*3 represent all the people who are in neither *V*1 nor *V*2. (*V*3 may be empty). By universality, the Arrow function must be applicable to any profile of individual preferences. Take three alternatives *x*, *y* and *z* and consider the following preferences regarding them:

For individuals in V_1: $x \; y \; z$
For individuals in V_2: $y \; z \; x$
For individuals in V_3: $z \; x \; y$

(At this point the close tie between Arrow and Condorcet is clear, for these are exactly the voting paradox preferences!)

Since *V* is by assumption decisive, *y* must be socially preferred to *z*, which we write yP_Sz. By the assumption of completeness for the social preference relation, either xR_Sy or yP_Sx must hold. If xR_Sy holds, since xR_Sy and yP_Sz, then xP_Sz must hold by transitivity. But now V_1 is decisive by the NIM assumption, contradicting *V*'s minimality. Alternatively, if yP_Sx holds, V_2 is decisive by the NIM assumption, again contradicting *V*'s minimality. In either case, the assumption that *V* has two or more people leads to a contradiction. Therefore *V* must contain just one person, who is, of course, a dictator! Q.E.D.

Since the third theorem was discovered, a whole literature of modifications and variations has been spawned. But the depressing conclusion has remained more or less the same: there is no logically infallible way to aggregate the preferences of diverse individuals into a social preference relation. Therefore, there are no logically infallible ways to vote, or to solve the problems of distribution of income and wealth in society.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Social welfare after Arrow

Social choice functions and strategy

The third fundamental theorem of welfare economics tells us that we cannot find an Arrow social welfare function satisfying certain reasonable requirements. An Arrow function maps preference profiles (that is, preference relations for each and every member of society) into social preference relations. But in order to make judgements about what alternative is *best* for society, it is not really necessary to have a social preference relation. Suppose we just had a rule that tells us, if the set of alternatives is *x*, *y*, *z* and so on, and the preference profile is $R = (R_1, R_2, \ldots, R_n)$, then the best alternative is such-and-such? Such a rule would be a mapping from preference profiles into alternatives, written symbolically as follows:

$$R_1, R_2, \ldots, R_n \rightarrow x.$$

Such a rule is called a *social choice function*, or SCF for short. An Arrow function produces a social ranking of all the alternatives; an SCF, in contrast, just produces a winner. As an example, think of plurality voting, with some kind of rule to break ties.

The essential difficulty with SCFs is that they may create obvious incentives for people to misrepresent their preferences, so as to obtain better (for them) social choices. As an example, consider again the Condorcet voting paradox preferences:

$$\begin{array}{lccc} 1: & x & y & z \\ 2: & y & z & x \\ 3: & z & x & y \end{array}$$

Suppose the three people use plurality voting (each person casts one vote for his favorite), and, in case of a tie, the social choice is the outcome closest to the beginning of the alphabet. Under this rule, if 1 votes for his favourite, *x*, and persons 2 and 3 do likewise, there is a three-way tie, which is resolved with the (alphabetical) choice of *x*. Now put yourself in the shoes of person 2. You will immediately see that, if persons 1 and 3 continue to vote for their favourites, and if you switch from your favourite *y* to your second favourite *z*, then social choice changes, from *x* to *z*, making you better off! You are in effect being asked ‘what is your preference relation?’ Instead of answering honestly (*y z x*), you offer, in effect, a false preference relation (*z y x*).

Reporting a false preference relation in order to bring about an SCF outcome that you prefer to the one you get if you are honest, is called *strategic behaviour*, or *strategizing*. It is obviously a bad thing if an SCF produces lots of opportunities for strategic behaviour: if individuals are commonly strategizing,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
there is no reason to believe that the outcome, based as it is on false reports, is truly best for society. If an SCF has the property that it is never advantageous for anyone to report a false preference relation it is called *strategy-proof*. For instance, suppose an SCF always chooses the alternative that is first in the alphabetical list of alternatives. This SCF would be frustrating and idiotic, but it would be strategy-proof.

The Gibbard – Satterthwaite th

This leads to a natural qst: are there SCFs that are immune to strategic behaviour, and that satisfy a few other reasonable conditions? Note that the question is very similar in style to the question that Arrow asked about Arrow SWFs. What would the reasonable conditions be? First (similar to Arrow), the SCF ought to be *universal*; that is, it should work no matter what the profile of individual preferences might be. Second (also similar to Arrow), there should be no dictator. In the SCF context person *i* is a *dictator* if the social choice is always a top-ranked alternative for person *i*. Third (and different from Arrow), the SCF should be *non-degenerate*. This means that, for any alternative *x*, there must be some preference profile which would give rise to *x*'s being the social choice. (This requirement excludes the SCF that always chooses the first alternative in the alphabetical list.) Now we can ask the qst: do there exist SCFs which are universal, non-degenerate, strategy-proof and non-dictatorial?

This question was asked and answered, independently, by Gibbard (1973) and Satterthwaite (1975). The Gibbard–Satterthwaite result turns out to be logically very close to third fundamental theorem of welfare economics; in fact, Gibbard uses Arrow's theorem to prove his theorem, and Satterthwaite shows that his theorem can be used to prove Arrow's. Following is the Gibbard–Satterthwaite theorem. The proof is omitted; a simplified and restricted version of the theorem, and a simple proof, can be found in Feldman and Serrano (2006):

Gibbard–Satterthwaite th: There is no social choice function that satisfies the conditions of universality, non-degeneracy, strategy-proofness and non-dictatorship.

Like the third fundamental theorem, the Gibbard–Satterthwaite theorem is starkly negative; it says that, if you want a decision-making process, an SCF to be precise, that has desirable characteristics, including being immune to strategic manipulation, you are bound to be disappointed. To put it differently, for any reasonable SCF, there will be circumstances under which some person will want to falsely report his preferences, resulting in a perversion of the process, and an outcome that may not be desirable for society.

If a decision-making process works in a way that offers each individual *no* incentive to misrepresent his preferences, no matter what preferences the other $n-1$ individuals might be reporting, we say that honestly reporting one's preferences (or telling the truth) is a *dominant strategy*. The Gibbard–Satterthwaite result then says that, if a social choice function satisfies the conditions of universality, non-degeneracy and non-dictatorship, truth-telling will not be a dominant strategy. That is, there will be some reported preference relations of all individuals except *i*, which will provide an incentive for individual *i* to lie. If everyone else is saying such-and-such (which might be true or false), person *i* will give a false report. This is what strategy-proofness excludes.

But what if we narrowed this broad notion of strategy-proofness; what if we required that *i* not have an incentive to lie when the others are reporting the truth, rather than requiring that *i* never have an incentive to lie, no matter what the others are reporting?

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Implementation and the Maskin th

If telling the truth is a best strategy when others are telling the truth, rather than always, then truth telling is a *Nash strategy*, rather than a *dominant strategy*. The theory of implementation, or mechanism design, provides a way out the negativity of the Gibbard–Satterthwaite theorem; it provides a way to implement an SCF, or support its choices, by incorporating truth telling about preferences into Nash equilibrium strategies in games.

The major theorem on implementation is due to Maskin (1999), whose paper first circulated in 1977. In Maskin's model, there is a social planner (or central authority) who wants to bring about choices according to a given SCF, which we now call *F*. The planner knows *F*, as do all the members of society. Given any preference profile *R*, the SCF produces an outcome $F(R) = x$. But the planner does not know the true preferences of the individuals. He must rely on the individuals to report their preferences, and they may lie. We assume for simplicity that every individual knows the true preference relation for himself *and* every other individual; that is, each *i* knows the true preference profile, but the social planner doesn't. (This is obviously a strong assumption.) From this point on, when one preference profile may be true and another may be false, we will use the unadorned *R* to represent the *true* profile. The social planner receives reports on preferences, or preference profiles, from the individuals, but they may be lies. We let $\hat{R}_i$ represent a reported preference *relation for* person *i*, which may be false; similarly $\hat{R} = (\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_n)$ represents a reported preference *profile*, which may be false; and $\hat{R}^i = (\hat{R}^i_1, \hat{R}^i_2, \ldots, \hat{R}^i_n)$ represents a preference *profile, reported by person i*, which may false. The social planner wants to devise a method, a mechanism, to induce individuals to honestly report preferences. That way he will get hold of the true preference profile *R*, and produce the desired outcome $F(R) = x$. How might this be done? The intuition is to ask each and every individual to report a preference profile. (Note that, since we assume all the individuals know each other's true preferences, it is no more challenging for an individual to report a preference profile, comprising preference relations for everyone in society, than it is to report his own preference relation.) If all the reported preferences profiles agree, there's a good chance they are all true, and the planner might accept the generally agreed-upon profile. If they all agree except for one, the one that's out of line probably comes from a liar, and he should be given a motive to avoid lying. (If the social planner were a despot, the out-of-line person would be shot. Note also that there must be three or more individuals in society to discover whose report is out of line.) Finally, when the reported preference profiles generally disagree, the social planner needs a way to avoid having the process stop at an inappropriate Nash equilibrium.

Let us be more precise. Maskin's algorithm for implementing an SCF works as follows. Each person *i* reports a message m_i, which is composed of an alternative *x*, a preference profile $\hat{R}^i$, and a non-negative integer. (i) If every message is the same, of the form $(x = F(\hat{R}), \hat{R}, 0)$, then the social planner chooses *x*. (ii) If every message but one is the same, of the form $(x = F(\hat{R}), \hat{R}, 0)$ for every person but *j*, while *j* reports a message of the form ($y \neq x, \hat{R}^j$, anything), then the social planner chooses *y*, unless person *j* would like *x less than y* according to $\hat{R}_j$, the person-*j* preference relation that all the other people are reporting. If this is the case the planner chooses *x*. (Person *j* is not shot. He simply does not gain, and

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
may lose, from his deviation.) (iii) In all other cases, the social planner finds the person who proposes the highest integer (with some method for resolving ties), and chooses the alternative named by that person.

Now the qsts can be framed. First, given this mechanism, would $m_1 = m_2 = \ldots = m_n = (F(R), R, 0)$, with *R* the true preference profile, constitute a Nash equilibrium? Second, if $(m_1, \ldots, m_n)$ is any Nash equilibrium list of messages in this mechanism, can we be sure the resulting chosen alternative will be *F* (*R*)?

The answers are ‘yes’ and ‘yes’, under certain general assumptions. The assumptions of Maskin's theorem are as follows. First, there must be three or more individuals (so that a deviant message can be spotted). Second, a mild diversity condition must be satisfied. Maskin uses a condition called *no veto*. Loosely speaking, this means that, if $n - 1$ people prefer *x* to all the other alternatives, then the SCF must choose *x*. Alternatively, one can assume the existence of a private economic good, that everyone values. This guarantees that individuals will disagree about what alternatives are best. In this article we will simply assume *diversity*, meaning the following: for any given alternative *x*, there exist at least two people, each of whom prefers something else to *x*.

Third, the social choice function *F* must satisfy an intuitive condition called *Maskin monotonicity*. (The condition is actually a distant relative of the NIM assumption used in the simple version of Arrow's theorem presented above.) Maskin monotonicity means the following. Let $\hat{R}$ and *R* be any two preference profiles. (These may be true or false; it does not matter in this context.) Suppose $F(\hat{R}) = x$, and suppose that, for all individuals *i* and all alternatives *y*, $x\hat{R}_i y$ implies $xR_i y$. Then $F(R) = x$. (In other words, in a hypothetical transition from $\hat{R}_i$ to R_i, for every person *i* the set of alternatives that *i* likes less than *x* or the same as *x* has expanded, or at least hasn't shrunk. Since *x* was the social choice under $\hat{R}$, *x* must continue to be the social choice under *R*.) With these three conditions, Maskin proved:

Maskin th: Assume $n \geq 3$. Assume diversity and Maskin monotonicity. Then the mechanism described above implements the SCF *F*, in the sense that truthful messages leading to *F*(*R*) comprise a Nash equilibrium, and in the sense that any Nash equilibrium list of messages results in the social planner choosing *F*(*R*).

We will not provide all of the proof, but the logic is as follows. First, establish that $m_1 = m_2 = \ldots = m_n = (F(R), R, 0)$ is a Nash equilibrium, where *R* is the true preference profile. This is rather obvious, given rules (i) and (ii) of the Maskin algorithm. Second, establish that under rules (ii) and (iii), there are no Nash equilibria. This follows rather easily from the diversity assumption. Third, establish that, if $m_1 = m_2 = \ldots = m_n = (F(\hat{R}), \hat{R}, 0)$ is any Nash equilibrium, then $F(\hat{R}) = F(R)$. That is, given a Nash equilibrium based on a universally reported, but possibly false, preference profile, the outcome implemented is the same as if the true preference profile had been reported. This follows from the assumption of Maskin monotonicity.

Maskin also provided a near converse this theorem, which says that Maskin monotonicity is a necessary condition for any SCF *F* to be implementable. Relatively simple proofs of both Maskin theorems are available in Feldman and Serrano (2006).

Last words

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Where does welfare economics now stand? The first and second theorems are encouraging results that suggest the market mechanism has great virtue: competitive equilibrium and Pareto optimality are firmly bound. The third theorem exposes impossibilities and paradoxes in economic choices, voting choices, and, in general, almost any choices made collectively by society. The Gibbard–Satterthwaite theorem, like the third theorem, is a starkly negative result: any plausible social choice function will, under some circumstances, produce incentives for someone to lie. But the Maskin theorem is a ray of hope; it suggests a way for a social planner to design a game, whose Nash equilibria will implement a desired social choice function.

See Also

- Arrow's theorem
- compensation principle
- consumer surplus
- externalities
- general equilibrium
- mechanism design
- social choice

Bibliography

Arrow, K.J. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy* 58, 328–46.

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J. 1963. *Social Choice and Individual Values*, 2nd edn. New York: John Wiley and Sons.

Bergson, A. 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52, 310–34.

Boadway, R. 1974. The welfare foundations of cost–benefit analysis. *Economic Journal* 84, 926–39.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11, 17–33.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Feldman, A.M. and Serrano, R. 2006. *Welfare Economics and Social Choice Theory*, 2nd edn. New York: Springer.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Groves, T. and Loeb, M. 1975. Incentives and public inputs. *Journal of Public Economics* 4, 211–26.

Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49, 549–52.

Kramer, G.H. 1973. On a class of equilibrium conditions for majority rule. *Econometrica* 41, 285–97.

Lange, O. 1942. The foundations of welfare economics. *Econometrica* 10, 215–28.

Lange, O. and Taylor, F.M. 1939. *On the Economic Theory of Socialism*. Minneapolis: University of Minnesota Press.

Lerner, A.P. 1934. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1, 157–75.

Lerner, A.P. 1944. *The Economics of Control*. New York: The Macmillan Company.

Lindahl, E. 1919. Just taxation – a positive solution. Trans. and reprinted In *Classics in the Theory of Public Finance*, ed. R.A. Musgrave and A.T. Peacock. New York: Macmillan, 1958.

Marshall, A. 1920. *Principles of Economics*, 8th edn. London: Macmillan.

Mas-Colell, A., Whinston, M.D. and Green, J.R. 1995. *Microeconomic Theory*. New York: Oxford University Press.

Maskin, E. 1999. Nash equilibrium and welfare optimality. *Review of Economic Studies* 66, 23–38.

McKelvey, R. 1976. Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic Theory* 12, 472–82.

Mises, L. von. 1922. *Socialism: An Economic and Social Analysis*. 3rd edn., trans., Indianapolis: Liberty Classics, 1981.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan, Part II.

Plott, C.R. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57, 787–806.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
387–9.

Satterthwaite, M.A. 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10, 187–217.

Scitovsky, T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* 9, 77–88.

Sen, A.K. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

How to cite this article

Feldman, Allan M. "welfare economics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000050>
doi:10.1057/9780230226203.1826

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

welfare state

Assar Lindbeck
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article starts out with a brief discussion of the historical background, the justifications and the political forces behind the creation of the modern welfare state. It also summarizes its major achievements in terms of economic efficiency and redistribution. The article also tries to identify some major problems of contemporary welfare-state arrangements, distinguishing problems caused by exogenous shocks from those related to endogenous behaviour adjustments by individuals to the welfare state itself. The latter include tax distortions, moral hazard, and endogenous changes in social norms concerning work and benefit dependency.

Keywords

advantageous selection; adverse selection; aging and retirement; altruism; Baumol, W.; birth rates; child care; civil society; dependency; direct provision vs subsidy; distribution of income; early retirement pensions; education; employment subsidies; enlightened self-interest; family; free riding; Gini coefficient; globalization; health care; human capital investment; human services; immigration; income effects; income insurance; income smoothing; individualization; industrialization; intergenerational contracting; intergenerational transfers; international trade (theory); intra-individual redistribution; job-security legislation; labour supply; labour-intensive human services; labour-market programmes; leisure; Luxembourg Income Study; marginal tax rates; market failure; moral hazard; myopia; non-governmental organizations; non-profit organizations; old-age care; political economy of income redistribution; preschooling; productivity growth; public social spending; redistribution of income and wealth; retraining; risk sharing; sickness benefits; social insurance; social norms; structural change; structural unemployment; substitution effect; Sweden; tax base; tax competition; tax distortions; tax wedge; transfer payments; unemployment benefits; universal vs. targeted transfers; urbanization; user fees; voucher systems; welfare state; working poor

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Article

According to a narrow definition, the welfare state comprises two types of government spending arrangements: (*a*) cash benefits to households (transfers, including mandatory income insurance) and (*b*) subsidies or direct government provision of human services (such as child care, preschooling, education, health care and old-age care). By broader definitions, the welfare state may also include price regulation (such as rent control and agricultural price support), housing policies, regulation of the work environment, job-security legislation, and environmental policies. This article is confined to the narrow definition.

Across developed Organisation for Economic Co-operation and Development (OECD) countries, total welfare-state spending (‘public social spending’), including spending on education, varies today (2006) from about a fifth of GDP to about a third. As we would expect, the share is tightly related to the degree of ‘universality’ of public social spending, that is, the extent to which benefits are extended to individuals in all income classes rather than largely targeted to particular groups of individuals, such as low-income groups. Broadly speaking, the lowest figures are currently found in Anglo-Saxon countries, while the highest appear in the Nordic countries – with other countries in Western Europe somewhere in-between. Indeed, nowadays welfare states are usually classified in the context of such geographical clusters rather than according to distinctions between Bismarck- and Beveridge-type welfare states, or distinctions in terms of ideological categories along the lines suggested by Esping-Andersen (1990).

Justifications and explanations

Urbanization has diminished the reliability of the family as a basis for reallocating income (or consumption) over the individual's life cycle, reducing income risk, and providing human services. Moreover, in connection with industrialization, new types of labour contracts emerged according to which unemployment and retirement became more discrete (abrupt) events than earlier (Atkinson, 1991). Industrialization and the subsequent increase in office work also required an expansion of education at all levels. Meanwhile, progress in health and medicine enhanced the usefulness of professional medical services.

Needless to say, such developments by themselves do not justify government intervention in the fields of income insurance and human services, rather than simply leaving people to rely on voluntary solutions via markets and private networks (‘civil society’). There are, however, well-known efficiency and distributional justifications for government intervention in these fields (see, for instance, Barr, 1998). It is useful to divide the *efficiency justifications* into three categories.

First, the microeconomic literature identifies a number of limitations (‘failures’) in markets for voluntary income insurance: advantageous selection (‘cream skimming’) of insurance applicants, when insurance providers can differentiate between low-risk and high-risk individuals; adverse selection, when insurance providers cannot do so; myopia, when individuals underestimate their future income needs; and free riding on the altruism of others, when individuals expect others to assist them in case of economic distress. Mandatory income insurance (‘social insurance’) helps solve all these problems. Moreover, poor individuals may simply believe that they cannot afford to save or to buy income insurance: their marginal evaluation of immediate consumption is higher than their marginal evaluation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of future income security. Paternalistic governments may prefer to deal with this issue by mandatory insurance rather than by cash transfers to such individuals. In addition, a monopoly provider may largely avoid marketing costs – although at the expense of individuals' freedom of choice.

Even if some of these problems may also be mitigated by group insurance, such arrangements are associated with well-known weaknesses. For instance, occupational income insurance often results in limited portability across jobs, and sometimes deficient financial viability, in particular when individual production firms or industries are in charge of the programmes. In some countries, however, such problems are avoided by institutional integration of occupational and government-operated arrangements ('corporatist' systems), such as in Germany and France.

Second, mandatory income insurance may also bring about risk sharing across generations. This is difficult to achieve by voluntary contracts alone since the potential parties of such contracts may not live simultaneously – both when the contract is signed and when it is supposed to be fulfilled.

Third, economists generally agree that investment in human capital (such as education and health care) tends to be suboptimal without government intervention (in the form of subsidies or direct government provision), either because of the difficulties in borrowing with expected future human capital as collateral or because of unexploited (positive) externalities in connection with such investment.

While the efficiency gains from government intervention in the context of the first two justifications show up in improved income smoothing and risk sharing, the efficiency gain according to the third justification takes the form of higher labour productivity and/or faster economic growth – provided disincentives due to higher government spending do not dominate these potential efficiency gains.

The *distributional justifications* for welfare-state arrangements also appear in different forms.

First, in the case of policies designed to fight poverty, it is natural to refer to genuine altruism or enlightened self-interest (a desire to mitigate negative externalities, such as ugly neighbourhoods and street crime). Intergenerational transfers in favour of old cohorts – for instance, via a pay-as-you-go (paygo) pension system – may also be justified by altruism, since lifetime income tends to be lower for older cohorts than for subsequent cohorts in growing economies.

Second, income insurance automatically reduces the overall dispersion of the *ex post* distribution of income. This holds for both yearly and lifetime income. Moreover, social insurance, as usually designed, may often reduce the *ex ante* dispersion of the distribution since such arrangements are seldom actuarially fair. A fairly common belief is that increased income security, and perhaps also a reduction in the overall dispersion of the distribution of income, up to a point tends to promote social peace, and that this in turn is favourable for economic growth; indeed, there is some empirical evidence in support of this view (Alesina and Rodrik, 1994). In other words, a distributional argument may, up to a point, be turned into an efficiency justification for income insurance and redistribution of lifetime income.

Of course, neither historical background factors nor theoretical justifications (rationales) by themselves can *explain* the actual emergence and expansion of welfare-state arrangements. References to the political processes are required. In countries where policies are based on electoral processes, the distribution of voting power across socio-economic groups is a natural starting point. It is also tempting to explain politically generated redistribution across generations by the distribution of voting power across cohorts. For instance, current generations may transfer resources to themselves at the expense of future generations, which (by definition) do not have voting rights, although they may later renege on political favours acquired by earlier generations. At the same time, young adults with children would be expected to push politically for education (and infrastructure investment), while older cohorts are

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
particularly likely to push for paygo pension systems and old-age care. The political outcome of such diverse interests, then, would be expected to depend on the relative power of different cohorts.
Indeed, some authors have tried to explain the emergence of modern social spending in western countries from the mid-19th century to the early 20th century by the gradual widening of franchise (Flora and Alber, 1981; Lindert, 2004). This is no doubt a realistic hypothesis. There are, however, obvious limitations to policy explanations in terms of relative voting powers of different interest groups. In the late 19th century (and even earlier), some welfare-state arrangements actually emerged in favour of individuals without voting rights; important examples include poor relief, mandatory and subsidized (even free) primary education, work-injury insurance and modest pensions. It is, therefore, tempting to assume that altruism and enlightened self-interest also help explain early welfare-state reforms – another example of how a justification may be turned into an explanation of actual development.
Moreover, the main expansion of welfare-state spending did not take place until half a century after the emergence of general franchise – indeed, mainly during the first three decades after the Second World War. One explanation for this apparent time lag may be that urbanization and industrialization were gradual processes, so that the political demand for new social arrangements likewise emerged gradually. It may also have taken considerable time to mobilize new groups of eligible voters. The time lags, and related gradualism in the expansion of welfare-state spending, could perhaps also be regarded as the result of an ‘experimental approach’ on the part of politicians or voters, due to uncertainty about the effects in various dimensions of higher welfare-state spending and related tax increases.

Achievements

Not only the level but also the composition of welfare-state spending, such as between transfers and human services, differs across countries. For instance, while about half of total public social spending consists of transfers in Western Europe as a whole (varying from 33 per cent in Iceland to 60 per cent in Austria), the corresponding figure is about 42 per cent in Anglo-Saxon countries outside Europe.

Transfers

What, then, is the relation between the size of aggregate government transfers on the one hand, and the degree of income security and government-induced redistribution of income across households on the other? To answer the first aspect of the question, it is important to consider the extent to which government-provided arrangements are a substitute for private income insurance. To answer the second aspect of the question, we would ideally also need to determine the extent to which government transfers have resulted in induced (endogenous) changes in the distribution of factor income (general equilibrium effects). Unfortunately, our knowledge on both issues is quite limited.
Scattered evidence suggests, however, that voluntary private income insurance and social insurance are rather close substitutes at the margin. In particular, government-provided benefits tend to be topped up by occupational pensions in countries with only modest public benefits (Pearson and Martin, 2005, pp. 8–10). As a result, *total* yearly per capita disposable income of retirees does not differ much across the eight west European countries studied by Forssell, Medelberg and Ståhlberg (2000), in spite of considerable differences in the replacement rates in government-operated pension systems. It is also

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

noticeable that total (public plus private) pensions are at least as large as a share of GDP in the United States as in western Europe (indeed, they are somewhat larger in the United States) in spite of the fact that the GDP share of public pensions is higher in western Europe, and that the population is younger in the United States (Table 1). Another example is that *total* per capita sick-pay benefits do not vary much among six west European countries studied by Kangas and Palme (1993), in spite of quite different replacement rates in government-operated systems – although the substitution is not complete.

Composition of total public social expenditures in 2001 (% GDP)

	United States			Western Europe*		
	Public	Private	Total	Public	Private	Total
Cash transfers	7.9	4.3	12.2	14.2	1.8	16.0
•*Pensions*	*6.1*	*3.8*	*9.9*	*8.5*	*1.0*	*9.5*
Human services	11.9	7.2	19.1	15.1	0.9	16.0
•*Health*	*6.2*	*5.0*	*11.1*	*6.4*	*0.4*	*6.8*
•*Education*	*5.1*	*2.3*	*7.3*	*5.4*	*0.4*	*5.8*
•*Active labour market programmes*	*0.1*		*0.1*	*0.9*		*0.9*
Total social expenditure	19.8	11.6	31.3	29.3	2.7	32.0

Notes: *Unweighted averages have been calculated for Austria, Belgium, Denmark, Finland, France, Germany, Iceland, Ireland, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom.

Figures for private health spending only cover private insurance programmes and exclude individual private health costs.

Sources: Adema and Ladaique (2005); OECD (2004).

There seems to be less substitution between public and private provision in countries where there is no government-operated system at all. The relatively low coverage of sick-pay insurance, sick-care insurance, and paid parental leave (‘parent insurance’) in the United States is a suggestive illustration. Thus, in areas where there is no government-operated system at all, it seems that the earlier discussed obstacles to the emergence of voluntary insurance arrangements ‘kick in’.

Since the distribution of disposable income is considerably more even than the distribution of factor income, it is natural to argue that welfare-state arrangements, and their financing, actually contribute to reducing the unevenness of the distribution of income. Moreover, based on data from the Luxembourg Income Study (LIS), Korpi and Palme (1998) found that the relative difference between the market-income Gini coefficient and the disposable-income Gini coefficient tends to be larger in countries with universal transfer systems than in those with a strongly targeted system. (Market income is then defined as factor income plus occupational pensions.) In this sense, universal systems tend in fact to be more redistributive than targeted systems. However, this conclusion does not hold concerning the redistribution *per unit* of aggregate public social spending; rather, the reverse tends to be the case (although the difference is not statistically significant).

The observation that welfare-state arrangements, in fact, have reduced the dispersion of the distribution of yearly income relies, of course, on the implicit assumption that an induced widening of the distribution of factor income has not offset the direct impact on the distribution of disposable income.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
One indicator that such adjustments have not taken place is that the distribution of yearly factor income did not become more uneven – at least not much – during the period when the generosity of public benefit systems increased the most, that is, from the late 1940s to the mid-1970s. Moreover, the subsequent widening of the distribution of yearly factor income in a number of countries (until about the mid-1990s) has been particularly pronounced in the United States and the United Kingdom, that is, in countries where welfare-state spending has increased *less* than in other countries. Thus, it seems reasonable to assume that government transfer systems (including social insurance) have, in fact, reduced the dispersion of the distribution of yearly disposable income.

Human services

In most developed countries, government intervention in the area of human services mainly takes the form of direct provision rather than general subsidies of such services. The effects of these policies would, however, be expected to differ systematically between low- and high-income citizens. One reason is that the per capita volume (or quality) provided by the government is often larger than what low-income individuals would have chosen themselves. Since human services cannot be resold in the market, the consumption of such services would be expected to increase among low-income groups. By contrast, it would be expected to fall among high-income groups, on the realistic assumption that human services, in contrast to income-insurance cash benefits, are difficult to supplement. (For instance, as a rule, parents do not divide up their children's attendance between a public and a private childcare centre or school.) Such a fall in consumption of human services among high-income groups would also be expected to take place among individuals who abstain from the public services offered and instead buy their services in the market. The reason, of course, is that their disposable income is reduced by the taxes they have to pay to finance the provision of human services to other citizens (basically reflecting an income effect).

There is a corollary to this reasoning: unless the volume provided is quite large, it is probably easier for the government to control the distribution than the aggregate volume of human services by direct provision. Total per capita consumption would therefore be expected to differ less across countries than the volume of government-provided services. Indeed, in spite of the fact that public-sector provision of human services is a larger share of GDP in western Europe than in the United States, 15.1 per cent as against 11.9 per cent, total (public plus privately provided) consumption of such services is larger in the United States than in Europe, 19.1 per cent as against 16.0 per cent (Table 1). In fact, this is the case for both education and health care – possibly partly reflecting a high income elasticity of demand for such services (with an ‘automatic’ supply response when such services are provided by markets).

It is probably easier to boost the aggregate consumption of human services by subsidies than by direct government provision – although the opposite is often asserted to be the case. (The government can be rather confident that general subsidies do increase the aggregate consumption of such services, in contrast to the case of direct government provision.) It is also *cheaper* for the government to boost such consumption by a certain volume by way of a subsidy than by way of direct provision. (While in the case of government provision the government has to finance the entire spending on such consumption, it has to finance only a fraction of total spending in the case of subsidies.)

There are other important differences between subsidies and direct government provision of human services. A subsidy allows the price to clear the market (zero excess demand), which implies that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
individuals will be able to choose volume and quality themselves, based on each individual's preferences and budget constraint. When judging the usefulness of allowing freedom of choice in the consumption of human services, it is, however, important to consider a number of other aspects as well, such as the efficiency of production and the quality of the services, the distribution of the services among households, and possible tendencies towards clustering (‘segregation’) of specific types of consumers (in terms of income, education, ethnicity, ideology and so on) on specific providers.

The age-specific nature of public social spending, of course, results in redistribution of resources (income as well as human services) over each individual's life cycle (intra-individual redistribution). Usually, resources are transferred to individuals below age 20–25 and above age 60–65, and extracted (via taxes) from individuals in the age groups in-between. Indeed, we may regard public financing of education as a (collectively decided) loan from the middle-aged to the young, and public financing of pensions as a subsequent payback of the loan via payroll taxes (Becker and Murphy, 1988). By these arrangements, two problems of intergenerational contracting are solved simultaneously: a liquidity constraint is removed for investment in human capital, and a universal pension system is created. Indeed, in countries with highly universal welfare-state arrangements, the *bulk* of social spending constitutes such intra-individual redistribution rather than inter-individual redistribution of lifetime income (‘wealth’), in contrast to countries with strongly targeted systems. For instance, the universal character of public social spending in Sweden and Italy helps explain the high shares of aggregate social spending that constitute intra-individual redistribution over the individual's life cycle in these countries (83 per cent and 76 per cent, respectively, according to Finance Department, Sweden, 2003, and O'Donoghue, 2001). The figure is, however, also boosted in countries, such as Sweden, where benefits usually are taxed. By contrast, the strongly targeted character of the social system in Australia helps explain its rather modest fraction of public social spending that consists of such intra-individual redistribution (38–52 per cent according to Falkingham and Harding, 1996). As pointed out above, in countries with large intra-individual redistribution over each individual's life cycle, the remaining part of public social spending (and its financing) is often sufficient, however, to generate considerable inter-individual redistribution of yearly income.

So far we know very little about the consequences of welfare-state arrangements for the distribution of *lifetime* disposable income. However, some simulations based on Swedish data indicate that lifetime income (‘wealth’) is to a considerable extent redistributed from the upper part of the distribution of lifetime income (the highest two quintiles) to the lower part (the lowest three quintiles) – if we abstract from conceivable general equilibrium effects (Finance Department, Sweden, 2003).

Problems

It is useful to classify major problems of contemporary welfare-state arrangements into (*a*) basically exogenous disturbances and (*b*) basically endogenous developments caused by the welfare state itself.

Exogenous factors

It is a commonplace that recent and predicted future changes in *demography* in developed countries, in particular the ‘graying’ of the population, simultaneously boost social spending and have a negative

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
influence on the tax base – since there are seldom automatic adjustments of social security contributions and benefit rules in response to changes in demography. Indeed, in the EU-19 the number of individuals above the statutory retirement age is already close to 25 per cent of the number of individuals of working age – and is projected to rise to about twice that figure or even more within three or four decades. It is difficult to alleviate this problem in the medium term except via immigration and tougher social-insurance legislation in the form of higher contribution rates, reduced benefits, stricter controls, and a higher effective retirement age.

The slowdown in the rate of *productivity growth* in the market sector in developed countries after the mid-1970s has created more or less the same financing problems, since neither the contribution rates nor the benefit rules in the social insurance systems are automatically (fully) adjusted to changes in productivity growth. So far, politicians have usually tried to deal with this problem in the same way as they have tried to adapt to demographic changes, that is, by ad hoc reductions in benefits and increases in social-insurance contributions. In recent years, the *internationalization* (globalization) of national economies has become perhaps the most hotly debated exogenous factor behind actual and predicted future welfare-state problems in developed countries. International trade theory predicts that the entry into the world economy of a number of countries with abundant low-wage labour (including China, India, the former Soviet republics and countries in eastern Europe) will reduce both the wage-income share of national income and the relative wages of low-skilled workers. Clearly, these consequences are bound to create problems for policy ambitions concerning the distribution of income in many developed countries. It is often also argued that the rate of structural change is likely to accelerate, thereby resulting in tendencies toward higher structural unemployment, due primarily to limited flexibility of the allocative mechanisms in national economies. With given social legislation, this would certainly boost transfer payments (including unemployment benefits) and give rise to an erosion of the tax base, thus threatening the financial sustainability of the welfare state.

If such problems were actually to arise, the standard policy advice is, of course, measures to promote the flexibility of domestic product and factor markets, for instance, along the lines of the so-called Lisbon Agreement among EU countries in 2000. Important examples are retraining of workers, easier entry and expansion of firms, less strict job-security legislation, and more flexible relative wage rates – possibly combined with employment subsidies for low-skilled workers (the ‘working poor’).

Another common worry in connection with the internationalization process is that important tax bases tend to become more internationally mobile. While, so far, this has occurred mainly for capital income, there is a possibility that similar (although less pronounced) consequences will emerge for other tax bases as well, possibly resulting in increased tendencies towards tax competition among governments. To the extent that such developments actually occur, increased international tax coordination (‘harmonization’) is perhaps the most frequently recommended, and predicted, policy response.

Moreover, increased migration to developed countries may place an additional strain on the financial position of various welfare-state arrangements, in spite of the fact that such immigration is likely to ‘improve’ the age structure of the population, since migrants may face difficulties in obtaining employment. Poorly functioning labour markets, partly as a result of regulated wages, would be an explanation. To the extent that governments are unable to alleviate these deficiencies, politicians will most likely remain under political pressure to stiffen the restrictions on immigration.

Internationalization is, however, not the main reason for the serious *unemployment problems* in Western Europe in recent decades, boosting welfare-state spending and damaging the tax base. Regardless of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
whether the background is a higher equilibrium unemployment rate or increased unemployment persistence (after unemployment-creating macroeconomic shocks or increased microeconomic turmoil), approximately the same types of structural reform are potentially useful. If the problems are caused by persistence mechanisms, there is, however, also a strong case for liberalizing job-security legislation, and adopting other policy measures that reduce the market power of labour-market insiders – both phenomena contributing to inertia of the employment level. Counter-cyclical demand management policies (monetary and fiscal policy) are also more useful if it is unemployment persistence (after unemployment-creating macroeconomic shocks), rather than higher equilibrium unemployment, that is the problem.

Baumol's ‘cost disease’ (Baumol, 1967) regarding labour-intensive human services – such as childcare, education and old-age care – is another largely exogenous threat to the financial viability of today's welfare-state arrangements. More specifically, since the relative costs of such services tend to increase over time (owing to slow productivity growth for such services), it will be necessary to raise tax rates gradually (without apparent limits) in countries where these services are tax-financed, even if the provision of such services is allowed to increase only rather slowly. The problem is somewhat different in the case of health care. After all, productivity in the health-care sector tends to rise rather rapidly along with advances in medicine and surgical techniques. However, since these improvements partly take the form of increased possibilities to treat health problems that could not be treated before, it is unavoidable that the demand for health care will also be boosted (at given incomes and prices). As a result, health care will, in fact, be exposed to similar financing problems as other human services, although partly for different reasons.

As a result of Baumol's cost disease, countries that today rely mainly on tax financing of human services will sooner or later have to limit the rate of expansion of such services (to the same rates as the increase in labour productivity of such services) or they will have to introduce complementary methods to finance human services, such as user fees and (voluntary or mandatory) insurance. Indeed, countries that are unwilling to accept such complementary financing methods may very well find themselves unable to finance equally large volumes of human services as countries with other financing methods. Perhaps these considerations help explain why both education and health spending, as mentioned above, are higher in the United States than in western European countries (although relatively high wages in the health sector in the United States is another explanation).

Endogenous factors

In contrast to the welfare-state problems discussed above, disincentive effects via tax distortions and moral hazard are (by definition) the result of endogenous adjustments of individuals to the welfare-state itself. In the case of income insurance, moral hazard (*ex post*) arises simply because the individual will be able to choose more leisure at a very low cost to himself in terms of lost income. It is also well known that health-care insurance induces some patients to ask for excessive medical tests and expensive treatment, demands that many physicians may be willing to satisfy.

Formally, the individual will (tautologically) choose work rather than benefits only if $u[w(1-t)] > u(bw) + a - f(n)$, where u is consumption utility, w the wage rate, t the average tax rate, b the benefit (replacement) rate, and a the *difference* between the utility of leisure and the intrinsic utility that one may derive from work. $f(n)$ denotes the disutility of stigmatization when breaking the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
prevailing work norm, where *n* is the aggregate number of individuals (or peers) who actually obey the work norm (or a norm against living on government benefits). I assume that the disutility of being stigmatized increases with the number of individuals who work rather than live on benefits; hence, $f'(n) > 0$. If we abstract, for the time being, from the social norms expressed by the stigmatization term *f*(*n*), the individual may prefer to live on benefits rather than on work already when the after-tax rate $(1 - t)$ is only modestly higher than the benefit rate (*b*), provided he evaluates leisure at least somewhat more than work (so that *a* is at least somewhat positive).

Of course, sufficiently strong social norms in favour of work (or against living on benefits), that is, a sufficiently high value of the $f(\cdot)$ function may prevent widespread and frequent reliance on benefits even if the difference in income when one works and when one lives on benefits is quite small. After a while, however, some ‘entrepreneurial’ individuals may be tempted to exploit the benefit systems. As a result, social norms in favour of work (against exploiting the benefit systems) may erode among others as well (Lindbeck, Nyberg and Weibull, 1999; Lindbeck and Persson, 2006). The long-term negative effects of more generous welfare-state arrangements on aggregate labour supply may then be stronger than suggested by traditional microeconomic studies of the elasticity of labour supply with respect to after-tax wage rates. (Empirical research on the role of social norms in favour of work or against living on benefits is, however, still at an early stage.)

As an illustration of the potential importance of moral hazard for per capita hours of work, we may note that about a fifth of the population of working age (15–64) in western Europe today (2006) lives on various cash transfers from the government – the most important examples include unemployment benefits, labour-market programmes, social assistance, sick-pay insurance, and early retirement pensions (OECD, 2003, pp. 188–90). Such moral hazard effects of generous welfare-state arrangements in western Europe are, therefore, an important explanation for the limited per capita hours of work in that part of the world. As a comparison, per capita hours of work (per year) in the United States are between 30 per cent and 50 per cent higher than in western Europe. (Prescott, 2004, has instead tried to explain this phenomenon by the higher *marginal* tax rates in western Europe, assuming quite high labour-supply elasticities with respect to after-tax wage rates.)

The character and size of the incentive effects of welfare-state arrangements depend, of course, on the specific rules of both the benefit arrangements and the financing of these. For instance, to the extent that tax-financed benefits are paid to retired individuals rather than to individuals of working age, the negative substitution effects of the tax wedges on labour supply are counteracted by positive income effects of the tax payment (since, in this case, the taxpayers of working age do not get anything back in exchange for the tax payments). It is also well known that the negative substitution effects of marginal tax wedges on the labour supply are mitigated if there is a (positive) link between the individual's contributions to various social-insurance systems and his expected future benefits – as in the case of actuarially fair or ‘quasi-actuarial’ social-insurance arrangements – provided the individual is aware of this link. It is also a commonplace that negative incentives to acquire education as a result of marginal (in particular, progressive) tax rates are often counteracted, or perhaps even overcompensated, by subsidies to investment in human capital. Moreover, in some countries tax revenues are used to finance services that are close substitutes for home production, and hence complements to work in the open labour market. Subsidies to childcare and old-age care outside the family are important examples. In this special case, the negative substitution effects of tax wedges on the labour supply would be counteracted

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
by positive cross-substitution effects on labour supply of the subsidized (or directly provided) services.
From an empirical point of view, the consequences of welfare-state spending on the efficiency and growth of the national economy are, of course, a perennial issue. In the case of countries with modest levels of such spending, economists generally agree that the positive effects of higher welfare-state spending on economic efficiency and economic growth are likely to dominate over the negative effects. This is particularly likely if increased public spending, starting from low levels, is concentrated on features such as sanitation, basic health care, elementary education and infrastructure, and if more comprehensive and generous income protection would further mitigate tendencies towards social unrest. However, there is also general agreement that, sooner or later, ever-increasing social spending will render the net effects on economic efficiency and growth negative, although it is difficult to identify the turning point.
The complexities of analysing and aggregating the effects of various types of benefit arrangements, and related taxes, have prompted many economists to try to find short cuts, by simply regressing either the level or (more often) the aggregate growth rate of per capita GDP on broad aggregates of taxes or government spending programmes. It is a fair summary of this huge literature that there is stronger support for the hypothesis that the effects of higher spending and taxes in today's developed countries are negative rather than positive. (Basically, studies conducted since around 1990 conclude that the effects are either negative or absent.) However, such aggregate studies suffer from well-known methodological problems.

New requirements

The modern welfare state is a success in the sense that it has contributed to solving a number of potentially serious social problems. It encounters, however, financial difficulties in several countries. Some welfare-state arrangements, and their financing, have also created *new* problems, including benefit dependency and other incentive effects. These developments are, of course, the background for ongoing and planned reforms of, and retreats from, existing welfare-state arrangements in a number of countries. At the same time, strong political demands have emerged for new or improved social arrangements in several areas. For instance, increased female labour-force participation has raised the demand for paid parental leave, subsidized childcare, and old-age care outside the family – basically to facilitate everyday life among families with two income earners. In some countries, such arrangements are also regarded as important methods for restoring rapidly falling birth rates. The reduced stability of the family has also generated a political demand for legislated property rights in spouses' social-insurance benefits, in particular pensions.
There is also evidence of increasing individualization of values and lifestyle in developed countries, as compared with a number of decades ago, when today's welfare-state arrangements were designed (for evidence of such value changes, see Inglehart et al., 2004). Obvious ways of adjusting various benefit systems to these new values are more individually differentiated and portable social entitlements (nationally as well as internationally), as well as increased freedom for the individual to choose type of (mandatory) income insurance and quality of various types of (subsidized) human services, for instance via voucher systems (in a wide sense of the term).
Moreover, the incidence of economic and social misery among specific minority groups has recently increased in several developed countries – partly as a result of rising long-term unemployment,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
immigration of low-skilled groups from poor countries, alcoholism, drug abuse and de-institutionalization of the mentally ill – the 'truly disadvantaged' individuals. These problems require more than a generally improved situation in the labour market; new types of *targeting* social policies are necessary to help specific minority groups. A generally accepted view among social workers seems to be that it is also important to integrate more closely the administration of social insurance, social assistance, labour-market exchange systems, health care, rehabilitation, labour-market training and so on. Moreover, in some cases non-governmental organizations, including non-profit organizations, seem to be more successful than governmental organizations in such endeavours. These observations raise the issue of the potential usefulness of new divisions of tasks among governments, markets, the family, and civil society.

See Also

- health insurance, economics of
- labour supply
- social insurance
- social norms
- taxation of the family

Bibliography

Adema, W. and Ladaique, M. 2005. Net social expenditure, 2005 edition: more comprehensive measures of social support. Social, Employment and Migration Working Papers No. 29. Paris: Directorate for Employment, Labour and Social Affairs, OECD.

Alesina, A. and Rodrik, D. 1994. Distributive politics and economic growth. *Quarterly Journal of Economics* 109, 465–90.

Atkinson, A. 1991. Social insurance: the Fifteenth Annual Lecture of the Geneva Association. *Geneva Papers on Risk and Insurance Theory* 16, 113–31.

Barr, N. 1998. *The Economics of the Welfare State*, 3rd edn. Oxford: Oxford University Press.

Baumol, W. 1967. Macroeconomics of unbalanced growth: the anatomy of urban crisis. *American Economic Review* 57, 415–26.

Becker, G. and Murphy, K. 1988. The family and the state. *Journal of Law and Economics* 31, 1–18.

Esping-Andersen, G. 1990. *The Three Worlds of Welfare Capitalism*. Cambridge: Polity Press.

Falkingham, J. and Harding, A. 1996. Poverty alleviation versus social insurance systems: a comparison

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of lifetime redistribution. Discussion Paper No. 12. National Centre for Social and Economic Modelling, Faculty of Management, University of Canberra.

Finance Department, Sweden. 2003. Fördelningen ur ett livscykelperspektiv [Distribution from a life-cycle perspective]. *Långtidsutredningen*, bilaga 9. Stockholm: Finansdepartementet.

Flora, P. and Alber, J. 1981. Modernization, democratization, and the development of welfare states in Western Europe. In *The Development of Welfare States in Europe and America*, ed. P. Flora and A. Hedenheimer. New Brunswick, NJ: Transaction Press.

Forssell, Å., Medelberg, M. and Ståhlberg, A.-C. 2000. Olika transfereringssystem men lika inkomster. De äldres ekonomiska situation i ett internationellt perspektiv [Different transfer systems but equal income: the economic status of the elderly in an international perspective]. *Ekonomisk Debatt* 28, 143–58.

Inglehart, R., Basanez, M., Diez-Medrano, J., Halman, L. and Luijkx, R., eds. 2004. *Human Beliefs and Values: A Cross-cultural Sourcebook Based on the 1999–2002 Values Surveys*. Mexico City: Siglo XXI Ed.

Kangas, O. and Palme, J. 1993. Statism eroded? Labor-market benefits and challenges to the Scandinavian welfare states. In *Welfare Trends in the Scandinavian Countries*, ed. E. Hansen et al. New York: M.E. Sharpe.

Korpi, W. and Palme, J. 1998. The paradox of redistribution and strategies of equality: welfare state institutions, inequality, and poverty in the Western countries. *American Sociological Review* 63, 661–87.

Lindbeck, A., Nyberg, S. and Weibull, J. 1999. Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics* 114, 1–35.

Lindbeck, A. and Persson, M. 2006. A model of income insurance and social norms. Working Paper No. 742. Stockholm: Institute for International Economic Studies.

Lindert, P. 2004. *Growing Public: Social Spending and Economic Growth Since the Eighteenth Century*. Cambridge, MA: Cambridge University Press.

O'Donoghue, C. 2001. Redistribution in the Irish tax-benefit system. Ph.D. thesis. London School of Economics.

OECD (Organisation for Economic Co-operation and Development). 2003. *OECD Employment Outlook*. Paris: OECD.

OECD. 2004. *Education at a Glance*. Paris: OECD.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

OECD. 2005. *Society at a Glance: OECD Social Indicators*. Paris: OECD.

Pearson, M. and Martin, J. 2005. Should we extend the role of private social expenditure? Discussion Paper No. 1544. Bonn: IZA.

Prescott, E. 2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28, 2–13.

How to cite this article

Lindbeck, Assar. "welfare state." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000053> doi:10.1057/9780230226203.1827

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wells, David Ames (1828 – 1898)

Warren J. Samuels
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

free trade; individualism; laissez-faire; Wells, D. A.

Article

Wells was born on 17 June 1828 in Springfield, Massachusetts, and died on 5 November 1898 in Norwich, Connecticut. Trained at Williams College and Lawrence Scientific School at Harvard, Wells first taught and published as a geologist and chemist. After newspaper work, Wells turned to economics in his mid-forties. After publishing on the national debt, he was appointed to a series of federal and state tax positions, where he issued influential reports, revised tax laws, and originated the stamp system for collecting taxes on tobacco and liquor. He lectured in economics at Yale, Harvard and elsewhere, succeeded John Stuart Mill in 1874 as foreign associate of the French Academy, was president of the American Social Science Association, and received honorary degrees from Oxford, Harvard and Williams. His economic interests were practical and empirical, rather than theoretical; his place was transitional between the popular writer and the technically trained professional investigator.
Politically active, he was a leading exponent of laissez-faire doctrine, which he equated with individualism (in the manner of William Graham Sumner, with whom he was associated), free trade and the gold standard. Although an early protectionist disciple of Henry C. Carey, he later actively wrote and campaigned in favour of free trade. He opposed fiat money, the greenbacks and free silver. At one point he proposed the conversion of the greenbacks to interest-bearing government bonds; at another, he advocated a ‘cremation theory of specie resumption’, with the Secretary of the Treasury to burn a volume of greenbacks each day until they attained parity with gold.
Considered by some to be so doctrinaire as to be impervious to the stresses brought by industrialization in the late 19th century, he was none the less concerned with economic instability. Here he departed from orthodox doctrine, emphasizing the existence of unemployment due to both technology and overproduction relative to present demand, aggravated by the decline of available public lands as an alternative open to labour. His remedy was freer trade.
Through his will, he established the David Ames Wells Prizes in economics at Harvard University.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Recipients have included John H. Williams, Jacob Viner, James W. Angell, Seymour E. Harris, Edward Chamberlin, Robert Triffin, Paul Samuelson, Lloyd Metzler, Robert Solow, Peter Kenen, Lester Thurow and Donald McCloskey.

Selected works

1864. *Our Burden and Our Strength*. New York: Loyal Publication Society.

1875. *The Cremation Theory of Specie Resumption*. New York: Putnam's.

1876. *Robinson Crusoe's Money*. New York: P. Smith.

1877. *The Silver Question*. New York: Putnam's.

1884a. *A Primer of Tariff Reform*. New York: New York State Revenue Reform League.

1884b. (With W.G. Sumner.) *Political Economy and Political Science*. New York: Society for Political Education.

1885. *Practical Economics*. New York: Putnam's.

1888. *Relation of The Tariff to Wages*. New York: Putnam's.

1889. *Recent Economic Changes*. New York: D. Appleton.

1900. *The Theory and Practice of Taxation*. New York: D. Appleton.

Bibliography

Joyner, F.B. 1939. *David Ames Wells: Champion of Free Trade*. Cedar Rapids, IA: Torch Press.

How to cite this article

Samuels, Warren J. "Wells, David Ames (1828–1898)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000054>
doi:10.1057/9780230226203.1828

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

West, Edward (1782 – 1828)

Mark Blaug
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

differential rent; diminishing returns; free trade; Malthus, T. R.; market price; natural price; Ricardo, D.; wages fund; West, E.

Article

Edward West is remembered – if he is remembered at all – for having stated the theory of differential rent based on the principle of diminishing returns in a long pamphlet just before Ricardo did so, and in virtually the same form and language. This has earned West the title of ‘the first, though not the name-father and greatest of the “Ricardian” school’ (Cannan, 1893, p. 219). However, it appears that Ricardo developed the principle of diminishing returns independently of West and even of Malthus (who also published the idea more or less simultaneously) and at any rate Ricardo's exposition in his *Essay on Profits* (1815) was clearer then anyone else's, was more carefully set out and went beyond West in spelling out its implications for the distribution of income between wages, profits and rent. In addition to his *Essay on the Application of Capital to Land, with Observations Shewing the Impolicy of any Great Restriction of the Importation of Corn* (1815), West only wrote one other work on economics, a short book entitled *Price of Corn and Wages of Labour, with Observations upon Dr. Smith's, Mr. Ricardo's, and Mr. Malthus's Doctrines Upon those Subjects* (1826). At the time of his death, he was working on a treatise in political economy, the manuscript of which has been lost.

West was born in 1782 near London, educated at Harrow and University College, Oxford (where he studied classics and mathematics), and then went on to study law. In 1817, two years after the *Essay on the Application of Capital to Land*, he published a major treatise on the law of ‘extents’ (indemnities against direct or indirect debts to royalty), which was instrumental in reforming the use of extents in the Court of Chancery. In 1822 he was knighted and appointed Recorder of Bombay, followed two years later by the post of Chief Justice of the Crown in the Bombay province of India. The publication of his book on the *Price of Corn* in 1826 showed that he maintained his interests in economics until his death in India in 1828.

The similarity between the ways in which both West and Ricardo expressed the principle of diminishing

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

returns in agriculture – in terms of diminishing average rather than marginal products of composite doses of capital-and-labour applied to a fixed quantity of heterogeneous land and inclusive, not exclusive, of technical progress in agriculture – is startling, and so is the fact that both of them employed it to deduce a falling rate of profit on capital that could be postponed, but not permanently reversed, by the abolition of the Corn Laws. The only striking difference between the two 1815 pamphlets lay in the implications the two authors deduced from diminishing returns: Ricardo inferred that rents per acre would rise as more capital and labour were applied to ever inferior land, while West inferred that rents would fall, so that free trade would benefit landlords as well as capitalists and workers. This was a point on which West later changed his mind: in the *Price of Corn*, he agreed with Ricardo's inference about both rents per acre and the rental share. Unlike Ricardo, West realized that free trade would not imply complete specialization as between manufacturing in Britain and agriculture in Britain's trading partners: diminishing returns would operate abroad to raise the price of exported corn even as free trade would diminish the pressure on the costs of raising corn at home, so that eventually ‘the actual price of both in the market must meet’. In this way, he met what was at the time a critical objection to the notion of free trade, namely that it would make Britain for ever dependent on foreign food supplies.

West's *Price of Corn* is a notable book if only because it was virtually the first work to attack the wages fund doctrine embedded in the writings of Adam Smith and Ricardo. ‘The opinion that the demand for labour is regulated solely by the amount of capital’, West asserted, ‘has led perhaps to more false conclusions in the science than any other cause’. The demand for labour, he insisted, is not governed by the stock of wage goods inherited from the past but by the total level of private and public investment and consumption spending in the economy. It followed, he concluded, that ‘the demand for the money wages of labour may be increased without any increase of the capital of the country’. The book contained a number of other insights, although opinions must differ as to how original these really were. There was the idea that price is determined by demand and supply, each of them considered as schedules of quantities at various hypothetical prices (an idea also found in Malthus); that the long-run ‘natural’ price of commodities is equal to average costs, including normal profits; that all manufacturing firms have identical cost functions; that the short-run market price of industrial goods cannot fall below average variable costs; and that the effect of a change in agricultural output on the price of corn depends on both the price and income elasticity of demand for corn. For some commentators these insights make him a ‘Marshallian before Alfred Marshall’ rather than a Ricardian before Ricardo (Grampp, 1970).

Bibliography

Cannan, E. 1893. *A History of the Theories of Production and Distribution in English Political Economy from 1776 to 1848*. London: Staples Press, 1953.

Grampp, W.D. 1970. Edward West reconsidered. *History of Political Economy* 2, 316–43.

How to cite this article

Blaug, Mark. "West, Edward (1782–1828)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000055>
doi:10.1057/9780230226203.1829

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wheatley, John (1772 – 1830)

Frank Whitson Fetter
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Bank Charter Act (1844); note issue; Wheatley, J.

Article

Lawyer and economist, Wheatley was born in Erith, Kent, of a prominent landed and military family, and died at sea on a voyage from South Africa to England. A memorial plaque to him is in the Wheatley chapel of the Erith parish church. At Oxford he was a member of Christ Church, and after receiving his BA in 1793 was admitted to Lincoln's Inn, but his activity in the law was limited, and his life was devoted largely to writing on economics and playing a small part in Whig politics. With him at Christ Church was Charles Watkin Williams Wynn, nephew of Lord Grenville, and Wheatley was active in support of Grenville's successful campaign in 1809 for Chancellor of Oxford University; he had correspondence with Wynn in 1812 about running for Parliament on the Whig ticket, but nothing came of this; a book of 1816 took the form of a letter to Lord Grenville, and his pamphlet of 1823 was a letter to Wynn.

Wheatley published ten books and brochures, two of these in India and one in South Africa. He lived in these two countries for the last nine years of his life, evidently to escape creditors. His works published in India and South Africa received little contemporary attention, and today are extremely rare. Of the others, *Remarks on Currency and Commerce* (1803), and the first volume of *An Essay on the Theory of Money and Principles of Commerce* (1807) received the most contemporary attention, and best stated his theoretical position on the monetary controversy that followed the suspension of cash payments by the Bank of England in 1797. Wheatley stated, in an even more extreme way than Ricardo did later, that exchange fluctuations were due exclusively to domestic price changes, and that the Bank of England, through its credit policy, could control prices, and thus exchange rates. These books of 1803 and 1807 criticized the Bank for its monetary expansion, but following the resumption of specie payments in 1821 Wheatley in his book of 1822 had become a severe critic of both the Bank and of the Tory government for the price deflation. His efforts, both in his book and in correspondence with Whig leaders, to launch an attack on the government's monetary policy, made no headway. In several publications he stressed the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
danger, for monetary stability, of permitting the issue of notes by banks other than the Bank of England. No economist of his period so well anticipated the note issue provisions of the Bank Act of 1844, which led to the elimination of all notes other than of the Bank of England.

Wheatley's views on political issues were something of a paradox. As a Whig he was frequently a voice for reform; with the background of the landed gentry he sometimes disagreed with Whig positions that threatened the supremacy of the landed aristocracy. He supported free trade, a commercial union with France, removal of restrictions on West Indian trade, and the abolition of slavery. He favoured primogeniture, maintenance of great landed estates, the political supremacy of the landed gentry, and an unreformed Parliament. In foreign policy his imperialist views foreshadowed the idea of the ‘white man's burden’.

Selected works

1803. *Remarks on Currency and Commerce*. London.

1805. *Thoughts on the Object of a Foreign Subsidy*. London.

1807, 1822. *An Essay on the Theory of Money and Principles of Commerce*. Vol. 1, London, 1807; vol. 2, London, 1822.

1807. *A Letter to Lord Grenville, on the Distress of the Country*. London.

1819. *A Report on the Reports of the Bank Committee*. Shrewsbury.

1821. *A Plan to Relieve the Country from Its Difficulties*. Shrewsbury. This short pamphlet is an extract from the book of 1822 that appeared shortly afterwards.

1823. *Letter to the Rt. Hon. Charles Watkin Williams Wynn, President of the Board of Control, on the Latent Resources of India*. Calcutta.

1824. *A Letter to his Grace the Duke of Devonshire on the State of Ireland, and on the General Effects of Colonization*. Calcutta.

1828. *Tempora praeterita: Or, More Currency and More Corn*. Cape Town. This was published anonymously, but in correspondence Wheatley admitted authorship.

The Wheatley letters to Lord Grey are in the Grey of Howick papers at the University of Durham; the Wheatley letters to Charles Watkin Williams Wynn are at the National Library of Wales at Aberystwyth.

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Fetter, F.W. 1942. The life and writings of John Wheatley. *Journal of Political Economy* 50, 357–76.

Viner, J. 1937. *Studies in the Theory of International Trade*. New York: Harper & Brothers.

How to cite this article

Fetter, Frank Whitson. "Wheatley, John (1772–1830)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000058>
doi:10.1057/9780230226203.1830

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Whewell, William (1799 – 1866)

G. Campanelli
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

fixed capital; mathematical economics; mathematical method in political economy; production price; Whewell, W.

Article

Whewell was born in Lancaster and died in Cambridge. He received his early education at Lancaster Grammar School and Heversham School, Westmoreland, and in 1812 he went up to Trinity College, Cambridge. In 1817 he became a Fellow of the college, in 1823 a tutor. In 1841 he was made Master, an appointment which he held until his death.

Whewell was at the centre of a ‘network’ of Cambridge scientists and exercised considerable influence upon scientific and philosophic circles in Victorian England. In 1820 he became a Lecturer in Mathematics, in 1828 he was appointed Professor of Mineralogy and in 1838 Professor of Moral Philosophy. He was active as an honorary member of 25 scientific, historical and philosophical societies in several countries. To mention a few of the most important in England: he was one of the founders of the Cambridge Philosophical Society in 1818; in 1820 he was elected a Fellow of the Royal Society; in 1831 he became a member of the British Association and in 1841 was appointed President.

Whewell was primarily a philosopher and mathematician, and he published his major works in these fields (Whewell, 1837; 1840). Political economy was one of the many other subjects dealt with by him. However, his contributions in this field – written over the whole period from 1829 to 1862 – give clear proof that his interest in economics was lifelong. Whewell's major works in political economy were four papers on mathematical economics which were read before the Cambridge Philosophical Society (Whewell, 1829; 1831; 1850a; 1850b) and a book – *Six Lectures on Political Economy* (1862) – which was composed for the edification of the Prince of Wales, the future Edward VII. In the *Six Lectures* Whewell presented, in a very elementary way, the principal ideas of Smith, Ricardo and Jones.

Whewell's four papers represent the earliest *systematic* application of mathematical symbols of political economy in England. Whewell believed that the arithmetic used by classical economists was inadequate, and that the more general language of algebra should take its place. He pointed out that the adoption of a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
mathematical method had two main advantages. Firstly, that many aspects of political economy could be presented in a more simple, clear and systematic form. Secondly – and more importantly – the use of mathematics could help to avoid the danger of drawing false conclusions for assumptions made. To illustrate this point, in his 1829 paper Whewell used mathematics to discuss Ricardo's theory of the incidence of a tax on wages. Ricardo had argued – against Smith – that a rise in the prices of goods due to a rise in wages would in turn affect wages and ‘the action and reaction first of wages on goods and then of goods on wages, will be extended without any assignable limits’ (Ricardo, 1821, p. 225).

Whewell, on the contrary, showed that if Ricardo had considered the mathematical implications of his theory, he would have found that an unlimited rise in prices and wages was impossible. Indeed, if it is assumed that only a part of the value of goods is wages, and only a part of the labourer's consumption consists of manufactured goods, then the paths that both prices and wages follow take the form of geometric series which converge.

But Whewell's most notable contribution to political economy was his mathematical formulation of Ricardo's theory, and in particular his analysis of fixed capital (Whewell, 1831). This analysis is important not only because it represents the first mathematical treatment of machinery in Ricardo's model, but also and mainly because it constitutes an original contribution to the subject. In 1831 Whewell had already provided an exact formulation for the reduction of fixed capital to dated quantities of labour. He also worked out a simple model to quantify the substitution effect between labour and machinery. Finally, through the annuity formula, he arrived at the equation which defines the production price in the presence of fixed capital.

These results also suggest that the dating of the genesis of fixed capital models may need to be reappraised, for it is usually thought that Bortkiewicz (1907) – on the basis of Dmitriev's contribution (1904) – was the first economist to treat fixed capital mathematically within the theory of production price.

Whewell has been consistently neglected in the history of economic analysis. The few authors who were acquainted with Whewell's work – of whom the most authoritative were Jevons (1871) and Schumpeter (1954) – considered his analysis as purely derivative: supposedly he merely translated into algebraic form results which others had previously expressed in non-mathematical language. The only exception was Walras, who regarded Whewell's contribution as ‘really remarkable’ (Walras, 1875, p. 32).

Whewell was in fact more than a translator: he was a major contributor to the early development of mathematical economics in England, and above all a pioneer in the general debate on fixed capital.

Selected works

1829. *Mathematical Exposition of Some Doctrines of Political Economy*. Cambridge: Cambridge Philosophical Society, 1830, Transaction 3.

1831. *Mathematical Exposition of Some of the Leading Doctrines of Political Economy: Second Memoir*. Cambridge: Cambridge Philosophical Society, 1833, Transaction 4.

1837. *History of the Inductive Sciences, From the Earliest to the Present Time*. 3rd edn, New York: Appleton, 1890.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1840. *The Philosophy of the Inductive Sciences, Founded Upon their History*. 2nd edn, London: Parker, 1847.

1850a. *Mathematical Exposition of Some Doctrines of Political Economy: Second Memoir*. Cambridge: Cambridge Philosophical Society, 1856, Transaction 9, Part I.

1850b. *Mathematical Exposition of Certain Doctrines of Political Economy: Third Memoir*. Cambridge: Cambridge Philosophical Society, 1856, Transaction 9, Part II.

1862. *Six Lectures on Political Economy*. Cambridge: Cambridge University Press.

1971. *Mathematical Exposition of Certain Doctrines of Political Economy*. New York: A.M. Kelley. A reprint of Whewell's four papers on mathematical economics (1829, 1831, 1850a and 1850b).

Bibliography

Bortkiewicz, L. 1907. Wertrechnung und Preisrechnung im Marxschen System. *Archiv für Sozialwissenschaft und Sozialpolitik*, vol. 25. The second and the third parts are translated into English as ‘Value and Price in the Marxian System’, *International Economic Papers* No. 2 (1952), 5–60.

Dmitriev, V.K. 1904. *Economic Essays on Value, Competition and Utility*. Ed. D.M. Nuti, Cambridge: Cambridge University Press, 1974.

Jevons, W.S. 1871. *The Theory of Political Economy*. London and New York: Macmillan.

Ricardo, D. 1821. *On the Principles of Political Economy and Taxation*. Ed. P. Sraffa, Cambridge: Cambridge University Press, 1951.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Walras, L. 1875. Une branche nouvelle de la mathématique: de l'application des mathématiques à l’économie politique. Italian trans. as ‘Un nuovo ramo della matematica. Dell'applicazione della matematiche alla economia politica’, *Giornale degli Economisti* 3 (1876).

How to cite this article

Campanelli, G. "Whewell, William (1799–1866)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000060>

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

doi:10.1057/9780230226203.1831

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wicksell effects

Edwin burmeister
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

capital deepening; capital goods; capital theory; capital theory paradoxes; numeraire; production possibility frontier; Wicksell effects; Wicksell, J. G. K.

Article

In realistic economic models with *n* different types of capital goods, the value of the capital stock is

$$V = \sum_{i=1}^{n} P_i K_i \tag{1}$$

where P_i is the price of the *i*th capital good in terms of some *numéraire*. The value of capital, however, is not an appropriate measure of the 'aggregate capital stock' as a factor of production except under extremely restrictive conditions. Wicksell (1893; 1934) originally recognized this fact, which subsequently was emphasized by Robinson (1956).

If attention is restricted to alternative steady-state comparisons, in constant-returns-to-scale economies without joint production *V* is a function of the interest rate, *r*; see, for example, Burmeister and Dobell (1970). The *Wicksell effect* is the change in the value of the capital stock from one steady state to another, namely

$$\frac{dV}{dr}. \tag{2}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The term ‘Wicksell effect’ was introduced by Uhr (1951), but its importance was not widely recognized until the writings of Robinson (1956) and Swan (1956).

The Wicksell effect is the sum of the *price Wicksell effect* (which is the revaluation of the inventory of capital goods due to new prices) and the *real Wicksell effect* (which is the price-weighted sum of the changes in the physical quantities of different capital goods):

$$\frac{dV}{dr} = \sum_{i=1}^{n} \frac{dP_i}{dr} K_i + \sum_{i=1}^{n} P_i \frac{dK_i}{dr}. \tag{3}$$

Numerical examples show that the price Wicksell effect can be negative, that is,

$$\sum_{i=1}^{n} P_i \frac{dK_i}{dr} < 0 \tag{4}$$

is possible, even when (i) the total Wicksell effect is positive $[dV/dr > 0]$, or (ii) particular capital stocks are increasing with $dK_i/dr > 0$ for some but not all *i*; see Burmeister and Dobell (1970, pp. 289–93). In neoclassical models with only one capital good (*n*=1), the real Wicksell effect is always negative. Moreover, the sign of the price Wicksell effect depends upon the choice of *numéraire*, and hence so does the total Wicksell effect given by (3). The sign of the real Wicksell effect, however, is independent of the choice of *numéraire*.

One central issue of the Cambridge controversies in capital theory involves Wicksell effects. Does a decrease (increase) in the steady-state interest rate always imply a rise (fall) in per capita steady-state consumption provided the rate of interest is greater (less) than the rate of growth of labour? In one-capital good models, the answer to this question is, ‘Yes’. In general, the answer is, ‘Yes’, if and only if the real Wicksell effect is negative; see Burmeister and Turnovsky (1972) and Burmeister (1976).

To establish this relationship between the behaviour of per capita consumption and the real Wicksell effect, consider a technology which can be represented by a *production possibility frontier*

$$Y_1 = T(Y_2, \ldots, Y_n; L, K_1, \ldots, K_n) \tag{5}$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

where Y_i is the output of commodity i, L is the labour which grows at the exogenous rate g, and K_i is the stock of commodity i used as a capital input.
It is assumed further that $T(\cdot)$ is twice continuously differentiable, exhibits constant returns to scale, and has a Hessian matrix $[T_{ij}]$ that is negative semi-definite and whose rank varies with the degree of joint production in the economy; see Samuelson (1966), Burmeister and Turnovsky (1971), and Kuga (1973). The analysis which follows can be generalized to non-differentiable technologies as in Burmeister (1976), but for simplicity only differentiable technologies are considered here.
In steady-state equilibria all quantities grow at the rate g, implying that the output of every commodity must satisfy

$$Y_i \equiv C_i + K_i = C_i + gK_i, \quad i = 1, \ldots, n, \tag{6}$$

where C_i denotes the consumption of commodity i. Substituting these steady-state restrictions into (5) and using lower-case letters to denote per capita quantities, we have

$$c_1 + gk_1 = T(c_2 + gk_2, \ldots, c_n + gk_n, 1, k_1, \ldots, k_n). \tag{7}$$

Let the prices of commodities and the rental rates for capital goods, both in terms of the wage rate as *numéraire*, be denoted by p_i and w_i respectively, i=1,•…,•n; also let r denote the interest or profit rate. It is well-known that intertemporal profit maximization and/or efficiency necessitates that

$$\frac{\dot{p}_i}{p_i} + \frac{w_i}{p_i} = r, \quad i = 1, \ldots, n. \tag{8}$$

Imposing the steady-state requirement that relative prices remain constant, (8), implies that

$$w_i = rp_i, \quad i = 1, \ldots, n.$$

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

(9)

Using the well-known marginal conditions

$$\frac{\partial T}{\partial (c_i + gk_i)} = -\frac{p_i}{p_1} \text{ and } \frac{\partial T}{\partial k_i} = \frac{w_i}{p_1}, \quad i = 1, \ldots, n,$$

(10)

we see that a vector

$$(c^*, k^*; r^*, p^*) = (c_1^*, \ldots, c_n^*; r^*, p_1^*, \ldots, p_n^*) \geq 0$$

(11)

satisfying (7) and (10) represents a steady-state solution at the growth rate *g*. It thus follows immediately from differentiation of (7) that almost everywhere.

$$\sum_{i=1}^{n} p_i \left(\frac{\mathrm{d}c_i}{\mathrm{d}r}\right)\Bigg|_{(r^*, p^*)} = (r - g) \sum_{i=1}^{n} p_i \left(\frac{\mathrm{d}k_i}{\mathrm{d}r}\right)\Bigg|_{(r^*, p^*)};$$

(12)

see Burmeister (1976) for details.
Now let *v* denote the per capita value of capital in terms of the wage rate as *numéraire*:

$$v = \sum_{i=1}^{n} p_i k_i .$$

(13)

The change in the per capita value of capital across alternative steady-state equilibria is obtained by differentiating (13); thus almost everywhere the per capita Wicksell effect is

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$\left.\frac{dv}{dr}\right|_{(r^*,p^*)} = \sum_{i=1}^{n}\left(\frac{dp_i}{dr}\right)\Bigg|_{(r^*,p^*)} \cdot k_i + \sum_{i=1}^{n} p_i \cdot \left(\frac{dk_i}{dr}\right)\Bigg|_{(r^*,p^*)}$$
(14)

Comparing (14) and (12), it is seen that it is the real Wicksell effect which determines whether or not ‘consumption’ is well-behaved across steady-state equilibria. That is, if the real Wicksell effect is negative and

$$\sum_{i=1}^{n} p_i \cdot \left(\frac{dk_i}{dr}\right)\Bigg|_{(r^*,p^*)} < 0,$$
(15)

then almost everywhere

$$\sum_{i=1}^{n} p_i \cdot \left(\frac{dc_i}{dr}\right)\Bigg|_{(r^*,p^*)} \square\ 0 \text{ as } r \square\ g.$$
(16)

In particular, when $c_2 = c_3 = \ldots = c_n$ and only commodity 1 is consumed, consumption as measured by c_1 rises (falls) as r is increased from r^* to $r^* + \Delta r^*$ when r^* is greater (less) than g. (The familiar golden rule condition giving maximum per capita steady-state consumption holds at $r^* = g$.)
It follows, then, that a negative real Wicksell effect is the appropriate concept of ‘capital deepening’ in a model with many heterogeneous capital goods. That is, when (15) and hence (16) hold, an economy with a low interest rate (but exceeding g) has ‘more capital’ than one with a higher interest rate in the sense that it is capable of providing more steady-state per capita consumption. Although (15) and (16) always hold in a neighbourhood of $r^* = g$, examples show that they do not generally hold everywhere. This possibility – that (16) does not hold everywhere – is perhaps the most interesting conclusion to emerge from the Cambridge controversies and has been termed a ‘paradox’. However, the ‘paradox’ involves comparisons of alternative steady-states rather than comparisons of alternative feasible paths; Bliss (1975) provides a lucid explanation of why such ‘paradoxes’ are in fact not surprising or damaging to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
the neoclassical paradigm.

Imposing some set of conditions on the technology $T(\cdot)$ should be sufficient to assure that the real Wicksell effect is always negative. Such conditions would be of interest – especially if they could be empirically tested – since they would validate the qualitative conclusions derived from the one-good models often used in macroeconomics without any theoretical justification for ignoring capital aggregation problems. Moreover, Burmeister (1977; 1979) has proved that a negative real Wicksell effect is a necessary and sufficient condition for the existence of an index of capital, κ, and a neoclassical aggregate production function $F(\kappa)$ defined across steady-state equilibria such that (i) $c=F(\kappa)$, (ii) $r=F'(\kappa)$, and (iii) $F''(\kappa)<0$. Unfortunately, no set of such sufficient conditions is known, but the literature on capital aggregation suggests that they would impose severe restrictions on the technology.

See Also

- capital theory (paradoxes)
- reswitching of technique
- reverse capital deepening

Bibliography

Bliss, C. 1975. *Capital Theory and the Distribution of Income*. Amsterdam: North-Holland.

Burmeister, E. 1976. Real Wicksell effects and regular economies. In *Essays in Modern Capital Theory*, ed. M. Brown, K. Sato and P. Zarembka. Amsterdam: North-Holland.

Burmeister, E. 1977. On the social significance of the reswitching controversy. *Revue d'économie politique* 87, 330–50.

Burmeister, E. 1979. Professor Pasinetti's 'unobtrusive postulate', regular economies, and the existence of a well-behaved aggregate production function. *Revue d'économie politique* 89, 644–52.

Burmeister, E. and Dobell, A.R. 1970. *Mathematical Theories of Economic Growth*. New York: Macmillan.

Burmeister, E. and Turnovsky, S.J. 1971. The degree of joint production. *International Economic Review* 12, 99–105.

Burmeister, E. and Turnovsky, S.J. 1972. Capital deepening response in an economy with heterogeneous capital goods. *American Economic Review* 62, 842–53.

Kuga, K. 1973. More about joint production. *International Economic Review* 14, 196–210.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Robinson, J. 1956. *The Accumulation of Capital*. London: Macmillan.

Samuelson, P.A. 1966. The fundamental singularity theorem for non-joint production. *International Economic Review* 7, 34–41.

Swan, T. 1956. Economic growth and capital accumulation. *Economic Record* 32, 334–61.

Uhr, C.G. 1951. Kunt Wicksell, a centennial evaluation. *American Economic Review* 41, 829–60.

Wicksell, K. 1893. *Value, Capital and Rent*. Reprinted New York: Augustus M. Kelley, 1970.

Wicksell, K. 1934. *Lectures on Political Economy*. 3rd edn, London: George Routledge and Sons Limited, 1938.

How to cite this article

burmeister, Edwin. "Wicksell effects." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000067>
doi:10.1057/9780230226203.1833

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wicksell, Johan Gustav Knut (1851 – 1926)

C.G. Uhr
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

aggregate demand and supply; Böhm-Bawerk, E. von; capital theory; Cassel, G.; Davidson, D.; Malthus's theory of population; marginal cost pricing; marginal productivity theory of distribution; marginal utility theory; market socialism; monetary equilibrium; period of production; socialism; Stockholm School; Wicksell effects; Wicksell, J. G. K.

Article

Life and career

Johan Gustav Knut Wicksell was born in Stockholm on 20 December 1851, the youngest of six children of Johan and Christina Wicksell. One child died in infancy, so Knut grew up with three sisters a few years older than he, and a brother, Axel, one year older.
Knut's mother died when he was not quite seven, an event that greatly affected the sensitive boy. His father, a produce dealer, remarried in 1861 but died five years later when Knut was 15. After that the children moved to live for a time with an aunt and their maternal grandmother. In the last decades of his life Knut's father had become moderately well-to-do by investing profits from his grocery business in rental properties. The estate that was left at his death yielded an income sufficient to provide for the children and their education through gymnasium (high school), and for the two boys a start at the University of Uppsala.
At the gymnasium Knut had already shown considerable aptitude for languages and an unusual ability at mathematics. Thus, when he enrolled at Uppsala University 1869, it was with the intention of becoming a mathematician with physics as a second field of study.
From about age 15 Knut came increasingly under the influence of a pietistic pastor of the Swedish Lutheran Church. This religious phase lasted about seven years, in the course of which he became a devout Christian; he withdrew from most social activities to study the Bible and meditate. At the same time he made rapid progress in his studies of mathematics, physics and astronomy, earning his first degree, BS *cum laude*, in 1871, after only two rather than the usual four years at the university, and then

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
proceeding to graduate studies.
However, doubts had begun to assail his faith, and in spring 1874 he had an emotional crisis from which he emerged, and for the rest of his life remained, a free thinker. He became a strictly a-religious philosophical rationalist who, later on, became known as an outspoken and witty critic of the Christian religion in all its forms.
Until 1873 Wicksell maintained himself at the university on the modest annual income he received as his share of his father's estate, on a small inheritance from his grandmother, and on a succession of grants from private foundations. Now the last were drying up and the money from his grandmother was nearly gone. To add to his scant resources, he filled a vacancy as a teacher at a secondary school at Uppsala, 1873–4. The year after that he worked as a private tutor to the son of an ironmaster. Also from time to time he borrowed various amounts from one of his sisters who had established herself in Stockholm as a physiotherapist.
In fact, Wicksell's financial condition remained precarious and often severely strained, except for the years 1885 and 1888–9 when he was studying abroad largely supported by grants, for most of his adult life until 1901. Then, at age 50 and supporting his wife and two school-age sons, he was finally appointed professor extraordinarius (about equivalent to associate professor) at Lund University, and then, from 1904, he served there as ordinarius or full professor for 12 years, until his retirement in 1916.
In 1875 he passed two of three required examinations for the degree *philosophiae licentiatus* in mathematics (the *phil. lic*. is a graduate degree taken prior to the student's beginning work on the doctoral dissertation). Soon after that he began to doubt that he would be able to make any significant contributions to mathematics. While contemplating a change of career either to humanities or to the emerging social sciences, he immersed himself, over a long transition period, in the activities of the students' organization, the Student Corps. He was elected as its curator, 1877–9. In that post he became well known for his critical social views and for his surprising effectiveness as a speaker. At this time he also wrote some 'social indignation' poetry as well as some plays, one of which proved popular and was performed at Uppsala and also in some other towns.
In 1879 two events, in themselves inconspicuous, occurred which strongly influenced Wicksell's subsequent career. He moved to share an apartment with two advanced graduate students, H. Ohrvall in medicine, and T. Frölander in law, and he acquired a book just recently released in Swedish translation, G. Drysdale's tome, *The Elements of Social Science*, with its challenging subtitle, 'Physical, Sexual, and Natural Religion; An Explanation of the True Causes and Cure of the Three Primary Evils of Society – Poverty, Prostitution, and Celibacy'. This work, published in England 1854, became very popular in the Swedish translation of 1878, and went through over 30 reprintings over the years.
The three men, whose outlook on society was in several respects similar, became lifelong friends. What cemented their friendship was that they set about on their own and jointly to study Drysdale's thoroughly neo-Malthusian treatise. It discussed frankly several subjects then regarded as unmentionable in 'polite society', such as sex, methods of birth control, the allegedly harmful psychological effects of celibacy if continued for a decade or more past puberty, prostitution as the only alternative for the young among the poor, the need for family planning, and the need to limit population growth in order to raise the standard of living for the working class above bare subsistence.
Wicksell treated this book as a revelation. It focused his mind on 'the social question', that is, on the social sciences towards which his inclination guided him more and more. As an early result of studying Drysdale, supplemented by some writings of J.S. Mill, in February 1880 he gave a lecture to a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
temperance lodge at Uppsala on ‘The Most Common Causes of Habitual Drunkenness and How to Remove Them’. His address got a mixed reception but was reported in the local newspaper, which led to an insistent demand for him to repeat his lecture two weeks later in a much larger hall which was filled to overflowing.

There he attributed alcoholism, widespread among factory workers, to the poverty and monotony of their lives, with wife and children crammed into crowded and often insanitary housing. For this the local inn offered almost the only relief and relaxation available. And with the young workers this led to the use of the services of prostitutes since these workers for years earned too little to marry and start a family. The remedies he urged were for the medical profession to be assigned the duty of disseminating information about birth control techniques, and for the public health authorities to set and enforce standards of sanitation and room space per occupant in housing in the factory districts of cities and towns. The reaction to his lecture was strong. Papers by the Young Socialists and by some student organizations praised him. Medical and temperance organizations either reviled or ridiculed him, and several newspapers questioned his competence to pronounce on some of the sensitive issues he had covered.

From now on the die was cast. There would always be one or more reporters present at his future appearances, because these were certain to be newsworthy. Reporters would summarize his talks and write longer accounts about how his audience reacted, especially the critics and opponents among them, and how he, in turn, responded to critics. Most of the reportage depicted him as a non-revolutionary radical social reformer, and that was how public opinion came to view him. We may add that he himself did nothing to modify and much to strengthen that impression.

Later in 1880 he issued his lecture as a tract of some 90 pages and along with it a pamphlet, ‘Answers To My Critics’, both of which sold in several thousand copies. In fact, this became something of a pattern. Between 1880 and 1885, and again in 1886–7, after his return from his first stay abroad, Wicksell had in substance turned into a radical public lecturer and journalist. This was how he earned his spartan maintenance, by paid public lectures sometimes followed by publishing tracts based on them, and by paid articles written in neo-Malthusian spirit on various ‘social questions’ for several, sometimes in a given week for as many as ten, different city and town newspapers.

In 1885 he set aside his journalistic work for a time and completed the last requirement for the *phil. lic*. degree in mathematics by a research paper, the other requirements he had met in 1875. Now, however, he wanted to shift into the social sciences rather than go on for a doctorate in mathematics. To do that at any level higher than the elementary meant study abroad, for at that time the social science disciplines were not separate fields but were elements of the curricula in law, philosophy, the humanities or theology in Sweden's universities. But he had no funds for going abroad. Then help came unexpectedly. His sisters had an opportunity to sell the rental properties of the Johan Wicksell estate to a buyer at a favourable price if Knut and his brother, Axel, who had emigrated to the United States, would agree, as they did. Knut's share of the proceeds was sufficient to pay off his old debts and also to maintain him for about a year abroad, and so in autumn 1885 he went to London.

In London Wicksell spent his days studying some of the classical economists and treatises by Cairnes, Jevons, Walras and Sidgwick, his first exposure in depth to economics, and his weekends in meetings with persons to whom he was introduced by Charles Drysdale, an engineer who continued the neo-Malthusian activities his father, George Drysdale, had initiated. Thus he met prominent British neo-Malthusians, Annie Besant among them, Karl Kautsky and some labour leaders, and the leading Fabians.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
By summer 1886 he returned to Uppsala and Stockholm to resume his public lecturing, writing for several newspapers, and composing tracts. This was now a matter of necessity, for he had used up his patrimony during his stay in Britain. In 1886–7 he delivered 42 public lectures in towns in central Sweden, in Copenhagen and Christiania, for fees which paid very little above his travel and maintenance expense. The subjects he spoke on were as follows:

Marriage	14 lectures
Population Control	10 lectures
Socialism	6 lectures
Prostitution	5 lectures
Spiritualism	2 lectures
Why Not A Free-Thinker?	2 lectures
Religion	1 lecture
Euthanasia	1 lecture
Impression of Britain	1 lecture

At the end of 1885 Victor Lorén, a wealthy young man, greatly interested in promoting the social sciences after studying them in Germany with Roscher, bequeathed his estate to a foundation bearing his name, with instructions that it should be used for the promotion of studies and research and publications by scholars devoted to economics and related social sciences. Wicksell was still in London when early in 1886 he was informed that the Lorén Foundation was awarding him a grant for up to three years to study economics at universities in Germany and Austria. Lorén's relatives unsuccessfully contested his will in court, but this held up the grant until the summer of 1887, when the suit was settled.

If the Lorén Foundation had not given him that large grant (and later smaller ones for each of the five treatises he published between 1893 and 1906) Wicksell could hardly have become an economist, much less a major figure in this discipline. As it was, he went first to London to renew acquaintances. In October 1887 he went to the University of Strassburg to follow lectures by Brentano on labour economics, on money and credit by both Brentano and Knapp, and on economic distribution by Singer. In spring 1888 he was in Vienna to listen to Carl Menger's lectures. In July he returned for a short stay in Sweden. On his way there he met Anna Bugge, a Norwegian gymnasium teacher, who later became his wife. By autumn 1888 he was at the University of Berlin to follow the lectures of Adolph Wagner on public finance. In spring 1889 he returned to Sweden to seek a lectureship in economics at the University of Stockholm. He was turned down as being ‘too notorious’ a person. Summer 1889 he decided to spend the rest of his grant studying economics in Paris. Before going there he took a trip to Christiania to see Anna Bugge, with whom he had corresponded while in Germany. There he proposed a common-law marriage to her, but out of consideration for her parents she turned him down, whereupon he left in a huff for Paris.

A word may be needed here about the romantic side of Wicksell's life. It is known that he was infatuated in his early twenties with two young ladies. But he was always shy and very hesitant in socializing with young women. So the first young woman moved to Switzerland and married there. The second one was a case of love at a distance, for he failed even to make contact with her. The third and last incident occurred years later. For a part of the summer 1886 he was invited to stay in his friend Frölander's household in Stockholm, where he gave most of his lectures. But there he soon found himself becoming

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
infatuated with his friend's wife. So before he might say or do something to jeopardize their friendship, he made the proper excuses and returned to his lonely lodgings in Uppsala.

Anna, however, did not want to give up Knut. She decided she would be happier with than without him even at the cost of estrangement from her parents. She joined him in Paris that summer; he was then 37 and she 26.

In Paris he attended lectures on public finance by Leroy-Beaulieau and on population theory by Desmoulin, and he began to publish in economics. His first article, the translated title of which is ‘Empty Stomachs – Full Stores’ came out in a Norwegian journal *Samtiden* in 1890. His second article, ‘Überproduktion oder Überbevölkerung’ (Excess production or excess population) appeared in *Zeitschrift für die gesamten Staatswissenschaften*, also in 1890. In both he argued that it was fluctuations in capital formation that made the difference between prosperity and depression. In recovery a rate of capital formation is generated which fails to be sustained because consumption demand, though rising, lags behind the rate at which productive capacity expands on a growing capital base.

In summer he and Anna returned to Stockholm. Though soon to be a father (their first son, Sven, was born in October 1890, and a second son, Finn, in 1893), Wicksell had no settled way of earning a living. Economics was then taught only in the faculties of law. Those teaching it in addition to a doctorate in economics also had to have at least an undergraduate degree in law in order to give courses on law and economics as related mainly to taxation and public finance. So he had no alternative but to return to being a freelance journalist and public lecturer.

During the years 1890–9 Wicksell had more trials and tribulations, only a few of which can be related here. He gave rather few public lectures, but some had a very negative effect on his public image.

In 1892 the government wanted to increase the duration of the compulsory military service to strengthen the country's defences. In November Wicksell lectured in Stockholm on the question, ‘Can Sweden Protect her Independence?’ He argued, and most of his listeners might have agreed with him, that no matter how long the draft were extended, it would not be adequate for defending Sweden against attack by a major military power. But they disagreed vehemently when he went on to say that since the country could not defend itself on its own resources, it would make better sense to disarm and use the resources set free from defence for other domestic purposes. Then Sweden ought to negotiate for incorporation into the Russian empire with its much greater military resources. In return for the protection thus provided, the Swedes with their long traditions of democracy ought then to play a civilizing role within and for the Russian empire.

This performance earned him the sobriquet of ‘defence nihilist’, which did not deter him 12 years later when another draft extension was proposed from repeating this same lecture, May Day 1904. At that time it occasioned even greater offence and ridicule than in 1892.

His article ‘Kapitalzins und Arbeitslohn’ (Interest and wages), published in the *Jahrbücher für Nationalökonomie und Socialwissenschaft und Statistik* 1892, formed the basis for the marginal productivity theory of distribution – one of Wicksell's main contributions to economic theory – which he developed in his first treatise, *Uber Wert Kapital und Rente* 1893 (*Value, Capital and Rent*, translated 1934). This remarkable work received initially almost no attention in Sweden, but was favourably reviewed by both Böhm-Bawerk and Walras.

Next he turned to an examination of Sweden's taxes in his popular tract, *Our Taxes – Who Pays and Who Ought to Pay Them*? (99 pp., 1894), issued under the pseudonym of Sven Trygg. He was outraged at the regressiveness of the country's taxes. That, he concluded, had to be a consequence of the fact that

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
only the well-to-do could vote, as income and property qualifications for the franchise excluded almost all the workers and most of the small farmers.

The analysis of that tract was extended and refined in his second treatise, *Finanztheoretische Untersuchungen* (Studies in the theory of public finance), 1896. There he urged that the major part of the revenue burden be shifted from indirect to direct progressive taxes on income and wealth. That treatise also embodied his design on an ‘equitable’ tax system based on an application of marginal utility theory to the public sector, and a methodology (essentially marginal cost pricing) for pricing pure and less than pure public goods, the services of public utilities, and the products of market-sharing oligopolies and cartels.

In fall 1894, Wicksell applied at Uppsala University to have *Value, Capital and Rent* evaluated as a doctoral dissertation. The answer was ‘no’, with the added advice to use it for a viva voce examination of a *phil. lic*. degree in economics. David Davidson was appointed examiner and Wicksell passed with high marks in May 1895. Next he needed the doctorate. In 1896 he submitted the first part of his *Finanztheoretische Untersuchungen*, ‘Theory of incidence of taxation’, as a dissertation. Again Davidson was chief examiner, and the degree was awarded Wicksell *magna cum laude*.

That done, he began research on monetary theory and policy, which he completed as his third treatise, *Geldzins und Güterpreise*, 1898 (*Interest and Prices*, translated 1936), the home of the Wicksellian ‘cumulative price level fluctuations or processes’, allegedly generated by a divergence between the rate of return on newly created real capital and the bank-dominated market rate of interest.

Now he applied both at Stockholm and Uppsala universities for a docentship but was rebuffed because he lacked a degree in law. From 1890 into 1897 he had maintained his family slightly above subsistence level by earnings from his newspaper articles and tracts and from a succession of Lorén grants. However, in autumn 1897 he decided at real hardship, with no more Lorén money, to move from Stockholm to Uppsala to devote his entire energy to cramming through law courses as fast as possible to a *juris candidatus* of LL.B. degree. To do this he had to maintain his family by borrowing from his friends Ohrvall, a physician, and Frölander, a banker-lawyer, both of whom were doing well. In 1899, in less than two years, he had earned his law degree, which usually takes undergraduates four years. He was appointed a docent at Uppsala University but without fixed salary. Consequently his income depended on how many law students came at a given fee per head to attend his tutorials.

At the Lund University faculty of law a professorial vacancy was created when an older professor's post, viewed as overloaded, was split to shift its courses in tax law and economics from the old position to the new one. But Parliament, in approving this, had voted less money for it than a full professor's salary. Wicksell and three others, including Gustav Cassel, competed for this post of professor extraordinarius. As the other candidates (Cassel for lack of a law degree) were eliminated as not sufficiently qualified, the appointment was offered to Wicksell in January 1900. For complex reasons the upgrading of this to ordinary or full professorship was delayed until January 1904, when Wicksell, at the age of 53, was finally securely established as a full professor.

At Lund, where his teaching of tax law courses required much more preparation than economics, he still found time to write *Föreläningar i Nationalekonomi* (*Lectures on Political Economy I*, 1901, translated 1934). *Lectures I* were an expansion and improvement, especially in capital theory, over what he had presented in *Value, Capital and Rent*.

His courses in law as related to taxation were well attended but those in economics attracted very few students, it being an elective subject. He soon found out that the students lacked the background to get

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
much out of a semester on *Value, Capital and Rent* and another on *Interest and Prices*. So he shifted his presentation from pure to applied economics to subjects such as agriculture and industry, commerce and consumption, social movements, social insurance, economic crises and inflation.

He had good relations with students. His approach to them was friendly. They, in turn, liked or were amused by his idiosyncrasies, and they admired his courage to fight for his convictions.

Unlike most professors, who at that time lectured in formal dress, swallowtail coats and all, Wicksell appeared in ordinary, rarely well-pressed, street clothing. Instead of a top hat or a Derby he wore a visored cap, much like a fisherman's. Since he lived some distance from the university, he did the family's marketing at the nearby open-air market before his morning lectures. Consequently, as he strode in to the lecture room, he would adorn one side of the lectern or the other with his market basket filled with produce, meats and fruits.

In 1905 he issued one of his best and last tracts, *Socialiststaten och nutidssamhället* (The socialist state and contemporary society, 40 pp.). He restated more systematically his perspective on socialism which he had lectured on in the 1880s. First he made it clear he considered a limited but not a complete achievement of a socialist economy (with all means of production other than labour collectively owned and administered) to be inevitable in the future. Under universal adult suffrage the workers would be the political majority. As such, they would not for long tolerate the great inequalities of income and wealth and the economic instability (of employment and economic insecurity and dependence in old age) of laissez-faire capitalism without seeking and taking remedial measures.

He warned against drastic measures of income redistribution taken by a workers' government suddenly come to power. That would only yield a temporary gain followed by loss as private capital accumulation would all but cease before the workers' regime would have developed the means to replace it by public accumulation. A socialist economy is best built gradually by peaceful means and under democratic governance. Nationalization initially of natural monopolies and cartels might suffice if followed by substantial expansion of tax supported social security and social insurance schemes. For the sake of efficiency, he held it was best to leave farming and most varieties of genuinely competitive enterprises in private and/or cooperative ownership.

Consequently he argued for a form of market socialism with a well developed welfare state. It is surprising to recognize the great extent to which his social vision has become a reality in Sweden (and in Scandinavia as a whole) after more than half a century of Social Democratic rule.

In 1906 Wicksell published *Lectures on Political Economy II*, the volume on money and credit. In part an expansion and revision of what he had put forth in *Interest and Prices*, yet *Lectures II* were much more than that. They were epoch-making less for their particular findings than for the broad framework and methodology they provided for analysis of money and credit. *Lectures II* were translated first into German in 1922 when, in the midst of the German hyperinflation they were read with greater than usual interest, and into English in 1935.

Wicksell's years at Lund were very productive. He wrote about 50 articles and took an active part in the tax reform of 1910, in the national pension legislation of 1913, and, after the outbreak of the First World War, along with Davidson, he played an important role in the legislation and policies relating to banking, currency and exchange controls.

His work had continued in a tranquil manner until 1908. Then a young 'anarchist agitator' was sentenced to prison for 'disturbing the religious peace' by public blasphemy. He had published a parody of the Wedding at Cana in a socialist newspaper. His case, and two or three similar ones that had

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
occurred earlier, impressed Wicksell as infringements by the courts of freedom of speech and press, guaranteed by the Swedish constitution. Against better advice he decided to make a test case of himself. Accordingly in November 1908 at Stockholm he lectured to a large audience on ‘The Throne, the Altar, the Sword, and the Bag of Money’ in which, inter alia, he satirized the story of the Immaculate Conception. He raised and answered the question:

> Why was not Joseph, the betrothed of the Virgin Mary, rather than the Holy Ghost allowed to father Jesus? Because then the world could not have been saved! Joseph's rights as an individual had to be set aside for the salvation of the many millions of souls in past centuries who would otherwise have gone to perdition and the further millions now and for all time to come until the Last Judgement.

Wicksell was tried and sentenced, against the protests of Social Democrats, organized workers and liberals, to two months in prison. He was allowed to select the jail where he would serve his time. Early in 1910, after a higher court had sustained the lower court's decision, he chose the jail, known to be better than most, at the small fisherman's town of Ystad in southern Sweden. There he suffered no hardship. His university salary was withheld as long as he was a guest of the government. He used his time to advantage by writing his last tract, *Laran om befolkningen, dess sammansättning, och förändringar* (The Theory of Population, its Composition, and Models of Change, 1910, 52 pp.). There, apart from the clear demographic analysis it presented, he reiterated the conclusion from his public lectures of the 1880s, that, because of partial depletion and increasing scarcity of natural resources (in Sweden's case primarily timber and iron ore), the country's optimum population should be three million instead of its five million inhabitants, and for Europe a reduction to three quarters of its population as of 1910. Like Malthus and many other writers on population, while he acknowledged the productivity increasing effect of technological progress, he failed to see, and greatly underestimated the fact, that some of the new technology virtually adds to existing resources, in part by turning former waste products to productive uses, in part by increasing the number of uses to which existing resources can be turned.

Wicksell's remaining years at Lund passed quietly. But as his retirement was approaching it threatened renewed hardships for him and Anna. Before coming to Lund they had no savings, and when leaving, they had very little more than their household effects in rented housing. Since Wicksell had served only 16 years at the university, compared with colleagues who at age 65 had usually served 25 or more years, he was barely entitled to two-thirds of the usual professional pension. As the First World War inflated prices, especially in Stockholm to which city he and Anna insisted on moving, two-thirds pension would not pay for much more than house rent.

Two of his friends who were members of parliament succeeded on a motion to obtain a supplementary allowance for him which raised his pension to 90 per cent of the usual amount. There still remained the problem of housing, which had become very expensive in the capital. So his two parliamentarian and several other close friends, including David Davidson and Eli Heckscher, gathered together and by their personal contributions they raised enough money to buy a lot for a house and garden in Mörby, a suburb of Stockholm, and to initiate construction to Anna's specifications. To complete the building of the house, Wicksell negotiated a small mortgage. By Christmas 1916 he and Anna moved into the first house they could call their own.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Now, in his last decade, a new phase of life began for both of them. Anna, who had taken a law degree in 1911 at Lund and had become a leader in the suffrage movement and later the peace movements, now had greater opportunity to be effective than at Lund. Wicksell, freed both from financial worries and the teaching of law courses, could devote himself full time to research and professional activities as an economist with the much greater resources at his disposal for research and opportunities for consultation of Stockholm as compared with Lund. The years 1917–26 were probably the most satisfactory and happiest in their lives.

Wicksell soon became very active. He wrote 29 articles from Mörby on wartime inflation and how to roll it back, on Scandinavia's post-war monetary problems, and on capital theory. From 1915 he had been a consultant to the governor of the Bank of Sweden. In 1916 he and Davidson were appointed to a parliamentary committee on banking and credit. Wicksell's involvement with its work and that of its successor committees lasted until his death. He and Davidson were both appointed as experts to another parliamentary committee on taxation of income and property which remained active from 1918 to 1922. These assignments improved Wicksell's finances, for he was paid somewhat more than his pension for his work with these committees.

Among achievements attributable to Wicksell's and Davidson's collaboration in these councils was the adoption in 1916 of the ‘gold exclusion policy’ for the Bank of Sweden (to limit inflation the Bank was relieved of the obligation to issue currency at the pre-war mint ratio to gold that had been turned in to it from Sweden's export surplus, and was given power to lower the price of gold in terms of currency). A second achievement was a thorough revision and improvement of the country's taxation of income and wealth.

In this decade, Wicksell also became a much sought-after adviser to young economists about their dissertations. At Lund he had only had three students in economics who took the intermediate graduate degree of *phil. lic*. under his guidance. In Stockholm, as a very active member of the Swedish Economics Association, and an indefatigable participant in the Economy Club, its inner circle of economists (as distinct from such members as bankers and industrialists), he had easy access to the club members’ graduate students. He was made president of that club, 1917–22. It was a source of satisfaction for him to be sought out to share in the problems of the young men.

Thus his teaching did not stop with his retirement, for Emil Sommarin, Erik Lindahl, the brothers Gustav and Johan Åkerman, Bertil Ohlin, and probably others such as Palander, Lundberg and Hammarskjöld consulted him about their dissertations in addition to benefiting from studying his treatises.

In the 1930s these persons, self-confessed ‘Wicksellians’, formed the core of the ‘Stockholm School of economists’. However, he remained estranged from Gustav Cassel, the third of Sweden's leading economists in the 1920s. This had nothing to do with Cassel's competition with him for the position at Lund in 1900; it was due to Cassel's character. Wicksell found him to be intellectually arrogant, rarely acknowledging the contributions or predecessors whose works he was using, and acting as if economics had been in its infancy prior to Cassel. Wicksell found some of his work to be superficial and his interpretations of several points in economic theory to be misleading. This he expressed clearly in his rather severe review in 1919, of Cassel's magnum opus, *The Theory of Social Economy* (Wicksell, 1919b).

After that the breach between them was complete. Cassel never replied to Wicksell's review. As a result, Cassel remained an outsider to ‘the Stockholm School’, although he was the mentor of one of its leading members, Gunnar Myrdal.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
In spring 1926, Wicksell was working on an article ‘Zur Zinstheorie’ (On the theory of interest) for a Festschrift for Friedrich von Wieser, when he fell ill with a stomach disorder which was further complicated by pneumonia. He died on 2 May 1926, at age 74.
His widow Anna, then a delegate from Sweden to the League of Nations, survived him until 1928. Their eldest son, Sven, became a professor of statistics at Lund University and died in 1939. Their young son, Finn, died in an accident in 1913 at the age of 19, when a medical student at Lund University.
Knut Wicksell would doubtless have objected to the elaborate funeral that was arranged for him, evidently with his widow's consent. Throughout life he had steadfastly rejected as meaningless and offensive to his sense of rationality all pomp and circumstance, academic formalities along with marriage ceremonies, baptism and confirmation for his children.

Contributions to economics

In his own lifetime Wicksell did not receive much recognition for his creative work, not even in Scandinavia. It was not until the 1930s, when at the initiative of R.F. Kahn and J.M. Keynes, *Geldzins as Interest and Prices* and *Vorlesungen* as *Lectures on Political Economy I and II* were translated, that economists generally heard of Wicksell. Yet is it clear that his stature in the annals of economics grew steadily after his death. In summary form, his main contributions were these.
In *Value, Capital, and Rent* he performed a remarkable labour of synthesis. He adopted the marginal utility and marginal productivity theory of value of Jevons, Menger and Marshall, added to it the Böhm-Bawerk analysis of capital, and fused the result in a Walrasian comparative static general equilibrium framework. In this process he became a founder of the marginal productivity (product exhausting) theory of distribution shortly ahead of Wicksteed. In his *Studies in the Theory of Public Finance* he pioneered a marginal utility approach to the public sector, synthesizing the benefit and ability principles of taxation, and urging that services of public sector enterprises and natural monopolies be provided on a marginal cost pricing basis.
In *Lectures I* he completed the restructuring, begun in *Value, Capital and Rent*, of Böhm-Bawerk's theory of capital and interest. He reduced Böhm's trinitarian ‘grounds’ for interest to the simpler, more realistic explanation as the marginal productivity of waiting. He relaxed Böhm's quantification of capital as an average period of production by a concept of capital as the time structure of inputs invested for various terms in production. He showed that this structure was capable of change in at least two dimensions, width and height. He endeavoured with partial success (on problems still unresolved about ‘Wicksell effects’ and ‘switching of techniques’) to develop a theory of the modes of change of this time structure of production, how it changes and interacts with variations in wages, rent, and interest, in conditions both of capital accumulation and technological change. He extended his treatment of these relationships from comparative static to dynamic analysis, using clear mathematical models for this purpose.
The greatest contribution to monetary analysis, both in terms of novelty back in 1898 and 1906, and in terms of eventual influence by fortifying the related analysis, independently worked up by J.M. Keynes three decades later in his *Treatise on Money* (1930), was Wicksell's work on monetary theory in *Interest and Prices* and especially in *Lectures on Political Economy II*.
Wicksell was a pioneer of applying an aggregate demand–supply approach with emphasis on the relations between investment and saving, to explain variations in value of money or fluctuations in

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
prices: ‘Any theory of money worthy of the name must be able to show why pecuniary demand for goods exceeds or falls short of the supply of goods in given conditions’ (*Lectures II*, p. 160).
Most versions of the quantity theory had the price level varying directly and proportionately to changes in the quantity of money. In that theory there was no link from the elasticity, as affected by bank credit, and quantity of money with individual income dispositions and entrepreneurial production decisions. Wicksell provided such a link in his hypothesis that in the absence of certain disturbances over which central banks have no control (such as large influx or efflux of gold, internal cash drains, large government deficits financed by loans from the central bank, fiat issues of inconvertible currencies, sudden and large changes in productivity or supply of goods), price level fluctuations were due to a persistent divergence between the bank rate or market rate of interest and the real rate, defined as the expected rate of return on newly produced capital goods.

> The fluctuations of commodity prices, which are *not* due to a change in gold production [a gold standard currency is assumed here] … have another cause … changes in the real rate of interest … [to which] … the loan-rate does not adapt itself quickly enough. (*Lectures II*, p. 205)

Thus Wicksell's analysis showed, contrary to that of the simple quantity theory, that it was the quantity of money that adapted itself to the movement of the price level, and in doing so affected the distribution of income and the dispositions to invest and save in the process.
In his analysis monetary equilibrium and stability of prices required the simultaneous fulfilment of the conditions that: (i) the money rate of interest correspond to the real rate; (ii) at that money rate demand for loans for investment and for cash for real balances equal the supply of savings by individuals and business enterprises; and (iii) that interest rate must be neutral in its effect on prices. Then: ‘… equilibrium must *ipso facto* obtain – if not disturbed by other causes – in the market for goods and services, so that wages and prices will remain unchanged’ (*Lectures II*, p. 193).
The consistency and compatibility of Wicksell's three criteria for monetary equilibrium and a critique of them in conditions of changing productivity by Davidson were given a thorough exegesis and analysis in the later 1920s and early 1930s by Lindahl, Myrdal and Ohlin. Their work combined with the efforts of younger colleagues such as Lundberg, Hammarskjold and Svennilson greatly expanded the heritage of Wicksellian economic theory and gave rise to the doctrines associated with the Stockholm School of economics.

See Also

- Stockholm School

Section I above relies heavily on both Blaug (1985) and Gårdlund (1958) and in addition on information obtained in correspondence with the late Professors Erik Lindahl and Emil Sommarin.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1890a. Tomme maver – og fulde magasiner [Empty stomachs and full stores]. *Samtiden* [Contemporary Times, a Norwegian periodical] 1, 245–7, 293–320.

1890b. Überproduction oder Überbevölkerung [Excess production or excess population]. *Zeitschrift für die gesamten Staatswissenschaften* 46.

1892. Kapitalzins und Arbeitslohn [Interest and wages]. *Jahrbücher für Nationalökonomie* 59, 552–74.

1893. *Über Wert, Kapital, und Rente*. Jena: G. Fischer. Trans. S.H. Frowein as *Value, Capital and Rent*, London: Allen & Unwin, 1954. Reprinted, New York: Augustus M. Kelley, 1970.

1894. Våra skatter: hvilka betalar dem, och hvilka border betala? [Our taxes: who pays them, and who ought to pay them?] Stockholm. This was one of Wicksell's early and very popular tracts, written under the pseudonym of Sven Trygg. It provides non-technical background and may be viewed as an introduction to Wicksell (1896).

1896. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steurewesens Schwedens*. Jena: G. Fischer. Pages iv–vi, 76–87 and 101–59 trans. J.M. Buchanan as Chapter 6 of *Classics of Public Finance*, ed. R.A. Musgrave and A.T. Peacock, London: Macmillan, 1958; 2nd edn, 1967. This treatise has only been translated in part; the untranslated sections deal with an historical sketch of the development of Sweden's system of taxation from the early 16th century up to the 1890s.

1898. *Geldzins und Güterpreise bestimmenden Ursachen*. Jena: G. Fischer. Trans. R.F. Kahn as *Interest and Prices. A Study of the Causes Regulating the Value of Money*, London: Macmillan, 1936.

1900. Om gränsproduktiviteten såsom grundval för den nationalekonomiska fördelningen. *Ekonomisk Tidskrift* 2, 305–37. Trans. as ‘Marginal productivity as the basis for distribution in economics’ in *Selected Papers on Economic Theory by Knut Wicksell*, ed. E. Lindahl, London: Allen & Unwin, 1958.

1901. The theory of exchange in its final form. In *Föreläsningar i nationalekonomi. Häft I*. Stockholm, Lund: Fritzes, Berlingska. The 3rd Swedish edn (1928) of this volume translated by E. Classen as *Lectures on Political Economy*. Vol. 1: *General Theory*, London: Routledge & Kegan Paul, 1934.

1902. Till fördelningsproblemet. *Ekonomisk Tidskrift* 4, 424–33. Translated as ‘On the problem of distribution’, in *Selected Papers on Economic Theory by Knut Wicksell*, ed. E. Lindahl, 1958.

1904. Mål och medel i nationalekonomien. *Ekonomisk Tidskrift* 6, 457–74. Translated as ‘Ends and means in economics’ in *Selected Papers on Economic Theory by Knut Wicksell*, ed. E. Lindahl, 1958.

1905. *Socialiststaten och nutidssamhället* [The socialist state and contemporary society]. Stockholm. A popular tract, 40 pp.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1906. *Föreläsningar i nationalekonomi. Häft II: Om penningar och kredit*. Stockholm and Lund: Fritzes, Berlingska. The 3rd Swedish edn (1929) trans. E. Classen, ed. L. Robbins, as *Lectures on Political Economy*. Vol. 2: *Money*, London: Routledge & Kegan Paul, 1935; reprinted 1946.

1907. Krisernas gåta. *Statsekonomisk Tidskrift*. Oslo. Trans. by C.G. Uhr as ‘The enigma of business cycles’ in *International Economic Papers* No. 3 (1953), 58–74. This article expands on and lifts to a higher level of analysis the rather primitive treatment of business cycles presented in Wicksell (1890a and 1890b) and in Wicksell's brief ‘Note on trade cycles and crises’ in *Lectures on Political Economy*, vol. 2.

1910. *Läran om Befolkningen, dess Sammansättning och Förändringar* [Theory of population, its composition, and modes of change]. Stockholm: A. Bonniers, 92 pp. A popular tract written while Wicksell served a two-month jail sentence in southern Sweden for disturbing the religious peace by public blasphemy in a public lecture he had given in Stockholm in 1908.

1919a. Växelkursernas gåta. *Ekonomisk Tidskrift* 21, 87–103. Trans. as ‘The riddle of the foreign exchanges’ in *Selected Papers on Economic Theory by Knut Wicksell*, ed. E. Lindahl, 1958.

1919b. Professor Cassels ekonomiska system. *Ekonomisk Tidskrift* 21, 195–226. This article, highly critical of several of Cassel's interpretations and formulations of economic theory, has been translated and added as Appendix 1 in *Lectures on Political Economy*, vol. 1, 1934.

1923. Realkapital och kapitalränta. *Ekonomisk Tidskrift* 25, 145–80. A review and a mathematical elucidation of the analysis in G. Åkerman's doctoral dissertation *Realkapital und Kapitalzins*, Lund, 1923. It has been translated and added to *Lectures on Political Economy*, vol. 1, as Appendix 2, ‘Real Capital and Interest’. Among other things, this article features demonstrations of both the ‘Wicksell effect’ and the reversal of that ‘effect’ in an effort to determine the relationship between the optimum durability of fixed capital and the rate of interest. As such, this article, long after Wicksell's death, has played an important role in the controversy over capital theory between the Cambridge (England) and Cambridge (Massachusetts) economists.

1925. Valutaspörsmålet i de skandinaviska länderna. *Ekonomisk Tidskrift* 27, 103–25. This article has been translated and added to *Interest and Prices* as an Appendix as ‘The Monetary Problem of the Scandinavian Countries’. It represents the several qualifications Wicksell was moved to add to his norm of price level stabilization for monetary policy, qualifications both to meet Davidson's criticism of this norm and to incorporate lessons from the monetary experiences and upheavals of the First World War and its aftermath in the early 1920s. The qualifications were introduced to make allowance for significant increases in productivity, and for something like its opposite, severe commodity shortages due to wartime blockades, crop failure and also for significant issues of fiat money by governments running large deficit budgets, and so on.

A full-scale bibliography of all of Wicksell's published writings is now available. Its author is Dr. E.D.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Knudtson, who has written *Knut Wicksells Tryckta Skrifter 1868–1950* [The published writings of Knut Wicksell, 1868–1950] edited by T. Hedlund-Nyström, and issued in the series Acta Universitatis Lundensis, Section I, Theologia-Juridica-et-Humaniora, No. 25, and published by the C.W.K. Gleerup Publishing House, Lund, Sweden, 1976. This bibliography runs to slightly more than 100 pages and accounts for 889 titles or items dating from Wicksell's student days in the later 1860s through his entire career, inclusive of his many popular articles for Sweden's leading newspapers, and beyond, to include also listings of the translations of his major works and reviews of these translations, which appeared between the decade or two after Wicksell's death.

Bibliography

Blaug, M. 1985. *Economic Theory in Retrospect*, 4th edn. Cambridge: Cambridge University Press. Chapter 12, especially ‘Reader's Guide to *Lectures on Political Economy*, vol. I’, pp. 546–565: see also some of the earlier part of this chapter, pp. 498–519. Chapter 15, ‘Reader's Guide to Wicksell's *Lectures*, vol. 2’, pp. 646–51, and also in the earlier part of this chapter, pp. 637–45.

Gårdlund, T. 1958. *The Life of Knut Wicksell*. Stockholm: Almqvist & Wiksell. The original Swedish version of this biography; *Kunt Wicksell – Rebell i Det Nya Riket* [Rebel in a new realm], Stockholm, 1956, is somewhat longer than the English version, largely because it contains, but the English omits, notes on source materials and valuable bibliographical data, pp. 374–405.

Hedlund-Nyström, T. and Jonung, L. 1986. Knut Wicksell's opublicerade manuscript. 2 vols. Mimeo, Department of Economics, University of Lund.

Jonung, L. 1987. Knut Wicksell's unpublished manuscripts. Mimeo, Department of Economics, University of Lund.

Uhr, C.G. 1951. Knut Wicksell – a centennial evaluation. *American Economic Review* 41, 829–60.

Uhr, C.G. 1960. *Economic Doctrines of Knut Wicksell*. Berkeley: University of California Press.

How to cite this article

Uhr, C. G. "Wicksell, Johan Gustav Knut (1851–1926)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000068>
doi:10.1057/9780230226203.1832

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wicksteed, Philip Henry (1844 – 1927)

Ian Steedman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

abstract utility; Aristotle; choice; cost of production; distribution theory; economic man; Euler's Theorem; exchange value; George, H.; intensive rent; interpersonal utility comparisons; Jevons, W. S.; labour theory of value; land nationalization; land reform; law of indifference; law of one price; marginal productivity theory; marginal productivity theory of the interest rate; marginal utility; marginalism; Marx, K. H.; opportunity costs; orderings; ophelemity; preference ordering; preferences; price discrimination; product exhaustion; real cost doctrine; reservation price; revealed preference theory; Shaw, G. B.; utility theory of value; Wicksteed, P. H.

Article

Wicksteed was born in October 1844 in Leeds, where his father, Charles Wicksteed, was a Unitarian minister. He died, at the age of 83, in March 1927, at Childrey in Berkshire. He attended Ruthin Grammar School in North Wales and then University College School, London, before studying at University College London (1861–4) and at Manchester New College (1864–7) in Gordon Square nearby. He received his Master's degree, with a gold medal for classics, in 1867. Wicksteed then became a Unitarian minister, first at Taunton in Somerset (1867–9), then at Dukinfield, east of Manchester (1870–4), and finally at Little Portland Street Chapel, London (1874–97). He left the ministry in 1897 and thereafter earned his living by writing and lecturing. From 1887 to 1918 Wicksteed was a most active University Extension Lecturer, lecturing on Wordsworth, Dante, Greek tragedy, Aristotle and Aquinas – and economics. He never held a university post.
The great breadth of Wicksteed's intellectual activity was far from being confined to his Extension lecturing. He had a considerable linguistic talent; whilst a minister in Dukinfield, for example, he learned Dutch for the express purpose of translation into English of Oort and Hooykaas's *Bible for Young People* (six volumes, 1873–9). And he completed a translation, with F.M. Cornford, of Aristotle's *Physics* only days before his death. Yet it was as a translator, expounder and interpreter of Dante that he became most widely known; his work as a Dante scholar, which extended over more than 40 years,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
included translations of and commentaries on the *Vita Nuova*, the *Convivio*, *De Monarchia* and the *Divina Commedia*. Combined with his theological and philosophical interests, this study of Dante led Wicksteed to Aquinas and thus to the writing of his *Dante and Aquinas* (1913) and his *Reactions between Dogma and Philosophy, illustrated from the Works of S. Thomas Aquinas* (1920). That a study of Aquinas' thought by a former Unitarian minister could be reviewed favourably in the *Blackfriars Review* is perhaps an indication of the catholicity of Wicksteed's interests and capacities. Nor did those interests extend only to the past; for example, Wicksteed publicly defended the poetry and drama of Ibsen at a time when Ibsen's work was the object of considerable hostility in England. And Wicksteed's numerous contributions to the *Inquirer*, the Unitarian newspaper, over a span of some 50 years, relate not only to theological and literary matters but also to many economic and political issues.

While he had earlier been influenced by the thought of Comte and of Ruskin, Wicksteed's first direct contact with political economy took the form of reading Henry George's *Progress and Poverty*, of corresponding with George in 1882 and 1883 and of being a co-founder, in 1883, of the Land Reform Union, which supported George's lecture tour of England and Scotland in 1883–5. (He continued to support some form of land nationalization long after this time.) It was probably late in 1882 that Wicksteed began to study the work of Jevons and thus to become 'Jevons's only disciple'. By early 1884, however, he was playing an active role in promulgating Jevonian theory in the Economic Circle, which met until 1888 or 1889 (to be followed by the Economic Club and the British Economic Association, later to become the Royal Economic Society). Wicksteed became a close friend of George Bernard Shaw and of Graham Wallas, and was well-informed about Fabian and other aspects of the 'social movements' of the 1880s and 1890s, but was generally an acute and sympathetic observer, rather than a direct participant in those movements. He was, however, a founder member, in 1891, of the Labour Church movement and continued to give that movement strong support even after other early supporters had withdrawn their active sympathy.

Wicksteed published three books in the field of economics. The first, *The Alphabet of Economic Science, Part I. Elements of the Theory of Value or Worth*, was published in 1888; the second, *An Essay on the Co-ordination of the Laws of Distribution*, was published in 1894, and the third work, *The Common Sense of Political Economy*, was first published in one volume in 1910; a second edition in two volumes, edited by L. Robbins and containing various papers and reviews by Wicksteed, was published in 1933.

Of Wicksteed's other writings in economics, the most important are probably his critique of *Das Kapital*, published in the socialist journal *To-Day* in 1884; his article on Jevons's *Theory of Political Economy* (1889); his various contributions to the first (1894) and second (1925) editions of *Palgrave's Dictionary of Political Economy*; and his 'Scope and Method of Political Economy' paper (1914), which originated as Wicksteed's Presidential Address to Section F of the British Association for the Advancement of Science in 1913. (All of these papers appear in the Robbins edition of the *Common Sense*.)

There are a few extant letters (Sturges, 1975, p. 128) and some handwritten sermons and letters at Manchester College, Oxford, but Wicksteed wrote to a correspondent (J.M. Connell) that 'I have never kept careful records of my life and have next to no documents'. As to secondary material, the following may be consulted: Herford's full biography (1931); Robbins's editorial introduction (1933); the relevant chapters in Hutchison (1953) and Stigler (1941); Steedman's editorial introduction (1987); and the relevant entries in the *Encyclopaedia of the Social Sciences* (by H.E. Batson, vol. 15, 1935) and in the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
International Encyclopedia of the Social Sciences (by W.D. Grampp, vol. 16, 1968).
Wicksteed's first contribution to economic theory was his October 1884 critique of *Das Kapital*, Volume 1. Resulting perhaps from a Fabian challenge within the Economic Circle, it was published in *To-Day*, which, in 1884, carried articles by many of the leading socialists of the time. Wicksteed's critique certainly converted George Bernard Shaw from the Marxian to the Jevonian theory of value and, since no effective reply was published, may have had a wider influence on the theory adopted by British socialists: some writers have regarded Böhm-Bawerk's later attack on the labour theory of value, of 1896, as inferior to that of Wicksteed. Displaying a firm grasp of many of the specific features of Marx's argument, Wicksteed was able to focus clearly on two central issues. Is the exchange value of ordinary commodities determined by labour time? And does Marx's argument apply to ‘labour force’ (as Wicksteed called it)?
With respect to the first question, Wicksteed follows Marx in saying that if two commodities are exchanged they must simultaneously differ from one another, to motivate the exchange, and have something in common, to make them commensurable. But he then seizes on Marx's point that labour time only ‘counts’ when producing something useful and argues that it was merely arbitrary for Marx to assert that commodities have only ‘abstract labour’ in common. On the contrary, Wicksteed insists, all commodities have ‘abstract *utility*, *i.e.*, power of satisfying human desires’ in common; moreover, this is just as true of exchangeable objects which are not freely reproducible. Thus, in a neat twist of the argument, he proposes ‘*abstract utility* as the measure of value’. Wicksteed argues, nevertheless, that for freely reproducible commodities equilibrium relative prices will *coincide* with relative labour costs – but this is not because labour quantities determine prices but because labour will be so allocated as to produce those quantities of the commodities which imply marginal utilities proportional to the given labour costs. For old masters, the products of monopolized industries, and so on, even this coincidence will not hold.
Turning to ‘the value of labour-force’, Wicksteed then observes that, in a non-slave society, labour is not allocated to the production of ‘labour-force’ under competitive pressures. He deduces that there is no reason to expect that the ratio of the money wage rate to the labour value of the necessary wage goods will be equal to the money price-embodied labour ratio for ordinary commodities. Consequently, he concludes, Marx has failed to show that ‘surplus labour’ is the source of profit. Neither George Bernard Shaw nor any other contributor to *To-Day*, or to the other British socialist periodicals of the period, provided a remotely effective reply to Wicksteed's argument.

The Alphabet

Wicksteed's *Alphabet of Economic Science*, of 1888, was dedicated to members of the Economic Circle who had ‘met to discuss the principles set forth in these pages’. Both the subtitle of the volume and certain remarks in Wicksteed's Introduction suggested that there might be successor volumes but this proved not to be the case. Although the work received the approbation of both Edgeworth and Pareto, it did not find a wide audience, which is perhaps not surprising given that it was simultaneously introductory and somewhat mathematical. As in his other books, Wicksteed disclaimed originality but showed himself to be, at the very least, a most careful and detailed thinker and expositor; in the case of the *Alphabet* a great many vivid examples are used to reinforce the reader's firm grasp of marginal principles. (The book's only index is indeed an index of examples.) As in his earlier reply to G.B. Shaw,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of 1885, and in the subsequent *Co-ordination of the Laws of Distribution*, of 1894, Wicksteed emphasized the importance of the mathematical expression of marginal economic theory.

For Wicksteed, the theory of value – or 'worth' – means essentially the theory of demand (the theory of supply he refers to as that of production – or 'making'). In both the discussion of 'individual worth' (pp. 1–67) and that of 'social worth' (pp. 68–138), stress is firmly laid on the distinction between total and marginal utility. (Wicksteed uses the latter term and avoids Jevons's 'final utility' and 'final degree of utility'.) While the analysis is based on utility rather than on choice or preference – and 'hedonistic value' is referred to (p. 54) – Wicksteed's later stress on choice between satisfactions which are rendered comparable at the margin is already foreshadowed in the *Alphabet*. It is suggested that all marginal utilities and disutilities, for an individual, may be measured in terms of the hedonistic value, to that individual, of foot-pounds of lifting work or perhaps of one hour of correcting examination papers. Although the exposition is elementary throughout the book, the careful reader will notice Wicksteed's remarks on indivisible commodities and marginal analysis, on the acquiring of preferences, on minimum perceived differences, on traditions and habits, on the desire to impress or to give to others, and on negative marginal (and even total) utilities.

Turning to 'social worth', Wicksteed asserts at once that interpersonal comparisons of utility are impossible; all that can be said is that the ratio of the marginal utilities of any two commodities is the same for any two individuals who possess some of each commodity. (Wicksteed gives a particularly clear account of why this proportionality rule does not hold for an individual whose possession of one or both of the commodities is zero.) Yet he is still ready to argue, on grounds of 'averages' and probabilities, that a more equal distribution of income will probably make the objective social scale of relative prices a more reliable guide to the relative social importance, at the margin, of the various commodities. Wicksteed then discusses the market demand curve, the law of indifference (that is, of one price) and various kinds of price discrimination.

As indicated above, Wicksteed considers that 'Strictly speaking [the allocation of productive resources] does not come within the scope of our present inquiry' (p. 109) but he nevertheless devotes pages 109–24 to the allocation of 'the labour (and other efforts or sacrifices, if there are any others) needful to production' (p. 109). As in the *To-Day* essay of 1884, he argues that the relative prices of freely reproducible commodities will, in equilibrium, be equal to their relative effort-and-sacrifice costs but that this is *not* because production costs give commodities their exchange value. Rather it is because resources are reallocated until the commodities are produced in those quantities for which the marginal utilities – which *are* the sources of exchange value – will be proportional to the constant costs. Given that Wicksteed argues here in terms of 'a unit of effort-and- sacrifice' or 'a unit of productive force' (p. 112 and n.), it is not surprising that no theory of distribution is offered or, indeed, even hinted at.

Co-ordination of the Laws of Distribution

Wicksteed's *QJE* article of the following year, 1889, nevertheless contained an important passage criticizing and extending Jevons's marginal productivity theory of the interest rate, and distribution theory became more prominent in Wicksteed's lectures in the following years. This development culminated with the publication, in 1894, of his famous *Essay on the Co-ordination of the Laws of Distribution*. A number of writers in the early 1890s began to extend the marginal theory of intensive rent into a more general theory of distribution but it was Wicksteed's *Essay* which most clarified the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
issues involved. He noted that the traditional exposition of intensive rent theory, in which varying amounts of ‘capital-and-labour’ were applied to a fixed amount of land, had two crucial properties. First, that the argument essentially concerned only the *proportions* between inputs, and not their absolute levels, and second that the whole argument was *reversible* – the logic is quite unchanged if varying amounts of land are applied to a fixed quantity of ‘capital-and-labour’. It was thus a mere matter of historical accident, Wicksteed argued, that the conventional diagram made one factor return appear as a ‘marginal product’ and the other as a ‘surplus’.

Having argued that it was in any case self-evident that a profit-maximizing entrepreneur would hire each input up to the point at which its marginal value product equalled its (given) price, Wicksteed set himself the task of demonstrating that marginal product pricing of all inputs would entail product exhaustion. (He did not show that there would be any objection in principle to a theory in which *one* return was determined residually – nor could he have done so.) This he did by a long and inelegant mathematical argument, which amounts to no more (and no less) than a proof of Euler's Theorem for homogeneous functions, in the two-variable case. (As was quickly pointed out by Flux in a review in the *Economic Journal* for June 1894; there is some evidence to suggest that Wicksteed was completely unaware of Euler's Theorem before reading Flux's review.) More interesting than the inelegance of Wicksteed's proof, however, is that he was not satisfied with the argument, for while he considered it to be a ‘truism’ that there are constant returns to scale in physical production, he insisted that there might well not be constant returns in terms of revenue. Even if such ‘commercial’ factors as ‘goodwill’, ‘travelling’ and ‘notoriety’ could be increased in the same proportion as all the inputs to physical production, he argued, total revenue might increase in a smaller proportion. Wicksteed was thus led first to consider a monopolist (and to present quite explicitly the marginal revenue formula – already known to Cournot – of the imperfect competition theory of some 40 years later) and then to show how, as the number of firms in an industry becomes ever larger, the product exhaustion theorem will become ‘virtually’ correct. In his later review of Pareto's *Manuale* (1906) and in *The Common Sense* (1910, p. 373, n. 1) Wicksteed appeared to withdraw the sixth section of the *Essay* dealing with product exhaustion in the presence of monopoly, and so on (although not the *Essay* as a whole) but there has been considerable discussion of just how that apparent withdrawal ought to be interpreted.

Wicksteed's *Essay* constituted a major contribution to marginal productivity theory, by raising and discussing the product exhaustion question and by setting the theory very firmly in a multi-product, multi-input setting. (The practice of treating capital, or ‘capital-and-labour’, as a single sum of value is sharply criticized.) It is to be clearly noted, however, that the *Essay* presented *partial* equilibrium analysis throughout; Wicksteed always takes input prices as given and, contrary to some commentators, he never asserts that input supplies are exogenously determined. The *Essay* is a major text in partial analysis; it does not present a general equilibrium argument.

The Common Sense

From 1894 to 1910 Wicksteed published very little in the field of economics but in 1906 he was ready to begin work on his magnum opus *The Common Sense of Political Economy*, published in 1910. In this 700-page book, he sought to expound in minute detail the consequences of ‘the revolution that has taken place’ (p. 2) in economic theory. Disclaiming originality yet again, as he had done in 1888 and in 1894, and making very few *explicit* references to the work of others, Wicksteed presented a consistently

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
subjective approach to all aspects of economic life. (Just five years earlier, in the *Economic Journal* , 1905, p. 435, he had written that ‘The school of economists of which Professor Marshall is the illustrious head may be regarded from the point of view of the thorough-going Jevonian as a school of apologists.’) Ranging from behaviour at the dining table to the significance of the division of labour in an advanced society, Wicksteed argued that attention to *selection between alternatives* was the key to understanding *all* aspects of allocation – whether of bread, of bricks, of friendship, of charity, of labour time or of prayers. Indeed, he even saw an intimate connection between careful marginal allocations and ‘the law formulated by Aristotle with reference to virtue’, that of the mean. The following discussion of Wicksteed's long, immensely detailed and occasionally prolix work will have to centre on his positive contributions and no reference will be made to weaker parts of his analysis (for example, that on increasing and diminishing returns in Book II, Chapter 5) or to his discussion of distribution theory, already referred to above in relation to the *Essay* of 1894. (Wicksteed's famous ‘Scope and Method’ paper, of 1914, presents an incisive epitome of the central themes of the *Common Sense* and may serve as an introduction to it.)

Wicksteed's analysis of choice, in the *Common Sense*, is firmly based on the concept of a scale of preferences, diminishing marginal significance and equivalence at the margin; it has been freed from the notions of utility and marginal utility as quantities, which are still evident in the earlier *Alphabet*. Moreover, while there is some room for doubt, in the *Alphabet*, whether the ‘marginal utility’ of a commodity depends only on the quantity of that commodity or on the quantities of all the commodities possessed, it is completely clear, in the *Common Sense*, that the ‘marginal significance’ of a quantity of a particular commodity depends on all the quantities in question. Indeed it depends not only on all those quantities but on all the circumstances of the choosing individual, for Wicksteed is insistent throughout that *all* objects of choice, and not just marketable commodities, have a bearing on each choice. The principles at work in the allocation of money between potatoes and milk are the same as those involved in the allocation of time between friendship and prayer: ‘whatever our definition of Economics and the economic life may be, the laws which they exhibit and obey are not peculiar to themselves, but are laws of life in its widest extent’ (p. 160). Wicksteed's firm refusal to draw boundaries is more readily understood when account is taken of his conviction that ‘these things, of which money gives us command, are, strictly speaking, never the ultimate objects of deliberate desire at all … as soon as we deliberately desire possession of any external object, it is because of the experiences or the mental states and habits which it is expected to produce or to avert’ (p. 152). In modern terms, the underlying preference ordering is over mental experiences, not over commodities, and there is no reason to expect that ‘economic’ choices will fall under different principles than do ‘other’ choices.

The individual's preference ordering, Wicksteed argues, will be complete but will not always be consistent (transitive), although reflection will increase its consistency. The ordering often will not be, and will not need to be, fully present to the agent's consciousness. Apparently ‘irrational’ behaviour based on impulse, habit or tradition certainly occurs but does not undermine the fundamental principles of rational behaviour; ‘Habit or impulse perpetually determines our selection between alternatives … But if [the terms on which alternatives are offered us] are altered beyond a certain point the habit will be broken or the unconscious impulse checked’ (pp. 28–9). Expectations, uncertainty and consumption loans are all discussed by Wicksteed, as is the fact that rational administration of one's resources is itself costly, in terms of time and effort, and thus should not be pursued beyond a certain point. Throughout his analysis of choice between alternatives Wicksteed returns repeatedly to the idea that the most

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
heterogeneous of satisfactions not only can be but actually are compared at the margin. He is thus led to consider how this analysis can represent ‘the martyr who has borne the rack [and] is ready to be burnt to death sooner than depart a hair's breadth from the formula of his confession’ (p. 404) or the man for whom there are ‘certain things which he would not do for any amount of money, however large’ (p. 405). Wicksteed's answer, in terms of all other considerations falling below a *minimum sensible* in such cases, appears to do little more than provide a polite reconciliation between his equality of marginal satisfactions and the presence of a *lexicographic* priority of honour over money, or of keeping the faith over escaping torture. Indeed, it is not clear how Wicksteed could maintain his own insistence that ethical considerations have priority over others (pp. 123–4), without allowing for at least some element of lexical ordering of alternatives. That said, Wicksteed's many subtle illustrations of how often disparate satisfactions *are* compared and equated at the margin remain highly instructive.

That Wicksteed pursued to the limit the concept of the rational maximizing individual is far from meaning that he had an asocial or ‘atomistic’ view of individual agents, or that he subscribed to the methodological fiction of the ‘economic man’. On the contrary, his most important contribution to marginal theory perhaps lies in his forceful rejection of the ‘economic man’ concept and his closely related demonstration that the marginal analysis of individual action is entirely compatible with the recognition of the intrinsically social nature of many, even most, of the individual agent's purposes and concerns. Whilst the whole of the *Common Sense* contributes powerfully to this ‘double’ argument, it is in Book I, Chapter 5, ‘Business and the Economic Nexus’, that these issues are confronted most directly. ‘But when we pass … to the phrase “the economic motive” … we are in the presence of one of the most dangerous and indeed disastrous confusions that obstruct the progress of Economics’ (p. 163), Wicksteed argues, for there can be no non-arbitrary way of distinguishing motives and considerations which do influence economic actions from those which do not. There are thus two coherent alternatives; ‘We may either ignore motives altogether, or may recognise all motives that are at work, according to the aspect of the matter with which we are concerned at the moment; but in no case may we pick and choose between the motives we will and the motives we will not recognise as affecting economic conditions’ (p. 165). (In fact Wicksteed very seldom adopts the former, external or behaviouristic analysis, even if there is one passage, p. 34, which strongly evokes the later ‘revealed preference’ approach.) If all motives are to be considered by the economic theorist, it follows, of course, that ‘The proposal to exclude “benevolent” or “altruistic” motives from consideration in the study of Economics is … wholly irrelevant and beside the mark’ (p. 179); the interests which an agent seeks to pursue may or may not be directly his own. (And motivations can very well be mixed.)

But if all motives are to be taken into account, and if the principles guiding economic activity are simply the principles guiding all human activity, what defines the particular object of study of the economist? For Wicksteed, the answer lies in the concept of ‘*economic relations*’; ‘economic investigation is concerned [with] the things a man can give to or do for another independently of any personal and individualised sympathy with him or with his motives or reasons’ (pp. 4–5). When persons A and B stand in an economic relation to one another, they may well be furthering each other's purposes in fact but A enters the relation with no thought or intention of promoting B's ends and B, likewise, is motivated by no desire to further the purposes of A; however rich and complex may be the motivations of A and of B, the economic relation between them is an impersonal one. ‘The economic relation does not exclude from my mind every one but me, it potentially includes every one but you’ (p. 174). To stress this point Wicksteed introduced the term ‘non-tuism’, which serves to focus attention upon the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
fact that, in an economic relation, A's lack of concern for the purposes of B (and vice versa), by no means entails that A acts from selfish motives. ‘The specific characteristic of an economic relation is not its “egoism” but its “non-tuism”•’ (p. 180).

With respect to the ‘supply side’ – a term which he might well have rejected – Wicksteed's central contributions lay in his stress on the conception of costs as opportunity costs and in his related views on reservation price and the supply curve as a ‘reverse’ demand curve. Wicksteed laid considerable emphasis on the idea that, no matter how indispensable productive inputs might be, ‘within limits, the most apparently unlike of these factors of production can be substituted for each other at the margins’ (p. 361). (Although it is noteworthy that, in the *Essay* of 1894, he had explicitly drawn attention to the possibility of completely dispensable inputs, p. 37, n. 1.) This emphasis no doubt facilitated – but did not, of course, entail – his insistence on the opportunity costs view of cost of production. ‘Cost of production’, he wrote, ‘is simply and solely “the marginal significance of something else”’ (p. 382) or, less abstractly, ‘By cost of production, or cost price, when the phrase is used without qualification, I mean the estimated value, measured in gold, of all the alternatives that have been sacrificed in order to place a unit of the commodity in question upon the market’ (p. 385). As he had done in 1884 and 1888, Wicksteed argued that ‘there is a constant tendency to equality between price and cost of production, but not because the latter determines the former’ (p. 358). The central thrust of the opportunity cost doctrine was thus directed against the ‘real cost’ doctrines. In his 1905 attack on the ‘apologetic’ school headed by Professor Marshall (referred to above), Wicksteed had written that ‘To scholars of this school the admission into the science of the renovated study of consumption leaves the study of production comparatively unaffected. As a determining factor of normal prices, cost of production is co-ordinate with the schedule of demands registered on the “demand curve”‘. His conclusion in 1910 was more explicit: ‘The only sense, then, in which cost of production can affect the value of one thing is the sense in which it is itself the value of another thing. Thus what has been variously termed utility, ophelemity, or desiredness, is the sole and ultimate determinant of all exchange values’ (p. 391). This was naturally a striking and challenging conclusion but Wicksteed did not give adequate consideration to the implications for the opportunity cost doctrine of limitations to factor mobility or of the presence of non-pecuniary benefits. (See the entry reservation price and reservation demand for further discussion of Wicksteed's ‘rejection’ of the supply curve.)

If Wicksteed's *Common Sense* is not flawless, it remains a brilliant demonstration that a writer who had a strongly ‘social’ conception of the individual agent, who was friendly to the socialist and labour movements of his time, and who was sometimes a sharp critic of the market system, could yet be a purist of marginal theory.

Selected works

1884. *Das Kapital*: a criticism. *To-Day* NS 2, October. Reprinted in *Common Sense* (1933).

1888. *The Alphabet of Economic Science*. Part I: *Elements of the Theory of Value or Worth*. London: Macmillan. Reprinted, New York: Augustus Kelley, 1955.

1889. On certain passages in Jevons's Theory of Political Economy. *Quarterly Journal of Economics*,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

April. Reprinted in *Common Sense* (1933).

1894a. *An Essay on the Co-ordination of the Laws of Distribution*. London: Macmillan. Reprinted as No. 12 in LSE Series of Reprints of Scarce Tracts in Economic and Political Science, 1932. Reprinted, ed. I. Steedman, London: Duckworth, 1987.

1894b. Degree of utility; Dimensions of economic quantities; Final degree of utility; W.S. Jevons; Political economy and psychology. In *Palgrave's Dictionary of Political Economy*. London: Macmillan. All reprinted in *Common Sense* (1933).

1905. Note on Jevons's economic work. *Economic Journal* 15, September. Reprinted in *Common Sense* (1933).

1906. Review of Pareto's *Manuale di Economia Politica*. *Economic Journal* 16, December. Reprinted in *Common Sense* (1933).

1910. *The Common Sense of Political Economy*. London: Macmillan. 2nd edn, 2 vols., ed. L.M Robbins, London: Routledge & Kegan Paul, 1933.

1914. The scope and method of political economy in the light of the ‘marginal’ theory of value and distribution. *Economic Journal* 24, March, 1–23. Reprinted in *Common Sense* (1933).

1925. Final utility. In *Palgrave's Dictionary of Political Economy*, 2nd, edn., ed. H. Higgs. London: Macmillan. Reprinted in *Common Sense* (1933).

Bibliography

Encyclopaedia of the Social Sciences, vol. 15, ed. E.R.A. Seligman and A. Johnson. London: Macmillan, 1935.

Flux, A.W. 1894. Review of Wicksell's Über Wert, *Kapital und Rente* and of Wicksteed's Essay. *Economic Journal* 4, June, 305–13.

Herford, C.H. 1931. *Philip Henry Wicksteed: His Life and Work*. London: Dent & Sons.

Hutchison, T.W. 1953. *A Review of Economic Doctrines, 1870–1929*. Oxford: Clarendon Press.

International Encyclopedia of the Social Sciences, vol. 16, ed. D.L. Sills, London and Glencoe: Macmillan and the Free Press, 1968.

Robbins, L. 1933. Editorial Introduction to Wicksteed (1910).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Steedman, I. 1987. Editorial Introduction to Wicksteed (1894a).

Stigler, G.J. 1941. *Production and Distribution Theories: The Formative Period*. New York: Agathon Press, 1968.

Sturges, R.P. 1975. *Economists' Papers 1750–1950*. London: Macmillan.

How to cite this article

Steedman, Ian. "Wicksteed, Philip Henry (1844–1927)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000070>
doi:10.1057/9780230226203.1834

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wiener process

A.G. Malliaris

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Brownian motion is the most renowned, and historically the first stochastic process that was thoroughly investigated. It is named after the English botanist, Robert Brown, who in 1827 observed that small particles immersed in a liquid exhibited ceaseless irregular motion. Brown himself mentions several precursors starting at the beginning with Leeuwenhoek (1632–1723). In 1905 Einstein, unaware of the existence of earlier investigations about Brownian motion, obtained a mathematical derivation of this process from the laws of physics. The theory of Brownian motion was further developed by several distinguished mathematical physicists until Norbert Wiener gave it a rigorous mathematical formulation in his 1918 dissertation and in later papers. This is why the Brownian motion is also called the Wiener process. For a brief history of the scientific developments of the process see Nelson (1967).

Keywords

Bachelier, L.; Brownian motion: *see* Wiener process; geometric Wiener process; stochastic calculus; uncertainty; Wiener process

Article

Having made these remarks we now define the process. A *Wiener process* or a *Brownian motion process*

$$\{Z(t, \omega): [0, \infty] \times \Omega \to R\}$$

is a stochastic process with index $t \in [0, \infty]$ on a probability space Ω and mapping to the real line *R*, with the following properties:

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1. (1) $Z(0, \omega) = 0$ with probability 1, that is by convention we assume that the process starts at zero.
2. (2) $0 \leq t_0 \leq t_1 \leq \ldots \leq t_n$ are time points then for any real set H_i

$$P[Z(t_i) - Z(t_{i-1}) \in H_i \text{ for } i \leq n] = \prod_{i \leq n} P[Z(t_i) - Z(t_{i-1}) \in H_i].$$

   This means that the increments of the process $Z(t_i) - Z(t_{i-1}), i \leq n$, are independent variables.
3. (3) For $0 \leq s < t$ the increment $Z(t)$–$Z(s)$ has distribution

$$P[Z(t) - Z(s) \in H] = \left(1 / \sqrt{2\pi(t-s)}\right) \int_H \exp\left[-x^2 / 2(t-s)\right] dx.$$

   This means that every increment $Z(t)$–$Z(s)$ is normally distributed with mean zero and variance (t –s).
4. (4) For each $\omega \in \Omega$, $Z(t, \omega)$ is continuous in t, for $t \geq 0$.

Note that condition (4) can be proved mathematically using the first three conditions. Here it is added because in many applications such continuity is essential. Although the sample paths of the Wiener process are continuous, we immediately state an important theorem about their differentiability properties.

Theorem: (non-differentiability of the Wiener process) Let $\{Z(t), t \geq 0\}$ be a Wiener process in a given probability space. Then for ω outside some set of probability 0, the sample path $Z(t, \omega), t \geq 0$ is nowhere differentiable.

Intuitively, a nowhere differentiable sample path represents the motion of a particle which at no time has a velocity. Thus, although the sample paths are continuous, this theorem suggests that they are very kinky, and their derivatives exist nowhere. The mathematical theory of the Wiener process is presented rigorously in Billingsley (1999, ch. 37) and more extensively in Knight (1981).

The first application of Brownian motion or the Wiener process in economics was made by Louis Bachelier in his dissertation ‘Théorie de la spéculation’ in 1900. Cootner (1964) collects several papers and cites additional references on the application of the Wiener process in describing the random character of the stock market. In the early 1970s Merton, in a series of papers, established the use of stochastic calculus as a tool in financial economics. The Wiener process is a basic concept in stochastic calculus and its applicability in economics arises from the fact that the Wiener process can be regarded as the limit of a continuous time random *walk* as step sizes become infinitesimally small. In other words, the Wiener process can be used as the cornerstone in modelling *economic uncertainty* in continuous time. For purposes of illustration consider the stochastic differential equation

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

$$dX(t) = \mu(t, x)dt + \sigma(t, x)dZ(t) \quad (1)$$

which appears in the economic literature describing asset prices, rate of inflation, quantity of money or other variables. In (1), changes in the variable $X(t)$, denoted as dX(t), are described as a sum of two terms: $\mu(t, x)$ which is the expected instantaneous change and $\sigma(t, x)\, dZ(t)$ which is the unexpected change. Furthermore, this unexpected change is the product of the instantaneous standard deviation $\sigma(t, x)$ and uncertainty modelled by increments in the Wiener process. See Merton (1990, ch. 3) for a methodological essay on continuous-time modelling, and Malliaris and Brock (1982) or Chang (2004, ch. 2) for numerous applications of the Wiener process in economics and finance.

Economists have constructed various processes based on the Wiener process. Let $\{Z(t), t \geq 0\}$ be a Wiener process and use it to construct a process $\{W(t), t \geq 0\}$ defined by $W(t) = Z(t) + \mu t, t \geq 0$ where μ is a constant. Then we say that $\{W(t), t \geq 0\}$ is a *Wiener process* or *Brownian motion process with drift* and μ is called the drift parameter. In this case the only modification that occurs in the definition of a Wiener process is in property (3) where $W(t) - W(s)$ is normally distributed with mean $\mu(t - s)$ and variance $(t - s)$. Finally, let *W*(*t*) be a Wiener process with drift as just defined. Consider the new process given by $Y(t) = \exp[W(t)], t \geq 0$. Then $\{Y(t), t \geq 0\}$ is called a *geometric Brownian motion* or *geometric Wiener process*.

The availability of an extensive mathematical literature on the Wiener process and the economists' fundamental goal to model economic uncertainty in continuous time suggest that this process will continue to be an important tool for economic theorists.

See Also

- martingales
- uncertainty

Bibliography

Billingsley, P. 1999. *Probability and Measure*. 3rd edn, New York: Wiley.

Chang, F.R. 2004. *Stochastic Optimization in Continuous Time*. New York: Cambridge University Press.

Cootner, P.H. 1964. *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press.

Knight, F.B. 1981. *Essentials of Brownian Motion and Diffusion*. Mathematical Surveys, Number 18. Providence, RI: American Mathematical Society.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Malliaris, A.G. and Brock, W.A. 1982. *Stochastic Methods in Economics and Finance*. Amsterdam: North-Holland Publishing Company.

Merton, R.C. 1990. *Continuous-Time Finance*. Malden, MA: Blackwell.

Nelson, E. 1967. *Dynamical Theories of Brownian Motion*. Princeton, NJ: Princeton University Press.

How to cite this article

Malliaris, A. G. "Wiener process." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000074> doi:10.1057/9780230226203.1835

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wieser, Friedrich Freiherr, (Baron) von (1851 – 1926)

E. Streissler

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Austrian economics; Austrian School; Böhm-Bawerk, E. von; countervailing power; efficient production; equimarginal principle in production; imputation; innovation; marginal utility; Menger, C.; Mises, L. E. von; natural value; opportunity cost; Schumpeter, J. A.; Wieser, F. F. von

Article

Wieser is commonly cited together with his senior, Carl Menger, and his exact contemporary, Eugen Böhm von Bawerk, as one of the founding trio of the Austrian School of Economics in the last quarter of the 19th century. The exact nature of his achievement, however, seems now practically forgotten: possibly because he produced an intractable mixture of deep and influential insights, very distinctly his own, intermingled with oratorical prose and often unpalatable value judgements; he was extremely successful in his own generation but appeared outdated in his attitudes half a century later.

Wieser was born on 10 July 1851, in Vienna. His father was Commissary-General of the Austrian army in the war of 1859, for which service he was ennobled, later becoming Vice President of the Austrian Court of Audit, a baron and a privy councillor (Geheimrat). But this high social status was only acquired after Friedrich Wieser's birth and very little money went with it so that the family lived in modest circumstances. Wieser went to the Benedictine Schottengymnasium in Vienna, one of the city's three elite schools. His classmate was Eugen Böhm von Bawerk, who became his close friend and brother-in-law. Together the two studied at Vienna University law faculty (which included courses in economics), together they entered the civil service in the fiscal division, and together they went on a two-year leave of absence to perfect themselves in economics at Heidelberg, Leipzig and Jena, with Knies, Roscher and Hildebrand. A little after Böhm, Wieser passed his ‘Habilitation’ in economics with Menger in 1883, was appointed associate professor in 1884 and full professor in 1889 at the University of Prague and was that university's Vice Chancellor in 1901–2. In 1903, he succeeded Menger in the chair of economic theory at Vienna University law faculty on the latter's early retirement, Böhm again joining him only a year later as extraordinarily appointed additional full professor. Böhm, Menger, and finally also Wieser served as members of the Austrian House of Lords (Herrenhaus). Wieser became Minister of Commerce

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
in 1917, holding this office up to the end of the monarchy in 1918. He died on 23 July 1926.
Apart from the short ministerial interlude Wieser thus taught from 1884 to 1926 at the largest universities of Austria. Basically, he must be considered the most successful teacher (especially of undergraduates) and the orator of the trio. His influence was pervasive through his lecturing to tens of thousands of law students, many of whom he examined in person, and at second and third remove on even vaster numbers in the intellectual melting pot of Vienna. In a true oral tradition Wieser influenced present-day economics through what these frequently very young students – an appreciable percentage of whom later became themselves important in Western intellectual life – picked up in a kind of intellectual osmosis, mostly without realizing it and therefore usually without attributing the ideas to their teacher.
Wieser's main works are: his thesis of ‘Habilitation’ (Wieser, 1884), which encompasses a large part of his original thought, particularly the marginal productivity valuation of factors of production and his cost theory; Wieser (1889), mainly an elaboration of the former together with an attempt to give the marginal utility concept normative distributional content; Wieser (1914), the definitive textbook of the Austrian School and (with its rival, G. Cassel's more up-to-date *Theoretische Sozialökonomie*, 1918) one of the two main theoretical textbooks in German of the early interwar period, a book still worth reading, especially the less well-known institutional chapters on large corporations and money and banking; finally Wieser (1926), a socio-philosophical tract in abject adulation of power (power being justified by mere ‘success’), whose lack of judgement can only be justified by the effect of the total breakdown of all established social and political order after the First World War on Wieser's own moral fibre.
Wieser prided himself on the invention of telling phrases, particularly the term ‘Grenznutzen’ in Wieser (1884), wherefore Marshall credits him, perhaps unjustly, with originating the term ‘marginal utility’. During his leadership the Austrian School had to sail under the flag ‘Grenznutzenschule’. In contrast to the purely analytic usage by the other members of the trio, ‘Grenznutzen’ had for Wieser a near mystic connotation and certainly normative content: more precisely, it is the average marginal utility in a competitive society with equality of incomes which is the ‘natural value’ of Wieser (1889). ‘Grenznutzen’ thus served Wieser, who was (unusual for an Austrian economist) a paternalistic interventionist, as a yardstick of policy evaluation.
In contrast to Menger and Böhm, Wieser was not a clear logical analyst but had influential visions. He was clearest in his cost and production theory, frequently being credited with introducing the opportunity cost principle that all costs are only utilities forgone, though Wieser's actual advance over Menger appears slight. Wieser certainly, however, gave what appears to be the first account of the principles of efficient production, which Menger had ignored (Wieser, 1884). Production is undertaken in expectation of the price the marginal valuation of consumers will allow, Wieser (1884) first formulating the equimarginal principle in production: the marginal product of each factor (or its cost) must be the same in all its different uses and as high as the least important marginal utility achievable from its given supply (Wieser's Law of Cost). In Wieser (1914) this is extended (contemporaneously with Wicksteed, *Common Sense of Political Economy*, 1910) to an analysis of differential quality rents on the lines of Ricardo: any more efficient factor earns as rent the additional value added over the least efficient equivalent factor. Some of his insights into capital and efficient production Wieser probably owes, as the terminology suggests, to Marx (Wieser never gives his authorities, apart from sparse references in Wieser, 1914): for example that the value of factors of production must reflect the socially necessary cost of production (the use of the best generally known technique); or that innovation brings

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
extra profits to the innovator, without changing the value of the (other) factors. (Marx, Engels, Ricardo, Jevons and Menger are the five authors Wieser acknowledges as inspiration in the Introduction to Wieser, 1884.) As to distribution, Wieser first posed the problem whether the marginal product reward of all factors of production would exactly exhaust the product (‘Zurechnungsproblem’, the problem of imputation), without being, despite many attempts, able to solve it.

These ideas were, however, all on the point of being discovered by others. Uniquely his own is the repeated stress, already in Wieser (1884), of the paramount importance of economic calculation and the need to have an economic measuring rod for all rational ‘planning’ for the future. (One is tempted to suspect this to be an obsession of the son of the Vice President of the Court of Audit.) The measuring rod for Wieser is marginal utility in its wide sense; but it was a small step, taken by Mises and Hayek, to make out of this need for a measuring rod in all economic planning the concept of the informative nature of prices. Economics may even owe the (then uncommon) term ‘planning’ for the rational activity in economics on the individual as well as the societal level to Wieser via Mises. Wieser already stated repeatedly that even a socialist economy would have to use the same economic measuring rod and basically the same principles of ‘planning’ as a capitalist one: out of which Mises developed the idea that, lacking prices, a socialist society could not plan rationally.

Besides his production and distribution-oriented ideas Wieser had a second influential vision: the importance of the creative individual in all economic processes. He felt deeply the basically contradictory nature of his two visions, the impersonal mass effects of efficient production on the one hand, an idea which he curiously traced to the influence of Herbert Spencer, and the elitist idea of the effects of the outstanding individual, which he attributed to the hero-worshipping teaching of history in the Schottengymnasium. In this vein, which he cultivated in his later years, he was again, above all, influential through the forceful and suggestive use of words, the terms ‘Führer’, ‘Pionier’, ‘Neuerung’ (German for innovation) being of his creation. Schumpeter adopted virtually all the terminology for his *Theory of Economic Development* (1912) from his acknowledged teacher Wieser and also the idea that economic ‘dynamics’ (in contrast to statics) is due to individual leadership activity. Wieser himself had developed relatively few concrete conclusions out of his leadership rhetoric, apart from remarks about the countervailing power of trade unions and the administrative and even innovative efficiency of large corporations in Wieser (1914), an idea taken up by Schumpeter only later. Wieser's second vision degenerated into the lurid prose of Wieser (1926), where, for example, the ‘Führer’ (a pet word of Wieser's), Adolf Hitler, is chided (in 1926!) for not quite making the grade. For Wieser, again in sharp contrast to the staunch liberal principles of Menger and Böhm, tended, in spite of his basic Catholic-conservative outlook, to flirt with any social movement that was new and appeared ‘great’, making commendatory references to socialism in his youth and to German nationalism and fascism in his old age.

See Also

- Austrian economics

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1884. *Über den Ursprung und die Hauptgestze des wirtschaftlichen Werthes*. Vienna: Hölder.

1889. *Der Natürliche Werth*. Vienna: Hölder. Trans. as *Natural Value*, ed. W. Smart, London: Macmillan, 1893. Reprinted, New York: Kelley & Millman, 1956.

1914. Theorie der gesellschaftlichen Wirtschaft. In *Grundriss der Sozialökonomik*, vols 1 and 2. Tübingen: Mohr-Siebeck. 2nd edn, 1923. Trans. as *Social Economics*, New York: Greenberg, 1927.

1926. *Das Gesetz der Macht*. Vienna: Springer.

Bibliography

Mayer, H. 1929. Friedrich Freiherr von Wieser. In *Neue österreichische Biographie*, vol. 6. Vienna: Amalthea.

Schumpeter, J.A. 1912. *The Theory of Economic Development. An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Trans. R. Opie, Cambridge, MA: Harvard University Press, 1934.

Stigler, G.J. 1941. *Production and Distribution Theories: The Formative Period*. New York: Macmillan.

Streissler, E.W. 1986. Arma virumque cano – Friedrich von Wieser, the bard as economist. In *Die Wiener Schule der Nationalökonomie*, ed. N. Leser. Vienna: Hermann Böhlau.

How to cite this article

Streissler, E. "Wieser, Friedrich Freiherr, (Baron) von (1851–1926)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000075> doi:10.1057/9780230226203.1836

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wilson, Edwin Bidwell (1879 – 1964)

Murray Milgate
From The New Palgrave: A Dictionary of Economics, First Edition, 1987
Edited by John Eatwell , Murray Milgate and Peter Newman

Article

While primarily a mathematician, Wilson's relatively few contributions to economics in the interwar years, particularly two short papers on demand theory (1935 and 1939) and one on business cycles (1934), were not without their influence in Harvard economic circles of the day. Schumpeter drew on the arguments of Wilson's paper on the periodicity of US business cycles in his *Business Cycles*, and Samuelson's *Foundations* contains an acknowledgement to Wilson (with Schumpter and Leontief) in its preface, and credits him with the suggestion of utilizing the Le Chatelier principle in economic analysis. The essay on cyclical fluctuations in business activity was an attempt to make deeper analytical use of the monthly index of US business activity prepared by Leonard Ayers and published in 1931. Using the device of the ‘periodogram’, invented by Arthur Schuster, Wilson is able to extract from Ayers' data ‘hidden’ cycles of different periodicities. The idea that behind any given aggregative series there might lurk different patterns of cyclical movement was, no doubt, a spur to Schumpeter's consideration of the simultaneous operation of Juglar, Kitchin and Kondratieff cycles in *Business Cycles*.

The two short essays on demand theory (1935 and 1939) are concerned with the derivation of the law of demand – that is, the inverse relationship between price and quantity demanded. The first generalizes Pareto's proof of the proposition, which had assumed additively separable utility functions. Wilson assumes instead only that $U_i(x_1 \ldots x_n)$, may take the form $U_i(x_1)+U_i(x_2 \ldots x_n)$ and derives from this the law of demand. The second paper is designed to show that Marshall's assumption of a constant marginal utility of money gave only a special case of the law of demand, and that the same result could be obtained without it. As Wilson observed, this ‘forces us over from the “index of ophelimity” to a utility definite except for a linear transformation, i.e., except for scale and origin’ (1939, p. 649). The importance of this result, especially given its relation to the Hicks–Allen theory of demand, for the subsequent debates over cardinal versus ordinal utility is readily apparent.

Wilson was born at Hartford, Connecticut, on 25 April 1879. After graduating from Harvard in 1899, he took his Ph.D. from Yale in 1901. From 1907 until 1922 he was on the faculty at MIT, first as professor of mathematics and later as professor of mathematical physics. From that date until his retirement, he was professor of vital statistics at the Harvard School of Public Health. He served as president of the American Statistical Association (1929), was vice-president of the National Academy of Sciences (1949–

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

53), and was an honorary fellow of the Royal Statistical Society. He died on 28 December 1964.

Selected works

1912. *Advanced Calculus*. Boston and New York: Ginn & Co.

1934. The periodogram and American business activity. *Quarterly Journal of Economics* 48, May, 375–417.

1935. Generalization of Pareto's demand theorem. *Quarterly Journal of Economics* 49, August, 715–17.

1939. Pareto versus Marshall. *Quarterly Journal of Economics* 53, August, 645–50.

How to cite this article

Milgate, Murray. "Wilson, Edwin Bidwell (1879–1964) ." The New Palgrave Dictionary of Economics. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000048>
doi:10.1057/9780230226203.1890

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Witte, Edwin Emil (1887 – 1960)

W. Cohen
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Witte, E. E.; Social Security in the United States

Article

Witte was Professor of Economics at the University of Wisconsin (1933–57), President of the American Economic Association (1955), first President of the Industrial Relations Research Association (1948), and Chief of the Wisconsin Legislative Reference Service (1922–33). His primary field was labour and social legislation. A student of John R. Commons, he was an institutional economist and a pragmatic social reformer.

His outstanding contribution was his significant role as Executive Director of President Franklin D. Roosevelt's Cabinet Committee on Economic Security (1934–5) which drafted the legislation that became the Social Security Act of 1935. Witte prepared the Committee's report and recommendations to the President. His *The Development of the Social Security Act* (1936) recounting the legislative history of the Act is an outstanding model of its type.

Witte published *The Government in Labor Disputes* (1932), assisted in the formulation of the Norris–LaGuardia Act (1932) restricting injunctions in labour disputes and he was Regional Director of the War Labour Board in Detroit (1942–5).

Witte received his Ph.D. in economics from the University of Wisconsin (1927) and was Chairman of the Department of Economics (1936–41 and 1946–57). Except for temporary assignments, he lived in Wisconsin all his life. He had a practical outlook on economic and political issues, coloured by LaFollette progressive populism.

Selected works

1932. *The Government in Labour Disputes*. New York and London: McGraw-Hill.

1936. *The Development of the Social Security Act: A Memorandum on the History of the Committee on*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Economic Security and the Drafting and Legislative History of the Social Security Act. Washington, DC. Reprinted, with a foreword by F. Perkins, Madison: University of Wisconsin Press, 1962.

Bibliography

[Anon.] 1960. Edwin E. Witte (1887–1960). Father of social security. *Industrial and Labor Relations Review* 14(1), 7–9.

Schlabach, T.F. 1969. *Edwin E. Witte: Cautious Reformer*. Madison: State Historical Society of Wisconsin.

How to cite this article

Cohen, W. "Witte, Edwin Emil (1887–1960)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000088> doi:10.1057/9780230226203.1837

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wold, Herman O.A. (1908)

E.J. Hannan

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Article

Wold was born at Skien in Norway on 25 December 1908 but the family migrated to Sweden in 1912. His university education was at Stockholm, where he took his doctoral degree in 1938, under H.A. Cramér. He became Professor of Statistics at Uppsala in 1942, moved to Gothenburg in 1970, and retired in 1975.

His doctoral thesis (Wold, 1934) dealt with the theory of stationary time series. Two theorems first proved in it are of lasting value. The first is the Wiener–Khintchine relation for a discrete–time series, but probably more important was the Wold Decomposition Theorem, which represents a stationary time series as the sum of an (infinite) moving average of past innovations (linear prediction errors) and a perfectly predictable component. Wold was also influential in time series analysis through his student, P. Whittle.

However, most of Wold's later work has been in econometrics. Membership of a 1938 committee to study consumer demand, rationing in case of war being the motivating force, led him to the study of general economic modelling. His work on consumer demand culminated in Wold and Juréen (1952). Economic modelling, in turn, led him to the work of Tinbergen (1939) on the statistical measurement of business cycles. Tinbergen's model was linear and connected a vector, $y(t)$, of endogenous variables to a vector, $z(t)$, of exogenous variables and lagged endogenous variables by an equation

$$y(t) = By(t) + Gz(t) + \varepsilon(t), \quad E\{\varepsilon(s)\varepsilon(t)'\} = \delta_{s,t}\Omega \tag{1}$$

where the $\varepsilon(t)$ are errors. Wold observed that Tinbergen's equations were recursive in that after a rearrangement of the rows of (1) the matrix B was lower triangular with zeros along the diagonal. If Ω is diagonal then (1) may be validly estimated by least squares and this will be the maximum likelihood method if the $\varepsilon(t)$ are also Gaussian. Wold sought to promote recursive modelling in contrast to the non-

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

causal modelling that became influential in econometrics following Haavelmo (1944). The recursive models are causal since, after the rearrangement of rows, elements of $y(t)$ can be regarded as arranged in a causal hierarchy. Wold's view does not seem to have prevailed. The complexity of economic phenomena, including the nonlinearity of economic behaviour, the poor quality of much data, and the large amount of aggregation, together with autocorrelation of $\varepsilon(t)$, make the issue seem somewhat removed from reality.

Wold (1959) also emphasized the distinction between prediction and structural estimation and has proposed an iterative estimation of (1), oriented towards prediction, where $y(t)$ on the right is replaced by $(I-\hat{B})^{-1}(\hat{C}z(t))$, with $\hat{B}, \hat{C}$ obtained from a previous iterations.

Selected works

1934. *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksell.

1952. (With L. Juréen.) *Demand Analysis: A Study in Econometrics*. New York: John Wiley.

1959. Ends and means ineconometric model building: basic considerations reviewed. In *Probability and Statistics*, The Harald Cramér volume, ed. U. Grenander. Stockholm: Almqvist and Wiksell, 355–434.

1960. *Forecasting on a Scientific Basis*. Lisbon: Gulbenkian Foundation.

1965. *Bibliography of Time Series and Stochastic Processes*. Edinburgh: Oliver & Boyd.

1980. *The Fix-Point Approach to Interdependent Systems*. Amsterdam: North-Holland.

Bibliography

Bentzel, R. and Wold, H. 1946. On statistical demand analysis from the viewpoint of simultaneous equations. *Skand. Aktuarietidskrift* 29, 95–114.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12, Supplement.

Tinbergen, J. 1939. *Statistical Testing of Business Cycle Theories, Vol. II: Business Cycles in the USA. 1919–9132*. Geneva: League of Nations.

How to cite this article

Hannan, E. J. "Wold, Herman O.A. (1908)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000090>
doi:10.1057/9780230226203.1838

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

women's work and wages

Francine D. Blau and Lawrence M. Kahn
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Since the early 1980s the gender wage gap has fallen in most economically advanced countries, although a gender wage differential remains in all countries. We first document for several industrialized countries recent trends in the gender gap in labour force participation and earnings. We then outline several explanations for the gender wage gap at a given point in time, changes in the gender gap over time, and differences in its extent across countries. Next, we consider the empirical evidence in support of various explanations. We conclude with some thoughts about future prospects for the gender wage gap.

Keywords

affirmative action; anti-discrimination law; black-white labour market inequality in the United States; centralized wage-setting; collective bargaining; comparable worth; compensating differentials; education; gender wage gap; human capital; labour force selectivity; labour market discrimination; labour supply; monopsony; occupational segregation; search costs; skill-biased technical change; statistical discrimination; technical change; trade unions; wage differentials; wage distribution; wage structure; women's work and wages

Article

Since the early 1980s the gender wage gap has fallen in most economically advanced countries, although a gender wage differential remains in all countries. While labour market outcomes for men and women may vary across a number of dimensions, economists have particularly focused on analysing gender differences in wages. This emphasis reflects a number of factors. The wage is a major determinant of economic welfare for employed individuals, as well as of the potential gain from market work for those not currently employed. Further, it affects decisions ranging from labour supply to marriage and fertility, as well as bargaining power and relative status within the family.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

This article begins with an overview across a number of economically advanced countries of labour force participation and earnings differences between men and women in the labour market, and delineates the recent trends. We then consider the explanations that have been offered for the gender wage gap at a given point in time as well as for changes in the gender gap over time and differences in its extent across countries. Next, we consider the empirical evidence in support of various explanations. We conclude with some thoughts about future prospects for the gender wage gap.

Overview of gender differences in labour force participation and wages

Although the focus of this article is the gender wage gap, it is useful to consider the evolution of labour force participation rates by gender. Women's rising labour force participation implies that women's wage gains, discussed below, apply to an increasing share of the female population. In addition, changing female participation can mean that the qualifications and experience of the typical employed woman may be changing as well. To the extent that women's rising participation is associated with rising labour force attachment of women over the life cycle, the average level of labour market experience of women will eventually increase. Changing participation rates may also be associated with changes in labour force selectivity of women, depending on the relative qualifications of entrants and incumbents. And these factors in turn may help us explain the evolution of the gender wage gap. Table 1 shows that, across ten economically advanced countries, women's labour force participation rates rose steadily between 1979 and 2000, both absolutely and relative to men, with a faster increase in the 1980s than the 1990s. For example, taking an unweighted average of the countries listed in the table, we see that women were 64 per cent as likely as men to be in the labour force in 1979; by 1990, this ratio had risen to 77 per cent and, by 2000, it was 83 per cent. Throughout this period, Scandinavian women had especially high participation rates.

Labour force participation rates by gender (ages 15–64), ten Western countries, 1979–2000

	1979			1990			2000		
	Men	Women	Ratio: (women/ men)	Men	Women	Ratio: (women/ men)	Men	Women	Ratio: (women/ men)
Australia	87.6	50.3	0.574	84.4	61.5	0.729	81.9	65.4	0.799
Finland	82.2	68.9	0.838	79.6	73.5	0.923	76.4	72.1	0.944
France	82.6	54.2	0.656	75.0	57.2	0.763	74.4	61.7	0.829
Germany	84.5	49.6	0.587	79.0	55.5	0.703	78.9	63.3	0.802
Japan	89.2	54.7	0.613	83.0	57.1	0.688	85.2	59.6	0.700
Netherlands	79.0	33.4	0.423	79.7	52.4	0.657	83.9	65.7	0.783
New Zealand	87.3	45.0	0.515	83.0	63.2	0.761	83.2	67.5	0.811
Sweden	87.9	72.8	0.828	86.7	82.5	0.952	81.2	76.4	0.941
United Kingdom	90.5	58.0	0.641	88.3	67.3	0.762	84.3	68.9	0.817
United States	85.7	58.9	0.687	85.6	67.8	0.792	83.9	70.7	0.843

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Average	85.7	54.6	0.636	82.4	63.8	0.773	81.3	67.1	0.827

Notes: For 1990 and 2000, ages are 16–64 for Sweden, the United Kingdom and the United States. Germany is defined as West Germany in 1979 and 1990 and unified Germany in 2000.

Sources: OECD (1990, p. 200; 2004, pp. 294 and 296).

Table 2 shows female to male pay ratios from the OECD across the same ten countries shown in Table 1. The entries in Table 2 are intended to show the price of labour; in some cases the data are available as weekly or monthly earnings of full-time workers (Panel A), and in others as annual earnings of full-time, year-round workers (Panel B). As may be seen in the table, women uniformly have lower wage rates than men. However, in all but one case, the gap fell between 1980 and 2000. (The exception is Sweden where the pay gap rose by one percentage point from an already low level in 1980.) For example, between 1980 and 2000, the ratio of women's to men's wages rose from an average of 69 to 77 per cent among the countries in Panel A, and from an average of 78 to 83 per cent among countries in Panel B. (US Current Population Survey data indicate that the gender gap in annual pay for full-time, year-round workers is slightly higher than the gender gap in weekly earnings for full-time workers, suggesting that the gap in weekly wages for the countries in Panel B may be even smaller than the figures shown in Table 2.) The gender wage gap is especially small in France, Australia, Sweden and New Zealand. The United States had one of the larger gaps in 1980, with only Japan having a lower female to male wage ratio. Over the 1980–2000 period, the gender wage gap fell by more, both absolutely and relatively, in the United States than in any of the other countries shown. Nonetheless, even in 2000 the gender wage gap remained relatively high in the United States, as other wage differentials such as those for education, cognitive ability or union membership have historically been (Blau and Kahn, 2002; 2005). As we shall see, the pattern of international differences in the gender wage gap is useful in shedding light on the impact of labour market institutions on the gender gap. This is because there are very large differences in the types of such institutions across the economically advanced countries that have consequences for the size of the gender wage gap.

Female to male ratios, median full time earnings: ten Western countries, 1980–2000

	1980	1990	2000	Changes: 1980–2000 Absolute	%
	A. Weekly or monthly earnings, full-time workers				
Australia	0.813	0.818	0.828	0.016	1.9%
W. Germany	0.705	0.738	0.793	0.088	12.5%
Japan	0.583	0.594	0.654	0.071	12.2%
New Zealand	0.733	0.773	0.815	0.082	11.2%
United Kingdom	0.647	0.688	0.761	0.113	17.5%
United States	0.634	0.715	0.748	0.114	17.9%
Average	0.686	0.721	0.766	0.081	11.8%
	B. Annual earnings, full-time, year-round workers				
Finland	0.734	0.771	0.796	0.062	8.4%

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

France	0.803	0.847	0.905	0.103	12.8%
Netherlands	0.744	0.750	0.783	0.039	5.3%
Sweden	0.855	0.804	0.845	–0.010	–1.2%
Average	0.784	0.793	0.832	0.048	6.2%

Notes: Actual years covered are: Australia, Finland, Sweden, United Kingdom and United States: 1980, 1990, 2000; France: 1980, 1990, 1998; W. Germany: 1984, 1990, 1998; Japan: 1980,1990, 1999; Netherlands: 1985, 1990, 1999; New Zealand: 1984, 1990, 1997.

All earnings are gross of taxes, except France which reports net earnings.

Source: OECD Earnings Database.

Explanations for the gender wage gap

Traditionally, economic analyses of the gender wage gap have focused on what might be termed gender-specific factors, that is, (*a*) gender differences in qualifications and (*b*) differences in the labour market treatment of men and women (or labour market discrimination). More recently, following the work of Juhn, Murphy and Pierce (1991) on trends in race differentials, some advances have been made by considering the gender wage gap and other demographic wage differentials in the context of the overall structure of wages. Wage structure is the array of prices determined for labour market skills and the rewards to employment in particular sectors. In addition, gender-specific factors and wage structure can interact to affect the gender wage gap.

Gender-specific factors

Gender differences in qualifications have primarily been analysed within the human capital model (Mincer and Polachek, 1974). Given the traditional division of labour by gender in the family, women tend to accumulate less labour market experience than men. Further, anticipating shorter and more discontinuous work lives, women have lower incentives to invest in market-oriented formal education and on-the-job training. Their resulting smaller human capital investments will lower their earnings relative to those of men. Working in a similar direction is Becker's (1985) model in which the longer hours women spend on housework lower the effort they put into their market jobs compared with men and hence reduce their wages.

Gender differences in occupations are also expected to result if women choose occupations for which on-the-job training is less important. Women may especially avoid jobs requiring large investments in firm-specific skills (that is, skills which are unique to a particular enterprise), because the returns to such investments are reaped only as long as one remains with a particular employer. At the same time, employers may also be reluctant to hire women for such jobs because they bear some of the costs of firm-specific training (see the discussion of statistical discrimination below).

To the extent that gender differences in outcomes are not fully accounted for by productivity differences derived from these and other sources or by compensating differences in non-wage job characteristics such as risk of injury, models of labour market discrimination offer an explanation. In Becker's (1957)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
model, discrimination is due to the discriminatory tastes of employers, co-workers or customers. Alternatively, in models of statistical discrimination (for example, Aigner and Cain, 1977), which assume a world of uncertainty, differences in the treatment of men and women arise from differences between the two groups in employer perceptions of the expected value of productivity or in the reliability with which productivity may be predicted. In either case, gender differences in wages or occupations may result. Another aspect of interest is the relationship between occupational segregation and a discriminatory wage gap formulated in Bergmann's (1974) overcrowding model. She argues that discriminatory exclusion of women from ‘male’ jobs results in an excess supply of labour in ‘female’ occupations, depressing wages there for otherwise equally productive workers. The same wage outcomes could also be observed if women voluntarily exclude themselves from male jobs due to gender differences in preferences for the jobs themselves or for various attributes of the jobs (for example, long hours or necessity for travel).

Two recently proposed models of discrimination suggest alternative motivations for male employees to discriminate against female coworkers than the personal prejudices assumed in the Becker model, particularly for resisting the introduction of women into traditionally male occupations. In Akerlof and Kranton (2000), occupations are associated with societal notions of ‘male’ and ‘female’, leading men to resist the entry of women due to the loss in male identity (or sense of self) that this would entail. In Goldin (2002), the entry of women is viewed as reducing the prestige of the occupation, based on perceptions that women are, on average, less productive.

An additional gender-specific factor potentially contributing to the observed gender wage gap is labour force selectivity. While one would ideally like to have evidence on the potential wage offers available to each individual in the population, we typically observe wages only for those who are actually employed. If there are unobserved differences in skills or labour market prospects between the non-employed and the employed, focusing on measured wages may give a misleading picture of the wage offers received by women relative to men (Heckman, 1979), both at a point in time and for trends over time and differences across countries.

Wage structure

The human capital model suggests that men and women tend to have different levels of labour market qualifications (especially work experience) and to be employed in different occupations and perhaps in different industries. Discrimination models too suggest that women may be segregated into different sectors of the labour market. This implies that the overall returns to skills and the size of premia for employment in particular sectors potentially play an important role in determining the gender wage gap. All else equal, the larger the returns to skills and the larger the rents received by individuals in predominantly-male sectors, the larger will be the gender wage gap. The framework provided by wage structure is particularly useful in analysing changes over time in gender differentials or differences across countries in gender gaps.

Interactions between gender-specific factors and wage structure

While gender specific factors and wage structure each potentially play a distinct role in affecting the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
gender wage gap, they are likely to interact, making it sometimes difficult to disentangle their separate effects. For example, as discussed in more detail below, since the 1970s, the labour market returns to skills such as education, specialized training and experience have risen in many countries, likely due in part to technological change, including computerization. To the extent that the prices of skills for which women have a relative deficit have risen, such changes in wage structure will raise the gender wage gap. However, technological change itself is not likely to have gender neutral effects on labour demand, given occupational and industrial segregation patterns by gender. So, for example, it is likely that computerization has reduced the demand for blue-collar production labour and therefore lowered the relative demand for sectors where men are disproportionately represented (Weinberg, 2000; Autor, Levy and Murnane, 2003).

These types of analyses raise the question of what the appropriate measure of wage structure or labour market prices is. For a number of reasons it is often viewed as appropriate to use male prices/returns as the measure of overall wage structure, with the maintained hypothesis that they therefore affect women's relative wages as well. For one thing, it is believed that men do not encounter discrimination and thus their returns are not contaminated by discrimination, although it is acknowledged that, were gender discrimination to be eliminated, male as well as female prices would likely change (for example, the supply of labour to some traditionally male occupations might increase). For another, the estimate of male prices is less likely than female prices to be influenced by selection bias or workforce interruptions. However, some research suggests that the connection between the male wage structure and women's labour market outcomes may be complicated. For example, Fortin and Lemieux (1998) present a model in which there is a fixed hierarchy of jobs. As women move up the hierarchy, they replace some men who previously would have had middle-level positions, bumping them down the hierarchy. Thus, increases in women's human capital or reductions in employment discrimination against women may cause increases in male wage inequality, and the male wage distribution may change even with no changes in the overall wage distribution. Topel (1994) makes a related argument to the effect that high-skill women compete with low-skill men in the labour market and, thus, that increases in the supply of high-skill women directly lower the real wages of low-skill men (through this increase in supply) and thereby raise male wage inequality.

Evidence on human capital, discrimination and the gender wage gap

The typical approach to analysing the sources of the gender wage gap is to estimate wage regressions specifying the relationship between wages and productivity-related characteristics for men and women. While it would be preferable to analyse total compensation (including non-wage benefits and compensating differentials for job amenities), virtually all studies focus on money wages since data on total compensation are generally not available. The gender wage gap may then be statistically decomposed into two components: one due to gender differences in measured characteristics, and another which is ‘unexplained’ and potentially due to discrimination. Such empirical studies provide evidence consistent with both human capital differences and labour market discrimination in explaining the gender wage gap.

One problem with this approach is that evidence for discrimination relies on the existence of a residual gender wage gap, which cannot be explained by gender differences in measured qualifications. This

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
accords well with the definition of labour market discrimination, that is, pay differences between groups that are not explained by productivity differences, but these may also reflect group differences in unmeasured qualifications. If men are more highly endowed with respect to these omitted variables then we would overestimate discrimination. And, conversely, to the extent that women are more highly endowed with respect to the omitted variables, discrimination would be underestimated. Another case in which discrimination would be underestimated would be if some of the factors controlled for (for example, occupation or tenure with the employer) themselves reflect the impact of discrimination.
Another challenge to empirically decomposing the gender wage gap into its constituent parts is the existence of feedback effects. The traditional division of labour in the family may influence women's market outcomes through its effects on their acquisition of human capital and on rationales for employer discrimination against them. But it is also the case that, by lowering the market rewards to women's human capital investments and labour force attachment, discrimination may reinforce the traditional division of labour in the family (for example, Weiss and Gronau, 1981). Even small initial discriminatory differences in wages may cumulate to large ones as men and women make human capital investment and time allocation decisions on the basis of them.

Representative findings from statistical analyses

Representative findings from analyses of this type may be illustrated by results from three recent studies of the gender wage gap in the United States (Blau and Kahn, 2006), Denmark (Datta Gupta, Oaxaca and Smith, 2006), and Sweden (Edin and Richardson, 2002). Each of these studies uses databases that have information on actual labour market experience, a variable that is crucial for the analysis and is often not available in nationally representative data sets.
For the United States, Blau and Kahn (2006) found a female–male ratio for average hourly earnings of 79.7 per cent in 1998. In light of the issues discussed above, they considered results when only human capital variables (that is, education and labour market experience) and race were taken into account, and results additionally controlling for occupation, industry and unionism. While gender differences in educational attainment were fairly small, and actually favoured women, men had more full-time work experience than women. Controlling for human capital, women earned 81 per cent of what men earned; the relatively small increase in the human-capital adjusted ratio compared with the raw ratio reflects the offsetting effects of adjusting for gender differences in education and experience. The gender ratio rose to 91 per cent when industry, occupation and union status were additionally controlled for. For Denmark, Datta Gupta, Oaxaca and Smith (2006) found an unadjusted gender ratio of 81.1 per cent in 1995. This rose to 83.2 per cent controlling for schooling and experience, and to 86.2 per cent additionally controlling for industry, occupation and region. Edin and Richardson (2002) found qualitatively similar results for Sweden in 1991. Thus, in all three countries measured characteristics explained some but not all of the gender wage gap.
Studies such as those discussed above suggest that gender differences in human capital (especially experience) can be an important factor helping to account for the gender wage gap at any given point in time. In the United States, improvements in women's relative experience were an important factor in explaining the rise in women's relative wages during the 1980s (Blau and Kahn, 1997; 2006; O'Neill and Polachek, 1993), while increases in women's relative experience and education both contributed to

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
female wage gains in the 1990s (Blau and Kahn, 2006). And since 1980 the unexplained wage gap in the United States has fallen, a finding consistent with a decline in discrimination or improvements in women's unmeasured characteristics, and also, as we shall see below, with shifts in relative demand favouring women. Sample selectivity can also affect measured gender wage gaps. For example, using different methodologies, Blau and Kahn (2006) and Mulligan and Rubinstein (2005) both find a role for selectivity in explaining these wage trends.

Additional evidence on discrimination

A problem with the types of statistical analyses just discussed is that evidence of discrimination is based on a residual or unexplained gender wage gap that is susceptible to a variety of interpretations, of which labour market discrimination is only one. Two lines of empirical research on discrimination pursue alternative approaches which lend additional support to the finding of discrimination.

First are two studies that use an experimental approach. Neumark (1996) analysed the results of a hiring ‘audit’ in which male and female pseudo-job seekers were given similar résumés and sent to apply for waiter or waitress jobs at the same set of Philadelphia restaurants. In high-priced restaurants where earnings of workers are generally higher than in the other establishments, a female applicant's probability of getting an interview was 40 percentage points lower than a male's, and her probability of getting an offer was 50 percentage points lower. A second study, by Goldin and Rouse (2000), examined the impact of the ‘natural experiment’ in which major symphony orchestras in the United States adopted ‘blind’ auditions. In a blind audition, a screen is used to conceal the identity of the candidate. Using data from actual auditions, the authors found that the screen substantially increased the probability that a woman would advance out of preliminary rounds and be the winner in the final round. Goldin and Rouse (2000) used their parameter estimates to conclude that the switch to blind auditions can explain one quarter of the increase in the female percentage in the top five symphony orchestras in the United States from less than five per cent of all players in 1970 to 25 per cent in 1996.

A second source of additional evidence on discrimination is provided by studies that examine predictions of Becker's (1957) discrimination model and obtain results which are consistent with the model and hence with discrimination against women. Becker and others have pointed out that competitive forces should reduce or eliminate employer discrimination in the long run because the least discriminatory firms, which hire more lower-priced female labour, would have lower costs of production and should drive the more discriminatory firms out of business. For this reason, Becker suggested that discrimination would be more severe in firms or sectors that are shielded to some extent from competitive pressures. Consistent with this reasoning, Hellerstein, Neumark and Troske (2002) found that, among plants with high levels of product market power, those employing relatively more women were more profitable. Similarly, Black and Strahan (2001) report that, with the deregulation of the banking industry beginning in the mid-1970s, the gender wage gap in banking declined. And Black and Brainerd (2004) found that increasing vulnerability to international trade reduced apparent gender wage discrimination in concentrated industries, again as predicted by the Becker model.

Possible sources of the unexplained gender wage gap

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

While there appears to be evidence from a variety of approaches that is consistent with discrimination against women in the labour market, this does not mean that the full unexplained gap estimated in traditional approaches may be attributed to discrimination. Some of the residual gap may be due to the impact of childbearing on women's wages. This is not a factor that can be examined simply by including a control for number of children in a wage regression, since the coefficient on children variables may be influenced by self-selection into motherhood and the endogeneity of number of children. Research that addresses some of these issues suggests a negative effect of children on wages, even when labour market experience is controlled for (for example, Waldfogel, 1998). This may reflect the fact that, in the past, the birth of a child often meant that a woman withdrew from the labour force entirely, breaking her tie to her employer and forgoing the returns to any firm-specific training she might have acquired, as well as any rewards for having made an especially good job match.

Another possible source of the unexplained wage gap is noncognitive skills/traits or what Fortin (2005) terms ‘soft factors’. For example, experimental evidence suggests there are gender differences in competitiveness (for example, Gneezy, Niederle and Rustichini, 2003), and negotiating skills (Babcock and Laschever, 2003). Fortin (2005) examines the impact of a number of noncognitive traits and attitudes in a wage regression context. While such findings are informative in elucidating the omitted factors that lie behind the unexplained gap in traditional wage equations, as Fortin acknowledges, the coefficients on soft factors in a wage equation cannot necessarily be given a causal interpretation. Both wages and attitudes, for example, may be determined by the same exogenous factor. And, as in the case of the traditional productivity proxies discussed above, there may be important feedback effects from differential treatment in the labour market to noncognitive traits. So, for example, income expectations may influence wages through negotiating behaviour or effort, but the source of women's lower income expectations could be, at least in part, anticipation of labour market discrimination. Nor is it clear that all such omitted factors favour men. Borghans, ter Weel and Weinberg (2005) argue for a female advantage in interpersonal interactions, which they proxy by altruism.

Just as the importance of gender differences in the traditional human capital variables may change over time, thus helping to account for the decline in the gender wage gap, so may the impact of noncognitive traits. In this regard, it is interesting that Fortin (2005) finds evidence that gender differences in work attitudes were much smaller in 2000 than in 1986. Further, Borghans, ter Weel and Weinberg (2005) find evidence of a growing importance of interpersonal interactions (in part due to increased computer use) in affecting wages that can help explain rapidly rising female relative wages in the 1980s as well as a slower rate of increase in the 1990s.

The impact of policy

Women's relative skills and the degree of employer discrimination can be affected by government policies directed at issues of combining work and family as well as equal employment opportunity laws. For example, many countries have enacted paid parental leave mandates which give parents who take time off to care for children or other relatives an entitlement to their jobs upon returning from the leave. While such policies may encourage firm-specific investments, thus raising women's relative wages (since parental leave is much more likely to be taken by women than men), they may also encourage labour force withdrawal for longer periods of time than otherwise, reducing women's accumulation of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
experience. Mandated paid leaves, particularly of long duration, may also diminish women's opportunities by increasing employer costs of hiring women and hence providing incentives to discriminate against them. Thus, the effect of parental leaves on the gender wage gap is theoretically ambiguous. Ruhm (1998) in fact finds in a study of 16 Western industrialized countries that, other things equal, short mandated paid parental leaves lead to higher relative wages for women, while longer leaves lead to a higher gender wage gap. These results suggest that a number of offsetting factors may be at work, with the positive impact dominating for short leaves and the negative effect dominating for long periods of mandated parental leave.

While virtually all industrialized countries have enacted legislation outlawing employment discrimination against women, in some countries government intervention is more dramatic than in others. The major approach in the United States involves enforcement of antidiscrimination legislation, including equal employment opportunity as well as equal pay for equal work. Further, under some circumstances, affirmative action, or ‘pro-active steps … to erase differences between women and men, minorities and nonminorities, etc.’ (Holzer and Neumark, 2000, p. 484), is also required or voluntarily adopted by employers. There is some evidence for the United States of a positive effect of government anti-discrimination policies on women's earnings and occupations. Studies focusing specifically on the impact of affirmative action also suggest modest employment and wage gains for women attributable to this programme. (For a summary, see Blau, Ferber and Winkler, 2006, pp. 240–5.)

‘Comparable worth’ or equal pay for work of equal value (that is, even if men and women are doing different jobs) constitutes a stronger form of government intervention. In evaluating the impact of such a policy it is interesting to look at studies focusing on Australia, which has adopted government mandates in this area nationwide, and the United States, where such policies are limited to selected state or local government employees (Gregory and Duncan, 1981; Killingsworth, 1990; O’Neill, Brien and Cunningham, 1989). One would expect that if such policies lower the gender wage gap, they might also lead to a decrease in women's relative employment due to employer demand effects. Gregory and Duncan (1981) in fact find such a pattern: the gender wage gap fell dramatically immediately after the Australian tribunal began implementing comparable worth policies in the early 1970s, but female employment grew less rapidly than one would have predicted in the absence of the wage intervention. Similarly, in studies of the impact of the comparable worth policies in state governments in the United States, small positive wage and negative employment effects for women have been found (Killingsworth, 1990; O’Neill, Brien and Cunningham, 1989).

While these results for the impact of comparable worth in Australia and the United States on women's employment are consistent with the existence of competitive labour markets, to the extent that the labour market is characterized by monopsony, government-mandated wage increases for women need not result in a reduction in women's employment levels. Manning (1996) interprets the impact of the UK Equal Pay Act of 1970 and the Sex Discrimination Act of 1975 in this light. Specifically, he shows that these laws led to a major reduction in the gender wage gap in the United Kingdom with no apparent employment losses for women: after the legislation, wage changes and employment changes within industries were strongly positively related for women but much less so for men.

Differential monopsony power facing men and women could help to explain the existence of the gender wage gap (Madden, 1973) in general, as well as Manning's (1996) results for the policy intervention. For this explanation of the wage gap to make sense, women's supply of labour to the *firm* must be less wage

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
elastic than men's, giving employers greater monopsony power over women than men. This might seem counter-intuitive at first, in that there is clear evidence that women have a *larger* own-wage elasticity of labour supply to the *labour market* than men, although in the United States the gender difference has been decreasing as women's elasticity has declined since 1980 (Blau and Kahn, 2007; Heim, 2007). However, a variety of factors could still potentially result in women having a smaller responsiveness to wage changes at the firm level. Perhaps the most intriguing possibility is discrimination itself. Black (1995) develops a model in which search costs give employers a degree of monopsony power. If there is discrimination against women, women will face higher search costs than men, increasing employers' monopsony power over them.

Evidence on gender differences in labour supply at the firm level is mixed. On the one hand, Viscusi (1980), Blau and Kahn (1981) and Light and Ureta (1992) all find that, for the United States, women's quit rates are at least as wage responsive as men's, suggesting that the monopsony model may have limited application in the United States. On the other hand, Barth and Dale-Olsen (1999) found that men's turnover in Norway is more wage-elastic than women's. Thus, Norwegian employers could potentially exercise differential monopsony power over women. Of course, the degree to which Norway's centralized wage-setting system would allow this to take place is an empirical question (Kahn, 1998).

Evidence on the impact of wage structure on the gender wage gap

The impact of wage structure on the gender wage gap is best studied in a comparative context. Since wage structure may differ across countries and change over time, investigations of the impact of wage structure have focused on (*a*) international differences in the gender wage gap at a specific point in time, and (*b*) changes in the gender wage gap in one country over time. A useful framework for analysing the impact of wage structure on demographic wage differentials was devised by Juhn, Murphy and Pierce (JMP) (1991) in their analysis of changes in black workers' relative wages in the United States. Blau and Kahn (for example, 1992, 1996b) have adapted their framework to studying international differences in the gender wage gap as follows.

Suppose we have for male worker *i* in country *j* the following wage equation:

$$Y_{ij} = B_j' X_{ij} + e_{ij} = B_j' X_{ij} + F_j^{-1}(\theta_{ij}), \quad (1)$$

where Y is log of wages, B is a coefficient vector, X is a vector of productivity-related characteristics, e is a disturbance term, $F^{-1}(\cdot)$ is the inverse cumulative distribution function of male log wage residuals, and θ is individual *i*'s percentile in the male residual distribution. Estimating eq. (1) separately for each of two countries, differences in the gender wage gap may be decomposed into components due to inter-country differences in: (*a*) gender differences in the X variables; (*b*) the male wage coefficients B; (*c*)

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
women's position in the male residual distribution (θ); and (*d*) the residual distribution *F*(·). Components (*a*) and (*c*) represent gender-specific factors: inter-country differences in women's relative measured productivity (component *a*) and in women's placement in the distribution of male wage residuals (component *c*). The latter can represent discrimination or unmeasured productivity differences. Components (*b*) and (*d*) represent the potential effects of wage structure: measured prices (component *b*) and the prices of unmeasured skills or rents due to unmeasured representation in favourable sectors (component *d*). Note that the sum of components (*c*) and (*d*) corresponds to the unexplained gap in a traditional decomposition of the gender wage gap, and component (*c*) may be viewed as the unexplained gap adjusted for differences in unmeasured prices.

Some examples may help to illustrate these components. It is straightforward that, if one country has a larger gender gap in experience, it will have a larger gender wage gap. But it is less obvious in the absence of this decomposition that, if the return to experience, which is part of the *B* vector, is higher in one country than another, then this difference will contribute to a higher gender wage gap in the first country, since women on average have less experience than men. Or suppose that *X* does not include data on the specific firm in which a worker is employed. If a country has especially high inter-firm wage differentials (part of the residual wage distribution *F*( )) and if women are employed in low-wage firms on average, then this unmeasured price effect will raise that country's gender wage gap. The same decomposition can also be used to explain changes in the gender wage gap over time within a country.

It may be noted that the interpretation of the residual proposed by JMP has been questioned by Suen (1997). For further discussion of this issue, see our discussion below on the assumption in the JMP decomposition that male prices and male residuals are relevant indicators of the prices facing women in the labour market. A fuller discussion is provided in Blau and Kahn (2003).

The JMP decomposition has been used by Blau and Kahn (1992; 1996b) and Kidd and Shannon (1996) to study international differences in the gender wage gap at a point in time. For example, Blau and Kahn (1996b) compared the US gender wage gap in the late 1980s with that in nine other countries (Australia, Austria, West Germany, Hungary, Italy, Norway, Sweden, Switzerland and the United Kingdom). They found that, on average, the ratio of women's to men's pay was 4.3 percentage points lower in the United States than in the other countries: 65.4 per cent as against 69.7 per cent. US women had better measured characteristics and were placed higher in the distribution of male residuals than women in the other countries, suggesting that gender-specific factors could not explain the higher US gender wage gap. However, measured and unmeasured prices together had large effects raising women's relative wages in the other countries compared with the United States. Wage structure was thus sufficient to explain more than the full amount of the difference between the US gender wage gap and that in other countries. Blau and Kahn (1992; 1996b) interpreted this pattern as reflecting the impact of international differences in labour market institutions. In the other countries, unions cover a much larger portion of the labour market than in the United States, and wage-setting is much more centralized. Centralized collective bargaining tends to reduce wage differentials through the negotiation of relatively high wage floors, which would raise the relative wages of anyone near the bottom of the distribution, including women (Blau and Kahn, 1996a).

Kidd and Shannon (1996) also found an important role for wage structure in their study of the gender wage gaps in Australia and Canada for 1989–90. Specifically, the gender gap in hourly wages was about 0.14 log points lower in Australia than in Canada. They found that 0.05–0.09 log points of this

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
difference was due to the combined effects of observed and unobserved prices. This result is similar to that in Blau and Kahn (1992; 1996b) in that Australia has much higher coverage by collective bargaining than Canada.

The JMP decomposition has also been used to study the impact of wage structure on changes in the gender wage gap over time within a country. For example, in Sweden between 1968 and 1974, the trade union movement engineered a major compression of wages. Edin and Richardson (2002) used the JMP decomposition to find that wage structure, especially unobserved prices, contributed to a reduction of the gender wage gap during this period. Moreover, Datta Gupta, Oaxaca and Smith (2006) used a version of the JMP decomposition to study changes in the Danish gender wage gap between 1983 and 1995. This was a period of increased decentralization of the wage determination process, a development that is expected to lead to rising labour market prices and therefore a rising gender wage gap. The authors indeed found that the gender wage gap in Denmark increased during this period and that most of the increase can be accounted for by rising unmeasured prices.

Finally, this approach has been applied to understanding the trends in the gender wage gap in the United States. Blau and Kahn (1997) used the JMP decomposition to study the apparent paradox of a substantial decrease in the gender wage gap in the United States during the 1979–88 period, a time of rising skill prices. They found that women were able to overcome the negative effects of these price changes by improving their measured human capital and by moving up the distribution of male residuals. The authors further noted that the process leading to higher skill prices in the United States might not have been gender-neutral. Specifically, it is likely that part of the explanation for this development involves skill-biased technical change in which the demand for white-collar labour rose relative to blue-collar labour, a change that, given gender differences in occupational distributions, in effect raises the demand for women workers. Thus, while skill prices were rising, contributing to a reduction in women's relative wages, developments such as computerization and perhaps outsourcing of production labour disproportionately lowered the demand for male labour (Welch, 2000; Weinberg, 2000; Autor, Levy and Murnane, 2003). In the context of the JMP decomposition, such changes in labour demand would be reflected in higher placement of women in the distribution of male residuals (leading to a decrease in the conventional unexplained gender wage gap). As noted previously, convergence in the gender wage gap in the United States slowed during the 1990s. Using the JMP decomposition, Blau and Kahn (2006) found that the major reason for the slowdown was the considerably smaller narrowing of the unexplained gender wage gap in the 1990s than in the 1980s. This raises the possibility that the types of demand shifts favouring women that we have outlined here were smaller in the 1990s than in the 1980s, and Blau and Kahn present some evidence that is consistent with this as at least a partial explanation for the smaller decrease in the unexplained gender gap in the latter period.

The JMP decomposition assumes that male prices and male residuals are relevant indicators of the prices facing women in the labour market. Some support for this assumption is provided by the fact that wage coefficients and residual distributions have changed similarly for men and women over time in the United States and are similar to each other within countries at a point in time (Blau and Kahn, 2002). But it is possible to directly test whether male wage compression leads to a smaller gender wage gap, and Blau and Kahn (2003) have done so by compiling a microdata-set for 22 countries over the 1985–94 period. They find, looking across countries, that the gender wage gap is positively affected by a country's male skill prices (that is, the level of male wage inequality adjusted for measured

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
characteristics), as well as by the relative net supply of women (that is, supply net of demand). A likely interpretation is that more compressed male wages are an indicator of smaller wage differentials in general, as suggested above in our discussion of centralized wage-setting institutions. Bolstering this interpretation is the authors' further finding that, other things equal, greater coverage by collective bargaining reduces the gender wage gap. It thus appears that high wage floors negotiated by unions serve to lower the gender wage gap.

If labour markets are competitive, then union-negotiated wage floors should lower female relative employment. And this is precisely what Bertola, Blau and Kahn (2007) find in a study of relative employment in 17 countries over the 1960–96 period. Specifically, they find that greater coverage by highly centralized unions lowers female employment and raises female unemployment compared with men's. This suggests that unionization can raise women's relative wages at the expense of lowering their employment. This in turn suggests that Manning's (1996) evidence in favour of monopsony in the United Kingdom may describe an exceptional case in the OECD.

Future prospects

While it is difficult to speculate about the future, Tables 1 and 2 do suggest some convergence across countries in both female labour force participation (absolutely and relative to men) and in the gender wage gap. For example, calculations based on the data in Tables 1 and 2 indicate that the standard deviation across countries in the ratio of women's to men's labour force participation rates fell from 0.128 in 1980 to 0.072 in 2000, and that the standard deviation of the gender wage ratio fell steadily from 0.086 in 1980 to 0.066 by 2000. While the gender wage ratio appears to be converging at around 80 per cent for several of the countries in Table 1, further changes are not precluded. Throughout the OECD, women's education has been rising relative to men's, a trend that shows no sign of ending (Goldin, Katz and Kuziemko, 2006). Technological change, which has likely raised women's relative wages through demand effects, will probably continue and could even accelerate. Going against these trends is the likely continued decentralization of wage-setting institutions in many Western countries, spurred in part by globalization (Katz, 1993).

See Also

- affirmative action
- family decision making
- feminist economics
- gender differences (experimental evidence)
- gender roles and division of labour
- inequality (international evidence)
- labour supply

Bibliography

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Aigner, D. and Cain, G. 1977. Statistical theories of discrimination in labor markets. *Industrial & Labor Relations Review* 30, 175–87.

Akerlof, G.A. and Kranton, R.E. 2000. Economics and identity. *Quarterly Journal of Economics* 115, 715–53.

Autor, D.H., Levy, F. and Murnane, R.J. 2003. The skill content of recent technological change: an empirical exploration. *Quarterly Journal of Economics* 118, 1279–333.

Babcock, L. and Laschever, S. 2003. *Women Don't Ask: Negotiation and the Gender Divide*. Princeton and Oxford: Princeton University Press.

Barth, E. and Dale-Olsen, H. 1999. Monopsonistic discrimination and the gender wage gap. Working Paper No. 7197. Cambridge, MA: NBER.

Becker, G.S. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Becker, G.S. 1985. Human capital, effort, and the sexual division labor. *Journal of Labor Economics* 3 (Supp.), S33–S58.

Bergmann, B. 1974. Occupational segregation, wages, and profits when employers discriminate by race or sex. *Eastern Economic Journal* 1, 103–10.

Bertola, G., Blau, F.D. and Kahn, L.M. 2007. Labor market institutions and demographic employment patterns. *Journal of Population Economics* 20, 833–67.

Black, D.A. 1995. Discrimination in an equilibrium search model. *Journal of Labor Economics* 13, 309–34.

Black, S.E. and Brainerd, E. 2004. The impact of globalization on gender discrimination. *Industrial & Labor Relations Review* 57, 540–59.

Black, S.E. and Strahan, P.E. 2001. The division of spoils: rent-sharing and discrimination in a regulated industry. *American Economic Review* 91, 814–31.

Blau, F.D., Ferber, M.A. and Winkler, A.E. 2006. *The Economics of Women, Men, and Work*, 5th edn. Englewood Cliffs, NJ: Prentice-Hall.

Blau, F.D. and Kahn, L.M. 1981. Race and sex differences in quits by young workers. *Industrial & Labor Relations Review* 34, 563–77.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Blau, F.D. and Kahn, L.M. 1992. The gender earnings gap: learning from international comparisons. *American Economic Review* 82, 533–8.

Blau, F.D. and Kahn, L.M. 1996a. International differences in male wage inequality: institutions versus market forces. *Journal of Political Economy* 104, 791–837.

Blau, F.D. and Kahn, L.M. 1996b. Wage structure and gender earnings differentials: an international comparison. *Economica* 63(Supp.), S29–S62.

Blau, F.D. and Kahn, L.M. 1997. Swimming upstream: trends in the gender wage differential in the 1980s. *Journal of Labor Economics* 15, 1–42.

Blau, F.D. and Kahn, L.M. 2002. *At Home and Abroad: U.S. Labor Market Performance in International Perspective*. New York: Russell Sage.

Blau, F.D. and Kahn, L.M. 2003. Understanding international differences in the gender pay gap. *Journal of Labor Economics* 21, 106–44.

Blau, F.D. and Kahn, L.M. 2005. Do cognitive test scores explain higher US wage inequality? *Review of Economics and Statistics* 87, 184–93.

Blau, F.D. and Kahn, L.M. 2006. The US gender pay gap in the 1990s: slowing convergence. *Industrial & Labor Relations Review* 60, 45–66.

Blau, F.D. and Kahn, L.M. 2007. Changes in the labor supply behavior of married women: 1980–2000. *Journal of Labor Economics* 25, 393–438.

Borghans, L., ter Weel, B. and Weinberg, B.A. 2005. People people: social capital and the labor-market outcomes of underrepresented groups. Discussion Paper No. 1494, Institute for the Study of Labor (IZA).

Datta Gupta, N., Oaxaca, R.L. and Smith, N. 2006. Swimming upstream, floating downstream: comparing women's relative wage position in the US and Denmark. *Industrial & Labor Relations Review* 59, 243–66.

Edin, P.-A. and Richardson, K. 2002. Swimming with the tide: solidarity wage policy and the gender earnings gap. *Scandinavian Journal of Economics* 101, 49–67.

Fortin, N. 2005. Greed, altruism, and the gender wage gap. Working paper, University of British Columbia.

Fortin, N.M. and Lemieux, T. 1998. Rank regressions, wage distributions, and the gender gap. *Journal*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
of Human Resources 33, 610–43.

Gneezy, U., Niederle, M. and Rustichini, A. 2003. Performance in competitive environments: gender differences. *Quarterly Journal of Economics* 118, 1049–74.

Goldin, C. 2002. A pollution theory of discrimination: male and female differences in occupations and earnings. Working Paper No. 8985. Cambridge, MA: NBER.

Goldin, C., Katz, L. and Kuziemko, I. 2006. The homecoming of American college women: the reversal of the college gender gap. *Journal of Economic Perspectives* 20 133–56.

Goldin, C. and Rouse, C. 2000. Orchestrating impartiality: the impact of 'blind' auditions on female musicians. *American Economic Review* 90, 715–41.

Gregory, R.G. and Duncan, R.C. 1981. Segmented labor market theories and the Australian experience of equal pay for women. *Journal of Post Keynesian Economics* 3, 403–28.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–61.

Heim, B.T. 2007. The incredible shrinking elasticities: married female labor supply, 1979–2003. *Journal of Human Resources* (forthcoming).

Hellerstein, J.K., Neumark, D. and Troske, K. 2002. Market forces and sex discrimination. *Journal of Human Resources* 37, 353–80.

Holzer, H. and Neumark, D. 2000. Assessing affirmative action. *Journal of Economic Literature* 38, 483–568.

Juhn, C., Murphy, K.M. and Pierce, B. 1991. Accounting for the slowdown in black-white wage convergence. In *Workers and Their Wages*, ed. M. Kosters. Washington, DC: AEI Press.

Kahn, L.M. 1998. Against the wind: bargaining recentralisation and wage inequality in Norway, 1987–1991. *Economic Journal* 108, 603–45.

Katz, H.C. 1993. The decentralization of collective bargaining: a literature review and comparative analysis. *Industrial & Labor Relations Review* 47, 3–22.

Kidd, M.P. and Shannon, M. 1996. The gender wage gap: a comparison of Australia and Canada. *Industrial & Labor Relations Review* 49, 729–46.

Killingsworth, M. 1990. *The Economics of Comparable Worth*. Kalamazoo, MI: W.E. Upjohn Institute

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

for Employment Research.

Light, A. and Ureta, M. 1992. Panel estimates of male and female job turnover behavior: can female nonquitters be identified? *Journal of Labor Economics* 10, 156–81.

Madden, J.F. 1973. *The Economics of Sex Discrimination*. Lexington, MA: DC Heath.

Manning, A. 1996. The Equal Pay Act as an experiment to test theories of the labour market. *Economica* 63, 191–212.

Mincer, J. and Polachek, S. 1974. Family investments in human capital: earnings of women. *Journal of Political Economy* 82(Part 2), S76–S108.

Mulligan, C.B. and Rubinstein, Y. 2005. Selection, investment, and women's relative wages since 1975. Working Paper No. 11159. Cambridge, MA: NBER.

Neumark, D.M. 1996. Sex discrimination in restaurant hiring: an audit study. *Quarterly Journal of Economics* 111, 915–41.

O'Neill, J., Brien, M. and Cunningham, J. 1989. Effects of comparable worth policy: evidence from Washington State. *American Economic Review* 79, 305–9.

O'Neill, J. and Polachek, S. 1993. Why the gender gap in wages narrowed in the 1980s. *Journal of Labor Economics* 11, 205–28.

OECD (Organisation for Economic Co-operation and Development). 1990. *Employment Outlook*. Paris: OECD.

OECD. 2004. *Employment Outlook*. Paris: OECD.

Ruhm, C.J. 1998. The economic consequences of parental leave mandates: lessons from Europe. *Quarterly Journal of Economics* 113, 285–317.

Suen, W. 1997. Decomposing wage residuals: unmeasured skill or statistical artifact? *Journal of Labor Economics* 15, 555–66.

Topel, R.H. 1994. Regional labor markets and the determinants of wage inequality. *American Economic Review* 84, 17–22.

Viscusi, W.K. 1980. Sex differences in worker quitting. *Review of Economics and Statistics* 62, 388–98.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Waldfogel, J. 1998. The family gap for young women in the United States and Britain: can maternity leave make a difference? *Journal of Labor Economics* 16, 505–45.

Weinberg, B. 2000. Computer use and the demand for female workers. *Industrial & Labor Relations Review* 53, 290–308.

Weiss, Y. and Gronau, R. 1981. Expected interruptions in labour force participation and sex-related differences in earnings growth. *Review of Economic Studies* 48, 607–19.

Welch, F. 2000. Growth in women's relative wages and in inequality among men: one phenomenon or two? *American Economic Review* 90, 444–9.

How to cite this article

Blau, Francine D. and Lawrence M. Kahn. "women's work and wages." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000129> doi:10.1057/9780230226203.1839

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Wood, Stuart (1853 – 1914)

George J. Stigler
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

factor substitution; marginal productivity theory; technological substitution; wages fund; Wood, S

Article

A member of a prominent Philadelphia Quaker family, Wood was briefly active in economics twice in his life. The first period was 1873–5, when he received at Harvard the first Ph.D. in economics in the United States. The second period was 1888–90, when he wrote three first-class articles on wage theory. His primary interests during his adult life were in business and finance.

In the two years Wood was at Harvard he took courses in economics and its history, chiefly from Professor Charles Dunbar, and wrote an essay on ‘A Review of the “Principles of Social Science” by Henry C. Carey’. It was not an impressive piece, even allowing for the time, the age (21) of the writer and the extreme vulnerability of the target.

It is all the more impressive that 13 years later he wrote two fine articles on the marginal productivity theory and one on the history of the wages-fund theory. Wood must be acknowledged to be an independent discoverer of the marginal productivity theory, an honour he shares with Marshall, Edgeworth, Barone, Wicksell, Clark and other major economists. Wood's version was not mathematical, but it synthesized two important dimensions of substitution between capital and labour: the substitution between industries with different capital–labour ratios, and the substitution within enterprises. The formulation was a skilful synthesis incorporating consumer demands and factor supplies as well as technological substitution.

Wood's final contribution was a history of the wages-fund doctrine (which was to be treated no more penetratingly by Harvard's second Ph.D. in economics, F.W. Taussig). Perhaps one should mention one other, involuntary role Wood played in the study of the history of economics: he was the victim of a thinly disguised, utterly unfounded charge of plagiarism (of Lord Lauderdale) in the *Journal of Political Economy* in 1894.

Selected works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

1888–9. A new view of the theory of wages. Pts I–II. *Quarterly Journal of Economics* 3, 60–86; 462–80.

1889. The theory of wages. *Publications of the American Economic Association* 4, 5–35.

1890. A critique of wage theories. *Annals of the American Academy of Political and Social Science* 1, 426–61.

Bibliography

Stigler, G.J. 1947. Stuart Wood and the marginal productivity theory. *Quarterly Journal of Economics* 61, 640–49.

How to cite this article

Stigler, George J. "Wood, Stuart (1853–1914)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000093>
doi:10.1057/9780230226203.1840

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

worker participation and profit sharing

John Pencavel

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

democracy; incentives; productivity; profit-sharing; worker participation

Article

Market economies are called ‘capitalist’ because in such economies most production is carried out in organizations owned by those who supply the firms' financial capital. A firm is ‘owned’ by its capital investors because, first, the capital investors claim the firm's net receipts or profits and, second, they have the authority to direct and manage (often indirectly) the firm's activities.

Yet in all market economies some production takes place in firms where these two dimensions of ownership are embodied in those who supply labour rather than capital. In this instance, workers enjoy as incomes the firm's net receipts and the workers hire individuals to supervise and organize production. Capital may be obtained from the workers' savings or from loans from financial intermediaries. Examples of worker cooperatives include the plywood companies in the Pacific Northwest of the United States, the kibbutzim in Israel, and the Mondragon group in the Basque country of Spain.

Many enterprises fall between these two limiting cases. These other firms are characterized by the owners either sharing net revenues with others – ‘profit-sharing’ – or sharing in the activities of management – ‘worker participation’. (For recent research on the general issues, see the essays in Blair and Roe, 1999, and Ichniowski et al., 2000.)

Profit-sharing occurs when those who have the right to consume all the firm's profits distribute a portion of them to others within the organization. Because most firms are owned by those who supply capital, profit-sharing usually occurs when some portion of profits is distributed among the rank-and-file workers.

With explicit profit-sharing, a clear formula is established linking profits and the pay of individuals. Profit-sharing is implicit when workers in firms that habitually enjoy higher profits are paid higher wages. Profit-sharing may take the form of deferred income, as when a portion of net receipts is placed in retirement accounts so that the firm's employees hold part of the assets of the firm in which they work. A principal goal of these various profit-sharing arrangements is to affect incentives: by linking workers'

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
compensation to the firm's success in making profits, the workers' interests are aligned more closely with the capital owners'. However, some economists reason that, when the firm's net earnings are divided among a large number of people and one individual's effort contributes little to total output, the incentive for a single individual to apply effort is meagre. What empirical evidence there is suggests that, with profit-sharing, workers monitor one another so that any tendency to shirk is checked.

When workers' pay is linked to profits, some automatic flexibility is imparted to a firm's payrolls so the effects of adverse shocks are communicated immediately and mechanically to the firm's costs. Some have suggested that, if profit-sharing payment schemes were widespread, recessions would be characterized by less unemployment. Kruse (1993) reviews profit-sharing.

‘Worker participation’ is a term embracing various arrangements by which workers are actively involved in the management of the enterprise where they work. These arrangements may include safety and health committees or panels to deal with worker grievances, or they may be more profound arrangements when workers are actively engaged in key management activities such as the organization of work and production. In Europe, works councils or workers' committees are empowered to be consulted and, sometimes, to share in determining any changes in the organization of production (also known as co-determination).

One argument in favour of worker participation is that participation begets productivity. Modern ‘flexible’ or ‘lean’ production techniques entail greater employee involvement in shop-floor decisions, greater teamwork, information-sharing between management and rank-and-file employees, and reduced task specialization. An extensive research literature quantifies the effect of greater worker participation on productivity. A general finding is that there are positive, though small, productivity benefits accompanying worker participation.

A second argument for worker participation is that it is the extension to the workplace of democratic governance in the political arena. In much the same way as citizens in political democracies have an important voice in choosing those who determine the provision of public goods in society so an enterprise's workers should have a voice in shaping their work environment when public goods are also prevalent. Worker participation is the application of the democratic principle to the workplace.

See Also

- industrial relations

Bibliography

Blair, M. and Roe, M., eds. 1999. *Employees and Corporate Governance*, Washington, DC: Brookings Institution Press.

Ichniowski, C., Levine, D., Olson, C. and Strauss, G., eds. 2000. *The American Workplace: Skills, Compensation, and Employee Involvement*. Cambridge: Cambridge University Press.

Kruse, D. 1993. *Profit Sharing: Does It Make a Difference?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

How to cite this article

Pencavel, John. "worker participation and profit sharing." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_C000550>
doi:10.1057/9780230226203.1841

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

World Bank

Albert Fishlow
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The World Bank, established in 1944, remained an important source of funding for developing countries generally through the early 1990s. Then an impressive increase in private flows reduced its overall significance. What remained was technical assistance on the one hand, and continued increasing credits of the International Development Association to the lowest income countries on the other. Despite criticism from both by the Right and the Left, the World Bank has survived, and has given voice to rising concerns about highly unequal distributions of income in the developing world, moving away from its earlier emphasis upon economic growth alone.

Keywords

conditionality; free trade; International Bank for Reconstruction and Development; International Development Association; International Finance Corporation; International Monetary Fund; Marshall Plan; non-governmental organizations; poverty alleviation; privatization; structural loans; World Bank; World Trade Organization

Article

The World Bank was founded in July 1944 as part of a new financial architecture for the post-Second World War period. At the inaugural meeting at Bretton Woods, the International Bank for Reconstruction and Development (IBRD) was proposed in order to satisfy demands in lesser-income countries for long-term capital, a market that had virtually disappeared with the Great Depression. The International Monetary Fund (IMF), much more a subject of debate at the time, was to satisfy anticipated short-term balance of trade needs, thereby avoiding competitive currency devaluations. A third component, the later proposed, but unfounded, International Trade Organization, was to resurrect freer trade. GATT, now the World Trade Organization (WTO), took on that function in 1947.
The Bank began in 1946 by focusing on reconstruction, which soon was successfully taken on by the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Marshall Plan, and then increasingly committed itself to the multiple and changing problems of economic development within the world's poorer countries. This conversion came about much more rapidly than John Maynard Keynes, and most other key economists in the immediate post-war period, had anticipated. Rapid European and Japanese economic recovery soon left space for loans entirely to developing countries.

The Bank utilized its multilateral resources in an ingenious way. The capital inputs originally established for member countries – among which the United States initially played a dominant role – required an actual contribution of only 20 per cent of the original capital of the Bank, with the remainder callable. Private financial markets were to be the real source of the money, and this they have remained despite modest increases in capital starting in 1959. By contrast, during its first decades, almost all the IMF's resources came directly from governments, and until the 1970s its focus remained the balance of payments problems of the developed countries.

Private direct investment within the developing countries also received early emphasis through the creation of the International Finance Corporation in 1956. Its performance did not live up to initial hopes. Only since 1995, through direct integration with the Bank, has much greater attention been given to the high cost of typical business procedures in many countries. The publication *Doing Business* has taken on a more significant role as private investment flows have multiplied.

The creation of the International Development Association (IDA) in 1960 fundamentally altered the initial conditions of Bank resources. Its funds, exclusively directed to the poorer countries on favourable terms, required regular triennial contributions. This circumstance allowed periodic legislative discussion of Bank policies, and pressures for policy changes, emanating principally from the United States. Its location in Washington, and the influence of an American president, reinforced that tendency. The initial volume of resources allocated was small. But already by 1980 IDA loans were amounting to equivalent net additions to total Bank lending as the conventional commitments. This part of the story, which was not an issue at the Bank's foundation, has now become the central feature of its decisions. (Other parts of the World Bank, such as the later International Center for the Settlement of Investment Disputes, ICSID, and Multilateral Investment Guarantee Agency, MIGA, should also be noted. These new components corresponded to new functions as the Bank expanded.)

The quality and size of Bank staff, which now totals more than 10,000, also merit mention. Begun at a time when ‘development economics’ was not yet a part of the economics curriculum within universities, the Bank soon began to employ a talented and professional group that much helped to elevate the status of the sub-discipline. Creation of the *World Development Report* and other publications, regular conference activities, increasingly held abroad, as well as training for many developing country economists through a period at the Bank, have made this intellectual role a positive highlight.

Finally, over a more than 60-year history, the Bank has altered its emphasis dramatically. Now it is best known for its commitment to the elimination of poverty, a subject of little import at its foundation, when its contribution to economic growth was the focus. Now, too, the Bank is exposed to increased opposition from both Left and Right as it searches to retain a role, not only intellectually but also practically. Its net financial contribution has continuously declined as a sophisticated international capital market has expanded, placing more of a burden upon Bank leadership.

In this article, I explore three subjects. First is an assessment of the changing pattern of Bank lending and its effects. Second is an exposition and evaluation of the critiques increasingly directed at the Bank since the early 1990s. Third, by way of conclusion, I raise an essential question: what role should the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Bank play in the future?

World Bank lending: 1946 – 2005

Table 1 provides the gross and net loan flows, as well as net transfers – with return interest payments subtracted – of the Bank and IDA at decadal intervals since the Bank's foundation. It also presents net private capital flows and net official flows to the developing countries. All have been converted to constant dollars. Three central conclusions immediately follow.

World Bank lending, total official flows and total private flows, 1960–2005 ($ billion, 2000)[a]

	Fiscal years						
	1960[b]	1970	1980	1990	1995	2000	2005
Disbursements		3.2	11.7	22.4	20.2	18.7	15.8
IBRD	1.8	2.5	8.6	17.0	13.9	13.5	8.7
IDA		0.6	3.1	5.4	6.3	5.2	7.0
Net flows		2.3	9.6	11.5	6.5	6.8	.5
IBRD	1.3	1.6	6.6	6.4	.8	2.9	–4.7
IDA		0.6	3.0	5.1	5.7	3.9	5.2
Net total official flows	22.3	20.0	65.4	69.5	59.2	23.2	–15.2
Net total private flows		20.3	78.4	54.4	226.0	189.7	443.4
Net transfers		1.3	6.0	2.7	–3.0	–2.0	–4.0
IBRD	.9	0.7	3.1	–2.1	–8.1	–5.4	–8.4
IDA		0.6	2.9	4.7	5.2	3.3	4.4

[a]The US chain-type GDP price index (averaging calendar years) has been used. Categories may not sum due to rounding.

[b]World Bank loans to developing countries only.*Sources*: For 1960, Mason and Asher (1973, pp. 208, 219). For other dates, World Bank, *Global Development Finance*, 2001 and 2005.

First, the gross real flows of regular disbursements, after an initial acceleration in the 1970s and continuing through to 1985, tend to stabilize thereafter. They then fall off considerably in later years. During the mid-1980s, net flows likewise began to diminish sharply, becoming negative by the early 21st century. Net transfers turn negative shortly after 1985 and become progressively more so thereafter. The Bank has ceased to be a source of resources for middle-income countries some time ago. That is the direct consequence of restricted gross outlays accompanied by amortization and market interest rate charges.

Additionally, the Bank altered its principal mandate in the 1980s, as it had done during the Robert McNamara presidency. Then, in the 1970s, the Bank launched new initiatives to deal with the extensive level of poverty found in the developing world. Income distribution, basic needs, reform of the agricultural sector all figured. With a debt crisis occurring soon after the oil shock of 1979, the Bank

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
underwent a transformation. Structural macroeconomic loans were introduced, moving away from the earlier sectoral emphasis. Conditionality loomed larger, and the Bank began to replicate the simultaneous involvement of the IMF. The Bank became notable for its emphasis upon the primary importance of market signals, as well as privatization and freer trade, sometimes to the exclusion of a positive role for the state that had figured importantly in the previous decade.

Second, IDA resources, regularly replenished and less diminished by repayments because of their longer term and lesser interest charges, grew substantially. But even these disbursements have not sustained their expansion. Now on a gross basis, they amount to nearly 80 per cent of regular Bank commitments, and will soon exceed them. On a net basis, the IDA proportion is much greater, as can be seen. The disparity is very much stronger when net transfers are recorded. Indeed, it is almost fair to say that the Bank's substantially increased commitment to the alleviation of poverty again in the 1990s was an almost inevitable consequence of its altered resource base. There was no other direction to take.

James Wolfensohn's active presidency, beginning in 1995, was also a major causal factor. Just like McNamara before him, Wolfensohn unleashed new programmes as an advocate of the poor. Coinciding with the rise of India and continuing rapid gains in China, and the beginning of recovery in Africa, his decade of engagement encountered a much better base for this renewed emphasis. Even the initial crises, in Mexico in 1995, in Asia in 1997, and in Russia in 1998 followed by Brazil within months, saw very rapid recovery; this was not a duplication of the lost decade of the 1980s. Wolfensohn transformed the Bank in terms of its managerial style, its relationship with non-governmental organizations (NGOs), its focus on institutions and governance and its emphasis upon concrete results.

Third, the Bank progressively became a marginal contributor of resources over this period, except to the very poor countries. This is clear if one compares the net private flows recorded in Table 1 with Bank, or even total official, lending. To compensate, the Bank's intellectual role has continuously had to be sharpened, redefined and extended, which helps to explain the expansion of branch offices abroad, the relocation of country directors, and the increase in activities within recipient borrowers. This is also why there has been such an emphasis upon information technology as an essential component for spreading knowledge about the development process. What started as a straightforward financial institution has been converted into a far more vocal and innovative participant in the advancement of the position of the poor. The Bank has led in the onerous task of reducing the accumulated debt that burdens many countries. It has also been active in defining a new vision, involving not only governments, but also civil society. But that transformation has not met with universal acclaim.

Should the Bank survive?

As the Bank approached its 50th anniversary in 1994, increasing unhappiness with its performance became evident. Since the McNamara years, there had been three successive presidents and a changed direction emphasizing macroeconomic programme loans to finance stabilization in a world increasingly adrift. A much broader range of internal reforms, going beyond the balance of payments and domestic savings, was now targeted. Stricter conditions were imposed. Almost 30 per cent of loans were allocated to stabilization objectives, and were concentrated among the highly indebted countries that were suffering from lack of private lending. Unfortunately, the record of accomplishment was not so high. The classification of risk in World Bank documents conveys that reality: in 1970, more than 70 per cent of loans were low risk; in 1980, 30 per cent. Arrears appeared. Positive IDA transfers compensated for

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
negative Bank flows in a number of countries newly eligible for these loans because of their declining income.

Critics multiplied, both on the Left and the Right. For both groups, the status quo was unacceptable. For the more radical, the correct solution was to close the Bank, but for most, the preferred outcome was an altered, more effective institution. Each side envisioned a redesigned Bank more able to accomplish its redesigned objectives.

For the Left, the Bank required nothing less than reinvention – as one critic would later put it – in its operations, concepts and distribution of power. Or, to put the matter another way, the object was a smaller, much more transparent, decentralized and pluralistic development bank. The new institution would be one where developing countries could exercise greater choice and have greater voice. Independent research and significant policy engagement would no longer be features of the Bank. Those functions would devolve to the developing countries themselves.

For the Right, the objective was an equally lesser institution, one that would provide grants to the poorest countries with limited alternative access to financial markets. These funds would be allocated to the conventional objectives identified at the Bank's outset, namely, health care, primary education and infrastructure. No attention would be directed to such issues as the environment, gender equality or labour standards. NGOs, which had become increasingly part of the developing community, would no longer be central participants. Shares of domestic contributions would vary, as a function of per capita income, from ten per cent to 90 per cent. Private capital markets could, and would, substitute for the very modest financial contribution of the Bank to the middle-income countries, and at much lesser cost. The Bank would no longer be engaged in lending to them. Instead, the Bank would limit itself to knowledge transfer and technical assistance for this group of countries. Independent auditors would conduct performance evaluation, emphasizing measured targets and results.

Neither of these directions of reform is now at the centre of discussion. The 60th anniversary has come and gone. One of the important reasons is the impressive acceleration of economic performance within the developing countries since the beginning of the 21st century. Another has been the ability of the Wolfensohn Bank to make itself more acceptable by adopting some of the suggestions from its critics on both sides. Thus the grant element in IDA loans has now risen to 30 per cent (as of 2007); there has been greater attention to the role of the state, as well as the private sector, within developing countries; transparency has increased; and there has been insistence upon country ownership, including broad participation of domestic groups, of the development projects being financed.

This may seem to work in the present. But there remains the question of what lies ahead.

The future of the Bank

Implicitly, and continuously, the Bank has throughout its history confronted the central issue of whether to give greater weight to growth or equity. During the crisis of the 1980s, the focus temporarily turned to economic recovery. Now, after the triumph of globalization, expansion of international trade and greater recognition of private sector importance, conditions have seemingly changed. They have also altered in a policy sense, with the establishment of the Millennium goals – UN-mandated objectives in a number of areas that developing countries are supposed to meet by 2015 – and a period of energetic commitment to expansion of social programmes to confront poverty.

As accelerating and generalized expansion has occurred for the first time in three decades, as of the early

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
21st century more countries are on the verge of graduation, and some of the IDA recipients are approaching their maximum income limit. At the same time, Bank evaluation has raised doubts about some of the new directions pursued: there has seemingly been too much effort directed to the social sectors, governance has not continuously improved nor has corruption been alleviated – despite the importance attached to these issues – and the numbers of people in poverty, other than in Asia, have since the mid-1990s been resistant to improvement. These difficulties are not easy to resolve.

An increase in international private resources as a source for investment has accompanied this expansion. These resources hardly show signs of stopping. Indeed, the speed of global recovery from the tumult at the end of the 1990s is a record accomplishment. Those crises were only a modest pause.

In this new world, the Bank will eventually have to adapt. As the data of Table 1 clearly reveal, neither Bank nor IDA lending is a central source of finance. Although much assistance is granted for political advantage, bilateral overseas development assistance regularly exceeds its net contribution. The Asian and Inter-American Development Banks dominate in their regions.

One direction of change, in a continuation with the recent past, may require an even greater degree of engagement with other agencies of the United Nations, whether the issue is the proliferation of new viruses or the extension of HIV. Another is likely to be a more active participation in global environmental efforts as increasing scientific research indicates the speed and importance of recent climate change. A third may involve efforts to finance infrastructure projects through shared participation of the private and public sectors, with the Bank engaged as a major contributor in poorer countries. A fourth may entail serious accommodation to the implications of changing global supplies of petroleum for the poorest countries.

These are just some possibilities. Many additional ones are sure to emerge. The Bank will have to take on a different role and function. As many developing countries begin to increase their income, the process of graduation implies a change in future leadership at the Bank. No longer will the United States influence choices and policy options as in the past. A new generation of executive directors and employees will debate such future Bank directions internally. External critics will again evaluate whether the Bank should finally cease and desist. Ultimately, however, international institutions have an instinct for survival. The World Bank is probably no exception.

See Also

- foreign aid
- international capital flows
- International Monetary Fund

Bibliography

Akyüz, Y. 2004. The rationale for multilateral lending: a critical assessment. Unpublished manuscript, July.

Asher, R.B. and Edward, M. 1973. *The World Bank Since Bretton Woods*. Washington, DC: Brookings Institution.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Gavin, M. and Dani, R. 1995. The World Bank in historical perspective. *American Economic Review* 85, 329–34.

Gill, I.S. and Todd, P. 2005. *At the Frontlines of Development: Reflections from the World Bank*. Washington, DC: World Bank.

Kagia, R. 2005. *Balancing the Development Agenda: The Transformation of the World Bank under James D. Wolfensohn, 1995–2005*. Washington, DC: World Bank.

Kapur, D. 2002. Do as I say not as I do: a critique of G7 proposals on reforming the MDBs. Working Paper No. 16. Washington, DC: Center for Global Development.

Kapur, D., John, L. and Richard, W. 1997. *The World Bank: Its First Half-Century*. Washington, DC: Brookings Institution.

Krueger, A.O. 1998. Wither the World Bank and the IMF? *Journal of Economic Literature* 36, 1983–2020.

Mason, E.S. and Asher, R.B. 1973. *The World Bank Since Bretton Woods*. Washington, DC: Brookings Institution.

Meltzer, A. et al. 2000. *International Financial Institution Advisory Commission Report*. Washington, DC.

Mohammed, A.A. 2004. Who pays for the World Bank? G24 Discussion Paper.

Pincus, J.R. and Jeffrey, A.W. 2002. *Reinventing the World Bank*. Ithaca: Cornell University Press.

How to cite this article

Fishlow, Albert. "World Bank." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_W000096> doi:10.1057/9780230226203.1842

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

World Trade Organization

Robert W. Staiger
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The success of the General Agreement on Tariffs and Trade (GATT)/World Trade Organization (WTO) as an international institution is widely acknowledged. Among multilateral institutions, the GATT/WTO has adopted a distinctive approach as a forum for international negotiation, based on reciprocal negotiations (over market access) that occur on a voluntary basis between pairs of countries or among small numbers of countries; the results of these bilateral negotiations are then ‘multilateralized’ to the full GATT/WTO membership under the GATT/WTO principle of non-discrimination. This article describes how recent economic research has attempted to understand and interpret these key design features of the GATT/WTO.

Keywords

commitment theories of trade agreements; cost shifting; free trade; General Agreement on Tariffs and Trade (GATT); mercantilism; most favoured nation (MFN); non-discrimination in trade; protection; reciprocity in trade; tariffs; terms-of-trade theories of trade agreements; unilateral and multilateral trade policies; World Trade Organization (WTO)

Article

The World Trade Organization (WTO), like its predecessor the General Agreement on Tariffs and Trade (GATT), has in effect served as the constitution of the post-war international trading system. (The GATT was created in 1947, and the WTO came into existence on 1 January 1995, as a result of the Marrakesh Agreement of April 1994, also known as the WTO Agreement. The WTO Agreement includes the text of GATT: GATT therefore continues to exist as a substantive agreement, but the WTO Agreement also includes a set of additional agreements that build on and extend GATT principles to new areas. Hoekman and Kostecki, 1995, provide an excellent institutional overview of GATT and the WTO.) Since 1947, membership in the GATT/WTO has grown from 23 countries to its present size of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
150 countries, and according to the WTO's World Trade Report (WTO, 2007), average *ad valorem* tariffs on industrial goods have been reduced from upwards of 30 per cent to below four per cent through eight multilateral rounds of negotiation (a ninth, the Doha round, is ongoing at this writing). The success of the GATT/WTO as an international institution is widely acknowledged. Among multilateral institutions, the GATT/WTO has adopted a distinctive approach to serving as a forum for international negotiation. This approach is based on reciprocal negotiations (over market access) that occur on a voluntary basis between pairs of countries or among small numbers of countries, and the results of these bilateral negotiations are then 'multilateralized' to the full GATT/WTO membership under the GATT/WTO principle of non-discrimination.

As an object of study, the GATT/WTO has attracted the attention of legal scholars since the late 1960s. But until relatively recently, the GATT/WTO has not been the subject of systematic and formal economic analysis. This might seem surprising, because the familiar economic arguments for free trade would seem to provide an obvious foundation for the economic analysis of the GATT/WTO. But this foundation immediately runs into a pair of impediments. First, the case for free trade is a unilateral case, and it therefore leaves no room for the existence of a trade agreement of any kind: from this starting point, the economic logic of the GATT/WTO is immediately suspect. And second, the liberalizing force that the GATT/WTO has harnessed does not appear to be the consumer gains that come from freer trade: rather, the GATT/WTO is driven by exporter interests. Traditionally, most economists have interpreted these observations as evidence that a mercantilist logic lies at the foundation of the GATT/WTO and that, as a result, economic analysis of the GATT/WTO is futile.

A growing body of theoretical and empirical literature has begun to challenge this view. There are two main branches of this literature (for recent attempts to articulate theories that would constitute a third branch, see Ethier, 2006 and Regan, 2006). A first branch (terms-of-trade theories) emphasizes the role of trade agreements in providing governments with an avenue of escape from a terms-of-trade driven Prisoner's Dilemma. A second branch (commitment theories) emphasizes the role of trade agreements in providing governments with a means of making commitments to their private sectors. Commitment theories of trade agreements have been developed by a number of authors, and there is also some empirical evidence that the GATT/WTO may play this role (see, for example, Conconi and Perroni, 2003; Maggi and Rodriguez-Clare, 1998; 2007; and Staiger and Tabellini, 1987; 1999). But most of the literature to date adopts the terms-of-trade perspective. So I will focus here on interpreting and evaluating some of the key design features of the GATT/WTO from the perspective of terms-of-trade theories. (Empirical evidence relating to the terms of trade theory of trade agreements is surveyed in Bagwell and Staiger, 2002, ch. 11. More recent evidence appears in Broda, Limao and Weinstein, 2006; and Bagwell and Staiger, 2006a.)

All theories of trade agreements must identify a means by which the negotiating governments can gain from the agreement. This entails identifying a 'problem' that would arise in the absence of an agreement, when governments make unilateral trade policy choices. The purpose of a trade agreement can then be viewed as providing a 'solution' to the problem, so that the negotiating governments may share in the associated benefits. The terms-of-trade theory posits that governments can gain from negotiations by correcting the international inefficiencies that occur under unilateral trade policy choices as a result of international cost shifting. This cost shifting arises whenever the government of an importing country increases its import barriers and the prices received by foreign exporters fall as a

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
result, thereby improving the importing country's terms of trade. In this way, a portion of the cost of each government's import protection is borne by foreigners, and as a consequence the unilateral best-response levels of import protection chosen by each government are overly restrictive relative to internationally efficient levels: starting from its best-response (reaction curve) tariffs, each government can therefore gain by negotiating reciprocal liberalization with its trading partners. From the perspective of the terms-of-trade theory, then, the problem associated with unilateral trade policy choices is the cost shifting that importing governments are able to achieve on to foreign exporters; and the purpose of negotiated trade agreements is to give foreign exporters (or their governments) a 'voice' in the trade policy choices of importing governments, so that the 'market access' that each country affords its trading partners can be expanded to internationally efficient levels. (The link between the terms-of-trade theory of trade agreements and the emphasis on market access found in GATT/WTO discussions is identified and formalized in Bagwell and Staiger, 2002, ch. 2.)

In this environment, internationally efficient policies can be achieved if each government agrees to adopt the policies it would have chosen had it 'ignored' its ability to shift costs on to foreigners. Accordingly, internationally efficient levels of market access may be delivered under multilateral free trade, but only if all governments seek to maximize national income with their trade policy choices: when governments have broader (for example, political/distributional) goals, international efficiency will generally *not* correspond to free trade. Nevertheless, according to the terms-of-trade theory, the purpose of a trade agreement remains the same independent of government objectives. This feature suggests that, despite the potential for wide diversity across the objectives of GATT/WTO member governments, the underlying structure of the cost-shifting problem central to the terms-of-trade theory may yield simple and robust insights concerning the logic of key design features of the GATT/WTO.

I now illustrate the basic structure of the international cost-shifting problem at the heart of the terms-of-trade driven Prisoner's Dilemma, and describe how it can account for two pillars of the GATT/WTO: reciprocity and non-discrimination. Broadly speaking, the principle of reciprocity in the GATT/WTO refers to the ideal of mutual changes in trade policy that trigger changes in the volume of each country's imports that are of equal value to changes in the volume of its exports. And according to the non-discrimination principle, a country must provide every other GATT/WTO member country with access to its markets on terms no less favourable than it provides the 'most-favoured' country: hence, under the non-discrimination principle, each GATT/WTO member country faces 'most-favoured-nation' (MFN) tariffs from all other GATT/WTO member countries.

I begin with reciprocity. The essential point can be understood from the perspective of a standard two-country/two-good competitive general equilibrium trade model, in which country *A* exports good *y* to country *B* in exchange for imports of good *x*. Following Bagwell and Staiger (1999; 2002), government preferences for the two countries can be represented very generally by the functions $W^i(p^i(\tau^i, \mathcal{P}^W), \mathcal{P}^W)$, where τ^i is 1 plus the ad valorem tariff in country $i \in \{A, B\}$, p^i is the relative price of good *x* to good *y* prevailing locally in country *i*, and $\mathcal{P}^W$ is the market-clearing 'world' relative price or terms of trade, which is itself a function of the two tariffs $\mathcal{P}^W(\tau^A, \tau^B)$. Under standard conditions $\mathcal{P}^W$ is decreasing in τ^A and increasing in τ^B, while p^A is increasing in τ^A and p^B is decreasing in τ^B. Apart from general concavity, the only condition that is imposed on government

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

welfare functions is that $W^A_{\tilde{p}^W} < 0$ and $W^B_{\tilde{p}^W} > 0$, meaning that each government would like more tariff revenue if it could have this without any change in its local prices (and therefore without any change in the distribution or levels of factor incomes within its economy). Because no restrictions are placed on the way in which governments feel about changes in local prices, this representation of government preferences is general enough to include, in addition to the traditional Johnson (1953–4) national-income maximizing government, the leading models of political economy of trade protection (each of which effectively defines government preferences over redistribution and hence local prices).

The non-cooperative (Nash) tariffs chosen in this environment are defined by the two first-order conditions $W^i_{p^i} + \lambda^i W^i_{\tilde{p}^W} = 0$ for $i \in \{A, B\}$, where $\lambda^i \equiv [\partial \tilde{p}^W / \partial \tau^i] / [dp^i / d\tau^i] < 0$. Notice that international cost shifting is embodied in the term $\lambda^i W^i_{\tilde{p}^W}$ which enters into the first-order conditions, and the presence of this cost-shifting term guarantees that $W^A_{p^A} < 0$ and $W^B_{p^B} > 0$ in the Nash equilibrium. The international efficiency frontier is defined by the (τ^A, τ^B) pairs from which it is not possible to adjust tariffs so as to help one country without hurting the other according to the government preferences W^A and W^B. Formally, this frontier takes the form $(1 - \Lambda^A W^A_{\tilde{p}}) = 1 / (1 - \Lambda^B W^B_{\tilde{p}})$, where $\Lambda^A \equiv (1 - \tau^A \lambda^A) / (W^A_{p^A} + \lambda^A W^A_{\tilde{p}^W})$ and $\Lambda^B \equiv (1 - \lambda^B / \tau^B) / (W^B_{p^B} + \lambda^B W^B_{\tilde{p}^W})$. From these expressions, a pair of observations can now be confirmed. First, the Nash tariff choices do not achieve the international efficiency frontier, and so there is indeed a ‘problem’ for an international agreement to solve. And second, politically optimal tariffs, defined by $W^i_{p^i} = 0$ for $i \in \{A, B\}$ and interpreted as the unilateral tariff choices governments would make if they were not motivated by terms-of-trade considerations, do achieve the international efficiency frontier, and so politically optimal tariffs represent a complete ‘solution’ to this problem. From these observations an important conclusion can be drawn: even in the presence of politically/distributionally motivated governments, the purpose of a trade agreement is simply to prevent terms-of-trade manipulation.

From this backdrop, we may now ask the question, ‘Why would the principle of reciprocity have appealing features?’ The answer, simply stated, is that reciprocity describes a fixed-terms-of-trade rule to which mutual tariff changes must conform. (Formally, this can be seen following Bagwell and Staiger, 1999; 2002. Define a set of tariff changes $\Delta\tau^A \equiv (\tau^A_1 - \tau^A_0)$ and $\Delta\tau^B \equiv (\tau^B_1 - \tau^B_0)$ as conforming to reciprocity whenever $\tilde{p}^W_0 [M^A(p^A_1, \tilde{p}^W_1) - M^A(p^A_0, \tilde{p}^W_0)] = [E^A(p^A_1, \tilde{p}^W_1) - E^A(p^A_0, \tilde{p}^W_0)]$, where $\tilde{p}^W_0 \equiv \tilde{p}^W(\tau^A_0, \tau^B_0)$, $\tilde{p}^W_1 \equiv \tilde{p}^W(\tau^A_1, \tau^B_1)$, $p^A_0 \equiv p^A(\tau^A_0, \tau^B_0)$ and $p^A_1 \equiv p^A(\tau^A_1, \tau^B_1)$, and where M^A and E^A denote A's imports and exports, respectively. Using balanced trade $(\tilde{p}^W M^A(p^A, \tilde{p}^W) = E^A(p^A, \tilde{p}^W))$, the condition for reciprocity simplifies to the fixed-terms-of-trade rule $[\tilde{p}^W_1 - \tilde{p}^W_0] M^A(p^A_1, \tilde{p}^W_1) = 0$.)

And in an environment where terms-of-trade manipulation is *the* problem to be fixed, a fixed-terms-of-trade rule is bound to have some attractive uses. (Bagwell and Staiger, 2002, ch. 4 describe and interpret a number of ways in which the principle of reciprocity appears in the GATT/WTO.) Intuitively, the nature of international cost shifting ensures that, beginning from their Nash tariff choices, each

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
government would desire tariff liberalization and the local price movements/greater trade volume that this would bring if this liberalization could be achieved at a fixed terms of trade p^W (that is, recall from above that $W^A_{p^A} < 0$ and $W^B_{p^B} > 0$ at Nash). The principle of reciprocity can be understood to harness this desire, and so to activate efficiency-enhancing tariff-liberalizing forces in this environment.

I now turn to the non-discrimination principle, as embodied in MFN. This requires an extension of the basic two-country model described above to a three-country setting. To this end, let country *C* have a similar trading pattern to *B*, in that *C* also exports good *x* to country *A* in exchange for imports of good *y*. An important feature of the MFN rule is that, in requiring country *A* to impose a common tariff on imports of *x* regardless of whether these imports of *x* originate in exporting country *B* or *C*, this rule ensures that a single market-clearing terms of trade $p^W(\tau^A, \tau^B, \tau^C)$ will prevail, and government preferences may continue to be expressed with the simple representation $W^i(p^i(\tau^i, p^W), p^W)$ for $i \in \{A, B, C\}$. Notice that, in the presence of MFN, countries *A* and *B* can still negotiate a reciprocal reduction in their respective tariffs τ^A and τ^B that provides each with more trade volume at a fixed terms of trade p^W, thereby ensuring that they each gain relative to Nash; and strikingly, as long as *A* and *B* abide by reciprocity, there will be no third-party effects of their bilateral negotiation on country *C*, whose welfare level $W^i(p^C(\tau^C, p^W), p^W)$ remains unaltered owing to the unchanged τ^C and the fixed terms of trade p^W. (For *C*'s welfare to remain unchanged, it is in fact not necessary that τ^C remain unchanged, but only that *C* remain on its tariff reaction curve and that p^W remain unchanged: see Bagwell and Staiger, 2006b.) Of course, *A* and *C* can engage in bilateral reciprocal negotiations that have the same property. This has an important implication: the MFN rule permits the liberalizing force of reciprocity to be harnessed in an essentially bilateral manner even in a multilateral world. (These and related points are developed in Schwartz and Sykes, 1997, and Bagwell and Staiger, 2005; 2006b.)

In this general manner, the GATT/WTO pillars of reciprocity and non-discrimination can be understood to underpin the architecture of an international negotiating forum in which the liberalizing force of reciprocity can be harnessed in bilateral negotiations with an assurance of minimal third-party spillovers, thereby permitting each member government – through a sequence of bilateral or small-numbers negotiations – to engineer its escape from a terms-of-trade driven Prisoner's' Dilemma.

See Also

- tariffs
- trade policy, political economy of

Bibliography

Bagwell, K. and Staiger, R.W. 1999. An economic theory of GATT. *American Economic Review* 89, 215–48.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bagwell, K. and Staiger, R.W. 2002. *The Economics of the World Trading System*. Cambridge, MA: MIT Press.

Bagwell, K. and Staiger, R.W. 2005. Multilateral trade negotiations, bilateral opportunism and the rules of GATT/WTO. *Journal of International Economics* 67, 268–94.

Bagwell, K. and Staiger, R.W. 2006a. What do trade negotiators negotiate about? Empirical evidence from the World Trade Organization. Working Paper No. 12727. Cambridge, MA: NBER.

Bagwell, K. and Staiger, R.W. 2006b. Backward stealing and forward manipulation in the WTO. Mimeo. Columbia University, Stanford University and University of Wisconsin.

Broda, C., Limao, N. and Weinstein, D. 2006. Optimal tariffs: the evidence. Working Paper No. 12033. Cambridge, MA: NBER.

Conconi, P. and Perroni, C. 2003. Self-enforcing international agreements and domestic policy credibility. Working Paper No. 114/03. Warwick University Center for the Study of Globalization and Regionalization.

Ethier, W. 2006. The theory of trade policy and trade agreements: a critique. PIER Working Paper 06–013. University of Pennsylvania.

Hoekman, B. and Kostecki, M. 1995. *The Political Economy of the World Trading System*. Oxford: Oxford University Press.

Johnson, H.G. 1953–4. Optimum tariffs and retaliation. *Review of Economic Studies* 1, 142–53.

Maggi, G. and Rodriguez-Clare, A. 1998. The value of trade agreements in the presence of political pressures. *Journal of Political Economy* 106, 574–601.

Maggi, G. and Rodriguez-Clare, A. 2007. A political-economy theory of trade agreements, *American Economic Review* 97, 1374–406.

Regan, D.H. 2006. What are trade agreements for? Two conflicting stories told by economists, with a lesson for lawyers, *Journal of International Economic* Law 9, 951–88.

Schwartz, W.F. and Sykes, A.O. 1997. The economics of the most-favored-nation clause. In *Economic Dimensions in International Law: Comparative and Empirical Perspectives*, ed. J.S. Bhandani and A.O. Sykes. Cambridge: Cambridge University Press.

Staiger, R.W. and Tabellini, G. 1987. Discretionary trade policy and excessive protection. *American*

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Economic Review 77, 823–37.

Staiger, R.W. and Tabellini, G. 1999. Do GATT rules help governments make domestic commitments? *Economics and Politics* 11, 109–44.

WTO (World Trade Organization). 2007. *World Trade Report*. Geneva: WTO.

How to cite this article

Staiger, Robert W. "World Trade Organization." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000189>
doi:10.1057/9780230226203.1843

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

World Wars, economics of

Stephen Broadberry and Mark Harrison
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

This article focuses on the role of economic factors in explaining the outcomes of the two world wars. In both wars, the scale of resources mobilized was decisive, leaving little room for other factors that feature prominently in narrative accounts, such as national differences in war preparations, war leadership, military organization and morale. The economic advantage of the Allies was not just in size, but also in the quality of their resources, reflected in average real incomes per head of their populations before the wars. We also quantify the economic effects of the wars within a national balance sheet framework.

Keywords

coalitions; collectivization; famine; fiscal mobilization; globalization; Lend-Lease; peasant economy; rationing; subsistence; tariffs; world wars, economics of the; war and economics

Article

The two world wars of the 20th century were events in a single process of reaction against globalization that was prolonged and, from time to time, violent.
From 1815 to 1914 trade and capital flows increased alongside global productivity. Everywhere, economic development tended to reduce local risks. At the same time, falling trade and transport costs exposed farmers, firms, and labouring households to new instabilities and risks that originated far away, in countries and markets across the world. Where governments and politicians embraced these long-range risks, liberalization fostered engagement in the global economy. Where political entrepreneurs mobilized reaction against them, however, resistance gained ground.
By the end of the 19th century, leaders of several newly industrializing countries were seeking to insulate their economies from global risks through tariff protection. German leaders, for example, aimed at national security through trade within a closed region based on a colonial empire. To secure this empire they launched a naval arms race; the arms race precipitated the formation of two Eurasian alliances that confronted each other in the First World War. On one side stood the Central Powers, primarily the German, Austro-Hungarian and Ottoman empires, joined in 1915 by Bulgaria; on the other side stood the Allies: the British, French and Russian empires, joined in 1915 by Italy, and in 1917 by the United States. But the war brought ruin to the three empires of the Central Powers and to the Russian empire too.
After the First World War, the instabilities intrinsic to the global economic order increased. The weakness of the formerly dominant British economy and the isolation of Germany and Russia undermined global market integration. The slump of 1929 sent deflationary ripples around the world and accelerated the disintegration. As the world market shrank, the great powers struggled over national shares. In the 1930s the world economy broke up into several relatively closed trading blocs. The British, French and Dutch reorganized their trade on colonial lines. With Hitler in power, Germany resumed the perspective of regulated trade within a colonial region in central and eastern Europe, and this led to rivalry with other interested regional powers. Italy established bilateral trade with the smaller states of the former Austro-Hungarian empire, and also set about winning an African empire. The Japanese competed with the Americans, the Dutch, the British and the Soviet Union for influence in east Asia and the Pacific. The Soviet Union developed a closed economic space behind the frontiers of the former Russian empire, and defended it against the Japanese.
The worldwide trade disintegration contributed to the causes of the Second World War. The economies of the Axis powers, Germany, Italy and Japan, were too small to prosper without specialization and external sources of food, fuel and other materials. A common thread in their course of external aggression was the attempt to secure these supplies by

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
imposing a colonial regime upon trading partners. In this sense the national policies that led to the Second World War were a continuation of those that had led to the First World War. The second war continued some of the themes of the first, but it was not just a repeat. The object of the first war was regional – to control Europe, the Atlantic and the Near East. The second war was a struggle for global domination in the full sense. The first war was certainly fuelled by racial identities, but no one aimed at genocide, as they did in the second. The first war ended inconclusively, with a ceasefire and a peace treaty that tried to punish the aggressors, but there was no unconditional surrender, and in Germany those who wanted to try again eventually took power. In 1945 the aggressors were crushed militarily and morally, their surviving leaders were put on trial, and what they stood for was excluded from public life.

In this article we pursue the similarities and differences of the two wars in terms of economic history. We have two main themes. First, what is the power of the economic factors compared with others in helping to explain the outcomes of the two wars? Second, of the possible economic factors that should be considered, which contribute most to the explanation of the results? These are not new questions, of course; here, we outline briefly some alternative views.

Historians of the two world wars tend to narrate their story as a complex interplay of forces that worked at many levels. They tell a story of warfare that was increasingly mechanized and waged for years on end by massed forces. Nonetheless, war was waged by people, not by numbers. Economists, in contrast, have tended to give the centre stage to the numbers, conceding less to aspects of warfare such as leadership, discipline, heroism and villainy. Raymond Goldsmith (1946, p. 69), an economist who helped to manage the United States war economy in his youth, once observed that:

> The cold figures of the output of airplanes, tanks, guns, naval ships, and ammunition, particularly when they are reduced to the still colder form of indices of aggregate munitions production of the major belligerents, probably tell the story of [the Second World War] as well as extended discussion or elaborate pictures … They back to the full the thesis, dear to the economist's ear, that whatever may have saved the United Nations from defeat in the earlier phases of the conflict, what won the war for them in the end was their ability – and particularly that of the United States – to produce more, and vastly more, munitions than the Axis.

To many historians this view remains unappealing; Richard Overy (1998), for example, has objected that it leaves no room for ‘a whole series of contingent factors – moral, political, technical, and organizational – [that] worked to a greater or lesser degree on national war efforts’.

The opposition between cold figures and hot blood is false to some extent. Of course, leadership and psychology mattered. But they mattered less than in previous eras because they had become problems that both sides could solve. In both world wars, multi-million armies took the field and stayed there for months and years, giving and taking appalling losses, without disintegrating. Since the moral fabric of military life could withstand the pressure, numbers of men and the volume of supplies assumed the decisive role.

If economics did matter, exactly what was it about the economies of the Allies that gave them superiority? In Goldsmith's tradition, size mattered and only size. Niall Ferguson is a historian who gives economics the attention it deserves. Noting the overwhelming size advantage of the Allies in population and production on the eve of the First World War, he remarks (1998, p. 248), ‘To the economic historian, the outcome of the First World War looks to have been inevitable from the moment [the British] opted for intervention’. Given this advantage, he argues, the war should have been over quickly; the only explanation for the Allied failure to conclude the war much sooner is Allied mismanagement, so Ferguson concludes that the Allied economic preponderance was ‘an advantage squandered’. As a result, economic advantage came into play only after much time had passed and the military advantage of the aggressors had almost won the day.

There is much truth in this, but we will take a more nuanced view of what it was about economic life that could be decisive in warfare. The belligerents’ economies differed not only in the volume of national resources but also in their quality. The main factor in quality was the level of peacetime economic development, which we measure by average real incomes per head of the population. Richer countries could mobilize production, public finance, soldiers and weapons out of proportion to their general economic capacities; in other words, the level of economic development acted as a multiplier of size. For Britain in both world wars, control of the vast but impoverished population and territory of India, for example, mattered little compared with access to the rich markets of the United States.

The First World War

From an economic viewpoint, the First World War can be divided into two phases. In the late summer of 1914, both sides hoped for a quick victory with a limited commitment of resources to the war effort. This first phase is summed up in the memorable phrase ‘business as usual’, which was common currency in Britain at the time (Lloyd, 1924). It was hoped that the war could be fought along similar lines to previous centuries, with a clear distinction between soldiers doing the fighting and civilians getting on with normal life. However, from late 1916 both Britain and Germany stepped up mobilization in the direction of ‘total war’. In total war, industry was mobilized to provide unprecedented amounts of munitions, and industrial workers became as vital to the war effort as soldiers. During this second phase, keeping up production and avoiding economic collapse became central to management of the war. The first economy to collapse was on the Allied side; the Bolshevik Revolution of 1917 took Russia out of the war and led to a Soviet republic (Gatrell, 2005). In 1918, falling output in Turkey, Austria and Germany led to the collapse of the Central Powers and the break-up of their empires (Pamuk, 2005; Schulze, 2005; Ritschl, 2005). France also suffered a late collapse of output, but was shored up by the other Allies (Hautcoeur, 2005).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Size and development

The Allies mobilized more soldiers and produced more of most weapons than the Central Powers, as can be seen in Table 1. Furthermore, the degree of Allied superiority increased with the complexity of the weapons. Only in guns did the Central Powers have numerical superiority, while the Allied superiority in tanks reached a factor of nearly 90:1.

Allies vs. Central Powers: soldiers and equipment in the First World War

	Allies	Central Powers	Ratio, 1:2
	(1)	(2)	(3)
Soldiers mobilized, million	41.0	25.6	1.6
Weapons produced:			
Guns, thousand	59.9	82.4	0.7
Rifles, million	13.3	12.1	1.1
Machine guns, thousand	656	319	2.1
Aircraft, thousand	124.5	47.3	2.6
Tanks	8,919	100	89.2

Source: Broadberry and Harrison (2005).

Table 2 shows that the balance between the two sides varied over time, as the alliances' compositions changed. In 1914, the Triple Entente of the United Kingdom, France and Russia could also draw on their colonies, and were joined by other countries including Serbia, the British Dominions, Liberia and Japan. By November 1916, the Allies had been joined by a second wave of countries, including Italy, Portugal and Rumania. By November 1918, although Russia had dropped out, the Allies had been strengthened by the United States and a further wave of countries. By this time, the Allied side included 70 per cent of the world's pre-war population and 64 per cent of its pre-war output. The scale of resources that could be mobilized by the Central Powers varied less over time. Austria-Hungary started the war, joined immediately by Germany and shortly after by the Ottoman Empire. By November 1915, Bulgaria had also joined, but Italy, defaulting on its treaty obligations, joined the Allies.

The alliances in the First World War: resources of 1913

	Population, million	Territory		GDP in 1990 prices	
		million sq. km	ha. per head	$ billion	per head, $
Allies					
November 1914					
Allies, total	793.3	67.5	8.5	1,093.6	1379
UK, France and Russia only	259.0	22.6	8.7	622.8	2405
November 1916					
Allies, total	853.3	72.5	8.5	1,210.5	1419
UK, France and Russia only	259.0	22.6	8.7	622.8	2405
November 1918					
Allies, total	1,271.7	80.9	6.4	1,760.6	1384
UK, France and USA only	182.3	8.7	4.8	876.6	4809
Central Powers					
November 1914					
Central Powers, total	151.3	5.9	3.9	376.6	2489

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Germany and Austria-Hungary only	117.6	1.2	1.0	344.8	2933
November 1915					
Central Powers, total	156.1	6.0	3.8	383.9	2459

Source: Broadberry and Harrison (2005).

It is important to consider the level of economic development of individual countries as well as the volume of output that the two alliances could draw upon. Britain, for example, with a prewar population of 46 million, had average incomes of nearly $5,000 (at 1990 prices), but its colonies, excluding the Dominions, had a pre-war population of 380 million, mostly in India, with average incomes of less than $700. Thus the colonies, with nearly eight times the population of Britain, produced only about the same volume of output. But the colonial output was less available for fighting the Germans because most of it was needed to meet the subsistence needs of the colonial population. Furthermore, this population was difficult to mobilize because of its distance from the theatre of war and the level of development of colonial administration. Even within the Triple Entente, the low level of development in Russia limited the Allied mobilization. The Central Powers were similarly hampered by the low level of development of the Ottoman Empire and Bulgaria, and even the Hungarian half of the Habsburg Empire.

By comparing the information for the two alliances in Table 2, it is possible to calculate size and development ratios for three benchmark dates in Table 3. The ratios are calculated for each alliance as a whole, and also for the great powers only. The rationale for the latter is that if, as we argue, poor colonies did not count for much, it is helpful to see how the ratios look if we do not count them at all. The table establishes a striking result: judging by economic size, the Central Powers were doomed to defeat. In November 1914, the Allies had access to five times the population, 11 times the territory and three times the output of the Central Powers. If we look only at great powers, the Allied advantages in population and output were smaller, but larger in territory, reflecting the fact that German and Turkish colonies tended to be in the sandy deserts of Africa and the Middle East. However, the Allied advantage was limited by relatively low average incomes in Russia and the British and French colonies. Allied incomes were less than two-thirds the average level of the Central Powers, or 80 per cent if attention is confined to the great powers, if Russia is counted as a great power.

Allies Versus Central Powers: size and development ratios

	Population	Territory	Territory per head	Gross domestic product	GDP per head
November 1914					
Total	5.2	11.5	2.2	2.9	0.6
Great powers only	2.2	19.4	8.8	1.8	0.8
November 1916					
Total	5.5	12.1	2.2	3.2	0.6
Great powers only	2.2	19.4	8.8	1.8	0.8
November 1918					
Total	8.2	13.5	1.7	4.6	0.6
Great powers only	1.6	7.5	4.8	2.5	1.6

Source: Calculated from Table 1.

By November 1916 the Allied advantage had grown moderately in terms of population, territory and output, but the Central Powers continued to have an advantage in average incomes. By November 1918, however, the situation had changed dramatically, largely as a result of the United States replacing Russia. The Allied advantages in population, territory and output all increased markedly, and for the first time the Allies enjoyed an average income advantage if attention is restricted to great powers. Although it took some time for the American presence to be felt on the battlefield, it sealed the Central Powers' fate.

Development and mobilization

The ratios in Table 3 are based on the assumption that during the war the real output of a given country did not change. The reason for this assumption is statistical: it is difficult to track GDP changes in wartime in the poorer countries. What information we have suggests that Table 3 must understate the actual swing in favour of the Allies during the war, because output increased in the United States and Britain but fell in the less developed economies of the Central Powers. This can be seen in Figure 1, which plots the change in GDP during 1913–17 against the level of per capita income in 1913 for nine countries. The relationship is strongly positive, reflecting the fact that rich countries were better able to mobilize output than poor countries. The biggest decline was in Russia, which was also the poorest amongst these countries in 1913, and collapsed in the Bolshevik Revolution of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
1917. There were also substantial falls in output in Austria-Hungary, France and Germany, which all started the war with lower average incomes than Canada, Britain, New Zealand and the United States.

Figure 1

Production mobilization: nine countries, 1913–1917. *Notes*: Observations are, from left to right, Russia, Austria-Hungary, France, Germany, Canada, UK, New Zealand, USA and Australia. *Source*: Broadberry and Harrison (2005).

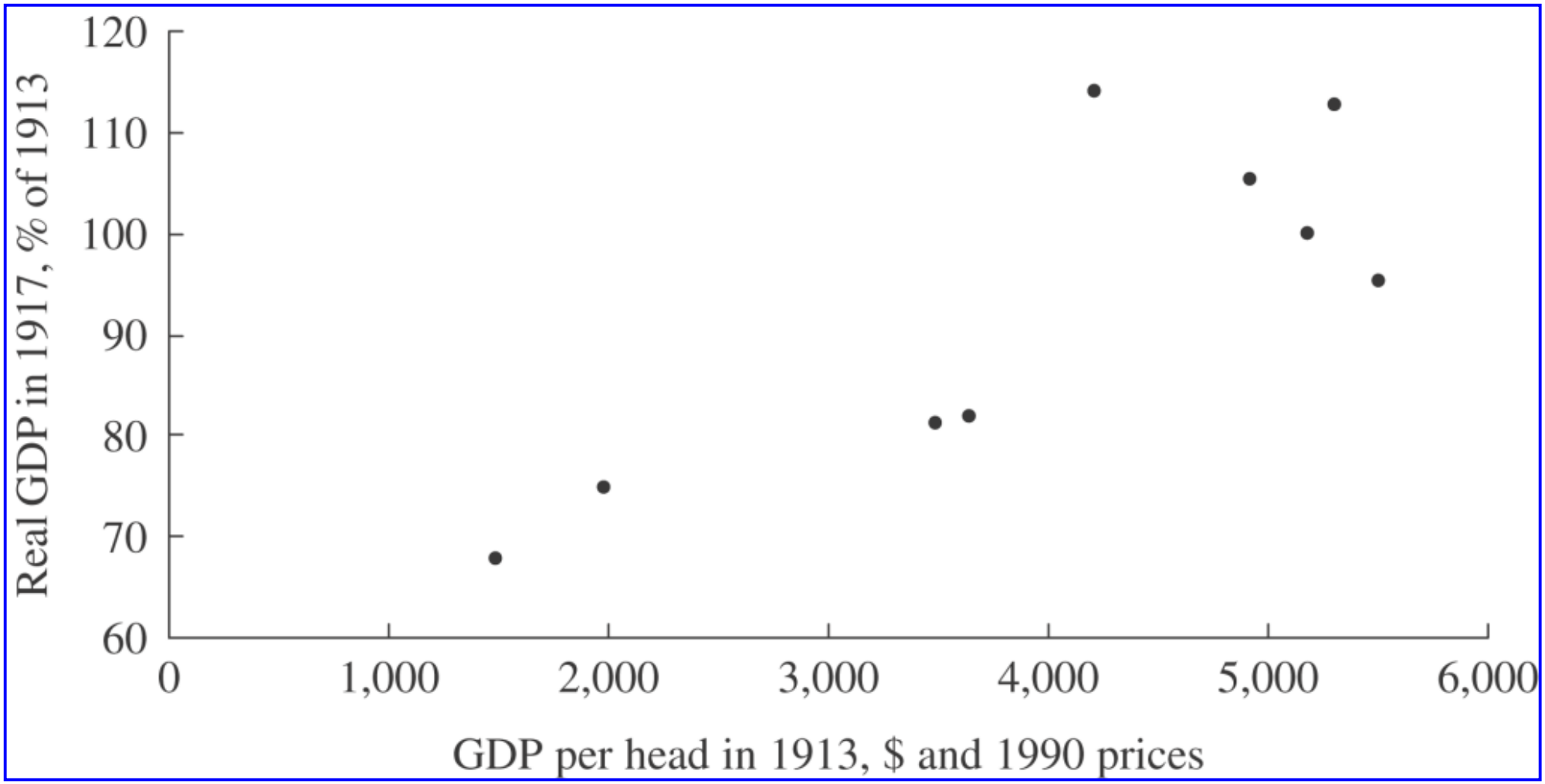

Another measure of mobilization which varied with the level of development is the change in government outlays as a share of GDP. This reflects the extent to which governments were able to convert output from peacetime uses to the war effort through taxation and spending. Figure 2 plots this measure of fiscal mobilization during the first year of the war against pre-war average incomes for eight countries. The relationship is positive, but not as clear cut as for production mobilization. In particular, it is necessary to control for distance from the main theatre of war in Europe, with the New World countries of Canada, Australia and the United States mobilizing fewer resources through taxation and public spending than the European countries.

Figure 2

Fiscal mobilization in the First World War: eight countries. *Notes*: Observations not labelled within the figure are, from left to right, Austria-Hungary, Italy, France, Germany and UK. *Source*: Broadberry and Harrison (2005), supplemented by Austria-Hungary from Schulze (2005).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

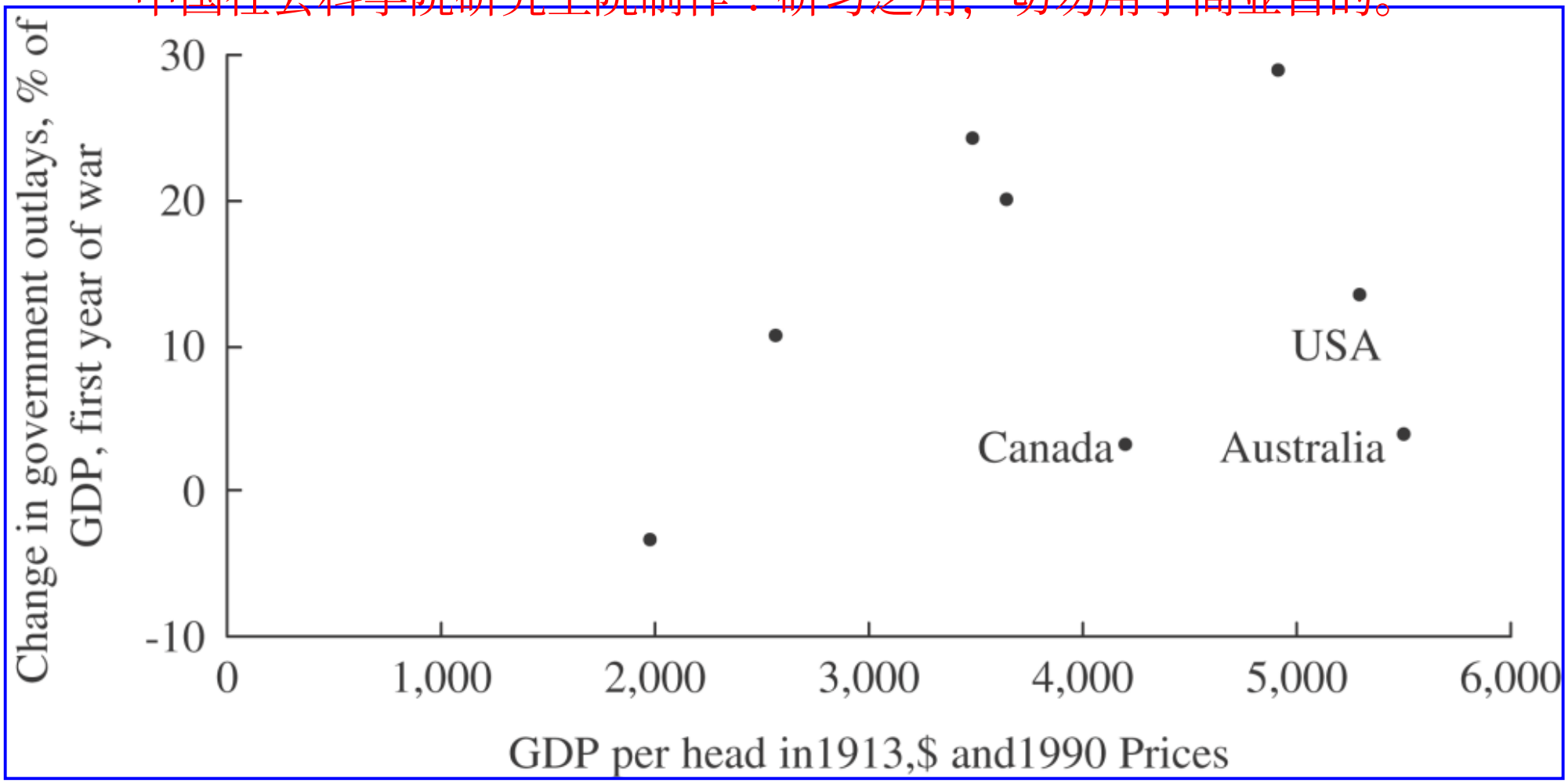

Perhaps the most direct measure of mobilization is the share of the prime-age male population recruited into the military. This measure, plotted in Figure 3 against average incomes in 1913, is available for a relatively large sample of countries. Again, we find a relationship that increases with pre-war prosperity and decreases with distance from the main theatre of war. The figure is plotted in three distance bands, comprising the frontline Eurasian states, peripheral European countries isolated from the frontline by land or sea (Britain and Portugal), and non-European states. Cumulative numbers mobilized are shown as a proportion of males aged 15–49. After we have controlled for distance (that is, within each distance band), there is a positive relationship between military mobilization and the level of development. But dropping a band also lowers the mobilization rate substantially.

Figure 3

Military mobilization in the First World War: 18 countries and the French colonies. *Note*: Observations are, from left to right: Front line Eurasia: Serbia, Turkey, Russia, Bulgaria, Roumania, Greece, Austria-Hungary, Italy, France, and Germany. European periphery: Portugal and UK. Non-European states: French colonies, India, South Africa, Canada, New Zealand, USA, Australia. *Sources*: GDPs per head in 1913 from Tables 1 and 2 or, if not listed there, from Maddison (2001, p. 185); cumulative mobilization rates, 1914–1918, from Urlanis (1971, p. 209).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

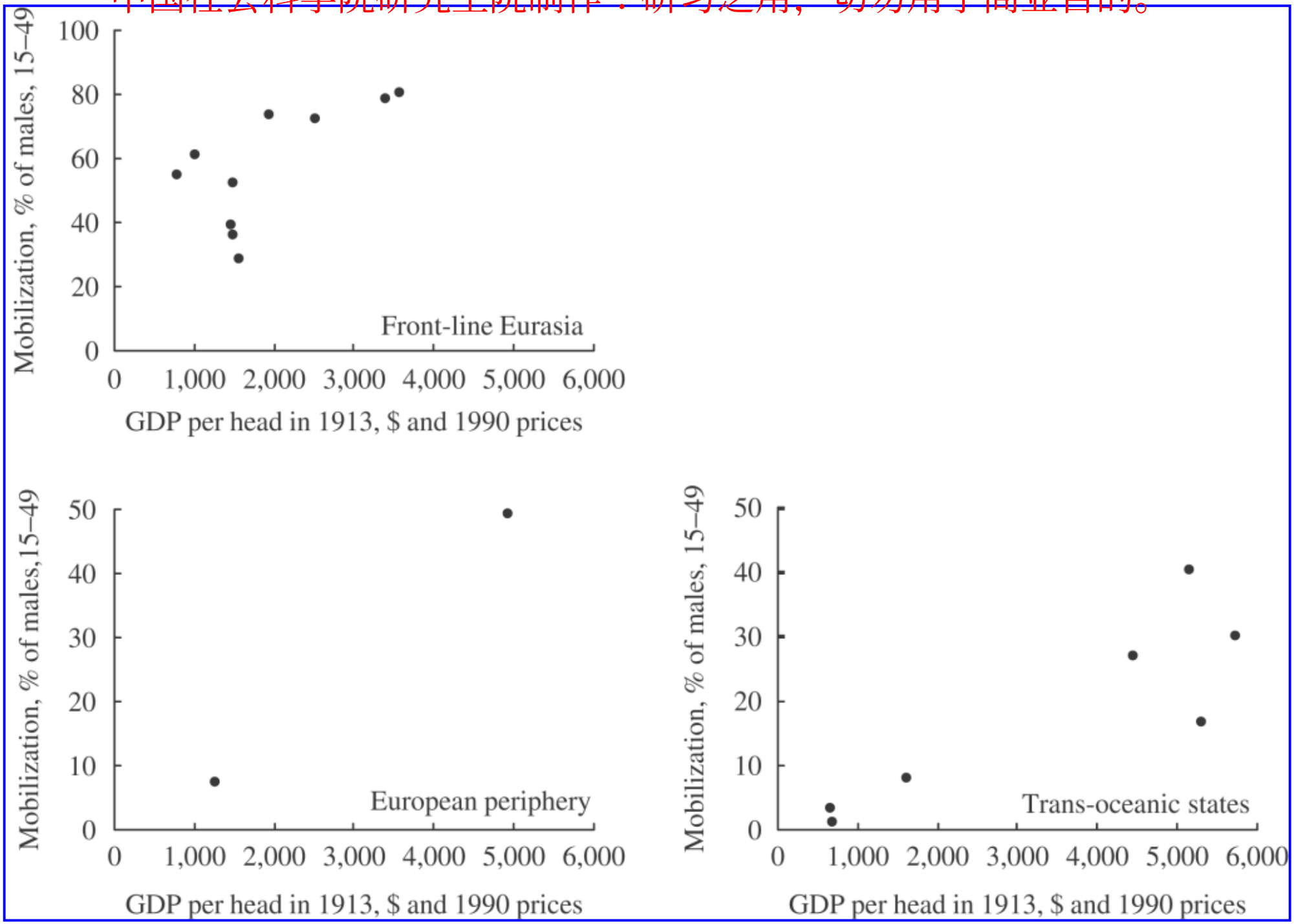

Figures 1 to 3 show us that the level of development acted as a multiplier of size. Rich countries were able to mobilize production, public finance and soldiers out of proportion to the size of their economies measured by GDP.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Mobilization and agriculture

Why did being poor matter for large countries like Russia, Austria-Hungary and Turkey? During the First World War, the answer can be found in the performance of the agricultural sector, since these countries all ran short of food long before they ran out of guns and shells (Offer, 1989). Broadberry and Harrison (2005) attribute this to the negative impact of peasant agriculture on mobilization.

One of the most striking attributes of relative poverty was the role of subsistence farming. Contemporary observers were aware of these differences and interpreted them as follows: when war broke out, a country such as Russia would have an immediate advantage in that most of the people could feed themselves; moreover, the diversion of food supplies from export to the home market would actually increase Russia's advantage. In contrast, Britain would quickly starve (Gatrell and Harrison, 1993). This diagnosis could not have been more wrong. In practice a large peasantry proved to be a great disadvantage in mobilizing resources for war. Meyendorff (cited by Gatrell, 2005) described what happened in Russia as ‘the Russian peasant's secession from the economic fabric of the nation’. And not only from Russia, for Italy, Austria-Hungary, the Ottoman Empire and Germany all had large peasant populations that proved extremely difficult to mobilize for much the same reason. In wartime peasant agriculture behaved like a neutral trading partner. Why should the Netherlands trade with Germany at war, given the latter's reduced ability to pay, except under threat of invasion and confiscation? Peasant farmers, trading with their own governments, made the same calculation. Thus the Russian economy looked large, but if the observers of the time had first subtracted its peasant population and farming resources they would have seen how small and weak Russia really was.

The peasant's propensity to secede is clearly visible from a comparison of the richer and poorer countries’ experience. When war broke out British and American farmers were offered higher prices, responded normally to incentives, and boosted production. The fact that British farming had already contracted to a small part of the economy made its expansion easier: there were plentiful reserves of land unused or little exploited, and the high productivity of farm labour meant that large increases in farm output could be achieved with few additional resources (Olson, 1963; Broadberry and Howlett, 2005).

In the poorer countries, in contrast, wartime mobilization took resources away from farming, particularly young men and horses for the army. Once in the army these young men and horses still needed to be fed, which required a diversion of food supplies from rural households to government purchasers. But the motivation for farmers in the countryside to sell food was reduced, not increased. These were subsistence farmers who grew food partly for their own consumption; what they sold, they took to the market primarily to buy manufactured commodities for their families. But war dried up the supply of manufactures to the countryside. The small industrial sectors of the poorer countries were soon wholly concentrated on supplying the army with weapons and kit. Little capacity was left to supply the countryside, which faced a steep decline in supplies.

Consequently, peasant farmers retreated into subsistence activities and the economy began to disintegrate. There might still be plenty of food, but it was locked in the countryside. The farmers preferred to eat it themselves than sell it for a low return. What food it could get, the government gave to the army for a simple reason: hungry soldiers will not fight. Between the army and the peasantry the urban workers were caught in a double squeeze. As the market supply of food dried up, urban food prices soared, and an urban famine set in. In terms of the economics of famines, the primary cause was not a failure of production but the urban society's loss of food entitlements (Sen, 1983; Offer, 1989).

Aware of this, public opinion might blame unpatriotic speculators or incompetent officials. But the truth was that a poor country had few genuine choices. The scope for policy to improve the situation was more apparent than real, and government action often made things worse: the Russian, Austrian and German governments all began to ration food to the urban population, for example, while attempting to buy food from the farmers at purchasing prices that were fixed low for budgetary reasons. To repeat: in richer countries the government paid *more* to the farmers, and this worked, but in poorer countries the government tried to pay *less* and this had entirely predictable results: the farmers’ willingness to participate in the market was further undermined.

In summary, in wartime poor countries suffered the consequences of peasant agriculture, which was essentially a deadweight on their mobilization efforts. Economic mobilization led to urban famine, revolutionary insurrection, and the downfall of emperors in Russia, Austria-Hungary, Germany and Turkey. The same process began in France, which still had a large peasant sector in 1914, but Allied support nipped it in the bud.

War losses

After the First World War, there were several attempts to calculate the costs of the war. However, these studies fell out of fashion, tainted by association with inflated demands for reparations, and because later writers became interested in any positive developments that could be identified as arising out of the carnage and destruction. Thus in his popular survey of the First World War, Hardach (1987, p. 286) argues that Bogart's (1920) estimates of the costs of the war have not been revised in the light of later evidence because ‘the whole basis of calculation has been recognized as inappropriate’.

There are good grounds to be sceptical, however, about the revisionist view that associates war with accelerated economic development. Milward (1984, pp. 17–18), a leading revisionist, cites Bowley (1930) as a pioneer of revisionism, but Bowley himself (1930, pp. 21–3) pointed out how difficult it is to show that any wider changes were actually the result of the war and would not have occurred anyway in its absence. Classifying developments as (*a*) mainly unconnected with the war, (*b*) accelerated or retarded by it or (*c*) apparently arising out of it, Bowley was himself reluctant to put anything other than the key elements of Bogart's ‘cost of war’ calculations such as loss of life and destruction of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

capital into category (*c*). He did mention the new economic relations between Europe and the United States in this category, but with hindsight we can see that the process of US overtaking was already under way well before 1914 (Abramovitz, 1986; Broadberry, 1998).

Table 4, accordingly, provides updated estimates of the destruction of human and physical capital as the costs of war within a national balance sheet framework provided by Broadberry and Howlett (1998). The first element, the destruction of human capital, is measured by war deaths relative to the population aged 15–49. This differs from the true proportion of human capital destroyed by the war to the extent that younger cohorts had more human capital investment, particularly through education. Although Germany suffered the most casualties in absolute numbers, the proportionate losses were higher in France, Serbia-Montenegro and Roumania among the Allies, and in Turkey and Bulgaria among the Central Powers.

Destruction of human and physical capital (per cent of pre-war assets)

	Human capital	Physical capital			
		Domestic assets	Overseas assets	Reparations bill	National wealth
Allies					
Britain	3.6	9.9	23.9	…	14.9
France	7.2	24.6	49.0	…	31.0
Russia	2.3	14.3	…	…	…
Italy	3.8	15.9	…	…	…
United States	0.3	…	…	…	…
Central Powers					
Germany	6.3	3.1	…	51.6	54.7
Austria-Hungary	4.5	6.5	…	…	…
Turkey and Bulgaria	6.8	…	…	…	…

Source: Derived from Broadberry and Harrison (2005, p. 28).

The domestic physical capital losses in Table 4 build upon the work of Bogart (1920), who estimated property losses on land and shipping and cargo losses. These are expressed as a proportion of physical capital from modern historical national accounting sources. The French figures draw on estimates of losses from the Reparations Commission and capital stock data from Carré, Dubois and Malinvaud (1976, p. 151). Although these probably overstate French losses, alternative estimates by Villa (1993) yield implausibly low ratios, given the concentration of fighting on French soil (Hautcoeur, 2005, p. 199). Russia's losses were proportionately high, more because of the small size of the pre-war capital stock than a large absolute amount of wartime destruction.

For some countries in Table 4, we can estimate the change in overseas assets and national wealth. Nearly a quarter of British overseas investments was liquidated during the war, so that the reduction of national wealth was proportionally much greater than the loss of physical capital. The loss of French overseas assets was proportionally very high due to heavy exposure to Russian loans, so that, as in Britain, the share of national wealth lost in the war was proportionally greater than the share of physical capital lost.

Finally in Table 4, we have added in Germany's reparations bill as a proportion of pre-war capital, since it represented an increase in foreign liabilities and hence a reduction in national wealth just as much as the liquidation of Britain's overseas assets meant a reduction in national wealth. Of course the extent to which Germany actually had to pay these reparations is much debated, but that does not alter the effect on the national balance sheet as it stood in 1919, immediately after the Treaty of Versailles. These figures include the A+B+C Bonds, which added up to a total of 132 billion Gold Marks.

The Second World War

Like its forebear, the Second World War may be divided into two periods. In the first period, economic considerations were less important than purely military factors. This was the phase of greater success for the powers of the Axis, and it lasted from 1937 when the war began in the Pacific, or from 1939 in Europe, until the end of 1941 or 1942; the exact turning point differed by a few months among the different regional theatres. In this first period, Germany and Japan had advantages of strategy and fighting power on their side. As a result, they were able to inflict overwhelming defeats upon an economically superior combination of powers. In early 1942, Richard Overy writes (1995, p. 15), ‘no rational man … would have guessed at the eventual outcome of the war’.

This phase ended, however, without the decisive victory that previously appeared within the Axis powers’ grasp. What ended it? On the surface it was the military failures, not

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

economic weaknesses, of the Axis. Beneath the surface, however, economic fundamentals reasserted themselves: while the Allies had given ground everywhere, their morale had stiffened, their economies were not exhausted, their cooperation was taking effect, and their industries were supplying the front with a rising flood of munitions that would eventually overwhelm the adversary.

In the second period of the war, which began in 1942, the early advantages of the Axis evaporated. There was a brief stalemate. A war of attrition developed in which the opposing forces ground each other down, with rising force levels and losses. Superior military qualities came to count for less than superior size, wealth and economic mobilization. Economic superiority let the Allies take risks, absorb the cost of mistakes, replace losses, and accumulate overwhelming force. This turned the balance against the Axis and won the war.

This narrative does not support the claim that only economics mattered. Economic factors were decisive, however, in the context of a simple fact. The Axis leaders had the chance to use the other factors to decide the war, and they failed. Their failure gave the Allies the chance to bring economics decisively into the equation.

Size and development

Table 5 shows the volumes of combat resources that each side delivered to the theatres of the Second World War. A comparison of the totals with those in Table 1 shows a staggering increase: a quarter of a million tanks and half a million aircraft, for example, compared with 170,000 aircraft and fewer than 10,000 tanks in the First World War. One thing remained the same, however, across the two wars: the Allies supplied a greater volume of combat resources than their combined adversaries in almost every respect.

Allies vs Axis: soldiers and equipment in the Second World War

	Allies	Axis	Ratio, 1:2
	(1)	(2)	(3)
Combatant-years, million	106.4	76.9	1.4
Weapons Produced:			
Rifles and carbines, million	25.3	13.0	1.9
Combat aircraft, thousand	370	144	2.6
Machine guns, thousand	4,827	1,646	2.9
Guns, thousand	1357	462	2.9
Armoured vehicles, thousand	216	51	4.3
Mortars, thousand	516	100	5.1
Major naval vessels	8,999	1,734	5.2
Machine pistols, thousand	11,604	1,185	9.8
Ballistic missiles	0	6,000	…
Atomic weapons	4	0	…

Source: Harrison (1998, pp. 14–16) except that numbers in the French armed forces in 1940 are corrected as noted by Harrison (2005). The number of ballistic missiles is an approximate upper limit based on Ordway and Sharpe (1979, pp. 405–7). Of the four bombs produced by the Manhattan Project one was tested at Alamogordo, two were exploded over Japanese cities, and one remained unused.

The Allied advantage did not hold at all points of time and place. As Goldsmith remarked, the pre-war rearmament of the Axis powers gave them an early start and this, combined with their purely military advantages, accounts for their early success. A balance struck at the end of 1940, for example, when France had dropped out, the United States remained neutral, and the Soviet Union was still Germany's partner of convenience, would show a picture of Allied disadvantage. By 1942, however, reinforced by America and Russia, the Allies outnumbered and outgunned the powers that they faced in every major theatre. This was true even on the eastern front where Germany and the USSR confronted each other. These two powers were of similar economic size measured by GDP and industrial production, but the Soviet Union was substantially poorer in terms of the average incomes of its much larger population. Although this disadvantage was enlarged by devastating military and territorial losses in 1941 and 1942, the Soviet Union fielded a bigger army and supplied it more generously. We return to this anomaly below.

The relative economic sizes of the powers and their colonial possessions are shown in Tables 6 and 7. If we consider the world as it was on the eve of the war, then the populations available to the Allies – principally Britain and France with their colonies and dominions, but also including Poland and Czechoslovakia – amounted to nearly 690 million people occupying nearly 48 million square kilometres. The total output of this territory is estimated at one trillion dollars in 1990 prices. Against them stood the nearly 260 million people

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

and more than six million square kilometres available to the Axis powers – Germany, Italy and Japan, with the Italian and Japanese empires and annexations at the time such as Austria, Korea and Manchuria. The people and their lands on the side of the Allies exceeded those available to the Axis power by several times.

The alliances in the Second World War: resources of 1938

	Population, million	Territory		GDP in 1990 prices	
		million sq. km	ha. per head	$ billion	per head, $
Allies					
1938					
Allies, total	689.7	47.6	6.9	1,024	1,485
UK and France only	89.5	0.8	0.9	470	5,252
1942					
Allies, total	783.5	68.0	8.7	1,749	2,232
UK, USA and USSR only	345.0	29.3	8.5	1,444	4,184
Axis					
1938					
Axis, total	258.9	6.3	2.4	751	2,902
Germany, Austria, Italy and Japan only	190.6	1.2	0.7	686	3,598
1942					
Axis, total	634.6	11.2	1.8	1,552	2,446
Germany, Austria, Italy and Japan only	190.6	1.2	0.7	686	3,598

Source: Harrison (1998, pp. 3–9).

Allies versus Axis: size and development ratios

	Population	Territory	Territory per head	Gross domestic product	GDP per head
1938					
Total	2.7	7.5	2.8	1.4	0.5
Great Powers only	0.5	0.6	1.4	0.7	1.5
1942					
Total	1.2	6.1	4.9	1.1	0.9
Great Powers only	1.8	23.5	13.0	2.1	1.2

Source: Calculated from Table 6.

The size of this advantage is more statistical than real, although a real advantage remains after the statistics are stripped out. Africa and South Asia, poor, undeveloped and relatively sparsely settled, made up the greater part of the Allied advantage in size. When we turn to total output, it turns out that the Allied GDP exceeded that of the Axis territories by only one-third; this is because average incomes across the Allied territories – less than $1,500 in modern prices – stood at only one half the $2,900 level of the Axis territories. Here is an ironic comment on the colonial aspirations of the Axis powers: what they wanted so much, and did not yet have, was access to millions of square kilometres of poorly integrated, low-yielding farmland and remote semi-desert.

As before, since poor colonies did not count for very much, we also count the resources on either side considering the great powers only. The Allied size advantage now disappears since Germany, Italy and Japan together had twice the population and one and a half times the territory of Britain and France – but it is replaced by a development advantage: the GDP per head of the Allied powers exceeded that of the Axis powers by one half.

Tables 6 and 7 also show how this balance evolved from 1938 to 1942, when the domains under control of the Axis powers had reached their maximal extent. As their forces swept across Europe and the Pacific region the population under Axis control tripled, while territory and peacetime output potential doubled; the addition of hundreds of millions of east European and east Asian farmers led the average development of the Axis empires to decline somewhat, from $2,900 to less than $2,500 in modern prices. Britain, in contrast, lost its

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

allies France – with its empire – and Poland, but was joined by the Soviet Union and the United States. Between 1938 and 1942, therefore, the Allied population and territorial sway increased somewhat, but its peacetime output rose by three quarters (from one trillion to 1.75 trillion dollars). This is *before* the wartime doubling of United States output is taken into account. Joining the richest and poorest of the great powers into a single coalition had a mixed effect, of course, but the net result was an increase in the measured average level of development across all the Allied territories from less than $1,500 in modern prices to more than $2,200.

Table 7 converts these figures into ratios of Allied advantage or disadvantage. We see that in 1942, when things were at their worst, the Allied powers alone had nearly twice the population, more than twice the output, and more than 20 times the territory of the Axis powers. All they had to do was not to lose; given enough time, this economic preponderance would surely bring victory. The weakest link in the Allied chain was poor Russia, with its hundred million low-productivity peasants and seven million square kilometres of permafrost. Germany had forced Russia out of the First World War; could the same not happen again?

Development and mobilization

As with the First World War we will consider three dimensions of mobilization: production (the increase in total output that was achieved during the war), the government's fiscal leverage (the extent that output was mobilized through government spending and taxation out of peacetime uses into the war effort), and military mobilization (the degree of mobilization of the population into uniform). Each of these was powerfully influenced by the pre-war level of development of the economy.

Figure 4 shows production mobilization plotted against pre-war average income. Under the pressure of war, rich countries expanded their economies; poor countries tended to collapse, and the collapse proceeded further, the poorer they were. Figure 5 shows the speed of fiscal mobilization. The slope of the relationship with pre-war economic development has the same positive sign: only rich countries achieved significant fiscal mobilization, but there is an exception: the Soviet Union. Some underlying figures are provided in Table 8: these confirm that the Soviet Union achieved a level of mobilization of GDP into the war effort – three-fifths at its peak – that was typical of much richer countries. Germany and Japan achieved similar degrees of mobilization only in the last spasm of the struggle that preceded immediate collapse and defeat.

The military burden 1939-44 (military outlays, per cent of national income)

	1939	1940	1941	1942	1943	1944
At current prices						
Allied powers						
USA	1	2	11	31	42	42
UK	15	44	53	52	55	53
USSR	…	…	…	…	…	…
Axis powers						
Germany	23	40	52	64	70	…
Italy	8	12	23	22	21	…
Japan	22	22	27	33	43	76
At constant prices						
Allied powers						
USA	1	2	11	32	43	45
UK	…	…	…	…	…	…
USSR	…	17	28	61	61	53
Axis powers						
Germany	23	40	52	63	70	…
Italy	…	…	…	…	…	…
Japan	…	…	…	…	…	…

Source: Harrison (1998, p. 21).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Figure 4
Production mobilization: 11 countries, 1938–1942. *Note*: Observations are, from left to right, the Soviet Union, Japan, Italy, Finland, Austria, Canada, Germany (excluding Austria), Australia, UK, USA and New Zealand. *Sources*: Harrison (1998, p. 10) and Maddison (2001); for Soviet GDP see also the sources listed under Table 5.

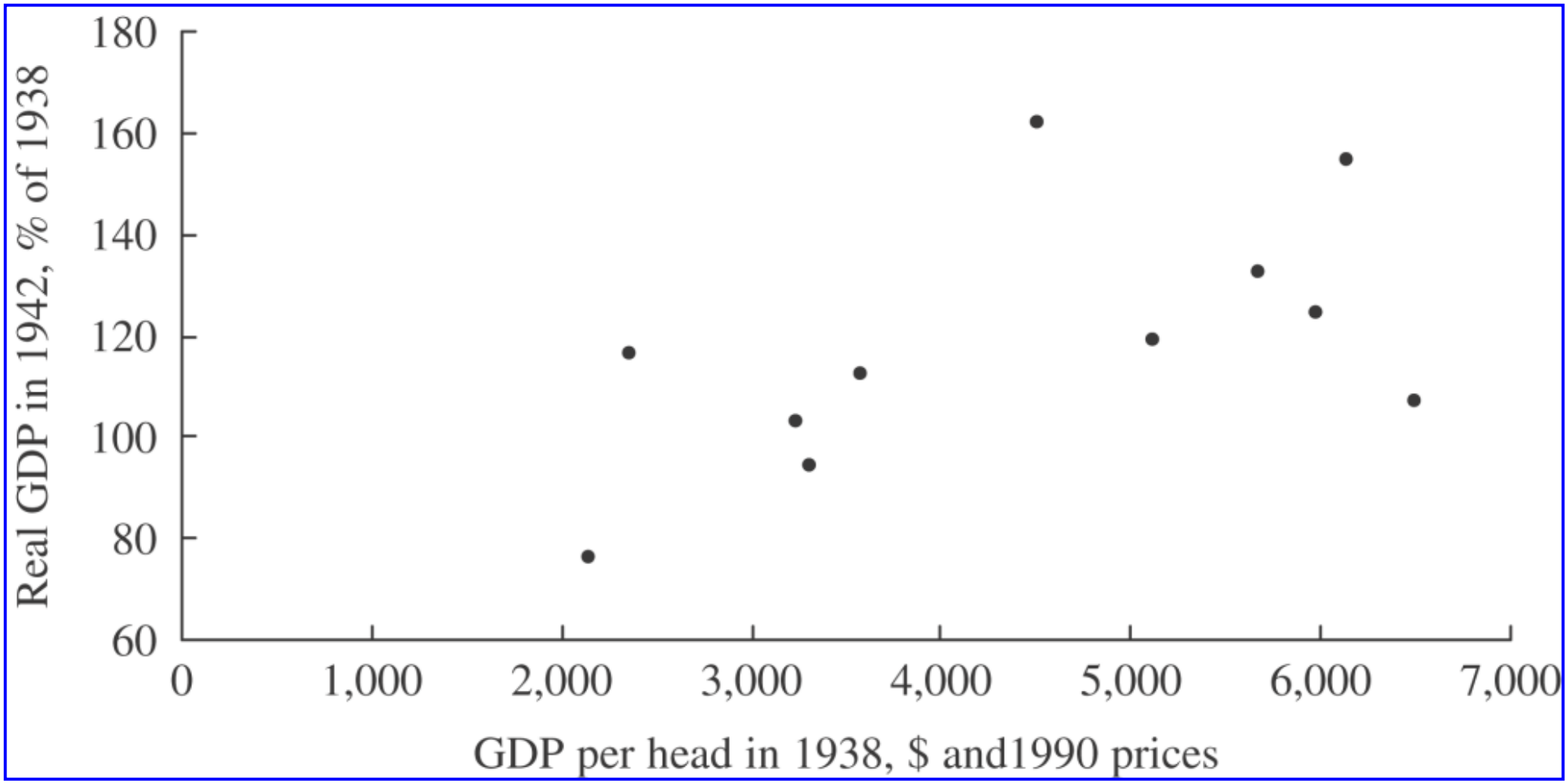

Figure 5
Fiscal mobilization in the Second World War: six countries. *Notes*: Observations are, from left to right, the Soviet Union, Japan, Italy, Germany, the UK and the USA. *Source*: Harrison (1998, p. 21).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

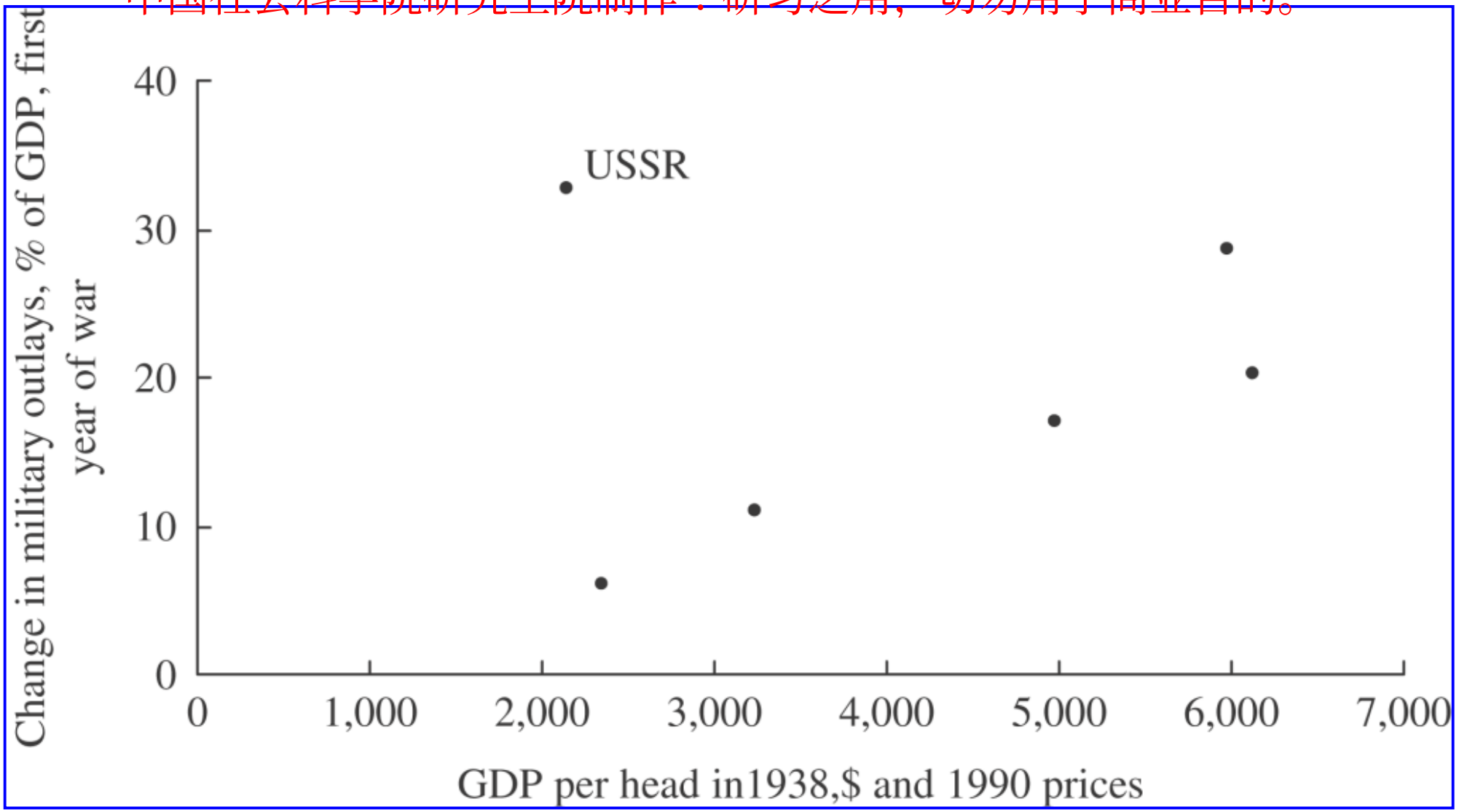

The Soviet anomaly demands explanation. A relatively poor country, Russia collapsed in the First World War, and the Soviet Union could have been expected to do so again in the Second World War, but did not. The course of inward-looking industrialization that Stalin pursued between the wars does not appear to be a sufficient explanation. More important was Stalin's victory in the destructive struggle to collectivize farming, which ensured state control over wartime food supplies and prevented the peasants from seceding from the war effort (Gatrell and Harrison, 1993). As a result the Soviet economy carried a disproportionately heavy economic burden in the Second World War without collapsing.

Finally, Figure 6 shows military mobilization; again, the rich countries mobilised much higher proportions of their population into military uniform. The figure also shows the moderating effect of distance: when we control for pre-war incomes, the countries separated from the fighting by oceanic distances put fewer men into the fighting forces. But the effect of distance was less in the Second World War than in the First World War, suggesting that the interwar decline of transport costs had brought about a more truly global struggle.

Figure 6

Military mobilization in the Second World War: 17 countries. *Note*: Observations are, from left to right: Front line Eurasia: China, Roumania, Bulgaria, USSR, Japan, Hungary, Greece, Italy, Finland, France, Germany and UK. Trans-Oceanic states: South Africa, Canada, Australia, USA and New Zealand. *Sources*: Harrison (1998, pp. 3–9, 14); Correlates of War data-set, version 2.1, at http://0-www.umich.edu.library.lemoyne.edu/~cowproj; Singer (1979; 1980).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

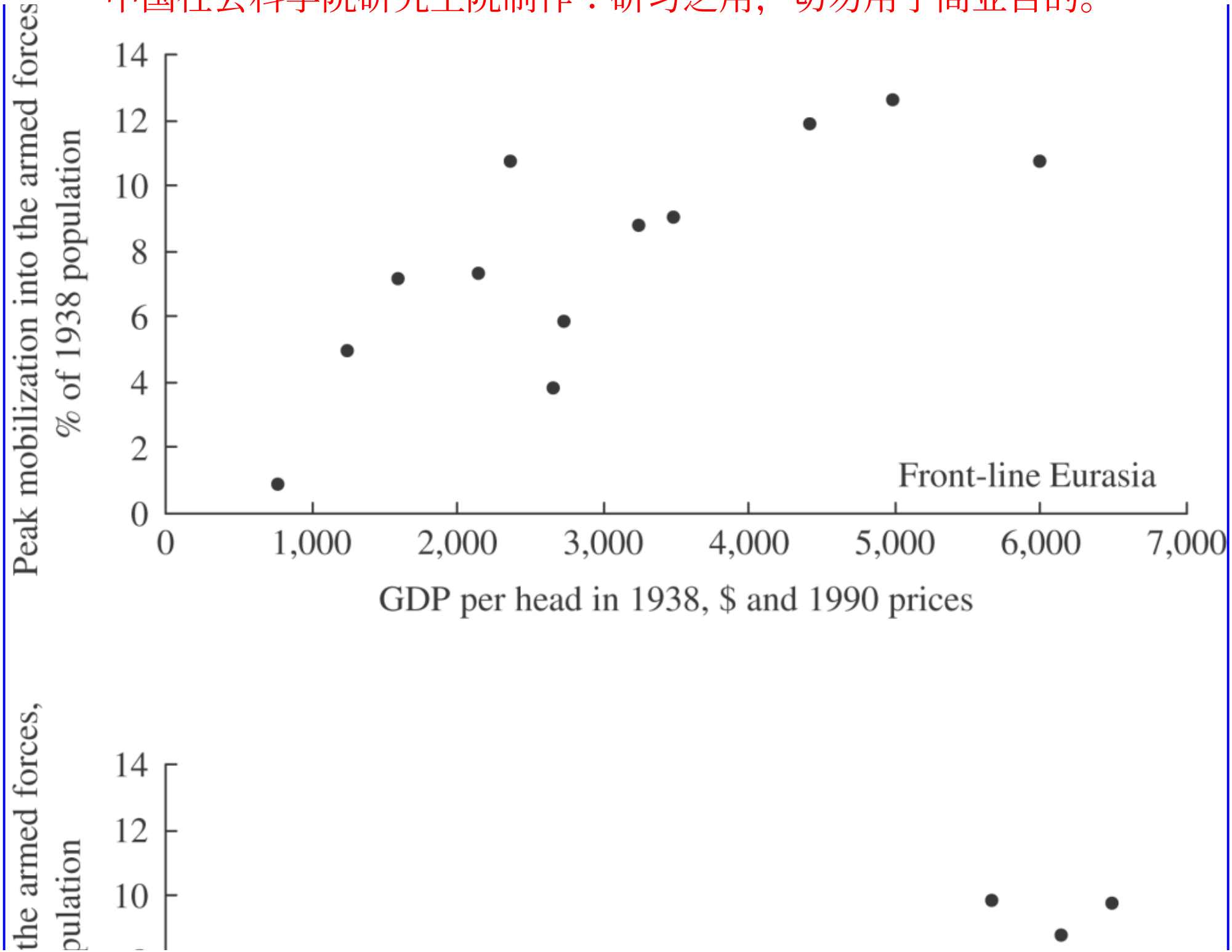

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

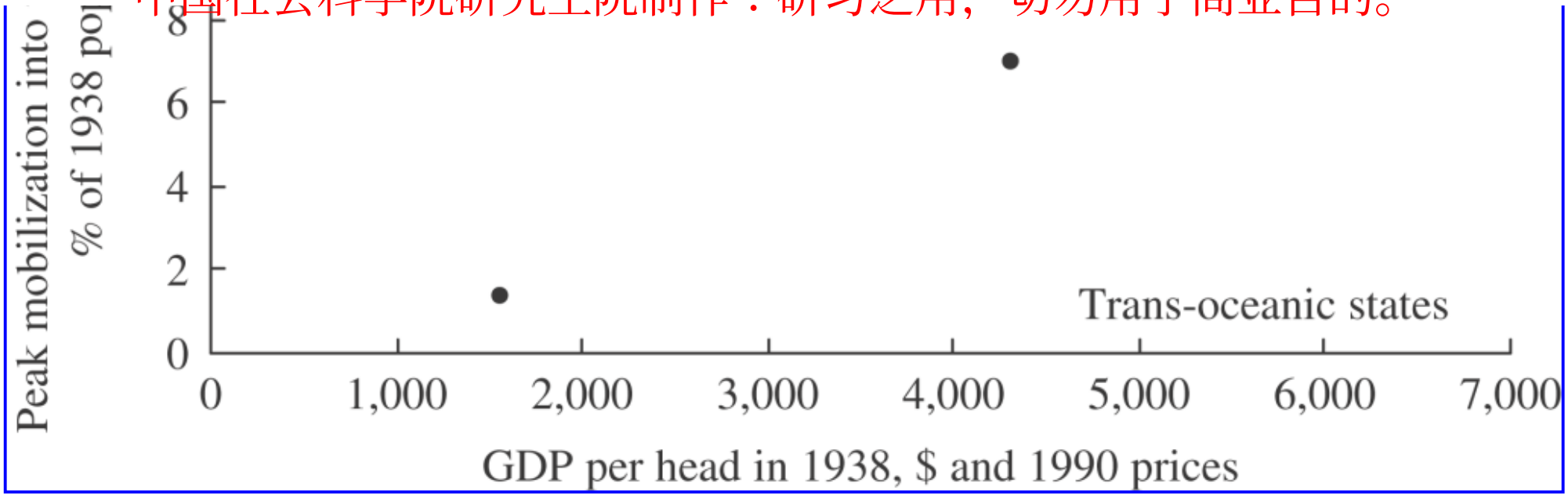

Inter-allied cooperation

In terms of cooperation within the opposing coalitions, the Second World War saw a repeat of the First World War with some differences. In both wars the German-led coalition failed to achieve significant economic cooperation among the powers, each of which aimed primarily to exploit its own internal and colonial spaces. The Allies, in contrast, achieved fuller cooperation. During the first war, this involved pooling the industrial and commercial resources of Britain and America with the fighting strength of France, Italy, and Russia; the result was to permit the aggregate military power of the Allies to be produced more efficiently. The main instruments of pooling were war credits from America to Britain, France and Italy, and from France and Britain to Italy and Russia. The amount was not enough to keep Russia in the war to the end, but enough that post-war repayments significantly complicated post-war international finance and trade.

The second time round, inter-allied cooperation assumed a larger scale. The main form it took was the transfer of industrial goods – equipment (including vehicles), materials, fuels and processed foodstuffs – from the United States to Britain and from both countries to the Soviet Union. Although the US legislative framework called it ‘Lend-Lease’, the goods were actually supplied free of financial charges, the aim being to promote the Allied partnership. Pooling of the resources counted in Tables 6 and 7 augmented their value, increased the Allied advantage, saved lives and resources, helped to prevent Soviet defeat, and brought forward the Allied victory.

Allied cooperation was not problem-free. The main issue was that, while it saved lives and brought forward victory, it did so asymmetrically. By keeping the Russians in the war, it saved primarily American and British lives, and the Russians felt this deeply. On the other hand, the victory that it brought forward was brought to Berlin by the Red Army, and was much more favourable to post-war Soviet power than would have been the case without western assistance – a source of wartime chagrin and post-war recriminations among the donors.

War losses

The Second World War was fought on a global scale but half a dozen countries saw most losses of wealth and population. Nearly all the 55 million premature deaths, for example, are accounted for between the USSR (25 million), China (10 million), Germany (6.5 million), Poland (5 million), Japan (2.4 million) and Yugoslavia (1.7 million). Table 9 summarizes the data for the great powers as percentages of prewar populations and assets.

War losses attributable to physical destruction (per cent of assets)

	Human assets	Physical assets	
		National wealth	Industry fixed assets

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Allied powers			
USA	1	0	…
UK	1	5	…
USSR	18–19	25	…
Axis powers			
Germany	9	…	17
Italy	1	…	10
Japan	6	25	34

Source: Harrison (1998, p. 37).

The figures show the heavy loss of life in the Soviet Union, followed by German and Japan, and also the widespread destruction of property in the same countries. Everywhere, it seems, human capital was destroyed at a higher rate than physical capital. The survivors were endowed, therefore, with a ratio of physical to human capital that was advantageous by pre-war standards, provided that mismatches resulting from the wartime distribution of combat could be smoothed out. Table 9 takes no account of accelerated wartime investments in industry; in western Germany, for example, industrial capacity was added at a faster rate than bombing took it away, so that West German industry ended the war with a larger and newer stock of equipment than before (Abelshauser, 1998, p. 168).

Economic growth

Evidently, wartime economic mobilization tended to make the rich richer and the poor poorer. Thus, both wars tended to polarize the global distribution of income. It is of some interest, therefore, to examine whether postwar recovery and long-term economic growth succeeded in reversing this pattern.

Figure 7 suggests that each war was followed by recovery and that those economies most damaged by the wartime experience recovered most rapidly. It takes 1929 as the benchmark date for measuring recovery from the First World War, and 1973 as the benchmark for recovery from the Second World War. It shows that, the more a country's average income fell during each war, the more it tended subsequently to rise. Thus, at least some of each war's negative effects were transitory.

Figure 7

Economic recovery following two world wars. *Note*: Observations are Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Italy, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, UK and USA. *Source*: Maddison (2001).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

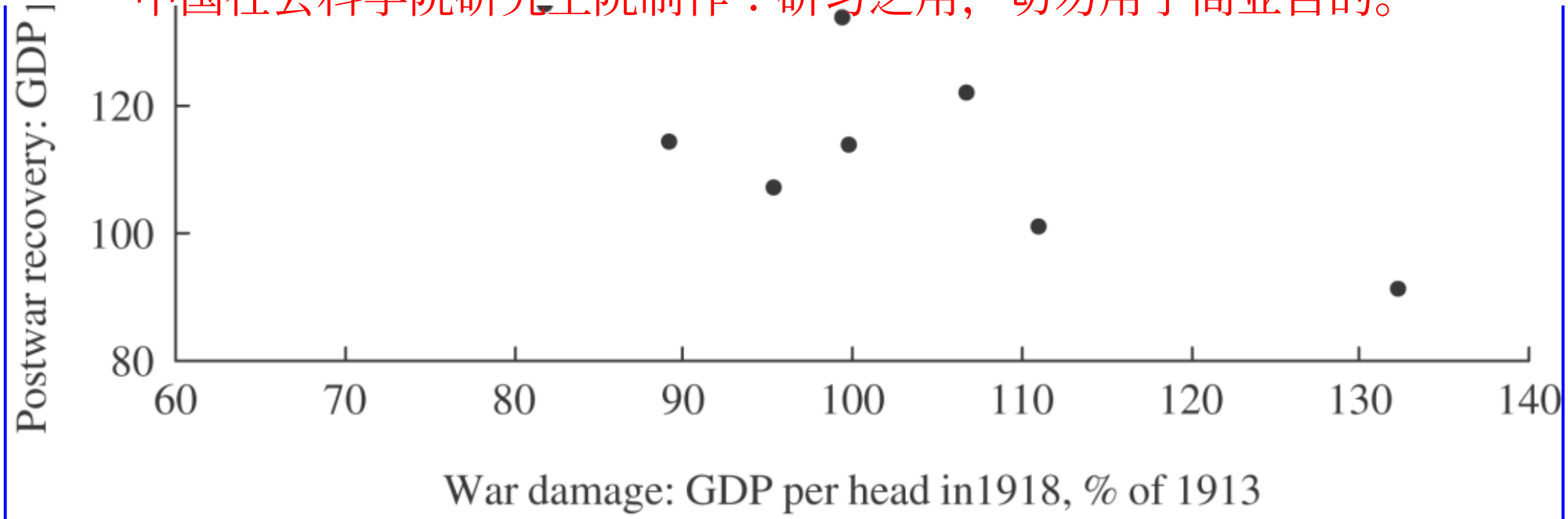

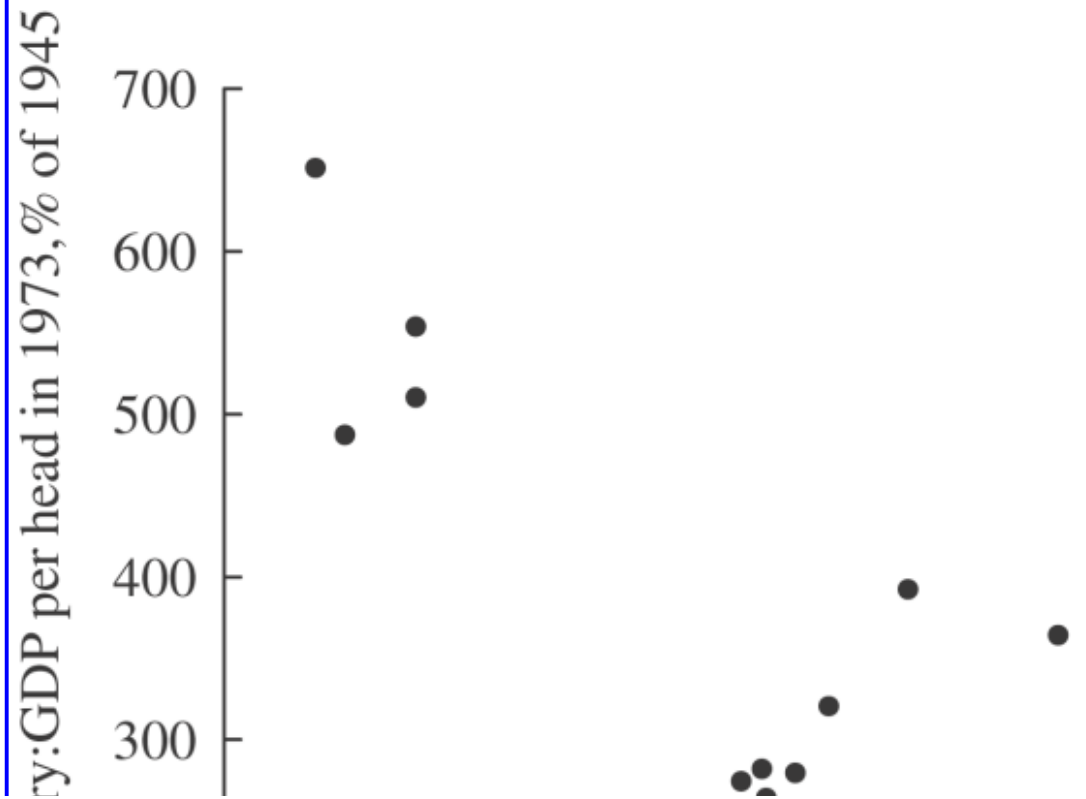

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

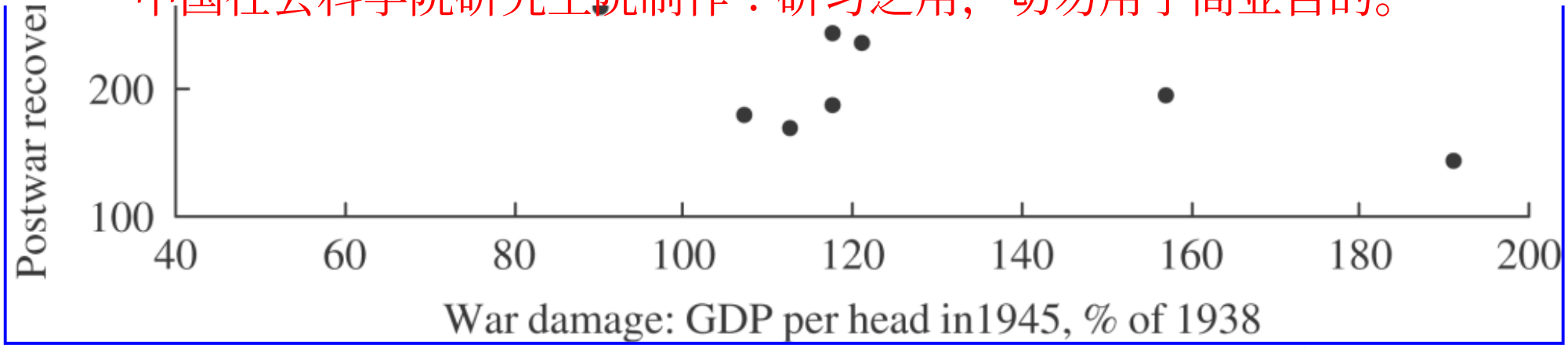

A more complex picture emerges when we turn to long-term economic growth. To what extent did the post-war recovery return each country to a path of convergence on the global productivity frontier? Figure 8 suggests that after the First World War there was little or no convergence. Some countries that were already rich did much better after the war than some countries that were already poor. In contrast, the Second World War was followed by convergent economic growth. This suggests that the Allies designed a much better international environment for genuine convergence after 1945 than after 1918.

Figure 8

Economic convergence through two world wars. *Notes* and *Sources*: as for Figure 7.

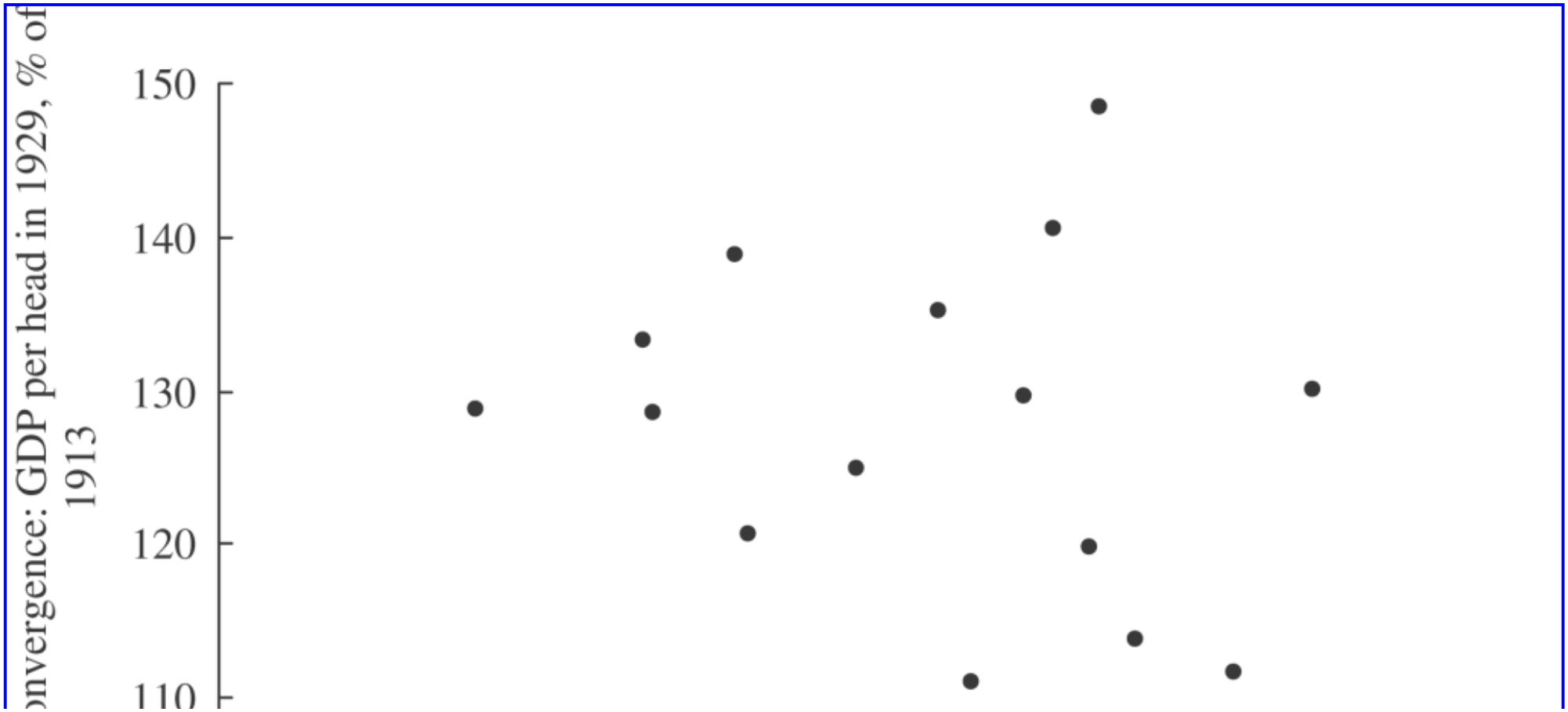

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

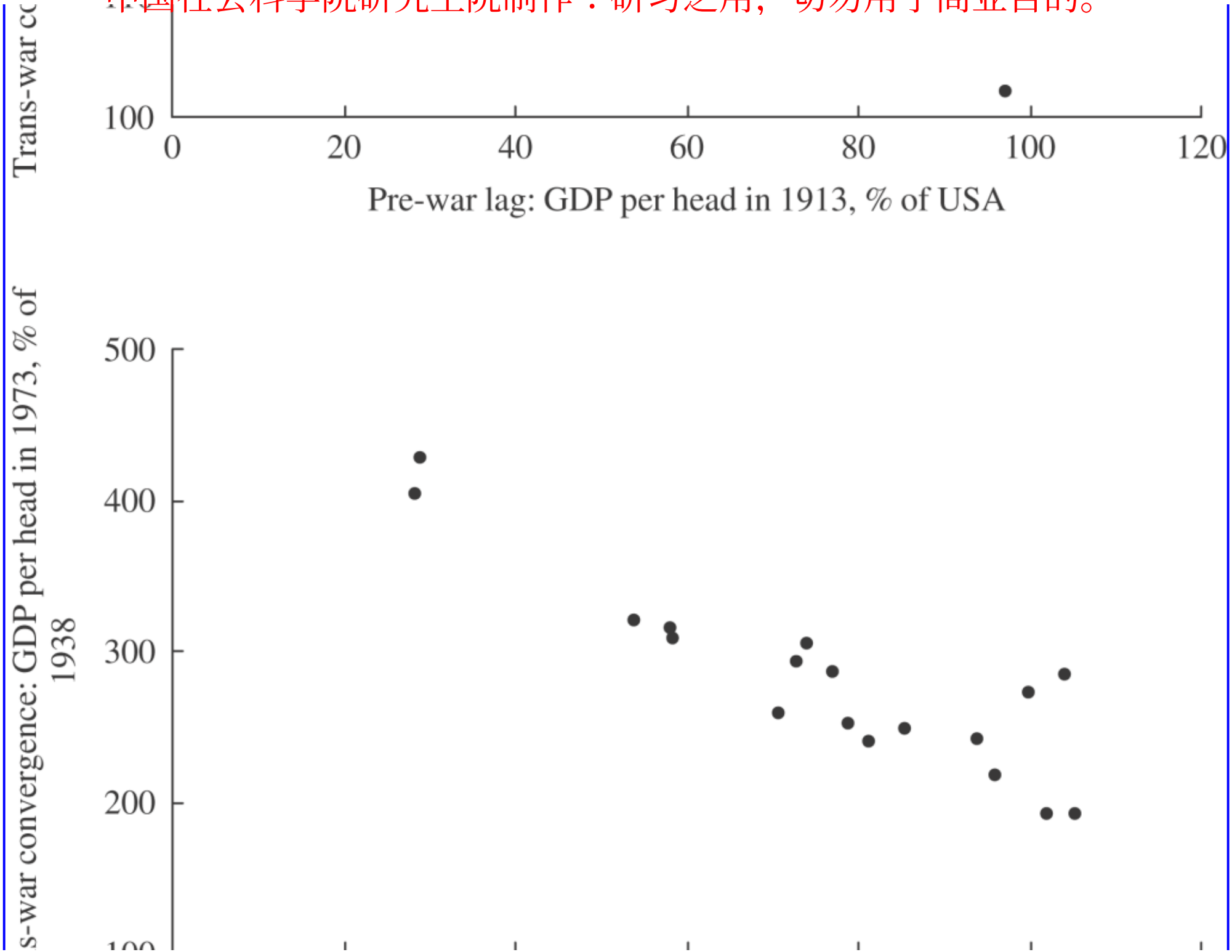

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

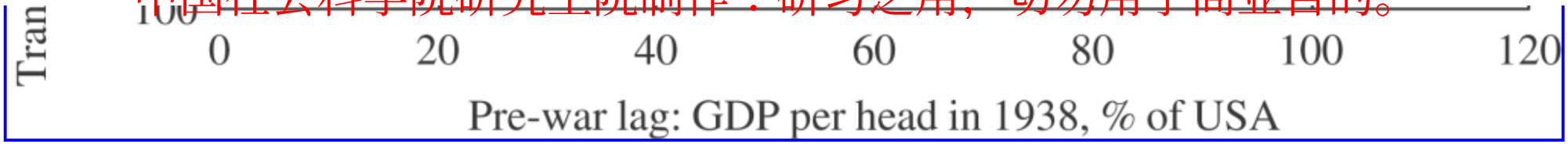

Conclusions

In this article we have shown that economics mattered, and we have shown how. Given time, resources won the two world wars. In mobilizing resources, the richer market economies had a significant advantage. It was more important to be rich than self-sufficient; probably, most pre-war efforts to protect jobs or diminish national dependence on trade in the name of strategic self-sufficiency were counterproductive. Poor economies, especially those with a large peasant population, tended to collapse under the stress of total war, although they tended to be less reliant on external trade. The main exception is the Soviet economy in the Second World War; its exceptional resilience is best explained by its rulers' exceptional degree of control over the peasant farmers.

The pattern should not be overgeneralized. Broadberry and Harrison (2005) have suggested that the power of the relationship between economic and military performance is confined to a relatively short historical period. The era of 'total war' from 1914 to 1945 seems to have been unique. In both world wars the main combatants were able to devote more than half of their national income to the war effort. This is likely to have been impossible before 1914 because until then most people were too poor to be taxed at such rates; most economies had the bulk of their resources locked up in forms of subsistence agriculture that were resistant to mobilization; before mass literacy and the telegraph, typewriter and duplicator, commercial and government services were too inefficient to do much about it. In short, in earlier stages of global development total war could not be staged because too many people were required to labour in the fields and workshops just to feed and clothe the population, and it cost too much for government officials to count, tax and direct them into mass combat.

Since 1945 the economic factors in warfare may have lost significance again. This is because nuclear weapons can give devastating military force to any rich country however small, or any large country however poor, for a few billion dollars. Hence the marshalling of economic resources played a much more vital role in the outcome of the two world wars than was likely in any period before or after.

See Also

- defence economics
- war and economics

Bibliography

Abelshauser, W. 1998. Germany: guns, butter, and economics miracles. In *The Economics of World War II: Six Great Powers in International Comparison*, ed. M. Harrison. Cambridge: Cambridge University Press.

Abramovitz, M. 1986. Catching up, forging ahead and falling behind. *Journal of Economic History* 46, 385–406.

Bogart, E.L. 1920. *Direct and Indirect Costs of the Great World War*, 2nd edn. New York: Oxford University Press.

Bowley, A.L. 1930. *Some Economic Consequences of the Great War*. London: Thornton Butterworth.

Broadberry, S.N. 1998. How did the United States and Germany overtake Britain? A sectoral analysis of comparative productivity levels, 1870–1990. *Journal of Economic History* 58, 375–407.

Broadberry, S. and Harrison, M., eds. 2005. *The Economics of World War I*. Cambridge: Cambridge University Press.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Broadberry, S.N. and Howlett, P. 1998. The United Kingdom: victory at all costs. In *The Economics of World War II: Six Great Powers in International Comparison*, ed. M. Harrison. Cambridge: Cambridge University Press.

Broadberry, S.N. and Howlett, P. 2005. The United Kingdom during World War I: business as usual? In *The Economics of World War I*, ed. S. N. Broadberry and M. Harrison. Cambridge: Cambridge University Press.

Carré, J.-J., Dubois, P. and Malinvaud, E. 1976. *French Economic Growth*. Oxford: Oxford University Press.

Ferguson, N. 1998. *The Pity of War*. London: Allen Lane.

Gatrell, P. 2005. Poor Russia, poor show: mobilising a backward economy for war, 1914–1917. In *The Economics of World War I*, ed. S.N. Broadberry and M. Harrison. Cambridge: Cambridge University Press.

Gatrell, P. and Harrison, M. 1993. The Russian and Soviet economy in two world wars. *Economic History Review* 46, 425–52.

Goldsmith, R.W. 1946. The power of victory: munitions output in World War II. *Military Affairs* 10(1), 69–80.

Hardach, G. 1987. *The First World War, 1914–1918*. Harmondsworth: Penguin.

Harrison, M., ed. 1998. *The Economics of World War II: Six Great Powers in International Comparison*. Cambridge: Cambridge University Press.

Harrison, M. 2005. The USSR and total war: why didn't the Soviet economy collapse in 1942? In *A World at Total War: Global Conflict and the Politics of Destruction, 1939–1945*, ed. R. Chickering, S. Förster and B. Greiner. Cambridge: Cambridge University Press.

Hautcoeur, P.-C. 2005. Was the Great War a watershed? The economics of World War I in France. In *The Economics of World War I*, ed. S.N. Broadberry and M. Harrison. Cambridge: Cambridge University Press.

Lloyd, E.M.H. 1924. *Experiments in State Control*. Oxford: Oxford University Press.

Maddison, A. 2001. *The World Economy: A Millennial Perspective*. Paris: OECD.

Milward, A.S. 1977. *War, Economy and Society, 1939–1945*. London: Allen Lane.

Milward, A.S. 1984. *The Economic Effects of the Two World Wars on Britain*, 2nd edn. London: Macmillan.

Offer, A. 1989. *The First World War: an Agrarian Interpretation*. Oxford: Clarendon Press.

Olson, M. 1963. *The Economics of the Wartime Shortage: A History of British Food Supplies in the Napoleonic War and in World Wars I and II*. Durham, NC: Duke University Press.

Ordway, F.I. and Sharpe, M.R. 1979. *The Rocket Team*. London: Heinemann.

Overy, R.J. 1995. *Why the Allies Won*. London: Pimlico.

Overy, R.J. 1998. Who really won the arms race? *The Times Literary Supplement* 13(November), 4–5.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Pamuk, S. 2005. The Ottoman economy in World War I. In *The Economics of World War I*, ed. S.N. Broadberry and M. Harrison. Cambridge: Cambridge University Press.

Ránki, G. 1993. *The Economics of the Second World War*. Vienna: Böhlau.

Ritschl, A. 2005. The pity of peace: Germany's economy at war, 1914–1918. In *The Economics of World War I*, ed. S.N. Broadberry and M. Harrison. Cambridge: Cambridge University Press.

Schulze, M.-S. 2005. Austria-Hungary's economy in World War I. In *The Economics of World War I*, ed. S.N. Broadberry and M. Harrison. Cambridge: Cambridge University Press.

Sen, A.K. 1983. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Oxford University Press.

Singer, J.D., ed. 1979; 1980. *The Correlates of War I: Research Origins and Rationale. II: Testing Some Realpolitik Models*. New York: Free Press.

Urlanis, B. 1971. *Wars and Population*. Moscow: Progress Publishers.

Villa, P. 1993. *Une analyse macroéconomique de l'économie française au XXe siècle*. Paris: CNRS.

How to cite this article

Broadberry, Stephen and Mark Harrison. "World Wars, economics of." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_T000224> doi:10.1057/9780230226203.1844

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

X-efficiency

Harvey Leibenstein
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

free-rider problem; incomplete contracts; Prisoner's Dilemma; social norms; X-efficiency.

Article

One has to distinguish the X-efficiency concept from the theory intended to explain it. As a concept X-inefficiency is similar to technical inefficiency. Leibenstein originated the concept of X-inefficiency because of a belief that there is nothing technical about the most substantial sources of non-allocative inefficiencies in organizations. At the time of the original article (Leibenstein, 1966), it seemed that no available concept, such as organizational inefficiency or motivational inefficiency, implied all the elements that could be involved in non-allocative inefficiencies. Hence, the comprehensive term, ‘X-inefficiency’, was used.

X-efficiency theory represents a line of reasoning based on postulates that differs from standard micro theory. A brief statement of the postulates and other elements of the theory follows. (*a*) *Relaxing maximizing behaviour*: it is assumed that some forms of decision making, such as habits, conventions, moral imperatives, standard procedures or emulation, can be and frequently are of a non-maximizing nature. They do not depend on careful calculation. Other decisions attempt at maximizing utility. In order to deal with the max/non-max mixture we use a psychological law, the Yerkes–Dodson Law, which essentially says that at low pressure levels individuals will not put much effort into carefully calculating their decisions, but as pressure builds they move towards more maximizing behaviour. At some point too much pressure can result in disorientation and a lower level of decision performance. (*b*) *Inertia*: we assume that functional relations are surrounded by inert areas, within which changes in certain values of the independent variables do not result in changes of the dependent variable. (*c*) *Incomplete contracts*: we assume the employment contract is incomplete in that the payment side is fairly well specified but the effort side remains mostly unspecified. (*d*) *Discretion*: we assume both that employees have effort discretion within certain boundaries, and that the firm, through its top management, has discretion with respect to working conditions and some aspects of wages.

Under these postulates the firm does not control all of the variables. Rather, the variables are controlled by employees on the one side, and management on the other; both jointly determine the outcome. Thus, this is a standard game-theory type problem. Given the postulates it is easy to suggest that a latent Prisoner's Dilemma problem exists. Employees have an incentive to move towards the minimum-tolerated effort level (E) and the firm has the incentive to move towards the minimum-tolerated working-condition-wage level (W). This is illustrated in Figure 1, where the discretionary effort options run from E_1 to E_n, $E_1<E_i \ldots<E_n$ and the discretionary working-condition – wage options run from W_1 to W_n, $W_1<W_i \ldots <W_n$. Under individual maximizing behaviour employees would want to end up at E_1, and the firm would want to offer W_1. This is the Prisoner's Dilemma solution. The optimal solution is E_nW_n. However, the theory argues that in general the Prisoner's Dilemma solution will be avoided. The reason is that a system of conventions, which depends on the history of human relations within the firm, is likely to lead to an outcome that is usually intermediate between the Prisoner's Dilemma outcome and the optimal solution. In Figure 1 the line with the arrow *MG* represents a locus (one of many) of ‘mutual gain’ situations. That is, for any point on the locus there is a point further up in the direction of the arrow that involves greater effort, greater firm revenues, and a division of the increase in quasi-rents such that both wages and profits are improved. In other words, both the employees and the firm can gain.

Figure 1

Source df Est. sd of coefficients

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

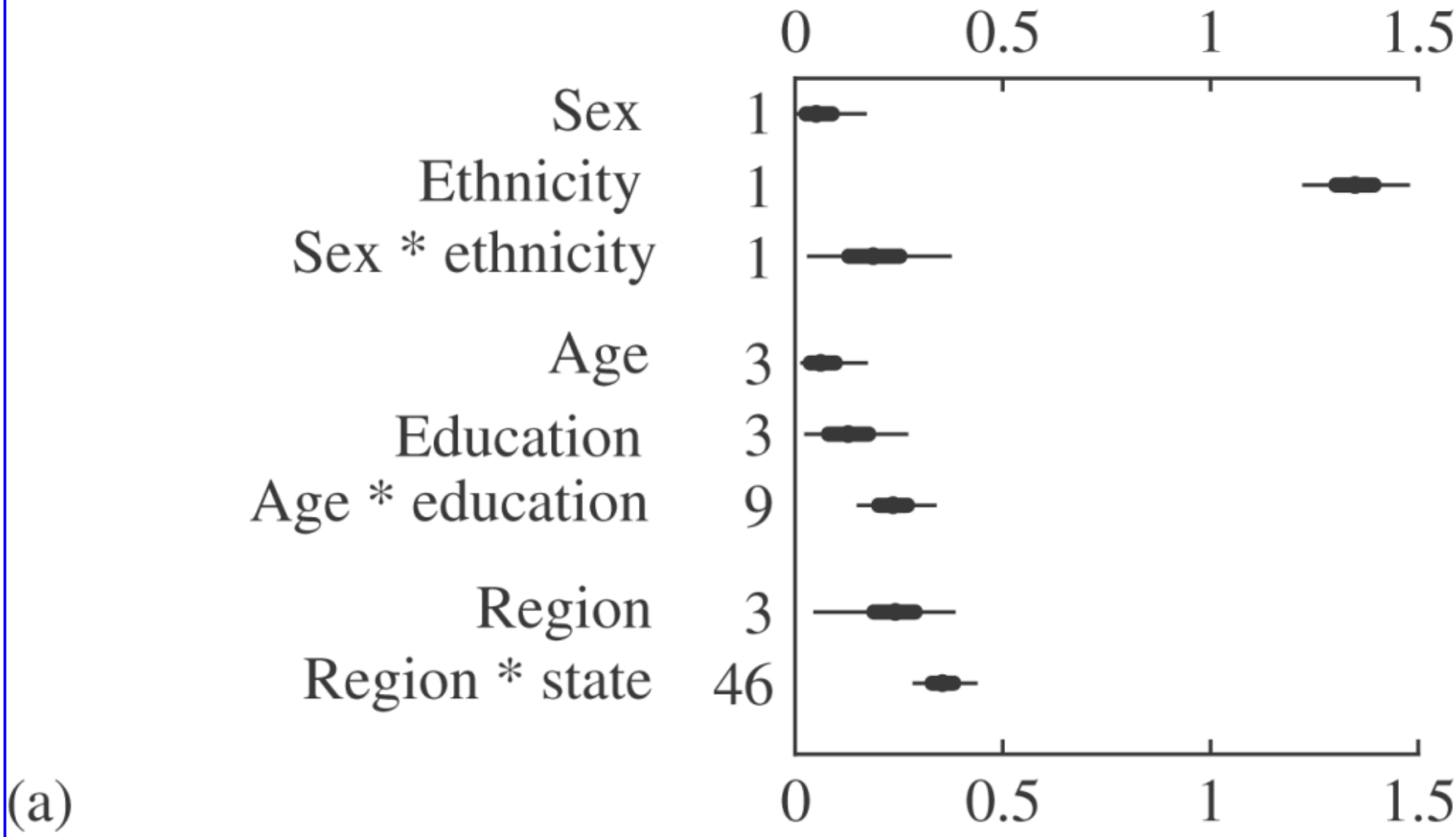

(a)

Source df Est. sd of coefficients

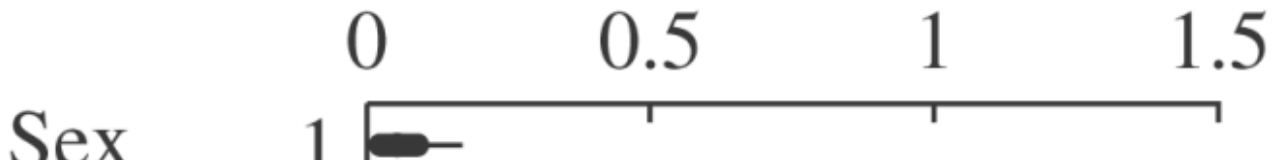

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

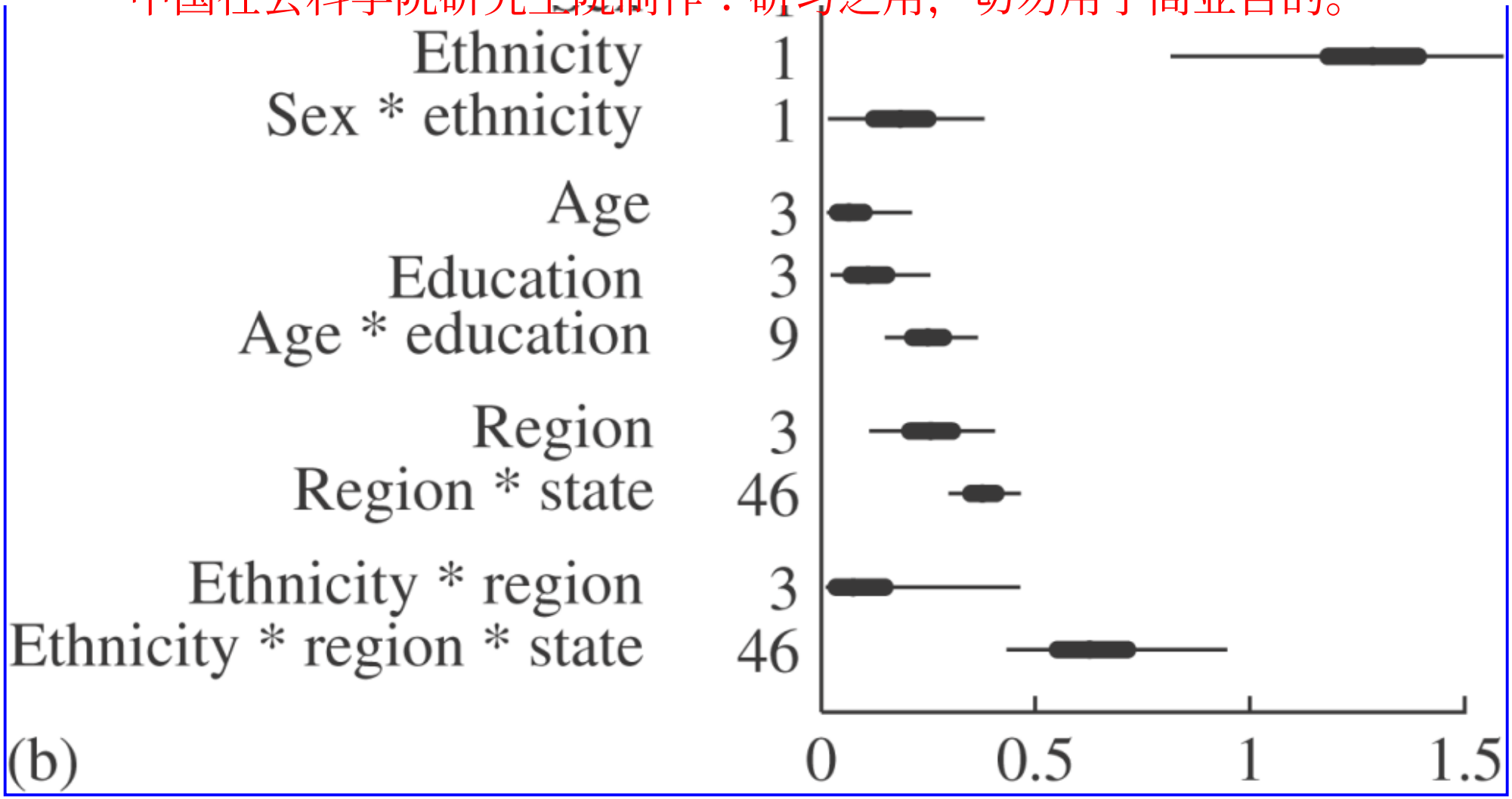

We should note that for every effort option that employees choose the firm will want to choose the minimum wages and working conditions, W_1. Similarly, for every W the firm chooses the employees will want to choose E_1. This is the Prisoner's Dilemma outcome, which the arguments that follow will suggest is not likely to occur. However, this adversarial-relations problem between employees and managers is compounded by another free-rider problem. Every employee has a free-rider incentive to move to the tolerated minimum level E_1, even though he or she might want others to work effectively. Since all employees and managers face these incentives, overall effort would be reduced to the minimum if they all followed their individual self-interest. Clearly, in this organizational situation individual rationality cannot solve the Prisoner's Dilemma problem. Something akin to ‘group rationality’ (see Rapoport, 1970) is required to achieve an improved solution.

A formal theory of conventions (social norms) has been developed in recent years based on the work of T. Schelling, D. Lewis, and E. Ullman-Margalit. The basic ideas are that conventions should be viewed as solutions to multi-equilibrium, coordination problems and that conventions can provide superior solutions to the Prisoner's Dilemma outcome. An example is whether automobiles should be driven on the left or the right. Everyone driving on the right is a desired outcome, and everyone driving on the left is a desired outcome, but a mixture of left-hand and right-hand driving has a negative payoff. Obviously, a convention is required to choose between all left-hand or all right-hand driving. A coordinated solution is superior to an uncoordinated outcome. However, the various coordinated solutions that are possible need not be equally good. Thus, different times of starting work may not be equally preferred, but a coordinated time may still be preferred to an uncoordinated time. Hence, the conventional hours of starting work need not be optimal.

The point of all this is that effort conventions and working-condition conventions can bring about a non-Prisoner's Dilemma solution. This is shown by the point C in Figure 1. The circle surrounding the point represents the inert area surrounding the solution. The distance between C and E_nW_n represents the degree of X-inefficiency in the system. Thus, the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
effort convention is a coordinated solution that is superior to uncoordinated individual behaviour. Similar remarks hold for managerial decisions. Of course, the value of *W* has to be viable in the sense that it must represent a long-run profitable outcome, although not necessarily the maximum profit level.

There is a difference between the creation of a convention and adherence to it. The creation may come about through various means, such as the leadership of some managers, or some employees, or by some initial effort levels being chosen arbitrarily. Once established, a convention reduces the flexibility of employees' behaviour. Thus, new employees will adhere to the convention, and possibly support it through sanctions on others.

Although stable to small changes of its independent variables, an effort convention need not stay at its initial level indefinitely. The concept of inert areas suggests that a large enough shock can destabilize a convention. Once destabilized it is no longer clear whether the dynamics of readjustment will lead to a superior or inferior situation for both sides, or a situation under which one side gains at the expense of the other. Such considerations (and fears) help to stabilize the convention.

It is of interest to note that under low-pressure conditions the postulate of non-maximizing behaviour enables us to recognize and understand why firm members may stick with their conventions and impose supporting sanctions even in situations where they would be better off not doing so. Non-calculating, situation-response behaviour helps to shore up the convention-solution to the Prisoner's Dilemma problem, and to shore up the persistence of non-optimal conventions. This helps to explain the existence and persistence of X-inefficient behaviour.

An illustration of X-inefficient behaviour was described in an article in the *New York Times* (13 October 1981) that compared two identically designed Ford plants, one in the UK and one in Germany, both designed to produce the identical automobile utilizing the same manpower and equipment. Nevertheless, the German plant produced 50 per cent more automobiles than its UK counterpart with 22 per cent *less* labour. Despite the identical plant design, the different effort conventions help to explain the X-inefficient result in the UK plant.

The theory permits a number of inferences to be drawn, some of which (stated without proof) are as follows. Firms generally operate *within* rather than *on* their production frontiers. Given the output, costs per unit are generally not minimized. Innovations are generally not introduced when it is optimal to do so. Less output is not necessarily associated with more desired leisure. The price of the product can have an influence on the cost of production.

There have been a number of measurements of X-inefficiency and empirical tests of its inferences. Professor Roger Frantz (1987) has estimated that over 50 empirical studies exist that either measure the degree of X-inefficiency or provide econometric results that help to confirm the theory.

Bibliography

Frantz, R. 1987. *X-Efficiency: Theory, Evidence, and Applications*. New York: Kluwer-Nyhoff.

Leibenstein, H. 1966. Allocative efficiency vs. ‘X-efficiency’. *American Economic Review* 56, 392–415.

Leibenstein, H. 1976. *Beyond Economic Man*. Cambridge, MA: Harvard University Press.

Leibenstein, H. 1982. The Prisoner's Dilemma in the invisible hand: an analysis of intrafirm productivity. *American Economic Review* 72, 92–7.

Leibenstein, H. 1987. *Inside the Firm: The Inefficiencies of Hierarchy*. Cambridge, MA: Harvard University Press.

Lewis, D. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Rapoport, A. 1970. *N-Person Game Theory*. Ann Arbor: University of Michigan Press.

Schelling, T. 1960. *The Strategy of Conflict*. Oxford: Oxford University Press.

Ullman-Margalit 1977. *The Emergence of Norms*. New York: Oxford University Press.

How to cite this article

Leibenstein, Harvey. "X-efficiency." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_X000001> doi:10.1057/9780230226203.1845

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Yntema, Theodore O. (1900 – 1985)

Jacob L. Mosak
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Committee for Economic Development; international trade theory; price theory; Yntema. T. O.

Article

Yntema was born on 8 April 1900 and died on 18 September 1985. He distinguished himself in the academic world as well as in the worlds of public policy and of business. His training was as varied as his career: he received an AM degree in chemistry in 1922, became a CPA in 1924, an AM in business in 1924, and a Ph.D. in economics in 1929. His Ph.D. thesis, *A Mathematical Reformulation of the General Theory of International Trade* – still in print in 1985 – is an elegant mathematical extension to the field of international trade, of Alfred Marshall's price theory for a domestic economy. The year after receiving his Master's degree in chemistry he began a 25-year academic career at the University of Chicago. After one decade in accounting, he served a second decade as Professor of Statistics, a post which for three years he combined with that of Director of Research of the Cowles Commission for Research in Economics. The last five years he served as Professor of Business and Economic Policy, an area in which he could draw upon his wide experience as economic consultant to United States Government agencies and to private companies. During 1942–9 he served as Director of Research, and, in 1961–7, as Chairman of the Research and Policy Committee, of the Committee for Economic Development (CED), shaping it into one of the most influential public organizations in the field of economic and social policy. In 1949 he embarked upon a new career, going to the Ford Motor Company as Vice-President and Director, and serving as Chairman of its Finance Committee during 1961–5. In that capacity, he was responsible for introducing highly innovative systems of financial management and for the recruitment and supervision of a group of so-called ‘whiz kids’ who helped to revitalize the company, two of whom subsequently became presidents of the company.

How to cite this article

Mosak, Jacob L. "Yntema, Theodore O. (1900–1985)." The New Palgrave Dictionary of Economics.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_Y000006>
doi:10.1057/9780230226203.1846

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Young, Allyn Abbott (1876 – 1929)

Peter Newman
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chamberlin, E. H.; division of labour; external economies; increasing returns; Keynes, J. M.; Knight, F. H.; Pigou, A. C.; roundabout methods of production; supply price; Young, A. A.

Article

Allyn Young's career presents a puzzle. He is best known to modern readers, if at all, as the author of one much-reprinted article on ‘Increasing Returns and Economic Progress’ (1928). With such a narrow base his present fame can of course hardly compare with that of some other leading American economists of his day, such as Irving Fisher, Frank Knight, Wesley Mitchell and Thorstein Veblen. Yet during his life he was very highly regarded indeed, and not just by a US economics profession more insular then than now. To Schumpeter, ‘his published work … [does] … not convey any idea of the width and depth of his thought and still less of what he meant to American economics’ (1954, p. 875, n23). To Keynes, in a letter of consolation to Young's widow, ‘His was the outstanding personality in the economic world and the most lovable’ (Blitch, 1983, p. 22). To Ohlin, he was ‘a man, who knew and thoroughly understood his subject – economics – better than anyone else I have ever met’ (Blitch, 1983, p. 14). The London School of Economics, in the mid-1920s flush with Rockefeller money and looking to make an ‘appointment to the new chair [that] should be so eminent as to be the basis of a major expansion’, chose Young ‘after a prolonged search of the English-speaking world’ (Robbins, 1971, p. 119).

Unlike those other American economists, he never wrote a book. Although he was an incorrigible contributor to such general compendia as the *Encyclopaedia Britannica* and *The Book of Popular Science*, and to such lesser magazines as the *Annalist* and *The Cornell Civil Engineer*, he wrote few major articles. One reason might have been that, as Keynes remarked in the same letter, he ‘would always share with others all his best ideas … it was his own work … which always came last’ (Blitch, 1983, pp. 22–3). Indeed, much of his best work was done through others, two of the great books in economic theory in the first half of the 20th century, Knight (1921) and Chamberlin (1933), originating

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
as doctoral dissertations written under his supervision. Thus like R.F. Kahn, though in the other Cambridge, Young remained an ‘elusive figure who hides in the prefaces of Cambridge books’ (Samuelson, 1947, p. 329).

His peculiar choices of what and how to publish are therefore one explanation of why today he is a minor figure; but they are not the whole, nor perhaps even the major, explanation.

Life

Allyn Abbott Young was born in Kenton, Ohio, on 18 September 1876, the eldest child of two schoolteachers; his given names were the surnames of his two grandmothers. His father was also a lawyer, and active in the Midwestern debates on ‘free silver’ which were part of the political phenomenon that was William Jennings Bryan (see Blitch, 1983, to whose article most of these biographical details are due; Dorfman, 1959, pp. 222–33, is also useful, but sometimes inaccurate). A prodigy, Young entered Hiram College in Ohio at 14 and graduated at 17, having studied languages as well as mathematics and physical sciences. After graduation he worked as a printer for some years in Ohio and Minnesota and saved enough money to further his education, so it was perhaps no accident that the chief example in his famous article (1928) was the printing trades.

In the fall of 1898 he entered the graduate programme in economics at the University of Wisconsin, then and for many years later a major centre of institutional economics. Richard Ely, his chief teacher, was so impressed by Young's ability that in 1899 he secured for him a 15 months’ internship with the staff of the Twelfth Census in Washington, where he met and formed fast friendships with, among others, Wesley Mitchell and Walter Willcox, a professor at Cornell. In the fall of 1900 he returned to Madison and two years later obtained the Ph.D. with a thesis on ‘age statistics’. Two years after that he married a girl from Madison.

Young was unusually restless for the whole of his academic career, which began in 1902. He first spent two years at Western Reserve University in Cleveland, then one at Dartmouth College, one back at Wisconsin, four at the new Stanford University (where he was the first chairman but failed twice to persuade President Jordan that Veblen should be promoted to full professor), one as a visitor at Harvard, and two as chairman at Washington University in St Louis. The pace slowed somewhat with his appointment in 1913 at Cornell, but even there he took leave of absence for two years in 1918–19 to work with the Wilson Administration on preparations for the peace conference (Blitch has a long account of this episode). He received the traditional ‘call from Harvard’ in 1920 but instead returned dutifully to Cornell. However, the next year he yielded to the renewed invitation of the Harvard department and soon became one of its most popular and respected members, among faculty and graduate students alike.

But Young was too complex a man to stay long even at Harvard, that absorbing barrier of so many academic random walks. Instead, he accepted the attractive offer from London, for a period of three years and at a salary well above the usual English professorial level (Blitch says that ‘It was the first time that a chair in a British university had been offered to an American’). Unfortunately, according to Robbins (1971, p. 121), Young ‘gave … the impression of a profoundly unhappy man’ in the job, and in fact decided to return to Harvard when his three years were up, in spite of a handsome offer from Chicago. Tragically, in the winter of 1928–9 he became a victim of a severe influenza epidemic and very

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
quickly succumbed to pneumonia, dying in London on 7 March 1929 at the early age of 52.

Character

In appearance Young was tall and thickset, looking according to E.S. Mason ‘more of a poet than an economist’ (Blitch, 1983, p. 13). He was well-read in his own subject and many others, a musician, ‘singularly unworldly’, and famously absent-minded, sometimes having to be summoned to his lectures by colleagues or students.

His reputation stood very high as a teacher, but perhaps more as a supervisor of dissertations and discussant of work in progress than as a formal lecturer, where apparently he was given to long silences while he puzzled out what he wanted to say. Lauchlin Currie said that he ‘gave the impression of *thinking* as he went along’ (Blitch, 1983, p. 14). Such teaching was more suited to graduates or brilliant undergraduates like Kaldor than to an average undergraduate audience, and Robbins's report of the same style as Young practised it in England was distinctly cool: ‘The more frivolous spirits … would compile betting books on the length in seconds of the longest interval.’ He was still harsher on Young's poor administrative ability: ‘after his untimely death … [there was] … a condition of almost unimaginable confusion, no order or system anywhere’ (1971, p. 120).

In spite of this alleged lack, Young seems to have been a successful chairman at Stanford and in St Louis. Moreover, he was a loyal member of his profession, serving as Secretary of the American Economic Association (AEA) from 1914 until 1920, as President of the American Statistical Association in 1917, and in 1925, after Veblen had refused the position (Dorfman, 1934, 491–2), as President of the AEA. His most famous paper (1928) was actually delivered as the Presidential Address of Section F of the British Association, ‘the first American to be so honored’ (Blitch, 1983, p. 18).

His feelings of loyalty may on occasion have affected his judgement. As a member of Wilson's delegation to the Peace Conference in 1919 he had independently arrived at much the same position towards the Treaty that Keynes developed with such force in *The Economic Consequences of the Peace*, but unlike Keynes could not bring himself to make a clean break with his government. On his return to the United States he reviewed Keynes’ book in the *New Republic* (1919–20), and privately protested to Keynes at the latter's account of Wilson's behaviour. Keynes wrote a placatory reply, saying that ‘I still believe that essentially the President played a nobler part at Paris than any of his colleagues’ (1977, p. 45). In what he later called ‘an indiscretion which I regret’ (1977, p. 48), at a public debate Young quoted without permission from Keynes's letter, inadvertently making it appear that in the *Consequences* Keynes had consciously distorted the truth about Wilson for his own propagandist purposes. Keynes was furious – ‘Young was very wrong … to make reference to a private letter’ – and threatened to publish the whole correspondence. However, Young finally wrote a letter of apology to the New York *Evening Post*, which to Keynes seemed ‘quite satisfactory … I am now quite content to let the matter drop’ (1977, p. 49).

Forgiven, then; but not forgotten. In *A Revision of the Treaty* (1922, p. 3n) Keynes first quoted the reference in Young's original review to the Treaty's ‘timorous failure to reckon with economic realities’, and then scathingly remarked: ‘Yet Professor Young has thought right, nevertheless, to make himself a partial apologist of the Treaty, and to describe it as a “forward-looking document”.’ Young reviewed the new book, wrote an irenic letter to Keynes (Harrod, 1951, p. 312) and the episode finally closed,

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
apparently (to judge from Keynes's letter to Young's widow) without seriously affecting their mutual regard.

Works

Young's range as an economist was unusually wide, even in an age when it was easier to be at the frontier of research in several fields than it is now. Beginning in what would now be called demography (1900–1), he worked and published in several areas of applied economics, such as public utility regulation, antitrust policy, banking, index numbers, public finance, income distribution, and problems of war finance and reparations (for example, 1922–3). In the nature of the case, however, little of this applied work was of lasting importance, and so rather unjustly it will be his work in theory and doctrine that receives attention here.

It is typical of him that Young should have made his first major theoretical impact with two review articles, rather than with independently original work. The first, appearing in 1912 when he was already 36 and an established member of the profession, was a review of the fourth edition of Jevons's *Theory of Political Economy*, appearing 40 years after the original edition of 1871. Thus Young's review (reprinted in 1927) was perforce an essay in the history of economic thought rather than current economics. As such it was a penetrating contribution, well worth reading today and by no means inferior to some modern assessments of Jevons.

Matters were quite different with the second review article (1913), on Pigou's recently published *Wealth and Welfare*. It was this that made Young's international reputation, for he was the first to point out a basic flaw in Pigou's reasoning. The excess of ‘marginal supply price’ over supply price that Pigou saw as a reason for taxing decreasing returns industries turned out to be in Young's argument almost entirely a matter of those increases in rents of the relatively scarcer factors by which necessary transfers of resources are accomplished, and certainly does not correspond to increased real usage of resources. The international impact of this fundamental criticism was probably all the greater because at that time one did not look for such subtle general equilibrium reasoning to come from the heavily empirically-minded US profession. However, for reasons which were hardly Young's fault, it failed to sweep away as it should all arguments of the Pigovian kind, so that later Young's student Knight (1924) felt impelled to repeat essentially the same point, in an article that is today much better known than Young's original criticism.

It is curious that Young did not see fit to reprint this article in his collection (1927), which included far inferior pieces. A kind and modest man, possibly he did not want to upstage his friend Knight's recently published article on the same subject. It is far more likely however, that he did not want to give renewed currency to a view which, in 1927, he almost certainly no longer held. To Pigou's claim (1912, p. 177) that ‘Provided that certain external economies are common to all the suppliers jointly, the presence of increasing returns in respect of all together is compatible with the presence of diminishing returns in respect of the special work of each severally’, Young had in 1913 made the terse dismissive comment that ‘I cannot imagine “external economies” adequate to bring about this result’ (p. 678n). But it was precisely Young's vivid and convincing vision of such external economies as the main vehicle of increasing returns and economic progress that was the centrepiece of his famous article a year later, in 1928.

Between 1913 and 1928 Young contributed no major articles in economic theory. One reason may have

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
been his work on Ely's *Outlines of Economics*, which with Thomas Adams and Max Lorenz he had helped to revise in 1908 and whose revisions he supervised in 1916 and 1923, writing the whole of ten chapters and parts of others (Dorfman, 1959, p. 222 n. 2). This work was by far the most popular college textbook in economics in America, selling a total of 350,000 copies and outstripping its chief rival (Taussig's *Principles*) by two to one (Dorfman, 1959, p. 211n). Of much greater importance must have been his work arising from the peace conference, which seems to have preoccupied him for several years from 1918 on.

His paper of 1928 is not quite the isolated phenomenon that it appears. Several of his contributions to the great 14th edition (1929) of the *Encyclopaedia Britannica* are consistent with the approach taken in (1928), particularly his entry on ‘Capital’. This is not surprising, for the chief analytical innovation in (1928), as distinct from its new ‘vision’ of economic progress, was to make the degree of roundaboutness depend primarily not on the rate of interest but on the scale of production, taken in a broad sense.

Although no brief summary can do justice to Young's vision and its details as set out in (1928), the following passage taken from the essay on ‘Capital’ (1929, vol. 4, p. 796) is a modest if inadequate substitute for reading the paper itself:

> There is nothing inherently economical in roundabout methods, but the most economical methods often happen to be roundabout. The degree of roundaboutness which is most economical generally depends upon the amount of a particular kind of work which is to be done. And also the making and use of instruments involves an extension of the principle of the division of labour, and the division of labour, as Adam Smith observed, depends upon the extent of the market. The use of capital on a large scale in industry came later than its use in commerce, for the reason that not until there were markets which were able to absorb large outputs of standard types of goods was it profitable to make any extensive use of roundabout methods of production. Once established, however, industrial capitalism showed that it had within itself the seeds of its own growth. Cheaper goods, improved means of transport, and the increased advantages of specialization led to larger markets, so that the economies of industrial capitalism grew in a cumulative way. The increasing division of labour, by breaking up complex industrial processes into simpler parts, not only invited a larger use of instruments, but also prompted the invention of new types of instrument.

Apart from an interesting discussion by Marx, Young's article was the first serious advance beyond Adam Smith on the relations between increasing returns and economic growth. However, the problems of formalizing that persuasive vision into a tractable model have proved formidable indeed, the chief technical problems being those of non-convex technologies and the introduction of new intermediate commodities. So, old as it is, his paper remains important for us precisely because there is not much else. Although there have recently been some encouraging signs that this long drought may be coming to an end, these are not yet sufficient either in number or in quality to predict definitely that it will.

Conclusion

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

As a critic, Young was knowledgeable and perceptive and possessed of that rare ability of entering into and appreciating minds that were quite unlike his own. This can be seen not only in his reviews of Jevons and Pigou and Edgeworth (1925) and his work with countless graduate students (not just stars like Knight and Chamberlin), but in a wider perspective. Thus although he had a populist upbringing in the Midwest and was trained in the very citadel of institutional economics, it is clear from his writings that by nature he had much greater affinities with the modes of thought of traditional economic theory. To that extent he was seriously handicapped in living in a time and place that was at best atheoretical. Nevertheless, his sympathies were so far extended that he could refer to Veblen as ‘the most gifted man I have know’ (Dorfman, 1961, p. 299) and could review much too kindly the institutionalist manifesto edited by Tugwell (1924), trying hard to see the mazy merits of its case.

Perhaps that is the key to the puzzle about Allyn Young. He was above all a great critic, and great critics, like great journalists and great wits, seldom survive into posterity.

See Also

- division of labour

Selected works

1900–1901. The comparative accuracy of different forms of quinquennial age groups. *Publications of the American Statistical Association* 7, 27–39.

1908 (With T.S. Adams and M.O. Lorenz.) Revised edn of R.T. Ely, *Outlines of Economics*. New York: Macmillan. Subsequent revisions in 1916, 1923 and 1930.

1913. Pigou's Wealth and Welfare. *Quarterly Journal of Economics* 27, 672–86.

1919–20. The economics of the treaty. *New Republic* 21, 388–9.

1922–3. The United States and reparations. *Foreign Affairs* 1, 35–47.

1925. Papers relating to political economy. *American Economic Review* 15, 721–4.

1927. *Economic Problems New and Old*. Boston: Houghton Mifflin. (This contains 14 papers, dated 1911 through 1927, that are not included in the present list.)

1928. Increasing returns and economic progress. *Economic Journal* 38, 527–42.

1929. Eleven articles, published posthumously: Capital; Economics; Labour; Land; Price; Rent; Supply and Demand; Utility; Wages; Wealth; Value. In *The Encyclopaedia Britannica*, 14th edn. London: Encyclopaedia Britannica Company.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Bibliography

Blitch, C.P. 1983. Allyn A. Young a curious case of professional neglect. *History of Political Economy* 15, 1–24.

Chamberlin, E.H. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Dorfman, J. 1934. *Thorstein Veblen and his America*. New York: Viking Press. Reprinted with new appendices, New York: Augustus M. Kelley, 1961.

Dorfman, J. 1959. *The Economic Mind in American Civilization*. Vol. 4. New York: Viking Press.

Harrod, R.F. 1951. *The Life of John Maynard Keynes*. London: Macmillan.

Keynes, J.M. 1919. *The Economic Consequences of the Peace*. London: Macmillan.

Keynes, J.M. 1922. *A Revision of the Treaty*. London: Macmillan.

Keynes, J.M. 1977. *Activities 1920–1922: treaty revision and reconstruction*. In *The Collected Writings of John Maynard Keynes*, vol. 17, ed. E. Johnson. London: Macmillan for the Royal Economic Society.

Knight, F.H. 1921. *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin.

Knight, F.H. 1924. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38, 582–606.

Pigou, A.C. 1912. *Wealth and Welfare*. London: Macmillan.

Robbins, L. 1971. *Autobiography of an Economist*. London: Macmillan.

Samuelson, P.A. 1947. ‘The General Theory’. In *The New Economics: Keynes’ Influence on Theory and Policy*, ed. S. Harris. New York: Alfred A. Knopf.

Schumpeter, J.A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Tugwell, R.G., ed. 1924. *The Trend of Economics*. New York: Alfred A. Knopf.

How to cite this article

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

Newman, Peter. "Young, Allyn Abbott (1876–1929)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_Y000007>
doi:10.1057/9780230226203.1847

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Young, Arthur (1741 – 1820)

K. Tribe

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

political arithmetic; Young, A.

Article

Born into a Suffolk clerical family in 1741, Arthur Young began his literary career at 17, writing novels and pamphlets. He began farming in his early twenties, and in 1767 he took on the tenancy of an Essex farm which was however beyond his means. Exchanging this tenancy for a smaller farm in Hertfordshire which proved equally unrewarding, his income during the 1770s was drawn as much from writing as farming. The publication of his accounts of travels in England and Ireland met with great success, but the long absences and expense involved led to the neglect of his farm. His response was to extend his literary activities: in 1784 he launched the *Annals of Agriculture*, a journal which rapidly gained international recognition. During the later 1780s he toured the Continent, and his observations of France on the eve of the Revolution remain a valuable source. On the establishment of the Board of Agriculture in 1793 Young became its Secretary, and it was his descriptive methods which were followed by the writers of the County Surveys for which the Board is best known. In 1811 Young became blind, and he died in 1820, leaving behind an autobiography which provided a social and personal record of a life devoted to farming and literary activities.

Young worked until his death on a book entitled *Elements of Agriculture*, a general survey of agricultural method and practice; but the work was never published, and indeed it is evident that Young's strength lay in his observational method and his systematic appraisal of rational agriculture, rather than his ability to produce a general survey of agricultural conditions. This is evident from his *Political Arithmetic* of 1774 and 1779, which, while combining an account of the agricultural state of England with commentary on the writings of Steuart, Davenant and the Physiocrats, lacks the originality of his agricultural writings. Even in the first of these writings, *The Farmer's Letters* (1767), he supported arguments for progressive husbandry and agrarian reform with the construction of ‘model farms’ which postulate ideal combinations of land, labour and capital for differing conditions of fertility and husbandry. This pattern was repeated in his *Six Weeks’ Tour* (1768) and also in many subsequent works

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
which record and comment upon the condition of farms and estates encountered on his route. Experiments carried out on his farm were also written up, while in *The Farmer's Guide* he produced a handbook for the establishment and conduct of a tenant farm – together with his *Rural Oeconomy*, an outline of capitalist farm management.

Visitors to Young's own farm were often struck by his failure to follow his own recommended ‘best practice’ – instead of order and efficient supervision, they found disarray and confusion. It may be that Young devoted too much time to writing to properly supervise the work of the farm (the true business of the farmer in his view), or he may simply have lacked the application necessary; but his practical shortcomings do not diminish his real achievement – the establishment of a system of agricultural observation which was a model for 19th-century studies of agricultural production.

Selected works

1767. *The Farmer's Letters to the People of England*. London.

1768. *A Six Weeks’ Tour through the Southern Counties of England and Wales*. London.

1770a. *The Farmer's Guide in Hiring and Stocking Farms*. 2 vols. London.

1770b. *A Course of Experimental Agriculture*. 2 vols. London.

1770c. *Rural Oeconomy*. London.

1771a. *The Farmer's Tour through the East of England*. 4 vols. London: W. Strahan.

1771b. *The Farmer's Calendar*. London.

1774, 1779. *Political Arithmetic*. Pt I, London: W. Nicoll, 1774. Pt II, London: T. Cadell, 1779.

1780. *A Tour in Ireland*. London.

1784–1815. *Annals of Agriculture* (1784–1815). 46 vols.

1792. *Travels during the Years 1787, 1788 and 1789 … in France*. Bury St. Edmunds.

1794–1809. *General View of the Agriculture of the County of Suffolk* (1794); also on Lincoln (1799), Hertfordshire (1804), Norfolk (1804), Essex (1807), Oxford (1809). London: Board of Agriculture.

1808. *General Report on Enclosures*. London: Board of Agriculture.

1809. *On the Advantages which have resulted from the Establishment of the Board of Agriculture*.

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

London.

1898. *The Autobiography of Arthur Young*, ed. M. Betham-Edwards. London: Smith, Elder & Co.

How to cite this article

Tribe, K. "Young, Arthur (1741–1820)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_Y000008>
doi:10.1057/9780230226203.1848

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Zawadzki, Wladyslaw Marian (1885 – 1939)

Stanislaw Swianiewicz

From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

economic sociology; general equilibrium; production theory; value theory; Zawadzki, W. M.

Article

The Polish economist W.M. Zawadzki, was born in Vilno, historic capital of Lithuania, and studied mathematics and social sciences in Moscow, Leipzig and Paris.
His first major work, *Les Mathématiques appliquées à l'économie politique*, was published in 1914 in Paris. After the First World War this book was recognized as a significant contribution to the development of general equilibrium theory as conceived by Vilfredo Pareto. At the beginning of his academic career, Zawadzki was also involved in the study of the role of theories of value in the history of economic thought. He came to the conclusion that, contrary to the prevailing view of the time, the idea of exchange value as something different from market prices should be discarded. This point of view is further developed in the book on *Value and Price*, containing extracts from the works of various economists and published in Polish under the editorship of Zawadzki in 1919. When the Econometric Society was founded in 1931 Zawadzki became one of its original members and was elected to the first committee of the society.
In 1919 Zawadzki contributed to the re-establishment of the ancient University of Vilno, closed down 80 years before by the Russians, and became there the Professor of Political Economy. At that time Zawadzki following the example of Vilfredo Pareto turned his attention to the problems of economic sociology. This found expression in his second major work *The Theory of Production*, published in 1923 in Polish and in 1925, with certain abbreviations, in French – an extensive study of production in various social, cultural and technical environments. Zawadzki distinguishes five types of environments in which regular production can take place: primitive, partriarchal, individualistic, based on compulsion and collectivist. The main part of the book is devoted to the analysis of the conditions of production in an individualistic environment. While admitting the feasibility of regular production in a collectivist society, Zawadzki is sceptical about the Marxist view that the conditions for collectivist systems can gradually emerge in the course of development of individualistic society. But he admits the possibility of

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
a revolutionary transition to collectivism as preached by George Sorel and the French syndicalist revolutionaries.

Zawadzki was the Polish Minister of Finance in 1931–5 and was to a very great extent responsible for keeping his country on the gold standard in spite of the abandonment of that system by Britain and the USA.

From 1936 until his death in 1939 Zawadzki was Professor of Economics in the Central School of Commerce, Warsaw. During that period he worked mainly in the field of monetary theory and had considerable influence as a teacher – several young economists adopted his ideas and tried to develop them.

Selected works

1914a. *Les Mathématiques appliqées à l'économie politique*. Paris.

1914b. *Zastosowanie matematyki do ekonomji politycznej* [The application of mathematics to political economy]. Wilno.

1919. *Wartosc i cena. Wypisy* [Value and price. An economic selection]. Warsaw.

1923. *Teorja produkcji* [The theory of production]. Warsaw.

1927. *Esquisse d'une théorie de la production*. Paris.

1938. *Manipulowanie pieniadzen jako narzedzie polityki gospodarczej* [The manipulation of money as an instrument of economic policy]. Krakow.

How to cite this article

Swianiewicz, Stanislaw. "Zawadzki, Wladyslaw Marian (1885–1939)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_Z000001> doi:10.1057/9780230226203.1849

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。

The New Palgrave Dictionary of Economics Online

Zeuthen, Frederik Ludvig Bang (1888 – 1959)

Hans Brems
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Keywords

Chamberlin, E. H.; collective bargaining; duopoly; product differentiation; Robinson, J. V.; Zeuthen, F. L. B.

Article

Zeuthen was born on 9 September 1888 in Copenhagen. He took a degree in economics at the University of Copenhagen in 1912 and spent the next 18 years in the service of the Danish social security system. That system, already then full-fledged by American standards, was the subject matter of a large number of books and articles published by Zeuthen in the period 1912–28. But then, at 40, Zeuthen published his *Fordeling* (1928) in which is found, among other things, his use of inequalities in a Walras system as well as his theory of collective bargaining. The following year he published his article (1929) on product differentiation under monopolistic competition. Zeuthen's treatment of collective bargaining and monopolistic competition appeared in English (1930) with a preface by Schumpeter, who called it a ‘bold raid into new and difficult country’. The new country would soon become part of mainstream economic theory. Also in English (1957), Zeuthen gave us his mature views on all this. He taught theory, labour economics and social security at his alma mater from 1930 to 1958 and died on 24 February 1959 in Copenhagen.

Inequalities in a Walrasian system

In a Walrasian system with fixed input-output coefficients, Zeuthen (1928, p. 27; 1932–3, pp. 2–3) saw that feasibility would require the sum of all inputs of any good absorbed in all processes to be smaller than or equal to the sum of all outputs of it supplied in all processes. By introducing a new variable, that is, the unused portion, Zeuthen could then turn his inequality into an equality and say that either the unused portion of the input or the price of the input would be equal to zero.
In a short paper Schlesinger (1935) agreed, but neither Zeuthen nor Schlesinger attempted to prove the

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
existence of a general equilibrium. Wald (1935; 1936) made the attempt for a stationary economy, and von Neumann (1937) succeeded for a growing one. Von Neumann formulated a primal and a dual problem. His primal problem was to maximize the rate of growth subject to the constraint that excess demand for any good must be non-positive. That constraint was what Zeuthen and Schlesinger had seen and Wald had worked on.

Von Neumann's dual problem was this. We must minimize the rate of interest subject to the constraint that in any time-consuming process profits must be non-positive. That constraint was seen by neither Zeuthen nor Schlesinger. Taking his primal and his dual together, von Neumann found his familiar existence proof in which the maximized rate of growth equalled the minimized rate of interest.

Collective bargaining

Inherent in labour–management bargaining is the threat of conflict. Zeuthen's (1928; 1930, pp. 104–50) point of departure was the net outcome of a possible conflict as expected by labour and management, respectively.

As seen by labour, let the net outcome be $w(L)$ defined as the money wage rate after conflict reduced by labour's cost of conflict, that is, lost wages. As seen by management, let the net outcome be $w(M)$ defined as the money wage rate after conflict raised by management's cost of conflict, that is, lost orders.

The lower bound to a negotiated money wage rate will then be what labour expects to live with after a possible conflict, that is, $w(L)$. A negotiated money wage rate equal to its lower bound $w(L)$ would leave labour indifferent and management eager to secure an agreement. To shake labour out of its indifference, management may be willing to raise the suggested wage rate. Thus suggestions at or near the lower bound $w(L)$ will very likely be abandoned in favour of higher ones.

The upper bound to a negotiated money wage rate will be what management expects to live with after a possible conflict, that is, $w(M)$. A negotiated money wage rate equal to its upper bound $w(M)$ would leave management indifferent and labour eager to secure an agreement. To shake management out of its indifference, labour may be willing to pare down the suggested wage rate. Thus suggestions at or near the upper bound $w(M)$ will very likely be abandoned in favour of lower ones.

Having established the existence of such centripetal forces – powerful near the bounds of the bargaining range, weaker towards its centre – Zeuthen found his negotiated money wage rate moving towards a point in which no party was more eager to secure an agreement than the other.

Zeuthen's theory of an economic conflict and its resolution may well have been the first ever.

Product differentiation

Like Bertrand, Zeuthen (1929; 1930, pp. 24–5) assumed his duopolists to have a price policy but cautiously removed Cournot's and Bertrand's assumption of ‘*qualité indentique*’. Zeuthen's product differentiation consisted in differences in product quality, geographical location, or advertising-generated image. Such product differentiation was not strong enough to protect a duopolist failing to match a price cut but strong enough to allow a duopolist matching any price cut to keep his old customers. Like Cournot and Bertrand duopolists, Zeuthen's duopolists would always sell at a common price but could attract new customers by lowering it. A ‘coefficient of expansion’ measured a duopolist's

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
ability to attract them by differences in product quality, geographical location, or advertising-generated image.

Unlike Joan Robinson's one-firm model three years later, Zeuthen's model had two firms interacting in a group equilibrium. Unlike Edward Chamberlin's group equilibrium three years later, Zeuthen's group equilibrium did not assume equal market shares: the more successful firm had the larger coefficient of expansion, hence the larger market share.

See Also

- bargaining

Selected works

1928. *Den konomiske Fordeling* [Economic distribution].Copenhagen: Busck.

1929. Mellem Konkurrence og Monopol [Between competition and monopoly]. *Nationalkonomisk Tidsskrift* 67, 265–305.

1930. *Problems of Monopoly and Economic Warfare*. With a preface by J.A. Schumpeter. London: Routledge.

1932–3. Das Prinzip der Knappheit, technische Kombination und ökonomische Qualität. *Zeitschrift für Nationalökonomie* 4, 1–24.

1957. *Economic Theory and Method*. Cambridge, MA: Harvard University Press.

Bibliography

Baumol, W.J. and Goldfeld, S.M., eds. 1968. *Precursors in Mathematical Economics: An Anthology*. London: London School of Economics and Political Science.

Neumann, J. Von. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums* 8. Trans. G. Morgenstern in Baumol and Goldfeld (1968).

Schlesinger, K. 1935. Über die Produktionsgleichungen der ökonomischen Wertlehre. *Ergebnisse eines mathematischen Kolloquiums* 6. Trans. W.J. Baumol in Baumol and Goldfeld (1968).

Wald, A. 1935. Über die eindeutige positive Löbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines mathematischen kolloquiums* 6. Trans. W.J. Baumol in Baumol and Goldfeld (1968).

中国社会科学院研究生院制作：研习之用，切勿用于商业目的。
Wald, A. 1936. Über die Produktionsgleichungen der ökonomischen Wertlehre. *Ergebnisse eines mathematischen Kolloquiums* 7. Trans. W.J. Baumol in Baumol and Goldfeld (1968).

How to cite this article

Brems, Hans. "Zeuthen, Frederik Ludvig Bang (1888–1959)." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 03 January 2009 <http://0-www.dictionaryofeconomics.com.library.lemoyne.edu/article?id=pde2008_Z000005> doi:10.1057/9780230226203.1850

Austrian economics: recent work

Mario J. Rizzo
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume
Back to top

Abstract

This article reviews research in Austrian economics over the last 25 years, relating it to (but not discussing in detail) earlier classic work in the Austrian tradition. Core issues are business cycle theory, entrepreneurship, market processes and economic institutions, the communication of knowledge in markets, spontaneous order, and issues related to law and economics.
Back to top

Keywords

Austrian Business Cycle Theory; Austrian economics; entrepreneurship; Hayek, F.A.; law and economics; market processes; monetary theory; spontaneous order; time
Back to top

Article

Back to top

Introduction

In the past 25 years, a large amount of new research in Austrian economics has developed and expanded the basic themes that are central to its unique identity (O'Driscoll and Rizzo, 1996). These highly interrelated themes are (1) the subjective, yet socially embedded, quality of human decision making; (2) the individual's perception of the passage of time ('real time'); (3) the radical uncertainty of expectations; (4) the decentralization of explicit and tacit knowledge in society; (5) the dynamic market processes generated by individual action, especially entrepreneurship; (6) the function of the price system in transmitting knowledge; (7) the supplementary role of cultural norms and other cultural products ('institutions') in conveying knowledge; and (8) the

中国社会科学院研究生院：研修之用

spontaneous - that is, not centrally directed - evolution of social institutions. The specific ways in which these themes have recently manifested themselves is the subject of this article.
Since our task is to discuss the developments in Austrian economics primarily since the last *New Palgrave* entry (1987) we shall not review the work of the many 'classic' Austrian authors. In addition, since our concerns are the substantive developments in the field, we omit many valuable contributions in the history of economic thought and in methodology.
Back to top

Macroeconomics and monetary theory

There have been many advances in Austrian macroeconomics. These include new work on business cycle theory and on alternative monetary institutions.
Each of these areas can be looked at from the general perspective of treating time, money and their related institutions seriously (Horwitz, 2000). Time is the medium of all action. Decisions are taken in time to produce consequences in the future. Taking time seriously means also taking the uncertainty that characterizes these decisions seriously. This applies to savings-investment choices, production plans, and the time structure of capital goods. In an Austrian (and Keynesian) perspective the pervasive uncertainty of the future makes money necessary. Thus, as time is the medium of all action, money is the medium of all exchange. All goods markets are accordingly affected by the supply and demand for money and the nature of monetary institutions.
The Austrian Business Cycle Theory (ABCT) received a major systemization and refinement in the work of Roger Garrison, culminating in his book, *Time and Money: The Macroeconomics of Capital Structure* (2001). The previous work in the subject was scattered in many articles by Friedrich Hayek and in the work of Ludwig von Mises. It was also very imperfectly linked to the brilliant, but underrated, work by Ludwig Lachmann, *Capital and its Structure* (1956). Garrison corrects these deficiencies and adds coherence to ABCT which had previously been unknown. In a sense, Garrison has done for ABCT what John Hicks and Alvin Hansen did for Keynes's macroeconomics, except that the Garrison's work is an accurate rendition of Hayek, Mises and Lachmann.
The subtitle of Garrison's book, 'The Macroeconomics of Capital Structure', expresses the important claim that Austrian macroeconomics cannot adequately be appreciated without understanding that 'investment' is not a homogeneous decision. This insight is developed at length by Peter Lewin in *Capital in Disequilibrium: The Role of Capital in a Changing World* (1999), the most important work in Austrian capital

中国社会科学院研究生院：研修之用

theory in many decades (see also Endres and Harper, 2008). The ABCT focuses on the inappropriateness of the capital structure (malinvestment) generated by artificially low real interest rates (that is, interest rates that are lower than the real supply of savings would allow). Thus, the term over-investment is, by itself, a misleading characterization of the ABCT process. While excessively low interest rates do increase the level of investment relative to its previous position, they do so in a biased way - those stages of production further from consumption are affected to a greater extent.

However, as Garrison's recent work (2004) has shown, there are even more widespread distortions in the production structure generated by artificially low interest rates. These include initial 'overconsumption' as the result of reduced savings and of increased incomes on the part of factors of production. Increased investment in close temporal proximity to the overconsumption is labeled the 'derived demand effect'. This is in addition to the 'discount effect', described above, which increases the profitability of new investment distant from consumption. These two contrary effects come at the expense of intermediate stages of production as well as reduced maintenance of existing capital at all stages. They may even result from the utilization of unused resources during periods of less than full employment. These effects show that the ABCT is a type of 'coordinationist macroeconomics' insofar as it describes the discoordination of various sectors of the economy, and is not simply a micro choice-theoretic approach to macroeconomics (Wagner, 2005).

Accordingly, in this Austrian view recessions are characterized not simply by low levels of aggregate economic activity but also by the misdirection of resources caused by previous boom-induced malinvestments. These systematic sectoral imbalances - too much investment in interest-sensitive areas of economic activity - must be corrected as recovery proceeds.

The Austrian theory, however, is not a complete theory of the business cycle. It accounts mainly for the process leading to and including the cycle's upper turning point. It is a theory of the crisis. How long the resulting recession lasts is not predicted by the theory or even, strictly speaking, by the degree to which resources were misallocated. The length of the recession will depend, for example, on those factors affecting the mobility of resources.

None of this implies that Hayek, Garrison or Horwitz are insensitive to the problems that would be induced by an aggregate increase in the demand to hold money (a fall in income velocity), which can accompany recessions. This 'secondary deflation' should be avoided by a concomitant increase in the supply of money by the relevant monetary institutions. Horwitz (2000) is the first to integrate Austrian macroeconomics with monetary

中国社会科学院研究生院：研修之用

disequilibrium theory to analyse deflationary processes. Nevertheless, recessions are not primarily deflationary phenomena (or at least need not be), but occasions for correction of the misdirection of resources. Some Austrians, however, argue that increases in the demand for money have significant negative consequences only in the presence of legal restraints on price flexibility (Salerno, 2003).

One of the most important possible obstacles to recovery from recessions may be in the behaviour of 'big players'. These are agents whose discretionary behaviour, insulated from the normal discipline of profit and loss, can significantly affect the course of economic effects (Koppl and Mramor, 2003; Koppl, 2002; Koppl and Yeager, 1996). Thus, discretionary behaviour on the part of monetary authorities (in the United States, the Fed), fiscal policy makers (Congress or the Executive), or even in some cases private monopolists, can increase uncertainty faced by most economic agents ('small players'). They will have to pay more attention to trying to guess the perhaps idiosyncratic behaviour of the big players. Economic variables will become contaminated with big-player influence. It will become more difficult to extract knowledge of fundamentals from actual market prices. And thus entrepreneurs will find it harder to determine where resources should be withdrawn and where they should be added in a way that is sustainable in the medium to long term.

An important variant of the ABCT in *Risk and Business Cycles: New and Old Austrian Perspectives* (1998), developed by Tyler Cowen, focuses on the integration of business cycle theory with developments in modern finance. The main sense in which this can be called a variant of ABCT is that changes in the riskiness of investment decisions are linked to the 'old Austrian' concern with the degree of futurity or roundaboutness in investments. For example, in Cowen's analysis, an increase in the acceptable level of risk will encourage undertaking more longer-term investments (as well as, of course, investments of any given length with more uncertain yields). These can be both investments in durable capital goods (that is, investments with a continuous flow of payoffs over a long period of time) and investments with a long period of gestation before the ultimate output is produced. Cowen associates less risky ('safe') investments with consumption and shorter-term investments.

Cowen's analysis is more general than the traditional ABCT because it allows many factors besides a fall in real interest rates to generate a lengthening of the capital structure. These include exogenous risk-preference shifts, increases in savings, easing financial constraints, and reductions in uncertainty (so as to reduce 'waiting' for acceptable investment opportunities). Any of these changes can generate an increase in the riskiness of investment. None of these changes must necessarily cause a cyclical boom and bust, but they might do so.

中国社会科学院研究生院：研修之用

Horwitz (2000) shows that the traditional business cycle concerns of Austrian macroeconomics quite naturally lead into comparative institutional analysis. Therefore, the obvious question is: What kind of institutional framework is necessary or conducive to avoiding the distortionary effects of inflation and deflation? Austrians have been critical of both discretionary central banking policies and rigid monetarist rules. Some have favoured free banking while others have favoured a 100 per cent (usually gold) reserve requirement and hence have opposed fractional reserve banking.

The free banking school, represented by Selgin and White (1994), Horwitz (2000), Dowd (1996) and Sechrest (1993), emphasizes the importance of adjusting to changes in the demand to hold money (income velocity). For prices in particular markets to do their work appropriately in transmitting knowledge and allocating resources they must be free of the distortions induced by inflation and monetarily induced deflation. Free-banking advocates argue that bank profit maximization, under sound institutional constraints, will lead banks to expand or contract deposits or currency *pari passu* with changes in the demand for money. Banks will receive signals about the demand for (their) money as their reserves expand or contract. When reserves expand, the demand to hold is increasing, and vice versa. Profit maximization leads banks to increase the supply of money when reserves expand beyond their desired levels. Thus, no explicit monetary policy is needed to avoid unwarranted expansion or contraction on the 'money market', just as on commodity markets no deliberate industrial policy is needed to avoided unwarranted expansion or contraction of resources in different areas.

The advocates of 100 per cent reserve money follow the work of Murray N. Rothbard (2008). These include Block (1988), Hoppe (1994), and Huerta de Soto (1995). They argue that free banking - to the extent that it is fractional reserve banking - is ethically suspect. Regardless of the merits of this argument, our concern here is solely with economics. They further argue that fractional reserve banks are inherently inflationary because any creation of fiduciary media beyond an increase in specie will generate a business cycle. (The word 'inflationary' is being used here either as a definition - an increase in money not covered by an increase in specie - or as an intellectual place-card to suggest the generation of a cycle.) Critiques are offered in Horwitz (2000) and in Selgin and White (1994).

Back to top

Entrepreneurship

The theory of entrepreneurship has been a subject of great importance in Austrian economics since the publication of Israel Kirzner's *Competition*

中国社会科学院研究生院：研修之用

and Entrepreneurship (1973). One could argue, of course, that this was implicit in the prior works of Ludwig von Mises and Friedrich Hayek. Nevertheless, there is an important difference between implicit and explicit ideas. Over a long period of time, Kirzner refined his theory of entrepreneurial discovery or alertness in many books.

Kirzner's approach is predominantly cognitive. There are roots of this cognitive approach in the early work of von Mises (Ebeling, 2007). However, quite curiously in this time of the resurgence of psychology and economics, it is a cognitive theory without explicit cognitive foundations. Kirzner is interested in the market implications of the fact that there is entrepreneurial alertness. He is not interested, beyond some very general observations, in the causal factors that give rise to or are conducive to alertness.

Alertness, or equivalently, entrepreneurial discovery, is hard to define. It is a creative, spontaneous, and to a certain extent idiosyncratic, mental act that goes beyond the mere apprehension of objective data. First, while it usually begins with objective data, it critically involves drawing connections with other data when those connections are not obvious or even the result of complex computations. Second, true discoveries are not the result of deliberate acts of search. They cannot reliably be attained by the simple deployment of resources. Something more is necessary. This is not to suggest, however, that they must be viewed as random shocks to the economic system. They can be cultivated and prepared for by deliberate decisions, but they cannot be mechanically produced by them. We might say that while deliberate search is not a sufficient condition for discovery, it is necessary. Even better, eschewing the excessively constraining categories of necessity and sufficiency, we might say that the serendipity of discovery favours the searching mind (Holcombe, 2007; Shane, 2000). Finally, it is likely that many individuals can be exposed to the same data and yet not make the discoveries that the alert individual does.

In a market context, entrepreneurial cognition is the discovery of profit opportunities. In Kirzner's perspective, this is based on noticing price inconsistencies, whether at a point in time or across time. Hence this is an arbitrage theory of profit. How well this conception of entrepreneurship takes uncertainty into account is a matter of some dispute (Kirzner, 1982). In the theory advanced by Young Back Choi (1993, 1999), however, uncertainty is more explicitly considered. In this perspective, related to Schumpeter's classic analysis (1934), entrepreneurs break through the conventional way of looking at the world. These conventions were originally adopted to reduce uncertainty. But as time goes on the world changes and they become less and less effective. Profit opportunities accumulate. Entrepreneurs adopt new paradigms that

中国社会科学院研究生院：研修之用

enable them to see the new profit opportunities that conventionalists cannot.

To the extent that entrepreneurial discovery is unconnected to any cognitive or psychological basis, it functions as a *deus ex machina* of the market process. It drives the processes that occur in response to errors and disequilibria. Ultimately, it is defined by what it does. This approach has been criticized because it presupposes empirical psychological processes that are not necessarily present in all circumstances (Jakee and Spong, 2003). Can we say anything systematic about the factors that, on an individual or social level, are conducive to discovery? If we can, we might begin to understand more precisely what it is, when it is successful and when it is not.

In the first unified analysis of the factors affecting entrepreneurship, David Harper focuses on the presence of a sense of personal agency as the primary factor. 'It comprises two cognitive elements – beliefs in the locus of control (or contingency expectations) and beliefs in self-efficacy (or competence expectations)' (Harper, 2003, p. 14). This means that the entrepreneurial agent believes that in a particular context results are contingent upon actions as opposed to luck or nature, and that he himself possesses the personal capabilities to effect these actions and thus to produce the overall results. Individual characteristics also interact with situations to make the development of a discovery propensity more likely. Harper goes on to show the ways in which economic, political and cultural institutions mediate the individual factors.

In most Austrian treatments entrepreneurial discovery is important because it drives the market process. Nevertheless, in path-breaking work Frederic Sautet (2000) shows that there are multiple levels of entrepreneurship. In the simple case, the entrepreneur is herself alert to profit opportunities outside of the firm. In the more complex case, the entrepreneur must face the fact that she often doesn't know what her employees know. They are often closer to the local facts and may have a superior insight in some respects about profit opportunities in the firm (from restructuring) as well as outside of the firm. Thus, the entrepreneur in a 'complex firm' will seek to structure the firm with abstract or loose rules – some relating to compensation schemes – that encourage employees to make discoveries and communicate those appropriately. The firm itself can be a locus of entrepreneurship. In related work, Harper (2008b) suggests that a team of individuals, either inside or outside firms, might also constitute an entrepreneurial unit. Randall Holcombe (2007) utilizes an idea of entrepreneurship beyond pure cognitive alertness, which includes, as well, acting upon the perception of novel opportunities. In this view, the entrepreneur can never be certain that she has correctly perceived a profit opportunity until she acts and assesses the consequences.

中国社会科学院研究生院：研修之用

Some Austrians have not followed Kirzner in their analysis of entrepreneurship. For example, Joseph Salerno (1993) rejects the characterization of alertness as the essence of entrepreneurship. He sees resource ownership as a necessary feature of entrepreneurial activity (Salerno, 2008). Along similar lines, K. Foss, N. Foss and P. Klein (2007, see also P. Klein, 2008; K. Foss and N. Foss, 2007) have weaved together aspects of Knight's uncertainty theory (1971) and Austrian heterogeneous capital theory (Lachmann, 1956) to create a theory of entrepreneurial judgment. This theory makes entrepreneurship inseparable from asset ownership. The entrepreneur's judgement is about the control of heterogeneous capital assets under conditions of radical uncertainty. These authors have applied their theory to understanding the internal operation of the firm.

Back to top

Market processes and economic institutions

The entrepreneurial function is closely related to market processes and economic institutions. These interrelations are both complex and important. It will help to somewhat artificially separate them for our consideration.

A. The Austrian approach to market processes is distinctive in a number of respects (Wagner, 2007, 2010). It is sometimes described as a genetic-causal theory (Cowan and Rizzo, 1996). First, markets are in process and not continually in equilibrium. Thus, most Austrians do not take interpersonal equilibria of any kind simply as given or as consequences of an axiom of rationality. (An exception is Salerno, 1994, who considers momentary market-clearing equilibrium as an implication of rationality.) Lack of alertness can be responsible for economic errors and inconsistencies (or lack of interpersonal coordination). The market process consists of those entrepreneurial responses to error. Kirzner and others take the view that market processes are generally coordinating: that is, that they generally correct market errors. Austrians accept this as an *empirical generalization*. The extent to which the empirical generalization can be traced to an a priori discovery tendency is a subject of debate. Kirzner appears to accept this view because he sees the tendency to discover as equivalent to, or tightly connected to, the tendency toward greater coordination (Kirzner, 1997). Rizzo, however, rejects this equivalence (Rizzo, 1996). Other authors have also expressed similar, though not identical, criticisms (D. Klein and Briggerman, 2009; D. Klein, 1997). The neoclassical view that equilibrium is an implication of rationality should not, in this author's opinion, be replaced with the view that a tendency to equilibrium is the implication of purposefulness. The former has empirical implications while the latter is not clearly

中国社会科学院研究生院：研修之用

defined, unless it is meant as the positive heuristic of an empirical research program (Rizzo, 1982).

Second, market processes are not instantaneous but take time. In the passage of time ('real time'), knowledge changes and unpredictable events occur (O'Driscoll and Rizzo, 1996). What were data at the start of a process may change because the process of 'equilibration' occurs in real time. Real time cannot elapse without knowledge changing.

Third, market processes take place in the context of radical uncertainty. This is to be distinguished from risk, in which all of the possibilities are known with objective probabilities. However, radical uncertainty is not simply a condition where the assigned probabilities are not objective, but one in which not all of the possibilities are known beforehand. (Still further complications ensue because sometimes individuals know that they don't know the possibilities and sometimes they do not.)

This leads to the fourth feature of market processes: they are relatively indeterminate. If market processes - in the form of entrepreneurial discovery - cannot be predicted, then the economist cannot know at the beginning where they will lead. In the process of adjusting to change, new 'data' will be discovered (Rizzo, 2000, 1990). How far to take this point about the indeterminacy of market processes is subject to debate and may, in part, depend on definitional issues (Holcombe, 2007). Some have argued that Kirzner, in particular, has incorrectly downplayed this indeterminacy (Jakee and Spong, 2003).

This is not to rule out the use of constructs in which equilibria are reached as heuristic devices when appropriate (Holcombe, 2007). However, since they are simply heuristic devices they can be thrown out when circumstances do not warrant such 'static' dynamics.

The fifth, and final, feature of market processes is the communication of decentralized or scattered knowledge. Markets enable individuals to act on more knowledge than they can ever hope to possess explicitly. They can do this through entrepreneurially produced market prices and through non-price manifestations of market behaviour. As Hayek showed, the man on the spot may be directly aware of certain economically relevant conditions. If he acts by taking advantage of this knowledge in profitably buying or selling he will ensure that market prices communicate what he knows (Hayek, 1948; Kirzner, 1992a).

Prices are not the only communicators of knowledge in markets. Capital goods also embody knowledge. First, the particular use and combination of capital goods can, under non-distortionary conditions, convey knowledge about efficient resource allocation and possible profit opportunities (Lachmann, 1956). Second, even the physical design of capital goods can convey accumulated knowledge about successful production techniques (Baetjer, 2000).

中国社会科学院研究生院：研修之用

In general, the communication of knowledge in market settings depends not only on catallactic phenomena but also crucially on the appropriate 'institutional' context. This includes legal and cultural products (Harper, 2003). In the latter category David Harper (2008a, forthcoming) has drawn attention to the role of numerical cognition - a product of both unique human biology and cultural development - in facilitating economic calculation. The development of conventionalized systems of number sequences and techniques of counting reduces transaction costs, and helps agents to make plans, compute values, scarcities, notice arbitrage opportunities, and ascertain the economically relevant aspects of capital goods.

B. Entrepreneurship does not simply operate within a familiar institutional structure like the market. It can also operate within structures like those involving social ties, philanthropy, non-profit organizations and so forth (Boettke and Coyne, 2009). There is also 'political entrepreneurship' within a given constitutional or governance structure, which seeks to create coalitions to effect specific legislation or transfers of wealth ('rent seeking'). These non-market structures determine the precise form that entrepreneurship takes. The common differentiating factor that separates the entrepreneurship of the market process from these other forms of entrepreneurship is the absence of the discipline of monetary profit and loss in the latter cases. Although money may change hands as a result of these forms of entrepreneurial activity, their outputs are not valued according to market prices. Whether effective feedback mechanisms exist in these contexts is an open question (Boettke and Coyne, 2009).

Some Austrians, however, have emphasized that non-market institutions can indeed provide feedback to entrepreneurs and can generate a social learning or knowledge-communication process similar to market prices and profit-loss signals (Chamlee-Wright, 2008; Chamlee-Wright and Myers, 2008; Lewis and Chamlee-Wright, 2008). In particular, reputation and status are forms of 'social capital' that convey information. Under conditions of competition and effective monitoring of standards, knowledge can be transmitted far beyond networks of individuals in direct communication with each other.

An important example of the communication of knowledge in a non-market context can be found in the scientific community (McQuade and Butos, 2003). We discuss this below in the section on spontaneous orders.

Entrepreneurship can also shape or create institutions. Rules of behaviour that surround and define markets, constitutional systems, social and cultural systems arise out of the previous framework of rules, whether it was de facto or de jure. (In fact, the distinction between de facto and de jure may not be all that important for the economics of institutions, aside from the possible issue of transaction costs.) There

中国社会科学院研究生院：研修之用

is path dependency in the development of institutions (Boettke, Coyne and Leeson, 2008). Those that develop as 'indigenously introduced endogenous institutions' are closely related to the informal practices and expectations of people, which in turn are grounded in local knowledge and values. Other institutions may be indigenously introduced but are exogenous in the sense that they are imposed by some formal authority, and do not gradually evolve from the informal traditions of a people. There is a risk that these institutions will not 'stick' because of conflict between the institution and the underlying norms. Externally (or foreign) introduced exogenous institutions exhibit the greatest probability of not succeeding because of the greater likelihood of conflict with underlying norms and expectations. Boettke, Coyne and Leeson (2008) refer to this analysis as an example of the 'regression theorem' first propounded by Ludwig von Mises (1953) in his analysis of the evolution of money.

Some of the evolution of framework institutions may simply be the undesigned outcome of individual behaviour that is not necessarily entrepreneurial, as when people follow each other in making a path through the snow (Kirzner, 1992b). In other cases, there may be alertness to possibilities of gain for the relevant acting parties in altering the political or social frameworks. Plausibly, the creation of the US Constitution was one such case.

Institutions exist at many levels. Perhaps the most basic are those that involve informal institutions like customs, traditions, norms and religion (Williamson, 2000). These take the longest time to change. They may also determine the standards by which lower-level institutions and behaviour within them are evaluated. A new political system is good or bad depending on the (more basic) norm structure in place.

Back to top

Spontaneous orders

Our discussion of entrepreneurship and of institutions leads naturally into a discussion of spontaneous order, an idea very closely associated with Austrian economics. Unfortunately, the term 'spontaneous order' is opaque. Somewhat more descriptive is the expression made famous by F.A. Hayek, 'the results of human action but not of human design' (Hayek, 1967), and even more descriptive is the idea of unintended social order produced by individually purposeful behaviour.

A spontaneous order is an organic or emergent form of coordination that manifests itself in social institutions, some organizations and clusters of individual plans. Orders of this kind arise without the design and maintenance (oversight) of a social planner. Nevertheless, spontaneous orders are generated by individual agents who do plan and carry out actions

中国社会科学院研究生院：研修之用

within their sphere of activity. Social order emerges as individuals adjust their plans to each other and to the environment over time.

Spontaneous order theories come in different varieties. Some refer to order produced on markets, while others concern order produced in non-market settings. These theories can be purely positive (descriptive) or they can also be normative. When they are normative their normativity can be relative to the society as a whole or simply to particular subgroups.

At the most basic level, spontaneous order can refer simply to the welfare-enhancing outcomes of competitive market processes operating within the 'fixed' constraints of property, contract and tort law. This is best studied within the context of market entrepreneurship.

Bruce Benson (1989), in his path-breaking study of the spontaneous evolution of commercial law, shows how market interactions, based on basic property constraints, can give rise to commercial (contract) law without a law-giver. The self-interested interactions of merchants lead them to develop and adhere to rules that increase their trade and hence overall social cooperation. These rules develop through a process of trial and error in which entrepreneurial alertness at a higher level - the level of rules of the game - doubtless plays an important role. Similarly, Stringham (2002, 2003) and Stringham and Boettke (2004) show that the self-interested interaction of participants in financial markets has generated useful regulations that govern the operation of these markets.

Peter Leeson (2007, 2009), in a number of studies of the organization of 18th-century pirate activity, shows how an outlaw subgroup of society developed maximizing (or 'rational' in a limited sense) rules of governance without central direction. Outside of a market context, pirates converged on a set of rules whereby their ability to steal wealth from the rest of society was enhanced. This involved rules within the pirate society itself as well as rules governing its treatment of others. Within their society 'democracy' was used; outside of it the use of brutality was constrained. This case is a good example of a spontaneous ordering process with 'good' consequences within the subgroup and yet negative consequences for society as a whole. Pirates steal resources from the rest of society. The success of any such rogue subgroup weakens the possibilities of voluntary exchange and other forms of peaceful interaction.

Thomas McQuade and William Butos (2003; see also Butos and Koppl, 2003; Butos and McQuade, 2006) further develop the spontaneous-order approach in the case of the organization of scientific research communities. Even where markets in the traditional sense may be missing, spontaneous - that is, non centrally directed - ordering processes are still present. They focus on the evolved non-market mechanism of publication - citation - reputation. Scientific knowledge is viewed as a

中国社会科学院研究生院：研修之用

'by-product' of the intentional activities of scientists to publish their results, get citations and enhance their reputations. Within this process competition among scientists tends to filter out inferior ideas. The resultant product ('science') is orderly in the sense that it tends to be reliable and codifiable. A set of procedures is put into place which acts as a filter to discriminate between rival claims. Furthermore, what comes out of the filter can be collected, integrated with other knowledge and transferred to other scientists.

These illustrations suggest the need for a more general theory of spontaneous order that would clarify the various conditions under which such ordering-processes will take place. Specifically, it should also explore the role of markets and market prices, since it is clear that spontaneous order can develop without markets. From the welfare point of view the research discussed above leaves us with a puzzle: When do spontaneous orders produce an enhancement of social welfare and when a reduction in it, as in the case of pirate societies?

Back to top

Law and economics

One of the most important areas of research in Austrian economics is the vibrant area of law and economics. Some of the contributions mentioned above in connection with spontaneous order and institutions could be included in this section. The field's uniquely Austrian features consist of attention to (1) the process of law and state intervention in markets; (2) the need for relatively stable law in a world of external change; (3) the influence of decentralized knowledge on the character and limits of law; and (4) the privatization of some of the basic functions of the state.

1. The most significant work on the processes generated by intervention since the classic analyses of Ludwig von Mises (1977), F.A. Hayek (1994) and Israel Kirzner (1985) can be found in Sanford Ikeda (1997, 2003, 2005) and in Mario Rizzo and Glen Whitman (2003). Ikeda's framework focuses on the deviation of the actual outcomes of intervention from the intended outcomes. This gap, based on an assumption of radical ignorance, generates price distortions, whether because the intervention takes the form of price regulations or because redistribution of wealth degrades incentives and thus individual responses to underlying economic data. These economic changes interact with largely, though not entirely, endogeneous changes in ideology to produce a tendency toward further policy intervention. Rizzo and Whitman, on the other hand, begin from the largely philosophical and jurisprudential literature of 'slippery slopes'. They construct a general approach that emphasizes the role of changes in ideas, or more precisely, in the arguments that rationalize or justify legislative policies or judicial decisions. The mechanism by which these arguments

中国社会科学院研究生院：研修之用

change is a combination of the largely unanticipated consequences of decisions and the higher-level theories in which acceptable arguments are embedded.

Recently, Rizzo and Whitman (2009; see also Whitman and Rizzo, 2007) have applied their slippery slope analysis in conjunction with many of the assumptions and findings of behavioral economics to demonstrate the expansive tendencies inherent in the supposedly moderate policies of new or 'libertarian' paternalism.

The Rizzo-Whitman and Ikeda approaches seem largely compatible. Ikeda stresses more traditional economic processes, while Rizzo and Whitman stress the details of the intellectual changes that occur in the context of economic or other processes. In neither of these approaches is the 'slippery slope' consequence of policies inevitable. They each describe tendencies that could be counterbalanced in specific cases, but which often have not been.

2. The classic work of Hayek (1960, 1973) on the rule of law simultaneously stresses the importance of stability in the legal framework and its adaptability to changing external circumstances. The solution to this paradox can be found in the level of abstraction of the relevant rules. For example, the abstract form of contract law can remain stable while the prices, conditions and content of exchanges vary at a point in time or over time. The consequences of abstraction in legal rules are examined in Whitman (2009). Whitman shows that an intermediate level of abstraction is optimal from the perspective of generating rules with predictable consequences.

From a slightly different perspective, Rizzo (1980a, 1980b, 1985) and Roy Cordato (2007) both criticize the cost-benefit framework in many conceptions of negligence law because it produces legal decisions that lack predictability to those for whom the particular law is relevant. Peter Lewin (1982) extends the critique to pollution externalities and social cost. The economic data upon which efficient legal decisions are to be made are often unavailable, complex or transient. This is especially true in a world characterized by radical uncertainty. Thus the so-called economic approach to tort law is defective on its own terms. Lack of predictability generates costs. In terms of the abstraction language of Whitman's analysis, the problem of the efficiency approach is that it enshrines a standard, rather than a set of specific rules, which is too abstract.

Similar criticisms of the so-called economic approach to property rights that derives from Ronald Coase and Harold Demsetz have been advanced by Walter Block. Block argues that the Coasian cost-benefit approach effectively abolishes property rights (1977, 1995, 2000). This view is extended to the analysis of the recent US Supreme Court eminent domain case, Kelo *v.* City of New London (Block, 2006).

中国社会科学院研究生院：研修之用

3. The decentralization of factual knowledge is a critically important factor limiting the feasibility of many forms of intervention. As in the earlier analysis of Mises (1977), the critiques discussed here begin from the announced goals of the interveners and do not challenge their worthiness. The approach is thus non-normative. It simply seeks to answer the question: Can the policies achieve the goals that their advocates have set? Rizzo (2005) tackles this question in the case of moral paternalism: that is, the form of paternalism that coerces the individual in the interests of her moral betterment. Using the internal standards of three major ethical approaches – utilitarianism, natural law and Kantianism – Rizzo argues that the factual knowledge needed to determine just what the moral course of action is in concrete cases is not available to the paternalist. Rizzo and Whitman (forthcoming) also apply this kind of analysis to a form of economic paternalism based on behavioural economics. They argue that the factual knowledge that behavioural economics claims is relevant to the crafting of policies designed to improve the decisions of individuals exceeds what is known to the policy makers.

4. Most economic analysis proceeds on the assumption that the state exercises at least its minimum functions: that is, provision of protection, enforcement of property rights and contracts, and the adjudication of disputes. Nevertheless some economists in the broad Austrian and spontaneous order tradition have argued that privatization of at least some of these functions is feasible and desirable. Bryan Caplan and Edward Stringham (2008) have compared the private and public adjudication of disputes. They find that private adjudication is more efficient in areas of commercial disputes, and more generally in those areas where prior relationships exist among the parties. They also speculate on a broader use of private adjudication. In a related area of public choice economics, Powell and Stringham (2009) survey a surprisingly large extant literature on the economics of a stateless society.

Back to top

See Also

- Austrian economics
- business cycle
- entrepreneurship
- spontaneous order

Back to top

Bibliography

中国社会科学院研究生院：研修之用

Baetjer, H. 2000. Capital as embodied knowledge: some implications for the theory of economic growth. *Review of Austrian Economics* 13, 147–74.

Benson, B. 1989. The spontaneous evolution of commercial law. *Southern Economic Journal* 55, 644–61.

Block, W. 1977. Coase and Demsetz on private property rights. *Journal of Libertarian Studies* 1, 111–15.

Block, W. 1988. Fractional reserve banking. In *Man, Economy and Liberty: Essays in Honor of Murray H. Rothbard*, ed. W. Block and L. Rockwell. Auburn Ala: Mises Institute, pp. 24–31.

Block, W. 1995. Ethics, efficiency, Coasian property rights, and psychic income: a reply to Demsetz. *Review of Austrian Economics* 8, 61–125.

Block, W. 2000. Private-property rights, erroneous interpretations, morality, and economics: reply to Demsetz. *Quarterly Journal of Austrian Economics* 3, 63–78.

Block, W. 2006. Coase and Kelo: ominous parallels and reply to Lott on Rothbard on Coase. *Whittier Law Review* 27, 997–1022.

Boettke, P. and Coyne, C. 2002. Entrepreneurship and development: cause or consequence? *Advances in Austrian Economics* 6, 67–88.

Boettke, P. and Coyne, C. 2009. Context matters: institutions and entrepreneurship. *Foundations and Trends in Entrepreneurship* 5, 135–209.

Boettke, P.J., Coyne, C.J. and Leeson, P.T. 2008. Institutional stickiness and the new development economics. *American Journal of Economics and Sociology* 67, 331–58.

Butos, W.N. and Koppl, R. 2003. Science as a spontaneous order. In *The Evolution of Scientific Knowledge*, ed. H.S. Jensen, L.M. Richter and M.T. Vendelø. Northampton, MA: Edward Elgar, pp. 164–88.

Butos, W.N. and McQuade, T.J. 2006. Government and science: a dangerous liaison? *Independent Review* 11, 177–208.

Caplan, B. and Stringham, E.P. 2008. Privatizing the adjudication of disputes. *Theoretical Inquiries in Law* 9, Article 8. Available at: http://www.bepress.com/til/default/vol9/iss2/art8.

中国社会科学院研究生院：研修之用

Chamlee-Wright, E. 2008. The structure of social capital: an Austrian perspective on its nature and development. *Review of Political Economy* 20, 41－58.

Chamlee-Wright, E. and Myers, J.A. 2008. Discovery and social learning in non-priced environments: an Austrian view of social network theory. *Review of Austrian Economics* 21, 151－66.

Choi, Y.B. 1993. *Paradigms and Conventions: Uncertainty, Decision Making, and Entrepreneurship*. Ann Arbor, MI: University of Michigan Press.

Choi, Y.B. 1999. Conventions and learning: a perspective on the market process. In *Economic Organisation and Economic Knowledge*, ed. S.C. Dow and P. Earl. London: Edward Elgar, pp. 57－75.

Cordato, R. 2007. *Efficiency and Externalities in an Open-Ended Universe*. Auburn, Ala.: Ludwig von Mises Institute, 1992.

Cowan, R. and Rizzo, M.J. 1996. The genetic-causal tradition and modern economic theory. *Kyklos* 49, 273－317.

Cowen, T. 1998. *Risk and Business Cycles: New and Old Austrian Perspectives*. London: Routledge.

Dowd, K. 1996. *Laissez-Faire Banking*. London: Routledge.

Dowd, K. 2000. *Money and the Market: Essays on Free Banking*. London and New York: Routledge.

Ebeling, R.M. 2007. Austrian economics and the political economy of freedom. *New Perspectives on Political Economy* 3, 87－104. Available at: http://pcpe.libinst.cz/nppe/3_1/nppe3_1.pdf.

Endres, A.M. and Harper, D.A. 2008. *Capital as a Layer Cake: Menger, Lachmann and the Nature of Capital*. Unpublished manuscript, New York University.

Foss, K. and Foss, N.J. 2006. The limits to designed orders: authority under 'distributed knowledge' conditions. *Review of Austrian Economics* 19, 261－274.

Foss, K. and Foss, N.J. 2007. The entrepreneurial organization of heterogeneous capital. *Journal of Management Studies* 44, 1165－86.

中国社会科学院研究生院：研修之用

Foss, K. , Foss, N.J. and Klein, P.G. 2007. Original and derived judgment: an entrepreneurial theory of economic organization. *Organization Studies* 28, 1–20.

Garrison, R. 2001. *Time and Money: The Macroeconomics of Capital Structure* . London: Routledge.

Harper, D.A. 1998. Institutional conditions for entrepreneurship. *Advances in Austrian Economics* 5, 241–75.

Harper, D.A. 2003. *Foundations of Entrepreneurship and Economic Development* . London: Routledge.

Harper, D.A. 2008a. A bioeconomic study of numeracy and economic calculation. *Journal of Bioeconomics* 10, 101–26.

Harper, D.A. 2008b. Towards a theory of entrepreneurial teams. *Journal of Business Venturing* 23, 613–26.

Harper, D.A. Forthcoming. Numbers as a cognitive and social technology: an economic ontology. *Journal of Institutional Economics*.

Hayek, F.A. 1948. The use of knowledge in society. In *Individualism and Economic Order*. Chicago: University of Chicago Press.

Hayek, F.A. 1960. *The Constitution of Liberty* . Chicago: University of Chicago Press.

Hayek, F.A. 1967. The results of human action but not of human design. In *Studies in Philosophy, Politics and Economics*. London: Routledge & Kegan Paul.

Hayek, F.A. 1973. *Law, Legislation and Liberty, Vol. 1: Rules and Order* . Chicago: University of Chicago Press.

Hayek, F.A. 1944. *The Road to Serfdom* . Chicago: University of Chicago Press, 1994.

Holcombe, R.G. 2007. *Entrepreneurship and Economic Progress* . London: Routledge.

Hoppe, H.-H. 1994. 'How is fiat money possible? or, the devolution of money and credit.' *Review of Austrian Economics* 7, 49–74.

Horwitz, S. 2000. *Microfoundations and Macroeconomics: An Austrian Perspective* . London: Routledge.

中国社会科学院研究生院：研修之用

Huerta de Soto, J. 1995. A critical analysis of central banks and fractional-reserve free banking from the Austrian perspective. *Review of Austrian Economics* 8, 25－38.

Huerta de Soto, J. 2006. *Money, Bank Credit, and Economic Cycles*, trans. M.A. Stroup, Auburn, Ala.: Ludwig von Mises Institute.

Ikeda, S. 1997. *Dynamics of the Mixed Economy: Toward a Theory of Interventionism*. London: Routledge.

Ikeda, S. 2003. How compatible are public choice and Austrian political economy?. *Review of Austrian Economics* 16, 63－75.

Ikeda, S. 2005. The dynamics of intervention: regulation and redistribution in the mixed economy. *Advances in Austrian Economics* 8, 21－57.

Jakee, K. and Spong, H. 2003. Praxeology, entrepreneurship and the market process: a review of Kirzner's contribution. *Journal of the History of Economic Thought* 25, 461－86.

Kirzner, I.M. 1973. *Competition and Entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, I.M. 1979. *Perception, Opportunity, and Profit: Studies in the Theory of Entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, I.M. 1982. Uncertainty, discovery, and human action: a study of the entrepreneurial profile in the Misesian system. In *Method, Process, and Austrian Economics*. Lexington, Ky.: Lexington Books.

Kirzner, I.M. 1985. The perils of regulation: a market process approach. In *Discovery and the Capitalist Process*. Chicago: University of Chicago Press.

Kirzner, I.M. 1992a. Prices, the communication of knowledge and the discovery process. In *The Meaning of Market Process*. London: Routledge.

Kirzner, I.M. 1992b. Knowledge problems and their solutions: some relevant distinctions. In *The Meaning of Market Process*. London: Routledge.

Kirzner, I.M. 1997. Entrepreneurial discovery and the competitive market process: an Austrian approach. *Journal of Economic Literature* 35, 60－85.

中国社会科学院研究生院：研修之用

Klein, D.B. 1997. Convention, social order, and the two coordinations. *Constitutional Political Economy* 8, 319–35.

Klein, D.B. and Briggerman, J. 2009. Israel Kirzner on coordination and discovery. *Journal of Private Enterprise* 25, 1–53.

Klein, P.G. 2008. Opportunity discovery, entrepreneurial action, and economic organization. *Strategic Entrepreneurship Journal* 2, 175–90.

Knight, F.H. 1921. *Risk, Uncertainty, and Profit*. Chicago: University of Chicago Press, 1971.

Koppl, R. 2002. *Big Players and the Economic Theory of Expectations*. Hampshire and New York: Palgrave Macmillan.

Koppl, R. and Mramor, D. 2003. Big players in Slovenia. *Review of Austrian Economics* 16, 253–69.

Koppl, R. and Yeager, L.B. 1996. Big players and herding in asset markets: the case of the Russian ruble. *Explorations in Economic History* 33, 367–83.

Lachmann, L.M. 1956. *Capital and its Structure*. London: G. Bell & Sons.

Leeson, P. 2007. An-arrgh-chy: the law and economics of pirate organization. *Journal of Political Economy* 115, 1049–94.

Leeson, P.T. 2009. *The Invisible Hook: The Hidden Economics of Pirates*. Princeton, N.J.: Princeton University Press.

Lewin, P. 1982. Pollution externalities: social cost and strict liability. *Cato Journal* 2, 205–29.

Lewin, P. 1999. *Capital in Disequilibrium: The Role of Capital in a Changing World*. London: Routledge.

Lewis, P. and Chamlee-Wright, E. 2008. Social embeddedness, social capital and the market process: an introduction to the Special Issue on Austrian Economics, Economic Sociology and Social Capital. *Review of Austrian Economic* s 21, 107–18.

McQuade, T.J. and Butos, W.N. 2003. Order-dependent knowledge and the economics of science. *Review of Austrian Economics* 16, 133–52.

Mises, L. von. 1912. *The Theory of Money and Credit*. New Haven, Conn.: Yale University Press, 1953.

中国社会科学院研究生院：研修之用

Mises, L. von. 1929. *A Critique of Interventionism*, trans. H.F. Sennhollz, New Rochelle, N.Y.: Arlington House, 1977.

O'Driscoll, G.P. Jr. and Rizzo, M.J. 1996. *The Economics of Time and Ignorance*. London: Routledge.

Powell, B.W. and Stringham, E.P. 2009. Public choice and the economic analysis of anarchy: a survey. *Public Choice* 140, 503–538.

Rizzo, M.J. 1980a. Law amid flux: the economics of negligence and strict liability in tort. *Journal of Legal Studies* 9, 291–318.

Rizzo, M.J. 1980b. The mirage of efficiency. *Hofstra Law Review* 8, 641–58.

Rizzo, M.J. 1982. Mises and Lakatos: a reformulation of Austrian methodology. In *Method, Process, and Austrian Economics*, ed. I. Kirzner. Lexington, KY: Lexington Books.

Rizzo, M.J. 1985. Rules versus cost–benefit analysis in the common law. *Cato Journal* 4, 865–84.

Rizzo, M.J. 1990. Hayek's four tendencies towards equilibrium. *Cultural Dynamics* 3, 12–31.

Rizzo, M.J. 1996. Introduction: time and ignorance after ten years. In *The Economics of Time and Ignorance*, ed. G.P. O'Driscoll Jr and M.J. Rizzo. London: Routledge.

Rizzo, M.J. 2000. Real time and relative indeterminacy in economic theory. In *Time in Contemporary Intellectual Thought*, ed. P. Baert. Amsterdam: Elsevier, pp. 171–88.

Rizzo, M.J. 2005. The problem of moral dirigisme: a new argument against moralistic legislation. *NYU Journal of Law and Liberty* 1, 789–843.

Rizzo, M.J. and Whitman, D.G. 2003. The camel's nose is in the tent: rules, theories, and slippery slopes. *UCLA Law Review* 51, 539–92.

Rizzo, M.J. and Whitman, D.G. 2009. Little brother is watching you: new paternalism on the slippery slopes. *Arizona Law Review* 51, 685–739.

Rizzo, M.J. and Whitman, D.G. 2009. The knowledge problem of new paternalism. *Brigham Young University Law Review* 2009, 905–968

中国社会科学院研究生院：研修之用

Rothbard, M.N. 1983. *The Mystery of Banking*. Auburn, Ala.: Ludwig von Mises Institute, 2008.

Salerno, J.T. 1993. Mises and Hayek dehomogenized. *Review of Austrian Economics* 6, 113–46.

Salerno, J.T. 2003. An Austrian taxonomy of deflation – with applications to the U.S. *Quarterly Journal of Austrian Economics* 6, 81–109.

Salerno, J.T. 2008. The entrepreneur: real and imagined. *Quarterly Journal of Austrian Economics* 11, 188–207.

Sautet, F.E. 2000. *An Entrepreneurial Theory of the Firm*. London: Routledge.

Schumpeter, J.A. 1911. *Theory of Economic Development*. Cambridge, Mass.: Harvard University Press, 1934.

Sechrest, L. 1993. *Free Banking: Theory, History, and a Laissez-Faire Model*. Westport, Conn.: Quorum.

Selgin, G.A. and White, L.H. 1994. How would the invisible hand handle money? *Journal of Economic Literature* 32, 1718–49.

Shane, S. 2000. Prior knowledge and the discovery of entrepreneurial opportunities. *Organization Science* 11, 448–69.

Stringham, E. 2002. The emergence of the London Stock Exchange as a self-policing club. *Journal of Private Enterprise* 17, 1–19.

Stringham, E. 2003. The extralegal development of securities trading in seventeenth-century Amsterdam. *Quarterly Review of Economics and Finance* 43, 321–44.

Stringham, E. and Boettke, P. 2004. Brokers, bureaucrats and the emergence of financial markets. *Managerial Finance* 30, 57–71.

Wagner, R. 2005. Austrian Cycle Theory and the prospect of a coordinationist macroeconomics. In *Modern Applications of Austrian Thought*, ed. J. G. Backhaus. London and New York: Routledge, pp. 77–92.

Wagner, R. 2007. Value and exchange: two windows for economic theorizing. *Review of Austrian Economics* 20, 57–68.

Wagner, R. 2010. *Mind, Society, and Human Action: Time and Knowledge in Social Economy*. London and New York: Routledge.

中国社会科学院研究生院：研修之用

Whitman, D.G. and Rizzo, M.J. 2007. Paternalist slopes. *NYU Journal of Law and Liberty* 2, 411–43.

Whitman, D.G. 2009. The rules of abstraction. *Review of Austrian Economics* 22, 21–42.

Williamson, O.E. 2000. The new institutional economics: taking stock, looking ahead. *Journal of Economic Literature* 38, 595–613.

Back to top

How to cite this article

Rizzo, Mario J. "Austrian economics: recent work." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 13 January 2010 <http://www.dictionaryofeconomics.com/article?id=pde2009_A000258>
doi:10.1057/9780230226203.1915

中国社会科学院研究生院：研修之用

Chandler, Alfred D. (1918–2007)

Walter A. Friedman
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume
Back to top

Keywords

big business; Business history; Chandler, Alfred D.; Chandlerian; economies of scale; invisible hand; management; visible hand
Back to top

Article

Alfred D. Chandler Jr. (15 September 1918–9 May 2007), a Pulitzer Prize-winning historian who pioneered the field of business history, was born in Guyencourt, Delaware, near Wilmington. He received a bachelor of arts degree from Harvard College in 1940 and served in the US Navy from 1941 to 1945. In the late 1940s, Chandler returned to school to study history, attending the University of North Carolina before going back to Harvard to complete his Ph.D. in 1952. He published a revision of his dissertation as his first book, *Henry Varnum Poor: Business Editor, Analyst, and Reformer* (1956). Poor (1812–1905), Chandler's paternal great-grandfather, was the long-time editor of the *American Railroad Journal* and *Poor's Manual of Railroads of the United States*. Chandler's book explained how Poor, through his detailed reports on individual railroad companies and their operations, helped to invent the role of the modern business analyst and investment advisor.

From 1950 to 1963, Chandler taught at the Massachusetts Institute of Technology, and then left to join the history department at Johns Hopkins, where he remained until 1970. While at Hopkins, he edited the papers of President Dwight D. Eisenhower. From 1970 to 1989, Chandler was a professor at Harvard Business School, where he held the Isidor Straus Chair in Business History. While there, Chandler inaugurated the course entitled 'The Coming of Managerial Capitalism: The United States'. He encouraged other historians to come to the school, including his successors in the Straus Chair, Thomas K. McCraw and Geoffrey Jones, and he sponsored research fellowships for graduate students and international scholars to travel to the business school's Baker Library. From the 1970s onward, he lived in a building within walking distance of the campus, in

中国社会科学院研究生院：研修之用

a large 17th-floor apartment overlooking the Charles River and filled with artwork, including paintings by his wife, the former Fay Martin.

Chandler was the sole author of six books and co-author or editor of more than 30 others. His most famous works focused on the rise of big business and the coming of a managerial class: *Strategy and Structure: Chapters in the History of Industrial Enterprise* (1962); *The Visible Hand: The Managerial Revolution in American Business* (1977); and *Scale and Scope: The Dynamics of Industrial Capitalism* (1990). As many commentators acknowledged, these books were so original in their approach and so impressive in their depth of research that they set the agenda for the entire field of business history for many years afterward.

The first of these, *Strategy and Structure*, analysed DuPont, General Motors, Sears and Standard Oil, and showed how each of these four companies came to adopt a multidivisional structure, or M-form, by the 1920s. No previous historian had provided such a rich account of how big businesses actually worked, or described how middle managers confronted the complexities of daily business life, filled as it was with committee meetings, budget decisions and forecasts. 'Only by showing these executives as they handled what appeared to them to be unique problems and issues can the process of innovation and change be meaningfully presented,' Chandler wrote in 1962. His detailed investigations were the basis for his influential argument that a company's strategy must shape its structure, not the other way around, as was often the case.

In *The Visible Hand*, Chandler sought to explain the rise of big business in the United States in the decades from 1840 to 1920, and to answer the question why large firms arose in some industries and not in others. Chandler argued that in industries whose firms were able to benefit from economies of scale and scope, the 'visible hand' of management came to replace the 'invisible hand' of the market in coordinating the production and distribution of goods. This was to become his most famous book, winning not only the Pulitzer, but also the Bancroft Prize and the Thomas Newcomen Book Award.

In *Scale and Scope*, Chandler branched out into comparative international history, comparing his story of the ascendancy of capitalism in the United States, from the late 19th to the mid-20th century, with the histories of Britain and Germany. Success in steel, chemicals, automobiles and other industries that emerged during the second industrial revolution, Chandler argued, was achieved through making a three-pronged investment: in mass-production facilities, in international marketing and distribution networks, and in proper management of resource allocation. While Chandler praised German industry, which had developed strong capacities in research engineering, banking, and the production of producer goods, he believed that Britain's tradition of 'personal capitalism' had

中国社会科学院研究生院：研修之用

prevented that country from making progress in developing large-scale industries.

These three books attracted their share of admirers throughout the world, and 'Chandlerian' business history quickly became the focus of conferences and academic papers. In 1973, Derek F. Channon published *Strategy and Structure of British Enterprise*; this was followed three years later by Gareth P. Dyas and Heinz T. Thanheiser's *The Emerging European Enterprise: Strategy and Structure in French and German Industry*. The First Fuji conference, held in Japan in January 1974, was devoted to the 'Strategy and Structure of Big Business'. Chandler's work was also central to curricula at business history units formed at the London School of Economics (in the late 1970s), and in the decades afterward at such places as the universities of Glasgow, Leeds and Reading in the United Kingdom, Bocconi University in Italy, and the Copenhagen Business School in Denmark.

Chandler also received his share of criticism, in part because of his narrow focus on the rise of big business and his relative neglect of the roles of politics, finance, and culture in explaining the growth of the American economy. Some, including Philip Scranton (1997) and Charles F. Sabel and Michael Piore (Pire and Sabel, 1984), argued that Chandler downplayed the contribution of small and medium-sized firms and overlooked the ways in which the supplanting of independent artisans and flexible manufacturers by middle managers created problems for the American economy. Chandler's most controversial book was *Scale and Scope*. British writers, in particular, bristled at Chandler's view that the preponderance of family-owned firms in the United Kingdom had contributed to that country's relative decline. Barry Supple, writing in the *Economic History Review* (1991), argued that Chandler's assumption that the American model should be the 'standard against which to assess the structural characteristics and achievements of the business systems of other countries has some pitfalls' (p. 512).

But Chandler was not an apologist for American industry, nor was he wholly enamoured with business success. He objected to many trends that were taking place in American management practice in the 1960s, including the conglomerate movement. Late in his career, he wrote admiringly of the triumph of Japanese industry over US competitors in the electronics industries in the final third of the 20th century. In the 1990s, Chandler became fascinated with the question of why some industries failed and others rose in their place. He completed his final two books, both touching on these themes, while in his 80s: *Inventing the Electronic Century: The Epic Story of the Consumer Electronics and Computer Industries* (2001); and *Shaping the Industrial Century: The Remarkable Story of the Modern Chemical and Pharmaceutical Industries* (2005).

中国社会科学院研究生院：研修之用

While most historians have focused on these core books, Chandler's other works should not be neglected. He co-wrote (with Stephen Salsbury) *Pierre S. du Pont and the Making of the Modern Corporation* (1971), a long and rich primary-source study that recounts Pierre's role in making DuPont the largest US chemical and explosives company and General Motors the world's biggest car manufacturer. Chandler also edited, or co-edited, many volumes, including, with Franco Amatori and Takashi Hikino, *Big Business and the Wealth of Nations* (1997); and with James W. Cortada, *A Nation Transformed by Information: How Information has Shaped the United States from Colonial Times to the Present* (2000). He published 60 articles, many of which are listed in the bibliography of Thomas K. McCraw's edited collection, *The Essential Alfred Chandler* (1988). One extremely insightful article, published in 1994 and hence not mentioned in McCraw's volume, is his 72-page international comparative study, published in *Business History Review*, 'The competitive performance of U.S. industrial enterprises since the Second World War.' Chandler was also the general editor of the scholarly monograph series Harvard Studies in Business History, published by Harvard University Press.

Throughout his career, Chandler's work attracted attention because he continued to ask and answer broad and challenging questions. In the 1950s and 1960s, he analysed the workings of firms, while most economists and historians at the time found them uninteresting. In the 1970s and 1980s, he turned his attention to international business and to comparative analysis. The influence of Chandler's work extended far beyond the discipline of history. He made vital contributions to organizational sociology, global business studies, and to the field of strategic management. Among his many honours he had the distinction of being listed as an eminent scholar by the Academy of International Business.

Chandler's significance to business history has been summarized in McCraw (1988) and in Richard John's 1997 essay "Elaborations, revisions, dissents: Alfred D. Chandler, Jr.'s, The Visible Hand after twenty years." In 2008, a year after Chandler's death, both *Business History Review* and *Enterprise & Society* published reflections by prominent scholars on his legacy

Back to top

Selected works

1956. *Henry Varnum Poor: Business Editor, Analyst and Reformer*. Cambridge, Mass.: Harvard University Press.

1962. *Strategy and Structure*. Cambridge, Mass.: MIT Press.

中国社会科学院研究生院：研修之用

1977. *The Visible Hand: The Managerial Revolution in American Business*. Cambridge, Mass.: Belknap Press.

1990. *Scale and Scope: The Dynamics of Industrial Capitalism*. Cambridge, Mass.: Belknap Press.

1994. The competitive performance of U.S. industrial enterprises since the Second World War. *Business History Review* 68, 255–98.

2001. *Inventing the Electronic Century: The Epic Story of the Consumer Electronics and Computer Industry*. New York: Free Press.

2005. *Shaping the Industrial Century: The Remarkable Story of the Evolution of the Modern Chemical and Pharmaceutical Industries*. Cambridge, Mass.: Harvard University Press.

Back to top

Bibliography

Business History Review. 2008. Essays on Alfred D. Chandler. Vol. 82.

Enterprise and Society. 2008. Essays on Alfred D. Chandler. Vol. 9.

John, R. 1997. Elaborations, revisions, dissents: Alfred D. Chandler, Jr.'s, *The Visible Hand* after twenty years. *Business History Review* 71, 151–200.

McCraw, T.K. 1988. *The Essential Alfred Chandler: Essays Toward a Historical Theory of Big Business*. Boston, Mass.: Harvard Business School Press.

Scranton, P. 1997. *Endless Novelty: Specialty Production and American Industrialization, 1865–1925*. Princeton, NJ: Princeton University Press.

Supple, B. 1991. *Scale and Scope*: Alfred Chandler and the dynamics of industrial capitalism. *Economic History Review* 44, 500–14.

Back to top

How to cite this article

中国社会科学院研究生院：研修之用

Friedman, Walter A. "Chandler, Alfred D. (1918 – 2007) ." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 13 January 2010 <http://www.dictionaryofeconomics.com/article?id=pde2009_C000610>
doi:10.1057/9780230226203.1916

Download Citation:

中国社会科学院研究生院：研修之用

international coordination of regulation

David Lazer
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume
Back to top

Abstract

Trade among states with diverse regulatory systems creates the possibility of taking advantage of the cost differentials in production that result. However, it is increasingly clear that what happens in one jurisdiction affects policy in other jurisdictions. It is often argued that this creates a 'race to the bottom' effect, where the most lax regulation gains an advantage, but the evidence on this is mixed, at best, and there is a plausible argument too for a 'race to the top' effect, where states set high regulatory standards as a barrier to entry.
Back to top

Keywords

comparative advantage; Delaware effect; California effect; policy information interdependence; race to the bottom; race to the top; regulation; regulatory competition
Back to top

Article

Regulation is a pervasive element of modern government. The regulatory state has its hand in everything from the food we eat to the couches we sit on. The regulations we live with are necessarily a collective affair – as a general proposition two people who live in the same jurisdiction cannot choose different regulatory regimes. However, it is increasingly clear that regulation is a collective affair more broadly – what happens in one jurisdiction affects other jurisdictions. This interdependence creates potential governance challenges, in part due to strategic dilemmas that these interdependencies may create, as well as accountability issues that occur when important policies affecting a polity originate outside of that polity.

Samuelson (1949) offered an evocative thought experiment that sheds some light on the potential dilemmas. Imagine, Samuelson asked, the integrated

中国社会科学院研究生院：研修之用

economy of a world with no boundaries, no transportation costs. Now assume angels come along and divide the world up into different nations, with different allocations of factors of production. The question Samuelson posed was how international trade might allow the world to recapture that lost paradise, where the trade of goods is a substitute of sorts for a homogeneous distribution of factors of production.

Interestingly, international trade creates an opportunity where none existed in Samuelson's integrated economy, if we assume that Samuelson's integrated economy had a single regulatory regime. Specifically, trade among states with diverse regulatory systems creates the possibility of taking advantage of the cost differentials in production resulting from that regulatory heterogeneity. For example, if we imagine that state A has strict regulation of public good 1 and relaxed regulation of public good 2, and state B has relaxed regulation of public good 1 and strict regulation of public good 2, then both states can benefit from trade. Those sectors that can produce more cheaply in A (because of its more relaxed regulation of public good 1) will naturally arise there (even in the absence of capital mobility), and similarly with respect to sectors that would produce more cheaply in B. Generally, the basis for trade (national and international) rests in significant part on the pillar of heterogeneity (that is, comparative advantage). Samuelson's angels create the possibility of exchange based on regulatory heterogeneity that did not exist in the integrated economy.

The preceding assumes that policy is exogenous and that all factors of production are immobile. What happens if we relax these assumptions? One possibility is that regulatory policies will diverge, because trade reduces the adverse economic effects of those policies. To take an extreme example, imagine a state with a preference for strong regulation of a particular sector that makes some good for which there is inelastic demand. In a closed economy the benefits of strict regulation need to be weighed against the fact that the costs of that regulation will be fully borne by consumers. In an open economy, it may be possible to regulate that sector out of existence, with fairly minimal welfare impacts on society – because that sector may locate in another jurisdiction where the demand for regulation in that sector is lower. That is, it is theoretically possible that trade will enable some jurisdictions to regulate more strictly, thus raising the average level of regulation in the international system.

Receiving far more attention, however, is the possibility of a 'race to the bottom' (RTB), where the combination of trade and factor mobility yields uniform downward pressure on regulation across all jurisdictions. RTB has been asserted in many settings, although proven in few. The essential intuition is fairly simple. Consider the following simple model of the world, where there are two factors of production, say labour and

中国社会科学院研究生院：研修之用

capital. Labour is assumed to be immobile and capital mobile. If we assume that labour and capital are complements in production at the national level, this would yield a competition at the national level for division of the surplus produced through their combination. This intranational competition, in turn, has an international dimension because capital can flow to jurisdictions that offer the biggest share of that surplus. That is, effectively, labour from different jurisdictions will compete to attract capital. More capital will yield a larger surplus; however, the competition among jurisdictions means that most of that surplus will go to capital. We can see this type of dynamic, for example, in the competition among states in the United States to attract automobile factories, where, as Donahue (1997) documents, foreign manufacturers garnered enormous incentives from US state governments.

There is also clearly a RTB effect when there are physical externalities from one jurisdiction to another - for instance, when power plants are placed close to borders, and pollution spills over to the neighbouring jurisdiction. Except where noted, for this paper the case of physical externalities will be bracketed, because analytically it is uninteresting. In other words, it is clear that there is a potential for a collective failure when there are physical externalities (the most obvious contemporary example is carbon emissions).

When applied to the regulatory context, regulation may be seen, in part, as an effort by a jurisdiction to reallocate some of the surplus towards immobile actors. Environmental protection offers a nice illustration. Efforts to protect the local environment increase costs to capital. As a result, capital is less likely to locate in a highly regulated jurisdiction, lowering wages (relative to a more lax regulatory jurisdiction) to the point that capital is indifferent between locating in jurisdictions with different levels of regulatory stringency. The preceding discussion would suggest that given a large number of jurisdictions, and a lack of collusion among those jurisdictions, the net effect of dividing the world up will be to redistribute from immobile factors of production to mobile factors of production. In the example above, this redistribution would take place from the environment and from labour to capital.

The above discussion notwithstanding, RTB has had far more currency among politicians than economists. While analytic models, following from the studies of regulatory federalism by Tiebout (1956), sometimes find the possibility of an RTB, it is far from the typical finding (Oates, 2002). It should be noted, however, that the federal context is an imperfect analogue for the international context. For example, these models assume strong sorting effects of citizens with respect to, for example, preferences for environmental protection. It is less likely that such significant sorting effects exist at the international level.

中国社会科学院研究生院：研修之用

Similarly, empirically, there is relatively little support for RTB dynamics. For example, there is little evidence that firms locate based on relaxed regulatory regimes (Bartik, 1988). There is also little evidence that increased capital mobility over the last few decades has created downward movement in regulation, and some evidence that supports the opposite proposition (such as Engel, 1997; Fredriksson and Millimet, 2002; Frankel and Rose, 2005).

The lack of RTB effects may in part reflect that costs imposed by a regulatory system are a fairly small factor in the decision on location (Jaffe et al., 1995). Further, one likely and important reason for the lack of observation of RTB dynamics is the inability of researchers to capture the full range of reasons why firms decide to locate in particular jurisdictions. The decision to locate an oil rig in a particular location may be affected by environmental regulations, but first and foremost is certainly driven mostly by whether there is oil present at that location. An oil company, all else being equal, may prefer to drill in locations with less stringent environmental rules. However, those locations with the most desirable locations for drilling oil are also therefore in a position to seek a larger share of the income produced from that oil. This share, in those jurisdictions with preferences for a cleaner environment, would likely in part be extracted through stronger environmental regulations. Empirically, this might yield the outcome that those locations with the most drilling will also have the strongest regulations. Such a snapshot might be misleading, since it might still be the case that regulatory competition results in more lax regulation than in a counterfactual world where capital were not mobile, or where policymakers colluded.

More generally, there are a variety of reasons why particular locations might offer particular advantages for capital. Some of these may be exogenous (such as those related to the location of particular natural resources). Others may be endogenous. For example locating particular physical capital (such as factories) in a specific location might facilitate the creation of human capital in that location, which might be relatively immobile. Closely related, there might be returns to scale at the industry-jurisdiction level. That is, the productivity of a firm might be positively related to the number of other firms in a jurisdiction or region. In fact, clustering of sectors in particular regions is quite common (Krugman, 1996). If the emergence of such a cluster reflects external economies, this creates the possibility that a jurisdiction containing such a cluster can increase the stringency of its regulations without a worry that capital will flee (Baldwin and Krugman, 2004).

The debate in the legal literature regarding the clustering of industry incorporations in Delaware offers an illuminating case study that illustrates this point. The starting point for this literature is the

中国社会科学院研究生院：研修之用

observation that Delaware towers over the rest of the United States in terms of number of incorporations. This was originally viewed as evidence of a lax regulatory regime in Delaware (Cary, 1974). The essential argument was that corporations located in Delaware because doing so minimized the burdens placed on corporate management, often at the expense of shareholders. The hypothesized RTB was branded, as a result, as the 'Delaware effect'. This argument came under sharp criticism in the 1990s. The first critique was that there would be shareholder pressure (exerted in part through the stock price) against locating in a jurisdiction that did not preserve shareholder value (Revesz, 1992). The dominance of Delaware thus reflects the capacity of Delaware to effectively manage corporate governance while still preserving shareholder value. That is, the dominance of Delaware reflects the benefits of regulatory competition, where Delaware simply provides superior governance to everyone else. A second critique (of both the benefits and the costs of regulatory competition) was that there are economies of scale in providing good corporate governance (Kahan and Kamar, 2003). For example, good corporate governance requires predictability, and predictability is facilitated by ample case law. Delaware, by dint of historical accident, had garnered an insurmountable lead in effectively producing good corporate law. The implication of this, in turn, is that Delaware has a fair degree of slack in extracting some surpluses from the regulated parties, as long as it does so in a way that does not threaten its competitive advantage in corporate governance.

Note that there are RTB debates in other domains, such as welfare benefits (Dahlberg and Edmark, 2008; Bailey and Rom, 2004) and corporate tax rates (Basinger and Hallerberg, 2004).

While these RTB debates rage on, with strong intuitions and political appeal ranged against modest analytic and empirical support from the economics literature, there is a stronger consensus about the potential capture of the regulatory system by domestic interests (Bartel and Thomas, 1987) - which I will label 'regulation as protection' (RAP). To paraphrase Clausewitz, regulation may be viewed as market competition through other means. It is rare for regulation to be neutral in its impact on producers in a given sector. A more restrictive regulation has the potential to benefit some producers at the expense of others. For example, requirements for greater efficiency in cars benefit producers that already produce high-efficiency cars, at the expense of producers of low-efficiency cars. In the context of international trade, the alignment of interests will often be domestic producers opposed to international producers interested in entering the domestic markets and domestic consumers who might benefit from increased competition in the home market. The potential scenario is for domestic producers to hijack the domestic regulatory apparatus as a means to block international competition.

中国社会科学院研究生院：研修之用

It is also clear that in many sectors there are strong drivers for convergence in regulation. A simple example is the mandate in a jurisdiction to drive on one side or the other of the road. Given significant traffic between two jurisdictions, it is clear that there would be a strong benefit to a consistent standard for driving. Thus, for example, Canada, which had a patchwork of standards in the first half of the 20th century at the provincial level, with some provinces mandating left-hand traffic, and others right-hand traffic, gradually converged to right-hand traffic, with uniformity emerging by the middle of the century. This presumably was driven in part by the need for internal consistency, and in part because of the degree of traffic between the United States and Canada. In the regulatory context, such convergence may be seen fairly routinely across policy areas, for example in food safety (Lazer, 2001), where small countries adopt the standards of their large export markets. Regulatory convergence more generally may take place to guard export markets. As Vogel (1995) argues, this type of push for convergence often occurs in international trade when there are increasing returns to scale in production, combined with potential divergence in product standards. Thus, for example, Vogel documents the flow of environmental standards for automobiles around the world, which he labels the race to the top (RTT) 'California effect' in contrast to the RTB 'Delaware effect'. Vogel argues that there is potentially a systematic bias toward higher standards because of the efficiency imperatives of having consistent standards. Such an imperative should lead toward convergence on the strictest standard on the system (or at least a standard that would encompass the large majority of the system), so as to achieve efficiencies in scale of production. Such a diffusion pattern should follow the reverse direction of exports, for example, as Prakash and Potoski (2006) found with respect to the spread of ISO 14001 adoption.

Finally, even in the absence of trade or factor mobility, there is significant potential for regulatory interdependence due to policy information interdependence (PII). Policy is necessarily experimentation, and novel policies even more so. Regulation in jurisdiction A creates insights in jurisdiction B as to what would be good or bad policy, where this information spreads through various informational networks (Wolman and Page, 2002; Lazer, 2005). Emulation may reflect lesson drawing or serve the function of policy legitimation (Bennett, 1997; Busch, Jorgens and Tews, 2005). There is a substantial literature on policy emulation (e.g. Rose, 1993; Haas, 1992), and it is clear that there is no reason to expect domestic regulatory policy to be exempt from emulation (e.g. Simmons and Elkins, 2004).

This array of interdependencies offers an array of empirical challenges for the researcher and governance challenges for the policy maker. For the researcher, the potential presence of different processes (which, if

中国社会科学院研究生院：研修之用

validated, have very different normative implications) creates a difficult but not impossible nut to crack (for various recent efforts to deal with exactly this issue, see Braun and Gilardi, 2006; Simmons, Dobbin and Garrett, 2006; Levi-Faur, 2005).
For the policy maker each of these kinds of regulatory interdependencies creates different types of collective strategic interaction problems (Lazer, 2001, 2006), which in turn translate into a collective governance challenge (Scharpf, 1997). RTB and RAP might be categorized as a prisoner's dilemma, where the potential dysfunctional equilibrium would be either suboptimally lax standards in the case of RTB, or suboptimally strict standards in the case of RAP. RTT may be viewed as a coordination game, where one potentially problematic outcome is the emergence of suboptimally strict standards (or perhaps worse would be the case where the RTT did not occur, where a critical mass of support did not emerge for any standard), with a handful of large jurisdictions driving the standards for the world. All of these issues might call for some type of negotiated standard. However, such centrally negotiated standards might eliminate some of the very benefits of trade in the first place. And the presence of PII would suggest a different kind of public goods issue than is usually associated with regulation - the public good of information. This construction of regulatory interdependence suggests a dual conundrum. The first aspect of this is how to take advantage of the publicness of the information, and the second is how to support continued production of this public good (Lazer, 2005).
Back to top

Acknowledgement

Thanks go to Elta Smith, for her research assistance with respect to this paper.
Back to top

See Also

- comparative advantage

Back to top

Bibliography

Bailey, M. and Rom, M.C. 2004. A wider race? Interstate competition across health and welfare programs. *Journal of Politics* 66, 326–47.

中国社会科学院研究生院：研修之用

Baldwin, R. and Krugman, P. 2004. Agglomeration, integration and tax harmonization. *European Economic Review* 48, 1－23.

Bartel, A. and Thomas, L.G. 1987. Predation through regulation: the wage and profit effects of the Occupations Safety and Health Administration and the Environmental Protection Agency. *Journal of Law and Economics* 30, 239－64.

Bartik, T. 1988. The effects of environmental regulation on business location in the United States. *Growth and Change* 19, 22－44.

Basinger, S. and Hallerberg, M. 2004. Remodeling the competition for capital: how domestic politics erases the race to the bottom. *American Political Science Review* 98, 261－76.

Bennett, C. 1997. Understanding ripple effects: the cross-national adoption of policy instruments for bureaucratic accountability. *Governance* 10, 213－33.

Braun, D. and Gilardi, F. 2006. Taking 'Galton's problem' seriously: towards a theory of policy diffusion. *Journal of Theoretical Politics* 188, 298－322.

Busch, P. , Jorgens, H. and Tews, K. 2005. The global diffusion of regulatory instruments: the making of a new international environmental regime. *Annals of the American Academy of Political and Social Science* 598, 146－67.

Cary, W. 1974. Federalism and corporate law: reflections upon Delaware. *Yale Law Journal* 83, 663－666.

Dahlberg, M. and Edmark, K. 2008. Is there a 'race-to-the-bottom' in the setting of welfare benefit levels? Evidence from a policy intervention. *Journal of Public Economics* 92, 1193－1209.

Donahue, J. 1997. *Disunited States* . New York: Basic Books.

Engel, K. 1997. State environmental standard-setting: is there race and is it to the bottom? *Hastings Law Journal* 48, 271－398.

Frankel, J. and Rose, A. 2005. Is trade good or bad for the environment? Sorting out the causality. *Review of Economics and Statistics* 87, 85－91.

中国社会科学院研究生院：研修之用

Fredriksson, P. and Millimet, D. 2002. Strategic interaction and the determination of environmental policy across U.S. States. *Journal of Urban Economics* 51, 101–22.

Haas, P. 1992. Introduction: epistemic communities and international policy co-ordination. *International Organization* 46, 1–35.

Jaffe, A. , Peterson, S.R. , Portney, P.R. and Stavins, R.N. 1995. Environmental regulations and the competitiveness of U.S. manufacturing: what does the evidence tell us? *Journal of Economic Literature* 33, 132–63.

Kahan, M. and Kamar, E. 2003. The myth of state competition in corporate law. *Stanford Law Review* 55, 679–750.

Krugman, P. 1996. *The Self Organizing Economy*. Cambridge, MA: Blackwell.

Lazer, D. 2001. Regulatory interdependence and international governance. *Journal of European Public Policy* , 474–92.

Lazer, D. 2005. Regulatory capitalism as a networked order: the international system as an informational network. In *The Rise of Regulatory Capitalism: The Global Diffusion of a New Order* , ed. D. Levi-Faur and J. Jordana . London: Sage.

Lazer, D. 2006. Global and domestic interdependence: modes of interdependence in regulatory policymaking. *European Law Journal* 12, 455–68.

Levi-Faur, D. 2005. The political economy of legal globalization: juridification, adversarial legalism, and responsive regulation. A comment. *International Organization* 59, 451–62.

Oates, W. 2002. Fiscal and regulatory competition: theory and evidence. *Perspektiven der Wirtschaftspolitik* 3, 377–90.

Prakash, A. and Potoski, M. 2006. Racing to the bottom? Trade, environmental governance, and ISO 14001. *American Journal of Political Science* 50, 350–64.

Revesz, R. 1992. Rehabilitating interstate competition: rethinking the 'race to the bottom' rationale for federal environmental regulation. *New York University Law Review* 67, 1210–54.

中国社会科学院研究生院：研修之用

Rose, R. 1993. *Lesson-Drawing in Public Policy* . New Jersey: Chatham House.

Samuelson, P. 1949. International factor price equalization once again. *Economic Journal* 58, 163 – 84.

Scharpf, F. 1997. Introduction: the problem-solving capacity of multi-level governance. *Journal of European Public Policy* 4, 520 – 38.

Simmons, B. , Dobbin, F. and Garrett, G. 2006. Introduction: the international diffusion of liberalism. *International Organization* 60, 781 – 810.

Simmons, B. and Elkins, Z. 2004. The globalization of liberalization: policy diffusion in the international political economy. *American Political Science Review* 98, 171 – 89.

Tiebout, C. 1956. A pure theory of local public expenditures. *Journal of Political Economy* 64, 416 – 24.

Vogel, D. 1995. *Trading Up: Consumer and Environmental Regulation in a Global Economy* . Cambridge, Mass.: Harvard University Press.

Wolman, H. and Page, E. 2002. Policy transfer among local governments: an information-theory approach. *Governance* 15, 477 – 501.

Back to top

How to cite this article

Lazer, David. "international coordination of regulation." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 13 January 2010 <http://www.dictionaryofeconomics.com/article?id=pde2009_I000306>
doi:10.1057/9780230226203.1917

中国社会科学院研究生院：研修之用

matching models: empirics

Jeremy T. Fox
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume
Back to top

Abstract

A matching model takes a set of payoffs or outputs for all possible matches and produces a set of matches where no couple would prefer to deviate and become matched, instead of their assigned matches. Matching models are increasingly being estimated in empirical work in industrial organization, labour economics, public economics, and other fields. This article surveys methods for and applications of structural estimation for two-sided matching games.
Back to top

Keywords

cooperative games; econometrics; family economics; industrial organization; labour market; marriage market; matching models; mergers; pairwise stability; structural estimation; two-sided matching games
Back to top

Article

Economists often observe data on relationships. We see who is in a relationship with whom: which firms merged with each other, which men are married to which women, and which bidders won which items in an auction. A matching model or matching game is one theoretical framework for modelling the equilibrium formation of these relationships. A relationship is termed a match. A matching model takes a set of payoffs or outputs for all possible matches and produces a set of matches where no couple would prefer to deviate and become matched, instead of their assigned matches. The robustness of the equilibrium to deviation by any potential couple suggests ‘pairwise stability’ as the term for an equilibrium.

The key economic idea in matching models is the rivalry to match with the most attractive partners. In marriage, men compete with each other to marry the most attractive women while women compete with each other to

中国社会科学院研究生院：研修之用

marry the most attractive men. There is scarcity on both sides of the market.

This entry focuses on research that formally estimates the parameters of a matching model using data on who matches with whom. Economists adopting this structural approach typically observe data on who matches with whom as well as exogenous characteristics of each agent. For example, economists studying marriage will observe the race, schooling level, religion, physical attractiveness and wage of each man and each women. The data also record which men married which women. Economists are willing to assume that the data represent a pairwise stable outcome to a particular matching game.

The structural approach means that researchers impose the structural model and estimate unknown parameters in the model. The advantages of the structural approach apply to more economic situations than just matching models, and have been explored elsewhere (Reiss and Wolak, 2007). A quick summary is that the structural approach allows the computation of counterfactuals and the measurement of economic parameters that cannot be directly observed. Using the example of marriage, one type of counterfactual would be to explore how the equilibrium set of matches is altered as demographics change. Measurement is also important: if men and women each have several characteristics, how important are each of the characteristics in the payoffs to a match?

This article focuses on the use of matching games in structural empirical work. Other literatures have focused on centralized market design (for example the medical resident matching programme in the United States) and the descriptive interpretation of matching patterns.

Back to top

Two-sided matching games

Most but not all empirical work has focused on two-sided matching games: agents can be divided into two sides, say men and women. Roth and Sotomayor (1990) is a useful text that explains simple matching games.

For the purposes of this article, I divide two-sided matching games into models with and without transfers. Gale and Shapley (1962) introduced the model where agents do not exchange money: men have preferences over women and women have preferences over men. Generically there will be a lattice of multiple pairwise stable outcomes in this model. Koopmans and Beckmann (1957), Shapley and Shubik (1972), and Becker (1973) study models where matched agents can exchange money and where agents have transferable utility. For one-to-one matching games such as marriage, generically there will be one set of pairwise stable physical matches in these models; there may be a continuum of transfers that support these physical matches.

中国社会科学院研究生院：研修之用

The choice of modelling framework depends on the researcher's understanding of the market in question.
The above papers allow each man, say, to marry at most one woman. There are extensions to many-to-one and many-to-many two-sided matching games. Complementarities between multiple matches involving the same agent are key to some of the empirical applications below (Fox, 2009a; Fox and Bajari, 2009). There are also one-sided and many-sided matching games.
There are more general matching games where other contract elements, such as the hours of work in a labour market, are determined as part of the pairwise stable outcome (Crawford and Knoer, 1981; Kelso and Crawford, 1982; Hatfield and Milgrom, 2005). Matching games are mathematically linked to hedonic equilibrium models, although I will not explore the link here (Rosen, 1974). There is also a clear link to models of frictions, such as search models, that also have observed agent heterogeneity (Shimer and Smith, 2000; Atakan, 2006).
Back to top

Estimation methods

Matching games share many similarities with the literature on estimating static, discrete Nash games, such as the well-known entry models of Berry (1992) and Bresnahan and Reiss (1991).
Matching games use pairwise stability and not Nash equilibrium, but many estimation challenges are similar. A key difficulty in matching games is that the number of agents in a market can be in the hundreds or thousands, compared to the three or four firms deciding to enter a market in some entry applications. The number of agents in matching empirical applications can make some estimators computationally infeasible.
Back to top

Nested solution methods

The most straightforward way to estimate a matching game is to use simulated maximum likelihood or the simulated method of moments. These estimators require solving the model a fixed number of times for each iteration of an outer optimization routine. Simulation estimators are conceptually straightforward but computationally burdensome.
Boyd and colleagues (2003) use the simulated method of moments to study the matching of public school teachers to schools in New York state. They use data on wages and assume the wages are exogenously determined. Their model without endogenous transfers has multiple stable matches, and they use the lattice structure of the equilibria to impose an equilibrium selection rule in estimation.
Back to top

中国社会科学院研究生院：研修之用

Full likelihood methods

In many cases, the full likelihood can be written down. In a study of the matching of venture capitalists to entrepreneurs, Sørensen (2007) uses an augmented likelihood approach where the unobserved payoffs of each match are treated as nuisance parameters and integrated out using a blocking structure in a Markov Chain Monte Carlo (MCMC) procedure. Sørensen does not use endogenous transfers and hence imposes an aligned preferences assumption that he proves generates a unique pairwise stable outcome.

The full likelihood approach can be computationally intractable in large matching markets. In a study of the matching between investment banks and firms undertaking an initial public offering, Akkus (2008) shows that the likelihood simplifies if the values of realized matches are recorded in the data. By using data on the payoffs of matches, estimation becomes easier.

Back to top

Inequality methods

With an application to automotive suppliers and automotive assemblers, Fox (2009a) introduces a maximum score estimator to estimate a many-to-many matching game where transfers are endogenous, but not in the data. The maximum score estimator maximizes the number of inequalities implied by pairwise stability that hold true. This approach breaks the computational curse of dimensionality because not all inequalities need to be included for the estimator to be consistent.

Back to top

Logit methods

Dagsvik (2000) and Choo and Siow (2006) study games with transfers, and assume that the payoffs to matches have error terms that satisfy the parametric logit property. To a large degree, the logit assumption allows researchers to derive closed-form equations that allow estimation, especially for very large markets that plausibly have a continuum of agents, such as the US national marriage market.

Back to top

Identification

In matching games, agents on the same side of the market are rivals to match with agents on the other side of the market. The fact that a man did not match with the most attractive woman does not mean that the man did not prefer that woman to his actual wife. The equilibrium budget set

中国社会科学院研究生院：研修之用

of each agent is unobserved. Thus, it is not clear what can be learned (identified) from data on who matches with whom.
Fox (2009b) studies identification in matching games with transfers and finds two sets of results. First, the relative importance of complementarities in payoffs for say schooling, compared with say wealth, is identified using data on matches but not the equilibrium transfers that are present in the model. Second, the ordering of production levels (which matches give higher payoffs) is identified using the same data.
Back to top

Selection correcting outcome equations

Sørensen (2007) explores the use of a matching model to parametrically selection correct an auxiliary outcome equation. The outcome, the success of an investment in his application, is not determined as part of the matching game, but the outcome is only observed in the data for realized matches. Sørensen's approach is analogous to using a single agent decision model to selection correct an outcome equation (Heckman, 1979). Sørensen (2009) extends the framework of Heckman (1990) to study identification in selection models where selection is induced by a matching game.
Back to top

Empirical applications

I now catalogue some of the many empirical applications of matching games.
Back to top

Marriage and family economics

Choo and Siow (2006) explore whether changing matching patterns in the US marriage market are due to changes in preferences or changes in the exogenous characteristics of potential spouses. They also explore the effects of the legalization of abortion. There have been a large number of marriage and family economics papers following up on the Choo and Siow framework, many by the original authors. See Siow (2008) for a complete survey of this material.
Bruze (2009) estimates a matching game where labour supply and the split of consumption between men and women are part of the pairwise stable contract terms. He explores the return to an agent for finding a higher-earning spouse in college.
Hitsch, Hortaçsu and Ariely (2009) use revealed preference information from an online dating site to avoid the need to use an equilibrium model to estimate preferences. They use the preference estimates to simulate a pairwise stable outcome and find it matches well with actual sorting on the site.

中国社会科学院研究生院：研修之用

Back to top

Industrial organization, corporate finance and marketing

Hall (1988) is an early paper that emphasizes the need for matching models to study mergers. She did not estimate a full matching game because of computational concerns.

It is common to have data on realized interfirm relationships. Sørensen (2007) studies the matching of venture capitalists to entrepreneurs with a focus on selection correcting an outcome equation where the success of an investment is regressed upon the experience of the venture capitalist. The basic approach of Sørensen allows for match-specific error terms, and he can allow for time-invariant characteristics of a venture capitalist by using fixed effects/panel data. Chen (2009) uses a similar selection-correction framework where the outcome equation of interest is the price of a bank loan. Akkus (2008) uses the selection-correction approach to regress the degree of first-day underpricing on the experience of an investment bank. Park (2008) uses a similar MCMC estimator to investigate the decision of a mutual fund manager to engage in a merger as a function of past returns.

Fox and Bajari (2009) was the first paper to estimate a many-to-one matching game where complementarities across multiple matches were allowed for. The authors look at auctions of multiple heterogeneous items, where each bidder can win multiple items. They study FCC spectrum auctions, where complementarities between the geographic territories being auctioned are estimated to be important for the efficient operation of the mobile phone industry. A key methodological challenge is showing how a potentially inefficient, dynamic Nash game could result in equilibria that satisfy pairwise stability. The estimator used is that of Fox (2009a) for matching games with transfers. Fox (2009a) studies the many-to-many matching of automotive suppliers to automotive assemblers, and measures the relative importance of specialization by suppliers in particular corporations, brands and car models. Further, Fox measures a potential benefit of suppliers matching with high-quality Asian assemblers, such as Toyota.

Levine (2008) uses the estimator of Fox (2009a) to explore the matching of biotechnology innovations to marketing firms. She explores whether the returns to scale of marketers might decrease the returns to innovators. Yang, Shi and Goldfarb (2009) use the same estimator to explore the matching of professional athletes to teams, with a focus on the potential marketing complementarities between players and teams from different-sized cities. Akkus and Hortaçsu (2007) extend the maximum score estimator to use data on equilibrium transfers. Akkus and Hortaçsu investigate geographic complementarities in the market for bank mergers

中国社会科学院研究生院：研修之用

after the removal of prohibitions against interstate banks. Mindruta (2009) studies the matching of university researchers and private firms.
Back to top

Development, public finance, labour economics and other

Boyd and colleagues (2003) investigate the matching of teachers to public schools, with a focus on learning how to attract qualified teachers to schools in impoverished areas.
Gordon and Knight (2009) investigate the consolidation decisions of Iowa school districts after the state passed incentives inducing such consolidation.
Ahlin (2009) uses the estimator of Fox (2009a) to study the matching of Thai villagers to other villagers in risk-sharing groups. He investigates whether villagers match by risk type or seek to diversify risk.
Fox (2008) estimates a repeated matching model for the labour market for engineers in Sweden. Each period state variables evolve, a matching model opens, prices are formed to equate supply and demand and workers choose jobs. The model is dynamic in that both firms and workers are forward looking: they consider the effect of the decision to switch today on future outcomes.
Baccara and colleagues (2009) study the matching of professors to offices, and estimate the importance of various types of professional networks in payoffs.
Back to top

See Also

- assortative matching
- econometrics
- marriage markets
- matching

Back to top

Bibliography

Ahlin, C. 2009. Matching for credit: risk and diversification in Thai micro-borrowing groups. Michigan State University working paper, September.

Akkus, O. 2008. Endogenous matching and the determinants of underpricing in IPO markets. University of Chicago working paper, November.

中国社会科学院研究生院：研修之用

Akkus, O. and Hortaçsu, A. 2007. The determinants of bank mergers: a revealed preference analysis. University of Chicago working paper, March.

Atakan, A.E. 2006. Assortative matching with explicit search costs. *Econometrica* 74, 667－81.

Baccara, M. , Imrohoroglu, A. , Wilson, A. and Yariv, L. 2009. A field study on matching with network externalities. New York University working paper, August.

Becker, G.S. 1973. A theory of marriage: Part I. *Journal of Political Economy* 81, 813－46.

Berry, S.T. 1992. Estimation of a model of entry in the airline industry. *Econometrica* 60, 889－917.

Boyd, D. , Lankford, H. , Loeb, S. and Wyckoff, J. 2003. Analyzing determinants of the matching of public school teachers to jobs: estimating compensating differentials in imperfect labor markets. University of Buffalo working paper.

Bresnahan, T.F. and Reiss, P.C. 1991. Empirical models of discrete games. *Journal of Econometrics* 48, 57－81.

Bruze, G. 2009. Schooling, marriage, and male and female consumption. Aarhus University working paper, March.

Chen, J. 2009. Two-sided matching and spread determinants in the loan market. University of California at Irvine working paper, May.

Choo, E. and Siow, A. 2006. Who marries whom and why. *Journal of Political Economy* 114, 175－201.

Crawford, V.P. and Knoer, E.M. 1981. Job matching with heterogeneous firms and workers. *Econometrica* 49, 437－50.

Dagsvik, J.K. 2000. Aggregation in matching markets. *International Economic Review* 41, 27－57.

Fox, J.T. 2008. An empirical, repeated matching game applied to market thickness and switching. University of Chicago working paper.

Fox, J.T. 2009a. Estimating matching games with transfers. University of Chicago working paper, August.

中国社会科学院研究生院：研修之用

Fox, J.T. 2009b. Identification in matching games. University of Chicago working paper, June.

Fox, J.T. and Bajari, P. 2009. Measuring the efficiency of an FCC spectrum auction. University of Chicago working paper.

Gale, D. and Shapley, L. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69.

Gordon, N. and Knight, B. 2009. A spatial merger estimator with an application to school district consolidation. *Journal of Public Economics* 93, 752–65.

Hall, B.H. 1988. Estimation of the probability of acquisition in an equilibrium setting. University of California at Berkeley working paper, August.

Hatfield, J.W. and Milgrom, P.R. 2005. Matching with contracts. *American Economic Review* 95, 913–35.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–62.

Heckman, J.J. 1990. Varieties of selection bias. *American Economic Review* 80, 313–18.

Hitsch, G. , Hortaçsu, A. and Ariely, D. 2009. Matching and sorting in online dating markets. *American Economic Review*.

Kelso, A.S. and Crawford, V.P. 1982. Job matching, coalition formation, and gross substitutes. *Econometrica* 50, 1483–1504.

Koopmans, T.C. and Beckmann, M. 1957. Assignment problems and the location of economic activities. *Econometrica* 25, 53–76.

Levine, A.A. 2008. Licensing and scale economies in the biotechnology pharmaceutical industry. Harvard University working paper, June.

Mindruta, D. 2009. Value creation in university-firm research collaborations: a matching approach. HEC Paris working paper, February.

Park, M. 2008. An empirical two-sided matching model of acquisitions: understanding merger incentives and outcomes in the mutual fund industry. University of Minnesota working paper, September.

中国社会科学院研究生院：研修之用

Reiss, P.C. and Wolak, F.A. 2007. Structural econometric modeling: rationales and examples from industrial organization. In *Handbook of Econometrics, Vol. 6A*. Elsevier.

Rosen, S. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* 82, 34.

Roth, A.E. and Sotomayor, M.A. O. 1990. Two-sided matching number 18. Econometric Society Monograph, Econometric Society.

Shapley, L.S. and Shubik, M. 1972. The assignment game I: the core. *International Journal of Game Theory* 1, 111–30.

Shimer, R. and Smith, L. 2000. Assortative matching and search. *Econometrica* 68, 343–69.

Siow, A. 2008. How does the marriage market clear? An empirical framework. University of Toronto working paper, July.

Sørensen, M. 2007. How smart is smart money? A two-sided matching model of venture capital. *Journal of Finance* 62, 2725–62.

Sørensen, M. 2009. Identification of multi-index selection models. Columbia University working paper.

Yang, Y. , Shi, M. and Goldfarb, A. 2009. Estimating the value of brand alliances in professional team sports. *Marketing Science*.

Back to top

How to cite this article

Fox, Jeremy T. "matching models: empirics." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 13 January 2010 <http://www.dictionaryofeconomics.com/article?id=pde2009_M000426> doi:10.1057/9780230226203.1919

Download Citation:

as RIS | as text | as CSV | as BibTex

中国社会科学院研究生院：研修之用

preindustrial inequality

Branko Milanovic
From The New Palgrave Dictionary of Economics Online, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume
Back to top

Abstract

This article considers inequality in preindustrial societies, defined as those prior to the industrial revolution and subsequent non-industrial societies that are not systematically integrated into the advanced world's economy. Although data on individual incomes and wealth in these societies are limited, increasingly they are becoming available. On the basis of these data, inequality as measured by the Gini coefficient is often on a par with modern industrialized societies, but the income gradient tends to be different, with a mass of people at subsistence level or marginally above, few at the mean, and a small affluent class. More work remains to be done, particularly on the relationship between income inequality and economic progress.
Back to top

Keywords

Gini coefficient; income distribution; inequality; Kuznets curve; preindustrial societies
Back to top

Article

Back to top

Defining preindustrial

We need to circumscribe the scope of preindustrial. At one level, it is easy: preindustrial economies are characterized by low urbanization rates, high share of agriculture in GDP, low literacy rates, and of course low overall GDP per capita. However, many of today's poor countries share precisely these features. They are however 'non-industrial' or 'non-industrialized' rather than 'preindustrial' economies: this is because they are part of the modern world, systematically included in

中国社会科学院研究生院：研修之用

trade and voluntary movements of factors of production ('globalization') and have social structures which are very different from those of preindustrial societies. The life expectancy of their populations as well as the immunization and school enrolment rates exceed many times those of 'true' preindustrial societies. Not the least important is the fact that political compulsion of slave or serf labour, so ubiquitous in all preindustrial societies, is - except in a few pockets - largely absent. Our definition of preindustrial includes all societies prior to the industrial revolution, and those that have not engaged with the industrial revolution, only up to a point when they began to be integrated systematically, rather than episodically, into the world economy. For many of them, integration coincides with colonization. Thus, broadly speaking - since we are painting with a very broad brush here - we can set limits around the end of the Napoleonic wars for Western Europe and the United States and Canada, and the end of the 19th century for everybody else. Twentieth-century societies, even when poor and hardly industrialized, belong to a different category.

A cut-off date around 1815–20 is convenient for at least three reasons. Politically, it coincides with a 'rearrangement' of Europe and, as later emerged, the world. It marks the beginning of the 'long 19th century'. Economically, it marks, according to the new English wage data series produced by Clark (2005), the beginning of a long-run rise in real wages which is continuing to this day. In terms of history of economic thought, Ricardo's *Principles* were published in 1817.

An obvious, but nevertheless important, clarification is that we are concerned here with income inequality: that is, inequality that includes all sources of income and reflects differences in households' and individuals' living standards. This, for example, rules out wage or rural–urban inequalities as such. (Wage inequality has meaning only if calculated across all wage-earners; income inequality includes the entire population.)

Back to top

Implicit theory

We do have an implicit theory about income inequality in preindustrial economies. The Kuznets hypothesis (formulated in 1955), the bread and butter of inequality economics, posits that inequality charts an inverted U shape as economy transforms from predominantly agricultural to predominantly industrialized or modern. In Kuznets' own words:

One might thus assume a long swing in the inequality characterizing the secular income structure: widening in the early phases of economic growth when the transition from the preindustrial civilization was most rapid, becoming stabilized for a while; and then narrowing in the later phases.

中国社会科学院研究生院：研修之用

(Kuznets, 1955, p. 276)
The same hypothesis, albeit without the mechanism that generates the inverted U-shaped curve, was formulated 120 years before Kuznets by Tocqueville:
If one looks closely at what has happened to the world since the beginning of society, it is easy to see that equality is prevalent only at the historical poles of civilization. Savages are equal because they are equally weak and ignorant. Very civilized men can all become equal because they all have at their disposal similar means of attaining comfort and happiness. Between these two extremes is found inequality of condition, wealth, knowledge-the power of the few, the poverty, ignorance, and weakness of all the rest.
(Tocqueville, 1835, pp. 42–3)
From both we should retain the sense that inequality is supposed to emerge only when societies are richer, and thus inequality in preindustrial societies may be expected to be low. But differently, we also have an image of preindustrial societies as combining abject poverty in the bottom with extravagant wealth on the top. For example, in ancient Rome, Goldsmith (1984, p. 287) notes the extraordinarily high income of the rulers relative to Great Britain in the early 19th century. Could both these images be right? As we shall argue below, yes – and this is one of the key features that distinguishes inequality in premodern times from inequality in modern times.
But in order to speak about inequality in preindustrial societies, we must also assume that preindustrial societies were 'modern' in the sense that they were (predominantly) market-oriented economies with non-negligible monetized sectors – and when they were non-monetized, goods and services given or received for political or power reasons could be valued at some meaningful 'market' prices. This is a position not universally accepted. In a famous debate about the later Roman Empire (and, by extension about all ancient economies) and 'modernity', there were two camps: that of 'primitivists' led by Polanyi (1944), Finley (1985) and Schiavone (1995), and that of 'modernists' (Rostovtzeff, 1926; Walbank, 1946). The first believed that Rome lacked most of the modern concepts that we associate with a market economy. Market relations, even when present, were of peripheral importance, and a market economy, itself a recent phenomenon, is perhaps, in a historical sense, only a brief episode (Polanyi, 1944). For the 'modernists', the links between a preindustrial society like Rome and modern capitalism were obvious. Both Rostovtzeff and Walbank write of Roman 'bourgeoisie'. Whatever our opinion about the respective merits of 'primitivists' and 'modernists', it is important to realize that once we attempt to make some tentative estimates of economic inequality in preindustrial societies, we *ipso facto* accept that, while preindustrial societies might have been poorer and with a different social

中国社会科学院研究生院：研修之用

structure from modern societies, the differences are of magnitude, not of kind. For if such key concepts of market economy as prices, wage-labour and private property are vague, insufficiently understood by the population, not sanctioned by custom or law, then applying modern economic categories may be meaningless. Every attempt to study preindustrial societies empirically using today's economists' tools must assume that 'ancient' and 'modern' are fundamentally the same - so that that the 'ancient' can be described and understood using economic concepts developed from Adam Smith onwards.

Private property must enter the list above with a caveat. No one would deny that socialist societies, where private property was limited, were not modern. Moreover, they regarded themselves as the epitome of modernity. Similarly, societies with largely communal ownership of land (as in Africa) are modern too. Thus, private property of the means of production seems to be less of a requirement for a modern society than for example monetization. Rawls (1971), who can hardly be seen as a non-modernist, allows in his *Theory of Justice* for both private and non-private ownership of the means of production (see pp. 54, 240 - 1).

Back to top

Data for preindustrial inequality

Where do data for preindustrial inequality come from? Since the Second World War, empirical studies of income distribution have been based on household surveys (nationally representative samples of households who are anonymously interviewed about their household characteristics, spending patterns and income). The earliest household surveys are from late 18th-century England. There were a few sporadic surveys in the 19th century (continental Europe, rural Russia) but they spread broadly only after the end of the Second World War, and as far as Africa and China are concerned, surveys became available only more recently, from the early 1980s. Obviously, such surveys were not conducted in any preindustrial society - even if censuses (driven by government tax needs) were. However, there are relatively abundant sources that economists can use to gauge income distribution in preindustrial societies, although the sources are often buried in hard-to-access archives and books, written in languages and alphabets that are not widely known, and requiring large amounts of both money and effort to be brought to light in a usable form. (For example, Ottoman censuses are written in Turkish but using Arabic script, rather than Latin as is used in today's Turkish. To process them requires knowledge of an often archaic Turkish and an alphabet into which this language is no longer written. See Cosgel, 2002, 2004.) And then lots of heroic assumptions are needed in order for them to be 'translated' into modern economic categories. This has severely limited the use of ancient

中国社会科学院研究生院：研修之用

sources, and this is probably why only a fraction of such sources has been used so far.

The most comprehensive contemporary sources are tax data and government censuses undertaken in order to supply governments with information about taxation and the war-waging capacity of the populace (number of men, houses, horses, grain). Early documentary evidence includes government edicts (such as Diocletian's edict on maximum prices and wages from 301, recently studied by Allen, 2007), as well as numerous Roman papyri preserved in the dry climate of Egypt. The English Domesday survey of 1086 is perhaps the best known of such sources.

From the Byzantine Empire, we have a few preserved *praktika* that provide descriptions of household characteristics, inventories of possessions and taxes paid, although they cover only limited areas (towns or ecclesiastical communities). (See the multi-volume *Economic History of Byzantium: From the Seventh through the Fifteenth Century* edited by Angeliki Laiou, 2002.) Ottoman censuses (*defterlar*) from approximately the 14th century onward, conducted to assess the wealth and military capacity of newly conquered territories, provide detailed information on settlements (hamlets, villages, small and larger towns) but then present it in average amounts for each settlement (not by individual household). If inequality within settlements is not huge, and the number of settlements included is large, censuses can be used to assess overall income distribution within a country or a region.

A much-used source is the Florentine *Catasto* from 1427. (The data were originally collated by Herlihy and Klapisch-Zuber, 1985. Currently, they are available on the Internet.) The Spanish *Ensenada Cadastre*, similar to modern-day household surveys, was carried out in the 1750s for the purposes of a never-implemented fiscal reform. It has recently been used by researchers, and will be no doubt analysed more once it is digitized. Inequalities for the cities of Paris, Amsterdam and London were studied from tax data for respectively 1292–1313 (Sussman, 2005), 1732–42 (McCants, 2007; Soltow, 1989) and 1797–1801 (Schwarz, 1979). However, they refer to wealth inequality (there is no attempted 'conversion' to income), cover very truncated data sets, focus either on the rich – those subject to taxation – or the poor (McCants, 2007), and of course include single cities only. Incidentally, all examples but one used by Pareto in the formulation of his famous 'iron law' of income distribution come from various European tax data from the end of the 19th century (see Pareto, 1896). The data on Latin America, produced by various Spanish *Visitas*, which collected detailed information on population, age, land ownership and agricultural output, have been published in numerous volumes but not used for estimates of income distribution. (For Peru, books with detailed notes from *Visitas* for the years 1562, 1567 and 1604-05 have been published.)

中国社会科学院研究生院：研修之用

What is common to these sources is that they are in principle surveys of stocks (people and wealth) and require a huge effort of price imputation; first, to 'transform' a stock into a meaningful annual yield (income), then to convert produced quantities, expressed in local 'natural' units (such as Egyptian *modii* of wheat), into kilograms, and finally to convert all of these into monetary units. Then the researcher needs to resort to even more heroic assumptions to calculate other sources of income, from husbandry, vineyards, honeybee cultivation, fruits and plants, services provided by farmers, and not least, from manufacturing activities like pottery, glass or cloth-making, or provision of urban services from the shoe-maker to the teacher (for which, at least some wage data are generally available). Particularly vexing is the issue of measurement units, volumes or weights with often confusingly similar or identical names, which nevertheless imply different physical amounts from one region to another; or when money units are provided, the issue of silver or gold conversion between them. But such sources, however frustrating. can and do provide very useful evidence about ancient living standards and distribution of income.

The second contemporary evidence is provided by social tables. This is what William Petty termed 'political arithmetick'. They aim to describe the structure of a society by listing all salient social classes (or professions) and estimating their average incomes (per household, or less often per person). For modern economists, these sources are much easier to use because the classification into presumably socially important groupings and estimates of their money-equivalent incomes provide us with most of what we need to know for the derivation of income distribution. England was the pioneer in the production of social tables, beginning with the famous one of Gregory King for 1688 (which contains 33 social groups with their population sizes and average incomes), and continuing with Massie (1759) and Colquhoun (1801 – 3). (None of the social tables, or the results obtained from them, is without its critics: for a critique of King's social table, see Arkell, 2006; for Colquhoun, see Schwarz, 1979; for a critique of Lindert – Williamson's use of English social tables, see Feinstein, 1988.) Much more recent authors have produced similar social tables for a number of countries (see, for example, Morrisson and Snyder, 2000, for France in 1788, Bértola, Castelnovo, Reis and Willebald, 2006, for Brazil in 1872, van Zanden, 2003, for Java in 1880, Berry, 1990, for Peru in 1876). These new social tables are of course not contemporary sources but they were produced, using bits of dispersed primary or most often secondary sources, by economic historians who specialize in various eras and countries, and they represent our best guess at social structure at remote points in time. The work of Milanovic, Lindert and Williamson (2009; hereafter MLW), who made the first systematic attempt to measure

中国社会科学院研究生院：研修之用

and analyse preindustrial inequality, is largely based on such (contemporary and recent) social tables.
Back to top

Empirical evidence

To translate preindustrial inequality into modern economics, we must not only hold that preindustrial societies were largely monetized (and whatever was not monetized could be ultimately expressed in money), but also hold that their inequality can be meaningfully handled by Gini, Theil or any other currently used inequality measure. Otherwise we lack a common yardstick with which to compare past and present.
Using mostly social tables from 30 preindustrial societies, MLW calculated Gini coefficients. They found that the preindustrial Ginis range from the mid-20s to around 65, with a mean of 45 and standard deviation of 11. (Gini is the most commonly used measure of inequality, and ranges from a theoretical zero (everybody has the same income) to a theoretical maximum of 100 (everybody but one person has a zero income, and the richest person takes the entire income of the community).) This is almost the same as the range of Ginis in modern societies. In fact, the modern equivalents of the preindustrial societies included in MLW sample (such as Turkey for Byzantium, Syria for the Levant, today's United Kingdom for the 1688 England and Wales) have an average inequality of 40 Gini points with a standard deviation of 10. However to make such a simple comparison and leave it at that would be erroneous. Preindustrial and modern societies were very different, even when compared in the language of modern economics.
First, it is very likely that the income gradient (how income increases as we move from poorer to richer income classes) was much flatter in preindustrial that in modern economies (see MLW, 2009). Using Jan Pen's (1971) metaphor of dwarfs and giants, where people are visualized as marching in a 60-minute parade, from the poorest to the richest, with everyone's height reflecting their income, preindustrial societies can be seen as societies of dwarfs who would take some 40 to 45 minutes to file past. They contained large groups of people (most of the time, the vast majority of the population) living at, or just above, the subsistence minimum. Percentage differences in income among this vast mass of people were small. The income gradient was flat up to a very high point in income distribution. But then, and quickly, as we approach the very end of the parade, the gradient would suddenly increase, much more so than in modern societies. Thus, unlike a modern parade which would be characterized by a steady increase of the gradient, in preindustrial societies the middle was not much different from the bottom. There was a dearth of people whom we would (using modern terminology) identify with the middle class. (It

中国社会科学院研究生院：研修之用

is worth pointing out that this 'middle class' is not defined in terms of absolute income, or what we would consider today to be middle-class requirements, but entirely in terms of the period average income.) We can thus see why both of our preconceived notions - of generalized equality and drastic income disparity among the ancient - are true: they just refer to different parts of income distribution.

This difference in structure implies that the same calculated measures of inequality have different meanings. Ginis, as we have already indicated, were broadly in the same range then and now. But a Gini of 40, estimated independently for the Roman Empire by MLW (2009) and Scheidel and Friesen (2009), had an altogether different meaning from the same Gini in the contemporary United States. (The MLW estimate refers to the year 14 (at the death of Octavian), Scheidel's estimate to the mid-second century.) The Roman Empire's mean income was about twice the physiological subsistence level (s). If we require that all members of a society have at least the subsistence minimum - for otherwise the society will tend to shrink and disappear - then a very low level of mean income, regardless of how tiny the upper class is, limits the extent of measured inequality. Simply put, the extent of inequality is limited by the size of average income. That ceiling is more binding when a society is poor. To realize this, assume that society's mean income is just a fraction above s. If all but a tiny elite live merely on s, the elite cannot be extravagantly rich because total income is low, and Gini or Theil indexes, which take into account incomes differences between all individuals, cannot be very high either. This is the idea underlying the Inequality Possibility Frontier (IPF: see Figure 1), defined by MLW (2009) and Milanovic (2006).

Figure 1

The Inequality Possibility Frontier. *Source*: MLW(2009).

中国社会科学院研究生院：研修之用

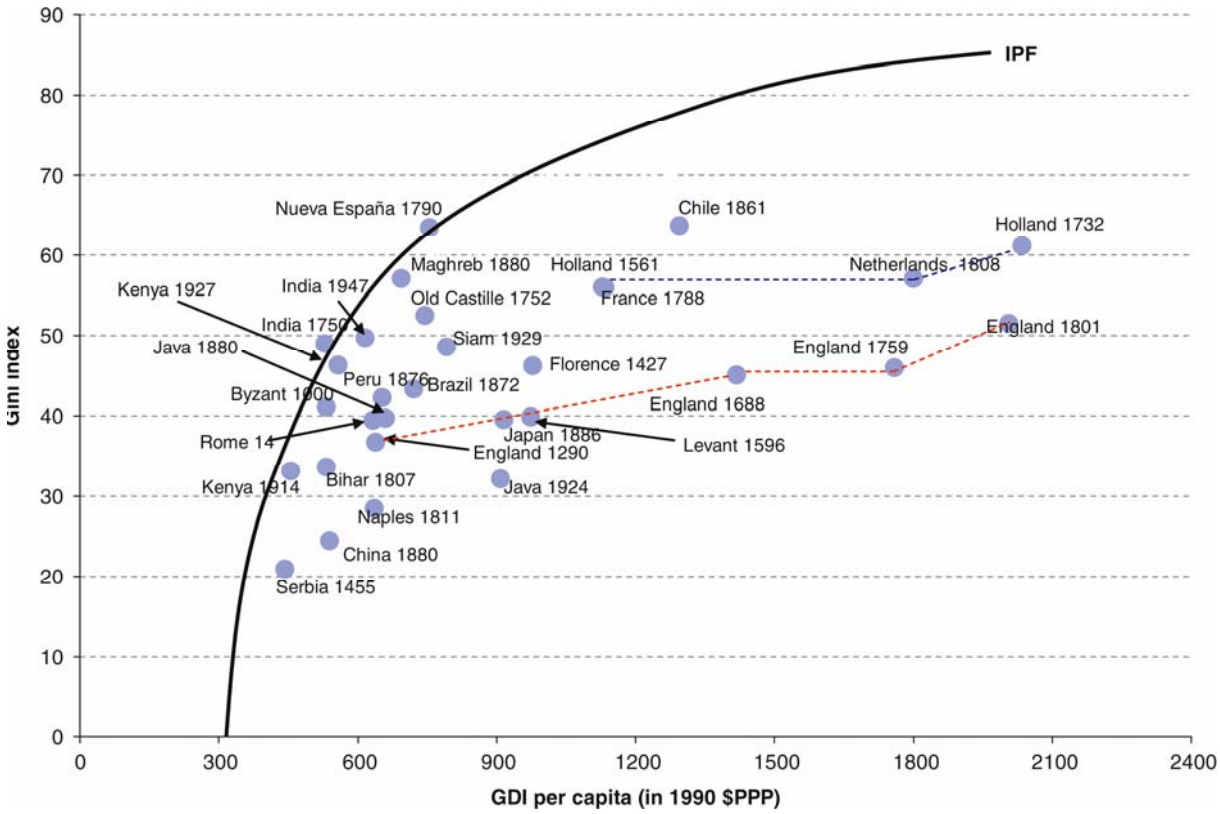

The frontier gives a maximum Gini (or Theil) coefficient which is compatible with a given level of mean income and maintenance of society as a going concern. The maximum Gini is equal to (a−1)/a where a=mean income divided by *s*, or the number of subsistence minima contained in the mean. As can be seen from the formula, the maximum feasible Gini rises with mean income (a), but at a decreasing rate. If average income is twice the subsistence (a=2), the maximum Gini will be 50. Thus, we see that the Roman inequality of 40 exhausted some 80 per cent of maximum feasible inequality. But for the modern-day United States, where the mean income stands at more than 100 *s*, the maximum Gini is 99. The actual inequality will have exhausted only 40 per cent of its maximum value. Hence, the social meaning of the same Gini is entirely different. To sustain high inequality, societies must be relatively rich.

We have left the issue of defining the subsistence minimum deliberately vague. Depending on whether we pitch this physiological (note: not social, not relative) minimum higher or lower, the IPF will move down or up, but the same logic will hold.

The difference in the income structure (income gradient) also shows why some other measures, like top-to-bottom ratio or top 1 per cent share, may not be very useful in the preindustrial context. They show the extent of the gap between the richest and the poorest, but they disregard the

中国社会科学院研究生院：研修之用

entire distribution in-between, which in the past has been much more equal than in today's societies.
IPF imposes a consistency check on our inequality calculations, a fact which is particularly useful for preindustrial societies where the evidence is scant. As illustrated in the figure, once we know the mean income of a society, and estimate its Gini, we know that this estimate must be within, or at the maximum on, the frontier. If it is not, there is something wrong with either the income or the inequality estimate, or the society is doomed to experience a dwindling population and ultimately extinction. It is not surprising that MLW found that all six cases of ancient societies with inequalities close to the frontier were colonies: India in 1750 and 1947, Kenya in 1914 and 1927, Nueva España (Mexico) in 1790, and Maghreb in 1880. Colonizers were clearly much less concerned about the welfare of the populations they ruled than, or did not have to fear them as much as, native rulers.
Back to top

Preindustrial inequality and modern debates

Empirical evidence on preindustrial inequality has a direct bearing on several contemporary debates. Evidence from the two most advanced economies at the time (England and Holland) paints a picture of increasing inequality from 16th century to the beginning of the Napoleonic wars. (The exception is Soltow, 1968, who found English inequality to have been flat throughout the 18th century.) Premodern growth seem to have exacerbated inequality even in the areas that were characterized by an already high inequality of wealth and income (such as the South Midlands in England, considered by Allen, 1992). Using social tables, Lindert (2000) and Lindert and Williamson (1982, 1983) document the increase of inequality in England between 1750 and 1801. All four observations available for England and Wales in the MLW database (1290 - from Campbell, 2007 - 1688, 1759 and 1801 - 3) show both mean income and inequality rising with time. Similarly, van Zanden (1995), and Soltow and van Zanden (1998) find that income inequality increased in Holland during its 'Golden age': between 1561 and 1732: the urban area Gini rose from 53 to 59, and the rural area Gini from 35 to 38. According to a pioneering study by Hoffman and colleagues (2002), 'real' European inequality between 1500 and the early 19th century increased even more because the prices of wage-goods, consumed by the poor, rose relative to the prices of 'luxuries'.
The upswing of the Kuznets curve seems to be strongly in evidence in all these cases. But what drove it? Was it a 'classical explanation' (as van Zanden, 1995, terms it), namely a shift in the functional distribution of income toward property owners (and their rising concentration) and away from labour - a mechanism that Marx would easily have recognized? (For

中国社会科学院研究生院：研修之用

Spain, Prados de la Escosura, 2008, uses functional distribution of income, and also finds a clear Kuznets upswing from 1850 to around 1914.) Or was it, as argued by Lindert and Williamson (1985) and Williamson (1982, 1985), caused by the 'wage-stretching' which continued well into the 19th century and involved labour-saving technological progress and increased pay-ratios for skilled labour in the presence of demographic pressure from mostly unskilled population? Education responded only very slowly, and the process continued for a couple of centuries until massive European emigration reversed it. The latter is a very neoclassical mechanism familiar to every economist working on poor or rich countries today. The focus is on the functioning of factor markets, not on the division of society into capitalists and workers.

If countries where the industrial revolution originated went through a period of sustained increase in inequality prior to the industrial revolution, does it shed some light on the relationship between higher inequality and the industrial take-off? A number of recent writings (most famously, Pomerantz, 2000; Frank, 1998; and more recently Wen, 2009; Shiue and Keller, 2007) have contrasted China and Western Europe in the 17th and 18th centuries, trying to understand why these two large areas that seemed in many respects similar (for instance in market integration, level of income, technological innovations) charted such different paths in the following three centuries. Does income distribution have to do something with it? Unfortunately, we do not yet have even the intimation of an answer because the historical data for China are not available. However a recent upsurge in archival research on Chinese sources might help throw some new light on this issue.

The work of Engerman and Sokoloff (1997) has profoundly affected our conception of the role of inequality in explaining the economic success of North America and relative decline of Latin America. But while there is little doubt that Latin America was more unequal (particularly in land ownership) that the North, recent historical evidence contrasting Western Europe and Latin America finds no perceptible difference in inequality between the two. Williamson (2009) thus wonders why Western Europe and Latin America have followed different growth trajectories. If the inequality explanation works for one set of regions (the two 'New Worlds'), why does it seem not to work for another (Europe and Latin America)? Moreover, it is not evident that Latin America was 'always' unequal. Prados de la Escosura (2007) and Bértola et al. (2009) argue that strong expansion of inequality occurred during the previous round of globalization (1870 – 1920). Prados de la Escosura (2007, p. 298) sees the explanation as consistent with the factor-price equalization theorem: opening up Latin America to trade raised land rents, and since land was unequally distributed, increased the concentration of incomes. The data prior to around 1870 are not available (although some estimates for 1870

中国社会科学院研究生院：研修之用

show inequality in the Southern Cone countries to be at the same level as in Spain: Prados de la Escosura, 2008, Figure 8, p. 307), but we could wonder whether our 'acquired idea' of an always high inequality in Latin America is not mistaken - or perhaps it was not inequality, but the inequality extraction ratio that was high. Recasting the issue in this way suggests that the Latin American problem was a low level of income rather than a high Gini.
Back to top

Conclusion

Studying inequality in its historical context, an area which will doubtlessly loom larger in economics as the search to uncover our economic past progresses, is important not only because it helps us learn about history but because it helps us understand today's economic problems. Actually, as every historian and politician knows, studying the past is about the future.
Back to top

Acknowledgements

I am grateful to Mihail Arandarenko, Peter Lindert, Leandro Prados de la Escosura and Jeffrey Williamson for excellent comments.
Back to top

See Also

- economic history
- inequality (international evidence)
- inequality (measurement)
- standards of living (historical trends)
- wage inequality, changes in

Back to top

Bibliography

Allen, R.C. 2007. How prosperous were the Romans? Evidence from Diocletian's Price Edict (ad 301). University of Oxford, Department of Economics, Discussion Paper Series 363, October.

Allen, R.C. 1992. *Enclosures and the Yeoman: Agricultural Development of the South Midlands, 1450–1850*. Oxford: Oxford University Press.

中国社会科学院研究生院：研修之用

Allen, R.C. 2003. Progress and poverty in early modern Europe. *Economic History Review* 56, 403 – 43.

Arkell, T. 2006. Illuminations and distortions: Gregory King's scheme calculated for the year 1688 and the social structure of later Stuart England. *Economic History Review* 59, 32 – 60.

Berry, A. 1990. International trade, government and income distribution in Peru since 1870. *Latin American Research Review* 25 (2), 31 – 59.

Bértola, L. , Castelnovo, C. , Reis, E. and Willebald, H. 2006. Income distribution in Brazil, 1839 – 1939. Paper presented at Session 116 of "A Global History of Income Distribution in the Long 20th Century," *XIV International Economic History Congress*, Helsinki-Finland 21-25 August.

Bértola, L. , Castelnovo, C. , Rodriguez, J. and Willeband, H. 2009. Income distribution in Latin American Southern Cone countries during the first globalization boom, ca. 1870 – 1920. *International Journal of Comparative Sociology*, forthcoming.

Campbell, B. 2007. Benchmarking medieval economic development: England, Wales, Scotland and Ireland c. 1209. *Economic History Review* 60, 1 – 50.

Clark, G. 2005. The condition of the working class in England, 1209 – 2004. *Journal of Political Economy* 115, 1307 – 1340.

Cosgel, M. 2004. Ottoman tax registers. *Historical Methods* 37 (2), Spring, 87 – 100.

Cosgel, M. 2006. Taxes, efficiency, and redistribution: discriminatory taxation of villages in Ottoman Palestine, Southern Syria, and Transjordan in the sixteenth century. *Explorations in Economic History* 43, 332 – 56.

Engerman, S. and Sokoloff, K. 1997. Factor endowments, institutions and differential paths of growth among New World economies. In *How Latin America Fell Behind: Essays on the Economic Histories of Brazil and Mexico, 1800 – 1914*, ed. S. Hager . Stanford, Calif.: Stanford University Press, pp. 260 – 304.

Feinstein, C. 1988. The rise and fall of the Williamson curve. *Journal of Economic History* 48, 699 – 729.

Finley, M. 1985. *The Ancient Economy*, 2nd edn. London: Penguin.

中国社会科学院研究生院：研修之用

Frank, A.G. 1998. *Re-Orient: Global Economy in the Asian Age*. Berkeley: University of California Press.

Goldsmith, R.W. 1984. An estimate of the size and structure of the national product of the early Roman Empire. *Review of Income and Wealth* 30, 263–88.

Herlihy, D. and Klapisch-Zuber, C. 1985. *Tuscans and their Families*. New Haven, Conn.: Yale University Press.

Herlihy, D., Klapisch-Zuber, C., Litchfield, R.B. and Molho, A. *The Online Catasto*, available at: http://www.stg.brown.edu/projects/catasto/overview.html, accessed January 2008. (a searchable online database of tax information for the city of Florence in 1427–29, based on D. Herlihy and C. Klapisch-Zuber, principal investigators, Census and property survey of Florentine dominions in the province of Tuscany, 1427–1480).

Hoffman, P.T., Jacks, D., Levin, P.A. and Lindert, P.H. 2002. Real inequality in Europe since 1500. *Journal of Economic History* 62, 322–55.

Kuznets, S. 1955. Economic growth and income inequality, presidential address to the 67th meeting of the American Economic Association, Michigan., December 1954; published in *American Economic Review* 45 (1955) and *Economic Growth and Structure: Selected Essays*, New Delhi: Oxford and IBH, 1965.

Laiou, A. 2002. *The Economic History of Byzantium: from the Seventh through the Fifteenth Century*. Washington DC: Dumbarton Oaks.

Lindert, P.H. 2000. Three centuries of inequality in Britain and the United Stares. In *Handbook of Income Distribution*, ed. A. Atkinson and F. Bourguignon. Amsterdam, North Holland: Elsevier.

Lindert, P.H. and Williamson, J.G. 1982. Revising England's social tables, 1688–1812. *Explorations in Economic History* 19, 385–408.

Lindert, P.H. and Williamson, J.G. 1983. Reinterpreting Britain's social tables, 1688–1913. *Explorations in Economic History* 20, 94–109.

Lindert, P.H. and Williamson, J.G. 1985. Growth, equality and history. *Explorations in Economic History* 22, 341–77.

Malanima, P. 2006. Pre-modern equality: income distribution in the Kingdom of Naples (1811), paper presented at 14th International Congress

中国社会科学院研究生院：研修之用

of Economic History, August 2006, Helsinki. Available at: http://www.helsinki.fi/iehc2006/papers3/Malanima.pdf.

Mayhew, N.J. 1995. Modeling medieval monetization. In *A Commercialising Economy: England 1086 to c. 1300*, ed. R.N. Britnell and B.M.S. Campbell. Manchester: Manchester University Press, pp. 55–77.

McCants, A. 2007. Inequality among the poor of eighteenth century Amsterdam. *Explorations in Economic History* 44, 1–21.

Milanovic, B. 2006. An estimate of average income and inequality in Byzantium around year 1000. *Review of Income and Wealth* 52 (3), 449–470.

Milanovic, B., Lindert, P.H. and Williamson, J.G. 2009. Preindustrial inequality. Unpublished ms, available at: http://econ.worldbank.org/projects/inequality. Previous version published as Measuring ancient inequality, National Bureau of Economic Research Working Paper No. 13550.

Morrisson, C. and Snyder, W. 2000. The income inequality of France in historical perspective. *European Review of Economic History* 4, 59–83.

Pareto, V. 1896. La courbe de la repartition de la richesse. Université de Lausanne. Republished as On the distribution of wealth and income, *Rivista di Politica Economica*, 645–660, 1997. (The same issue of *Rivista di Politica Economica* contains English translations of five other articles by Pareto on the same topic.)

Pen, J. 1971. *Income Distribution: Facts, Theories, Policies*. New York/Washington DC: Praeger.

Polanyi, K. 1944. *The Great Transformation*. Boston, Mass.: Beacon Press.

Pomerantz, K. 2000. *The Great Divergence: China, Europe, and the Making of the Modern World Economy*. Princeton, N.J.: Princeton University Press.

Prados de la Escosura, L. 2007. Inequality and poverty in Latin America: a long-run exploration. In *The New Comparative Economic History: Essays in Honor of Jeffrey G. Williamson*, ed. T.J. Hutton, K.H. O'Rourke and A.M. Taylor. Boston, Mass.: MIT Press, pp. 291–315.

Prados de la Escosura, L. 2008. Inequality, poverty, and the Kuznets curve in Spain, 1850–2000. *European Review of Economic History* 12, 287–324.

中国社会科学院研究生院：研修之用

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Rostovtzeff, M. 1926. *The Social and Economic History of the Roman Empire*. Oxford: Oxford University Press, 1957.

Scheidel, W. and Friesen, S.J. 2009. The size of the economy and the distribution of income in the Roman Empire. Princeton/Stanford Working Papers in Classics, January.

Schiavone, A. 1995. *The End of the Past: Ancient Rome and the Modern West*. Cambridge, Mass.: Harvard University Press, 2000.

Schwarz, L.D. 1979. Income distribution and social structure in London in the late eighteenth century. *Economic History Review* 32, 250–9.

Shiue, C. and Keller, W. 2007. Markets in China and Europe on the eve of the industrial revolution. *American Economic Review*, September.

Soltow, L. 1968. Long-run changes in British income inequality. *Economic History Review* 21, 17–29.

Soltow, L. 1989. Income and wealth inequality in Amsterdam, 1585–1805. *Economisch-en-Social Historisch* 58, 72–95.

Soltow, L. and van Zanden, J.-L. 1998. *Income and Wealth Inequality in the Netherlands, 16th–20th Century*. Amsterdam: Het Spinhuis.

Sussman, N. 2005. Income inequality in Paris at the heyday of the commercial revolution. Hebrew University, Jerusalem. Available at: http://economics.huji.ac.il/facultye/sussman/sussman_gobaleconomies.pdf.

Tocqueville, A. de. 1835. *Memoir on Pauperism*. Chicago: Ivan R. Dee, 1997.

Van Zanden, J.L. 1995. Tracing the beginning of the Kuznets curve: western Europe during the early modern period. *Economic History Review* 48, 643–64.

Van Zanden, J.L. 2003. Rich and Poor before the Industrial Revolution: A Comparison between Java and the Netherlands at the Beginning of the 19th Century. *Explorations in Economic History* 40 (1), 1–23.

Walbank, F.W. 1946. *The Decline of the Roman Empire in the West*. London: Cobbett Press.

中国社会科学院研究生院：研修之用

Wen, J.G. 2009. Why was China trapped in an agrarian society - an economic-geographical approach to the Needham puzzle. *New History*, forthcoming.

Williamson, J.G. 1980. Earnings inequality in nineteenth century England. *Journal of Economic History* 40, 457–76.

Williamson, J.G. 1982. The Structure of pay in Britain, 1710–1911. *Research in Economic History* 7, 1–54.

Williamson, J.G. 1985. *Did British Capitalism Breed Inequality?* Boston, Mass.: Allen and Unwin.

Williamson, J.G. 2009. History without evidence: Latin American inequality since 1491. Mimeo.

Back to top

How to cite this article

Milanovic, Branko. "preindustrial inequality." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 13 January 2010 <http://www.dictionaryofeconomics.com/article?id=pde2009_I000307>
doi:10.1057/9780230226203.1918

Download Citation:

中国社会科学院研究生院：研修之用

reparations

Kim Oosterlinck
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume
Back to top

Abstract

Reparations for damage caused, paid by the loser following wars, have been known since Antiquity, although much of the literature focuses on the First World War. There has been much debate, both politically and among economists, on the appropriate basis on which to pay.
Back to top

Keywords

reparations; transfer problem; war
Back to top

Article

From time immemorial, wars have led to international financial transfers. During Antiquity, the practice of imposing a payment of a tribute on the vanquished was already common practice (Livy V, pp. 48–9). War indemnities were common during the Middle-Ages. From a legal point of view, few authors questioned the practice, and victory was seen as a sufficient motivation to extract tribute. With the development of theories on international law in the 16th and 17th centuries, the idea of ‘repairing’ the torts done by the war came to be discussed.
During the 19th century various legal arguments were invoked to justify war indemnities. The amounts to be paid could, for example, be meant to cover war expenditures incurred by the victor, to make good losses as a result of the war, or to guarantee the victor’s future safety. Most treaties remained silent or at best vague regarding the exact computation of the amounts due, and the final figures were often the result of long negotiations between the parties. To secure payments, occupation of part of the defeated country was common. France had to cope with a German occupation after the Second Treaty of Paris (1815) and as a consequence of the Treaty of Frankfurt (1871) ending the Franco-Prussian war. For both wars, France would eventually pay in full (White, 2001). The amounts

中国社会科学院研究生院：研修之用

(close to FF1.9 billion for the Napoleonic war and equal to FF5 billion for the Franco-Prussian one) were at the time viewed as exorbitant, and produced heated debates and controversies in France. Devising the means to pay for the reparations also proved to be an arduous task, which has been analysed in detail by economic historians. Loans would eventually provide the bulk of the payment, even though, in 1871, economists such as Louis Wolowsky and bankers such as Henri Germain suggested instead creating a progressive income tax.

The importance of war indemnities before the First World War should not be downplayed. For example, the 230 million silver taels extracted from China after the Sino-Japanese War played a preeminent role in Japan's decision to join the gold standard (Metzler, 2006, p. 3). The First World War would however act as a real turning point regarding the literature on reparations. Several elements contributed to this change. First, the intensity of the conflict and its huge material and human costs rendered the question of reparations of foremost importance. Second, the Treaty of Versailles (art. 232) directly referred to the notion of reparations. Since the payments were only meant to 'repair' the damages done, economists would attempt to evaluate the exact size of this damage. This would raise two questions: which were the losses for which Germany was to pay, and how could one evaluate these? Third, the phrasing of the Treaty of Versailles opened the door to further discussions regarding the amounts due. Indeed, it left the determination of the amounts to be paid to the Reparation Commission, an ad-hoc Inter-Allied Commission, which had to provide an estimate before 1 May 1921. Fourth, the Treaty (art. 232 and art. 234) alluded to the fact that Germany's capacity to pay had to be taken into account, which meant this capacity had somehow to be estimated.

Eventually, the World War reparations would certainly not have been so central in the literature had John Maynard Keynes not attacked them vigorously in his *Economic Consequences of the Peace* (1919). Keynes's subsequent fame led the broader public to accept his point of view, even though it was firmly contested when first expressed (Bainville, 1920; Ohlin, 1929).

Keynes attacked the treaty on several fronts. First, he suggested a limited interpretation regarding the scope of the damages. During the discussions related to the amounts to be paid, France and Great Britain at first pleaded that Germany should be made to pay all the costs of the war. For the French finance minister, Louis-Lucien Klotz, it was clear that Germany would pay ('*le boche paiera*'). Eventually however, the position defended by the United States and Belgium would prevail, and the treaty (art 232) would state that reparations were to cover 'the damage done to the civilian population of the Allied and Associated Powers and to their property'. Keynes defended a narrow interpretation of this article, and considered that only direct damages caused by the war should

中国社会科学院研究生院：研修之用

be taken into account. To Keynes's consternation, costs such as pensions and separation allowances were included in the total. Keynes found the terms so harsh that he would qualify the Treaty of Versailles as a 'Carthaginian Peace', referring to the extremely hard conditions imposed upon defeated Carthage by Rome after the Second Punic War (218 to 201 bc) (Livy XXX, 16 and 37).

Keynes then proceeded to criticize the first tentative evaluations of the amounts that had been put forward during the Paris Peace Conference. According to him, French estimates regarding reconstruction costs were much too high. Once converted into pounds, they ranged from £2.6 to 5.63 billion, whereas Keynes's own computation led him to suggest they should be rounded up to £2 billion. Mantoux (1946), probably one of the fiercest critics of Keynes's estimates, showed that Keynes erroneously used the pre-war exchange rate to convert the French figures into pounds, therefore largely inflating these once converted. He further suggested that the amounts really spent for the reconstruction were actually quite close to the French estimates.

Eventually, the Reparation Commission concluded on 5 May 1921 that Germany would have to pay a total of 132 billion marks (£6.6 billion), and cover Belgium's war debt, worth the equivalent of 5.6 billion marks. In practice, Germany would pay interest and amortization on two bonds series: the A one representing 12 billion marks and the B one worth 38 billion. The remaining 82 billion would be issued in a third series of bonds (the C one) only if Germany ever became prosperous enough to service them on top of the series A and B. As a matter of fact, most Allied politicians never believed this last issue would ever take place. Furthermore Germany was bound to pay a fixed annuity of 2 billion marks and a variable amount worth 26 per cent of its exports.

According to Keynes, Germany did not have the capacity to pay these amounts. He viewed the requested sums as so high that they would 'reduce Germany to servitude for a generation, degrade the lives of million of human beings and deprive a whole nation of happiness' (Keynes, 1929, p. 17). In his eyes, two different issues were to be taken into account when one considered Germany's capacity to pay: there was both a budgetary problem linked to German capacity to extract the required sums from its citizens and a transfer problem linked to the conversion of the obtained marks into hard currency.

According to Keynes the transfer problem was as important an issue as the budgetary problem (Keynes, 1929). The macroeconomic impact of international transfers had already been addressed before the war. The traditional view was that the transfer-paying country would suffer deterioration in its terms of trade. In the case of the First World War, Keynes believed that creditor countries would not allow Germany to run a huge trade surplus and would thus force Germany to let its terms of trade

中国社会科学院研究生院：研修之用

deteriorate. This would in turn reduce Germany's real income and in a sense add a further burden to the defeated country. Ohlin (1929) attacked this view on the grounds that income effects could induce the recipient countries to buy more German goods and thus improve its terms of trade. The transfer problem in the framework of reparations has since then led to a large literature, and has recently been analysed using recent macroeconomics approaches (Morrison, 1992; White, 2001; Devereux and Smith, 2007 for example).

This transfer problem would partially be addressed in the framework of the Dawes plan. At the end of 1921, Germany started negotiations to get a moratorium on its debts. In exchange for this moratorium, France requested productive pledges, which Great Britain opposed. On 11 January 1923 French and Belgian troops entered the Ruhr to seize part of Germany's mines and industrial production. German passive resistance to the occupation, its dramatic economic situation characterized by hyperinflation, and geopolitical changes all led to the resumption of negotiations. Eventually, the plan drafted by the former US Director of the Bureau of the Budget, Charles G. Dawes, was agreed upon in August 1924. The Ruhr was to be evacuated, and it was hoped that the transfer problem would be resolved by accepting that Germany pay the annuities either in gold marks or in its German monetary equivalent on a special account of the Reichsbank. The Dawes plan was meant to be a provisional settlement but remained in application for five years. It also opened the door to US investment in Germany, which quickly boomed and rendered the payment of reparations much easier for Germany (Klug, 1990).

Reparations represented but part of the international debts stemming from the war. The sums lent by the United States to its European allies were huge, and had been a major point of contention since the end of the war. Allied European countries wanted them suppressed or reduced. In 1928, they pleaded to link their reimbursement to the reparations. This would lead to a revision of the Dawes plan by a team of experts led by Owen D. Young. The Young report was finalized on 7 June 1929 and adopted shortly afterwards. The amounts due by Germany were reduced to 121 billion marks and were to be administered by the Bank for International Settlements. The Young plan was short lived. The great depression forced Germany to ask for a moratorium, which lead to negotiations and the Lausanne agreements which stated that Germany would pay 3 billion gold marks as final settlement for the reparations. The Lausanne agreements were never formally signed by all parties, but payments nonetheless stopped after 1932. Depending on the sources, the estimation of the amounts paid differs greatly, ranging from close to 21 billion marks by the Reparation Commission to close to 68 billion according to the pre-Second World War German government.

中国社会科学院研究生院：研修之用

In the 1970s the Keynesian view became the subject of more and more critiques by historians and economists who had access to new archival material. As mentioned by Schuker (1988), however, 'not only did the Reich entirely avoid paying net reparations to its wartime opponents; it actually extracted the equivalent of reparations from the Allied powers, and principally from the United States'.

In view of the preeminence of the post-First World War German reparations in the literature, many economists still base their views on reparations on this single example, and conclude that reparations are in essence an inefficient mechanism. Contrasting with the German case, history however provides many examples when reparations were paid in full. Table 1 provides a comparison of war reparations for four historical episodes. Even though Germany's indemnities were higher in percentage of one year's GDP than the ones imposed on defeated France in 1819 and 1871, they still remained far below the indemnities paid by occupied France during the Second World War. Noteworthy, amounts were paid in full for all the episodes mentioned except the First World War.

A comparison of war reparations

	Indemnities (billions)	Percentage of one year's GDP	Share of debt service to GDP
France 1815–19	FF 1.65 to 1.95	18 to 21	1.2 to 1.4
France 1871	FF 5.0	25	0.7
Germany 1923–31	DM 50	83	2.5
Vichy 1940–4	FF 479 (633)	111 (147)	2.6 (3.4)

Source: White (2001), Klug (1990) and Occhino, Oosterlinck and White (2008). Note: the numbers in parentheses include occupation payments and looting.

Even though the experience of the First World War appeared as negative to most, the Allied countries, anticipating victory, started discussing reparations even before the end of the Second World War. In the framework of the Yalta Conference, which took place on 11 February 1945, the Allied countries agreed to make Germany pay for the 'losses caused by her to the Allied nations in the course of the war'. The protocol further suggested that reparations in kind should be extracted from Germany (by requesting equipment, machine tools, ships, rolling stock, German investments abroad, and shares of industrial, transport and other enterprises in Germany, or by asking for annual deliveries of goods, or even by using German labour). Exact details were to be determined by a Commission of Damage to be seated in Moscow.

Following Germany's surrender, the Commission of Damage met to estimate the reparations. Heated debates, fuelled by geopolitical considerations,

中国社会科学院研究生院：研修之用

animated the discussions. Since the Soviet Union had suffered dramatically from the war, generous reparations would have allowed it to recover quickly, a prospect that did not please countries from the Western bloc. Despite the diverging views, the Commission managed to present a series of principles which were agreed upon at the Potsdam Conference on 2 August 1945. Most importantly, dismantling of German factories was to represent one of the main sources of reparations. The Cold War and Germany's partition would soon lead to a separation of responsibilities regarding the reparations: West Germany would pay reparations to all countries but Poland and the Soviet Union, which were to be covered by East Germany. More than 650 factories (or part of factories) were dismantled from West Germany and transferred for an estimated value of US$130 million, representing 25 per cent of the payment eventually made by West Germany. In East Germany, dismantling was soon abandoned in favor of the transfer of part of its industrial production. Estimates of the magnitude of the transfers are tentative at best, but tend to indicate that they were of large magnitude.

During the Cold War few conflicts led to reparations, most probably because each belligerent could most of the time count on the support of one of the superpowers and therefore avoid a painful settlement (d'Argent, 2002). The practice did not however disappear, and as a consequence of the first Gulf War, a UN Security Council resolution forced Iraq to pay reparations to the victims of its aggression. The first Gulf War case would, on top of the 'traditional' questions, raise an additional one. Since payments were expected to come from oil sales, we could wonder, as does Morrision (1992), 'how to proceed when the guilty nation has the ability to affect world prices - even to the point where it may be able to reduce its reparations burden and inflict real income losses on those seeking compensation'.

Back to top

See Also

- war and economics
- World Wars, economics of

Back to top

Bibliography

Bainville, J. 1920. *Les Conséquences Politiques de la Paix* . Paris: Nouvelle librairie nationale.

中国社会科学院研究生院：研修之用

D'Argent, P. 2002. *Les réparations de guerre en droit international public. La responsabilité internationale des Etats à l'épreuve de la guerre*. Bibliothèque de la Faculté de Droit de Louvain, 36, Bruylant, Brussels.

Devereux, M.B. and Smith, G.W. 2007. Transfer problem dynamics: macroeconomics of the Franco-Prussian war indemnity. *Journal of Monetary Economics* 54, 2375–98.

Keynes, J.M. 1919. *The Economic Consequences of the Peace*. London: Macmillan.

Keynes, J.M. 1922. *A Revision of the Treaty*. London: Macmillan.

Keynes, J.M. 1929. The German transfer problem. *Economic Journal* 39, 1–7.

Klug, A. 1990. The theory and practice of reparations and American loans to Germany, 1925-1929. Princeton Working Papers in International Economics, G-90-03, International Finance Section.

Livy. *Ab urbe condita* (The Early History of Rome, books I–V, and The History of Rome from its Foundation, books XXI–XXX: The War with Hannibal), London; Penguin Classics, 2002 and 1976.

Mantoux, E. 1946. *The Carthaginian Peace or The Economic Consequences of Mr. Keynes*. London: Oxford University Press.

Metzler, M. 2006. *Lever of Empire: The International Gold Standard and the Crisis of Liberalism in Prewar Japan*. Berkeley and Los Angeles, Calif.: University of California Press.

Morrison, R.J. 1992. Gulf war reparations: Iraq, OPEC, and the transfer problem. *American Journal of Economics and Sociology* 51, 385–99.

Occhino, F., Oosterlinck, K. and White, E. 2008. How much can a victor force the vanquished to pay? *Journal of Economic History* 68, 1–45.

Ohlin, B. 1929. The reparation problem: a discussion. *Economic Journal* 39, 172–82.

Schuker, S.A. 1988. 'American reparations' to Germany, 1919–33: implications for the third-world debt crisis. Princeton Studies in International Finance no. 61.

White, E.N. 2001. Making the French pay: the cost and consequences of the Napoleonic reparations. *European Review of Economic History* 5, 337–65.

中国社会科学院研究生院：研修之用

Back to top

How to cite this article

Oosterlinck, Kim. "reparations." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 13 January 2010 <http://www.dictionaryofeconomics.com/article?id=pde2009_R000279> doi:10.1057/9780230226203.1920

Download Citation:

中国社会科学院研究生院：研修之用

Wal-Mart, economics of

Emek Basker
From The New Palgrave Dictionary of Economics, Online Edition, 2009
Edited by Steven N. Durlauf and Lawrence E. Blume
Back to top

Abstract

Wal-Mart is the largest retailer in the world, with stores in 16 countries (including the United States) and annual revenues exceeding $400 billion. Wal-Mart owes its success primarily to its early and persistent investments in technology. Technology has allowed Wal-Mart not only to grow - adding stores in new markets and adding a broad range of products over the past half century - but also to cut its costs, making it a formidable competitor in almost every retail sector. Wal-Mart's competitive effect lowers prices in local markets, in the process driving some of its competitors to contract or shut down.
Back to top

Keywords

big box; chain store; general merchandise; one-stop shopping; price competition; retail; Wal-Mart
Back to top

Article

Wal-Mart is a discount general-merchandise retail chain based in the United States, and is the largest retailer in the world. As of 2006, Wal-Mart accounted for 28 per cent of Playtex's sales, 25 per cent of Clorox's, 21 per cent of Revlon's, 13 per cent of Kimberly Clark's, and 17 per cent of Kellogg's (Weinswig and Tang, 2006).
Started in 1962 by Sam and Bud Walton as a five-and-dime store in Rogers, Arkansas, the chain had 18 stores when it incorporated in 1970, more than 1,700 stores by the end of 1990, and 7,873 stores as of 31 January 2009: 891 US Wal-Mart 'Discount Stores' (Wal-Mart's traditional format, selling apparel, housewares, toys, electronics, prescription drugs, and more), 2,612 US Supercenters (which include a full grocery store in addition to general merchandise), 602 Sam's Clubs (membership clubs selling a wide range of products), 153 Neighborhood Markets (smaller

中国社会科学院研究生院：研修之用

formats which sell mostly groceries) and 3,615 international stores in 15 countries: Argentina, Brazil, Canada, Chile, China, Costa Rica, El Salvador, Guatemala, Honduras, India, Japan, Mexico, Nicaragua, Puerto Rico and the United Kingdom. From 1 February 2008 to 31 January 2009, Wal-Mart's sales exceeded $400 billion; the US divisions alone had revenues exceeding $255 billion, or nearly 1.8 per cent of US GDP. Figure 1 shows the number of stores, by division, at the end of each calendar year (1962 – 2008).

Figure 1

Wal-Mart stores, 1962 – 2008

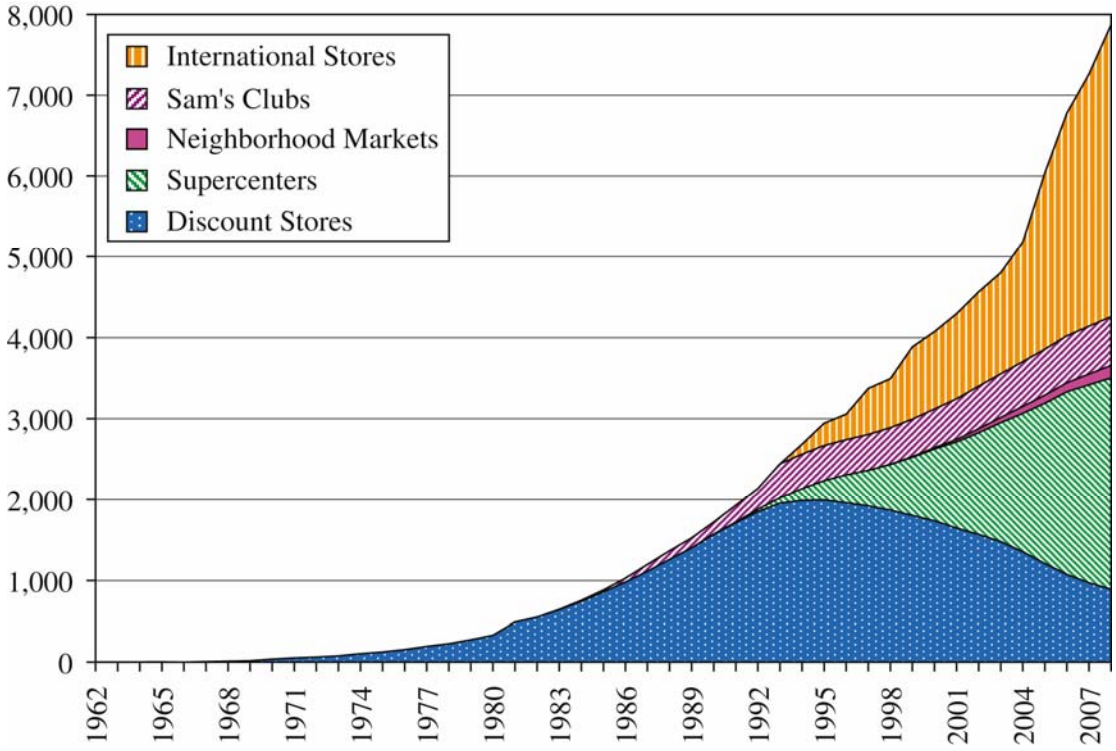

Wal-Mart's main competitors are Target and Kmart, the two largest US discount general merchandisers, but because it sells such a wide range of products it effectively competes with supermarkets, toy stores, electronics stores, apparel stores and much more. Over 99 per cent of Wal-Mart stores have pharmacies; and most Wal-Mart stores, including 'discount stores', carry at least some groceries.

Not coincidentally, the first Kmart and Target also date to the early 1960s; Wal-Mart's rapid expansion has been emblematic of the widespread rise of chain retailing in the 20th century. Chains accounted for less than 30 per cent of all retail sales in 1948 and over 60 per cent of sales by 2002 (Basker, Klimek and Van, 2008). The growth of retail chains owes much to technology, which has made it possible for a single firm to manage complex supplier relationships, personnel, logistics and distribution.

Popular opinion about Wal-Mart is mixed. The criticisms of Wal-Mart tend to vacillate between two contradictory views: that it competes too

中国社会科学院研究生院：研修之用

aggressively to maximize its profit, for instance by placing extreme conditions on suppliers, aggressively fighting unions, or charging lower prices in markets with heavier competition (see, for example, Norman, 2004); and that it uses its market power in non-profit-maximizing ways, such as to reduce access to birth control or to censor music (Bianco, 2006, pp. 248–50).

Despite widespread claims of negative employment effects and low wages, research has found very small, and generally positive, effects of Wal-Mart on employment. On the other hand, Wal-Mart has either settled or lost a large number of legal challenges by current and former employees alleging they were required to work off the clock, denied breaks, or denied overtime pay.

Back to top

Wal-Mart's advantages: technology and scale

Wal-Mart invested in computers early and has continued to make large investments in technology, which accounts for much of its success. In 1969, Wal-Mart installed a computer in its first distribution centre, and it later connected all its stores and distribution centres, along with company headquarters, to a computer network. In the 1980s, Wal-Mart was at the forefront of bar-code adoption (Vance and Scott, 1994), just as it is currently a leader in radio frequency identification (RFID) technology, which works by embedding radio transmitters in individual products or cases of products, and reduces much of the cost involved in tracking shipments, inventory and sales. Its Retail Link software, introduced in 1990, connects its stores, distribution centres and suppliers, providing detailed up-to-the-minute inventory data to its suppliers.

These investments have increased Wal-Mart's productivity and made it a formidable competitor. Basker (2007) offers a back-of-the-envelope calculation of Wal-Mart's labour productivity growth compared with the productivity growth in the rest of the retail sector from 1982 to 2002. Wal-Mart's sales per employee grew, in real terms, by 55 per cent over this period; other general merchandisers increased their sales per worker by only 18.5 per cent, and productivity in the rest of the retail sector grew by only 9 per cent.

Economies of scale also play an important role in Wal-Mart's success and its ability to charge low prices to consumers. There is anecdotal evidence that Wal-Mart asks for, and receives, price discounts from suppliers (see Fishman, 2006). In addition, the benefit that Wal-Mart is able to squeeze from its investments in technology owes much to its size. Tracking purchases alone would not have enabled Wal-Mart to discover that Strawberry Pop Tarts are a popular item consumers stock before hurricanes

中国社会科学院研究生院：研修之用

(Leonard, 2005). A large volume of purchases – from many stores, over a long period of time, in hurricane-prone areas – was also essential. Finally, there is some evidence that Wal-Mart imports disproportionately to its size, which also lowers its costs compared with its smaller competitors. But Wal-Mart was not always a major importer. From 1985 to 1992, Wal-Mart gained popular acclaim with its 'Buy American' campaign. This campaign ended abruptly in late 1992 after *Dateline NBC* aired a segment accusing Wal-Mart of producing private-label goods in Bangladesh, smuggling textiles into the United States in excess of quotas, and placing imported clothes on racks marked 'Made in the USA' (Gladstone, 1992). By 2004, Wal-Mart's imports from China were valued at $18 billion, or 15.4 per cent of US imports of consumer goods from China, more than twice Wal-Mart's share of retail sales (Basker and Van, 2008).

Back to top

Local effects

An immediate effect of Wal-Mart's entry into a local market is increased price competition. Estimates of Wal-Mart's food prices range from 10 per cent below the competition (Basker and Noel, 2009) to 30 per cent below the competition (Hausman and Leibtag, 2007b). Basker and Noel (2009) estimate that competing supermarkets and grocery stores cut their prices by about 1 per cent on average when a Wal-Mart Supercenter opens in town. Hausman and Leibtag (2007a) calculate that prices paid by consumers fall by 3 per cent on average, accounting for product and outlet substitution. An earlier study focusing on prices of drugstore products such as shampoo and toothpaste found that these also fall (Basker, 2005b).

Increased price competition reduces profits at some incumbent stores and may cause them to contract or exit. On average, almost as many people lose their jobs at other retail establishments as are hired by a new Wal-Mart store. Using publicly available data on Wal-Mart stores' opening dates to 1999, Basker (2005a) estimated that the number of retail jobs in a county increases by 100 the year Wal-Mart opens a new store (relative to what would have happened had Wal-Mart stayed out of the county), and by 50 after five years. Since the average Wal-Mart store over the period of the study employed about 250 workers, this estimate implies that approximately 200 workers at other stores lose their jobs. In addition, the number of wholesale jobs declines by about 30 in the long run, reflecting the fact that Wal-Mart is vertically integrated: unlike the merchants it replaces, Wal-Mart does not rely on local wholesalers. Drewianka and Johnson (2009) find somewhat larger positive effects on employment in general merchandising. In contrast, Neumark, Zhang and Ciccarella (2008) estimate net job losses of 150 workers per Wal-Mart

中国社会科学院研究生院：研修之用

store (implying that 400 workers lost their jobs, on average, as a result of a new Wal-Mart store hiring 250 workers).
Competitors whose profits fall sufficiently eventually shut down. Basker (2005a) estimates a net closing of about five stores in a county as a result of Wal-Mart's opening. Jia (2008) finds a net reduction of two or three small general-merchandise stores, such as dollar and variety stores, if Wal-Mart or Kmart enter a market, compared with if either of these large retailers stays out. Drewianka and Johnson (2009) find little effect on the number of competitors.
The wide range of estimates of Wal-Mart's effect on prices, employment and competitors reflects a fundamental problem of identification: that is, disentangling cause and effect. Basker (2005a, 2005b) addresses this problem by exploiting the time lag between Wal-Mart's initial decision to open a store and its actual opening. Drewianka and Johnson (2009) instead control for pre-existing trends in the outcome variables of interest (employment, wages and so on). Finally, Neumark, Zhang and Ciccarella (2008) use Wal-Mart's geographic pattern of expansion to predict when Wal-Mart will open in each location. These diverse identification strategies - and the different estimates they produce - are the source of controversy and debate. In the absence of a criticism-proof identification strategy, the precise impact of Wal-Mart on employment and other outcomes remains somewhat uncertain. For a comparison of the store-planning and geographic instruments, see Basker (2006).
Back to top

Business cycles

Wal-Mart's low prices are particularly attractive to consumers during economic troughs. Basker (forthcoming) shows that Wal-Mart sells 'inferior goods' in the technical sense of the term: goods for which demand increases (relative to trend) when incomes fall. This analysis uses quarterly data on average sales per store at Wal-Mart over a ten-year period; the same analysis shows that Target's products are 'normal', meaning demand for them increases when incomes rise. Using a monthly data set and Granger causality tests, Jantzen, Pescatrice and Braunstein (2009) also find that Wal-Mart's sales growth falls when the overall economy's growth rate accelerates.
Back to top

Conclusion

As the largest, and arguably most efficient, chain retailer in the world, Wal-Mart leads the way in both technological innovation and arousing public opposition. Ultimately, however, Wal-Mart is just one of many

中国社会科学院研究生院：研修之用

retail chains making these investments, made possible by the same advances in computing technologies that have transformed other sectors of the economy.

Back to top

Bibliography

Basker, E. 2005a. Job creation or destruction? Labor market effects of Wal-Mart expansion. *Review of Economics and Statistics* 87, 174–83.

Basker, E. 2005b. Selling a cheaper mousetrap: Wal-Mart's effect on retail prices. *Journal of Urban Economics* 58, 203–29.

Basker, E. 2006. When good instruments go bad. Unpublished paper, University of Missouri.

Basker, E. 2007. Causes and consequences of Wal-Mart's expansion. *Journal of Economic Perspectives* 21, 177–98.

Basker, E. forthcoming. Does Wal-Mart sell inferior goods? *Economic Inquiry*.

Basker, E., Klimek, S. and Van, P.H. 2008. Supersize it: the growth of chains and the rise of the 'big box' retail format. Unpublished paper, University of Missouri.

Basker, E. and Noel, M. 2009. The evolving food chain: competitive effects of Wal-Mart's entry into the supermarket industry. *Journal of Economics and Management Strategy* 18, 977–1009.

Basker, E. and Van, P.H. 2008. Putting a smiley face on the dragon: Wal-Mart as catalyst to U.S.-China trade. Unpublished paper, University of Missouri.

Bianco, A. 2006. *Wal-Mart: The Bully of Bentonville*. New York: Doubleday.

Drewianka, S. and Johnson, D. 2009. Wal-Mart and local labor markets, 1990–2005. Unpublished paper, University of Wisconsin–Milwaukee.

Fishman, C. 2006. *The Wal-Mart Effect*. New York: Penguin.

Gladstone, R. 1992. No. 1 retailer slammed in 'Dateline NBC' show, stock falls. Associated Press, 21 December.

中国社会科学院研究生院：研修之用

Hausman, J. and Leibtag, E. 2007a. Consumer benefits from increased competition in shopping outlets: measuring the effect of Wal-Mart. *Journal of Applied Econometrics* 22, 1157–77.

Hausman, J. and Leibtag, E. 2007b. Wal-Mart effects and CPI construction. In *Price and Productivity Measurement*, ed. W.E. Diewert, B.M. Balk, D. Fixler, K.J. Fox, and A.O. Nakamura. Victoria, Canada: Trafford Press.

Jantzen, R., Pescatrice, D. and Braunstein, A. 2009. Wal-Mart and the U.S. Economy. *Eastern Economic Journal* 35, 297–308.

Jia, P. 2008. What happens when Wal-Mart comes to town: an empirical analysis of the discount industry. *Econometrica* 76, 1263–1316.

Leonard, D. 2005. The only lifeline was the Wal-Mart. *Fortune*, 3 October.

Neumark, D., Zhang, J. and Ciccarella, S. 2008. The effects of Wal-Mart on local labor markets. *Journal of Urban Economics* 63, 405–30.

Norman, A. 2004. *The Case against Wal-Mart*. Atlantic City, N.J.: Raphel Marketing.

Vance, S.S. and Scott, R.V. 1994. *Wal-Mart: A History of Sam Walton's Retail Phenomenon*. New York: Twayne.

Weinswig, D. and Tang, C. 2006. Equity research: Wal-Mart Stores, Inc. (WMT). Citigroup Global Markets, 24 April.

Back to top

How to cite this article

Basker, Emek. "Wal-Mart, economics of." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 13 January 2010 <http://www.dictionaryofeconomics.com/article?id=pde2009_W000140> doi:10.1057/9780230226203.1921

Download Citation:

中国社会科学院研究生院：研修之用